ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

PULP: A Parallel Ultra Low Power platform for next generation IoT Applications

Davide Rossi[1]

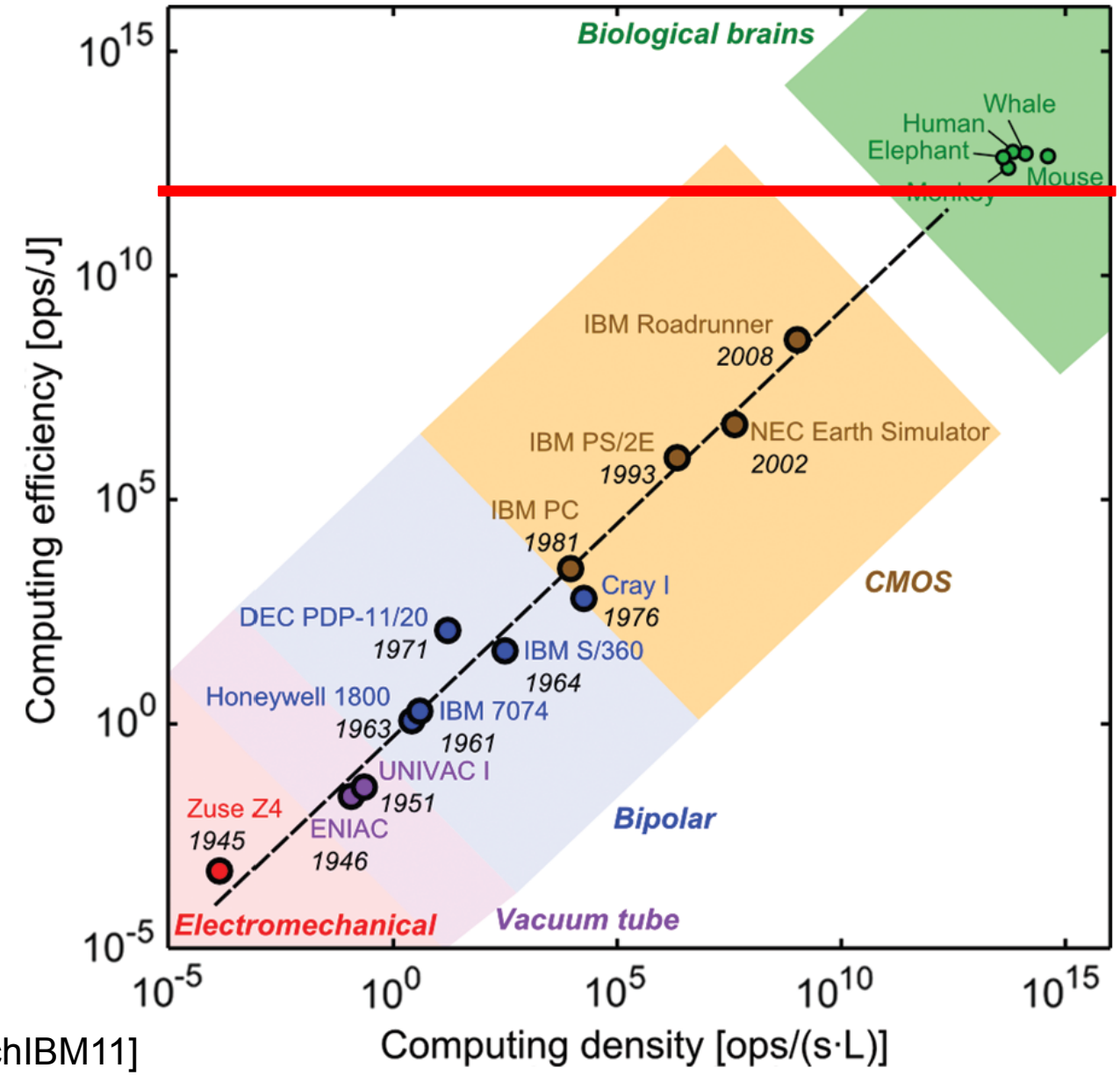
Francesco Conti[1], Andrea Marongiu[1,2], Antonio Pullini[2], Igor Loi[1], Michael Gautschi[2], Giuseppe Tagliavini[1], Alessandro Capotondi[1], Philippe Flatresse[3], Luca Benini[1,2]

[1]DEI-UNIBO, [2]IIS-ETHZ, [3]STMicroelectroncis

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

How efficient do we need to be?

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA



10^{12}ops/J

↓

1pJ/op

↓

1GOPS/mW

[*RuchIBM11]

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

System View

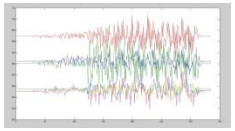
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

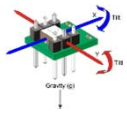
Sense

Analyze and Classify

Transmit

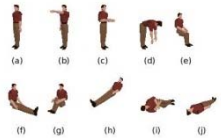
MEMS IMU



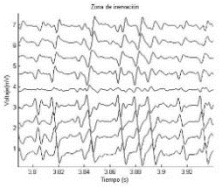


MEMS Microphone

ULP Imager



EMG/ECG/EIT



µController

L2 Memory

IOs

ULP Parallel Processor

1 ÷ 2000 MOPS
1 ÷ 10 mW

Short range, BW

Low rate (periodic) data

SW update, commands

Long range, low BW

LoRa SEMTECH

100 µW ÷ 2 mW

Battery + Harvesting powered
→ a few mW power envelope

Idle: ~1µW
Active: ~ 50mW

Near-Sensor Processing

	INPUT BANDWIDTH	COMPUTATIONAL DEMAND	OUTPUT BANDWIDTH	COMPRESSION FACTOR
Image				
Tracking: [*Lagroce2014]	80 Kbps	1.34 GOPS	0.16 Kbps	500x
Voice/Sound				
Speech: [*VoiceControl]	256 Kbps	100 MOPS	0.02 Kbps	12800x
Inertial				
Kalman: [*Nilsson2014]	2.4 Kbps	7.7 MOPS	0.02 Kbps	120x
Biometrics				
SVM: [*Benatti2014]	16 Kbps	150 MOPS	0.08 Kbps	200x

- ***Extremely compact output (single index, alarm, signature)***
- ***Computational power of ULP µControllers is not enough***
- ***Parallel worloads***

