Buffer Management in High-performance Routers

Dragi Kimovski[1] and Atanas Hristov[2]

Abstract – **Buffer management methods are crucial for achieving high utilization of the communicational resources provided by the high-performance routers. Ineffective management of the available buffers can limit the channel bandwidth and increase the backpressure on the reverse communicational channels. This could lead to degradation of the communicational performance of the interconnection network. In this paper a novel buffer management method has been suggested. The proposed method manages the buffer space by coordinating the timing when the flits can be transferred from the current router to the input buffers of the next-in-line router.**

Keywords – ***Buffer management, High-performance Routers, Interconnection Networks.***

I. Introduction

The most important component of every cluster based high-performance computing system (HPCS) is its interconnection network. In a concurrent HPCS, the interconnection network connects thousands of computing and I/O nodes. As the number of nodes increases, the effective allocation of the shared communicational resources can decline dramatically, resulting in disruption in the intra-cluster communication [2].

Congestion control, which is performed by the flow control mechanism, ensures the effective allocation of the network resources among the competing network agent. In general, the flow control mechanism should allocate the communicational resources in highly effective way so the interconnection network can achieve high fraction of its maximal throughput and can deliver the packets with constant low latency.

The flow control mechanisms can be divided in to two main categories: bufferless and buffered flow control. In the area of HPCS the usage of routers that are based on buffered flow control mechanisms is mandatory, mainly because they provide low latency and high channel throughput [4, 5]. All of the routers that use agent buffering technique require a means to communicate the availability of the downstream buffers. By exchanging information about the available buffer space the upstream nodes can determinate if the transmitting flits could be saved in the buffers of the appropriate routers.

This type of buffer management provides significant backpressure, i.e. network communication in reversed direction. Usually, the intensity of the backpressure depends on the type of the buffered flow control mechanism that is implemented in the router. The buffers of the routers that implement packet-buffered flow control are relatively simple to manage. Large buffer space allows flexible allocation of the available buffers to the appropriate network agents - packets. On the other hand, the routers that support flit-buffered flow control mechanisms have small input buffers. This limits the available storage space for the network agents, which translates into complex and less flexible buffer management.

Buffer management methods are crucial for achieving high utilization of the communicational resources. Ineffective management of the available buffers can limit the channel bandwidth and increase the backpressure on the reverse communicational channels.

The goal of this paper is to suggest a highly effective strategy for buffer management in flit-buffered high performance routers. The proposed buffer management method combines the advantages of the “on/off” and “credit-based” managements, whilst reducing the backward communication traffic.

II. Background

The network routers require an effective means for management of the available buffers. In order to achieve adequate allocation of the expensive buffers it is mandatory to provide information about the storage space availability between the neighboring nodes. Only then the routers will have clear image of the free buffers in the routes of the traversing agents.

When implementing a buffer management strategy it is important to take into account the specifics of the flow control mechanisms and the architecture of the input modules of the routers [7]. The relation between them can be grasped by examining the concurrent buffer management strategies.

Concurrent high performance routers usually use “credit-based” or “on/off” management methods.

A. Credit Based Buffer Management

With the “Credit-based” strategy, the upstream routers are keeping information about the number of available buffers on each input channel downstream [6]. Therefore, each time an upstream router forwards a flit, it decrements the appropriate count. If the count on the upstream router reaches zero, then all of the downstream buffers are full and cannot receive any additional flits until a buffer space become available. On Figure 1 a time line illustrating the “Credit-based” strategy is shown. The round delay time $\boldsymbol{t_{rt}}$ for this method is the minimum time interval $\boldsymbol{t_1}$-$\boldsymbol{t_5}$ between the sending of two consecutive credits from the same buffer [1].

Flit-buffered flow control mechanisms, like “Wormhole” and “Step-back-on-blocking” are limiting the router buffers to only few flits per input port [3]. This means that on each step every traversing flit would have to wait for a credit or acknowledgment for available buffer space before it can advance to the upstream router. This would greatly decrease

[1] Dragi Kimovski is with the University for Information Science and Technology, Partizanska bb, Ohrid 6000, Republic of Macedonia, E-mail: dragi.kimovski@uist.edu.mk.

