Static scheduling algorithms for allocating directed task graphs to multiprocessors
Yu-Kwong Kwok , Ishfaq Ahmad
ACM Computing Surveys (CSUR) December 1999
Volume 31 Issue 4

Static Scheduling Algorithms for Allocating Directed Task Graphs to Multiprocessors

YU-KWONG KWOK

The University of Hong Kong

AND

ISHFAQ AHMAD

The Hong Kong University of Science and Technology

Static scheduling of a program represented by a directed task graph on a multiprocessor system to minimize the program completion time is a well-known problem in parallel processing. Since finding an optimal schedule is an NP-complete problem in general, researchers have resorted to devising efficient heuristics. A plethora of heuristics have been proposed based on a wide spectrum of techniques, including branch-and-bound, integer-programming, searching, graph-theory, randomization, genetic algorithms, and evolutionary methods. The objective of this survey is to describe various scheduling algorithms and their functionalities in a contrasting fashion as well as examine their relative merits in terms of performance and time-complexity. Since these algorithms are based on diverse assumptions, they differ in their functionalities, and hence are difficult to describe in a unified context. We propose a taxonomy that classifies these algorithms into different categories. We consider 27 scheduling algorithms, with each algorithm explained through an easy-to-understand description followed by an illustrative example to demonstrate its operation. We also outline some of the novel and promising optimization approaches and current research trends in the area. Finally, we give an overview of the software tools that provide scheduling/mapping functionalities

Categories and Subject Descriptors: C.1.2 [**Processor Architectures**]: Multiple Data Stream Architectures (Multiprocessors); *Parallel processors*; D.1.3 [**Programming Techniques**]: Concurrent Programming; *Parallel programming*; D.4.1 [**Operating Systems**]: Process Management—*Multiprocessing/multiprogramming*; *Scheduling*; F.1.2 [**Conputation by Abstract Devices**]: Modes of Computation—*Parallelism and concurrency*

General Terms: Algorithms, Design, Performance, Theory

Additional Key Words and Phrases: Automatic parallelization, DAG, multiprocessors, parallel processing, software tools, static scheduling, task graphs

---

This research was supported by the Hong Kong Research Grants Council under contract numbers HKUST 734/96E, HKUST 6076/97E, and HKU 7124/99E.

Authors' addresses: Y.-K. Kwok, Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong; email: ykwok@eee.hku.hk; I. Ahmad, Department of Computer Science, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong.

CONTENTS

1. Introduction
2. The DAG Scheduling Problem
   2.1 The DAG Model
   2.2 Generation of a DAG
   2.3 Variations in the DAG Model
   2.4 The Multiprocessor Model
3. NP-Completeness of the DAG Scheduling Problem
4. A Taxonomy of DAG Scheduling Algorithms
5. Basic Techniques in DAG Scheduling
   Computing a *t-level*
   Computing a *b-level*
   Computing ALAP
6. Description of the Algorithms
   6.1 Scheduling DAGs with Restricted Structures
   6.2 Scheduling Arbitrary DAGs Without Communication
   6.3 UNC Scheduling
   6.4 BNP Scheduling
   6.5 TDB Scheduling
   Constructing the CPN-Dominant Sequence
   6.6 APN Scheduling
   The BSA Algorithm
   6.7 Scheduling in Heterogeneous Environments
   6.8 Mapping Clusters to Processors
7. Some Scheduling Tools
   7.1 Hypertool
   7.2 PYRROS
   7.3 Parallax
   7.4 OREGAMI
   7.5 PARSA
   7.6 CASCH
   7.7 Commercial Tools
8. New Ideas and Research Trends
   8.1 Scheduling Using Genetic Algorithms
   8.2 Randomization Techniques
   8.3 Parallelizing a Scheduling Algorithm
   8.4 Future Research Directions
9. Summary and Concluding Remarks

1. INTRODUCTION

Parallel processing is a promising approach to meet the computational requirements of a large number of current and emerging applications [Hwang 1993; Kumar et al. 1994; Quinn 1994]. However, it poses a number of problems that are not encountered in sequential processing such as designing a parallel algorithm for the application, partitioning of the application into tasks, coordinating communication and synchronization, and scheduling of the tasks onto the machine. A large body of research efforts addressing these problems has been reported in the literature [Amdahl 1967; Chu et al. 1984; Gajski and Peir 1985; Hwang 1993; Lewis and El-Rewini 1993; Lo et al. 1991; Lord et al. 1983; Manoharan and Topham 1995; Pease et al. 1991; Quinn 1994; Shirazi et al. 1993; Wu and Gajski 1990; Yang and Gerasoulis 1992]. Scheduling and allocation is a highly important issue since an inappropriate scheduling of tasks can fail to exploit the true potential of the system and can offset the gain from parallelization. In this paper we focus on the scheduling aspect.

The objective of scheduling is to minimize the completion time of a parallel application by properly allocating the tasks to the processors. In a broad sense, the scheduling problem exists in two forms: *static* and *dynamic*. In static scheduling, which is usually done at compile time, the characteristics of a parallel program (such as task processing times, communication, data dependencies, and synchronization requirements) are known before program execution [Chu et al. 1984; Gajski and Peir 1985]. A parallel program, therefore, can be represented by a node- and edge-weighted directed acyclic graph (DAG), in which the node weights represent task processing times and the edge weights represent data dependencies as well as the communication times between tasks. In dynamic scheduling only, a few assumptions about the parallel program can be made before execution, and thus, scheduling decisions have to be made on-the-fly [Ahmad and Ghafoor 1991; Palis et al. 1995]. The goal of a dynamic scheduling algorithm as such includes not only the minimization of the program completion time but also the minimization of the scheduling overhead which constitutes a significant portion of the cost paid for running the scheduler. We address only the static scheduling problem. Hereafter, we refer to the static scheduling problem as simply scheduling.

The scheduling problem is NP-complete for most of its variants except for a few simplified cases (these cases will be elaborated in later sections) [Chretienne 1989; Coffman 1976; Coffman and

Graham 1972; El-Rewini et al. 1995; Garey and Johnson 1979; Gonzales, Jr. 1977; Graham et al. 1979; Hu 1961; Kasahara and Narita 1984; Papadimitriou and Ullman 1987; Papadimitriou and Yannakakis 1979; 1990; Rayward–Smith 1987b; Sethi 1976; Ullman 1975]. Therefore, many heuristics with polynomial-time complexity have been suggested [Ahmad et al. 1996; Casavant and Kuhl 1988; Coffman 1976; El-Rewini et al. 1995; El-Rewini et al. 1994; Gerasoulis and Yang 1992; Khan et al. 1994; McCreary et al. 1994; Pande et al. 1994; Prastein 1987; Shirazi et al. 1990; Simons and Warmuth 1989]. However, these heuristics are highly diverse in terms of their assumptions about the structure of the parallel program and the target parallel architecture, and thus are difficult to explain in a unified context.

Common simplifying assumptions include uniform task execution times, zero inter-task communication times, contention-free communication, full connectivity of parallel processors, and availability of unlimited number of processors. These assumptions may not hold in practical situations for a number of reasons. For instance, it is not always realistic to assume that the task execution times of an application are uniform because the amount of computations encapsulated in tasks are usually varied. Furthermore, parallel and distributed architectures have evolved into various types such as distributed-memory multicomputers (DMMs) [Hwang 1993]; shared-memory multiprocessors (SMMs) [Hwang 1993]; clusters of symmetric multiprocessors (SMPs) [Hwang 1993]; and networks of workstations (NOWs) [Hwang 1993]. Therefore, their more detailed architectural characteristics must be taken into account. For example, intertask communication in the form of message-passing or shared-memory access inevitably incurs a non-negligible amount of latency. Moreover, a contention-free communication and full connectivity of processors cannot be assumed for a DMM, a SMP or a NOW. Thus, scheduling algorithms relying on such assumptions are apt to have restricted applicability in real environments.

Multiprocessor scheduling has been an active research area and, therefore, many different assumptions and terminology are independently suggested. Unfortunately, some of the terms and assumptions are neither clearly stated nor consistently used by most of the researchers. As a result, it is difficult to appreciate the merits of various scheduling algorithms and quantitatively evaluate their performance. To avoid this problem, we first introduce the directed acyclic graph (DAG) model of a parallel program, and then proceed to describe the multiprocessor model. This is followed by a discussion about the NP-completeness of variants of the DAG scheduling problem. Some basic techniques used in scheduling are introduced. Then we describe a taxonomy of DAG scheduling algorithms and use it to classify several reported algorithms.

The problem of scheduling a set of tasks to a set of processors can be divided into two categories: job scheduling and scheduling and mapping (see Figure 1(a)). In the former category, independent jobs are to be scheduled among the processors of a distributed computing system to optimize overall system performance [Bozoki and Richard 1970; Chen and Lai 1988a; Cheng et al. 1986]. In contrast, the scheduling and mapping problem requires the allocation of multiple interacting tasks of a single parallel program in order to minimize the completion time on the parallel computer system [Adam et al. 1974; Ahmad et al. 1996; Bashir et al. 1983; Casavant and Kuhl 1988; Coffman 1976; Veltman et al. 1990]. While job scheduling requires dynamic run-time scheduling that is not a priori decidable, the scheduling and mapping problem can be addressed in both static [El-Rewini et al. 1995; 1994; Gerasoulis and Yang 1992; Hochbaum and Shmoys 1987; 1988; Khan et al. 1994; McCreary et al. 1994; Shirazi et al. 1990] as well as dynamic

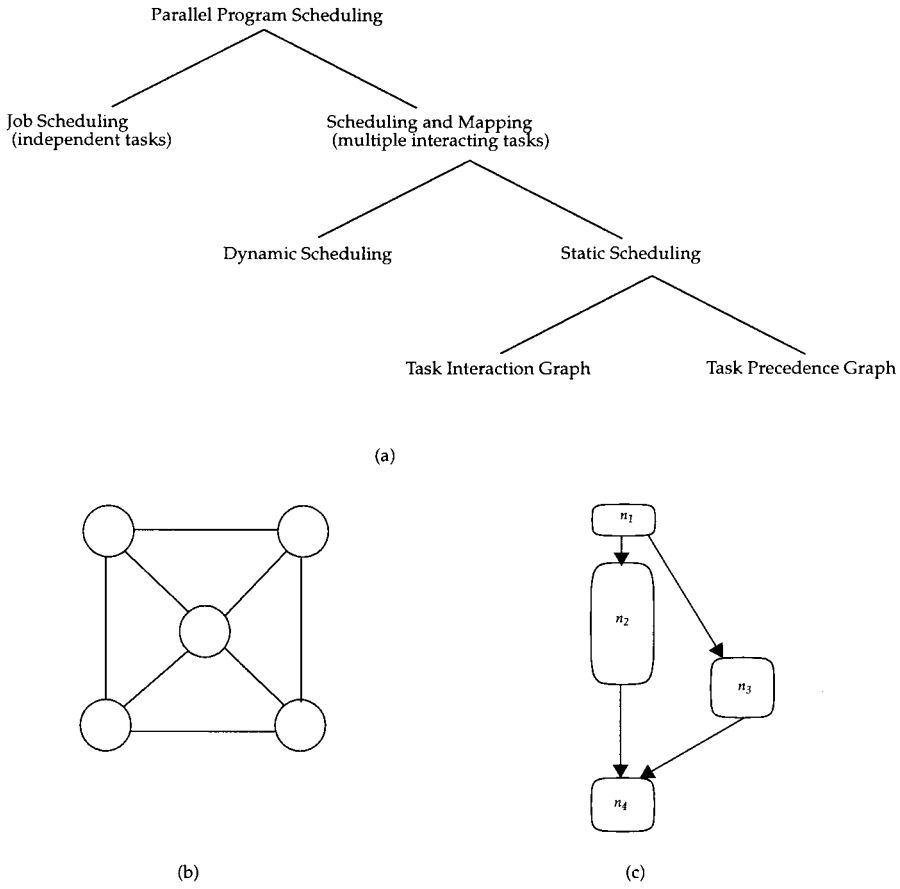


Figure 1. (a) A simplified taxonomy of the approaches to the scheduling problem; (b) a task interaction graph; (c) a task precedence graph.

contexts [Ahmad and Ghafoor 1991; Norman and Thanisch 1993]. When the characteristics of the parallel program, including its task execution times, task dependencies, task communications and synchronization are known a priori, scheduling can be accomplished off-line during compile-time. On the contrary, dynamic scheduling in the absence of a priori information is done on-the-fly according to the state of the system.

Two distinct models of the parallel program have been considered extensively in the context of static scheduling: the task interaction graph (TIG) model and the task precedence graph (TPG) model (see Figure 1(b) and Figure 1(c)).

The task interaction graph model, in which vertices represent parallel processes and edges denote the interprocess interaction [Bokhari 1981], is usually used in static scheduling of loosely coupled communicating processes (since all tasks are considered as simultaneously and independently executable, there is no temporal execution dependency) to a distributed system. For example, a TIG is commonly used to model the finite element method (FEM) [Bokhari 1979]. The objective of scheduling is to minimize parallel program completion time by properly mapping the tasks to the processors. This requires balancing the computation load uniformly among the processors while

simultaneously keeping communication costs as low as possible. The research in this area was pioneered by Bokhari [1979] and Stone [1977]: Stone [1977] applied network-flow algorithms to solve the assignment problem, whereas Bokhari [1981] described the mapping problem as being equivalent to graph isomorphism, quadratic assignment, and sparse matrix bandwidth reduction problems.

The task precedence graph model (or simply the DAG) in which the nodes represent the tasks and the directed edges represent the execution dependencies as well as the amount of communication, is commonly used in static scheduling of a parallel program with tightly coupled tasks on multiprocessors. For example, in the task precedence graph shown in Figure 1(c), task n_4 cannot commence execution before tasks n_1 and n_2 finish execution and gathers all the communication data from n_2 and n_3. The scheduling objective is to minimize the program completion time (or maximize the speed-up, defined as the time required for sequential execution divided by the time required for parallel execution). For most parallel applications, a task precedence graph can model the program more accurately because it captures the temporal dependencies among tasks. This is the model we use in this paper.

As mentioned above, earlier static scheduling research made simplifying assumptions about the architecture of the parallel program and the parallel machine, such as uniform node weights, zero edge weights, and the availability of an unlimited number of processors. However, even with some of these assumptions, the scheduling problem has been proven to be NP-complete except for a few restricted cases [Garey and Johnson 1979]. Indeed, the problem is NP-complete even in two simple cases: (1) scheduling tasks with uniform weights to an arbitrary number of processors [Ullman 1975] and (2) scheduling tasks with weights equal to one or two units to two processors [Ullman 1975]. There are only three special cases for which there exists optimal polynomial-time algorithms. These cases are (1) scheduling tree-structured task graphs with uniform computation costs on an arbitrary number of processors [Hu 1961]; (2) scheduling arbitrary task graphs with uniform computation costs on two processors [Coffman and Graham 1972]; and (3) scheduling an interval-ordered task graph [Fishburn 1985] with uniform node weights to an arbitrary number of processors [Papadimitriou and Yannakakis 1979]. However, even in these cases, communication among tasks of the parallel program is assumed to take zero time [Coffman 1976]. Given these observations, the general scheduling problem cannot be solved in polynomial-time unless $P = NP$.

Due to the intractability of the general scheduling problem, two distinct approaches have been taken: sacrificing efficiency for the sake of optimality and sacrificing optimality for the sake of efficiency. To obtain optimal solutions under relaxed constraints, state-space search and dynamic programming techniques have been suggested. However, these techniques are not useful because most of them are designed to work under restricted environments and most importantly they incur an exponential time in the worst case. In view of the ineffectiveness of optimal techniques, many heuristics have been suggested to tackle the problem under more pragmatic situations. While these heuristics are shown to be effective in experimental studies, they usually cannot generate optimal solutions, and there is no guarantee about their performance in general. Most of the heuristics are based on a list scheduling approach [Coffman 1976], which is explained below.

2. THE DAG SCHEDULING PROBLEM

The objective of DAG scheduling is to minimize the overall program finish-

time by proper allocation of the tasks to the processors and arrangement of execution sequencing of the tasks. Scheduling is done in such a manner that the precedence constraints among the program tasks are preserved. The overall finish-time of a parallel program is commonly called the schedule length or makespan. Some variations to this goal have been suggested. For example, some researchers proposed algorithms to minimize the mean flow-time or mean finish-time, which is the average of the finish-times of all the program tasks [Bruno et al. 1974; Leung and Young 1989]. The significance of the mean finish-time criterion is that minimizing it in the final schedule leads to the reduction of the mean number of unfinished tasks at each point in the schedule. Some other algorithms try to reduce the setup costs of the parallel processors [Sumichrast 1987]. We focus on algorithms that minimize the schedule length.

2.1 The DAG Model

A parallel program can be represented by a directed acyclic graph (DAG) $G = (V, E)$, where V is a set of v nodes and E is a set of e directed edges. A node in the DAG represents a task which in turn is a set of instructions which must be executed sequentially without preemption in the same processor. The weight of a node n_i is called the computation cost and is denoted by $w(n_i)$. The edges in the DAG, each of which is denoted by (n_i, n_j), correspond to the communication messages and precedence constraints among the nodes. The weight of an edge is called the communication cost of the edge and is denoted by $c(n_i, n_j)$. The source node of an edge is called the parent node while the sink node is called the child node. A node with no parent is called an entry node and a node with no child is called an exit node. The communication-to-computation-ratio (CCR) of a parallel program is defined as its average edge weight divided by its average node weight. Hereafter, we use the terms node and task interchangeably. We summarize in Table I the notation used throughout the paper.

The precedence constraints of a DAG dictate that a node cannot start execution before it gathers all of the messages from its parent nodes. The communication cost between two tasks assigned to the same processor is assumed to be zero. If node n_i is scheduled to some processor, then $ST(n_i)$ and $FT(n_i)$ denote the start-time and finish-time of n_i, respectively. After all the nodes have been scheduled, the schedule length is defined as $max_i\{FT(n_i)\}$ across all processors. The goal of scheduling is to minimize $max_i\{FT(n_i)\}$.

The node and edge weights are usually obtained by estimation at compile-time [Ahmad et al. 1997; Chu et al. 1984; Ha and Lee 1991; Cosnard and Loi 1995; Wu and Gajski 1990]. Generation of the generic DAG model and some of the variations are described below.

2.2 Generation of a DAG

A parallel program can be modeled by a DAG. Although program loops cannot be explicitly represented by the DAG model, the parallelism in data-flow computations in loops can be exploited to subdivide the loops into a number of tasks by the loop-unraveling technique [Beck et al. 1990; Lee and Feng 1991]. The idea is that all iterations of the loop are started or fired together, and operations in various iterations can execute when their input data are ready for access. In addition, for a large class of data-flow computation problems and many numerical algorithms (such as matrix multiplication), there are very few, if any, conditional branches or indeterminism in the program. Thus, the DAG model can be used to accurately represent these applications so that the scheduling techniques can be applied. Furthermore, in many numerical appli-

Table I. Notation

Symbol	Definition
n_i	The node number of a node in the parallel program task graph
$w(n_i)$	The computation cost of node n_i
(n_i, n_j)	An edge from node n_i to n_j
$c(n_i, n_j)$	The communication cost of the directed edge from node n_i to n_j
v	Number of nodes in the task graph
e	Number of nodes in the task graph
p	Number of edges in the task graph
CP	The number of processors or processing elements (PEs) in the target system
CP	A critical path of the task graph
CPN	Critical Path Node
IBN	In-Branch Node
OBN	Out-Branch Node
sl	Static level of a node
$b\text{-}level$	Bottom level of a node
$t\text{-}level$	Top level of a node
$ASAP$	As soon as possible start time of a node
$ALAP$	As late as possible start time of a node
$T_s(n_i)$	The actual start time of a node n_i
$DAT(n_i, P)$	The possible data available time of n_i on target processor P
$ST(n_i, P)$	The start time of node n_i on target processor P
$FT(n_i, P)$	The finish time of node n_i on target processor P
$VIP(n_i)$	The parent node of n_i that sends the data arrive last
$Pivot\_PE$	The target processor from which nodes are migrated
$Proc(n_i)$	The processor accommodating node n_i
L_{ij}	The communication link between PE i and PE j
CCR	Communication-to-computation Ratio
SL	Schedule Length
UNC	Unbounded Number of Clusters scheduling algorithms
BNP	Bounded Number of Processors scheduling algorithms
TDB	Task Duplication Based scheduling algorithms
APN	Arbitrary Processors Network scheduling algorithms

cations, such as Gaussian elimination or fast Fourier transform (FFT), the loop bounds are known during compile-time. As such, one or more iterations of a loop can be deterministically encapsulated in a task and, consequently, be represented by a node in a DAG.

The node- and edge-weights are usually obtained by estimation using profiling information of operations such as numerical operations, memory access operations, and message-passing primitives [Jiang et al. 1990]. The granularity of tasks usually is specified by the programmers [Ahmad et al. 1997]. Nevertheless, the final granularity of the scheduled parallel program is to be refined by using a scheduling algorithm, which clusters the communication-intensive tasks to a single processor [Ahmad et al. 1997; Yang and Gerasoulis 1992].

2.3 Variations in the DAG Model

There are a number of variations in the generic DAG model described above. The more important variations are: preemptive scheduling vs. nonpreemptive scheduling, parallel tasks vs. non-parallel tasks, and DAG with conditional branches vs. DAG without conditional branches.

Preemptive Scheduling vs. Nonpreemptive Scheduling: In preemptive scheduling, the execution of a task may be interrupted so that the unfinished portion of the task can be re-allocated to a different processor [Chen and Lai 1988b; Gonzales and Sahni 1978; Horvath et al. 1977; Rayward-Smith 1987a]. On the contrary, algorithms assuming nonpreemptive scheduling must allow a task to execute until completion on a single processor. From a theoretical perspective, a preemptive scheduling approach allows more flexibility for the scheduler so that a higher utilization of processors may result. Indeed, a preemptive scheduling problem is commonly reckoned as "easier" than its nonpreemptive counterpart in that there are cases in which polynomial time solutions exist for the former while the latter is proved to be NP-complete [Coffman and Graham 1972; Gonzalez, Jr.

1977]. However, in practice, interrupting a task and transferring it to another processor can lead to significant processing overhead and communication delays. In addition, a preemptive scheduler itself is usually more complicated since it has to consider when to split a task and where to insert the necessary communication induced by the splitting. We concentrate on the nonpreemptive approaches.

Parallel Tasks vs. Nonparallel Tasks: A parallel task is a task that requires more than one processor at the same time for its execution [Wang and Cheng 1991]. Blazewicz et al. [1986; 1984] investigated the problem of scheduling a set of independent parallel tasks to identical processors under preemptive and nonpreemptive scheduling assumptions. Du and Leung [1989] also explored the same problem but with one more flexibility: a task can be scheduled to no more than a certain predefined maximum number of processors. However, in Blazewicz et al. 's approach, a task must be scheduled to a fixed predefined number of processors. Wang and Cheng [1991] further extended the model to allow precedence constraints among tasks. They devised a list scheduling approach to construct a schedule based on the earliest completion time (ECT) heuristic. We concentrate on scheduling DAGs with nonparallel tasks.

DAG with Conditional Branches vs. DAG without Conditional Branches: Towsley [1986] addressed the problem of scheduling a DAG with probabilistic branches and loops to heterogeneous distributed systems. Each edge in the DAG is associated with a nonzero probability that the child will be executed immediately after the parent. He introduced two algorithms based on the shortest path method for determining the optimal assignments of tasks to processors. El-Rewini and Ali [1995] also investigated the problem of scheduling DAGs with conditional branches. Similar to Towsley's approach, they also used a two-step method to construct a final schedule. However, unlike Towsley's model, they modeled a parallel program by using two DAGs: a branch graph and a precedence graph. This model differentiates the conditional branching and the precedence relations among the parallel program tasks. The objective of the first step of the algorithm is to reduce the amount of indeterminism in the DAG by capturing the similarity of different instances of the precedence graph. After this preprocessing step, a reduced branch graph and a reduced precedence graph are generated. In the second step, all the different instances of the precedence graph are generated according to the reduced branch graph, and the corresponding schedules are determined. Finally, these schedules are merged to produce a unified final schedule [El-Rewini and Ali 1995]. Since modeling branching and looping in DAGs is an inherently difficult problem, little work has been reported in this area. We concentrate on DAGs without conditional branching in this research.

2.4 The Multiprocessor Model

In DAG scheduling, the target system is assumed to be a network of *processing elements* (PEs), each of which is composed of a processor and a local memory unit so that the PEs do not share memory and communication relies solely on message-passing. The processors may be heterogeneous or homogeneous. Heterogeneity of processors means the processors have different speeds or processing capabilities. However, we assume every module of a parallel program can be executed on any processor even though the completion times on different processors may be different. The PEs are connected by an interconnection network with a certain topology. The topology may be fully connected or of a particular structure such as a hypercube or mesh.

Table II. Summary of Optimal Scheduling Under Various Simplified Situations

Researcher(s)	Complexity	p	$w(n_i)$	Structure	$c(n_i, n_j)$
Hu [1961]	$O(v)$	—	Uniform	Free-tree	NIL
Coffman and Graham [1972]	$O(v^2)$	2	Uniform	—	NIL
Sethi [1976]	$O(v\alpha(v) + e)$	2	Uniform	—	NIL
Papadimitrious and Yannakakis [1979]	$O(ve)$	—	Uniform	Interval-ordered	NIL
Ali and El-Rewini [1993]	$O(ev)$	—	Uniform (=c)	Interval-ordered	Uniform (=c)
Papadimitrious and Yannakakis [1979]	NP-complete	—	—	Interval-ordered	NIL
Garey and Johnson [1979]	Open	Fixed, >2	Uniform	—	NIL
Ullman [1975]	NP-complete	—	Uniform	—	NIL
Ullman [1975]	NP-complete	Fixed, >1	=1 or 2	—	NIL

3. NP-COMPLETENESS OF THE DAG SCHEDULING PROBLEM

The DAG scheduling problem is in general an NP-complete problem [Garey and Johnson 1979], and algorithms for optimally scheduling a DAG in polynomial-time are known only for three simple cases [Coffman 1976]. The first case is to schedule a uniform node-weight free-tree to an arbitrary number of processors. Hu [1961] proposed a linear-time algorithm to solve the problem. The second case is to schedule an arbitrarily structured DAG with uniform node-weights to two processors. Coffman and Graham [1972] devised a quadratic-time algorithm to solve this problem. Both Hu's algorithm and Coffman et al.'s algorithm are based on node-labeling methods that produce optimal scheduling lists leading to optimal schedules. Sethi [1976] then improved the time-complexity of Coffman et al.'s algorithm to almost linear-time by suggesting a more efficient node-labeling process. The third case is to schedule an interval-ordered DAG with uniform node-weights to an arbitrary number of processors. Papadimitriou and Yannakakis [1979] designed a linear-time algorithm to tackle the problem. A DAG is called interval-ordered if every two precedence-related nodes can be mapped to two nonoverlapping intervals on the real number line [Fishburn 1985].

