

# Performance Evaluation of an Optical Burst Switching Scheme for Grid Networks

Xenia Mountroudou

Harry G. Perros

Computer Science Department,  
North Carolina State University,  
Raleigh, NC 27695

Email: {pmountr, hp}@csc.ncsu.edu

Maged Beshai

Ottawa, Canada

Email: beshai@sympatico.ca

**Abstract**—We propose and evaluate a novel scheme, referred to as Bimodal Burst Switching (BBS) scheme, for Optical Burst Switching (OBS) star networks, which can be used in Grid networks. The main feature of this scheme is that it has zero burst loss and a large geographical coverage. Using simulation techniques we show that the proposed scheme provides high bandwidth and outperforms OBS.

## I. INTRODUCTION

The Grid is a mechanism (software/middleware) which allows the use of multiple computers and multiple data sources for general purpose computation [6]. This is already achieved using clusters or distributed computing, but the Grid goes one step further to perform these computations in multiple domains. The main characteristics of the Grid are:

- Coordination of resources that are not subject to centralized control: This is a new feature of the Grid compared to distributed computing. It integrates different administrative domains.
- Use of standard, open general purpose interfaces and protocols: This is used in order to avoid an application specific system. Open standards also ensure that worldwide research can contribute in the development of the Grid.
- High quality of service: This provides a high quality system that overcomes the weaknesses of the individual parts that compose it. Quality of service in the Grid stands for high bandwidth, low response delays, low loss of data, as well as adaptability to a highly dynamic environment.

The main applications of a Grid are in the field of e-sciences such as, high-energy physics, biology, environmental science, and engineering [7]. Most of these applications use different scientific methods, simulation and data analysis, modeling and remote sensor data collection. They typically require large amounts of bandwidth, transfer of large amounts of data reliably without losses, and user initiated end-to-end connections. Additionally Grids may be used in multimedia applications such as distribution of HDTV, where low-latency and high-capacity connections are required.

Grids impose certain requirements on the underlying networking infrastructure, such as high bandwidth availability, data granularity, user control of connectivity and high quality

of service. Optical Burst Switching appears to be a suitable technology that fulfills these requirements.

Optical fibers offer high bandwidth (50 Tbps/fiber) with relatively low latency [12]. Burst switching is a promising technique for transporting bursty traffic in an optical network. The structural unit in an OBS network is a burst: a collection of packets into a group of a size that may vary according to the characteristics of the specific network. There are several burst aggregation algorithms used in order to form a burst at an edge node [2]. The choice of a burst assembly algorithm shapes the traffic of the OBS network. This variable granularity offers the flexibility needed in a Grid network.

The main characteristic of an OBS network is the separation between the data and the control planes. Data is collected and assembled to bursts at each edge node of the network. This is performed in the electronic domain since we use buffers to collect the packets. In order to send the data optically, we first need to establish a connection along a path through the bufferless optical network. This is done by sending a control packet that includes the information needed for the transfer: the source and the destination of the burst as well as its size. The control packet is transmitted optically in-band or out-of-band and it is processed by the OBS node electronically. The optical control plane is the infrastructure and distributed intelligence that controls the establishment and maintenance of connections, and it provides dynamic control and allocation of bandwidth per user.

The trends that lead us to believe that the demanding Grid computing applications can be supported by the optical control plane are [7]:

- The low-cost, high-capacity optical connections that derive from the continuous advances in optical networks.
- The sharing of resources that is more affordable when using the optical network infrastructure. This is valuable in e-science applications: sharing of computational resources and storage is one of the main milestones in Grid networking that can be affordable using optical networks.
- Interdisciplinary research in various fields is another goal of Grid networking that can be provided by the optical control networks.

The interactions of the control plane with the Grid middleware can offer a viable solution for the Grid computing demands in combination with the high speed, refined granularity, low-cost and quality of service offered by an OBS network.

An Optical Cross Connect (OXC) can be configured to switch through an optical burst using two schemes: *immediate (on-the-fly) setup* and *delayed setup* as described in [9]. There are also two methods in order to release the resources of an OXC: *timed release* and *explicit release*. The following combinations are possible:

- 1) Immediate Setup with Timed Release
- 2) Immediate Setup with Explicit Release
- 3) Delayed Setup with Timed Release
- 4) Delayed Setup with Explicit Release

A connection can be setup using either *on-the-fly* connection setup or *confirmed* connection setup as described in [9]. These methods of configuring the OXC and the connection are used in order to transmit data in the optical plane without buffering. They provide dynamic end-to-end connection setup that is highly required on Grid applications [7]. On the other hand, setting up a path without waiting for a confirmation message from the network is one of the weaknesses of OBS networks. This is because the connection request may be refused at an intermediate OXC, which will cause the burst to be lost. It has been demonstrated that in order to keep the burst loss rate within acceptable levels, the per wavelength utilization has to be very low [10]. Burst loss is a characteristic that we would like to avoid especially in high QoS demanding e-science applications such as particle physics [12].

Several solutions to this problem have been proposed [9], such as fiber delay lines (FDL), wavelength conversion and deflection routing. A small number of FDLs could be used in order to reduce burst loss rate. Fiber delay lines require lengthy pieces of fiber, and therefore they cannot be commercialized. Wavelength conversion is a viable solution to the burst loss problem. An incoming burst on a wavelength that is currently in use at the destination output fiber, can be converted to a free wavelength. Finally deflection routing can offer an alternative path to the destination device and divert a burst that would be lost otherwise. This path may include more hops making deflection routing an ineffective method.

Sophisticated scheduling algorithms have been proposed in order to reduce burst loss. In the case of the delayed setup/timed release scheme, it is possible to do *void filling*. Assume that a burst is due to arrive at  $t_1$ . Now, let us assume that in the meantime the OXC receives a setup message for a burst at time  $t_2$ , where  $t_2 < t_1$ . The burst will be accepted if there is sufficient time between  $t_2$  and  $t_1$  for the transmission of the second burst plus the necessary OXC configurations. This mechanism is presented by Xiong et al. in [14] Another scheduling method called *Horizon*, was described by Turner [13]. The scheduler can schedule the reservation of the resources at the shortest time (horizon) according to the offset and the availability. An analogous technique is used for releasing the resources; in this case it is defined by the size

of the burst. Horizon can be easily implemented, as opposed to void filling which is CPU intensive.

