Janusz KLEBAN*
Piotr PIETRUSZEWSKI*

PERFORMANCE EVALUATION OF CRRD-OG PACKET DISPATCHING SCHEME UNDER NONUNIFORM TRAFFIC DISTRIBUTION PATTERNS

The three-stage Clos switching fabric has attractive scalability features thanks to a modular architecture. That makes it appealing as an alternative for high-performance, scalable switches and routers. To avoid internal blocking and output port contention in a Clos-network switch the scheduling and contention resolution schemes have to be employed. Algorithms which can assign a route between input and output modules are usually called packet dispatching schemes. This paper presents new results obtained for the CRRD-OG (Concurrent Round-Robin Dispatching with Open Grants) packet dispatching scheme under the nonuniform traffic distribution patterns in the MSM (Memory-Space-Memory) Clos switching fabric. The performance of the CRRD-OG is compared with results obtained for the CRRD, CMSD (Concurrent Master-Slave Round-Robin Dispatching) and the SRRD (Static Round-Robin Dispatching) schemes. We show via simulation that the CRRD-OG algorithm gives better performance results than other packet dispatching schemes.

Keywords: Clos-network, Dispatching Algorithm, Packet Switching, Packet Scheduling

1. INTRODUCTION

The continued growth of Internet Protocol-based service subscribers requires a much more robust, highly scalable core routers switches to handle the expected annual doubling of bandwidth in the United States and Europe and the expected tripling and possibly quadrupling of bandwidth in Asia. To meet these demands, service providers will need to deploy a new class of core routers that have taken a major leap forward in design. While the bandwidth of external connections on core routers has increased in recent years from STM-1 to STM-16 and STM-64, tomorrow’s core routers will need to support STM-256 connections operating at 40 Gbps. In addition, the number of line cards that the core router will need to

* Poznan University of Technology. Scientific work financed from science funding resources in the years 2005-2008 as a research project (grant 3T11D 003 29).
support will grow dramatically to handle the aggregate subscriber and backbone bandwidth growth. To meet these new demands, tomorrow’s router architectures will have to function very differently from those of today. They will require distributed memories and multi-stage switching fabrics that replace single-stage crossbars, allowing extraordinary scalability.

The high-performance switches internally operate on fixed-size data units, called cells from the ATM jargon. This means that in the case of variable-size packets on transmission lines, as it is normally the case in the Internet, packets must be segmented into cells at switch inputs, and cells must be reassembled into packets at switch outputs [1].

The multiple-stage Clos switching fabric was proposed as a scalable architecture for the implementation of large-capacity switches. It is a potential solution to overcome the limited scalability of single stage switches, in terms of number of I/O chip pins and the number of switching elements. In a Clos-network switch packet scheduling is needed as there is a large number of points where contention may occur. Cells that have lost contention must be either discarded or buffered. Generally, buffers for storing cells and solve the contention problems can be placed at inputs, outputs, inputs and outputs, and/or within the switching fabric. Depending on the buffer placement, respective switches are called input queued (IQ), output queued (OQ), combined input/output queued (CIOQ) and combined input/crosspoint queued (CICQ) [2].

One way to ease the complexity of scheduling in Clos-network switches is by allocating memory in the first and third stages. In this way, if contention for an internal link occurs, loser cells are stored in the buffers in the first stage modules. These switches can be referred to as the Memory-Space-Memory (MSM) Clos-network switches. As the memory technology evolves, the memory amount that can be embedded into a chip is no longer a strict limitation.

In the MSM Clos-network switch the input modules have virtual output queues (VOQs), where one queue per output port is allocated to store cells for that output. Thanks to VOQs the switch avoids the Head-Of-Line (HOL) blocking problem. While cells are being routed in a switching fabric, it is very likely that more than one cell is destined to the same output port or for a physical link inside the switching fabric. The fast arbitration schemes have to be employed to solve internal blocking and output port contention problems. The arbitration scheme decides which items of information should be passed from inputs to arbiters, and – based on that decision – how each arbiter picks one cell from among all input cells destined for the output. Algorithms which can assign a route between input and output modules are usually called packet dispatching schemes. Considerable work has been done on scheduling algorithms for VOQ switches. Most of them achieve 100% throughput under the uniform traffic, but the throughput is usually reduced under the nonuniform traffic [1, 3-13].
In this paper new results obtained for the CRRD-OG packet dispatching scheme under the nonuniform traffic distribution patterns in the MSM Clos switching fabric are presented. The idea of the open grants was introduced by us in [11], where the performance of the CRRD-OG scheme under the uniform traffic with Bernoulli arrivals was also evaluated. The results presented in this paper cover the bi-diagonal, trans-diagonal, and Chang’s nonuniform traffic distribution patterns. The performance of CRRD-OG scheme under the bursty traffic is also presented. The simulation results are compared with the findings of the CRRD, CMSD and SRRD packet dispatching algorithms [4, 6]. These algorithms also use the effect of desynchronization of arbitration pointers in the Clos-network switch and common request-grant-accept handshaking scheme.

