Headfirst Sliding Routing: A Time-Based Routing Scheme for Bus-NoC Hybrid 3-D Architecture

Takahiro Kagami[1], Hiroki Matsutani[1], Michihiro Koibuchi[2], and Hideharu Amano[1]

[1]Keio University
3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Japan
blackbus@am.ics.keio.ac.jp

[2]National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
koibuchi@nii.ac.jp

***Abstract*—A contact-less approach that connects chips in vertical dimension has a great potential to customize components in 3-D chip multiprocessors (CMPs), assuming card-style components inserted to a single cartridge communicate each other wirelessly using inductive-coupling technology. To simplify the vertical communication interfaces, static Time Division Multiple Access (TDMA) is used for the vertical broadcast buses, while arbitrary or customized topologies can be used for intra-chip networks. In this paper, we propose the Headfirst sliding routing scheme to overcome the simple static TDMA-based vertical buses. Each vertical bus grants a communication time-slot for different chips at the same time periodically, which means these buses work with different phases. Depending on the current time, packets are routed toward the best vertical bus (elevator) just before the elevator acquires its communication time-slot. Network simulations show that Headfirst sliding routing reduces the communication latency by up to 32.7%, and full-system CMP simulations show that it reduces application execution time by 9.9%. Synthesis results show that the area and critical path delay overheads are modest.**

I. Introduction

The three-dimensional integration is a promising VLSI architecture that stack several smaller wafers or dies in order to reduce the wire length and wire delay, and three-dimensional Network-on-Chip (3-D NoC) [28] has been extensively studied in terms of its network topology [18][7][23], router architecture [12][14][22], and routing strategy [25].

Various interconnection techniques have been developed to connect multiple chips in a 3-D IC package: wire-bonding, micro-bump [2][13], wireless (e.g., capacitive- and inductive-coupling) [6][10][20][21] between stacked dies, and through-silicon via (TSV) [6][3] between stacked wafers. These 3-D IC technologies are compared in [6]. Many recent studies on 3-D IC architectures focus on micro-bump and TSV techniques that offer the highest level of interconnect density. On the other hand, as another 3-D integration technique, the inductive-coupling can connect more than two examined dies without wire connections.

The wireless contact-less approach that connects chips in vertical dimension has a great potential to customize components in 3-D chip-multiprocessors (CMPs), assuming card-style components inserted to a single cartridge communicate each other wirelessly using inductive-coupling technology. Although power supplies are provided by bonding wires at this moment, wireless power transmission techniques using inductive-coupling have been improved recently [30][29][24]. The inductive-coupling power transmission can be used for these card-style components inserted to a cartridge [4]. In this case, adding, removing, and swapping chips in a package after the chips have been inserted to a cartridge are possible, which will bring us a great flexibility of "field stackable systems" using the card-style components in the future.

Toward this purpose, the vertical communication interfaces should be simplified, while arbitrary or customized topologies should be used for intra-chip networks; thus, we focus on static Time Division Multiple Access (TDMA) buses for the inter-chip communication. In this paper, we propose the Headfirst sliding routing scheme to overcome the simple static TDMA-based vertical buses. The static TDMA-based vertical buses grants a communication time-slot for different chips at the same time periodically, which means they are working with different periodic scheduling. For example, at a certain moment, vertical bus 0 gives a time-slot for chip 1, vertical bus 1 allows chip 2, and vertical bus 2 allows chip 0. At the next phase, vertical bus 0 gives a time-slot for chip 2, vertical bus 1 allows chip 0, and vertical bus 2 allows chip 1. Each vertical bus behaves just like an elevator in an office building.

Fortunately, a waiting time to obtain the time-slot of vertical bus (elevator) is predictable for each chip, thus a key design of packet routing is to select the best elevator that minimizes the waiting time. The best elevator to route packets to a destination depends on the current time. The proposed Headfirst sliding routing routes packets toward the best elevator so that the packets acquire their communication time-slot just when they arrive at the elevator.

In this paper, Headfirst sliding routing and a conventional minimal routing are compared in terms of the zero-load latency, network performance (latency vs. offered workload), and application execution time using a full system CMP simulator. The area and critical path delay overheads are also evaluated.

The rest of this paper is organized as follows. Section II overviews the 3-D bus-NoC architecture that uses buses for vertical dimension and NoC for horizontal dimension. It also introduces wireless inductive-coupling technology. Section III proposes Headfirst sliding routing for wireless 3-D NoC-Bus architecture. Section IV analyzes the zero-load latency and Section V evaluates the performance and cost. We conclude the paper in Section VI.

II. Related work

A. Wired 3-D Interface and Bus Structure

Micro-bump [2][13] and **Through-silicon via (TSV)** [6][3] have been mature techniques utilized in actual products. Although the micro-bump is mostly limited for face-to-face connections of two dies, the TSV is used to connect a number of chips and so 3-D bus structure is utilized. Especially, the 3-D bus is efficient for connecting processors and memory systems in 3-D multi-core systems[15]. Dynamic Time Division Multiple Access (dynamic TDMA) [26][1] bus is introduced for distribution of bus mastership between multiple chips, while point-to-point Network-on-Chips (NoCs) are used inside the chip. HIBS[5] adds paralleism to the bus structure. Such 3-D architecture is called Bus-NoC Hybrid architecture, and it has become a conventional way to build 3-D NoCs. However, as described later, these architectures is difficult to be used in inductive coupling bus.

B. Wireless Inductive Coupling

Techniques on wireless chip interconnection are classified into two: capacitive-coupling and inductive-coupling. Here, we concentrate inductive coupling, since capacitive coupling is mainly used for face-to-face connecting with two chips.

