Performance Analysis of Multistage Interconnection Networks determining optimal parameter values for data-intensive applications

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

Abstract

Multistage Interconnection Networks (MINs) are frequently used for connecting processors in parallel computing systems or constructing high speed networks such as ATM (based on Asynchronous Transfer Mode) and Gigabit Ethernet Switches. New applications require distributed computing implementations, but old networks are too slow to allow efficient use of remote resources. Moreover, multimedia are considered as applications with high bandwidth requirements. Some of them are also sensitive to packet loss and claim reliable data transmission. Specific applications require bulk data transfers for database replication or load balancing and therefore packet loss minimization is necessary in order to increase the performance of them. The demand for high performance multimedia services such as full motion video on demand is becoming an increasingly important driving force in the communication market in the Digital Age. Thus, the performance of MINs is a crucial factor, which we have to take into account in the design of new applications. Their performance is mainly determined by their communication throughput and cell latency, which have to be investigated either by time-consuming simulations or approximated by mathematical models. In this paper we investigate the performance of MINs in order to determine optimal values for hardware parameters under deferent operating conditions.

Keywords: *Banyan switches, multistage switches, performance of switching systems.*

1. Introduction

The growing need for new services has increased the demand for new networks that can support a variety of services, which include both traditional bandwidth-hungry data services and others that require a guaranteed quality of service *(QoS)* from the networks. The main advantage of MINs is their low cost, taking into account the performance they offer. So, *MINs* have received considerable interest in the development of networks. Nowadays there is a great interest about Switching Systems and especially for self-routing systems called *Banyan Switches*. Their performance has a direct effect in the overall performance of communication between networks, thus predicting their performance before actual network implementation and understanding the impact of parameter settings in a MIN setup are indispensable for delivering efficient networks.

To this end, a number of studies and approaches have been published. [1] and [2] assume uniform arriving traffic on inputs. [3] addresses non-Markovian processes, which are approximated by Markov models. Markov chains are also used in [4] to compare MIN performance under different buffering schemes. Hot spot traffic performance in MINs is examined in [5, 6] deals with multicast in CLOS networks as a subclass of MINs. [7] investigates group communication in circuit switched MINs by applying Markov chains as a modeling technique, calculating the throughput of finite and infinite buffered MINs under uniform and non uniform traffic. In the literature, there are also other approaches that focus only on non uniform arriving traffic [8, 9]. [10] discusses approaches that examine the case of Poisson traffic on inputs of a MIN. [11] analyzes communication throughput of single-buffered multistage interconnection networks consisting of *2X2* switches with maximum arrivals of packets *(100 %)*, using relaxed blocking model. In this work it is shown that the throughput is $\approx N / \sqrt{\log_2 N}$, where *N* is the size of a MIN. Furthermore, there are studies that deal with self-similar traffic on inputs.

In this paper we investigate the performance of MINs in terms of throughput, delay, and loss packet probability. We provide a comprehensive study analyzing the performance impact of network parameters. We study a typical *8X8* Banyan Switch, one of the most widely used classes of MINs, with internal queuing under variable length buffer size. The buffer management scheme uses the "back-pressure blocking" model, which is a more realistic approach than using the "block-and-lost" model. According to the back-pressure blocking model, when a packet finds that the next buffer position is occupied, it cannot be routed and is thus blocked. We assume a uniform load for each network input link in order to better analyze the behavior of a MIN.

The remainder of this paper is organized as follows: in section 2 we briefly analyze a typical 8X8 Banyan Switch. Then, in section 3 we explain the performance criteria and parameters related to the Banyan switch. Section 4 presents the results of our simulation experiments, while section 5 provides the concluding remarks.

