Optical Switching Speed Requirements for Terabit/Second Packet Over WDM Networks

D. Sadot, *Member, IEEE,* and I. Elhanany, *Student Member, IEEE*

***Abstract*—Optical dynamics requirement for packet-over-WDM networks is analyzed. Optimization among optical switching speed, global resource availability, and local queuing considerations is performed, yielding multiterabit/second throughput capability by employing submicrosecond switching technology.**

***Index Terms*—High-speed scheduling algorithms, packet switching architectures, WDM networks.**

I. Introduction

THE exponential growth of the Internet and demand for datacom, telecom, and multimedia services has forced carriers and communication equipment providers to look for a new scale of solutions for the backbone bandwidth and quality of service (QoS) infrastructure. Multiterabit/second (Tb/s) routers and switches are already being developed. At the same time, several switching schemes and scheduling algorithms are being designed in order to support the high capacity, large number of ports, and low latency requirements of such networks [1], [2]. Conversely, presently available switches and routers are confined to an aggregated capacity of up to 200 Gb/s. Promising solutions rely on recent developments in high-speed optical switching technology such as free-space electroholographic switches [3], free-space micromechanical switches [4], and fast tunable wavelength-division multiplexing (WDM) transmitters [5], [6] that can be used as switching elements. As high-speed optical switching technology is still in its research and early development stages, it is extremely important to characterize and define the requirements of its dynamics from the packet-switching perspective. While most commercially available millisecond-range elements are limited to support circuit-switching functionality, it is not yet clear what the adequate optical switching speed requirements are for packet switching functionality, nor whether microsecond devices are sufficient or nanosecond switches should be developed.

II. Scheduling Approach

Here, the issue of switching speed is being analyzed from a three-lag tripod perspective. The first lag is the optical technology viewpoint. Switching schemes in the range of microseconds or even hundreds of nanoseconds seem to be already in hand [3]–[6], while it will still take a few more years to reach commercial large-scale nanosecond switching, e.g., via current injection or electrooptic devices. The second lag of the tripod is the global router resource availability, which is imperative for overall switching performance, e.g., maximizing throughput and minimizing mean packet delay. If it were practical, for example, to consider requests for all available switch resources (channels), employing a weighted maximum matching algorithm would be efficient. Unfortunately, the global centralized approach is impractical since it can not be implemented in real time; calculations scale proportionally to n^3, with n being the number of channels. In addition, under nonuniform traffic conditions, maximum matching algorithms are known to perform poorly and cause queue starvation, resulting in limited QoS provision.

Consequently, the third lag of the tripod is the local queuing perspective. Due to practical VLSI issues, and in order to meet queue dynamic requirements, most pragmatic algorithms utilize distributed schemes. Decisions are made concurrently at each port, based merely on local information, such as queue length and average packet delay. Contention is typically avoided using backpressure mechanisms or by updating scheduling pointers. The major advantage of such approaches is their capability of reaching real-time decisions on a "packet-by-packet" basis, while the dominant disadvantages are lack of global optimization of resources and a normalized priority mechanism for the provision of QoS.

Summarizing the three-lag strategy:

a) realistic optical switching technology limits switching speeds to submicroseconds or longer;
b) decision time for global considerations does not scale and is therefore infeasible on a "packet-by-packet" basis;
c) shorter scheduling decisions yield finer switching resolution and hence potentially better performance from the local queue perspective.

It is shown here that although cell-by-cell switching is commonly perceived as the goal of any scheduling scheme, it is beneficially preferred to produce optimized switching decisions once every several cell times providing the latter yields improved overall performance. As packet duration in high-speed networks is approximately 50 ns (i.e., asynchronous transfer mode (ATM) cells at oc-192 are 42.61 ns in length), the notion of "several cell times" translates to a few hundred nanoseconds allowed per switching event. It is thus the aim of the switching architecture to guarantee performance at the expense of lengthier scheduling decisions.

Manuscript received June 21, 1999; revised October 22, 1999. This work was supported by the Israeli Ministry of Science under Contract 8535-2-96 and Contract 9404-1-97, by the Wolfson Foundation, and by the Rector's Office of Ben-Gurion University.

The authors are with the Electrical and Computers Engineering Department, Ben-Gurion University, Beer-Sheva 84105, Israel (e-mail: sadot@ee.bgu.ac.il).

Publisher Item Identifier S 1041-1135(00)02877-9.

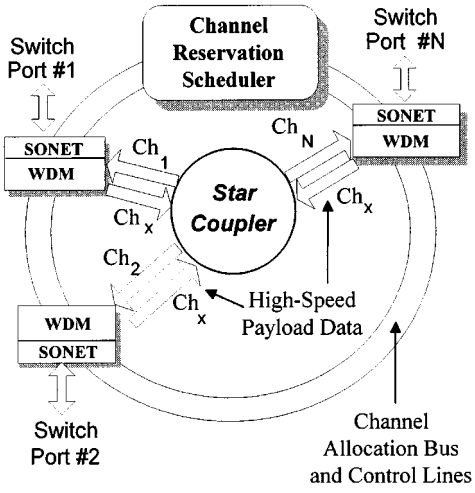

Fig. 1. The proposed switch architecture.