The remainder of this paper is organized as follows. Section 2 introduces some background knowledge concerning the MSM Clos switching fabric; we refer to that knowledge throughout the paper. Section 3 presents the CRRD-OG packet dispatching scheme. Section 4 is devoted to performance evaluation of the CRRD-OG algorithm. We conclude this paper in section 5.

2. MSM CLOS SWITCHING NETWORK

Clos-networks are well known and widely analyzed in the literature [14]. The three-stage Clos-network architecture is denoted by $C(m, n, k)$, where parameters $m$, $n$, and $k$ entirely determine the structure of the network. There are $k$ input switches of capacity $n \times m$ at the first stage, $m$ switches of capacity $k \times k$ at the second stage, and $k$ output switches of capacity $m \times n$ at the third stage. The capacity of this switching system is $N \times N$, where $N = nk$. The three-stage Clos switching fabric is strictly nonblocking if $m \geq 2n-1$ and rearrangeable nonblocking if $m \geq n$. We define the MSM Clos switching fabric based on the terminology used in [4] (see Fig. 1 and Tab. 1).

In the MSM Clos switching fabric architecture the first stage consists of $k$ IMs, and each of them has an $n \times m$ dimension and $nk$ VOQs to eliminate HOL blocking. The second stage consists of $m$ bufferless CMs, and each of them has a $k \times k$ dimension. The third stage consists of $k$ OMs of capacity $m \times n$, where each $OP(j, h)$ has an output buffer. Each output buffer can receive at most $m$ cells from $m$ CMs, so a memory speedup is required here.

Generally speaking, in the MSM Clos switching fabric architecture each $VOQ(i, j, h)$ located in $IM(i)$ stores cells going from $IM(i)$ to the $OP(j, h)$ at $OM(j)$. In one cell time slot VOQ can receive at most $n$ cells from $n$ input ports and send one cell to any CM. A memory speedup of $n$ is required here because the rate of memory work has to be $n$ times higher than the line rate. Each $IM(i)$ has $m$ output links connected to each $CM(r)$, respectively. A $CM(r)$ has $k$ output links $LC(r, j)$, which are connected to each $OM(j)$, respectively.
Janusz Kleban, Piotr Pietruszewski

3. CRRD-OG PACKET DISPATCHING SCHEME

The CRRD-OG packet dispatching scheme is an enhanced version of the CRRD scheme thanks to implementation of the open grant rules. An open grant is sent by CM to IM and contains information about unmatched link from the second to the third stage. In other words IM(i) is informed about unmatched output link LC(r, j) to OM(j). Because the architecture of the Clos-network is well-defined, it is also information about the switching system outputs, which can be reached from

| IM | Input module at the first stage |
| CM | Central module at the second stage |
| OM | Output module at the third stage |
| i | IM number, where 0 ≤ i ≤ k-1 |
| j | OM number, where 0 ≤ j ≤ k-1 |
| h | Input/output port number in each IM/OM, where 0 ≤ h ≤ n-1 |
| r | CM number, where 0 ≤ r ≤ m-1 |
| IM(i) | The (i+1)th input module |
| CM(r) | The (r+1)th central module |
| OM(j) | The (j+1)th output module |
| IP(i, h) | The (h+1)th input port at IM(i) |
| OP(j, h) | The (h+1)th output port at OM(j) |
| LI(i, r) | Output link at IM(i) that is connected to CM(r) |
| LC(r, j) | Output link at CM(r) that is connected to OM(j) |
| VOQ(i, j, h) | Virtual output queue at IM(i) that stores cells from IM(i) to OP(j, h) |
output $j$ of $CM(r)$. On the basis of this information $IM(i)$ looks up through VOQs and search for a cell which is destined to any output of $OM(j)$. If such cell exists it will be sent in the next time slot.

