PAPER *IEICE/IEEE Joint Special Issue on Recent Progress in ATM Technologies*

A Contention-Free Tbit/sec Packet-Switching Architecture for ATM over WDM Networks

Itamar ELHANANY[†] ***and*** **Dan SADOT**[†], ***Nonmembers***

SUMMARY Future high-speed switches and routers will be expected to support a large number of ports at high line rates carrying traffic with diverse statistical properties. Accordingly, scheduling mechanisms will be required to handle Tbit/sec aggregated capacity while providing quality of service (QoS) guarantees. In this paper a novel high-capacity switching scheme for ATM/WDM networks is presented. The proposed architecture is contention-free, scalable, easy to implement and requires no internal “speedup.” Non-uniform destination distribution and bursty cell arrivals are examined when studying the switching performance. Simulation results show that at an aggregated throughput of 1 Tbit/sec, low latency is achieved, yielding a powerful solution for high-performance packet-switch networks.

key words: *Tbit routers, scheduling, QoS (quality-of-service), global considerations*

1. Introduction

The exponential growth of Internet and datacom traffic encourages research and development of high-performance networking infrastructure consisting of both increased media transmission speeds and performance enhancement of switches and routers [1]–[5], [8]. Optical technologies, such as wavelength division multiplexing (WDM), are widely accepted as the technology for such high capacity, scalable and cost effective networks. Based on WDM technology, Tbit/sec point-to-point transmission speeds have been demonstrated. Hence, the main bottleneck of the network has shifted towards designing high-speed switches and routers, which are expected to efficiently regulate large amounts of diversely characterized traffic. Accordingly, considerable attention has been paid to WDM contention resolution and wavelength assignment issues, commonly targeted at packet-switched networks such as asynchronous transfer mode (ATM). It is generally acknowledged that the two main goals of network switches are (1) to optimally utilize the available bandwidth in order to achieve maximum throughput, and (2) to support quality of service (QoS) requirements. Constraints derived from these goals typically contradict in the sense that the maximal bandwidth provision may not necessarily correlate with the most urgent traffic flow.

Recently, several switching schemes were designed in order to support high capacity, large number of ports and low latency requirements [1], [3], [5]. Many of these scheme employ output-queuing mechanisms, which means that cells arriving at the input node are transmitted through the cross-connect fabric to a designated output queue. In order to avoid transmission contention in an $N \times N$ switch, either N^2 independent channels or N times faster circuitry are required by the switch core. Considering today’s high line rates, N times faster circuitry is becoming infeasible. Typical designs implement either centralized or output queuing mechanisms in order to maximize switch bandwidth. However, as the line rates and number of ports increase, output queuing is found impractical for high-performance switches.

An alternative to output queuing is input queuing, whereby cell buffering is located at the input nodes. It has been shown that an input-queued switch employing a single FIFO at each input, may achieve a maximum of 58.6% throughput due to the head-of-line (HOL) blocking phenomenon [6]. An alternative technique, which entirely eliminates the HOL blocking, is known as the *virtual output queuing* (VOQ). In VOQ each input node contains a separate queue for each output. Several scheduling algorithms have been proposed for VOQ switches [6], [7], [9]. As indicated by Chuang et al. [5], most maximum matching algorithms known to-date are too complex to be implemented in hardware on real-time, therefore, are found unsuitable for switches with a large number of nodes at high line rates. Moreover, the algorithms proposed are frequently examined under uniform traffic assumptions, which clearly does not represent real life traffic.

One method of enhancing VOQ based switching is to increase the “speedup” of the switch. A switch with a speedup of L can transmit L cells in a single cell-time. However, the switch-core speed is a paramount resource making speedup a drawback of any scheduling approach. In this paper, we introduce a new packet-switching architecture that offers Terabit/sec capacity along with support of over 100 ports and low switching latency. Scheduling is based on a sequential-reservation scheme with prioritized-matching that was initially introduced in [13], [14], and is broadened here to comply with diverse traffic characteristics. We show that the new architecture is contention-free, requires no speedup

Manuscript received May 28, 1999.
Manuscript revised August 27, 1999.

