

On the interaction between TCP-like sources and throughput-efficient scheduling policies

Technical report - Politecnico di Torino - July 2006

Paolo Giaccone, Emilio Leonardi, Fabio Neri
Dipartimento di Elettronica, Politecnico di Torino (Italy)

Abstract

We focus on the dynamic interaction in packet networks between regulated Additive-Increase Multiplicative-Decrease (AIMD) traffic sources and *max-scalar* scheduling policies (such as the popular Maximum Weight Matching – MWM) at switches. The latter were proved to be optimal in terms of throughput for stationary unregulated traffic sources.

We describe the average dynamics of both traffic sources and switch queues through a system of Delay Differential Equations (DDEs), whose properties are thoroughly analyzed. Our study allows to gain important insights both on the system efficiency and on the long-term bandwidth share among connections.

Our main finding is that AIMD sources and *max-scalar* switches co-exist well, despite several claims in the technical literature against the greedy behavior of *max-scalar* policies.

1 Introduction

In recent years a significant effort has been devoted by the networking research community to the definition of efficient scheduling policies that maximize the system throughput in several application contexts, such as wireless, satellite networks and high-capacity switching architectures [1, 4, 11, 18, 20, 21, 26, 30]. This problem dates back to the early '90, when Tassiulas and Ephremides, in their seminal work [28], have first shown that the maximization of throughput in a *network of interacting queues* (also called *constrained queueing systems*) can be achieved with a dynamic scheduling policy according to which the selection of packet transmissions, at servers, is driven by the instantaneous queues state.

It is worth noticing that the scheduling policy proposed in [28], i.e., the so-called *max-scalar policy*, and its later extensions [1, 11, 18, 20, 21, 30] (such as the popular Maximum Weight Matching – MWM), do not require any a priori knowledge of the long-term traffic behavior, thereby appearing amenable for implementation in contexts in which traffic is highly dynamic and unpredictable.

Optimality of the *max-scalar* policy and its extensions has been proved, however, only under assumptions of stationar-

ity and admissibility for the traffic flowing through the system of queues. It is not clear how optimal policies behave in the case of either non stationary, or rate-adaptive traffic sources, which may induce temporary overloads of some system architectural elements. The suspect that the *max-scalar* scheduling policy and its extensions may be strongly unfair in the latter case (several such claims appeared in the technical literature) has probably refrained a massive deployment of such policies in commercial systems.

Only very recently the attention has been turned to the analysis of the interaction between optimal dynamic policies and regulated traffic sources. Results in this field can have a great practical significance in consideration of the fact that the majority of Internet traffic sources adopt the Transmission Control Protocol (TCP) and dynamically adapt their sending rate to the estimated traffic congestion level according to an Additive-Increase Multiplicative-Decrease (AIMD) scheme.

In [12, 22] the behavior of max-scalar policies under regulated sources has been analyzed. However, the rate adaptation algorithms considered in [12, 22] significantly differ from the AIMD source behavior, since they require the sources to gather detailed and updated information about the network status.

We consider the dynamical behavior of max-scalar policies under rate-controlled sources executing an idealized TCP-like algorithm, driven only by losses and delay information as observed by the sources. The average dynamics of both sources and queues are described through a fluid model, i.e., a system of Delay Differential Equations (DDEs), whose qualitative properties are thoroughly analyzed. Our study allows to gain important insights both on the system efficiency and on the long-term bandwidth sharing among traffic flows.

Our findings are rather surprising and intriguing; the adoption of *max-scalar* scheduling policies along with carefully designed Active Queue Management (AQM) schemes permits to efficiently exploit the bandwidth of complex systems such as either Input Queued (IQ) switches or wireless cells, without negatively affecting the fairness of TCP flows.

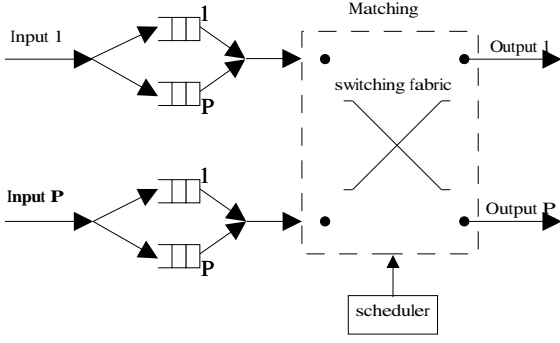


Figure 1: A $P \times P$ IQ switch architecture with VOQ

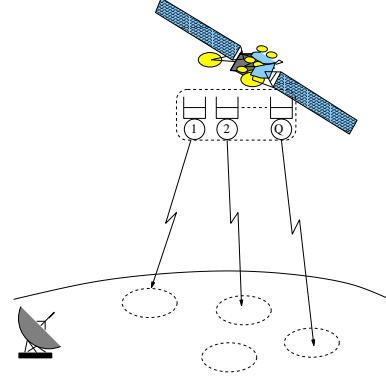


Figure 2: A wireless station

2 Systems of interacting queues

We consider a system of Q discrete time, interacting queues. This provides an abstract model for several different communication scenarios such as the system of transmission queues either at a wireless access point or a satellite, or the Virtual Output Queueing (VOQ) system in an IQ switch.

In discrete time, the queue evolution is described by:

$$x_q(n+1) = [x_q(n) + a_q(n) - \mu_q(n)]^+ \quad 1 \leq q \leq Q,$$

where $x_q(n)$ represents the queue length, $a_q(n)$ represents the number of arrivals at the queue, $\mu_q(n)$ represents the amount of service received by queue q during time $(n, n+1]$ and $[x]^+$ denotes $\max\{0, x\}$. These quantities can be expressed either in packets or in bytes. The set of achievable queue service rates is subject to a set of physical constraints, such as those expressing either capacity limits or the effects of possible interference between signals.

We formalize the previous concepts saying that the vector of service rates $\mu(n) = (\mu_1(n), \dots, \mu_Q(n))$ belongs to a convex set of achievable rates \mathcal{S} , i.e., $\mu(n) \in \mathcal{S}$ with $\mathcal{S} \subset \mathbb{R}_+^Q$, for every n .

We consider three possible application scenarios:

- **Work-conserving server:** in this first simple case, the bandwidth $\hat{\mu}$ of a work-conserving server is dynamically shared among Q queues. The sum of service rates allocated to queues is bounded by the transmission capacity $\hat{\mu}$ of the server: $\mathcal{S} = \{\mu : \mu_q \geq 0 \text{ and } \sum_q \mu_q \leq \hat{\mu}\}$.

We emphasize that this simple toy case has limited real interest, but its analysis can provide important insights on the behavior of more complex systems.

- **IQ switch:** Fig. 1 describes a $P \times P$ input queued switching architecture which represents the forwarding engine of a modern high-performance Internet router. Upon their arrival at the router, IP packets are chopped into fixed-sized cells and stored at input interfaces. Each interface maintains P separate queues, one per

output port ($Q = P^2$). Cells are independently transferred through the switching fabric toward the output ports where packets are reassembled. The switching fabric operates in a synchronous fashion. At time slot n , a set $\pi(n)$ of non contending cells, called *matching*, is selected for transfer through the switching fabric. Matchings can comprise no more that one cell per input port and no more than one cell per output port. Denoting with $IQ(i)$ ($1 \leq i \leq P$) the set of VOQs at input i , and with $OQ(j)$ ($1 \leq j \leq P$) the set of queues directed to output j , in order to be feasible (i.e. a matching), the service vector $\mu(n)$ must satisfy the following capacity constraints:

$$\begin{cases} \mu_q(n) \in \{0, 1\} & \forall q \\ \sum_{q \in IQ(i)} \mu_q(n) \leq 1 & 1 \leq i \leq P \\ \sum_{q \in OQ(j)} \mu_q(n) \leq 1 & 1 \leq j \leq P \end{cases}$$

Finally, we denote with \mathcal{S} the Convex Hull generated by the feasible service vectors, i.e., $\mathcal{S} = \{\mu : \mu_q \geq 0, \sum_{q \in IQ(i)} \mu_q \leq 1, \sum_{q \in OQ(j)} \mu_q \leq 1\}$

- **Wireless scenario:** it represents a multi-beam satellite which transmits data to Q different ground locations (as depicted in Fig. 2) or a terrestrial wireless station, such as a node of a large-bandwidth WiMAX mesh network. Packets destined for each location are stored in separate queues. In this case the transmission rate vector $\mu(n)$ depends on the coding scheme and on the power used for transmitting the signals. Using an access scheme which orthogonalizes transmitted signals, we can assume that the transmission rate μ_q depends only on the power P_q used for transmitting information from q [21]. We further assume $\mu_q(P_q)$ to be a regular concave function. In addition, we assume that the total transmission power P_{tot} is bounded.

Hence, the set of possible transmission rates \mathcal{S} is defined as the convex region defined by all the vectors $\mu = (\mu_1(P_1), \dots, \mu_q(P_q), \dots, \mu_Q(P_Q))$ being $\sum_q P_q \leq P_{tot}$.

