A Wide Dynamic Range Sparse FC-DNN Processor with Multi-Cycle Banked SRAM Read and Adaptive Clocking in 16nm FinFET

Sae Kyu Lee[1,2], Paul N. Whatmough[1,3], Niamh Mulholland[1], Patrick Hansen[1], David Brooks[1], Gu-Yeon Wei[1]

[1] Harvard University, Cambridge, MA
[2] IBM Research, Yorktown Heights, NY
[3] Arm Research, Boston, MA

***Abstract*—Always-on classifiers for sensor data require a very wide operating range to support a variety of real-time workloads and must operate robustly at low supply voltages. We present a 16nm always-on wake-up controller with a fully-connected (FC) Deep Neural Network (DNN) accelerator that operates from 0.4–1V. Calibration-free automatic voltage/frequency tuning is provided by tracking small non-zero Razor timing-error rates, and a novel timing-error driven sync-free fast adaptive clocking scheme provides resilience to on-chip supply voltage noise. The model access burden of neural networks is relaxed using a multi-cycle SRAM read, which allows memory voltage to be reduced at iso-throughput. The wide operating range allows for high performance at 1.36GHz, low-power consumption down to 750µW and state-of-the-art raw efficiency at 16-bit precision of 750 GOPS/W dense, or 1.81 TOPS/W sparse.**

Keywords—deep neural networks; always on; adaptive clocking;

I. INTRODUCTION

Mobile and IoT devices increasingly require "always-on" intelligence: sensor-data classification workloads running all the time, such as audio keyword spotting (KWS), face detection/matching, and human activity recognition (HAR). Such workloads typically consist of pre-processing, feature extraction, and a fully-connected deep neural network (FC-DNN) classifier (Fig. 1). To achieve this within a realistic energy budget, these workloads require a self-contained subsystem with: dedicated HW support for energy efficient classification; wide operating range to support varying throughput, latency, and energy budgets; on-chip memory access only; calibration-free automatic V_{DD}/F_{CLK} tuning; and resilience to V_{DD} noise from workload transients and aggressive clock/power gating.

This paper presents a 16nm flip-chip packaged test chip for a complete DNN processor to enable energy efficient always-on operation. The processor achieves wide dynamic range (0.4-1.0V) to meet the varying workload demands, operating down to near-threshold supply voltage for ultra-low-power applications. Energy efficiency is improved across the operating range via multi-cycle banked SRAM reads, while a novel low-latency adaptive clocking scheme provides resilience to voltage noise and reduces unnecessary margins. The processor achieves low power consumption down to 750µW and 151nJ/inference for MNIST at 98.51% accuracy.

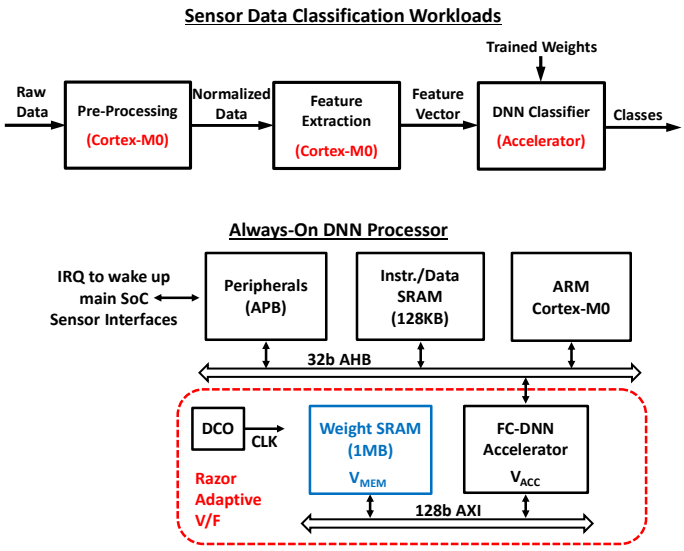


Fig. 1. Overview of sensor classification workloads (top) and 16nm test chip block diagram (bottom).

