Hierarchical Time-Sliced Optical Burst Switching

Vijay Sivaraman and Arun Vishwanath
School of Electrical Engineering and Telecommunications
University of New South Wales
Sydney, NSW 2052, Australia
Emails: {vijay@unsw.edu.au, arunv@ee.unsw.edu.au}

Abstract*—To overcome the need for large buffers to store contending bursts in optical burst switched (OBS) networks, a recent variant called time-sliced OBS (TSOBS) suggested that bursts be sliced and spread across multiple frames of fixed-length time-slots. Since TSOBS is rigid in its frame structure, this paper generalises TSOBS to allow a hierarchy of frames. Termed hierarchical TSOBS (HiTSOBS), this scheme supports several granularities of rates, and permits multiple traffic classes with different loss-delay requirements to efficiently share the network. Our contributions are as follows: First, we present an architecture for HiTSOBS and offer it as a viable option for the realisation of flexible and cost-effective OBS networks. Second, we develop mathematical analysis to study the loss and delay performance of the proposed HiTSOBS system. Finally, we present simulation results that captures these loss-delay tradeoff values. Our HiTSOBS architecture gives network operators the freedom to choose the right mix of traffic with desired loss-delay requirements to coexist in the network.

[1]

I. Introduction

Wavelength Division Multiplexing (WDM) optical networks provide enormous bandwidth and are promising candidates for information transmission in next-generation high-speed networks. It is possible to realise 10-40 Gbps bandwidth on a single wavelength in commercial WDM networks today. However, a fundamental concern in the continued scalability of optical networks is the huge disparity in the switching speeds between optical and electronic switches in the core of the network. With a vision towards evolving to an all-optical Internet, optical switching can be classified into three categories - optical circuit switching (OCS), optical packet switching (OPS) and optical burst switching (OBS).

In OCS networks, lightpaths are used to transmit data between two end nodes [1], [2], where a lightpath is defined as an all-optical circuit switched medium with possible wavelength conversion at the intermediate nodes along the transmission path. Although OCS is easy to implement, it suffers from poor statistical multiplexing gains if the source node does not have any data to send, thereby leading to poor resource and bandwidth utilisation.

OPS [3], [4] on the other hand is similar to traditional electronic packet switching wherein packets are switched directly in the optical domain without the need for any electronic conversion. However, the most important concern is contention, which occurs at a switching node whenever two or more packets try to leave on the same output interface, on the same wavelength, at the same time. Unlike in electronic RAMs where as many as a million packets can be buffered during times of contention, buffering in the optical domain remains a very complex and expensive operation. Spools of fibre can implement fibre delay lines (FDL) that can buffer light by delaying the signal, however the size of the optical crossbar increases with bigger FDLs, thereby making all-optical switches very expensive. Recent research work [5], [6], [7], [8] explores the feasibility and performance of transport protocols for realising OPS networks in routers equipped with very small buffers, i.e., only a few dozen packet buffers that can be implemented in an on-chip optical memory. This remains an active area of research, and if successful, could lead to commercial large-scale deployment of OPS networks in the future.

Optical Burst Switching (OBS) [9], [10] is a hybrid of circuit and packet switching: aggregates of packets, called bursts, are switched atomically within the network, while a control packet is sent ahead of the burst to set up a short-lived end-to-end circuit for the burst. OBS thus combines the scalability of optics for fast data plane switching with the flexibility of electronics for switching decision control. An unfortunate consequence of this architecture, similar to OPS, is that the optical buffering required for contention resolution grows in proportion to the burst size. The control-plane advantage of large bursts is thus tempered by the larger buffers required in the data-plane.

A. Time Sliced Optical Burst Switching

A variant of optical burst switching, called Time Sliced OBS (TSOBS), was proposed in [11] to overcome this problem of having larger buffers. Time is divided into frames that contain a given number of fixed-length slots. TSOBS slices a burst, and transports successive slices in the same slot location of successive frames. This preserves the control-plane scalability of OBS (since only one switching decision is required to switch all slices belonging to a burst), while drastically reducing the optical buffering required at switching nodes (since a contending burst need only be buffered a slice at a time, independent of burst size). In addition, since switching is entirely done in the time domain rather than the wavelength domain, TSOBS eliminates the need for having wavelength converters, which substantially reduces the cost of designing

[1] A short summarized version of this paper was presented at the First Symposium on Advanced Networks and Telecom Systems (ANTS) 2007 in Mumbai, India, in December 2007.

such a network. Further, the authors identify three important factors that affect the cost and performance of optical time-slot interchangers (OTSI), which is a key component of the TSOBS system. They are the size of its internal crossbar, amount of fibre needed for the FDLs to reorder the timeslots and the number of switching operations that a burst may be subjected to when passing through the OTSI. Several blocking and non-blocking architectures for implementing the OTSIs are also proposed and analysed.