The geometry of region \mathcal{S} depends on the specification of functions $\mu_q(P_q)$, which in turn depend on the physical layer specification [21]. In this paper, just as matter of example, we assume \mathcal{S} to be a circular region, defined by the following constraints:

$$\begin{cases} \mu_q(n) \geq 0 & \forall q \\ \sum_q \mu_q^2(n) \leq \hat{\mu} \end{cases} \quad (1)$$

being $\hat{\mu}$ a positive constant.

2.1 Max-Scalar policy definition

Under unregulated traffic sources, the problem of defining the optimal dynamic scheduling policy and the associated throughput region in complex systems of infinite-size interacting queues has attracted significant attention in the last decade from the research community since the pioneering work [28]. By assuming $a_q(n)$ to form a sequence of i.i.d. random variables, and applying the Lyapunov function methodology, it has been shown that a system of interacting queues achieve maximum throughput¹ if the max-scalar scheduling policy \mathcal{P}_{MS} is applied. According to \mathcal{P}_{MS} , at each time slot n , the service vector is selected as follows:

$$\mu(n) = \arg \max_{\gamma \in \mathcal{S}} \sum_{q=1}^Q \gamma_q x_q(n) \quad (2)$$

The result in [28] has been generalized and adapted to different application contexts in recent years. As matter of example, we just briefly recall some of the related works. In the packet switching context, several studies aimed at the definition of the stability region in IQ switching architectures built around a buffer-less crossbar have recently appeared: papers [1, 11, 18, 26, 30] have proposed different extensions of \mathcal{P}_{MS} , which have been shown to achieve the maximum throughput; stability properties for simpler scheduling policies have been also studied in [10, 30]; in [2, 4, 11], finally, the problem of the definition of the stability region in networks of IQ switches has been considered. In the context of satellite and wireless networks, generalizations of \mathcal{P}_{MS} have been recently proposed and shown to achieve the maximum throughput in [17, 20, 21, 29]. Finally, recently [9] generalized the result in [28] under more general exogenous arrival processes applying a different analytical technique called fluid models. All previous works, however, have considered unregulated stationary traffic sources.

2.2 \mathcal{P}_{MS} in three application scenarios

It is not difficult to particularize the policy \mathcal{P}_{MS} for the previous three scenarios:

¹A scheduling policy achieves maximum throughput if the system of queues is stable under any i.i.d. sequence $a_q(n)$ such that $(E[a_1(n)], E[a_2(n)], \dots, E[a_Q(n)]) \in \mathcal{S}$, i.e. the average arrival rates are within the set of admissible service rates.

- **Work-conserving server:** \mathcal{P}_{MS} selects simply the queue with the largest queue size i.e., this policy is usually referred as Longest Queue First (LQF) scheduling.
- **IQ switch:** \mathcal{P}_{MS} selects the matching $\pi(n)$ to maximize $\sum_{q \in \pi(n)} x_q(n)$, thus degenerates in the popular maximum weight matching.²
- **Wireless scenario:** Assuming \mathcal{S} to be defined according to (1), \mathcal{P}_{MS} selects the maximum service vector $\mu(n) \in \mathcal{S}$ parallel to the queue size vector $X(n)$:

$$\mu_q(n) = \frac{x_q(n)}{\|X(n)\|_2} \sqrt{\hat{\mu}} \quad (3)$$

being $\|X(n)\|_2$ the square norm of vector $X(n)$, i.e., $\|X(n)\|_2 = \sqrt{\sum_q x_q(n)^2}$.

2.3 \mathcal{P}_{MS} under regulated sources: previous work

Only very recently the attention has been turned to the analysis of the interactions between optimal dynamical policies and adaptive traffic sources. Papers [12, 22] have shown that \mathcal{P}_{MS} behaves well in presence of regulated sources, and guarantees an acceptable degree of fairness to flows also in case of temporary overload. However, these papers have focused on congestion control mechanisms significantly different from TCP, which require that traffic sources strictly interact with the network and gather detailed and up-dated information about the queues status.

Let $\rho_q(n)$ be the aggregate arrival rate at queue q , equal to the overall sending rate at the corresponding sources. In [12] sources were assumed to adjust $\rho_q(n)$ on the basis of the instantaneous queue size $x_q(n)$:

$$\rho_q(n) = \frac{\alpha_q K}{x_q(n) + \gamma_q}$$

being α_q , K and γ_q suitable positive constants. Similarly, in [22] the rate of sources must be dynamically adapted on the basis of instantaneous queues lengths. According to one of the proposals in [22]:

$$\rho_q(n) = \min \left\{ \left[\frac{V}{2x_q(n)} - 1 \right]^+, \rho_{max} \right\}$$

being V a control parameter and ρ_{max} the maximum allowed source sending rate.

In both cases above, sources must be made aware of queue sizes at switches.

3 Our system

We consider a TCP/IP infrastructure comprising a set of hosts interconnected through a network of switches/routers

²Note that, according to \mathcal{P}_{MS} , $\mu(n) = \arg \max_{\gamma \in \mathcal{S}} \sum_{q=1}^Q \gamma_q x_q(n)$ can always be selected to be an integer-valued vector; this is a consequence of the unimodularity of region \mathcal{S} .

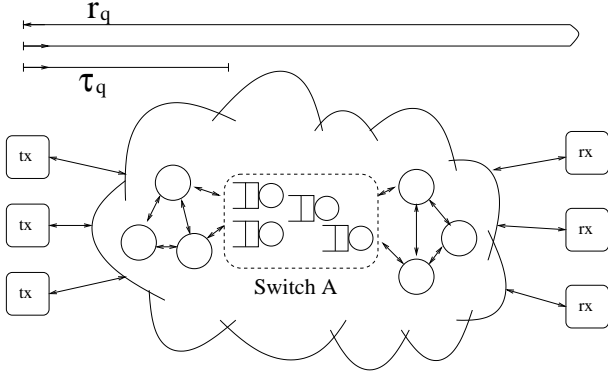


Figure 3: The system under study

as depicted in Fig. 3. In particular we focus on the network element identified in the figure as switch A. Switch A represents either an IQ switch or a wireless station implementing the max-scalar scheduling policy, and acts as bottle-neck for traffic flows traversing it (i.e., queuing and losses processes at switches/routers different from A, depicted as circles in Fig. 3, introduce negligible effects on TCP dynamics).

3.1 The system model

Aim of our work is to study the behavior of AIMD-based congestion control mechanisms in networks of interacting queues.

In our analysis, each queue q is fed with traffic originated in a set of M_q TCP sources. To study the interactions between sources and queues, we adopt a continuous time fluid approach [19] in which the average dynamics of both sources and queues are described by deterministic delay differential equations.

We assume that all the M_q TCP sources feeding queue q experience a constant round trip time r_q (see Fig. 3). The fluid evolution of the average window size $w_q(t)$ is driven by the classical, well-known AIMD fluid equation [19]:

$$\frac{dw_q(t)}{dt} = \frac{1}{r_q} - \frac{w_q(t)}{2} \phi_q(t) \quad (4)$$

where $\phi_q(t)$ represents the rate of congestion indications experienced at time t by sources. The first term on the right-hand side represents the additive increase mechanism, while the second term represents the multiplicative decrease contribution.

We denote with $W(t) = (w_1(t), w_2(t), \dots, w_Q(t))$ the vector whose elements represent the average transmitter window sizes (modeling TCP congestion windows) at time t for sources feeding queue q .

The fluid evolution of queue lengths $x_q(t)$ is driven by the

following equation:

$$\frac{dx_q(t)}{dt} = \left[\frac{M_q}{r_q} w_q(t - \tau_q) + \lambda_q(t) \right] (1 - d_q(t)) - \mu_q(t) 1_{\{x_q(t) > 0\}} \quad (5)$$

where the first term on the right represents the aggregate arrival rate at queue q ; being τ_q the average propagation delay between sources and queue q , $M_q w_q / r_q$ the overall average sending rate of the M_q sources, $\lambda_q(t)$ the aggregate arrival rate of unregulated traffic, and $d_q(t)$ the dropping probability at buffer q ; $\mu_q(t)$ is the service rate of queue q at time t . We denote with $X(t) = (x_1(t), x_2(t), \dots, x_Q(t))$ the vector whose elements represent queues lengths at time t . Furthermore, in the following denote with k_q the ratio M_q / r_q .