In all of the above three cases, communication between tasks is ignored. Ali and El-Rewini [1993] showed that interval-ordered DAG with uniform edge weights, which are equal to the node weights, can also be optimally scheduled in polynomial time. These optimality results are summarized in Table II.

Ullman [1975] showed that scheduling a DAG with unit computation to p processors is NP-complete. He also showed that scheduling a DAG with one or two unit computation costs to two processors is NP-complete [Coffman 1975; Ullman 1975]. Papadimitriou and Yannakakis [1979] showed that scheduling an interval-ordered DAG with arbitrary computation costs to two processors is NP-complete. Garey et al. [1983] showed that scheduling an opposing forest with unit computation to p processors is NP-complete. Finally, Papadimitriou and Yannakakis [1990] showed that scheduling a DAG with unit computation to p processors possibly with task-duplication is also NP-complete.

4. A TAXONOMY OF DAG SCHEDULING ALGORITHMS

To outline the variations of scheduling algorithms and to describe the scope of our survey, we introduce in Figure 2 a taxonomy of static parallel scheduling

[Ahmad et al. 1996; Ahmad et al. 1997]. Note that unlike the taxonomy suggested by Casavant and Kuhl [1988], which describes the general scheduling problem (including partitioning and load balancing issues) in parallel and distributed systems, the focus of our taxonomy is on the static scheduling problem, and therefore is only partial.

The highest level of the taxonomy divides the scheduling problem into two categories, depending upon whether the task graph is of an arbitrary structure or a special structure such as trees. Earlier algorithms have made simplifying assumptions about the task graph representing the program and the model of the parallel processor system [Coffman 1976; Gonzalez Jr. 1977]. Most of these algorithms assume the graph to be of a special structure such as a tree, forks-join, etc. In general, however, parallel programs come in a variety of structures, and as such, many recent algorithms are designed to tackle arbitrary graphs. These algorithms can be further divided into two categories. Some algorithms assume the computational costs of all the tasks to be uniform. Others assume the computational costs of tasks to be arbitrary.

Some of the scheduling algorithms that consider the intertask communication assume the availability of an unlimited number of processors, while some other algorithms assume a limited number of processors. The former class of algorithms are called the UNC (*unbounded number of clusters*) scheduling algorithms [Kim and Browne 1988; Kwok and Ahmad 1996; Sarkar 1989; Wong and Morris 1989; Yang and Gerasoulis 1994] and the latter the BNP (*bounded number of processors*) scheduling algorithms [Adam et al. 1974; Anger et al. 1990; Kim and Yi 1994; Kwok and Ahmad 1997; McCreary and Gill 1989; Palis et al. 1996; Sih and Lee 1993b]. In both classes of algorithms, the processors are assumed to be fully connected, and no attention is paid to link contention or routing strategies used for communication. The technique employed by the UNC algorithms is also called *clustering* [Kim and Browne 1988; Liou and Palis 1996; Palis et al. 1996; Sarkar 1989; Yang and Gerasoulis 1994]. At the beginning of the scheduling process, each node is considered a cluster. In the subsequent steps, two clusters are merged if the merging reduces the completion time. This merging procedure continues until no cluster can be merged. The rationale behind the UNC algorithms is that they can take advantage of using more processors to further reduce the schedule length. However, the clusters generated by the UNC need a postprocessing step for mapping the clusters onto the processors because the number of processors available may be less than the number of clusters. As a result, the final solution quality also highly depends on the cluster-mapping step. On the other hand, the BNP algorithms do not need such a postprocessing step. It is an open question as to whether UNC or BNP is better.

We use the terms *cluster* and *processor* interchangeably since, in the UNC scheduling algorithms, merging a single node cluster to another cluster is analogous to scheduling a node to a processor.

There have been a few algorithms designed with the most general model in that the system is assumed to consist of an arbitrary network topology, of which the links are not contention-free. These algorithms are called the APN (*arbitrary processor network*) scheduling algorithms. In addition to scheduling tasks, the APN algorithms also schedule messages on the network communication links. Scheduling of messages may be dependent on the routing strategy used by the underlying network. To optimize schedule lengths under such unrestricted environments makes the design of an APN scheduling algorithm intricate and challenging.

The TDB (Task-Duplication Based) scheduling algorithms also assume the availability of an unbounded number of processors but schedule tasks with duplication to further reduce the schedule

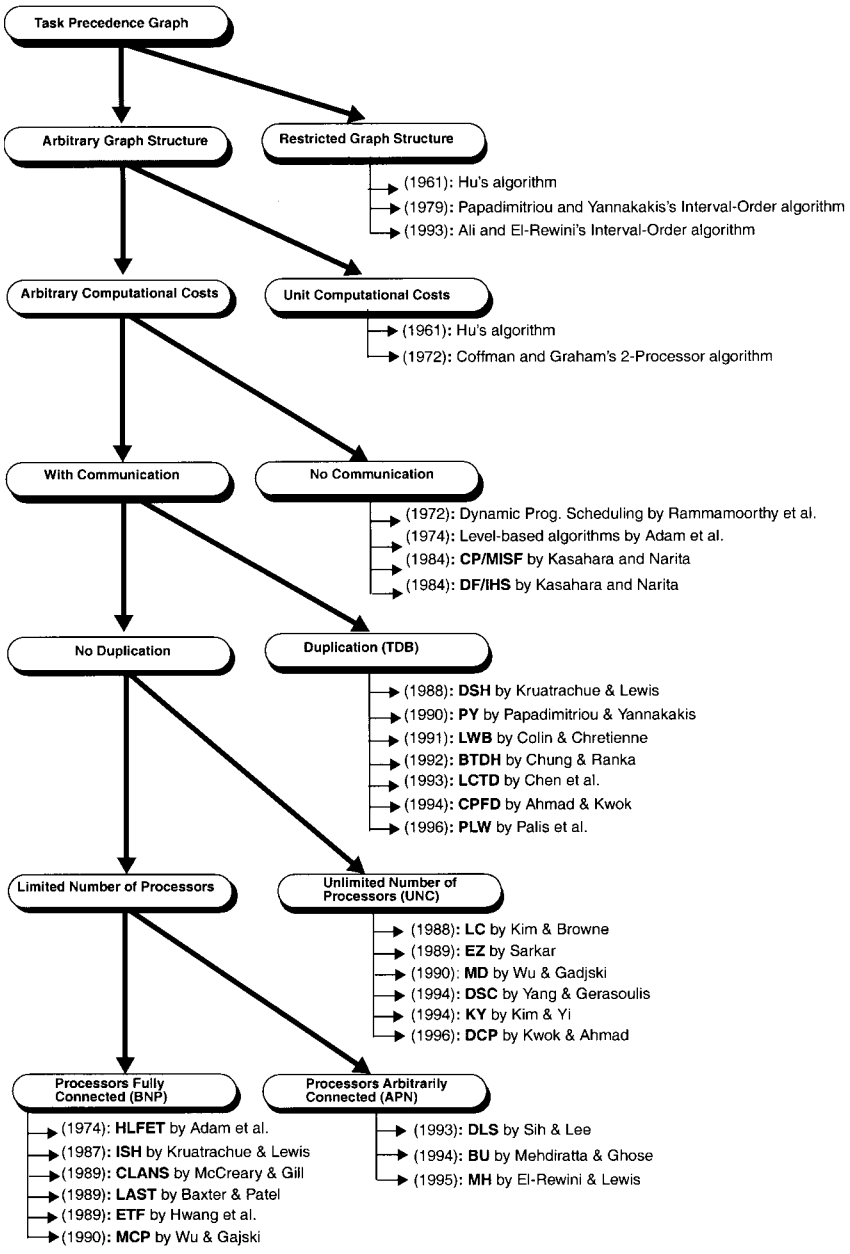


Figure 2. A partial taxonomy of the multiprocessor scheduling problem.

lengths. The rationale behind the TDB scheduling algorithms is to reduce the communication overhead by redundantly allocating some tasks to multiple processors. In duplication-based scheduling, different strategies can be employed to select ancestor nodes for duplication. Some of the algorithms duplicate only the direct predecessors while others try to duplicate all possible ancestors. For a recent quantitative comparison of TDB scheduling algorithms, the reader is referred to Ahmad and Kwok [1999].

5. BASIC TECHNIQUES IN DAG SCHEDULING

Most scheduling algorithms are based on the so-called list scheduling technique [Adam et al. 1974; Ahmad et al. 1996; Casavant and Kuhl 1988; Coffman 1976; El-Rewini et al. 1995; El-Rewini 1994; Gerasoulis and Yang 1992; Khan et al. 1994; Kwok and Ahmad 1997; McCreary et al. 1994; Shirazi et al. 1990; Yang and Miller 1988]. The basic idea of list scheduling is to make a scheduling list (a sequence of nodes for scheduling) by assigning them some priorities, and then repeatedly execute the following two steps until all the nodes in the graph are scheduled:

(1) Remove the first node from the scheduling list;

(2) Allocate the node to a processor which allows the earliest start-time.

There are various ways to determine the priorities of nodes, such as HLF (Highest Level First) [Coffman 1976]; LP (Longest Path) [Coffman 1976]; LPT (Longest Processing Time) [Friesen 1987; Gonzalez, Jr. 1977]; and CP (Critical Path) [Graham et al. 1979].

Recently a number of scheduling algorithms based on a *dynamic* list scheduling approach have been suggested [Kwok and Ahmad 1996; Sih and Lee 1993a; Yang and Gerasoulis 1994]. In a traditional scheduling algorithm, the scheduling list is statically constructed before node allocation begins, and most importantly, the sequencing in the list is not modified. In contrast, after each allocation, these recent algorithms recompute the priorities of all unscheduled nodes, which are then used to rearrange the sequencing of the nodes in the list. Thus, these algorithms essentially employ the following three-step approaches:

(1) Determine new priorities of all unscheduled nodes;

(2) Select the node with the highest priority for scheduling;

(3) Allocate the node to the processor which allows the earliest start-time.

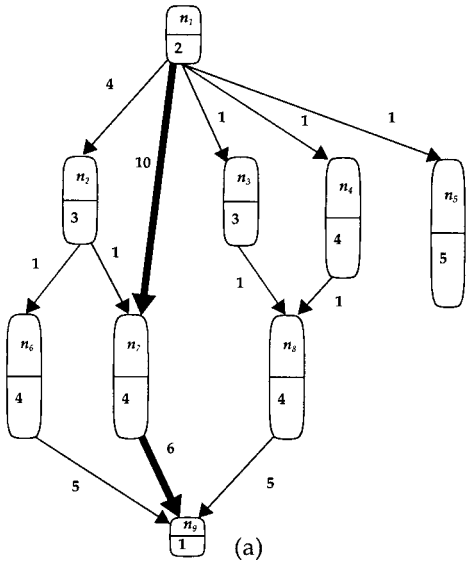
Scheduling algorithms that employ this three-step approach can potentially generate better schedules. However, a dynamic approach can increase the time-complexity of the scheduling algorithm.

Two frequently used attributes for assigning priority are the *t-level* (top level) and *b-level* (bottom level) [Adam et al. 1974; Ahmad et al. 1996; Gerasooulis and Yang 1992]. The *t-level* of a node n_i is the length of a longest path (there can be more than one longest path) from an entry node to n_i (excluding n_i). Here, the length of a path is the sum of all the node and edge weights along the path. As such, the *t-level* n_i highly correlates with n_i's *earliest start-time*, denoted by $T_s(n_i)$, which is determined after n_i is scheduled to a processor. This is because after n_i is scheduled, its $T_s(n_i)$ is simply the length of the longest path reaching it. The *b-level* of a node n_i is the length of a longest path from n_i to an exit node. The *b-level* of a node is bounded from above by the length of a *critical path*. A critical path (CP) of a DAG, which is an important structure in the DAG, is a longest path in the DAG. Clearly, a DAG can have more than one CP. Consider the task graph shown in Figure 3(a). In this task graph, nodes n_i, n_7, and n_8 are the nodes of the only CP and are called CPNs (Critical-Path Nodes). The edges on the CP are shown with thick arrows. The values of the priorities discussed above are shown in Figure 3(b).

Below is a procedure for computing the *t-levels*:

Computing a *t-level*

(1) Construct a list of nodes in topological order. Call it *TopList*.
(2) **for** each node n_i in *TopList* **do**
(3) $max = 0$
(4) **for** each parent n_x of n_i **do**
(5) **if** $t\text{-}level(n_x) + w(n_x) + c(n_x, n_i) > max$ **then**



node	sl	t-level	b-level	ALAP
$*n_1$	11	0	23	0
n_2	8	6	15	8
n_3	8	3	14	9
n_4	9	3	15	8
n_5	5	3	5	18
n_6	5	10	10	13
$*n_7$	5	12	11	12
n_8	5	8	10	13
$*n_9$	1	22	1	22

(b)

Figure 3. (a) A task graph; (b) the static levels (sls), t-levels, b-levels and ALAPs of the nodes.

(6) max=*t-level*(n_x) + $w(n_x)$ + $c(n_x, n_i)$
(7) **endif**
(8) **endfor**
(9) *t-level*(n_i) = max
(10)**endfor**

The time-complexity of the above procedure is $O(e + v)$. A similar procedure, which also has time-complexity $O(e + v)$, for computing the *b-levels* is shown below:

Computing a *b-level*

(1) Construct a list of nodes in reversed topological order. Call it *RevTopList*.
(2) **for** each node n_i in *RevTopList* **do**
(3) max = 0
(4) **for** each child n_y of n_i **do**
(5) **if** $c(n_i, n_y)$ + *b-level*(n_y) > max **then**
(6) max = $c(n_i, n_y)$ + *b-level*(n_y)
(7) **endif**
(8) **endfor**
(9) *b-level*(n_i) = $w(n_i)$ + max
(10) **endfor**

In the scheduling process, the *t-level* of a node varies while the *b-level* is usually a constant, until the node has been scheduled. The *t-level* varies because the weight of an edge may be zeroed when the two incident nodes are scheduled to the same processor. Thus, the path reaching a node, whose length determines the *t-level* of the node, may cease to be the longest one. On the other hand, there are some variations in the computation of the *b-level* of a node. Most algorithms examine a node for scheduling only after all the parents of the node have been scheduled. In this case, the *b-level* of a node is a constant until after it is scheduled to a processor. Some scheduling algorithms allow the scheduling of a child before its parents, however, in which case the *b-level* of a node is also a dynamic attribute. It should be noted that some scheduling algorithms do not take into account the edge weights in computing the *b-level*. In such a case, the *b-level* does not change throughout the scheduling process. To distinguish this definition of *b-level* from the one we described above, we call it the *static b-level* or simply *static level* (sl).

Different algorithms use the *t-level* and *b-level* in different ways. Some algo-

rithms assign a higher priority to a node with a smaller *t-level* while some algorithms assign a higher priority to a node with a larger *b-level*. Still some algorithms assign a higher priority to a node with a larger (*b-level* — *t-level*). In general, scheduling in a descending order of *b-level* tends to schedule critical path nodes first, while scheduling in an ascending order of *t-level* tends to schedule nodes in a topological order. The composite attribute (*b-level—t-level*) is a compromise between the previous two cases. If an algorithm uses a static attribute, such as *b-level* or *static b-level*, to order nodes for scheduling, it is called a *static* algorithm; otherwise, it is called a *dynamic* algorithm.

Note that the procedure for computing the *t-levels* can also be used to compute the start-times of nodes on processors during the scheduling process. Indeed, some researchers call the *t-level* of a node the *ASAP* (As-Soon-As-Possible) start-time because the *t-level* is the earliest possible start-time.

Some of the DAG scheduling algorithms employ an attribute called *ALAP* (As-Late-As-Possible) start-time [Kwok and Ahmad 1996; Wu and Gajski 1990]. The ALAP start-time of a node is a measure of how far the node's start-time can be delayed without increasing the schedule length. An $O(e + v)$ time procedure for computing the ALAP time is shown below:

Computing ALAP

(1) Construct a list of nodes in reversed topological order. Call it *RevTopList*.
(2) **for** each node n_i in *RevTopList* **do**
(3) $min\_ft = CP\_Length$
(4) **for** each child n_y of n_y **do**
(5) **if** $alap(n_y) - c(n_i, n_y) < min\_ft$ **then**
(6) $min\_ft = alap(n_y) - c(n_i, n_y)$
(7) **endif**
(8) **endfor**
(9) $alap(n_i) = min\_ft - w(n_i)$
(10) **endfor**

After the scheduling list is constructed by using the node priorities, the nodes are then scheduled to suitable processors. Usually a processor is considered suitable if it allows the earliest start-time for the node. However, in some sophisticated scheduling heuristics, a suitable processor may not be the one that allows the earliest start-time. These variations are described in detail in Section 6.

6. DESCRIPTION OF THE ALGORITHMS

In this section, we briefly survey algorithms for DAG scheduling reported in the literature. We first describe some of the earlier scheduling algorithms that assume a restricted DAG model, and then proceed to describe a number of such algorithms before proceeding to algorithms that remove all such simplifying assumptions. The performance of these algorithms on some primitive graph structures is also discussed. Analytical performance bounds reported in the literature are also briefly surveyed where appropriate. We first discuss the UNC class of algorithms, followed by BNP algorithms and TDB algorithms. Next we describe a few of the relatively unexplored APN class of DAG scheduling algorithms. Finally, we discuss the issues of scheduling in heterogeneous environments and the mapping problem.

6.1 Scheduling DAGs with Restricted Structures

Early scheduling algorithms were typically designed with simplifying assumptions about the DAG and processor network model [Adam et al. 1974; Bruno et al. 1974; Fujii et al. 1969; Gabow 1982]. For instance, the nodes in the DAG were assumed to be of unit computation and communication was not considered; that is, $w(n_i) = 1$, $\forall i$, and $c(n_i, n_j) = 0$. Furthermore, some algorithms were designed for specially structured DAGs such as a free-tree [Coffman 1976; Hu 1961].

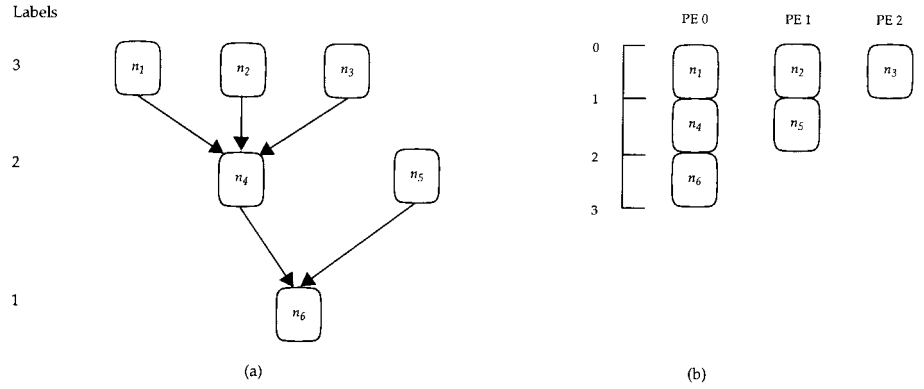


Figure 4. (a) A simple tree-structured task graph with unit-cost tasks and without communication among tasks; (b) the optimal schedule of the task graph using three processors.

6.1.1 *Hu's Algorithm for Tree-Structured DAGs*. Hu [1961] proposed a polynomial-time algorithm to construct optimal schedules for in-tree structured DAGs with unit computations and without communication. The number of processors is assumed to be limited and is equal to p. The crucial step in the algorithm is a node labelling process. Each node n_i is labelled α_1 where $\alpha_1 = x_i + 1$ and x_i is the length of the path from n_i to the exit node in the DAG. Here, the notion of *length* is the number of edges in the path. The labelling process begins with the exit node, which is labelled 1.

Using the above labelling procedure, an optimal schedule can be obtained for p processors by processing a tree-structured task graph in the following steps:

(1) Schedule the first p (or fewer) nodes with the highest numbered label, i.e., the entry nodes, to the processors. If the number of entry nodes is greater than p, choose p nodes whose α_i is greater than the others. In case of a tie, choose a node arbitrarily.

(2) Remove the p scheduled nodes from the graph. Treat the nodes with no predecessor as the new entry nodes.

(3) Repeat steps (1) and (2) until all nodes are scheduled.

The labelling process of the algorithm partitions the task graph into a number of levels. In the scheduling process, each level of tasks is assigned to the available processors. Schedules generated using the above steps are optimal under the stated constraints. The readers are referred to Hu [1961] for the proof of optimality. This is illustrated in the simple task graph and its optimal schedule shown in Figure 4. The complexity of the algorithm is linear in terms of the number of nodes because each node in the task graph is visited a constant number of times.

Kaufman [1974] devised an algorithm for preemptive scheduling that also works on an in-tree DAG with *arbitrary* computation costs. The algorithm is based on principles similar to those in Hu's algorithm. The main idea of the algorithm is to break down the non-unit-weighted tasks into unit-weighted tasks. Optimal schedules can be obtained since the resulting DAG is still an in-tree.

6.1.2 *Coffman and Graham's Algorithm for Two-Processor Scheduling*. Optimal static scheduling have also been addressed by Coffman and Graham

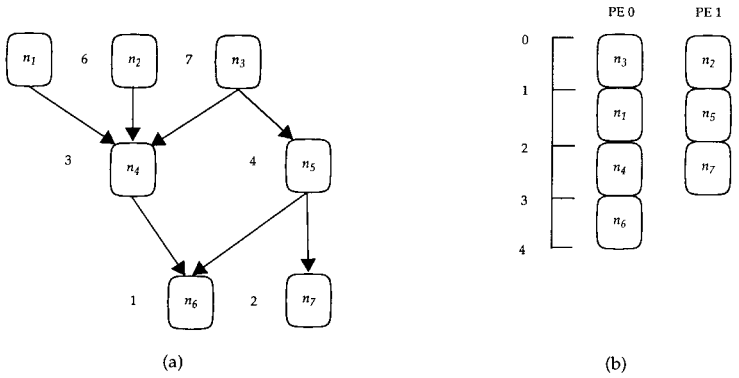


Figure 5. (a) A simple task graph with unit-cost tasks and no-cost communication edges; (b) the optimal schedule of the task graph in a two-processor system..

[1972]. They developed an algorithm for generating optimal schedules for arbitrary structured task graphs with unit-weighted tasks and zero-weighted edges to a two-processor system. The algorithm works on similar principles as in Hu's algorithm. The algorithm first assigns labels to each node in the task graph. The assignment process proceeds "up the graph" in a way that considers as candidates for the assignment of the next label all the nodes whose successors have already been assigned a label. After all the nodes are assigned a label, a list is formed by ordering the tasks in decreasing label numbers, beginning with the last label assigned. The optimal schedule is then obtained by scheduling ready tasks in this list to idle processors. This is elaborated in the following steps.

(1) Assign label 1 to one of the exit node.

(2) Assume that labels 1, 2, . . . , $j - 1$ have been assigned. Let S be the set of unassigned nodes with no unlabeled successors. Select an element of S to be assigned label j as follows. For each node x in S, let $y_1, y_2, \ldots, y_k$ be the immediate successors of x. Then, define $l(x)$ to be the decreasing sequence of integers formed by ordering the set of y 's labels. Suppose that $l(x) \leq l(x')$ lexicographically for all x' in S. Assign the label j to x.

(3) After all tasks have been labeled, use the list of tasks in descending order of labels for scheduling. Beginning from the first task in the list, schedule each task to one of the two given processors that allows the earlier execution of the task.

Schedules generated using the above algorithm are optimal under the given constraints. For the proof of optimality, the reader is referred to Coffman and Graham [1972]. An example is illustrated in Figure 5. Through counter-examples, Coffman and Graham also demonstrated that their algorithm can generate sub-optimal solutions when the number of processors is increased to three or more, or when the number of processors is two and tasks are allowed to have arbitrary computation costs. This is true even when computation costs are allowed to be one or two units. The complexity of the algorithm is $O(v^2)$ because the labelling process and the scheduling process each takes $O(v^2)$ time.

Sethi [1976] reported an algorithm to

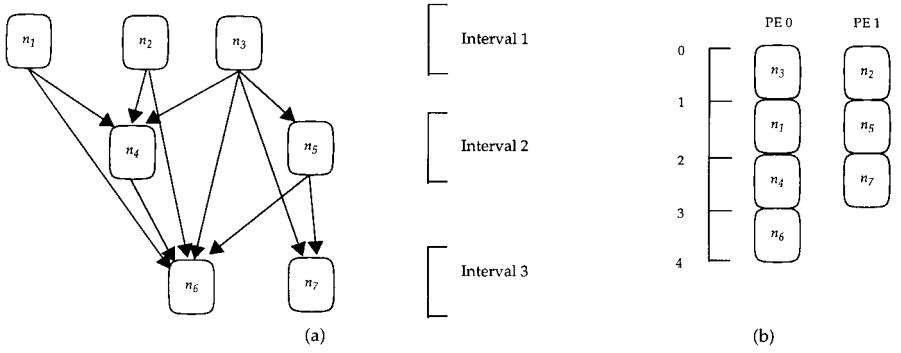


Figure 6. (a) A unit-computational interval ordered DAG; (b) an optimal schedule of the DAG.

determine the labels in $O(v + e)$ time and also gave an algorithm to construct a schedule from the labeling in $O(v\alpha(v) + e)$ time, where $\alpha(v)$ is an almost constant function of v. The main idea of the improved labeling process is based on the observation that the labels of a set of nodes with the same height only depend on their children. Thus, instead of constructing the lexicographic ordering information from scratch, the labeling process can infer such information through visiting the edges connecting the nodes and their children. As a result, the time-complexity of the labeling process is $O(v + e)$ instead of $O(v^2)$. The construction of the final schedule is done with the aid of a set data structure, for which v access operations can be performed in $O(v\alpha(v))$, where $\alpha(v)$ is the inverse Ackermann's function.