B. Related Work

Following the TSOBS system, the work in [12], [13] proposes a variant called Time-Synchronized Optical Burst Switching (SynOBS), which not only assumes the presence of fibre delay lines, but also considers the impact of full wavelength conversion. Several FDL reservation mechanisms - core node without FDLs, separated, shared and multi-length FDLs, are proposed and analysed using discrete time Markov chains to compute the burst drop probability. Their study also suggests that timeslot size must be chosen with care to achieve best timeslot utilisation, which subsequently reduces the burst blocking probability. In [14], the authors consider a slotted optical burst switching network (SOBS) and argue that such a network improves the overall link utilisation. They claim that their work is the first to point out the advantage of SOBS in supporting Quality of Service (QoS) requirements, and also propose a new cost-effective method for aligning packets at core nodes.

Akin to TSOBS but termed all-optical cell switching, [15] proposes FDL assignment algorithms to achieve low cell-loss rate to support both guaranteed and best-effort traffic. In [16], an analytical model is developed to estimate the overall blocking probability for a multi-fibre TSOBS network. The model is able to compute the overall blocking probability for circuit switched, best effort and multi-class traffic services in the network. Their results indicate that multi-fibre TSOBS can achieve the same level of performance (with respect to blocking probability) as a conventional OBS network (employing just-in-time reservation protocol [17]) with wavelength conversion functionality.

To address the fairness issue in OBS networks, [18] presents a new scheduling algorithm using round-robin scheduling, termed Almost Strictly Proportional Fair Scheduling (ASPFS), for SOBS networks with full wavelength conversion capability. SOBS is chosen to overcome the difficulty of the lack of large optical buffering in today's optical networks. Analytical and simulation results indicate that ASPFS is a promising candidate to provide fairness in future OBS networks. Slot allocation for TSOBS networks using centralised control is discussed in [19]. Request to calculate a path and an appropriate slot for a burst from an ingress OXC (Optical Cross Connect) is delivered to a centralised controller, which then computes these values. At the expense of an increased queueing delay at the ingress node, their scheme is able to improve channel utilisation, which is derived using both analysis and simulation. In [20], the authors propose a scheme to balance the loss-delay tradeoff in a slotted optical packet switched network. Using analysis and experimental results, the authors study the effect of ingress traffic conditioning, i.e., the effect of spacing out optical packets that feed into an OPS core node. They demonstrate that such a scheme can effectively bring down the packet loss probability to acceptable levels even when only minimal buffering is available at the core node. However, this low loss comes at the cost of an increased end-to-end delay of the conditioned traffic flow. The resulting strategy allows network service providers to choose the appropriate loss-delay values for operating their networks.

C. Our Contributions

While TSOBS successfully addresses the scalability of optical burst switching systems, it is excessively rigid in its frame structure. The frame size (i.e. number of slots per frame) is a key parameter that has to be universally pre-configured at all switches. A small frame size increases contention probability since overlapping bursts are more likely to pick the same slot number, while large frame sizes induce larger end-to-end delays due to each flow having access to a reduced fraction of the link capacity (one slot per frame), leading to significant queueing delay at the ingress edge node. This loss-delay trade-off, determined by frame size, is uniform across all traffic flows, and cannot be dynamically adjusted to provide differentiated QoS, making TSOBS too rigid for practical use.

We overcome these limitations of TSOBS by generalising the frame structure to a flexible hierarchy. Our idea draws inspiration from the hierarchical round-robin (HRR) packet scheduler proposed in [21], and is termed hierarchical TSOBS (HiTSOBS). As we will elaborate in the following section, HiTSOBS allows multiple frame sizes to concurrently co-exist, with slots lower in the hierarchy progressively offering lower rate service. This allows delay-sensitive traffic classes to operate at higher levels of the hierarchy while concurrently supporting loss-sensitive traffic at the lower levels. Along with the ability to support differentiated services to different traffic classes, HiTSOBS dynamically adapts the frame hierarchy as the traffic mix changes, thus obviating network-wide pre-configuration.

The rest of this paper is organised as follows. In Section II, we describe the HiTSOBS architecture in detail and examine its control and data plane operations. In Section III, we develop a mathematical model to study the performance of the HiTSOBS system, namely to estimate the loss probability and average delay. Simulation results of the proposed system are presented in Section IV. We conclude the paper in Section V.

II. Architecture

In this section, we first give an overview of how the frames are structured in HiTSOBS and subsequently explain the control and data plane operation.

