Cluster Computing 3 (2000) 113–124

Link contention-constrained scheduling and mapping of tasks and messages to a network of heterogeneous processors

Yu-Kwong Kwok [a] and Ishfaq Ahmad [b]

[a] *Department of Electrical and Electronic Engineering, The University of Hong Kong, Pokfulam Road, Hong Kong*
[b] *Department of Computer Science, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong*

In this paper, we consider the problem of scheduling and mapping precedence-constrained tasks to a network of heterogeneous processors. In such systems, processors are usually physically distributed, implying that the communication cost is considerably higher than in tightly coupled multiprocessors. Therefore, scheduling and mapping algorithms for such systems must schedule the tasks as well as the communication traffic by treating both the processors and communication links as equally important resources. We propose an algorithm that achieves these objectives and adapts its task scheduling and mapping decisions according to the given network topology. Just like tasks, messages are also scheduled and mapped to suitable links during the minimization of the finish times of tasks. Heterogeneity of processors is exploited by scheduling critical tasks to the fastest processors. Our experimental study has demonstrated that the proposed algorithm is efficient and robust, and yields consistent performance over a wide range of scheduling parameters.

1. Introduction

One of the major goals of using a heterogeneous system is to minimize the completion time of a parallel application by exploiting the heterogeneous processing requirements within the application [7,22]. To achieve this goal, a judicious scheme is needed to properly schedule and allocate the tasks of the application to the most suitable processors. In this study, we are interested in the static scheduling of precedence-constrained tasks to a network of heterogeneous processors. Static scheduling is normally done at compile-time with available information about the structure of the parallel application in terms of its task execution times, task dependencies, communication, and synchronization [5,17,20]. The goal of static scheduling is to allocate a set of tasks to a set of processors such that the overall completion time of the application, called the *schedule length*, is minimized[1] while the precedence constraints among the tasks are preserved. Since this scheduling problem is NP-complete [5,8,21] it is commonly tackled by using heuristics [12,13]. While each heuristic may perform well under different circumstances, there are three important criteria that must be considered for evaluating a heuristic:

(1) Does the heuristic make realistic assumptions about the application and architecture of the system?

(2) Is it problem-specific or can it work under a wide range of parameters without compromising the solution quality?

(3) Does the complexity of the heuristic permit it to be practically used for compile-time scheduling?

[1] It should be noted that balancing the computational load among processors does not necessarily minimize the program completion time due to the inter-task communications.

The first criterion relates to the assumptions made by the scheduling algorithm about the program tasks and architecture models. Indeed, to simplify the design of the scheduling method, earlier approaches usually rely on simplifying assumptions such as assuming all tasks to have equal execution times, or ignoring the communication delays among tasks altogether [5,17]. With the emergence of a wide variety of architectures in recent years, the architectural attributes such as system topology, message routing strategy, overlapped communication and computation, and processors heterogeneity, must also be taken into account by a scheduling algorithm. The second criterion dictates that the scheduling algorithm should generate good solutions for a variety of applications and target systems. A scheduling algorithm tailored for one particular application and architecture may not generate efficient solutions on another architecture [15,17]. The third criterion, which is related to the execution time of the heuristic itself, is an important consideration for effectively using it for compile-time scheduling of large-scale applications [2].

The majority of algorithms that take into account inter-task communication assume the availability of unlimited number of processors [24,25]. These algorithms are called the UNC (unbounded number of clusters) scheduling algorithms [14–17,25]. The algorithms that consider limited number of processors are called the BNP (bounded number of processors) scheduling algorithms [3,15,17]. In UNC and BNP scheduling algorithms, the processors are assumed to be fully-connected, and no attention is paid to link contention or routing strategies used for communication. Processors and links heterogeneity are also ignored. The UNC and BNP algorithms are also called clustering algorithms because they merge tasks into clusters (bounded or unbounded) [19]; the clusters may need to be mapped onto the processors using a mapping algorithm [25]. A few algorithms, which relax all of the above mentioned assump-

tions and can handle a system connected via an arbitrary network topology, are called the APN (arbitrary processor network) scheduling algorithms. In APN algorithms, the mapping of tasks to processors is implicit, and messages are also scheduled while considering link contention.

Most algorithms belonging to the above three classes use different forms of the classical *list scheduling* approach [1,5,15,18,20,25]. The basic idea of list scheduling is to assign priorities to the tasks, and then repeatedly execute the following two steps until a valid schedule is obtained: select from the list the task with the highest priority for scheduling; and select a processor to accommodate this task. The main drawback of this approach is that the static priority assignment does not always lead to an optimized task sequence for scheduling. Indeed, the weakness of a list scheduling approach is that each task is scheduled independently without regard to the scheduling of subsequent tasks. This adverse effect is particularly severe if message scheduling, which is a difficult problem for a heterogeneous system, has to be considered. A later task may not be able to occupy an earlier time slot because its incoming messages cannot be scheduled earlier due to the inefficient scheduling of previous messages. The idea of determining priorities dynamically has been proposed [17,23,27] but it increases the time complexity and still may not be able to avoid making suboptimal scheduling decisions.

We are interested in APN algorithms that both schedule tasks and messages on arbitrary networks consisting of heterogeneous processors and communication links. Scheduling tasks while considering link contention for a heterogeneous system is a relatively less explored research topic and very few algorithms for this problem have been designed. One well-known algorithm is the *dynamic level scheduling* (DLS) algorithm [23], which employs a dynamic list scheduling approach. In this paper, we propose a new algorithm, the primary objective of which is to generate efficient solutions while simultaneously handles arbitrary communication and execution costs in the parallel application, schedules tasks and messages by considering link contention as well as processors heterogeneity, and adapts to arbitrary network topology. The algorithm has a practicable complexity and is suitable for regular and irregular parallel program structures.

The remainder of this paper is organized as follows. In the next section, we provide a formal problem statement, followed by a detailed description and explanation of the proposed algorithm. An illustrative example is used throughout the section to explicate the features of the algorithm. Section 3 includes some performance analysis of the algorithm. Section 4 presents the experimental results. The last section concludes the paper.

2. The proposed algorithm

In this section, we first formally define the scheduling problem and the model used. We then explicate our proposed algorithm, called *Bubble Scheduling and Allocation* (BSA), by describing its constituent procedures. A small example is used for illustrating the algorithm's characteristics.

2.1. The scheduling and mapping model

A parallel program is composed of n tasks $\{T_1, T_2, \ldots, T_n\}$ in which there is a partial order: $T_i < T_j$ implies that T_j cannot start execution until T_i finishes due to the data dependency between them. Thus, a parallel program can be represented by a *directed acyclic task graph* [3]. Parallelism exists among independent tasks – T_i and T_j are said to be independent if neither $T_i < T_j$ nor $T_j < T_i$. Each task T_i is associated with a nominal execution cost τ_i which is the execution time required by T_i on a reference machine in the heterogeneous system. Similarly, a nominal communication cost c_{ij} is associated with the message M_{ij} from T_i to T_j. These nominal costs are obtained by estimation techniques such as profiling and analytic benchmarking [2,6,9–11,26]. Assume there are e messages where $(n-1) \leqslant e < n^2$ so that the task graph is a connected graph.

To model heterogeneity of the target system which consists of m processors $\{P_1, P_2, \ldots, P_m\}$, *heterogeneity factors* are used. For example, if a task T_i is scheduled to a processor P_x, then its actual execution cost is given by $h_{ix}\tau_i$ where h_{ix} is the heterogeneity factor which is determined by measuring the difference in processing capabilities (e.g., speed) of processor P_x and the reference machine with respect to task T_i. Similarly, if a message M_{ij} is scheduled to the communication link L_{xy} between processors P_x and P_y, its actual communication cost is given by $h'_{ijxy}c_{ij}$. An example parallel program graph and a heterogeneous processor network are shown in figure 1.

