SESSION

MODELING AND APPLICATIONS

Chair(s)

TBA

Lifecycle Modeling Methodology for Service Architecture: Framework-of-Frameworks for Context-to-Code Modeling

R. William Maule, Ph.D.
Information Sciences Department, Naval Postgraduate School, Monterey, CA, USA

Abstract - *Applications and services deployed across complex, multi-layered architectures require comprehensive modeling frameworks to effectively manage the software lifecycle—from concept to code. The typical process evolves from high-level concepts to requirements and specifications and then design, development, deployment and maintenance. Too often the modeling process breaks with the code, especially in the deployment and maintenance phases. This makes the understanding of complex architectures exceedingly difficult, especially as we evolve into the "Internet of Everything" with the resultant requirement to address diverse end-user devices. This paper presents a methodology that integrates best practices from government and commercial modeling and provides for integration of commercial tooling to support system and software lifecycles from concept models through to technical engineering models suitable for code generation.*

Keywords: Modeling, Lifecycle, Enterprise, Systems

1 Introduction

Modeling frameworks can support the design and development of complex enterprise systems. Frameworks can be especially advantageous for system-of-systems architectures where a modeler can quickly be overwhelmed with the complexity of integration. Comprehensive frameworks might address distributed enterprises and assist with current efforts to support "bring your own device" (BYOD) in "Internet of Things" (IoT) scenarios—which collectively require that future enterprises be flexible enough to integrate any device yet robust enough to secure every device. This paper examines some of the current issues faced by modelers of diverse enterprise systems and discusses some tools to help mitigate complex modeling problems.

2 Model Frameworks

IEEE Standard 1516-2000 establishes modeling and simulation high-level architecture rules to help ensure consistent framework deployment in the community [1]. For example, the Object Model Template provides a documentation standard for describing data used by a particular model. The Federate Interface Specification describes a generic communications interface for interoperability between simulation models. In addition to industry frameworks are frameworks that are more specific and often aligned with functional areas or disciplines.

Even within specializations or disciples these framework-based models can be developed from a number of complementary perspectives. Object-oriented frameworks work at the code level and offer an efficient means for code reuse [2]. Frameworks have been developed to support multi-dimensional engineering at a macro level that includes mission areas, technical domains, and supporting lifecycles [3]. Model frameworks have been targeted to address issues of abstraction in sequence-based activities [4]. Frameworks have also been narrowly focused to address issues within specific disciplines, such as a framework to support concurrent design refinement from behavioral and functional modeling in rapid prototyping systems architecture [5].

Recognizing the need for frameworks to model components in complex heterogeneous systems, researchers have proposed meta-model frameworks to integrate semantics across models to provide consistent designs [6]. The addition of services into models with modular self-contained components and requirements to support loose coupling, dynamic runtime service discovery, and late binding has given rise to frameworks specific to service architecture [7, 8].

3 Services and Processes

The need to address services within a modeling framework adds to complexity. Service-based engineering tools and algorithms tend toward fine-grain process analysis and offer an additional means for framework and model development. In this context the focus would be on frameworks for component assessment and methods for component performance modeling.

Cumulative or composite services multiply the number of component interactions that must be modeled due to this increased number of components and processes [9, 10]. This need is accentuated in widely distributed system-of-system architectures. Resource availability in composite, cumulative services becomes a primary concern as the paradigm shifts from single process to integrated services [11, 12]. Adding to the complexity of modeling composite services is the cumulative impact of operations in each of the OSI layers, e.g., process, service, routing, transformation, etc. [13, 14].

This requires that attributes be modeled at different levels of abstraction [15] to address coupling between services and clients [16] and to support measurements to include process throughput, reliability and availability [17]. Heuristics must

address the impact of component variations in service reuse across different architectures [18, 19, 20] and the impact of these service interactions on user requirements [21, 22] in component relationships, object interactions, and rules.

4 DoD Model Framework

The U.S. Department of Defense (DoD) uses a highly detailed modeling framework with dozens of integrated models to help with lifecycle modeling. While not sufficiently comprehensive to provide end-to-end lifecycle modeling, from concept to code, the DoD Architecture Framework (DoDAF) provides a base into which other needed frameworks and models can interface. DoDAF applies the IEEE 1471 definition of an architecture description to define a standard approach to describing, presenting, and integrating architecture.

Services represent operational functionality, with information exchanges modeled as data types that traverse functionality through Capability Viewpoints (CV) that model requirements, Data and Information Viewpoints (DIV) to model data relationships, Project Viewpoints (PV) to model relationships between operational and capability requirements in services design; and Services Viewpoints (SvcV) to model performers, activities, services, and their exchanges [23].

Enhancements to DoDAF can help with the sequencing of modeling products [24] by validating consistency in the products [25], enhancing event specification capabilities [26], and providing support for simulation-based testing [27]. Integrations with Activity Based Methodology (ABM), System Modeling Language (SysML), and Business Process Modeling Notation (BPMN) have strengthened DoDAF.

DoDAF-2 shifted the underlying modeling paradigm to data-driven "viewpoints" with an extensible data model [28]. DoDAF-2 also added DoDAF Meta-Model (DM2) which provides a high-level view of data elements and enables the modeler to describe a model in XML (vice a physical/visual model) to speed model exchange to better support reuse of architectural information [29].

Processes for integration with coding frameworks and models such as the Unified Modeling Language (UML), UML Profile-based Integrated Architecture (UPIA), and SOA Modeling Language (SoaML) have been advanced [30]. The research herein extends this research through a first step at a workflow for the integration of models and supporting frameworks as required for a comprehensive end-to-end lifecycle and toward an eventual Framework-of-Frameworks for Context-to-Code (FoF-CtC) modeling.

5 Framework Integration

Any discussion of industry model integration must begin with the Unified Modeling Language (UML). A history of UML is not needed for the readers of this paper but some notes on the extensions as applied to larger meta-models is appropriate. Specifically, the UML Profile-based Integrated Architecture (UPIA) provides a useful extension to DoDAF in a FoF-CtC methodology. UPIA can be integrated with Model-Driven Software Development (MDSD) methodologies to enable generation of operational code from models [31].

The UPIA SOA Design viewpoint can model service specifications, service ports, and service consumers and providers at both operational and system levels of abstraction. UPIA is DoDAF-2 XML-complaint and can import and export DoDAF-2 Physical Exchange Specification (PES) architectural data. Most of the concepts defined in DM2 can be modeled in UPIA and there are utilities for model validation. Models advanced in this paper will implement UPIA to bridge UML and DoDAF and therein provide a data point for framework integration. Comprehensive architecture models will evolve from high-level operational and systems models to low-level code.

The required framework integration for end-to-end lifecycle modeling is further strengthened through SoaML which can address low-level variables for modeling code development. SoaML is an open source specification project from the Object Management Group (OMG) to describe a UML profile and meta-model for services within a SOA. Existing models and meta-models (e.g., TOGAF [32]) for describing system architectures were considered insufficient to describe SOA in a precise and standardized way. UML was too general and needed clarification and standardization specific to SOA. Additionally, a means was required to operationalize SOA as advanced in the OASIS Reference Model (OASIS-RM). SoaML was adopted as a means to provide SOA-specific tooling and to instantiate OASIS-RM.

Researchers have found that SoaML adds value to UML for large-scale SOA deployment [33]. Extensions have been added to provide support for multi-agent systems [34], mobile applications [35], service requirement variability [36], security-aware invocations [37, 38], and pattern propagation between design and performance models [39]. However, while designed to bridge IT and business models, SoaML is somewhat vague regarding implementation methodology to other business-level languages, such as Business Process Modeling Notation (BPMN) [40] – hence, the need for framework integration.

6 Model Integration

The following discussion extends methodologies for model integration through examples, moving from high-level concept models, through performance-based models, to low-level code models. Whenever possible the models and their supporting frameworks are referenced against DoDAF to show how commercial tools and their model outputs can interface in day-to-day, practical usage.

For example, SoaML is modeled through UPIA and referenced against DoDAF to address anticipated future requirements for secure distributed SOA clouds. In other words, UPIA is presented as a means to associate SoaML models within the DoD modeling framework. The author has taken liberties in the application of commercial models to DoD frameworks and this does not represent DoD policy.

First is a commercial modeler that has been adopted and modified to meet DoD-specific requirements. The Joint

Communication Simulation System (JCSS) uses DoD-specific systems nomenclature. JCSS is a Defense Information Systems Agency (DISA)-endorsed Commercial Off-the-Shelf (COTS) Modeling and Simulation (M&S) tool based on the OPNET Modeler product line. The intent in providing this software to the DoD community is to develop a common modeling base across Command, Control, Communications, and Computer (C4) systems communities [41]. As the name implies, this software is focused on communications planning and on simulation of communications effects on networks.

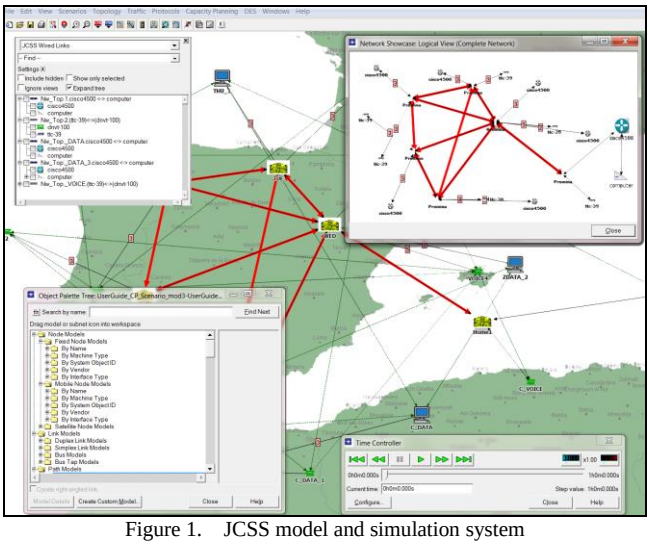


Figure 1. JCSS model and simulation system

In operation, modeling begins with construction of a network topology. Then, information systems, databases, and communication-specific devices are added—such as routers, satellite antennas, multiplexors, etc. Expected transmission capacity is modeled, including bandwidth that will be available and expected demands on that bandwidth. Simulations using "what-if" analysis help determine if communications capacity is sufficient to accommodate expected network traffic. Of course, this implies that the modeler has access to accurate data—such as past network performance for similar environmental constraints and transmission conduit, the load generated by the applications, and the degree of latency tolerated by the systems.

JCSS can generate reports to help analyze models with utilization statistics, failure reports, and network optimization analysis. Discrete event simulation can be run as a "state machine" in which change in the state of a machine, network connection, or transmission capacity can help predict impact on systems and applications—with traffic metrics for jitter, latency, and queuing delay.

JCSS reports can be rendered into a DoDAF OV-3 Operational Information Exchange Matrix or DoDAF SV-6 Systems/Services Data Exchange Matrix. The higher-level views generated in JCSS can serve as DoDAF OV-1 High-Level Operational Concept Description models, and the more detailed views can present DoDAF SV-2 Systems Communications Description models (see figure 1).

Another tool, not technically a DoDAF modeler but a used regularly within the DoD community to help understand complex mission-based relationships is the System Tool Kit (STK) from AGI Corporation which excels in model development for space defense industries and is widely used for satellite and aircraft modeling and simulation. Recent product enhancements have introduced opportunities to include air, land, and sea assets in the models with high-resolution 3-D visual simulation [42]. As an output format in a DoD environment the STK would be used for operational modeling.

If properly constructed, through STK, we can create precise, realistic events. The reports that we can generate are more in the area of resource allocation vice the JCSS communication performance reports. In a perfect world, with time permitting, we would use both tools. We would model then simulate the scenario in STK with primary resources and assets for 3D visual analysis, and we would simultaneously model communication specifics within JCSS. We would use STK for asset evaluation—for example, Unmanned Aircraft System (UAS) sensor packages within geographic areas. Or, area coverage of Unmanned Aerial Vehicle (UAV) flight patterns for search and rescue. Or, analysis of communications coverage for mobile users in isolated terrain with geo-stationary satellite constellations supplemented by Beyond Line-of-Sight (BLOS) radios. Prior to execution we would model the communications in JCSS and during execution capture the traffic data, import that data to JCSS, and use that data to refine our STK event simulations and resource allocation models.

STK does not render directly to DoDAF but the visual representations can be captured for OV models. The reports can be used as data points for DoDAF resource matrix models. Figure 2 shows the STK user interface with rendering in 2D and 3D windows—which are active for both modeling and simulation.

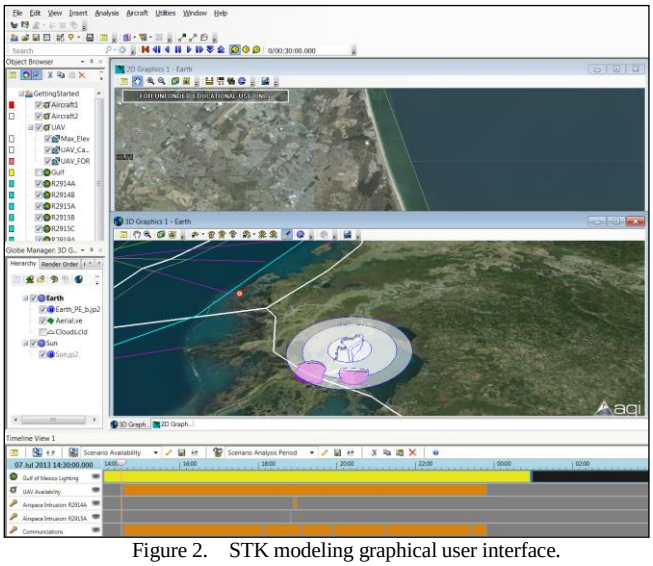


Figure 2. STK modeling graphical user interface.

We can integrate STK with a large number of complementary modeling packages for an easy exchange of

ideas, such as Keyhole Markup Language (KML) for visualization to Google Earth. Or, we can export model specifications to the open source System Modeling Language (SysML) for systems engineering specification and model validation. We can use QualNet software from Scalable Network Technologies to capture live communications data into STK simulations. As such, STK fits nicely within the concept of an end-to-end lifecycle modeling framework.

As referenced earlier, UML-based modeling tools, suites and frameworks would be a critical component of a comprehensive end-to-end lifecycle analysis. These tools and their models represent the far end of the modeling spectrum—code generation. So, in evaluating a FoF-CtC approach, modelers such as STK would simulate the concept or mission and receive performance data from QualNet. JCSS and the supporting OpNet modelers and tool suites would occupy the middle tier for communications performance assessment. UML-based frameworks and modeling suites would be for low-level code modeling. In the FoF-CtC proposed herein all are integrated with DoDAF-2 such that each model and approach is mapped to a corresponding DoDAF model or viewpoint.

While UML itself is an open standard and governed accordingly, the company that brought UML to market is Rational which is now part of IBM. The System Architect (SA) suite of modeling tools from Telelogic is also now part of IBM, positioned under Rational. SA has the most robust support for DoDAF of any model tooling available and excels at the lower-level modeling. In the FoF-CtC, SA will serve as the technical means to integrate low-level system and service models with UML-compliant code-generation models as applied to DoDAF.

Figure 3 shows the DoDAF model options within the SA graphical user interface. There are two SA versions optimized for DoDAF architectures: SA for DoDAF with the MITRE Activity Based Method (ABM) option, and SA C4ISR—which has been renamed to SA for DoDAF—to build open architecture models around structured IDEF techniques. Integration DEFinition (IDEF) models are also significant in the DoD and would be a component of a FoF-CtC. Similar to DoDAF, IDEF models range from high-level functional models to low-level object-oriented models.

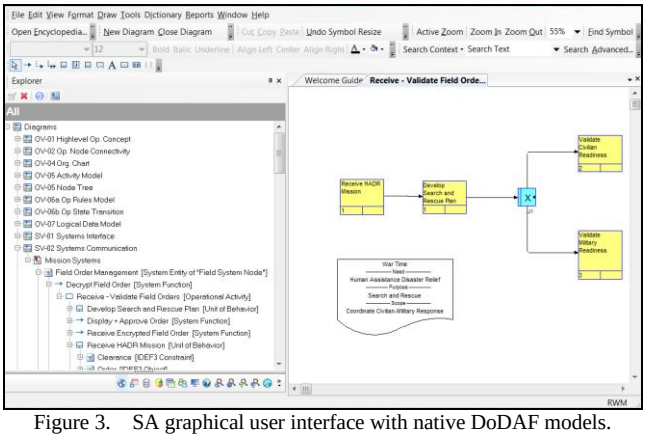


Figure 3. SA graphical user interface with native DoDAF models.

Rational Software Architect (RSA) supports the software engineering phase of code modeling and is also built on UML and is capable of DoDAF-like models but without pre-defined DoDAF output modes such as SA. In other words, we need to understand the DoDAF model that we are designing and then use RSA to develop that model vice the SA process where we would select the DoDAF model and then automatically receive the necessary tooling. RSA is built on the Eclipse open-source software framework which is the industry standard for Model-Driven Development (MDD) in both the DoD and commercial sectors.

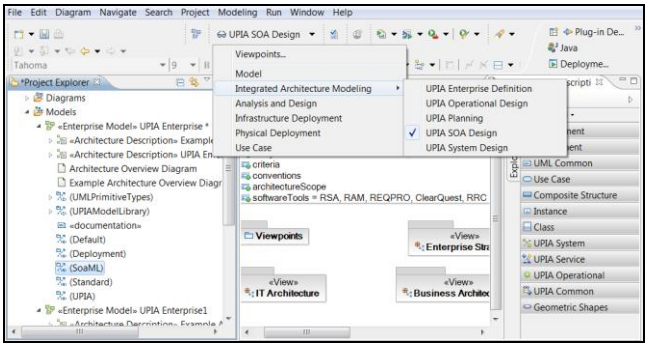


Figure 4. RSA UPIA and SoaML modeling options.

Figure 4 shows the RSA user interface in the traditional Eclipse design pattern—with the model in the center and the development tooling around the model. In the figure we have selected Integrated Architecture Modeling (IAM) with UPIA SOA design options to develop SoaML models—as indicated in the selection in the left column. RSA can be applied in the requirements specification phase and models simulated to help communicate system dynamics and evaluate software against different network topologies [43].

Another level of modeling valuable in FoF-CtC lifecycle assessment but without a direct correlation to DoDAF viewpoints addresses Return on Investment (ROI) analysis wherein processes are modeled, assigned cost variables, simulated and assessed. Figure 5 shows a Business Process Management (BPM) model specific to a SOA processes using WebSphere Business Modeler (WBM) [44]. The simulation component includes a module that calculates ROI for a proposed innovation—in this instance, for a complex role- and attribute-based security process.

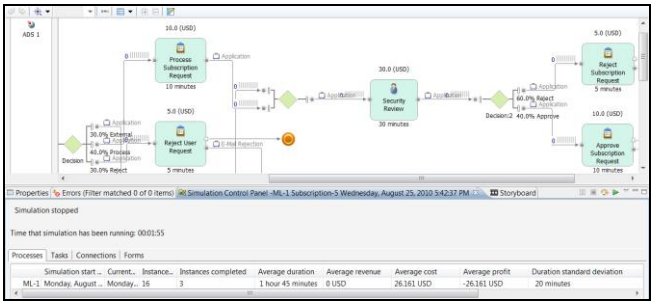


Figure 5. WBM in simulation mode with ROI analysis.

7 DoDAF Integration

A discussion of DoDAF is beyond the scope of this paper. There are many excellent online references, and the source DoDAF site is available online. The framework is extensive with multiple models in model categories that address capabilities, data and information, operations, projects, services, standards, and systems [45].

The intent herein is to use DoDAF as a baseline for framework integration since it seems to be the most comprehensive framework available. Yet, DoDAF is insufficient at the end-to-end lifecycle extremes, both at the highest conceptual levels and the need to visualize and simulate complex mission threads, and at the other extreme for low level code development, deployment, monitor, management and maintenance.

As such, DoDAF serves as the baseline and our starting point for FoF-CtC lifecycle integration. Since our interests for this paper are in service architecture, a couple of examples that apply the commercial tooling, models and frameworks previously discussed to DoDAF will be given.

To start, figure 6 presents a UML use-case for a distributed service architecture. While not really a UML use-case, and not really a DoDAF Operational Viewpoint (OV), the integration provides value in our FoF-CtC integration since a parallel development of our DoDAF modeling with the commercial tool suite for code development will enable an evolution from concept to code—all within the confines of a reference framework.

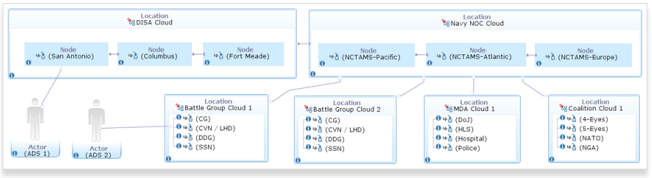


Figure 6. Use-case for secure distributed service architecture.

The System Viewpoint (SV)-2 is a prominent DoDAF model to show how a system operates and its interfaces. In the SV-2 we model data flows between systems, identify the protocols and the networks, and specify the system ports. A model may be created for each resource flow, or we can model all resource flows on one diagram. For example, figure 7 models a Navy cloud which consists of telecommunications centers and user systems. Within the cloud are nodes which provide services. A source node hosts authoritative data sources and physical servers host virtual machines.

Services Viewpoint (SvcV) models offer a precise means to model interactions in a service architecture. In these models we visualize interactions and describe services and resources required for development and execution [46]. As such, the SvcV series of models become a backbone for a FoF-CtC service architecture lifecycle.

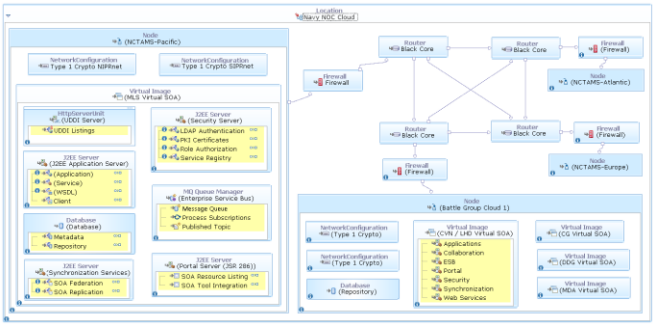


Figure 7. SV-2 for backbone cloud with subordinate clouds.

SvcV models can articulate use-case performers, system activities, and data exchanges in support of operational capabilities. Both structural and behavioral factors, and the impact of such factors on the architecture, can be modeled. The SvcV models work hand-in-hand with the SV models discussed earlier. The SV models focus on system and component interaction, the SvcV models a step lower to service, object and process interaction.

A final example of framework integration is the hybrid model in figure 8 which builds from a SvcV-10b to model a resource (or function) response to events—taking action to move to a new state as a function of a current state—with each transition an event and an action. This can be correlated with a UML state chart to model change in a sequence of service functions. Behavior is modeled as the traversal of a graph of specific states, all interconnected by transition arcs that are triggered by events. A SoaML model represents the SvcV-10b state machine. State change in this instance refers to change in our SOA due to changes in data elements or attributes—such as publication topics or subscriptions—or changes in security for any process.

The ovals in the diagrams are UML collaboration use-cases for participant interaction—which is where our state change occurs. The connection boxes with handles we use to identify service interface ports and protocols. The service contract is the means to ensure QoS and enforce contracts. The collaboration shows how participants work together to provide services. Each participant plays a role in the service contract to help verify and ensure that underlying constraints are honored.

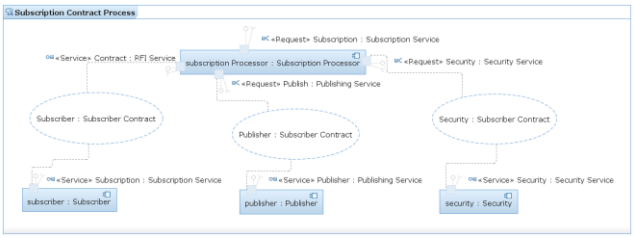


Figure 8. SvcV-10b SoaML service state diagram.

	Concept	Design	Architect	Code	Deploy	Monitor	Manage
	Objectives Requirements Specifications	*Network Systems Services*	*Data Flows Systems Services*	*Data Models System Specifications Service Properties*	*Budgets Timelines Personnel*	*Performance Interoperability Throughput, Latency*	*Code Systems Applications*
DoDAF	Operational Standards Viewpoints	Information Capability Viewpoints	Systems Services Viewpoints	Data Viewpoints	Project Viewpoints		
UML	Use-Cases	Activity Diagrams	States Interactions Sequences	Classes Objects Components	Deployment Diagram		
Tools		Mission Visualization Scenarios	BPM BPMN	BPEL SoaML UPIA		BAM Packet Capture	System Management Tools

Figure 9. Framework-of-Frameworks Concept-to-Code Lifecycle Methodology

8 FoF-CtC Lifecycle

Figure 9 provides the cumulative FoF-CtC integration lifecycle as developed throughout this paper and presented as an initial step toward model and framework integration across disciplines and between modeling communities. The intent is to address limitations of current approaches which do not really provide a comprehensive end-to-end modeling solution from concept to code—to include deployment, monitor, performance assessment and maintenance.

In the examples presented, UML-based frameworks were integrated with the DoD framework as a step toward comprehensive end-to-end modeling. For example, a SvcV model represented data flows in standard DoD notation, and then a corresponding UML model was integrated to address tooling requirements for the generation of computer code from the models. Together the integrated models, along with complementary tooling and frameworks, become the core of our service architecture and a baseline for FoF-CtC lifecycle integration. Supporting models and commercial tools were presented as extensions that provide visualizations of the capabilities advanced in the collective models, as represented in the FoF-CtC lifecycle.

Future research might integrate additional frameworks or commercial tooling to fill identified gaps in the current FoF-CtC or strengthen the base of applicable models in the identified categories. Additionally a workflow for model application, drawing from the FoF-CtC, might present a means for ready integration and usage of models across system and software development lifecycles.

9 Conclusion

The FoF-CtC lifecycle methodology provides a means to model from high-level operational concepts to an integrated enterprise, evolving from goals or objectives to modeling specifics which realize specific functions, processes and services. Examples were presented to illustrate a means to integrate government and commercial modeling approaches to optimize the strengths of each and begin a dialog on framework and model integration.

10 Acknowledgment

The author wishes to acknowledge OPNAV for support of this research.

11 References

[1] IEEE-SA Standards Board. "IEEE Standard for Modeling and Simulation (M&S) High Level Architecture (HLA) — Framework and Rules," New York: Simulation Interoperability Standards Committee (SISC) of the IEEE Computer Society, 21 September 2000.

[2] Zhang, T., Xiao, X., Wang, H. and Qian, L. "A Feature-Oriented Framework Model for Object-Oriented Framework: An MDA Approach," IEEE Ninth International Conference on Computer and Information Technology, 2009, pp. 199-204.

[3] Wells, B. "A Multi-dimensional Hierarchal Engineering Competency Model Framework," 2008 2nd Annual IEEE Systems Conference, 2008, pp. 1-6.

[4] Sun, S., and Wang, N. " Formalizing the Multiple Abstraction Process within the G-KRA Model Framework," 2010 International Conference on Intelligent Computing and Integrated Systems (ICISS), 2010, pp. 281-284.

[5] Ragavan, S., Shanmugavel, M., Shirinzadeh, B., and Ganapathy, V. "Unified modelling framework for UAVs using Bond Graphs," 2012 12th International Conference on Intelligent Systems Design and Applications (ISDA), 2012, pp. 21-17.

[6] King, C., and Alexander, P. "The Rosetta Meta-Model Framework," 10th IEEE International Conference and Workshop on the Engineering of Computer-Based Systems, 2003, pp. 133-140.

[7] Kaviani, N., Mohabbati, B., Lea, R., Gasevic, D., Hatala, M., Blackstock, M. "ReCoIn: A framework for dynamic integration of remote services into a service-oriented component model," IEEE Asia-Pacific Services Computing Conference (APSCC), 2009, pp. 502-507.

[8] Yuan-sheng, L., Xiao, H., Yu, W., and Si-xin, S. "Research on a Web Services Discovery Model Framework, 2010 International Conference on Computational Intelligence and Software Engineering (CiSE), 2010, pp. 1-4.

[9] Parsons, T., Mos, A., Trofin, M., Gschwind, T., and Murphy, J. "Extracting Interactions in Component-Based Systems", IEEE Transactions on Software Engineering, Vol. 34, No. 6, 2008, pp. 783-799.

[10] Liu, G., Zhu, Z., Li, Y., Li, D., and Cui, J. "A New Web Service Model Based on QoS", International Symposium on Intelligent Ubiquitous Computing and Education, 2009, pp. 395-399.

[11] Luo, J., Li, Y., Pershing, J.; Xie, L., and Chen, Y. "A Methodology for Analyzing Availability Weak Points in SOA Deployment Frameworks", IEEE Transactions on Network and Service Management, Vol. 6, No. 1, 2009, pp. 31-44.

[12] Maule, R. "SoaML and UPIA Model Integration for Secure Distributed SOA Clouds", Proceedings of the IEEE 2012 World Congress on Services Computing (SERVICES 2012), 24-29 June, Honolulu, HI, 2012.

[13] Lee, Y. "Event-driven SOA Test Framework Based on BPA-Simulation," First International Conference on Networked Digital Technologies, 2009, pp. 189-194.

[14] Her, J., Choi, S., Oh, S., and Kim, S. "A Framework for Measuring Performance in Service-Oriented Architecture", Third International Conference on Next Generation Web Services Practices (NWeSP 2007), 2007, pp. 55-60.

[15] Shim, B., Choue, S., Kim, S., and Park, S. "A Design Quality Model for Service-Oriented Architecture", 15th Asia-Pacific Software Engineering Conference (APSEC 2008), 2008, pp. 403-410.

[16] Xiao-jun, W. "Metrics for Evaluating Coupling and Service Granularity in Service Oriented Architecture", International Conference on Information Engineering and Computer Science (ICIECS 2009), 2009, pp. 1-4.

[17] Gao, J., Wu, Y., Chang, L., and Meldal, S. "Measuring Component-Based Systems using a Systematic Approach and Environment", Second IEEE International Workshop on Service-Oriented System Engineering (SOSE 2006), 2006, pp. 121-129.

[18] Kumari, G., Kandan, B., and Mishra, A. "Experience Sharing on SOA Based Heterogeneous Systems Integration", 2008 IEEE Congress on Services, Honolulu, HI, 7-11 July 2008.

[19] Hau, T., Ebert, N., Hochstein, A., and Brenner, W. "Where to Start with SOA: Criteria for Selecting SOA Projects", Proceedings of the 41st Hawaii International Conference on System Sciences, Waikoloa, HI, 7-10 January 2008.

[20] Roach, T., Low, G., D'Ambra, J. "CAPSICUM–A Conceptual Model for Service Oriented Architecture", 2008 IEEE Congress on Services, Honolulu, HI, 7-11 July 2008.

[21] Choi, S., Her, J., and Kim, S. "Modeling QoS Attributes and Metrics for Evaluating Services in SOA Considering Consumers' Perspective as the First Class Requirement", IEEE 2nd Asia-Pacific Services Computing Conference, Tsubuka Science City, Japan, 11-14 December 2007.

[22] Maule, R. "Quality of Service Assessment in SOA Synchronous Networked Communications", Proceedings of the 2007 International Conference on Computing, Communications and Control Technologies (CCCT 2007), Orlando, FL, 12-15 July 2007.

[23] URL:http://en.wikipedia.org/wiki/Department_of_Defense_ Architecture_Framework#DoDAF_V2.0_Viewpoints

[24] Thal, A., Havlicek, J., Chambal, S., and Osgood, J. "Sequencing the Development Order of Architecture Products: An Application to DoDAF", Proceedings of the 2010 43rd Hawaii International Conference on System Sciences System Sciences (HICSS), IEEE Press, 2010, pp. 1-10.

[25] Zhiping, J., Hongda, W., Ming, H., and Guangyu, B. "The Consistency Validation Method of the DODAF Described Models Based on Meta-Data Model", Proceedings of the International Conference on E-Business and E-Government (ICEE), 2011, pp. 1-6.

[26] Zeigler, B., and Mittal, S. "Enhancing DoDAF with a DEVS-based System Lifecycle Development Process", Proceedings of the IEEE International Conference on Systems, Man and Cybernetics. Vol. 4, 2005, pp. 3244-3251.

[27] Baumgarten, E., and Silverman, S. "Dynamic DoDAF Power Tools", Proceedings of the 2008 IEEE Military Communications Conference (MILCOM), 2008, pp. 1-6.

[28] Hughes, T., and Tolk A. "Orchestrating Systems Engineering Processes and System Architectures within DoD: A Discussion of the Potentials of DoDAF", Journal of Reliability, Maintainability, & Supportability in Systems Engineering, Winter 2013, pp. 13-19.

[29] Mittal, S., and Martin, J. Netcentric System of Systems Engineering with DEVS Unified Process. Boca Raton, FL: CRC Press, 2013.

[30] Xing, P., Baoshi, Y., and Jianmi, H. "Modeling and Simulation for SoS based on the DoDAF Framework", Proceedings of the 9th International Conference on Reliability, Maintainability and Safety (ICRMS), 2011, pp. 1283-1287.

[31] IBM. UML Profile-Based Integrated Architecture (UPIA) DoDAF 2.01 Mapping Reference, June 2010, URL: http://www.ibm.com/developerworks/wikis/download/attachments/133497801/UPIA_DoDAF2.01_Reference_v8.0.pdf

[32] The Open Group. The Open Group Architecture Framework (TOGAF), URL: http://pubs.opengroup.org/architecture/togaf8-doc/arch/

[33] Todoran, I., Hussain, Z., and Gromov, N. "SOA Integration Modeling: An evaluation of How SoaML Completes UML Modeling", Proceedings of the 15th IEEE International Enterprise Distributed Object Computing Conference Workshops (EDOCW), 2011, pp. 57-66.

[34] Jacobi, S., Hahn, C., and Raber, D. "Integration of Multiagent Systems and Service Oriented Architectures in the Steel Industry", Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), Vol. 2, 2010, pp. 479-482.

[35] Ali, N. and Babar, M. "Modeling Service Oriented Architectures of Mobile Applications by Extending SoaML with Ambients," Proceedings of the 35th Euromicro Conference on Software Engineering and Advanced Applications (SEAA 2009), 2009, pp. 442-449.

[36] Abu-Matar, M., and Goma, H. "Feature Based Variability for Service Oriented Architectures," Proceedings of the 9th Working IEEE/IFIP Conference on Software Architecture (WICSA), 2011, pp. 302-309.

[37] Hoisl, B., and Sobernig, S. "Integrity and Confidentiality Annotations for Service Interfaces in SoaML Models", Proceedings of the Sixth International Conference on Availability, Reliability and Security (ARES), 2011, pp. 673-679.

[38] Ayed,G., and Ghernaouti-Helie, S. "Digital Identity Management within Networked Information Systems: From Vertical Silos View into Horizontal User-Supremacy Processes Management", Proceedings of the 14th International Conference on Network-Based Information Systems (NBIS), 2011, pp. 98-103.

[39] Mani, N., Petriu, D., and Woodside, M. "Studying the Impact of Design Patterns on the Performance Analysis of Service Oriented Architecture", Proceedings of the 37th EUROMICRO Conference on Software Engineering and Advanced Applications (SEAA), 2011, pp. 12-19.

[40] Sadovykh, A., Desfray, P., and Elvesaeter, B., Berre, A., and Landre, E. "Enterprise Architecture Modeling with SoaML using BMM and BPMN - MDA Approach in Practice," Proceedings of the 6th Central and Eastern European Software Engineering Conference (CEE-SECR), 2010, pp. 79-85.

[41] DISA. Joint Communication Simulation System. Available at http://disa.mil/Services/Enterprise-Engineering/JCSS.

[42] AGI. Systems Tool Kit. Available: http://www.agi.com.

[43] Mohlin, M. (2010). Model Simulation in Rational Software Architect: Simulating UML Models. Cupertino, CA: IBM.

[44] IBM. (undated). IBM WebSphere Business Process Management Version 7.0 Information Center. Available: http://pic.dhe.ibm.com/-infocenter/dmndhelp/v7r0mx/index.jsp

[45] DoD CIO. DoDAF Architecture Framework Version 2.02. Available: http://dodcio.defense.gov/dodaf20/dodaf20_background.aspx.

[46] DoD CIO. DoD Architecture Framework Version 2.0: Volume 1: Introduction, Overview and Concepts. Available: http://jitc.fhu.disa.mil/jitc_dri/pdfs/dodaf_v2v1.pdf.

A Meta Model for Predictive Analysis of Modifications on HPDC Infrastructures

Christian Straube[1], **and Dieter Kranzlmüller**[1]
[1]Munich Network Management (MNM) Team
Ludwig-Maximilians-Universität München, Germany
[straube,kranzlmueller]@nm.ifi.lmu.de

Abstract— *Modifications of High Performance Distributed Computing (HPDC) infrastructures can have intended and unintended effects. The more complex the infrastructure or the modification itself, the more difficult are predictions about their consequences. Understanding modifications on infrastructures requires accurate planning, which can be supported by corresponding predictive models and analysis activities. The analysis needs to incorporate multiple attributes and their dimensions, and can itself become very complex. This paper presents mandatory criteria for a meta model to facilitate simulating modification effects on HPDC infrastructures. The model itself represents the original infrastructure and the intended changes, and delivers the predicted behavior after the modifications have been introduced. We describe the model and its utilization for HPDC infrastructures and an outlook on potential applications of the model.*

Keywords: HPDC Infrastructure, Analytic, Modeling, Optimization

1. Introduction

Hardware-related modifications, such as changes to the CPU's clock rate or the throughput of the interconnect, induce two groups of effects: the (by definition) positive intended effects and the often negative but unavoidable side effects [1]. This conflict is especially challenging for typically very complex *High Performance Distributed Computing* (HPDC) infrastructures, as the following example illustrates: given a typical HPDC infrastructure, increasing redundancy to address short-time breakdown and to improve reliability (positive intended effect) [2], simultaneously increases energy consumption and degrades performance due to redundancy overhead (negative side effects) [3]. As side effects can have serious impact, analysis and prediction of modification effect conflicts should receive special attention during modification planning [4]: a *predictive modification effect analysis* (PMEA) could avoid a harmful execution or a costly but spare modification exploration, if it is investigated in advance, whether the (negative) side effects will outweigh the (positive) intended effects or in other words, whether the effects will lead to an improvement with respect to the entire HPDC infrastructure. *How* to finally accomplish the modifications, can be neglected in this consideration and passed over to expert groups who apply conventional approaches, like focused analytic [5] or simulation [6] models.

PMEA addresses two **problems**: 1) Most HPDC infrastructure attributes are combined by several aspects, e.g., system performance is built upon computing cores, communication interconnect, and I/O performance [7]. If only the intended effects are analyzed, possible negative side effects may be overlooked. 2) Contemporary HPDC infrastructures provide functionality through complex qualitative and quantitative component interplay. *High Performance Computing* (HPC) and distributed systems, each built from manifold (often heterogeneous) compute, storage, interconnect, and other specialized sub components, interact to deliver sophisticated and non-trivial computation and storage functionality. This close interplay hardens identifying the specific contribution of a single component to HPDC infrastructure's functionality [8], and hence, predicting effect cascading within the HPDC infrastructure is a difficult task.

The **challenge** of facilitating PMEA is the development of a sufficient model. The general need for a model is subsequently justified: PMEA must cover the entire HPDC infrastructure in a domain-spanning way [9]. Even if there are predictions based on empirical data, like performance prediction [10], those data neither cover the entire HPDC infrastructure nor contain all domains of the intended and side effects. The necessity of (empirical) data could by avoided by physically testing a modification before it is executed, but HPDC infrastructures central role for scientist's work prohibits testing on productive systems [11], [12], and their scale and complexity make physically copying hardly possible. Hence, a *predictive* modification effect analysis requires a *model*, which must fulfill a series of demands, e.g., covering the entire HPDC infrastructure to avoid a delusive interpretation or addressing the aforementioned HPDC infrastructure complexity. Furthermore, consideration of both effects groups results in a complex optimization problem spanning multiple attribute dimensions, mostly involving but not limited to monetary and algorithmic *cost*, *energy efficiency*, *reliability*, and *performance* [12]. Fostering (a seamless) integration into an existing model and tool landscape and keeping the model in sync with the considered real

world HPDC infrastructure require interfaces for information exchange.

The **contribution** of this paper is two-fold: 1) derivation of a criteria set for a model to facilitate the outlined PMEA, and 2) development of a meta model that provides language constructs to model arbitrary HPDC infrastructure settings compliant to the derived criteria set. To the contrary, accomplishing content to be stored by the model is discussed in other work, like attribute calculation and aggregation [1], or currently under investigation, like system workload calculation. The meta model is currently used to conduct PMEA onto the the *SuperMUC*, a 3 PetaFlop/s HPC system in the top 10 of the Top 500 Supercomputer Sites list. The paper is structured as follows: Section 2 fosters a common understanding by providing some non-formal definitions. Section 3 develops a set of model criteria, which are employed to develop the meta model presented in Section 4. Section 5 examines existing approaches and compares them to our approach, before Section 6 concludes the paper and presents future work.

2. Terminology and Investigated Elements

The lack of a characteristic feature set to delimit HPDC infrastructures and related topics [13] requires ensuring a common understanding of important terms and concepts. The subsequently provided non-formal definitions aim at fostering this common understanding, and not at developing a universal definition.

An **HPDC infrastructure** is a distributed system, consisting of several heterogeneous, geographically distributed components, which are connected via communication components. Examples for HPDC infrastructures are the TeraGrid [14] or PRACE [15]. HPDC infrastructure components can be either *atomic components*, like a CPU, or *composite components*, like a compute node, a compute island or even an entire data center. Since composite components can be defined at a very high level, HPC infrastructures, like the SuperMUC, are implicitly a subset of HPDC infrastructures and hence, are also considered. HPDC infrastructures are described by the three following terms: *capability*, *attribute*, and *property*. A **capability** is a well-defined functionality the HPDC infrastructure exposes to a user or (scientific) application. They describe low-level functionality, like computation, storage or transfer. High-level capabilities, like sophisticated *reliable* file transfer, are in our understanding *services* that are built on top of an HPDC infrastructure. **Attributes** describe the quality of the exposed capabilities, e.g., *performance*, *energy efficiency*, or *reliability*. A **property** represents a specific aspect of the HPDC infrastructure, which is influenced by the applied hardware and software configuration. Properties can be used as "tool" to influence attributes, e.g., *latency* is a property that can be modified to alter the attribute *performance*. In summary, modifications target at properties and alter hardware and software configurations. Thereby they have an effect on attributes of the HPDC infrastructure capabilities.

Scientific applications are executed on HPDC infrastructures and usually consist of several heterogeneous and distributed (specialized) applications, processes, and data items, which are mostly contributed by different cooperating disciplines [13], [16]. These elements are composed, orchestrated, and managed by *scientific workflows*. The close relationship of scientific applications and scientific workflows causes a disappearance of separation, resulting in a fusion of both terms, e.g., bioinformatics applications can be viewed as bioinformatics workflows [16]. Following this direction, we use both terms synonymously. We focus on scientific applications as they mostly 1) describe complex workflows and 2) use HPC systems. Both characteristics put high demands on the HPDC infrastructure and our model and include several less complex systems.

System workload is caused by executing a scientific application on an HPDC infrastructure. Since PMEA is conducted on a model (cf. Section 1), workload is also considered theoretically. It must be calculated for every application execution on a particular HPDC infrastructure [17], because the same scientific application might cause different system workloads on different HPDC infrastructures, dependent on their properties.

3. Mandatory Model Criteria

To provide a benchmark for related work analysis and instructions for model development, this section develops a criteria set (*C1–C5*), strongly guided by 1) the general aim of modeling, i.e. getting a *valid* model ("a model that is sufficiently accurate for the purpose at hand" [18]), 2) the objective of our work, i.e. facilitating a predictive analysis of modification effects onto HPDC infrastructures, and 3) the aim of keeping the model as straight forward as possible.

C1 – Real World Element Selection Our "purpose at hand" is a predictive and domain-spanning evaluation of the expected modification effects onto the entire HPDC infrastructure, including effect cascading and interplay. Hence, achieving a valid model requires coverage of the subsequently itemized real world elements.

> **C1.1** *Component types and scale* – The unpredictability of effect cascading (cf. Section 1) and the strong influence of communication components on attributes [19] bans focusing on particular component types or sub systems. Instead, all components must be considered, independently of their types and administrative and geographic scale.
>
> **C1.2** *Redundancy* – As redundancy "allows a function to be performed on more than one node" [20] it is one of the major tools to improve HPDC infrastructure resilience [2]. Aiming at a valid model requires coverage

of redundancy in a sufficient level of detail, which is in particular not achieved by summing up redundant components to one component, but by providing enhanced redundancy characterization facilities, e.g., to describe boolean functions that define component availability [3].

C1.3 *System Workload* – Nearly all HPDC infrastructure attributes are influenced by system workload [21], e.g., system's response time (performance attribute) is correlated to workload [22], or energy consumption/efficiency (energy attribute) can be calculation based upon system workload [23].

C1.4 *Capabilities* – Even if capabilities describe the exposed functionality of an HPDC infrastructure (cf. Section 2), they are of minor importance to PMEA and therefore can be considered only coarsely.

C1.5 *Properties* – A property vector per component, e.g., describing latency and throughput of a communication component, is necessary for system workload and attribute calculation. The former applies the vector for mapping application elements to HPDC infrastructure components, the latter incorporates the vector into attribute values: considering the attribute *performance*, for instance, calculates the transfer time for a given amount of data and the available maximum throughput. In this way, properties come up to being the "tools" to influence attributes (cf. Section 2).

C1.6 *Attributes* – There are manifold ways to calculate attribute values, e.g., applying a probability distribution like Weibull or Pareto, or using an arbitrary mathematical function like the bathtub curve. Since there is no "one-size-fits-all" calculation rule, but all ways have specific advantages dependent on the objective [1], calculation rule modeling should be as generic as possible to cover at least most of them.

C2 – Multiple Granularity Levels Different aims might require analysis on different levels of detail, e.g., considering a compute node very detailed, while other components remain at an abstract level. Increasing model granularity to the desired level globally would quickly lead to an unmanageable complex model due to HPDC infrastructures' scale and diversity. Hence, multiple levels of granularity within the same model are mandatory to facilitate a good trade-off between accuracy, complexity and time to solution [12].

C3 – Reusability PMEA consists of several tasks (cf. Section 1 and 2), e.g., HPDC infrastructure topology description or attribute calculation. The re-use of task results, e.g., applying the same attribute calculation rules to different HPDC infrastructure component types or using the same HPDC infrastructure model for different scientific applications, would be extremely hindered by mixing up fields of different tasks and hence, a *Separation of Concerns* (SoC) is required. An ancillary effect of applying SoC is the simplified independent adaption of different (model) parts, which can change differently over time. For instance, the model part describing HPDC infrastructure components can be enhanced to respect the fast development of microprocessor technology [24], while keeping the attribute calculation rules.

C4 – Interfacing The high dynamic of real world HPDC infrastructures requires an interface to existing management and description approaches and tools to ease keeping the model up to date. An interface would also be beneficial to interact with special purpose approaches, like discrete event simulators, and hence, gaining acceptance.

C5 – Simplicity Since a model's size or bulkiness can outweigh its benefits [18], even if it is perfectly valid, the model should be as simple as possible. Besides a reduced run time, further advantages are a fast development, easy result interpretation [25], and being less error prone [26].

4. A Meta Model for Predictive Modification Effect Analysis

Implementing the required SoC (*C3(✓)*), our model is split up into three parts, i.e. the *HPDC infrastructure components*, *property and attribute description*, and *analysis execution*, which are described in detail in Section 4.1. The concepts of the model are formalized in a meta model in Section 4.2.

4.1 Model Parts in Detail

Model Part 1 – HPDC Infrastructure Components

Fig. 1 depicts the usage of our HPDC infrastructure component model to describe a simple exemplary infrastructure. It illustrates the tree structure, interfacing, and the black box approach, as explained below.

Tree-based HPDC Infrastructure Component Modeling – HPDC infrastructure components and topology are represented by a graph-based tree, which supports or even facilitates some features of our approach, e.g., the black box approach or attribute aggregation, as subsequently explained. Additionally, it fosters interaction with other models, like workflow models, which are mainly also represented as graphs.

Tree *vertices* represent type-independent all *atomic components*, like local memory or internal bus, which ensures an equal consideration of *all* components during PMEA (*C1.1(✓)*). Consequently, *edges* represent *logical communication dependencies* between components, instead of physical communication hardware or communication components. Since the lion's share of communication is bi-directional and for the sake of simplicity, edges are non-directed. Composite components are described by a sub tree, containing all atomic components that compose the composite component. This is depicted in Fig. 1: the composite component *Compute Node*, depicted by the dotted rectangle on the left side, is composed of five atomic components, i.e. *Local Memory*,

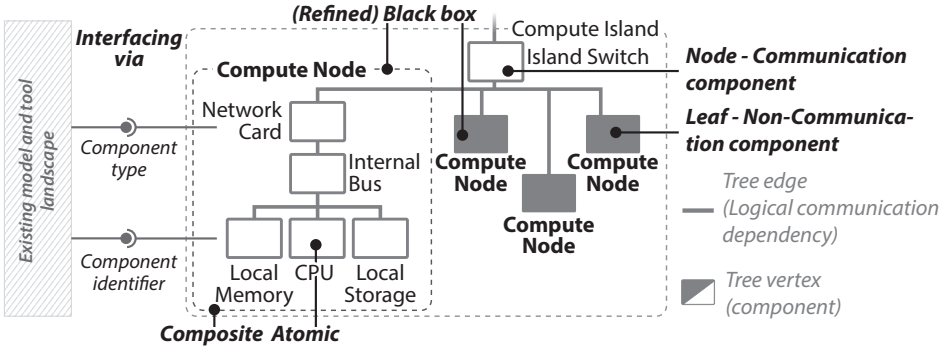


Fig. 1: Exemplary usage of the presented HPDC infrastructure model to describe a simplified infrastructure. It illustrates the model's key features, i.e. a tree structure, interfacing to other models, and a black box approach.

CPU, *Local Storage*, *Internal Bus*, and *Network Card*, depicted by the white boxes within the rectangle.

Coarsely following the concept of syntax trees, communication components apply the *Connect* operation to other components, which means that communication components are regarded as *operators*, and all connected (leaf) components as *operands*. Consequently, nodes in the tree exclusively represent communication components, which connect two or more components to enable information exchange. For instance, the *Island Switch* (operator) on the right side in Fig. 1 connects several *Compute Nodes* (operands) and enables information exchange between them. The following informal proof shows that this premise can always be fulfilled ($n = |vertices|$): $n = 1$ – The tree contains only a root component. $n = 2$ – This situation is more or less meaningless, because two communicating components (operands) need at least one additional component for interaction (operator). $n = 3$ – A simple tree, containing a communication component as root and two communicating components as leafs. $n > 3$ – The three cases described above can be applied inductively to the leafs of the tree with $n = 3$, because a leaf can describe a composite component that is refined by applying the black box approach, as described later.

Interfacing to Other Models – Components are identified by a *globally unique identifier* and the already mentioned *component type*, both acting as interface to the existing model and tool landscape by applying the same naming conventions or even storing the same values as the corresponding model (*C4(✓)*). In CIM, for instance, every object is identified by an `InstanceID`, inherited from the global super class `ManagedElement.InstanceID`. Using this value for component identification facilitates information exchange and keeping the model in sync with alterations in the considered HPDC infrastructure, as components can easily be related to corresponding objects, e.g., in an existing CIM based management tool. The very generic identifier definition provides an interface to all models that use String based identifiers, e.g., the URI formatted *Document Object Identifier* (DOI) applying *Simple Network Management Model* (SNMP) [27]. The component type value range is extracted from the corresponding model and applied to describe components, e.g. CIM would set the possible value range to the class inheritance hierarchy, like $ManagedElement > ... > NetworkAdapter$ or $ManagedElement > ... > DiskDrive$.

Black Box Approach – A composite component, whose internal matters are hidden in the model, is called a *black box*, which can be refined whenever it is necessary to know these internal details, resulting in a *composite component* described by a sub tree. This is depicted in Fig. 1: the *Compute Island* consists of four *Compute Nodes*, of whom the outer left one was refined, resulting in a sub tree describing a composite component. The black box approach facilitates multiple levels of granularity simultaneously, because refinement can be done punctually, while keeping other components on an abstract level (*C2(✓)*).

Assigning types to HPDC infrastructure components facilitates an automatic propagation of a black box refinement, because the refining sub tree can replace all other black boxes having the adequate component type. Algorithm 1 describes this automatic propagation in pseudo code: In a preparation step, the refinement is stored in $refinedBlackBox$ for further usage (line 1). Afterwards, it is iterated over all other black boxes having the same component type as the refined black box (line 2). The located black boxes ($currentBB$) are replaced by the refinement in two consecutive steps: **First**, the black box refinement $refinedBlackBox$ is copied and placed into the tree by reconnecting the edges that were previously connected to the replaced black box (line 3). **Second**, the black box identifier value of all vertices in the refining tree is set to facilitate recognition of initially modeled black boxes (line 4).

Combining the black box approach and the component types facilitates very fast model development (*C5(✓)*), because an HPDC infrastructure model can be created very quickly on an abstract level and afterwards punctually refined. This advantage is further strengthened by the usually high amount of components having the same type, especially in HPC infrastructures and clusters.

Model Part 2 – Property and Attribute Description

To each component, a specific *property vector* is assigned, consisting of an arbitrary amount of key-value pairs: the key is the property identifier, e.g., *throughput*, and the value is the respective value accordant to the property's dial.

Attribute description builds upon two pillars: *calculation rule definition per component* and *guidance by calculation result types*. Dependent on the objective, the component type, and the varying quality, uncertainty, and incompleteness of data, each calculation rule has its specific (dis)advantages, e.g., a Weibull distribution is often advocated for various resource availability data [28] or there are reliable vendor specifications to describe CPU reliability, but only incomplete empirical observations for the local memory. Our model allows a particular attribute calculation rule for (potentially) every component to cover this situation. As shown in [1], attribute calculations end up in a *binary*, *fraction*, and *probability* result type. Calculation of *composite attribute values* is heavily eased by *focusing on calculation result types*, because only the same result types can be aggregated.

Model Part 3 – Analysis Execution

For each analysis run, a different property vector can be assigned to the components, which facilitates *what-if* analysis and is required to solve the introduced optimization problem (cf. Section 1).

Algorithm 1: Automatic propagation of black box refinement

```
1 refinedBlackBox = describes refinement of blackbox
2 while currentBB = getUnrefinedBlackboxOfType(blackbox.type) do
3 |   placeRefinementIntoTree();
4 |   setBlackboxIdentifierTo(currentBB.identifier);
5 end
```

4.2 Formalization

The UML class diagram depicted in Fig. 2 formalizes the outlined model concepts and acts as meta model to describe arbitrary settings for PMEA. Formalization follows the above applied SoC implementation, illustrated by the numbered rectangles in Fig. 2. This SoC is of special importance to property and attribute representation, because both are separated into *concept* and *value* representation, as subsequently described in *Meta Model Part 2* and *Meta Model Part 3*, respectively.

Meta Model Part 1 – HPDC Infrastructure Components The class `Component` and the recursive data structure *composite pattern* represent atomic and composite components as well as the tree structure. The composite pattern is used not only because of its suitability for modeling tree structures, but also because 1) it intends to treat leafs and components similarly, which is mandatory to fulfill *C1.1*, and 2) it facilitates reusability, especially of attribute calculation rules (*C3*). Due to object oriented modeling, the `Component` class can be extended, e.g., to store individual fields like operating system, if no existing external model is attached. The `Identifier` class defines the component identifier syntax and takes care for a globally unique identification. Together with the taxonomy describing `Type` class, it serves as interface to external models and tools. Since a description in the meta model of the black box approach's main features, i.e. the applied perspective and the algorithm, is not useful, there is only the field `Component.blackboxIdentifier` to store the former black box identifier (cf. Algorithm 1). The component's system workload caused by application execution is stored in the field `Component.currentWorkload`.

Meta Model Part 2 – Property and Attribute Representation Static property and attribute concepts are described by the classes `Property` and `Attribute`, respectively. Both classes contain a globally unique `identifier`, the respective `dial`, and a `description` field to explain the concepts. For instance, an object describing the property *throughput* in *bit/s* could have the following (simplified) class field values: {`identifier=throughputBitPerSecond, dial=bit/s, description=...`}. Property objects are used for attribute calculation and for mapping scientific applications onto HPDC infrastructures.

Meta Model Part 3 – Analysis Execution The static attribute and property concept definitions are assigned to components with different (dynamic) values per analysis run. The value is stored in the `PropertyValue` class for properties and calculated by the `AttributeValue` class for attributes, respectively. Both classes contain an `analysisRunId` to explicitly identify the analysis run the value is defined or calculated for. Due to its extend, attribute value calculation is further explained: the three possible result types (cf. Section 4.1) are represented by interfaces, which certainly can be further sub classed, e.g., sub classing the `Probability` interface to distinguish Weibull and Pareto probability distribution. These interfaces are implemented by `AttributeValue` objects, each calculating a particular attribute value for a component. For instance, calculating the probability of finding a component in the up state at time t (*point availability* $A(t)$) [29] for a certain component, is implemented as follows: the calculating `AttributeValue` object extends the `Attribute` object `Reliability` to inherit the "semantics", and it implements the interface `Probability`, because a probability is calculated. The concrete calculation rule is implemented in the method `AttributeValue.getValue()`, which handles different parameter sets per calculation rule by a key-value map parameter. The same approach is applied

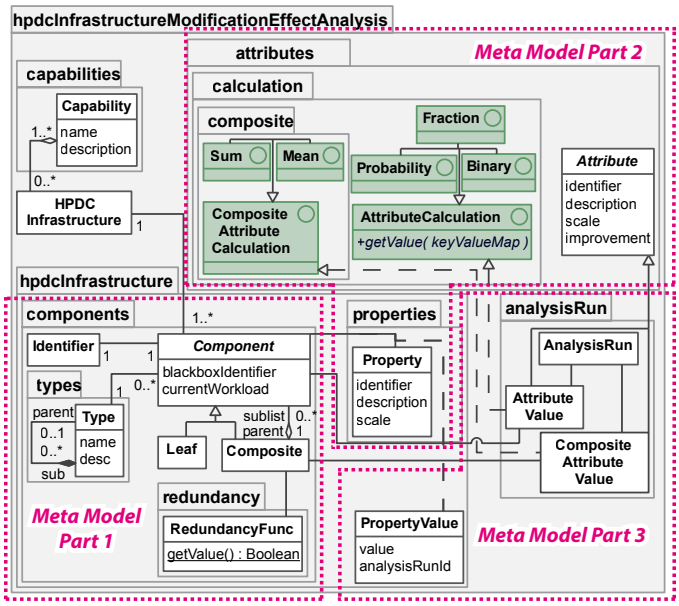


Fig. 2: UML class diagram formalizing the fundamental concepts of our meta model

for calculating composite attribute values: the calculating class `CompositeAttributeValue` extends the accordant concept `Attribute` class and implements one of the `CompositeAttributeCalculation` sub classing interfaces.

5. Related Work

Investigation of related work is split up into two groups, one group containing tools as possible candidates for PMEA, and one group containing models to serve as possible extension base for our meta model. Benchmarks are not considered at all, because they require a physical system by design, which is not possible (see Section 1).

At a first glance, the **first group** contains a bulk of sufficient tools, since there are a lot of mature approaches to plan and analyze modifications. Some examples are *Dramsim2* [30] describing cycle accurate DDR2/3 memory systems, or [5] analyzing the impact of sub-optimal checkpointing on application efficiency, or the *Structural Simulation Toolkit* (SST) [12], combining "a collection of hardware component models including processors, memory, and networks". Even if these tools are very advanced in analyzing the modifications in a certain domain or area, they do not fulfill *C1* partly or completely: either they do not cover all HPDC infrastructure components (*C1.1*(×)), like Dramsim2, focusing on specific memory systems, or they focus only on (a few) attributes while omitting other attributes and effect cascading (*C1.6*(×)), like [5], focusing only on efficiency, or they do not cover workload sufficiently (*C1.3*(×)), like SST, not providing a dedicated scientific application interface. Obviously, there are lots of other not named tools, but to the best of our knowledge, they do not fulfill *C1* either, as they all try to copy the physical real world as accurate as possible by modeling the properties underlying hardware and software configuration in detail (cf. Section 2), while omitting attributes and effects and their interplay partly or completely. Since this group does not provide the mandatory valid model, further evaluation of criteria *C2*–*C5* is omitted.

Extending an existing model would benefit from third party experience and ease integration in and interaction with existing (management) tools. Hence, the **second group** contains possible extension base candidates. There are manifold and widespread used (meta) models and languages to describe a network, computer system or a distributed system, e.g., the SNMP, the *Grid Laboratory Uniform Environment* (GLUE) [31], or the *Common Information Model* (CIM) [32], just to mention some of the most prominent ones. All these models have a specific focus, but no model focuses on PMEA (*C1*(×)). Since a potential model should only serve as an extension base, this non-fulfillment is not that sever as for the above described tools. Nevertheless, extending one of these models in a way that the resulting model would completely fulfill *C1* would strongly contradict the required simplicity (*C5*) due to the overhead. For instance, CIM contains lots of hardware related details, like `MediaAccessDevice.MaxBlockSize`. Additionally, most of the models do not strictly implement a SoC, since classes mix up hardware details and properties, e.g., the CIM class `NetworkAdapter` describes `PermanentAddress` (hardware detail) and `MaxSpeed` (property). In summary, the benefits of extending an existing model would be outweighed by the overhead.

6. Conclusion and Future Work

The paper emphasizes the necessity of a *predictive modification effect analysis* (PMEA), especially during modification planning, to investigate, whether the (negative) side effects will outweigh the (positive) intended effects, both induced by the considered modification. The paper presents an elaborated meta model to describe HPDC infrastructures for PMEA, based on a derived criteria set. In future work, runtime performance of the overall model will be evaluated and improved, a graphical editor based upon the presented meta model will be shown, and the currently developed scientific application model will be combined with the presented model to calculate system workloads.

References

[1] C. Straube and D. Kranzlmüller, "Model-Driven Resilience Assessment of Modifications to HPC Infrastructures," in *Proceedings of the 6th Workshop on Resiliency in High Performance Computing (Resilience) in Clusters, Clouds, and Grids in Conjunction with Euro-Par 2013*, 2013.

[2] J. Elliott, K. Kharbas, D. Fiala, F. Mueller, K. Ferreira, and C. Engelmann, "Combining Partial Redundancy and Checkpointing for HPC," in *Proceedings of the 32th International IEEE Conference on Distributed Computing Systems (ICDCS)*, 2012, pp. 615–626.

[3] I. Eusgeld, B. Fechner, F. Salfner, M. Walter, P. Limbourg, and L. Zhang, "Hardware Reliability," in *Dependability Metrics*, R. Reussner, F. C. Freiling, and I. Eusgeld, Eds. Springer, 2008, vol. 4909, pp. 59–103.

[4] C. Straube and D. Kranzlmüller, "An IT-Infrastructure Capability Model," in *Proceedings of the 10th ACM Conference on Computing Frontiers (CF'13)*. ACM, May 2013.

[5] W. Jones, J. Daly, and N. DeBardeleben, "Impact of Sub-Optimal Checkpoint Intervals on Application Efficiency in Computational Clusters," in *Proceedings of the 19th International ACM Symposium on High Performance Distributed Computing (HPDC'10)*. ACM, 2010, pp. 276–279.

[6] R. Murphy, "On the Effects of Memory Latency and Bandwidth on Supercomputer Application Performance," in *Proceedings of the 10th International IEEE Symposium on Workload Characterization (IISWC'07)*, 2007, pp. 35–43.

[7] D. Chen, N. Eisley, P. Heidelberger, S. Kumar, A. Mamidala, F. Petrini, R. Senger, Y. Sugawara, R. Walkup, B. Steinmacher-Burow, A. Choudhury, Y. Sabharwal, S. Singhal, and J. Parker, "Looking Under the Hood of the IBM Blue Gene/Q Network," in *Proceedings of the International ACM/IEEE Conference on High Performance Computing, Networking, Storage and Analysis (SC'12)*, 2012, pp. 1–12.

[8] M. Al-Mashari and M. Zairi, "Creating a Fit Between BPR and IT Infrastructure: A Proposed Framework for Effective Implementation," *Journal of Flexible Manufacturing Systems*, vol. 12, no. 4, pp. 253–274, 2000.

[9] A. Sabetta and H. Koziolek, "Measuring Performance Metrics: Techniques and Tools," in *Dependability Metrics*, R. Reussner, F. C. Freiling, and I. Eusgeld, Eds. Springer, 2008, vol. 4909, pp. 226–232.

[10] S. Chen, Y. Liu, I. Gorton, and A. Liu, "Performance Prediction of Component-Based Applications," *Journal of Systems and Software*, vol. 74, no. 1, pp. 35–43, 2005.

[11] K. L. Calvert, M. B. Doar, and E. W. Zegura, "Modeling Internet Topology," *IEEE Communications Magazine*, vol. 35, no. 6, pp. 160–163, 1997.

[12] A. F. Rodrigues, K. S. Hemmert, B. W. Barrett, C. Kersey, R. Oldfield, M. Weston, R. Risen, J. Cook, P. Rosenfeld, E. CooperBalls, and B. Jacob, "The Structural Simulation Toolkit," *ACM SIGMETRICS Performance Evaluation Review - Special Issue on the 1st International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computing Systems (PMBS'10)*, vol. 38, no. 4, pp. 37–42, 2011.

[13] B. Ludäscher, M. Weske, T. McPhillips, and S. Bowers, "Scientific Workflows: Business as Usual?" in *Proceedings of the Business Process Management*. Springer, 2009, vol. 5701, pp. 31–47.

[14] C. Catlett, W. Allcock, P. Andrews, R. Aydt, R. Bair, N. Balac, B. Banister, T. Barker, M. Bartelt, P. Beckman, *et al.*, "Teragrid: Analysis of Organization, System Architecture, and Middleware Enabling New Types of Applications," IOS Press, Tech. Rep., 2008.

[15] (2013) Partnership for Advanced Computing in Europe (PRACE). [Online]. Available: www.prace-project.eu

[16] L. Meyer, S. Rössle, P. Bisch, and M. Mattoso, "Parallelism in Bioinformatics Workflows," in *Proceedings of the High Performance Computing for Computational Science (VECPAR'04)*, J. Palma, V. Hernández, and M. Daydé, Eds. Springer, 2005, vol. 3402, pp. 583–597.

[17] W. T. Kramer and C. Ryan, "Performance Variability of Highly Parallel Architectures," in *Proceedings of the Computational Science (ICCS'03)*. Springer, 2003, vol. 2659.

[18] S. Robinson, *Simulation: The Practice of Model Development and Use*. John Wiley & Sons, Ltd., 2004.

[19] D. Molka, D. Hackenberg, T. Minartz, R. Schöne, and W. Nagel, "Flexible Workload Generation for HPC Cluster Efficiency Benchmarking," *Computer Science - Research and Development*, vol. 27, no. 4, pp. 235–243, 2012.

[20] C. P. Pfleeger and S. L. Pfleeger, *Security in Computing*. Prentice Hall, 2007, vol. 4.

[21] D. Hackenberg, R. Schöne, D. Molka, M. Müller, and A. Knüpfer, "Quantifying Power Consumption Variations of HPC Systems Using SPEC MPI Benchmarks," *Computer Science - Research and Development*, vol. 25, no. 3, pp. 155–163, 2010.

[22] H. Koziolek, "Introduction to Performance Metrics," *Dependability Metrics*, vol. 4909, November 2008.

[23] T. Talbot and H. Davis, "Verizon NEBS TM Compliance: Energy Efficiency Requirements for Telecommunications Equipment," Verizon, Tech. Rep. VZ.TPR.9205, October 2011.

[24] G. E. Moore, "Cramming More Components onto Integrated Circuits," *Electronics*, vol. 38, no. 8, April 1965.

[25] L. Chwif, M. Barretto, and R. Paul, "On Simulation Model Complexity," in *Proceedings of the IEEE Winter Simulation Conference*, 2000, pp. 449–455.

[26] O. Khalili, J. He, C. Olschanowsky, A. Snavely, and H. Casanova, "Measuring the Performance and Reliability of Production Computational Grids," in *Proceedings of the 7th IEEE/ACM International Conference on Grid Computing*, 2006, pp. 293–300.

[27] A. Pilz and J. Swoboda, "Network Management Information Models," *AEU - International Journal of Electronics and Communications*, vol. 58, no. 3, pp. 165–171, 2004.

[28] D. Nurmi, J. Brevik, and R. Wolski, "Modeling Machine Availability in Enterprise and Wide-Area Distributed Computing Environments," in *Proceedings of the Euro-Par 2005 Parallel Processing*, P. Medeiros and J. Cunha, Eds. Springer, 2005, vol. 3648, pp. 432–441.

[29] K. Trivedi, *Probability and Statistics with Reliability Queuing and Computer Science Applications*. John Wiley & Sons, Inc., 2001.

[30] P. Rosenfeld, E. Cooper-Balis, and B. Jacob, "DRAMSim2: A Cycle Accurate Memory System Simulator," *Computer Architecture Letters*, vol. 10, no. 1, pp. 16–19, 2011.

[31] S. Andreozzi, S. Burke, F. Ehm, L. Field, G. Galang, B. Konya, M. Litmaath, P. Millar, and J. Navarro, "GLUE Specification v. 2.0," Open Grid Forum, Tech. Rep. GFD-R-P.147, 2009.

[32] DMTF, "Common Information Model (CIM) Metamodel (Specification)," Distributed Management Task Force (DMTF), Tech. Rep. DSP0004, 2012.

Comparing Collective Behaviour of Sociophysical Models

C. Butt[1], **D. P. Playne**[1] **and K. A. Hawick**[2]
[1]Computer Science, Massey University, Auckland, New Zealand
[2]Computer Science, University of Hull, Hull, United Kingdom
Email: d.p.playne@massey.ac.nz, k.a.hawick@hull.ac.uk
Tel: +64 9 414 0800 Fax: +64 9 441 8181

Abstract— *Many complex systems such as crowds, financial and economic markets and other sociological systems can be modelled as a collection of microscopically simple individual agents. These models have properties in common with some physical systems in that they exhibit critical behaviour or phase transitions where a small change in a controlling parameter can give rise to a sudden and dramatic change. Such systems are difficult to study using conventional analytic mathematical techniques and generally require numerical simulation to allow study of the complex emergent effects that occur. Three sociophysical models based on opinion formation and cultural dissemination are presented and discussed. The behaviour of these three models are analysed and compared to show the effect of the social forces present in each of the models.*

Keywords: Voter, Sznajd, Axelrod, Sociophysical.

1. Introduction

Sociophysical models can be used to describe the interactions of many complex systems such as crowds, opinion formation, cultural dissemination and collective decision making [1], [2]. The microscopic details of such models can often be described by simple rules which can result in complex macroscopic behaviour. The study of such emergent behaviour can provide useful insights into the nature of such systems and have applications in Sociology, Economics and Politics.

These models have similarities to physical systems that exhibit critical behaviour or phase transitions where a seemingly minor adjustment or change in parameter and lead to a significant effect in the large-scale macroscopic behaviour of the entire system [1]. Many of the simulation and analysis techniques developed for physical simulations can be applied to this type of sociophysical model.

The Voter [3], [4], Sznajd [5] and Axelrod [2] models can all be used to approximate the dynamics of opinions within a population. Each of these models use different microscopic rules to govern the interactions between agents. The rules of these models are based on taking different social forces and aspects into consideration. The three main social forces approximated by the models investigated are - Social Influence, Social Validation and Homophily [6], [7].

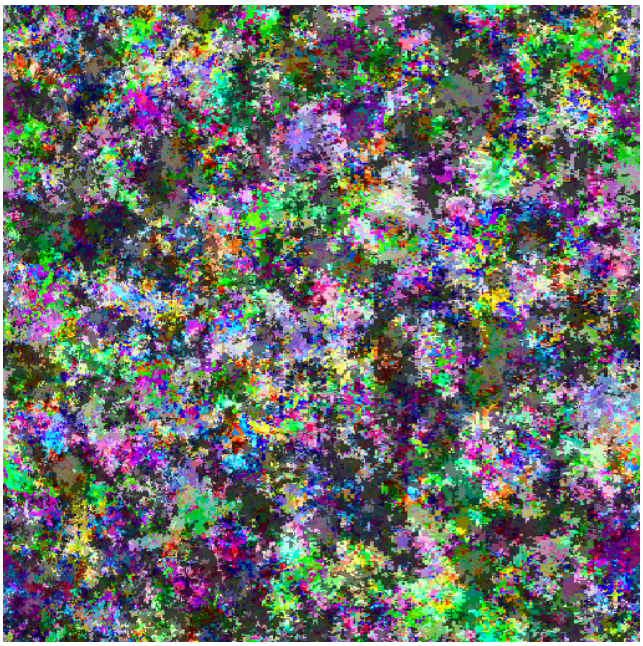


Fig. 1: A 512x512 Axelrod system with $F = 10$ Features and $q = 2$ opinions. Details on the visualisation method are given in Section 4.

Social Influence is the effect that individuals are likely to be influenced by the others that they interact with and are more likely to agree on the same opinion. This effect is present in all of the three models and is the only social force present in the Voter Model where individuals adopt the opinions of one of their neighbours.

Social Validation considers the influence that multiple agreeing individuals can have on their surrounding neighbours [6]. This social effect is only represented in the Sznajd model which requires two neighbouring individuals to agree on an opinion for their neighbours to be influenced.

Homophily allows the opinions two individuals hold to affect the strength of their social interaction. Two individuals that have a lot in common are more likely to interact and

exchange opinions on topics they disagree on. The Axelrod Model allows individuals to hold opinions on a range of cultural features and includes the effects of Homophily.

These three models are described in detail in Sections 2, 3 and 4. The analysis and results of these three models are presented and compared in Section 5. Finally a discussion is offered and some conclusions are drawn in Section 6.

2. Voter Model

The Voter Model is a simple stochastic process that can be used to simulate the dynamics of opinions in a population. It was independently developed by Clifford and Sudbury to model two species competing for territory [3] and Holley and Liggett [4]. The traditional model considers a population of individuals which may have one of two possible opinions. The individuals in the population are updated at random and each individuals always adopts the opinion of one of its neighbours. A number of variations of this model have been investigated and analysed, including - Heterogenous Networks [8], Biased Voters [9], Zealous Voters [10], Noisy Voter Model [11].

Fig. 2: In the Voter model, a randomly chosen individual adopts the opinion of one of its four neighbours chosen at random.

The Voter Model investigated in this work allows each individual to adopt any one of q-opinions rather than just two. These opinions are considered to be completely independent and do not represent varying opinions along a scale such as left-wing to right-wing. This also means there is no concept of similarity between two individuals and two neighbours are equally likely to interact regardless of what opinions they hold. By allowing individuals to have one of q-opinions, the behaviour of the Voter model can be compared to the Sznajd and Axelrod models across different values of q.

The update rules follow the normal Voter Model pattern where each individual is updated by randomly selecting a neighbour to interact with and adopting its opinion. This is how the Voter model implements the effects of social influence. In the simulations in this work, the individuals are arranged on a two-dimensional rectilinear lattice and are connected to their four nearest neighbours. The process of updating an individual is shown in Figure 2 and the algorithm for simulating a q-opinion Voter model system is summarised below:

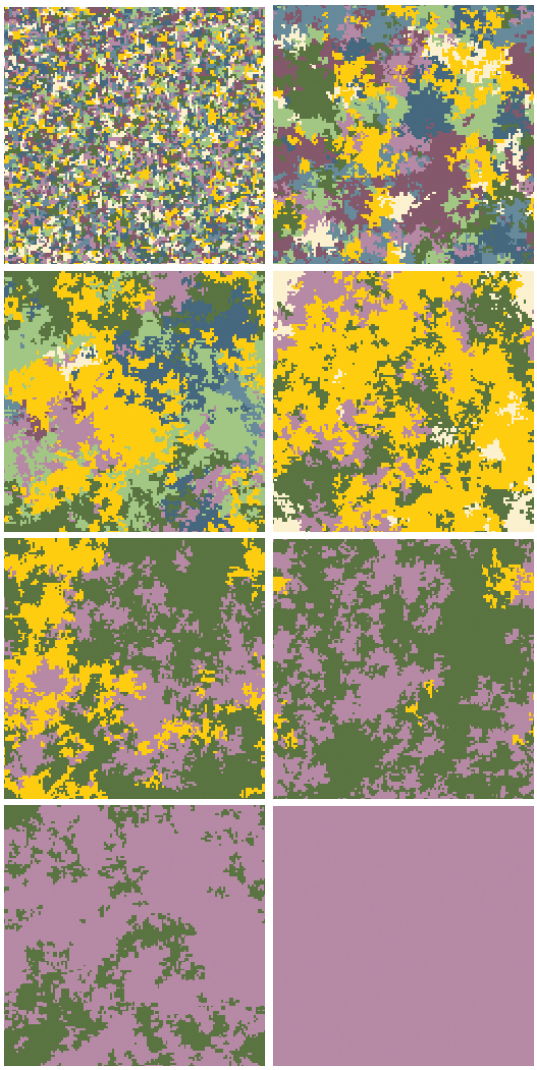


Fig. 3: Eight snapshots of a 128x128 Voter model simulation with $q = 8$ opinions from initialisation to consensus. Opinion clusters slowly grow over time and tend to form very 'jagged' structures.

1) Initialise each voter with a random opinion [0..q)
2) Select an individual at random
3) Select a single neighbour at random
4) The individual adopts the opinion of the neighbour
5) Return to 2. until system reaches consensus

The Voter Model describes the dynamics of opinions within a population through a different microscopic process to the Sznajd and Axelrod (See following sections). It considers each individual to be easily swayed by any neighbour. It does not require two neighbouring voters to agree like the Sznajd or consider the possibility that voters may hold opinions on a number of features like the Axelrod model. These three models are compared in Section 5.

The time evolution of a 128x128 Voter model simulation with $q = 8$ opinions is shown in Figure 3. Each individual in the simulation is represented as a coloured square. As an individual may hold only a single opinion at a time, a look-up table was used to determine which colour should be rendered corresponding to each individual's opinion.

3. Sznajd Model

The "United we Stand, Divided we Fall" or Sznajd Model was developed to model the evolution of opinions in a closed community [5], [6], [12]. The original Sznajd Model considers individual that have the option to hold one of two opinions. Unlike the Voter Model where each voter adopts the opinion of a neighbouring voter, the Sznajd model allows two individuals to influence their neighbours. The model requires two individuals to agree on an opinion to be able to convince their neighbours. This means that social validation is required for neighbouring individuals to be affected by social influence [6].

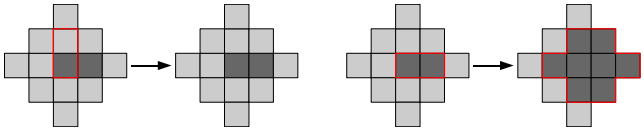


Fig. 4: The Sznajd looks at one of its four closest neighbours. If they share an opinion, that opinion spreads to their six closest neighbours.

The individuals in the Sznajd model simulations are allowed to hold one of q opinions which allows the evolution of opinions to be compared across range of values for q. Like the Voter model, the individuals are arranged in a two-dimensional, rectilinear lattice and are allowed to interact with each of their four nearest-neighbours.

When an individual is updated, a neighbour will be chosen at random to interact with. If the two voters agree (have the same opinion) then they convince all of their surrounding neighbours. In these two-dimensional simulations, when the two voters do not agree then they have no effect on their neighbours [13], [14]. This update is shown in Figure 4.

This does mean that there are two possible stable states of the Sznajd Model. The first is when the system reaches consensus and all voters agree on the same opinion. This is the only stable state for the voter model. The second is the scenario that every single voter in the population disagrees with each and every neighbour. This scenario is unlikely for low values of q but becomes possible for small systems with large q values. The process of initialising and simulating the Sznajd model is:

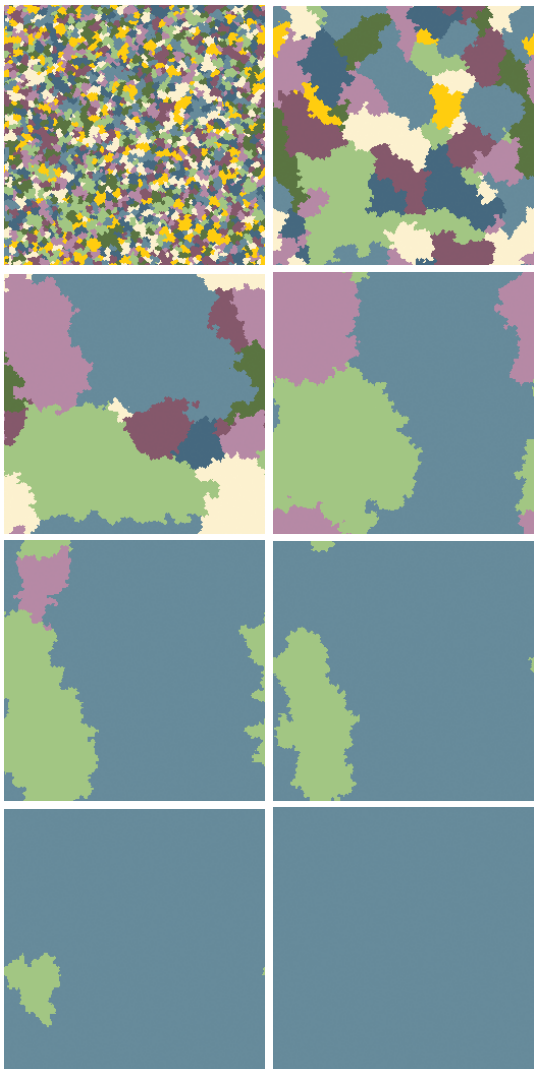


Fig. 5: System snapshots of the evolution of Sznajd opinion clusters over time. The simulation is of size 128x128 and is initialised with 8 opinions. Opinion clusters form quickly after initialisation and grow into large clusters with more clearly defined edges (compared to the Voter Model).

1) Initialise each individual with a random opinion
2) Select an individual at random
3) Select a single neighbour at random
4) If the voters agree, set all neighbouring voters to that opinion
5) Repeat until system reaches a stable state

The Sznajd model describes the behaviour of voters changing their opinion under the influence of peer pressure from their surrounding neighbours. This opinion 'pushing' where information flows outwards contrasts with the Voter model where individuals adopt a neighbouring opinion. This process can present some significant challenges when it comes

to computing the simulation in parallel [15].

The evolution from initialisation to consensus of a $q = 8$ Sznajd Model simulation of size 128x128 is shown in Figure 5. The simulation is visualised using the same look-up table as the Voter model. The main visual difference seen in the Sznajd model as compared to the Voter model is the distinct large clusters of single colours that emerge quickly. The 'jagged' structures present in the Voter Model simply cannot survive in the Sznajd model that requires social validation for an opinion to propagate.

4. Axelrod Model

The Axelrod model is a sociophysical model that allows individuals to hold opinions on different cultural features which are related and influence each other [2]. These cultural features could represent real-world features such as Social Policy, Economics, Music etc. The Axelrod model takes into account the social factor known as homophily where individuals who hold similar opinions are more likely to interact and adopt each other's opinions on other cultural features [7], [16]. These different cultural features is a major point of difference between the Axelrod model and the Voter and Sznajd models. The Axelrod simulations investigated and compared in this work consider systems with $F = 10$ cultural features, each of which may have q opinions.

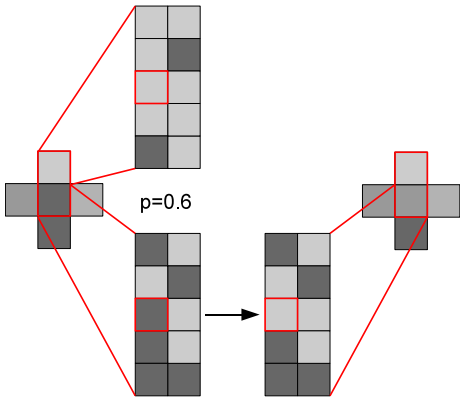


Fig. 6: In the Axelrod model, an individual adopts a single random opinion from a neighbour with a probability determined by how many features the two individuals currently share.

When an individual (A) in an Axelrod model is updated it will select a random neighbour (B) and interact with it based on the probability give in equation 1.

$$P^{AB} = \frac{1}{F}\sum_{i=1}^{F}\delta_{\sigma_i^A\sigma_i^B} \quad (1)$$

If an interaction occurs, the individual A will adopt the opinion of its neighbour B on a feature that they currently disagree on. This is significant point of difference compared to the Voter and Sznajd models where an individual and its neighbour will always interact. The interaction between two individuals is shown in Figure 6 and the process of simulating an Axelrod system is as follows:

1) For each individual, initialise every feature with an opinion $[0..q)$
2) Select an individual at random
3) Select a neighbour at random
4) Calculate P^{AB} and generate a random number to determine whether the two individuals interact
5) If an interaction occurs, adopt the neighbour's opinion on a feature the two disagree on
6) Repeat until a stable state is reached

Unlike the Voter and Sznajd models that drive the system towards a monocultural equilibrium, the Axelrod system can result in a disordered state where the population does not reach consensus and all individuals either agree completely with their neighbours or disagree completely and therefore do not interact [7], [17]. The Axelrod model is more likely to result in a disordered state as the number of possible opinions q increases. The two-dimensional Axelrod system with $F = 10$ features exhibits a transition from a ordered final state to a disordered final state when the number of opinions is in the range $q = 30 - 70$ [17]. The values of $q = \{2..8\}$ used in this work are well within the range of $q << 30$ where the two-dimensional Axelrod model with $F = 10$ features will reach a final state of consensus. Within this range the behaviour of the three models can be compared in the way they reach a final global consensus.

The Axelrod model is rendered procedurally, as opposed to the look-up table used by the Sznajd and Voter models. For the Axelrod model with $F = 10$ and $Q = 2$, each individual's colour is calculated in HSV colour space. The opinions for features 0 and 1 were used to calculate the Value, the opinions for features 2 and 3 used for the Saturation and finally the remaining opinions for features 4 to 9 were used calculate the hue.

This gives a total number of 2^{10} possible shades of colour to represent each individual for a system with $F = 10$ and $q = 2$. This visualisation method allows opinion clusters in different features to be visible at the same time. These clusters can be seen in Figure 7. As the model evolves, the system converges towards a single set of opinions until the system reaches consensus or the individuals are no longer able to interact.

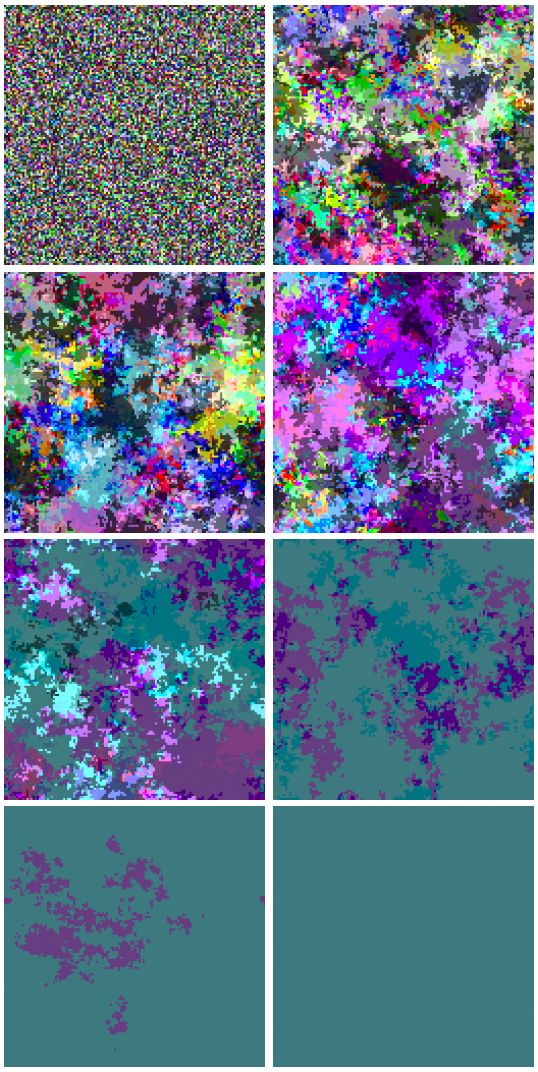


Fig. 7: A 128x128 Axelrod model simulation with $F = 10$ features and $q = 8$ opinions. As the system evolves in time opinion clusters form in each of the features and finally reach consensus with the entire population holding the same opinion on all feature.

5. Results

To track the evolution of the opinions through the population, the number of opinion clusters is measured. An opinion cluster is a collection of individuals that are all connected together and share the same opinion. For the Axelrod model, individuals are considered to be in the same cluster if they agree on every cultural feature. Figures 8, 9 and 10 show the evolution of the number of opinion clusters in the Voter, Sznajd and Axelrod models. The results are shown for three different sizes $L = \{16, 32, 64\}$ and with the number of opinions $q = \{2..8\}$.

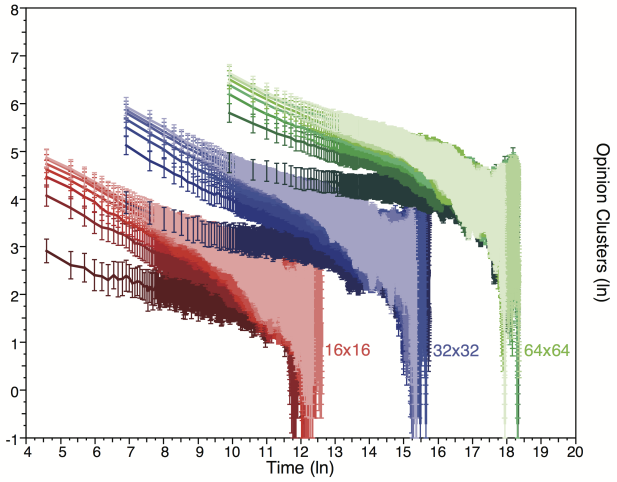


Fig. 8: The evolution of opinion clusters in the Voter Model. Results are shown for sizes $L = \{16, 32, 64\}$ and number of opinions $q = \{2..8\}$ in ln-ln scale. The colour of the lines is dependent on the number of opinions ranging from the darkest ($q = 2$) through to the lightest ($q = 8$).

