A Power-Efficient 3-D On-Chip Interconnect for Multi-Core Accelerators with Stacked L2 Cache

Kyungsu Kang, Sangho Park and Jong-Bae Lee
Design Technology Team
Memory Business, Samsung
Hwaseong, Korea

Luca Benini
Integrated Systems Laboratory
ETH Zurich
Zurich, Switzerland

Giovanni De Micheli
Integrated Systems Laboratory
EPFL
Lausanne, Switzerland

***Abstract*—The use of multi-core clusters is a promising option for data-intensive embedded applications such as multi-modal sensor fusion, image understanding, mobile augmented reality. In this paper, we propose a power-efficient 3-D on-chip interconnect for multi-core clusters with stacked L2 cache memory. A new switch design makes a circuit-switched Mesh-of-Tree (MoT) interconnect reconfigurable to support power-gating of processing cores, memory blocks, and unnecessary interconnect resources (routing switch, arbitration switch, inverters placed along the on-chip wires). The proposed 3-D MoT improves the power efficiency up to 77% in terms of energy-delay product (EDP).**

I. Introduction

Parallel computing architectures have recently taken the center stage in research and development from embedded systems to workstations as they can provide high-performance computing with good power-efficiency. The most visible examples in this trend are GP-GPUs such as NVIDIA Kepler, HyperCore, STMicroelectronics Platform 2012, ADAPTEVA Epiphany, KALRAY MPPA, CAVIUM OCTEON, Tilera TILE-Gx, and Intel Xeon Phi. In such architectures consisting of multiple cores with on-chip shared (cache) memory units, it is crucial to implement a high-throughput and low-latency on-chip interconnect to connect the on-chip components (i.e., cores and memory units) that are tightly coupled with each other [1].

For 3-D integrated circuits (ICs), many 3-D on-chip interconnects have been studied. Most 3-D on-chip interconnects employ packet-switching, which deliver packets among homogeneous nodes through packet routers. Since the major difference between 3-D and 2-D on-chip interconnects is the presence of short vertical links, many researches have focused mainly on optimizing the vertical communications, for example, designing a NoC-Bus hybrid interconnect [2] to reduce vertical hop counts, designing more energy-efficient routers to increase energy-delay product (EDP) [3], and decomposing router components into the third dimension to reduce wire latency within the router [4]. However, despite the previous efforts, packet-switched 3-D on-chip interconnects still suffers from the inherent long network latency mainly due to the hop-by-hop packet communications. On the contrary, some researchers proposed circuit-switched 3-D on-chip interconnects that reduce the interconnect latency [5] [6].

G. Beanato [7] first proposed an extension of circuit-switched Mesh-of-Tree (MoT) topology [1] into 3-D integration for multi-core clusters with 3-D stacked shared L1 scratchpad memory. E. Azarkhish [8] [9] presented two synthesizable

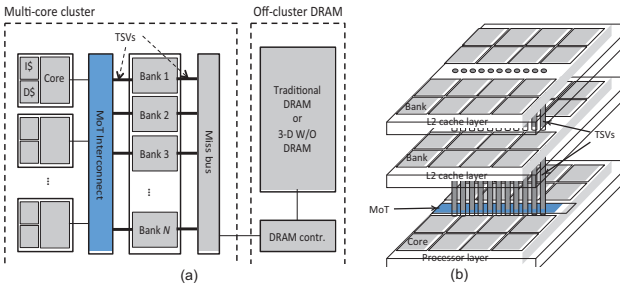


Fig. 1. 3-D multi-core cluster with MoT interconnect: (a) schematic view; (b) geometric view.

3-D MoT interconnects for multi-core clusters with stacked L1 scratchpad memory. From the results of physical implementation, the two interconnects were evaluated and compared with 2-D MoT in terms of chip area and interconnect latency. In [10], a circuit-switched 3-D MoT topology for stacked L2 scratchpad memory was proposed. Pipelining routing switches was used in order to exploit the delay asymmetry (i.e., longer horizontal delay than vertical delay) of 3-D ICs and, thus, increase the interconnect performance.

