Photon Netw Commun (2009) 17:157–169
DOI 10.1007/s11107-008-0150-1

Scheduling in overlaid star all-photonic networks with large propagation delays

N. Saberi · M. J. Coates

Received: 29 May 2008 / Accepted: 29 July 2008 / Published online: 22 August 2008

Abstract This paper describes a framework for fixed-length frame scheduling in all-photonic networks with large propagation delays. We introduce the Fair Matching Algorithm a novel scheduling approach that results in weighted max-min fair allocation of extra slots, achieves zero rejection for admissible demands, and minimizes the maximum percentage rejection of any connection. We also propose the Minimum Rejection Algorithm, which minimizes total rejection but treats non-critical connections in a fair manner. Finally, we introduce a feedback control system based on Smith's principle that reduces the effect of prediction errors and increases the speed of the response to the sudden changes in traffic arrival rates. Simulations performed using OPNET Modeler explore the performance of the scheduling and control algorithms we propose.

Keywords All-photonic networks · Scheduling · Max-min fairness · Star topology

1 Introduction

In modern high speed networks, electronic switches and the associated opto-electronic conversion limit the optical capacity to a few gigahertz, so the insertion of all-photonic switches in the network cores is attractive. The primary disadvantage is that all-photonic switches are currently incapable of performing queuing, so packet transmissions must be carefully controlled. Burst switching and just-in-time reservation approaches, and routing and wavelength assignment techniques address this challenge in general mesh topologies [23,34]. Using a simpler architecture such as an (overlaid) star topology reduces the complexity of the control plane.

In this article, we consider the Agile All-Photonic Network (AAPN), which is an overlaid star topology [4,16]. This architecture (see Fig. 1) consists of edge nodes equipped with buffers and optical electronic convertors and fast, reconfigurable and buffer-less photonic core crossbar switches which connect the edge nodes. The star topology facilitates global network synchronization [12], enabling the adoption of optical time-division multiplexing (OTDM) approaches such as wavelength-specific scheduling of time-slots. To avoid collision a source edge-node must be aware of when it has ownership of a given time-slot and is allowed to transmit to a specific destination edge node.

In this study, we assume that the traffic has been divided among the stars using some form of load-balancing, for example one of the techniques outlined in [35]. Therefore, the core switches act independently and the control problem is reduced to the task of scheduling one switch configuration to achieve a good match with the traffic arrival pattern at the edge nodes.

Bandwidth allocation in networks with substantial signaling delay is normally based on the prediction of traffic arrival rates. In wide-area networks, it is much more efficient to schedule blocks of slots (frames)[1] than single slots [13]. In frame-based scheduling algorithms, the edge nodes report their predicted bandwidth requirements for each

N. Saberi (✉)
Division of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA
e-mail: nahid@deas.harvard.edu

M. J. Coates
Department of Electrical and Computer Engineering, McGill University, Montreal, QC, Canada
e-mail: mark.coates@mcgill.ca

[1] In this article, the term "frame" refers to a set of time-slots containing multiple packets (for example, slots of 10 μs duration, which can hold up to 100 packets of 1000 bits on average on a 10 Gbps optical channel).

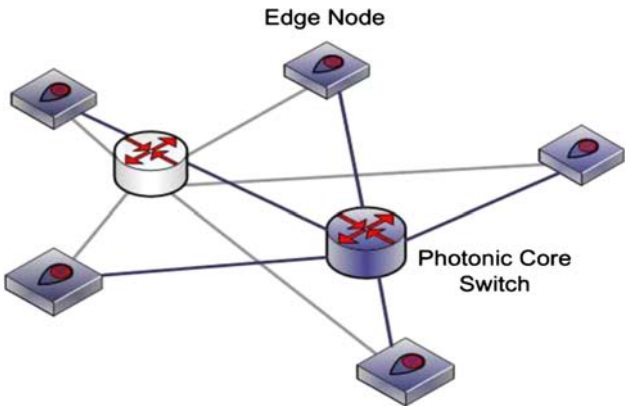

Fig. 1 Architecture of the agile all-photonic network described in [4,16]. Edge nodes perform electronic-to-optical conversion and transmit scheduling requests to the core photonic node(s). Each buffer-less photonic core switch provides connectivity between any pair of edge switches establishing a star topology. The overlay of several stars provides resilience to link or core failures

frame duration to the central scheduler. Many techniques can be adopted for performing this prediction, ranging from a naive predictor (the prediction is equal to the current traffic arrival rate) to more elaborate techniques based on sophisticated traffic models [31]. Since traffic prediction is beyond the scope of this paper, we simply consider a naive predictor.

Contribution: We study the problem of fixed-length frame scheduling in an overlaid star topology all-photonic network. The minimization of rejection is the priority. Fairness in the max-min sense is also a desirable criterion and plays an important role in achieving minimum average end-to-end delay. We therefore propose the Fair Matching Algorithm (FMA), an algorithm based on the weighted max-min fairness criterion. This algorithm provides zero rejection in the case of admissible traffic and a fair allocation of extra bandwidth for the under-loaded links in the network. We show that FMA minimizes the maximum percentage rejection experienced by any connection. Subsequently, we propose the Minimum Rejection Algorithm (MRA), which minimizes total rejection but treats non-critical connections in a fair manner. This algorithm has much lower average time complexity compared to the straightforward approach of solving a max-flow problem.

Finally, we introduce a closed-loop control architecture designed to interact with our proposed open-loop scheduling mechanisms. We employ Smith's principle to design a linear feedback controller that compensates for the sources of error (prediction, rounding, and rejection), resulting in a stable and fair system. The feedback control system we propose allocates spare capacity in a fair manner and responds to traffic variations faster than the open-loop scheduling algorithm alone. This controller acts as an illustration of a general framework for combining a closed-loop controller with a centralized scheduler.

Related work: Scheduling in the AAPN is similar to scheduling of an input queued switch (see [1,18,19]), with the difference that there is a large propagation delay between the input buffers at the edge nodes and the switch (photonic core). This leads to superior performance for frame-based scheduling algorithms. Of the frame-based algorithms proposed for star topologies in optical and satellite networks, the majority have focused on *variable-length* frames [9–11,22,33]. Using fixed-length frames reduces computational complexity, and simplifies control and signaling, particularly slot synchronization and bandwidth request management. The authors of [3,5,14,25] have considered the problem of scheduling a frame of fixed length for star-coupled networks with tunable transmitters/receivers, but do not address the allocation of unused time-slots or rejection of inadmissible demand. We note that the general principles employed in our algorithms, water-filling and max-flow formulations, have been used in various scheduling contexts, e.g., [3,14], but never for scheduling fixed-length frames in all-photonic networks (where wavelength-tunability of transmitters/receivers is not a primary consideration).

The most closely related work is that of Peng et al. [21], who also address scheduling in a star-topology agile all-photonic network. Their procedure focuses on determining a service matrix that is similar to the original demand matrix through a process of iterated projection. This procedure achieves clamping of the demand matrix, but the authors make no claim regarding what the projection procedure achieves in terms of network performance. In contrast, the algorithms we propose in Sect. 3, FMA and MRA, explicitly achieve fairness properties or minimize total rejection.

Feedback congestion control has been examined from a control theoretic perspective by many authors, with the primary focus being controlling the rates at which sources inject best-effort traffic into a network in order to reduce the congestion at bottleneck queues while maintaining high utilization. In the work most closely related to the controller design presented in this article, Mascolo combines classical control theory and Smith's principle to design a simple congestion control law that guarantees no packet loss and efficient use of bandwidth [15]. In related work, Bauer et al. propose a new class of time-variant Smith predictors using time-variant network delay models [2]. Although the theoretical techniques we adopt in our design are similar to those used by Mascolo, the problem we address differs significantly. We assume that we have no control over arrival rates; instead we can adjust, through scheduling, the resources allocated in the network. This results in an inverted form of the standard congestion control problem: switch resources are controlled rather than source rates.

Finally, we should note that the research reported in this article is a compilation and extension of the material presented in the conference papers [28–30].

Structure of the paper: Section 2 provides a statement of the scheduling problem that we address. Section 3 details our proposed frame-based scheduling algorithms, FMA and MRA. Section 4 illustrates how the frame-based scheduling algorithms act as feed-forward control systems. Section 5 describes the design of a modified Smith controller that interacts with the FMA scheduling algorithm to produce a stable resource allocation mechanism for AAPNs. Section 6 describes the simulation experiments we have executed to assess performance. Finally, Section 7 summarizes the proposed algorithms and results. Proofs appear in the appendices.

