UCLA
UCLA Previously Published Works

Title
Fine-Pitch Integration Technology for Cognitive System Scaling

Permalink
https://escholarship.org/uc/item/7wc938pz

Authors
Wan, Zhe
Iyer, Subramanian

Publication Date
2018-09-01

Peer reviewed

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

Fine-Pitch Integration Technology for Cognitive System Scaling

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 discuss the use of fine-pitch integration technologies to build and scale-out cognitive systems up to the scale of a human brain (10^{10} neurons, 10^{13} synapses). Simulation result of cognitive systems based on different system scaling technologies is reported to evaluate the impact of system scaling. Fine-pitch integration technology based on three-dimensional wafer-scale integration (3D-WSI) shows promising advantage over the traditional integration scheme using printed circuit boards. To the extent of human brain scale system, the 3D-WSI system reduces the communication latency by about 10X, while consuming at least 100X less communication power.**

Keywords—system scaling, fine-pitch integration, ultra-large scale system, cognitive system.

I. INTRODUCTION

Artificial intelligence based on cognitive functionalities, such as image recognition, has attracted enormous attention due to recent advancement of the performance of the neural network with its derivatives. However, neural network is inspired by the biological brain, whose architecture is distributed, parallel while running slowly – very different from the centralized and fast traditional computing. As a result, corresponding paradigm shift in the hosting hardware is needed to optimize performance and energy efficiency. One example is the IBM TrueNorth [1] chip, which imitates the brain architecture using a spike-based distributed design. However, scale of a single system on chip (SoC) is still limited. While being a gigantic and expensive SoC, one TrueNorth chip contains one million neurons and 256 millions of synapses, which is at least 10,00X away from the human brain. A scaled-out version (NS16e) based on 16 TrueNorth chips has been design and implemented [2] on a printed circuit board (PCB), but it results in tremendous overhead in energy. More than 70% of the power in the NS16e system is consumed by the supporting peripheral for communication. Therefore, novel integration schemes are needed to reduce the communication power consumption to pave the way for large-scale cognitive systems to make them practical and affordable.

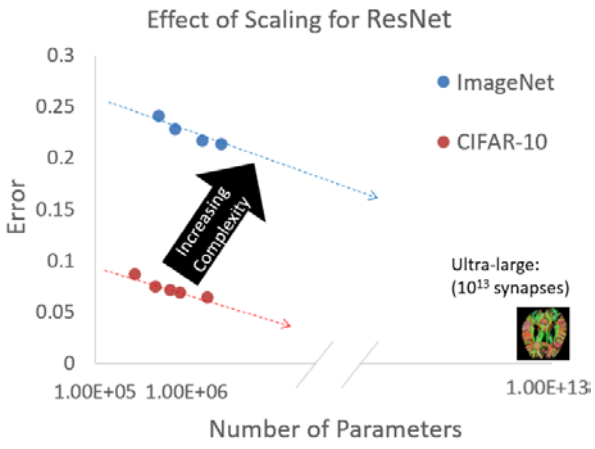

Fig. 1. Error rate of different sized ResNets against two image recognition benchmarks CIFAR-10 (classification among 10 classes of objects) and ImageNet (classification among 1,000 classes of objects). For the same ResNet architecture, a larger ResNet performs better. In the meanwhile, a more complex problem would require a larger network as well

In this paper, we explore various system integration technologies, model the communication events based on each technology, and simulate the respective systems. The simulation results show the advantage of the fine-pitch integration technologies for ultra-large scale system integration.

II. COGNITIVE SYSTEM SCALING

Many of the cognitive applications today are enabled by neural networks, whose performance depend on the scale. In general, a larger neural network, properly trained, tends to be able to achieve lower error rate for a classification problem. Although the issue of overfitting kicks in when one tries to use a very large network for a simple problem, a problem with greater complexity would nevertheless require a larger network. Such trait of neural networks can be seen from the performance of ResNet[11] (Fig. 1) which is one of the state-of-the-art neural networks on image recognition benchmarks. Naively, a brute-force solution to such problems is an enormously large network which simply maps every possible input to the correct answer.

A larger neural network means not only more neurons (processing element) but also more synapses, which are the interconnect links among the neurons. The best cognitive machine we know, the brain, has a highly-interconnected and distributed architecture. The brain is also very energy efficient. It consumes only around 20W of power and performs better than the supercomputers in various challenges. The goal of this work is to explore the methods to scale the cognitive system up to the "size" of the human-brain in terms of the number of neurons and synapses. At the meantime, the system built by this method should also be energy-efficient.