A. Frame Hierarchy

Assume that time-slots are numbered consecutively, starting at 0. We select radix r which defines the number of slots in

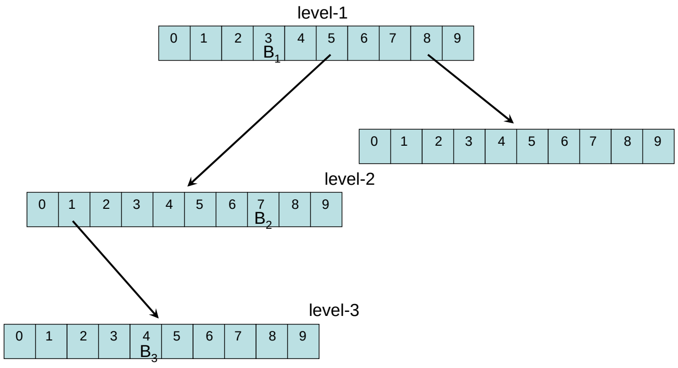


Fig. 1. HiTSOBS frame hierarchy

each frame in the HiTSOBS hierarchy. The top-level (level-1) frame therefore repeats every r slots. A burst transmitted at this level would occupy slots k, $k+r$, $k+2r$, $\dots$, $k+(B-1)r$ where k is the time-slot at which burst transmission starts and B is the size of the burst in slot units. For example, for radix $r = 10$, the burst B_1 of size 22 slots, shown in Fig. 1 to occupy the 3-rd slot in the level-1 frame, may be transmitted over time-slots 8043, 8053, 8063, $\dots$, 8253. Note that a given flow of bursts transmitted at level-1 has access to $1/r$ of the link capacity.

A slot in the level-1 frame may expand into an entire level-2 frame. For example, the 5-th slot in the level-1 frame in Fig. 1 expands into a level-2 frame. Successive slots in this level-2 frame are served in each successive turn of the 5-th slot of the level-1 frame. The burst B_2, shown to occupy the 7-th slot in this level-2 frame, may therefore be transmitted in time-slots 8175, 8275, 8375, and so on. Note that a burst transmitted at level-2 therefore has access to $1/r^2$ of the link bandwidth. Consequently, we can expect flows transmitting their bursts at level-2 of the hierarchy to have larger queueing delay at the edge compared to flows at level-1. However, the larger spacing between burst slices leaves more room for contention resolution using small optical buffers, making the losses for level-2 flows lower than for level-1 flows.

The reader can extend the above structure to more levels; in general a slot in a level-i frame transports the burst at $1/r^i$ of the link capacity. It is also easy to map a time-slot number to its position in the frame hierarchy: the k-th digit of the time-slot number read backwards denotes its position in the level-k frame, and the process terminates when a leaf node is encountered. Returning to our example with radix $r = 10$ illustrated in Fig. 1, if we are asked to determine the contents of time-slot 8415, we would traverse the 5-th slot in the level-1 frame, the 1-st slot in the level-2 frame, leading to the 4-th slot in the level-3 frame, which is a leaf showing that a slice of burst B_3 is carried in that slot. Such an operation will be required for the control pane operation described next.

B. Control Plane Operation

The HiTSOBS ingress edge node accumulates data into bursts, and classifies them into an appropriate QoS class. For illustration purposes, say there are two classes: real-time traffic that needs low latency and is not very sensitive to loss, and TCP traffic that is not very sensitive to latency but requires low loss. It would then be appropriate to transmit a real-time traffic burst at level-1, and a TCP traffic burst at a lower level, say level-2. Like the TSOBS network, HiTSOBS also sends a burst header control packet prior to the arrival of a data burst and contains three pieces of information: the level in the hierarchy at which the burst will be transmitted, the start slot, and the burst length. A core node receiving this control packet would first deduce the outgoing link for the bursts, and then determine where the slot lies in its hierarchy corresponding to that output link. There are three possible outcomes:

- A frame does not exist at the requested level in the hierarchy: For example, say Fig. 1 denotes the current hierarchy at the core node, and say the new burst is arriving at level-2 starting in slot 8234. The 4-th slot in the level-1 frame does not have a level-2 frame under it, so there are two options: either create a new level-2 frame under this slot (if the slot is unoccupied), or use a delay line to delay the burst slices by one slot, moving it to the 5-th slot in the level-1 frame, which already has a level-2 frame underneath, and in which the 3-rd slot may be used if available.
- A frame exists at the requested level but the required slot is unavailable: Again using Fig. 1 as an example, a new burst arriving at level-2 starting in slot 8375 collides with scheduled burst B_2. The new burst could be delayed using fibre loops by 10 slots to move it to the 8-th slot in the same level-2 frame. Alternatively, the new burst could be delayed by 3 slots to move it to the other level-2 frame if it has its 7-th slot available.
- A frame exists and the requested slot is available: In this case the burst is assigned the requested slot and passes through the switch in a cut-through manner without any delays.

It is important to note that the complexity of control plane operations does not depend on the burst length; much like OBS (and TSOBS), bursts are scheduled atomically (not slice-by-slice) by finding an appropriate free slot in the hierarchy for the entire burst. This preserves the control plane scalability of OBS.

C. Data Plane Operation

The data plane uses the hierarchy constructed by the control plane for each output link. A counter is maintained for each frame in the hierarchy, corresponding to the slot last served in that frame. Each time-slot, the counter for the level-1 frame is incremented by one, and the corresponding slot entry checked. If it is a leaf entry containing a burst, the optical crossbar is configured so that the input line corresponding to that burst is switched to the output link under consideration. If on the other hand the slot entry points to a lower level frame, the counter for the lower-level frame is incremented, and the process recurses. Note that the optical delay lines are also scheduled by this process by treating them as output ports on the optical crossbar (e.g. in a "shared memory" architecture [22] where all fibre delay lines are connected to the central crossbar).

