Harvard SoC Designs

Paul Whatmough

Arm ML Research Lab

paul.whatmough@arm.com

Harvard SoC Designs

…how to win friends and influence people with successful Arm-based SoC tape outs

Paul Whatmough

Arm ML Research Lab

paul.whatmough@arm.com

Why build test chips in academia?

If I just do simulation I can write twice as many papers…

Circuits research

- Measured test chip essential for tier-1 publication
- Often need to characterize devices and passives

Architecture research

- Understand the whole stack, soup to nuts
- Know you are solving a real problem
- Add real impact to your work; not just another academic paper

All models are wrong, some are useful

- Many things are hard to simulate convincingly
- Data to build models to design better circuits

Training for researchers

- Build a deep understanding of real computers
- Problem solving, team work, time management
- Extremely valuable depth of experience

Harvard / Arm collaboration on ML hardware

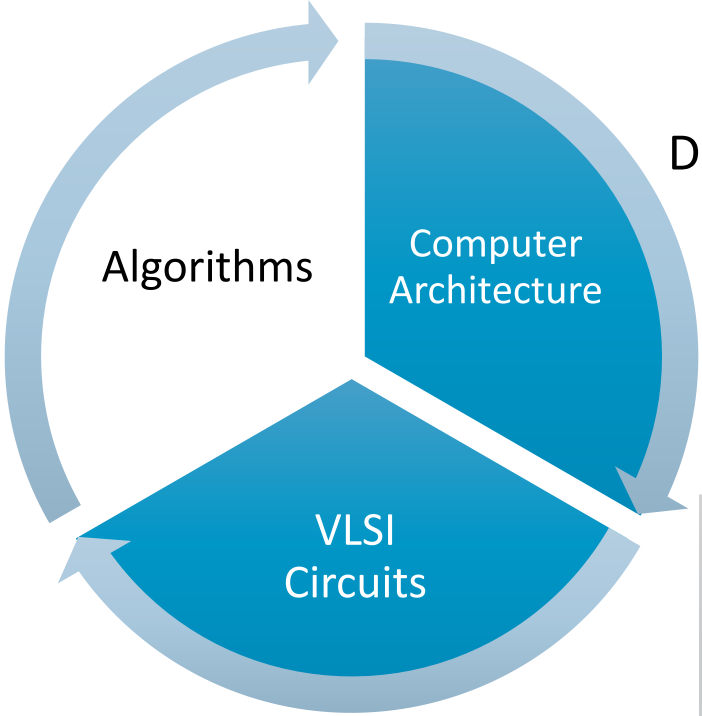
Harvard VLSI-Architecture group

- vlsiarch.eecs.harvard.edu

Harvard NLP group

- nlp.seas.harvard.edu

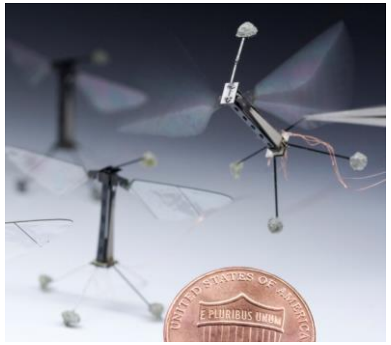
Sasha Rush
Harvard NLP



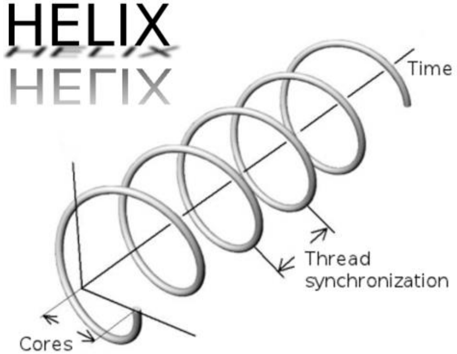
David Brooks

Paul Whatmough
(Arm)

