SMIV: A 16nm SoC with Efficient and Flexible DNN Acceleration for Intelligent IoT Devices

Paul N. Whatmough (Arm Research / Harvard)

S.-K. Lee (IBM, Harvard), S. Xi, U. Gupta, L. Pentecost,

M. Donato, H.-C. Hseuh, D. Brooks and G.-Y. Wei (Harvard)

Deep learning enables intelligent IoT devices

Making sense of sensor data

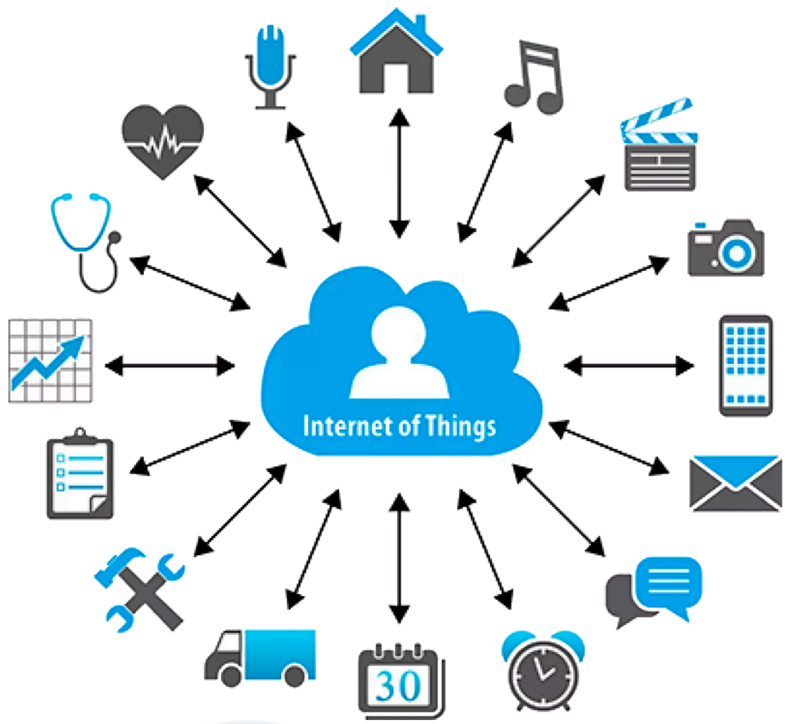
- Classification/regression problems in diverse domains
- Multi-modal data (sensor fusion)
- Unsupervised learning (e.g. anomaly detection)

Next-generation UI/UX

- Small form-factor without a big screen or input device
- Speech detection/recognition/synthesis, gaze detection, biometric authentication (e.g. face detection)
- Personalize interface and predict user decision-making

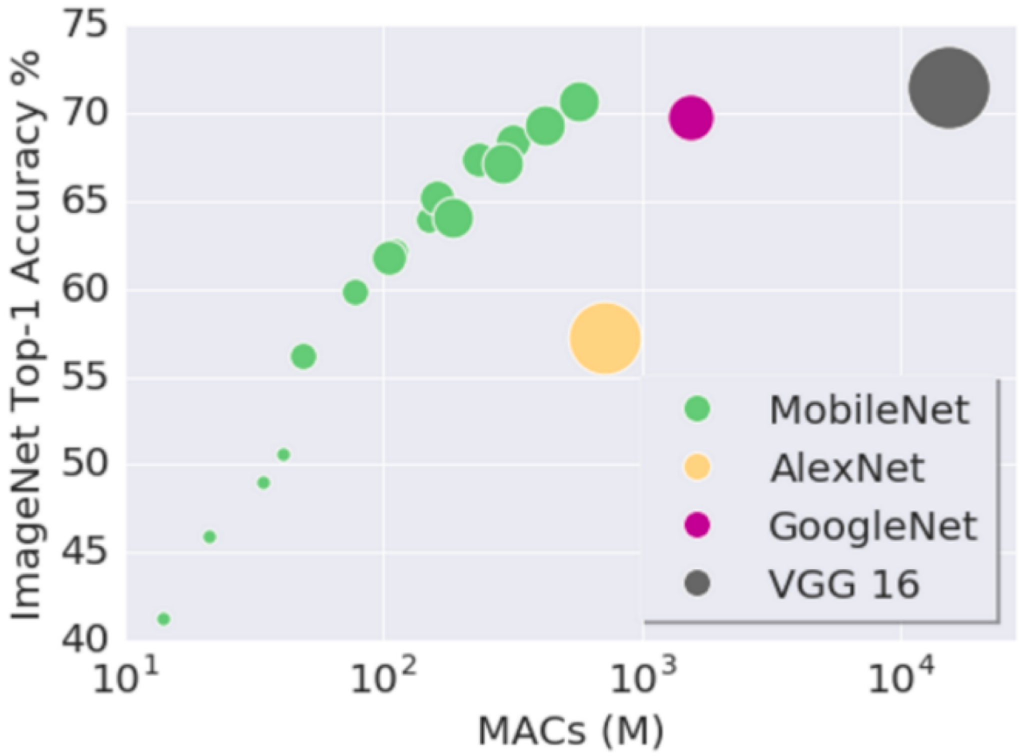
DNN inference on edge device

- Privacy, latency and energy issues with transmitting data
- Demands a large compute and storage capability…



