Conference Proceedings

Title: An efficient 2D router architecture for extending the performance of inhomogeneous 3D NoC-based multi-core architectures

Creators: Opoku Agyeman, M. and Zong, W.

Example citation: Opoku Agyeman, M. and Zong, W. (2016) An efficient 2D router architecture for extending the performance of inhomogeneous 3D NoC-based multi-core architectures. In: *SBAC-PAD Workshop on Applications for Multi-Core Architectures.* USA: IEEE . (Accepted)

Version: Accepted version

An Efficient 2D Router Architecture for Extending the Performance of Inhomogeneous 3D NoC-Based Multi-Core Architectures

Michael Opoku Agyeman[1],Wen Zong[2]
[1]Department of Computing and Immersive Technologies, University of Northampton, UK. Email: Michael.OpokuAgyeman@northampton.ac.uk
[2]Department of Computer Science and Engineering, The Chinese University of Hong Kong, HK SAR

Abstract—**To meet the performance and scalability demands of the fast-paced technological growth towards exascale and Big-Data processing with the performance bottleneck of conventional metal based interconnects, alternative interconnect fabrics such as inhomogeneous three dimensional integrated Network-on-Chip (3D NoC) has emanated as a cost-effective solution for emerging multi-core design. However, these interconnects trade-off optimized performance for cost by restricting the number of area and power hungry 3D routers. Consequently, in this paper, we propose a low-latency adaptive router with a low-complexity single-cycle bypassing mechanism to alleviate the performance degradation due to the slow 2D routers in inhomogeneous 3D NoCs. By combining the low-complexity bypassing technique with adaptive routing, the proposed router is able to balance the traffic in the network to reduce the average packet latency under various traffic loads. Simulation shows that, the proposed router can reduce the average packet delay by an average of** 45% **in 3D NoCs.**

I. INTRODUCTION

Recently, three-dimensional Network-on-Chip (3D NoC) has been proposed to solve the communication demands of modern multi-core architecture design. However, 3D ICs have alignment issues along with low yield and high temperature dissipation, which affect the reliability of the implemented on-chip cores. Specifically, the 3D routers have a larger area and power consumptions than a 2D router with a similar architecture. Moreover, Through Silicon Via (TSV) which has been accepted as a viable inter-layer wiring technique has a complex and expensive manufacturing process [1]. To optimize the performance and manufacturing cost of 3D NoCs with minimal distortion to the modularity, inhomogeneous architectures have been proposed to combine 2D and 3D routers in 3D NoCs [2]–[4]. Several inhomogeneous 3D architectures focusing on different NoC router architectures, minimal hop-count between 2D and 3D routers in each layer, and uniform distribution of 2D and 3D routers have been proposed [5]. However, due to the limited number of 3D routers and vertical links, inhomogeneous 3D NoCs have a performance trade-off.

While inhomogenous 3D NoCs promises to resolve the poor scalability and performance issues of conventional traditional 2D NoCs, the multi-hop among the long wired 2D routers is still a performance bottleneck. Our goal is to mitigate the performance reduction in inhomogeneous 3D NoCs by proposing an efficient router architecture that accounts of the manufacturing cost in terms of area and power consumption.

We make the following observations within the layers of the 3D NoC: 1) On paths without contention, a packet traverses through routers' pipeline without stall and experiences solely the zero-load delay. Contrarily, on congested paths, a packet needs to compete for NoC resources to proceed. 2) A packet that fails to acquire desired resources is stalled, adding a non-deterministic queuing delay to its packet latency. To improve the performance of inhomogeneous 3D NoCs, both router pipeline and queuing delay should be minimized to efficiently reduce the communication delay of multi-core workload. We exploit the uneven utilization of resources under different traffic intensities and replace the 2D routers in inhomogeneous 3D NoCs with an efficient router architecture that employs bypassing and adaptive routing to significant reduce the average packet delays within the layers of the NoC.

In this paper, we propose SlideAcross, a 3-stage adaptive virtual channel (VC) compatible router with single-cycle bypassing mechanism to meet the communication needs of emerging communication fabrics for modern multi-core architecture. The proposed router integrates adaptive routing with low-latency bypassing in a cost-effective way to overcome the drawbacks of existing adaptive routing and low-latency architectures. A packet that takes advantage of bypass datapaths does not need to wait for crossbar setup and experiences single-cycle delay per hop (including link traversal). If bypassing is not available or not applicable for a packet, the packet is stored in an input buffer, and then follows the adaptive routing pipeline. SlideAcross uses a simple VC allocation (VA) scheme to allow VA be performed after switch allocation (SA) in the same cycle non-speculatively.

