A 1.15 TOPS/W, 16-Cores Parallel Ultra-Low Power Cluster with 2b-to-32b Fully Flexible Bit-Precision and Vector Lockstep Execution Mode

A. Garofalo, G. Ottavi, A. Di Mauro, F. Conti, L. Benini, D. Rossi
DEI
University of Bologna, Italy
angelo.garofalo@unibo.it

September 6-9, 2021

Outline

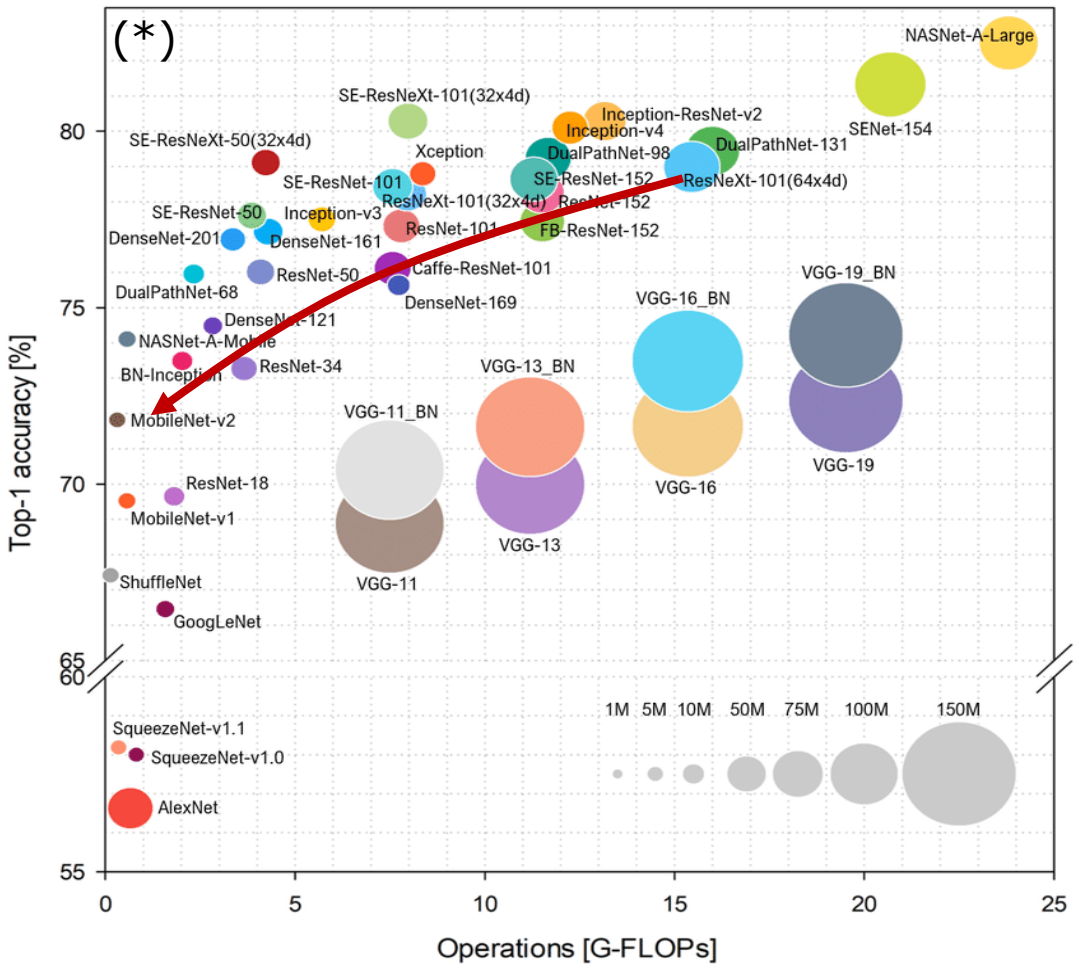
- **Introduction & Motivation**
- Dustin Architecture Overview
  - Tunable Mixed-Precision Computation
  - Vector Lockstep Execution Mode
- Chip Results Summary
- Comparison with the State-of-the-art
- Conclusion

Introduction & Motivation

- **Extreme Edge AI and TinyML**
  - Low latency and network load compared to cloudML;
  - Eases privacy concerns
- **Challenges**
  - High computational and memory requirements (ML + DL);
  - Limited resources on IoT End-Nodes (microcontrollers).
- **Opportunities**
  - Reduce DL/ML model size;
  - Low-Bitwidth Mixed-Precision computation.
  - Reduce instruction fetch and decode overhead exploiting data-parallel execution




Quantized Neural Networks (QNNs)



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

Mixed-precision Quantized Neural Networks (QNNs) are the natural target for execution on constrained edge platforms.

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

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	**Low**	Medium	**High**
Cost	**High**	Medium	**Low**

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
```

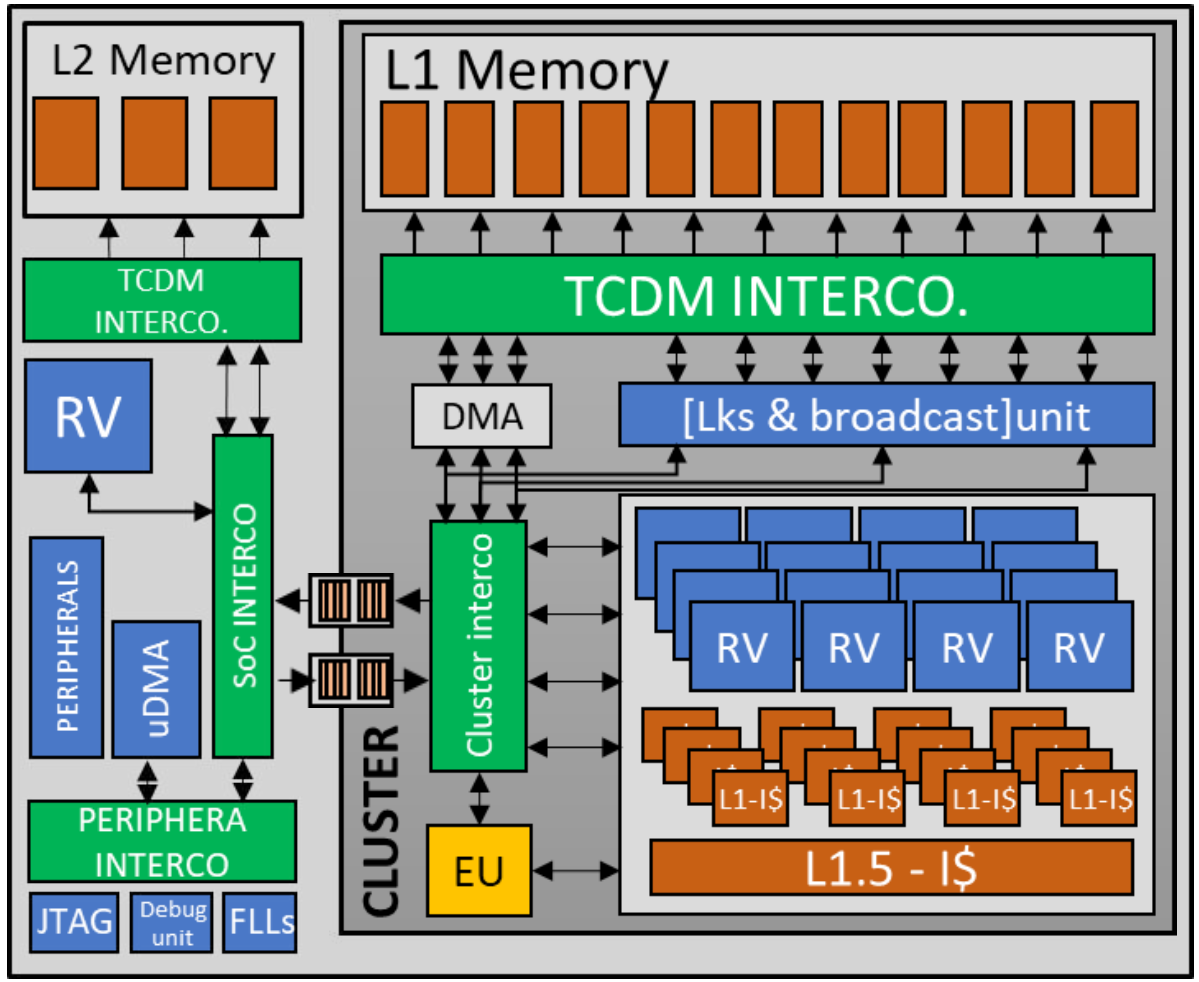
- **IoT End-Nodes scenario:**
  - Must be inexpensive and software programmable (MCUs);
  - SoA ISAs (RISCV (*), ARM (**)) support only integer uniform arithmetic (with SIMD);
  - Huge overhead to perform mixed-precision computation for data casting and packing;
- **This Work:**
  - Low-Power IoT End-Node with a fully programmable RISCV accelerator cluster;
  - Mixed precision 2b-to-32b SIMD instructions in the ISA of the cores;
  - *Vector Lockstep Exec. Mode* to boost the Efficiency on data-parallel DL/ML algorithms.