II. ALWAYS-ON DNN PROCESSOR

A. Processor Architecture

The DNN processor, shown in Fig. 1, integrates an ARM Cortex-M0 with 128KB SRAM, with a FC-DNN accelerator and on-chip 1MB low-power banked SRAM. The M0 acts as the wake-up controller to the sensor interfaces and implements sensor data pre-processing and feature extraction. Fig. 2 shows the FC-DNN accelerator architecture. The FC-DNN accelerator is a 2nd generation design, with a five-stage pipeline, 8-way fixed-point SIMD datapath and optimizations for 16b/8b operands and sparse activation data. The accelerator architecture includes sequencing that can exploit sparse activation data to reduce the number of SRAM accesses and MAC operations. During the activation stage, output activations are compared to a programmable threshold to eliminate small neurons from downstream processing, achieved by storing a list of active neurons in a small SRAM. A 1MB on-chip SRAM stores the model weights, allowing the processor to operate autonomously without incurring energy and latency overheads of off-chip memory access (c.f. [1]).

978-1-5386-5404-0/18/$31.00 ©2018 IEEE

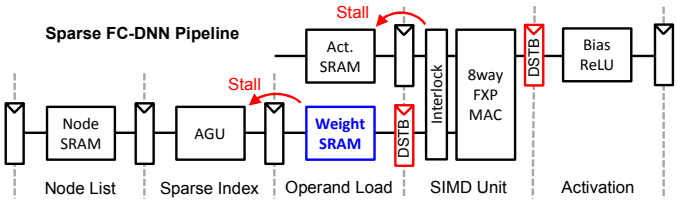


Fig. 2. Sparse FC-DNN pipeline with DSTB Razor latches and dynamic sequencing for sparse activations. Both the activation load and the weight address generation are stallable.

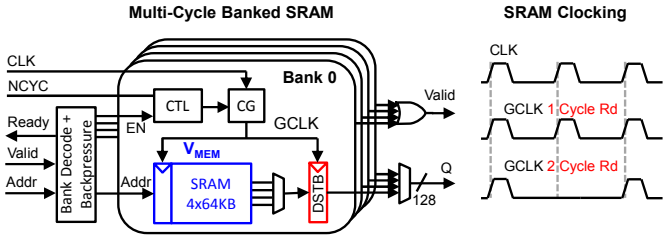


Fig. 3. Simplified multi-cycle banked weights SRAM architecture and clocking

To maximize energy efficiency across a wide operating range, it is important to automatically tune supply voltage (V_{DD}) to match clock frequency (F_{CLK}) demand across process, voltage, temperature and aging (PVTA) delay variations. The accelerator and the weights SRAM operate on independent voltage domains to enable optimum V_{DD}/F_{CLK} tuning. The Supply voltages for the accelerator (V_{ACC}) and weights SRAM (V_{MEM}) are automatically tuned to match clock frequency (F_{CLK}) demand across delay variations by using "Razor" distributed in-situ timing error detection circuits on critical paths to track a very small non-zero timing error rate, without costly calibration.

B. *Multi-Cycle Banked SRAM Architecture*

Accessing the large FC-DNN model stored in on-chip SRAM dominates power consumption. To further improve energy efficiency across the entire operating range, we propose a multi-cycle banked SRAM architecture to aggressively scale down memory voltage without impacting throughput. The proposed design, shown in Fig. 3, splits the DNN model across four autonomous SRAM banks. This relaxes the timing requirements of any individual bank, which can be combined with aggressive voltage scaling to increase energy-efficiency at iso-throughput. A counter and clock gating (CG) logic enables one- or two-cycle read latencies for the SRAM. As shown in Fig. 4(a), when executing in dense mode, bank contention does not occur as the weights are striped across the banks and are read sequentially. However, in sparse operation, bank collisions occasionally occur due to non-contiguous reads that trigger stalls. Nevertheless, performance degradation is negligible relative to the benefits gained. Fig. 4(b) plots the stall cycles incurred for MNIST inference against network sparsity, by sweeping the activation threshold. Even at 90% sparsity, the SRAM reads incur <2.6% additional stall cycles, at which point the classification accuracy degrades due to aggressive pruning of the activation nodes. At iso-accuracy, the stall cycle penalty is less than 1%.

