Performance Analysis of a Robust Scheduling Algorithm for Scalable Input-Queued Switches

Itamar Elhanany†, Dan Sadot
Department of Electrical and Computer Engineering
Ben-Gurion University, Israel
†Email: itamar@ieee.org

Abstract— **In this paper a high-performance, robust and scalable scheduling algorithm for input-queued switches, called distributed sequential allocation (DISA), is presented and analyzed. Contrary to pointer-based arbitration schemes, the algorithm proposed is based on a synchronized channel reservation cycle whereby each input selects a designated output, considering both its local requests as well as global channel availability information. The distinctiveness of the algorithm is in its ability to offer high-performance when multiple cells are transmitted within each switching intervals. The subsequent relaxed switching-time requirement allows for the utilization of commercially available crosspoint switches, yielding a pragmatic and scalable solution for high port-density switching platforms. The efficiency of the algorithms and its robustness is established through analysis and computer simulations addressing diverse traffic scenarios.**

Keywords- Packet scheduling algorithms, Switch fabric, Input-queued switches, Non-uniform destination distribution

I. INTRODUCTION

Scalable packet scheduling algorithms that offer high-performance for input buffered switches and routers have been the focus of many academic and industry studies in recent years. The ever growing need for additional bandwidth and more sophisticated service provisioning in next generation networks requires the introduction of scalable packet scheduling solutions that go beyond legacy schemes. As port densities and line rates increase, input buffered switch architectures are acknowledged as a pragmatic approach for implementing scalable switches and routers. In these architectures, arriving packets or cells are stored at queues at the ingress port until scheduled for transmission. With the introduction of commercially available high port-density crosspoint switches ([1], [2]), it has become more compelling to propose packet scheduling techniques that can control crosspoint switches to offer a low chip-count, high-performance and cost-efficient switch fabric solutions.

The majority of the proposed scheduling algorithms targeting high port density switches are based on an iterative request-accept-grant process [3], [4], [5]. Although such algorithms offer ease of implementation, they typically experience degradation in performance at high loads in cases where the traffic is correlated or non-uniformly distributed between the destinations. As means of overcoming the inherent limitations of pointer-based schemes, speedup is usually introduced such that the internal fabric bandwidth is S times higher than that of incoming traffic, where if N is the number of ports then $1<S<N$. A speedup of N yields performance equivalent to an output queued switch. However, speeding up the fabric carries enormous implementation ramifications that motivate the search for switching architectures and scheduling algorithms that would overcome the need for a large speedup. An additional drawback of many pointer-based schemes is the connectivity complexity which is known to be $O(N^2)$. Consequently, these algorithms have been proposed primarily for switches with small number of ports (i.e. $N\leq 32$) [3], and are generally unsuitable for next-generation switches where hundreds and even thousands of ports are considered.

In this paper we proposed a scheduling algorithm called distributed sequential allocation (DISA) that represents a shift from legacy scheduling schemes, by constituting a non-pointer based approach in which the contention resolution process takes into consideration both local and global resource availability [6]. Moreover, the DISA algorithm requires only $O(NlogN)$ connectivity and no speedup.

This paper is organized as follows. Section II describes the DISA scheduling algorithm and a complementing switch architecture. Section III outlines the queueing notation and model formulations utilized throughout the paper for performance analysis. Section IV presents analysis for Bernoulli i.i.d. uniformly distributed arrival patterns, while sections V focuses on analysis of non-uniform destination distribution. In section VI performance results under correlated arrivals are shown, while in section VII the conclusions are drawn.

II. THE DISA SCHEDULING ALGORITHM

A. Switch Architecture

Figure 1 shows a diagram of the proposed switch fabric architecture. Packets received from the switch port at each line card flow into a network processor or traffic manager. The processed packets are subsequently forwarded to an ingress fabric interface device which typically segments the packets into smaller fixed-size cells and provides a buffering stage to the fabric. In order to avoid head-of-line blocking [7], a queueing technique called virtual output queueing (VOQ) is deployed whereby a unique queue is maintained at the ingress for each of the outputs. We interchange the terns channel and output in the context of allocation by an input port.

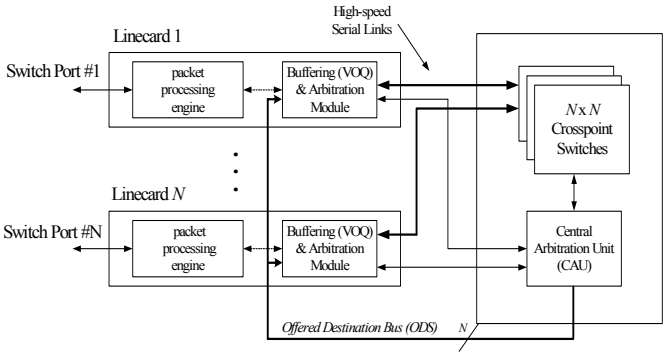

Figure 1. The proposed crosspoint-based, single-stage switch fabric architecture.