2. Analysis of a MIN

An *NXN MIN* is constructed by $n=\log_c N$ stages of *cxc Switching Elements (SEs)* where *c* is the degree of the SEs. Each SE consists of *c* input ports and *c* output ports. There are exactly *m=N/c* SEs at each stage, thus the total number of *SEs* of a MIN is $(N/c)*\log_c N$ and there are $N*\log_c N$ interconnections among all stages, as opposed to the crossbar network which requires $O(N^2)$ *SEs* and links. In this paper we

study the Banyan Networks which were defined by [12] and are characterized by the property that there is exactly one path from any input to any output. In the case of a typical 8X8 Banyan Switch that consists of *2X2 SEs*, the number of stages is *n=3* and there are *m=4 SEs* at each stage. Banyan Switches are typical multistage self-routing switching fabrics. This means that every switch that accepts a packet in one of its input ports can decide in which of its output ports to forward this packet depending only on the packet's destination address. In the case of a typical 8X8 Banyan Switch, every *SE* of stage *k* can decide in which output port to send it based on the k^{th} bit of the destination address. If this bit is *0*, then the packet is forwarded to the upper output port and if the bit is *1* packet is forwarded to the lower output port. Multiple configurations are possible for a Banyan Switch. One possible configuration for 3-stage Banyan Switches is shown below (figure 1). It is assumed to operate under the following conditions:

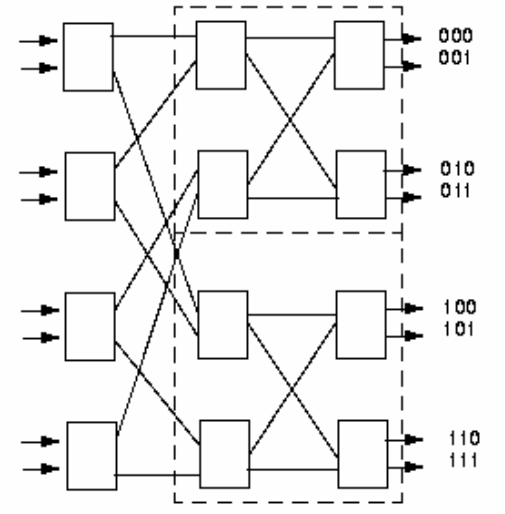

Fig 1. An 8X8 Banyan Switch

- We consider a typical 8X8 Banyan Switch that consists of 2X2 SEs (2 input ports and 2 output ports). It has the ability to accept packets in every input port and send packets randomly to one of two output ports. This grid operates with switching packets and static routing. The messages are transferred as simple cells and the route is based on k^{th} bit of the destination address for each stage (k=0, 1, 2) of the MIN. The flow of packets follows one direction, although in fact there are some acknowledgements to the opposite direction.

- The whole grid operates with in discrete time slots. In other words the packets are sent in specific time cycles (τ, 2τ, 3τ, …). Each SE has the ability to send only one packet to the next stage queues in a time cycle. A time cycle is considered to include the total transmission time plus the total queuing delays. The packets are uniformly distributed across all the destinations and each queue uses a FIFO policy for all output ports.

- When two packets in the i^{th} stage contend for the same buffer in the $(i+1)^{th}$ stage and there is not adequate free space for both of them to be stored, there is a conflict. In this case, one of them will be accepted at random and the other will be rejected by means of upstream control signals. The rejected packet stays in a buffer of i^{th} stage and has to try again to be forwarded in the next time slot. This is the mechanism of the back-pressure blocking model. Additionally, when contention arises on an output port for the head of internal queue and a packet is blocked, all other packets must wait behind that (the head-of-the-line blocking).

- All packets in input ports contain both the data to be transferred and the routing tag. As soon as they reach a destination port they are removed from the MIN. So, they cannot be blocked at the last stage.

- We assume oblivious routing algorithms, i.e. algorithms in which the path of a packet through the network is fixed at the source node issuing it. The path can be encoded as a sequence of labels of the successive switch outputs of the path. The routing logic at each switch is assumed to be fair, as conflicts are randomly resolved. Finally, we consider that the departure of a packet (if one exists) takes place at first in each time slot and then follows the arrival of packets (if they exist) in every queue of the grid. So, the inputs for simulation are the fixed probability p_a of arrivals of packets on inputs and constant β that shows the length of buffers for every queue.

3. Performance criteria and parameters

In order to evaluate the performance of an NXN MIN with $n=\log_2 N$ intermediate stages of 2X2 SEs, we use the following metrics. Let *T* be a relatively large time divided into υ discrete time intervals $(1\tau, 2\tau,\ldots,\upsilon\tau)$.

1. *Packet loss probability* (p_l) is the probability of lost packets on a queue on inputs.