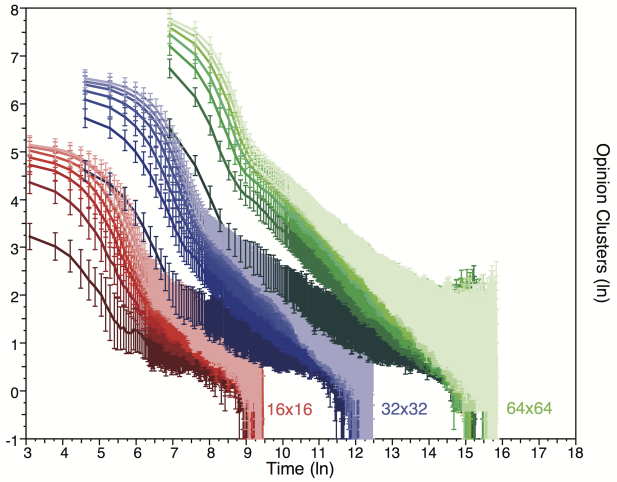


Fig. 9: The number of Sznajd model opinion clusters over time from initialisation to consensus shown for system sizes $L = \{16, 32, 64\}$. All results are shown in ln-ln scale for shaded from darkest-lightest for number of opinions $q = \{2..8\}$.

The evolution of the opinion clusters in the Voter Model shows similar behaviour across the different system sizes and number of opinions. This can be contrasted with the behaviour of the Sznajd model which shows a fast initial decrease and approaches global consensus faster. This fast initial decrease is due to the fact that two neighbours in agreement can quickly form a cluster by convincing all of

their collective neighbours. It also means that small clusters are more likely to die out because they require at least two neighbouring individuals for the cluster to survive - a voter model cluster can survive with one one.

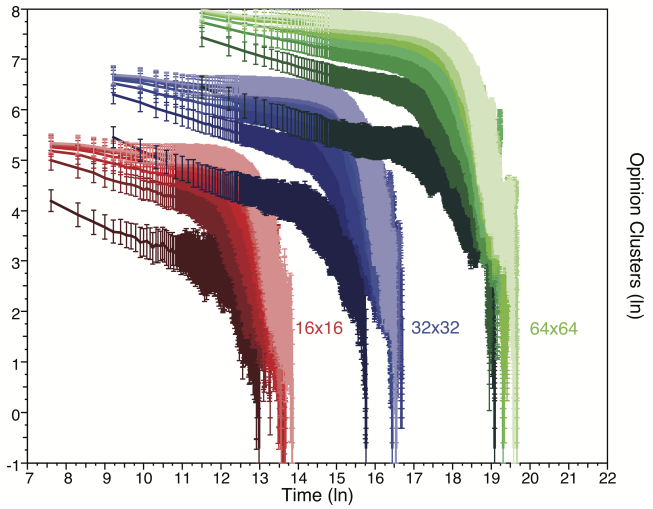


Fig. 10: The number of opinion clusters over time present in Axelrod simulations. All simulations have $F = 10$ features with $q = \{2..8\}$ opinions for each feature. Results are shown for systems of size $L = \{16, 32, 64\}$ on ln-ln scale and are shaded based on q. Note that two individuals are only considered to be part of the same opinion cluster if they hold the same opinion on all features.

The number of opinion clusters in the Axelrod model shows different behaviour again, the population takes a long time to form initial opinion clusters as individuals are only considered to be in the same cluster if they agree of every cultural feature. However, once these initial clusters have started to form and grow the system quickly approaches global consensus.

Figure 11 shows the average time for each model to reach consensus for system sizes $L = \{16, 32, 64\}$ and number of opinions $q = \{2..8\}$ (Note that the steps to consensus are shown in ln-scale). From this graph it can be seen that Sznajd model is the fastest to reach consensus and the Axelrod is the slowest. Each of the models shows similar scaling behaviour as the size of the system is increased.

Finally the evolution of the number of opinion clusters for the models are compared by scaling time by the average number of steps it takes each model to reach consensus. This allows the opinion evolution profiles of the three models to be seen on the same time scale and the patterns of behaviour of the models to be more easily compared. These results are shown in Figure 12 which is shown with and without error bars to allow the average to be clearly seen but also to indicate the variation in the results.

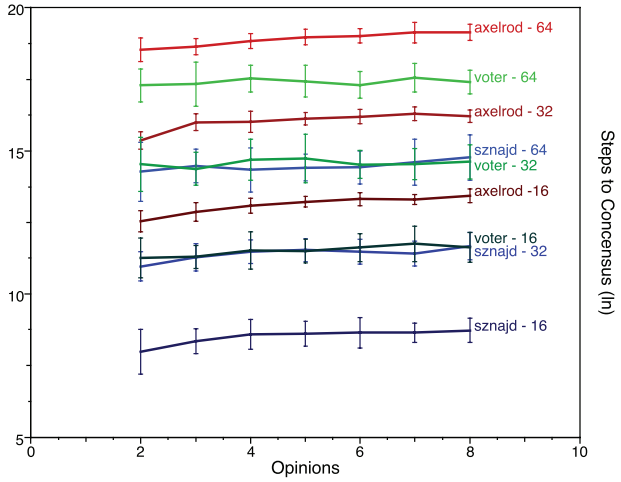


Fig. 11: A comparison of the average number of steps the Voter, Sznajd and Axelrod models take to reach consensus. The number of steps are shown in ln-scale for all system sizes and number of opinions.

6. Discussion and Conclusions

Three models of opinion propagation, the Voter, Sznajd and Axelrod models on a two-dimensional rectilinear domain have been discussed and compared. All show the effect of social influence driving the population towards a global consensus, however the average time (number of steps) required for this consensus to be reach differs greatly between models.

The effect of social validation is seen in the results of the Sznajd model which reaches consensus the fastest as opinions held only by a single individual are unable to survive. The effect of homophily in the Axelrod model is also clear as initial clusters take a long time to form.

The scaling of the average time to consensus across different domain sizes appears to be the same for each of the models suggesting that the scaling is dependent on the system size and not dependent on the particular model. By rescaling the results based on the average time to consensus, the formation of opinion clusters in the different models can be compared and contrasted.

The microscopic differences between the Voter, Sznajd and Axelrod models manifest as both significant structural differences in the clusters formed and the manner in which these clusters evolve as the system works towards consensus.

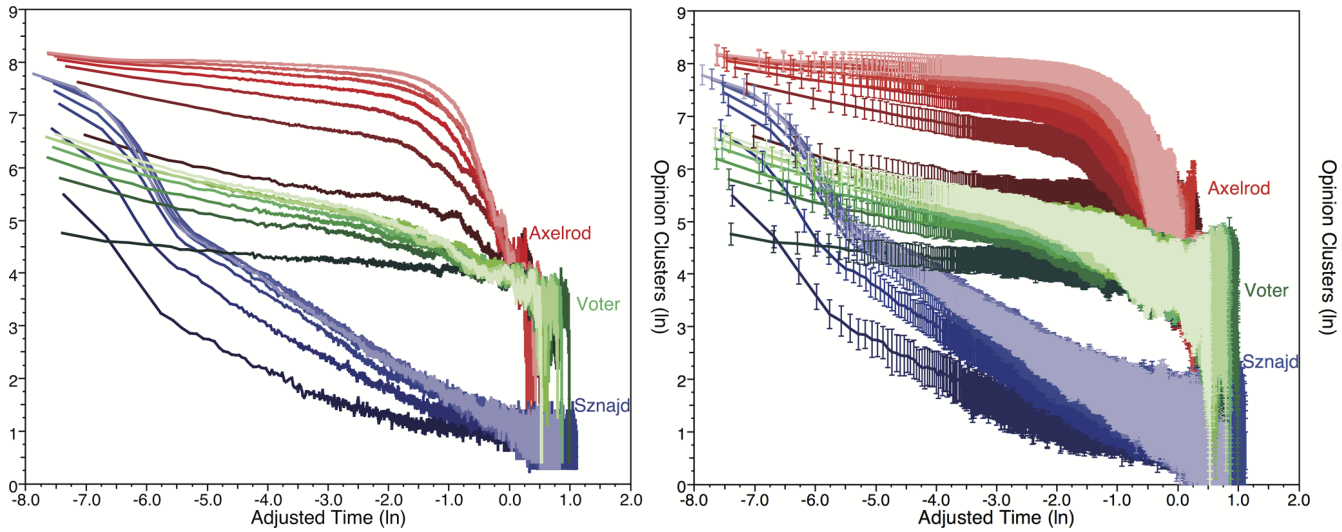


Fig. 12: Evolution of the opinion clusters of the Voter, Sznajd and Axelrod models scaled by the average time to reach system consensus for each model. These results are shown for a system size of $L = 64$ only.

References

[1] Sen, P., Chakrabarti, B.K.: Sociophysics - An Introduction. 1 edn. Oxford University Press (2014)

[2] Axelrod, R.: The dissemination of culture: a model with local convergence and global polarization. J. Conflict Resolution **41** (1997) 203–226

[3] Clifford, P., Sudbury, A.: A model for spatial conflict. Biometrika **60** (1973) 581–588

[4] Holley, R., Liggett, T.: Ergodic theorems for weakly interacting infinite systems and the voter model. The Annals of Probability **3** (1975) 643–663

[5] Sznajd-Weron, K., Sznajd-Weron, J.: Opinion evolution in closed community. Int. J. Modern Physics C **11** (2000) 1157–1165

[6] Sznajd-Weron, K.: Sznajd model and its applications. Acta Physica Polonica B **36** (2005) 2537–2547

[7] Lanchier, N., Schweinsberg, J.: Consensus in the two-state axelrod model. Stochastic Processes and their Applications **122** (2012) 3701–3717

[8] Sood, V., Antal, T., Redner, S.: Voter models on heterogeneous networks. Physical Review E **77** (2008) 041121

[9] Ferreira, I.: The probability of survival for the biased voter model in a random environment. Stochastic Processes and their Applications **34** (1990) 25–38

[10] Mobilia, M., Petersen, A., Redner, S.: On the role of zealotry in the voter model. Journal of Statistical Mechanics: Theory and Experiment (2007) P08029

[11] Granovsky, B., Madras, N.: The noisy voter model. Stochastic Processes and their Applications **55** (1995) 23–43

[12] Timpanaro, A.M., Prado, C.P.C.: Generalized sznajd model for opinion propagation. Phys. Rev. Lett. **80** (2009) 021119

[13] Staffer, D.: Monte carlo simulations of sznajd models. Journal of Artificial Societies and Social Simulation **5** (2001) 1–9

[14] Hawick, K.A.: Multi-party and spatial influence effects on opinion formation models. In: Proc. IASTED International Conference on Modelling and Simulation (MS 2010). Number 696-035, Banff, Canada, IASTED (2010) 235–240

[15] Hawick, K.A., Playne, D.P.: Halo Gathering Scalability for Large Scale Multi-dimensional Sznajd Opinion Models Using Data Parallelism with GPUs. In: Proc. Int. Conf. on Parallel and Distributed Processing Techniques and Applications (PDPTA'12), Las Vegas, USA, CSREA (2012) 95–101

[16] Lanchier, N.: The axelrod model for the dissemination of culture revisited. The Annals of Applied Probability **22** (2012) 860–880

[17] Dybiec, B., Mitarai, N., Sneppen, K.: Axelrod model: accepting or discussing. Eur. Phys. J. B **85** (2012) 357–1–5

An Improved Pseudo-parallel Genetic Algorithm for Process Mining

A. Hong Li[1], B. Gang Xue[2], C. Kunman Li[3], and D. Shaowen Yao[2]
[1]School of Information Science and Engineering, Yunnan University, Kunming, Yunnan, China
[2]National Pilot School of Software, Yunnan University, Kunming, Yunnan, China
[3]School of Computing and Engineering, University of Huddersfield, Queensgate, Huddersfield

Abstract *- Process mining is helpful for deploying new business processes as well as auditing, analyzing and improving the already enacted ones. An improved pseudo-parallel genetic algorithm is proposed with an asexual reproduction for avoiding crossover operators' breach to nice gene patterns. The initial population is produced by greedy algorithm in order to enhance convergence velocity. Information exchange between subgroups employs island model in pseudo- parallel genetic algorithm. These measures are of great significance on reducing complexities and enhancing convergence velocity, as well as increasing global searching ability of the algorithm.*

Keywords: process mining, genetic algorithm, parallel, log

1 Introduction

With the rapid development of information nowadays, workflow technology has been widely used. Among them, enterprise modeling is an important field in workflow technology application. Process mining is an important method of Enterprise modeling. It can be used for analyzing, auditing, optimizing the existing system, and help to improve the existing business processes. It can also help enterprises to realize the business process modeling and recycling then enhance the enterprise's competitive ability.

Research on process mining are mainly focused on mining the heuristic rules based on the binary sequence relationship which was recorded in the logs, and then extract the process model based on these rules. But the structures exist in actual workflow models have many complex structures which increase the difficulty of process mining, such as duplicate tasks, hidden tasks, implicit places, non-free-choices.

The scholars at home and abroad have done a lot of work in the process mining [1, 2]. So far, experts have put forward many meaningful and understanding languages to definite the model. For example, ADONIS model language, Petri nets model language, and the block structure model language. They all have their own characteristics. They have been able to finish the process modeling well basically. In the process mining, Agrawal firstly use the process mining algorithm applied to workflow management system. Initially, this algorithm requirements process does not exist circulation, the model dig out by the process illustrate by diagram which has direction but no circulation. But Agrawal disposed part of the circulation problem when he improved the algorithm. Herbst and others proposed another algorithm [7, 8] based on the ADONIS language. This algorithm deal well with the structure which containing the same activity name in the process, and find simple circulation structure. But to finding hide structure it haven't realized. Aalst and Weijter posed α algorithm [10, 11] that based on the extended Petri nets to find equivalent substitution with And-split、And-join structure. But it only uses for the relationship between each activity relationship. Thus, it cannot find the structures which need considering the global structure like NFC (no-free-choice) structure. And also it is helpless in two conditions that the same process contains the same activity name or complex circulation structures.

The purpose of the process mining from existing event log is to automatically find valuable information. The first dedicated process mining algorithms were proposed in [1][2]. The discovered information can be used to deploy new systems that support the execution of business processes or as an analytical tool that helps in auditing. At present many domestic and foreign scholars devoted to the research of process mining. A lot of people dig workflow model using WF nets, and based on this, puts forward α algorithm, β algorithm. The limitations [3] of α algorithm cannot mining the complex structure, such as loop, invisible task, non-free-choice constructs and so on. α+ algorithm, α++ algorithm extended some function based on it and solved part problem, but they did not consider the interval between the start events and end events. The β algorithm considers the interval, but cannot mine the looped and invisible tasks.

2 Related works

Since the mid-nineties several groups have been working ontechniques for process mining [4], [5], [10] discovering process models based on observed events. In [3] an overview is given of the early work in this domain. The idea to apply process mining in the context of workflow processes was introduced in [5]. In parallel data [12] looked at the discovery of business process models. Cook et al. investigated similar issues in the context of software engineering processes [10]. Herbs was one of the first to

tackle more complicated processes, which contain duplicate tasks.

Most of the classical approaches have problems in dealing with concurrency. The αalgorithm [4] is an example of a simple technique that takes concurrency as a starting point. However, this simple algorithm has problems in dealing with complicated routing constructs and noise (like most of the other approaches describedin literature). In [14] a more robust but less precise approach is presented. Heuristics or genetic algorithms [13] have been proposed to deal with issues such as noise.

Genetic algorithm (GA) is a kind of random search method reference the law of the biological evolution, its main algorithm steps are as Fig.1.

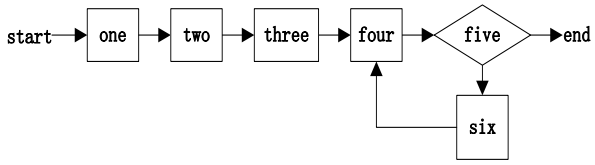


Fig.1. Genetic algorithm

Main steps of our genetic algorithm:

Step one: Read the event log.
Step two: Calculate the dependency relations among activities.
Step three: Build the initial population.
Step four: Calculate individuals' fitness.
Step five: Stop and return the fittest individuals
Step six: Create next population by using the genetic operators.

Genetic algorithm has been widely used in many engineering fields such as complex function optimization, structure design, system control, machine learning, and image processing and so on. It has global searching ability, high efficiency of group search strategy, irrelevance of the goal of gradient information and simple universal. But it has existent problem like the slow of efficiency and resource consumption. There put forward a Pseudo-Parallel Genetic Algorithm with asexual reproduction, it reduces the complexity of the algorithm, improves the algorithm convergence speed and improves the operation efficiency of algorithm.

3 Foundation technologies

3.1 Population initialization

First of all we should define the causal matrix.

Definition 1.A causal matrix is a tuple CM=(A,C,I,O), where: A is a finite set of activities, $C \subseteq A \times A$ is the causality relation, $I \in A \to p(p(A))$ is the input condition function, and $O \in A \to p(p(a))$ is the output condition function; such that: $C = \{(a1, a2) \in A \times A | a1 \in \cup\, I(a2)\}$, $C = \{(a1, a2) \in A \times A | a2 \in \cup\, O(a1)\}$, and $C \cup \{(ao, , ai) \in A \times A | ao = \emptyset \wedge ai = \emptyset\}$ is a strongly connected graph.

The initial population is randomly built by the genetic algorithm. When building the initial population, we roughly follow Definition 1. Given a log, all individuals in any population of the genetic algorithm have the same set of activities (or tasks) A. This set contains the tasks that appear in the log. However, the causality relation C and the condition functions I and O are randomly built for every individual in the population. As a result, the initial population can have any individual in the search space defined by a set of activities A. Note that the higher the amount of tasks that a log contains, the bigger this search space.

3.2 Fitness Calculation

If an individual in the genetic population correctly describes the registered behavior in the event log, the fitness of that individual will be high. In our approach the fitness is strongly related to the number of correctly parsed traces from the event log. Note that in case of noisy situation, we cannot aim at mining a process model that can correctly parse all traces, because the traces with noise cannot also be parsed by the desired model.

We use the naive semantics with silent transitions that only fire when needed and simply play the "token game". When the activity to be parsed is not enabled, the parsing process does not stop. The problem is registering and the parsing proceeds as if the activity was enabled (conceptually, this is equivalent to adding the necessary missing tokens in the Petri net to enable the activity and, then, firing it). We adopt this parsing semantics because it is more robust to noisy logs and it gives more information about the fitness of the complete process models. In a noise-free situation, the fitness of a model can be 1 (or 100%) (i.e. all traces can be parsed). In practical situations, the fitness value ranges from 0 to 1. The exact fitness of an individual to a given log is given by the formula:

$$\text{Fitness} = 0.40 \times \frac{\text{allParsedActivities}}{\text{numberOfActivitiesAtLog}} + 0.60 \times \frac{\text{allProperlyCompletedLogTraces}}{\text{numberOfTracesAtLog}} \quad (1)$$

numberOfActivitiesAtLog is the number of activities in the log. numberOfTracesAtLog is the number of log traces, allParsedActivities is the sum of parsed activities for all log traces. allProperlyCompletedLogTraces is the number of logtraces that were properly parsed

3.3 Stochastic tournament selection operator

To select a parent, the tournament selection algorithm randomly selects 5 individuals and returns the fittest individual among the fiveones.

3.4 Asexual reproduction

Here, the sexual reproduction crossover operator as a genetic operator may not meet certain constraints, so we must use "fix" algorithm or some special crossover operator such as partial match cross(PMC)、order crossover(OC) and circular cross (CC), But these crossover operators often have large damage to good genes mode. It can't be avoided to increase the complexity of the algorithm. So, in this paper use the asexual reproduction; It is said that to cancel the crossover operator, replaced by gene recombinant operator operation on individual. It simplifies the process of genetic operations, raises the calculation efficiency, and does not require the initial population diversity; there isn't a premature convergence problem.

3.5 Gene reconstruction

Adjusted gene symbol that Swap mutation、Insert mutation and Invert mutation the Relative position of chromosome be called gene recombinant operator. Individual fitness function calculation method can be further simplified because using asexual reproduction, that reorganization value is equal to the individual's fitness before the reorganization of fitness value plus the value of the fitness changed.

3.5.1 Swap mutation

Swap mutation is the exchange of encoded string of individual loci between two randomly selected gene values, resulting in a new tour route, for example:

$$T_X = (1,2\cdots,\mathrm{i}-1,\mathrm{i},\mathrm{i}+1,\cdots,\mathrm{j}-1,\mathrm{j},\mathrm{j}+1,\cdots,\mathrm{n}-1,\mathrm{n}) \rightarrow (1,2,\cdots,\mathrm{i}-1,\mathrm{j},\mathrm{i}+1,\cdots,\mathrm{j}-1,\mathrm{i},\mathrm{j}+1,\cdots,\mathrm{n}-1,\mathrm{n}) = T'_X$$

3.5.2 Insert mutation

Insert mutation is at first elected randomly two loci in the individual encoding string, then put a gene of loci after another geneloci. For example:

$$T_X = (1,2\cdots,\mathrm{i}-1,\mathrm{i},\mathrm{i}+1,\cdots,\mathrm{j}-1,\mathrm{j},\mathrm{j}+1,\cdots,\mathrm{n}-1,\mathrm{n}) \rightarrow (1,2,\cdots,\mathrm{i}-1,\mathrm{i},\mathrm{j},\mathrm{i}+1,\cdots,\mathrm{j}-1,\mathrm{j}+1,\cdots\mathrm{n}-1,\mathrm{n}) = T'_X$$

3.5.3 Invert mutation

Invert mutationis in reverse order the gene that selected randomly between the two loci of individual coding string, resulting in a new tour route, for example:

$$T_X = (1,2\cdots,\mathrm{i}-1,\mathrm{i},\mathrm{i}+1,\cdots,\mathrm{j}-1,\mathrm{j},\mathrm{j}+1,\cdots,\mathrm{n}-1,\mathrm{n}) \rightarrow (1,2,\cdots,\mathrm{i}-1,\mathrm{j},\mathrm{j}-1,\cdots,\mathrm{i}+1,\mathrm{i},\mathrm{j}+1,\cdots\mathrm{n}-1,\mathrm{n}) = T'_X$$

3.6 Island model

Pseudo-Parallel Genetic Algorithm is that run multi group parallel genetic algorithm on one processor. This article will initial population into three sub groups. Each child group independently evolves according to certain mode, at the right time, sub group can change some information, which can maintain the diversity of the population, and keep the different evolution direction, so as to suppressing precocious phenomena. Island model is also called coarse-grained model. Transfer strategy is the operator that changes information between sub groups. The migration can speed up in the propagation of the individual groups, improve convergence speed. This article uses the migration model as below in Fig.2.

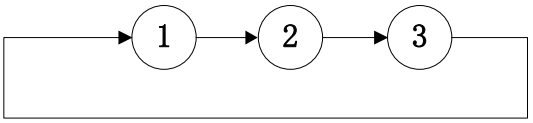


Fig.2. migration model

3.7 Stop criteria

The mining algorithm stops when (i) it finds an individual with fitness equals 1; or (ii) it computes n generations, where n is the maximum number of generation that is allowed; or (iii) the fittest individual has not changed for n/2 generations in a row.

4 Genetic Operations

We use elitism, crossover and mutation to build the individuals of the next generation. A percentage of the best individuals (the elite) are directly copied to the next population. The other individuals in the population are generated via crossoverand mutation. Two parents produce two offspring. To select parents, a tournament is played in which five individuals in the population are randomly drawnand the fittest one always wins. The crossover rate determines the probabilitythat two parents undergo crossover. Crossover is a genetic operator that aims atrecombining existing material in the current population. In our case, these materials are the set of current causality relations in the population. The crossover operation should allow the complete search of the space defined by the existing causality relation in a population. Given a set of causality relations, the search space contains all the individuals that can be created by any combination of a subsetof the causality relations in the population. Thus, our crossover operator allowsan individual to: lose tasks from the subsets in its I/O condition functions (butnot necessarily causality relations because a same task may be in more than onesubset of an I/O condition function), add tasks to the subsets in its I/O conditionfunctions (again, not necessarily causality relations), exchange causality relationswith other individuals, incorporate causality relations that are in the populationbut are not in the individual, lose causality relations, decrease the number of sub-sets in its I/O condition functions, and/or increase the number of subsets in its I/O condition functions. The crossover point of two parents is a randomly chosentask. Note that, after crossover, the number of causality relations for the wholepopulation remains constant, but how these relations appear in the offspring maybe different from the parents.

After the crossover, the mutation operator takes place. The mutation operator aims at inserting new material in the current population. In our case, this means that the mutation operator may change the existing causality relations of a population. Thus, our mutation operator performs one of the following actions to theI/O condition functions of a task in an individual: (i) randomly choose a subset and add a task (in A) to this subset, (ii) randomly choose a subset and remove from this subset, or (iii) randomly redistribute the elements in the subsetsof I/O into new subsets. Every task in an offspring may undergo mutation withthe probability determined by the mutation rate.

Improved pseudo parallel genetic algorithm process description as follows:

Step1: Initialize the genetic algebra counter: $t \leftarrow 0$

Step2: Create initial group P(t) based on the method of greed, perform preprocessing to P(t) and individual fitness calculation.

Step3: Copy P(t) into three son-group: P(t)= $P_1(t),P_2(t),P_3(t)$, perform three different genetic restructuring operator, swap mutation, insert mutation, and invert mutation respectively.

Step4: $P_i(t)(i = 1,2,3)$ evolution independent: $P'_i(t) \leftarrow \text{Selection}[P_1(t)](i = 1,2,3)$; do variation operation by gene recombinant operator$P''_i(t) \leftarrow Mutation[P'_1(t)](i = 1,2,3)$.

Step5: Group calculation individual fitness in different $P''_i(t)(i = 1,2,3)$.

Step6: Change the information between different $P''_i(t)(i = 1,2,3)$ by transfer strategy, and get the offsprings $P_i(t+1) \leftarrow \text{Exchange}[P''_1(t), P''_2(t), P''_3(t)]$.

Step7: Judge the conditions for the termination:

If do not meet the conditions for the termination then $t \leftarrow t+1$, go to Step4;

If meet the termination conditions, the optimal results, the output is over.

5 Experiment

As a first test of our Pseudo-Parallel Genetic Algorithm (PPGA), we applied it for noise-free event logs and checked if it could mine process models that contain all the behavior in these logs. In other words, the mined model should have the fitness F = 1. During the experiment, the genetic algorithm mined event logs from nets that contain 5, 7, 8, 12 and 22 tasks. These nets contain short loops, parallelism and/or non-free-choice constructions. Every event log has 1000 random executions of the nets. For each noise-free event-log, 10 runs of the genetic algorithm were executed. The populations had 500 individuals and were iterated for at most 100000 generations. The crossover rate was 1.0 and the mutation rate was 0.01. The elitism rate was 0.01. The initial population might contain duplicate individuals.

The generic mining algorithm presented in this paper is supported by a plugin in the ProM framework. Fig.3 shows screenshot of the plugin.

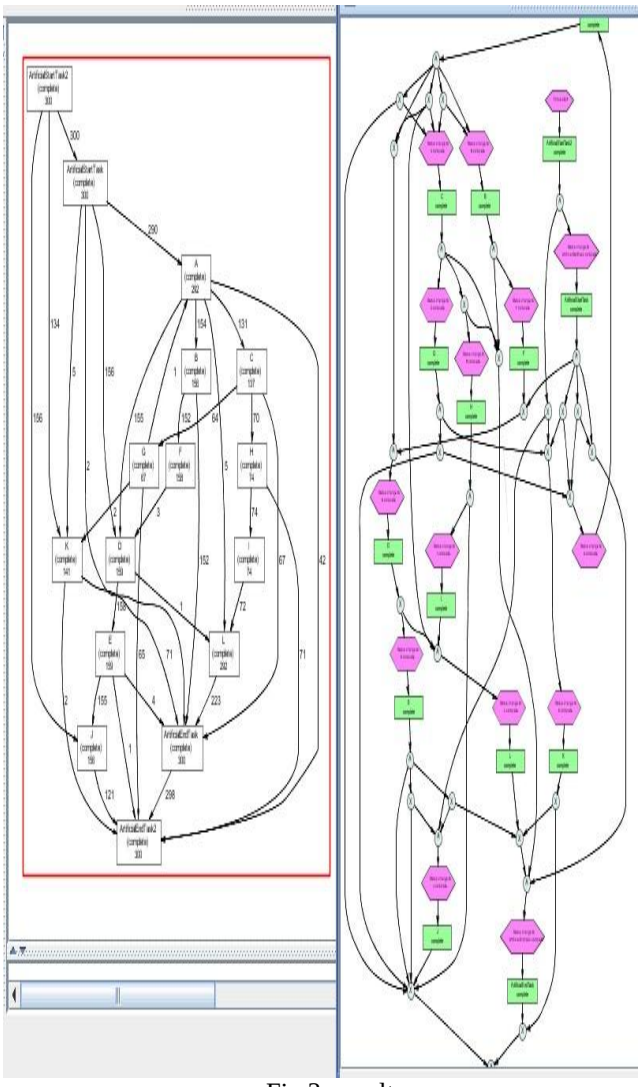


Fig.3. result

6 Conclusion

In this paper we presented our first experiences with a more global mining technique. The experiment shows that Pseudo-Parallel Genetic Algorithm can mining the process very well. These measures are of great significance on reducing complexities and enhancing convergence velocity, as well as increasing global searching ability of the algorithm.

7 ACKNOWLEGEMENT

This paper is supported by Laboratory of Network Intelligent Computing @ Yunnan University.

8 References

[1] J.E. Cook and A.L. Wolf. Automating process discovery through event-data anal-ysis. In Proceedings of the 17th international conference on Software engineering,pages 73{82. ACM, New York, NY, USA, 1995.

[2] J.E. Cook and A.L. Wolf. Discovering Models of Software Processes from Event-Based Data. ACM Transactions on Software Engineering and Methodology,7(3):215{249, 1998.

[3] R. Agrawal, D. Gunopulos, and F. Leymann. Mining Process Models from Workflow Logs. In Sixth International Conference on Extending Database Technology, pages 469–483, 1998.

[4] J.E. Cook and A.L. Wolf. Discovering Models of Software Processes from Event-Based Data. ACM Transactions on Software Engineering and Methodology, 7(3):215–249, 1998.

[5] J. Dehnert and W.M.P. van der Aalst. Bridging the Gap Between Business Models and Workflow Specifications. International Journal of Cooperative Information Systems, 13(3):289–332, 2004.

[6] K. van Hee, N. Sidorova, and M. Voorhoeve. Soundness and Separability of Workflow Nets in the Stepwise Refinement Approach. In W.M.P. van der Aalst and E. Best, editors, Application and Theory of Petri Nets 2003, volume 2679 of Lecture Notes in Computer Science, pages 335–354. Springer-Verlag, Berlin, 2003.

[7] Joachim H，DimitrisKaragiannis．Workflow mining with InWoLvE[J]、Computers in Industry,2004，53(4)：245—264．

[8] Guido Schimm．Mining exact models of concurrent workflows[J]．Computers in Industry，2004，53(2)：265—281

[9] R. Boumen, I.S.M. de Jong, J.W.H. Vermunt, J.M. van de Mortel-Fronczak, and J.E. Rooda. Test sequencing in complex manufacturing systems. Accepted for IEEE Transactions on Systems,Man,andCybernetics-Part A: Systems and Humans, 2006.

[10] J.E. Cook and A.L. Wolf. Discovering Models of Software Processes from Event-Based Data. ACM Transactions on Software Engineeringand Methodology, 7(3):215–249, 1998.

[11] Chipworkscorporation. Advanced semiconductor manufacturing handbook. Technical report, Chipworkscorporation, Januari 2000.

[12] A. Datta. Automating the Discovery of As-Is Business Process Models: Probabilistic and Algorithmic Approaches. Information SystemsResearch, 9(3):275–301, 1998.

[13] A.K. Alves de Medeiros. Genetic Process Mining. PhD thesis, Department of Technology Management, Technical University Eindhoven, 2006.

[14] B.F. van Dongen and W.M.P. van der Aalst. Multi-Phase Process Mining: Building Instance Graphs. In P. Atzeni, W. Chu, H. Lu, S. Zhou, and T.W. Ling, editors, International Conference on Conceptual Modeling(ER 2004), volume 3288 of Lecture Notes in Computer Science, pages 362–376. Springer-Verlag, Berlin, 2004.

Accurate Calculation of LDO IR-Drop for SRAMs Using Redhawk EDA Tool

Emad Abu-Shama
Intel Corporation
Austin, TX 78746
emad.y.abu-shama@intel.com

Ziad Sadi
Intel Corporation
Austin, TX 78746
ziad.sadi@intel.com

Amr Elchouemi
Hewlett-Packard Company
Austin, TX 78746
amr.elchouemi@hp.com

Azam Beg[1]
United Arab Emirates University
Al Ain, UAE
abeg@uaeu.ac.ae

Abstract—The purpose of an on-chip *low drop-out* (LDO) voltage regulator is to supply steady voltage to static random access memory (SRAM). With the explosive growth of mobile technologies, the LDOs are becoming more prevalent in the semiconductor industry. However, due to its sensitivity, there is a strict voltage (IR) drop requirement over the power grid network from the regulator to the SRAM. The Redhawk EDA tool allows the designers to model the analog behavior of the LDO – a capability that is not available with other traditional tools. Still Redhawk does not consider the LDOs as voltage sources and can miss the IR drops during analysis. In this paper, we discuss a method that enables Redhawk to view and treat the LDOs as voltage sources to ensure robust power-grid connectivity and to meet the specifications of IR-drop between the LDO and the SRAM.

Keywords—low drop-out voltage regulator, SRAM, IR drop, Redhawk, ploc

1 Introduction

All designs for today's mobile devices are targeting low-power and low-voltage operation to improve the battery life. According to International Technology Roadmap for Semiconductors [1], the oxide thickness in CMOS devices is predicted to drop to 0.5 nm. As there is an exponential relationship between the oxide thickness and the gate-leakage currents, the leakage power would take up a larger share in a chip's total power consumption. In addition, the memory portion of the memory on a *system-on-chip* (SoC) is expected to increase significantly thus occupying much larger die area. Therefore, keeping the leakage-power dissipation of the memory (SRAM) in check would be very vital.

Low drop-out (LDO) voltage regulators are essential components of power management systems on modern chips. Such regulators use feedback to ensure an accurate voltage level at a given current. It becomes imperative for SoC designs to have LDO in order to supply steady voltage to SRAMs to ensure high yields and low-cost design alternatives.

Fig. 1 shows the schematic of an on-chip LDO voltage regulator. An operational transconductance amplifier (OTA) and the voltage divider (R_{fb1} and R_{fb2}) make up the negative feedback mechanism for the regulator. A capacitor in parallel with the divider provides an extra 'zero' for higher stability. V_{REF} is a stable voltage reference, and M_p is a power transistor which can be either pMOS or nMOS type. For low V_{DD}'s, a pMOS transistor is preferable but it needs to cater to precise frequency characteristics.

An LDO is an analog circuit that has several performance metrics such as power consumption, noise, line/load regulation, and settling times. Therefore, designing an LDO that meets the desired performance criteria becomes a multi-objective optimization problem.

It is highly desirable to have the LDO to contribute as little noise as possible. Main source of the noise are: the OTA, the resistive divider, and the voltage reference. Usually, the voltage reference is the largest contributor of the noise, and to suppress it, an external capacitor is commonly used. In scaled LDOs, the noise from the voltage dividers becomes significant as well.

For high output regulation and for better *power-supply rejection ratio* (PSRR), the loop-gain needs to be increased which requires a large-gain amplifier; this in turn needs large output resistance. Additionally, for large load-current and lower-dropout, M_P (pass transistor), the W/L ratio needs to be increased. An adverse effect of these design steps is the reduction in the bandwidth of the regulator. Another method of noise reduction is to use a cross-coupled design that utilizes matched resistors and capacitors. In this case, the PSRR improves but degraded slew-rate negatively affects the transient response of the regulator.

[1]Corresponding author

2 Using Redhawk EDA Tool

As discussed in the previous section, the design of an LDO involves consideration of many characteristics, such as the output levels, stability, noise, etc., simultaneously. Unfortunately, at the full-chip level, LDO has been traditionally modeled as an ideal voltage source, which ignores the circuit's power noise. The Redhawk EDA tool from ANSYS-Apache is capable of precise LDO modeling in order to provide the designers an accurate view of the transient behavior for full-chip power analysis. On-chip LDOs deliver needed voltages to the voltage islands of the chips without requiring extra power pins. LDOs being analog circuits conventionally require Spice simulations for design. Redhawk not only facilitates LDO design but also provides full-chip physical power integrity solutions for SoCs at advanced technology nodes.

In this paper, we have used Redhawk to analyze the IR drop as it is one of the industry standards for power integrity solution. It is fairly easy to setup, use and efficient in runtime and memory. Redhawk tool reports IR drop analysis starting from voltage source (bump) to the ground. These voltage sources are normally defined in the Redhawk *ploc* input file. (ploc contains pad location information based on a chip's floor plan). For circuits that generate voltage sources such as an LDO, there are no associated ploc for them. Therefore, Redhawk tool reports these voltage sources as disconnected from the simulation network and no IR drop analysis is done on the SRAMs. LDOs output pins are normally on higher metal pins which supply a grid that reaches the top metal layer and then descends down on metal until it reaches the SRAMs input pins on specific metal. An LDO designer would like to ensure that the IR-drop from LDOs voltage source to SRAMs input pins does not exceed a certain limit [3]. Although virtual ploc can be placed on higher metal segment of the LDO grid, the results could be overly pessimistic or optimistic. This paper introduces a method for measuring this IR drop in an accurate way.

3 Redhawk's Virtual Voltage Source Approaches

3.1 Virtual Ploc on Top Metal Approach to Model Voltage Source

Traditionally, the user adds virtual ploc on top metal layer that is not connected to real bump when the actual bump is residing outside the design boundary. The user also should consider how far the real bump from the boundary in order to account for the IR drop (R and C value from real bump location to the virtual ploc location). But for LDO, since there is no real bump locations push down from the full chip level, then adding plocs on the top metal will be a challenge if accuracy is concerned. Take for example the top metal LDO to SRAM in Fig. 2, the user placed virtual ploc in the middle of each of the top metal segment while making sure the distances between adjacent ploc will not exceed certain length. Other methods would place a single virtual ploc on the whole top metal grid. The deciding factor is not the virtual ploc

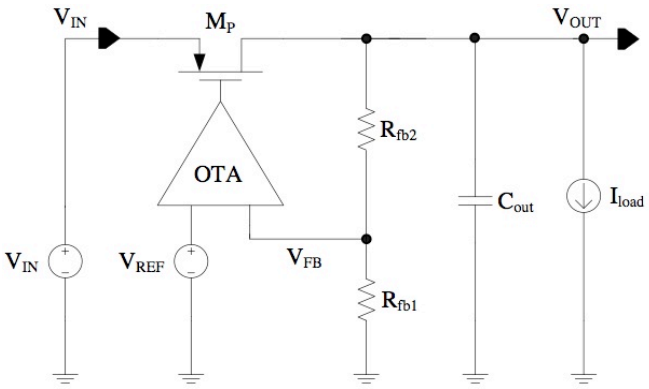


Fig. 1. Schematic of an on-chip LDO voltage regulator [1]

count or location but which approach can be closer to reality and give more accurate results. The approach in Fig. 2 could be optimistic in if the SRAM at the top of the image while the actual LDO at the bottom of image since virtual ploc are very close to destination and may miss some IR drop from the actual LDO source. If we choose to follow the other scenario by placing a single ploc; then the question will be where to place it? Placing it to or bottom could be pessimistic if the SRAM resides on lef or right. Therefore, there will not systematic or automated way to decide the ploc location using these approaches. It has to be hand crafted which is tedious, time consuming and not prune to errors.

3.2 Virtual Ploc on LDO Output Voltage Pins Approach to Model Voltage Source

The perfect voltage source from LDO outputs to supply the SRAM grid would be the LDO output pins. These LDO output pins reside on a specific metal layer by design such as Metal (n), this metal in order to reach to the top metal layer, then it will be connected to Metal (n+1) and so on until reach the metal top then start descending down to reach SRAM destination Metal (m) pins. The perfect location of the virtual ploc will in all the LDO output pins on Metal (n) connecting to next metal layer Metal (n+1) through the VIAs. The size of LDO is relatively large and thus there will be large numbers of VIAs that could be in range of 2K to 5K depending on technology, size or LDO and many other factors. The bottom line is there will be a lot of ploc, and the more the plocs used the more the accurate the results.

In our test case, we have selected a method to select 20 plocs randomly from all these plocs generated in fully automated way. You can experiment increasing or decreasing the number of plocs to be used based on technology you are using to determine what the best coverage you need is. From some experiments we ran on based on specific technology, voltage and IR limit target. We found the IR accuracy loss from having all plocs defined on all the VIAs verse selecting only random 20 was less than 2%, while accuracy loss compared to the first method, the accuracy loss was as high

as 20% of the target IR limit. Below is the pseudo-code of finding these plocs and for randomly selecting 20 plocs:

```
ldo_output_pins = list of possible LDO output voltage pin names if using more than one LDO
open output file
set virtual ploc starting counter
for every pin of ldo_output_pins do the following
    get the LDO instance name for current pin and on which metal layer this pin is on
    find the cell location
    find the net name driven by this instance pin
    get the coordinates of all the VIAs connected from instance pin to next metal layer and calculate the VIA center point
    calculate how many VIAs we have
    determine the 20 VIAs to be used as virtual ploc
    write out to the file virtual plocs
end for loop
```

Fig. 3 shows the top metal layer in blue and LDO cell boundary in yellow. You can see the 20 virtual plocs distributed randomly over the LDO output voltage pins.

At this point, by appending these virtual plocs to the master ploc file that used in Redhawk run, then Redhawk will start reporting the entire IR drop from LDO virtual plocs to the SRAM cells. But since we are interested only on reporting the IR drop from LDO pin to SRAM pin, we used the Redhawk built-in command which automatically provides a sorted file:

```
report dvd -net <the net driven by LDO output voltage pins> -o <output file name> -metal <SRAM input voltage pin metal layer>
```

4 Conclusion

In this paper, we discussed some of the methods that have been used in the industry to model IR drop from LDO to SRAM cells. We presented a new approach that involved Redhawk tool to analyze the IR drop from LDO to SRAM with a higher level of accuracy. We anticipate that the proposed method will be useful for both the industry practitioners and the researchers alike.

5 References

[1] 2007 International Technology Roadmap for Semiconductors, [online]. Available http://public.itrs.net.

[2] G. Razavipour, A. Afzali-Kusha, and M. Pedram, "Design and Analysis of Two Low-Power SRAM Cell Structures," IEEE Trans. VLSI, vol. 17, no. 10, 1999, pp. 1551–1555.

[3] Milliken, R.J.; Silva-Martinez, J.; Sanchez-Sinencio, E., "Full On-Chip CMOS Low-Dropout Voltage Regulator," IEEE Trans. Circ. and Syst. I: Regular Papers, vol. 54, no. 9, 2007, pp. 1879–1890.

[4] Lv Xiaopeng, Bian Qiang, Yue Suge, "An On-Chip Low Drop-Out Voltage Regulator with 150 mA Driving Capability," Proc. World Congr. Eng. and Comp. Sci. (WCECS), Oct. 2011, San Francisco, CA, pp. 1–4.

[5] Kanj, R.; Joshi, R.; Nassif, S., "SRAM Yield Sensitivity to Supply Voltage Fluctuations and Its Implications on Vmin," IEEE Int. Conf. Integ. Circ. Des. and Techn. (ICICDT '07), 2007, pp. 1–4.

[6] Chun-Yen Tseng; Po-Chiun Huang; Li-Wen Wang, "An integrated linear regulator with fast output voltage transition for SRAM yield improvement," Proc. IEEE Asian Solid-State Circ. Conf. (A-SSCC), 2009, pp. 329–332.

[7] ANSYS Apache Redhawk tool, [online]. Available http://www.apache-da.com/products/redhawk

[8] Chen, H.H.; Neely, J.S., "Interconnect and circuit modeling techniques for full-chip power supply noise analysis," IEEE Trans. Components, Packaging, and Manufacturing Tech., Part B: Advanced Packaging, vol. 21, no. 3, 1998, pp. 209–215.

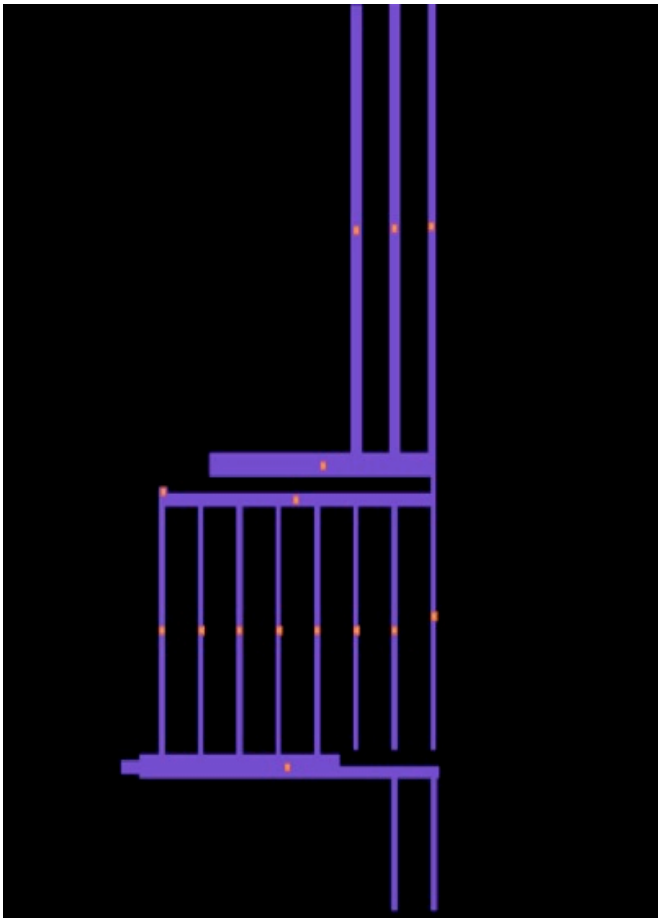


Fig. 2. Virtual ploc on top metal layer

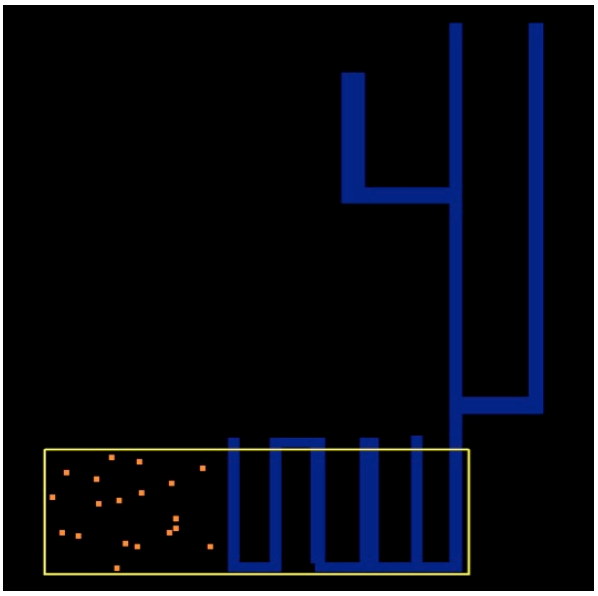


Fig. 3. Twenty virtual ploc's over the LDO

Reducing the Modeling Error in Different Observer Based Regulators

Cs. Bányász and L. Keviczky

Computer and Automation Research Institute and
MTA-BME Control Engineering Research Group
Hungarian Academy of Sciences
H-1111 Budapest, Kende u 13-17, HUNGARY

Abstract - *An equivalent transfer function representation (*TFR*) is introduced to study the state-feedback/observer (*SFO*) topologies of control systems. This approach is used to explain why an observer can radically reduce even large model errors. Then the same principle is combined with YOULA-parameterization (*YP*) introducing a new class of regulators.*

Keywords: Observer, state-feedback, model error, YOULA-parameterization

1 Introduction, the state feedback (*SF*)

It is a well known methodology to use the state variable representations (*SVR*) of linear time invariant (*LTI*) single input - single output (*SISO*) systems. The *SVR* proved to be excellent tool to implement both *LQR* (*Linear system - Quadratic criterion - Regulator*) control and pole placement design. Thousands of theoretical considerations mostly concentrate on the irregularities and special structures in the *SVR* appearing and almost no publications deal with the model error properties of these systems. It is possible to find a proper way to discuss and investigate the limitations if someone replaces the *SVR* by their transfer function representations (*TFR*).

Consider a *SISO* continuous time (t) *LTI* dynamic plant described by the *SVR*

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}u \quad ; \quad P = \frac{\mathcal{B}}{\mathcal{A}} = \boldsymbol{c}^{\mathrm{T}}\left(s\,\boldsymbol{I} - \boldsymbol{A}\right)^{-1}\boldsymbol{b} \qquad (1)$$
$$y = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}$$

Here u, y and $\boldsymbol{x}$ are the input, output and state variables of the process to be controlled and $^{\mathrm{T}}$ stands for transposition. Furthermore P is the *TFR* of the open-loop system with the numerator and denominator polynomials

$$\mathcal{B}(s) = s^n + b_1 s^{n-1} + \ldots + b_{n-1}s + b_n \qquad (2)$$

$$\mathcal{A}(s) = s^n + a_1 s^{n-1} + \ldots + a_{n-1}s + a_n \qquad (3)$$

If the feedback is restricted to a linear *SF*, then the classical solution can be written as

$$u = k_{\mathrm{r}}\, r - \boldsymbol{k}^{\mathrm{T}}\boldsymbol{x} \qquad (4)$$

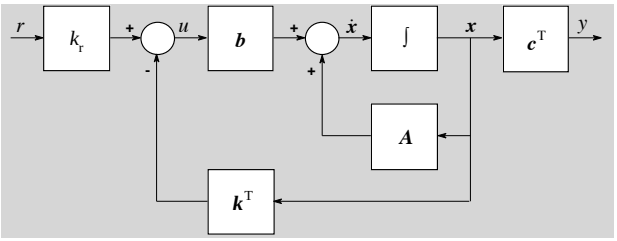


Figure 1. Block diagram of the *SF* of *SISO LTI* system

The resulting closed-loop system is shown in Fig. 1, where r is the reference signal, k_{r} is a calibrating constant and $\boldsymbol{k}^{\mathrm{T}}$ is the linear *SF* vector. It is easy to check that the complementary sensitivity function (*CSF*) from the reference signal r to the output y is

$$T_{\mathrm{ry}}(s) = k_{\mathrm{r}}\boldsymbol{c}^{\mathrm{T}}\left(s\,\boldsymbol{I} - \boldsymbol{A} + \boldsymbol{b}\,\boldsymbol{k}^{\mathrm{T}}\right)^{-1}\boldsymbol{b} = \frac{k_{\mathrm{r}}\,P}{1 + \boldsymbol{k}^{\mathrm{T}}\left(s\,\boldsymbol{I} - \boldsymbol{A}\right)^{-1}\boldsymbol{b}} \qquad (5)$$

where k_{r} is obtained by requiring that the static gain of T_{ry} should be equal to one.

The usual classical design goal is to determine the feedback gain $\boldsymbol{k}^{\mathrm{T}}$ so that the closed-loop system has the characteristic polynomial

$$\mathcal{R}(s) = s^n + r_1 s^{n-1} + \ldots + r_{n-1}s + r_n \qquad (6)$$

The solution formally makes the characteristic polynomial of the closed-loop equal to the desired polynomial ("placed poles")

$$\det\left(s\,\boldsymbol{I} - \boldsymbol{A} + \boldsymbol{b}\,\boldsymbol{k}^{\mathrm{T}}\right) = \mathcal{R}(s) \qquad (7)$$

The solution always exists if P is controllable. If the *TFR* of the process is known then one can easily form a controllable canonical form, and the feedback gain is obtained from

$$\boldsymbol{k}_{\mathrm{c}}^{\mathrm{T}} = \left[r_1 - a_1, r_2 - a_2, \ldots, r_n - a_n\right] \qquad (8)$$

It is easy to derive from equation (6) that $T_{\mathrm{ry}}(s)$ is now

$$T_{\mathrm{ry}}(s)=\frac{k_{\mathrm{r}}\,\mathcal{B}(s)}{\mathcal{R}(s)} \tag{9}$$

i.e., besides reaching the desired pole-placement the *SF* leaves the open-loop zeros untouched.

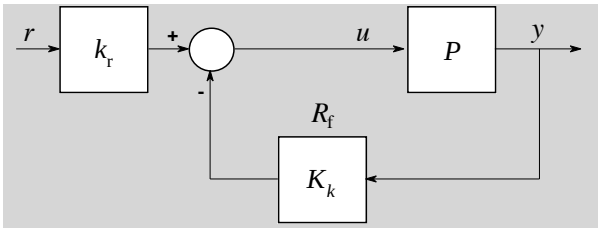


Figure 2. Equivalent schemes of *SF* using *TFR* forms

If we want to express the operation of the *SF* by equivalent scheme using *TFR* forms, Fig. 3 can be used, where the feedback regulator $R_{\mathrm{f}}=K_k$ is obtained from the basic equation of the closed-loop

$$T_{\mathrm{ry}}(s)=\frac{k_{\mathrm{r}}\,\mathcal{B}(s)}{\mathcal{R}(s)}=\frac{k_{\mathrm{r}}\,\mathcal{B}(s)}{\mathcal{A}(s)+\mathcal{K}(s)}=\frac{k_{\mathrm{r}}\,P}{1+K_k\,P} \tag{10}$$

which clearly shows, that the open-loop poles remain unchanged and the closed-loop poles will be the required ones. Here it is obtained that

$$R_{\mathrm{f}}=K_k(s)=\frac{\mathcal{K}(s)}{\mathcal{B}(s)}=\frac{\mathcal{R}(s)-\mathcal{A}(s)}{\mathcal{B}(s)}=\frac{\boldsymbol{k}^{\mathrm{T}}\left(s\boldsymbol{I}-\boldsymbol{A}\right)^{-1}\boldsymbol{b}}{\boldsymbol{c}^{\mathrm{T}}\left(s\boldsymbol{I}-\boldsymbol{A}\right)^{-1}\boldsymbol{b}} \tag{11}$$

It is not a very frequently discussed question using *SF* whether the stability of the process polynomials in P are required or not for the closed-loop stability. Figure 2b shows the real operation of the *SF*. The polynomial $\mathcal{R}(s)-\mathcal{A}(s)$ in the feedback path stabilizes $1/\mathcal{A}(s)$, the denominator of the process, even if it is an unstable one, and places the required poles via the design polynomial $\mathcal{R}(s)$. The numerator $\mathcal{B}(s)$ of the plant is outside of this process and remains unchanged. Observe that this regulation can be interpreted and realized only if *SF* is used. Figure 3a shows another interpretation of the controller using only *TFR*s. This equivalent scheme, however, can be realized by the indicated transfer functions, only if the process P itself is inverse stable (*IS*), i.e., if $\mathcal{B}(s)$ is a stable polynomial.

The final conclusion is that the *SF* stabilizes all observable plants, however, the zeros of the process remain unchanged.

2 Observer-based state-feedback

The practical applicability of the *SF* theory was introduced by the development of the observers capable to calculate the unmeasured state variables. The most general *SF*/observer (*SFO*) topology is shown in Fig. 3.

In the general basic *SFO* scheme the controller consists of two parts: one observer and one *SF*. The observer calculates the estimated state variable $\hat{\boldsymbol{x}}$ and the estimated process output $\hat{y}$.

The feedback gain $\boldsymbol{k}^{\mathrm{T}}$ is computed, if all state variables could be measured, using $\hat{\boldsymbol{x}}$, so

$$u=k_{\mathrm{r}}\,r-\boldsymbol{k}^{\mathrm{T}}\,\hat{\boldsymbol{x}} \tag{12}$$

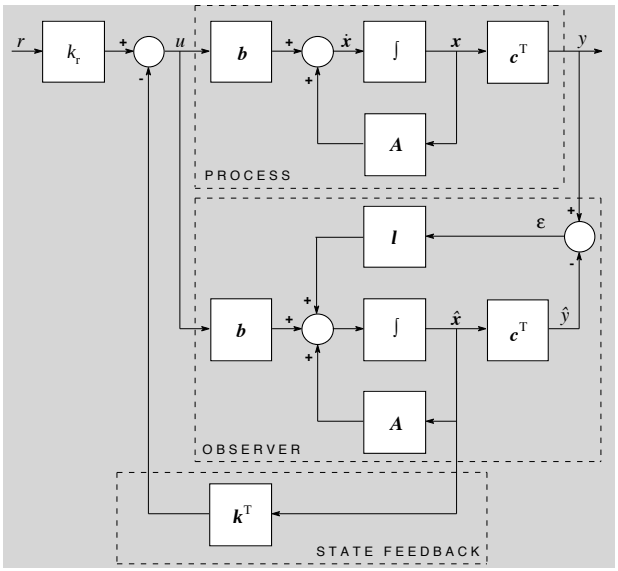


Figure 3. The general basic *SFO* scheme

The observer modifies the internal model of the process by introducing a proportional feedback $\boldsymbol{k}$ from the error $\varepsilon=y-\hat{y}$. It is long and not easy derivation to prove that the *CSF* from the reference signal r to the output y is

$$T_{\mathrm{ry}}(s)=\frac{k_{\mathrm{r}}\,\boldsymbol{c}^{\mathrm{T}}\left(s\boldsymbol{I}-\boldsymbol{A}\right)^{-1}\boldsymbol{b}}{1+\boldsymbol{k}^{\mathrm{T}}\left(s\boldsymbol{I}-\boldsymbol{A}\right)^{-1}\boldsymbol{b}}=k_{\mathrm{r}}\,\mathcal{B}(s)/\mathcal{R}(s) \tag{13}$$

thus surprisingly exactly the same as it was for the simple *SF* case (see equation (5) and [3] for a nice short derivation). It is also possible to compute an equivalent feedback regulator, as

$$R_{\mathrm{f}}=\frac{\boldsymbol{k}^{\mathrm{T}}\left(s\boldsymbol{I}-\boldsymbol{A}+\boldsymbol{b}\,\boldsymbol{k}^{\mathrm{T}}\right)^{-1}\boldsymbol{l}}{1+\boldsymbol{c}^{\mathrm{T}}\left(s\boldsymbol{I}-\boldsymbol{A}+\boldsymbol{b}\,\boldsymbol{k}^{\mathrm{T}}\right)^{-1}\boldsymbol{l}} \tag{14}$$

which has a much more complex structure than what was in (11). Introducing the state error

$$\tilde{\boldsymbol{x}}=\boldsymbol{x}-\hat{\boldsymbol{x}} \tag{15}$$

the dynamics of the observer is basically determined by the state error equation

$$\frac{\mathrm{d}\,\tilde{\boldsymbol{x}}}{\mathrm{d}\,t}=\left(\boldsymbol{A}-\boldsymbol{l}\,\boldsymbol{c}^{\mathrm{T}}\right)\hat{\boldsymbol{x}} \tag{16}$$

The usual classical design goal for the observer is to determine the observer feedback gain $\boldsymbol{k}$ so that the dynamic system (16) has the characteristic polynomial

$$\mathcal{Q}(s)=s^n+q_1s^{n-1}+\ldots+q_{n-1}s+q_n \tag{17}$$

The solution formally means making the characteristic polynomial equal to the desired polynomial

$$\det\left(s\boldsymbol{I}-\boldsymbol{A}+\boldsymbol{l}\,\boldsymbol{c}^{\mathrm{T}}\right)=\mathcal{Q}(s) \tag{18}$$

The solution always exists if P is observable. If the

TFR of the process is known then one can easily form an observable canonical form, when the feedback gain is obtained from

$$\boldsymbol{l}_{\mathrm{o}} = \left[q_1 - a_1, q_2 - a_2, \ldots, q_n - a_n\right]^{\mathrm{T}} \tag{19}$$

The *SVR* of the entire *SFO* closed-loop system can be described by

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix}\boldsymbol{x}\\ \tilde{\boldsymbol{x}}\end{bmatrix} = \begin{bmatrix}\boldsymbol{A}-\boldsymbol{b}\boldsymbol{k}^{\mathrm{T}} & \boldsymbol{b}\boldsymbol{k}^{\mathrm{T}}\\ \boldsymbol{0} & \boldsymbol{A}-\boldsymbol{l}\boldsymbol{c}^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\boldsymbol{x}\\ \tilde{\boldsymbol{x}}\end{bmatrix} + \begin{bmatrix}k_{\mathrm{r}}\boldsymbol{b}\\ \boldsymbol{0}\end{bmatrix} r$$

$$e = y - \hat{y} = \boldsymbol{c}^{\mathrm{T}}\tilde{\boldsymbol{x}} \tag{20}$$

Since the matrix on the right-hand side is block diagonal the characteristic equation of the closed loop system is

$$\det\left(s\boldsymbol{I} - \boldsymbol{A} + \boldsymbol{b}\boldsymbol{k}^{\mathrm{T}}\right)\det\left(s\boldsymbol{I} - \boldsymbol{A} + \boldsymbol{l}\boldsymbol{c}^{\mathrm{T}}\right) = \mathcal{R}(s)\mathcal{Q}(s) \tag{21}$$

This polynomial is the product of two terms: one which is used for the SF design and the other which is used for the observer design. In spite of (21) it is interesting to observe that $\mathcal{Q}(s)$ does not appear in the $T_{\mathrm{ry}}(s)$ given by (13). The explanation of this phenomenon can be given by the investigation of the internal topology of the equivalent TFR forms of the *SFO* scheme.

3 Equivalent *TFR* forms of the *SFO* scheme

Introducing Figs. 2a-b the technique using equivalent *TFR* forms of *SF* has been discussed above. To get a more general procedure consider Figs. 2a-b again in the forms presented in Figs. 4a-b. It follows from Fig. 5 that the serial compensator R_{s} can be calculated by

$$R_{\mathrm{s}} = \frac{1}{1+R_{\mathrm{s}}P} = \frac{1}{1+K_k P} = \frac{\mathcal{A}(s)}{\mathcal{A}(s)+\mathcal{K}(s)} = \frac{\mathcal{A}(s)}{\mathcal{R}(s)} \tag{22}$$

Observe that this serial compensator can not be applied for unstable processes because of the full pole cancellation in R_{s}, inspite the fact that $k_{\mathrm{r}}R_{\mathrm{s}}P$ ensures the same overall transfer function T_{ry}. Finding the equivalent *TFR* form an auxiliary internal signal $\bar{x}$ is introduced and used (which is not equal to $\boldsymbol{x}$) indicating that finally both the *SF* and the observer use a *SISO* filter realizing their effect. The difference is that they use internal state variable vectors ($\boldsymbol{x}$, $\hat{\boldsymbol{x}}$, $\tilde{\boldsymbol{x}}$, etc.) instead of scalar ones. It is always possible to find input/output equivalence between these representations. Using this approach the general basic *SFO* scheme in Fig. 3 can be redrawn into another topology shown in Fig. 5.

After some long, but straightforward block manipulations the equivalent *SFO* scheme can be transformed into another unity feedback closed-loop form given in Fig. 6.

As a consequence of the previous discussions the following *TFR* forms are selected:

$$K_k(s) = \frac{\mathcal{K}(s)}{\mathcal{B}(s)} \text{ and } K_l(s) = \frac{\mathcal{L}(s)}{\mathcal{B}(s)} \tag{23}$$

where the pole-placement design goals for both the *SF* and the observer dynamics require

$$\mathcal{K}(s) = \mathcal{R}(s) - \mathcal{A}(s) \quad \text{and} \quad \mathcal{L}(s) = \mathcal{Q}(s) - \mathcal{A}(s) \tag{24}$$

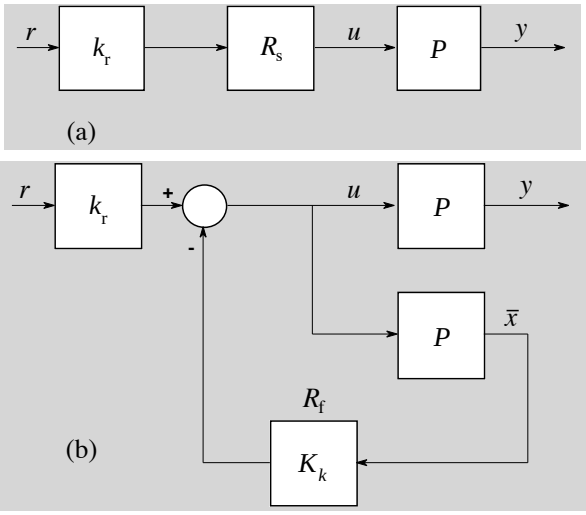


Figure 4. Equivalent schemes of *SF* using *TFR* forms

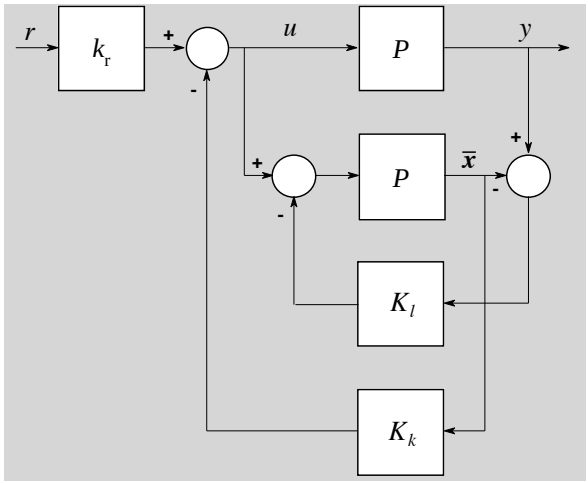


Figure 5. Equivalent topology of the general basic *SFO* scheme using *TFR* forms

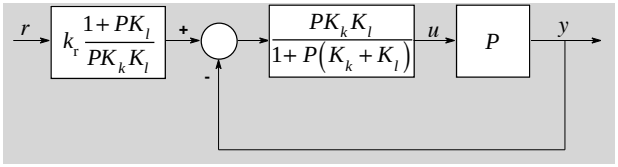


Figure 6. Reduced equivalent topology of the general basic *SFO* scheme

It is interesting to observe that the transfer function of the closed-loop in Fig. 6 has a very special structure

$$\frac{P^2K_kK_l}{1+P(K_k+K_l)+P^2K_kK_l} = \frac{PK_k}{1+PK_k}\frac{PK_l}{1+PK_l} = \frac{\mathcal{K}}{\mathcal{R}}\frac{\mathcal{L}}{\mathcal{Q}} \tag{25}$$

which is formally two simpler closed-loops cascaded, which dynamically completely corresponds to the characteristic equation (21). The overall transfer function of the *SFO* system is

$$T_{\mathrm{ry}}(s) = k_{\mathrm{r}}\frac{1+PK_l}{PK_kK_l}\frac{PK_k}{1+PK_k}\frac{PK_l}{1+PK_l} = \frac{k_{\mathrm{r}}P}{1+PK_k} = \frac{k_{\mathrm{r}}\mathcal{B}}{\mathcal{R}} \tag{26}$$

which is equal to (13) as expected: the poles introduced by the observer do not appear in the tracking dynamics of the *SFO* system. This behavior can be well seen in Fig. 7.

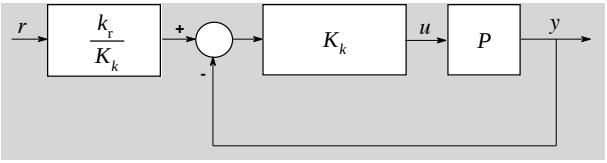


Figure 7. The overall tracking structure of the *SFO* scheme

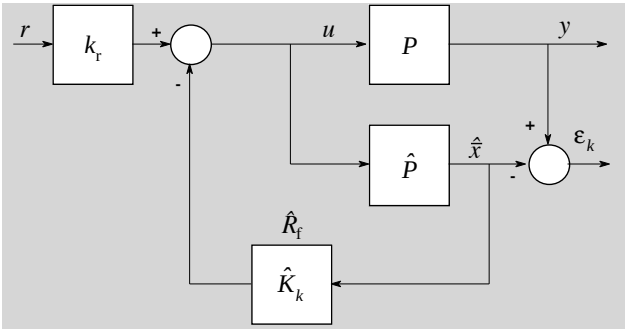


Figure 8. The model based *SF* scheme and error

4 Model error properties

The above widely applied methodology has a common problem, that in all regulator and observer equations the true process P is used instead of the estimated model $\hat{P}$ of the process. A logical notation should be if $\hat{\boldsymbol{A}},\hat{\boldsymbol{b}},\hat{\boldsymbol{c}}^{\mathrm{T}}$, belonging to $\hat{P}$, would be used instead of $\boldsymbol{A},\boldsymbol{b},\boldsymbol{c}^{\mathrm{T}}$, representing P. The equivalent *TFR* form of the *SF* using the model of the process is shown in Fig. 8.

Inspite of the practical reality of this scheme a major drawback raises with the parallel model *TFR* of the *SF* method, because it can not stabilize unstable plants. A model-based equivalent serial compensator (22) can not provide stable operation with pole cancellation. The *SF* is operable if the exact states are available for feedback! So the parallel scheme in Fig. 8 is used only to compute the model error. Using (13) the $\hat{T}_{\mathrm{ry}}$ model-based version of T_{ry} is

$$\hat{T}_{\mathrm{ry}} = \frac{k_{\mathrm{r}} P}{1+K_k \hat{P}} = \frac{k_{\mathrm{r}} \mathcal{B}}{\mathcal{R}} \frac{\hat{\mathcal{A}}}{\mathcal{A}} = T_{\mathrm{ry}} \frac{\hat{\mathcal{A}}}{\mathcal{A}} \tag{27}$$

and its relative uncertainty

$$\ell_{\mathrm{T}} = \frac{\hat{T}_{\mathrm{ry}} - T_{\mathrm{ry}}}{\hat{T}_{\mathrm{ry}}} = \frac{\hat{\mathcal{A}}-\mathcal{A}}{\mathcal{A}} = \ell_{\mathcal{A}} \tag{28}$$

which shows that $\ell_{\mathrm{T}} = 0$ for $\ell_{\mathcal{A}} = 0$. Introducing the additive $\Delta = P - \hat{P}$ and relative plant model error

$$\ell = \frac{\Delta}{\hat{P}} = \frac{P-\hat{P}}{\hat{P}} \tag{29}$$

the modeling error ε_k in Fig. 8 can be expressed as

$$\varepsilon_k = \frac{k_{\mathrm{r}} \hat{\mathcal{B}}}{\mathcal{R}} \ell r = T_{\mathrm{ry}} \frac{\hat{\mathcal{B}}}{\mathcal{B}} \ell r = \hat{P}\, \ell u \tag{30}$$

After some long but straightforward computations

$$\varepsilon_l = \frac{\hat{P}}{1+K_l \hat{P}} \ell u = \frac{\hat{\mathcal{B}}}{\mathcal{Q}} \ell u = \frac{1}{1+K_l \hat{P}} \varepsilon_k \tag{31}$$

is obtained. Equation (31) clearly shows the influence of the *SFO* scheme, because it decreases the modeling error ε_k by $(1+K_l \hat{P})$. Selecting fast observer poles, one can reach quite small "virtual" modeling error ε_l in the major frequency domains of the tracking task.

Inspite of the above analysis the *SFO* scheme is widely applied in the practice with model-based *SVR*, so it is interesting how the model-based scheme in Fig. 9 influences the original modeling error ε_k.

Besides the radical model error attenuating behavior of the model-based *SFO* scheme, unfortunately it has a very important drawback, the nice cascade structure changes to

$$\left.\frac{\hat{P}^2 K_k K_l \left(1+\ell\right)}{1+\hat{P}\left(K_k+K_l\right)+\hat{P}^2 K_k K_l \left(1+\ell\right)}\right|_{\ell \to 0} = \frac{\mathcal{K}}{\mathcal{R}} \frac{\mathcal{L}}{\mathcal{Q}} \tag{32}$$

which form is not factorable except for the exact model matching case, when $\ell \to 0$. On the basis of Fig. 9 and (32) it is easy to see that the poles of the observer feedback loop remains unchanged using the placement design equation form model-based *SFO* (19). However, in this case the pole placement equation (8) is no longer valid. The only solution is to use the available model of the process, in this case $\hat{\mathcal{A}}$, and

$$\hat{\boldsymbol{k}}_{\mathrm{c}}^{\mathrm{T}} = \left[\, p_1 - \hat{a}_1, p_2 - \hat{a}_2, \ldots, p_n - \hat{a}_n \,\right] \tag{33}$$

for the pole placing equation.

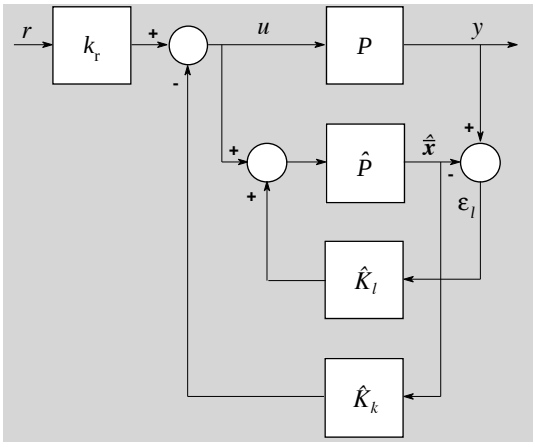


Figure 9. Model based *SFO* scheme with *TFR* forms

Because this design ensures the required poles only for small ℓ (see (33)), a serious robust stability investigation is required first. Next it is important to investigate where the actual pole is located for non zero ℓ, so how big the performance loss is coming from the model based *SFR*. These steps are usually neglected in most of the published papers, books and applications.

5 Introducing the observer based YOULA-regulator

For open-loop stable processes the all realizable stabilizing (*ARS*) model based regulator $\hat{C}$ is the *YOULA-parameterized* one:

$$\hat{C}\left(\hat{P}\right)=\left.\frac{Q}{1-Q\hat{P}}\right|_{\hat{P}\to P}=\frac{Q}{1-QP}=C\left(P\right) \tag{34}$$

where the "parameter" Q ranges over all proper ($Q\left(\omega=\infty\right)$ is finite), stable transfer functions [5], [6], see Fig. 10a.

It is important to know that the *Y-parameterized* closed-loop with the *ARS* regulator is equivalent to the well-known form of the so-called *Internal Model Control* (*IMC*) principle [6] based structure shown in Fig. 11b.

Q is anyway the transfer function from r to u and the closed-loop transfer function (i.e., *CSF*) for $\hat{P}=P$, when $\ell\to 0$

$$\hat{T}_{\mathrm{ry}}=\frac{\hat{C}P}{1+\hat{C}P}=QP\left.\frac{1+\ell}{1+\left(1-QP\right)\ell}\right|_{\ell\to 0}=QP=T_{\mathrm{ry}} \tag{35}$$

is linear (and hence convex) in Q.

It is interesting to compute the relative error ℓ_{T} of $\hat{T}_{\mathrm{ry}}$

$$\ell_{\mathrm{T}}=\frac{T_{\mathrm{ry}}-\hat{T}_{\mathrm{ry}}}{\hat{T}_{\mathrm{ry}}}=Q\left(P-\hat{P}\right)=QP\frac{\ell}{1+\ell}=T_{\mathrm{ry}}\frac{\ell}{1+\ell} \tag{36}$$

The equivalent *IMC* structure performs the feedback from the model error ε_Q. Similarly to the *SFO* scheme it is possible to construct an internal closed-loop, which virtually reduces the model error to

$$\varepsilon_l=\frac{1}{1+\hat{K}_l\hat{P}}\left(y-\hat{P}u\right)=\frac{1}{1+\hat{L}_l}\varepsilon_Q=\hat{H}\,\varepsilon_Q\,;\,\hat{L}_l=\hat{K}_l\hat{P} \tag{37}$$

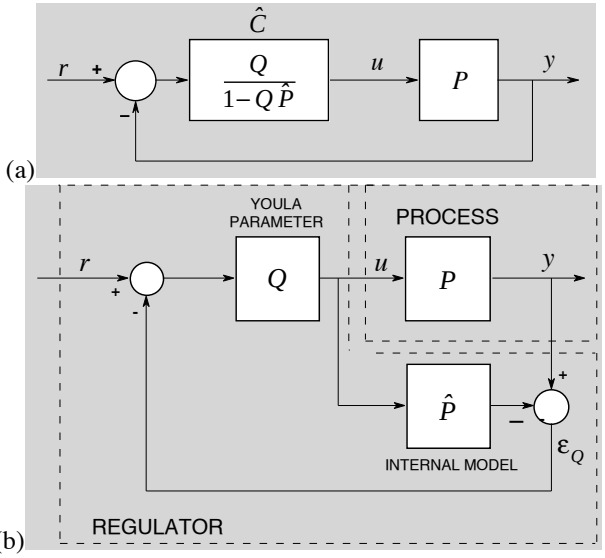


Figure 10. The equivalent *IMC* structure of an *ARS* regulator

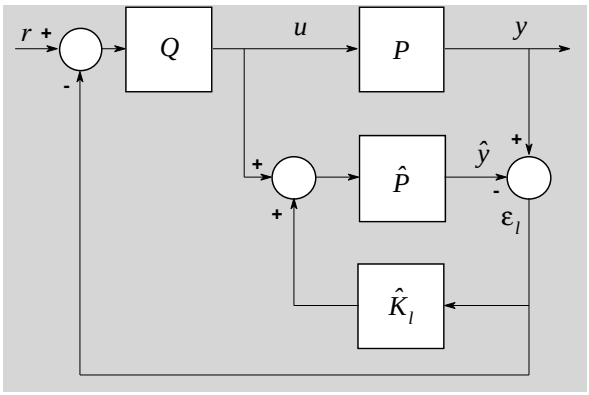


Figure 11. The observer-based *IMC* structure

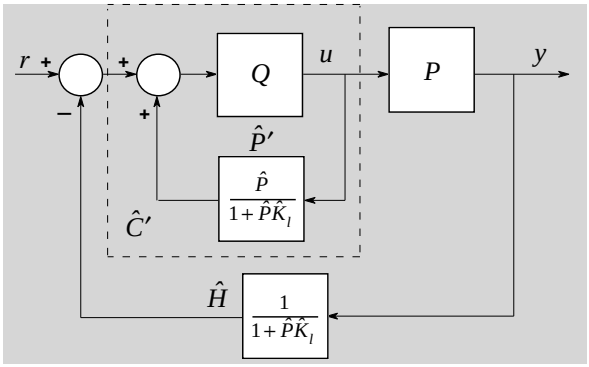


Figure 12. Equivalent closed-loop for the observer-based *IMC* structure

and performs the feedback from ε_l (see Fig. 11), where $\hat{L}_l$ is the internal loop transfer function. In this case the resulting closed-loop will change to the scheme shown in Fig. 12.

This means that the introduction of the observer feedback changes the *YOULA-parameterized* regulator to

$$\hat{C}'\left(\hat{P}'\right)=\frac{Q}{1-Q\dfrac{\hat{P}}{1+\hat{K}_l\hat{P}}}=\frac{Q\left(1+\hat{K}_l\hat{P}\right)}{1+\hat{K}_l\hat{P}-Q\hat{P}} \tag{38}$$

The form of $\hat{C}'$ shows that the regulator virtually controls a fictitious plant $\hat{P}'$, which is also demonstrated in Fig. 12. Here the fictitious plant is

$$\hat{P}'=\frac{\hat{P}}{1+\hat{K}_l\hat{P}}=\frac{\hat{P}}{1+\hat{L}_l} \tag{39}$$

The closed-loop transfer function is now

$$\hat{T}'_{\mathrm{ry}}=\frac{\hat{C}'P}{1+\hat{C}'P}=\left.\frac{QP}{1+QP\dfrac{1}{1+\hat{K}_l\hat{P}}\dfrac{\ell}{1+\ell}}\right|_{\ell\to 0}=QP=T_{\mathrm{ry}} \tag{40}$$

The relative error ℓ'_{T} of $\hat{T}'_{\mathrm{ry}}$ becomes

$$\ell'_{\mathrm{T}}=\frac{T_{\mathrm{ry}}-\hat{T}'_{\mathrm{ry}}}{\hat{T}'_{\mathrm{ry}}}=QP\frac{\ell}{1+\ell}\frac{1}{\left(1+\hat{K}_l\hat{P}\right)}=\ell_{\mathrm{T}}\frac{1}{1+\hat{L}_l} \tag{41}$$

what is smaller than ℓ_{T}. The reduction is by $\hat{H} = 1/(1+\hat{L}_l)$.

6 An observer based *PID* regulator

The ideal form of a YOULA-regulator based on reference model design [5] is

$$C_{\mathrm{id}} = \frac{\left(R_{\mathrm{n}}P^{-1}\right)}{1-\left(R_{\mathrm{n}}P^{-1}\right)P} = \frac{Q}{1-QP} = \frac{R_{\mathrm{n}}}{1-R_{\mathrm{n}}}P^{-1} \tag{42}$$

when the inverse of the process is realizable and stable. Here the operation of R_{n} can be considered a reference model (desired system dynamics). It is generally required that the reference model has to be strictly proper with unit static gain, i.e., $R_{\mathrm{n}}(\omega = 0) = 1$.

For a simple, but robust *PID* regulator design method assume that the process can be well approximated by its two major time constants, i.e.,

$$P \cong \frac{A}{\mathcal{A}_2} \tag{43}$$

where

$$\mathcal{A}_2 = (1+sT_1)(1+sT_2) \tag{44}$$

According to (42) the ideal YOULA-regulator is

$$C_{\mathrm{id}} = \frac{R_{\mathrm{n}}P^{-1}}{1-R_{\mathrm{n}}} = \frac{R_{\mathrm{n}}(1+sT_1)(1+sT_2)}{A(1-R_{\mathrm{n}})} \quad ; \quad T_1 > T_2 \tag{45}$$

Let the reference model R_{n} be of first order

$$R_{\mathrm{n}} = \frac{1}{1+sT_{\mathrm{n}}} \tag{46}$$

which means that the first term of the regulator is an integrator

$$\frac{R_{\mathrm{n}}}{1-R_{\mathrm{n}}} = \frac{\dfrac{1}{1+sT_{\mathrm{n}}}}{1-\dfrac{1}{1+sT_{\mathrm{n}}}} = \frac{1}{1+sT_{\mathrm{n}}-1} = \frac{1}{sT_{\mathrm{n}}} \tag{47}$$

whose integrating time is equal to the time constant of the reference model. Thus the resulting regulator corresponds to the design principle, i.e., it is an ideal *PID* regulator

$$C_{\mathrm{PID}} = A_{\mathrm{PID}}\frac{(1+sT_{\mathrm{I}})(1+sT_{\mathrm{D}})}{sT_{\mathrm{I}}} = A_{\mathrm{PID}}\frac{(1+sT_1)(1+sT_2)}{sT_1} \tag{48}$$

with

$$A_{\mathrm{PID}} = \frac{T_1}{AT_{\mathrm{n}}} \quad ; \quad T_{\mathrm{I}} = T_1 \quad ; \quad T_{\mathrm{D}} = T_2 \tag{49}$$

The YOULA-parameter Q in the ideal regulator is

$$Q = R_{\mathrm{n}}P^{-1} = \frac{1}{A}\frac{(1+sT_1)(1+sT_2)}{1+sT_{\mathrm{n}}} \tag{50}$$

It is not necessary, but desirable to ensure the realizability, i.e., it is reasonable to use

$$Q = R_{\mathrm{n}}P^{-1} = \frac{1}{A}\frac{(1+sT_1)(1+sT_2)}{(1+sT_{\mathrm{n}})(1+sT)} \tag{51}$$

where T can be considered as the time constant of the derivative action ($0.1T_{\mathrm{D}} \le T \le 0.5T_{\mathrm{D}}$). The regulator $\hat{C}'$ and the feedback term $\hat{H}$ must be always realizable. In the practice the *PID* regulator and the YOULA-parameter is always model-based, so

$$\hat{C}_{\mathrm{PID}}(\hat{P}) = \hat{A}_{\mathrm{PID}}\frac{(1+s\hat{T}_1)(1+s\hat{T}_2)}{s\hat{T}_1}\,;\ \hat{A}_{\mathrm{PID}} = \frac{\hat{T}_1}{\hat{A}T_{\mathrm{n}}} \tag{52}$$

$$\hat{Q} = R_{\mathrm{n}}\hat{P}^{-1} = \frac{1}{\hat{A}}\frac{(1+s\hat{T}_1)(1+s\hat{T}_2)}{1+sT_{\mathrm{n}}} \tag{53}$$

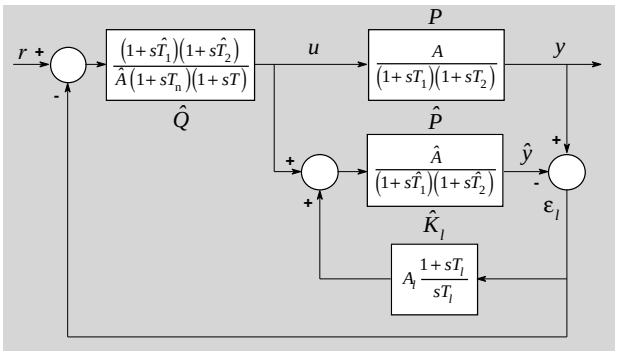


Figure 13. An observer based PID regulator

The sheme of the observer based *PID* regulator is shown in Fig. 13, where a simple *PI* regulator

$$\hat{K}_l = A_l\frac{1+sT_l}{sT_l} \tag{54}$$

is applied in the observer-loop. Here T_l must be in the range of T, i.e., considerably smaller than T_1 and T_2.

Note that the frequency characteristic of $\hat{H}$ cannot be easily designed to reach a proper error suppression. For example, it is almost impossible to design a good realizable high cut filter in this architecture. The high frequency domain is always more interesting to speed up a control loop, so the target of the future research is how to select $\hat{K}_l$ for the desired shape of $\hat{H}$.

7 Simulation experiments

The simulation experiments were performed in using the observer based *PID* scheme shown in Fig. 13.

Example 1

The process parameters are: $T_1 = 20$, $T_2 = 10$ and $A = 1$. The model parameters are: $\hat{T}_1 = 25$, $\hat{T}_2 = 12$ and $\hat{A} = 1.2$. The purpose of the regulation is to speed up the basic step response by 4, i.e., $T_n = 5$ is selected in the first order R_n. In the observer loop a

simple proportional regulator $\hat{K}_l = 0.01$ is applied. The ideal form of Q (50) was used. Figure 14 shows some step responses in the operation of the observer based *PID* regulator.

It is easy to see that the $\hat{T}'_{\mathrm{ry}}$ very well approximates R_n in the high frequencies (for small time values) inspite of the very bad model $\hat{P}$.

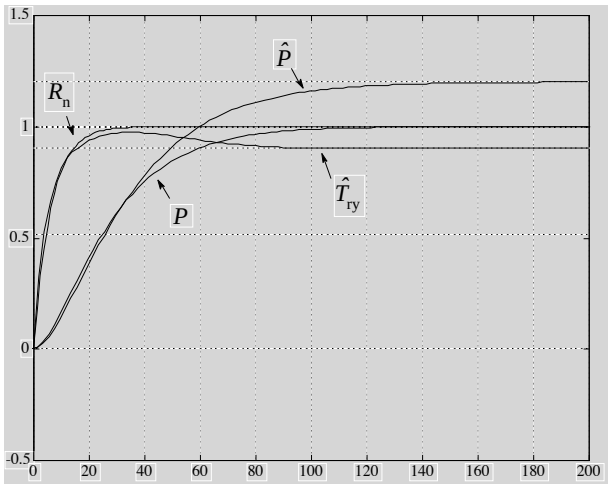


Figure 14. Step responses using the observer based *PID* regulator

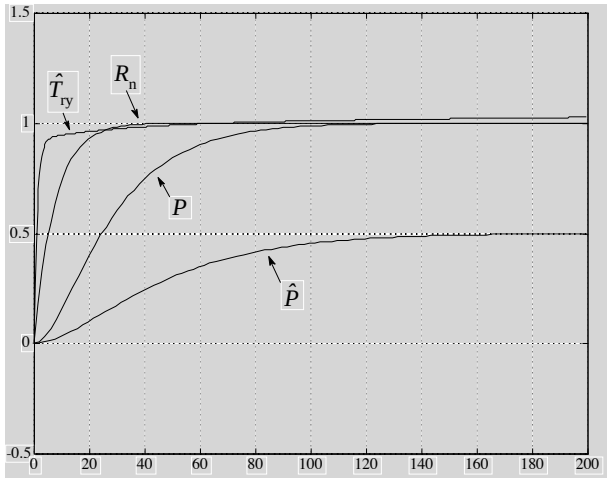


Figure 15. Step responses using the observer based *PID* regulator

Example 2

The process parameters and the selected first order R_n are the same as in the previous example. The model parameters are: $\hat{T}_1 = 30$, $\hat{T}_2 = 20$ and $\hat{A} = 0.5$. In the observer loop a *PI* regulator (54) is applied with $A_l = 0.001$ and $T_l = 2$. The ideal form of Q (50) was used. Figure 15 shows some step responses in the operation of the observer based *PID* regulator.

It is easy to see that the $\hat{T}'_{\mathrm{ry}}$ well approximates R_n in the high frequencies (for small time values) inspite of the very bad model $\hat{P}$.

8 Conclusions

It was shown that the *SFO* methodology results in such a control system, which leaves the open-loop zeros untouched for the tracking properties and unfortunately the disturbance rejection (regulatory) properties can only be partly designed, because they are not independent of the tracking design.

The *TFR* of these classical methods are introduced to get a simple and useful tool to analyze and explain further behaviors, which are difficult to obtain using *SVR*. Using *TFR* it was shown, if the *SVR* used in the *SFO* scheme is model-based then the original (without observer) model error decreases by the sensitivity function of the observer feedback loop. This model error reducing capability gives the theoretical background of the success of practical model-based *SFO* applications.

Finally the *SFO* method was applied for the classical *IMC* structure, opening a new class of methods for open-loop stable processes. This new method combines the classical YOULA-parameterization based regulators with the *SFO* scheme. Using this new approach an observer based *PID* regulator was also introduced. This regulator works well even in case of large model errors as some simulations showed.

9 References

[1] Åström K.J. and B. Wittenmark (1984). *Computer Controlled Systems* (Prentice-Hall, 1984).

[2] Åström K.J. (2002). *Control System Design Lecture Notes*, U of California, Santa Barbara.

[3] Kailath T. (1980). *Linear Systems* , Prentice Hall.

[4] Keviczky L. (1995). Combined identification and control: another way. (Invited plenary paper.) *5th IFAC Symp. on Adaptive Control and Signal Processing, ACASP'95*, Budapest, H, pp. 13-30.

[5] Keviczky L. and Cs. Bányász (2001). Iterative identification and control design using K-B parameterization, In: Control of Complex Systems, Eds: K.J. Åström, P. Albertos, M. Blanke, A. Isidori, W. Schaufelberger and R. Sanz, Springer, pp. 101-121.