Closing the DNN gap on embedded devices

More efficient networks



[Howard et al., CoRR 2017]

More efficient hardware

Reduced precision

4x improvement from INT8 versus FP32

Data re-use

Drives overall microarchitecture

Data compression

Reduce mem footprint, bandwidth, and power

Transforms

Winograd fast convolution (N^2 not N^3)

Sparse computation

Static weight pruning + dynamic ReLU activation

Efficient and flexible hardware acceleration

Architecture specialization

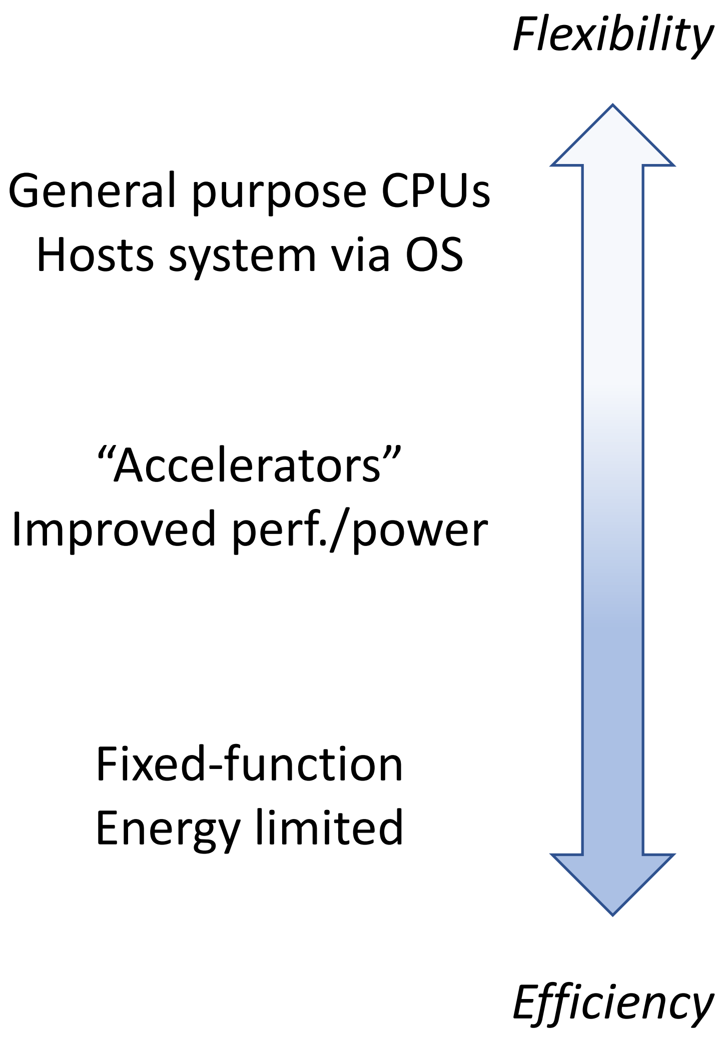
- Accelerators trade flexibility for efficiency
- Still care about silicon area in embedded
- DNNs pose risk of premature obsolescence

No silver bullet

- “Fluid” workloads suit more programmable accelerators
- Always-on sensing and monitoring – energy limited
- Reconfigurable architectures – FPGA and CGRA

It’s the system, stupid!

- CPU interfacing and memory systems for accelerators
- Abstractions and tools to map workloads to rich SoC



