Reducing external speedup requirements for input-queued crossbars

Berger, Michael Stübert

Published in:
Workshop on High Performance Switching and Routing, 2005. HPSR.

Link to article, DOI:
10.1109/HPSR.2005.1503227

Publication date:
2005

Document Version
Publisher's PDF, also known as Version of record

Reducing External Speedup Requirements for Input-queued Crossbars

M. S. Berger, *Member, IEEE*

Abstract—**This paper presents a modified architecture for an input queued switch that reduces external speedup. Maximal size scheduling algorithms for input-buffered crossbars requires a speedup between port card and switch card. The speedup is typically in the range of 2, to compensate for the scheduler performance degradation. This implies, that the required bandwidth between port card and switch card is 2 times the actual port speed, adding to cost and complexity. To reduce this bandwidth, a modified architecture is proposed that introduces a small amount of input and output memory on the switch card chip. This architecture allows for internal speedup in the switch card and the external speedup between port card and switch card can be reduced significantly. A simulation study is used for buffer dimensioning and demonstrates the feasibility of the proposed architecture.**

Index Terms—**Input queued switch. Scheduling, Speedup.**

I. Introduction

CROSSBAR switch fabrics have been studied extensively in the literature. In combination with Virtual Output Queuing (VOQ) the architecture provides a scalable solution with respect to memory access bandwidth. An input queued bufferless crossbar requires a complex scheduling mechanism that matches inputs with outputs. The scheduling algorithm is typically classified as either a maximum weight match or a maximal match type. A maximum weight match algorithm assigns a weight to each pair of inputs and outputs, and the maximal weight match pairs the inputs and outputs that result in the highest total weight. The weight could indicate the age of a cell or occupation of a VOQ. In the simplest case, the weight just indicates, by a one or a zero, whether there is a packet available or not. In this case, the scheduler calculates a maximum size match because it pairs the maximum number of inputs and outputs. On the other hand, a maximal match algorithm is characterized by the property that all unmatched inputs has no cell in any queue destined to an unmatched output. This implies that no further matches can be added unless the existing matches are rearranged.

A number of Maximum weight matching algorithms have been presented in [1]. Their main disadvantage is timing complexity, leading to an interest in maximal matching algorithms such as PIM and iSLIP [2]. SLIP matches input with output by having a round robin scheduler for each input and output. The input schedulers independently select an output, and the output scheduler selects among contending inputs. The iterative SLIP, iSLIP, performs a number of iterations of SLIP. To compensate for the lower performance of a maximum matching algorithm, speedup is introduced between the VOQs and the crossbar. A speedup of 2 is sufficient to obtain 100 % throughput [3].

The basic input queued switch is composed of port cards containing VOQ buffers and a switch card containing the bufferless switch chip(s) and the scheduler chip. Typically, the large physical size of a packet switch system implies that port cards and the central switch cards are separated and located in different shelves and racks. This leads to a high Round Trip Time (RTT) between port cards and switch cards. Potentially, this might affect the performance of the switch system, because the central scheduler calculates a match based on delayed VOQ state information. Solutions to this problem are discussed in [4]. One approach is to implement small VOQs close to the switch core and putting large buffers on the remote line cards. This approach is in fact utilized in the tiny-tera concept (LCS protocol) [5]. However, additional chips are required to implement the additional switch card VOQs, and this will add to complexity and power consumption. Another solution proposed in [4] is to have the VOQs on the line cards communicate arrivals instead of state information to the central scheduler. In this case, the scheduler will calculate the state information for all VOQs. The performance of this approach for a specific scheduling algorithm iSLIP has been evaluated in [6]. The scheme denoted ΔiSLIP was shown to have good performance close to that of iSLIP, and much better performance than in the case where delayed state information is communicated from the line cards. ΔiSLIP is however susceptible to loss of arrival information and furthermore, the system still requires a speedup in the order of two between port cards and switch card. Furthermore, it is proposed in [4] to integrate VOQs and switch fabric on the same chip, however this will only work for very small systems.

Another way of improving performance by adding buffer capacity to the switch chip is the buffered crossbar with VOQ, first introduced in [7]. Buffered crossbars have several advantages compared to non-buffered crossbars including

Manuscript received March 18, 2005. This work was supported in part by the European Union (EU) project Ethernet Switching at Ten gigabit and Above (ESTA). IST-2001-33182.