6.1.3 *Scheduling Interval-Ordered DAGs*. Papadimitriou and Yannakakis [1979] investigated the problem of scheduling unit-computational interval-ordered tasks to multiprocessors. In an interval-ordered DAG, two nodes are precedence-related if and only if the nodes can be mapped to non-overlapping intervals on the real line [Fishburn 1985]. An example of an interval-ordered DAG is shown in Figure 6. Based on the interval-ordered property, the number of successors of a node can be used as a priority to construct a list. An optimal list schedule can be constructed in $O(v + e)$ time. However, as mentioned earlier, the problem becomes NP-complete if the DAG is allowed to have arbitrary computation costs. Ali and El-Rewini [1993] worked on the problem of scheduling interval-ordered DAGs with unit computation costs and unit communication costs. They showed that the problem is tractable and devised an $O(ve)$ algorithm to generate optimal schedules. In their algorithm, which is similar to that of Papadimitriou and Yannakakis [1979], the number of successors is used as a node priority for scheduling.

6.2 Scheduling Arbitrary DAGs Without Communication

In this section, we discuss algorithms for scheduling arbitrary structured DAGs in which computation costs are arbitrary but communication costs are zero.

6.2.1 *Level-based Heuristics*. Adam et al. [1974] performed an extensive simulation study of the performance of a number of level-based list scheduling heuristics. The heuristics examined are:

- HLFET (Highest Level First with Estimated Times): The notion of level is the sum of computation costs of all

the nodes along the longest path from the node to an exit node.

- HLFNET (Highest Levels First with No Estimated Times): In this heuristic, all nodes are scheduled as if they were of unit cost.
- Random: The nodes in the DAG are assigned priorities randomly.
- SCFET (Smallest Co-levels First with Estimated Times): A co-level of a node is determined by computing the sum of the longest path from an entry node to the node.
- A node has a higher priority if it has the smaller co-level.
- SCFNET (Smallest Co-levels First with No Estimated Times): This heuristic is the same as SCFET except that it schedules the nodes as if they were of unit costs.

In Adam et al. [1974], an extensive simulation study was conducted using randomly generated DAGs. The performance of the heuristics were ranked in the following order: HLFET, HLFNET, SCFNET, Random, and SCFET. The study provided strong evidence that the CP (critical path) based algorithms have near-optimal performance. In another study conducted by Kohler [1975], the performance of the CP-based algorithms improved as the number of processors increased.

Kasahara et al. [1984] proposed an algorithm called CP/MISF (critical path/most immediate successors first), which is a variation of the HLFET algorithm. The major improvement of CP/MISF over HLFET is that when assigning priorities, ties are broken by selecting the node with a larger number of immediate successors.

In a recent study, Shirazi et al. [1990] proposed two algorithms for scheduling DAGs to multiprocessors without communication. The first algorithm, called HNF (Heavy Node First), is based on a simple local analysis of the DAG nodes at each level. The second algorithm, WL (Weighted Length), considers a global view of a DAG by taking into account the relationship among the nodes at different levels. Compared to a critical-path-based algorithm, Shirazi et al. showed that the HNF algorithm is more preferable for its low complexity and good performance.

6.2.2 *A Branch-and-Bound Approach*. In addition to CP/MISF, Kasahara et al. [1984] also reported a scheduling algorithm based on a branch-and-bound approach. Using Kohler and Steiglitz's [1974] general representation for branch-and-bound algorithms, Kasahara et al. devised a depth-first search procedure to construct near-optimal schedules. Prior to the depth-first search process, priorities are assigned to those nodes in the DAG which may be generated during the search process. The priorities are determined using the priority list of the CP/MISF method. In this way the search procedure can be more efficient both in terms of computing time and memory requirement. Since the search technique is augmented by a heuristic priority assignment method, the algorithm is called DF/IHS (depth-first with implicit heuristic search). The DF/IHS algorithm was shown to give near optimal performance.

6.2.3 *Analytical Performance Bounds for Scheduling without Communication*. Graham [1966] proposed a bound on the schedule length obtained by general list scheduling methods. Using a level-based method for generating a list for scheduling, the schedule length SL and the optimal schedule length SL_{opt} are related by the following:

$$SL \leq \left(2 - \frac{1}{p}\right) SL_{opt.}$$

Rammamoorthy et al. [1972] used the concept of precedence partitions to generate bounds on the schedule length and the number of processors for DAGs with unit computation costs. An earliest pre-

cedence partition E_i is a set of nodes that can be started in parallel at the same earliest possible time constrained by the precedence relations. A latest precedence partition is a set of nodes which must be executed at the same latest possible time constrained by the precedence relations. For any i and j, $E_i \cap E_j = \emptyset$ and $L_i \cap L_j = \emptyset$. The precedence partitions group tasks into subsets to indicate the earliest and latest times during which tasks can be started and still guarantee minimum execution time for the graph. This time is given by the number of partitions and is a measure of the longest path in the graph. For a graph of l levels, the minimum execution time is l units. In order to execute a graph in the minimum time, the absolute minimum number of processors required is given by $\max_{1 \le i \le l}\{|E_i \cap L_i|\}$.

Rammamoorthy et al. [1972] also developed algorithms to determine the minimum number of processors required to process a graph in the least possible amount of time, and to determine the minimum time necessary to process a task graph given k processors. Since a dynamic programming approach is employed, the computational time required to obtain the optimal solution is quite considerable.

Fernandez and Bussell [1983] devised improved bounds on the minimum number of processors required to achieve the optimal schedule length and on the minimum increase in schedule length if only a certain number of processors are available. The most important contribution is that the DAG is assumed to have unequal computational costs. Although for such a general model similar partitions as in Rammamoorthy et al. 's work could be defined, Fernandez et al. [Fernandez and Bussell 1983] used the concepts of activity and load density, defined below.

Definition 1. The activity of a node n_i is defined as

$$f(\tau_i, t) = \begin{cases} 1, & t \in [\tau_i - w(n_i), \tau_i], \\ 0, & \text{otherwise} \end{cases}$$

where τ_i is the finish n_i.

Definition 2. The load density function is defined by: $F(\tau, t) = \Sigma_{i=1}^{v} f(\tau_i, t)$.

Then, $f(\tau_i, t)$ indicates the activity of node n_i along time, according to the precedence constraints in the DAG, and $F(\tau, t)$ indicates the total activity of the graph as a function of time. Of particular importance are $F(\tau_e, t)$, the earliest load density function for which all tasks are completed at their earliest times, and $F(\tau_l, t)$, the load density function for which all tasks are completed at their latest times. Now let $R(\theta_1, \theta_2, t)$ be the load density function of the tasks or parts of tasks remaining within $[\theta_1, \theta_2]$ after all tasks have been shifted to form minimum overlap within the interval. Thus, a lower bound on the minimum number of processors to execute the program (represented by the DAG) within the minimum time is given by:

$$p_{min} = \left\lceil \max_{[\theta_1, \theta_2]} \left[\frac{1}{\theta_2 - \theta_1} \int_{\theta_1}^{\theta_2} R(\theta_1, \theta_2, t)dt \right] \right\rceil.$$

The maximum value obtained for all possible intervals indicate that the whole computation graph cannot be executed with a number of processors smaller than the maximum. Supposing that only p' processors are available, Fernandez and Bussell [1973] also showed that the schedule length will be longer than the optimal schedule length by no less than the following amount:

$$\left\lceil \max_{w(n_1) \le w(n_k) \le CP} \left[-w(n_k) + \frac{1}{p'} \int_{0}^{w(n_k)} F(\tau_l, t)dt \right] \right\rceil.$$

In a recent study, Jain and Rajaraman [1994] reported sharper bounds using the above expressions. The idea is that the intervals considered for the integration is not just the earliest and

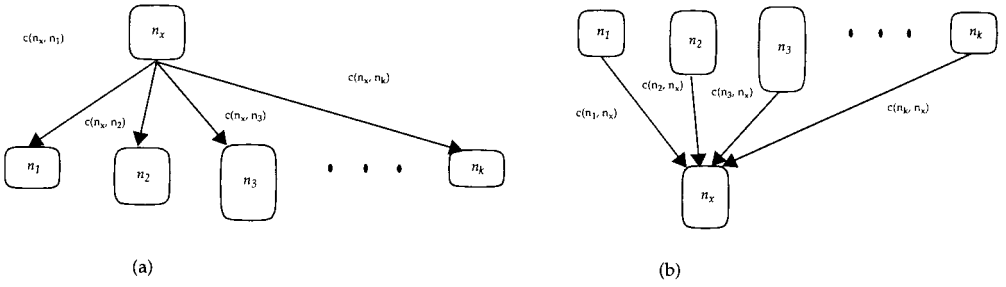


Figure 7. (a) A fork set; and (b) a join set.

latest start-times but are based on a partitioning of the graphs into a set of disjoint sections. They also devised an upper bound on the schedule length, which is useful in determining the worst case behavior of a DAG. Not only are their new bounds easier to compute but are also tighter, because DAG partitioning strategy enhances the accuracy of the load activity function.

6.3 UNC Scheduling

In this section, we survey the UNC class of scheduling algorithms. In particular, we will describe in more details five UNC scheduling algorithms: the EZ, LC, DSC, MD, and DCP algorithms. The DAG shown in Figure 3 is used to illustrate the scheduling process of these algorithms. In order to examine the approximate optimality of the algorithms, we will first describe the scheduling of two primitive DAG structures: the *fork* set and the *join* set. Some work on theoretical performance analysis of UNC scheduling is also discussed in the last subsection.

6.3.1 *Scheduling of Primitive Graph Structures*. To highlight the different characteristics of the algorithms described below, it is useful to consider how the algorithms work on some primitive graph structures. Two commonly used primitive graph structures are *fork* and *join* [Gerasoulis and Yang 1992], examples of which are shown in Figure 7. These two graph primitives are useful for understanding the optimality of scheduling algorithms because any task graph can be decomposed into a collection of forks and joins. In the following, we derive the optimal schedule lengths for these primitive structures. The optimal schedule lengths can then be used as a basis for comparing the functionality of the scheduling algorithms described later in this section.

Without loss of generality, assume that for the fork structure, we have:

$$c(n_x, n_1) + w(n_1) \geq c(n_x, n_2) + w(n_2) \geq \ldots \geq c(n_x, n_k) + w(n_k).$$

Then the optimal schedule length is equal to:

$$\max\left\{w(n_x) + \sum_{i=1}^{j} w(n_i),\ w(n_x) + c(n_x, n_{j+1}) + w(n_{j+1})\right\},$$

where j is given by the following conditions:

$$\sum_{i=1}^{j} w(n_i) \leq c(n_x, n_j) + w(n_j)$$

and

$$\sum_{i=1}^{j+1} w(n_i) > c(n_x, n_{j+1}) + w(n_{j+1}).$$

In addition, assume that for the join structure, we have:

$$w(n_1) + c(n_1, n_x) \geq w(n_2) + c(n_2, n_x) \geq \ldots \geq w(n_k) + c(n_k, n_x).$$

Then the optimal schedule length for the join is equal to:

$$\max\left\{\sum_{i=1}^{j} w(n_i) + w(n_x), w(n_{j+1}) + c(n_{j+1}, n_x) + w(n_x)\right\},$$

where j is given by the following conditions:

$$\sum_{i=1}^{j} w(n_i) \leq w(n_j) + c(n_j, n_x)$$

and

$$\sum_{i=1}^{j+1} w(n_i) > w(n_{j+1}) + c(n_{j+1}, n_x).$$

From the above expressions, it is clear that an algorithm has to be able to recognize the longest path in the graph in order to generate optimal schedules. Thus, algorithms which consider only *b-level* or only *t-level* cannot guarantee optimal solutions. To make proper scheduling decisions, an algorithm has to dynamically examine both *b-level* and *t-level*. In the coming subsections, we will discuss the performance of the algorithms on these two primitive graph structures.

6.3.2 *The EZ Algorithm*. The EZ (Edge-zeroing) algorithm [Sarkar 1989] selects clusters for merging based on edge weights. At each step, the algorithm finds the edge with the largest weight. The two clusters incident by the edge will be merged if the merging (thereby zeroing the largest weight) does not increase the completion time. After two clusters are merged, the ordering of nodes in the resulting cluster is based on the static *b-levels* of the nodes. The algorithm is briefly described below.

(1) Sort the edges of the DAG in a descending order of edge weights.

(2) Initially all edges are *unexamined*.
Repeat

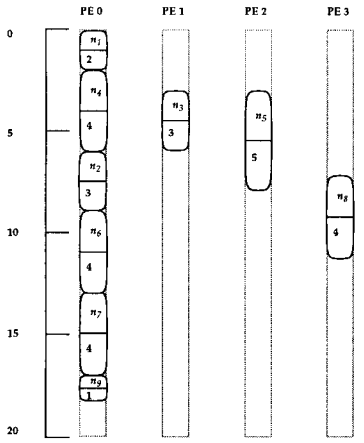
(3) Pick an *unexamined* edge which has the largest edge weight. Mark it as *examined*. Zero the highest edge weight if the completion time does not increase. In this zeroing process, two clusters are merged so that other edges across these two clusters also need to be zeroed and marked as *examined*. The ordering of nodes in the resulting cluster is based on their static *b-levels*.
Until all edges are *examined*.

The time-complexity of the EZ algorithm is $O(e(e + v))$. For the DAG shown in Figure 3, the EZ algorithm generates a schedule shown in Figure 8(a). The steps of scheduling are shown in Figure 8(b).

Performance on fork and join: Since the EZ algorithm considers only the communication costs among nodes to make scheduling decisions, it does not guarantee optimal schedules for both fork and join structures.

6.3.3 *The LC Algorithm*. The LC (Linear Clustering) algorithm [Kim and Browne 1988] merges nodes to form a single cluster based on the CP. The algorithm first determines the set of nodes constituting the CP, then schedules all the CP nodes to a single processor at once. These nodes and all edges incident on them are then removed from the DAG. The algorithm is briefly described below.

(1) Initially, mark all edges as *unexamined*.
Repeat



(a)

(b)

Step	Edge zeroed
1	$(n_1 - n_7)$
2	$(n_7 - n_9)$
3	$(n_6 - n_9)$
4	$(n_8 - n_9)$
5	$(n_1 - n_2)$
6	$(n_1 - n_5)$
7	$(n_1 - n_5)$
8	$(n_1 - n_4)$
9	$(n_1 - n_3)$

Figure 8. (a) The schedule generated by the EZ algorithm (schedule length = 18); (b) a scheduling trace of the EZ algorithm.

(2) Determine the critical path composed of *unexamined* edges only.

(3) Create a cluster by zeroing all the edges on the critical path.

(4) Mark all the edges incident on the critical path and all the edges incident to the nodes in the cluster as *examined*.

Until all edges are *examined*.

The time-complexity of the LC algorithm is $O(v(e + v))$. For the DAG shown in Figure 3, the LC algorithm generates a schedule shown in Figure 9(a); the scheduling steps are shown in Figure 9(b).

Performance on fork and join: Since the LC algorithm does not schedule nodes on different paths to the same processor, it cannot guarantee optimal solutions for both fork and join structures.

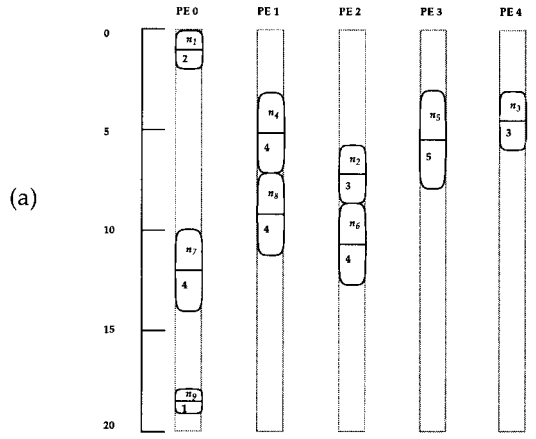
6.3.4 *The DSC Algorithm*. The DSC (Dominant Sequence Clustering) algorithm [Yang and Gerasoulis 1993] considers the Dominant Sequence (DS) of a graph. The DS is the CP of the partially scheduled DAG. The algorithm is briefly described below.

(1) Initially, mark all nodes as unexamined. Initialize a ready node list L to contain all entry nodes. Compute *b-level* for each node. Set *t-level* for each ready node.

Repeat

(2) If the head of L, n_i, is a node on the DS, zeroing the edge between ni and one of its parents so that the *t-level* of n_i is minimized. If no zeroing is accepted, the node remains in a single node cluster.

(3) If the head of L, n_i, is not a node on the DS, zeroing the edge between n_i and one of its parents so that the *t-level* of n_i is minimized under the constraint called Dominant Sequence Reduction Warranty (DSRW). If some of its parents are entry nodes that do not have any child other than n_i, merge part of them so that the *t-level* of n_i is minimized. If no zeroing is accepted, the node remains in a single node cluster.



Step	CP Merged
1	(n_1, n_7, n_9)
2	(n_4, n_8)
3	(n_2, n_6)
4	(n_5)
5	(n_3)

(b)

Figure 9. (a) The schedule generated by the LC algorithm (schedule length = 19); (b) a scheduling trace of the LC algorithm.

(4) Update the *t-level* and *b-level* of the successors of n_i and mark n_i as examined.

Until all nodes are examined.

DSRW: *Zeroing incoming edges of a ready node should not affect the future reduction of t-level* (n_y)*, where* n_y *is a not-yet ready node with a higher priority, if t-level* (n_y) *is reducible by zeroing an incoming DS edge of* n_y.

The time-complexity of the DSC algorithm is $O((e + v)\log v)$. For the DAG shown in Figure 3, the DSC algorithm generates a schedule shown in Figure 10(a). The steps of scheduling are given in the table shown in Figure 10(b). In the table, the start-times of the node on the processors at each scheduling step are given and the node is scheduled to the processor on which the start-time is marked by an asterisk.

Performance on fork and join: The DSC algorithm dynamically tracks the critical path in the DAG using both *t-level* and *b-level*. In addition, it schedules each node to start as early as possible. Thus, for both fork and join structures, the DSC algorithm can guarantee optimal solutions.

Yang and Gerasoulis [1993] also investigated the granularity issue of clustering. They considered that a DAG consists of $fork(F_x)$ and $join(J_x)$ structures such as the two shown in Figure 7. Suppose we have:

$$g(F_x) = \frac{\min\{w(n_i)\}}{\max\{c(n_x, n_i)\}},$$

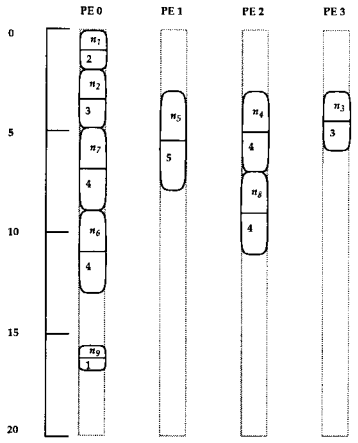
$$g(J_x) = \frac{\min\{w(n_i)\}}{\max\{c(n_i, n_x)\}}$$

Then the granularity of a DAG is defined as $g = min\{g_x\}$ where $g_x = min\{g(F_x), g(J_x)\}$. A DAG is called coarse grain if $g \geq 1$. Based on this definition of granularity, Yang and Gerasoulis proved that the DSC algorithm has the following performance bound:

$$SL_{DSC} \leq \left(1 + \frac{1}{g}\right) SL_{opt}$$

Thus, for a coarse grain DAG, the DSC algorithm can generate a schedule length within a factor of two from the optimal. Yang and Gerasoulis also

(a)



(b)

Step	n_x (prio)	n_y (prio)	Parent	PE0	PE1	PE2	PE3
1	n_1 (23)	NIL	NIL	0*	N.C.	N.C.	N.C.
2	n_2 (21)	n_7 (23)	n_1	2	6*	N.C.	N.C.
3	n_7 (23)	NIL	n_1	5	N.C.	N.C.	N.C.
4	n_5 (8)	n_9 (16)	n_1	9	3*	N.C.	N.C.
5	n_4 (18)	n_9 (16)	n_1	9	N.C.	3*	N.C.
6	n_3 (17)	n_8 (18)	n_1	9	N.C.	N.C.	3*
7	n_6 (20)	n_9 (16)	n_2	9*	N.C.	N.C.	N.C.
8	n_8 (18)	n_9 (19)	n_4	N.C.	N.C.	7*	N.C.
9	n_9 (19)	NIL	n_6	16*	N.C.	N.C.	N.C.

Figure 10. (a) The schedule generated by the DSC algorithm (schedule length = 17); (b) a scheduling trace of the DSC algorithm (N.C. indicates "not considered").

proved that the DSC algorithm is optimal for any coarse grain in-tree, and any single-spawn out-tree with uniform computation costs and uniform communication costs.

6.3.5 *The MD Algorithm*. The MD (Mobility Directed) algorithm [Wu and Gajski 1990] selects a node n_i for scheduling based on an attribute called the relative mobility, defined as:

$$\frac{\text{Cur\_CP\_Length} - (\text{b-level}(n_i) + \text{t-level}(n_i))}{w(n_i)}$$

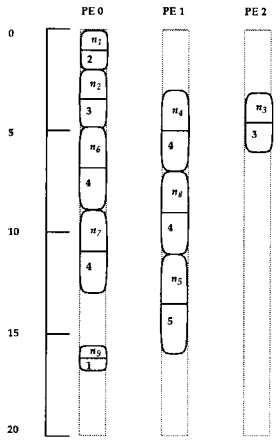
If a node is on the current CP of the partially scheduled DAG, the sum of its *b-level* and *t-level* is equal to the current CP length. Thus, the relative mobility of a node is zero if it is on the current CP. The algorithm is described below.

(1) Mark all nodes as *unexamined*. Initially, there is no cluster.
Repeat

(2) Compute the relative mobility for each node.

(3) Let L' be the group of unexamined nodes with the minimum relative mobility. Let n_i be a node in L' that does not have any predecessors in L'. Start from the first cluster, check whether there is any cluster that can accommodate n_i. In the checking process, all idle time slots in a cluster are examined until one is found to be large enough to hold n_i. A large enough idle time slot may be created by pulling already scheduled nodes downward because

(a)



(b)

Step	n_i	Rel. Mob.	PE0	PE1	PE2	PE3
1	n_1	0.0	0*	N.C.	N.C.	N.C.
2	n_7	0.0	10*	N.C.	N.C.	N.C.
3	n_2	0.0	2*	N.C.	N.C.	N.C.
4	n_4	0.0	N.R.	3*	N.C.	N.C.
5	n_8	0.0	N.R.	7*	N.C.	N.C.
6	n_3	0.0	N.R.	N.R.	3*	N.C.
7	n_9	0.0	16*	N.C.	N.C.	N.C.
8	n_6	0.25	5*	N.C.	N.C.	N.C.
9	n_5	1.8	N.R.	11*	N.C.	N.C.

Figure 11. (a) The schedule generated by the MD algorithm (schedule length = 17); (b) a scheduling trace of the MD algorithm (N.C. indicates "not considered," N.R. indicates "no room").

the start-times of the already scheduled nodes are not fixed yet. If n_i cannot be scheduled to the first cluster, try the second cluster, and so on. If n_i cannot be scheduled to any existing cluster, leave it as a new cluster.

(4) When n_i is scheduled to cluster m, all edges connecting n_i and other nodes already scheduled to cluster m are changed to zero. If n_i is scheduled before node n_j on cluster m, add an edge with weight zero from n_i to n_j in the DAG. If n_i is scheduled after node n_j on the cluster, add an edge with weight zero from n_j to n_i, then check if the adding edges form a loop. If so, schedule n_i to the next available space.

(5) Mark n_i as examined.
Until all nodes are examined.

The time-complexity of the MD algorithm is $O(v^3)$. For the DAG shown in Figure 3, the MD algorithm generates a schedule shown in Figure 11(a). The steps of scheduling are given in the table shown in Figure 11(b). In the table, the start-times of the node on the processors at each scheduling step are given and the node is scheduled to the processor on which the start-time is marked by an asterisk.

Performance on fork and join: Using the notion of relative mobility, the MD algorithm is also able to track the criti-

cal path of the DAG in the scheduling process. Thus, the algorithm can generate optimal schedules for fork and join as well.

6.3.6 *The DCP Algorithm*. The DCP (Dynamic Critical Path) algorithm is proposed by Kwok and Ahmad [1996] and is designed based on an attribute which is slightly different from the relative mobility used in the MD algorithm. Essentially, the DCP algorithm examines a node n_i for scheduling if, among all nodes, n_i has the smallest difference between its ALST (Absolute-Latest-Start-Time) and AEST (Absolute-Earliest-Start-Time). The value of such difference is equivalent to the value of the node's mobility, defined as:

$$(\text{Cur\_CP\_Length} - (\text{b-level}(n_i) + \text{t-level}(n_i))).$$

The DCP algorithm uses a lookahead strategy to find a better cluster for a given node. The DCP algorithm is briefly described below.

Repeat

(1) Compute (Cur_CP_Length − (b-level (n_i) + t-level(n_i))) for each node n_i.

(2) Suppose that n_x is the node with the largest priority. Let n_c be the child node (i.e., the *critical child*) of n_x that has the largest priority.

(3) Select a cluster P such that the sum $T_s(n_x) + (T_x(n_c))$ is the smallest among all the clusters holding n_x's parents or children. In examining a cluster, first try not to pull down any node to create or enlarge an idle time slot. If this is not successful in finding a slot for n_x, scan the cluster for suitable idle time slot again possibly by pulling some already scheduled nodes downward.

(4) Schedule n_x to P.

Until all nodes are scheduled.

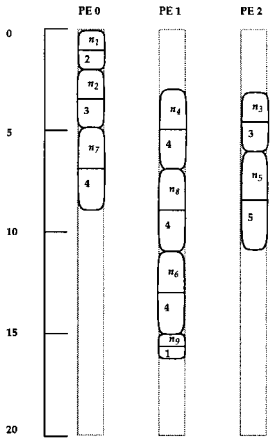
The time-complexity of the DCP algorithm is $O(v^3)$. For the DAG shown in Figure 3, the DCP algorithm generates a schedule shown in Figure 12(a). The steps of scheduling are given in the table shown in Figure 12(b). In the table, the composite start-times of the node (i.e., the start-time of the node plus that of its critical child) on the processors at each scheduling step are given and the node is scheduled to the processor on which the start-time is marked by an asterisk.

Performance on fork and join: Since the DCP algorithm examines the first unscheduled node on the current critical path by using mobility measures, it constructs optimal solutions for fork and join graph structures.