In this paper, we propose a power-efficient 3-D on-chip interconnect that is suitable for a multi-core cluster consisting of multiple cores and a shared multi-banked L2 cache memory where the L2 cache memory is stacked onto the multi-core die. The contributions of this paper are as follows. First, we propose a new design of routing switch for the 3-D MoT interconnect so that the interconnect is reconfigurable to support power-gating of processing cores, memory units, and interconnect circuits (e.g., inverters placed along the on-chip wires). The reconfigurability of on-chip interconnect makes it possible to adjust power states of the on-chip interconnect to applications running on the system in a power-efficient manner. Second, we investigate several packet-switched 3-D on-chip interconnects recently proposed in literatures and compare them with the proposed circuit-switched 3-D MoT interconnect in terms of performance. Performance result of each on-chip interconnect is measured using multi-core system-level simulator [11] with real parallel benchmark programs [12].

II. Circuit-Switched 3-D MoT Interconnect

Figure 1 (a) shows a schematic view of a multi-core cluster with 3-D stacked L2 cache using through-silicon vias (TSVs).

978-3-9815370-6-2/DATE16/

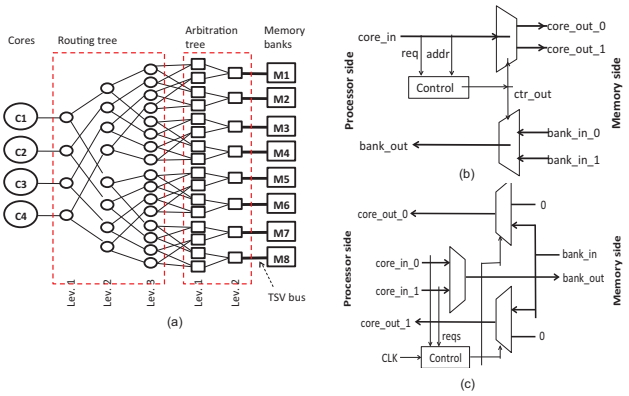
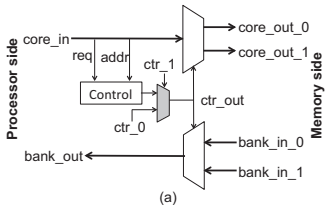


Fig. 2. Circuit-switched 3-D MoT interconnect: (a) an example of 4×8 3-D MoT interconnect, where empty circles represent routing switches and empty squares represent arbitration switches; (b) routing switch; (c) arbitration switch.



ctr_1	ctr_0	Packet directions
GND	X	Depends on conventional "Control" output
VDD	GND	All packets go/come through port 0
VDD	VDD	All packets go/come through port 1

(b)

Fig. 3. Modified routing switch and control scheme: (a) modified routing switch; (b) packet routing based on the control signals.

The cluster consists of simple cores each with its own private L1 instruction and data caches. The multi-banked stacked L2 cache consists of multiple SRAM banks. Each stacked SRAM bank is connected with the 3-D MoT interconnect through a TSV bus (i.e., a set of TSVs for address, data, and control signals). In case of instruction miss, *Miss bus* handles line refills in a round-robin manner towards the off-cluster DRAM. Figure 1 (b) shows a geometry view of the 3-D multi-core cluster. MoT interconnect is placed in the middle of the core tier, which makes it easier that memory access latency from each core is well balanced.

Figure 2 (a) shows a 3-D MoT interconnect consisting of four cores and eight L2 cache banks stacked on the cores. When a core accesses its target cache bank, a combinational path is created through two kinds of binary trees, i.e., *routing tree* and *arbitration tree*. The combinational path is able to support low-latency and non-blocking communications between cores and cache banks [1]. As shown in Figure 2 (b), routing switch consists of a MUX, a DEMUX, and a combinational control logic which routes packets individually based on the packet's address field. In order the packet to arrive the target cache bank, it must be arbitrated among the other simultaneous packets heading for the same cache bank through the arbitration switch shown in Figure 2 (c). In the control logic, a round-robin algorithm is implemented for a starvation-free arbitration.

