Enabling Vertical Wormhole Switching in 3D NoC-Bus Hybrid Systems

Changlin Chen, Marius Enachescu, and Sorin D. Cotofana
Computer Engineering, Delft University of Technology, Delft, the Netherlands
Email: {C.Chen-2, M.Enachescu, S.D.Cotofana}@tudelft.nl

***Abstract*—In Networks-on-Chip (NoC) systems Wormhole Switching (WS) enables lower packet transmission latency and requires less silicon real estate than the Packet Switching (PS). However, enabling vertical WS in conventional 3D NoC-Bus hybrid systems requires a large amount of TSVs, which have low yield in state of the art 3D stacking technology. In this paper, we alleviate this issue by introducing a Bus Virtual Channel (VC) Allocation (BVA) mechanism, which assigns to at most one cross layer packet a free input VC in its target router before injecting it into the bus. In this way, a routing path is reserved by the head flit, and the rest of the packet flits can be WS transmitted through the vertical buses. Given that VC allocation is performed only once per packet per hop BVA can be performed in such a way that it doesn't become a system bottleneck. We evaluated our proposal with both synthetic and real application traffics and the experimental results indicate that when vertical WS is implemented, the bus critical path length is reduced by at least 31% and the average packet transmission latency is reduced by at least 22%, when compared with conventional pipelined bus or TDMA bus based systems. Moreover, the area cost and power consumption of the output buffer incident to the bus are reduced by 47% and 43%, respectively.**

Keywords*—3D NoC-Bus hybrid system; Wormhole Switching; Bus virtual channel allocation; Pipelined bus.

I. Introduction

To efficiently utilize the ever-increasing transistor budget brought by transistor dimension scaling and packaging innovation, multi-/many-core systems integrated with a Networks-on-Chip (NoC) communication infrastructure are widely investigated [1]. However, in 2 dimensional (2D) chips, the NoC induced system performance enhancement is still limited due to several aspects, e.g., restricted floor planning choices and large clock tree network [2], [3]. With the emerging of 3 dimensional (3D) IC stacking various 3D NoC architectures have been proposed [4], which can alleviate those issues. As most 2D NoC principles can be applied to each silicon layer, the main challenge in 3D NoCs relates to the vertical links implementation and utilization. In 3D chips, silicon tiers are vertically stacked and connected with Through Silicon Vias (TSVs) [5], which, when compared with moderate size planar wires, exhibit extremely low data transmission delay, but suffer from low manufacturing yield [6]. Thus in 3D NoC designs one should exploit the benefit of negligible TSV delay while reducing the TSVs amount.

Among state of the art 3D NoC proposals, the 3D NoC-Bus hybrid structure serves this purpose well. When the vertical links are implemented as buses, it is possible to run the buses at higher speed than the routers due to the low delay of TSVs and the simplicity of the bus structure. Moreover, the buses can be shared by multiple routers on each silicon layer to reduce the amount of TSVs. The bus can be accessed by the routers incident to it with a Time Division Multiple Access (TDMA) strategy [7], and can be pipelined to enable concurrent data transmission [3].

It is well known that, when compared with Packet Switching (PS), Wormhole Switching (WS) requires less silicon and enables lower packet transmission latency [8]. In symmetric NoC systems, each router port is solely connected with one neighboring router thus the allocation of one output Virtual Channel (VC), i.e., an input VC in the downstream router, to a packet can be easily done, because the output VCs are only used by packets from the current router and their status can be actively maintained by the output port [9]. However, in a NoC-Bus hybrid system, a vertically traveling packet can be destined to any other layer, and an input VC in the UPDOWN port, i.e., the port connected with the bus, can be competed by packets from different layers. Thus, for each UPDOWN output port, maintaining the VCs status in the UPDOWN input ports in other layers and assigning each cross layer packet a free output VC becomes much more complicated. Due to this existing 3D NoC-Bus hybrid structures postpone the vertical packet VC assignation for the moment when the packet reaches its target layer, i.e., PS instead of WS is utilized for data transmission over the vertical buses.