Inductive-coupling [20] [19] [24] uses square or hexagon coils as data transmitters.The coils can be implemented with common metal layers of the chip, and no special process technology is required for building them.

Inductive-coupling has potential as an interconnection technology for custom building-block SiPs, since addition, removal, and swapping of chips become possible after the chips have been fabricated and stacked in a package with a low cost. A contact-less interface without ESD protection device has been shown to be able to handle bit rate of more than 1GHz with a low energy dissipation (0.14pJ per bit) and a low bit-error rate (BER $< 10^{-12}$) [20].

In the inductive-coupling approach, data modulated by a driver are transferred between two coils that are exactly superimposed on each other. The driver and inductor pair for sending data is called the TX channel, while the receiver and inductor pair is called the RX channel. Since a coil can be used both for transmitter and receiver, the TX channel and RX channel can be switched quickly, that is, a half-duplex bi-directional channel can be formed by using a single coil. Also, data multicast can be used if a TX channel is placed at the same location of multiple RX channels in different chips.

By switching of TX/RX channel on a multicast channel, an inductive coupling bus can be formed on multiple chips. Although a lot of practical systems are available by using point-to-point networks using inductive coupling, there is few report to use bus. Since it takes relatively large latency for arbitration of multiple chips, a simple TDMA is preferred for inductive coupling bus. In MuCCRA-Cube [27], TDMA bus is used for 3-D links between PE (Processing Element) array of dynamically reconfigurable processors. In this system, four time-slots are assigned into each chip, and a chip can send the data in the term of the assigned time-slot. The 3-D links were not efficiently used because of the low utilization ratio for 3-D direction [27].

III. Wireless 3-D NoC-Bus Hybrid Architecture

Toward the practical wireless 3-D ICs that allow us to add, remove, and swap the chips in the field, hardware complexity of vertical communication lines (e.g., number of inductors) should be minimized. Thus, static TDMA buses are preferred for the inter-chip communication compared to the dynamic one that requires additional control lines (i.e., dedicated inductors) for dynamic arbitration.

To fill in the gap between static and dynamic TDMA schemes while keeping the hardware simplicity of the static scheme, the following two ideas are combined in this paper.

- Phase-shift static TDMA control for multiple vertical broadcast buses
- Headfirst sliding routing scheme that routes packets to the best vertical bus (elevator) depending on the current time, in order to minimize the waiting time at the elevator

Note again that arbitrary or customized topologies can be used for intra-chip networks.

A. Phase-Shifted Static TDMA Buses

The major performance bottleneck of the static TDMA scheme is the waiting time at the bus to acquire a time-slot. The waiting time for a time-slot increases as the number of chips increases. The negative impact will also increase when state-of-the-art low-latency single-cycle routers are employed in the horizontal dimension. To minimize the waiting time at the buses, all the vertical buses grant a communication time-slot for different chips at the same time periodically, which means they are working with different periodic scheduling(See Figure 1). Let n is the number of chips and T_{slot} is the length of a time-slot. At time T, vertical bus i gives a time-slot for chip $(T/T_{slot}+i) mod\, n$, When the number of vertical buses is greater than or equal to n, at least one vertical bus is available for sending on each chip, which can reduce the worst-case waiting time from $nT_{slot}-1$ to $T_{slot}-1$ assuming no packet contentions at the selected elevator.

B. Routing Schemes

To route packets in the wireless 3-D NoC in which vertical links employ the phase-shift static TDMA buses while horizontal networks employ arbitrary or customized topologies, we propose to use two routing schemes: Minimum-hop (MH) routing and Headfirst sliding (HS) routing. MH routes packets using a minimal path between a source and a destination via an elevator.

Our observation is that MH routing achieves a high saturated throughput while its zero-load communication latency is longer than that of the dynamic TDMA (ideal case) due to the waiting time at elevators. Another observation is that the best