M. S. Berger is with Research Center COM, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark. Phone: +45 45 25 38 53. e-mail: msb@com.dtu.dk.

simpler arbitration, synchronization relaxation and better performance. The main drawback, however, is the total amount of crossbar memory that is proportional to the square of the number of input/output ports and RTT.

This paper presents a modification to the basic input queued switch architecture. The goal of the proposed architecture is to be able to support reasonable large RTT values and on the same time reduce the required speedup between port cards and switch card. Speedup is expensive, especially given that switch chips are currently most limited by the IO bandwidth across chip and card boundaries. In fact, high-speed serial link communication adds significantly to the overall power consumption. In the proposed architecture, small VOQ input buffers and output buffers are added to the switch chip, and this allows for decoupling of port speed and scheduling speed. Internally, a speedup of two between input and output buffers can be realized, but externally, the port speed can be reduced compared to the basic architecture. The size of the added buffer capacity in the switch chip will impact the tradeoff between performance and implementation complexity. To small buffers would lead to poor performance and to large buffers would not be feasible to implement. In this paper, a simulation study has been performed to quantitatively assess the tradeoff between performance and buffer size. Actually, it will be shown that a significant reduction in the required speedup can be obtained with a reasonable and feasible amount of switch chip buffer capacity.

A detailed description of the switch model is given in sec. II. The simulation study presented in sec. III compares the performance of this switch architecture to a basic input queued system. The simulation study is furthermore used as a guideline for system dimensioning, and the memory requirements will be compared to a buffered crossbar. Finally, concluding remarks are given in sec. IV.

II. SWITCH MODEL

The basic bufferless crossbar architecture is shown in Fig. 1. The system is usually denoted Combined Input and Output Buffered (CIOB) switch. A switch system of size $N \times N$ consists of N Input/Output port cards and a switch card implementing the N^2 crosspoint matrix. Each input port card contains VOQs with one buffer for each of the N outputs. The output port card contains a buffer to store cells in case of speedup. Queue status information is sent to the central scheduler in the switch card. In this paper, the iSLIP scheduler is assumed, even though improved algorithms have been proposed, e.g. [8]. Typically, a large round trip time of several cells between port card and switch card due to transmission time and synchronization. One of the main drawbacks of the CIOB architecture is the required speedup of 2 between port cards and switch card. Increasing the number of high-speed serial links significantly increases power consumption and chip pin count. Fig. 2 presents the modified switch card architecture with a small amount of buffer memory introduced on the switch card chip.

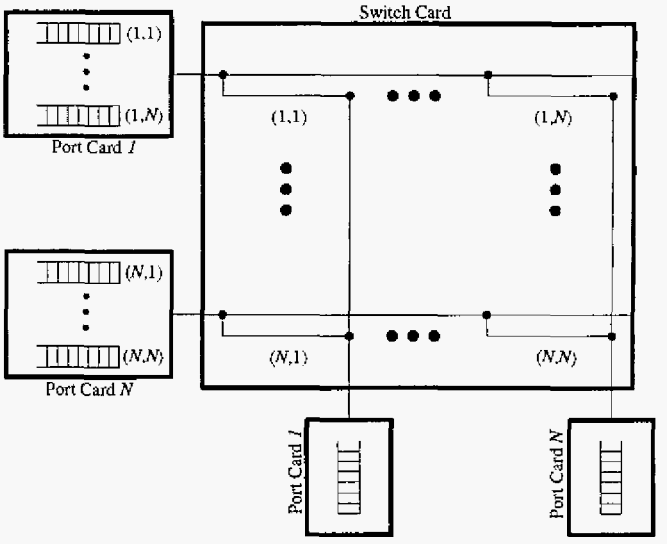

Fig. 1. Bufferless crossbar with combined input and output queuing. This architecture typically requires an external speedup of 2 between port cards and switch card to obtain 100 % throughput.

The main motivation for this architecture is to reduce external speedup between port cards and switch card, and this is achieved by introducing an internal speedup between switch card input and output buffers. Each new input queue system has a dedicated VOQ for each output. The VOQs are implemented in a shared memory following e.g. a linked list approach. A speedup of 2 can now be performed internally between switch card input and output buffers, and the switch card output buffers are therefore required to perform rate adaptation between internal and external speed. Since the switch card VOQ buffers have limited capacity, backpressure signals towards the port card VOQs are required.

