Prioritized Contention Resolution Scheme for Grid Services over OBS Networks

Burak Kantarci, Sema Oktug, and D.Turgay Altilar
Istanbul Technical University,
Department of Computer Engineering, Istanbul, Turkey
(bkantarci, oktug, altilar@itu.edu.tr)

Abstract

In this paper a novel contention resolution scheme, namely Expanded Priority Vector (EPV) approach is proposed to support differentiated Grid service over OBS networks. EPV is an enhancement to a previously proposed application-aware contention resolution scheme for the OBS-based-Grid services in the literature. By using the EPV, we keep all the advantages of the application-aware contention resolution. Moreover, we introduce fairness among the contending jobs of the same class by taking care of the two least significant digits of the vector. Those two digits pay attention to the remaining distance to the resource and the job size. By simulations, we show that EPV based policy reduces the blocking probability for all classes of submitted jobs.

1. Introduction

As a result of the high computational and storage demand in scientific applications in the technical community, local resources became suffering from less efficient computational power. This phenomenon has caused the philosophy of Grid Computing to develop [1]. The most favorable example of an high computational demand application is the particle physics experiments. Those experiments require a computational resource of several petabytes/year and expected to rise up to several exabytes/year. Therefore the users are expected to access and use remote resources that are distributed in different locations forming a grid.

Basically, the grid consist of heterogeneous computing and storage units that are connected to each other via LANs, WANs or metro networks, and that are distributed into multiple administrative domains. Besides these, the grid is also supposed to have no central administration unit.

Several types of data and processing intensive grid applications can be considered in order to make the proposed architecture be based on. The first one is high performance computing and visualization applications. Multimedia video editing is a good example for this kind of computationally intensive jobs and that require TBytes of storage and TFlops of computational resources. An online visualization application is another type of job in which the computational capacity requirement is as much as a few thousands of GFlops, and a few tens of milliseconds of transmission and processing latency even though the storage requirement is not as much as the previous application example.

Today, in order to build the communication backbone for the grid, optical networking technologies are employed [2]. Most of the optical networking technologies are based on optical circuit switching (OCS). In OCS, wavelength division multiplexing is used on the fibers. In order to transmit a job from a source to destination, the source has to set up a connection with the destination over a pre-determined routing path and a wavelength. Unless the connection is terminated, the transmission resources are dedicated to the source.

For a high performance computing and visualization application, the required transmission times of jobs are as short as a few hundred microseconds or tens of milliseconds. However, in an OCS architecture, the main overhead of the job submission into the grid comes from the connection setup and release times that are in hundreds of milliseconds. Besides these, since the resources are devoted to a connection until the connection is released, OCS leads to a waste of the grid resources.

The data intensive user applications need high access bandwidth. As we state above, dedicating the network resources to a job submission in a connection oriented structure increases the cost of constructing a grid dramatically. Although the dedicated grid architecture provides a job to be submitted at full wavelength capacity, as a result of the grid service requests' non-deterministic arrival, this may not be the optimal case. In many of the grid applications, the job sizes are significantly short, which leads to the holding time of

978-0-7695-3135-9/08 $25.00 © 2008 IEEE
DOI 10.1109/CNSR.2008.55

wavelengths to be significantly lower than the connection setup delays. Besides these, current applications require a job scheduler to queue jobs and allocate resource for them which causes another delay in the response time. Therefore a decentralized scheduling structure has to be embedded into the grid architecture. As a result of these points, a non-dedicated, high bandwidth offering, flexible infrastructure that is integrated with a decentralized scheduling intelligence is emergent for the next generation Grids.

At this point optical burst switching (OBS) appears as the futuristic infrastructure for the next generation grid architecture. OBS has been recently proposed as the transmission technology for the grid [3].

The rest of the paper is organized as follows: In Section 2, we give a brief information on OBS-Based Grid, together with the the contention resolution schemes in Grid services over OBS. The details of the proposed contention resolution scheme are explained in Section 3. In Section 4, we give the simulation results obtained. Finally, we conclude the paper by giving future directions in Section 5.