The start time and finish time of a message M_{ij} from T_i to T_j on a communication link L_{xy} are denoted by *MST*(M_{ij}, L_{xy}) and *MFT*(M_{ij}, L_{xy}), respectively. Obviously, we have *MFT*(M_{ij}, L_{xy}) = *MST*$(M_{ij}, L_{xy}) + h'_{ijxy}c_{ij}$. The start time of a task T_i on processor P_x is denoted by *ST*(T_i, P_x) which critically depends on the task's *data ready time* (DRT). The DRT of a task is defined as the latest arrival time of messages from its predecessors. The finish time of a task T_i is given by *FT*(T_i, P_x) = *ST*$(T_i, P_x) + h_{ix}\tau_i$. The objective of scheduling is to minimize the maximum *FT*, which is called the schedule length (SL).

2.2. Overview of the BSA algorithm

In a traditional scheduling algorithm, the tasks are first arranged as a list using some priority measure and then each task is scheduled one after another to a processor which allows the earliest finish time. To find such a processor in a heterogeneous target system where message scheduling has to be handled, a routing table is also needed, as in the DLS algorithm [23], for determining the most suitable route for messages in order to minimize the DRT. The

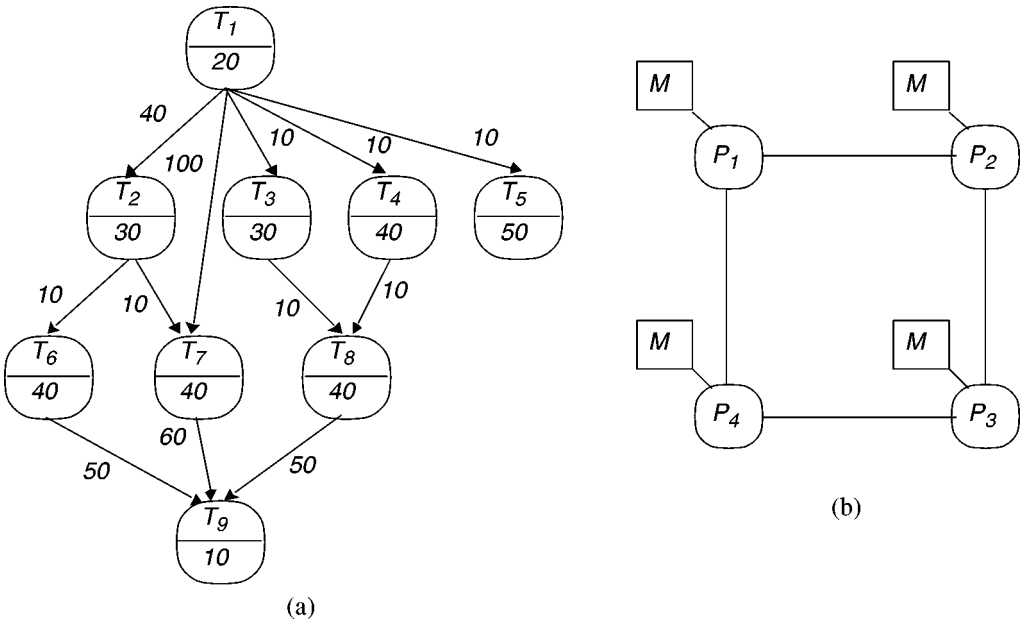

Figure 1. (a) A parallel program task graph; (b) a 4-processor ring heterogeneous system.

problem with using a routing table is two-fold: (1) the routing table has to be pre-determined, usually using shortest-path algorithm, for the input target topology; (2) during the scheduling process, the routing table, which has to be frequently updated, may not give optimized routes. Checking such routing information for every candidate processors inevitably results in high time complexity. Furthermore, the routing information is usually maintained for only a few common network topologies which may not be useful for an arbitrary network.

The proposed BSA algorithm is different from traditional scheduling schemes in several aspects. First, in the BSA algorithm, the tasks are not fixed in one single list throughout the entire scheduling process as in the traditional approach. Initially, the tasks are all scheduled to a single processor – effectively the parallel program is serialized. Then, each task is considered in turn for possible migration to the neighbor processors. The objective of this process is to improve the finish times of tasks because a task migrates only if it can "bubble up". If a task is selected for migration, the communication messages from its predecessors (some of which may remain in the original processor while others may have also migrated) are scheduled to the communication link between the new processor and the original processor. After all the tasks in the original processor are considered, the first phase of scheduling completes. In the second phase, the same process is repeated on one of the neighbor processor. Thus, a task migrated from the original processor to a neighbor processor may have an opportunity to migrate again to a processor one more hop away from the original processor. This incremental scheduling by migration process is repeated for all the processors in a breadth-first fashion. The advantage of this incremental approach is that no pre-specified routing table is needed because the algorithm adapts its scheduling decisions to each input topology, which may be arbitrary. More importantly, the incremental scheduling of tasks and messages can lead to optimized routes.

2.3. Serialization

The serialization process, which determines the order of subsequent task migration, is a crucial step of the algorithm. A parallel program can be serialized using many different methods because there are many total orders which do not violate the original partial order. In the BSA algorithm, the serialization process is centered around a *critical path* of the parallel program.

Definition 1. A critical path (CP) is defined as the set of tasks and messages forming a path with the largest sum of execution costs and communication costs.

In the case that there are multiple CPs, we select the one with a larger sum of execution costs and ties are broken randomly. The CP is a crucial structure of a parallel program because it is the longest execution path and thus, timely scheduling of its tasks can potentially lead to a shorter schedule length. However, to preserve the precedence constraints among tasks, we cannot arrange all the CP tasks first. Instead, in the serialization process, we have to first consider a CP task's predecessors, which need not be CP tasks themselves. Such predecessors are called *in-branch* (IB) tasks. The remaining tasks, which are neither CP tasks nor IB tasks, are called *out-branch* (OB) tasks. This partitioning of the tasks into three disjoint categories induces a serial order of the parallel program, in which CP tasks are arranged to occupy the earliest possible positions, with IB tasks inserted among them, and OB tasks are appended at the end.

To determine whether a task is a CP task, we can use two attributes: *t-level* (top level) and *b-level* (bottom level). The *b-level* of a task is the length of the longest path beginning with the task. The *t-level* of a task is the length of the longest path reaching the task. Thus, all tasks on the CP have the same value of (*t-level* + *b-level*), which is equal to the length of the CP. Based on this observation, we can easily partition the parallel program into CP, IB, and OB tasks by in O(e) time because the *t-level* and *b-level* of all tasks can be computed by using depth-first search. A task with a larger *b-level* implies that it is followed by a longer chain of tasks, and thus, is given a higher priority. The serialization process can be performed by an O(e) time algorithm outlined below.

Algorithm Serialization
Input: a program task graph with n tasks $\{T_1, T_2, \ldots, T_n\}$.
Output: a serial order of the tasks.

1. compute the *t-level* and *b-level* of each task by using depth-first search;
2. identify the CP; if there are multiple CPs, select the one with the largest sum of execution cost and ties are broken randomly;
3. put the CP task which does not have any predecessor to the first position of the serial order;
4. $i \leftarrow 2$; $T_x \leftarrow$ the next CP task;
5. while not all the CP tasks are included do
6. if T_x has all its predecessors in the serial order then
7. put T_x at position i and increment i;
8. else let T_y be the predecessor of T_x which is not in the serial order and has the largest *b-level* (ties are broken by choosing the predecessor with a smaller *t-level*);
9. if T_y has all its predecessors in the serial order then put T_y at position i and increment i;
otherwise, recursively include all the ancestors of T_y in the serial order such that the tasks with a larger *b-level* are included first;
10. repeat the above step until all the predecessors of T_x are in the serial order;
11. put T_x at position i and increment i;
12. $T_x \leftarrow$ the next CP task;
13. append all the OB tasks to the serial order in descending order of *b-level*.

