Performance Modelling of the Multicast Balanced Gamma Switch

Cheng Li[†], *Member, IEEE,* R. Venkatesan[†], *Senior Member, IEEE,* and H. M. Heys[†], *Member, IEEE*
Faculty of Engineering and Applied Science
Memorial University of Newfoundland
St. John's, NL, A1B 3X5, Canada

Abstract— **This paper presents an analytical model for the performance analysis of a new cell-based multicast switch for broadband communications. Using distributed control and a modular design, the Balanced Gamma switch features a high performance for unicast, multicast and combined traffic under both random and bursty conditions. Although it has buffers on input and output ports, the multicast BG switch follows predominantly an output-buffered architecture. The analytical model follows the three phase switching operation. The performance is evaluated under multicast random traffic in terms of cell loss ratio and cell delay. Performance under bursty traffic is studied through simulation and the results are compared to those of an ideal pure output-buffered multicast switch.**

Index Terms— **Multicast, Balanced Gamma (BG) switch, performance analysis, analytical modelling, cell loss ratio, cell delay.**

I. Introduction

The fast growing Internet and multimedia applications over the past two decades have generated huge demand for network capacity and service provisioning. In addition to the high speed requirement, many real-time applications, such as video-conferencing, remote diagnostic, and music/video on demand require messages to be sent to more than one destination. As a result, supporting multicast has become a necessary requirement for any switch designed for future broadband communication networks.

Many possible multicast switch architectures have been explored since the late 1980s, such as the PINIUM switch [1], the ABACUS switch [2], and the input-queued (virtual output queued) switch [3]. Due to the many desirable features such as self-routing, distributed control, modularity, constant delay for all input-output pairs and suitability for VLSI implementation, multistage interconnection network (MIN) design has become an attractive solution for broadband switch architecture. A multicast switch fabric using the implicit cell replication is preferred because it combines the routing and replication functions into a single unified network. Since cell replication is performed only when necessary as the cell is routed through, the load that multicast cells bring to the switch can be minimized, especially for the early stages. This kind of design will inherit most of the attractive features of the MIN design.

ATM-like fixed-size packet switching attracts much interest because of its application in high-speed Internet routers and switches. The switch fabric internally operates on the fixed-size packets called cells. The incoming variable-size IP packet (datagram) is internally segmented into cells that are transmitted to the output port, where they are reassembled into the IP datagram. In this paper, we use the term cell to identify the fixed-size packet used in the switch, which can be ATM cells, or any other convenient data format [4].

In this paper, we study the performance of a new cell-based multicast switch architecture that has input and output buffers, a backpressure mechanism and a very high throughput. The switch is called the multicast Balanced Gamma (BG) switch and utilizes a multi-path MIN design. Cell replication to achieve multicast is integrated into the functionality of each switch element and is performed in a distributed fashion along with the routing.

II. Switch Architecture

A. Switch Architecture

The multicast Balanced Gamma network utilizes a 4×4 switch element (SE) as the basic building block. Figure 1 shows the architecture of an 8×8 multicast BG switch. The basic architecture of an $N \times N$ BG multicast switch consists of N input port controllers (IPCs), an $N \times N$ multistage interconnected switch fabric that supports self-routing, self-replication and delivery acknowledgement, and N output port controllers (OPCs). No dedicated copy network to support the replication functionality is required. The IPC terminates the input signals from the network, strips the information contained in the cell header, and uses a lookup table to determine the destinations. The switch fabric is the core of the multicast BG switch. An $N \times N$ BG switch fabric consists of $n+1$ stages, where $n = log_2 N$, with each stage consisting of N SEs numbered from 0 to $N-1$. In stage 0, 1×2 SEs are used and in stage 1, 2×4 SEs are used. Each of the following $n-2$ stages is comprised of 4×4 SEs. The last stage is the output buffer stage, which can accept up to 4 cells per output port in one switching cycle. Network bandwidth is expanded through the first two stages and then remains the same for all subsequent stages. Through internal bandwidth expansion, the multicast BG switch can achieve better performance while keeping the hardware complexity reasonable. The OPC includes a regulator and scheduler. It updates each arrived cell with a new cell header and sends onto the output link. Details on the architecture design and the design choices justification can be found in [5], [6].

†. This work is supported in part by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

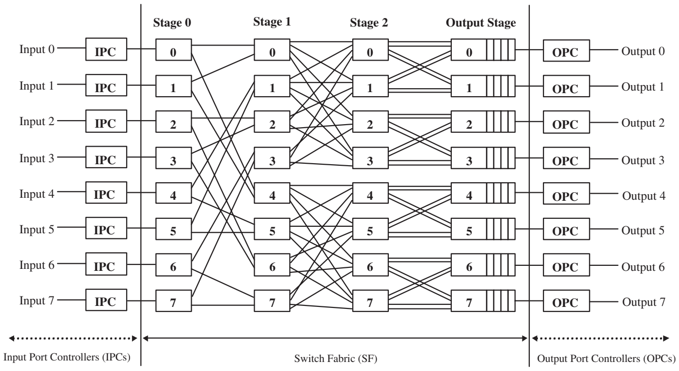

Fig. 1. The architecture of an 8 × 8 multicast BG switch

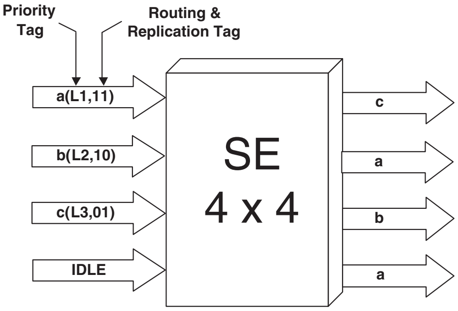

Fig. 2. Self-routing and cell replication in the 4 × 4 SE.