2. OBS-based Grid

The most significant advantage of the OBS in comparison to the other optical networking technologies is that the network resources are held only within the duration of a burst. By using the Just-Enough-Time (JET) protocol, the OBS network gains a bufferless qualification. By employing OBS in the Grid Network infrastructure, the user jobs can be transmitted efficiently at full wavelength capacity. Besides these, by separating the transmission of the job (burst) between a control plane and a data plane, the latency to
set-up/release a connection between the resource server and the user is eliminated. Since the burst header is processed electronically, the intermediate routers can perform intelligence scheduling and resource discovery mechanisms. Therefore the proposed infrastructure is enabled to develop efficient Grid protocols.

The center of the Grid consists of interconnected intelligent OBS routers. The intelligence of the OBS routers comes from determining the job that is carried in an optical burst to be destined to the optimal destination.

The main difference between the Grid request and an Internet application is that the Grid request is submitted without a pre-determined destination. The Grid user is not interested in the location of the computational or storage resource while he needs the submitted request to be handled and answered with the specified requirements and within an acceptable delay. The Grid user accesses the grid by submitting his job by coding into an optical burst and just waits for the job to be handled. The resource allocation is performed by the intelligence of the OBS routers. Therefore, user to Grid traffic is generated based on the *anycast* paradigm since no destination is specified by the user.

Signaling mechanism in OBS-based grid consists of two-step Just Enough Time (JET). Burst header and payloads are separated into two categories as active and passive. The first step is the resource recovery stage. An active burst header informs the intermediate active routers on the incoming resource request. The payload of the active burst header arrives at the intermediate active switches an offset time later. The user is informed on the result of the resource discovery process. Whenever a route is discovered for the resource allocation, the job itself is transmitted in a passive optical burst by means of optical burst forwarding [4]. In the rest of the paper, the terms *job* and *passive burst* are used interchangeably.

2.1. Contention resolution schemes in Grid-OBS network

In the Grid-OBS upon the arrival of the resource discovery, the user releases the request in the passive burst [5]. In order to support the QoS differentiation the OBS core protocols are concerned with the contention resolution problem. There are previously proposed contention resolution schemes namely *shortest drop policy (SDP), latest drop policy (LDP), deadline-based drop policy (DDP)* [6].

In the SDP scheme, the burst scheduler searches for an unscheduled timeslot among the wavelengths. If an unscheduled timeslot is found, the burst carrying the job is scheduled in the available timeslot. Otherwise, the shortest one among the contending bursts is dropped, and the burst is scheduled on that wavelength.

In the LDP scheme drops the burst that arrives the latest among the contending group.

DDP is proposed in [6] and uses the *tolerate time($\in$)* to determine which burst to be dropped and retransmitted. Each burst is transmitted with its QoS vector. One of the members of the QoS vector is the job completion time, *C*. The job completion time is determined and specified in the service level agreement. The computation time of a job is represented by *E*. When a job is released at time *t*, then the tolerate time is calculated as follows:

$\in = C - (t + E)$

DDP uses this tolerate time to resolve contention and support QoS. Therefore, a job with a short tolerate time is supposed to have high priority and vice versa.

Thus, whenever a burst contends with the bursts that are scheduled on all the wavelengths on a fiber, the burst that has the longest tolerate time is dropped and retransmitted to resolve contention. The three schemes are illustrated in Figure 1. In the figure, B_i and B_j represent the bursts that arrive at time i and j respectively. The uppermost scenario in the figure is the LDP. Here, $B(t)_i$ and $B(t)_j$ are the arrival times of the bursts respectively. Since $j < i$, B_j is scheduled in case of an overlap. The second policy in the figure is a sample scenario from the SDP. Since size of B_j is greater than the size of B_i, B_j is supposed to be scheduled on the channel when they overlap. The last policy is the DDP. The deadlines $\in$ are compared, and B_j is scheduled on the channel as the burst having closer deadline time $B(\in)_j$ to the timeslot of overlap.

