ELectronics EXpress

IEICE Electronics Express, Vol.8, No.20, 1743–1750

A hybrid multicast deadlock-free scheme for virtualization at the NoC level

Wenmin Hu[a)], **Hengzhu Liu, and Botao Zhang**

Computer School, National University of Defense Technology, Changsha, P.R. China

a) *huwenmin@nudt.edu.cn*

Abstract: A new hybrid multicast deadlock-free scheme is proposed to enhance the multicast capability. In the proposed scheme, the small-sized multicast packet is routed in a deadlock-free way by packet-buffered and asynchronous replication, and the large-sized multicast packet is transferred under the control of centralized allocator, which restricts the number of concurrent multicast with large-sized packet transmission. The virtualization at the Network-on-Chip (NoC) level is also taken into consideration, that the allocator reserves respective counters for each sub-network. According to result of the experiment under the real workload traces, the performance reduction caused by centralized allocator is negligible. The router and the allocator are synthesized in Chartered 90 nm CMOS technology. Compared with the only packet-buffered scheme, the allocator only consumes extra 0.079% to 0.2972% area overhead according to the different setting (routing mechanism, network size, number of supported sub-network) while offering the ability of large-sized multicast packet transmission.

Keywords: multicast, centralized allocator, Network-on-Chip, virtualization

Classification: Other communication hardwares

References

[1] C. J. Glass and L. M. Ni, "The turn model for adaptive routing," *Proceedings of the 19th annual international symposium on Computer architecture (ISCA)*, pp. 278–287, 1992.

[2] D. R. Kumar, W. A. Najjar, and P. K. Srimani, "A new adaptive hardware tree-based multicast routing in k-ary n-cubes," *IEEE Trans. Parallel Distrib. Syst.*, vol. 50, no. 7, pp. 647–659, 2001.

[3] M. Malumbres, J. Duato, and J. Torrellas, "An efficient implementation of tree-based multicast routing in distributed shared-memory multiprocessors," *Proc. 8th IEEE Symp. Parallel and Distributed Processing*, pp. 186–189, 1996.

[4] H. K. Young, S. Jeff, and D. Jeff, "Multicast routing with dynamic packet fragmentation," *Proc. 19th ACM Greatlakes Symp. VLIS*, pp. 113–116, 2009.

DOI: 10.1587/elex.8.1743
Received September 04, 2011
Accepted September 14, 2011
Published October 25, 2011

[5] J. Duato, S. Yalamanchili, and L. Ni, "Interconnection Network: An Engineering Approach," *Morgan Kaufmann Publishers*, 2003.
[6] S. Murali, "Design Reliable and Efficient Networks on Chips," *Springer verlag*, 2009.
[7] V. Varavithya and P. Mohapatra, "Asynchronous tree-based multicasting in wormhole-switched mins," *IEEE Trans. Parallel Distrib. Syst.*, vol. 10, pp. 1159–1178, 1999.
[8] N. E. Jerger, L. S. Peh, and M. Lipasti, "Virtual circuit tree multicasting: A case for on-chip hardware multicast support," *Proc. 35th annual international Symp. Computer architecture*, 2008.
[9] W. Hu, Z. Lu, A. Jantsch, and H. Liu, "Power-efficient tree-based multicast support for networks-on-chip," *Proceedings of 16th Asia and South Pacific Design Automation Conference (ASP-DAC11)*, 2011.
[10] X. Wang, M. Yang, Y. Jiang, and P. Liu, "On an efficient NoC multicasting scheme in support of multiple applications running on irregualr sub-networks," *Microprocessors and Microsystems*, vol. 35, pp. 119–129, 2011.
[11] P. S. Magnusson, M. Christensson, J. Eskilson, D. Forsgren, G. Hallberg, J. Hogberg, F. Larsson, A. Moestedt, and B. Werner, "A full system simulation platform," *Computer*, vol. 35, pp. 50–58, 2002.