6.3.7 *Other UNC Approaches*. Kim and Yi [1994] proposed a two-pass scheduling algorithm with time-complexity $O(v \log v)$. The idea of the algorithm comes from the scheduling of intrees. Kim and Yi observed that an intree can be efficiently scheduled by iteratively merging a node to the parent node that allows the earliest completion time. To extend this idea to arbitrary structured DAGs, Kim and Yi devised a two-pass algorithm. In the first pass, an independent v-graph is constructed for each exit node and an iterative scheduling process is carried out on the v-graphs. This phase is called *forward-scheduling*. Since some intermediate nodes may be assigned to different processors in different schedules, a *backward-scheduling* phase—the second pass of the algorithm—is needed to resolve the conflicts. In their simulation study, the two-pass algorithm outperformed a simulated annealing approach. Moreover, as the principles of the algorithm originated from scheduling trees, the algorithm is optimal for both fork and join structures.

6.3.8 *Theoretical Analysis for UNC Scheduling*. In addition to the granularity analysis performed for the DSC algorithm, Yang and Gerasoulis [1993] worked on the general analysis for UNC

(a)



(b)

Step	n_i	AEST	ALST	Cri. Child	PE0	PE1	PE2	PE3
1	n_1	0	0	n_7	0+10*	N.C.	N.C.	N.C.
2	n_7	12	12	n_9	10+19*	12+19	N.C.	N.C.
3	n_2	6	6	n_7	2+5*	6+9	N.C.	N.C.
4	n_4	3	3	n_8	N.R.	3+7*	N.C.	N.C.
5	n_8	8	8	n_9	N.C.	7+15*	8+15	N.C.
6	n_3	3	3	n_8	N.R.	N.R.	3+7*	N.C.
7	n_9	16	16	NIL	16+0	15+0*	N.C.	16+0
8	n_6	6	6	n_9	N.R.	11+0*	N.C.	6+15
9	n_5	3	11	NIL	9+0	N.R.	6+0*	N.C.

Figure 12. (a) The schedule generated by the DCP algorithm (schedule length = 16); (b) a scheduling trace of the DCP algorithm (N.C. indicates "not considered," N.R. indicates "no room").

scheduling. They introduced a notion called δ-lopt which is defined below.

Definition 2. Let DL_i^{lopt} be the optimum schedule length at step i of a UNC scheduling algorithm. A UNC scheduling algorithm is called δ-lopt if $max_i\{SL_i - SL_i^{lopt}\} \leq \delta$ where δis a given constant.

In their study, they examined two critical-path-based simple UNC scheduling heuristics called RCP and RCP*. Essentially, both heuristics use *b-level* as the scheduling priority but with a slight difference in that RCP* uses $(b\text{-}level - w(n_i))$ as the priority. They showed that both heuristics are δ-lopt, and thus demonstrated that critical path-based scheduling algorithms are near-optimal.

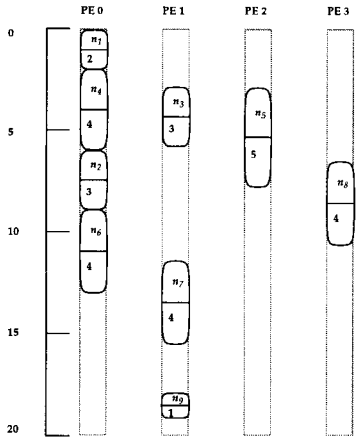
6.4 BNP Scheduling

In this section we survey the BNP class of scheduling algorithms. In particular we discuss in detail six BNP scheduling algorithms: the HLFET, ISH, MCP, ETF, DLS, and LAST algorithms. Again, the DAG shown in Figure 3 is used to illustrate the scheduling process of these algorithms. The analytical performance bounds of BNP scheduling will also be discussed in the last subsection.

6.4.1 *The HLFET Algorithm.* The HLFET (Highest Level First with Estimated Times) algorithm [Adam et al. 1974] is one of the simplest list scheduling algorithms and is described below.

(1) Calculate the static *b-level* (i.e., *sl* or static level) of each node.

(a)



(b)

Step	n_i	PE0	PE1	PE2	PE3
1	n_1	0*	N.C.	N.C.	N.C.
2	n_4	2*	3	N.C.	N.C.
3	n_2	6*	6	N.C.	N.C.
4	n_3	9	3*	N.C.	N.C.
5	n_5	9	6	3*	N.C.
6	n_6	9*	10	10	10
7	n_7	13	12*	12	12
8	n_8	13	16	8	7*
9	n_9	22	18*	22	22

Figure 13. (a) The schedule generated by the HLFET algorithm (schedule length = 19); (b) a scheduling trace of the HLFET algorithm (N.C. indicates "not considered").

(2) Make a ready list in a descending order of *static b-level*. Initially, the ready list contains only the entry nodes. Ties are broken randomly.

Repeat

(3) Schedule the first node in the ready list to a processor that allows the earliest execution, using the non-insertion approach.

(4) Update the ready list by inserting the nodes that are now ready.

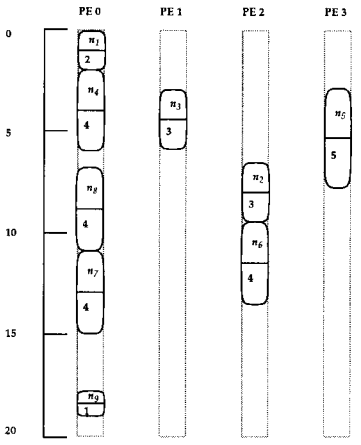
Until all nodes are scheduled.

The time-complexity of the HLFET algorithm is $O(v^2)$. For the DAG shown in Figure 3, the HLFET algorithm generates a schedule shown in Figure 13(a). The steps of scheduling are given in the table shown in Figure 13(b). In the table, the start-times of the node on the processors at each scheduling step are given and the node is scheduled to the processor on which the start-time is marked by an asterisk.

Performance on fork and join: Since the HLFET algorithm schedules nodes based on *b-level* only, it cannot guarantee optimal schedules for fork and join structures even if given sufficient processors.

6.4.2 *The ISH Algorithm*. The ISH (Insertion Scheduling Heuristic) algorithm [Kruatrachue and Lewis 1987] uses the "scheduling holes"—the idle time slots—in the partial schedules. The algorithm tries to fill the holes by scheduling other nodes into them and uses static *b-level* as the priority of a

(a)



(b)

Step	n_i	PE0	PE1	PE2	PE3	Idle Slot	Hole Tasks (start-time)
1	n_1	0*	N.C.	N.C.	N.C.	NIL	NIL
2	n_4	2*	3	N.C.	N.C.	NIL	NIL
3	n_3	6	3*	N.C.	N.C.	[0..3]	n_2 (6), n_5 (3)
4	n_2	6	6	6*	N.C.	[0..6]	n_5 (3), n_8 (7)
5	n_5	6	6	9	3*	[0..3]	n_6 (10), n_7 (12), n_8 (7)
6	n_8	7*	7	9	8	[6..7]	n_6 (10), n_7 (10)
7	n_7	11*	12	12	12	NIL	NIL
8	n_6	15	10	9*	10	NIL	NIL
9	n_9	18*	21	21	21	[15..18]	NIL

Figure 14. (a) The schedule generated by the ISH algorithm (schedule length = 19); (b) a scheduling trace of the ISH algorithm (N.C. indicates "not considered").

node. The algorithm is briefly described below.

(1) Calculate the static *b-level* of each node.

(2) Make a ready list in a descending order of static *b-level*. Initially, the ready list contains only the entry nodes. Ties are broken randomly.

Repeat

(3) Schedule the first node in the ready list to the processor that allows the earliest execution, using the non-insertion algorithm.

(4) If scheduling of this node causes an idle time slot, then find as many nodes as possible from the ready list that can be scheduled to the idle time slot but cannot be scheduled earlier on other processors.

(5) Update the ready list by inserting the nodes that are now ready.

Until all nodes are scheduled.

The time-complexity of the ISH algorithm is $O(v^2)$. For the DAG shown in Figure 3, the ISH algorithm generates a schedule shown in Figure 14(a). The steps of scheduling are given in the table shown in Figure 14(b). In the table, the start-times of the node on the processors at each scheduling step are given and the node is scheduled to the processor on which the start-time is marked by an asterisk. Hole tasks are the nodes considered for scheduling into the idle time slots.

Performance on fork and join: Since the ISH algorithm schedules nodes based on b-level only, it cannot guarantee optimal schedules for fork and join

structures even if given sufficient processors.

6.4.3 *The MCP Algorithm*. The MCP (Modified Critical Path) algorithm [Wu and Gajski 1990] uses the ALAP of a node as the scheduling priority. The MCP algorithm first computes the ALAPs of all the nodes, then constructs a list of nodes in an ascending order of ALAP times. Ties are broken by considering the ALAP times of the children of a node. The MCP algorithm then schedules the nodes on the list one by one such that a node is scheduled to a processor that allows the earliest start-time using the insertion approach. The MCP algorithm and the ISH algorithm have different philosophies in utilizing the idle time slot: MCP looks for an idle time slot for a given node, while ISH looks for a hole node to fit in a given idle time slot. The algorithm is briefly described below.

(1) Compute the ALAP time of each node.

(2) For each node, create a list which consists of the ALAP times of the node itself and all its children in a descending order.

(3) Sort these lists in an ascending lexicographical order. Create a node list according to this order.
Repeat

(4) Schedule the first node in the node list to a processor that allows the earliest execution, using the insertion approach.

(5) Remove the node from the node list.
Until the node list is empty.

The time-complexity of the MCP algorithm is $O(v^2 \log v)$. For the DAG shown in Figure 3, the MCP algorithm generates a schedule shown in Figure 15(a). The steps of scheduling are given in the table shown in Figure 15(b). In the table, the start-times of the node on the processors at each scheduling step are given and the node is scheduled to the processor on which the start-time is marked by an asterisk.

Performance on fork and join: Since the MCP algorithm schedules nodes based on ALAP (effectively based on *b-level*) only, it cannot guarantee optimal schedules for fork and join structures even if given sufficient processors.

6.4.4 *The ETF Algorithm*. The ETF (Earliest Time First) algorithm [Hwang et al. 1989] computes, at each step, the earliest start-times for all ready nodes and then selects the one with the smallest start-time. Here, the earliest start-time of a node is computed by examining the start-time of the node on all processors exhaustively. When two nodes have the same value in their earliest start-times, the ETF algorithm breaks the tie by scheduling the one with the higher static priority. The algorithm is described below.

(1) Compute the static *b-level* of each node.

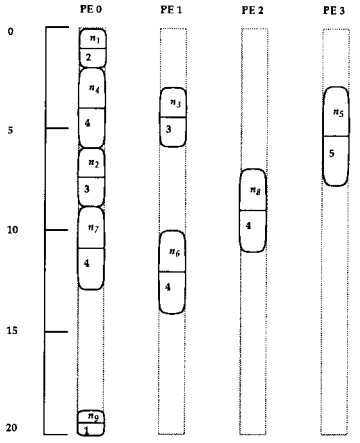
(2) Initially, the pool of ready nodes includes only the entry nodes.
Repeat

(3) Calculate the earliest start-time on each processor for each node in the ready pool. Pick the node-processor pair that gives the earliest time using the non-insertion approach. Ties are broken by selecting the node with a higher static *b-level*. Schedule the node to the corresponding processor.

(4) Add the newly ready nodes to the ready node pool.
Until all nodes are scheduled.

The time-complexity of the ETF algorithm is $O(pv^2)$. For the DAG shown in Figure 3, the ETF algorithm generates a schedule shown in Figure 16(a). The scheduling steps are given in the table shown in Figure 16(b). In the table, the start-times of the node on the processors at each scheduling step are given and the node is scheduled to the proces-

(a)



(b)

Step	n_i	PE0	PE1	PE2	PE3
1	n_1	0*	N.C.	N.C.	N.C.
2	n_4	2*	3	N.C.	N.C.
3	n_2	6*	6	N.C.	N.C.
4	n_3	9	3*	N.C.	N.C.
5	n_7	9*	12	12	N.C.
6	n_6	13	10*	10	N.C.
7	n_8	13	14	7*	N.C.
8	n_5	13	14	11	3*
9	n_9	19*	19	19	19

Figure 15. (a) The schedule generated by the MCP algorithm (schedule length = 20); (b) a scheduling trace of the MCP algorithm (N.C. indicates "not considered").

sor on which the start-time is marked by an asterisk.

Performance on fork and join: Since the ETF algorithm schedules nodes based on *b-level* only, it cannot guarantee optimal schedules for fork and join structures even if given sufficient processors.

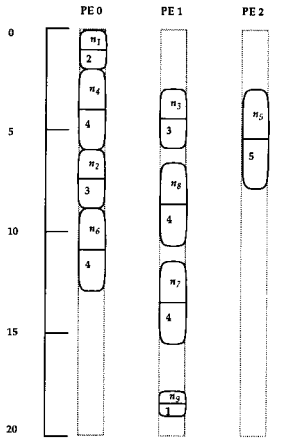
Hwang et al. [1989] also analyzed the performance bound of the ETF algorithm. They showed that the schedule length produced by the ETF algorithm SL_{EFT} satisfies the following relation:

$$SL_{EFT} \leq \left(2 - \frac{1}{p}\right) SL_{opt}^{nc} + C,$$

where SL_{opt}^{nc} is the optimal schedule length without considering communication delays and C is the communication requirements over some parent-parent pairs along a path. An algorithm is also provided to compute C.

6.4.5 *The DLS Algorithm*. The DLS (Dynamic Level Scheduling) algorithm [Sih and Lee 1993a] uses an attribute called *dynamic level (DL)*, which is the difference between the *static b-level* of a node and its earliest start-time on a processor. At each scheduling step, the algorithm computes the DL for every node in the ready pool on all processors. The node-processor pair which gives the largest value of DL is selected for scheduling. This mechanism is similar to the one used by the ETF algorithm. However, there is one subtle difference between the ETF algorithm and the DLS algorithm: the ETF algorithm always schedules the node with the minimum

(a)



(b)

Step	n_i	PE0	PE1	PE2	PE3
1	n_1	0*	N.C.	N.C.	N.C.
2	n_4	2*	3	N.C.	N.C.
3	n_3	6	3*	N.C.	N.C.
4	n_5	6	6	3*	N.C.
5	n_2	6*	6	8	N.C.
6	n_8	9	7*	8	N.C.
7	n_6	9*	11	10	N.C.
8	n_7	13	12*	12	N.C.
9	n_9	22	18*	22	N.C.

Figure 16. (a) The schedule generated by the ETF algorithm (schedule length = 19); (b) a scheduling trace of the ETF algorithm (N.C. indicates "not considered").

earliest start-time and uses *static b-level* merely to break ties. In contrast, the DLS algorithm tends to schedule nodes in a descending order of *static b-levels* at the beginning of a scheduling process but tends to schedule nodes in an ascending order of *t-levels* (i.e., the earliest start-times) near the end of the scheduling process.The algorithm is briefly described below.

(1) Calculate the *b-level* of each node.

(2) Initially, the ready node pool includes only the entry nodes.
Repeat

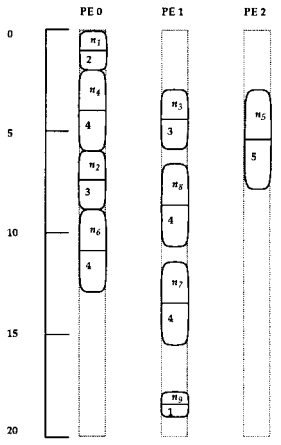
(3) Calculate the earliest start-time for every ready node on each processor. Hence, compute the DL of every node-processor pair by subtracting the earliest start-time from the node's static *b-level*.

(4) Select the node-processor pair that gives the largest DL. Schedule the node to the corresponding processor.

(5) Add the newly ready nodes to the ready pool
Until all nodes are scheduled.

The time-complexity of the DLS algorithm is $O(pv^3)$. For the DAG shown in Figure 3, the ETF algorithm generates a schedule shown in Figure 17(a). The steps of scheduling are given in the table shown in Figure 17(b). In the table, the start-times of the node on the processors at each scheduling step are given and the node is scheduled to the

(a)



(b)

Step	n_i	sl	PE0	PE1	PE2	PE3	DL
1	n_1	11	0*	N.C.	N.C.	N.C.	11
2	n_4	9	2*	3	N.C.	N.C.	6
3	n_3	8	6	3*	N.C.	N.C.	5
4	n_2	8	6*	6	N.C.	N.C.	2
5	n_5	5	9	6	3*	N.C.	2
6	n_8	5	9	7*	8	N.C.	-2
7	n_6	5	9*	11	10	N.C.	-4
8	n_7	5	13	12*	12	N.C.	-7
9	n_9	1	22	18*	22	N.C.	-17

Figure 17. (a) The schedule generated by the DLS algorithm (schedule length = 19); (b) a scheduling trace of the DLS algorithm (N.C. indicates "not considered").

processor on which the start-time is marked by an asterisk.

Performance on fork and join: Even though the DLS algorithm schedules nodes based on dynamic levels, it cannot guarantee optimal schedules for fork and join structures even if given sufficient processors.

6.4.6 *The LAST Algorithm*. LAST (Localized Allocation of Static Tasks) algorithm [Baxter and Patel 1989] is not a list scheduling algorithm, and uses for node priority an attribute called D_NODE, which depends only on the incident edges of a node. D_NODE is defined below:

$$D\_NODE(n_i) = \frac{\sum c(n_k,n_i) D\_EDGE(n_k,n_i) + \sum c(n_i,n_j) D\_EDGE(n_i,n_j)}{\sum c(n_k,n_i) + \sum c(n_i,n_j)}$$

In the the above definition, D_EDGE is equal to 1 if one of the nodes on the edge is assigned to some processor. The main goal of the LAST algorithm is to minimize the overall communication. The algorithm is briefly described below.

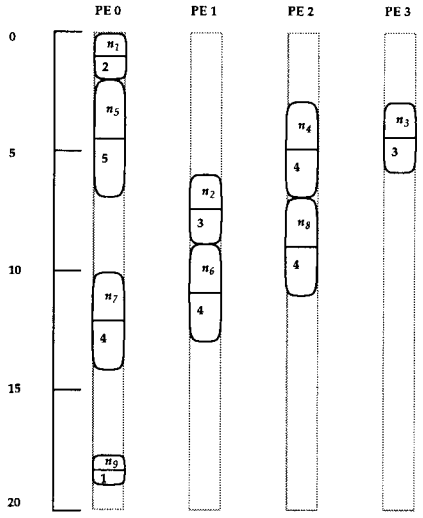
(1) For each entry node, set its D_NODE to be 1. Set all other D_NODEs to 0. **Repeat**

(2) Let *candidate* be the node with the highest D_NODE value.

(3) Schedule *candidate* to the processor which allows the minimum start-time.

(4) Update the D_EDGE and D_NODE values of all adjacent nodes of *candidate*.

(a)



(b)

Step	n_i	D_NODE	PE0	PE1	PE2	PE3
1	n_1	1.00	0*	N.C.	N.C.	N.C.
2	n_5	1.00	2*	3	N.C.	N.C.
3	n_2	0.67	7	6*	N.C.	N.C.
4	n_7	0.59	10*	12	12	N.C.
5	n_4	0.50	14	9	3*	N.C.
6	n_3	0.50	14	9	7	3*
7	n_8	0.29	14	9	7*	8
8	n_6	0.17	14	9*	11	10
9	n_9	1.00	18*	20	20	20

Figure 18. (a) The schedule generated by the LAST algorithm (schedule length = 19); (b) a scheduling trace of the LAST algorithm (N.C. indicates "not considered").

The time-complexity of the LAST algorithm is $O(v(e + v))$. For the DAG shown in Figure 3, the LAST algorithm generates a schedule shown in Figure 18(a). The steps of scheduling are given in the table shown in Figure 18(b). In the table, the start-times of the node on the processors at each scheduling step are given and the node is scheduled to the processor on which the start-time is marked by an asterisk.

Performance on fork and join: Since the LAST algorithm schedules nodes based on edge costs only, it cannot guarantee optimal schedules for fork and join structures even if given sufficient processors.

6.4.7 *Other BNP Approaches*. McCreary and Gill [1989] proposed a BNP scheduling technique based on the clustering method. In the algorithm, the DAG is first parsed into a set of CLANs. Informally, two nodes n_i and n_j are members of the same CLAN if and only if parents of n_j outside the CLAN are also parents of n_i, and children of n_i outside the CLAN are also children of n_j. Essentially, a CLAN is a subset of

nodes where every element outside the set is related in the same way to each member in the set. The CLANs so derived are hierachically related by a parse tree. That is, a CLAN can be a subset of another CLAN of larger size. Trivial CLANs include the single nodes and the whole DAG. Depending upon the number of processors available, the CLAN parse tree is traversed to determine the optimal CLAN size for assignment so as to reduce the schedule length.

Sih and Lee [1993b] reported a BNP scheduling scheme which is also based on clustering. The algorithm is called declustering because upon forming a hierarchy of clusters the optimal cluster size is determined possibly by cracking some large clusters in order to gain more parallelism while minimizing schedule length. Thus, using similar principles as in McCreary and Gill's approach, Sih and Lee's scheme also traverses the cluster hierarchy from top to bottom in order to match the level of cluster granularity to the characteristic of the target architecture. The crucial difference between their methods is in the cluster formation stage. While McCreary and Gill's method is based on CLANs construction, Sih and Lee's approach is to isolate a collection of edges that are likely candidates for separating the nodes at both ends onto different processors. These *cut-edges* are temporarily removed from the DAG and the algorithm designates each remaining connected component as an elementary cluster.

Lee et al. [1991] reported a BNP scheduling algorithm targeted for dataflow multiprocessors based on a vertical layering method for the DAG. In their scheme, the DAG is first partitioned into a set of vertical layers of nodes. The initial set of vertical layers is built around the critical path of the DAG and is then optimized by considering various cases of accounting for possible interprocessor communication, which may in turn induce new critical paths. Finally, the vertical layers of nodes are mapped to the given processors in order to minimize the schedule length.

Zhu and McCreary [1992] reported a set of BNP scheduling algorithms for trees. They first devised an algorithm for finding optimal schedules for trees, in particular, binary trees. Nonetheless the algorithm is of exponential complexity since optimal scheduling of trees is an NP-complete problem. They then proposed a number of heuristic approaches that can generate reasonably good solutions within a much shorter amount of time. The heuristics are all greedy in nature in that they attempt to minimize the completion times of paths in the tree and exploit only a small number of possible paths in the search of a good schedule.

Varvarigou et al. [1996] proposed a BNP scheduling scheme for in-forests and out-forests. However, their algorithm assumes that the trees are with unit computation costs and unit communication costs. Another distinctive feature of their algorithm is that the time-complexity is pseudopolynomial, $O(v^{2p})$, which is polynomial if p is fixed and small. The idea of their algorithm is to first transform the trees into delay-free trees, which are then scheduled using an optimal merging algorithm. This transformation step is crucial and is done as follows. For each node, a successor node is selected to be scheduled immediately after the node. Then, since the communication costs are unit, the communication costs between the node and all other successors can be dropped. Only an extra communication free edge is needed to add between the chosen successor and the other successors. The successor node is so selected that the resulting DAG does not violate the precedence constraints of the original DAG.

Pande et al. [1994] proposed a BNP scheduling scheme using a thresholding technique. The algorithm first computes the earliest start-times and latest start-times of the nodes. A threshold for a node is then the difference between its earliest and the latest start-times. A

global threshold is varied between the minimum threshold among the nodes to the maximum. For a node with threshold less than the global value, a new processor is allocated for the node, if there is any available. For a node with threshold above the global value, the node is then scheduled to the same processor as its parent which allows the earliest start-time. The rationale of the scheme is that as the threshold of a node represents the tolerable delay of execution without increasing overall schedule length, a node with smaller threshold deserves a new processor so that it can start as early as possible. Depending upon the number of given processors, there is a trade-off between parallelism and schedule length, and the global threshold is adjusted accordingly.

6.4.8 *Analytical Performance Bounds of BNP Scheduling*. For the BNP class of scheduling algorithms, Al-Mouhamed [1990] extended Fernandez and Bussell's work [1973] (described in Section 6.2.3) and devised a bound on the minimum number of processors for optimal schedule length and a bound on the minimum increase in schedule length if only a certain smaller number of processor is available. Essentially, Al-Mouhamed extended the techniques of Fernandez et al. for arbitrary DAGs with communication. Furthermore, the expressions for the bounds are similar to the ones reported by Fernandez and Bussell, except that Al-Mouhamed conjectured that the bounds need not be computed across all possible integer intervals within the earliest completion time of the DAG. However, Jain and Rajaraman [1995] in a subsequent study found that the computation of these bounds needs to consider all the integer intervals within the earliest completion time of the DAG. They also reported a technique to partition the DAGs into nodes with non-overlapping intervals so that a tighter bound is obtained. In addition, the new bounds can take lesser time to compute. Jain and Rajaraman also found that using such a partitioning facilitates all possible integer intervals to be considered in order to compute a tighter bound.

6.5 TDB Scheduling

In this section ,we survey the TDB class of DAG scheduling algorithms. We describe in detail six TDB scheduling algorithms: the PY, LWB, DSH, BTDH, LCTD, and CPFD algorithms. The DAG shown in Figure 3 is used to illustrate the scheduling process of these algorithms.

In the following, we do not discuss the performance of the TDB algorithms on fork and join sets separately because with duplication the TDB scheduling schemes can inherently produce optimal solutions for these two primitive structures. For a fork set, a TDB algorithm duplicates the root on every processor so that each child starts at the earliest possible time. For a join set, although no duplication is needed to start the sink node at the earliest time, all the TDB algorithms surveyed in this section employ a similar recursive scheduling process to minimize the start-times of nodes so that an optimal schedule results.

6.5.1 *The PY Algorithm*. The PY algorithm (named after Papadimitriou and Yannakakis[1990]) is an approximation algorithm which uses an attribute, called e-value, to approximate the absolute achievable lower bound of the start-time of a node. This attribute is computed recursively beginning from the entry nodes to the exit nodes. A procedure for computing the e-values is given below.

(1) Construct a list of nodes in topological order. Call it *TopList*.
(2) **for** each node n_i in *TopList* **do**
(3) **if** n_i has no parent **then** $e(n_i) = 0$
(4) **else**
(5) **for** each parent n_x of n_i do $f(n_x) = e(n_x) + c(n_x, n_i)$ **endfor**
(6) Construct a list of parents in decreasing f. Call it **ParentList**.

