Analysis of PLR for Shared Switch Fabrics

Chankyun Lee and June-Koo Kevin Rhee

A modular upgrade design for packet transport switch nodes is presented where packet loss is dramatically reduced by intra-module and inter-module buffer sharing. This modular design offers significant cost and power reduction in a high-data-rate system where buffers are highly costly and power-greedy.

Keywords: Packet switching, contention resolution, shared buffering, modular switch upgrade.

I. Introduction

Novel design concepts of high-end tera-class packet transport switches have become a major research interest with concerns of power and cost efficiency. For packet transport switch networks, power and cost optimizations become critical to success in addition to graceful switch capacity upgrade capability. In a typical modular system design, sharing of buffers in switch modules can improve power and cost merits significantly. This letter proposes an efficient design to share power-greedy and costly buffer modules, especially for an optical switch network, and reports on an analytical model of packet loss rate (PLR) for a switch design with *intra-module* and *inter-module buffer sharing*.

In a shared buffer design, a switch module with intra-module buffer sharing can enhance PLR by simply increasing the buffer sharing ratio [1]. One can consider a graceful switch node upgrade in a modular switch design as illustrated in Fig. 1. In this design, an additional benefit in the PLR performance can be gained if an inter-module buffer sharing scheme is combined. We consider such two-fold sharing gains in PLR performance. Figure 1 illustrates partially shared $N \times N$ switch architecture with N links. Each link can have W channels to provide an aggregate link capacity. Switch capacity is upgraded by the addition of channels. In this design, W switch modules can share buffers via partial connections with (N–B) ports. Remaining B ports connect the switch fabric to its buffer.

The sharing mechanism is achieved when W switch modules are connected in a directional cyclic pattern as shown in Fig. 1. Consider a switch as the *current switch*. When packet contentions occur in the current switch, contended packets are scheduled to be forwarded to its buffer ports first. If the number of contended packets is more than the number of buffer ports, the rest are forwarded to the *first next neighbor switch* in the cycle. Then, this switch schedules the received contended packets to be forwarded to its buffers; later, these packets are forwarded to the destined link at the wavelength of this switch. If there are not enough buffer ports, the excessive packets are forwarded again to the *second next neighbor switch*. This process can be repeated for (W–1) steps. In this way, inter-module buffer sharing is achieved, which can dramatically reduce the total number of buffer interfaces. This scheme, of course, leaves a packet sequencing penalty that needs to be resolved at the destined end node. With a large W, the buffer sharing ratio B/N requirement can be greatly reduced, and the

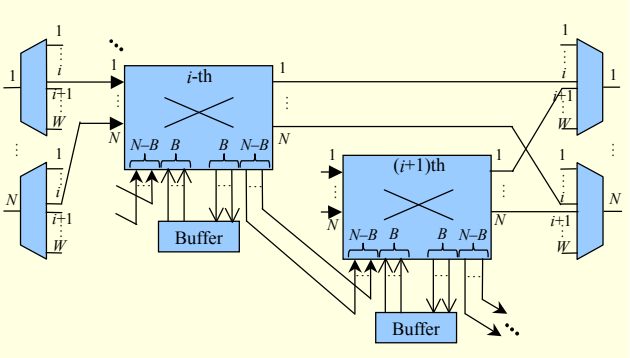

Fig. 1. Architecture of partially shared switch fabrics.

Manuscript received May. 17, 2010; revised July 23, 2010; accepted Aug. 9, 2010.

This research was supported by the Ministry of Knowledge Economy (MKE), Rep. of Korea, under the Information Technology Research Center (ITRC) support program supervised by the National IT Industry Promotion Agency (NIPA) (NIPA-2010-(C1090-1011-0004)).

Chankyun Lee (phone: +82 42 350 7516, email: ck.lee@kaist.ac.kr) and June-Koo Kevin Rhee (corresponding author, email: rhee.jk@kaist.ac.kr) are with the Department of Electrical Engineering, KAIST, Daejeon, Rep. of Korea.

doi:10.4218/etrij.11.0210.0168

switch can still achieve sufficiently low PLR performance. This is a key factor to reduce the power and cost of the high performance packet transport switch when the buffer interface is power-greedy and costly, such as in an optical system.