2 Problem statement

We investigate the scheduling problem in the AAPN, an overlaid star topology network which connects a large number of edge nodes using optical fibers and photonic switches. We assume that load is divided between the stars and we are concerned with scheduling each star independently. Every star is composed of N edge nodes connected through an all-photonic switch. We also assume that there are W available wavelengths and every edge node is able to transmit/receive on every wavelength simultaneously. To isolate the wavelengths, we assume that the load is distributed among the wavelengths as well, and our task in this article is to schedule one single wavelength on a single star network.

During each frame, every edge node records the number of packets that arrived for each destination node and reports these values to the scheduler. For simplicity, we assume that the scheduler is located at the core photonic switch. If the maximum one-way signaling delay is T frames, then the scheduler must design a schedule T frames into the future. It uses all information at its disposal to predict the demand for each source–destination pair.

Suppose that D_{ij} is the predicted number of slots needed for transmission from source node i to destination node j. We consider a frame of length L time-slots. Our aim is to devise a schedule S such that the element S_{jk} identifies the source node allocated to the k-th time-slot associated with destination j in the frame.

The rejection for any individual connection (i, j) is denoted by:

$$REJ_{ij} = \max\left(0, D_{ij} - \sum_{k=1}^{L} \mathbb{I}[S_{jk} = i]\right) \tag{1}$$

where $\mathbb{I}$ is the indicator function. The total number of rejections is defined as: $TREJ(S, D, L) = \sum_i \sum_j REJ_{ij}$. We identify two scheduling problems for frames of fixed length L with demand matrix D. The first strives to minimize total rejection; the second strives to minimize worst-case percentage rejection. Suppose that S_1^* and S_2^* are the schedules obtained from solving the first and second problems, respectively. Therefore, we have:

MINREJ(D,L): $S_1^* = \arg\min_S TREJ(S, D, L)$.
PERMIN(D,L): $S_2^* = \arg\min_S \max_{(i,j)} REJ_{ij}/D_{ij}$.

2.1 Terminology and definitions

We now define some terminology that will be used throughout the article and recall some definitions. We denote the line sum of line ℓ of the demand matrix D by LS_ℓ. Note that line ℓ consists of a set of source–destination demands which correspond to the connections passing through link ℓ of the network. Each of these connections belongs to two lines, a row and a column. The i-th row represents a link from source i to the optical switch at the core, and the j-th column represents the link from the core to destination node j. The *row-sum*, $r_i(D) = \sum_{j=1}^{N} D_{ij}$, is the total demand at source i, and the *column-sum*, $c_j(D) = \sum_{i=1}^{N} D_{ij}$, is the total demand for destination j.

Definition 1 *Admissibility.* A demand matrix D is *admissible* if

$$\max\left\{\max_i\{r_i(D)\}, \max_j\{c_j(D)\}\right\} \le L.$$

For an inadmissible demand matrix, we denote the set of overflowing rows of the demand matrix (rows with $r_i(D) > L$) as O_r, and the set of overflowing columns ($c_j(D) > L$) as O_c. The set of overflowing lines, $O_\ell = \{\ell : LS_\ell > L\}$ is the union of O_r and O_c. We define a *critical connection*, or critical demand element, as any demand entry D_{hp} such that $h \in O_r$ and $p \in O_c$. The remaining entries constitute *non-critical* connections/demands.

Definition 2 *Feasibility.* Consider an arbitrary network as a set of links $\mathcal{L}$ where each link $\ell \in \mathcal{L}$ has a capacity $C_\ell > 0$. Let $\{1, \ldots, \zeta\}$ be the set of network connections, and H_ℓ the set of all connections passing through link ℓ. Let D_u be the demand (request) of connection u and υ_u be its assigned rate. A rate allocation $\{\upsilon_1, \upsilon_2, \ldots, \upsilon_\zeta\}$ is *feasible* if for every link $\ell \in \mathcal{L}$ we have $\sum_{u \in H_\ell} \upsilon_u \le C_\ell$.

Definition 3 *Weighted max-min fairness.* Let $\omega_u(\upsilon_u)$ be an increasing function representing the weights assigned to connection u at rate υ_u. A feasible allocation $\{\upsilon_1, \upsilon_2, \ldots, \upsilon_\zeta\}$ is *weighted max-min fair* if for each connection u any increase in υ_u would cause a decrease in transmission rate of connection z satisfying $\omega_z(\upsilon_z) \le \omega_u(\upsilon_u)$. The special case of max-min fairness is obtained by $\omega_u(\upsilon_u) = \upsilon_u$.

3 AAPN scheduling algorithms

This section introduces two scheduling algorithms. The FMA, addresses the *PERMIN* problem. FMA achieves weighted max-min fairness in sharing the bandwidth between the communicating source–destination pairs. For inadmissible traffic, FMA minimizes the maximum percentage rejection experienced by any demand. The second algorithm, the MRA efficiently solves *MINREJ(D,L)*.

3.1 Fair Matching Algorithm (FMA)

Fair Matching Algorithm is a combination of a clamping procedure and the *EXACT* algorithm. The *EXACT* algorithm, presented in [7,33], was designed for a variable-length frame and it achieves the minimum number of slots for this case. It is an iterative procedure that repeatedly performs maximum cardinality bipartite matching (MCBM) to obtain the schedule. When applied to the problem of scheduling a fixed-length frame with an admissible demand matrix, the *EXACT* algorithm generates a schedule S that has length less than L, and therefore zero rejection. If the demand matrix is inadmissible, or the demand is lower than the capacity of a frame, then it is desirable to modify the demand matrix to control the way in which rejection occurs or free slots are assigned.

Clamping modifies the demand matrix to ensure that all of the frame resources are assigned properly. If the demand matrix is admissible, FMA performs water-filling, incrementally assigning additional demands to all elements until all of the links reach capacity (their line-sums are equal to L). This algorithm can be implemented by processing one line at a time. We first choose the most constrained line (the line that would reach its capacity first under the water-filling procedure) and increase its demand to capacity. Then we choose the next most constrained line and increase its demand to capacity. We repeat until all lines have reached capacity. FMA assigns extra capacity *in proportion to the original demand*.

A similar procedure can be used for the case of an inadmissible demand matrix (containing one or more overloaded lines). In this case, FMA identifies the most overloaded line and reduces the demands on that line such that they sum to capacity (L). Demand reduction is proportional to the original demand, i.e. each adjusted demand experiences the same *percentage reduction*. When there are both overloaded and under-utilized lines, the overloaded lines are adjusted first.

Here we describe how FMA treats demands belonging to the adjustable lines in the set $U_\ell = \{\ell : \ LS_\ell(0) \neq L\}$, where $LS_\ell(0)$ is the line sum of line ℓ at the beginning of calculations. We define $\mathcal{A}_D \subseteq U_\ell$ as the set of unmodified lines and $\mathcal{B}_D \subseteq U_\ell$ as the set of modified lines. Initially $\mathcal{A}_D$ contains all lines in U_ℓ and $\mathcal{B}_D$ is empty. Similarly, we define a_ℓ as the set of unmodified demands in line ℓ and b_ℓ as the set of modified demands. Initially, a_ℓ contains all the demands and b_ℓ is empty. In each iteration we adjust the unmodified demands in line ℓ as follows:

$$D'_{ij} = D_{ij} \times \frac{L - S_{b_\ell}}{S_{a_\ell}} \quad \forall\,(i,j) \in a_\ell, \tag{2}$$

where $S_{a_\ell} \triangleq \sum_{(i,j)\in a_\ell} D_{ij}$ and $S_{b_\ell} \triangleq \sum_{(i,j)\in b_\ell} D'_{ij}$. We always have $S_{a_\ell} + S_{b_\ell} = LS_\ell$. Note that when demand D_{ij} belongs to an overloaded line, $\frac{L-S_{b_\ell}}{S_{a_\ell}} < 1$, and when D_{ij} belongs to an under utilized line $\frac{L-S_{b_\ell}}{S_{a_\ell}} > 1$. Define for each of line in $\mathcal{A}_D$ the value $G_\ell \triangleq \frac{L-LS_\ell}{S_{a_\ell}}$.