In our study, we evaluate the performance of an innovative OBS-based scheduling architecture that provides zero burst loss and low overheads. This architecture referred to as Bimodal Burst Switching (BBS) was reported in [3]. It uses the delayed setup timed-release scheme in conjunction with two modes of operation. The first mode (Mode 0) applies to an edge node that is close to its ingress OXC, the second mode (Mode 1) applies to a distant edge node. BBS can cover a large geographical area (1,000 km radius) and it is implementable in hardware as described in [3]. The rapid growth of scientific data in combination with its wide geographic distribution makes the characteristics of this architecture quite attractive for a Grid network. In this paper we analyze the performance of this new scheme and we compare it to the OBS scheme as described in [1].

The paper is organized as follows. In Section II, we review the proposed architecture, in Section III, we present the simulation results, and finally, in Section IV, we provide some concluding remarks.

## II. THE BIMODAL BURST SWITCHING ARCHITECTURE

The architecture under study is a star network consisting of  $N$  edge nodes and an OXC, hereafter referred to as the core node. Each edge node is linked to the core node via an upstream and a downstream fiber, each carrying  $W$  wavelengths. The core node has a number of parallel switching planes equal to  $W$ , the number of wavelengths in a WDM link. In this scheme we assume full wavelength conversion. The number of edge nodes that the core node can support is equal to the number of dual ports per switch plane. More than one core node can be used in order to provide larger geographic coverage. The core nodes of a star network are independent and, hence, it suffices to consider only one core node to illustrate the salient features of the proposed BBS technique.

Integral part of the proposed architecture is time-locking. As described in [4], this is a technique for time coordination and it uses time-counters to enable time alignment of signals received at connecting core nodes. A node is considered time-locked to another node, if a signal transmitted at an instant of time indicated by a time counter at the first node, arrives at the second node at the same instant of time as indicated by an identical time counter at the second node.

Each edge node receives data in the electronic domain and transmits it out to the other edge nodes via the core node. The incoming traffic at each edge node is queued to  $N - 1$  different output queues (buffers), one per destination. We use a burst aggregation algorithm that combines a pre-set timer and minimum/maximum burst size parameters in order to form a burst. When the pre-set timer expires we check the destination queues of every edge node. A data burst is formed using the packets of a destination queue, and its size is limited by a minimum and a maximum burst size. This aggregation technique prevents from forming very small bursts that would

be undesirable in a Grid network, where we want to avoid too many control packets contending for resources. It also avoids very large data bursts, that would occupy the network resources for long periods. This would block the transmission of other bursts, leading to burst loss [9].

This architecture introduces an innovative scheduler that performs the flow regulation of the traffic that arrives at each edge node. The scheduler is embedded in the controller of the core node. The core node does not use buffers on either inputs or outputs. The main structures that are used by the controller in order to make a scheduling decision is the *Calendar* and the *M-element array*. The calendar keeps track of the time when the uplink wavelengths of each edge node are going to be free. It consists of  $K$  elements. Preferably  $K$  is equal to  $N * W$ , where  $W$  is the number of wavelengths per fiber and  $N$  the number of edge nodes. We may also use a two-dimensional array with  $N$  rows and  $W$  columns to store the calendar structure. If an element that belongs to the  $i^{th}$  row and  $w^{th}$  column of the calendar is equal to  $j$ , then this means that the  $w^{th}$  wavelength of the edge node  $i$  is free at time slot  $j$ . The M-element array keeps track of the availability of the output edge nodes. It consists of  $M$  elements, where  $M$  is  $N * W$ . We use a two-dimensional array with  $N$  rows and  $W$  columns for the M-element array elements, that are stored in the same way as the calendar elements. The scheduler is described in detail below.

#### A. The Bimodal Burst Switching Modes

The scheduling algorithm consists of two modes, Mode 0 and Mode 1, based on the proximity of the edge node to the core node. The proximity is determined by measuring the round trip propagation delay between the edge node and the core node. Mode 0 is used for edge nodes that are at a small distance from the core node and consequently have a small round trip propagation delay. On the other hand, Mode 1 is more suitable for distant edge nodes that have a large propagation delay. As will be explained below the main difference between these two modes is that in the first mode the flow rate regulation is provided to waiting bursts whereas in the second mode it is provided to anticipated bursts.

1) *Mode 0 Burst Switching*: This mode of operation implements the OBS scheme, whereby an edge node sends requests to transmit bursts to its ingress OXC. As will be seen below, these requests are sent by the edge node at fixed intervals. The operation of Mode 0 is as follows:

- **Transmission:** Edge node  $i$  receives packets which it then buffers to the appropriate destination queues. Every  $T$   $\mu$ sec the edge node checks the input queues and forms bursts subject to a minimum and a maximum burst size. For each burst it issues a burst request that is stored at the burst request queue. Each request consists of a number of fields such as: source, destination, size and an ID number. Every  $T$   $\mu$ sec edge node  $i$  sends all the burst requests it has stored until that moment in a single control packet to the core node. This procedure needs time equal to one way propagation delay to be completed. The one-way

propagation delay is small for short distance edge nodes and this is why the OBS scheme is chosen. Additionally, it is more efficient to send all the requests gathered during a period of  $T$   $\mu$ secs in one control packet rather than send one request per burst.

- **Scheduling:** Once the control packets reach the core node, the controller that implements the scheduler decides when the burst will be transmitted using the *shortest horizon* scheduling policy. This decision is formed using the calendar and the M-element array. The scheduler scans the calendar to find the first uplink wavelength of any edge node that is free, then calculates the horizon for the specific edge node requests. Assume that the first uplink wavelength that is free is wavelength  $w_{i1}$ , that is wavelength 1 of edge node  $i$ . We compute the horizon for each burst request of the edge node  $i$  as the difference of the time slot the downlink wavelengths of the requested destination are free, minus the time slot that the uplink wavelength 1 of edge node  $i$  is free. Let us assume that there is one request to destination  $j$ . Then we would have to calculate all the horizons for all the downlink wavelengths that correspond to  $j$  and pick the minimum of them. Let us also assume that the same edge node  $i$  has a request for edge node  $k$ . Then again we have to calculate all the horizons for the downlink wavelengths to  $k$  and pick the minimum. Finally, we would have to compare these two minimum horizons and decide which of the two requests would be served. According to the proposed scheduling policy it is preferred to schedule the minimum value that is negative, because this means that the destination is available earlier than the source and the source can start transmitting immediately. For example if source edge node  $i$  has a free uplink wavelength at time slot 10 and requests to transmit to destination edge nodes  $j$  and  $k$  that have free downlink wavelengths at times 5 and 15, the horizons are -5 and 5. We prefer to schedule the request destined to  $j$  since the source can start transmitting to it immediately. After we serve one request we update the calendar and the M-element array. We then scan the calendar for the next available wavelength and repeat the horizon scheduling, until we schedule all the requests that were sent at this time. Scheduling all the requests every fixed time period satisfies another Grid requirement: high bandwidth combined with low latency. Additionally, this scheduling scheme provides the required determinism for high demand grid applications: it guarantees that no burst is lost since it provides the precise scheduling time that the burst can be sent. When all the requests are scheduled, the core node sends permits to the edge nodes containing information as to when they can transmit their bursts.
- **Reception:** Each destination edge node receives bursts from the core node which are buffered electronically and then delivered to its users through other interfaces.

