DAReS: Deflection Aware Rerouting between Subnetworks in Bufferless On-Chip Networks

Rose George Kunthara
Rekha K James
Division of Electronics
School of Engineering, CUSAT
Cochin, India
{rosegeorgekunthara,rekhajames}@cusat.ac.in

Simi Zerine Sleeba
Dept. of Electronics & Communication Engineering
Rajagiri School of Engineering and Technology, Cochin, India
simizs@rajagiritech.edu.in

John Jose
Department of Computer Science and Engineering
Indian Institute of Technology Guwahati, India
johnjose@iitg.ac.in

ABSTRACT

Network on Chip (NoC) is an effective intercommunication structure used in the design of efficient Tiled Chip Multi Processor (TCMP) systems as they improve system performance manifold. Bufferless NoC has emerged as a popular design choice to address area and energy concerns associated with buffered NoC systems. For low to medium injection rate applications, both bufferless and buffered routers show similar network performance. As the network load rises, network performance of bufferless router based designs deteriorate due to increased deflections. This paper proposes a subnetwork based bufferless design, DAReS, to minimize deflections by redirecting contending flit in one subnetwork to unoccupied productive ports of other subnetwork without incurring any extra cycle delay. From evaluations, we observe that our proposed design approach improves network performance by minimizing deflection rate, power dissipation and shows better throughput in comparison to state-of-the-art bufferless router.

CCS CONCEPTS

• **Computer systems organization** → **Multicore architectures**.

KEYWORDS

Network on Chip, deflection routing, bufferless router, average latency, deflection rate, throughput

ACM Reference Format:
Rose George Kunthara, Rekha K James, Simi Zerine Sleeba, and John Jose. 2022. DAReS: Deflection Aware Rerouting between Subnetworks in Bufferless On-Chip Networks. In *Proceedings of the Great Lakes Symposium on VLSI 2022 (GLSVLSI '22), June 6–8, 2022, Irvine, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3526241.3530332

1 INTRODUCTION

The advancements in IC technology have culminated in massive transistor integration. This has led to emergence of Multi-Processor System-on-Chip (MPSoC) where a number of computational cores can be integrated on a single chip. Traditional intercommunication frameworks such as bus based and point-to-point structures cannot effectively handle communication needs as the number of cores is scaled. NoC, a packet switched network, is the most preferred interconnect solution in TCMP design due to its scalability, better parallelism and load handling capability. A regular TCMP based NoC generally employs mesh topology to connect various Processing Elements (PE) [1].

Recent studies show that with rise in number of cores in multicore chips, NoC design can pose a communication bottleneck. Almost 40% of chip power is due to NoC, of which input flit buffers of the router is a major contributor [2], [3]. Thus NoC designs require efficacious router architecture, topology, routing algorithms, power-aware designs, etc. Bufferless NoC is an alternative design approach to overcome rising area and power issues of conventional buffered NoC routers by removing buffers altogether [4], [5]. As there are no buffers to hold flits in a bufferless router, all the incoming flits must move forward through router pipeline and come out through output ports after every cycle. Port conflicts in bufferless routers are handled by deflection routing [4], [5] or by dropping mechanism [6], [7]. Simpler control, absence of flow control, minimal router area and power dissipation are some of the advantages of bufferless designs over buffered counterpart.

BLESS [4] and CHIPPER [5] are the two most popular bufferless deflection routers. BLESS employs an age based sorting network and a sequential port allocation unit to minimize deflections at the expense of a router with larger critical path delay. Compared to BLESS, CHIPPER is considered to be a superior bufferless router as its parallel port allocation technique leads to reduced critical path and smaller hardware footprint. Experiments show that at low to medium network traffic, bufferless deflection routers have performance at par with buffered NoC designs. But as network load rises, network performance suffers due to increased flit deflection rates [4], [8]. In this paper, we address high deflection rates of bufferless designs through **D**eflection **A**ware **Re**routing between **S**ubnetworks (DAReS) that employs CHIPPER router microarchitecture to improve network performance. Thus our subnetwork based design approach reroutes deflected flits in one subnetwork to vacant productive ports of other subnetwork to limit unbounded deflections without any additional cycle latency.