[6] Maciejowski J.M. (1989). *Multivariable Feedback Design*, Addison Wesley.

This work was supported in part by the MTA-BME Control Engineering Research Group of the HAS, at the Budapest University of Technology and Economics and by the project TAMOP 4.2.2.A-11/1/KONV-2012-2012, at the Széchenyi University of Győr.

Using Video Modeling in Virtual Learning Environment to Develop some of Mathematical Skills among Children with Down Syndrome in the State of Kuwait

Muntaha S. TH. F. ALHendal
Distance Teaching and Training Program
Arabian Gulf University
Manama, Kingdom of Bahrain
m.alhindal@hotmail.com

Hamdy A. Abdulaziz
Distance Teaching and Training Program
Arabian Gulf University
Manama, Kingdom of Bahrain
hamdyaaa@agu.edu.bh

Ahmed M. Nouby
Distance Teaching and Training Program
Arabian Gulf University
Manama, Kingdom of Bahrain
ahmedmns@agu.edu.bh

Elsayed S. Elkhamisi
Mental Disabilities and Autism program
Arabian Gulf University
Manama, Kingdom of Bahrain
saidsmk@agu.edu.bh

***Abstract*—The objective of this research is to investigate the effect of Using Video Modeling in a Virtual Learning Environment to Develop some of Mathematical Skills among children with Down syndrome in some primary stage schools in the State of Kuwait. The measuring tool for the research was an achievement test and interview with the parents to assess their satisfaction. The sample was of 16 children. They were aged between 7 -13 years, in the scholastic year 2012-2013. The research designed semi-experimental, pre-test/post-test control group.**

Keywords—: video modeling - Mathematics - Virtual Learning Environment - Down Syndrome - Kuwait .

i. Introduction

E-Learning is the use of technology to enable people to learn anytime and anywhere. E-Learning can include training, the delivery of just-in-time information and guidance from experts. It enables people to learn in many different ways and at different times. To support these different learning needs, you will need different e-learning delivery methods. Additionally, you will need a way to develop and manage e-learning (Lucas,L. , 2013).

A virtual Learning Environment (VLE), is an e-learning education system based on the web that models conventional in-person education by providing equivalent virtual access to classes, class content, tests, homework, grades, assessments, and other external resources such as academic or museum website links. It is also a social space where students and teacher can interact through threaded discussions or chat (Wikipedia,2013).

Many theories were applied in e-learning and one of them is the Social Learning Theory. Social Learning Theory proposed that learning can occur through simple passive observation of behavior (Bandura, 1971). Recent evidence suggested that passive observational learning may be more effective than interactive modeling as an instructional technique (Biederman, Davey, Ryder &Franchi, 1994; Biederman, Ryder, Davey & Gibson, 1991). His theory added a social element, arguing that people can learn new information and behaviors by watching other people. Known as observational learning(or modeling), this type of learning can be used to explain a wide variety of behaviors and knowledge (Cherry).

The Modeling Process

Not all observed behaviors are effectively learned. Factors involving both the model and the learner can play a role in whether social learning is successful. Certain requirements and steps must also be followed. The following steps are involved in the observational learning and modeling process:

Attention:

In order to learn, you need to be paying attention. Anything that detracts your attention is going to have a negative effect on observational learning. If the model interesting or there is a novel aspect to the situation, you are far more likely to dedicate your full attention to learning.

- **Retention:**
  The ability to store information is also an important part of the learning process. Retention can be affected by a number of factors, but the ability to pull up information later and act on it is vital to observational learning.
- **Reproduction:**
  Once you have paid attention to the model and retained the information, it is time to actually perform the behavior you observed. Further practice of the learned behavior leads to improvement and skill advancement.
- **Motivation:** Finally, in order for observational learning to be successful, you have to be motivated to imitate the behavior that has been modeled. Reinforcement and punishment play an important role in motivation. While experiencing these motivators can be highly effective, so can observing other experience some type of reinforcement or punishment. For example, if you see another student rewarded with extra credit for being to class on time, you might start to show up a few minutes early each day (Cherry).

A. Types of modeling for children with special needs

Nikopoulos & Keenan (2006) mentioned the types of the educational modeling for children with special needs.

- Exact and behavior-feature imitation
- Generalized imitation
- Peer modeling
- Self-modeling
- Video modeling.

In this research we focused on video modeling. Dowrick (1991) defined the video modeling as a type of modeling, where the modeling person is distant from the learner but is seen on video to change or teach a target skill. Video modeling is defined as a form of observational learning in which desired behaviors are learned by watching a video demonstration and then imitating the behavior of the model. (Ellis & Marietta, 2011).

A. Possible skills to improve via video modeling

Has your child struggled with a task or behavior that might be easily demonstrated on a video? Here are some modeling ideas to get you thinking:
• Self-help skills, including brushing teeth, getting dressed, tying shoes or making a sandwich;
• Fine motor skills such as tracing and coloring;
• Gross motor skills, like jumping or catching a ball;
• Academic skills, such as object recognition, yes/no questions and even long division; Social skills in clips such as a dentist visit, meeting and greeting people or playing board games. (Ellis & Marietta, 2011).One type of the special needs who gets benefit from video modeling is Down's syndrome.

B. Down Syndrome

The National Association for Down Syndrome (NADS) described Down syndrome as "… a well known genetic disorder … caused by a chromosome abnormality that occurs before birth. Typically babies have 46 chromosomes, but those with Down Syndrome are born with 47 chromosomes, which causes abnormal changes in the development of the child's body and brain" (NADS, 2010: in Becker).
• Children with Down Syndrome are visual learners and should be involved in topic work and curriculum on the same subject- at their own level-with individual targets within the subject.

• Break down reading, numeracy and social skills to the pre-reading and pre-number level

• Present tasks that will help develop the problems with short term memory and poor co-ordination.

• Develop reading in order to develop speech.

• Develop a whole word visual sight vocabulary before breaking words down phonically.

• Use flash cards even before speech has developed.

• Use signing as a conceptual bridge to motivate communication and encourage productive speech (Black 1998).

The specialized needs and skills of children with Down Syndrome demand software specifically designed to meet those needs and skills. The benefits that instructional technologies offer students with Down Syndrome convincingly reinforce that demand. Present tasks that will help develop the problems with short term memory and poor co-ordination. (backer,2010) Ellis & Marietta (2011) found out that they could teach a child with Down syndrome how to write letters by video modeling. They recorded a video of how the teacher can write the /C/ with song for the letter C. The boy Rayan who is with down syndrome liked the song then tried to write the letter after observing his teacher in the video.

As a matter of fact the research team knew there was no usage of virtual learning environment or e-learning in teaching or in training via video modeling, though e-learning had significant benefits for children with Down syndrome.

C. Benefits of computer-assisted learning for people with Down Syndrome.

The instructional technologies and e-learning can be more appropriate and effective than other approaches for students

with Down syndrome due to their numerous benefits that were suggested by a variety of authors and compiled by Black as listed below:

• Improving motivation: The learning experience was enhanced with pictures, sounds and animation which may increase a child's interest and attention.

• Multi-sensory experience: Computers provided both visual and auditory input. Children with Down syndrome are 'visual learners' who learn best through visual means and find audio means more difficult. ICT is particularly well suited to this learning style.

• Non-verbal mode of response: Children were able to give non-verbal responses, which enabled them to demonstrate their understanding without having to produce a spoken response, which may be particularly difficult for them due to their troubles with articulation, word finding and intelligibility.

• Being in control: Children began to understand that they can have an effect on their surroundings through 'cause and effect' software; this sense of being in control developed further as children started to use familiar programs unassisted; self-esteem developed as they became more independent in their learning thus their presentation skills were improved.

• Immediate feedback: Children are rewarded for their successes immediately, e.g. with pictures, sound effects or music, or prompted if they need to try again. The computer never gets impatient or frustrated by the repeated errors, feedback is non-threatening and non-judgmental.

• Errorless learning: Software can be designed in such a way that the child is supported in order to achieve repeated success. The child is supported at each step as necessary, before they commit a mistake. This allowed the child to learn a sequence of steps to achieve success every time

• Opportunities for practice: Children with Down syndrome need much more practice to acquire new skills and ICT can provide as many opportunities as necessary to repeat the same objective in exactly the same way

• Self-paced learning: The child was able to proceed as fast or as slow as he or she wishes; the computer will 'wait' for the child to respond without prompting them before they have had time to fully process the information and construct their response

• Clutter free working environment: Computer programs provided a highly organized and predictable working environment which focused the child on specific learning targets. (Black, 1998).

D. children with Down syndrome and Mathematics.

Hughes (2006) mentioned that numeracy system is a system of logic. Considering the characteristics of children with Down syndrome, there must be an indication of how to help them learn Mathematics effectively where:

- They learn better when mixed with normal healthy children.
- Solving Mathematical problems came in a later stage after getting the skills of reading and writing. The cause is not known so far.
- Children with Down syndrome can be taught as ordinary healthy children only with a slower process.
- They have a short memory below their age which made them need longer time to solve problems and participate in the class.
- Teaching Mathematics depends on visual support.

Teaching process should be organized, divided into small steps. Students should be given enough time to practice more in order to develop and improve new skills.

ii. The Research

A. Research Problem.

E-learning has been demonstrated to be effective with students with special needs. This method is specifically appropriate for use with the children with Down Syndrome. Therefore, depending on the difficulties of learning mathematics for Down's syndrome due to their cognative characteristics, we considered the significant advantages of video modelig that can address pervasive difficulties in educating and training students with special needs. Accordingly, the research addressed : Using Video Modeling in Virtual Learning Environment to Develop some Mathematical Skills among Children with Down Syndrome in The State of Kuwait.

B. The design of the research

Using semi-experimental, pre-test/post-test control group.

C. Participants

1. Control group included seven children (4 males, 3 females; 9-10 years of age, mean 10) from schools in The State of Kuwait.

2. Experimental group included nine children (2 males, 7 females; 7-13 years of age, mean 9,78) from schools in the states of Kuwait.

3. The sample of people with Down syndrome, with Mild Intellectual Disability,IQ55-70, and any susceptible to education. The research team had the approval of the Ministry of Education and parents to implement the strategy of video modeling in virtual learning environment in order to develope the Mathematical skills among these children.

D. Research Hypothesis

1. There were no statistically significant differences in the mean average between the experimental group and the control group in the post achievement test due to the use of video modeling.
2. There were no statistically significant differences in the mean average of the learners in the experimental group between the pre and the post achievement test scores.
3. There were no statistically significant differences in the mean average of the learners in the experimental group between the post and the achievement test scores.

iii. Instructional design (ADDIE)

The generic term for the five-phase instructional design model consists of Analysis, Design, Development, Implementation, and Evaluation. Each step had an outcome that feeds into the next step in the sequence. There were probably over 100+ different variations of the generic ADDIE model (learning-theories.com, 2013).

A. Analysis

The research team analyzed the following:

1. Needs Analysis: It was carried out by interviewing a group of Down's syndrome mathematics teachers and reaching to the difficulties in combining skills.
2. The personal characteristics of the learners: It was carried out by considering literature, observing children with Down syndrome, and survey teachers' opinion; it was noticed that they like imitation, music, and games.
3. Content Analysis: This process was formulated after reviewing the mathematics curriculum for second-grade Down's syndrome children in The State of Kuwait and selecting "addition number seven" lesson; moreover.
4. Analyzing the educational environment: It was carried out by reviewing schools, and checking the availability of the internet access. Finally, the research team provided the experimental group with computer and internet in order to apply the video modeling strategy in virtual learning.

B. Design

The research team designed the program by Formulating educational objectives: Setting objectives after viewing the curriculum and analyzing the educational content. The objectives were defined according to Bloom's Taxonomy of Educational Objectives, and were included as knowledge, comprehension, and application.

1. Describing the script of the educational content: Preparing the script for video modeling in virtual learning was designed after writing the video recorded Clips, the most important thing to be considered is that the clips must be identical to the child's reality, such as using household tools or shopping at vegetables and fruit store and the similar child experiences.

- Determining the educational strategy: Identifying the educational strategy as a four basic stages applied at the video modeling; observation, application, re-production, and motivation.
- The content was divided into five lessons.
- The strategy was commenced by an acting model who explained the first lesson's targeted skill "how to count seven items"; count the red flowers and replaces them at the appropriate place.
- The educational video was controllable as it can be stopped or restarted.
- The process of modeling started through observing the clip as the student observed during applying the targeted skill.
- The student pressed the orange icon (next) down the video that contrast the page color.
- The following screen displayed a text of the activities and simple instructions about the lesson accompanied by a recorded voice that read the text in accordance with the cognitive characteristics of the learners.
- The student continued the modeling activity. As for the feedback, it started during the performance of the video modeling.
- After modeling, the feedback appeared as music to reward the student for the completion of the task, and then appeared the text and the accompanied voice that confirmed the performance of the targeted skill.
- After completing the targeted skills, the evaluation process started as the student applied hard applications and simple e-test through the virtual learning environment.

Specifying the standards for the design of the educational site as befits the characteristics of Down's syndrome: Identifying the standards for the educational design of the site according the characteristics of Down's syndrome.

When designing the educational website, the research team considered identifying the most necessary standards that are qualified for the characteristics of the Down's syndrome and their needs.

1- The educational objectives of the curriculum fit their mental and academic abilities.
2- The website content was modeled slowly and clearly and is presented in attractive way identical to the educational objectives.

3- To increase interaction and integration; colors were comfortable and icons were big, clear and the background was contrasted. Depending on the needs of Down's syndrome children, the site contained few texts read by accompanied voice to help them understand their tasks or get the feedback. Additionally, the movement items seemed to be easy by using the mouse.
4- Page design was focusing on the constancy of the size and the place of the videos and the icons in all screens.
5- The website contained the four stages of observing and modeling.
6- Integrally, there was a great use of multimedia (text, image, sound, video, color)
7- Video modeling (electronic activities) is a video-based intervention involving feedback during and after the modeling. Time designation was not required at that process because it is based on the child's abilities and the other characteristics.
8- The variety of the feedbacks on the website were considered.
2. Preparing the script for the educational web to create video modeling in the virtual environment learning.

C. Development

The third level in the instructional design, the research team developed the website by:

1. Compilation of media (video production and video modeling in virtual learning environment).

a-They gathered the appropriate images and tools applied at the modeling and videotaped the model while performing the targeted mathematical skills by using Canon EOS 550D, up to 30 fps shooting ,so that the videos were created with more flexibility besides the multi frames. Moreover, the external microphones socket was used in order to add an additional microphone for better quality sound.

b- The model recorded the sound for all the Clips of home page, instructions, final page and feedbacks. These sounds were connected by using Free Sound Recorder, free available software on the internet, due to its ability to record sound from a microphone, line-in, and just about any media program. Additionally, sound recorder, is offering professional recording features with full mp3, wma, and wav support.

c- Prior to performing the modeling, the e- activities were created identical to the educational video by using Adobe Flash program due to its high quality in producing stunning media experiences and delivering console-quality games and content to the browser.

2. Compilation of video screens inside the program and arranged.

To design the storyboard and develop Clips, the researchers used Photoshop Program to create the banner of the educational website and the other pages. Depending on the needs of Down's syndrome children, icons were modified to be big, clear, and contrastive the background. Designing the home page that contained a picture of a train moving on one direction accompanied by its sound that was displayed for once time. The lesson's name appeared on the train's carriages. To exit the lesson and back to the home page, the student pressed the small familiar train image set as the exit icon. Accordingly, selecting the lessons was interesting to the Down's syndrome children who like and interact with colors and sounds.

3. Determine the interactions within the environment.

To increase the impact of the visually-based interventions, the researchers considered stimuli such as connecting the voice with texts, activities, and feedbacks. In addition, the constancy of their size and place. Consequently, the researchers used the easy available Adobe Dreamweaver program in order to provide intuitive visual interface to address the needs of the Down's syndrome children.

Fig. 1. Example of the education website in the Virtual Learning Environment

D. Implementation

1. Prior to learning the target mathematical skills, the researchers applied the pre-test on the experimental and the control group.
2. Researchers corrected the tests and monitored the grades.
3. The control group studied some mathematical skills traditionally.

4. To acquire the targeted mathematical skills, the experimental group was studied via video modeling strategy in virtual learning environment.
5. After the implementation of the research, the children in both experimental and control group were set for post-test to evaluate the acquisition of some mathematical skills.
6. The test has been corrected and the grades had been monitored for each sample.
7. After applying the electronic processing, a followed test was applied to evaluate the learning acquisition of the experimental group.
7. Non parametric statistic test Mann-Whitne, and Wilcoxon was used for data analysis.

E. Evaluation

there were three levels of the evaluation:
1. Formative assessment.
2. Summative assessment.
3. Satisfaction of the parents.

iv. Research results

The findings from this research included the following:

1. There were statistically significant difference at the significant level 0.05 between the experimental and control group , favor the experimental group.
2. There were statistically significant differences at the significant level 0.05 between the pre and post test in the experimental group.
3. There were no significant difference at the significant level 0.05 between post and followed test in the experimental group.

The researcher obtained feedback from parents though phone calls. Collected feedback was classified into descriptive and coded.

The researcher founded the following:
1- Parents of the participating students were satisfied with the application of the modeling video strategy in the virtual learning environment and thought it made the students interested in learning.
2- Parents expressed finding the strategy applied serving all subjects and can help educate and train pupils with Down Syndrome.
3- Parents cleared that the students' relatives and siblings were impressed with the educational website and the application of the virtual learning environment.
4- Parents showed that interaction with the teacher and communication among each other has encouraged them to continue logging into the virtual learning environment.
5- Parents assured that using video modeling in a virtual learning environment had a significant positive impact on learning mathematical skills.

v. Final conclusions and recommendations

1) School administration and supervision departments are recommended to give more care for this technique and its requirements to be applied at the schools of special needs on pupils with qualifying abilities to deal with it.
2) Holding training courses for teachers of Mathematics, especially for Down's syndrome teachers to enable them of using this technique appropriately.
3) Utilizing visually based interventions with students with special needs, especially Down syndrome, is highly recommended as it has been demonstrated to be effective with their needs.

Recommendation
Accordingly, the researchers recommend the following:

1- Due to the effectiveness of the virtual learning for students with special needs, their teachers should be trained to use this method properly.
2- To address the needs of Down's syndrome children as a visually- based learners, mathematics teachers should be trained to apply video modeling in virtual learning environment.
3- To increase interaction and communication, Down's syndrome parents should be trained to use virtual learning environment.

ACKNOWLEDGMENT

I owe a debt of gratitude to my dissertation committee. To the main supervisor, Dr. Hamdy Abdelaziz , thank you for your untiring support and guidance to present an outstanding scientific research . I am especially grateful to Dr. Ahmed Nouby, thank you for reinforcing and supporting the development of the research. I am also grateful to Dr. Elsayed Elkhamisi, thank you for your constant promoting and encouraging for the researches on the people with special needs. Many thanks to Conference Secretariat, CSCI'14 for accepting the publication of this research during this considerable conference which is a great honor for the research team. Finally, special thanks to my sister Maha Al Hendal for her cooperation with the research team.

REFERENCES

Bandura, A. (1970). Modeling theory: Some traditions, trends, and disputes. In W. S. Sahakian (Ed.), *Psychology of learning: Systems, models, and theories*. Chicago: Markham.

Becker, T. (2010). E-Learning for children with Down Syndrome; The need for additional instructional technology titles. Duquesne University Instructional Technology EDD Program. http://tombecker245.blogspot.com/

Biederman, G.B., Ryder, C., Davey, V.A., & Gibson, A. (1991). Remediation strategies for developmentally delayed children: Passive vs. active modeling interventions. *Canadian Journal of Behavioural Science*, 23, 611-618.

Black, B. (1998). Inclusive *Education Culture of Down Syndrome. Downloaded* 5\10\2013 from:

http://www.inclusive.co.uk/infosite/bblack.shtml

Cherry, K. Social Learning Theory " an overviw of Bandura's social learning theory. from:
http://psychology.about.com/od/developmentalpsychology/a/sociallearning.htm

Dowrick, P. W. (Ed.) (1991). *Practical guide to using video in the behavioral sciences***.** New York: John Wiley.

Ellis, S., Marietta, G. (2011). Video Modeling: Teaching Through Television. *Down Syndrome News 34*,(3) 35-36. Retrieved Sep 20,2013 http://ndsccenter.org/worpsite/wp-content/uploads/2012/03/NDSC-Down-Syndrome-News-June-2011.pdf

Hughes, J. (2006). Learning about number and maths. *Down Syndrome News and Update*, 6(1);10-13. Retrieved March 15,2012 http://www.down-syndrome.org/practice/374/

Learning-theories.com. (2013). Retrieved NOV. 15,2013(http://www.learning-theories.com/addie-model.html)

Lucas, L**.**(2013). *Why do you need e-Learning Consulting Services?* Retrieved NOV. 15,2013 (http://www.e-learningconsulting.com/consulting/what/e-learning.html).

National Association for Down Syndrome http://www.nads.org/

Nikopoulos, C. , & Keenan, M. (2006). *Video Modeling and behavior analysis: A guide for teaching social skills to children with autism*. Philadelphia, PA: Jessica Kingsley Publishers.

Robertson, A. and Biederman, B. (1989). Modeling, imitation and observational learning in remediation experimentation 1979-1988: An analysis of the validity of research designs and outcomes. *Canadian Journal of Behavioural Science*, 21, 174-197.

The University of Kansas. Retrieved March 15,2013 http://www2.ku.edu/~kucrl/cgi-bin/drupal/?q=node/225

Model Design for Scheduling Multiple Factory Cranes in Manufacturing Floor

Hee-jin Lee[1], Kwang-yeol Ryu[1], Sang-il Lee[1], Ho-yeun Ryu[2], and Hui-chul Lee[2]
[1]Department of Industrial Engineering, Pusan National University, Busan, South Korea
[2]Korea Institute of Industrial Technology, Jinju, South Korea

Abstract - *Unpredictable and ever changing customer demands make manufacturers to make quality products with cheaper price and shorter delivery. To survive in the global market by meeting customer requirements dynamically changed, the manufacturing industry needs to respond quickly to the customer needs and various demands, an efficient schedule takes more important role in manufacturing system than ever. The problems for scheduling multiple factory cranes are very complex due to several factors, such as crane movement types, time factor, and available weight. In this paper, we propose model design and scheduling methods by using heuristic algorithm for multiple factory cranes in manufacturing floor. In other words, this paper proposes a model to minimize the time of factory crane movements required in manufacturing orders, divides crane operations into empty crane and connect movement in order to approach realistic factory crane scheduling.*

Keywords: Multiple Factory Cranes, Crane Movement Types, Moving Time, Heuristic Algorithm

1 Introduction

Recently in a globalized competitive manufacturing environment, product life-cycle becomes shorter and the type of manufacturing system has been changed in order to deal with small quantity of products in batch. As a result, many manufacturers are trying for various methods of cost reduction, faster delivery and quality improvement to enhance competitiveness in the international market. Thus, in order to respond quickly to the customer needs and various demands, an efficient schedule plays a more important role in manufacturing system than ever.

Many manufacturers use various machines to produce specific parts and products in manufacturing floor. The material transport systems for handling the parts and products include the movement, storage, protection and control of materials throughout the manufacturing process. In that process, the material handling equipment can be classified into four categories, such as material transport equipment, storage systems, equipment, utilization, identification and tracking systems (Groover, 2008). The material movement is processed by much equipment: cranes and hoist, industrial trucks, conveyors, automated guided vehicles (AGV), etc. Usually, the cranes and hoist in manufacturing floor are operated by experienced workers who carry and process the materials. As the results of cranes operation depends highly on the workers' experience, this operation experience are not easy to plan.

Manufacturing facilities frequently rely on track-mounted cranes to move in-process materials or equipment from one location to another. A typical configuration allows on or more hoists to move along a single horizontal track that is normally mounted on the ceiling as depicted in Figure 1. Each hoist may be mounted on a crossbar that permits lateral movement as the crossbar itself moves longitudinally along the track. A cable suspended from the crossbar raises and lowers a lifting hook or other device (Aron *et al*., 2008).

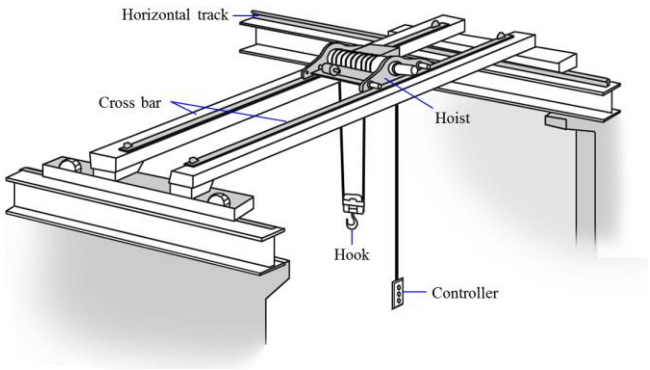


Figure 1. A typical configuration of factory crane

Planning the movement of a crane to follow a manufacturing schedule is a real-world combinational problem. For a single-crane factory, the problem consists of sequencing the tasks on the crane in an order that minimizes their completion times subject to some constraints. For multiple-crane factory, there are additional problems of assigning cranes to tasks and selecting movement priority when the cranes interfere (Peterson, 2010). The factory crane movement is defined as material handing operations between machine (or process) and machine (or process).

This paper proposes a model to minimize the time of factory crane movements required in manufacturing orders,

divides crane operations into empty crane and connect movement in order to approach realistic factory crane scheduling.

2 Literature review

The crane scheduling problem is mainly classified as container terminal (port) and factory (manufacturing floor), as shown in Figure 2. The researches for crane scheduling in port container terminal are to load and unload containers using quay cranes, yard cranes and rail cranes. The factory crane scheduling problem has been researched mostly discussing of hoist scheduling for an electroplating line, and the scheduling problem focused on only factory cranes is hard to find. In this paper, the factory crane denotes overhead crane because most of the manufacturing floors for machining have to use the overhead crane for safety between jobs.

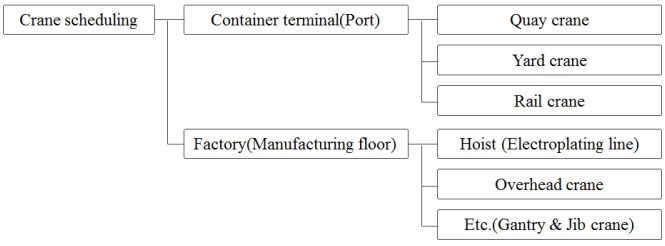


Figure 2. Classification of crane scheduling problems

The problem of scheduling quay cranes in port container terminals is to determine the schedule of each crane assigned to a vessel for loading and unloading containers. Previous researches have suggested methods for determining the number of quay cranes, the operation sequence and the minimized makespan (Daganzo, 1989; Peterkofsky and Daganzo, 1990; Kim *et al*., 2004; Kim and Park, 2004). The case of yard cranes runs on tracks in a single storage block between the tracks and can move among storage blocks. This researches have proposed methods in which the times and routs of yard crane movements among blocks are determined (Kim and Kim, 1999; Narasimhan and Palekar, 2002; Zhang *et al*., 2002; Kim *et al*., 2003; Ng and Mak, 2005; Liang *et al*., 2011). Researches related to rail cranes have been proposed for minimizing transshipment and movement times in rail container terminals. (Bostel and Dejax, 1998; Kozan, 2000; Jeong and Kim, 2011).

Manier and Bloch (2003) dealt a classification for hoist scheduling problems in electroplating lines. This problems were classified into four trees: cyclic, predictive, dynamic and reactive hoist scheduling problems, depending on the specific features. Most of this research deals with single-hoist problems (Phillips and Unger, 1976; Lei and Wang, 1994; Chen *et al*., 1998). However, even this single-hoist problem is considered as NP-complete problem (Lei and Wang, 1989). Other research for electroplating lines has dealt with cyclic multi-hoist problems. The method partitions the tanks into groups so that assigns a hoist to each task, and schedules each hoist within its partition (Lei and Wang, 1991; Zhou and Li, 2009). Some research has been studied the various methods of the problem in which the hoists are allowed to avoid collisions. (Manier *et al*., 2000; Che and Chu, 2002; Leung *et al*., 2004; Che and Chu, 2004; Leung and Levner, 2006). This paper allows also to move a crane to an arbitrary location in order to avoid interference to another crane).

3 Crane movement types

The production system comprises manufacturing support systems and facilities (Groover, 2008); facilities typically use

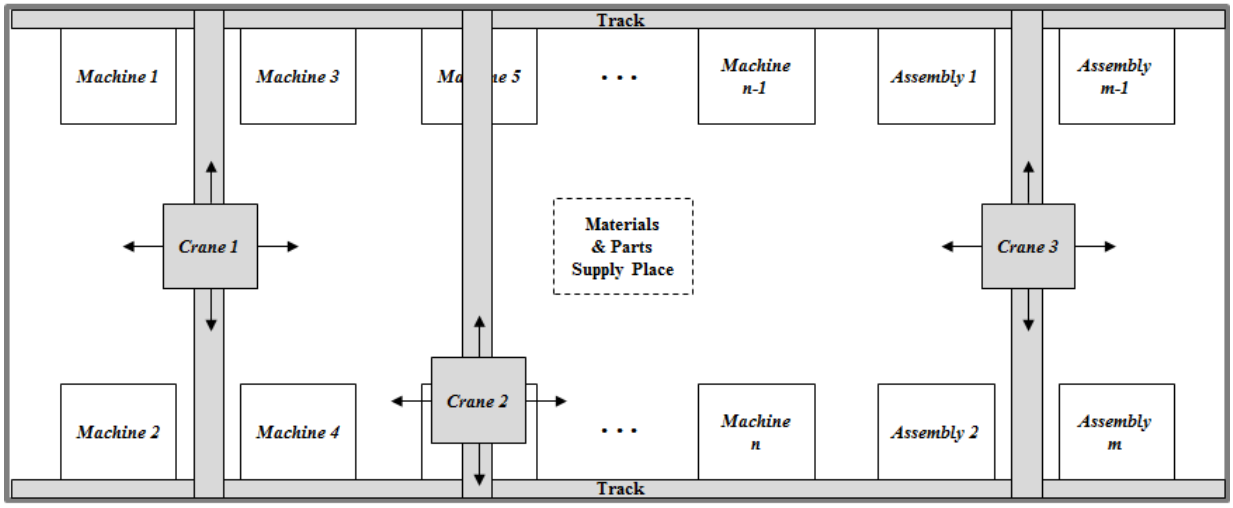


Figure 3. Layout of a manufacturing floor for multiple factory cranes

manufacturing systems, automation and control technologies, processes and assembly operations, and material handling technologies. This material handling is defined as movement of materials, parts and products between the processes (or machines). We consider factory cranes of material handling to be that between processes.

This paper proposes a method to minimize the moving times of multiple factory cranes required in manufacturing orders. A manufacturing floor has machines, assembly processes and cranes for producing specific parts and products, as shown in Figure 3. Generally a factory crane loads the materials on the pick-up point, and operates vertically the hoist on crossbar if the crane meets an obstacle when moving horizontally, and then the factory crane moves to the delivery point to unload the materials. (Figure 4). Black-color rectangle denotes material loading on pick-up point, and just another rectangle denotes material unloading on delivery point.

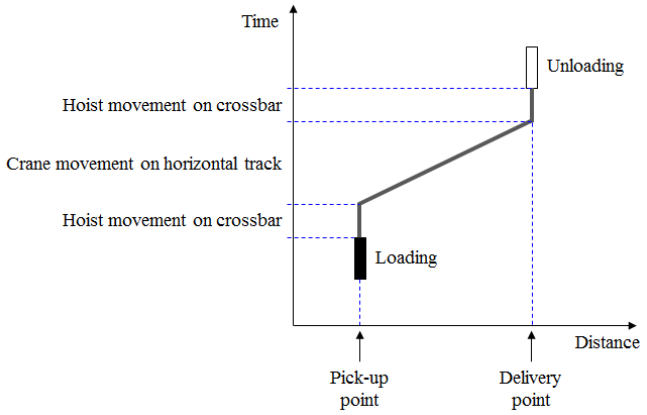


Figure 4. Basic route of a factory crane movement

This paper divides factory crane movement into empty and connect movement in order to approach realistic crane movement scheduling, based on the manufacturing order types. Empty movement means only to move an empty crane for avoiding interference between cranes regardless of material handling. Crane 1, in Figure 5, loads some material on a pick-up point and then moves to a delivery point; crane 2 moves to the arriving distance of crane 1. The empty crane again moves to other point to avoid interference or moves to other order who require it. Circles mean the empty cranes in Figure 5.

After unloading the material from the crane, the crane can be connected with another order for loading through connect movement. The crane loading of material becomes an empty crane after unloading the material and can then be moved to another order point. In other words, if the other order has a loading task, the empty crane can be used for material handling. However, the other order must be located within an area where movement is available. If the material is unloaded to the delivery point, the crane becomes an empty crane and is then available for loading a material. In addition, the empty crane can be connected to other task orders on the pick-up point.

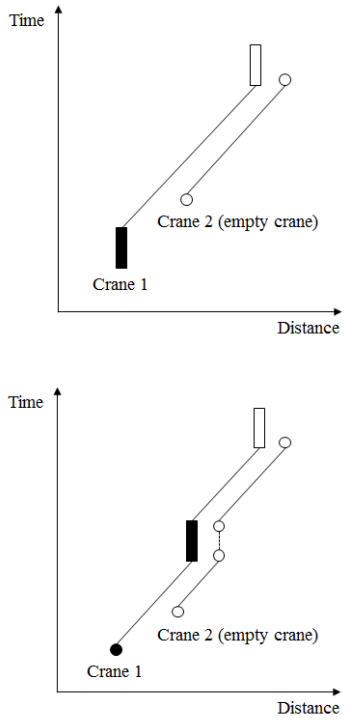


Figure 5. Examples of an empty crane movement to avoid interference

4 Heuristic algorithm model

It is very difficult to determine the optimal solution for crane scheduling since it is an NP-complete problem. Therefore, this problem has a property wherein the time required to find the optimal solution rapidly increases with the number of processes. The mathematical model cannot be used to solve large problems. Therefore, we propose a heuristic algorithm model that reduces the calculation time required for solving the problem. The procedure of the heuristic algorithm model for this problem is as follows:

Step 0.

1. Input the constraints allowances – crane moving time *MT*, passage time *MP*, waiting time *MH*, and weight *MW*.

2. Set crane $k = 0$ and moving sequence $e = 0$.

3. Sort all orders in ascending order based on the request time.

Step 1.

1. The prior order selected by request time sequence is $k = k + 1$, $e = e + 1$,

   $s_0 = q_j - t_{0j}$, $a_j = q_j$ and $s_j = q_j + v_j$.

   (si : starting time of order i, a_j : arrival time of order j, t_{ij} : moving time, v_j : loading or unloading time)

2. Search for an order from the same process that can connect to the previous selected order. The connected process order must satisfy the constraints with the previous selected process.

   The constraints are $t_{ij} = 0$, $w_k \le MW$. (w_k : loaded weight in crane k)

3. If the searched order can be connected with the previous selected order, go to Step 2. However, if the connected orders have the filled crane, $a_0 = s_j + t_{j0}$, go to Step 4.

4. Otherwise, go to Step 3.

Step 2.

1. The connected order for the same process is $e = e + 1$, $a_j = s_i + t_{ij}$ and $s_j = q_j + v_j$.

2. Search for an order from the same process that can connect to the previous process order. The connected process order must satisfy the constraints with the previous process order.

   The constraints are $t_{ij} = 0$, $w_k \le MW$.

3. If the searched process order can be connected with the previous process order, go to Step 2. However, if the connected process orders have the filled crane, $a_0 = s_j + t_{j0}$, go to Step 4.

4. Otherwise, go to Step 3.

Step 3.

1. Search for a process order that can connect to the previous process order and satisfy the constraints of the previous process order.

   The constraints are $t_{ij} \le MT$, $p_{ij} \le MP$, $q_j - a_j \le MH$, $w_k \le MW$. (p_{ij} : passage time between process orders)

2. If the searched process order can be connected with the previous process order, the searched process order is $e = e + 1$, $a_j = s_i + t_{ij}$ and $s_j = q_j + v_j$, go to Step 3.

3. If the connected process orders have the filled crane, or there are no more process orders that can connect to the previous process order,

   $a_0 = s_j + t_{j0}$, go to Step 4.

Step 4.

1. If all process orders are completed using the crane scheduling method, terminate the heuristic algorithm.

2. Otherwise, go to Step 1.

The constraints of this procedure are the crane moving time, passage time, waiting time, and weight. The crane moving time between the processes must be less than the allowed time. Passage time implies the crane moving time from a process to another process. The rate of crane moving time from a process to the final process is equal to 100%, and the passage time with another process is possible until the allowed rate, which is summed to 100%. The waiting time, which is the gap between the request time and arrival time, must be less than the allowed time. In addition, the loaded weight in each crane must be less than the allowed weight.

The procedure of the heuristic algorithm shows the iteration structure that can be connected by the crane states. In this procedure, if a crane state is the filled crane, the crane must finish the search for the connect movement. If all process orders are completed by using crane scheduling, the heuristic algorithm is terminated.

5 Conclusions

The problems for scheduling multiple factory cranes are very complex due to several factors, such as crane movement types, time factor, and available weight. We introduced a classification of crane scheduling problems, crane movement types and a heuristic algorithm to perform crane scheduling in manufacturing floor. We did attempt to find a research direction for crane scheduling, and computational experiment and decision support system need additionally. Future research might be to find an optimal solution using mathematical and meta-heuristic models with extended approach in this study.

6 References

[1] Groover, M., "Automation, Production Systems, and Computer-Integrated Manufacturing," 3rd Edition, Paperback, 2008.

[2] Aron, I., Genç-Kaya, L., Harjunkoski, I., Hoda, S. and Hooker, J. N., "Optimal Movement of Factory Cranes," Tepper School of Business, Paper 143, 2008.

[3] Peterson, B. K., "Transportation Scheduling Methods," Tepper School of Business, 2010.

[4] Daganzo, C. F., "The Crane Scheduling Problem," Transportation Research Part B: Methodological 23(3), 159-175, 1989.

[5] Peterkofsky, R. I. and Daganzo, C. F., "A Branch and Bound Solution Method for the Crane Scheduling Problem," Transportation research Part B: Methodological 24(3), 159-172, 1990.

[6] Kim, K. H., Kang, J. S. and Ryu, K. R., "A Beam Search Algorithm for the Load Sequencing of Outbound Containers in Port Container Terminals," OR Spectrum 26, 93-116, 2004.

[7] Kim, K. H. and Park, Y. M., "A Crane Scheduling Method for Port Container Terminals," European Journal of Operational Research 156, 752-768, 2004.

[8] Kim, K. H. and Kim, K. Y., "An Optimal Routing Algorithm for a Transfer Crane in Port Container Terminals," Transportation Science 33(1), 17-33, 1999.

[9] Narasimhan, A. and Palekar, U. S., "Analysis and Algorithms for the Transtainer Routing Problem in Container Port Operations," Transportation Science 36(1), 63-78, 2002.

[10] Zhang, C., Wan, Y., Liu, J. and Linn, R. J., "Dynamic Crane Deployment in Container Storage Yards," Transportation Research Part B 36, 537-555, 2002.

[11] Kim, K. H., Lee, K. M. and Hwang, H. "Sequencing Delivery and Receiving Operations for Yard Cranes in Port Container Terminals," International Journal of Production Economics 84, 283-292, 2003.

[12] Ng, W. C. and Mak, K. L., "Yard Crane Scheduling in Port Container Terminals," Applied Mathematical Modelling 29, 263-276, 2005.

[13] Liang, C., Ma, X. and Chen, M., "Study on Yard Crane Scheduling with Multiple Container Flows in a Container Terminal," Journal of Quality 18(4), 375-392, 2011.

[14] Bostel, N. and Dejax, P., "Models and Algorithms for Container Allocation Problems on Trains in a Rapid Transshipment Shunting Yard," Transportation Science 32(4), 370-379, 1998.

[15] Kozan, E., "Optimizing Container Transfers at Multimodal Terminals," Mathematical and Computer Modelling 31, 235-243, 2000.

[16] Jeong, B. J. and Kim, K. H., "Scheduling Operations of a Rail Crane and Container Deliveries between Rail and Port Terminals," Engineering Optimization 43(6), 597-613, 2011.

[17] Manier, M. A. and Bloch, C. "A Classification for Hoist Scheduling Problems," The International Journal of Flexible Manufacturing Systems 15, 37-55, 2003.

[18] Phillips, L. W. and Unger, P. S. "Mathematical Programming Solution of a Hoist Scheduling Problem," AIIE Transactions 8(2), 219-321, 1976.

[19] Lei, L. and Wang, T. J. "A proof: The cyclic Hoist Scheduling Problem is NP-complete," Working paper, Rutgers University, 1989.

[20] Lei, L. and Wang, T. J. "Determining Optimal Cyclic Hoist Schedules in a Single-hoist Electroplating Line," IIE Transactions 26, 25-33, 1994.

[21] Chen, H., Chu, C. and Proth, J. M. "Cyclic Scheduling of a Hoist with Time Window Constraints," IEEE Transactions on Robotics and Automation 14(1). 144-152, 1998.

[22] Lei, L. and Wang, T. J. "The Minimum Common-cycle Algorithm for Cycle Scheduling of Two Material Handling Hoists with Time Window Constraints," Management Science 37, 1629-1639, 1991.

[23] Zhou, Z. and Li, L. "A Solution for Cyclic Scheduling of Multi-hoists without Overlapping," Annals of Operations Research 168, 5-21, 2009.

[24] Manier, M. A., Varnier, C. and Baptiste, P. "Constraint-based Model for the Cyclic Multi-hoists Scheduling Problem," Production Planning & Control 11(3), 244-257, 2000.

[25] Che, A., Chu, C. and Chu, F. "Multicyclic Hoist Scheduling With Constant Processing Times," IEEE Transactions on Robotics and Automation 18(1), 69-80, 2002.

[26] Leung, J. M. Y., Zhang, G., Yang, X., Mak, R. and Lam, K. "Optimal Cyclic Multi-Hoist Scheduling: A Mixed Integer Programming Approach," Operations Research 52(6), 965-976, 2004.

[27] Che, A. and Chu, C. "Single-track Multi-hoist Scheduling Problem: A Collision-free Resolution based on a Branch-and-bound Approach," International Journal of Production Research 42(12), 2435-2456, 2004.

[28] Leung, J. M. Y. and Levner, E. "An Efficient Algorithm for Multi-hoist Cyclic Scheduling with Fixed Processing Times," Operations Research Letters 34, 465-472, 2006.

A Non-linear Parametric Second-Degree Model for the Lifetime of Ultraviolet Cu+ Ne-CuBr Laser

I. Iliev[1] **and S. Gocheva-Ilieva**[2]
[1]Department of Physics, Technical University of Sofia –branch Plovdiv, Plovdiv, Bulgaria
[2]Department of Applied mathematics and Modeling, Plovdiv University "Paisii Hilendarski", Plovdiv, Bulgaria

Abstract - *A non-linear statistical parametric model of second degree for determining the influence of 10 input laser operational characteristics over the average service life of ultraviolet (UV) copper ion excited neon copper bromide (Cu+ Ne-CuBr) laser is developed. The model is built on the basis of available experimental data. It describes in an explicit form the dependence of the basic laser characteristics such as laser tube geometry, supplied electric power, pulse repetition frequency, gas pressures and others on the laser lifetime. The adequacy of the regression model is established. Analysis and physical interpretation of the obtained results is performed. The model is applied for estimation and simulation of the laser lifetime, which is important for further development of the laser device.*

Keywords: UV Cu+ Ne-CuBr Laser, Stepwise Regression Analysis, Non-linear Parametric Model, Second-degree Model, Laser Lifetime

1 Introduction

It is considered that copper vapor and copper compound vapor lasers are well-known and thoroughly studied. These have a number of unique properties - they continue to be the most powerful laser sources in the visible spectrum (516.6 nm and 578.2 nm), demonstrating high coherence and convergence of the laser beam. For this reason, they continue to be the subject of active development and commercial interest. In particular, this type of lasers are developed as a source of ultraviolet (UV) radiation. The copper ion laser with copper bromide vapor (UV Cu+ Ne-CuBr) is a promising innovative product with stable operation and unique characteristics. The first lasers of this type were constructed at the Metal Vapor Lasers Laboratory at the Georgi Nadjakov Institute of Solid State Physics of the Bulgarian Academy of Sciences in 1999 [1]. The copper ion laser with copper bromide vapor emits in the deep ultraviolet spectrum at five spectral lines - 248.6, 252.9, 259.7, 260.0 and 270.3 nm. Through experimentation, the maximum average output power of 1.3 W was achieved for all five lines and 0.85 W for the 248.6 nm line. It was also found that adding small quantities of hydrogen (0.02-0.04 Torr) doubles the average output laser power.

Due to the narrow emitting range in some spectral lines and the high coherence of the beam, the copper ion laser with copper bromide vapor is used in processes which require high resolutions, such as recording information, fluorescence, fine drilling, cutting, cleaning, modifying new materials, physics-chemistry, etc. [1, 2]. It has various applications in medicine, microelectronics, microbiology, photolithography, genetic engineering, etc. The laser is to be developed for industrial purposes. In order to accelerate this process, experimental studies are being conducted as well as mathematical and statistical modeling studies [3 - 5]. Some simulation results have been obtained using the software prototype LasSim [6].

Laser output power is the most important output characteristic of laser devices, and this is also true for metal vapor and metal compound vapor lasers. This characteristic is the subject of numerous experimental and theoretical studies. Although laser devices are not mass produced, another output characteristic is often examined - laser efficiency. There are practically no studies related to the laser lifetime of the laser device. This information is very important for industrial applications and is always given in the laser device specifications.

The goal of this paper is to use regression analysis to further the investigation of the lifetime of a UV Cu+ Ne-CuBr laser. In our previous paper the lifetime of this laser was modeled by using multivariate statistical techniques such as cluster analysis, factor analysis and principal component regression [7]. In essence, this paper is a continuation of [7] aiming to extract non-linear dependences. Since laser sources are highly non-linear, it is important to establish the non-linear relationships and their influence on the service life of the laser source.

The statistical analysis was performed by using SPSS software package [8].

2 Subject of investigation

The subject of investigation is a copper bromide vapor ion laser [1, 2]. A general diagram of the laser source is given in Fig. 1.

The following 10 independent quantities (variables, predictors) are considered: *D* (mm) – inner diameter of the laser tube, *DR* (mm) – inner diameter of the rings, *L* (cm) – distance between electrodes (active volume length), *PIN* (kW) – input power taking losses into account, *PH*2 (Torr) – hydrogen pressure, *PL* (W/cm) – specific power per unit length, *PRF* (kHz) – pulse repetition frequency, *PNE* (Torr) – neon pressure, *C* (nF) – equivalent capacitance of the condenser battery, *TR* (°C) – temperature of the copper bromide reservoir.

The dependent variable is *Ltime* – average lifetime of the laser source.

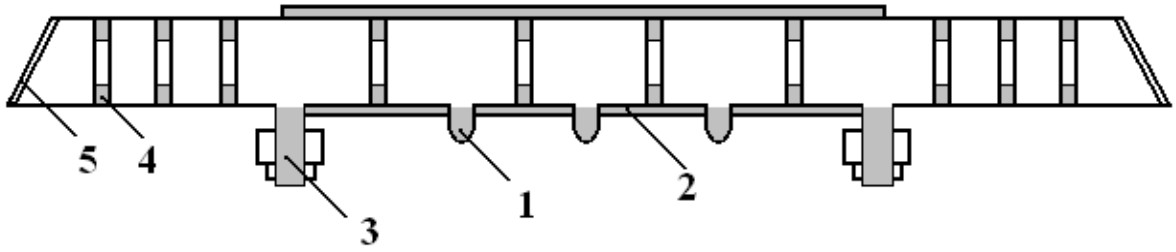


Fig. 1. Conceptual schematic of the laser tube: 1 - copper bromide reservoirs, 2 - heat insulation of the active volume, 3 - copper electrodes, 4 - inner rings, 5 - output windows.

3 Building a model of laser lifetime using the multiple stepwise regression method

Multivariate regression analysis is used to build models which provide a quantitative description of the relationships between several independent variables $x_1, x_2, ..., x_p$ and one or more dependent variables *y*. In the common case, an explicit functional relationship of the type $\hat{y} = f(x_1, x_2, ..., x_p; a_1, a_2, ..., a_m)$ is needed to express the influence of individual independent variables on the dependent quantity, where $a_1, a_2, ..., a_m$ are the unknown parameters.

In a linear case, under the condition for normal distribution of dependent variable, the regression equation is searching in the following form:

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_p x_{ip}, \quad i = 1, 2, ..., n \quad (1)$$

where $y_i = \hat{y}_i + \varepsilon_i$, $i = 1, 2, ..., n$, $b_0, b_1, ..., b_p$ are regression coefficients, $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)$ is the error vector, *n* is the sample size.

The relative influence of each quantity is determined using a z-scores transformation

$$z_j = \frac{x_j - \overline{x}_j}{s_j}, \qquad j = 1, 2, ..., p$$

where $\overline{x}_j$ is the mean value and s_j is the standard deviation of the variable x_j. The resulting standardized regression equation has the form:

$$\hat{y} = \beta_1 z_1 + \beta_2 z_2 + ... + \beta_p z_p . \quad (2)$$

In (2) coefficients $\beta_1, \beta_2, ..., \beta_p$ are called standardized coefficients. These indicate the relative influence of each independent variable on the dependent one within the equation. The least squares method is used to determine the regression coefficients.

As noted above, the goal of this paper is to determine the degree of influence each of the ten independent variables has on the dependent quantity *Ltime*. The studies carried out until now on the output characteristics - laser output power and laser efficiency - indicate that these quantities are non-linearly dependent on the ten independent quantities. For this reason, we are looking for the coefficients of the second order regression equation in the following form:

$$Ltime = b_0 + \sum_{j=1}^{10} b_j x_j + \sum_{\substack{j,k=1 \\ k \le j}}^{10} b_{j,k} x_j x_k . \quad (3)$$

Equation (3) is comprised by first order quantities and all possible combinations of the two elements, including repetitions. A total of 66 unknown coefficients need to be defined. Out of all unknown coefficients, only those which are statistically significant at a level of Sig ≤ 0.05 are to be chosen. To this end, stepwise regression is performed using the SPSS software package [8]. The latter has the advantage

of monitoring the level of significance of unknown coefficients. Each coefficient for which the index Sig. ≤ 0.05 is not fulfilled is removed from the equation and no longer participates in the regression analysis.

The results are given in Table 1. In this case, 10 steps were needed to calculate all statistically significant coefficients and to remove insignificant ones.

Based on the obtained results from Table 1, the following unstandardized equation can be written:

$$\begin{aligned}\hat{L}time = {} & 4182.369 - 99.837PL + 1.859D.PNE \\ & + 0.493D.TR - 1.132DR.L - 4.445DR.PNE \\ & - 23.277DR.PL - 0.038PIN.PRF \\ & - 789.072PNE.PH2 + 0.104PHE.C \\ & + 27.291PH2.TR - 0.020TR.C - 7.827D^2 \\ & + 2.439DR^2 + 0.056L^2 - 0.463PH2^2 \\ & + 0.562PRF^2 - 0.006TR^2 + 197.974PL^2 \\ & + 0.006C^2\end{aligned} \quad . (4)$$

The resulting regression equation (4) includes all 10 independent variables with only one being of first degree - *PL*. All other terms are of the second degree. This once again shows that the physical processes which occur in the laser tube are strongly nonlinear and in particular this applies to their influence on the service life of the laser device. Practically none of the independent variables influence the dependent variable *Ltime* on their own, rather this is the result of complex interactions of the former.

In view of the statistical analysis, it needs to be noted that the ANOVA for the whole model (4) is statistically significant at a level of 5% and a significance index Sig. =0.000. The examination of the residuals showed that they are normally distributed. It can be concluded that there is no process which has been excluded when building model (4). It is also important to note that all obtained errors are within 5% which fits well with the 5% measurement error. The latter applies to all variables participating in the modeling process.

In order to check the adequacy of (4), some additionally calculated specific values of the *Ltime* according to (4) are given in column $\hat{L}time$ of Table 2. Table 2 shows that the average relative error for the values of *Ltime* predicted using the model and experiment measurements is within 2.5%.

Table 1. Regression analysis results by the stepwise regression. Dependent variable: *LTime*.

Model Variables	Unstandardized Model Coefficients		Standardized Coefficients	Critical t-Value	Sig.
	B	Std. Error	Beta (β_j)		
Constant (b_j)	4182.369	171.650		24.366	0.000
PL	-99.837	30.456	-0.517	-3.278	0.001
D.PNE	1.859	1.078	0.178	1.724	0.046
D.TR	0.493	0.017	5.563	29.379	0.000
DR.L	-1.132	0.118	-1.007	-9.609	0.000
DR.PNE	-4.445	1.080	-0.429	-4.115	0.000
DR.PL	-23.277	3.164	-0.616	-7.358	0.000
PIN.PRF	-0.038	0.007	-0.650	-5.314	0.000
PNE.PH2	-789.072	203.875	-0.765	-3.870	0.000
PNE.C	0.104	0.019	0.969	5.504	0.000
PH2.TR	27.291	6.800	0.789	4.013	0.000
TR.C	-0.020	0.001	-3.257	-19.829	0.000
D.D	-7.827	0.141	-4.800	-55.630	0.000
DR.DR	2.439	0.445	0.515	5.481	0.000
L.L	0.056	0.010	0.189	5.524	0.000
PH2.PH2	-0.463	0.116	-0.289	-4.001	0.000
PRF.PRF	0.562	0.113	0.269	4.979	0.000
TR.TR	-0.006	0.000	-0.685	-15.131	0.000
PL.PL	197.974	24.432	1.124	8.103	0.000
C.C	0.006	0.000	2.445	17.668	0.000

Table 2. Adequacy test of the developed non-linear model by comparing the experiment data with the predicted values.

D, mm	*DR*, mm	*L*, cm	*PIN*, W	*PNE*, Torr	*PH*2, Torr	*PRF*, KHz	*TR*, ^{0}C	*PL*, W/cm	*C*, pF	*Ltime* , Hour	$\hat{Ltime}$, Hour	Relative Error, %
7	7	87	1300	17	0.03	20	495	7.47	735	180	187.6	4.2
7	7	87	1300	16	0.03	20	495	7.47	735	180	176	2.2
7	7	87	1300	17	0.03	20	495	7.47	735	180	187.6	4.2
7	7	87	1300	17	0.04	20	495	7.47	372.2	700	702.4	0.34
7	7	87	1300	18	0.028	20	495	7.47	372.2	700	679.1	3
7	7	87	1400	17	0.026	20	495	8.05	372.2	700	688.4	1.66
7	7	87	1400	17	0.021	20	495	8.05	372.2	700	685.9	2.01
7	7	87	1400	17	0.031	20	495	8.05	372.2	700	690.8	1.31
7	7	87	1400	17	0.03	20	490	8.05	372.2	700	730.7	4.38
7	7	87	1600	17	0.03	25	490	8.05	372.2	700	711	1.57

This indicates that the developed regression model describes adequately the data obtained from the experiment.

The coefficients of the standardized equation are given in Table 1, column Standardized coefficients (Beta). As mentioned, every coefficient indicates the degree of influence of each pair on the independent variable on the laser lifetime variable *Ltime* . Practically, it is impossible to estimate the individual influence of 10 independent variables because each of these does not affect directly *Ltime* but only in combination with some of the other quantities, sometimes even more than once.

4 Simulation, analysis and physical interpretation of the obtained results

In order to estimate the individual influence of each of the 10 independent variables, a simulation is to be carried out using the unstandardized equation (4). Based on the last row of Table 2 each of the 10 quantities is increased within 10% of the newly-calculated value for *Ltime* and its relative change in percent.

Table 3 shows the results from the simulation, namely - the relative change of quantities which indicates the degree of influence each has on the quantity *Ltime* . We will consider in more detail the influence of just 4 variables, which can be physically interpreted - these are the quantities *TR*, *C*, *PH*2 and *PRF*.

The temperature of the copper bromide reservoirs *TR* has a negative influence on *Ltime* . When the temperature of the reservoirs go up, the concentration of CuBr vapors increases along with laser generation. At the same time, CuBr losses also increase due to diffusion in the cold zones of the laser tube. Since the laser tube is unsoldered, the high rate of loss of CuBr leads to a rapid fall in laser generation over time and therefore the period of laser lifetime *Ltime* goes down, too.

The pulse repetition frequency *PRF* also has a negative influence. When the pulse repetition frequency goes up, the sputtering of the electrodes and the contamination of the laser medium also increase. This reduces laser generation and the laser lifetime duration.

When the equivalent capacitance of the condenser battery *C* increases, so does the electric power in the laser tube according to the expression $E = U^2 C / 2$, where *U* is the voltage between tube electrodes. This leads to the physical erosion of the laser tube and reduces the service life of the laser device.

Table 3. Simulation estimating the relative influence of 10 independent laser parameters on the service life of a UV laser.

Independent variables, increased by 10%	Relative change of *Ltime* in %
D	0.720641
DR	-1.60659
L	-17.3644
PIN	20.83068
PNE	-0.03434
*PH*2	0.00119
PRF	-0.01495
TR	-2.98427
PL	21.44909
C	-0.80259

The hydrogen pressure *PH*2 influences positively the laser lifetime *Ltime* . When *PH*2 is increased, at a constant supplied electric power *PIN*, the current passing through the

tube decreases at the expense of a higher applied voltage. The reduced current limits the sputtering of electrodes extending the lifetime of the laser tube, so *Ltime* increases.

5 Conclusions

For the first time, an explicit parametric second-order model is developed for the service life of UV Cu+ Ne-CuBr lasers. A regression equation is obtained, which includes all 10 independent laser operational variables. An adequacy test is performed for the statistical model with known experiment results. Using the model, a computer simulation is carried out, providing a qualitative assessment of the relative influence of each independent quantity on the behavior of the laser lifetime. The obtained results are analyzed and interpreted physically.

6 Acknowledgment

This paper is supported by the Plovdiv University "Paisii Hilendarski" grant NPD-NI13 FMI-002.

7 References

[1] Nikolay K. Vuchkov, Krasimir A. Temelkov, Peter V. Zahariev, and Nikola V. Sabotinov. "UV Lasing on Cu+ in a Ne-CuBr Pulsed Longitudinal Discharge," IEEE J. Quantum Electron., Vol. 35, Issue 12, pp. 1799-1804, 1999.

[2] Nikolay Vuchkov. "High Discharge Tube Resource of the UV Cu+ Ne-CuBr Laser and Some Applications," in: New Development in Lasers and Electric-Optics Research, W. T. Arkin, ed. Nova Science Publishers, pp. 41-74, 2007.

[3] Snezhana G. Gocheva-Ilieva and Iliycho P. Iliev. "Statistical Models of Characteristics of Metal Vapor Lasers". Nova Science Publishers Inc., 2011.

[4] Snezhana G. Gocheva-Ilieva, Desislava S. Voynikova, and Ilycho P. Iliev. "Modeling of Output Characteristics of a UV Cu+ Ne-CuBr Laser", Math. Probl. Eng., Vol. 2012 (420782), pp. 1-22, 2012.

[5] Desislava S. Voynikova. "Application of Multivariate Nonparametric Statistical Methods, PhD Thesis, Plovdiv University 'Paisii Hilendarski', Plovdiv, Bulgaria, 2013 (in Bulgarian).

[6] Snezhana G. Gocheva-Ilieva and Chavdar P. Kulin. "Development of LasSim Software Prototype for Simulating Physical Characteristics of Laser Devices", Scientific Works of Plovdiv University, Vol. 37, Book 3-Mathematics, pp. 45-52, 2010.

[7] Iliycho P. Iliev and Snezhana G. Gocheva-Ilieva, "Study of UV Cu + Ne – CuBr Laser Lifetime by Statistical methods", Quantum Electronics, Vol. 43, Issue 11, pp. 1014 – 1018, 2013.

[8] SPSS IBM Statistics 22, [Online]. Available: http://www-01.ibm.com/software/analytics/spss/

Automatic Generation of Animatable 3D Personalized Model Based on Multi-view Images

Seong-Jae Lim, Ho-Won Kim, Jin Sung Choi
CG Team, Contents Division
ETRI
Daejeon, South Korea
sjlim@etri.re.kr

Bon-Ki Koo
Contents Division
ETRI
Daejeon, South Korea

***Abstract*—We propose a fully-automatic 3D model generating method of individual people from multi-view images. An automatically constructed 3D model can be deformed and animated by controlling the joint. An animatable 3D personalized model of individual people is generated by transferring the joint-skeleton structure and appearance of a generic 3D human model to an individual 3D volumetric model reconstructed from multi-view images. We automatically estimate the joint position of individual people by computing a weight function consisting of the kinematic joint-skeleton structure of the generic 3D human model and anthropometric information. Our generic 3D human model approximates the whole body using sweep surfaces. The vertices on the object boundary are bound to the sweep surfaces and follow their deformation. Thus, the animatable 3D individual model can be generated by transferring the sweep surface of the generic 3D model to the individual 3D volumetric model.**

Keywords-Personalized modeling; animatable model; joint-skeleton; sweep surface; transfering

I. INTRODUCTION

A. Motivation

There has been increasing demand to reconstruct moving 3D shapes and motions of individual people. Properly estimated 3D human models are thus very useful for a variety of applications including augmented reality, free-viewpoint video [1], media production for 3D television, sports industry, surveillance, virtual education, virtual shopping, gaming, and so on. Current methods for human modeling and motion capture are based on an active 3D sensing to build the shape of the body surface and on an optical motion capture system to capture the movements of the body. Such systems are prohibitively expensive and require interactive manual work to build high-quality models. In addition, meshes of a 3D model are often animated by hand using keyframing and procedural deformations. Current procedural approaches can generate mesh deformations with greater ease and efficiency but they are difficult to control when the goal is to match a particular motion or real performance.

Current systems for whole-body shape and motion capturing are based on shape-from-silhouette approaches and model-based approaches. However, for the generation of artificial renditions of a scene from arbitrary novel viewpoints and for reusability of the reconstructed shape and motion, a model-based approach is more powerful. In order to achieve reliable motion capture of individual people, an articulated 3D human model that is very similar to the moving subject is essential. Modeling of 3D is becoming much easier than before, but it is still not simple and requires tedious tasks because the user must rig the model manually.

In this paper, we present a fully-automated method for constructing an animatable 3D mesh model of individual people. A 3D human model of an individual person is generated by transferring the joint-skeleton structure and appearance of a generic 3D model to an individual 3D volumetric model reconstructed from multi-view images.

B. Background

Most prior related research mainly deals with constructing a surface skin model. 3D laser-scanner systems [2] capture the entire surface of a whole body, but such systems are highly priced and require interactive manual work. On the other hand, reconstructing systems of whole-body models from captured multi-view images are much cheaper and more easily available. The shape-from-silhouette approaches [3-4] are a popular method that can generate a 3D volumetric model from multi-view silhouette images. However, shape-from-silhouette approaches depend on the segmentation results and have a 'blocky' problem if there are an insufficient number of views.

There are researches that have attempted to estimate the joint locations [5] in order to estimate a skeleton from a sequence of volume data of rigid bodies. However, the resultant skeleton is an estimated stick-figure-like structure that is not for realistic character animation or skinning. There is also a 3D modeling method using shape feature points, limb outlines, and a generic 3D model to yield a final customized 3D model [6]. However, user interaction for feature points is needed and skinning/animation using anatomical skeleton of 3D generic model is not easy or simple. Automatic rigging presents a method for animating characters automatically [7].

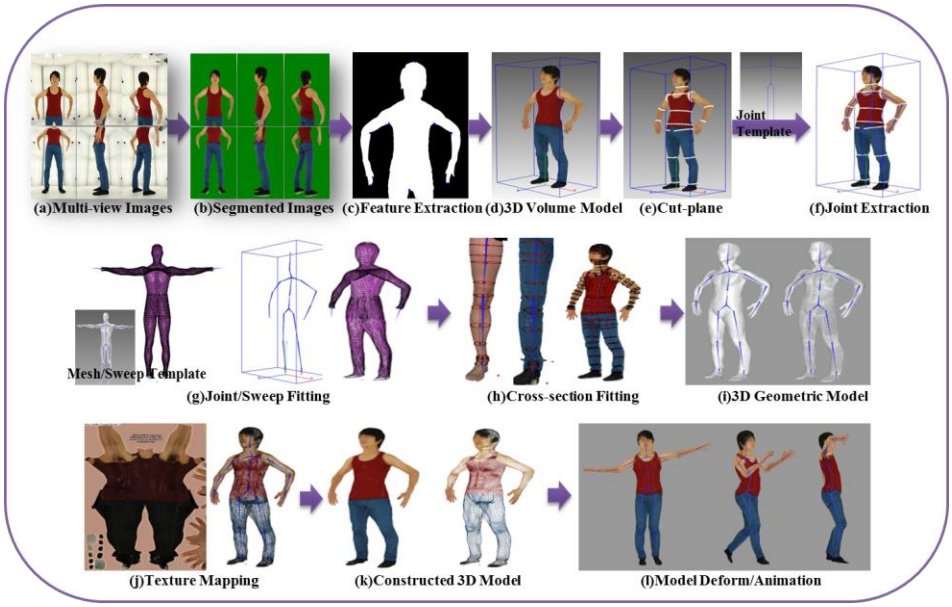


Figure 1. Overview of our method.

This algorithm adapts the skeleton to a character by minimizing a penalty function, and attaches it to the surface, allowing skeletal motion data to animate the character. However, this method assumed there are 3D mesh models for rigging and animation. Park and Hodgins[8] used skin motion-capture data for transferring template mesh model to individual people and they captured skin deformation of individual people. This algorithm has some limitation that should use motion-capture system and use a lot of markers(approximately 350 markers) on skin of human.

C. *Outline of Our Approach*

In this paper, we address the automatic construction of an articulated 3D human model that can be deformed and animated. An animatable 3D human model of individual people can be generated by transferring a joint-skeleton structure and appearance of the generic 3D human model to an individual 3D volumetric model reconstructed from multi-view images. Our algorithm consists of two main steps: joint positioning and appearance transfer. We automatically estimate the joint position of individual people in multi-view images by computing a weight function consisting of the kinematic joint-skeleton structure of the generic 3D human model and anthropometric information. The generic 3D human model is called the template model. In order to create a deformable and animatable 3D model of individual people, we present the transferring method of the sweep-based 3D human model. The sweep-based approach [9-10] represents the appearance of a 3D model as sweep surfaces with elliptic cross-sections. A sweep surface is generated by approximating these elliptic 2D cross-sections with swept 2D cross-sectional shapes of various human body parts, and the vertices of a human model are bound to the sweep surfaces and then follow their transformations. By controlling the size and orientation of one key ellipse assigned to each joint, we can deform the sweep surfaces, and then the vertices of the human model bound to the sweep surfaces follow the change. In [12], a star-shaped cross-sectional closed curve is used instead of elliptic cross-sections. This is further extended for freeform deformation. The important advantage of a sweep-based approach for deformation and animation of a 3D model is volume preservation. We can generate an animatable 3D human model of individual people through the transferring its appearance, and by sweeping the surfaces of the sweep-based 3D human model into 2D silhouettes and reconstructing the silhouettes into a 3D volumetric model.

An overview of the automatic generating method of a 3D human model is illustrated in Fig. 1.

The rest of this paper is organized as follows. Section II introduces the preprocessing of 3D modeling, including silhouette extraction, feature point extraction, and reconstruction of a 3D volumetric model. Section III describes the joint positioning. The appearance transfer of a sweep-based 3D model is presented in section IV. Section V describes the post-processing of 3D modeling, including refinement and texturing. Finally, section VI contains the experimental results of automatic generation of a 3D human model.

II. PRE-PROCESSING OF 3D MODELING

A. *Sihouette Extraction and Feature Extraction*

Multi-view images are provided from several angles using multiple calibrated cameras. General background subtraction using CIELAB color space and a standard chroma-key technique are used to identify the background pixels. To maintain a consistent environment, we use the limited lighting condition and a static background. This consistent environment can allow the general background subtraction technique to provide a good segmentation result. In addition, appropriate threshold values and using the Lab color space characteristics help distinguish between foreground and background. A silhouette curve is extracted following an 8-connected-pixel chain on the border of the foreground.

The objective of feature extraction is to represent each human part such as the head, neck, arms, legs and torso of individual people in multi-view images. Extracted feature points are used to estimate each joint position of the 3D model. We therefore perform robust feature extraction for a wide range of changes in body shape, size, and clothing by constraining of the initial pose and wearing clothing. To achieve this we use a pre-specified pose, as represented in Fig. 2, and clothing that allows both the armpits and crotch to be visible.

The algorithm for extracting feature points [11] from the frontal upper, frontal lower, and side binary images is presented in Fig. 2. Given a foreground mask, an edge pixel list $\{e_i\}_{i=1}^{N}$ is extracted according to the foreground boundary in a clockwise direction.

We traverse the edge pixel list to locate extremum points that correspond to the head, hands, and feet, as shown in Fig. 2(a), and to identify feature points located with large changes in shape and pose. Those points correspond to the crotch, arm-pits, and elbows shown in Fig. 2(c).

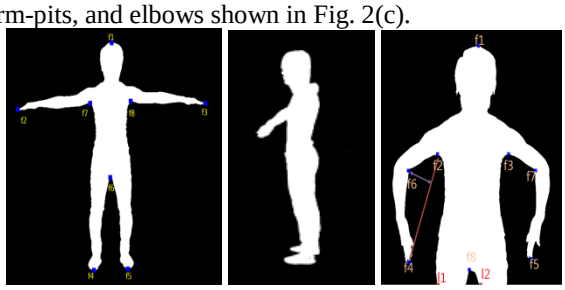


Figure 2. Segmented Foreground Images: (a) front of T-pose, (b) side, (c) front of N-pose.

B. 3D Volumetric Reconstruction

We reconstruct a 3D volumetric model of individual people for the search of 3D joint position, and a reference to transfer the appearance of a generic model into the shape of individual people through 3D-to-3D appearance mapping. We use the photo-consistent scene recovery method [13] for reconstruction of the 3D volumetric model. This methodology solves many problems associated with previous stereo-based and volumetric methodologies by introducing of a self-constrained greedy-style optimization technique based on the probabilistic shape photo-consistency measure.

III. JOINT POSITIONING

We use a kinematic joint-structure with 21 joints including the head, neck, spine, root, shoulders, elbows, wrists, hips, knees, ankles, head-end, hands-end, and ends of the feet. Figure 3 depicts the sweep-based 3D human model used for the generic 3D model of this paper. The joint-skeleton structure for the generic humanoid model is illustrated in Fig. 4. To estimate each joint position, landmarks, approximate 3D coordinates of each joint, are defined by estimating the initial joint position based on the feature points, described in section II, and the evaluating process is then performed on the initial joint position with a weight function consisting of the configuration ratio of the template's joint structure and anthropometric information.

For the initial joint position, we obtain 3D feature points corresponding to the 2D feature points extracted in section II through a 2D-to-3D linear mapping between the silhouette image and reconstructed 3D volumetric model. Most of the 3D feature points are located in an approximate x-y coordinate of each joint because the joints exist inside the 3D model. For the z coordinate of the 3D feature points, we use a 3D cut-plane on the reconstructed 3D volumetric model. First, we set the 3D cut-plane on each feature point. The orientation of each cut-plane is determined by the orthogonal directions between two neighboring feature points and each body part of the 3D volumetric model. Then, we can find the voxels intersected with each cut-plane. The centroid of the intersected voxels is defined as the initial joint position.

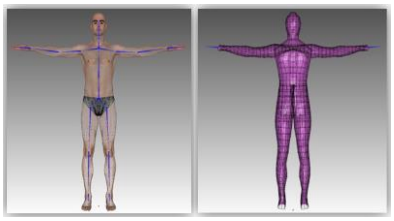


Figure 3. Sweep-based 3D human model.

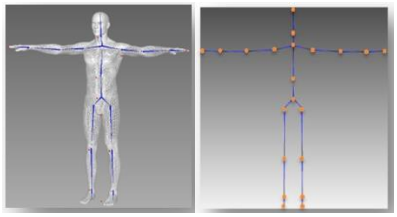


Figure 4. Joint-skeleton structure of 3D generic model.

Figure 5 depicts 3D cut-planes on 3D feature points, and the initial joint position from the centroid of the cut-plane.

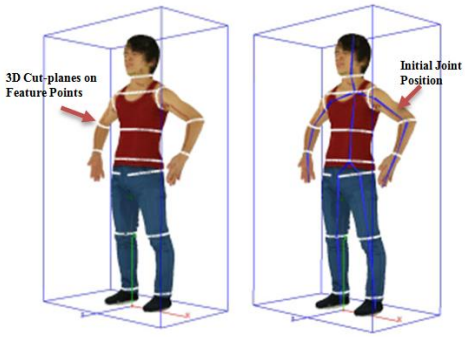


Figure 5. Cut-planes and initial joint position on reconstructed 3D volume model.

Anthropometric information, which is the measurement of the human body, exists. This information is the statistical data regarding the distribution of body dimensions in a population [13]. We use this anthropometric information for our evaluation function. The goal of our evaluation function is to evaluate and refine the initial joint position by using the kinematic joint-skeleton features of the generic 3D model and anthropometric information. The equations are given below, where $Init(i)$ is the initial ^{ith} joint position, $E_t(i)$ is the distortion ratio between template model and initial joint position, $E_a(i)$ is the distortion ratio between anthropometric information and initial position, $w_i(i)$ is the weight of the initial joint position, $w_t(i)$ is the weight of the template model, and $w_a(i)$ is the weight of the anthropometric feature on ^{ith} joint. Actually, $w_i(i)$ is defined as the reliability level

of the 3D feature points. The total sum of the weight should be 1.

$$\mathrm{Eval(i)} = \left(\mathrm{w_i(i)} \times \mathit{Init(i)}\right) \times \left(\mathrm{w_t(i)} \times \mathrm{E_t(i)} + \mathrm{w_a(i)} \times \mathrm{E_a(i)}\right)$$

After evaluating the initial joint position, we solve the forward kinematics with the newly evaluated joint position. We solve the forward kinematics by computing the link-length, global and local position, and rotation of each joint based on the estimated joint position. Solving the forward kinematic provides estimates of the joint position and rotation of several poses of individual people acquired from multi-view cameras.

IV. Transfer of Appearance and Sweep Surface

We use a sweep-based 3D human model that can be simply deformed by a joint angle change and animated with motion capture data. To create an animatable 3D personalized model of individual people captured from multi-view cameras, we transfer the appearance of the template model to a reconstructed 3D volumetric model, as shown in section II, based on the extracted joint-skeleton structure shown in section III. A reconstructed 3D volumetric model is called the target model. Since the mesh of the template model is bound to the sweep surface, we fit the appearance of the template model to the shape of the target model by controlling the parameter of the sweep surface.

A. *Global fitting*

Global fitting of the template model is started with the joint-skeleton structure described in section III, which is called the target joint-skeleton structure. Position fitting of a template's joint is first performed to the corresponding target joint as shown in Fig. 6. After fitting the position, the forward kinematic structure solves the connectivity, link-length between each joint, and each joint's orientation. At this point, each sweep surface of the body corresponding to each joint will be transferred by fitting the joint.

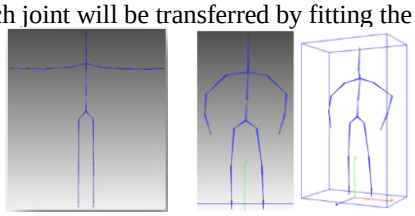


Figure 6. Joint-skeleton structureof template target model.

In this paper, we use star-shaped 2D cross-sections that approximate the cross-sectional shapes of various human body parts [12]. Figure 7 represents cross-sections of the whole body and the left arm of the template model. When a star-shaped cross-sectional closed curve of variable size $O_t(\theta)$ is moving under rotation $R(t)$ and translation $C(t)$, it generates a sweep surface $S(\theta,t)$ by interpolating cross-sections $O_t(\theta)$(detailed in [12].

$$S(\theta,t) = C(t) + R(t) \cdot O_t(\theta)$$

$$= \begin{bmatrix} x(t) \\ y(t) \\ z(t) \end{bmatrix} + \begin{bmatrix} r_{11}(t) & r_{12}(t) & r_{13}(t) \\ r_{21}(t) & r_{22}(t) & r_{23}(t) \\ r_{31}(t) & r_{32}(t) & r_{33}(t) \end{bmatrix} \cdot \begin{bmatrix} r(\theta,t)\cos\theta \\ r(\theta,t)\sin\theta \\ 0 \end{bmatrix}$$

This sweep surface represents a time-variant star-shaped cross-section using scalar radius function $r(\theta,t)$.

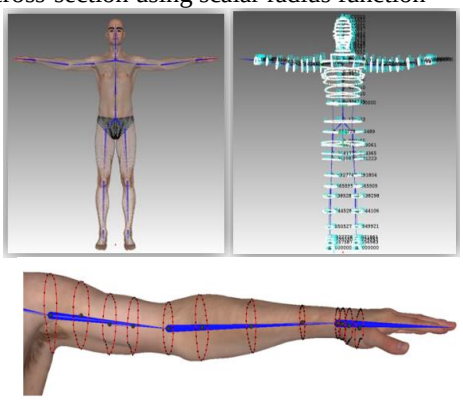


Figure 7. Star-shaped cross-sections of the template model.

Joint fitting that fits the position, orientation, and link-length of the joint change the rotation and orientation of the related cross-sections, which are called keyframes. Changing of the keyframes deforms the shape of the sweep surface. Thus, sweep fitting follows the joint fitting. Results of joint and sweep fitting are shown in Fig. 8.

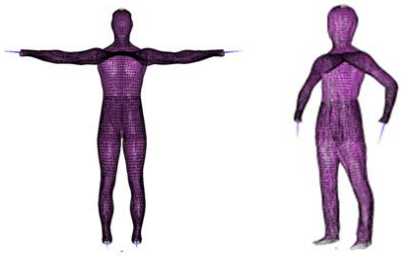


Figure 8. Joint/Sweep fitting.

B. *Local fitting*

To achieve local fitting, we can fit the cross-sectional shape of a sweep surface to the corresponding cross-section of the reconstructed 3D volumetric model by modifying the scalar radius function $r(\theta,t)$. After joint fitting, each cross-section of the target model is computed by cutting the 3D volumetric model of the corresponding part with planes. The center of each cross section is computed, and then corresponding radii from the center to the boundary voxels of the cross-section are sampled. The new radius function $r(\theta,t)$ for local fitting is computed by the difference ratio of corresponding radii between the template and target model. At this time, we assign the weight for each feature to the new radius function $r(\theta,t)$ for considering the characteristic of the body part of the template model and complementing the drawbacks of the reconstructed 3D volumetric model such as occlusions, specular problems, and so on.

$$r_i'(\theta,t) = r_i(\theta,t) \times \left\{ \left(w_r(i) \times \frac{i\ th\ radius\ of\ templat}{i\ th\ radius\ of\ target} \right) \right\}$$

After fitting the joint and cross-section to the target model, a 3D geometric mesh model of the individual is constructed. Figure 9 shows the results of the cross-section fitting. Figure 9(a) represents cross-sections of a template model, and Fig. 9(b) shows overlapping cross-sections of a template model on a 3D volumetric model (target model). The fitted cross-sections after local fitting are represented in Fig. 9(c).

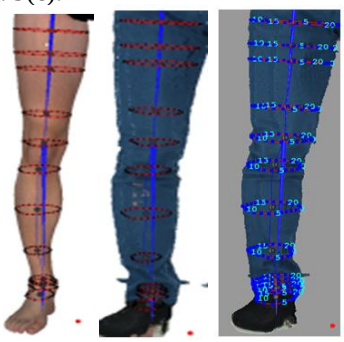


Figure 9. Cross-section of (a)template model, (b)3D volumetric model, and (c)3D volumetric model after local fitting.

V. POST-PROCESSING OF 3D MODELING

The transferred 3D model from the template model in section III and IV is an articulated 3D mesh model that can be deformable and animatable. However, displacements exist between the bounding vertices and the sweep surface. The displacement of the constructed 3D mesh model is from the template model, so we can adjust those vertices to the surface of the target model in detail.

For a whole body, we obtain a single texture map by back projecting and integrating all overlapping images from multi-view cameras. Integration of the texture map is based on the approximate 3D shape information for the reconstructed model. The texture map is depicted in Fig. 10.

Figure 10. Color texture map of the template and target model.

VI. EXPERIMENTAL AND RESULTS

The algorithm presented in this paper was evaluated in several sets of experiments with multi-view images containing people of different heights and body types, along with different types of poses such as ‘a T’ pose and ‘A’ pose.

Figure 11 shows the results of our 3D model constructing algorithm. Figure 11(a) shows the frontal image among multi-view images acquired from several individual people, Fig. 11(b) shows the feature in a segmented image, and Fig. 11(c) shows the reconstructed 3D volumetric model.

In addition, Fig. 11(d) shows the joint extraction by the cut-plane on the landmarks of each body part, Fig.11(e) shows the joint and sweep transfer, and lastly Fig.11(f) shows the final constructed 3D model of an individual.

Figure 12 illustrates the snapshots of the deformation and animation of the constructed 3D human model.

Figure 13 shows the evaluation result of the reconstructed mesh model. For comparison, we projected the reconstructed mesh model to silhouette boundary images and we could check the difference between the silhouette boundary(true) and the projected boundary of the reconstructed mesh model.

The experimental timing of our automatic constructing method for a 3D individual human model is within 40s on a 3.5 GHz Intel Core Duo with 8GB of RAM. The most time-consuming step is to reconstruct the 3D volumetric model from multi-view images.

VII. CONCLUSION

We have presented a fully-automated method for constructing an animatable 3D mesh model of individual people. The 3D human model is generated by transferring the joint-skeleton structure and appearance of a generic 3D model to an individual 3D volumetric model reconstructed from multi-view images. We automatically estimate the joint position of individual people in multi-view images by computing a weight function consisting of the kinematic joint-skeleton structure of the generic 3D human model and anthropometric information. In order to create a deformable and animatable 3D model of individual people, we present a transferring method of a sweep-based 3D human model. The sweep surface, interpolated by approximating a cross-section with a swept star-shaped cross-sectional closed curve, binds the vertices of the 3D generic model. By controlling the size and orientation of one key ellipse assigned to each joint, we can deform the sweep surfaces, and then the vertices of a human model bound to the sweep surfaces follow their changes. We can generate an animatable 3D human model by transferring the appearance and sweep surfaces of the sweep-based 3D human model into 2D silhouettes and reconstructing a 3D volumetric model from the silhouettes. Our proposed technique has unique specialty in contrast to the other method such as user-defined 3D model generation, only appearance deformation, respectively making 3D individual mesh model and skinning that model for animation.

In future work, we will investigate the optimal radius function $r(\theta,t)$ of the cross-sectional curve between the template and target model for local fitting in detail. In addition, we will use this customized model to track and capture the motion of individual subjects.

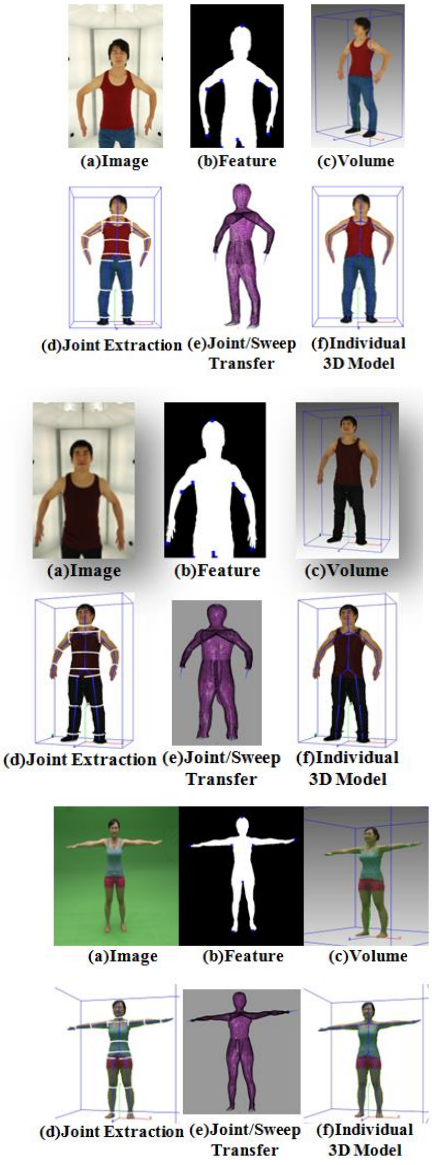


Figure 11. Experimental results of 3D modeling of a human subject.

Figure 12. Deformation and animation of a constructed 3D personalized model: its (upper)joint and appearance and its (lower) mesh model.

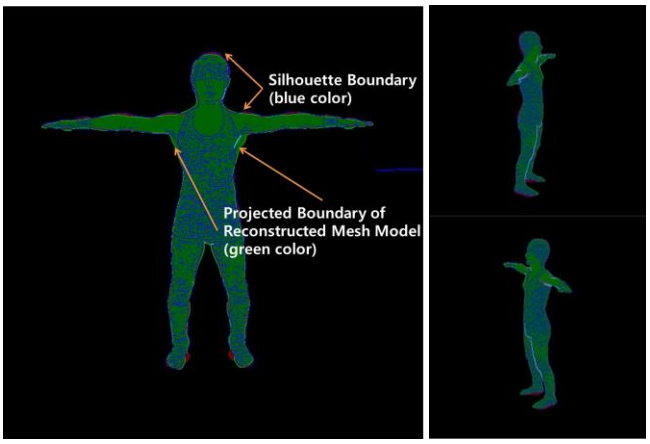


Figure 13. Comparing results of silhouette and projected boundary.

ACKNOWLEDGMENT

The research was supported by the strategic technology development program of MSIP/KEIT. [10047093, 3D Content Creation and Editing Technology Based on Real Objects for 3D Printing].

REFERENCES

[1] J. Carranza, C. Theobalt, M. Magnor, et al., "Free-Viewpoint Video of Human Actors," ACM Trans. Graph. , vol. 22, no. 3, 2003, pp. 569-577.

[2] Cyberware, http://www.cyberware.com

[3] A. Hilton, D. Beresford, T. Gentils, et al., "Whole-body Modeling of People from Multi-View Images to Populate Virtual Worlds," The Visual Computer, vol. 16, no. 7, 2000, pp. 411-436.

[4] S. Weik, "A Passive Full Body Scan using Shape from Silhouette," Proc. ICPR'00, 2000, pp. 99-105.

[5] C. Theobalt, E. Aguiar, M. Magnor, et al., "Marker-free kinematic skeleton estimation from sequence of volume data," Proc. ACM Virtual Reality Software and Technology'04, 2004, pp. 57-64.

[6] C. K. Quah, A. Gagalowicz, R. Roussel, et al, "3D Modeling of Humans with Skeleton from Uncalibrated Wide Baseline Views," Proc. CAIP'05, 2005, pp. 379-389.

[7] I. Baran and J. Popovic, "Automatic Rigging and Animation of 3D Characters," ACM Trans. On Graphics, vol. 26, no. 3, 2007, pp. -.

[8] S. I. Park, Jessica K. Hodgins, "Capturing and Animating Skin Deformation in Human Motion," ACM Trans. On Graphics, vol. 25, no. 3, 2006, pp.881-889.

[9] D. E. Hyun, S. H. Yoon, M. S. Kim, et al., "Modeling and Deformation of Arms and Legs based on Ellipsoidal Sweeping," Proc. PG'03, 2003, pp. 204-212.

[10] D. E. Hyun, S. H. Yoon, J. W. Chang, et al., "Sweep-based Human Deformation," The Visual Computer, vol. 21, no. 8, 2005, pp. 542-550.

[11] S. J. Lim, H. B. Joo, H. W. Kim, et al., "Automatic Rigging of 3D Human Models," Proc. FCV'10, 2010, pp. 223-226.

[12] S. H. Yoon and M. S. Kim, "Sweep-based Freeform Deformation," Proc. EG'06, 2006, pp. -.

[13] H. W. Kim and I. S. Kweon, "Appearance-Cloning: Photo-Consistent Scene Recovery from Multi-View Images," IJCV, vol. 66, no. 2, 2006, pp. 163-192.

[14] Anthropometry,http://www.wikipeida.com

Fourth-Order Compact Formulation Applied to the Resolution of Transmission Lines

Fatima ALLALI [1], **N. ALAA** [2], **A. GHAMMAZ** [1], **H. ROUIJAA** [1]
[1]LSET, University Cadi Ayyad, Marrakesh, Morocco
[2]LAMAI, University Cadi Ayyad, Marrakesh, Morocco

Abstract - *In this work a new beta method of fourth order applied to the resolution of homogeneous transmission lines is developed. A comparison with conventional methods used for this type of problems like β-method or order fourth is also given. Furthermore, various numerical experiments are presented to confirm the accuracy, efficiency and stability of our proposed method. In particular, these simulations show that our new schema is unconditionally stable and fourth-order accurate in space and time.*

Keywords: Transmission lines, modeling, electromagnetic wave, finite difference, high order Compact, beta method.

1 Introduction

One of the basic technologies of the twentieth century is without doubt electromagnetism which is the study of electric and magnetic fields and their interaction [1].

In this study we are concerned with the transmission lines which serve to guide the propagation of electromagnetic energy from a source terminal to a load terminal. These lines take many physical forms, including a twisted pair line used for telephone or internet connections, coaxial cable, or any of a number of systems wave guiding multi-conductor. Transmission Lines is a multidisciplinary field.

From the theory of electromagnetism, methods of numerical simulations to geometric modeling and visualization. Computing and numerical algorithms all have important roles to play. In this study, the focus is on the proposal of new numerical simulation methods.

Typically, modeling of the transmission line is one-dimensional and is represented by voltages and currents according to the line axis of the transmission line.

A transmission line is a specialized cable or other structure designed to carry alternating current. A transmission line can be modeled "Fig. 1" with a resistance (R) and inductance (L) The time-dependent transmission line governing the voltages and currents are expressed as follows:

$$\frac{\partial V(z,t)}{\partial z} + L\frac{\partial I(z,t)}{\partial t} + RI(z,t) = 0 \quad (1)$$

$$\frac{\partial I(z,t)}{\partial z} + C\frac{\partial V(z,t)}{\partial t} + GV(z,t) = 0 \quad (2)$$

Where the axis Oz corresponds to the direction of the line V(z,t) is the line voltage at position z along the transmission line axis at time t, I(z,t) is the line current, and L and C are the per unit length inductance (H/m) and capacitance (F/m).

