Real-Time Divisible Load Scheduling for Cluster Computing

Xuan Lin, Ying Lu, Jitender Deogun, Steve Goddard
Department of Computer Science and Engineering
University of Nebraska - Lincoln
Lincoln, NE 68588
{lxuan, ylu, deogun, goddard}@cse.unl.edu

Abstract

Cluster computing has emerged as a new paradigm for solving large-scale problems. To enhance QoS and provide performance guarantees in cluster computing environments, various real-time scheduling algorithms and workload models have been investigated. Computational loads that can be arbitrarily divided into independent pieces represent many real-world applications. Divisible load theory (DLT) provides insight into distribution strategies for such computations. However, the problem of providing performance guarantees to divisible load applications has not yet been systematically studied. This paper investigates such algorithms for a cluster environment. Design parameters that affect the performance of these algorithms and scenarios when the choice of these parameters have significant effects are studied. A novel algorithmic approach integrating DLT and EDF (earliest deadline first) scheduling is proposed. For comparison, we also propose a heuristic algorithm. Intensive experimental results show that the application of DLT to real-time cluster-based scheduling leads to significantly better scheduling approaches.

1 Introduction

The dawn of the information age has changed how we solve important problems. Emerging computing and data intensive applications cannot be solved by a single stand-alone machine. This has led to the emergence of cluster computing—which harnesses the power of hundreds and thousands of machines—as a new paradigm for computing. However, as the size of a cluster increases, so does the complexity of resource management and maintenance. Automated performance control and resource management is crucial to achieve continued evolution of cluster computing. Current cluster scheduling practice is similar in sophistication to early supercomputer batch scheduling algorithms, and no consideration is given to desired quality-of-service (QoS) attributes. To fully avail the power of computational clusters, new scheduling theory that provides high performance, QoS assurance, and streamlined management of the cluster resources needs to be developed.

The challenge, however, in developing real-time scheduling theory for cluster computing is to support various types of cluster applications.Broadly speaking, computational loads submitted to a cluster are structured in two primary ways: indivisible and divisible. An indivisible load is essentially a sequential job and thus must be assigned to a single processor. The divisible loads are comprised of tasks that can be executed in parallel and can be further divided into two categories: modularly divisible and arbitrarily divisible loads. Modularly divisible loads are divided a priori into a certain number of subtasks and are often described by a task (or processing) graph. Arbitrarily divisible loads can be partitioned into an arbitrarily large number of load fractions. Examples of arbitrarily divisible loads can be easily found in high energy and particle physics. For example, the CMS (Compact Muon Solenoid) [10] and ATLAS (A toroidal LHC Apparatus) [6] projects, which are associated with the Large Hadron Collider (LHC) at CERN (European Laboratory for Particle Physics), execute cluster-based applications with arbitrarily divisible loads. Usually all elements in such computational loads demand an identical type of processing, and relative to the huge total computation, the processing on each individual element is infinitesimally small.

The problem of providing QoS or real-time guarantees for sequential and modularly divisible jobs in distributed systems has been studied extensively. Similarly, significant progress has been made in divisible load theory (DLT)[28]. However, despite the increasing importance of arbitrarily divisible applications [24], to the best of our knowledge, the real-time scheduling of arbitrarily divisible loads has not been systematically investigated.

Scheduling of arbitrarily divisible loads represents a problem of great significance for cluster-based research computing facilities such as the U.S. CMS Tier-2 sites [26]. (The CMS project will not be fully operational until 2007.) One of the management goals at the University of Nebraska-Lincoln (UNL) Research Computing Facility (RCF) is to provide a multi-tiered QoS scheduling framework in which applications “pay” according to the response time requested for each job [26]. Existing real-time cluster scheduling algorithms assume the existence of a task graph for all applications, which are not appropriate for arbitrarily divisible loads. To better manage these high-end clusters and control their performance, we need new real-time scheduling algorithms for arbitrarily divisible applications.

Four contributions are made in this paper. First, DLT is extended to compute the minimum number of processors required to meet an application’s deadline. Second, based on this, a novel algorithmic approach integrating DLT and EDF (earliest deadline first) scheduling is proposed. For comparison, we also propose a heuristic algorithm. Third, important design parameters are identified that affect the performance of real-time divisible-
load scheduling algorithms. Fourth, we systematically investigated the effects of these design parameters on a set of real-time scheduling algorithms, and show that the application of DLT to real-time, cluster-based scheduling leads to significantly better scheduling approaches.

The remainder of this paper is organized as follows. Related work is presented in Section 2. Section 3 describes both task and system models. In Section 4, real-time scheduling algorithms investigated in this paper are discussed. We evaluate the performance of algorithms in Section 5 and conclude the paper in Section 6.

2 Related Work

Development of commodity-based clusters and Grid computing has recently gained considerable momentum. By linking a large number of computers together, a cluster provides a cost-effective facility for solving complex problems. In a large-scale Grid, the resource management system (RMS), which provides real-time guarantees or QoS, is central to its operation.

Research has been carried out in utility-driven cluster computing [29, 25] to improve the value of utility delivered to users. Proposed cluster RMSs [9, 3] have addressed the scheduling of both sequential and parallel loads. The goal of those schemes is similar to ours: to harness the power of resources based on user objectives.

