Abstract - Many existing token-based, distributed mutual exclusion algorithms can be generalized by a single algorithm. The number of messages required to provide mutual exclusion is dependent upon the logical topology imposed on the nodes and the policy used to forward requests. This paper extends the generalized algorithm to support prioritized requests and presents models that can be used to analyze the performance and verify the correctness of the prioritized algorithm. Both safety and liveness properties are verified. Model checking can also be used to analyze performance. Using the best topology, the generalized algorithm performs better than a centralized one when the star topology is used.

Keywords: distributed algorithm, model checking, mutual exclusion, real-time, token-based

1 Introduction

Many distributed mutual exclusion algorithms have been proposed [1, 4, 5, 6, 8, 12, 13, 16, 17, 18, 19]. These algorithms can be classified into two groups [14, 17]. The algorithms in the first group are called permission-based [1, 5, 6, 16]. A node enters its critical section only after receiving permission from a set of nodes. The algorithms in the second group are called token-based [4, 8, 12, 17, 18, 19]. The possession of a system-wide unique token gives a node the right to enter its critical section.

Lamport proposed one of the first distributed mutual exclusion algorithms [5]. Lamport’s algorithm is permission-based and requires \( 3 \times (N - 1) \) messages to provide mutual exclusion. Another permission-based algorithm, proposed by Ricart and Agrawala, reduces the number of required messages to \( 2 \times (N - 1) \) messages per critical section entry [15]. Maekawa proposed a permission-based algorithm in which the number of messages required is \( O(\sqrt{N}) \) [6]. Sanders generalized permission-based algorithms [16].

Ricart and Agrawala proposed a token-based algorithm which is essentially the same as Suzuki and Kasami’s algorithm [18]. Based on Suzuki and Kasami’s algorithm, Singhal proposed a heuristically-aided algorithm that uses state information to more accurately guess the location of the token [17]. The maximum number of messages required by these three algorithms is \( N \).

By imposing a tree-based logical structure on the nodes, another class of token-based algorithms has been obtained. In this class of algorithms, all of the nodes, except for the root node, are on a path to the root node (a sink node) in the logical structure. The logical structure determines the path along which a request message travels. There are two different types of logical structures: dynamic and static.

An algorithm, based on a dynamic logical structure, was proposed by Trehel and Naimi [19]. The basic notion underlying this algorithm is path reversal. Path reversal at each node is performed as a request from node \( x \) travels along the path from node \( x \) to the root node. As the request travels, node \( x \) becomes the new parent of each node on the path, except for node \( x \). Thus, node \( x \) becomes the new root node. A complete analysis of path reversal has been given by Ginat [3]. The average number of messages required per critical section is \( O(\log(N)) \).

If a static logical structure is used, the basic notion underlying the algorithm is what we call edge-reversal [11]. Edge reversal at each node is performed as the request from node \( x \) travels along the path from node \( x \) to the root node. At each node, the direction of each edge on the path is changed to point towards node \( x \); that is, to the neighboring node who sent the request on behalf of node \( x \). However, the shape of the logical structure never changes. Suprisingly, this small change results in algorithms which have a small fixed upper bound on the number of messages required per critical section, and the upper bound only depends on the logical structure. Algorithms based on edge-reversal were proposed by Neilsen and Mizuno [10] and Raymond [12]. Raymond’s algorithm assumes that the static logical structure is an unrooted tree. If the ra-
2 Overview

We assume that the system consists of \(N\) nodes, which are uniquely numbered from 0 to \(N-1\). At any given time, each node can have at most one outstanding request to enter its critical section. Physically, the nodes are fully connected by a reliable network, but logically, the nodes at each priority level are organized in a directed acyclic graph (dag). Nodes can be assigned different priorities. We start by describing the basic generalized algorithm for each priority level.

Two types of messages, called REQUEST and TOKEN, are exchanged among nodes. When a node requests to enter its critical section, it initiates a REQUEST message. A TOKEN message represents the token; when a node receives a TOKEN message, it may enter its critical section.

Each node maintains three simple variables: integer variables LAST and NEXT, and a boolean variable HOLDING or SINK. The logical directed acyclic graph (dag) structure indicates the path along which a REQUEST message travels and is imposed by the LAST variables in the nodes. When a node initiates or receives a REQUEST message, the node forwards the request to the neighboring node pointed at by its LAST variable (unless the node is a sink, in which case its LAST variable is -1).