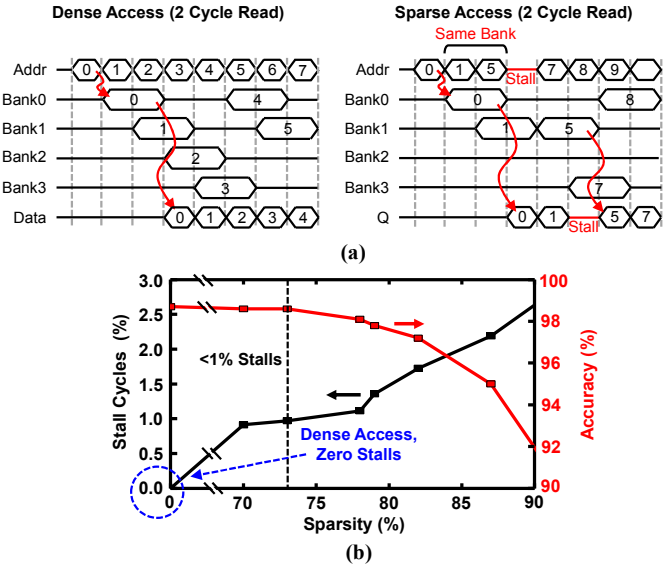


Fig. 4. Multi-cycle read: (a) timing, (b) stalls from contention

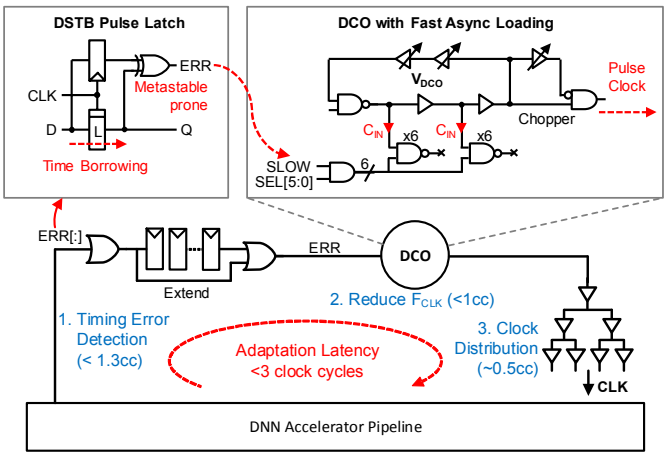


Fig. 5. Razor-driven adaptive clocking: DSTB pulse latch and DCO with metastability-safe fast loading stage removes sync flop latency

C. *Low-Latency Razor Adaptive Clocking*

Since DNN algorithms are probabilistic and inherently robust to noise, the DNN accelerator can tolerate small amounts of error with minimal impact on classification accuracy [3]. However, fast voltage droops are still problematic, as large bursts of timing violations may not be tolerable and also risk corrupting the control plane logic. To guard against this, we propose a fast adaptive clocking scheme specifically for error-tolerant accelerators, triggered by in-situ razor circuits.

Adaptive clocking schemes must have very low latency to react to nanosecond-scale V_{DD} droops. While voltage-triggered schemes often require complex calibration, in-situ Razor circuits are ideal for triggering adaptive clocking since they can be distributed to capture fast local variation and do not require calibration or significant margin. However, like other schemes [4], the output trigger signal (ERR) is asynchronous to F_{CLK}, and prone to metastability, typically mandating a synchronizer that increases latency, with the fastest reported adaptation

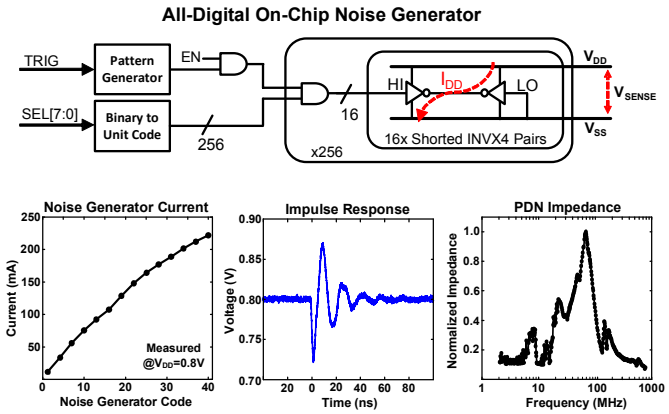


Fig. 6. All-digital on-chip noise generator circuits (top). Current vs noise generator code, PDN impedance and impulse response (bottom).