1 Introduction

How to maintain deadlock-free is a key issue in designing routing scheme on wormhole Network, which covers unicast as well as multicast. The deadlock caused by the packets waiting for each other in a cycle is a common issue in unicast and multicast, and it can be solved perfectly when some special turns are forbidden according to turn model [1]. The other kind of deadlock only occurs in multicast transmission, caused by several concurrent multicast transmissions. Fig. 1 shows one configuration of this kind of deadlock. To solve this problem, several approaches have been proposed, such as detecting and recovering [2], pruning [3, 4]. However, these approaches either needs a large buffer to hold the entire packet or highly complicates router logic, which are not suitable for the Network-on-Chip. In distributed shared-memory (DSM) systems, since the invalidation and update commands only require a few flits, it is possible for the wormhole router to hold the entire packet and forward it to available branch, which can avoid deadlock efficiently [5]. Barrier synchronization, as a key operation in data-parallel programming model, also just needs a few flits [5]. However, with the development of the multi-core technology and downsizing of the feature sizes in semiconductor, thousands of cores are possibly integrated on one die. To efficiently utilize these vast computation resources, virtualization at NoC level is necessary, where a single NoC-based CMP may be shared by multiple applications, each of which is mapped to different sub-networks [6]. It is hard to conclude that all the multicast operations of these application just need a few flits. Hence, a hybrid multicast deadlock-free scheme is more promising in common use.

In this paper, we focus on the solution scheme for the second kind of

DOI: 10.1587/elex.8.1743
Received September 04, 2011
Accepted September 14, 2011
Published October 25, 2011

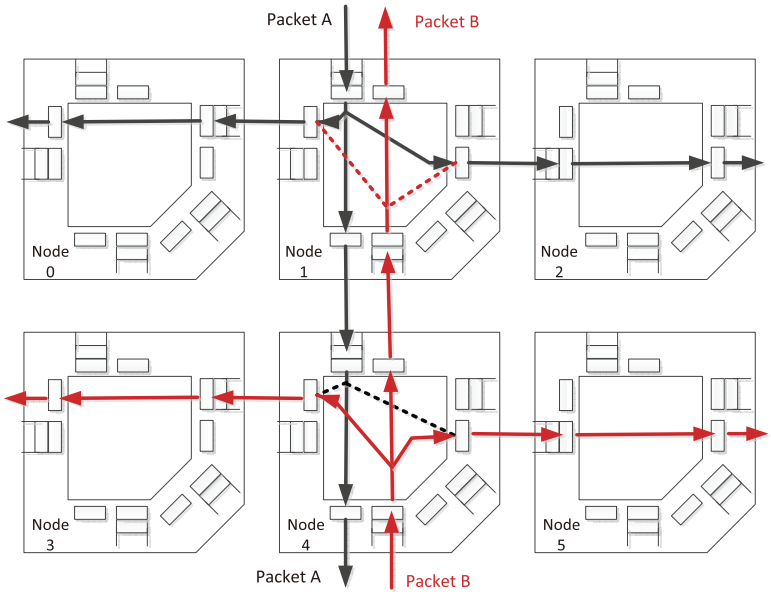


Fig. 1. A deadlock scenario (when two multicasts exist concurently)

deadlock, and propose a hybrid multicast deadlock-free scheme to support the virtualization at the NoC level. In our scheme, for the small-sized multicast packet, packet-buffered [5] and asynchronous replication [7] (PBaAR) are used to avoid deadlock, while for the large-sized multicast packet, the number of concurrent multicast transmission is restricted to constant through a centralized allocator. The second scheme can be viewed as deadlock prevention, which reserves all the required resources before packet transmission [5]. This scheme works efficient on the assumption that the frequency of large-sized multicast packet transmission is low both in space and time. Both scheme were implemented in Verilog HDL and synthesized in Chartered 90 nm CMOS technology. Compared with only PBaAR scheme, proposed hybrid scheme only increases the hardware-overhead by 0.079% to 0.2972% while extending the ability to support large-sized multicast packet transmission.

