International Journal of Pure and Applied Mathematics
Volume 19 No. 1 2005, 95-105

STABLE SCHEDULING SCHEMES FOR PACKET SWITCHES

G. Baklavas[1], M. Roumeliotis[2] §
[1,2]Department of Applied Informatics
University of Macedonia
P.O. Box 1591, 156 Egnatia Str., Thessaloniki, 54006, GREECE
[1]e-mail: grigoris@uom.gr
[2]e-mail: manos@uom.gr

Abstract: This paper discusses packet scheduling algorithms similar to the ones used in modern IP routers. Two new scheduling techniques (schemes) are proposed. We describe them formally and then prove the conditions under which they keep their stability property, meaning that they do not experience packet loss even under the most severe traffic conditions. It is shown that every packet scheduling algorithm that follows any of the two proposed schemes remains stable under any admissible traffic pattern.

AMS Subject Classification: 68R10
Key Words: input queue switches, packet scheduling, queuing system stability

1. Introduction

One of the most important issues in today's networking is the delay problem. The time it takes for a packet to travel through Internet and reach its destination can sometimes make executing an application unfeasible. This is very common especially in multimedia applications.

Received: January 27, 2005

§Correspondence author

To deal with this problem, there are two general approaches. The first one suggests to constantly enhance networks with more bandwidth. This could provide more resources to the average user, thus improving network performance. The disadvantage of this approach is mainly its high cost and the difficulty of changing the existing network infrastructure.

On the other hand, there is the approach that seeks for the optimization of the current infrastructure. This will mean better use of the existing network resources. To achieve this, new algorithms have to be introduced. This paper discusses a set of such algorithms that deal with the scheduling of packets within the main network relay, the router. We formally prove that both scheduling schemes remain stable, i.e. they are not losing packets under all admissible traffic loads.

A similar study was performed in [2]. The difference is that now we deal with packets instead of cells. The importance of this variation is immense as currently in Internet there is a huge number of IP routers that operate using IP datagrams and do not split packets into cells.

A brief description of an Input-Queue (IQ) switch is given in Section 2. We also propose a variation on the operation of the device, so that packets are accommodated instead of cells. After defining the notations used, and providing the basic definitions of our study in Section 3 and Section 4, we examine the stability properties of our scheduling schemes. Section 5 ends with the main result of the paper.

2. IQ Switches

The logical architecture for an IQ packet switch is shown in Figure 1. At each input there is a segmentation of the incoming packet into cells. The switch operates in store-and-forward mode, is equipped with enough memory to store a maximum-size packet, and starts the segmentation process only after the complete reception of the packet. The cells resulting from the segmentation are transferred to the cell-switch input. The capacity of each input queue at the cell-switch is finite, hence losses can occur. We assume that the entire packet is discarded if the input queue of the cell-switch does not have enough free space to store all the cells deriving from the segmentation of the packet when the first of these cells hits the queue. This is clearly a pessimistic assumption, but has the advantage of case of implementation, and of avoiding the transmission of incomplete packet fractions through the switch.

We consider a switch with K inputs and K outputs. We also assume for

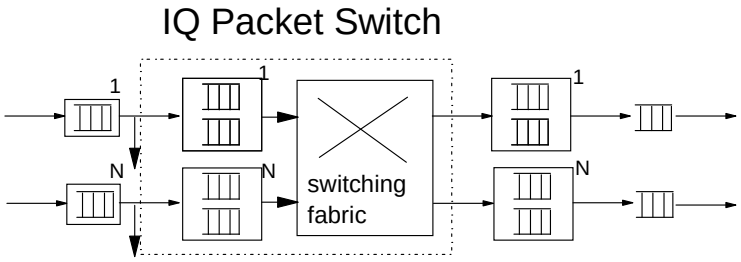

Figure 1: Input Queue (IQ) switch

simplicity that all input and output lines run at the same speed. Each input manages one queue for each output, hence a total of $K \times K = K^2$ queues are present. Each queue can store up to Q_{max} cells and excess packets are dropped. This queue separation technique avoids performance degradations due to "head-of-the-line blocking" [5], and is called Virtual Output Queuing (VOQ) or Destination Queuing [1, 7].