To scale out the cognitive systems, a straightforward approach is to make very large SoCs. One example is the IBM TrueNorth neuromorphic chip [1], which is a huge and expensive SoC. But the 1 million neurons contained in this SoC is far from enough comparing with the brain, which has tens of billions of neurons. As a result, such chips are integrated on circuit boards for system scaling. Unfortunately, the circuit board integration is also costly in terms of energy efficiency. The high-speed Serializer/Deserializer (SerDes) links on the circuit board are power-hungry and slow. In an event when two neurons try to communicate, neuron1 sends a package or a spike to neurons2. The energy and latency associated with this event depend on the spatial locations of neuron1 and neuron2. Such cost increases tremendously when these two neurons are on different chips.

The performance and energy efficiency of a scaled cognitive system based on the TrueNorth chip [2] is shown in Fig. 2. In this case, neural networks are designed for the CIFAR-100 dataset to classify 100 classes of objects. As the network contains more chips and grows larger, the classification accuracy increases, but the energy efficiency of such system, defined by frames per second per Watt (equivalent to frames/Joule), decreases because now more communication events are out-of-chip and cost more energy.

III. 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. 3. 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.

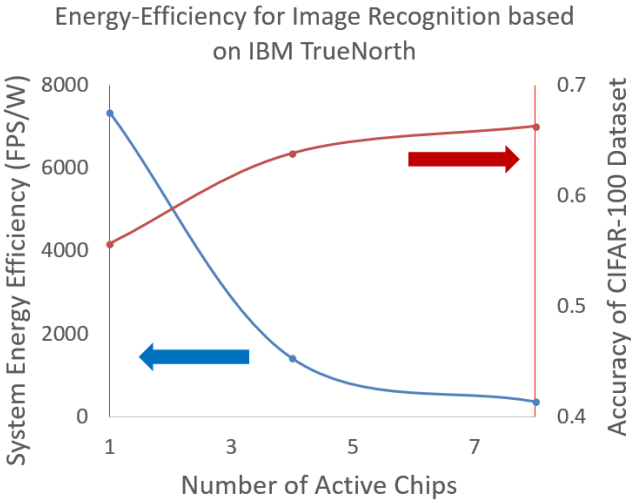

Fig. 2. Energy-efficiency and image recognition accuracy of a scaled cognitive system based on the IBM TrueNorth chip. As the neural network becomes larger and use more chips, the image recognition accuracy increases, but the energy-efficiency decreases due to the inefficient chip-to-chip communication events.

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

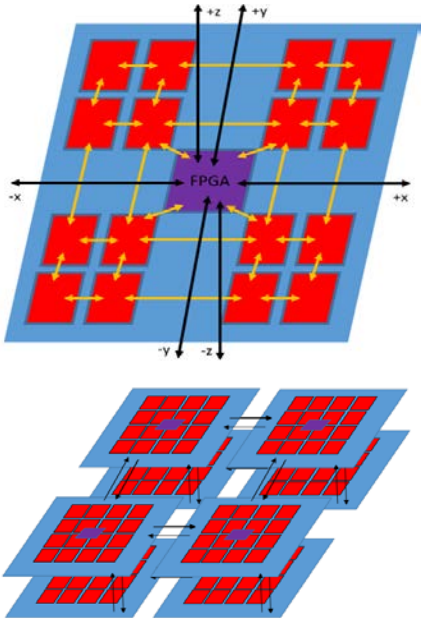

Fig. 3. 2DI System integrated on the PCBs. Top: on-board chip-to-chip connection is via SerDes (yellow arrows). Bottom: board-to-board connection is via the FPGA (purple chip) and SerDes (black arrows).

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)

to the logic) and the inter-board power (due to the SerDes links), as summarized in Table 2.

B. Fine-Pitch Wafer-Scale Integration

Intensive chip-to-chip communication within the cognitive system requires massive chip-to-chip 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 among the reticles. Another approach is to use the Si interconnect fabric (Si-IF) to integrate known good dies (KGDs), which addresses yield challenge on the wafer and grants heterogeneity for the system. We have demonstrated 10μm-pitch FPI on the Si-IF (Fig. 4) at UCLA CHIPS, which can be further reduced to 2μm pitch [5].

