IEEE Asian Solid-State Circuits Conference
November 7-9, 2016/Toyama, Japan

A 5.5GHz 0.84TOPS/mm^2 Neural Network Engine with Stream Architecture and Resonant Clock Mesh

Shengshuo Lu, Zhengya Zhang, Marios Papaefthymiou
University of Michigan, Ann Arbor, MI, USA

***Abstract*—This paper presents an ultra-high-performance neural network engine fabricated in a 65nm CMOS technology. The 0.9mm^2 core relies on an energy-efficient resonant clock mesh running at 5.5GHz to achieve 0.76 8-bit TOPS, improving throughput by over 4x, area efficiency by over 8x, and energy-delay-area product by over 1.8x compared to previous state-of-the-art neural network designs. Achieving a charge recovery rate of 63%, the resonant clock mesh enables the deployment of a deeply-pipelined stream architecture and high-speed stream buffers with a sub-5W power consumption.**

Keywords—Neural Network, Application-Specific Integrated Circuits, High-Performance Clocking, Stream Architecture.

I. INTRODUCTION

Large-scale deep learning methods are now being widely used in big data analytics. One important factor that has enabled the rapid advancement of deep learning is the continued improvement in computational performance and power efficiency. Deep learning accelerators [1-3] have the potential to offer superior performance and power efficiency over conventional general-purpose processors.

Previously reported accelerators deliver high computational performance through massive parallelism. Typically, parallelism is combined with supply voltage scaling to improve power efficiency, albeit at a significant area penalty [1-3]. An area-efficient alternative to achieve high computational performance is to operate at an increased clock frequency. This approach may lead to excessive power densities, however, and imposes tight clock jitter/skew constraints. The focus of this work is on maximizing computational performance and area efficiency within a sustainable power envelope.

We present a neural network accelerator test-chip for deep learning that relies on a multi-GHz clock to achieve high performance and high area efficiency. An ultra-low-skew clock is distributed across the entire chip through an optimized clock mesh. To attain power efficiency, resonant clocking is used to recover the bulk of clock mesh power, which is the largest single component of the core's power consumption.

Fabricated in a 65nm process, the 0.9mm^2 accelerator test-chip runs at 5.5GHz. It achieves a performance of 0.76TOPS and an area efficiency of 0.84TOPS/mm^2, outperforming the latest deep learning accelerators by at least 4x [1-3] and improving on area efficiency by at least 8x, as shown in Fig. 1. By taking advantage of energy-efficient resonant clocking, we use high-speed resonant stream buffers to provide weights to neurons without stalling. The resulting stream architecture with multi-GHz resonant clocking demonstrates a new design point that optimizes both area efficiency and power efficiency. Measured in a combined energy-delay-area (EDA) product, i.e., the inverted product of area efficiency and energy efficiency, the resulting figure of merit is 0.13TOPS^2/W·mm^2, surpassing state-of-the-art designs by at least 1.8x.

The test-chip incorporates a number of architecture and circuit techniques to overcome the challenges associated with a 5.5GHz design, including high datapath speed, high memory bandwidth, ultra-low clock skew, and high power efficiency. First, to enable multi-GHz datapath operation, we adopt a stream architecture with gate-level pipelining. By structuring the datapath in a wide and shallow fashion, we reduce the cost of data alignment and decrease the number of registers by 1.5x over typical serial datapath designs. Second, to provide an ultra-high memory bandwidth of 5.6Tb/s in support of the high-speed parallel datapath, we utilize a resonant stream buffer memory. Third, we propose a novel high-speed dual single-phase resonant clock mesh for global clock distribution to eliminate extra decap overhead, resulting in a worst-case clock skew of 2ps and 63% charge-recovery rate.

