DARKSIDE: 2.6GFLOPS, 8.7mW Heterogeneous RISC-V Cluster for Extreme-Edge On-Chip DNN Inference and Training

A. Garofalo, M. Perotti, L. Valente, Y. Tortorella,
A. Nadalini, L. Benini, D. Rossi and F. Conti

DEI, University of Bologna, Italy &
IIS, ETH Zurich, Switzerland

angelo.garofalo@unibo.it

Outline

- **Introduction and Motivation**
- Darkside: Heterogeneous SoC Architecture
  - RVNN Cores
  - Depth-wise Engine
  - DataMover
  - Tensor Product Engine
- Chip Results Summary
  - Implementation Results
  - Benchmarking
  - Comparison with State-of-the-Art
- Conclusion

Introduction and Motivation

TinyML: Deploy DL and ML at the Extreme-Edge of IoT

- AI-enhanced IoT Applications;
- Reduced privacy issues, lower transmission power …

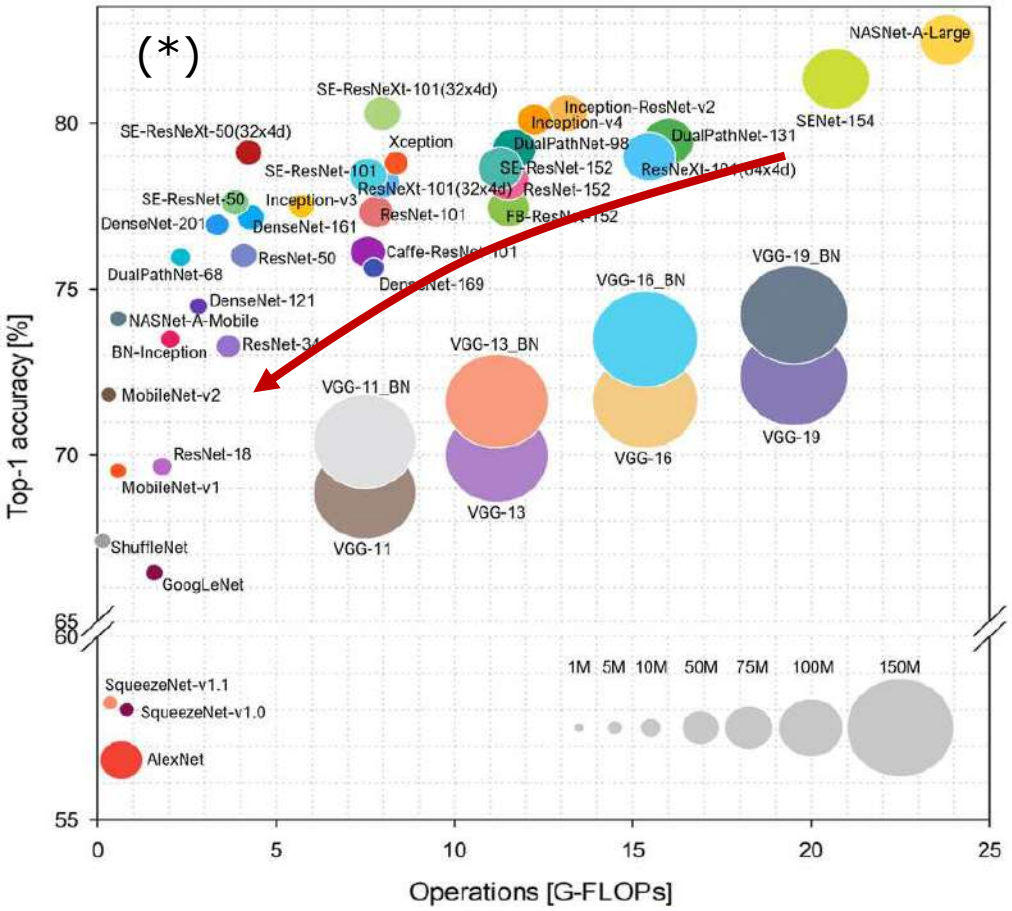
Challenges

- High computational and memory requirements (ML + DL);
- On-Chip Inference and Training within a power budget typical of MCU-class of devices (few hundreds mW).

Opportunities

- Reduced precision ML & DL models (both integer and floating-point);
- Low-Bitwidth Mixed-Precision integer computation;
- Specialized acceleration solutions (MAC, SIMD, Vectors, systolic arrays..).

Reduced Precision DL Models



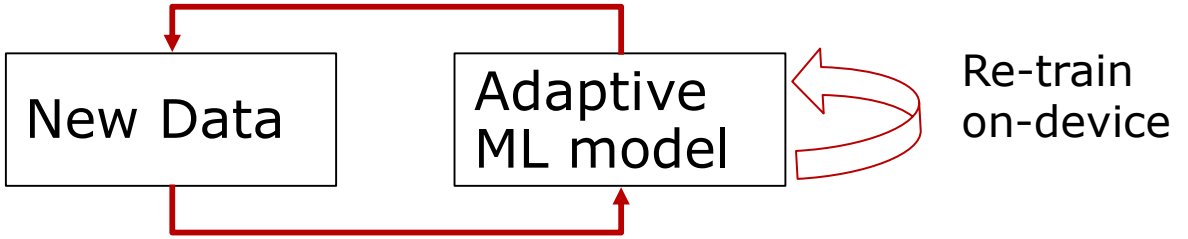
(*) Bianco, Simone, Remi Cadene, Luigi Celona, and Paolo Napoletano. "Benchmark analysis of representative deep neural network architectures." IEEE Access 6 (2018): 64270-64277.

❑ **DNN Inference**: Mixed-precision integer arithmetic

Quantization Method	Top1 Accuracy		Weight Memory Footprint	
Full-Precision	70.9%		16.27 MB	
INT-8	70.1%	**0.8%**	4.06 MB	**4x**
INT-4	66.46%	**4.4%**	2.35 MB	**7x**
Mixed-Precision	68%	**2.9%**	2.09 MB	**8x**

Courtesy of Rusci M. «Example on MobilenetV1_224_1.0.»C

❑ **On-Chip Training** (Continual, Federated Learning..):



❑ Floating-Point Arithmetic (16-bits) required

Extreme-Edge AI Computing Platforms

	ASICs	FPGAs	MCUs
Throughput [Gop/s]	**1 K – 50 K**	10 – 200	0.1 – 2
Energy Efficiency [Gop/s/W]	**10 K – 100 K**	1 - 10	1 – 50
Flexibility/Programmability	**Low**	Medium	**High**

