Communication Contention Reduction in Joint Scheduling for Optical Grid Computing
(Invited Paper)

Yaohui Jin, Yan Wang, Wei Guo, Weiqiang Sun, and Weisheng Hu

State Key Laboratory on Advanced Optical Communication Systems and Networks,
Shanghai Jiao Tong University, Shanghai 200240, China
jinyh@sjtu.edu.cn

Abstract. Optical network, which can provide guaranteed quality of service (QoS) connections, is considered as a promising infrastructure for grid computing to solve more and more complex scientific problems. When optical links are regarded as resources and jointly scheduled with other grid resources, communication contention must be taken into consideration for efficient task scheduling. This paper models the optical grid computing as a communication-aware Directed Acyclic Graph (DAG) scheduling problem. To reduce the communication contention, we propose to use hop-bytes metric (HBM) heuristic to select computing resource. Simulation results show that the HBM approach combined with the adaptive routing scheme can achieve better performance in terms of normalized schedule length and link utilization.

Keywords: Optical Grid, DAG, Communication Contention.

1 Introduction

By using open middleware technologies [1], Grid computing enables the sharing, selection, and aggregation of a wide variety of geographically distributed computational resources (e.g. supercomputers, data sources, storage systems, instruments etc) together to solve more and more complex problems for scientific research. Optical circuit-switched (OCS) networking technologies are considered better suited to fulfill the QoS requirements, i.e., to offer huge capacity and relatively low latency, as well as dynamic control and allocation of bandwidth at various granularities [2][3]. Thus optical networking is expected to play an important role in creating an efficient infrastructure for supporting such advanced Grid computing applications, which is called optical Grid or photonic Grid [4].

Recently some significant researches have been done on the testbeds or architectures for optical Grid applications [5-9]. These efforts mainly aim to the integration of optical networks as Grid services, or to make optical circuit-switched networks more suitable to meet the Grid requirements, such as user-controlled capability, fast lightpath provisioning, and flexible dynamic control. However, few works focus on the scheduling problem for optical Grid computing in theoretical

P. Primet et al. (Eds.): GridNets 2008, LNICST 2, pp. 206 – 214, 2009.

details. A Grid computing application can be modeled as a Directed acyclic graph (DAG) [10]. The scheduling of Grid computing is to map the DAG onto the computational resources with efficient resource utilization. Although many algorithms have been proposed for DAG scheduling [11-14], these algorithms cannot be directly used for optical Grid computing applications. Because most of them assume an ideal communication system in which the Grid resources are fully connected and the communication between any two Grid resources can be provisioned whenever they need. These assumptions are not consistent with those of practical OCS networks in which a lightpath should be first setup before each communication and tore down after the communication finishes. When the lightpath is occupied by one communication, other communication cannot use it and therefore the contention in communication arises.

There are some attempts to incorporate communication contention awareness into DAG scheduling. A few algorithms ware proposed that consider network link contention [14] or end-port contention [17]. Sinnen and Sousa [18] propose a new graph model of the system network which is capable of capturing both end-point and network link contention. Agarwal et al [16] propose a Hop-bytes metric for task mapping in large parallel machines. Hop-bytes is the total size of inter-processor communication in bytes weighted by distance between the respective end-processors. Using a Hop-bytes metric based estimation function, in each iteration in the mapping algorithm, the more heavily communication task can be mapped onto the nearby processor.

A joint scheduling model of computing and networking resources for optical Grid application was proposed in [15] by incorporating the link communication contention of the optical networks into DAG scheduling. The optical network resource takes the form of a lightpath composed of a series of links which are viewed as network resources to be shared among Grid users like other traditional computing resources. In this paper, we investigate how to reduce the communication contention so as to minimize the schedule length in DAG scheduling under the joint scheduling model. In extending the classic list scheduling algorithm to fulfill the joint scheduling, we find there are basically two ways to reduce the communication contention: (1) Using adaptive routing scheme to detour the heavy traffic; and (2) Mapping task object to nearby grid resources to avoid long-hop communications. In this paper, we focus on the second contention reduction scheme and incorporate a Hop-bytes metric [16] into the grid resource selection phase in the algorithm.

