UCLA
UCLA Previously Published Works

Title
Three-Dimensional Wafer Scale Integration for Ultra Large Scale Cognitive Systems

Permalink
https://escholarship.org/uc/item/4mv925fz

Authors
Wan, Zhe
Iyer, Subramanian S.

Publication Date
2017-10-01

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Three-Dimensional Wafer Scale Integration for Ultra-Large-Scale Cognitive Systems

Zhe Wan and Subramanian S. Iyer
Center for Heterogenous Integration and Performance Scaling (CHIPS), Electrical and Computer Engineering Department
University of California, Los Angeles, Los Angeles, CA, USA
z.wan@ucla.edu

***Abstract*—In this paper, we present a method to use three-dimensional wafer scale integration (3D-WSI) to build large-scale cognitive systems up to the scale of a human brain (10^{10} neurons, 10^{13} synapses). We analyze the effect of scaling by simulating some cognitive systems. The result exhibits attractive properties of the 3D-WSI technology, as compared with the traditional integration scheme using printed circuit boards. At the scale of a human brain, the 3D-WSI system reduces the communication latency by about 10X, while consuming at least 100X less communication power.**

Keywords—system scaling, three-dimensional-wafer-scale-integration, cognitive system.

I. INTRODUCTION

Since cognitive applications, such as image recognition, have been successfully demonstrated using neural-network based algorithms, more efforts are devoted to the associated hardware to improve performance and energy-efficiency. For example, the IBM TrueNorth [1] chip has reduced power consumption using a distributed spike-based design. For ultra-large neural networks, multi-chip cognitive systems are required. Unfortunately, the conventional integration scheme, using the printed circuit boards (PCB), results in a significant power consumption from communication. For example, more than 70% of the power in the NS16e system (16 TrueNorth chips integrated on PCBs) is consumed by the supporting peripheral for communication [2]. Therefore, novel integration schemes are needed to reduce the communication power consumption of large-scale cognitive systems.

In this work, we model and simulate the operation of cognitive systems in two scenarios: integrated using the PCBs, or integrated using three-dimensional wafer scale integration (3D-WSI). The simulated result exhibits advantages of the 3D-WSI technology for ultra-large-scale system integration.

II. ULTRA-LARGE-SCALE COGNITIVE SYSTEM INTEGRATION

A. Two-Dimensional Printed Circuit Board Integration (2DI)

Most state-of-the-art cognitive systems utilize PCB(s) and backplane(s) to integrate multi-chip systems [2,3]. A model of such (2DI) system is shown in Fig. 1. This model is a graph, where the chips and the chip-to-chip interconnects are represented by the nodes and the edges, respectively. Chips on the same board are interconnected by high-speed Serializer/Deserializer (SerDes) links. Board-to-board interconnect is supported by SerDes links and FPGA chips. The boards are arranged in a cubic 3D-mesh.

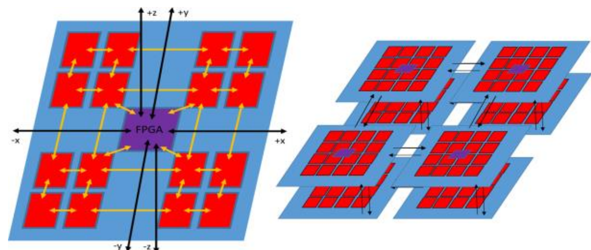

Fig. 1. 2DI System integrated on the PCBs. On-board chip-to-chip connection is via SerDes. Board-to-board connection is via the FPGA and SerDes.

The latency of communication (spiking events) has three components, namely the time spent in the router (t_r), in the SerDes circuit (t_{SD}), and in the physical channel (t_{phy}), as summarized in Table 1. The power consumption due to communication is the sum of the intra-board power (mainly due to the logic) and the inter-board power (due to the SerDes links), as summarized in Table 2.

