The role of priority mechanisms on performance metrics of double-buffered Switching Elements

D. C. Vasiliadis, G. E. Rizos and C. Vassilakis

Department of Computer Science and Technology
Faculty of Sciences and Technology
University of Peloponnese
GR-221 00 Tripolis Greece
dvas@uop.gr , georizos@uop.gr , costas@uop.gr

Abstract. The main concerns in designing the multistage switching fabrics are speed, throughput, delay and variance of delay for a given bandwidth. The rationale behind using various priority mechanisms is either to offer different quality of service levels to packets or to optimize performance parameters of the network, e.g. minimize internal blocking in the Switching Elements (SEs). We investigated the performance parameters of an enhanced priority (EP) mechanism versus a single priority (SP) one. In the EP scheme, packet priority was computed dynamically and was directly proportional to the transmission queue length of the SE that the packet is currently stored in. Finally, we extended the idea of the priority scheme by proposing a multi-priority (MP) mechanism. In the MP scheme, each SE has two transmission queues per link, with one queue dedicated to high priority packets and the other dedicated to low priority ones. We simulated a multistage network under the uniform traffic condition and concluded that the proposed double-buffered SEs provide higher throughput, and decreased latency.

Keywords: Multistage Interconnection Networks, Banyan Switches, Delta Networks, Packet Switching, Multi-Priority Networks, Performance Analysis
PACS: 07.05.Tp, 42.79.Ta

INTRODUCTION

Multistage Interconnection Networks (MINs) with crossbar Switching Elements (SEs) are frequently proposed for interconnecting processors and memory modules in parallel multiprocessor systems [1] [2], [3]. MINs have been recently identified as an efficient interconnection network for communication structures such as gigabit Ethernet switches, terabit routers, and ATM switches [4], [5], [6]. Significant advantages of MINs include their low cost/performance ratio and their ability to route multiple communication tasks concurrently. MINs with the Banyan [7] property are proposed to connect a large number of processors to establish a multiprocessor system; they have also received considerable interest in the development of packet-switched networks. Non-Banyan MINs, are in general, more expensive than Banyan networks and more complex to control.

During the last decades, much research has been performed in investigating the performance of parallel and distributed systems, particularly in the area of networks and communications. In order to evaluate their performance different methods have been used. These methods mainly include Markov chains [8], queuing theory [9], Petri nets [10], [11] and simulation [12], [13].

In the industry domain, Cisco has built its new CRS-1 router [14], [15] as a multistage switching fabric. The switching fabric that provides the communications path between line cards is a 3-stage, self-routed architecture.

Packet priority is a common issue in networks and can be used at application level and/or within the system level. Applications may specify different priority classes for packets to designate to the network that some packets need to be offered better quality of service than others. The system, on the other hand, may exploit the priority mechanism to improve system performance, adapting to the current traffic conditions. Thus, in this paper we analyze the role of various priority mechanisms in order to achieve satisfactory levels for the two most important network performance factors, namely packet throughput and the mean time a packet needs to traverse the MIN, improving the QoS offered to high-priority packets.

The remainder of this paper is organized as follows: in section 2 we present the model for single, enhanced, and multi-priority schemas. Subsequently, in section 3 we present the results of our performance analysis, which has been conducted through simulation experiments, while section 4 provides the concluding remarks.

THE MODEL

A MIN can be defined as a network used to interconnect a group of N inputs to a group of M outputs using several stages of small size Switching Elements (SEs) followed (or leaded) by link states. It is usually defined by, among others, its topology, routing algorithm, switching strategy and flow control mechanism. An (N X N) MIN with the Banyan property [7] can be constructed by $n=\log_c N$ stages of (cxc) SEs, where c is the degree of the SEs. At each stage there are exactly N/c SEs, consequently, the total number of SEs of a MIN is $(N/c)*\log_c N$. Thus, there are O(N*log$N$) interconnections among all stages, as opposed to the crossbar network which requires O(N^2) links. These MINs are characterized by the fact that there is exactly one unique path from each source (input) to each sink (output) providing efficient multistage self-routing switching fabrics.