B. Self-Routing and Self-Replication Algorithm

In the multicast BG switch, there are three types of SEs: the 1×2 and 2×4 SEs are used for the first two expanding stages while the 4×4 SEs are used for all subsequent stages. Functionally, the first two types of SEs can be treated as a special (simpler) case of the 4×4 SE. Therefore, in the following discussion, the more general 4×4 SE is used, which is shown in Figure 2. The four output links are numbered 0 to 3 from top to bottom. Among the four links, link 0 and link 2 are called upper and lower regular links, respectively, while link 1 and link 3 are called upper and lower alternate links, respectively. Both the regular link and its alternate link have the same capability of reaching the same destination. Upon switching, the regular links are always used first. The alternate link is used only when the regular link has already been assigned to a connection.

In the multicast environment, tag design becomes more challenging because not only the routing information should be carried but also the cell replication information, and the tag length should be minimized to minimize the delay in the reservation phase. In the BG switch, for each SE to make the right routing and replication decision, a 2-bit tag is used by each SE for each input link. Four different actions can be taken by the SE and these are summarized in Table I.

Bit 1	Bit 0	Routing Action	Replication Action
0	0	Idle (no action)	Idle (no action)
0	1	Lower link	No replication
1	0	Upper link	No replication
1	1	Both links	Replication

TABLE I
ROUTING AND REPLICATION ACTIONS BASED ON TAG PAIR INFORMATION.

Priority switching is a feature considered in the multicast BG switch, with up to 8 priority levels currently supported. The SE will make its decision in two steps. Firstly, the SE decides the processing order of incoming cells based on the priority level associated with each cell. Secondly, incoming cells are switched following the order determined in the first step. Cells with higher priority are always processed first until all incoming cells are processed or all the sources are used up. In the latter case, the remaining low priority cells will be blocked. An example is provided in Figure 2 in which cells are coming in from the top three input links. By sorting on the priority tag, the process order is $c \rightarrow b \rightarrow a$. Following the routing and replication table, cell c is a unicast cell which requests an upper output link, it is switched to output 0 and similarly cell b is switched to output 2. For cell a, the tag bit pair '11' indicates that replication is required. The available outputs are checked and cell a is replicated and sent to both upper and lower alternative output links, links 1 and 3, respectively.

III. MULTICAST TRAFFIC MODEL

An important part of any performance analysis is an accurate traffic model which will be used to generate traffic for both simulation and analytical purposes. The multicast traffic model can be described by three random processes: the arrival process, the fanout process, and the destination selection process. The arrival process specifies the correlation among the successive cells. It can be random or bursty. For random traffic, the cell arrival is randomly selected based on the link load and is independent of cell arrival during the previous switching cycle. For bursty traffic, the ON-OFF model is used [1], [7]. The source generates cells in a bursty manner: one active period (ON period) followed by an idle period (OFF period). The durations of ON and OFF periods are independently evaluated from two geometric distributions with the period length L in cells derived from

$$L = 1 + \left\lceil \frac{\ln(1-R)}{\ln(1-p)} - 1 \right\rceil , \tag{1}$$

where $R,\ 0 \leq R < 1$, is the random number generated, and $p,\ 0 < p < 1$, is the reciprocal of the average period length in cells. The cells arriving at each input line in a burst have the same fanout number and are destined to the same output ports.

The fanout process describes the fanout distribution of a multicast cell, i.e., the distribution of the number of copies of an incoming cell. The truncated geometric distribution is used to model the fanout distribution [1], [7]. Given a switch size N, parameter q can be calculated numerically for any given mean fanout $\overline{F}$ following the equation

$$\overline{F} = \sum_{i=1}^{N} i \times \frac{(1-q) \times q^{i-1}}{1-q^N} = \frac{1}{1-q} - \frac{N \times q^N}{1-q^N}. \tag{2}$$

With parameter q, the probability of having a fanout value f, denoted by $P_{tg}(f)$, can be calculated by using

$$P_{tg}(f) = \begin{cases} \frac{(1-q)\times q^{f-1}}{1-q^N} & \text{for } 1 \leq f \leq N \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

The destination selection process describes how cell destination will be selected. In this paper, cell destinations are considered to be uniformly distributed. Under uniform destination selection, all output ports are equally likely to be requested. Therefore, the input and output load of the switch can be represented by the load of each of the input and output links, denoted by ρ_{in} and ρ_{out}. Given an ideal strict-sense non-blocking switch fabric, for multicast traffic, when cell replication occurs inside the switch fabric, the offered load ρ_{out} can be associated with ρ_{in} via the mean traffic fanout $\overline{F}$ by using

$$\rho_{out} = \overline{F} \times \rho_{in}. \tag{4}$$

From basic queuing theory, a queue will become unstable when the data arrival rate is greater than the departure rate. For each output queue, the departure rate is assumed to be one cell per switching cycle. To avoid overflow, the offered load ρ_{out} should normally be kept below one. Even though the average load is kept below one, due to the statistical nature of the traffic, it is possible that the instantaneous load momentarily exceeds one. To accommodate different fanout situations and carry out reasonable comparison in multicast traffic, the offered load to the switch is defined at the switch output. The load is converted to the input load via the mean fanout $\overline{F}$. As long as the load used does not cause the output queue to overflow, it is guaranteed that there is no overflow problem at the BG switch input. Unless otherwise stated, the traffic load reported in the paper refers to the offered output load.