The complexity of the data plane operation per time-slot at most equals the number of levels in the frame hierarchy, which can be capped at a small constant. This preserves the scalability of OBS to high data plane rates.

III. Analytical Model

In this section we develop an idealised analytical model for the loss and delay in a HiTSOBS network transporting flows at several levels of the frame hierarchy. For our analysis, time is measured in units of timeslots, while service is measured in units of slices (which corresponds to the amount of data that can be transported by the HiTSOBS system in one timeslot). The core optical links thus operate at unit rate, i.e. one slice per timeslot. We assume bursts corresponding to flow i arrive according to any arrival process, with mean arrival rate $\lambda_i/\bar{B}$ bursts per timeslot, where $\bar{B}$ denotes the average burst size. The arrival rate of flow i measured in slices per timeslot is therefore λ_i. Further, we denote by k_i the level of the frame hierarchy at which the edge node transmits burst slices of flow i; therefore, fraction $f_i = r^{-k_i}$ of the link capacity is available to flow i, where r denotes the radix (number of slots in each frame) of the frame hierarchy. For stability, $\lambda_i < f_i$ should hold for all flows.

A. Estimating Loss

We first analyse the loss at an arbitrary core node in the HiTSOBS network. Our loss estimate uses a **fluid** approximation. Namely, though arrival of bursts to the edge remains a point process, departures (which happen one slice at a time) can be approximated as a fluid process, particularly when slices are much smaller than bursts (in the limiting case when timeslots become infinitesimally small, all core traffic is indeed fluid). Under this assumption, the traffic contribution of flow i to the HiTSOBS core at a random point in time can be denoted by a random variable X_i given by:

$$X_i = \begin{cases} f_i & \text{with probability } \lambda_i/f_i \\ 0 & \text{with probability } 1 - \lambda_i/f_i \end{cases} \tag{1}$$

In words, this states that the flow either contributes fluid traffic into the HiTSOBS core (at service rate $f_i = r^{-k_i}$ available at the edge node server), which happens with probability equal to the utilisation of the edge node server, or contributes nothing when the edge node server is idle. Note that $E[X_i] = \lambda_i$ and $var(X_i) = \lambda_i(f_i - \lambda_i)$.

Having quantified the instantaneous traffic load contributed by each flow to the core node under consideration, we estimate the fluid loss at the core under a **bufferless** assumption. Loss estimates under bufferless fluid approximations have been used extensively in literature (e.g. [23], [24]) for analysing packet switching systems, and are justified in this case given the very small buffering (OTSIs) that core HiTSOBS nodes are likely to have. If N denotes the number of flows multiplexed at the core link, and X the overall arrival rate to the core link at a random instant of time, then $X = \sum_{i=1}^{N} X_i$. If N is sufficiently large, and the traffic from different flows is independent, we can apply the central limit theorem to approximate X by a normal distribution whose mean and variance are the sums over all flows:

$$X \approx \mathcal{N}\left(\sum_{i=1}^{N} \lambda_i, \sum_{i=1}^{N} \lambda_i(f_i - \lambda_i)\right) \tag{2}$$

Under the bufferless fluid approximation, loss happens when the aggregate arrival rate to the core link exceeds its (unit) service capacity, i.e. when $P[X > 1]$. Based on the normal approximation of Equation (2), the loss can be estimated using

$$L = \phi\left(\frac{1 - \sum_{i=1}^{N} \lambda_i}{\sqrt{\sum_{i=1}^{N} \lambda_i(f_i - \lambda_i)}}\right) \tag{3}$$

where $\phi(.)$ denotes the complementary cumulative distribution function of the normal distribution.

We note that the loss estimate in Equation (3) makes no assumption about the arrival process itself, other than that it has a mean and that the mean is lower than the service rate given to this flow by the edge node (i.e. the edge node queueing system is stable).

