A Prioritized Packet Scheduling Architecture for Provision of Quality-of-Service in Tbit/sec WDM Networks

I. Elhanany, D. Sadot
Electrical and Computers Engineering Department
Ben-Gurion University
Beer-Sheva 84105, ISRAEL
itamar@ieee.org

Abstract-**The provision of Quality-of-Service (QoS) is a predominant requirement from modern high-performance switches and routers. Port densities of tens and even hundreds of ports characterize ATM and IPv6 switches where each port carries high-speed traffic with diverse statistical characteristics. In this paper a scheduling scheme for Tbit/sec, input-queued, WDM packet-switch networks is presented. The proposed architecture is contention-free, scalable, easy to implement and requires no internal "speedup". Moreover, the scheme inherently supports per-class QoS. Non-uniform destination distribution and bursty cell arrivals are studied for a switch with an aggregated capacity of 1 Tbit/sec. Simulation results show that class-differentiated low latency is achieved, yielding a powerful solution for high-performance packet-switch networks.**

I. INTRODUCTION

High-speed cell-switching devices are employed as building blocks for asynchronous transfer mode (ATM) and Internet Protocol (IP) routers and switches. As Internet traffic volume increases at an exponential rate, the search for high-performance and scalable packet-switching technologies is broadening. Traffic traversing the Internet is not only increasing in volume but also becoming more demanding in terms of delay and packet loss requirements. Multimedia applications, such as voice and video, which are growing more prevalent, demand strict delay boundaries. Accordingly, switches and routers support the provision of Quality of Service (QoS) as means of offering such differentiated services.

Broadband network infrastructures are coarsely composed of two basic building blocks: (1) high-speed point-to-point links and (2) high-performance switching devices for the network's nodes. Wavelength division multiplexing (WDM) is widely accepted as the physical-layer solutions for future broadband networks. While reliable WDM-based Tbit/sec point-to-point communication has been demonstrated, switches and routers that can manage the extensive amounts of diversely characterized traffic loads are not yet available. Hence, the bottleneck of the network has shifted towards designing such high-performance switches and routers.

It is generally acknowledged that the two main goals of network switches are 1) to optimally utilize the available internal bandwidth while 2) support quality of service (QoS) requirements. Constraints derived from these goals typically contradict in the sense that maximal bandwidth allocation may not necessarily mutually relate to giving preference to the most urgent traffic flows. This concept has spawn a vast range of adaptations, each seeking to support high capacity, large number of ports and low latency requirements [1],[2],[3],[6].

Many of these schemes employ output-queueing 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 overcome collision in an *NxN* switch, either N^2 independent channels or *N* times faster circuitry may be implemented. Considering today's high line rates, *N* times faster circuitry is infeasible. Typical designs apply either centralized or output queueing mechanisms in order to maximize switch bandwidth. However, as the line rates and number of ports increase, output queueing is found impractical for high-performance switches.

An alternative to output-queueing is input-queueing whereby cell buffering is located at the input nodes. It has been shown that an input-queued switch with a single FIFO at each input may achieve a maximum of 58.6% throughput due to the head-of-line (HOL) blocking phenomenon [4]. A well-practiced technique, which entirely eliminates the HOL blocking, is *virtual output queueing* (VOQ). In VOQ each input node maintains a separate queue for each output as illustrated in figure 1.

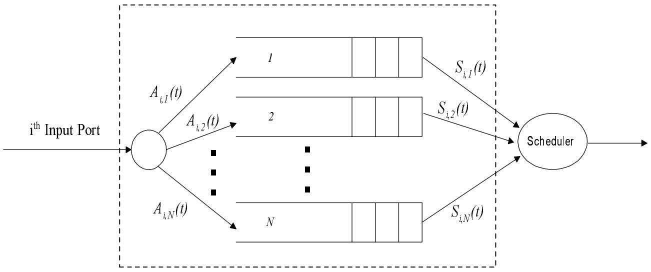


Fig. 1. Virtual output queueing: cells are classified to queues according to their destination.