(*) Garofalo et al. "XpulpNN: Enabling Energy Efficient and Flexible Inference of Quantized Neural Networks on RISC-V based IoT End Nodes." IEEE Transactions on Emerging Topics in Computing (2021).
(**) D. E. Joseph Yiu, "Introduction to the arm cortex-m55 processor. : https://pages.arm.com/cortex-m55-introduction.html," Feb. 2020.

Outline

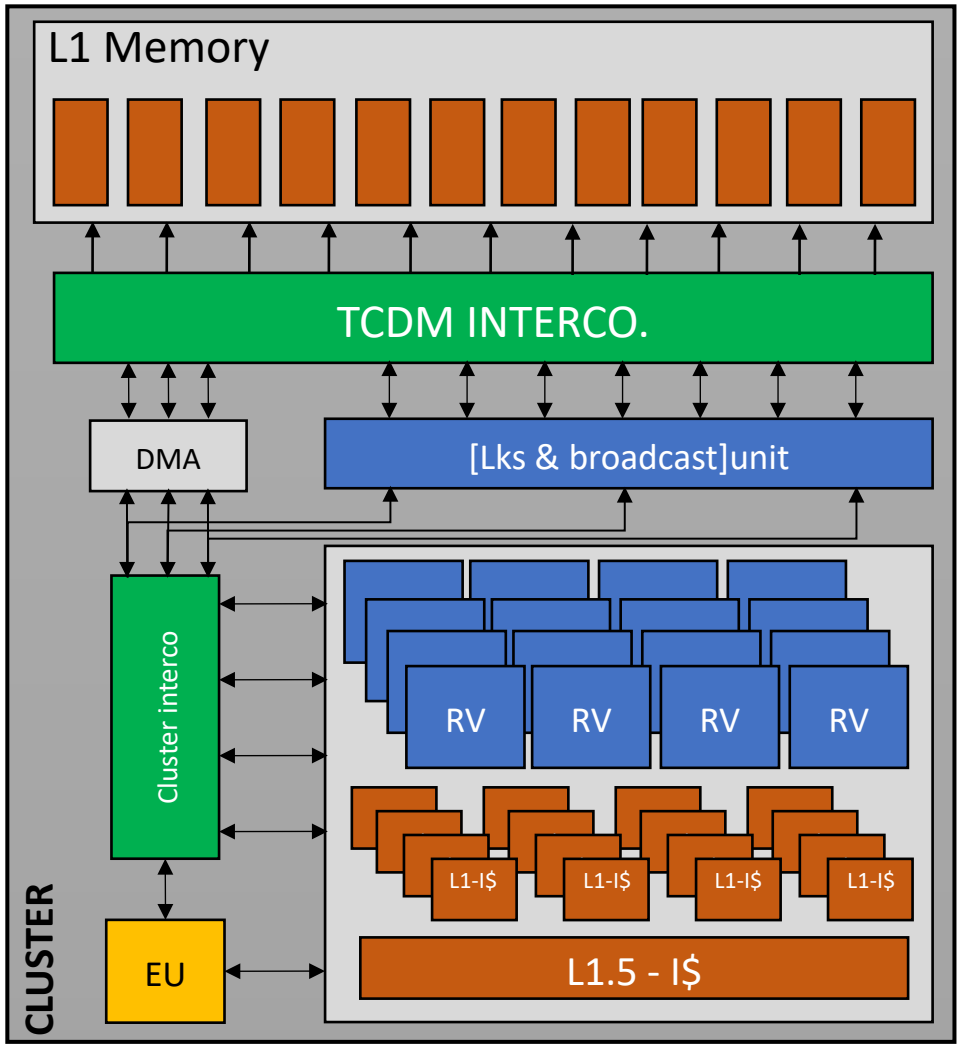
- Introduction & Motivation
- **Dustin Architecture Overview**
  - Tunable Mixed-Precision Computation
  - Vector Lockstep Execution Mode
- Chip Results Summary
- Comparison with the State-of-the-art
- Conclusion

Dustin: Architecture Overview



- **The SoC is a Low-Power IoT End-Node with AI edge computing capabilities**
  - Microcontroller
    - 1 RISCV core
    - 112 kB of L2 memory
    - Rich sets of peripherals (UART, I2C, CAM itf..)
    - JTAG (Debug), GPIOs, ROM
    - Interrupt Controller
  - Parallel cluster accelerator of fully programmable RISC-V cores

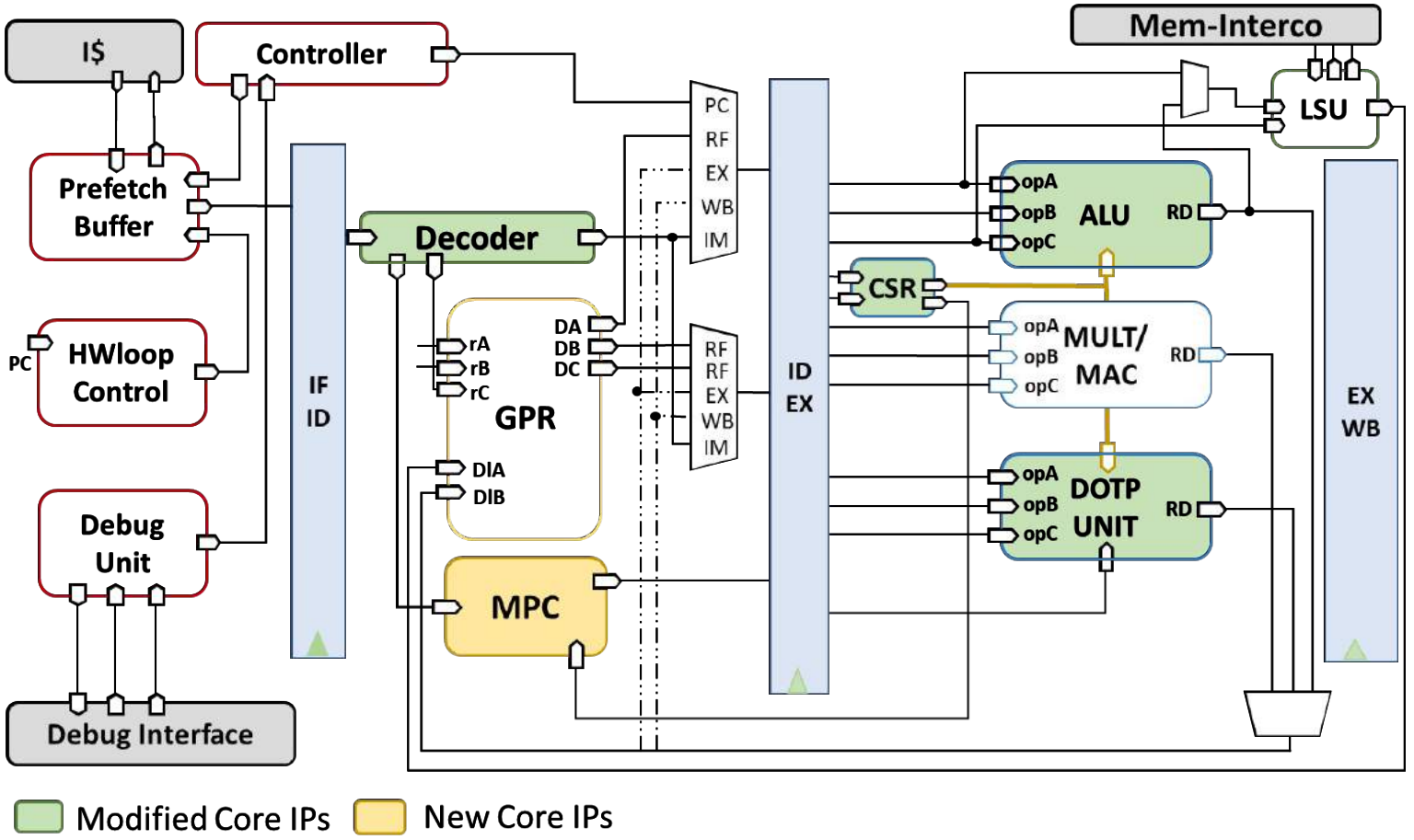
Dustin: Cluster



- **Accelerator Cluster**
  - 16 RI5CY (*) cores augmented with 2b-to-32b SIMD instructions;
  - Software Configurable *Vector Lockstep Execution Mode (VLEM);*
  - Single-cycle latency TCDM interco. leveraging a req/gnt protocol, word-level interleaved scheme.
