A dynamic adaptive arbiter for Network-on-Chip

Yanhua Liu1,2, Jie Jin2,3, Zongsheng Lai1

1Institute of Microelectronics Circuit & System, East China Normal University, Shanghai, China
2Jiangsu Provincial Key Lab of ASIC Design, Nantong University, Nantong, China
3Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai, China

Abstract: Network-on-chip (NoC) is considered as a promising paradigm to overcome the communication bottleneck of future multicore systems. As a basic component in on-chip router, arbiter has a big impact on the performance of router. In this paper, we propose a novel dynamically adaptive arbiter which is based on the round robin mechanism. The proposed arbiter detects buffer status of input ports and changes priorities of the input port dynamically to enhance the performance of the router. Simulation results show that the proposed arbiter can achieve 7.3% improvement in saturation packet injection rate and 13.3% improvement in saturation throughput of NoC on average, when compared with round robin arbiter. Using Synopsys design tools with 0.18-μm technology, implementation results show that a router with the proposed arbiter needs additional 4.8% area compared to a router with round robin arbiter.

Key words: Network-on-chip (NoC), on-chip router, dynamically adaptive arbiter, round robin mechanism,
to output port contention or blocking. Increasing buffer slots for heavy-load input ports can improve performance of router. However, more buffer slots will cause more area and power consumption. If heavy-load input ports can be authorized with high priority when they have requests for connecting output port, the pressure on these input port will be decreased and the performance of the router can be improved. So, among the three major aspects, arbiter plays an important role to enhance the router performance.

In this paper, we propose a novel dynamically adaptive arbiter (DAA) which can change the priority of input port dynamically according to its buffer status. Buffer full signal is used as high priority signal of input port. The input port with high priority is allowed to occupy its desired output port prior to other ports, and thus the buffer pressure of the input port can be decreased efficiently. Simulation results show that NoC using proposed arbiter achieves higher performance.

This paper is organized as follows: Section 2 describes related works for arbiters of on-chip router. Section 3 describes details of the proposed arbiter. Section 4 gives experimental and comparison results. Finally, conclusions are made in section 5.

2 Related Works

Many researchers focused on developing various arbitration schemes in order to achieve an efficient allocation and reduce packet latency. A lot of arbiters have been proposed in the routers of computer network such as round robin arbiter [4], fixed priority arbiter [5], lottery arbiter [6], token ring arbiter [7] and so on.

Round robin arbiter treats each input port fairly and guarantees fairness in scheduling. Using round robin arbiter, each input port have an equal chance to own the output port and the starvation problem can be solved. However, round robin arbiter is too fair and may cause low efficiency for some input ports. Fixed priority arbiter always authorizes the requiring input port with the highest priority when requiring contention happens. The input ports with lower priority may rarely be authorized which results in extremely unfair. Lottery arbiter offers input ports certain numbers of lottery as their priority level. Input port with more lotteries has bigger probability to win the output port. However, if the number of lotteries is static, some input ports which have little lotteries may be hardly responded under heavy traffic load. Token ring arbiter cannot guarantee correctness and may miss some requests from different input ports.

All above arbiters can be used in computer network but they are unfit for on-chip router. For on-chip router, resource consumption, average packet latency and complexity of priority strategy should be considered. Recently, some new arbitration methods based on round robin and lottery mechanisms were proposed for the on-chip router. A priority based output arbiter was proposed in [8] which could eliminate the congestion state of NoC. Before arbitrating, the arbiter counted the number of output port requirements for packets in each input port and gave a higher priority to the input port which had more requirements. Zhu et al. [9] presented three new scheduling methods based on round robin mechanism for the on-chip router. The three scheduling methods used different heuristic information to determine the scheduling sequence. However, these arbitration mechanisms needed to detect all packets in input ports and compute their routing paths before delivering them. This resulted in difficult hardware implementation of control logics and totally unsuitable for on-chip router.

Zhang [10] designed a statistic-based lottery arbiter which did not cause starvation problem. However, this arbiter needed many registers, which led to a large amount of resource consumption. A customized priority arbiter based on lottery mechanism for the on-chip router was proposed by Wu et al. [11]. The arbitral priorities were customized according to the communication cases among PEs in NoC. This arbiter used static priority which was not fit for dynamic real applications in NoC. Wang [12] improved the lottery arbiter and presented a dynamic lottery arbiter. The dynamic lottery arbiter detected the loads of input ports in every clock cycle and adjusted the priority of each input port dynamically. The proposed arbiter did not work efficiently under the uniform traffic. Even under non-uniform traffics, some heavy-load input ports might always occupy the output ports which caused starvation problem.

