A 16nm 25mm^2 SoC with a 54.5x Flexibility-Efficiency Range from Dual-Core Arm Cortex-A53 to eFPGA and Cache-Coherent Accelerators

Paul N. Whatmough[1,2*], Sae Kyu Lee[1,3*], Marco Donato[1], Hsea-Ching Hsueh[1], Sam (Likun) Xi[1], Udit Gupta[1], Lillian Pentecost[1], Glenn G. Ko[1], David Brooks[1], Gu-Yeon Wei[1]
[1]Harvard University, Cambridge, MA, [2]Arm Research, Boston, MA, [3]IBM Research, Yorktown Heights, NY

Abstract

This paper presents a 25mm^2 SoC in 16nm FinFET technology targeting flexible acceleration of compute intensive kernels in DNN, DSP and security algorithms. The SoC includes an always-on sub-system, a dual-core Arm A53 CPU cluster, an embedded FPGA array, and a quad-core cache-coherent accelerator cluster. Measurement results demonstrate the following observations: 1) moving DSP/cryptography kernels from A53 to eFPGA increases energy efficiency between 5.5x–28.9x, 2) the use of cache coherency for datapath accelerators increases throughput by 2.94x, and 3) accelerator flexibility-efficiency (GOPS/W) range spans from 3.1x (A53+SIMD), to 16.5x (eFPGA), to 54.5x (CCA) compared to the dual-core CPU baseline on comparable tasks. The energy per inference on MobileNet-128 CNN shows a peak improvement of 47.6x.

Introduction

Low-power hardware acceleration of deep neural network (DNN) inference is a key enabling technology for a broad array of applications and use-cases in embedded Internet-of-Things (IoT) devices. Previous work on hardware for DNNs has focused on maximizing raw throughput and energy efficiency by emphasizing low-precision computation [1]. However, in a real SoC programmability and flexibility is vital to avoid fragile over-optimization as applications and algorithms (e.g. DNN architectures [2]) change rapidly over product life cycles. In addition to this, the programming model and memory system design are also essential to achieving high energy efficiency under a full software stack.

SoC Architecture

Fig. 1 shows the 16nm SoC with four main blocks that span the flexibility-efficiency spectrum: 1) always-on sub-system (AON), 2) dual-core Arm Cortex-A53 CPUs, 3) 2x2 embedded FPGA (eFPGA) array with hard-DSP 4) quad-core cache-coherent datapath accelerators (CCA). The SoC memory system includes a 4MB 4-way software-managed SRAM, and a wide off-chip interface to an FPGA board with DRAM and other peripherals.

The AON sub-system (Fig. 1) performs housekeeping and autonomous continuous sensing tasks (e.g. small DNNs at 151nJ/inf. [3]), while the remainder of the SoC is powered down. To perform more complex tasks, AON boots the A53 cluster. The A53 CPUs implement a rich 64-bit ISA with a dual-issue pipeline and wide 128-bit SIMD units, with a private 64KB L1 I/D-cache and a large 2MB shared L2 cache. The A53 cluster is connected to the rest of the SoC via a 128-bit interconnect, with an Accelerator Coherency Port (ACP) providing direct access into the large L2 cache.

Fig. 2 shows the FlexLogix eFPGA which sits in the middle ground between fully software programmable CPUs and specialized hardware accelerators [4]. The Flex Logix eFPGA is integrated as a first-class citizen on the SoC, which makes it amenable to a huge range of potential roles within the system, including data movement, compression, encryption/decryption, and custom datapath accelerators. The 2x2 array includes two logic tiles and two DSP tiles (Fig. 1). The logic tile (Fig. 2) includes 2.5K 6-input LUTs composed into logic compute elements (CE) and interconnected with a boundary-less radix interconnect [4]. The DSP tiles include 40x 22-bit DSP datapaths with less programmable logic (1.88K 6-input LUTs).

Finally, the CCA cluster (Fig. 3) provides the highest performance and energy efficiency, and consists of SRAM and datapaths dedicated to 2D convolution, dot-product and reduction operations common to DNNs and DSP algorithms. Datapath for 2D convolution is shown in Fig. 4. CCA is attached to the A53 L2 cache via the accelerator coherency port (ACP), which eradicates cache flushes when sharing data and significantly improves the software model [5]. With a low data migration cost, individual kernels can be accelerated on the CCA, composed by software on the CPU, while sharing data in the L2 cache.