SMIV: motivation and chip overview

SoC platform for architecture and systems research

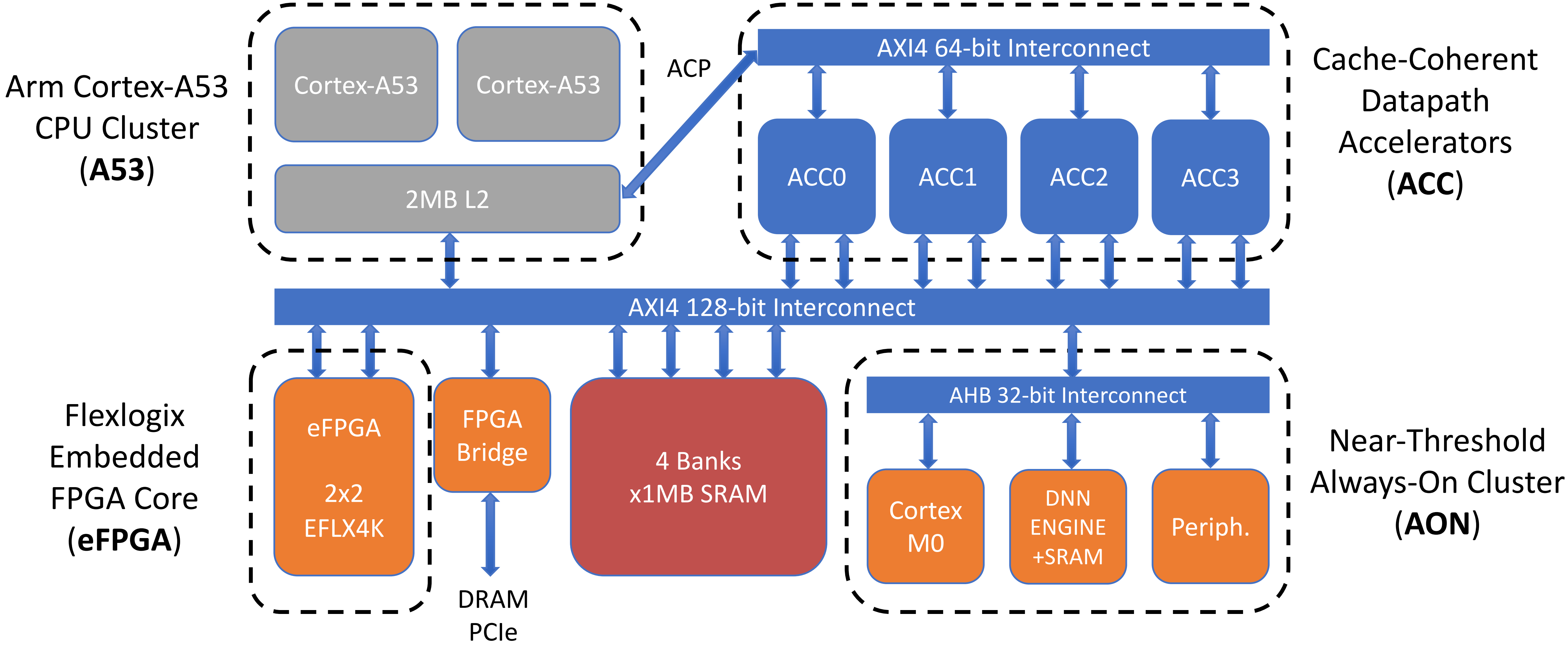
Test chip details

- $25mm^2$ die area (5mm x 5mm), TSMC 16nm FFC
- Half a billion transistors, 72.2 Mbit of SRAM, 7 clocks, 5 power domains
- First academic chip to feature Arm Cortex-A class CPUs
- All-digital GHz+ on-chip clock generation and chip-chip link
- Custom 672-pin flip-chip BGA package

Very short design, validation and implementation cycle

- 7 people (4 PhDs and 3 post-docs) in about 9 months (final IP came in 6 weeks before tape out)

SMIV: SoC platform



Near-Threshold Always-On Cluster (AON)