Arriving cells, $A_{i,j}(t)$, are classified at a primal stage to a queue that corresponds to their designated destination. The scheduler determines which queue is served for transmission ($S_{i,j}(t)$). Several scheduling algorithms have been proposed for VOQ switches [5]. As indicated by Chuang *et al.* [3], most algorithms known to-date are too complex to be implemented in hardware and are found unsuitable for switches with large number of nodes at high line rates. Moreover, the algorithms proposed are frequently examined under uniform traffic conditions, which clearly does not represent real life traffic. One method of enhancing VOQ based switching is to increase the internal "speedup" of the switch. A switch with a speedup of L can transmit L cells in a single cell-time. However, the switching-core speed is a paramount resource making speedup a drawback of any scheduling approach. In order to support QoS, VOQ may be expanded by assigning r different queues for each destination, whereby a distinct queue is assigned to each QoS class. Contention for transmission is now carried out not only between queues associated with the same destination, but also between QoS class-queues designated for the same destination.

Although this work focuses on packets of fixed length, many network protocols, such as IP, have variable length packets. Most switching engines 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 performance. Accordingly, the methods described here may apply to both fixed and variable length packets. In this paper, a packet-switching architecture along with a scheduling algorithm for high-performance WDM networks is presented. We show that the architecture is contention-free, requires no speedup and has the advantages of high-throughput, low implementation complexity and scalability. Scheduling is based on a sequential-reservation scheme with prioritized-matching that was initially introduced in [10], and broadened to comply with diverse QoS requirements.

In section 2 the proposed switch architecture is described. Section 3 focuses on the wavelength allocation discipline and its implementation considerations. The investigated traffic models are described in Section 4. Simulation results are presented and discussed in Sections 5. Section 6 draws the main conclusions.

II. Scheduling discipline

A. 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 may 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 are a most 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.
- *QoS support*: Differentiating between traffic flows is a substantial requirement in order to provide time-critical services in modern networks. Fairness is directly derived from this criterion and exists to assure that queue starvation is avoided. An evolved mechanism for supporting such differentiation and fairness should be practiced.
- *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 future 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, such as Asynchronous Transfer Mode (ATM) and Internet Protocol (IP), and of line speeds, require that a good switch architecture will support asymmetric ports carrying different data flows.
- *Flexible high-level manageability* – Decisions regarding distributing available bandwidth should be left to the service provider as a business decision, requiring highly evolved tools for dynamically managing the nodes.
- *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.

A. Switch Architecture

The structure of the proposed switch is shown in figure 2. The nodes, corresponding to the switch ports, have bi-directional optical data links interconnected via an optical passive star coupler. The passive star topology acts as a single-stage non-blocking cross-connect fabric. The optical transmission is based on a TT-FR setup, whereby each transmitter can be tuned to any of the N wavelengths, while each receiver is assigned a distinct and fixed wavelength. Utilizing virtual output queueing with differentiated classes, data packets received at each port are distributed to designated queues within the input node on a cell-by-cell basis, where each queue corresponds to a wavelength and a QoS class, in accordance with the cell header information.

Two electronic bus systems are employed in the proposed architecture. All nodes have read access to a node signaling