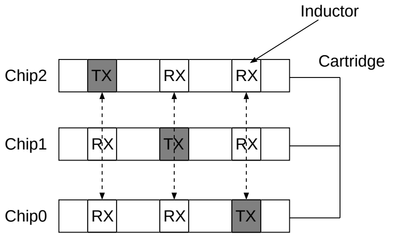


Fig. 1. Phase-shifted static TDMA Buses

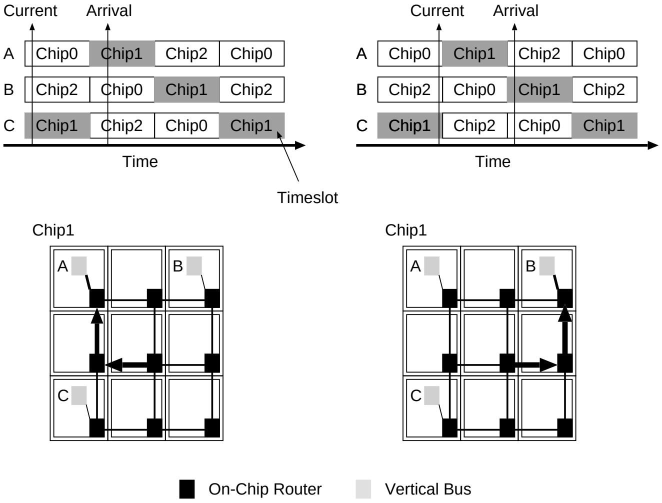


Fig. 2. Concept of Headfirst sliding routing

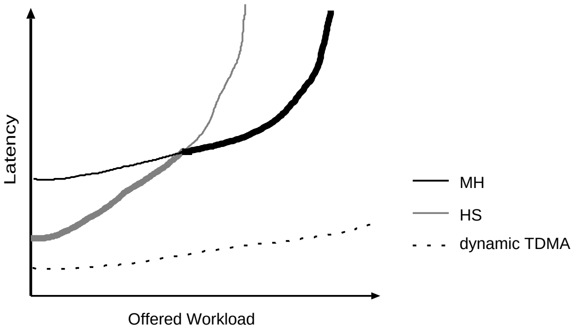


Fig. 3. Contribution of this work

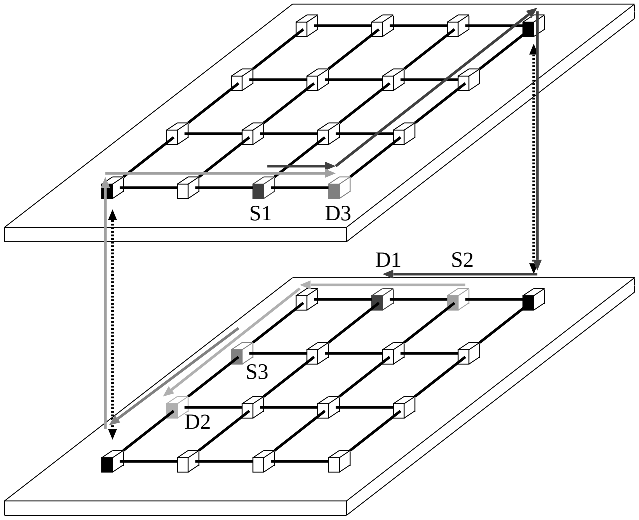


Fig. 4. An example of deadlock situation

elevator to route packets to a destination is depending on the current time. HS routes packets toward the best elevator, based on the current time, so that they acquire their communication time-slot when they arrive at the elevator (see Figure 2).

To fill in the gap between static and dynamic TDMA schemes while keeping the hardware simplicity of the static scheme, MH and HS routing schemes are used as follows.

- When a offered workload is less than a certain threshold value, HS routing is used to further reduce the communication latency.
- When a offered workload is more than the threshold, MH routing is used to achieve a higher saturated throughput.

Figure 3 illustrates the contribution of this work. The x-axis shows the offered workload and the y-axis shows the communication latency. By using MH and HS routing schemes depending on the workload, the expected throughput vs. latency curve is emphasized with a bold line. Note that there is no throughput degradation compared to MH routing and its communication latency can be close to the ideal dynamic TDMA scheme when the workload is not high by using the time-based HS routing scheme.

The following sections will illustrate MH and HS routing schemes.

1) Minimum Hop Routing: Packets are routed based on the following rules.

- **Transfer rule 1:** If the source and destination are on the same chip, packets are routed based on arbitrary deadlock-free routing on the chip (e.g., XY routing on 2-D mesh topology).
- **Transfer rule 2:** If the source and destination are on different chips, packets are first routed to an elevator on the source chip, moved to the destination chip, and routed to the destination. An elevator is selected so that the hop count is minimized.

MH routing does not guarantee deadlock-freedom without virtual channels. Figure 4 illustrates an example of deadlock situation. Each chip employs 4×4 2-D mesh topology, in which XY routing is used for intra-packet transfers. In this case, S1 sends a message to D1, S2 sends a message to D2, and S3 sends a message to D3; thus they cause the cyclic dependency which introduces deadlocks.

To avoid such structural deadlocks, two VCs are required for all the routers, and the following rule is imposed to MH routing.

- **Transfer rule 3:** If the source and destination are on different chips, packets are transferred with VC-0 on the source chip, while VC-1 is used on the destination chip after an elevator is used. If the source and destination are on the same chip, only VC-1 is used for the packet transfer.

2) Headfirst Sliding Routing: Depending on the current time, packets are routed toward the best elevator so that an expected transfer time is minimized. The above-mentioned Transfer rule 2 is updated as follows.

- **Transfer rule 2':** If the source and destination are on different chips, packets are first routed to an elevator on the source chip, moved to the destination chip, and routed to the destination. An elevator is selected so that the expected transfer time T is minimized. T is formulated as follows.

$$T = RH_{sd} + T_{wait} \tag{1}$$

where R is a flit transfer time at a router, H_{sd} is the number of hops from source to destination, and T_{wait} is an expected waiting time at an elevator.

T_{wait} is calculated as follows. First the arrival time of a packet to an elevator T_{arrive} is calculated as follows, assuming no packet contentions.

$$T_{arrive} = CurrentTime + RH_{sb}, \tag{2}$$

where H_{sb} is the number of hops from source to elevator, which is depending on the routing algorithm. The transfer start and finish times can be estimated based on this T_{arrive}.

Let T_{alloc} is a time-slot allocation time and T_{slot} is the length of a time-slot. If a packet transfer start time is greater than or equal to T_{alloc} and a packet transfer finish time is less than $T_{alloc} + T_{slot}$, T_{wait} is zero. Otherwise, T_{wait} is set to the next time-slot allocation time.

C. Run time Routing Policy Switching

In order to switch HS and MH routing at run time, each on-chip router employs both routing policy at the local input port. The router selects either routing policy depending on the offered network load. For measurement of the network load, the number of packets the local input buffer receives is counted with a packet counter which is reset to zero for the every *m* cycles.

Here, the following simple selection algorithm is adopted. First, the HS routing is used. The counter is incremented when a header flit arrives at the local input buffer. If the counter value reaches the threshold before resetting to zero, the MH routing is selected. Otherwise, the HS routing is selected. Threshold is selected depending on the simulation as described later.

