Exploring partitioning methods for multicast in 3D bufferless Network on Chip

Chaoyun Yao\textsuperscript{1(a)}, Chaochao Feng\textsuperscript{1}, Minxuan Zhang\textsuperscript{1}, Wei Guo\textsuperscript{1}, Shouzhong Zhu\textsuperscript{1}, and Shaojun Wei\textsuperscript{2}

\textsuperscript{1} College of Computer, National University of Defense Technology, Changsha, 410073, P.R. China
\textsuperscript{2} The Department of Electronic Systems, Tsinghua University, Beijing, P.R. China
\textsuperscript{a} ycy021417@163.com

Abstract: In this paper, we first propose two multicast partitioning methods named TBP (two block partitioning) and LBP (layer block partitioning), each of which has a different level of efficiency in 3D bufferless network on chip. In addition, we present a Recursive Partitioning (RP) method in which the network is recursively partitioned until all partitions contain comparable number of multicast destination nodes. By this method, the multicast destination nodes are distributed evenly and the network latency is significantly decreased. Simulation results illustrate that the RP scheme achieves 39\%, 45\% and 41\% less latency on average than that of the TBP scheme and 8\%, 17\% and 12\% less latency on average than that of the LBP scheme under three synthetic traffic patterns respectively.

Keywords: 3D, bufferless network on chip, multicast, partitioning methods

Classification: Integrated circuits

References

[1] W. J. Dally and B. Towles: DAC (2001) 684. DOI:10.1109/DAC.2001.156225
[2] L. Benini and G. De Micheli: Computer \textbf{35} [1] (2002) 70. DOI:10.1109/2.976921
[3] N. E. Jerger, L. S. Peh and M. Lipasti: ISCA (2008) 229. DOI:10.1109/ISCA.2008.12
[4] L. P. Carloni, P. Pande and Y. Xie: ACM/IEEE NoCs (2009) 93. DOI:10.1109/NOCS.2009.5071456
[5] B. S. Feero and P. P. Pande: IEEE Trans. Comput. \textbf{58} [1] (2009) 32. DOI:10.1109/TC.2008.142
[6] H. Matsutani, M. Koibuchi and H. Amano: ICPP (2007) 75. DOI:10.1109/ICPP.2007.79
[7] H. Matsutani, M. Koibuchi and Y. Yamada: IEEE Trans. Parallel Distributed Syst. \textbf{20} (2009) 1126. DOI:10.1109/TPDS.2008.233
[8] C. Feng, Z. Lu, A. Jantsch and M. Zhang: IEICE Trans. Inf. Syst. \textbf{95} (2012) 1519. DOI:10.1587/transinf.E95.D.1519
[9] R. S. Ramanujam and B. Lin: CAL \textbf{7} [2] (2008) 37. DOI:10.1109/L-CA.2008.6
[10] T. Moscibroda and O. Mutlu: ACM SIGARCH Computer Architecture News \textbf{37} [3] (2009) 196. DOI:10.1145/1555815.1555781
1 Introduction

As technology scaling with Moore’s law, a single die can integrate several hundreds of cores. When the number of cores increases, interconnect becomes more and more complicated in System-on-Chip design. Networks on chip (NoC) were shown to be feasible and easy to scale for supporting a large number of cores and memory [1]. NoC can provide high throughput and low latency for one-to-one (unicast) traffic at present. In order to run many kinds of parallel applications on Multi-Processors System-on-Chip (MPSoc), collective communications must be supported. In an MPSoc system, the cache-coherent shared memory protocols (such as directory-based or token-based) require one-to-many or broadcast communications to obtain shared data or invalidate shared data on different cache blocks [2]. It has been proved that only 2–5% multicast traffic in the total network traffic will have a serious impact on system performance [3]. As a result, supporting efficient multicast in NoC can make the performance of the MPSoc system better.