For the simulation of this type of equations, the full-wave techniques are used which are finite difference time domain method (FDTD); the method of moments (MoM); and the finite element method (FEM). There is another class of numerical methods for solving these equations and are called asymptotic techniques. These methods require approximations developed. These methods require approximations developed. They are applied in physical optics (PO), geometrical optics (GO) and uniform theory of diffraction (UTD). It is a field of study in itself. These methods are very powerful, but in the case of our problem, the underlying approximations of the physical limit their use. Generally, the accuracy of the numerical methods is related to the discretization (i.e. mesh size). The finer is the mesh, the better is the accuracy of the methods.

The method of moments (MoM) is preferred for the radiation in the frequency domain and distribution problems involving perfectly or highly son and conductive surfaces. If the problem involves non-homogeneous dielectric materials, it is unlikely that the best formulation.

The finite element method (FEM) has been widely used in structural mechanics and thermodynamics. It is the most practical method for device simulation microwave and analysis eigenproblem. His association with MoM can be accurate and efficient for scattering problems involving penetrable electromagnetic media and specialized antenna problems.

The finite difference time domain (FDTD) method is of a similar vintage to the MoM and FEM in electromagnetics. Unlike the FEM, the FDTD method does not use variational

functionals or weighted residuals, it directly approximates the differential operators in (1)-(2) on a grid staggered in time and space. The FDTD method is an *explicit* finite difference approach, i.e. no matrix equation is set up and solved. There are several very good texts on the FDTD method Kunz and Luebbers was the first [6], appearing in 1993, but Taflove's 1995 volume [7] (and his revised version with Hagness) are the standard reference for the FDTD. For wideband systems, the FDTD method is preferred. It is also effective for any problems or radiation diffusion electromagnetic. But it does not give great accuracy.

For that we will adopt the β-schema that was introduced in 1987 by Desideri et al [11] in finite volumes decentered. Several studies show that we can control the degrees of dispersion and diffusion of the schema by the choice of beta parameter. Despite their details these methods produce oscillations near such discontinuities because they are linear greater than 1. These oscillations order schema have serious consequences on the quality of our results, especially since our system is hyperpolic. We will see this flaw in our numerical simulations of the following paragraphs.

To remedy these problems, several studies have looked at first to use schema without oscillations (TVD) [10]. However, these schemas (TVD) proved very small because they do not contain high order schema. Then a deputy nonlinear [10] these schemas to make nonlinear limiters.

Where the original idea of this work is to couple the β-methods of high order with Rung Kutta method. This part will be presented in section 3. We get High order schemas without oscillations. This will prove our numerical simulations of section 4.

In this work, we will present a new fourth-order-beta method for our transmission line (1)-(2). Our paper is organised as follows. In the first section we give the discretization of classical fourth-order for our system (1)-(2). In the third section we present the principle of our compact fourth-order-beta-method adapted to transmission lines. The last section is devoted to numerical results and comparing our code with a classical method of fourth-order.

2 Principle of classical fourth-order Method applied to line transmission

To analyze the transmission line as a whole and to better predict its output, a finite difference method for fourth order compact [9] may be applied. This method has a satisfactory accuracy but, as we will see in the numerical simulations, it also presents oscillations. Now explain the principle:

The first step in obtaining a β-method solution is to set up a regular grid in space and time. The points on this grid can be designated as (zk , tn), where

$$z_k = (k-1)*\Delta z \ , \ t_n = n*\Delta t \quad 1 \le k \le NDZ+1, \ \ 0 \le n \le NDT$$

As has already been noted in previous work, additional points of grid points at halftime and half space are put in place. These additional points can be designated as (zk+1/2, tn+1/2), where:

$$z_{k+1/2} = (k-\frac{1}{2})*\Delta z \ , \ t_{n+1/2} = (n-\frac{1}{2})*\Delta t \quad \text{for } 1 \le k \le NDZ, \ \ 0 \le n \le NDT$$

We shall compute V(z, t) at the points (zk , tn), and I(z, t) at the points (zk+1/2, tn+1/2), i.e. the voltage and currents are computed at offset locations in space and also in time.

For that we will adopt the β-shema that were introduced in 1987 by Desideri and al in finite volumes decentered. Their adaptation to the system (1)-(2) can be written as follows:

2.1 Discretization of the system of equation

Compute of Voltage and Current

- for k=1

$$V_1^{n+1} = (R_s C \frac{\Delta_z}{\Delta_t} + 1)^{-1} \left[(R_s C \frac{\Delta z}{\Delta t} - 1) V_1^n - 2R_s I_1^n + (V_s^{n+1} + V_s^{n+1}) \right] \quad (3)$$

- for k=2

$$V_2^{n+1} = V_1^n + \frac{\Delta_t}{\Delta z C} \left[I_2^n - I_1^n \right] \quad (4)$$

- for k=2:NDZ

$$V_{k+1}^{n+1} = V_k^{n+1} + \frac{\Delta_t}{\Delta_z C} \left[-\frac{\beta}{4} I_{k+2}^n + \frac{1+\beta}{4} I_{k+1}^n - \frac{1+\beta}{4} I_{k-1}^n + \frac{\beta}{4} I_{k-2}^n \right] (5)$$

$$I_{k+1}^{n+1} = I_k^{n+1} + \frac{\Delta_t}{\Delta_z L} \left[-\frac{\beta}{4} V_{k+2}^n + \frac{1+\beta}{4} V_{k+1}^n - \frac{1+\beta}{4} V_{k-1}^n + \frac{\beta}{4} V_{k-2}^n \right] (6)$$

- For k=NDZ+1

$$V_{NDZ+1}^{n+1} = (R_L C \frac{\Delta_z}{\Delta_t} + 1)^{-1} \left[\begin{array}{l} V_{NDZ+1}^n (R_L \, c \frac{\Delta_z}{\Delta_t} - 1) + \\ 2R_L I_{NDZ}^n + (V_L^{n+1} + V_L^n) \end{array} \right] (7)$$

Where n is the time sign, k is the space sign, F=Δt/Δz and 0<b<1 is a fixed parameter.

2.2 Stability and order of the method:

It is shown for example see [7], that when β<1/3 and *F = sqrt (4 (1-3β)* then the beta schema is of order 4 in time and space and that the dispersion is of order 4. Where F=Δt/Δz

3 The proposed fourth order beta method applied to line transmission

In order to eliminate the oscillations that are a result of the hyperbolic characteristic of our system, We have proposed a solved system by coupling the β-method schema with the Runge Kutta method. The formulation is described in the following system.

Compute of voltage

- for k=1

$$V_1^{n+1} = (R_s C\frac{\Delta_z}{\Delta_t}+1)^{-1}\left[(R_s C\frac{\Delta_z}{\Delta_t}-1)V_1^n - 2R_s I_1^n + (V_s^{n+1}+V_s^{n+1})\right] \quad (8)$$

- for k=2

$$V_2^{n+1} = \left[V_2^n - \frac{\Delta_t}{2\Delta_z C}\left(I_3^n - I_1^n\right)\right] \quad (9)$$

- for k=3 : NDZ

$$V_k^{n+1} = V_k^n - \frac{\Delta_t}{4C\Delta_z}\left[-\frac{\beta}{2}I_{k+2}^n + 1 + \beta I_{k+1}^n - 1 + \beta I_{k-1}^n + \frac{\beta}{2}I_{k-2}^n\right] \quad (10)$$

- for k=NDZ+1

$$V_{NDZ+1}^{n+1} = (R_L C\frac{\Delta_z}{\Delta_t}+1)^{-1}\left[(R_s C\frac{\Delta_z}{\Delta_t}-1)V_1^n + 2R_L I_{NDZ}^n + (V_L^{n+1}+V_L^n)\right] \quad (11)$$

Compute of Current

- For k=1

$$I_1^{n+1} = I_1^n - \frac{\Delta_t}{\Delta_z L}\left[V_2^{n+1} - V_1^{n+1}\right] \quad (12)$$

- For k=2

$$I_2^{n+1} = I_2^n - \frac{\Delta_t}{2L\Delta_z}\left[V_3^{n+1} - V_1^{n+1}\right] \quad (13)$$

- For k=3 : NDZ

$$I_k^{n+1} = I_k^n - \frac{\Delta_t}{4\Delta_t L}\left[-\frac{\beta'}{2}V_{k+2}^{n+1} + (1+\beta')V_{k+1}^{n+1} - (\beta'-1)V_{k-1}^{n+1} + \frac{\beta'}{2}V_{k-2}^{n+1}\right] \quad (14)$$

Before turning to the application of the method of the β-method high order in the equations governing our problem we will combine the Rung Kutta expression (15) and (16)

- For k=3 : NDZ

$$V_k^{n+1} = V_k^n - \frac{\Delta_t}{2\Delta_z C}(I_{k+1}^{n+1} - I_{k-1}^{n+1}) \quad (15)$$

- For k=3 : NDZ

$$I_k^{n+1} = I_k^n - \frac{\Delta_t}{2\Delta_z L}(V_{k+1}^{n+1} - V_{K-1}^{n+1}) \quad (16)$$

4 GRID TESTING AND CODE VALIDATION

This section presents several results obtained by the model that we developed. These preliminary results allow us to explore the possibilities of this model shows different simulation results obtained with a transmission line without loss and for a PCB.

For this we consider a line (400 m length) with two conductors. The line at the input is charged by the excitation source (30V) and impedance at the output. The characteristic impedance of the line =50Ω and a velocity ν = 200m/μs. The line is characterized by significant metal loss and dielectric loss (R=G=0). With:

L = 0, 25 μH/m; C = 0.1 μF/m; = 44.46 pF/m. In our model, we use an excitation characterized by Tr=0.1μs and amplitude v=30V.

4.1 Validation result for lossless line discretized with classical b-method Method

We chose as the following example of simulation for values NDZ = 200 and NDT= 4000, the signal appears after 2μs, the result shows oscillations.

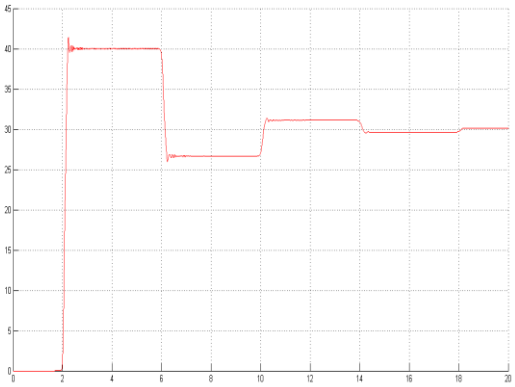


Figure 1. Induced voltage on the line excited by a time step = 0.2 microseconds for NDZ=200 and NDT=2000 using classical β-method

4.2 Validation result for lossless line discretized with our proposed β-method

We choose as the following example of simulation for values NDZ = 200 and NDT= 2000, the signal appears after 2μs, the spatial and time steps don't affect the stability of the signal.

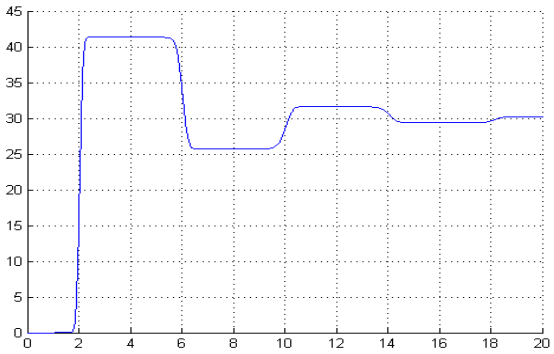


Figure2. Induced voltage on the line excited by a time step = 0.2 microseconds for NDZ=200 and NDT=2000 using our proposed β-method

The following figure will present the comparison of the two schema. We note that our proposed method is more robust and stable and the oscillations are completely neutralized

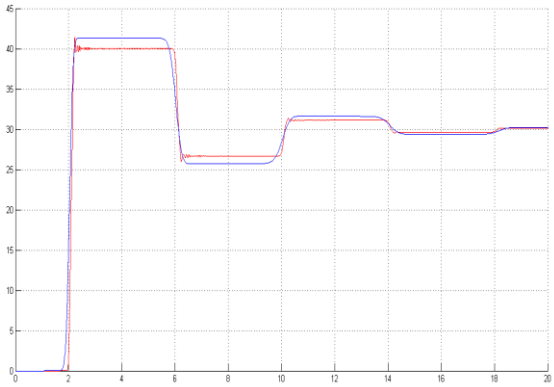


Figure3. Comparison of the two schemas Induced voltage on the line excited by a time step = 0.2 microseconds for NDZ=200 and NDT=2000

5 Conclusions

In this paper a new β-method of fourth order applied to the resolution of homogeneous transmission lines is developed. The conventional methods commonly used for this type of phenomenon simulate like FDTD are limited because they do not contain high order schemas. To obtain high compact order schemas must be used β-method. This method has a satisfactory accuracy but, as we observed in the numerical simulations, it also presents oscillations. Furthermore, various numerical experiments are presented to confirm the accuracy, efficiency and stability of our proposed method. In particular, these simulations show that our new schema is unconditionally stable and fourth-order accurate in space and time.

To demonstrate the interest and compatibility of our simulator with classical β-method, we developed a model in C programming language which simulates our new schema.

6 Acknowledgment

We are grateful to the anonymous referee for a number of corrections and useful suggestions that will improve this article.

7 References

[1] David B. Davidson, ‘’Computational Electromagnetics for RF and Microwave’’, Engineering Second Cambridge university Press, 2011.

[2] J. Clerk Maxwell, ‘’A Treatise on Electricity and Magnetism’’, 3rd ed, vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] Erturk E, Corke TC, ‘’Numerical solutions of 2- D steady incompressible driven cavity flow at high Reynolds

numbers''. International journal for Numerical Methods in fluids 2005; 48:747-774.

[4] Gupta MM, Manohar RP, Stephenson JW. A single cell high order schema for the convection-diffusion equation with variable coefficients. International journal for numerical Methods in fluids 1984;4:641-651.

[5] Spotz WF, Carey GF. ''High order compact schema for the steady streamfunction vorticity equations''. International journal for numerical Methods in engineering 1995; 38:3497-3512

[6] Dennis SC, Hudson JD. ''Compact h4 finite difference approximation to operators of Navier-stokes type''. Journal of computational physics 1989;85:390-416.

[7] Clayton R. Paul , ''Introduction to ElectroMagnetic Compatibility'', (second Edition), Wiley Series in Microwave and Optical Engineering, Kai Chang, Series Editor1992.

[8] Haiyan Xie, Jianguo Wang, Ruyu Fan, and Yinong Liu, ''A Hybrid FDTD-SPICE Method for Transmission Lines Excited by a Nonuniform Incident Wave'' IEEE TRANSACTIONS ON ELECTROMAGNETIC COMPATIBILITY, VOL. 51, NO. 3, (AUGUST 2009).

[9] N. Canouet, L. Fézoui, S. Piperno, ''Méthode volumes finis pour la résolution du système de Maxwell 1D sur des grilles raffinées localement'', INRIA report N°4301, (October 2001).

[10] S. Piperno, ''Schémas TVD d'ordre élevé pour la résolution de l'équation de Burgers'', INRIA report N°4301, - (December 1996)

[11] J. A. Desideri, A. Goudjo, V. Selmin, ''Third order schems for hyperbolic problems'', INRIA report N°607, 1987.

SESSION

SIMULATION, NUMERICAL METHODS, AND APPLICATIONS

Chair(s)

TBA

Interoperability of DEVS and C-BML Simulation Models

Elizabeth Hosang, Gabriel A. Wainer
Department of Systems and Computer Engineering, Carleton University, Ottawa, Ontario, Canada

Abstract - *Simulation Interoperability allows systems that have been developed using different technologies on different operating systems to work together. The Joint Command, Control and Communications Information Exchange Data Model, Coalition Battle Management Language and Military Scenario Definition Language define formatted schemas that can be used to communicate information during a scenario. The Discrete Event System Specification (DEVS) is used to create models whose behavior is defined in response to events in the simulated environment. RISE provides a web-based interface for executing DEVS models. We present an application that uses formatted messaging to interact with a DEVS model running on the RISE server. The purpose of this is to demonstrate that a DEVS model can be executed as part of a larger, web-enabled synthetic environment.*

Keywords: DEVS, RISE, CD++, C-BML.

1 Introduction

Computer simulation can be used to evaluate the behavior of systems, resources and perform planning for large-scale operations. Simulation allows planners to evaluate the behavior of complex systems without deploying and consuming the actual resources required in the scenario. Military organizations use simulation and modeling for mission rehearsal to evaluate planned courses of action. Simulation programs called Computer Generated Force (CGF) systems such as VR Forces and VBS2 [43] are used to allow commanders to develop mission plans and evaluate them without requiring the physical deployment of actual personnel and equipment to the field. These simulations may be used in training of new command staff personnel. From the point of view of the command staff, there needs to be no difference between the physical forces that are deployed in the field and the simulated forces, either when being tasked by command staff or when reporting back on progress or observations. [42].

Military organizations rarely operate in isolation. In order for their systems to interact, standards have evolved that can be used to allow the C2 systems to exchange data [42]. Coalition Battle Management Language (C-BML) is one such standard. It is based on the Joint Command, Control and Communication Information Exchange Data Model (JC3IEDM) which was developed by the MIP program as a basis for exchanging data. C-BML is designed to unambiguously communicate a commander's intent. The Military Scenario Definition Language (MSDL) is a companion schema to C-BML, using the same syntax for initialization of systems prior to an exercise.

Civilian emergency response teams uses simulation for training and planning, although to a lesser extent. Most simulations involve first responder personnel responding to simulated emergencies with actors portraying victims in an emergency. Natural disasters such as floods and fires may require cooperation of military and civilian agencies. Planning for this type of large-scale operation could benefit from the use of M&S currently used by militaries.

In all these scenarios, the use of Web-enabled systems has evolved, bringing together more and more systems, including both military and civilian applications. As Web services become more complex and offer more services, the interfaces they present to users has become more complex. The Representational State Transfer (REST) architecture was introduced as a way to simplify the interactions of Web entities. The use of a RESTful interface on a service simplifies the interface of the service, thus simplifying the logic required in the clients.

The Discrete Event Systems Specification formalism (DEVS) has gained popularity for modeling such complex systems. The RESTful Interoperability Simulation Environment (RISE) allows clients to access simulations using a RESTful interface. It hosts DEVS and Cell-DEVS models, allowing them to be run on behalf of a client. During execution the client can access the simulations using RESTful messages. In this paper we will explore the use of a web-enabled application written in C# accessing a DEVS model created using C++ on a RISE server. This demonstrated the successful interaction of a mix of technologies. Through the use of a scripted scenario, the application demonstrated how to generate initial state data using a structured format, a Report message of activities in the synthetic environment using a structured message format, querying a DEVS model on the RISE server for its response to the Report and parsing the results from the DEVS model to update the state of the modeled entity.

2 Background

High Level Architecture (HLA) is an architecture which allows multiple systems to coordinate. The individual systems, known as federates, coordinate by publishing formatted text messages on a Run Time Infrastructure (RTI), which manages the messages. The Base Object Model (BOM) [8] defines this common base. Each federate defines a Federation Object

Model (FOM) which must be understood by any other federate that needs to know about those objects. Distributed Interactive Simulation (DIS) is an older standard that served the same purpose as HLA. It defines a data delivery architecture with a strictly controlled message format. Instead of sending messages in a text-based format it defines Protocol Data Units (PDU) which specifies message formats in binary for communicating between simulations [1]. Binary data is more efficient to process than XML, but more difficult to understand.

While DIS and HLA are both standards meant to allow Command and Control systems to communicate, they are not immediately compatible. A DIS gateway may be used to allow a DIS model to communicate to an RTI. Alternately, the Real-time Platform Reference Federation Object Model (RPR FOM) organizes attributes and interactions of DIS models into an HLA hierarchy [44].

The Joint Command, Control and Communication Information Exchange Data Model (JC3IEDM) are a logical data model that defines concepts that are common in a C3 environment. The model was developed by the Multinational Interoperability Program (MIP), a consortium of 29 NATO and Non-NATO nations. The goal of the MIP is to promote international interoperability of Command and Control Information Systems at all levels of command [3]. Specialized functional areas such as fire support operations served as sources of requirements for the original development of the data [45]. The intent was to allow commanders to send instructions to units, and allows them to report back observed entities and their status. The specification divides category codes for actions into two types: Action-Tasks, which are instructions for units being tasked by a commander, and Action-Events, which are activities which are observed being performed by entities outside the control of the commander. Location definitions include points, sequences of points, polygonal lines, circles, etc [45].

The Coalition Battle Management Language (C-BML) standard is used to represent battle management doctrine in the Command and Control (C2) environment [1]. It was defined based on the JC3IEDM [2] and uses the category codes defined in that model. The C-BML definition is not a specific schema. Instead it is a set of building blocks that can be used to define custom schemas to be used during an exercise. Objects that are produced during an exercise featuring Command and Control are never updated or deleted. Instead, updates are created. For this reason they can be modeled as web resources. This makes C-BML suitable for implementation using RESTful web services [4]. The Command, Control, Communications, Computing and Intelligence (C4I) Center of Excellence at George Mason University (GMU) has developed components for use with C-BML. They have created a web service that serves as a repository for C-BML messages, including orders, reports and requests [6]. They have also developed the Scripted BML Server (SBML), which is a form of middleware that is intended to allow rapid development of web services as BML evolves [6]. The SBML provides a RESTful interface as well as a SOAP interface [7, 5].

The Military Scenario Definition Language (MSDL) is a standard for specifying the start state of entities in a scenario. It can be used to describe organizational hierarchies, initial deployments, terrain and weather, and plan objectives. It is closely related to C-BML [1]. It reuses the Base Object Model (BOM) SISO standard and JC3IEDM category codes. It has been used to initialize systems in exercises where C-BML is used for communications, such as the NMSG-085 Land Operation demonstration [5].

Data that resides in a fixed field within a record or file is referred to as structured data [46]. Unstructured data is data that does not reside in a traditional row-column [47]. The JC3IEDM, C-BML and MSDL schemas all define structured data. Because the structure and all possible values of the data are well defined this makes it possible for clients of this data to map the data to a format that can be understood by users.

Discrete Event System Simulation (DEVS) is a technique for modeling the behavior of systems as a series of behaviors in response to stimuli. The system is broken down into a set of one or more atomic models. Multiple atomic models can be coupled to form larger models [21]. The CD++ tool is a supports the definition and execution of DEVS models using C++. Each atomic model is defined as a subclass of the Atomic class. Ports are defined as member variables of the atomic models. The methods in the DEVS formal definition, as well as an initialize function, are inherited from the Atomic class and must be overridden in the subclasses.

Representational state transfer (REST) is a style of software architecture for distributed systems. The intent of REST was to provide a simplified protocol using only four of the HTTP methods: PUT, POST, GET and DELETE. REST emphasizes scalability, generic interfaces, and independent deployment of components [9]. The RESTful Interoperability Simulation Environment (RISE) provides a framework for executing DEVS simulations. It provides a RESTful interface that allows the creation, execution and examination of simulations. RISE is a server with namespaces, arranged in a hierarchy. The root of the server is accessed as <machine-URI>/cdpp/sim/workspaces. Under this level workspaces are created for individual users. Under the user workspace further workspaces may be created for each type of DEVS model. Models are executed by POSTing them to the RISE system under an appropriate namespace [30].

3 A RESTful C-BML environment

Discrete Event Simulation has been proven to be a versatile technique for modeling a variety of systems. The RISE

framework provides a RESTful framework for executing DEVS models. This interface provides the opportunity to integrate a DEVS model into a larger synthetic environment via simulation integration techniques such as the use of C-BML messaging. In this section we show how to build a proof of concept system to demonstrate the interoperability of a DEVS model running under RISE. This is done by using an external application to interact with a RISE server to execute a multi-part scenario.

In the Conceptual Architecture for this experiment, a DEVS model would execute in Real Time. Prior to the start of execution of the scenario, a DEVS Bridge would load the DEVS model on the RISE server, and supply the initial state information for the DEVS models. Once the simulation started, C-BML messages from other entities in the larger synthetic environment would be received by the DEVS Bridge, which executes as an independent process and can communicate using web-protocols. The DEVS Bridge would then forward the messages to the DEVS model on the RISE server, which would start execution of the DEVS model. Once the model had completed executing, the DEVS Bridge would detect the output from the DEVS model and publish it to the synthetic environment, either via the C-BML Server or an RTI. For the purposes of this scenario the DEVS model represents Emergency Services. The atomic models are a central Dispatcher which tasks a Police Car/Unit and a Fire Truck/Unit. The Dispatcher receives messages from the external DEVS Bridge, and determines the tasks that the units should perform. The Police and Fire Units determine the time at which they would start the task and passivate. Since the DEVS model uses a simulated clock the output from the DEVS model is generated right away and includes the time of the output. The output is then parsed by the DEVS Bridge and reported to the rest of the larger simulated environment.

The key conceptual entities are shown in the figure below.

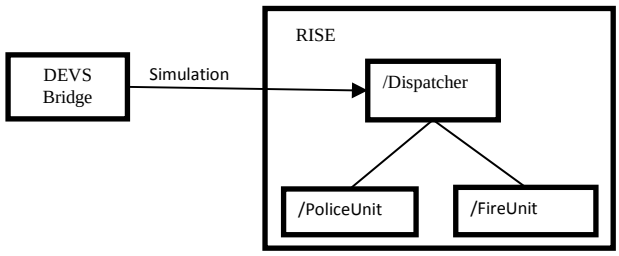


Figure 1 Conceptual Entities in the Emergency Services Dispatch scenario

The Dispatcher entity is responsible for responding to messages from external systems and sending Tasks to the Fire and Police units. However, in implementation it must also bridge the technology gap between the external systems and the DEVS model. Rather than assign the Dispatcher model two distinctly different responsibilities, a Gateway model was added. The Gateway does not participate in the simulated behavior of the system. Instead, its job is to parse incoming messages from the external system, and translate them into the data model used by the Atomic model. It also parses the initial data state information. The Gateway then distributes this information to the DEVS model, which triggers the execution of the DEVS model.

Because the intent of C-BML schema is to unambiguously communicate the commander's intent, it is highly structured data and uses a number of category codes that are used to define the entities in the simulation, their locations, tasks, equipment, etc. However, the category codes which are used are those defined in the JC3IEDM schema. This schema was developed to direct the behavior of military entities. Where civilian entities are represented, they are represented incidentally, as the targets of activities or entities in the environment. Part of the motivation for this experiment is to explore integrating new simulations into a larger synthetic environment. For that reason civilian entities were selected for this experiment. However, there are few category codes in JC3IEDM that can be used to represent civilian agencies, and no category codes that could be used to task the civilian agencies modeled in this scenario. For example, there is no category code for fire-fighting units, nor is there an Action-Task code for "Fight Fire".

The DEVS Bridge application uses a number of classes to manage the data used to generate the Report messages as input to the DEVS model. The data classes are also used to create the RESTful messages that are used to post the model and retrieve the results. The message classes are shown in Figure 2 below.

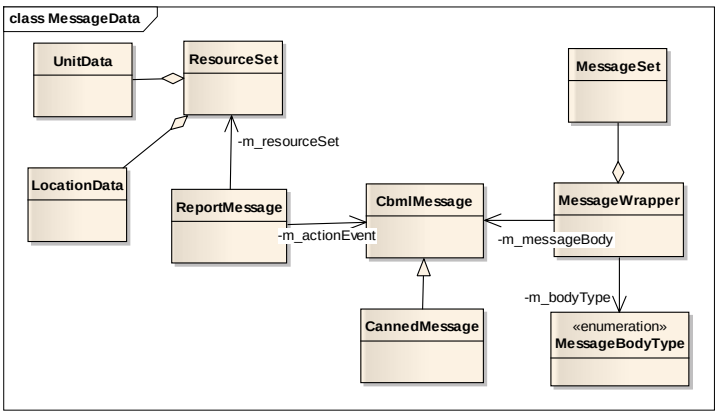


Figure 2 Message Classes

The ResourceSet class contains all of the resources used in a given scenario. The class contains two lists: a list of UnitData instances and a list of LocationData instances. The DEVS Bridge maintains one instance of ResourceSet during processing. All initial state data is generated using the ResourceSet, and when the output from the DEVS model is parsed it is used to update the location and current tasks in the Resource Set. In addition to maintaining the two lists of data, the ResourceSet class also provides a number of methods that

support the use of the class and hide its internals from users of the data. The key responsibilities of the class are to maintain the data, and to support users iterating through the list.

The MessageSet class contains the Report messages used in a given scenario. The class contains a single list of MessageWrapper instances, each of which has a CbmlMessage as its MessageBody attribute. The entire collection can be serialized into or deserialized from XML format using the SerializeUtility. This is done to save the messages to file, and to read them back into the DEVS Bridge from file. The MessageWrapper class holds information for a single message. It can be used for RESTful messages that are sent to the server, or for Report messages which are sent to the executing DEVS model. The responsibilities of the Message Wrapper class include methods for formatting the message, and for generating an HTTP message.

The CbmlMessage contains an ActionEvent and a LocationId. These are used to generate the Report which is sent to the DEVS model. Instead of generating a serialized message based on the attributes of the class, the message string is used as-is. This message type is used to generate the RESTful messages to create, post, execute, and delete the model, and to get the results or error files. The ReportMessage class holds the data that is used to generate all input for the current MA file, including a single MessageWrapper for the Report and the ResourceSet for the initial state data.

The Messages and Resource Set editors display data in a table form. This data is displayed using the standard Data Grid View control supplied as part of the .NET Windows Form toolkit. Data Grid Views provide support for formatting displayed data, re-arranging columns, hiding and showing columns, and selecting rows or columns in the display. Data Grid Views require a DataSource to supply the data in the table.

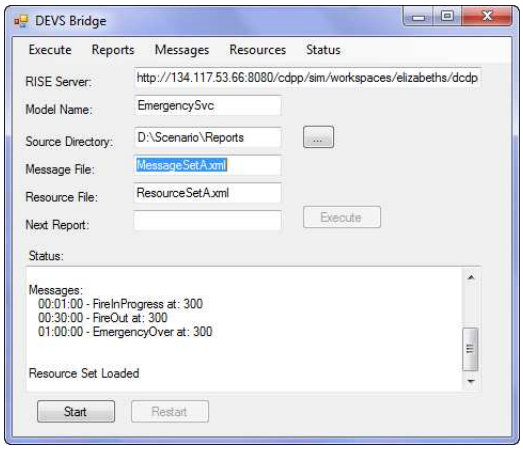


Figure 3 DEVS Bridge Main Form

The DEVS Bridge form has two main display areas: an area describing the scenario information, and a text window where the results of running the tool are displayed. The fields in the top half of the form are:

- RISE Server – address and directory where the model will be created/posted.
- Model Name – name of the model that will be used to create the model on the server. The string "_Model" is appended to the text in the Model Name field. The value of this field may be entered manually or will be set after zipping the model. This is done by selecting the "Model Prep" option under the "Execute" menu. This function zips up the DEVS model files and generates the XML file required to create the model on the server.
- Source Directory – the location of the zipped model that will be uploaded to the server. This value is also set after executing the "Model Prep" function. It can also be edited manually or the browse button to the right of the field ("…") can be used to locate the directory where a zipped model has previously been stored. This value is appended to the RISE Server address on the main DEVS Bridge window.
- Message File – the name of the file containing the loaded message set for the scenario. This may be blank if no file has been loaded or saved. The Execute button is not enabled if no messages have been loaded.
- Resource File – the name of the file containing the loaded set of resources (Locations and Units) that are used in the scenario. This may be blank if no file has been loaded or saved. The Execute button is not enabled if no resources have been loaded.
- Next Report – a description of the next Report message to be processed by the DEVS model.

The Status text pane at the bottom of the window displays output from operations performed by the application. An example of the output in the Status window is shown in Figure 4 below.

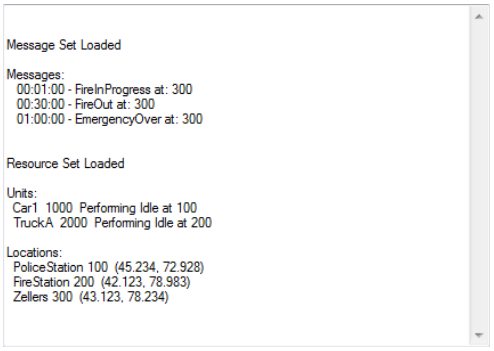


Figure 4 Status window displaying Messages and Resources

The DEVS model provides a simplified model of a civilian emergency response management system. It represents a centralized Dispatch office, such as a 911 operator, which receives reports of an Action Event and assigns an Action Task to multiple types of responders, such as Police and Fire.

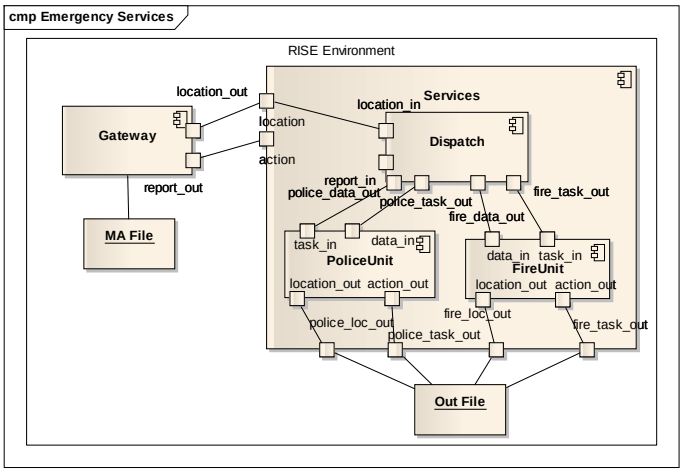


Figure 5 DEVS Atomic models

The Gateway Model represents the interface between the DEVS model and external simulations. It is responsible for receiving data from the DEVS Bridge and parsing it into the internal data format used by the atomic models.

The initFunction method loads the event, units and locations strings from the MA file and passes them to the MessageHolder instance to populate it. After all of the models have initialized, the Gateway model sends the ActionEvent and a Location Id to the Dispatch model.

The Dispatch Model collects the Action Event and Location Id from the Gateway. When it has them both it examines the Action Event. It determines which tasks to assign the Police and Fire units based on the Action Event. It sends the Action Tasks and Location Id to the Police and Fire units. It does not need to load initial state data.

The Police Unit model loads its initial state information from the MessageHolder on the Gateway. When it receives its task and a location id from the Dispatch model it queries the MessageHolder for the details of the Location. If the Task is ReturnToBase, the Police Unit requests the location of its home base. The model then calculates the time required to travel from its current location to its next location.

The Fire Unit model loads its initial state information from the MessageHolder on the Gateway. When it receives its task and a location id from the Dispatch model it queries the MessageHolder for the details of the Location. If the Task is ReturnToBase, the Fire Unit requests the location of its home base. The model then calculates the time required to travel from its current location to its next location.

The Action Task enumeration class corresponds to the Action Task category code defined in the DEVS Bridge project. It is used to send tasks to the units. The enum values are cast as integers when they are sent to the out ports. In the externalFunctions on the receiving classes, the msg.value() is cast to the enumerated type.

The Action Event enumeration class corresponds to the Action Event category code defined in the DEVS Bridge project. It is used to send tasks to the Dispatch class. The enum value is cast as an integer when it is sent to the out port. In the externalFunction on the Dispatch class the msg.value is cast to the enumerated type.

The Location class represents a single location. It stores the unique location id for the location and the latitude and longitude of the location. The UnitRecord class represents the state of a unit. It holds the unique location id for the unit, the unique id for the unit's location, and the current ActionTask of the unit.

The Gateway passivates for a few milliseconds, but for a longer time than the PoliceUnit and FireUnit classes. This is done so that it does not send any output until after the classes have loaded their initial values. When the internalFunction executes the Gateway exports the location and action event data to the Dispatch class. The sequence of behavior is shown in the diagram below.

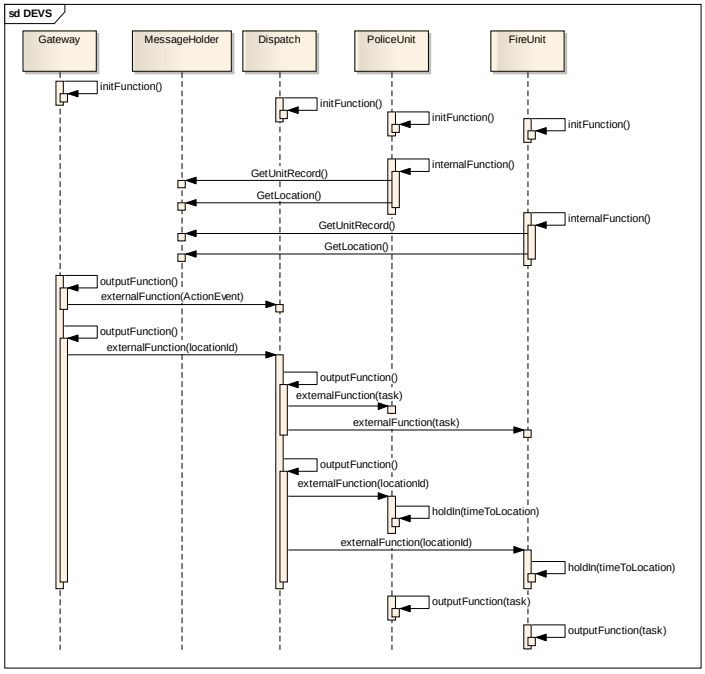


Figure 6 DEVS Messaging Sequence

The main functions are shown in the figure. Atomic Models do not invoke each other directly. The sending model invokes sendOutput and passes in the name of the port, the value of the message, and the time. The externalFunction method is invoked on the receiving model. However, for simplicity the diagram shows the sending model invoking the externalFunction method on the receiving model. At the end of the externalFunction and internalFunction methods the model either calls the passivate or the holdIn method. However, the diagram only shows the holdIn methods on the PoliceUnit and

FireUnit at the ends of the sequences. These methods are shown because they represent the calculation of the time required for the units to arrive at their new locations. When the holdIn method ends the units send their new task and location id information to their output files. The time, task, location and output port information is written to the out file.

The DEVS Bridge was implemented and tested with the DEVS model. The results were successfully parsed and used to generate new messages. The DEVS Bridge was also used to post a pre-existing model to the server and download the results, proving that the tool can also be used as an interface for posting models to the server.

4 Conclusions

We showed the use of a web-enabled application written in C# accessing a DEVS model created using C++ on a RISE server. This demonstrated the successful interaction of a mix of technologies. The application shows that Discrete Event Simulations running under the RISE system can participate in a simulation with other systems that are coordinated by exchanging C-BML messages, using subscriptions and notification to be alerted to the presence of data. The application demonstrates a proof of concept system showing that a DEVS model executing on a RISE server can participate as a part of a larger synthetic environment with the use of a bridge application that acts as a front end to the RISE system.

A custom set of Action Event, Action Task and Unit Type category codes were defined. This issue is one of the reasons why a custom message format was created. This issue cannot easily be addressed. The C-BML schema is based on the JC3IEDM definition. Both are defined by international committees based on the requirements of military organizations and the JC3IEDM model has been in used for many years.
If civilian agencies are to be modeled in simulations that use C-BML for interoperability, they will have to work around the current schema definition. There are several possible options for this:

- Use the "Not Otherwise Specified" / NOS value defined as a value in most category codes.
- Extend the schema and define custom category codes, or extend the existing ones.
- Select existing category codes and map them internally to appropriate values for the simulation. For example, sending an "ATTACK" tasking to a Fire Unit could be interpreted as performing their main function, such as "FightFire". However, this goes against the principle of unambiguously communicating the commander's intent.

5 References

[1] *An Architecture for Demonstrating the Interplay of Emerging SISO Standards*; Blais, Gustavsson, Gustavson, Reichenthal, Fall Simulation Interoperability *Workshop* (*FallSIW*): *10-15*
[2] Simulation Interoperability Standards Organization (SISO) Guide for: Coalition Battle Management Language (C-BML) Phase 1, (SISO-GUIDE-002-2012-DRAFT); 4 April 2012
[3] Multilateral Interoperability Programme (MIP) Implementation Guidance (MIG); Edition 3.1.10, 25 April 2013
[4] *Coalition Battle Management Language (C-BML) Phase I Standard: Trial Use Findings and Next Steps*; Blais, Alstad, NPS Papers and Presentations for SISO Simulation Interoperability Workshop
[5] Lessons learned from NMSG-085 CIG Land Operation demonstration, Gautreau, Khimeche, Remmersmann, Martinet, Muniz, Serrano, Pedersen, Lillesoe, Henderson, Liberg
[6] Maturing Supporting Software for C2-Simulation Interoperation; 2011 15th IEEE/ACM International Symposium on Distributed Simulation and Real Time Applications
[7] Standardization for C2-Simulation Interoperation (MSG-085) Web Site, retrieved September 2013. http://www.cso.nato.int/ACTIVITY_META.asp?ACT=1957,
[8] *Simulation Interoperability Standards Organization (SISO) Guide for Base Object Model (BOM) Use and Implementation SISO-STD-003.1-2006*, SISO Base Object Model Product Development Group, 31 March 2006
[9] Architectural Styles and the Design of Network-based Software Architectures, Roy Fielding, 2000
[10] MIL-STD-2525B, Department Of Defence Interface Standard: Common Warfighting Symbology (30 Jan 1999)
[11] *Scripted Battle Management Language Web Service Version 1.0 Operation and Mapping Description Language*, Pullen, Corner, Singapogu, C4i Center, George Mason University (GMU)
[12] *Adding Publish/Subscribe to the Scripted Battle Management Language Web Service*; Corner, Pullen, Singapogu, Bulusu
[13] *Performance and Usability Enhancements to the Scripted BML Server*, Pullen, Corner, Nicklas, C4I Center, GMU
[14] *Coalition Battle Management Language (C-BML) Phase I Standard: Trial Use Findings and Next Steps*; Blais, Alstad
[15] Lessons learned from NMSG-085 CIG Land Operation demonstration, Gautreau, Khimeche, Remmersmann, Martinet, Muniz, Serrano, Pedersen, Lillesoe, Henderson, Liberg
[16] *Using Coalition Battle Management Language* [160-UK Experiences Using C-BML DS RT 11- RAB.pdf]
[17] Maturing Supporting Software for C2-Simulation Interoperation Pullen, M and Nicklas, L, C4I Center, George Mason University, Fairfax VA, USA
[19] Yong-Ju Lee; Chang-Su Kim, "Building semantic ontologies for RESTful web services," *Computer Information Systems and Industrial Management Applications (CISIM), 2010 International Conference on* , vol., no., pp.383,386, 8-10 Oct. 2010
[20] Bing Pan; Kaifa Liang, "An aggregation search engine based On RESTful Web services and Mashup," *Computer Science and Automation Engineering (CSAE), 2011 IEEE*

International Conference on , vol.3, no., pp.142,146, 10-12 June 2011
[21] *Discrete-Event Modeling and Simulation: A Practitioner's Approach*, Gabriel A. Wainer, CRC Press, 2009
[22] Bowers, F.A., III; Prochnow, D.L., "JTLS-JCATS federation support of emergency response training," *Simulation Conference, 2003. Proceedings of the 2003 Winter* , vol.1, no., pp.1052,1060 Vol.1, 7-10 Dec. 2003
[23] Cowdale, A.C., "Simulation modelling in support of emergency fire-fighting in Norfolk," *Simulation Conference, 2003. Proceedings of the 2003 Winter* , vol.2, no., pp.1707,1710 vol.2, 7-10 Dec. 2003
[24] Brady, T.F., "Emergency management: capability analysis of critical incident response," *Simulation Conference, 2003. Proceedings of the 2003 Winter* , vol.2, no., pp.1863,1867 vol.2, 7-10 Dec. 2003
[25] Chen Tao; Yuan Hong-yong; Yang Rui; Chen Jianguo, "Integration of GIS and Computational Models for Emergency Management," *Intelligent Computation Technology and Automation (ICICTA), 2008 International Conference on* , vol.2, no., pp.255,258, 20-22 Oct. 2008
[26] Yanyan Huang; Jianyu Wang; Rong Jiang; Mei Liu; Lei Dong, "Simulation and Evaluation for the Emergency Management under the Situation of Fatal Disaster," *Multimedia Communications (Mediacom), 2010 International Conference on* , vol., no., pp.258,262, 7-8 Aug. 2010
[27] Aringhieri, R., "An integrated DE and AB simulation model for EMS management," *Health Care Management (WHCM), 2010 IEEE Workshop on* , vol., no., pp.1,6, 18-20 Feb. 2010
[28] Cappelaere, P., S. Frye, D. Mandl. 2009. Flow-enablement of the NASA SensorWeb using RESTful (and secure) workflows. In Proceedings of the 2009 IEEE Aerospace conference. Big Sky, Montana, USA.
[29] Ribault, J.; Wainer, G., "Simulation Processes in the Cloud for Emergency Planning," *Cluster, Cloud and Grid Computing (CCGrid), 2012 12th IEEE/ACM International Symposium on* , vol., no., pp.886,891, 13-16 May 2012
[30] Al-Zoubi, K.; Wainer, G., "Using REST Web-Services Architecture for Distributed Simulation," *Principles of Advanced and Distributed Simulation, 2009. ACM/IEEE/SCS 23rd Workshop on* , pp.114,121, 22-25 June 2009
[31] Rahmat, M.H.; Annamalai, M.; Halim, S.A.; Ahmad, R., "Agent-based modelling and simulation of emergency department re-triage," *Business Engineering and Industrial Applications Colloquium* 219,224, 2013
[32] Zhan Qian; Wei Xianglong, "The research and implementation of a RESTful map mashup service," *Second International Conference on Communication Systems, Networks and Applications,* , pp.401,4032010
[33] Lanthaler, M.; Gutl, C., "Aligning Web Services with the Semantic Web to Create a Global Read-Write Graph of Data," *Web Services (ECOWS), 2011 Ninth IEEE European Conference on* , vol., no., pp.15,22, 14-16 Sept. 2011
[34] Tosic, M.; Manic, M., "A RESTful technique for collaborative learning content transclusion by Wiki-style mashups," *e-Learning in Industrial Electronics (ICELIE), 2011 5th IEEE International Conference on* , vol., no., pp.38,43, 7-10 Nov. 2011
[35] Bing Pan; Kaifa Liang, "An aggregation search engine based On RESTful Web services and Mashup," *Computer Science and Automation Engineering (CSAE), 2011 IEEE International Conference on* , vol.3, no., pp.142,146, 10-12 June 2011
[36] Cheng Bo; Qiao Xiuquan; Wu Budan; Wu Xiaokun; Shi Ruisheng; Chen Junliang, "RESTful Web Service Mashup Based Coal Mine Safety Monitoring and Control Automation with Wireless Sensor Network," *Web Services (ICWS), 2012 IEEE 19th International Conference on* , vol., no., pp.620,622, 24-29 June 2012
[37] Rouached, M.; Baccar, S.; Abid, M., "RESTful Sensor Web Enablement Services for Wireless Sensor Networks," *Services (SERVICES), 2012 IEEE Eighth World Congress on* , vol., no., pp.65,72, 24-29 June 2012
[38] Hessam S. Sarjoughian,Bernard P. Zeigler The Role of Collaborative DEVS Modeler in Federation Development, Proceedings of the 1999 Fall Simulation Interoperability Workshop.
[39] Saurabh Mittal, José L. Risco, Bernard P. Zeigler DEVS-Based Simulation Web Services for Net-Centric T&E, Summer Computer Simulation Conference 2007 (SCSC 2007)
[40] Misagh Tavanpour, Gabriel A. Wainer, Gary Boudreau, and Ronald Casselman. Simulation of Mobile Networks using Discrete Event System Specification Theory, SpringSim CNS, ACM, (2013),
[41] Hypertext Transfer Protocol – Wikipedia. (2014, February 24). http://en.wikipedia.org/wiki/Http.
[42] MacQuarrie, D, Taff, C., Asselstine, B., Hans, R., Reid, S. Simulation-C2 Interoperability Through Data Mediation: the Virtual Command and Control Interface. Proceedings of the 2008 Summer Computer Simulation Conference
[43] Canadian Forces Purchase VBS2 Enterprise License, Bohemia Interactive Simulations (June 22, 2010.) http://www.army-technology.com/contractors/training/bohemia-interactive/press2.html
[44] Guidance, Rationale, and Interoperability Modalities for the Real-time Platform Reference Federation Object Model (RPR FOM), Version 1.0, 10 September 1999
[45] The Joint C3 Information Exchange Data Model (JC3IEDM Main), 14 February 2012
[46] Structured Data, Retrieved February 28, 2012 http://www.webopedia.com/TERM/S/structured_data.html
[47] Unstructured Data, Retrieved February 28, 2012. http://www.webopedia.com/TERM/U/unstructured_data.html,
[48] Simulation Interoperability Standards Organisation (SISO) Standard for Military Scenario Definition Language (MSDL), SISO-STD-007-2008, 14 October 2008.

Reliability of Nano-Scaled Logic Gates Based on Binary Decision Diagrams

Azam Beg[1] **and Ajmal Beg**[2]
[1]College of Information Technology, United Arab Emirates University, Al-Ain, UAE
[2]Cortex Business Solutions, Calgary, AB, Canada

Abstract—*Binary decision diagrams (BDDs) have been useful for synthesis and verification of digital circuits. This paper, for the first time, looks into the reliability of a few logic gates implemented using BDDs. The gates were designed using an advanced CMOS technology node and subject to threshold-voltage variations. The results of the Monte Carlo Spice simulations show that BDD-based gates are significantly more reliable than their conventional CMOS counterparts.*

Keywords: Reliability, static noise margin (SNM), nano-circuits, binary decision diagrams (BDDs), threshold voltage (V_{TH}) variation, Monte Carlo simulations

1 Introduction

1.1 Nano-scaled circuits and their reliability

Nanotechnology is science, engineering, and technology conducted at the scale between 1 to 100 nm [1]. Nanotechnology is an important technology of the 21st century, and is used for producing different products including integrated circuits (ICs). New nanodevices and nanomaterials are expected to continue improving of the quality of human life.

Scaling the semiconductor technology deep into the nano-scale allows novel applications such as wireless sensor networks, wearable computers, implantable devices, etc. The emergence of real-life applications depends to a great extent on the ability to fabricate small, ultra-low power/energy, yet reliable circuits. However, as the transistor/device sizes are aggressively shrunk, the manufacturing of ICs becomes more complex and unavoidably introduces more defects. The nano-scaled devices require small amount of energy for switching, but are highly susceptible to transient failures [2], [3]. Parameter variation is yet another challenge for reliable operation of circuits and systems [4], [5]. This means that the reliability needs to be added to the list of traditional design vectors (power, delay, and area) [6].

As transistors are scaled to nano-dimensions, it becomes harder to limit the variation of threshold voltages (V_{TH}) of the myriads of transistors on a large IC. In MOS transistors, the variation of V_{TH} occurs primarily due to the randomness of the count and the positions of the dopant atoms. The V_{TH}–variation can be estimated using normal distribution with a standard deviation of [7]:

$$\sigma_{V_{TH}} \approx 3.19\times10^{-8}\cdot\frac{t_{ox}\times N_A^{0.4}}{\sqrt{L_{eff}\times W_{eff}}} \quad (1)$$

where t_{ox} is the oxide thickness, N_A is the channel doping, W_{eff} is the effective channel width, and L_{eff} is the effective channel length. Understandably, the V_{TH} fluctuation can be alleviated by increasing either L_{eff} or W_{eff}, or both [8].

1.2 Binary decision diagrams

The high complexity of today's digital ICs necessitates that the design (CAD) tools handle and manipulate the Boolean functions, as efficiently as possible. In the past two decades, the *binary decision diagrams* (BDDs) have gained popularity for efficient synthesis and verification of digital circuits. The BDDs were introduced by Lee [9], and popularized by Akers [10] and Boute [11].

The BDDs are based on the principle of Shannon expansion [12], and can be used to represent any Boolean function $f(x_1, x_2, \ldots x_n)$ [13]. In general, a BDD is a *levelized* acyclic graph of nodes. The nodes on a given level represent one input of the function. There is an additional level (at the bottom) with nodes 0 and 1. All nodes (except the 0- and 1-nodes) have two outgoing edges connecting to the lower level nodes. An upper level node can have multiple incoming inputs from the lower levels. A set of inputs results in a path leading 0- or 1-node to the top-level node(s), i.e., the output(s). A fully-specified input-set takes exactly one complete path in the BDD [14].

The typical BDD-based design procedure consists of the following steps:

1. A truth table of the Boolean function is determined.
2. Outputs with 1's in the truth table are mapped to a BDD.
3. The BDD is *reduced* (while retaining its logical properties).
4. Each node in the BDD is implemented using a multiplexer circuit.

Generally, the BDDs can be simplified using these two techniques:

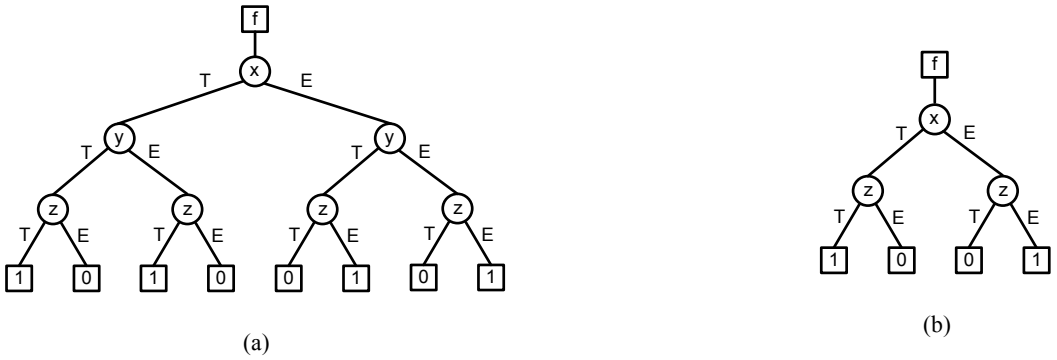


Fig. 1. Example of a BDD (a) in its non-simplified form, and (b) in the reduced form (T=true; E=else/false)

1. *Redundant nodes* are deleted. A redundant node is a node whose two *child* nodes are identical.
2. *Equivalent nodes* are shared. The equivalent nodes are two or more nodes that have the same variable and the same pair of child nodes.

A sample BDD representing the following Boolean function (*sum-of-products*/SOP) is shown in Fig. 1(a).

$$f(x,y,z) = \sum m(0,2,5,7)$$

After simplification, the BDD reduces to the form drawn in Fig. 1(b).

1.3 About this paper

Ours is the first known effort to evaluate the reliability of BDD-based circuits realized with an advanced (22nm CMOS) technology node. This paper presents our findings about the reliability of three basic logic gates implemented with multiplexers. We use *static noise margin* (SNM) as the metric for the gate reliability [15].

2 Related work

The BDDs have been extensively used for designing optimum logic circuits [14] [16]. Simple Boolean functions like adder and comparator has been implemented using BDDs [17]–[19]. More complex computing circuits such as cipher and crypto have also been designed [20].

The evaluation time of a BDD is related to its *path length* and this relationship was investigated and modeled in [21][22].

Low power BDD-based designs have been investigated in [23]–[26]. In addition to power, [25] also considers delays of the circuits. Power and area-efficient design of an asynchronous adder was looked at in [27].

The BDD representations have been used to model the reliability of different systems [28]–[30] but the reliability of BDD-based circuits themselves has not been reported in any research literature.

The reliability of gates using traditionally- and non-conventionally-sized transistors have been looked into [31]–[36]. This paper presents the results of our investigation of the reliability of three basic logic gates implemented using BDDs.

3 Experimental results and analysis

For estimating V_{TH} and its variations, we have used the MOS model from BSIM4 v4.7 level 54 [37]. For modelling the conventional and BDD-based gates, we have used 22 nm PTM HP v2.1 (high-*k*/metal gate and stress effect) transistor models [38], [39]. We have simulated the gates using NGSpice [40].

The BDDs for the three fundamental gates, INV, NAND2, and NOR2 are shown in Figs. 2 (a)–(c), and the corresponding multiplexer-based implementations are shown in Figs. 2 (d)–(f). The multiplexers are built using the well-known *compound-AOI*22 configuration of Fig. 3. The transistor dimensions in all the experiments are fixed at: $L_{nMOS} = L_{pMOS} = L_{min}$ = 22 nm, W_{pMOS} = 88 nm, and W_{nMOS} = 44 nm. The supply voltage (V_{DD}) is set at 0.8V (i.e., the nominal voltage specified in 22nm PTM HP models [38], [39]).

We used the SNM as a measure of reliable operation of a logic gate. In order to measure the SNM, we used the test setup of Fig. 4(a) for the INV, and the setup of Fig. 4(b) for both NAND2 and NOR2. Understandably, for the INV, a single ramp stimulus is sufficient, while the other gates need different combinations of constant and ramp inputs that result in transitioning of the outputs. For NAND2, the inputs are: 10→11, 01→11, and 00→11; and for NOR2, the inputs are: 00 →01, 00→10, and 00→11.

We use Unix-*shell* and Matlab scripts to find the SNMs. The shell script parses the simulation log files to extract *clean*,

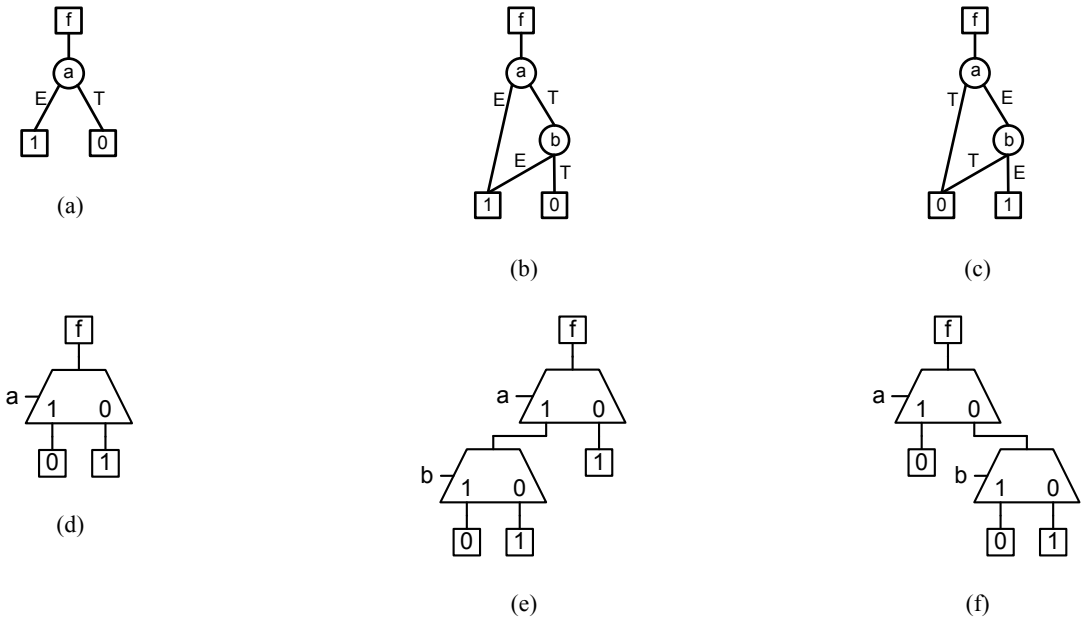


Fig. 2. BDDs for (a) INV, (b) NAND2, and (c) NOR2; multiplexer-based implementation of (d) INV, (e) NAND2, and (f) NOR2

Matlab-friendly text files. The Matlab script automatically determines: the highest allowed input voltage for logic *low* (V_{IL}); the lowest allowed input voltage for logic *high* (V_{IH}); the highest output voltage for logic *low* (V_{OL}); and the lowest output voltage for logic *high* (V_{OH}). These four voltage levels are then used to calculate the SNM as follows:

$$SNM_{low} = V_{IL} - V_{OL} \qquad (2)$$

$$SNM_{high} = V_{OH} - V_{IH} \qquad (3)$$

$$SNM = \min (SNM_{low}, SNM_{high}). \qquad (4)$$

An INV has a single value of SNM; see the *voltage transfer curves* (VTCs) of Fig. 5(a). In comparison, NAND2 and NOR2 have three different sets of VTCs, as shown in Fig. 5(b), resulting in three different SNMs – we choose the worst/least of the three values as the gate's SNM.

In this paper, we have selected 20% (a common threshold) as the minimum acceptable SNM. Stated differently, a gate has *failed*, if its SNM is less than 20% of V_{DD}.

We conducted 1000 Monte Carlo simulations for each of three gates (INV, NAND2, and NOR2) in two different configurations (in other words, a total of 6000 simulations): (1) the normal/conventional CMOS design [41], and (2) BDD-based design (Figs. 2(d)–(f)). The V_{TH}'s of the transistors in the gates (of both types) were varied according to equation (1).

A numerical comparison of the SNM variations for the three gates in the normal- and the BDD-implementations is shown in Tables 1 and 2. The histograms of the SNMs of the gates are shown in Figs. 6–8.

From Table 1, we observe that BDD-based gates have consistently lower failures than the normal ones. The normal INV has 6% failures as compared to none for the BDD version. NAND2's failures rates are 25% and 2.7% for the normal and the BDD-types, respectively. Normal NOR2 fails 17.5% of the time, while the BDD-version fails in just 0.2% of the simulations. These statistics show that a BDD-based gate is appreciably more reliable than a normal one. Additionally, the INV due to its simpler construction has lower failure rates than the two-input gates. Both types of NOR2s have higher reliability than NAND2s. Besides the structural differences between the NAND2 and NOR2, the reliability of NAND2 and NOR2 also depends on the input vectors [42].

Table 2 summarizes the effect of V_{TH} variations on the SNMs of the three gates. INVs of both types, normal and BDD, have the highest SNMs. NOR2s have higher average-SNMs (5-7%) than NAND2s. The BDD-based INV has 71% higher average-SNM than the normal one, whereas NAND2 and NOR2 have 85% and 83% higher SNMs, respectively. The standard deviations for SNMs of normal and BDD-gates are quite comparable. BDD-NAND2 shows higher *spread* than the normal one, while NOR2 is just the opposite.

The results of our experiments show that the BDD-based gates can be more robust and reliable alternatives to the conventional gates. However, it is obvious that the BDD-manifestations of the three gates occupy more area than the conventional ones. One also has to compare the power and the delay of the BDD and the normal gates – a focus of our current work.

4 Conclusions

We have explored the reliabilities (measured in terms of SNM) of BDD-based gates and compared them to the conventional CMOS gates. The BDD-gates have significantly higher SNMs than the conventional gates.

The higher noise margins come at the higher area cost and possibly the delay and the power – a subject of our continued investigation. Use of different multiplexer circuits for the realization of the BDD-nodes is also being looked into.

The three BDD-based basic gates in this paper have been *handcrafted*. To facilitate reliability studies of other gates/cells and larger circuits, we would need an automated mechanism for creating both the BDD-descriptions and the corresponding multiplexer-based implementations – yet another planned extension to our work.

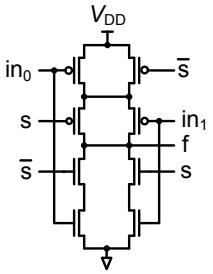


Fig. 3. Schematic of a 2-1 multiplexer

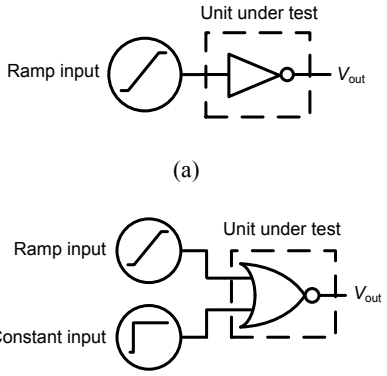


Fig. 4. Test setups for measuring the SNMs of (a) an INV, and (b) a NOR2 gate

5 References

[1] "http://www.nano.gov/nanotech-101/what/definition." .

[2] C. Constantinescu, "Trends and challenges in VLSI circuit reliability," *IEEE Micro*, vol. 23, no. 4, pp. 14–19, 2001.

[3] P. Shivakumar, M. Kistler, S. W. Keckler, D. Burger, and L. Alvisi, "Modeling the effect of technology trends on the soft error rate of combinational logic," in *International Conference on Dependable Systems and Networks*, 2002, pp. 389–398.

[4] SIA, "International Technology Roadmap for Semiconductors (ITRS)," Semiconductor Industry

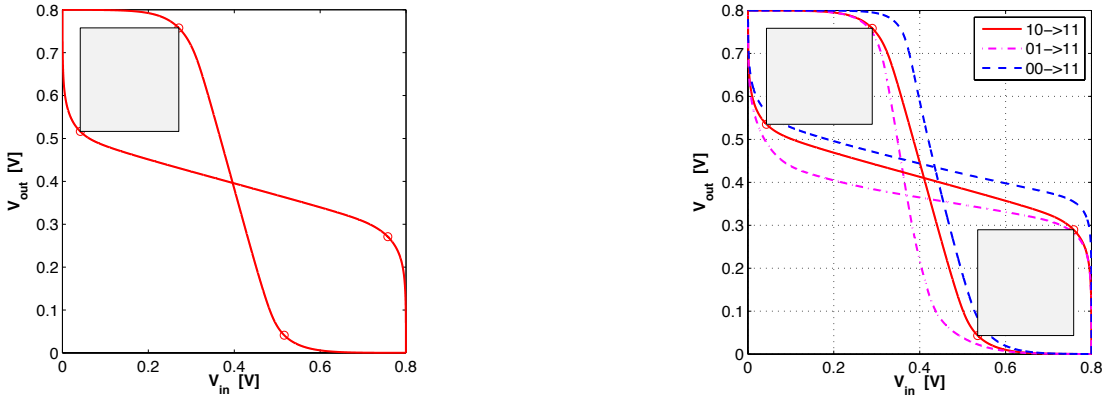


Fig. 5. Voltage transfer curves for measuring the SNM of (a) an INV, and (b) a NAND2,

Association, SEMATECH, Austin, TX, USA, 2009.

[5] Y. Cao and L. T. Clark, "Mapping Statistical Process Variations Toward Circuit Performance Variability: An Analytical Modeling Approach," *IEEE Trans. Comput. Des. Integr. Circuits Syst.*, vol. 26, no. 10, pp. 1866–1873, Oct. 2007.

[6] A. Beg and W. Ibrahim, "Relating reliability to circuit topology," in *IEEE North Eastern Workshop on Circ. & Syst. (NEWCAS'09)*, 2009, pp. 1–4.

[7] A. Asenov, A. R. Brown, J. H. Davies, S. Kaya, and G. Slavcheva, "Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs," *IEEE Trans. Electron Devices*, vol. 50, no. 9, pp. 1837–1852, 2003.

[8] A. Beg, V. Beiu, and W. Ibrahim, "Unconventional transistor sizing for reducing power alleviates threshold voltage variations," in *Int. Semiconductor Conf. (CAS)*, 2012, pp. 429–432.

[9] C.Y. Lee, "Representation of Switching Circuits by Binary-Decision Programs," *Bell Syst. Tech. J.*, vol. 38, pp. 985–999, 1959.

[10] Sheldon B. Akers, "Binary Decision Diagrams," *IEEE Trans. Comput.*, vol. 27, no. 6, pp. 509–516, 1978.

[11] R. T. Boute, "The binary decision machine as programmable controller," *Euromicro Newsl.*, vol. 2, no. 1, pp. 16–22, Jan. 1976.

[12] C. E. Shannon, "A symbolic analysis of relay and switching circuits," *Trans. Am. Inst. Electr. Eng.*, vol. 57, no. 12, pp. 713–723, Dec. 1938.

[13] R. E. Bryant, "Graph-Based Algorithms for Boolean Function Manipulation," *IEEE Trans. Comput.*, vol. C–35, no. 8, pp. 677–691, Aug. 1986.

[14] M. Ciesielski, "BDS: a BDD-based logic optimization system," *IEEE Trans. Comput. Des. Integr. Circuits Syst.*, vol. 21, no. 7, pp. 866–876, Jul. 2002.

[15] A. Beg, "Enhancing Static Noise Margin While Reducing Power Consumption."

[16] R. I. Bahar, E. A. Frohm, C. M. Gaona, G. D. Hachtel, E. Macii, A. Pardo, and F. Somenzi, "Algebraic decision diagrams and their applications," in *Proceedings of 1993 International Conference on Computer Aided Design (ICCAD)*, pp. 188–191.

[17] P. R. Panda, B. V. N. Silpa, A. Shrivastava, and K. Gummidipudi, "Power-efficient System Design," 2010.

[18] and M. B. Singh, Preeti, Chetan Gupta, "Power optimization in a 4-bit magnitude comparator circuit using BDD and pre-computation based strategy," *Int. J. Appl. Eng. Res.*, vol. 7, no. 11, pp. 1482–1485, 2012.

[19] M. Bansal and A. Agarwal, "Ordering and reduction of BDDs for multi-input adders using evolutionary algorithm," in *2013 International Conference on Advanced Electronic Systems (ICAES)*, 2013, pp. 127–130.

[20] S. Morioka and A. Satoh, "A 10-Gbps full-AES crypto design with a twisted BDD S-Box architecture," *IEEE Trans. Very Large Scale Integr. Syst.*, vol. 12, no. 7, pp. 686–691, Jul. 2004.

[21] A. Beg and P. W. C. Prasad, "Prediction of area and length complexity measures for binary decision diagrams," *Expert Syst. With Appl.*, vol. 37, no. 4, pp. 2864–2873, 2010.

[22] P. W. C. Prasad, A. Assi, and A. Beg, "Binary decision diagrams and neural networks," *J. Supercomput.*, vol. 39, no. 3, pp. 301–320, 2007.

[23] P. Lindgren, M. Kerttu, M. Thornton, and R. Drechsler, "Low power optimization technique for BDD mapped circuits," in *Asia and South Pacific Design Automation Conference 2001 (ASP-DAC 2001)*, 2001, pp. 615–621.

[24] S. N. Pradhan, G. Paul, A. Pal, and B. B. Bhattacharya, "Power Aware BDD-based Logic Synthesis Using Adiabatic Multiplexers," in *2006 International Conference on Electrical and Computer Engineering*, 2006, pp. 149–152.

[25] G. P. R. Reddy, J. Ghosh, A. P. C. R. Mandal, and B. B. Bhattacharya, "Power-delay efficient technology mapping of BDD-based circuits using DCVSPG cells," in *2008 3rd International Design and Test Workshop*, 2008, pp. 123–128.

[26] Q. Dinh, D. Chen, and M. D. F. Wong, "BDD-based circuit restructuring for reducing dynamic power," in *2010 IEEE International Conference on Computer Design*, 2010, pp. 548–554.

[27] G. Paul, R. Reddy, C. R. Mandal, and B. B. Bhattacharya, "A BDD-Based Design of an Area-Power Efficient Asynchronous Adder," in *2010 IEEE Computer Society Annual Symposium on VLSI*, 2010, pp. 29–34.

[28] K. S. Trivedi, "A BDD-based algorithm for reliability analysis of phased-mission systems," *IEEE Trans. Reliab.*, vol. 48, no. 1, pp. 50–60, Mar. 1999.

[29] T. Tsuchiya, "A BDD-Based Approach to Reliability Optimal Module Allocation in Networks," in *2012 IEEE 18th Pacific Rim International Symposium on Dependable Computing*, 2012, pp. 121–126.

[30] R. Hu, J. Mi, T. Hu, M. Fu, and P. Yang, "Reliability research for PV system using BDD-based fault tree analysis," in *2013 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering (QR2MSE)*, 2013, pp. 359–363.

[31] A. Beg and W. Ibrahim, "On teaching circuit reliability," in *2008 38th Annual Frontiers in Education Conference*, 2008, pp. T3H–12–T3H–17.

[32] A. Beg, "Reviewing high-level estimation of reliability of nanometric digital circuits," in *9th Int. IEEE Conf. Nanotech. (NANO 2009)*, 2009.

[33] A. Beg and W. Ibrahim, "Relating reliability to circuit topology," in *Circuits and Systems and TAISA Conference, 2009. NEWCAS-TAISA '09. Joint IEEE North-East Workshop on*, 2009.

[34] W. Ibrahim, V. Beiu, and A. Beg, "GREDA: A Fast and More Accurate Gate Reliability EDA Tool," *IEEE Trans. Comput. Des. Integr. Circuits Syst.*, vol. 31, no. 4, pp. 509–521, 2012.

[35] W. Ibrahim, V. Beiu, and A. Beg, "On optimum reliability sizing for complementary metal oxide semiconductor gates," *IEEE Trans. Reliab.*, vol. 61, no. 3, pp. 675–686, 2012.

[36] W. Ibrahim, A. Beg, and H. Amer, "A Bayesian based EDA tool for accurate VLSI reliability evaluations," in *International Conference on Innovations in Information Technology, 2008 (IIT 2008)*, 2008.

[37] B. J. Sheu, D. L. Scharfetter, P.-K. Ko, and M.-C. Jeng, "BSIM: Berkeley short-channel IGFET model for MOS transistors," *IEEE J. Solid-State Circuits*, vol. 22, no. 4, pp. 558–566, Aug. 1987.

[38] W. Zhao and Y. Cao, "New generation of predictive technology modeling for sub-45nm early design exploration," in *7th Int. Symp. Quality Electron. Des.*, 2006, pp. 585–590.

[39] W. Zhao and Y. Cao, "New generation of predictive technology modeling for sub-45nm early design exploration," *IEEE Trans. Electr. Dev.*, vol. 53, pp. 2816–2823, 2006.

[40] "NGSpice – Mixed mode – Mixed level circuit simulator," *ngspice.sourceforge.net*, 2013. [Online]. Available: http://ngspice.sourceforge.net/.

[41] J. M. Rabaey, A. P. Chandrakasan, and B. Nikolić, *Digital integrated circuits: A design perspective*, 2nd ed. Prentice Hall, NJ, USA, 2003.

Table 1. Number of failures (defined as SNM < 0.2 x V_{DD}) for normal and BDD-based gates. (Number of simulations for each gate = 1000)

Gate type	Normal	BDD
INV	60	0
NAND2	250	27
NOR2	173	2

Table 2. SNM (volts) for normal and BDD-based gates

Gate type	Normal		BDD	
	Mean	Std. dev.	Mean	Std. dev.
INV	0.2027	0.0251	0.346	0.0242
NAND2	0.1745	0.0332	0.322	0.0924
NOR2	0.1859	0.0419	0.3393	0.028

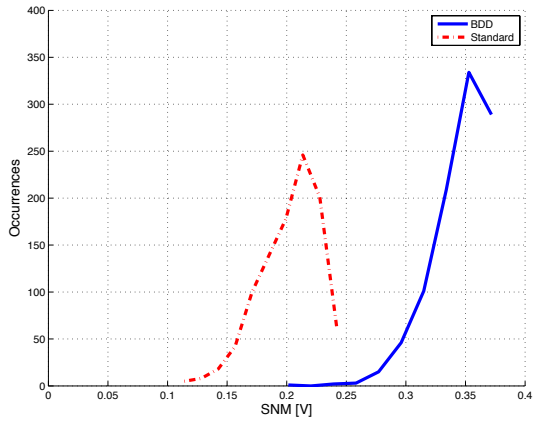


Fig. 6. Histograms of SNM for normal and BDD-based INV's

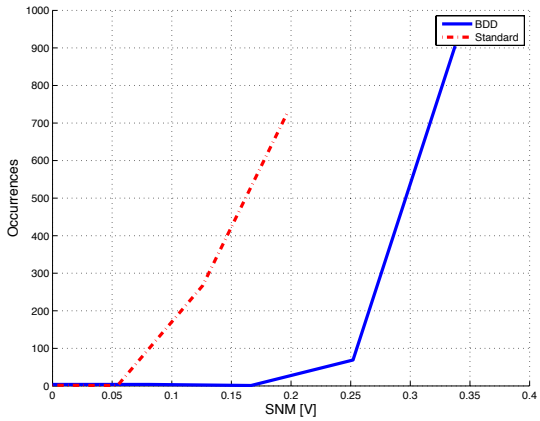


Fig. 7. Histograms of SNM for normal and BDD-based NAND2's

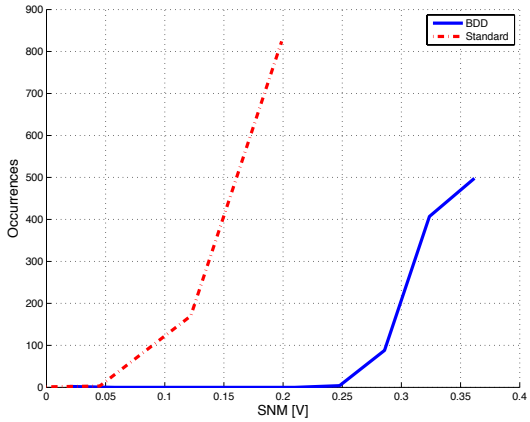


Fig. 8. Histograms of SNM for normal and BDD-based NOR2's

Numerical Algorithms for Solving a Generalized Cancer Chemotherapy Problem

Frank Nani and Mingxian Jin
Department of Mathematics and Computer Science
Fayetteville State University, Fayetteville, North Carolina, USA

Abstract - *In this paper, two elaborate numerical algorithms are presented for solving the Nani- Oğuztöreli functional differential equations associated with cell-cycle specific cancer chemotherapy. The generalized cell-cycle specific cancer chemotherapy model of Nani- Oğuztöreli contains discrete time delays which represent the times that the cancer cells spend in each cell-cycle phase. The model also takes into account that inter-cell cycle phase transition rate constants, recruitment of resting cells from the G0 phase, and effect of chemotherapy drug on cells in each phase. The algorithms utilize a modified version of the Method of Steps algorithms. The constructed numerical schemes can be implemented in C++ or FORTRAN codes in order to generate clinically plausible simulations based on the model equations.*

Keywords: numerical algorithm, time-delays, cell-cycle specific, cancer chemotherapy, modeling

1. Introduction

It has been demonstrated using experimental studies and flow cytofluorometric techniques that proliferating normal and cancerous cells undergo cytokinetic cell transitions through cell-cycle phases S, G2, M, G1, and G0. [1]. Most chemotherapy drugs are also cell-cycle specific. A mathematical method of cancer chemotherapy based on cell-cycle specificity has been presented by Woo-Wiig in [5] and Oguztöreli in 1983 [4]. A generalized model was presented by Nani and Oguztöreli in 1999[2]. The equations used in these models are delay-differential equations which belong to a class of functional differential equations.

In this paper, the Method of Steps will be used to construct algorithms for solving the delay-differential equations of the Nani- Oguztöreli model.

2. Model

2.1 Model Parameters

The cytokinetic parameters used in this generalized model are explicitly defined as follows:

t: normalized time measured in units of multiples of the average cell-cycle time.

$v\,(t)$: the total number of proliferating clonogenic cancer cells that constitute the tumor. In particular, $v\,(t)$ is the sum of the proliferating clonegenic cells in the cell cycle phases: S, G_1, M_1, and G_2.

$x_k\,(t)$: The number of cancer cells in compartment k at time T, where $k \in [1, 2, 3, 4, 5]$.

$x_6\,(t)$: number of nonclogenic cells at time t.

$A_{jk}(t)$: The intercompartmental cytokinetic transition rate.

In particular, it is the rate at time t, at which cells from compartment j migrate to compartment k. The value of $A_{j6}(t)$: denotes the time course of the tumor cell kill produced by the anti-cancer drug, whereas $A_{jk}(t)$: for $k \in [1, 2, 3, 4, 5]$ depicts transition rates between various cell cycle phases.

τ_{jk}: The average cell cycle phase residence times (due to natural cell-cycle processing and repair of damage done by the drug) for the cells in compartment j before migrating to compartment k. The value of τ_{j6} denotes the time required by the cancer cell to lose its clonogenicity and die due to the lethal action of the anti-cancer drug. The values for τ_{jk} for $k \in [1, 2, 3, 4, 5]$ denotes the average residence times expressed mathematically as $\tau_{jk} = \tau_{j\hat{k}} + \sigma_{jk}$ $j \in [1, \ldots, 6]$, $k \in [1, \ldots, 5]$ where $\tau_{j\hat{k}}$ is the average residence time due to natural cell cycle processing only, and σ_{jk} is the drug induced progression delay.

B_{i6}, r_{i6}: Respectively depict the rate and time course associated with cell death due to natural events.

2.2 Model Diagram

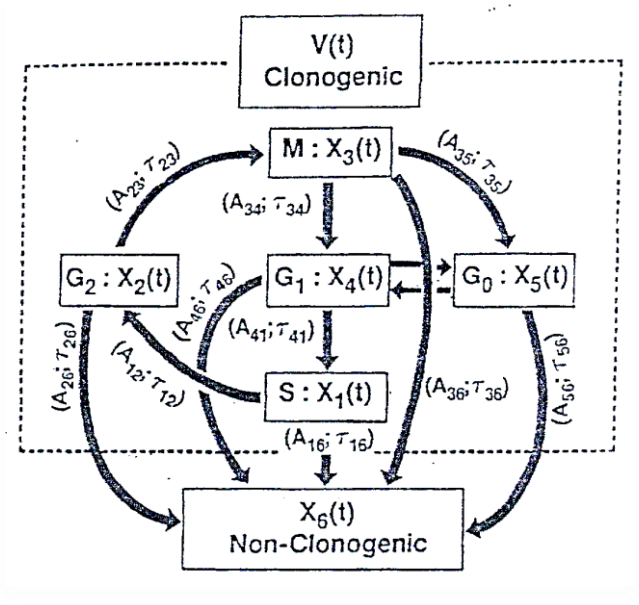


Figure 1. Compartmental tumor cytokinetic model. t: Normalized time, $x_k\,(t)$, k = 1… 5, depicts respectively the number of cells at time t respectively in the S, G_2, M, G_1 and G_0 phases.

2.3 Mathematical Description of the Model

The basic characteristic features of the model are exhibited in Fig. 1. The compartmental description of the model is as follows [2]:

Suppose that at time $t = 0$, there exists x_{k_0} clonogenic cancer cells in the respective k compartments. x_k, k =1, 2, 3, 4, 5 depict respectively cells in : $S, G_2, M\ G_1$, and G_0 phases. Prior to the initial time $t = 0$, these cells have pre-therapeutic distributions, $\emptyset_k\ (t)$ defined over a prescribed, normalized pre-therapeutic interval [-τ, 0].

At time $t > 0$, drug therapy starts and there exist $A_{41}\ (t)x_4$(t -τ_{41}) clonogenic cells in the S-phase, i.e. in compartment #1, which come from compartment #4, i.e. the G_1-phase. After residing in compartment #1 for times$\tau_{12}, r_{16}, \tau_{16}$, these cells respectively move to the cellular compartments #2 and #6.

The $A_{12}(t)x_1$(t - τ_{12}) cells arriving in compartment #2 at t$\geq \tau_{12}$, after residing in compartment #2 for times $\tau_{23}, \tau_{26},$ and r_{26}, move respectively to compartments #3 and #6.

Thus $A_{23}(t)x_2$(t -τ_{23}), cells after $t \geq \tau_{23}$ arrive in compartment #3, which is the mitotic phase where the number of clonogenic tumor cells are duplicated. The newborn clonogenic cancer cells $2A_{34}(t)x_3$(t -τ_{34}) which were produced after a resident time of τ_{34}, more at $t \geq \tau_{34}$ into compartment #4. Also after residence times τ_{35}, τ_{36}, r_{36} more respectively to compartments #5 and #6.

Thus at time $\geq \tau_{34}$, $2A_{34}(t)x_3(t - \tau_{34})$ clonogenic cancer cells arrive in compartment #4, which is the G_1- phase. In particular, the G_1- phase also has recruitment of cells from the G_0 phase, i.e. compartment #5. After the residence times of τ_{41}, $\tau_{45}, \tau_{46}, r_{46}$, cells in compartment #4 move respectively into compartments #1, #5 and #6.

The $A_{45}(t)x_4$(t -τ_{45}) clonogenic tumor cells arrive in compartment #5, which is the non-proliferating clonogenic phase. Also contributions to the G_0 phase come from the compartment #3, i.e. the mitotic phase. After drug induced residence times of $\tau_{54}, \tau_{56}, r_{56,}$, the cells in the compartment #5 move into the respective compartments #4, and #6. This cycle of events is repeated with cyclo-sequential periodicity.

The preceding cytokinetic description can be reformulated into the following mathematical equations which depict tumor cell cytokinetics under pharmacologic drug interactions.

$$\dot{x}_1\,(t) = A_{41}\,(t)A_{41}(t)x_4(t-\tau_{41}) - A_{12}(t)x_1(t-\tau_{12}) - A_{16}\,(t)x_1(t-\tau_{16}) - \mathrm{B}_{16}(t)x_1(t-r_{16})$$

$$\dot{x}_2\,(t) = A_{12}(t)x_1(t-\tau_{12}) - A_{23}(t)x_2(t-\tau_{23}) - A_{26}\,(t)x_2(t-\tau_{26}) - \mathrm{B}_{26}(t)x_2(t-r_{26})$$

$$\dot{x}_2\,(t) = A_{23}(t)x_2(t-\tau_{23}) - A_{34}(t)x_3(t-\tau_{34}) - A_{35}(t)x_3(t-\tau_{35}) - \mathrm{A}_{36}(t)x_3(t-r_{36}) - B_{36}(t)x_3(t-r_{36})$$

$$\dot{x}_4\,(t) = 2A_{34}(t)x_3(t-\tau_{34}) + A_{54}(t)x_5(t-\tau_{54}) - A_{41}(t)x_4(t-\tau_{41}) - \mathrm{A}_{45}(t)x_4(t-\tau_{45}) - A_{46}(t)x_4(t-\tau_{46}) - \mathrm{B}_{46}(t)x_4(t-r_{46})$$

$$\dot{x}_5\,(t) = A_{35}(t)x_3(t-\tau_{35}) + A_{45}(t)x_4(t-\tau_{45}) - A_{54}(t)x_5(t-\tau_{54}) - \mathrm{A}_{56}(t)x_5(t-\tau_{56}) - \mathrm{B}_{56}(t)x_5(t-r_{56})$$

$$\dot{x}_6\,(t) = \textstyle\sum_{i=1}^{6} A_{i6}\,(t)x_i\,(t-\tau_{i6}) + \sum_{i=1}^{6} B_{i6}\,x_i\,(t-r_{i6})$$

where for $t \geq 0, \dot{x}(t)$ denotes the derivative of $x(t)$.

Suppose that the onco-kinetic history of the cancer for a pre-therapeutic normalized time $t \in [-\tau, 0]$ is known. Thus the pre-therapeutic cytokinetic functions:

$$x_k(t) = \emptyset_k\,(t)\ (-\tau \leq t \leq 0\,)$$

$$\tau = \max_{ij} \tau_{ij} \qquad \textit{and} \quad k = 1,2,3,4,5$$

can be specified. In particular, the $\phi(t)$ are given continuous functions such that:

$$\emptyset_k\ (t) \geq 0\ \ k = 1,2,3,4,5$$

$$\emptyset_6\ (t) \equiv 0\ -1\ \leq\ t\ \leq 0$$

and $\phi_k\ (0) =\ x_{k0}\ k = 1,2,3,4,5$ with $x_{60} \equiv 0$, at $t = 0$.

3. Numerical Algorithms for Construction of Solutions

In this section, the mathematical algorithms for construction of solutions to the Cauchy problem for cancer chemotherapy will be presented under various simplifying assumptions. The construction of solution to such functional differential equations (FDE) can be accomplished by various algorithms such as the Euler Polygonal Method, Picard Successive Approximations ([3]) and the Method of Steps ([3]). A modified version of the Method of Steps ([4]) will be used in the subsequent constructions. If the cytokinetic transition coefficients $A_{jk,}\ B_{jk}$ are constant, then it is possible to invoke the theory of the Laplace Transform in constructing some of the solutions to the Cauchy problem.

3.1 Solutions Algorithms

Consider the Cauchy Problem for cancer chemotherapy under the following simplifying assumptions:

$S_{(\mathrm{I})}$: $\quad \tau_{i6} \cong 0\ (i = 1,2,3,4,5)$

This implies that the anti-cancer drug is such that the cancer cells lose their reproductive and clonogenic potential almost instantaneously on contacting the drug. (Drug is hyper-cytotoxic).

$S_{(\mathrm{II})}$: $\quad A_{35} = A_{54} = A_{45}\ \equiv 0.$

This assumption means that there is no recruitment of cancer cells from the non-proliferating G_0 phase.

$S_{(\mathrm{III})}$ $\quad B_{ij} \equiv 0.$ This implies that there is negligible cancer cell loss due to natural mechanisms such as immunological control, necrosis, etc.

Under the assumptions S(I→ III), the cancer chemotherapy equations take the form:

$$\begin{cases} \dot{x}_{1+}\ A_{16}(t)x_1(t) = A_{41}(t)x_4(t-\tau_{41}) - A_{12}(t)x_1(t-\tau_{12}) \\ \dot{x}_{1+}\ A_{26}(t)x_2(t) = A_{12}(t)x_1(t-\tau_{12}) - A_{23}(t)x_2(t-\tau_{23}) \\ \dot{x}_{3+}\ A_{36}(t)x_3(t) = A_{23}(t)X_2(t-\tau_{23}) - A_{34}(t)x_3(t-\tau_{34}) \\ \dot{x}_{4+}\ A_{46}(t)x_4(t) = 2A_{34}(t)x_3(t-\tau_{34}) - A_{41}(t)x_4(t-\tau_{41}) \\ \dot{x}_6 = A_{16}(t)x_1(t) + A_{26}(t)x_2(t) + A_{36}(t) + A_{46}(t)x_4(t) \end{cases} \tag{1}$$

Let

$$\psi_k(t;\tau) = e^{-\int_\tau^t A_{k6(\sigma)d\sigma}} \tag{2}$$

Where $k = 1,2,3,4.$

The preceding system of equations (1) can be treated as a first-order linear inhomogeneous differential equation. Using the transform (2), the following solution are obtained:

$$\begin{aligned} x_1(t) &= x_{10}\psi_1(t;0) + \int_0^t \psi_1(t;\tau)[A_{41}(\tau)x_4(\tau-\tau_{41}) - A_{12}(\tau)x_1(\tau-\tau_{12})]d\tau \\ x_2(t) &= x_{20}\psi_2(t;0) + \int_0^t \psi_2(t;\tau)[A_{12}(\tau)x_1(\tau-\tau_{12}) - A_{23}(\tau)x_2(\tau-\tau_{23})]dr \\ x_3(t) &= x_{30}\psi_3(t;0) + \int_0^t \psi_3(t;\tau)[A_{23}(\tau)x_2(\tau-\tau_{23}) - A_{34}(\tau)x_3(\tau-\tau_{34})]dr \\ x_4(t) &= x_{40}\psi_4(t;0) + \int_0^t \psi_4(t;\tau)[2A_{34}(\tau)x_3(\tau-\tau_{34}) - A_{41}(\tau)x_4(\tau-\tau_{41})]dr \\ x_6(t) &= \int_0^t [A_{16}(\tau)x_1(\tau) + A_{26}(\tau)x_2(\tau) + A_{36}(\tau) + A_{46}(\tau)x_4(\tau)]dr \end{aligned} \tag{3}$$

where

$$\psi_k(t,\tau) = e^{-\int_\tau^t A_{k6}(\sigma)d\sigma}$$

and

$$\begin{cases} x_k(t) = \phi_k(t)\ (-1 \leq t \leq 0, k = 1,2,3,4,6,\} \\ x_{60} \equiv 0\ at - 1 \leq t \leq 0 \end{cases} \tag{4}$$

Starting with the initial condition (4), the system of equations (3) can be solved by successive continuations into the future. The principles of such solution algorithms have been discussed by M.N. Oğuztöreli ([3, 4]).

