MINIMALLY BUFFERED ROUTER USING WEIGHTED DEFLECTION ROUTING FOR MESH NETWORK ON CHIP

Simi Zerine Sleeba and Mini M.G.

Department of Electronics Engineering, Government Model Engineering College,
Cochin University of Science and Technology
Kochi, India

ABSTRACT

The scalability, modularity and massive parallelism exhibited by Network on chip(NoC) interconnects make them highly suitable for the inter core communication framework of multiprocessor system-on-chip (MPSoC) designs. Routers play the most vital role in transferring flits through the network, hence efficient microarchitecture and cost effective routing algorithms are highly essential for modern NoC routers. Elimination of buffers and deflection routing help to achieve energy and area efficiency of these routers. The advantages of bufferless and buffered designs can be combined by using a minimum number of side buffers to store a fraction of deflection flits in the router. In this paper, we propose a routing algorithm based on weighted deflection of flits for minimally buffered deflection routers. Evaluations on 4x4 and 8x8 mesh NoC using synthetic workloads as well as benchmark applications demonstrate that deflection rate and average network latency are significantly reduced in comparison with the state of the art NoC routers. Performance analysis of the newly proposed algorithm shows that the network saturation point improves by 26% compared to earlier designs in this domain.

KEYWORDS

Network on Chip, Deflection routing, Minimal buffering, Average Latency

1. INTRODUCTION

As technology scaling reduces the feature size to nanolevels, a large number of intellectual property cores are being integrated on to a single chip[1]. Such chips can execute applications that demand extensive amounts of parallel processing. Hence there is an ever increasing demand for on chip interconnects capable of handling huge amount of data. Traditional bus based interconnects are no longer suitable for MPSoCs with hundreds and thousands of cores. NoCs emerge as a promising design choice for realising efficient on chip interconnections as they largely alleviate the limitations of bus based interconnects.

Generally, MPSoCs employ a regular mesh topology as in Figure 1. Each Processing Element (PE) is connected to a local switching element called router(R) by means of a network interface. Each PE is either a processor core (with built-in L1 cache) or a slice of shared L2 cache. Requests

DOI : 10.5121/vlsic.2015.6306

for data transfer between processor and shared cache inject packets into the network. These packets start from source node and traverse multiple routers and links before reaching the destination node. Each router has five input ports and five output ports, four of them are connected to neighbouring routers in the east, west, north and south directions. The fifth input port accepts packets injected from the PE to the network and output port ejects packets destined to the PE from the network. The packets arriving at the input ports of a router move through various functional units and get access to output ports depending on their routing choice and prioritisation.

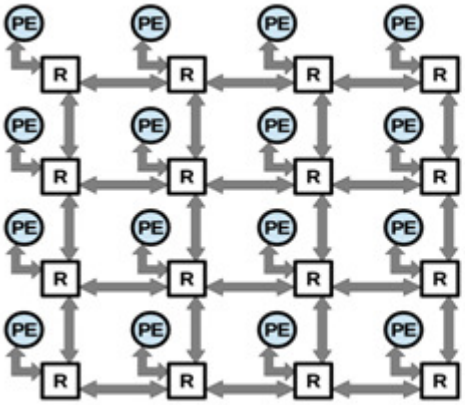

Figure 1. Mesh Interconnection Topology

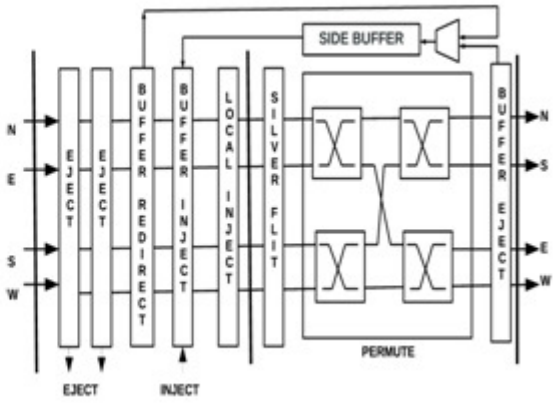

Figure 2. MinBD router microarchitecture[8]

In the traditional virtual channel router (VCR), large number of buffers are employed at the input ports so that flits belonging to different packet transmissions can proceed simultaneously through the same physical channel[2][3]. Packets that cannot find a productive output port from the router are temporarily buffered. Thus buffers prevent unnecessary wastage of link bandwidth and increase saturation throughput of the network. But these buffers consume significant amount of dynamic power when active and static power when idle[4][5]. Recent research findings indicate that static power could constitute 80% to 90% of interconnect power in future systems[6][7]. Hence buffer size is an important parameter that affects area, power and performance of NoCs.