Furthermore, with the increasing number of cores, the 2D plane is not an efficient way due to its long wire interconnections [4]. The emerging of the 3D integrated circuit stacks several dies to reduce the long wire delay, which results in lower power consumption and higher performance [5, 6]. The 3D integrated circuit is also an attractive way for Network on chip. The 3D NoC architecture is widely studied in the network topology [7], router architecture [8], and routing algorithms [9].

Previous researches on NoC focus on the wormhole/virtual-channel router which contains buffers in each input port to buffer the packets. Though buffers can improve the bandwidth efficiency of the network, they also consume much energy and chip area [10]. Bufferless NoC has become a potential alternative to the traditional wormhole based NoC lately [10, 11, 12]. The bufferless router makes use of deflection routing to eliminate the buffers, which has two benefits: simplicity in design and reduce hardware cost. However, in bufferless NoC, packets cannot wait in a router and deflection routing makes the routing path unpredictable. Therefore it is impossible to employ existing multicast algorithms used in networks with buffers.

A method related to our work is discussed in [13] and [14] which propose partitioning methods for the 3D mesh network with buffers. However, it is not suitable for 3D bufferless network due to the fact that unpredictable routing path exists in bufferless network. To the best of our knowledge, there is no work addressing multicasting in 3D bufferless NoC systems. In this paper, we first...
propose two multicast partitioning methods named TBP (two block partitioning) and LBP (layer block partitioning), each of which has a different level of efficiency in 3D bufferless network on chip. In addition, we present a Recursive Partitioning (RP) method in which the network is recursively partitioned until all partitions contain comparable number of multicast destination nodes. By this method, the multicast destination nodes are distributed evenly and the network latency is significantly decreased. Simulation results illustrate that the RP scheme achieves 39%, 45% and 41% less latency on average than that of the TBP scheme and 8%, 17% and 12% less latency on average than that of the LBP scheme under three synthetic traffic patterns respectively.

2 3D bufferless router architecture

The baseline 3D bufferless router is a 1-cycle router with a 3-stage permutation network [8]. The structure of the baseline 3D bufferless router is show in Fig. 1. There are 7 input/output ports in each router. Each input port has only one input register, therefore the packet is not buffered in the router. It utilizes deflection routing to route packets. In the case of two or more flits contending for the same output of the permutation cell, the permutation cell selects the flit with the higher priority to the shortest output port and misroutes the other flit to the other output. The priority is decided according to the number of hops the packet has been routed in the network which can limit the number of misroutings to avoid livelock. The router can handle at most six packets at the same cycle and make output port allocation for each packet from high priority to low.

3 Partitioning methods for 3D bufferless mesh architecture

In this section, we propose three different partitioning methods for 3D bufferless mesh architecture.

3.1 Two-block partitioning (TBP) method

As the number of multicast destination increases, the long path of baseline 3D bufferless deflection routing without packets replication will lead to large multicast latency. To solve this problem, we propose TBP scheme with adaptive packets replication at the source node. Fig. 2 shows an example of this partitioning policy.
and the portions of each partition depending on the source node position. It is assumed that the source node is node 25 while the destination node set is $D = \{9, 16, 17, 24, 39, 44\}$. The multicast based on TBP method is shown in Fig. 2(a). First the multicast packet in Node 25 is copied into two and sent to the two different partitions separately, if the source router has more than two empty ports. Node 24 and 39 are chosen as the first best candidate since they have the minimum Manhattan distance to the source node 25. After the packet is sent to node 24 and 39, node 16 and 9 are chosen as the second best candidate. Without contention, the minimum multicast latency is equal to 9 hops. The path shown in Fig. 2(a) is not the only one path since the packet may be deflected due to contention. As illustrated in Fig. 2(a), the source node is located at the middle layer, two partitions cover comparable number of nodes. However in Fig. 2(b), the source is located at the first layer, one partition contains much more nodes than the other. Now suppose that the multicast message $m = (6, \{1, 16, 17, 24, 39, 44\})$ is generated by the source node 6. The destination nodes are split into two groups according to the labels: $G_H = \{16, 17, 24, 39, 44\}$ and $G_L = \{1\}$. The packet created for $G_H$ uses the manhattan path as follows: $\{6, 25, 24, 39, 40, 47, 16, 17, 46, 45, 44\}$ where 10 hops are needed to reach the last destination. The packet path for the $G_L$ is: $\{6, 1\}$ only one hop is required for delivering the packet to the destination. This is the drawback of TBP method.