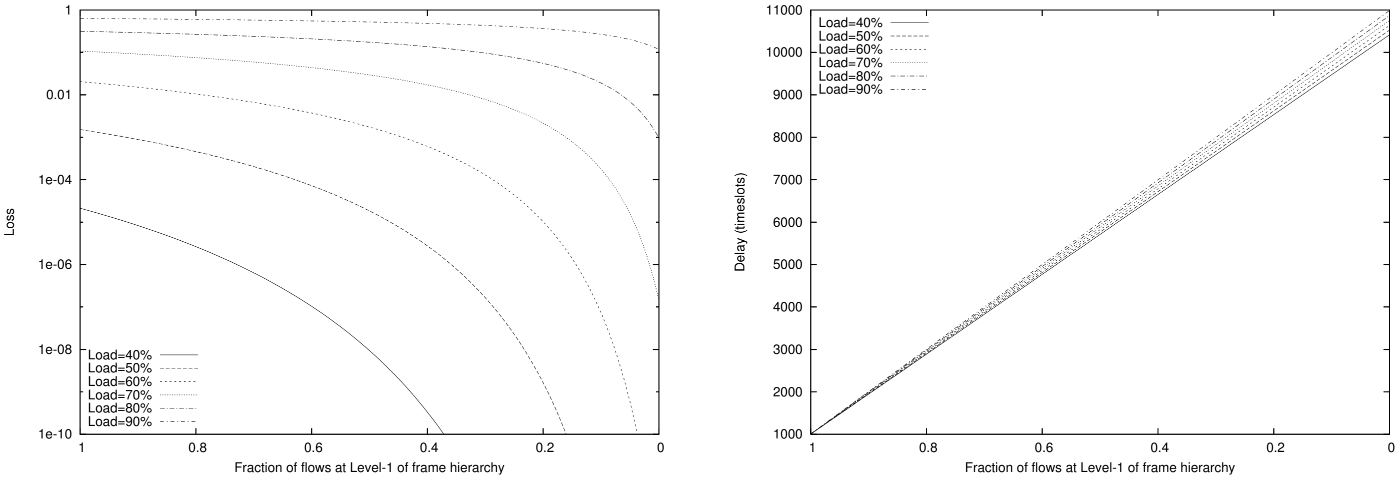
To illustrate the significance of Equation (3), we consider a two-class HiTSOBS system (i.e. with $K = 2$ levels in the frame hierarchy) with radix $r = 10$: class-1 flows, N_1 in number, each with arrival rate λ_1 (slices per timeslot), have their burst slices transported at level-1 in the frame hierarchy, thereby getting rate $1/r = 0.1$ of the link capacity, while class-2 flows, N_2 in number, each with arrival rate λ_2 slices per timeslot, have their bursts transported at level-2 of the hierarchy, thereby getting fraction $1/r^2 = 0.01$ of the link capacity. For this two-class system, the loss estimate of Equation (3) becomes:

$$L = \phi\left(\frac{1 - (N_1\lambda_1 + N_2\lambda_2)}{\sqrt{N_1\lambda_1(1/r - \lambda_1) + N_1\lambda_2(1/r^2 - \lambda_2)}}\right) \tag{4}$$

Fig. 2(a) shows the loss at a core node as the total number of flows is constant at $N = N_1 + N_2 = 1000$ while the fraction of class-1 flows N_1/N is varied in $[1, 0]$. The different curves in the plot correspond to different loadings $N_1\lambda_1 + N_2\lambda_2 = 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$ of the core link. The figure shows clearly that as the fraction of class-1 flows decreases, namely as more flows are pushed to the lower level of the hierarchy, the aggregate loss probability in the HiTSOBS core node steadily decreases.

B. Delay Estimate

The end-to-end delay for bursts of any particular flow include the fixed propagation delay on the flow path, the queueing delay at the edge node, and the delay incurred at the



(a) Aggregate Loss

(b) Average Delay

Fig. 2. Analytical estimate of aggregate loss and average delay as a function of fraction of high priority flows for $N = 1000$ flows and $\rho = 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$.

optical buffers (OTSI) in the core nodes. The last component is negligible, since the HiTSOBS core performs cut-through switching for the most part, and has very little optical buffers (OTSIs). We ignore the propagation delays as they are fixed, and focus only on the queueing delays at the edge.

The edge queueing delay for a burst is measured from the time the burst arrives at the edge node till it completes service, i.e. its last slice has been sent into the core. The mean delay for bursts of flow i is independent of other flows, and can be computed using classical queueing theory – if flow i bursts arrive as a Poisson process with arrival rate $\lambda_i/\bar{B}$ (bursts per timeslot), have exponentially distributed lengths with mean $\bar{B}$ (slices), and on average f_i slices are released every timeslot, the average delay (in timeslots) is given by the M/M/1 expression

$$\bar{D_i} = \frac{1}{f_i/\bar{B} - \lambda_i/\bar{B}} = \frac{\bar{B}}{f_i - \lambda_i} \tag{5}$$

Though delay expressions for more general arrival processes (e.g. G/G/1) have been derived in the literature, the M/M/1 expression above suffices for the illustration in this paper. The above expression can be used to compute the delay for each individual class.

Considering again the two-class system described in the previous subsection, we can compute the average delay D across all N flows in the system as:

$$\bar{D} = \sum_{i=1}^{N} D_i/N = \frac{N_1\bar{B}}{N(f_1 - \lambda_1)} + \frac{N_1\bar{B}}{N(f_2 - \lambda_2)} \tag{6}$$

Fig. 2(b) shows the queueing delay averaged over all flows as the fraction of class-1 flows varies in $[1, 0]$ while the total number of flows is held constant at 1000. As expected, as more flows are moved to the lower level in the hierarchy, the delays increase. Thus increasing the fraction of traffic carried in lower levels of the hierarchy reduces network losses (Fig. 2(a)) but increases end-to-end delays (Fig. 2(b)). Understanding this trade-off can assist the network operator in deciding the right mix of traffic to put at each level of the hierarchy based on the performance requirements of various traffic flows.

