A De-centralized Scheduling and Load Balancing Algorithm for Heterogeneous Grid Environments

Manish Arora and Sajal K. Das
Dept. of Computer Science & Engineering
The University of Texas at Arlington
Arlington, TX 76019-0015
{arora, das}@cse.utexas.edu

Rupak Biswas
NASA Advanced Supercomputing Division
NASA Ames Research Center
Moffett Field, CA 94035-1000
rbiswas@nas.nasa.gov

Abstract

In the past two decades, numerous scheduling and load balancing techniques have been proposed for locally distributed multiprocessor systems. However, they all suffer from significant deficiencies when extended to a Grid environment: some use a centralized approach that renders the algorithm unscaleable, while others assume the overhead involved in searching for appropriate resources to be negligible. Furthermore, classical scheduling algorithms do not consider a Grid node to be N-resource rich and merely work towards maximizing the utilization of one of the resources. In this paper, we propose a new scheduling and load balancing algorithm for a generalized Grid model of N-resource nodes that not only takes into account the node and network heterogeneity, but also considers the overhead involved in coordinating among the nodes. Our algorithm is de-centralized, scalable, and overlaps the node coordination time with that of the actual processing of ready jobs, thus saving valuable clock cycles needed for making decisions. The proposed algorithm is studied by conducting simulations using the Message Passing Interface (MPI) paradigm.

1. Introduction

Computational Grids [1, 6] are typically a conglomeration of various resources with different owners, but make it possible for users to develop complex applications that access remote sites. Each of these sites (or nodes) could be a uni-processor machine, a symmetric multiprocessor cluster, a distributed memory multiprocessor system, or a massively parallel supercomputer. Each node consists of a number of heterogeneous resources; the heterogeneity being in the type and capability of each of its N-resources (e.g., number of processors, CPU speed, amount of memory, and so on).
formation Power Grid (IPG) [2, 7]. The overhead involved in capturing the resource utilization status of a given node's neighbors before making a scheduling decision can be a major issue negating the advantages of job migration. Our algorithm therefore overlaps the time spent looking for appropriate nodes with the actual execution of the ready jobs, thus saving precious clock cycles. Also, since each Grid node (whether a single uni-processor machine or a multiprocessor system) can have its own independent scheduling algorithm, our technique does not overrule the local schedulers' job assignment policy. The class of problems we address is where jobs are computation-intensive and can be divided into totally independent sub-tasks with no communication between them.

We have conducted extensive experiments using the Message Passing Interface (MPI) paradigm and by simulating the job arrival rate. We compared the quality of load balance with the ideal case (where no overheads are involved) and found that our algorithm performs remarkably well in an heterogeneous Grid environment and gives encouraging results. The remainder of this paper is organized as follows: Section 2 describes our algorithm and presents pseudo codes of the key procedures; Section 3 discusses the experimental setup that we used to test and substantiate our claims, and interprets the results; and Section 4 concludes the paper.

2. Scheduling and load balancing

Two important aspects of any wide area network scheduler are its transfer [4, 15] and location [8, 9] policies. The transfer policy decides if there is a need to initiate load balancing across the system, and is typically threshold based. Using workload information, it determines when a node becomes eligible to act as a sender (transfer a job to another node) or as a receiver (retrieve a job from another node). The location policy selects a partner node for a job transfer transaction. In other words, it locates complementary nodes to/from which a node can send/receive workload to improve overall system performance.

Location policies can be broadly classified as sender-initiated [4, 21], receiver-initiated [4, 12], or symmetrically-initiated [5, 15, 19]. Sender-initiated policies are those where heavily-loaded nodes search for lightly-loaded nodes while receiver-initiated policies are those where lightly-loaded nodes search for suitable senders. Symmetrically-initiated policies combine the advantages of these two by requiring both senders and receivers to look for appropriate partners.

Load balancing policies can also be classified on the basis of how up-to-date each node's knowledge is about the state of the system. Dynamic [16, 17] policies make decisions based on the current system state and can rapidly adapt to workload fluctuations. On the other hand, policies that use static information and are not amenable to changes in the workload are known as static [3] policies. However, dynamic policies incur the overhead of communicating among the system nodes to keep them informed about the state of the system.

In this section, we describe our scheduling and load balancing algorithm for N-resource Grid environments. It is dynamic, sender-initiated, and completely de-centralized. The last feature makes it extremely scalable for Grid environments. A remarkable property of our algorithm is that it uses a smart search strategy for finding partner nodes. It also overlaps this decision making process of a node with the actual execution of ready jobs, thereby saving precious processor cycles.

