Two Priority Buffered Multistage Interconnection Networks

Galia Shabtai[1], Israel Cidon and Moshe Sidi
Electrical Engineering Department
Technion, Haifa 32000, Israel

***Abstract*— This paper presents a novel architecture of internally two priority buffered Multistage Interconnection Network (MIN). First, we compare by simulation the new architecture against a single priority MIN and demonstrate up to N times higher high-priority throughput in a hot spot situation, when N is the number of inputs. In addition, under uniform traffic assumption we show an increase in the low priority throughput, without any change in the high priority throughput. Moreover, while in the single priority system the high priority delay and its standard deviation are increased when low priority traffic is present, it is kept constant in the dual priority system. Finally, we introduce a new approach of long Markovian memory performance model to better capture the packets dependency in a single priority MIN under uniform traffic and extend this model for a dual priority MIN. Model results are shown to be very accurate.**

I. INTRODUCTION

A priority service scheme in a multistage interconnection networks can be defined in terms of a policy determining: (a) which of the arriving packets are admitted to the buffer(s); and/or (b) which of the admitted packets is served next. The former priority service schemes are typically referred to as space priority (or discarding) schemes and attempt to minimize the packet loss of loss-sensitive traffic, such as data. An overview and classification of some space priority strategies can be found in [1,2]. The latter priority service schemes are typically referred to as time priority (or priority scheduling) schemes and attempt to guarantee acceptable delay boundaries to delay-sensitive traffic, such as voice and video. Several types of time priority schemes, such as Weighted-Round-Robin and Weighted-Fair-Queueing, have been proposed and analyzed, each with their own specific algorithmic and computational complexity, see for example [1,3] and the references therein.

There are already several commercial switches which accommodate traffic priority schemes, see for example [4,5]. These switches consist of an internally single priority switch fabric and employ two priority queues for each input port. Packets are queued based on their priority level and packets with higher priority are allowed to pass first.

The internal switch structure used in previous studies [6,7] is a single priority fabric with controlled inputs. In contrast to these previous works, our paper considers for the first time an internal two priority switch fabric architecture and focus on the effect of a two priority input buffered Multistage Interconnection Network (MIN) on the performance of high and low priority traffic. We also suggest a new Markovian model for analyzing the performance of the two priority traffic types, assuming uniform traffic, and present numerical results.

II. SINGLE VS. DUAL PRIORITY MIN

In this section we introduce our novel architecture of internally two priority MIN and compare its performance to a single priority MIN. Our work concentrates on an (N×N) delta-2 network, i.e. n stages of N/2 (2×2) crossbar switches, where $N=2^n$, as illustrated in Fig. 1.

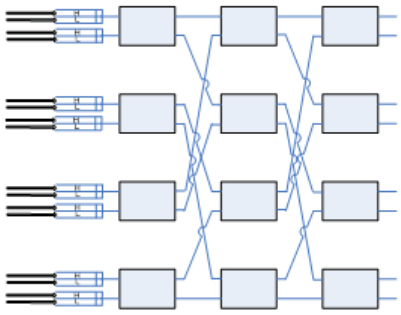

Fig. 1 An 8×8 system: delta-2 network with input FIFOs.

A. Single Priority MIN

Fig. 2 shows the basic model of a 2×2 single buffered single priority switching element, which mainly consists of two input and two output ports, a single buffer for each incoming link and a non blocking switching matrix to connect the input buffers to the output ports. We assume that a maximum of one packet can be sent from each output port during one clock cycle and therefore a maximum of one packet can be received at each SE input link.

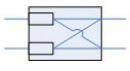

Fig. 2 Basic model of a 2×2 Single Buffered Single Priority Switching Element.

[1] Galia Shabtai is also with Cisco Systems Inc.

The assumptions of a network under a synchronous uniform traffic model with global flow control mechanism are described in [8-12].

B. Dual Priority MIN

In the previous section, we assumed that all packets are treated identically, i.e., there is no traffic classification. In this section we extend the model for two traffic classifications: high priority traffic and low priority traffic.