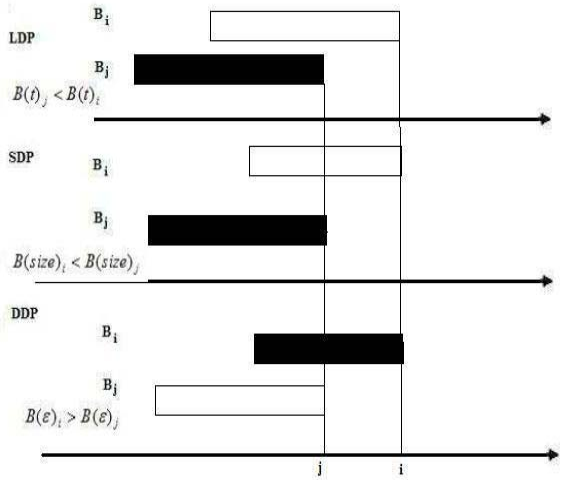

Figure 1. Previously proposed contention resolution schemes

In [6], it is shown that SDP causes the highest blocking probability. On the other hand, in DDP, the jobs are differentiated based on three CoS categories according to their deadlines. The authors define the *1st Class* bursts (highest priority) with 0-3 ms deadline, *2nd Class* bursts with 6-9 ms deadline, and the *3rd Class* bursts with 9-12 ms deadline. The simulations performed under a small sample topology show that DDP decreases the blocking probability of the high priority bursts when compared to LDP.

3. Expanded Priority Vector (EPV) for Contention Resolution

In order to support QoS between user jobs, we introduce a new contention resolution scheme that also considers fairness between different service classes. We enhance the DDP scheme by using the so called expanded priority vector (EPV) mechanism. The EPV has three digits, EPV_2, EPV_1, EPV_0, and it is coded into the passive bursts. The three digits of EPV are coded as follows:

EPV_2: CoS value of the job
EPV_1: Residual distance to the resource domain.
EPV_0: A discrete value for the job size (burst length)

The jobs are classified according to their tolerate time; the longer tolerate time leads to the less priority and vice versa. In our work, we assume that the CoS value can have 3 different values as follows: 3 for Class-1 jobs, 2 for Class-2 jobs, and 1 for Class-3 jobs. Residual distance to the resource domain is in terms of hop count. However, the bursts that are closer to the resource should be prioritized when compared to the bursts that are further from the resource domain. Therefore, we set the EPV_1 field to a value of *(N- hop_count)* where N is the number of routers in the Grid. We make a similar differentiation in job size as it is done in [7]. We classify the bursts in which the jobs are coded as *long, medium, and short* size. Since longer bursts are more likely to contend [8], in order to increase to ratio of handled job submissions, we prioritize the shorter jobs when the jobs of the same class and the of the same residual distance contend. Therefore for we define three threshold values, namely STh_{short}, STh_{medium}, *and* STh_{long} to determine whether a job is *short, medium or long* respectively. If the submitted job is of short size, then the EPV_0 field of the EPV is set to 3, else if it is of medium size, then the EPV_0 field is set to 2, otherwise it is set to 1. By using these values a hexadecimal priority factor is produced for each job at its release time, and the time that it is switched at the intermediate routers. The hexadecimal priority factor is as follows:

$$EPV_{Factor} = \sum_{i=0}^{2} 16^i . EPV_i$$

In computation of the EPV_{Factor}, the factor *16* is determined empirically based on the topology we use to employ the Grid. At each intermediate router, this EPV_{Factor} parameter is re-computed since the distance traveled changes. For an incoming passive burst, an intermediate router searches for an unscheduled timeslot in each wavelength. If it cannot find an available timeslot, it detects contention on the fiber. In case of a contention, the router discards the job that has the lowest EPV_{Factor} value. The tolerate time of the discarded job is updated and dilated through another outgoing port of the router. If the dilated path violates the tolerate time, then the source node is informed that the job submission is blocked.