In this paper, we propose a Bus VC Allocation (BVA) mechanism that enables vertical WS in 3D NoC-Bus hybrid systems. Because the VC allocation is performed only once per packet per hop, the possibility that multiple BVA requests are asserted along the same bus in the same cycle is low. Thus in each cycle, the BVA mechanism forwards at most one request to the packet target router and picks a free input VC there. In this way the routing path is reserved, and the packet flits can be transmitted with the WS technique on the buses. The BVA mechanism is evaluated on a $4 \times 4 \times 4$ 3D NoC-Bus hybrid system with both synthetic and real benchmarks traffic. The experimental results indicate that when vertical WS is implemented, the bus critical path in the pipelined and TDMA bus based hybrid systems are shortened by 31% and 34%, respectively, and the average packet transmission latency are reduced by 22% and 24%, respectively. Moreover, the area cost and power consumption of the output buffer incident to the bus are reduced by 47% and 43%, respectively.

The rest of the paper is organized as follows. Section II presents a brief related work survey. Section III introduces the proposed NoC-Bus hybrid system. Section IV evaluates our proposal and compares it with tightly related work. Section V concludes the presentation.

978-3-9815370-4-8/DATE15/©2015 EDAA

II. Related Work

The most intuitive way to implement a 3D NoC is to simply stack 2D Mesh NoCs and utilize TSVs to connect vertically adjacent routers. Despite its simplicity, such 3D symmetric NoC is not taking advantage of the negligible inter-layer TSV delay [10]. To reduce the TSV amount, Hwang et al. [11] propose to connect only a few number of TSV routers to the vertical link, while the rest just connect with routers on the same silicon layer. This strategy substantially diminishes the TSV count but the vertically connected routers can easily cause communication hot spots and become the system bottleneck.

In [12], Kim et al. propose to utilize TSVs as intra-router connections to implement the 3D crossbar in a 3D Dimensionally-Decomposed router structure. Although the energy-delay product characteristic is enhanced, this approach requires many TSVs, which are much larger than planar wires and have low manufacturing yields [6], thus such designs are not really applicable for state of the art 3D stacking technology.

Noticing that a large proportion of the network traffic takes place between the Processing Units (PUs) and the closest cache memories in the same pillar [7], Feero et al. [2] proposed a ciliated 3D NoC architecture, in which the routers locate on only one layer or a small number of layers, and each router is connected with multiple PUs residing in its Z pillar but on different layers. Such design offers an advantage in terms of energy dissipation when traffic is localized within a pillar. However, it has much lower throughput than 3D symmetric NoC systems and requires a large amount of TSVs.

In [7], Li et al. utilize TSVs to implement dynamic TDMA (dTDMA) buses to achieve one hop data transmission from a source layer to any destination layer. To alleviate the dTDMA bus drawback that it can only be occupied by one source-destination pair at any given time, Ebrahimi et al. [3] proposed a High-performance Inter-layer Bus Structure (HIBS), they pipeline the bus to enable concurrent data transmission. In such NoC-Bus hybrid systems, each bus can be shared by multiple routers on each layer to reduce the TSV amount. The buses are operated at higher clock frequency to provide enough data throughput and prevent that they become the system bottleneck.

Many other 3D NoC structures exist [4] but when compared with the NoC-Bus hybrid structure, they have drawbacks, e.g., large amount of TSVs, obvious system performance degradation, incompatibility with existing 2D NoC technologies, etc. Conversely, the existing NoC-Bus hybrid designs have the drawback that do not allow WS on the vertical buses, and thus rely on complicated bus structures and exhibit long packet transmission latency. In the following we address those issues by proposing a novel BVA mechanism.

III. Bus VC Allocation

The 3D NoC-Bus hybrid structure embedding the proposed BVA mechanism is illustrated in Fig. 1 where only 3 layers are depicted, for the sake of simplicity. We note that the scheme is general and can be applied for more silicon layers and that the BVA arbiter always locates in the middle layer to reduce the TSV number. The zones marked with different colors can run at different frequencies, thus the requirement for universal

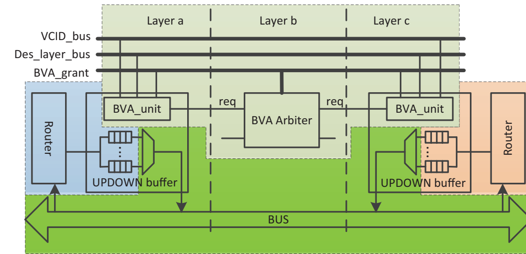


Fig. 1. NoC-Bus hybrid system structure. The BVA arbiter locates in the middle layer and the colored zones can run at different clock frequencies.

clock distribution throughout the 3D chip is released and each layer can work at its optimized speed. The BVA units only exist around buses, thus their clock zone only occupies a small portion of the silicon area.