The rest of this paper is organized as follows: In Section 2 we describe the joint scheduling model for optical grid computing. In Section 3, we redefine the HBM for DAG scheduling and elaborate on how to incorporate it into the resource selection phase. In Section 4, we provide simulation results to evaluate the performance. Finally, section 5 concludes this paper.

2 Joint Scheduling for Optical Grid Computing

In traditional DAG scheduling, networks are seldom thought of as resources. In this section, the optical networks are considered as network resources in the same way as processing and storage resources and all these resources can be jointly scheduled for DAG scheduling.

2.1 Resource Model

In optical networks, possible resources mainly include optical switch nodes and fiber links. Fig. 1 (b) depicts an example of optical Grid extended resource model in which there are 7 Grid resources and 4 optical switches. The adjacent optical switches are connected via the WDM fiber links. Each grid resource is connected to the optical switch via the access link. Therefore, the traffic from and to the end grid resources can be mapped onto wide-area SONET/SDH circuits or all-optical light-paths. In our model, the optical switch is assumed to be equipped with all-wavelength converters, thus there is no wavelength continuity constraint for routing.

Then our optical grid resources model can be formulated as an extended resource graph OGR = (**N**, **L**, *type*, *bw*, *d*), where **N** is a set of network nodes and **N**=**R**+**S**, where a node $r \in \mathbf{R}$ represents a grid resource and a node $s \in \mathbf{S}$ represents an optical switch. **L** is a set of undirected links and $\mathbf{L}=\mathbf{L_A}+\mathbf{L_T}$, where each link $l \in \mathbf{L_A}$ represents the access link between a grid resource and an optical switch, while a link $l \in \mathbf{L_T}$ represents the transmission link between two optical switches. The notation *type*(*r*) is the type of *r*, for example, 1 represents computer, 2 storage and 3 I/O device etc. The weight *bw*(*l*) represents the link *l*'s bandwidth and the weight *d*(*l*) denotes the distance of the link *l*.

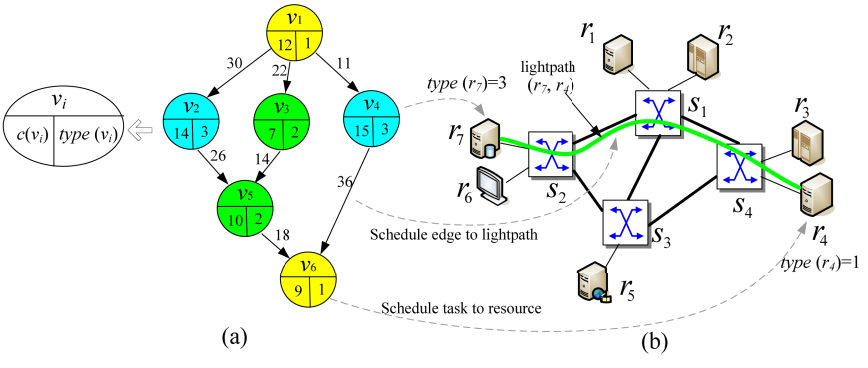

Fig. 1. Optical grid joint scheduling example. (a) DAG-modeled Grid application and (b) optical grid extended resource model.

2.2 Task Model

A Grid computing application can be modeled by a directed acyclic graph (DAG) [10]. We formulate the task model as G_{DAG} = (**V**, **E**, *type*, *c*, *w*), where **V** is a set of *v* tasks and **E** is a set of *e* edges between the tasks. Each edge $e_{mn} \in \mathbf{E}$ represents the precedence constraint such that task v_n can not start execution until v_m finishes. The notation *type*(*v*) represents the type of task *v*. Note that a task can only be scheduled onto the Grid resource of the same type. The weight *c*(*v*) denotes the average execution time required by *v* on a reference resource in the heterogeneous system. The weight *w*(*e*) denotes the data volume transmitted on the edge *e*.