To build a multi-wafer system that provides energy-efficient bandwidth between wafers, three-dimensional wafer-scale integration (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 3D-WSI technology has achieved 2μm fine pitch TSVs [8]. A simplified process flow of 3D-WSI is presented in Fig. 5. The silicon wafer strata1 is first processed at the front-side. Then, the wafer is bonded to a handle wafer at the same side. Afterwards, the strata1 is thinned from the backside by grinding from around 700μm (300mm wafer thickness) to about 7μm to accommodate 1μm diameter (2μm pitch) through-silicon-vias (TSVs) with proper aspect ratio (7μm: 1μm) for the Cu plating process. The thinning is followed by a backside oxide/dielectric deposition and CMP. Strata1 is then bonded (using the back side) to the front-side of the strata0. Finally, the handle wafer of strata1 is removed prior to further fabrication of the metal layer and TSVs at the front-side. The top of this stacked wafer pair is similar to the top of strata0, and therefore additional wafers can be bonded to this stack in series.

A schematic of the interconnect within a 3D-WSI based system is shown in Fig. 6. SerDes links used in the PCBs are replaced by the horizontal metal wires or 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

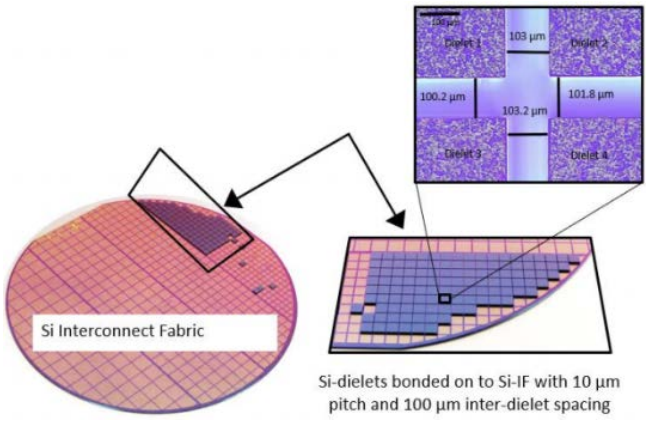

Fig. 4. CHIPS Si interconnect fabric (Si-IF) based on 100mm wafer made at UCLA CHIPS.

efficiency (<0.2pJ/bit), while supporting high aggregate bandwidth [9], as summarized in Table 2.

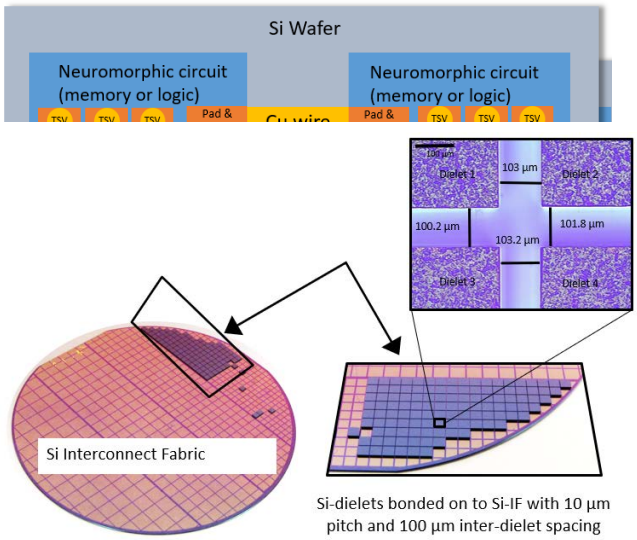

Fig. 4. CHIPS Si interconnect fabric (Si-IF) based on 100mm wafer made in UCLA.

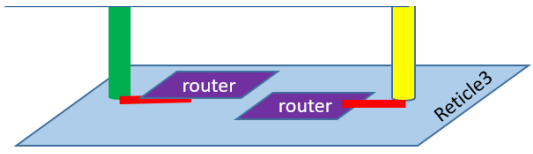

Fig. 7. Vertical Express Lanes (VEL) in the 3D-WSI system.

