Rate optimal scheduling schemes for asynchronous input-queued packet switches

Thomas Bonald, Davide Cuda

Rate-Optimal Scheduling Schemes for Asynchronous Input-Queued Packet Switches

Thomas Bonald, Davide Cuda*
Telecom ParisTech
23 avenue d'Italie
Paris, France
{thomas.bonald,davide.cuda}@telecom-paristech.fr

ABSTRACT

The performance of input-queued packet switches critically depends on the scheduling scheme that connects the input ports to the output ports. We show that, when packets are switched asynchronously, simple scheduling schemes where contention is solved locally at each input or output can achieve rate optimality, without any speed-up of the internal transmission rate.

1. INTRODUCTION

We consider an input-queued packet switch with virtual output queueing and no speed-up (i.e., the internal transmission rate of the switching fabric is equal to the line card rates). Current switches typically operate synchronously, the connections between input ports and output ports changing at the ticks of a common clock. Since packets have variable sizes, they must be segmented into fixed-size data units, which are switched independently and reassembled at the switch output.

The growth of the line card rates questions the viability of such an architecture [2]. In this paper, we consider an *asynchronous* switch, which does not require any packet segmentation. The connections between input ports and output ports are changed at arbitrary times. We describe some simple, distributed scheduling schemes that are provably rate-optimal, i.e., stabilize the queue lengths whenever possible. The contentions at each input or output are solved locally, based on the lengths of the corresponding virtual output queues (VOQ).

This stability result is remarkable in that synchronous switches are known to require complex scheduling schemes like maximum-weight matching to achieve rate optimality [15, 13, 5], except for the specific case of 2×2 switches [3]. In fact, practical schedulers like PIM [1] or iSLIP [12] only guarantee *maximal* matchings (i.e., both the input and the output associated with a non-empty VOQ are matched) and, as such, are suboptimal [6]. It turns out that, in the limit of large queue lengths, *maximum* matchings (i.e., the largest number of input-output connections associated with non-empty VOQs) are imposed by the switch asynchrony, which stabilizes the system.

Our result differs from those derived for wireless networks, that either rely on random access algorithms to approximate maximum-weight scheduling [9] or apply simple, greedy algorithms under restrictive assumptions on the underlying interference graph [6, 7, 11]. Note in particular that the stability result of Maguluri, Hajek and Srikant [11] on Longest Queue First requires the *local pooling* condition of [6], while that of Feuillet, Proutière and Robert [7] on random capture algorithms assumes multipartite interference graphs. In both cases, the condition on the underlying conflict graph of an $N \times N$ switch imposes that $N = 2$, as in [3]. We here use the bipartite structure of the graph connecting the inputs to the outputs to prove rate optimality for all N.

*The work presented in this paper has been carried out at LINCS, see www.lincs.fr

2. SCHEDULING SCHEME

Consider an $N \times N$ input-queued switch. Denote by I the set of inputs and by J the set of outputs. A feasible schedule is a matching on the bipartite graph $G = (V, E)$ where $V = I \cup J$ and $E = I \times J$.

Packets arriving at input i and destined to output j are directed to VOQ ij. If the VOQ is empty at the packet arrival and both the input and the output are unmatched, a new edge is added to the matching and the packet is directly transmitted to the output. Otherwise, the packet is queued. We denote by Q_{ij} the length of VOQ ij.

It remains to determine the change of matching at the end of a packet transmission, say from VOQ ij. Edge ij is then removed from the matching so that both input i and output j may select a new edge. By convention, we assume that the input takes the first decision. Specifically, for some $\alpha > 0$:

- input i selects, if any, unmatched output j' with probability proportional to $Q_{ij'}^{\alpha}$;
- if there is no such output or if $j' \neq j$, output j selects, if any, unmatched input i' with probability proportional to $Q_{i'j}^{\alpha}$.

The parameter α controls the sensitivity of the scheduling scheme to the queue lengths, from a purely random decision when $\alpha \to 0$ to a greedy algorithm based on Longest Queue First when $\alpha \to \infty$. Note that contentions are solved locally at input i then at output j. There are only two steps, unlike usual scheduling schemes in synchronous switches that are iterative and may involve, in theory, up to N steps [1, 12].