The cell-based switching fabric transfers cells from input to output queues, according to a scheduling algorithm. Then, packets (i.e., IP datagrams) are reassembled. In general, cells belonging to the same packet are contiguous in the input queue of the internal IQ cell-switch. By using packet scheduling mode (described in the following section) cells belonging to the same packet are kept contiguous also in the output queue, and the reassembly modules are no longer necessary (or at most one per output is used). Once a packet is complete, it is logically added to an global output packet First-In-First-Out (FIFO) queue, from which packets are sequentially transmitted onto the output line. Note that the queue's functionality is typically implemented by imposing a sequential transfer from the suitable output to the output line of all the cells belonging to the same reassembled packet.

2.1. Packet Scheduling Mode

Packet scheduling mode algorithms introduce the additional constraint of keeping the cells belonging to the same packet contiguous also in output queues. To achieve this, the scheduling algorithm must enforce that, once the transfer through the switching fabric of the first cell of a packet has started towards the corresponding output port, no cells belonging to other packets can be transferred to that output, i.e., when an input is enabled to transmit the first cell of a packet comprising of m cells, the input/output match must persist for the

following $m-1$ slots.

We propose to extend the above mentioned IQ architecture to operate in packet scheduling mode. The only complexity increase in the implementation is to add a Boolean variable at each input to flag over-prioritized connections.

3. Notations

The notation used in the proofs is the following:

— Let K is the number of switch ports and Q is the number of queues (input and output). Clearly, $Q = K^2$.

— Let t and ν are two discrete time variables.

— Let B_t be a vector showing the number of cells currently waiting in the system at time t: it has K^2 elements and the i-th element is the number of cells currently waiting in the i-th queue.

— Let A_t be a vector showing the arrivals at time t: it has K^2 elements, it is a binary vector and a 1 in the i-th element implies the arrival of a cell at the i-th queue at time t.

— Let Δ_t be a vector showing the departures at time t: it has K^2 elements, it is a binary vector and a 1 in the i-th element implies the departure of a cell from the i-th queue at time t. It also corresponds to a matching Δ between input and output ports.

— Let $B_{t+1} = [B_t + A_t - \Delta_t]^+$ be the evolution of the system. We assume that first the arrival take place and then cell depart from the queues.

— $E(Z)$ is the expected value of the random variable Z.

4. Theoretical Analysis

In the literature, there exists a number of analytical approaches for studying IQ switches. Usually, an IQ switch is modelled as a controlled queueing system and therefore it can be studied using stochastic modelling techniques.

The major component of a queuing system of this kind is its stability. A queueing system is stable, when there is no queue growing to infinity, assuming that the arrival process is "admissible".

Definition 1. We call $A_{\chi\psi}$ the arrival process from port χ to port ψ and $\lambda_{\chi\psi}$ the average arrival rate. The aggregation of all process is $A = \{A_\chi, 1 \leq \chi \leq K\}$. An arrival process A is considered admissible when no

port is overloaded, i.e.:

$$\sum_{\chi=1}^{K} \lambda_{\chi\psi} < 1, \qquad 1 \le \psi \le K\,,$$

$$\sum_{\psi=1}^{K} \lambda_{\chi\psi} < 1, \qquad 1 \le \chi \le K\,.$$

To prove that a queuing system is stable, a special function, called the "Lyapunov" function, is used. In [6], there is also an analysis for estimating delays using this method.

5. Stability of Packet Scheduling Mode

We start with two well-known definitions of the stability of controlled queueing systems.

Definition 2. A system of queues is stable, or achieves 100% throughput, if

$$\lim_{t\to\infty} \frac{B_t}{t} = \lim_{t\to\infty} \frac{1}{t} \sum_{\chi=0}^{t-1} (A_x - \Delta_x) = 0,$$

with probability 1.

Definition 3. A system of queues is strongly stable if:

$$\lim_{t\to\infty} \sup E||B_t|| < \infty\,.$$

Consider an IQ packet switch, and suppose that all input packet lengths are multiples of some unit length called UL (UL may correspond to a bit, a byte, or a cell). Consider the system of discrete-time queues comprising all input queues of the packet switch. The discrete time unit corresponds to a continuous time increment equivalent to UL.