ALADDIN



RoboBees



Gu-Yeon Wei

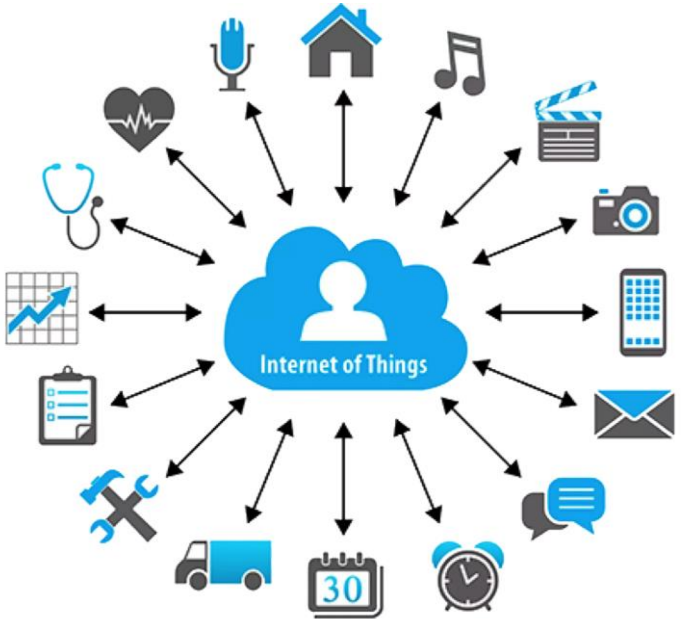
Deep learning hardware for IoT

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

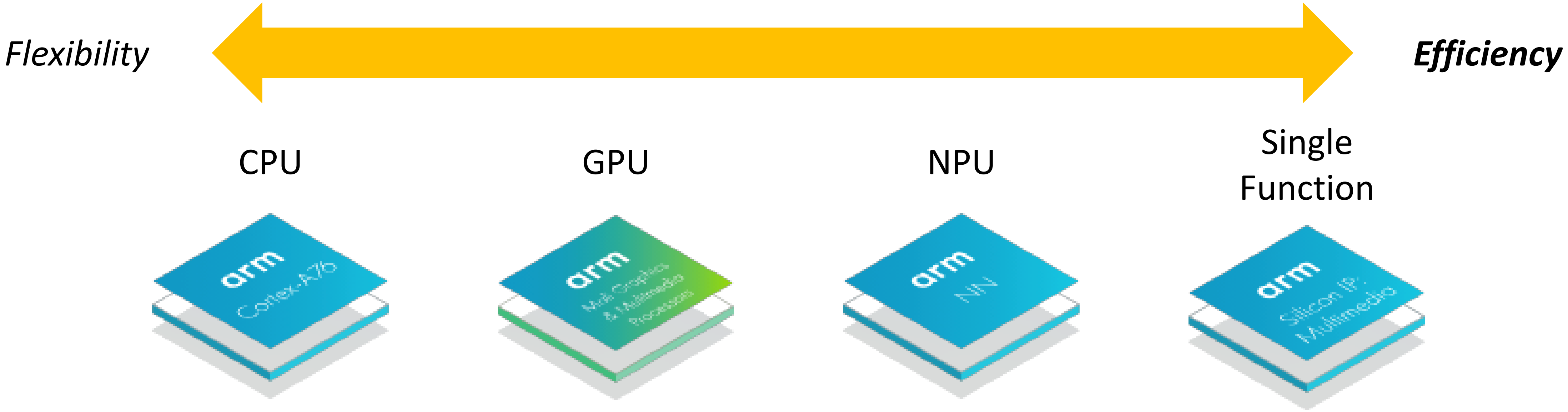
Efficient DNN inference hardware on embedded devices

- Privacy, latency and energy issues with transmitting data
- Efficiency to handle large compute and storage demand
- But, still care about programmer efficiency and flexibility



arm

Hardware architecture specialization



Flexibility

Efficiency

CPU

GPU

NPU

Single Function

What is the gap in energy efficiency and flexibility?

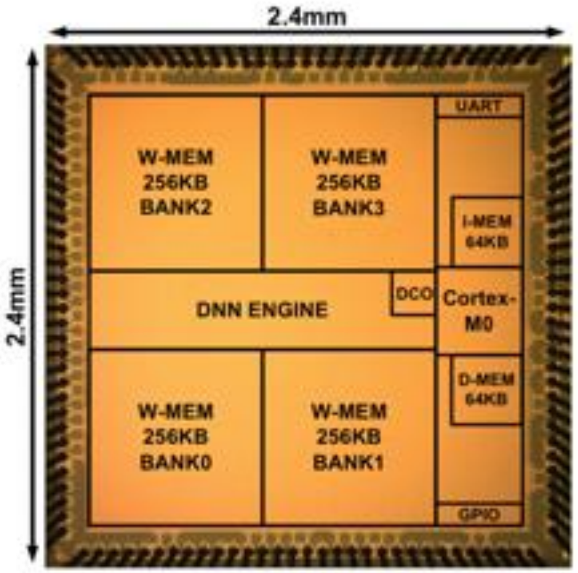
- Measured results comparing energy efficient across compute sub-systems

How can we allow hardware accelerators and CPUs to share data efficiently?

- Demonstrate lightweight L2$ coherent data sharing via accelerator coherency port (ACP)