Our performance analysis on a MIN is exemplified through its application on a typical (NXN) Delta Network, providing both benefits of Omega [16] and Generalized Cube Networks [17], i.e. destination routing, partitioning and expandability. A typical configuration of an (NXN) Delta Network, one of the most widely used classes of Banyan MINs, which were proposed by Patel [18], is shown at Fig. 1, and 2.

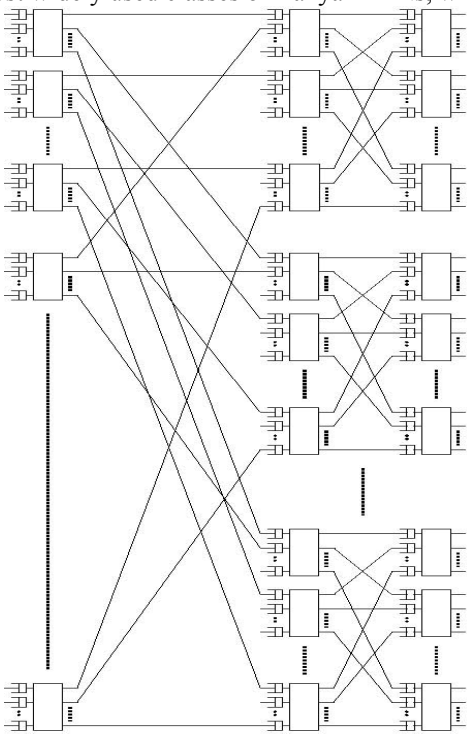

FIGURE 1. An N X N Delta Network employing a single/enhanced-priority scheme

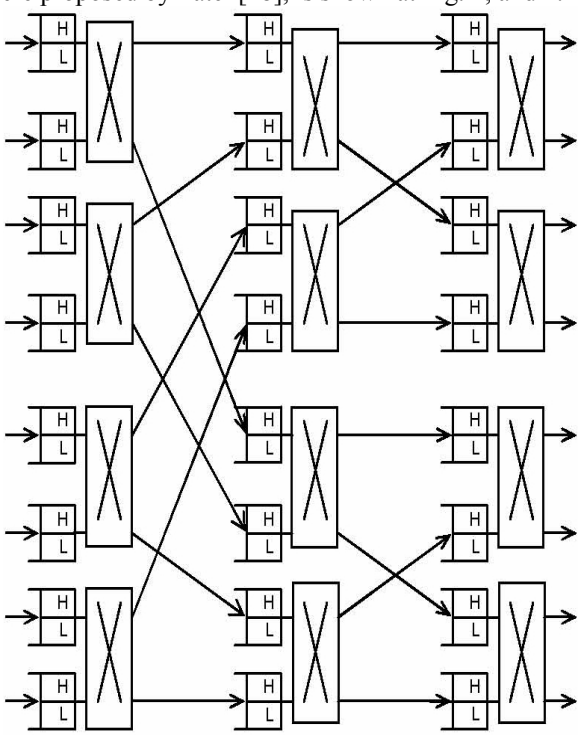

FIGURE 2. An 8 X 8 Delta Network employing a multi-priority scheme

In our paper, we consider a MIN that operates under the following assumptions:

- 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 one. In the second phase, packets flow from one stage to the next in accordance with the flow control information.
- 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. We will denote this probability as λ.
- A packet arriving at the first stage is discarded if the buffer of the corresponding SE is full.
- All SEs have deterministic service time.
- A packet is blocked at a stage if the destination buffer at the next stage is full.