2) *Mode 1 Burst Switching*: In Mode 1 burst switching, data is still transmitted in bursts, but the initial phase in OBS where an edge node sends a request to its ingress OXC has been eliminated. Mode 1 scheduling is preferable when the propagation delay is large. Unlike Mode 0, a Mode 1 edge node does not issue burst requests. Rather, the edge node requests and is allocated a fixed bandwidth for each destination. This bandwidth is calculated based on the traffic between the edge node and each destination edge node, and it is made available to the edge node in bursts. These bursts are fixed in time and they repeat periodically.

Let  $t_{ij}$  be the transmission time allocated to the traffic from  $i$  to  $j$ , and let this be repeated every  $T$  units of time. Then the bandwidth allocated to edge node  $i$  for transmitting traffic to edge node  $j$  is  $(t_{ij}/T) * V$  where  $V$  is the transmission speed. The edge node communicates the values  $t_{ij}$ ,  $j = 1, \dots, N$ ,  $j \neq i$  and  $T$  to the controller. The controller issues automatically a burst request of duration  $t_{ij}$  every  $T$  units of time for each destination. These burst requests are then scheduled following the same procedure as in Mode 0 operation. The scheduler then issues permits which are sent to the edge node. For simplicity we have assumed that  $T$  is equal to the polling period for Mode 0.

The Mode 1 operation is summarized as follows:

- **Transmission:** The edge node may transmit the data it has gathered up to this moment based on the permit information. In this case we do not have a minimum or maximum burst size that define the size of a burst. The burst size is defined by the transmission time  $t_{ij}$ . When a Mode 1 edge node  $i$  receives a permit it will transmit data for the duration  $t_{ij}$ . Assume, for instance, that the data it has requires  $112 \mu\text{sec}$  to transmit out. Assume also that  $t_{ij} = 100 \mu\text{sec}$ . Then in this case, it will not be able to transmit all the data, and  $12 \mu\text{sec}$  worth of data will remain in its buffer. On the other hand, if it has  $80 \mu\text{sec}$  worth of data, then it will transmit all its data and the remaining  $20 \mu\text{sec}$  of the  $t_{ij}$  period will go unused.
- **Scheduling:** The controller creates a burst request of duration  $t_{ij}$  for every destination  $j \neq i$ , and for every Mode 1 edge node  $i$ , every  $T$  units of time. These requests are then placed in the scheduler's queue, and they are scheduled in the same manner as Mode 0 burst requests. Transmission permits are then sent to the Mode 1 edge nodes.
- **Reception:** The destination edge node  $j$  receives bursts which are buffered electronically and then delivered to its users.

The main difference between the two scheduling modes is that in Mode 0 we schedule already existent bursts whereas in Mode 1 we issue permits for anticipated bursts. As a result, it is possible that a Mode 1 node may receive permits during periods of time that it has no data to transmit.

We also note that the bandwidth allocated for each Mode 1 edge node by the controller can be renegotiated using specially designed messages. Such renegotiation will take place when the traffic arriving at an edge node changes significantly. This

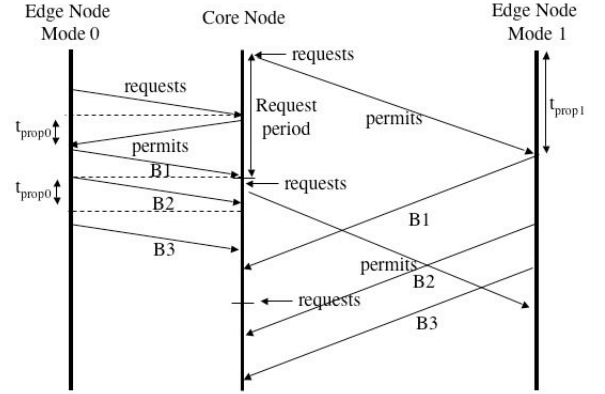


Fig. 1. Simulation timing diagram

is expected to take place less frequently compared to the time scales of the burst transmission operation, and it has not been considered in this study. Renegotiation of the bandwidth allocated is a useful feature since Grid applications require near-real-time reaction to events and changing environments.

3) *The Bimodal Burst Switching Scheduler*: The operation of the bimodal scheduler is depicted in Figure 1. In the case of a nearby edge node (Mode 0), the edge node sends the requests it has accumulated up to this moment every fixed time period, say every  $256 \mu\text{sec}$ . The core receives the requests, schedules them according to the shortest horizon scheme and then sends permits to the edge nodes. Finally, the edge nodes transmit their bursts according to the permits received from the core. The fixed period used to send requests is short and as a result it provides a continuous supply of permits to the edge nodes.

In the case of a distant node (Mode 1) the core node creates burst requests periodically which are then scheduled according to shortest horizon. The main difference in this scheme is that there are no requests from the edge nodes to the core node. This provides a more efficient scheme since the one way propagation delay is large. When a Mode 1 edge node receives a permit, it transmits data for a fixed period of time. In this study we have assumed out-of-band signalling. The signalling messages can also be implemented in-band, but this was not considered in this study.

### III. SIMULATION RESULTS

In this section, we first describe the simulation model that we developed in order to evaluate the performance of the Bimodal Burst Switching architecture. We then provide numerical results and discuss its performance.