2. *Packet blocking probability* (p_b^i) is the probability of blocking packets on a queue of an intermediate stage $(i=0, 1,\ldots, n-1)$ of the MIN.

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

$$\Theta_{avg} = \lim_{u \to \infty} \frac{\sum_{i=1}^{u} n_i}{u}$$

where n_i denotes the number of packets that reach their destinations during the i^{th} time interval τ.

4. *Normalized throughput (Θ)* is the ratio of the average throughput Θ_{avg} to network size N. Formally, Θ can be expressed by

$$\Theta = \frac{\Theta_{avg}}{N}$$

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

$$D_{avg} = \lim_{u \to \infty} \frac{\sum_{i=1}^{N_u} d_i}{N_u}$$

where N_υ denotes the total number of packets accepted within υ time intervals and d_i represents the total delay for the i^{th} packet. We consider $d_i = q_i + tr_i$ where q_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 tr_i denotes the total transmission delay for i^{th} packet at each stage of the network, that is $n*\tau$, where n is the number of stages and τ is the network cycle.

6. *Normalized packet delay (D)* is the ratio of the D_{avg} to the minimum packet delay which is simply the transmission delay $n*\tau$. Formally, D can be defined as

$$D = \frac{D_{avg}}{n * t}$$

The following parameters affect the packet loss and blocking probability, the throughput and the packet delay of the MINs.

- *Probability of arrivals (p_a)* is the 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, 0.99.

- *Buffer size (β)* is the maximum number of packets that an input buffer of an SE can hold. In our case β is assumed to be β*=0, 2, 4, 8.*

4. Simulation and performance results

The performance of MINs is usually determined by modeling, using simulation [13] or mathematical methods [14]. We implemented a simulator for an 8X8 Banyan Switch using internal queuing with the mechanism of Back-pressure blocking. To achieve synchronously operating SEs, a MIN is internally clocked. In each stage *k (0<=k<=n-1)* of non-shared buffer MINs, there is a FIFO buffer of size *(β=0, 2, 4, 8)* in front of each SE input. We performed extensive simulations to validate our results. The number of simulation runs was adjusted to ensure a *steady-state* operating condition for the MIN. In this paper, we present the results of simulation experiments for buffer size (β=0, 2, 4, 8) with a fixed probability of arrivals (p_a=0.1, 0.2, …, 0.9, 0.99).

Figures 2 and 3 present the packet blocking probability at intermediate stages. We notice here that it is higher at first stage, while at last stage it is assumed to be 0. We can also notice that under very heavy traffic (p_a>=0.9) buffer size must be increased (β=8) in order to keep the packet blocking probability at low level (p_b<0.01).

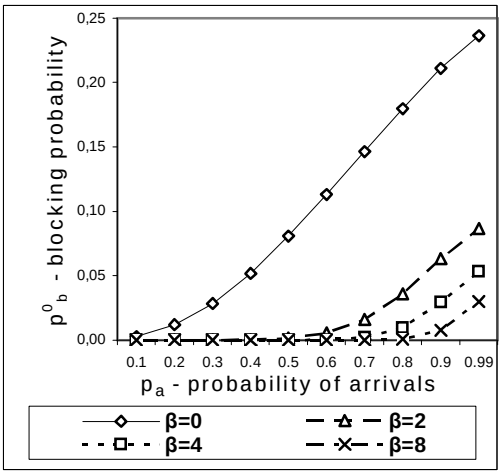

Fig 2. Blocking probability at stage 0

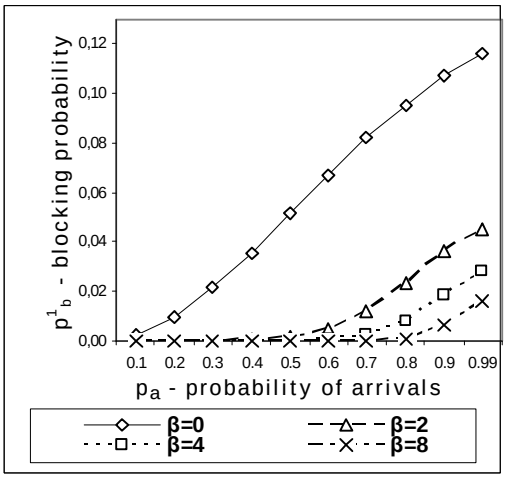

Fig 3. Blocking probability at stage 1

Figure 4 presents the packet loss probability on inputs and figure 5 illustrates the variation of normalized throughput. We notice here that throughput is satisfactory (Θ>0.80) only for (β>=2), because the values of probabilities of lost packets are high for unbuffered setups. We can also notice that packet loss probabilities are higher than

corresponding packet blocking probabilities at intermediate stages for all buffer size configurations.