- The packets are uniformly distributed across all the destinations and each queue uses a FIFO policy for all output ports.
- Conflict resolution under the Single Priority Mechanism operates under the following scheme: when two packets at a stage contend for the same buffer at the next stage and there is not adequate free space for both of them to be stored (i.e. only one buffer position is available at the next stage), there is a conflict. One of them will be accepted at random, and the other will be blocked by means of upstream control signals.
- When the Enhanced Priority Mechanism is employed, the system dynamically assigns priority to packets, favoring those stored in lengthier transmission queues. Conflict resolution under the Enhanced Priority Mechanism is performed as follows: when a conflict occurs it is resolved by examining the number of packets within the transmission queue of the SEs from which the contending packets originate. For such a decision, however, to be taken, the receiving SE needs to have available the queue lengths of the transmitting SEs, a piece of information which is not available to the receiving SE in typical MINs. To make this information available, SEs operating under the Enhanced Priority MIN scheme send the length of their transmission packet queue at the start of the packet header, as a preamble. When receiving SEs detect a conflict situation (i.e. two incoming transmissions and only one free buffer slot), they compare the queue sizes of the transmitting SEs and proceed in receiving the packet preambled with the largest value for the queue size. The other packet will be blocked, and the transmitting SE will be notified by means of an upstream control signal during the next network cycle. Since buffer sizes in SEs are usually in the range 1 to 8, the length of the preamble can vary from 1 to 3 bits which is quite small compared to the packet length The preamble need not be checksumed (which would increase its size), since any error in these bits would simply lead to accepting the wrong (with respect to the priority policy) packet, a case that would only marginally affect the gains obtained by the introduction of the enhanced priority scheme.
- When the Multi Priority Mechanism is used, when applications enter a packet to the network they specify its priority, designating it either as *high* or *low*. The criteria for priority selection may stem from the nature of packet data (e.g. packets containing streaming media data can be designated as high-priority while FTP data can be characterized as low-priority) or from protocol intrinsics (e.g. TCP out-of-band/expedited data vs. normal connection data).
- Conflict resolution under the Multi Priority Mechanism operates under the following scheme: when two high- or low-priority packets at a stage contend for the same buffer at the next stage and there is not adequate free space for both of them to be stored, there is a conflict. One of them will be accepted at random, and the other will be blocked by means of upstream control signals, with high priority packets having precedence over low priority packets at the transmission process. The priority of each packet is indicated through a priority bit in the packet header.
- Finally, all packets in input ports contain both the data to be transferred and the routing tag. In order to achieve synchronously operating SEs, the MIN is internally clocked. As soon as packets reach a destination port they are removed from the MIN, so, packets cannot be blocked at the last stage.

PERFORMANCE ANALYSIS

The performance metrics of a double-buffered MIN was evaluated through extensive simulations by a generic simulator which was developed for packet-oriented communication environments. The simulator can handle several switch types, inter-stage interconnection patterns, load conditions, switch operation policies, and priorities. The simulation was performed at packet level, assuming fixed-length packets transmitted in equal-length time slots, where the slot was defined to be equal to the time required to forward a packet from one stage to the next. The parameters for the packet traffic model were varied across simulation experiments to generate different offered loads and traffic patterns. Metrics such as packet throughput and packet delays were collected at the output ports. We performed extensive simulations to validate our results. All statistics obtained from simulation running for 10^5 clock cycles. The number of simulation runs was adjusted to ensure a steady-state operating condition for the MIN. There was a stabilization process in order the network be allowed to reach a steady state by discarding the first 10^3 network cycles, before collecting the statistics. In the following paragraphs, we present performance studies based on selected simulation results.

Before collecting performance parameters for the proposed priority schemes, we validated the accuracy of our simulator by using it to compute the *normalized throughput* (see the next paragraph for a definition) for a standard single-buffered 6-stage MIN and comparing the results obtained against the results already published for three classical models [19] [20] [21]. All models are very accurate at low loads. The accuracy reduces as input load increases. Especially, when input load approaches the network

maximum throughput, the accuracy of Jenq's model is insufficient. One of the reasons is the fact that many packets are blocked mainly at the network first stages at high traffic rates. Thus, Mun introduced a "blocked" state to his model to improve accuracy. The consideration of the dependencies between the two buffers of an SE in Theimer's model leads to further improvement. Our simulation provides accurate statistics which comparing with the results of Theimer's model were found to be in close agreement (differences are less than 1%).