On each silicon layer, a 2D mesh NoC is implemented. Each router in the NoC has 6 physical ports, i.e., one UP-DOWN port connected with the bus, and 5 ports connected with the local PU and the 4 neighboring routers located on its layer. On each NoC layer data are transmitted WS wise. For packets whose next hop is on another silicon layer, their head flits wait in the UPDOWN buffer until a free input VC in the target UPDOWN port is assigned to them by the BVA mechanism. Only after that, the packet flits can be injected into the bus along with the target layer number and the assigned VC index (VCID).

A. Problem Description

In symmetric NoCs each router port is solely connected with one neighboring router. As illustrated in Fig. 2, the state of each output VC is actively maintained in the output port by a Finite State Machine (FSM). When a packet needs to be transmitted to another router, it first applies for an output VC by asserting the VC Allocation (VA) request. If the request won both the local and the global arbitration, a free output VC is picked from the free VC list and assigned to the packet.

With this conventional VA strategy, a routing path can be reserved by the head flit without waiting until the entire packet is received, which is time efficient and can operate with small input buffers, as only few flits are locally stored instead of integral packets. Such VA mechanism can be easily implemented in 2D NoCs as the output VCs are only used by packets from the local router. However in 3D NoC-Bus hybrid systems, a packet can be destined to any other layer, and packets from different layers can compete for the same input VC in an UPDOWN input port. For each UPDOWN output port, maintaining the VCs status in UPDOWN input ports in other layers and assigning each cross layer packet a free output VC becomes much more complicated.

In the existing 3D NoC-Bus hybrid structures, the packets are usually buffered in the UPDOWN buffers (see Fig. 1) before they are injected into the bus. As the matter of fact, a bus and the UPDOWN buffers incident to it can be considered as belonging to a virtual 3D router, i.e., the UPDOWN buffers are actually the input ports of the virtual 3D router, and the bus works as the crossbar. Implementing the aforementioned conventional VA in the virtual 3D router requires a large

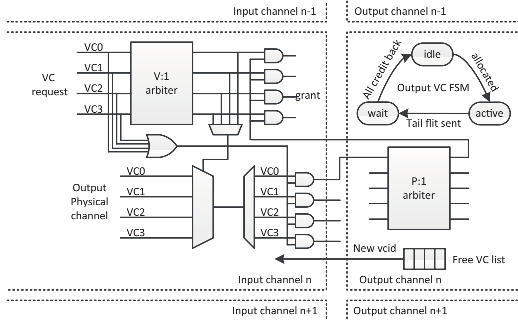


Fig. 2. Conventional VC allocation mechanism. The number of VCs is v. The number of physical ports is p.

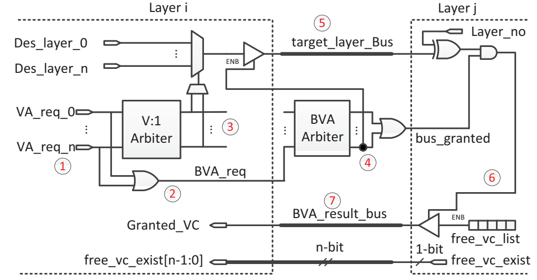


Fig. 3. The proposed BVA scheme.

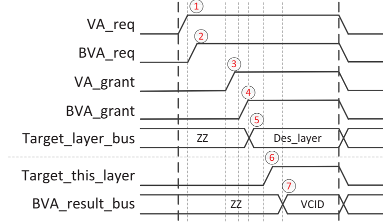


Fig. 4. Timing diagram of the BVA mechanism.

amount of TSVs. As illustrated in Fig. 2, each input port needs to send p request wires to the p output ports, and each output port needs to send p grant wires to the p input ports and $\log_2(v)$ wires to broadcast the assigned VCID, where v is the number of VCs. Moreover, each output port needs 1 extra signal to inform the input ports whether free output VCs exist. In an n-layer 3D NoC-Bus hybrid systems this implies that $2n^2 + n\log_2(v) + n$ TSVs are required and, as demonstrated in the next section, the proposed BVA mechanism significantly reduce this value.

B. Bus VC Allocation

The conventional VA mechanism assumes that multiple VA requests can be asserted by different input ports in the same cycle. However, at each hop, the VA application is done once per packet which means that for a packet length l the probability that a VA request is asserted at each port is $1/l$. Considering that packets do not arrive at the same port continuously, the actual probability is much smaller, thus the chance that multiple VA requests are asserted by different ports in the same cycle is also very small. In view of this analysis, we choose a BVA mechanism that assign free downstream input VC to only one packet per cycle.