Mixed-Precision kernel: RISC-V Assembly

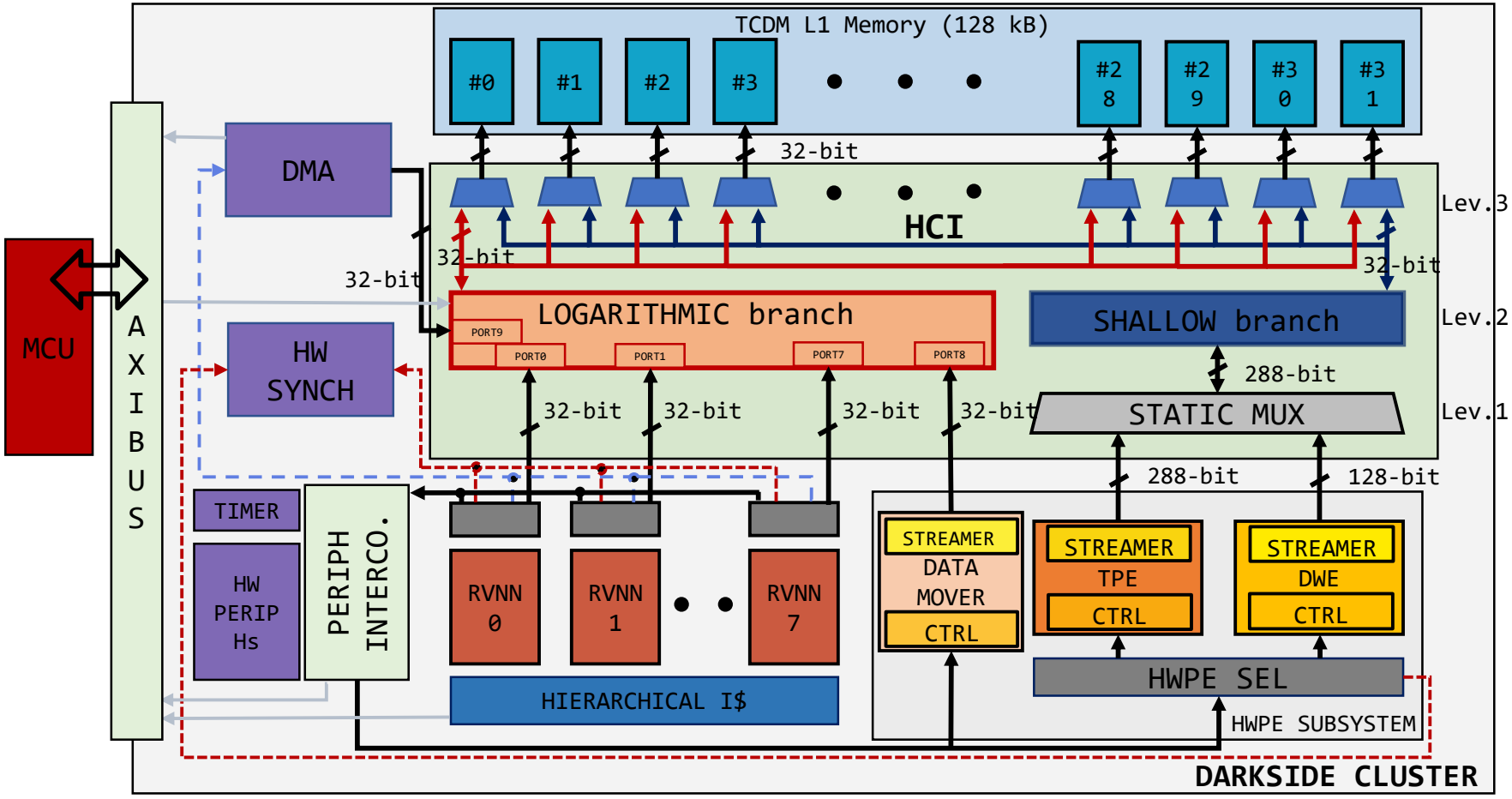
```
    p.lw    x10,4(x4!)
    p.lw    x11,4(x5!)
→   p.extract x5, x11, 4, 0
→   p.extract x6, x11, 4, 4
→   p.extract x7, x11, 4, 8
→   p.extract x8, x11, 4, 12
→   pv.packlo.b x15, x5, x6
→   pv.packhi.b x15, x7, x8
    pv.sdotsp.b x20, x15, x10
```

- **IoT End-Nodes scenario (MCUs):**
  - Lack of support for mixed-precision integer arithmetic at ISA level (RISCV, ARM); → Huge Overhead!
  - Missing-low power specialized solutions to speed-up low-reuse kernels, compute-intensive floating-point workloads.
- **This Work:**
  - Heterogeneous Compute Cluster:
    - Enhanced RISC-V cores with advanced integer mixed-precision capabilites;
    - Tightly-Coupled Specialized accelerators to boost heavy kernels dominating the workload;

Outline

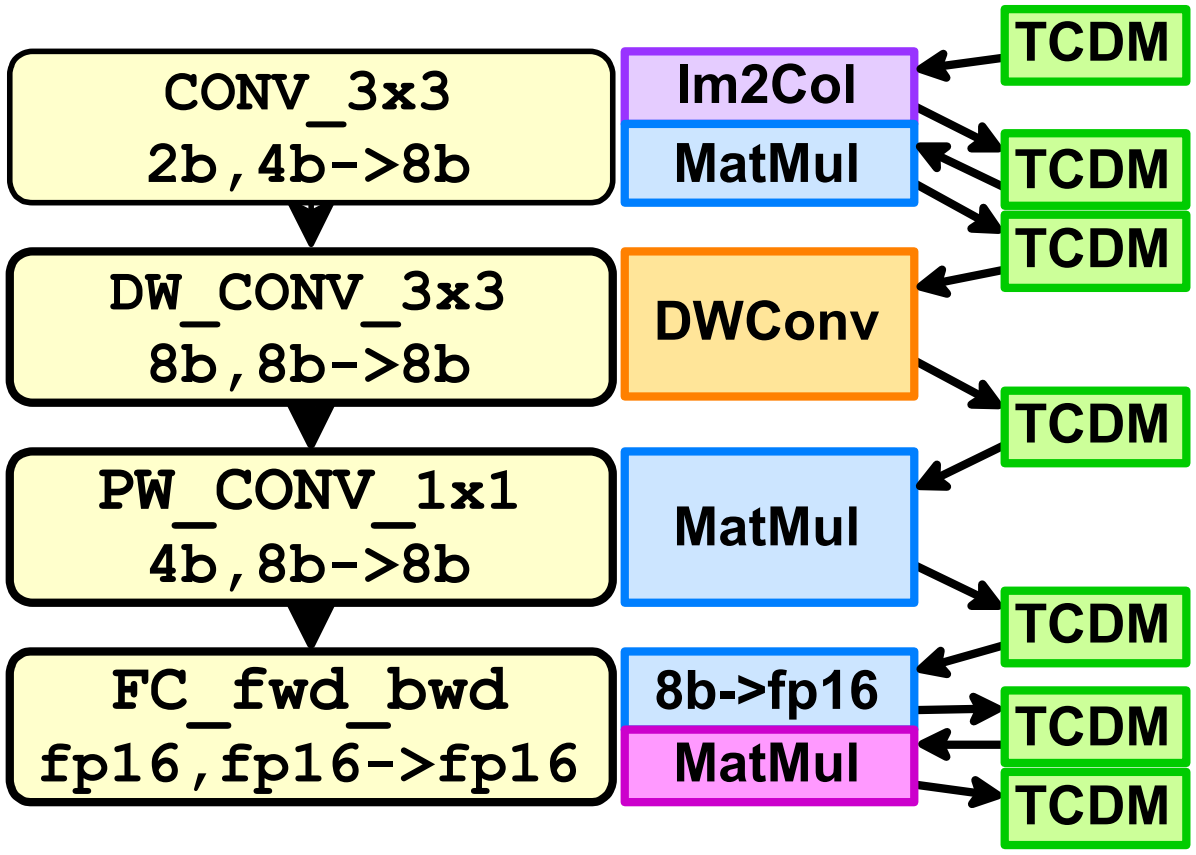
- Introduction and Motivation
- **Darkside: Heterogeneous SoC Architecture**
  - RVNN Cores
  - Depth-wise Engine
  - DataMover
  - Tensor Product Engine
- Chip Results Summary
  - Implementation Results
  - Benchmarking
  - Comparison with State-of-the-Art
- Conclusion