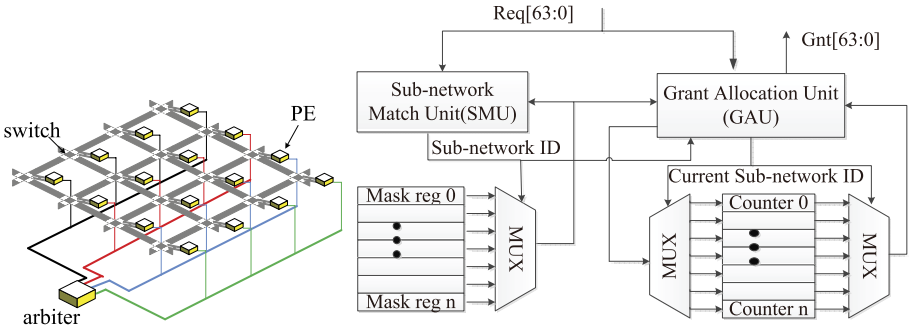
2 Proposed scheme

2.1 Packet-buffered scheme

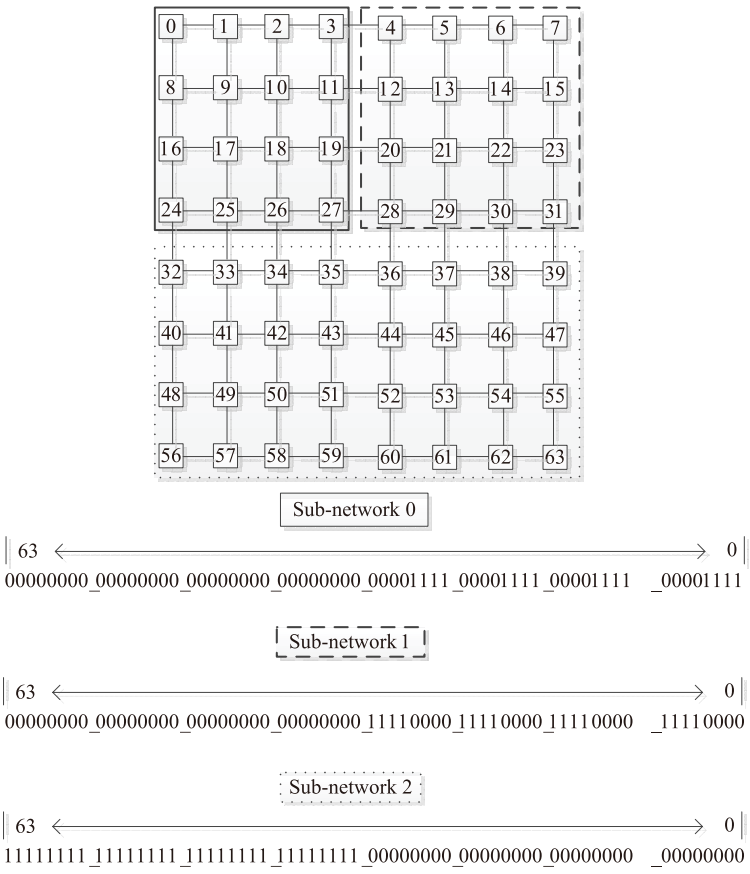
In many multicast scenarios, such as barrier synchronization, invalidation and update operations in DSM systems, only a few flits are required for a message [5]. Thus, it is possible to hold the entire packet in wormhole router when the packet is blocked. In the proposed scheme, asynchronous replication is adopted to forward the flit at the branch router, which means flit could be forwarded to the desired outgoing port independently [7]. If some outgoing port is blocked, the other outgoing port (s) can continue to forward the remaining flits in the buffer. Since the message only contains a few flits and the blocked router can hold the entire packet, the remaining flits of the packet can be forwarded to the available branch and the released virtual channel in downstream router after a few cycles.

DOI: 10.1587/elex.8.1743
Received September 04, 2011
Accepted September 14, 2011
Published October 25, 2011

As can be seen in Fig. 1, Packet A is blocked at Router 4 while Packet B is blocked at Router 1. Since asynchronous replication is adopted, the remaining flits of Packet A can be forwarded to the east-port and west-port. After the tail flits leave the west input buffer in Router 2 and east input buffer in Router 0, the virtual channels are released, then Packet B can get the grants of the virtual channel and continue to be forwarded to Router 0 and Router 2. In a similar way, the rest flits of Packet B are forwarded to Router 3 and Router 5, and then the corresponding virtual channels are released. Hence, Packet A can go on. The deadlock caused by multi-port requirements is solved perfectly.



(a) Centralized Allocator on 4×4 Mesh-based Network (b) Structure of Proposed Allocator (req: request signal vector, gnt: grant signal vector)



(c) Mapping Relation between the Node and the Mask register

Fig. 2. Proposed Allocator

DOI: 10.1587/elex.8.1743
Received September 04, 2011
Accepted September 14, 2011
Published October 25, 2011

2.2 Centralized allocate scheme

If the value of the multicast packet size is bigger than that of the depth of the buffer, the whole packet cannot be buffered at a single router, usually spanning several routers. In this case, if the packet is blocked, the asynchronous replication scheme cannot forward the whole packet to the available branch, and the virtual channel in the downstream router cannot be released. Fortunately, the multicast only occupies a little part of traffic in most application [8], and the large-sized packet transmission should be even less. It should be a simple and effective way to restrict the number of concurrent large-sized multicast. Furthermore, the increasing number of virtual channels and virtualization at NoC level allow more concurrent large-sized multicast transmission.