IV. Analysis

First, we analyzed the zero-load latencies for MH routing and HS routing when the source and destination are different chips. For obtaining baseline results, uniform traffic is assumed.

Given that a packet which consists of L flits goes through H routers, its zero-load latencies, T_0, is calculated as

$$T_0 = H(T_{router} + T_{link}) + T_{bus} + L/BW + T_{block}, \tag{3}$$

where T_{router}, T_{link}, and T_{bus} are latencies for transferring a header flit on a router, a link, and a bus, and T_{block} is the waiting time in the case when a packet misses the time-slot.

TABLE I
Configurations of different bus placement policies

Pattern	B (Number of buses)	Placement method
sparse2	2	sparse
dense2	2	dense
sparse4	4	sparse
dense4	4	dense
sparse8	8	sparse
dense8	8	dense

MH and HS routing schemes show different zero-load latencies since their T_{block} differs. Under uniform traffic,T_{block} for MH ($T_{block,mh}$) is calculated as follows

$$T_{block,mh} = \frac{\sum_{t=1}^{T_{max}} t}{MT_{slot}}, \tag{4}$$

where M is the number of stacked chips and T_{max} is the maximum time of waiting for the valid time-slot assignment.

On the other hand, T_{block} for HS, $T_{block,hs}$ is described as

$$T_{block,hs} = f(X_{bus}, Y_{bus}, T_{slot}, M), \tag{5}$$

where X_{bus} and Y_{bus} are 2-D coordinates of a bus, T_{slot} is the length of a time-slot, M is the number of chips.

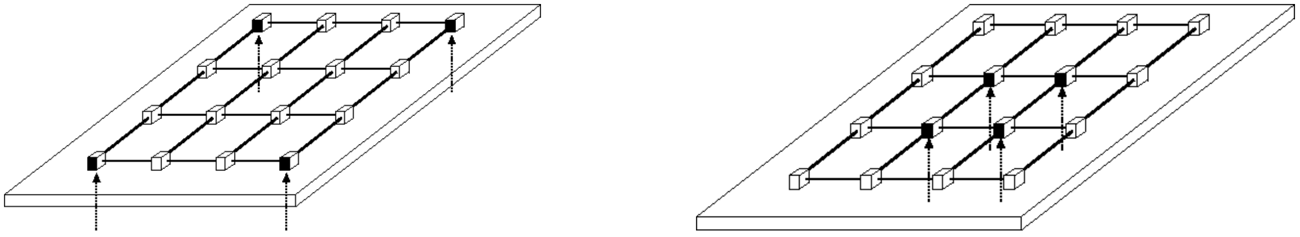
Based on Equations 4 and 5, the zero-load latencies for MH and HS routing schemes are analyzed by using a flit-level simulator. The 3-D stacked chips each of which uses a 4×4 mesh topology are selected as targets. The number of chips M is assumed as $2 \leq M \leq 8$. Six configurations of bus placement are summarized in Table I. The results are depending on location of 3-D buses. Here, two placement methods, *sparse* and *dense* are examined. In *sparse* method, the buses are distantly located along the edges of the chip. On the other hand, in *dense* method, they are located near the center of the chip locally. As an example, Figure 5(a) and 5(b) show *sparse* and *dense*, respectively given that $B = 4$.

Figure 6 shows the zero-load latencies for MH and HS, assuming $T_{router} = 2$, $T_{link} = 1$, $L = 5$, and $T_{slot} = 8$. Here, *dense*4 is adopted as the bus location pattern. The similar tendency is observed when other bus location patterns are used.

The result shows that T_{block} has a great impact on overall latency. HS reduces the zero-load latency compared with MH, since it can reduce T_{block} drastically. As the number of stacked chip increases, the reduction of T_{block} increases. In contrast, it slightly increases H compared with MH, since it is not a minimal routing.

V. Evaluations

In this section, the hardware overhead of the router with the HS routing is evaluated first. Then, we compare the MH routing with the HS routing in terms of their communication latency and impact to application performance.



(a) sparse4

(b) dense4

Fig. 5. An example of bus placement method

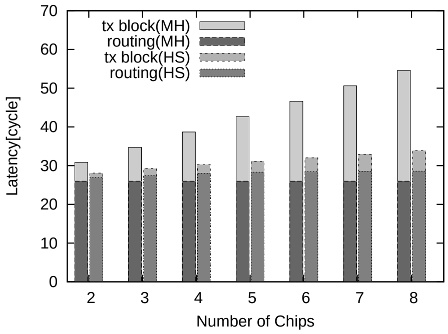


Fig. 6. Zero-load latency

TABLE II
BASELINE ROUTER

Port	5 input/output port
Buffer	5 flit
Routing	XY Routing
Switching	Wormhole 2 VCs
Pipeline Stage	[RC/VSA] [ST] [LT]
Flit Size	128 bit

A. Hardware Overhead

Here, we designed the HS router based on a 5-port common router with 3-stage pipeline structure shown in Table II, and evaluated the hardware amount and critical path delay. The HS routing algorithm is implemented on the RC stage of the baseline router. The design is described in Verilog HDL, and synthesized at operating frequency of 500MHz by using the Synopsys Design Compiler version D-2010.3. Fujitsu's e-shuttle 65nm CMOS process with 12-layer, and a standard cell library CS202SZ are used.

Figure 7 compares the area of the baseline router and the HS router. The area overhead of the HS is only 2.3% compared with the baseline router. Note that if the number of VCs or the size of the buffer increases, the relative overhead becomes small.

The critical paths of both routers are on the LT stage, and their maximum delays are the same, although the RC Stage of the HS router is heavier than that of the baseline router. Thus, the operational frequency is not influenced.