III. Reconfigurable 3-D MoT Interconnect for Power Management of Multi-Core Cluster

Power management (i.e., power-gating) of cores, memory units, and on-chip interconnect links can be supported by modifying a circuit of routing switch of 3-D MoT interconnect. Figure 3 shows the proposed modified routing switch that is basically the same circuit structure as the original one except for the additional multiplexor (i.e., gray one in Figure 3 (a)). The new multiplexor with the corresponding control signals (shown in Figure 3 (b)) makes packet routing reconfigurable so that packets can traverse the interconnect network either in the conventional way or a user-defined way. In the conventional way, packet direction (i.e., either port 0 or port 1) is determined based on the packet destination address (i.e., the target L2 cache bank index). In the user-defined way, packet direction is determined based on the two control signals (i.e., ctr_0 and ctr_1), not related to the destination address.

Figure 4 shows an example of how to adapt the modified routing switch for the use of power-gating in 3-D MoT interconnect, where four cores and eight memory banks are connected to each other. In Figure 4, half cache banks (M0, M1, M6, and M7) and the corresponding interconnect circuits (i.e., routing switches, arbitration switches, and inverters placed along the on-chip wires) are turned off. For that, the routing switches at the second level of the routing tree run on the user-defined mode, whereas other routing switches run on the conventional mode. This architecture does not need to modify the conventional cache architecture and packet routing scheme, because the cache data that is supposed to be stored at the power-gated cache banks will evenly be distributed the rest of cache banks based on the packet destination address (i.e., cache bank index). For example, in Figure 4, the cache data for M0 (bank index of 000) and M1 (bank index of 001) will be stored at M2 (bank index of 010) and M3 (bank index of 011), respectively, because the routing switches in the user-defined mode at the second level of routing tree make the second digit of cache bank index ignored for packet routing. For the same reason, the cache data for M6 (bank index of 110) and M7 (bank index of 111) will be stored at M4 (bank index of 100) and M5 (bank index of 101), respectively. If cache banks are turned off at runtime, dirty cache blocks in the power-off banks must be written back to the off-cluster memory for data coherency. After turning on the cache banks again, the old cache data that does not belong to cache banks any more will be removed by the cache replacement policy.

The reconfigurable 3-D MoT interconnect supports different interconnect delays because of the inherent asymmetry of 3-D integration in the wire lengths. Figure 5 shows an example of wire lengths comparison between the two power states, i.e., 1) where all cores and cache banks run, and 2) where four cores and eight cache banks are run and the rest are turned off. A wide disparity of wire lengths between the two power states makes a difference of several clock cycles in cache access latency. As 3-D integration makes it possible to stack DRAM main memory and, thus, reduces the access latency of the main memory, the miss penalty of last-level cache might be decreased. Then, the reduction in the L2 cache access latency, in conjunction with power-gating some cache resources, gives more effects on the power efficiency.

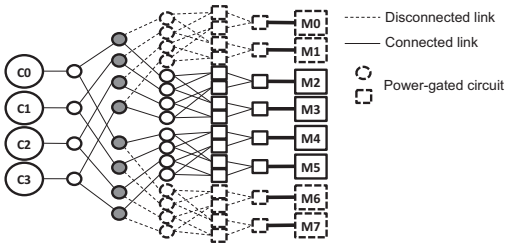


Fig. 4. Example of the use of power-gating in a 3-D MoT interconnect where four cores and eight memory banks are connected each other. Most routing switches (i.e., the white circles) route packets in the conventional way, while the routing switches at the second level of the routing tree (i.e., the gray circles) route packets in the user-defined mode.

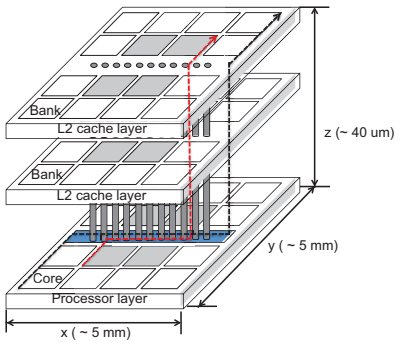


Fig. 5. Example of wire lengths comparison between two power-states, i.e., 1) where all cores and cache banks run, and 2) where four cores and eight cache banks are run and the rest are turned off.