The scheduling models investigated for distributed or multiprocessor systems most often (e.g., [23, 22, 14, 1, 20, 13, 5]) assume periodic or aperiodic sequential jobs that must be allocated to a single resource and executed by their deadlines. With the evolution of cluster computing, researchers have begun to investigate real-time scheduling of parallel applications on a cluster [31, 21, 12, 2, 4]. However, most of these studies assume the existence of some form of task graph to describe communication and precedence relations between computational units called subtasks (i.e., nodes in the task graph).

The most closely related work to our problem is scheduling algorithms for “scalable real-time tasks” running in a multiprocessor system presented in [16]. In that paper, like divisible loads, it is assumed that a task can be executed on more than one processor and as more processors are allocated to it, its pure computation time decreases monotonically. The paper notes that the decision on the number of processors allocated to tasks is an important factor in the design of parallel scheduling algorithms. However, the simulations described in the paper are limited and are favorably biased towards their proposed schemes. Therefore, their conclusions on comparing their proposed MWF (Maximum Workload derivative First) schemes with the EDF and FIXED algorithms [19, 7] hold true only in certain scenarios.

Our work differs significantly from previous work in real-time as well as cluster computing in both the task model assumed and in the comprehensiveness of our study. In this paper unlike the previous study [16], we do not assume task execution times are known a priori. Instead, we apply DLT to guide task partitioning, to derive its execution time function, and to compute the minimum number of processors required to meet its deadline.

DLT provides an in-depth study of distribution strategies for arbitrarily divisible loads [8, 24, 28]. The goal of DLT is to exploit parallelism in computational data so that the workload can be partitioned and assigned to several processors such that execution completes in the shortest possible time [8]. DLT has been previously applied to and implemented in Grid computing [30, 15, 27]. Complimentary to that work, our paper applies DLT in the design of real-time scheduling algorithms for cluster computing; specifically, DLT is applied in the partitioning of applications, such as CMS [10] and ATLAS [6], that execute on a large cluster.

3 Task and System Models

In this section we describe our task and system models briefly, and state assumptions related to these models.

Task Model. We assume a real-time aperiodic task model in which each aperiodic task \( T_i \) consists of a single invocation specified by the tuple \( (A_i, \sigma_i, D_i) \), where \( A_i \geq 0 \) is the arrival time of the task, \( \sigma_i > 0 \) is the total data size of the task, and \( D_i > 0 \) is its relative deadline. The absolute deadline of the task is given by \( A_i + D_i \). Section 4.2 presents, in detail, how task execution time is dynamically computed based on total data size \( \sigma_i \), resources allocated (i.e., processing nodes and bandwidth) and the partitioning method applied to parallelize the computation.

System Model. A cluster consists of a head node, denoted by \( P_0 \), connected via a switch to \( N \) processing nodes, denoted by \( P_1, P_2, \ldots, P_N \). We assume that all processing nodes have the same computational power and all links from the switch to the processing nodes have the same bandwidth. The system model assumes a typical cluster environment in which the head node does not participate in computation. The role of the head node is to accept or reject incoming tasks, execute the scheduling algorithm, divide the workload and distribute data chunks to processing nodes. Since different nodes process different data chunks, the head node sequentially sends every data chunk to its corresponding processing node via the switch. We assume that data transmission does not happen in parallel, although it is straightforward to generalize our model and include the case where some pipelining of communication may occur. For the divisible loads we assume that tasks and subtasks are independent. Therefore, there is no need for processing nodes to communicate with each other.

According to divisible load theory, linear models are used to represent processing and transmission times [28]. In the simplest scenario, the computation time of a load \( \sigma \) is calculated by a cost function \( C_P(\sigma) = \sigma C_{ps} \), where \( C_{ps} \) represents the time to compute a unit of workload on a single processing node. The transmission time of a load \( \sigma \) is calculated by a cost function \( C_I(\sigma) = \sigma C_{tm} \), where \( C_{tm} \) is the time to transmit a unit of workload from the head node to a processing node. For many applications the output data is just a short message and is negligible, particularly considering the very large size of the input data. Therefore, in this paper we only model transfer of applica-
tion input data but not the transfer of output data. The extension to consider the output data transfer using DLT is straightforward.

4 Algorithms

This section presents real-time scheduling algorithms for divisible loads. To develop the algorithms, we need to make three important decisions. The first is to adopt a scheduling policy to determine the order of execution for tasks (Section 4.1). The second decision is to determine the number of processing nodes to allocate to each task and the third is to choose a strategy to partition the task among the allocated nodes (Section 4.2).

4.1 Scheduling Policies

Three scheduling policies to determine the execution order of tasks are investigated: FIFO, EDF and MWF (Maximum Workload derivative First) [16]. The FIFO scheduling algorithm executes tasks following their order of arrival. EDF, a well-known real-time scheduling algorithm, orders tasks by their absolute deadline. MWF, a cost-effective scheduling algorithm, orders tasks following their order of arrival. EDF, a well-known real-time scheduling algorithm, orders tasks by their absolute deadline. FIFO, EDF and MWF (Maximum Workload derivative First) [16]. The FIFO scheduling algorithm executes tasks following their order of arrival. EDF, a well-known real-time scheduling algorithm, orders tasks by their absolute deadline. MWF, a cost-effective scheduling algorithm, orders tasks following their order of arrival. EDF, a well-known real-time scheduling algorithm, orders tasks by their absolute deadline.