II. NEURAL NETWORK ACCELERATOR DESIGN

The neural network accelerator implements a 128x16 (128 visible nodes and 16 hidden nodes) fully-connected layer. This fully-connected network can implement the connectivity of any same-size neural network, for example, any convolutional neural network. Typical uses of such an accelerator include mapping of multiple neural network modules, a fully-connected module (up to 128x16), part of a fully-connected

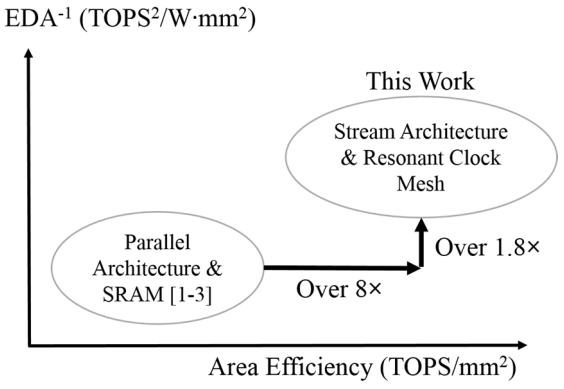

Fig. 1: Design scheme comparison.

This work was supported in part by NSF under grants No. CCF 1320027 and No. CCF 1161505.

978-1-5090-3700-1/16/$31.00 ©2016 IEEE

module, or any of the above followed by a small classifier (up to 16x16). The proof-of-concept design can be scaled up in a straightforward manner to support much larger networks.

As shown in Fig. 2, inputs are provided to the visible layer. Each visible node value V is multiplied by the weights W associated with its hidden layer connections, and the sum of the products forms H for each hidden node. Each value H is then passed through a nonlinear activation function, and the resulting value H' goes through another multiplication-summation step to reach the output layer to produce classification results.

III. Stream Architecture

A. Stream Buffer Memory

The neural network accelerator is implemented using a stream architecture, shown in Fig. 3, along with a 16Kbit stream buffer memory. Input data is shifted into the visible nodes registers. The stream buffers store the weights and continuously rotate them to line up with the corresponding inputs for multiplication. The products are passed to an adder tree for summation. As the stream buffers provide a new set of weights every cycle, the datapath processes input data as they arrive, yielding one hidden node output every cycle. In the stream buffers, data rotation is performed through local shifting to match the speed of the datapath, providing an ultra-high bandwidth of 5.6Tb/s.

For this ultra-high-speed design, a stream buffer is more suitable than conventional memories, including SRAM and registers. Specifically, achieving multi-GHz operation with standard SRAM arrays poses significant design challenges. To achieve fast SRAM operation, significant overheads are introduced in the SRAM cell and peripheral circuitry. Moreover, SRAM efficiency depends on size. The peripheral overhead can be easily amortized if SRAM array size is sufficiently large. In our case, however, memory depth is kept to 16 rows, so the area and energy overhead introduced by peripheral circuitry will be dominating. In the case of

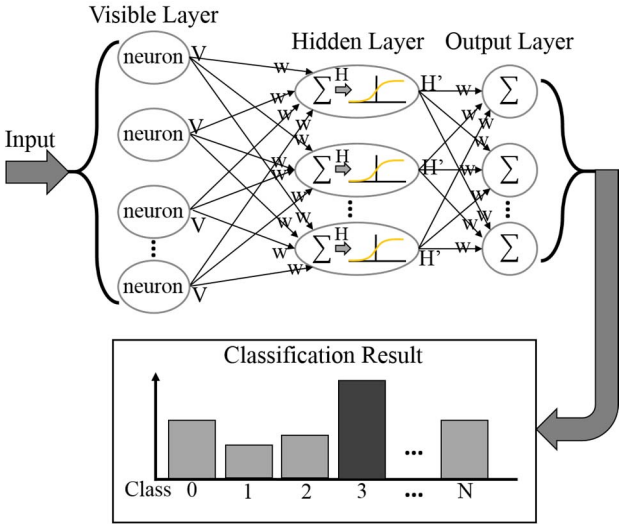

Fig. 2: Fully-connected neural network.