This process of successive continuations shall be illustrated for the case

$$\tau_{23} < \tau_{34} < \tau_{41} < \tau_{12}.$$

The other cases can be formulated in analogous manner. Initially, the following auxiliary functions are introduced.

$$\left\{ \begin{array}{c} x_{m,1}^{[k]} = x_k(t)\ (m\tau_{12} \leq t \leq m\tau_{12} + \tau_{23}) \\ x_{m,2}^{[k]} = x_k(t)\ (m\tau_{12} + \tau_{23} \leq\ t \leq m\tau_{12} + \tau_{34}) \\ x_{m,3}^{[k]} = x_k(t)\ (m\tau_{12} + \tau_{34} \leq\ t \leq m\tau_{12} + \tau_{41}) \\ x_{m,4}^{[k]} = x_k(t)\ (m\tau_{12} + \tau_{41} \leq\ t \leq (m+1)\tau_{12} \\ k = 1,2,3,4,6, \qquad m = 0,1,2,3,4, \ldots\ldots \\ \\ i.e.\ k \in \mathbb{Z}^+,\ \ m\ \in\ \mathbb{N}_m \end{array} \right\} \tag{5}$$

Let

$$\left\{\begin{array}{c} x_{m,2}^{[k]}(t) = x_{m,1}^{[k]}(t)\ (m\tau_{12} \le t \le m\tau_{12} + \tau_{23}) \\ x_{m,3}^{[k]}(t) = x_{m,2}^{[k]}(t)\ (m\tau_{12} + \tau_{23} \le\ t \le m\tau_{12} + \tau_{34}) \\ x_{m,4}^{[k]}(t) = x_{m,3}^{[k]}(t)\ (m\tau_{12} + \tau_{34} \le\ t \le m\tau_{12} + \tau_{41}) \\ x_{m+1,1}^{[k]}(t) = x_{m,4}^{[k]}(t)\ (m\tau_{12} + \tau_{41} \le\ t \le (m+1)\tau_{12} \end{array}\right\} \quad (6)$$

Next, the function $x_{m,n}^{[k]}(t), k = 1,2,3,4$ shall be constructed. In particular, the function $x_{m,n}^{[6]}(t)$ can be computed in terms of the functions $x_{m,n}^{[k]}(t), k = 1,2,3,4$ and by invoking the expressions specified in (3).

Let M be the largest integer smaller than t/τ_{12}, t > 0. In particular

$$M = \left[\frac{t}{\tau_{12}}\right]$$

and

$$M\tau_{12} \le t < (M+1)\tau_{12}. \quad (7)$$

The solution of the system (3)-(4) can be constructed recursively on the intervals:

$$I_i = [0,\tau_{12}],[\tau_{12},2\tau_{12}],\ldots\ldots,[m\tau_{12},(m+1)\tau_{12}],\ldots,[M\tau_{12},(\mathrm{M}+1)\tau_{12}] \quad (8)$$

In particular, the preceding inequalities imply that

$$[m\tau_{12},(m+1)\tau_{12}] = [m\tau_{12},m\tau_{12}+\tau_{23}] \cup [m\tau_{12}+\tau_{23},m\tau_{12}+\tau_{34}] \cup [m\tau_{12}+\tau_{34},m\tau_{12}+\tau_{41}] \cup [m\tau_{12}+\tau_{41},(m+1)\tau_{12}] \quad (9)$$

The construction of solutions on the interval $[m\tau_{12},(m+1)\tau_{12}]$ will be accomplished by Algorithm A which involves constructing solutions in the interval $[0,\tau_{12}]$ and then Algorithm B which involves establishing recurrence relations for the intervals $[m\tau_{12},(m+1)\tau_{12}]$. On each interval, of (m+1) $\tau_{12}]$, the solutions will be constructed in four steps.

Algorithm A. Solutions in the Interval $0 \le t \le \tau_{12}$

The algorithm for recursive solutions of the system (3)-(4) is presented as follows:

Step 1:

The solutions in the interval $0 \le t \le \tau_{23}$ are given by the recursive relations:

$$x_{0,1}^{[1]}(t) = x_{10}\psi_1(t;0 + \int_0^t \psi_1(t,\tau)\,[A_{41}(\tau)\phi_4(\tau-\tau_{41}) - A_{12}(\tau)\phi_1(\tau-\tau_{12})]d\tau$$

$$x_{0,1}^{[2]}(t) = x_{20}\psi_2(t;0) + \int_0^t \psi_2(t,\tau)\,[A_{12}(\tau)\phi_1(\tau-\tau_{12}) - A_{23}(\tau)\phi_2(\tau-\tau_{23})]d\tau$$

$$x_{0,1}^{[3]}(t) = x_{30}\psi_3(t;0) + \int_0^t \psi_3(t,\tau)\,[A_{23}(\tau)\phi_2(\tau-\tau_{23}) - A_{34}(\tau)\phi_3(\tau-\tau_{34})]d\tau$$

$$x_{0,1}^{[4]}(t) = x_{40}\psi_4(t;0) + \int_0^t \psi_4(t,\tau)\,[2A_{34}(\tau)\phi_3(\tau-\tau_{34}) - A_{41}(\tau)\phi_4(\tau-\tau_{41})]d\tau \quad (10)$$

Step 2.

The solutions in the interval $\tau_{23} \le t \le \tau_{34}$ are expressible in the form:

$$x_{0,2}^{[1]}(t) = x_{10}\psi_1(t;0) + x_{0,1}^{[1]}(\tau_{23}) - x_{10}\psi_1(\tau_{23};0) + \int_{\tau_{23}}^t \psi_1(t,\tau)\,[A_{41}(\tau)\phi_4(\tau-\tau_{41}) - A_{12}(\tau)\phi_1(\tau-\tau_{12})]d\tau$$

$$x_{0,2}^{[2]}(t) = x_{20}\psi_2(t;0) + x_{0,1}^{[2]}(\tau_{23}) - x_{20}\psi_2(\tau_{23};0) + \int_{\tau_{23}}^t \psi_2(t,\tau)\left[A_{12}(\tau)\phi_1(\tau-\tau_{12}) - A_{23}(\tau)x_{0,1}^{[2]}(\tau-\tau_{23})\right]d\tau$$

$$x_{0,2}^{[3]}(t) = x_{30}\psi_3(t;0) + x_{0,1}^{[3]}(\tau_{23}) - x_{30}\psi_3(\tau_{23};0) + \int_{\tau_{23}}^t \psi_3(t,\tau)\left[A_{23}(\tau)x_{0,1}^{[2]}(\tau-\tau_{23}) - A_{34}(\tau)\phi_3(\tau-\tau_{34})\right]d\tau$$

$$x_{0,2}^{[4]}(t) = x_{40}\psi_4(t;0) + x_{0,1}^{[4]}(\tau_{23}) - x_{40}\psi(\tau_{23};0) + \int_{\tau_{23}}^t \psi_4(t,\tau)\,[2A_{34}(\tau)\phi_3(\tau-\tau_{34}) - A_{41}(\tau)\phi_4(\tau-\tau_{41})]d\tau \quad (11)$$

Step 3.

The solutions in the interval $\tau_{34} \le t \le \tau_{41}$ are presented as follows:

$$x_{0,3}^{[1]}(t) = x_{10}\psi_1(t;0) + x_{0,2}^{[1]}(\tau_{34}) - x_{10}\psi_1(\tau_{34};0)$$

$$+\int_{\tau_{34}}^{t} \psi_1(t,\tau)\,[A_{41}(\tau)\phi_4(\tau-\tau_{41}) - A_{12}(\tau)\phi_1(\tau-\tau_{12})]dr$$

$$x_{0,3}^{[2]}(t) = x_{20}\psi_2(t;0) + x_{0,2}^{[2]}(\tau_{34}) - x_{20}\psi_2(\tau_{34};0)$$

$$+\int_{\tau_{34}}^{t} \psi_2(t,\tau)\left[A_{12}(\tau)\phi_1(\tau-\tau_{12}) - A_{23}(\tau)x_{0,1}^{[2]}(\tau-\tau_{34})\right]dr$$

$$x_{0,3}^{[3]}(t) = x_{30}\psi_3(t;0) + x_{0,2}^{[3]}(\tau_{34}) - x_{30}\psi_3(\tau_{34};0)$$

$$+\int_{\tau_{34}}^{t} \psi_3(t,\tau)\left[A_{23}(\tau)x_{0,2}^{[2]}(\tau-\tau_{23}) - A_{34}(\tau)x_{0,1}^{[3]}(\tau-\tau_{34})\right]dr$$

$$x_{0,3}^{[4]}(t) = x_{40}\psi_4(t;0) + x_{0,2}^{[4]}(\tau_{34}) - x_{40}\psi_4(\tau_{34};0)$$

$$+\int_{\tau_{34}}^{t} \psi_4(t,\tau)\left[2A_{34}(\tau)x_{0,1}^{[3]}(\tau-\tau_{34}) - A_{41}(\tau)\phi_4(\tau-\tau_{41})\right]dr \qquad (12)$$

Step 4.

The solutions in the interval $\tau_{41} \le t \le \tau_{12}$ are expressible in the form:

$$x_{0,4}^{[1]}(t) = x_{10}\psi_1(t;0) + x_{0,3}^{[1]}(\tau_{41}) - x_{10}\psi_1(\tau_{41};0)$$

$$+\int_{\tau_{41}}^{t} \psi_1(t,\tau)\left[A_{41}(\tau)x_{0,1}^{[4]}(\tau-\tau_{41}) - A_{12}(\tau)\phi_1(\tau-\tau_{12})\right]dr$$

$$x_{0,4}^{[2]}(t) = x_{20}\psi_2(t;0) + x_{0,3}^{[2]}(\tau_{41}) - x_{20}\psi_2(\tau_{41};0)$$

$$+\int_{\tau_{41}}^{t} \psi_2(t,\tau)\left[A_{12}(\tau)\phi_1(\tau-\tau_{12}) - A_{23}(\tau)x_{0,3}^{[2]}(\tau-\tau_{23})\right]dr$$

$$x_{0,4}^{[3]}(t) = x_{30}\psi_3(t;0) + x_{0,3}^{[3]}(\tau_{41}) - x_{30}\psi_3(\tau_{41};0)$$

$$+\int_{\tau_{41}}^{t} \psi_3(t,\tau)\left[A_{23}(\tau)x_{0,3}^{[2]}(\tau-\tau_{23}) - A_{34}(\tau)x_{0,2}^{[3]}(\tau-\tau_{34})\right]dr$$

$$x_{0,4}^{[4]}(t) = x_{40}\psi_4(t;0) + x_{0,3}^{[4]}(\tau_{41}) - x_{40}\psi(\tau_{41};0)$$

$$+\int_{\tau_{41}}^{t} \psi_4(t,\tau)\left[2A_{34}(\tau)x_{0,2}^{[3]}(\tau-\tau_{34}) - A_{41}(\tau)x_{0,1}^{[4]}(\tau-\tau_{41})\right]dr$$

(13)

<u>Algorithm B.</u> Solutions in the Interval $(m+1)\tau_{12} \le t \le (m+2)\tau_{12}$

Let the function $x_{m,n}^{[k]}(t), k, n = 1,2,3,4. m = 0,1,2 \dots, M-1$ be constructed the algorithm exhibited in A. Then the functions $x_{m+1,n}^{[k]}(t)$ can be constructed in the following four steps.

Step 1.

The solutions in the interval

$$(m+1)\tau_{12} \le t \le (m+1)\tau_{12} + \tau_{23}, m = 0,1,2 \dots, M-1$$

Are given by the recurrence formula:

$$x_{m+1,1}^{[1]}(t) = x_{10}\psi_1(t;0) + x_{m,4}^{[1]}((m+1)\tau_{12}) - x_{10}\psi_1((m+1)\tau_{12};0))$$

$$+\int_{(m+1)\tau_{12}}^{t} \phi\ \psi_1(t,\tau)\left[A_{41}(\tau)x_{m,4}^{[4]}(\tau-\tau_{41}) - A_{12}(\tau)x_{m,1}^{[1]}(\tau-\tau_{12})\right]dr$$

$$x_{m+1,1}^{[2]}(t) = x_{20}\psi_2(t;0) + x_{m,4}^{[2]}((m+1)\tau_{12}) - x_{20}\psi_2((m+1)\tau_{12};0))$$

$$+\int_{(m+1)\tau_{12}}^{t} \psi_2(t,\tau)\left[A_{12}(\tau)x_{m,1}^{[1]}(\tau-\tau_{12}) - A_{23}(\tau)x_{m,3}^{[2]}(\tau-\tau_{23})\right]dr$$

$$x_{m+1,1}^{[3]}(t) = x_{30}\psi_3(t;0) + x_{m,3}^{[3]}((m+1)\tau_{12}) - x_{30}\psi_3((m+1)\tau_{12};0))$$

$$+\int_{(m+1)\tau_{12}}^{t} \psi_3(t,\tau)\left[A_{23}(\tau)x_{m,4}^{[2]}(\tau-\tau_{23}) - A_{34}x_{m,4}^{[3]}(\tau-\tau_{34})\right]dr$$

$$x_{m+1,1}^{[4]}(t) = x_{40}\psi_4(t;0) + x_{m,3}^{[4]}(\tau_{41}) - x_{40}\psi_4((m+1)\tau_{12});0))$$

$$+\int_{(m+1)\tau_{12}}^{t} \psi_4(t,\tau)\left[2A_{34}(\tau)x_{m,4}^{[3]}(\tau-\tau_{34}) - A_{41}(\tau)x_{m,4}^{[4]}(t-\tau_{41})\right]dr$$

(14)

Step 2.

The solutions in the interval

$$(m+1)\tau_{12} + \tau_{23} \le t \le (m+1)\tau_{12} + \tau_{34}, m = 0,1,2 \dots, M-1$$

are presented as follows:

$$x_{m+1,2}^{[1]}(t) = x_{10}\psi_1(t;0) + x_{m+1,1}^{[1]}((m+1)\tau_{12} + \tau_{23}) - x_{10}\psi_1((m+1)\tau_{12} + \tau_{23};0)$$

$$+\int_{(m+1)\tau_{12}+\tau_{23}}^{t} \psi_1(t,\tau)\left[A_{41}(\tau)x_{m,4}^{[4]}(\tau-\tau_{41}) - A_{12}(\tau)x_{m,2}^{[1]}(\tau-\tau_{12})\right]dr$$

$$x_{m+1,2}^{[2]}(t) = x_{20}\psi_2(t;0) + x_{m+1,1}^{[2]}((m+1)\tau_{12}+\tau_{23}) - x_{20}\psi_2((m+1)\tau_{12}+\tau_{23};0)$$

$$+\int_{(m+1)\tau_{12}+\tau_{23}}^{t}\psi_2(t,\tau)\left[A_{12}(\tau)x_{m,2}^{[1]}(\tau-\tau_{12}) - A_{23}(\tau)x_{m+1,1}^{[2]}(\tau-\tau_{23})\right]dr$$

$$x_{m+1,2}^{[3]}(t) = x_{30}\psi_3(t;0) + x_{m+1,1}^{[3]}((m+1)\tau_{12}+\tau_{23}) - x_{30}\psi_3((m+1)\tau_{12}+\tau_{23};0)$$

$$+\int_{(m+1)\tau_{12}+\tau_{23}}^{t}\psi_3(t,\tau)\left[A_{23}(\tau)x_{m+1,1}^{[2]}(\tau-\tau_{23}) - A_{34}(\tau)x_{m,4}^{[3]}(\tau-\tau_{34})\right]dr$$

$$x_{m+1,2}^{[4]}(t) = x_{40}\psi_4(t;0) + x_{m+1,1}^{[4]}((m+1)\tau_{12}+\tau_{23}) - x_{40}\psi_4((m+1)\tau_{12}+\tau_{23};0)$$

$$+\int_{(m+1)\tau_{12}+\tau_{23}}^{t}\psi_4(t,\tau)\left[2A_{34}(\tau)x_{m,4}^{[3]}(\tau-\tau_{34}) - A_{41}(\tau)x_{m,4}^{[4]}(\tau-\tau_{41})\right]dr \quad (15)$$

Step 3.

The solution in the interval

$$(m+1)\tau_{12}+\tau_{34} \le t \le (m+1)\tau_{12}+\tau_{41}, m = 0,1,2\ldots,M-1$$

are given by the recurrence relations:

$$x_{m+1,3}^{[1]}(t) = x_{10}\psi_1(t;0) + x_{m+1,2}^{[1]}((m+1)\tau_{12}+\tau_{34}) - x_{10}\psi_1((m+1)\tau_{12}+\tau_{34};0)$$

$$+\int_{(m+1)\tau_{12}+\tau_{34}}^{t}\psi_1(t,\tau)\left[A_{41}(\tau)x_{m,4}^{[4]}(\tau-\tau_{41}) - A_{12}(\tau)x_{m,3}^{[1]}(\tau-\tau_{12})\right]dr$$

$$x_{m+1,3}^{[2]}(t) = x_{20}\psi_2(t;0) + x_{m+1,2}^{[2]}((m+1)\tau_{12}+\tau_{34}) - x_{20}\psi_2((m+1)\tau_{12}+\tau_{34};0)$$

$$+\int_{(m+1)\tau_{12}+\tau_{34}}^{t}\psi_2(t,\tau)\left[A_{12}(\tau)x_{m,3}^{[1]}(\tau-\tau_{12}) - A_{23}(\tau)x_{m+1,2}^{[2]}(\tau-\tau_{23})\right]dr$$

$$x_{m+1,3}^{[3]}(t) = x_{30}\psi_3(t;0) + x_{m+1,2}^{[3]}((m+1)\tau_{12}+\tau_{34}) - x_{30}\psi_3((m+1)\tau_{12}+\tau_{34};0)$$

$$+\int_{(m+1)\tau_{12}+\tau_{34}}^{t}\psi_3(t,\tau)\left[A_{23}(\tau)x_{m+1,2}^{[2]}(\tau-\tau_{23}) - A_{34}(\tau)x_{m+1,1}^{[3]}(\tau-\tau_{34})\right]dr$$

$$x_{m+1,3}^{[4]}(t) = x_{40}\psi_4(t;0) + x_{m+1,2}^{[4]}((m+1)\tau_{12}+\tau_{34}) - x_{40}\psi_4((m+1)\tau_{12}+\tau_{34};0)$$

$$+\int_{(m+1)\tau_{12}+\tau_{34}}^{t}\psi_4(t,\tau)\left[2A_{34}(\tau)x_{m+1,1}^{[3]}(\tau-\tau_{34}) - A_{41}(\tau)x_{m,4}^{[4]}(\tau-\tau_{41})\right]dr \quad (16)$$

Step 4.

The solutions in the interval

$$(m+1)\tau_{12}+\tau_{41} \le t \le (m+2)\tau_{12}, m = 0,1,2\ldots,M-1$$

are exhibited as follows:

$$x_{m+1,4}^{[1]}(t) = x_{20}\psi_2(t;0) + x_{m+1,3}^{[2]}((m+1)\tau_{12}+\tau_{41}) - x_{10}\psi_1((m+1)\tau_{12}+\tau_{41};0)$$

$$+\int_{(m+1)\tau_{12}+\tau_{41}}^{t}\psi_1(t,\tau)\left[A_{41}(\tau)x_{m+1,1}^{[4]}(\tau-\tau_{41}) - A_{12}(\tau)x_{m,4}^{[1]}(\tau-\tau_{12})\right]dr$$

$$x_{m+1,4}^{[2]}(t) = x_{20}\psi_2(t;0) + x_{m+1,3}^{[2]}((m+1)\tau_{12}+\tau_{41}) - x_{20}\psi_2((m+1)\tau_{12}+\tau_{41};0)$$

$$+\int_{(m+1)\tau_{12}+\tau_{41}}^{t}\psi_2(t,\tau)\left[A_{23}(\tau)x_{m+1,3}^{[2]}(\tau-\tau_{23}) - A_{34}(\tau)x_{m+1,2}^{[3]}(\tau-\tau_{34})\right]dr$$

$$x_{m+1,4}^{[3]}(t) = x_{30}\psi_3(t;0) + x_{m+1,3}^{[3]}((m+1)\tau_{12}+\tau_{41}) - x_{30}\psi_3((m+1)\tau_{12}+\tau_{41};0)$$

$$+\int_{(m+1)\tau_{12}+\tau_{41}}^{t}\psi_3(t,\tau)\left[A_{23}(\tau)x_{m+1,3}^{[2]}(\tau-\tau_{23}) - A_{34}(\tau)x_{m+1,2}^{[3]}(\tau-\tau_{34})\right]dr$$

$$x_{m+1,4}^{[4]}(t) = x_{40}\psi_4(t;0) + x_{m+1,3}^{[4]}((m+1)\tau_{12}+\tau_{41}) - x_{40}\psi_4((m+1)\tau_{12}+\tau_{41};0)$$

$$+\int_{(m+1)\tau_{12}+\tau_{41}}^{t}\psi_4(t,\tau)\left[2A_{34}(\tau)x_{m+1,2}^{[3]}(\tau-\tau_{34}) - A_{41}(\tau)x_{m+1,1}^{[4]}(\tau-\tau_{41})\right]dr \quad (17)$$

Suppose the preceding recurrence relations outlined in A and B have been constructed. Let t be such that $M_{\tau 12} \le t \le (M=1)\tau_{12}$.

Then starting with the relations exhibited in formulas (1) to (13) and invoking successively the recursion relations (14) to (17) for m=0,1,2,3,4…, M-1, the expressions for $x_{M,n}^{[k]}(t)$,

$(k, 0 = 1,2,3,4)$ can be computed. In particular, $x_{M,n}^{[k]}(t)$, according to (5), gives the desired values of $x_k(t)$ at time t. This completes the construction of solutions to the Cauchy problem for cancer chemotherapy exhibited in (3)-(4). The algorithms presented in Algorithms A and B are readily computer programmable using C++ or FORTRAN languages.

4. References

[1] R. Baserga, Cell cycle of cancer cells in man. The cell Cycle and Cancer Cells. The Biochemistry of Disease (1971), Vol. 1 (R. Baserga, ed.), pp.104-105. New York: Marcel Dekker.

[2] F. Nani and M.N. Oğuztöreli, Modeling and simulation of chemotherapy of haematological and gynaecological cancers, IMA Journal of Mathematics Applied in Medicine and Biology (1999) 16, 39-91.

[3] M. N. Oğuztöreli, Time Lag Control Systems(1966), New York: Academic Press.

[4] M. N. Oguztöreli, C. P. Tsokos, and J. Akabutu. A kinetic study of chemotherapy. Applied Math Comput.(1983), 12, 255-300.

[5] K.B. Woo and K.M. Wiig, Quantifying the progress of tumor by bio-marker kinetics. The cell cycle in malignancy and Immunity: Proceeding of the thirteen Annual Hanford Biology Symposium at Richland, Washington, 1-3 October 1973. Technical Information Center, US Energy Research.

A Simulation Approach for Traffic Management

Karmel Manaa
[1]Engineering and Environment Faculty, Northumbria University, Newcastle, England

Abstract- *The traffic congestion problems have been aggravated lately; becoming an annoying daily phenomenon accompanied negative impacts on human health, environment and economy. Conventional traffic management methodologies failed to accommodate the ever-increasing of mobility needs. Hence, the necessity of modern approaches was raised to handle the problem; technology has the power of solving our road congestions dilemmas, it's challengeable but worthwhile. This paper presents a notion of traffic management methodology based on a smart control system that varies the street flow by alterable luminous road lanes and signs, according to circumstances in that time to re-distribute the road usage priority. Three main scenarios were considered and simulated on Witness Software to find the efficiency of the notion, a noticeable enhancement in the traffic flow was found. Finally, a value and risk assessment performed to find the possibilities and challenges of the projects realistic implementation.*

Keywords: Transportation, Simulation, Solar Roadways, alterable lanes.

1 Introduction

Different traffic theories were developed and adopted to handle the traffic elements and conditions, the approaches went on continuous improvement alongside the growing of mobility needs, until a saturation point where roads aren't able to handle the vehicles increment, causing subsequential problems related to human health, environment and economy.

The challenge now is utilizing the available technological researches and employing them in finding feasible and practical traffic enhancement solutions. This paper offers a notion of traffic managing approach based on technological real-time monitoring of traffic using a video camera at the site, the captured scene is analyzed using image processing algorithm, a smart control system varies the street flow by alterable luminous road lanes and signs, according to the analyzed circumstance in that time interval to re-distribute the road usage priority by re-aligning the lanes of the street, using variable e-signs, text panels and solar roadway displaying facility.

2 Philosophical managing perspective

The use of recent technologies and studies will give a hand in organizing road priorities, by distributing infrastructures according to circumstances that's instantly updated [1].

Such a way, a lane is reserved for bike users, and public transports have the priority of the 'No Car Lane', to *terminate the boasted private cars preference*, and to assure a fair allocation between all clients, these priorities remain until a certain level of congestion where it turns to be shared. Similarly, re-aligning each of the road side's width, to give the congested side more width on account of the free side is also a kind of justice satisfying.

3 Solar Roadways

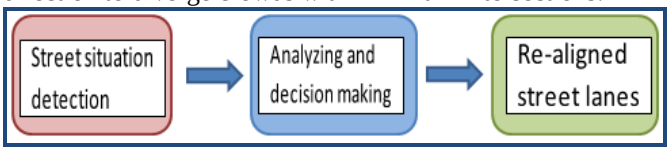
The electrical engineer Scott Brusaw started a notion of making a section road as panels made out of materials that air plane black box was made of, capable of protecting sensitive electronics from the worst airline crashes [2]; these panels could encompass spectacular features as: Housed solar cells to collect energy, creating a road that would pay for itself over time, Added LEDs to provide luminous road lines that light up the road for safer night driving, Added heating elements on the surface of the panels, as the car's rear window defrosting wire to prevent snow/ice accumulation on streets, and continuous possibilities are rolled in.

Figure 1 Conception of Solar Roadway [2].

4 Methodology

Video detection system will be used, a camera at an appropriate placement as an input, recording a video of the road as a sequential frames to be examined. Image processing software (vision processor) will analyze these

frames in order to extract road traffic instantaneous circumstances like LOS, density, mean speed and incidents.
After analyzing the street situation, the road lanes and separators will be re-aligned accordingly, to vary the path width increasing congested side space, or varying traffic flow direction to diverge crowds with minimum intersections.



Video processing using MATLAB is selected to provide a low cost method of traffic conditions detection with no need for using electrical sensors.
Witness traffic simulator will demonstrate the different output actions, to simulate all scenarios as a virtual vision, and evaluate the idea feasibility and efficiency in improving the road situation.

4.1 First Scenario

: A highway with two direction sides, each side with two lanes or above, Fig 2.

Figure 2 First Scenario illustration.

The remarkable output is varying the congested side width on account of the relatively free side, by lending a lane of the free side to the congested one.

Orderly, a methodical approach will be followed, according to data analyzed from the camera (input) or the predictable *transit mass* [3], of the two sides circumstances. If the flow difference between the two sides reaches a threshold level, an output decision will be triggered.
For more than two lanes per side roads, the decision turn to be more complicated pattern, where free sides could lend more than one lane to the congested side.

When the decision of re-aligning lanes taken, the middle yellow line separator will be shifted one step toward the free side, producing a new lane to the congested side as shown below, an informing signs announcing a warning of the process should be displayed on both of the e-board and the Solar Roadway LEDs before sufficient time for drivers to take the appropriate action, a virtual vision of this case is shown in Figure 3.

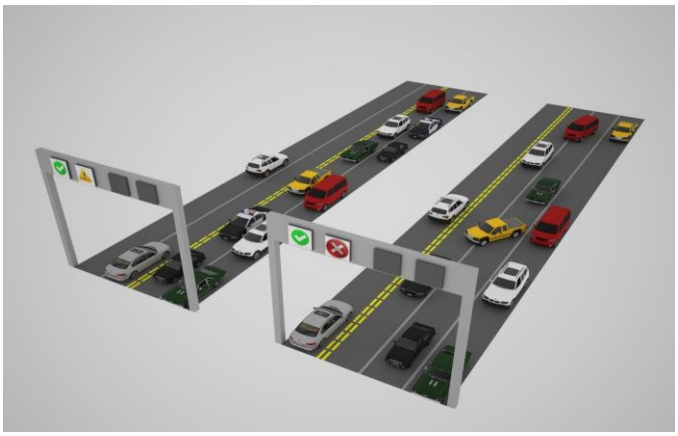


Figure 3 a) road before b) after

4.2 Second Scenario

No Car Lanes are lanes reserved for public transport vehicles, which give public transport priority, under public transportation encouragement policy [4], Fig 4.


Figure 4 Locked 'No Car Lane' in congested road.

While Mully unfold that 'No Car Lanes' are preferable for all motorized modes [4], Cresham is investigating the legality and enforceability of these lanes, in private cars prohibition non-sensibility, especially in congestion cases [5].

As an intermediate solution for both perspectives, same technology could be applied, public transports have the priority until a certain congestion threshold where the lane usage right transforms to all road users, or at least when there is no programmatic buses existence.

The idea is to refer to the road condition and the buses timetable, in order to inspect the 'No Car Lane' priority, the output is represented as varying the lane LED colors to indicate the decision made as following:

- Maroon (already in use): No Car Lanes
- Green: Unlocked for all users
- Red: emergency vehicles only

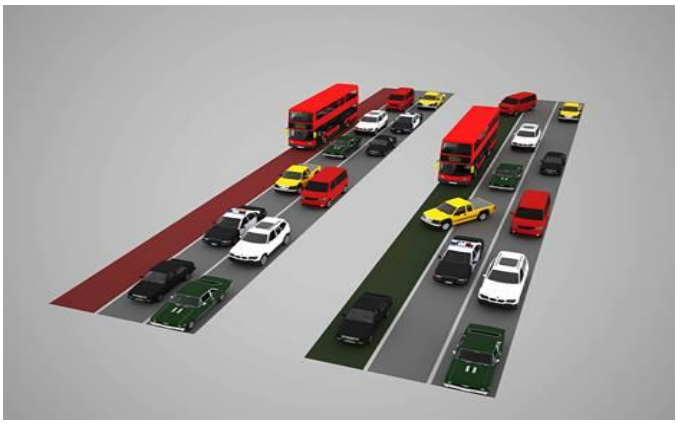


Figure 5 Second scenario visualization.

4.3 Third Scenario

Manipulating the streets flow dynamically, can release predictable congestions around locations that is active during specified intervals of time, such as schools, stadiums, huge organizations, appointed big events, etc., Altering roads from dual-directional to uni-directional (one way traffic) around these locations in course of calming traffic can enhance traffic flow and potentially reduce congestions there, traffic is typically smoother on one way streets since there is no worries about oncoming traffic [6] and the traffic-carrying capacity of the system is greater.

Road directional system remain as it is during the peak-off times, that assures shorter paths around, while over predicted peak-hours (as schools attending and leaving hours, time of games, special events), uni-directional rearrangements performed, to control the flow, by variable e-signs and luminous notes on the Solar Road LEDs, to inform drivers with directions at that time.

The following portrays shows an example of organizing the streets according to that principle, an added e-signs will display direction instructions, as well, luminous roadway LEDs display yellow arrows of the direction.

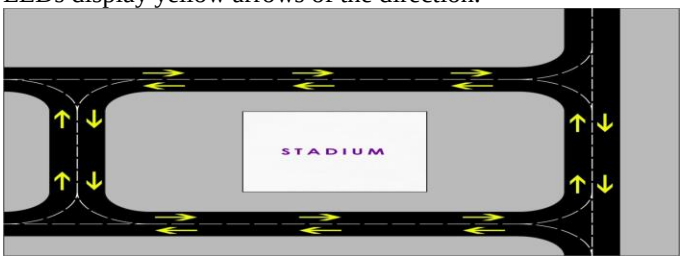


Figure 6 Off-peak hours arrangements around a stadium

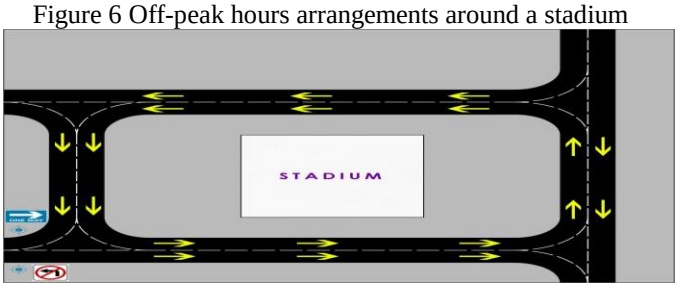


Figure 7 Arrangements at peak times

5 Witness Simulation

5.1 First Scenario

The Witness simulation software is used to simulate the first scenario in two analogous models, the first model is a section of two sides' highway, each side with two lanes, where one of the sides is congested with cars, while the other one is with much lower traffic density.

Witness software is primarily designed to simulate industrial manufacturing, but logical reflections could be assumed to present the transportation elements [7].

These reflections are as follows:

- Vehicles modelled as *Parts*
- Lanes modelled as *Machines*

Thus, each of the 4 lanes is represented as 100 *machines* in a row, each machine represents a fixed length of road, equivalent to an average vehicle length plus some extra space that other cars couldn't enter it (Safe distance), in this model 10 meters is used as the length of each machine. Therefore, each lane length is 1Km (100 machine*10m).

There are two Car *parts* (one for each direction of flow, CarI and CarO), each are sampled from a distribution to make the speed anywhere between 60 and 120 Km/hr.

5.1.1 First Model

Figure 8 First model in simulating the first scenario

In the model shown above, CarI will be Pushed out into LaneI1 and laneI2 at a defined rate higher than the rate in which CarO are Pushed out to LaneO1 and LaneO2, in order to initiate a congested side (RoadI) and a normal side (RoadO). The rate for CarI is settled to be 3600/5000, that denotes 5000 cars entered the road per 1 hour, while the rate for CarO is 3600/1200.

The cars enter the first segment of the lane (the first *machine)* at the pre described rates, then, they move forward by inserting an output rule for the machines as followed:

```
IF N < NQty (LaneI1)                ; while the car in the lane (NQty= NO. of segments)
    PUSH to LaneI1(N+1),LaneI2(N+1)  ; move the car to the next segment
    ELSE                             ; at the end of the road
      PUSH to SHIP
    ENDIF
```

The line PUSH to LaneI1 (N+1), LaneI2 (N+1), moves the car forwardly, if the next segment in the inside lane is blocked by a slower driver, the car will move to the other side; that represents the Overtaking process.

The same output rule applies for the other three lanes by just varying the Lane name accordingly, and without the overtaking move on the outside lanes.

The time that each car spends from entering the road until getting out is recorded to be monitored and analyzed.

A snap shot after 3 minutes of running the model is taken as followed:

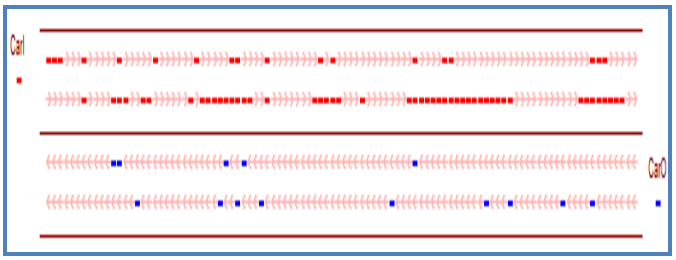


Figure 9 First model while running.

We can see a congested side compared to the other due to the different entering rates of the cars.

5.1.2 Second Model

The second model represents the highway after the free side lends a lane to the congested side, as below:

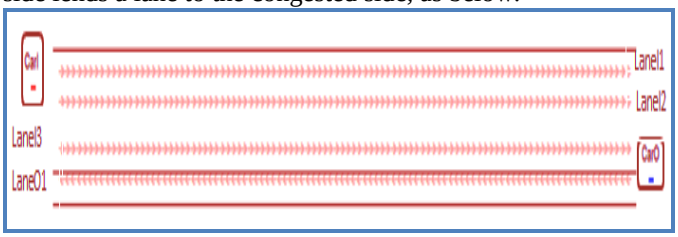


Figure 10 Second model

Everything remains the same as the first model unless the LaneO1 turned to be LaneI3, since it's a lane for opposite side now, while the this side has just one lane (LaneO1).

Conceptually, this lane (LaneI3) is to be reversed in movement, to make it ingoing rather than outgoing.

A snap shot after 3 minutes of running the model is taken as followed:

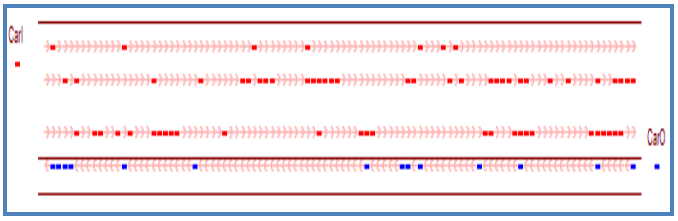


Figure 11 Second model while running.

The times spent by each car from entering the road till reaching the end are recorded. After 1000 cycles of running the model, these values are displayed in separate two histograms, one for RoadI (the congested), and the other one for RoadO, as shown.

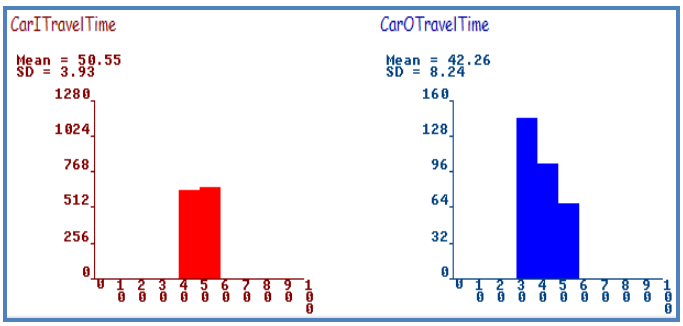


Figure 12 Travel time histograms, first model-first scenario

X-axes of each histogram represent times spent, while Y-axes is the number of cars. The average time spent by cars on RoadI (CarI) is shown to be 50.55s with a 3.93 standard deviation, this means that most cars in this side spend 50.55s to pass, which is a relatively higher value than the 42.26s average spent by cars in the other side (RoadO) for just 1Km section of the road, it's also remarkable to mention the 8.24 standard deviation for RoadO that indicates a freedom of wider range of speeds.

In the second model, where one more lane is added to the congested side, the reading was as followd after 1000 cycles running:

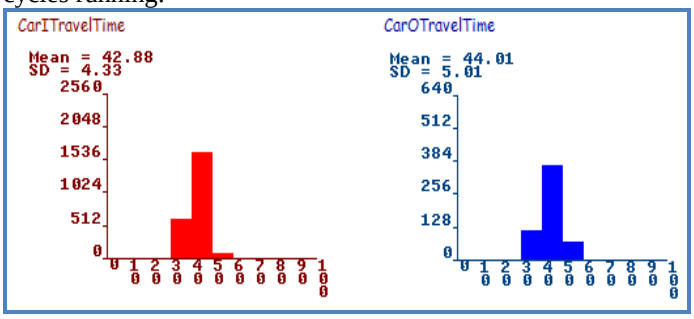


Figure 13 Travel time histograms, second model-first scenario

The average travel time for RoadI was dropped significantly to 42.88s, whilst it rose slightly to 44.01s for RoadO.

Theoretically, each driver will save 7.67s in the 1Km congested side, with only 1.75s increase on the other side, this satisfies the priority organizing tactic from the democratic perspective previously mentioned.

5.2 Second Scenario

The same simulating approach as first scenario is used here; however, the road consists of three lanes in this case, where one of them is a 'No Car Lane' that is used for public buses.

The lanes specifications are the same, with 100 segments (*machines*) each of 10m length.

Two vehicle *Parts* are used, one represents cars, and the other for buses, with speeds boundaries between 60 and 120 Km/hr.

5.2.1 First Model

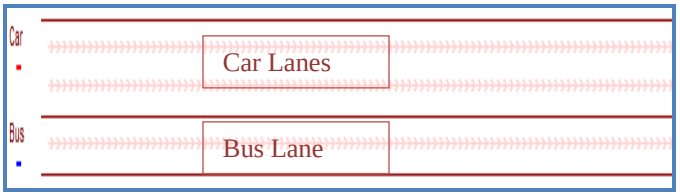


Figure 14 First model in simulating the second scenario

In the above shown model, Car *Part* will be Pushed out into the two Car Lanes at a 3600/5000 rate, while Bus is Pushed at a rate of 3600/240 into the Bus Lane, then, they move forward by the same lanes output rule mentioned previously.

The time which cars and buses take to cross the section is also recorded to be observed and discussed.

After running the model, this snap is captured:

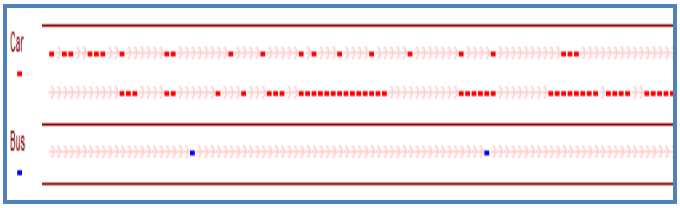


Figure 15 First model while running

5.2.2 First Model

The lanes alignment is maintained, but now, cars are allowed to use the 'No Car Lane'. The model in operation looks like:

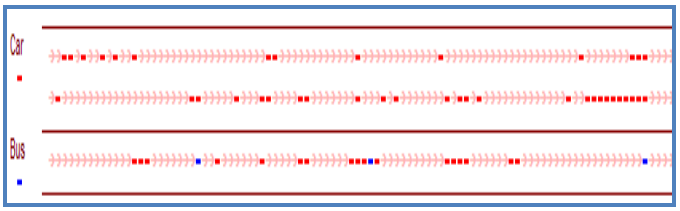


Figure 16 Second model while running

After 1000 cycles, the histogram below shows a 47.65s spent by cars on average to cross the section, whilst 38.1s is spent by buses.

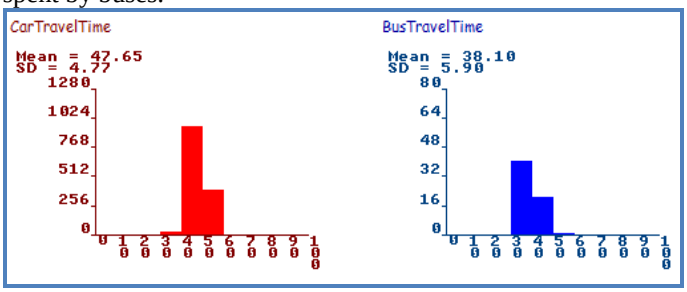


Figure 17 Travel time histograms, first model-second scenario

After allowing cars to use the 'No Car Lane' they are capable of passing the road within 43.03s, this enhancement was on the expense of buses speeds with 5.19 added seconds.

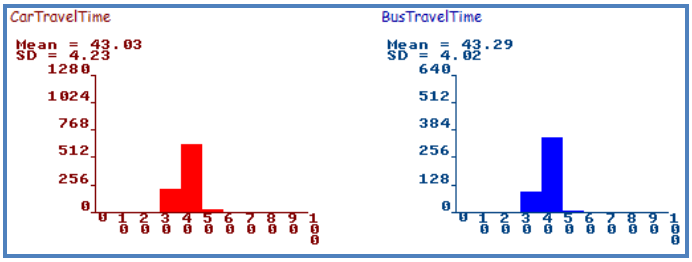


Figure 18 Travel time histograms, second model-second scenario

If we considered the number of vehicles affected in both histograms (Y-axes), it's obvious that the positive gained time affects much more people than the negatively affected (the buses), which mean that the democratic priority of the lane is satisfied rationally.

Generally, these results have strongly upheld the notion that real-time traffic managing could enhance the whole transportation experience in organizing priorities, by distributing infrastructure to enhance the life value with a fair allocation between all clients.

6 Case Study

A case study will be developed here to carry out the previous implementation approach as a realistic scene, the location is at Ponteland Road, Newcastle upon Tyne, the UK, it's a two directional road with double lanes in each side which perfectly matches our case, a preview of the street is as shown:

Figure 19 Preview of Ponteland Road

Practically implementing the principle, the regular asphaltic road is to be replaced with the three layers of the Solar Roadway, additionally, a warning display panel is to be added showing the accessibility of the lane. The realistic vision of that is shown below:

Figure 20 The road after the added features. a) Case one b) Case two

Brusaw claims the Solar Roadway costs of a 1 mile four-lane highway to be $10,000 [8], Ponteland Road of 1.3 Km will cost around £5105 accordingly, and there is an added expenses of variable alert signs, cameras and control system. On the other hand, a typical four-lane road will produce 13.376MWhr per mile, based on sunlight of four hours per day [9], that's equivalent to 10.7 MWhr for our case, a value enough for 350 homes to go off-grid according to American average home consumption [9].

7 Risk Management

7.1 Technical

New technology brings new risks, said Reilly [10], new types of accidents are raised by the new sophisticated technologies [11], since the complexity of the whole processing algorithms with an increasing number of interrelated steps, and the lack of experience due to novelty. Therefore, more complex technology requires specialized resources and support, in an environment of vulnerabilities and risks breeding [12]. A lengthy period of deep planning and investigations of all possible failures or errors that could presence, on both hardware and software tools should be ensured, where this period is capable to increase the proficiency and expertise of all the incorporated human resources.

7.2 Physical

All the video detection, image processing, e-signs and moveable barriers techniques are already in used with quite limited risks. The extreme challengeable issue is concentrated with the solar roadway technology, Dr Carlo Pantano, claims that the biggest obstacle for the project is choosing the right glass for the panels [13], the right balance between the strength of the glass (holding out heavy loads) and it's transparency (allowing UV rays accessing the solar cells), the hardened glass would decrease the photovoltaic cells efficiency, Further, traction between the glass surface and the rubber wheels regardless of weather conditions must be confirmed.

Thus, prober studies of the optimal glass type, strength and texture to be researched, with appropriate testing and examination to develop the required proprieties for the glass surface.

7.3 Financial

The initial cost is the primary facing challenge in this technology, it cost between £4,500 and £7,000 for a single photo-voltaic cell of 4ft X 4ft. that is, the M25 London orbital would cost £93 Billion pounds to construct from simple calculations [13]. Furthermore, manufacturing of the solar cells to create enough cells for M25, Sharp Electronics manufactures would need to work non-stop for 83 years to build cells enough for project completing.

So governments alongside citizens necessitated to re-evaluate in planning terms how our cities would be if such projects performed, to encourage private sector finance and raise number of corporate sponsors, to support the development pursuit, which will lead to greener and more efficient future.

7.4 Cultural

A possibility that conception will go terribly improper as a result of a lack of understanding about methodology in a new culturally outlandish context or venture [14], Without clients (the public) knowing how that way of interacting will be received or perceived by the new prevision cultural risk can ultimate in cultural shock and hence productivity and efficiency losses until getting accustomed.

Handling that, several consciousness phases of media, to enhance the public realization and understanding of the renovation stages, defining the interacting rules and behavior with the new technology elements, hence, smoothing the transition to the new environment.

8 Conclusion

In summary, there is an apparent necessity for modern and originative approaches to manage and control our roads traffic, in order to overcome the raised related dilemmas. This piece of work suggests a new managing technique using the most effective traffic tools deduced in the research; video traffic detection provides load of real time road characteristics

processed by powerful image processing software (MATLAB) to derive the best decision for the instant circumstances by varying the traffic flow dependently.

The innovation of Solar Roadway by Brusaw provides a considerable addendum to the study, from the luminous lanes delineation, flexible managing preferences and investment payback by solar energy saving.

MATLAB image processing code was developed; capable for several street characteristics detection as traffic density, speeds, and incidents, using RGB pixels values variation when cars passed by pre-defined detection points. The MATLAB code detection capability analyzed and showed reasonable and satisfying results compared with the reference video, qualified for providing a feedback of the street condition for the decision making.

Witness software simulates the three suggested scenarios, by referring the traffic elements as a manufacturing elements embedded in the software, and unfolded an enhanced traffic flow after varying the directions stream or priority depending on the case.

Overall, the study proved the real-time managing importance in providing updated information for instantaneous controlling decisions. Moreover, the value and risk assessment demonstrates the idea's feasibility, the risks are surmountable with a great environmental, technical, and economical added value.

REFERENCES

[1] M. Freudendal, Mobility in daily life -between freedom and unfreedom, Denmark: Ashgate, 2009.

[2] Solar Roadways, "Solar ROadways introduction," 2012. [Online]. Available: http://solarroadways.com/intro.shtml. [Accessed 5 July 2013].

[3] UK IGN, "Traffic," 2013. [Online]. Available: http://uk.ign.com/wikis/simcity/Traffic. [Accessed 10 July 2013].

[4] C. Mulley, "No car lanes or bus lanes: which gives public transport the better priority? An evaluation of priority lanes in Tyne and Wear," The Australian Key Centre in Transport and Logistics Management, Sydney , 2011.

[5] P. Cresham, "Legality of 'No Car' lanes," 2011. [Online]. Available: https://www.whatdotheyknow.com/request/legality_of_no_car_lanes. [Accessed 10 July 2013].

[6] R. Rivero, "One Way Signs: Helping Improve Traffic Flow With Single Direction Traffic," 2013. [Online]. Available: http://www.seton.com/blog/2013/05/one-way-signs-helping-improve-traffic-flow-with-single-direction-traffic. [Accessed 12 July 2013].

[7] M. Krempl, "The Traffic Simulation in the Public Transport Terminal," Technical University of Ostrava, Ostrava-Poruba, 2011.

[8] J. Landers , "Solar Roadways: An Idea that Could Change the Energy Landscape," Energy Trend, 2011.

[9] S. Brusaw, "Solar Roadways," 2012. [Online]. Available: http://www.solarroadways.com/faq.shtml. [Accessed 3 Augest 2013].

[10] W. Ashford, "Study finds businesses unready for risks from new technology," 2010. [Online]. Available: http://www.computerweekly.com/news/1280094558/Study-finds-businesses-unready-for-risks-from-new-technology. [Accessed 25 July 2013].

[11] "New technologies, new risks," ICRP, vol. 39, no. 4, pp. 3-5, 2009.

[12] Y. Qian, Y. Fang and J. Gonzalez, "Managing information security risks during new technology adoption," Computers & Security, vol. 31, no. 8, p. 859–869, 2012.

[13] Smartcity, "The Road Ahead for Solar Roadways," 2013. [Online]. Available: http://www.smartcity-magazine.com/2013/solar-roadways-road-surface-solar-panels.html. [Accessed 28 July 2013].

[14] M. Murrell, "Understanding Culturally Driven Operational Risk," 2012. [Online]. Available: http://theforeigninvestmentgroup.com/culturalriskmngt/risk/. [Accessed 30 July 2013].

SESSION

VISUALIZATION, GRAPHICS, TOOLS AND TECHNIQUES

Chair(s)

TBA

Clustered Data Separation via Barycentric Radial Visualization

Adam Russell[1] **and Robert Marceau**[1] **and Franck Kamayou**[1]
and Karen Daniels[1] **and Georges Grinstein**[1]
[1]Computer Science Department, University of Massachusetts Lowell, Lowell, MA, USA

Abstract— *This paper addresses visualizing clusters of multi-dimensional data using barycenters as cluster representatives within the RadViz radial visualization technique. RadViz is a composition of two mappings. Where in this two-stage mapping the cluster barycenters are formed is a key decision. Motivated by the nature of the second mapping, we form cluster barycenters at the end of the first stage, rather than at the start of the first stage. In the second stage we must select an appropriate configuration of dimensional representatives (dimensional anchors). Since this problem is intractable we present a heuristic to: 1) separate clusters and 2) move clusters away from the barycenter of the dimensional anchors. The heuristic uses our prior Voronoi quality assessment technique and our recent observation that circular motion of dimensional anchors confines each data image to an annulus. We demonstrate the benefit of our barycentric approach for a variety of clustered datasets.*

Keywords: Visualization, clustering

1. Introduction

1.1 Motivation, Contribution and Overview

Data visualization is becoming more prominent as the dimensionality and number of records to be explored becomes large. This paper aims to visualize multi-dimensional clustered data, with scalability in mind, for situations in which cluster barycenters (i.e. averages, centers, or centroids) are cluster representatives. We use this framework:

1) Preprocess data using a clustering technique of choice.
2) Preprocess data to help preserve clustering structure under the chosen visualization mapping.
3) Apply the visualization mapping.

The input to this process is a collection of n records in Euclidean space $\mathbb{E}^d$. For step 1, the user selects and applies a clustering technique, and after step 1 the data records are grouped into k disjoint clusters: $C = \{C_1, \ldots, C_k\}$. The literature on clustering is extensive; relevant clustering research is summarized in Section 1.2.

Steps 2 and 3 require choosing and applying a visualization mapping to the clustered data. Among the many existing techniques, radial visualizations stand out as often providing useful views of multi-dimensional data [1]. Radial visualization has its origins in the 19^{th} century [2]. In this paper we use RadViz [3], [4], one type of radial visualization, as the primary example. For a collection of records in the data space, RadViz maps each record to a point within a unit disk in a 2D image space. A dimensional anchor (DA) representing each of d dimensions in the data space is placed on the unit circle. In an interpretation using a spring force analogy with Hooke's law, the image of each of the records is an "equilibrium" point if we imagine a collection of d springs with each spring attached to the point at one end and a DA at the other end. Figure 1 shows a RadViz result for 9D data with 5 clusters. A Voronoi diagram [5] proximity structure is overlayed where each DA and the barycenter of the DAs are the Voronoi sites (Section 1.5). In this example only 4 of the 5 clusters are distinguishable, and all data falls within the region associated with the barycenter of the DAs, where cancellation of DA influences can occur. Our goal is to improve on this type of situation by visually separating clusters and trying to move them away from the barycenter of the DAs. RadViz and related work in radial visualizations are discussed in more detail in Section 1.3.

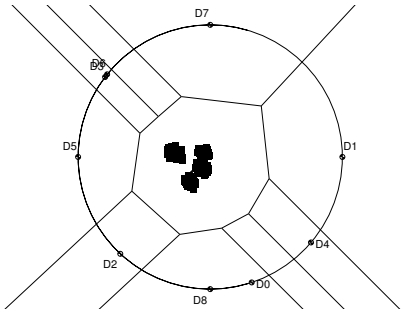


Figure 1: Dimensional anchor placement in RadViz for 1800 records of our 9D synthetic data with 5 clusters (4 visible), with Voronoi regions.

For step 2 in the visualization process, in the RadViz case one would preprocess the data to help prevent losing clustering structure early in the transformation process (see Section 1.3). For scalability reasons, during step 3 we operate on a set of cluster representatives rather than the data images themselves, with one representative for each of the k disjoint clusters: $R = \{R_1, \ldots, R_k\}$. RadViz has a 2-stage visualization pipeline, as it is the composition of two mappings [6] (see Section 1.3). The first stage normalizes the data while the second applies DA weights in the image space. A key design question is: where in the pipeline to form the cluster representatives? We show in this paper that the choice makes a difference (Section 2), and that when

using cluster barycenters as representatives, clustering quality is improved when forming the representatives after the first stage of the transformation and before the second stage rather than forming them in the original data space.

Selecting a DA configuration is another crucial design question in the RadViz visualization, and is equivalent to applying the second part of the RadViz mapping. We explain in Section 2 how the second part of the RadViz mapping guarantees that, regardless of the chosen DA configuration, in the image space the transformed barycenter clusters remain barycenters of the transformed data clusters. This provides strong justification for forming barycentric cluster representatives after the first stage of the pipeline rather than in the original data space. It also provides motivation for working with cluster representatives that are cluster barycenters rather than other types of cluster representatives (see Section 1.2).

For the remainder of step 3, cluster representatives are used in place of the data to select RadViz DA positions with the goal of preserving clustering structure in the visualization. Because we are working here with cluster representatives, many different DA configurations can quickly be examined, allowing better exploration of the search space and hence better preservation of clustering structure.

As discussed in Section 1.4, deciding where to place the DAs on the RadViz circle is a complicated challenge. In Section 3 this paper offers a heuristic for DA placement that supports 2 cluster visualization goals: 1) separate cluster representatives from each other and 2) move cluster representatives away from the (bary)center of the DAs in the RadViz circle to minimize "cancellation of forces" from opposing DAs. Figure 1 highlights the need for these goals. Our heuristic moves cluster representatives away from the barycenter of the DAs using a quality score (see Section 1.5) and our recent results on how data image points move as DAs move (see Section 1.6). Due to our usage of barycenters as cluster representatives and the importance of the barycenter of the DAs in finding a good DA configuration, we refer to our overall approach as barycentric.

Results are presented in Section 4 for a variety of clustered datasets, including the one from Figure 1. There we experimentally show that forming barycentric cluster representatives after the first stage of the RadViz pipeline is better than forming them in the original data space and we demonstrate the benefit of our DA placement heuristic. In this paper we assume that barycenters are appropriate representatives for our data clusters. Future work will attempt to preserve clustering structure for elongated or very irregularly shaped clusters; some preliminary thoughts on treating this difficult issue appear in our discussion of future work in Section 5.

1.2 Clustering

Jain *et al.* [7], [8] provides an overview of various clustering algorithms. It should be noted that clustering is subjective. These algorithms may be divided into several broad categories. Certain issues which need to be considered across several broad categories of clustering algorithms are discussed first. Although several different similarity measures may be used, we will utilize Euclidean distance between points in the 2-dimensional RadViz circle.

The algorithm may operate in either a divisive or an agglomerative mode. In the divisive mode, all data points are initially placed in a single cluster which is then partitioned until a suitable stopping criterion is met. For the agglomerative mode, each data point is placed in a singleton cluster. Clusters are then merged until a stopping criterion is met.

The algorithms may operate by only considering the distance between two data points along a single dimension (monothetic). Once a partitioning has been performed using one dimension, each cluster is partitioned again, using the distance along another dimension. Alternatively, the distance between data points may consider all dimensions (polythetic).

The membership of a data point in a particular cluster may be either "hard" or "fuzzy". The algorithm may be either deterministic of stochastic. For very large data sets an incremental algorithm may be desirable. Also of interest for very large data sets is the selection of a smaller set of representative points. Agarwal *et al.* [9] discuss a "core set" of representatives. Another possible choice of representative is the barycenter (centroid) of a cluster.

The popular k-means partitional algorithm seeks to minimize a squared error (distance from the centroid) function. To be able to employ k-means clustering, the value of k, the number of clusters must be known *a priori*.

Concepts from Support Vector Machines have been used for Support Vector Clustering [10]. This technique can handle even highly interleaved clusters [11]. Challenges here include selecting an appropriate kernel function and its associated parameter values.

Several measures that exist for evaluating the "quality" of cluster separation are described by Desgraupe [12]. Of particular interest is the Dunn index. The Dunn index is the quotient of the minimal distance between any two points in distinct clusters over the largest diameter of any cluster. The Dunn index, $\mathscr{I}$ for a set of clusters C_i is calculated as:

$$d_{min} = \min_{i \neq j} \min \|\overline{pq}\| \qquad \forall p \in C_i, q \in C_j$$

$$d_{max} = \max_{p,q \in C_i} \|\overline{pq}\|, \qquad \mathscr{I} = \frac{d_{min}}{d_{max}}$$

1.3 Radial Visualization

Early examples of radial visualizations are William Playfair's pie charts and Florence Nightingale's polar plots [2]. Other radial visualization work includes Circle Segments[13], Star Coordinates [14], [15], and SphereViz [16]. Yi *et al.* [17] describe a radial visualization that employs "magnets." Additional RadViz research includes integration of RadViz with Parallel Coordinates by Bertini *et al.* [18], Vectorized

RadViz [19], [20], [21] and using RadViz to visualize time series data [22]. Tominski *et al.* [23] describe several different visualization methods which take high dimensional data and map it to a $2D$ image space. A thorough survey of radial visualizations is provided by Draper *et al.*[24]. Diehl *et al.* empirically evaluate the strengths and weaknesses of radial visualization [25].

The RadViz visualization introduced above can be expressed as follows. Given the n records in Euclidean space $\mathbb{E}^d$, where the i^{th} record is v_i and that record's j^{th} dimensional value is $v_{i,j}$, we follow the notation from [6], where $\vec{S}_j$ is the position on the RadViz circle of the j^{th} DA and $\vec{x}_i$ represents the location of the image of v_i:

$$\eta_i = \frac{1}{\sum_{j=0}^{d-1} v_{i,j}}, \qquad \vec{x}_i = \eta_i \sum_{j=0}^{d-1} \vec{S}_j v_{i,j}. \tag{1}$$

Daniels *et al.* show that RadViz is a special case of a normalized radial visualization (NRV) formulation. A variety of NRV properties are established there. Previously other authors, such as Nováková [26] and McCarthy *et al.* [27], had offered informal observations on properties formally stated and proved in Daniels *et al.* [6].

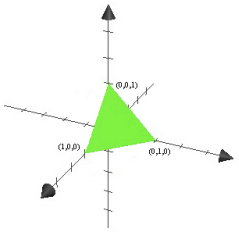


Figure 2: Triangular simplex facet onto which data points are projected under N mapping for $d = 3$ [6].

For the current paper we use 3 especially important aspects of the NRV mapping [6]: 1) it is a composition of two mappings and, 2) for cluster separation preservation some (rotational) preprocessing may be necessary prior to the application of the first mapping, and 3) the second mapping preserves barycentric combinations [28].

The first aspect decomposes the NRV mapping (in matrix form) as: $VN\Psi$, where:

- V is a $1 \times d$ vector of coordinates of a data record v_i;
- N is a $d \times d$ diagonal matrix with η_i in each diagonal entry and 0 elsewhere;
- Ψ is a $d \times d'$ matrix with each row equal to $\vec{S}_j$.

N is a perspective projection of the original data from the positive orthant of $\mathbb{E}^d$ to $\mathbb{E}^d$, with the center of projection at the origin and a projection hyperplane: $\sum_{j=0}^{j=d-1} D_j = 1$, where D_j is a variable associated with the j^{th} dimension of the data space. See Figure 2 for a $d = 3$ example. In this paper we refer to the clipped projection hyperplane as η-space. Ψ is an affinity [29], [28] from $\mathbb{E}^d$ to $\mathbb{E}^{d'}$; for RadViz we have $d' = 2$. (Note: This affinity preserves barycentric coordinates, but is not guaranteed to be 1-1 or invertible since $d' \leq d$.)

The second aspect stems from the definition of N. If two or more clusters are intersected by the same ray from the origin, then the intersection points all project to the same point, making them indistinguishable. However, [6] provides preliminary results on rotational preprocessing of the data that helps mitigate this effect. We assume in this paper that the data has not only been clustered, but it has also been rotationally preprocessed. The third aspect of the NRV mapping will be exploited in Section 2.

1.4 Dimensional Anchor Placement

The focus of much of the other previous theoretical work was the reordering and redistributing of the DAs on the circle [20], [1], [27]. Ankerst showed that the problem of optimally ordering DAs on a circle is NP-complete [30], which motivates our heuristic method for DA placement in Section 3; this appeared in earlier form (for unclustered data) in Russell's thesis [31]. Prior work on DA placement in RadViz either assumed that the DAs are uniformly distributed around the circle or else assumed that class membership of the data is known. For example, McCarthy *et al.*[27] cluster the data, extract features, use the t-statistic for feature reduction and then use a Class Discrimination Layout technique [20] that uses class structure to divide the circle into equal sized sectors; one for each class.

1.5 Q-Score

Relevant research on visual clustering metrics is surveyed in Russell [31]. This includes work by Bertini and Santucci[18] and Peng *et al.*[32]. As introduced in [33] and subsequently explored in [34], [31], the Q-score is based on the importance of the barycenter b_P of the DAs:

$$b_P = \frac{\sum_{j=0}^{d-1} \vec{S}_j}{d}, \tag{2}$$

where the P subscript refers to the configuration of DA positions. Any data record with all equal dimensional values maps to b_P [31], [35]. Furthermore, records whose standard deviation, computed across dimensional values, is small tend to map close to the barycenter of the DAs [31], [35]. Adding b_P to the set of DAs yields a set of $d+1$ points that can be interpreted as sites for a Barycentric Voronoi Diagram (BVD), which is a variant of the famous proximity construct known as the Voronoi diagram [5]. This partitions the plane into $d+1$ regions, one for each site. When the DAs are cocircular, as in most RadViz work, we obtain a Cocircular BVD (CBVD; Figure 1).

We view the region associated with b_P (denote this region by B_P) as an undesirable place for data because it can reflect lowly expressed dimensional values or cancellation effects of opposing DAs. For this reason Russell *et al.* [33] introduced the Q score as a way of measuring visualization quality. First

form a ratio $U = (\#\text{data images in } B_P)/n$. Then $-1 \leq Q \leq 1$ is defined:

$$Q = \left(\sum_{j=0}^{d-1} \frac{\#\text{data images in region}_j}{n} \right) - U = 1 - 2U. \quad (3)$$

Russell *et al.* [33], [34] show that the Q-score is a beneficial quality metric for RadViz. In this paper we use Q within our dimensional anchor placement heuristic in Section 3.

1.6 Point Sensitivity

Since one goal of this paper is to determine useful dimensional anchor placement, prior work on how the data images move as dimensional anchors move along the RadViz circle is of particular interest. Preliminary research by Russell [31] and later enhanced in [35] uses a blend of computational geometry and 2D robotic motion planning to show that when some of the DAs are fixed on the RadViz circle and the remaining DAs are allowed to move in 360° arcs around the circle the locus of points obtainable for a given v_i is an annulus a_i (see Figure 3 for 3 moving DAs). The center of the annulus depends on the fixed DAs and the inner and outer radii depend on the moving DAs. Furthermore, given any point in a_i, it is shown in [31], [35] that a corresponding DA configuration can be "reverse-engineered" from the point within the annulus. These ideas are used within our heuristic in Section 3.

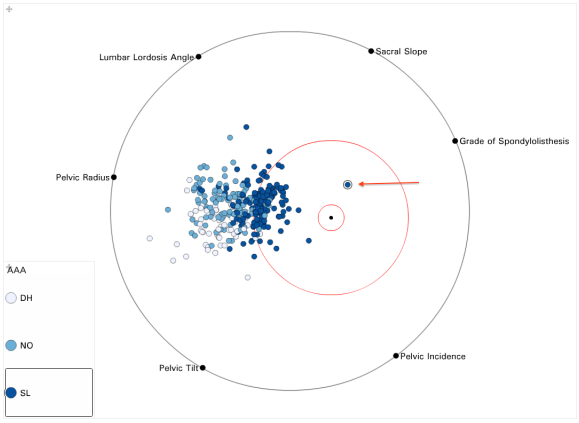


Figure 3: From [35]: the annular path traced by a single point (indicated by the arrow) when the DAs for Lumbar Lordosis Angle, Pelvic Radius, and Grade of Spondyloisthesis are moved around the circle. Data points represent records from the Vertebral Column dataset [36] shown in RadViz within the Weave visualization system [37]. Center = $(0.27, 0.006)$ and inner radius = 0.07 and outer radius=0.43.

2. Cluster Barycenter Formation

This section has 2 goals: 1) show that forming barycenters in the original data space can produce different results than forming barycenters in the η-space (Section 1.3), and 2) demonstrate that, regardless of how DAs are chosen during the second (Ψ) part of the mapping, transformed cluster representatives (if they are barycenters) are still cluster representatives of the transformed data.

For the first goal we have the following result so that the barycenter in η-space is well-defined:

Proposition 2.1: Any convex combination of a set of points in η-space is a point in η-space.

Proof: For arbitrary dimension d, a convex combination of a set of q points, $x^{(1)}, x^{(2)}, \ldots, x^{(q)}$ is $c = \sum_{h=1}^{q} \lambda^{(h)} x^{(h)}$, where $0 \leq \lambda^{(h)}$ and $\sum_{h=1}^{q} \lambda^{(h)} = 1$.

Since each $x^{(h)}$ is in the η-space we have:

$$\sum_{j=1}^{d} x_j^{(h)} = 1$$

For the convex combination of the points, we have:

$$\begin{aligned} \sum_{j=1}^{d} c_j &= \sum_{j=1}^{d} \left(\sum_{h=1}^{q} \lambda^{(h)} x_j^{(h)} \right) \\ &= \sum_{h=1}^{q} \left(\lambda^{(h)} \sum_{j=1}^{d} x_j^{(h)} \right) \\ &= \sum_{i=1}^{q} \lambda^{(h)} \cdot 1 = 1. \end{aligned}$$

■

Corollary 2.1: The barycenter of a set of points in η-space is a point in η-space.

Proof: This follow immediately from Proposition 2.1 since for arbitrary dimension d, the barycenter of a set of q points, $x^{(1)}, x^{(2)}, \ldots, x^{(q)}$ is $b = (b_1, b_2, \ldots, b_d)$ where each $b_j = \frac{1}{q} \sum_{h=1}^{q} x_j^{(h)}$. We have $\sum_{h=1}^{q} \frac{1}{q} = 1$ and $\frac{1}{q} \geq 0$. The barycenter is a convex combination of the q points. ■

We present an example (for $d = 2$) below of the difference between a cluster barycenter transformed into η-space versus the barycenter of points that have been transformed into η-space. For $d = 2$ the clipped projection hyperplane is the 2D line segment L connecting point (0,1) and (1,0). So, under the N transformation, each point p in the original data space maps to a point p' on L. The point p' on L is the intersection of L with a ray through the origin and p. Note that, as mentioned in Section 1.3, all points that lie on the same ray through the origin map to the same point on L.

Proposition 2.2: The barycenter of the points in the η-space might not coincide with the N-transformation of the barycenter of the data points.

Proof: Consider the case where $d = 2$ and the data points are $(1,3)$ and $(3,1)$. The barycenter of these two points is $(2,2)$. Applying the N-transformation to each of these points, we obtain the points $(\frac{1}{4}, \frac{3}{4}), (\frac{3}{4}, \frac{1}{4})$, and $(\frac{1}{2}, \frac{1}{2})$ for the barycenter. If the point $(3,1)$ is moved along a line segment

that passes through the origin and $(3,1)$, the N-transformation of this new point will remain $(\frac{3}{4},\frac{1}{4})$. The barycenter of the new pair of points will be the midpoint of a line segment connecting $(1,3)$ and the new point. By moving the new point away from the origin, the N-transformation of the new barycenter may be moved arbitrarily close to $(\frac{3}{4},\frac{1}{4})$. ■

A key contribution of these results is that they are independent of how the Ψ mapping is performed because Corollary 2.1 and the example in Proposition 2.2 only refer to the original data space and η-space.

For the second goal of this section, we recall that Section 1.3 stated (from [6]) that Ψ preserves barycentric combinations [28]. This implies that, if s represents a point that has been transformed from the data space via the N mapping into the η-space, and if we express s as a barycentric combination of points in the η-space: $s = \sum_{h=1}^{q} \gamma_h g_h$ for $\sum_{h=1}^{q} \gamma_h = 1$ and $(g_h \in \eta - space)$, then:

$$\Psi(s) = \Psi(\sum_{h=1}^{q} \gamma_h g_h) = \sum_{h=1}^{q} \gamma_h \Psi(g_h). \quad (4)$$

Now we observe that the barycenter of a set of points is a barycentric combination of that set of points; in this special case all γ_h are equal to $1/q$. In the clustering context this implies that we can do the following. Apply N to each cluster C_m to obtain $N(C_m)$. Now calculate the barycenter of $N(C_m)$ and let this serve as the representative R_m of cluster C_m. With the goal of finding a good configuration of DAs, the preservation of barycenter combinations under Ψ guarantees that $\Psi(R_m)$ is the barycenter of $\Psi(N(C_m))$. So, working with $\Psi(R_m)$ provides a barycentric representative of cluster $N\Psi(C_m)$ in the NRV image space. This provides execution time scalability, which is especially useful if the total number of data records is large. Working with $\Psi(R_m)$ for different Ψ possibilities, we can try many different DA configurations quickly inside of our heuristic, with each Ψ representing a different DA configuration.

3. Annular DA Placement Heuristic

We now combine our point sensitivity results (Section 1.6) with insights from our barycenter investigations (Section 2) to inform a DA placement heuristic. Our annular heuristic proceeds as follows: we first remove records with 0 standard deviation, rank the dimensions by t-statistic and select the highest ranking half as our moving set of DAs. Thereby we have removed records which are known to map exactly to the barycenter and also selected the dimensions which correspond to DAs which have a larger influence on the representative's image point position. Our heuristic then iterates over all the cluster representatives, determining a DA configuration using an annulus corresponding to the potential representative positions, and computing Q at each pass. At the completion of this loop we select the DA configuration corresponding to the highest Q computed over all iterations. Finally, we compute the Dunn index as a measure of how well we've separated the clusters. We've chosen the Dunn index as we feel it matches well to our intuitive sense of what we hope to achieve. The ratio of the minimum of inter-cluster distance to the maximal intra-cluster distance, a high Dunn index indicates denser, well separated clusters.

Note that steps 1, 2, and 17-18 operate on the complete set of original data points (untransformed), whereas inside the **for** loop we work with k transformed cluster representatives. To locate an extreme point p_m on annulus a_m $(1 \le m \le k)$ with respect to the barycenter b_P we shoot a ray through the annulus center and select the intersection of this ray with a_m's outer boundary that is furthest from b_P. We justify this in Russell [31].

ANNULAR_HEURISTIC(C,R)

1: Remove from consideration in C the low standard deviation points according to Section 1.5.
2: Apply t-statistic to C to identify lowly ranked dimensions.
3: Fix lowly ranked dimensions uniformly on circle.
4: $b_P \leftarrow$ barycenter of fixed dimensional anchors
5: $best_{DA} \leftarrow$ initial uniform configuration of moving DAs
6: $best_Q \leftarrow -1$
7: **for** $m = 1$ to k **do**
8: Make annulus a_m for cluster representative R_m.
9: Identify extreme point p_m on a_m with respect to b_P.
10: Reverse-engineer DA positions A_m for p_m.
11: Using A_m, calculate Q-score Q_m using CBVD built from A_m.
12: **if** $Q_m > best_Q$ **then**
13: $best_Q \leftarrow Q_m$
14: $best_{DA} \leftarrow A_m$
15: **end if**
16: **end for**
17: $I \leftarrow$ Dunn index using $best_{DA}$ and all the data.
18: $Q \leftarrow Q$-score using $best_{DA}$ and all the data.
19: Report $best_{DA}$ and $Q+I$.

4. Results

Here we assess clustering visualization results for our 2 competing ways to form cluster representatives using our heuristic from Section 3, implemented in Perl. For this, we compare $Q+I$ for calculating barycenters in the data space with the same measure for calculating barycenters in the η-space, where Q is the best Q-score (Section 1.5) and I is the Dunn index for clustering quality (Section 1.2).

In Table 1 the number of dimensions (d), number of clusters (k), number of data records (n) are represented for each dataset. For each dataset Q_1+I_1 is $Q+I$ for barycenters calculated in data space and Q_2+I_2 is $Q+I$ for barycenters calculated in η-space. Datasets 1-4 are generated by starting with one randomly generated point for each of the desired number of clusters and then randomly perturbing that point

Table 1: Q_1+I_1 is for barycenters calculated in data space; Q_2+I_2 is for barycenters calculated in η-space.

Dataset #	d	k	n	Q_1+I_1	Q_2+I_2
1	9	6	1800	-0.282	.31
2	9	5	1800	-0.887	.191
3	12	6	600	-0.223	.966
4	8	8	2048	-0.088	1.014
5	6	2	100310	1.152	1.256
6	9	5	1800	0.167	0.558
7	12	6	600	0.009	1.139
8	9	6	1800	-0.717	0.070
9	8	8	1000000	-.825	-.523

to generate additional points in each cluster. Dataset 5 was formed by adding to the Vertebral Column dataset (from the UCI machine learning repository [38]) a second cluster which is a 6d hyper-sphere. Datasets 6-9 are similar to datasets 1-4 with some of the dimensions more highly expressed for some of the clusters. For these 9 datasets, the average improvement in $Q+I$ for calculating barycenters in the η-space over calculating barycenters in the data space is 0.74. Furthermore, for every dataset $Q_2+I_2 > Q_1+I_1$.

Figure 4 shows the same dataset (row 2 in Table 1) as Figure 1 in Section 1.1. In Figure 4 the representatives for the heuristic were the cluster barycenters in the η-space. In Figure 1 the barycenters of the clusters in the data space were used. Note that all 5 clusters are now visible and 2 are almost out of the region for the DA barycenter. Figures 5-8 show promising results for datasets 8 and 9.

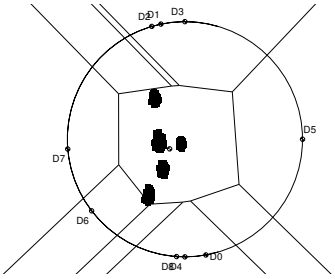


Figure 4: Dimensional anchor placement in RadViz for 1800 records of 9D data from Figure 1 with 5 clusters (dataset 2); η-space barycenters used for the heuristic.

5. Conclusion and Future Work

This paper has presented a barycentric approach for visualizing clustered datasets within the context of the RadViz visualization tool. Barycenters represent the clusters and the barycenter of the dimensional anchors is a key part of our dimensional anchor placement heuristic. We show where in the 2-stage RadViz pipeline the cluster barycenters are formed makes a difference, and, guided by the nature of part of the RadViz mapping, we form the barycenters at the end of the first stage.

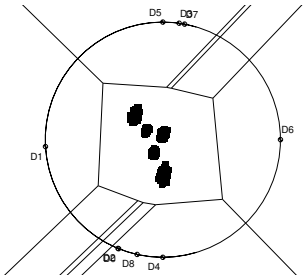


Figure 5: Dimensional anchor placement in RadViz for 1800 records of 9D data with 6 clusters, but only 5 visible (dataset 8); data space barycenters used for the heuristic.

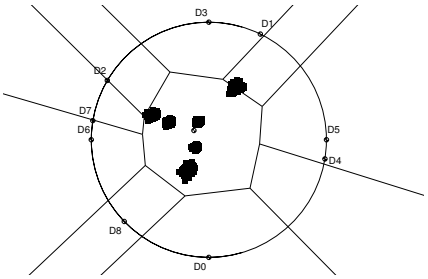


Figure 6: Dimensional anchor placement in RadViz for 1800 records of 9D data with all 6 clusters visible (dataset 8); η-space barycenters used for the heuristic.

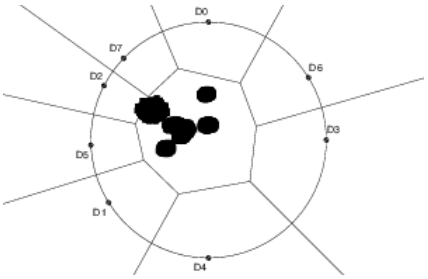


Figure 7: Dimensional anchor placement in RadViz for 10^6 records of 9D data with 8 clusters, but only 5 or 6 visible (dataset 9); data space barycenters used for the heuristic.

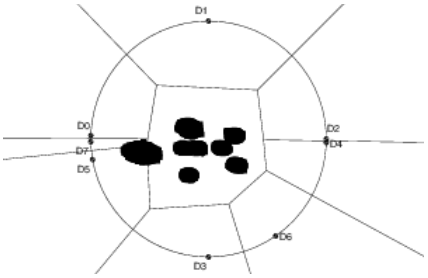


Figure 8: Dimensional anchor placement in RadViz for 10^6 records of 9D data with 8 clusters and 7 visible (dataset 9); η-space barycenters used for the heuristic.

Section 1.1 alluded to difficulty in preserving cluster separation for elongated or very irregularly shaped clusters.

One way to cope with that would be to assign a set of representatives for each of the clusters. Another intriguing direction for future work might be to use other NRV visualization invariants, such as convexity preservation, to guide the design decisions within the 2-stage visualization process.

References

[1] L. Di Caro, V. Frias-Martinez, and E. Frias-Martinez, "Analyzing the Role of Dimension Arrangement for Data Visualization in Radviz," in *Advances in Knowledge Discovery and Data Mining*, ser. Lecture Notes in Computer Science, M. Zaki, J. Yu, B. Ravindran, and V. Pudi, Eds. Springer Berlin / Heidelberg, 2010, vol. 6119, pp. 125–132.

[2] M. Ward, G. Grinstein, and D. Keim, *Interactive Data Visualization: Foundations, Techniques, and Applications*. A. K. Peters Ltd, 2010.

[3] P. Hoffman, G. Grinstein, K. Marx, I. Grosse, and E. Stanley, "DNA Visual and Analytic Data Mining," in *VIS '97: Proceedings of the 8th conference on Visualization '97*. Los Alamitos, CA, USA: IEEE Computer Society Press, 1997, pp. 437–441.

[4] ——, "DNA Visual and Analytic Data Mining," in *Visualization '97., Proceedings*, October 1997, pp. 437–441.

[5] F. Aurenhammer, "Voronoi diagrams - A Survey of a Fundamental Geometric Data Structure," *ACM Computing Surveys*, vol. 23, pp. 345–405, September 1991.

[6] K. Daniels, G. Grinstein, A. Russell, and M. Glidden, "Properties of Normalized Radial Visualizations," *Information Visualization*, vol. 11, no. 4, pp. 273–300, October 2012.

[7] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, Sept. 1999. [Online]. Available: http://portal.acm.org/citation.cfm?doid=331499.331504

[8] A. K. Jain and E. Lansing, "Data Clustering : 50 Years Beyond K-Means 1 Anil K . Jain Michigan State University," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[9] P. K. Agarwal and S. Har-peled, "Geometric Approximation via Coresets," *Discrete and Computational Geometry*, vol. 52, pp. 1–30, 2005.

[10] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik, "Support Vector Clustering," *Journal of Machine Learning Research*, vol. 2, pp. 125–137, 2001.

[11] S. Lee and K. Daniels, "Gaussian Kernel Width Exploration and Cone Cluster Labeling for Support Vector Clustering," *Pattern Recognition Letters*, vol. 13, pp. 327–344, 2012.

[12] B. Desgraupes, "Clustering Indices," pp. 1–34, April 2013. [Online]. Available: http://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf

[13] M. Ankerst, D. Keim, and H.-P. Kriegel, "Circle Segments: A Technique for Visually Exploring Large Multidimensional Data Sets," *Human Factors*, pp. 5–8, 1996.

[14] E. Kandogan, "Star Coordinates: A Multi-dimensional Visualization Technique with Uniform Treatment of Dimensions," *In Proceedings of the 7th ACM International Conference on Knowledge Discovery and Data Mining*, pp. 107–116, 2000.

[15] J. Shaik and M. Yeasin, "Visualization of High Dimensional Data using an Automated 3D Star Co-ordinate System," in *Neural Networks, 2006. IJCNN '06. International Joint Conference on*, July 2006, pp. 1339–1346.

[16] M. Soldati, M. Doulis, and A. Csillaghy, "SphereViz - Data Exploration in a Virtual Reality Environment," in *International Conference on Information Visualization*. Los Alamitos, CA, USA: IEEE Computer Society, 2007, pp. 680–683.

[17] J. S. Yi, R. Melton, J. Stasko, and J. a. Jacko, "Dust & Magnet: Multivariate Information Visualization Using a Magnet Metaphor," *Information Visualization*, vol. 00, no. April, pp. 1–18, June 2005. [Online]. Available: http://ivi.sagepub.com/lookup/doi/10.1057/palgrave.ivs.9500099

[18] E. Bertini, L. D. Aquila, G. Santucci, S. Universit, R. La, and V. Salaria, "SpringView : Cooperation of Radviz and Parallel Coordinates for View Optimization and Clutter Reduction," in *Third International Conference on Coordinated and Multiple Views in Exploratory Visualization*, 2005, pp. 22–29.

[19] J. Sharko, "RadViz Extensions with Applications," Ph.D. dissertation, University of Massachusetts Lowell, 2004.

[20] J. Sharko, G. Grinstein, and K. A. Marx, "Vectorized Radviz and Its Application to Multiple Cluster Datasets," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, 2008, pp. 1444–1451.

[21] M. Zimmerman, *Radviz Extensions with Applications*. BiblioLabsII, 2011.

[22] L. Nováková and O. Ŝtêpánková, "Visualization of Trends Using RadViz," *Journal of Intelligent Information Systems*, vol. 37, no. 3, pp. 355–369, Apr. 2011. [Online]. Available: http://link.springer.com/10.1007/s10844-011-0157-4

[23] C. Tominski and J. Abello, "Axes-Based Visualizations with Radial Layouts," in *Proceedings of ACM Symposium on Applied Computing*. ACM Press, 2004, pp. 1242–1247.

[24] G. M. Draper, Y. Livnat, and R. F. Riesenfeld, "A Survey of Radial Methods for Information Visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 15, pp. 759–776, September 2009.

[25] S. Diehl, F. Beck, and M. Burch, "Uncovering Strengths and Weaknesses of Radial Visualizations–an Empirical Approach." *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 935–42, 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20975130

[26] L. Nováková, "Visualization data for data mining," Ph.D. dissertation, Czech Technical University in Prague, 2009, supervisor-Ŝtêpánková, Olga.

[27] J. F. McCarthy, K. A. Marx, P. E. Hoffman, A. G. Gee, P. O'Neil, M. L. Ujwal, and J. Hotchkiss, "Applications of Machine Learning and High-Dimensional Visualization in Cancer Detection, Diagnosis, and Management," *Annals of the New York Academy of Sciences*, vol. 1020, no. 1, pp. 239–262, 2004.

[28] G. Farin, *Curves and Surfaces for CAGD: A Practical Guide, fifth edition*. Morgan Kaufmann, 2002.

[29] F. Flohr and F. Raith, *Fundamentals of Mathematics, Volume II: Geometry: Affine and Euclidean Geomety*, H. Behnke, F. Backmann, K. Fladt, and H. Kunle, Eds. MIT Press, 1987.

[30] M. Ankerst, S. Berchtold, and D. A. Keim, "Similarity Clustering of Dimensions for an Enhanced Visualization of Multidimensional Data," in *Proceedings of the IEEE Symposium on Information Visualization*, G. Wills and J. Dill, Eds. IEEE Computer Society Press, 1998, pp. 52–60.

[31] A. Russell, "Formulation and Application of Radial Visualization Properties," Ph.D. dissertation, University of Massachusetts, Computer Science Department, 2013.

[32] W. Peng, M. O. Ward, and E. A. Rundensteiner, "Clutter Reduction in Multi-Dimensional Data Visualization Using Dimension Reordering," in *INFOVIS '04: Proceedings of the IEEE Symposium on Information Visualization*. Washington, DC, USA: IEEE Computer Society, 2004, pp. 89–96.

[33] A. Russell, K. Daniels, and G. Grinstein, "Voronoi Diagram Based Dimensional Anchor Assessment for Radial Visualizations," in *Information Visualization (IV), 2012 16th International Conference on*, July 2012, pp. 229–233.

[34] ——, "Voronoi Diagram Based Dimensional Anchor Assessment for Radial Visualizations," in *Information Visualization - Techniques, Usability and Evaluation*, Banissi and et al, Eds. Cambridge Scholar Publishing, to appear, 2014.

[35] A. Russell, F. Kamayou, R. Marceau, K. Daniels, and G. Grinstein, "Point Sensitivity for Radial Visualization under Dimensional Anchor Motion," in *22nd International Conference on Computer Graphics, Visualization and Computer Vision (WSCG)*, to appear in 2014.

[36] K. Bache and M. Lichman, "UCI Machine Learning Repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[37] A. Dufilie, P. Stickney, J. Fallon, and G. Grinstein, "Weave: A Web-Based Architecture Supporting Asynchronous and Real-Time Collaboration," in *Proceedings of the International Conference on Advanced Visual Interfaces*, Capri, 2012.

[38] A. Frank and A. Asuncion, "UCI machine learning repository," 2010, http://archive.ics.uci.edu/ml.

Parallel Coordinates for Visualization of Rules Developed using Grammatical Evolution

M. Abdulgader[1] **and D. Kaur**[1]
[1] Department of Electrical Engineering and Computer Science ,University of Toledo, Toledo, Ohio, USA

Abstract- *Information visualization has become an increasingly interesting research area. Such visualization techniques are found to be useful in many applications. This paper illustrates the visualization of fuzzy rules generated by grammatical evolution using parallel coordinates. The application chosen was the well-studied machine learning problem of iris data and the fuzzy rules generated to classify it. The paper presents a novel technique using parallel coordinates to visually verify the correctness of each rule.*

1. Introduction

The use of visualization techniques to view large scale multi-dimensional datasets is a well-known approach in data mining field. Visualization of data higher than three dimensions is not an easy and straight forward task using Cartesian coordinates. Representing n-dimensional dataset using Cartesian coordinates will require scatter plots taking two dimensional representation at time which makes the complexity of $O(n^2)$. Inselberg [1] proposed the parallel coordinates method where the n-dimensional data is represented using n parallel lines. Therefore reducing the complexity of the n-dimensional data set to be O(n). With the N axes aligned as parallel lines; there is a correspondence of dimensions with their respective axis. The set of connected points on parallel axes is a polyline which represent one entry of the n-dimensional data set [1]. For illustrating the visualization with parallel coordinate, the Iris data by R.A. Fisher has been used. The three species of Iris flower have four input parameters which are: sepal width, sepal length, pedal width and pedal length [2]. Fig.1 illustrates the representation of Iris data using five parallel coordinates, where four coordinates are for inputs and the fifth is for the output which represents the three categories of the flower, *Setosa*, *Versicolour* and *Virginica*.