[†] The authors are with the Electrical and Computers Engineering Department, Ben-Gurion University, Beer-Sheva 84105, Israel.

and has the advantages of low implementation complexity and high scalability. In addition, the architecture can easily be adapted to comply with diverse quality of service (QoS) requirements.

Although this work focuses on packets having fixed length, many network protocols, such as IP, have variable length packets. Most switches and routers today segment these packets into fixed-length packet (or "cells") prior to entering the switch fabric. The original packets are reconstructed at the output stage. This methodology is commonly practiced in order to achieve high switching performance. Hence, the techniques described in this paper can be applied to both fixed and variable length packet sizes.

In Sect. 2 the proposed switch architecture is described. Section 3 focuses on the WDM perspective, channel (wavelength) allocation discipline and its VLSI implementation considerations. Traffic modeling is described in Sect. 4, while simulation results and switch metrics are presented and discussed in Sect. 5. Section 6 draws the main conclusions.

2. Switch Architecture

2.1 The WDM Perspective

In the proposed switch we focus on a tunable-transmitter fixed-receiver (TT-FR) configuration whereby a node tunes its transmitter to a predefined wavelength according to the desired destination. By the same token, a fixed-transmitter tunable-receiver (FT-TR) configuration can be applied with little changes in the proposed switching architecture. Employed scheduling schemes for Tbit/sec WDM-based switches are required to possess the following properties:

- *Maximal bandwidth utilization*: the algorithm should maximize the use of the switch internal bandwidth in order to provide maximal switching throughput.
- *Contention-resolution*: Since the wavelengths constitute an expensive resource, no more than N internal channels should be assumed. Two or more simultaneous transmissions over the same wavelength result in collision. Hence, packet losses due to optical collision should be avoided.
- *No speedup*: In the optical context, speeding up internal transmission rates is a wasteful act. Schemes that require no speedup (speedup factor of 1) are more pragmatic for scalable networks.
- *Non-blocking*, *Single-stage*: WDM cross-connect technology is commonly deployed in a single-stage, non-blocking constellation whereby optical signals are not re-routed or re-transmitted within the switching core. The ultra-fast decision processes require that minimal latency is contributed by the cross-connecting element.
- *Large number of ports*: The diversity of communication protocols and line speeds imply that future switches will have port densities of hundreds and even thousands of ports. Therefore, a good switch architecture must support a large number of asymmetric ports.

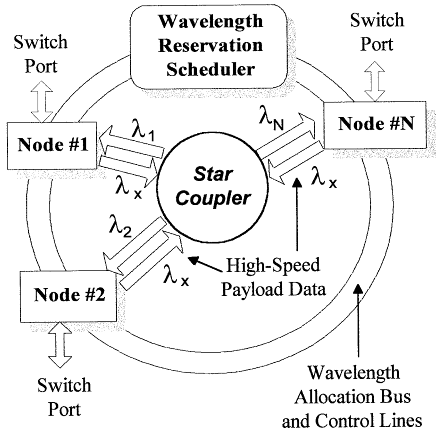

Fig. 1 Terabit/sec ATM over WDM switch architecture.

- *Simple to implement*: The switching architecture and respective algorithm should be plainly implemented in designated custom hardware. Performance in terms of speed is directly affected by the simplicity of the hardware design.

2.2 The Scheduling Algorithm

Figure 1 depicts the proposed switch architecture. The nodes, corresponding to the switch ports, have bidirectional optical data links interconnected via an optical passive star coupler. The passive star topology acts as a single-stage non-blocking cross-connect fabric. Each transmitter can be tuned to any of the N wavelengths, while each receiver is assigned a fixed and unique wavelength. ATM traffic received at each port is distributed to various buffers within the node on a cell-by-cell basis, where each buffer relates to a single wavelength according to the desired destination node.