In the CRRD-OG algorithm two phases are necessary to complete the matching process. Phase one is the same as in the CRRD algorithm. In detail, the CRRD-OG algorithm works as follows:

- **PHASE 1**: Matching within IM
  
  **First iteration**
  
  - Step 1. Request: Each nonempty $VOQ(i, v)$ sends a request to every output
    link $LI(i, r)$ arbiter within $IM(i)$.  
  
  - Step 2. Grant: Each output link $LI(i, r)$ chooses one VOQ request in a round-robin fashion and sends the grant to the selected VOQ. It starts searching from the position of $PL(i, r)$.  
  
  - Step 3. Accept: Each $VOQ(i, v)$ arbiter chooses one grant in a round-robin fashion and sends the accept to the matched output link $LI(i, r)$. It starts searching from the position of $PV(i, v)$.
  
  **$i$-th iteration ($i>1$)**:
  
  - Step 1. Each unmatched $VOQ(i, v)$ at the previous iterations sends another request to all unmatched output link arbiters.  
  
  - Step 2 and 3. These steps are the same as in the first iteration.

- **PHASE 2**: Matching between IM and CM
  
  - Step 1. Request: Each selected in phase one IM output link $LI(i, r)$ sends the request to $CM(r)j$th output link $LC(r, j)$.  
  
  - Step 2. Grant: Each round-robin arbiter associated with output link $LC(r, j)$ chooses one request by searching from the position of $PC(r, j)$, sends the grant to the matched $LI(i, r)$ of $IM(i)$.  
  
  - Step 3. Open Grant: If after step 2 still exist requests, which are not granted, and unmatched output links $LC(r, j)$, each unmatched output link $LC(r, j)$ selects one request and sends open grant to the output link $LI(i, r)$ of $IM(i)$. The open grant contains the number of an idle output of the CM module, and simultaneously determine $OM(j)$ to which it is possible to send a cell.
  
  - Step 4. If the arbiter associated with $LI(i, r)$ receives the grant from $LC(r, j)$ it sends a cell at the next time slot, from the matched $VOQ(i, v)$ to $OP(j, h)$ through $CM(r)$. If the arbiter associated with $LI(i, r)$ receives the open grant from $LC(r, j)$ it has to choose one cell, which is destined to $OM(j)$ and sends it at the next time slot. The IM cannot send the cell without receiving the grant or the open grant. Not granted requests will be attempted to be matched at the next time slot because the pointers are updated only if the matching is achieved.
4. SIMULATION EXPERIMENTS

Two packet arrival models are considered in the paper: the Bernoulli packet arrival model arrival model and the bursty traffic model. Under the Bernoulli arrival process the probability that a cell may arrive in a time slot is denoted by $p$ and is referred to as the load of the input.

In the bursty traffic model, each input alternates between active and idle periods. During active periods, cells destined for the same output arrive continuously in consecutive time slots. The average burst (active period) length is set to 16 cells.

We consider several nonuniform traffic distribution models which determine the probability that a cell which arrives at an input will be directed to a certain output. The considered traffic models are:

- **Trans-diagonal traffic** – in this traffic model some outputs have a higher probability of being selected, and respective probability $p_{ij}$ was calculated according to the following equation:

$$
\begin{align*}
    p_{ij} &= \begin{cases} 
    \frac{p}{2} & \text{for } i = j \\
    \frac{p}{2(N-1)} & \text{for } i \neq j
    \end{cases}
\end{align*}
$$

(1)

- **Bi-diagonal traffic** – is very similar to the nonuniform traffic but packets are directed to one of two outputs, and respective probability $p_{ij}$ was calculated according to the following equation:

$$
\begin{align*}
    p_{ij} &= \begin{cases} 
    \frac{2}{3} & \text{for } i = j \\
    \frac{1}{3} & \text{for } j = (i+1) \mod N \\
    0 & \text{otherwise}
    \end{cases}
\end{align*}
$$

(2)

- **Chang’s traffic** – this model is defined as:

$$
\begin{align*}
    p_{ij} &= \begin{cases} 
    0 & \text{for } i = j \\
    \frac{p}{N-1} & \text{otherwise}
    \end{cases}
\end{align*}
$$