Algorithm 1 FMA

Set $D' = D$.
while $\mathcal{A}_D \neq \mathcal{O}$ **do**
  Identify the line $\ell^* = \arg\min_{\ell\in\mathcal{A}_D} G_\ell$.
  Apply (2) to line ℓ^*.
  Transfer ℓ^* from $\mathcal{A}_D$ to $\mathcal{B}_D$.
  Update a_ℓ and b_ℓ for all lines $\ell \in \mathcal{A}_D$.
  Re-evaluate LS_ℓ for all lines in $\mathcal{A}_D$.
  Transfer lines γ with $LS_\gamma = L$ from $\mathcal{A}_D$ to $\mathcal{B}_D$.
end while
Apply *EXACT* to $\lfloor D' \rfloor$ to generate S.

The following theorem states that prior to rounding, FMA achieves weighted max-min fair allocation of capacity (weighted relative to the original demand). See Appendix A for the proof.

Theorem 1 *FMA generates an adjusted demand matrix D' with weighted max-min fair allocation, where the weight is $\omega(D'_{ij}) = \frac{D'_{ij}}{D_{ij}}$.*

If the demand matrix contains zero entries, then an algorithm that adjusts requests multiplicatively (such as FMA) cannot always generate full utilization; there can be *natural blocking* because there is no demand. After all of the demands are adjusted FMA uses *EXACT* to allocate the time-slots and generate the schedule. We now present some properties of FMA and the demand matrix $D' = \{D'_{ij}\}$ obtained by FMA prior to rounding.

Property 1: FMA guarantees full allocation of all links provided D contains no zero elements.

Property 2: If there is no natural blocking the maximum total throughput of the network is obtained: $\sum_i \sum_j D'_{ij} = NL$.

Property 3: The while-loop in FMA has $O(N^2)$ computational complexity in terms of the number of edge nodes ($2N$ iterations with a minimization over N elements in each iteration). The best current implementation of the *EXACT* algorithm has complexity $O(N^{\frac{5}{2}})$, and hence this is also the complexity of FMA.

Define the *percentage rejection* as $1 - \frac{D'_{ij}}{D_{ij}}$ for the lines which were initially overloaded. Consider the set of demands that experience the highest percentage rejection (i.e., the demands on the most overloaded line). Since the weight ω is a monotonically increasing function of allocated rate D'_{ij}, weighted max-min fairness implies that it is impossible to increase the rate allocated to these demands (or decrease the maximum percentage rejection) without violating feasibility. Decreasing the rejection of any of those demands requires increasing the rejection of another demand on the same line, and hence the maximum percentage rejection increases. We thus have the following corollary of Theorem 1:

Corollary 1 *Subject to the capacity constraints, FMA generates a schedule that minimizes the maximum percentage rejection experienced by the connections.*

$$\max_{ij}\left\{\frac{D_{ij}-D'_{ij}}{D_{ij}}\right\}_{FMA} = \min_{CL}\left\{\max_{ij}\left\{\frac{D_{ij}-D'_{ij}}{D_{ij}}\right\}_{CL}\right\}, \tag{3}$$

where CL is any clamping algorithm that clamps the overloaded lines down to L.

3.2 Minimum Rejection Algorithm (MRA)

We are now in a position to define an algorithm that (i) minimizes overall rejection, and (ii) subsequently, fairly allocates any necessary residual rejection or free slots.

We commence by considering a decomposition of the demand matrix, $D = D' + R$. Here D' is the pruned demand matrix with line sums not exceeding the schedule length L and R shows the resulting rejections of every demand after pruning. We define the sets $\mathcal{B} \triangleq \{(h, p) : h \in O_r \text{ or } p \in O_c\}$, and $\mathcal{C} \triangleq \{(h, p) : h \in O_r \text{ and } p \in O_c\}$, where O_r and O_c are the set of overflowing input and output links of the optical network, respectively. The minimization of total rejection can be formulated as the following max-flow problem:

$$\begin{aligned}
&\text{Maximize} \sum_{(h,p)} D'_{hp} \text{ subject to}\\
&\quad 0 \le D'_{hp} \le D_{hp} && \forall\,(h,p)\in\mathcal{B},\\
&\quad r_h(D') \le L, \quad c_p(D') \le L && \forall (h,p).
\end{aligned}$$

Ford and Fulkerson presented a solution to max-flow problems of this kind in 1954 [8]. Note that the max-flow solution is in general not unique. The fastest maximum flow algorithms to date are preflow-push algorithms, which work in a more localized manner than the Ford–Fulkerson method [6]. In the straightforward formulation of the max-flow problem above, there are $2N$ active nodes in the corresponding $s \to t$ network (see [26] for details), so the complexity of the preflow-push algorithm for finding a max-flow solution is $O(N^3)$ [6].

We now outline a procedure for solving the minimum rejection problem that can result in significant computational savings. We commence by defining a related but simpler max-flow linear programming problem, *MAXREJFLOW(D,L)*:

$$\begin{aligned}
&\text{Maximize} \sum_{(h,p)\in\mathcal{C}} Y_{hp} && \text{subject to}\\
&\quad Y_{hp} = 0 && \text{if } (h,p)\notin\mathcal{C}, && (4)\\
&\quad Y_{hp} \le D_{hp} && \forall (h,p), && (5)\\
&\sum_{p\in O_c} Y_{hp} \le r_h(D) - L && \forall\, h \in O_r, && (6)\\
&\sum_{h\in O_r} Y_{hp} \le c_p(D) - L && \forall\, p \in O_c. && (7)
\end{aligned}$$

In order to find an efficient approach for solving *MINREJ(D,L)*, we identify a relationship to a solution of *MAXREJFLOW(D,L)* with the following theorem. The proof is in Appendix B.

Theorem 2 Set $A = MAXREJFLOW(D,L)$. *Construct a rejection matrix* $R = A + Q$, *where* Q *is a non-negative matrix such that* $Q_{hp} = 0 \;\; \forall (h,p) \notin \mathcal{B}$, $Q_{hp} \le D_{hp} - A_{hp} \;\; \forall (h,p)$, $r_h(Q) = r_h(D) - L - r_h(A) \;\; \forall h \in O_r$, *and* $c_p(Q) = c_p(D) - L - c_p(A) \;\; \forall p \in O_c$. *Then if S is a schedule that generates the decomposition* $D = D' + R$, *it is a solution to the problem MINREJ(D,L).*

The identification of a solution to *MINREJ(D,L)* thus requires us to (i) find a solution A to *MAXREJFLOW(D,L)*; and (ii) determine a suitable Q. The *MAXREJFLOW* problem is a max-flow problem, and a solution can also be determined using the Ford–Fulkerson algorithm or one of the preflow-push algorithms. The FMA algorithm can be used to determine a suitable Q. Note that A only has non-zero entries on the critical connections. By using FMA to determine the remaining rejection, we are introducing weighted max-min fairness in rejection allocated to the non-critical connections. The combined MRA is specified in Algorithm 2.

Algorithm 2 Minimum Rejection Algorithm

1: Apply the Ford–Fulkerson algorithm (or an alternative preflow-push algorithm) to solve $A = MAXREJFLOW(D,L)$.
2: Generate the modified demand matrix $D' = FMA(D - A, L)$.
3: Apply *EXACT* to $\lfloor D' \rfloor$ to generate S.

The complexity of the MAXREJFLOW problem is $O(|O_\ell|^3)$. In the worst case all $2N$ lines are overflowing, and the complexity is $O(N^3)$. In general, only a fraction of the lines are overflowing, and $|O_\ell| \ll N$, so there is a substantial reduction in computational complexity. In the MRA

algorithm, this reduction is offset, however, by the incorporation of the FMA algorithm, which has complexity $O(N^{5/2})$. The primary advantage of the MRA algorithm is the introduction of weighted max-min fairness in rejection and residual slot allocation for the non-critical connections.

4 Queue control and stability

The scheduling techniques outlined in the previous sections can be interpreted as open-loop control algorithms. If the system relies on only open-loop, feed-forward control then the effect of errors is ignored, leading to instability and unfairness. These errors arise primarily from mistakes in the traffic prediction and the fact that the scheduling algorithms involve rounding and do not remember past rejection. A closed-loop control system is needed to achieve stability (i.e., bounded steady state queue size variation), fairness, and faster response to traffic variations.

We now develop a control system model for resource allocation in an AAPN. Initially, we adopt a continuous-time model, but since scheduling is performed once per frame, we later sample the data with period T_s (the frame duration) to obtain a discrete-time system. Figure 2 shows a feedback control model for an agile all-photonic network with a central controller. Note that this figure depicts the control loop for one source–destination pair, or virtual output queue (VOQ), (i, j). There is a similar control loop for every source–destination pair, and all of these loops are coupled through the FMA scheduler.