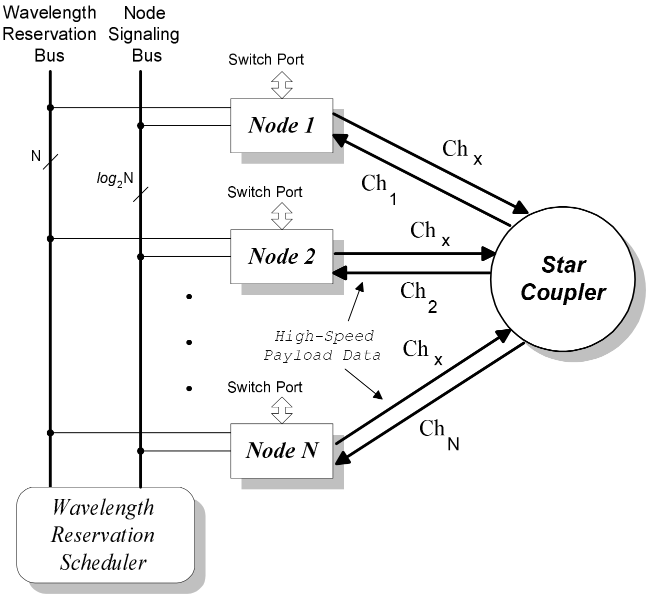


Fig. 2. WDM based Input-Queued Switch Architecture

bus (NSB) and read/write access to a common wavelength reservation bus (WRB). The WRB contains *N* lines that indicate the reservation status of each of the *N* available wavelengths. A logical '1' on a bus line denotes an available wavelength, while logical '0' denotes a reserved one. The NSB consists of a binary word of length *log (*N) representing the index of the addressed node. When a node is addressed it executes a wavelength reservation procedure intended to allocate an available wavelength.

III. Wavelength reservation

In order to properly differentiate between the state of various queues and QoS classes a prioritization system must be practiced. Many scheduling schemes, which apply binary request matrices and not prioritized queueing, are limited by their ability to provide global channel allocation fairness. It has previously been shown that the importance of prioritization is paramount in the QoS context [6]. In figure 3, a block diagram of the wavelength reservation logic performed by each node in a 4×4 switch with 2 QoS classes is depicted. $P_{i,j}$ denotes queue priority related to the j^{th} QoS class of the i^{th} desired wavelength (destination), and S_i is the WRB line for the i^{th} wavelength. Upon receiving a signal via the NSB from the wavelength reservation scheduler, each node performs wavelength reservation according to two primary guidelines: (a) global switch resources status, i.e., available wavelengths at the reservation instance, and (b) local considerations, i.e., the status and priorities of the node's internal queues.

At the highest level, the WRB lines either grant or discard queue priorities using designated AND logic. Consequently, only queue indices and priorities relating to available wavelengths advance to the lower levels. Each level consists of a row of comparators that concurrently receive as input a pair of priorities, along with their respective indices, and output the higher priority and corresponding index. The output of the bottommost comparator determines the "prevailing" queue, i.e. the queue that held the highest priority out of the subset of queues relating to unreserved wavelengths. Node wavelength selection is instantly followed by assertion of the relevant WRB line. Any number of parameters, such as queue load, accumulated delay and required QoS can affect the queueing priority metrics. Moreover, contention for wavelengths is carried out not only between queues relating to different destinations but also between queues of different QoS classes designated for the same destination. In this way, both fair contention as well as QoS class differentiation are guaranteed. At that point, utilizing a weighted Round Robin procedure, the central scheduler signals the next node to commence wavelength reservation. It should be noted that the Round Robin signaling is employed not as a core scheduling discipline but rather to guarantee fairness and quality differentiation between nodes.

In the proposed scheme 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. While nodes transmit their data, wavelength reservation is carried out for the succeeding timeslot, therefore no transmission dead time is introduced.

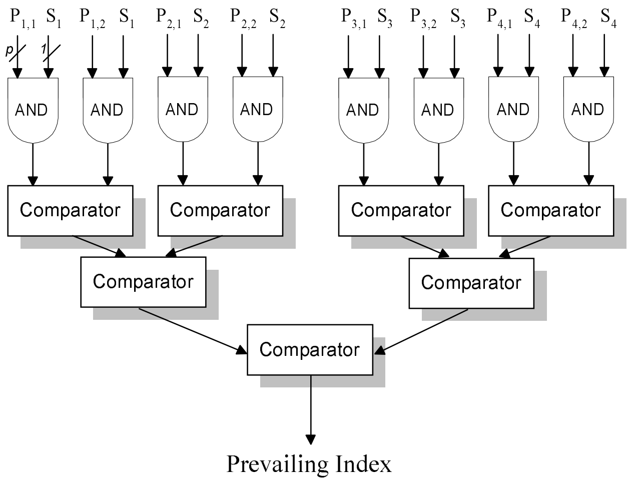
Assuming *N* destinations and *r* QoS classes, the time slot duration can be expressed as:

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

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 queueing time delays. For a 64 priority-levels coding, t_c=1 *ns* is attainable using current CMOS technology. Consequently, an extremely short processing time for resource allocation is attained, yielding high switching performance. Figure 4 shows the minimal number of cells required to be transmitted during each time slot for various number of QoS classes, versus switch aggregated throughput. The line speed is assumed to be 10 Gbps (oc-192). By determining the nodes that are participating in a given time slot, along with their preference, flexible network-level manageability is attained.