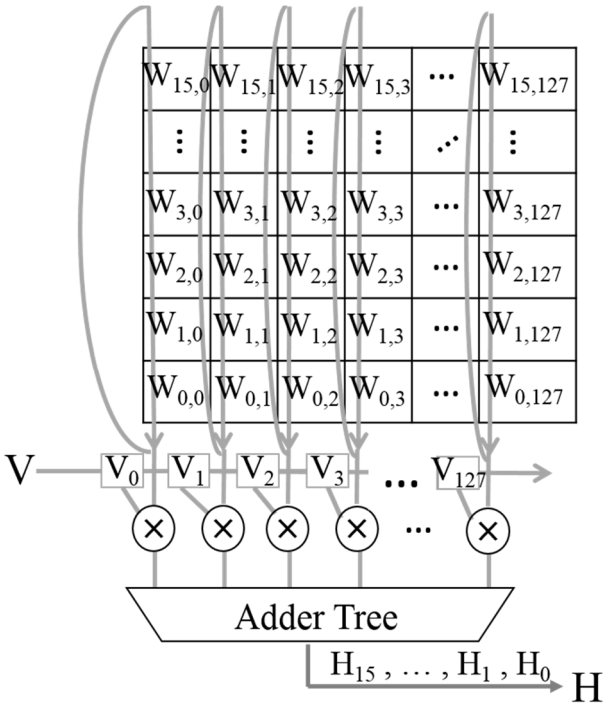

Fig. 3: High-speed stream architecture using stream buffers and a wide and shallow datapath.

conventional registers, although it is feasible to provide the necessary memory bandwidth, the complexity of the clock distribution network to the registers is considerable, and the place/route of distributed registers to meet a 5.5GHz timing constraint introduces significant challenges.

B. Stream Datapath

Gate-level pipelining is used to enable an aggressive 5.5GHz clock frequency. To keep register overheads at a minimum, a wide (parallel) and shallow datapath is designed with low pipeline depth, as opposed to a narrow (serial) and deep datapath, as the latter incurs disproportionately high register overheads to store and align inputs. Wide and shallow datapaths feature significantly lower register overheads, as they operate on all data simultaneously upon their arrival. This insight led to a Ladner-Fischer adder [4] designed with maximum fanout of 3 to minimize register count subject to a fanout load constraint. Compared to a typical serial adder, this structure cuts pipeline depth in half, decreasing register count by 1.5x with minimal combinational logic gates overhead.

IV. Resonant Clock Mesh

To enable 5.5GHz operation, a clock mesh has been designed to minimize clock skew. Unlike conventional designs [5-6], which use mostly high-layer metal for the clock mesh, this test-chip uses both high-layer and layer-3 metal for the clock mesh to further minimize skew. Fig. 4 shows the simulated clock waveform. In Fig. 5, resonant clock mesh simulation results indicate that for every register, clock insertion delay from the clock root is between 4.4ps and 6.4ps, yielding a worst-case clock skew of 2ps across the entire core.

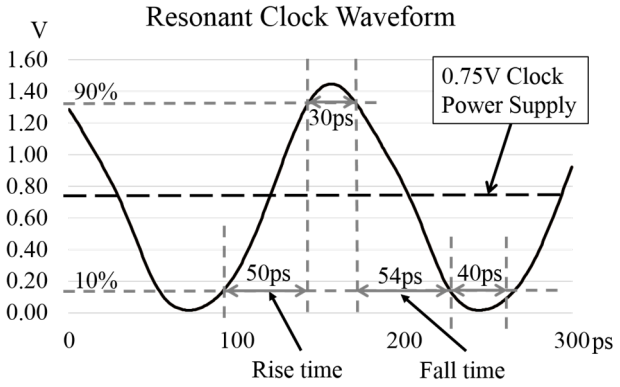

Fig. 4: Simulated clock waveform.

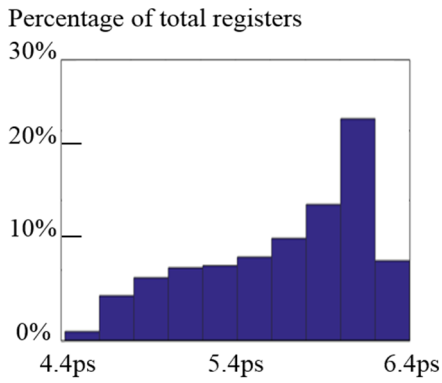

Fig. 5: Histogram of simulated clock insertion delay.

A significant portion of the power in this design is consumed by the high-speed clock mesh due to its high wire loading and large number of registers. To reduce power, we use resonant clocking to recycle charge using LC resonance [5]. Based on simulation results, the clock distribution charge recovery rate is 63%. Moreover, compared to a clock mesh driven by conventional drivers which consumes 1.563W based on simulation, the resonant clock consumes 710mW based on measurements, resulting in at least 54% power savings over conventional clock mesh.