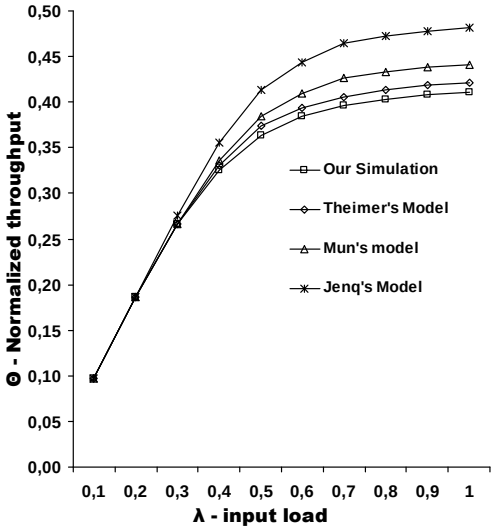

FIGURE 3. Normalized throughput of a single-buffered 6-stage MIN

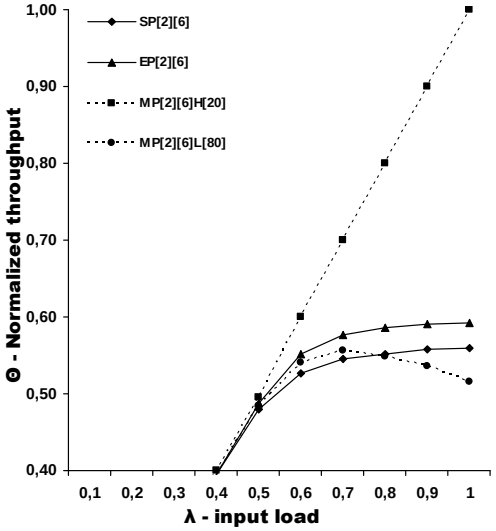

FIGURE 4. Normalized throughput of a double-buffered 6-stage MIN

Normalized throughput Θ is the ratio of the *average throughput* Θ_{avg} (or *bandwidth*) to network size *N,* where *average throughput* is the average number of packets accepted by all destinations per network cycle. Figures 4-6 present *the normalized throughput* of a double-buffered *k*-stage MIN versus *offered load* under three priority mechanisms. In the diagrams, curve SP[*b*][*k*] depicts the *normalized throughput* of a *k*-stage MIN with queues of buffer-length *b*, employing a single priority scheme. Similarly, curve EP[*b*][*k*] shows the corresponding *normalized throughput* of a *k*-stage MIN with queues of buffer-length *b*, employing an enhanced priority scheme. Finally, curves MP[*b*][*k*]H[*x*], and MP[*b*][*k*]L[*y*] depict the *relative normalized throughputs* of high and low priority packets respectively using a multi priority scheme. The *relative normalized throughputs* of high and low priority packets are the ratios of *normalized throughput* to the corresponding *input load* of high and low priority packets respectively. In this configuration scheme the length of both high and low priority queues is *b*, the probability of high-priority packet appearance is *x*%, while the low-priority one is *y*%, holding that $x + y = 100$.