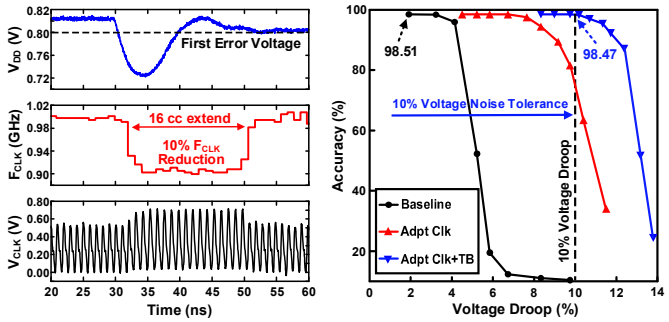


Fig. 7. Measured trace of VDD droop waveform demonstrating adaptive clocking operation (left). Adaptive clocking with time borrowing provides ~10% droop tolerance (right).

latency at 6ns [4]. We propose a metastability-safe adaptive clocking circuit to remove the latency overhead associated with synchronizers.

Our approach, shown in Fig. 5, comprises timing-error detection via double-sampling with time borrowing (DSTB), a digitally-controlled oscillator (DCO) with a metastability-safe control input, and a conventional synthesized (buffered) clock tree. The worst-case droop response latency is 2.8 clock cycles (cc). In the accelerator pipeline (Fig. 2), the SRAM load and SIMD datapath unit have DSTB latches at critical timing end-points. Timing error outputs from all DSTB latches are reduced to a single ERR control signal. Initial timing errors that trigger adaptation are masked with time-borrowing (TB) in the DSTB. The DCO is an all-digital, single-ended design with muxed delay stages for conventional coarse-grained frequency tuning, and achieves fast metastability-safe frequency reduction via binary varactor loads before the ring-oscillator output (simplified schematic in Fig. 5). The varactors use the two inputs of NAND2 gates: the input closest to the output in the NAND2 stack directly loads the delay path (C_{IN} in Fig. 5), while the other input modulates the effective capacitance of C_{IN}. Hence, an asynchronous event that results in a metastable signal on the NAND2 ERR input cannot cause glitches on F_{CLK}, and hence a synchronizer is not required. Multiple NAND2 stages, close to the DCO output, minimize charge injection and adaptation latency. The duration of the F_{CLK} reduction is programmable via an extend block.

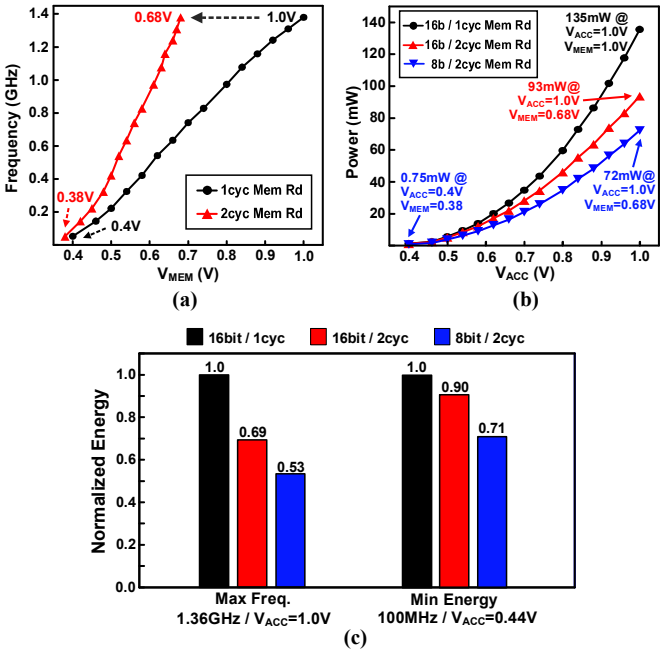


Fig. 8. Measurements of multi-cycle SRAM banking benefits: (a) reduction in SRAM operating voltage, (b) DNN accelerator power savings, and (c) energy savings for two operating points.