[2] Atanas Hristov is with the University for Information Science and Technology, Partizanska bb, Ohrid 6000, Republic of Macedonia, E-mail: atanas.hristov@uist.edu.mk.

the maximal throughput of the output channel. The channel bandwidth will be limited to one flit per round trip time t_{rt}. In other words, this would mean that the maximal bit rate will be equal to S_{flit}/ t_{rt}., where S_{flit} is the size of the flit. In order to prevent the buffer management method from limiting the maximal throughput over a channel d, we would require a buffers size:

$$G > \frac{t_{rt}d}{S_{flit}} \qquad (1)$$

It is immediately clear that the "Credit-based" strategy could greatly reduce the channel through, when a flit-buffered flow control is implemented in the network routers.

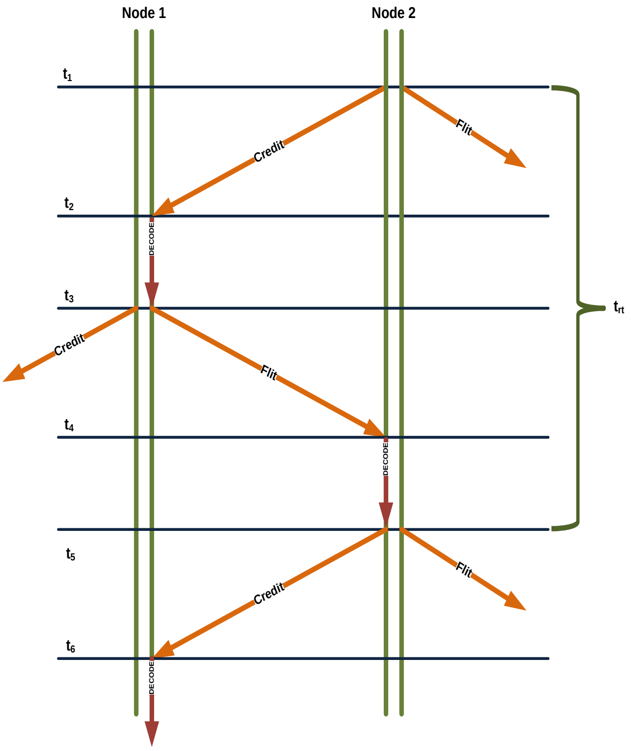

Fig. 1. Buffer management with a "credit-based" method

B. "On/Off" management strategy

"On/Off" management introduces an approach that greatly reduces the amount of reverse traffic. With this strategy the state of the upstream router is represented as a single bit control flag, which can be switched between two states: permitted to send-***on*** or not permitted to send-***off***. A control signal is sent via the reverse channel only when the state of the upstream router has changed. The control state is always changed from ***on*** to ***off*** when the number of available buffers falls below predefined threshold T_{off}. The opposite ***on*** state is activated only when the number of available buffers rises above the threshold T_{on}.

Even though, this buffer management method appears to be highly effective it suffers from frequent buffer overflows, especially in routers with small buffers. The buffer overflow usually appears due to the wire latency and computational time that is needed for sending and decoding of the on/off signals by the downstream router. To prevent buffer overflow the "On/Off" buffer management requires minimal buffer size G_m of:

$$G_M > \frac{t_{ss}d}{S_{flit}} \qquad (2)$$

Where t_{ss} is the time needed for receiving and decoding of the on/off signal. If this equation is not satisfied then the buffers can easily overflow. This can result in degradation of the communicational performance of the interconnection network.

The routers that support "Wormhole" and "Step-back-on-Blocking" flow control usually have small input buffers. This means that their size is not adequate for the "On/Off" buffer management. The implementation of this buffer management method in flit-buffered routers will surely lead to frequent buffer overflows and poor network performance.

III. Sliding interval buffer management

In order to be fully functional, both "credit-based" and "on/off" buffer management methods require predefined minimal buffer size. On the other hand the flit-buffered routers cannot provide the adequate buffers size for effective implementation of the stated management methods. We suggest a novel sliding interval buffer management (SIBM) method that primary addresses buffer allocation and management in flit-buffered routers.