The rest of the paper is organized as follows. Section II introduces the background and existing efforts to reduce network latency. Section III discusses the state-of-the-art high performance inhomogeneous 3D NoCs and formulates the problem of improving their performance. Section IV shows the overview of the proposed SlideAcross router. Section V presents the adaptive routing pipeline and the proposed VA and deadlock avoidance technique. Section VI, evaluates the performance of the proposed router. In Section VII, we conclude this work.

II. RELATED WORK

Adaptive routing helps reduce packet queuing delay effectively [6]. However, at low-load conditions, adaptive routing has negligible improvement. Moreover, the required pipeline of adaptive routing has higher complexity. On the other hand, some operations can be executed in parallel. [7] does VA and SA in parallel speculatively, and prioritizes non-speculative packets in SA to increase resource utilization. [8] exploits the abundant bandwidth inside router and multicast flits to output ports speculatively rather than waiting for SA. Parallel processing of a packet can also happen on different routers with the help of control flits which goes ahead of data flits [9]. SA for a flit is done based on the control flit while the data flit is traversing the link on previous router. When the data flit arrives, it can bypass SA stages and goes directly to ST. However, the sideband network for control flits introduces extra wiring and power overhead.

Low-swing signaling [10] and asynchronous link [11], [12] have been adopted in NoC to allow multiple-hop traversal in one cycle. Low-swing signaling has poor bandwidth density, and asynchronous link can have signal skew issues due to interference [13]. In chips operating at high frequency the signal traversal length can be limited due to the small clock cycle. The simplicity of ring topology allows router to have simple and low-latency micro-architecture [14]. In 2D mesh topology, dimension-sliced router (DSR) is proposed to reduce router cost and latency [15]. DSR abandons the input buffers of routers and also decouples datapath of the two dimensions to reduce cost. Intra-dimension traversal in DSR incurs single-cycle delay (including link traversal).

CMP workloads require low-latency adaptive routers to reduce communication latency and also requires VC support in NoC to achieve message isolation. Existing work reduces latency of NoC routers by either enhancing the classic router [16] or developing simpler micro-architectures. Approaches such as lookaheads [17] add wiring and logic complexity to routers, and increase NoC's area overhead and power consumption. Speculation [7] does not reduce the worst case pipeline delay.

Simple NoC micro-architectures like [14], [15] are not adaptive and have no VCs. High radix routers [18], [19] usually have higher serialization delay, and do not work well under adversarial traffic [12]. NoC with multi-hop traversal in single cycle capability such as SMART [12] shows significant latency reduction. However, such feature may not sustain in chips operating at high frequency or with long links (*e.g.* hierarchical topology), or in the combination of two. In contrast, single-cycle-per-hop routers are still good candidates for such scenarios. In this paper, we propose a 3-stage non-speculative adaptive VC router for CMP and develop a low-complexity single-cycle bypassing mechanism to reduce low-load latency without using sidebanded lookahead signals.

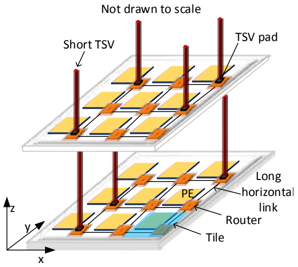


Fig. 1. Inhomogeneous 3D NoC

III. INHOMOGENEOUS 3D NOCS

A. 3D Network-on-Chip Architectures

The evolution of SoC design to the third dimension offers a lot of opportunities such as integration of inhomogeneous cores which results in several challenges [20]. A 3D router has a larger area and power consumption than a 2D router with similar architectures [21]. Particularly, the 7 port symmetric router has an area and power overhead of 36% and 158%, respectively compared to a conventional 5 port router [5].

Existing inhomogeneous architectures (Fig. 1) [2], [22]–[25] however, do not consider the dynamics of application traffic load in their architectures generation. Applications in such 3D NoCs are not optimized as communication bandwidth and performance constraints of the applications were not considered in the architecture generation. To resolve this, a systematic approach for generating inhomogeneous 3D NoC architectures where the TSV and buffer utilization of the given application are exploited is proposed in [3]. Though inhomogeneous 3D NoC architectures, reduce the number of power (up to 67%) and area hungry 3D routers as well as the number of TSVs, they inhibit the total performance of the NoC. Particularly, by reducing the number of 3D routers to 25% the average hop-count and delay can increase up to 28% and an average of 45%, respectively in 4×4×4 3D NoCs [26], [27]. This paper aims to resolve the performance degradation introduced by the heterogeneity in router architectures of existing inhomogeneous 3D NoCs while maintaining the small area of the 2D routers by introducing bypass links and adaptivity to escape the intra-layer multi-hop and congested regions.