The chip floorplan is shown in Fig. 6. Using 28 on-chip center-tapped inductors, clock signals CLKA and CLKB are generated with 180 degrees phase difference. Each clock signal is supplied to its corresponding half of the core by a separate clock mesh. The two meshes have similar capacitance, and function as decap to each other. Thus, unlike conventional single-phase resonant clocking [5-6], this novel dual single-phase design does not require any additional decap, saving area and simplifying design. A total of 108 cross-coupled NMOS pairs serve as negative resistance to compensate for the energy loss of the clock distribution network. The 5.5GHz clock rate in this test-chip achieves the same level of speed as the fastest resonant clock design in 22nm [6].

V. Measurement Results & Conclusion

The chip has been fabricated in a 65nm CMOS technology, and total die area is 0.903mm^2, including 0.509mm^2 for datapath and memory, and 0.394mm^2 for the 28 on-chip inductors. The die photo is shown in Fig. 7. Measurement results are given in Table 1. The chip operates at a maximum clock rate of 5.5GHz. With a 1.35V supply, the datapath and memory consume 4.21W, while the resonant clock mesh consumes 710mW with a 0.75V supply. The engine is capable of classifying MNIST digits [7] with 86% accuracy, which is comparable to that of a state-of-the-art neural network design that used the same benchmark [1].

The combination of stream architecture with a resonant clock mesh results in a high-performance neural network

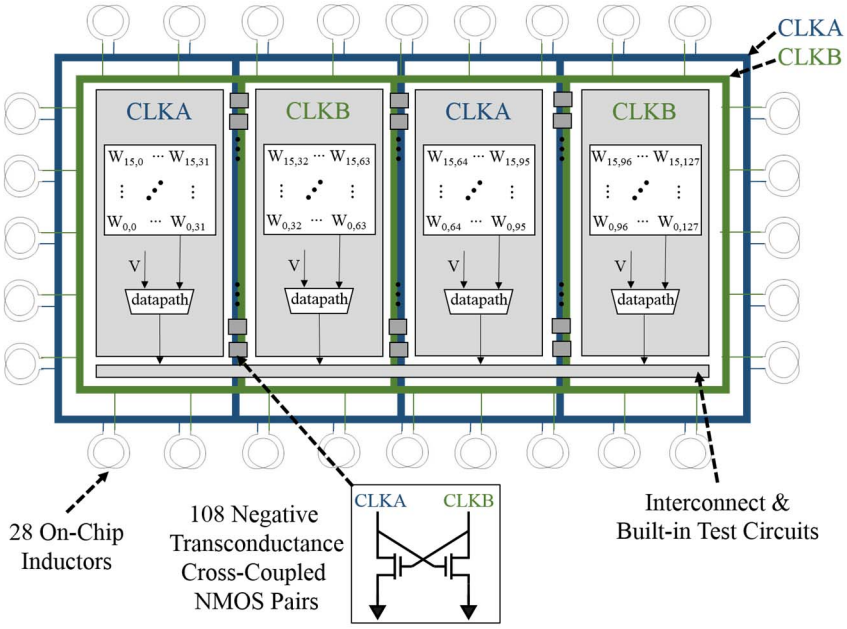

Fig. 6: Test-chip floor plan.

Table 1: Chip characteristics.

	This work
Technology	65nm
Clock Frequency	5.5GHz
Core Supply	1.35V
Core Power	4.21W
Clock Supply	0.75V
Clock Power	0.71W
Total Power	4.92W
Core Area	0.509mm^2 (Datapath: 0.260mm^2 Memory: 0.204mm^2 Periphery: 0.045mm^2)
Inductor Area	0.394mm^2
Total Area	0.903mm^2

engine with improved throughput, area efficiency, and EDA product. As shown in Table 2, running at 5.5GHz, the test-chip achieves 0.76TOPS, 0.84TOPS/mm^2, and 0.13TOPS^2/W·mm^2. Compared to the designs published to date [1-3], this design improves throughput by over 4x, area efficiency by over 8x, and EDA product by over 1.8x. Given that inductor area accounts for approximately half of total area, area efficiency and EDA product could be further improved by approximately 2x through embedding of the inductors in the package, as demonstrated in [8].