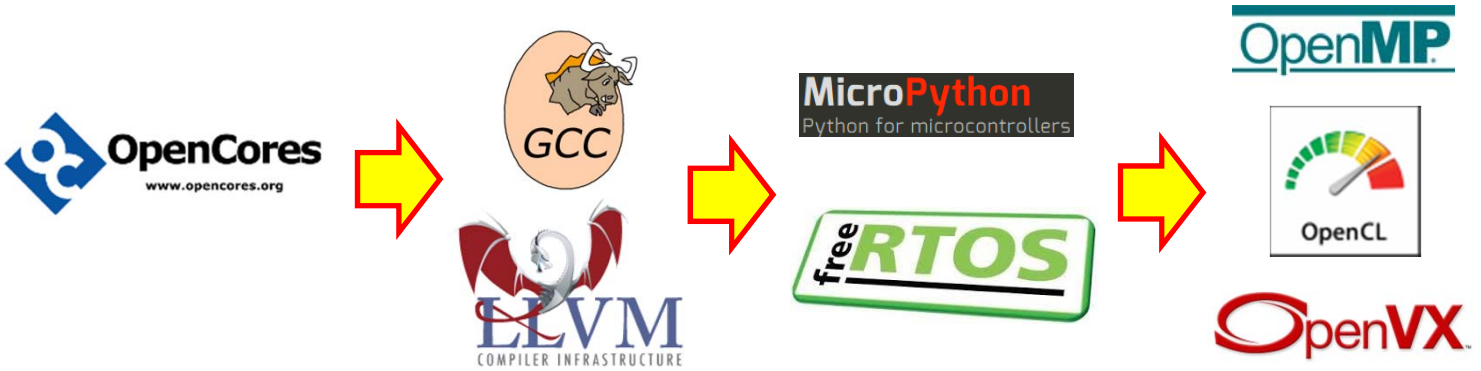
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

PULP:
pJ/op Parallel ULP computing

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

pJ/op is traditionally the target of ASIC + µControllers

- Scalable: to many-core + heterogeneity
- Best-in-class LP silicon technology
- Programmable: OpenMP, OpenCL, OpenVX
- Open: Software & HW



Processor & Hardware IPs	Compiler Infrastructure	Virtualization Layer	Programming Model

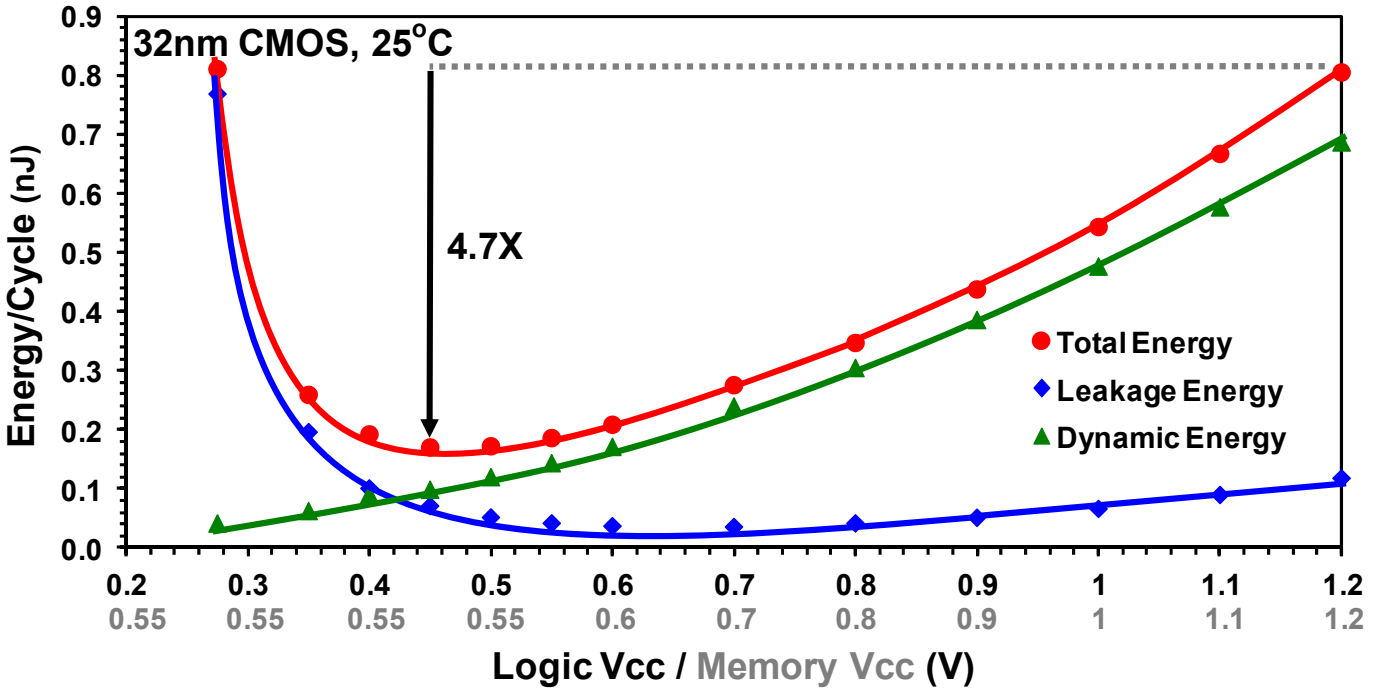
From ULP computing to parallel + heterogeneous ULP computing
1mW-10mW active power

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Near-Threshold Multiprocessing

