Performance Analysis of blocking Banyan Switches

D. C. Vasiliadis , G. E. Rizos , 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*—Banyan Networks are a major class of Multistage Interconnection Networks (MINs). They have been widely used as efficient interconnection structures for parallel computer systems, as well as switching nodes for high-speed communication networks. The performance of them is mainly determined by their communication throughput and their mean packet delay. In this paper we use a model that is based on a universal performance factor, which includes the importance aspect of each of the above main performance factors (throughput and delay) in the design process of a MIN. The model can also uniformly be applied to several representative networks. The complexity of the model requires to be investigated by time-consuming simulations. In this paper we study a typical (8X8) Baseline Banyan Switch that consists of (2X2) Switching Elements (SEs). The objective of this simulation is to determine the optimal buffer size for the MIN stages under different conditions.**

***Index Terms*—Multistage interconnection networks, baseline networks, delta networks, crossbar switches, packet switching, performance analysis.**

I. Introduction

MINs have been recently identified as an efficient interconnection network for a switching fabric of communication structures such as gigabit Ethernet switch, terabit router, and ATM switch. They are also frequently used for connecting processors in parallel computing systems. They have received considerable interest in the development of networks. The main parameter is their low cost, taking into account the overall performance they offer. The important thing about an interconnection system is that it has the capacity to route many communication tasks concurrently. A conflict occurs when more than one packet insist on the same communication resource. When a packet meets the next buffer position occupied then it cannot be routed and is thus blocked. The primary purpose of buffers in a SE is to prevent loss of packets due to routing conflicts.

Thus, insofar, a number of studies and approaches have been published. There are studies with uniform arriving traffic on inputs like [1,2] . [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 by [5,6] deals with multicast in Clos networks as a subclass of MINs. [7] uses mathematical methods. Group communication in circuit switched MINs is investigated by applying Markov chains as a modeling technique. Merchant calculates 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. Rehrmann [11], makes an analysis of communication throughput of single-buffered multistage interconnection networks consisting of (2X2) switches with maximum arrivals of packets 100%, using relaxed blocking model. Furthermore, there are studies that deal with self-similar traffic on inputs.

In this paper, we assume that packets are uniformly distributed across all the destinations and each queue uses a *FIFO* policy for all output ports. We study the performance of *a Baseline Banyan Switch* with blocking SEs that operates under different conditions. At first we present and analyze a typical (8X8) *Baseline Banyan Switch*. Then, we explain the performance criteria and parameters of this. Finally we present the results of our simulation experiments and provide the concluding remarks.

II. Analysis of a (NxN) Banyan Switch

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 *Banyan Network* was defined by [12] and is characterized by the property that there is exactly a unique path from each source (input) to each sink (output).

The path can be encoded as a sequence of labels of the successive outputs of the SEs. Thus, *Banyan Switches* are multistage self-routing switching fabrics. That means, each *SE* 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 destination address of it. A *SE* of stage k can decide in which output port to send it based on the k^{th} bit of the destination address and the *k-bit* shuffle algorithm. If the corresponding 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.

A (*NXN*) *Banyan Switch* can be constructed by $n=\log_c N$ stages of (*cxc) SEs*, where *c* is the degree of them. 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(NlogN*) interconnections among all stages, as opposed to the crossbar network which requires $O(N^2)$ links.

In this paper we study a typical *Baseline Banyan Switch* of dimension *(8X8)* that consists of 12 small *SEs* each of degree *(2X2).* This type of *Banyan Switches* provides both benefits of *Omega* and *Generalized Cube Switches* (destination routing, partitioning and expandability). A configuration with finite size non-shared buffer queues is shown below in the *figure 1*. It is assumed to operate under the following conditions:

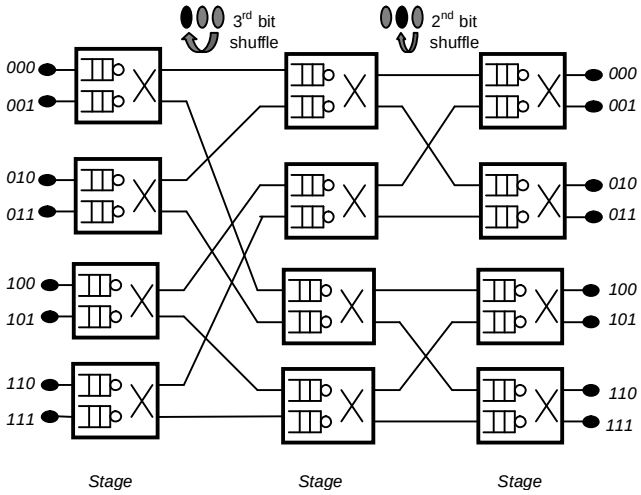

Fig.1 A (8X8) Baseline Banyan Switch

- The network clock cycle consists of two phases. In the first face 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.