The SIBM method manages the buffer space by coordinating the timing when the flits can be transferred from the current router to the input buffers of the next-in-line router. In fact, the SIBM method is based upon the "On/Off" strategy, but it adds a function for discrete timing of the flit forwarding and storing.

On Figure 2 the buffer management with SIBM method is shown.

The buffer management is performed per virtual channel i.e. the discrete forwarding timings are calculated individually for each packet. The process begins with the injection of the head flit from the NIC to first in line router. When the head flit is ready to be forwarded to the upstream router, the average passing time t_{avg} through the switch is calculated. This time is used in order to define the forwarding intervals for the next flits of the same packet. The forwarding intervals between two consecutive flits are calculated using a so-called sliding interval coefficient k. The forwarding intervals are not fixed, they are continuously decreasing until the upstream router returns a signal for possible appearance of overflow or blocking. The forwarding interval T_{for} is calculated by multiplying the average passing time of the flit with the sliding coefficient $t_{avg}*k$. The initial value of the sliding coefficient should be predefined with value $t_{avg}>1$. Choosing a value below 1 can induce frequent buffer overflow, because the upstream router will not have time and buffers to forward the current flits before accepting new ones. Selecting the initial value of the sliding coefficient can be problematic. Low values can lead to buffer overflows, but higher values can limit the channel bandwidth. From the conducted simulation experiments we can conclude that the best value for the sliding window is 1.4.

Subsequently, each time a flit is sent to the upstream router the value of sliding coefficient is decreased by a predefined constant ***c***. The decreasing constant is defined in accordance with the packet size $\boldsymbol{S_{packet}}$. The value of the constant is determined by the following equation:

$$c = \frac{k-1}{S_{packet}} \tag{3}$$

With each forwarded flit, the initial value of the sliding coefficient is reduced by the predefined constant. This means that the current router will forward the flits in continuously decreasing intervals, until the value of the sliding coefficient becomes 1, i.e. when all flits from the packet have been forwarded. This will guarantee reliable flit transfer and buffer management in low traffic conditions. But in reality, the interconnection networks and routers are usually working with high offered communicational load. This means that the head flit can easily become blocked, which will lead to congestion of the whole packet and buffer overflow.