4.2 Node Assignment and Task Partitioning

We study two primary strategies for node assignment. First, assign a task all N nodes and thus try to finish the current task as early as possible. Second, assign a task the minimum number of nodes needed to meet its deadline, and thereby save resources for new tasks.

Similarly, two different partitioning methods are investigated: Optimal Partitioning Rule (OPR) (analyzed in Section 4.2.1), and Equal Partitioning Rule (EPR) (analyzed in Section 4.2.2). OPR is based on divisible load theory (DLT), which states that the optimal execution time is obtained when all nodes allocated to the task complete their computation at the same time [28]. For comparison, we propose EPR, based on a common practice of dividing a task into n equal-sized subtasks when the task is to be processed by n nodes.

The following notations, partially adopted from [28], are used in the analysis.

- $T = (A, \sigma, D)$: A divisible task, where $A$ is the arrival time, $\sigma$ is the data size, and $D$ is the relative deadline.

- $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$: Data distribution vector, where $n$ is the number of processing nodes allocated to the task, $\alpha_j$ is the data fraction allocated to the $j$th node, i.e., $\alpha_j \sigma$, is the amount of data that is to be transmitted to the $j$th node for processing, $0 < \alpha_j \leq 1$ and $\sum_{j=1}^{n} \alpha_j = 1$.

- $C_{ms}$: Cost of transmitting a unit workload.

- $C_{ps}$: Cost of processing a unit workload.

Based on our system model (Section 3) we have the following cost functions: the data transmission time on the $j$th link is $C_{ms}(\alpha_j \sigma) = \alpha_j \sigma C_{ms}$ and the data processing time on the $j$th node is $C_{ps}(\alpha_j \sigma) = \alpha_j \sigma C_{ps}$. We assume that the setup costs for initializing data transmission and processing are negligible. We have also investigated the case where the setup costs may be significant [17], but those results are omitted from this paper due to space limitations.

4.2.1 Optimal Partitioning Rule (OPR)

For a given task, we first analyze its execution time function, $E(\sigma, n)$, assuming $n$ nodes are to be allocated to process a total data size of $\sigma$. Then, we use the function to derive the minimum number, $n_{opt}$, of nodes needed to meet the task’s deadline.

Task Execution Time Analysis. Figure 1 shows an example task execution time diagram following OPR when $n$ nodes are allocated to process the task. Let $E$ denote Task Execution Time, which is a function of $\sigma$ and $n$. We have,

\[
E(\sigma, n) = \alpha_1 \sigma C_{ms} + \alpha_2 \sigma C_{ps} + \sum_{j=2}^{n} \alpha_j \sigma C_{ms} + \alpha_j \sigma C_{ps} + \ldots + \alpha_n \sigma C_{ms} + \alpha_n \sigma C_{ps}.
\]

From (2) and (3), we get

\[
\alpha_1 = \frac{\sigma C_{ms} + \sigma C_{ps}}{\sigma C_{ps}} = \frac{C_{ps}}{C_{ms} + C_{ps}} \quad \text{and} \quad \alpha_2 = \beta
\]

where

\[
\beta = \frac{\sigma C_{ps}}{\sigma C_{ms} + \sigma C_{ps}} = \frac{C_{ps}}{C_{ms} + C_{ps}}
\]
Note that $0 < \beta < 1$. It follows that $\alpha_2 = \beta \alpha_1$. Similarly, from (3) and (4), we have $\alpha_3 = \beta \alpha_2$, and therefore, $\alpha_3 = \beta^2 \alpha_1$. This leads to a general formula: $\alpha_j = \beta^{j-1} \alpha_1$. Since $\alpha_j$ is the data fraction distributed to the $j^{th}$ processing node, we have $\sum_{j=1}^{n} \alpha_j = 1$. Substituting $\alpha_j$ with $\beta^{j-1} \alpha_1$ in this equation, we obtain
\[
\alpha_1 + \beta \alpha_1 + \beta^2 \alpha_1 + \ldots + \beta^{n-1} \alpha_1 = 1.
\]
Solving this equation, we get $\alpha_1 = \frac{1 - \beta}{1 - \beta^n}$. Substituting it into (2), we have
\[
\mathcal{E}(\sigma, n) = \frac{1 - \beta}{1 - \beta^n} \sigma(C_{ms} + C_{ps}).
\]  

Derivation of $n_{min}$. Given $\mathcal{E}(\sigma, n)$, we can calculate the minimum number $n_{min}$ of nodes required to meet a task’s deadline.

Let $C(n)$ denote the task completion time function. Assuming that the task $T = (A, \sigma, D)$ has a start time $s$, then its completion time is $C(n) = s + \mathcal{E}(\sigma, n)$, which leads to
\[
C(n) = s + \frac{1 - \beta}{1 - \beta^n} \sigma(C_{ms} + C_{ps}).
\]  

To meet a task’s deadline means its completion time should satisfy the constraint that $C(n) \leq A + D$. It follows that
\[
s + \frac{1 - \beta}{1 - \beta^n} \sigma(C_{ms} + C_{ps}) \leq A + D, \quad \text{that is}
\]
\[
\frac{1 - \beta}{1 - \beta^n} \sigma(C_{ms} + C_{ps}) \leq A + D - s.
\]  

Since $1 - \beta^n > 0$. Multiplying both sides of (9) by $(1 - \beta^n)$, we get
\[
(1 - \beta) \sigma(C_{ms} + C_{ps}) \leq (1 - \beta^n)(A + D - s).
\]  