We assume that customers correspond to cells to be transferred from input to output ports. Since we consider an IQ switch, each element δ_i^{ν} of the departure vector Δ_ν, can only assume the values 0 and 1 $\forall i$ and $\forall \nu$. The arrival of a packet corresponds to the arrival of a group of customers, whose cardinality equals the packet length in UL units. Therefore, α_n^i can be larger than 1. However, if the traffic is admissible, $E[\alpha_n^i] \le 1, \forall i$.

Let $t_\nu \in \mathbb{N}^+$ be a non-defective sequence of regeneration instants (or stopping times) for the evolution of the system of queues, i.e., for any t_ν, the evolution of the system following t_ν is conditionally independent of the evolution of the system before t_ν given the state $Y(t_\nu)$; moreover, $\zeta_\nu = t_{\nu+1} - t_\nu$.

Definition 4. An IQ packet switch follows a renewal MWM scheme if at each stopping time t_ν a new switching configuration is selected according to the outcome of a Maximum Weight Matching (MWM) algorithm whose weights are proportional to queue lengths, and the switching configuration is kept constant until $t_{\nu+1}$.

Definition 5. An IQ packet switch follows an incremental MWM scheme if at each stopping time t_ν a new matching is selected according to the outcome of a MWM algorithm whose weights are proportional to queue lengths. Between two consecutive stopping times t_ν and $t_{\nu+1}$, partial updates of the switching configuration are allowed. These reconfigurations are performed according to the outcome of a MWM algorithm whose weights are proportional to queue lengths, operating on a subset of input and output ports.

Lemma 1. *An IQ packet switch following a renewal MWM scheme is stable under any admissible i.i.d. input traffic pattern A_ν such that $E[A_\nu A_\nu^T] < \infty, \forall \nu$.*

Proof. The evolution of the system of discrete-time queues in the IQ packet switch is represented by a Discrete Time Markov Chain whose state is defined by the vector of queue lengths B_{t_ν}; between consecutive stopping times, the system evolution satisfies the following equation:

$$B_{t_{\nu+1}} = B_{t_\nu} + \sum_{i=0}^{\zeta_\nu - 1} (A_{t_\nu+i} \Delta_{t_\nu+i}) \,.$$

Note that all $\Delta_{t_\nu+i}, i < \zeta_\nu$ refer to the same matching; however, they need not be all equal, since some queue scheduled for transmission at time t_ν may become empty before the next stopping time. If this happens, no packet can be transferred from empty queues.

By using the Lyapunov function $V(B_{t_\nu}) = B_{t_\nu} B_{t_\nu}^T$:

$$\begin{aligned} &E[V(B_{t_{\nu+1}}) \mid B_{t_\nu}] - V(B_{t_\nu}) \\ &\qquad = E[2 \sum_{i=0}^{\zeta_\nu - 1} (A_{t_\nu+i} \Delta_{t_\nu+i}) B_{t_\nu}^T + \sum_{i=0}^{\zeta_\nu - 1} (A_{t_\nu+i} \Delta_{t_\nu+i}) \sum_{i=0}^{\zeta_\nu - 1} (A_{t_\nu+i} \Delta_{t_\nu+i})^T] \,. \end{aligned}$$

Thus, under the assumption that $E[A_\nu A_\nu^T]$ is finite (which corresponds to assuming finite packet length variances), since also $E[\Delta_{t_{\nu+i}}$

$\Delta^T_{t_{\nu+i}}]$ is finite:

$$\lim_{\|B_\nu\|\to\infty} \frac{E[V(B_{t_{\nu+1}})|B_{t_\nu}]-V(B_{t_\nu})}{\|B_{t_\nu}\|} = \lim_{\|B_\nu\|\to\infty} \frac{2E[\sum_{i=0}^{\zeta_\nu-1}(A_{t_\nu+i}-\Delta_{t_\nu+i})B^T_{t_\nu}]}{\|B_{t_\nu}\|}.$$