IV. Simulations

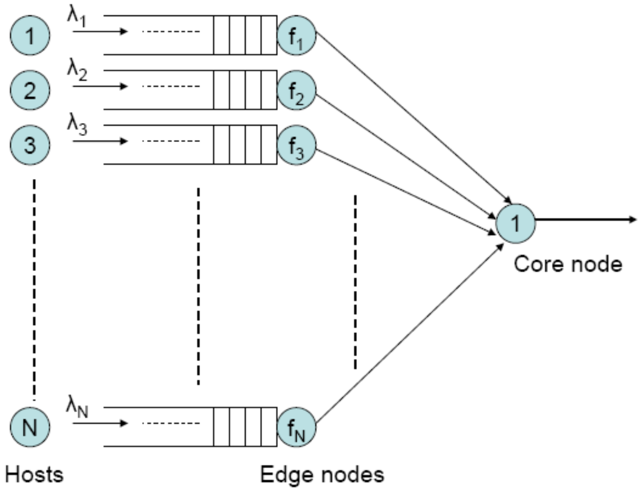
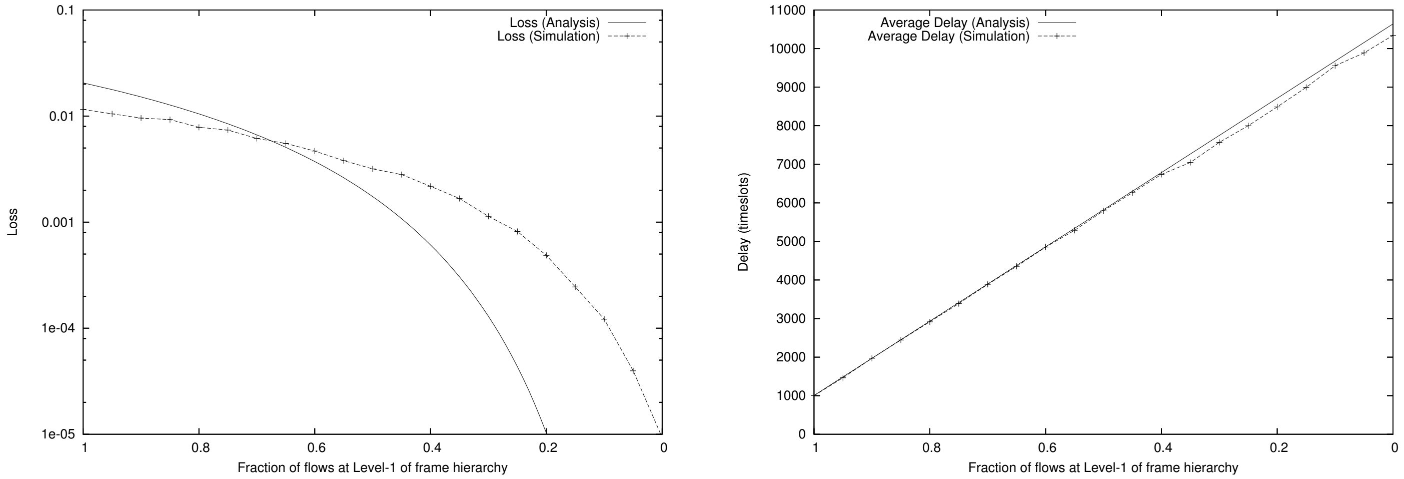


Fig. 4. Simulation topology with edge nodes and core node

To validate the analytical model developed above, and to verify the effectiveness of the HiTSOBS architecture in supporting traffic classes with different loss-delay requirements, we developed our own discrete-event simulation model in the C programming language. Our simulation considered a very simple topology, shown in Fig. 4, consisting of $N = 1000$ flows, each originating at its own edge node, being multiplexed at a single core node. Bursts for flow i arrive as a Poisson process at rate $\lambda_i/\bar{B}$ bursts per timeslot. Our timeslot was



(a) Aggregate Loss

(b) Average Delay

Fig. 3. Aggregate loss and average delay as a function of fraction of high priority flows from analysis and simulation for $N = 1000$ flows and $\rho = 0.6$.

chosen to correspond to 1μs, which is consistent with the switching speeds of solid-state optical switching technologies available today [25]–[27]. Our link rate was chosen at 10 Gbps, which makes the burst slice of size 1250 Bytes that can be transported in a timeslot. We chose the burst sizes to have an exponential distribution with mean $\bar{B} = 100$ slices or 125 KBytes. We kept the loading of the core link at $\rho = 0.6$, namely 60% of the timeslots at the core link carry slices, which we believe is reasonable. For simplicity, we make each flow contribute equally to this load; thus each flow generates $\rho/N = 0.6 \times 10^{-3}$ slices per slot, which corresponds to a burst arrival rate of $\lambda_i = \frac{\rho/N}{\bar{B}} = 6 \times 10^{-6}$ for $i = 1, 2, \ldots, N$. Our simulation only supports two levels of frame hierarchy, and uses a frame radix size of $r = 10$. Thus a slot in a level-1 frame gives a flow 0.1 of the link capacity, while a level-2 slot gives a flow 0.01 of the link capacity.

