Performance Study of Multi-Layered Multistage Interconnection Networks under Hotspot Traffic Conditions

D. C. Vasiliadis[a,b], G. E. Rizos[a,b], C. Vassilakis[a]

[a]Department of Computer Science and Technology
University of Peloponnese
Tripolis, Greece
[b]Technological Educational Institute of Epirus,
Arta, Greece
dvas@uop.gr, georizos@uop.gr, costas@uop.gr

Abstract— **the performance of Multistage Interconnection Networks (MINs) under hotspot traffic, where some percentage of the traffic is targeted at single nodes, which are also called hot-spots, is of crucial interest. The prioritizing of packets has already been proposed at previous works as alleviation to the tree saturation problem, leading to a scheme that natively supports 2-class priority traffic. In order to prevent hotspot traffic from degrading uniform traffic we expand previous studies by introducing multi-layer Switching Elements (SEs) at last stages in an attempt to balance between MIN performance and cost. In this paper the performance evaluation of dual-priority, double-buffered, multi-layer MINs under single hotspot setups is presented and analyzed using simulation experiments. The findings of this paper can be used by MIN designers to optimally configure their networks.**

Keywords-Multistage interconnection networks; performance evaluation; hotspot traffic; multi-layer networks

I. Introduction

Multistage Interconnection Networks (MINs) with crossbar Switching Elements (SEs) are often used as communications infrastructure in the domains of networked systems and multiprocessor systems. In the former, MINs are employed to construct the communication backplane of high-performance networking elements, including terabit routers and gigabit Ethernet switches; in the latter, MINs are used for interconnecting processor nodes with memory chips. The spread of MINs can be attributed to their potential to concurrently route multiple packets, as well as to their cost/performance ratio, which is quite small, compared to other approaches.

MINs can be distinguished in two major sub-categories, those which exhibit the Banyan [3] property (including Delta Networks [10], Generalized Cube Networks [1] and Omega networks [6]), and those that do not. Banyan MINs are generally preferred over their non-Banyan counterparts, since they are cheaper and simpler to build and control.

The increase of MIN technology adoption has attracted considerable research efforts, which target to investigate the performance of MINs under various traffic load, traffic patterns and configurations. In these efforts, researchers have considered the parameters of offered load volume, switching elements' buffer size (e.g. [4], [21], [27]), overall size of the MIN network (number of stages – e.g. [27], [28]), priority schemas, policies and mechanisms (e.g. [12], [13], [22]) and traffic patterns (including uniform vs. hotspot e.g. [2], [8], [12], [13], [27] and unicast vs. broadcast/multicast e.g. [9], [16]). Issues related to MIN architecture, such as multilayer configurations [17] and wiring [7] and routing algorithms (e.g. [20]) have also been considered in research efforts.

In order to assess the performance of MINs, researchers have followed mainly two approaches. The first approach uses analytical methods, such as queuing theory, Perti-nets and Markov chains, while the second is simulation-based. The simulation-based approach [18] has been preferred against mathematical modeling [19] since it has a number of desired properties, including the accuracy of the results that can be obtained using simulation, the increased flexibility and the ability to capture all the aspects of an architecture requiring fewer abstractions in the model [29],[30].

Handling traffic with hotspot traffic shape and different priorities are two issues that attract the attention of researchers, due to the fact that these issues are frequently encountered in real-world systems. Hotspot traffic shape occurs when a considerable amount of the overall communication volume is targeted to a specific endpoint, typically occurring when a network contains a server accessed by numerous clients or when trunk ports are used to interconnect different network devices (e.g. if two switches A and B are interconnected via trunk ports t_a and t_b, all communication originating from nodes attached to switch A and directed to a node attached to B is effectively routed to t_a, to be then forwarded to switch B for delivery to the destination; similarly, all communication originating from nodes attached to switch B and directed to a node attached to A is effectively routed to t_b).

Allowing the specification of packet priorities and offering different classes of service to packets with different priority designation is another important issue in contemporary networks. The IEEE 802.1p standard designates four "normal" application priorities (best effort, background, excellent effort and critical applications), reserving two additional priorities for real-time media (video and voice) and two more for management (network and internetwork control). The TCP protocol [14] also distinguishes between ordinary and out-of-band data.

While these two issues have been researched independently in the context of MIN performance, the joint effect of packet priorities and hotspot traffic on the performance of MINs has not been adequately explored insofar. Two notable works that address these two issues together are [12] and [13], but they discuss an *extreme hotspot situation*, where all inputs send traffic to a specific output link and, additionally, all high-priority traffic is sent by a single input. Moreover, the MINs considered in these works are *single-buffered*, while more recent works (e.g. [23] and [24]) have shown that using double buffering or asymmetric buffering is beneficial for performance. [25] studies the joint effect of hotspot traffic and priorities in a MIN, showing that the performance of communication endpoints "near" the hotspot (cf. figure 1) is poor, especially regarding packet delay, even for modest loads ($\lambda \geq 0.25$).

In this paper, we consider multi-layer MIN architecture [17] as a solution to the performance bottlenecks observed under the hotspot traffic pattern, and examine the performance aspects of multi-layer MINs under different rates of offered load. Taking into account the fact that the performance of MIN outputs under hotspot traffic is *not* uniform [11], but depends on the amount of overlapping that the path to the specific output has with the path to the hotspot output, we classify MIN outputs into groups according to this characteristic and collect performance metrics for each group individually. Our study also takes into account the existence of two different priority classes, namely high-priority and low-priority, and performance metrics are collected and presented individually for each priority class. Our study is performed using simulation, and we present metrics for the two most important network performance factors, namely *throughput* and *delay*. We also adopt the metric of *universal performance factor* introduced in [21], which combines *throughput* and *delay* into a single metric, allowing the designer to express the perceived importance of each individual factor through *weights*.

The rest of this paper is organized as follows: in section II we briefly analyze the operation a Delta Network operating under hotspot traffic conditions and natively supporting 2-class routing traffic. In the same section we also describe the environment and operation of a MIN comprising of an initial single-layer segment having the Banyan property [3] and a subsequent multi-layer segment which sacrifices the Banyan property in order to achieve higher performance. Subsequently, in section III we illustrate the performance criteria and parameters related to this network. The results from our simulation experiments are presented in section IV, while section V concludes the paper and outlines future work.