The NEXT variable indicates the node which will be granted mutual exclusion after this node. If the node is currently the last node to be granted mutual exclusion, its NEXT variable is -1. Thus, by following the NEXT variables from the token holder to the node whose NEXT variable is -1, the implicit waiting queue of the system can be deduced. When a node leaves its critical section, it forwards the token to the node at the front of the waiting queue and also performs a dequeue operation. That is, it sends a TOKEN message to the node indicated by its NEXT variable (this is the node at the front of the queue) and clears the variable (this corresponds to the dequeue operation), unless NEXT is -1. If NEXT is -1, the node continues to hold the token if it is at the highest priority level by setting HOLDING to true, otherwise the token is returned to the highest priority level as described below.

Semantically, a sink node in the system is (1) the last node in the implicit waiting queue (i.e., its NEXT variable is -1), and (2) the last node on the path along which a request travels within a given priority level (i.e., its LAST variable is -1). When a sink node receives a REQUEST message, it enqueues the request into the implicit waiting queue and becomes a non-sink. The node initiating the request becomes the new sink since it is now the last node in the queue. Each edge in the path must change direction to point in the direction of the new sink. This is done by the nodes along the path in a distributed manner as follows:

- When a node initiates a new REQUEST message, it forwards the message to its neighboring node indicated by its LAST variable and sets its LAST variable to -1 to become a new sink. It remains a sink until it receives a subsequent request.
- When an intermediate (non-sink) node receives a REQUEST message from a neighboring node \(X\), it passes the message to the neighboring node indicated by its LAST variable. Then, the node sets its LAST variable to any node on the path traveled by the REQUEST message. Thus, if it
receives a subsequent request, it forwards the request in the direction of the new sink. In Trehel and Naimi’s algorithm, the LAST variable is set to the node that initiated the request; this is called path reversal. In Neilsen and Mizuno’s algorithm, the LAST variable is set to point to the neighboring node from which it received the REQUEST message; this is called edge reversal.

• When a sink node receives a REQUEST message from a node \(X\), it sets its NEXT variable to the identifier of the node initiating the request. This corresponds to an enqueue operation. The node also sets its LAST variable to any node on the path traveled by the REQUEST message. Note that if a sink node holds the token, but is not in its critical section (this state is indicated by a boolean variable HOLDING) when it receives a request, it immediately forwards the token to the node initiating the request.

Because of message delay, there may be several sink nodes in the system while requests are in transit. The system is initialized so that only one node at the highest priority level possesses the token. Initially, there is only one sink node at each priority level, and its LAST variable is initialized to -1. In all other nodes, LAST is set to point to the neighbor which is on a path to the node holding the token at the highest priority level (or the sink at all lower priority levels indicated by setting SINK to true).

```plaintext
const I = node identifier
var HOLDING : boolean;
LAST, NEXT : integer;

proc ProcessWork;
begin
if (not HOLDING) then
begin
send REQUEST(<I>) to LAST;
LAST := -1;
wait until a TOKEN message is received;
end;
HOLDING := false;

critical section (CS)
if (NEXT ≠ -1) then
begin
send TOKEN message to NEXT;
NEXT := -1;
end;
else HOLDING := true;
end;
```

```plaintext
proc ProcessRequest; (receive REQUEST(X_1, ···, X_k))
begin
if (LAST = -1) then
begin
if HOLDING then
begin
send TOKEN message to X_1;
HOLDING := false;
end;
else NEXT := X_1;
end;
else send REQUEST(X_1, X_2, ···, X_k, I) to LAST;
LAST := X_i for some 1 ≤ i ≤ k;
end;
```

**Figure 1. Single Priority Algorithm**

The complete generalized algorithm for a single priority level is shown in Figure 1. There are two procedures at each node: ProcessWork and ProcessRequest. Procedure ProcessWork is executed when a high priority node \(I\) requests for entry into its critical section, and procedure ProcessRequest is executed when a high priority node \(I\) receives a request from some other high priority node. In the algorithm, REQUEST messages are of form REQUEST \((X_1, X_2, ···, X_k)\) where \(X_1, X_2, ···, X_k\) denotes the path on which the request traveled and \(X_1\) denotes the node where the request originated. Each node executes procedures ProcessWork and ProcessRequest in local mutual exclusion. The only exception is that a node does not have to execute in mutual exclusion while waiting for a TOKEN message to arrive or while in its critical section.