Minimum Energy Operation

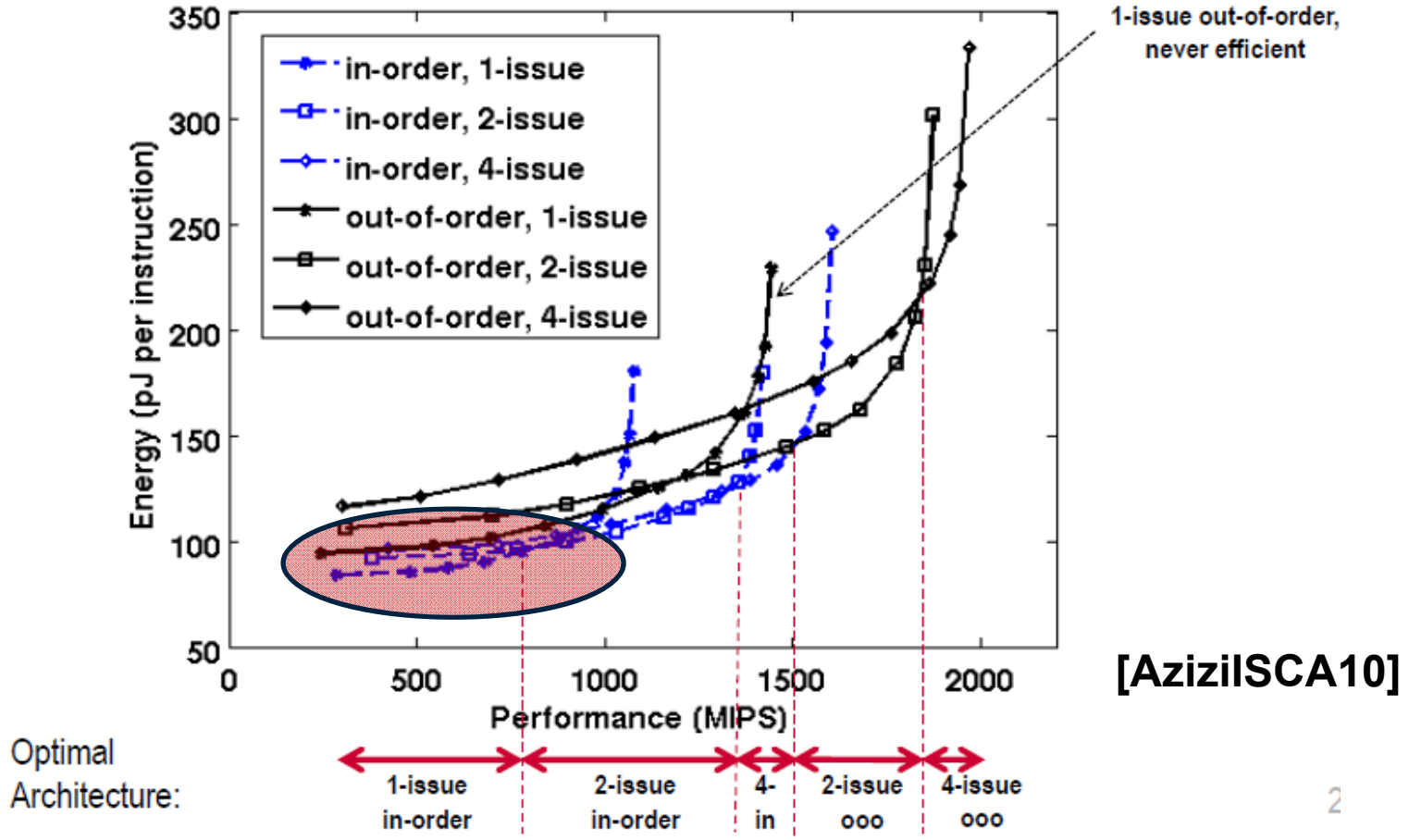


[VivekDeDATE2013]

Near-Threshold Computing (NTC):

1. Don't waste energy pushing devices in strong inversion
2. Recover performance with parallel execution
3. Aggressively manage idle power (switching, leakage)

The "best" Processor



[AzizilSCA10]

- **Single issue in-order is most energy efficient**
- **Put more than one + shared memory to fill cluster area**

