Performance Analysis of dual priority single-buffered blocking Multistage Interconnection Networks

D. C. Vasiliadis , G. E. Rizos , C. Vassilakis
Department of Computer Science and Technology, University of Peloponnese
{dvas, georizos, costas}@uop.gr

Abstract

In this paper a novel architecture of dual priority single-buffered blocking Multistage Interconnection Networks (MINs) is presented. We analyzed their performance in the uniform traffic condition under various loads using simulations. We compared the dual priority architecture against a single priority MIN, by gathering metrics for the two most important network performance factors, namely packet throughput and the mean time a packet needs to traverse the network. We demonstrated the gain of the high priority packets against the low priority packets under different configuration schemas. In this paper we focus on studying the influence of the priority bit in the header field of transmitted packets on the performance of high and low priority traffic of a MIN. Performance prediction before actual network implementation and understanding the impact of parameter settings in a MIN setup are valuable assets for network designers for minimizing overall deployment costs and delivering efficient networks.

1. Introduction

Multistage Interconnection Networks (MINs) with crossbar Switching Elements (SEs) are frequently proposed for interconnecting processors and memory modules in parallel multiprocessor systems. MINs have been recently identified as an efficient interconnection network for communication structures such as gigabit Ethernet switches, terabit routers, and ATM switches. A significant advantage of MINs is their low cost, taking into account the overall performance they offer. A significant advantage of MIN-type interconnection systems, is their ability to route multiple communication tasks concurrently. MINs with the Banyan [9] 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, queuing theory, Petri nets and Simulation.

Markov chains have been extensively used by many researchers. Bolch [2] and Merchant [16] used Markov chains in order to approximate the behavior of MINs under different buffering schemes. Bolch [2] also enhances Markov chains, with elements from queuing theory. Some authors that dealt with Markov chains also employed Petri nets as a modeling method. In the literature, there are several approaches that focus on Petri nets as those of German [8], Haas [10] and Lindermann [15]. Hsiao [11] and Theimer [21] studied MINs with uniform load traffic on inputs. Hot-spot traffic performance was also examined by Jurczyk [13], and Turner [22] dealt with multicast in Clos networks, as a subclass of MINs. Atiquzzaman [1] focused only on non-uniform arriving traffic schemes. Furthermore, Kleinrock [14] discusses approaches that examine the case of Poisson traffic on inputs of a MIN. In the industry domain, Cisco has built its new CRS-1 router [3, 4] as a multistage switch fabric. The switch fabric that provides the communications path between line cards is a 3-stage, self-routed architecture.

Packet priority is a common issue in networks, arising when some packets need to be offered better quality of service than others. Packets with real-time requirements (e.g. from streaming media) vs. non real-time packets (e.g. file transfer), and out-of-band data vs. ordinary TCP traffic [20] are two examples of such differentiations. There are already several commercial switches which accommodate traffic priority schemes, such as [6, 7]. These switches consist internally of single priority SEs and employ two priority queues for

each input port, where packets are queued based on their priority level. Chen and Guerin [5] studied an (*N X N)* non-blocking packet switch with input queues, built using one-priority SEs. Ng and Dewar [18] introduced a simple modification to load-sharing replicated buffered Banyan Networks to guarantee priority traffic transmission.

The internal switch structure used in all the above listed studies was a single-priority fabric with controlled inputs. In contrast to these previous works, our paper deals with an internal dual-priority switch fabric architecture in order to improve the performance parameters of high-priority packets. Packet priority is designated using a *priority bit* in the header field of transmitted packets, and we study the effect of this priority-handling scheme on the performance of high and low priority traffic, as well as on the overall network performance.

The remainder of this paper is organized as follows: in section 2 we briefly analyze a typical single-buffered blocking MIN with internal dual priority SEs. Subsequently, in section 3 we explain the performance criteria and parameters related to the network. Section 4 presents the results of our performance analysis, which has been conducted through simulation experiments, while section 5 provides the concluding remarks

2. Analysis of aN (*n x n*) MIN

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. A MIN with the Banyan property is defined in [9] and is characterized by the fact that there is exactly a unique path from each source (input) to each sink (output). Banyan MINs are multistage self-routing switching fabrics. Thus, each SE of k^{th} stage can decide in which output port to route a packet, depending on the corresponding k^{th} bit of the destination address.

An (N X N) MIN 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.