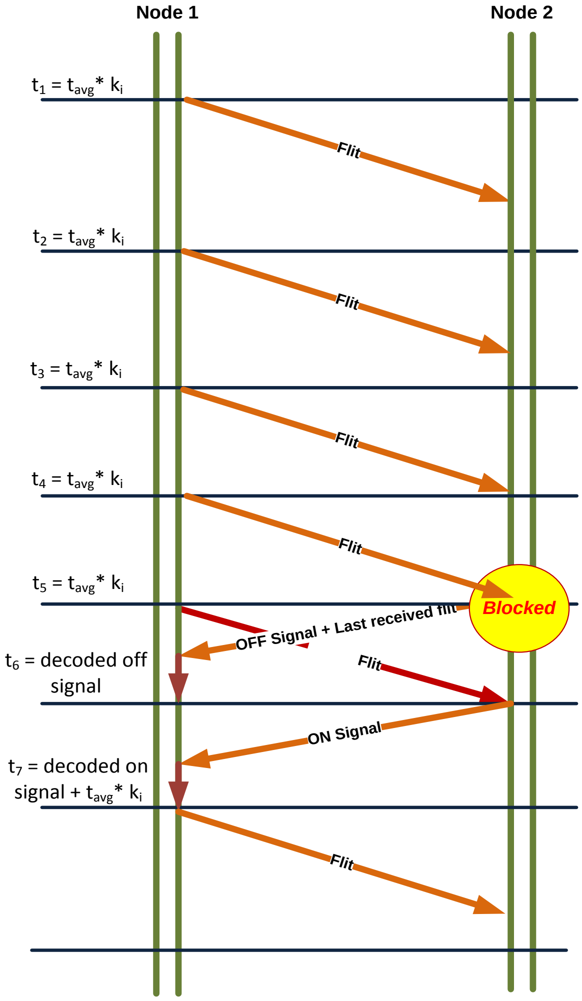

Fig. 2. Buffer management with the SIBM method

In order to overcome the buffer overflows, in the cases where the packet is blocked, the SIBM method introduces the usage of the “on/off” strategy. So, if the packet is blocked then all upstream routers will immediately send an ***off*** signal to the previous routers. The off signal carries the unique number of the currently stored flit in the buffers of the upstream router. This will provide information to the downstream routers about the status of the last successfully forwarded flit. After the off signal has been received end decoded, the current router will stop with flit transfer until the blocking has been overcome. Only then the upstream routers will send an ***on*** signal in reversed direction. This will reset the sliding interval to its initial value and the whole process will start again.

The sliding interval will be also reinitialized if more than one flit has been received on the input ports of the upstream router. In that case the “on/off” strategy is applied.

IV. Experimental Evaluation

On the basis of simulation experiments conducted in the OMNet++ discrete event simulation environment the efficiency of the SIBM method has been evaluated. OMNet++ simulation environment has been chosen because it provides extensible component based simulation framework [8].

The effectiveness and the behavior of the proposed buffer management method were verified indirectly, by evaluation of the communicational performance of an interconnection network where it has been implemented. More specifically, the experimental evaluation included measuring of the reverse and forward traffic between neighboring nodes, comparing the maximal latency to the offered load and measuring the effect of the value of the sliding coefficient to the network performance.

On Figure 3 the reverse traffic between neighboring nodes is given. The figure exhibits the effectiveness of the SIBM over the concurrent methods.

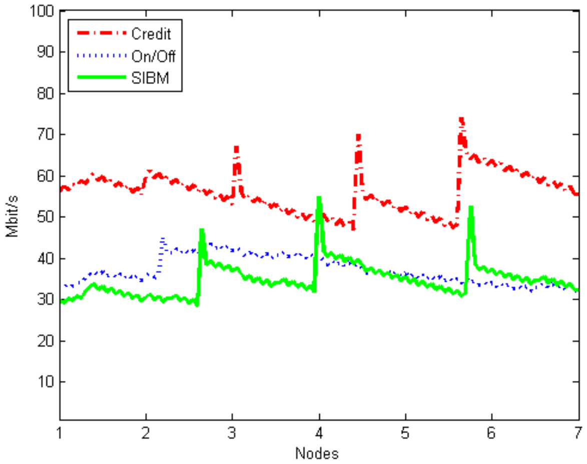

Fig. 3. Reverse traffic between the given nodes

On Figure 4 the maximal achieved forward bandwidth per node’s physical channel is given. Once more, we can see that the prosed buffer management method achieves higher bandwidth than “credit-based”, while maintaining lower reverse communication.

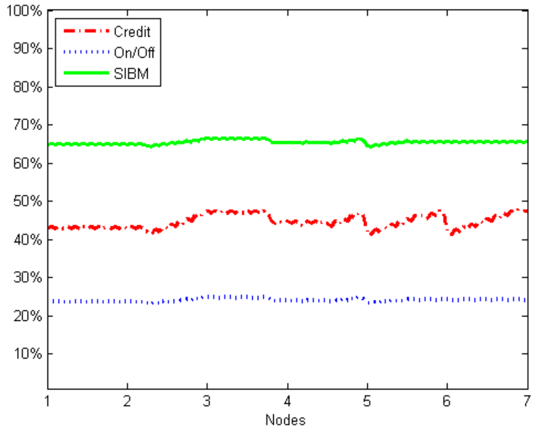

Fig. 4. Forward traffic between the given nodes. Given as a fraction of the ideal bandwidth

Figure 5 exhibits the relation between the latency and the offered load of an interconnection networks where the SIBM method has been implemented. The flit size is 32.