The efficiency of the scheduler has a paramount impact on the amount of buffering required at the linecards and, consequently, the latency through the system. Two types of control links are utilized in the proposed architecture between the nodes and the fabric. The first is an N-bit bus called the *offered destination set* (ODS), which all nodes have read and write access to, indicating the reservation status of each of the N outputs. We denote a logical '1' on the ODS as representing an available output port while a logical '0' a previously reserved one. In addition to the ODS, each node receives a dedicated signal from a *central arbitration unit* (CAU) which indicates when the node is to perform reservation, as will be clarified in the following section. The switching interval is a fixed multiple of the time slot duration, where a time slot represents a single cell time.

B. Scheduling Algorithm

The proposed scheduling algorithm is carried out as sequential selection process. For each switching interval, which may span over one or more time slots, a single output is allocated by each of the ports. Using the dedicated links between the CAU and the ports, the CAU signals each port, in turn, to perform allocation. A random order of signaling the ports is drawn at the beginning of each interval. Once the last port completes allocation, the crosspoint switches are configured and each port transmits cells to its designated output in accordance with the selected configuration while, concurrently, a new allocations interval (for the following transmission cycle) begins. Since transmission and scheduling are carried out simultaneously, there is no transmission "dead-time" involved,.

In figure 2, a block diagram of the channel allocation scheme performed by each node is depicted. We let w_j denote the weight for queue j within the VOQ, and φ_i the i^{th} bit element in the ODS indicating the availability status of output i. Upon receiving a signal from the CAU, each node performs output reservation according to two primary guidelines: (a) global switch resources status, i.e., available outputs and (b) local considerations reflected by the priorities map of the node's queues. We note that as will be elaborated in subsequent sections, the term queue weight may refer either to a single bit denoting queue occupancy state (empty or non-empty), or a value corresponding to the queue size.

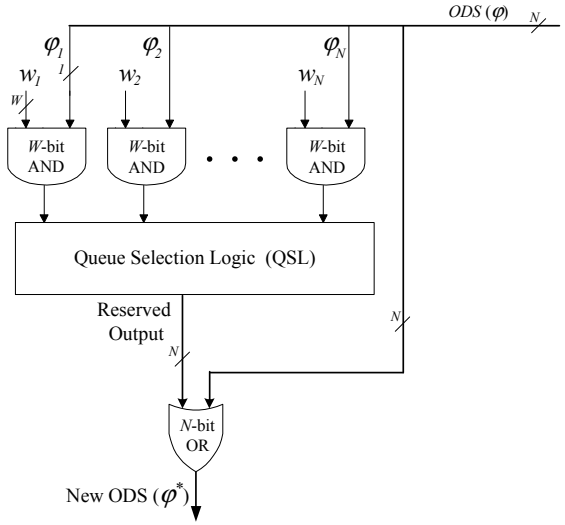

Figure 2. The output allocation function performed by each ingress port.

At the initial step, the ODS lines either grant or discard queue requests using a layer of W-bit AND gates, where W is the number of bits per weight. As a result, only weights corresponding to available outputs advance to the next level. The following step is comprised of a queue selection logic (QSL) block which determines the output to be reserved. In the single-bit-weight case, this level may be implemented using a randomized priority decoder circuit or if queue sizes are considered as weights then a maximal value circuit should be deployed. The output of the QSL is an N-bit vector with at most a single bit set to '1' denoting the index of the selected output. This word is inserted into a N-bit OR gate together with φ to yield φ^* - the new ODS. In typical staged designs, the propagation delay of such QSL circuits is of $O(\log_d N \cdot t_c)$, where d is the number of comparisons performed in each level and t_c is the propagation delay contributed by each comparison block. Using current 0.13μ CMOS VLSI technology, for 128 ports with 2 comparisons per level and t_c~500psec, the critical path duration is approximately 3.5nsec.

Upon completion of allocation by the first input node, the CAU randomly signals the next node to commence output allocation. That node will perform the same allocation process only with the new ODS as a mask of previously allocated outputs. Contention is inherently avoided since at any given time only one node attempts to reserve an output. After all N nodes perform reservation, the CAU configures the crosspoint switches and data traverses the fabric. When multiple classes of service are deployed, a different queue (and hence a unique weight) is associated with each class of service. A basic precondition for supporting the analysis is that the algorithms is stable. To that end, we have the following theorem addressing the issue of stability:

Theorem 1: (Stability) The DISA scheduling algorithm with no speedup is strictly stable for any admissible, uniformly distributed traffic.

The reader may refer to appendix A for a detailed proof.