B. Problem Formulation

Routing packets along the long horizontal links to access the limited number of vertical links in inhomogeneous 3D NoCs may result in significant packet latency due to the buffering, hop-by-hop traversal and the destribution of the 3D nodes. To this end, the *M/M/1/B* queueing model is employed as a closed-form expression or the average packet latency. Here the number of nodes in a transmission queue can be derived as [28]:

$$\zeta^{h}_{(i_k,i_{k+1})} = \frac{\rho^{h}_{i_k,i_{k+1}} + (\beta_{h\rho_{i_k,i_{k+1}}} - \beta_h - 1)(\rho^{h}_{i_k,i_{k+1}})^{\beta_h+1}}{(\rho^{h}_{i_k,i_{k+1}} - 1)((\rho^{h}_{i_k,i_{k+1}})^{\beta_h+1} - 1)}, \tag{1}$$

Adopting Little's results [29], the average time spent over any path q_{ij} can be given by:

$$T^h_{q_{ij}} = \sum_{i_k,i_{k+1}\in q_{ij}} \left(\frac{\zeta^h_{(i_k,i_{k+1})}}{\lambda^h_{i_k,i_{k+1}}(1-P^{block}_{(i_k,ik+1),h})}\right) \tag{2}$$

where $P^{block}_{(i_k,ik+1),h}$ is the blocking probability:

$$P^{block}_{(i_k,ik+1),h} = \frac{((\rho^h_{i_k,i_{k+1}})^{\beta_h})(\rho^h_{i_k,i_{k+1}}-1)}{(\rho^h_{i_k,i_{k+1}})^{\beta_{h+1}}-1}, \tag{3}$$

β_h is the relative buffer length of the router with respect to application with β_h= $\frac{\beta}{L_h}$. L_h and β are the packet length [flits] of application h and buffer size, respectively. $\rho^h_{i_k,i_{k+1}}$ is the intensity of the traffic at link (i_k, i_{k+1}) which is given by:

$$\rho^h_{i_k,i_{k+1}} = \frac{\lambda^h_{i_k,i_{k+1}}}{\mu^h_{i_k,i_{k+1}}} \tag{4}$$

where $\lambda_{i_k,i_{k+1}}$ is the aggregated incoming traffic of application h [packets/s] traversing link (i_k, i_{k+1}) including the traffic from previous nodes that are either directly or indirectly connected to the node. $\mu^h_{i_k,i_{k+1}}$[packets/s] is the service rate, which is expressed as:

$$\mu^h_{i_k,i_{k+1}} = \frac{\log(1+\gamma_{k,k+1})}{8L_h}. \tag{5}$$

Here, W is the available bandwidth at node i_k.

Hence to solve the problem of improving the performance efficiency of inhomogeneous 3D NoCs, our objective is to design a router micro-architecture that is able to reduce the average time T packets spend along the slow horizontal wires such that:

$$\min_{\forall(i_k,i_{k+1},h)} (T^h_{q_{ij}}) \tag{6}$$

subject to:

$$\psi = (A_{new_R} - A_{old_R}) + (P_{new_R} - P_{old_R}) \tag{7}$$

where

$$\psi \leq min \tag{8}$$

where new_R and old_R are the proposed new router and conventional 2D routers micro-architecture, respectively. A_x and P_x represents the area and power consumption of router x. The most efficient design has a ψ = 0 and hence the minimum (min in Eq 8) must be as close to zero as possible.

IV. PROPOSED ROUTER ARCHITECTURE

We proposed to replace the slow 2D routers with, SlideAcross, an adaptive virtual-channel router equipped with single-cycle bypass datapaths. SlideAcross contains two types of datapaths, one optimized for low latency, the other optimized for adaptivity. Fig. 2(a) shows the micro-architecture of proposed router. Input buffers are connected to output ports through the crossbar which forms the adaptive routing pipeline. For adaptive routing, each input port has a dedicated VC for bypassing, Slide Virtual Channel (SVC) buffer reserved for fast packet traversal. The crossbar is composed of input multiplexers and output multiplexers to be cost effective [7], [30]. The red bold arrow in the figure is a bypass datapath that connects *West* input link directly to the *East* output multiplexer. *Mux2* connects the red arrow with *East* output port when there's no request for *East* output port, forming the one of the pre-setup intra-dimension bypass datapaths.