II. ANALYSIS OF 2-CLASS PRIORITY DELTA NETWORKS UNDER HOTSPOT ENVIRONMENT

Multistage Interconnection Networks (MINs) are 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 preceded) by link states. All different types of MINs ([1], [6], [10]) with the Banyan property [3] are self-routing switching fabrics and they are characterized by the fact that there is exactly a unique path from each source (input) to each sink (output).

Under a multiple-priority scheme, when a packet is entered in the MIN, its priority is specified by the application or the architectural module that has produced the packet. The priority is henceforth reflected into a field in the packet header and is maintained throughout the lifetime of the packet within the MIN. This field should have an amble size of bits to accommodate all priority classes (e.g. 1 bit for 2 priorities, 3 bits for 8 priority classes and so forth). An example (8X8) MIN is illustrated at figure 1, supporting natively 2-class priorities. In order to support priority handling, each SE has two transmission queues per link, accommodated in two (logical) buffers, with one queue dedicated to high priority packets and the other dedicated to low priority ones ([12], [13], [22]). During a single network cycle, the SE considers all its links, examining for each one of them firstly the high priority queue. If this is not empty, it transmits the first packet towards the next MIN stage; the low priority queue is checked only if the corresponding high priority queue is empty. Packets in all queues are transmitted in a first come, first served basis. In all cases, at most one packet per link (upper or lower) of an SE will be forwarded for each pair of high and low priority queues to the next stage. Each queue is assumed to have two buffer positions for incoming packets.

The traffic pattern in the MIN depicted in figure 1 is hotspot, with a single hotspot output, namely output 0: this output is termed *hotspot*, because it receives a higher share of the overall MIN traffic than other outputs. More formally if we denote as $p_{i,j}$ the probability that a packet appearing in input port i has output port j as its destination, then $p_{i,0} > p_{i,j}$ $\forall i: 0 \leq i \leq 7$, $\forall j: 1 \leq j \leq 7$. Thus all input ports (0-7) direct to single hotspot output an increased share of the traffic they generate.

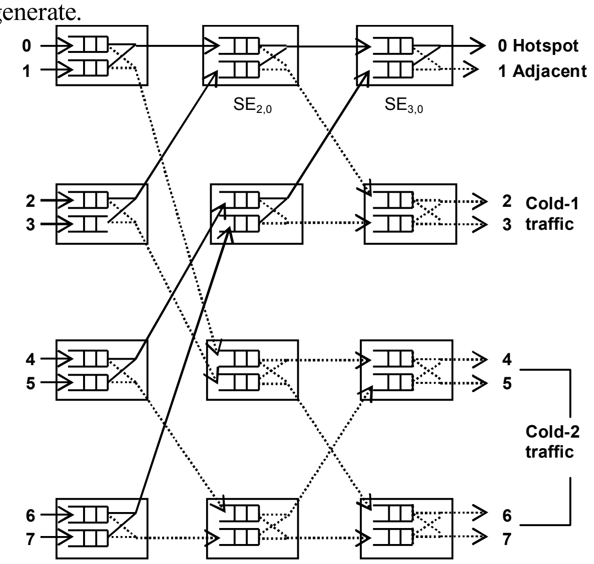

Figure 1. A single-layer 8X8 MIN under hotspot traffic

Within a hotspot environment, all SEs of MIN can be classified into two different groups: *Group-hst* and *Group-nt*, where *hst* (**h**ot**s**pot **t**raffic) stands for those SEs which receive and forward hotspot traffic, while *nt* (**n**ormal **t**raffic) stands for those SEs in which receive only normal traffic; i.e. they are free of hotspot traffic. In figure 1 we can distinguish the following categories of outputs (in accordance to [11]):

- *output* 0, which is the hotspot output.
- *output* 1, which is the output adjacent to the hotspot output. Packets directed to this output have to contend with packets addressed to the hotspot output at all stages of the MIN, and they are free of such contention only when traversing the output link.
- *outputs* 2 and 3, which are free of contention with packets addressed to the hotspot output when they traverse the last stage of the MIN. These outputs are termed as Cold-1, since they are free of contention with hotspot traffic for one stage.
- *outputs* 4-7, which are free of contention with packets addressed to the hotspot output when they traverse the last two stages of the network and thus are termed as Cold-2.

Generalizing, in an *i*-stage MIN, its output ports can be classified into the following (i+1) zones: *hotspot*, *adjacent*, and Cold-j ($1 \leq j \leq i$-1).

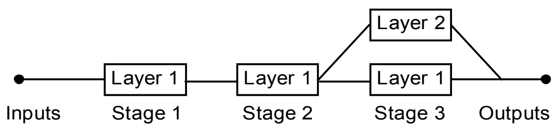

Figure 2. A lateral view of an 8X8 multi-layer MIN

In this paper we also extend previous studies by considering multi-layer MINs. Figure 2 illustrates the lateral view of an (8X8) multi-layer MIN, which employs multiple layers only at the final stage. Thus, the example network consists of two segments, an initial single-layer one and a subsequent multi-layer one (with 2 layers). The rationale behind choosing such an architecture is to have switching elements and more paths (and therefore more routing power) available at the final stages of the MIN, where the hotspot traffic from all inputs converges towards the hotspot output, creating bottlenecks. It is worth noting that, in the architecture presented in figure 2, packet forwarding from stage 2 to stage 3 is blocking-free, since packets in stage-2 SEs do not contend for the same output link. To make this more clear, consider the case that *both* queues in the topmost SE of stage 2 in figure 1 ($SE_{2,0}$) need to forward a packet towards output 0 ($SE_{3,0}$ - the SE containing the hotspot). In a single-layer MIN only one packet would be forwarded through the link connecting $SE_{2,0}$ to $SE_{3,0}$, and the other packet would be blocked. In the multi-layer MIN, however, of figure 2, there would exist *two* $SE_{3,0}$ elements (one for each layer, $SE_{3,0L1}$ and $SE_{3,0L2}$), and there would be *two* available links, one connecting $SE_{2,0}$ to $SE_{3,0L1}$ and one connecting $SE_{2,0}$ to $SE_{3,0L2}$. Therefore the packet from the upper queue of $SE_{2,0}$ would be forwarded to $SE_{3,0L1}$ through the first link and the packet from the lower queue would be forwarded to $SE_{3,0L2}$ through the second link, resulting in absence of contention.