The proposed BVA scheme is depicted in Fig. 3 and the associated timing diagram in Fig. 4. In each UPDOWN input port, an 1-bit *free_vc_exist* signal is asserted when at least one idle input VC exists, and sent to all the other routers along the bus. When a packet is destined to another layer, it is forwarded to the UPDOWN buffer, where if the *free_vc_exist* signal from the target router is high, its head flit asserts the VA request ① in Fig. 3 and 4. If multiple VCs are implemented in an UPDOWN buffer, each VC has the possibility to assert the VA request. The BVA request is generated by OR-ing the VA requests ②. The VA requests are sent to the local arbiter and the BVA request is sent to the BVA arbiter. After a certain delay the arbitration results are generated and as the BVA request and grant signals have to traverse several silicon layers, a packet receives the local *VA_grant* ③ earlier than the *BVA_grant* ④. The granted packet places its target layer number on the *Target_layer_bus* ⑤ while each UPDOWN input port monitors the BVA arbitration results and the *Target_layer_bus*. On the target layer of the granted BVA request ⑥, the UPDOWN input port chooses one idle VC from its free VC list and places the VCID on the *BVA_result_bus* ⑦. The *free_vc_exist* signal is also updated according to the remained free VCs. The source router reads the VCID from the *BVA_result_bus* and stores it into a dedicated register.

The number of TSVs (N_{TSV}) required by the BVA mechanism is calculated by (1), where n is the number of silicon layers, and v is the number of VCs in each physical channel. We note that the n *free_vc_exist* signals, the 1-bit *bus_granted* signal, the *Target_layer_bus* (width is $\log_2(n)$), and the *BVA_result_bus* (width is $\log_2(v)$) are penetrating TSVs, while the n *BVA_req* signals and the n *BVA_grant* signals are just half way due to the fact that the BVA arbiter is always placed in the middle of the 3D stack. Assuming, for example, a 4 layer 3D NoC-Bus hybrid with 4 VCs per physical channel the BVA requires only 13 TSVs for each pillar while the conventional VA requires 42.

$$N_{TSV} = 2n + \log_2(n) + \log_2(v) + 1. \tag{1}$$

C. Bus Data Transmission Policy

After a free input VC in the target router is reserved, a packet can transmit its flits to it via the bus with the WS technique. Each flit is transmitted together with the target layer number and the assigned VCID, thus flits belonging to different packets can be processed by the bus indiscriminately. In symmetric NoC systems, the VC buffer depth is usually less than the packet length if WS is implemented, and the credit return mechanism is used to inform the upstream router whether free buffer slots exist in the input VC. To save the TSVs required by the credit return mechanism, we set the input VC buffer depth in the UPDOWN port to the packet length, thus there is no need to inform the source router how many flits can still be transmitted. This can also guarantee that all flits injected into the bus will be absorbed by their target routers. We note that the buffer depth in the other ports can be smaller than the packet length.

In the dTDMA bus [7], a time slot is allocated to the packet that won the bus arbitration until all its flits are transmitted which means that the integral packet must be reassembled

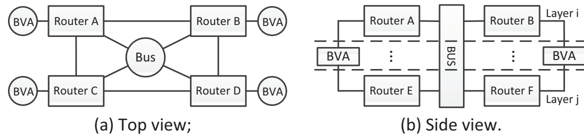


Fig. 5. Cluster Mesh Inter-layer topology.

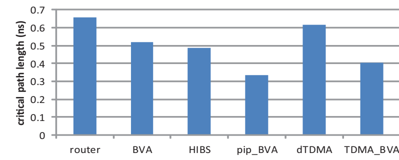


Fig. 6. Critical path length of routers and buses.

before the request is asserted. Otherwise, the reserved time slot is wasted by waiting for the un-arrived flits. When the BVA mechanism is utilized a bus request is asserted by each valid flit and the bus arbiter allocates time slots to individual flits instead of entire packets, which means that flits from different packets can be transmitted whenever the bus is available.