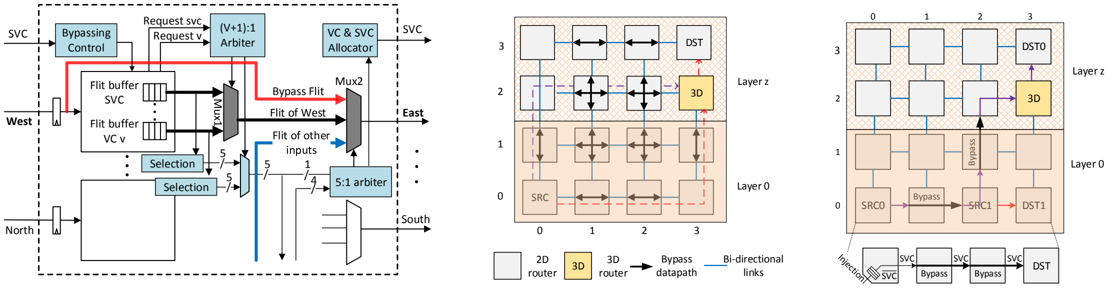
Control modules are colored with blue in this figure, including the bypassing control, input multiplexer arbiter, output multiplexer arbiter, VC allocator and SVC allocator. VC and SVC allocator absorb the arbitration result of the 5:1 arbiter (SA-II) and allocate VC and SVC tag to the winning packet accordingly. Selection units automatically select the less congested for buffered packets by masking the congested output port in the output port request vector.

The bypass datapath is developed from the single-cycle-per-hop router DSR [15]. Packets traversing through the bypass datapath maintains its progress on current dimension and incurs a single-cycle delay. The adaptive datapath is similar to existing adaptive routers [6] but with a simplified VA scheme. We modify the VA by forcing packets retain their original VC. Moreover, VA is performed after SA in the same cycle non-speculatively. There is a single-bit tag in each flit to notify a downstream router if this flit can utilize the bypass datapath. If the tag bit is set, upon receiving the flit, a router will try to use bypass datapath to transmit the flit, otherwise the router lets it follow the adaptive routing datapath. Packets from all VCs have chances to utilize the bypass datapath using the SVC tagging mechanism proposed in this paper.

A. *In-Layer Intra-dimension Bypass Datapath*

We add a set of bypass paths on top of a 2D VC router to achieve single-cycle intra-dimension traversal to provide shorter paths between 2D and 3D routers. During SA if an output port receives no requests (indicating that the output port will be idle in next cycle), the output port is connected directly to the input channel of the opposite side in a router. Thus, an incoming packet can directly traverse to the corresponding output port without waiting for switch allocation. We assume a 128-bit 1.5mm long bypass datapath (including crossbar and link). DSENT [31] reports that the bypass datapath can satisfy a delay constraint of 0.2ns with proper repeater insertion. Traversing through a bypass path skips the buffering procedure as well as multi-stage allocation procedures and incurs a single-cycle delay.

We use an example to demonstrate how these pre-setup datapaths can be utilized to transmit any packets. The thick arrows in Fig. 2(b) represent the pre-setup bypass paths in a 4×4×z mesh network under zero-load conditions. Suppose a packet is injected to router *SRC* and targets destination *DST* on layer *z*. At *SRC*, the router selects an output direction for the packet according to congestion status. If it chooses *East* output, the packet will go to Router (1,0,0) in next hop. Router (1,0,0) has a bypass path from *West* to *East*. Hence, on Router (1,0,0), this packet can go to Router (2,0,0) directly without arbitration. The same procedure of bypassing works on Router (2,0,0) which sends the packet to Router (3,0,0). On Router



(a) Router micro-architecture

(b) Bypassing technique under zero-load conditions

(c) Example of packet path

Fig. 2. Improved 2D Router for Inhomogeneous 3D NoCs

(3,0,0), the packet needs to make a turn, and is buffered and then sent to *North* output through the crossbar (in DSR [15] it is through a shared intermediate buffer). The *South* to *North* bypass path on Router (3,1,0) sends the packet to 3D Router (3,2,0) for interlayer traversal to the destination in layer z. The red dashed line shows the complete path for this packet if it selects *East* output at *SRC*, which is actually an XYZ routing path. Similarly, if *North* output is selected on *SRC*, the path of this packet will be the purple dashed line which is a YXZ routing path.