Fig. 2 (a) illustrates the proposed centralized allocate (CA) scheme based on a simple 4 mesh architecture. In CA scheme, before a large-sized multicast packet is issued at the sender, a request signal connected to the allocator is set to 1. Only when the returned grant signal is 1, can the packet be injected into network. Otherwise, the sender will continue to wait. As previously mentioned, the waiting occurs infrequently, but being deadlock-free is necessary. After the sender finishes the packet transferring, the request signal is set back to 0.

Fig. 2 (b) depicts the hardware architecture of the Centralized Allocator supporting for 8×8 mesh network, which has integrated two main functions. One is the sub-network matching. The Sub-network Match Unit (SMU) runs to find the sub-network hit by the request signals, which means that the node from the sub-network wants to transfer a large-sized multicast packet. If succeeds, it will stop and wait for feeding the value of sub-network mask register (SMR) to Grant Allocation Unit (GAU). Once the GAU gets the value of SMR, the SMU continues to search the next one.

GAU tries to grant the requests in a fair way (Round Robin) according to the value of respective counter register. Once the granting process of current sub-network is done, GAU will get the value of the next hit SMR (if it is ready) or wait the SMU to finish matching. The number of concurrent multicast transmission is determined by the number of virtual channels. For example, if a wormhole router contains four virtual channels for each input port, this network could be viewed as four virtual networks. For each virtual network, one multicast never causes the second kind of deadlock. Hence, the number of concurent multicast (Counter) is set to 4. Fig. 2 (c) shows the relationship between the nodes and SMR value in 64-core mesh network, which is partitioned to three sub-networks for different applications. Each node is labelled as shown in the left of Fig. 2 (c), the SMR value is represented by a 64-bit vector. If the bit is set to 1, the node belongs to the sub-network.

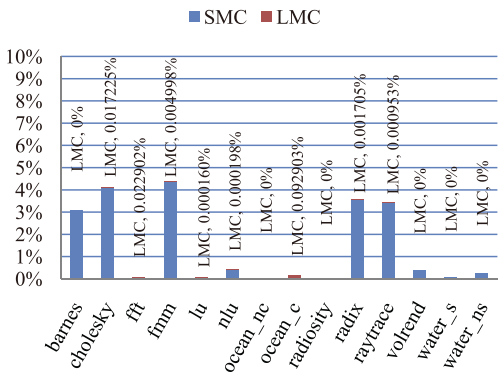
As mentioned previously, one virtual sub-network allows one large-sized multicast packet transmission. As shown in Fig. 1, we assume Packet A is a small-sized packet which can buffer at a router while Packet B is a large-sized packet which spans several router. Packet A can be forwarded to east branch

DOI: 10.1587/elex.8.1743
Received September 04, 2011
Accepted September 14, 2011
Published October 25, 2011

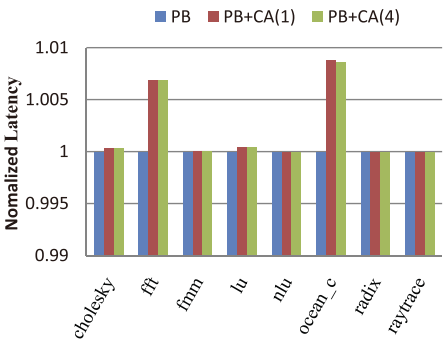
and west branch, and be blocked at the south branch. After the Packet A leaves the Node 0 and Node 2, the virtual channels are released, Packet B continues to be transferred. Once the tail flit of B leaves Node 3 and Node 5, Packet A can go on. Thus,the deadlock issue is solved by the hybrid scheme.

3 Evaluation

To evaluate the multicast performance under the influence of CA, we have implemented a cycle-accurate NoC simulator with SystemC. We evaluate the performance using the traces of real applications, which are extracted from SPLASH-2 by Simics [11]. Our CMP configuration has 64 processors, of which each has a L1 I/D Cache (64 KB, 4 way set-associative) and L2 Cache (256 KB, 8 way set-associative). Each processor is attached to one router in a 8×8 mesh, and 8 memory controllers are distributed at the borderline of the mesh. We use bit string encoding to indicate the destinations, where the routing computation for multicast packet can be finished in one cycle. The data width per flit is set as 8B. Hence, the control message needs 2 flits (1 for routing information and 1 for command), and the data message needs 10 flits (1 for routing information, 1 for command and 8 for cache line). Packet-buffered (PB) scheme is evaluated as the baseline, and the fifo depth is set to 10. In our experiment, the control message is classified as small-sized packet, while the data message is large-sized packet. The traces are extracted during the parallel period of the applications, where each core runs about 10 million instructions.