If $A + D - s \leq 0$, the task will miss its deadline no matter how many nodes we assign to it and how we partition it. Such a task will be rejected because it fails the schedulability test of our scheduling algorithms (for details see Sec. 4.3). Thus, $A + D - s > 0$, and dividing both sides of (10) by $(A + D - s)$ we have
\[
(1 - \beta^n) \geq \frac{(1 - \beta) \sigma(C_{ms} + C_{ps})}{A + D - s}, \quad \text{thus}
\]
\[
\beta^n \leq 1 - \frac{(1 - \beta) \sigma(C_{ms} + C_{ps})}{A + D - s} = 1 - \frac{(C_{ps} C_{ms} + C_{ps}) \sigma(C_{ms} + C_{ps})}{A + D - s} = 1 - \frac{C_{ms} \sigma(C_{ms} + C_{ps})}{A + D - s} = 1 - \frac{C_{ps}}{A + D - s}.
\]

Therefore, following EPR, the minimum number of processing nodes that the task needs at time $s$ to complete before its deadline is $n_{min} = \lceil \frac{\ln \gamma}{\ln \beta} \rceil$ where $\gamma$ is defined above and $\beta$ in (6).

4.2.2 Equal Partitioning Rule (EPR)

To understand the merits of divisible load theory (DLT) in practical real-time cluster-based scheduling, we analyze EPR as a comparison. Similar to the analysis for DLT-based OPR, we derive the task execution time function and $n_{min}$ for EPR.

Task Execution Time Analysis. Assuming $n$ nodes are allocated to a task, an example task execution time diagram following the EPR is shown in Figure 2. By analyzing the diagram, we have $\mathcal{E}(\sigma, n) = \sigma C_{ms} + \alpha_n \sigma C_{ps}$, where $\alpha_n = \frac{1}{n}$. Thus,
\[
\mathcal{E}(\sigma, n) = \sigma C_{ms} + \frac{\sigma C_{ps}}{n}.
\]

Derivation of $n_{min}$. Assuming that the task $T = (A, \sigma, D)$ has a start time $s$, then the task completion time is $C(n) = s + \mathcal{E}(\sigma, n)$, which must satisfy the constraint that $C(n) \leq A + D$. That is,
\[
s + \sigma C_{ms} + \frac{\sigma C_{ps}}{n} \leq A + D.
\]  

Thus,
\[
\frac{\sigma C_{ps}}{A + D - s - \sigma C_{ms}} \leq n.
\]  

Therefore, following EPR, the minimum number of processing nodes that the task needs at time $s$ to complete before its deadline is $n_{min} = \lceil \frac{\ln \gamma}{\ln \beta} \rceil$.

4.3 Algorithm Framework

As is typical for dynamic real-time scheduling algorithms [11, 18, 22], when a task arrives, the scheduler dynamically determines if it is feasible to schedule the new task without compromising the guarantees for previously admitted tasks. The pseudocode for a general Schedulability Test is shown in Figure 3. It could be configured to generate various real-time divisible load scheduling algorithms by giving the design decisions on:
Data structure:
- \( n_{\text{min}} \): the minimum number of processing nodes needed to finish \( T_i \), before its deadline, assuming it is dispatched at time \( t \).
- AvailableNodesList: \( \langle t_i, AN_i \rangle \): a list of number of available nodes along with the time, where \( t_i \) is the time and \( AN_i \) is the number of available nodes.

Pseudocode:
```plaintext
bool Schedulability-Test(T):
    TempTaskList = T + AdmittedTasksQueue:
    // According to the chosen scheduling policy:
    // EDF, FIFO or MWF (Decision #1)
    order TempTaskList:
    // Obtain the available nodes’ information
    generate AvailableNodesList \( \langle t_i, AN_i \rangle \):
    // Introduction
    ScheduledTaskList = \-
    while TempTaskList != T:
        remove \( T(A_n, n, d) \) from TempTaskList:
        // According to the chosen node assignment policy:
        // assigning a task \( \sigma_i \) or \( N \) nodes (Decision \#2)
        identify the earliest time \( t_i \) when the
        available nodes \( AN_i \) \( n_{\text{min}} \)’s \( t_i \) or \( AN_i \times N \)
        // Set scheduled starting time
        \( s_i = t_i \):
        // According to the chosen partitioning rule:
        // OPR or EPR (Decision \#3), calculate, following
        // Eq (4.2) in Section 4.21 or Eq (4.11) in Section 4.22,
        // and set the expected completion time
        \( e_i = s_i + d_i \):
        if \( s_i + d_i > T \) // Deadline misses
            return false
        put \( T(A_n, n, d, s_i, e_i) \) into ScheduledTaskList
        update AvailableNodesList
    // End algorithm
```

Figure 3: Schedulability Test.

1) scheduling policy (FIFO, EDF or MWF), 2) node assignment method (assigning a task all \( N \) or its \( n_{\text{min}} \) nodes), and 3) task partitioning rule (OPR or EPR). Upon completion of the test, if all tasks are schedulable a feasible schedule is developed and the new task is accepted, otherwise, it is rejected.