IV. Traffic model

Next generation networks will carry diverse traffic embodying a wide range of statistical properties. Two principal criteria that strongly affect switching performance



P1,1 S1 P1,2 S1 P2,1 S2 P2,2 S2 P3,1 S3 P3,2 S3 P4,1 S4 P4,2 S4
p
1
AND
Comparator
Prevailing Index

Fig. 3. Block diagram of node wavelength reservation logic for a 4×4 switch with 2 QoS Classes

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 two hundred Gbit/sec, it is difficult to predict how valid will these assumptions be in future multi-Tbit/sec interconnections. The simplest traffic load model consists of a Bernoulli i.i.d. arrival pattern whereby cell destinations are uniformly distributed. According to the Binomial distribution model, the probability of k packets arriving during n cell slots is given by:

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

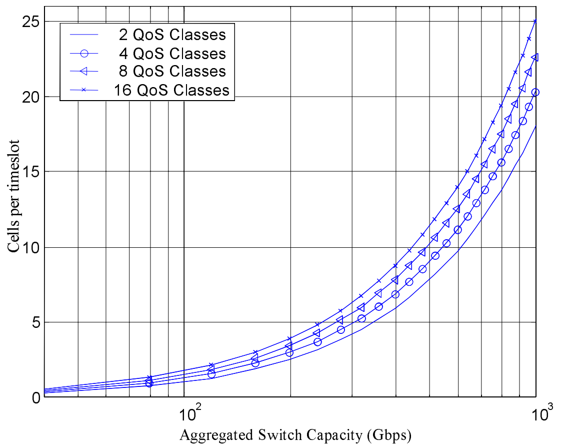


Fig. 4. Transmission time slot duration (in oc-192 cells) for various numbers of QoS classes as a function of the switch aggregated throughput

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. Accordingly, packet destinations vary dynamically and are non-uniformly distributed. Many models for modeling bursty traffic have been proposed. 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:

$$Offerd\ Load = \frac{q}{p+q} \qquad (3)$$

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 mean burst length is *1/p*.

In addition to the bursty arrival characteristics, 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 under these conditions. We introduce here a destination distribution model named Zipf's law. The Zipf law was proposed by G. K. Zipf [7]-[9]. 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. 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. We use the Zipf distribution to model packet destination distribution. The probability that an arriving packet is heading destination i is given by:

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

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 represents 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 flows arriving from all input ports to any destination port must not exceed 100%.

V. SIMULATION RESULTS

Simulations were carried out for a 100×100 switch with several QoS classes, where each port receives data, in ATM cells, at a line rate of 10 Gbps (oc-192). The WDM-based switch fabric is assumed to be single-staged, non-blocking with *N* available wavelengths. It is important to note that in this case the fabric speedup is 1, i.e. the internal transmission speeds are also 10 Gbps per port. Destination distributions of both uniform and Zipf were examined under Bernoulli and bursty cell arrival conditions. For the Zipf distribution, *k*=1 was found to represent best real traffic behavior. The priority function is defined as:

$$P_{i,j} = QO(i,j) \times j \qquad i = 1 \ldots N,\ j = 1,2,3 \qquad (5)$$

where *QO(i,j)* denotes the cell occupancy of the j^{th} QoS-class queue relating to the i^{th} destination.

Figure 5 shows the mean queue occupancy under Bernoulli i.i.d. cell arrival with uniform destination distribution for port densities of 100×100, 64×64 and 32×32 with a single class (i.e. no QoS differentiation) as a function of the offered load. The curves exhibit a logarithmic pattern, i.e. the increase in the queue length between 32 and 64 ports is logarithmically proportional to the difference between 64 and 100 ports.

Figures 6 (a) and (b) depict, respectively, the mean queue occupancy (MQO) for uniformly and Zipf distributed cells with Bernoulli i.i.d. arrival for with QoS classes as a function of the offered traffic load. The switch size is 100×100. A clear difference in the mean delay is observed between the four classes with lower delays attributed to queues with higher priorities.