As a result of the employment of the EPV_1 and EPV_0 fields, fairness is guaranteed among the jobs

belonging to the same priority level. Therefore, this contention resolution scheme also leads to a decrease in the blocking priority of Class-2 jobs together with Class-1 jobs.

4. Simulation Work

We develop our simulations in Visual C++ and run on a Pentium 4 3.00GHz with 3.50 GB memory space. We use the NSFNET topology shown in Figure 2 as the Grid infrastructure. The link weights indicate the distance in kilometers between the resource domains. It is assumed that, each link consists of 16 wavelengths of 10Gbps bandwidth. It is also assumed that the destination resource domains are determined by the active bursts before the transmission of passive bursts carrying the jobs. The average passive burst header processing time is taken as $25\mu s$ while the average switch reconfiguration time is taken as 100 ns [9].

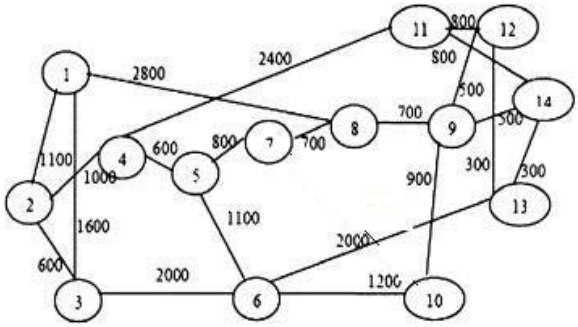

Figure 2. NSFNET topology is used as the infrastructure for the Grid

An initial offset time of $100\mu s$ is introduced to the passive burst between itself and its corresponding passive burst header. User requests are carried in IP packets under a self-similar traffic with H=0.8. The IP packets are generated based on the realistic internet measurements [10]. Each node represents a resource domain in the NSFNET topology. Each node can source user requests at line rate 80 Gbps. Simulation scenarios are run for virtual durations of five minutes. Each data point represent the average of five runs and for each point 90 % confidence interval is also shown in the graphs.

As we state in the previous section, to set the EPV_0 field to an appropriate value, we use three size threshold values, STh_{short}, STh_{medium}, and STh_{long} equal to 500KB (200 μs on 10 Gbps), 750 (600 μs on 10 Gbps) KB, and 1MB (1ms on 10 Gbps) respectively. When coding the jobs into the bursts the time-and-size threshold based hybrid burst assembly algorithm [11] is used where we set the time threshold value to 1ms, and the size threshold to 1MB.

Considering the propagation delays on the links of the NSFNET topology, we re-define the tolerate time intervals for the three classes as [0-5]ms, (5-10]ms, and (10,15] ms. We take our results for two different scenarios. The first part of the results are taken when the jobs are distributed among the three classes with equal probability (uniform CoS distribution). The second part presents a heterogenous CoS distribution where the Class-1, Class-2, and Class-3 jobs contribute the 25%, 42%, and 33% of the total requests respectively.

4.1. Results Taken under uniform CoS distribution

The simulation results that are taken under uniform CoS distribution, we compare the performance of the proposed scheme that uses EPV with two previously defined contention resolution schemes, namely LDP and DDP. Figure 3 represents the overall burst blocking probability for the three schemes. As expected, LDP has the highest blocking probability since it is the simplest scheduling mechanism. DDP and EPV lead to significantly less blocking probability. As it is seen in the Figure, EPV decreases the overall probability more than DDP since it considers the residual distance to the resource and the job size.