In a given DAG, the set of all direct predecessor of task *v* is denoted by **pred**(*v*) and the set of all direct successors of *v* is denoted by **succ**(*v*). A task vertex *v* without

predecessors, **pred**(v)= ϕ, is named source node and if it is without successors, **succ**(v)= ϕ, it is named sink node. Fig. 1 (a) shows an example DAG.

2.3 Communication Contention Aware DAG Scheduling

To schedule a DAG onto the optical grid extended resource system, the awareness for communication contention can be achieved by edge scheduling, i.e., the scheduling of the edges of the DAG onto the links of the extended resource graph, in a similar manner how the task nodes are scheduled on the processing resources. Fig.1 exemplifies an optical grid scheduling which consists of two parts: One is task scheduling which is to schedule the computation tasks onto grid resources (e.g., $v_4 \rightarrow r_7$); the other is communication scheduling which is to schedule the edges onto each link along the lightpath (e.g., $e_{46} \rightarrow (r_7, r_4)$). The scheduled communication should start and end on all the links along the route simultaneously since the route is a cutting-though lightpath in the network without any store and forwarding.

```
1: Step 1: Determine the scheduling list
2:  Determine each task's bottom level of the DAG
3:   Sort each task  v∈V into a list LIST by decreasing order of
their bottom levels
4: Step 2:  Sequential scheduling over the list
5:  for each task v_n∈LIST do
6:    for each resource r∈R with type(v_n) = type(r) tentatively do
7:        for each v_m∈pred(v_n) in a definite order do
8:            if task v_m and v_n scheduled on two distinct resources
then
9:                   Find a route Rt = 〈l_1, l_2, …, l_k〉for edge e_mn
between the two resources
10:               Schedule e_mn on each link along Rt
11:           else
12:               neglect the communication cost of e_mn
13:           end if
14:       end for
15:       Schedule task v_n on resource r tentatively
16:    end for
17:    Select the resource r_min on which task v_n has earliest finish
time
18:     Schedule each incoming edge e_mn of v_n, v_m∈pred(v_n), on its
determined route
19:    Schedule task v_n on resource r_min
20:  end for
```

Fig. 2. The extended list scheduling (ELS) algorithm

The objective of DAG scheduling is to minimize the schedule length. The scheduling problem under our communication contention model has been proved to be NP-hard [18]. The heuristics therefore try to produce near optimal solutions in acceptable solving time. As list scheduling is one of the most common heuristics for the DAG scheduling, we extend the classic list scheduling algorithm to implement the joint schedule. The extended list scheduling (ELS) algorithm is outlined in Fig. 2.

Since the communication scheduling is included in the DAG scheduling, communication contention will naturally increase the schedule length. However, the ELS algorithm only describes how to implement communication contention aware DAG scheduling without any means to reduce the contention. From Fig.2, we can find there are basically two ways to alleviate the network resource contention. One way is to improve routing scheme (line 9 in Fig.2) and the other way is to improve computing resource selection scheme (line 17 in Fig.2). There are generally three approaches to establish lightpath in optical networks, fixed routing, fixed-alternate routing and adaptive routing [20]. We will discuss resource selection scheme in the following section.

3 Hop-Bytes Metric Based Grid Resource Selection Scheme

To improve the resource selection scheme, we have a motivation that is to try to map the task onto the nearby resource to reduce this link contention by introducing a Hop-Bytes Metric (HBM) [17] into the grid resource selection phase for DAG scheduling. As HBM is originally used to judge the quality of the solution produced by the independent job mapping algorithm, we now redefine the HBM for the communication contention aware DAG scheduling.

Definition: The HBM of task v_i scheduled on computing resource r is defined as the total size of the data volume in bytes carried by each incoming edge of v_i weighted by the hops of the route the edge scheduled on it

$$hb(v_i, r) = \begin{cases} \sum_{v_j \in \mathbf{pred}(v_i)} w(e_{ji}) \times h\left(rsc(v_j), rsc(v_i)\right) & \mathbf{pred}(v_i) \neq \phi \\ 0 & otherwise \end{cases} \quad (1)$$

where e_{ji} is the incoming edge from predecessor task v_j to the current unscheduled task v_i, $w(e_{ji})$ is the weight, i.e., the data volume in bytes, of the incoming edge e_{ji}, $rsv(v_j)$ and $rsv(v_i)$ denote the grid resources where task v_j and v_i are allocated respectively, $h(rsv(v_j), rsv(v_i))$ is the number of hops of the route between resource $rsv(v_j)$ and $rsv(v_i)$.

