Achieving Fairness in Generalized Processor Sharing for Network Switches

Raymond Yim, Natasha Devroye, Vahid Tarokh and H. T. Kung
Division of Engineering and Applied Sciences, Harvard University
33 Oxford Street, Cambridge, MA 02138, USA

Abstract— **In Generalized Processor Sharing (GPS), the goal of the scheduler is to maximize the processor utilization, while maintaining a rate guarantee and fairness in the bandwidth allocation to each input stream according to the quality of service (QoS) criteria. While in the single server case, these requirements can be satisfied simultaneously by scheduling service rate according to a scale version of the rate request, the result does not generalize to the case when multiple processors are used. This paper extends the notion of max-min and proportional fairness for single node Generalized Processor Sharing in Network Switches (GPS-SW), and provides rate augmentation algorithms in achieving the two performance criteria.**

Keywords— **Switching and Scheduling, Generalized Processor Sharing, Max-Min Fairness, Proportional Fairness**

I. Introduction

Generalized Processor Sharing (GPS) [8] involving a single server has been studied extensively over the past decade. Given a quality of service (QoS) contract, the GPS problem computes a fair fluid-based service schedule for a work conserving server. Such a server must be busy if there are packets waiting in the system. If service is demanded from a particular traffic input, but that input does not have any packet waiting to be served, GPS scales the rate allocated to the remaining inputs accordingly so that the server remains work conserving. Once we understand the idealized fluid-based service schedule, we can then derive an optimal packetizing algorithm that allows data to be served fairly in a packet-by-packet fashion. In single server systems, it is well known that Weighted Fair Queuing (WFQ) [6], otherwise known as Packet-by-Packet Generalized Processor Sharing (PGPS) [8], can closely approximate the GPS policy, differing by no more than one maximum size packet.

There have been many attempts to generalize the single-server GPS result to multiple server systems. For example, Chang *et al.* [5] provide an online algorithm that packetizes any time-invariant rate request for each input-output pair in an $N \times N$ switch. However, their rate augmentation algorithm does not capture any performance criterion in the bandwidth gain of each stream. In other attempts [4], [9], fair scheduling is considered for multiple identical servers. In particular, packets have no preferrence over the servers to which they are sent for processing. This is different from the scenario for packet switches, where each packet is sent from a specific input to a specific output. In [10], fairness is addressed when a single packet demands service from multiple different processors simultaneously; however, they do not consider any service rate guarantee in their formulation.

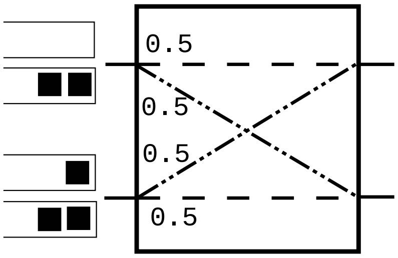

Fig. 1. Scenario in which a switch cannot achieve work conservation and satisfy service rate guarantees simultaneously

Unlike traditional GPS, in general, it is impossible to ensure work conservation and service rate guarantees simultaneously in multiple-server switches. In this article, we call the traffic going through each input-output port pair as a *stream*. Consider the scenario in Figure 1. Each stream demands a service rate guarantee of 50%. However, there is no traffic to be serviced from input port 1 to output port 1. Nonetheless, if the service rate of any other stream is increased, it must be done at the cost of reducing the service rate of another stream. This would violate the rate guarantee constraints. Hence, in such scenrio, output port 1 can maximally be using only 50% of its capacity if rate guarantees are to be satisfied.

The contribution of this work is twofold. First, we extend the definition of two performance criteria, max-min and proportional fairness, in such a way that it will capture the rate guarantee constraint of each traffic stream in the switch. Furthermore, these performance criteria will allow the resulting service rate to acheive maximal server utilization. Secondly, we provide two rate augmentation algorithms that satisfy the performance criteria. In Section II, we formulate the concept of Generalized Processor Sharing for Network Switches (GPS-SW). In Section III and IV, we extend the notion of max-min and proportional fairness in GPS-SW respectively. Also, we provide rate augmentation algorithms to satisfy the two criteria. Finally, we conclude in Section V.

II. Problem Formulation

Let $\Phi = (\phi_{ij})$ be an $N \times N$ substochastic matrix[1] with ϕ_{ij} being the rate requested by the traffic from input i to output j in an $N \times N$ per-stream input-buffered crossbar switch. Furthermore, if no packet is available to be

This material is based upon research supported by the National Science Foundation under the Alan T. Waterman Award, Grant No. CCR-0139398. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation.