Figure 4 presents the blocking probabilities for the three schemes when $\in$ is between 0 and 5 ms. The QoS satisfying schemes (DDP and EPV) significantly lead to less blocking probabilities when compared to LDP. It is also observed that, the EPV scheme performs as well as the DDP scheme. At some load levels, EPV performs even better than DDP since it considers additional parameters other than the tolerate time. Therefore it can be said that, using an expanded priority vector leads to as low blocking probability level as the DDP for the Class-1 jobs.

In Figure 5, the blocking probabilities for the Class-2 jobs are given for the three techniques when $\in$ is between 5 and 10 ms. As expected, the difference between DDP and LDP is less than the one in Figure 4 since DDP aims to handle as much Class-1 requests as it can. However, EPV also serves better to the Class-2 jobs since it uses a drop policy based on an expanded priority vector factor. This factor also introduces intra-class fairness together with the inter-class QoS

assurance.

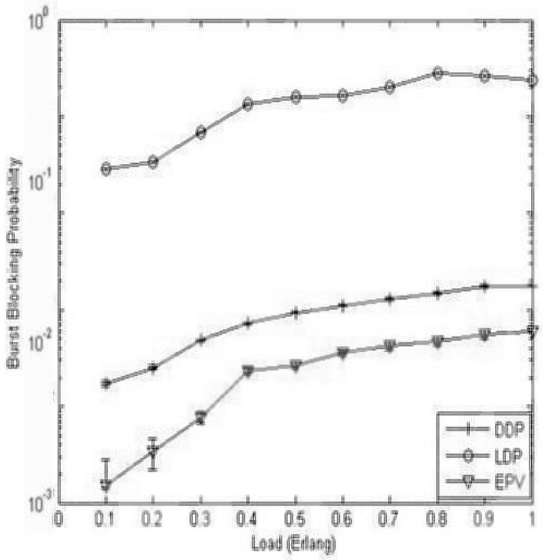

Figure 3. Overall Burst Blocking Probability in uniform CoS distribution

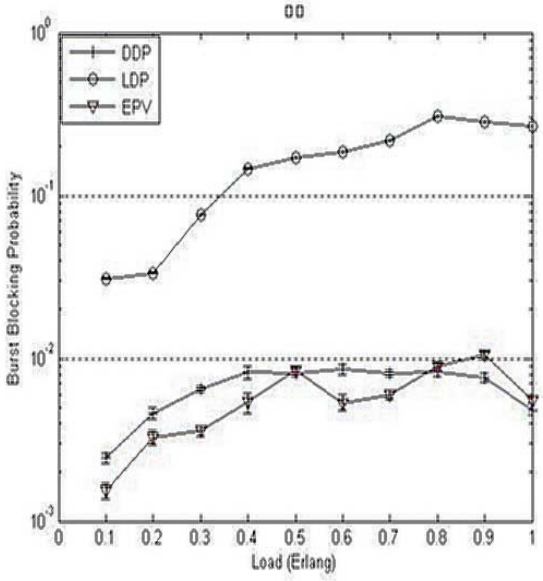

Figure 4. Burst Blocking Probability in uniform CoS distribution when ∈=0-5 ms

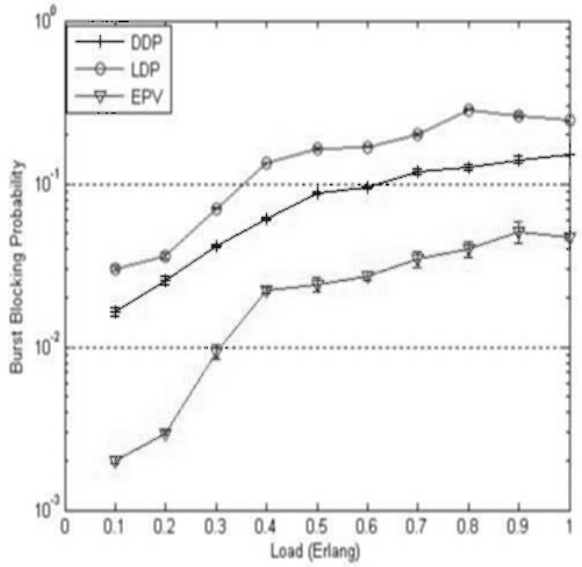

Figure 5. Burst Blocking Probability in uniform CoS distribution when ∈=5-10 ms

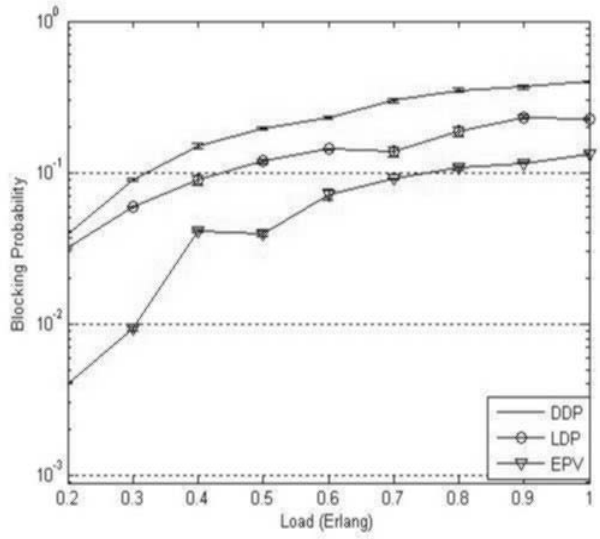

Figure 6. Burst Blocking Probability in uniform CoS distribution when ∈=10-15 ms

As it is seen in Figure 6, DDP performs worse than LDP. The reason is due to LDP's equivalent service to all of the classes. Since DDP tends to drop as much as Class-3 jobs to reserve as much as Class-1 jobs, it leads to a high blocking probability for the low priority level classes. EPV also lead to higher blocking probability when it is compared with the Class-1 and Class-2 blocking probabilities. However, due to the

employment of additional parameters, it also serves better to even low priority level jobs.