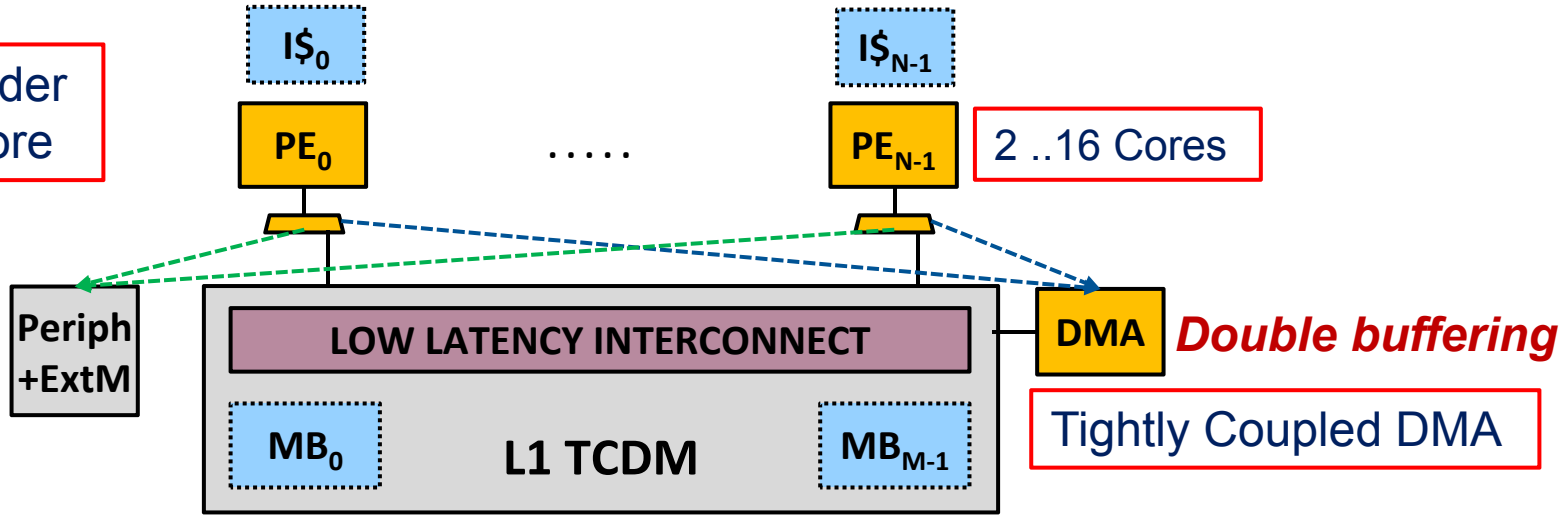
Building PULP

SIMD + MIMD + sequential

Private per-core instruction cache

4-stage, in-order OpenRISC core

DEMUX



Double buffering

Tightly Coupled DMA

1 Cycle Shared Multi-Banked L1 Data Memory
+ Low Latency Interconnect

"GPU like" shared memory → low overhead data sharing

Near Threshold but parallel → Maximum Energy efficiency when Active

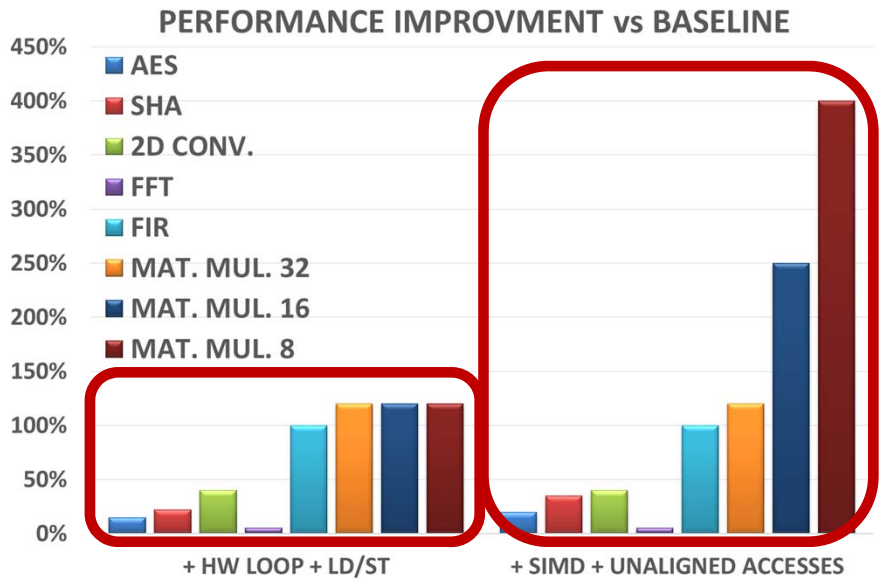
+ strong power management for (partial) idleness

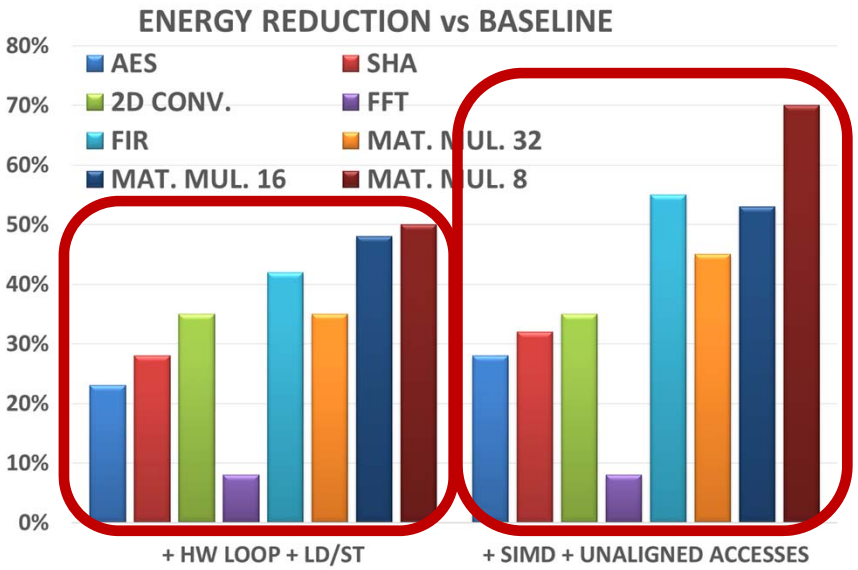
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

OR1ON: Extended OpenRISC Core

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

- 4-stage OpenRISC
- IPC ~ 1
- DSP extensions:
  - Hardware loops
    - Eliminates branching overhead
  - LD/ST + post-increment
    - Enhanced vector indexing
  - Small vector support (SIMD)
    - 2x 16-bit operations
    - 4x 8-bit operations
  - Unaligned memory accesses
    - To better exploit SIMD





UP TO 5x performance improvement and 3x reduction of energy!!!

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

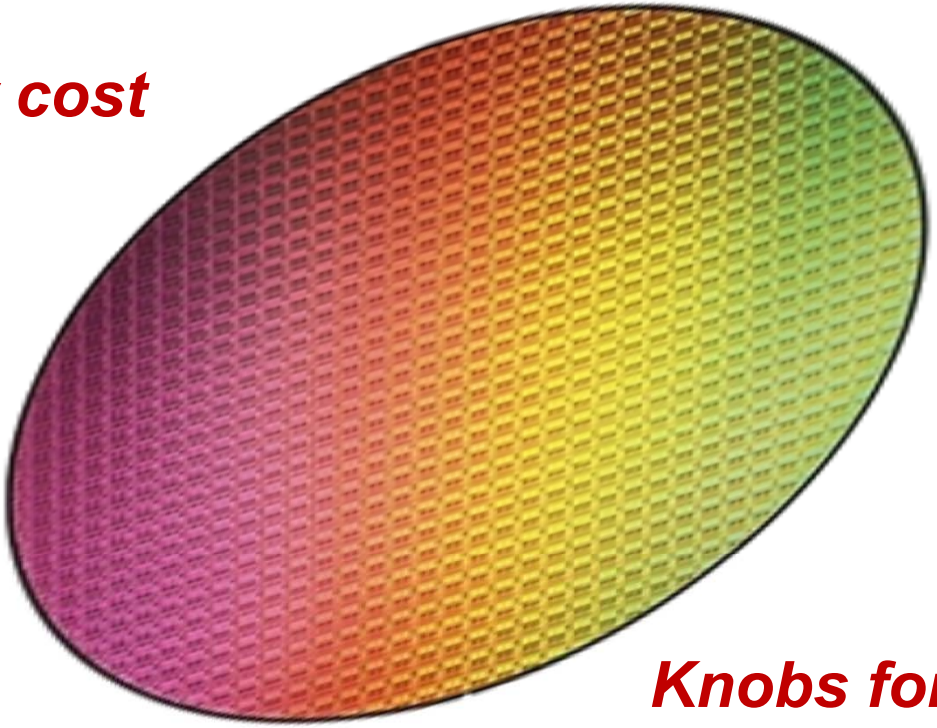
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Silicon Implementation

Technology For ULP

Low VDDMIN

Low cost



High ION/IOFF @ LowVDD

Knobs for variability management

Knobs for power management

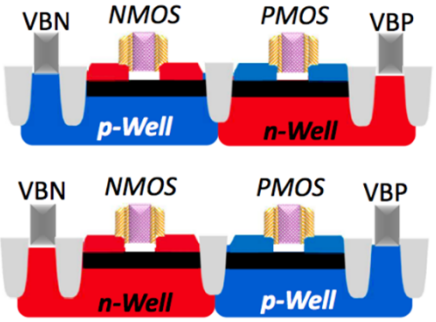
UTBB FD-SOI provides good features for ULP design:

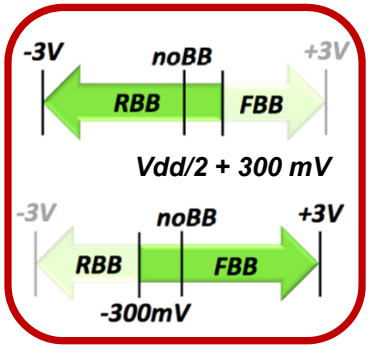
- ***Good behavior at low voltage***
- ***Body bias for power and variability management***

Body biasing with UTBB FD-SOI technology

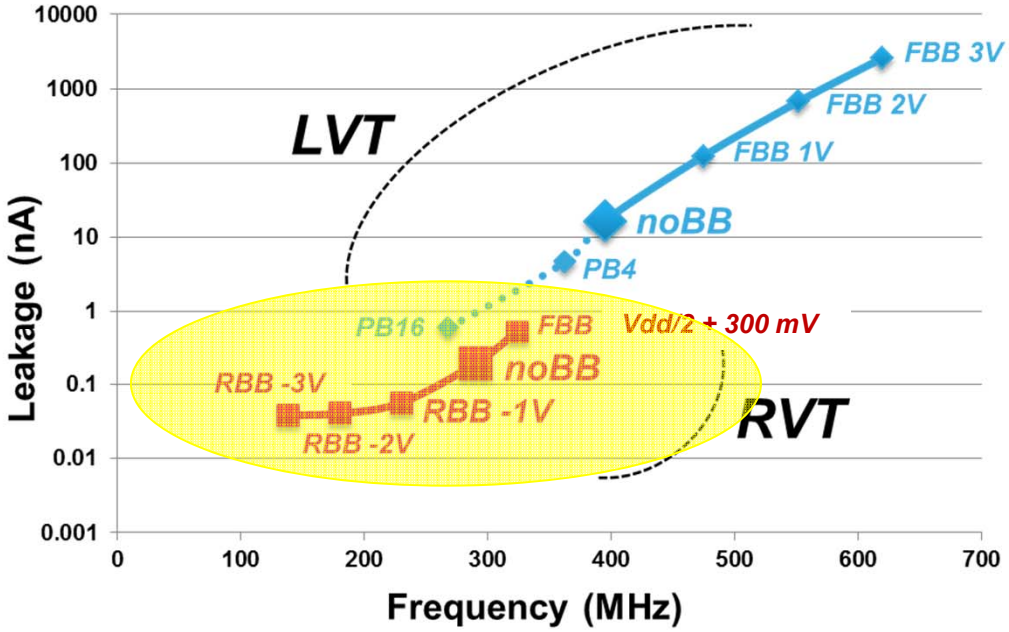
RVT transistor (conventional-well)

LVT transistor (flip-well)





BODY BIAS WINDOWS



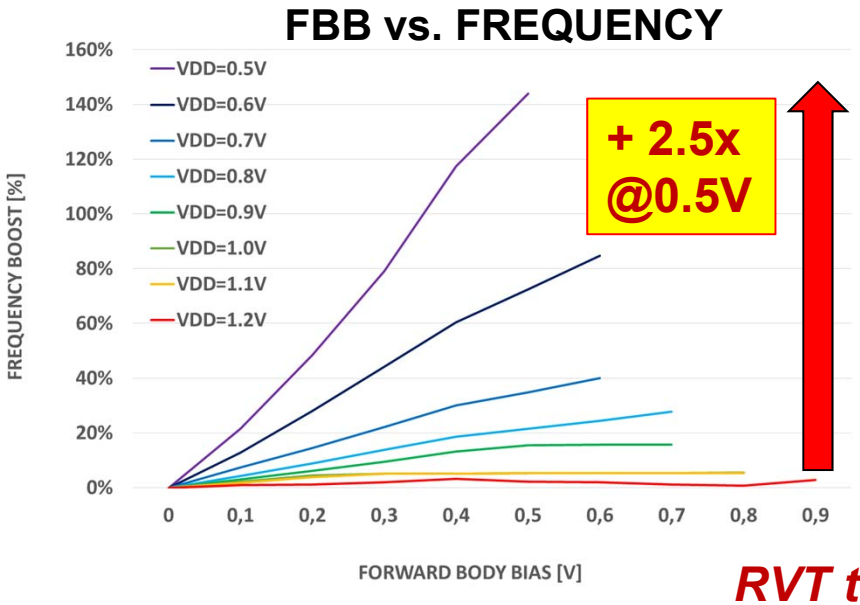
RVT: Regular Voltage Threshold
LVT: Low Voltage Threshold

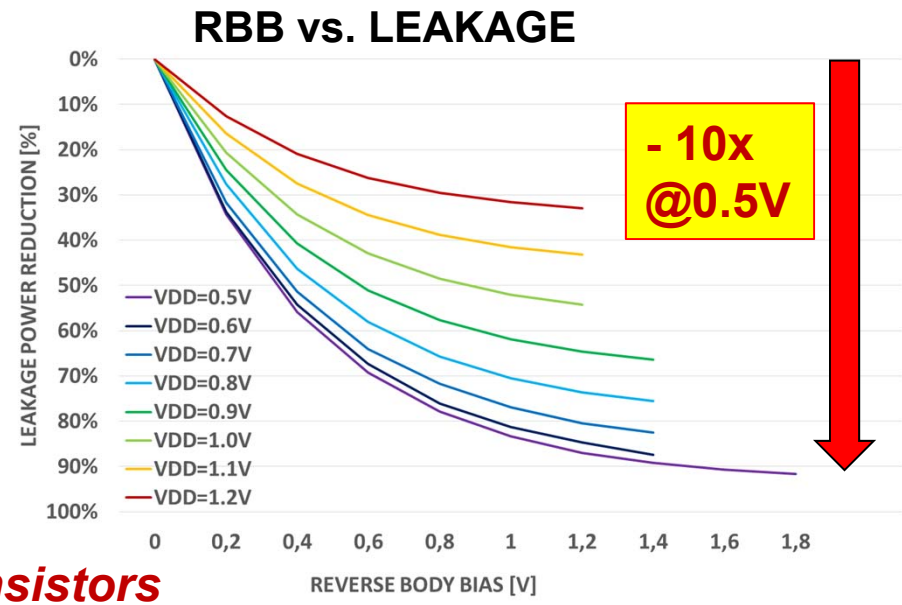
FBB: Forward Body Bias
RBB: Reverse Body Bias