For example, consider the parallel program graph shown earlier in figure 1(a). Based on the nominal execution and communication costs, the *t-level*s and *b-level*s of the tasks are shown in table 1 and the tasks $\{T_1, T_7, T_9\}$ form the CP. Since T_1 is the first CP task, it is placed in the first position in the serial order. The second task is T_2 because it is another unexamined predecessor of the next CP task T_7. After T_2 is appended to the serial order, all predecessors of T_7 have been considered and, therefore, it can also be added. Now, the last CP task, T_9 is considered. It cannot be appended to the serial order because some of its predecessors (i.e., the IB tasks) have not been examined yet. Since both T_6 and T_8 have the same value of *b-level* and T_8 has a smaller *t-level*, T_8 is considered first. However, both predecessors of T_8 have not been examined. Thus, its two predecessors, T_3 and T_4 are appended to the list first. Next, T_8 is appended followed by T_6. The only OB task, T_5, is the last task in the serial order. The final serialized list is as follows: $\{T_1, T_2, T_7, T_4, T_3, T_8, T_6, T_9, T_5\}$.

Table 1
The *t-level* and *b-level* of each task in the task graph based on the nominal execution costs.

Task	*t-level*	*b-level*
T_1	0	230
T_2	60	150
T_3	30	140
T_4	30	150
T_5	30	50
T_6	100	100
T_7	120	110
T_8	80	100
T_9	220	10

Table 2
The task execution cost $h_{ix}\tau_i$ of each task T_i on the four heterogeneous processors.

Task	P_1	P_2	P_3	P_4
T_1	39	7	2	6
T_2	21	50	57	56
T_3	15	28	39	6
T_4	54	14	16	55
T_5	45	42	97	12
T_6	15	20	57	78
T_7	33	43	51	60
T_8	51	18	47	74
T_9	8	16	15	20

In the serialization process, the tasks are all scheduled to a single processor, called the *pivot* processor, which is selected as follows. The first processor in the heterogeneous system is considered and the corresponding heterogeneity factor is multiplied to the nominal execution cost of each task. Based on the set of actual execution costs, the CP is constructed. This process is repeated for other processors and eventually the processor that gives the shortest CP length based on actual execution costs is selected as the first pivot processor. To illustrate, consider the actual execution costs of the tasks on the four processor heterogeneous system as shown in table 2. Given the actual execution costs, the CPs with respect to P_1, P_2, P_3, and P_4 are $\{T_1, T_7, T_9\}$, $\{T_1, T_2, T_6, T_9\}$, $\{T_1, T_2, T_7, T_9\}$, and $\{T_1, T_2, T_6, T_9\}$, respectively. The CP lengths are 240, 226, 235, and 260, respectively. Thus, the first pivot processor is P_2 because the CP is shortest with respect to this proces-

sor. The serial order is $\{T_1, T_2, T_6, T_7, T_3, T_4, T_8, T_9, T_5\}$, which is different from that determined earlier using nominal execution costs.

2.4. Tasks migration

After the parallel program is serialized to the first pivot processor, tasks have to be considered for possible migration to the neighbor processors in order to improve their finish times (bubble up). To determine whether a migration is beneficial, we have to compute the finish time of the task on a neighbor processor. To compute the start time, we need to know the DRT of the task, which in turn depends on the scheduling of messages. We outline below an algorithm for computing the finish time of a message on a communication link between two processors.

Algorithm ComputeMFT
Input: a message M_{ij}, a communication link L_{xy} on which k messages $\{M_{i_1j_1}, M_{i_2j_2}, \ldots, M_{i_kj_k}\}$ have been scheduled, and $FT(T_i, P_z)$ (note that P_z may be P_x).
Output: $MFT(M_{ij}, L_{xy})$.

1. check if there exists some s such that:

$$MFT(M_{i_{s+1}j_{s+1}}, L_{xy}) - \max\{MFT(M_{i_xj_x}, L_{xy}), FT(T_i, P_z)\} \geqslant h'_{ijxy}c_{ij}$$

where $s = 0, 1, \ldots, k$, $MST(M_{i_{k+1}j_{k+1}}, L_{xy}) = \infty$, and $MFT(M_{i_0j_0}, L_{xy}) = 0$;

2. if such s exists, use the smallest one to compute:

$$\max\{MFT(M_{i_sj_s}, L_{xy}), FT(T_i, P_z)\} + h'_{ijxy}c_ij$$

which is returned as $MFT(M_{ij}, L_{xy})$ otherwise, return ∞.

Using *ComputeMFT*, we can determine the finish times of every incoming messages of the task on a neighbor processor. The maximum finish time is then the DRT of the task. The corresponding predecessor which sends this latest message is called the *very important predecessor* (VIP) of the task. Specifically, we use the algorithm outlined below for finding the DRT and VIP.

Algorithm ComputeDRT
Input: a neighbor processor P_y, a task T_j, its set of predecessors, and their messages to T_j
Output: $DRT(T_j, P_y)$.

1. $DRT(T_j, P_y) = 0$;
2. for each predecessor T_i of T_j do:
3. let P_x be the processor currently accommodating T_i;
4. if $P_x = P_y$ then arrival-time $= FT(T_i, P_x)$;
5. else call*ComputeMFT* and
   arrival-time $= MFT(M_{ij}, L_{xy})$;
6. if arrival-time $> DRT(T_j, P_y)$ then
   $DRT(T_j, P_y) =$ arrival-time; VIP $= T_i$.

After the DRT of the task on a neighbor processor is computed, the potential finish time of the task can also be determined, using the algorithm outlined below.

Algorithm ComputeFT
Input: a task T_j, a neighbor processor P_y on which l tasks $\{T_{P_y^1}, T_{P_y^2}, \ldots, T_{P_y^l}\}$ have been scheduled, and $DRT(T_j, P_y)$.
Output: $FT(T_j, P_y)$.

1. check if there exists some t such that:

$$ST(T_{P_y^{t+1}}, P_y) - \max\{FT(T_{P_y^t}, P_y), DRT(T_j, P_y)\} \geqslant h_{jy}\tau_j$$

where $t = 0, 1, \ldots, l$, $ST(T_{P_y^{t+1}}, P_y) = \infty$, and $FT(T_{P_y^0}, P_y) = 0$;

2. if such t exists, use the smallest one to compute:

$$\max\{FT(T_{P_y^t}, P_y), DRT(T_j, P_y)\} + h_{jy}\tau_j$$

which is returned as $FT(T_j, P_y)$; otherwise, return ∞.

Using *ComputeFT*, we can determine whether a task can improve its finish time through migrating to a neighbor processor of the pivot processor. If the finish time does improve, the task is rescheduled to the neighbor processor and its incoming and outgoing messages are also rearranged. If the finish time does not improve, nevertheless a task will also migrate provided that its VIP is scheduled to that neighbor processor. The rationale behind this heuristic decision is that if a task and its VIP are scheduled to the same processor, the successors of the task may subsequently improve their finish times also. This process is repeated for all the remaining tasks on the pivot. Then a neighbor processor is chosen to be a new pivot. Thus, each processor in the heterogeneous system in turn will be assigned as the pivot in a breadth-first manner. Throughout the entire bubbling up process, messages are automatically routed in the migration process of tasks from the pivot processor to other processors. There is no need to use a routing table. If the routing of messages has to be static (as in some commonly used networks, such as a hypercube that uses the E-cube routing method), we can just put a constraint on the destinations a task can migrate to. Moreover, the routes taken by such messages are optimized routes in that, at every step, a task migrates if its finish time is not increased.