  - 128 kB of Shared Tightly-Coupled L1 Data Memory;
  - Hierarchical Instruction Cache;
  - High performance DMA (L2 <-> L1);
  - Event Unit supporting efficient synchronization among the cores;

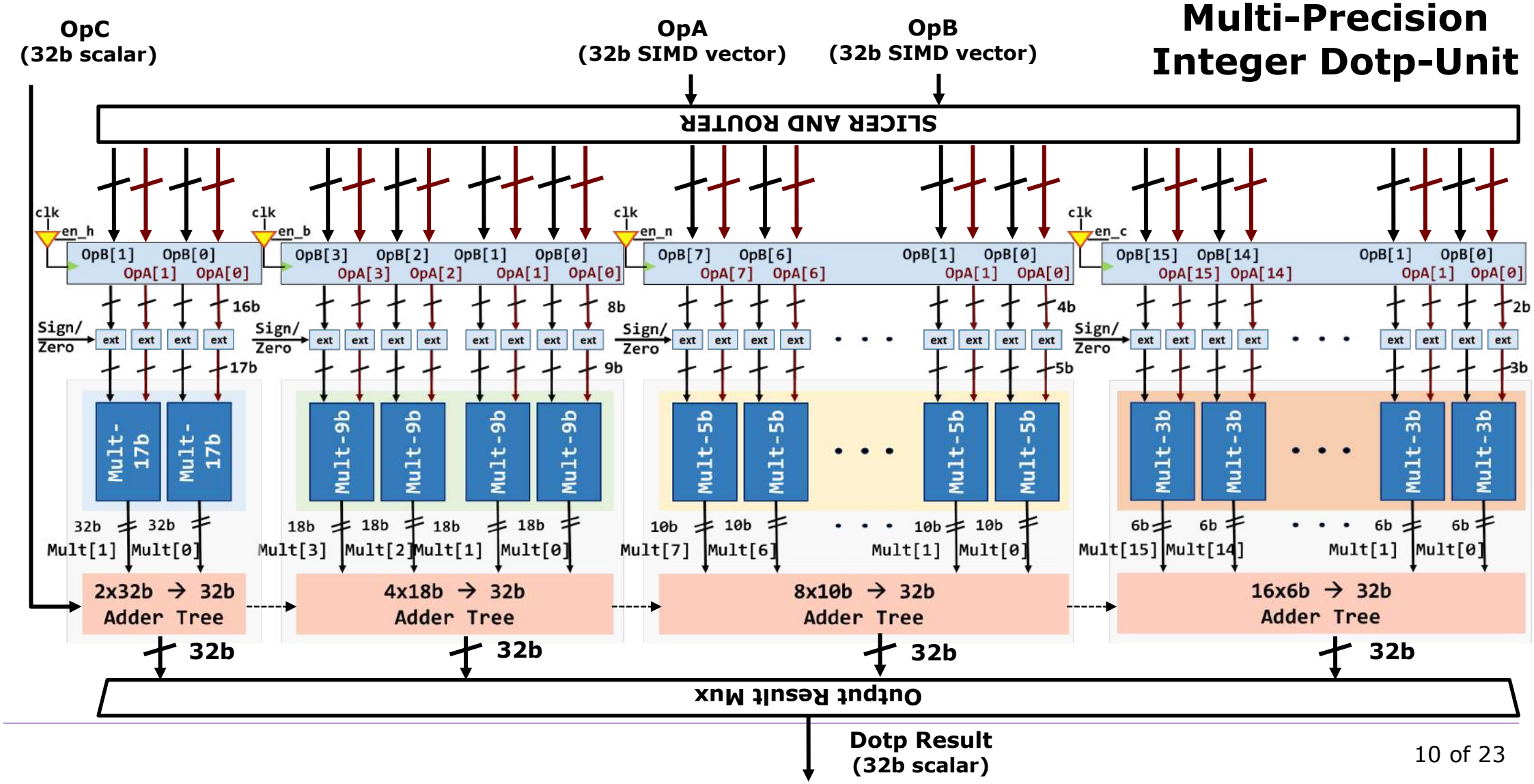
(*) Gautschi et al., Near-threshold RISC-V core with DSP extensions for scalable IoT endpoint devices, IEEE VLSI, 2017.

Core Enhancements

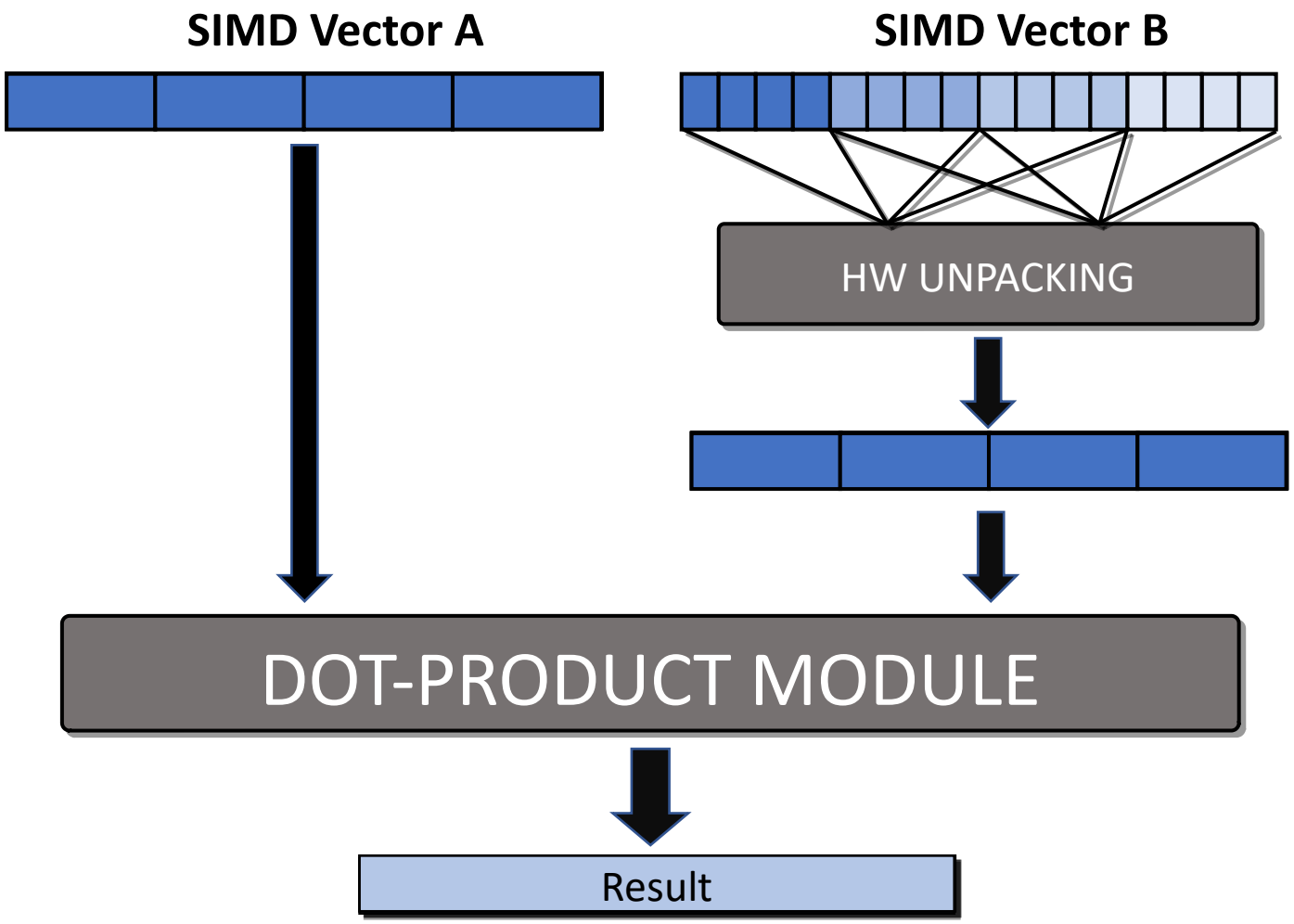


- **RI5CY**: 4-stage in order single-issue pipeline
- **ISA**: RV32IMCXpulpV2
- **XpulpV2** extensions:
  - HW Loops;
  - Post-Increment LD/ST;
  - 16-/8-bit SIMD insns;
  - Bit Manip. insns.
- Goal
  - HW support for **mixed-precision** SIMD instructions;
- Challenge
  - Enormous number of instructions to be encoded in the ISA;
- Solution
  - **Dynamic Bit-Scalable Precision**

Extended Dot-Product Unit



Mixed-Precision Controller



- Mixed-Precision operations require a controller: selection of the correct sub-word of the lowest precision operand (Vector B) to be used in current SIMD op.