A configuration of an 8 X 8 delta network, one of the most widely used classes of Banyan MINs, which were proposed by Patel [19], is shown below:

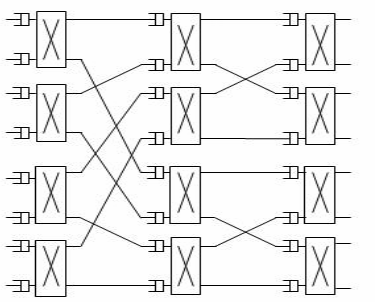

Figure 1. An 8 X 8 delta-2 network

A MIN is assumed to operate under the following conditions:

- 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.
- 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 p_a. This probability can be further broken down to p^h_a and p^l_a, which represent the arrival probability for high and low priority packets, respectively. It holds that $p_a = p^h_a + p^l_a$.
- A packet arriving at the first stage (k=1) 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.
- 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, there is a conflict. One of them will be accepted at random, with high priority packets having precedence over low priority packets, and the other will be blocked by means of upstream control signals.
- 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.

3. Performance Evaluation Methodology

In order to evaluate the performance of a (N X N) MIN with $n=\log_c N$ intermediate stages of (cxc) SEs,

we use the following metrics. Let T be a relatively large time period divided into u discrete time intervals ($\tau_1, \tau_2, ..., \tau_u$).

- *Average throughput* Th_{avg} is the average number of packets accepted by all destinations per network cycle. This metric is also referred to as *bandwidth*. Formally, Th_{avg} can be defined as

$$Th_{avg} = \lim_{u \to \infty} \frac{\sum_{i=1}^{u} n(i)}{u} \quad (1)$$

where $n(i)$ denotes the number of packets that reach their destinations during the i^{th} time interval.

- *Normalized throughput Th* is the ratio of the *average throughput* Th_{avg} to network size N. Formally, Th can be expressed by

$$Th = \frac{Th_{avg}}{N} \quad (2)$$

- *Relative normalized throughput* of high priority packets $RTh(h)$ is the *normalized throughput* $Th(h)$ of high priority packets divided by the *probability of arrivals* p^h_a of high priority packets.

$$RTh(h) = \frac{Th(h)}{p^h_a} \quad (3)$$

- *Relative normalized throughput* of low priority packets $RTh(l)$ is the *normalized throughput* $Th(l)$ of low priority packets divided by the *probability of arrivals* p^l_a of low priority packets.

$$RTh(l) = \frac{Th(l)}{p^l_a} \quad (4)$$

- *Average packet delay* D_{avg} is the average time a packet spends to pass through the network. Formally, D_{avg} can expressed by

$$D_{avg} = \lim_{u \to \infty} \frac{\sum_{i=1}^{n(u)} t_d(i)}{n(u)} \quad (5)$$

where $n(u)$ denotes the total number of packets accepted within u time intervals and $t_d(i)$ represents the total delay for the i^{th} packet.

We consider $t_d(i) = t_w(i) + t_{tr}(i)$ where $t_w(i)$ denotes the total queuing delay for i^{th} packet waiting at each stage for the availability of an empty buffer at the next stage queue of the network. The second term $t_{tr}(i)$ denotes the total transmission delay for i^{th} packet at each stage of the network, that is just $n*nc$, where n is the number of stages and nc is the network cycle.

- *Normalized packet delay* D is the ratio of the D_{avg} to the minimum packet delay which is simply the transmission delay $n*nc$. Formally, D can be defined as

$$D = \frac{D_{avg}}{n * nc} \quad (6)$$

The following parameters affect the above performance aspects of a MIN.

- *Buffer size* (b) is the maximum number of packets that an input buffer of a SE can hold. In our paper we consider a single-buffered (b=1) MIN. Double-buffered SEs lead to better exploitation of network, while the increase in delay can be tolerated [25].
- *Probability of arrivals* (p_a) is the steady-state fixed probability of arriving packets at each queue on inputs. In our simulation p_a is assumed to be p_a = 0.1, 0.2, …, 0.9, 1.
- *Ratio of high priority offered load* (r_h), where $r_h = p^h_a / p_a$. In our simulation r_h is assumed to be r_h =0.05, 0.10, 0.15, 0.20, 0.25.
- *Network size n*, where $n=\log_2 N$, is the number of stages of an (N X N) MIN. In our simulation n is assumed to be n=6, 8, 10.

4. Simulation and Performance Results

The performance of MINs is usually determined by modeling, using simulation [23] or mathematical methods [24]. In this paper we estimated the network performance using simulations. We developed a general simulator for MINs in a packet communication environment. The simulator can handle several switch types, inter-stage interconnection patterns, load conditions, switch operation policies, and priorities. We focused on an *(N X N)* Delta Network that consists of (2 X 2) SEs, using internal queuing. Each (2 X 2) SE in all stages of the MIN was modeled by two non-shared buffer queues. Buffer operation was based on FCFS principle. When there was a contention between the packets in a *SE*, it was solved randomly. The simulation was performed at packet level, assuming fixed-length packets transmitted in equal-length time slots, where the slot was 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.