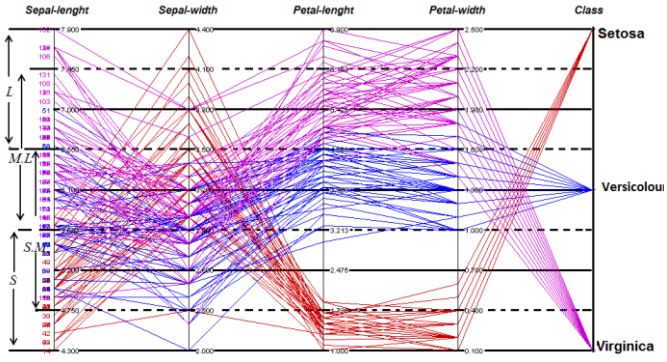


Fig.1: Representation of Iris Data in Parallel Coordinates

This paper presents a technique to verify the correctness of each fuzzy rule evolved using grammatical evaluation [2] visually by using parallel coordinates. The paper is organized into 5 sections. Section 2 explains the grammatical evolution and the fuzzy rules evolved using it. The two-dimensional Cartesian plane for the data is presented in Section 3. Section 4 focuses on demonstrating the visual representation of Iris dataset. Section 5 displays the fuzzy rules in parallel coordinates and analyses the correctness of the rules for each category.

2. Grammatical Evolution

Grammatical Evolution is a process of evolving a grammar template which is defined for each problem uniquely using Backus Naur Form (BNF) template [3, 4, 5]. For the Iris data the template was chosen in the form of fuzzy rules. The input parameters of *Sepal Length(SL), Petal Length(PL), Sepal Width(SW) and Petal Width(PW)* are uniformly divided into four fuzzy sets of *Small*, Small Medium, *Medium Large* and *Large*, as shown in Fig. 2. [2]

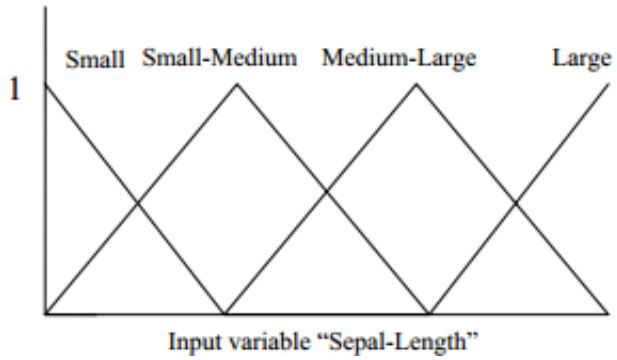


Fig.2: Triangular membership functions for Sepal-Length [2]

The representative tuple of BNF grammar is {N, T, P S}, where N represents an assortment of non-terminals, T is for the set of terminals, P shows the production rules which are necessary for the conversion of N to T, and S, an element of N, is the start symbol[6,7]. Fig. 3 represents the BNF template for the fuzzy rules to be evolved for iris data [2].

The difference between using GE and other evolutionary approaches in the literature is that in GE the rule structure is evolvable within the constraint of the BNF. Also the transcription from bit string to evolved code uses redundancy to improve search efficiency[2,8]

Non_terminals = {<statement>,<ant>,<cons>}
Terminals = {Small, Small-Medium, Medium-Large, Large, Virginica, Setosa, Versicolour}
(Start symbol) S = <statement>
(Production rules) P=
<statement> ::=
If (SL is <ant>) and (SW is <ant>) and (PL is <ant>)and (PW is <ant>) Then <cons>;
If (SL is <ant>) and (SW is <ant>) and (PL is <ant>)and (PW is <ant>) Then <cons>;
.
.
. *(repeat n times for n rules)*
.

If (SL is <ant>) and (SW is <ant>) and (PL is <ant>)and (PW is <ant>) Then <cons>;
<ant> ::= Small (0)
| Small-Medium (1)
|Medium-Large (2)
|Large (4)
|All (5)
<cons>::= *Virginica* (0)
|*Setosa* (1)
|*Versicolour* (2)

Fig. 3: Actual BNF used for Rule base [2]

The following section of the BNF template defines the template of rules:

If (SL is <ant>) and (SW is <ant>) and (PL is <ant>)and (PW is <ant>) Then <cons>;
If (SL is <ant>) and (SW is <ant>) and (PL is <ant>)and (PW is <ant>) Then <cons>;
.
. *(repeat n times for n rules)*

Each input parameter such as SL, PL SW and PW is associated with and <ant> which represents the antecedent to be evolved. The choices for <ant> are the fuzzy sets from Fig.3 and are shown below.

<ant> ::= Small (0)
| Small-Medium (1)
|Medium-Large (2)
|Large (4)
|All (5)

The options for <cons> are

<cons>::=
Virginica (0)
|Setosa (1)
|Versicolour (2)

Eight fuzzy rules were obtained through the use of "Fuzzy Classification using Grammatical Evolution for Structure Identification" [2]. The classification success rate was 97.7%. Population size was 100 and it was randomly initialized, Number of generations was 50 and Crossover rate was 0.8.

Table1. A Rule base on Fuzzy Classification using GE [2]

	Antecedent	
1	If (SL is Large) and (PL is Large)	*Virginica*
2	If (SL is Large) and (PL is Medium -Large) and (PW is Large)	*Virginica*
3	If (SL is Small) and (SW is Medium-Large) and (PL is Small)	*Setosa*
4	If (PL is Small)	*Setosa*
5	If (SL is Medium-Large) and (PL is Large) and (PW is Small-Medium)	*Virginica*
6	If (SL is Small-Medium) and (SW is Small) and (PL is Medium -Large)and (PL is Large)	*Virginica*
7	If (SL is Small) and (SW is Small-Medium) and (PL is Large)and (PW is Medium-Large)	*Virginica*
8	If (SW is Medium-Large) and (PL is Medium-Large)and (PW is Large)	*Versicolour*

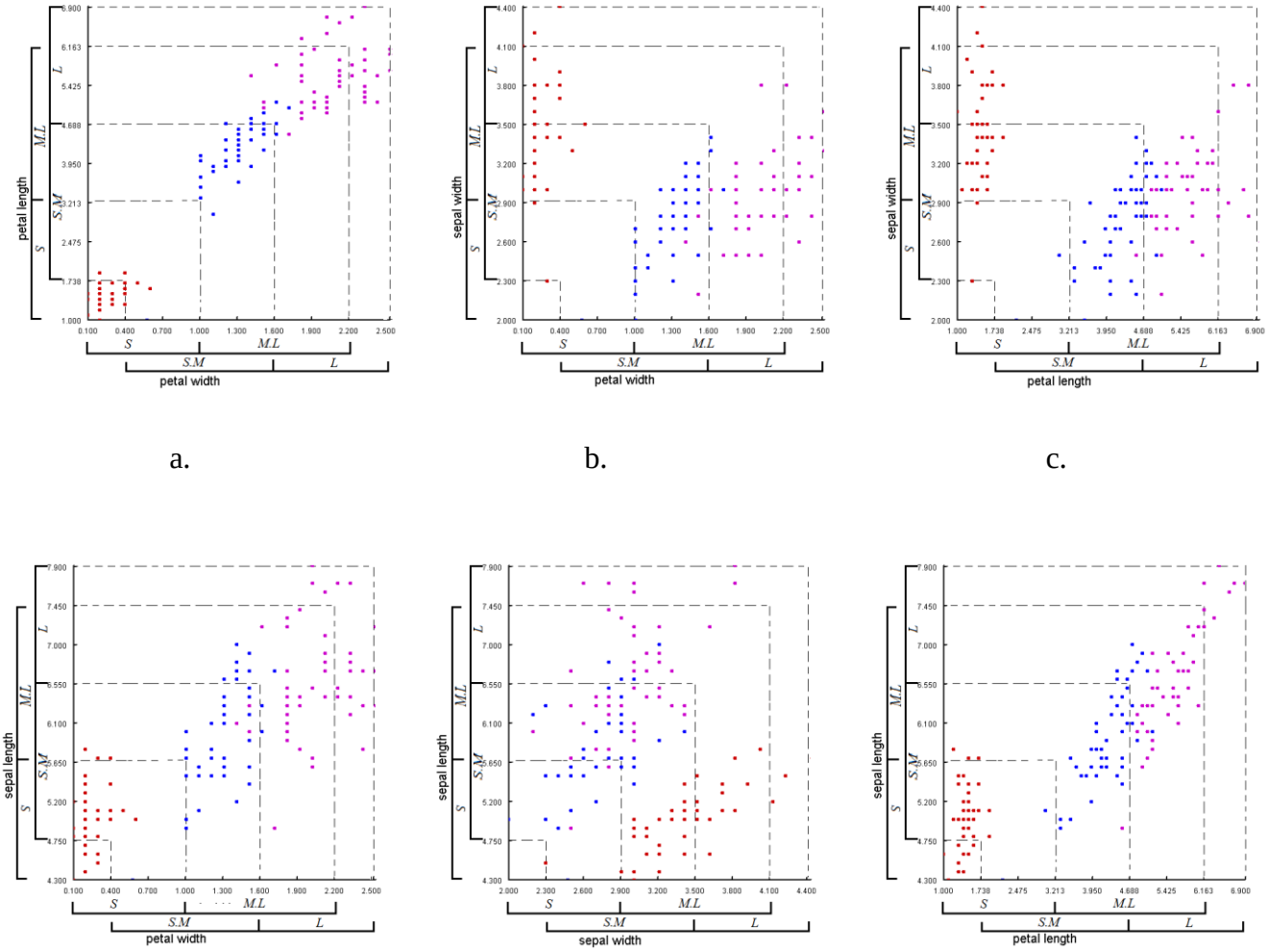
3. Two-dimensional Visualization

The conventional way of visualization of 4-dimensional data would be to generate scatter plots taking two parameters at a time. This would generate six plots as shown in Fig.4. From the scattered graphs if there is no overlap of colored dots (colors represent different categories of iris flower), it is easy to come up with rules using two antecedents which would classify the data correctly.

For instance from Fig. 4a it can be inferred that if *Petal Length* is small and *Petal Width* is small then the category is *Setosa* irrespective of sepal length and sepal width. In other words *Sepal Length* and *Sepal Width* are Don't Care. Fig 4b and 3c also validate the same rule for category *Setosa*. This rule was again verified in parallel coordinates as shown in Fig. 4a.

However, there is a considerable overlapping among the categories of V*irginica* and *Versicolour* (among blue and magenta colors). Fig. 3b illustrates that if petal width is *Small-Medium* or *Medium Large* the category is always *Versicolour*. Fig. 4a illustrates that if *Petal width* is Small Medium and *Petal Length* is small medium the category is *Versicolour*. However from Fig. 4b it is also clear that if *Petal Width* is *Medium Large* and *Sepal Width* is *Medium Large* then the output category can be *Versicolour* or *Virginica*. In this case the boundaries are fuzzy and therefore no crisp rule can classify it correctly. Hence, in order to classify iris correctly we had evolved fuzzy rules.

Thus, it can be concluded that scatter plots are not enough to classify the overlapping iris data. The disadvantage of implementing two-dimensional representation obscures the efficient analysis of other parameters regarding their impact on the classification. Therefore it can be concluded that Cartesian representation is not suitable for classifying the overlapping data of iris flower.



a. b. c.

d. e. f.

Fig. 4: Two-dimensional Scatter Plots for Iris

4. Visualization of Rules in Parallel Coordinates

By applying the filters in the parallel coordinate representation of the iris data, figures 5a, 5b and 5c were generated. The data represented by burgundy are the instances of class *Setosa*. Fig. 5b shows the data which represent *Veriscolor* in blue lines. The purple lines represent the data that belong to *Virginica* in Fig. 5c. Fig 5 represents the following three observations.

1. If PL is small, or PW is small, it is always *Setosa.*
2. If PL is small-medium or medium-large, and PW is small-medium or medium-large, it will be always recognized as *Versicolour*.
3. If PL is medium-large or large and PW is medium-large or large, then the class will be *Virginica.*

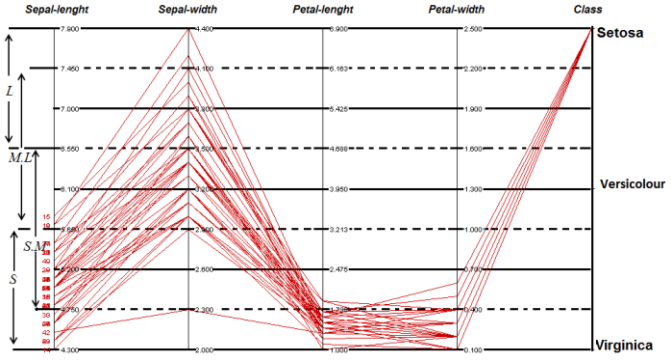
From the above three observations it is clear that the first observation classifies the *Setosa* correctly. However, for the classification of *Versicolour* and *Virginica*, there is overlap of input parameters in several sets like Large, Medium Large and Small Medium.

Therefore, it is clear there are no crisp rules which can classify the iris data correctly. We evolved fuzzy rules using grammatical evolution to classify the iris data.

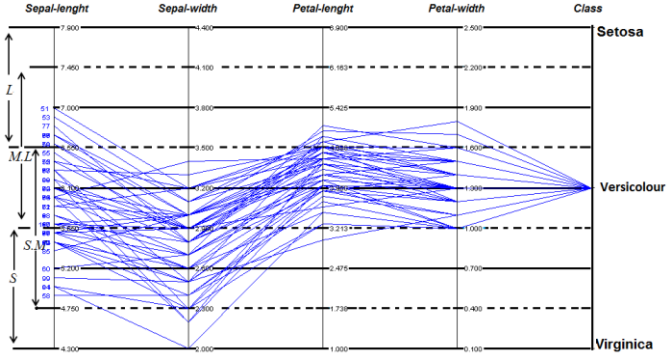
Table 2: Range of four uniformly distributed sets

	SL		SW		PL		PW	
S	4.300	5.650	2.000	2.900	1.000	3.213	0.100	1.000
SM	4.750	6.550	2.300	3.500	1.738	4.688	0.400	1.600
ML	6.650	7.450	3.900	4.100	3.213	6.163	1.000	2.200
L	6.550	7.900	3.500	4.400	4.688	6.900	1.600	2.500

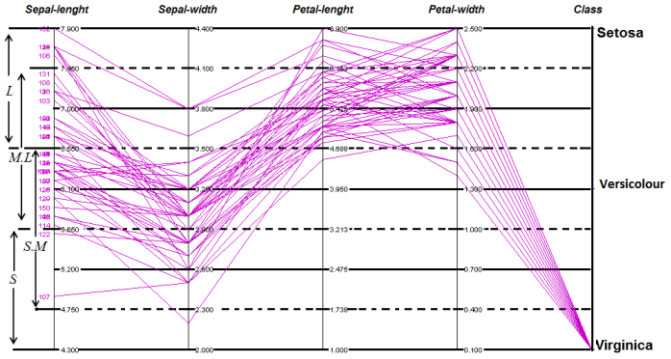
In order to view the correctness of the fuzzy rules evolved using grammatical evolution as shown in Table 1, we divided the vertical axis of each parameter into four uniformly distributed and overlapping fuzzy sets of small , small medium, medium large and large. Table 2 shows the range of fuzzy sets S, SM, ML and Large for each of the four attributes such as SL, SW, PL and PW.



a. Setosa



b. Versicolour



c. Virginica

Fig5. parallel coordinates

5. Analyzing the Rules in Parallel Coordinates

Our next objective is to verify the correctness of the fuzzy rules evolved by grammatical evolution. In this section we are visually verifying the accuracy of eight fuzzy rules as shown in Table 1. To verify the classification of each rule, we have transformed the input tuples to segmental parallel coordinate representation, which represent the fuzzy sets of the input parameters.

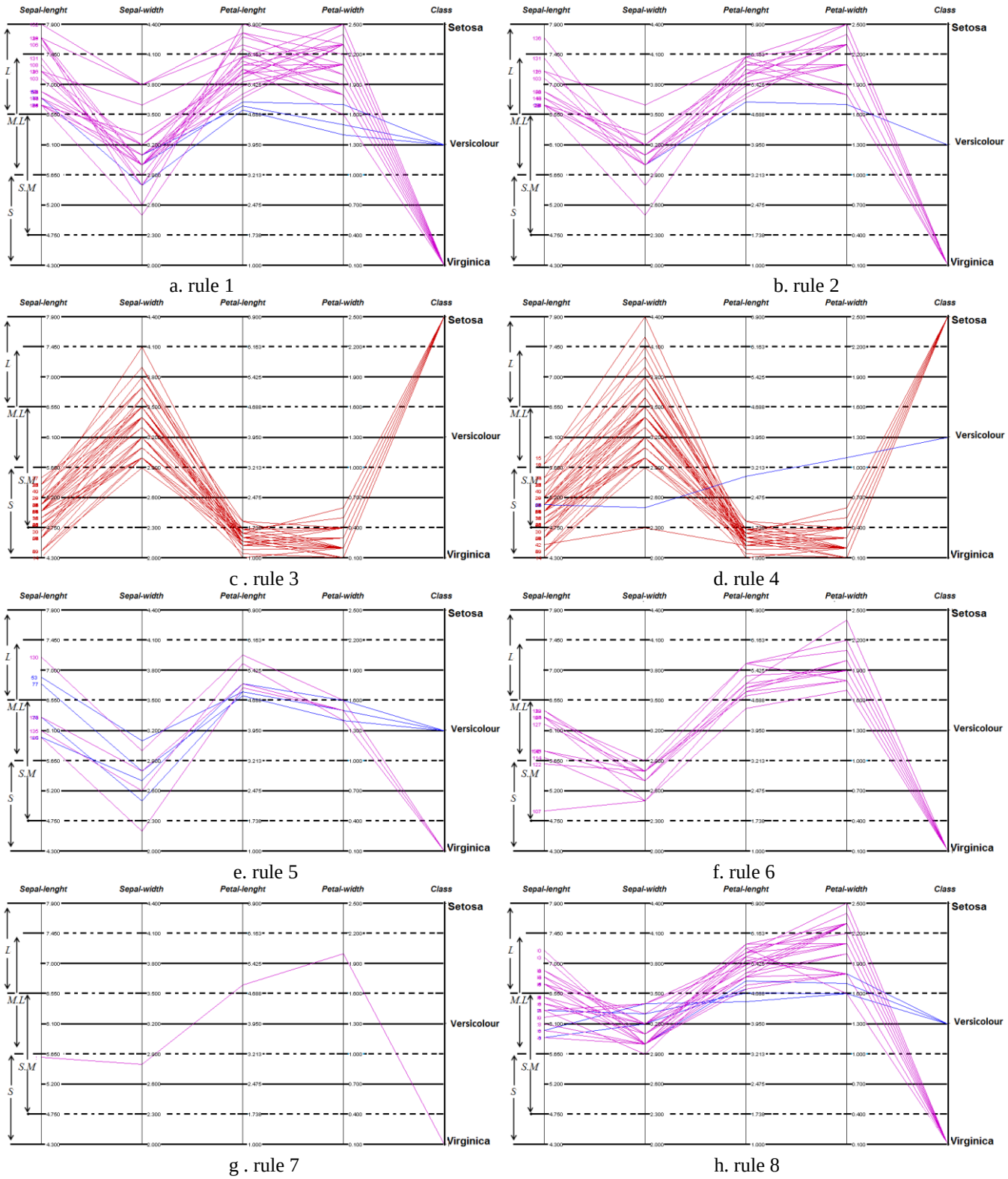
Using parallel coordinates, we have divided the range of each input attribute such as SL, PL, PW and SW into fuzzy sets of S, SM, ML and Large as shown in Fig 6. All the eight rules are shown in Fig 6a-6h. Fig 6a represents rule 1 and was created using filters where SL is Large and PL is Large

While the other two parameters, namely SW and PW are don't care. The output category of this rule is *Virginica*. However, Fig. 6a shows three misclassifications giving the category as *Versicolour*, which indicates the strength of the rule is not 100%.
In fuzzy rules the strength of each rule varies between the interval 0 and1. Therefore this only reflects that the strength of rule 1 is less than 1.

All the eight rules are taken together and applied to 150 data entries to see the overall classification success rate.
Similarly, Fig 6b. illustrates the correctness of the rule 2. The petal length belongs to the range of small medium and petal width ranges in the large scale. The rule 3 can be verified by applying filters on PL small, SW is medium-large SL is small and regardless of petal width as shown in Fig 6.c.

Fig. 6d verifies the correctness of rule 4 to predict Setosa. Fig. 6e represents the rule 5 and three input patterns converge to *Virginica* and other three patterns converge to *Versicolour*, which illustrates the rule strength of this rule, is 50%. On the contrary, Fig 6f illustrates the rule strength of rule 6 to be 100% as all the input patterns have been classified correctly to *Virginica*.

Furthermore, Fig. 6g shows the correctness of the rule 7. Rule 8 is represented in Fig. 6h. Here there are 8 patterns which are misclassified and only three patterns are correctly classified. This indicates the rule strength of this rule is 27%.



a. rule 1 b. rule 2

c . rule 3 d. rule 4

e. rule 5 f. rule 6

g . rule 7 h. rule 8

Fig.6. parallel coordinates of Fuzzy Classification Rule base on GE

6. Conclusion

In this paper, we presented the Parallel Coordinate Technique to visualize the fuzzy rules for classification of iris. This serves as a proof of concept for visualization of fuzzy classification system. By applying filters to the output, it is possible to generate some knowledge in the form of rules. This knowledge is utilized to classify the different species of Iris flower. Future work will include the application of this technique to achieve solutions for more complicated classification problems and investigating ways of obtaining higher classification rates.

7. References

[1] A. Inselberg, Parallel Coordinates Visual Multidimensional Geometry and Its applications, Heidelberg, London, New York: Springer Dordrecht, 2009.

[2] D. Wilson and D. Kaur, "Fuzzy Classification using Grammatical Evolution for Structure Identification", Proceedings of NAFIPS 2006, Montreal, June 3-6, 2006. IEEE Catalog No. 05TH 8815C, ISBN: 0-7803-9188-8, Library of Congress: 2006924074.

[3] Goldberg, D.E. Genetic Algorithms in Search, Optimization and Machine Learning, pp. 1-14. Reading, MA: Addison Wesley, 1989.

[4] Ryan, C., Collins, J.J. and Michael O'Neill, Grammatical Evolution: Evolving Programs for an Arbitrary Language. Lecture Notes in Computer Science: Proceedings of the First European Workshop on Genetic Programming, pp. 83-95, 1998.

[5] Ryan, C., Michael O'Neill and Collins, J.J. Grammatical Evolution: Solving Trigonometric Identities. Proceedings of Mendel 1998: 4th International Mendel Conference on Genetic Algorithms, Optimization Problems, Fuzzy Logic, Neural Networks, Rough Sets, pp. 111-119, 1998.

[6] Ryan, C. and Michael O'Neill, "Grammatical Evolution: A Steady State Approach". Proceedings of the Second International Workshop on Frontiers in Evolutionary Algorithms, pp. 419-423, 1998.

[7] Michael O'Neill and Conor Ryan, Grammatical Evolution: Evolutionary Automatic Programming in an Arbitrary Language, Kluwer Academic Publishers, 2003.

[8] S.Y. Ho, H.M. Chen, S.J. Ho and T.K. Chen, "Design of accurate classifiers with a compact fuzzy-rule base using an evolutionary scatter partition of feature space", IEEE Transactions on Systems, Man and Cybernetics, Part B 34:2 (2004) 1031-1044

The singular properties of the Hermit curve on Unit Circle

Ching-Shoei Chiang [1], **Ching-Tsorng Tsai** [2]

[1]Computer Science and Information Management, Soochow University, Taipei, Taiwan

[2]Computer Science, Tunghai University Taichung, Taiwan

Abstract - *The design of curves, surface, and solid are important in computer aided geometric design (CAGD). Images, surround by boundary curves, are also investigated by many researchers. One way to describe an image is using the medial axis transform. Under this consideration, the properties of the boundary curves tangent to circles become important to design for 2D images. In this paper, we define a special class of curves, which is the curves whose end points is on a circle, and endpoints tangent vectors are parallel to the tangent line of circle at the end points. There are 4 parameters to design this curve, two are the length of the tangent vectors α_0 and α_1 on the boundary curve, one is the angle θ, associated with the circle, between two endpoints, and the final parameter is the variable t for the curves. Different curves design for singular point can be achieved by given two values or constraints among these 4 parameters. We can use the singular point design in calligraphic strokes with cusp to describe the edge contour.*

Keywords: Computer Graphics, computer-aided geometric design, Hermit Curve, geometric modeling

1 Introduction

The 2D image can be stored by many different ways, including its boundary curves, a medial axis curves with radius function, union of many primitive figures, and so on. There are many researchers use different curves to simulate different images. For example, Cinque , Levialdi and Malizia [1] uses cubic Bezier curve to do the shape description. Yang, Lu and Lee[2] use Bezier curve to approach the shape description for Chinese calligraphy characters. Chang and Yan[3] derived an algorithm to approach the hand-drawn image by using cubic Bezier curve. Cao and Kot[4] derived an algorithm to do data embedding in electronic inks without losing data. The boundary curves can be also used to derived the offset curves and medial axis of an images[5], and also to simulate the nature objects, such as flowers[6].

In this paper, we would like to investigate the cubic Hermit curves, where the end points are on a unit circle, and its end points tangent line is parallel to the tangent line of circles. Using these properties, with more constraints on the singular point or maximum curvature at specified parameter, we survey the couture of the curves, so that the design of the curve may be simplify by giving constraint.

2 Definition and Theorem

Let introduce the Hermit curve first. Given two points P_0 and P_1, with two associated tangent vectors V_0 and V_1, the Hermit curve C(t) defined as (See figure 1):

$$C(t) = [t^3, t^2, t, 1] \begin{bmatrix} 2 & -2 & 1 & 1 \\ -3 & 3 & -2 & -1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} P_0 \\ P_1 \\ V_0 \\ V_1 \end{bmatrix} \quad (1)$$

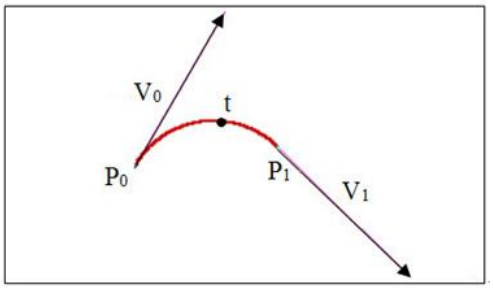


Figure 1. Hermite Curve

To simplify our problem, we assume that the end points P_0 and P_1 are on the unit circle, the first point P_0=(-1,0), and the second point P_1=(-cosθ, sinθ). It's associated tangent vectors parallel to (0,1) and (sinθ, cosθ). We have the following definition for this curve.

Definition1：Unit Circle Hermit Curve

Two points P_0=(-1, 0)、P_1=(-cosθ, sinθ) on the unit circle with its associated tangent vector $V_0=\alpha_0(0, 1)$、$V_1=\alpha_1(\sin\theta, \cos\theta)$ produced a Hermite curve, we call it unit circle hermit curve, denoted $H_1(t;\theta,\alpha_0, \alpha_1)$，where $\alpha_0>0$、$\alpha_1>0$、$0<t<1$、$0<\theta<2\pi$。

Notice that for the general case, we can always convert the design problem into the design problem for $H_1(t;\theta,\alpha_0,\alpha_1)$. We can always translate the center to the origin, scale the radius into one and rotate the circle so that the first point is on (-1,0). We call this process the standardization of the problem. After we solve the problem, produce the curve we want, we can always inverse the process, which is rotate, scale, and translate to see the designed curves for the original design problem.

Let us consider the cusp on the curve.

Theorm1： Singular Point

Let $C(t)=H_1(t;\theta,\alpha_0,\alpha_1)$，then

$C'(t)=(x'(t),y'(t))=(0,0)\Leftrightarrow\alpha_0B(t)=\alpha_1D(t)=A(t)\tan\frac{\theta}{2}$，where $A(t)=6t(t-1)$、$B(t)=(3t-1)(t-1)$、$D(t)=t(3t-2)$。

Proof：

We separate our proof into two parts, including if ("⇒") and only if ("⇐").

(1) $C'(t)=(x'(t),y'(t))=(0,0)\Rightarrow\alpha_0B(t)=\alpha_1D(t)=A(t)\tan\frac{\theta}{2}$

After the first derivative of the curve, we have

$$C'(t)=[A(t),-A(t),B(t),D(t)]\begin{bmatrix}(-1,0)\\(-\cos\theta,\sin\theta)\\\alpha_0(0,1)\\\alpha_1(\sin\theta,\cos\theta)\end{bmatrix}\quad(2)$$

Where $A(t)=6t(t-1)$、$B(t)=(3t-1)(t-1)$、$D(t)=t(3t-2)$。

From $C'(t)=(0,0)$，the following equations holds：

$$\begin{cases}-A(t)+\cos\theta A(t)+\alpha_1\sin\theta D(t)=0 & (3)\\ -A(t)\sin\theta+\alpha_0B(t)+\alpha_1\cos\theta D(t)=0 & (4)\end{cases}$$

From Cramer's rule to solve $\sin\theta$ and $\cos\theta$, with the properties $\sin^2\theta+\cos^2\theta=1$, we have:

$$\text{or}\begin{cases}A^2(t)+{\alpha_1}^2D^2(t)=0 & (5)\\ {\alpha_0}^2B^2(t)-{\alpha_1}^2D^2(t)=0 & (6)\end{cases}$$

Notice that equation (5) never holds under the assumption that $\alpha_0>0$、$\alpha_1>0$、$0<t<1$。

We can simplify the equation (6) into $\alpha_0B(t)=\pm\alpha_1D(t)$：

$$\text{or}\begin{cases}\alpha_0B(t)=\alpha_1D(t) & (6.a)\\ \alpha_0B(t)=-\alpha_1D(t) & (6.b)\end{cases}$$

Solve the system of equation (3)、(4) and(6.a)，we have $\alpha_0B(t)=\alpha_1D(t)=A(t)\tan\frac{\theta}{2}(0<\theta<2\pi)$。On the other hand, solve the system of equation (3)、(4) and (6.b)，we have $\alpha_1D(t)=A(t)\tan\frac{\theta}{2}=-A(t)\cot\frac{\theta}{2}$，conclude $\sec^2\frac{\theta}{2}=0$，since we cannot find any θ so that $\sec^2\frac{\theta}{2}=0$，so $\alpha_0B(t)=-\alpha_1D(t)$ will never happened。

(2) $C'(t)=(x'(t),y'(t))=(0,0)\Leftarrow\alpha_0B(t)=\alpha_1D(t)=A(t)\tan\frac{\theta}{2}$

Via equation (2), we have：

$(x'(t),y'(t))=(-A(t)+A(t)\cos\theta+\alpha_1D(t)\sin\theta,-A(t)\sin\theta+\alpha_0B(t)+\alpha_1D(t)\cos\theta)=-2\left(A(t)\sin\frac{\theta}{2}-\alpha_1D(t)\cos\frac{\theta}{2}\right)\left(\sin\frac{\theta}{2},\cos\frac{\theta}{2}\right)=-2\cos\theta\left(A(t)\tan\frac{\theta}{2}-\alpha_1D(t)\right)\left(\sin\frac{\theta}{2},\cos\frac{\theta}{2}\right)$。

Since $\alpha_0B(t)=\alpha_1D(t)=A(t)\tan\frac{\theta}{2}$，we got $A(t)\tan\frac{\theta}{2}-\alpha_1D(t)=0$，so $C'(t)=(0,0)$。

◆

Observation: Notice that there are some values of t degenerate the case. Assume t can be 0 ($A(t)=D(t)=0$) or 1($A(t)=B(t)=0$), then there are no cusp in the defined curve.

Now we have 4 variables (t、θ、α_0、α_1) with two constraints ($x'(t)=0$、$y'(t)=0$)，in general case, we have solutions by given the values of two variables or two constraints. These two variables can be (θ, t)、(θ, α_0)、(θ, α_1)、(t, α_0)、(t, α_1)、(α_0, α_1)。Constraints can be something like $\alpha_0=\alpha_1$. So, we have the following corollaries:

2.1 Corollary1：θ、t are given

Let $C(t)=H_1(t;\theta,\alpha_0,\alpha_1)$，and gives the value of θ、t，then $\alpha_0=\frac{6t}{3t-1}\tan(\frac{\theta}{2})$、$\alpha_1=\frac{6(t-1)}{3t-2}\tan(\frac{\theta}{2})$，where $t\neq\frac{1}{3}$、$t\neq\frac{2}{3}$。

From Corollary1, we can know whether α_0 and α_0 are positive, negative, or zero by given the range of θ and t，as shown in Table 1.

Let $0<t<1$, $0<\theta<2\pi, t\neq\frac{1}{3}, t\neq\frac{2}{3}, \theta\neq\pi$, we have 4 different cases concerning the positive/negative values of α_0 and α_1, that is, ($\alpha_0>0\wedge\alpha_1>0$)、($\alpha_0>0\wedge\alpha_1<0$)、($\alpha_0<0\wedge\alpha_1>0$) and ($\alpha_0<0\wedge\alpha_1<0$). These cases, associated with the range of θ and t, are list in Table 2.

2.2 Corollary 2：θ、α_0 are given

Let $C(t)=H_1(t;\theta,\alpha_0,\alpha_1)$，if we know the value of θ *and* α_0，then $\alpha_1=4\frac{\alpha_0-3\tan\frac{\theta}{2}}{\alpha_0-4\tan\frac{\theta}{2}}\tan\frac{\theta}{2}$、$t=\frac{\alpha_0}{3(\alpha_0-2\tan\frac{\theta}{2})}$ where $\theta\neq0$。

2.3 Corollary 3：θ、α_1 are given

Let $C(t)=H_1(t;\theta,\alpha_0,\alpha_1)$，if we know the value of θ *and* α_1，

then $\alpha_0 = 4\frac{\alpha_1-3\tan\frac{\theta}{2}}{\alpha_1-4\tan\frac{\theta}{2}}\tan\frac{\theta}{2}$、$t=\frac{2}{3}\frac{\alpha_1-3\tan\frac{\theta}{2}}{\alpha_1-2\tan\frac{\theta}{2}}$，where $\theta \neq 0$。

Notice that the similar equation for α_1 in corollary 2 and α_0 in corollary 3.

2.4 Corollary4：Given t、α_0

Let $C(t)=H_1(t;\theta,\alpha_0,\alpha_1)$, if we know the value of t、α_0，then

$\alpha_1 = \frac{\alpha_0(3t-1)(t-1)}{t(3t-2)}$、$\theta = 2\tan^{-1}(\frac{\alpha_0(3t-1)}{6t})$，where $t \neq 0$ 、$t \neq \frac{2}{3}$。

Table 1 Positive/Negative of α_0, α_1

	>0	=0	＜0
$\tan\frac{\theta}{2}$	$0<\theta<\pi$	θ=0	$\pi<\theta<2\pi$
$\frac{6t}{3t-1}$	$t>\frac{1}{3}$ or $t<0$	t=0	$0<t<\frac{1}{3}$
$\frac{6(t-1)}{3t-2}$	t>1 or $t<\frac{2}{3}$	t=1	$\frac{2}{3}<t<1$
α_0	$(t>\frac{1}{3})\wedge(0<\theta<\pi)$ $(t<0)\wedge(0<\theta<\pi)$ $(0<t<\frac{1}{3})\wedge(\pi<\theta<2\pi)$	t=0 θ=0	$(t>\frac{1}{3})\wedge(\pi<\theta<2\pi)$ $(t<0)\wedge(\pi<\theta<2\pi)$ $(0<t<\frac{1}{3})\wedge(0<\theta<\pi)$
α_1	$(t>1)\wedge(0<\theta<\pi)$ $(t<\frac{2}{3})\wedge(0<\theta<\pi)$ $(\frac{2}{3}<t<1)\wedge(\pi<\theta<2\pi)$	t=1 θ=0	$(t>1)\wedge(\pi<\theta<2\pi)$ $(t<\frac{2}{3})\wedge(\pi<\theta<2\pi)$ $(\frac{2}{3}<t<1)\wedge(0<\theta<\pi)$

Table 2 Range of α_0, α_1 from range of θ and t

Case A	$(\frac{1}{3}<t<\frac{2}{3})\wedge(0<\theta<\pi)$	⇔	$(\alpha_0>0\wedge\alpha_1>0)$
Case B	$(\frac{2}{3}<t<1)\wedge(0<\theta<\pi)$ or $(0<t<\frac{1}{3})\wedge(\pi<\theta<2\pi)$	⇔	$(\alpha_0>0\wedge\alpha_1<0)$
Case C	$(0<t<\frac{1}{3})\wedge(0<\theta<\pi)$ or $(\frac{2}{3}<t<1)\wedge(\pi<\theta<2\pi)$	⇔	$(\alpha_0<0\wedge\alpha_1>0)$
Case D	$(\frac{1}{3}<t<\frac{2}{3})\wedge(\pi<\theta<2\pi)$	⇔	$(\alpha_0<0\wedge\alpha_1<0)$

2.5 Corollary5：given t、α_1

Let $C(t)=H_1(t;\theta,\alpha_0,\alpha_1)$，if we know the value of t、α_1，

then $\alpha_0 = \frac{\alpha_1 t(3t-2)}{(3t-1)(t-1)}$、$\theta = 2\tan^{-1}(\frac{\alpha_1(3t-2)}{6(t-1)})$，where

$t \neq 1$ 、$t \neq \frac{1}{3}$。

2.6 Corollary6：Given α_0、α_1

Let $C(t)=H_1(t;\theta,\alpha_0,\alpha_1)$，if we know the value of α_0、α_1，then

$t = \frac{-\alpha_1+2\alpha_0\pm\sqrt{{\alpha_1}^2-\alpha_0\alpha_1+{\alpha_0}^2}}{3(\alpha_0-\alpha_1)}$、 $\theta = 2\tan^{-1}(\frac{\alpha_0(3t-1)}{6t})$，where

$\alpha_0 \neq \alpha_1$ 、$t \neq 0$，or $t=\frac{1}{2}$ 、$\theta = 2\tan^{-1}(\frac{\alpha}{6})$，where $\alpha_0=\alpha_1=\alpha$。

2.7 Corollary7：given θ and α_0=α_1

Let $C(t)=H_1(t;\theta,\alpha_0,\alpha_1)$，assume $\alpha_0=\alpha_1$，and give the value of θ, then $\alpha_0=\alpha_1=6\tan\frac{\theta}{2}$、$t=\frac{1}{2}$。

3 Experimental Result

From the theorem and corollaries describes in the last section, we give examples in this section.

3.1 Example 1：θ、t are given

In corollary 1, given θ、t, we can find α_0 and α_0, the result is shown in Table 3 and Figure 2. The last column of Table lists the different cases shown in Table 2.

Table 3 Findingα_0 and α_0 by giving θ, t

#	Given value		Obtained value		Cases
	θ	t	α_0	α_1	
(a)	140	0.2	-8.24243	9.41992	C
(b)	140	0.4	32.96973	12.36365	A
(c)	140	0.5	16.48486	16.48486	A
(d)	140	0.6	12.36365	32.96973	A
(e)	140	0.8	9.41992	-8.24243	B
(f)	200	0.5	-34.02769	-34.02769	D

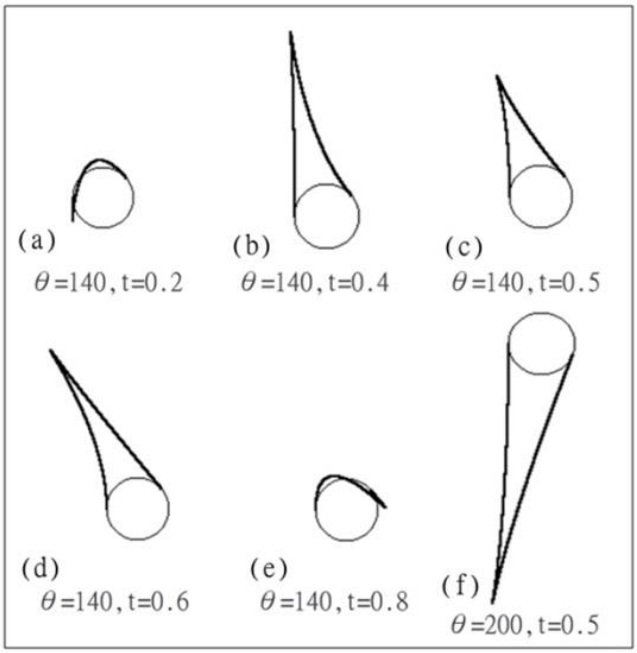


Figure 2 Fixed θ, vary t

Table 4 and Figure 3 shows the cases that we fixed the value t, and vary the value θ.

Table 4 Fixed t, θ vary

#	Given value		Obtained value		Cases
	θ	T	α_0	α_1	
(a)	50	0.4	5.59569	2.09838	A
(b)	100	0.4	14.30104	5.36289	A
(c)	150	0.4	44.78461	16.79423	A
(d)	200	0.4	-68.05538	-25.52077	D
(e)	250	0.4	-17.13778	-6.42667	D
(f)	300	0.4	-6.9282	-2.59808	D

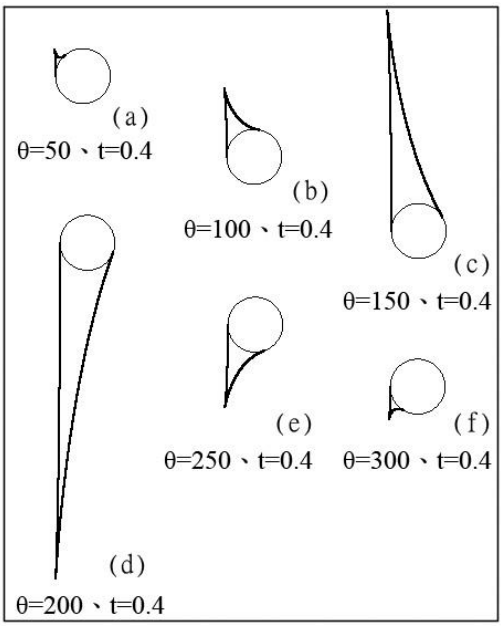
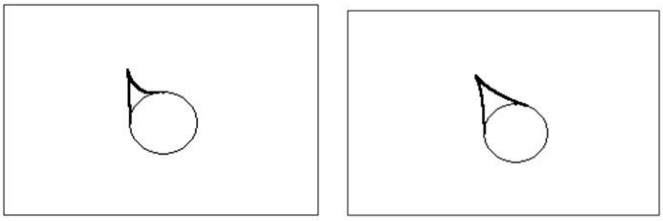


Figure 3 Fixed t, θ vary

3.2 Example 2：

Let θ=80 and α_0=10，we find t≈0.40055, α_1≈3.78032，as shown in Figure 4(a). Let θ=110, α_1=8，we find t≈0.48157、α_0≈9.27926，as shown in Figure 4(b).

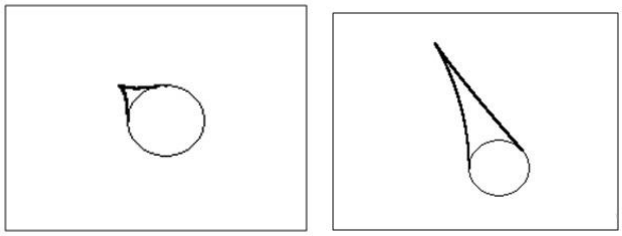


(a) (b)

Figrue 4 (θ, α_0) or (θ, α_1) are given

3.3 Example 3：

Let t=0.5、α_0=5 we find α_1=5、θ≈79.61114 as shown in Figure 5(a). Let t=0.57、α_1=25，we have α_0=13.53587、θ≈140.82257，as shown in Figure 5(b).

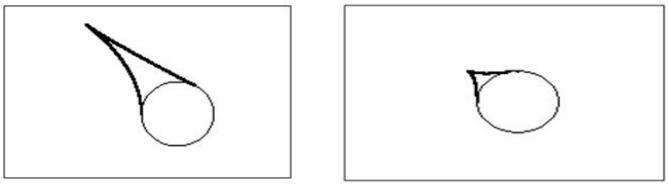


(a) (b)

Figure 5 (t, α_0) or (t, α_1) are given

3.4 Example 4：α_0、α_1 are given

Let $C(t)=H_1(t;\theta,\alpha_0,\alpha_1)$. If $\alpha_0=8$、$\alpha_1=20$，we find $t\approx-0.37321$、0.59543，when $t=0.59543$, we find $\theta\approx120.81162$，as shown in Figure 6(a). If $\alpha_0=5$、$\alpha_1=5$，we find $t=\frac{1}{2}$、$\theta\approx79.61114$，as shown in Figure 6(b)。



(a) (b)

Figure 6 (α_0, α_1) or (α_0, α_1) are given

We consider the case that $\boldsymbol{\alpha_0=\alpha_1}$ and the value of $\boldsymbol{\theta}$ are given

3.5 Example 5：

Let $C(t)=H_1(t;\theta,\alpha_0,\alpha_1)$, when the value of θ are given, we can find the value of α_0、α_1、t，we list the result in Table 5 and Figure 7. For example, when $\theta=50$，we find $\alpha_0=\alpha_1=2.79785$、$t=1/2$, as shown in Figure 7(a).

Table 5 Given θ with $\alpha_0=\alpha_1$

#	Given value	Obtained value		
	Θ	α_0	α_1	t
(a)	50	2.79785	2.79785	0.5
(b)	100	7.15052	7.15052	0.5
(c)	150	22.3923	22.3923	0.5
(d)	200	-34.02769	-34.02769	0.5
(e)	250	-8.56889	-8.56889	0.5
(f)	300	-3.4641	-3.4641	0.5

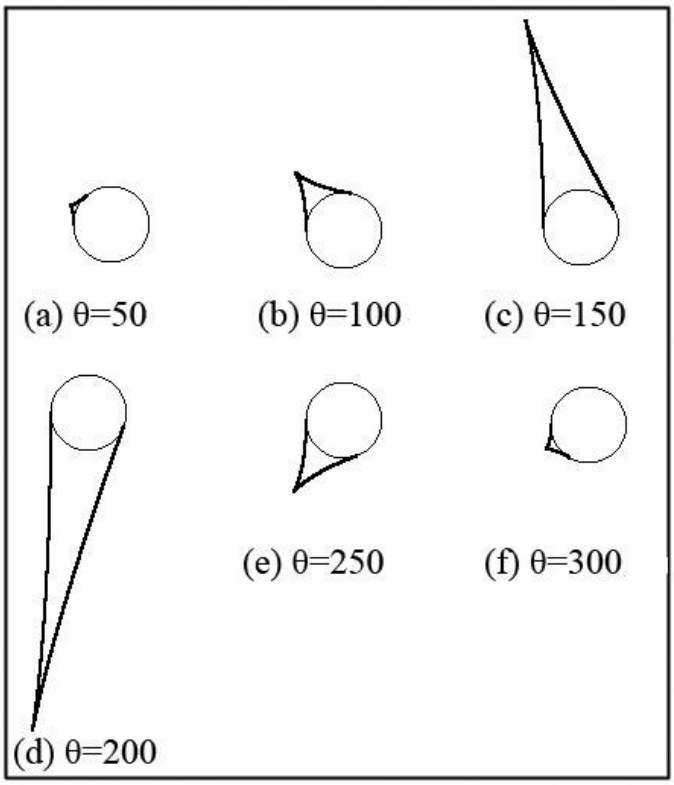


Figure 7 Given θ with $\alpha_0=\alpha_1$

4 Conclusion and future research

The unit circle hermit curve, $H_1(t;\theta,\alpha_0,\alpha_1)$, can be used to design curves and images. On many cases, the singular point may be needed to be considered. For example, the Chinese characters may have cusp on the boundary. Before the design of the character, the relationship between the parameter value and the contour of the boundary curve is important. When we give two values of the parameters for $H_1(t;\theta,\alpha_0,\alpha_1)$ curve, the can find the other two values with simple computation. When we define a curve, the only memory we need is 7 locations to store these 4 parameters, and other 3 parameters for the standardization process. So, the design process saved not only the time, but also the memory

There are more constraints we can used to design the curves. For example, the maximum curvature happened at t_0, the curve passing through a point p_0, the curve tangent to a line L_0, and so on. We believe the result will as simple as the case we introduced here.

Acknowledge

This work was supported in part by the National Science Council in Taiwan under Grants 102-2221-E-031-003.

Reference

[1] Cinque, L., Levialdi, S. and Malizia, A. "Shape description using cubic polynomial Bezier curves"; Pattern Recognition Letters, Vol. 19., pp.821-828, 1998.

[2] Yang, H.-M., Lu, J.-J. and Lee, H.-J.. "A Bezier curve-based approach to shape description for Chinese calligraphy characters". ; In "Proceedings of the Sixth International Conference on Document Analysis and Recognition", pp.276-280. (2001)

[3] Chang, H.-H. and Yan, H. "Vectorization of hand-drawn image using piecewise cubic Bezier curves fitting". ; Pattern Recognition, Vol. 31, No. 11, pp.1747-1755. (1998)

[4] Cao, H., Kot, A.C. "Lossless Data Embedding in Electronic Inks"; IEEE Transactions on Information Forensics and Security, Vol. 5, No.2, pp.314-323. (2010)

[5] Cao, L., Jia, Z., Liu, J. "Computation of medial axis and offset curves of curved boundaries in planar domains based on the Cesaro's approach". ; Computer Aided Geometric Design, Vol. 26, No.4, pp.444-454. (2009)

[6] Qin, Peiyu and Chen, Chuanbo "Simulation Model of Flower Using the Integration of L-systems with Bezier Surfaces"; International Journal of Computer Science and Network Security, Vol.6, No.2, pp.65-68. (2006)

SESSION

POSTERS

Chair(s)

TBA

Exponential dissipativity of discrete-time switched systems with time-varying delays

J.H. Park[1], **K. Mathiyalagan**[1], **S.M. Lee**[2], **J.H. Yoo**[2]
[1]Department of Electrical Engineering, Yeungnam University,
Kyongsan, Republic of Korea. Email: jessie@ynu.ac.kr.
[2]College of Information and Communication Engineering, Daegu University,
Kyongsan, Republic of Korea. Corresponding author: joonhyuk@daegu.ac.kr (J.H. Yoo)

Abstract— *In this brief, the problem of exponential dissipativity of discrete-time switched system with time-varying delays is proposed by using the average dwell time approach and multiple Lyapunov-Krasovskii functional (LKF).*

Keywords: Dissipativity, Switched system, Time-varying delays.

1. Introduction

The concept of dissipative dynamical system includes passivity as a special case, it was first introduced by Willems in 1972 [1]. Dissipative systems have wide range of applications in areas such as stability, estimation, chaos and synchronization, and robust control. The passivity framework is a concept from electrical network theory which relates the stored energy in an electrical network with the supplied energy into the network. Over the past decades, dissipativity theory turned out to be a very useful concept in systems and control theory. The problem of the robust dissipativity analysis for delayed neural networks with randomly occurring uncertainties has been investigated in [5]. The problem of dissipative control for uncertain discrete singular systems has been studied in [4]. On the other side, a switched system is composed of family of subsystems described by continuous- or discrete-time dynamics and a rule coordinating the switching between them. It is known that many real-world systems can be modeled as switched systems, for example, chemical process systems, transportation systems, computer controlled systems aircraft, air traffic control and communication systems [3]. In the last decade, substantial progress has been made for switched systems, for example, $\mathcal{L}_2$ gain control synthesis problem via off-line-type switched state feedback is proposed in [6]. Sun *et al.,* [2] has studied the problem of the exponential stability of networked control systems with large delay periods. To the best of the authors' knowledge, the problem of exponential dissipativity has not been fully investigated, which motivates our present investigation.

Our main aim is focused on obtaining the sufficient conditions for the exponential dissipative performance for switched system with time-varying delays. The results are obtained by using the multiple LKF and linear matrix inequalities (LMIs). Further, the average dwell time method is used to control the dwell time of a switching between systems.

2. Problem description

Consider the following discrete-time switched system with time-varying delays:

$$\begin{aligned}\bar{x}(k+1)&=A_{\sigma(k)}\bar{x}(k)+B_{\sigma(k)}\bar{x}(k-d(k))+D_{\sigma(k)}v(k),\\ \bar{z}(k)&=C_{\sigma(k)}\bar{x}(k)+E_{\sigma(k)}v(k),\end{aligned} \tag{1}$$

where $\bar{x}(k)$ is the state vector; $v(k)$ is the disturbance input vector in $l_2[0,\infty)$; $\bar{z}(k)$ is the output signal; A_i, B_i, C_i, D_i, E_i are known real constant matrices; $d(k)$ is the time-varying delay satisfying $0 < d_m \le d(k) \le d_M$; $\sigma(k)=i \in \mathbb{N}=\{1,2,\ldots,N\}$ is the switching signal; N is the number of subsystem.

Definition 1: System (1) is strictly $(\mathcal{Q},\mathcal{S},\mathcal{R})$ exponentially dissipative, if for a sufficiently small scalar $\beta > 0$ and every non-zero $v(k)$, for energy function $E(\bar{z}(s),v(s)) = \bar{z}^T(s)\mathcal{Q}\bar{z}(s) + 2\bar{z}^T(s)\mathcal{S}v(s) + v^T(s)\mathcal{R}v(s)$, the output $\bar{z}(k)$ under zero initial condition satisfies, $\sum_{s=k_0}^{\infty}(1-\alpha)^s E(\bar{z}(s),v(s)) \ge \beta\sum_{s=k_0}^{\infty} v^T(s)v(s)$, where $\mathcal{Q},\mathcal{S}$ and $\mathcal{R}$ are real matrices with symmetric $\mathcal{Q}$ and $\mathcal{R}$. For convenience, assume $\mathcal{Q} \le 0$, then $-\mathcal{Q} = (\bar{\mathcal{Q}}^{1/2})^2$.

3. Main results

Theorem 1: For given constants $0 < \alpha < 1$, $\mu > 1$ and matrices $\mathcal{Q}$, $\mathcal{S}$, $\mathcal{R}$, switched system (1) is globally exponentially stable when $v(k) = 0$ and strictly $(\mathcal{Q},\mathcal{S},\mathcal{R})$ exponentially dissipative, if there exist symmetric positive matrices $\hat{P}_i$, $\hat{Q}_i$, $\hat{R}_i$, X_i, and for any switching signal $\sigma(k)$ with average dwell time $T_a > T_a^* = -\frac{\ln\mu}{\ln(1-\alpha)}$, such that the following LMIs holds for any $i,j \in \mathbb{N}=\{1,\ldots,N\}$, $i \ne j$,

$$\begin{bmatrix}\Theta_{i(1,1)} & * & * & * & * & *\\ B_i^T X_i^T & -\bar{\alpha}Q_i & * & * & * & *\\ 0 & 0 & -\bar{\alpha}R_i & * & * & *\\ \Theta_{i(4,1)} & X_iB_i & 0 & \Theta_{i(4,4)} & * & *\\ \Theta_{i(5,1)} & 0 & 0 & D_i^T X_i^T & \Theta_{i(5,5)} & *\\ \mathcal{Q}^{1/2}C_i & 0 & 0 & 0 & \mathcal{Q}^{1/2}E_i & -I\end{bmatrix} < 0, \tag{2}$$

$$P_i \le \mu P_j,\ Q_i \le \mu Q_j,\ R_i \le \mu R_j, \tag{3}$$

where $\Theta_{i(1,1)} = (d_M - d_m + 1)Q_i + R_i + \alpha P_i + \mathrm{sym}(X_i(A_i - I))$, $\Theta_{i(4,1)} = P_i - X_i^T + X_i(A_i - I)$, $\Theta_{i(4,4)} = P_i - \mathrm{sym}(X_i)$, and $\Theta_{i(5,1)} = -\mathcal{S}C_i + D_i^T X_i^T$, $\Theta_{i(5,5)} - \mathcal{R} -$

sym$(E_i^T\mathcal{S})$, $\bar{\alpha}=1-\alpha^{d_M}$. Further, estimate for state decay is $\|\bar{x}(k)\|^2 < \frac{b}{a}\chi^{(k-k_0)}\|\phi\|_L^2$, where $a=\min_{\forall i\in S}\lambda_{\min}(P_i)$, $b=\max_{\forall i\in S}\lambda_{\max}(P_i)+d_M\max_{\forall i\in S}\lambda_{\max}(R_i)+d_M\max_{\forall i\in S}\lambda_{\max}(Q_i)$ and $\chi=(1-\alpha)\mu^{\frac{1}{T_a}}$.

Proof: Consider the following piecewise LKF:

$$V_i(k)=\bar{x}^T(k)P_i\bar{x}(k)+\sum_{s=k-d(k)}^{k-1}(1-\alpha)^{k-s-1}\bar{x}^T(s)Q_i\bar{x}(s)+\sum_{s=k-d_M}^{k-1}(1-\alpha)^{k-s-1}\bar{x}^T(s)R_i\bar{x}(s)+\sum_{j=-d_M+1}^{-d_m}\sum_{s=k+j}^{k-1}(1-\alpha)^{k-s-1}\bar{x}^T(s)Q_i\bar{x}(s). \quad (4)$$

Calculating $\triangle V_i(k)=V_i(k+1)-V_i(k)$ along the solutions of (1), and considering the following zero equality for $v(k)=0$ and $\eta(k)=\bar{x}(k+1)-\bar{x}(k)$:

$$2(\eta^T(k)+\bar{x}^T(k)X_i\left[(A_i-I)\bar{x}(k)+B_i\bar{x}(k-d(k))-\eta(k)\right]=0,$$

it is easy to get $\Delta V_i(k)+\alpha V_i(k)\le\xi(k)^T\Theta_i\xi(k)$, where $\xi(k)=\left[\bar{x}^T(k)\ \bar{x}^T(k-d(k))\ \bar{x}^T(k-d_M)\ \eta^T(k)\right]^T$,

$$\Theta_i=\begin{bmatrix}\Theta_{i(1,1)} & * & * & *\\ B_i^TX_i^T & -\bar{\alpha}Q_i & * & *\\ 0 & 0 & -\bar{\alpha}R_i & *\\ \Theta_{i(4,1)} & X_iB_i & 0 & \Theta_{i(4,4)}\end{bmatrix}.$$

In view of (2), we get $\Theta_i<0$. It is easy to obtain $\Delta V_i(k)+\alpha V_i(k)\le 0$, then $V_i(k+1)-V_i(k)\le-\alpha V_i(k)$, which implies that $V_{\sigma(k)}(k)\le(1-\alpha)^{k-k_t}V_{\sigma(k_t)}(k_t)$. Using the definition of average dwell time, we get $V_{\sigma(k)}(k)\le\left((1-\alpha)\mu^{\frac{1}{T_a}}\right)^{k-k_0}V_{\sigma(k_0)}(k_0)$. Also, one can get from LKF (4), that $V_{\sigma(k)}(k)\ge a\|\bar{x}(k)\|$ and $V_{\sigma(k_0)}(k_0)\le b\|\phi\|_L^2$. Then, from above two terms, and taking $\chi=(1-\alpha)\mu^{\frac{1}{T_a}}$, it is easy to get (**??**). Therefore, for time-varying delay $d(k)$ and any switching rule $\sigma(k)$ with the average dwell time T_a, one can easily obtain $\chi<1$. Hence, system (1) with $v(k)=0$ is globally exponentially stable. Next, for all non zero $v(k)$, we can get $\Delta V_i(k)+\alpha V_i(k)-E(\bar{z}(s),v(s))\le\xi_1^T(k)(\Pi_i+\Pi_{1i}^T\mathcal{Q}\Pi_{1i})\xi_1(k)$, where $\xi_1^T(k)=\left[\xi^T(k)\ v^T(k)\right]$ and

$$\Pi_i=\begin{bmatrix}\Theta_{i(1,1)} & * & * & * & *\\ B_i^TX_i^T & -\bar{\alpha}Q_i & * & * & *\\ 0 & 0 & -\bar{\alpha}R_i & * & *\\ \Theta_{i(4,1)} & X_iB_i & 0 & \Theta_{i(4,4)} & *\\ \Theta_{i(5,1)} & 0 & 0 & D_i^TX_i^T & \Theta_{i(5,5)}\end{bmatrix},\ \Pi_{1i}=\begin{bmatrix}C_i^T\\0\\0\\0\\E_i^T\end{bmatrix}^T.$$

By defining $-\mathcal{Q}=(\bar{\mathcal{Q}}^{1/2})^2$, it is easy to get (2). Hence, if matrix inequality (2) holds, we get

$$\Delta V_i(k)+\alpha V_i(k)-E(\bar{z}(s),v(s))\le 0. \quad (5)$$

For a sufficiently small scalar $\beta>0$, it can always be found such that $\Delta V_i(k)+\alpha V_i(k)-E(\bar{z}(s),v(s))+\beta v^T(k)v(k)\le 0$. To prove the exponential dissipative performance, from (5) and above inequality, we get $V_i(k_1)<(1-\alpha)V_i(k_0)+E(\bar{z}(k_0),v(k_0))-\beta v^T(k_0)v(k_0)$. Iterating the above inequality and using the switching rule, it is easy to obtain (1). Thus, from Definition 1, it is clear that system (1) has strictly $(\mathcal{Q},\mathcal{S},\mathcal{R})$ exponential dissipative performance. ■

Example: Consider system (1) with two subsystems and parameters below:

$$A_1=\begin{bmatrix}0.1 & 0.2\\ 0.5 & -0.2\end{bmatrix},B_1=\begin{bmatrix}-0.2 & -0.3\\ 0.1 & 0\end{bmatrix},D_1=\begin{bmatrix}-0.2\\-0.1\end{bmatrix},$$
$$A_2=\begin{bmatrix}-0.1 & -0.2\\ 0.4 & 0.8\end{bmatrix},B_2=\begin{bmatrix}-0.1 & 0.2\\ -0.1 & -0.2\end{bmatrix},D_2=\begin{bmatrix}-0.8\\-0.1\end{bmatrix},$$
$$C_1=\begin{bmatrix}0.4 & -0.2\end{bmatrix},\ E_1=0.5,\ C_2=\begin{bmatrix}-0.2 & 0.1\end{bmatrix},\ E_2=0.1.$$

For $1\le d(k)\le 3$, $\alpha=0.07$, $\mu=1.5$, and letting $\mathcal{Q}=-0.56$, $\mathcal{S}=-1.4$, $\mathcal{R}=3$, then we can find a LMI solution set in Theorem 1, but it is not omitted here. Also, the average dwell time obtained as $T_a>T_a^*=5.5872$, then for $T_a=6$, the decay rate is $\chi=0.9950<1$, and the state decay as $\|\bar{x}(k)\|\le 1.4443e^{-0.0025(k-k_0)}\|\phi\|_L$. It means that the considered switched system is exponentially dissipative. Simulation results are shown in Fig. 1.

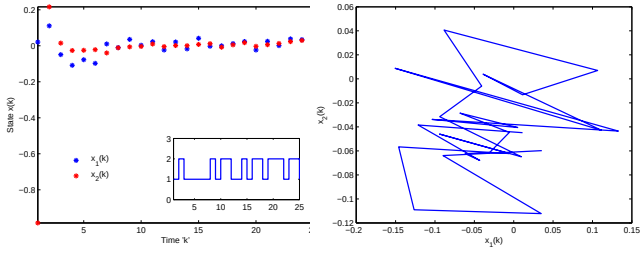


Fig. 1
STATE TRAJECTORIES AND RESPONSES.

Acknowledgements: The work was supported by Basic Science Research Program through the NRF funded by the Ministry of Education (2013R1A1A2A10005201).

References

[1] J.C. Willems, Dissipative dynamical systems, part I: General theory. *Archive for Rational Mechanics and Analysis*, 45:321–351, 1972.
[2] Y. Sun and S. Qin, Stability of networked control systems with packet dropout: an average dwell time approach. *IET Contr. Theory Appl.*, 5:47–53, 2011.
[3] L. Wu, D.C. Ho and C.W. Li, Sliding mode control of switched hybrid systems with stochastic perturbation. *Systems & Control Letters*, 60:531–539, 2011.
[4] X. Dong, Robust strictly dissipative control for discrete singular systems.*IET Contr. Theory Appl.*, 1:1060-1067, 2007.
[5] J. Wang, J.H. Park, H. Shen and J. Wang, Delay-dependent robust dissipativity conditions for delayed neural networks with random uncertainties.*Appl. Math. Comput.*, 221:710–719, 2013.
[6] D. Xie Q. Wang and Y. Wu, Average dwell-time approach to $\mathcal{L}_2$ gain control synthesis of switched linear systems with time delay in detection of switching signal. *IET Contr. Theory Appl.*, 3:763–771, 2009.

The Virtual Device Managing Module of the Metaverse Assisted Living Support System

EunJin Ko
Real & Emotional Sense Convergence Service Research Section
Electronics and Telecommunications Research Institute
Daejeon, Republic of KOREA
ejko@etri.re.kr

JongHyun Jang
Real & Emotional Sense Convergence Service Research Section
Electronics and Telecommunications Research Institute
Daejeon, Republic of KOREA
jangjh@etri.re.kr

Abstract— **The number of people who should have to need help from other people is increasing. If an IT service or solution can provide an easy way to help them, the handicapped people do more and live more comfort without other people's help. In recent years, a lot of approaches focus on the physical devices to provide many functions. By the way, there are many other people who do not use these devices or IT solutions easily cause of unusual user interface. We concentrate our efforts to find a way to use usable solutions or devices of people and focus on the virtualization of real world using metaverse server which simulates the real world and creates new cyber space. Using metaverse server, we can provide and change dynamically virtual devices and IT solutions to each person considering individual condition by relatives or system manager remotely.**

This paper presents the Metaverse Assisted Living Support System (MALSS) to support handicapped people who do not use IT service or devices easily. In Metaverse Assisted Living Support System (MALSS), there are two major function modules: one is virtual device managing module and the other is individual profile managing module. To provide individual virtual device control interface, it is needed to collect individual living pattern data and analyze it to recommend the best adaptable personal device interfaces.

In this paper, we present the virtual device managing module in the Metaverse Assisted Living Support System. The virtual device managing module gathers device profile and status data, creates virtual devices based on the individual preference managed by an authorized person or relatives remotely through metaverse client program[1].

Keywords— Assisted Living Care, Metaverse, Device Virtualization, Remote Management, IT Remote Helper

I. INTRODUCTION

The best solution to care the handicapped person at home is helper who lives at same space, watches and supports his/her life. But this is dream. To overcome this problem in IT industry, there are many home automation services including sensors, devices and monitoring solutions and real test-bed to test devices and solutions. In these approaches, there is one precondition: all participants should know the functional mechanisms of all devices and solutions and deeply understand why these services are needed [1]. Unfortunately, most of cared peoples are old and poor and they didn't use these services well. Even some of them turned off or broken devices and told the service providers that the functions of devices are hard to understand.

To prove these requirements, we design and develop the Metaverse Assisted Living Support System that is consisted of the virtual device managing module and the individual profile managing module. The many reason to use metaverse service is flexible and easy to understand. With the advantage of virtualization, we consider the way to access and control the remote site where there is person needed to care.

Normally, we use web camera to monitor the remote site. But many of them show us resistance to use web camera mentioning their privacy. So, we use metaverse service to avoid the violence of privacy problem and indirectly monitor the current status of the challenged person mapping the real situation to virtual world.

II. METAVERSE SERVICE

A virtual world is an instantiation of a metaverse and a 3D virtual space in which people interact with one another through avatars and metaverse client. This metaverse resembles the actual world but without its original limitations [2]. Recently, many people thought of the metaverse as social or gaming environments. The metaverse service provides technology capabilities that can transform education, learning, virtual project management, and conversation [3]. And the scope of the metaverse is expanding to home automation, sports game, sports simulation, and so on. In metaverse service, the important feature is an avatar. Normally, customers can control the virtual world with an action of an avatar and most projects focused on the relations between an avatar and a virtual world to provide reality without physical limitations of the real world [4].

[1] This work was supported in part by the IT R&D programs of The MISP[10043430, Development of Metaverse based Collaborative Family Safety Service Virtualization Technology for Remotely Resider Family Memebers]

[텍스트 입력]

III. The Virtual Device Managing Module of the Metaverse Assisted Living Care System

We designed the several modules in the MALSS in order to provide virtualization of remote monitoring and controlling functions and the figure 1 shows the structure of the MALSS.

To map and manage the virtual device and physical device, we divide functions of physical device into unit functions. Each unit function defined as the structured xml format which presents the pre/post conditions, input/output parameters and function description. The unit functions are stored in the repository and controlled by manager web browser. If anyone wants to create a virtual device which presents only several simple functions to challenged person remotely, he/she loads the list of unit functions from the repository, arranges the order of function operation to create a new virtual device using an authoring program for his/her cared person and tells him/her about the usage of a new virtual device remotely. In an authoring program, a creator finds the current feature of a virtual device in the left side and the list of unit functions in the right side.

After these processes, a challenged person accesses the Metaverse Assisted Living Support System using a client program and can see the virtual home/office world. In virtual world, there are several physical devices replicated from real world and several virtual devices with simple function icons managed by relatives or service managers. He/She just touches the screen of his/her display device to active the function of a virtual device and a metaverse client program transfers the action information of a customer to the metaverse server system. The metaverse server system verifies these data and translates from the information of the virtual to the data of a physical device. After that, the metaverse server system sends control data to the physical device and the physical device acts based on the received action order from the metaverse server system.

In these processes, a challenged person who is not familiar to the IT devices or solutions can easily access and use the IT devices or solutions because of simplified icon using virtual world technologies.

IV. Conclusions

In this paper, we look around the concept of assisted living care approaches and metaverse service at first. To provide metaverse living support service, there is needed to create the mapping information between the physical devices and the virtual devices. With mapping information, customers can create and modify the virtual devices with simple icons and delivery these features to some other people who have problems to use IT devices or IT solutions because of the complicated functions of them.

In the future, we will have a plan to design the living pattern collecting module and the recommendation the preference function icons of the virtual devices based on the analyzing the living pattern of customers through the living pattern collecting module.

References

[1] Sangwook Park, Noh-Sam Park, Jung-Tae Kim and Eui-Hyun Paik, "Provising of the Express Multisensory Adaption Platform for Heterogeneous Multimedia Devices in the Ubiquitous Home, pp 126-131, Feb. 2009

[2] W.S. Bainbridge, "The Scientific 1. Research Potential of Virtual Worlds," Science, pp. 472–476, Jul. 2007.

[3] Dawn Owens, Alanah Davis, John D. Murphy, Deepak Khazanchi, and Ilze Zigurs, "Real-World Opportunities for Virtual-World Project Management y," IEEE Computer Society, pp. 34-41, Mar. 2009.

[4] W.S. Bainbridge, "The Scientific 1. Research Potential of Virtual Worlds," Science, 27 July 2007, pp. 472–476.

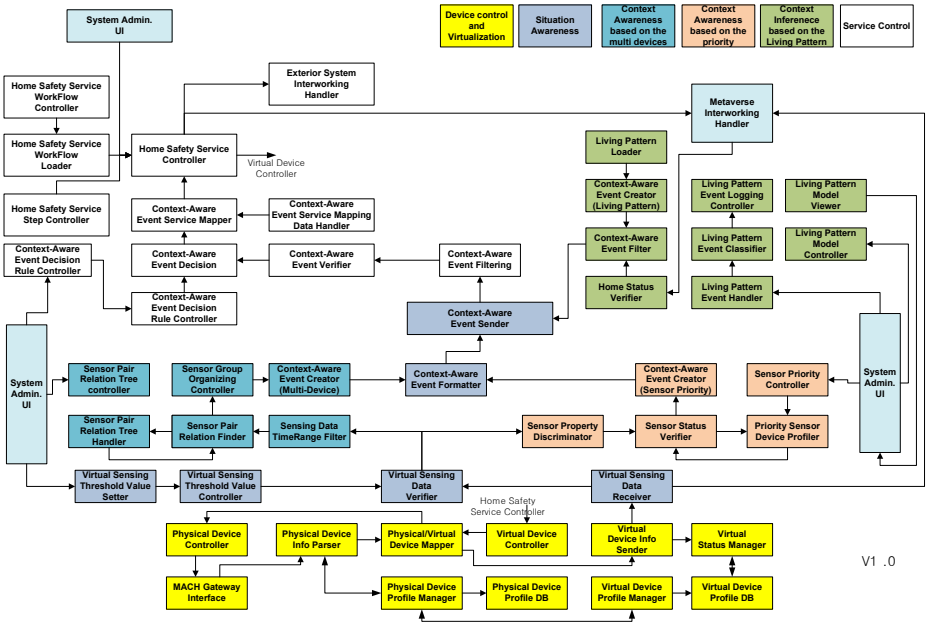


Figure 1. The Structure of the MALSS

[텍스트 입력]

Real-time Radar Data Visualization

Giti Javidi[1], Ehsan Sheybani[1], Danielle Mason[1], Akbar Eslami[2] and Jamiiru Luttamaguzi[2]
[1]College of Engineering and Technology, Virginia State University, Petersburg, VA
[2]Department of Engineering, Elizabeth City State University, Elizabeth City, NC

Abstract - *To have a fully functional FMCW X-band radar for the SMARTLabs ACHIEVE trailer, it is necessary to (a) produce code to retrieve data from an FPGA board linked to the radar, (b) calculate Fourier transforms and (c) display the power spectrum in near-real time using a computer code based on freely available scientific development tools. So that the communication between the FPGA board and the computer is reliable and accurate, developing a specific format through the use of C was an initial step. This was followed by the development of a method to visualize data efficiently. In this case, Python, along with its matplotlib, SciPy, and NumPy modules, were used. Both programs were then integrated together within a graphical user interface.*

Keywords: Radar, Processing, FPGA, Fourier transform, Python.

1 Introduction

The use of radar is important in the detection of both stationary objects, such as buildings, and moving objects, such as clouds and aircraft [1, 2]. Radar has varying levels of frequency with the measurements ranging anywhere from megahertz to gigahertz [5]. An X-Band radar frequency can range anywhere between 8 and 12.5 GHz [6]. The radar that is focused on throughout the course of this project was an X-Band, frequency modulated continuous wave (FMCW) radar with frequency of 10 GHz. FMCW radar utilizes frequency modulation of a continuous signal to acquire range information. One prior example of such a radar includes the PILOT radar which was "used by warships for navigation where the ability to perform accurate navigation in poor weather is essential for the accomplishment of the ships' tactical mission" [7]. Another example is the use of scanning X-band radar, paired with FMCW K-band radar, in an experiment by Joel Van Baelan, Frederic Tridon, and Yves Pointen to retrieve accurate rainfall rate estimates [9].

Benefits of this type of radar are that it is not just safe and inexpensive, but also serves as a means in filling in gaps of higher powered pulse-doppler radars when used in conjunction with them [3]. This proves especially important for SMARTLabs (Surface-based Mobile Atmospheric Research and Testbed Laboratories), which consists of the three mobile laboratories [8]. The mobile trailers are SMART (Surface-sensing Measurements for Atmospheric Radiative Transfer), COMMIT(Chemical, Optical & Microphysical Measurements of In-situ Troposphere), and ACHEIVE (Aerosol-Cloud-Humidity Interaction Exploring & Validating Enterprise). The SMART trailer utilizes active and passive sensors to collect data on the atmosphere, gaining more knowledge about surface energy. The focus of the COMMIT trailer is to collect and measure information on the microphysics of aerosols, such as particle size, in addition to gathering information about optical properties of aerosols, such as absorption and scattering. For the focus of this paper, the ACHEIVE trailer will be referenced. The purpose of the ACHEIVE trailer is to further the understanding of aerosol-cloud interactions by being able to probe cloud properties for SMARTLabs as a whole. Within the trailer, there are three different radars that work together: the W-band (94 GHz), the K-band(24 GHz), and the X-band. To successfully use this X-Band radar. It was important to first effectively retrieve data in a proper format and then visualize the data in an efficient and reliable manner.

2 Methods

Prior to the start of this project, there were a few steps that had been completed already in regards to preparing the radar-computer communication. One such step included development and installation of a packet sniffer [4] program to the Field Programmable Gate-Array (FPGA) board. A packet sniffer program is "a network monitoring tool that captures data packets and decodes them using built-in knowledge of common protocols" ("packet"). In this case, the packet sniffer program (named Xserver program) was written in C and accepts data packets with the size of 416 bytes, with 16 bytes of header and 400 bytes of actual data. Because data coming from the FPGA board is in a 24-bit format and the data that the Xserver and plotting program expect is in a 32-bit format, the data needed to be properly converted by way of another program or additional code added to Xserver.

3 Summary and Conclusions

Results that have been found at the time of this writing have shown that while the methods produced are functional, there are several areas that can be improved or further integrated. For instance, there were numerous times where the speed was not fast enough, such as plotting the data. As a compromise to plot the data in a way that is close to real-time, only a fraction of the data is plotted to the screen. The

reason for that is because every two seconds of data displayed to the screen is an average of two seconds of data, instead of it being a second for second plot. Because the computer that was being used had a small amount of memory and a slower processor, this method had to be done or else the plot would run slower. Also, instead of having both of the programs centralized to the one GUI window, all of the programs run outside of the GUI in their own separate windows. This was not a desired effect because it leaves some disorganization for the user even though the programs do accomplish the task intended. This was proven to be true during the field calibration tests conducted at the Wallops Island Facility in Virginia.

4 References

[1] Barrick, Donald E. "FM/CW Radar Signals and Digital Processing." *National Oceanic and Atmospheric Administration*. (1973): 1-22. Print.

[2] Cohen, Danny. "On holy wars and a plea for peace." *Computer Magazine* 14.10 (1981): 48-54. Web. 17 July 2012.

[3] Gabriel, Philip M., Philip W. Lee, and Graeme L. Stephens. "The Development of an Inexpensive, X-Band High-Spatial Resolution, Frequency-Modulated, Continuous-Wave Radar." 2009.

[4] "packet sniffer." The Free On-line Dictionary of Computing. Denis Howe. 12 Jul. 2012. <http://dictionary.reference.com/browse/packet sniffer>.

[5] Rashed, Md. Golam and Raquib Ahsan. "Python in Computational Science: Applications and Possibilities." *International Journal of Computer Applications* 46.20 (2012): 26-30. Web. 17 July 2012.

[6] Skolnik, Merrill I. "Chapter 1: An Introduction to Radar." *Radar Handbook.*(1980): 1-21. Print.

[7] Stove, A.G. "Linear FMCW radar techniques." *IEE Proceedings-F*. 139.5 (1992): 343-350. Web. 16 July 2012.

[8] Tsay, Si-Chee, Philip M. Gabriel, Colby S. Goodloe, Richard A. Hansell, Q. Jack Ji, Can Li, and Peter Pantina. "SMARTLabs: A Ground-based Mobile Supersite/Network Facility: Continuing Enrichment of NASA Earth Science Missions."

[9] Van Baelen, Joel, Frederic Tridon, and Yves Pointin. "Simultaneous X-band and K-band study of precipitation to derive specific Z-R relationships." *Atmospheric Research* 94 (2009): 596-605. Web. 16 July 2012.

Simulated Spaceborne Numerical Convergence Study

Akbar Eslami[1], **Jamiiru Luttamaguzi**[1], **Ka'Ren Byrd**[1], **Giti Javidi**[2], **Ehsan Sheybani**[2]
[1]Department of Engineering, Elizabeth City State University, Elizabeth City, NC
[2]College of Engineering and Technology, Virginia State University, Petersburg, VA

Abstract - *This paper describes a NASA internship case study in which the outcome can be implemented in a classroom setting. Through this case study, the students can learn numerical interpolation and integration of functions in real world project as well as the error analysis. A tradeoff expected to be observed is between the speed of obtaining results and getting accurate results. As a result, recommendations are expected for various data sets to ensure fast accurate results. The simulation can also be carried out using Matlab. The case study discusses numerical convergence of simulated space-borne microwave radiometer measurements of earth brightness temperatures so as to get fast results. The results are obtained by numerically evaluating a double integral. The integral relies on antenna pattern measurements and observed brightness temperature distribution over the earth's surface. Accurate antenna temperatures are obtained by modifying the step sizes while getting faster results. The accuracy of the numerical methods is analyzed and recommendations are given to improve the process. Such recommendations will be seen to vary for different data patterns. The study will also include antenna theory to understand its parameters, and antenna equations that affect the accuracy of the results as well as antenna equipment, radiation patterns and radiation propagation.*

Keywords: *antenna, temperature, directivity, beam solid, parameter calculation.*

1 Introduction

The main goal of this project is to enable students to realize the applications of mathematic and numerical techniques in antenna theory. Students that have a background in mathematic topics such as log scale, frequency, wavelength, spherical coordinates, integrals of several variables, and numerical integration can be introduced to antenna theory and measurements.

Antennas are devices used to efficiently transmit and/or receive electromagnetic waves. They serve as interface between wireless channels and circuits. Most antennas reversibly link radiation fields to currents flowing in wires at frequencies ranging from sub-audio through the far-infrared region [1]. Each antenna is designed for a certain frequency band. Beyond the operating band, the antenna rejects the signal. Antenna is an essential part in a communication system that is installed in satellite to receive the signals. Special instruments inside the satellite use these signals to measure air temperatures, soil moisture, wind, water currents, and much more. One of the instruments installed in a specific satellite is the Soil Moisture Active Passive (SMAP) mission that observes soil moisture and freeze/thaw state from space to improve estimates of water, energy and carbon transfers between the land and the atmosphere [2]. This will reduce uncertainties in quantifying the global energy and carbon balance. Areas directly addressed by SMAP are weather and climate forecasting, droughts, floods, agricultural productivity, human health, and national security. These areas, as well as others, help to improve the understanding of linkages between water, energy and carbon cycles. Data for measurements in SMAP are carried out by antennas. Specifically the SMAP concept utilizes L-band radar and radiometer instruments sharing a rotating 6-m mesh reflector antenna to provide high-resolution and high-accuracy global maps of soil moisture and freeze/thaw state every two to three days [3, 4].

2 Methodology

The recent development in computational capabilities, along with increased software reliability, made the numerical method and simulation approach more favorable. Examples of radiation patters can be used to evaluate the integrals that reflect different kinds of antennas such as traditional versus a focused antenna. The study includes the numerical evaluations of antenna temperature, directivity, and beam solid angle.

3 Module Implementation

This paper can be used as a three-week module to apply application of mathematics in antenna theory. The module can be presented to students with a background in Calculus III, Numerical Methods, Complex Variables, and Physics I. It can be presented in a weekly basis as follows:

Week 1: Special emphasis can be paid to define antenna and the fact that how an antenna transfer power to/from a receiver/transmitter (energy is contained in voltages and currents) into electromagnetic radiation (where the energy is contained in the E- and H-fields) travelling away from the antenna. Types of antennas can be introduced such as wire antennas, aperture antennas, reflector antennas, lens antennas, and etc. Finally the question "Why Antennas Radiate?" is to be addressed.

Week 2: Topics in surface spherical coordinates integration, numerical interpolation, and numerical integration such as Simpsons and midpoint rules need to be

reviewed. Terminology is introduced such as frequency and frequency bands, radiation patterns, fields, gain, directivity, antenna temperature and antenna efficiency.

Week 3: Antenna measurement formulas finally are introduced and their calculations are done. The solution to some formulas can be obtained exactly or numerically. Numerical integration schemes will be based on both one-dimensional and two-dimensional independent variable data. Graphical representation of measured quantities such as radiation pattern and gain are presented in software such as Mathematica or Matlab. Feedback from the students can be obtained via exercises that test ability to carry out calculations to get directivity, beam solid angle approximations, and radiation efficiency.

4 Conclusion

This paper presents a novel method to introduce antenna theory and antenna parameter calculations to Engineering, Technology, Science, and Mathematics students. The topics in spherical coordinate systems, double integrals, numerical interpolation, and numerical integration have been applied to yield the results. In particular, antenna patterns, antenna temperature, directivity, and beam solid angle are introduced for various antennas. Antenna pattern data that is collected from various earth topologies are used to measure the resulting earth heat pattern.

5 References

[1] David H. Staelin, Ann W. Morgenthaler and Jin Au Kong, *Electromagnetic Waves*, Prentice Hall, 1993.

[2] Entekhabi, D., Njoku, E. G., & O'Neill, P. E. (2010), *The Soil Moisture Active Passive (SMAP) mission*, Proceedings of the IEEE, 98(5), 704-716.

[3] Piepmeier, J. R., Long, D. G., & Njoku, E. G. (2008),*Stokes Antenna Temperatures*, IEEE Transactions on Geoscience and Remote Sensing, 46(2).

[4] Fawwaz T. Ulaby, Richard K. Moore, Adrian K. Fung, *Microwave Remote Sensing: Active and Passive Reading*, MA: Addison-Wesley, 1986.

SESSION

LATE BREAKING PAPERS AND POSITION PAPERS: MODELING, SIMULATION AND VISUALIZATION METHODS

Chair(s)

Prof. Hamid R. Arabnia

Simulink Model of a Full State Observer for a DC Motor Position, Speed, and Current

Louiza Sellami, Department of Electrical and Computer Engineering

US Naval Academy, Annapolis, MD, 21402, USA

***Abstract*— In this paper we develop a state observer model for the armature of a DC motor based on the well-known equivalent circuit model, and torque and speed equations. Towards this end, and as a first step, we derive a state space representation for the circuit model, and demonstrate its controllability and observability properties. Using the Luenberger full state observer technique, we derive and implement the latter in MATLAB/Simulink for position control of the motor, and verify its operation.**

***Keywords*—Circuit Model, State Space, DC Motors, Armature, Rotor, Luenberger Sate Observer, Simulink.**

I. INTRODUCTION

DC motors are classified into two categories: the permanent magnet type and the electromagnet type, based on how the magnetic field is created. The latter category is further subdivided into self-excited and separately excited, depending on whether there is a physical connection or not between the field windings and the armature windings. If the two windings are connected in series, this is referred to as a series motor. These are known for their variable speed and high starting torque. Applications include cranes, conveyors, elevators, and electric locomotives. If the two windings are connected in parallel, this gives rise to a shunt DC motor, which has a fairly constant speed and a medium starting torque [1]. These are used in fans, pumps, controlled fabrication machines, automated equipment such as industrial robots, and smart printers and plotters. In such applications, it is imperative that the predetermined position be acquired from the preceding position within a short period of time. Hence it becomes necessary to control the input voltage supplied to the motor by continuously detecting the position and speed of the rotor shaft.

An observer is a dynamic system that is used to estimate the state of a system or some of the states of a system. A full-state observer is used to estimate all the states of the system. The observer can be designed as either a continuous-time system or a discrete-time system. The characteristics are the same, and the design processes are at least very similar and in some cases identical. The purpose of the observer is to generate an estimate of the state based on measurements of the system output and the system input. The input and output signals are assumed to be exactly measurable. Also, the observer uses a mathematical model of the state space realization of the system, and is software implemented [2]-[4].

In this paper a full state observer is designed for a DC motor, based on the actual electrical equivalent circuit of the armature winding and the relationship between position and voltage. The observer is simulated via MATLAB/Simulink and the results and performance are compared with those of the actual system.

The paper is organized as follows. First the theory for the full observer is presented in section two. In sections three the armature electrical circuit is presented and the state space representation is derived. In section four, the design of the observer for the position is carried out. In section five we present a Simulink implementation of the system, as well as the simulation results.

II. STATE OBSERVER THEORY

A. The Physical System

The assumptions here are that the real system is a deterministic, linear time-invariant (LTI), continuous control system that is observable and controllable, and whose internal states may not be determined by direct observation. Its dynamics are described by the following state space equations:

$$\dot{x}(t) = Ax(t) + Bu(t) \quad (1)$$

$$y(t) = Cx(t) + Du(t) \quad (2)$$

Where x is the state vector, u the control input, y the output, and A, B, C, D are constant system matrices of appropriate dimensions.

B. System Controllability and Observability

A system is said to be controllable if there exists a control input that transfers any state of the system to zero in finite time. It can be shown that a LTI system is controllable if and only if its controllability matrix, given in (3), has full rank, i.e. its rank is equal to the number of states [5]. Note that the rank of the controllability matrix of an LTI system can be readily

determined in MATLAB using the commands *rank(ctrb(A,B))* or *rank(ctrb(sys))*.

$$CO = [B \mid AB \mid A^2B \mid \ldots \mid A^{n-1}B] \tag{3}$$

All the state variables of the system may not be directly measurable if, for instance, one or more components of the system is in an inaccessible location. In these cases it is necessary to estimate the values of the unknown internal state variables using only the available system output.

A system is said to be observable if the initial state, $x(t_0)$, can be determined from the system output, $y(t)$, over some finite time $t_0 \le t \le t_f$. Mathematically, a LTI system is observable if and only if the observability matrix, given in (4), has full rank, i.e. its rank is equal to the number of states [5]. Note here also that in MATLAB this can easily be checked by the command *rank(obsv(A,C))* or *rank(obsv(sys))*. Also, it is worth mentioning that controllability and observability are dual concepts. A system (A, B, C, D) is controllable if and only if the system (A^T, C, B^T, D) is observable. Here A^T and B^T are the transpose matrices of A and B, respectively. This fact will be useful when designing an Observer [5][6].

$$OB = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-1} \end{bmatrix} \tag{4}$$

C. Full Observer Model

There are several ways to derive the state equations for the full-state observer. One approach is to model the observer state equations as a model of the actual system plus a correction term based on the measured output and the estimate of what that output is expected to be. With the actual system described by (1) and (2), the observer is modeled as [2][6]

$$\dot{\hat{x}}(t) = A\hat{x}(t) + Bu(t) + L[y(t) - \hat{y}(t)] \tag{5}$$

$$\hat{y}(t) = C\hat{x}(t) + Du(t) \tag{6}$$

where L is the n × m gain matrix for the observer. The state equation in (5) is seen to model the actual state equation (1), with the true state, $x(t)$, replaced by the estimate, $\hat{x}(t)$, and a correction term which is the difference between the actual measured output $y(t)$ and its estimate $\hat{y}(t)$. Similarly, the output equation in (6) is also seen to be a model of the system's output equation, with $x(t)$ replaced by $\hat{x}(t)$.

Substituting (6) in (5) yields the following alternative form for the observer:

$$\dot{\hat{x}}(t) = (A - LC)\hat{x}(t) + (B - LD)u(t) + Ly(t) \tag{7}$$

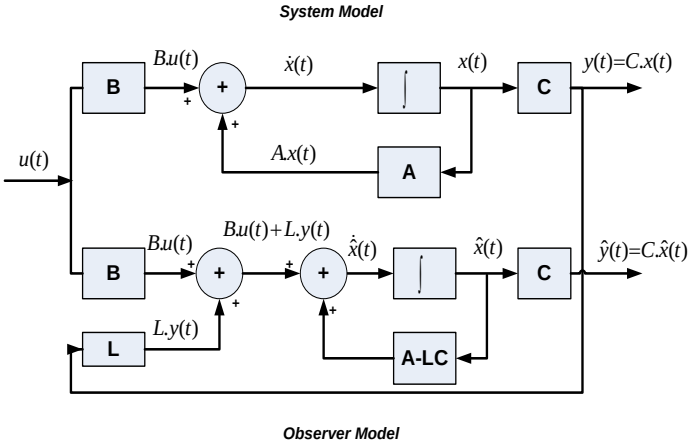


Figure 1: Block Diagram of System and Observer Models.