The buffered routers provide an overprovisioned buffer size to accomodate the worst case network traffic. Real applications are found to have a low injection rate[8][9] compared to synthetic traffic and in such situations, buffer utilisation is very less. Hence the concept of bufferless and minimally buffered router in NoC are gaining acceptability as they can deliver similar performance with a lesser power consumption. In these routers, packets are broken down into smaller flow control units or flits which are independently routed. In bufferless deflection routers, when more than one incoming flit competes for the same output port, one wins and traverses through the desired output port, the others are assigned non productive ports ie. such flits are deflected. So eventhough power and area due to buffers is minimum in bufferless designs, the flit deflection rate is high and network saturates earlier compared to buffered routers. Minimally buffered routers combine the advantages of buffered and bufferless designs by buffering a small fraction of the deflection flits, achieving performance close to that of NoCs with buffered routers. The area and power consumed by the side buffers are not significantly higher than that of bufferless designs.

In this work, we propose a novel weighted deflection algorithm for output port selection in minimally buffered routers for mesh NoCs. From evaluations, we observe that the proposed algorithm outperforms the native routing method of MinBD in terms of reduced deflection rate and average flit latency. Reduction in deflection rate leads to reduced dynamic power consumption due to unproductive flit movement through the network.

The rest of the paper is organised as follows. Sections 2 and 3 give an overview of the existing deflection routers employing minimal buffering and the motivation behind this work. Section 4 and 5 explain the Weighted Deflection routing algorithm and router pipeline in detail. Section 6 describes the evaluation method. We consolidate our experimental results in Section 7 and then we conclude the paper.

2. RELATED WORK

First among the network on chip routers is the buffered virtual channel router that attains the required network performance by providing large number of virtual channels and buffers [2] [3]. Bufferless routers are first proposed in [10][11]. Bufferless routing mechanism is either based on dropping and retransmission of packets or on deflection of packets. SCARAB [12] uses the policy of packet dropping which incurs a high retransmission overhead. BLESS [13] proposes a bufferless router microarchitecture using deflection routing. In a deflection router, all flits arriving at the input ports pass through one of the output ports from the router pipeline [14]. BLESS routers use sequential allocation of output ports to flits which are sorted on the basis of age; this increases the critical path length inside the router and makes the router slow. A much superior router microarchitecture is that of CHIPPER [15] which employs parallel port allocation of flits. CHIPPER is faster compared to BLESS but exhibits higher flit deflection rate. In our earlier work, WeDBless [16],we use a routing policy which reduces deflection rate significantly by prioritising output ports for a flit based on Deflection Weights(DW) and ranking flits inside the router based on Weighted Deflection Count (WDC).

MinBD [8] which is the first minimally buffered router design employs a small side buffer to store a few number of deflection flits in the router. DEBAR [9] and SLIDER [17] also use minimal buffering at the routers along with innovative methods for flit ejection, injection and buffering. Another work using minimal buffering inside router is mentioned in [18]. The study of various NoCs using deflection routing comes out with the opinion that a priority based deflection policy which uses global or history related criteria is most suited for boosting the performance of networks [19].

3. MOTIVATION

The MinBD router is efficient in terms of lower deflection rate compared to CHIPPER and consumes lesser power and area compared to virtual channel router. In deflection routers, each incoming flit is transferred to an output port or a side buffer in two cycles of operation (we use two cycle routers). Deadlock problem does not arise in deflection routers since cyclic dependency of resources will not occur [13]. In bufferless routers, it is equally important to ensure livelock freedom. Both CHIPPER and MinBD use the golden flit priority scheme to resolve livelock problem. In this scheme, flits belonging to one packet are marked as the highest priority flits (golden flits) in the entire network and it is guaranteed to win a productive output port at each router until the packet finally reaches the destination and golden status is passed on to another

packet. This scheme is not very efficient because more that 90% of the flits are delivered without becoming golden. The conflict for port allocation of non golden flits is resolved randomly. MinBD uses an additional silver flit priority for prioritising a flit within a router. We analyse certain drawbacks of MinBD which motivate us to propose a new router using weighted deflection algorithm.

3.1. Problem of complex flit prioritisation scheme

The golden priority scheme requires that all routers in the network incorporate a functional unit for golden flit identification while most of them donot recieve golden flits. Silver flit scheme is also inefficient since it cannot ensure that high priority flits in one router get similar priority in the succeeding routers. This leads to increased number of deflections per flit resulting in early saturation of the network. The silver flit prioritisation unit also increases the critical path delay inside router as can be seen from Figure 2.

3.2. Inefficiency of dimension order routing

MinBD uses the Permutation Deflection Network(PDN) proposed in CHIPPER [15] for allocating output port to flits. Computation of desired output port is done by XY routing method which computes only one productive output for a flit. In a mesh network, majority of the flits have two productive output ports in a router. If one of the output ports chosen by the flit is in demand by a higher priority flit, the lower priority flit can occupy the next productive output port. But this capability of PDN is not utilised by the existing routing algorithm in MinBD. Hence we need a routing algorithm which adaptively chooses output ports for flits based on local congestion in the neighbouring routers and global congestion in the network.

3.3. Additional power and area consumption due to duplicate ejection units

As shown in Figure 2, MinBD router uses two ejection units placed one after another for ejecting two locally destined flits in a cycle. This costs additional power and area and also increases the critical path length inside the router.