Fig. 2 shows the *normalized throughput* of a single buffered MIN with 6 stages as a function of the *probability of arrivals* for the three classical models [12, 17, 21] and our simulation. All models are very accurate at low loads. 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.

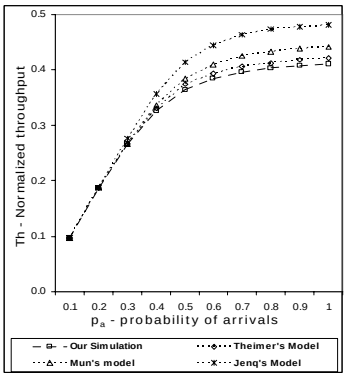

Figure 2. Normalized throughput of a single buffered 6-stage MIN

We performed extensive simulations to validate our results. Our simulation was also tested by comparing the results of the Theimer's model with those of our simulation experiments which were found to be in close agreement (differences are less than 1%).

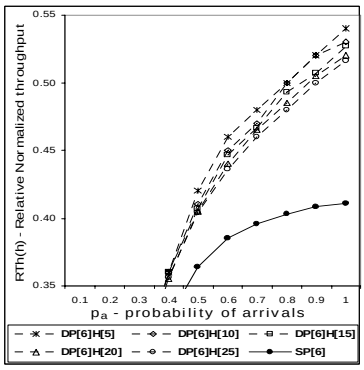

Figure 3. Normalized throughput of high priority packets of a 6-stage MIN

Fig. 3 illustrates the gains on *normalized throughput* of high priority packets at various rates (r_h=0.05, 0.10, 0.15, 0.20, 0.25) of high priority offered loads. In the diagram, curve SP[6] depicts the *normalized throughput* of a 6-stage MIN employing a single priority scheme, while curves DP[6]H[*X*] show the *relative normalized throughput* of a 6-stage MIN employing using a dual-priority scheme, where the probability of high-priority packet appearance is *X*%. In the case of r_h is 0.05 the *normalized throughput* approaches to *RTh(h)*=0.54 (gain 0.13 over the *normalized throughput* in the single-priority case) under full load traffic.

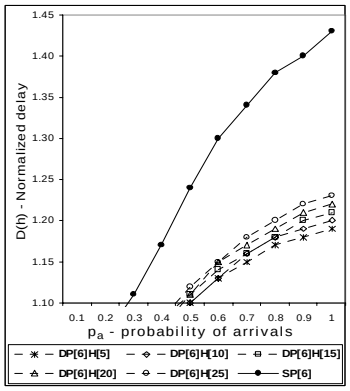

Figure 4. Normalized delay of high priority packets of a 6-stage MIN

Fig. 4 illustrates the minimization of *normalized delay* of high priority packets at various *rates* (r_h=0.05, 0.10, 0.15, 0.20, 0.25) of high priority offered loads. The decrements of packet delays are considerable for all setups. Especially, when r_h=0.05 the *normalized delay* is reduced to *D(h)*=1.19 (gain 0.24 as compared to the single-priority scheme).

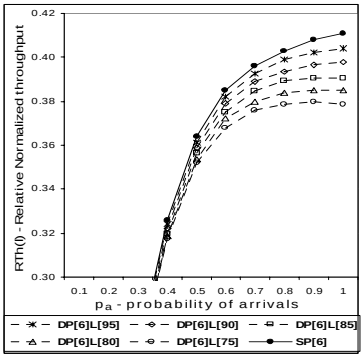

Figure 5. Normalized throughput of low priority packets of a 6-stage MIN

Fig. 5 presents the opposite case, where the loss of *normalized throughput* of low priority packets depends on the *rates* (r_l=0.95, 0.90, 0.85, 0.80, 0.75) of low priority offered loads. Especially, when the *rate* of high priority packets is low r_h=0.05 (or r_l=0.95) the loss of *normalized throughput* is negligible (0.007), while in the case of r_h=0.25 (or r_l=0.75) the loss is higher, but tolerable (0.032).