We consider a simple integrator as the dynamic model for a VOQ. Let $q_{ij}(t)$ be the length of the virtual queue of packets at edge node i destined to edge node j. Let a_{ij} be the input rate to VOQ_{ij}, and dep_{ij} the depletion rate of this queue. In the control model the length of each VOQ is compared with a reference signal, $r_{ij}(t)$, and the difference is the input to the controller. The controller then calculates how to adjust the predicted traffic arrival rate $\hat{a}_{ij}(t)$ to account for past prediction errors, rejections, and rounding errors (this adjustment rate is $ac_{ij}(t)$).

We model the depletion rate dep_{ij} as constant throughout a frame period:

$$dep_{ij}(t+T) = \frac{D'_{ij}(k)C}{L} \quad kT_s \le t \le (k+1)T_s.$$

The predicted arrival rate is used as the demand signal $d_{ij}(t)$. Therefore, we have:

$$\hat{a}_{ij}(t) = d_{ij}(t) = \frac{D_{ij}(k)C}{L} \quad kT_s \le t \le (k+1)T_s.$$

Here $D_{ij}(k)$ is the predicted number of time slots demanded for a source–destination pair (i, j) during frame k, $D'_{ij}(k)$ is the adjusted number of allocations based on the FMA algorithm, C is the line rate in bits-per-second, L is the frame-length in slots, and T is the propagation (signaling) delay.

Provided that the queue does not empty ($q_{ij} > 0$), the depletion rate is the sum of the predicted arrival rate $\hat{a}_{ij}$ and the feedback adjustment ac_{ij}, suitably delayed in time, i.e., $dep_{ij}(t) = \hat{a}_{ij}(t-T) - ac_{ij}(t-T)$. Based on the flow conservation equation [15] the queue length, with initial condition $q_{ij}(0) = 0$, is $q_{ij}(t) = \int_0^t [a_{ij}(\tau) - dep_{ij}(\tau)]d\tau$. We model the queues as always-occupied to avoid the need for non-linear components.

Demand matrix adjustment is performed by a clamping algorithm (e.g., FMA) which clamps the line sums of the demand matrix up or down to L. FMA multiplies the predicted arrival rate $\hat{a}_{ij}$ by a factor, $\mathbf{x}_{ij} = \frac{D'_{ij}}{D_{ij}}$. Since this factor changes with the overall arrival rates the gain of the controller is tuned each frame.

For this control system, we aim to minimize the error between the queue length and a *desired queue length* shown by the reference signal, which may be calculated based on

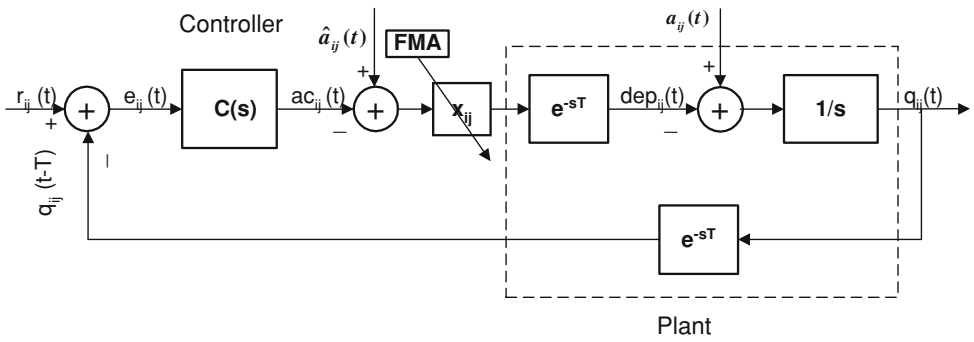

Fig. 2 The provision of a feedback signal results in bandwidth allocation in an AAPN becoming a simple closed-loop control system. Inputs to the system are a reference signal r_{ij}, the estimated arrival rate $\hat{a}_{ij}$ and the true arrival rate a_{ij}, and the feedback is the information from the VOQ, indicated by q_{ij}. The propagation delay from the controller to the plant is T

the state of the network. For example, if the desired state is equal queue lengths for all of the VOQs, then the reference signal should be the average of the VOQ lengths. This is also the effect of FMA [28], so this choice of reference signal aligns the feedback controller with the feed-forward controller. Note that because FMA is a clamping algorithm, the combination of the controller and FMA never acts to artificially increase queue lengths. FMA always allocates the full capacity of the switch (provided there is non-zero demand).

5 AAPN controller design based on the Smith predictor

Instability is a common problem in delayed systems, since the addition of delays introduces extra phase lag, resulting in a less stable system. If the controller is not properly tuned to consider this delay (deadtime), it can overcompensate substantially. The Smith predictor, introduced by Smith [32], makes the controller aware of the deadtime and adjusts its behavior based on prediction of the effect of controller on the output during this delay. Our controller design is an extension of the modified Smith predictor developed by Mataušek and Micić [17].

Figure 3 shows the modified version of this controller for the AAPN network. The inputs to the system are $r_{ij}(t)$, $a(t)$ and $\hat{a}(t)$, and the output is $q_{p_{ij}}(t) = q_{ij}(t - T)$. We consider the arrival rate $a(t)$ and its prediction $\hat{a}(t)$ as disturbances. The reference signal, r_{ij}, represents the desired VOQ length. The setpoint and disturbance responses of the system are:

$$H_r(s) = \frac{\mathbf{x}_{ij} K_r e^{-sT}}{s + \mathbf{x}_{ij} K_r}, \tag{8}$$

$$H_d(s) = \frac{e^{-sT}[s - \mathbf{x}_{ij} K_r (1 - e^{-2sT})]}{(s + \mathbf{x}_{ij} K_r)(s + K_0 \mathbf{x}_{ij} e^{-2sT})},$$

$$\hat{H}_d(s) = \frac{\mathbf{x}_{ij} e^{-2sT}[s - \mathbf{x}_{ij} K_r (1 - e^{-2sT})]}{(s + \mathbf{x}_{ij} K_r)(s + K_0 \mathbf{x}_{ij} e^{-2sT})}$$

$$= \mathbf{x}_{ij} e^{-sT} H_d(s). \tag{9}$$

We strive to eliminate the steady-state effect of variations in the traffic arrival rate on the VOQ lengths. This corresponds to eliminating the load disturbance steady-state response and requires that $\lim_{s\to 0} H_d(s) = 0$, which is possible if $K_0 \neq 0$. Based on the final value theorem:

$$\lim_{t\to\infty} q_{p_{ij}}(t) = \lim_{s\to 0} R_{ij}(s) H_r(s) = r_{ij}. \tag{10}$$

The stability of the system depends on the roots of the characteristic equation:

$$(s + \mathbf{x}_{ij} K_r)(s + K_0 \mathbf{x}_{ij} e^{-sT}) = 0. \tag{11}$$

The first term implies that $\mathbf{x}_{ij} K_r > 0$ must be satisfied. We can apply the same analysis as that employed in [17] to derive the range of values for K_0 for which the system is stable (the phase margin $\phi M > 0$). We require:

$$K_0 < \frac{1}{4\mathbf{x}_{ij} T}. \tag{12}$$

It is highly likely that there is additional error in the control system, because the data are subject to queueing delay which is not explicitly included in our control system model. The system estimates the dead-time as T, which corresponds only to the propagation (signaling) delay, and this can be a significant underestimate. We, therefore, must examine the robustness of the system to this type of error. This analysis, conducted in Appendix C, reveals that the proposed system is robust to such errors, even if they are as large as the anticipated dead-time itself.