Define now $\Delta_\delta = \sum_{i=0}^{\zeta_\nu-1} \Delta_{t_\nu+i} - \zeta_\nu \Delta_{t_\nu}$; as noted before, this difference is due to the fact that some queues may become empty before changes in the switch configuration. Thus:

$$\frac{E[\sum_{i=0}^{\zeta_\nu-1}(A_{t_\nu+i}-\Delta_{t_\nu+i})B^T_{t_\nu}]}{\| B_{t_\nu} \|} = \frac{E[\sum_{i=0}^{\zeta_\nu-1}(A_{t_\nu+i}B^T_{t_\nu}-\zeta_\nu\Delta_{t_\nu}B^T_{t_\nu}-\Delta_\delta B^T_{t_\nu})]}{\| B_{t_\nu} \|}.$$

Wald's Lemma [8] can be applied, since t_ν is a sequence of stopping times, therefore obtaining:

$$\frac{E[\sum_{i=0}^{\zeta_\nu-1}(A_{t_\nu+i}B^T_{t_\nu}-\zeta_\nu\Delta_{t_\nu}B^T_{t_\nu}-\Delta_\delta B^T_{t_\nu})]}{\|B_{t_\nu}\|} = \frac{E[\zeta_\nu](E[A_\nu]-\Delta_{t_\nu})B^T_{t_\nu}-E[\Delta_\delta]B^T_{t_\nu}}{\|B_{t_\nu}\|}.$$

Note that $E[\Delta_\delta]B^T_{t_\nu} \geq -QE[\zeta^2_\nu]$, since at most Q components of Δ_δ can be non-null, no component of Δ_δ can exceed the value ζ_ν, and, finally, a component of Δ_δ can be non-null only if the corresponding queue length at time t_ν is smaller than ζ_ν. Moreover, for each admissible load and non-null queue length vector, $(E[A_\nu]-\Delta_{t_\nu})B^T_{t_\nu} < 0$ as proved in [8]. Thus:

$$\begin{aligned}\lim_{\|B_\nu\|\to\infty} & \frac{E[V(B_{t_{\nu+1}}) \mid B_{t_\nu}] - V(B_{t_\nu})}{\| B_{t_\nu} \|} \\ &= \lim_{\|B_\nu\|\to\infty} \frac{E[\zeta_\nu](E[A]-\Delta_{t_\nu})B^T_{t_\nu} - 2E[\Delta_\delta]B^T_{t_\nu}}{\| B_{t_\nu} \|} \\ &= 2E[\zeta_\nu] \lim_{\|B_\nu\|\to\infty} \frac{(E[A]-\Delta_{t_\nu})B^T_{t_\nu}}{\| B_{t_\nu} \|} < -E[\zeta_\nu]\epsilon\,. \quad \square\end{aligned}$$

Lemma 2. *An IQ packet switch following an incremental MWM scheme is stable under any admissible i.i.d. input traffic pattern such that* $E[A_\nu A^T_\nu] < \infty, \forall\nu$.

Proof. The proof can be easily obtained by applying the same Lyapunov function used in Lemma 1.

Consider an IQ packet switch with a given packet arrival process running an incremental MWM scheme with stopping times $\{t_\nu\}$. A particular renewal MWM scheme can be defined under the same arrival process and with the same set of stopping times. The latter scheme is stable due to Lemma 1.

Since for the two schemes we have the same set of stopping times, we get:

$$\sum_{i=0}^{\zeta_\nu-1} \Delta^{Inc}_{t_\nu+i} B_{t_\nu+i} \geq \sum_{i=0}^{\zeta_\nu-1} \Delta_{t_\nu+i} B_{t_\nu+i}\,,$$

where $\Delta^{Inc}_{t_\nu+i}$ is the departure vector at time $t_\nu + i$ for the incremental MWM scheme, and $\Delta_{t_\nu+i}$ is the departure vector for the renewal MWM scheme. □

Definition 6. An IQ packet switch follows a packet MWM scheme if a new switching configuration is selected according to a MWM algorithm, whose weights are proportional to queue lengths, whenever either:

— all packet transmissions end at the same time, or

— all the queues selected for transfer become empty.