B. Three-Dimensional-Wafer-Scale-Integration (3D-WSI)

Intensive chip-to-chip communication within the cognitive systems requires massive bandwidth, which is not energy-efficient when realized by the SerDes links. Wafer-scale integration emerges as a promising candidate to provide energy-efficient bandwidth, by utilizing the fine pitch interconnect (FPI). One example of wafer-scale integration is the FACETS project [4] in which FBEOL process is used to make 10μm-pitch FPI to interconnect the reticles. Another approach is to use the Si interconnect fabric (Si-IF) to integrate known good dies (KGDs), which addresses the issue of yield on the wafer and grants heterogeneity of the system. We have demonstrated 10μm-pitch FPI on the Si-IF [5], which can be further reduced to 2μm pitch.

Table 1. Components of communication latency in 2DI and 3D-WSI systems.

	t_r	t_{SD}	t_{phy}
2DI	20ns	130ns	1ns (on-board), 5ns (inter-board)
3D-WSI	20ns	0ns	1ns (one TSV), 1ns*# of repeaters (VEL)

Table 2. Comparison between SerDes in 2DI systems and FPI in 3D-WSI systems.

	SerDes (2DI) [6]	Fine pitch interconnect (3D-WSI)
Wire pitch	$400\mu m$ (chip-board)	$2\mu m$
Area/link	$> 1mm^2$	$2\mu m^2$
Data rate/link	$30Gbps$	$1Gbps$
Energy efficiency	$20pJ/bit$ (high speed), $136pJ/bit$ (low speed)	$< 0.2pJ/bit$ @1V (with ESD capacitors)

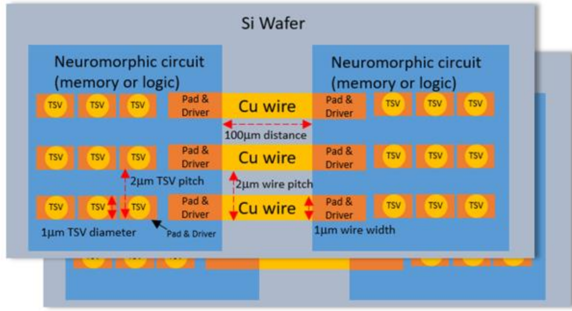

Fig. 2. A schematic of the interconnect within a 3D-WSI system.

To build a multi-wafer system that provides energy-efficient bandwidth between wafers, 3D-WSI is a viable solution [7]. In 3D-WSI technology, wafers are bonded through fusion bonding; through silicon vias (TSVs) are then made to connect wafers. State-of-the-art TSVs have achieved 2µm fine pitch [8]. In addition, bonded wafers are thinned to less than 10µm in order to electroplate the TSVs properly. In this technology, yield of the wafers depends on the wafer bonding process. Therefore, high quality low-temperature wafer bonding process (<400C) is needed.

A schematic of the interconnect within a 3D-WSI system is shown in Fig. 2. SerDes links used in the PCBs are replaced by the metal wires/TSVs. Also, FPGA is not needed in the 3D-WSI systems. Consequently, in 3D-WSI systems, $t_{SD} = 0$ (Table 1). Since FPI provides abundant wires, they can run at a lower frequency (1Gbps) to improve energy efficiency (<0.2pJ/bit), while supporting high aggregate bandwidth [9], as summarized in Table 2.

C. Vertical Express Lanes (VEL) in 3D-WSI

Because 2µm-pitch TSVs are short (≤ 10µm), vertical express lanes can be designed to accelerate long-distance communications in 3D-WSI systems. A schematic of a VEL, with a comparison with normal TSVs, is shown in Fig. 3. While a normal TSV (the blue and yellow TSVs in Fig. 3) terminates at routers at both ends, a VEL (in green) travels across intermediate node(s) by connecting to a repeater and skipping the router of that node. The latency benefit of VELs is shown in Table 2. To maximize the benefit from the VELs, they are designed for every two wafers. For example, a 100-wafer stack has 9,900 unique pairs of wafers. If each TSV link is realized by a 10-TSV bundle (20µm^2), the maximum area occupied by the VELs on a reticle is smaller than 0.2mm^2.