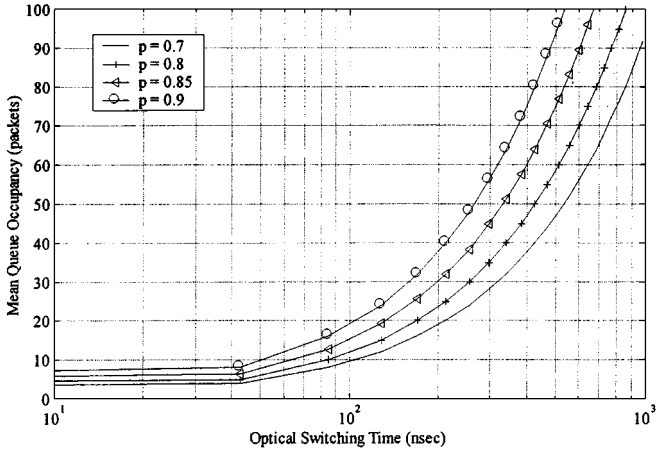

Fig. 2. Mean queue occupancy for uniform distributed Bernoulli traffic.

III. Switch Architecture

Here, a switch architecture that produces a switching decision once in several cell times is proposed. Accordingly, a switching-time valuation metric for optical cross-connect implementations is derived. In the proposed switch architecture, depicted in Fig. 1, each node includes a SONET module and a WDM module. The SONET module connects the high-speed SONET data stream, which originates from the switch port, with the optical WDM interconnection. Data are optically transmitted and routed through a passive star to the WDM receiver module at the relevant destination. In turn, at the destination node, the data are retransmitted via the SONET module to the desired remote destination. The optical switching functionality is performed by means of a tunable-transmitters and fixed-receivers (TTFR) strategy. The header in each packet indicates the desired destination relating to a specific WDM module at the switch. According to a scheduling procedure, the local controller sets the wavelength of the fast tunable laser at the WDM module of the source port to a specific wavelength corresponding to the wavelength of the fixed-tuned optical filter at the WDM module of the destination port. By modulating the tunable laser, packets are transmitted through the star coupler and received at the relevant WDM module. To meet the foregoing goals, the optical dynamic WDM portion of the network is physically organized in a small area, i.e., a single rack or central office. All 2.5/10-Gb/s streams, independent of their distance, are connected to the central WDM interconnection site. Switching-time requirements thus apply to the wavelength tuning duration. Submicrosecond tuning speeds together with wide quasi-continuous tuning range of over 60 nm have been demonstrated, for example, by utilizing gain coupler sampled reflector laser technologies [5], [6]. Using currently available 50-GHz channel spacing WDM filtering technology, our proposed architecture can support more than 150 channels, each running at 10 Gb/s, exceeding an aggregate switch throughput of 1.5 Tb/s. Power budget does not limit the switch performance in this architecture, as conventional 30-dB power margin can support up to 1000 nodes.

Fig. 1 depicts the switch control architecture, described in detail in [7]. Packet traffic received at each port is distributed to different queues within the node on a packet-by-packet basis, where each queue relates to a desired destination node. All nodes are connected to a central channel reservation scheduler via a common electronic channel reservation bus and individual control lines. The N bus lines are accessible to all nodes and indicate the reservation status of each of the N channels. The individual control lines are used by the scheduler to signal each node, in turn, to commence the channel reservation procedure.

All nodes transmit data concurrently via channels reserved during the preceding time slot. Finding the maximal priority between the contending queues may be achieved by utilizing comparison-tree hierarchy implementation. Assuming N nodes, the switching time slot period is calculated as

$$t_{\rm ts} = N \cdot \log_2(N) \cdot t_c + t_{\rm switch} \tag{1}$$

where t_c is the processing time for a single comparator and $t_{\rm switch}$ is the optical cross-connect switching time. Accordingly, the optimal switching time slot $t_{\rm ts}$ dictates the queuing time delay and hence the number of cells required to be transmitted during each time slot.

IV. Computer Simulation Results

Considering present VLSI technology, $t_c = 0.5$ ns was found feasible. As a result, extremely short processing time for resource allocation is attained, yielding high switching performance. Inclusive sets of simulations demonstrate that using the proposed architecture, Tb/s switching capacity with 100 OC-192 (9.95 Gb/s) ports can be pragmatically achieved.