The remaining part of this paper is structured as follows: Section II discusses about related work on multiple on-chip network designs and Section III describes motivation for our proposed design. Section IV gives details about DAReS-CHIPPER design. Section V gives

GLSVLSI '22, June 6–8, 2022, Irvine, CA, USA

ACM ISBN 978-1-4503-9322-5/22/06…$15.00
https://doi.org/10.1145/3526241.3530332

an account of experimental methodology followed and Section VI details the results. Section VII finally concludes the paper.

2 RELATED WORK

The network performance of NoC system can be increased manifold by adoption of a second parallel network, as processor count is scaled. Thus multiple networks can scale bandwidth, improve path diversity and efficiency with negligible impact on area and energy efficiency [9], [10]. TILERA [11], Xeon Phi [12] and OpenPiton [13] are some prominent on-chip network designs which employ multiple networks to separate different message classes to guarantee protocol-level freedom as well as quality of service.

Volos et al. proposes a heterogeneous architecture that employs specialized subnets for each message class in multi-NoC design for better efficiency [14]. Yoon et al. provides a comparative analysis of virtual channel design and multiple physical network approach as two distinct solutions that can improve NoC performance and quality of service [15]. Chameleon, a heterogeneous Multi-NoC design integrates a congestion-aware traffic allocation scheme and a fine-grained power gating algorithm at various levels of granularity simultaneously for better performance [16]. Yadav et al. describes about the design of energy efficient multiple NoCs by customizing hardware implementation for static power with fine-grain traffic distribution exploration [17].

Kwon et al. proposes a universal ordered NoC platform that uses a separate ordering bufferless network to form global request order for delivering requests. Thus in addition to communication, their design can impart ordering of coherence requests [18]. Runahead NoC is a lightweight, lossy network that follows single-cycle hops by combining routing and port arbitration mechanism with link traversal. Their design employs a bufferless NoC, which is lossy in nature such that packets are dropped during contentions [19]. An approximate multiplane NoC (AMNoC) employ buffered subnetwork for lossless transmission of nonapproximable packets whereas a lossy bufferless subnetwork is used to transfer approximable packets for power efficiency [20].

Majority of the work on mulitple networks for NoC systems assigns packets to a fixed network when they are injected into network. Thus packets cannot change the subnetwork during the course of their traversal from source node to destination node. Deflection Containment (DeC) router is a bufferless NoC with bridged multiple subnetworks that minimizes excessive deflections for energy savings. At every router, multiple subnetworks are bridged together to form a bypass ring so as to forward contending flit in one subnetwork to another with an extra latency cycle [21].

3 MOTIVATION

Existing baseline bufferless deflection router, CHIPPER follows golden flit mechanism for flit prioritization. This will lead to increased deflections as only golden flit, which is the highest priority flit, will get required output port and other flits are randomly assigned to remaining output ports. At high network loads, deflections rise rapidly as a large number of low-priority flits are assigned to non-productive output links as depicted in Figure 1. The better path diversity offered by subnetwork based design approach can improve network performance by increasing throughput and reducing deflections considerably.

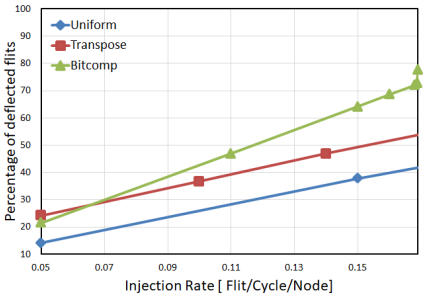

Figure 1: Percentage of flits moving in non-productive paths for various synthetic workloads in a 8×8 mesh based CHIPPER NoC.

We perform simulations on CHIPPER that employ two subnetworks (i.e., S-CHIPPER) for various synthetic traffic workloads. On comparison with CHIPPER, we observe latency and deflection rate reduction of 16% and 59% in S-CHIPPER for uniform traffic. An exhaustive flit flow analysis across S-CHIPPER shows that 76% of traffic is occurring due to flit movement through productive path and flit deflections contribute to the remaining 24%. Also 50% of above flits deflections can be again minimized by reallocating deflected flits in one subnetwork to free productive output links of other subnetwork. In this paper, we make minor alterations in S-CHIPPER router microarchitecture for rerouting deflected flits to improve path diversity.