The Round Trip time for backpressure *RT_BP* is defined as the number of timeslots it takes to stop the cell flow to a specific switch card VOQ measured from the time when backpressure was asserted by that VOQ. The round trip time

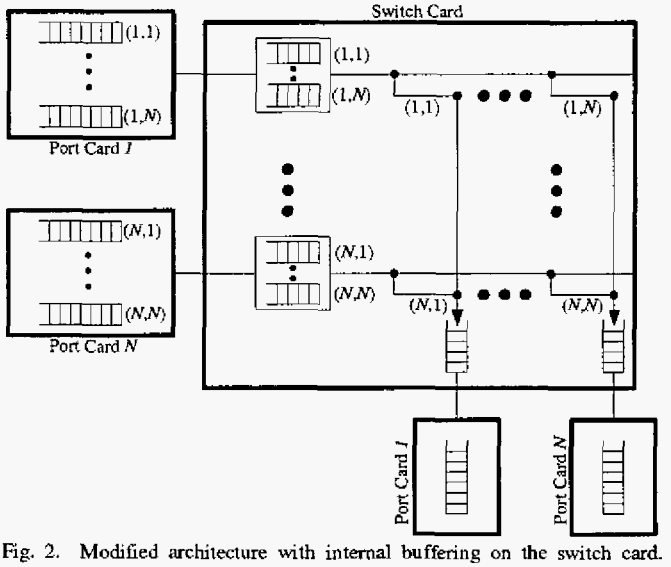

Fig. 2. Modified architecture with internal buffering on the switch card. There is internal speedup between switch card input and output buffers, which reduces external speedup requirements significantly.

is composed of a propagation delay for the backpressure signal, the time it takes before the port card scheduler is blocked and the data path delay from the port card scheduler to the switch card VOQ. The port card scheduler in this study performs a simple Round Robin (RR) arbitration.

In the following, the number of cells in VOQ number i in the shared memory is denoted Q_i. The backpressure threshold for queue i is B_i, that is, a backpressure signal is generated if $Q_i \geq B_i$. Due to the round trip time for backpressure signals, the size of queue i can grow to $Q_{i,\max} = B_i + RT\_BP$. The total number of cells in the shared buffer is $Q = \sum Q_i$. The total capacity of the shared memory S is typically much smaller than $\sum Q_{i,\max}$, therefore a global backpressure threshold B is introduced to avoid queue overflow. The global backpressure signal is then asserted if $Q \geq B$. The global threshold must be selected such that $B + RT\_BP \leq S$ in order to avoid overflow in the shared buffer.

Backpressure is asserted if the VOQ buffer level equals or exceeds RT_BP ($B_i = RT\_BP$). The total occupancy of a switch card input buffer could potentially reach $2*RT\_BP*N$, but the size is typically less than that, and the global backpressure signal is required, that blocks all port card VOQs. The size should be large enough to reduce the global backpressure to a minimum.

The switch card output buffer could potentially become congested as well. In this case, a backpressure signal is transmitted to the scheduler such that requests to this output are ignored.

In addition to benefits from reduced external speedup, the architecture has other advantages, including synchronization relaxation between port cards and switch card. Furthermore, communication between port cards and switch card is simplified because the scheduler works on local information from the switch card VOQs. Only simple backpressure signals are required between port cards and switch card.

III. Simulation and Results

A simulation study has been carried out in order to compare the new architecture in Fig. 2 with the well-known bufferless crossbar in Fig. 1. Each port card receives cells from a source. In each timeslot, the source generates a cell with probability equal to the load ρ. The switch size is 32x32. The destination is selected randomly according to a uniform distribution. Assigning the same destination to a number of consecutive cells generates bursty traffic. Fig. 3 shows the average delay as a function load for a burst length of 0, 10 and 20 respectively.

The round trip time for backpressure is set to four, $RT\_BP = 4$. The size of switch card input and output buffer is set to 100 and 20 respectively. It is concluded that the average delay for the modified switch architecture with internal speedup of 2 is close to the delay of an output buffered switch.