It can be seen that a slight difference exists between the two destination distributions. Since the priority is defined as the queue length, the Zipf distribution, which favors explicit queues, offers more correlative queue servicing and hence slightly lower mean queue occupancies.

Figures 7 (a) and (b) depict, respectively, the MQO for bursty traffic arrival (mean burst length is 10 cells), under uniformly and Zipf distributed cells in a 100×100 switch, as a function of the offered traffic load. The results show that under bursty traffic conditions, the MQO increases by 20%-30% with respect to uniformly distributed traffic. In addition, it is noted that the higher priority classes undergo less degradation due to the non-uniformity of the traffic distribution while the lower classes are more affected. The latter is an imperative property in the context of supporting delay sensitive traffic.

The MQO directly relates, and hence may be translated, to the mean cell delay (MCD). Under non-uniform distributed traffic, translation to the MCD should refer to a specific queue/traffic channel. Our simulations are easily compared with other high-speed buffer management

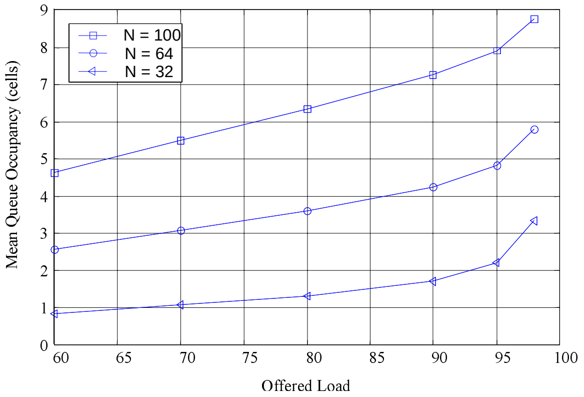
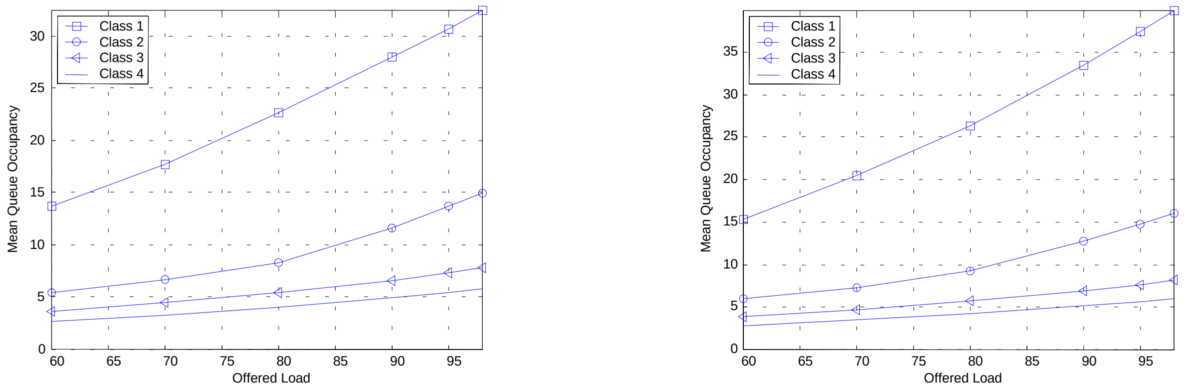


Fig. 5. Mean queue occupancy with no QoS differentiation for 100, 64 and 32 ports switches as a function of load.



(a)

(b)

Fig. 6. Mean queue occupancy for a 100×100 switch with Bernoulli arrivals and 4 QoS classes for (a) Uniform distribution and (b) Zipf distribution as a function of the offered load