Scheduling and signaling are only performed once per frame. In order to obtain the equivalent discrete-time system equations a simple approach is to design a digital control system using the *Delta transform*. Since the plant is continuous the input to the plant is then converted to continuous

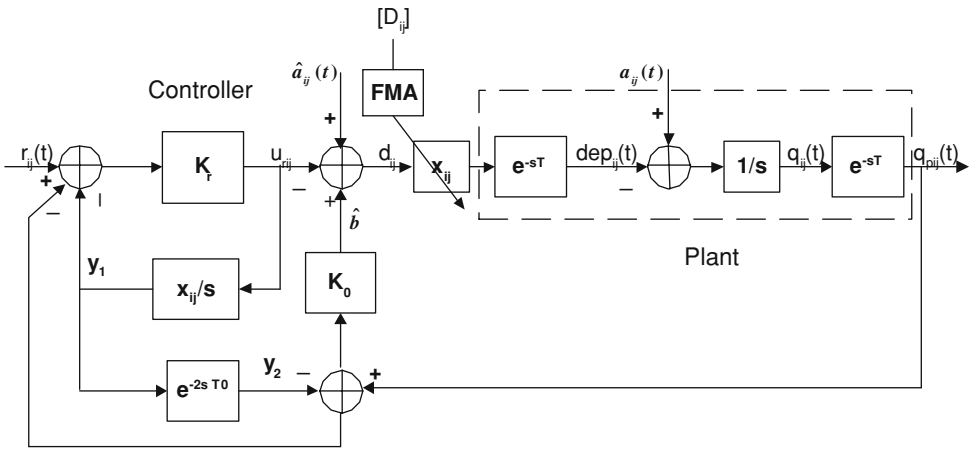

Fig. 3 Schematic of a modified Smith predictor for bandwidth allocation in a wide-area AAPN with large signaling delay. The terms K_r and K_0 represent gains (control parameters) and $T0$ is the estimated dead-time. In our analysis we assume $T0 = T$

form with zero-order-hold. The discrete time equations are approximated from the continuous form as:

$$\begin{aligned} d_{ij}(k) &= \hat{a}_{ij}(k) - u_{rij}(k) + K_0 q_{pij}(k) - K_0 y_2(k), \\ y_1(k) &= y_1(k-1) + \mathbf{x}_{ij}(k-1) u_{rij}(k-1) T_s, \\ y_2(k) &= y_1\left(k - \frac{2T}{T_s}\right), \\ u_{rij}(k) &= K_r(-y_1(k) + y_2(k) - q_{pij}(k) + r_{ij}(k)). \end{aligned} \tag{13}$$

Defining $\lambda \triangleq \frac{T}{T_s}$, we have:

$$u_{rij}(k) = K_r \left(-\sum_{p=1}^{\lambda} \mathbf{x}_{ij}(k-p) u_{rij}(k-p) T_s - q_{pij}(k) + r_{ij}(k) \right). \tag{14}$$

The rate adjustment thus depends, through the controller parameters K_0 and K_r, on the divergence of each queue length from the average queue length, r_{ij}, as well as the amount of the queue backlog $q_{pij}(k)$. The role of the Smith controller is to take into account the effect of rate adjustment on the queues during the λ previous frames for which there is no feedback available.

The gain of the controller K_r is designed based on the Nyquist–Shannon sampling theorem which states that the sampling period should be at most half the time constant of the continuous system $(1/\mathbf{x}_{ij} T_s)$. Using a fixed controller gain can result in undesirable behavior. A small gain does not provide sufficiently fast response to traffic changes, but a large gain results in overreaction to minor fluctuations. An adaptive gain can provide a good compromise. We design the controller such that the gain K_r adapts to the size of the queue variations:

$$K_r(k) = \min \left\{ A \exp(C \Delta q_p), \frac{1}{2\mathbf{x}_{ij} T_s} \right\}, \tag{15}$$

where $\Delta q_p = q_p(k) - q_p(k-1)$. The choice of the constants A and C determines how fast the system reacts to traffic changes and whether there are residual oscillations. To avoid overcompensation due to a large control gain we use a *fast-start slow-finish* procedure in which we reduce the gain of the controller by a factor of 0.05 two frames after activation of the Smith controller.

6 Simulation performance

In this section, we report the results of simulations of the scheduling approaches performed using OPNET Modeler [20]. We performed simulations on a 16 edge-node star topology network. The links in the network have capacity 10 Gbps and the propagation delay between each edge node and the optical switch is 5 ms. A time-slot is of length 10 μs, and a frame has a fixed length of 1 ms (or 100 slots). Recall that each time-slot contains multiple packets (e.g., 100 IP packets on average) and a frame refers to a set of time-slots. Each experiment was run for a duration of 0.2 s (equal to 200 frame durations) and the results were averaged over five repetitions of the simulations. The VOQs in the simulations have fixed buffer size (90,000 packets). Whenever the buffer is full, arriving packets are dropped. A summary of the network parameters is presented in Table 1.

Table 1 Network parameters

Parameter	Value
Number of nodes	16, 32, or 64
Link capacity	10 Gbps
Propagation delay	5 ms
Time-slot duration	10 μs
Frame duration	1 ms
Frame length	100 time-slots
Average number of packets per slot	100
Simulation time	0.2–0.5 s
Pareto shape parameter (α)	1.9

Our simulations involve bursty traffic using on/off traffic sources. Every edge node is equipped with 6 on/off sources. The "on" and "off" periods have Pareto distributions with a shape parameter $\alpha = 1.9$. The mean of the "off" periods is five times greater than the mean of the "on" periods. During "on" periods the sources generate packets with an average rate up to the full link capacity (10 Gbps). The rate distribution is exponential.

In the first experiment, we compare the performance of FMA with that of the algorithm proposed by Peng et al. [21]. We use a non-uniform traffic pattern; each destination receives on average the same amount of traffic, but each source sends five times as much traffic to one specific destination as compared to the others. As Fig. 4, top panel, shows the average rejection when FMA is used is less than that when the projection method is used. The advantage of using FMA is more apparent when we compare the maximum rejection percentages of the two algorithms in Fig. 4, middle panel. Figure 4, bottom panel, indicates that FMA achieves lower average queueing delay especially at higher loads. Note that propagation delay is not included in the figure.

Our second experiment compares FMA and MRA. Since these only differ when there are critical elements in the demand matrix, we investigate scenarios where critical demands are likely to exist. In order to do this, in each frame we choose one arbitrary source i and one arbitrary destination j. Each source generates z times as many packets for destination j compared to other destinations. Similarly, source

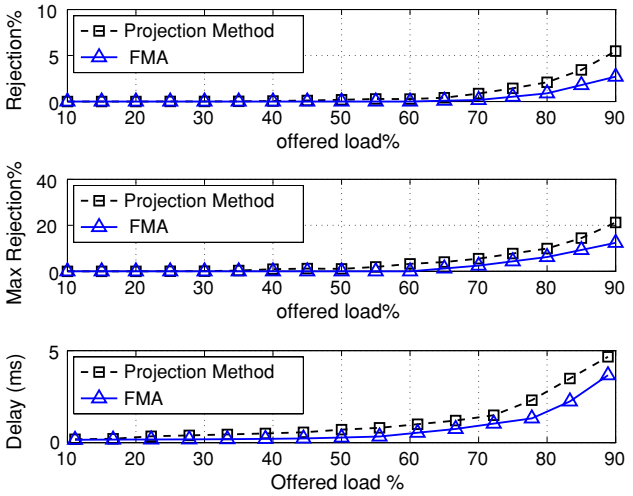

Fig. 4 Comparison between FMA and projection method under varying non-uniform traffic load in terms of rejection percentage and average queuing delay. Network has a propagation delay of 5 ms and 16 edge nodes. *Top panel*: Rejection percentage. *Middle panel*: Maximum of rejection percentage. *Bottom panel*: Queuing delay

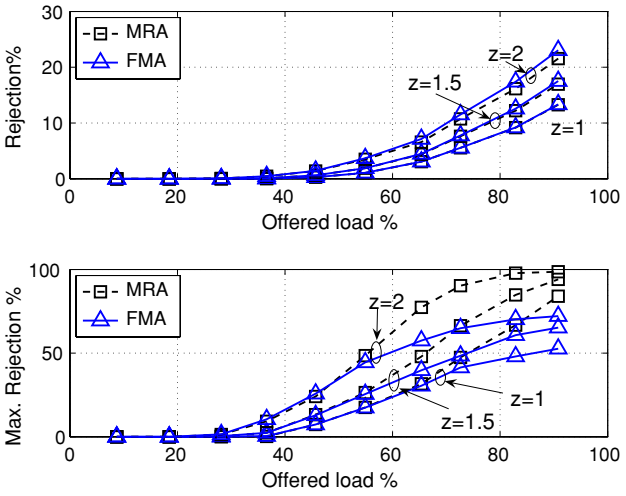

Fig. 5 Comparison between the rejection (*top panel*) and rejection percentage (*bottom panel*) obtained by FMA and MRA under varying offered load for different factors of imbalanced load (z). Traffic is bursty (generated by on–off sources) and has uniform distribution, aside from the impact of z

i generates z times as many packets (to all destinations) as any other source. As z increases, the elements of the demand matrix corresponding to these two edge nodes are more likely to be critical connections; the demand element D_{ij} has even higher likelihood of being critical.