Measurement Results

The 25mm^2 test chip was fabricated in a 16nm FinFET process (Fig. 8), and packaged in a 671-pin custom flip-chip substrate. Fig.5 shows a comparison between software (Dual-A53) and Verilog (eFPGA) implementation of different kernels. For the 2D convolution kernel we compare a systolic array implementation on eFPGA with an optimized (SIMD) GEMM implementation running on the A53 cores. Despite the highly optimized software implementation, energy efficiency and throughput increase by 5.5x and 27x respectively when implemented on eFPGA. The two FIR implementations (40-tap and 80-tap) show how the eFPGA utilization is critical to optimize the energy efficiency. For an 80-tap design, which maximizes DSP utilization (100%), the energy efficiency increases by 17.36x compared to software. For the 40-tap design, which use half the number of DSPs the energy efficiency improvement drops to 13.4x. The cryptography kernel is an AES128 ECB encryption/decryption block, which does not require DSPs and can be efficiently implemented in LUTs. The eFPGA implementation provides up to 28.9x and 120x improvement for energy efficiency and throughput.

The benefits of the ACP interface was described on FPGA in [5]. Results for ASIC are given in Fig. 6, which shows speedup of 2.7—3.1x arising from avoiding costly off-chip access and flushing to DRAM. The fastest results are achieved using software-managed on-chip SRAM, but this is a very expensive software model.

Fig. 7 shows a comparison of raw throughput and energy efficiency for the four different compute clusters. CCA achieves the highest energy efficiency at 1.04 TOPS/W. Better efficiency can be traded off with flexibility, as shown by the AON FC-only accelerator which peaks at 2.44 TOPS/W [3]. The energy per inference for MobileNet-128 is compared for three operating points, namely minimum energy (MEP), nominal (NOM) and max frequency (F_{MAX}). The energy/inf. on the entire model relative to the CPU baseline shows an improvement of 3.1x (SIMD), 22.7x (eFPGA), and 47.6x (CCA).

Acknowledgments

This work is supported in part by JUMP ADA, DARPA CRAFT and DSSoC programs, NSF Awards 1704834 and 1718160, and Intel Corp. We are grateful to ARM and Flex Logix for providing IP.

References

[1] Lee *et al.*, *ISSCC*, pp. 218-220, 2018.
[2] Howard *et al.*, *CVPR*, 2017.
[3] Lee *et al.*, *ESSCIRC*, pp. 158-161, 2018.
[4] Yuan *et al.*, *JSSC*, vol. 50, no. 1, pp. 137-149, 2015.
[5] Sadri *et al.*, *FPGAworld Conf.*, 2013.

* These authors have equal contributions

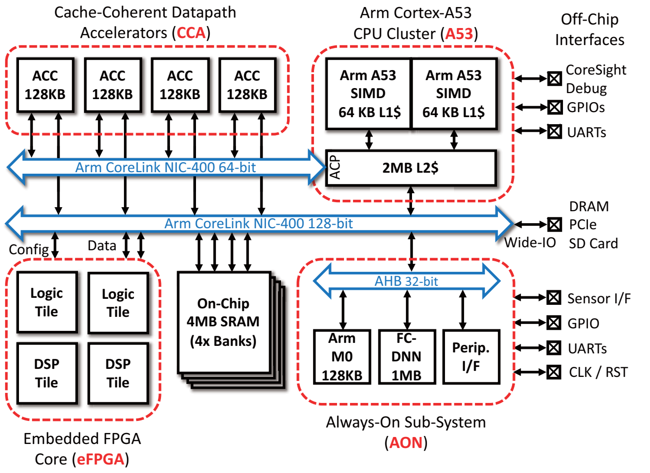


Fig. 1: Simplified block diagram of 16nm heterogeneous embedded DNN inference SoC.

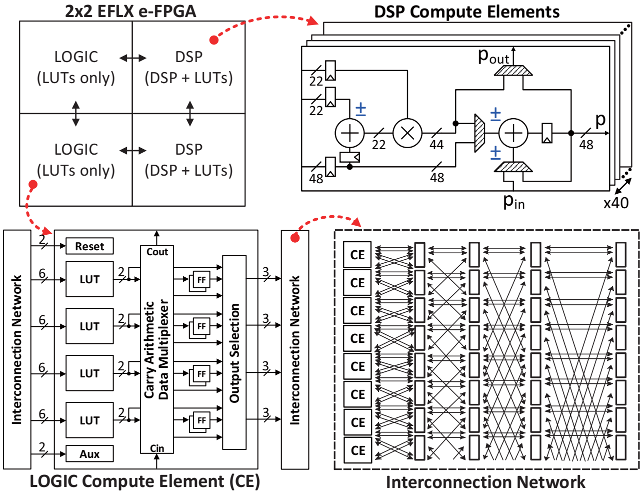


Fig. 2: Embedded FPGA (eFPGA) circuit architecture.