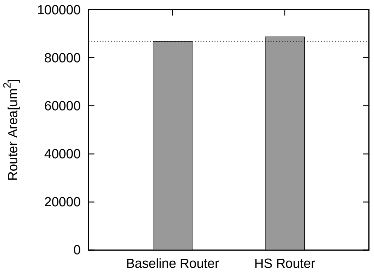


Fig. 7. Hardware amount

B. Network Performance

In order to measure the baseline network performance, we used a flit-level simulator written by C++ [9]. In the simulation, HS, MH and the case of switching the HS and the MH (HS + MH) are compared. Also, the network performance on the ideal dynamic TDMA scheme is evaluated. The target 3-D stacked chips are as described in Section IV, and in the case of HS + MH, the counter reset interval is 512 cycles. The threshold value of switching routing policys are determined based on the cross point of HS and MH shown in Figure 8 later. Let TH be the offered workload of cross point, m be the reset interval of packet counter, and S be the packet size. The threshold value is given as mTH/S.

Figure 8 shows simulation results under the uniform traffic, matrix traffic and reversal traffic. Here, *sparse*4 (4-chip) and *dense*4 (4-chip) are adopted as the bus location pattern. The similar tendency is observed when other bus location patterns are used. As shown in figures, the HS improves latency, compared with the MH when the offered workload is low. In paticular, the *dense*8 (8-chip) reduces the latency by up to 32.7% compared with MH. As the workload increases, the difference becomes gradually small, and when it is more than a certain threshold value, the latency of the HS is larger than that of the MH. This tendency is just the same as expected in Section III-B.

Also, the results demonstrate that the HS + MH latency is close to the HS routing scheme when workload is not high, and in the region of high workload, the latency of the MH is close to the HS + MH routing scheme. Moreover, the HS + MH latency is close to the dynamic scheme in the region of low workload. Thus, by switching the HS and the MS

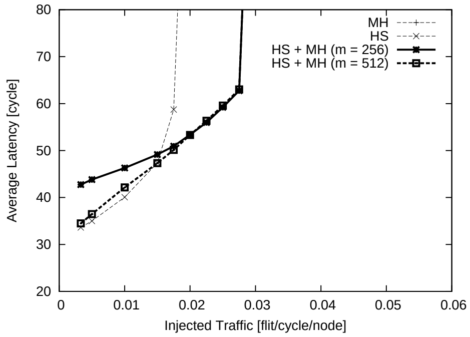


Fig. 9. Change the reset interval

schemes depending on the workload, we can further reduce the communication latency compared to MH routing even by using the simple static TDMA buses.

The interval of counter reset is important to select the routing scheme. As shown in Figure 9, the latency of the MH + HS is close to the MH routing scheme when the short interval is used because of the low resolution of the workload measurement. As the interval increases, the resolution is improved. However, the large interval value makes the switching slow. Thus, there is a trade-off between the measure resolution and switching speed.

C. Application Performance

In order to measure the impact to real application performance, full system simulations of wireless 3-D CMPs are performed.

Table III shows target 3-D shard-memory CMPs consisting of four chips. They are connected with multiple vertical buses as shown in Table I. SNUCA[11] is adopted as a cache architecture, and four memory controllers are placed in the four corners of the bottle chip. Other parameters are listed in Table IV.

For simulation, we used a full system multi-processor simulator: GEMS[17] and Wind River Simics[16]. We modified a detailed network model of GEMS, called Garnet, in order to precisely simulate behavior of the MH and HS.

A directory-based MOESI coherence protocol that uses three message classes is used. Six VCs are, thus required for each input port since each message class requires two VCs to avoid structural deadlocks.

To evaluate the application performance of these routing schemes, we used eight parallel programs using OpenMP from NAS Parallel Benchmarks[8] on Sun Solaris 9 operating system. These benchmark programs are compiled by Sun Studio 12 and executed on Solaris 9. The number of threads is set to eight.

Figure 10 shows the application execution cycles of eight benchmark programs in the case of four buses. The application execution time (Y-axis) is normalized to the execution time using the MH. As shown, HS improves execution cycles by 5.9% - 9.9%, compared with those of the MH. This comes from that in the assumed CMPs and application programs, the workload is not so heavy. In this evaluation, the advantages of the HS is large when the location *dense*4 is used.