4 PROPOSED DESIGN

The router pipeline architecture of DAReS-CHIPPER consists of two stages as depicted in Figure 2. Ejection and injection blocks constitute the first stage. The second pipeline stage comprises of Permutation Deflection Unit (PDU) for parallel port allotment and a Subnetwork Reallotment Unit (SRU). The working of various blocks in our proposed router is described below. The internal flit channels carry incoming flits through different functional blocks of router structure. Flits get stored in appropriate pipeline registers at the end of every clock cycle.

4.1 Ejection and Injection Unit

Incoming flits from neighbouring routers arrive at input pipeline register A at the onset of every clock cycle. Ejection unit ejects flits that are destined to local processing core by removing them from internal flit links. Only a single ejection is allowed per router per cycle. When there are multiple flits to be ejected at a router, only the highest priority flit will get ejected and the remaining flit gets deflected to nearby router, which will reach same router in ensuing clock cycles for ejection.

Flit injection happens only when there are vacant internal flit channels as there are no buffers in bufferless routers to hold incoming flits. When there are no free internal flit links, flits will remain queued up at the core level. Injection mechanism of DAReS-CHIPPER works in such a way that injection unit injects flit into subnetworks adaptively depending on the availability of vacant

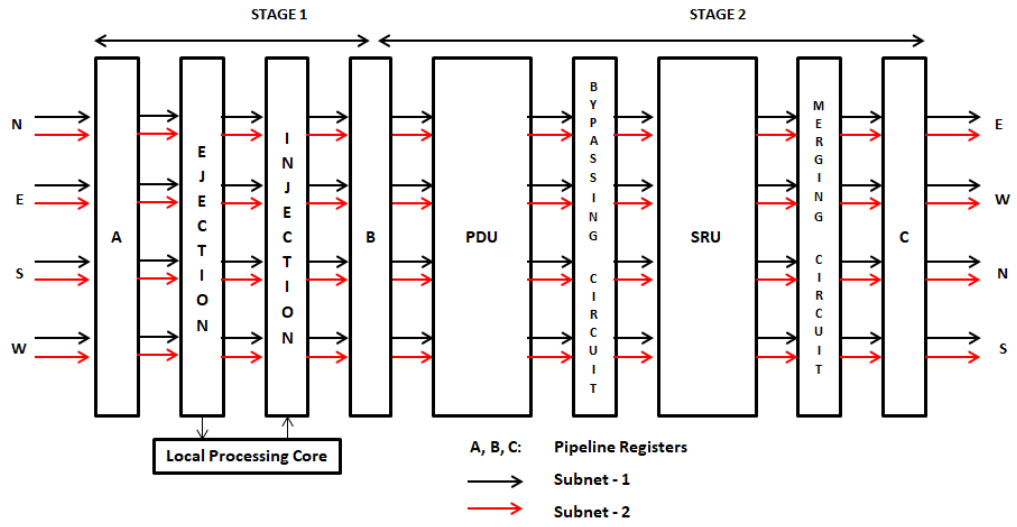

Figure 2: Two stage router pipeline architecture of DAReS-CHIPPER.

internal flit links. XY routing algorithm is employed to find required output port for every flit. Also, the absence of buffers prevent occurrence of deadlock in bufferless design as cyclic dependency of resources will not occur.

4.2 Permutation Deflection Unit (PDU)

Our proposed router uses PDU for parallel output port allocation of flits residing in pipeline register B. Similar to PDN (Permutation Deflection Network) unit in CHIPPER, PDU allocates output ports to every flit on the basis of flit priority and desired output port. As in CHIPPER, our router also follows golden flit prioritization technique to prevent occurrence of livelock. PDU consists of four permuter units for each subnetwork as shown in Figure 3(a). Each permuter unit has two input and two output ports. Incoming flits through North and East ports are connected to permuter P1 whereas incoming flits through South and West internal flit channels are linked to P2. The highest priority flit acquires desired port while remaining flit gets the other port at each permuter. For both P1 and P2, highest priority flit is routed to permuter P3 or P4 based on desired output port and the other flit is directed to remaining permuter unit. Thus each permuter unit efficiently directs inputs to outputs in parallel for both the subnetworks.