All nodes are connected to a central wavelength reservation scheduler via a common electronic wavelength 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 wavelengths. The individual control lines are used by the scheduler to signal each node, in turn, to commence the wavelength reservation procedure.

3. Wavelength Reservation

Upon receiving a control signal from the wavelength reservation scheduler, each node performs wavelength reservation according to two major guidelines: (a) global switch resources status, i.e., available wavelengths at the reservation time, and (b) local consid-

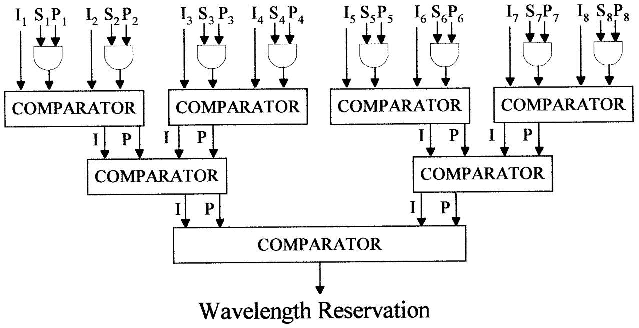

Fig. 2 Block diagram of node wavelength reservation logic for an 8 × 8 switch.

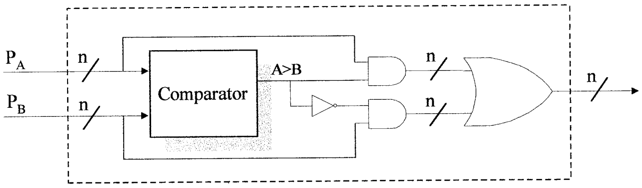

Fig. 3 Single-level comparator logic with n-bits per priority value.

erations, i.e., the status and priorities of the node's internal buffers. Figure 2 illustrates a block diagram of the wavelength reservation hardware of a single node in an 8 × 8 switch.

At the highest level, the wavelength reservation status lines, denoted by S_i, either grant or discard buffer priorities, denoted by P_i, via designated AND logic. Consequently, only buffer indices, denoted by I_i, relating to available wavelengths advance to the lower levels. Each level consists of a set of comparators, which concurrently receive as input a pair of indices, along with their respective priorities, and output the higher priority and its corresponding index. Figure 3 depicts the logic comprising each comparator unit. The output of the last comparator determines the "prevailing" buffer, which held the highest priority out of the subset of buffers relating to unreserved wavelengths. Any number of parameters, such as buffer load, accumulated delay and required QoS can affect the buffer priority metrics. Node wavelength selection is instantly followed by assertion of the relevant line within the wavelength reservation bus. At that point, utilizing a weighted Round Robin procedure, the central scheduler signals the next node to begin wavelength reservation. Contention is inherently avoided, since at any given time only one node attempts to reserve a wavelength. After all N nodes complete wavelength reservation, high-speed data is optically transmitted via the star coupler.

The wavelength reservation and data transmission are conducted in a time slot discipline. Nodes transmit data concurrently at wavelengths reserved during the previous time slot. Assuming N nodes, the time slot period can be calculated as

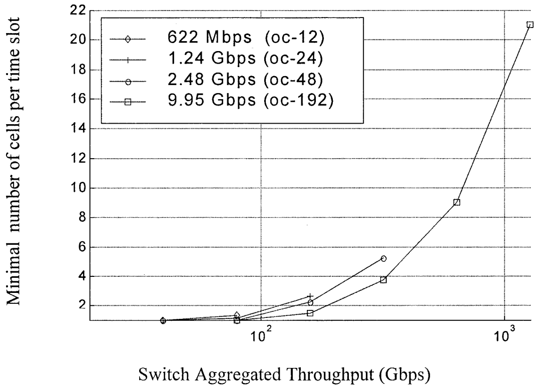

Fig. 4 Minimal number of cells per time slot vs. switch aggregated throughput.