The prioritized algorithm for low priority nodes is similar, except that the boolean variable HOLDING is replaced with SINK. Initially, a node in the highest priority level holds the token, and HOLDING for that node is set to true. Likewise, the root node at each lower priority level has its SINK variable set to true. When a request from a low priority node reaches the sink node, a PROXY REQUEST is sent up to a node at the highest priority level, called the proxy, and eventually the PROXY REQUEST reaches a node that holds the token or will receive the token in the future. Low priority requests are eventually enqueued in the token. Also, proxy requests are forwarded, at the highest priority level, just like a regular request, except that the edges are not reversed; i.e., higher priority nodes never request the token from lower priority nodes so there is no need to dynamically adjust the edges. When a node that is holding the token passes the token to a low priority node, a PROXY TOKEN message is used to pass the token, and the token is returned to the high priority node who sent the token using a PROXY RETURN message.
3 Verification Model

In this section we provide models that can be used to verify the correctness of the generalized algorithm with respect to guaranteed mutual exclusion, deadlock freedom, and starvation freedom using UPPAAL [2]. The model for a single priority level consists of two templates, ProcessWork and ProcessRequest, corresponding to proc ProcessWork and proc ProcessRequest, respectively, in the algorithm shown above in Figure 1. Channels are used to model the exchange of request messages and the token. Global arrays are used to model the state at each node using the arrays Holding, Sink, Next, and Last as defined above. Initially, one node will hold the token, so Holding[i] = true at that node. Also, the topology is defined by the initial values assigned to Last.

The first UPPAAL template, ProcessWork, is shown below in Figure 2. It models the work performed at each node. Initially, all nodes are in the Idle state. The template is parameterized using id to identify each node where id ∈ {0, 1, ..., N − 1}. At node 0, id = 0, etc. Also, Holding[id] is set to true at the node currently holding, but not using, the token. This node can enter its critical section immediately after setting Holding[id] to false to indicate that the token is in use.

All other nodes must send a request message to the node identified by LAST[id], and set LAST[id] = -1. Upon receipt of the token via a Token message, the requesting node may enter its critical section. A local clock, x, is used to prevent a node from remaining in its critical section forever. The model limits critical sections to be at most 10 time units through the location invariant x ≤ 10.

To process requests, the ProcessRequest template is used as shown in Figure 3. When a request message is received, at node id from node p, using Request[p][id][s]? , the source node requesting to enter its critical section is node s.

If the node receiving the request is a sink node, Last[id] == -1, the request can be satisfied immediately if the node receiving the request is holding, but not using, the token; that is, if Holding[id] == true. In which case, the Token message can be sent immediately. On the other hand, if the token is currently in use, then the request is simply enqueued by setting Next[id] = t which is a local meta variable assigned to a when the request is received.

If the node receiving the request is not a sink node, Last[id] ≥ 0, then the request is forwarded on to the node indicated by Last[id].

In all cases, Last[id] is set to the identifier of the neighboring node which sent the request; that is, edge reversal is used.

The generalized algorithm is slightly more complex because each node can set its Last[id] value to be any node visited from the requesting node to the sink. Consequently, a list or queue of visited node numbers must be carried with the Request message. This is modeled with a set of global queues that get updated as the Request travels. One queue is assigned to each node and used to enqueue the nodes on the path from the given requesting node to a sink. Since each node can have at most one outstanding request to enter its critical section, the queues can be initialized when a request is first initiated, and each requesting node can be associated with a single queue. The queues can be initialized to impose any logical topology. For example, to impose the star topology, with node 0 as the root, each node could execute the following code:

```c
void Initialize(int id)
{
    if (id == 0)
    {
        Holding[id] = true;
        Last[id] = -1;
    }
    else
    {
        Holding[id] = false;
        Last[id] = 0;
    }
    Next[id] = -1;
}
```

Figure 4. Initialization Routine
The generalized ProcessWork template is shown in Figure 5. By making Init a committed state, all nodes will enter the Ready state before any nodes start requesting. The only other change required is to add a function, initRequest(id), that is used to initialize the queue passed with the Request message to contain a single element <id>.

The generalized ProcessRequest template is shown in Figure 6. As the Request message travels from a requesting node to a sink node, the identifier of each node receiving the request must be enqueued in the request message. Since UPPAAL messages have zero capacity, this is modeled using a set of global queues and a function call enQueue(t, id) to enqueue id on the queue for the requesting node t. Also, each node on the path sets Last[id] to be some element in the queue. Recall that if it is set to the node at the front of the queue (the requesting source node), then this is just path reversal; at the other extreme, setting Last[id] to the neighboring node sending the request, namely node p, then this results in edge reversal.

The function, someQueue(), relies on a random number generator, modeled by the RandomValue template shown below, to randomly select a random element from the queue carried with the request.

Finally, to support different priority levels, we can add the notion of a proxy node at the highest priority level, and use algorithms similar to the above to request the token.

When no requests are pending at the highest priority level, and a request is pending at some lower priority level, the token is passed to allow a node at a lower priority level to obtain the token. The first pending request at each lower priority level is enqueued in priority order in the token. Of course, this may lead to starvation; if there is always a pending request at the highest priority level, then lower priority requests will not be satisfied.