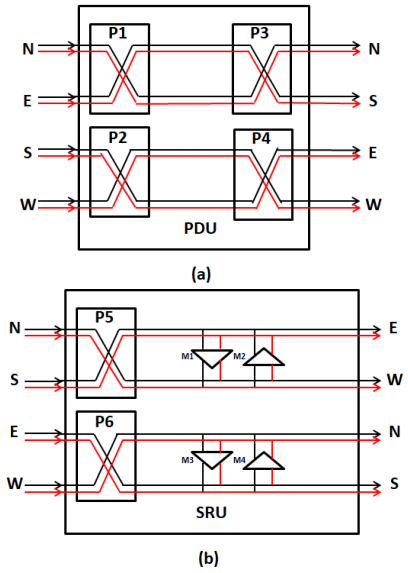

Figure 3: Internal structure of (a) Permutation Deflection Unit (PDU) (b) Subnetwork Reallotment Unit (SRU).

4.3 Subnetwork Reallotment Unit (SRU)

Flits coming out of PDU block enters SRU block through a bypassing circuit as depicted in Figure 2. The purpose of SRU block is to reallocate deflected flits of one subnetwork to unoccupied productive output ports of other subnetwork in the same router without causing any additional cycle delays. This guarantees that flits are not unnecessarily deflected as long as free productive ports are present in any subnetwork of the router. Thus the flits which are already allotted with desired output ports are not altered by rerouting logic of SRU. Flit from PDU block moves into SRU block only if following requirements are satisfied.

1. Flit does not get desired output port.
2. A vacant productive port exists in the other subnetwork within the current router that can route flit to a neighbouring node.

Thus the bypassing circuit affirms that flits that do not follow above requirements evade SRU block by meeting timing constraints. SRU block comprises of two permuter blocks for each subnetwork as given in Figure 3(b). When there are flits that satisfy above mentioned requirements, permuter P5 reallots flits of North and South PDU outputs of first subnetwork to East and West output lines associated with second subnetwork of same router. Similarly, North and South output links of second subnetwork are connected

Table 1: Various benchmark mixes

Mix #	SPEC CPU 2006 Benchmark Mix Pattern	Mix Characteristics
M1	calculix(16) gobmk(16) gromacs(16) h264ref(16)	100% Low MPKI
M2	bwaves(16) bzip2(16) gamess(16) gcc(16)	100% Medium MPKI
M3	hmmer(16) lbm(16) mcf(16) leslie3d(16)	100% High MPKI
M4	hmmer(16) lbm(16) gromacs(16) h264ref(16)	50% Low MPKI, 50% High MPKI
M5	bwaves(16) bzip2(16) mcf(16) leslie3d(16)	50% Medium MPKI, 50% High MPKI
M6	calculix(16) gobmk(16) gamess(16) gcc(16)	50% Low MPKI, 50% Medium MPKI
M7	calculix(10) gromacs(10) bwaves(10) gamess(10) hmmer(12) mcf(12)	31% Low MPKI, 31% Medium MPKI, 38% High MPKI

with East and West output ports of first subnetwork. Permuter P6 is used to reassign East and West PDU outputs of first subnetwork to North and South output lines of second subnetwork while East and West outputs of second subnetwork are associated with North and South output lines of first subnetwork. Flit reallotment across North and South ports or East and West ports of corresponding subnetworks are done using multiplexers (M1, M2, M3, M4) that are connected between permuter output lines as depicted in Figure 3(b). Finally, merging circuit ensures that output links of SRU block are multiplexed with corresponding PDU outputs before flits exit router pipeline.

We consider following example cases to explain functionality of SRU block.