Absence of contention is always possible for cases where the degree of replication of succeeding stage i+1 (which we will denote as l_{i+1}) is equal to $2*l_i$ (i.e. stage i+1 contains twice as many SEs as stage i). If, for some MIN with n stages there exists some nb ($1 <= nb < n$) such that $\forall$ k: $l_{k+1} = 2 * l_k$ ($nb <= k < n$), then the MIN operates in a non-blocking fashion for the last (n-nb) stages. Note that according to [17], blocking *can* occur at the MIN outputs, where SE outputs are multiplexed, if either the multiplexer or the data sink do not have enough capacity; in this paper however we will assume that both multiplexers and data sinks have adequate capacity.

We also note that the addition of multiple layers in the final stages effectively creates multiple paths between sources and destinations; therefore the MIN as a whole does not have the Banyan property. The MINs considered in this study retain the Banyan property within the initial, single-layer segment, while this property is dropped in the final, multi-layer one.

In this study, we consider a Multistage Interconnection Network that operates under the following assumptions at hotspot environment:

- Routing is performed by all SEs in parallel, thus the MIN can be considered to operate in a pipeline fashion. The pipeline is synchronized using an internal clock and operates in a slotted time model [15]. The service time for all SEs is deterministic.
- Each input of the MIN accepts only one packet within each time slot. A packet entering the MIN comprises of (a) the routing tag, which effectively contains the routing instructions for all SEs that the packet will traverse (b) the packet priority specification [only under multi-priority schemes] since in this paper we consider a dual-priority scheme, the priority specification is a single bit designating the packet as high- or low-priority one and (c) the packet payload, i.e. the actual data that are sent to the destination.
- All packets have the same size, arrivals are independent of each other and packets arrive with equal probability at all inputs.
- SEs operate in a store-and-forward fashion, i.e. each packet received by an SE is stored in a buffer until it can be forwarded to the next SE (or sent to the MIN output). To enable its store-and-forward operation, each SE incorporates one FIFO buffer per incoming link. When the FIFO buffer within an SE is full, the SE cannot accept further input packets from its predecessor SEs (or the MIN input), and a backpressure mechanism is employed to force packets to remain in the previous MIN stage until amble buffer space is available. Under this scheme, no packets are lost inside the MIN.
- When a multiple priority scheme is employed, each SE has a distinct FIFO buffer dedicated to each priority class per incoming link. In the packet-forwarding phase of operation, the SE examines its FIFO buffers successively, starting from the highest priority queue and working towards the lowest priority one. When a queue containing a packet is found, it is forwarded towards the successive MIN stage. For the two-priority scheme, in particular, the low-priority queue is only checked if the high-priority

queue contains no packets. In all cases, at most one packet per link (upper or lower) of an SE will be forwarded to the next stage. If both the upper and the lower incoming link buffers hold packets of the same priority to be forwarded to the next MIN stage through the same output link, the contention is solved randomly with equal probabilities.

- Hotspot traffic is modelled by having a fraction f_{hs} of the total offered load λ that is routed to the single hotspot output port. In this study, we consider this fraction to comprise solely of low-priority packets. The remaining load, i.e. $\lambda*(1- f_{hs})$ comprises both high- and low-priority packets and is uniformly distributed across all destinations, *including the hotspot*. Therefore, each MIN output except for the hotspot one, receives a load equal to $[\lambda*(1- f_{hs})/N]$ (with N being the number of outputs), while the hotspot output receives a load equal to $[f_{hs} * \lambda + \lambda*(1- f_{hs})/N] = [((N-\lambda)*. f_{hs} + \lambda)/N]$. Note that the hotspot output receives high-priority packets which are contained in the load fraction $[\lambda*(1- f_{hs})/N]$ that is addressed to it (whereas the fraction $\lambda*f_{hs}$ does not contain high-priority packets).
- Packets are removed from their destinations immediately upon arrival, thus packets cannot be blocked at the last stage.

While a number of packet switching techniques have been proposed in the literature and are used in commercial products (store and forward, virtual cut-through, wormhole, pipelined circuit switching and adaptive cut through switching [31]), in this paper we choose the store and forward technique for conducting the performance evaluation mainly because the performance of the store and forward technique has been more extensively studied in the literature and is therefore a better basis for comparing a situation that has been studied (hotspot traffic in single-layer MINs) with a situation that has been not investigated insofar (hotspot traffic in multi-layer MINs). Other switching techniques are not investigated in this paper, since our primary goal is to gain insight on how the MIN performance is affected by the introduction of multiple layers, and not the particular performance characteristics of different switching techniques, such as decreased *latency*. Finally, store and forward has been found to be more predictable, more resistant to saturation and free of issues such as deadlock, as compared to other techniques (e.g. wormhole- [32][33]), and these features would facilitate the interpretation of the performance analysis results.

III. PERFORMANCE EVALUATION PARAMETERS AND METHODOLOGY

A. MIN configuration and traffic load parameters

In this paper we extend our study on performance evaluation of MINs by comparing the performance metrics of 2-class priority MINs at a multi-layer architecture versus a single-layer one under hotspot traffic. All presented MINs are constructed by either single- or multi-layer SEs. Recall from section 2, since the second segment of multi-layer architecture is blocking-free, all SEs within the multi-layer segment are considered to have only the buffer space needed to store and forward a single packet. On the other hand, the SEs of the single-layer segment may employ different *buffer sizes* in order to improve the overall MINs performance. Under these considerations, the operational parameters of the MINs evaluated in this chapter are as follows:

- *Buffer size* (b) of a queue is the maximum number of packets that an input buffer of a SE can hold. In this study, symmetric double-buffered SEs (b=2) are considered for both high- and low-priority packets at the single-layer segment of MIN, where blockings can occur and thus additional buffers may be needed to store blocked packets and newly arriving packets. We note here that the particular *buffer size* has been chosen since it has been reported [21] to provide optimal overall network performance.
- *Number of stages n* is the number of stages of an (N X N) MIN, where $n=\log_2 N$. In our case study n is assumed to be 6, thus the MIN size is (64x64).
- *Offered load* (λ) is the steady-state fixed probability of arriving packets at each queue on inputs. In our simulation λ is assumed to be λ = 0.1, 0.2… 0.9, 1. λ can be further broken down to λ_{hs}, λ_{hp} and λ_{lp}, which represent the arrival probability of the initial hotspot traffic, and the high- and low- priority traffic of the rest offered load respectively. It holds that $\lambda = \lambda_{hs} + \lambda_{hp} + \lambda_{lp}$.
- *Hotspot fraction* (f_{hs}) is the fraction of the initial hotspot traffic which is considered to be f_{hs} =0.05. We fix f_{hs} to this value, since using a higher value for a network of this size would lead to quick saturation of the paths to the hotspot output.
- *Ratio of high priority packets* (r_{hp}), is the ratio of high priority offered load for the normal traffic –i.e. excluding the traffic addressed to the initial hotspot- which is uniformly distributed among all output ports and it is assumed to be r_{hp} =0.20. This ratio is generally adopted in works considering multiple priorities ([12], [13], [22]).
  Consequently, $\lambda_{hs} = f_{hs} * \lambda$,
  $\lambda_{hp} = r_{hp} * (1-f_{hs}) * \lambda$
  $\lambda_{lp} = (1-r_{hp}) * (1-f_{hs}) * \lambda$
- *Number of single-layer stages s* is the number of stages within the MIN where "traditional", single-layer SEs are employed. These stages are *always* the initial ones in the MIN, more routing power is required towards the last stages, due to the convergence of hotspot traffic. In our work, we consider the number of layers within each subsequent stage to be doubled, i.e. $nl(i+1) = 2* nl(i)\ \forall i$: $s \leq i < n$ [$nl(i)$ denotes the number of layers at stage i]. Doubling the number of layers in each subsequent stage guarantees that the last segment of the MIN operates in a blocking-free fashion, in the general case however, the number of layers in each stage i+1 within the multi-layer segment is subject to the constraint $nl(i) \leq nl(i+1) \leq 2*nl(i)$ [17]. Under the assumption that the number of layers within each subsequent stage doubles, the *Number of layers at the final stage l* will be equal to $2^{(n-s)}$. In this work, we consider s=4 and therefore l=4.

B. MIN Performance metrics

In order to evaluate the performance of a dual-priority MIN under hotspot environment the following metrics are used.

Average throughput $Th_{avg}(zone)$ of a specific output zone of MIN, where *zone*={hotspot, adjacent, Cold-1, …, Cold-(n-1)} is the mean number of packets accepted by all destination ports of this *zone* per network cycle. Formally, $Th_{avg}(zone)$ is defined as

$$Th_{avg}(zone) = \lim_{u \to \infty} \frac{\sum_{i=1}^{u} n_{zone}(i)}{u} \tag{1}$$

where $n_{zone}(i)$ denotes the total number of packets routed to this specific output *zone* that reach their destinations during the i^{th} time interval.

Normalized throughput $Th(zone)$ of a specific output *zone* of MIN is the obtained by dividing the throughput of a zone $Th_{avg}(zone)$ by the number of output ports within the particular zone $N(zone)$, giving thus a *per port* throughput metric. This is required, since the number of nodes within different zones may greatly vary from 1 to 2^{n-1}. Formally, $Th(zone)$ can be expressed by

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

where $N(zone)$={1, 1, 2, …, 2^{n-1}} for *zone*={hotspot, adjacent, Cold-1…Cold-(n-1)}.

Average packet delay $D_{avg}(zone)$ of packets having their destination within a specific output *zone* of MIN is the mean time that these packets require to traverse the network. Formally, $D_{avg}(zone)$ is expressed by

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

where $n(zone,u)$ denotes the total number of packets reaching their destinations in zone *zone* within u time intervals, while $t_d(zone,i)$ represents the delay of the i^{th} packet to travel from an input port to a port of the specific output *zone*. $t_d(zone,i)$ can be broken down to $t_w(zone,i)+t_{tr}(zone,i)$ where $t_w(zone,i)$ denotes the total waiting time of the i^{th} packet, i.e. the queuing delay for it waiting at each stage for the availability of an empty buffer at the next stage of the network, while $t_{tr}(zone,i)$ represents the total transmission time for i^{th} packet for all stages of the network. Since the network has deterministic service time, equal to the network cycle nc, $t_{tr}(zone,i)$ will be equal to $n*nc$, where n is the number of stages.

Normalized packet delay is used to eliminate the impact of the network size from the average packet delay metric, allowing for comparisons of delays between networks of different sizes. The need for introducing normalized packet delay stems from the fact that networks of different sizes have different *minimum delays* for the packets to traverse them, with the minimum delay for a network of size n being $n*nc$ (i.e. queuing delay equal to zero). Normalized packet delay is computed by dividing average packet delay with the minimum delay of the network, and can be formally expressed as

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

Relative normalized throughput of hotspot traffic RTh_{hs} is the *normalized throughput* $Th(hotspot)$ of the hotspot output port divided by the corresponding ratio of packets on all input ports which are routed to single hotspot output port.

$$RTh_{hs} = \frac{Th(hotspot)}{\mathrm{N} * \mathrm{f_{hs}} + (1 - \mathrm{r_{hp}}) * (1 - \mathrm{f_{hs}})} \tag{5}$$

Relative normalized throughput of high priority traffic RTh_{hp} is the *normalized throughput* Th_{hp} of high priority packets routed to all output zones divided by the corresponding ratio of high priority packets on input ports.

$$RTh_{hp} = \frac{Th_{hp}}{\mathrm{r_{hp}} * (1 - \mathrm{f_{hs}})} \tag{6}$$

We do not report different RTh_{hp} for each zone, since our experiments have shown that this parameter is not affected by the zone when the MIN operates under the parameter ranges listed above (section III.A).

Relative normalized throughput of low priority traffic $RTh_{lp}(zone)$ routed to a specific *zone* of output ports is the *normalized throughput* $Th_{lp}(zone)$ of such packets divided by the corresponding ratio of low priority packets on input ports.

$$RTh_{lp}(zone) = \frac{Th_{lp}(zone)}{(1 - \mathrm{r_{hp}}) * (1 - \mathrm{f_{hs}})} \tag{7}$$