The basic model of a 2×2 single buffered dual priority switching element in shown in Fig. 3. The main difference from the single priority SE is that each input buffer is composed of two single *queues*: one for high priority packets and one for low priority packets. The assumption of sending a maximum of one packet from each SE output port during one clock cycle is still valid and therefore each SE input link can still receive a maximum of one packet during each clock cycle. On the other hand, we do allow an input buffer to send up to two packets, one high priority and one low priority, during a clock cycle, if each packet is sent to a different output port and the other buffer of the same SE does not send any packet during this particular clock cycle.

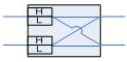

Fig. 3 Basic model of a 2×2 Single Buffered Dual Priority Switching Element.

Following are the assumptions for the dual priority model.

1. The network clock cycle consists of two phases. In the first phase, flow control information passes through the network from the last stage to the first stage. In the second phase, packets flow from one stage to the next in accordance with the flow control information.
2. A switch input is able to accept a high priority packet if it has an empty high priority queue or if the high priority packet in its high priority queue will leave during the second phase of the current clock cycle.
3. A switch input is able to accept a low priority packet if it has an empty low priority queue or if the low priority packet in its low priority queue will leave during the second phase of the current clock cycle.
4. There is no blocking at the output links of the network.
5. The arrival process of each input of the network is a simple Bernoulli process, i.e., the probability that a packet arrives within a clock cycle is constant and the arrivals are independent of each other. Moreover, there is a fixed probability for each packet to be either high or low priority.
6. The routing logic within each priority at each SE is fair, i.e., same priority conflicts are randomly resolved.
7. High priority packets have a fixed priority over the low priority packets.
8. Packets are of fixed size.

If a uniform traffic model is considered, then the following assumption is added:

9. Each input link is offered the same traffic load and the same high to low priority ratio. In addition, the destination addresses of the packets are distributed uniformly over all output links of the network.

Since the high priority packets have strict priority over the low priority packets, and since we still allow a maximum of one packet into each SE input link and out of each SE output link, the performance (both throughput and delay) of the high priority traffic in the dual priority MIN is identical to the performance of the single priority traffic in the single priority MIN. Moreover, the low priority traffic is getting served only in those clock cycles in which no high priority traffic is able to move to the desired destination. Therefore, the overall throughput of the dual priority MIN under specific total input load (low priority + high priority) should be at least as high as the single priority MIN throughput under the same total input load and can be even higher.

C. System Description

As in most of contemporary commercial switches, see for example [4,5], we added two input buffers (FIFOs) in front of each MIN input: one is designated for the low priority packets and the other for the high priority packets. Each low priority packet that arrives to a system input is enqueued to the low priority input FIFO, and each high priority packet that arrives is enqueued to the high priority input FIFO.

An N×N single priority system comprises of N high priority input FIFOs and N low priority input FIFOs which are connected to an N×N single priority MIN's inputs, as illustrated in Fig. 1. A high priority packet leaves the high priority input FIFO and enters the MIN input if the corresponding SE input is able to accept a packet. On the other hand, a low priority packet can enter the MIN, only if the high priority input FIFO is empty and the corresponding SE input is able to accept a packet. This strict priority admission of high priority packets over low priority packets suggests that the throughput of the high priority traffic is not affected by the presence of low priority traffic. However, the total delay of the high priority traffic in the single priority system is affected by the presence of low priority traffic, since it increases the congestion probability inside the MIN, and hence increases the delay and its standard deviation.

The dual priority system is obtained by replacing the single priority MIN in the single priority system with a dual priority MIN. In this system, a high priority packet leaves the high priority input FIFO and enters the MIN input if the corresponding SE input is able to accept a high priority packet. On the other hand, a low priority packet can enter the MIN, only if there is no high priority packet that can enter and the corresponding SE input is able to accept a low priority packet. As in the single priority system, the throughput of the high priority traffic is not affected by the presence of low priority traffic. However, unlike the single priority system, the delay of the high priority traffic in the dual priority system is also not affected by the low priority traffic.

D. Simulations Results

In order to isolate the input FIFOs size from the system performance, we used infinite input FIFOs in front of each MIN input, so there was actually no packet loss. Nevertheless, it is obvious that a system with low throughput and finite input FIFOs will suffer from higher packet loss than a system that can reach higher throughput with the same input FIFOs size.

Therefore, we concentrated on both the delay and the throughput measurements in our simulations.