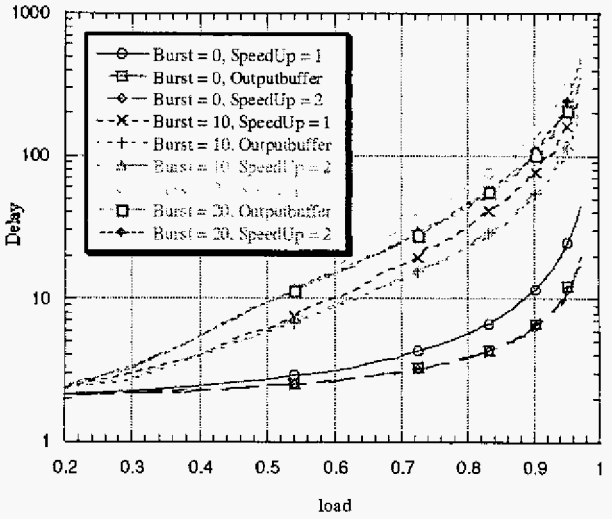

Fig. 3. Average delay vs. load for different values of burstiness. Delay is measured in number of timeslots. With an internal speedup of 2, the modified architecture has an average delay close to that of an output buffered switch.

In order to determine the required switch card input buffer capacity, the input buffer occupancy has been examined for various load and burst values. The results are shown in Fig. 4, for load values of 85 %, 90 % and 95 %. The switch card output buffer size is 20. This value is explained and justified later. The occupancy increases rather slowly with the burst size. A detailed investigation shows less than logarithmic growth, and this result is used to dimension the buffer by taking only the system load into account. Assuming a load of 95 %, the average occupancy is below 40, and by allocating 80 buffer locations, global backpressure is practically eliminated. The total number of switch card buffer locations becomes (80+20)*32 = 3200, feasible to implement on a single chip.

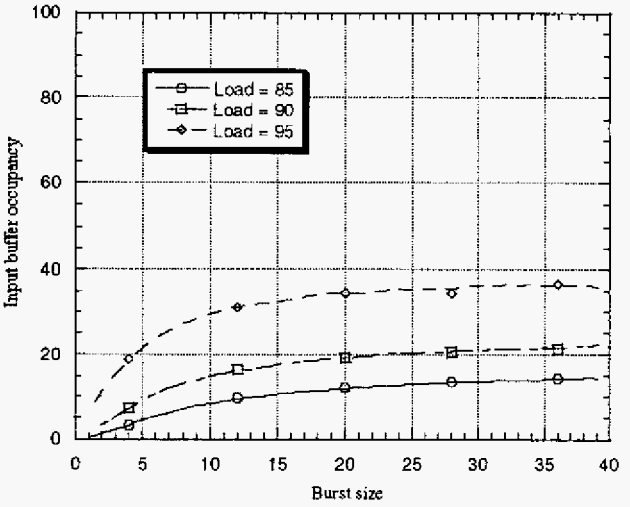

Fig. 4. Average input buffer occupancy vs. burst size. The occupancy grows slowly with burst size, and the buffer can therefore be dimensioned based on load values only.

To determine the switch card output buffer size, the switch performance under unbalanced traffic is investigated. Maximal matching algorithms do not provide 100 % throughput for unbalanced traffic, and speedup and output buffering is therefore required. The model for unbalanced traffic from [6][9] is used below: The unbalance weight w defines the degree of unbalance. The load from input port s to output port d is denoted $\rho_{s,d}$:

$$\rho_{s,d} = \begin{cases} \rho\left(w + \frac{1-w}{N}\right) & \text{if } s=d \\ \rho\left(\frac{1-w}{N}\right) & \text{otherwise.} \end{cases}$$

Note that

$$\sum_s \rho_{s,d} = \sum_d \rho_{s,d} = \rho .$$

The traffic matrix is thus admissible, and all input and output have a load equal to ρ. If w=0 there is no unbalance, and if w=1 the traffic is completely unbalanced. Fig. 5 shows the performance degradation under unbalanced traffic for different sizes of the switch card output buffer and also the case with no speedup. The throughput penalty depends on the output buffer size, and with a buffer size of 20, the degradation is small, below 4 %. At this point, it can be concluded from Fig. 4 and Fig. 5 that the switch can support full throughput if the load is 95 % with an input buffer size of 80 and output buffer size of 20. The switch can therefore support 100 % throughput with a small external speedup of 5 %. Also, the behavior under unbalanced and bursty traffic has been investigated to ensure that 20 output buffer locations are sufficient for bursty traffic. For a given degree of unbalance, the throughput was actually improved by increasing the burstiness of traffic. This is mainly because the switch card input buffer occupancy decreases due to a lower average number of simultaneous flows.