Using the techniques discussed above, the BSA algorithm can be formalized below. In the following, the procedure *BuildProcessorList* constructs a list of processors in a breadth-first order from the first pivot processor.

Algorithm BSA
Input: a parallel program graph with n tasks $\{T_1, T_2, \ldots, T_n\}$ and a heterogeneous system with m processors $\{P_1, P_2, \ldots, P_m\}$.
Output: a program schedule.

1. initial Pivot $\leftarrow$ the processor that gives the shortest CP length;

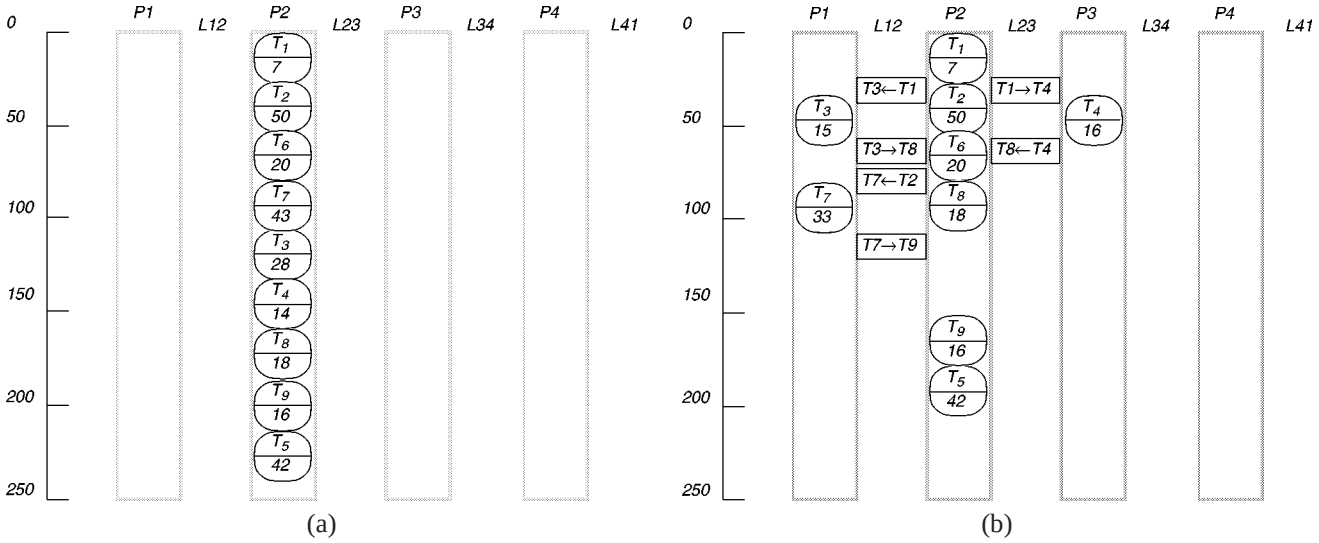

Figure 2. (a) Intermediate schedule generated by the BSA algorithm after *Serialization* (schedule length = 238, total communication costs = 0); (b) intermediate schedule after T_3, T_4, and T_7 migrate to neighbor processors (schedule length = 218, total communication costs = 110).

2. *Serialization*(Pivot);
3. *BuildProcessorList*(Pivot);
4. while *ProcessorList* is not empty do:
5. Pivot ← remove the first processor from *ProcessorList*;
6. for each T_i on Pivot do:
7. if *FT*(T_i, Pivot) > *DRT*(T_i, Pivot) or VIP of T_i is not scheduled to Pivot then
8. for each neighbor processor P_y of Pivot, compute *DRT*(T_i, P_y) and *FT*(T_i, P_y);
9. if there is a neighbor processor P'_y such that *FT*(T_i, P'_y) < *FT*(T_i, Pivot) then
10. make T_i migrate from Pivot to P'_y;
11. else if *FT*(T_i, P'_y) = *FT*(T_i, P'_y) and VIP of T_i is scheduled to P'_y then
12. make T_i migrate from Pivot to P'_y.

The time complexity of the BSA algorithm is derived as follows. The procedure *BuildProcessorList* takes O(m^2) time while *Serialization* takes O(n^2) time. Thus, the dominant step is the while-loop, which takes O(e) time to compute the FT and DRT values of the task on each neighbor processor. If migration is done, it also takes O(e) time. Since there are O(n) tasks on the Pivot and there are O(m) neighbor processor, each iteration of the while loop takes O(men) time. Thus, the BSA algorithm takes O(m^2en) time. The correctness of the BSA algorithm is formalized in the following theorem.

Theorem 2. The BSA algorithm generates schedules in which precedence constraints are preserved.

Proof. We sketch a proof of this theorem by induction on the number of migrations. First observe that in the procedure *Serialization*, the program task graph is serialized by partitioning the graph into three categories: CP tasks, IB tasks, and OB tasks. In the process, every IB tasks of a CP task are considered first before the CP task itself is put into the serial order. The OB tasks are also appended according to their levels. Thus, *Serialization* produces a serial order that preserves the precedence constraints. In the task migration process, a task migrates only if its finish time improves. And in the computation of potential finish times, the message scheduling and task scheduling are examined according to the procedures *ComputeMFT* and *ComputeFT*, in which there are inequalities that determine schedulability based on precedence constraints. Note that the inequalities capture the scheduling state of previous migrations, which according to the induction assumption do not lead to violation of precedence constraints. Thus, a task or a message will not be inserted in a slot which leads to a violation of precedence constraints. As such, both the *Serialization* procedure and the task migration process will not violate the precedence constraints, and therefore, the BSA algorithm generates valid schedules. □

2.5. An example

To illustrate the novel characteristics of the BSA algorithm, let us consider applying it to schedule the parallel program graph shown in figure 1(a) to the heterogeneous ring system shown in figure 1(b) with the actual execution costs depicted in table 2. For simplicity, we assume that the communication links are homogeneous; that is, $h'_{ijxy} = 1$ for all messages M_{ij} and links L_{xy}. Initially, the tasks are injected by the procedure *Serialization* to the first pivot processor P_2 in the order: {T_1, T_2, T_6, T_7, T_3, T_4, T_8, T_9, T_5}, as we have shown in section 2.3. The resulting intermediate schedule is depicted in figure 2(a). Note that the