To emphasize the "immunity" of the high priority traffic over the low priority traffic in the dual priority system vs. the single priority system, we considered an extreme case in which: (a) all inputs send traffic to output link 0, which describes an extreme hot spot situation; (b) all inputs send the same input load; (c) all inputs, except input 0, send low priority traffic, while input 0 sends high priority traffic. While this scenario does not represent a realistic long term steady state, it demonstrates a transient load situation that should be taken into account in the design of contemporary systems. The high priority throughput in both systems is depicted in Fig. 4 for 6 stages networks, with 64 inputs and outputs.

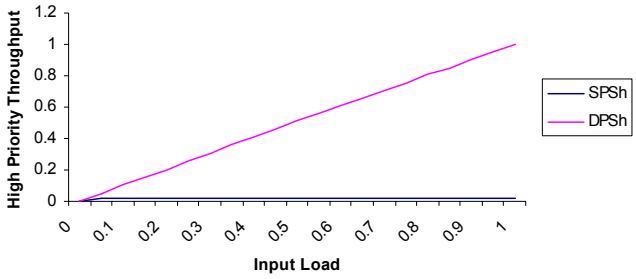

Fig. 4 High priority throughput in both single and dual priority systems with 6 stages under hot spot traffic. SPSh represents the high priority throughput in the single priority system, while DPSh represents the high priority throughput in the dual priority system

In general, all packets are destined to output 0, which yields throughput of 1 for all inputs together. In the single priority system all packets are treated equally and therefore each input, including input 0 which sends high priority traffic, is able to send throughput of 1/64=0.015. However, in the dual priority system high priority traffic has strict priority over low priority traffic and therefore the high priority throughput equals the high priority input load, while the low priority throughput equals 1-high priority throughput.

The results in the rest of this section consider a uniform traffic model, as describes in sections II.A and II.B.

The total throughput of both systems under full input load is illustrated in Fig. 5.

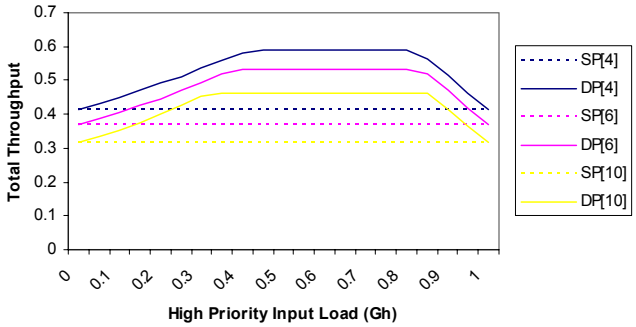

Fig. 5 Maximum total throughput of both single and dual priority systems under dual priority traffic (G = Gh + Gl = 1) as a function of the high priority input load (Gh) for various MIN sizes. SP[k] represents a single priority system with k stages MIN. Similarly, DP[k] represents a dual priority system with k stages MIN.

As implied earlier, we can see that the maximum throughput of the dual priority system is higher than that of the single priority system when more than one priority traffic enters the system (up to 47% increase in the 1024×1024 system). The source of this extra throughput in the dual priority system is the advance of low priority packets when high priority packets cannot move forward, i.e. this is exactly the low priority throughput difference between the two systems.

As discussed in the previous sub-section, the high priority traffic throughput is identical in both the single and the dual priority systems under the same dual priority input load. Moreover, unlike the high priority throughput the high priority total delay and its standard deviation are affected by the low priority input load in the single priority system.

Fig. 6 depicts the average high priority total delay in 64×64 single and dual priority systems under dual priority traffic. We can see that in the single priority system the average delay increases with the increase of the low priority input load, but the increase stops when the total input load reaches the maximum throughput of that system. At this point, the low priority load inside the MIN stops increasing and therefore the high priority delay stays constant. As high priority input load increases, the probability that a high priority packet arrives to an empty input FIFO decreases and therefore, the total delay increases. The high priority delay in the dual priority system is not affected by the low priority input load and remains constant.