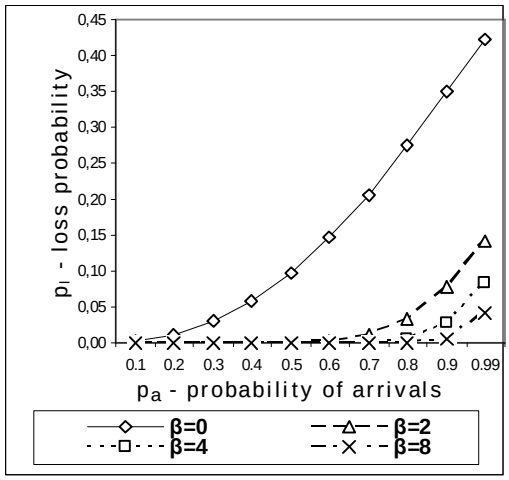

Fig 4. Loss probability at inputs

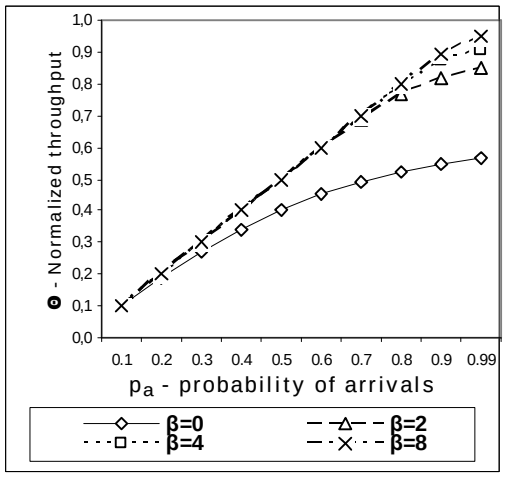

Fig 5. Normalized throughput

Figures 6 and 7 present the average and normalized packet delays. We notice that latency begins to increase as the buffer size increases.

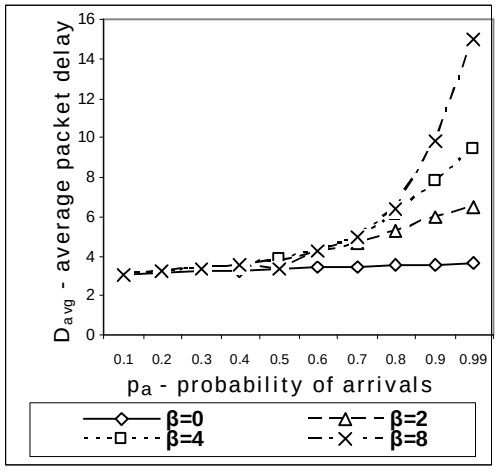

Fig 6. Average packet delay

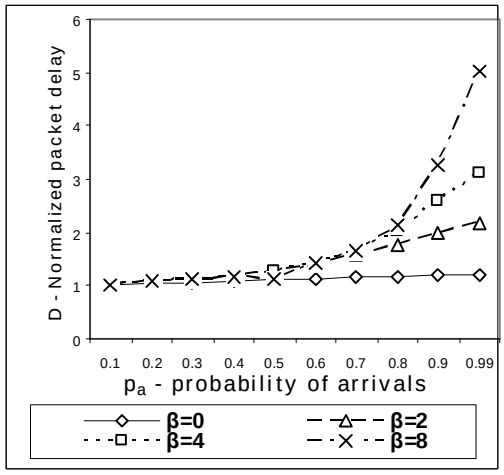

Fig 7. Normalized packet delay

5. Conclusion and future work

In this paper we analyzed the throughput (figure 5), the average and normalized packet delays (figures 6 and 7) of an 8X8 Banyan Switch under different loads (packet arrival probabilities) and various sizes of SE buffers. We also studied the loss and blocking probabilities in all stages (figures 2, 3, and 4). All performance parameters are accurately approximated in the simulation environment.