By following the aforementioned framework, we generate ten algorithms: EDF-OPR-MN, EDF-OPR-AN, EDF-EPR-MN, EDF-EPR-AN, FIFO-OPR-MN, FIFO-OPR-AN, FIFO-EPR-MN, FIFO-EPR-AN, MWF-OPR-MN, and MWF-EPR-MN. The nomenclature of the algorithms, include three parts corresponding to the three algorithm design decisions. The first part denotes the scheduling policy adopted: EDF, FIFO or MWF. The second part represents the choice of the partitioning rule: DLT-based OPR or heuristic EPR. In the third portion of the name, MN means the algorithm assigns a task the minimum number of nodes needed to meet its deadline, and AN means the algorithm assigns all nodes. Since MWF always allocates a task \( n_{\text{min}} \) nodes, the algorithm only has the MN versions.

5 Performance Evaluation

The previous section proposed various real-time cluster-based scheduling algorithms for divisible loads. This section evaluates their performance relative to each other and to changes of various configuration parameters.

Cluster Configuration. We use a discrete simulator to simulate a range of clusters that are compliant to the system model presented in Section 3. For every simulation, three parameters, \( N, C_{ms} \), and \( C_{ps} \), are specified for a cluster.

Workload Generation. To generate task \( T_i = (A_i, \sigma_i, D_i) \), we assume that the interarrival times follow an exponential distribution with a specified mean of \( 1/\lambda \), and task data sizes \( \sigma_i \) are assumed to be normally distributed with a specified mean of \( Avg_{\sigma} \) and a standard deviation equal to the mean. Task relative deadlines are assumed to be uniformly distributed in the range of \( [\frac{Avg_{D}}{2}, 2 \frac{Avg_{D}}{2}] \), where \( Avg_{D} \) is the mean relative deadline. To specify \( Avg_{D} \), a new term \( DCRatio \) is introduced. It is defined as the ratio of mean deadline to mean minimum execution time (cost), that is, \( DCRatio = \frac{Avg_{D}}{Avg_{\sigma}, N} \), where \( E(Avg_{\sigma}, N) \) is the task execution time computed with Eq (7) assuming the task has an average data size \( Avg_{\sigma} \) and runs on all \( N \) processing nodes. Given \( DCRatio \), the cluster size \( N \) and the average data size \( Avg_{\sigma}, Avg_{D} \) is implicitly specified as \( DCRatio \times E(Avg_{\sigma}, N) \). In this way, by \( DCRatio \), task relative deadlines are specified relating to the average task execution time. In addition, a task relative deadline \( D_i \) is chosen to be larger than its minimum execution time \( E(\sigma_i, N) \). In summary, we specify the following parameters for every simulation:

\[
(N, C_{ms}, C_{ps}, 1/\lambda, Avg_{\sigma}, DCRatio)
\]

To analyze how loaded a cluster is for a simulation, we define another metric \( SystemLoad \). It is derived from the specified parameters as

\[
SystemLoad = \frac{E(Avg_{\sigma}, N)}{\lambda},
\]

which corresponds to

\[
SystemLoad = \frac{TotalTaskNumber \times E(N, Avg_{\sigma})}{TotalSimulationTime}.
\]

Sometimes, we specify \( SystemLoad \) for the simulation instead of average interarrival time \( 1/\lambda \). Configuring \( (N, C_{ms}, C_{ps}, SystemLoad, Avg_{\sigma}, DCRatio) \) is equivalent to specifying \( (N, C_{ms}, C_{ps}, 1/\lambda, Avg_{\sigma}, DCRatio) \), because

\[
1/\lambda = \frac{SystemLoad}{E(Avg_{\sigma}, N)}.
\]

To evaluate the performance of the real-time scheduling algorithms, we use the metric, \( Task\ Rejection\ Ratio \), which is the ratio
of the number of tasks rejected by a real-time scheduling algorithm to the total number of tasks arriving at the cluster. The smaller the Task Reject Ratio, the better the real-time scheduling algorithm.

For all figures in this paper, a point on a curve corresponds to the average performance value of ten simulations.\(^2\) In the ten runs, the same parameters \((N, C_{ms}, C_{ps}, 1/\lambda, \text{Avg}_\sigma, D\text{CRatio})\) are specified but different random numbers are generated for task arrival times \(A_i\), data sizes \(\sigma_i\), and deadlines \(D_i\). For each simulation, the TotalSimulationTime is 10,000,000 time units, which is sufficiently long.

In Section 4, we identified three important decisions on Task Partitioning, Node Assignment, and Scheduling Policy in designing real-time, cluster-based scheduling algorithms for divisible loads. The next three subsections evaluate and compare the algorithms proposed in Section 4 and respectively investigate the scenarios where each of these three decisions matters.

5.1 OPR vs. EPR Partitioning

We first evaluate the performance of the following real-time scheduling algorithms with respect to the two proposed partitioning rules (OPR and EPR): EDF-OPR-MN v.s. EDF-EPR-MN, EDF-OPR-AN v.s. EDF-EPR-AN, FIFO-OPR-MN v.s. FIFO-EPR-MN, FIFO-OPR-AN v.s. FIFO-EPR-AN, and MWF-OPR-MN v.s. MWF-EPR-MN. Due to space limitations, we only report the comparison of EDF-OPR-MN v.s. EDF-EPR-MN and EDF-OPR-AN v.s. EDF-EPR-AN here. The performance results for the other pairs are similar (refer to [17] for details).