- 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 the i^{th} stage contend for the same buffer at 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 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 *(k=3).*

III. Performance Evaluation Methodology

In order to evaluate the performance of a *(NXN) MIN* with $n=\log_c N$ intermediate stages of *(cxc) SEs*, we use the following metrics. Let *T* be a relatively large time divided into *u* discrete time intervals ($\tau_1, \tau_2,\ldots,\tau_u$).

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

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

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

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

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

$$D_{avg} = \lim_{u\to\infty} \frac{\sum_{i=1}^{n(u)} t_d(i)}{n(u)} \qquad (3)$$

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

$$D = \frac{D_{avg}}{n * nc} \qquad (4)$$

- *Universal performance (U)* is defined by the following relation of two above normalized contrary factors: one must be minimized *(D)* and the other must be maximized *(Th)*. Formally, U can be expressed by

$$U = \sqrt{D^2 + \frac{1}{Th^2}} \qquad (5)$$

It is obvious that, when the packet delay factor becomes smaller or/and throughput factor becomes larger the universal performance factor *(U)* becomes smaller. Consequently, as the universal performance factor *(U)* becomes smaller, the performance of a MIN is considered to improve. Because the above factors (parameters) have different measurement units and scaling, we normalize them to obtain a common value domain. Normalization is performed by dividing the value of each factor by the (algebraic) maximum value that this factor may attain. Thus, the equation (5) can be replaced by the following equation:

$$U = \sqrt{\left(\frac{D}{D^{\max}}\right)^2 + \left(\frac{Th^{\max}}{Th}\right)^2} \qquad (6)$$

where D^{max} is the maximum value of normalized packet delay *(D)* and Th^{max} is the maximum value of normalized throughput.

- *Universal performance ($U_{wd,wt}$)* with weight factors w_d,w_t includes the importance aspect of each factor in the design process of a MIN. Formally , $U_{wd,wt}$ can be expressed by

$$U_{wd,wt} = \sqrt{w_d * \left(\frac{D}{D^{\max}}\right)^2 + w_t * \left(\frac{Th^{\max}}{Th}\right)^2} \qquad (7)$$

Effectively, the values of w_d and w_t will be chosen by the MIN designers to reflect the significance that the corresponding factor (delay and throughput respectively) has in the particular MIN.

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

- *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.

- *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 $p_a = 0.1,0.2,\ldots,0.9,0.99$.

IV. Simulation and Performance Results

The performance of *MINs* is usually determined by modeling, using simulation [13] or mathematical methods [14]. 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, loading conditions, and switch operation policies. We focused on an *(8X8) Banyan Switch* that consists of *(2X2) SEs,* using internal queuing. Each *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 the 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. Statistics 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 .

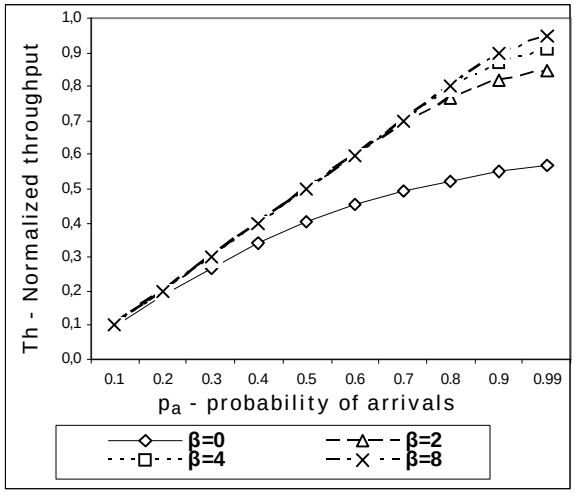

Fig.2 Normalized throughput vs. probability of arrivals

This section summarizes the results obtained from simulating the behavior of the *MIN* using various performance parameter value combinations. The objective of the simulation is to determine the optimal buffer size for the *MIN* stages under different conditions.