- The controller is programmed by control status registers (CSRs).

Dynamic Bit-Scalable Execution

Standard Instructions

pv.dotsp.h	pv.dotup.h	pv.dotusp.h	pv.sdotsp.h	pv.sdotup.h	pv.sdotusp.h
pv.dotsp.b	pv.dotup.b	pv.dotusp.b	pv.sdotsp.b	pv.sdotup.b	pv.sdotusp.b
pv.dotsp.n	pv.dotup.n	pv.dotusp.n	pv.sdotsp.n	pv.sdotup.n	pv.sdotusp.n
pv.dotsp.c	pv.dotup.c	pv.dotusp.c	pv.sdotsp.c	pv.sdotup.c	pv.sdotusp.c
pv.dotsp.m4x2	pv.dotup.m4x2	pv.dotusp.m4x2	pv.sdotsp.m4x2	pv.sdotup.m4x2	pv.sdotusp.m4x2
pv.dotsp.m8x2	pv.dotup.m8x2	pv.dotusp.m8x2	pv.sdotsp.m8x2	pv.sdotup.m8x2	pv.sdotusp.m8x2
pv.dotsp.m8x4	pv.dotup.m8x4	pv.dotusp.m8x4	pv.sdotsp.m8x4	pv.sdotup.m8x4	pv.sdotusp.m8x4
pv.dotsp.m16x8	pv.dotup.m16x8	pv.dotusp.m16x8	pv.sdotsp.m16x8	pv.sdotup.m16x8	pv.sdotusp.m16x8
pv.dotsp.m16x4	pv.dotup.m16x4	pv.dotusp.m16x4	pv.sdotsp.m16x4	pv.sdotup.m16x4	pv.sdotusp.m16x4
pv.dotsp.m16x2	pv.dotup.m16x2	pv.dotusp.m16x2	pv.sdotsp.m16x2	pv.sdotup.m16x2	pv.sdotusp.m16x2
pv.dotsp.sc.h	pv.dotup.sc.h	pv.dotusp.sc.h	pv.sdotsp.sc.h	pv.sdotup.sc.h	pv.sdotusp.sc.h
pv.dotsp.sc.b	pv.dotup.sc.b	pv.dotusp.sc.b	pv.sdotsp.sc.b	pv.sdotup.sc.b	pv.sdotusp.sc.b
pv.dotsp.sc.c	pv.dotup.sc.c	pv.dotusp.sc.c	pv.sdotsp.sc.c	pv.sdotup.sc.c	pv.sdotusp.sc.c
pv.dotsp.sc.n	pv.dotup.sc.n	pv.dotusp.sc.n	pv.sdotsp.sc.n	pv.sdotup.sc.n	pv.sdotusp.sc.n
pv.dotsp.sc.m4x2	pv.dotup.sc.m4x2	pv.dotusp.sc.m4x2	pv.sdotsp.sc.m4x2	pv.sdotup.sc.m4x2	pv.sdotusp.sc.m4x2
pv.dotsp.sc.m8x2	pv.dotup.sc.m8x2	pv.dotusp.sc.m8x2	pv.sdotsp.sc.m8x2	pv.sdotup.sc.m8x2	pv.sdotusp.sc.m8x2
pv.dotsp.sc.m8x4	pv.dotup.sc.m8x4	pv.dotusp.sc.m8x4	pv.sdotsp.sc.m8x4	pv.sdotup.sc.m8x4	pv.sdotusp.sc.m8x4
pv.dotsp.sc.m16x8	pv.dotup.sc.m16x8	pv.dotusp.sc.m16x8	pv.sdotsp.sc.m16x8	pv.sdotup.sc.m16x8	pv.sdotusp.sc.m16x8
pv.dotsp.sc.m16x4	pv.dotup.sc.m16x4	pv.dotusp.sc.m16x4	pv.sdotsp.sc.m16x4	pv.sdotup.sc.m16x4	pv.sdotusp.sc.m16x4
pv.dotsp.sc.m16x2	pv.dotup.sc.m16x2	pv.dotusp.sc.m16x2	pv.sdotsp.sc.m16x2	pv.sdotup.sc.m16x2	pv.sdotusp.sc.m16x2
pv.dotsp.sci.h	pv.dotup.sci.h	pv.dotusp.sci.h	pv.sdotsp.sci.h	pv.sdotup.sci.h	pv.sdotusp.sci.h
pv.dotsp.sci.b	pv.dotup.sci.b	pv.dotusp.sci.b	pv.sdotsp.sci.b	pv.sdotup.sci.b	pv.sdotusp.sci.b
pv.dotsp.sci.c	pv.dotup.sci.c	pv.dotusp.sci.c	pv.sdotsp.sci.c	pv.sdotup.sci.c	pv.sdotusp.sci.c
pv.dotsp.sci.n	pv.dotup.sci.n	pv.dotusp.sci.n	pv.sdotsp.sci.n	pv.sdotup.sci.n	pv.sdotusp.sci.n
pv.dotsp.sci.m4x2	pv.dotup.sci.m4x2	pv.dotusp.sci.m4x2	pv.sdotsp.sci.m4x2	pv.sdotup.sci.m4x2	pv.sdotusp.sci.m4x2
pv.dotsp.sci.m8x2	pv.dotup.sci.m8x2	pv.dotusp.sci.m8x2	pv.sdotsp.sci.m8x2	pv.sdotup.sci.m8x2	pv.sdotusp.sci.m8x2
pv.dotsp.sci.m8x4	pv.dotup.sci.m8x4	pv.dotusp.sci.m8x4	pv.sdotsp.sci.m8x4	pv.sdotup.sci.m8x4	pv.sdotusp.sci.m8x4
pv.dotsp.sci.m16x8	pv.dotup.sci.m16x8	pv.dotusp.sci.m16x8	pv.sdotsp.sci.m16x8	pv.sdotup.sci.m16x8	pv.sdotusp.sci.m16x8
pv.dotsp.sci.m16x4	pv.dotup.sci.m16x4	pv.dotusp.sci.m16x4	pv.sdotsp.sci.m16x4	pv.sdotup.sci.m16x4	pv.sdotusp.sci.m16x4
pv.dotsp.sci.m16x2	pv.dotup.sci.m16x2	pv.dotusp.sci.m16x2	pv.sdotsp.sci.m16x2	pv.sdotup.sci.m16x2	pv.sdotusp.sci.m16x2

- Virtual Instructions
  - [illegible] ed at
  - [illegible] ID insn
  - [illegible]
- No e[illegible] ats
- Reus[illegible]

Virtual Instructions

pv.dotsp.v	pv.sdotsp .v
pv.dotsp.sc .v	pv.sdotsp.sc .v
pv.dotsp.sci .v	pv.sdotsp.sci .v
pv.dotup .v	pv.sdotup .v
pv.dotup.sc .v	pv.sdotup.sc .v
pv.dotup.sci .v	pv.sdotup.sci .v
pv.dotusp .v	pv.sdotusp .v
pv.dotusp.sc .v	pv.sdotusp.sc .v
pv.dotusp.sci .v	pv.sdotusp.sci .v

[illegible]ic Bit-[illegible]le Execution

[illegible]10,4(x4!)
[illegible]11,4(x5!)