We suppose that each queue q implements a RED/ECN [27] AQM scheme, according to which packets are in general either dropped or marked. Both dropping probability $d_q(t)$ and marking probabilities $m_q(t)$ are driven by the buffer level. In particular, AQM schemes usually maintain a moving average \hat{x}_q of the instantaneous queue size x_q , updated whenever a packet arrives according to the rule:

$$\hat{x}_q \leftarrow (1 - z) \hat{x}_q + z x_q$$

The instantaneous mark/drop probability is computed as a function of \hat{x}_q according to some relation $d_q(t) = f_d(\hat{x}_q(t))$ and $m_q(t) = f_m(\hat{x}_q(t))$ (for example, $m_q(t) = 0$ in the case of a pure dropping policy). For fluid modeling, we need a characterization of the temporal evolution of the moving average $\hat{x}_q(t)$ as a continuous function of time. This was originally done in [19], where the authors have shown that the evolution is represented by the differential equation:

$$\frac{d\hat{x}_q(t)}{dt} = \frac{\log(1 - z)}{\delta(t)} \hat{x}_q(t) - \frac{\log(1 - z)}{\delta(t)} x_q(t) \quad (6)$$

if $z < 1$ and $\hat{x}_q(t) = x_q(t)$ if $z = 1$; $\delta(t)$ is the average packet inter-arrival time, i.e., $1/\delta(t) = k_q w_q(t - \tau_q) + \lambda_q(t)$.

We denote with $\hat{X}(t) = (\hat{x}_1(t), \hat{x}_2(t), \dots, \hat{x}_Q(t))$ the vector whose elements represent the moving average of the instantaneous queue size.

The rate of congestion indications $\phi_q(t)$ experienced by sources at time t is given by:

$$\phi_q(t) = \frac{w_q(t - r_q)}{r_q} \left[m_q(t - r_q + \tau_q) + d_q(t - r_q + \tau_q) \right] \quad (7)$$

where $\frac{w_q(t - r_q)}{r_q}$ is the packet sending rate of sources at time $t - r_q$ and $m_q(t - r_q + \tau_q) + d_q(t - r_q + \tau_q)$ is the sum of packet marking and dropping probabilities at time $t - r_q + \tau_q$.

Finally, according to the definition of \mathcal{P}_{MS} , queue service rates are determined in the fluid model as:

$$\mu(t) = \arg \max_{\gamma \in S} \sum_{q=1}^Q \gamma_q x_q(t) \quad (8)$$

being $\mu(t) = (\mu_1(t), \mu_2(t), \dots, \mu_Q(t))$ the fluid service vector, and \mathcal{S} the set of feasible service vectors.

Any functional vector $(X(t), \hat{X}(t), W(t))$ satisfying (4), (5), (6) and (8) represents a solution of the above dynamic system.

3.2 A critical discussion of the assumptions

In this sub-section, we critically discuss the assumptions and approximations of our model. First of all, we adopt a fluid approach to model both sources and queues dynamics. This is because several recent works have clearly shown that the fluid approach is a viable alternative to detailed packet-level simulations for the analysis of large-bandwidth TCP/IP networks [3, 7, 13, 14, 19]. Moreover, fluid models were proved to be effective for the parameter design of AQM/ECN schemes in TCP/IP networks [16]. Finally the system of fluid equations can be analytically approached, enlightening in a concise and elegant way important structural properties of the system dynamics [6, 23].

Since our goal is to analytically study the interaction between TCP sources and max-scalar scheduling at nodes, enlightening structural properties of the system as a whole, we have tried to simplify as much as possible the description of every architectural element. This is the reason why we have modeled just the basic AIMD mechanism of TCP, ignoring slow-start, time-outs, etc. We notice, however, that this basic description of idealized AIMD sources is usually able to capture the dominant dynamics of the system, providing fairly accurate results in several scenarios [7, 13].

We restrict our analysis to long-lived connections, neglecting short-lived connections. This is essentially due to the fact that short-lived flows can be assimilated in the fluid model to unregulated flows (term $\lambda_q(t)$ in (5)) as recently shown in [8, 15], since the effect of AIMD congestion control feedback is not effective due to their limited durations.

We have neglected the effects of queuing delay on the round-trip time, considering the latter to be constant. This assumption is justified by the fact that in the considered scenarios we expect the propagation delay to represent the dominant component of the round trip time. In addition, despite of the traditional rule of thumb according to which routers need a bandwidth-delay product of buffering in order to fully utilize the bottleneck link, recent studies have pointed-out that moderate-size buffer (two order of magnitude shorter) at routers lead to good performance [5, 24], significantly reducing router costs. Furthermore, AQM parameters are usually selected to make system buffers operating with a moderate number of packets [27].

Finally, in (4) and (7) we have implicitly assumed that all the M_q TCP sources feeding queue q experience the same propagation delay. This assumption can be relaxed to the case in which sources feeding queue q may experience different propagation delays, but the dispersion in their values is not too large. In this case r_q and τ_q appearing in (4) and (7)

can be reinterpreted in terms the average values. The model can be generalized when several classes of TCP source with significant different propagation delay coexist among the M_q sources feeding queue q . In this case an equation for each class of sources [13] has to be written. However in this paper we do not address this extension.

4 Qualitative study of the model solutions

Now, we characterize the qualitative properties of the model solutions of the above system of differential equations. Our study will be carried out using both analytical (where possible) and numerical tools.

First, we investigate on the existence of equilibrium points (i.e., stationary solutions) under the assumption of stationary traffic conditions, i.e. $\lambda_q(t) = \lambda_q \forall q$. We show that under mild assumptions a unique equilibrium point always exists. Then we turn our attention to the problem of the equilibrium point attractiveness (stability). In general we conjecture that, by carefully designing the AQM scheme equilibrium, attractiveness can be obtained. In simple cases we were able to analytically prove the equilibrium point attractiveness, while in more complex cases we report numerical results in support of our thesis.

We emphasize that a unique attractive equilibrium point unequivocally determines the long-term behavior (i.e., for $t \rightarrow \infty$) of system dynamics. As a consequence, by looking at the equilibrium point, we gain important insights on the system efficiency and long-term bandwidth share among connections, as shown in the next section.

4.1 System equilibrium point

The following statement fully characterizes the equilibrium points of our dynamical system.

Theorem 1 *Under the following assumptions: i) for every q the arrival unregulated traffic rate is stationary, i.e., $\lambda_q(t) = \lambda_q$; ii) \mathcal{S} is a convex compact set in \mathbb{R}_+^Q with non null interior; iii) $f_m(y)$ and $f_d(y)$ are non decreasing continuous and derivable functions; iv) for some finite B_q , $f_d(B_q) = 1$; v) $f(y) = f_m(y) + f_d(y)$ is strictly increasing for $0 \leq y \leq B_q$, with $f(0) = 0$; the system of differential equations (4), (5), (6) and (8) admits a unique stationary solution (X^*, \hat{X}^*, W^*) satisfying the following conditions:*

$$\mu^* = \arg \max_{\mu \in \mathcal{S}} G(\mu) \quad (9)$$

$$\hat{x}_q^* = x_q^* \quad (10)$$

$$x_q^* = f^{-1}(h_q(\mu_q^*)) \quad (11)$$

$$w_q^* = \frac{2}{\sqrt{f(x_q^*)}} \quad (12)$$

where

$$G(\mu) = \sum_q \int_0^{\mu_q} h_q(\alpha) d\alpha$$

and $h_q(\alpha)$ is the only positive solution of the equation:

$$\left(\frac{2k_q}{\sqrt{f(x_q^*)}} + \lambda_q \right) (1 - f_d(x_q^*)) = \alpha$$

The proof is given in Appendix A.

4.2 Stability analysis of the equilibrium point

Our claim is that, under reasonable traffic conditions, *the equilibrium point can be made attractive by carefully designing $f_d(x)$, $f_m(x)$ and z .*

Even if we are unable to provide a general formal proof of our claim, we report a wide range of partially numerical and analytical results in support of our thesis.

We start analyzing the simplified case in which delays to propagate packets from the sources to the queue in (5) are neglected along with delays to propagate congestion signals in (7). In this way the dynamical system described by (4), (5), (6) and (7) becomes a more treatable pure ordinary differential system (with no delays). Furthermore we assume $z = 1$; hence, $\hat{x}_q(t) = x_q(t)$ and system solutions are unequivocally determined by the vector $(X(t), W(t))$ describing queues and windows dynamics. Under these assumptions we are able to formally prove local stability of the equilibrium point in the three previously considered scenarios.

4.2.1 Work-conserving server

The local asymptotic stability of the equilibrium point can be proved, in this case, by using the Lyapunov function technique.

Suppose that the system is at $t = 0$ in an initial state $(X(0), W(0))$ sufficiently close to the equilibrium point (X^*, W^*) . We denote with $(X(t), W(t))$ the trajectory of the system and consider the following functional (Lyapunov function):

$$\mathcal{L}(X(t), W(t)) = \max_q (x_q - x_q^*)^2 + \beta \sum_q (w_q - w_q^*)^2$$

which represents a sort of “distance” between the current state and the equilibrium point. Note that by definition: i) $\mathcal{L}(X(t), W(t)) \geq 0$; ii) $\mathcal{L}(X(t), W(t)) = 0$, if and only if $(X(t), W(t)) = (X^*, W^*)$. Since $\frac{d\mathcal{L}(X(t), W(t))}{dt} < 0$, for almost every $t > 0$ (as discussed in Appendix B), we can conclude that the “distance” between the current system state and the equilibrium point is reducing with time (i.e., the trajectory gets closer and closer to the equilibrium point).