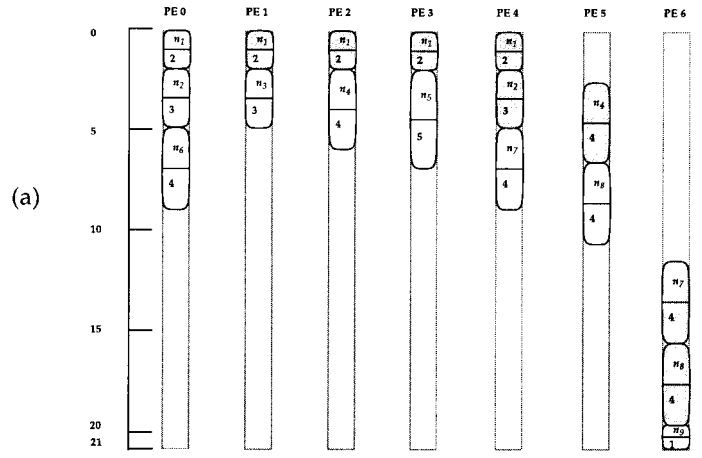


Figure 19. (a) The schedule generated by the PY algorithm (schedule length = 21); (b) the *3*-values of the nodes computed by the PY algorithm.

(7) Let min_e = the f value of the first parent in *ParentList*
(8) Make n_i as a single node cluser. Call it $Cluster(n_i)$.
(9) **for** each parent n_x in *ParentList* **do**
(10) Include $Cluster(n_x)$ in $Cluster(n_i)$.
(11) Compute the new min_e (i.e., start-time) of n_i in $Cluster(n_i)$.
(12) if new min_e > original min_e **then** exit this for-loop **endif**
(13) **endfor**
(14) $e(n_i) = min\_e$
(15) **endif**
(16) **endfor**

After computing the *e*-values, the algorithm inserts each node into a cluster, in which a group of ancestors are to be duplicated such that the ancestors have data arrival times larger than the e-value of the node. Papadimitriou and Yannakakis also showed that the schedule length generated is within a factor of two from the optimal. The PY algorithm is briefly described below.

(1) Compute *e*-values for all nodes.
(2) **for** each node n_i **do**
(3) Assign n_i to a new processor PEi.
(4) **for** all ancestors of n_i, duplicate an ancestor n_x if:

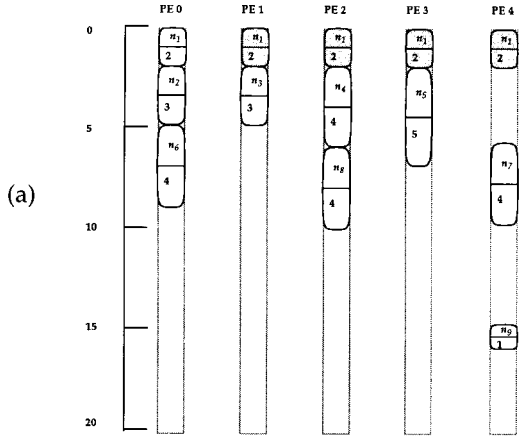
$$e(n_x) + w(n_x) + c(n_x, n_i) > e(n_i)$$

(5) Order the nodes in PEi so that a node starts as soon as all its data is available.
(6) **endfor**

The time-complexity of the PY algorithm is $O(v^2(e + v\log v))$. For the DAG shown in Figure 3, the PY algorithm generates a schedule shown in Figure 19(a). The e-values are also shown in Figure 19(b).

6.5.2 *The LWB Algorithm*. We call the algorithm the LWB (Lower Bound) algorithm [Colin and Chretienne 1991] based on its main principle: it first determines the lower bound start-time for each node, and then identifies a set of critical edges in the DAG. A critical edge is the one in which a parent's message-available time for the child is greater than the lower bound start-time of the child. Colin and Chretienne [1991] showed that the LWB algorithm can generate optimal schedules for DAGs in which node weights are strictly larger than any edge weight. The LWB algorithm is briefly described below.

(1) For each node n_i, compute its lower bound start-time, denoted by $lwb(n_i)$, as follows:



(b)

n_i	lwb
n_1	0
n_2	2
n_3	2
n_4	2
n_5	2
n_6	5
n_7	5
n_8	6
n_9	15

Figure 20. (a) The schedule generated by the LWB algorithm (schedule length = 16); (b) the lwb (lower bound) values of the nodes computed by the LWB algorithm.

(a) For any entry node n_i, $lwb(n_i)$ is zero.

(b) For any node n_i other than an entry node, consider the set of its parents. Let n_x be the parent such that $lwb(n_x) + w(n_x) + c(n_x, n_i)$ is the largest among all parents. Then, the lower bound of n_i, $lwb(n_i)$, is given by, with $n_y \neq n_x$,

$$\text{MAX}\{lwb\{n_x\} + w(n_x), \text{MAX}\{lwb(n_y) + w(n_y) + c(n_y, n_i)\}\}$$

(2) Consider the set of edges in the task graph. An edge (n_y, n_i) is labelled as "critical" if $lwb(n_x) + w(n_x) + c(n_x, n_i) > lwb(n_i)$.

(3) Assign each path of critical edges to a distinct processor such that each node is scheduled to start at its lower bound start-time.

The time-complexity of the LWB algorithm is $O(v^2)$. For the DAG shown in Figure 3, the LWB algorithm generates a schedule shown in Figure 20(a). The lower bound values are also shown in Figure 20(b).

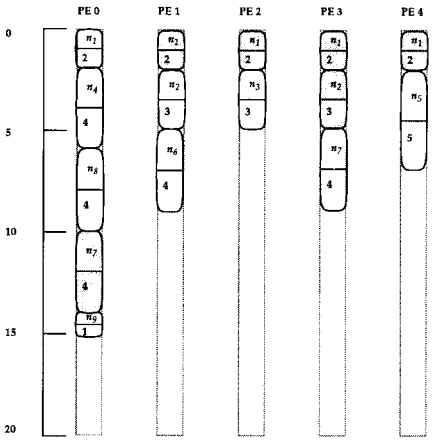
6.5.3 *The DSH Algorithm*. The DSH (Duplication Scheduling Heuristic) algorithm [Kruatrachue and Lewis 1988] considers each node in a descending order of their priorities. In examining the suitability of a processor for a node, the DSH algorithm first determines the start-time of the node on the processor without duplication of any ancestor. Then, it considers the duplication in the idle time period from the finish-time of the last scheduled node on the processor and the start-time of the node currently under consideration. The algorithm then tries to duplicate the ancestors of the node into the duplication time slot until either the slot is used up or the start-time of the node does not improve. The algorithm is briefly described below.

(1) Compute the *static b-level* for each node.
Repeat

(2) Let n_i be an unscheduled node with the largest static *b-level*.

(3) For each processor P, do
(a) Let the ready time of P, denoted by RT, be the finish-time of the last node on P. Compute the

(a)



(b)

Step	n_i	PE 0	PE 1	PE 2	PE 3	PE 4	PE 5	Nodes Dup.
1	n_1	0*	N.C.	N.C.	N.C.	N.C.	N.C.	NIL
2	n_4	2*	2	N.C.	N.C.	N.C.	N.C.	NIL
3	n_2	6	2*	N.C.	N.C.	N.C.	N.C.	n_1
4	n_3	6	5	2*	N.C.	N.C.	N.C.	n_1
5	n_6	N.C.	5*	6	N.C.	N.C.	N.C.	NIL
6	n_8	6*	N.C.	7	7	N.C.	N.C.	NIL
7	n_7	10	9	6	5*	N.C.	N.C.	n_1, n_2
8	n_5	10	9	5	9	2*	N.C.	n_1
9	n_9	14*	15	N.C.	15	N.C.	15	n_7

Figure 21. (a) The schedule generated by the DSH algorithm (schedule length = 15); (b) a scheduling trace of the DSH algorithm.

start-time of n_i on P and denote it by ST. Then the duplication time slot on P has length ($ST - RT$). Let candidate be n_i.

(b) Consider the set of candidate's parents. Let n_x be the parent of n_i which is not scheduled on P and whose message for candidate has the latest arrival time. Try to duplicate n_x into the duplication time slot.

(c) If the duplication is unsuccessful, then record ST for this processor and try another processor; otherwise, let ST be *candidate*'s new start-time and *candidate* be n_x. Go to step (b).

(4) Let P' be the processor that gives the earliest start-time of n_i. Schedule n_i to P' and perform all the necessary duplication on P'

Until all nodes are scheduled.

The time-complexity of the DSH algorithm is $O(v^4)$. For the DAG shown in Figure 3, the DSH algorithm generates a schedule shown in Figure 21(a). The steps of scheduling are given in the table shown in Figure 21(b). In the table, the start-times of the node on the processors at each scheduling step are given and the node is scheduled to the processor on which the start-time is marked by an asterisk.

6.5.4 *The BTDH Algorithm*. The BTDH (Bottom-Up Top-Down Duplication Heuristic) algorithm [Chung and Ranka 1992] is an extension of the DSH algorithm described above. The major improvement of the BTDH algorithm

over the DSH algorithm is that the algorithm keeps on duplicating ancestors of a node even if the duplication time slot is totally used up (i.e., the start-time of the node temporarily increases) with the hope that the start-time will eventually be minimized. That is, the BTDH algorithm is the same as the DSH algorithm except for step (3)(c) of the latter in that the duplication of an ancestor is considered successful even if the duplication time slot is used up. The process stops only when the final start-time of the node is greater than before the duplication. The time-complexity of the BTDH algorithm is also $O(v^4)$. For the DAG shown in Figure 3, the BTDH algorithm generates the same schedule as the DSH algorithm which is shown in Figure 21(a). The scheduling process is also the same except at step (5) when node n_6 is considered for scheduling on PE 2, the start-time computed by the BTDH algorithm is also 5 instead of 6 as computed by the DSH algorithm. This is because the BTDH algorithm does not stop the duplication process even though the start-time increases.

6.5.5 *The LCTD Algorithm*. The LCTD (Linear Clustering with Task Duplication) algorithm [Chen et al. 1993] is based on linear clustering of the DAG. After performing the clustering step, the LCTD algorithm identifies the edges among clusters that determines the completion time. Then, it tries to duplicate the parents corresponding to these edges to reduce the start-times of some nodes in the clusters. The algorithm is described below.

(1) Apply the LC algorithm to the DAG to generate a set of linear clusters.

(2) Schedule each linear cluster to a distinct processor and let the nodes start as early as possible on the processors.

(3) For each linear cluster C_1 do:
  (a) Let the first node in C_1 be n_x.
  (b) Consider the set of n_x's parents. Select the parent that allows the largest reduction of n_x's start-time. Duplicate this parent and all the necessary ancestors to C_1.
  (c) Let n_x be the next node in CP_i. Go to step (b).

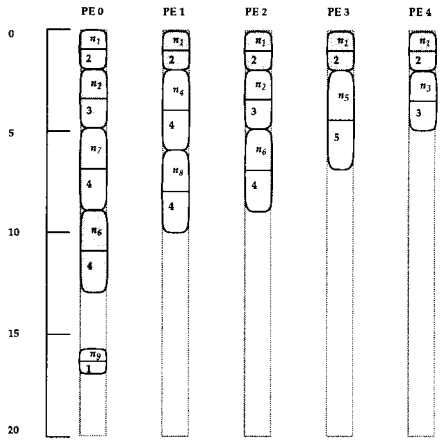
(4) Consider each pair of processors. If their schedules have enough common nodes so that they can be merged without increasing the schedule length, then merge the two schedules and discard one processor.

The time-complexity of the LCTD algorithm is $O(v^3 \log v)$. For the DAG shown in Figure 3, the LCTD algorithm generates a schedule shown in Figure 22(a). The steps of scheduling are given in the table shown in Figure 22(b). In the table, the original start-times of the node on the processors after the linear clustering step are given. In addition, the improved start-times after duplication are also given.

6.5.6 *The CPFD Algorithm*. The CPFD (Critical Path Fast Duplication) algorithm [Ahmad and Kwok 1998a] is based on partitioning the DAG into three categories: critical path nodes (CPN), in-branch nodes (IBN), and out-branch nodes (OBN). An IBN is a node from which there is a path reaching a CPN. An OBN is a node which is neither a CPN nor an IBN. Using this partitioning of the graph, the nodes can be ordered in decreasing priority as a list called the CPN-Dominant Sequence. In the following, we first describe the construction of this sequence.

In a DAG, the CP nodes (CPNs) are the most important nodes since their finish-times effectively determine the final schedule length. Thus, the CPNs in a task graph should be considered as early as possible for scheduling in the scheduling process. However, we cannot consider all the CPNs without first considering other nodes because the start-times of the CPNs are determined by their parent nodes. Therefore, before we can consider a CPN for scheduling, we

(a)



(b)

Step	n_i	Org. ST	New ST	Nodes Dup.
1	n_1	0	0	NIL
2	n_7	10	5	n_2
3	n_9	18	16	n_6
4	n_4	3	2	n_1
5	n_8	7	7	NIL
6	n_2	6	2	n_1
7	n_6	9	5	NIL
8	n_5	3	2	n_1
9	n_3	3	2	n_1

Figure 22. (a) The schedule generated by the LCTD algorithm (schedule length = 17); (b) a scheduling trace of the LCTD algorithm.

must first consider all its parent nodes. In order to determine a scheduling order in which all the CPNs can be scheduled as early as possible, we classify the nodes of the DAG into three categories given in the following definition.

Definition 4. In a connected graph, an In-Branch Node (IBN) is a node, which is not a CPN, and from which there is a path reaching a Critical Path Node (CPN). An Out-Branch Node (OBN) is a node, which is neither a CPN nor an IBN.

After the CPNs, the most important nodes are IBNs because their timely scheduling can help reduce the start-times of the CPNs. The OBNs are relatively less important because they usually do not affect the schedule length. Based on this reasoning, we make a sequence of nodes called the CPN-Dominant sequence which can be constructed by the following procedure:

Constructing the CPN-Dominant Sequence

(1) Make the entry CPN to be the first node in the sequence. Set *Position* to 2. Let n_x be the next CPN.

Repeat

(2) **If** n_x has all its parent nodes in the sequence **then**

(3) Put n_x at *Position* in the sequence and increment *Position*.

(4) **else**

(5) Suppose n_y is the parent node of n_x which is not in the sequence and has the largest *b-level*. Ties are broken by choosing the parent with a smaller *t-level*. If n_y has all its parent nodes in the sequence, put n_y at *Position* in the sequence and increment *Position*. Otherwise, recursively include

all the ancestor nodes of n_y in the sequence so that the nodes with a larger communication are considered first.

(6) Repeat the above step until all the parent nodes of n_x are in the sequence. Put n_x in the sequence at *Position*.

(7) **endif**

(8) Make n_x to be the next CPN.

Until all CPNs are in the sequence.

(9) Append all the OBNs to the sequence in a decreasing order of *b-level*.

The CPN-Dominant sequence preserves the precedence constraints among nodes as the IBNs reaching a CPN are always inserted before the CPN in the CPN-Dominant sequence. In addition, the OBNs are appended to the sequence in a topological order so that a parent OBN is always in front of a child OBN.

The CPN-Dominant sequence of the DAG shown in Figure 3 is constructed as follows. Since n_1 is the entry CPN, it is placed in the first position in the CPN-Dominant sequence. The second node is n_2 because it has only one parent node. After n_2 is appended to the CPN-Dominant sequence, all parent nodes of n_7 have been considered and can, therefore, also be added to the sequence. Now the last CPN, n_9 is considered. It cannot be appended to the sequence because some of its parent nodes (i.e., the IBNs) have not been examined yet. Since both n_6 and n_8 have the same *b-level* but n_8 has a smaller *t-level*, n_8 is considered first. However, both parent nodes of n_8 have not been examined, thus, its two parent nodes, n_3 and n_4 are appended to the CPN-Dominant sequence first. Next, n_8 is appended followed by n_6. The only OBN, n_5, is the last node in the CPN-Dominant sequence. The final CPN-Dominant sequence is as follows: n_1, n_2, n_7, n_4, n_3, n_8, n_6, n_9, n_5 (see Figure 3(b); the CPNs are marked by an asterisk). Note that using *sl* (static level) as a priority measure will generate a different ordering of nodes: n_1, n_4, n_2, n_3, n_5, n_6, n_7, n_8, n_9.

Based on the CPN-Dominant sequence, the CPFD algorithm is briefly described below.

(1) Determine a critical path. Partition the task graph into CPNs, IBNs, and OBNs. Let *candidate* be the entry CPN.

Repeat

(2) Let P_SET be the set of processors containing the ones accommodating the parents of *candidate* plus an empty processor.

(3) For each processor P in P_SET, do:
   (a) Determine *candidate*'s start-time on P and denote it by ST.
   (b) Consider the set of *candidate*'s parents. Let m be the parent which is not scheduled on P and whose message for *candidate* has the latest arrival time.
   (c) Try to duplicate m on the earliest idle time slot on P. If the duplication is successful and the new start-time of *candidate* is less than ST, then let ST be the new start-time of *candidate*. Change *candidate* to m and go to step (a). If the duplication is unsuccessful, then return control to examine another parent of the previous *candidate*.

(4) Schedule *candidate* to the processor P' that gives the earliest start-time and perform all the necessary duplication.

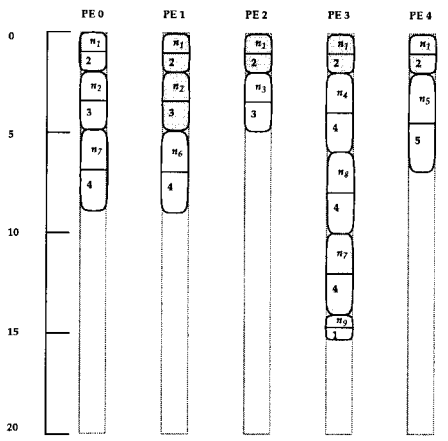
(5) Let *candidate* be the next CPN.

(6) Repeat the process from step (2) to step (5) for each OBN with P_SET containing all the processors in use together with an empty processor. The OBNs are considered one by one topologically.

Until all CPNs are scheduled.

The time-complexity of the CPFD algorithm is $O(v^4)$. For the DAG shown in

(a)



(b)

Step	n_i	PE 0	PE 1	PE 2	PE 3	PE 4	Nodes Dup.
1	n_1	0*	N.C.	N.C.	N.C.	N.C.	NIL
2	n_2	2*	2	N.C.	N.C.	N.C.	NIL
3	n_7	5*	5	N.C.	N.C.	N.C.	NIL
4	n_6	9	5*	N.C.	N.C.	N.C.	n_1, n_2
5	n_3	9	9	2*	N.C.	N.C.	n_1
6	n_4	9	9	5	2*	N.C.	n_1
7	n_8	N.C.	N.C.	7	6*	6	NIL
8	n_9	14	15	14	14*	N.C.	n_7
9	n_5	9	9	5	15	2*	n_1

Figure 23. (a) The schedule generated by the CPFD algorithm (schedule length = 15); (b) a scheduling trace of the CPFD algorithm.

Figure 3, the CPFD algorithm generates a schedule shown in Figure 23(a). The steps of scheduling are given in the table shown in Figure 23(b). In this table, the start-times of the node on the processors at each scheduling step are given and the node is scheduled to the processor on which the start-time is marked by an asterisk.

6.5.7 *Other TDB Approaches*. Anger et al. [1990] reported a TDB scheduling scheme called JLP/D (Joint Latest Predecessor with Duplication). The algorithm is optimal if the communication costs are strictly less than any computation costs, and there are sufficient processors available. The basic idea of the algorithm is to schedule every node with its latest parent to the same processor. Since a node can be the latest parent of several successors, duplication is necessary.

Markenscoff and Li [1993] reported a TDB scheduling approach based on an optimal technique for scheduling in-trees. In their scheme, a DAG is first transformed into a set of in-trees. A node in the DAG may appear in more than one in-tree after the transformation. Each tree is then optimally scheduled independently and hence, duplication comes into play.

In a recent study, Darbha and Agrawal [1995] proposed a TDB scheduling algorithm using similar principles as the LCTD algorithm. In the algorithm, a DAG is first parsed into a set of linear clusters. Then each cluster is examined to determine the critical nodes for duplication. Critical nodes are the

nodes that determine the data arrival time of the nodes in the cluster but are themselves outside the cluster. Similar to the LCTD algorithm, the number of processors required is also optimized by merging schedules with the same set of "prefix" schedules.

Palis et al. [1996] also investigated the problem of scheduling task graphs to processors using duplication. They proposed an approximation TDB algorithm which produces schedule lengths at most twice from the optimal. They also showed that the quality of the schedule improves as the granularity of the task graph becomes larger. For example, if the granularity is at least 1/2, the schedule length is at most 5/3 times optimal. The time-complexity of the algorithm is $O(v(v\log v + e))$, which is v times faster than the PY algorithm proposed by Papadimitriou and Yannakakis [1990]. In Palis et al. [1996], similar algorithms were also developed that produce: (1) optimal schedules for coarse grain graphs; (2) 2-optimal schedules for trees with no task duplication; and (3) optimal schedules for coarse grain trees with no task duplication.

6.6 APN Scheduling

In this section, we survey the APN class of DAG scheduling algorithms. In particular we describe in detail four APN algorithms: the MH (Mapping Heuristic) algorithm [Rewini and Lewis 1990], the DLS (Dynamic Level Scheduling) algorithm [Sih and Lee 1993a], the BU (Bottom Up) algorithm [Mehdiratta and Ghose 1994], and the BSA (Bubble Scheduling and Allocation) algorithm [Kwok and Ahmad 1995]. Before we discuss these algorithms, it is necessary to examine one of the most important issues in APN scheduling–the message routing issue.

6.6.1 *The Message Routing Issue*. In APN scheduling, a processor network is not necessarily fully-connected and contention for communication channels needs to be addressed. This in turn implies that message routing and scheduling must also be considered. Recent high-performance architectures (nCUBE-2 [Hwang 1993], iWarp [Hwang 1993], and Intel Paragon [Quinn 1994]) employ wormhole routing in which the header flit of a message establishes the path, intermediate flits follow the path, and the tail flit releases the path. Once the header gets blocked due to link contention, the entire message waits in the network, occupying all the links it is traversing. Hence, it increasingly becomes important to take link contention into account as compared to distance when scheduling computations onto wormhole-routed systems. Routing strategies can be classified as either *deterministic* or *adaptive*. Deterministic schemes, such as the *e-cube* routing for hypercube topology, construct fixed routes for messages and cannot avoid contention if two messages are using the same link even when other links are free. Yet deterministic schemes are easy to implement and routing decisions can be made efficiently. On the other hand, adaptive schemes construct optimized routes for different messages depending upon the current channel allocation in order to avoid link contention. However, adaptive schemes are usually more complex as they require much state information to make routing decisions.

Wang [1990] suggested two adaptive routing schemes suitable for use in APN scheduling algorithms. The first scheme is a greedy algorithm which seeks a locally optimal route for each message to be sent between tasks. Instead of searching for a path with the least waiting time, the message is sent through a link which yields the least waiting time among the links that the processor can choose from for sending a message. Thus, the route is only locally optimal. Using this algorithm, Wang observed that there are two types of possible blockings: (i) a later message blocks an earlier message (called LBE blocking), and (ii) an earlier message blocks a later message (called EBL blocking).

LBE blocking is always more costly than EBL blocking. In the case that several messages are competing for a link and blocking becomes unavoidable, LBE blockings should be avoided as much as possible. Given this observation, Wang proposed the second algorithm, called the least blocking algorithm, which works by trying to avoid LBE blocking. The basic idea of the algorithm is to use Dijkstra's shortest path algorithm to arrange optimized routes for messages so as to avoid LBE blockings.

Having determined routes for messages, the scheduling of different messages on the links is also an important aspect. Dixit-Radiya and Panda [1993] proposed a scheme for ordering messages in a link so as to further minimize the extent of link contention. Their scheme is based on the Temporal Communication Graph (TCG) which, in addition to task precedence, captures the temporal relationship of the communication messages. Using the TCG model, the objective of which is to minimize the contention on the link, the earliest start-times and latest start-times of messages can be computed. These values are then used to heuristically schedule the messages in the links.

6.6.2 *The MH Algorithm.* The MH (Mapping Heuristic) algorithm [El-Rewini and Lewis 1990] first assigns priorities by computing the *sl* of all nodes. A ready node list is then initialized to contain all entry nodes ordered in decreasing priorities. Each node is scheduled to a processor that gives the smallest start-time. In calculating the start-time of node, a routing table is maintained for each processor. The table contains information as to which path to route messages from the parent nodes to the node under consideration. After a node is scheduled, all of its ready successor nodes are appended to the ready node list. The MH algorithm is briefly described below.

(1) Compute the sl of each node n_i in the task graph.

(2) Initialize a ready node list by inserting all entry nodes in the task graph. The list is ordered according to node priorities, with the highest priority node first.

Repeat

(3) $n_i \leftarrow$ the first node in the list.

(4) Schedule n_i to the processor which gives the smallest start-time. In determining the start-time on a processor, all messages from the parent nodes are scheduled and routed by consulting the routing tables associated with each processor.

(5) Append all ready successor nodes of n_i, according to their priorities, to the ready node list.

Until the ready node list is empty.

The time-complexity of the MH algorithm is shown to be $O(v(p^3v + e))$, where p is the number of processors in the target system.

For the DAG shown in Figure 3(a), the schedule generated by the MH algorithm for a 4-processor ring is shown in Figure 24. Here, L_{ij} denotes a communication link between PE i and PE j. The MH algorithm schedules the nodes in the following order: n_1, n_4, n_3, n_5, n_2, n_8, n_7, n_6, n_9. Note that the MH algorithm does not strictly schedule nodes according to a descending order of sls(static levels) in that it uses the sl order to break ties. As can be seen from the schedule shown in Figure 24, the MH algorithm schedules n_4 first before n_2 and n_7, which are more important nodes. This is due to the fact that both algorithms rank nodes according to a descending order of their sls. The nodes n_2 and n_7 are more important because n_7 is a CPN and n_2 critically affects the start-time of n_7. As n_4 has a larger static level, both algorithms examine n_4 first and schedule it to an early time slot on the same processor as n_1. As a result, n_2 cannot start at the earliest

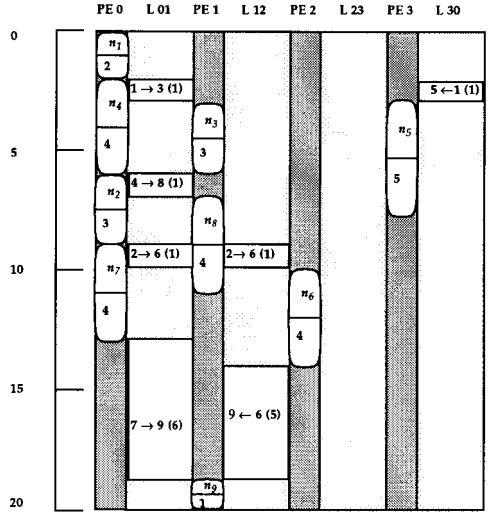


Figure 24. The schedule generated by the MH and DLS algorithm (schedule length = 20, total comm. costs incurred = 16).

possible time–the time just after n_1 finishes.