III. Measurement Results

The test chip, fabricated in a 16nm FinFET technology, achieves a wide operating range (~24×) in F_{MAX} (Fig. 8), from 1.36GHz (136mW, 1.0V, 16-bit) to operating at near-threshold, consuming 750µW (57MHz, 0.4V, 8-bit). In all cases, Razor error rate counters are used to automatically find the optimal V_{DD}/F_{CLK} settings at the first error point, without calibration.

An all-digital on-chip noise generator, shown in Fig. 6, consisting of an array of shorted inverter pairs induces resonant V_{DD} droops by driving one of the inverter inputs with an on-chip pattern generator while the other input is tied low, causing short-circuit current between V_{DD} and V_{SS}. This simple arrangement allows easy implementation of tunable noise generators using standard cells and avoids custom layout. The resonant frequency of the power delivery network (PDN) for the DNN accelerator and 1MB SRAM is ~68MHz, with the measured impulse response shown in Fig. 6. Measured scope traces shown in Fig. 7 confirm proper operation of the Razor adaptive clocking architecture. After inducing a V_{DD} droop using the on-chip noise generator, a timing violation occurs when V_{DD} falls below the "first error" voltage, which then triggers a 10% F_{CLK} reduction in the DCO within 3 clock cycles (cc). F_{CLK} reduction lasts for a configurable number of cycles via the extend block (16 cc in Fig. 7). Without adaptive clocking, accuracy plummets as voltage noise droops beyond 4%. Adaptive clocking with time borrowing tolerates up to 10% droop with no accuracy loss, which translates to 8.4% energy savings at nominal operating voltage of 0.8V by reclaiming the voltage guardband.

In addition, the proposed banked SRAM architecture allows for a two-cycle SRAM read latency, which relaxes read timing constraints and enables aggressive V_{MEM} scaling at iso-throughput to provide significant power savings, as shown in

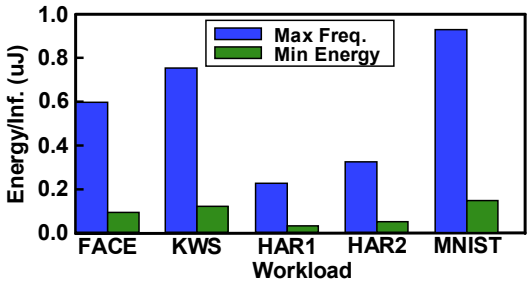


Fig. 9. Energy/Inference for IoT workloads at Maximum Frequency and Minimum Energy points.

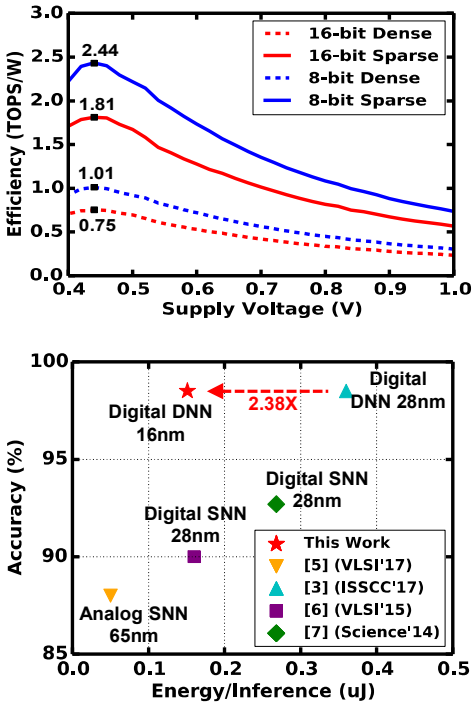
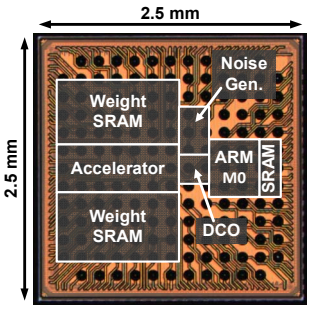


Fig. 10. Measured efficiency over voltage (top), and energy vs accuracy comparison for published MNIST results measured in silicon (bottom).