$N$  edge nodes and one core node are simulated. We assume that edge nodes 1 to  $N/2$  are within a small distance  $d$  from the core node, where  $10 \text{ km} < d < 100 \text{ km}$ , which means that they are served using Mode 0 scheduling. The remaining edge

nodes  $N/2+1$  to  $N$  are more than 100 km away, which means that the core node serves them using the Mode 1 scheduling mechanism. The minimum burst size and maximum burst size were fixed to 16 kB and 112 kB respectively. Furthermore, the period  $T$  was set to 256  $\mu\text{sec}$ . The one way propagation delay between an edge node and the core node for Mode 0 edge nodes was set to 500  $\mu\text{sec}$ , which means they are at a 100 km distance from the core node, and for Mode 1 edge nodes to 5,000  $\mu\text{sec}$ , which means they are at 1,000 km from the core node.

Each edge node has a 10 MB electronic buffer to store the packets that arrive from external sources. The arrival process is an Interrupted Poisson Process (IPP) as described in [5]. This IPP arrival process is an ON/OFF process, where both the ON and OFF periods are exponentially distributed. Packets arrive back to back during the ON period at a rate of 10 Gbps. Packets do not arrive during the OFF period. The packet length is assumed to be exponentially distributed with an average of 500 bytes. The last packet of the ON period may be truncated so that its last bit arrives at the end of the ON period. We used the squared coefficient of variation,  $c^2$ , of the packet interarrival time to characterize the burstiness of the packet arrival process. This coefficient is defined as the variance of the packet inter-arrival time divided by the squared mean packet inter-arrival time. Assuming that the distribution of the ON period is exponential with average  $1/\mu_1$  and the distribution of the OFF period is exponential with average  $1/\mu_2$  we have:

$$c_{IPP}^2 = 1 + \frac{2\lambda\mu_1}{(\mu_1 + \mu_2)^2}$$

where  $\lambda$  is the arrival rate of a packet during the ON period and  $\frac{1}{\lambda} = \frac{(500\text{Bytes})}{(10\text{Gbps})} = 0.4\mu\text{sec}$ . Finally to characterize completely the arrival process we use the *average arrival rate*, given by:

$$\text{average arrival rate} = \frac{(10\text{Gbps})\mu_2}{\mu_1 + \mu_2}$$

Given the  $c^2$  and the average arrival rate we calculate the quantities  $\mu_1$  and  $\mu_2$ . In our simulation experiments  $c^2$  was set to 5 and 20, and the arrival rate was varied from 6 Gbps to 100 Gbps. Packets arriving at an edge node were assigned to a destination using the uniform distribution. Results for non-uniform destinations (hot-spots) were also obtained.

The simulation outputs consist of the mean delay per packet per edge node for nearby and distant edge nodes, the mean overall delay per packet for all nodes and the percentage of utilization of an uplink or a downlink wavelength. In all the figures provided, the results are plotted with 95% confidence intervals estimated by the method of the batch means [8]. The number of batches is set to 30 and each batch consists of at least 10,000 bursts/edge node. The confidence intervals are very narrow and as a result are barely visible in the figures.

We compared the Bimodal Burst Switching (BBS) scheme against the case where all  $N$  edge nodes operate under the Mode 0 scheme, indicated in the graphs as "Mode 0", and also against the case where all  $N$  edge nodes operate under

the Mode 1 scheme, indicated in the graphs as "Mode 1". We recall that in the BBS scheme edge node 1 to  $N/2$  operate under Mode 0 and edge nodes  $(N/2+1)$  to  $N$  under Mode 1.

As described above Mode 0 implements the OBS scheme as analyzed in [1]. The calculation of the intervals  $t_{ij}$  for Mode 1 was based on the average arrival rate. These intervals can also be calculated using the Equivalent Bandwidth (EB) as we present below. Full wavelength conversion was assumed.

Figures 2, 3 and 4 give the average delay per packet versus the number of wavelengths for 10 edge nodes when  $c^2 = 5$  and 20. The average arrival rate at every edge node is 6 Gbps. The delay of a packet is the time elapsed from the moment it fully arrives at an edge node to the moment it is delivered to a destination edge node. That is, it consists of the queuing delay at edge node plus the propagation delay from the transmitting edge node to the destination edge node. Edges nodes 1 to 5 are at short distance from the core node (i.e. they have 500  $\mu\text{sec}$  one-way propagation delay) and edge nodes 6 to 10 are far away (i.e. they have a 5,000  $\mu\text{sec}$  one-way propagation delay). Packets arriving at each edge node were assigned to a destination node using the uniform distribution, and the intervals  $t_{ij}$  for Mode 1 were calculated based on the average arrival rate.

Figure 2 gives the average delay per packet calculated only for nodes 1 to 5. In Figure 2(a)  $c^2 = 5$  and in Figure 2(b)  $c^2 = 20$ . We observe that the BBS scheme guarantees the lowest average delay when the number of wavelengths per edge node is less than 15. This is useful information for the implementation of a low cost, high bandwidth Grid network, since increasing the number of wavelengths increases the cost of implementation. As the number of wavelengths increases, the average packet delay for all three schemes becomes almost identical. An interesting observation is that the average delay per packet for the BBS and the Mode 1 are very close. This leads us to the conclusion that differentiating our scheduling technique between distant and closeby edge nodes does not offer a large improvement on the average delay per packet. We also observe that the mean delay for the closeby edge nodes is almost the same as we increase the burstiness of the input traffic from  $c^2 = 5$  to  $c^2 = 20$ .