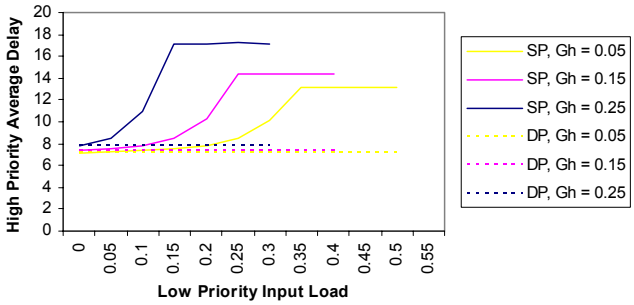

Fig. 6 Average high priority total delay in 64×64 single priority (SP) and dual priority (DP) systems under dual priority traffic. Gh represents the high priority input load.

The high priority maximum delay graph has a similar shape with a more extreme effect: a difference of up to 55 time units.

The standard deviation of the high priority delay is illustrated in Fig. 7.

III. Performance Model

The performance model is focused on uniform traffic. Due to space limitations, the performance model is only briefly sketched below. The full model can be found in [13]. Previous work for modeling and analyzing MINs under uniform traffic [8-12], [14-16] used short Markovian memory (the last clock cycle). We propose to extend the Markovian memory to the last two consecutive clock cycles to better capture the packets dependency.

A. *Single Priority Model*

The basic model of the single priority SE is presented in section II.A "Single Priority MIN". Following Jenq [8] and later works (such as [12] and [15]), we assume that under the synchronous uniform traffic model the state of an SE at stage k is statistically indistinguishable from that of another SE of the same stage. Moreover, the two buffers in the same SE are statistically independent and therefore the state of a stage can be reduced to that of a single buffer.

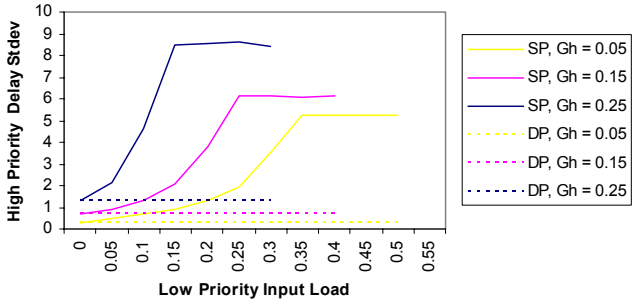

Fig. 7 Standard deviation of the high priority delay in 64×64 single priority (SP) and dual priority (DP) systems under dual priority traffic. Gh represents the high priority input load.

Following are the five possible states of a buffer in the Single Priority Model:

- "00": buffer was empty at the beginning of the previous clock cycle and is empty at the beginning of the current clock cycle as well, i.e., no new packet has been received during the previous clock cycle.
- "01": buffer was empty at the beginning of the previous clock cycle and contains a new packet at the beginning of the current clock cycle, i.e., a new packet has been received during the previous clock cycle.
- "10": buffer had a packet at the beginning of the previous clock cycle but has no packet at the beginning of the current one, i.e., a packet has been sent from this buffer during the previous clock cycle, but no new packet has been received.
- "11n": buffer had a packet at the beginning of the previous clock cycle and has a new one at the beginning of the current clock cycle, i.e., a packet has been sent from this buffer during the previous clock cycle, and a new packet has been received.
- "11b": buffer had a packet at the beginning of the previous clock cycle and has a blocked one at the beginning of the current clock cycle, i.e., no packet has been sent from this buffer during the previous clock cycle.

The state transition diagram is shown in Fig. 8.

Fig. 9 shows the throughput of a single buffered single priority delta-2 network for various network sizes. It can be seen that the model's accuracy decreases as network size increases. This is due to the fact that every additional stage introduces further collisions and the inaccuracy of one stage accumulates to the previous stage. Nevertheless, our model seems to be very accurate: the maximum deviation is only 10.9% for a fully loaded 1024×1024 network.

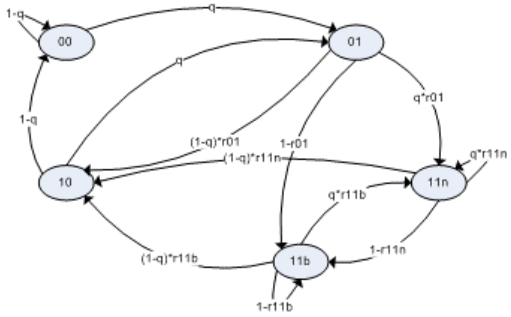

Fig. 8 The state transition diagram of a single priority SE buffer.