Each flow is assigned a priori to one of the two levels in the HiTSOBS frame hierarchy. Upon arrival of a flow's burst at the edge node, the following processing happens: if the arriving burst encounters a non-empty queue, the burst is queued and awaits service. If on the other hand the arriving burst encounters an empty queue, the edge node reserves the first available slot in the appropriate level of the frame hierarchy for the duration of the burst (i.e. the slot is reserved over a number of frames equal to the burst length), and the burst is transmitted on to the core node. If the queue for the flow is non-empty, namely there are more bursts awaiting service, the edge node reserves the same slot for the subsequent burst, and the process continues. It is important to note that the slot positions for burst slices for any given flow vary each time the flow becomes newly backlogged; this randomness helps prevent synchronisation and phase locking amongst the various flows.

The core node operates on a single wavelength (wavelength conversion is not considered in this paper, so wavelengths operate independently), and is equipped with a very small buffer of capacity $B = 10$ slices. The buffer capacity is chosen to be 10 slices so that the lossless scenario from analysis corresponds to the lossless scenario from simulations. In other words, as losses occur at the core node only when the aggregate arrival rate exceeds its capacity of unity, we need more than 10 sources at level-1 of the frame hierarchy to be active at any instant of time to induce losses, since level-1 frames transport bursts at a rate of 0.1. To capture this in simulations, the buffer capacity is chosen to be 10 slices.

When the core node receives burst slices it schedules them on the output link in a cut-through fashion. Contending slices, namely ones that request the same slot in the same level of the frame hierarchy, are buffered if space is available, and are dropped otherwise. We measure the fraction of slices dropped at the core, as well as the delay incurred by bursts of each flow at the edge nodes.

We keep the total number of flows constant at $N = 1000$, and vary with each simulation run the fraction of flows that are transported at level-1 of the hierarchy. Each run simulated the operation of the HiTSOBS system for 100 million timeslots. Fig. 3(a) shows the aggregate loss, while Fig. 3(b) the overall average delay in the network as the traffic mix changes. The delay estimate from the M/M/1 analysis matches very well with the average delay observed in simulation, as expected. The loss prediction from analysis shows the same general shape as obtained from simulation, though the numerical match is not as close, particularly at very low loss values. This is because the normal approximation is not very accurate when applied to the tail of the distribution (i.e., when we try to estimate the probability of values far from the mean). The bufferless fluid assumption in the analysis also makes it approximate. The difference not withstanding, we believe the analysis is able to capture fairly well the shape of the loss curve, and shows clearly that losses in the HiTSOBS network decrease dramatically as more and more flows are moved to lower levels in the hierarchy. The accompanying cost is the increase in end-to-end delays (averaged over all flows), as shown in Fig. 3(b). This verifies that the HiTSOBS architecture does support flows with different loss-delay trade-off requirements: an operator can move flows with loose delay requirements to lower levels in the hierarchy, thereby improving loss performance in the network. Our analytical model can be used by the operator to adjust the mix of traffic at the various levels of the HiTSOBS hierarchy to operate the network at the desired loss-delay trade-off point.

V. Conclusions

In this paper we have presented an architecture for hierarchical time-sliced optical burst switching (HiTSOBS). HiTSOBS preserves the data and control plane scalability of OBS, while introducing a flexible frame hierarchy that allows different traffic classes to operate at different loss-delay trade-off points, which was not feasible in the TSOBS architecture. We also presented an approximate analytical model to estimate the loss and average delay, and evaluated the performance of HiTSOBS via simulation.

References

[1] R. Ramaswami and K. N. Sivarajan, "Routing and Wavelength Assignment in All-Optical Networks," *IEEE/ACM Transactions on Networking*, vol. 3, pp. 489–499, Oct. 1995.

[2] D. Banerjee and B. Mukherjee, "A Practical Approach for Routing and Wavelength Assignment in Large Wavelength-Routed Optical Networks," *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 5, pp. 903–908, June 1996.

[3] L. Xu, H. G. Perros and G. N. Rouskas, "A Survey of Optical Packet Switching and Optical Burst Switching," *IEEE Communications Magazine*, vol. 39, no. 1, pp. 136–142, Jan. 2001.

[4] G. N. Rouskas and L. Xu, "Optical Packet Switching," Book chapter in *Emerging Optical Network Technologies: Architectures, Protocols and Performance*, pp. 111–127, Springer, 2004.

[5] M. Enachescu, Y. Ganjali, A. Goel, N. McKeown and T. Roughgarden, "Routers with Very small Buffers," *Proc. IEEE INFOCOM*, Barcelona, Spain, Apr. 2006.