Darkside Architecture

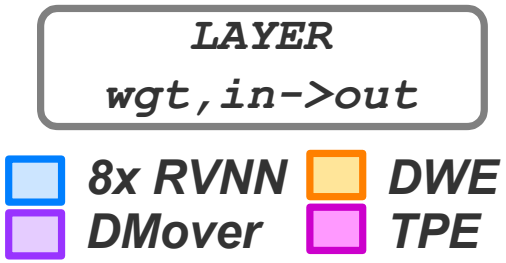


- 8 **RVNN** cores (32-b custom RISC-V ISA) ;
- Depth-wise Engine (**DWE**) to boost low-reuse depthwise;
- Tensor Product Engine (**TPE**) to boost IEEE FP16 MatMuls;
- **DataMover** for efficient data marshalling;
- Accelerators encapsulated within standardized Hardware Processing Engine (HWPE) interface;
- Heterogeneous Cluster Interconnect (**HCI**) for tightly-coupled integration;
- **DMA** Controller for Double Buffering/DNN model tiling;
- **HW synchronization** unit for efficient parallelization and event-based execution.

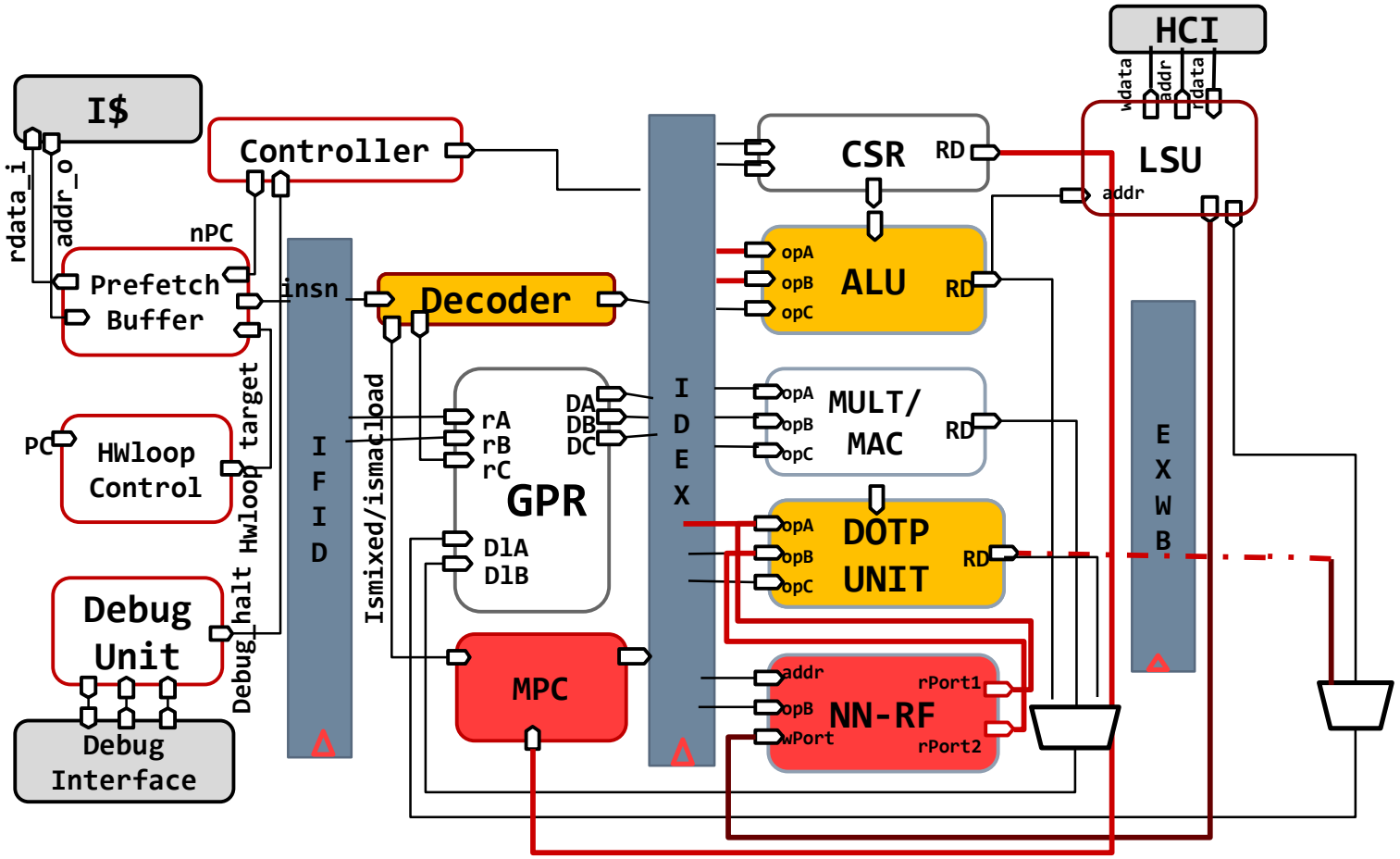
Darkside Execution Model



- TCDM as tightly-coupled memory buffer for all the compute units of the cluster:
- Efficient cooperation among hardware compute units;
- Support complex ML and DL execution models (e.g. full MobileNetV2, FC Autoencoder);



RVNN Cores μArchitecture



- **RI5CY:** 4-stage in order single-issue pipeline;
  - **69 kGE**;
- **ISA:** RV32IMCFXpulp
  - Support for HW Loops;
  - Auto-Increment LD/ST instructions;
  - 16-/8-bit SIMD operations;
  - Bit Manipulation instructions.
- **RVNN core** (XpulpNN ISA)
  - Hardware support for SIMD mixed-precision operations (2b-to-32b);
  - Fused Mac-Load (M&L) instruction operating on a dedicated NN-RF;
  - **81 kGE** (**17%** overhead w.r.t. RI5CY)
  - Only **3%** power overhead:
    - Clock-gating cells, operand isolation gates.