Figs. 2 and 3 summarize the simulation results [8], [9], in which the mean queue occupancy (in packets) versus optical switching time is depicted. In Fig. 2, a switch with binomial packet arrival process and uniform destination distribution traffic is assumed, while in Fig. 3, the case of *nonuniform* Zipf destination distribution [10] is examined. For both models, the full three-lag tripod optimization is utilized, while assuming a fixed overhead of 10% due to optical switching time. It is assumed that both the packet transmission duration and the

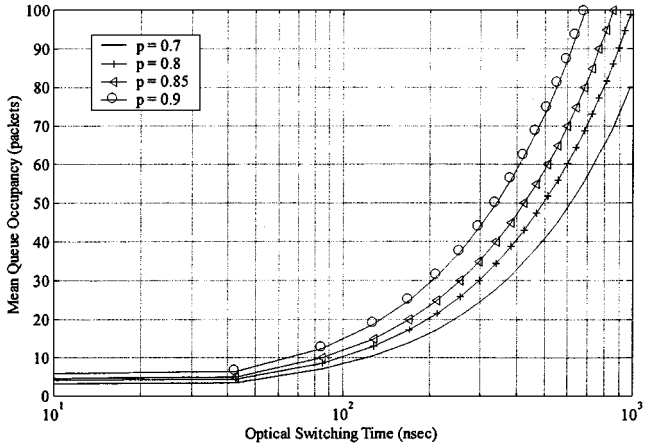

Fig. 3. Mean queue occupancy for nonuniform Zipf distributed Bernoulli traffic.

algorithm processing time (for the preceding transmission time slot) last ten times longer than the optical switching time. When the optical switching time limitation is negligible $(t_{\text{switch}} \sim 0)$, the mean queue occupancy is bounded only by the efficiency of the scheduling algorithm. As optical switching time increases (prolonging transmission "dead time"), the queues become heavily loaded, resulting in significant degradation of overall switching performance. Consequently, the generalized optimization scheme, which merges between physical-layer and algorithm-layer considerations, implies that optical switching time of 100 ns in a Tb/s switch results in typical queues occupancy of less than 20 packets for the case of 53-bytes packet length. This is equivalent to packet delay of below 80 μs. Accordingly, in the case of ATM cells' arriving at 10 Gb/s, switching granularity is on the order of 20 cells per switching event. As can be derived from Figs. 2 and 3, reducing the switching time bellow 100 ns provides limited improvement in terms of mean packet delay due to limitations of the scheduling algorithm. In other words, "clever" switching with thicker granularity performs better than nonoptimized packet-by-packet switching. It should be noted that by setting the switching time overhead to a fixed 10%, an inherent limitation of 90% throughput is obtained. Yet, 100% throughput can be achieved by using a 10% internal speed-up scheme. Also, although in this letter, we discuss dynamic-WDM based optical switching, the switching mechanism can be replaced by other multiplexing technology, e.g., optical space-division multiplexing or time-division multiplexing.

V. Conclusion

The optical dynamics requirement for packet-over-WDM networks is presented and analyzed. An advanced three-lag tripod optimization scheme that merges between physical-layer and algorithmic-layer considerations is presented, leading to a novel design for next-generation multiterabit/second routers based on optical cross-connects with submicrosecond switching speed requirements. Such optical technology is already in hand.

References

[1] F. M. Chiussi, J. G. Kneuer, and V. P. Kumar, "Low-cost scalable switching solutions for broadband networking: The ATLANTA architecture and chipset," *IEEE Commun. Mag.*, vol. 25, pp. 44–53, Dec. 1997.

[2] N. McKeown, M. Izzard, A. Mekkittikul, B. Ellersick, and M. Horowitz, "The tiny tera: A packet switch core," *IEEE Micro*, pp. 26–33, Jan./Feb. 1997.

[3] A. J. Agranat, G. Bartal, J. Krupnic, B. Pessah, and D. Sadot, "The electroholographic optical switch," in *Proc. IEEE/OSA Eur. Conf. Optical Communications (ECOC'99)*, vol. I, Sept. 1999, pp. 334–335.

[4] J. E. Ford, V. A. Aksyuk, D. J. Bishop, and J. A. Walker, "Wavelength add-drop switching using tilting micromirrors," *J. Lightwave Technol.*, vol. 17, pp. 904–911, May 1999.

[5] P. J. Rigole *et al.*, "Fast wavelength switching in a widely tunable GCSR laser using a pulse pre-distortion technique," in *IEEE OFC'97 Tech Dig.*, Feb. 1997, pp. 231–232.

[6] Altitun., Stockholm, Sweden. [Online]. Available: http:\\www.altitun.com

[7] J. Nir, I. Elhanany, and D. Sadot, "A new Tbit/sec switching scheme for ATM/WDM networks," *Electron. Lett.*, vol. 35, no. 1, pp. 30–31, Jan. 1999.

[8] D. Sadot and I. Elhanany, "Optical switching speed requirements for Terabit/sec packet over WDM networks," in *Proc. IEEE/OSA Eur. Conf. Optical Communications (ECOC'99)*, vol. I, Sept. 1999, pp. 444–445.

[9] I. Elhanany and D. Sadot, "A novel Tbit/sec switch architecture for ATM/WDM high-speed networks," presented at the IEEE/IEICE ATM Workshop'99, Japan, 1999.

[10] R. E. Wyllis, "Measuring scientific prose with rank-frequency ('Zipf') curves: A new use for an old phenomenon," in *Proc. Amer. Soc. Information Science*, vol. 12, Washington, DC, 1975, pp. 30–31.