Universal performance $U(zone)$ is defined through a formula involving the two major above normalized factors, namely $D(zone)$ and $RTh(zone)$: the performance of a *zone* of MIN is considered optimal when $D(zone)$ is minimized and $RTh(zone)$ is maximized, thus the formula for computing the *universal* factor arranges so that the overall performance metric follows that rule. Formally, $U(zone)$ can be expressed by

$$U(zone) = \sqrt{D(zone)^2 + \frac{1}{RTh(zone)^2}} \tag{8}$$

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 MIN is considered to be improved. Because the above factors (parameters) have different measurement units and scaling, we normalize them to eliminate the effect of the network size from these factors, similarly to the case of normalizing throughput and delay. Normalization is performed by dividing the value of each factor by the (algebraic) minimum or maximum value that this factor may attain. Thus, equation (8) can be replaced by:

$$U(zone) = \sqrt{\left(\frac{D(zone) - D(zone)^{\min}}{D(zone)^{\min}}\right)^2 + \left(\frac{RTh(zone)^{\max} - RTh(zone)}{RTh(zone)}\right)^2} \quad (9)$$

where $D(zone)^{min}$ is the minimum value of *normalized packet delay D(zone)* and $RTh(zone)^{max}$ is the maximum value of *relative normalized throughput*. Consistently to equation (8), when the *universal performance* factor *U*, as computed by equation (9) is close to 0, the performance of the specific *zone* of MIN is considered optimal whereas, when the value of *U* increases, its performance deteriorates. Finally, taking into account that the values of both *delay* and *throughput* appearing in equation (9) are normalized, $D(zone)^{\min} = RTh(zone)^{\max} = 1$, thus the equation can be simplified to:

$$U(zone) = \sqrt{(D(zone) - 1)^2 + \left(\frac{1 - RTh(zone)}{RTh(zone)}\right)^2} \quad (10)$$

IV. Simulation and Performance Results

In order to obtain the simulation results presented in this section, we developed a special simulator in C++, capable of handling 2-class priority MINs with multi-layer architecture, whose traffic follows the hotspot pattern. Each (2X2) SE was modeled by four non-shared buffer queues, where buffer operation was based on the first come first serviced principle; the first two buffer queues for high priority packets (one per incoming link), and the other two for low priority ones. Thus, at the simulator several parameters such as the *buffer-length,* the *number of input and output ports*, the initial *hotspot fraction*, the *ratio of high priority packets,* and the *number of layers* was considered.

Finally, the simulations were performed at packet level, assuming fixed-length packets transmitted in equal-length time slots, while the number of simulation runs was again adjusted at 10^5 clock cycles with an initial stabilization process 10^3 network cycles, ensuring a steady-state operating condition.

A. Simulator validation

To validate our simulator, we compared the results obtained from our simulator against the results reported in other works, selecting among them the ones considered most accurate under dual-priority and hotspot traffic in single-layer environments.

Since no other related works supporting performance evaluation metrics for dual priority MINs under hotspot traffic environment within a multi-layer architecture have been reported insofar in the literature, we validated our simulator against those that have been made available; i.e. single-priority under hotspot environment and dual-priority under uniform traffic conditions.

In the case of hotspot environment, the measurements reported in table 1 of [5] and those obtained by our simulator in the marginal case of single-priority traffic, where r_{hp} =0, f_{hs} =0.10, and N=8, have found to be in close agreement (all differences were less than 2%).

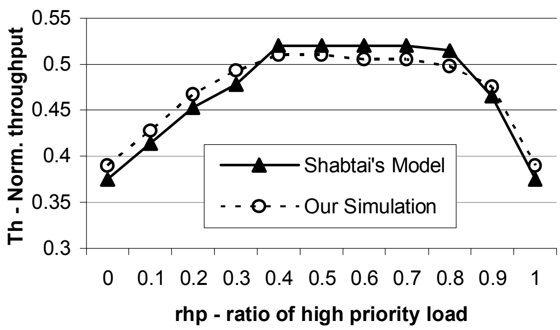

Figure 3. Total normalized throughput of a dual-priority, single-buffered, 6-stage MIN

On the other hand, the priority mechanism was tested under uniform traffic conditions; this was done by setting the parameter f_{hs} =0. We compared our measurements against those obtained from Shabtai's Model reported in [12], and have found that both results are in close agreement (the maximum difference was only 3.8%).

Figure 3 illustrates this comparison, involving the *total normalized throughput* for all packets (both high and low priority) of a dual-priority, single-buffered, 6-stage MIN vs. the *ratio of high priority packets* under full *offered load.*

B. 2-Class Priority Single-Layer MINs Performance under Hotspot Environment

In previous studies [25] we examined the performance of MINs natively supporting two priorities, when these operate under hotspot traffic conditions at single-layer architecture. We found that when the hotspot conditions were not extreme and the high priority packet ratio was moderate (r_{hp} =0.20), high priority packets received almost optimal quality of service, whereas the QoS offered to low priority packets varied, depending on the *zone* they were addressed to (figure 4). Another interesting finding was that while *normalized throughput* for some *zones* was found to be identical, the same *zones* exhibited variations of behavior regarding the *normalized delay* metric (figure 5). In all cases, performance indicators of low-priority packets for *zones* that were “close” to the hotspot output appeared to quickly deteriorate even for light loads ($\lambda \geq 0.3$), whereas low-priority packets addressed to *zones* “far” from the hotspot output exhibited a performance similar to that of MINs under uniform input load.

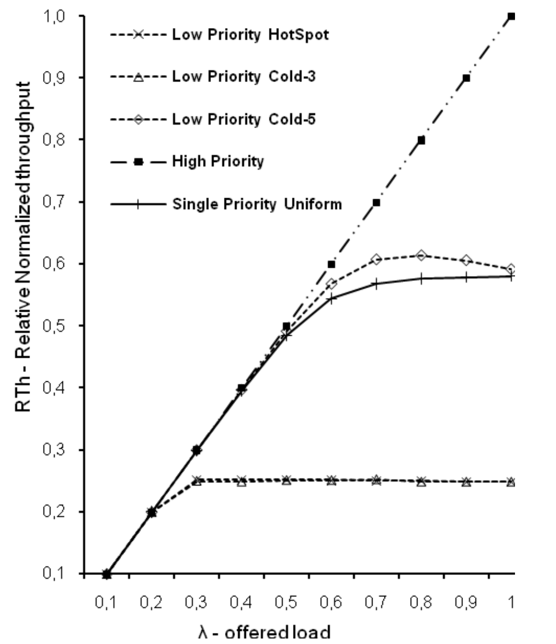

Figure 4. Relative normalized throughput of a dual-priority, double-buffered, 6-stage MIN under hotspot traffic at a single-layer architecture