2.1. Preliminaries

Before discussing the algorithm, let us introduce the concepts of Internal and External queues, which we assume exist in each Grid node. The Internal Queue of a node consists of the ready jobs which would be executed by this particular node only. Let τ be the time the tasks were last mapped, a(tj) be the arrival time of task tj, and e(tj) be the time tj starts executing. Then, the jobs in the Internal Queue are those that have been mapped and scheduled to this node, and are either being executed (Eq. 1) or are ready to be executed (Eq. 2); they would never be delegated to any other node:

\[ \{ t_j | a(t_j) \leq \tau, e(t_j) \leq \tau \} \]  

\[ \{ t_j | a(t_j) \leq \tau, e(t_j) > \tau \} \]  

Instead, the External Queue of a node consists of jobs which have been initially submitted to this node by a user, but are yet to be mapped and scheduled for execution (Eq. 3):

\[ \{ t_j | a(t_j) > \tau, e(t_j) > \tau \} \]  

Let us now enumerate the key notations we will be using throughout the paper to explain our algorithm:

- \( P_i \): Grid node i
- \( P^j_i \): The j-th resource of \( P_i \)
- \( J_k \): Job k
- \( J^j_k \): Ideal requirement for the j-th resource by \( J_k \)
- \( Neigh(P_i) \): Immediate neighbors of \( P_i \)
- \( Compl(t) \): Time needed by \( P_i \) to empty its Internal Queue assuming no more jobs are assigned to it after time t
- \( Comm^j_i \): Communication latency between \( P_i \) and \( P_j \)
- \( ExQ \): Number of jobs in the External Queue of \( P_i \)

We assume that each Grid node has knowledge about the communication latency between itself and all of its neighbors; i.e., each node \( P_i \) knows \( Comm^j_i \), \( \forall j \in Neigh(P_i) \). Not only does this make the algorithm highly scalable, it
also allows the network to conveniently accommodate any changes in its topology.

We also postulate that each incoming job knows its requirements for each of the resources available at a node. In order to generalize this concept, we define $N$-resource jobs and $N$-resource nodes/servers. Each job $J_k$ looks for a node $P_i$ with resources $P_i^0, P_i^1, ..., P_i^{N-1}$, such that it meets its requirement for each resource type, $J_k^0, J_k^1, ..., J_k^{N-1}$. The algorithm described below would be executing on every node of the Grid.

2.2. Proposed algorithm

Whenever a job is submitted by a user to a node $P_i$, procedure Main (Fig. 1) invokes procedure NeedForTriggering (Fig. 2) to make a decision whether the job need to be migrated. If the job ought to be migrated to another node, a request is sent to all nodes $j \in \text{Neigh}(P_i)$, provided $2 \times \text{Comm}_j \leq \text{Time}_{IQ}$. This condition implies that the status request to the neighboring nodes and their responses should be received before the Internal Queue is emptied (denoted by $\text{Time}_{IQ}$). This strategy avoids any wastage of the node’s resources; the inequality overlaps the task of looking for appropriate nodes with the actual processing of the Internal Queue, thus hiding the overhead.

![Figure 1. Procedure Main](image1)

Refer to Figures 2 and 3 for the triggering policy we have incorporated into our algorithm. It is based on the simple heuristic that greater the load at a node, the less inclined would it be to accept future loads. Within a time window of $\text{Comp}_i(\tau)$, triggering is initiated if the traffic burst is more than admissible; however, higher the resource usage, the smaller is the traffic burst that a node will accommodate (Fig. 3).

![Figure 2. Procedure NeedForTriggering](image2)

A node, having received a request to send the status of its resources, packs the information about their current utilization and sends it back to the requesting node along the route the request came (Fig. 4). This route is also piggybacked to the node which needs to migrate load. Besides replying to requests, a node also recursively pings its neighbors for their resource status if its database says that the total round-trip latency between the sender and its neighbor would be less than $\text{Time}_{IQ}$. This allows the time required to look for additional resources be hidden under processing.

![Figure 3. Value of $\delta$ when Job Queue is x% full](image3)

![Figure 4. Procedure Request](image4)
Figure 5 shows the pseudo code for procedure Receive. The sender waits for time \( T_{\text{SEQ}} \) to get replies from the nodes that have been queried for the status of their resources.

```plaintext
Procedure Receive(T\( \text{SEQ} \)):
    While (Time \( \leq \) T\( \text{SEQ} \))
        MPI.Recvive (S)
    End
    R \( \leftarrow \) Number of replies
    Return R
End Receive
```