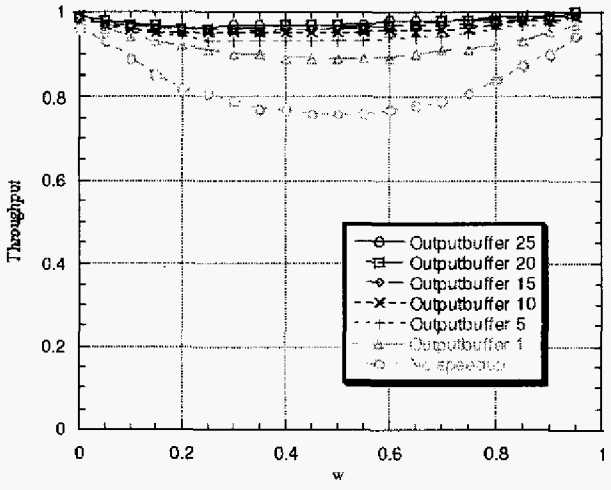

Fig. 5. Throughput with unbalanced traffic patterns. The throughput reduction depends on the switch card output buffer size. The largest reduction is obtained in the case with no internal speedup.

Now, as discussed previously, good performance can be obtained with 80 input buffer locations and 20 output buffer locations, which is 100 in total. The average number of buffer locations pr crosspoint is thus 100/32 = 3.125. In a buffered crossbar with VOQ, the crosspoint buffer size must be at least equal to *2RT_BP*, which equals 8 positions. The memory reduction is thus approximately 60%, but at the cost of a more complex scheduler operating at a speedup of 2 and therefore also a higher memory access bandwidth. Comparing Fig. 5 with performance figures in [9], it is concluded that the behavior from a performance perspective is similar.

IV. Conclusion

The modified switch architecture presented in this paper has several advantages compared to basic bufferless switch card architecture. By introducing small switch card buffers, the required iSLIP speedup can be performed internally on-chip, and external speedup can be reduced significantly. The simulation study shows that only a small external speedup of 5 % was required with 80 locations in the switch card input buffer and 20 locations in the switch card output buffer. The reduction in speedup also reduces the number of high-speed serial links, and thereby a reduction in power consumption is also obtained. In this paper, only round robin scheduling in port cards was assumed. However, it is expected that other scheduling schemes in the port cards, e.g. Oldest Cell First (OCF), might increase performance. OCF is difficult to implement in the traditional input buffered crossbar, but easy in the new architecture because the scheduling is performed independently among port cards. This is for further study to determine the impact on performance by utilizing other port card scheduling algorithms.

References

[1] N. McKeown, A. Mekkittikul, V. Anantharam, J. Walrand, "Achieving 100% throughput in an input-queued switch", *IEEE Transactions on Communications*, Vol.47 Issue.8 1999

[2] N. McKeown, "The iSLIP scheduling algorithm for input-queued switches", *IEEE/ACM Transactions on Networking*, p 188 -201, Vol.7 Issue.2, 1999.

[3] J.G. Dai, B.Prabhakar, "The throughput of data switches with and without speedup", *Proceedings IEEE INFOCOM 2000*, p556-64 vol.2 2000.

[4] C. Minkenberg, R. P. Luijten, F. Abel, W. Denzel, M. gusat, "Current issues in packet switch design", ACM SIGCOMM Computer Communication Review, Volume 33, Issue 1, January 2003.

[5] N. McKeown et al. "Packet Switch System", United States Patent 6,647,019.

[6] C. Minkenberg, "Performance of iSLIP Scheduling with Large Round-Trip Latency", Proceedings of *IEEE Workshop on High Performance Switching and Routing*, Torino, Italy, 2003.

[7] M. Nabeshima, "Performance evaluation of a combined input- and crosspoint-queued switch", *IEICE Transactions on Communications*, Vol.E83-B Issue.3, p737-41, 2000

[8] R. Manivasakan, M. Hamdi, D.H.K. Tsang, "The Dual Round Robin Pseudo-grant Matching for high-speed packet Switches", Proceedings of *HPSR 2002*, Kobe, Japan, 2002.

[9] R. Rojas-Cessa, E. Oki, H.J. Chao, "CIXOB-k: combined input-crosspoint-output buffered packet switch", Proceedings of IEEE GLOBECOM 2001, p 2654-2660, Vol.4