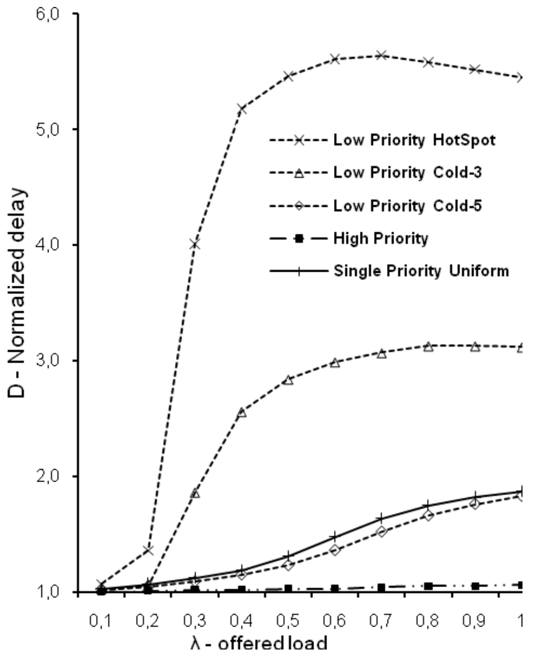

Figure 5. Normalized delay of a dual-priority, double-buffered, 6-stage MIN under hotspot traffic at a single-layer architecture

C. 2-Class Priority Multi-Layer MINs Performance under Hotspot Environment

In this paper, we extend previous studies by introducing a multi-layer architecture for a 6-stage multi-layer MIN, where *the number of layers* at the last stage is equal to l=4, i.e. the first four stages are single-layer and multiple layers are only used at the last two stages, in an attempt to balance between MIN performance and cost. It is also worth noting that for the first 4 stages, double-buffered SEs are considered, whereas at the last two stages (which are non-blocking), single-buffered SEs are used, as the absence of blockings removes the need for larger buffers.

Figures 6 and 7 depict the *relative normalized throughput* and *normalized delay* metric respectively for a dual-priority, double-buffered, 6-stage, multi-layer MIN versus a corresponding single-layer one, when the initial hotspot traffic is set to f_{hs}=0.05, while the *ratio of high priority packets* is considered to be r_{hp}=0.20. Curves represent the performance of low priority traffic for single hotspot output port and Cold-5 *zone*, as well as the performance of high priority traffic, routed to all output zones, since our experiments have shown that this parameter is not affected by the forwarding *zone* of such packets. According to figure 6, the *relative normalized throughput* of hotspot traffic for multi-layer MIN is found to be dramatically improved in comparison to the single-layer one. *Relative normalized throughput* reaches its peak performance RTh_{hs}=0.575 when the *offered load* is λ=0.8 -throughput gain 130%-, indicating that the additional bandwidth offered by the multi-layer SEs is exploited to a great extent. It is also noticed that the throughput gain for Cold-5 *zone* is considerable, i.e. 17.3% under full load traffic, while the performance of high priority packets remains optimal.

In figure 6 we can observe that high-priority packets are serviced optimally, both in the single- and the multi-layer case. This is expected, since the MIN gives precedence to servicing high-priority packets, and there is always amble bandwidth to serve all high-priority packets appearing at the inputs. Regarding the throughput of low-priority packets addressed to the hotspot output, we can observe that the single-layer MIN is quickly saturated (at *offered load* λ=0.3), while the multi-layer MIN, exploits to a very good extent the additional switching capacity, reaching its saturation point much later, at *offered load* λ=0.8. Beyond this point we observe a small drop in the hotspot output throughput, owing to the increased number of high-priority packets, which consume MIN bandwidth in the expense of the quality of service offered to low priority packets. Finally, regarding the throughput of the Cold-5 zone, the single-layer MIN exhibits approximately the same throughput with the multi-layer one at *offered loads* $\lambda\leq$0.6, where the MIN has amble power to route packets and the contention between packets is low. This can also be concluded from the fact that at this range the low-priority Cold-5 throughput increases almost linearly with the offered load. Beyond that point, the number of contentions between packets increases, but in the multi-layer MIN contentions are limited to the four initial stages only, and are thus smaller in number than the contentions in the single-layer MIN case where contentions may occur at any stage; this explains the performance gains exhibited for the multi-layer MIN for *offered loads* $\lambda\geq$0.7.

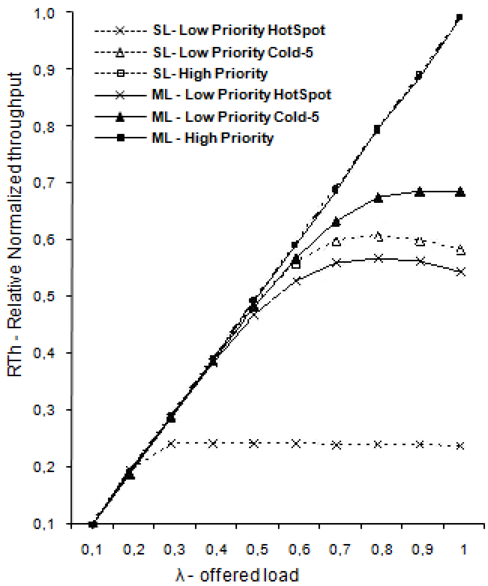

Figure 6. Relative normalized throughput of a dual-priority, double-buffered, 6-stage MIN under hotspot traffic at a multi-layer architecture

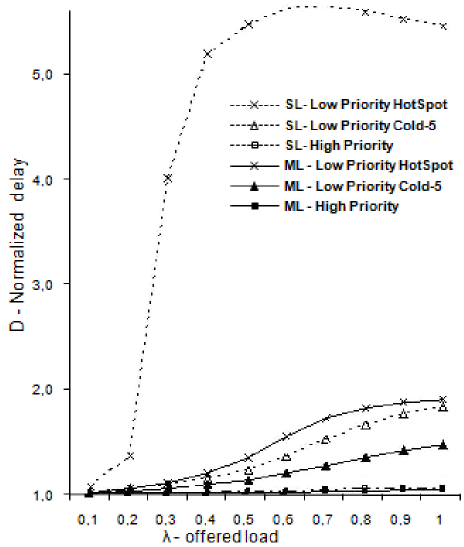

Figure 7. Normalized delay of a dual-priority, double-buffered, 6-stage MIN under hotspot traffic at a multi-layer architecture

Although Multistage Interconnection Networks (MINs) are fairly flexible in handling varieties of traffic loads, their performance considerably degrades by hotspot traffic, especially at increasing network sizes. Packet prioritization, through a scheme that natively supports dual-priority traffic, is a solution for providing better QoS to packets designated as "high priority". It was noticed that both performance metrics for high priority packets *relative normalized throughput* and *normalized delay* approached their optimal values Th_{hp}^{max}=1 and D_{hp}^{min}=1 respectively under r_{hp}=0.20 *ratio of high priority packets*. The rationale behind using multiple layers at the last two stages is to improve as well as the performance of low priority packets. Thus, in an attempt to balance between MIN performance and cost, in a 4-layer MIN configuration, we found again the second major performance metric, namely *normalized delay* to be dramatically improved in terms of hotspot traffic (figure 7); the peak value of this metric was reduced from the value D_{hs}=5.64 to D_{hs}=1.9. Finally, it is also observed that the decrement of *normalized delay* for low priority packets of Cold-5 *zone* is also considerable - e.g. 20% under full load traffic.