Definition 7. An IQ packet switch follows a packet incremental MWM scheme if:

— whenever either all packet transmissions end at the same time, or all the queues selected for transfer become empty, a new switching configuration is selected according to a MWM algorithm, whose weights are proportional to queue lengths, as in a packet MWM scheme;

— whenever some queues selected for transfer become idle (i.e., either they are empty, or packet transmissions end) a partial update of the switching configuration is allowed, according to an MWM algorithm among idle ports, whose weights are proportional to queue lengths.

Lemma 3. *Consider an IQ packet switch, following either a packet MWM scheme, or a packet incremental MWM scheme, whose input traffic is formed by variable length packets with i.i.d. random size. Packet sizes are expressed in integer multiples of UL. Assume that the average packet size is* Λ *and the packet size variance is* σ^2 *(both being finite). Assume that the transmission of packets from all queues selected by the MWM algorithm starts at the same time with exactly the same rate. Consider the sequence of instants* t_ν *at which either the transmission of all the packets at the head of the selected queues ends at the same time, or all selected queues become empty. The sequence of stopping times* t_ν *is non-defective, i.e.* $\zeta_\nu = t_{\nu+1} - t_\nu$*, are such that* $E[\zeta_\nu] < \infty$ *and* $E[\zeta_\nu^2] < \infty$.

Proof. For simplicity, assume that the packet length distributions at all queues are aperiodic, i.e., the maximum common divisor of all possible packet lengths expressed in UL is equal to 1. The proof can be easily extended to the case of periodicity. For simplicity we consider here a switch operating according to a packet MWM scheme, but the the proof can be easily extended for a switch operating according to a packet incremental MWM scheme.

We suppose that switch queues have infinite length, so that we neglect the probability that switch queues become empty near traffic saturation; thus, we obtain an overestimate of $E[\zeta_\nu]$ and $E[\zeta_\nu^2]$, since t_ν are defined by only the sequence of instants in which transmission of all the packets at the head of the selected queues ends at the same time.

Each sequence of instants at which transmissions of packets end at queue k forms a discrete-time aperiodic renewal point process, thanks to the independence of packet lengths. Thus, for Blackwell Theorem, see [8], the average number $E[f_\nu^\kappa]$ of packets whose transmissions end at queue at time satisfies the following equation:

$$\lim_{\nu\to\infty} E[f_\nu^\kappa] = \frac{1}{\Lambda}\,. \tag{1}$$

However, no more than one packet transmission can end at each queue at each time (assuming no packet is of length zero); thus $E[f_\nu^\kappa]$ equals the probability that a packet ends:

$$E[f_\nu^\kappa] = \Pr\{\text{transmission ends at time } \nu \text{ and at queue } \kappa\}.$$

Limit (1) implies that, for any integer $m > 1$, there exists an instant ν_κ such that, $\forall \nu > \nu_\kappa$:

$$\Pr\{\text{transmission ends at time } \nu \text{ and at queue } \kappa\} > \tfrac{1}{m\Lambda} > 0.$$

The probability that at instant ν the transmission of packets at the head of all queues selected for transmission (be their number N_S) ends can be easily computed, since no correlation exists among queues behavior. Thus, given m, for $\nu > \nu_\kappa, \forall \kappa$:

$$\Pr\{\text{all transmissions end at time } \nu\} = \prod_{k=1}^{N_S} \Pr\{\text{transmission ends at time } \nu \text{ and at queue } \kappa\} \geq \prod_{k=1}^{N_S} \tfrac{1}{m\Lambda} = \tfrac{1}{(m\Lambda)^{N_S}} > 0.$$

Consider now the sequence of instants t_ν at which either all packet transmissions end, or selected queues become empty. The sequence t_ν forms a renewal process; thus Blackwell's Theorem applies:

$$\Pr\{\text{all transmissions end at time } \nu\} = E[f_\nu] = \frac{1}{E[\zeta_\nu]},$$

where $E[f_\nu]$ is the average number of regenerations at time n; since

$$\Pr\{\text{all transmissions end at time } \nu\} > 0,$$

we obtain $E[\zeta_\nu] < \infty$.