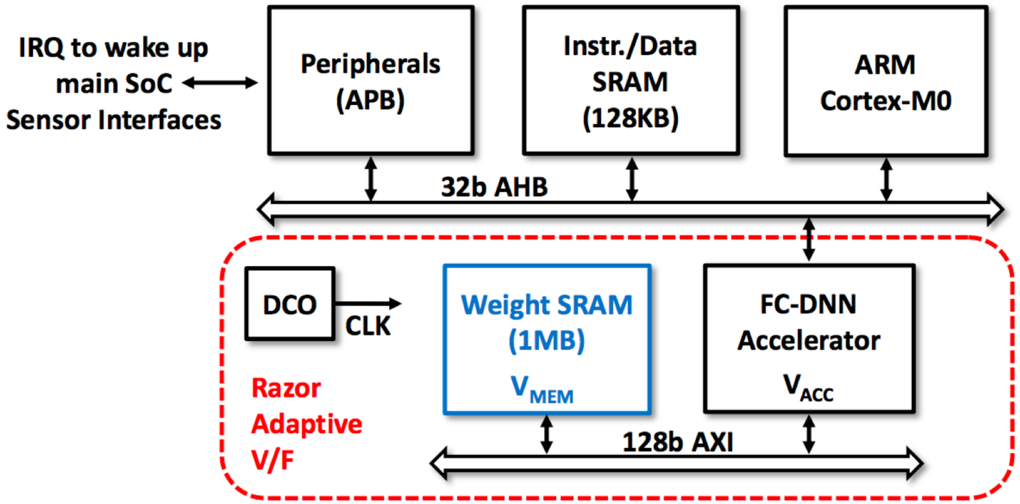
Cortex-M0, peripherals and accelerators

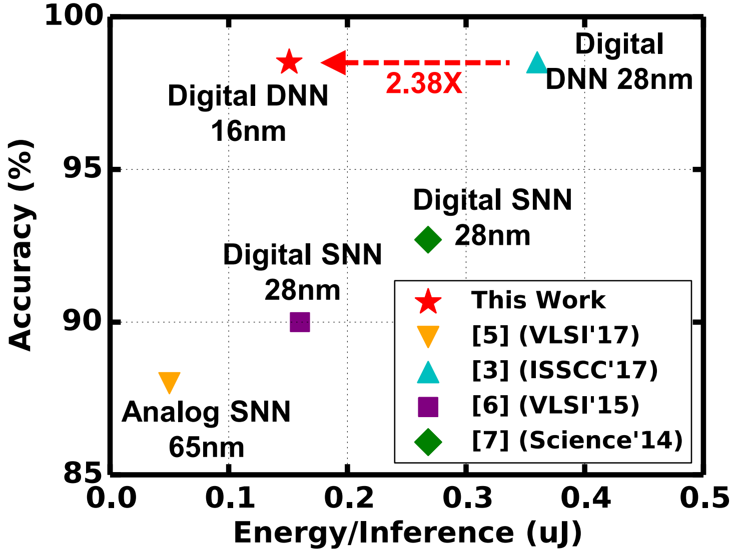
- Runs firmware for system control
- Performs low-energy always-on tasks
- Optimized for robust low voltage operation

DNN ENGINE programmable classifier

- Second generation design (ISSCC'17)
- Parallelism, data-reuse, optimized data-types, sparse computation, algorithmic resilience
- Model stored in on-chip SRAM
- Energy as low as 150nJ/inference for MNIST @98.5%

[Lee et al., ESSCIRC'18]





Arm Cortex-A53 CPU cluster

High efficiency embedded processor

Mature product with high volume

In-order pipeline

- Lower power consumption

Extensive dual-issue capability

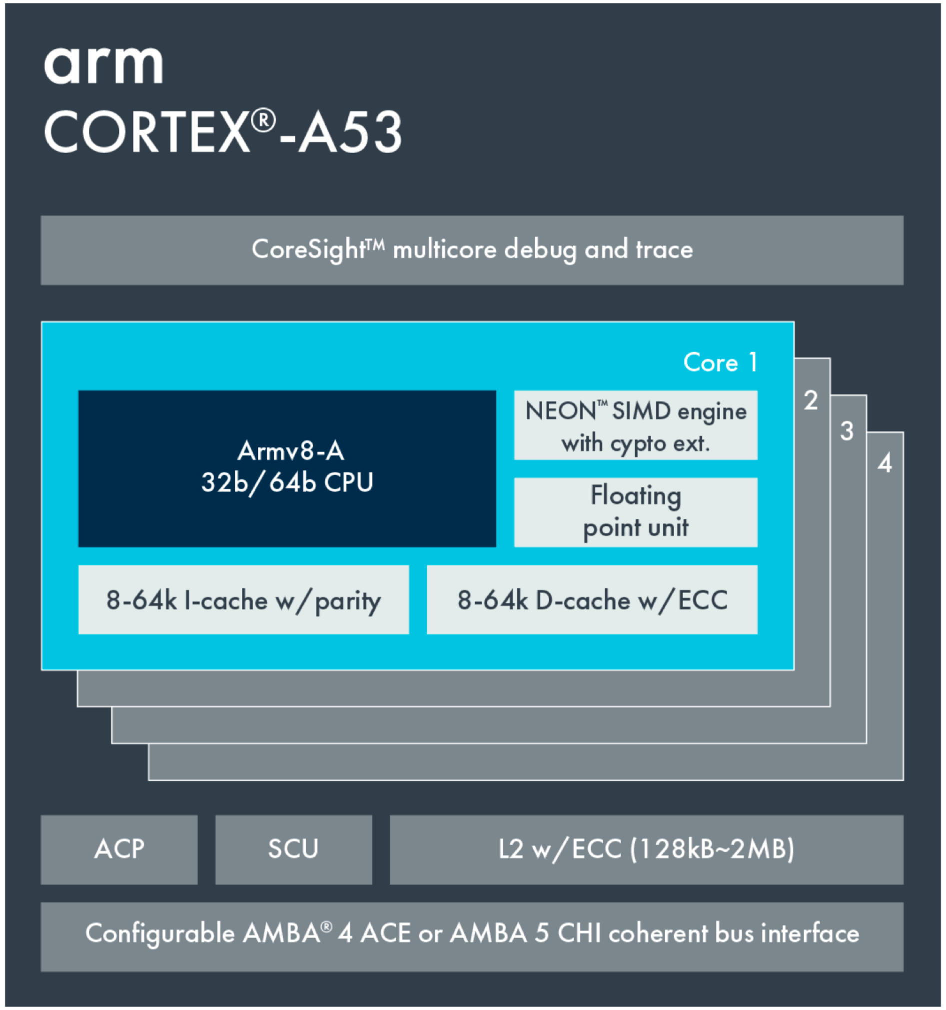
- Increased peak instruction throughput via dual instruction decode and execution