Mac-Load: MatMul Inner Kernel

Without M&L (Xpulp – RI5CY)

```
lp.setup
p.lw        w1,4(aw1!)
p.lw        w2,4(aw2!)
p.lw        w3,4(aw3!)
p.lw        w4,4(aw4!)
p.lw        x1,4(ax1!)
p.lw        x2,4(ax2!)
pv.sdotsp   s1,x1,w1
pv.sdotsp   s2,x1,w2
pv.sdotsp   s3,x1,w3
pv.sdotsp   s4,x1,w4
pv.sdotsp   s5,x2,w1
pv.sdotsp   s6,x2,w2
pv.sdotsp   s7,x2,w3
pv.sdotsp   s8,x2,w4
```

HW LOOP

LD/ST WITH POST INCREMENT

MIXED-PRECISION SIMD MAC

8 SIMD MACs with 6 explicit LOADs

INIT NN-RF

16 mldotp

With M&L (XpulpNN – RVNN)

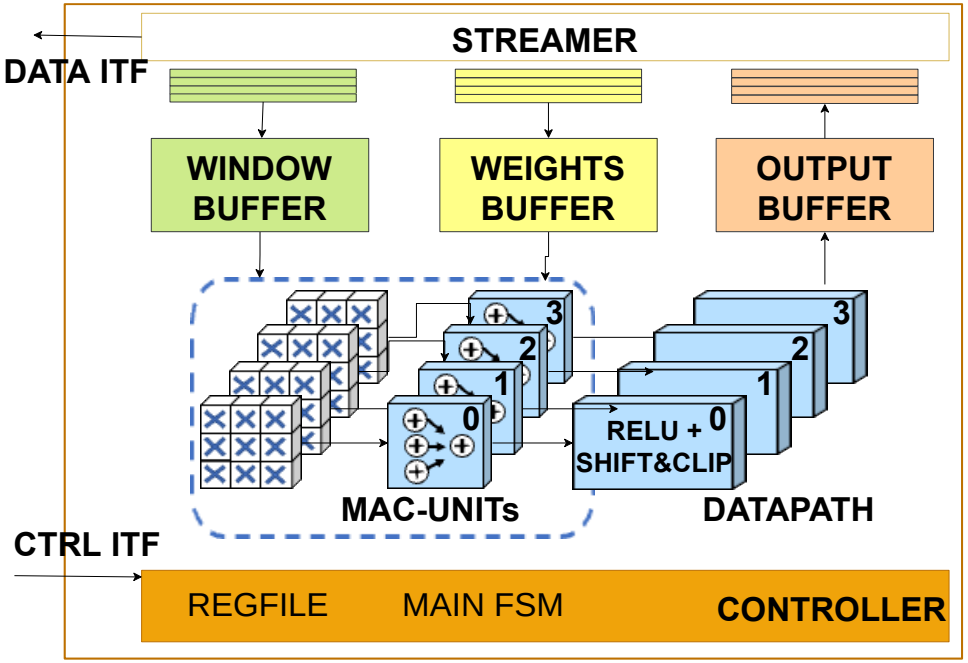
```
pv.ml_load   aw1,16
pv.ml_load   aw2,18
pv.ml_load   aw3,20
pv.ml_load   aw4,22
pv.ml_load   ax1,8
lp.setup
pv.ml_load   ax2,9
pv.mlsdotsp  s1,aw2,0
pv.mlsdotsp  s2,aw4,2
pv.mlsdotsp  s3,aw3,4
        . . .
        . . .
pv.mlsdotsp  s14,aw2,19
pv.mlsdotsp  s15,aw3,21
pv.mlsdotsp  s16,aw4,23
```

- MAC operands reside into NN-RF;
- More registers available in the GP-RF to implement kernels with higher data-reuse;

16 SIMD MACs with 1 explicit LOAD

- ❑ Up to **94%** of SIMD Dotp Unit Utilization on MatMul kernels
- ❑ Up to **12.7x** performance improvements over RI5CY on mixed-precision MatMul kernels

Depth-Wise Engine & DataMover

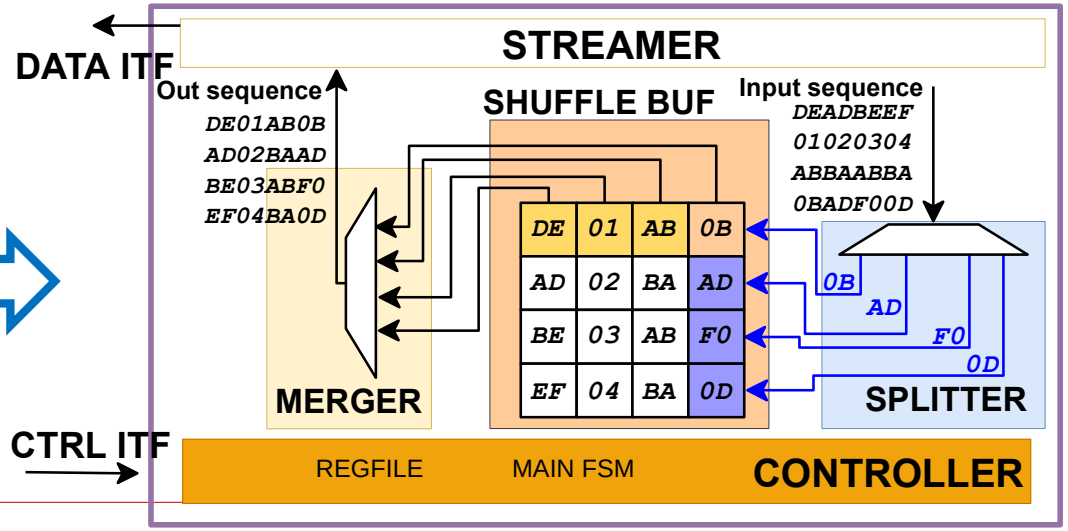


Depth-Wise Engine (DWE):