Figures 3(a) and (b) give the average delay per packet for edge nodes 6 to 10 for  $c^2 = 5$  and 20 respectively. BBS and Mode 1 exhibit almost the same average delay per packet. Though this is not surprising, since they use the same scheduling technique, we observe that BBS does not improve the average delay over Mode 1 by differentiating the scheduling of requests between the two modes of operation. In the case of Mode 0 the average delay is much higher, almost 20 times than in BBS, especially when the number of wavelengths per edge node is low. This indicates that the OBS scheme is not effective for distant stations. Additionally we observe that the mean delay for all three schemes is very close when we increase  $c^2$  from 5 to 20. Only in Mode 1 we observe an increase of 107  $\mu\text{sec}$ . This is justified since this scheme allocates for every edge node the same bandwidth for  $c^2 = 5$  and  $c^2 = 20$ , but the burstiness of the input traffic

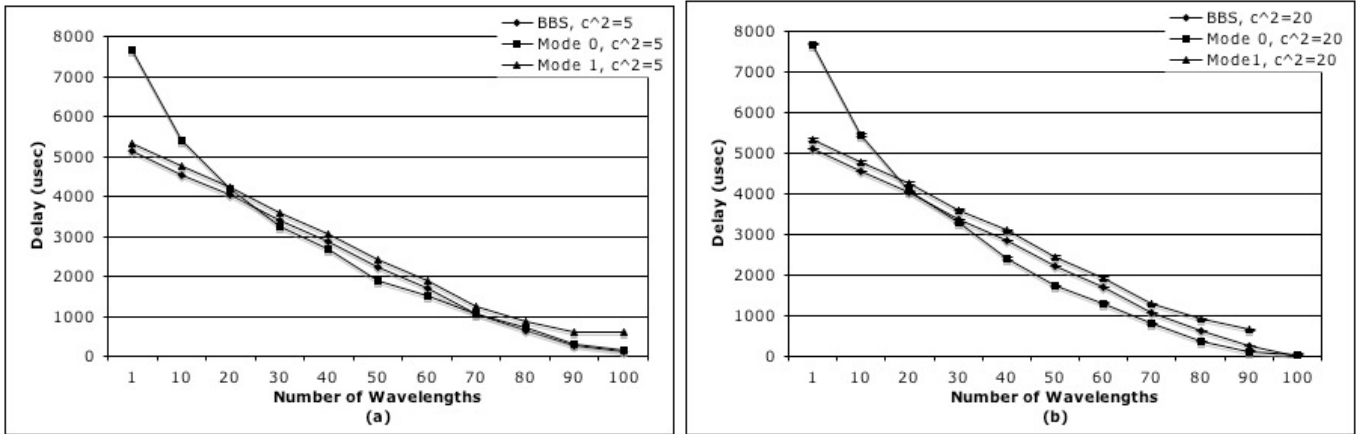


Fig. 2. (a) Mean packet delay for edge nodes 1 to 5 vs. number of wavelengths for  $c^2=5$ , (b) Mean packet delay for edge nodes 1 to 50 vs. number of wavelengths for  $c^2=20$

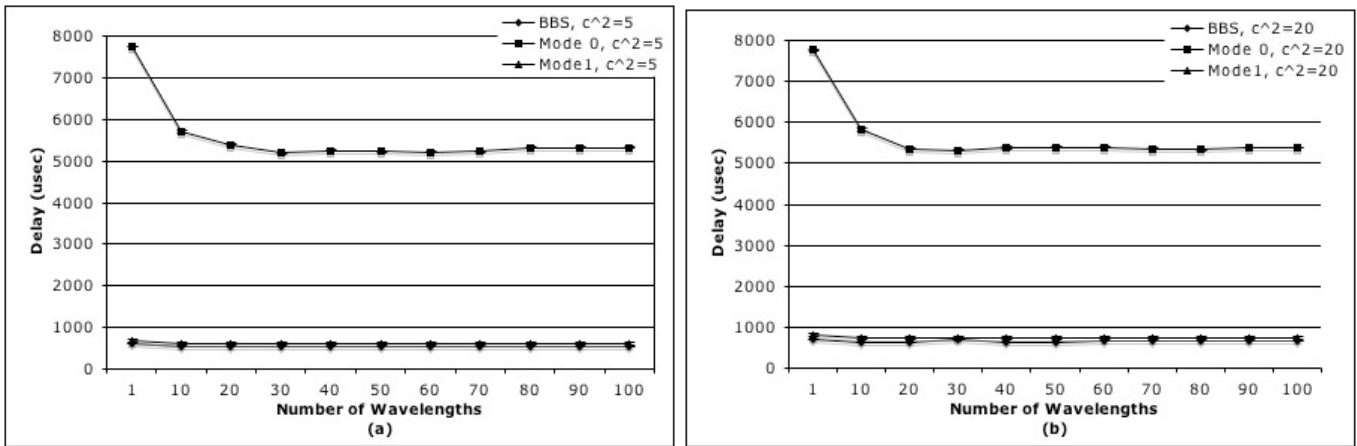


Fig. 3. (a) Mean packet delay for edge nodes 6 to 10 vs. number of wavelengths for  $c^2=5$ , (b) Mean packet delay for edge nodes 6 to 10 vs. number of wavelengths for  $c^2=20$

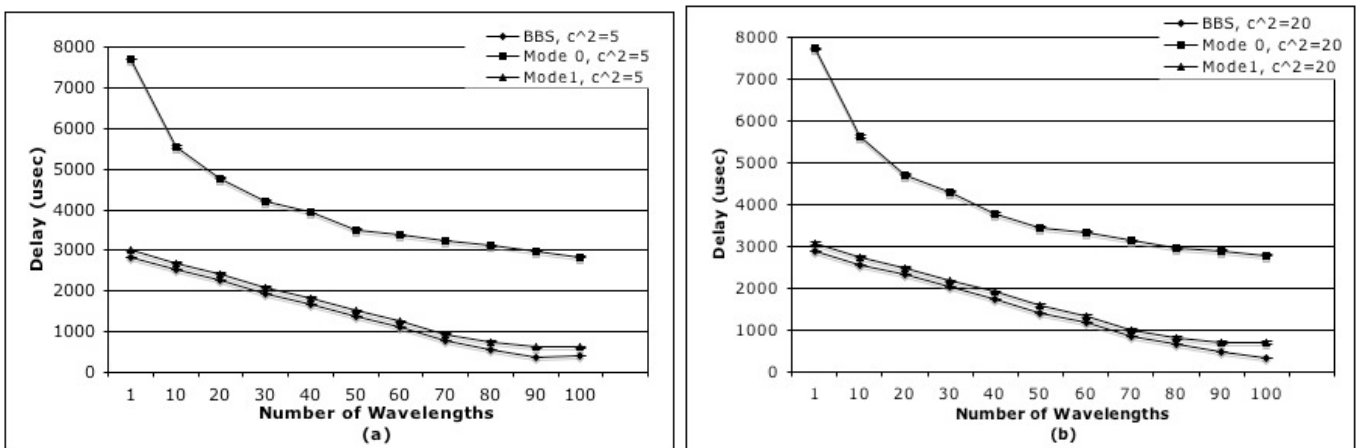


Fig. 4. (a) Mean packet delay for all 10 edge nodes vs. number of wavelengths for  $c^2=5$ , (b) Mean packet delay for all 10 edge nodes vs. number of wavelengths for  $c^2=20$