3. TRAFFIC MODEL

We assume that packets arrive at VOQ ij according to a Poisson process with intensity λ_{ij} and have exponential sizes with parameter μ_{ij} (these assumptions are discussed

in Section 6 below). We denote by $\rho_{ij} = \lambda_{ij}/\mu_{ij}$ the corresponding load. Traffic is said to be *admissible* if:

$$\forall i \in I,\ \sum_{j\in J} \rho_{ij} < 1 \quad \text{and} \quad \forall j \in J,\ \sum_{i\in I} \rho_{ij} < 1. \qquad (1)$$

The vector $Q(t)$ of VOQ lengths and the matching $M(t)$ at time t define a Markov process. The scheduling scheme is said to be rate-optimal if this Markov process is ergodic for any admissible traffic.

4. EXAMPLE OF THE 3×3 SWITCH

To illustrate the stability result, we start with the simple example of a 3×3 switch with unit mean packet sizes and bidiagonal traffic pattern, i.e., $\lambda_{ii} = \lambda_{i|i+1|3} = \lambda$ for all $i = 1, 2, 3$. The load of each input or output is $\rho = 2\lambda$. Figure 1 shows the evolution of the length of an arbitrary VOQ at load $\rho = 0.95$, starting from the initial state $Q_{ij}(0) = 100$ for all VOQs ij receiving traffic and a matching $M(0)$ of maximum size. Here $\alpha = 1$ so that the input and output selection probabilities of the scheduling scheme are proportional to the queue lengths. Packet sizes are assumed to be constant and equal to the size of switched data units in the synchronous case, with a centralized scheduling scheme that iteratively selects active edges with probabilities proportional to the queue lengths.

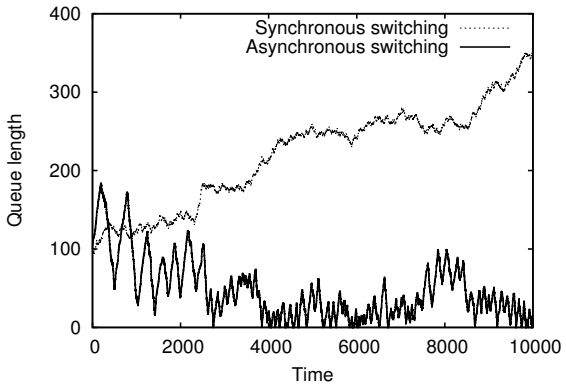
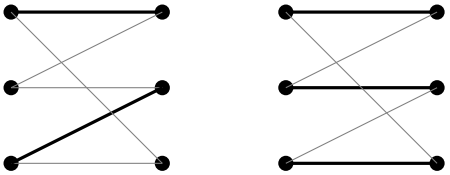


Figure 1: The stabilization effect of asynchrony.

We observe that, while the VOQs eventually empty in the asynchronous case, they grow continuously in the synchronous case. The reason is that, in the latter case, the scheduling scheme regularly selects suboptimal matchings of size $m = 2$ (refer to Figure 2). Specifically, such matchings are selected with probability $1/3$ when the lengths of the VOQs are equal and then limit the total service rate to $2/3$ of the maximum service capacity. We obtain a maximum load of $(1/3) \times (2/3) + (2/3) \times 1 = 8/9$, as observed in [6]. Since $\rho > 8/9$, the system is unstable.

In the asynchronous case, matchings of size $m = 2$ are *unstable* equilibriums: at the end of a packet transmission, the matching size increases to $m = 3$ whenever the input selects the other available output. In contrast, maximum matchings are *stable* equilibriums since packets are transmitted asynchronously: at the end of a packet transmission, the input cannot select another output, so that the same matching is used until one of the corresponding VOQs empties. In this sense, the scheduling scheme behaves as a random capture algorithm [7].



(a) $m = 2$ (b) $m = 3$

Figure 2: Maximal matchings of size m in the 3×3 switch with bidiagonal traffic pattern.

5. RATE OPTIMALITY

We now state and prove the main result of the paper, namely the rate optimality of the scheduling scheme described in Section 2 for asynchronous packet switches. We start with two monotonicity properties of the matching size. We denote by $E(t) = \{(i,j) \in E, Q_{ij}(t) > 0\}$ the set of active edges at time t.