- Case 1: Assume that there are two flits (Flit1 and Flit2) at PDU output corresponding to first subnetwork where North port is assigned to Flit1 and East port is assigned to Flit2. Suppose East is the required productive port for both the flits. Now if there is an unoccupied East port in the second subnetwork, permuter P5 will reassign Flit1 to vacant East port of second subnetwork whereas Flit2 will bypass the SRU unit as it is already assined with its required East port.
- Case 2: Suppose two flits, Flit1 and Flit2 are present at PDU output of second subnetwork where Flit1 is assigned to North port while Flit2 is assigned to South port. Assume South is the required port for both the flits. As Flit2 is already assigned with its required port, it will circumvent over the SRU block. Meanwhile if there is vacant South port in the first subnetwork, Flit1 will be rerouted to unoccupied South port of first subnetwork by activating multiplexer M3.

5 EXPERIMENTAL METHODOLOGY

We make modifications on Booksim [22], an open source cycle accurate input buffered NoC simulator, to prototype a two cycle bufferless deflection router design as suggested in CHIPPER [5]. As every flit is independently routed in deflection routers, requisite control information is added to each flit. We make use of essential reassembly mechanism to handle out-of-order flit delivery. Modifications are done on this baseline bufferless router to model subnet based CHIPPER design and our proposed design for experimental evaluations.

5.1 Synthetic Traffic

The destination router of every generated flit in an NoC is determined by traffic pattern. We employ standard synthetic traffic patterns such as uniform, transpose and bit-complement to show superior performance of DAReS-CHIPPER in an 8×8 mesh network. Network performance parameters like average latency, deflection rate and throughput are captured after adequate warm up time for various traffic patterns by changing injection rate from zero to network saturation.

5.2 Real Traffic

To compare performance of DAReS-CHIPPER against other designs for real application workloads, we use Gem5 simulator [23] to prototype a 64-core multiprocessor setup. Each processing core contains an out-of-order x86 processing unit with 2 levels of cache: 4-way set associative, 64KB private L1 cache with 32 byte block size and a 16-way set associative, 512KB shared distributed L2 cache with 64 byte block size. Every processing core runs a SPEC CPU2006 benchmark application program.

The benchmark applications are categorized into various injection intensity groups based on misses per kilo instructions (MPKI) values computed on a 64KB L1 cache as: Low MPKI (less than 5), Medium MPKI (between 5 and 25) and High MPKI (greater than 25). For our experimental evaluations, we take *calculix, gobmk, gromacs* and *h264ref* under Low MPKI group, *bwaves, bzip2, gamess* and *gcc* under Medium MPKI group, and *hmmer, lbm, leslie3d* and *mcf* under High MPKI group. Based on this categorization, we create 7 categories of workload mixes, each having 64 application instances as shown in Table 1. Network traffic produced by running real workloads is given to Booksim simulator to generate network statistics.

6 RESULTS AND ANALYSIS

We compare and evaluate the performance of our proposed design approach against baseline CHIPPER and subnetwork based CHIPPER. Simulations are performed on 8×8 mesh network for all designs under consideration to analyse the effect on various network parameters.

6.1 Effect on Average Flit Latency

The average flit latency comparison for CHIPPER, S-CHIPPER and DAReS-CHIPPER is shown in Figure 4 for various synthetic traffic patterns. For low injection rate traffic, it can be seen that latency is more or less same for all three designs. This is because deflections are less at low network loads and flits have same number of hop traversals for all the three designs under consideration. Across all synthetic workloads, we can observe that our proposed design

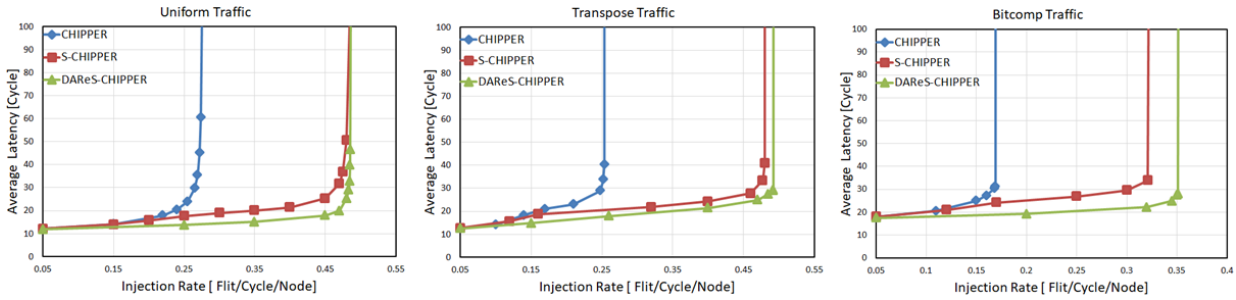

Figure 4: Flit latency comparison for various synthetic workloads in 8×8 mesh network.