III. QUEUEING MODEL FORMULATION

We begin our analysis by establishing the notation and queueing formulation framework, which will be used to derive the performance metrics. We assume a discrete-time VOQ system with a single-server and infinite buffer capacity. The number of queues, *N*, corresponds to the number of distinct destinations in the system. All events occur at discrete time slot intervals in which at most a single arrival and a single service event may occur. Let $\lambda_{ij}(n)$ denote the probability of arrival to virtual output queue *j* in port *i* at time step *n*. We further label λ_{ij} as the corresponding average arrival rate. In order to guarantee admissibility of the traffic, we require that $\sum_i \lambda_{ij} \le 1, \quad \sum_j \lambda_{ij} \le 1$. For convenience, we employ the early-arrival model whereby an arrival will precede a service event within any given time slot. As will be explicitly shown in the following sections, the service discipline governing each VOQ system (ingress port) can be approximated by a Bernoulli i.i.d. process, resulting in geometrically distributed interservice times. In the context of the proposed architecture, the task of the scheduler is to determine which of the virtual output queues in each port is to be granted transmission during the next switching interval.

Let μ denote the probability of service to the VOQ during each interval. Once a service event occurs, the internal arbitration scheme determines which of the queues is granted transmission. Let $Q_{ij}(n)$ designate the queue occupancy of queue *j* in port *i* at time step *n*, such that

$$Q_{ij}(n) = \max\{Q_{ij}(n-1) + A_{ij}(n) - D_{ij}(n), 0\} \tag{1}$$

where $A_{ij}(n) \in \{0,1\}$ and $D_{ij}(n) \in \{0,1\}$ are the number of arrivals and departures during time step *n*, respectively. To ensure the existence of a stochastic equilibrium of the queueing system, the arrival rate for each queue should converge to the departure rate, yielding the condition

$$\lim_{n\to\infty}\left\{\frac{\sum_{t=1}^{n} A(t)}{n}\right\} = \lim_{n\to\infty}\left\{\frac{\sum_{t=1}^{n} D(t)}{n}\right\} \tag{2}$$

Utilizing the fact that $\mu_0 \equiv 0$ together with the statement of convergence between the arrival and departure rates, a general expression for the queueing balance equation can be written as

$$\lambda = \sum_{m=1}^{\infty} \mu_m \pi_m \tag{3}$$

where λ is the mean arrival rate, π_m is the steady-state queue size distribution (i.e. $\pi_m = P\{Q = m\}$) and μ_m is the probability of service given that the queue size is *m*. For a *Geo/Geo/1* system in which $\mu_m = \mu$, a direct outcome of the steady-state balance equation for the early arrival model can thus be written as

$$\lambda = \sum_{m=1}^{\infty} \mu\left(\pi_m(1-\lambda) + \pi_{m-1}\lambda\right) = \mu\left(1-\pi_0(1-\lambda)\right), \tag{4}$$

where the term $\pi_0(1-\lambda)$ expresses the probability that the queue was empty prior to the arrival phase and no cell has arrived. When rearranged, (4) yields the well-known result for *Geo/Geo/1* systems [8],

$$\pi_0 = \frac{\mu - \lambda}{\mu(1-\lambda)}. \tag{5}$$

The common goal for each examined scenario is to find a closed-form expression for the steady-state queue size distribution, π_m. Of particular interest is π_0, which represents the share of time that the queue is empty of cells. In some cases, π_0 is sufficient to derive the queue size distribution while in others additional information is required. Based on the queue size distribution, important performance metrics can be directly obtained including the mean queue size, $E[Q] = \sum_{m=1}^{\infty} m\pi_m$ and using Little's theorem [8], the mean queue latency, $E[\tau] = E[Q]/\lambda$. For any given stable queueing system we necessitate that π_m satisfy $\sum_{m=0}^{\infty} \pi_m = 1$.

Throughout the paper we call attention to the distinction between two independent attributes of traffic patterns: the *destination distribution* and the *arrival process*. Destination distribution corresponds to the probability of an arriving cell to be destined to each of the output ports. The most common and simple case is that of uniform distribution, whereby a packet has an identical likelihood of being destined to each output, thereby implying that $\lambda_i = \lambda / N, \; i, = 1,2,..N$. However, real life traffic has been shown to be characterized by non-uniform destination distribution throughout all levels of the network hierarchy [9], [10].

IV. UNIFORM BERNOULLI ARRIVALS

The most commonly examined traffic is uniformly distributed between the outputs and obeys a Bernoulli i.i.d. arrival process. Recalling that υ_k denotes the mean size of the ODS at the beginning of the k^{th} phase (k = 1,2,…N), we note that $\upsilon_1 = N$ since at the beginning of the first phase all outputs are unreserved. Moreover, υ_k forms a monotonically non-increasing series given that during each phase either a node selects an output, resulting in $\upsilon_{k+1} = \upsilon_k - 1$ or, alternatively, it does not select any output implying that $\upsilon_{k+1} = \upsilon_k$.