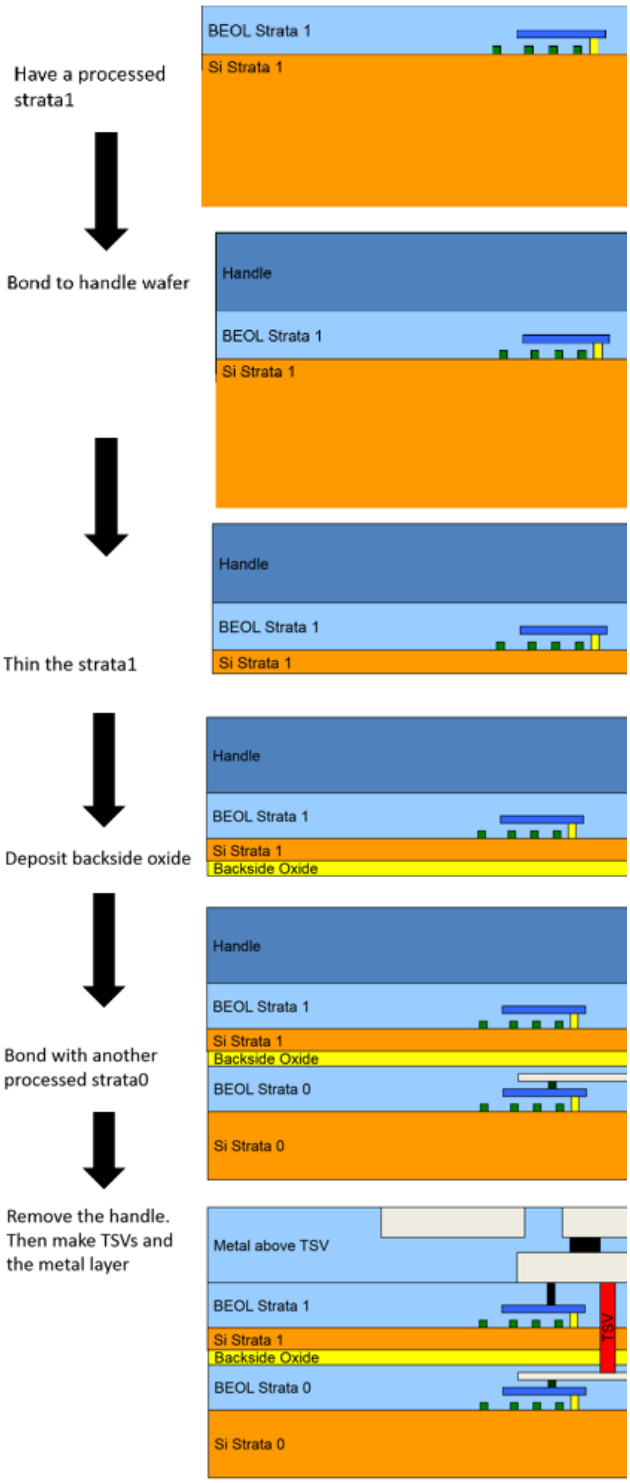

Fig. 5. A brief process flow of the three-dimensional wafer-scale integration (3D-WSI). It is a via-last technology which can be used for face-to-face, and face-to-back wafer bonding. By doing a rigorous thinning on the wafer to less than $10\mu m$ thickness before making the TSVs, TSVs of $1\mu m$ diameter can be fabricated in FBEOL process.

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

Due to the short distance in the vertical direction in the stacked wafers, concatenating multiple TSVs will still end up with rather short channels. Therefore, vertical express lanes (VEL) can be made to accelerated long-distance communications in the 3D-WSI systems. Fig. 7 shows the VEL with a comparison to the normal TSVs. A VEL (in green) is made from concatenated TSVs, which goes across intermediate node(s) by connecting to a repeater and skipping the router of that node(s). Because the physical transit time (t_{phy} from Table 2) is a small part of the total latency, VEL works as a long-link in the system without significant latency penalty. Thanks to the small on-wafer area consumed by the TSVs, dedicated VELs can be designed between for every two wafers..

IV. Results and Conclusion

Experimentally measured neural connectivity in the Macaque monkey brain [10] is used to simulate the systems of different sizes using different integration technologies as listed in Table 3. During the simulation, each node of the system is assigned a part of the Macaque monkey brain region. Communication events are triggered by the probability distribution equivalent to the measured connectivity data. The average latency and longest-path communication latency are shown in Fig. 8, showing that 3D-WSI systems improve the latency by a factor of 4 to 10. Fig.9 shows the communication power consumption extracted from the same simulations, where 3D-WSI systems reduce the communication power consumption by a factor of 100 to 1,000.

We have demonstrated that fine-pitch integration technology provides fast and energy-efficient interconnects, and significantly enhances the performance and energy-efficiency of ultra-large-scale cognitive systems.

Acknowledgment

This work was supported in part by Semiconductor Research Corporation (SRC). We would also like to thank members of the CHIPS consortium, DARPA, UC MRP-17-454999 and IBM for their support of this work.