IV. Analytical Modelling

Analytical modelling provides a good method to validate simulation results using theoretically calculated values. In this section, the performance model of the BG switch under multicast random traffic is studied. Generally speaking, the analysis follows the three-phase switching operation of the multicast BG switch. Firstly, the cell blocking probabilities at SEs of different stages are analyzed. With this information, the cell blocking probability for the whole switch fabric and the traffic arrival probability on the four links feeding each output queue are obtained. Secondly, the output queue is analyzed using a discrete-time Markov chain. The cell blocking probability, queue occupancy and queueing delay can be obtained through the output queue analysis. Then, the overall cell blocking probability for the combined switch fabric and output queue is calculated, which is also the probability of cells being kept in the head-of-the-line (HOL) position of the input queue. Finally, the input queueing analysis is performed to get the cell loss probability and other performance measures. The following conditions are assumed for the switch and the traffic:

1) The input queue, switch fabric, and output queue operate independently.
2) Cell arrival is identically and independently distributed (i.i.d) among all input links.
3) The arrivals of incoming cells follow a Bernoulli distribution with probability ρ on each input link.
4) Incoming multicast traffic has a mean fanout of $\overline{F}$.
5) Cell Destinations are uniformly distributed.
6) All cells are of the same priority.
7) All SEs within the switch fabric operate independently.
8) Cell arrival only occurs at the beginning of each cycle.
9) Cells are served on the first-come-first-served (FCFS) discipline in the input and output queues.

The following notation is introduced for the analysis:

$\rho =$ Offered traffic load at the switch input.
$p_{ir} =$ Pr{Regular input to SE at Stage i is active}.
$p_{ia} =$ Pr{Alternative input to SE at Stage i is active}.
$p_a(i) =$ Pr{Upper regular output at stage i is active}.
$p_b(i) =$ Pr{Upper alternative output at stage i is active}.
$p_c(i) =$ Pr{Lower regular output at stage i is active}.
$p_d(i) =$ Pr{Lower alternative output at stage i is active}.
$k_i =$ Mean fanout factor for Stage i, where $1 \leq k_i \leq 2$.
$P_{blk_i} =$ Pr{Cell being blocked in Stage i}.
$a_j =$ Pr{j incoming cells request the output link group}.

A. Analysis of Stage 0 (1 × 2 SEs)

The inputs to the switch fabric are connected directly to the input links of SEs in Stage 0. The load to the input of each 1 × 2 SE, p, is equal to the offered load to the switch fabric

$$p = \rho. \tag{5}$$

The bandwidth is expanded through this stage. Because of the random traffic assumption, both output links have the same probability of being requested by the incoming cell. Thus,

$$p_a(0) = p_b(0) = p \cdot k_0/2. \tag{6}$$

It is proved that Stage 0 is a non-blocking stage [8]. Hence,

$$P_{blk_0} = 0. \tag{7}$$

B. Analysis of Stage 1 (2 × 4 SEs)

Due to the random traffic assumption, all links between Stage 0 and Stage 1 have the same probability to be active. Therefore, instead of distinguishing the load on the two input links to each SE in Stage 1 as $p_a(0)$ and $p_b(0)$, $p(0)$ is used. The input link load of the 2×4 SE in Stage 1 is equal to the output link load of the 1×2 SE in Stage 0. Hence,

$$p(0) = p_{0a} = p_{0b} = p_a(0) = p_b(0). \tag{8}$$

As described in Section II, for all 2×4 and 4×4 SEs, because each regular link and its alternative link have the same capability of delivering cells to their destination, they are represented using the name "link group". Therefore, the four output links are divided into two link groups, the upper link group and the lower link group. The regular link in each link group is always used first when there is a request. The alternative link is used only when the regular link is occupied and there is another cell request. Because any incoming active cell will not request both links within the same link group and the two link groups are equally likely to be selected due to the random traffic condition, for the 2 × 4 SE, the regular link will be active when one or both of the two input links are active, and the alternative link will be active only when both of the input links are active. Therefore, the probability that two regular output links are active is given by

$$\begin{aligned} p_a(1) &= p_c(1) \\ &= p(0) \cdot k_1 \cdot \left(1 - \frac{p(0) \cdot k_1}{2}\right) + \left(\frac{p(0) \cdot k_1}{2}\right)^2 \end{aligned} \quad (9)$$

As well, the two alternative links $p_b(1)$ and $p_d(1)$ have the same probability to be active and the probability is:

$$p_b(1) = p_d(1) = \left(\frac{p(0) \cdot k_1}{2}\right)^2. \quad (10)$$

Again, Stage 1 is a non-blocking stage [8]. Therefore,

$$P_{blk_1} = 0. \quad (11)$$

C. Analysis of Stage 2 to Stage n - 1 (4 × 4 SEs)

From the interconnection pattern between stages, it is not difficult to find that among the four input links of each 4 × 4 SE, two of them are from the regular output links of the previous stage and the other two are from the alternative output links. The random traffic assumption ensures that the load for both regular links of the same stage is the same; similarly, the load for all alternative links of the same stage is also the same, i.e.,

$$p_{ir} = p_a(i-1) = p_c(i-1), \quad (12)$$

$$p_{ia} = p_b(i-1) = p_d(i-1). \quad (13)$$

Similar to the analysis for the 2 × 4 SEs, the probability of the upper link group being requested by a cell is equal to that of the lower link group under random traffic. Therefore, either of those two link groups can be chosen as the targeted link group for analysis. The probability that there are i cells requesting the targeted link group can be calculated as:

1. Probability of 0 cell requesting the targeted link group:

$$a_0 = (1 - \frac{p_{ia} \cdot k_i}{2})^2 \cdot (1 - \frac{p_{ir} \cdot k_i}{2})^2 \quad (14)$$