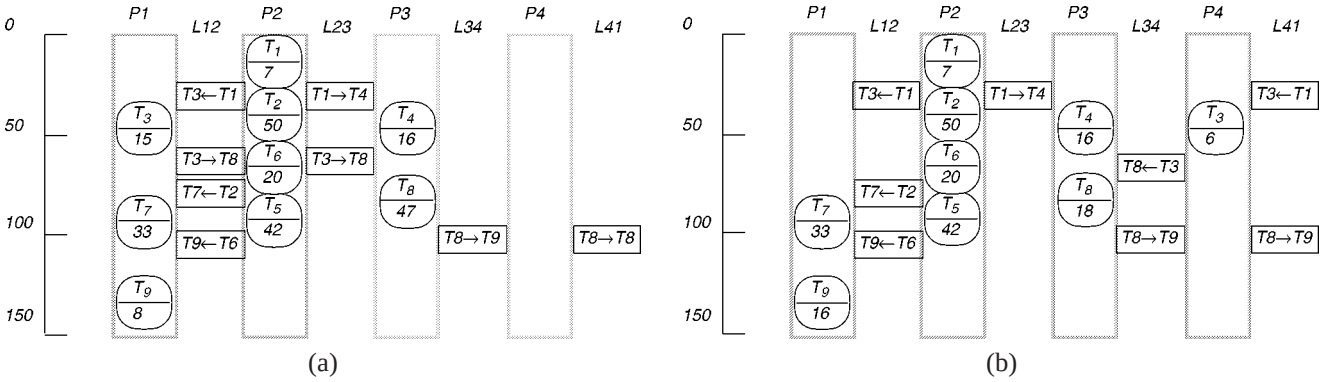

Figure 3. (a) Intermediate schedule after T_8 and T_9 migrate to neighbor processors (schedule length = 147, total communication costs = 200); (b) final schedule after T_3 migrates from P_1 to P_4 (schedule length = 138, total communication costs = 200).

actual execution costs on P_2 are quite different from the nominal execution costs. Then, tasks are considered for possible migration. In the first phase, T_1, being the first CP task, does not migrate because its migration is not beneficial. Also, T_2 and T_6 do not migrate because their finish times cannot be improved by migration. However, T_3 and T_2 migrate to P_1 and P_3, respectively, as their finish times are improved. Note that the reduction of T_3's finish time is contributed not only by the "bubbling up" process but also by the heterogeneity of the processors – the execution cost of T_3 on P_2 is 28 while on P_1 is only 15. Similarly, T_7 also migrates to P_1 since it can also be "bubbled up" and its execution cost is reduced. The resulting intermediate schedule is shown in figure 2(b). After two more migrations from the first pivot processor P_2, the first phase is completed; the intermediate schedule at this point is shown in figure 3(a). In the second phase, the pivot processor is P_1. Only T_3 migrates while the other tasks cannot improve their finish times. No more migration can be performed after this stage and the final schedule is shown in figure 3(b). The schedule length is only 138 which is considerably smaller than that can be achieved on homogeneous processors.

3. Analytical performance on primitive structures

The BSA algorithm considers a large number of systems parameters, in particular, the processor heterogeneity which leads to the lack of symmetry in the problem. Thus, exact performance analysis of the algorithm on general task graphs and processor networks is very difficult. In this section, nevertheless, we analyze the performance of the algorithm on a basic graph structure: the *fork* set. An example of a fork set is shown in figure 4. The fork set is a basic building block of many general task graphs.

Given a fork set with parent T_x and k children (labeled from T_1 to T_k), without loss of generality, we assume that for the fork set F_x,

$$c_{x1} + \tau_1 \geqslant c_{x2} + \tau_2 \geqslant c_{x3} + \tau_3 \geqslant \cdots \geqslant c_{xk} + \tau_k.$$

First, we assume that the target system has m homogeneous processors which are fully connected with homogeneous

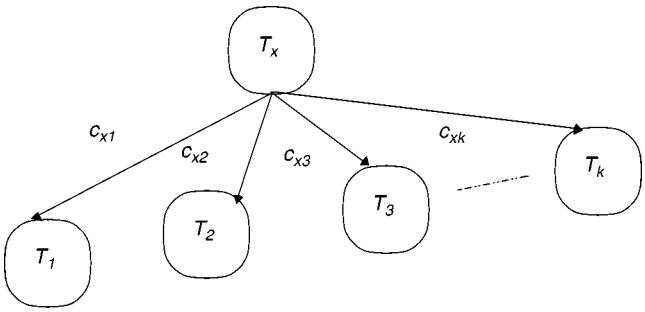

Figure 4. A fork set F_x.

communication links and $m \geqslant k$. The optimal schedule length $SL_{\text{opt}}^{\text{full}}$ for the fork set on this system is given by

$$SL_{\text{opt}}^{\text{full}} = \max\left\{\left(\tau_x + \sum_{i=1}^{j} \tau_i\right), (\tau_x + c_{xj} + \tau_{j+1})\right\},$$

where j is given by the following conditions:

$$\sum_{i=1}^{j} \tau_i \leqslant c_{xj} + \tau_j$$

and

$$\sum_{i=1}^{j+1} \tau_i > c_{xj+1} + \tau_{j+1}.$$

Intuitively, the above optimal schedule length is obtained by including the tasks to the processor holding the parent task until the accumulated execution costs exceed the finish time of the next task.

Next we analyze the optimal schedule length for the fork set for a hypercube network with m processors ($m = 2^d$). If $d \geqslant k - j$ (call this condition C1), then the optimal schedule length $SL_{\text{opt}}^{\text{hc}}$ for a hypercube is equal to $SL_{\text{opt}}^{\text{full}}$. This follows from the fact that all the tasks can be scheduled within one-hop away from the processor P_x to which the root task is scheduled.

On the other hand, if $d < k - j \leqslant C_r^d$, where C_r^d denotes the number of combination of choosing r objects from d objects, some of the tasks have to be scheduled

on processors at two hops from P_x. In this scenario, it is difficult to obtain a closed form expression for the optimal schedule length. Here, we impose one more constraint, called C2, on the task graph:

$$c_{xt} = K, \quad 1 \leqslant t \leqslant k.$$

That is, the communication costs are the same. With this constraint, the ordering of tasks is preserved even though some tasks are to be scheduled two-hop away from P_x. The optimal schedule length is then given by

$$SL_{\text{opt}}^{\text{hc}} = \max\left\{\left(SL_{\text{opt}}^{\text{full}} + \sum_{i=j+d+1}^{s} \tau_i\right), (\tau_x + 2K + \tau_{s+1})\right\},$$

where s is given by the condition

$$SL_{\text{opt}}^{\text{full}} + \sum_{i=j+d+1}^{s} \tau_i \leqslant \tau_x + 2K + \tau_s$$

and

$$SL_{\text{opt}}^{\text{full}} + \sum_{i=j+d+1}^{s+1} \tau_i > \tau_x + 2K + \tau_{s+1}.$$

The above optimal schedule length is obtained by including the tasks with index larger than $j + d$ in the subset of processors holding the tasks with indices less than or equal to $j + d$ until the schedule length is longer than the finish time of the next task to be scheduled two-hop away from P_x. Given the constraint C1, the analysis for the cases in which some tasks have to be scheduled on processors which are three hops away from P_x is similar. We have the following result on the BSA algorithm.

Theorem 3. The BSA algorithm gives optimal schedule length for a fork set F_x if:

(1) the processors are homogeneous and fully-connected with homogeneous links; or

(2) the processor network is a homogeneous hypercube with C1 satisfied; or

(3) the processor network is a homogeneous hypercube with C2 satisfied.

Proof. First, we observe that according to procedure *Serialization*, the tasks are injected in an increasing order of indices to the pivot processor, call it P_x. The tasks are then examined for possible migration to the neighbor processors one by one. Thus, for cases (1) and (2), only those tasks with indices larger than j will migrate to the neighbor processors. The resulting schedule length is optimal. For case (3), assume that $d < k - j \leqslant C_2^d$. The tasks with indices from $j + 1$ to $j + d$ will migrate to the neighbor processors of P_x. Afterwards, tasks with indices from $j + d + 1$ to s will not migrate to the neighbor processors because they cannot start earlier according to the expressions given earlier in the above discussion. A task T_t with $t > s$, however, will migrate to the neighbor processors because it can start earlier. The task may eventually migrate to processors which are at two hops from P_x depending upon whether it can start at a time $\tau_x + 2K + \tau_t$. As a result, the BSA algorithm constructs an optimal schedule. Similar arguments can be applied to the case when some tasks have to be scheduled to processors at three hops in an optimal schedule. □

4. Performance results

In this section, we present the experimental performance of the BSA algorithm and also compare it with the DLS algorithm [23], which was also designed for heterogeneous systems. The DLS algorithm is also a greedy algorithm in that it chooses a task for scheduling if its potential start time is the earliest and it has the largest *b-level*.