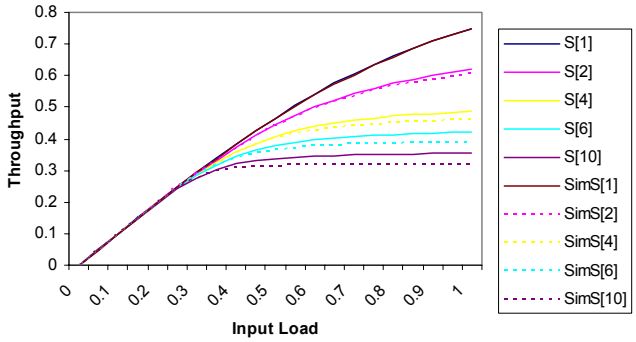

Fig. 9.Throughput of a single buffered single priority delta-2 network for various network sizes. S[k] is the analyzed throughput for a network with k stages. SimS[k] is the simulated throughput for a network with k stages.

B. *Dual Priority Model*

The basic model of the dual priority SE and its assumptions are presented in section II.B "Dual Priority MIN".

Since high priority packets have strict priority over the low priority packets and since the low priority traffic is getting served only in those clock cycles in which no high priority traffic is able to move to the desired destination, our model includes two separate Markov chains. The first one is a stand alone chain, which represents the high priority traffic queue and is identical to the single priority model, presented in the previous section. On the other hand, since the service of the low priority traffic depends on the high priority service, the transitions of the second chain, which represents the low priority traffic queue, depends on the transitions of the first chain. Moreover, the low priority model is an extended version of the single priority model and includes six states. The states "00", "01", "10" and "11n" are identical to the states of the single priority model, while state "11b" is split into two states as follows.

- "11hb": queue had a low priority packet at the beginning of the previous clock cycle and this packet has been blocked by a high priority packet and stayed at least till the beginning of the current clock cycle.

- "11lb": queue had a low priority packet at the beginning of the previous clock cycle and this packet has been blocked by a low priority packet and stayed at least till the beginning of the current clock cycle.

The state transition diagram for the low priority traffic queue is shown in Fig. 10.

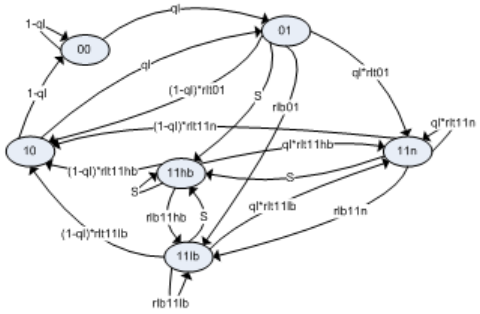

Fig. 10 The state transition diagram of a low priority queue in an SE(k) buffer.

Fig. 11 shows the normalized low priority throughput of a single buffered dual priority Delta-2 network for various network sizes as a function of the high priority input load. The offered load is 1, therefore the low priority input load equals to 1-high priority input load. The high priority throughput is identical to the one shown in Fig. 9. There seems to be no specific direction to the model: sometimes optimistic and sometimes pessimistic. Nevertheless, the maximum deviation of our model is only 16.9% for fully loaded delta-2 network with 10 stages, i.e. a 1024×1024 delta-2 network.

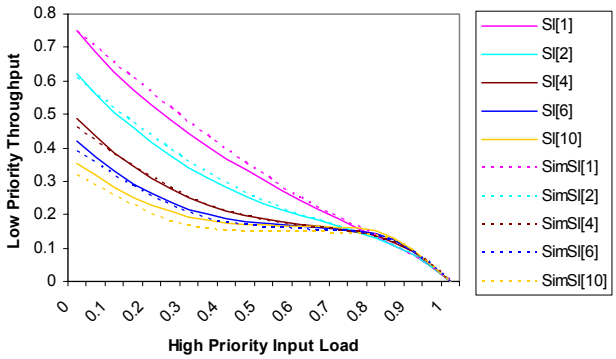

Fig. 11 Low priority throughput of a single buffered dual priority Delta-2 network for various network sizes as a function of the high priority input load. Sl[k] is the analyzed low priority throughput for a network with k stages. SimSl[k] is the simulated low priority throughput for a network with k stages. Low priority input load equals to 1-high priority input load.