2. Probability of 1 cell requesting the targeted link group:

$$\begin{aligned} a_1 &= p_{ia} \cdot k_i \cdot (1 - \frac{p_{ia} \cdot k_i}{2}) \cdot (1 - \frac{p_{ir} \cdot k_i}{2})^2 \\ &+ p_{ir} \cdot k_i \cdot (1 - \frac{p_{ir} \cdot k_i}{2}) \cdot (1 - \frac{p_{ia} \cdot k_i}{2})^2 \end{aligned} \quad (15)$$

3. Probability of 2 cells requesting the targeted link group:

$$\begin{aligned} a_2 &= p_{ia} \cdot p_{ir} \cdot k_i^2 \cdot (1 - \frac{p_{ia} \cdot k_i}{2}) \cdot (1 - \frac{p_{ir} \cdot k_i}{2}) \\ &+ \frac{1}{4} \cdot p_{ia}^2 \cdot k_i^2 \cdot (1 - \frac{p_{ir} \cdot k_i}{2})^2 \\ &+ \frac{1}{4} \cdot p_{ir}^2 \cdot k_i^2 \cdot (1 - \frac{p_{ia} \cdot k_i}{2})^2 \end{aligned} \quad (16)$$

4. Probability of 3 cells requesting the targeted link group:

$$\begin{aligned} a_3 &= \frac{1}{4} \cdot p_{ia}^2 \cdot p_{ir} \cdot k_i^3 \cdot (1 - \frac{p_{ir} \cdot k_i}{2}) \\ &+ \frac{1}{4} \cdot p_{ia} \cdot p_{ir}^2 \cdot k_i^3 \cdot (1 - \frac{p_{ia} \cdot k_i}{2}) \end{aligned} \quad (17)$$

5. Probability of 4 cells requesting the targeted link group:

$$a_4 = \frac{1}{16} \cdot p_{ia}^2 \cdot p_{ir}^2 \cdot k_i^4 \quad (18)$$

6. Probability of having 5 or more cells requesting the targeted link group:

$$a_j = 0 \qquad \text{for j} \geq 5 \quad (19)$$

Therefore, the probability that any regular output link (upper / lower) carries an active cell is given by

$$p_a(i) = p_c(i) = \sum_{j=1}^{4} a_j. \quad (20)$$

The probability for the alternative output link is given by:

$$p_b(i) = p_d(i) = \sum_{j=2}^{4} a_j. \quad (21)$$

Blocking occurs when more than two incoming cells request the same link group. The cell blocking probability for stage i, where $2 \leq i \leq n-1$, is given as:

$$\begin{aligned} P_{blk_i} &= \frac{\text{E\{Number of blocked copies\}}}{\text{E\{Number of copies\}}} \\ &= \frac{\sum_{j=3}^{4} (j-2) \cdot a_j}{\sum_{j=1}^{4} j \cdot a_j}. \end{aligned} \quad (22)$$

D. Switch Fabric Blocking Probability Analysis

It has been assumed that traffic toward all input ports of the switch is identically and independently distributed multicast traffic with a load of ρ and a mean fanout of $\overline{F}$. Let N_i, where $0 \leq i \leq n-1$, denote the average number of cells arriving at stage i. For an $N \times N$ switch fabric, the average number of active cells that arrive at stage 0, N_0, is

$$N_0 = N \cdot \rho. \quad (23)$$

Taking the fanout factor (k_0) and the probability that a cell is blocked at stage 0 (P_{blk_0}) into consideration, the average number of cells that will be sent to stage 1 is

$$N_1 = N_0 \cdot k_0 \cdot (1 - P_{blk_0}). \quad (24)$$

Similarly, a recursive relation can be established between any adjacent subsequent stages, i.e.,

$$N_{i+1} = N_i \cdot k_i \cdot (1 - P_{blk_i}), \qquad \text{for } 0 \leq i \leq n-1. \quad (25)$$

The average number of cells that manage to arrive at the output buffer stage, N_n, is given by

$$N_n = N \cdot \rho \cdot \prod_{i=0}^{n-1} k_i \cdot \prod_{i=0}^{n-1} (1 - P_{blk_i}). \quad (26)$$

The blocking probability inside the switch fabric for each incoming cell copy is then given by

$$P_{blk_{SF}} = \frac{N \cdot \rho \cdot \overline{F} - N \cdot \rho \cdot \prod_{i=0}^{n-1} k_i \cdot \prod_{i=0}^{n-1} (1 - P_{blk_i})}{N \cdot \rho \cdot \overline{F}}. \quad (27)$$

Because of

$$\overline{F} = \prod_{i=0}^{n-1} k_i, \tag{28}$$

the blocking probability inside the SF for each copy is

$$P_{blk_{SF}}(copy) = 1 - \prod_{i=0}^{n-1} (1 - P_{blk_i}). \tag{29}$$

E. Finite Output Queueing Analysis

Among the four links coming into each output buffer, two links are from the regular output of the switch element while the other two are from the alternative output links. Let $p_{(n-1)r}$ and $p_{(n-1)a}$ denote the loads on the regular and alternative output link from SEs in the last stage of the switch fabric, respectively. The traffic load arriving at the output queue, λ_{OQ}, is the sum of the traffic on all the four incoming links to an output queue, which is in the unit of cells per switching cycle:

$$\lambda_{OQ} = 2 \cdot (p_{(n-1)r} + p_{(n-1)a}). \tag{30}$$

For simplicity, we assume that the destinations for the copies of a multicast cell can be repeated. Similar to the previous analysis, the probability of i copies requesting an output queue, where $0 \leq i \leq 3$, can be approximated by

$$a_i = \left(\begin{array}{c} N \\ i \end{array} \right) \cdot (\frac{\rho \cdot \overline{F}}{N})^i \cdot (1 - \frac{\rho \cdot \overline{F}}{N})^{(N-i)}. \tag{31}$$

Under multicast random traffic, the probability of having more than four cells requesting the same output port is very low [5], [6]. Assuming that it is negligible, the probability of four cell arrivals to the output queue can be approximated by

$$a_4 = 1 - \sum_{j=0}^{3} a_j, \tag{32}$$

and the probability of more than four cell arrivals is zero.