6.6.3 *The DLS Algorithm*. The DLS (Dynamic Level Scheduling) algorithm [Sih and Lee 1993a] described in Section 6.4.5 can also be used as an APN scheduling algorithm. However, the DLS algorithm requires a message routing method to be supplied by the user. It then computes the earliest start-time of a node on a processor by tentatively scheduling and routing all messages from the parent nodes using the given routing table.

For APN scheduling, the time-complexity of the DLS algorithm is shown to be $O(v^3pf(p))$, where $f(p)$ is the time-complexity of the message routing algorithm. For the DAG shown in Figure 3(a), the schedule generated by the DLS algorithm for a 4-processor ring is the same as that generated by the MH algorithm shown in Figure 24. The DLS algorithm also schedules the nodes in the following order: n_1, n_4, n_3, n_5, n_2, n_8, n_7, n_6, n_9.

6.6.4 *The BU Algorithm*. The BU (Bottom-Up) algorithm [Mehdiratta and Ghose 1994] first determines the critical path (CP) of the DAG and then assigns all the nodes on the CP to the same processor. Afterwards, the algorithm assigns the remaining nodes in a reversed topological order of the DAG to the processors. The node assignment is guided by a load-balancing processor selection heuristic which attempts to balance the load across all processors. The BU algorithm examines the nodes at each topological level in a descending order of their *b-levels*. After all the nodes are assigned to the processors, the BU algorithm tries to schedule the communication messages among them using a channel allocation heuristic which tries to keep the hop count of every message roughly a constant constrained by the processor network topology. Different network topologies require different channel allocation heuristics. The BU algorithm is briefly described below.

(1) Find a critical path. Assign the nodes on the critical path to the

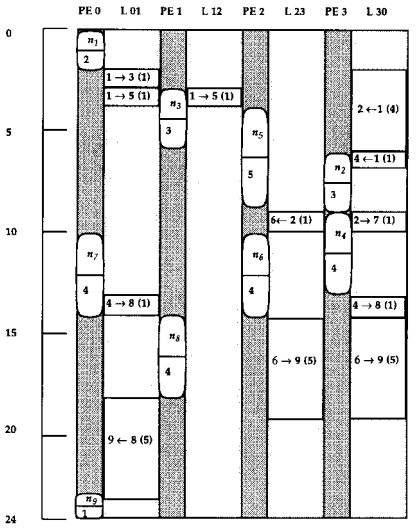


Figure 25. The schedule generated by the BU algorithm (schedule length = 24, total comm. costs incurred = 27).

same processor. Mark these nodes as assigned and update the load of the processor.

(2) Compute the *b-level* of each node. If the two nodes of an edge are assigned to the same processor, the communication cost of the edge is taken to be zero.

(3) Compute the *p-level* (precedence level) of each node, which is defined as the maximum number of edges along a path from an entry node to the node.

(4) In a decreasing order of *p-level*, for each value of *p-level*, do:
   (a) In a decreasing order of *b-level*, for each node at the current p-level, assign the node to a processor such that the processing load is balanced across all the given processors.
   (b) Re-compute the *b-levels* of all nodes.

(5) Schedule the communication messages among the nodes such that the hop count of each message is maintained constant.

The time-complexity of the BU algorithm is shown to be $O(v^2 \log v)$.

For the DAG shown in Figure 3(a), the schedule generated by the BU algorithm[1] for a 4-processor ring is shown in Figure 25. As can be seen, the schedule length is considerably longer than that of the MH and DLS algorithms. This is because the BU algorithm employs a processor selection heuristic which works by attempting to balance the load across all the processors.

6.6.5 *The BSA Algorithm*. The BSA (Bubble Scheduling and Allocation) algorithm [Kwok and Ahmad 1995] is proposed by us and is based on an incremental technique, which works by improving the schedule through migration of tasks from one processor to a neighboring processor. The algorithm first allocates all the tasks to a single processor which has the highest connectivity in the processor network and is

[1]In this example, we have used the PSH2 processor selection heuristic with $p = 1.5$. Such a combination is shown [Mehdiratta and Ghose 1994] to give the best performance.

called the pivot processor. In the first phase of the algorithm, the tasks are arranged in the processor according to the CPN-Dominant sequence discussed earlier in Section 6.5.6. In the second phase of the algorithm, the tasks migrate from the pivot processor to the neighboring processors if the start-times improve. This task migration process proceeds in a breadth-first order of the processor network in that after the migration process is complete for the first pivot processor, one of the neighboring processors becomes the next pivot processor and the process repeats.

In the following outline of the BSA algorithm, the *Build_processor_list()* procedure constructs a list of processors in a breadth-first order from the first pivot processor. The *Serial_injection()* procedure constructs the CPN-Dominant sequence of the nodes and injects this sequence to the first pivot processor.

The BSA Algorithm

(1) Load processor topology and input task graph
(2) *Pivot_PE*← the processor with the highest degree
(3) *Build_processor_list(Pivot_PE)*
(4) *Serial_injection(Pivot_PE)*
(5) **while** *Processor_list_not_empty* **do**
(6) *Pivot_PE*← first processor of *Processor_list*
(7) **for** each n_i on *Pivot_PE* **do**
(8) **if** $ST(n_i, Pivot\_PE) > \mathrm{DAT}(n_i, Pivot\_PE)$ or $Proc(VIP(n_i)) = Pivot\_PE$ **then**
(9) Determine DAT and ST of n_i on each adjacent processor PE′
(10) **if** there exists a PE' s.t. $ST(n_i, PE') < ST(n_i, Pivot\_PE)$ **then**
(11) Make n_i to migrate from $Pivot\_PE$ to PE'
(12) Update start-times of nodes and messages
(13) **else if** $ST(n_i, PE') = ST(n_i, Pivot\_PE)$ and $Proc(VIP(n_i))$ **then**
(14) Make n_i to migrate from $Pivot\_PE$ to PE' **then**
(15) Update start-times of nodes and messages
(16) **end if**
(17) **end if**
(18) **end for**
(19) **endwhile**

The time-complexity of the BSA algorithm is $O(p^2ev)$.

The BSA algorithm, as shown in Figure 26(a), injects the CPN-Dominant sequence to the first pivot processor PE 0. In the first phase, nodes n_1, n_2, and n_7 do not migrate because they are already scheduled to start at the earliest possible times. However, as shown in Figure 26(b), node n_4 migrates to PE 1 because its start-time improves. Similarly, as depicted in Figure 26(c), node n_3 also migrates to a neighboring processor PE 3. Figure 26(d) shows the intermediate schedule after n_8 migrates to PE 1 following its VIP n_4. Similarly, n_6 also migrates to PE 3 following its VIP n_3, as shown in Figure 27(a). The last CPN, n_9, migrates to PE 1 to which its VIP $n_8$is scheduled. Such migration allows the only OBN n_5 to bubble up. The resulting schedule is shown in Figure 27(b). This is the final schedule as no more nodes can improve the start-time through migration.

6.6.6 *Other APN Approaches*. Kon'ya and Satoh [1993] reported an APN scheduling algorithm for the hypercube architectures. Their algorithm, called the LST (Latest Starting Time) algorithm, works by using a list scheduling approach where the priorities of nodes are first computed and a list is constructed based on these priorities. The priority of a node is defined as its latest starting time, which is determined before scheduling starts. Thus, the list is static and does not capture the dynamically changing importance of nodes, which is crucial in APN scheduling.

In a later study, Selvakumar and Murthy [1994] reported an APN scheduling scheme which is an extension of Sih and Lee's DLS algorithm. The distinctive new feature in their algorithm is that it exploits schedule holes in processors and communication links in or-

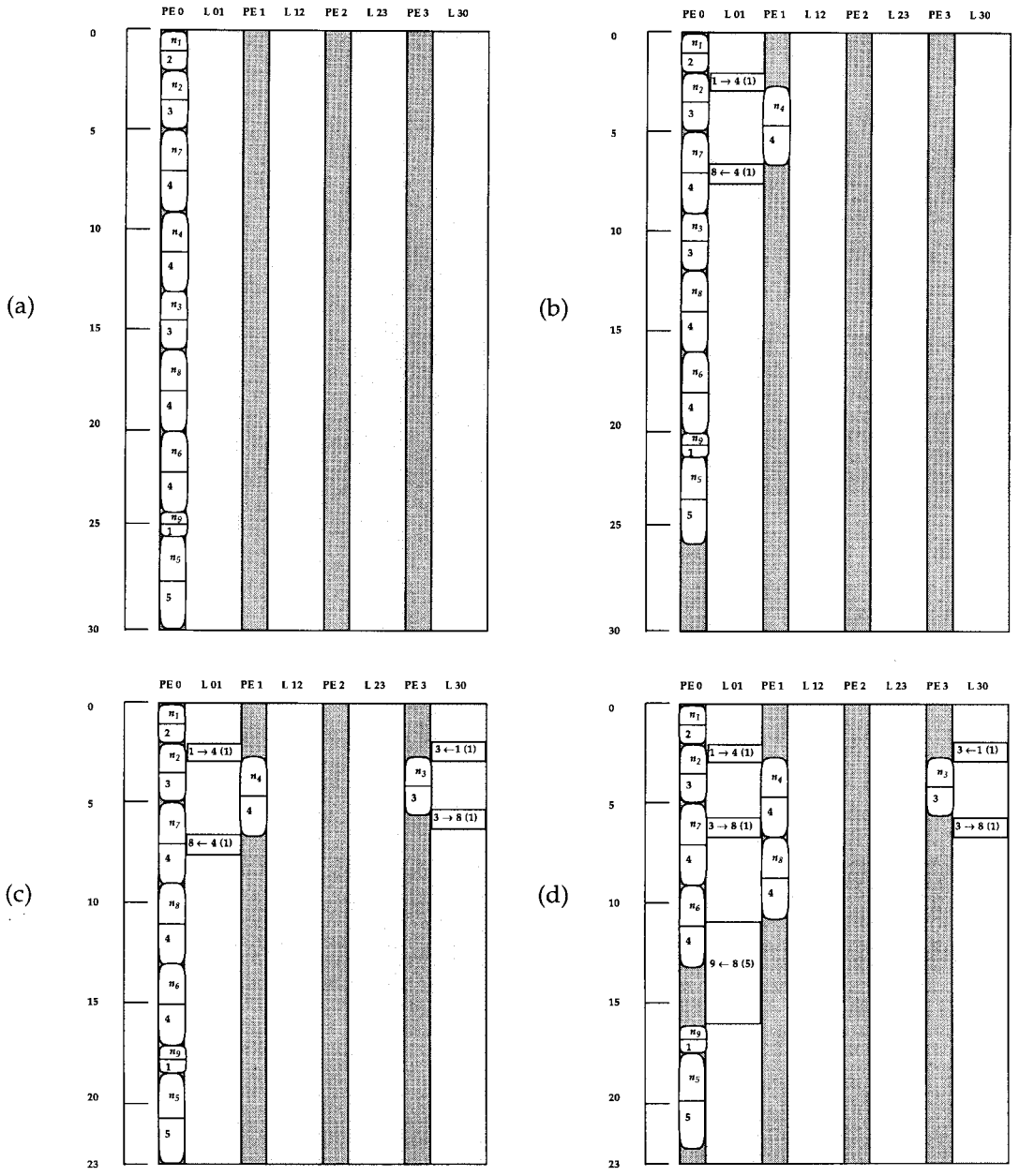


Figure 26. Intermediate schedules produced by BSA after (a) serial injection (schedule length = 30, total comm. cost = 0); (b) n_4 migrates from PE 0 to PE 1 (schedule length = 26, total comm. cost = 2); (c) n_3 migrates from PE 0 to PE 3 (schedule length = 23, total comm. cost = 4); (d) n_8 migrates from PE 0 to PE 1 (schedule length = 22, total comm. cost = 9).

der to produce better schedules. Essentially, it differs from the DLS algorithm in two respects: (i) the way in which the priority of a task with respect to a processor in a partial schedule; and (ii) the way in which a task and all communica-

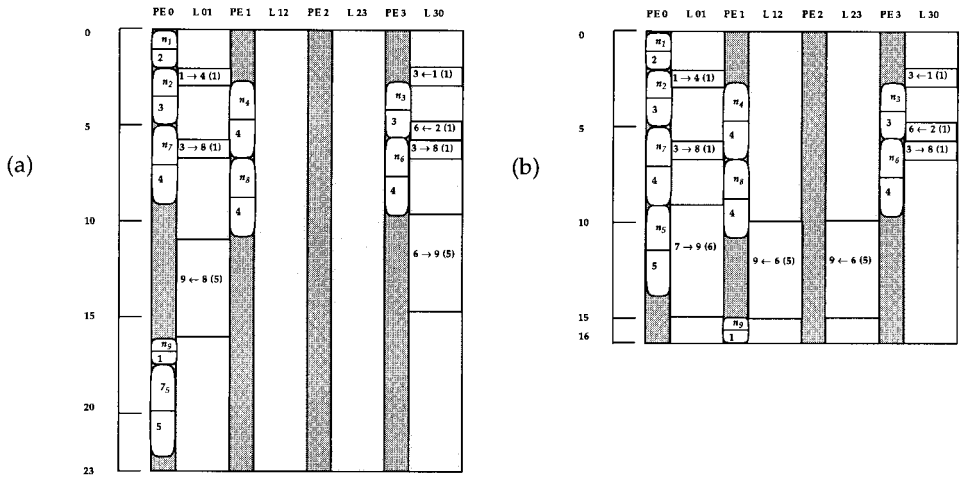


Figure 27. (a) Intermediate schedule produced by BSA after $n6$ migrates from PE 0 to PE 3 (schedule length = 22, total comm. cost = 15); (b) final schedule produced by BSA after n_9 migrates from PE 0 to PE 1 and n_5 is bubbled up (schedule length = 16, total comm. cost = 21).

tions from its parents are scheduled. The priority of a node is modified to be the difference between the static level and the earliest finish-time. During the scheduling of a node, a router is used to determine the best possible path between the processors that need communication. In their simulation study, the improved scheduling algorithm outperformed both the DLS algorithm and the MH algorithm.

6.7 Scheduling in Heterogeneous Environments

Heterogeneity has been shown to be an important attribute in improving the performance of multiprocessors [Ercegovac 1988; Freund and Siegel 1993; Menasce and Almeida 1990; Siegel et al. 1992; Siegel et al. 1996; Wang et al. 1996]. In parallel computations, the serial part is the bottleneck, according to Amdahl's law [Amdahl 1967]. In homogeneous multiprocessors, if one or more faster processors are used to replace a set of cost-equivalent processors, the serial computations and other critical computations can be scheduled to such faster processors and performed at a greater rate so that speedup can be increased.

As we have seen in earlier parts of this section, most DAG scheduling approaches assume the target system is homogeneous. Introducing heterogeneity into the model inevitably makes the problem more complicated to handle. This is because the scheduling algorithm has to take into account the different execution rate of different processors when computing the potential start-times of tasks on the processors. Another complication is that the resulting schedule for a given heterogeneous system immediately becomes invalid if some of the processing elements are replaced even though the number of processors remain the same. This is because the scheduling decisions are made not only on the number of processors but also on the capability of the processors.

Static scheduling targeted for heterogeneous environments was unexplored until recently. Menasce et al. [Menasce and Porto 1993; Menasce et al. 1994; Menasce et al. 1992; Menasce et al. 1995] investigated the problem of sched-

uling computations to heterogeneous multiprocessing environments. The heterogeneous environment was modeled as a system with one fast processor plus a number of slower processors. In their study, both dynamic and static scheduling schemes were examined, but nevertheless DAGs without communication are used to model computations [Almeida et al. 1992]. Markov chains were used to analyze the performance of different scheduling schemes. In their findings, out of all the static scheduling schemes, the LTF/MFT (Largest Task First/Minimizing finish-time) significantly outperformed all the others including WL (Weighted Level), CPM (Critical Path Method) and HNF (Heavy Node First). The LTF/MFT algorithm works by picking the largest task from the ready tasks list and schedules it to the processor which allows the minimum finish-time, while the other three strategies select candidate processors based on the execution time of the task. Thus, based on their observations, an efficient scheduling algorithm for heterogeneous systems should concentrate on reducing the finish-times of tasks. Nonetheless, if communication delays are also considered, different strategies may be needed.

6.8 Mapping Clusters to Processors

As discussed earlier, mapping of clusters to physical processors is necessary for UNC scheduling when the number of clusters is larger than the number of physical processors. However, the mapping of clusters to processors is a relatively unexplored research topic [Lee and Aggarwal 1987]. In the following we discuss a number of approaches reported in the literature.

Upon obtaining a schedule by using the EZ algorithm, Sarkar [1989] used a list-scheduling based method to map the clusters to physical processors. In the mapping algorithm, each task is considered in turn according to the static level. A processor is allocated to the task if it allows the earliest execution, then the whole cluster containing the task is also assigned to that processor and all the member tasks are marked as assigned. In this scheme, two clusters can be merged to a single processor but a cluster is never cracked. Furthermore, the allocation of channels to communication messages was not considered.

Kim and Browne [1988] also proposed a mapping scheme for the UNC schedules obtained from their LC algorithm. In their scheme, the linear UNC clusters are first merged so that the number of clusters is at most the same as the number of processors. Two clusters are candidates for merging if one can start after another finishes, or the member tasks of one cluster can be merged into the idle time slots of another cluster. Then a dominant request tree (DRT) is constructed from the UNC schedule which is a cluster graph. The DRT consists of the connectivity information of the schedule and is, therefore, useful for the mapping stage in which two communicating UNC clusters attempt to be mapped to two neighboring processors, if possible. However, if for some clusters this connectivity mapping heuristic fails, another two heuristics called perturbation mapping and foster mapping are invoked. For both mapping strategies, a processor is chosen which has the most appropriate number of channels among currently unallocated processors. Finally, to further optimize the mapping, a restricted pairwise exchange step is called for.

Wu and Gajski [1990] suggested a mapping scheme for assigning the UNC clusters generated in scheduling to processors. They realized that for best mapping results, a dedicated traffic scheduling algorithm that balances the network traffic should be used. However, traffic scheduling requires flexible-path routing, which incurs higher overhead. Thus, they concluded that if network traffic is not heavy, a simpler algorithm which minimizes total network traffic can be used. The algorithm they used is a heuristic algorithm designed by Hanan and Kurtzberg [1972]

to minimize the total communication traffic. The algorithm generates an initial assignment by a constructive method and the assignment is then iteratively improved to obtain a better mapping.

Young and Gerasoulis [1993] employed a *work profiling method* for merging UNC clusters. The merging process proceeds by first sorting the clusters in an increasing order of aggregate computational load. Then a load-balancing algorithm is invoked to map the clusters to the processors so that every processor has about the same load. To take care of the topology of the underlying processor network, the graph of merged clusters are then mapped to the network topology using Bokhari's algorithm.

Yang et al. [1993] reported an algorithm for mapping cluster graphs to processor graphs which is suitable for use as the post-processing step for BNP scheduling algorithms. The mapping scheme is not suitable for UNC scheduling because it assumes the scheduling algorithm has already produced a number of clusters which is less than or equal to the number of processors available. The objective of the mapping method is to reduce contention and optimize the schedule length when the clusters are mapped to a topology which is not fully connected as assumed by the BNP algorithms. The idea of the mapping algorithm is based on determining a set of critical edges, each of which is assigned a single communication link. Substantial improvement over random mapping was obtained in their simulation study.

In a recent study, Liou and Palis [1997] investigated the problem of mapping clusters to processors. One of the major objectives of their study was to compare the effectiveness of one-phase scheduling (i.e., BNP scheduling) to that of the two-phase approach (i.e., UNC scheduling followed by clusters mapping). To this end, they proposed a new UNC algorithm called CASS-II (Clustering And Scheduling System II), which was applied to randomly generated task graphs in an experimental study using three clusters mapping schemes, namely, the LB (load-balancing) algorithm, the CTM (communication traffic minimizing) algorithm, and the RAND (random) algorithm. The LB algorithm uses processor workload as the criterion of matching clusters to processors. By contrast, the CTM algorithm tries to minimize the communication costs between processors. The RAND algorithm simply makes random choices at each mapping step. To compare the one-phase method with the two-phase method, in one set of test cases the task graphs were scheduled using CASS-II with the three mapping algorithms while in the other set using the mapping algorithms alone. Liou and Palis found that two-phase scheduling is better than one-phase scheduling whereas the utilization of processors in the former is more efficient than the latter. Furthermore, they found that the LB algorithm finds significantly better schedules than the CTM algorithm.

7. SOME SCHEDULING TOOLS

Software tools providing automated functionalities for scheduling/mapping can make the parallel programming task easier. Despite a vast volume of research on scheduling that exists, building useful scheduling tools is only recently addressed. A scheduling tool should allow a programmer to specify a parallel program in certain textual or graphical form, and then perform automatic partitioning and scheduling of the program. The tool should also allow the user to specify the target architecture. Performance evaluation and debugging functions are also highly desirable. Some tools provide interactive environments for performance evaluation of various popular parallel machines but do not generate an executable scheduled code [Pease et al. 1991]. Under the above definition, such tools provide other functionalities but cannot be classified as scheduling tools.

In the following, we survey some of the recently reported prototype scheduling tools.

7.1 Hypertool

Hypertool takes a user-partitioned sequential program as input and automatically allocates and schedules the partitions to processors [Wu and Gajski 1990]. Proper synchronization primitives are also automatically inserted. Hypertool is a code generation tool since the user program is compiled into a parallel program for the iPSC/2 hypercube computer using parallel code synthesis and optimization techniques. The tool also generates performance estimates including execution time, communication and suspension times for each processor as well as network delay for each communication channel. Scheduling is done using the MD algorithm or the MCP algorithm.

7.2 PYRROS

PYRROS is a compile-time scheduling and code generation tool [Yang and Gerasoulis 1992]. Its input is a task graph and the associated sequential C code. The output is a static schedule and a parallel C code for a given architecture (the iPSC/2). PYRROS consists of a task graph language with an interface to C, a scheduling system which uses only the DSC algorithm, a X-Windows based graphic displayer, and a code generator. The task graph language allows the user to define partitioned programs and data. The scheduling system is used for clustering the task graph, performing load balanced mapping, and computation/communication ordering. The graphic displayer is used for displaying task graphs and scheduling results in the form of Gantt charts. The code generator inserts synchronization primitives and performs parallel code optimization for the target parallel machine.

7.3 Parallax

Parallax incorporates seven classical scheduling heuristics designed in the seventies [Lewis and El-Rewini 1993], providing an environment for parallel program developers to find out how the schedulers affect program performance on various parallel architectures. Users must provide the input program as a task graph and estimate task execution times. Users must also express the target machine as an interconnection topology graph. Parallax then generates schedules in the form of Gantt charts, speedup curves, and processor and communication efficiency charts using X-Windows interface. In addition, an animated display of the simulated running program to help developers to evaluate the differences among the scheduling heuristics is provided. Parallex, however, is not reported to generate an executable code.

7.4 OREGAMI

OREGAMI is designed for use in conjunction with parallel programming languages that support a communication model, such as OCCAM, C*, or C and FORTRAN with communication extension [Lo et al. 1991]. As such, it is a set of tools that includes a LaRCS compiler to compile textual user task descriptions into specialized task graphs, which are called TCG (Temporal Communication Graphs) [Lo 1992]. In addition, OREGAMI includes a mapper tool for mapping tasks on a variety of target architectures, and a metrics tools for analyzing and displaying the performance. The suite of tools are implemented in C for SUN workstations with an X-Windows interface. However, precedence constraints among tasks are not considered in OREGAMI. Moreover, no target code is generated. Thus, like Parallax, OREGAMI is rather a design tool for parallel program development.

7.5 PARSA

PARSA is a software tool developed for automatic scheduling and partitioning of sequential user programs [Shirazi et al. 1993]. PARSA consists of four components: an application specification tool, an architecture specification tool, a partitioning and scheduling tool, and a performance assessment tool. PARSA does not generate any target code. The application specification tool accepts a sequential program written in the SISAL functional language and converts it into a DAG, which is represented in textual form by the IF1 (Intermediate Form 1) acyclic graphical language. The architecture specification tool allows the user to interactively specify the target system in graphical form. The execution delay for each task is also generated based on the architecture specification. The partitioning and scheduling tool consists of the HNF algorithm, the LC algorithm, and the LCTD algorithm. The performance assessment tool displays the expected runtime behavior of the scheduled program. The expected performance is generated by the analysis of the scheduled program trace file, which contains the information on where each task is assigned for execution and exactly where each task is expected to start execution, stop execution, or send a message to a remote task.

7.6 CASCH

CASCH(Computer- Aided SCHeduling) tool [Ahmad et al. 1997] is aimed to be a complete parallel programming environment including parallelization, partitioning, scheduling, mapping, communication, synchronization, code generation, and performance evaluation. Parallelization is performed by a compiler that automatically converts sequential applications into parallel codes. The parallel code is optimized through proper scheduling and mapping, and is executed on a target machine. CASCH provides an extensive library of state-of-the-art scheduling algorithms from the recent literature. The library of scheduling algorithms is organized into different categories which are suitable for different architectural environments.

The scheduling and mapping algorithms are used for scheduling the task graph generated from the user program, which can be created interactively or read from disk. The weights on the nodes and edges of the task graph are computed using a database that contains the timing of various computation, communication, and I/O operations for different machines. These timings are obtained through benchmarking. An attractive feature of CASCH is its easy-to-use GUI for analyzing various scheduling and mapping algorithms using task graphs generated randomly, interactively, or directly from real programs. The best schedule generated by an algorithm can be used by the code generator for generating a parallel program for a particular machine–and the same process can be repeated for another machine.

7.7 Commercial Tools

There are only a few commercially available tools for scheduling and program parallelization. Examples include ATEXPERT by Cray Research [1991]; PARASPHERE by DEC [Digital Equipment Corp.]; IPD by Intel [1991]; MPPE by MasPar [1992]; and PRISM by TMC [1991]. Most of these tools provide software development and debugging environments. Some of them also provide performance tuning tools and other program development facilities.