Figure 2 presents the relation between the normalized throughput performance metric and the arrival probability under different buffer sizes. This diagram clearly shows that using no buffer *(β = 0)* is not a good option, since *42%* approximately of the network capacity is lost, mainly due to the excessive number of dropped packets. Analytical results of our simulation were validated by comparing them with earlier works. S.H. Hsiao and R.Y. Chen [1] in *figures* 5 and 7 represent the *normalized throughput* (*Th*) of an *(NXN) Banyan Switch.* It was investigated either by time-consuming simulations or approximated by mathematical models. In those figures there is a comparison in *normalized throughput* (*Th*) with respect to number of stages under maximum value of *probability of arrivals* (p_a*=1*) with *buffer size* (*β=0 ;* only the processors of SEs have a single buffer). We notice that in the case of a 3-stage *MIN*, the *normalized throughput* ranges from 0.5 to 0.6. In our simulation the corresponding *normalized throughput* is (*Th≈0.57*).

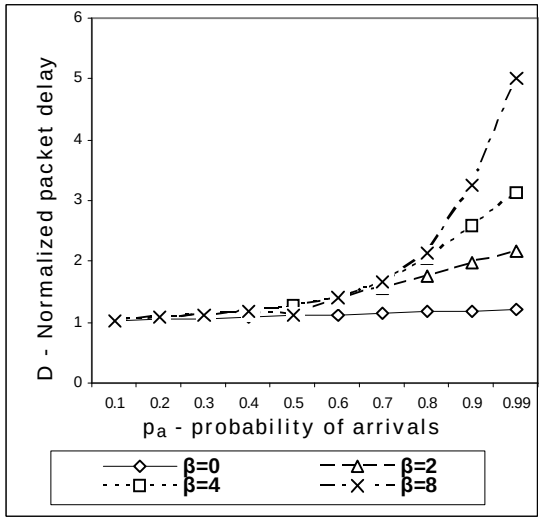

Fig.3 Normalized packet delay vs. probability of arrivals

Figure 3 illustrates the normalized packet delay for the various buffer sizes (*0, 2, 4* and *8*), when the arrival probability ranges from *0.1* to *0.99*. It is clear that the normalized packet delay significantly increases for large buffer sizes (*4* and *8*) when the arrival probability exceeds *80%*; however we should note that for small buffer sizes, the probability that a packet is dropped under heavy load (arrival probability > *80%*) is also considerable [15]

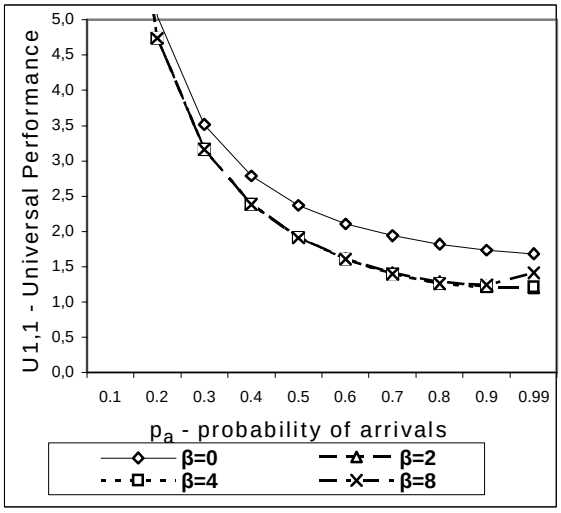

Fig.4 Universal performance factor with equal weights for individual factors

Figures 4-6 illustrate the relation of the combined performance indicator *U* to the arrival probability under different buffer sizes. Recall from *section 3* that the combined performance indicator is itself parametric, allowing *MIN* designers to designate the importance of each individual factor (packet delay and throughput) through the use of weights. Thus, *figure 4* depicts the case when the two factors are considered of equal importance ($w_d = w_t = 1$).

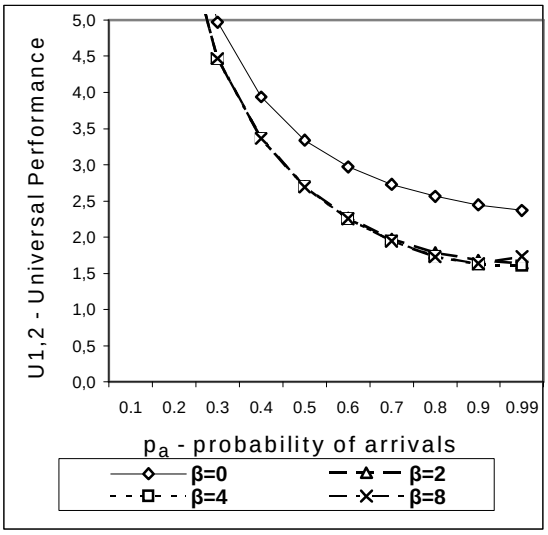

Fig.5 Universal performance factor with delay weight = 1 and throughput weight = 2

Figure 5 presents the case of a *MIN* where the overall throughput (and consequently, the exploitation of the available network capacity) is considered of greater importance; in this case w_d is set to *1*, while w_t is set to *2*.