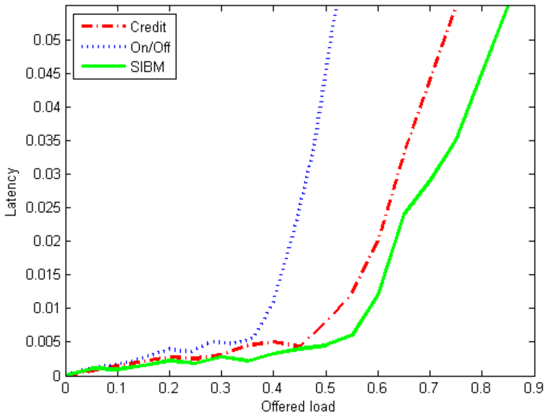

Fig. 5. Latency vs. Offered Load

On Figure 6 the relation between the value of the sliding coefficient and the maximal channel bandwidth is given.

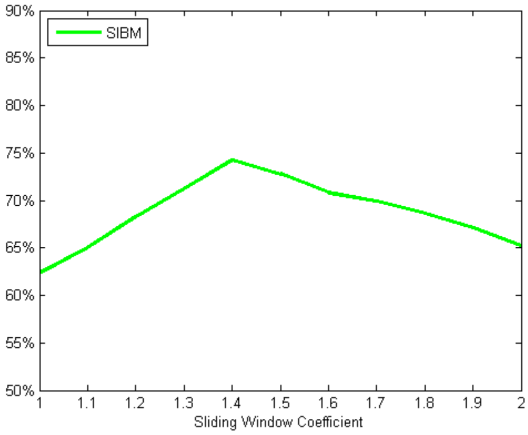

Fig. 6. Relation between the sliding interval coefficient and the maximal achieved bandwidth

V. Conclusion

Buffer management methods are crucial for achieving high utilization of the communicational resources. Ineffective management of the available buffers can limit the channel bandwidth and increase the backpressure on the reverse communicational channels.

In this paper a novel buffer management method has been suggested. The proposed SIBM method manages the buffer space by coordinating the timing when the flits can be transferred from the current router to the input buffers of the next-in-line router. In fact, the SIBM method is based upon the “On/Off” strategy, but it adds a function for sliding interval timing of the flit forwarding and storing.

The effectiveness and the behavior of the proposed buffer management method were verified indirectly, by evaluation of the communicational performance of an interconnection network where it has been implemented.

In future, we intend to verify and explore the efficiency of the proposed buffer management method in respect to its behavior for a wide spectrum of network topologies, flow control mechanisms and router architectures.

References

[1] W. J. Dally, B. Towles “Principles and Practices of Interconnection Networks”, Morgan Kaufmann Publishers, 2004.

[2] D.Abts, J. Kim, “High Performance Data Center Networks”, Morgan and Claypool Publishers, 2011.

[3] Plamenka Borovska, Dragi Kimovski, Atanas Hristov „Step-Back-On-Blocking Flow Control Mechanism For High Performance Interconnection Networks“, Xth International Conference CHER 21, 2012.

[4] Seiculescu, C., Murali, S. ; Benini, L. ; De Micheli, G., “A method to remove deadlocks in Networks-on-Chips with Wormhole flow control”, Design, Automation & Test in Europe Conference & Exhibition, 2010, Page(s): 1625 – 1628.

[5] Rubio, J.M.M. Lopez, P. Duato, J. “Flow control-based distributed deadlock detection mechanism for true fully adaptive routing in wormhole networks”, Parallel and Distributed Systems, IEEE Transactions, Volume: 14 , Issue: 8 Page(s): 765- 779, 2003.

[6] H. T. Kung, R. Morris, “Credit-Based Flow Control for ATM Networks”, IEEE Network Magazine, March 1995.

[7] Ximing Hu, Jing Qu, Yinhai Li, Binqiang Wang, “VOIQ: a practical high-performance architecture for the implementation of single-buffered routers”, High-Performance Computing in Asia-Pacific Region, 2005. Proceedings. Eighth International Conference, 2005.

[8] Lencse, Gábor and Varga, András, "Performance Prediction of Conservative Parallel Discrete Event Simulation", In Proceedings of the 2010 Industrial Simulation Conference: 214-219.