The bypass datapath applies DoR on packets so to utilizes the pre-setup intra-dimension crossbar connections. Utilizing these bypass paths skips the long adaptive routing pipeline and effectively reduces packet delay at low-loads.

B. Dedicated Virtual Channel for Bypassing

An incoming flit may belong to an arbitrary VC. Deciding whether a flit can bypass current router. The VC must be decoded and then the availability of corresponding credits for downstream router must be checked. Here, we assume the flit retains its VC ID when bypassing. Suppose the VC ID of a received flit is vc, and the output port of DoR is o. If the following two conditions are met, the received flit can bypass current router in one cycle. Bypassing must not cause overshooting to the destination (minimal routing). Moreover, the vc at output o must be idle (ensuring a successful VA). Implementing this bypassing logic requires using the VC ID as the input to index corresponding information. This control logic will inevitably increase the critical path length of bypassing logic compared to the one in [15] due to VC decoding. Preliminary synthesis result shows that the path delay for this decision making on 16 VCs is 0.1ns on 45nm standard cell library. In this implementation, the decision making speed slows down as the number of VC increases.

To speedup this process, we introduce a dedicated VC for bypassing. Suppose the special VC introduced is called *slide virtual channel* (SVC). We now only perform bypassing for flits belonging to SVC. To check if SVC flit can bypass current router, a router only needs to check if SVC of output o is idle. Bypass decision making is faster because we do not need to use VC ID as index to absorb credit information or other information. The processing speed is invariant to the number of VCs. Therefore the path delay for the SVC logic is reduced to 0.05ns using the same 45nm standard cell library.

Only SVC packets are considered for bypassing, and there is also dedicated buffer space reserved for SVC in each router. This design reduces the complexity of bypass decision making. Bypassing with SVC is faster, and more importantly, invariant to the number of VCs. Adding an extra VC does not necessarily increase buffer space in router because most NoC routers use shared buffer between VCs [32].

C. SVC Tagging Mechanism

Packets of SVC can enjoy bypassing. In this work, all VCs have the chance to be tagged with SVC to reduce overall packet delay and increase link utilization. SVC can be allocated to any packet that wins the output port. All packets are injected to network with the SVC tag being zero. A router updates the SVC tag of a packet after it wins the output port. A packet has the first chance to be tagged with SVC when leaving the its source router. Each output port (excluding ejection port) has a tagging unit. The principle to tag a head flit with SVC is simple, meeting the following two conditions:1) The SVC tag of the output port is not assigned to any packet. 2) The SVC buffer at corresponding downstream is empty. Otherwise, the SVC tag bit is set to zero. The two rules work together as a lightweight SVC allocator which assigns the SVC tag to packets. A body flit of a packet follows the SVC tag of its head flit, and the tail flit releases the possession of the SVC tag of that output port.

Fig. 2(c) shows an example of SVC tagging. As long as a packet can win an output port, it can be tagged with SVC if the two conditions for SVC tagging are met. Proposed SVC and its tagging mechanism is a fast and scalable solution for single-cycle bypassing in virtual-channel adaptive routers. SVC tagging is transparent to CPUs or upper level applications. Any packet that wins switch allocation on current router has a chance to to be tagged with SVC. The packet tagged SVC can enjoy bypassing in the next hop.

V. ADAPTIVE ROUTING

Packet that cannot utilize bypass datapath are routed through the adaptive routing datapath in SlideAcross. If a received

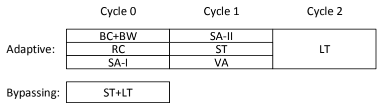


Fig. 3. Bypassing and adaptive routing pipeline stages

packet cannot bypass current router, it is written to input buffer (BW) and meanwhile route computation (RC) is performed. Adaptive selection is done automatically by masking the congested output port similar to [6]. The crossbar in this router is implemented using two sets of multiplexers like those in [7] to be cost-effective. The SA process thus contains the arbitration for multiplexer of input buffer (SA-I) and that of the output port (SA-II). The winner of SA-II will then transmit a flit to the output link (LT). An idle VC of the output port is also assigned to the the SA-II winner which forms VA procedure. Fig. 3 shows the pipeline of this adaptive routing process. To support bypassing, upon receiving a packet, we need to perform bypassing control (BC) to determine if the packet should be written to buffer, so there is a BC procedure before BW operation in pipeline. If the packet can bypass current router, it follows the single stage bypassing traversal (ST+LT). Fig. 2(c) shows an example of how adaptive routing and bypassing determine the path of a packet.