The probability that a node selects no outputs may be interpreted as the probability that all non-empty queues belonging to the ODS are empty. In all other cases one output is selected. Employing the early arrival model described in chapter 2, the probability of a queue not selecting any outputs equals the probability that all queues within the ODS were empty at the beginning of the time slot *and* no cell has arrived during the current time slot. Expressing the above mathematically, gives us the following recursive expression for υ_{k+1}:

$$\upsilon_{k+1} = \begin{cases} \upsilon_k & \text{with prob. } \pi_0^{\upsilon_k}\left(1-\frac{\upsilon_k}{N}\lambda\right) \\ \upsilon_k - 1 & \text{with prob. } 1-\pi_0^{\upsilon_k}\left(1-\frac{\upsilon_k}{N}\lambda\right) \end{cases} \tag{6}$$

and hence, combining the probabilistic terms in (6) into an expected value expression yields

$$E[\upsilon_{k+1}]=\begin{cases} N & k=0 \\ E[\upsilon_k]-1+\pi_0^{E[\upsilon_k]}\left(1-\dfrac{[\upsilon_k]}{N}\lambda\right) & 0\le k\le N-1 \end{cases} . \quad (7)$$

By applying summation and rearranging (7), it can be shown that an approximation of the mean ODS size, $\overline{\upsilon}=E[\upsilon_k]$, is given by

$$\overline{\upsilon}=\frac{N+1}{2}+\frac{\pi_0^2\left[N(1-\pi_0)-\lambda\right]}{N(1-\pi_0^{-1})^2} \quad (8)$$

For high load conditions ($\lambda\to 1$) the probability of a queue being empty is low, resulting in the intuitive conclusion that $\overline{\upsilon}\to(N+1)/2$, which is the mean of the arithmetic series decreasing from N to 1. Our initial examination is that of *random selection* arbitration where non-empty queues contend for transmission in an equal manner. In other words, a queue containing 3 cells and another containing 10 cells have identical probabilities of prevailing for transmission. In the context of the DISA algorithm, there are three conditions that must be met for a queue to be selected: (1) the queue must reside within the ODS, (2) the queue must be non-empty and (3) the queue must prevail when contending against the other non-empty queues in the ODS such that

$$P_{departure}=P\{Q_{ij}\in\psi\}P\{Q_{ij}>0\}P\left\{\begin{matrix}Q_{ij}\text{ prevails for}\\ \text{transmission}\end{matrix}\middle|\, Q_{ij}\in\psi, Q_{ij}>0\right\} \quad (9)$$

From (8), an estimated average probability of being in the ODS is $\overline{\upsilon}/N$ and the probability that a queue is non-empty following an arrival phase is $(1-\pi_0(1-\lambda/N))$. The probability of prevailing in the internal contention for transmission can be expressed as

$$P\left\{\begin{matrix}Q_{ij}\text{ prevails for}\\ \text{transmission}\end{matrix}\middle|\, Q_{ij}\in\psi, Q_{ij}>0\right\}=\frac{1}{(\overline{\upsilon}-1)(1-\pi_0)+1} \quad (10)$$

where $(\overline{\upsilon}-1)(1-\pi_0)$ represents the mean number of non-empty queues in Ψ (other than the queue analyzed) to which we add 1 to account for the analyzed queue. Substituting these terms in (9) yields

$$P_{arrival}=\frac{\lambda}{N}=P_{departure}=\frac{\overline{\upsilon}\left(1-\pi_0(1-\lambda/N)\right)}{N[(\overline{\upsilon}-1)(1-\pi_0)+1]} \quad (11)$$

The service discipline is governed by a memoryless process, primarily since within each time slot decisions are made regardless of the outcome of previous time slots. To that end, the queueing behavior may be described using a *Geo/Geo/1* model whereby both the arrival and service processes are the outcomes of Bernoulli i.i.d. trials with parameters λ and μ, respectively. Equation (11) states that the probability of departure equals the probability of the queue being non-empty multiplied by a term for the probability of service. Accordingly, we may isolate the probability of service, μ, as

$$\mu=\frac{\overline{\upsilon}}{N[(\overline{\upsilon}-1)(1-\pi_0)+1]} \quad (12)$$

Substituting (8) in (11) and assuming that $1/N<<1$, we finally obtain

$$\pi_0\approx\frac{(1-\lambda)(N-1)}{(1-\lambda)(N+1)+\lambda}. \quad (13)$$

Utilizing the results from the *Geo/Geo/1* model we find the mean queue occupancy to be

$$E[Q]=\frac{1-\pi_0}{\pi_0}=\frac{\lambda}{(1-\lambda)(N+1)} \quad (14)$$

and, accordingly, the mean queueing delay is

$$E[\tau]=\frac{E[Q]}{\lambda/N}=\frac{N}{(1-\lambda)(N+1)}. \quad (15)$$

The latter implies that the mean queueing delay, for large values of N, does not depend on N and is roughly $1/(1-\lambda)$. Figure 3 depicts the mean queueing delay of a 128-port switch with Bernoulli i.i.d. uniformly distributed arrivals, compared to the theoretical limit of an output queued switch.