In HIBS [3] each bus stage is actually implemented as a 3 port router and a cross layer packet is only assigned a free input VC in the target UPDOWN input port when it arrives at the target layer. If all input VCs are occupied already, the packet has to wait in the bus stage buffer and causes Head of Line (HOL) blocking in the bus. A non blocking scheme was proposed to partially solve this issue by enabling the transmission of single hop packets when multiple hop packets are blocked, or vice versa, the blocking still can happen as it is possible that both kinds of packets are blocked. This scheme also requires that each bus stage must be able to store at least 2 integral packets for both upward and downward data flow direction. When the BVA mechanism is embedded, all flits are guaranteed to be absorbed by the target router, thus the HOL blocking is totally removed. Moreover, each bus stage just need to store several flits as the flits from different packets are equally processed by the bus. Thus the logic area cost of each bus stage is significantly reduced.

The buses can be shared by multiple routers, e.g., 2 or 4, on each layer to reduce the number of TSVs. As BVA grants only 1 BVA request in each cycle, we prohibit the inter-layer or intra-layer diagonal data transmission to maintain the simplicity and efficiency of the BVA mechanism. Each pillar still has a dedicated BVA arbiter, and when a router transmits flits via the bus, the target router must be right above or below it. Take the Cluster Mesh Inter-layer Topology (CMIT) [3], in which each bus is shared by 4 routers per layer, illustrated in Fig. 5 as an example, router A has to communicate with D and F via B or C, and B or E, respectively.

IV. Evaluation

To put the implications of our BVA mechanism in a better practical prospective we embed it in TDMA and pipelined bus based 3D NoC-Bus hybrid systems labeled as TDMA_BVA and pip_BVA, respectively, and compare their figure of merit against that of dTDMA bus [7] and HIBS [3] based systems. We implemented $4 \times 4 \times 4$ NoCs with 4 VCs per physical port at RTL level by using Verilog HDL. The VC buffer depth is the same with the packet length, i.e., 8-flits, in the UPDOWN input ports and 4-flits in other input ports. The flit width and the link width in 3D symmetric NoC are all 32-bits. For the sake of fairness, the buses, either pipelined or TDMA based, are composed of 2 unidirectional data lanes, each is 32-bits wide in charge of the upward and downward data flow, respectively. In HIBS, the input buffer in each bus stage is able to store 2 packets for every data flow direction. While in pip_BVA, the two Bus_FIFOs, upward and downward, in each bus stage are only set to 4-flits deep. The packets are transmitted by WS in each planar NoC and routed with the XYZ algorithm in the 3D NoC system.

The Routers and buses are synthesized using Cadence RTL Compiler with TSMC 45nm technology to estimate the critical path length, area cost, and power consumption. We assumed a TSV pad size of 5 μm with a $3\mu m$ pitch [6], and analysis with Cadence Virtuoso Spectre indicate a TSV delay of 20 ps per layer.

A. Critical Path Length

The critical path length of routers and buses in different NoC-Bus hybrid systems are illustrated in Fig. 6. We can observe that the BVA has shorter critical path than the routers. If the NoC layers number increases the BVA delay increases too and eventually surpasses the router delay. We note that given that BVA can work at different speed than the routers and buses this has no consequences on their implementation.