Figure 5, top panel, compares the percentage of rejected demand achieved by FMA and MRA as the offered load changes for various values of z. At high load (>70%) with $z = 2$, there are numerous critical elements and MRA begins to achieve less rejection than FMA. The discrepancy is still only two percent at 90% load. Figure 5, bottom panel, compares the maximum percentage rejection experienced by any demand when scheduling is performed by FMA and MRA. As the offered load increases, MRA concentrates rejection on the critical elements; the maximum percentage rejection is thus much (up to 25%) higher than that achieved by FMA, which distributes rejection fairly amongst all competing connections. The average queuing delay experienced by packets when scheduling is performed using FMA and MRA are similar, and so not shown here.

Our third experiment explores how increasing the network size affects the performance of FMA. The simulation settings are the same as in the previous experiment (with $z = 1$). Figure 6, top panel, compares utilization for networks of 16, 32, and 64 edge nodes and uniform traffic. The utilization is not affected by network size. The bottom panel compares the average queuing delays. For lower offered loads the queuing delay multiplies by a factor close to 2 (and 4) for 32 (and 64) edge nodes. This is the expected scaling behavior, because the injected traffic is kept constant per node, so the total traffic doubles (and quadruples). For higher loads the queuing

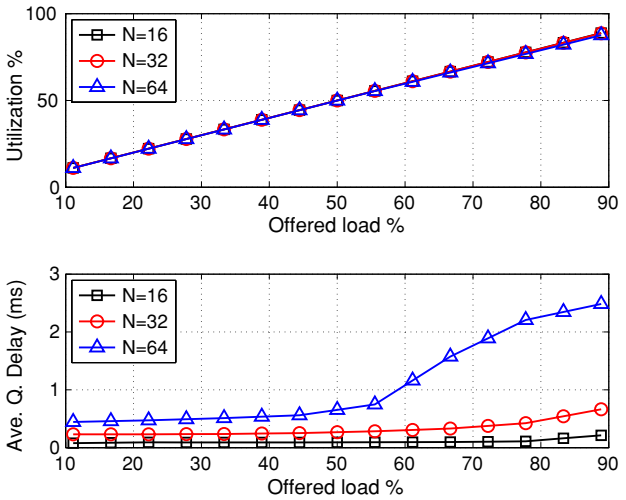

Fig. 6 Network performance (using FMA) with uniform traffic as a function of offered load with a propagation delay of 5 ms for different number of edge nodes. *Top panel*: Utilization. *Bottom panel*: Queuing delay

delay increases dramatically because the frame length is too small to support the increased number of nodes fairly. Sixty-four edge nodes with similar traffic arrivals cannot share 100 time-slots in a fair fashion.

Our fourth experiment investigates how the incorporation of the Smith controller impacts the response time of our system when there is a sudden change in traffic arrival rates. We are also interested in exploring the effect on the fairness in the system. We measure an average relative fairness factor (*divergence*), defined for source node j as:

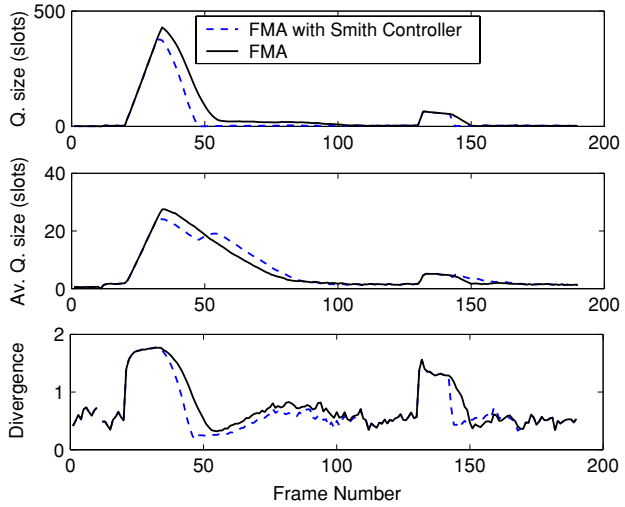

Fig. 7 The impact of the feedback controller with adaptive gain and fast-start slow-finish compensation for the simulation conditions of Scenario A. *Top panel*: Average queue length for VOQ experiencing the heavy load. *Middle panel*: Average queue lengths of all VOQs. *Bottom panel*: Relative fairness factor (divergence) as defined by (16)

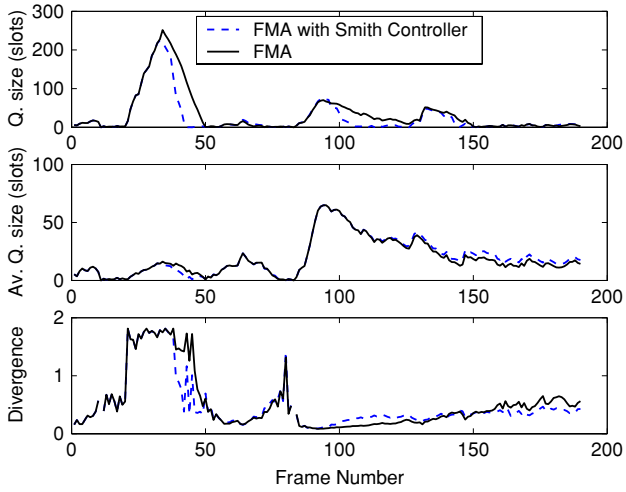

Fig. 8 The effect of the Smith controller under the bursty traffic conditions of Scenario B. *Top panel*: Average queue length for VOQ experiencing the heavy load. *Middle panel*: Average queue lengths of all VOQs. *Bottom panel*: Relative fairness factor (divergence) as defined by (17)

$$\delta_j = \frac{\sum_{i,i\neq j} |q_{p_{ji}} - \frac{\sum_{i,i\neq j} q_{p_{ji}}}{(n-1)}|}{\sum_{i,i\neq j} q_{p_{ji}}} \tag{16}$$

This factor measures the average divergence of the queue lengths of all VOQs at source node j from the overall average. It thus provides a good indication of the degree of equality of waiting times for packets in different queues (a value closer to zero indicates better fairness).

In this experiment, we employ two traffic scenarios. In both scenarios, the average arrival rates to the VOQs are equal except for two periods (frames 20–32 and frames 130–132) during which the arrival rate of traffic from one source to one destination increases by a factor of 10. The two traffic scenarios are:

Scenario A: The arrival distribution of the data packets is Poisson with average arrival rate of 9 Gbps during the baseline periods.

Scenario B: Six Pareto ($\alpha = 1.9$) on–off sources are connected to each edge node. The mean on-period is 0.33 ms and mean off-period is 1.6 ms. The average rates are 9 Gbps during the on-period.

The top panel of Fig. 7 compares the queue lengths of the VOQ carrying the heavy connection when using FMA with and without the Smith controller for the case of adaptive gains with $A = 63/\mathbf{x}_{ij}$ and $C = 0.08$ in (15). The Smith controller decreases the response time substantially, reducing the queue length of the heavy connection much faster than FMA alone. The middle panel shows that there is little impact on the other queues. The bottom panel compares average divergences. During the initial periods of heavy traffic, the fast draining of the long queue improves fairness.

Figure 8 examines the performance in response to bursty traffic as described in Scenario B, which is more unpredictable and thus poses a greater challenge for the Smith controller. The simulations indicate that the Smith controller still provides better drainage of the queues experiencing severe congestion. There is minimal negative effect on other queues or fairness. Similar results are observed for the case of several queues experiencing a sudden change.

7 Conclusion

We investigated bandwidth allocation and scheduling problem in single-hop all-photonic networks with cross-connect switches and large propagation delays. We proposed the FMA, a novel scheduling algorithm that achieves zero rejection for admissible demands and provides weighted max-min fair allocation of free capacity. When the demand matrix is inadmissible, FMA minimizes the maximum percentage rejection experienced by any connection. We subsequently proposed the MRA, which ensures minimum global rejection and provides weighted max-min fair allocation/rejection to non-critical connections. Finally, we described a feedback control system that compensates for scheduling errors due to mispredictions and rejection. The controller design is based on the Smith principle, which removes the destabilizing delays from the feedback loop by using a "loop cancelation" technique.

OPNET Modeler simulations indicate that FMA and MRA achieve similar performance in terms of total rejection, but there is a major difference in the fairness of the allocation of

rejection. A comparison with an alternative algorithm proposed for AAPNs, the projection method proposed by Peng et al. [21], suggests that FMA achieves better performance in terms of both rejection and queuing delay. Simulations of the feedback control system indicate that it reduces the response time to sudden changes in traffic intensity and imparts fairness by controlling the divergence from the average queue length.