II. Analytical Packet Loss Rate Model

The PLR is considered a key performance measure of a network. Previous research groups have introduced analytical PLR models for packet switching models [2], [3]. Most of these approaches are formulated based on a macroscopic point of view of a particular output port. In this section, we derive a PLR model that can evaluate the PLR of a W-fold shared switch fabric and buffer model, which estimates the reduced PLR of the model in Fig. 1.

1. Single Bufferless Switch Module

Let us consider a packet traffic model with offered load ρ that follows a Bernoulli process of arriving at a bufferless $N\times N$ switch fabric. Then, the PLR of a single bufferless switch model in a macroscopic particular port approach is calculated by an average total number of lost packets at the particular output port over an average number of packets arriving at the particular output port [2]:

$$\frac{1}{N\mu}\sum_{k=2}^{N}(k-1)\binom{N}{k}\mu^k(1-\mu)^{N-k}, \quad (1)$$

where μ is the offered load for particular output port, ρ/N.

PLR is quite simply calculated with a macroscopic analysis of a particular port approach. However, in order to analyze the packet-forwarding on partially shared switch modules, one needs to take into account the probability of individual events of packet arrivals, buffering, and forwarding. Let us consider an event of K number of packets arriving at the core of a switch fabric with probability $P(K)$, where $0\le K\le N$. The probability of X number of packet losses $P(X)$ due to contention can be calculated as

$$P(X)=\sum_{k=2}^{N}P(K=k)P(X\mid K=k), \quad (2)$$

where $P(K=k)={}_N C_k\rho^k(1-\rho)^{N-k}$ for the Bernoulli process. $P(X\mid K)$ is the probability of X number of packet losses among K input packets. This probability can be modeled by mapping K input packets to (K–X) output ports. The number of possible selection of (K–X) output ports out of N output ports can be written as ${}_N C_{K-X}$. Then, the number of cases of mapping K input packets to particular (K–X) output ports is denoted as $F^{\text{parts}}_{K,X}$, which is the number of permutations of (K–X) indistinguishable objects to K such that [4]:

$$F^{\text{parts}}_{K=k,X}=\sum_{a_1+\cdots+a_{k-X}=k}\frac{k!}{a_1!\cdots a_{k-X}!}\ \text{for all}\ a_i>0, \quad (3)$$

where a_i is the number of packets competing for the i-th output port among particular (K–X) output ports. Considering the total number of possible partitions of K input packets to N output ports, the probability of X packets lost out of K packets is written as

$$P(X\mid K)=\binom{N}{K-X}\frac{F^{\text{parts}}_{K,X}}{N^K}. \quad (4)$$

Consequently, the PLR for a simple bufferless switch derived by the microscopic approach can be written as

$$PLR_0=\frac{1}{N\rho}\sum_{k=2}^{N}P(K=k)\sum_{x=1}^{k-1}xP(X=x\mid K=k). \quad (5)$$

The denominator $N\rho$ in (5) is the average total number of packets arriving at the core of a switch fabric.

2. Partially Shared Switch Module

There are three causes of packet loss events in partially shared switch modules of the switch model in Fig. 1. Such situations happen when the contended packets cannot be resolved as follows.

i) If the sum of packets passed from previous switches and contended packets of the current switch is larger than the number of drop ports N of the current switch, denoted as set I;

ii) If all buffer ports at all switch modules are so busy the excessive contended packets are not served by any buffer, denoted as set II; or

iii) If the queue length of a buffer has reached the buffer size, denoted as set III.

The queue length limit consideration has been studied in [3]. When the buffer size is large enough, we consider only cases of I and II for packet loss events.