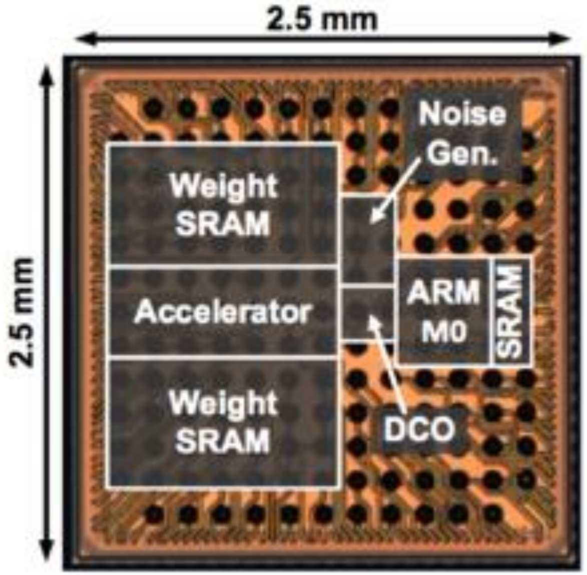
Recent tape outs at Harvard

TSMC 28HPC



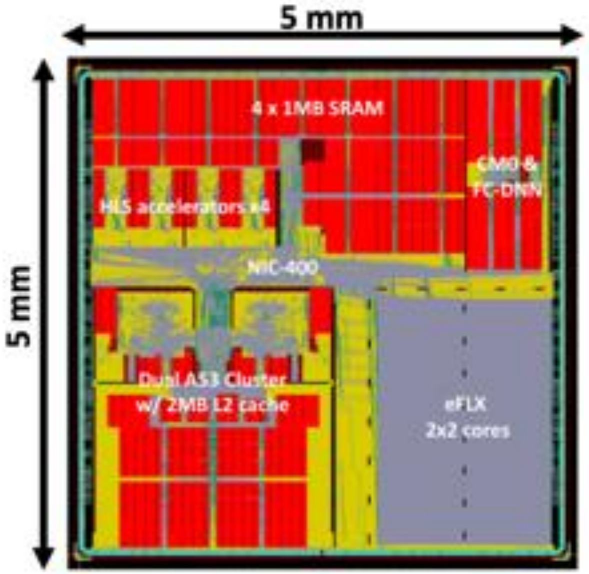
Tapeout: May 2016

TSMC 16FF+



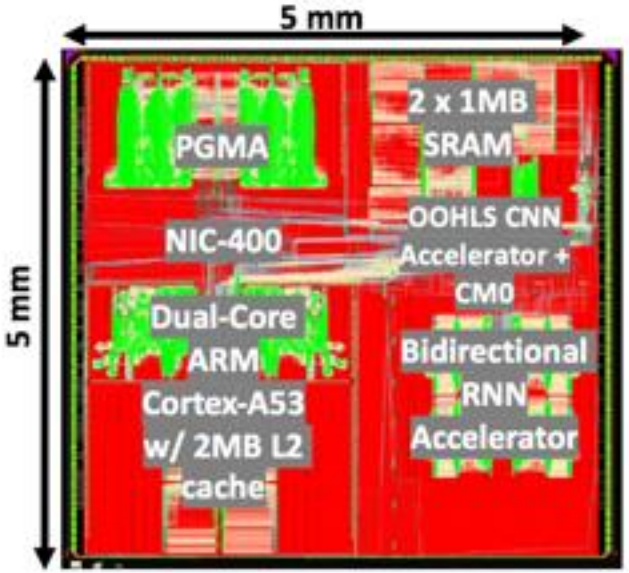
Tapeout: July 2016

TSMC 16FFC



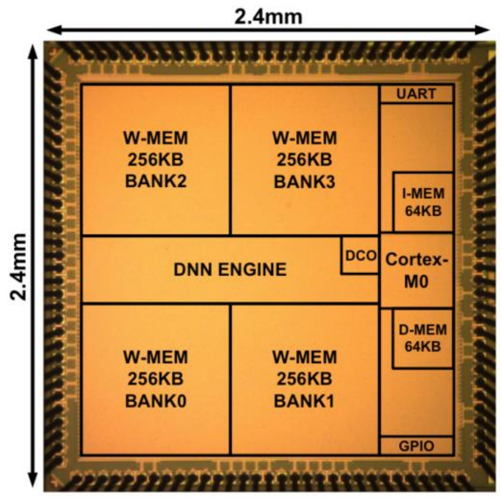
Tapeout: May 2017

TSMC 16FFC


5 mm
5 mm
PGMA
2 x 1MB SRAM
NIC-400
OOHLS CNN Accelerator + CM0
Dual-Core ARM Cortex-A53 w/ 2MB L2 cache
Bidirectional RNN Accelerator