5.1.1 Simulation Modeling

For our basic simulation model we chose the following parameters: number of processing nodes in the cluster, \(N = 16\); unit data transmission time, \(C_{txs} = 1\); unit data processing time, \(C_{ps} = 100\); SystemLoad changes in the range \([0.1, 0.2, \ldots, 1.0]\); Average data size, \(\text{Avg}_\sigma = 200\); and the ratio of the average deadline to the average execution time, \(D\text{CRatio} = 2\).

Our simulation has a three-fold objective. First, we want to verify our hypothesis that it is advantageous to apply DLT in real-time cluster-based scheduling. Second, we study the effects of \(D\text{CRatio}\), and third, we want to investigate effects of the processing speed.

5.1.2 Merits of DLT for Cluster Scheduling

To study the merits of DLT we employ our basic simulation model without any changes. The four curves in Figure 4a show the Task Reject Ratio of the four algorithms: EDF-OPR-MN, EDF-EPR-MN, EDF-OPR-AN, and EDF-EPR-AN. Observe that EDF-OPR-MN always leads to a lower Task Reject Ratio than EDF-EPR-MN. Similarly, observe that EDF-OPR-AN always achieves a lower Task Reject Ratio than EDF-EPR-AN. These simulation results confirm our hypothesis that it is advantageous to apply DLT in real-time, cluster-based scheduling algorithms. DLT provides an optimal task partitioning, which leads to minimum task execution times, and as a result the cluster can satisfy a larger number of task deadlines.

We carried out the same type of simulations by changing, one at a time, the following cluster or workload parameters: cluster size \(N\), unit transmission time \(C_{txs}\), and average data size \(\text{Avg}_\sigma\). Results are similar to Figure 4a, where algorithms with OPR partitioning always perform better than algorithms with EPR partitioning (refer to [17] for details).

5.1.3 Effects of \(D\text{CRatio}\)

To study the effects of the \(D\text{CRatio}\), we use the same configuration as the basic simulation except that we vary the \(D\text{CRatio}\) over the range \([2, 3, 10, 20, 100]\). For sake of readability, Figure 4b only shows the performance of EDF-OPR-AN and EDF-EPR-AN with \(D\text{CRatio} = 2, 10,\) and 100. Corresponding to different combinations of algorithm and \(D\text{CRatio}\), six curves are produced. Again, Figure 4b shows that the algorithm with OPR partitioning performs better. In addition, we can see as \(D\text{CRatio}\) increases, the performance of EDF-EPR-AN becomes closer to that of EDF-OPR-AN. This is because the higher the \(D\text{CRatio}\), the looser the task relative deadlines are. Consequently, the worse execution times caused by a non-optimal partition, like EPR, will have less impact on the algorithms’ performance. In particular, when \(D\text{CRatio}\) is extremely high (100), the two algorithms perform almost the same.

5.1.4 Effects of Processing Speed

To study effects of processing speed, we vary \(C_{ps}\) over \([10, 50, 100, 500, 1000, 5000, 10000]\) range. The larger \(C_{ps}\), the slower the computation. Figure 4c shows results of EDF-OPR-MN and EDF-EPR-MN with \(C_{ps} = 10\) and 10000 respectively.

From the figure, we can see that the algorithm with OPR partitioning (EDF-OPR-MN) still outperforms the algorithm with EPR partitioning (EDF-EPR-MN). However, as the processing speed decreases, i.e., \(C_{ps}\) increases, the differences between the two algorithms becomes less and less significant. In particular, when the computation is extremely slow (\(C_{ps} = 10000\)), we can see that the curves for the two algorithms are almost overlapped, indicating non-differentiable Task Reject Ratios. To demonstrate this point, let us assume \(C_{ps}\) is so large that the ratio of \(C_{ms}\) to \(C_{ps}\) is approaching 0. In the analysis of OPR, we showed that \(\beta\) from Eq (6) will approach 1 in this case, causing the data fractions allocated to processing nodes, \(\alpha_1, \alpha_2, \ldots, \alpha_n\), to all be close to \(\frac{1}{n}\). Therefore, OPR and EPR will perform the same in this case.

From the aforementioned intensive experiments, we conclude that no matter what the system parameters are, the algorithms with DLT-based partitioning (OPR) always perform better than the ones with the equal-sized partitioning heuristic (EPR). This shows that it is beneficial to apply divisible load theory in real-time, cluster-based scheduling.

5.2 \(N\) vs. \(n_{\text{min}}\) Nodes

In this subsection, we compare and analyze the real-time scheduling algorithms with different node assignment methods. We investigate the performance difference in algorithms assigning all \(N\) nodes to every task (ALG-AN) v.s. those assigning...
Figure 4: Performance Evaluation.
the minimum number $n^{\text{min}}$ of nodes needed to meet a task’s deadline (ALG-MN).

The relative performance of EDF-OPR-MN v.s. EDF-OPR-AN is shown in Section 5.2.1. It is noteworthy that in contrast to the results in [16] comparing MWF-(MN) and FIXED-(AN) algorithms, our initial data seem to indicate that EDF-OPR-AN outperforms EDF-OPR-MN most of the time. To gain insight into the performance results, in Section 5.2.2 we carry out rigorous analysis of a simplified scenario where a scheduling algorithm always assigns $K$ nodes ($K < N$) to a periodic divisible task. This analysis sheds new light on possible scenarios where algorithms assigning $n^{\text{min}}$ nodes (ALG-MN) perform better than those assigning all $N$ nodes (ALG-AN).