Beyond local stability, for different parameters settings, starting from 100 randomly chosen initial conditions, in all

cases we have numerically observed the convergence toward the equilibrium point of the solutions of the simplified (no delays) dynamical system of equations.

4.2.2 IQ switch

In case of the IQ switch, a formal proof can be done only in the special case of a 2×2 IQ switch by using the following Lyapunov function:

$$\mathcal{L}(X(t), W(t)) = \max_{\pi} (\omega_{\pi} - \omega^*)^2 + \beta \sum_q M_q (w_q - w^*)^2 + \gamma \sum_q (x_q - x^*)^2$$

and repeating arguments similar to the previous case. Also in this scenario, for several different parameters settings, starting from 100 randomly chosen initial conditions we have always observed the numerical convergence of the solutions toward the equilibrium point. The experiment was repeated both for 2×2 and 4×4 IQ switches.

4.2.3 Wireless scenario

In this case the local asymptotic stability of the system can be proved by linearizing the system around the equilibrium point and checking the stability of the linearized system. We report a sketch of the stability proof for the linearized system in Appendix C.

Also in this case, numerical experiments have shown that solutions converge to the equilibrium point starting from randomly chosen initial conditions, suggesting a global form of attractiveness for the equilibrium point.

4.3 Considering delays

When delays to propagate packets from sources to queues in (5), and to propagate congestion signals in (7), are considered, the problem of the definition of general conditions under which the equilibrium point is attractive becomes harder.

In the simple case in which TCP sources interact with a single FIFO server implementing an AQM scheme, sufficient conditions for local stability have been obtained linearizing the system fluid equations in [16]. As a result, paper [16] provides guidelines for the design of AQM parameters based upon simple relations to the physical parameters of the system such as the number of interacting TCP connections, the round-trip time and the capacity of the queue. In this paper, we propose to follow the guidelines specified in [16] for each queue q , then selecting for all queues the most conservative set of parameters found at the previous step.

Fig. 4 reports numerical results for the simple scenario in which a work conserving server manages two queues ($Q = 2$). Physical parameters are $M_1 = 20$, $M_2 = 40$, $r_1 = 20$ ms, $r_2 = 40$ ms, $\hat{\mu} = 100$ Mbit/s. The RED parameters, consistently with guidelines in [16], have been set to $min_th = 5$, $max_th = 50$, $p_max = 0.1$, $z = 10^{-4}$.

Four model trajectories corresponding to four initial conditions are plotted in Fig. 4. All trajectories converge to the equilibrium point.

We have checked the effectiveness of the proposed methodology in the three considered scenarios for 20 choices of physical parameters, and we always numerically verified the attractiveness of the equilibrium point.

5 System performance and fairness

In this section we explicitly characterize the equilibrium point for the three previously defined scenarios, analyzing system performance and fairness.

Before proceeding, however, we need to agree on an acceptable definition of fair bandwidth allocation to TCP flows. This choice is rather critical in light of the fact that no global consensus exists in the networking community on what a fair allocation is. Our opinion is that, in the Internet context, a simple First-In-First-Out (FIFO) server constitutes a good reference model for the distribution of bandwidth to connections. In a FIFO server, bandwidth is evenly distributed by the system to homogeneous connections (i.e., connections with the same round-trip time), while bandwidth is distributed among inhomogeneous connections (i.e., connections with different round-trip times) proportionally to the inverse of the connection round trip time, thereby achieving a rough form of proportional fairness.

In the following, we will qualify the bandwidth allocation provided by a FIFO server as the “fair allocation”.

5.1 Work-conserving server

In this simple case, it is rather straightforward to derive that the equilibrium point defined by (9)-(12) shows very surprising properties:

$$\begin{aligned} x_q^* &= x^* & w_q^* &= w^* & \forall q \\ \mu_q^* &= (k_q w^* + \lambda_d)(1 - f_d(x^*)) \end{aligned}$$

i.e., at the equilibrium queues have the same length, sources have the same average window size, and service is provided to regulated traffic aggregates proportionally to parameter ³ k_q .

The values for x^* and w^* can be explicitly computed; for example, in case $f_m(x_q) = 0$ and $\lambda_d = 0$; they are given by:

$$\begin{aligned} x^* &= f_d^{-1}\left(\frac{4 + \beta^2 - \sqrt{\beta^4 + 8\beta^2}}{4}\right) \\ w^* &= \frac{\beta + \sqrt{\beta^2 + 8}}{2} \end{aligned}$$

³The fact that all queues are of the same length at the equilibrium immediately derives from the fact that all service rates μ_q are different from 0 at the equilibrium (see proof of Theorem 1). As a consequence sources experience the same marking/loss probability, and thus, have the same window size.

where $\beta = \frac{\hat{\mu}}{\sum_q k_q}$.

The long-term throughput s_q of connections is determined by the parameters at the equilibrium point through the simple relation $s_q = w^*/r_q(1 - f_d(x^*))$; hence, the system bandwidth is distributed among connections proportionally to the inverse of the round-trip time, guaranteeing the same average share to homogeneous connections. *Thus, LQF provides the same long term bandwidth share among connections that we expect when adopting a conventional FIFO policy at the buffer!*

As final remark we notice that, since $x_q^* > 0$ (see proof of Theorem 1), it follows $\sum_q \mu_q^* = \hat{\mu}$, and consequently, the system is always able to efficiently exploit the available bandwidth.

5.2 IQ switch

In this case, the analytical characterization of the equilibrium point requires the solution of a system of $2P^2 + 2P - 1$ non-linear equations. To simplify the analysis, we assume that all the Virtual Output Queues are fed by some regulated traffic sources (i.e., for every q , $k_q > 0$); we however emphasize that the analysis can be easily extended to the more general case.

First we point out that the equilibrium point must satisfy the following important property: *all the $\omega(\pi)$ weights, associated with all $P!$ possible matchings π , have the same value ω^* at the equilibrium, i.e.:*

$$\omega(\pi) = \sum_{q \in \pi} x_q^* = \omega^* \quad \forall \pi$$

This property generalizes the property exhibited by LQF in the work-conserving queue case.

As a consequence, X^* lies in the span of $\mathcal{M} = \{I^1, I^2, \dots, I^P, O^1, O^2, \dots, O^P\}$, being I^i a vector whose element i_q^p is one if $q \in IQ(i)$ and null otherwise; and O^j be a vector whose element o_q^j is one if $q \in OQ(j)$, and null otherwise. Dimension of $\text{span}(\mathcal{M})$ is $2P - 1$, hence X^* can be expressed as a linear combination of $2P - 1$ vectors selected within \mathcal{M} . Choosing the first $2P - 1$ vectors in \mathcal{M} , we can write:

$$X = \sum_{p=1}^P \alpha_p I^p + \sum_{p=1}^{P-1} \beta_p O^p \quad (13)$$

for some positive values of the parameters α_p and β_p .

On the other hand rates μ_q^* and queue size x_q are deterministically related by the following systems of non linear equations:

$$\left(\frac{2k_q}{\sqrt{f(x_q^*)}} + \lambda_d \right) (1 - f_d(x_q^*)) = \mu_q^* \quad 1 \leq q \leq Q \quad (14)$$

We notice that, since at the equilibrium point all the queues are non empty (i.e., $x_q^* > 0, \forall q$), the service rate vector at the equilibrium maximizes the global throughput i.e.,

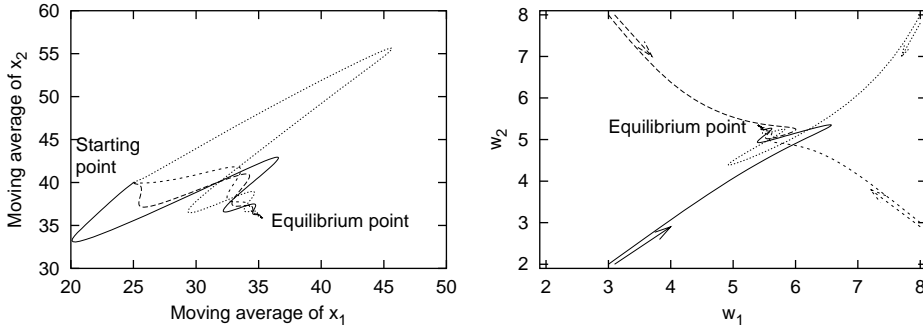


Figure 4: A work conserving server managing two queues; (on the left) trajectories projections on the coordinate plain representing the moving averages of queue lengths, (on the right) trajectories projections on the coordinate plane representing the window sizes.

$\sum_{q \in IQ(i)} \mu_q(n) = 1$ and $\sum_{q \in OQ(j)} \mu_q(n) = 1$ for every input i and output j .

The properties of the equilibrium point suggest that also in this case the system tends to evenly distributing the bandwidth among homogeneous TCP connections (at the equilibrium), while it tends to distribute the bandwidth among inhomogeneous connections proportionally to the inverse of their round-trip delay. This feeling is confirmed by our numerical experiments; we have focused on a 4×4 IQ switch, loaded with unbalanced traffic patterns. Table 1 describes the first family of considered traffic patterns; the numbers of connections feeding VOQs, M_q , are reported on the left while the round trip times r_q are reported on the right. We have tried several cases for different values of α and β ranging in the interval $[1, 3]$. In all cases the numerical results confirm that relative bandwidth obtained by connections at the equilibrium is perfectly proportional to $1/\beta$.