4.2. Results Taken under heterogeneous CoS distribution

The second part of our simulation results are taken under heterogeneous CoS distribution where 25% of the submissions have tolerance time between 0-5ms, 42% of the submissions have a tolerance time between 5-10ms, and 33% of the submissions have a tolerance time between 10-15ms. Since we have seen that LDP has the worst performance for overall blocking probability, Class-1, and Class-2, here, we focus on the performance comparison of DDP and EPV.

In Figure 7, the overall blocking probabilities of DDP and EPV are given. The results are so similar to those in Figure 3. The reason is the same as we state in the previous subsection. Since EPV considers QoS assurance together with the intra-class fairness, the degraded blocking probability for each class is projected to overall blocking probability as a decrease.

In Figure 8, comparison of EPV and DDP is shown when ϵ is between 0 and 5 ms. EPV seems to show a better performance in comparison to DDP except the load levels 0.7 and 0.8 Erlang. However, even in these load levels, blocking probability of EPV does introduce increase in the blocking probability. It seems that, under heterogeneous CoS distribution scenario, EPV performs better than that under uniform CoS distribution scenario. The main reason for this performance enhancement is the slight decrease in the ratio of Class-1 jobs.

Figure 9 and Figure 10 show the performance of DDP and EPV in terms of blocking probability when ϵ is between 5-10ms and 10-15 ms respectively. As we state above, EPV again serves better to both Class-2 and Class-3 jobs. The same reason holds here for this behavior. A burst discard policy which prioritizes the bursts of the same class, first based on their residual path lengths, and then their sizes provides a significant amount of the bursts to survive. Therefore for all classes, EPV serves significantly better in terms of burst blocking probability.