### 5.2.1 Initial Comparison

In Section 4.2, we have explained the rational behind the two node-assignment strategies: an ALG-AN tries to finish the current task as soon as possible, while an ALG-MN saves resources for new tasks.

For the $N$ node assignment strategy, the problem is it causes higher parallel execution overhead than the $n^{\text{min}}$ node assignment counterpart. For the cluster model investigated (Section 3), the higher the transmission cost ($C_{\text{ms}}$), the greater the overhead. As shown in Figure 1, the node idle time due to data transmission is one type of parallel execution overhead.

On the other hand, the $n^{\text{min}}$ node assignment strategy, trying to save resources for new tasks, will not improve performance if there is no task coming in the near future. In that case, some number of nodes (potentially $N - n^{\text{min}}$ of them) are left idle and their processing cycles are wasted. Since the larger the interarrival time $1/\lambda$, the less frequent are the tasks arrivals, we believe that ALG-MN will lose its performance gain over ALG-AN as the interarrival time $1/\lambda$ increases.

We conducted intensive simulations, comparing EDF-OPR-MN v.s. EDF-OPR-AN, to verify the aforementioned analysis and present some conclusive simulation results.

As explained, EDF-OPR-AN leads to higher overhead than EDF-OPR-MN, and the larger $C_{\text{ms}}$, the higher the parallel overhead. Thus, we expect that as $C_{\text{ms}}$ increases, the performance of EDF-OPR-AN should be affected more than that of EDF-OPR-MN. Figure 4d shows the relative performance of the two algorithms ($\text{Task Reject Ratio}$ of EDF-OPR-MN $-$ \text{Task Reject Ratio} of EDF-OPR-AN) in a simulation where we gradually increase the value of $C_{\text{ms}}$. We can see as $C_{\text{ms}}$ gets larger, the difference between the \text{Task Reject Ratios} of EDF-OPR-MN and EDF-OPR-AN decreases, indicating that the relative performance of EDF-OPR-MN v.s. EDF-OPR-AN improves. Interestingly, for this simulation, the curve is above 0 for most of the cases, indicating EDF-OPR-MN performs worse than EDF-OPR-AN.

The next group of simulations verifies our prediction that EDF-OPR-MN will perform worse relative to EDF-OPR-AN as task interarrival time $1/\lambda$ increases. As demonstrated in Figure 4e, with the increasing of the interarrival time, \text{Task Reject Ratio} of EDF-OPR-MN $-$ \text{Task Reject Ratio} of EDF-OPR-AN also increases. The relative \text{Task Reject Ratio} curve once again lies above 0, meaning EDF-OPR-MN performs worse than EDF-OPR-AN.

The results from our initial simulations contradict the conclusion drawn by [16] that the $n^{\text{min}}$ node assignment strategy performs better than the maximum node assignment strategy.

We believe that there should be some scenarios where ALG-MN performs better than ALG-AN, while in the other scenarios (for instance, Figures 4d and 4e) the reverse is true. To show that there are cases where ALG-MN outperforms ALG-AN, we purposely configure a simulation where ALG-MN should have improved performance.

For this simulation, we choose a similar configuration as that in Figure 4e except changing the $D_{\text{CRatio}}$ from 2 to 10. By increasing the $D_{\text{CRatio}}$, we have longer relative deadlines compared to the mean execution time. For an ALG-MN, a longer deadline leads to a smaller $n^{\text{min}}$ of nodes allocated to a task, thus smaller overhead. While for an ALG-AN, its node assignment and resulting overhead will not be affected by deadlines, since a task is always assigned all $N$ nodes. Therefore, we believe, as $D_{\text{CRatio}}$ increases and ALG-MN’s overhead decreases, ALG-MN might perform better than ALG-AN.

Figure 4f validates our analysis. We can successfully create some scenarios (those with interarrival time smaller than 1000 time units), by increasing $D_{\text{CRatio}}$ from 2 to 10, where EDF-OPR-MN outperforms EDF-OPR-AN.

### 5.2.2 Analysis

In this subsection, we study a simplified scenario where there is only one periodic divisible task. A new algorithm, ALG-K that always assigns $K$ nodes ($K < N$) to every task is investigated, which is compared with an ALG-AN. We demonstrate that with this simple analysis we could identify some scenarios where an ALG-MN will be better than its corresponding ALG-AN when handling aperiodic divisible tasks.

Here, we assume the system model described in Section 3, while the task model is simplified: only one periodic task $T$ is running in the system. We assume that in every period $P$ a subtask with a fixed data size $\sigma$ and a relative deadline $D$ arrives at the cluster. For such system and task models, we have proved the following theorems [17].

**Theorem 5.1** When an ALG-AN is applied, if $P < E(\sigma, N)$ and $D$ is finite, some subtasks are doomed to miss their deadlines.

**Theorem 5.2** For $N > 1$, $1 \leq K < N$, if $P \geq \frac{KE(\sigma, K)}{N - K}$, $C_{\text{ps}} > (N - 1)C_{\text{ms}}$, and $D \geq E(\sigma, K)$, no subtask will miss its deadline when an ALG-K is applied.

From the above theorems, it is further derived that if $P \in \left[\frac{E(\sigma, K)}{N - K}, E(\sigma, N)\right]$, $C_{\text{ps}} > (N - 1)C_{\text{ms}}$, $D \geq E(\sigma, K)$, and $1 \leq K < N$, an ALG-K will perform better than a corresponding ALG-AN.