The output queue analysis employed here is modified from the approach described in [9]. Letting Q_m denote the number of cells in the output queue at the end of the m^{th} switching cycle, A_m denote the number of cell arrivals during the m^{th} switching cycle, and B_o denote the output queue size, the control function of the output queue is:

$$Q_m = min\{max\{0, Q_{m-1} + A_m - 1\}, B_o\}. \tag{33}$$

When $Q_{m-1} = 0$ and $A_m > 0$, one of the arriving cells is immediately transmitted to the output link of the switch without experiencing any delay.

Similar to the infinite queue analysis described in [9], Q_m is modeled by a finite-state, discrete-time Markov chain, which is shown in Figure 3. The state transition probabilities $p_{_{ij}} \equiv Prob[Q_m = j | Q_{m-1} = i]$ are modified and given as:

$$p_{_{ij}} = \begin{cases} a_0 + a_1 & i = 0,\ j = 0 \\ a_0 & 1 \leq i \leq B_o,\ j = i - 1 \\ a_{j-i+1} & 1 \leq j \leq B_o - 1,\ 0 \leq i \leq j. \\ \sum\limits_{k=j-i+1}^{4} a_k & j = B_o,\ 0 \leq i \leq j \\ 0 & \text{otherwise} \end{cases} \tag{34}$$

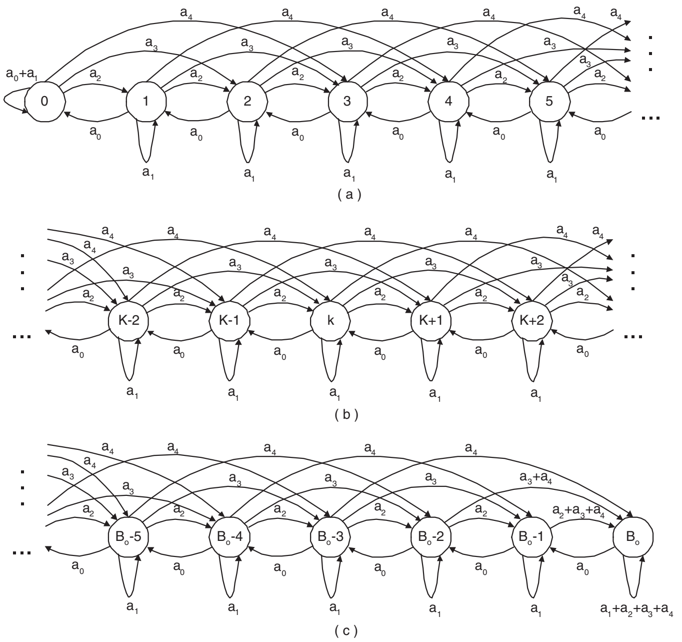

Fig. 3. The output queue state transition diagram.

Let π_i denote the probability of the output queue being in state i, where $0 \leq i \leq B_o$. By using the Markov chain balance equation [10], a recursive relation between the queue state probabilities is obtained:

$$\pi_1 = \frac{1 - a_0 - a_1}{a_0} \cdot \pi_0 \tag{35}$$

$$\pi_2 = \frac{1 - a_1}{a_0} \cdot \pi_1 - \frac{a_2}{a_0} \cdot \pi_0 \tag{36}$$

$$\pi_3 = \frac{1 - a_1}{a_0} \cdot \pi_2 - \frac{a_2}{a_0} \cdot \pi_1 - \frac{a_3}{a_0} \cdot \pi_0 \tag{37}$$

$$\pi_i = \frac{1 - a_1}{a_0} \cdot \pi_{i-1} - \sum_{k=2}^{4} \frac{a_k}{a_0} \cdot \pi_{i-k}, \quad \text{for } 4 \leq i \leq B_o \tag{38}$$

$$\text{where } \pi_0 = \frac{1}{1 + \sum\limits_{i=1}^{B_o} \pi_i / \pi_0}. \tag{39}$$

It is very difficult to get an analytical expression for the output queue state probability for an arbitrary queue size B_o. Therefore, a numerical analysis approach using Maple [11] was adopted. With all state probabilities known, the analysis of the output queue becomes straightforward.

The average number of cells in the output queue, $\overline{n}_{OQ}$, can be calculated through the sum of the products of each state of the output queue and the corresponding probability. That is,

$$\overline{n}_{OQ} = \sum_{i=0}^{B_o} i \cdot \pi_i. \tag{40}$$

The probability of a cell which manages to arrive at the output queue getting blocked is equal to the overflow probability of the output queue while there are cell arrivals, i.e.:

$$P_{blk_{OQ}} \quad = \quad \frac{\sum_{i=0}^{3}(\pi_{B_o-i} \cdot \sum_{j=i+1}^{4} j \cdot a_j)}{\sum_{i=1}^{4} i \cdot a_i}. \qquad (41)$$

By applying the well-known Little's formula, the average cell waiting time in the output queue, $\overline{T}_{OQ}$, is given by:

$$\overline{T}_{OQ} = \frac{\overline{n}_{OQ}}{\lambda_{OQ} \cdot (1 - P_{blk_{OQ}})}. \qquad (42)$$

F. Finite Input Queueing Analysis

By considering the blocking effect inside the switch fabric and at the output queue together, the blocking probability for a copy of the HOL cell at the input queue is

$$P_{blk_{SF\&OQ}}(copy) = 1 - (1 - P_{blk_{OQ}}) \cdot \prod_{i=0}^{n-1}(1 - P_{blk_i}). \quad (43)$$