ACKNOWLEDGMENT

We thank Chester Liu, Phil Knag, Jung Kuk Kim and Yajing Chen for their help with the design.

REFERENCES

[1] Kim, Jung Kuk, Phil Knag, Thomas Chen, and Zhengya Zhang. "A 640M pixel/s 3.65 mW sparse event-driven neuromorphic object recognition processor with on-chip learning." In 2015 Symposium on VLSI Circuits, pp. C50-C51, 2015.

[2] Sim, Jaehyeong, Jun-Seok Park, Minhye Kim, Dongmyung Bae, Yeongjae Choi, and Lee-Sup Kim. "14.6 A 1.42 TOPS/W deep convolutional neural network recognition processor for intelligent IoE systems." In 2016 IEEE International Solid-State Circuits Conference (ISSCC), pp. 264-265, 2016.

[3] Chen, Yu-Hsin, Tushar Krishna, Joel Emer, and Vivienne Sze. "14.5 Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks." In 2016 IEEE International Solid-State Circuits Conference (ISSCC), pp. 262-263, 2016.

[4] Ladner, Richard E., and Michael J. Fischer. "Parallel prefix computation." Journal of the ACM (JACM) 27, no. 4 (1980): 831-838.

[5] Sathe, Visvesh, Srikanth Arekapudi, Charles Ouyang, Marios Papaefthymiou, Alexander Ishii, and Samuel Naffziger. "Resonant clock design for a power-efficient high-volume x86–64 microprocessor." In 2012 IEEE International Solid-State Circuits Conference (ISSCC), pp. 68-70, 2012.

[6] Shan, David, Phillip Restle, Doug Malone, Rob Groves, Eric Lai, Michael Koch, Jason Hibbeler et al. "Resonant clock mega-mesh for the IBM z13 TM." In 2015 Symposium on VLSI Circuits, pp. C322-C323, 2015.

[7] LeCun, Yann, Corinna Cortes, and Christopher J. C. Burges. MNIST database. http://yann.lecun.com/exdb/mnist.

[8] Ou, Tai-Chuan, Zhengya Zhang, and Marios C. Papaefthymiou. "A 934MHz 9Gb/s 3.2 pJ/b/iteration charge-recovery LDPC decoder with in-package inductors." In 2015 IEEE Asian Solid-State Circuits Conference (A-SSCC), pp. 1-4, 2015.

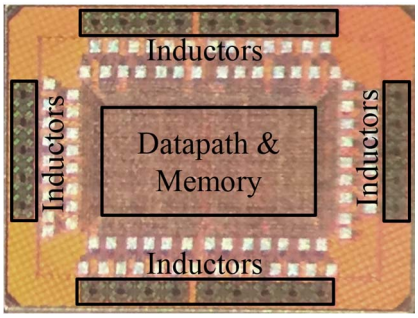

Fig. 7: Die photo.

Table 2: Comparison with previously published neural network chips.

	This work		VLSI 15 [1]	ISSCC 16 [2]	ISSCC 16 [3]
Function	Fully Connected Neural Network		Sparse Coding Neural Network	Convolutional Neural Network	Convolutional Neural Network
Technology	65nm		65nm	65nm	65nm
Clock Frequency	5.5 GHz		635 MHz	125 MHz	250MHz
Power	4.92 W		268.2 mW	45 mW	45 mW
Area	0.903 mm^2		1.82 mm^2	16 mm^2	12.25 mm^2
Throughput	41.6 GPixelPS	0.76 TOPS	10.2 GPixelPS	0.064 TOPS	0.042 TOPS
Area Efficiency	46.1 GPixelPS/mm^2	0.84 TOPS/mm^2	5.6 GPixelPS/mm^2	0.004 TOPS/mm^2	0.0034 TOPS/mm^2
Energy Efficiency	8.46 GPixelPS/W	0.154 TOPS/W	37.3 GPixelPS/W	1.42 TOPS/W	0.933 TOPS/W
Area Efficiency × Energy Efficiency (EDA^{-1})	390 GPixelPS^2/W·mm^2	130 ×10^{-3} TOPS^2/W·mm^2	208 GPixelPS^2/W·mm^2	5.68 ×10^{-3} TOPS^2/W·mm^2	3.17 ×10^{-3} TOPS^2/W·mm^2