Figs. 8(a) and 8(b). For example, with V_{ACC} at 1V, operating at 1.36GHz, V_{MEM} can reduce from 1V to 0.68V, reducing total power by 31%. Fig. 8(c) shows the energy improvements for two operating points: maximum frequency and minimum energy. Multi-cycle read provides 31% and 10% energy improvement respectively at the two operating points. This scheme improves energy efficiency across the entire operating range when the minimum energy point (MEP) is limited by throughput. Five always-on applications were mapped onto the DNN processor: binary face matching (FACE), keyword spotting (KWS), two human activity recognition scenarios (HAR1/2), and handwritten digit classification (MNIST). Fig. 9 shows the measured energy/inference for these workloads at two operating points of maximum frequency and MEP, with all five operating well below 1 µJ/inference.

Fig. 10 reports raw energy efficiency with 16b data and no sparsity optimizations of 750 GOPS/W at the MEP, which includes all memory power. This is the highest reported raw GOPS/W to date at this precision, even amongst CNN accelerators with very high data reuse [1] not possible with FC



Test Chip Summary

Technology	16nm FinFET
On-chip Model Size	1 MB
Layer Support	FC-DNN
Voltage Range	0.4 – 1 V
Frequency Range	57 – 1360 MHz
Peak Eff. 16-bit	750 GOPS/W
MNIST Energy/Acc.	151 nJ / 98.51 %
Resilience Features	Razor + Adaptive Clock

Fig. 11. Annotated die photo and test chip summary

layers. Compared to other works without DRAM accesses [2–3], this is a ~2× improvement on the state-of-the-art. Switching to 8b operands with sparse data predication increases efficiency to 2.44 TOPS/W, yielding the lowest reported MNIST energy of 151nJ at 98.51% accuracy, 2.3X lower than [3] (Fig. 10). Fig. 11 shows the annotated die photo and test chip summary.

IV. Conclusion

This paper presents a sparse DNN processor to enable energy efficient always-on classification for IoT applications. Energy efficiency is optimized across a wide operating range using a multi-cycle banked SRAM architecture and a novel razor adaptive clocking scheme. Measurement results from a 16nm test chip demonstrate energy efficiency improvements across the wide operating range, with state-of-the-art minimum energy of 151nJ/inference for MNIST at 98.51% accuracy.

Acknowledgment

This work was supported in part by the U.S. Government, under DARPA CRAFT and DARPA PERFECT programs. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. We are grateful to ARM for providing physical IP.

References

[1] B. Moons *et al.*, "ENVISION: A 0.26-to10TOPS/W Subword-Parallel Dynamic-Voltage-Accuracy-Frequency-Scalable Convolutional Neural Network Processor in 28nm," *IEEE International Solid-State Circuits Conference (ISSCC)*, 2017.

[2] S. Bang *et al.*, "A 288uW Programmable Deep-Learning Processor with 270KB On-Chip Weight Storage Using Non-Uniform Memory Hierarchy for Mobile Intelligence," *IEEE International Solid-State Circuits Conference (ISSCC)*, 2017.

[3] P. N. Whatmough *et al.*, "A 28nm SoC with a 1.2GHz 568nJ/Prediction Sparse Deep-Neural-Network Engine with >0.1 Timing Error Rate Tolerance for IoT Applications," *IEEE International Solid-State Circuits Conference (ISSCC)*, 2017.

[4] M. Floyd *et al.*, "Adaptive Clocking in the POWER9 Processor for Voltage Droop Protection," *IEEE International Solid-State Circuits Conference (ISSCC)*, 2017.

[5] F. Buhler *et al.*, "A 3.43TOPS/W 48.9pJ/Pixel 50.1nJ/Classification 512 Analog Neuron Sparse Coding Neural Network with On-Chip Learning and Classification in 40nm CMOS," *IEEE Symp. on VLSI Circuits*, 2017.

[6] J. Kim *et al.*, "A 640M pixel/s 3.65mW Sparse Event-Driven Neuromorphic Object Recognition Processor with On-Chip Learning," *IEEE Symp. on VLSI Circuits*, 2015.

[7] P. Merolla *et al.*, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, 2016.