IV. Discussion

This paper presents a novel internally two priority buffered MIN architecture. It compares its performance with a single priority MIN. Simulation results show increase in high priority throughput of up to N times under hot spot traffic. For uniform traffic, we show an increase in low priority throughput, without any change in the high priority throughput. Moreover, while high priority delay and its standard deviation are increased when low priority traffic present in the single priority system, it is kept constant in the dual priority system. Finally, we introduce a new approach of long Markovian memory performance model to better capture the packets dependency in a single priority MIN under uniform traffic and extend this model for a dual priority MIN. Model results seems to be very accurate. Non-homogenous traffic study via simulation and analysis is yet to be studied.

Acknowledgments

The authors would like to thank Michael Laor from Cisco Systems for fruitful discussions on switch fabric architectures.

References

[1] K. Liu, D. W. Petr and V. S. Frost, "Design and analysis of a bandwidth management framework for ATM-based broadband ISDN", *IEEE Communications Magazine*, vol. 35, No. 5, pp. 138-145, May 1997.

[2] I. Cidon, R. Guerin and A. Khamisky, "On protective buffer policies", *IEEE/ACM Transactions on Networking*, vol. 2, No. 3, pp. 240-246, June 1994.

[3] S. P. Morgan, "Queueing disciplines and passive congestion control in byte-stream networks", *IEEE Trans. Comnun.*, vol. 39(7), pp. 1097-1106, July 1991.

[4] D-Link DES-3250TG 10/100Mbps managed switch, http://www.dlink.co.uk/DES-3250TG.htm

[5] Intel® Express 460T standalone switch, http://www.intel.com/support/express/switches/460/30281.htm

[6] J. S. C. Chen and R. Guerin, "Performance study of an input queueing packet switch with two priority classes", *IEEE Trans. Comnun.*, vol. 39, No. 1, pp. 117-126, Jan. 1991.

[7] S. L. Ng and B. Dewar, "Load sharing replicated buffered banyan networks with priority traffic", unpublished.

[8] Y-C. Jenq, "Performance analysis of a packet switch based on single-buffered banyan network", *IEEE Journal Selected Areas of Commun.*, vol SAC-1, No. 6, pp. 1014-1021, Dec. 1983.

[9] T. Szymanski and S. Shaikh, "Markov chain analysis of packet-switched banyans with arbitrary switch sizes, queue sizes, link multiplicities and speedupds", in *Proc. INFOCOM 89*, Apr. 1989.

[10] H. Yoon, K. Y. Lee, and M. T. Lui, "Performance analysis of multibuffered packet-switching networks in multiprocessor systems", *IEEE Trans. Comput.*, vol. 39, pp. 319-327, Mar. 1990.

[11] J. S. Turner, "Queueing analysis of buffered switching networks", *IEEE Trans. Commun.*, vol. 41, no. 2, pp. 412-420, Feb. 1993.

[12] H. Mun and H. Y. Youn, "Performance analysis of finite buffered multistage interconnection networks", *IEEE Trans. Comput.*, vol. 43, no. 2, pp. 153-161, Feb. 1994.

[13] G. Shabtai, I. Cidon and M. Sidi, "Two priority buffered multistage interconnection networks", CCIT Report. 2003.

[14] S. H. Hsiao and C. Y. R. Chen, "Performance analysis of single-buffered multistage interconnection networks", *IEEE Trans. Commun.*, vol. 42, no. 9, pp. 2722-2729, Sep. 1994.

[15] T. H. Theimer, E. P. Rathgeb and M. N. Huber, "Performance analysis of buffered banyan networks", *IEEE Trans. Commun.*, vol. 39, no. 2, pp. 269-277, Feb. 1991.

[16] K. S. Chan, K. L. Yeung and S. C. H. Chan, "A refined model for performance analysis of buffered banyan networks with and without priority control", *IEICE Trans. Commun.*, vol. E82-B, no. 1, pp. 48-59, Jan. 1999..