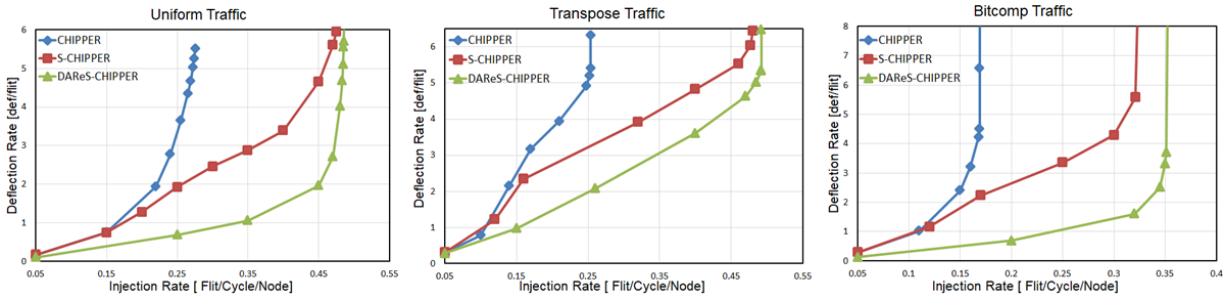

Figure 5: Deflection rate comparison for various synthetic workloads in 8×8 mesh network.

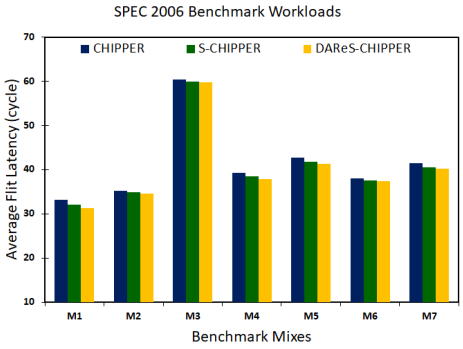

Figure 6: Flit latency comparison for real workloads.

shows lowest latency at higher network loads and also saturates later than other designs. This is because of the extra logic included after permutation deflection unit by which deflected flits are reallocated to unoccupied productive ports of other subnetwork. Thus for high injection rate workloads, DAReS-CHIPPER will be an adequate design option.

From average flit comparison graph for real traffic depicted in Figure 6, it is quite evident that DAReS-CHIPPER has lowest latency across all benchmark mixes. As expected Mixes M3 and M5 have higher latency values as they are formed by high MPKI benchmark programs. Compared to CHIPPER, S-CHIPPER has a latency reduction of 1% - 3% while DAReS-CHIPPER has a latency reduction of 1% - 5.4% across various mixes.

6.2 Effect on Deflection Rate

Deflection rate is calculated as average deflections undergone by a flit. The deflection rate comparison of CHIPPER, S-CHIPPER and DAReS-CHIPPER is given in Figure 5. For all the synthetic workloads, CHIPPER has highest deflection rate due to random port assignment of non-golden flits which lead to more deflections. Compared to CHIPPER, S-CHIPPER has smaller deflection rate as it has several channels which can reduce flit contentions by better path diversity. Our proposed design has lowest deflection rate due to subnetwork reallocation mechanism by which flit deflections are considerably reduced. Figure 7 depicts deflection rate comparison for various real workloads. On comparison with CHIPPER, S-CHIPPER has a deflection rate reduction of 1% - 15% whereas DAReS-CHIPPER shows deflection rate reduction of 28% - 52% for different benchmark mixes. Thus there is considerable reduction in flit deflections due to which more number of flits can successfully reach their destination routers.