![Fig. 2. The TBP multicast routing method](image)

**3.2 Layer block partitioning (LBP) method**

In this method, the network is partitioned into layer $0, 1, \ldots, n$ subnetworks. Destination are sorted based on each layer. The nodes with the same $z$ value will be put into the same group. The scheme has several advantages over the TBP scheme as it achieves a high degree of parallelism and avoids the creation of long multicast paths. For the LBP scheme, when a multicast packet is being injected into the network from the local port, if free output ports are plane ports, the packet will not be replicated and routed the same way as the path-based 3D bufferless multicast.
scheme. If the number of free output ports is more than one and is the up or down port and destinations fall into different layer corresponding to the free output ports, the destinations will be grouped according to the layer partition and the multicast packet will be replicated and routed through different output ports. The pseudo code of LBP scheme is shown in Table I.

**Table I.** Adaptive packets replication for LBP algorithm

| Input: Multicast destination address dst_add |
| Number of free output ports num_free_port |
| Free output port status free_port[6] |
| Output port priority output_prio[6] |
| Output: Replicated packet address replicate_pkt_addr[6] |

1: \(\text{copy_num} \leftarrow 0\)
2: if num_free_port = 0 then
3: Do not replicate the multicast packet
4: elsif free_port[6] = N or E or S or W then
5: Do not replicate the multicast packet
6: else
7: for i in 0 to 5 loop
8: if (Freeport[output_prio[i]] = 1 and copy_num < free_port_num) then
9: (replicate_pkt_addr[output_prio[i]]) \leftarrow \text{replicate_addr_gen}(dst_add, output_prio[i])
10: dst_addr := dst_addr and (not replicate_addr_gen(dst_add, output_prio[i]))
11: if replicate_pkt_addr[output_prio[i]] \neq 0 then
12: copy_num \leftarrow (copy_num + 1)
13: end if
14: end if
15: end loop
16: end if

Free_port is a 6-bit vector, each bit of which indicates the corresponding output port is free or not. Output_prio sorts the priority of the output port from the highest priority to the lowest. The router always chooses a free output port with the highest priority to replicate and deliver a copy of the multicast packet. The function replicate_addr_gen acquires the address of the replicated packet. After replication, the original destination address will be updated and the number of packet copies will increase by 1.

Fig. 3 shows an example of the multicast routing algorithm based on the LBP method. For the multicast message \(m = (25, \{9, 16, 17, 18, 24, 39, 46\})\), three groups are formed: layer0 = \{9\}, layer1 = \{16, 17, 18, 24\}, layer2 = \{39, 46\}. Without contention, the multicast packet requires 5 link traversals in total. The path shown in Fig. 3(a) is not the only one path since the packet may be deflected due to contention. The unbalanced partitions are shown in Fig. 3(b). The multicast message \(m = (6, \{1, 9, 16, 17, 18, 24, 39, 46\})\) requires 6 hops in total to reach the final destination without contention.
3.3 Recursive partitioning (RP) method

The objective of the recursive partitioning method is to optimize the number of destination nodes that can be included in a partition. In this method, the network is recursively partitioned until each partition includes less than \( n \) destination nodes. In the worst case, the network is partitioned into \( c \) layer partitions like the LBP method in an \( a \times b \times c \) network. So, we have considered the value \( n \) as the maximum number of destination nodes in a partition of the LBP method. The pseudo code of RP scheme is shown in Table II. \text{Num\_dst} indicates whether a partition of the LBP method includes more than \( n \) destination nodes. The number of \( n \) will change if the number of destinations of the multicast packets is changed or the size of the network is changed. When \text{free\_port} and \text{num\_dst} are equal to 1, the router will choose a free output port to replicate destination address and delivery a copy of the multicast packet. The function \text{replicate\_addr\_gen\_1} acquires the address of the replicated packet in a partition of the LBP method. After replication, the original destination address will be updated and the number of packet copies will increase by 1.