Advanced branch predictor

- Increased branch hit rate with 6Kb Conditional Predictor and 256 entry indirect predictor

Extensive power-saving features

- Hierarchical clock gating, power domains, advanced retention modes



Accelerator coherency port (ACP)

Efficient accelerator interfacing

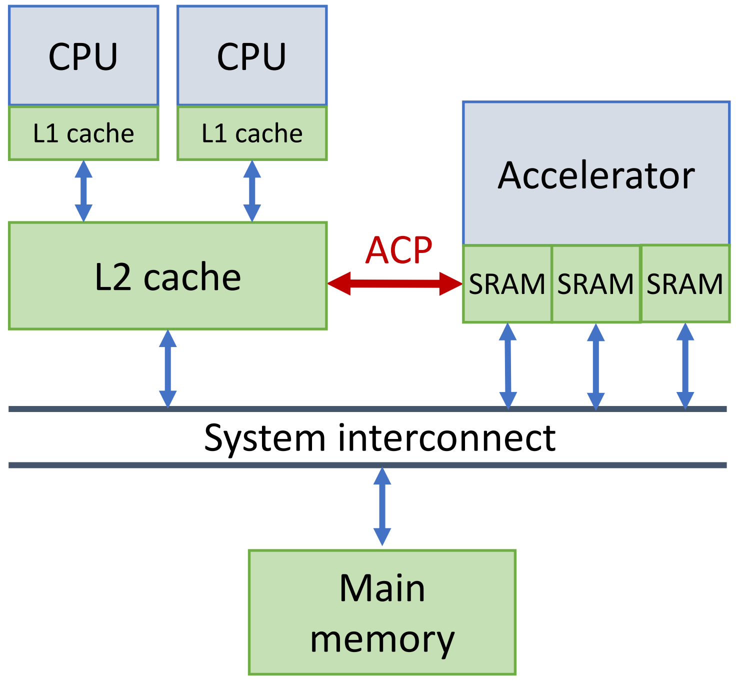
- Avoids CPU cache flush when accessing cached data
- Coherent memory simplifies programming model
- Very low hardware cost for accelerator