6.3 Effect on Throughput

Throughput is calculated as the number of ejected flits per router per cycle. Figure 8 depicts throughput variations for different synthetic traffic patterns. S-CHIPPER and DAReS-CHIPPER have nearly double the throughput of standard CHIPPER design as they employ parallel networks. Also compared to S-CHIPPER, DAReS-CHIPPER shows a throughput improvement of 1% - 8% for various synthetic workloads. This is due to the additional logic which reallocates deflected flits in one subnetwork to vacant productive ports of another subnetwork.

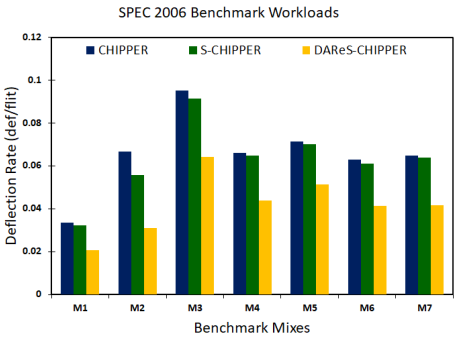

Figure 7: Deflection rate comparison for real workloads.

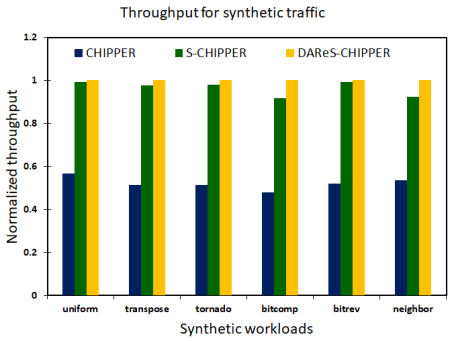

Figure 8: Throughput comparison for synthetic workloads.

6.4 Power and Thermal Analysis

Higher flit deflection rate leads to larger power dissipation across NoC links due to increased network activity. Orion 3.0 tool [26] is used to estimate dynamic power dissipation across inter-router links. We assume 65nm technology with an operating frequency of 1GHz and one cycle inter-router link delay. Dynamic power dissipation of DAReS-CHIPPER decreases by 10% - 35% for various synthetic workloads when compared to S-CHIPPER. In comparison with CHIPPER, DAReS-CHIPPER has a dynamic power reduction of 32% - 54% for similar traffic patterns.

We employ Hotspot 6.0 tool [25] to analyse thermal distribution across NoC. The power traces found with Orion tool is fed into Hotspot tool to estimate transient temperature deviation across all the routers. The empirical analysis show that across central routers DAReS-CHIPPER has a temperature reduction of 6°K and 3°K when compared to CHIPPER and S-CHIPPER respectively.

6.5 Hardware Overhead

We implement pipelined router microarchitecture of CHIPPER, subnetwork based CHIPPER and our proposed router in Verilog HDL and is synthesized using Xilinx Vivado 2018.3 targeted to Xilinx Zynq UltraScale+, XCZU7EV device to obtain router pipeline latency, area and power overhead. The latency of first pipeline stage prevails over second stage for all designs under analysis. S-CHIPPER and DAReS-CHIPPER incur same delay in first stage as similar functional blocks are used in both router architectures. When compared to S-CHIPPER, DAReS-CHIPPER has an extra logic delay of 18% in second stage of router pipeline due to the additional logic block (SRU) after permutation deflection unit. As latency of first stage is more than the second stage, router pipeline frequency of our proposed design will be same as that of S-CHIPPER. The additional rerouting logic also incurs area and power overhead of 4% and 7% respectively when compared to S-CHIPPER.

7 CONCLUSION

Bufferless on-chip networks experience higher deflection rates with increase in network load. A subnetwork based deflection rerouting technique is proposed to improve overall network performance of bufferless mesh NoC. An additional logic unit is inserted after permutation deflection unit to curb deflections by reassigning deflected flits of one subnetwork to unused productive ports of other subnetwork without any extra cycle delays. Hardware overhead of DAReS-CHIPPER is balanced by notable enhancement in network performance with a more distributed traffic across the network.