We emphasize that not always a perfectly “fair” (in the previously specified sense) distribution of the bandwidth is achieved in IQ switches. Bandwidth shares among TCP flows deviate from the “fair” distribution when traffic asymmetries among inputs or outputs ports are established (note that previous traffic patterns were completely symmetrical with respects to both inputs and outputs ports). We notice that, in the latter cases, forcing a “fair” distribution of bandwidth among TCP flows would cause a not complete exploitation of the switch bandwidth.

To better understand the behavior of the max-scalar policy, consider a traffic scenario comprising homogeneous TCP flows distributed according to Table 2 (on the left). The throughputs obtained by connections are reported in Table 2 (on the right). We notice that flows traversing input port 4 are penalized in throughput with respect to other flows; this effect however has an easy explanation: connections traversing input 4 are bottlenecked at the input port where they obtain the maximum possible “fair” share. All the other connections evenly share the residual switch bandwidth. Note that a perfectly “fair” distribution of bandwidths is possible only

at the cost of reducing the throughput of some connections (those not traversing input 4) without any beneficial effect on the other connections (those traversing input 4). Thus the system, in this case, distributes bandwidth according to a max-min scheme.

Several other numerical experiments (whose results are not reported for brevity) have confirmed that the max-scalar policy at the equilibrium distributes bandwidth among connections according to a weighed max-min fairness scheme in which connection weights are proportional to the inverse of the connection round-trip time.

5.3 Wireless scenario

In this case, as already observed, queue services are provided by the max-scalar proportionally to the queue lengths; at the equilibrium:

$$\mu_q(t) = \frac{x_q^*}{\|X^*\|_2} \hat{\mu}$$

The quantities x_q^* and w_q^* can be obtained solving the following system of $2q$ nonlinear equations, $\forall q$:

$$(k_q w_q^* + \lambda_q)(1 - f_d(x_q^*)) = \frac{x_q^*}{\|X^*\|_2} \hat{\mu} \quad (15)$$

$$w_q^* = \sqrt{\frac{2}{f(x_q^*)}} \quad (16)$$

Note that, differently from what happens for a work-conserving server, in this scenario the rule according to which bandwidth subdivided among connection aggregates have an impact on the global system throughput. This is due to the fact that vectors which lie on the boundary of region \mathcal{S} (i.e., vectors satisfying $\sum_q \mu_q^2 = \hat{\mu}$) do not correspond to the same global system throughput $S = \sum_q \mu_q$. Maximum system throughput is achieved when all the aggregates receive the same bandwidth ($\mu_q = \mu$ for every q), irrespectively of their parameters M_q and r_q . We notice, however, that in this case a significantly unfair distribution of throughputs to individual connections may result when parameters M_q, r_q or

I/O	1	2	3	4
1	M_0	$M_0\alpha$	$M_0\alpha^2$	$M_0\alpha^3$
2	$M_0\alpha^3$	M_0	$M_0\alpha$	$M_0\alpha^2$
3	$M_0\alpha^2$	$M_0\alpha^3$	M_0	$M_0\alpha$
4	$M_0\alpha$	$M_0\alpha^2$	$M_0\alpha^3$	M_0

I/O	1	2	3	4
1	r_0	$r_0\beta$	$r_0\beta^2$	$r_0\beta^3$
2	$r_0\beta^3$	r_0	$r_0\beta$	$r_0\beta^2$
3	$r_0\beta^2$	$r_0\beta^3$	r_0	$r_0\beta$
4	$r_0\beta$	$r_0\beta^2$	$r_0\beta^3$	r_0

Table 1: Traffic pattern: distribution of the number of connections at VOQs (on the left), distribution of overage round-trip times (on the right)

I/O	1	2	3	4
1	M_0	$M_0\alpha$	$M_0\alpha^2$	$M_0\alpha^3$
2	$M_0\alpha^3$	M_0	$M_0\alpha$	$M_0\alpha^2$
3	$M_0\alpha^2$	$M_0\alpha^3$	M_0	$M_0\alpha$
4	$\gamma M_0\alpha$	$\gamma M_0\alpha^2$	$\gamma M_0\alpha^3$	γM_0

I/O	1	2	3	4
1	ζ	ζ	ζ	ζ
2	ζ	ζ	ζ	ζ
3	ζ	ζ	ζ	ζ
4	ζ/γ	ζ/γ	ζ/γ	ζ/γ

Table 2: The distribution of the number of connections at VOQs (on the left); the connection throughputs (on the right); being $\gamma > 1$ and $\zeta = \frac{(\alpha - 1)}{M_0(\alpha^4 - 1)}$

λ_q are strongly unbalanced. On the contrary the system that fairly distributes bandwidth to connections (i.e. obeys to the law $\frac{\mu_q - \lambda_q}{\mu_{q'} - \lambda_{q'}} = k_q/k_{q'}$, $\forall q, q'$) may be significantly inefficient in terms of system throughput.

We consider a wireless station hosting two queues (i.e., $Q = 2$), to ease the graphic presentation of numerical results. However all considerations apply to the more general case $Q > 2$.

First we consider the case in which the wireless server is fed by regulated traffic, only i.e., $\lambda_1 = \lambda_2 = 0$, and we focus on the fairness index η defined as the ratio between throughput-delay products of TCP flows belonging to first and the second aggregate $\eta = r_1 s_1 / r_2 s_2$. Note that $\eta = 1$ corresponds to a “fair” distribution of bandwidth among TCP flows. In Fig. 5 η (on the right) is reported as function of the ratio M_1/M_2 , for different values of the ratio r_1/r_2 . To better understand the behavior of the system, the ratio between the bandwidths obtained by aggregates $\chi = \mu_1^*/\mu_2^*$ is also reported in Fig. 5 (on the left)⁴.

Bandwidth is distributed between aggregates in a rather complex way, in this case; an aggregate gets more and more bandwidth by the system when the relative number of its flows increases. However since χ exhibits a sub-linear dependence from parameter M_1/M_2 , flows belonging to more numerous aggregates are penalized with respect to flows belonging to less numerous aggregates, as shown in Fig. 5. Bandwidth shares obtained by connections depends also on round trip-time: larger bandwidth are obtained by aggregates of connections with shorter round-trip times. However, since the flow bandwidth share increases sub-linearly with the inverse of round-trip time, flows with shorter round-trip get less than their “fair” share.

We consider now the case in which unregulated traffic ar-

⁴Plots in Figs. 5 and 6 are obtained for $\sqrt{\mu} = 50$ Mbit/s, $M_1 = 10$ and $r_1 = 20$ ms

rives at queue 2 ($\lambda_2 = \sqrt{\mu}/3$). Fig. 6 reports χ and η . With respect to the previous case, the presence of unregulated traffic at queue 2 induces a moderate perturbation of the relative connections share, penalizing flows belonging to aggregate 2.

In conclusion, in the wireless scenario the max-scalar policy exhibits a complex behavior reaching an operational point which is middle way between the maximum throughput operational point ($\mu_1 = \mu_2$) and the point corresponding to a fair distribution of bandwidth to connections ($\frac{m\mu_1 - \lambda_1}{m\mu_2 - \lambda_2} = k_1/k_2$). So doing, a reasonable trade off is achieved between the conflicting requirements of optimizing global performance (system throughput) and fairly distributing bandwidths among TCP flows.

6 Validation of the model

As final step we validate the prediction of our model against packet level simulations. To this end, we developed in OMNeT++[25] modules representing systems of queues implementing the *max-scalar* scheduling policy. In simulations, queues have been fed by standard TCP new-Reno sources.

First, we consider a simple scenario comprising a 1 Gbit/s work-conserving server managing three queues ($Q = 3$), and fed by 420 TCP flows. These TCP flows are distributed among the three queues as follows: $M_1 = 240$, $M_2 = 120$, $M_3 = 60$. Round-trip times of TCP flows have been randomly chosen according to a uniform distribution with support in the interval [26, 30] ms. A RED AQM mechanism is implemented at the queues ($min_th = 10$, $max_th = 500$, $p_max = 0.05$, $z = 10^{-5}$). A comparison between model predictions (labeled TEO) and simulation results (labeled SIM) is reported in Fig. 7, where the bandwidth shares evolution for the three traffic aggregates are plotted. Simula-

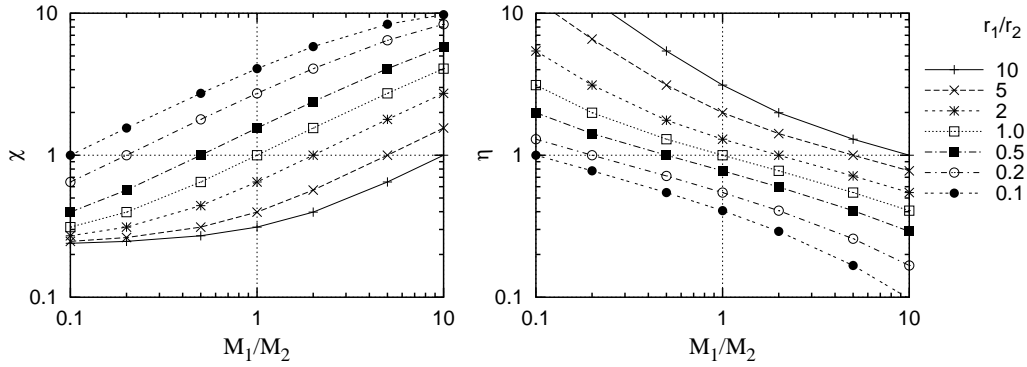


Figure 5: Wireless scenario with TCP traffic only: χ (on the left) and η (on the right) Vs M_1/M_2 , for different values of the ratio r_1/r_2 .