[illegible]p.v x20,x11,x10

[illegible]ne instruction per type

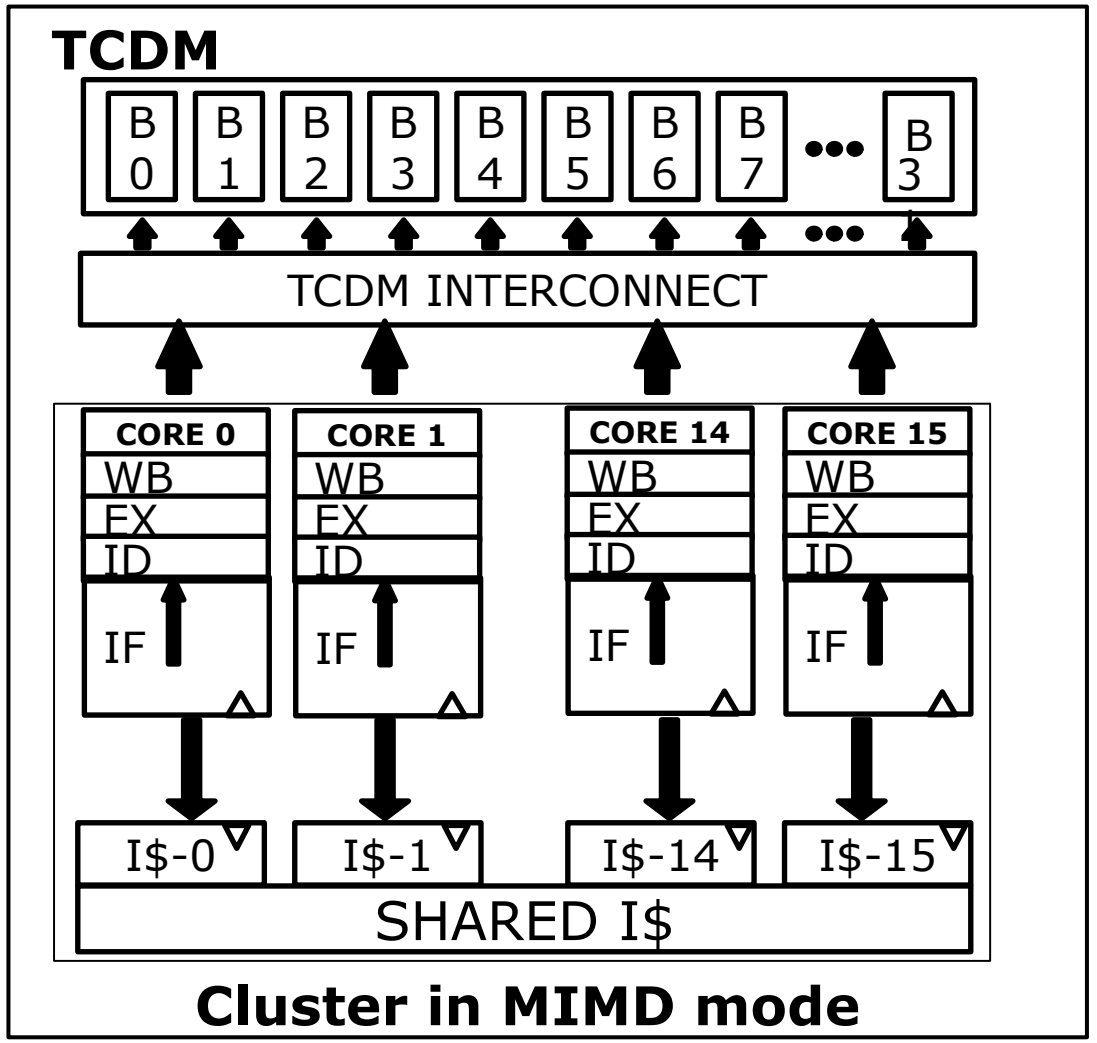
[illegible]ode

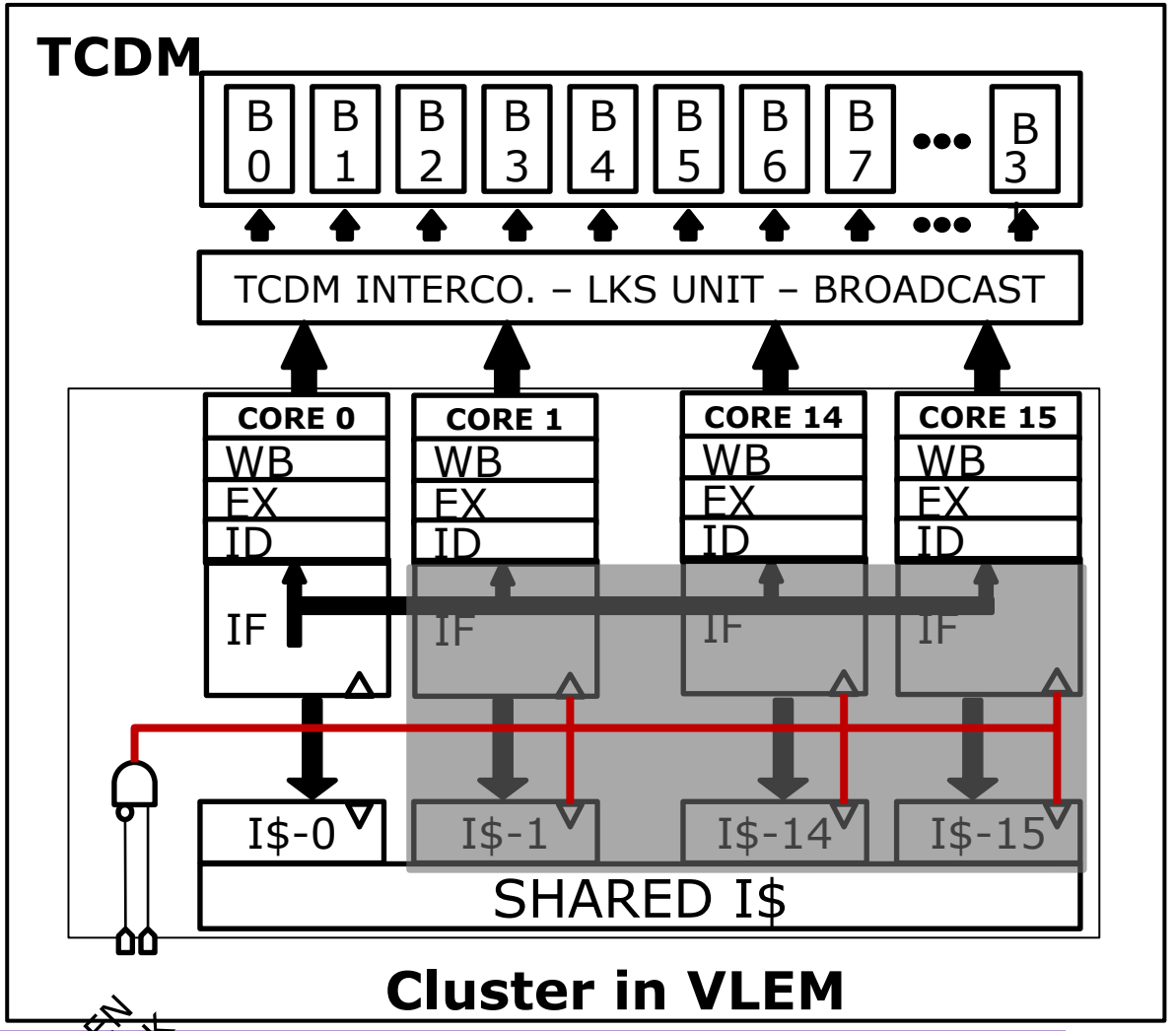
```
int main()
{
....
SIMD_FMT(M8x4);
convolution(A, W, Res);
....
SIMD_FMT(M8x2);
convolution(A, W, Res);
....}
```

[illegible]
pv.packlo.b x15, x5, x6
pv.packhi.b x15, x7, x8
pv.sdotsp.b x20, x15, x10

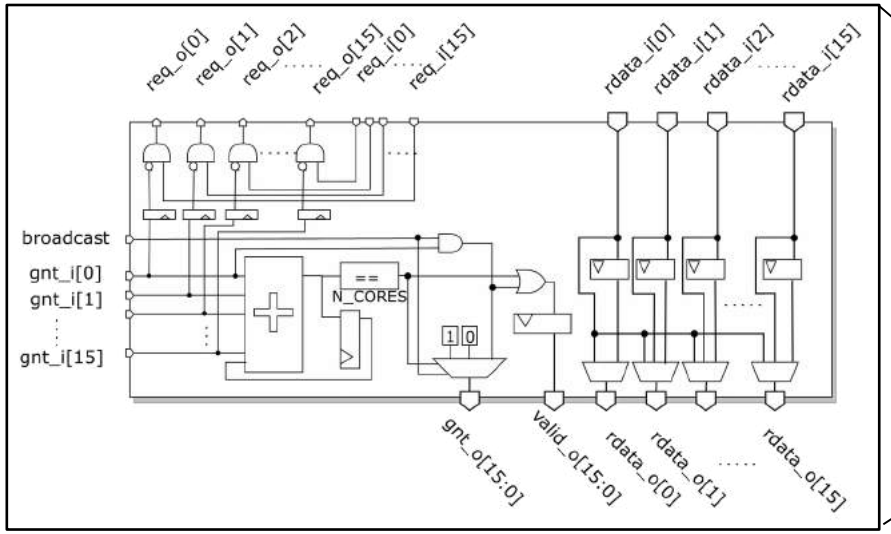
instruction per format and type

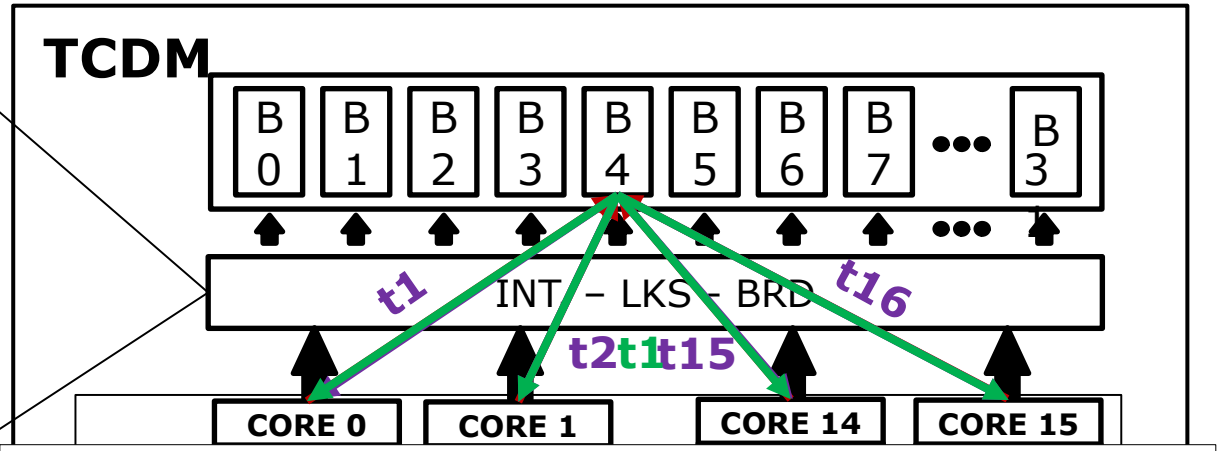
Vector Lockstep Exec. Mode (VLEM)

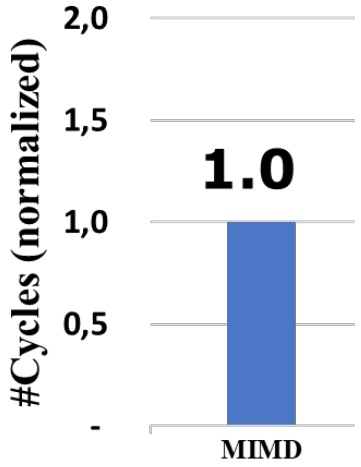




VLEM: Broadcast Unit







MatMul exec. kernels

- **Overhead**: at least 16 clk cycles to unlock the execution in case of concurrent accesses;
- **Solution**:
  - eliminate the overhead in case of access to the same mem address → **BROADCAST UNIT**.