An example of the multicast routing algorithm based on the RP method is illustrated in Fig. 4, where a multicast message is generated at the source node 25. The steps of the RP algorithm can be expressed as follow:

1. The value \( n \) is set to 4 in a \( 4 \times 4 \times 3 \) network;
2. The network is divided into three partitions using LBP method. The Fig. 4(a) shows the partitions when the source node is 25;
3. If the number of nodes in a partition exceeds the predefined value \( n \), the function \text{replicate\_addr\_gen\_1} acquires the address of the replicated packet in a partition of the LBP method. The partition is divided into new partitions according to the location of the destination nodes. A region partition policy is proposed to divide the layer into at most 4 regions (P1,P2,P3,P4) according to the position of the source node, as shown in Fig. 4(a) layer1. If there is only one free output port, the packet will not be replicated and will be routed the same way as the LBP scheme. If the number of free output ports is more than
one and destinations fall into different regions corresponding to the free output ports, the RP scheme will work. Fig. 4(a) shows the example of recursive partitioning. There are 5 destination nodes in layer1 which is greater than \( n = 4 \). So, the multicast packet will be replicated according to the region partition in layer1. Without contention, the multicast path requires 5 hops. However, the multicast path will require 12 hops when it using the LBP method.

The same partitioning technique is applied to the low layer subnetwork. Fig. 4(b) shows another example of the RP method where the source node is 6. In this example, the multicast path requires 6 hops as the maximum latency.

4 Experimental evaluation

In this section, we evaluate the performance of the three proposed multicasting partitioning mechanism for the 3D mesh interconnection network with a cycle-accurate NoC simulator which is developed in VHDL.

4.1 Methodology

The experiments are performed on a \( 4 \times 4 \times 4 \) 3D mesh network. The routing process is partitioned into two steps. First of all, the routing computation in company with the input and output priority sorting are executed in parallel. Then the output port allocator allocates each packet based on the results of the first step. For the LBP and RP routers with adaptive packets replication, multicast packets are replicated at the moment of output allocation when at least one free output port left for replication. The three multicast routers are more complicated than the regular

| Table II. Adaptive packets replication for RP algorithm |
|---------------------------------------------------------|
| **Input:** Multicast destination address dst.add |
| whether a partition includes more than \( n \) destination nodes num.dst |
| **Output:** Replicated packet address replicate_pkt_addr[6] |

1: copy.num ← 0
2: num.dst ← num.dst.addr(row, col, layer, dst.add)
3: if num.free.port = 0 then
4: Do not replicate the multicast packet
5: elsif num.dst = 0 then
6: Do not replicate the multicast packet
7: else
8: for i in 0 to 5 loop
9: if (Freeport[output.prio[i]] = 1 and copy.num < free.port.num) then
10: (replicate_pkt_addr[output.prio[i]]) ← replicate_addr_gen_1(dst.add.output.prio[i])
11: dst.addr←dst.addr and ((not replicate_addr_gen_1(dst.add.output.prio[i]))
12: (if replicate_pkt_addr[output.prio[i]] ≠ 0 then)
13: copy.num ← (copy.num + 1)
14: end if
15: end if
16: end loop
17: end if
unicast bufferless router as a result of increasing route computation for multicast packets and packets replication. For the purpose of simplify the design, we use a 1-cycle router [8] for three kinds of multicast routers (TBP, LBP, RP routers) which is shown in Fig. 1. To perform the simulations, a combination of unicast and multicast traffic is used. For unicast traffic, we use three synthetic traffic patterns: uniform random, bit complement, transpose. For multicast packet, the destination positions are uniformly distributed.