For the prioritized case, four templates are used, nodes at the highest priority level use ProcessWork and ProcessRequest as shown in Figures 8 and 10, and the lower priority nodes use ProcessWorkLow and ProcessRequestLow as shown in Figures 9 and 11. Due to space constraints, we only include the prioritized models for edge reversal (extending the templates shown in Figures 2 and 3), but the generalized case is similar.
4 Performance Analysis

To verify the correctness of the algorithm we have developed analytical proofs of correctness. We have also verified both safety and liveness properties using UPPAAL [2]. To verify that the algorithm satisfies mutual exclusion, the property: \( A[] (\text{not (ProcessWork}(i).\text{CS and ProcessWork}(j).\text{CS})) \) must be verified for different values of \( i \) and \( j \); that is, only one process can be in its critical section (state \( \text{CS} \) at any time. By symmetry, only a few combinations need to be checked.

To verify liveness properties, we first limit the amount of time each process can remain in its critical section; otherwise, one node could remain in its critical section forever. The \( \text{CS} \) state has a location invariant of \( x \leq 10 \) for a real-valued clock \( x \) which is initialized to 0 upon entry to the critical section state. The choice of ten time units is arbitrary. To verify that a node that wants to enter its critical section can enter, we verify the property \( \text{ProcessWork}(1).\text{Requesting} \rightarrow \text{ProcessWork}(1).\text{CS} \); that is, requesting the critical section “leads to” entry. Formally, a process, at the highest priority level, in the \text{Requesting} state eventually reaches the \text{CS} state. Starvation freedom is not satisfied by lower priority processes.

The performance of the algorithm depends on the topology of the logical structure. The best topology is the star topology, with one node in the center and all other nodes as leaf nodes. For the analysis, we define the diameter \( D \) of a logical structure to be the length of the longest path in the structure. As the logical structure evolves, the value of \( D \) may also change.

First, we consider a single priority level. The upper bound is equal to \((D + 1)\) messages per critical section entry. This occurs when a requesting node and a sink node are at opposite ends of the longest path: \( D \) messages for the request to travel to the sink node and one message for the token to be sent back to the requesting node. Thus, using the straight line topology, the upper bound is \( N \), where \( N \) is the number of nodes in the system. Using the best topology, the upper bound is 3, since the diameter of a star is 2. This is the same as the performance of a centralized mutual exclusion algorithm. To verify this upper bound, counters can be used, as shown in Figures 4 and 5. Once a request is satisfied, the counter is set back to zero. Then, the model can be verified to see if there exists a path on which the number of messages required for a request from a given node to be satisfied is possible; e.g., \( E<>((\text{Count}[1]==3)) \). For the star topology with edge reversal, each leaf node can require up to 3 messages per critical section; thus, the property \( E<>((\text{Count}[1]==3)) \) is satisfied if node 1 is a leaf node, and the property \( E<>((\text{Count}[1]==4)) \) is never satisfied. Likewise, the central node requires at most two messages per critical section. Not surprisingly, for the worst topology – a straight line – the worst-case for path reversal and the generalized algorithm is \( N \) messages, \( N - 1 \) \text{Request} messages and a \text{Token} message to forward the token. However, this is not true for edge reversal because the underlying topology doesn’t change. Consequently, the number of \text{Request} messages required is the maximum distance from the requesting node to the end of the line. Roughly, \( N/2 \) for the node in the middle of the line.

If there is more than one priority level, then the number of messages required at the highest priority level is only dependent on the number of nodes in the highest level with the same analysis as above. For lower priority messages, after the request reaches a sink node, by traversing at most the diameter of the topology used for the given priority level, a proxy request may need to be forwarded up to the highest priority level, and to a sink node – again, at most the diameter of the nodes at the highest priority level. The
token is passed to the lower priority node with a single message, and returned with a single message. Thus, the worst-case message complexity is $D_H + D_L + 2$ for low priority requests, where $D_H$ is the diameter of high priority nodes, and $D_L$ is the diameter of low priority nodes. If the low priority level consists of a single node, then $D_L = 0$. For the star topology, and the proxy node set as the root, at most 3 messages are required per critical section.

5 Summary

This paper presented a prioritized token-based algorithm for distributed mutual exclusion. The algorithm imposes very little storage overhead on each node and in each message. Furthermore, the algorithm generalizes several existing token-based algorithms and shows how they can be extended for prioritized systems. Using the best topology and edge reversal, the algorithm attains comparable performance to a centralized mutual exclusion algorithm; that is, three messages per critical section entry. In the average case, the algorithm attains the best performance of any known algorithm.

References