Appendix A: Proof of Theorem 2

We first define a *bottleneck link* and state a lemma relating weighted max-min fairness and the existence of bottleneck links; the proof of the lemma appears in [27].

Definition 4 *Bottleneck Link* Given a feasible rate vector υ and a weight vector ω, we say that link ℓ is a *bottleneck link* with respect to $(\upsilon\ ,\ \omega)$ for a connection u crossing ℓ, if $C_\ell = \sum_k \upsilon_k \triangleq F_\ell$ and $\omega_u \geq \omega_k$ for all connections k crossing ℓ.

Lemma 1 *A feasible rate vector υ with weight vector $\omega = \{\frac{\upsilon_u}{R_u}\}$ is weighted max-min fair if and only if each connection has a bottleneck link with respect to $(\upsilon\ ,\ \omega)$.*

Proof (*Proof of Theorem 2*) Let $u \in \{(i, j), 1 \leq i, j \leq N\}$ index the source–destination connections specified by the demand matrix. We focus on the properties of the modified demand matrix and associated sets at various iterations of the while loop in Algorithm 1, so we index entities by iteration number and note that this indicates the value of the entity at the *start* of the iteration. For example, $\mathcal{A}_D(h)$ denotes the set of unmodified overloaded lines at the start of iteration h of the algorithm.

We prove that FMA achieves weighted max-min fair allocation of the overloaded demand. During each iteration h of the while-loop, FMA identifies the line $\gamma \in \mathcal{A}_D(h)$ such that $G_\gamma(h) = \min\{G_\ell(h);\ \ell \in \mathcal{A}_D(h)\}$. It alters the demands in $a_\gamma(h)$ according to (2) and after this modification, there is no subsequent modification of these demands. Substituting (2) into the definition of the weight, we have $\omega_u = 1 + G_\gamma(h)$ for all $u \in a_\gamma(h)$.

We demonstrate that the adjustment at iteration h leads to γ being a bottleneck link (line) for $u \in a_\gamma(h)$, i.e., after this adjustment it holds that $\omega_z \leq \omega_u$ for $u \in a_\gamma(h)$ and $z \in b_\gamma(h)$. Equivalently, we prove that $\min\{G\}$ is monotonically increasing with respect to the iteration number, i.e., $\min\{G(h)\} \leq \min\{G(h+1)\}$. The equivalence follows since the ω_z are obtained from adjustments prior to iteration h.

Suppose that line β has minimum G at iteration $h + 1$. Lines γ and β have at most one connection (demand) in common. If there is no common connection, then $G_\beta(h + 1) = G_\beta(h) \geq G_\gamma(h)$. If there is a common connection k, then:

$$LS_\beta(h + 1) = LS_\beta(h) + D_k(\omega_k - 1) \tag{17}$$

$$S_{a_\beta}(h + 1) = S_{a_\beta}(h) - D_k \tag{18}$$

and hence

$$\begin{aligned} G_\beta(h + 1) &= \frac{L - LS_\beta(h) - D_k(\omega_k - 1)}{S_{a_\beta}(h) - D_k} \\ &= \frac{S_{a_\beta}(h)G_\beta(h) - D_k(\omega_k - 1)}{S_{a_\beta}(h) - D_k} \\ &\geq G_\gamma(h) \end{aligned} \tag{19}$$

where the last inequality follows from substitution based on $G_\beta(h) \geq G_\gamma(h) = \omega_k - 1$.

Thus the application of FMA upon an inadmissible demand matrix D leads to the generation of a bottleneck link for each connection u with weight $\omega_u = \frac{D'_u}{D_u}$. By Lemma 1, this establishes that FMA achieves weighted max-min fair allocation of adjusted demands D'. □

Appendix B: Proof of Theorem 3

Proof We approach the proof by contradiction. Consider a matrix R^* that achieves minimum rejection and suppose that it cannot be decomposed in the form $R^* = A + Q$ outlined in the theorem statement. Let $R^*_{\mathcal{C}}$ denote the matrix formed by setting all elements of R^* to zero except those where $(h, p) \in \mathcal{C}$.

If there exists an element $(h, p) \in \mathcal{C}$ such that $r_h(R^*_{\mathcal{C}}) < r_h(D) - L$, $c_p(R^*_{\mathcal{C}}) < c_p(D) - L$ and $R^*_{\mathcal{C}}(h, p) < D(h, p)$, then it is clear that we can form a new rejection matrix R' by (i) setting $R'(i, j) = R^*(i, j)$ for all $(i, j) \neq (h, p)$; (ii) setting $R'(h, p) = R^*(h, p) + \delta$ for some $\delta > 0$; and (iii) reducing one or more of the non-critical elements of the line $R'(h, \cdot)$ by a sum total of δ, and doing the same for the column $R'(\cdot, p)$. The total rejection of R' is less than R^*, contradicting the assumption that R^* is a minimum-demand matrix.

We must therefore be able to construct a decomposition $R^* = A^* + Q^*$, where Q^* satisfies the same properties as Q (if we replace A by A^*), and A^* is a matrix that satisfies the constraints of the $MAXREJFLOW$ problem, and at least one of (5–7) with equality. With this decomposition, we can write the following expression for $|R^*|$:

$$|R^*| = \sum_h \sum_p (A^* + Q^*) \tag{20}$$

$$= |A^*| + \sum_{h \in O_r} \left(r_h(D) - L - r_h(A^*)\right) + \sum_{p \in O_c} \left(c_p(D) - L - c_p(A^*)\right) \tag{21}$$

$$= \sum_{h \in O_r} (r_h(D) - L) + \sum_{p \in O_c} (c_p(D) - L) - |A^*| \tag{22}$$

Since A^* satisfies one of the constraints for each (h, p) with equality, the matrix Q^* can contribute additional rejection on either the row h or the column p, but not both, so we do not double-count rejection in (21).

Now consider an alternative, arbitrary rejection matrix R that can be decomposed as $R = A + Q$. This is always possible because a water-filling procedure (such as FMA) can be used to identify a satisfactory matrix Q. Since A also satisfies at least one of the three latter constraints of *MAXREJFLOW* with equality, an equivalent expression to (22) is possible with $|R|$ replacing $|R^*|$, and $|A|$ replacing $|A^*|$. The first two terms are only dependent on D and L, and $|A| > |A^*|$, since A is the maximum flow solution satisfying the specified constraints. It follows that $|R| < |R^*|$, contradicting the assumption that R^* is a minimum rejection matrix. □

Appendix C: Robustness analysis

Gain margin and phase margin only measure robustness with respect to model parameters, which are independent of frequency ω. Since systems perform differently at different frequencies, we need to find a tighter bound on the phase and gain margins with respect to the frequency of the system. As a widely accepted and more useful robustness indicator, we define $M \triangleq \max_\omega |H_r(j\omega)|$, the maximum of the closed-loop transfer function. The following relationships establish lower bounds on ϕM and GM [24]:

$$GM \geq 1 + \frac{1}{M} \tag{23}$$

$$\phi M \geq 2 \sin^{-1}\left(\frac{1}{2M}\right) \simeq \frac{1}{M} \tag{24}$$

In the proposed control model the major error occurs due to the mismatch between the dead-time model and the actual delay experienced by the data. This may cause the system to cross its stability limits. Suppose that the dead-time of the actual plant exceeds the dead-time T in our model by the quantity δ. This error introduces a phase lag of $\omega\delta$ at frequency ω. Therefore, the system remains stable if $\delta < \frac{\phi M}{\omega_c}$, where ω_c is the crossover frequency at which the open-loop system gain drops to unity. When (24) is substituted into this equation, a more conservative condition $\delta < \frac{1}{\omega_c M}$ is obtained.

For the transfer function obtained in (8) $M = 1$, and the above condition is transformed to $\delta < \frac{1}{\omega_c}$. Since it is not possible to obtain the crossover frequency for our system explicitly (due to the time delay in the transfer function), one approach is to represent the dead-time as a first-order Padé approximation [24]:

$$e^{-Ts} = \left(\frac{1 - \frac{T}{2}s}{1 + \frac{T}{2}s}\right) \tag{25}$$

Then the crossover frequency can be approximated as $\omega_c = \frac{1}{T + \frac{1}{x_{ij}K_r}}$. Substituting this equation into $\delta < \frac{1}{\omega_c}$ leads to:

$$\delta < T + \frac{1}{x_{ij}K_r} \tag{26}$$

This equation confirms that our designed controller is stable for errors as large as the actual dead-time. It also indicates that smaller values of the gain K_r make the system more resilient to error, but this of course has the disadvantage of slowing the system response.