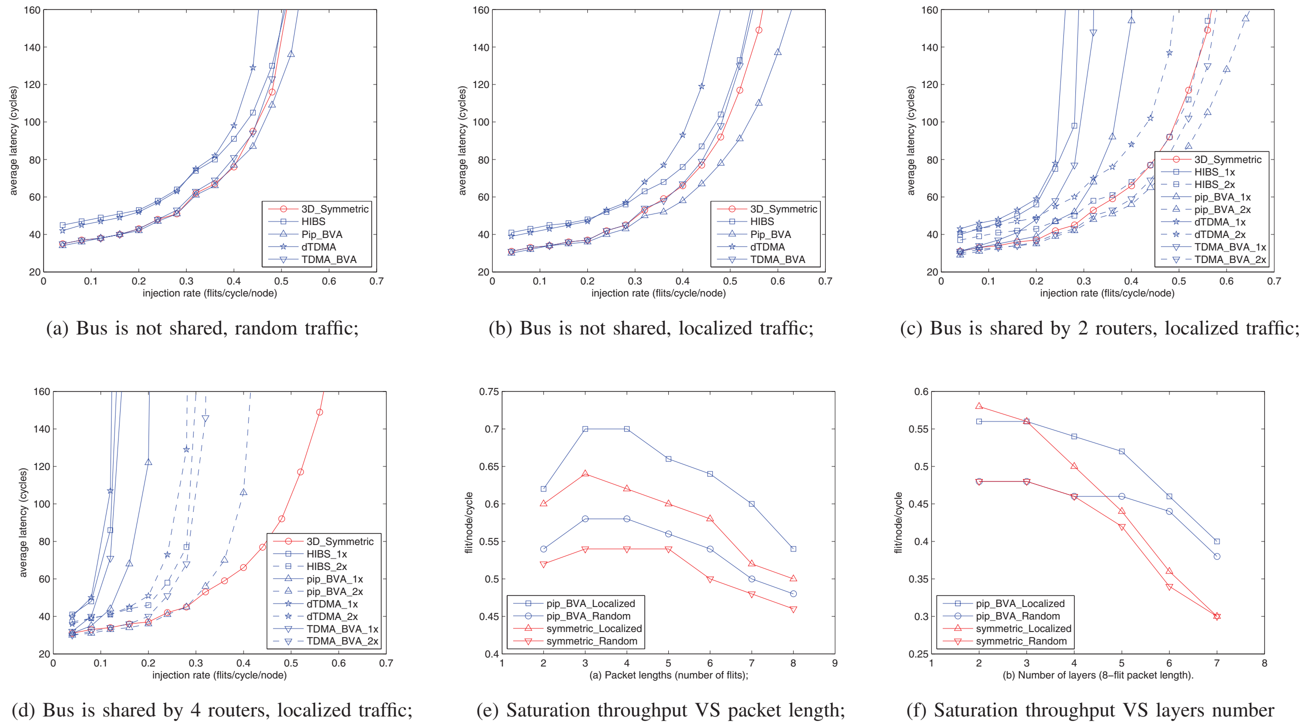
In the HIBS based system, each bus stage is actually a three port router and it makes use of VCs to partially solve the HOL blocking issue. When vertical WS is enabled, each bus stage just need to decide which of the two flits, one from the previous bus stage and one from the UPDOWN buffer, will be forwarded. Consequently, in pip_BVA, the critical path length of each bus stage is 31% shorter than that of HIBS, which means the bus can be twice faster than the routers.

In the dTDMA bus based system, each cross layer packet is assigned a free input VC when its head flit arrives at the target router. The "free input VC exists" flag in the UPDOWN input port is also updated in the meanwhile. This flag is then used by the centralized bus arbiter to decide whether another bus request can be granted. Consequently the dTDMA bus has a long critical path. In TDMA_BVA, the arbiter grants a request just according to the request current priority, thus the buses have shorter critical path than that of dTDMA and can be 1.6 times faster than the routers.

B. Synthetic Traffic

The performance of different 3D NoC systems at various Flit Injection Rates (FIRs) under random and localized traffic patterns is illustrated in Fig. 7(a-d). In the localized traffic, 50% of the packets are destined to the nodes in the same pillar, to simulate the practical case when task mapping is optimized [7]. The 3D symmetric NoC system embedding WS is also evaluated as a reference. The packet transmission latency is counted since the packet is generated in the source node till the tail flit is received by the destination node, i.e., the queuing time in the source node is included.

Our simulation results (Fig. 7(a-b)) when the bus is not shared on each layer indicate that the average packet transmission latencies in HIBS and dTDMA bus based hybrid systems



(a) Bus is not shared, random traffic; (b) Bus is not shared, localized traffic; (c) Bus is shared by 2 routers, localized traffic; (d) Bus is shared by 4 routers, localized traffic; (e) Saturation throughput VS packet length; (f) Saturation throughput VS layers number

Fig. 7. (a-d) Average packet transmission latency in different 3D NoC system. 1x, 2x means the bus frequency is 1, or 2 times higher than the NoC router frequency, respectively. (e-f) The system saturation throughput at different packet length and layers number. The packet length is 8-flits.

are on average 22% and 24% longer than that of 3D symmetric NoC system, respectively. This is because when WS is used in planar NoC and PS is used in the buses, extra delay is required to reassemble the integral packets in the UPDOWN buffer. When vertical WS is enabled, the extra delay is removed and both TDMA_BVA and pip_BVA can achieve similar or even better performance than the 3D symmetric NoC.