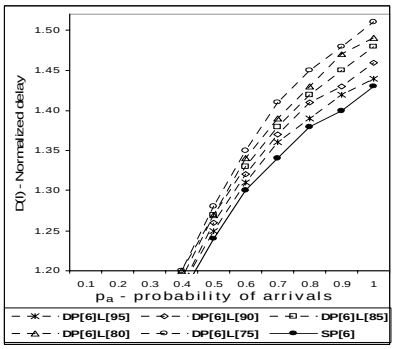

Figure 6. Normalized delay of low priority packets of a 6-stage MIN

Fig. 6 presents the corresponding increments in *normalized packet delays* of low priority packets, which are almost negligible for all configuration schemas. In the worst case, i.e. when high-priority packets are the 25% of the overall MIN traffic, the *normalized delay* of low-priority increases from 1.43 to 1.51.

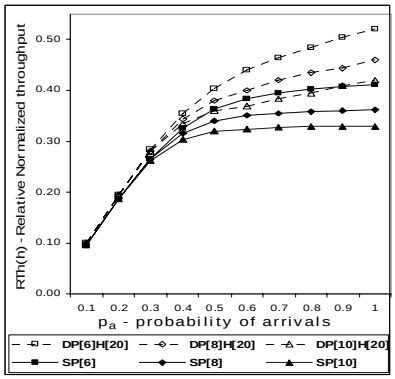

Figure 7. Normalized throughput of high priority packets of a k-stage MIN

Fig. 7 illustrates the gains on *normalized throughput* of high priority packets at different *network sizes* (number of stages n=6, 8, 10), when the *rate* of high priority packets r_h is 0.20. In the diagram, curves SP[X] depict the *normalized throughput* for single-priority MINs with X stages, while curves DP[X]H[20] show the *relative normalized throughput* for high-priority packets in dual-priority MINs with X stages. It is clear that the gain regarding the *normalized throughput* metric is considerable (≈ 0.10) for all *network size* configurations.

Fig. 8 illustrates the benefits related to the *normalized delay* metric for high priority packets at different *network sizes* (number of stages n=6, 8, 10), when the rate of high priority packets r_h is 0.20. The decrements of *normalized delays* are considered satisfactory (≈ 0.21) for all *network size* configurations.

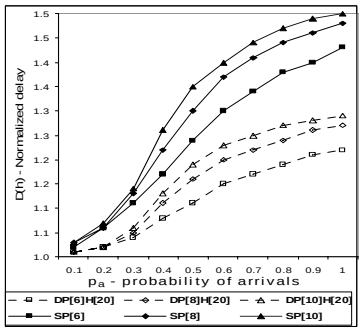

Figure 8. Normalized delay of high priority packets of a k-stage MIN

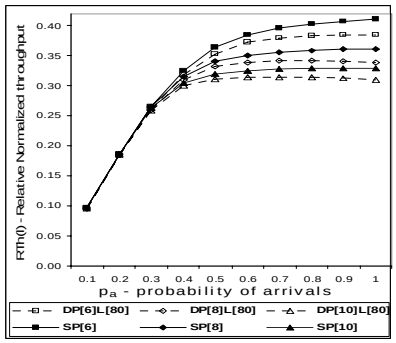

Figure 9. Normalized throughput of low priority packets of a k-stage MIN

Fig. 9 presents the opposite case, where *normalized throughput* deteriorates for low priority packets; the performance loss is however negligible (≈ 0.02) for all *network size* configurations (number of stages n=6, 8, 10), when the *rate* of low priority packets *is* r_l=0.80.

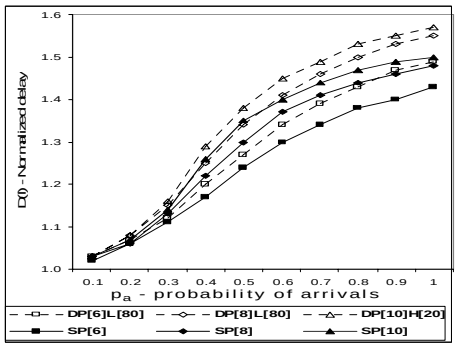

Figure 10. Normalized delay of low priority packets of a k-stage MIN

Finally, fig. 10 presents the increments in *normalized packet delays* of low priority packets, which are quite small and thus tolerable (≈0.07) for all *network sizes* (number of stages n=6, 8, 10), when the rate of low priority packets is r_l=0.80 or (r_h =0.20). It

is noteworthy that for higher values of r_l (or, equivalently, lower values of r_h).