In this work, we propose a Weighted Deflection routing algorithm for a minimally buffered router which eliminates flit ranking based on complex global timing and provides local priority for a flit inside the router based on the value of Weighted Deflection Level. Flits with higher WDL win the arbitration for output port in a router. Choice of output port for a flit is based on Weighted Distance to Destination which provides more than one productive routes for a flit in majority of the cases. Dual ejection is facilitated through a single ejection port by providing an Eject Buffer which buffers one ejection ready flit. The path delay inside the router caused by duplicate ejection units and the additional wiring overhead is reduced in this design. We also propose a simple logic that precomputes the productive routes of outgoing flits in the succeeding router by adjusting the WDDs of the flit.

4. WEIGHTED DEFLECTION ROUTING ALGORITHM

In this routing method, flits in a router are ranked based on the value of Weighted Deflection Level(WDL)and output ports for a flit are prioritised based on Weighted Distance to Destination(WDD). WDL and WDD values are incorporated in the enhanced header of a flit.

4.1. Weighted Deflection Level

The WDL is 6 bits long and has an initial value of 0 when injected into the network. When a flit makes a hop from one router to another, its WDL value changes. When the flit moves through a productive output port, its WDL is decremented and when it is deflected through non productive port, its WDL is incremented. So frequently deflected flits have higher WDL value compared to less deflected ones. During arbitration for port allocation, if two flits compete for the same output port, the one with higher WDL wins and moves through the desired port whereas the other flit is deflected through the remaining output port. As the flit moves out of a router, its WDL is updated with the new value . Since the WDL of deflected flits are incremented, these flits get higher priority to win productive output ports in subsequent routers. The WDL field maintains an account of the flit's deflection and this priority scheme effectively replaces other methods based on global timing information.

4.2. Weighted Distance to Destination

For a flit, the productive output ports of a router are those which move the flit towards its destination router. Each flit has four WDDs, each of which are 2 bit values. They represent the

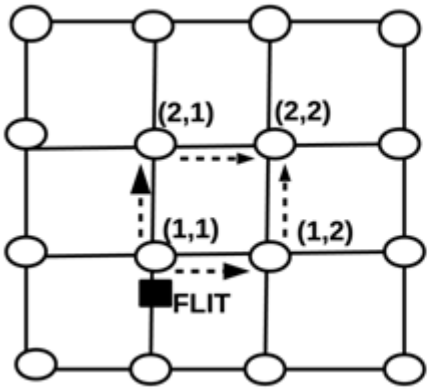

Figure 3. Example 4x4 Mesh Network

Figure 4. Enhanced Flit Header

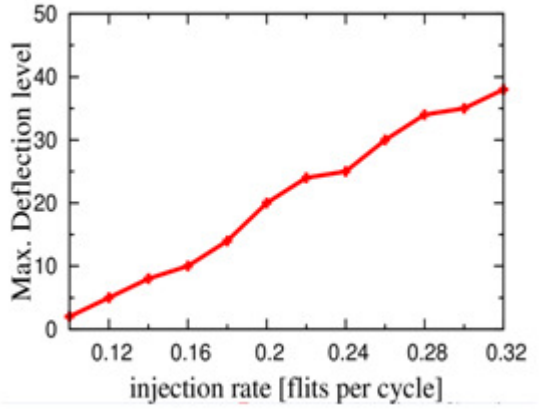

Figure 5: Maximum value of WDL for various flit injection rates for 8x8 mesh

relative distance of the flit's destination from the current router in the North, South, East, West directions. Each WDD can have a value of -1(least distance and highly preferred output port), +1 (second preference output port), +2 (non productive or least preferred output port). WDD of -1,+1 and +2 are coded as 00, 01 and 10 respectively. We use 8 bits in total to accomodate the four WDDs in the enhanced flit header. When a flit is injected into the network the WDDs are calculated based on the position of its destination and stored in its header. During the arbitration for output port selection, each flit gets its routing preference from the WDD values in its header. A flit tries to occupy an output port with the least WDD value. Consider the 4 x 4 mesh NoC shown in Figure 3. For a flit at south input port of router(1,1) whose destination is at router(2,2), north and east directions are equally productive, hence WDD values in these directions will be -1. south and west output ports carry the flit away from the destination and hence WDD values in those directions will be +2 . The 8 bit WDD values of this flit at router(1,1) will be North(00), South(10), East(00), West(10). As a second example, if the flit destination is at router(1,2), only the East output port has a WDD of -1, North and South WDDs are +1 and west WDD is +2.