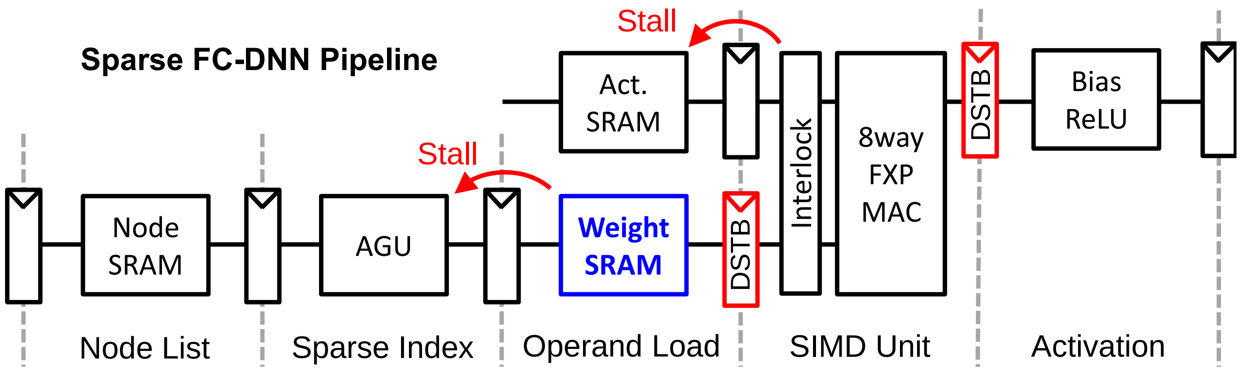
Tapeout: May 2018

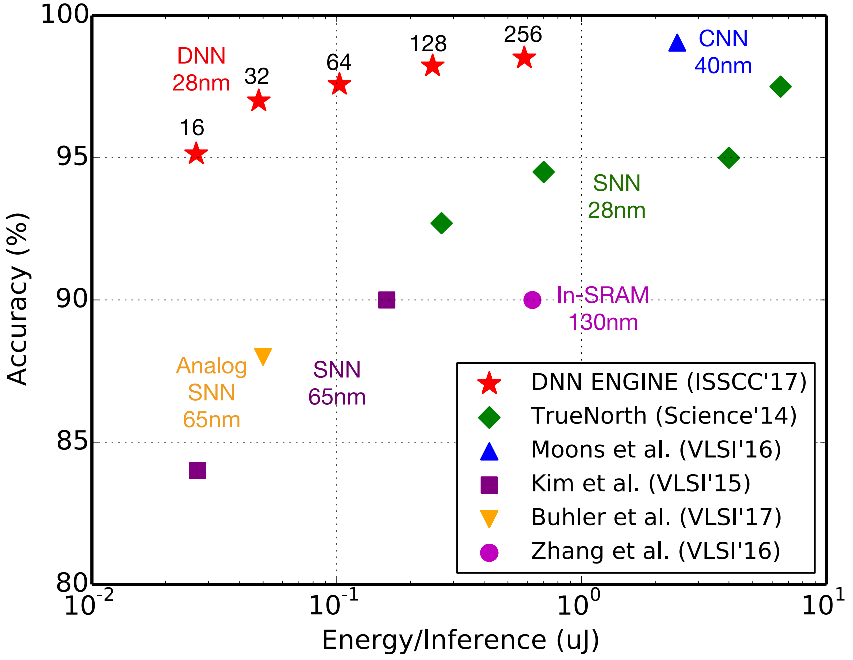
SM2 – 28nm DNN ENGINE



Programmable DNN Classifier for IoT

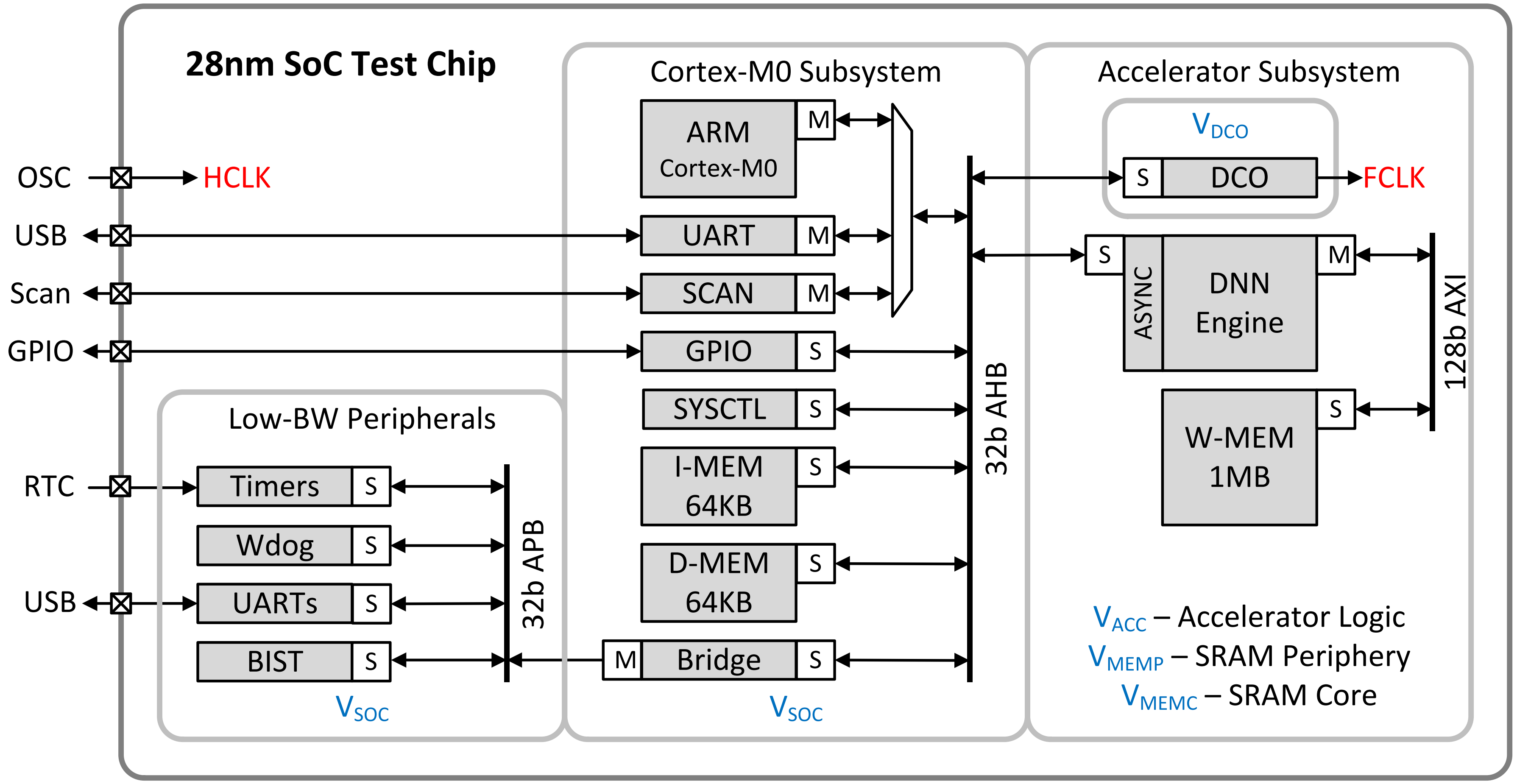
- Parallelism/reuse 8-way SIMD, 10X data reuse @ 128b/cycle BW
- Small data-types 8-bit weights, –30% energy
- Sparse activation data +4X throughput and –4X energy
- Algorithmic resilience +50% throughput or –30% energy





[Whatmough et al., ISSCC ’17, JSSC ’18]