Enables fine grained datapath acceleration

- Focus on accelerating key composable kernels
- Increases flexibility…

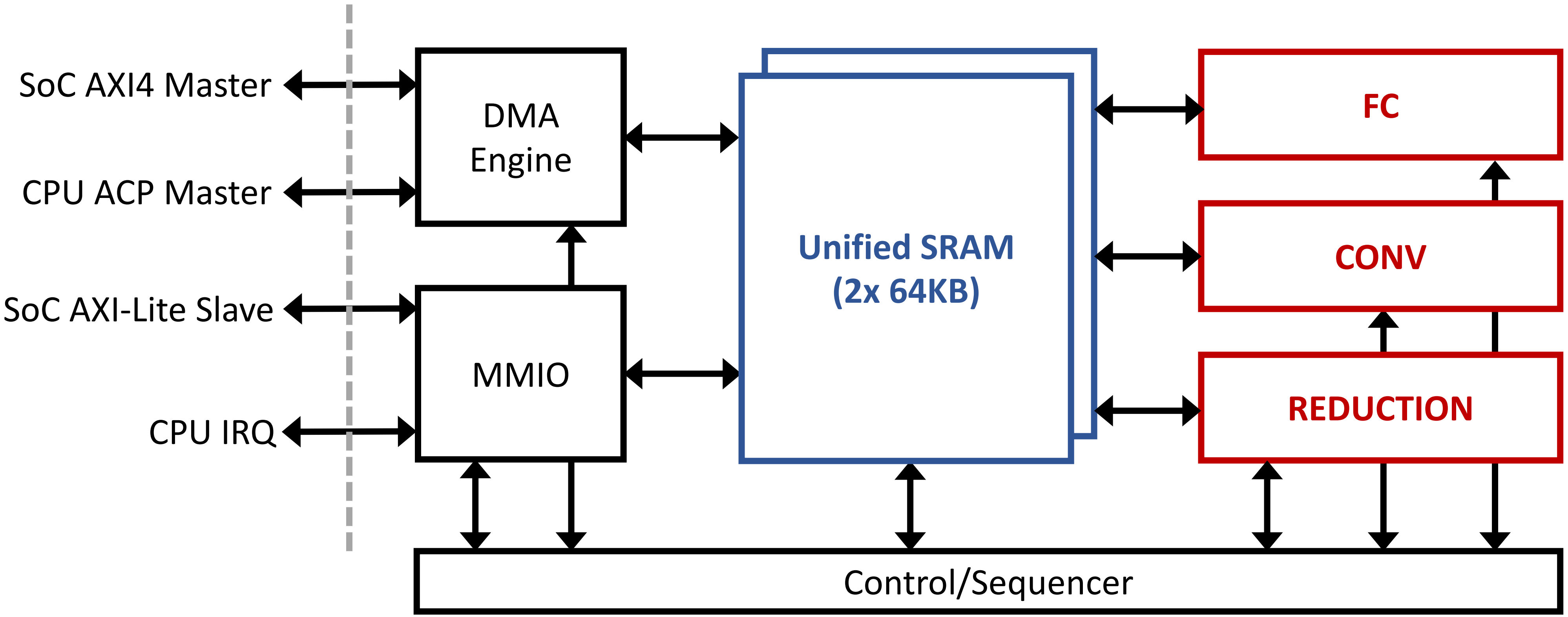
Cache a larger working set in L2

- Workload dependent - your mileage may vary
- Exploit lower energy of uncached loads (e.g. FC weights)



Cache-Coherent Datapath Accelerators (ACC)

Modeled and implemented using high-level synthesis (HLS) methodology



Flexlogix embedded FPGA (eFPGA)

Embedded FPGA IP

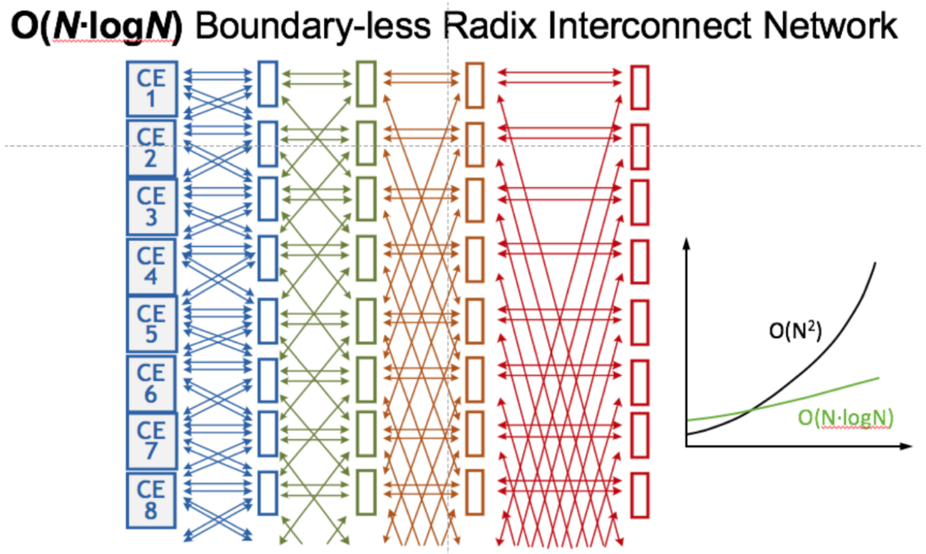
- Efficient interconnect enables density and scalability
- Density & performance similar to full custom FPGA

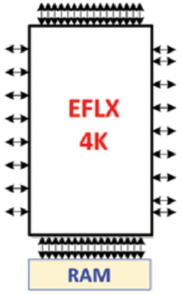
Scalable and flexible array size and layout

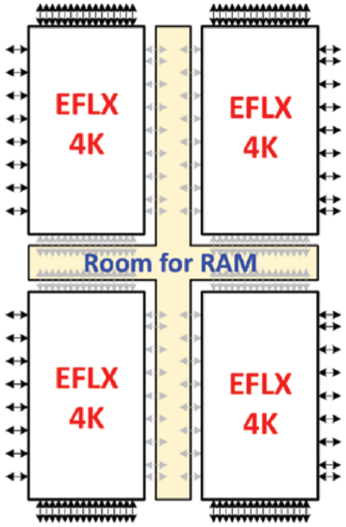
- Logic tile: 2,520 LUT6 + 21 kbits distributed mem
- DSP tile: 40 22b hardware MACs + 1,880 LUT6

SMIV contains a 2x2 eFPGA array

- 2x logic tiles and 2x DSP tiles
- Total ~9,000 LUT6 logic, 80 hardware MACs, 44kbits RAM
- Attached as a first-class citizen on the SoC interconnect