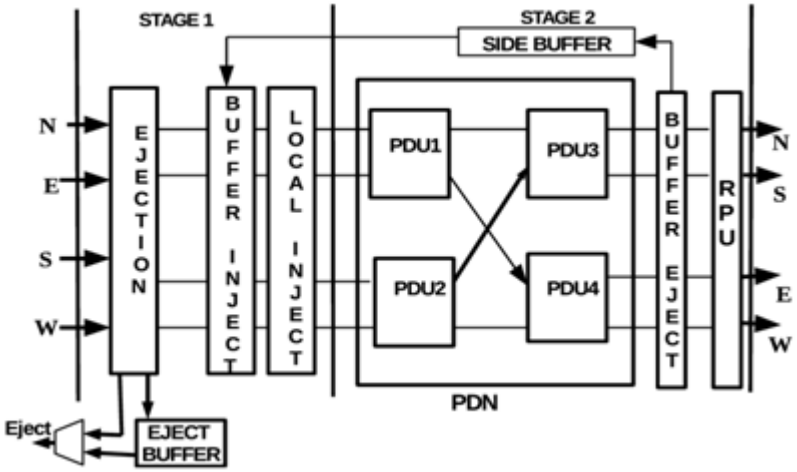

Figure 6. Pipeline diagram of proposed router

When the flit leaves the current router, the flit header is updated with a new set of WDDs corresponding to the next router. This is done in the route precomputation stage at the end of the router pipeline. The WDL of a flit is updated by adding the WDD corresponding to the allocated output port to it. In the first example mentioned above, if north or east output port is alloted to the flit (WDD value is -1), its new WDL value is obtained by adding -1 to the current WDL value. If south or west output port (WDD value is +2) is alloted, new WDL is obtained by adding +2 to the current WDL value. As a flit gets deflected away from the destination, the WDL value increases due to addition of larger WDDs. We maintain the value of WDL between 0 and 26 by decrementing the value as the flit approaches destination and incrementing it as it goes away. This priority scheme based on Weighted Deflection largely increases the probability of the older flits reaching the destination earlier. Thus, the algorithm guarantees livelock freedom of flits since increase in WDL raises the priority of deflected flits and prevents them from being deflected continuously.

4.3. Enhanced Flit Header

In order to incorporate the WDL and WDDs, we use an enhanced header of 14 bits for each flit . The header format is shown in Figure 4. Among the 14 bits; 8 bits represent the four WDDs and 6 bits represent WDL. In Figure 5, we plot the maximum values of WDL obtained for various flit

injection rates in an 8x8 mesh network. The WDL value increases from 0 to 38 as the injection rate is increased. Hence a 6 bit field is sufficient to represent WDL in the flit header.

5. ROUTER PIPELINE

In the router pipeline shown in Figure 6, the incoming flits experience two cycle latency. At the beginning and end of a cycle, flits are stored in pipeline registers which are shown by solid vertical lines in Figure 6. One cycle latency is constituted by ejection unit, flit injection from side buffers and injection from the local PE. The next cycle includes output port allocation using Permutation Deflection Network(PDN), side buffering of deflection flits and route precomputation. Function of each of these units is explained in detail.

5.1. Ejection Unit

The function of the Ejection Unit is to remove flits destined for the local core from the network. The ejection unit consists of a flit identification circuit, an Eject Buffer (EB) and Eject Multiplexer. We refer to the flits destined for the local core as Ejection Ready Flits (ERF). The flit identification circuit identifies the ERF from among the flits arriving at the input ports of the router. We simulate the network using real application traffic and observe that for 10% of the operation cycles, more than one ERF occur. For very few cases, more than one ERF is present in two consecutive cycles. In the newly proposed router, we use the dual ejection mechanism mentioned in [9] and [16] where dual ejection is made possible using a single ejection port. If more than one ERF is present, the one with the highest WDL qualifies for ejection through the local port and one is buffered in the EB. The ERF in the Eject Buffer is ejected from the router in the next cycle. In the next cycle, newly arrived ERF are considered for ejection only if EB is empty. If more than two ERF are present, one is ejected, one is buffered and the others are deflected. But such cases occur rarely as we see in evaluations. Buffering of one ERF in a cycle substantially reduces deflections due to them.

5.2. Injection

There are two injection blocks placed one after another. One is for buffer injection and the other is for injection from local core. In this router, some of the deflection flits are forced into a side buffer after port allocation. The Buffer Injection unit re-injects these flits back into the router pipeline in a subsequent clock cycle, when any of the flits in the four input channels are not present. In a cycle, atmost one flit can be re-injected into the router pieline. A FIFO side buffer is used in the routers ie. the flit that entered the side buffer first will be re-injected into the pipeline first. The local PE can inject flits into the network only when at least one input channel is free even after re injection from side buffer. In a cycle, atmost one flit can be injected from the local core.

5.3. Permutation Deflection Network

We use the PDN structure mentioned in CHIPPER and MinBD for parallel output port allocation. This PDN efficiently maps every input port to every output port of the router. The PDN consists of four 2x2 Permutation Deflection Units (PDU1, PDU2, PDU3, PDU4) arranged in two stages, two units per stage. In the first stage, the North and East input channels are connected to PDU1 and South and West input channels are connected to PDU2. In each permuter block, the flit with higher WDL value has the priority to choose the desired output port and the flit with lower WDL gets the other port. The winning flit chooses the output port with lesser WDD; the WDD values are obtained from the flit header.

At the output of Permutation units PDU3 and PDU4, an output port is alloted to each flit. Most of the flits have two output ports with the same WDD values. So even if a flit does not win an arbitration in the first permutation stage, it can still acquire an equally desirable output port in the second stage. For example, if a flit from the north input port is destined for a router in the south east direction, both south and east output ports are equally productive for it. If the flit fails to get to south output port at PDU1 due to lesser WDL value, it will be deflected to PU4 which is connected to east and west output ports. In this block the flit has a chance of getting the east output port which is also equally desirable for it. If the flit again loses in this arbitration and is deflected through the west output port, its WDL value is incremented by + 2 (WDD of west output port is +2 for this flit). Then the flit has an increased probability of winning the desired output port in the next router.