IV. Experimental Result

We performed experiments using a 3-D multi-core cluster with a multi-banked shared L2 cache memory stacked on top of the multi-core die, as shown in Figure 1. Sixteen processing cores are integrated in the multi-core cluster and each core is considered to be ARM Cortex-A5 with 16KB/16KB instruction and data caches. The core clock frequency is assumed to be 1GHz. The stacked L2 cache consists of 32 SRAM banks of two tiers. Each bank has a capacity of 64KB. The size of a cache bank and the propagation delay from bank I/Os to memory core cells within a SRAM cache bank are estimated from CACTI [13]. For TSV bonding, Micro-bumps bonding with a minimum pitch of 40 μm by 50 μm is used [14]. In order to estimate the latency of 3-D MoT interconnect, the delay for the longest possible link between cores and cache banks is estimated by using Elmore distributed RC delay model [15]. To estimate power consumption of core, L2 cache, and interconnect, we used power models in [19], [13], and [20], respectively. For the performance evaluation of real applications, we employed Graphite [11]. Table I shows the details of configuration. For simulation benchmarks, SPLASH-2 benchmark suite [12] was used. For the performance comparisons among the popular packet-switched 3-D on-chip interconnects, which have been studied earlier in interatures, and our MoT interconnect, we chose the three packet-switched 3-D on-chip interconnects, i.e., *True 3-D Mesh*, *3-D Hybrid Bus-Mesh* [2], and, *3-D Hybrid Bus-Tree* [21].

Figure 6 shows the experimental results of L2 cache access latency and execution time of real benchmarks for each

TABLE I. Architecture Configurations

Feature	Discription
Core	1GHz, 4 - 16 cores, in-order execution
L1 I/D cache	Private, 4KB capacity (per-core), 32B line, 4-way associative, LRU replacement, 1 cycle latency
L2 cache	Shared, 32B line, 8-way associativity, 64KB capacity (per bank), - Full connection: 32 banks, 12 cycle latency - PC16-MB8: 8 banks, 9 cycle latency - PC4-MB32: 32 banks, 9 cycle latency - PC4-MB8: 8 banks, 7 cycle latency
DRAM	One controller, 2Gb capacity, 4KB page size, Latency; - 200ns (off-chip 2-D DRAM) [18] - 63ns (on-chip 3-D DRAM from JEDEC) [17] - 42ns (on-chip 3-D DRAM from Weis work) [16]

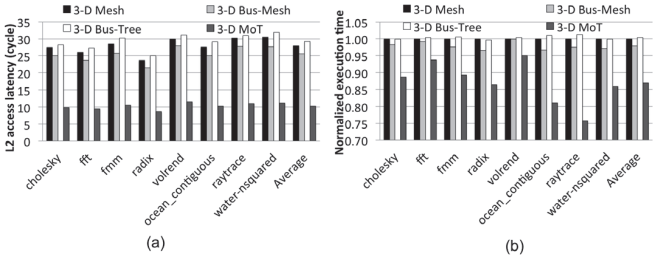


Fig. 6. Performance comparisons of four 3-D on-chip interconnects, i.e., True 3-D Mesh, 3-D Hybrid Bus-Mesh, 3-D Hybrid Bus-Tree, and 3-D MoT: (a) L2 cache access latency in terms of clock cycle; (b) application execution time where DRAM access latency is 200ns.

interconnect. 3-D Hybrid Bus-Mesh shows better performance (i.e., lower L2 cache access latency) than True 3-D Mesh, which proves that the use of a bus structure for vertical communications may reduce the L2 cache access latency by exploiting the short vertical links, in conjunction with the reduction in the number of hop accesses. 3-D Hybrid Bus-Tree shows the worst performance among the other interconnects. Even though 3-D Hybrid Bus-Tree reduces the number of hop accesses even more than 3-D Hybrid Bus-Mesh, the increased vertical bus accesses in 3-D Hybrid Bus-Tree may offset the benefit from hop access reduction or make the performance even worse. 3-D MoT reduces the execution time by 13.01%, 11.16%, and, 13.34%, on average, compared with 3-D Mesh, 3-D Hybrid Bus-Mesh, and 3-D Hybrid Bus-Tree, respectively.