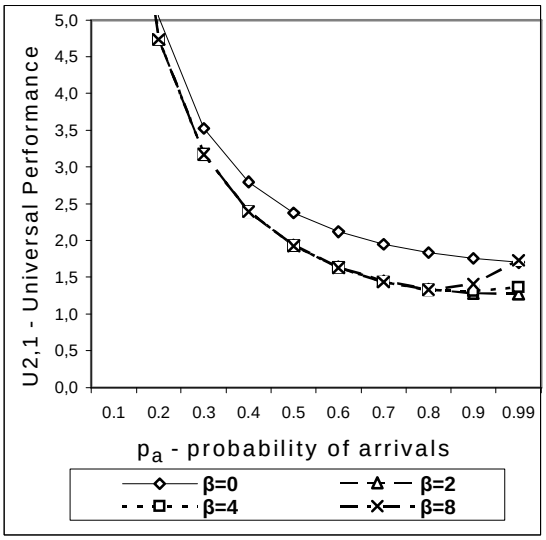

Fig.6 Universal performance factor with delay weight = 2 and throughput weight = 1

Finally, *figure 6* illustrates the opposite case, where the minimization of packet delays is the primary consideration of the *MIN* designers.

V. Conclusion

The simulation results presented above provide useful insights for MIN designers regarding the network throughput and performance parameters, under different loads and buffer sizes. The combined metric *Universal Performance Factor* ($U_{wd,\ wt}$) introduced in this paper gives an overall, single-dimension estimate of the network performance by allowing MIN designers to assign weights to individual performance factors; it is expected that MIN designers will choose weights accordingly to reflect the importance of each performance factors in the MIN operation.

An important finding from the simulation results is that the Universal Performance Factor deteriorates significantly when the switching element buffer size increases from 4 to 8. This happens because the throughput gains from increasing the buffer size from 4 to 8 are almost negligible, while the corresponding increment in the average packet delay within the MIN is considerable. This becomes more apparent when the w_d factor (i.e. the weight assigned to the delay performance parameter) is set higher than the w_t factor (the throughput factor weight).

At an application level, multimedia and streaming-oriented communications typically require small end-to-end packet delays, thus it is expected that in such contexts MIN designers will opt for small buffer sizes, with the values of 2 and 4 being the prevalent candidates. Especially for heavily loaded MINs, the choice of buffer size = 2 leads to the optimal value for the Universal Performance Factor. For MINs that do not exhibit such real-time requirements (and thus w_d will be equal to or smaller than w_t), a choice of buffer size = 4 is acceptable, since the network throughput is better exploited, while the additional end-to-end packet delay can be tolerated.

References

[1] 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.

[2] 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.

[3] A. Merchart, A Markov chain approximation for analysis of Banyan networks, in Proc. ACM Sigmetrics Conf. 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, no. 1: 17.25, 2002.

[5] M.Jurczyk. Performance Comparison of Wormhole-Routing Priority Switch Architectures. In 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, no. 10: 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, no. 3: 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”, 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“, 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, Pages: 68 - 78, 1991.

[11] R. Rehrman, B. Monien, R.. Luling, R. Diemann, On the communication throughput of buffered multistage interconnection networks, in ACM SPAA '96 pp. 152-161.
[12] G. F. Goke, G.J. Lipovski. Banyan Networks for Partitioning Multiprocessor Systems, Proc. 1st Ann. Symp. on Computer Architecture, 1973, pp. 21-28
[13] D. Tutsch, M.Brenner. .MIN Simulate. A Multistage Interconnection Network Simulator.. In 17th European Simulation Multiconference: Foundations for Successful Modelling & Simulation (ESM'03); Nottingham, SCS, 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, no. 2: 228.240, 2002.
[15] D.C. Vasiliadis, G..E.Rizos Simulation for Multistage Interconnection Networks using relaxed blocking model. Proceedings of the ICCMSE 2006 conference, Greece, 2006.