[6] V. Sivaraman, H. Elgindy, D. Moreland and D. Ostry, "Packet Pacing in Short Buffer Optical Packet Switched Networks," *Proc. IEEE INFOCOM*, Barcelona, Spain, Apr. 2006.

[7] A. Vishwanath, V. Sivaraman and G. N. Rouskas, "Are Bigger Optical Buffers Necessarily Better?," *Proc. IEEE INFOCOM Student Workshop*, Phoenix, Arizona, USA, Apr. 2008.

[8] A. Vishwanath and V. Sivaraman, "Routers With Very Small Buffers: Anomalous Loss Performance for Mixed Real-Time and TCP Traffic," *Proc. 16th IEEE International Workshop on Quality of Service (IWQoS)*, Netherlands, June 2008.

[9] C. Qiao and M. Yoo, "Optical Burst Switching (OBS) – A New Paradigm for an Optical Internet," *Journal of High Speed Networks*, vol. 8, no. 1, pp. 69–84, 1999.

[10] J. S. Turner, "Terabit Burst Switching," *Journal of High Speed Networks*, vol. 8, no. 1, pp. 3–16, 1999.

[11] J. Ramamirtham and J. Turner, "Time Sliced Optical Burst Switching," *Proc. IEEE INFOCOM*, San Francisco, California, USA, Mar. 2003.

[12] A. Rugsachart and R. A. Thompson, "An Analysis of Time-Synchronized Optical Burst Switching," *Proc. IEEE Workshop on High Performance Switching and Routing*, Poland, June 2006.

[13] A. Rugsachart, "Time-Synchronized Optical Burst Switching," *PhD Thesis*, University of Pittsburgh, 2007.

[14] Z. Zhang, L. Liu and Y. Yang, "Slotted Optical Burst Switching (SOBS) Networks," *Proc. IEEE International Symposium on Network Computing and Applications*, USA, July 2006.

[15] H. J. Chao and S. Y. Liew, "A New Optical Cell Switching Paradigm," *Proc. First International Workshop on Optical Burst Switching*, Dallas, Texas, USA, Oct. 2003.

[16] O. Liang, T. Xiansi, M. Yajie and Y. Zongkai, "A Framework to Evaluate Blocking Performance of Time-slotted Optical Burst Switched Networks," *Proc. IEEE LCN*, Sydney, Australia, Nov. 2005.

[17] J. Y. Wei and R. I. McFarland, "Just-in-Time Signaling For WDM Optical Burst Switching Networks," *IEEE Journal of Lightwave Technology*, vol. 18, no. 12, pp. 2019–2037, Dec. 2000.

[18] L. Liu and Y. Yang, "Fair scheduling in optical burst switching networks," *Proc. 20th International Teletraffic Congress (ITC-20)*, Ottawa, Canada, June 2007.

[19] T. W. Um, J. K. Choi, S. G. Choi and W. Ryu, "Performance Analysis of a Centralized Resource Allocation Mechanism for Time-Slotted OBS Networks," *Proc. 9th APNOMS*, Korea, Sep. 2006

[20] V. Sivaraman, D. Moreland and D. Ostry, "Ingress Traffic Conditioning in Slotted Optical Packet Switched Networks," *Proc. ATNAC*, Sydney, Australia, Dec. 2004.

[21] C. R. Kalmanek, H. Kanakia and S. Keshav, "Rate Controlled Servers for Very High-Speed Networks," *Proc. IEEE GLOBECOM*, San Diego, California, USA, Dec. 1990.

[22] S. Yao, S. Dixit and B. Mukherjee, "Advances in Photonic Packet Switching: An Overview," *IEEE Communications Magazine*, vol. 38, no. 2, pp. 84–94, Feb. 2000.

[23] A. Elwalid and D. Mitra, "Design of Generalized Processor Sharing Schedulers Which Statistically Multiplex Heterogeneous QoS Classes," *Proc. IEEE INFOCOM*, New York, NY, USA Apr. 1999.

[24] M. Reisslein, "Measurement-Based Admission Control for Bufferless Multiplexers," *International Journal of Communication Systems*, vol. 14, no. 8, pp. 735–761, June 2001.

[25] T. McDermott and T. Brewer, "Large-Scale IP Router Using a High-Speed Optical Switch Element," *Journal of Optical Networking*, vol. 2, no. 7, pp. 229–240, Jul. 2003.

[26] F. Masetti, D. Chiaroni, R. Dragnea, R. Robotham, and D. Zriny, "High-Speed High-Capacity Packet-Switching Fabric: A Key System for Required Flexibility and Capacity," *Journal of Optical Networking*, vol. 2, no. 7, pp. 255–265, Jul. 2003.

[27] J. Gripp *et al.*, "Optical Switch Fabrics for Terabit-Class Routers and Packet Switches," *Journal of Optical Networking*, vol. 2, no. 7, pp. 243–254, Jul. 2003.