5.4. Side Buffering of Deflection Flits

In minimally buffered router, after port allocation by the PDN, some of the flits donot win productive output ports to their destinations. Deflection of these flits through the links is reduced by moving some of them into a side buffer. We choose the side buffer size as 4. Atmost one flit can be side buffered in a cycle. If the side buffer is full, flits in the subsequent cycles are deflected.

Table 1. Precomputed WDD values for a flit at router (1,1)

Assigned Output	Destination at (2,2)				Destination at (3,3)			
	N	S	E	W	N	S	E	W
North	1	1	-1	2	-1	2	-1	2
South	-1	2	-1	2	-1	2	-1	2
East	-1	2	1	1	-1	2	-1	2
West	-1	2	-1	2	-1	2	-1	2

5.5. Route Precomputation Unit

The WDL and WDD values in the header of the outgoing flits are recomputed and updated in this unit which is placed after the port allocation stage. The new WDD values correspond to preferred output routes in the next router. Pre-computation of routes using this method is less complex compared to that of MinBD where XY route computation is done for incoming flits. This can be explained with the help of the example in Figure 3. If south or west output port (WDD is +2) is alloted to the sample flit at router (1,1), the flit is being moved further away from destination. So in the south or west neighbouring router of (1,1), the WDD values have no change ie. they need not be re-computed. Table 1 shows the precomputed WDD values for a flit at router(1,1) when destination is at router(2,2) (in columns 2 to 5) and destination is at router(3,3) (in columns 6 to 9). Since the current WDD values of the flit in north, south, east and west directions are 0, +2, 0 and +2 respectively, WDDs need to be re-computed only when a flit having two productive output ports is alloted one of them and destination is at two hops distance from the current router. In all other cases, the new values of WDDs are equal to the previous values.

6. EXPERIMENTAL METHODOLOGY

We use the traditional cycle accurate NoC simulator, Booksim [20] which models the VC router mentioned in [2]. We make suitable modifications on Booksim to model the newly proposed minimally buffered router using weighted deflection (abbreviated as MinBWD in the evaluation section) as mentioned in Section 4. We enhance the flit header with 14 bits for representing WDD

(8 bits) and WDL values (6 bits). We also model the CHIPPER and MinBD architectures as mentioned in the literature and use MinBD as a baseline in our evaluations. The router delay is two cycles in all the three architectures. We conduct all evaluations using single flit packets.

6.1. Synthetic Workloads

We use synthetic traffic patterns to analyse the robustness of our algorithm. Real applications exhibit self throttling, hence synthetic traffic is best suited to test the network saturation point of an algorithm. Uniform random traffic is used to assess the adaptability and load balancing capability of the routing algorithm where as patterns like transpose, bit-complement and tornado are network intensive. Using these synthetic workloads, we conduct simulations on 4x4 and 8x8 mesh network. After sufficient warm up time, we collect the average flit latency, deflection rate and throughput values by varying the flit injection rates from 0 to network saturation point.

Table 2. Percentage of different network injection intensity applications in various benchmark mixes.

Benchmark Mix	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**
% of Low	100	0	0	50	0	31
% of Medium	0	100	0	0	50	31
% of High	0	0	100	50	50	38

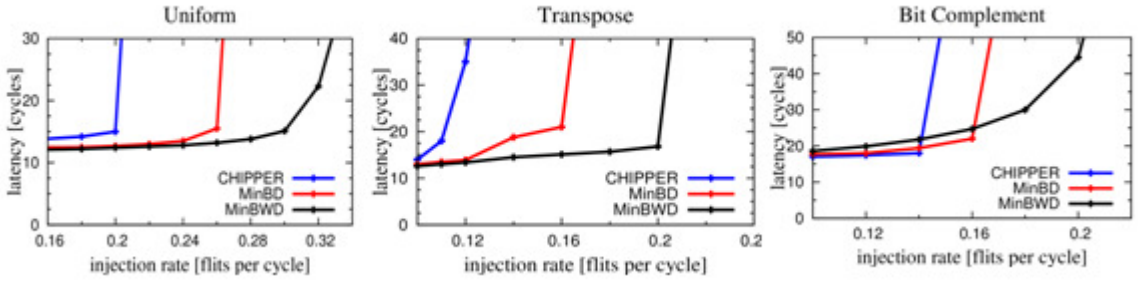

Figure 7. Average flit latency comparison under various synthetic traffic patterns in 8 × 8 mesh network.