In our experiments, we applied the two algorithms to two suites of task graphs using a Sun Ultrasparc workstation. The first suite consisted of regular graphs representing a number of parallel applications including the mean value analysis [2], Gaussian elimination [4], Laplace equation solver [2], LU-decomposition [4], containing regular patterns of tasks and communication messages. Since these applications operate on matrices, the number of tasks (and messages) in their task graphs depends on the matrix dimension N. Each application has its own equation in terms of N for determining the exact number of tasks but all of the equations are O(N^2). We generated ten graphs for each application by varying N such that the graph size varies from approximately 50 to 500 with increments of 50. The average execution cost each task of the applications is about 150. Note that the graph structure and relative magnitudes of the execution costs in these applications are fixed according to the underlying algorithm modeled by the graph. However, the communication costs can be varied. We used a parameter called *granularity*, which is defined as the average execution cost divided by the average communication cost in a graph. Within each type of graph, we used three granularities: 0.1, 1.0 and 10.0. Thus, in a fine-grained (i.e., granularity = 0.1) application, the average communication cost is about ten times the average task execution cost. On the other hand, in a coarse-grained (i.e., granularity = 10.0) application, the average communication cost is only about 10% of the average task execution cost. In summary, the regular graphs suite contained 90 graphs (three graph types, ten sizes, and three granularities). The second suite of task graphs consisted of randomly structured graphs with sizes also varied from 50 to 500 with increments of 50. The execution cost of each task was randomly selected from a uniform distribution with range [100–200]. Again, three granularities (0.1, 1.0 and 10.0) were selected for each graph size. Unless otherwise state, the heterogeneity factors (i.e., h_{ix} and h'_{ijxy}) were selected randomly from a uniform distribution with range [1–50]. Thus, the nominal execution and commu-

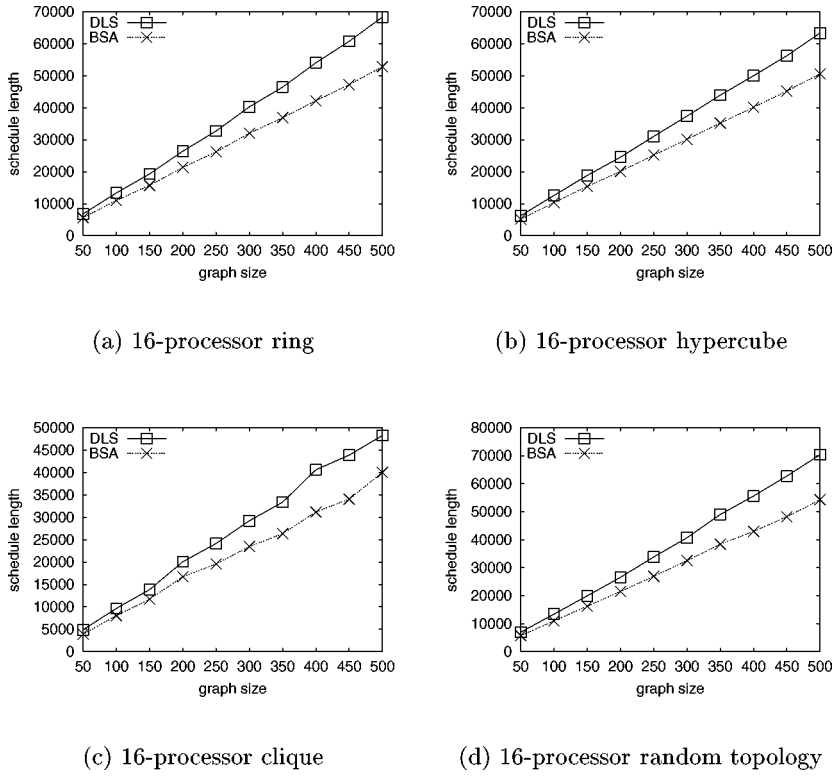

Figure 5. Average schedule lengths for the regular graphs with different graph sizes using four different network topologies.

nication costs in each graph represented the costs of the fastest processor.

To investigate the effect of processor network topology (i.e., processor connectivity), we used four different topologies in the experiments: 16-processor ring, 16-processor hypercube, 16-processor fully-connected network, and 16-processor randomly structured topology. The random topology was generated such that the degree of each processor ranged from two to eight.

In our first experiment, we compared the schedule lengths produced by the BSA algorithm with those by the DLS algorithm. For the regular graphs, it turned out that each algorithm generated similar performance for the three types of applications and thus, we computed the average schedule lengths across different applications. To examine the effect of graph size, we also computed the average schedule lengths across the three granularities. These average schedule lengths for the four topologies are shown in figure 5. From the plots, we make a number of observations:

- the BSA algorithm consistently outperformed the DLS algorithm;
- the improvement was about 20% and increased slightly with graph size;
- the improvement was slightly larger for lower processor connectivity (e.g., a ring); and
- both algorithms gave shorter schedule lengths for higher processor connectivity (e.g., a clique).

These observations can be explained as follows. First, notice that the DLS algorithm selects a task for scheduling if its start time is the earliest. This greedy decision is made without regard to the scheduling of subsequent tasks and hence, such a decision may be too "local" in that the communication links are not properly utilized leading to inefficient scheduling of communication messages of subsequent tasks. Indeed, when we looked into the schedules produced by the DLS algorithm more closely, we found that there were many cases in which a task could not be scheduled to a better time slot due to the inefficient scheduling of messages of previous tasks. The adverse effect of inefficient scheduling of messages and tasks was also more profound for increasing graph size and decreasing processor connectivity. In this aspect, the BSA algorithm has a better strategy because the messages are incrementally scheduled to suitable slots such that the finish times of tasks can be improved. When the connectivity was high, both algorithms generated shorter schedules because the message scheduling was easier to handle.

The results for randomly structured graphs are shown in figure 6. From these results, we can see that the BSA algorithm is robust in that it also consistently outperformed the DLS algorithm, despite that both algorithms generated longer schedules compared with the regular graphs. Next, we investigated the effect of granularity by computing the average schedule lengths across the graph sizes. The results for regular graphs are shown in figure 7. We can see that the granularity had significant impact on the performance of the scheduling algorithms. First, the schedule lengths increased