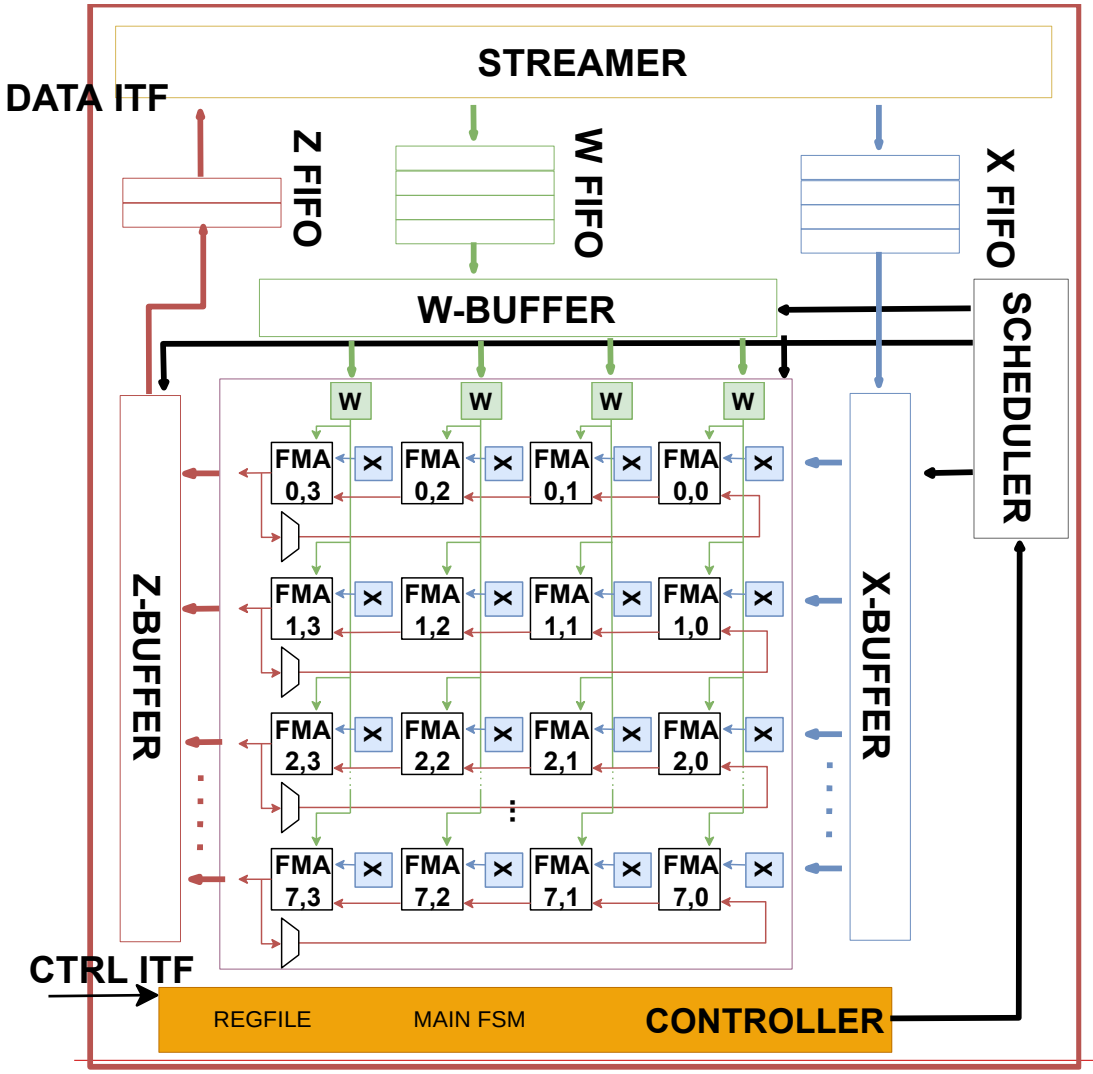
- Boost low-reuse 8-bit (integer) 3x3 Depth-wise convolutions;
- Weight-Stationary Data Flow to maximize data reuse;
- Fully exploit memory bandwidth of 36B/cyc through *shallow* HCI branch;
- Peak performance of **30 MAC/cycle**;
- **131 kGE**.

DataMover:

- 1b-32b configurable precision on-the-fly efficient data transposition;
- Up to **100x** less transposition time than SW (scales with precision of data to transpose);
- **54 kGE**.