Figure 7 (a) and (b) show the experimental results of energy delay product (EDP) and application's execution time with respect to the power states, respectively, where DRAM access latency is 200ns. In Figure 7 (a), in cases of cholesky, fft, volrend, and raytrace, PC4-MB32 reduces EDP up to 66% (by 44% on average) compared with Full connection. As shown in Figure 7 (b), cholesky, fft, volrend, and raytrace show the reduction in the execution time up to 33% (by 19% on average) as the number of cores increases from 4 to 16, while fmm, radix, ocean_contiguous, and water-nsquared show the reduction in the execution time up to 69% (by 64% on average). Thus, for the applications that have limited scalability for the parallelism, power efficiency will be achieved by assigning smaller number of cores to the applications. When compared with Full connection, PC16-MB8 reduces EDP up to 18% (by 13% on average) in cases of fft, fmm, volrend, raytrace, and water-nsquared owing to the reduction in the number of L2 cache banks (and the corresponding decrease in the interconnect power and L2 cache access latency). As shown in Figure 7 (b), when compared with Full connection,

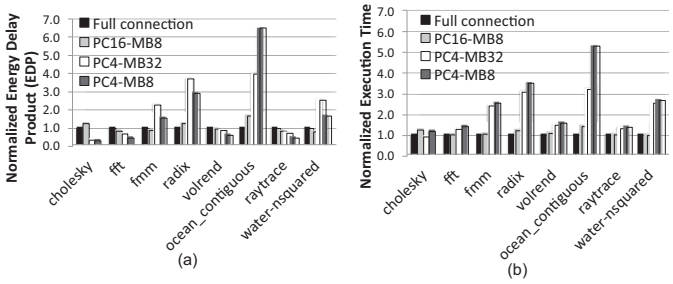


Fig. 7. Power efficiency of the four power states of target architecture, i.e., Full connection, PC16-MB8, PC4-MB32, and PC4-MB8; (a) energy delay product (EDP) and (b) application execution time where DRAM access latency is 200ns.

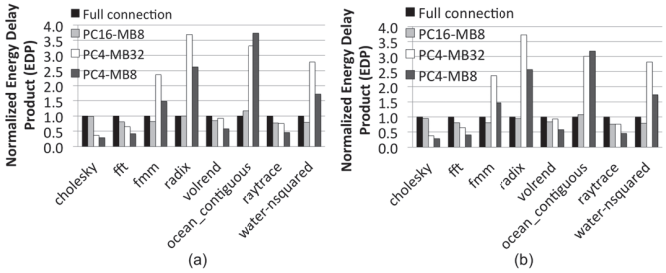


Fig. 8. Power efficiency of the four power states of target architecture, i.e., Full connection, PC16-MB8, PC4-MB32, and PC4-MB8; (a) energy delay product (EDP) where DRAM access latency is 63ns; (b) EDP where DRAM access latency is 42ns.

PC16-MB8 increases the execution time up to 8.6% (4.7% on average) for fft, fmm, volrend, raytrace, and water-nsquared, while PC16-MB8 increases the execution time up to 31% (24% on average) for the other programs (i.e., cholesky, radix, and ocean_contiguous). PC4-MB8 reduces EPD up to 77% (52% on average) for cholesky, fft, volrend, and raytrace owing to the reduction in both the number of cores and the number of L2 cache banks, compared with Full connection. All these experimental results show that the reconfigurable 3-D MoT interconnect capable of power-gating technique is necessary to exploit various programs characteristics such as parallelism scalability and L2 cache demand. Figure 8 shows power efficiency of the four power states where the DRAM access latency is 63ns and 42ns, respectively, which proves that power efficiency resulting from power-gating of cache banks increases as the DRAM access latency decreases. Compared with PC16-MB8 when DRAM access latency is 200ns, PC16-MB8 reduces EDP for more benchmark programs when DRAM access latency is 63ns and 42ns.

V. Conclusion

We presented a power-efficient 3-D on-chip interconnect for multi-core clusters with stacked L2 cache memory. The new design of routing switch for 3-D MoT interconnects can make the interconnects reconfigurable to support power-gating of processing cores, cache memory banks, and the corresponding interconnect links. This reconfigurability makes it possible to adjust power states of the interconnects to application's characteristics such as scalability for parallelism and L2 cache demand. The experimental results showed that proper power state on the 3-D MoT interconnect reduces energy-delay product (EDP) up to 77% (by 48% on average). In this paper, we also investigated several packet-switched 3-D on-chip interconnects and compared them with the 3-D MoT interconnect. The low latency of 3-D MoT interconnect is suitable for an on-chip interconnect within a multi-core cluster where multiple cores and the heavily shared multi-banked L2 cache communicates each other with ultra-low latency.