Poly biasing allow to trade performance/leakage At design time

RVT transistors: low leakage + flexible power management (FBB + RBB)

Near Threshold + Body Biasing Combined





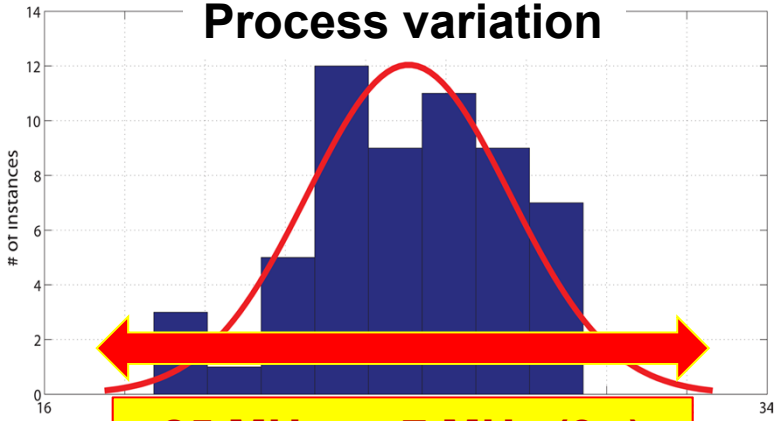
RVT transistors

- State retentive (no state retentive registers and memories)
- Ultra-fast transitions (tens of ns depending on n-well area to bias)
- Low area overhead for isolation (3μm spacing for deep n-well isolation)
- Thin grids for voltage distribution (small transient current for wells polarization)
- Simple circuits for on-chip VBB generation (e.g. charge pump)

But even with aggressive RBB leakage is not zero!