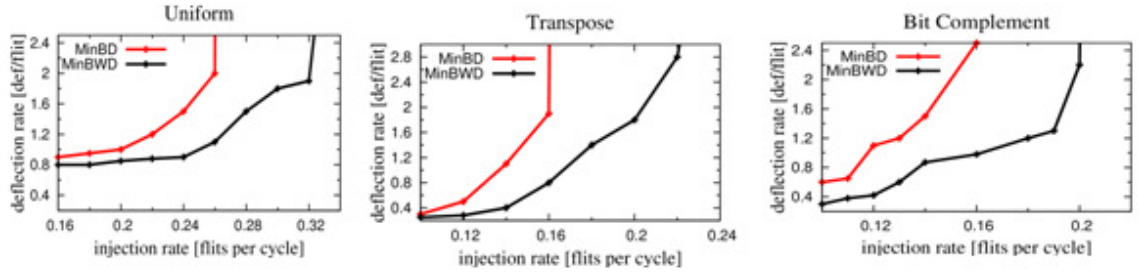

Figure 8: Average flit deflection comparison under various synthetic traffic patterns in 8 × 8 mesh network

6.2. Real Applications

We simulate a 64 (8x8) core multiprocessor system using Multi2sim [21]. Each core has an out-of-order x86 processor with 64KB 4-way set associative L1 cache and 512KB 16-way set associative shared L2 cache. Each core runs one of the applications from SPEC CPU 2006 benchmark application suite [22]. Based on the misses per kilo instructions(MPKI) from L1 cache, these applications are classified as low, medium or high MPKI . We run low MPKI applications like calculix, gombk and h264ref; medium MPKI applications like bwaves,

bzip2,gcc and high MPKI applications such as hmmer, leslie3d and matlab in our simulations. We generate 6 workload mixes by combining applications from the suite as given in Table 2. The network packets generated by these workloads are used in our NoC simulator and average deflection rate and average latency of MinBWD and MinBD are compared.

7. EXPERIMENTAL ANALYSIS

We compare the average deflection rate of MinBWD with that of MinBD for synthetic and real workloads to analyse the role of weighted deflection algorithm in reducing deflections. We also compare the average flit latencies of MinBWD with MinBD as well as a bufferless router, CHIPPER. We analyse average latencies for 4x4 and 8x8 mesh inorder to study the variation in performance with network scaling. Various performance parameters like average deflection rate, average latency and average throughput are analysed in detail.

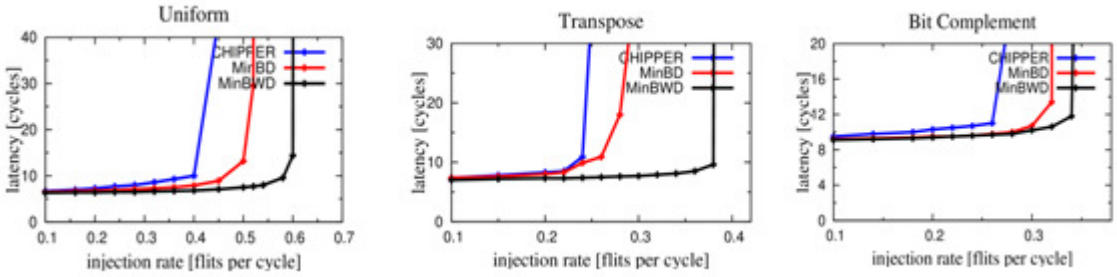

Figure 9. Average flit latency comparison under various synthetic traffic patterns in 4 × 4 mesh network.

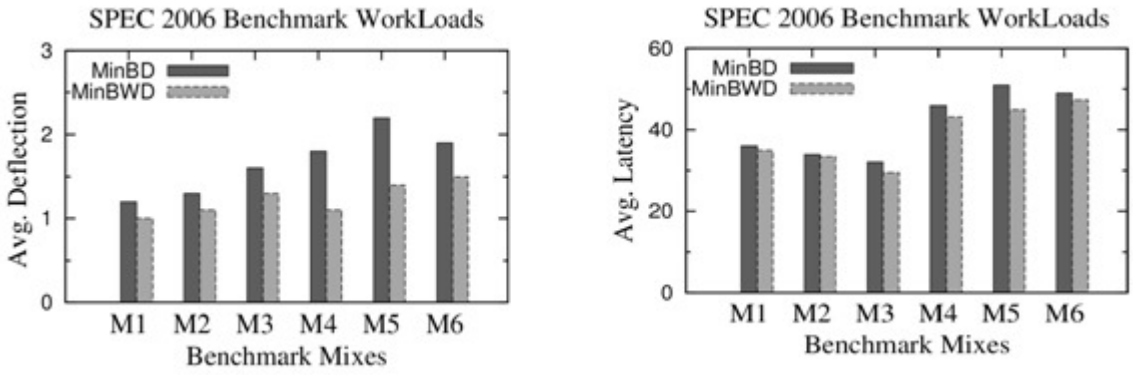

Figure 10: Average deflection rate and latency for real applications.