SM2 – System architecture

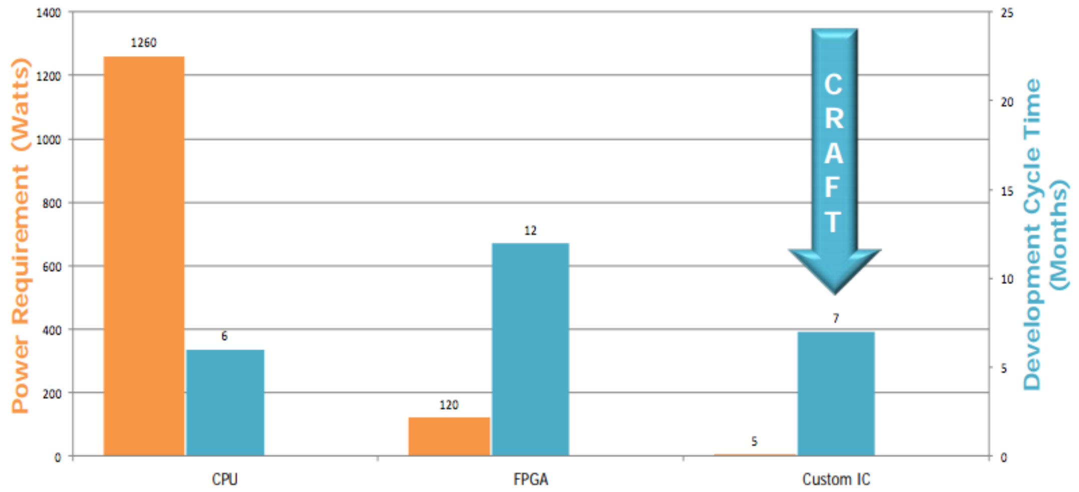


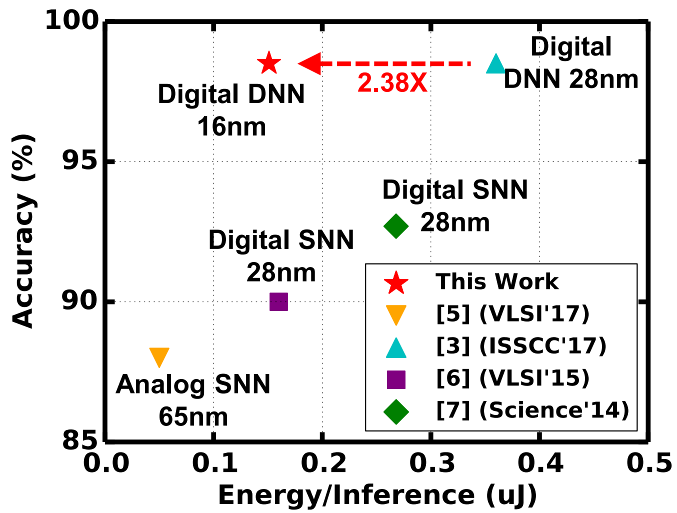
[Whatmough et al., ISSCC ’17, JSSC ’18]

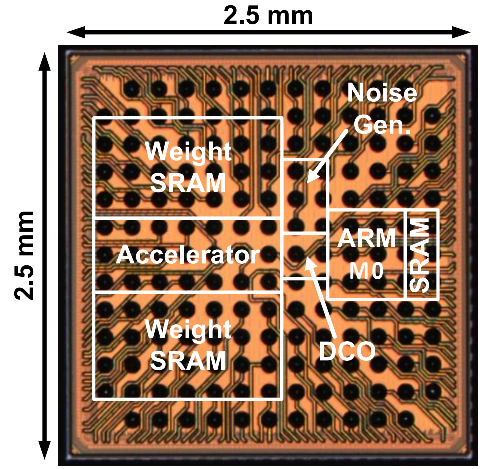
SM3 – 16nm DNN ENGINE v2

CRAFT Vision

To sharply reduce the barriers to DoD use of custom integrated circuits built using leading-edge CMOS technology while maintaining the high level of performance at power promised by this technology.







[Whatmough et al., Hot chips ‘17, Lee et al. ESSCIRC ’18, JSSC ‘19]

SM3 – IoT application demos

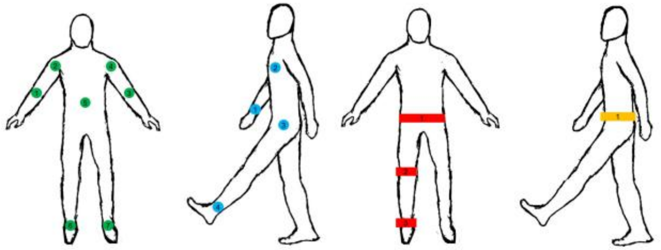
Labeled Faces in the Wild (FACE)

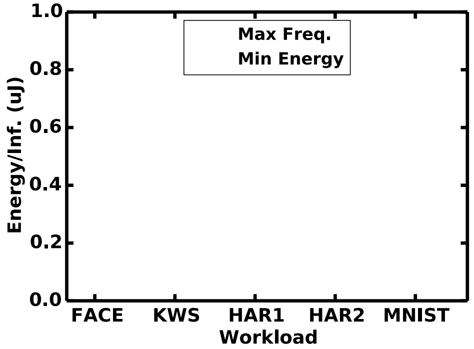
Keyword Spotting (KWS)

Human Activity Recognition (HAR)

- Opportunity (HAR1): Detect 18 gestures w/ 7 sensors
- Smartphone (HAR2): Detect basic activities w/ smartphone inertial measurement unit (IMU)
- PAMAP2: Detect activities of daily living w/ 4 sensors
- Daphnet Freezing of Gait: Detect freezing incidents for Parkinson’s patients