3 Design of dynamic adaptive arbiter

In our proposed arbiter, the input buffer full signals are detected as the high priority signals. A high priority will be given to the input port if the buffer of port is full. In order to prevent packet starvation, we design the proposed arbiter based on round robin mechanism. In this section, we first present a typical NoC platform and head-of-line blocking problem. Then, we describe the working principle of the proposed arbiter.

3.1 NoC platform and head-of-line blocking problem

Fig. 1 shows a typical 2D mesh NoC and the corresponding router architecture. 2D mesh is the most popular
topology for Network-on-Chip which has good scalability. The communication data named packet can be transferred by the on-chip routers and links in 2D mesh NoC. On-chip router is the core component in the NoC and has big impact on system performance [13].

Wormhole router is one of the most commonly used on-chip routers in NoC. It is easy for implementing and suitable for on-chip network. A typical wormhole on-chip router with dedicated buffer per input port is shown in fig. 1 (b). In the wormhole router, each packet is divided into small unit called flit. Head flit proceeds through all the stages while body and tail flits skip route computation and output arbitration stages. Body and tail flits inherit the output port allocated to the head flit. The last flit in a packet, called tail flit, releases the reserved output port that have been reserved by the header flit of that packet, when it departs the current router. The route computation determines deliver direction of the packet according to the destination of the head flit and the routing algorithm. Route computation sends request to the arbiter after determining deliver direction. Arbiter grants the request and connects output port with input port by MUX.

Arbiter is one of the key components in the on-chip router. It can determine the output sequence when output contention happens. If packets arrive at different input ports but need to be dispatched into the same output port simultaneously, a output contention will happen. In order to solve the contention, an arbitration mechanism is necessary to allow only one input port to access the output port. Most arbiters are unconcerned with buffer status and authorize the input port by a determinate mechanism. These arbiters may cause head-of-line blocking problems if the input port with full buffer cannot be authorized preferentially. Fig.2 shows an example of head-of-line blocking problem. As shown in fig. 2, west and south input ports request for east output port simultaneously in router R2. If the east output port is connected to the south instead of west input port in R2, packet p1 in west input buffer of R1 cannot advance. The packets behind p1 will occupy west input buffer in R1 and decrease the network performance. Therefore, head-of-line blocking is a key factor when evaluating different arbiters for on-chip router.

In order to alleviate head-of-line blocking problem, we propose a dynamic arbitration priority for each input port according to the buffer status of the port. If some input ports request the same output port simultaneously, the proposed arbiter will detect the buffer status of these input ports and check whether their buffers are full. Heavy-load input ports makes their buffer easily full. An input port with a full buffer means it cannot hold incoming packets any more. If packets in the full buffer cannot be delivered as soon as possible, the packets will be halted and cause packet latency. For this reason, the input port with full buffer should be authorized with high priority. However, under heavy-load traffic distribution in NoC, there is always more than one input port in a router whose buffer is full. If these input ports request the same output port, these ports should be treated fairly and have an equal chance to win the desired output port. Under the light-load distribution in NoC, there may be no input port in a router whose buffer is full. In other words, no input port has high priority and can be authorized preferentially. So, all input ports should be authorized with an equal chance. Fig. 3 shows a block diagram of one output arbiter in on-chip router which uses the proposed arbitration mecha-
nism. In fig.3, Buffer_full and Input_Port_Req are buffer full signals and request signals from input port of different directions.

\[\text{Input_Port1} \quad \text{Input_Port2} \quad \text{Input_Port3} \quad \text{Input_Port4} \quad \text{Output_Port} \]

\[\text{Input_Port_Req} \quad \text{Buffer_full} \quad \text{DAA} \quad \text{Gnt} \quad \text{Any_Gnt} \]

**Figure 3**: Block diagram for one output port arbiter

Fig. 4(a) gives detailed structure of DAA. As shown in fig. 4(a), DAA is based on two round robin arbiters. In the previous NoC research, many on-chip routers used round robin arbiter as their output arbiter [2] [14]. Round robin arbiter performs well for uniform distribution traffic, but it is not flexible for unbalanced distribution traffics when there are hotspots or priority requirements in NoC. In DAA, if some requests for output ports are generated by the input ports with full buffers, high priorities will be granted to these requests. These input ports will be scheduled first by round robin arbiter1. Other request of input ports with low priorities will be disabled when the high priority ports are scheduled.