7.1. Effect on Average Deflection Rate

Deflection of flits through the network causes unnecessary dissipation of dynamic power. The aim of our algorithm is to minimise these deflections and achieve energy efficiency for the NoC. Average deflection rate is computed as the average number of deflections encountered per flit. The average deflection plot for three typical synthetic traffic patterns ie., uniform-random, transpose and bit complement are shown in Figure 8. MinBWD reduces the average deflection rate by 56% compared to MinBD for uniform-random traffic. For non-uniform traffic distribution

like transpose and bit complement, MinBWD reduces deflection rate by 33% and 65% respectively. Unique features of the proposed algorithm which lead to reduction in deflection rate are: (1) providing more than one productive paths for a flit (2) incrementing WDL value for deflected flits giving them high priority to win arbitration in successive routers (3)providing Eject Buffer to store ERF for dual flit ejection. From evaluations using real applications as shown in 10, MinBWD shows a maximum of 36% reduction in deflection rate which is obtained for work-loads with a larger share of high MPKI applications. This explains that the effectiveness of the MinBWD algorithm is more evident under higher network load.

7.2. Average Flit Latency

It is obvious that the reduction in deflection rate obtained for MinBWD reflects in the average flit latency also. Flit latency is measured as the average number of cycles taken by the flit from the time it is generated at the PE upto its ejection at the destination core. A lower value of latency means that the data flits traverse the network in lesser number of cycles and this results in application speed up. We plot the average latencies for synthetic traffic patterns for 8x8 and 4x4 mesh in Figure 7 and Figure 9 respectively. For uniform traffic, the adaptivity and reduced deflections of MinBWD improves the network saturation point by 26% compared to MinBD.

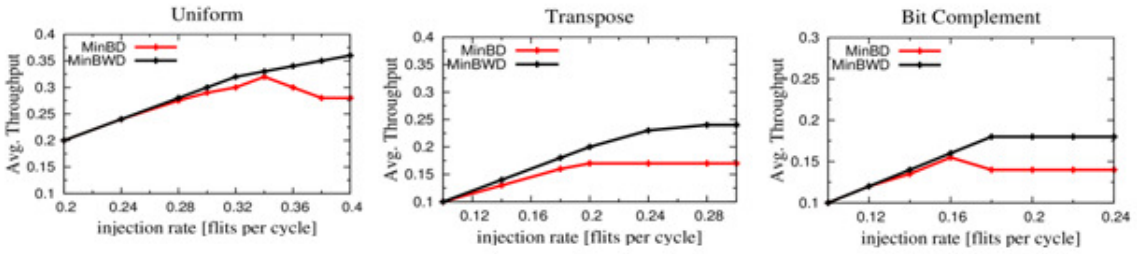

Figure 11: Average throughput comparison under various synthetic traffic patterns in 8×8 mesh network.

7.3. Average Throughput

The throughput is computed as the number of flits ejected from the network per router per cycle. Average throughput variations for various synthetic traffic patterns is shown in Figure 11. At lower injection rates, the throughput delivered by MinBWD and MinBD are similar. MinBD reaches saturation earlier and a dip in throughput is observed for injection rates higher than saturation point.The average throughput of MinBWD is higher for all synthetic workloads and a steady value is maintained further beyond saturation.

8. ROUTER TIMING, POWER AND AREA

We implement each of the functional units of the minimally buffered router using Verilog HDL and synthesise using an EDA tool. We also model additional control logic required for weighted deflection algorithm and compare the overall timing latency of MinBWD and MinBD router architectures. Since MinBWD uses dual ejection with Eject Buffer, the first pipeline stage has 37% gain in timing latency compared to MinBD which uses two ejection Units. Silver flit prioritisation stage of MinBD is not used in our design and the timing delay of PDN in both MinBD and MinBWD is the same. The delay of the Route precomputation unit at the end of the second pipeline stage in MinBWD is lesser than that of silver flit priority block of MinBD, hence we conclude that the timing latency of the second pipeline stage is same for MinBWD and

MinBD. Among the two pipeline stages, second stage dominates the timing latency and hence it decides the router's operating frequency.

We use Orion 2.0 tool[23] for modelling the power and area consumed by MinBWD as well as MinBD. We choose the operating frequency as 1GHz, technology parameter as 65nm and link delay as one cycle. The enhanced flit header in MinBWD uses additional 14 bits for transmitting WDL and WDD values. Hence static power and wiring area at each of the links increases by 10% compared to MinBD. The scheme for dual ejection using single ejection port reduces channel wiring overhead in MinBWD by 26%.

9. CONCLUSION

In this paper we propose a novel routing algorithm for minimally buffered deflection routers which reduces the flit deflection rate and average latency of flits using a unique mechanism of Weighted Deflection. Flits are prioritized using Weighted Deflection Level which is directly proportional to the amount of deflections encountered by the flit. Route computation for a flit based on Weighted Distance to Destination and predetermination of these values for a router contribute significantly to adaptive routing and reducing critical path lengths. From this work, we conclude that the new algorithm combined with minimal buffering for deflection routers promise better performance for NoCs in terms of higher network saturation points and lower latencies.

ACKNOWLEDGEMENTS

This work is supported in part by grant from UGC under MOMA-MANF scheme.