In the previous ELS algorithm, there is only one objective in the grid resource selection phase, i.e., $\text{minimize}\{t_f(v,r)\}$, where $t_f(v,r)$ denotes the finish time of task $v \in \mathbf{V}$ on grid resource $r \in \mathbf{R}$. When the Hop-bytes metric is taken into consideration for the reduction of long-hop communication, we will have tow objectives, i.e., $\min\left[t_f(v,r), hb(v,r)\right]$, for all $r \in \mathbf{R}$.

There are many approaches to solve multi-objective or multi-criteria problem [21] and here we use a basic weighted sum method. Since the finish time and HBM are different metrics, we first normalize them to the same measurement dimension. At each resource selection phase, we tentatively schedule the current task v on all the resources and record the maximum and minimum finish time of v, denoted as $t_{f\max}(v)$ and $t_{f\min}(v)$, as well as the maximum and minimum HBM of v, denoted as

$hb_{\max}(v)$ and $hb_{\min}(v)$. Then the new normalized finish time and HBM are given as below,

$$t_f'(v,r)=\frac{t_{f\max}(v)-t_f(v,r)}{t_{f\max}(v)-t_{f\min}(v)}\in[0,1],\ hb'(v,r)=\frac{hb_{\max}(v)-hb(v,r)}{hb_{\max}(v)-hb_{\min}(v)}\in[0,1] \quad (2)$$

The weighted sum objective function is written as $FH(v,r)=\lambda\cdot t_f'(v,r)+(1-\lambda)\cdot hb'(v,r)$, $\lambda\in[0,1]$. In the following step, we can directly select the best resource $r_{\min}$ with the maximum $FH(v,r)$, i.e.

$$FH(v,r_{\min})=\max_{r\in\mathbf{R}}\left[FH(v,r)\right],\quad r_{\min}\in\mathbf{R},\lambda\in[0,1] \quad (3)$$

When there is more than one resource having the same maximum $FH(v,r)$, the best one is selected using first-fit strategy as mentioned in section 5.1.

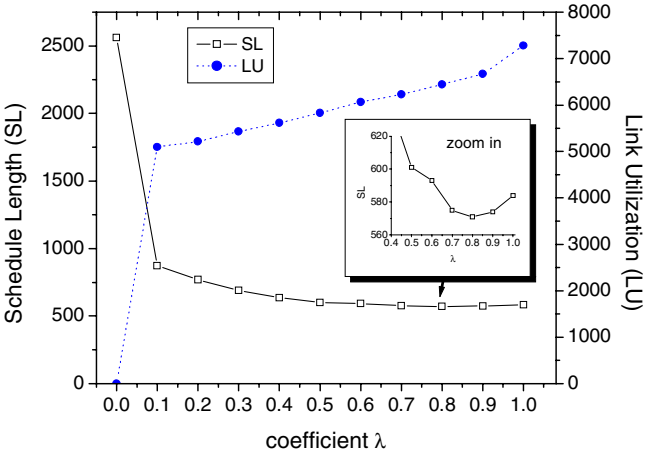

Fig. 3. Schedule results under various coefficient λ for weighted sum method in terms of schedule length and link utilization

For the weighted sum method the problem is how to determine the value of coefficient λ. There are two extreme values for λ, λ=0 and λ=1. if λ=0, the objective becomes to minimize the HBM for resource selection. If λ=1, the objective is to minimize the finish time. As we have mentioned before, the importance of $t_f(v,r)$ is higher than $hb(v,r)$ in order to get the minimal schedule length, then λ should be more than 0.5. So we have $0.5<\lambda<1$.