For W switches with B buffers per switch, we first derive the average number of packet losses of I. We define x_n as the number of unserved packets in the previous n-th switch, and thus x_0 indicates those of the current switch. An average number of packet losses of I in the current switch due to the first to (W–1)th previous switches is calculated as nested summations of

$$\lambda_I^{(W-1)}=\sum_{x_0=0}^{N-1}P(X=x_0)\cdots\sum_{x_{W-1}=0}^{N-1}P(X=x_{W-1})\xi_I^{(0,W-1)}, \quad (6)$$

where $P(X)$ is given in (2). The number of lost packets in the n-th switch position due to its first to w-th previous switches in I is represented by $\xi_I^{(n,w)}$. For example, the number of lost

packets in the current switch due to its first and second previous switches in I can be expressed as

$$\xi_I^{(0,2)} = [x_0 + [\min(x_1 + [x_2 - B]^+, N) - B]^+ - N]^+, \quad (7)$$

where $[r]^+ \equiv \max(r, 0)$ for real number r. Here, $[x_2 - B]^+$ corresponds to the number of unserved packets at the second previous switch, which are passed to the first previous switch. Since every switch has N drop ports, the first previous switch can serve $\min(x_1 + [x_2 - B]^+, N)$ contended packets. Then, the current switch will receive $[\min(x_1 + [x_2 - B]^+, N) - B]^+$ unserved packets from the previous switches. This formulation can be extended for $\xi_I^{(0,Z)}$:

$$\begin{aligned}\xi_I^{(0,Z)} = [x_0 + [\min(x_1 + [\min(x_2 + \cdots [\min(x_{z-2} \\ + [\min(x_{Z-1} + [x_Z - B]^+, N) - B]^+ \\ , N) - B]^+ \cdots, N) - B]^+, N) - B]^+ - N]^+.\end{aligned} \quad (8)$$

The number of switch modules W is multiplied in order to consider all switch modules as

$$\lambda_I = W \lambda_I^{(W-1)}. \quad (9)$$

The average number of packet loss in II is written as

$$\lambda_{II} = \sum_{x_0=0}^{N-1} P(X = x_0) \cdots \sum_{x_{W-1}=0}^{N-1} P(X = x_{W-1}) \left[\sum_{n=0}^{W-1} x_n - WB \right]^+. \quad (10)$$

Since the packet loss events in I affect those of II, one needs to calculate the relative complementary packet loss events $\lambda_{II \setminus I}$ as

$$\begin{aligned}\lambda_{II \setminus I} = \sum_{x_0=0}^{N-1} P(X = x_0) \cdots \sum_{x_{W-1}=0}^{N-1} P(X = x_{W-1}) \\ \times \left[\sum_{n=0}^{W-1} x_n - \sum_{n=0}^{W-1} \xi_I^{(n,W-1)} - WB \right]^+.\end{aligned} \quad (11)$$

Consequently, the overall PLR of partially shared W switch modules with B buffers per $N \times N$ switches can be calculated as

$$PLR_{N,B,W} = \frac{\lambda_I + \lambda_{II \setminus I}}{WN\rho}. \quad (12)$$

III. Performance Evaluation

Figure 2 shows the PLR performance data of W number of 8×8 switch cards in the partially buffer sharing configuration of our proposal. As expected, the intra-module buffer sharing method reduces the PLR efficiently. Moreover, it is remarkable that inter-module buffer sharing reduces the PLR dramatically.

It shows the 'many hands make light work' effect in that switch modules help to resolve contention troubles of neighbors.