Body Biasing for Variability Management

Process variation

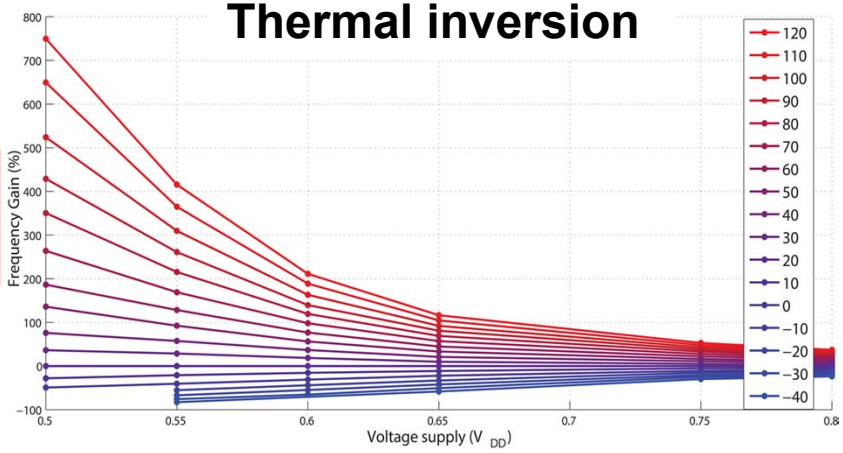


120° C

100x @0.5V

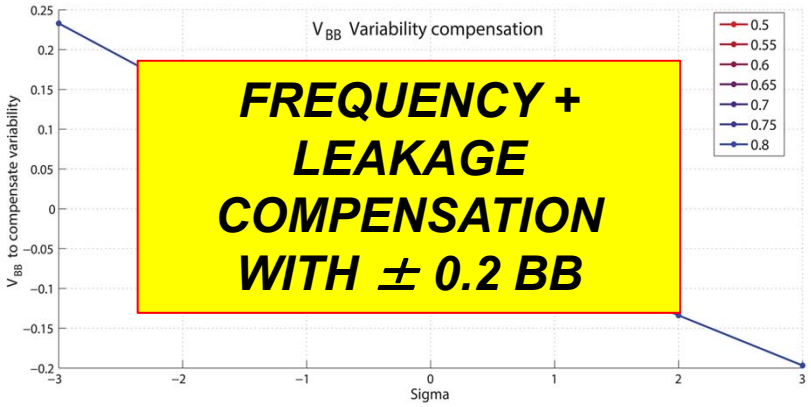
-40° C

Thermal inversion

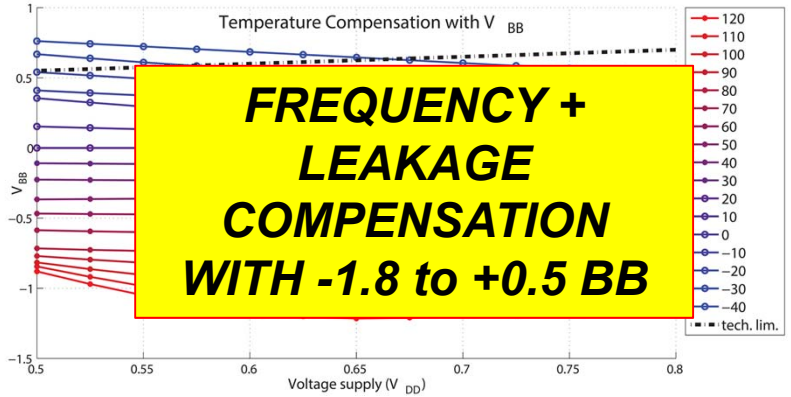


RVT transistors

FBB/RBB



FREQUENCY + LEAKAGE COMPENSATION WITH ± 0.2 BB

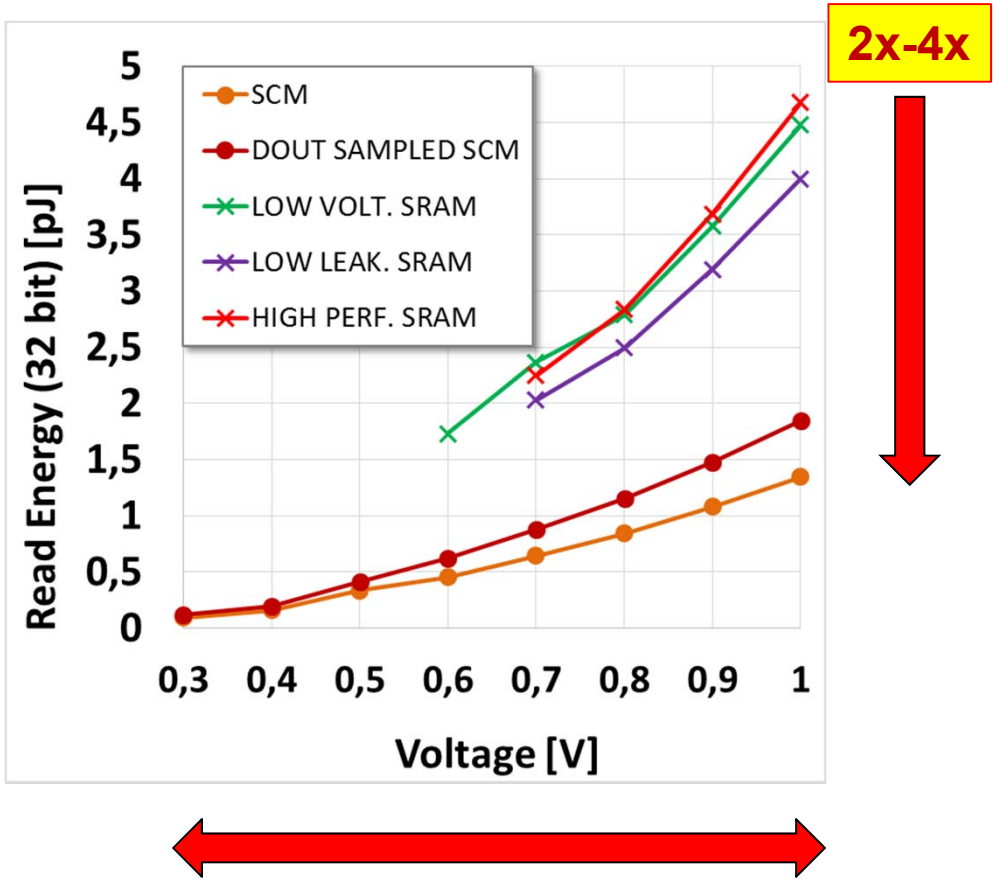


FREQUENCY + LEAKAGE COMPENSATION WITH -1.8 to +0.5 BB

ULP memory implementation: latch-based SCM

- "Standard" 6T SRAMs:
  - High VDDMIN
  - Bottleneck for energy efficiency
- Near-Threshold SRAMs (8T)
  - Lower VDDMIN
  - Area/timing overhead (25%-50%)
  - High active energy
  - Low technology portability
- Standard Cell Memories:
  - Wide supply voltage range
  - Lower read/write energy (2x - 4x)
  - Easy technology portability
  - Controlled P&R mitigates area overhead

256x32 6T SRAMS vs. SCM



ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

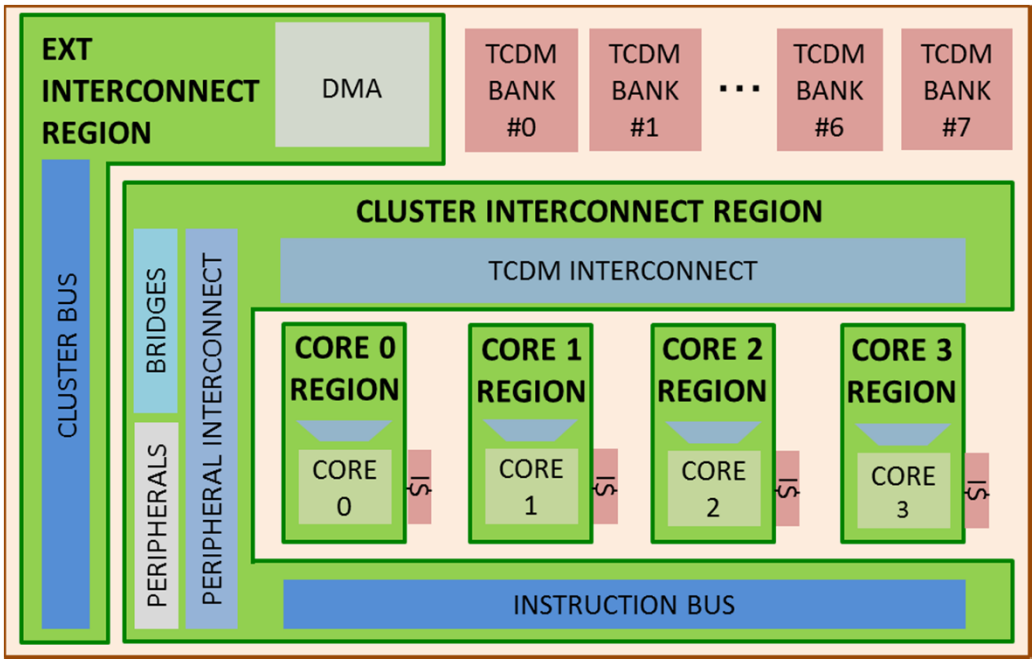
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

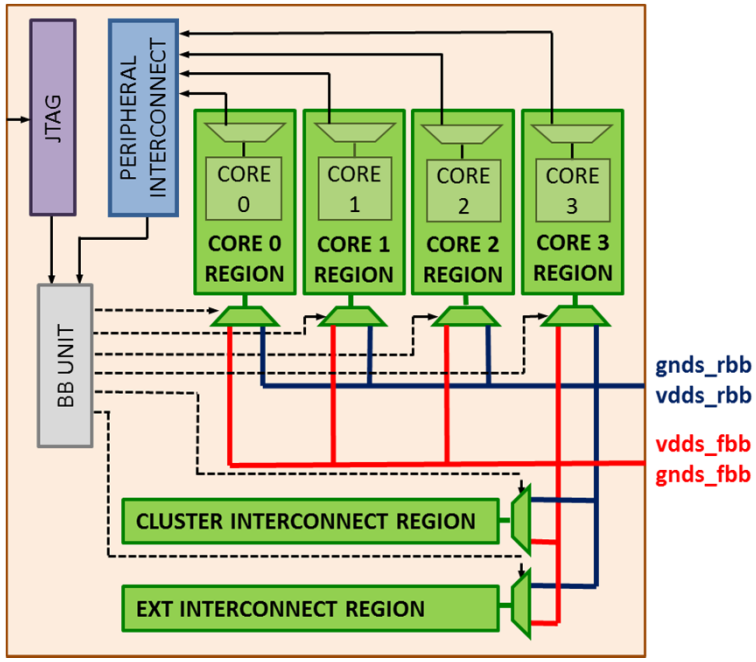
Architectural Technology Awareness

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Exploiting body biasing

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA





- The cluster is partitioned in separate clock gating and body bias regions
- Body bias multiplexers (BBMUXes) control the well voltages of each region
- Each region can be **active** (FBB) or **idle** (deep RBB → low leakage!)