LEMMA 1. *The matching size $|M(t)|$ is a non-decreasing function of time t over any time interval where $E(t)$ is constant.*

Proof. Since $E(t)$ is constant, ouput j can still be selected by input i at the end of the transmission of a packet from VOQ ij. □

LEMMA 2. *If the matching $M(t)$ is not maximum, the probability that $|M(t+1)| > |M(t)|$ is lower bounded by some constant which only depends on the ratios $Q_{ij}(t)/Q_{ik}(t)$, for all $(i,j),(i,k) \in E(t)$.*

Proof. Let $\lambda = \sum_{(i,j)\in E} \lambda_{ij}$, $\mu = \sum_{(i,j)\in E} \mu_{ij}$ and

$$\theta = \min_{(i,j)\in E(t)} \left(\frac{\mu_{ij}}{\mu} \frac{Q_{ij}(t)^\alpha}{\sum_{k:(i,k)\in E(t)} Q_{ik}(t)^\alpha} \right).$$

If the matching $M(t)$ is not maximum, there exists some augmenting path P such that the matching

$$M(t)\Delta P \equiv (M(t) \setminus P) \cup (P \setminus M(t))$$

has size $|M(t)| + 1$, see, e.g., [10]. The new matching is obtained by flipping the edge states (matched or unmatched) along path P. Since the length of any augmenting path is at most $2N$, this can be done in one time unit with probability at least:

$$((1 - e^{-\frac{\nu}{2N}})\theta)^{2N} \times e^{-(\lambda+\mu)}$$

with $\nu = \min_{(i,j)\in E} \mu_{ij}$, where the first factor accounts for the flipping of the edge states along path P, each over a time interval of duration $1/(2N)$, starting from the extreme edge of path P with matched input, and the second factor guarantees that no other event occurs. □

THEOREM 1. *The Markov process $(Q(t), M(t))$ is ergodic under condition* (1).

Proof. We prove the result by fluid limits, see, e.g. [4]. Consider a sequence of initial states $\{(Q^{(n)}(0), M^{(n)}(0))\}_{n\geq 1}$ such that:

$$\lim_{n\to\infty} \frac{Q^{(n)}(0)}{n} = \bar{Q}(0), \quad \text{with} \quad |\bar{Q}(0)| = 1.$$

Standard arguments based on the strong law of large numbers show the existence of an absolutely continuous, piecewise deterministic, stochastic process $\bar{Q}(t)$ on $\mathbb{R}_+^{N\times N}$, referred to as the *fluid limit*, such that, for any such sequence of initial states, there is some subsequence indexed by n_k such that, almost surely and uniformly on compact sets:

$$\frac{Q^{(n_k)}(n_k t)}{n_k} \to \bar{Q}(t) \quad \text{when} \quad k \to \infty.$$

The corresponding fluid model is *stable* if there exists some finite time t_0 such that $\bar{Q}(t) = 0$ a.s. for all $t \geq t_0$. This is a sufficient condition for the ergodicity of the Markov process $(Q(t), M(t))$, see [4, 14].

Denote by $\bar{E}(t) = \{(i,j) \in E, \bar{Q}_{ij}(t) > 0\}$ the set of active edges in the fluid model. Let $[u, v)$ be some time interval such that $\bar{E}(t)$ is constant and not empty. We denote this set by $\bar{E}$. By Lemmas 1 and 2, the fluid volumes are served by maximum matchings of the graph $\bar{G} = (V, \bar{E})$. Let m be the maximum matching size of $\bar{G}$ and define the workload of the fluid model at time t as:

$$W(t) = \sum_{i\in I, j\in J} \frac{1}{\mu_{ij}} \bar{Q}_{ij}(t).$$

Note that $\bar{Q}_{ij}(t) = 0$ for all $(i,j) \notin \bar{E}$ and $t \in [u, v)$ so that:

$$\begin{aligned} \forall t \in [u, v), \quad \frac{dW(t)}{dt} &= \sum_{(i,j)\in\bar{E}} \frac{1}{\mu_{ij}} \frac{d\bar{Q}_{ij}(t)}{dt}, \\ &= \left(\sum_{(i,j)\in\bar{E}} \rho_{ij} \right) - m. \end{aligned} \tag{2}$$

