Cyberinfrastructure: A Case Study of IT Infrastructure for Next Generation Bioinformatics and Computational Biology

Haiqing Li and Yate-Ching Yuan
Bioinformatics Core, Department of Molecular Medicine, Beckman Research Institute,
City of Hope Medical Center, Duarte, CA 91010, USA

Abstract - *This presentation will share our experiences in establishing cost-effective translational bioinformatics platforms using an integrated cyberinfrastructure to support high-throughput data analysis, management, and integration in order to streamline analysis pipelines for predictive, preventive, personalized and participatory medicine. In this case study, we present the architecture of cyberinfrastructure and the challenges we face during design, deployment and management of cyberinfrastructure. We also show how new computational technologies, such as GPGPU and Cloud Computing, can help to speed up the bioinformatics analysis and data management. At the end, we discuss the future directions of IT infrastructure to support bioinformatics and computational biology.*

Keywords: Cyberinfrastructure, Bioinformatics, Computational Biology, Infrastructure, charge back, GPGPU, Cloud Computing

1 Introduction

IT infrastructure is the essential foundation for the bioinformatics and computational biology. The different types of biological computing have different computer utilization requirements[1]. The next generation bioinformatics and computational biology needs scalable and flexible IT resources to support the analysis of the next generation high throughput data, such as next generation sequencing, mass spectroscopy, HTS, high content screening technologies, etc. In spite of the broad spectrum of growing fields of OMICs technologies, the traditional IT data center has not yet evolved to provide adequate scalable cyberinfrastructure. In this study, we will demonstrate our efforts to integrate new paralllezation approaches of using shared meory ScaleMP and GPGPU servers to leverage the growing needs of IS&T support.

2 Cyberinfrastructure

Cyberinfrastructure is a project to design and deploy an IT infrastructure for next generation bioinformatics and computational biology in a national comprehensive cancer research center. The motivation of Cyberinfrastructure is to establish the IT infrastructure to support the integration bioinformatics of genomics, proteomics, cheminformatics, imaging, animal study, to enable our support to translational research.

2.1 Architecture

The architecture of Cyberinfrastructure (figure 1) includes two layers. The first layer is the system management and usage monitoring. The second layer is the IT infrastructure, which includes three components: the internal IT, scientific grid, and external cloud computing[2]. The internal IT is the core of cyberinfrastructure to support the on-demand high performance computing with several thousands of processors connected with Petabyte tiered of disk storage connected with infinity band network to processing TBs raw data generated from scientific high throughput instruments. We adopt several new technologies within the internal IT, which include the private cloud for application virtualization, high performance computing system using GPGPU and cluster technology, shared TBs memory computation, the tiered cluster storage system, high performance network system using Infiniband, and integrated lab information management system (LIMS). The second component of cyberinfrastructure is the scientific grid resources, which plays an important role in computational biology[2]. Cyberinfrastructure provides the interface to connect the scientific grid resources. The last component is external cloud computing, which steadily grows in demand for collaborative research by integrating large open source database such as EBI, NCBI, and UCSC Genome browser, etc. The challenges are to reconcile various of high throughput data analysis workflow in the area of genomics, proteomics, imaging, cheminformatics, small animal studies, and translational research and provide the computational and storage on demand Cloud computing for collaborative research. We have evaluated several Pay-As-You-Go external cloud services such as Amazon and Penguin-On-Demand for our current needs. We hope to establish a global platform that we can dynamically and cost effectively to support multidisciplinary translation research.

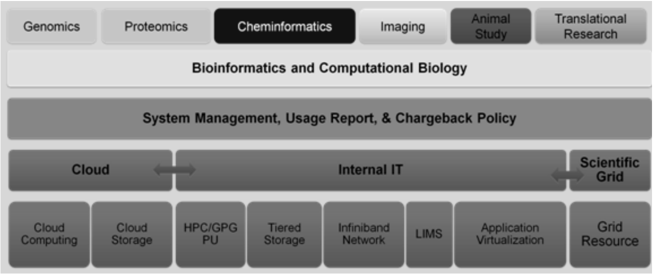

Figure 1. Architecture of Cyberinfrastructure

Figure 2 shows some highlights of the current deployment of Cyberinfrastructure at City of Hope. We adopt several new IS&T (information system and technology) technologies, which include high demand of CPU and memory intensity SMP system, virtual SMP system, high performance computing system using GPGPU and beowulf cluster technology, the tiered Isilon storage system, application virtualization using CITRIX and VMware, and integrated lab information management system (LIMS). It provides a powerful bioinformatics platform for the cancer research at City of Hope. Figure 3 shows the benchmark of the GPGPU cluster system with different GPU/CPU ratio configurations and how it speeds up Molecular Dynamics simulation for just few days running time that took months using super computational center resource.