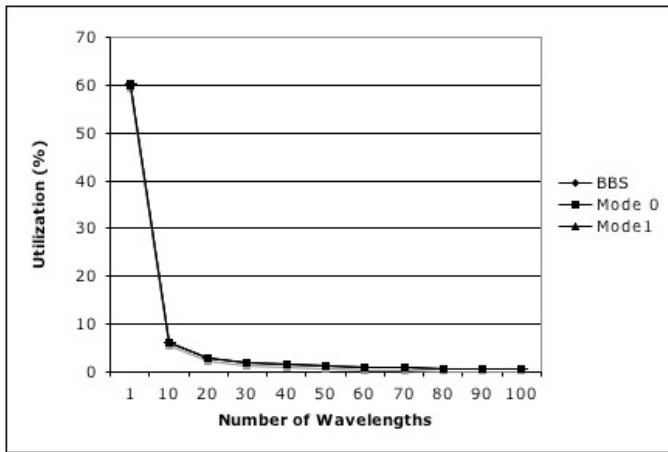


Fig. 5. Mean utilization for all 10 edge nodes vs. number of wavelengths

per edge node increases. Then the edge node does not have sufficient bandwidth to send all the packets that accumulate in its destination queues, leading to longer delays.

Figure 4(a) shows the total average delay per packet for all 10 edge nodes for  $c^2=5$ . As the number of wavelengths increases, we observe an almost linear decrease in the packet delay for the BBS and Mode 1 schemes. The difference between Mode 0 and BBS is evident: Mode 0 overall average delay per packet is much higher. Finally Mode 1 and BBS scheduling have almost the same average delay per packet. This leads us to the conclusion that Mode 1 is a good scheduling technique even though it does not differentiate between distant and nearby edge nodes. Figure 4(b) gives the average delay per packet when the input traffic is burstier, i.e.  $c^2 = 20$ . Again we observe that there is no significant difference for the BBS and Mode 0 schemes when burstiness increases.

Figure 5 shows the percentage of utilization of an uplink/downlink wavelength. The uplink and downlink wavelength utilization is the same. That is because we have assumed the same arrival process to each edge device and uniformly chosen destination nodes. When we use only one wavelength in our model, wavelength utilization approaches 60%. All three schemes have the same utilization. As mentioned in Section II Mode 1 scheduling scheme is used to schedule bursts that are not yet formed at the edge node. If the wavelength utilization is high, this means that there is always a burst formed for each destination in every edge node that is scheduled using this scheme. Then the bandwidth that is allocated periodically is not wasted. On the other hand, the utilization per wavelength decreases since the number of wavelengths increases and there are more alternative paths for a data burst. This means that lower per wavelength utilization does not affect Mode 1 and BBS schemes if the input traffic and the overall utilization remains the same (60% for all wavelengths in one fiber link).

Figure 6 shows how the average delay per packet is affected when we vary the average arrival rate of the input traffic and set the rate of packet arrivals to 100 Gbps for all

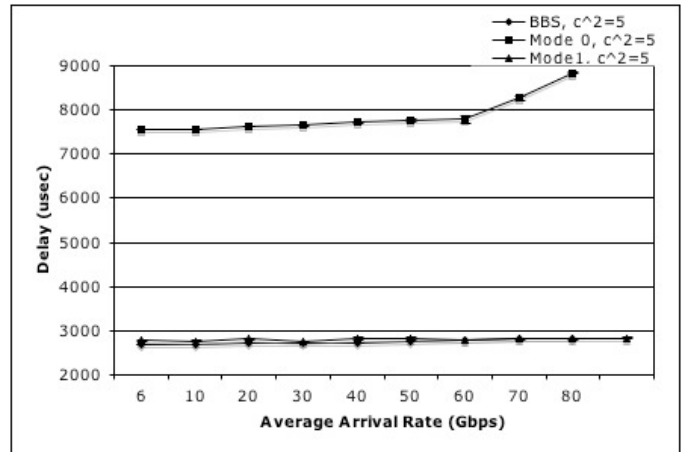


Fig. 6. Mean packet delay for all 10 edge nodes vs. average arrival rate

three scheduling schemes, with all other parameters remaining the same as above. There is one uplink and one downlink wavelength for every fiber link. When the average arrival rate is  $>80$  Gbps we get very high delays for Mode 0 and BBS, whereas Mode 1 gives very high delay when it is  $>90$  Gbps. These delays are not drawn in this Figure. The BBS and Mode 1 schemes scale well when the average arrival rate increases. Mode 0 on the other hand has a high increase in the mean delay when the average arrival rate is  $>60$  Gbps. This proves that BBS and Mode 1 are suitable for the high bandwidth demands of Grid networks.

In Figure 7(a) we plot the average delay per packet for all edge nodes when we vary the number of edge nodes for  $c^2 = 5$ . We assume that  $W = 1$ , i.e. one uplink and one downlink wavelength. The BBS scheme scales well as we increase the number of edge nodes, whereas Mode 0 has large delays. We also observe that Mode 1 scales very well, remaining almost constant. The scalability of BBS and Mode 1 is a good feature for Grid networks where high geographic coverage and sharing of multiple resources is required. The low delays of Mode 1 and BBS is contrasted to the high utilization percentage, that is about 60% when only one wavelength is used. When traffic burstiness increases, the average delay per packet remains almost the same for all three schemes, as plotted in Figure 7(b) for  $c^2 = 20$ . In this case BBS and Mode 1 scale well with additional edge nodes, and Mode 0 appears still to have high average delays per packet when the edge nodes attached to the core node increase. This is an interesting observation since the traffic and the resources used in Grid networks are highly dynamic parameters. An architecture that has low delays, no loss and can scale well with bursty traffic is required in such cases.