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

where t_c is the propagation delay of a single-level comparator logic. Accordingly, t_{ts} dictates the minimal number of cells required to be transmitted during each time slot and hence the queuing time delays.

Utilizing current CMOS technology $t_c = 1\,\text{ns}$ is found feasible. Consequently, extremely short processing time for resource allocation is attained, yielding high switching performance. Figure 4 shows the relationship between the minimal number of cells to be transmitted during each time slot and the switch aggregated throughput. Various port bitrates are presented as a parameter. The marks on each curve represent standard number of ports, e.g., 8, 16, 32, 64 and 128.

4. Traffic Modeling

Next generation networks will carry diverse traffic embodying a wide range of statistical properties. Two principal criteria that strongly affect switching performance are the packet arrival statistics and destinations distribution. Typically, it is assumed that arriving packets obey a binomial process and are uniformly distributed to all destinations. Since current high-speed routers and switches are limited to an aggregated capacity of several hundred Gbit/sec, it is difficult to predict how valid will these assumptions be in future multi-Tbit/sec interconnections. It is also well known that real-life traffic tends to burstiness due to modern applications such as compressed video and multimedia services. Accordingly, packet destinations vary dynamically and hence are non-uniformly distributed.

4.1 Traffic Arriving Process

4.1.1 Binomial Arrivals

Networks that contain IP packets or ATM cells are regarded as discrete systems. At each cell time there is a probability p that a cell will arrive, consequently a probability $q = 1 - p$ that no cell arrives. According to the Binomial i.i.d. distribution model, the probability of k packets arrivals during n cell slots is given by:

$$P\{k\} = \frac{n!}{k!(n-k)!} p^k q^{n-k}, \tag{2}$$

It should be noted that for $n \to \infty$ and $p \ll 1$, the Binomial distribution converges to the continuous Poisson distribution.

4.1.2 Bursty Arrivals

It is well known that real-life traffic tends to burstiness due to the prevalence of modern multi-media applications such as compressed video and sound. Correspondingly, packet destinations vary dynamically and are non-uniformly distributed. Many models for bursty traffic have been proposed [15]–[17]. In [17] it has been claimed that network traffic is bursty at any level making bursty models an essential testbed for switching performance. We employ an on-off arrival process modulated by a two-state Markov chain. The result is a train of bursty cell arrivals, each containing cells with identical destination, followed by periods of empty cell slots. The expression for the mean offered load under such bursty traffic is:

$$\mathit{Offerd\ Load} = \frac{q}{p+q} \tag{3}$$

where p and q are the transition probabilities with regard to the active and idle periods, respectively. Due to the geometric distribution of the duration of active and idle periods, the average burst length is $1/p$. From the average burst length and offered load, an idle parameter q can easily be determined.

4.2 Destination Distribution

Real-life traffic destinations are not uniformly distributed; traffic tends to be focused on preferred or popular destinations. Unfortunately, the performance of many scheduling algorithms degrades under non-uniform traffic conditions, where not all queues are evenly and heavily loaded. Maximum matching algorithms are known to perform poorly and cause queue starvation [2] under these conditions. We introduce here a destination distribution model named Zipf's law. The Zipf law was proposed by G. K. Zipf [10]–[12]. This model may be used as a reference to investigate and compare the performance of different scheduling algorithms. The Zipf law states that frequency of occurrence of some events (P), as a function of the rank (i), where the rank is determined by the above frequency of occurrence, is a power-law function: $P_I \sim 1/i^k$, with the exponent k close to unity. The most famous example of Zipf's law is the frequency of English words in a given text. Most common is the word "the," then "of," "to" etc. When the number of occurrence is plotted as the function of the rank ($i = 1$ most common, $i = 2$ second most common, etc.), the functional form is a power-law function with exponent close to 1. Figure 5 illustrates the Zipf distribution for different values of the k parameter. It was shown that many natural and human phenomena such as Web access statistics, company size and biomolecular sequences all obey the Zipf law with k close to 1 [11]. We use the Zipf distribution to model packet destination distribution. The probability that an arriving packet is heading destination i is given by:

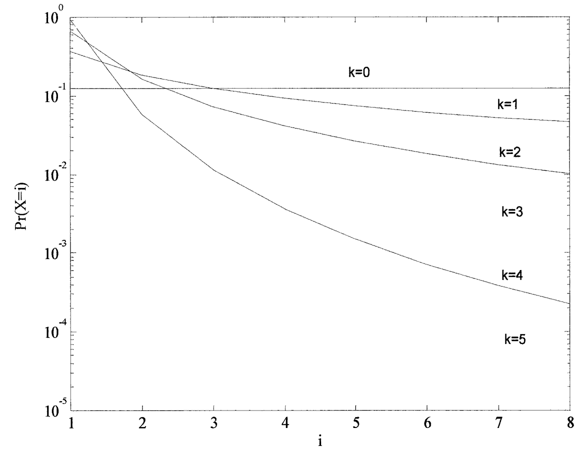

Fig. 5 The Zipf distribution function for various orders.

$$\mathit{Zipf}(i) = \frac{\frac{1}{i^k}}{\sum_{j=0}^{N} \frac{1}{j^k}} \tag{4}$$

where i is the packet destination, k is the Zipf order and N is the system order, i.e., the number of switch ports. While $k = 0$ corresponds to uniform distribution, as k increases the distribution becomes more biased towards preferred destinations. In order to generate a stable and realistic traffic model, the average steady state load at each input port must not exceed 100%. Similarly, the average steady state aggregated traffic rate arriving from all input ports to any destination port must not exceed 100%.

5. Simulation Results and Discussion

Simulations were carried out for various switch sizes (N), each assumed to receive data at a line rate of 10 Gbps (oc-192). The WDM-based switch fabric is single-staged, non-blocking with N available internal channels (wavelengths), as described earlier. It is important to note that in this case the fabric speedup is 1. Destination distributions both uniform and Zipf were examined under Bernoulli and bursty cell arrival

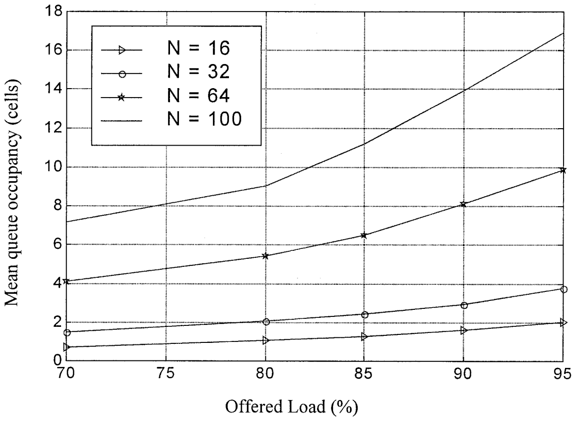

Fig. 6 Mean queue occupancy for uniformly distributed Bernoulli traffic loads.

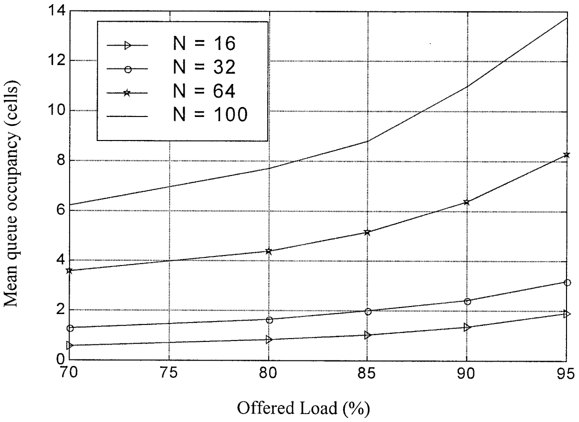

Fig. 7 Mean queue occupancy for Zipf distributed Bernoulli traffic loads.

conditions. For the Zipf distribution, $k = 1$ was selected in accordance with the claims of Sect. 4.2. Any other order of k has lead to improved switching performance results, therefore $k = 1$ can be regarded as the worst case analysis. Each simulation experiment included transmission of two million cells obeying the specific arrival process and destination distribution that was initially selected. It was observed that simulation results have converged to a steady-state value significantly prior to accomplishing the two million events, leading to the conclusion the simulation error was essentially negligible.