Next we will get the more precise value of λ through simulation. We randomly generate DAG with 250 task nodes. Then the DAG is scheduled onto 16 resources interconnected by a 16-node NSFNET. Each link has only one unit bandwidth. We get the average result over 100 simulations. Fig. 3 depicts the schedule length and link utilization that is defined as the sum of multiplication of occupied time and occupied

bandwidth on all the links under different λ. We can get relatively smaller schedule length when λ is around 0.8, as can be seen in the zoom-in inset of Fig. 3. When we change the network topology (e.g. 16-node Mesh-torus) or change the DAG size (the number of DAG nodes ranging from 128 to1024), we can get the similar simulation results and λ =0.8 can always produce relatively good results in terms of both schedule length and link utilization.

4 Simulation Study

In this section, we evaluate the performance of the ELS algorithm with Hop-bytes metric based grid resource selection scheme through simulations.

Two typical network topologies are employed in the following simulations, one is a 64-node mesh-torus and the other is a 46-node USNET. The purpose of minimizing the link capacity is to maximize the communication contention. We also employ two routing schemes and three resource selection schemes in the following simulations (see Table 1).

Table 1. Different routing and resource selection schemes in the simulation

FR	Fixed shortest path routing scheme in which each route is pre-computed
AR	Adaptive routing scheme which can find an earliest start route for current communication according to current network state.
EFT	Earliest finish time method for resource selection scheme
HB_WS	HBM based weighted sum method for resource selection scheme

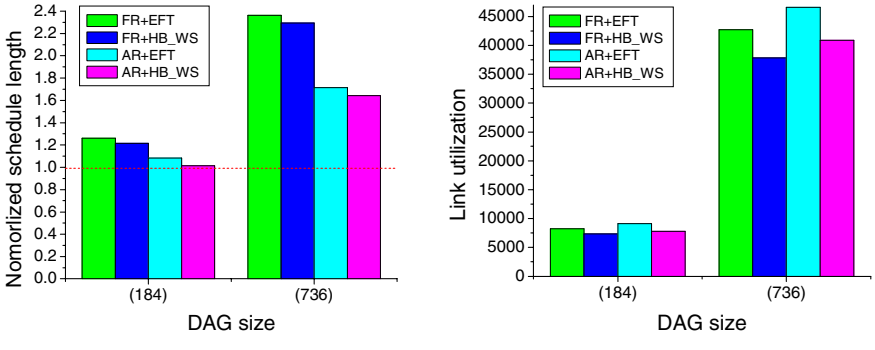

Fig. 4. Scheduling results of 4 combinations of two routing and two resource selection schemes in terms of (a) normalized schedule length and (b) link utilization vs. DAG size (184, 736)

We use the same random DAG generator in [15]. The average out-degree of DAG is 2. The DAG node weight is taken randomly from a uniform distribution [18] around 10, thus the average node weight is 10. The communication-computation-ratio (CCR) is selected to be 2 to simulate the application with more communications.

Moreover, it is also assumed that all the DAG nodes and grid resources have identical type and all the grid resources are homogeneous. The performance results are the average over 100 simulations.

We compare 4 scenarios of combination of different routing and resource selection schemes. The scheduling results in terms of normalized schedule length and link utilization are given in Fig. 4. Compared with the results of FR+EFT, we can find that WS scheme has more contribution in reducing link utilization, but little effects in reducing schedule length. AR scheme can contribute to much shorter schedule length, but at the cost of higher link utilization (which is not desirable for public shared network). When we combine the two contention reduction schemes together (i.e., AR+WS) in the DAG scheduling, it can be seen that most of the communication contention can be removed, producing the minimal schedule length with relatively lower link utilization (even lower than F1R+EFT by 4.2%).

5 Conclusions

Optical grid computing can be modeled as communication contention aware DAG scheduling over optical circuit switched networks. There are basically two ways to reduce the communication contention in the ELS algorithm: adaptive routing scheme and Hop-bytes based grid resource selection scheme. This paper mainly focused on the latter problem. We incorporated a Hop-bytes metric into the resource selection and proposed two methods, multilevel method and weighted sum method, to schedule the task onto the nearby resources. Simulation shows that the weighted sum method is better than multilevel method in terms of network resource utilization. We also demonstrated that the Hop-bytes based resource selection scheme contributes in lower resource utilization, while the adaptive routing scheme has the advantage in the reduction of schedule length. When we employ the two schemes together, both of the merits can be achieved and most of the communication contention can be avoided, leading to the smallest schedule length with relatively lower link utilization.

