WDM Burst Switching for Petabit Data Networks

Jonathan S. Turner
Washington University, St. Louis
jst@cs.wustl.edu

Introduction. Bandwidth usage in the Internet is doubling every six to twelve months [1] and data network capacities are now exceeding voice network capacities. The emergence of WDM technology is unlocking more bandwidth, leading to lower costs further fueling demand. While ATM switches and IP routers can switch data using the individual channels within a WDM link, this approach implies that tens or hundreds of switch interfaces must be used to terminate a single link. Moreover, there can be a significant loss of statistical multiplexing efficiency when WDM channels are used simply as a collection of independent links, rather than as a shared resource. The growing mismatch between optical and electronic technologies is creating an opportunity for more extensive use of optical components within switching systems.

WDM burst switching is an attempt at a new synthesis of optical and electronic technologies, which seeks to exploit the tremendous bandwidth of optical technology, while using electronics for management and control. It uses sophisticated routing and resource management mechanisms to enable efficient usage of the optical bandwidth, while providing the flexibility needed for direct support of standard data communication protocols, like IP. This paper briefly summarizes the operational principles of WDM burst switching networks, explains the key performance and cost issues that must be addressed by burst switching systems, and describes algorithms for efficient assignment of data bursts to channels on WDM links.

Operational Principles. Burst switching systems assign user data bursts to channels in WDM links on-the-fly in order to provide efficient statistical multiplexing of high rate data channels. The transmission links in a burst network carry multiple WDM *channels*, which can be dynamically assigned to user data bursts. One channel is reserved for control information. The information in the control channel is interpreted by the control portion of the switching systems, while the data channels are switched through transparently with no examination or interpretation of the data. This separation of control and data simplifies the data path implementation, facilitating optical implementation.

When an end system has a burst of data to send, an idle channel on the access link is selected, and the data burst is sent on that idle channel. Shortly before the burst transmission begins, a *Burst Header Cell* (BHC) is sent on the control channel, specifying the channel on which the burst is being transmitted and the destination of the burst. A burst switch, on receiving the BHC, selects an outgoing link leading toward the desired destination with an idle channel available, and then establishes a path between the specified channel on the access link and the channel selected to carry the burst.

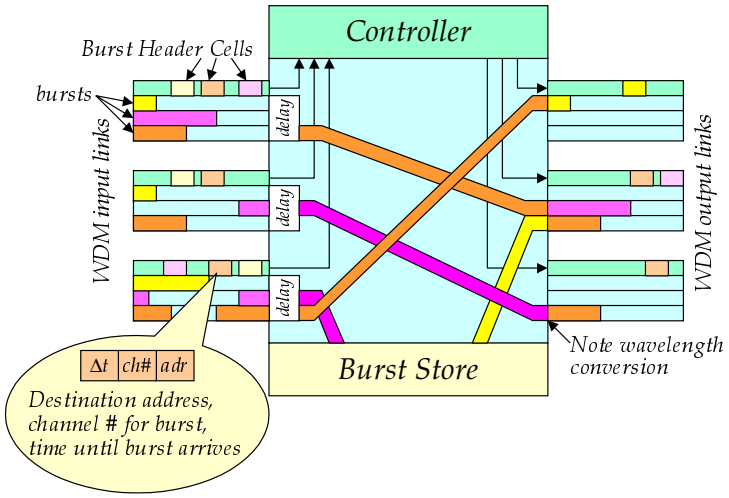

Figure 1: Operation of a Burst Switch

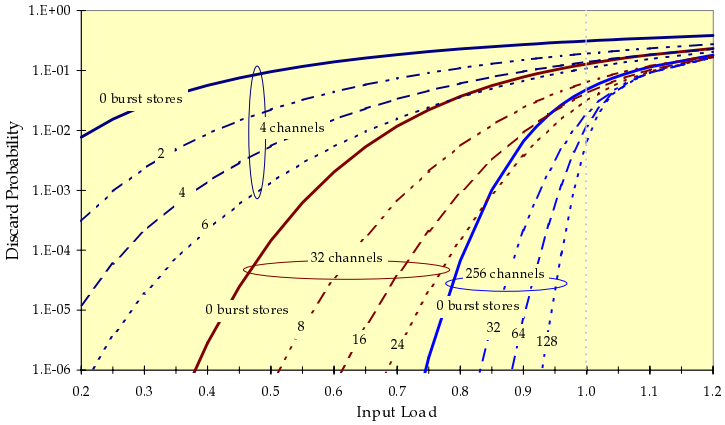

Figure 2: Statistical Multiplexing Performance of Burst Switching

It also forwards the BHC on the control channel of the selected link, after modifying the cell to specify the channel on which the burst is being forwarded. This process is repeated at every switch along the path to the destination. The BHC includes an *Offset Field*, specifying the time between the transmission of the first bit of the BHC and the first bit of the burst, plus a *Length Field* specifying the amount of data in the burst. These are used to schedule the setup and release of paths through the switch and the outgoing channels. As bursts and BHCs propagate, the offset is updated to reflect any differences in the data and control path delays. The operation of a small burst switch is illustrated in Figure 1.

Performance and Cost Considerations. One of the key performance concerns for a burst switching network is the statistical multiplexing performance. We need to know how to engineer a burst-switched network so that it can be operated at reasonably high levels of utilization, with acceptably small probability that a burst is discarded due to lack of an available channel or storage location. Figure 2 shows the results of a first-order analysis of the statistical multiplexing performance of a link in a burst-switched network. The chart shows burst discard probability for links with 4, 32 and 256 channels, with several different choices for the number of burst storage locations. Note that when the number of channels is large, reasonably good statistical multiplexing performance can be obtained with no burst storage at all. With 32 channels, 8 burst stores is sufficient to handle bursty data traffic at a 50% load level (an average of 16 of the 32 channels are busy), with discard probabilities of less than one in a million. With 256 channels and 128 burst stores, loading levels of over 90% are possible with low loss rates. Note that the statistical multiplexing characteristics of burst switching systems are nicely matched to the technology trends for WDM links. As the number of channels per link increases, the statistical multiplexing efficiency rises, and the need for storage drops.