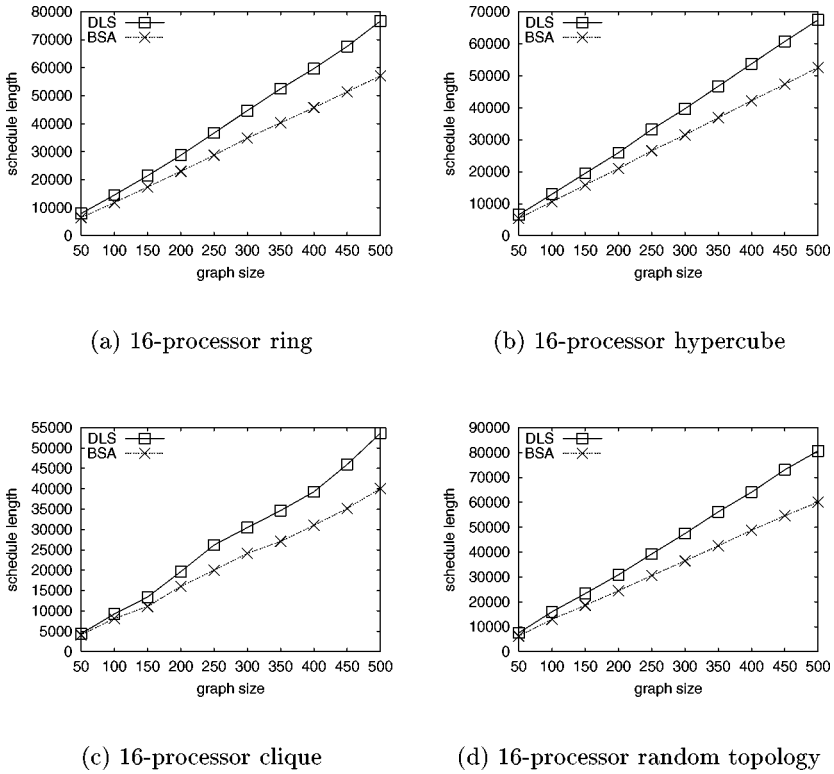

Figure 6. Average schedule lengths for the random graphs with different graph sizes using four different network topologies.

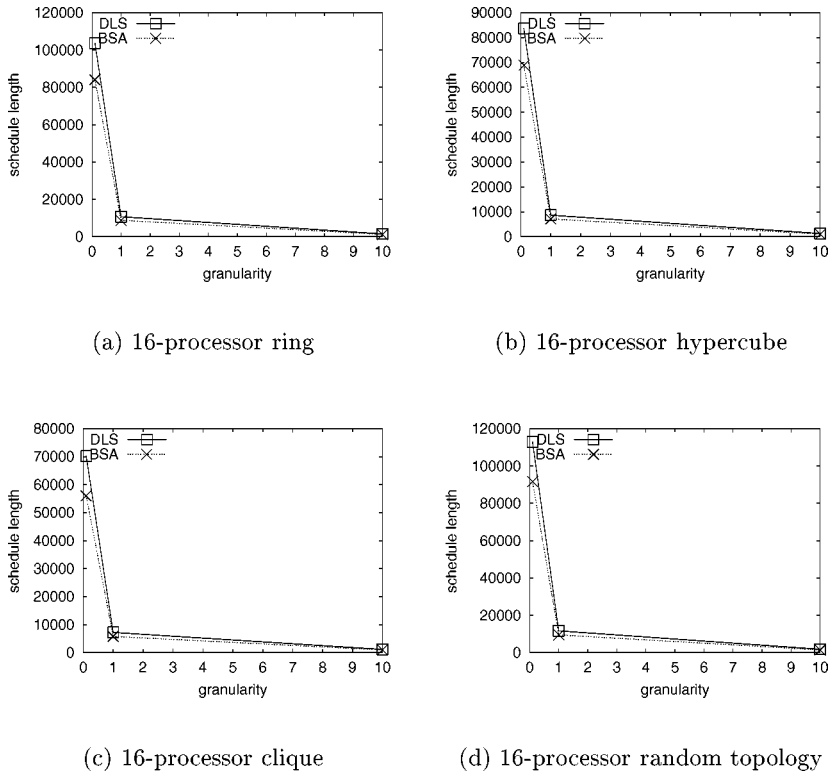

Figure 7. Average schedule lengths for the regular graphs with different granularities using four different network topologies.

sharply with decreasing granularity. At a low granularity (e.g., 0.1), the message scheduling was a dominant factor in determining the schedule length. Thus, the improvement of the BSA algorithm over the DLS algorithm was also larger for lower granularity. Finally, it is interesting to note that the effect of network topology was less significant from a

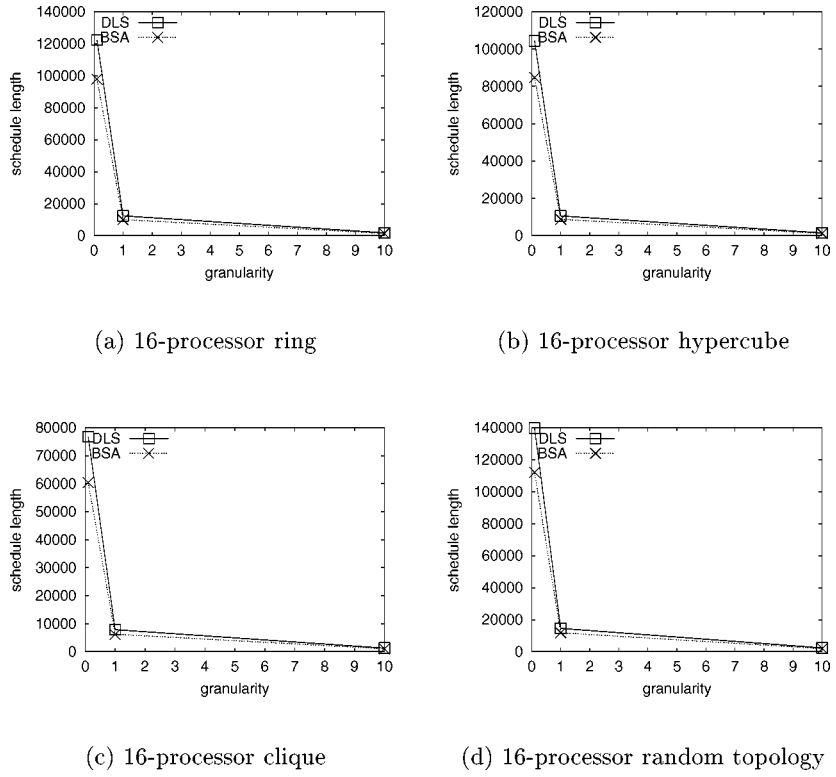

Figure 8. Average schedule lengths for the random graphs with different granularities using four different network topologies.

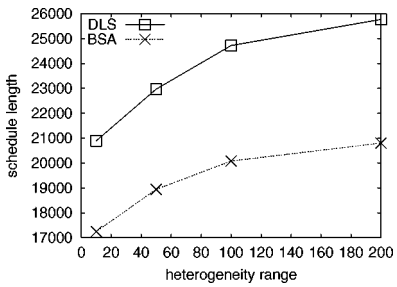

Figure 9. Effect of heterogeneity.

granularity perspective. Similar conclusions can be drawn from the results for randomly structured graphs, which are shown in figure 8.

We also investigated the effect of heterogeneity. For this purpose, we used ten different randomly structured task graphs with 500-task each (the granularity was 1.0). We chose the 16-processor hypercube topology and varied the range of heterogeneity as follows: [1–10], [1–50], [1–100], and [1–200]. Thus, a large range implies that there are more slow processors in the system. Again we computed the average schedule lengths, which are shown in figure 9. As can be seen, both algorithms generated longer schedules as the heterogeneity range increased. However, the rate of increase in schedule lengths generated by the BSA algorithm was lower than that of the DLS algorithm. This indicates that the BSA algorithm is more adaptive to a highly heterogeneous system. We also measured the running times of both algorithms, which were about the same because the two algorithms are of comparable time complexity.

5. Conclusions

In this paper we have presented a new algorithm, called the BSA algorithm, for scheduling and allocation of parallel tasks onto message-passing heterogeneous architectures using a novel task ordering strategy. The objective is to generate efficient solutions while simultaneously taking into account realistic parameters such as arbitrary execution and communication costs, network topology, contention on communication links, and heterogeneity of processors. The distinctive feature of the BSA algorithm is that it can adapt its tasks and messages scheduling decisions according to the given network topology. Messages are incrementally scheduled to suitable links during the optimization of the finish times of tasks. Heterogeneity of processors is also exploited by scheduling critical tasks to the fastest processors. Our performance evaluation study has demonstrated that the BSA algorithm is efficient, robust, and able to give consistent performance over a wide range of parameters.