Note that although the matrix D explicitly appears in (7), it has no bearing on the state estimate produced by the observer. The reason is because in (5) the term $Du(t)$ cancels out in $y(t) - \hat{y}(t)$. The block diagram for the system described by (1) and (20), and its corresponding observer described by (6) and (7) are shown in Figure 1 for the case D=0.

D. Error Estimation

The purpose of the observer is to produce an estimate of the true state $x(t)$ of the real system. It is reasonable to assume that there will be some error in the estimate at the initial time, but it is hoped that the error would decrease over time. The estimation error is defined as

$$e(t) = x(t) - \hat{x}(t) \tag{8}$$

Using (1) and (7), it can easily be shown that this estimation error signal satisfies the differential equation

$$\dot{e}(t) = (A - LC)e(t) \tag{9}$$

Thus, the state equation for the estimation error is a homogeneous differential equation governed by the n×n matrix $A - LC$. The solution to this equation is

$$e(t) = e(0)\,\mathrm{e}^{-(A-LC)t} \tag{10}$$

The eigenvalues of the matrix (A – LC) can be made arbitrary by appropriate choice of the observer gain, L, when the pair (A,C) is observable. So the observer error e(t) goes to zero as t goes t infinity. If the gain matrix L is chosen so that the eigenvalues of $A - LC$ are strictly in the left-half of the complex plane, then the error equation is asymptotically stable, and therefore the estimation error will decay to zero over time. Also, if the system (A, C) is completely observable, then L can be chosen so that the eigenvalues of $A - LC$ are

placed at arbitrary locations in the plane, provided that complex eigenvalues occur in complex conjugate pairs.

E. Computation of Gain Matrix L

The gain matrix L of the full-state observer can be computed using any of the methods used to compute the control gain matrix K for a control system [5][7]. For the control problem with full-state feedback, the closed-loop system matrix of interest is $A-BK$. Comparing that with the observer problem, the closed-loop system matrix is $A-LC$. The structure of those two matrices is similar; only the order of the unknown matrix differs between BK and LC. Since the eigenvalues of a matrix and its transpose are the same, the observer problem can be formulated the same way as the control problem by considering the transpose matrix $(A-LC)^T = A^T - C^T L^T$. Therefore, the gain matrix L can be computed using the Row-Reduced Echelon (RRE) method, Singular Value Decomposition (SVD), or the MATLAB *place* function in the same way as the control gain matrix K by replacing (A,B) by (A^T, C^T). By doing this, the result from any of these methods will give the matrix L^T.

III. ARMATURE CIRCUIT MODEL

A common actuator in control systems is the DC motor. It directly provides rotary motion and, coupled with wheels or drums and cables, can cause translational motion of another machine. In such a motor (separately excited DC motor), the field windings are excited by a DC current in order to create a magnetic field. In turn, the armature windings receive current from a separate DC source which results in the creation of a torque by Lenz's Law and a back electromotive force (EMF) by Faraday's law [1].

A. Motor Equations

The electric equivalent circuit of the armature and the free-body diagram of the rotor are shown in the Figure 2, where R_A and L_A are the equivalent resistance and inductance, respectively, of the armature winding, i_A the armature current, V the input voltage, E_A the induced back electro-motive force (emf) created as a result of injecting a current into a magnetized coil.

It is assumed that the input of the system is the voltage *V* applied to the motor's armature, while the output is the position θ of the shaft. It is further assumed a viscous friction model, that is, the friction torque is proportional to shaft angular speed. Referring to Figure 2, the corresponding governing Kirchoff's voltage law and Newton's second law equations are given by (11) [1].

In general, the developed torque generated by a DC motor is proportional to the armature current and the strength of the magnetic field. Here we assume that the magnetic field is constant and, therefore, the motor torque is proportional to the armature current i_A as shown in (12). This is referred to as an armature-controlled motor.

Symbol	Unit	Definition
V	*Volts (V)*	Input voltage
i_A	*Ampere (A)*	Armature current
E_A	*Volts (V)*	Back EMF
R_A	*Ohm (Ω)*	Armature Winding Resistance
L_A	*Henry (H)*	Inductance of Armature Windings
K_m	*Volts/radians/s*	Machine Constant
T_d	$N.m$	Developed Torque
θ	*Radians*	Shaft angular position
$\omega = \dot{\theta}$	*Radians/s*	Angular speed
$\ddot{\theta}$	$Radians/s^2$	Angular acceleration
J_m	$kg.m^2$	Moment of Inertia
B_m	$N.m.s$	Viscous Frictional Constant
T_L	$N.m$	Load Torque

Table1: Motor Parameters and Constants.

$$T_d = K_m i_A \tag{11}$$

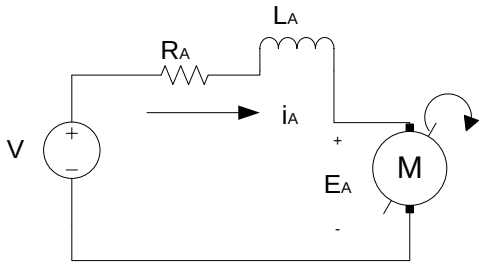


Figure 2: Electrical Equivalent Circuit of Motor Armature.

Here K_m is the machine constant. Also, the back emf, E_A, is proportional to the angular velocity of the shaft

$$E_A = K_m \omega_m = K_m \dot{\theta} \tag{12}$$

Referring to Figure 2, the corresponding governing Kirchoff's voltage law and Newton's second law equations are given by

$$V = i_A R_A + L_A \frac{di_A(t)}{dt} + K_m \dot{\theta} \tag{13}$$

$$J_m \ddot{\theta} = -B_m \dot{\theta} + K_m i_A - T_L \tag{14}$$

Where J_m is the moment of inertia of the rotor and B_m the motor viscous frictional constant [6][9].

B. Transfer Function

We take equations (13) and (14) as a basis for deriving two transfer functions for the motor under no load conditions, i. e, $T_L = 0$, with input being the voltage and the output the angular speed for the first one and the position for the second. Taking the Laplace transforms of (13) and (14) [8] gives:

$$I_A(s)(R_A + L_A s) + [V(s) - sK_m\theta] = 0 \quad (15)$$

$$s^2 J\theta + sB_m\theta = K_m I_A(s) \quad (16)$$

Solving for $I_A(s)$ from (15) and substituting in (16) yields

$$\frac{\theta(s)}{V(s)} = \frac{K_m}{s[(sJ_m + B_m)(L_A s + R_A) + K_m^2]} \quad (17)$$

and

$$\frac{\dot{\theta}(s)}{V(s)} = \frac{K_m}{(sJ_m + B_m)(L_A s + R_A) + K_m^2} \quad (18)$$

C. State Space Representation

By defining the state vector x, the output y and the input u as follows:

$$x = \begin{bmatrix} i_A \\ \theta \\ \dot{\theta} \end{bmatrix}, y = \theta, u = V \quad (19)$$

the state space equations for the motor are derived using (13) and (14) as follows:

$$\frac{d}{dt}\begin{bmatrix} i_A \\ \theta \\ \dot{\theta} \end{bmatrix} = \begin{bmatrix} -\frac{R_A}{L_A} & 0 & -\frac{K_m}{L_A} \\ 0 & 0 & 1 \\ \frac{K_m}{J_m} & 0 & -\frac{B_m}{J_m} \end{bmatrix}\begin{bmatrix} i_A \\ \theta \\ \dot{\theta} \end{bmatrix} + \begin{bmatrix} \frac{1}{L_A} \\ 0 \\ 0 \end{bmatrix} V \quad (20)$$

and

$$y = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} i_A \\ \theta \\ \dot{\theta} \end{bmatrix} \quad (21)$$

IV. Observer Design for DC motor position

For the Observer design, we consider a motor with the following parameters:

R_A	L_A	K_m	J_m	B_m
1	10^{-3}	0.1	5.10^{-3}	10^{-4}

Table 1: Values of Motor Constants and Parameters

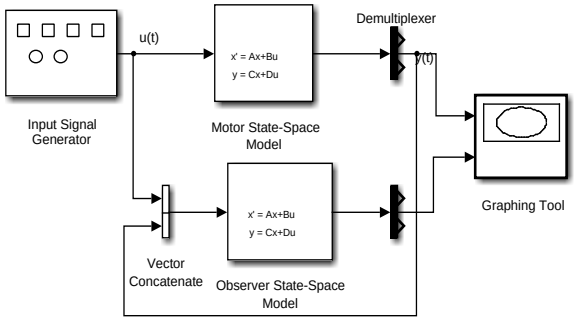


Figure 3: Simulink Block Diagram for the DC Motor Model and the Full Observer Model.

Then the systems matrices A, B, C (D=0) are constructed using (20 and (21), and observability and controllability are checked using MATLAB. The Observer is designed by calculating the matrix L such that the eigenvalues of the matrix (A-LC) are placed at -500+j250, -500-j250, and -200 respectively, and calculating the corresponding matrices "A", "B", and "C" for the observer model using (5) and (6).

V. Simulink simulation Results

The Simulink block diagram for the system and the Observer is shown in Figure 3. The input signal generator block generates the signal u(t) which serves as the excitation voltage for the motor, modeled by the motor state-space block which produces the state x and y as its output. The system output y, being equal to the second component of the state vector, is extracted from x using a de-multiplexer block. Both the motor input u(t) and the output y serve as inputs to the observer system, as shown in Figure 3. Note that, since the Simulink state space block requires the system it simulates to be in the form of equations (1) and (2), the observer equation (7) needs to be reformulated to match (1), as follows:

$$\dot{\hat{x}}(t) = \hat{A}\hat{x}(t) + \hat{B}\hat{u}(t) \quad (22)$$

$$\hat{y}(t) = \hat{C}\hat{x}(t) + \hat{D}\hat{u}(t) \quad (23)$$

Where the observer matrices can be calculated from the system matrices, as given below:

$$\hat{A} = A - LC \quad \hat{B} = [B \vdots L]$$
$$\hat{C} = C \quad \hat{D} = [D \vdots D] \quad (24)$$

and $\hat{u}(t) = \begin{bmatrix} u(t) \\ y(t) \end{bmatrix}$.

Two simulations were run for two different input voltages: a pulse of amplitude 100V, and a sinusoid of 100V amplitude and 60Hz frequency, with an initial state of [0, 0, 0]. The

actual states, namely the current i_A, the angular position θ, the angular speed $\dot{\theta}$, and their estimates are plotted in Figure 4 for a time frame of 5*sec* for the pulse, and in Figure 5 for the sinusoid. The plots of Figure 5 show that the state estimates almost match the actual states, whereas Figure 5 indicate that in the sinusoidal case, though initially and up to about 15 *msec* after the start of the simulation, the estimates diverged from the actual quantities being estimated, they did converge very quickly after the initial 15 *msec*, thus verifying the design.

VI. Conclusion

An asymptotic algebraic state estimation method known as Luenberger Observer model has been successfully applied to estimate the current, position, and angular speed of a motor. Further, examining the performance of such an observer shows that this method provides satisfactory estimates even in the presence of noise levels, and different initial conditions.

References

[1] Chapman, S. J., *Electrical Machinery and Power System Fundamentals*, McGraw Hill, 2002.

[2] Luenberger D. G., "An introduction to observers," IEEE Transaction on Automatic Control, 1971; AC-16: 596-602.

[3] Clark, R. N., Fosth, D. C., Walton, V. M., "Detecting Instrument Malfunctions in Control Systems," IEEE Transactions on Aerospace and Electronic Systems, Vol. 11, No. 4, July 1975, pp. 465-473.

[4] Heredia, G., Ollero, A., Bejar, M., and Mahtani, R., "Sensor and Actuator Fault Detection in Small Autonomous Helicopters," Mechatronics, Vol. 18, pp. 90-99, 2008.

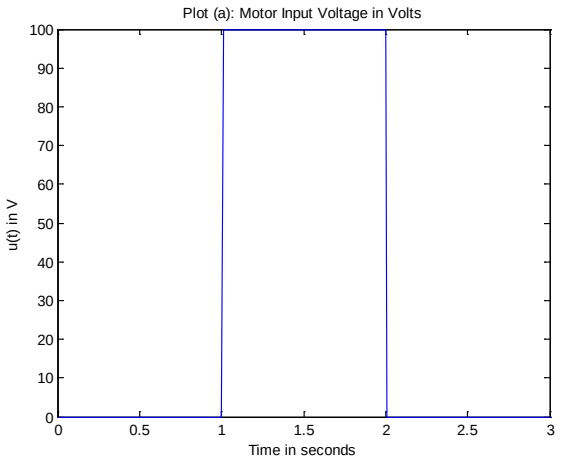
[5] T. Kailath, Linear Systems, Prentice Hall, 1980.

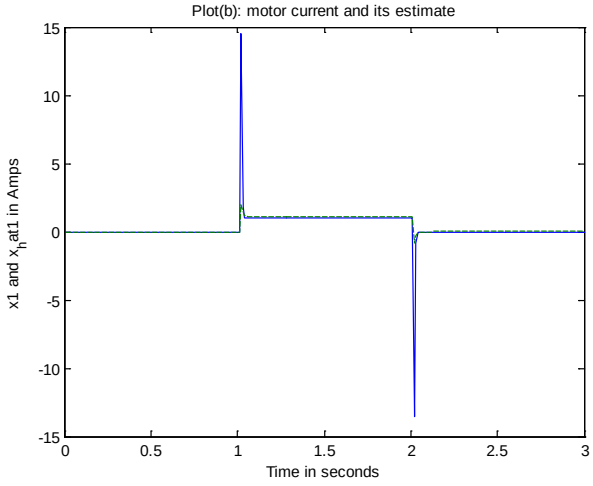
[6] H. Temeltas, G. Aktas, *Friction State Observation in Positioning Systems using Leunberger Observer*, Proceedings of the 2003 IEEBASME International Conference on Advanced Intelligent Mechatronics (AIM 2003).

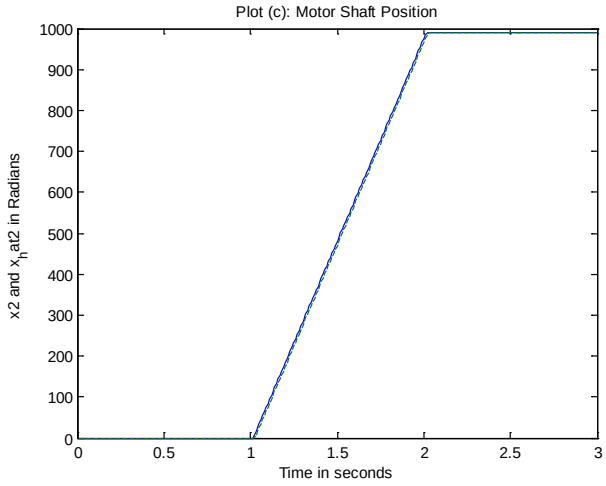
[7] Friedland B., *Control system design- an introduction to state space methods*, New York: McGraw-Hill, 1986.

[8] H. Olsson, K. AmstrÄom , C.C. de Wit, and Fric-W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993.

[9] V. Feliu, K. S. Rattan, and H. B. Brown, "Control of flexible arms with friction in the joints," IEEE Trans. on Robotics and Automation,Vol.9, N.4, pp.467{475 (1993).







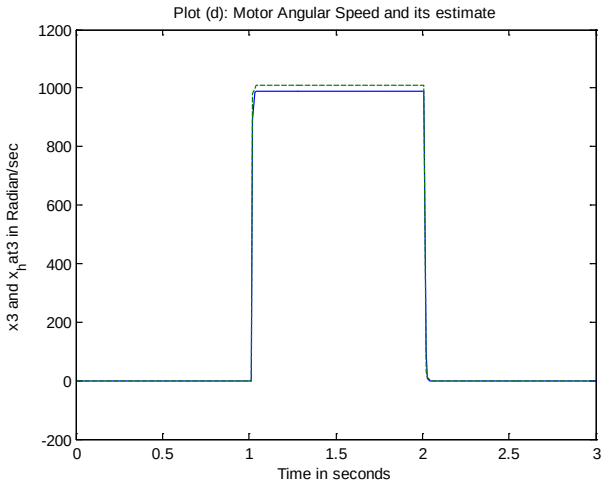
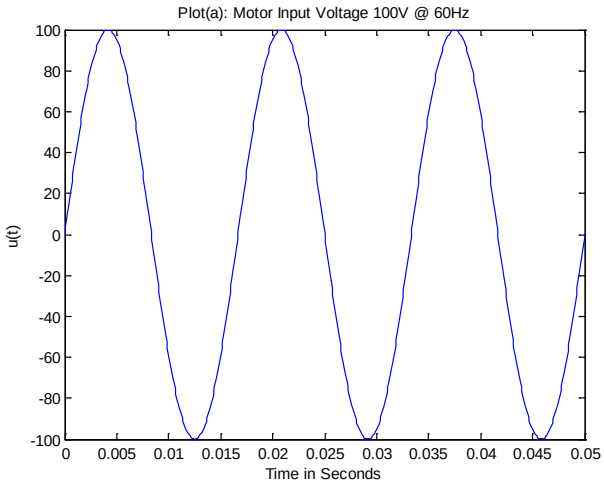
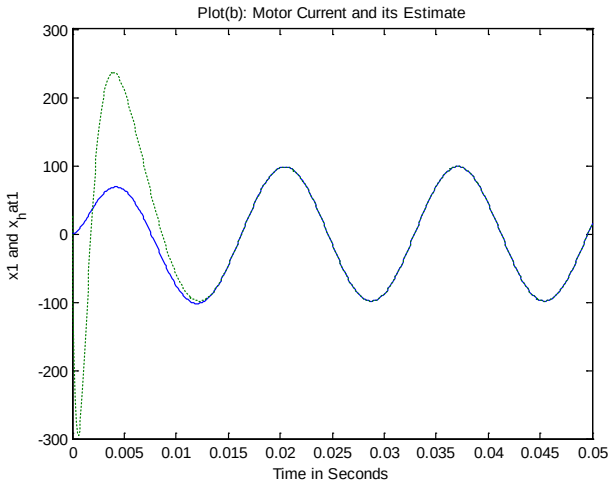
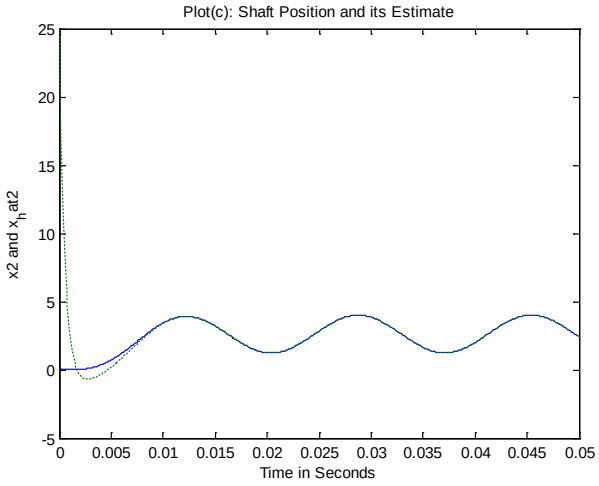


Figure 4: (a): Input pulse Voltage, (b)-(d): Plots of armature current, shaft position , and motor angular speed and their respective estimates, as produced by the observer system..







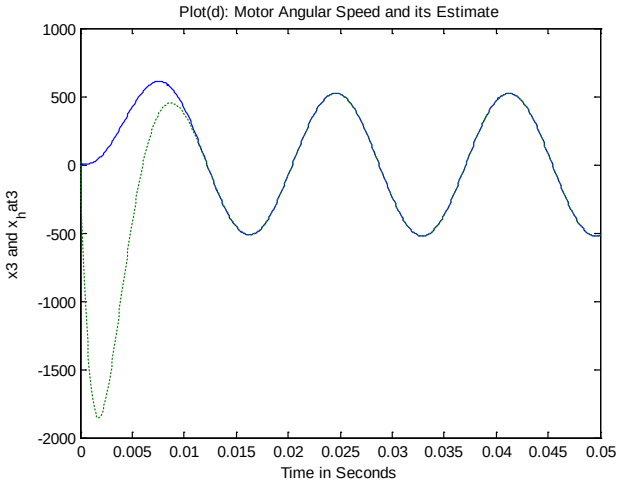


Figure 5: (a) Sinusoidal input, (b)-(d): Plots of armature current, shaft position, and motor angular speed and their respective estimates, as produced by the observer system.

Meaningful Touch and Gestural Interactions with Simulations Interfacing via the Dart Programming Language

T. H. McMullen and K. A. Hawick
Department of Computer Science, University of Hull, Cottingham Road, Hull HU6 7RX, UK.
Email: t.h.mcmullen@hull.ac.nz; k.a.hawick@hull.ac.uk,
Tel: +44 01482 465181 Fax: +44 01482 466666

June 2014

ABSTRACT

Interactive technologies are improving the way in which we are able to communicate with devices. The rise in availability of products, such as the Leap Motion, Kinect and touch screen devices, means that we are able to program applications that significantly increase the potential input from human users. These devices are able to be programmed to work together allowing for situations where one device may improve the functionality of another. In this paper we discus use of the Leap Motion and touch screens in our research into interactive simulations. We use the Dart programming language as a suitable vehicle to integrate together interface components to these two different technologies. We show how combinations of these devices can lead to new and meaningful ways to communicate with applications over and beyond the capabilities of conventional input methods for simulations.

KEY WORDS
Dart; HCI; touch-screens; LEAP; gestural interaction.

1 Introduction

Interactive simulation continues to be an important tool for exploring highly complex models and systems that are not easily explored with analytical methods. Exploring the parameter space and emergent behaviours of complex system models can benefit enormously from a rich set of user interaction capabilities.

Human-Computer Interaction (HCI) [8, 15, 32] is a relatively mature research field [9] but is one that continually benefits from the emergence of new interaction technologies [1, 4, 7, 34]. Devices that support gestures and human touch interaction are not new in principle [2] but devices such as the Leap Motion controller and desktop and larger touch sensitive screens have become commodity priced and widely available recently. Such widespread availability naturally opens up possibilities for various programs and applications that might not otherwise justify custom use of expensive domain specific tools.

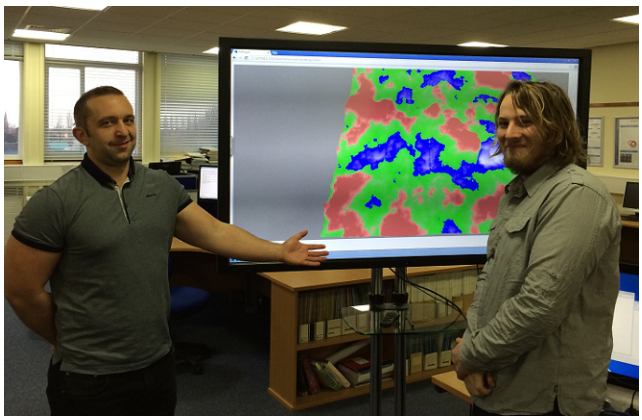


Figure 1: Touch interactions with a large touch screen

Figure 1 shows ways in which a group of collaborators might interact using a large sized touch screen driven by a desktop computer.

In this article we explore the hybrid use of gestural recognition devices such as the Leap Motion controller alongside desktop (27 inch) and larger (65 inch) touch sensitive displays. We investigate how these can be used to support interaction with simulation software for studying complex systems and models.

We describe our implementation experiments in interacting with procedurally generated 3D landscapes [21]. This is an important area for interactive simulation and makes use of heavy duty computational graphics performance [18] on web platforms [22]. The touch and gestural devices allow a user to explore and navigate through such generated scenes, as well as having the potential to allow parameter adjustment and regeneration in interactive time.

Google's Dart programming language [14, 31, 37] is a convenient platform independent [19] language and environment for this work. Dart supports device libraries [6] for interaction with both of these interaction paradigms and these devices. Work has been reported elsewhere on Dart's

use as a simulation programming language [20] and on its library [10], interfacing [16] and performance capabilities [23, 33].

Gestural computing [5] is not a new field, but the availability of commodity devices and the software interfaces that allow their rapid uptake in applications, such as navigating bulk image sequence data [26], opens up many new aspects of HCI and also reasons to consider some of the long standing rules and conventions of HCI [24]. In particular, touch control that has recently been largely dominated by mobile devices such as tables and mobile phones [27], and gesture control [29] that is of a high enough quality and accuracy [38] to be a relied upon interaction mode [36] may give a new lease of life into the desktop computing market, that has been widely discussed as a a potentially declining one [11].

Touch computing in the context of tablet-sized displays has already been explored as a platform and vehicle for interactive simulation models [28]. Desktop devices with a similar interaction capability means that high performance processing power can be brought to bear on the field of interactive simulations.

The Leap Motion gestural controller [12] is a commodity-priced device that plugs into a conventional desktop via a USB interface. It detects hand and finger gestures in the air space immediately above it and can recognize non-touching gestures at hand and finger resolution levels. The Leap Motion has successfully been employed for a number of applications including: data navigation [3]; medical rehabilitation [13] art and sculpting [17]; speech-impaired communication via sign language [25]; and augmented reality systems [30].

The new influx of devices designed to aid with improving human computer interaction (HCI) has changed the way in which we are able to interact with an application or simulation. Through the use of various new and emerging technologies multi-touch devices are becoming commonplace, ranging in size and application. Recently the Leap Motion came to market, allowing for hand tracking in 3D space. This leads to a new area for exploration in HCI, along with the idea of combining these technologies. To create this we look at using Dart, a new programming language designed for web applications. Dart allows us to interface with multiple devices, across different platforms.

Figure 2 shows the set up we used for testing the touch input using a more conventionally sized (27 inch) Dell touch monitor along with the Leap Motion gestural device.

The rise in touch screen enabled devices has helped change how we interact with various pieces of technologies. Small devices such as smart phones and tablets have been greatly improved by the use of touch screens. Larger screens, such as TVs or monitors, are now able to be interacted with in a similar manner as to that of our smart phones. Though these larger devices are capable of such feats, very few applications are designed with that in mind. We look at the use of large multi user touch screens and how to optimize some simulations to take advantage of these new features.

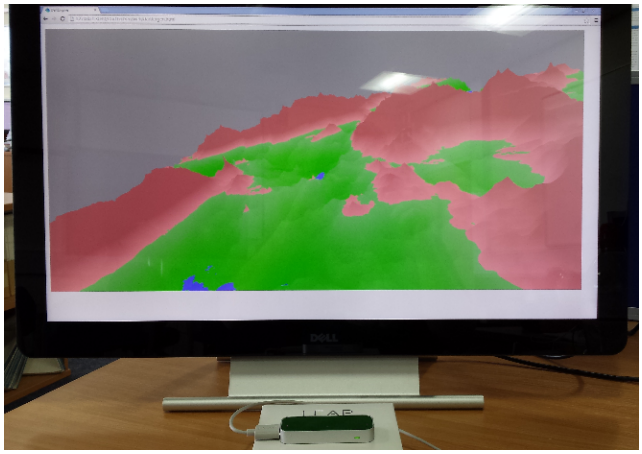


Figure 2: Dell touch screen with a Leap Motion device lying in front of it.

Newer devices such as the Leap Motion, have allowed for more unique interactive communication with applications. The Leap Motion is a USB peripheral which is used to find out information about a user's hand in 3D space. This data includes finger information, location, and rotations, amongst others. It is with this that we are able to experiment with new methods of interacting with simulations. These devices and approaches will continue to become more popular since they are becoming commonly embedded into other hardware, such as laptops and keyboards, and so are expected to become somewhat of an expected standard.

This research looks at the ability to use both touch screens along with the Leap Motion to create new methods to interact with simulations. To do this we must first look at each one's strengths and weaknesses. The touch screen offers users an easy way to interact with devices. It is accurate in the data passed, and simple to design for in many applications. The Leap Motion is also easy to use, but is not as well established as touch screens. Using the Leap Motion we are able to track more than just a point on a screen, but instead a hand in 3D space. The downside of the Leap Motion is that it requires giving the user constant feedback as to the position of their hand and how that is relevant to the current screen.

In Section 2 we describe our approach to using Dart software to integrate simulation code, touch screen interaction code and Leap device interaction code with some implementation details given in Section 3. We present some selected results in Section 4. We discuss our findings and offer some conclusions and suggested areas for further work in Section 5.

2 Method

Designing for multiple inputs to work simultaneously can prove to be challenging, because of this an application needed to be created. To create this application for testing, we looked at a range of languages which support both the Leap Motion and touch screens, and decided to work using Dart. For the touch screen input this is built into Dart, while the Leap Motion is not so common needing its own specialized API for Dart. Using both these inputs can be designed to work using just one hand, leaving the other free.

For an application to be created on these devices a suitable programming language needs be selected. In this situation we chose to use Dart, as it is easy to program and can be run on a range of devices. Code written in Dart also has the advantage of being able to be converted to JavaScript, to increase the range of platforms which it can be run on. Dart's performance is also something which needed to be looked at as we require a responsive application, which does not leave the user waiting. The performance aspect also played a role in the design of the application, as to maximize the user's experience while minimizing the effect on the simulation. The resulting application had two separate loops, one for logic and another for rendering. The logic would occur at set intervals while the rendering would be as fast as possible. The input was taken into account during the logic step so that if rendering lagged behind the inputs would still be recorded.

With Dart, using the touch screen becomes a simple method of using an event listener like that for a keyboard or mouse. Using multi touch becomes a bit more difficult, as we need to know more than a simple input such as a key press. To manage the multi touch input Dart creates a listener which stores an array for each touch that is encountered. The problems which can occur with multi touch is that on larger screens it becomes necessary to differentiate between multiple users. To account for multiple users we take the coordinates and based on how close they are together treat them differently. Another aspect which we looked at was that of what these interactions meant, and how each one would interact with the simulation. For testing the touch screen interactions we implemented several known gestures which are discussed later on in this paper. These include: scaling; rotation; and movement within the generated simulation.

Creating an input for use with the Leap Motion proved to be a little more difficult than that of a simple touch screen. The Leap Motion requires the use of external drivers, which there is a library for. The drivers can be easily installed and implemented to work in a similar manner to that of other input devices. The Leap Motion can then be used by creating a new controller which implements a listener. The listener in turn requires being set up as to correctly interpret user input. With the listener we are able to take data for various input such as finger counts, angles and rotation of the hand relative to the Leap Motion, along with some gestures built in. To explore this we ran some simple tests in which the output from the Leap Motion was used to directly manipulate an object in three dimensional space within a simulation. These test are covered in more detail further on in this paper.

When designing an application to have meaningful inputs from both a touch screen and Leap Motion, we need to look at what use each part fulfils. In our case we have created a 3D simulation in which the user is able to interact with by using the touch screen in conjunction with the Leap Motion. Each part of the inputs were compared to find what they would be most suited for. In our situation we use the Leap Motion to control the camera and how we view the scene. The touch screen was used to aid in creating more accurate control over how we move through the scene as well as manipulating an object within. By using these together we are able to create a user experience when one is able to use one hand to move through and control 3D space. To allow for these to work together we created a camera class which is able to directly take data from the Leap Motion and, in turn, use that to change camera angles. The touch input was then passed and used based on the new view angle of the scene.

3 Implementation

Implementing a software system that integrated both the touch screen paradigm and the Leap Motion gestural paradigm was accomplished using the Dart programming language. To implement both of these we needed to look at how they interact with each other, and how to maximize their usefulness. Each device had different uses so to take advantage of the full potential we needed to design our application to use what was best for a set task. Handling multiple listeners alongside creating of custom classes to hand the input proved to be a slight challenge due to technical issues which will be discussed. To fully take advantage of such a interface a 3D environment was created. This environment needed to be designed to test the implementation of our system, along with not hindering the user's experience.

For the more refined user input the touch screen is used. This is due to the refined accuracy and ease of understanding for the user. Using the touch screen we have the camera position move based on view angle. When the screen is swiped an event is called causing the placement of the camera to move in that direction, such as swiping left to move left. For multi touch events a loop is used to go through each touch and based on that perform an action. The action performed was also reliant on the screen size, as with larger screens it was not unusual for a user to use multiple fingers when one would work. This worked by first taking into account the screen size, then based on that finding the distance between the two inputs. Then if the inputs were of a set distance apart they would be handled as two touches or

```
Hand sphere radius:49.8036 mm, palm position:[Vector3 x:65.2385 y:282.062 z:325.446]
Hand pitch:-6.739971866753232 degrees, roll:-135.1264313948533 degrees, yaw:26.44358775300438 degrees

Hand sphere radius:51.4572 mm, palm position:[Vector3 x:65.531 y:280.128 z:330.244]
Hand pitch:-0.9820868367789749 degrees, roll:-122.2840871409539 degrees, yaw:24.945750361295993 degrees
```

Figure 3: Basic output from the Leap Motion.

if not they would act as a single touch event with the central value between acting as the input. For multi-touch purposes we implemented a zoom feature, which is a common gesture amongst many Apps.

The Leap Motion implementation was not as simple as using touch events. For this we required specialized drivers which have not been updated alongside Dart, causing them to slowly become outdated. This required a little bit of updating as some core functions were no longer working, causing crashes if set gestures occurred. Once the Leap Motion was functioning correctly we were able to create a class to translate between the sensor input and the simulation. This input could range from various gestures to hand location, rotation, or how many fingers were being held up. For our application we only wanted to use the rotation of the hands to apply this to our camera. Once a rotation value was read we would pass that to our custom class which would then be converted into a more relevant value for the view matrix. As the data is in control of manipulating the view matrix, it creates an environment where the user's hand is able to determine how and when we move around the scene.

4 Results

The resulting application partnered alongside a touch screen and Leap Motion, allowing for new and unique interactions to take place. We tested our setup on a range of screen sizes, with few limitations found. Support for smart phones unfortunately was limited in terms of input.

Algorithm 1

```
declare cameraMatrix, hand
Launch Leap Motion Listeners
onFrameUpdate
if hands.length != 0;
   declare Vector3 handDirection = hand.direction;
   updateCamera(handDirection)
end if
updateCamera(Vector3 handDirection)
    declare rotationVector = handDirection;
    check data, edit values to keep in a set range if needed;
    cameraMatrix.rotate(rotationVector)
```

Algorithm 1 shows the process of updating the cameraMatrix based on the inputs from the Leap Motion.

Figure 3 shows the output to the console from the Leap Motion.

Figure 4 shows a user moving around a scene using the

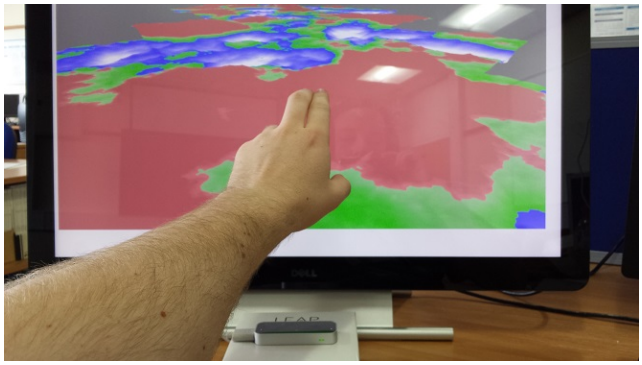


Figure 4: User interaction

touch screen and Leap Motion

5 Discussion and Conclusions

We managed to develop a simulation and terrain navigation application which works on a range of screen sizes, from 22 inch to 65 inch, with the limitations only being those arising from the physical geometry and accuracy of the Leap Motion device. The Leap is designed - and works well - in front of a conventional desktop-sized display.

At present our software and application could not be used on smart phones and tablets, as there is as yet no support for Leap Motion on such platforms. This may change in the future and it may be that the Leap technology will be an integrated part of future tablet devices. Our Dart code base makes it likely that we would be able to port our system to any future platform, particularly using a Web platform interface to Dart.

There are some issues connected with Dart and its conversion to JavaScript code to be run on the Web client. Our application cannot be run as converted JavaScript, like some Dart code can - this is due to the need to use the Dart interfaces to the devices. Nevertheless, this work has demonstrated that a modern language like Dart can provide a very convenient integration language to incorporate input from two very different device paradigms.

HCI as a field is continually developing and there is still considerable scope for new paradigms and associated standards and terminologies to describe the sort of gestures and touch patterns that an application developer would wish to have supported for simulations or the closely related field of gaming applications [35].

In summary, we believe integration of new and emerging HCI technologies such as direct touch, with its positional accuracy, and free moving gestural interactions, with their natural fluidity of motion to be a powerful and useful addition to a simulation model such as the procedural terrain generation and navigation tool we have developed.

We anticipate that these device-level technologies will become even more widespread and that they will find uses in many other applications. Finally, we conclude that Dart is a

very promising integration language for all these operations as well as providing a good platform for implementing the simulation application itself.

Acknowledgments

One of us (THM) thanks Massey University for financial assistance in travelling to the UK, where this work was carried out.

References

[1] Ali, S.I., Jain, S., Lal, B., Sharma, N.: A framework for modeling and designing of intelligent and adaptive interfaces for human computer interaction. Int. J. Applied Information Systems 1(2), 20–25 (January 2012)

[2] Antle, A.N., Marshall, P., van den Hoven, E. (eds.): Workshop Proceedings Embodied Interaction: Theory and Practice in HCI. 29th Int. Conf on Human Factors in Computing Systems, ACM, Vancouver, BC, Canada (7-12 May 2011), iSBN 978-1-4503-0268-5/11/05

[3] Bacim, F., Nabiyouni, M., Bowman, D.A.: Slice-n-swipe: A free-hand gesture user interface for 3d point cloud annotation. In: Proc. IEEE Symp. on 3D User Interfaces. Minneapolis, Minnesota, USA (29-30 March 2014)

[4] Capra, R., Golovchinsky, G., Kules, B., Russell, D., Smith, C.L., Tunkelang, D., White, R.W.: Hcir 2011: The fifth international workshop on human-computer interaction and information retrieval. ACM SIGIR Forum 45(2), 102–107 (December 2011)

[5] Deshayes, R., Mens, T.: Statechart modelling of interactive gesture-based applications. In: Proc. First International Workshop on Combining Design and Engineering of Interactive Systems through Models and Tools (ComDeis- Moto),. Lisbon, Portugal (September 2011), iNTERACT 2011, 13th IFIP TC13 Conf. on HCI

[6] Dhiman, K., Quach, B.: Google?s go and dart: Parallelism and structured web development for better analytics and applications. In: Proc. 2012 Conference of the Center for Advanced Studies on Collaborative Research. pp. 253–254. ACM (2012)

[7] Diaper, D., Stanton, N. (eds.): The Handbook of Task Analysis for Human-Computer Interaction. IEA (2004)

[8] Dix, A., Finlay, J., Abowd, G., Beale, R.: Human-Computer Interaction. Prentice Hall (1993)

[9] Gouin, D., Lavigue, V.: Trends in human-computer interaction to support future intelligence analysis capabilities. In: Proc. 16th Int. Command and Control Research and Technology Symposium. No. Paper 130, Quebec City, Canada (21-23 June 2010)

[10] Griffith, R.: The dart programming language for non-programmers - overview. Web Page (12 December 2012)

[11] Hesseldahl, A.: It's official: The era of the personal computer is over. Web (15 September 2012), http://allthingsd.com/250907

[12] Inc, L.M.: Leap motion controller. Web (2013), https://www.leapmotion.com/

[13] Khademi, M., Dodakian, L., Hondori, H.M., Lopes, C.V., McKenzie, A., Cramer, S.C.: Free-hand interaction with leap motion controller for stroke rehabilitation. In: Proc. CHI 2014, One of a CHInd. pp. 1663–1668. Toronto, Ontario, Canada (2014)

[14] Kopec, D.: Dart for Absolute Beginners. Apress Springer (2014), iSBN 978-1-4302-6481-1

[15] Kortum, P.: HCI Beyond the GUI - Design for Hapric, Speech, Olfactory and other Nontraditional Interfaces. Morgan Kaufmann (2008)

[16] Ladd, S., Wren, J.: Bullseye: Your first dart app. Web (27-29 June 2012), google Developer Conference

[17] Lee, J.: Motionsculpt: A gestural-based 3d content authoring tool. Tech. rep., School of Engineering, Univ. Pennsylvania (2013), http://motionsculpt.tumblr.com/

[18] McMullen, T.H., Hawick, K.A.: Webgl for platform independent graphics. Tech. Rep. CSTN-185, Computer Science, Massey University, Auckland, New Zealand (October 2012), in 8th IIMS Postgraduate Conference

[19] McMullen, T.H., Hawick, K.A.: Improving platform independent graphical performance by compressing information transfer using json. In: Proc. 12th Int. Conf. on Semantic Web and Web Services (SWW'13). p. SWW4052. No. CSTN-174, WorldComp, Las Vegas, USA (22-25 July 2013)

[20] McMullen, T.H., Hawick, K.A.: Performance and capabilities of dart programming language for web based simulation and visualisation. Tech. Rep. CSTN-253, Computer Science, Massey University, Auckland, New Zealand (2013), iNMS Postgraduate Conference, October 2013

[21] McMullen, T.H., Hawick, K.A.: Procedural generation of terrain within highly customizable javascript graphics utilities for webgl. In: Proc. 10th Int. Conf. on Modeling, Simulation and Visualization Methods (MSV'13). p. MSV7287. No. CSTN-219, WorldComp, Las Vegas, USA (22-25 July 2013), http://www.massey.ac.nz/~kahawick/cstn/219/cstn-219.html

[22] McMullen, T.H., Hawick, K.A., Preez, V.D., Pearce, B.: Graphics on web platforms for complex systems modelling and simulation. In: Proc. International Conference on Computer Graphics and Virtual Reality (CGVR'12). pp. 83–89. WorldComp, Las Vegas, USA (16-19 July 2012), cSTN-157

[23] Ortiz, S.: Computing trends lead to new programming languages. IEEE Computer July, 17–20 (2012)
[24] Pavlus, J.: Does gestural computing break fitts? law. Web (8 February 2013), http://www.technologyreview.com/view/511101/does-gestural-computing-break-fitts-law/, mIT Technology Review
[25] Potter, L.E., Araullo, J.A., Carter, L.A.: The leap motion controller: A view on sign language. In: Proc. 25th Australian Computer-Human Interaction Conference. pp. 175–178. Adelaide, Australia (25-29 November 2013)
[26] Preez, V.D., Clarkson, E.P., Innes, S.A., Quach, D.Q., Hawick, K.A.: Comparing hand-gesture and finger-touch interfacing to navigate bulk image-sequence data. In: Proc. 17th Int. Conf. on Image processing, Computer Vision and Pattern Recognition. p. IPC7298. No. CSTN-232, WorldComp, Las Vegas, USA (22-25 July 2013), http://www.massey.ac.nz/~kahawick/cstn/232/cstn-232.html, submitted to IPCV'13, Las vegas
[27] Preez, V.D., Pearce, B., Hawick, K.A., McMullen, T.H.: Human-computer interaction on touch screen tablets for highly interactive computational simulations. In: Proc. International Conference on Human-Computer Interaction. pp. 258–265. IASTED, Baltimore, USA. (14-16 May 2012)
[28] Preez, V.D., Pearce, B., Hawick, K.A., McMullen, T.H.: Software engineering a family of complex systems simulation model apps on android tablets. In: Proc. Int. Conf. on Software Engineering Research and Practice (SERP'12). pp. 215–221. SERP12-authors.pdf, CSREA, Las Vegas, USA (16-19 July 2012)
[29] Rautaray, S.S., Agrawal, A.: Real time multiple hand gesture recognition system for human computer interaction. Int. J. Intelligent Systems and Applications 5, 56–64 (2012)
[30] Regenbrecht, H., Collins, J., Hoermann, S.: A leap-supported, hybrid ar interface approach. In: Proc. Adelaide, Australia (November 2013)
[31] Ridjanovic, D., Balbaert, I.: Learning Dart. PACKT (2014), iSBN 9781849697422
[32] Scogings, C.J.: The Integration of Task and Dialogue Modelling in the Early Stages of User Interface Design. Ph.D. thesis, Massey University (2003)
[33] Shankland, S.: Dart, google's controversial web language, turns 1.0. Web (14 November 2013), http://www.cnet.com/uk/news/dart-googles-controversial-web-language-turns-1-0/
[34] Tesoriero, R., Montero, F., Lozano, M.D., Gallud, J.A.: Hci design patterns for pda running space structured applications. In: Proc. 12th Int. Conf. on Human-Computer Interaction: Interaction Design and Usability (2007)
[35] Turner, J., Browning, D.: Workshop on hci and game interfaces: A long romance. In: Proc. OZCHI 2010 : Design, Interaction, Participation. Queensland University of Technology, Brisbane, Queensland, Australia (22-26 November 2010 2010)
[36] Veeriah, V., Swaminathan, P.L.: Robust hand gesture recognition algorithm for simple mouse control. Int. J. Computer and Communication Engineering 2(2), 219–221 (2013)
[37] Walrath, K., Ladd, S.: Dart - Up and Running. O Reilly (2012), iSBN 978-1-4493-3089-7
[38] Weichert, F., Bachmann, D., Rudak, B., Fisseler, D.: Analysis of the accuracy and robustness of the leap motion controller. Sensors 13, 6380–6393 (2013)

Exergaming Simulator for Gym Training, Fitness Testing and Rehabilitation

Nurkkala, V-M., Kalermo, J. & Järvilehto T.
Kajaani University of Applied Sciences
Kajaani, Finland

***Abstract*—*Exergaming products are gradually becoming more common in gyms. Evidence of the benefits the exergaming products can provide for people of different ages has been widely published. During the last 1.5 years we have developed a new exergaming simulator for gym training, fitness testing and rehabilitation. As a next generation exergaming device, it combines various gym and rehabilitation equipment (treadmill, exercise bike, etc.) with virtual environments, games, sports applications, immersive gaming view and advanced motion controllers.*

Keywords: Exergaming, simulator, gym training, virtual environments, games, CAVE environment

I. INTRODUCTION

Exergaming solutions have been mostly targeted to home markets where dance pads and Wii boards merge exercising and gaming together. Recently exergaming solutions for gyms, rehabilitation centers, and amusement and activity parks have started to become more common. There is only a limited amount of high quality exergaming products on the market at the moment, but it is easy to anticipate that this market will expand in increasing pace in the fight against obesity and inactivity. Another growing field is exergaming training spaces in which all activities and devices are exergames. One of the pioneers in this field is the company called Exergame Fitness USA which provides a large variety of different products and training space solutions.

Exergaming has also become a part of the traditional exercising equipment. Modern gym devices, such as treadmills and exercise bikes, are often equipped with a screen and an integrated computer. The user can choose to play games or use virtual and/or video landscapes during the training. These devices with additional integrated applications, such as social media and web browser, may motivate and bring fun for some users. Nevertheless, the lack of immersion due to the small screen and the limitations in the control of the game (most often it is only the speed that can be controlled) may restrict their usability and leave space for innovations, in which these features are developed to a new level. There are also some products in which an exercise device is connected to a computer and a screen (e.g., TacX, Spivi Indoor Cycling Training system) or the exercise device can be connected to a mobile application (e.g., Kickr, BitGym).

Extensive research has been carried out concerning the possible physical and psychological benefits of exergames for different ages (e.g., children, seniors) [1, 2, 3] and for different target groups (e.g. inactive children, rehabilitation groups) [4]. Also the use of virtual environments in exercising has been studied [5]. These results indicate generally that exergames have positive psychological and physical impact to the studied groups. The studies have shown, e.g., increase of exercise motivation [6], physical activity [4] and energy expenditure while playing exergames [3, 7], and improvement of the balance [2], mood and attention after playing exergames [8]. However, some studies exist with no clear evidence of the benefits [9], but no harmful effects concerning the use of the exergames have been reported.

Several research groups and laboratories have concentrated on the scientific study, development, and/or testing of exergaming products in order to get better understanding and evidence of the possible benefits of the use of exergame devices. Such groups are, e.g., Exergame lab in Coastal Carolina University, University of South Florida Research Labs, the Canadian Exergaming Research Centre, and ExerGame Lab in in SUNY Cortland's Kinesiology Department.

During the last 1.5 years, we have been working in Kajaani University of Applied Sciences with a new kind of exergaming simulator for gym training, fitness testing and rehabilitation. Our aim has been to develop a simulator which integrates different exercising and rehabilitation devices, immersive virtual environments, games and advanced motion controllers, in order to bring these devices to a new level.

Our core team worked previously in the University of Oulu, where our focus was to develop different kinds of driving simulator solutions [10], and carry out psychological research in the field of simulator sickness [11] and driving skills [12, 13]. During the research and development of driving simulators we learned, for example, how to create high quality virtual environments, how to build and use a multi-screen CAVE environment and how to control and fine tune different motion platforms. In 2012 we continued our work at Kajaani University of Applied Sciences, where we built a high-end driving simulator with a cockpit of a real car, a real wheel, gauges, a gearshift and pedals as well as with a 6DOF motion platform. During the development of the high-end driving simulator we started to plan how to apply our knowledge and expertise of simulators to the area of sports,

games, well-being, health and tourism, which are strengths of the Kajaani University of Applied Sciences, and in which game education is internationally well-known.

In the beginning of the year 2013 the development of the new exergaming solution started. The project received funding from Cemis Development Programme (EU funding), and expanded six months later with a parallel game intensive project, supported by the Finnish Funding Agency for Technology and Innovation and 17 project partners. Game companies, such as Supercell and Remedy, joined the project together with partners from the areas of tourism, sports, healthcare, education and technology, and provided resources and expertise for the project.

II. Development of the Exergaming Simulator

In the beginning we did not know exactly which group will be mainly targeted in our project. We designed a three step roadmap in order to find out the potential customers for the product, and to recognize the product's prospects in the exergaming market.

The first step was to carefully select the project partners from many different areas that could help us 1) in finding the existing solutions, potential needs and markets for exergaming simulators, and 2) in providing information about the business potential in the field they represented. The collaboration has been fruitful especially with sports, gaming, technology and health care sector.

The second step consisted of mapping carefully existing products, creating networks, and testing different exergaming solutions, which helped us to get acquainted with the state of the art of the exergaming field. Based on all this information and our analyses we decided to focus primarily in developing the exergaming simulator for gyms and secondarily for fitness testing and rehabilitation.

The third step was to build several pilot setups in which the simulators were tested for different purposes with different kinds of test groups.

Our first setup presenting Athene Exergaming was an orienteering simulator, in which a treadmill, Kinect motion controller, high quality virtual environment and an in-house programmed orienteering game were integrated together. This pilot case was chosen because the World Orienteering Championships were held in our region, which provided a unique environment to pilot and seek for validation for our solution. During the Championships, the sportsmen and other visitors were able to try the simulator and the orienteering game (Figure 1). The task in the game was to find as many checkpoints as possible in three minutes by using real map of the city of Kajaani. Using the real map was possible because virtual Kajaani is an exact copy of real city environment. We attached a HALL-sensor in the treadmill and used it to calculate the speed of the treadmill and transferred the speed information into the game in order to get the gameplay follow the speed of the user. We used the Kinect motion sensor device to recognize the user's gestures to define the direction where the player wanted to go. The user was able to choose between using head tilt or hand movement to turn.

Development of Athene Exergaming has been funded by European Social Fund, European Regional Development Fund, Finnish Funding Agency for Innovation (TEKES), Joint Authority of Kainuu and numerous companies.

Figure 1 Over 300 people tested Athene Exergaming for Orienteering at WOC2013.

After each test run we asked the user to answer a questionnaire and give us feedback by using the iSurvey iPad application. We got over 300 answers from professional and national level athletes, which was very useful for further development of the simulator. An example of the questionnaire's results is presented in Figure 2.

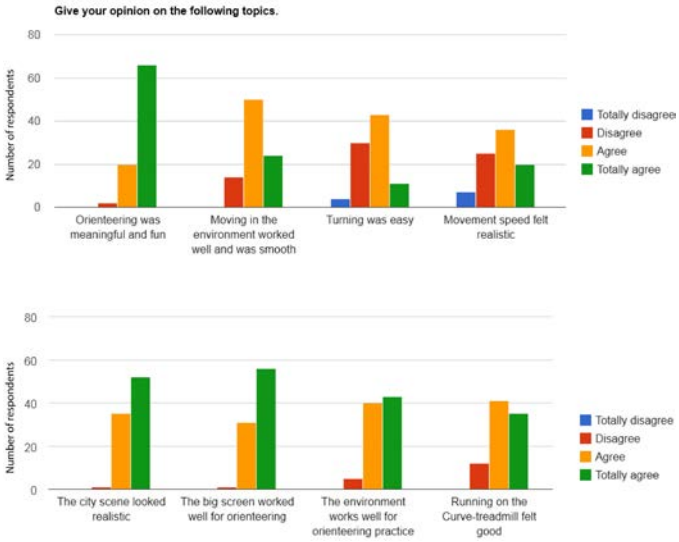


Figure 2 Distribution of user opinions when testing Athene Exergaming during the World Orienteering Championships 2013. Data from 305 users.

The first pilot was followed by a series of different pilot cases (e.g., gym, rehabilitation, fitness testing, and adventure contest pilots) which were defined with specialists of each area. The pilot cases were carefully analyzed afterwards.

Athene Exergaming Team consists of specialists from different areas. The project organization is divided to Lead, Software and Technology Teams which are working closely together. During the project special attention has been paid to improve the communication between different teams. It is difficult to develop efficiently a new product together if vocabularies do not match. The team uses an iterative and incremental agile software development framework for managing product development (SCRUM). Each morning the software team has their daily standup in which everyone reports on the progress made since the last meeting, goals for the next one, and if they have confronted any impediments.

The progess is followed in one week periods by weekly sprint reviews in which the team reports their accomplishments to the whole project organization.

Recently an important step in our product development process and QA management was taken when we started to use a new testing tool called the Deeper Game Testing Software (Figure 3) which was developed in our previous *Pelitys* project. This software makes it possible to monitor the test person, the virtual environment view and the physiological signals simultaneously. It is also possible to add comments and mark important findings in the software. The software shows all data and video records synchronized together.

Figure 3 The interface of the Deeper Game Testing Software.

III. PRODUCT PROTOTYPES

Athene software includes beautiful virtual environments with a free run option (user can explore the area freely) and several routes of different lengths and difficulty levels. Finnish forest (Virtual Vuokatti) and city (Virtual Kajaani) environments, a tropical island, and mountain scenery are the first available scenes.

The software includes variety of exercising modes, such as jogging, mountain biking, orienteering, and adventure. Gaming plays a major role in all exercising modes. As an example, it is possible to challenge a friend in an adventure game or, in future, participate in competitions in which numerous people are participating via Internet. Also in fitness testing, the users may challenge themselves or their friends by competing against avatar characters based on their previous exercise results. The software includes also support for different add-ons, such as heart rate monitor, stride and cadence sensors, and Oculus Rift.

At the present, there are three prototype products in the Athene Exergaming product family. The Athene Basic is a light version of Athene Exergaming. It has been designed for limited space requirements. It includes the Athene software (based on the Unity3D game engine), PC, television or projector, Kinect motion sensor and an Athene communication device (ACD), a sensor to exercise machine or connection to exercise device's interface. It also includes Athene wall, which provides a safe place for devices and wires, and polishes the product looks. In the Athene Advanced (Figure 4), three televisions are used to create an immersive exergaming experience. The setup is provided with an Athene triple wall structure.

Figure 4 Prototype of the Athene Advanced with the Kickr device.

In the Athene Premium (Figure 5), the CAVE environment provides an ultimate exergaming experience. Three projectors and the CAVE structure enhance the virtual or video environment to the next level. There is also a more lightweight CAVE solution which can be easily passed from one place to another.

Figure 5 Prototype of Athene Premium with Curve treadmill.

Athene Communication Device (ACD) has a size of a match box. It enables communication between the exercise device and the Athene software, and includes also different sensors (for recording of heart rate, cadence, stride, and acceleration). It has support for the most typical protocols, such as ANT+, Bluetooth 2.0 and 4.0, but it is also possible to integrate the device by USB serial port.

ACD is the key for the integration of different exercising devices into Athene Exergaming even wirelessly. Full Integration can be made directly by using device's interface which is usually done in collaboration with the manufacturer.

In this case information can be read from the device (speed, RPM, etc.) and data can be sent to the device (angle of actuators, resistance etc.). Semi-integration means that the data can be read by the device interface, but the data cannot be sent to the device. The third option is to attach, e.g., a HALL sensor to the device and then calibrate the received data from the sensor in order to get the device speed synchronized with the speed of the virtual environment. Currently the integration has been made with Woodway Curve and Tunturi Pure Run 10 treadmills, HUR Monark FCM exercise bike, Kickr bike trainer and a couple of other devices.

IV. Discussion

Exergames have good prospects to motivate people from all age groups to be more physically active. They are often perceived as fun to play, but the idea behind exergames also promotes a healthier lifestyle. Sedentary people may find joy in moving through the gameplay. In addition to positive effects in physical activity, it is as important to consider the positive effects on the mental side of the people, as has been presented in several publications mentioned above.

Exergaming, especially with features that allow the users also to travel virtually, can bring fascinating exercising experiences for many people. There are several places in the world that are too dangerous to visit. The climate can be too hot or polluted for exercising outdoors. In urban cities the nature is just too far for many people to reach. Exercising in virtual environments may bring people a unique experience that is realistic, but without great deal of trouble.

The exergaming industry is still small and in its early phase, but many interesting products for different purposes can be found and new products are created with increasing pace. One major player in the exergaming business is Exergame Fitness USA which is the global leader in exergaming products. Also, one recently recognized product takes advantage of arcade game devices. NBA Baller Beats is a prototype of an arcade machine in which Kinect motion sensor is used to turn a basketball computer game into a very promising exergame that aims to get lots of American children and adults to move.

Athene Exergaming takes advantage of existing technologies and equipment. What makes a difference to the other similar products on the market is the innovative way of combining existing exercising and rehabilitation devices with the virtual environment, application of games, advanced motion control and immersive visual systems and wall structures. Also, the Unity 3D game engine provides an ideal platform to apply the software for developing, e.g., more game content for different target groups.

In April 2014 the innovation was awarded as the second best innovation in the world wide Leading Trade Fair of Fitness, Wellness and Health (FIBO). Good feedback from different pilots and success in FIBO Innovation competition has encouraged the developmental work of the whole team for new endeavors. The Athene Exergaming will be launched in the markets in the fall 2014 by a spin-off company from the University.

Acknowledgment

Athene Exergaming has been developed at Kajaani University of Applied Sciences in collaboration with University of Jyväskylä and University of Oulu. Numerous companies have provided their expertise and resources for our R&D. We are grateful of all the support we have received from our partners, as well as encouraging feedback we have got from people all over the world.

References

[1] Brox, E., Luque, L. F., Evertsen, G. J., & Hernández, J. E. G. (2011). Exergames for elderly: Social exergames to persuade seniors to increase physical activity. Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2011 5th International Conference on IEEE, 546-549.

[2] Lamoth, C. J., Caljouw, S. R., & Postema, K. (2011). Active video gaming to improve balance in the elderly. Stud Health Technol Inform, 167, 159-164.

[3] Graf D. L., Pratt L. V., Hester C. N. & Short K. R. (2009). Playing Active Video Games Increases Energy Expenditure in Children, Pediatrics, 124(2) 534-540.

[4] Fogel, V. A., Miltenberger, R. G., Graves, R., & Koehler, S. (2010). The effects of exergaming on physical activity among inactive children in a physical education classroom. Journal of applied behavior analysis, 43(4), 591-600.

[5] Smith, B. K. (2005). Physical fitness in virtual worlds. Computer, 38 (10) 101-103.

[6] Sanders, S. & Hansen, L. (2008). Exergaming: New directions for fitness education in physical education [Policy Brief]. Tampa: University of South Florida, College of Education, David C. Anchin Center.

[7] Graves, L., Stratton, G., Ridgers, N. D., & Cable, N. T. (2007). Comparison of energy expenditure in adolescents when playing new generation and sedentary computer games: Cross sectional study. British Medical Journal, 335, 1282–1284.

[8] Russell, W. D., Newton, M. (2008). Short-term psychological effects of interactive video game technology exercise on mood and attention. Educational Technology & Society, 11 (2), 294-308.

[9] Daley, A. J. (2009). Can exergaming contribute to improving physical activity levels and health outcomes in children? Pediatrics, 124(2), 763-771.

[10] Koskela, K., Nurkkala, V. M., Kalermo, J. & Järvilehto, T. (2011). Low-cost Driving Simulator for Driver Behavior Research. In Proceedings of the 2011 International Conference on Computer Graphics and Virtual Reality, CGVR, Vol. 11, 18-21.

[11] Nurkkala, V. M., Koskela, K., Kalermo, J., Nevanperä, S., Järvilehto, T. & Honkanen, R. (2012). A Method to Evaluate Temporal Appearances of Simulator Sickness during Driving Simulation Experiments. Actes INRETS, 41-49.

[12] Kalermo, J., Nurkkala, V. M., Koskela, K. & Järvilehto, T. (2012). Driving Characteristics and Development of Anticipation of Experienced and Inexperienced Drivers When Learning a Route in a Driving Simulator. Actes INRETS, 105-111.

[13] Järvilehto, T., Nurkkala, V. M., Koskela, K. & Kalermo, J. (2013). Anticipation, Neural Function and Mastering Driving. In Dorn, L. & Sullman M. (eds), Driver Behaviour and Training, Volume VI, 3-11. Ashgate Publishing Limited.

Procedural Generation of Landscapes for Interactive Environments using the Dart Programming Language

T.H.McMullen[1] **and K.A. Hawick**[2]
[1]Computer Science, Massey University, North Shore 102-904, Auckland, New Zealand
[2]Computer Science, University of Hull, Robert Blackburn, Hull, HU6 7RX, United Kingdom
email: [1]t.h.mcmullen@massey.ac.nz, [2]k.a.hawick@hull.ac.uk
Tel: +64 9 414 0800 Fax: +64 9 441 8181

June 2014

ABSTRACT

Dart is a relatively new programming language development by Google, designed for use with web-based applications. In this paper we discuss performance aspects of Dart with the aim of using it to create unique and dynamic procedurally generated visual landscapes. Currently many new programming systems are emerging to improve upon existing web-based languages, but most are built upon existing JavaScript. This makes programming easier as there is less for a developer to re-learn, but unfortunately does not dramatically improve the performance or scalable code structure of web applications. While Dart is still under constant development it is able to outperforms JavaScript on several benchmarks, and is still improving. We describe our Dart implementation of a procedural terrain generation application and discuss implications of the use of Dart for similar simulation and game-related applications.

KEY WORDS
computer games; scene generation; procedural generation; Dart; fractals.

1 Introduction

Programming languages are continually developing [5, 7] and many powerful features have been introduced into modern languages [8, 22] that go beyond the scope of conventional performance-based languages [2, 3] such as the imperative and object-oriented systems widely used for simulation applications [15,26]. One area of particular recent interest has been the introduction by Google and other organizations of web-based programming languages and systems Micoch on Ajax tech [16]. A programming language that is optimized to run well within a web client - and which therefore offers a platform-independent approach [12] - is very attractive for running visual simulations that can be used to demonstrate a model or idea such as procedural terrain generation [13].

Web applications [1] are becoming more commonplace in business as they become a viable solution to many computational problems and especially for those applications with a high graphical performance. This generally requires close integration with the supporting capabilities of the web client with some sort of well developed virtual machine architecture [25, 27]. requirement [11, 14]. We aim to explore the new language Dart [9, 20, 28], and how it is able to improve upon existing web languages. Dart brings many new features which have not been natively implemented in other web languages by means of an object oriented design. Visualization also plays a large role in the design of some web applications, with HTML5 technologies [24] and WebGL being supported across most browsers. The performance improvements of Dart [18, 21] when compared to JavaScript [19] need to be looked at as a developer would clearly not want to swap to using a slower language. Dart is by no means Google's first new programming language and the Go language [23] has already attracted a considerable following. Dart however has some particularly powerful features [17] and capabilities [4] that make it well suited to the area of interactive performance graphical simulations that we consider and describe in this present article.

Figure 1 shows the creation of a grid made up of several tiles, which link together to create a landscape using the diamond square algorithm.

With the release of Dart as a language designed for web-based applications [6, 10] comes many new features which otherwise are not available, or are difficult to obtain in other web languages such as JavaScript. In this research we use Dart as a way to gain improved performance for applications, while also making use of new data structures and computational methods. Darts design allows for the use of many common data structures such as classes, lists etc., which would otherwise need to be recreated by a developer

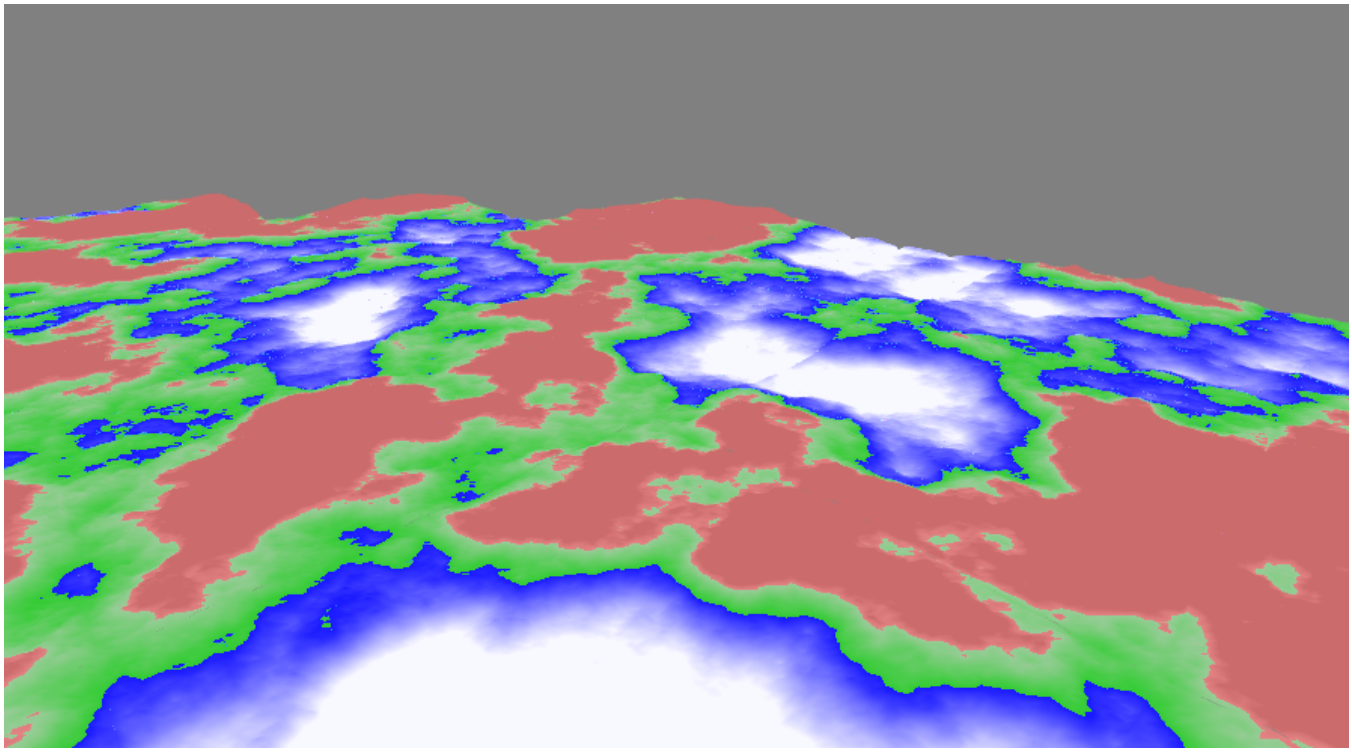


Figure 1: Grid made up of several smaller tiles

in JavaScript through various workarounds. These new features, linked with WebGL, allow for 3D graphics on web-based devices to have increased performance, and allow for newer, more complex simulations and environments to be completed in real time.

Visualization is key to increasing realism in many interactive simulations and virtual reality environments. With many applications using software such as OpenGL to render scenes, newer devices are able to make use of web clients to render an environment or system. To achieve this using web technology, such as Dart, JavaScript and WebGL many problems arise which otherwise would not impact an application that was not web-based. From all these problems which occur, one of the most significant is bandwidth restraints; we want quality, detailed models, but without the large file size that accompanies them. Another commonly faced dilemma is the performance aspect of web-based languages as they are not run natively on the client.

Creating new and unique landscapes can become time consuming, that is why we propose using procedural generation to create them at run time. By using this approach we are able to reduce the bandwidth needed to transfer files between client and server. One of the key aims of this research is to minimize the amount of unnecessary data stored, transferred, and computed when visualizing a simulation. We achieve this through the use of the diamond square algorithm to create landscapes, while applying a grid system which we will talk about later in the paper.

Firstly in Section 2 we will discuss the landscape generation algorithms used to create our visualization, and simulations, such as the diamond square mentioned above. We also summarize the shallow water model system used to generate realistic water behaviour in 3. In the results Section 4 we measure the performance of the simulations, in terms of time to create the scene, then again the render. In Section 5 we review some prior work to improve upon generated landscapes, and describe how Dart can help allow for these to become platform independent. We also suggest some areas for future research in 6.

2 Landscape Generation

For this research we create virtual landscapes to help test the performance of Dart's rendering ability along with computational speed up it provides over JavaScript. This is tested using the diamond square algorithm to generate a virtual landscape. This algorithm allows for use to create a new and unique landscape each time, or by using pseudo random numbers, recall a previously created one. By using the diamond square algorithm mixed with a grid based system, we are able to greatly increase not only the size of the landscapes, but also the quality. When implementing the grid based system this can lead to errors and discontinuation throught the system if not properly taken care of.

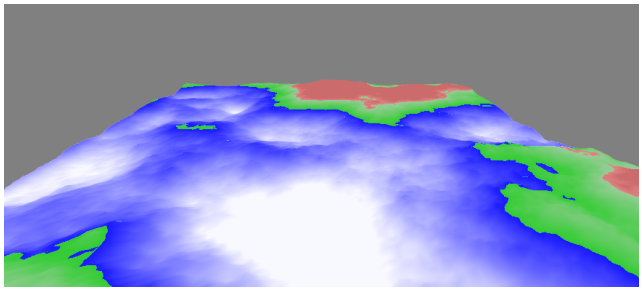


Figure 2: A small tile created using the diamond square algorithm

To set up the environment required for testing we first must make a grid which is able to accommodate the results from the diamond square algorithm. We then must pass this information to a mesh to manipulate the given height, which is used to produce the 3D landscape. From this point the information is passed to the GPU to be rendered. We try to minimize creating and processing unnecessary data to allow for a faster start up and run time. The performance of a simulation is based on how fast it is able to run as JavaScript vs Dart.

Unfortunately like JavaScript, Dart does not have native support for multi dimensional arrays. This leads to the need to work within the confines of a single array, or create a custom array class. For this we worked with both array types, using multidimensional to simplify the simulation, while converting to a single array when passing the data to be rendered. By creating our own array class we were able to use that to help implement our grid system, which is what allows for the continued spreading of the landscape within a scene.

The grid system works by first creating the core part of the landscape which all connecting pieces are based on. From the central piece we are able to add other tiles to the grid based on the existing edges. By using the existing edges as the seeds for the new tile we can ensure that they match up correctly, while also having the style of the existing landscape lead over to the new tile. This grid setup allows for the rapid creation of a range of landscapes in terms of size and detail. Once a tile piece has been created and its data passed to the GPU for storage only the outermost edge pieces are saved. This is designed to help minimize memory usage to improve performance of large landscapes on newer devices, such as smart phones.

Figure 2 shows the creation of a small tile, used to create a landscape and generated by the diamond square algorithm.

Algorithm 1 shows the creation of a tile for the grid based on the size of the grid and a given location. Also shows some simple logic behind the creation of the tile to fit in the grid

Algorithm 1

```
declare grid = List[], size;
for i less than size do
   grid.add(new tile(i, size);
end for
part of tile class
tile(location, gridsize)
    check edges, to seed base values
    create new tile at set location
    save newer edge values
    pass vertices data to GPU
```

3 Shallow Water Simulation

To test how Dart would run on a more intensive simulation with ongoing updates to the GPU we implemented an instance of the shallow water model. This model is designed to simulate a small area of water, which is not deep. The simulation works by creating a grid which is then initialized with some value as to the height of water plus the height of the ground below. As the simulation runs each part needs to be updated to reflect the movement of the water based on the simplified NavierStokes equations. After each update the grid needs to be passed to the GPU to update the current scene. We chose to run this simulation on the CPU rather than a GPU to more accurately display the benefits of each language, Dart and JavaScript.

We create our shallow water by firstly initializing a grid so not every part is level, in our case we raise the centre of the grid to create ripples. The velocity of the X and Y are both set to 0 as to not affect the initial drop of the water in either direction. Once initialized we start the simulation; this works by taking the current height of the first spot on the grid (in our case 0,0). Based on the height we work out how it is affected by the velocity in the X or Y direction of the grid. The new height value is stored in a temporary array as not to affect any future calculations which require the previous value. Once the height has been updated we then look at the velocity in the X direction and update that based on the drop of the height, by checking it is based on the ground value at that point. The Y velocity is also updated in the same manner, after which we have the new values for the velocities and height which need to then be passed to the GPU to update the simulation.

Algorithm 2 shows steps to update the shallow water simulation.

This simulation was chosen as it tests the speed of Dart and JavaScript in terms of performance with floating point numbers. It is also able to demonstrate the rate at which they are able to update the GPU to show the flow of water.

Figure 3 shows the creation of the shallow water simulation, before any updates are applied.

Algorithm 2

```
declare h = List[], vx = list[], vy = list[];
initialize h, vx, vy;
while runing do
    update h based on vx and vy;
    update vx based on change in h;
    update vy based on change in h;
    push new h values to the GPU;
    drawWater();
end while
```

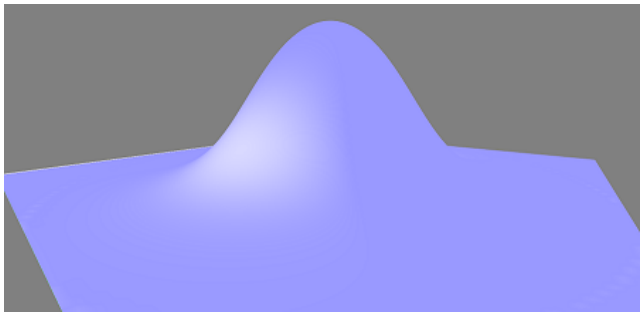


Figure 3: Shallow water at initial state

Figure 4 shows the progress of the flow of water after several time steps.

Figure 5 shows the progress of the flow of water, as the velocity in X and Y have both been lowered and the height of the water has levelled off.

The images above show the progress of the flow of water from the initial stage, where the water has rapid movement. This continues until the velocity of the X and Y have both dropped down significantly and the height of the water levels off. Once this has happened the small ripples are present within the water.

4 Experimental Results

We found that by using Dart we were able to have substantial speed ups in terms of time generating the landscape, as well as the time taken to render a frame. We tested our generation techniques by running three lots of code. Firstly the Dart code source, then the Dart code which Dart's editor is able to

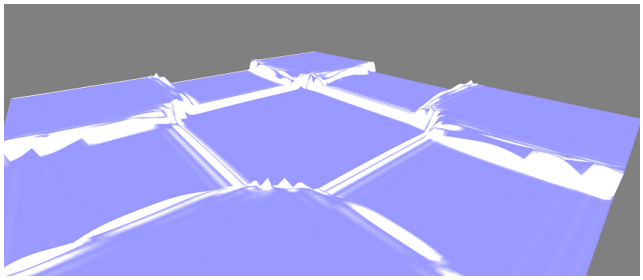


Figure 4: Shallow water after several time steps

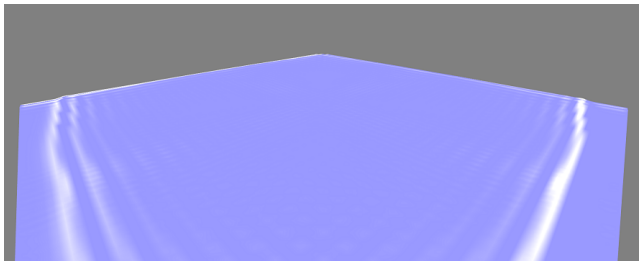


Figure 5: Shallow water once the water has settled down

convert to run as JavaScript. Finally we tested a JavaScript implementation we had created. The tests covered the time taken to create a tile in the grid along with the start up times for each implementation.

From our tests we were able to find that Dart is the fastest in terms of creating a single tile while also being faster to start. We found that in all cases the first tile took the longest to be created, taking an average of 12 ms in Dart to 52 ms in JavaScript. The subsequent tiles took a fraction of the original time. To generate a tile in Dart would take around 3 ms, while in JavaScript it would take about 5 ms. The Dart code which had been converted to JavaScript was also found to be faster with an average of 4.4 ms to create a tile.

By using our grid system we are able to test our methods on a range of devices and a range of environmental sizes, with each one showing the same trend as above. This led to the opinion that Dart had a performance advantage over JavaScript.

To further test Dart's performance in terms of an ongoing simulation we tested using the shallow water simulation as mentioned above. In this we look at the rate at which a new frame was able to be produced after each logic step as well as the time taken to set up each variable and initialize them.

When setting up the shallow water simulation, Dart was much faster taking around the time taken by JavaScript. The updating of a scene based on the logic showed Dart being slightly faster than JavaScript, but both only taking a fraction of a second.

It is our belief that the performance increase from Dart for both the Diamond Square algorithm and shallow water simulation comes from how each handle data types. Dart is able to work with a range of data types such as integer's doubles and floats etc. JavaScript on the other hand is limited as it treats all numbers as doubles, increasing the computational intensity of some calculations. As the shallow water simulation works with primarily doubles it is no wonder that Dart only had a slight performance increase while with the the landscape generation it was found to have a significant advantage.

Figure 6 shows two images, one showing the process of creating a small gird adding new tiles to the list, then the next image showing how to render each element within the grid.

```
setup(){

  containerClass = new List<object>();

  //First create the landscape
  for(int i = 0; i < 4; i++){
    for(int j = 0; j < 4; j++){
      containerClass.add(new land(gl, i, j));

      containerClass[0].updateMesh(containerClass[j+(i*4)]);
    }
  }
```

```
draw(locX, locY){
  //gets our current view matrix
  Matrix4 viewMat = camera.getViewMat();
  gl.clear(webgl.RenderingContext.COLOR_BUFFER_BIT | webgl.RenderingContext.DEPTH_BUFFER_BIT);

  for(int i = 0; i < containerClass.length; i++){

    containerClass[i].draw(viewMat, projectionMat);
  }
}
```

Figure 6: The process of creating the grid, and rendering each element within.

(It is worth noting in this grid not every element is necessarily a tile, and could be a instance of water or another object we add to the environment.)

5 Discussion

In this paper we have looked at the performance of Dart as a web language and compared it to JavaScript. We looked at several aspects of the languages and how they affect the applications produced. Creating multiple simulations lead us to find possible improvements for each simulation based on the language. Several key advantages to Dart were found in terms of performance, and structure.

These tests worked as they helped show the drawbacks of various methods and features in JavaScript, while showing how they have been improved upon in Dart. These improvements include allowing for a class structure which was missing from JavaScript. The use of multiple data types in Dart allowed for a great speedup in performance when integers were been used as JavaScript still needs to work using doubles.

This research shows how new languages such as Dart are becoming more powerful and useful than older ones. One main finding of this research is not that Dart is a better language but why it is. The one large factor holding Dart back is the limitation of devices it is able to be run on. It can overcome this drawback to varying success by converting to JavaScript. By converting the code source this allows Dart code to be run on multiple devices at improved performance to that of JavaScript.

6 Conclusions

Dart is a new language that is essentially just out of beta-release stage. It is being constantly updated and improved with support for a range of new technologies. By adapting to use these newer technologies a developer is able to improve the performance of an application, by the way of a range of devices. Dart's ability to be converted into JavaScript sets it above that of many of the newer languages being released, and allows for it to overcome many of it shortcomings.

In the future we would like to look at taking advantage of many of the newest features in Dart. SIMD is one aspect which offers a great performance increase based on how an application or simulation can be divided.

In summary, we believe Dart has great potential as a platform-independent programming language, capable of supporting both computational and graphical performance for interactive simulations and game applications.

References

[1] Balat, V.: Client-server web applications widgets. In: Proc. WWW 2013 Companion. ACM, Rio de Janeiro, Brazil (13-17 May 2013), iSBN 978-1-4503-2038

[2] Bracha, G., Ungar, D.: Mirrors: Design principles for meta-level faciloities of object-oriented programming languages. In: Proc. OOPSLA'04. ACM, Vancouver, British Columbia, Canada (24-28 October 2004)

[3] Chisnall, D.: A new objective-c runtime: From research to production. Communications of the ACM 55(9), 44–47 (2012)

[4] Dhiman, K., Quach, B.: Google?s go and dart: Parallelism and structured web development for better analytics and applications. In: Proc. 2012 Conference of the Center for Advanced Studies on Collaborative Research. pp. 253–254. ACM (2012)

[5] Emerick, C., Carper, B., Grand, C.: Clojure Programming - Practical Lisp for the Java World. O'Reilly (2012), iSBN 978-1-449-39470-7

[6] Griffith, R.: The dart programming language for non-programmers - overview. Web Page (12 December 2012)

[7] Halloway, S., Bedra, A.: Programming Clojure. Pragmatic Bookshelf (2012), iSBN 978-1-93435-686-9

[8] Horstmann, C.S.: Scala for the Impatient. Pearson (2012), iSBN 978-0-321-77409-5

[9] Kopec, D.: Dart for Absolute Beginners. Apress Springer (2014), iSBN 978-1-4302-6481-1

[10] Ladd, S., Wren, J.: Bullseye: Your first dart app. Web (27-29 June 2012), google Developer Conference

[11] McMullen, T.H., Hawick, K.A.: Webgl for platform independent graphics. Tech. Rep. CSTN-185, Computer Science, Massey University, Auckland, New Zealand (October 2012), in 8th IIMS Postgraduate Conference

[12] McMullen, T.H., Hawick, K.A.: Improving platform independent graphical performance by compressing information transfer using json. In: Proc. 12th Int.

Conf. on Semantic Web and Web Services (SWW'13). p. SWW4052. No. CSTN-174, WorldComp, Las Vegas, USA (22-25 July 2013)

[13] McMullen, T.H., Hawick, K.A.: Procedural generation of terrain within highly customizable javascript graphics utilities for webgl. In: Proc. 10th Int. Conf. on Modeling, Simulation and Visualization Methods (MSV'13). p. MSV7287. No. CSTN-219, WorldComp, Las Vegas, USA (22-25 July 2013), http://www.massey.ac.nz/~kahawick/cstn/219/cstn-219.html

[14] McMullen, T.H., Hawick, K.A., Preez, V.D., Pearce, B.: Graphics on web platforms for complex systems modelling and simulation. In: Proc. International Conference on Computer Graphics and Virtual Reality (CGVR'12). pp. 83–89. WorldComp, Las Vegas, USA (16-19 July 2012), cSTN-157

[15] Miller, J.A., Han, J., Hybinette, M.: Using domain specific language for modeling and simulation: Scalation as a case study. In: Proc. 2010 Winter Simulation Conference. pp. 741–752 (2010)

[16] Mlcoch, Z.: Technology for the application based on ajax. AWERProcedia Info. Tech and Computer Science 03, 335–342 (2013), proc. 3rd World Conf. on IT (WCIT-2012)

[17] Mokkapati, S.: Static Type Checker Tools for DART. Master's thesis, San Jose State University (December 2012), http://scholarworks.sjsu.edu/etd_projects

[18] Ortiz, S.: Computing trends lead to new programming languages. IEEE Computer July, 17–20 (2012)

[19] Powell, T.A., Schneider, F.: JavaScript: the complete reference. McGraw-Hill (2012), iSBN 9780071741200

[20] Ridjanovic, D., Balbaert, I.: Learning Dart. PACKT (2014), iSBN 9781849697422

[21] Shankland, S.: Dart, google's controversial web language, turns 1.0. Web (14 November 2013), http://www.cnet.com/uk/news/dart-googles-controversial-web-language-turns-1-0/

[22] Suereth, J.D.: Scala in Depth. Manning (2012), iSBN 978-1-935182-70-2

[23] Summerfield, M.: Programming in Go. Addison-Wesley (2012)

[24] Taraldsvik, M.: Exploring the future: is html5 the solution for gis applications on the world wide web? Tech. rep., Norwegian University of Science and Technology (NTNU) (12 March 2012)

[25] Thamsen, L., Gulenko, A., Thomas, D.A.: Orca: A single-language web framework for collaborative development. In: Proc. 10th Int. Conf. on Creating, Connecting and Collaborating through Computing (2012)

[26] Todd, A.B., Keller, A.K., Lewis, M.C., martin G. Kelly: Multi-agent system simulation in scala: An evaluation of actors for parallel simulation. In: Proc. Int. Conf of Parallel and Distributed processing and Applications (PDPTA'11) (2011)

[27] Vilk, J., Carey, C.J., Ng, J., Berger, E.D.: The doppio jvm: Building a complete language in the browser. Tech. rep., UMass, USA (2013)

[28] Walrath, K., Ladd, S.: Dart - Up and Running. O Reilly (2012), iSBN 978-1-4493-3089-7

Semiotic Transposition Method for Biomimetic Software Modeling: an Overview

C. E. P. Camargo[1] **and I. S. Vega**[1]
[1]Intelligence Technologies and Digital Design, Pontifícia Universidade Católica, São Paulo, SP, Brazil

Abstract— *Modern computers have Turing machine as their model and, as such, they are only able to perform calculations. In some sense, biological systems also perform calculations, but at a much higher level of complexity. So, how to simulate natural phenomenon complexity on a limited computational environment? This paper seeks to answer this question by establishing the Semiotic Transposition Method (STM), an auxiliary tool applicable to the development of biomimetic softwares. The fundamental hypothesis of STM is that semiotic theory can be used as an intermediate field to the transposition of natural phenomenon from its original biological field to computational field.*

Keywords: biomimetic software, modeling methodology, simulation.

1. Introduction

The word *biomimetics* was coined by Schimitt in the decade of 1950 and was used in the title of his paper of 1969 [1]. He used this terminology to refer to the transfer of ideas and analogues from biology to technology [2]. For Benyus, biomimetics or biomimicry, as she prefers, is a new science that studies natural models in order to imitate them, or their processes, in search of inspiration to solve human problems [3].

The most common applications of biomimetics relate to the development of tangible products inspired by concrete models. However, natural processes and their abstract relations may also serve as inspiration sources in solving complex human problems. The fields of natural computation [4], biorobotic softwares [5] and artificial intelligence [6], among others, are examples of biomimetic approaches.

According to biorobotics, biomimetic approaches can serve two different and mutually exclusive purposes, as tools for biologists to study animal behavior and as support for engineers to study and evaluate biological algorithms for potential engineering applications [5]. Both start from similar bases, but seek different goals: While the first one tries to accurately simulate all the organic aspects of a given natural phenomenon, the latter seeks to simulate the underlying biological algorithm, modeling just the essential abstract aspects necessary for its computational representation. Regardless of different purposes, the overall question to be answered concerns the transposition between the two fields of interest, namely, how to do this? How to transpose a natural phenomenon from its original biological field to the computational field?

This is a fundamental issue considering the difference, in complexity, between these fields. Modern computers have Turing machine as their model and, as such, they are only able to perform calculations [7], [8], [9], [10]. In some sense, biological systems also perform calculations, but at a much higher level of complexity. So, how is it possible, then, to simulate the complexity of natural phenomenon in a limited computational environment? This article seeks to answer this question by establishing the *Semiotic Transposition Method (STM)*, applicable to the abstraction of meta-models to be used in the development of biomimetic softwares which purpose coincides with the second type described above, i.e., the study of biological algorithms for potential applications in engineering.

The fundamental hypothesis of STM is that the semiotic theory can be used as an intermediate field in the transposition process between biological and computational fields. This method seeks to recognize the *semiosis*, or sign processes[1], operating in the biological field, and transpose them as algorithmic functions to the computational field. Each semiotic triad (object/representamen/interpretant) corresponds to a computational triad (input/processing/output), which responsibility is to make an interpretation.

After the description of the steps to be followed in implementing the STM, a real case is presented: the habituation and sensitization capabilities of the *Aplysia californica*, a species of sea slug which behavior was widely documented by Kandel in the development of his research on memory and learning [11], [12]. This biological phenomenon is reduced by STM to a generic model (meta-model) that expresses the essential functionality of such phenomenon. The scope of this paper ends with the presentation of this meta-model. Afterwards, in future developments, this meta-model can be used as clue to develop biomimetic softwares to solve problems concerning to similar functionalities in human domains.

2. Theoretical grounding

Semiotics is the science of signs and their processes in nature and culture. It is the study of signification and a sign is the element responsible for this phenomenon. In the

[1]In this paper, semiosis and sign process are synonymous.

theory of Peirce [13], [14], a sign is represented by a triadic and inseparable relation (object/representamen/interpretant) and its actuation in a system is called semiosis or sign process [15], [16].

According to Nöth [17], in the last years of the twentieth century, semiotics expands its cultural context toward the possibilities of a semiotic of Nature. In this context, Ziemke [18] believes that the theory of Uexküll can be useful not only in understanding the use of signs and representations by living organisms, but also in understanding the possibilities and limitations of autonomy and semiosis in artificial organisms. For Uexküll [19], [20], [21], [22], [23], the sign relations among species, between organisms and their environment, and also between cellular structures, occur through what he called *functional circles* (or habitable-integuments, for cells)

In some sense, semiosis can be taken as an interpretation process: when something (object) is presented to an entity capable of interpretation, this happening is internally represented (representamen) and this fact causes an interpretation (interpretant) in accordance to the characteristics of the entity, them this interpretant can present itself as a new object to another semiosis that will make a new interpretation and so on, establishing a semiotic chain. By analogy, it is assumed here that each semiosis can be considered as a simple Turing machine: the object can be taken as the input, the representamen as processing and the interpretant as the output. Then, the output acts as a new input for another Turing machine and so on, performing some kind of a computational chain, similar to a semiotic chain. Hence, the STM seeks to reduce complex biological phenomenon to its essential semiotic triads (semiosis or sign processes) in order to transpose these triads to computational triads (input/processing/output).

But, when choosing a biological phenomenon of interest, however, the researcher must identify the hierarchical level where the relevant semiotic processes occur. STM uses the hierarchical structuralism of Salthe in this task. For him, an observer (researcher) should choose the focal level according to his/her own interests. A complete picture of hierarchical levels can be described by the following levels: atomic, molecular, intracellular organelles, cells, tissues, organs, systems, body, ecological niche, ecosystem and environment. In accordance with specific needs, some of these levels may be deleted and other condensed, e.g., a possible structure can consider the level of cells and their organelles as the focal level. In this case, molecular level could represent the lower level and the body level could represent the upper level, suppressing the tissue, organs and system levels [24]. According to El-Hani and Queiroz [25], the hierarchical structuralism assists in describing complex systems by drawing up a basic triadic relation based on the semiotic theory of Charles Sanders Peirce; the focal level is responsible for the semiotic chain that corresponds to a certain phenomenon, and settles between two other levels, the lower level or the level of the potentialities, and the upper level or restriction level.