TABLE III
CHIP CONFIGURATION

Topology	4×4 mesh
Routing	XY Routing
Processor/L1Cache	2
L2Cache	16
Router	16

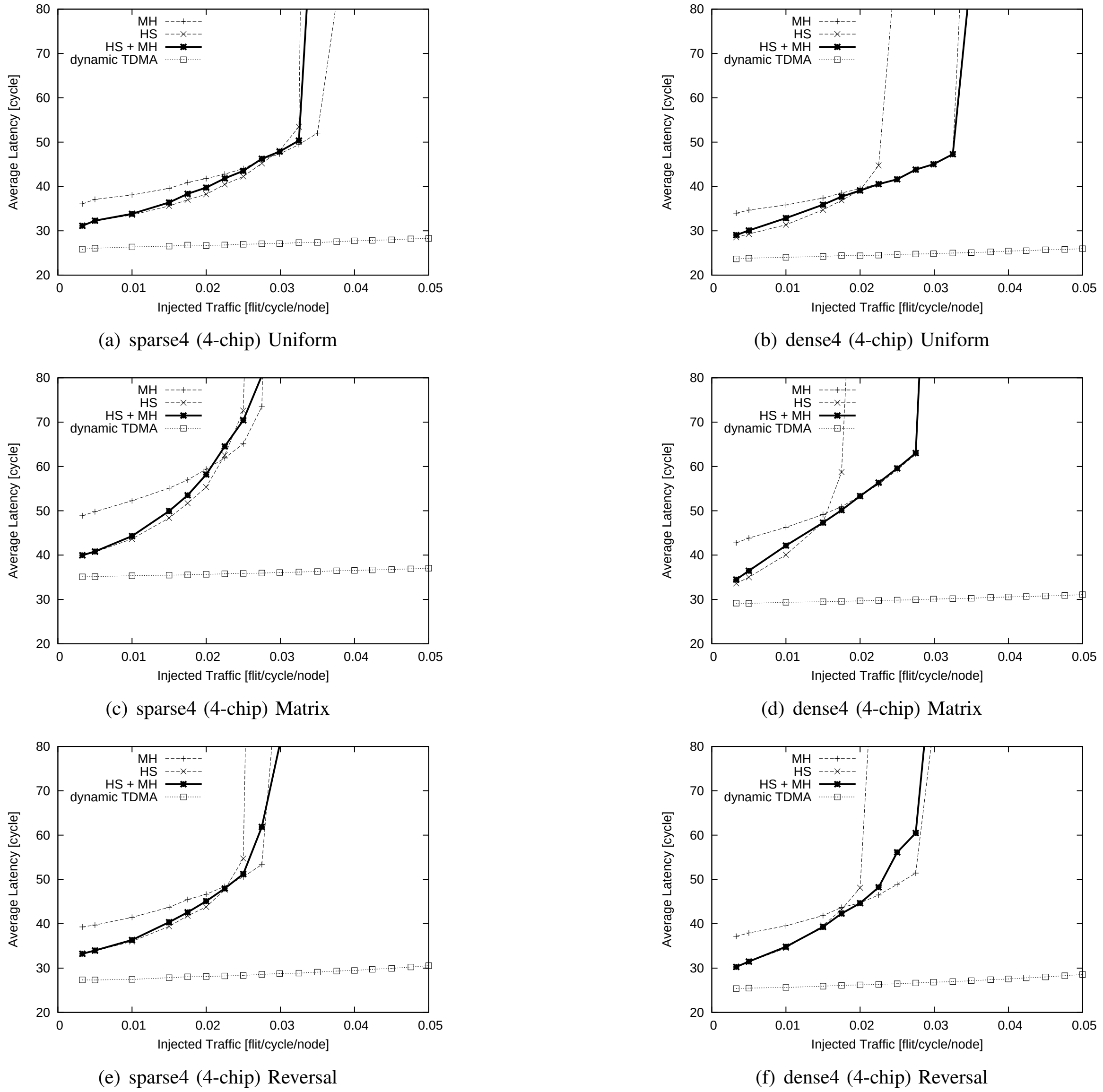
TABLE IV
SIMULATION PARAMETER

Processor	UltraSPARC-III
L1 I/D cache size	64 KB (line:64B)
L1 cache latency	1 cycle
L2 cache bank size	256 KB (assoc:4)
L2 cache latency	6 cycle
Memory size	4 GB
Memory latency	160 (± 2) cycle
Router pipeline	[RC/VSA][ST][LT]
Buffer size	5-flit per VC (default)
Flit size	128 bit
Protocol	MOESI directory
# of Message Classes	3
Control / data packet size	1 flit / 5 flit

VI. Conclusions

Toward the practical wireless 3-D ICs that allows us to add, remove, and swap the chips in the field, hardware complexity of vertical communication lines (e.g., number of inductors) should be minimized. Thus, static TDMA buses are preferred for the inter-chip communication compared to the dynamic one that requires additional control lines (i.e., dedicated inductors) for dynamic arbitration. To fill in the gap between static and dynamic TDMA schemes while keeping the hardware simplicity of the static scheme, we propose Headfirst sliding routing scheme, which routes a packet toward the best vertical bus (elevator) just before the elevator acquires its communication time-slot, depending on current time. In addition, we propose to switch Headfirst sliding routing and a conventional minimal routing (Minimum hop routing) in accordance with a offered workload.

In this paper, we analyzed zero-load latencies of Headfirst sliding routing and Minimum Hop routing, and evaluated them in terms of the network performance (latency vs offered workload) and the application performance. According to the analysis, we confirmed that the time when the packet transfer wait at buses have a heavy impacts on the whole latency, and Headfirst sliding routing can drastically reduce it. The result of the network performance showed that Headfirst sliding routing reduces latency up to 32.7% at a low workload, although Minimum hop performs better than Headsliding first at a high workload. It present that the selection between Headfirst sliding and Minimum hop in accordance with a offered workload is the best approach in static TDMA-based bus architectures. Moreover, at a low workload, Headfirst sliding performs nearly as good as a minimal routing with