5. Conclusions

In this paper we have presented a performance evaluation for dual-priority MINs. The study has been performed using simulation and considers different network loads, high to low priority *packet ratios* and MIN *sizes*. In all cases, it has been found that the gains for high-priority packets are considerable, both in terms of *throughput* and *delay*; the quality of service delivered for low priority packets, on the other hand, has been found to slightly deteriorate, but losses are quantified from negligible to tolerable. In all cases, reduction of delays has been found to adversely affect throughput (and vice versa), so MIN designers should carefully select related parameters (buffer sizes and use of packet priority designations) to best suit the needs of the applications that the particular MIN will support. The overall network performance is not affected by the introduction of the dual-priority scheme, since performance indicators (*throughput* and *packet delay*) appear to be very close to the indicators computed in other studies for single-priority MINs. Finally, the dual-priority scheme needs only one extra bit in the packet header (to indicate whether the packet is high- or low-priority), thus overhead in terms of additional control information is very small.

Future work will focus on assessing the role of SE *buffer sizes* in the performance of MINs and examining both the feasibility and the performance issues related to handling multiple (more than two) packet priority classes.

6. References

[1] M. Atiquzzaman and M.S. Akhatar. Efficient of Non-Uniform Traffic on Performance of Unbuffered Multistage Interconnection Networks. IEE Proceedings Part-E, 1994.

[2] 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.

[3] CISCO Systems, http://www.cisco.com/application/pdf/en/us/guest/products/ps5763/c1031/cdccont_0900aecd800f8118.pdf

[4] CISCO Systems, http://newsroom.cisco.com/dlls/2004/next_generation_networks_and_the_cisco_carrier_routing_system_overview.pdf

[5] J.S.C. Chen and R. Guerin. Performance study of an input queueing packet switch with two priority classes. IEEE Trans. Commun. 39(1) (1991) 117–126.

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

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

[8] R. German. Performance Analysis of Communication Systems. John Wiley and Sons, 2000

[9] G. F. Goke, G.J. Lipovski. Banyan Networks for Partitioning Multiprocessor Systems. Proc. 1st Ann. Symp. on Computer Architecture, 1973, pp. 21-28

[10] P. J. Haas. Stohastic Petri Nets. Springer Verlag, 2002.

[11] S.H. Hsiao and R. Y. Chen. Performance Analysis of Single-Buffered Multistage Interconnection Networks. 3rd IEEE Symposium on Parallel and Distributed Processing, pp. 864-867, December 1-5, 1991.

[12] Y.-C.Jenq. Performance analysis of a packet switch based on single-buffered banyan network IEEE Journal Selected Areas of commun. , pp. 1014-1021, 1983.

[13] M.Jurczyk. Performance Comparison of Wormhole-Routing Priority Switch Architectures. Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications 2001 (PDPTA'01); Las Vegas, pp. 1834-1840, 2001.

[14] T. Lin, L. Kleinrock. Performance Analysis of Finite-Buffered Multistage Interconnection Networks with a General Traffic Pattern. Joint International Conference on Measurement and Modeling of Computer Systems. Proceedings of the 1991 ACM SIGMETRICS conference on Measurement and modeling of computer systems, San Diego, California, United States, pp. 68 - 78, 1991.

[15] C. Lindermann. Performance Modelling with Deterministic and Stohastic Petri Nets. John Wiley & Sons, 1998.

[16] 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.

[17] H. Mun and H.Y. Youn. Performance analysis of finite buffered multistage interconnection networks, IEEE Trans. Comput., pp. 153-161, 1994.

[18] S. L. Ng and B. Dewar, Load sharing replicated buffered banyan networks with priority traffic. Connecting the System: Australian Telecom. Networks and Application Conference, Monash University, Clayton, Victoria, 1995, pp. 77-82.

[19] J.H. Patel. Processor-memory interconnections for mutliprocessors. Proceedings of 6th Annual Symposium on Computer Architecture New York, pp. 168-177, 1979.

[20] Stevens W. R., TCP/IP Illustrated: Volume 1. The protocols, 10th Ed., Addison-Wesley Pub Company, 1997.

[21] 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.

[22] J.Turner, R. Melen. Multirate Clos Networks. IEEE Communications Magazine, 41, no. 10, pp. 38-44., 2003

[23] D. Tutsch, M.Brenner. MIN Simulate. A Multistage Interconnection Network Simulator. 17th European Simulation Multiconference: Foundations for Successful Modelling & Simulation, Nottingham, SCS, pp. 211-216, 2003.

[24] D.Tutsch, G.Hommel. Generating Systems of Equations for Performance Evaluation of Buffered Multistage Interconnection Networks. Journal of Parallel and Distributed Computing, 62, no. 2, pp. 228-240, 2002.

[25] D.C. Vasiliadis, G.E. Rizos, C. Vassilakis. Performance Analysis of blocking Banyan Swithces. Proceedings of the IEEE sponsored CISSE 06, December, 2006.