flexlogix
EFLX embedded FPGA ®

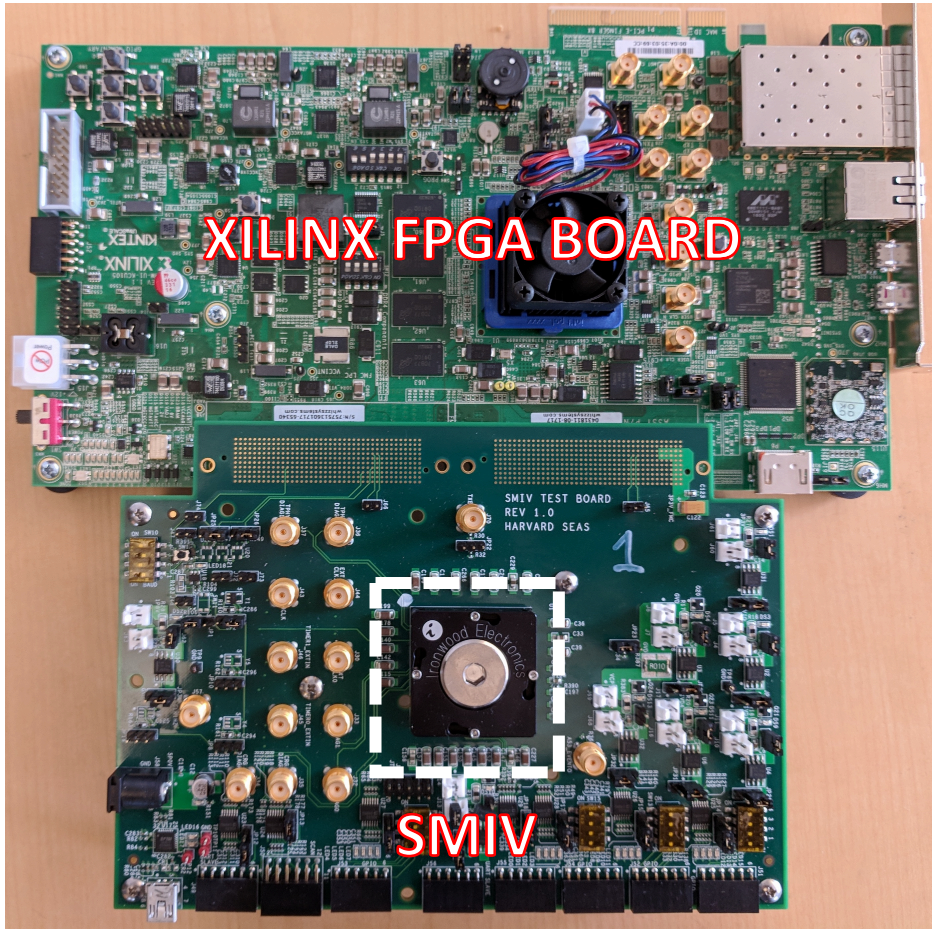
SMIV in action

Initial silicon bring up successful

- All pre-silicon tests passed
- Off-chip interface to FPGA board working, providing DRAM main memory and peripherals
- eFPGA programming fully-functional

Using SMIV SoC platform for research

- Scheduling accelerators sharing L2 over ACP
- eFPGA as first class citizen on an SoC
- Incremental wakeup from AON sub-system
- Real applications!