When each bus is shared by multiple routers, i.e., 2 and 4, per layer, the advantage of our proposal becomes more obvious (see Fig. 7(c-d)). Due to limited space only the experimental results for localized traffic are depicted. Note that due to the high traffic load, the buses in the hybrid systems saturate quickly when they work at the same frequency with the routers, especially when CMIT is implemented. However, according to the analysis in Section IV-A, the buses can be operated at higher frequencies, case in which we obtain substantial improvements. Note that although the maximum bus frequencies of the HIBS, dTDMA, and TDMA_BVA designs are only 35%, 7%, and 63% higher than that of the routers, we still evaluated their performance at 2x bus speed for comparison purpose. It is worth to note that pip_BVA always achieves the best performance, in terms of both average packet latency and saturation throughput, in all simulation contexts.

C. BVA Efficiency

The proposed BVA mechanism grants only one BVA request from one silicon layer in each cycle. Thus the BVA efficiency is expected to decrease as the packet becomes shorter and the number of layers becomes higher. The impact of those issues on the pip_BVA design and the 3D symmetric NoC saturation throughput is illustrated in Fig. 7(e-f).

Against expectation, the experimental results indicate that the saturation throughput increases as the packet length decreases from 8 to 3. This can be explained by the fact that although shorter packets assert BVA request more frequently at the same FIR, they also release the VCs faster. The packet length has more system performance influence, e.g., when the packet length decreases from 3 to 2, the saturation throughput has a 11% and 7% reduction for random and localized traffic, respectively. We note that the same trend exists in the 3D symmetric NoC systems.

As expected, the system saturation throughput decreases as the silicon layers number increases. But the decreasing speed in 3D symmetric NoC systems is faster than that in pip_BVA. Thus the decreasing is mainly caused by the layers number increase but not the BVA mechanism. Note that the planar NoC size is always 4×4. This proves that the BVA is not the system performance bottleneck.

D. PARSEC benchmarks

In this subsection, we evaluate our proposal with PARSEC benchmarks [13] traffic traces, which are recorded with the Netrace [14] tool on the M5 full system simulator [15]. We replay the traffic traces according to each packet time flag while maintaining the packets dependencies. The average packet transmission latencies for different benchmarks are illustrated in Fig. 8. The results indicate that when vertical WS is enabled, all 3D NoC-Bus hybrid systems provide similar packet transmission latency with the 3D symmetric NoC system, even when CMIT is implemented. Without BVA, the latencies in the HIBS and dTDMA bus based systems are on average 18% and 13% longer, respectively.

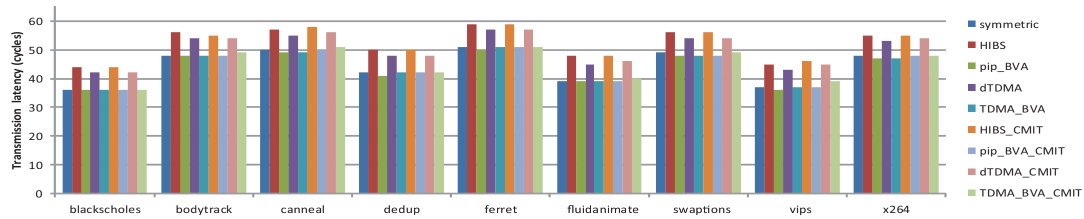


Fig. 8. Average packet transmission latency of PARSEC benchmarks. The buses work at the same frequency with the routers.

TABLE I. AREA AND POWER OF EACH ROUTER AND EACH BUS STAGE IN DIFFERENT 3D NOC SYSTEMS.

	Router	UPDOWN buffer		Bus stage			
		PS	WS	HIBS	pip_BVA	dTDMA	TDMA_BVA
Power (mW)	43.18 / 100%	10.28 / 24%	5.48 / 13%	13.06 / 30%	2.86 / 7%	–	–
Logic area (μm^2)	68207 / 100%	13701 / 20%	7854 / 12%	15740 / 23%	3644 / 5%	–	–
TSV number/area	–	–	–	76 / 1900	91 / 2275	86 / 2150	97 / 2425

E. Area and Power

The area cost and power consumption of routers and buses are presented in Table I. The bus power consumption is derived when buses and routers work at the same frequency. Note that the routers are implemented in the same way in different NoC-Bus hybrid systems.

In each UPDOWN buffer, 4 VCs are implemented in our experiments. For HIBS and dTDMA bus based system, integration packets need to be reassembled in the UPDOWN buffer before they are injected into the bus. While when vertical WS is enabled, such requirement is released and the buffer depth can be reduced to, for example, 4-flits. Consequently the area and power cost of the UPDOWN buffer is reduced by 47% and 43%, respectively. For each data flow direction, the HIBS bus stage is required to register at least 2 packets to partially solve the HOL blocking issue. While when BVA is implemented, the bus stage FIFO just need to register several flits, e.g., 4 flits in our case, and the HOL blocking is completely removed. Thus the implementation cost is also significantly reduced.