To prove that also $E[\zeta_\nu^2] < \infty$, consider all packets transmitted from queue k between two subsequent regenerations; let W be the number of such packets, and Λ_j be their lengths expressed in UL. We can write:

$$\begin{aligned} E[\zeta_\nu^2] - E^2[\zeta_\nu] &= E[(\sum_{j=1}^{W}(\Lambda_j - E[\Lambda_j]))^2] \\ &= E[\sum_{j=1}^{W}(\Lambda_j^2 - E^2[\Lambda_j])] + E[\sum_{j=1}^{W}\sum_{i=1,i\neq j}^{W}(\Lambda_j\Lambda_i - E[\Lambda_j\Lambda_i])]\,. \end{aligned}$$

The second term in the sum can be easily shown to be null by conditioning on the value of W; it can thus be eliminated. As a consequence:

$$E[\zeta_\nu^2] - E^2[\zeta_\nu] = E[\sum_{j=1}^{W}\Lambda_j^2] - \sum_{j=1}^{W}E^2[\Lambda_j].$$

and by Wald's Lemma, since regeneration points are stopping times for the sequence Λ_j:

$$E[\zeta_\nu^2] - E^2[\zeta_\nu] = E[W]E[\Lambda^2] - E[W]E^2[\Lambda_j] = E[W]\sigma^2.$$

Being $E[W]$ finite (otherwise $E[\zeta_\nu]$ would be infinite), it results $E[\zeta_\nu^2] < \infty$. □

We can now state our main result.

Theorem 1. *Any IQ packet switch following either a packet MWM scheme or a packet incremental MWM scheme is strongly stable, provided that:*

— the input traffic is admissible;

— the input traffic is formed by variable length packets with i.i.d. random size having finite average and variance;

— the transmission of packets from all queues selected by the MWM algorithm starts at the same time with the same rate.

Proof. The proof is quite straightforward from Lemma 1 (for packet MWM schemes), or Lemma 2 (for packet incremental MWM schemes), and Lemma 3 since the assumptions of Theorem 1 satisfy the conditions under which Lemma 3 holds. □

6. Conclusions

In this paper, we have examined scheduling in packet switches. We proposed two new schemes in the area and formally proved that they are stable, thus providing 100% throughput under all admissible traffic patterns. The base of our approach was a similar study performed for cell switches.

To strengthen the results of this paper even more, we conducted a series of simulations that tested the schemes. We developed a network simulator, using *MODSIM III*, a special programming language for simulations. The simulator is described in detail in [4]. Some of the results are presented in [3]. Under all scenarios tested, the schemes proven to have comparable results with other algorithms in the area and also remained stable, especially in situations, where most other schemes were loosing packets.

References

[1] T. Anderson, S. Owicki, J. Saxe, C. Thacker, High speed switch scheduling for local area networks, *ACM Transactions on Computer Systems*, **11**, No. 4 (Nov. 1993), 319-351.

[2] G. Baklavas, M. Roumeliotis, Stability conditions of a recursive routing algorithm, *International Journal of Pure and Applied Mathematics*, **14**, No. 1 (2004), 75-89.

[3] G. Baklavas, M. Roumeliotis, Comparison of comparison of input-queue and output-queue cell switch architectures, In: *Proceedings of WSEAS AIC*, Tenerife (Dec. 2004).

[4] G. Baklavas, S. Souravlas, M. Roumeliotis, Simulating queue formations on a multi-port router, *IEEE CSMA*, Orlando (Oct. 2000).

[5] M. Karol, M. Hluchyj, S. Morgan, Input versus output queuing on a space division switch, *IEEE Transactions on Communications*, **35**, No. 12 (Dec. 1987), 1347-1356.

[6] E. Leonardi, M. Mellia, F. Neri, M. Ajmone Marsan, Bounds on average delays and queue size averages and variances in input queued cell-based switches, *IEEE INFOCOM 01*, Anchorage, AK, **3** (Apr. 2001), 1095-1103.

[7] Y. Tamir, H.-C. Chi, Symmetric crossbar arbiters for VLSI communication switches, *IEEE Transaction on Parallel and Distributed Systems*, **4**, No. 1 (Jan. 1993), 13-27.

[8] R.W. Wolff, *Stochastic Modeling and the Theory of Queues*, Prentice-Hall, NJ (1989).