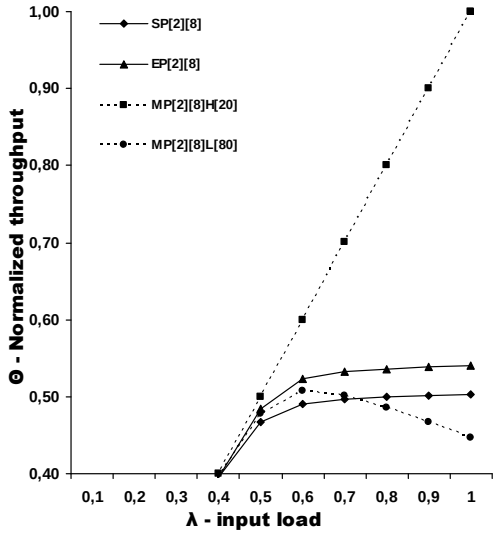

FIGURE 5. Normalized throughput of a double-buffered 8-stage MIN

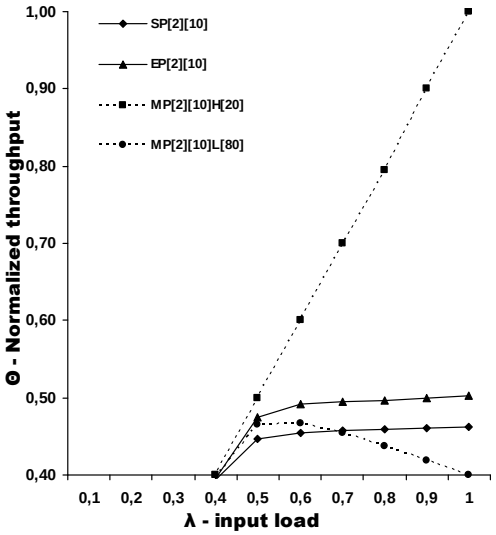

FIGURE 6. Normalized throughput of a double-buffered 10-stage MIN

Throughput performance (Figs. 4-6) depicts the benefits obtained form adopting the multi-priority mechanism. It is noteworthy that the *normalized throughput* for high-priority packets has an optimal value under all traffic loads (low, moderate, high and full load) at all configurations. Thus, all high priority packets that enter the MIN are forwarded to outputs without loss. The improvement of the *normalized throughput* for high-priority packets comes partially at the expense of normalized throughput for low-priority packets, which however is quantified to be tolerable to negligible for all network configuration setups. It is also clear that the adoption of the enhanced priority mechanism also improves considerably the normalized throughput.

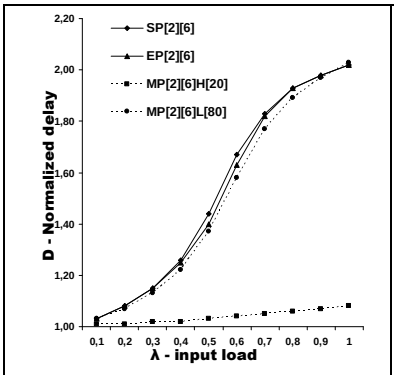

FIGURE 7. Normalized delay of a double-buffered 6-stage MIN

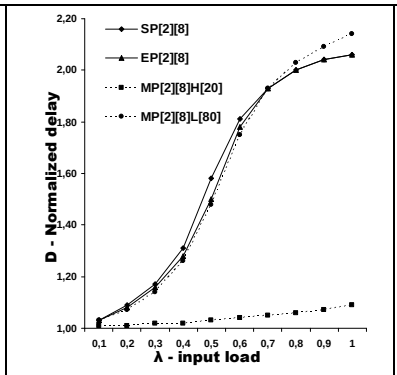

FIGURE 8. Normalized delay of a double-buffered 8-stage MIN

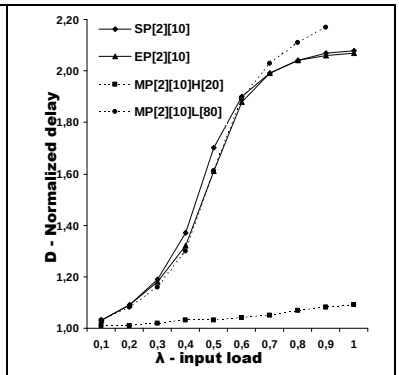

FIGURE 9. Normalized delay of a double-buffered 10-stage MIN

Figures 7-9 present the *normalized delay* of a double-buffered *k*-stage MIN versus *offered load* under three priority mechanisms. Besides the *throughput* gains achieved by employing the multi-priority scheme, it is worth noting that the *normalized delay* of high-priority packets falls dramatically, approaching the optimal value D_{min}=1. It is also seen that *normalized delay* of high-priority packets does not exceed the value of D<=1.09, under full load traffic (λ=1) at all network configurations. On the other hand, the *normalized delay* of low-priority packets deteriorates slightly. Finally, the performance metric of *normalized delay* was not affected considerably, when the enhanced priority scheme was incorporated.

CONCLUSIONS

In this paper we presented three different priority mechanisms and estimated their impact on performance in the context of a double-buffered MIN. Initially, we assessed the efficiency of the proposed priority schemes in terms of *throughput* and *delay* metrics. Regarding the enhanced priority mechanism, it was found that the gains for MINs in terms of *throughput* were considerable, while the respective *delay* was affected slightly in all network setups. Moreover, when a multi-priority scheme was employed for supporting high and low priorities on the input packets, it was found that both throughput and delay for high-priority packets approached their optimal values, whereas the respective deterioration for low-priority packets ranged from negligible to tolerable. According to Figures 4-9, the *throughput* and *delay* metrics for high-priority packets were found to be very close to the optimal values, when the high-priority packets were 20% of the overall network traffic. As depicted in all these figures, the multi-priority scheme excels in *bandwidth* and *latency*, while the enhanced priority mechanism also achieves satisfactory levels for *throughput*.

The proposed priority mechanisms can also be uniformly applied to several representative networks providing a basis for fair comparison and the necessary data for network designers to select optimal values for network operation parameters.