REFERENCES

[1] S. S. I. Association, International Technology Roadmap for Semiconductors Interconnect, Technical Report, 2011 (http://public.itrs.net/).
[2] W. Dally, B. Towles, Principles and Practices of Interconnection Networks, Morgan Kaufmann Publishers Inc., USA, 2003.
[3] W. Dally, Virtual-channel flow control, IEEE Transactions on Parallel and Distributed Systems 3 (2) (1992) 194–205.
[4] G. Kim, J. Kim, S. Yoo, Flexibuffer : Reducing Leakage Power in On-Chip Network Routers, in: Proceedings of the 48th Design Automation Conference, San Diego, USA, 2011, pp. 936–941.
[5] S. Vangal et al., An 80-tile sub-100-W TeraFLOPS processor in 65-nm CMOS, IEEE Journal of Solid State Circuits 43 (1) (2008) 29–41.
[6] M. B. Taylor et al., Evaluation of the Raw Microprocessor: An Exposed-Wire-Delay Architecture for ILP and Streams, in: Proceedings of the 31st Annual International Symposium on Computer Architecture,Munchen, Germany, 2004, pp. 2–13.
[7] Y. Hoskote et al., A 5-GHz Mesh Interconnect for a Teraflops Processor, IEEE Micro 27 (5)(2007) 51–61.
[8] C. Fallin et al., MinBD: Minimally-Buffered Deflection Routing for Energy-Efficient Interconnect, in: Proceedings of the 6th ACM/IEEE International Symposium on Networks on Chip, Lyngby, Denmark, 2012, pp. 1–10.
[9] J. Jose, B. Nayak, M. Mutyam, DeBAR: Deflection Based Adaptive Router with Minimal Buffering, in: Proceedings of the Design Automation and Test in Europe Conference, Grenoble, France, 2013, pp. 1583–1588.
[10] E. Nilsson et al., Load distribution with the proximity congestion awareness in a network-on-chip, in: Proceedings of the Design Automation and Test in Europe Conference, Munich, Germany, 2003, pp. 1126–1127.

[11] P. Baran, On distributed communication networks, IEEE Transactions on Communications 12(1)(1964) 1–9.
[12] M. Hayenga, N. E. Jerger, M. Lipasti, SCARAB: A Single Cycle Adaptive Routing and Bufferless Network, in: Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture,NewYork, USA, 2009, pp. 244–254.
[13] T. Moscibroda, O. Mutlu, A case for bufferless routing in onchip networks, in: Proceedings of the 36th International Symposium on Computer Architecture, Texas, USA, 2009, pp. 196–207.
[14] G. Michelogiannakis et al., Evaluating bufferless flow control for on-chip networks, in: Proceedings of the 4th ACM/IEEE International Symposium on Networks on Chip, 2010, pp. 9–16.
[15] C. Fallin, C. Craik, O. Mutlu, CHIPPER: A Low complexity Bufferless Deflection Router, in: Proceedings of the 17th IEEE International Symposium on High Performance Computer Architecture, Texas, USA, 2011, pp. 144–155.
[16] S. Z. Sleeba, J. Jose, M. G. Mini, WeDBless: Weighted Deflection Bufferless Router for Mesh NoCs, in: Proceedings of the 24th Great Lakes Symposium on VLSI, Houston, Texas, USA, ACM, 2014, pp. 77–78.
[17] B. Nayak, J. Jose, M. Mutyam, SLIDER: Smart Late Injection DEflection Router for Mesh NoCs, in: Proceedings of the 31st IEEE International Conference on Computer Design,Asheville, USA, 2013, pp. 377–383.
[18] J. Lin, X. Lin, L. Tang, Making-a-stop: A new bufferless routing algorithm for on-chip network, Journal of Parallel and Distributed Computing 72 (2012) 515–524.
[19] Z. Lu, M. Zong, A. Jantsch, Evaluation of on-chip networks using deflection routing, in: Proceedings of the 16th Great Lakes Symposium on VLSI, Philadelphia, USA, 2006, pp. 296–301.
[20] N.Jiang et al., Booksim 2.0 user's guide (http://nocs.stanford.edu).
[21] R. Ubal et al., Multi2sim: A simulation framework to evaluate multicore-multithreaded Processors, in: Proceedings of the 26th International Symposium on Computer Architecture and High Performance Computing, Munich, Germany, 2007, pp. 62–68.
[22] SPEC2006 CPU Benchmark suite: (http://www.spec.org).
[23] A. B. Kahng et al., Orion 2.0: A fast and accurate NoC power and area model for early stage design space exploration., in Proceedings of the Design Automation and Test in Europe Conference, Nice, France, 2009, pp. 423–429.

AUTHORS

Simi Zerine Sleeba recieved her B.Tech degree in Electronics & Communication Engineering from Mahatma Gandhi University, India in 1997 and M.Tech degree in VLSI & Embedded Systems from Cochin University of Science and Technology, India in 2010. Currently, she is pursuing her PhD research at Cochin University of Science and Technology. Her research interests include On chip interconnection network architectures and algorithms and low power MPSoC design.

Mini M.G. recieved her B.Tech degree in Electronics & Communication Engineering from Kerala University, M.Tech degree in Digital Electronics and PhD in Image Processing from Cochin University of Science and Technology, India. She is presently working as the Academic Dean at Govt. Model Engineering College, Kochi, India. Her research interests include low power MPSoC design, network on chip architectures and digital image processing.