When the bus is not shared on each silicon layer, enabling vertical WS requires 15 and 11 more TSVs than the original HIBS and dTDMA based design in each pillar, respectively. However, the area overhead induced by the TSVs is still negligible even they are much bigger than planar wires.

V. CONCLUSIONS

In this paper, we proposed a Bus Virtual Channel (VC) Allocation (BVA) mechanism to enable vertical Wormhole Switching (WS) in 3D NoC-Bus hybrid systems. Because the VC allocation is performed only once per packet per hop, the possibility that multiple BVA requests are asserted along the same bus in each cycle is low. Thus in each cycle, the BVA mechanism forwards at most one request to its target router and picks a free input VC there. In this way, a routing path is reserved by the head flit, and the next flits in the packet can be transmitted with the WS technique on the buses. The evaluation results indicated that when vertical WS is enabled, the bus critical path length is reduced by at least 31% and the average packet transmission latency is reduced by at least 22%, when compared with the conventional pipelined bus and TDMA bus based systems. Moreover, the area cost and power consumption of the output buffer incident to the bus are reduced by 47% and 43%, respectively.

REFERENCES

[1] G. Blake, R. G. Dreslinski, and T. Mudge, "A survey of multicore processors," *IEEE Signal Process. Mag.*, vol. 26, pp. 26–37, Nov. 2009.

[2] B. Feero and P. Pande, "Networks-on-chip in a three-dimensional environment: A performance evaluation," *IEEE Trans. Comput.*, vol. 58, no. 1, pp. 32–45, Jan. 2009.

[3] M. Ebrahimi, M. Daneshtalab, P. Liljeberg, J. Plosila, and H. Tenhunen, "Cluster-based topologies for 3d networks-on-chip using advanced inter-layer bus architecture," *Journal of Computer and System Sciences*, vol. 79, no. 4, pp. 475–491, Jun. 2013.

[4] A. Rahmani, K. Latif, P. Liljeberg, J. Plosila, and H. Tenhunen, "Research and practices on 3d networks-on-chip architectures," in *Proc. NORCHIP*, Nov. 2010, pp. 1–6.

[5] K. Bernstein and et al., "Interconnects in the third dimension: Design challenges for 3d ics," in *Proc. DAC*, Jun. 2007, pp. 562–567.

[6] I. Loi, F. Angiolini, S. Fujita, S. Mitra, and L. Benini, "Characterization and implementation of fault-tolerant vertical links for 3-d networks-on-chip," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 30, no. 1, pp. 124–134, Jan. 2011.

[7] F. Li, C. Nicopoulos, T. Richardson, and Y. Xie, "Design and management of 3d chip multiprocessors using network-in-memory," in *Proc. ISCA*, 2006, pp. 130–141.

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

[9] J. Duato, S. Yalamanchili, and L. Ni, *Interconnection Networks: An Engineering Approach*. San Francisco, CA: Morgan Kaufmann Publishers Inc., 2003.

[10] Y. Xie, J. Cong, , and S. Sapatnekar, *Three-Dimensional Integrated Circuit Design: EDA, Design and Microarchitectures*. New York, NY: Springer, 2010.

[11] Y. Hwang, J. Lee, and T. Han, "3d network-on-chip system communication using minimum number of tsvs," in *Proc. ICTC*, Sep. 2011, pp. 517–522.

[12] J. Kim and et al., "A novel dimensionally-decomposed router for on-chip communication in 3d architectures," in *Proc. ISCA*, May 2007, pp. 138–149.

[13] C. Bienia, "Benchmarking modern multiprocessors," Ph.D. dissertation, Princeton University, Jan. 2011.

[14] J. Hestness and S. W. Keckler, "Netrace: Dependency-tracking traces for efficient network-on-chip experimentation," Department of Computer Science, The University of Texas at Austin, Austin, Texas, Tech. Rep. TR-10-11, May 2010.

[15] N. Binkert, R. Dreslinski, L. Hsu, K. Lim, A. Saidi, and S. Reinhardt, "The m5 simulator: Modeling networked systems," *IEEE Micro*, vol. 26, no. 4, pp. 52–60, Jun./Aug. 2006.