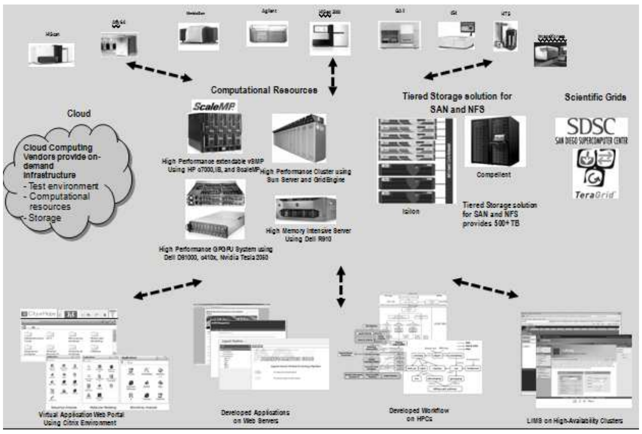
Computational Resources
ScaleMP
Cloud
Cloud Computing Vendors provide on-demand infrastructure
- Test environment
- Computational resources
- Storage
High Performance extendable vSMP Using HP c7000,IB, and ScaleMP
High Performance Cluster using Sun Server and GridEngine
High Memory Intensive Server Using Dell R910
High Performance GPGPU System using Dell C6100, c410x, Nvidia Tesla 2050
Tiered Storage solution for SAN and NFS
Isilon
Compellent
Tiered Storage solution for SAN and NFS provides 500+ TB
Scientific Grids
SDSC
SAN DIEGO SUPERCOMPUTER CENTER
TeraGrid
Virtual Application Web Portal Using Citrix Environment
Developed Applications on Web Servers
Developed Workflow on HPCs
LIMS on High-Availability Clusters

Figure 2. Current development of Cyberinfrastructure

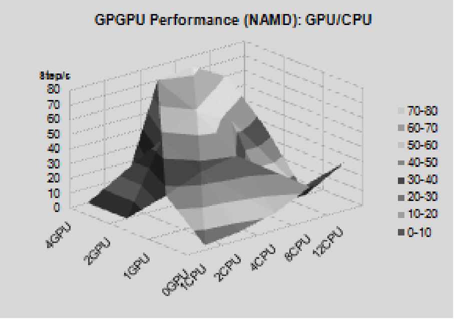

Figure 3. Benchmark of GPGPU

2.2 User access and Charge Back

As a core service, Cyberinfrastructure is open to all bioinformatics core subscribers. Subscribers can access cores Cyberinfrastructure resources, which include high performance servers, large scale tiered data storage, CITRIX web portal, and high performance workstations on campus. Tiered subscriber fee schema was established and fits all different type PIs and their research projects. Usage Metrics reports help PIs to cost effectively strategize their resources, and help administration team to strategic planning IS&T infrastructure needs.

3 Conclusions and Discussions

Strategic Planning for IS&T infrastructure is critical to support modern high throughput technologies such as next generation sequencing, high throughput screening, imaging, and high content screening. This presentation share our experiences in establishing cost-effective translational bioinformatics platforms using an integrated cyber-infrastructure to support high-throughput data analysis, management, and integration in order to streamline analysis pipelines for predictive, preventive, personalized and participatory medicine.

4 References

[1] Eric Jakobsson. *"Specifications for the Next-Generation Computational Biology Infrastructure," CTWatch Quarterly*, Volume 2, Number 3, August 2006. http://www.ctwatch.org/quarterly/articles/2006/08/specifications-for-the-next-generation-computational-biology-infrastructure/

[2] Brian Hayes, "*Cloud computing*". *Commun. ACM* 51, 7 (July 2008), 9-11

[3] El-Ghazali Talbi , Albert Y. Zomaya, *"Grid Computing for Bioinformatics and Computational Biology"*, Wiley, 2007