In practice, completing the scheduling cycle within a single cell time (approximately 50 nsec for 10 Gbps links) for $N>16$ is impractical. In view of the latter, we next investigate the performance implications of increasing the switching interval duration beyond one time slot. It is apparent from the nature of a multi-time-slot service discipline that we are still considering a memoryless process, particularly in view of the fact that consecutive service (switching) events are independent of previous ones. Accordingly, we may continue to utilize the *Geo/Geo/1* model providing that we find an expression for π_0 which reflects on the lengthy switching intervals. It is intuitive that as these intervals increase in duration, more cells accumulate in the queues and hence the mean queue occupancy is expected to increase and π_0 to decrease. Using similar rationale to that of (3), we may write

$$B\frac{\lambda}{N}=\mu'\sum_{i=1}^{B}\alpha^i \quad (17)$$

where $\alpha=\left(1-\pi_0(1-\lambda/N)\right)$ representing the probability of a queue being non-empty, B is the duration of the switching interval in time slots and

$$\mu'=\frac{\overline{\upsilon}}{N[(\overline{\upsilon}-1)(1-\pi_0)+1]}. \quad (18)$$

The left hand side of (17) corresponds to the mean number of arrivals that are expected within B time slots, while the right hand side pertains to the probability of a service event occurring (μ') multiplied by the probabilities of having B non-empty time slots in succession. Equation (18) is a polynomial expression of order B with only a single solution for α in the region (0,1). Using α we find π_0, from which the mean queue occupancy and mean waiting time are derived. Figure 4 depicts the mean queueing delay as a function of the offered load for various switching interval durations.

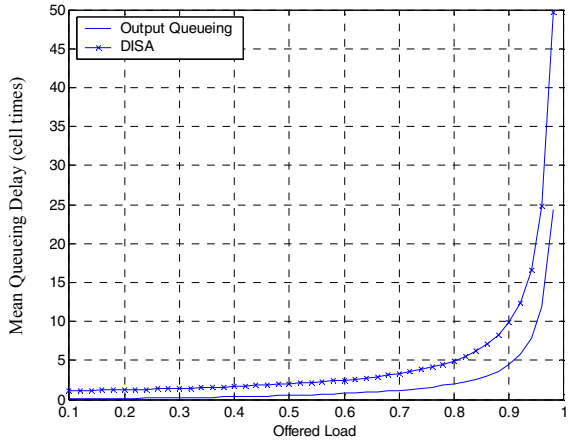

Figure 3. Mean queueing delay for DISA and an output queued switch with N=128 and Bernoulli i.i.d. uniformly distributed arrivals

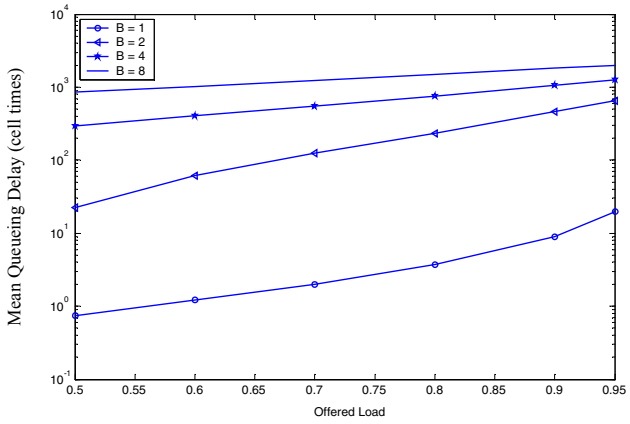

Figure 4. Mean queueing delay for N=128 with random selection arbitration and various switching interval durations.

We next examine longest-queue-first (LQF) arbitration in which, in contrast to random selection, the queue with the largest occupancy is selected for transmission. The significance of considering the queue size becomes apparent when the switching interval is larger than one time slot thus allowing for the accumulation of cells. Figure 5 illustrates a Markov chain that describes the queueing behavior with LQF and a switching interval of B time slots. The states correspond to queue occupancy where the n-step forward and reverse transition probabilities are defined as

$$\alpha_n = \binom{B}{n}\left(\frac{\lambda}{N}\right)^n\left(1-\frac{\lambda}{N}\right)^{B-n}, \quad n = 1,2,\ldots,B$$
$$\beta_n = \binom{B}{n}\left(\frac{\lambda}{N}\right)^{B-n}\left(1-\frac{\lambda}{N}\right)^{n} \qquad (19)$$

Assuming there are J+1 states in the system, implying that $P(Q > J) - 0$, we obtain J+1 equations from the Markov chain. The variables included in this set of equations are π_m (m=0,1,…J) and μ_m (m=1,2,…J), amounting to 2J+1 independent variables. Since the Markov chain is assumed to be in equilibrium, an additional equation is $\sum_{m=0}^{\infty}\pi_m = 1$. The remaining J equations are given by the approximation

$$\mu_m = \frac{1}{N}\sum_{\upsilon=1}^{N}\binom{\upsilon}{N}\left[\frac{P(Q_i \le m, i \in \nu)}{(\upsilon-1)\pi_m + 1}\right] \qquad (20)$$

which is the expected probability that the VOQ is serviced multiplied by the probability that the size of all other υ-1 queues in the ODS is smaller or equal to m *and* the queue prevails in contending against the set of other queues with size equal to m. The inner probabilistic term in (20) an be coarsely approximated as

$$P(Q_i \le m, i \in \nu) \approx \left(\sum_{j=0}^{m}\pi_j\right)^{\upsilon-1} \qquad (21)$$