The development of biomimetic softwares is the final objective of the STM. Thus, it should be added an organizing element that can help to achieve more appropriate software architectures. For this task, STM uses a non-semiotic theory, the subsumption architecture developed by Brooks at MIT and used by him to construct robots. This biologically inspired approach divides the implementation into subsumption behavioral layers. The lower layers are subordinated to the upper layers [26], [27], [28].

3. The Semiotic Transposition Method

For Vincent, biomimetics is still an empirical approach. Thus, he proposed in 2006 [2] a method to develop biomimetic products. He called it BioTRIZ, a derivation of a Russian solution system used to solve engineering problems called TRIZ (acronym of *Teorija Reshenija Izobretatel'skih Zadach* (loosely translated as *Theory of Inventing Solve Problem*). A possible step forward in this method towards biomimetic software modeling is presented by Korecki in 2008 [29]. He proposes a new framework for software development based on interdisciplinarity and biomimetics, believing that this approach could bring creativity and innovation to computer science. The STM presented below is not a rival for BioTRIZ or for the Korecki's approach; rather, it can be used not only as a standalone tool, but also work jointly with these other approaches.

For Benyus [30], the practice of biomimetics can occur from biology to engineering or vice versa. In the first case, a biological phenomenon may suggest new ways to overcome engineering challenges. In the second case, starting from a real human problem, researchers try to identify processes in nature that can help in finding a solution for such problem. The STM was developed considering the first of these cases and consists of the six steps described below.

- *Step 1 - Primary concerns:* The goal is to find characteristics in the phenomenon under study that may indicate its potential as inspiration for the development of algorithms to be used in human domains. A very important issue here is to determine the function that the biological algorithm satisfies. By analogy, this function will work as a clue to find human problems with similar characteristics.

- *Step 2 - Defining the subsumption architecture:* The goal is to abstract the subsumption architecture of the phenomenon under study. The subsumption architecture is organized in incremental layers. Each layer connects perceptions to actions. The goal here is to find the perceptions and actions in the phenomenon under study, dividing them into incremental layers that

will be represented in the final software.

- *Step 3 - Choosing the focal level:* The goal is to choose the appropriate hierarchical structure that match the phenomenon under study. It must be identified: the focal level in which semiotic chains take place, the lower level of potentialities and the upper restriction level.

- *Step 4 - Finding relevant sign processes:* The goal is to identify all relevant sign processes that take place in focal level. A biological structure in operation presents a great number of entities that perform sign processes, as much as necessary to attend its evolutionary context. Since the goal of STM is to transpose only the biological strategy underlying to a biological algorithm and not the total number of biological entities involved in the phenomenon, this step also works as a heuristic process eliminating redundancies and irrelevant parts to the strategy abstraction.

- *Step 5 - Semiotic modeling:* The goal is to abstract a semiotic model that shows all relevant sign processes. Here, a diagrammatic model must be designed to represent the relevant sign processes recognized in the focal level, its distribution among subsumption layers, its elements and its relations.

- *Step 6 - Computational meta-modeling:* The goal is to abstract a computational meta-model based on the semiotic model (UML based diagrams [31]). This is the transposition step. It is a meta-model because it is a generic diagram representing the relational dynamics between sign processes, and refers no more to a biological algorithm but to computational one. This diagram intends to represent the essence of the phenomenon under study seeking its implementation in digital environment.

4. Case study

The case below is inspired by the learning capabilities of the *Aplysia californica*. The works of Kandel [11], [12], [32], [33], [34], [35] concerning the habituation and sensitization trainings are the basis for this study. Starting from Pavlov protocols, Kandel opens the behaviorist "black box" in order to recognize the underlying neural pathways to processes of learning and memory retention. According to Kandel, habituation allows aplysia to focus on important events. Immature animals normally overreact to non threating stimuli. Getting used to such stimuli causes the animal to pay attention only to relevant events. This is important for the organization of its perceptions and thus for its survival. On the other hand, sensitization is the mirror image of habituation. Its function is to accentuate animal reactions - even to harmless stimulus - after being exposed to a truly threatening event. It is a kind of learned fear which increases the level of attention of the animal in specific contexts [12].

Basically, to train habituation, Kandel touches the siphon of the aplysia (harmless event), which causes the retraction of its gill. However, the harmless stimulus repetition contributes to the gradual decrease of the retraction intensity. The habituated behavior can last from minutes to weeks depending on the dynamics of the training sections. Sensitization is a bit more complex than habituation. A harmful stimulus applied to certain neural circuit produces a change in the reflex strength in another neural circuit. To train sensitization, Kandel causes a single shock on the tail of the aplysia, which results not only in a strong reaction to this stimulus, but also to other stimuli, even harmless ones. With specific training sections, sensitization can last from minutes to weeks. Thus, when a single event occurs (touch on siphon) three different responses can follow: normal (for untrained animals); short and long term habituation (animal takes the stimulus as noise), and short and long term sensitization (animal takes the stimulus as signal) [11], [12], [32], [33], [34], [35].

4.1 Step 1: Primary concerns

Habituation and sensitization suggest a very effective signal-and-noise differentiation system (differentiation between useful and useless information). Similarly, there are several human systems in which the identification of relevant events (signals) among a multitude of random events with low differentiation (noise) is a hard problem to solve: big data analysis [36], earthquake prediction [37], cybersecurity [38], among others, could benefit from an efficient system that can learn to recognize when a particular input should be considered or not. This signal-and-noise analogy points to interesting possibilities for the development of softwares inspired by the learning ability of aplysia.

4.2 Step 2: Defining the Subsumption Architecture

Taking into account the perceptions and actions involved in the experiments of Kandel, three incremental layers can be recognized: normal behavior layer, related to untrained animals; habituation layer, that prevails to normal behavior layer; and sensitization layer, that prevails over all [11], [12].

4.3 Step 3: Choosing the focal level

Seeking to recognize the neural pathways underlying the behaviors of habituation and sensitization of aplysia, the work of Kandel points to an appropriate focal level, i.e., the cellular level and its organelles. Consequently, the lower or micro-semiotic level, initiator of the process, is the molecular level; and the upper or macro-semiotic level, responsible for natural constraints, is the organism level [11], [12].

4.4 Step 4: Finding relevant sign processes

Heuristically, 10 sign process could be abstracted from the subsumption layers of the aplysia (5 at normal behavior layer, 2 at habituation layer and 3 at sensitization layer), as follows:

4.4.1 Normal behavior layer

- *Sign process #1 (S1) - Trigger mechanism of the siphon:* when the siphon is touched, an ionic rearrangement occurs in the trigger zone of its sensor neuron. If the touch strength is sufficient to reach the threshold of the equilibrium potential at the trigger zone (-55 mV), it settles the action potential that leads to the retraction of the gill. If the threshold is not reached, nothing happens. In semiotic terms (triad: object/representamen/interpretant): reaching the threshold designates the object (O), the re-arrangement on the ions designates the representamen (R), and the action potential designates the interpretant (I) [11], [12], [19], [20], [23], [39].
  Here, it is worth mentioning the heuristic character of this step. After the first action potential, the process continues toward semiosis S2, and an intermediary semiotic chain settles along the axon. This set of semiosis, however, only conveys the characteristics of the first semiosis; they are necessary repetitions only to the animal due to the physical characteristics of the axon (length), being not a new semiosis in qualitative sense. So, as STM seeks to abstract only the essential aspects of biological algorithm, this intermediary chain can be ignored.

- *Sign process #2 (S2) - Exocytosis mechanism:* the action potential, generated in the trigger zone, reaches the active zone of the sensor neuron (siphon). Action potential acts as the object (O) to a new sign process which representamen (R) is the flow of large amounts of ions Ca^{++} into the cell, resulting in exocytosis (I). The consequence of exocytosis is the presence of glutamate in the synaptic cleft, which amount is proportional to the amount of mobilized vesicles in the active zone (influence of semiosis 4, below) and the amount of interneurons connections [11], [12], [19], [20], [23], [39].

- *Sign process #3 (S3) - Motor mechanism:* the glutamate in the synaptic cleft is picked up by the receivers of motor neuron (gill) and acts as the object (O) of a new sign process. This leads to a specific organic configuration (R) in the motor neuron (gill), which causes physical reactions in the innervation region of the gill (I) and, thus, its retraction. The retraction is as strong as the amount of glutamate in the synaptic cleft [11], [12], [19], [20], [23], [39].

- *Sign process #4 (S4)- Endocytosis mechanism:* the exocytosis eliminates the glutamate vesicles of the active zone of the sensor neuron (siphon). Then, an opposite mechanism (endocytosis) takes place in order to restore such vesicles[11], [12][39]. Thus, the absence of vesicles (O) causes a specific organic configuration, setting the active zone (R) which result is the recovery of such vesicles (Endocytosis) (I). Here it is considered a rate of replenishment of vesicles (Rr) that can be normal or modulated (habituation or sensitization), in accordance to its respective subsumption layers [11], [12], [19], [20], [23], [39].

- *Sign process #5 (S5) - Connections mechanism:* this mechanism controls the amount of interneuron connections and contributes to the formation of long-term memory for both habituation and sensitization. Its object (O) is the presence of some specific protein concentration that leads to new organic configurations (R). The increase or decrease in the number of connections is the interpretant (I) of this sign process. Here it is considered a rate of connection construction (Rc) that can be normal or modulated (habituation or sensitization), in accordance to its respective subsumption layers [11], [12], [19], [20], [23], [39].

4.4.2 Habituation layer

- *Sign process #6 (S6) - Short term habituation mechanism:* The object (O) of this semiosis is the exocytosis of glutamate. The new organic configuration on active zone (sensor neuron of the siphon) is the representamen (R) and an action of short term habituation, the interpretant (I). Consequently, the Rr (at sign process #4) decreases. It happens under a short term habituation rate (Rhs) [11], [12], [19], [20], [23], [39].

- *Sign process #7 (S7) - Long term habituation mechanism:* A certain frequency of glutamate exocytosis (Fg) is the object (O) of this sign process; it results in a correspondent organic configuration (R) in the active zone (sensor neuron of the siphon) and an action of long term habituation occurs (I). Consequently, the Rc (at sign process #5) decreases. It happens under a long term habituation rate (Rhl) [11], [12], [19], [20], [23], [39].

4.4.3 Sensitization layer

- *Sign process #8 (S8) - Trigger mechanism of the interneurons:* The shock on the tail causes reactions in modulatory interneuron that connects the sensor

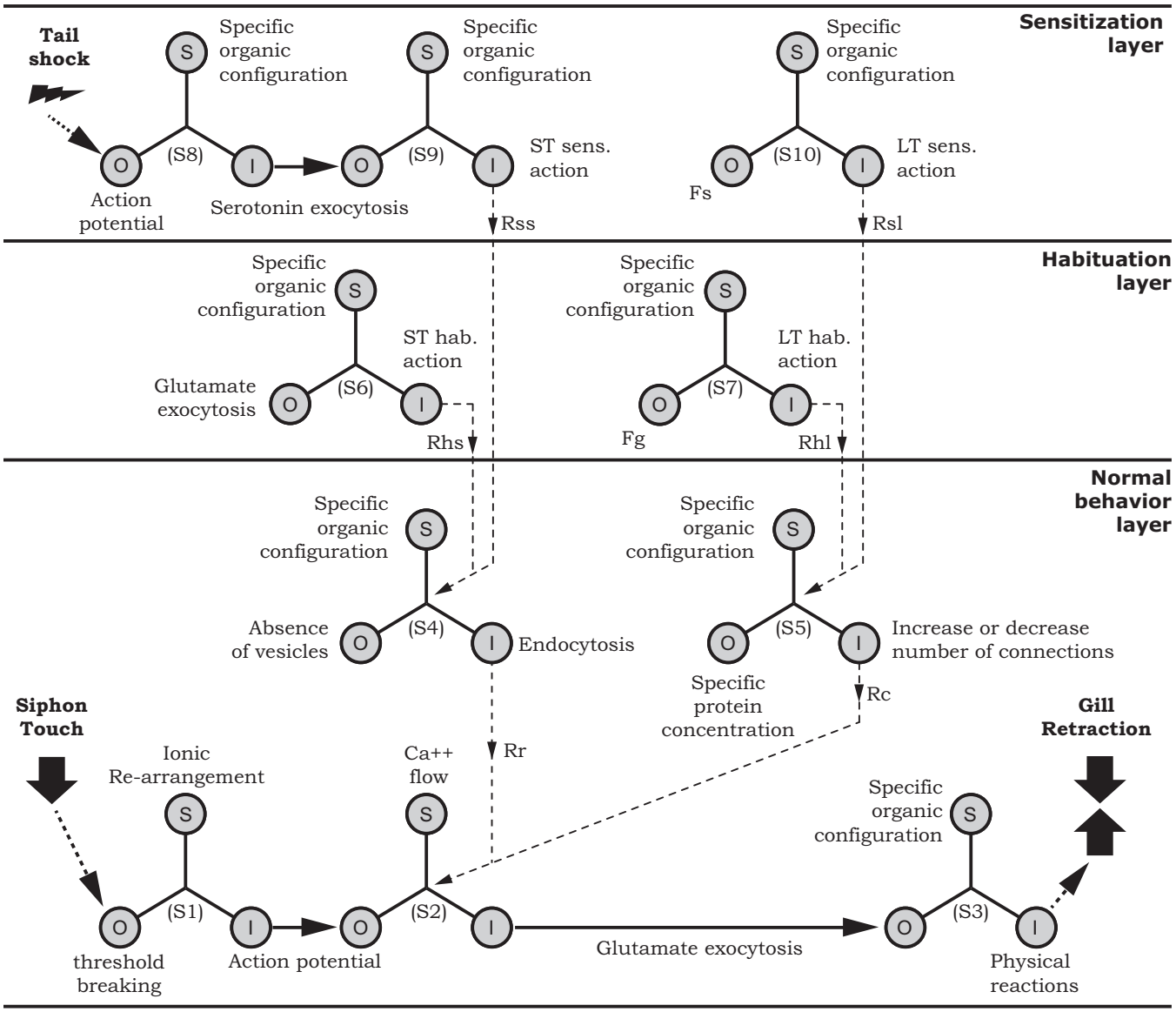


Fig. 1: Semiotic model inspired by the habituation and sensitization behaviors of the aplysia.

neuron of the tail to the sensor neuron of the siphon, modulating the action of the latter. The object (O) of this sign process is the presence of action potential in the active zone of the modulatory interneuron, which causes an organic re-configuration in the active zone (R) resulting in exocytosis of serotonin at the synaptic cleft (I). The amount of serotonin is proportional to the intensity of the shock [11], [12], [19], [20], [23], [39].

- *Sign process #9 (S9) - Short term sensitization mechanism:* The object (O) of this sign process is the presence of serotonin at the synaptic cleft. The new organic configuration on active zone (sensor neuron of the siphon) is the representamen (R) and an action of short term sensitization, the interpretant (I). Consequently, the Rr (at sign process #4) increases. It happens under a short term sensitization rate (Rss) [11], [12], [19], [20], [23], [39].

- *Sign process #10 (S10) - Long term sensitization mechanism:* A certain frequency of serotonin exocytosis (Fs) is the object (O) of this sign process, which results in a correspondent organic configuration (R) in the active zone (sensor neuron of the siphon) and an action of long term sensitization occurs (I). Consequently, the Rc (at sign process #5) increases. It happens under a long term sensitization rate (Rsl) [11], [12], [19], [20], [23], [39].

4.5 Step 5: Semiotic modeling

The sign processes recognized in step #4 give rise to a diagram that represents the behavior of the aplysia regarding to its learning and memory retention ability. The underlying biological strategy is translated in terms of semiosis and their relations (figure 1).

4.6 Step 6: Computational meta-modeling

The computational meta-model (UML based diagrams) follows the previous step. It represents the transposition between biological field and computational field for the phenomenon under study. It is a set of generic diagrams that represents the general strategy abstracted from the biological algorithm. Figure 2, as an example, is a detailed activity diagram corresponding to semiosis S1. Any other semiosis should be represented in a similar way.

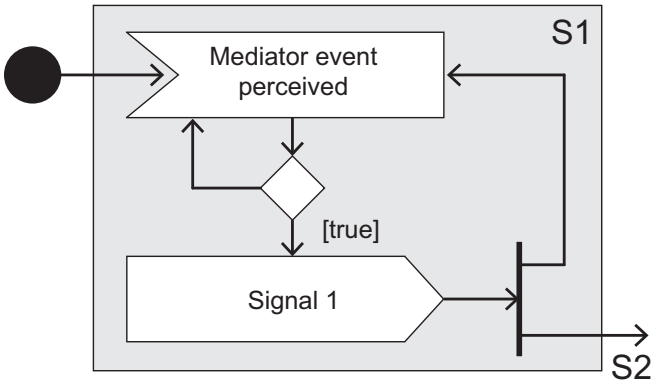


Fig. 2: Activity diagram corresponding to semiosis S1.

Figures 3 and 4 are simplified activity diagrams representing the semiotic chains on modulatory circuit and mediator circuit respectively. And finally, figure 5 is a machine state diagram representing the interactions between both circuits.

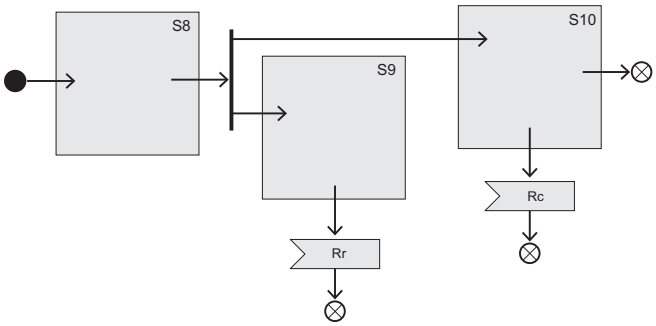


Fig. 3: Semiotic chain on modulatory circuit.

5. Beyond the STM

The target of STM is the meta-model, step 6. It means that a biological algorithm, responsible for a natural behavior

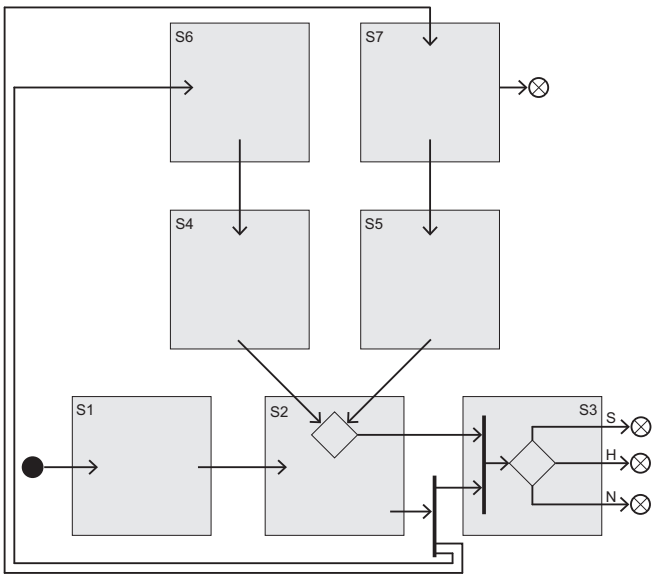


Fig. 4: Semiotic chain on mediator circuit.

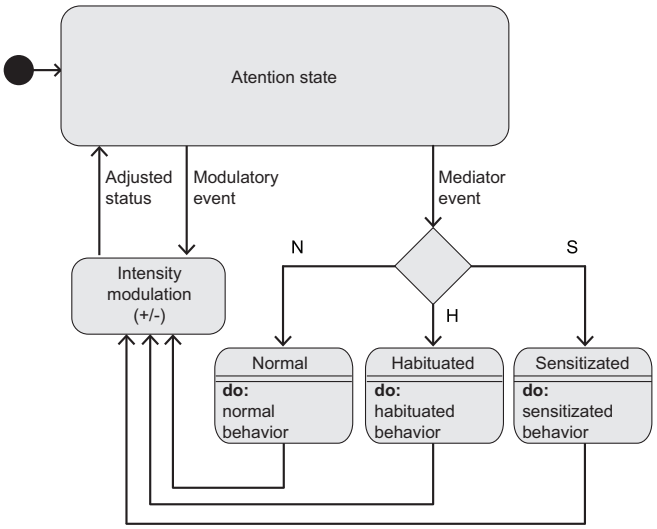


Fig. 5: Machine state diagram.

(function), was abstracted. This abstraction, represented by the UML diagrams, can be used as a clue to find solutions for computational applications in domains where similar behavior is expected. Concerning the case showed above, the learning capabilities of the *Aplysia californica*, an interesting signal-and-noise differentiation system was abstracted, and many domains can be pointed as candidates to receive implementations based on this meta-model. Beyond the domains on step #1, internet of things, prevision of failure on industrial equipment and support for decisions are some possibilities.

6. Conclusion

Biomimetics provides numerous insights to solve complex human problems. The use of natural knowledge, developed over billion years, could be decisive factor in getting most effective and economical systems in all technological areas, including Computer Science. Biological algorithms can be sources of inspiration for the development of softwares to meet the most important issues of the digital world. The Semiotic Transposition Method can be regarded as an interesting tool for developing these biomimetic software, and can be used separately or in conjunction with other tools such as BioTRIZ.

This paper presents the outline of STM, highlighting the importance of the semiotic theory as an intermediary field in the transposition of natural phenomenon from their original biological field to computational field. Future studies and research can contribute to strengthen the STM. Its multidisciplinary characteristic remains opened to contributions from other areas of knowledge and also to even deeper insights based on semiotics theory. The STM was not developed as a closed method, contained in itself, but as a tool opened to future enhancements.

7. Acknowledgements

The authors would like to thank Lucia Santaella and João José Neto for their comments and suggestions.

References

[1] O. H Schmitt, "Some Interesting and Useful Biomimetic Transforms," in *Third Int. Biophysics Congress*, 1069, p. 197.

[2] J. F. Vincent, O. A. Bogatyreva, N. R. Bogatyrev, A. Bowyer and A. K. Pahl, "Biomimetics: its Practice and Theory," *J. R. Soc. Interface*, vol. 3, pp. 471-482, 2006.

[3] J. M. Benyus, *Biomimicry: innovation Inspired by Nature*. New York, USA: Prennial, 2002.

[4] L. N. Decastro, "Fundamentals of Natural Computing: an Overview," *Physics of Life Reviews*, vol. 4, pp. 1-36, 2007.

[5] B. Webb and T. R. Consi, *Biorobotics, Methods and Applications*. Cambridge, USA: MIT Press, 2001.

[6] J. McCarthy et al., "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," *AI Magazine*, vol. 27, n. 4, 1955.

[7] C. Gere, *Digital Culture*. London, England: Reakton Books, 2002.

[8] M. Sipser, *Introduction to the Theory of Computation*. Boston, USA: Thomson, 2006.

[9] J. F. Teixeira, *Mentes e Máquinas*. São Paulo, Brazil: Artes Médicas, 1998.

[10] A. Turing, "On Computable Numbers, with an Application to the Entscheidungsproblem," Princeton University, New Jersey, USA, 1936.

[11] E. R. Kandel, J. H. Schwartz and T. M. Jessel, *Principles of Neural Science*. New York, USA: McGraw-Hill, 2000.

[12] E. R. Kandel, *In Search of Memory: the Emergence of a New Science of Mind*. New York, USA: W. W. Norton & Company, 2006.

[13] C. S. Peirce, *Essencial Peirce*, vol. 1. Bloomington, USA: Indiana University Press, 1992.

[14] C. S. Peirce, *Essencial Peirce*, vol. 2. Bloomington, USA: Indiana University Press, 1998.

[15] W. Nöth, *Panorama da Semiótica*. São Paulo, Brazil: Annablume, 1995.

[16] W. Nöth, *A Teoria Geral dos Signos: Como as Linguagens Significam as Coisas*. São Paulo, Brazil: CENGAGE Learning, 2000.

[17] W. Nöth, *Semiótica no Século XX*. São Paulo, Brazil: Annablume, 1996.

[18] T. Ziemke and E. Sharkey, "A Stroll Through the Worlds of Robots and Animals: Applying Jakob von Uexküll's Theory of Meaning to Adaptative Robots and Artificial Life," *Semiotica*, vol. 1/4, n. 134, pp. 701-746, 2001.

[19] J. von Uexküll, *A Stroll Through the Worlds of Animals and Men: a Picture Book of Invisible Worlds*, In: Instinctive Behavior: the Development of a Modern Concept, pp. 5-80, C. H. Schiller, Ed. New York, USA: International Universities Press, 1934.

[20] J. von Uexküll, "A Theory of Meaning," *Semiotica*, vol. 42(1), pp. 19-48, 1982.

[21] T. von Uexküll, "The Sign Theory of Jakob von Uexküll," *Semiotica*, vol. 89(4), pp. 279-235, 1992.

[22] T. von Uexküll, "A Teoria da Umwelt de Jakob von Uexküll," *Galaxia*, vol. 7, pp. 25-82, 2004.

[23] W. Nöth, *Handbook of Semiotics*. Bloomington, USA: Indiana University Press, 1995.

[24] S. N. Salthe, *Evolving hierarchical systems: their structure and representation*. New York, USA: Columbia University Press, 1985.

[25] C. N. El-Hani and J. Queiroz, *Estruturalismo Hierárquico, Semiose e Emergência*, In: Computação, Cognição, Semiose, pp. 93-128, J. Queiroz, A. Loula, R. Gudwin, Ed. Salvador, Brazil: EDUFBA, 2007.

[26] R. A. Brooks, *Intelligence Without Representation*, In: Cambrian Intelligence, the Early History of the New AI, pp. 79-101, Cambridge, USA: MIT Press, 1999.

[27] R. A. Brooks, *Intelligence Without Reason*, In: Cambrian Intelligence, the Early History of the New AI, pp. 133-186, Cambridge, USA: MIT Press, 1999.

[28] R. A. Brooks, *Elephants Don't Play Chess*, In: Cambrian Intelligence, the Early History of the New AI, pp. pp. 110-131, Cambridge, USA: MIT Press, 1999.

[29] S. A. Korecki, "Inspired Design: Using Interdisciplinarity and Biomimicry for Software Innovation," M. Eng. thesis, School of Computing and Information Systems, Grand Valley State University, Allendale, USA, 2008.

[30] (2014) Biomimicry 3.8. [Online]. Available: http://biomimicry.net/about/biomimicry/conversation-with-janine/

[31] (2014) Unified Modeling Language. [Online]. Available: Unified Modeling Language

[32] E. R. Kandel et al., "Molecular Mechanisms of Memory Storage in Aplysia," *The Biological Bulletin, Marine Biological Laboratory*, 210, pp. 174-191, 2006.

[33] E. R. Kandel et al., "Dishabituation in Aplysia Can Involve Either Reversal of Habituation or Superimposed Sensitization," *Learn. Mem.Cold Spring Harbor Laboratory Press*, 13, 397-403, 2006.

[34] E. R. Kandel et al., "The Molecular Biology of Memory Storage: a Dialogue Between Genes and Synapses," *Science*, 294, pp. 1030-1038, 2004.

[35] E. R. Kandel et al., "Neural Mechanisms of Habituation and Dishabituation of the Gill-Withdrawal Reflex in Aplysia," *Science*, 167(3926), 1745-1748, 1970.

[36] P. C. Zikopoulos et al., *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. USA: McGraw Hill, 2012.

[37] N. Silver, *The Signal and the Noise: Why So Many Predictions Fail - But Some Don't*. New York, USA: Penguin Press, 2013.

[38] P. W. Singer and A. Friedman, *Cybersecurity and Cyberwar: What Everyone Needs to Know*. New York, USA: Oxford University Press, 2014.

[39] R. Lent, *Cem Bilhões de Neurônios: Conceitos Fundamentais de Neurociência*. Rio de Janeiro, Brazil: Atheneu, 2001.

A Generic Simulation Model For High Performance Computing Systems

Taner Dursun
TÜBİTAK BİLGEM UEKAE
taner.dursun@tubitak.gov.tr

Hasan Dağ
Kadir Has University
hasan.dag@khas.edu.tr

Keywords: high performance computing, discrete event simulation, heterogeneous simulation, policy based management, heterosim

Abstract

Most of the academic research on High Performance Computing (HPC) systems has been conducted with simulation platforms because they are generally too expensive and hard to construct systems. Moreover, the research has been generally focused on a specific type of HPC system. This paper however, introduces a general-purpose simulation model that can be used for constructing simulations of the most well-known HPC system types.

In this paper, we propose a new approach that leverages usage of simulation systems for constructing hybrid simulations. In order to arrange heterogeneous simulation executions, the simulation tools are required to allow easy and fast creation of simulation sessions by employing real-time software components beside simulation codes. Although there have been considerable amount of research activity in simulation community, the current simulation tools are not capable of supporting such a cooperation between components working in real-time and simulation-time. We introduce HeteroSim, a simulation model that can execute discrete event simulations by employing both simulation entities and real world software entities. This model offers great potential for many research areas. For instance, we are able to rely on this model in order to build simulations that combine both the simulated elements of an HPC system and previously implemented elements of our Policy Based Management (PBM) framework [1][2]. In this way, it may be possible to study the efficiency and usability of PBM concept on the management of HPC systems.

1. INTRODUCTION

Having been the subject of much research, simulation systems provide a way to evaluate distributed systems under different scenarios such as varying the number of resources and users. In real environments, it may not always be possible, to perform evaluation in a repeatable and controllable way for different scenarios. Because the status of entities continuously varies with time and it may not be possible to control activities of them. Simulation systems, however, enable to perform various studies, such as behavior and performance without building the actual system. In contrast to analytic models, simulation models represent the run time behaviors of the real systems being worked on. In some cases as listed in [3] simulation is preferable to analytical modeling.

HPC systems are expensive and generally have a distributed nature. It is not always possible to take the opportunity of academic research and performing experiments with them. Therefore, simulations are essential for carrying out research experiments in HPC systems. Thus, a number of simulation tools have already been developed. On the other hand, most of these simulation tools have dedicated only to a specific type of HPC system or a sub set of its components. Thus, there is no generic tool to support simulations for different types of HPC systems. Moreover, the existing tools are not flexible and modular enough, because they are not a result of a research effort to develop a generic model for the implementation of different HPC scenarios. Therefore, we have developed a universal simulation model for easy building simulations of well-known HPC systems in order to fill this gap. Section 3 explains the proposed model with a set of execution results.

For more realistic results by involving more accurate data, simulations may need to interact with real-time systems. Based on the degree of human and real system participation, there are three main styles of simulation. These styles are described by [4] as follows:

- ***Live simulation*** involves humans for operating real systems in order to rehearsal or practice with "go-to-war" systems. As in the real world, time is continuous. Testing a real car battery using an electrical tester is an example of live simulation.
- ***Virtual simulation*** includes simulated systems operated by humans. Time is in discrete steps. Human is in a central role (e.g. a flight simulator).
- ***Constructive simulation*** refers to classical computerized simulation model that involves simulated people operating simulated systems. Science-based simulations are constructive in nature. Humans just make inputs to such simulations and are driven by the proper sequencing of events, but have no effect on the outcomes.

Most of the research activities in simulation community have generally focused on simulation frameworks in these three styles. As criticized in [5], this categorization suffers from no clear division between categories and a missing category for simulated entities interacting with real equipment (e.g. smart vehicles). Besides, current simulation tools do not sufficiently support such cooperation between real-time and simulation-time software components. Therefore, tools that enable easy and fast creation of simulations employing both real world applications and simulation codes would be useful. In this paper, as a solution to fill this gap, we also introduce a model rather falling into that missing category. In our model, interactions are allowed between not only real world systems and simulated entities but also humans and the real world systems. That's both simulated people and humans are interacting simultaneously with the same real systems.

A proof-of-concept implementation of the infrastructure for building simulations by involving existing real systems having JMX [6] or JMS-based [7] communication interfaces has already been accomplished. As the second contribution, Section 3 explains our approach to facilitate building of hybrid simulations.

Another aim of our study is to show how the PBM concept [8] responds the needs for effective management of the HPC systems and thus to prove the usability and performance raising effect of our already PBM framework [1],[2]. Therefore, in this study, we have integrated our simulation model with the POLICE PBM framework as described in Section 3.3.

Rest of this paper is organized as follows. Section 2 introduces some background for the reader. The model is discussed in Section 3. Section 4 includes evaluation of the related work and the last section presents our conclusions.

2. SIMULATION BACKGROUND

Simulation concept has been used to model and evaluate real world systems and understand their behaviors. The output of the simulation indicates how a real system behaves without need for its actual implementation.

In recent years, simulation has emerged as an important research area and many simulation tools and technologies have been developed. The simulation tools enable constructing repeatable and controllable environments for feasibility and performance studies.

2.1. Simulation Models

A simulation abstracts all the entities and their time dependent interactions in the real system. Simulation models are generally encoded as discrete event-driven programs. Events are time-stamped messages processed in their temporal order as the simulator progresses through simulated time interval [9]. However, there are many other concepts for the realization of the simulation models. Buyya provides a collection of these concepts and a comprehensive categorization of simulation tools according to an inclusive set of criteria [10]. In this section, we emphasize some classifications related to our work.

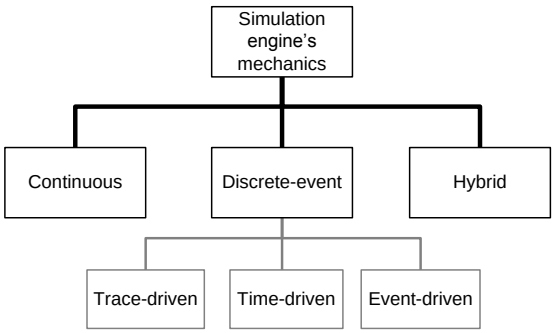


Figure 1. Simulation engine mechanics [10]

Each simulation tool has a simulation engine to execute the simulation model. Figure 1 depicts the Buyya's taxonomy for the mechanics of simulation engines according to their behavior against time. These categories are explained in [10] as follows:

- In a continuous time simulation, state changes occur continuously with time; such systems are usually described and solved by sets of differential equations.
- In a discrete-event/time simulation (DTS/DES), state variables change at instants in time and the system is only considered at these time points (the *observation points*). A DTS is further subdivided into a trace-driven, time-driven or event-driven simulation. A **trace-driven** DTS proceeds by reading a set of events that are collected before from another environment for modeling a system that has executed in another environment. A **time-driven** DTS advances by fixed time increments and is useful for modeling events that occur at regular time intervals. An **event-driven** DTS advances by irregular time increments and is useful for modeling events that may occur at any time.
- In a hybrid (continuous time-discrete event) simulation, the time is (conceptually) continuous and the observation period is a (equally or arbitrarily spaced) real time interval. The discrete changes in the system state take place at these event times. In between consecutive event times the system state may vary continuously.

The Buyya's *modeling framework* criterion [10] describes how the target systems are modeled and how the events are scheduled. Buyya defines three categories for scheduling operations of a simulator as follows:

- An **entity-based** (Process-oriented) modeling involves simulation entity concept in order to logically encapsulate processes of target system. Entities communicate with others via messaging to perform their tasks. HeteroSim falls into this category.
- In an **event-based** modeling framework, there is a procedure associated with each type of event in the system. The system performs the action required to handle that type of event and every time such an event occurs the same action is performed.
- Hybrid models of the categories above are possible.

2.2. Simulation Tools

In order to achieve realistic simulations, a simulation model has to be as identical as possible to simulated system. For satisfaction of this requirement, simulation tools should support object-oriented approach, which is well-suited for modeling real world entities in software. There are already various frameworks that combine the concept of objects with the concurrent computation in which simulation entities run concurrently. This process oriented approach is addressed by most of the discrete event simulation tools. Because of its strong support for multi-threaded programming and native schedulers, many simulation tools[1] have been developed with Java language [11].

JavaSim [12] is an open source tool developed at the Department of Computing Science, University of Newcastle upon Tyne for building discrete event process-based simulation. JSIM [13] is another Java-based simulation and animation environment supporting Web-Based Simulation. Simulation models may be built using either the event or the process oriented approach. In addition, a visual designer allows process models to be built graphically.

J-Sim [14] is a Java-based sequential network simulator developed at the Department of Computing Science of the Ohio State University. It is based on the component-based software architecture, Autonomous Component Architecture (ACA). Unlike JavaBeans, CORBA, and COM/DCOM, the components in J-Sim are loosely coupled, communicate with one another by "wiring" their ports together (Figure 2). J-Sim can be used for both discrete event simulation and real-time process-based simulation. **Silk** [15] and **SimJava** [16][17] are two early Java-based libraries for process-oriented discrete-event simulation. **Silk** is a complete package including JavaBeans components for visual modeling and a variety of methods for entity generation, resource scheduling and output post-processing.

[1] In the academical simulation world relying on Java language, there have been naming ambiguity. Many tools have names including words of "Java" and "Sim", such as JavaSim, SimJava, and J-Sim.

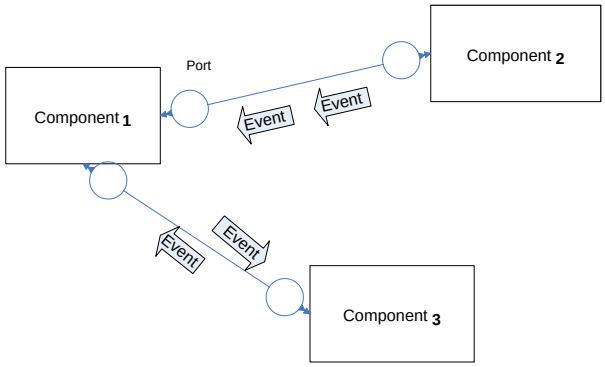


Figure 2. Simulation elements

SimJava is a general-purpose package that is a Java implementation of a C++ library called HASE++ (Hierarchical computer Architecture design and Simulation Environment). A SimJava simulation contains a number of simulation entities each of which runs in parallel in its own thread. As in J-Sim, these entities are connected to each other via ports and can communicate by sending and receiving passive event objects through these ports.

3. PROPOSED SIMULATION MODEL

3.1. Generic HPC Simulation Model

The first members of HPC family are super-computers (or parallel systems), which are generally composed of a great number of processors, large memory modules and a high speed communication bus. Due to high cost of super-computers, an alternative family, named as Distributed Computing Environment (DCE), has been emerged. As being member of DCE family, networked virtual supercomputers (meta-computers) and distributed network computing (NC) systems are virtual computers that are composed with heterogeneous machines for sharing their resources for a common purpose. Grids, however, are very large scale, Internet wide, distributed NC systems, whose machines belong to different administrative domains. They combine characteristics of both distributed and parallel systems.

A generic, abstract model that can be used for definition of both distributed computing environments (DCEs) and Grids is already proposed in [18]. This abstract model contains several functional units and four interfaces: resource consumer interface (RCI), resource provider interface (RPI), resource manager support interface (RMSI), and resource manager peer interface (RMPI). The resource is a shared entity or capability that is employed to fulfill scheduled job or resource requests. It could be a machine, network, or some service that is under control of a Resource Management System (RMS). An RMS is defined as a service that is provided by a distributed NC system that manages a pool of resources. *The resource consumers* that

interface via RCI can be either actual applications or another RMS that represents a 'higher' layer. It is an agent that controls the consumer. The *resource provider (broker)* that interfaces via the RPI can be an actual resource or another RMS that represents a lower layer. The *support functions* such as naming and security can be accessible through the RMSI. The RMPI is intended for interaction with other RMSs and may support several protocols including resource discovery, resource dissemination, trading, resolution, and co-allocation. Figure 3 shows the abstract model and a sample system with multiple interconnected multiple levels RMSs.

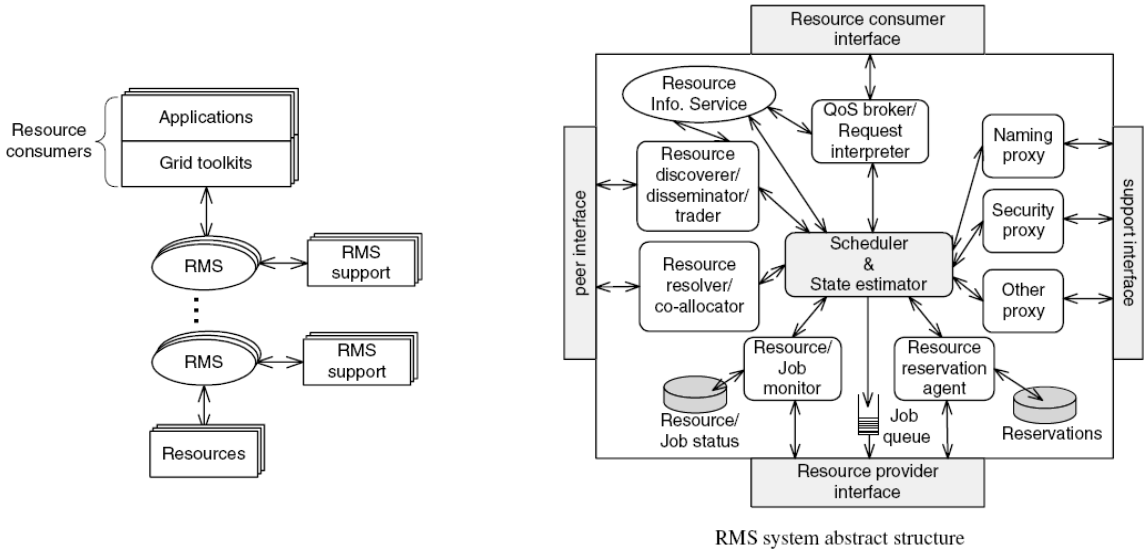


Figure 3. An abstract RMS model and its use for Grid modeling [18]

Although many HPC architectures could be specified with this abstract model, to be generic enough it needs some improvements such as support for distributed ownership of the Grid resources. Traditional DCEs are generally established in a single administrative domain and accept jobs of clients that belong to the same organization. Thus, job priorities could be simply used as QoS mechanism in traditional DCEs. However, in a Grid, providing QoS to clients from different administrative domains is more challenging. The issues such as access privileges of jobs, and resource requirements shall also be taken into account [19]. Thus, different administrative domains concept and handling semantic relationships between components of HPC systems should also be supported in this model.

Therefore, as the first contribution, we developed a general purpose, policy-based manageable simulation framework for HPC systems by realizing an enhanced version of the abstract model above. The following main improvements and simplifications are applied to this abstract model:

- Adding observation capability to Resource/Job Monitor in order to observe resources of a peer RMS.
- Adding an Inter-Domain Interface for interaction with RMSs belonging to different administrative domains by using various protocols.
- For simplification purpose, by combining some functionally close modules into one module, an easily implementable model has been obtained.
- To achieve policy based management capability for each component of RMS, adding PEP (Policy Enforcement Point) interface as a support service. Therefore, each component of the HeteroSim RMS is instrumented for policy enforcement.

Each component of our abstract RMS simulation model shown in
Figure **4** can be replaced with a different implementation thanks to the abstract interfaces we implemented. Furthermore, by connecting our RMS nodes, hierarchically or peer-wise and then grouping them into administrative domains, it is possible to construct policy based manageable simulations of centralized, distributed and hierarchical HPC systems. Figure 5 shows a sample HPC scenario including mixed types of peer and hierarchical RMS relationships that is built with triple RMS nodes.

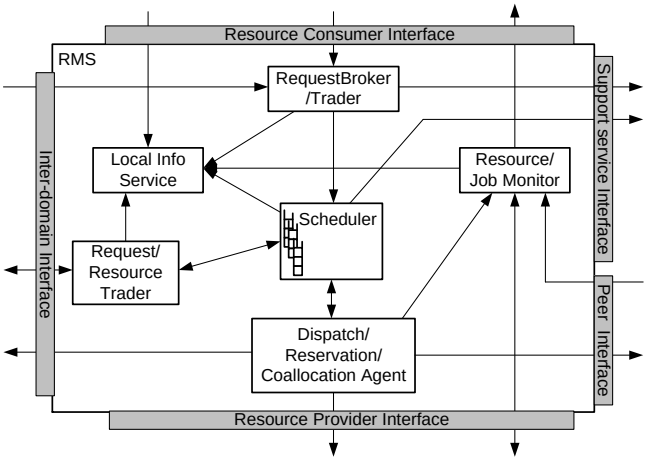


Figure 4. RMS abstraction model (revised version of the model proposed by [18])

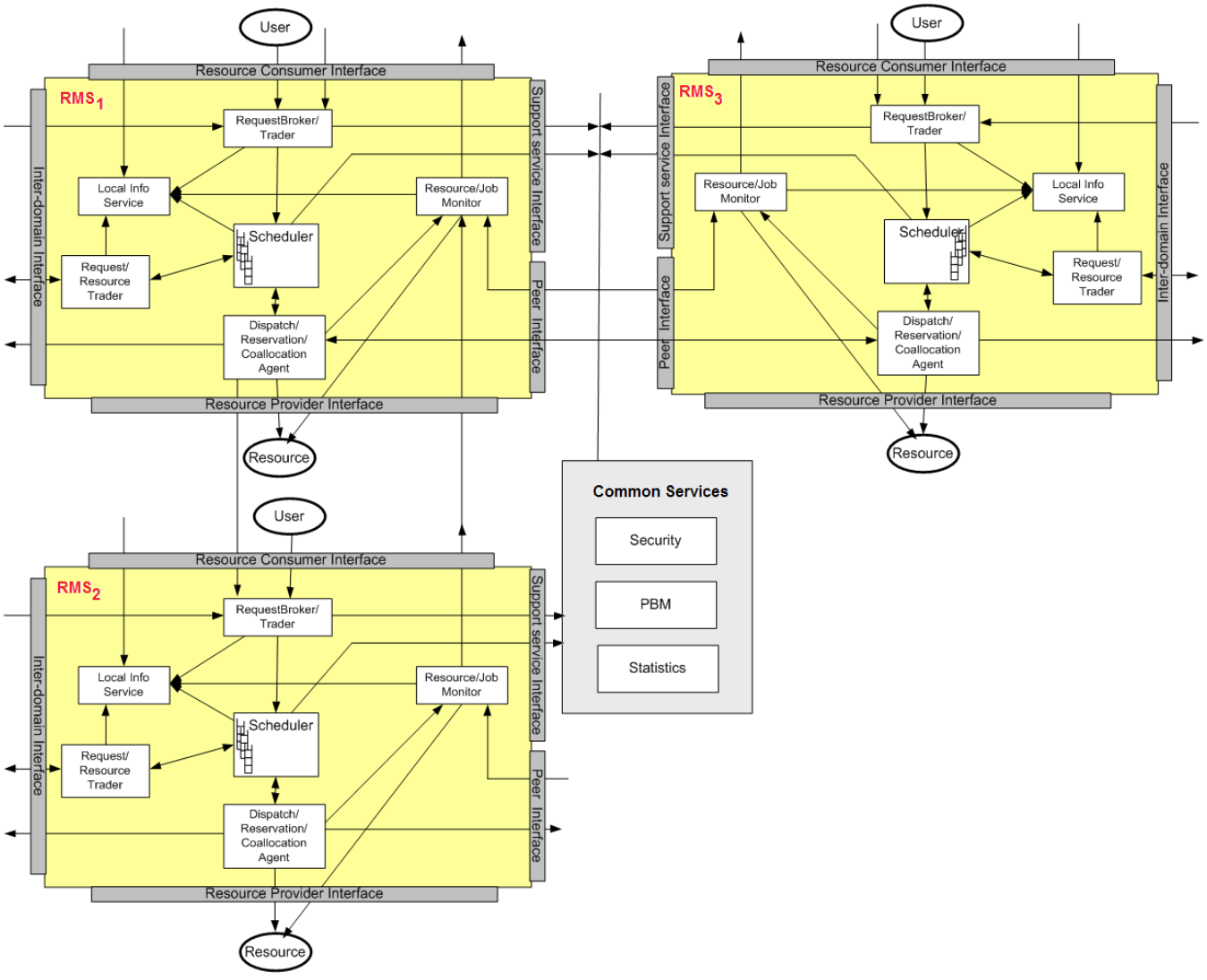


Figure 5. A sample HPC scenario established with HeteroSim RMS nodes

Another significant feature of the HeteroSim framework is its support for hybrid simulation in which both simulation codes and real application software can participate. In this manner, for example, we are able to make our already developed POLICE PBM application (real world application) communicates with entities of HeteroSim simulations.

As the first step of the development of HeteroSim platform, we have surveyed academic and open-source simulation tools that offer process-based discrete event simulation base, and then we have selected GridSim [20] toolkit as the starting point of our development work. Having larger academical background, the GridSim has the benefits of relative simplicity, extensibility, and Grid based HPC domain support. Then, we have developed a Java-based discrete-event simulation toolkit called HeteroSim that is a redesign of GridSim Toolkit [20] from a new point of view. The refactoring results in the following main improvements:

- Support for modeling and simulation of heterogeneous types of entities, from both simulation world and real world was added.
- More detailed resource modeling (finer granularity). The resource represents a set of machines in GridSim framework and is managed by a Scheduler. However, in HeteroSim each machine can be modeled as a separate resource. By means of this capability, the resource concept is changed from "Resource including Machines with Scheduler" to "the machine with a number of PEs (Processing Elements)".
- Management interface for POLICE PEP was added and enforcement activities were implemented. Policy based manageable versions of all RMS components were developed. Therefore, it is possible to change behaviors of simulation entities with policies.
- Local and remote job separation was added
- Added new entity types of Trader, RMS, Dispatcher, and Local Information Service. Therefore, complexity of the Scheduler is decreased by moving the functionalities to where they must exist. Jobs are not submitted to directly Scheduler anymore but to the Trader.
- Automatic RMS topology construction according to configuration specified before simulation start.
- Support for RMS hierarchy deeper than two levels was added.
- Mechanism for adjusting the simulation execution speed mechanism was implemented
- Entity synchronization and multi-thread support were improved.
- Allocation of jobs to multiple PEs belonging to different resources is now possible. Moreover, instant mapping a job to multiple RMS resources is also possible.
- For resource allocation, a number of predefined strategies such as LONGEST_FIRST, LOCAL_FIRST, EQUAL, LEAST_REMAINING_FIRST and REMOTE_FIRST can be triggered via policies.
- A new resource addressing schema is used such as HPC1/RMS2/Resource6.
- Virtual Organization (VO) attribute was added to the Resources and to the Users entities.
- ResourceLoader was developed and tested with real workload traces from various HPC centers.
- Mechanisms implemented for on the fly calculation of metrics based on HPC system, RMS and users:

based on / metrics	RMS	User	HPC
AET (Average Execution Time)	√	√	√
AWT (Average Waiting Time)	√	√	√
AJL (Average Job Length)	√	√	√
TJL (Current/Cumulative Total Job Length)		√	
MakeSpan			√
Max, Min, Average Job length			√
System workload ratio			√

- Calling the added metrics within the policies was supported.
- Execution results are shown with charts implemented with JFreeChart [21] library for each statistical metric defined above. In addition, ResourceAllocation chart was also implemented. The sample charts are shown in Figure 6 and Figure 7.

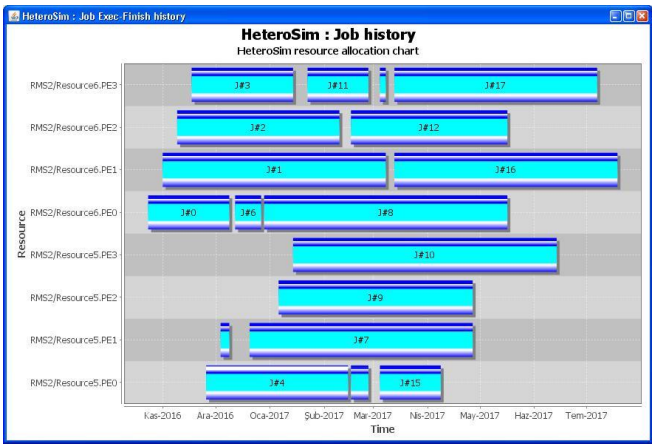


Figure 6. HeteroSim resource allocation history chart

HeteroSim follows the process-oriented approach of SimJava, where each SE (Simulation Entity) can be considered as a separate process. The SEs are the Java

objects which have an independent thread of control associated with them (pseudo-parallel execution) and they use event-based messaging.

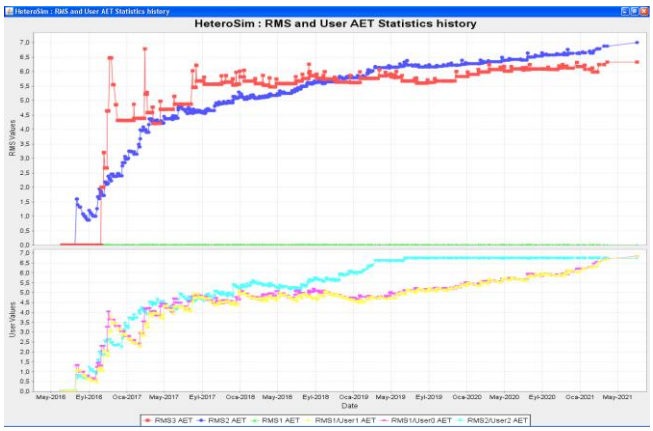


Figure 7. HeteroSim AET metric chart

There is a central event queue, called FEL (Future Event List) that contains timestamp ordered events and is operated by a scheduler. The scheduler observes the FEL and finds the event with the smallest time stamp and invokes the entity related to that event. After all entities have been executed for the current instance of simulation time, the scheduler pops the next event off the queue and advances the simulation clock. This flow continues until no more events are generated.

In the rest of this section, we first outline the basic system architecture developed to implement the HeteroSim model, followed by a detailed explanation of mechanism through which the interactions take place between simulation and real world entities. We then, present how to employ the model for the study on PBM of HPC systems.

3.2. Heterogeneous Simulation Architecture

As shown in Figure 8, HeteroSim architecture consists of four types of components: an adaptation layer, a simulation runtime, simulation entity and real world entity [22]. The real world entities (REs) represent external real time systems and are able to interact with the simulation entities (SEs) during a run. The actions of REs can influence the outcomes of the simulation. Controversially, the SEs may also cause state changes in REs. This bidirectional interaction ability makes it possible, for example, to create SEs including test logic for testing the REs or to exploit existing REs to build simulations quickly instead of implementing them as SE.

The interactions between simulation entities and real system components occur through a common communication mechanism, so called Adaptation Layer (AL) in our model. The AL facilitates expansion of simulation to external real world applications. Instead of establishing direct connection from SEs to the real world applications, AL communicates with the real world applications on behalf of SEs. It supports multicast, broadcast and unicast communications.

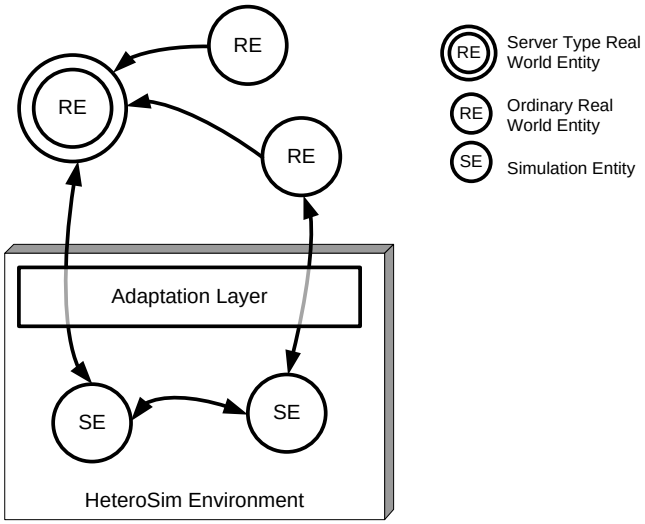


Figure 8. HeteroSim simulation architecture

Similar to SEs, the REs are also parts of the simulation sessions. The REs can be either an ordinary or server type entity. The Server type RE (SRE) can be any kind of server application such as an RDBMS Server, an SNMP Server, a Web Server, a Messaging Server, and so on that SEs or other REs can consume its service. For use with different types of external applications, the AL provides various alternative communication technologies for the SE and RE interaction as shown in Figure 9.

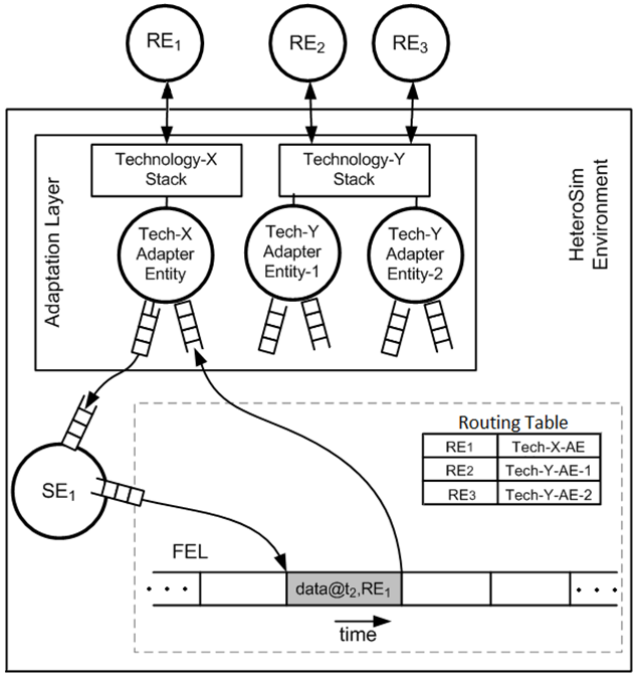


Figure 9. Interaction between SEs and REs through AL

In order to obtain a communication path, a proxy entity (so-called Adapter Entity) for each RE is automatically created and registered in a routing table when the simulation starts. The proxy entities are responsible for translating the communication between SEs and REs transparently. While the proxy entities communicate with REs via proper technology such as SMTP, JMS [23], and RPC (Remote Procedure Call), they also play SE role in order to interact with SEs (i.e. via messaging). Thus, the codes of SEs that interact with REs do not include any code which is specific to the communication technology.

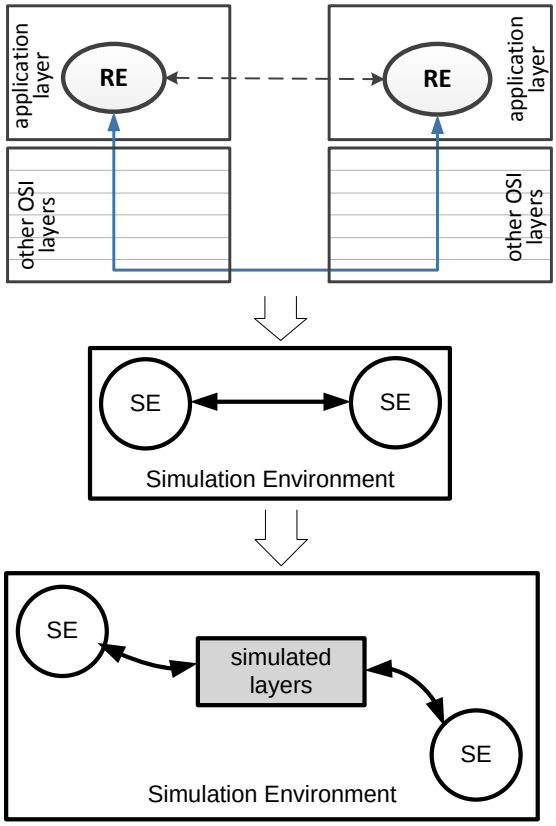


Figure 10. Obtaining a more accurate simulation

For simplicity purpose, the simulation frameworks generally focus on modeling only application-layer level activities of real world applications. However, accurate simulations are achieved only by simulating as much communication layer as possible, in addition to application-layer behaviors, as depicted in Figure 10. For increasing accuracy of simulations, the AL can also be used for communication between SEs.

Because the AL can already employ the original protocol stacks related to communication technologies within simulation sessions, HeteroSim can also implicitly help to simulate the communication mechanisms through which the REs communicate without writing any additional code.

A sample simulation scenario for the case of the Figure 10 can be arranged as shown in Figure 11. In this sample scenario, the SE_2 should select the path over AL so that the messages are to be transmitted to SE_1 via communication layers associated to the Technoloy-X.

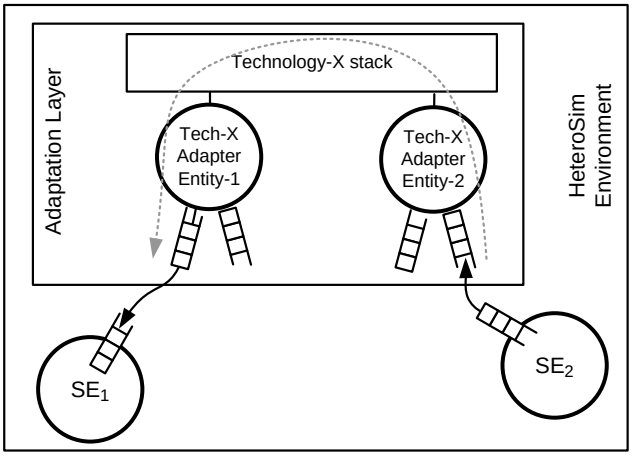


Figure 11. SEs communicating via Technology-X

Another capability provided to users by HeteroSim is the BeanShell [24] scripts based graphical user interface for interacting with simulation run.

For the validation purpose, modeling results must be compared with the corresponding field observations to ensure that the model realistically represents the real system. Therefore, we have used various real workload traces obtained from well-known supercomputing centers to test our HPC simulation model.

3.3. Issue of Simulation Time

As summarized in [9] and [5], there are two main types of simulation execution model according to time treatment:

- *Logical-time simulations* explicitly model the passage of time inside and allow the rate of their advancement to be dictated by the granularity of simulated events. This category is also named as simulation time execution (dual/application clock) in which the progress of time depends on the progress of the application [9]. The simulation time does not advance to the next discrete time point unless all codes for the current simulation time have been executed. HeteroSim falls into this category. Logical-time simulations have the classical discrete event or continuous simulation data structures and algorithms.

- *Real-time simulations* using the real clock values to drive their execution are further divided into two sub categories [9] :
  - In the actual time execution (standard Java execution) simulations, the progress of the application and the passing of time are independent. Due to unpredictable parameters such as interrupts, cpu load, and I/O events, the program can advance at a variable rate. Moreover, for the interpreted languages, program is not guaranteed to progress timely due to survival activities.
  - Real time execution: the application progress depends on the passing of time. The runtime guarantees that instructions will meet given deadlines. Real-time execution simulations resemble real-time systems and their execution is measured by hertz frequency.

The HeteroSim simulations permit interactions with the real-world applications and the events related to them are processed by REs in the simulation time. Therefore, the real time clock should also be taken into account, in addition to the simulation clock. Otherwise, the execution of simulation may cause inconsistencies for REs. During an interaction between external applications and SEs, any delay in simulation side may infect the operations of real world systems negatively, or vice versa. As an example, the enforcement of scheduling policies including Date/Time constraints may not be performed in a consistent way due to two different time domains.

In fact, no system can *guarantee* simultaneous faithfulness to both simulation time and real time. When the required time to compute the next state exceeds the amount of real time available before the next state should occur, changing execution rate at some ratio to real time (for example, by injecting specialized events on the event list or by adjusting frame rate), degrading or abandoning next-state computation, ignoring the delay and, if capability exists, attempting to catch up later by running faster are the possible choices for keeping synchronization with real time.

Simulation time and real time combination is still an open issue for HeteroSim. However, in the current implementation, the events related to the real world systems are given the highest priority and are processed immediately. Additionally, as a workaround, it is possible to define bi-directional timeout values for interaction with real time applications.

On the other hand, the authors of [9] propose a new approach, called JiST (Java in Simulation Time) for constructing discrete-event simulator by embedding simulation time semantics into the standard Java language to achieve performance. Thus, it can be possible to write simulations by using simplified JiST API wherein the simulation codes may contain functions related directly to real time. Then, the JiST framework modifies the standard Java-byte codes of the simulation programs written in Java and embeds its execution semantic. As a future work, we plan to study to involve the actual time concept mentioned in [9] to solve the possible performance problem and to help the solution of dual-time-domains problem.

3.4. Using HeteroSim in PBM simulation

The number of parameters that should be configured in an optimal way and consequently management overhead increase in parallel with capabilities and size of traditional HPC systems. Especially, due to highly heterogeneous components, distributed ownership of the resources, and very dynamic conditions, Grids have much more management requirements. In order to deal with this need, employing PBM tools should be the first alternative that comes to mind. The easily handling of processes such as access control for resources, allocating jobs, and sharing resources, according to management goals can only be realized with policy based management of HPC systems. In fact, for the management of HPC systems many tools have been developed as a built-in or supplementary mechanism. Despite policy and strategy words are addressed frequently by these tools, they do not actually fulfill the characteristics of a real PBM tool as defined in Literature. Therefore, for the management of HPC systems in conformance with PBM concept, we have intended to use the output of our previous study which is a general purpose PBM framework, so-called POLICE [1] [2] whose architecture is shown in Figure 13.

The PBM offers an effective way for management of systems wherein the desired system behavior is specified as policies by administrators. Containing high-level, human-friendly terms, policies are rules to administer, manage, and control the resources. The PBM system automatically translates policies into commands and configuration parameters that are understandable to the managed devices.

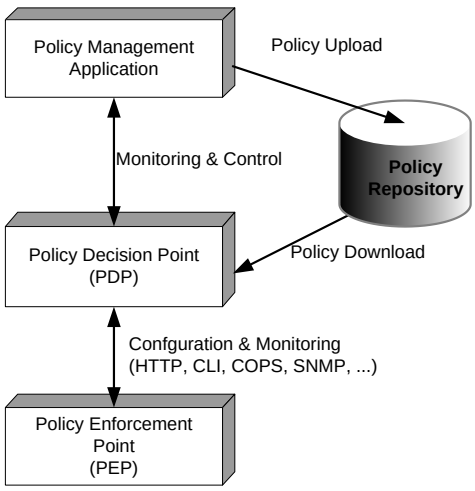


Figure 12. General Architecture of a PBM System [25]

The information model and architectural blocks published by IETF to describe network policies, and services [8], [25] have been commonly accepted in the PBM community (Figure 12)

The policy management applications allow administrators to specify the policies, translate the input into a common format and store them in the policy repository. A PDP, on the other hand, retrieves policies from the policy repository, interprets the policies, sends them to the PEPs for enforcement and replies the policy decision requests from PEPs with policy decisions. PEPs (residing on managed resources) act according to the PDP's decisions and actual system conditions in order to enforce policies.

HPC systems need to be evaluated under different scenarios such as varying the number of resources and users. However, in a real HPC environment, it is hard, and perhaps even impossible, to perform evaluation in a repeatable and controllable manner for different scenarios. Because the availability of resources and their loads continuously vary with time and it is impossible to control activities of users especially the ones belonging to different administrative domains. Moreover, HPC systems are very expensive and it is not always possible to have opportunity for academic research with them. That's why, we selected to use simulation instead of real HPC systems. Therefore, we have developed the generic simulation model for HPC systems whose details are given in Section 3.1.

Then, thanks to our simulation model allowing simulation entities to interact with real software applications, we have been able to exploit our PBM framework in order to manage the simulated HPC components with policies. In this simulation setup, POLICE components, as the real application elements, interact with the simulation entities representing HPC components. In this manner, we are able to show the effectiveness of POLICE framework on HPC throughput and the usability of POLICE for management of HPC systems.

RMS nodes typically contain various functionally related components. That's why, with our RMS model, not only scheduling components but also all functions and components of an HPC system can be managed as a whole. This is a must for consistent management of HPC systems. Otherwise various inconsistencies may occur. For example, in case for which access control is performed by policies and scheduling is not, if the jobs are assigned the resources that they do not have access privilege, this fact remains unknown until the jobs are sent to the related RMS. All management functions shall be performed with enterprise wide policies which are collection of access control policies, scheduling policies, and so on.

Thanks to this PEP expansion, RMS hierarchies belonging to different administrative domains can be constructed. By this way, site autonomy can be achieved. Each management server (PDP) allows specification of policies for its own administrative domain. Inter-domain resource sharing can be performed according to the policy negotiation, SLA, etc. This negotiation concept is out of scope of this paper but it is noted as a future work.

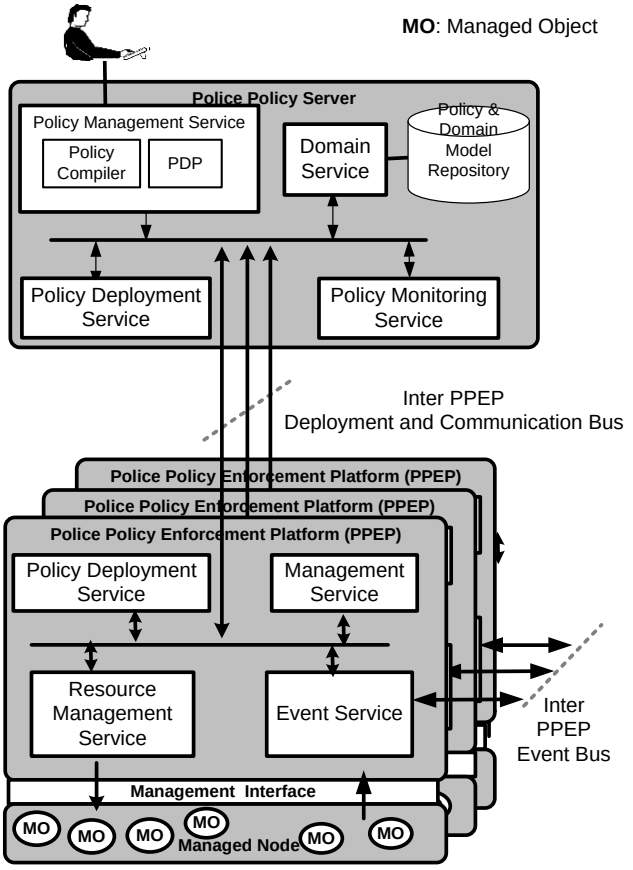


Figure 13. Architecture of POLICE framework

By employing the mechanisms mentioned above, only resource allocation policies can be defined. However, semantic relationships (acyclic task graph, etc.) between resources and jobs that may pose limitations for resource allocation and job execution also must be considered. Handling this kind of semantic relationships for an error free resource sharing can be performed with E-Code concept which is already provided in POLICE PBM framework [1]. For this purpose, HPC system administrators can specify the actions for semantic constraints as E-Code that must be executed during resource management.

In order to enforce policies, we have integrated HeteroSim environment with POLICE framework by using the AL mechanism of HeteroSim. Figure 15 shows interconnection schema with a sample scenario including an RMS and its dedicated PPEP. In this way, the real world components of POLICE interact with the SEs representing HPC components. Figure 15 includes more details on integration schema of POLICE PEP and HeteroSIM.

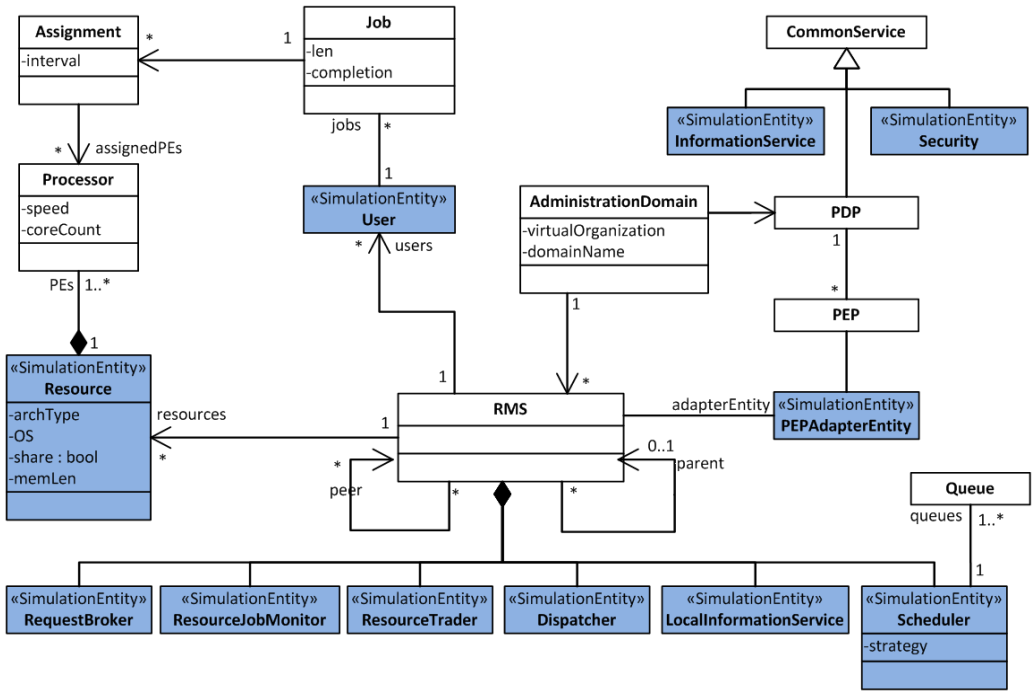


Figure 14. The information model for HPC modeling

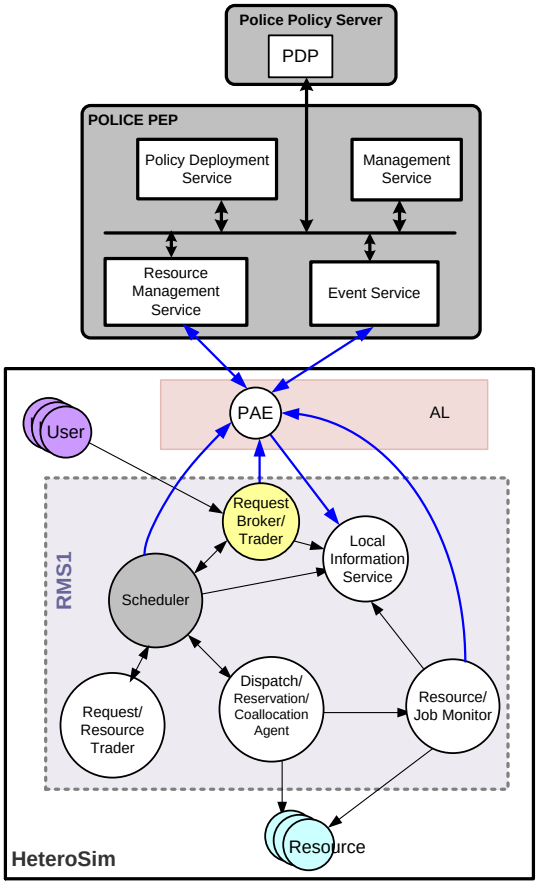


Figure 15. Integration POLICE PEP and an RMS within HeteroSIM

Figure 14 shows the summary information model of our HPC simulation framework. Each RMS is associated with a PAE (PEP Adapter Entity) that plays proxy entity role for the communication with POLICE PEP which is designed for this RMS. The managed node to be controlled by PEP is the RMS and the Managed Objects (MOs) are the components of this RMS. During simulation startup, the PAE registers the MOs with the PEP. The primary actor objects are the users and virtual organizations whereas the target objects are the resources. With this model, the following components and processes can be referenced within POLICE policies:

- Job (job start, stop, remove, migrate, priorities and scheduling)
- Job queues (strategy, priority, etc.)
- Consumers (Users)
- Resources
- Reservations
- Scheduler
- Broker (for QoS/SLA)
- Trader
- Resource allocations
- Access Control

Police policy language is extended with the keywords of HPC, Statistics, RMS, and Scheduler. Sample statements that can be used in policies as follows:

- `rms1.scheduler.awt(user)`
- `rms1.scheduler.awt()`

- `rms1.hpc.awt()`
- `getResponsiblePEPofMO(rms1.hpc)`
- `findQueueName("rms1/hpc")`
- `rms1.user.awt(user3)`
- `hpc.rms2.user1.aet`
- `rms1.scheduler.awt()`
- `rms1.scheduler.hpcAWT()`
- `rms1.scheduler.userAWT(userX)`
- `hpc.awt`
- `hpc.loadRatio`
- `rms1.scheduler.`
  `setJobSelectionStrategy("LocalFirst")`

For example, the following POLICE policy changes the job selection strategy followed by the scheduler component of the RMS3 when the average job waiting time for User1 is greater than zero so that the longest jobs could have priority.

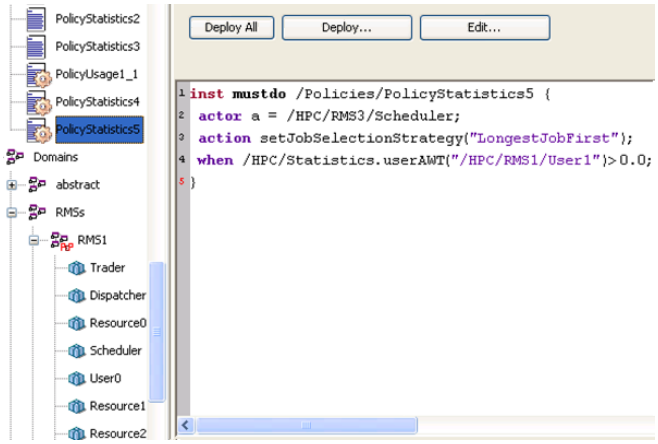


Figure 16. Sample policy for HPC management

4. RELATED WORK

In recent years, modeling and simulation has emerged as an important research area and various standards, methods, tools and technologies have been developed. However, there have been quite a few studies [11][26] dealing with bringing simulation and real world codes together in a simulation session. Except few tools which are already hard to integrate with external real-time systems, most of the simulation tools contain only simulation codes. A mature solution has not appeared yet. Distributed SimJava [26] is the model proposed to make distributed simulation sessions possible. It extends SimJava with distributed simulation capability by employing the Java RMI (Remote Method Invocation). In this way, the SEs belonging to different simulation nodes may interact with each other. It is also possible for the SEs to interact via RMI interface with the real world components which are actually running the real application codes. The main difference of this model to our approach is that it requires the real-world participants of the simulation to implement the RMI interface. Moreover, the real applications can only be seen as external systems providing services to be called by SEs and they cannot participate in a simulation as a simulation entity.

Another study [27] also proposes a distributed simulation model including a DES-based remote simulation service and a representation service interface for real applications *to* interact *concurrently* with that system. Simulation objects interact with remote, asynchronous subscribed clients in order to produce representations of the simulated system. Although this model provides mechanisms to external users for either *passive* (the user only visually monitors the output of the simulation) or *active* (the user is able to interact with the model during a run and will then influence the outcomes) interaction with a simulation, it does not provide any software interface for bi-directional communication between simulation codes and existing real applications. The simulation service accesses external information systems just to obtain data related to the simulation.

On the other hand, there are several works in Literature which bring heterogeneous and simulation words together as we do. However, none of them does it for the same purpose as ours. The authors, for instance, proposed [28] a simulation model consisting of a **heterogeneous simulation environment** (HSE) that integrates a variety of simulation and analytical models of a manufacturing system. The HSE executes each model sequentially so that the output from one model can be used in the input for the next model.

Another work [29] describes a heterogeneous simulation framework in which a collection of heterogeneous simulation models such as conventional simulation models and the DEVS (Discrete Event System Specification) models communicate with each other via DEVS bus. It includes a DEVS/CSIM simulation protocol converter for communication between DEVS model and conventional model. This work only focused on simulation time domain but it didn't deal with the adaptation problem of time domains.

The last example [30], which uses the same name as our model, describes a simulation model named, HETEROSIM to replicate the field conditions of heterogeneous (car) traffic flow that does not follow traffic lanes. That's why they named their model as heterogeneous.

For simulation of HPC systems, a number of simulation tools have already been developed, such as GridSim [20], SimGrid [31], and MicroGrid [32]. However, most of these simulation tools have dedicated only to a specific type of HPC system or a sub set of its components such as scheduler, broker, and so on. Thus, there is no generic tool to support simulations for different types of HPC systems. For example, GridSim[20] and ChicSim [33] data grid

simulation system, were developed only for simulation of Grid systems. **Ganglia** discrete event simulator [34], popular **Ganglia Cluster Monitor** [34], and GangSim simulator [35] are all tools having rather focus on the cluster-wide scheduling. Moreover, they are not flexible and modular enough, because they are not a result of a research effort to develop a generic model for the implementation of different HPC scenarios.

5. CONCLUSION AND FUTURE WORK

Simulation designers may need to involve real-time systems into simulations for several purposes such as reducing building time of simulations, achieving more realistic simulations, arranging simulations of partially implemented systems, and so on. Therefore, it would be useful to have simulation models that enable easy and fast creation of simulation sessions by employing real software components besides simulation codes. Current simulation tools are not sufficiently capable of supporting such cooperation. Our proposition in this paper, however, is able to overcome this shortage with its extensible model for heterogeneous simulations. When preparing a HeteroSim simulation, the implementation of the interactions between SEs and REs are hidden from the modeler's view. Unlike the other proposed systems, in our model, starting connections from both simulation and real world sides is possible.

Having an extensible Adaptation Layer makes our proposition suitable to be used with any real world application. For the time being, we have only implemented and tested a JMS (Java Message Service) [23] based proxy entity for real world applications. Other types of proxy entities can be easily added to the architecture.

Many research areas can take the advantage of the infrastructure proposed in this paper. Scenario-based software testing [36] would be an interesting area to apply our model. For a typical use, all testing scenario can be easily implemented within simulation entities (SEs). Then, through the AL, the SEs can interact with the real world applications to be tested.

Simulations are essential for research experiments in HPC systems. We have developed a universal simulation model for building simulations of HPC systems. With this model, well-known HPC architectures can be created easily either by specifying the configuration of the target architectures manually or by providing workload trace files. Beside the manual experiments, we have also successfully performed simulations with the workload traces of various supercomputing centers. We have been able to execute simulations and observed their results on the graphical charts produced automatically during the simulations. This universal model also contains structures related to policy based management so that by defining policies in POLICE PBM system, we have investigated the effects of them on the behaviors of the simulation entities. Our model can be used for future research on both PBM and HPC domains.

Cloud computing is a new and emerging form of distributed resource sharing. Thus, the same challenges as the Grids have to be overcome. The following *list* is *not exhaustive*, but for the majority issues to be handled in cloud domain:

- Cloud service auditing, monitoring, and metering.
- Mobility management in cloud scenarios.
- New models and paradigms for cloud service management.
- Novel and emerging standards for interoperability between clouds and management of cloud federations.
- QoS/QoE and SLA management in the cloud.
- Secure and private management of cloud data.

Thus, enlarging our model to allow modeling the cloud computing architectures would be a good research opportunity.

On the other hand, our model can be extended and investigated further. We plan to concentrate on especially in the following issues:

- Developing more types of proxy entities for the AL related to mostly used technologies area such as RMI-IIOP, JDBC, and so on,
- Adding distributed simulation support,
- Performance analysis of Adapter Entities,
- Further exploration of the problem of integrating applications running in different time-domains, and investigating solution alternatives,
- Finding a way to synchronize the events occurred in different time domains such as real clock and simulation clock.
- Testing how the proposed model meets the requirements of various domains and applications,
- Improving existing monitoring tools in order to follow visually the activities performed within AL.
- Developing SE templates and structures for scenario based software testing,
- Taking the advantages of heterogeneous simulation sessions for software testing might be a study target. If the test logic is prepared as the simulation codes interacting with the real-time application being tested, it could be possible to benefit from the simulation tools.
- Real world implementation of the RMS node whose simulation model described in this paper.
- Providing a visual tool for preparation of HPC simulation scenarios.

- Using outcomes of this study for developing SLA management capability for Load Sharing Facility (LSF) [37],
- Comparing the performances of the policy based manageable RMS with those of traditional RMSs, and those of heuristic based RMSs such as exploiting evolutionary algorithms and genetic programming. For this purpose, comparing the RMSs via experiments that use the same workload traces
- Developing the mechanisms for SLA negotiations between administrative domains into HeteroSIM and extending POLICE model for SLA management,
- Improving statistical metrics charts so that they do not affect the performance of HeteroSim simulations.

REFERENCES

[1]. Dursun, T., Örencik, B., 2003,: "POLICE: A Novel Policy Framework", Lecture Notes in Computer Science, LNCS 2869, 819-827.

[2]. Dursun, T., 2005, "A Generic Policy Conflict Handling Model", Lecture Notes in Computer Science, LNCS 3733, 193-204

[3]. J. Banks, J.S. Carson, B.L. Nelson, and D. M. Nicol. 2001. "*Discrete-Event System Simulation*", Prentice Hall.

[4]. U.S. Department of Defense. 1997b. DoD modeling and simulation (M&S) glossary, DoD 5000.59-M, December.

[5]. Ernest H. Page, Roger Smith, Introduction To Military Training Simulation: A Guide For Discrete Event Simulationists, *Proceedings of the 1998 Winter Simulation Conference*

[6]. Java Management Extensions (JMX) Technology, http://docs.oracle.com/technetwork/java/javase/tech/javamanagement-140525.html

[7]. Mark Richards, R. Monson-Haefel, Java Message Service, O'Reilly Media, May 2000

[8]. IETF, Internet Engineering Task Force, Policy Working Group, available on WWW from http://www.ietf.org/html.charters/policy-charter.html

[9]. Rimon Barr, Zygmunt J. Haas, and Robbert van Renesse, 2004, "JiST: An Efficient Approach to Simulation Using Virtual Machines", Software—Practice and Experience, 00:1–7.

[10]. Sulistio A., Yeo C. S. and Rajkumar B., 2004, "A Taxonomy of Computer-based Simulations and its Mapping to Parallel and Distributed Systems Simulation Tools", Software Practice And Experience, 34:653–673 (DOI: 10.1002/spe.585)

[11]. Kilgore R. A., Healy K. J., and Kleindorfer G. B., 1998. "The Future of Java-based Simulation", In Winter Simulation Conference, pages 1707–1712, December.

[12]. M.C. Little. 2004. "JavaSim User Guide", public release 0.3. http://javasim.ncl.ac.uk/, June.

[13]. J. A. Miller, R. S. Nair, Z. Zhang, and H. Zhao. 1997. "JSIM: A JAVA-based Simulation and Animation Environment", In 30th Annual Simulation Symposium (ANSS'97), Atlanta, USA.

[14]. H.-Y. Tyan and C.-J. Hou. 2001. "J-Sim JavaSim: A Component based Compositional Network Simulation Environment", In Western Simulation Multiconference, January

[15]. Healy K. J., and Kilgore R. A., 1997. "Silk : A Java-based Process Simulation Language", In Winter Simulation Conference, 475–482, December.

[16]. F. Howell and R. McNab. 1998. "SimJava: A Discrete Event Simulation Package For Java With Applications In Computer Systems Modelling", First International Conference on Web-based Modelling and Simulation, San Diego, CA, Society for Computer Simulation, January

[17]. Kreutzer W., Hopkins J., Mierlo M. van. 1997. "SimJAVA - A Framework For Modeling Queueing Networks In Java", Proceedings of the Winter Simulation Conference

[18]. Krauter K., Buyya R. and Maheswaran M., 2002, "A Taxonomy and Survey of Grid Resource Management Systems for Distributed Computing", Software Practice and Experience, 32:135–164, Feb.

[19]. Sinha PK. *Distributed Operating Systems: Concepts and Design*. IEEE Press: New York, NY, 1997.

[20]. Buyya R, Murshed M., 2002, "GridSim: A Toolkit for the Modeling and Simulation of Distributed Resource Management and Scheduling for Grid Computing", Concurrency and Computation: Practice and Experience, 14(13–15):1175–1220.

[21]. JFreeChart library, http://www.jfree.org/jfreechart

[22]. T.Dursun, H.Dağ, "HeteroSim: Heterogeneous Simulation Framework", Communications and Networking Simulation Symposium, CNS 2009

[23]. Java Message Service, http://java.sun.com/products/jmso

[24]. JSR 274: The BeanShell Scripting Language

[25]. Strassner, J., Ellesson, E. and Moore, B. 1999. Eds., "Policy Framework Core Information Model", *Internet draft, draft-ietf-policy-core-schema-02.txt*, Feb. 1999.

[26]. Page, E.H., Moose, R.L. and Griffin, S.P. 1997. "Web-Based Simulation in SimJava Using Remote Method Invocation", Winter Simulation Conference, Atlanta, GA, 7-10 December.

[27]. Jacobs P.H.M., Niels A. Lang, Verbraeck A. 2002. "D-SOL; A Distributed Java Based Discrete Event Simulation Architecture", Proceedings of the Winter Simulation Conference, pages 793-800.

[28]. J. W. Herrmann, B. F. Conaghan, and L. H. Lecordier, Understanding The Impact Of Equipment And Process Changes With A Heterogeneous Semiconductor Manufacturing Simulation Environment, Proceedings of the 2000 Winter Simulation Conference

[29]. Yong Jae Kim, Tag Gon, Kim, A Heterogeneous Simulation Framework Based on the DEVS BUS and the High Level Architecture, Simulation Conference Proceedings, 1998, Winter, 13-16 Dec 1998

[30]. V. Thamizh Arasan, P. Vedagiri, Estimation Of Saturation Flow Of Heterogeneous Traffic Using Computer Simulation, Proceedings 20th European Conference on Modeling and Simulation, ECMS, 2006

[31]. Legrand A, Marchal L, Casanova H. 2003. "Scheduling Distributed Applications: The SimGrid Simulation Framework", Proceedings 3rd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid2003), Tokyo, Japan, 12–15 May.

[32]. Song HJ, Liu X, Jakobsen D, Bhagwan R, Zhang X, Taura K, Chien A. 2000. "The MicroGrid: A scientific tool for modeling computational Grids", IEEE Supercomputing (SC2000), Dallas, TX, 4–10 November

[33]. Ranganathan, K.; Foster, I.,"Decoupling computation and data scheduling in distributed data-intensive applications", 11th IEEE International Symposium on High Performance Distributed Computing, 2002. HPDC-11 2002, 23-26 July 2002 Page(s):352 –358

[34]. Dumitrescu, C.L., Foster,I,,"Usage Policy-based CPU Sharing in Virtual Organizations" 12/2004, In proceeding of Fifth IEEE/ACM International Workshop on Grid Computing, 2004.

[35]. Dumitrescu, C.L.; Foster, I., "GangSim: a simulator for grid scheduling studies", IEEE International Symposium on Cluster Computing and the Grid, 2005. CCGrid 2005, Volume 2, 9-12 May 2005 Page(s):1151 - 1158 Vol. 2

[36]. Kaner C, "Cem Kaner on Scenario Testing: The Power of "What If…" and Nine Ways to Fuel Your Imagination", Software Testing & Quality Engineering (STQE), October, 2003

[37]. Platform LSF Datasheet, http://www.platform.com/Products/platform-lsf