For multicast traffic, each HOL cell will be removed only when all its copies are delivered, although each copy might be switched at different cycles. Considering each new cell is expected to carry $\overline{F}$ copies and assuming that the copies contained in a master cell are independent to each other, the probability that a multicast cell stays in the HOL position during the next switching cycle can be approximated by

$$\begin{aligned} P_{blk_{SF\&OQ}}(cell) &\approxeq 1 - [1 - P_{blk_{SF\&OQ}}(copy)]^{\overline{F}} \\ &\leq \overline{F} \cdot P_{blk_{SF\&OQ}}(copy). \end{aligned} \qquad (44)$$

The two extreme cases, i.e., the unicast case where the fanout factor k_i is 1 and the broadcast case where k_i is constant 2, constitute the best and worst blocking condition that the switch fabric will have. The corresponding two bounds can be used as the indication for the best and worst performances.

Let $P_{success}$ denote the probability that all copies of the HOL master cell are successfully switched during a switching cycle. The probability $P_{success}$ is given by

$$P_{success} = 1 - P_{blk_{SF\&OQ}}(cell). \qquad (45)$$

This is the probability that the HOL cell will be removed and the next cell in the input queue will become the HOL cell.

Let Q_m denote the number of cells in the input queue at the end of the m^{th} switching cycle, A_m and D_m denote the number of cell arrivals and departures during the m^{th} switching cycle, respectively, and B_{in} denote the input queue size. Both A_m and D_m must be 0 or 1. The control function of the input queue is given as [9]:

$$Q_m = min\{max(0, Q_{m-1} + A_m - D_m), B_{in}\}. \qquad (46)$$

Based on incoming traffic load p, given by (5), and the resulting cell blocking probability $q = P_{blk_{SF\&OQ}}(cell)$, the input queue can be modeled by a finite state discrete-time Markov chain as in Figure 4. Similar to the approach described in [9], the probability of different queue states π_i can be obtained from the Markov chain balance equations:

$$\pi_i = \frac{p \cdot q}{(1-p) \cdot (1-q)} \cdot \pi_{i-1} \qquad \text{for } 1 \leq i \leq B_{in} - 1, \quad (47)$$

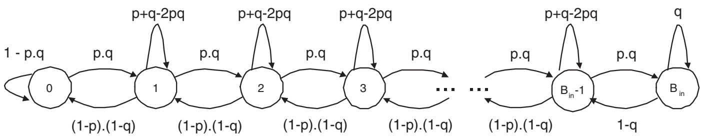

Fig. 4. The input queue state transition diagram.

$$\pi_{B_{in}} = \frac{p \cdot q}{1-q} \cdot \pi_{B_{in}-1}, \qquad (48)$$

and the sum of the probabilities for all states is

$$\sum_{i=0}^{B_{in}} \pi_i = 1. \qquad (49)$$

We define t as:

$$t = \frac{p \cdot q}{(1-p) \cdot (1-q)}. \qquad (50)$$

Substituting variables in (47) and (48) with variable t, and by solving the recursive equations, the probability of the input queue having zero cells, π_0, is given by

$$\pi_0 = \frac{1-t}{1 - p \cdot t^{B_{in}} - (1+p) \cdot t^{B_{in}+1}}, \qquad (51)$$

and other input queue state probabilities as:

$$\pi_i = \begin{cases} t^i \cdot \pi_0 & \text{for } 1 \leq i \leq B_{in} - 1 \\ (1-p) \cdot t^{B_{in}} \cdot \pi_0 & \text{for } i = B_{in}. \end{cases} \qquad (52)$$

Because of the use of backpressure algorithm, the cell loss probability for the switch fabric is determined by the overflow probability of the input queue, which is:

$$P_{loss} = (1-p) \cdot t^{B_{in}} \cdot \pi_0. \qquad (53)$$

With the state probability π_i, where $0 \leq i \leq B_{in}$, the average number of cells in the queue can be calculated by

$$\overline{n}_{in} = \sum_{i=0}^{B_{in}} i \cdot \pi_i, \qquad (54)$$

and the average cell delay in the input queue, $\overline{T}_{in}$, can be obtained by using Little's formula

$$\overline{T}_{in} = \frac{\overline{n}_{in}}{p \cdot (1 - P_{loss})}. \qquad (55)$$

Delay through the switch fabric is a constant. Assuming it is negligible, the average cell delay in the switch, $\overline{T}$, is given by:

$$\overline{T} = \overline{T}_{in} + \overline{T}_{OQ}. \qquad (56)$$

G. Constraints

To ensure a stable system for analysis, the system under study is constrained by the following conditions:

1. The mean fanout factor for each stage k_i, satisfies

$$1 \leq k_i \leq 2 \quad \text{for } 0 \leq i \leq n-1. \qquad (57)$$

2. The mean fanout of the multicast traffic, $\overline{F}$, satisfies

$$\overline{F} = \prod_{i=0}^{n-1} k_i. \qquad (58)$$

3. For multicast random traffic, we have

$$k_i = \sqrt[n]{\overline{F}} \quad \text{for } 0 \leq i \leq n-1. \tag{59}$$

4. The effective offered output load λ_{OQ} must be kept below one to ensure a stable output queueing system

$$\lambda_{OQ} = \rho \cdot \overline{F} = \rho \cdot \prod_{i=0}^{n-1} k_i < 1. \tag{60}$$

V. Performance Analysis

In this section, performance results from the analytical model are compared to those from simulation for the multicast BG switch under multicast random traffic. The loss and delay performance are examined for the 128×128 switch. Because ideal switch will have zero cell loss with enough output buffering, only the delay performance are compared to the ideal switch for various burstiness and fanout conditions under bursty traffic. Simulation results, which are obtained using the simulator developed by the authors and their students over the past ten years, are used for comparison. All simulation results provided are based on simulations that have run through a period of switching one billion cells.