Figure 8 shows the average delay per packet for all three schemes when non-uniform destinations are used. We direct 10% of the input traffic to edge node 0 which represents a hot spot. The rest of the input traffic (90%) is uniformly distributed to all destinations. The average arrival rate is 6 Gbps and  $c^2 = 5$ . The bandwidth allocated to each Mode 1 edge

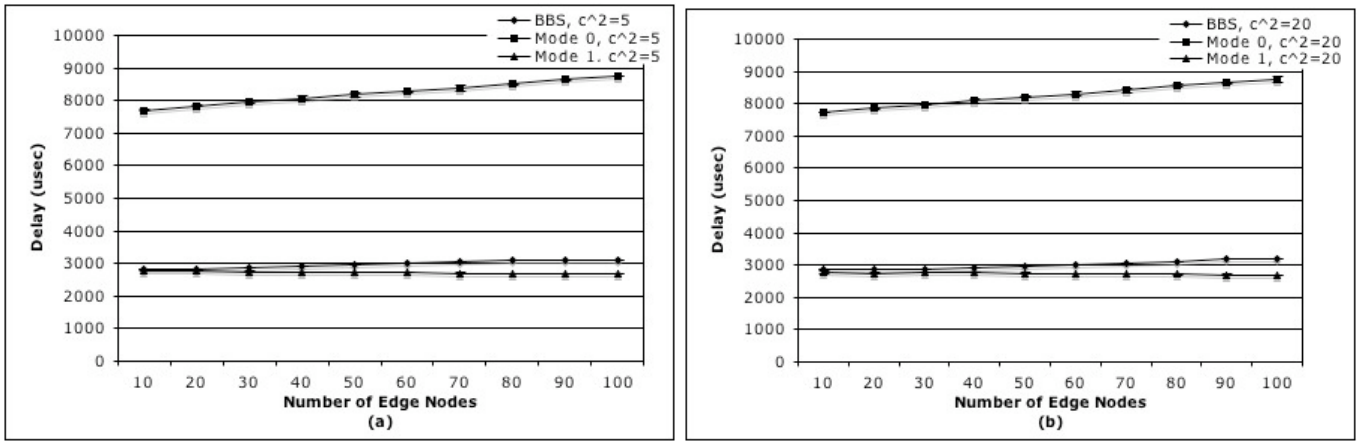


Fig. 7. (a) Mean packet delay for all edge nodes vs. number of edge nodes for  $c^2=5$ , (b) Mean packet delay for all edge nodes vs. number of edge nodes for  $c^2=20$

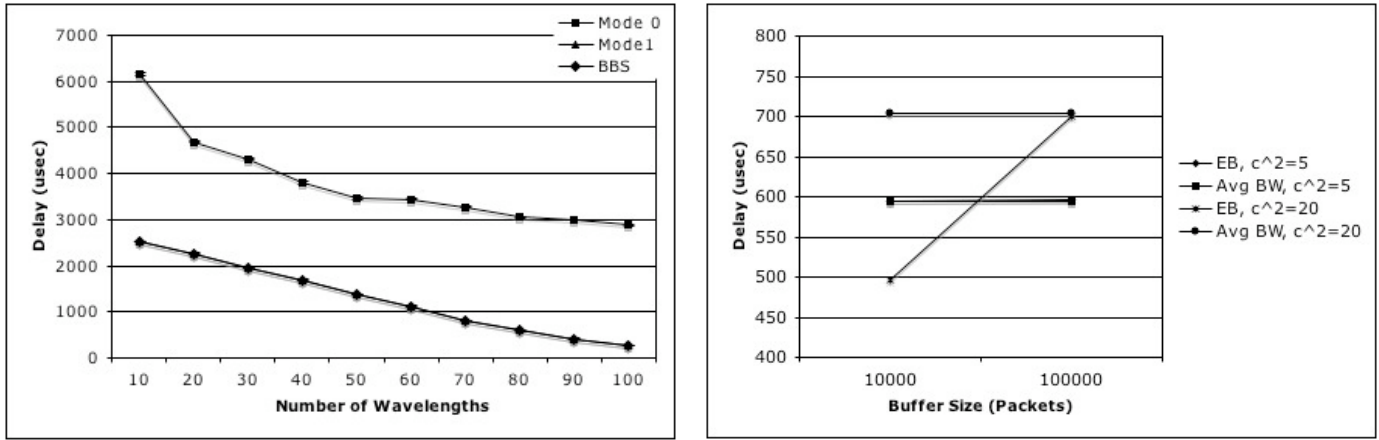


Fig. 8. Mean packet delay for all 10 edge nodes vs. number of wavelengths (non-uniform destinations)

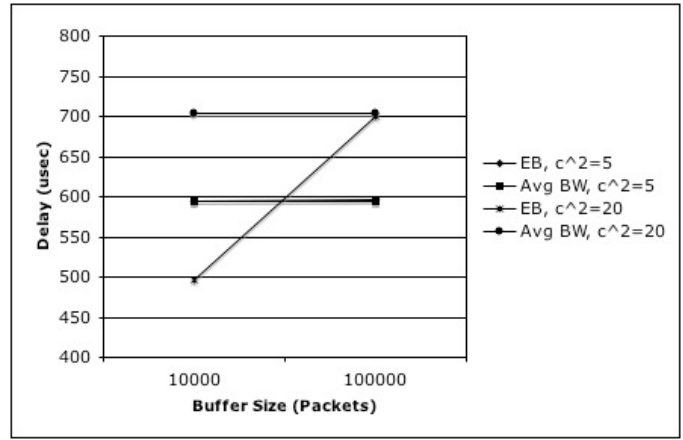


Fig. 9. Mean packet delay vs. Buffer Size

device was recalculated to account for the hot spot. This means that we allocate 10% more to  $t_{i0}$ , which is the transmission time from every edge node  $i$  to edge node 0. Simulation results with one wavelength are not shown in this figure because all three schemes result to very high delays. This proves that these scheduling schemes are inefficient for hot-spot implementation with only one wavelength. In Figure 8 the mean delay for the BBS and the Mode 1 scheme is identical. Comparing the BBS and Mode 1 schemes' average delay per packet for uniform destinations as shown in Figure 4 and non-uniform (Figure 8) we do not observe large differences when we use more than 10 wavelengths. In this case BBS and Mode 1 are efficient schemes for a non-uniform destination. This is a useful characteristic for Grid networks that are characterized by a highly dynamic structure where traffic to some very powerful resources (mainframes) may be much higher than traffic to other nodes. Mode 0 exhibits a little higher average delays than in the uniform destinations model. All three schemes give very high delays when the downlink utilization is 100%, which is the case for the hot spot when

there is only one uplink/downlink wavelength.