  - Misalign static data and stacks to avoid accesses to the same mem bank;

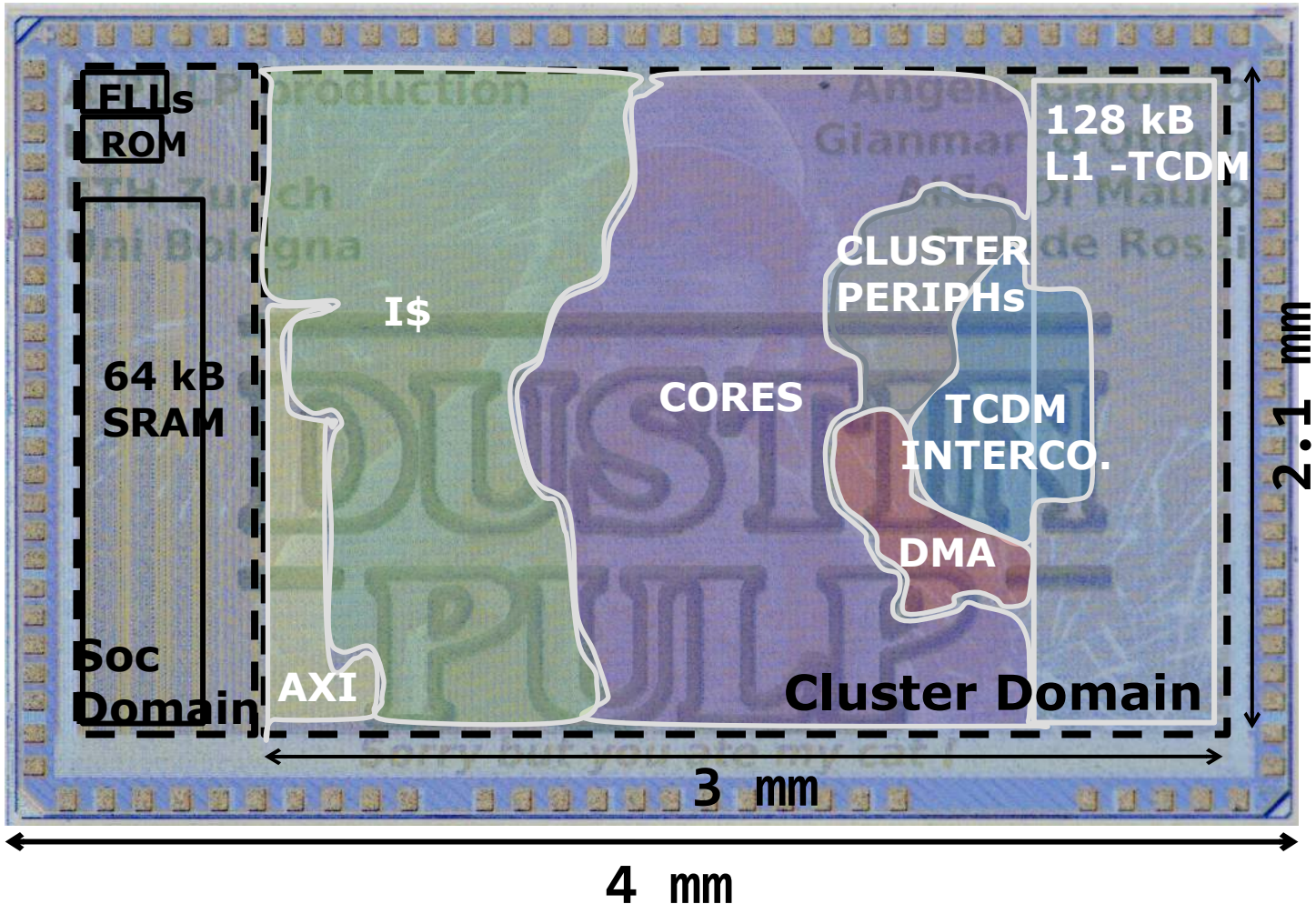
Cluster in VLEM

LKS EN
CLK

Outline

- Introduction & Motivation
- Dustin Architecture Overview
  - Tunable Mixed-Precision Computation
  - Vector Lockstep Execution Mode
- **Chip Results Summary**
- Comparison with the State-of-the-art
- Conclusion

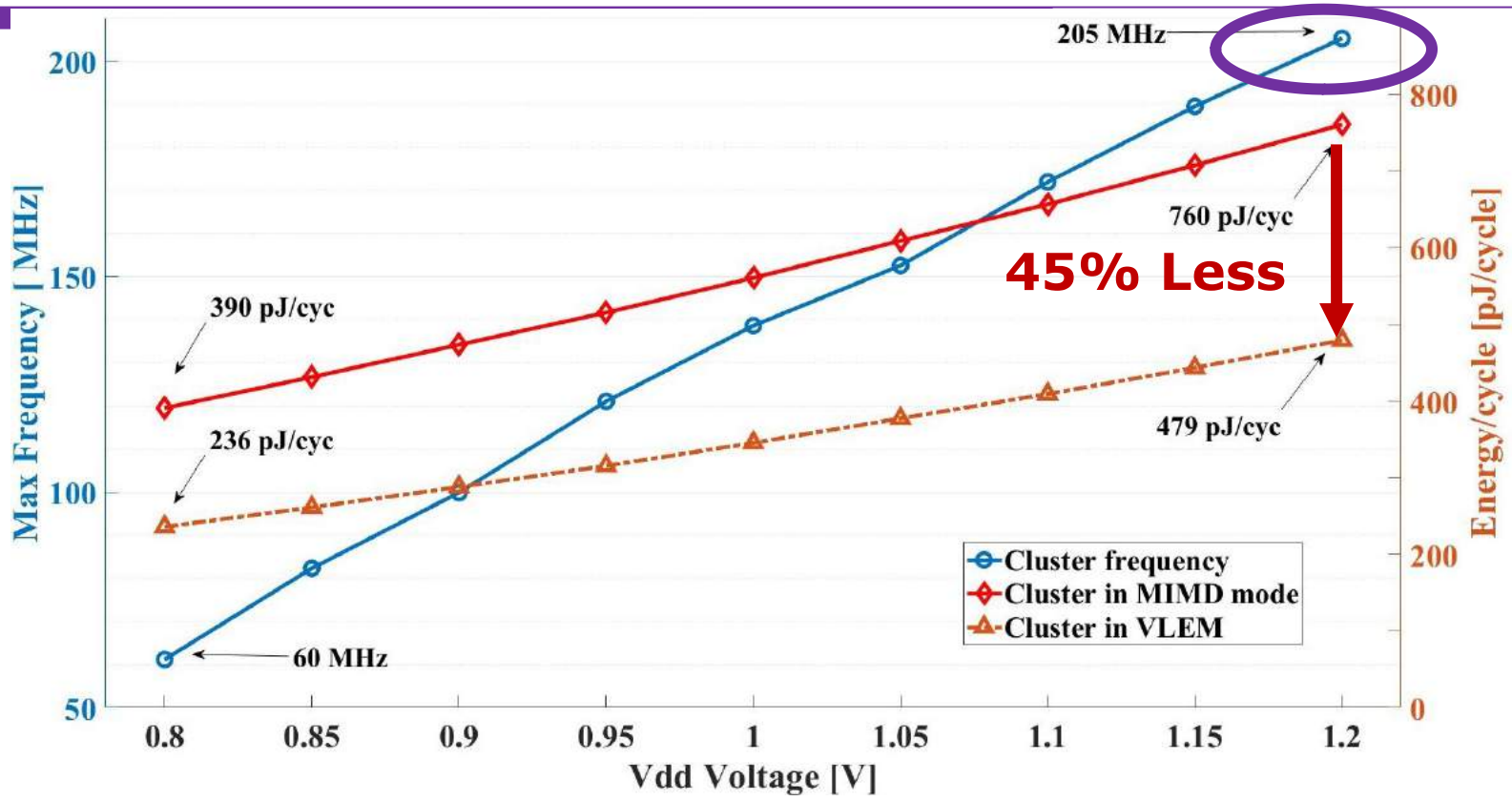
Chip Results Summary



Technology	CMOS TSMC 65nm
Chip area	10 mm²
Total SRAM	208 kB
VDD range	0.8V - 1.2V