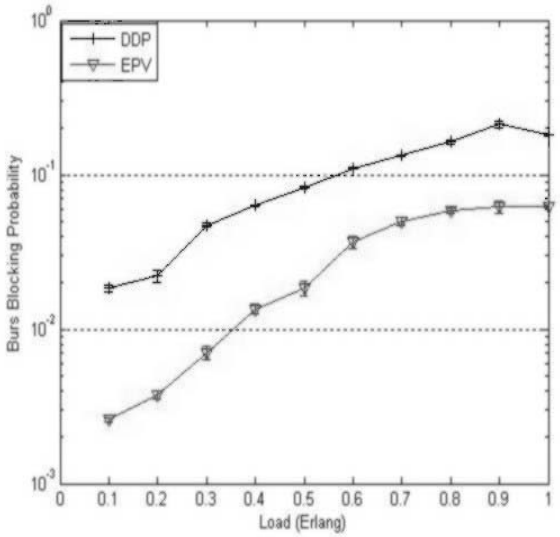

Figure 7. Burst Blocking Probability in heterogeneous CoS distribution

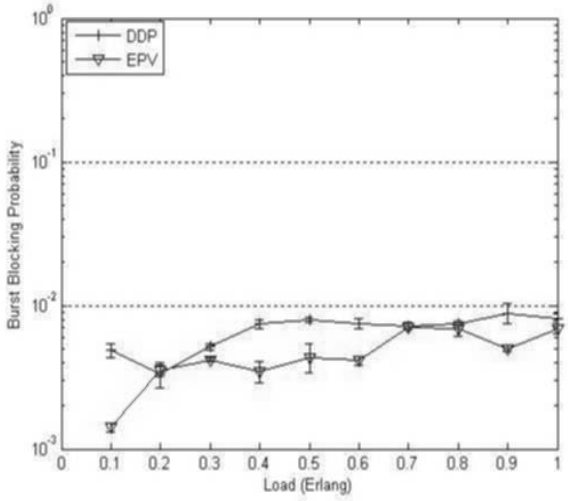

Figure 8. Burst Blocking Probability in heterogeneous CoS distribution when ϵ=0-5 ms

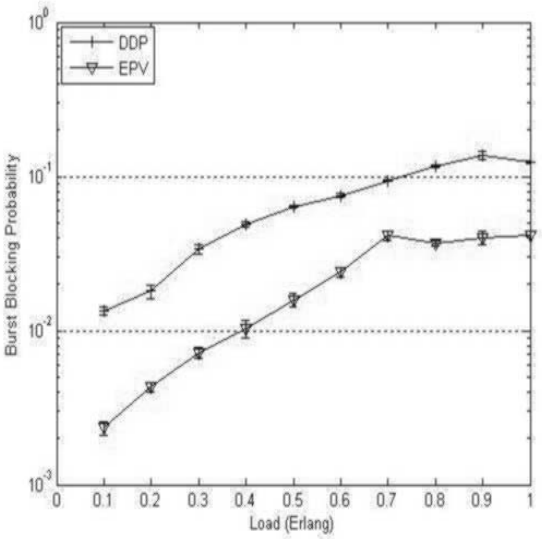

Figure 9. Burst Blocking Probability in heterogeneous CoS distribution when ∈=5-10 ms

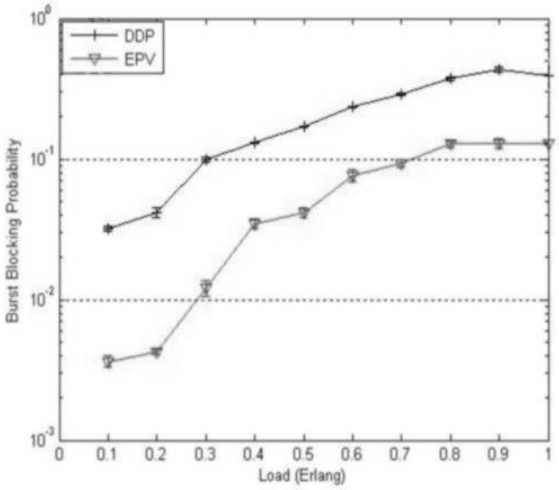

Figure 10. Burst Blocking Probability in heterogeneous CoS distribution when ∈=10-15 ms