Figures 6 and 7 depict, respectively, the mean queue occupancy (MQO) for uniformly and Zipf distributed cells, with Bernoulli arrival characteristics, as a function of offer traffic load. As can be seen, regardless of the destination distribution behavior, the resulting queue occupancies amount to several cell times. A switch of 100 × 100 ($N = 100$) under 90% traffic load, has an MQO of less than 20 cells. The metrics are mildly affected by the traffic destination characteristics, as can be derived from Fig. 7.

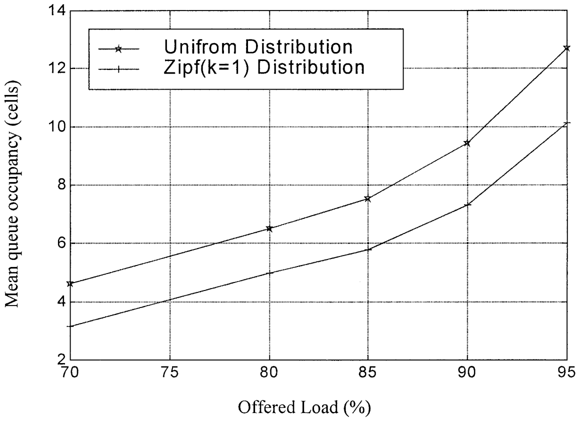

Fig. 8 Mean queue occupancy for Zipf and uniformly distributed bursty traffic loads (average burst length is 10 cells) in a 100 × 100 switch.

Figure 8 depicts the MQO for bursty traffic (mean burst length is 10 cells) arriving at a 100 × 100 switch, under both distribution conditions as a function of the offered load. The results show that bursty traffic, which typically degrades switching performance, has no degrading impact. On the contrary, due to the time slot discipline applied by the architecture, bursty traffic yields better results. The MQO directly relates, and hence may be translated, to the mean cell delay (MCD). Under non-uniform destination distribution conditions, translation to the MCD should refer to a specific queue/traffic channel.

Our simulations can be compared to other high-speed buffer management approaches such as in [1] and [2]. In such architectures, various types of real-time Round Robin algorithms are employed and are typically limited in performance under non-uniform and bursty traffic. In [2], for example, a 16 × 16 switch under 95% offered load of uniform Bernoulli arrivals yields an average latency of approximately 100 cells, as compared to a latency of 2 cells under the same conditions using our proposed scheme. The fact that no Round Robin pointers are employed in our scheme results in low variance of the service-time distribution and hence the lower mean cell delays.

Although in this paper we discuss WDM as the physical layer switching fabric and fixed-size packets, the proposed method can be extended to other multiplexing technology, e.g., optical SDM (space division multiplexing), and to variable-size packets, e.g., IP traffic.

6. Conclusions

A novel Tbit/sec switch architecture for ATM over WDM packet-switched networks has been proposed. By applying an ultra-high speed scheduling discipline, extremely high capacity and low latency switching is achieved for fabrics of up to 100 × 100. Simulation results demonstrate that bounded queue lengths and low latency can be achieved utilizing the proposed scheme, even when applying complex traffic models with non-uniform destination distributions and bursty cell arrivals. Using current CMOS technology, Tbit/sec switching throughput can be attained with delay in the range of tens of microseconds. The design is contention-free, simple to implement, scalable and can easily be adapted to various multiplexing technologies.

Acknowledgement

This work was supported by the Israeli Ministry of Science under contract no. 9404-1-97, the Wolfson foundation and the Rector's office at the Ben-Gurion University of the Negev, Israel.