A. Deadlock Avoidance

Routing in this router is minimal and fully adaptive and is hence prone to be deadlock. To break the cycles in resource dependency graph [16], we require at least two VCs (*VC0* and *VC1*) in each VN. A packet is assigned to a VC during injection according to the position of its destination. Packets with destination locating at the left and right side of its source node are assigned to *VC0* and *VC1* respectively. If a packet's destination is on the same column with the source node, the packet can be assigned to either VC randomly or according to congestion status. As the routing is minimal, turns in neither VC form a circle. So both *VC0* and *VC1* are deadlock-free.

Packets from all VCs have chances to use the SVC buffer, so SVC can potentially be a shared media that chains the turns of *VC0* and *VC1* to form a circle. To prevent this deadlock configuration, we only allow one packet to stay in SVC buffer. This is achieved by controlling SVC tagging, a head flit will be tagged SVC only when the downstream SVC buffer is empty as imposed by the second rule in section IV-C. Because SVC contains at most one packet, it will not chain up the turns of different VCs. The rules above all together guarantee a deadlock-free network. Sharing the SVC is also protocol-level deadlock-free. Suppose all SVCs are occupied by a certain class of message, a message of other classes can still reach their destination through the normal VCs, which are guaranteed to drain. So there will not be dependency between different classes of messages making the network protocol-level deadlock-free.

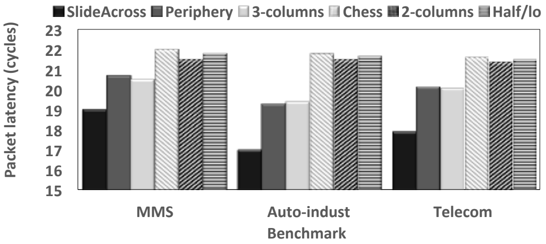


Fig. 4. Average packet latency of various inhomogeneous 3D NoCs

VI. EVALUATION

In order to evaluate the performance of the proposed bypassing technique in 3D NoCs and to facilitate correlation with existing work, an extended version of $Worm\_sim$, a cycle-accurate NoC simulator [3] is used. Our extended simulator employs wormhole packet switching flow control to accurately simulate 3D NoCs with any configuration of 3D and 2D routers. In the simulation, a fixed packet size of 5 flits is used in the NoC model. In order to evaluate the performance sustainability and energy of the NoC in real-world scenarios: a complex multimedia traffic (MMS) [33], Auto-indust and Telecom (from the E3S benchmark suite) [34] and an AV (Audio-visual) benchmark [35]. The setup is running for a warm-up period of 2000 cycles and performance statistics are collected after a simulation length of $200,000$ simulation cycles. Hence, by introducing different delay models of 2D and 3D routers in the system, we have compared the average packet latency.

To analyse the performance benefits of inhomogeneous 3D NoC architectures implemented with the proposed SlideAcross router, Branch-and-Bound [3] mapping algorithm algorithm is used to map the applications to various inhomogeneous architectures for comparison. For inhomogeneous 3D NoCs with bypass techniques (a.k.a SlideAcross), we replace the conventional 2D routers by the SlideAcross routers. Hence packets destined for other layers could be routed to 3D router either via the bypass links or by the proposed deadlock-free adaptive routing to get access to the destination layer. Moreover, in the destination layer, packets can either exploit the bypass links or adaptive routing depending on the traffic conditions of the network, to the destination node. For a fair comparison the performance efficient Buffer-Nearest Vertical Hub (Buff NVH) [24] routing algorithm which always forwards packets towards the 3D whose path provides the maximum output channel buffer space on the current core and has the closest Cartesian x,y to the current core as well as minimum Manhattan distance to the destination, is employed for routing in existing inhomogeneous 3D NoCs.