A. Performance Under Multicast Random Traffic

1) Loss Performance: Because of the backpressure algorithm, a blocked cell is buffered in the input queue for further switching. Cell loss occurs only when the input queue is full and a new cell arrives. In that case, all copies implicitly contained in the new cell will be dropped. Therefore, the cell loss performance of the multicast BG switch is tightly associated with the size of the input queuing space, measured in cells. During this analysis, the output queue is assumed to have enough capacity to receive any cell appearing at its input. Therefore, the only reason for the HOL cell to be kept in the input queue is the internal blocking of the switch fabric.

Figure 5 shows the cell loss ratio versus the size of input queue for a 128×128 BG switch under 90% multicast random traffic with a mean fanout of 2. It is observed that the input buffering requirement is very low. With an input queue of six cell spaces, a cell loss ratio of around 10^{-8} can be achieved. The lower bound represents a best-case scenario which corresponds to unicast traffic that has the same input load, while the upper bound represents the worst-case condition which is from the result of broadcast traffic. The simulation results fit well between the two bounds from the analytical model. The simulation result is very close to but slightly better than the analytical approximation. It is obvious that with a lighter traffic load, even smaller queue sizes are sufficient to achieve the desired performance.

Fanout is the most important characteristic for multicast traffic. Even under the same load, multicast traffic loads with different fanouts behave differently. The analytical approximation and simulation results for different fanouts are plotted and compared in Figure 6 for heavy load situation, i.e., 90% offered load, to demonstrate the high performance of the BG switch. Once the performance under high load conditions are acceptable, with similar or even less buffering resources, it is guaranteed that the performance requirements under the low traffic load will be satisfied. The simulation and analytical model results are consistent for different loads and fanouts. The input buffering requirements are very low for different fanouts. With the same offered load, the larger the mean fanout, the better the loss performance of the BG switch. This is because the offered load is defined at switch output and multicast cell replication is done within the switch as late as required, thereby reducing load and blocking at the earlier switch stages.

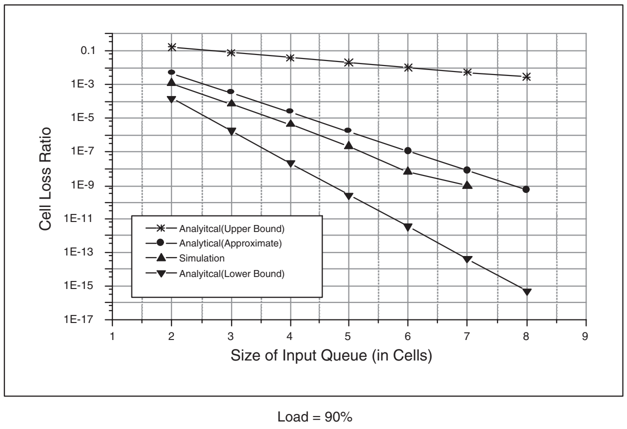

Fig. 5. Cell loss performance comparison under 90% load for 128×128 BG switch under multicast random traffic with mean fanout 2

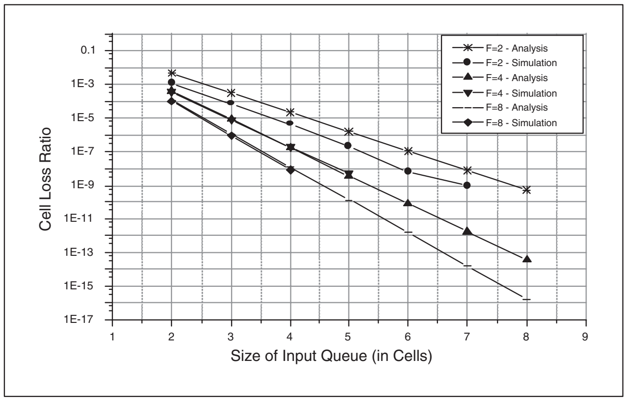

Fig. 6. Loss performance comparison between analytical approximation and simulation results under different fanout for 128×128 BG switch

2) Delay Performance: Because the SEs do not contain internal buffers, cells are delayed either at the input queue or at the output queue. The delay associated with the overhead transfer during the reservation phase, which is a constant value and applies to every cell, is not included. At the input queue, only the master cell is stored. Multiple destination requests are contained in the cell header. When reaching the output queue, each copy becomes an independent cell. Therefore, in delay performance analysis, the input queueing delay is measured in terms of the master cell while the output queueing delay and total delay are calculated based on an individual copy.

The average delay break down, measured in number of switching cycles, is presented in Table II for various switch sizes under 90% multicast random traffic with a mean fanout of 2. Enough buffering resources are provided at both input ports and output ports to ensure virtually no cell loss. The trends for different switch sizes are almost the same. The input queueing delay is much smaller than the output queueing delay due to the high switching capability of the switch. Most of the cells can be switched immediately without being buffered at the switch input. This also indicates that the input buffering requirement will be very small when compared to that of output buffering when a real switch is constructed using this architecture.