(3)

The experiments have been carried out for the MSM Clos switching fabric of size $64 \times 64 - C(8, 8, 8)$, and for a wide range of traffic load per input port: from $p = 0.05$ to $p = 1$, with the step 0.05. The 95% confidence intervals that have been calculated after $t$-student distribution for ten series, per 55000 cycles each (after the starting phase comprising 15000 cycles, which enables to reach the stable state of the switching fabric), are at least one order lower than the mean value of the simulation results, therefore they are not shown in the figures. We have evaluated
two performance measures: the average cell delay in time slots and the maximum VOQs size. The results of the simulation are shown in the charts (Fig. 2-9). Fig. 2, 4, 6 show the average cell delay in time slots obtained for the Chang’s, trans-diagonal and bi-diagonal traffic patterns, whereas Fig. 3, 5, 7 show the maximum VOQ size in a number of cells. Fig. 8 and 9 show the results for the bursty traffic with the average burst length set to 16 cells. The results obtained for the CRRD, CMSD and SRRD algorithms are also shown in the charts for comparison.

We can see that for the bursty traffic and all investigated traffic distribution patterns the CRRD-OG algorithm provides better performance than the CRRD, CMSD and SRRD algorithms. In many cases the CRRD-OG algorithm with one iteration delivers better performance than other algorithms with four iterations.

The Chang’s distribution traffic pattern is very similar to the uniform distribution traffic pattern. Under this traffic pattern all algorithms receive 100% throughput and CRRD-OG scheme with one iteration delivers better performance than other algorithms with four iterations. (Fig. 2, 3). The trans-diagonal and bi-diagonal traffic distribution patterns are highly demanding and the investigated packet dispatching schemes cannot provide the 100% throughput for the MSM Clos switching fabric. The best results have been obtained for the CRRD-OG scheme. These are respectively: under trans-diagonal traffic pattern - 80% throughput for one iteration and 85% throughput for four iterations (Fig. 4) and under bi-diagonal traffic pattern – 95% (Fig. 5). Under the bursty packet arrival model the CRRD-OG scheme provides much better performance than other algorithms especially for the very high input load (Fig. 8). The same relationship as for the cell delay we can observe for the maximal VOQs size (Fig. 3, 5, 7, 9). It is obvious that for small cell delay the size of VOQs will be also small.

The simulation experiments have shown that the CRRD-OG scheme with one iteration provides a noticeable improvement in the average cell delay and VOQs size. Any increase in the number of iterations do not produce further improvement, quite the opposite to other iterative algorithms. Particularly more than \(n/2\) iterations do not change significantly the performance of all investigated iterative schemes.

The investigated packet dispatching schemes are based on the effect of desynchronization of arbitration pointers in the Clos-network switch. The authors have made an attempt to improve the desynchronization method for the CRRD-OG scheme to ensure the 100% throughput for the nonuniform traffic distribution patterns. Additional pointers and arbiters for open grants had been added to the MSM Clos switching fabric, but the scheme was not able to provide 100% throughput for the nonuniform traffic distribution patterns. To our best knowledge it is not possible to achieve very good desynchronization of pointers using the methods implemented in the iterative packet dispatching schemes. In our opinion
the decisions of the distributed arbiters have to be supported by the central arbiter, but the implementation of such solution in the real equipment will be very complex.

Fig. 2. Average cell delay, Chang's traffic

Fig. 3. Maximum VOQ size, Chang's traffic

Fig. 4. Average cell delay, trans-diagonal traffic

Fig. 5. Maximum VOQ size, trans-diagonal traffic

Fig. 6. Average cell delay, bi-diagonal traffic

Fig. 7. Maximum VOQ size, bi-diagonal traffic
CONCLUSIONS

In this paper new results of simulation studies carried out for the CRRD-OG packet dispatching scheme under the nonuniform packet distribution patterns are presented. This scheme uses the distributed arbiters and common request-grant-accept handshaking scheme. Simulation experiments have shown that the proposed scheme is not able to achieve the 100% throughput for all kind of nonuniform traffic distribution patterns. The scheme produces very good results for the uniform and Chang’s traffic patterns with Bernouli arrivals and for the bursty traffic. In general, the CRRD-OG scheme provides the best performance from among all investigated algorithms.

REFERENCES