8. NEW IDEAS AND RESEARCH TRENDS

With the advancements in processors and networking hardware technologies, parallel processing can be accomplished in a wide spectrum of platforms ranging from tightly-coupled MPPs to a loosely-coupled network of autonomous workstations. Designing an algorithm for

such diverse platforms makes the scheduling problem even more complex and challenging. In summary, in the design of scheduling algorithms for efficient parallel processing, we have to address four fundamental aspects: performance, time-complexity, scalability, and applicability. These aspects are elaborated below.

Performance: A scheduling algorithm must exhibit high performance and be robust. By high performance we mean the scheduling algorithm should produce high quality solutions. A robust algorithm is one which can be used under a wide range of input parameters (e.g., arbitrary number of available processors and diverse task graph structures).

Time-complexity: The time-complexity of an algorithm is an important factor insofar as the quality of solution is not compromised. As real workload is typically of a large size [Ahmad et al. 1997], a fast algorithm is necessary for finding good solutions efficiently.

Scalability: A scheduling algorithm must possess problem-size scalability, that is, the algorithm has to consistently give good performance even for large input. On the other hand, a scheduling algorithm must also possess processing-power scalability, that is, given more processors for a problem, the algorithm should produce solutions with comparable quality in a shorter period of time.

Applicability: A scheduling algorithm must be applicable in practical environments. To achieve this goal, it must take into account realistic assumptions about the program and multiprocessor models such as arbitrary computation and communication weights, link contention, and processor network topology.

It is clear that the above mentioned goals are conflicting and thus pose a number of challenges to researchers. To combat these challenges, several new ideas have been suggested recently. These new ideas, which include genetic algorithms, randomization approaches, and parallelization techniques, are employed to enhance the solution quality, or to lower the time-complexity, or both. In the following, we briefly outline some of these recent advancements. At the end of this section, we also indicate some current research trends in scheduling.

8.1 Scheduling Using Genetic Algorithms

Genetic algorithms (GAs) [Davis 1991; Filho et al. 1994; Forrest and Mitchell 1993; Goldberg 1989; Holland 1975; Srinivas and Patnaik 1994] have recently found many applications in optimization problems including scheduling [Ali et al. 1994; Benten and Sait 1994; Chandrasekharam 1993; Dhodhi et al. 1995; Hou et al. 1994; Schwehm et al. 1994]. GAs use global search techniques to explore different regions of the search space simultaneously by keeping track of a set of potential solutions of diverse characteristics, called a population. As such, GAs are widely recognized as effective techniques in solving numerous optimization problems, because they can potentially locate better solutions at the expense of longer running time. Another merit of a genetic search is that its inherent parallelism can be exploited so as to further reduce its running time. Thus, a parallel genetic search technique in scheduling is a viable approach in producing high quality solutions using short running times.

Ali et al. [1994] proposed a genetic algorithm for scheduling a DAG to a limited number of fully connected processors with a contention-free communication network. In their scheme, each solution or schedule is encoded as a chromosome containing v alleles, each of which is an ordered pair of task index and its assigned processor index. With such encoding the design of genetic operators is straightforward. Standard crossover is used because it always produces valid schedules as offsprings and is computationally efficient. Mutation is simply a swapping of the assigned processors between two randomly chosen

alleles. For generating an initial population, Ali et al. use a technique called "prescheduling" in which N_p random permutations of numbers from 1 to v are generated. The number in each random permutation represents the task index of the task graph. The tasks are then assigned to the PEs uniformly: the first v/p tasks in a permutation are assigned to PE 0, the next v/p tasks to PE 1, and so on. In their simulation study using randomly generated task graphs with a few tenths of nodes, their algorithm was shown to outperform the ETF algorithm proposed by Hwang et al. [1989].

Hou et al. [1994] also proposed a scheduling algorithm using a genetic search in which each chromosome is a collection of lists, and each list represents the schedule on a distinct processor. Thus, each chromosome is not a linear structure but a two-dimensional structure instead. One dimension is a particular processor index and the other is the ordering of tasks scheduled on the processor. Using such an encoding scheme poses a restriction on the schedules being represented: the list of tasks within each processor in a schedule is ordered in ascending order of their topological height, which is defined as the largest number of edges from an entry node to the node itself. This restriction also facilitates the design of the crossover operator. In a crossover, two processors are selected from each of two chromosomes. The list of tasks on each processor is cut into two parts, and then the two chromosomes exchange the two lower parts of their task lists correspondingly. It is shown that this crossover mechanism always produces valid offsprings. However, the height restriction in the encoding may cause the search to be incapable of obtaining the optimal solution because the optimal solution may not obey the height ordering restriction at all.

Hou et al. incorporated a heuristic technique to lower the likelihood of such a pathological situation. Mutation is simpler in design. In a mutation, two randomly chosen tasks with the same height are swapped in the schedule. As to the generation of the initial population, N_p randomly permuted schedules obeying the height ordering restriction are generated. In their simulation study using randomly generated task graphs with a few tenths of nodes, their algorithm was shown to produce schedules within 20 percent degradation from optimal solutions.

Ahmad and Dhodhi [1995] proposed a scheduling algorithm using a variant of genetic algorithm called simulated evolution. They employ a problem-space neighborhood formulation in that a chromosome represents a list of task priorities. Since task priorities are dependent on the input DAG, different sets of task priorities represent different problem instances. First, a list of priorities is obtained from the input DAG. Then the initial population of chromosomes are generated by randomly perturbing this original list. Standard genetic operators are applied to these chromosomes to determine the fittest chromosome, which is the one giving the shortest schedule length for the original problem. The genetic search, therefore, operates on the problem-space instead of the solution-space, as is commonly done. The rationale of this approach is that good solutions of the problem instances in the problem-space neighborhood are expected to be good solutions for the original problem as well [Storer et al. 1992].

Recently, we have proposed a parallel genetic algorithm for scheduling [Kwok and Ahmad 1997], called the Parallel Genetic Scheduling (PGS) algorithm, using a novel encoding scheme, an effective initial population generation strategy, and computationally efficient genetic search operators. The major motivation of using a genetic search approach is that the recombinative nature of a genetic algorithm can potentially determine an optimal scheduling list leading to an optimal schedule. As such,

a scheduling list (i.e., a topological ordering of the input DAG) is encoded as a genetic string. Instead of generating the initial population totally randomly, we generate the initial set of strings based on a number of effective scheduling lists such as ALAP list, *b-level* list, *t-level* list, etc. We use a novel crossover operator, which is a variant of the order crossover operator, in the scheduling context. The proposed crossover operator has the potential to effectively combine the good characteristics of two parent strings in order to generate a scheduling string leading to a schedule with shorter schedule length. The crossover operator is easy to implement and is computationally efficient.

In our experimental studies [Kwok and Ahmad 1997], we have found that the PGS algorithm generates optimal solutions for more than half of all the cases in which random task graphs were used. In addition, the PGS algorithm demonstrates an almost linear speedup, and is therefore scalable. While the DCP algorithm [Kwok and Ahmad 1996] has already been shown to outperform many leading algorithms, the PGS algorithm is even better, since it generates solutions with comparable quality while using significantly less time due to its effective parallelization. The PGS algorithm outperforms the well-known DSC algorithm in terms of both the solution quality and running time. An extra advantage of the PGS algorithm is scalability, that is by using more parallel processors, the algorithm can be used for scheduling larger task graphs.

8.2 Randomization Techniques

The time-complexity of an algorithm and its solution quality are in general conflicting goals in the design of efficient scheduling algorithms. Our previous study [Kwok and Ahmad 1999b] indicates that not only does the quality of existing algorithms differ considerably but their running times can vary by large margins. Indeed, designing an algorithm which is fast and can produce high quality solutions requires some low-complexity algorithmic techniques. One promising approach is to employ randomization. As indicated by Karp [1991], Motwani and Raghavan [1995], and other researchers, an optimization algorithm which makes random choices can be very fast and simple to implement. However, there has been very little work done in this direction.

Recently, we [Kwok 1997; Kwok and Ahmad 1999a; Kwok et al. 1996] proposed a BNP scheduling algorithm based on a random neighborhood search technique [Johnson et al. 1988; Papadimitriou and Steiglitz 1982]. The algorithm is called the Fast Assignment and Scheduling of Tasks using an Efficient Search Technique (FASTEST) algorithm, which has a time-complexity of only $O(e)$, where e is the number of edges in the DAG [Kwok and Ahmad 1999a]. The FASTEST algorithm first constructs an initial schedule quickly in linear-time and then refines it by using multiple physical processors, each of which operates on a disjoint subset of blocking-nodes as a search neighborhood. The physical processors communicate periodically to exchange the best solution found thus far. As the number of search steps required is a small constant, which is independent of the size of the input DAG, the algorithm effectively takes linear-time to determine the final schedule.

In our performance study [Kwok 1997; Kwok and Ahmad 1999a], we compared the FASTEST algorithm with a number of well-known efficient scheduling algorithms. The FASTEST algorithm has been shown to be better than the other algorithms in terms of both solution quality and running time. Since the algorithm takes linear-time, it is the fastest algorithm to our knowledge. In experiments using random task graphs for which optimal solutions are known, the FASTEST algorithm generates optimal solutions for a significant portion of all the test cases, and close-

to-optimal solutions for the remaining cases. The FASTEST algorithm also exhibits good scalability in that it gives a consistent performance when applied to large task graphs. An interesting finding of the FASTEST algorithm is that parallelization can sometimes improve its solution quality. This is due to the partitioning of the blocking-nodes set, which implies a partitioning of the search neighborhood. The partitioning allows the algorithm to explore the search space simultaneously, thereby enhancing the likelihood of getting better solutions.

8.3 Parallelizing a Scheduling Algorithm

Parallelizing a scheduling algorithm is a novel as well as natural way to reduce the time-complexity. This approach is novel in that no previous work has been done in the parallelization of a scheduling algorithm. Indeed, as indicated by Norman and Thanisch [1993], it is strange that there has been hardly any attempt to parallelize a scheduling and mapping process itself. Parallelization is natural in that parallel processing is realized only when a parallel processing platform is available. Furthermore, parallelization can be utilized not only to speed up the scheduling process further but also to improve the solution quality. Recently there have been a few parallel algorithms proposed for DAG scheduling [Ahmad and Kwok 1999; Kwok 1997; Kwok and Ahmad 1997].

In a recent study [Ahmad and Kwok 1998b], we have proposed two parallel state-space search algorithms for finding optimal or bounded solutions. The first algorithm which is based on the A* search technique uses a computationally efficient cost function for quickly guiding the search. The A* algorithm is also parallelized, using static and dynamic load-balancing schemes to evenly distribute the search states to the processors. A number of effective state-pruning techniques are also incorporated to further enhance the efficiency of the algorithm. The proposed algorithm outperforms a previously reported branch-and-bound algorithm by using considerable less computation time. The second algorithm is an approximate algorithm that guarantees a bounded deviation from the optimal solution, but executes in a considerably shorter turnaround time. Based on both theoretical analysis and experimental evaluation [Ahmad and Kwok 1998b] using randomly generated task graphs, we have found that the approximate algorithm is highly scalable and is an attractive choice, if slightly degraded solutions are acceptable.

We have also proposed [Ahmad and Kwok 1999; Kwok 1997] a parallel APN scheduling algorithm called the Parallel Bubble Scheduling and Allocation (PBSA) algorithm. The proposed PBSA algorithm is based on considerations such as a limited number of processors, link contention, heterogeneity of processors, and processor network topology. As a result, the algorithm is useful for distributed systems including clusters of workstations. The major strength of the PBSA algorithm lies in its incremental strategy of scheduling nodes and messages together. It first uses the CPN-Dominant sequence to serialize the task graph to one PE, and then allows the nodes to migrate to other PEs for improving their start-times. In this manner, the start-times of the nodes, and hence, the schedule length, are optimized incrementally. Furthermore, in the course of migration, the routing and scheduling of communication messages between tasks are also optimized. The PBSA algorithm first partitions the input DAG into a number of disjoint subgraphs. The subgraphs are then scheduled independently in multiple physical processors, each of which runs a sequential BSA algorithm. The final schedule is constructed by concatenating the subschedules produced. The proposed algorithm is, therefore, the first attempt of its kind in that it is a parallel algorithm and it also solves the scheduling problem by con-

sidering all the important scheduling parameters.

We have evaluated the PBSA algorithm [Ahmad and Kwok 1999; Kwok 1997] by testing it in experiments using extensive variations of input parameters including graph types, graph sizes, CCRs, and target network topologies. Comparisons with three other APN scheduling algorithms have also been made. Based on the experimental results, we find that the PBSA algorithm can provide a scalable schedule, and can be useful for scheduling large task graphs which are virtually impossible to schedule using sequential algorithms. Furthermore, the PBSA algorithm exhibits superlinear speedup in that given q physical processors, the algorithm can produce solutions with comparable quality with a speedup of roughly $O(q^2)$ over the sequential case.

Other researchers have also suggested techniques for some restricted forms of the scheduling problem. Recently, Pramanick and Kuhl [1995] proposed a paradigm, called Parallel Dynamic Interaction (PDI), for developing parallel search algorithms for many NP-hard optimization problems. The PDI method is applied to the job-shop scheduling problem in which a set of independent jobs are scheduled to homogeneous machines. De Falco et al. [1997] have suggested using parallel simulated annealing and parallel tabu search algorithms for the task allocation problem, in which a Task Interaction Graph (TIG), representing communicating processes in a distributed systems, is to be mapped to homogeneous processors. As mentioned earlier, a TIG is different from a DAG in that the former is an undirected graph with no precedence constraints among the tasks. Parallel branch-and-bound techniques [Ferreira and Pardalos 1996] have also been used to tackle some simplified scheduling problems.

8.4 Future Research Directions

Research in DAG scheduling can be extended in several directions. One of the most challenging directions is to extend DAG scheduling to heterogeneous computing platforms. Heterogeneous computing (HC), using physically distributed diverse machines connected via a high-speed network for solving complex applications, is likely to dominate the next era of high-performance computing. One class of HC environment is a suite of sequential machines known as a network of workstations (NOWs). Another class, known as the distributed heterogeneous supercomputing system (DHSS), is a suite of machines comprising a variety of sequential and parallel computers–providing an even higher level of parallelism. In general, it is impossible for a single machine architecture with its associated compiler, operating system, and programming tools to satisfy all the computational requirements in an application equally well. However, a heterogeneous computing environment that consists of a heterogeneous suite of machines, high-speed interconnections, interfaces, operating systems, communication protocols and programming environments provides a variety of architectural capabilities, which can be orchestrated to perform an application that has diverse execution requirements. Due to the latest advances in networking technologies, HC is likely to flourish in the near future.

The goal of HC using a NOW or a DHSS is to achieve the minimum completion time for an application. A challenging future research problem is to design efficient algorithms for scheduling and mapping of applications to the machines in a HC environment. Task-to-machine mapping in a HC environment is beyond doubt more complicated than in a homogeneous environment. In a HC environment, a computation can be decomposed into tasks, each of which may have substantially different processing requirements. For example a signal processing task may strictly re-

quire a machine possessing DSP processing capability. While the PBSA algorithm proposed [Ahmad and Kwok 1999] is a first step toward this direction, more work is needed. One possible research direction is to first devise a new model of heterogeneous parallel applications as well as new models of HC environments. Based on these new models, more optimized algorithms can be designed.

Another avenue of further research is to extend the applicability of the existing randomization and evolutionary scheduling algorithms [Ali et al. 1994; Hou et al. 1994; Kwok 1997]. While they are targeted to be used in BNP scheduling, the algorithms may be extended to handle APN scheduling as well. However, some novel efficient algorithmic techniques for scheduling messages to links need to be sought, lest the time-complexity of the randomization algorithms increase. Further improvements in the genetic and evolutionary algorithms may be possible if we can determine an optimal set of control parameters, including crossover rate, mutation rate, population size, number of generations, and number of parallel processors used. However, finding an optimal parameters set for a particular genetic algorithm is hitherto an open research problem.

9. SUMMARY AND CONCLUDING REMARKS

In this paper, we have presented an extensive survey of algorithms for the static scheduling problem. Processors and communication links are in general the most important resources in parallel and distributed systems, and their efficient management through proper scheduling is essential for obtaining high performance. We first introduced the DAG model and the multiprocessor model, followed by the problem statement of scheduling. In the DAG model, a node denotes an atomic program task, and an edge denotes the communication and data dependency between two program tasks. Each node is labeled a weight denoting the amount of computational time required by the task. Each edge is also labeled a weight denoting the amount of communication time required. The target multiprocessor systems is modeled as a network of processing elements (PEs), each of which comprises a processor and a local memory unit, so that communication is achieved solely by message-passing. The objective of scheduling is to minimize the schedule length by properly allocating the nodes to the PEs and sequencing their start-times so that the precedence constraints are preserved.

We have also presented a scrutiny of the NP-completeness results of various simplified variants of the problem, thereby illustrating that static scheduling is a hard optimization problem. As the problem is intractable even for moderately general cases, heuristic approaches are commonly sought.

To better understand the design of the heuristic scheduling schemes, we have also described and explained a set of basic techniques used in most algorithms. With these techniques the task graph structure is carefully exploited to determine the relative importance of the nodes in the graph. More important nodes get a higher consideration priority for scheduling first. An important structure in a task graph is the critical path (CP). The nodes of the CP can be identified by the nodes' *b-level* and *t-level*. In order to put the representative work with different assumptions reported in the literature in a unified framework, we described a taxonomy of scheduling algorithms which classifies the algorithms into four categories: the UNC (unbounded number of clusters) scheduling, the BNP (bounded number of processors) scheduling, the TDB (task duplication based) scheduling, and APN (arbitrary processor network) scheduling. Analytical results as well as scheduling examples have been shown to illustrate the functionality and characteristics of the surveyed algorithms. Tasks scheduling for heterogeneous systems,

which are widely considered as promising platforms for high-performance computing, is briefly discussed. As a postprocessing step of some scheduling algorithms, the mapping process is also examined. Various experimental software tools for scheduling and mapping are also described.

Finally, we have surveyed a number of new techniques which are recently proposed for achieving these goals. These techniques include genetic and evolutionary algorithms, randomization techniques, and parallelized scheduling approaches.

ACKNOWLEDGMENT

The authors would like to thank the referees for their comments.

REFERENCES

ADAM, T. L., CHANDY, K. M., AND DICKSON, J. R. 1974. A comparison of list scheduling for parallel processing systems. *Commun. ACM 17*, 12 (Dec.), 685–690.

AHMAD, I. AND DHODHI, M. K. 1995. Task assignment using a problem-space genetic algorithm. *Concurrency: Pract. Exper. 7*, 5 (Aug.), 411–428.

AHMAD, I. AND GHAFOOR, A. 1991. Semi-distributed load balancing for massively parallel multicomputer systems. *IEEE Trans. Softw. Eng. 17*, 10 (Oct. 1991), 987–1004.

AHMAD, I. AND KWOK, Y.-K. 1998a. On exploiting task duplication in parallel program scheduling. *IEEE Trans. Parallel Distrib. Syst. 9*, 9, 872–892.

AHMAD, I. AND KWOK, Y.-K. 1998b. Optimal and near-optimal allocation of precedence-constrained task to parallel processors: Defying the high complexity using effective search technique. In *Proceedings of the 1998 International Conference on Parallel Processing* (Aug.),

AHMAD, I. AND KWOK, Y.-K. 1999. On parallelizing the multiprocessor scheduling problem. *IEEE Trans. Parallel Distrib. Syst. 10*, 4 (Apr.), 414–432.

AHMAD, I., KWOK, Y.-K., AND WU, M.-Y. 1996. Analysis, evaluation, and comparison of algorithms for scheduling task graphs on parallel processors. In *International Symposium on Parallel Architectures, Algorithms, and Networks* (June), 207–213.

AHMAD, I., KWOK, Y.-K., WU, M.-Y., AND SHU, W. 1997. Automatic parallelization and scheduling of programs on multiprocessors using CASCH. In *Proceedings of the International Conference on Parallel Processing* (ICPP, Aug.), 288–291.

ALI, H. H. AND EL-REWINI, H. 1993. The time complexity of scheduling interval orders with communication is polynomial. *Para. Proc. Lett. 3*, 1, 53–58.

ALI, S., SAIT, S. M., AND BENTEN, M. S. T. 1994. GSA: Scheduling and allocation using genetic algorithm. In *Proceedings of the Conference on EURO-DAC'94*, 84–89.

AL-MOUHAMED, M. A. 1990. Lower bound on the number of processors and time for scheduling precedence graphs with communication costs. *IEEE Trans. Softw. Eng. 16*, 12 (Dec. 1990), 1390–1401.

ALMEIDA, V. A. F., VASCONCELOS, I. M. M., ÁRABE, J. N. C., AND MENASCÉ, D. A. 1992. Using random task graphs to investigate the potential benefits of heterogeneity in parallel systems. In *Proceedings of the 1992 Conference on Supercomputing* (Supercomputing '92, Minneapolis, MN, Nov. 16–20), R. Werner, Ed. IEEE Computer Society Press, Los Alamitos, CA, 683–691.

AMDAHL, G. 1967. Validity of the single processor approach to achieving large scale computing capability. In *Proceedings of the on AFIPS Spring Joint Computer Conference* (Reston, Va.), AFIPS Press, Arlington, VA, 483–485.

ANGER, F. D., HWANG, J.-J., AND CHOW, Y.-C. 1990. Scheduling with sufficient loosely coupled processors. *J. Parallel Distrib. Comput. 9*, 1 (May 1990), 87–92.

BASHIR, A. F., SUSARLA, V., AND VAIRAVAN, K. 1983. A statistical study of the performance of a task scheduling algorithm. *IEEE Trans. Comput. C-32*, 8 (Aug.), 774–777.

BAXTER, J. AND PATEL, J. H. 1989. The LAST algorithm: A heuristic-based static task allocation algorithm. In *Proceedings of the International Conference on Parallel Processing* (ICPP '89, Aug.), Pennsylvania State University, University Park, PA, 217–222.

BECK, M., PINGALI, K., AND NICOLAU, A. 1990. Static scheduling for dynamic dataflow machines. *J. Parallel Distrib. Comput. 10*, 4 (Dec. 1990), 279–288.

BENTEN, M. S. T. AND SAIT, S. M. 1994. Genetic scheduling of task graphs. *Int. J. Electron. 77*, 4 (Oct.), 401–415.

BLAZEWICZ, J., DRABOWSKI, M., AND WEGLARZ, J. 1986. Scheduling multiprocessor tasks to minimize schedule length. *IEEE Trans. Comput. C-35*, 5 (May 1986), 389–393.

BLAZEWICZ, J., WEGLARZ, J., AND DRABOWSKI, M. 1984. Scheduling independent 2-processor tasks to minimize schedule length. *Inf. Process. Lett. 18*, 5 (June 1984), 267–273.

BOKHARI, S. H. 1979. Dual processor scheduling with dynamic reassignment. *IEEE Trans. Softw. Eng. SE-5*, 4 (July), 341–349.

BOKHARI, S. H. 1981. On the mapping problem. *IEEE Trans. Comput. C-30*, 5, 207–214.

BOZOKI, G. AND RICHARD, J. P. 1970. A branch-and-bound algorithm for continuous-process

task shop scheduling problem. *AIIE Trans. 2*, 246–252.

BRUNO, J., COFFMAN, E. G., AND SETHI, R. 1974. Scheduling independent tasks to reduce mean finishing time. *Commun. ACM 17*, 7 (July), 382–387.

CASAVANT, T. L. AND KUHL, J. G. 1988. A taxonomy of scheduling in general-purpose distrbuted computing systems. *IEEE Trans. Softw. Eng. 14*, 2 (Feb.), 141–154.

CHANDRASEKHARAM, R., SUBHRAMANIAN, S., AND CHAUDHURY, S. 1993. Genetic algorithm for node partitioning problem and applications in VLSI design. *IEE Proc. Comput. Digit. Tech. 140*, 5 (Sept.), 255–260.

CHEN, G. AND LAI, T. H. 1988a. Scheduling independent jobs on hypercubes. In *Proceedings of the Conference on Theoretical Aspects of Computer Science*, 273–280.

CHEN, G.-I. AND LAI, T.-H. 1988b. Preemptive scheduling of independent jobs on a hypercube. *Inf. Process. Lett. 28*, 4 (July 29, 1988), 201–206.

CHEN, H., SHIRAZI, B., AND MARQUIS, J. 1993. Performance evaluation of a novel scheduling method: Linear clustering with task duplication. In *Proceedings of the 2nd International Conference on Parallel and Distributed Systems* (Dec.), 270–275.

CHENG, R., GEN, M., AND TSUJIMURA, Y. 1996. A tutorial survey of job-shop scheduling problems using genetic algorithms—I: representation. *Comput. Ind. Eng. 30*, 4, 983–997.

CHRETIENNE, P. 1989. A polynomial algorithm to optimally schedule tasks on a virtual distributed system under tree-like precedence constraints. *Europ. J. Oper. Res. 43*, 225–230.

CHU, W. W., LAN, M.-T., AND HELLERSTEIN, J. 1984. Estimation of intermodule communication (IMC) and its applications in distributed processing systems. *IEEE Trans. Comput. C-33*, 8 (Aug.), 691–699.

CHUNG, Y.-C. AND RANKA, S. 1992. Applications and performance analysis of a compile-time optimization approach for list scheduling algorithms on distributed memory multiprocessors. In *Proceedings of the 1992 Conference on Supercomputing* (Supercomputing '92, Minneapolis, MN, Nov. 16–20), R. Werner, Ed. IEEE Computer Society Press, Los Alamitos, CA, 512–521.

COFFMAN, E. G. 1976. *Computer and Job-Shop Scheduling Theory*. John Wiley and Sons, Inc., New York, NY.

COFFMAN, E. G. AND GRAHAM, R. L. 1972. Optimal scheduling for two-processor systems. *Acta Inf. 1*, 200–213.

COLIN, J. Y. AND CHRETIENNE, P. 1991. C.P.M. scheduling with small computation delays and task duplication. *Oper. Res. 39*, 4, 680–684.

COSNARD, M. AND LOI, M. 1995. Automatic task graph generation techniques. *Para. Proc. Lett. 5*, 4 (Dec.), 527–538.

CRAY RESEARCH, INC. 1991. UNICOS Performance Utilities Reference Manual, SR2040. Cray Supercomputers, Chippewa Falls, MN.

DALLY, W. J. 1992. Virtual-channel flow control. *IEEE Trans. Parallel Distrib. Syst. 3*, 3 (Mar.), 194–205.