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, no.12, pp.44–53, Dec. 1997.

[2] N. McKeown, "The iSLIP scheduling algorithm for input-queued switched," IEEE/ACM Trans. Networking, vol.7, no.2, pp.188–201, April 1999.

[3] V. Sivaraman and G.N. Rouskas, "HiPeR-1: A high performance reservation protocol with look-ahead for broadcast WDM networks," Proc. IEEE INFOCOM'97, vol.3, pp.1270–1277, April 1997.

[4] C. Salisbury and R. Melhem, "A high speed scheduler/controller for unbuffered Banyan networks," Proc. IEEE ICC'98, Session S18-6, June 1998.

[5] S. Chuang, A. Goel, N. McKeown, and B. Prabhakar, "Matching output queuing with a combined input output queued switch," Proc. IEEE INFOCOM'99, March 1999.

[6] M. Karol, M. Hluchyj, and S. Morgan, "Input versus output queuing on a space division switch," IEEE Trans. Commun., vol.12, no.35, pp.1347–1356, 1987.

[7] A. Mekkittikul and N. McKeown, "A practical scheduling algorithm to achieve 100% throughput in input-queued switches," Proc. IEEE INFOCOM'98.

[8] H. Obara, "Optimum architecture for input queueing ATM switches," IEE Electronic Letters, pp.555–557, March 1991.

[9] N. McKeown, V. Anantharam, and J. Walrand, "Achieving 100% throughput in an input-queued switch," Proc. IEEE INFOCOM'96, San Francisco, March 1996.

[10] G.K. Zipf, Psycho-Biology of languages, MIT Press, Houghton-Miffin, 1965.

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

[12] B.M. Hill, "A simple general approach to inference about the tail of a distribution," Anals. of Statistics, vol.3, pp.1163–1174, 1975.

[13] 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.

[14] I. Elhanany and D. Sadot, "A novel Tbit/sec switch architecture for ATM/WDM high-speed networks," Proc. IEEE ATM Workshop'99, Kochi City, Japan, 1999.

[15] H. Heffes and D.M. Lucantoni, "A Markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance," IEEE J. Sel. Areas Commun., vol.4, pp.856–868, 1988.

[16] R. Jain and S.A. Routhier, "Packet trains: Measurements and a new model for computer network traffic," IEEE J. Sel. Areas Commun., vol.4, pp.986–995, April 1986.

[17] W.E. Leland, W. Willinger, M. Taqqu, and D. Wilson, "On the self-similar nature of ethernet traffic," Proc. SIGCOMM, pp.183–193, San Francisco, CA, Sept. 1993.

Itamar Elhanany received his B.Sc. and M.Sc. degrees both in Electrical & Computers Engineering from the Ben-Gurion University in Israel in 1995 and 1998, respectively. He is currently pursuing his Ph.D. degree at the Optical Communications Lab in the field of Switching-Management Algorithms for High-Speed Network. His fields of interest are scheduling algorithms, switching architectures, Quality-of-Service and WDM networks.

Dan Sadot received his B.Sc., M.Sc., and Ph.D. (Summa Cum Laude) from the Ben Gurion University of the Negev, Beer Sheva, Israel, all in Electrical Engineering, in 1988, 1990, and 1994, respectively. During 1994–1995, he was a Post-Doctorate associate in the Optical Communication Research Laboratory at the Department of Electrical Engineering of Stanford University. His Ph.D. studies were supported by the Clore scholarship. His post-doctorate was supported by both the Fulbright and the Rothchild scholarships. In October 1995, Dr. Sadot joined the Ben Gurion University as a senior lecturer in the Electrical and Computer Engineering Department, where he has founded the Optical Communications Laboratory. Currently, his main activities include development of a Tbit/sec optical WDM/ATM network, ultra-fast tunable fiber-loop filters and lasers, information security within optical fiber, and optical CDMA. Dr. Sadot is the founder and chair of the IEEE/LEOS chapter in Israel.