We carry out simulations to verify this conclusion. A periodic task $T$ that includes subtasks with $\sigma = 200$ and $D = E(\sigma, 2)$ is simulated to run in a $(N = 16, C_{\text{ms}} = 1, C_{\text{ps}} = 100)$ cluster by EDF-OPR-2 or EDF-OPR-AN. According to the above
Table 1: For a \((N = 16, C_{ms} = 1, C_{ps} = 100)\) cluster, the verification of the derived range where EDF-OPR-K outperforms EDF-OPR-AN.

<table>
<thead>
<tr>
<th>(K)</th>
<th>Range</th>
<th>EDF-OPR-K TRR</th>
<th>EDF-OPR-AN TRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[1263,1359)</td>
<td>0.0187</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>[1269,1359)</td>
<td>0.0251</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>[1282,1359)</td>
<td>0.0251</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>[1307,1359)</td>
<td>0.0187</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: For a \((N = 64, C_{ms} = 1, C_{ps} = 100)\) cluster, the derived range where EDF-OPR-K outperforms EDF-OPR-AN.

<table>
<thead>
<tr>
<th>(K)</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[316,425)</td>
</tr>
<tr>
<td>2</td>
<td>[318,425)</td>
</tr>
<tr>
<td>3</td>
<td>[335,425)</td>
</tr>
<tr>
<td>4</td>
<td>[321,425)</td>
</tr>
<tr>
<td>5</td>
<td>[350,425)</td>
</tr>
<tr>
<td>6</td>
<td>[357,425)</td>
</tr>
<tr>
<td>7</td>
<td>[366,425)</td>
</tr>
<tr>
<td>8</td>
<td>[327,425)</td>
</tr>
</tbody>
</table>

5.3 FIFO, EDF and MWF

In this subsection, we examine the effects of different execution order policies and compare algorithms FIFO-OPR-MN, EDF-OPR-MN v.s. MWF-OPR-MN.

Recall that the MWF (Maximum Workload derivative First) algorithm, proposed in [16], executes the task with highest workload derivative \((DC_i)\) first, and thus reduces the total workload (cost) of all scheduled tasks. In [16] MWF is compared with EDF and shown that MWF performs better than EDF. Moreover, the authors claim that MWF is likely to be the best choice for on-line scheduling of divisible tasks.

We conducted intensive simulations and a systematic study of the three execution order strategies. Our data cast some doubts on the conclusion drawn in [16] that the MWF algorithm is the best choice. Our hypothesis is that MWF performs well when the parallel execution overhead (workload) of the tasks is significant compared to their pure computation. To test our theory, a group of simulations is designed to study how changing parallel overhead affects the performance of scheduling algorithms. In the 20 simulations, we gradually change the data transmission \((C_{ms})\) from 1 to 20, while keeping the data processing cost \((C_{ps})\) constant. Since the bigger the \(C_{ms}\), the higher the parallel execution overhead, for the 20 simulations with \(C_{ms}\) changing from 1 to 20, the task overhead increases. According to our theory, MWF should perform better than EDF and FIFO when \(C_{ms}\) increases.

Figures 4h, 4i and 4j show the results for simulations where \(C_{ms} = 2, 10\) and 20 respectively. As observed, when \(C_{ms}\) is small (Figure 4h), the Task Reject Ratio curve of EDF-OPR-MN lies below that of MWF-OPR-MN, indicating EDF execution order performs better. As \(C_{ms}\) increases (Figures 4i and 4j), the relative performance of the two algorithms changes. When \(C_{ms}\) increases to 20 (Figure 4j), MWF-OPR-MN outperforms EDF-OPR-MN. These data match our analysis and verify our hypothesis that MWF performs better than EDF and FIFO as workload parallel overhead increases.

Interestingly, for all 20 simulations, EDF-OPR-MN always performs better than FIFO-OPR-MN, while in some scenarios like the one in Figure 4h, MWF-OPR-MN performs even worse than FIFO-OPR-MN.

Another observation is that as the interarrival time increases, the performance of the three algorithms become non-differentiable (see Figure 4h). As the interarrival time gets larger and the system load smaller, either EDF and MWF generate the same execution order as FIFO, or the load is so light that the choice of execution order does not matter anymore.

In summary, our data indicate that the best choice of execution order policy depends on individual system and workload parameters.
conditions. It might be appropriate and desirable to have an adaptive scheduling algorithm where policies can be configured with changing conditions. In the future, we plan to study real cluster workload traces and further investigate this problem.

6 Conclusion
In this paper, we address the problem of providing deterministic QoS to arbitrarily divisible applications executing in a cluster. Four contributions are made. First, we extend DLT to compute the minimum number of processors required to meet an application deadline. Second, based on this, a novel algorithmic approach integrating DLT and EDF scheduling is proposed. Third, important design parameters are identified that affect the performance of real-time divisible-load scheduling algorithms. Finally, we systematically investigated the effects of these design parameters on a set of real-time scheduling algorithms, and show that the application of DLT to real-time, cluster-based scheduling leads to significantly better scheduling approaches.

References
[26] D. Swanson. Personal communication. Director, UNL Research Computing Facility (RCF) and UNL CMS Tier-2 Site, August 2005.