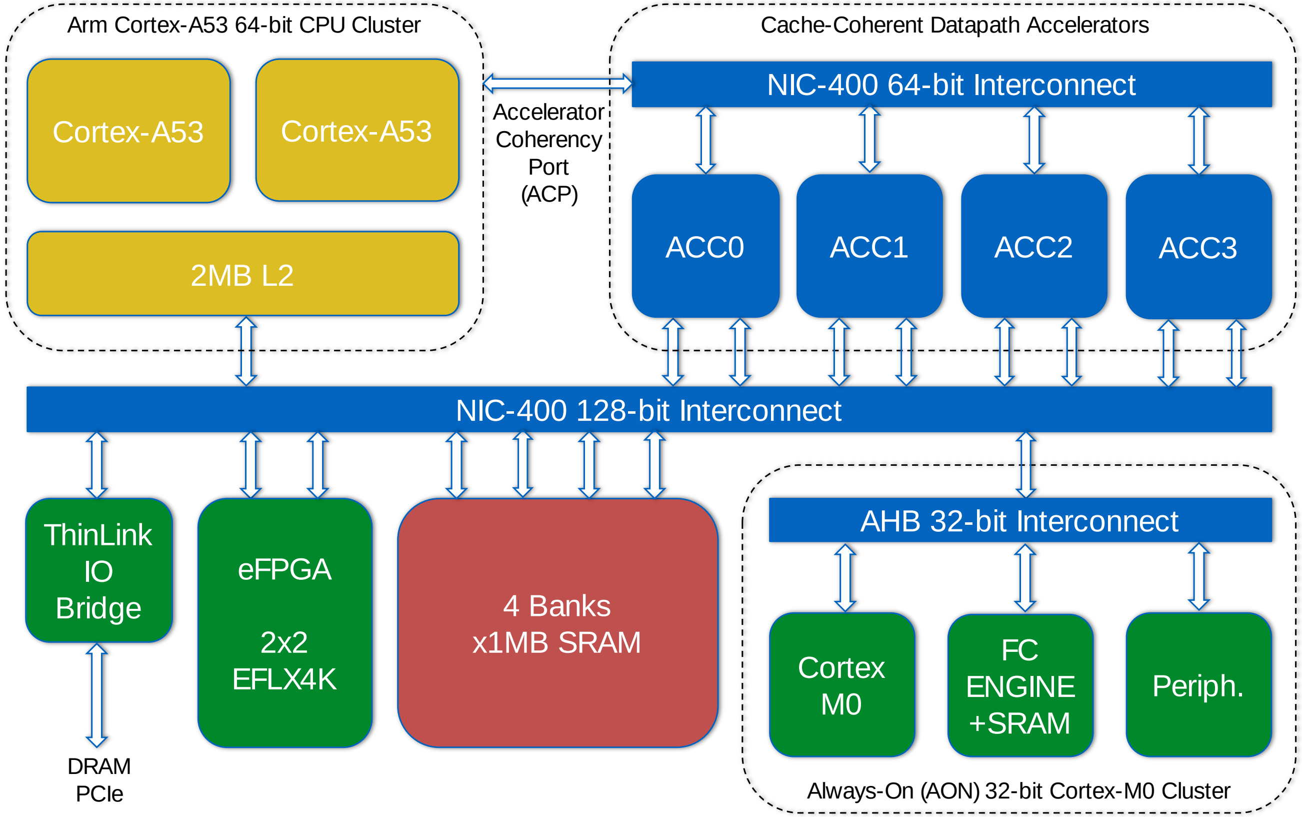
Hand written digit classification (MNIST)


(a) OPP (b) PAMAP2 (c) DG (d) Smartphone



[Kodali et al., ICCD’17]

SMIV – 16nm SoC to evaluate ML hardware



Arm Cortex-A53 64-bit CPU Cluster

Cortex-A53

Cortex-A53

2MB L2

Accelerator Coherency Port (ACP)

Cache-Coherent Datapath Accelerators

NIC-400 64-bit Interconnect

ACC0

ACC1

ACC2

ACC3

NIC-400 128-bit Interconnect

ThinLink IO Bridge

DRAM PCIe

eFPGA

2x2 EFLX4K

4 Banks x1MB SRAM

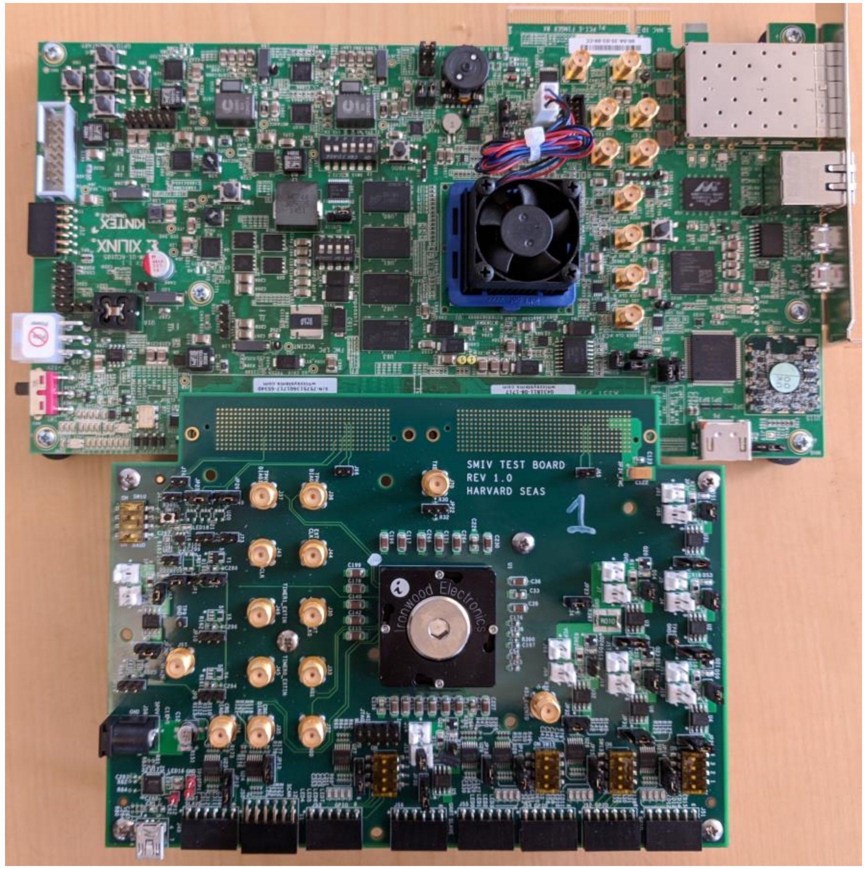
AHB 32-bit Interconnect

Cortex M0

FC ENGINE +SRAM

Periph.