5. Conclusion

In this paper we present a new application-aware contention resolution policy, that satisfies the QoS requirements of the users for a Grid service over OBS networks. We define a structure called Expanded Priority Vector (EPV) to prioritize the submitted jobs. The performance of this policy is compared with those of Deadline Based Contention Resolution Policy (DDP), and LAtest Arrival Drop Policy (LDP). EPV keeps the CoS value of a submitted job based on its tolerance time, residual path to the resource domain, and the job size. The aim of keeping such a vector is to implement a 3-step prioritization scheme. The first step is related to the tolerate time, namely the CoS value. The second step is related with the residual path length to the resource domain. Finally the third step is concerned with the job size. In case of a contention, the priority of a contending burst is computed based on the parameters carried in this vector. The burst to be discarded is

The results obtained show that EPV based three-step prioritization scheme enhances the performance of the Grid service and satisfies the QoS requirements. Moreover, it introduces fairness to the jobs of the same service class. Therefore, it is also observed that the overall burst blocking probability is reduced due to this behavior.

As a future work, we plan to focus on adapt this approach to resource discovery mechanisms and differentiating the intermediate routers based on their intelligence to contribute to resource discovery by active bursts.

10. References

[1] I. Foster, C. Kesselman, and S. Tuecke, "The anatomy of the Grid: Enabling scalable virtual organizations," *Int. Journal of High Performance Comput. Appl*., vol. 15, no. 3, pp. 200-222, 2001.

[2] L. Battestilli et al., "EnLIGHTened Computing: An Architecture for Co-scheduling and Co-allocating Network, Compute, and other Grid Resources for High-End Applications", *http://www.enlightenedcomputing.org*.

[3] M. de Leemheer et. al, "An OBS-based Grid Architecture", *in Proc. IEEE Global Telecommunications Conference (GLOBECOM)*, pp. 390-394, 29 Nov- 3 Dec 2004.

[4] D. Simeonidou et. al., "Dynamic Optical-Network Architectures and Technologies for Existing and emerging Grid Services", *IEEE Journal of Lightwave Technology*, vol 23, No, 10, pp.3347-3357, October 2005.

[5] R. Nejabati, et. al, "Programmable Optical Burst Switch Network: A Novel Infrastructure for Grid Services", *5th IEEE/ACM International Symposium on Cluster Computing Grid (CCGrid)*, Cardiff, UK, pp.993-999, May 2005.

[6] Kim, M-G. Kim, M. Kang, "Application-aware contention resolution scheme for Grid services over OBS networks", *in Proc. International Conference on Advanced Communication Technology (ICACT)*, pp.1425-1428, February 2006.

[7] B. Kantarci, S. Oktug, "Loss Rate Based Burst Assembly to Resolve Contention in OBS Networks", *IET Communciations*, vol 2, Issue 1, Jan 2008, pp 137-143.

[8] T. Battestilli, H. Perros, "A performance study of an Optical Burst Switched Network with Simultaneous Dynamic Link Possession", *Computer Networks*, vol 50, Issue 2, pp.219-236, February 2006.

[9] N. Barakat, E. H. Sargent, "Separating resource reservations from service requests to improve the performance of optical burst switching networks", *IEEE Journal on Selected Areas in Comunications*, vol 24, No 4, pp. 95-107, April 2006.

[10] M.D. Cano, J. Malgosa-Sanahuja, F. Cerdan, J. Garcia-Haro, "Internet Measurements and Data Study over the Regional", *in Proc. IEEE Pacific RIM Conference on Communications, Computers, and Signal Processing*, vol 2, pp.393-396, August 2001.

[11] J. Luo et al, "The impacts of burst assembly on the traffic properties of optical burst switching networks", *in Proc. IEEE International Conference on Communication Technology*, Vol 1, pp. 521-524, April 2003.