Table 3. Scales of the simulated cognitive systems.

Integration Tech.	1% brain (432 nodes)	10% brain (4,256 nodes)	90% brain (34,048 nodes)
PCB	27 boards	266 boards	2,128 boards
3D-WSI	4 wafers	32 wafers	266 wafers

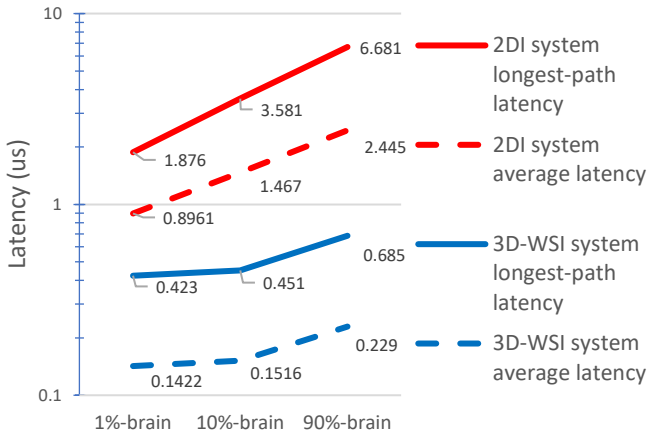

Fig. 8. Simulated average and longest-path communication latency.

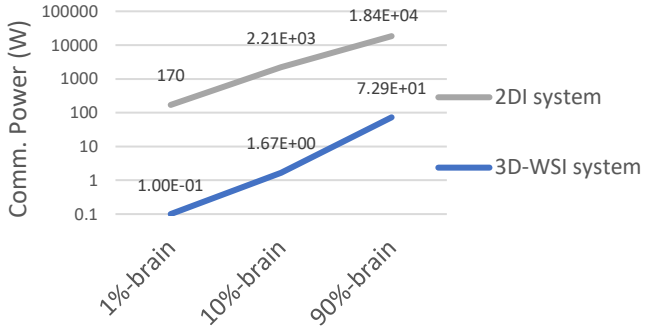

Fig. 9. Simulated communication power consumption.

References

[1] Esser, Steven K., et al. "Convolutional networks for fast, energy-efficient neuromorphic computing." *Proceedings of the National Academy of Sciences* (2016): 201604850.

[2] Sawada, Jun, et al. "Truenorth ecosystem for brain-inspired computing: scalable systems, software, and applications." *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE Press, 2016. N. P. Jouppi et al., arXiv:1704.04760 (2017).

[3] Jouppi, Norman P., et al. "In-datacenter performance analysis of a tensor processing unit." Computer Architecture (ISCA), 2017 ACM/IEEE 44th Annual International Symposium on. IEEE, 2017.

[4] Schemmel, Johannes, et al. "A wafer-scale neuromorphic hardware system for large-scale neural modeling." *Circuits and systems (ISCAS), proceedings of 2010 IEEE international symposium on*. IEEE, 2010.

[5] Bajwa, Adeel A., et al. "Heterogeneous Integration at Fine Pitch (≤ 10 µm) Using Thermal Compression Bonding." 2017 IEEE 67th Electronic Components and Technology Conference (ECTC). IEEE, 2017.

[6] Kimura, Hiroshi, et al. "A 28 Gb/s 560 mW multi-standard SerDes with single-stage analog front-end and 14-tap decision feedback equalizer in 28 nm CMOS." *IEEE Journal of Solid-State Circuits* 49.12 (2014): 3091-3103.

[7] Kumar, Arvind, et al. "Toward Human-Scale Brain Computing Using 3D Wafer Scale Integration." *ACM Journal on Emerging Technologies in Computing Systems (JETC)* 13.3 (2017): 45.

[8] Lin, Wei, et al. "Prototype of multi-stacked memory wafers using low-temperature oxide bonding and ultra-fine-dimension copper through-silicon via interconnects." *SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S), 2014 IEEE*. IEEE, 2014.

[9] Jangam, SivaChandra, et al. "Latency, bandwidth and power benefits of the SuperCHIPS integration scheme." Electronic Components and Technology Conference (ECTC), 2017 IEEE 67th. IEEE, 2017.

[10] Modha, Dharmendra S., and Raghavendra Singh. "Network architecture of the long-distance pathways in the macaque brain." *Proceedings of the National Academy of Sciences* 107.30 (2010): 13485-13490.

[11] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.