Always-On (AON) 32-bit Cortex-M0 Cluster



[Whatmough et al., Hotchips ’18, VLSI ’19]

SMIV – GEMM performance across accelerators

Dual-Core CPU (A53)

- C/C++ or BLAS library
- 18.9 GOPS/W (1x)

Dual-Core CPU with SIMD enabled (A35)

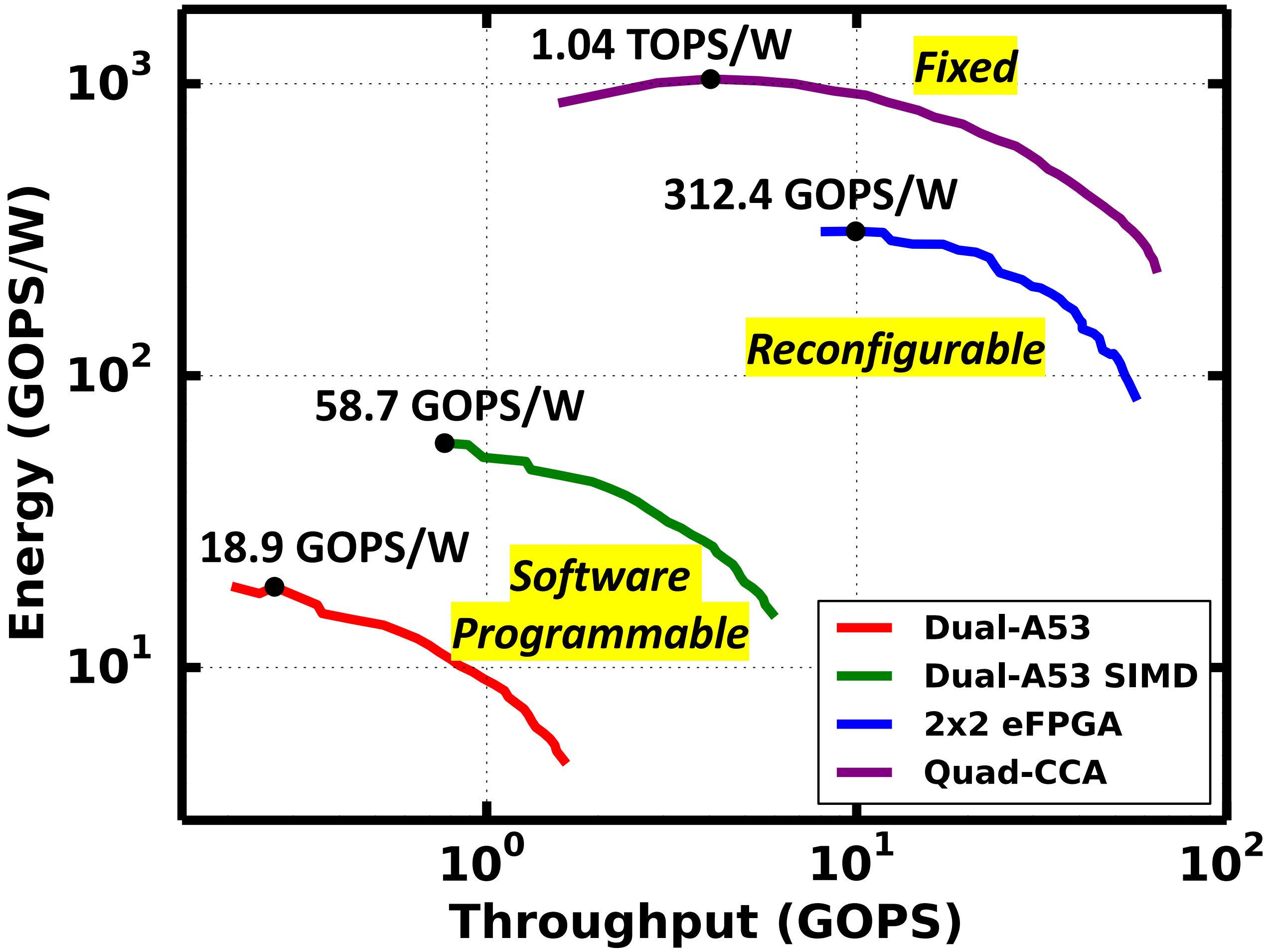
- Assembly or BLAS library
- 58.7 GOPS/W (3.1x)

4-Tile Embedded FPGA (eFPGA)

- Verilog/VHDL/HLS, reconfigurable
- 312 GOPS/W (16.5x)

Quad-Core Datapath Accelerator (CCA)

- Verilog/VHDL/HLS, fixed functionality
- >1 TOPS/W (54.9x)



[Whatmough et al., Hotchips ’18, VLSI ’19]