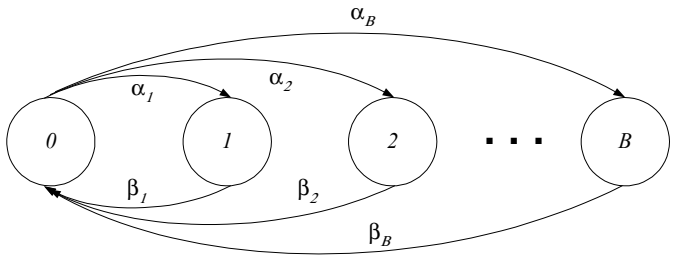

Figure 5. Markov chain for a system with LQF arbitration where α_n and β_n are the n-step forward and reverse transition probabilities, respectively.

reflecting on the probability that the size of υ - 1 queues are independently smaller or equal to m. Numerical techniques may be applied to obtain approximations of π_m from which the mean queue occupancy and mean delay are derived.

V. NON-UNIFORM DESTINATION DISTRIBUTION

Let λ_k denote the non-homogeneous probability that a cell is destined to output k, such that for a given VOQ system $\sum_{k=1}^{N}\lambda_k = \lambda$. The service discipline is independent of the arrival process and remains the same as before. We assume random selection between non-empty queues as the VOQ arbitration scheme. We further define Q_k {k=1,2,..N} as the size of the k^{th} queue having the arrival rates λ_k.

Utilizing the *Geo/Geo/1* model for the same reasons described in the case of uniform distribution, we focus our analysis on expressing the steady-state probability of Q_k being empty, i.e. $\pi_0^k = \Pr\{Q_k = 0\}$. Recalling the three conditions for transmission and accordingly rewriting (3) by replacing the generic term λ with λ_k, we obtain

$$\lambda_k = \hat{\mu}\frac{\left(1-\pi_o^k(1-\lambda_k)\right)}{\sum_{j\ne k}(1-\pi_0^j)+1} = \hat{\mu}\frac{\left(1-\pi_o^k(1-\lambda_k)\right)}{N-\sum_{j\ne k}\pi_0^j} \qquad (22)$$

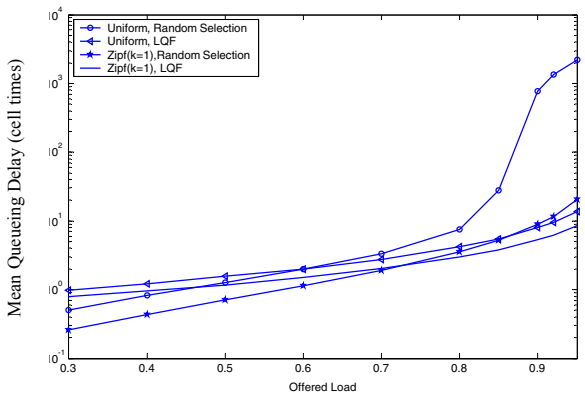

Figure 6. A comparison of the mean queue occupancy for a 16-port switch with Bernoulli i.i.d. arrivals obeying uniform and $Zipf_{k=1}$ destination distributions.

where $\hat{\mu} = \bar{v}/N$. Isolating π_0^k gives us

$$\pi_0^k = \frac{1}{1-\lambda_k} - \frac{\lambda_k}{\hat{\mu}(1-\lambda_k)}\left[N - \sum_{j \neq k} \pi_0^j\right] \tag{23}$$

Letting $\Pi_N = \{\pi_0^1, \pi_0^2, ..., \pi_0^N\}^T$ and rearranging (23) for each of the queues, we arrive at a matrix solution in the form of

$$\begin{pmatrix} 1 & \alpha_1 & \alpha_1 & \alpha_1 & ... & \alpha_1 \\ \alpha_2 & 1 & \alpha_2 & \alpha_2 & ... & \alpha_2 \\ \alpha_3 & \alpha_3 & 1 & \alpha_3 & ... & \alpha_3 \\ ... & ... & ... & ... & ... & ... \\ \alpha_N & \alpha_N & \alpha_N & ... & \alpha_N & 1 \end{pmatrix} \Pi_N = \begin{pmatrix} \alpha_1 - \beta_1 N \\ \alpha_2 - \beta_2 N \\ \alpha_3 - \beta_3 N \\ ... \\ \alpha_N - \beta_N N \end{pmatrix} \tag{24}$$

where

$$\alpha_k = -\frac{\lambda_k}{\hat{\mu}(1-\lambda_k)}, \beta_k = -\frac{1}{1-\lambda_k} \tag{25}$$