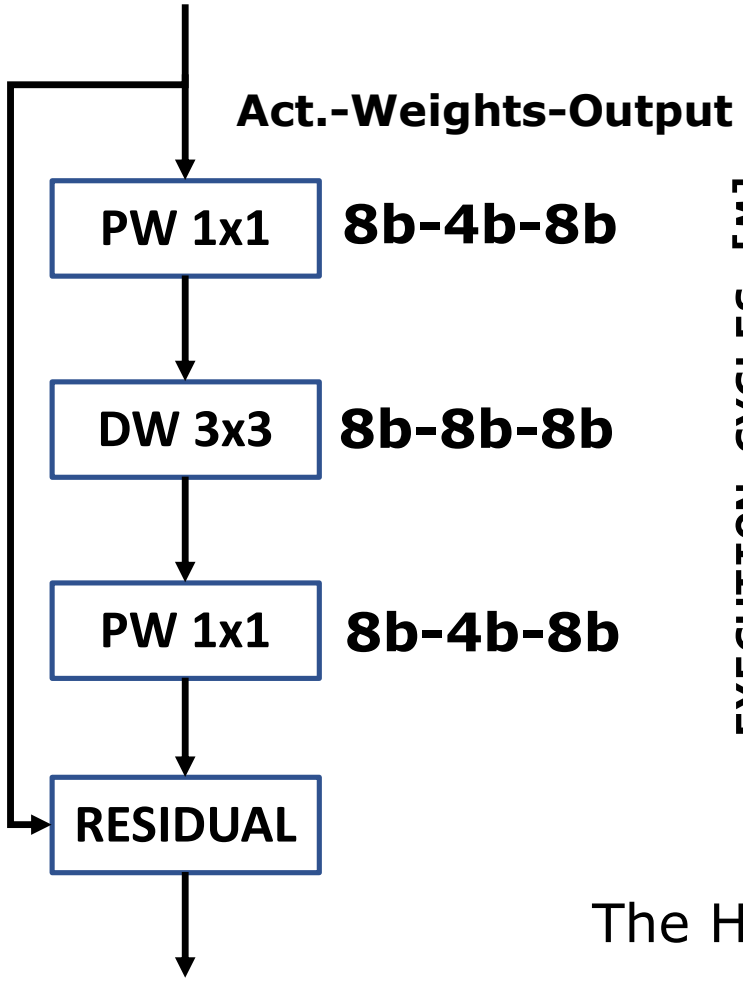


Tensor Product Engine

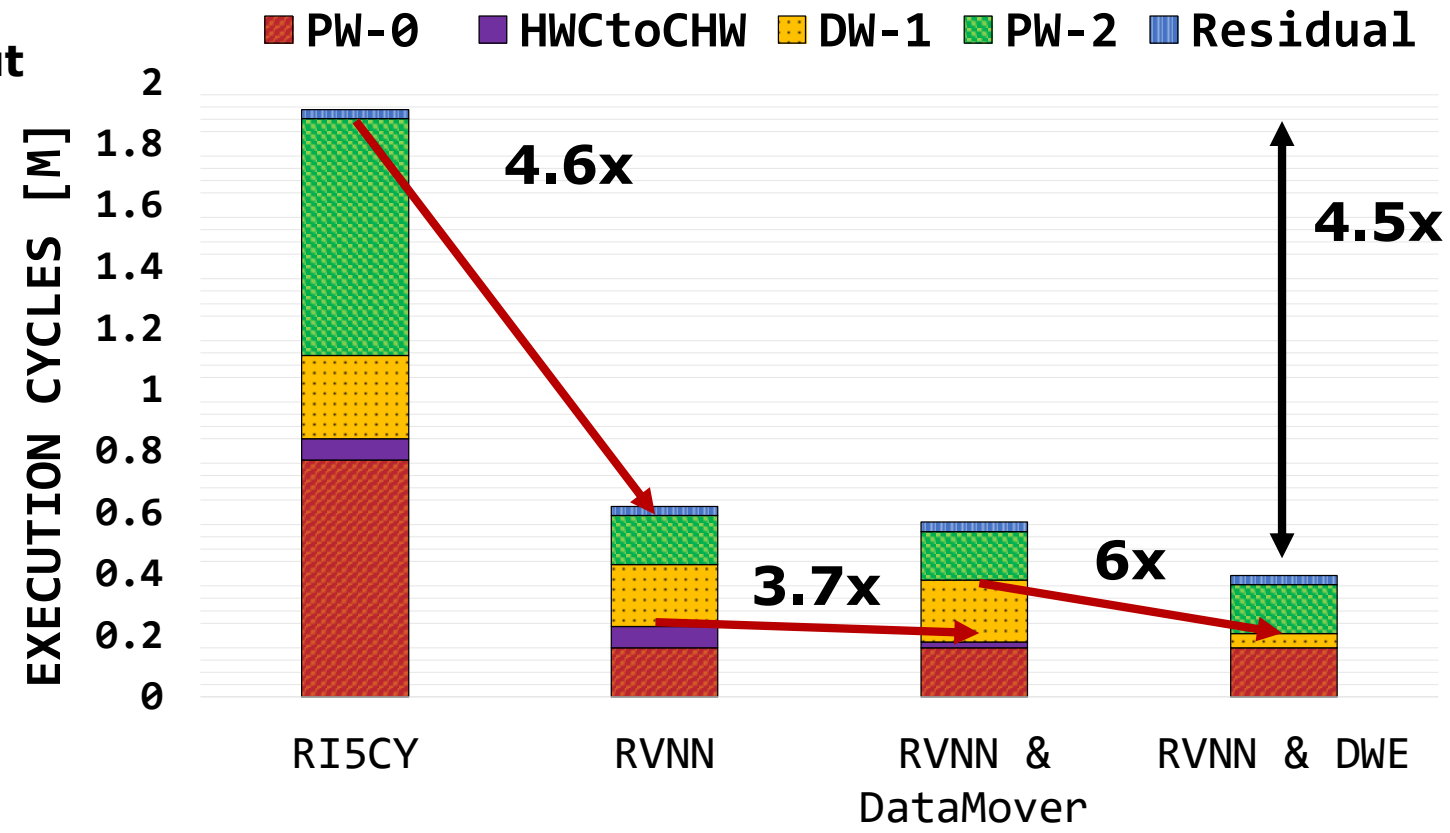


- Boost IEEE FP16 Matrix Multiplications;
- Datapath:
  - array of 32 FMA Units, organized over 8 rows and 4 columns;
  - FMAs cascaded along the rows;
  - Trade-off between performance and area;
- Execution scheduling optimized:
  - Streaming always overlaps computation;
  - Data reuse is maximized;
- **98%** of FMAs Utilization;
- Near-to-ideal performance: **31.6 GMAC/cycle** (ideal is 32 GMAC/cycle);
- Up to 22x Speed-Up over SW MatMuls execution;
- **292 kGE**.