REFERENCES

1. Chang-hoon Choi, Sung-chun Kim. Hierarchical multistage interconnection network for shared-memory multiprocessor system. Proceedings of the 1997 ACM Symposium on Applied Computing, pp. 468-472.
2. Josep Torrellas, Zheng Zhang. The Performance of the Cedar Multistage Switching Network. IEEE Transactions on Parallel and Distributed Systems, 8(4), April 1997, pp. 321-336
3. Gheith A. Abandah and Edward S. Davidson. Modeling the communication performance of the IBM SP2. In Proceedings of the 10th International Parallel Processing Symposium (IPPS'96); Hawaii. IEEE Computer Society Press, 1996.

4. Elizabeth Suet Hing Tse. Switch fabric architecture analysis for a scalable bi-directionally reconfigurable IP router. Journal of Systems Architecture: the EUROMICRO Journal, 50(1), 2004, pp. 35-60.
5. Toshio Soumiya, Koji Nakamichi, Satoshi Kakuma, Takashi Hatano, and Akira Hakata. The large capacity ATM backbone switch ”FETEX-150 ESP”. Computer Networks, 31(6):603–615, 1999.
6. Ra’ed Y. Awdeh and H. T. Mouftah. Survey of ATM switch architectures. Computer Networks and ISDN Systems, 27:1567–1613, 1995.
7. G. F. Goke, G.J. Lipovski. Banyan Networks for Partitioning Multiprocessor Systems. Proc. 1st Annual Symposium on Computer Architecture, 1973, pp. 21-28
8. A. Merchart. A Markov chain approximation for analysis of Banyan networks. Proceedings of the ACM Sigmetrics Conference on Measurement and Modelling of Computer systems, 1991.
9. G. Bolch, S. Greiner, H. de Meer, K. S. Trivedi. Queueing Networks and Markov Chains - Modeling and Performance Evaluations with Computer Science Applications. John Wiley and Sons, New York, 1998
10. R. German. Performance Analysis of Communication Systems. John Wiley and Sons, 2000
11. P. J. Haas. Stohastic Petri Nets. Springer Verlag, 2002.
12. D.C. Vasiliadis, G.E. Rizos, C. Vassilakis. Performance Analysis of blocking Banyan Switches. Proceedings of the IEEE sponsored International Joint Conference on Telecommunications and Networking CISSE 06, December, 2006.
13. D. C. Vasiliadis, G. E. Rizos, and C. Vassilakis. Performance Analysis of dual priority single-buffered blocking Multistage Interconnection Networks. Proceedings of the third International Conference on Networking and Services (ICNS’07), IEEE Computer Society pres, June 2007.
14. Cisco Systems, http://www.cisco.com/application/pdf/en/us/guest/products/ps5763/c1031/cdccont_0900aecd800f8118.pdf.
15. Cisco Systems, http://newsroom.cisco.com/dlls/2004/next_generation_networks_and_the_cisco_carrier_routing_system_overview.pdf.
16. Siu-Cheung Chau, Tiehong Xiao, Ada Wai-Chee Fu. Routing and Scheduling for a Novel Optical Multistage Interconnection Network. Proceedings of the Euro-Par 2005 Parallel Processing Conference, LNCS Vol. 3648/2005, pp. 984-993
17. G. B. Adams and H. J. Siegel, The extra stage cube: A fault-tolerant interconnection network for supersystems. IEEE Transactions on Computers,. 31(4)5, pp. 443-454, May 1982.
18. J.H. Patel. Processor-memory interconnections for mutliprocessors. Proceedings of 6th Annual Symposium on Computer Architecture New York, pp. 168-177, 1979.
19. T.H. Theimer, E. P. Rathgeb and M.N. Huber. Performance Analysis of Buffered Banyan Networks. IEEE Transactions on Communications, vol. 39, no. 2, pp. 269-277, February 1991.
20. H. Mun and H.Y. Youn. Performance analysis of finite buffered multistage interconnection networks IEEE Transactions on Computers, pp. 153-161, 1994.
21. Y.-C.Jenq. Performance analysis of a packet switch based on single-buffered banyan network IEEE Journal Selected Areas of Communications, pp. 1014-1021, 1983.