Solving this linear system, we find expressions for which, by letting ρ_k=1-π_0^k and utilizing the results for the *Geo/Geo/1* model, yields

$$E[\tau_k] = \frac{E[Q_k]}{\lambda_k} = \frac{\rho_k}{\lambda_k(1-\rho_k)} \cdot \tag{26}$$

Figure 6 illustrates the mean delay for a 16-port switch with Bernoulli i.i.d. arrivals which are distributed both uniformly and according to $Zipf_{r=1}$, where

$$Zipf_r(k) = \lambda_k = k^{-r}\left(\sum_{j=0}^{N} j^{-r}\right)^{-1} \tag{27}$$

and r is a model parameter, pragmatically assumed to equal approximately 1. As can be noted from the results, random selection leads to significant degradation in the performance when non-uniform destination distribution is introduced, particularly at high loads. However, when utilizing LQF the algorithm exhibits notable robustness to the destination distribution.

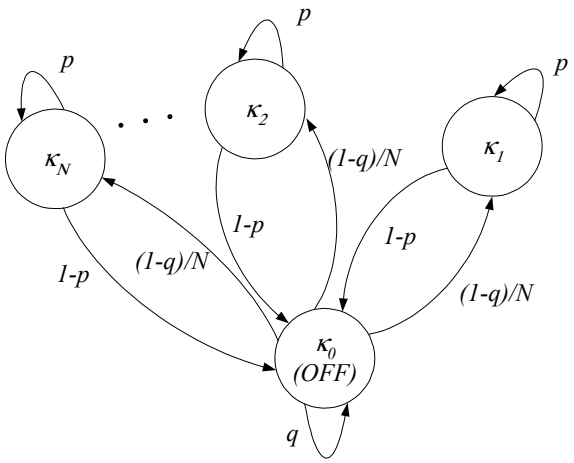

Figure 7. Markov chain characterizing the behavior of a uniformly distributed ON/OFF arrival process to a VOQ with N ports. Each port receives a mean offered load of λ/N.

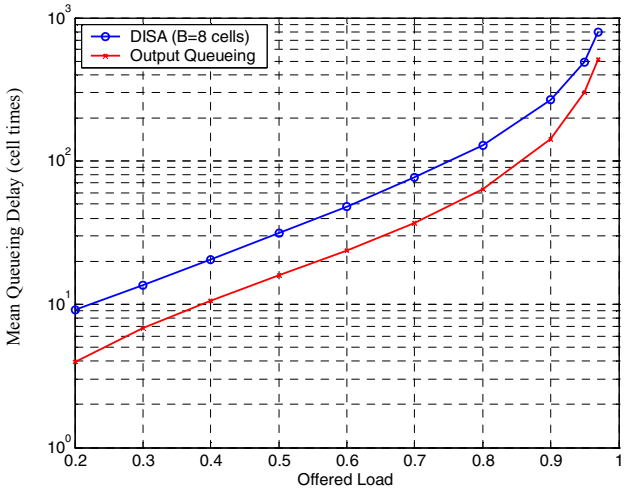

Figure 8. Performance of the DISA scheduling algorithm in a 128-port switch under bursty traffic with mean burst size of 32 cells. LQF arbitration is employed utilizing switching intervals of 8 cells in duration.

VI. Correlated Arrivals

In an aim to evaluate the DISA algorithm under traffic patterns that more accurately portray the nature of Internet traffic, we focus our attention on scenarios with ON/OFF [10] correlated arrival patterns. We extend the basic two-state ON/OFF model to accommodate for an N-queue VOQ system. Within each burst all cells are destined to the same output. The bursts are geometrically distributed in size, as with the classic two-state ON/OFF model. The process alternates equally between the various queues such that each queue receives, on average, an equivalent portion of the traffic, i.e. $\lambda_i = \lambda/N$ for i=1,2,…N. The Markov chain describing the behavior of this arrival process is depicted in Figure 7.

Figure 8 illustrates the mean latency of the switch, where bursty traffic is applied to a 128-port switch with envelope switching intervals and LQF arbitration employed. The reader will note that the performance is relatively close to that of an output queued switch. The reason for the exceptionally low delay lies in the fact that correlated traffic results in temporarily un-evenly populated queues, allowing the scheduler to more efficiently utilize lengthy switching intervals. As a result, the latency obtained under bursty scenarios is lower than that of Bernoulli i.i.d. traffic. The deployment of LQF allows the scheduler to grant service to the most populated queue, hence improving the switching capacity utilization. Since real-life traffic does tend to be correlated on several levels, the robustness exhibited by the DISA algorithm under bursty arrivals is perhaps one of its key properties.