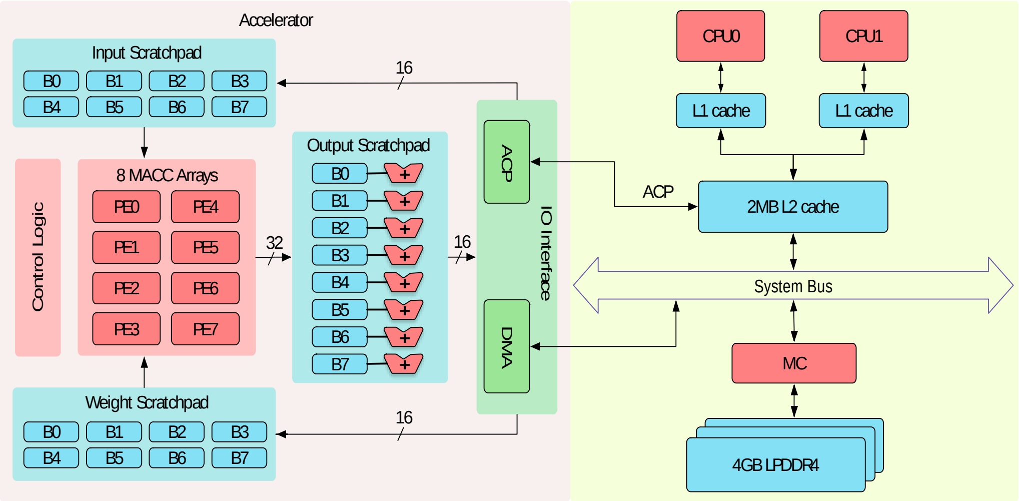
SMIV – Gem5-Aladdin model

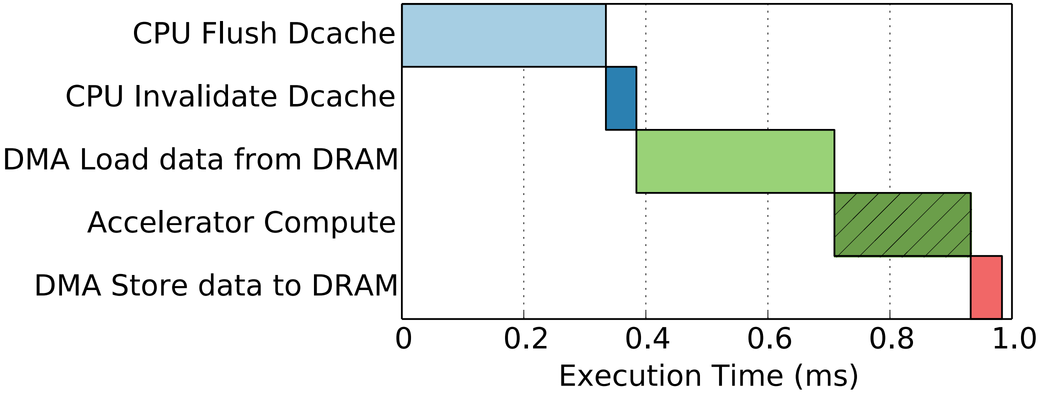
Leverages gem5-Aladdin framework

Compares different IO interfaces for interfacing accelerator to CPU across variety of DNN workloads running on NVDLA-like accelerator

Shows ACP with direct interface to CPU’s L2 cache can improve energy-delay product over using DMA

Multiple accelerators + other sys questions





[Xi et al., in review]

CHIPKIT: Tutorial on Agile Research Test Chips @MICRO’19

MICRO 2019

52nd IEEE/ACM International Symposium on Microarchitecture®

October 12-16, 2019

Columbus, Ohio, USA

microarch.org/micro52/program/workshops.html#chipkit

CHIPKIT: Tutorial on Agile Research Test Chips @MICRO’19

Outline

- Research test chips: fabrication routes, process technologies, project planning
- Test chip architectures: CPUs, peripherals, memories, interconnects and frameworks
- Design methodologies for custom blocks: Verilog, SystemVerilog, HLS and beyond
- Physical design flow: linting, synthesis, place and route, DRC/LVS, timing closure
- Bring up and test: packaging, PCBs, clocking, testing flows

Supporting materials

- CHIPKIT – a collection of generators and RTL glue for rapid SoC design
- SM2 – a simple Arm Cortex-M0 based scaffold for rapid (and functional!) test chips

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

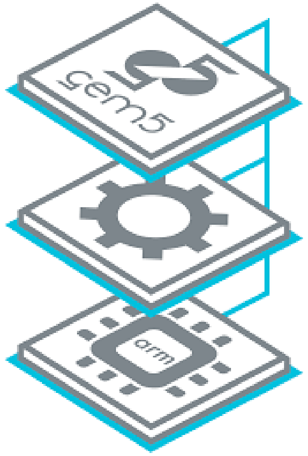
Compute systems modelling and architecture exploration

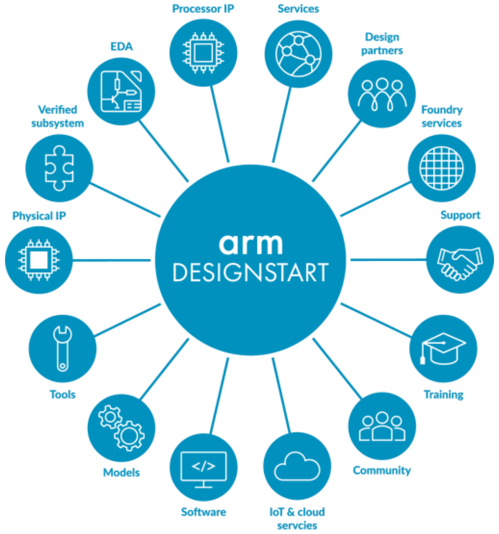
- Gem5 - CPU and system modelling

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

Acknowledgments

Harvard University

- Faculty: Gu-Yeon Wei, David Brooks, Sasha Rush
- Many talented PhD students and post-docs
- Generous sponsors

People, papers, software

- vlsiarch.eecs.harvard.edu
- nlp.seas.harvard.edu

P. N. Whatmough, S. K. Lee, H. Lee, S. Rama, D. Brooks and G. Wei, IEEE Int. Solid-State Cir. Conf. (ISSCC), 2017

P. N. Whatmough, S. K. Lee, D. Brooks and G. Wei, IEEE Journal of Solid-State Circuits (JSSC), 2018

S. K. Lee, P. N. Whatmough, N. Mulholland, P. Hansen, D. Brooks and G. Wei, IEEE Euro. Solid State Cir. Conf. (ESSCIRC), 2018

S. K. Lee, P. N. Whatmough, D. Brooks and G. Wei, IEEE Journal of Solid-State Circuits (JSSC), 2019

P. N. Whatmough, S. K. Lee, M. Donato, H.-C. Hsueh, S. L. Xi, U. Gupta, L. Pentecost, G. Ko, D. Brooks and G.-Y. Wei, Symposium on VLSI Circuits (VLSI), 2019

Thank You!

Danke!

Merci!

谢谢!

ありがとう!

Gracias!

Kiitos!

감사합니다

धन्यवाद