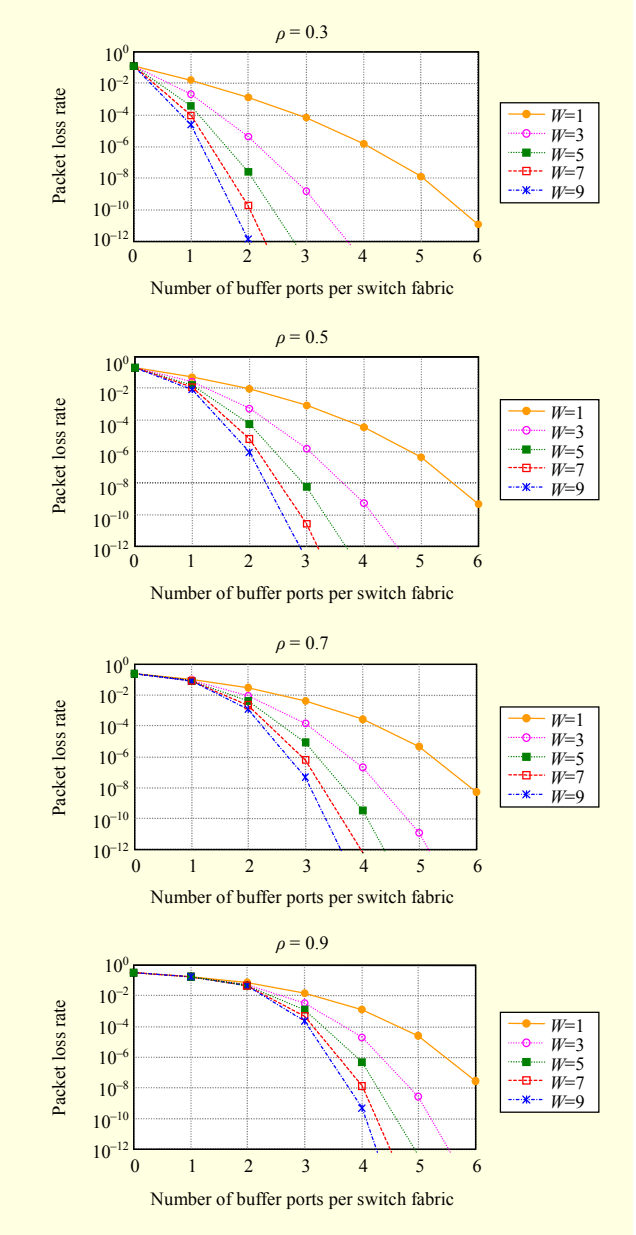

Fig. 2. Packet loss rate graphs for the number of shared switch (W) and buffers (B) at 8×8 switch fabric with various offered loads (ρ).

Table 1. Required buffer sharing ratio (S) to guarantee the 10^{-6} packet loss rate for W shared 8×8 switch fabrics.

Number of shared switches (W)	Offered load (ρ)				
	0.1	0.3	0.5	0.7	0.9
1	35%	51%	60%	65%	68%
2	22%	36%	48%	56%	62%
4	12%	24%	34%	44%	52%
8	7%	17%	26%	36%	45%

The PLRs of the bufferless cases in Fig. 2 are calculated by both microscopic and macroscopic approaches. Exactly the same results of the PLR of bufferless cases are yielded to show the validity of suggested microscopic approaches for PLR expression of (12). Table 1 summaries buffer sharing ratios for a PLR of 10^{-6}. In the implementation, $\lceil S \times N \rceil$ number of buffer ports can be installed per switch module. For example, when an offered load ρ and a number of switch modules W are 0.5 and 4, respectively, three buffer ports per switch module can achieve the PLR of 10^{-6} or less. The required buffer sharing ratio table manifests the possibility of substantial power and cost savings with modular upgrades.

IV. Conclusion

In this letter, we introduced a modular switch upgrade system solution that utilizes two-fold sharing by intra-module and inter-module buffer sharing to handle traffic increase and achieve PLR reduction. We presented a numerical analysis of a PLR model based on a microscopic understanding of packet loss statistics to verify the superb packet loss performance of our two-fold sharing technique. We presented buffer sharing ratio metrics for acceptable PLR. In particular, the proposed model is expected to bring great gains when large, expensive, and power-consuming buffer modules or buffer interfaces are used, as is found, for example, in photonic switch fabrics with optical interfaces.

References

[1] J. Kim et al., "Design of Novel Passive Optical Switching System Using Shared Wavelength Conversion with Electrical Buffer," *IEICE Electron. Express*, vol. 3, no. 24, Dec. 2006, pp. 546-551.

[2] X. Gu et al., *Control and Performance in Packet, Circuit, and ATM Networks*, Boston: Kluwer Academic, 1995.

[3] M. Hluchyj and M. Karol, "Queueing in High Performance Packet Switching," *IEEE JSAC*, vol. 6, no. 9, Dec. 1988, pp.1587-1597.

[4] J.S. Milton and J.C. Arnold, *Introduction to Probability and Statistics: Principles and Applications for Engineering and the Computing Science,* New York: McGraw-Hill, 1995.