Large capacity burst switching systems can be constructed using multistage interconnection networks composed of moderate-sized *Burst Switch Elements* (BSE), as described in [3]. That system architecture uses a $2k-1$ stage Beneš network constructed from BSEs with d inputs and outputs. The system supports $n = d^k$ external links, with h channels per link. The BSEs include optical crossbars, wavelength selectors (a selector can propagate a specified one of h input wavelengths, while rejecting the rest) and wavelength converters (a converter uses an optical input signal of arbitrary wavelength to modulate a fixed wavelength optical carrier, to produce an output at the carrier wavelength). The crossbars and wavelength selectors must be capable of switching times of 10 ns or less to provide efficient handling of short data bursts. The architecture requires $nh(2k-1)d$ optical crosspoints plus $nh(2k-1)$ wavelength selectors and wavelength converters or $(2k-1)d$ optical crosspoints and $(2k-1)$ wavelength selectors and converters, per external channel. So, for $d = 8$ and $k = 3$ (implying $n = 512$) we require 40 optical crosspoints per external channel and 5 wavelength selectors and converters. Substantial progress is needed in optical component technology to make such systems cost-effective relative to electronic counterparts. A large electronic router can be constructed at a parts cost of less than $1,000 per Gb/s of throughput. For the example burst switch configuration cited above to be competitive, the combined cost of 40 optical crosspoints and 5 wavelength selectors and converters must drop well below $10,000 (assuming 10 Gb/s per WDM channel).

Link Scheduling. In order for burst switching systems to be effective in modern data networks, they must be

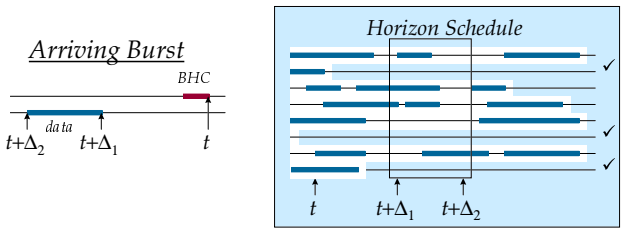

Figure 3: Horizon Channel Scheduling

capable of efficient operation, even in the presence of traffic mixes that include a large percentage of fairly short data bursts (1-10 KB). This means that the control subsystem must be capable of high burst processing rates, and that it cannot simply reserve a channel for a burst when the Burst Header Cell (BHC) arrives, since the time between the arrival of the BHC and the burst itself may be as long or longer than the burst duration. *Lookahead Resource Management* avoids the inefficiency of early reservation, by maintaining a schedule of future channel usage for a link, and assigning bursts to channels based on the projected usage.

One simple technique for channel assignment is *horizon scheduling*. In horizon scheduling, the controller for a link maintains a *time horizon* for each of the channels of an outgoing link. The horizon is defined as the earliest time after which there is no planned use of the channel. The horizon scheduler assigns arriving bursts to the channel with the latest horizon that is earlier than the arrival time of the burst, if there is such a channel. If there is no such channel, the burst is assigned to the channel with the smallest horizon and is diverted to the burst storage area where it is delayed until the assigned channel is available. Horizon scheduling is straightforward to implement in hardware, but because it does not keep track of time periods before a channel's horizon when the channel is unused, it cannot insert bursts into these open spaces. The horizon scheduler can provide good performance if the time between the arrival of a BHC and the subsequent arrival of a burst is subject to only small variations. However if the variations are as large or larger than the time duration of the bursts, the performance of horizon scheduling can deteriorate significantly.

We can improve the performance of horizon scheduling by processing bursts out-or-order. Rather than process bursts as soon as their BHCs arrive, one can delay processing, and then process the bursts in the order of expected burst rrival, rather than the order in which the burst header cells arrived. As the burst header cells arrive, they are inserted into a resequencing buffer, in the order in which the bursts are to arrive. A horizon scheduler then processes requests from the resequencing buffer. The processing of a request is delayed until shortly before the burst is to arrive, reducing the probability that we later receive a burst header cell for a burst that will arrive before any of the bursts that have already been scheduled. If the *lead time* for processing bursts is smaller than the burst durations, optimal performance can be obtained.

Acknowledgements. This work is supported by the Advanced Research Projects Agency and Rome Laboratory (contract F30602-97-1-0273).

References

[1] Coffman, K. G. and A. M. Odlyzko. "The Size and Growth Rate of the Internet," unpublished technical report, available at http:// www.research.att.com/ amo/doc/networks.html, 1998.

[2] Turner, Jonathan S. "Terabit Burst Switching," *Journal of High Speed Networks*, 1999.

[3] Turner, Jonathan S. "WDM Burst Switching," *Proceedings of INET*, 6/99. Also, on the web at http:// www.arl.wustl.edu/~jst/pubs/inet99/inet99.html.

[4] Turner, Jonathan S. "WDM Burst Switching for Petabit Capacity Routers," *Proceedings of Milnet*, 11/99.

WDM Burst Switching for Petabit Data Networks

Jonathan S. Turner
Washington University, St. Louis
jst@cs.wustl.edu

Abstract

WDM burst switching combines optical data paths and electronic control to enable very high capacity routing switches. This paper shows how expected advances in optical component technology and electronics can lead to petabit capacity routers.