Figure 5. Procedure Receive

Figure 6 shows our procedure to schedule the jobs soon after \( T_{\text{SEQ}} \) elapses. Without loss of generality, we can assume that \( 0.0 \leq P^j_i, J^j_k \leq 1.0, 0 \leq j \leq N - 1 \). Let \( M^k_i \) be a match variable which defines the number of resources in node \( P_i \) that fulfill the requirements of job \( J_k \). If \( \text{bool}(J^j_k \leq P^j_i) = 1 \) and \( \text{bool}(J^j_k > P^j_i) = 0 \), then we can formally define \( M^k_i \) as

\[
M^k_i = \sum_{j=0}^{N-1} \text{bool}(J^j_k \leq P^j_i) \tag{4}
\]

Clearly, \( 0 \leq M^k_i \leq N \). Now, let us define matrices \( T \) and \( C \), and vector \( V \), as described in steps 1, 2, and 3 of procedure Balance (Fig. 6). Intuitively, the \( u \)-th row and \( k \)-th column of \( T \) gives the number of resources in node \( P_i \) that meets the requirements of job \( J_k \); the \( k \)-th entry of \( V \) gives the number of nodes which satisfy the minimum requirements of \( J_k \); and element \( C_{u,j} \) denotes that there is a common node that satisfies the requirements of both \( J_u \) and \( J_j \), and that there might be a conflict while scheduling them.

Another possible scenario is when the set of nodes that satisfy the requirements of \( J_u \) is a subset of the set which satisfies the requirements of \( J_j \); in such cases, giving preference to \( J_j \) might leave \( J_u \) with no viable option. To avoid such cases, our algorithm first schedules jobs that have the fewest choices. \( T(u, \min(j)) \) in step 4.1 of Fig. 6 corresponds to the job \( J_{\min} \) that has the minimum number of nodes it can be mapped to. The variable \( z \) indicates the node \( P_z \) to which \( J_{\min} \) can be delegated. Step 4.2 checks matrix \( C \) and, in case there is another job that can be mapped to \( P_z \), chooses a different \( z \) for \( J_{\min} \), if possible. Finally, \( J_{\min} \) is mapped and scheduled to \( P_z \). This mechanism continues until all jobs have been scheduled or until no more can be mapped because of the lack of resources.

3. Experimental study

Here we describe the metrics used to gauge the performance of our scheduling and load balancing algorithm, the setup we had for our experiments, the simulation results that were obtained, and the conclusions we can draw from them.

3.1. Performance metrics

We analyze the performance of our algorithm using a parameter called Normalized Performance, \( \eta \) (defined in Eq. 5). Basically, \( \eta \) is the effectiveness of the load balancing strategy. It is a comprehensive metric as it considers both the initial load balance as well as the load balancing overheads:

\[
\eta = \frac{T_{\text{no}} - T_{\text{lb}}}{T_{\text{no}} - T_{\text{lb}}} \tag{5}
\]

Here, \( T_{\text{no}} \) is the time to completely process all the jobs on a uniprocessor machine; \( T_{\text{lb}} \) is the time required by one processor divided by the total number of processors, thus providing the runtime with ideal load balancing; and \( T_{\text{lb}} \) is the time needed by our algorithm to balance the load and execute all the jobs. Clearly,

\[
\text{if } T_{\text{lb}} \rightarrow T_{\text{lb}}, \text{ then } \eta \rightarrow 1 \tag{6}
\]

\[
\text{if } T_{\text{lb}} \rightarrow T_{\text{no}}, \text{ then } \eta \rightarrow 0 \tag{7}
\]

These two conditions imply that higher the value of \( \eta \), the better is the load balancing; the ideal case being \( \eta = 1 \).

3.2. Experimental setup

The experimental results reported in this paper were obtained by using an MPI implementation of our proposed algorithm. It is worth mentioning here that the various parameters of our algorithm were varied following a Poisson distribution. Their respective mean values are given in Table 1.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Mean</th>
<th>Simulated by</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processing Power</td>
<td>2-16</td>
<td>50 floatingpoint multiplications per unit</td>
</tr>
<tr>
<td>Requirements</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Memory Requirements</td>
<td>2-16</td>
<td>1KB of memory allocated &amp; freed per unit</td>
</tr>
<tr>
<td>I/O Requirements</td>
<td>2-16</td>
<td>1KB of data written to disk per unit</td>
</tr>
<tr>
<td>Network Latency</td>
<td>5-11</td>
<td>sleep(3) per unit</td>
</tr>
<tr>
<td>Node Degree</td>
<td>5</td>
<td>number of neighboring nodes</td>
</tr>
</tbody>
</table>
Procedure Balance(S, R)
1. Using S, define matrix T of dimensions ExQi × R where T_{u,k} ← M_{k}^u
2. Define vector V of dimension ExQi where V_{k} ← ∑_{u=1}^{R} bool(T_{u,k} = N), 1 ≤ k ≤ ExQi,
3. Define matrix C of dimensions ExQi × ExQi where C_{i,k} ← C_{u,k} + 1, if T_{k,j} = T_{i,j} = N; 0, otherwise; 1 ≤ i, k ≤ ExQi, 1 ≤ j ≤ R
4. Repeat until (no more jobs can be mapped)
   4.1 z ← u | T_{i,m(min(V))} = N, 1 ≤ u ≤ R, 1 ≤ j ≤ ExQi
   4.2 If (C_{min(V),k}) = 1, 1 ≤ k ≤ ExQi)
       Choose another z, if possible
   4.3 Assign j_{min(V)} to node P_{i}, 1 ≤ j ≤ ExQi
   4.4 Remove row min(V), 1 ≤ j ≤ ExQi, and column z from T
End Balance