Acknowledgements

This research was jointly supported by the Research Grants Council of the Hong Kong SAR under contract numbers HKUST619/94E and HKU7124/99E, and by a research initiation grant from the HKU CRCG. A preliminary version of this paper appeared in the Proceedings of the 1999 International Conference on Parallel Processing, Aizu-Wakamatsu, Fukushima, Japan, September 1999.

References

[1] I. Ahmad and Y.-K. Kwok, On exploiting task duplication in parallel program scheduling, IEEE Transactions on Parallel and Distributed Systems 9(9) (September 1998) 872–892.

[2] I. Ahmad, Y.-K. Kwok, M.-Y. Wu and W. Shu, CASCH: A software tool for automatic parallelization and scheduling of programs on message-passing multiprocessors, IEEE Concurrency (2000) to appear.

[3] M. Cosnard and M. Loi, Automatic task graphs generation techniques, Parallel Processing Letters 5(4) (December 1995) 527–538.

[4] M. Cosnard, M. Marrakchi, Y. Robert and D. Trystam, Parallel Gaussian elimination on an MIMD computer, Parallel Computing 6 (1988) 275–296.

[5] H. El-Rewini, T.G. Lewis and H.H. Ali, *Task Scheduling in Parallel and Distributed Systems* (Prentice-Hall, Englewood Cliffs, NJ, 1994).

[6] T. Fahringer, Estimating and optimizing performance for parallel programs, IEEE Computer 28(11) (November 1995) 47–56.

[7] R.F. Freund and H.J. Siegel, Heterogeneous processing, IEEE Computer 26(6) (June 1993) 13–17.

[8] M.R. Garey and D.S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness* (Freeman, 1979).

[9] A. Ghafoor and J. Yang, A distributed heterogeneous supercomputing management system, IEEE Computer 26(6) (June 1993) 78–86.

[10] K. Hwang, Z. Xu and M. Arakawa, Benchmark evaluation of the IBM SP2 for parallel signal processing, IEEE Transactions on Parallel and Distributed Systems 7(5) (May 1996) 522–536.

[11] M.A. Iverson, F. Ozguner and L.C. Potter, Statistical prediction of task execution times through analytic benchmarking for scheduling in a heterogeneous environment, in: *Proceedings of Eighth Heterogeneous Computing Workshop (HCW'99)* (1999) pp. 99–111.

[12] D. Kim and B.G. Yi, A two-pass scheduling algorithm for parallel programs, Parallel Computing 20 (1994) 869–885.

[13] Y.-K. Kwok and I. Ahmad, Dynamic critical path scheduling: An effective technique for allocating tasks graphs to multiprocessors, IEEE Transactions on Parallel and Distributed Systems 7(5) (May 1996) 506–521.

[14] Y.-K. Kwok and I. Ahmad, Efficient scheduling of arbitrary task graphs to multiprocessors using a parallel genetic algorithm, Journal of Parallel and Distributed Computing 47(1) (November 1997) 58–77.

[15] Y.-K. Kwok and I. Ahmad, FASTEST: A practical low-complexity algorithm for compile-time assignment of parallel programs to multiprocessors, IEEE Transactions on Parallel and Distributed Systems 10(2) (February 1999) 147–159.

[16] Y.-K. Kwok and I. Ahmad, Benchmarking and comparison of the task graph scheduling algorithms, Journal of Parallel and Distributed Computing 59(3) (December 1999) 381–422.

[17] Y.-K. Kwok and I. Ahmad, Static scheduling algorithms for allocating directed task graphs to multiprocessors, ACM Computing Surveys 31(4) (December 1999).

[18] M.G. Norman and P. Thanisch, Models of machines and computation for mapping in multicomputers, ACM Computing Surveys 25(3) (September 1993) 263–302.

[19] M.A. Palis, J.-C. Liou and D.S.L. Wei, Task clustering and scheduling for distributed memory parallel architectures, IEEE Transactions on Parallel and Distributed Systems 7(1) (January 1996) 46–55.

[20] V. Sarkar, *Partitioning and Scheduling Parallel Programs for Multiprocessors* (MIT Press, Cambridge, MA, 1989).

[21] B. Shirazi, M. Wang and G. Pathak, Analysis and evaluation of heuristic methods for static scheduling, Journal of Parallel and Distributed Computing 10(3) (November 1990) 222–232.

[22] H.J. Siegel, H.G. Dietz and J.K. Antonio, Software support for heterogeneous computing, ACM Computing Surveys 28(1) (March 1996) 237–239.

[23] G.C. Sih and E.A. Lee, A compile-time scheduling heuristic for interconnection-constrained heterogeneous processor architectures, IEEE Transactions on Parallel and Distributed Systems 4(2) (February 1993) 75–87.

[24] M.-Y. Wu and D.D. Gajski, Hypertool: A programming aid for message-passing systems, IEEE Transactions on Parallel and Distributed Systems 1(3) (July 1990) 330–343.

[25] T. Yang and A. Gerasoulis, List scheduling with and without communication delays, Parallel Computing 19 (1993) 1321–1344.

[26] J. Yang, A. Khokhar, S. Sheikh and A. Ghafoor, Estimating execution time for parallel tasks in heterogeneous processing (HP) environment, in: *Proceedings of the Fourth Heterogeneous Computing Workshop (HCW'94)* (1994) pp. 23–28.

[27] A.Y. Zomaya, M. Clements and S. Olariu, A framework for reinforcement-based scheduling in parallel processor systems, IEEE Transactions on Parallel and Distributed Systems 9(3) (March 1998) 249–260.

Yu-Kwong Kwok is an Assistant Professor in the Department of Electrical and Electronic Engineering at the University of Hong Kong. Before joining the University of Hong Kong, he was a visiting scholar for one year in the Parallel Processing Laboratory at the School of Electrical and Computer Engineering at Purdue University. His research interests include software support for parallel and distributed computing, heterogeneous cluster computing, and distributed multimedia systems. He is a member of the IEEE Computer Society and the ACM. He received his B.Sc. degree in computer engineering from the University of Hong Kong in 1991, the M.Phil. and Ph.D. degrees in computer science from the Hong Kong University of Science and Technology in 1994 and 1997, respectively.

E-mail: ykwok@eee.hku.hk

WWW: http://www.eee.hku.hk/∼ykwok

Ishfaq Ahmad is an Associate Professor in the Department of Computer Science at the Hong Kong University of Science and Technology. His research interests are in the areas of parallel programming tools, scheduling and mapping algorithms for scalable architectures, video compression, and interactive multimedia systems. He is director of Multimedia Technology Research Center at HKUST, where he and his colleagues are working on a number of research projects related to information technology, in particular in the areas of video coding and interactive multimedia systems in a distributed environment using high-performance computing for emerging applications. He has published over 100 technical papers in refereed journals and conferences. He has served on the program committees of numerous international conferences, and has guest-edited several journals. He is serving on the editorial board of IEEE Concurrency, Cluster Computing, and IEEE Transactions on Circuits and Systems for Video Technology. He is a member of the IEEE Computer Society. He received a B.Sc. degree in electrical engineering from the University of Engineering and Technology, Lahore, Pakistan, in 1985. He received his M.S. degree in computer engineering and Ph.D. degree in computer science, both from Syracuse University, in 1987 and 1992, respectively.

E-mail: iahmad@cs.ust.hk

WWW: http://www.cs.ust.hk/faculty/iahmad