[1] A matrix $M = (m_{ij})$ is said to be *substochastic* if, for each matrix index (p, q), $m_{pq} \geq 0$, $\sum_{i=1}^{N} m_{iq} \leq 1$ and $\sum_{j=1}^{N} m_{pj} \leq 1$.

switched from input port p to output port q, ϕ_{pq} is assumed to be zero.

The goal of GPS-SW is to choose an $N \times N$ service rate matrix $R = (r_{ij})$ that satisfies three criteria: i) it must provide at least the rate requested from each stream; ii) it allocates bandwidth to each stream fairly according to their rate requests; and iii) it allows the server to be used maximally by minimizing server idle time. The following definitions are introduced to formally define R.

Definition 1: A matrix $M = (m_{ij})$ is said to *dominate* a matrix $P = (p_{ij})$ if, for each matrix index (i, j), $m_{ij} \geq p_{ij}$.

Definition 2: Consider a substochastic matrix $M = (m_{ij})$. We say a non-zero matrix element at position (p, q) is *non-augmentable* if an arbitrary increase in m_{pq} will cause another matrix element to decrease in value in order for M to remain substochastic. A row or a column of M is said to be *non-augmentable* if all non-zero valued matrix elements in the row or column are non-augmentable. A matrix is said to be *non-augmentable* if all non-zero valued matrix elements in the matrix are non-augmentable.

Using these definitions, our objective is to find a non-augmentable substochastic service rate matrix R that dominates the rate request matrix Φ while satisfying some performance criterion.

The following lemma characterizes the property of a non-augmentable matrix.

Lemma 1: A substochastic matrix $M = (m_{ij})$ is non-augmentable if and only if, for all $m_{pq} \neq 0$,

$$\left(\sum_{j=1}^{N} m_{pj} - 1\right)\left(\sum_{i=1}^{N} m_{iq} - 1\right) = 0. \tag{1}$$

Proof: Proof is omitted as it is easy to derive from the definition of augmentability. ■

III. Max-Min Fairness

We extend the notion of max-min fairness [1] to take into account the rate request from each traffic stream. In particular, the modified max-min fair augmentation aims to obtain a service rate matrix that is closely related to a scale version of the service request, and it does so by giving priority to the streams that can be scaled the least.

Definition 3: Given the rate request matrix $\Phi = (\phi_{ij})$, we say $M = (m_{ij})$ is a *feasible matrix* if, by letting $r_{ij} = m_{ij}\phi_{ij}$, $R = (r_{ij})$ is non-augmentable and substochastic, and $m_{ij} \geq 1$. Furthermore, we say R is the *max-min fair* augmentation of Φ if, for each non-zero valued matrix element at position (p, q) in Φ, and for any other feasible matrix $\hat{M} = (\hat{m}_{ij})$ for which $m_{pq} < \hat{m}_{pq}$, there exits some (p', q') with $m_{pq} \geq m_{p'q'}$ and $m_{p'q'} > \hat{m}_{p'q'}$.

Note that max-min fairness usually applies to multiple node networks [1]. However, in this study, the above definition is a natural verion of max-min fairness for single node case. The following algorithm achieves max-min fairness.

Algorithm 1:

```
1. Initialize $R = \Phi$.
2. If $R$ is non-augmentable, stop.
3. Choose a maximum possible $k > 1$ such that, after
   setting $r_{pq} \leftarrow k r_{pq}$ for each augmentable matrix element
   at position $(p, q)$, $R$ would remain substochastic.
   Then, set $r_{pq} \leftarrow k r_{pq}$ for each augmentable element.
4. Go to Step 2.
```

We show that Algorithm 1 produces a service rate matrix that satisfies the max-min fair criterion. At initialization, there are at most N rows and N columns that can be non-augmentable. At each iteration, the algorithm scales all augmentable elements in the matrix in a such way that at least one new row or column becomes non-augmentable. Hence, the total number of non-augmentable row and column is reduced by at least one. The algorithm will terminate in at most $2N$ steps.

We demonstrate this algorithm by the following example.