References

[1] A. Rahimi et al., "A fully-synthesizable single-cycle interconnection network for shared-L1 processor clusters," in Proc. DATE, 2011, pp. 1-6.

[2] F. Li et al., "Design and management of 3D chip multiprocessors using network-in-memory," in Proc. ISCA, 2006, pp. 130-141.

[3] J. Kim et al., "A novel dimensionally-decomposed router for on-chip communication in 3-D architecture," in Proc. ISCA, 2007, pp. 138-149.

[4] D. Park et al, "MIRA: a multi-layered on-chip interconnect router architecture," in Proc. ISCA, 2008, pp. 251-261.

[5] H. Saito et al., "A chip-stacked memory for on-chip SRAM-rich SoCs and processors," IEEE Journal of Solid-State Circuits, vol. 45, no. 1, Jan. 2010.

[6] S.-H. Chou et al., "No cache-coherence: a single-cycle ring interconnection for multi-core L1-NUCA sharing on 3D chips," In Proc. DAC, 2009, pp. 587-592.

[7] G. Beanato et al., "3D-LIN: a configurable low-latency interconnect for multi-core clusters with 3D stacked L1 memory," In Proc. VLSI-SoC, 2012, pp. 30-35.

[8] E. Azarkhish, I. Loi, and L. Benini, "3D logarithmic interconnect: stacking multiple L1 memory dies over multi-core clusters," In Proc. NoCS, 2013, pp. 1-2.

[9] E. Azarkhish, I. Loi, and L. Benini, "A case for three-dimensional stacking of tightly coupled data memories over multi-core clusters using low-latency interconnects," IET Computers & Digital Techniques, vol. 7, no. 5, Sep. 2013.

[10] K. Kang, L. Benini, and G. D. Micheli, "A high-throughput and low-latency interconnection network for multi-core clusters with 3-D stacked L2 tightly-coupled data memory," In Proc. VLSI-SoC, 2012.

[11] J. E. Miller et al., "Graphite: A distributed parallel simulator for multicores," in Proc. HPCA, 2010, pp. 1-12.

[12] S. C. Woo et al., "The SPLASH-2 programs: characterization and methodological considerations," In Proc. ISCA, 1995, pp. 24-36.

[13] D. Tarjan, S. Thoziyoor, and N. P. Jouppi, "CACTI 4.0," HP Laboratories, Palo Alto, CA, Tech. Rep. HPL-2006-86, Jun. 2006.

[14] E. J. Marinissen et al., "Wafer probing on fine-pitch micro-bumps for 2.5D- and 3D-SICs," Int. Report IMEC [online]. Available: http://www.swtest.org/swtw_library/2011proc/PDF/S04_03_Marinissen_SWTW2001.pdf.

[15] G. Katti et al., "Electrical modeling and characterization of through silicon via for three-dimensional ICs," IEEE Transactions on Electron Devices, vol. 57, no. 1, Jan. 2010.

[16] C. Weis, I. Loi, and L. Benini, "Exploration and optimization of 3-D integrated DRAM subsystems," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 32, no 4, Apr. 2013.

[17] JEDEC Standard, "Wide I/O single data rate (WIDE I/O SDR)," JESD229, Dec. 2011.

[18] MICRON DDR3 SDRAM [online]. Available: http://www.micron.comproductsdramddr3-sdram.

[19] S. Li et al., "McPAT: an integrated power, area, and timing modeling framework for multicore and manycore architectures," in Proc. MICRO-42, 2009, pp. 469-480.

[20] W. Liao and L. He, "Full-chip interconnect power estimation and simulation considering concurrent repeater and flip-flop insertion," in Proc. ICCAD, 2003, pp. 574-580.

[21] N. Madan et al., "Optimizing communication and capacity in a 3D stacked reconfigurable cache hierarchy," In Proc. HPCA, 2009, pp. 262-274.