Now by König's theorem [10], the maximum matching size m is the minimum size of a vertex cover of $\bar{G}$ (set of vertices such that all edges of $\bar{E}$ are incident to at least one element of this set). Let $\bar{C}$ be a vertex cover of $\bar{G}$ of size m. We have:

$$\sum_{(i,j)\in\bar{E}} \rho_{ij} \leq \sum_{i\in\bar{C}\cap I} \sum_{j\in J} \rho_{ij} + \sum_{j\in\bar{C}\cap J} \sum_{i\in I} \rho_{ij} \leq m\delta, \tag{3}$$

with

$$\delta = \max \left\{ \max_{i\in I} \sum_{j\in J} \rho_{ij}, \max_{j\in J} \sum_{i\in I} \rho_{ij} \right\}.$$

The proof then follows from (2) and (3) on noting that $\delta < 1$ and $W(t) = 0$ if and only if $\bar{Q}(t) = 0$. □

6. DISCUSSION

The assumption of Poisson packet arrivals is reasonable if traffic is generated by a large number of flows, each with a low peak rate relative to the line card rates. In contrast, flows with high peak rates may be limited by the line card rates and thus generate bursty packet arrivals with state-dependent intensity due to the closed-loop control of TCP. The stability result easily extends to the case where each such flow maintains a constant number of packets in the queue, mimicking the behavior of TCP, until the completion of the data transfer. Such flows may, for instance, be assumed to arrive according to a Poisson process.

It proves more difficult to relax the assumption on the packet size distribution. In practice, this distribution is discrete with high weights on the minimum (40B) and maximum (around 1500B) sizes, which may synchronize packet departures. To keep rate optimality, it is then necessary to maintain asynchrony in the scheduling scheme by removing and adding edges successively, in some arbitrary order, as if packets would leave at different times. Note that the randomness in packet arrivals prevents the switch from the cyclic, pathological behavior described in [8].

In future work, we plan to study the packet delay performance of the proposed scheduling schemes, under various traffic patterns.

7. REFERENCES

[1] T. Anderson, S. Owicki, J. Saxe, and C. Thacker. High-speed switch scheduling for local-area networks. *ACM Transactions on Computer Systems*, 11(4):319–352, Nov. 1993.

[2] A. Bianco, D. Cuda, P. Giaccone, and F. Neri. Asynchronous vs synchronous input-queued switches. In *IEEE GLOBECOM*, 2010.

[3] A. Brzezinski and E. Modiano. Greedy weighted matching for scheduling the input-queued switch. In *CISS*, 2006.

[4] J. Dai. On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Annals of Applied Probabilities*, 5:49–77, 1995.

[5] J. Dai and B. Prabhakar. The throughput of data switches with and without speedup. In *IEEE INFOCOM*, 2000.

[6] A. Dimakis and J. Walrand. Sufficient conditions for stability of longest-queue-first scheduling: Second-order properties using fluid limits. *Advances in Applied Probability*, 38, 2006.

[7] M. Feuillet, A. Proutière, and P. Robert. Random capture algorithms: Fluid limits and stability. In *Information Theory and Applications Workshop*, 2010.

[8] Y. Ganjali, A. Keshavarzian, and D. Shah. Cell switching versus packet switching in input-queued switches. *IEEE/ACM Transactions on Networking*, 13(4):782–789, 2005.

[9] L. Jiang, D. Shah, J. Shin, and J. Walrand. Distributed random access algorithm: Scheduling and congestion control. *IEEE Transactions on Information Theory*, 56(12):6182 – 6207, 2010.

[10] L. Lovász and M. Plummer. *Matching theory*. North-Holland: Elsevier Science Publishers, 1986.

[11] S. T. Maguluri, B. Hajek, and R. Srikant. The stability of longest-queue-first scheduling with variable packet sizes. In *IEEE CDC*, 2011.

[12] N. McKeown. The iSLIP scheduling algorithm for input-queued switches. *IEEE/ACM Transactions on Networking*, 7(2):188–201, 1999.

[13] N. McKeown, A. Mekkittikul, V. Anantharam, and J. Walrand. Achieving 100% throughput in an input-queued switch. *IEEE Transactions on Communications*, 47(8):1260–1267, Aug 1999.

[14] P. Robert. *Stochastic networks and queues*. Springer, 2003.

[15] L. Tassiulas and A. Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE Transactions on Automatic Control*, 37:1936–1948, 1992.