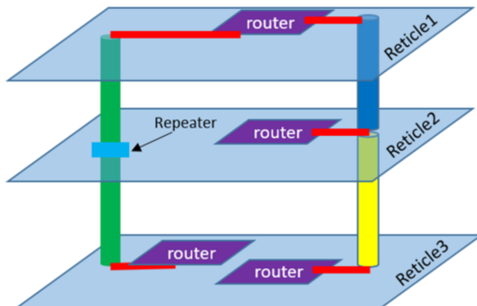

Fig. 3. Vertical Express Lanes (VEL) in the 3D-WSI system.

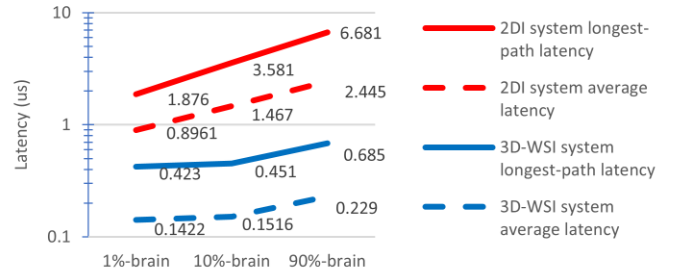

Fig. 4. Simulated average and longest-path communication latency.

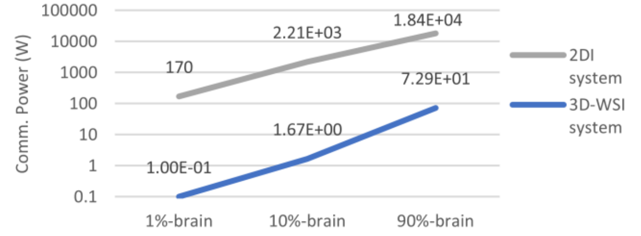

Fig. 5. Simulated communication power consumption.

III. Results and Conclusion

Simulation is performed by assigning a part of the monkey brain region to each node of the system, and using experimentally measured biological connectivity data [10] to operate this system. Systems of different sizes (1%-brain, 10%-brain, 90%-brain) are simulated. The average latency and longest-path communication latency are shown in Fig. 4. Compared with the 2DI systems, the 3D-WSI systems reduce the communication latency by a factor of 4 to 10. This reduction increases as the system size grows. The power consumed by the interconnect, including SerDes links and FPGAs in the 2DI system, or FPI in the 3D-WSI systems, is shown in Fig. 5. 3D-WSI systems reduce the communication power consumption by a factor of 100 to 1,000.

We have demonstrated that 3D-WSI is a superior solution for scaled-out cognitive systems. It provides fast and energy-efficient interconnects, and significantly enhances the performance and energy-efficiency of ultra-large-scale cognitive systems.

Acknowledgment

We would like to thank the CHIPS consortium, DARPA, UCOP (MRP-17-454999) and IBM for their support.

References

[1] S K. Esser et al., PNAS (2016): 201604850.
[2] J. Sawada et al., SC16, 2016, pp. 130-141.
[3] N. P. Jouppi et al., arXiv:1704.04760 (2017).
[4] J. Schemmel et al., ISCAS, 2010, pp. 1947-1950.
[5] A. A. Bajwa et al., ECTC, 2017, pp. 1276-1284.
[6] H. Kimura et al., JSSCC 49.12 (2014): 3091-3103.
[7] A. Kumar et al., JETC 13.3 (2017): 45.
[8] W. Lin et al., S3S, 2014, pp. 1-3.
[9] S. Jangam et al., ECTC, 2017, pp. 86-94.
[10] D. S. Modha and R. Singh, PNAS 107.30 (2010): 13485-1349