Loss and blocking probabilities have been found to increase when packet arrival probability increases. Under heavy load conditions, we can improve the MIN's behavior regarding the loss and blocking probabilities and increase its throughput, by using queues with greater buffer size. Simulation results show that normalized throughput is satisfactory ($\Theta>0.80$), only when buffer size is set to 2 or higher ($\beta>=2$). Especially, in the case of ($\beta=8$) the normalized throughput is ($\Theta>0.95$) under ultra heavy ($p_a>=0.99$) traffic.

On the other hand, the average and normalized delays have been found to increase as the buffer size increases. Larger buffers introduce larger delays as packets fill the buffers and they stay in the MINs longer. In the case of ($\beta=8$) normalized packet delay rises to ($D\approx5$) under ultra heavy ($p_a=0.99$) traffic conditions.

This study shows how to determine optimal values for hardware parameters under different operating conditions. In cases of modestly loaded MINs buffer size $\beta=2$ is considered to be the best choice. On heavily loaded MINs buffer size $\beta=4$ can be used for achieving better throughput ($\Theta>0.90$) with an acceptable ($D<3$) packet delay.

The results of this paper can also be applied in analyzing and evaluating the use of MINs as an intercommunication medium which will able to satisfy future data-intensive applications, especially in symmetric multiprocessor systems. Finally, a MIN

that consists of *cxc SEs (c>2)* can be analyzed, in order to make clear the role of the *c* parameter in the performance of MINs.

6. References

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

[2] T.H. Theimer, E. P. Rathgeb, and M.N. Huber, "Performance Analysis of Buffered Banyan Networks", *IEEE Transactions on Communications*, (39:2), pp. 269-277, February 1991.

[3] A. Merchart, "A Markov chain approximation for analysis of Banyan networks", *in Proc. ACM Sigmetrics Conference On Measurement and Modelling of Computer systems*, 1991.

[4] B.Zhou, M.Atiquzzaman. "A Performance Comparison of Four Buffering Schemes for Multistage Interconnection Networks". *International Journal of Parallel and Distributed Systems and Networks*, (5:1), pp. 17-25, 2002.

[5] 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, 1834.1840, 2001.

[6] J.Turner, R. Melen. "Multirate Clos Networks". *IEEE Communications Magazine*, (41:10), pp. 38-44., 2003

[7] Y. Yang, J. Wang. "A Class of Multistage Conference Switching Networks for Group Communication". *IEEE Transactions on Parallel and Distributed Systems*, (15:3), pp. 228.243, 2004.

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

[9] M. Atiquzzaman and M.S. Akhatar, "Effect of Non-Uniform Traffic on the Performance of Multistage Interconnection Networks", *Proceedings of the 9th International Conference on System Engineering*, Las Vegas, pp. 31-35, July 1993.

[10] T. Lin, L. Kleinrock, "Performance Analysis of Finite-Buffered Multistage Interconnection Networks with a General Traffic Pattern", *Proceedings of the 1991 ACM SIGMETRICS conference on Measurement and modeling of computer systems*, San Diego, California, United States, Pages, pp. 68 - 78, 1991.

[11] R. Rehrmann, B. Monien, R.. Luling, R. Diemann, "On the communication throughput of buffered multistage interconnection networks", in *Proceedings of the ACM SPAA* '96 pp. 152-161.

[12] G. F. Goke, G.J. Lipovski. "Banyan Networks for Partitioning Multiprocessor Systems", *Proceedings of the 1st Ann. Symposium on Computer Architecture*, 1973, pp. 21-28

[13] D. Tutsch, M.Brenner. "MIN Simulate. A Multistage Interconnection Network Simulator". *Proceedings of the 17th European Simulation Multiconference: Foundations for Successful Modelling & Simulation* (ESM'03); Nottingham, SCS, pp. 211-216, 2003.

[14] D.Tutsch, G.Hommel. "Generating Systems of Equations for Performance Evaluation of Buffered Multistage Interconnection Networks". Journal of Parallel and Distributed Computing, (62:2), pp. 228-240, 2002.