Acknowledgements This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and industrial and government partners through the Agile All-Photonic Networks (AAPN) Research Network.

References

[1] Anderson, T., Owicki, S., Saxe, J., Thacker, C.: High-speed switch scheduling for local-area networks. ACM Trans. Comp. Syst. **11**(4), 319–352 (1993)

[2] Bauer, P.H., Sichitiu, M.L., Ernst, R., Premaratne, K.: A new class of Smith predictors for network congestion control. In: Proceedings of International IEEE Conference on Electronics, Circuits, and Systems (ICECS), St. Julian's, Malta, pp. 685–688 (2001)

[3] Bianco, A., Careglio, D., Finochietto, J., Galante, G., Leonardi, E., Neri, F., Solé-Pareta, J., Spadaro, S.: Multiclass scheduling algorithms for the DAVID metro network. IEEE J. Sel. Area Comm. **22**(8), 1483–1496 (2004)

[4] Bochmann, G., Coates, M., Hall, T., Mason, L., Vickers, R., Yang, O.: The agile all-photonic network: An architectural outline. In: Proceedings of Queens' Biennial Symposium on Communications, Kingston, Canada, pp. 217–218 (2004)

[5] Bogineni, K., Sivalingham, K.M., Dowd, P.W.: Low-complexity multiple access protocols for wavelength-division multiplexed photonic networks. IEEE J. Sel. Area Comm. **11**(4), 590–604 (1993)

[6] Cormen, T., Leiserson, C., Rivest, R., Stein, C.: Introduction to Algorithms. 2nd edn. MIT Press, Cambridge, MA (2001)

[7] Crescenzi, P., Deng, X., Papadimitriou, C.H.: On approximating a scheduling problem. J. Comb. Optim. **5**(3), 287–297 (2001)

[8] Ford, L.R. Jr., Fulkerson, D.R.: Maximal flow through a network. Can. J. Math. **8**, 399–404 (1956)

[9] Ganz, A., Gao, Y.: Efficient algorithms for SS/TDMA scheduling. IEEE Trans. Comm. **40**(6), 1367–1374 (1992a)

[10] Ganz, A., Gao, Y.: A time-wavelength assignment algorithm for a WDM star network. In: Proceedings of IEEE INFOCOM, Florence, Italy, vol. 3, pp. 2144–2150 (1992)

[11] Gopal, I.S., Wong, C.K.: Minimizing the number of switchings in an SS/TDMA system. IEEE Trans. Comm. **33**, 1497–1501 (1985)

[12] Keslassy, I., Kodialam, M., Lakshman, T., Stiliadis, D.: Scheduling schemes for delay graphs with applications to optical packet networks. In: Proceedings of IEEE Workshop High Performance Switching and Routing, Phoenix, AZ (2003)

[13] Liu, X., Saberi, N., Coates, M., Mason, L.: A comparison between time-slot scheduling approaches for all-photonic networks. In: Proceedings of International Conference on Inf., Comm. and Signal Processing (ICICS), Bangkok, Thailand, pp. 1197–1201 (2005)

[14] Marsan, M., Bianco, A., Leonardi, E., Neri, F., Nucci, A.: Simple on-line scheduling algorithms for all-optical broadcast-and select networks. IEEE Eur. Trans. Telecom. **11**(1), 109–116 (2000)
[15] Mascolo, S.: Congestion control in high-speed communication networks using the Smith principle. Automatica **35**(12), 1921–1935 (1999)
[16] Mason, L., Vinokurov, A., Zhao, N., Plant, D.: Topological design and dimensioning of agile all photonic networks. Comput. Netw. **50**(2), 268–287 (2006)
[17] Mataušek, M., Micić, A.: A modified Smith predictor for controlling a process with an integrator and long dead-time. IEEE Trans. Automat. Cont. **41**(8), 1199–1203 (1996)
[18] McKeown, N. Scheduling algorithms for input-queued cell switches. Ph.D. thesis, University of California at Berkeley (1995)
[19] McKeown, N., Anantharam, V., Walrand, J.: Achieving 100% throughput in an input-queued switch. In: Proceedings of IEEE INFOCOM, San Francisco, CA, vol. 1, pp. 296–302 (1996)
[20] Opnet Technologies, Inc: OPNET modeler 12.1. http://www.opnet.com (2008)
[21] Peng, C., Paredes, S., Hall, T.J., von Bochmann, G.: Constructing service matrices for agile all-optical cores. In: Proceedings of IEEE International Symposium on Computers and Communication (ISCC), Sardinia, Italy, pp. 967–973 (2006)
[22] Pomalaza-Raez, C.A.: A note on efficient SS/TDMA assignment algorithms. IEEE Trans. Comm. **36**, 1078–1082 (1988)
[23] Ramaswami, R., Sivarajan, K.: Routing and wavelength assignment in all-optical networks. IEEE/ACM Trans. Netw. **3**(5), 489–500 (1995)
[24] Rivera, D.E., Morari, M., Skogestad, S.: Internal model control. 4. PID controller design. Ind. Eng. Chem. Proc. Design Dev. **25**, 252–265 (1986)
[25] Rouskas, G.N., Ammar, M.H.: Analysis and optimization of transmission schedules for single-hop WDM networks. In: Proceedings of IEEE INFOCOM, San Francisco, CA, vol. 3, pp. 1342–1349 (1993)
[26] Saberi, N.: Bandwidth allocation and scheduling in photonic networks. PhD thesis, McGill University (2007)
[27] Saberi, N., Coates, M.: Fair matching algorithm: An optimal scheduling algorithm for the AAPN network. Technical report, McGill University, Montreal, Canada (2005), available at http://www.tsp.ece.mcgill.ca/Networks/publications.html
[28] Saberi, N., Coates, M.: Fair matching algorithm: Fixed-length frame scheduling in all-photonic networks. In: Proceedings of IASTED International Conference Optical Communication Systems and Networks, Alberta, Canada, pp. 213–218 (2006)
[29] Saberi, N., Coates, M.: Minimum rejection scheduling in all-photonic networks. In: Proceedings of IEEE BROADNETS, San Jose, CA, pp. 1–10 (2006)
[30] Saberi, N., Coates, M.: Feedback control system for scheduling of wide-area all-photonic networks. In: Proceedings of IEEE International Symposium Computers and Communication (ISCC), Aveiro, Portugal, pp. 115–120 (2007)
[31] Sang, A., Li, S.Q.: A predictability analysis of network traffic. In: Proceedings of IEEE INFOCOM, Tel Aviv, Israel, vol. 1, pp. 342–351 (2000)
[32] Smith, O.: A controller to overcome dead-time. J. ISA **6**(2), 28–33 (1959)
[33] Towles, B., Dally, W.J.: Guaranteed scheduling for switches with configuration overhead. IEEE/ACM Trans. Netw. **11**(5), 835–847 (2003)
[34] Xu, L., Perros, H., Rouskas, G.: Techniques for optical packet switching and optical burst switching. IEEE Comm. Mag. **39**(1), 136–142 (2001)
[35] Zheng, J., Peng, C., von Bochmann, G., Hall, T.J.: Load balancing in all-optical overlaid-star tdm networks. In: Proceedings of IEEE Sarnoff Symposium, Princeton, NJ, pp. 1–4 (2006)

Author Biographies

N. Saberi received her Ph.D. degree in Electrical Engineering from McGill University in 2007. She was a Visiting Scholar at the Johns Hopkins University from April to August 2007. Currently, she is a Postdoctoral Fellow in the School of Engineering and Applied Sciences at Harvard University. Her research interests include network scheduling algorithms, congestion control, and resource allocation in wireless and photonic networks.

M. J. Coates received the B.E. degree (first class honors) in Computer Systems Engineering from the University of Adelaide, Australia, in 1995, and a Ph.D. degree in Information Engineering from the University of Cambridge, U.K., in 1999. Currently, he is an Associate Professor at McGill University, Montreal, Canada. He was awarded the Texas Instruments Postdoctoral Fellowship in 1999 and was a Research Associate and Lecturer at Rice University, Texas, from 1999 to 2001. His research interests include communication and sensor/actuator networks, statistical signal processing, causal analysis, and Bayesian and Monte Carlo inference.