Bottleneck Layer



Mixed-Precision integer *Bottleneck* Layer

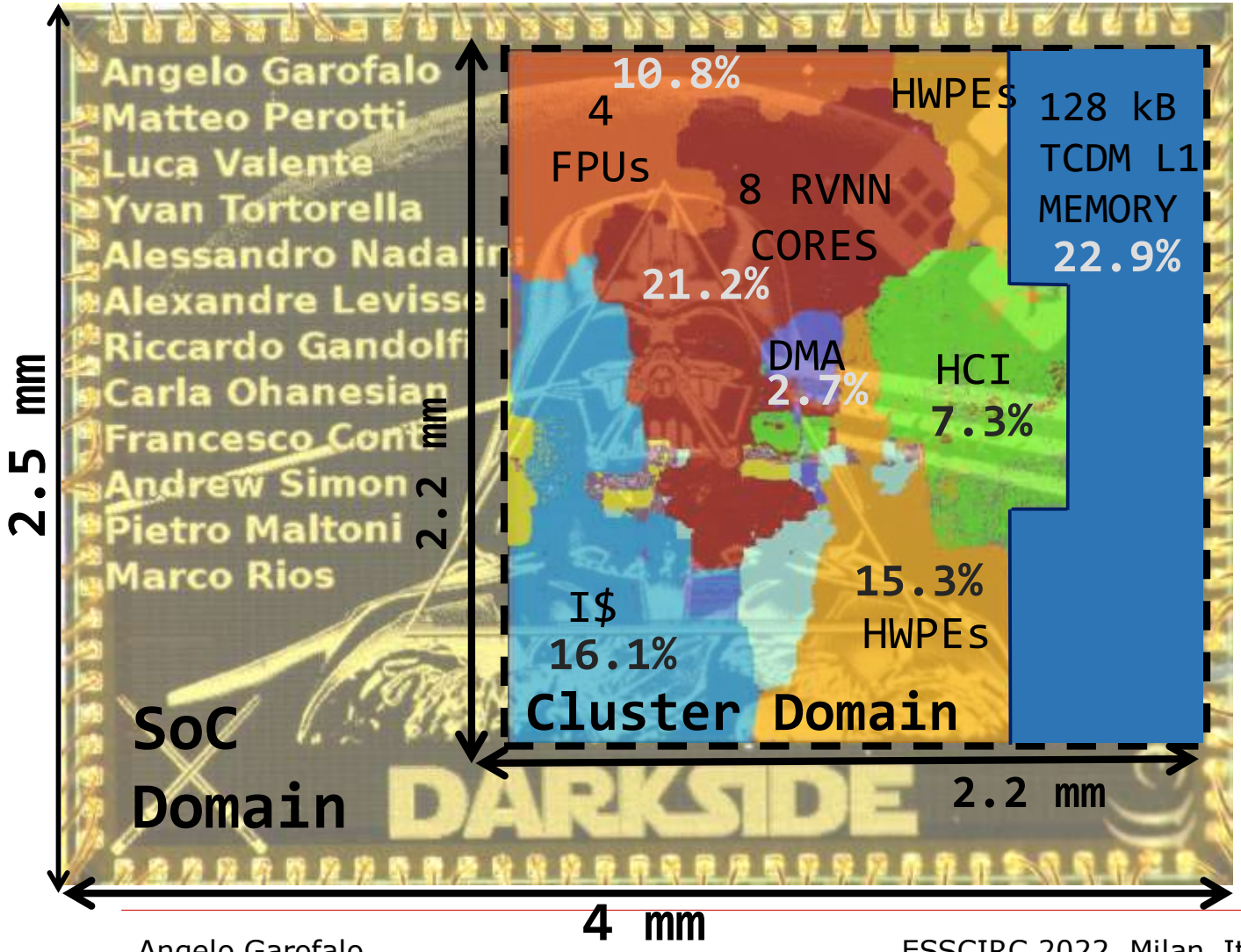


The Heterogeneous approach mitigates Amdahl's effects on heterogeneous workloads

Outline

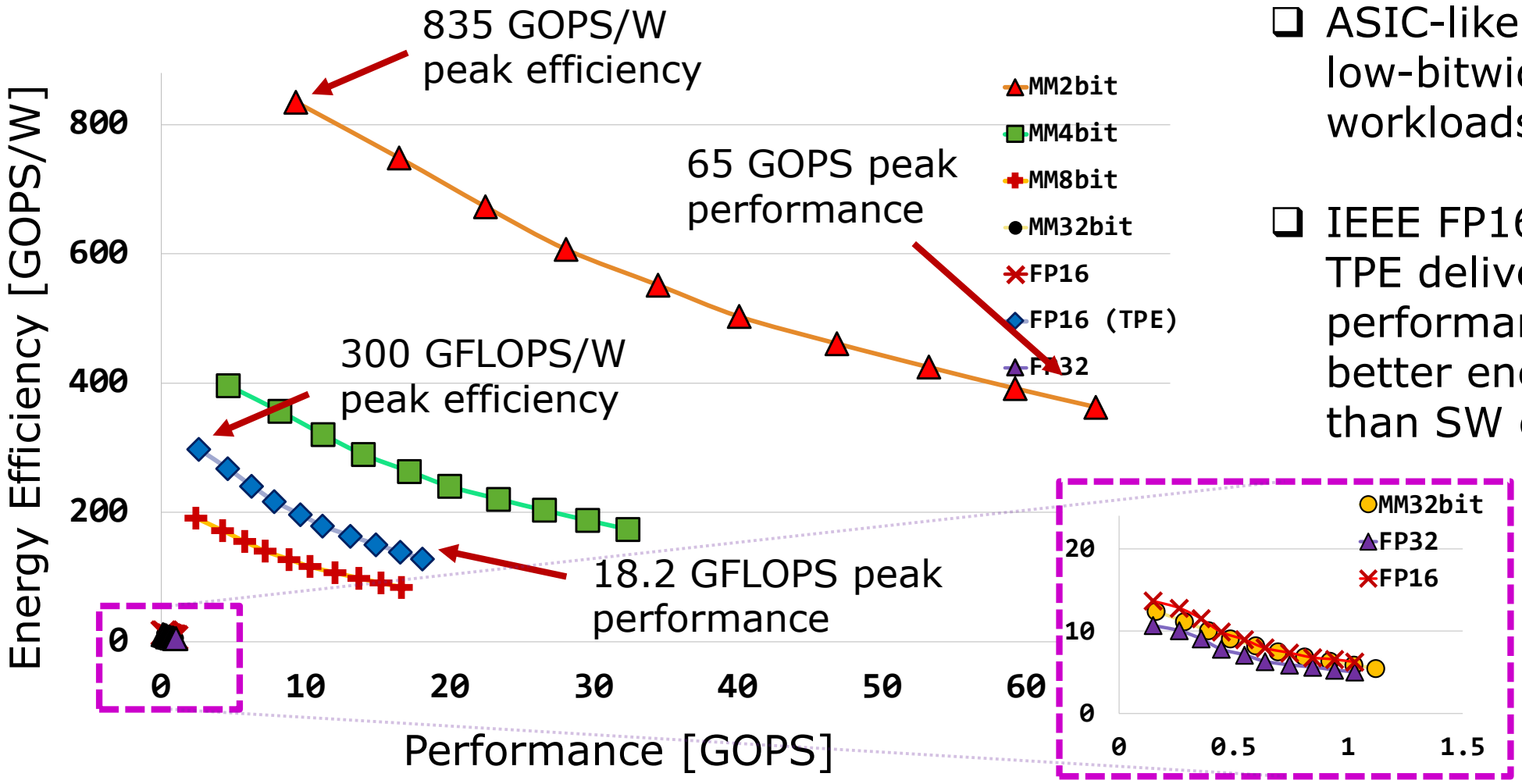
- Introduction and Motivation
- Darkside: Heterogeneous SoC Architecture
  - RVNN Cores
  - Depth-wise Engine
  - DataMover
  - Tensor Product Engine
- **Chip Results Summary**
  - Implementation Results
  - Benchmarking
  - Comparison with State-of-the-Art
- Conclusion

Chip Results Summary



Technology	CMOS 65nm
Chip Area	12 mm^2
Cluster Area	**4.84 mm^2**
Total SRAM	384 kB
Cluster SRAM	128 kB
Frequency Range	40 – **290 MHz**
Vdd Range	0.75 – 1.2 V
Power Envelope	**213 mW**

Performance vs. Energy Efficiency



- ASIC-like efficiency on low-bitwidth integer workloads;
- IEEE FP16 MatMuls on TPE delivers **17.7x** better performance, **21.8x** better energy efficiency than SW execution.