REFERENCES

[1] W. Dally and B. Towles, *Principles and Practices of Interconnection Networks*, Morgan Kaufmann, USA,2004.
[2] Y. Hoskote et al., "A 5-GHz mesh interconnect for a teraflops processor," in *IEEE MICRO*, vol. 27, no. 5, pp. 51-61, 2007.
[3] F. Alazemi et al., "Routerless network-on-chip," in *HPCA*, 2018.
[4] T. Moscibroda and O. Mutlu, "A case for bufferless routing in on-chip networks," in *ISCA*, pp. 196-207, 2009.
[5] C. Fallin et al., "CHIPPER: A low complexity bufferless deflection router," in *HPCA*, pp. 144-155, 2011.
[6] C. G. Requena et al., "An Efficient Switching Technique for NoCs with Reduced Buffer Requirements," in *ICPADS*, pp. 713-720, 2008.
[7] M. Hayenga et al., "Scarab: A single cycle adaptive routing and bufferless network," in *IEEE MICRO*, 2009.
[8] C. Fallin et al., "MinBD: Minimally-buffered deflection routing for energy-efficient interconnect," in *NOCS*, 2012.
[9] J. Balfour, and W. Dally, "Design tradeoffs for tiled CMP on-chip networks," in *ACM International conference on supercomputing 25th anniversary volume*, pp. 390-401, 2006.
[10] R. Das et al., "Catnap: Energy proportional multiple network-on-chip," in *ACM SIGARCH Computer Architecture News*, 41(3), 2013, pp. 320-331.
[11] D. Wentzlaff et al., "On-Chip Interconnection Architecture of the Tile Processor," in *IEEE MICRO*, vol. 27, no. 5, pp. 15-31, Sept.-Oct. 2007.
[12] A. Sodani et al., "Knights Landing: Second-Generation Intel Xeon Phi Product," in *IEEE MICRO*, vol. 36, no. 2, pp. 34-46, Mar.-Apr. 2016.
[13] J. Balkind et al., "OpenPiton: An open source manycore research framework," in *ACM SIGPLAN Notices*, pp. 217-232, 2016.
[14] S. Volos et al., "CCNoC: Specializing On-Chip Interconnects for Energy Efficiency in Cache-Coherent Servers," in *NOCS*, 2012, pp. 67-74.
[15] Yoon et al., "Virtual channels and multiple physical networks: Two alternatives to improve NoC performance," in IEEE TCAD, 32, no. 12, pp. 1906-1919, 2013.
[16] J. Wu et al., "Energy-efficient NoC with multi-granularity power optimization" in *The journal of Supercomputing* 73, no. 4, pp. 1654-1671, 2017.
[17] S. Yadav et al., "Multiple-NoC Exploration and Customization for Energy Efficient Traffic Distribution," in *IFIP/IEEE VLSI-SOC*, 2020.
[18] Kwon et al., "A universal ordered NoC design platform for shared-memory MP-SoC," in *ICCAD*, pp. 697-704, 2015.
[19] Z. Li et al., "The runahead network-on-chip," in *HPCA*, pp. 333-344, 2016.
[20] L. Wang et al., "An approximate multiplane network-on-chip," in *DATE*, 2020, pp. 234-239.
[21] X. Xiang et al., "Bufferless Network-on-Chips With Bridged Multiple Subnetworks for Deflection Reduction and Energy Savings," in *IEEE TC*, vol. 69, no. 4, pp. 577-590, April 2020.
[22] N. Jiang et al., "A Detailed and Flexible Cycle-Accurate Network-on-Chip Simulator," in *ISPASS*, 2013.
[23] N. Binkert et al., "The gem5 simulator," *SIGARCH Computer Architecture News*, vol. 39, no. 2, pp. 1-7, 2011.
[24] "SPEC2006 CPU benchmark suite," http://www.spec.org.
[25] W. Huang et al., "Compact Thermal Modeling for Temperature-Aware Design," in *DAC*, 2004.
[26] A. B. Kahng et al., "Orion 2.0: a fast and accurate NoC power and area model for early stage design space exploration,"in *IEEE TVLSI*, 20, (1), pp. 191-196, 2012.