In order to prevent starvation for the low priority input ports, a counter and a comparator are used in DAA. The counter is used to record the number of authorizing times for high priority input ports. If the number of authorizing times is bigger than \(T_{\text{threshold}}\), all input ports should be scheduled with equal priority by round robin arbiter2. \(T_{\text{threshold}}\) can be customized by NoC designer according to the characteristics of traffics. Moreover, it can be found that, when one of the round robin arbiters works, the request signals for the other round robin arbiter will be disabled. This will ensure that no request signal is missed by DAA in each clock cycle.

Fig. 4(b) shows the typical block diagram of the round robin arbiters. It mainly consists of two barrel shifters, one simple arbiter and one shifter pointer coder. Fig. 4(c) and (d) give the hardware structure of the barrel shifter1 and simple arbiter, respectively. The barrel shifter2 has a similar hardware structure with barrel shifter1 but offers opposite shift direction. The shifter pointer coder generates the shift pointer for two barrel shifters according the previous grant. The detailed working mechanism of the round robin arbiter will not described in this paper because it has been researched in many papers [15] [16] [17]. For clarity, the description of input ports scheduling by DAA is illustrated in fig. 5 by pseudo-C language.

Following the scheduling method described in fig. 5, DAA assigns high priorities for input ports whose buffers are full, and priorities are changed dynamically according to the Buffer_full signal. When a buffer of input port is not full, the corresponding buffer full signal will be disabled, and then the input port will lose its high priority. However, with the advance of time, some light-load input ports will accumulate more and more packets if they have no chance to occupy the desired output port. When buffers of these input ports become full, they will be given high priority to connect with their desired output ports. Even for some input ports whose buffers are never full, they will also have a chance to occupy their desired output port by the appropriate \(T_{\text{threshold}}\). This scheduling method avoids the starvation problem thoroughly.

### 4 Results

#### 4.1 Performance evaluation

In this section, A cycle-accurate NoC simulator implemented in SystemC is used to evaluate the performance
of DAA for different traffics. In order to demonstrate the effectiveness of DAA, latency and throughput are chosen as performance metrics of NoC. We compare the performance of NoC based on round robin arbiter [4] (RRA-NoC), lottery arbiter (LA-NoC) [12] and the proposed arbiter (DAA-NoC). All simulations are carried in a 4×4 mesh network for 20000 cycles with X-Y routing and wormhole switching. Four traffic patterns are simulated including three synthetic traffic patterns (Uniform, Bit-complement and Transpose) [18] and one real benchmark (VOPD application) [19]. For synthetic traffic patterns, each packet contains 4~8 flits randomly and for real benchmark, the number of flits in each packet is determined by the bandwidth requirement of the VOPD application.

Latency is defined as the time (in clock cycles) that elapses between the occurrence of a header flit injection into a network at the source node and the occurrence of a tail flit reception at the destination node. We use the average latency, \( L \), as a performance metric like follows:

\[
L = \frac{1}{K} \sum_{i=1}^{K} L_i
\]

where \( K \) is the total number of packets reaching their destination nodes and \( L_i \) is the latency (cycles) of packet \( i \). We define throughput \( TP \) as follows:

\[
TP = \frac{\text{total received flits}}{\text{number of PEs} \cdot \text{total cycles}}
\]

For each traffic scenario, we first give the average packet latency with various packet injection rates (PIRs). Fig. 6 shows the packet latencies for NoCs with different arbiters under the four traffic patterns.

As shown in fig. 6, RRA-NoC, DAA-NoC and LA-NoC perform almost the same when the networks are under light to moderate packet injection rate because there are few full buffers. So, most input ports have an equal chance to occupy the desired output ports. Although some input ports have less chance to win the desired output ports, they have free buffer slots for holding the incoming packets that results in no harm to packet latency. When the networks start to approach saturation, DAA-NoC provides much lower latency compared with RRA-NoC and LA-NoC. For example, when packet injection rate is 0.013 under VOPD pattern, the latencies of RRA-NoC, LA-NoC and DA-NoC are 19 cycles, 18 cycles and 14 cycles respectively. As the packet injection rate is increase to 0.019 under VOPD pattern, the latencies of RRA-NoC, LA-NoC and DA-NoC are 376 cycles, 130 cycles and 91 cycles respectively.