SMIV measured efficiency

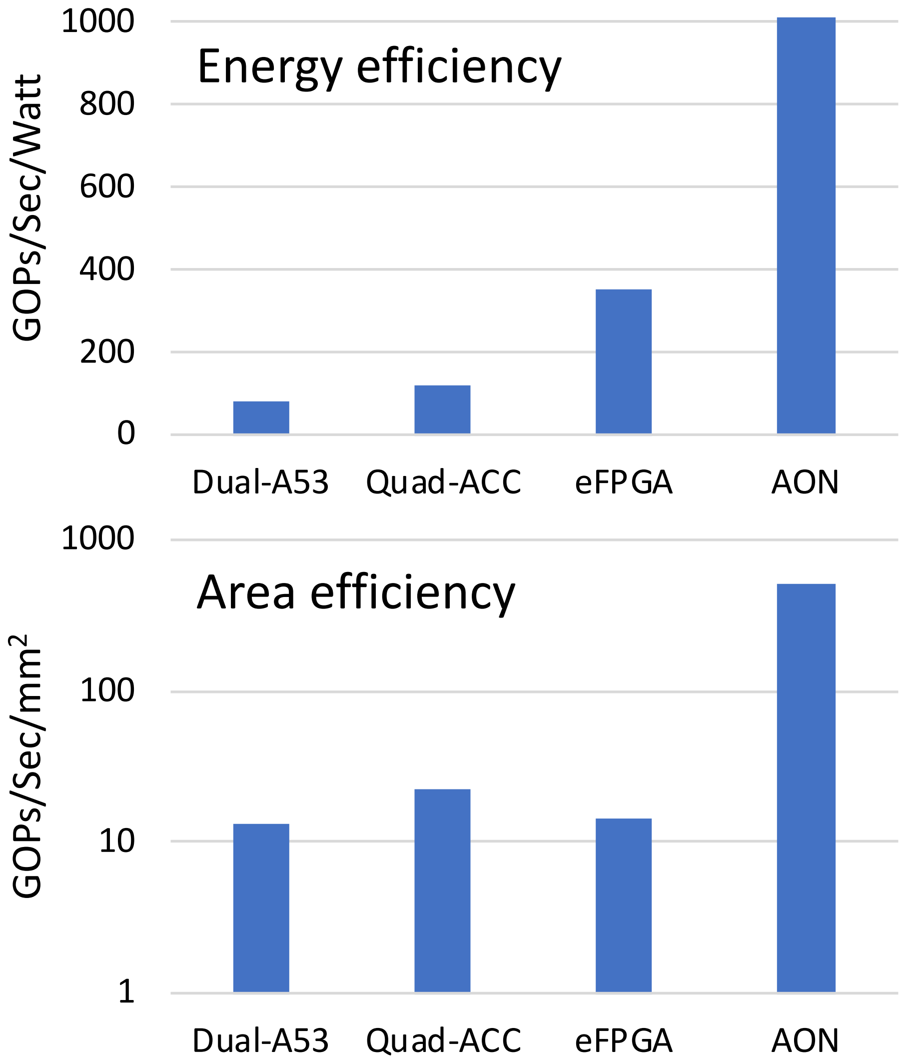
Measured on representative DNN kernels

Energy efficiency range is >10x

- CPU to fixed-function AON
- This also spans the whole flexibility spectrum
- AON will not benefit from new algorithm advances
- eFPGA is 4.5x energy efficiency of CPUs

Area efficiency heavily impacted by SRAM

- Important to share on-chip SRAM resources efficiently
- ACP allows accelerators to use large L2 cache
- eFPGA has the area overhead of reconfiguration



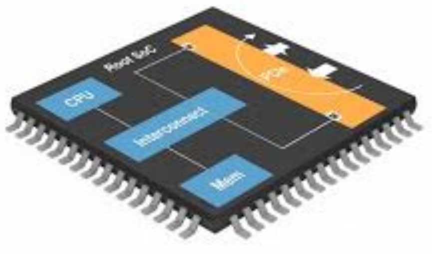
Rapid SoC design and implementation

DARPA program - Circuit Realization at Faster Timescales (CRAFT)

- 7 people (4 PhDs and 3 post-docs) in about 9 months (final IP came in 6 weeks before tape out)

How did we make a complex SoC in a short timescale?

- Industry strength IP, interface standards and software eco-system
- Arm Socrates tool for rapid iteration on SoC integration and IP configuration
- Scripted methodologies for generating memory-mapped registers, IO pad-ring, clock domains etc.
- No custom layout - entirely std-cells + SRAM, including clock generation and off-chip link PHY
- High-level synthesis (HLS) from a SystemC design, using Nvidia methodology



Minimize the long tail of validation and timing closure

Arm research enablement offerings

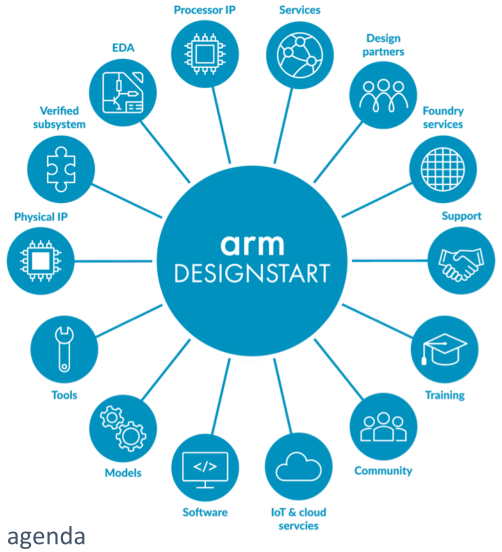
SoC HW/SW co-development with DesignStart

DesignStart Eval - Cortex M0/ M3 based systems, evaluation with obfuscated RTL

DesignStart Pro Academic - Cortex M0/ M3 based systems, RTL for SoC design

Compute systems modelling and architecture exploration

Gem5 - CPU and system modelling



IP Building blocks*

Design IP – CPUs, Interconnects, peripherals

Physical IP – Standard cells, Memory compilers, POP IP

www.arm.com/resources/research/enablement

*Any logic Arm IP that is not part of DesignStart will be provided on a case by case basis, depending on the research project scope, objectives and alignment with Arm research agenda

Summary

Deep learning on edge devices is driving new IoT use cases

Efficient and flexible DNN acceleration

It’s the system, stupid!

SMIV - a 16nm SoC platform for architecture and systems research

First academic chip to feature Arm Cortex-A class CPUs

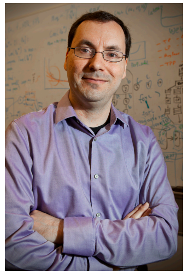
Near-threshold always-on cluster

Cache-coherent multi-core datapath accelerators with ACP attach

Embedded FPGA cluster

Rapid SoC design and implementation

Acknowledgments

We are very grateful to our sponsors, including the DARPA PERFECT and CRAFT programs, and to Arm, Flex Logix and TSMC for IP support