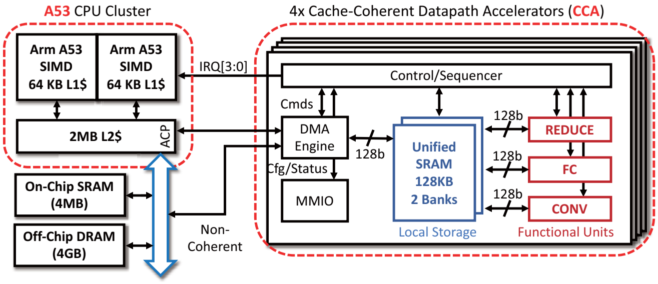


Fig. 3: Cache-coherent datapath accelerator (CCA) architecture and SoC integration with CPU cluster and memory system.

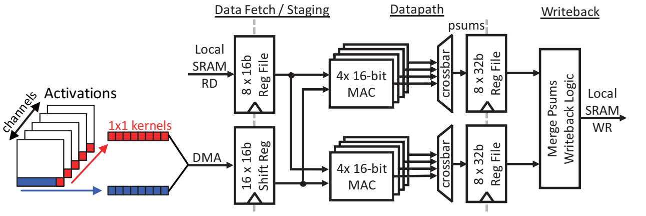


Fig. 4: Block diagram of 2D convolution datapath within the cache-coherent accelerator (CCA).

Kernel	eFPGA Utilization	Fmax [MHz]	Improvement [Dual-A53/eFPGA]
2D Conv	99.6% LUTs 100% DSPs	353.8	5.5x Energy 27x Throughput
FIR 40-tap	13.4% LUTs 50% DSPs	595.8	13.4x Energy 41.9x Throughput
FIR 80-tap	27.4% LUTs 100% DSPs	521.6	17.36x Energy 79.9x Throughput
AES128 ECB Enc.	37.2% LUTs	734	19.23x Energy 64x Throughput
AES128 ECB Dec.	37.2% LUTs	732.48	28.9x Energy 120x Throughput

Fig. 5: Software (Dual-A53) and Verilog (eFPGA) implementation of DSP and cryptography kernels. Improvements in terms of energy efficiency and throughput are shown at nominal operating voltage (0.8V).

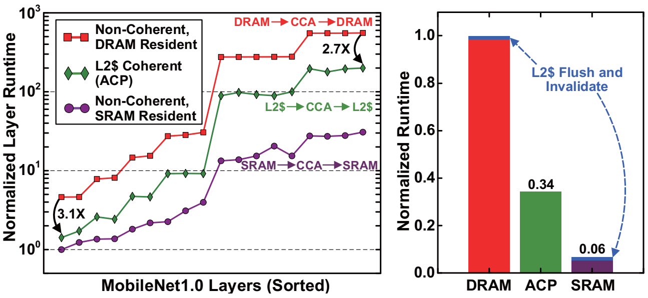


Fig. 6: Runtime benefits for ACP interface for individual MobileNet layers (left), and cumulative benefit for entire MobileNet inference (right).

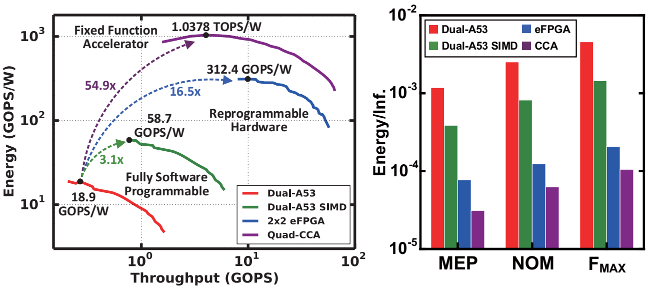
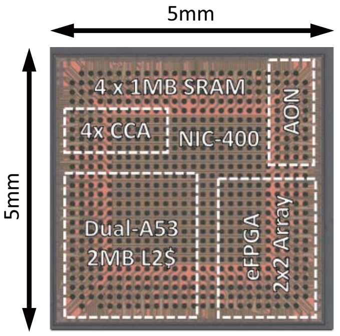


Fig. 7: Energy vs throughput (left) and energy/inference for three operating points (right) across the different compute clusters (Dual-A53, Dual-A53 SIMD, eFPGA, and Quad-CCA).



Technology	TSMC 16nm FFC
Size	25mm^2
Xtor Count	> 0.5 Billion
Total SRAM	72.2 Mbits
Packaging	Flip-chip BGA-672
Clock Domains	7
Power Domains	5
Supply Voltage	0.5 – 1.0 V
Fmax	>1 GHz

Fig. 8: Die photo of the fabricated 25 mm^2 test chip in 16nm FinFET.