(a) sparse4 (4-chip) Uniform

(b) dense4 (4-chip) Uniform

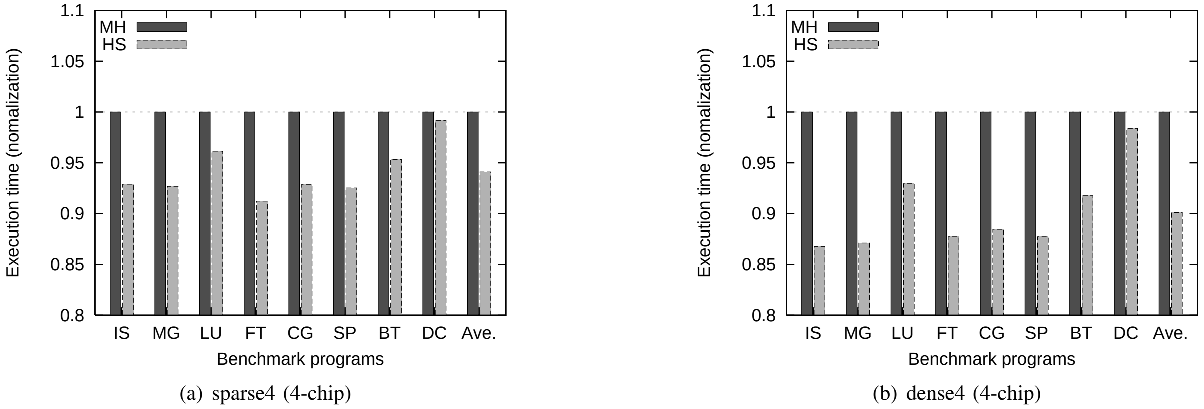
(c) sparse4 (4-chip) Matrix

(d) dense4 (4-chip) Matrix

(e) sparse4 (4-chip) Reversal

(f) dense4 (4-chip) Reversal

Fig. 8. Average packet latency



(a) sparse4 (4-chip)

(b) dense4 (4-chip)

Fig. 10. Application execution time

dynamic TDMA-based buses, although Headfirst sliding works on static TDMA-based buses. Thus, Headfirst sliding fill in the gap between static and dynamic TDMA schemes, while keeping the hardware simplicity of the static scheme. As the result of the application performance, Headfirst sliding routing improves by 9.9%, compared by Minimum hop routing. We also confirmed that the area and critical path delay overheads are modest.

As a future work, we are planning to investigate routing schemes to improve the saturated throughput at a high workload. We are going to evaluate the thermal In addition, we will extend Headfirst sliding routing scheme to improve the latency within a range of low to middle workload.

References

[1] M. O. Agyeman, A. Ahmadinia, and A. Shahrabi. Low power heterogeneous 3D Networks-on-Chip architectures. In *Proceedings of the International Conference on High Performance Computing and Simulation (HPCS'11)*, pages 533–538, July 2011.

[2] B. Black, M. Annavaram, N. Brekelbaum, J. DeVale, L. Jiang, G. H. Loh, D. McCaule, P. Morrow, D. W. Nelson, D. Pantuso, P. Reed, J. Rupley, S. Shankar, J. P. Shen, and C. Webb. Die Stacking (3D) Microarchitecture. In *Proceedings of the International Symposium on Microarchitecture (MICRO'06)*, pages 469–479, Dec. 2006.

[3] J. Burns, L. McIlrath, C. Keast, C. Lewis, A. Loomis, K. Warner, and P. Wyatt. Three-Dimensional Integrated Circuits for Low-Power High-Bandwidth Systems on a Chip. In *Proceedings of the International Solid-State Circuits Conference (ISSCC'01)*, pages 268–269, Feb. 2001.

[4] H. Chung, A. Radecki, N. Miura, H. Ishikuro, and T. Kuroda. A 0.025-0.45 W 60%-Efficiency Inductive-Coupling Power Transceiver with 5-bit Dual-Frequency Feedforward Control for Non-Contact Memory Cards. *IEEE Journal of Solid-State Circuits*, 47(10):2496–2504, Oct. 2012.

[5] M. Daneshtalab, M. Ebrahimi, and J. Plosila. HIBS - Novel Inter-layer Bus Structure for Stacked Architectures. In *Proceedings of International Conference on 3D Systems Integration Conference (3DIC'11)*, 2011.

[6] W. R. Davis, J. Wilson, S. Mick, J. Xu, H. Hua, C. Mineo, A. M. Sule, M. Steer, and P. D. Franzon. Demystifying 3D ICs: The Pros and Cons of Going Vertical. *IEEE Design and Test of Computers*, 22(6):498–510, Nov. 2005.

[7] B. Feero and P. P. Pande. Networks-on-Chip in a Three-Dimensional Environment: A Performance Evaluation. *IEEE Transactions on Computers*, 58(1):32–45, Jan. 2009.

[8] H. Jin, M. Frumkin, and J. Yan. The OpenMP Implementation of NAS Parallel Benchmarks and Its Performane. In *NAS Technical Report NAS-99-011*, Oct. 1999.

[9] A. Jouraku, M. Koibuchi, and H. Amano. An Effective Design of Deadlock-Free Routing Algorithms Based on 2-D Turn Model for Irregular Networks. In *IEEE Transactions on Parallel and Distributed Systems(TPDS)*, volume 18, pages 320–333, Mar. 2007.

[10] K. Kanda, D. D. Antono, K. Ishida, H. Kawaguchi, T. Kuroda, and T. Sakurai. 1.27-Gbps/pin, 3mW/pin Wireless Superconnect (WSC) Interface Scheme. In *Proceedings of the International Solid-State Circuits Conference (ISSCC'03)*, pages 186–187, Feb. 2003.

[11] C. Kim, D. Burger, and S. W. Keckler. An Adaptive, Non-Uniform Cache Structure for Wire-Delay Dominated On-Chip Caches. In *Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'02)*, pages 211–222, Oct. 2002.

[12] J. Kim, C. Nicopoulos, D. Park, R. Das, Y. Xie, N. Vijaykrishnan, M. Yousif, and C. Das. A Novel Dimensionally-Decomposed Router for On-Chip Communication in 3D Architectures. In *Proceedings of the International Symposium on Computer Architecture (ISCA'07)*, pages 138–149, 2007.