Power Envelope	156 mW
Frequency Range	60 – 205 MHz

Integer Performance (8-bit)	15 GOPS
Integer Efficiency (8-bit)	303 GOPS/W
Integer Performance (4-bit)	30 GOPS
Integer Efficiency (4-bit)	562 GOPS/W
Peak Performance	58 GOPS
Peak En. Efficiency	1.15 TOPS/W

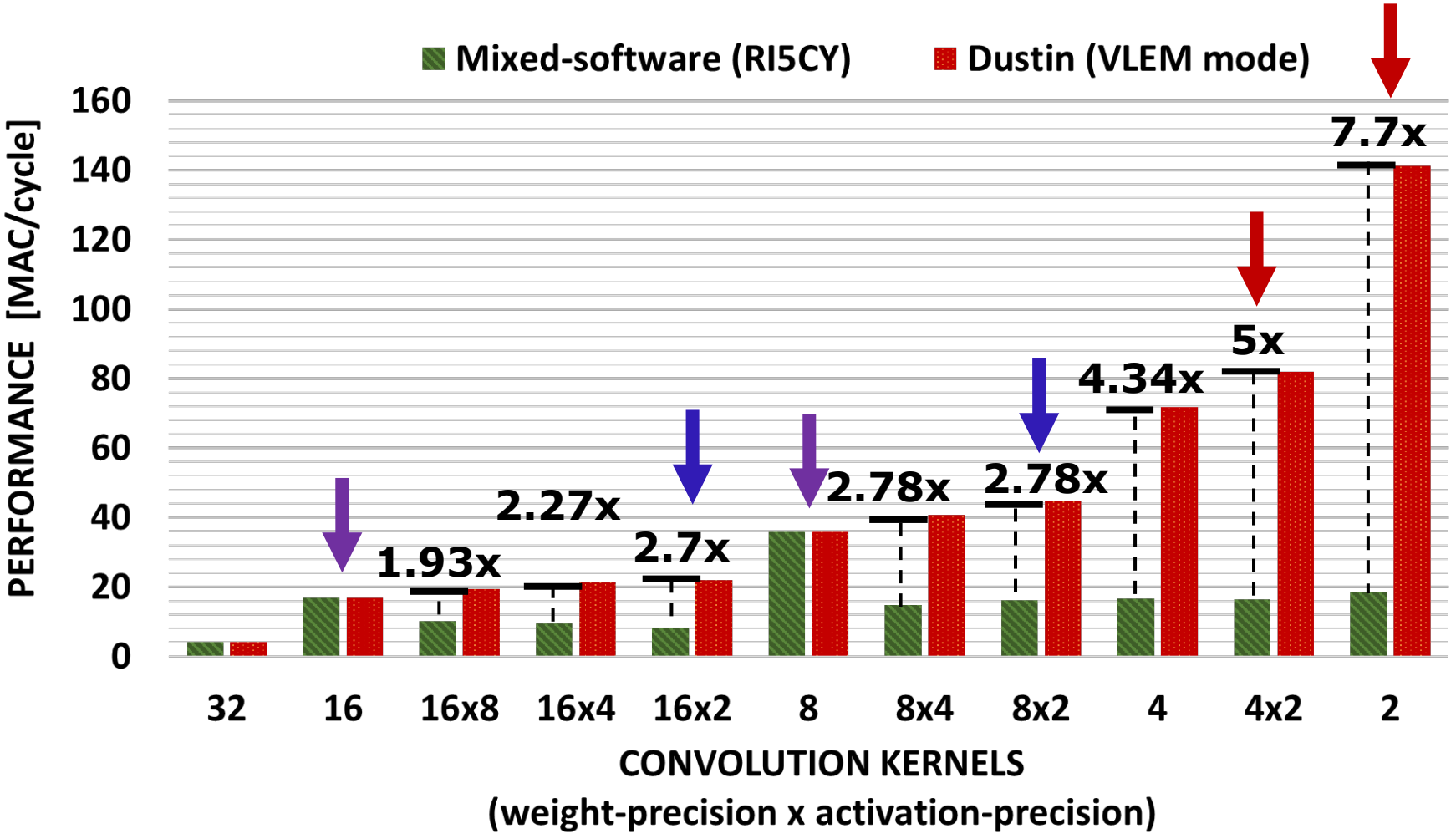
Voltage vs. Frequency



VLEM, no impact on max operating freq. of the cluster

- Measurements of the Cluster;
- Maximum frequency 205 MHz @ 1.2 V;
- **~45% energy saving** with the Cluster in VLEM wrt MIMD mode.