Switch Size	Simulation			Analysis		
	Total	Input	Output	Total	Input	Output
16×16	4.2379	0.0368	4.2031	4.1208	0.0133	4.1075
32×32	4.3731	0.0575	4.3356	4.2311	0.0185	4.2126
64×64	4.4655	0.0799	4.4067	4.2911	0.0234	4.2677
128×128	4.5199	0.1023	4.4376	4.3278	0.0282	4.2996
256×256	4.5503	0.1256	4.4548	4.3534	0.0329	4.3205
512×512	4.5824	0.1496	4.4628	4.3735	0.0376	4.3359

TABLE II

AVERAGE DELAY PERFORMANCE FOR 90% MULTICAST RANDOM TRAFFIC WITH A MEAN FANOUT OF 2 FOR VARIOUS SWITCH SIZES

Average Burst Length	Mean Fanout	BG Switch			Ideal Switch
		Total Delay	Input Delay	Output Delay	Total Delay
5	2	49.5026	0.6428	48.7291	48.8855
	4	49.3910	0.7787	48.3376	48.7632
	8	49.3592	0.9459	48.0373	48.3810
10	2	94.1450	1.2324	92.5919	93.0086
	4	94.3925	1.5435	92.1378	93.9522
	8	94.6992	1.9188	91.7577	93.6238
15	2	139.1580	1.8302	136.8090	137.3470
	4	139.3400	2.3427	135.8240	136.8180
	8	138.8480	2.9285	134.1840	137.2500

TABLE III

AVERAGE DELAY PERFORMANCE BREAKDOWN FOR 128×128 BG AND IDEAL MULTICAST SWITCH UNDER 90% MULTICAST BURST TRAFFIC

B. Performance Under Multicast Bursty Traffic

Performance under realistic traffic is difficult to model analytically. Simulation experiment is used. The ideal switch is used for comparison. Detailed analysis and comparison under bursty and non-uniform traffic conditions can be found in [6].

Table III presents the delay breakdown for the 128×128 BG switch under 90% load for various fanout and burstiness conditions. It is obvious that the output queueing delay is the dominant part for the BG switch and the change in burstiness affects the delay performance significantly while the impact of fanout is negligible. The burst length increase means the correlation between successive cells increases because all cells belonging to the same burst go to the same destinations. Even though in the long run, destination selection is uniformly distributed, on a cycle by cycle basis, the possibility of having more cells coming to the same output increases with longer bursts. Having more than two arrivals to the same output will cause output queue buildup. As a result, each cell will experience longer delay. At the same time, correlation will likely increase the chance of internal blocking. This will result in more blocked cells retained at the input queue, thus increases the input queueing delay. Even though the internal blocking becomes worse as the traffic burstiness increases, the switching capability of the multicast BG switch ensures that most of the cells manage to reach their destinations. Therefore, although the input queueing delay increases along with the traffic burst length, it is always a small fraction of the output queueing delay. In general, average cell delay in the BG switch is only marginally higher than that of the ideal switch.

VI. CONCLUSION

In this paper, we have presented the architecture and an analytical model for the performance of the multicast BG switch. The switch adopts a multipath MIN design. A distributed control and a modular architecture are used in the design to fulfill the high-speed requirement. No dedicated copy network is needed to support multicast switching. Switch performance from the analytical model was studied and compared to the simulation results under multicast random traffic. Performance under multicast bursty traffic was investigated through simulation and the results were compared to the ideal switch. It was shown that the simulation result matches the analytical model well and the BG switch maintains high performance under both random and bursty traffic conditions. It was also shown that the BG switch achieves a performance close to an ideal pure output-buffered architecture while keeping hardware complexity reasonable. The high performance and easy hardware implementation properties make the multicast BG switch a promising candidate in future high-speed packet switching networks.

REFERENCES

[1] K. L. E. Law and A. Leon-Garcia, “A Large Scalable ATM Multicast Switch,” *IEEE Journal on Selected Area in Communications*, vol. 15, pp. 844–854, July 1997.

[2] H.J. Chao, B.S. Choe, J.S. Park, and N. Uzun, “Design and Implementation of Abacus Switch: A Scalable Multicast ATM Switch,” *IEEE J. on Selected Area in Communications*, vol. 15, pp. 830–843, June 1997.

[3] M. A. Marsan, A. Bianco, P. Giaccone, E. Leonardi and F. Neri, “Multicast traffic in input-queued switches: optimal scheduling and maximum throughput,” *IEEE/ACM Transactions on Networking*, vol. 11, Issue 3, pp. 465–477, June 2003.

[4] M. A. Marsan, A. Bianco, P. Giaccone, E. Leonardi and F. Neri, “Packet Scheduling in Input-Queued Cell-Based Switches,” in *Proceedings of IEEE INFOCOM’2001*, vol. 2, pp. 1085-1094, Alaska, USA, April, 2001.

[5] C. Li, H. M. Heys and R. Venkatesan, “Design and Scalability of the Multicast Balanced Gamma (BG) Switch,” *Proceedings of the Eleventh International Conference on Computer Communications and Networks (IEEE ICCCN’2002)*, p.p. 518-521, Miami, Florida, USA, October 2002.

[6] Cheng Li, *Design, Modelling, and Analysis of the Balanced Gamma Multicast Switch for Broadband Communications.* PhD Dissertation, Memorial University of Newfoundland, October, 2004.

[7] S. H. Dyun and D. K. Sung, “A General Expansion Architecture for Large-Scale Multicast ATM Switches,” *IEICE Transactions on Communications*, vol. E80-B, pp. 1671–1679, November 1997.

[8] Harinath Sivakumar, *Performance, Fault Tolerance and Reliability of Multistage Interconnection Networks for Broadband Packet Switch Architectures*, Master’s thesis, Memorial University of Newfoundland, 1995.

[9] M. Hluchyj and M. Karol, “Queuing in High-performance Packet Switching,” *IEEE Journal on Selected Area in Communications*, vol. 6, pp. 1587–1597, December 1988.

[10] L. Kleinrock, *Queueing System.* New York: John Wiley & Sons, 1975.

[11] Waterloo Maple web site: http://www.maplesoft.com/main.shtml.