DARBHA, S. AND AGARWAL, D. P. 1995. A fast and scalable scheduling algorithm for distrbuted memory systems. In *Proceedings of 7th Symposium on Parallel and Distributed Processing* (Oct.), 60–63.

DAVIS, T., Ed. 1991. *The Handbook of Genetic Algorithms*. Van Nostrand Reinhold Co., New York, NY.

DE FALCO, I., DEL BALIO, R., AND TARANTINO, E. 1997. An analysis of parallel heuristics for task allocation in multicomputers. *Computing 59*, 3, 259–275.

DHODI, M. K., AHMAD, I., AND AHMAD, I. 1995. A multiprocessor scheduling scheme using problem-space genetic algorithms. In *Proceedings of the IEEE International Conference on Evolutionary Computation*, IEEE Computer Society Press, Los Alamitos, CA, 214–219.

DIGITAL EQUIPMENT CORP. 1992. PARASPHERE User's Guide. Digital Equipment Corp., Maynard, MA.

DIXIT-RADYA, V. A. AND PANDA, D. K. 1993. Task assignment on distrbuted-memory systems with adaptive wormhole routing. In *Proceedings of the 2nd International Conference on Parallel and Distributed Systems* (Dec.), 674–681.

DU, J. AND LEUNG, J. Y.-T. 1989. Complexity of scheduling parallel task systems. *SIAM J. Discrete Math. 2*, 4 (Nov. 1989), 473–487.

EL-REWINI, H. AND ALI, H. H. 1995. Static scheduling of conditional branches in parallel programs. *J. Parallel Distrib. Comput. 24*, 1 (Jan. 1995), 41–54.

EL-REWINI, H., ALI, H. H., AND LEWIS, T. G. 1995. Task scheduling in multiprocessor systems. *IEEE Computer 28*, 12 (Dec.), 27–37.

EL-REWINI, H. AND LEWIS, T. G. 1990. Scheduling parallel program tasks onto arbitrary target machines. *J. Parallel Distrib. Comput. 9*, 2 (June 1990), 138–153.

EL-REWINI, H., LEWIS, T. G., AND ALI, H. H. 1994. *Task scheduling in parallel and distributed systems*. Prentice-Hall series in innovative technology. Prentice-Hall, Inc., Upper Saddle River, NJ.

ERCEGOVAC, M. D. 1988. Heterogeneity in supercomputer architectures. *Parallel Comput. 7*, 367–372.

FERNANDEZ, E. B. AND BUSSELL, B. 1973. Bounds on the number of processors and time for multiprocessor optimal schedules. *IEEE Trans. Comput. C-22*, 8 (Aug.), 745–751.

FERREIRA, A. AND PARDALOS, P., Eds. 1996. *Solving Combinatorial Optimization Problems in Parallel: Methods and Techniques*. Lecture Notes in Computer Science, vol. 1054.. Springer-Verlag, New York, NY.

FILHO, J. L. R., TRELEAVEN, P. C., AND ALIPPI, C. 1994. Genetic-algorithm programming environments. *IEEE Computer 27*, 6 (June 1994), 28–43.

FISHBURN, P. C. 1985. *Interval Orders and Interval Graphs*. John Wiley and Sons, Inc., New York, NY.

FORREST, S. AND MITCHELL, M. 1993. What makes a problem hard for a genetic algorithm?: some anomalous results and their explanation. *Mach. Learn. 13*, 2/3 (Nov./Dec. 1993), 285–319.

FREUND, R. F. AND SIEGEL, H. J. 1993. Heterogeneous processing. *IEEE Computer 26*, 6 (June), 13–17.

FRIESEN, D. K. 1987. Tighter bounds for LPT scheduling on uniform processors. *SIAM J. Comput. 16*, 3 (June 1987), 554–560.

FUJII, M., KASAMI, T., AND NINOMIYA, K. 1969. Optimal Sequencing of Two Equivalent Processors. *SIAM J. Appl. Math. 17*, 1.

GABOW, H. 1982. An almost linear algorithm for two-processor scheduling. *J. ACM 29*, 3 (July), 766–780.

GAJSKI, D. D. AND PIER, J. 1985. Essential issues in multiprocessors. *IEEE Computer 18*, 6 (June).

GAREY, M. AND JOHNSON, D. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Co., New York, NY.

GAREY, M. R., JOHNSON, D., TARJAN, R., AND YANNAKAKIS, M. 1983. Scheduling opposing forests. *SIAM J. Algebr. Discret. Methods 4*, 1, 72–92.

GERASOULIS, A. AND YANG, T. 1992. A comparison of clustering heuristics for scheduling DAG's on multiprocessors. *J. Parallel Distrib. Comput. 16*, 4 (Dec.), 276–291.

GERASOULIS, A. AND YANG, T. 1993. On the granularity and clustering of directed acyclic task graphs. *IEEE Trans. Parallel Distrib. Syst. 4*, 6 (June), 686–701.

GOLDBERG, D. E. 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Publishing Co., Inc., Redwood City, CA.

GONZALEZ, M. J., JR. 1977. Deterministic processor scheduling. *ACM Comput. Surv. 9*, 3 (Sept.), 173–204.

GONZALEZ, T. AND SAHNI, S. 1978. Preemptive scheduling of uniform processor systems. *J. ACM 25*, 1 (Jan.), 92–101.

GRAHAM, R. L. 1966. Bounds for certain multiprocessing anomalies. *Bell Syst. Tech. J. 45*, 1563–1581.

GRAHAM, R. L., LAWLER, E. L., LENSTRA, J. K., AND RINNOY KAN, A. H. G. 1979. Optimization and approximation in deterministic sequencing and scheduling: A survey. *Ann. Discrete Math. 5*, 287–326.

HA, S. AND LEE, E. A. 1991. Compile-time scheduling and assignment of data-flow program graphs with data-dependent iteration. *IEEE Trans. Comput. 40*, 11 (Nov. 1991), 1225–1238.

HANAN, M. AND KURTZBERG, J. 1972. A review of the placement and quadratic assignment problems. *SIAM Rev. 14* (Apr.), 324–342.

HOCHBAUM, D. S. AND SHMOYS, D. B. 1987. Using dual approximation algorithms for scheduling problems: theoretical and practical results. *J. ACM 34*, 1 (Jan. 1987), 144–162.

HOCHBAUM, D. S. AND SHMOYS, D. B. 1988. A polynomial approximation scheme for scheduling on uniform processors: Using the dual approximation approach. *SIAM J. Comput. 17*, 3 (June 1988), 539–551.

HOLLAND, J. H. 1992. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. 2nd MIT Press, Cambridge, MA.

HORVATH, E. C., LAM, S., AND SETHI, R. 1977. A level algorithm for preemptive scheduling. *J. ACM 24*, 1 (Jan.), 32–43.

HOU, E. S. H., ANSARI, N., AND REN, H. 1994. A genetic algorithm for multiprocessor scheduling. *IEEE Trans. Parallel Distrib. Syst. 5*, 2 (Feb.), 113–120.

HU, T. C. 1961. Parallel sequencing and assembly line problems. *Oper. Res. 19*, 6 (Nov.), 841–848.

HWANG, K. 1993. *Advanced Computer Architecture: Parallelism, Scalability, Programmability*. McGraw-Hill, Inc., New York, NY.

HWANG, J.-J., CHOW, Y.-C., ANGER, F. D., AND LEE, C.-Y. 1989. Scheduling precedence graphs in systems with interprocessor communication times. *SIAM J. Comput. 18*, 2 (Apr. 1989), 244–257.

INTEL SUPERCOMPUTER SYSTEMS DIVISION. 1991. iPSC/2 and iPSC/860 Interactive Parallel Debugger Manual.

JAIN, K. K. AND RAJARMAN, V. 1994. Lower and upper bounds on time for multiprocessor optimal schedules. *IEEE Trans. Parallel Distrib. Syst. 5*, 8 (Aug.), 879–886.

JAIN, K. K. AND RAJARAMAN, V. 1995. Improved lower bounds on time and processors for scheduling precedence graphs on multicomputer systems. *J. Parallel Distrib. Comput. 28*, 1 (July 1995), 101–108.

JIANG, H., BHUYAN, L. N., AND GHOSAL, D. 1990. Approximate analysis of multiprocessing task graphs. In *Proceedings of the International Conference on Parallel Processing* (Aug.), 228–235.

JOHNSON, D. S., PAPADIMTRIOU, C. H., AND YANNAKAKIS, M. 1988. How easy is local search?. *J. Comput. Syst. Sci. 37*, 1 (Aug. 1988), 79–100.

KARP, R. M. 1991. An introduction to randomized algorithms. *Discrete Appl. Math. 34*, 1-3 (Nov. 1991), 165–201.

KASAHARA, H. AND NARITA, S. 1984. Practical multiprocessor scheduling algorithms for efficient parallel processing. *IEEE Trans. Comput. C-33*, 11 (Nov.), 1023–1029.

KAUFMAN, M. 1974. An almost-optimal algorithm for the assembly line scheduling problem. *IEEE Trans. Comput. C-23*, 11 (Nov.), 1169–1174.

KHAN, A., MCCREARY, C. L., AND JONES, M. S. 1994. A comparison of multiprocessor scheduling heuristics. In *Proceedings of the 1994 International Conference on Parallel Processing*, CRC Press, Inc., Boca Raton, FL, 243–250.

KIM, S. J. AND BROWNE, J. C. 1988. A general approach to mapping of parallel computation upon multiprocessor architectures. In *Proceedings of International Conference on Parallel Processing* (Aug.), 1–8.

KIM, D. AND YI, B.-G. 1994. A two-pass scheduling algorithm for parallel programs. *Parallel Comput. 20*, 6 (June 1994), 869–885.

KOHLER, W. H. 1975. A preliminary evaluation of the critical path method for scheduling tasks on multiprocessor systems. *IEEE Trans. Comput. C-24*, 12 (Dec.), 1235–1238.

KOHLER, W. H. AND STEIGLITZ, K. 1974. Characterization and theoretical comparison of branch-and-bound algorithms for permutation problems. *J. ACM 21*, 1 (Jan.), 140–156.

KON'YA, S. AND SATOH, T. 1993. Task scheduling on a hypercube with link contentions. In *Proceedings of International Parallel Processing Symposium* (Apr.), 363–368.

KRUATRACHUE, B. AND LEWIS, T. G. 1987. Duplication Scheduling Heuristics (DSH): A New Precedence Task Scheduler for Parallel Processor Systems. Oregon State University, Corvallis, OR.

KRUATRACHUE, B. AND LEWIS, T. G. 1988. Grain size determination for parallel processing. *IEEE Software 5*, 1 (Jan.), 23–32.

KUMAR, V., GRAMA, A., GUPTA, A., AND KARYPIS, G. 1994. *Introduction to Parallel Computing: Design and Analysis of Algorithms*. Benjamin/Cummings, Redwood City, CA.

KWOK, Y.-K. 1997. High-performance algorithms for compile-time scheduling of parallel processors. Ph.D. Dissertation. Department of Computer Science, Hong Kong University of Science and Technology, Hong Kong.

KWOK, Y.-K. AND AHMAD, I. 1995. Bubble scheduling: A quasi-dynamic algorithm for static allocation of tasks to parallel architectures. In *Proceedings of 7th Symposium on Parallel and Distributed Processing* (Oct.), 36–43.

KWOK, Y.-K. AND AHMAD, I. 1996. Dynamic critical-path scheduling: An effective technique for allocating task graphs to multiprocessors. *IEEE Trans. Parallel Distrib. Syst. 7*, 5, 506–521.

KWOK, Y.-K. AND AHMAD, I. 1997. Efficient scheduling of arbitrary task graphs to multiprocessors using a parallel genetic algorithm. *J. Parallel Distrib. Comput. 47*, 1, 58–77.

KWOK, Y.-K. AND AHMAD, I. 1999a. FASTEST: A practical low-complexity algorithm for compile-time assignment of parallel programs to multiprocessors. *IEEE Trans. Parallel Distrib. Syst. 10*, 2 (Feb.), 147–159.

KWOK, Y.-K. AND AHMAD, I. 1999b. Benchmarking and comparison of the task graph scheduling algorithms. *J. Parallel Distrib. Comput. 59*, 3 (Dec.), 381–422.

KWOK, Y.-K., AHMAD, I., AND GU, J. 1996. FAST: A low-complexity algorithm for efficient scheduling of DAGs on parallel processors. In *Proceedings of 25th International Conference on Parallel Processing* (Aug.), 150–157.

LEE, S.-Y. AND AGGARWAL, J. K. 1987. A mapping strategy for parallel processing. *IEEE Trans. Comput. C-36*, 4 (Apr. 1987), 433–442.

LEE, B., HURSON, A. R., AND FENG, T. Y. 1991. A vertically layered allocation scheme for data flow systems. *J. Parallel Distrib. Comput. 11*, 3 (Mar. 1991), 175–187.

LEUNG, J. Y.-T. AND YOUNG, G. H. 1989. Minimizing schedule length subject to minimum flow time. *SIAM J. Comput. 18*, 2 (Apr. 1989), 314–326.

LEWIS, T. G. AND EL-REWINI, H. 1993. Parallax: A tool for parallel program scheduling. *IEEE Parallel Distrib. Technol. 1*, 2 (May), 64–76.

LIOU, J.-C. AND PALIS, M. A. 1996. An efficient task clustering heuristic for scheduling DAGs on multiprocessors. In *Workshop on Resource Management, Symposium on Parallel and Distributed Processing*,

LIOU, J.-C. AND PALIS, M. A. 1997. A comparison of general approaches to multiprocessor scheduling. In *Proceedings of the 11th International Parallel Processing Symposium* (Apr.), 152–156.

LO, V. M. 1992. Temporal communication graphs: Lamport's process-time graphs augmented for the purpose of mapping and scheduling. *J. Parallel Distrib. Comput. 16*, 4 (Dec.), 378–384.

LO, V. M., RAJOPADHYE, S., GUPTA, S., KELDSEN, D., MOHAMED, M. A., NITZBERG, B., TELLE, J. A., AND ZHONG, X. 1991. OREGAMI: Tools for mapping parallel computations to parallel architectures. *Int. J. Parallel Program. 20*, 3, 237–270.

LORD, R. E., KOWALIK, J. S., AND KUMAR, S. P. 1983. Solving linear algebraic equations on an MIMD computer. *J. ACM 30*, 1 (Jan.), 103–117.

MANOHARAN, S. AND TOPHAM, N. P. 1995. An assessment of assignment schemes for dependency graphs. *Parallel Comput. 21*, 1 (Jan. 1995), 85–107.

MARKENSCOFF, P. AND LI, Y. Y. 1993. Scheduling a computational DAG on a parallel system with communication delays and replication of node execution. In *Proceedings of International Parallel Processing Symposium* (Apr.), 113–117.

MASSPAR COMPUTER. 1992. MPPE User's Guide.

MCCREARY, C. AND GILL, H. 1989. Automatic determination of grain size for efficient parallel processing. *Commun. ACM 32*, 9 (Sept. 1989), 1073–1078.

MCCREARY, C., KHAN, A. A., THOMPSON, J. J., AND MCARDLE, M. E. 1994. A comparison of heuristics for scheduling DAG's on multiprocessors. In *Proceedings of International Parallel Processing Symposium*, 446–451.

MEHDIRATTA, N. AND GHOSE, K. 1994. A bottom-up approach to task scheduling on distributed memory multiprocessor. In *Proceedings of the 1994 International Conference on Parallel Processing*, CRC Press, Inc., Boca Raton, FL, 151–154.

MENASCÉ, D. AND ALMEIDA, V. 1990. Cost-performance analysis of heterogeneity in supercomputer architectures. In *Proceedings on Supercomputing '90* (New York, NY, Nov. 12–16, 1990), J. L. Martin, Ed. IEEE Computer Society Press, Los Alamitos, CA, 169–177.

MENASCÉ, D. A. AND PORTO, S. C. 1993. Scheduling on heterogeneous message passing architectures. *J. Comput. Softw. Eng. 1*, 3.

MENASCÉ, D. A., PORTO, S. C., AND TRIPATHI, S. K. 1994. Static heuristic processor assignment in heterogeneous message passing architectures. *Int. J. High Speed Comput. 6*, 1 (Mar.), 115–137.

MENASCÉ, D. A., PORTO, S. C., AND TRIPATHI, S. K. 1992. Processor assignment in heterogeneous parallel architectures. In *Proceedings of International Parallel Processing Symposium*.

MENASCÉ, D. A., SAHA, D., PORTO, S. C. D. S., ALMEIDA, V. A. F., AND TRIPATHI, S. K. 1995. Static and dynamic processor scheduling disciplines in heterogeneous parallel architectures. *J. Parallel Distrib. Comput. 28*, 1 (July 1995), 1–18.

MOTWANI, R. AND RAGHAVAN, P. 1995. *Randomized Algorithms*. Cambridge University Press, New York, NY.

NORMAN, M. G. AND THANISCH, P. 1993. Models of machines and computation for mapping in multicomputers. *ACM Comput. Surv. 25*, 3 (Sept. 1993), 263–302.

PALIS, M. A., LIOU, J.-C., RAJASEKARAN, S., SHENDE, S., AND WEI, D. S. L. 1995. Online scheduling of dynamic trees. *Para. Proc. Lett. 5*, 4 (Dec.), 635–646.

PALIS, M. A., LIOU, J.-C., AND WEI, D. S. L. 1996. Task clustering and scheduling for distributed memory parallel architectures. *IEEE Trans. Parallel Distrib. Syst. 7*, 1, 46–55.

PANDE, S. S., AGRAWAL, D. P., AND MAUNEY, J. 1994. A threshold scheduling strategy for Sisal on distributed memory machines. *J. Parallel Distrib. Comput. 21*, 2 (May 1994), 223–236.

PAPADIMITRIOU, C. H. AND STEIGLITZ, K. 1982. *Combinatorial Optimization: Algorithms and Complexity*. Prentice-Hall, Inc., Upper Saddle River, NJ.

PAPADIMITRIOU, C. H. AND ULLMAN, J. D. 1987. A communication-time tradeoff. *SIAM J. Comput. 16*, 4 (Aug. 1987), 639–646.

PAPADIMITRIOU, C. H. AND YANNAKAKIS, M. 1979. Scheduling interval-ordered tasks. *SIAM J. Comput. 8*, 405–409.

PAPADIMITRIOU, C. H. AND YANNAKAKIS, M. 1990. Towards an architecture-independent analysis of parallel algorithms. *SIAM J. Comput. 19*, 2 (Apr. 1990), 322–328.

PEASE, D., GHAFOOR, A., AHMAD, I., ANDREWS, D. L., FOUDIL-BEY, K., KARPINSKI, T. E., MIKKI, M. A., AND ZERROUKI, M. 1991. PAWS: A performance evaluation tool for parallel computing systems. *IEEE Computer 24*, 1 (Jan. 1991), 18–29.

PRAMANICK, I AND KUHL, J. G. 1995. An inherently parallel method for heuristic problem-solving: Part I–General framework. *IEEE Trans. Parallel Distrib. Syst. 6*, 10 (Oct.), 1006–1015.

PRASTEIN, M. 1987. Precedence-constrained scheduling with minimum time and communication. Master's Thesis. University of Illinois at Urbana-Champaign, Champaign, IL.

QUINN, M. J. 1994. *Parallel computing (2nd ed.): theory and practice*. McGraw-Hill, Inc., New York, NY.

RAMAMOORTHY, C. V., CHANDY, K. M., AND GONZALEZ, M. J. 1972. Optimal scheduling strategies in a multiprocessor system. *IEEE Trans. Comput. C-21*, 2 (Feb.), 137–146.

RAYWARD-SMITH, V. J. 1987a. The complexity of preemptive scheduling given interprocessor communication delays. *Inf. Process. Lett. 25*, 2 (6 May 1987), 123–125.

RAYWARD-SMITH, V. J. 1987b. UET scheduling with unit interprocessor communication delays. *Discrete Appl. Math. 18*, 1 (Jan. 1987), 55–71.

SARKAR, V. 1989. *Partitioning and Scheduling Parallel Programs for Multiprocessors*. MIT Press, Cambridge, MA.

SCHWEHM, M., WALTER, T., BUCHBERGER, B., AND VOLKERT, J. 1994. Mapping and scheduling by genetic algorithms. In *Proceedings of the 3rd Joint International Conference on Vector and Parallel Processing* (CONPAR '94), Springer-Verlag, New York, NY, 832–841.

SELVAKUMAR, S. AND MURTHY, C. S. R. 1994. Scheduling precedence constrained task graphs with non-negligible intertask communication onto multiprocessors. *IEEE Trans. Parallel Distrib. Syst. 5*, 3 (Mar.), 328–336.

SETHI, R. 1976. Scheduling graphs on two processors. *SIAM J. Comput. 5*, 1 (Mar.), 73–82.

SHIRAZI, B., KAVI, K., HURSON, A. R., AND BISWAS, P. 1993. PARSA: A parallel program scheduling and assessment environment. In *Proceedings of the International Conference on Parallel Processing*, CRC Press, Inc., Boca Raton, FL, 68–72.

SHIRAZI, B., WANG, M., AND PATHAK, G. 1990. Analysis and evaluation of heuristic methods for static task scheduling. *J. Parallel Distrib. Comput. 10*, 3 (Nov. 1990), 222–232.

SIEGEL, H. J., ARMSTRONG, J. B., AND WATSON, D. W. 1992. Mapping computer-vision-related tasks onto reconfigurable parallel-processing systems. *IEEE Computer 25*, 2 (Feb. 1992), 54–64.

SIEGEL, H. J., DIETZ, H. G., AND ANTONIO, J. K. 1996. Software support for heterogeneous computing. *ACM Comput. Surv. 28*, 1, 237–239.

SIH, G. C. AND LEE, E. A. 1993a. A compile-time scheduling heuristic for interconnection-constrained heterogeneous processor architectures. *IEEE Trans. Parallel Distrib. Syst. 4*, 2 (Feb.), 75–87.

SIH, G. C. AND LEE, E. A. 1993b. Declustering: A new multiprocessor scheduling technique. *IEEE Trans. Parallel Distrib. Syst. 4*, 6 (June), 625–637.

SIMONS, B. B. AND WARMUTH, M. K. 1989. A fast algorithm for multiprocessor scheduling of unit-length jobs. *SIAM J. Comput. 18*, 4 (Aug. 1989), 690–710.

SRINIVAS, M. AND PATNAIK, L. M. 1994. Genetic algorithms: A survey. *IEEE Computer 27*, 6 (June 1994), 17–26.

STONE, H. S. 1977. Multiprocessor scheduling with the aid of network flow algorithms. *IEEE Trans. Softw. Eng. SE-3*, 1 (Jan.), 85–93.

SUMICHRAST, R. T. 1987. Scheduling parallel processors to minimize setup time. *Comput. Oper. Res. 14*, 4 (Oct. 1987), 305–313.

STORER, R. H., WU, S. D., AND VACCARI, R. 1992. New search spaces for sequencing problems with application to job shop scheduling. *Manage. Sci. 38*, 10 (Oct. 1992), 1495–1509.

THINKING MACHINES CORPORATION. 1991. PRISM User's Guide. Thinking Machines Corp., Bedford, MA.

TOWSLEY, D 1986. Allocating programs containing branches and loops within a multiple processor system. *IEEE Trans. Softw. Eng. SE-12*, 10 (Oct. 1986), 1018–1024.

VARVARIGOU, T. A., ROYCHOWDHURY, V. P., KAILATH, T., AND LAWLER, E. 1996. Scheduling in and out forests in the presence of communication delays. *IEEE Trans. Parallel Distrib. Syst. 7*, 10, 1065–1074.

VELTMAN, B., LAGEWEG, B. J., AND LENSTRA, J. K. 1990. Multiprocessor scheduling with communication delays. *Parallel Comput. 16*, 173–182.

ULLMAN, J. 1975. NP-complete scheduling problems. *J. Comput. Syst. Sci. 10*, 384–393.

WANG, M.-F. 1990. Message routing algorithms for static task scheduling. In *Proceedings of the 1990 Symposium on Applied Computing* (SAC '90), 276–281.

WANG, Q. AND CHENG, K. H. 1991. List scheduling of parallel tasks. *Inf. Process. Lett. 37*, 5 (Mar. 14, 1991), 291–297.

WANG, L., SIEGEL, H. J., AND ROYCHOWDHURY, V. P. 1996. A genetic-algorithm-based approach for task matching and scheduling in heterogeneous computing environments. In *Proceedings of the '96 Workshop on Heterogeneous Computing*, IEEE Computer Society Press, Los Alamitos, CA, 72–85.

WONG, W. S. AND MORRIS, R. J. T. 1989. A new approach to choosing initial points in local search. *Inf. Process. Lett. 30*, 2 (Jan. 1989), 67–72.

WU, M.-Y. AND GAJSKI, D. D. 1990. Hypertool: A programming aid for message-passing systems. *IEEE Trans. Parallel Distrib. Syst. 1*, 3 (1990), 330–343.

YANG, C.-Q. AND MILLER, B. P. 1988. Critical path analysis for the execution of parallel and distributed programs. In *Proceedings of the 8th International Conference on Distributed Computing Systems* (ICDCS '88, Washington, D. C., June), IEEE Computer Society Press, Los Alamitos, CA, 366–373.

YANG, T. AND GERASOULIS, A. 1993. List scheduling with and without communication delays. *Parallel Comput. 19*, 12 (Dec. 1993), 1321–1344.

YANG, T. AND GERASOULIS, A. 1992. PYRROS: Static task scheduling and code generation for message passing multiprocessors. In *Proceedings of the 1992 international conference on Supercomputing* (ICS '92, Washington, DC, July 19–23, 1992), K. Kennedy and C. D. Polychronopoulos, Eds. ACM Press, New York, NY, 428–437.

YANG, T. AND GERASOULIS, A. 1994. DSC: Scheduling parallel tasks on an unbounded number of processors. *IEEE Trans. Parallel Distrib. Syst. 5*, 9 (Sept.), 951–967.

YANG, J., BIC, L., AND NICOLAU, A. 1993. A mapping strategy for MIMD computers. *Int. J. High Speed Comput. 5*, 1, 89–103.

ZHU, Y. AND MCCREARY, C. L. 1992. Optimal and near optimal tree scheduling for parallel systems. In *Proceedings of Symposium on Parallel and Distributed Processing*, IEEE Computer Society Press, Los Alamitos, CA, 112–119.

Received: December 1997; revised: July 1998; accepted: September 1998