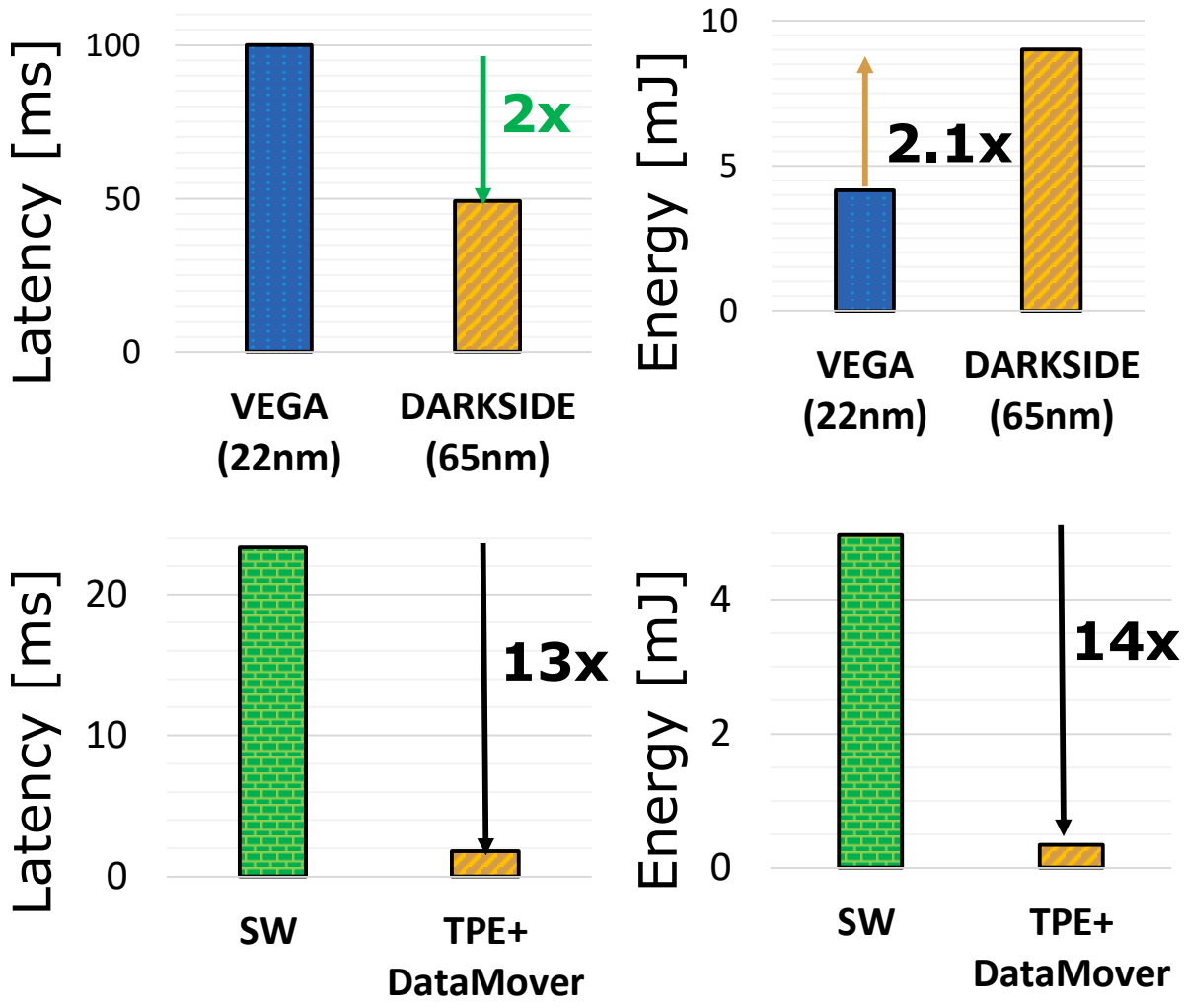
TinyML Benchmarks

End-to-End Inference

- Mixed-Precision MobileNetV2:
  - ~1MB footprint;
  - 69.4% Top-1 Accuracy;
- Optimized execution flow for efficient L2-L1 data movements [Burr21];
- Performance: **20 frame/s** (@290 MHz)**;**
- Energy per Inference: **9.1 mJ** (65nm).

On-Chip Training

- FC TinyML AutoEncoder
- One training epoch benchmarked
  - Full forward, backward steps and weight updates;
- Latency: **1.8 ms** (@290 MHz);
- Energy: **345 µJ**;



[Burr21]: Burrello, A., et al.. Dory: Automatic end-to-end deployment of real-world dnns on low-cost iot mcus. *IEEE Transactions on Computers*

Comparison with SoA

	SleepRunner	SamurAI	VEGA	Dustin	This work
Technology	28nm	28nm	22nm	65nm	**65nm**
CPU	**1x**CM0DS	**1x**RI5CY	**10x**RI5CY	**16x**MPIC (RV)	**8x**RV-NN (RV)
INT Precision	32b	8b-32b	8b-32b	2b-32b mixed-precision	2b-32b mixed-precision
FP Precision	--	--	FP32, FP16, bfloat	--	FP32, FP16
Best Int Perf. Best.Int Eff. @ Perf. (**8-bit**)	31 MOPS, 97MOPS/mW @18 MOPS	1.5 GOPS, 230 GOPS/W @110 MOPS	15.6 GOPS, 614 GOPS/W @7.6 GOPS	15 GOPS, 303 GOPS/W @ 4.4 GOPS	**17 GOPS, 191 GOPS/W @2.4 GOPS**
Best FP32 Perf. Best. FP32 Eff. @ Perf.	--	--	2 GFLOPS, 79 GFLOPS/W @ 1 GFLOPS	--	1.03 GFLOPS, 12 GFLOPS/W @ 0.4 GFLOPS
Best IEEE **FP16** Perf. Best. IEEE **FP16** Eff. @ Perf.	--	--	3.3 GFLOPS, 129 GFLOPS/W @1.27 GFLOPS	--	**18.2 GFLOPS, 300 GFLOPS/W @2.6 GFLOPS**

Outline

- Introduction and Motivation
- Darkside: Heterogeneous SoC Architecture
  - RVNN Cores
  - Depth-wise Engine
  - DataMover
  - Tensor Product Engine
- Chip Results Summary
  - Implementation Results
  - Benchmarking
  - Comparison with State-of-the-Art
- **Conclusion**

Conclusion

- Darkside: Low-Power Heterogeneous Compute Cluster for in **65nm**;
- RVNN cores with 2b-to-32b mixed-precision integer computing capabilities and Mac-Load instructions (**2x** to **12.7x** speed-up over RI5CY on linear kernels);
- Tightly-Coupled Accelerators to boost depth-wise kernels (up to **10x** speed-up over SW) and data marshalling operations (up to **100x** speed-up over SW);
- Tensor Product Engine (TPE) to boost FP16 MatMuls, achieving **300 GFLOPS/W within only 8.7 mW**;
- End-to-end inference and training workloads (full MobileNetV2, FC AutoEncoder) with better or comparable metrics than SoA solutions;
- Darkside is competitive with IoT end-nodes using much more scaled technology nodes (Peak Integer Perf. **65 GOPS**, En. Eff. **835 GOPS/W**; Peak FP Perf. **18.2 GFLOPS**, En. Eff. **300 GFLOPS/W**);