Example 1: Let $\Phi = \begin{bmatrix} \frac{1}{5} & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{10} \\ \frac{2}{5} & 0 & \frac{2}{5} \end{bmatrix}$. The algorithm first set $R = \Phi$. In the first iteration, the algorithm multiplies each matrix element in R by $\frac{5}{4}$, as this causes the last row of R to become non-augmentable. In subsequent steps, the elements on the last row are non-augmentable, thus they will not get multiplied further. The algorithm proceeds as follows. Note that the last matrix is non-augmentable.

$$\begin{bmatrix} \frac{1}{4} & 0 & 0 \\ 0 & \frac{5}{8} & \frac{1}{8} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} \xrightarrow{\times\frac{4}{3}} \begin{bmatrix} \frac{1}{3} & 0 & 0 \\ 0 & \frac{5}{6} & \frac{1}{6} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} \xrightarrow{\times\frac{3}{2}} \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{5}{6} & \frac{1}{6} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}$$

IV. Proportional Fairness

The traditional proportional fairness as defined by Kelly *et al.* [7] does not capture rate guarantee constraints. We propose a new definition of proportional fairness in the following way.

Definition 4: Let $R = (r_{ij})$ and k be the solution to the following optimization problem:

$$\min \sum_{i,j} \left| r_{ij} - k\phi_{ij} \right|, \tag{2}$$

subject to

$$r_{pq} = 0 \quad if \quad \phi_{pq} = 0, \tag{3}$$

$$r_{pq} \geq \phi_{pq} \quad \forall \quad p, q, \tag{4}$$

$$k \geq 0, \tag{5}$$

and R is non-augmentable. Then, we say R is a *proportional fair* solution to $\Phi = (\phi_{ij})$, and k is the proportional increase of Φ.

This optimization problem is non-linear, as the service rate matrix must satisfy at most N^2 non-augmentability constraints in (1). Naively, a solution can be found by considering 2^{N^2} linear subproblems and selecting the solution with minimal objective value, but such solution method has a forbidding complexity.

The speed of the simplex algorithm [2] is limited primarily by the number of constraint equations. Hence, we propose an alternative formulation that will reduce the number of constraints in the optimization problem. The following definitions are needed for the formulation.

Definition 5: A matrix $M = (m_{ij})$ is said to be a *zero-enforced permutation matrix* of matrix $B = (b_{ij})$ if, letting $P = (p_{ij})$ be a permutation matrix[2],

$$m_{ij} = p_{ij} 1_{\{b_{ij} \neq 0\}}, \tag{6}$$

[2] An $N \times N$ matrix P is said to be a *permutation matrix* if $P \in \{0,1\}^{N \times N}$, and each row and column of P sums to one.

where $1_{\{\cdot\}}$ is the indicator function.

Definition 6: A zero-enforced permutation matrix M of B is said to be *non-absorbable* if it does not have another zero-enforced permutation matrix $\hat{M}$ of B, $\hat{M} \neq M$, such that $\hat{M}$ dominates M. Otheriwse, we say the zero-enforced matrix is *absorbable.*

The two following lemmata allow us to transform the non-augmentability constraints by considering the service rate matrix R as a convex combination of non-absorbable zero-enforced permutation matrices of Φ.

Lemma 2: Let $R = (R_{ij})$ be a matrix such that $r_{ij} = 0$ whenever $\phi_{ij} = 0$. If a matrix R is non-augmentable, then it can be decomposed as a convex combination of non-absorbable zero-enforced permutation matrices of Φ.

Proof: We first show that if R is a non-augmentable matrix, it can always be decomposed into a convex sum of zero-enforced permutation matrices. Consider a doubly stochastic matrix $\hat{R}$ that is created by augmenting only the zero-valued elements in R. By Birkoff decomposition [3], $\hat{R}$ can be decomposed as a convex combination of permutation matrices. Now, if we set every element at position (p, q) in each permutation matrix to zero whenever $r_{pq} = 0$, R can be decomposed as a convex combination of these zero-enforced permutation matrices.

We now show that if some of the zero-enforced permutation matrices in the decomposition are absorbable, then matrix R is augmentable. By definition, every absorbable zero-enforced matrix can be dominated by a non-absorbable matrix. Hence, by replacing the absorbale matrices by the corresponding non-absorbable matrices in the decomposition, the new matrix will dominate R. This implies R is augmentable. ■