[13] K. Kumagai, C. Yang, S. Goto, T. Ikenaga, Y. Mabuchi, and K. Yoshida. System-in-Silicon Architecture and its application to an H.264/AVC motion estimation fort 1080HDTV. In *Proceedings of the International Solid-State Circuits Conference (ISSCC'06)*, pages 430–431, Feb. 2006.

[14] F. Li, C. Nicopoulos, T. Richardson, Y. Xie, V. Narayanan, and M. Kandemir. Design and Management of 3D Chip Multiprocessors Using Network-in-Memory. In *Proceedings of the International Symposium on Computer Architecture (ISCA'06)* , pages 130–141, June 2006.

[15] F. Li, C. Nicopoulos, T. Richardson, Y. Xie, V. Narayanan, and M. Kandemir. Design and Management of 3D Chip Multiprocessors Using Network-in-Memory. In *Proceedings of the International Symposium on Computer Architecture (ISCA'06)*, pages 130–141, 2006.

[16] P. S. Magnusson et al. Simics: A Full System Simulation Platform. *IEEE Computer*, 35(2):50–58, Feb. 2002.

[17] M. M. K. Martin, D. J. Sorin, B. M. Beckmann, M. R. Marty, M. Xu, A. R. Alameldeen, K. E. Moore, M. D. Hill, and D. A. Wood. Multifacet General Execution-driven Multiprocessor Simulator (GEMS) Toolset. *ACM SIGARCH Computer Architecture News (CAN'05)*, 33(4):92–99, Nov. 2005.

[18] H. Matsutani, M. Koibuchi, Y. Yamada, D. F. Hsu, and H. Amano. Fat H-Tree: A Cost-Efficient Tree-Based On-Chip Network. *IEEE Transactions on Parallel and Distributed Systems*, 20(8):1126–1141, Aug. 2009.

[19] H. Matsutani, Y. Take, D. Sasaki, M. Kimura, Y. Ono, Y. Nishiyama, M. Koibuchi, T. Kuroda, and H. Amano. A Vertical Bubble Flow Network using Inductive-Coupling for 3-D CMPs. In *Proceedings of the International Symposium on Networks-on-Chip (NOCS'11)* , pages 49–56, May 2011.

[20] N. Miura, H. Ishikuro, T. Sakurai, and T. Kuroda. A 0.14pJ/b Inductive-Coupling Inter-Chip Data Transceiver with Digitally-Controlled Precise Pulse Shaping. In *Proceedings of the International Solid-State Circuits Conference (ISSCC'07)*, pages 358–359, Feb. 2007.

[21] N. Miura, D. Mizoguchi, M. Inoue, K. Niitsu, Y. Nakagawa, M. Tago, M. Fukaishi, T. Sakurai, and T. Kuroda. A 1Tb/s 3W Inductive-Coupling Transceiver for Inter-Chip Clock and Data Link. In *Proceedings of the International Solid-State Circuits Conference (ISSCC'06)*, pages 424–425, Feb. 2006.

[22] D. Park, S. Eachempati, R. Das, A. K. Mishra, V. Narayanan, Y. Xie, and C. R. Das. MIRA: A Multi-layered On-Chip Interconnect Router Architecture. In *Proceedings of the International Symposium on Computer Architecture (ISCA'08)*, pages 251–261, 2008.

[23] V. F. Pavlidis and E. G. Friedman. 3-D Topologies for Networks-on-Chip. *IEEE Transactions on Very Large Scale Integration Systems*, 15(10):1081–1090, Oct. 2007.

[24] A. Radecki, H. Chung, Y. Yoshida, N. Miura, T. Shidei, H. Ishikuro, and T. Kuroda. 6W/25mm2 Inductive Power Transfer for Non-Contact Wafer-Level Testing. In *Proceedings of the International Solid-State Circuits Conference (ISSCC'12)*, pages 230–232, Feb. 2012.

[25] R. S. Ramanujam and B. Lin. Randomized Partially-Minimal Routing on Three-Dimensional Mesh Networks. *IEEE Computer Architecture Letters*, 7(2):37–40, July 2008.

[26] T. D. Richardson, C. Nicopoulos, D. Park, V. Narayanan, Y. Xie, C. Das, and V. Degalahal. A Hybrid SoC Interconnect with Dynamic TDMA-Based Transaction-Less Buses and On-Chip Networks. In *Proceedings of International Conference on VLSI Design (VLSID'06)*, pages 657–664, Jan. 2006.

[27] S. Saito, Y. Kohama, Y. Sugimori, Y. Hasegawa, H. Matsutani, T. Sano, K. Kasuga, Y. Yoshida, K. Niitsu, N. Miura, T. Kuroda, and H. Amano. MuCCRA-Cube: a 3D Dynamically Reconfigurable Processor with Inductive-Coupling Link. In *Proceedings of the Field-Programmable Logic and Applications (FPL'09)*, pages 6–11, Sept. 2009.

[28] A. Sheibanyrad, F. Petrot, and A. Janstch. *3D Integration for NoC-Based SoC Architectures*. Springer, 2010.

[29] Y. Yuan, A. Radecki, N. Miura, I. Aikawa, Y. Take, H. Ishikuro, and T. Kuroda. Simultaneous 6Gb/s Data and 10mW Power Transmission using Nested Clover Coils for Non-Contact Memory Card. In *Proceedings of the Symposium on VLSI Circuits (VLSIC'10)*, pages 199–200, June 2010.

[30] Y. Yuan, Y. Yoshida, N. Yamagishi, and T. Kuroda. Chip-to-Chip Power Delivery by Inductive Coupling with Ripple Canceling Scheme. In *Proceedings of the International Conference on Solid State Devices and Materials (SSDM'07)*, pages 502–503, Sept. 2007.