VII. Conclusions

We have presented the DISA scheduling algorithm along with a complementary switch architecture. Analytical foundations coupled with simulation results have been provided for evaluating the performance of the system under different traffic scenarios, including non-uniform destination distribution and correlated arrivals. Low latency and maximal throughput are exhibited for admissible traffic loads. Ease of implementation in conjunction with the ability to utilize commercially available crosspoint switches with moderate configuration rates, further accentuate the attractiveness of the proposed scheme for high-port density switching platforms.

References

[1] Information available at: http://www.mindspeed.com.

[2] Information available at: http://www.vitesse.com.

[3] N. McKeown, "The iSLIP Scheduling Algorithm for Input-Queued Switches," *IEEE/ACM Trans. on Networking*, Vol. 7, No. 2, pp. 188-201, April 1999.

[4] K. Yamakoshi, K. Nakai, E. Oki, N. Yamanaka, "Dynamic Deficit Round-Robin Scheduling Scheme for Variable-Length Packets," *Electronics Letters*, Vol. 38, No. 3, pp. 148-149, Jan. 2002.

[5] R. Rojas-Cessa, E. Oki, H. J. Chao**, "**CIXOB-k: Combined Input-Crosspoint-Output Buffered Packet Switch," *Proc. IEEE GLOBECOM 2001*, Vol. 4, pp. 2654-2660, Nov. 2001.

[6] I. Elhanany, D. Sadot, "A Contention-Free Tbit/sec Packet Switching Architecture for ATM over WDM Networks," *IEICE Trans. on Communications*, Vol. E83-B, No. 2, Feb. 2000.

[7] M. J. Karol, M. G. Hluchyj, and S. P. Morgan, "Input versus Output Queueing in a Space Division Switch," *IEEE Trans. Communications*, Vol. COM-35, pp. 1347-1356, Dec. 1987.

[8] J. J. Hunter, *Mathematical Techniques of Applied Probability: Discrete Time Models*: *Techniques and Applications*, Vol. 2, Academic Press, 1983.

[9] C. Williamson, "Internet Traffic Measurement," *IEEE Internet Computing*, Vol. 5, No. 6, pp. 70-74, Nov.-Dec. 2001.

[10] I. Elhanany, M. Kahane, D. Sadot, "On Uniformly Distributed ON/OFF Arrivals in Virtual Output Queued Switches with Geometric Service Times," to appear in *Proc. IEEE ICC 2003*, Anchorage, Alaska.

Appendix A: Proof of Stability for the DISA Algorithm

In this appendix we provide a proof for theorem 1. We being by noting that each of the virtual output queues can be analyzed, without loss of generality, as a *GI/G/1* queueing system with $Q(n)$ denoting the queue occupancy at time step n. The latter implies generic arrival and service processes. We label $E[T_n]$ as the mean interarrival time and $E[S_n]$ as the mean interservice time. It has been shown that for a *GI/G/1* queue, if $E[T_n]>E[S_n]$ then $\lim_{n\to\infty}\{Q(n)\}=C$ exists, and thus the queueing system is stable. An alternative and identical interpretation of this stability criterion is that the mean probability of arrival, $\lambda=(E[T_n])^{-1}$ is smaller than the mean probability of service, $\mu=(E[S_n])^{-1}$. Under the assumption of uniformly distributed arrivals, the mean probability of arrival is λ/N. Recalling that υ_k denotes the mean size of the ODS at the beginning of the i^{th} phase ($i = 1,2,\ldots N$). By observing that at most a single output is selected by each input during the reservation process, we have $\upsilon_k \geq (N - (i - 1))/N$. Accordingly, we can write

$$\Pr\{service \mid phase = i\} \geq \frac{1}{r_i}\left(\frac{N-(i-1)}{N}\right). \tag{28}$$

where r_i represents the mean number of non-empty queues at the i^{th} phase. The latter pertains to the worst case scenario whereby in each phase an output is matched to an input and thus removed from the ODS. If this is not the case, the probability of service increases since more destinations are offered resulting from a larger ODS. Since the probability that a queue contends for transmission during each of the phases is uniform and equal to $1/N$, we have

$$\frac{\lambda}{N} < E[\Pr\{service\}] = \frac{1}{N}\left(\frac{1}{r_1}+\frac{(N-1)}{N}\frac{1}{r_2}+\frac{(N-2)}{N}\frac{1}{r_3}+\ldots+\frac{1}{N}\right). \tag{29}$$

However, we know that $r_i \leq (N-(i-1)) < N$, leading to

$$\frac{1}{N}\left(\frac{1}{r_1}+\frac{(N-1)}{N}\frac{1}{r_2}+\frac{(N-2)}{N}\frac{1}{r_2}+\ldots+\frac{1}{N}\right) \geq \frac{1}{N}\left(\frac{1}{N}+\frac{(N-1)}{N}\frac{1}{N-1}+\frac{(N-2)}{N}\frac{1}{N-2}+\ldots+\frac{1}{N}\right)=\frac{1}{N} \tag{30}$$

Since we require for admissibility that $\lambda<1$, the system is always stable and thus we conclude the proof.