Acknowledgements

This work is supported by the China 863 Program and National Nature Science Foundation Committee of China under grants 2006AA01Z247, 60672016, 60502004.

References

1. Froster, I., Grossman, R.: Data integration in a bandwidth-rich world. Commun. ACM 46, 50–57 (2003)
2. Veeraraghavan, M., et al.: On the use of connection-oriented networks to support grid computing. IEEE Commun. Mag. 44, 118–123 (2006)
3. Barker, K.J., et al.: On the Feasibility of Optical Circuit Switching for High Performance Computing Systems. Supercomputing (2005)
4. Simeonidou, D., et al.: Optical network infrastructure for grid. Global Grid Forum Document, GFD.36. Grid High Performance Networking Group (2004)

5. Baroncelli, F., et al.: A Service Oriented Network Architecture suitable for Global Grid Computing. In: Conference on Optical Network Design and Modeling (2005)
6. Figueira, S., et al.: DWDM-RAM: Enabling Grid Services with Dynamic Optical Networks. In: IEEE International Symposium on Cluster Computing and the Grid (2004)
7. Lehman, T., et al.: DRAGON: A Framework for Service Provisioning in Heterogeneous Grid Networks. IEEE Commun. Mag. 44, 84–90 (2006)
8. Zheng, X., Veeraraghavan, M., Rao, N.S.V., Wu, Q., Zhu, M.: CHEETAH: Circuit-switched high-speed end-to-end transport architecture testbed. IEEE Commun. Mag. 43, 11–17 (2005)
9. Smarr, L.L., et al.: The OptIPuter. Commun. ACM 46(11), 59–67 (2003)
10. Sarkar, V.: Partitioning and Scheduling Parallel Programs for Execution on Multiprocessors. MIT, Cambridge (1989)
11. Topcuoglu, H., et al.: Performance-Effective and Low-Complexity Task Scheduling for Heterogeneous Computing. IEEE Trans. on Parallel Distrib. Syst. 13, 3 (2002)
12. Wu, M.Y., et al.: Hypertool: a programming and aid for message-passing systems. IEEE Trans. Parallel Distrib. Syst. 1, 330–343 (1990)
13. Yang, T., et al.: DSC: scheduling parallel tasks on an unbounded number of processors. IEEE Trans. Parallel Distrib. Syst. 5, 951–967 (1994)
14. Kwok, K., Ahmad, I.: Link Contention-Constrained Scheduling and Mapping of Tasks and Messages to a Network of Heterogeneous Processors. Cluster Computing 3, 113–124 (2000)
15. Wang, Y., et al.: Joint scheduling for optical grid applications. Journal of Optical Networking 6, 3 (2007)
16. Agarwal, T., et al.: Topology-aware Task Mapping for Reducing Communication Contention on Large Parallel Machines. In: IEEE Parallel and Distributed Processing Symposium (2006)
17. Beaumont, O., et al.: A Realistic Model and an Efficient Heuristic for Scheduling with Heterogeneous Processors. In: Proc. 11th Heterogeneous Computing Workshop (2002)
18. Sinnen, O., Sousa, L.A.: Communication contention in task scheduling. IEEE Trans. Parallel Distrib. Syst. 16, 503–515 (2005)
19. He, E., et al.: AR-PIN/PDC: Flexible Advance Reservation of Intradomain and Interdomain Lightpaths. In: IEEE Global Telecommunications Conference (2005)
20. Zang, H., et al.: Dynamic Lightpath Establishment in Wavelength-Routed WDM Networks. IEEE Commun. Mag. 39, 100–108 (2001)
21. Das, I.: Multi-Objective Optimization (1997), http://www-fp.mcs.anl.gov/otc/Guide/OptWeb/multiobj/