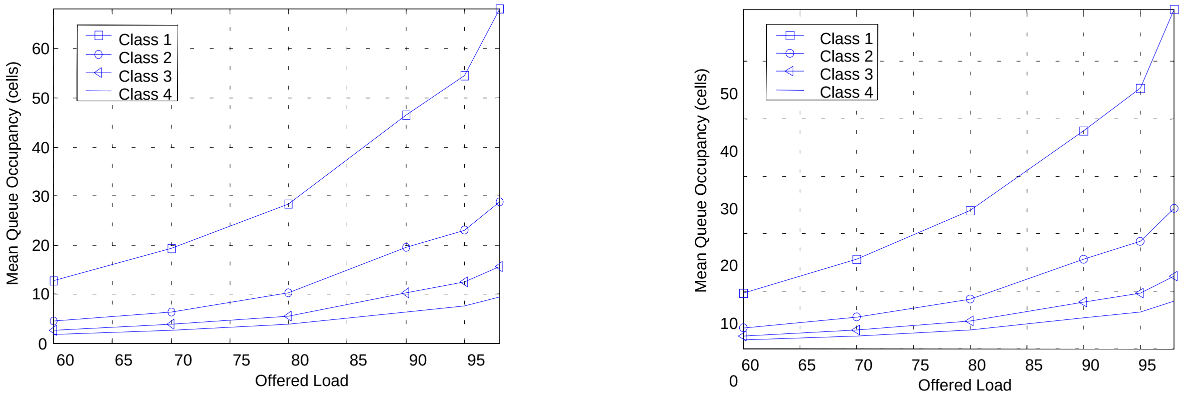


Fig. 7. Mean queue occupancy for a 100×100 switch with bursty arrivals and 4 QoS classes for (a) Uniform distribution and (b) Zipf distribution as a function of offered load

approaches such as in [1] and [2]. In such architectures, various types of real-time Round Robin algorithms are employed. These algorithms are limited in performance under non-uniform and bursty traffic due to a pointer synchronization phenomenon.

Furthermore, contention is typically avoided using backpressure mechanisms or by locally updating input and output queue-scheduling pointers leading to sub-optimal performance and scalability.Although in this work we discuss WDM as the physical layer switching fabric, the proposed method can be extended to other multiplexing technology, e.g., optical space division multiplexing (SDM).

I. CONCLUSIONS

A contention-free Tbit/sec switch architecture and scheduling algorithm for guaranteed QoS in WDM packet-switched networks has been proposed. By applying an ultra-high speed scheduling discipline, together with the deployment of Virtual Output Queueing, extremely high throughput and low latency switching is achieved for fabrics of up to 100×100 (1 Tbit/sec). Simulation results demonstrate that bounded queue lengths and low latency are attained utilizing the proposed scheme, even when applying complex traffic models with non-uniform destination distributions and bursty cell arrivals. QoS class differentiation has been demonstrated with little affect due to traffic characteristics. The design is simple to implement, scalable and can easily be adapted to various multiplexing technologies.

ACKNOWLEDGMENT

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 Communications Magazine*, vol. 25, no. 12, 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*, vol. 17, pp. 27-40, Feb. 1997.

[3] Chuang, A. Goel, N. McKeown and B. Prabhakar, "Matching Output Queueing with a Combined Input Output Queued Switch," *Proc. IEEE INFOCOM '99, March 1999.*

[4] M. Karol, M. Hluchyj and S. Morgan, "Input Versus Output Queueing on a Space Division Switch," *IEEE Trans. Communications*, No. 35, Vol. 12, pp. 1347-1356, 1987.

[5]A. Mekkittikul and N. McKeown, "A Practical Scheduling Algorithm to Achieve 100% Throughput in Input-Queued Switches," *Proc. IEEE INFOCOM '98.*

[6] R. Schoenen, G. Post, G. Sander, "Prioritized Arbitration for Input-Queued Switches with 100% throughput," *Proc. IEEE ATM Workshop 1999*, pp. 253-258.

[7] G. K. Zipf, *Psycho-Biology of Languages,* Houghton-Miffin, MIT press, 1965.

[8] R. E. Wyllis, "Measuring Scientific Prose with Rank-Frequency ('Zipf') Curves: A New Use for an Old Phenomenon," *Proceedings of the American Society for Information Science* **12**, pp. 30-31, Washington, DC: 1975.

[9] B. M. Hill, "A Simple General Approach to Inference About the Tail of a Distribution," *Anals. of Statistics*, **3**, pp. 1163-1174, 1975.

[10] J. Nir, I. Elhanany and D. Sadot, "Tbit/s Switching Scheme for ATM/WDM Networks," *Electronics Letters*, Vol. 35, No. 1, 7th Jan. 1999.