We measure the average packet latency $T$ which is calculated by the equation and measured in cycle, where $T_{src}$ is the time that a packet startup latency and $T_{net}$ is the network delivery time. We also evaluate the link utilization, which is the average percentage of used links in total links. Link utilization can reflect the power consumption. The more used links, the higher power consumption is consumed.

\[
T_{latency} = T_{src} + T_{net}
\]

### 4.2 Latency analysis

We measure the performance of the three multicast schemes in this subsection. Fig. 5(a)–(c) illustrate the average packet latency of the three multicast schemes. The multicast traffic is 10% of the total traffic and the number of destinations is 8. It can be observed that the RP scheme achieves the smaller average latency than the other two multicast schemes. However, the network with the RP scheme reaches saturation earlier than the other two schemes. As the network is not saturated, the average latency of the LBP scheme is 32%, 40% and 35% less than that of TBP scheme under three synthetic traffic patterns respectively. The RP scheme achieves 39%, 45% and 41% less latency on average than that of the TBP scheme and 8%, 17% and 12% less latency on average than that of the LBP scheme under three synthetic traffic patterns respectively. The results are consistent with our expectations. The RP scheme outperforms the LBP scheme and TBP scheme at low or medium traffic loads. The RP scheme performs similar as the LBP scheme when the network reaches saturation point. The fact is that they can replicate multicast packets at both source and intermediate nodes which increases the network loads.
so the throughput of the network is a little bit lower than TBP scheme. However, under real application workloads the network is never fully loaded, so the RP scheme is a better alternative for bufferless NOC than the other two schemes. We synthesize three routers using Synopsys Design Compiler with 65 nm technology at the typical condition (25°C and 1.0 v). The critical path delay of the TBP multicast routers is 2 ns while the LBP and RP methods are at most 8% higher than this value, thus all the above conclusions still hold.

4.3 Link utilization

Fig. 6 shows the link utilization with the number of multicast destinations increasing from 4 to 16 under random traffic with 10% multicast traffic. In the case of the network not saturated, compared with TBP scheme, the RP scheme can save 12%, 16% and 23% of link utilization on average for different number of multicast destinations respectively. Compared with LBP scheme, the RP scheme can save 3%, 5% and 10% of link utilization. It can also be concluded that as the number of destinations increases, the RP scheme can save more power consumption than that of the other two multicast schemes.

4.4 Scalability

We also evaluate the performance of the three multicast schemes with different portion of multicast traffic and different number of multicast destinations. Fig. 7(a) shows the performance variation of the three multicast schemes with different
portion of multicast traffic from 10% to 40%. The number of multicast destinations is 8. It can be seen from the figure, the RP scheme achieves much less average latency than the TBP and LBP schemes with the increasing portion of multicast traffic. Fig. 7(b) reveals the average latency of the three multicast schemes with the increasing number of multicast destinations. The portion of multicast destinations is 10%. As the number of multicast destinations increases from 4 to 32, the average latency of the RP scheme does not vary so much, while the average latency of the LBP scheme increases by up to 1 times and the average latency of the TBP scheme increases by up to 3 times.

5 Conclusion

This paper proposes a novel idea of balanced partitioning which leads the 3D mesh architecture to be partitioned effectively. We first presented TBP and LBP partitioning methods for 3D bufferless mesh architecture, each of which has a different level of efficiency. In order to achieve higher performance, the recursive partitioning (RP) method was introduced. This method partitions the network recursively until all partitions contain considerable number of nodes. Our simulations with different traffic profiles showed that the RP method reduces the average packet latency and provides higher degree of parallelism compared with the other two methods.

Acknowledgements

This research is supported by National Natural Science Foundation of China with No. 61303066 and 61373032 and by Specialized Research Fund for the Doctor Program of Higher Education of China with Grant No. 20124307110016.