Fig. 4 shows the average packet latency of various inhomogeneous architectures under different realistic benchmarks. By bypassing the links between 2D and 3D layers, SlideAcross has reduced average hop-count with less traffic loads within the layers and exploits the performance benefits of short vertical wires for inter-layer traversal. Hence, inhomogeneous 3D NoCs with SlideAcross have much lower packet latencies compared to existing inhomogeneous architectures. This is expected as though existing hop-count based inhomogeneous architectures have evenly distributed 3D routers and the efficient Buff NVH adaptive routing algorithm is adopted, the extra delays introduced by the multi-hops between 2D routers and 3D routers reduces the performance of the NoC by increasing delays in the network which consequently causes contention.

VII. CONCLUSION

In this paper, an efficient router with reduced low-load latency is proposed to improve the performance of inhomogeneous 3D NoCs. The proposed router architecture has a cost-effective dual datapath design that is able to minimize packet delay under both low-loads and high loads. A fast bypass datapath is proposed to alleviate the performance degradation due to multi-hops along the long horizontal wires. Furthermore, a deadlock-free adaptive routing algorithm is proposed to avoid congested paths when the NoC is heavily loaded with traffic. The performance effect of replacing conventional 2D routers with the proposed router architecture in inhomogeneous 3D NoCs is evaluated by cycle-accurate simulations. The experimental results show significant reductions in the average packet delay compared to existing high-performance inhomogeneous 3D NoCs even when efficient adaptive routing is used.

REFERENCES

[1] D. Velenis, M. Stucchi, E. Marinissen, B. Swinnen, and E. Beyne, "Impact of 3d design choices on manufacturing cost," in *IEEE International Conference on 3D System Integration (3DIC)*, 2009, pp. 1 – 5.

[2] M. O. Agyeman, A. Ahmadinia, and A. Shahrabi, "Heterogeneous 3d network-on-chip architectures: area and power aware design techniques," *Journal of Circuits, Systems and Computers*, vol. 22, no. 4, p. 1350016, 2013.

[3] M. O. Agyeman, A. Ahmadinia, and N. Bagherzadeh, "Performance and energy aware inhomogeneous 3d networks-on-chip architecture generation," *IEEE Transactions on Parallel and Distributed Systems*, vol. PP, no. 99, pp. 1–1, 2015.

[4] M. O. Agyeman, "A study of optimization techniques for 3d networks-on-chip architectures for low power and high performance applications," *International Journal of Computer Applications*, vol. 121, no. 99, pp. 1–8, 2015.

[5] J. Kim, C. Nicopoulos, D. Park, R. Das, Y. Xie, V. Narayanan, M. S. Yousif, and C. R. Das, "A novel dimensionally-decomposed router for on-chip communication in 3D architectures," *SIGARCH Comput. Archit. News*, vol. 35, no. 2, pp. 138–149, 2007.

[6] J. Kim *et al.*, "A low latency router supporting adaptivity for on-chip interconnects," in *Proceedings of DAC*. ACM, 2005, pp. 559–564.

[7] L.-S. Peh and W. J. Dally, "A delay model and speculative architecture for pipelined routers," in *Proceedings of HPCA*. IEEE, 2001, pp. 255–266.

[8] Y. He *et al.*, "Mcrouter: Multicast within a router for high performance network-on-chips," in *Proceedings of PACT*. IEEE, 2013, pp. 319–330.

[9] T. Krishna *et al.*, "Swift: A swing-reduced interconnect for a token-based network-on-chip in 90nm cmos," in *Proceedings of ICCD*. IEEE, 2010, pp. 439–446.

[10] C.-H. O. Chen *et al.*, "Smart: a single-cycle reconfigurable noc for soc applications," in *Proceedings of DATE*. EDA Consortium, 2013, pp. 338–343.

[11] T. N. Jain, P. V. Gratz, A. Sprintson, and G. Choi, "Asynchronous bypass channels: Improving performance for multi-synchronous nocs," in *Proceedings of NOCS*. IEEE, 2010, pp. 51–58.

[12] T. Krishna *et al.*, "Breaking the on-chip latency barrier using smart," in *Proceedings of HPCA*. IEEE, 2013, pp. 378–389.

[13] R. Kumar, Y. S. Yang, and G. Choi, "Intra-flit skew reduction for asynchronous bypass channel in nocs," in *Proceedings of VLSI Design*. IEEE, 2011, pp. 238–243.

[14] R. Ausavarungnirun *et al.*, "Design and evaluation of hierarchical rings with deflection routing," in *Proceedings of SBAC-PAD*. IEEE, 2014, pp. 230–237.