Experiments were conducted for three different values of Max (15, 20, and 25) (see Fig. 3), and repeated for 1-, 2-, and 3-resource nodes. The following three inequalities give the relationships between the m_{i}'s, where each m_{i} refers to the slope of the line joining the co-ordinates (0, Max) and (100, 0) (Fig. 3):
\begin{align*}
    m_1, m_2, m_3 &< 0 \quad (8) \\
    m_1 < m_2 < m_3 \quad (9) \\
    |m_1| > |m_2| > |m_3| \quad (10)
\end{align*}

3.3. Simulation results

We have conducted extensive experiments to evaluate the performance of our algorithm and help us substantiate our approach. Figures 7 through 9 illustrate the results obtained from the study.

To verify that our algorithm works well for completely heterogeneous systems, we divided the experiments into three groups. The first set of experiments was run on systems where heterogeneity was in the capabilities of the N-resources of a node; thus, the communication latency between all neighboring nodes was constant. The second set involves keeping the node capability constant and varying only the communication latency between the nodes. Finally, the third set of experiments combined the above two approaches, thereby exposing a totally heterogeneous setup to various load conditions (that were varied by changing the job arrival rate and the load associated with each job). Each set of experiments was repeated for 1-, 2-, and 3-resource nodes. The objective was to evaluate the algorithm thoroughly by taking various scenarios of heterogeneity into consideration.

Results for the first set (where only the capabilities of the N-resources of a node are varied while keeping all other factors unchanged) are summarized by the graphs in Fig. 7. The horizontal axis represents the Mean Node Capacity of the network which can be defined as the mean value used for the capacity of each of the resources in a node (all resources having the same mean). Increasing the resource capability of the nodes without changing the job resource requirements effectively reduces the granularity of the latter. As depicted, any increase in node capability increases \( \eta \). However, as the threshold slopes (m_{i}'s) become steeper, \( \eta \) decreases. This is because the frequency of triggering the load balancing algorithm is reduced.

In the second set of experiments, the Mean Node Capacity was held constant while varying the communication latency. The results presented in Fig. 8 show that \( \eta \) decreases with increasing communication cost. As in the previous set, the algorithm performs best when the absolute value of the threshold slope is the smallest (m_{3} in this case).

For the final set of experiments, we vary the input load for a setup which has a heterogeneous mix of resource capability and communication latency. This was repeated for 1-, 2-, and 3-resource job specification for a 3-resource node. Figure 9 shows that the execution time decreases as we get more specific about job requirement.

4. Conclusions

In this paper, we presented a highly de-centralized, distributed, and scalable algorithm for scheduling tasks and load balancing resources in heterogeneous Grid environments. Our algorithm takes into consideration the overheads of coordination and communication between the Grid nodes which were assumed to be N-resource servers that varied in their respective capacities across resources. The goal was to assign each node a job which would utilize its resources in the best possible manner, thus providing an effective scheduling and resource management strategy. We introduced a new load balance triggering policy based on the endurance of a node reflected by its current queue length. Also, our algorithm overlaps the time needed for
Figure 7. Normalized Performance ($\eta$) vs. Mean Node Capacity for 1-, 2-, and 3-resource nodes (top to bottom)

Figure 8. Normalized Performance ($\eta$) vs. Mean Communication Time for 1-, 2-, and 3-resource nodes (top to bottom)
various communication overheads with that of executing the jobs already committed to the nodes, making the effective time for overheads virtually zero. The algorithm has been discussed in detail with pseudo codes being provided for all the major modules of the algorithm.

To substantiate our claims, a comprehensive experimental study was conducted using the Message Passing Interface (MPI) paradigm. Heterogeneity in resource capabilities and communication latency was maintained while repeating the set of experiments for 1-, 2-, and 3-resource jobs and nodes. The Normalized Performance parameter was 0.79 for 3-resource nodes and as high as 0.85 for 1-resource nodes. These excellent performance levels could be attained only by overlapping the various overheads with the actual execution of the jobs.

Acknowledgements

The work of the first two authors was partially supported by NASA Ames Research Center under Cooperative Agreement Number NCC 2-5395.

References