(a) Percentage of Multicast in Each Workload of SPLASH-2 (SMC: small-sized multicast, LMC: large-sized multicast)



(b) Performance(PB: only packet-buffered scheme, PB + CA (1): proposed scheme with 1 VC, PB + CA (4): proposed scheme with 4 VCs

Fig. 3. Performance Evaluation

As can be seen in Fig. 3 (a), multicast traffic occupies a small fraction of the workload of application, and the large-sized multicast takes even less (0.01% on average, 0.09% at most). However, the small fraction of the large-sized packet cannot prohibit multiple large-sized multicast occurrence in short period, which may cause deadlock. Fig. 3 (b) shows the performance

DOI: 10.1587/elex.8.1743
Received September 04, 2011
Accepted September 14, 2011
Published October 25, 2011

of multicast under the traces of SPLASH-2. Compared with PB, the latency increased by CA is negligible (0.2% on average, 0.88% at most). However, PB+CA reduces the required buffer size from 10 to 2 for each VC, which is more area-efficient.

Multicast router using asynchronous replication [9] and proposed allocator have been modeled in Verilog HDL, and synthesized using a 90-nm CMOS stand-cell technology library from Chartered Semiconductor Manufacturing with Synopsys Design Compiler. The multicast router using the asynchronous replication consists of VA (virtual channel allocator), SA (switch allocator), FM (Flow Manager), RC (Routing Computation Unit), SW (Switch), VC (virtual channels) and Multicast table. The centralized allocator is modeled as version for 8×8 and 16×16 mesh, and the number of supported sub-networks varies form 8 to 32. The mechanism of multicast routing may vary in different routers. For example, some are table-based [9]and other are logic-based [10]. Hence, to compare the area in a fair way, the area of logic-based router synthesized using the same 90-nm CMOS stand-cell technologies is also taken into consideration. We cite the area result of AL+RPM [10] and show it in Table. I. Table. I shows the synthesis results with 90-nm CMOS stand-cell technologies. Compared with 64-core mesh network, the allocator only increases area by 0.1297% for the logic-based network while 0.0818% for the table-based network. Compared with 256-core mesh network, the increased area is from 0.079% to 0.2972% according to the different number setting of SMRs and routing scheme. From the results, it can be conclude that the centralized allocator provides the ability to support certain large-size multicast packet transferring while the increased area overhead is negligible.

Table I. Area of the multicast router and allocator

Router	table-based [9]	logic-based (AL+RPM) [10]		
$Area_{router}(\mu m^2)$	341469	215447		
Network Size	8×8	16×16		
$Area_{net}(\mu m^2)$ [9]	21854016	87416064		
$Area_{net}(\mu m^2)$ [10]	13788608	55154432		
Allocator (bit, SMRs)	(64, 8)	(256, 8)	(256, 16)	(256, 32)
$Area_{alloc}$ (μm^2)	17882	69059	100413	163936
$Area_{alloc}/Area_{net}$[9]	0.0818%	0.079%	0.1149%	0.1875%
$Area_{alloc}/Area_{net}$[10]	0.1297%	0.1252%	0.1821%	0.2972%

4 Conclusions

This paper proposed a hybrid multicast deadlock-free scheme to enhance the multicast support on Network-on-Chip. Packet-buffered and asynchronous replication scheme was introduced to avoid deadlock for small-sized packet,

DOI: 10.1587/elex.8.1743
Received September 04, 2011
Accepted September 14, 2011
Published October 25, 2011

while the centralized allocator controlling the number of concurrent multicast is adopted to prevent deadlock for large-sized packet. What is more important, the number restriction is based on the sub-network, which is suitable for virtualization at NoC level. The evaluation result shows that both the latency and area overhead increased by CA is negligible.

Acknowledgments

This work is supported by the National Science Foundation of China, under Grant No. 60970037, and Doctor Program Foundation of Education Ministry of China, under Grant No. 20094307110009.

DOI: 10.1587/elex.8.1743
Received September 04, 2011
Accepted September 14, 2011
Published October 25, 2011