Besides that, the saturation PIR of DAA-NoC also shows an improvement no matter under uniform or unbalanced traffic patterns. For above four traffic patterns, the saturation PIR of DAA-NoC increase by 1.5%, 9%, 7.1% and 11.7% respectively, compared with RRA-NoC. Compared to RRA-NoC, LA-NoC performs worse under uniform traffic pattern, but better under the other three traffic patterns. For uniform traffic pattern, lottery arbiter cannot distinguish the high priority input port clearly. Thus, all input ports win the desired output ports based on luck in LA-NoC. The average saturation PIR of LA-NoC is about 4.2% higher than that of RRA-NoC.

After comparing the latencies and saturation PIRs, fig.7 presents the saturation throughput of different NoCs. Considering the four traffic patterns utilized in this work, the average saturation throughputs of LA-NoC and DAA-NoC increase by 8.5% and 13.3% respectively, compared with RRA-NoC. Even under uniform traffic pattern, the DAA-NoC also achieves about 3.2% saturation throughput improvement compared to RRA-NoC.
4.2 Hardware implementation

Hardware overhead is also an important metric during NoC design. In this section, we first compare the characteristics of round robin arbiter, lottery arbiter and the proposed arbiter. All these arbiters are described in verilog HDL, and synthesized by Synopsys Design Compiler tool with 0.18-μm CMOS library. Table 1 shows the size and the timing results of these designs.

As shown in Table 1, among all the designs, round robin arbiter runs the fastest with its the smallest size. However, as we pointed out and showed before, round robin arbiter is not efficient enough for most communications. Lottery arbiter runs the slowest because of the most complex logic. Compared to round robin arbiter, the size and critical path delay of lottery arbiter increase by 322.2% and 58.6%, respectively. Dynamic adaptive arbiter is better than lottery arbiter, but worse than round robin arbiter. It consumes 180.6% additional hardware resources and causes 31% extra critical path delay compared with round robin arbiter.

Arbiters are small components in a router. Although there are five output arbiters in 2D mesh on-chip router, the area of arbiters makes up only a small portion...
of the whole router area. The router using round robin arbiter (RRA-Router), lottery arbiter (LA-Router) and dynamic adaptive arbiter (DAA-Router) are implemented with 200MHz clock frequency. Table 2 gives the detailed area consumptions of these routers.

Table 2: Area overheads for different arbiters in router design

<table>
<thead>
<tr>
<th>design</th>
<th>Area for five arbiters (um2)</th>
<th>Area for the whole router (um2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>RRA-Router</td>
<td>3,250</td>
<td>121,334</td>
</tr>
<tr>
<td>LA-Router</td>
<td>13,675</td>
<td>131,762</td>
</tr>
<tr>
<td>DAA-Router</td>
<td>9,105</td>
<td>127,192</td>
</tr>
</tbody>
</table>

As table 2 shows, arbiters consume approximately 2.7% to 10.4% of the total area in the different routers. Compared with RRA-Router, LA-Router and DAA-Router require 8.6% and 4.8% additional chip area for the whole router, respectively. However, using DAA-Router instead of RRA-Router, DAA-NoC achieves average 7.3% improvement on saturation PIR and 13.3% improvement on saturation throughput at the cost of additional 4.8% chip area overhead. Thus, the area overhead of the proposed arbiter for on-chip router is acceptable. Compared with LA-Router, DAA-Router pays less area overhead but achieves more performance improvement for NoC. Besides that, the maximum frequencies of these three routers are similar because the critical paths in these routers are determined by the buffer read/write control logics and arbiters have no influence on critical path delay.

5 Conclusions

In this paper, we proposed a dynamic adaptive arbiter based on round robin mechanism. The proposed arbiter detects buffer status of input ports in every clock cycle and adjusts priority of each input port dynamically. It can authorize the input port for transferring data preferentially if the buffer of port is full. Under uniform traffic and non-uniform traffic patterns in NoC, we compared the performance and hardware overhead of NoC based on round robin arbiter, lottery arbiter and the proposed arbiter. The comparison results showed that the proposed arbiter can improve the performance of NoC with an affordable hardware overhead.

Acknowledgments

The research work in this paper is financially supported by National Natural Science Fund (Grant No.61201244), Key Project of Chinese Ministry of Education (Grant No.210080) and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant No. 12KJA51002 ).

References

13. J. Kim et al., A Gracefully Degrading and Energy-Efficient Modular Router Architecture for On-Chip


Arrived: 03. 01. 2013
Accepted: 08. 05. 2013