[15] J. Kim, "Low-cost router microarchitecture for on-chip networks," in *Proceedings of Micro*. ACM, 2009, pp. 255–266.

[16] W. Dally and B. Towles, *Principles and Practices of Interconnection Networks*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003.

[17] S. Park, T. Krishna, C.-H. Chen, B. Daya, A. Chandrakasan, and L.-S. Peh, "Approaching the theoretical limits of a mesh noc with a 16-node chip prototype in 45nm soi," in *Proceedings of the 49th Annual Design Automation Conference*. ACM, 2012, pp. 398–405.

[18] J. Kim, J. Balfour, and W. Dally, "Flattened butterfly topology for on-chip networks," in *Proceedings of the 40th Micro*. IEEE Computer Society, 2007, pp. 172–182.

[19] B. Grot, J. Hestness, S. W. Keckler, and O. Mutlu, "Express cube topologies for on-chip interconnects," in *HPCA 2009*. IEEE, 2009, pp. 163–174.

[20] P. Vivet, Y. Thonnart, R. Lemaire, E. Beigne, C. Bernard, F. Darve, D. Lattard, I. Miro-Panades, C. Santos, F. Clermidy, S. Cheramy, F. Petrot, E. Flamand, and J. Michailos, "8.1 a 4x4x2 homogeneous scalable 3d network-on-chip circuit with 326mflit/s 0.66pj/b robust and fault-tolerant asynchronous 3d links," in *IEEE International Solid-State Circuits Conference (ISSCC)*, 2016, pp. 146–147.

[21] L. P. Carloni, P. Pande, and Y. Xie, "Networks-on-chip in emerging interconnect paradigms: Advantages and challenges," in *ACM/IEEE International Symposium on Networks-on-Chip (NOCS)*, 2009, pp. 93–102.

[22] T. C. Xu, G. Schley, P. Liljeberg, M. Radetzki, J. Plosila, and H. Tenhunen, "Optimal placement of vertical connections in 3d network-on-chip," *Journal of Systems Architecture*, vol. 59, no. 7, pp. 441 – 454, 2013.

[23] C. Liu, L. Zhang, Y. Han, and X. Li, "Vertical interconnects squeezing in symmetric 3D mesh Network-on-Chip," in *Asia and South Pacific Design Automation Conference (ASP-DAC)*, 2011, pp. 357 –362.

[24] M. O. Agyeman, A. Ahmadinia, and A. Shahrabi, "Efficient routing techniques in heterogeneous 3d networks-on-chip," *Parallel Computing*, no. 0, pp. –, 2013.

[25] A. Bose, P. Ghosal, and S. P. Mohanty, "A low latency scalable 3d noc using bft topology with table based uniform routing," in *IEEE Computer Society Annual Symposium on VLSI (ISVLSI),*, 2014, pp. 136–141.

[26] M. OpokuAgyeman, *3D Networks-on-Chip Architecture Optimization for Low Power Design*. LAP LAMBERT Academic Publishing, 2015.

[27] M. O. Agyeman and A. Ahmadinia, "Optimised application specific architecture generation and mapping approach for heterogeneous 3d networks-on-chip," in *IEEE International Conference on Computational Science and Engineering*, 2013, pp. 794–801.

[28] J. MacGregor Smith, "Properties and performance modelling of finite buffer m/g/1/k networks," *Comput. Oper. Res.*, vol. 38, no. 4, pp. 740–754, 2011.

[29] S. K. Bose, *An Introduction to Queuing Systems*. Springer Press, 2001.

[30] R. Mullins, A. West, and S. Moore, "Low-latency virtual-channel routers for on-chip networks," in *Proceedings of ISCA*, vol. 32. IEEE, 2004, p. 188.

[31] C. Sun *et al.*, "Dsent-a tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling," in *Proceedings of NOCS*. IEEE, 2012, pp. 201–210.

[32] D. U. Becker, "Efficient microarchitecture for network-on-chip routers," Ph.D. dissertation, Stanford University, 2012.

[33] J. Hu and R. Marculescu, "Energy- and performance-aware mapping for regular NoC architectures," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, no. 4, pp. 551–562, 2005.

[34] R. Dick, "Embedded system synthesis benchmarks suite(e3s)," ziyang.eecs.umich.edu/dickrp/e3s.

[35] V. Dumitriu and G. Khan, "Throughput-oriented noc topology generation and analysis for high performance socs," *VLSI*, vol. 17, no. 10, pp. 1433 –1446, 2009.