Regarding low-priority traffic routed to the hotspot output, we can observe a rapid increase at load λ=0.25, where the paths to the hotspot output become saturated. The increase in the delay becomes less sharp beyond the point of offered load λ=0.5, but we have to note that beyond the point of λ=0.4, a big number of blockings occurs at the network inputs (because the buffers at stage 1 are full), therefore less low-priority packets enter the network and are serviced. The multi-layer MIN exhibits significantly better performance, since it avoids blockings at the last two stages.

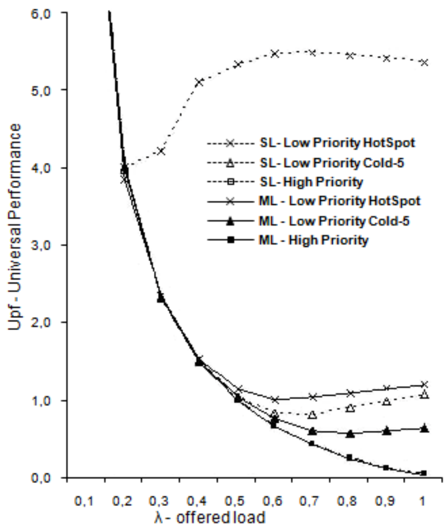

Figure 8. Universal performance metric of a dual-priority, double-buffered, 6-stage MIN under hotspot traffic at a multi-layer architecture

Figure 8 depicts the behavior of the *universal performance factor* of dual-priority multi-layer MINs vs. single-layer one, under hotspot traffic conditions. We notice that the introduction of multi-layer architecture greatly improves the performance of the MIN under hotspot traffic, with performance gains ranging from 94% (Cold-5, low-priority at full input load) to over 500% (hotspot, low-priority at high input load). These gains do not affect the quality of service offered to high priority packets, with this service remaining close to the optimal value of zero under full *offered load*. Note that the universal performance metric for low-priority packets at high loads appears poor, mainly due to the increased delay that these packets exhibit at the specific load range.

While the analysis above clearly shows that the multi-layer MIN has a clear performance advantage against its single-layer counterpart, the adoption of the multi-layer MIN is associated with increased costs, therefore network designers should carefully balance between the elevated switching capacity offered by the multi-layer MIN and the increased network deployment costs, to achieve the best cost/performance ratio. If we consider the 6x6 multi-layer MIN studied in this paper, with four layers at the final stage, it consists of 320 SEs in overall (4 layers * 32 SEs/layer =128 SEs for the final stage + 2 layers * 32 SEs/layer = 64 SEs for the 5th stage + 4 stages * 32 SEs/stage =128 SEs), an increase of 66% as compared with the 192 SEs needed for the implementation of a single-layer 6x6 MIN (6 stages * 32 SEs/stage =192 SEs).

Figure 9 depicts the performance analysis results for hotspot traffic, considering different number of layers at the final stage of the MIN. Effectively, the curve for l=1 corresponds to the single layer MIN, the curve for l=4 pertains to the MIN with four layers at the final stage studied in this paper and the curve for l=2 corresponds to an intermediate solution, limiting the number of layers at the last stage to lessen the infrastructure implementation cost. From these results, we can observe that for *offered loads* $\lambda \leq 0.4$ the MIN with l=2 has adequate switching power to service packets, and the introduction of more layers at the final stage offers no performance enhancement. For *offered load* λ=0.5, the performance gains of the MIN with l=4 against the MIN with l=2 are limited to 6.45%, which may not justify the 42.85% increment in the required SEs (the MIN with l=2 requires 2 layers * 32 SEs/layer =64 SEs for the final stage + 5 stages * 32 SEs/stage = 160 SEs, summing up to 224 SEs in overall). For loads $\lambda \geq 0.6$, the performance gains of the MIN with l=4 against the MIN with l=2 range between 18.1% (λ=0.6) and 35.7% (λ=0.9), therefore the infrastructure designers may opt for bearing the increased cost of adding more layers to favor performance.

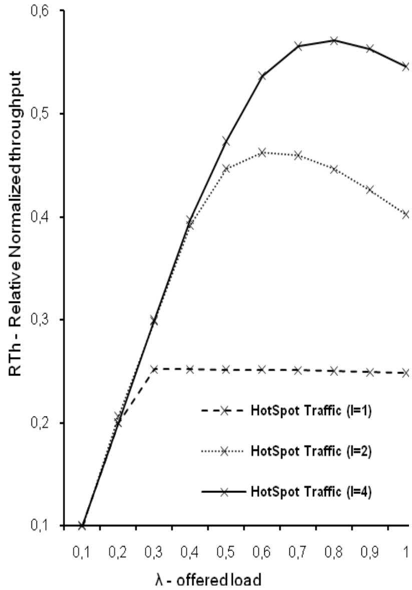

Figure 9. Relative normalized throughput for different numbers of layers at the final stage of the MIN.

V. CONCLUSIONS

In this paper we have examined the introduction of multi-layer architecture as a solution to the problem of performance degradation due to hotspot traffic under the presence of a dual-priority scheme. Since multi-layer architectures are associated with high costs, we have limited the multi-layer portion of the network to the final two stages (over a total of six stages), balancing thus between performance and cost. Performance gains were found to be considerable, both in terms of *throughput* and *delay*, with higher gains observed for the outputs "near" the hotspot.

Future work will include further experimentation with operating parameters of the MIN, including the overall network size, the high/low priority packet ratio and the hotspot/normal traffic ratio. The introduction of an adaptive scheme, altering buffer allocation to different priority classes according to current traffic load and high/low priority ratios will be investigated as well.