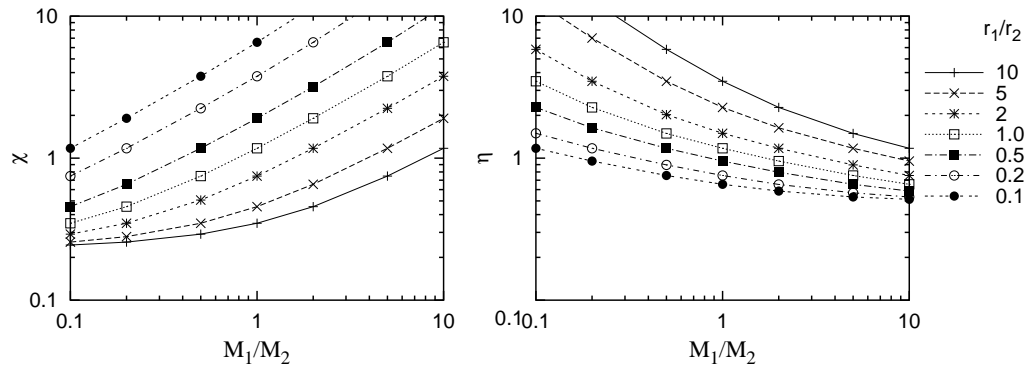


Figure 6: Wireless scenario with unregulated traffic ($\lambda_2 = \sqrt{\mu}/3$): χ (on the left) and η (on the right) Vs M_1/M_2 , for different values of the ratio r_1/r_2 .

tion points are obtained by averaging the bandwidth obtained by aggregates within 200 ms windows. The good agreement between model predictions and simulation results confirms that bandwidth shares obtained aggregates are roughly proportional to parameters M_q .

As second scenario, we consider a case in which all server queues are fed by the same number of TCP flows (140); however flows traversing different queues have different round-trip times. Round-trip times of flows traversing queue 1 are uniformly distributed in the interval [18, 22] ms; those traversing queue 2 are distributed in the interval [36, 44] ms; those traversing queue 3 are distributed in the interval [54, 66] ms. The same parameters of the previous scenario were used to tune the RED mechanism. Also in this case, a good agreement between model predictions and simulation is shown in Fig. 8. The server bandwidth is distributed to the aggregates proportionally to the inverse of their round-trip times.

As third scenario we have considered a 4×4 IQ switch with ports running at 1 Gbit/s. The analyzed traffic scenario, already considered in the previous section, is represented in Table 1 where $M_0 = 28$, $\alpha = 2$, $\beta = 1$ and $r_0 = 28$ ms. The

I/O	1	2	3	4
1	73	141	268	518
2	524	71	139	266
3	270	531	67	132
4	133	257	526	84

Table 3: Scenario 3: Bandwidth shares in Mbit/s among TCP aggregates (in $MBit/s$), averaged over interval [20, 35] s

same RED parameters of the previous scenarios were used. Table 3 shows the average bandwidth shares obtained by aggregates traversing the switch. The results confirm that IQ switch bandwidth is efficiently exploited by the max-scalar policy as predicted by the model; moreover the long-term bandwidth shares are almost exactly proportional to M_q .

At last, we emphasize that we have validated the model predictions against simulations in several other scenarios, which are not described in details for brevity. In all cases the simulation results have confirmed model predictions.

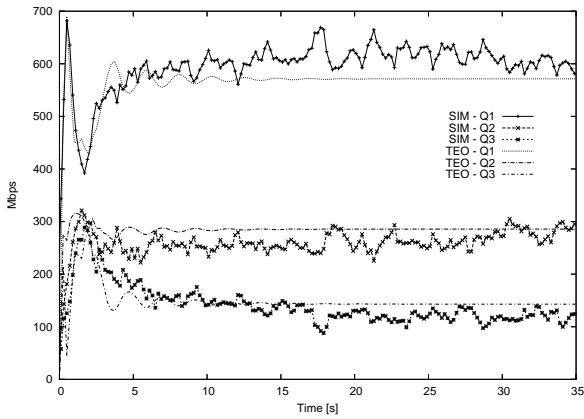


Figure 7: Scenario 1: bandwidth shares among TCP aggregates. Bandwidth obtained by aggregates, averaged over interval [20, 35] s, are 604, 271 and 125 Mbit/s, respectively.

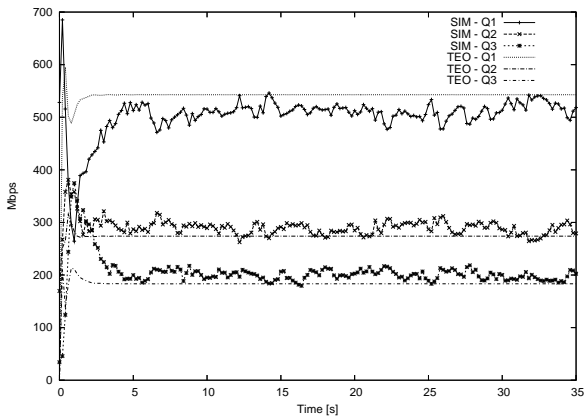


Figure 8: Scenario 2: bandwidth shares among TCP aggregates. Bandwidth obtained by aggregates, averaged over interval [20, 35] s, are 530, 282 and 188 Mbit/s, respectively.

7 Conclusions

Max-scalar scheduling policies have been proposed to optimize the global system performance in several application contexts such as wireless networks, satellite networks and high-capacity router architectures.

Optimality of such scheduling policies was proved, however, only under assumptions of stationarity and admissibility for the traffic flowing through the system of queues. It is unclear how they behave in the case of either non stationary, or rate-adaptive traffic sources, that may induce temporary overloads of some system architectural elements.

In this paper we investigated how max-scalar scheduling policies behave under TCP traffic sources. To this end, we have described the average dynamics of both traffic sources and switch queues through a system of Delay Differential Equations (DDEs), whose properties were thoroughly ana-

lyzed.

Our findings were rather surprising and intriguing; the adoption of max-scalar scheduling policies along with carefully designed AQM schemes permits to efficiently exploit the bandwidth of complex systems such as either IQ switches or wireless stations without negatively affecting the fairness of TCP flows.

We recognize that research on max-scalar scheduling policies has been driven so far mainly by speculative interest; being such policies largely ignored in products implementations. Still, we believe that max-scalar scheduling policies offer interesting potentialities in application contexts where bandwidth over-provisioning has significant costs (like in wireless networks). For these reasons the results of our investigation, shedding some light on important aspects that are often neglected, can stimulate a debate on the implementability of such scheduling policies at nodes.

References

- [1] M. Ajmone Marsan, A. Bianco, P. Giaccone, E. Leonardi, F. Neri, "Packet-Mode Scheduling in Input-Queued Cell-Based Switch", *IEEE/ACM Transactions on Networking*, vol. 10, n. 5, Oct. 2002, pp. 666-678
- [2] M. Ajmone Marsan, P. Giaccone, E. Leonardi, F. Neri, "On the stability of local scheduling policies in networks of packet switches with input queues", *IEEE Journal on Selected Areas in Communications*, vol. 21, n. 4, May 2003, pp. 642-655
- [3] M. Ajmone Marsan, M. Franceschinis, P. Giaccone, E. Leonardi, E. Schiattarella, A. Tarello, "Using Partial Differential Equations to Model TCP Mice and Elephants in Large IP Networks", *IEEE Transactions on Networking*, vol. 13, n. 6, pp. 1289-1301
- [4] M. Andrews, L. Zhang, "Achieving Stability in Networks of Input-Queued Switches", *IEEE Infocom 2001*, Anchorage, Alaska, USA, Apr. 2001
- [5] G. Appenzeller, I. Keslassy, N. McKeown, "Sizing router buffers", *ACM Sigcomm'04*, Portland, OR, USA, August 2004
- [6] F. Baccelli, D. Hong, "Interaction of TCP Flows as Billiards", *IEEE Infocom 03*, San Francisco, CA (USA), March 2003
- [7] F. Baccelli, D. Hong, "Flow Level Simulation of Large IP Networks", *IEEE Infocom 2003*, San Francisco, CA (USA), March 2003
- [8] C. Barakat, P. Thiram, G.F. Iannaccone, C. Diot "Modelling Internet Backbone Traffic at Flow Level", *IEEE Transactions on Signal Processing*, Vol 51, n. 8, Aug. 2003