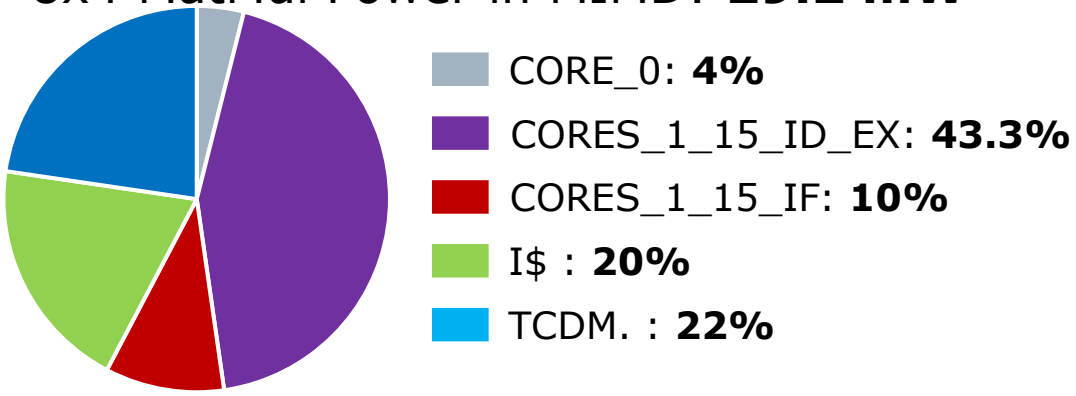
Performance on MatMul kernels



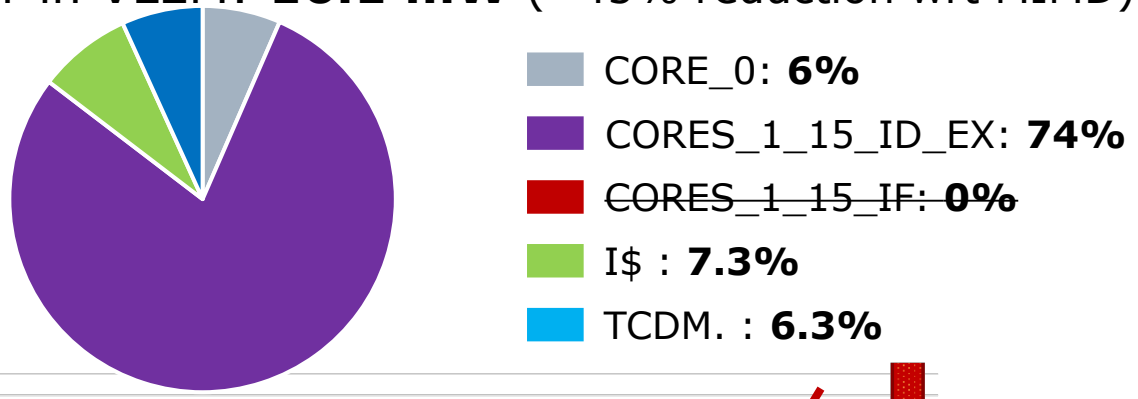
- Benchmarked mixed-precision convolution kernels
- Mixed-precision ISA extensions provide significant advantage: 2x to 7.7x

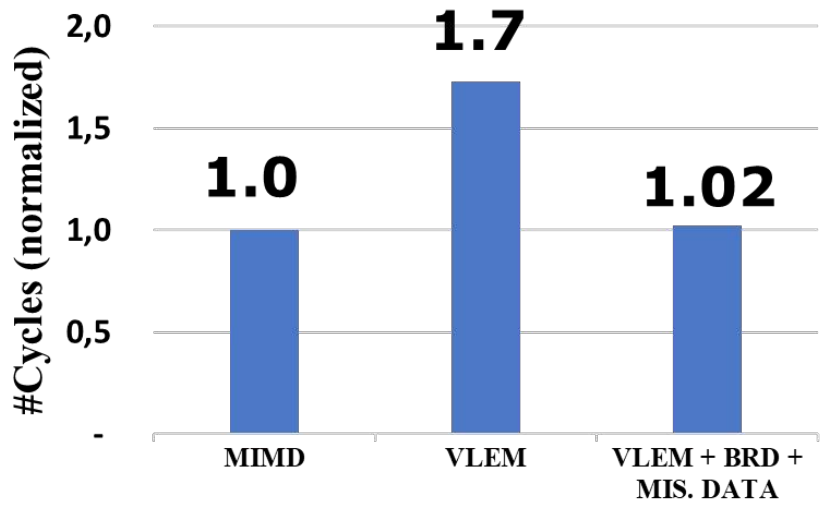
Energy Efficiency on MatMul kernels

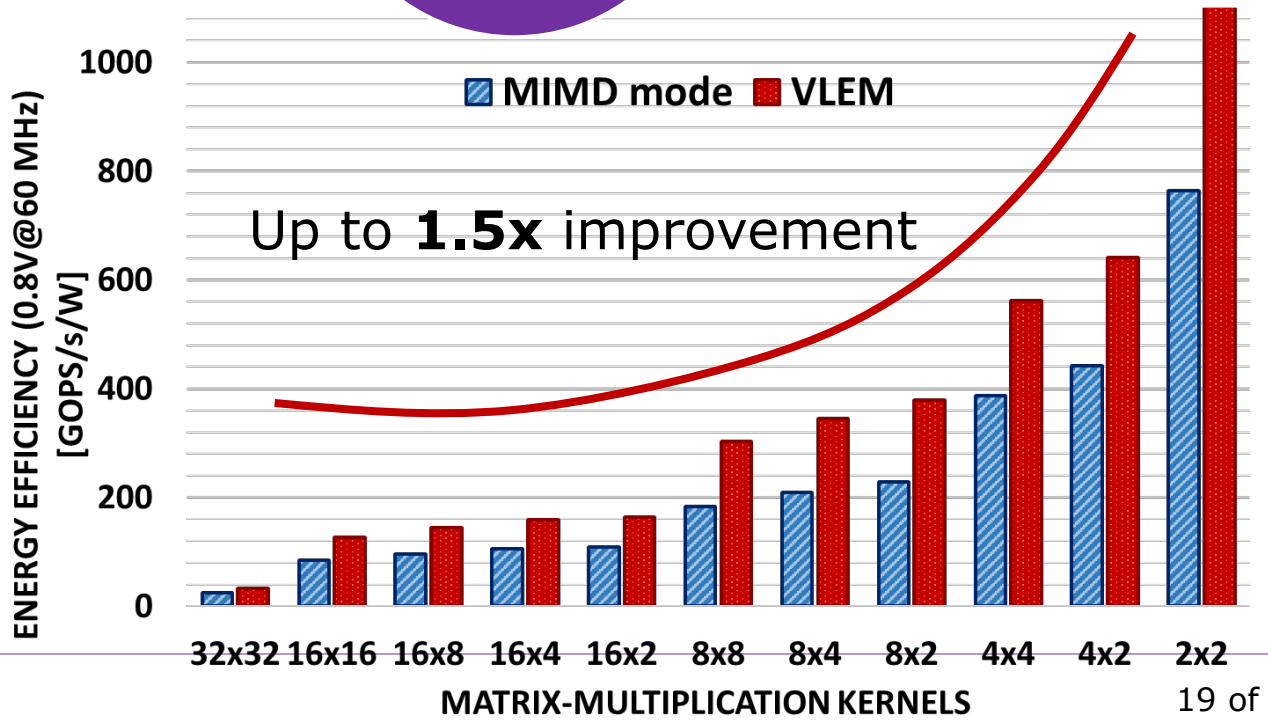
8x4 MatMul Power in MIMD: **29.2 mW**



Power in VLEM: **16.1 mW** (~45% reduction wrt MIMD)







Outline

- Introduction & Motivation
- Dustin Architecture Overview
  - Tunable Mixed-Precision Computation
  - Vector Lockstep Execution Mode
- Chip Results Summary
- **Comparison with the State-of-the-art**
- Conclusion

Comparison with the SoA

	SleepRunner [6]	SamurAI [7]	Mr.Wolf [8]	Vega [9]	**Dustin (this work)**
Technology	CMOS 28nm FDSOI	CMOS 28nm FDSOI	CMOS 40nm LP	CMOS 22nm FDSOI	CMOS 65nm
Die Area	0.68 mm^2	4.5 mm^2	10 mm^2	12 mm^2	10 mm^2
Applications	IoT GP	IoT GP + DNN	IoT GP + DNN	IoT GP + NSA+DNN	IoT GP + DNN + QNNs
CPU/ISA	CM0DS Thumb-2 subset	1x RI5CY RVC32IMFXpulp	9 x RI5CY RVC32IMFXpulp	10 x RI5CY RVC32IMFXpulp+SF	16 x MPIC CORES (RISC-V)
Int Precision (bits)	32	8, 16, 32	8, 16, 32	8, 16, 32	2, 4, 8, 16, 32 (plus Mixed-Precision)
Supply Voltage	0.4 - 0.8 V	0.45 - 0.9 V	0.8 - 1.1 V	0.5 – 0.8 V	0.8 - 1.2 V
Max Frequency	80 MHz	350 MHz	450 MHz	450 MHz	205 MHz
Power Envelope	320 µW	96 mW	153 mW	49.4 mW	156 mW
[1]Best Integer Performance	31 MOPS (32b)	1.5 GOPS (8b) [2]	12.1 GOPS (8b)	15.6 GOPS (8b)	**15 GOPS (8b)** **30 GOPS (4b)** **58 GOPS (2b)**
[1]Best Integer Efficiency	97 MOPS/mW @ 18.6 MOPS (32b)	230 GOPS/W @110 MOPS (8b) [2]	190 GOPS/W @ 3.8 GOPS (8b)	614 GOPS/W @ 7.6 GOPS	**303 GOPS/W @4.4 GOPS (8b)** **570 GOPS/W @8.8 GOPS (4b)** **1152 GOPS/W@17.3 GOPS(2b)**

Dustin supports Mixed-Precision computation in HW

Better efficiency wrt solutions in 28nm and 40nm tech node

Comparable efficiency wrt Vega (22 nm)

[1] 2 OPs = 1 8-bit (or 4-bit or 2-bit) MAC on MatMul benchmark unless differently specified.

[2] For fair comparison we consider the execution on software programmable cores.

Outline

- Introduction & Motivation
- Dustin Architecture Overview
  - Tunable Mixed-Precision Computation
  - Vector Lockstep Execution Mode
- Chip Results Summary
- Comparison with the State-of-the-art
- **Conclusion**

Conclusion

- Dustin SoC: IoT end-node with AI computing capabilities in **tsmc 65 nm** tech node;
- RISC-V cores featuring 2b-to-32b bit-precision instruction set architecture (ISA) extensions enabling fine-grain tunable mixed-precision computation (**2x** to **7x** speed-up w.r.t. RI5CY);
- Software reconfigurable cluster in Vector Lockstep Execution Mode (**~45% energy saving** w.r.t. MIMD mode);
- Dustin is competitive with IoT end-nodes using much more scaled technology nodes (Peak Perf. **58 GOPS**, Peak Eff. **1.15 TOPS/W**).
- Despite a less scaled tech node, we reach energy efficiency in the order of **TOPS/W** → Comparable with ASIC solutions.