Lemma 3: Let $P = \{P^1, \cdots, P^z\}$ be the set of all non-absorbable zero-enforced permutation matrices of Φ, and let p^n_{ij} be the entries of matrix P^n for $n \in \{1, \cdots z\}$. Let $\lambda = (\lambda_1, \cdots, \lambda_z)$ be a vector of dimension $z = |P|$. Suppose λ and k are the solution to the optimization problem:

$$\min \sum_{i,j} \left| \sum_{n:p^n_{ij}=1} \lambda_n - k\phi_{ij} \right|, \tag{7}$$

subjected to

$$\sum_{n:p^n_{ij}=1} \lambda_n \geq \phi_{ij} \quad \forall \quad i, j, \tag{8}$$

$$\sum_n \lambda_n = 1, \tag{9}$$

$$\lambda_n \geq 0 \quad \forall \quad n, \tag{10}$$

$$k \geq 0, \tag{11}$$

and the service rate matrix $R = \sum_{n=1}^{z} \lambda_n P^n$ satisfies the non-augmentability constraint. Then, we say R is the proportional fair solution to Φ, and k is the proportional increase of Φ.

Proof: It follows by transforming the definition of proportional fairness. ■

The fact that λ needs to satisfy the non-augmentability constraint still causes the optimization problem to be nonlinear. However, this constraint is a disjunction of at most $2N$ terms, one for each row and column. This allows speed improvement in obtaining the result. We apply basic algorithms to achieve the solution:

Algorithm 2:

1. Use the simplex algorithm to solve the linear programming problem by ignoring the non-augmentability constraints.
2. For each non-augmentability constraint, use the dual simplex algorithm [2] to obtain a revised optimal λ, or the problem becomes infeasible.
3. Choose λ that gives the minimum cost among all solutions.

This algorithm is a direct application of optimization techniques, and it can be shown that it will gives the proportional fair solution.

We illustrate the algorithm using the following example.

Example 2: Using Φ from Example 1. We first write the service rate matrix R as a convex combination of non-absorbable zero-enforced permutation matrices. That is, $R = \lambda_1 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \lambda_2 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} + \lambda_3 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} + \lambda_4 \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$. The non-augmentability constraint is $\lambda_2\lambda_3 = 0$. Solve the two linear optimization problems, one with $\lambda_2 = 0$ and one with $\lambda_3 = 0$, the minimum cost result from $\lambda_2 = 0$. The resulting $R = \begin{bmatrix} 0.4 & 0 & 0 \\ 0 & 0.85 & 0.15 \\ 0.6 & 0 & 0.4 \end{bmatrix}$.

V. Conclusions

We have defined two performance criteria for single node Generalized Processor Sharing for Network Switches (GPS-SW) in the context of $N \times N$ network switches. These performance criteria take into account the service request from each stream while maximizing server utilization. Furthermore, we have proposed two rate augmentation algorithms that can achieve the two criteria. As demonstrated in the examples, the two solutions are not the same in general. Hence, a network switch designer should choose the appropriate criterion that is best suited to their application.

References

[1] D. Bertsekas and R. Gallager, *Data Networks,* Englewood Cliffs, NJ: Prentice Hall, 1987.

[2] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization,* Belmont, MA: Athena Scientific, 1997.

[3] G. Birkhoff, "Tres observaciones sobre el algebra lineal," *Univ. Nac. Tucumán Rev. Ser. A,*, vol. 5, pp. 147-151, 1946.

[4] J. M. Blanquer and B. Ozden, "Fair queuing for aggregate multiple links," *Proc. ACM SIGCOMM'01*, pp. 189-197, August 2001.

[5] C. S. Chang, W. J. Chen and H. Y. Huang "Birkhoff-von Neumann Input Buffered Crossbar Switches,", *Proc. IEEE INFOCOM'03*, vol. 3, pp. 1624-1633, March 2000.

[6] A. Demers, S. Keshav and S. Shenker, "Analysis and simulation of a fair-queuing algorithm," *Proc. ACM SIGCOMM'89*, pp. 1-12, September 1989.

[7] F. Kelly, A. Maulloo and D. Tan, "Rate control in communication networks: shadow prices, proportional fairness and stability," *J. Operational Research Society*, pp. 237-252, 1998.

[8] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: The single-node case", *Proc. IEEE INFOCOM'92*, vol. 2, pp. 915-924, May 1992.

[9] A. Srinivasan, P. Holman, J. H. Anderson and S. Banruah, "The case for fair multiprocessor scheduling", *Proc. Parallel and Distr. Processing Symp.*, pp. 22-26, April 2003.

[10] Y. Zhou and H. Sethu, "On achieving fairness in the joint allocation of processing and bandwidth resources", submitted to *IEEE Trans. Networking.*