In Figure 9 we show the average delay per packet for the edge nodes that are served with Mode 1 in the BBS scheme (i.e. edge nodes 6 to 10) when we use average bandwidth allocation and equivalent bandwidth allocation [11]. We vary the input traffic by changing the burstiness, defined by  $c^2$ . We show the equivalent bandwidth allocation only for a 10,000 and a 100,000 packet buffer, because for lower buffer sizes our model has high packet loss, and for buffer size  $\geq 100,000$  packets the equivalent bandwidth allocation is the same as the average bandwidth allocation. When we use the equivalent bandwidth allocation scheme the average delay per packet is not affected considerably. We observe that for small burstiness ( $c^2 = 5$ ) there is no difference in the average and the equivalent bandwidth allocation even for smaller than 100,000 packets buffer sizes. As the burstiness increases, the average delay per packet increases for the edge nodes that are served using the Mode 1 scheme. In this case, allocating bandwidth using the equivalent bandwidth scheme causes the mean delay per packet to decrease, when the buffer size is less than 100,000 packets. The edge nodes that are served using the Mode 0

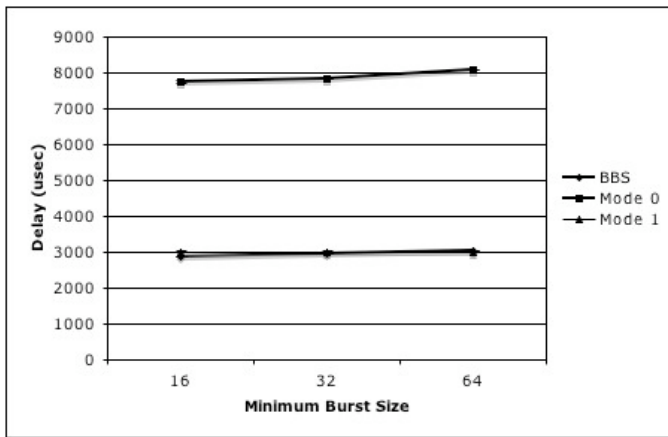


Fig. 10. Mean packet delay vs. Minimum Burst Size

scheme are not affected, since the burst size in this case is controlled by the minimum/maximum burst size.

Finally, in Figure 10 we varied the minimum burst size, while keeping the average arrival rate at 6 Gbps. The remaining assumptions are the same as discussed above. We observe that the average delay per packet is slightly affected for the BBS and the Mode 0 scheduling. For minimum burst sizes greater than 64 KB (i.e. 128 KB) very high mean delays were observed (not shown in this figure). This leads us to the conclusion that a good value for the minimum burst size is 64 KB. Mode 1 is not affected by the variation of the minimum burst size since the average bandwidth allocation remains the same.

#### IV. CONCLUSIONS

In this paper, we presented a novel bimodal scheduling architecture for OBS networks that could be used to support Grid networks. This architecture consists of two scheduling schemes: Mode 0 and Mode 1. We evaluated this architecture by comparing it to OBS scheduling and Mode 1 scheduling. We concluded that this novel architecture provides low average packet delays compared to OBS scheduling but does not offer any advantage in performance from Mode 1 by differentiating service into different modes.

The BBS and the Mode 1 scheme appear to be slightly affected by traffic burstiness. They also give low average delays per packet when the packet destinations are non-uniform and the number of wavelengths is greater than or equal to the number of edge nodes attached to the core node. Finally, the use of equivalent bandwidth allocation gives better results when traffic burstiness is high. In Mode 1, the transmission window can be adjusted to follow fluctuations in the arrival rate. This is an issue that will be explored at a later time.

Grid networks require low delays as well as zero burst losses and large geographic coverage. All these are offered by the BBS architecture. Another advantage of this is that it provides good granularity, which is an important issue in e-science applications. Finally this architecture is highly scalable

and this can provide good sharing of resources required in Grid computing.

#### REFERENCES

- [1] I. Baldine, G. Rouskas, H. Perros, and D. Stevenson. Jumpstart: A just-in-time signaling architecture for wdm burst-switched networks. *IEEE Magazine on Communications*, page 82, February 2002.
- [2] Tzvetelina Battestilli and Harry Perros. An introduction to optical burst switching. *IEEE Comm. Optical Magazine*, 41:S10–S15, 2003.
- [3] Maged Beshai. Temporal-spatial burst switching. U.S. Patent Application, Publication number 2005-0078666, April 14, 2005.
- [4] Maged Beshai et al. Burst switching in a high capacity network. U.S. Patent, 6,907,002, June 14, 2005.
- [5] W. Fischer and K. Meier-Hellstern. The Markov-Modulated Poisson Process (MMPP) cookbook. *Performance Evaluation*, 18:149–171, 1992.
- [6] Ian Foster. What is the Grid? A three point checklist. Available at: <http://www-fp.mcs.anl.gov/foster/Articles/WhatIsTheGrid.pdf>, July 2002.
- [7] Gigi Karmous-Edwards. Global e-science collaboration. *Computing in Science & Engineering [see also IEEE Computational Science and Engineering]*, 7(2):67–74, March-April 2005.
- [8] Harry G. Perros. Computer simulation techniques: the definitive introduction. Available at: <http://www.csc.ncsu.edu/faculty/perros/books.html>, 2003.
- [9] Harry G. Perros. *Connection-oriented networks: SONET/SDH, ATM, MPLS, Optical Networks*. Wiley, 2005.
- [10] Vishwas Puttasubba and H. Perros. An approximate queueing model for limited-range wavelength conversion in an OBS switch. In *Networking 2005*, pages 697–709. Springer-Verlag, 2005.
- [11] Mischa Schwartz. *Broadband Integrated Networks*. Prentice Hall, 1996.
- [12] D. Simeonidou, R. Nejabati, B. St. Arnaud, M. Beck, P. Clarke, D. B. Hoang, D. Hutchison, G. Karmous-Edwards, T. Lavian, J. Leigh, J. Mambretti, V. Sander, J. Strand, and F. Travostino. Optical Network Infrastructure for Grid. Informational GHPN, 2003.
- [13] J. Turner. Terabit burst switching. *Journal of High Speed Networks*, 8(1):3–16, 1999.
- [14] Yijun Xiong, M.M. Vandenhouete, , and H. Cankaya. Control architecture in optical burst switched WDM networks. *IEEE Journal on Selected Areas in Communications*, 18(10):1838–1851, October 2000.