VI. REFERENCES

[1] G. B. Adams and H. J. Siegel, "The extra stage cube: A fault-tolerant interconnection network for supersystems", IEEE Trans. on Computers, 31(4)5, pp. 443-454, May 1982.

[2] H.K. Chang. "Nonuniform memory reference of multistage interconnection networks". Computer Standards & Interfaces, pp. 221-227, 2004.

[3] G.F. Goke, G.J. Lipovski. "Banyan Networks for Partitioning Multiprocessor Systems" Procs. of 1st Annual Symposium on Computer Architecture, pp. 21-28, 1973.

[4] M. Ilyas and M. A. Syed. "An efficient multistage switching node architecture for broadband ISDNs", Telecommunication Systems, pp. 229-241, 1998.

[5] J. Kim, T. Shin, and M. Yang. "Analytical modeling of a Multistage Interconnection Network with Buffered axa Switches under Hot-spot Environment", Procs. of PACRIM'07.

[6] D.A. Lawrie. "Access and alignment of data in an array processor", IEEE Transactions on Computers, vol. 24 (12), Dec. 1975.

[7] B.M. Maggs, "Randomly-wired multistage networks", Statistical Science vol. 8(1), pp. 70-75, 1993.

[8] H. Mun and H.Y. Youn. "Performance analysis of finite buffered multistage interconnection networks", IEEE Transactions on Computers, pp. 153-161, 1994.

[9] J. Park and H. Yoon. "Cost-effective algorithms for multicast connection in ATM switches based on self-routing multistage networks", Computer Communications, vol. 21, pp. 54–64, 1998.

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

[11] M. Saleh, M. Atiquzzaman. Analysis of shared buffer multistage networks with hot spot. IEEE First International Conference on Algorithms and Architectures for Parallel Processing, vol. 2, pp. 799-808, 1995.

[12] G. Shabtai, I. Cidon, and M. Sidi, "Two priority buffered multistage interconnection networks", Journal of High Speed Networks, pp.131-155, 2006

[13] G. Shabtai, I. Cidon, and M. Sidi, "Two Priority Buffered Multistage Interconnection Networks", IEEE High Performance Switching and Routing Conference HPSR'04, pp.75-79, 2004

[14] W.R. Stevens, "TCP/IP Illustrated", vol. 1. The protocols, (10th Ed), Addison-Wesley Pub Company, 1997.

[15] 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.

[16] D. Tutsch, G.Hommel. “Comparing Switch and Buffer Sizes of Multistage Interconnection Networks in Case of Multicast Traffic”, Procs. of the High Performance Computing Symposium, (HPC 2002); San Diego, SCS, pp. 300-305, 2002.

[17] D. Tutsch and G. Hommel. “Multilayer Multistage Interconnection Networks”, Proceedings of 2003 Design, Analysis, and Simulation of Distributed Systems (DASD'03). Orlando, USA, pp. 155-162, 2003.

[18] D. Tutsch, M. Brenner. “MIN Simulate. A Multistage Interconnection Network Simulator” Procs. of 17th European Simulation Multiconference (ESM'03); Nottingham, SCS, pp. 211-216, 2003.

[19] 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.

[20] E. Upfal, S. Feleprin and M. Snir, “Randomized routing with shorter paths”, Proceedings of the 5th ACM Symposium on Parallel Systems, pp. 283-292, 1993.

[21] D.C. Vasiliadis, G.E. Rizos, and C. Vassilakis. “Performance Analysis of blocking Banyan Swithces”, Procs. of CISSE 06, December, 2006.

[22] D.C. Vasiliadis, G.E. Rizos, C. Vassilakis, and E.Glavas. “Performance evaluation of two-priority network schema for single-buffered Delta Network”, Procs. of IEEE PIMRC' 07, 2007.

[23] D.C. Vasiliadis, G.E. Rizos, C. Vassilakis. “Improving Performance of Finite-buffered Blocking Delta Networks with 2-class Priority Routing through Asymmetric-sized Buffer Queues”, Proceedings of the Fourth Advanced International Conference on Telecommunications AICT08, IEEE Press, 2008.

[24] D.C. Vasiliadis, G.E. Rizos, C. Vassilakis, E. Glavas. “Routing and Performance Analysis of Double-Buffered Omega Networks Supporting Multi-Class Priority Traffic”, Proceedings of International Conference on Systems and Networks Communications ICSNC08, IEEE Press, 2008.

[25] D.C. Vasiliadis, G.E. Rizos, and C. Vassilakis. “Routing and Performance Evaluation of Dual Priority Delta Networks under Hotspot Environment”, Proceedings of the First International Conference on Advances in Future Internet AFIN09, IEEE Press, pp. 24-30, 2009.

[26] Wikipedia, IEEE 802.1p. http://en.wikipedia.org/wiki/IEEE_802.1p

[27] J. Garofalakis, E. Stergiou, “An approximate analytical performance model for multistage interconnection networks with backpressure blocking mechanism”, Journal of Communications (JCM), Academy, vol. 5, no 3, March 2010, pp. 247-261.

[28] J. Garofalakis, and E. Stergiou “An analytical performance model for multistage interconnection networks with blocking”, Proceedings of Communication Networks and Services Research Conference CNSR 2008, IEEE Press, May 2008.

[29] S. Chakraborty, S. Künzli, L. Thiele, A. Herkersdorf, P.Sagmeister. Performance evaluation of network processor architectures: combining simulation with analytical estimation. Computer Networks vol. 41(5) April 2003, pp. 641-665.

[30] T. Issariyakul, E. Hossain. Introduction to Network Simulator NS2. Springer, 2008. ISBN-13: 978-0387717593.

[31] H. Gu, Z. Qiu, Z. Liu, G. Kang, K. Wang, F. Hong. Choice of Inner Switching Mechanisms in Terabit Router. Proceedings of International Conference on Networking (ICN) 2005, P. Lorenz and P. Dini (Eds.), LNCS 3420, pp. 826–833.

[32] D. Holman, D. Lee. A Survey of Routing Techniques in Store-and-Forward and Wormhole Interconnects. Sandia National Laboratories Albuquerque, New Mexico 87185 and Livermore, California 94550, 2008.

[33] J. Duato, S. Yalamanchili, L. Ni. Interconnection networks: an engineering approach. Morgan Kaufmann Publishers, 2003. ISBN: 1-55860-852-4.