- [9] J. G. Dai, W. Lin, “Maximum Pressure Policies in Stochastic Processing Networks”, *Operations Research*, vol. 53, 2005, pp. 197-218
- [10] J.G. Dai, B. Prabhakar, “The throughput of data switches with and without speedup”, *IEEE Infocom 2000*, Tel Aviv, Israel, Mar. 2000, pp. 556-564
- [11] E. Leonardi, M. Mellia, M. Ajmone Marsan, F. Neri, “On the Throughput Achievable by Isolated and Interconnected Input-Queueing Switches under Multiclass Traffic”, *IEEE Infocom 2002*, New York, NY (USA), June 2002
- [12] A. Eryilmaz, R. Srikant, “Fair resource allocation in wireless networks using queue-length-based scheduling and congestion control”, *IEEE Infocom 2005*, March 2005, Miami, FL (USA), March 2005
- [13] Y. Gu, Y. Liu, F. Lo Presti, V. Misra, D. Towsley, “Fluid Models and Solutions for Large-Scale IP Networks”, *ACM Sigmetrics 03*, San Diego, CA (USA), June 2003
- [14] Y. Gu, Y. Liu, D. Towsley, “On Integrating Fluid Models with Packet Simulation”, *IEEE Infocom 04*, Hong Kong, China, March 2004
- [15] C.V. Hollot, Y. Liu, V. Misra, D. Towsley, “Unresponsive flows and AQM performance,” *IEEE Infocom 2003*, San Francisco, CA (USA), March 2003
- [16] C.V. Hollot, V. Misra, D. Towsley, W.B. Gong, “On the Design of Improved Controllers for AQM Routers Supporting TCP Flows”, *IEEE Infocom 01*, Anchorage, Alaska (USA), April 2001
- [17] X. Lin, N.B. Shroff, “The Impact of Imperfect Scheduling on Cross-Layer Rate Control in Wireless Networks”, *IEEE Infocom 2005*, Miami, FL (USA), March 2003
- [18] N. McKeown, A. Mekkittikul, V. Anantharam, J. Walrand, “Achieving 100% throughput in an input-queued switch”, *IEEE Trans. on Communications*, vol. 47, n. 8, Aug. 1999, pp. 1260-1272
- [19] S. Misra, W.B. Gong, D. Towsley, “Fluid-Based Analysis of a Network of AQM Routers Supporting TCP Flows with an Application to RED”, *ACM Sigcomm’00*, Stockholm, Sweden, August 2000
- [20] M.J. Neely, E. Modiano, C.E. Rohrs, “Dynamic power allocation and routing for time varying wireless networks”, *IEEE Infocom 2003*, San Francisco, CA (USA), Mar. 2003
- [21] M.L. Neely, E. Modiano, C.E. Rohrs “Power Allocation and Routing in Multibeam Satellites with Time-Varying Channels”, *IEEE/ACM Trans. on Networking*, vol. 11, n. 1, Feb. 2003, pp. 138-152
- [22] M.J. Neely, E. Modiano, Li Chih-Ping “Fairness and optimal stochastic control for heterogeneous networks”, *IEEE Infocom 2005*, Miami, FL (USA), March 2005
- [23] R. Pan, B. Prabhakar, K. Psounis, D. Wischik, “SHRiNK: A method for scaleable performance prediction and efficient network simulation”, *IEEE Infocom 2003*, San Francisco, CA (USA), March 2003
- [24] N. McKeown, D. Wischik, et al., “Making router buffers much smaller (Parts I , II and III)”, *ACM SIGCOMM Computer Communication Review*, vol. 35, n. 3, July 2005, pp.73-89
- [25] “OMNeT ++ Discrete Event Simulation System”, available at <http://www.omnetpp.org>
- [26] K. Ross, N. Bambos, “Local Search Scheduling Algorithms for Maximal Throughput in Packet Switches”, *IEEE Infocom 2004*, Hong Kong, China, Mar. 2004
- [27] S. Floyd, V. Jacobson, “Random Early Detection Gateways for Congestion Avoidance”, *IEEE/ACM Transactions on Networking*, vol. 1, n. 4, Aug. 1993, pp. 397-413
- [28] L. Tassiulas, A. Ephremides, “Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks”, *IEEE Trans. on Automatic Control*, vol. 37, n. 12, Dec. 1992, pp. 1936-1948
- [29] L. Tassiulas, “Scheduling and performance limits of networks with constantly changing topology”, *IEEE Transactions on Information Theory*, vol. 43, n. 3, May 1997, pp. 1067-1073
- [30] L. Tassiulas, “Linear complexity algorithms for maximum throughput in radio networks and input queued switches”, *IEEE Infocom 1998*, San Francisco, CA, USA, Apr. 1998

A Proof of Theorem 1

The equilibrium point, (X^*, \hat{X}^*, W^*) , by definition, satisfies the set of algebraic equations obtained by: (4), (5), (6), (8) setting to zero the time derivatives. As a consequence, (X^*, \hat{X}^*, W^*) is an equilibrium point of the dynamical system if it satisfies:

$$(k_q w_q^* + \lambda_q)(1 - f_d(x_q^*)) = \mu_q^* 1_{\{x_q^* > 0\}} \quad (17)$$

$$\hat{x}_q^* = x_q^* \quad (18)$$

$$\frac{1}{r_q} = \frac{(w_q^*)^2}{2r_q} f_q(x_q^*) \quad (19)$$

$$\mu^* = \arg \max_{\alpha \in \mathcal{S}} \sum \alpha_q x_q^* \quad (20)$$

Note that, at the equilibrium point, for every q , it must be $f_q(x_q^*) > 0$, otherwise (19) can not be satisfied; as a consequence necessarily from assumption iv) it results $x_q^* > 0$ and thus (17) can be simplified since $\mathbb{1}_{\{x_q^* > 0\}} = 1$.

Observe that (20) can be rewritten in the following way:

$$\sum_q (\mu_q^* - \eta_q) x_q^* \leq 0 \quad \forall \eta \in \mathcal{S} \quad (21)$$

Now (17) and (19) allow to algebraically relate μ_q^* and x_q^* , indeed from (19):

$$w_q^* = \frac{2}{\sqrt{f(x_q^*)}}$$

and thus substituting in (17) we obtain:

$$\left(\frac{2k_q}{\sqrt{f(x_q^*)}} + \lambda_q \right) (1 - f_d(x_q^*)) = \mu_q^* \quad (22)$$

the function the left hand side is weakly decreasing with respect to its argument, $x_q^* \geq 0$, moreover for $x_q^* \rightarrow 0$ it tends to $+\infty$, while for $x_q^* = B_q$, by assumption v), it is null. As a consequence for every value of μ_q^* , (22) always admits one and only one solution in x_q^* . We denote with $h_q(\mu_q^*)$ this solution. By construction $x_q^* = h_q(\mu_q^*)$, and $0 < h_q(\mu_q^*) \leq B_q$. At last, $h_q(\mu_q^*)$ is not increasing with its argument.

From above calculations, μ^* satisfies:

$$\sum (\mu_q^* - \eta_q) h_q(\mu_q^*) \leq 0 \quad \forall \eta \in \mathcal{S}$$

Define:

$$G(\mu) = \sum_q \int_0^{\mu_q} h_q(x) dx$$

where $G(\mu)$ by construction is a concave function as it can be easily verified by computing its Hessian (we recall that $h_q(\cdot)$ is a weakly decreasing function, and from assumption iii) it is derivable). Thus condition (21) can be rewritten as:

$$(\eta - \mu^*) \nabla G(\mu^*) \leq 0 \quad \forall \eta \in \mathcal{S} \quad (23)$$

We conclude our proof, invoking the following lemma.

Lemma 1 *In order to satisfy condition (23) μ^* must be the only solution of the following optimization problem:*

$$\mu^* = \arg \max_{\mu \in \mathcal{S}} G(\mu) \quad (24)$$

First observe that since $G(\mu)$ is concave and \mathcal{S} is compact and convex (24) admits one and only one solution.

Second we prove the satisfaction of condition (24) provides a necessary condition for the satisfaction of (23). Indeed assume (μ^*) does not satisfy (24), then there exists $\eta \in \mathcal{S}$ such that:

$$G(\eta) > G(\mu^*)$$

However due to the concavity of $G(\mu)$:

$$G(\eta) \leq G(\mu^*) + (\eta - \mu^*) \nabla G(\mu^*)$$

from which $(\eta - \mu^*) \nabla G(\mu^*) > 0$.

Finally the fact the the solution of (24) also satisfies (23) is an immediate consequence of the Karush Kuhn-Tucker conditions.

B Stability of the equilibrium point: work conserving server

In this appendix we prove that $\frac{d\mathcal{L}(X(t), W(t))}{dt} < 0$ almost for every $t > 0$. To simplify the calculations we suppose that the AQM scheme adopts a pure marking policy at equilibrium. However the arguments reported in this proof can be extended in a straightforward way to the more general case.

Since $x_q(t)$ and $w_q(t)$ are by definition absolutely continuous functions, $\mathcal{L}(X(t), W(t)) = \max_q (x_q(t) - x^*)^2 + \beta \sum_q M_q (w_q(t) - w^*)^2$ is an absolutely continuous function, and thus it is derivable almost for every $t > 0$. In the following we will show that whenever $\frac{d\mathcal{L}(X(t), W(t))}{dt}$ exists, it is negative.

We start by recalling some results from Section 5.1 which will be useful in this proof: at the equilibrium point every queue has the same length length, and every source have the same average window size, i.e., $\forall q, x_q^* = x^*$ and $w_q^* = w^*$; moreover $\mu_q^* = (\frac{M_q}{r_q} w^* + \lambda_q) (1 - f_d(x^*))$. As a consequence, $\mu_q^* > 0$ for every q . We denote with μ_{\min} the minimum achieved rate at the equilibrium: $\mu_{\min} = \min_q \mu_q^*$.

Consider time t , and suppose $\max_q |x_q(t) - x^*| < \delta$ and $\max_q |w_q(t) - w^*| < \delta$ for some strictly positive δ . Let \mathcal{Q}_0 be the set of queues which are at maximum distance from the equilibrium point at time t , i.e., $\mathcal{Q}_0 = \arg \max_q (x_q - x^*)^2$, first we assume that $|\mathcal{Q}_0| < Q$. If $|\mathcal{Q}_0| = 1$, no problems of derivability for $\mathcal{L}(X(t), W(t))$ arise, however if $|\mathcal{Q}_0| > 1$, $\mathcal{L}(X(t), W(t))$ is not guaranteed to be derivable at t . A sufficient and necessary condition for derivability is that: i) all $x_{q_0}(t)$ in \mathcal{Q}_0 have the same length i.e., $x_{q_0}(t) = x_{q'_0}(t)$ for every q_0 and q'_0 belonging to \mathcal{Q}_0 ; ii) $x_{q_0}(t)$ for $q_0 \in \mathcal{Q}_0$ are derivable at t with the same derivative, i.e., $\frac{dx_{q_0}(t)}{dt} = \frac{dx_{q'_0}(t)}{dt}$ for every q_0 and q'_0 belonging to \mathcal{Q}_0 .

In the latter case there are two cases:

$$\begin{cases} x_{q_0} > x^* \rightarrow \sum_{q_0 \in \mathcal{Q}_0} \mu_{q_0}(t) = \hat{\mu} \\ x_{q_0} < x^* \rightarrow \sum_{q_0 \in \mathcal{Q}_0} \mu_{q_0}(t) = 0 \end{cases}$$

This implies that

$$\sum_{q_0 \in \mathcal{Q}_0} \frac{dx_{q_0}}{dt} = \sum_{q_0 \in \mathcal{Q}_0} \left[\frac{M_q}{r_q} w_q(t) + \lambda_q \right] - \sum_{q_0 \in \mathcal{Q}_0} \mu_{q_0}(t)$$

from which, exploiting the fact that all queues $q_0 \in \mathcal{Q}_o$ must have the same derivative, after some calculations we obtain:

$$\begin{cases} x_{q_0} > x^* \rightarrow \frac{dx_{q_0}}{dt} < \frac{M_q \delta}{r_q} - \frac{1}{|\mathcal{Q}_0|} (-\hat{\mu} + \sum_{\mathcal{Q}_0} \mu_q^*) \\ x_{q_0} < x^* \rightarrow \frac{dx_{q_0}}{dt} > -\frac{1}{|\mathcal{Q}_0|} \mu_q^* - \frac{M_q \delta}{r_q} \end{cases}$$

In both the above case $\frac{d \max_q (x_q(t) - x^*)^2}{dt}$ for small values of δ , satisfies to the following inequality:

$$\begin{aligned} \frac{d \max_q (x_q(t) - x^*)^2}{dt} &= 2 \frac{dx_{q_0}(t)}{dt} (x_{q_0}(t) - x^*) < \\ &< -\frac{1}{|\mathcal{Q}_0|} \mu_{\min} (x_q(t) - x^*) \end{aligned}$$

Now consider source q :

$$\begin{aligned} \frac{d(w_q(t) - w^*)^2}{dt} &= 2 \frac{dw_q(t)}{dt} (w_q(t) - w^*) = \\ &- \frac{2w^* f(x^*) (w_q(t) - w^*)^2}{r_q} - \\ &- \frac{f'(x^*) (w^*)^2 (w_q(t) - w^*) (x_q(t) - x^*)}{r_q} \end{aligned}$$

being $f'(x^*) = \frac{df(x)}{dx} \Big|_{x=x^*}$. Combining previous results, we obtain:

$$\begin{aligned} \frac{d\mathcal{L}(X(t), W(t))}{dt} &< -\frac{1}{2|\mathcal{Q}_0|} \mu_{\min} (x_{q_0}(t) - x^*) - \\ &2\beta w^* f(x^*) \sum_q \left[\frac{M_q (w_q(t) - w^*)^2}{r_q} \right] - \\ &\beta f'(x^*) (w^*)^2 \sum_q \left[\frac{M_q (w_q(t) - w^*) (x_q(t) - x^*)}{r_q} \right] \end{aligned}$$

Note that the first two terms are negative, while the latter has an indefinite sign. However the whole is negative if:

$$\delta f'(x_q^*) (w_q^*)^2 \sum_q \frac{M_q}{r_q} < \frac{1}{2|\mathcal{Q}_0|} \frac{\mu_{\min}}{\beta}$$

Now we consider the case $|\mathcal{Q}_0| = Q$ i.e., all the queues at time t are of the same length $x_q(t) = x(t)$. Repeating similar consideration as before we obtain:

$$\frac{dx_{q_0}}{dt} = \frac{1}{Q} \sum_{q=1}^Q \frac{M_q}{r_q} (w_q(t) - w^*)$$

From which it follows that:

$$\begin{aligned} \frac{d\mathcal{L}(X(t), W(t))}{dt} &= \frac{2}{Q} \sum_q \left[\frac{M_q}{r_q} (w_q(t) - w^*) \right] (x(t) - x^*) - \\ &2\beta w^* f(x^*) \sum_q \left[\frac{M_q (w_q(t) - w^*)^2}{r_q} \right] - \\ &\beta f'(x^*) (w^*)^2 \left[\sum_q \frac{M_q}{r_q} (w_q(t) - w^*) \right] (x(t) - x^*) \end{aligned}$$

can be made always negative choosing

$$\beta = \frac{2}{Q f'(x^*) (w^*)^2}$$

C Stability of the equilibrium point: wireless station

Denoting with $\Delta w_q(t) = w_q(t) - w_q^*$ and $\Delta x_q(t) = x_q(t) - x_q^*$, the linearized dynamical system of equations comprises $2Q$ equations. The first set of Q equations are obtained linearizing the equations describing the TCP window dynamic, we obtain:

$$\frac{d\Delta w_q(t)}{dt} = -\frac{2w_q^* f(x_q^*) \Delta w_q}{2r_q} - \frac{(w_q^*)^2 f'(x_q^*) \Delta x_q}{2r_q}$$

The other Q equations are obtained linearizing the queue dynamics equations:

$$\begin{aligned} \frac{d\Delta x_q(t)}{dt} &= \frac{M_q (1 - f_d(x_q^*)) \Delta w_q}{2r_q} - \\ &\frac{(M_q w_q^* + \lambda_q) f_d'(x_q^*) \Delta x_q}{2r_q} - \\ &- \frac{1}{2\|X^*\|_2} \Delta x_q + \frac{1}{2\|X^*\|_2} \sum_{q' \neq q} \Delta x_{q'} \end{aligned}$$

being $f'(x_q^*) = \frac{df(x)}{dx} \Big|_{x=x_q^*}$ and $f_d'(x_q^*) = \frac{df_d(x)}{dx} \Big|_{x=x_q^*}$.

The asymptotic stability of the equilibrium point $(0, 0)$ in the above linearized system of equations can be studied using standard techniques. The first step consists in rewriting the above system of equations in its standard form:

$$\frac{dY(t)}{dt} = AY(t)$$

being $Y(t)$ the vector state

$$(\Delta w_1(t), \Delta x_1(t), \Delta w_2(t), \Delta x_2(t) \dots \Delta w_Q(t), \Delta x_Q(t))$$

A sufficient and necessary condition for $(0, 0)$ to be asymptotically stable is that matrix A is stable, i.e., all eigenvalues of A have real part strictly negative.

We have verified the stability of matrix A computing the characteristic polynomial of A and applying the standard Routh-Hurwitz criterion (we do not report the details for brevity).