State-Retentive + Low Leakage + Fast transitions

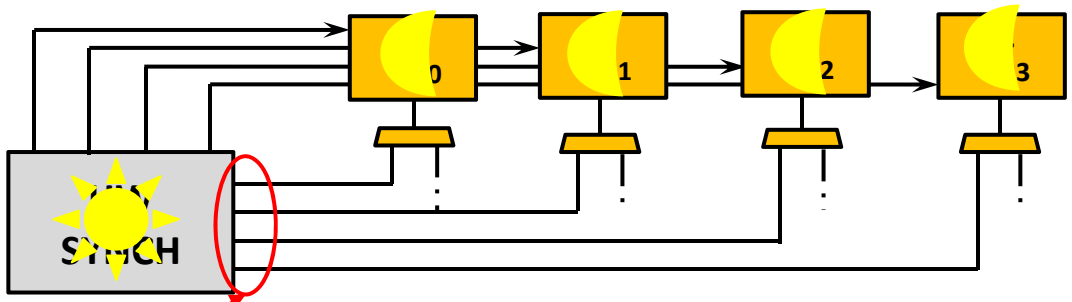
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Power Management: Hardware Synchronization

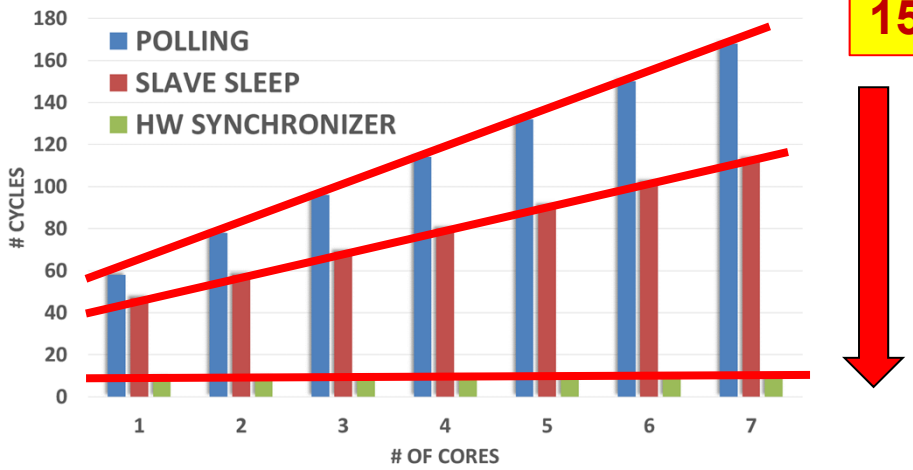
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Core shut-down sequence:
1) Disable fetching
2) Wait outstanding transactions
2) Clock gating
3) Reverse Body Biasing

Private, per core port
→single cycle latency
→no contention



BARRIER COST

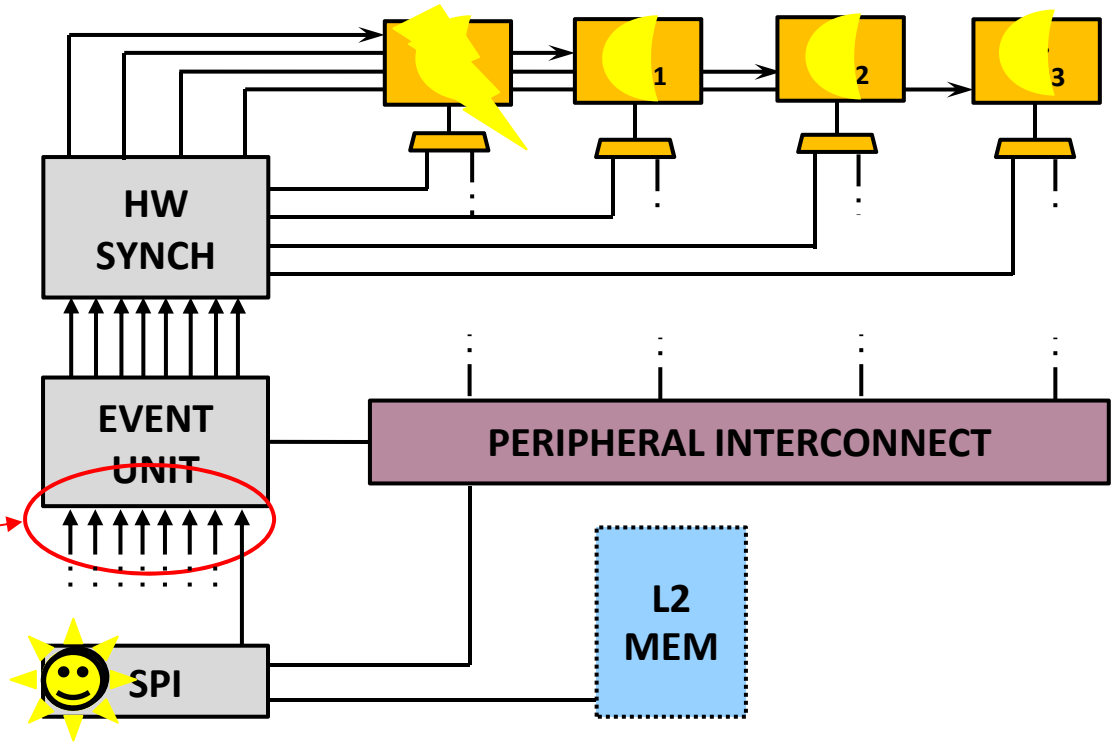


15x

GOALS:
→ Reduce parallelization overhead
→ Accelerate common OpenMP and OpenCL patterns (e.g. Task creation)
→ Automatically manage shut down of idle cores

Power Management: External Events

Programming sequence:
1) Set events mask
2) Program transfer
3) Trigger transfer
4) Shut down cores

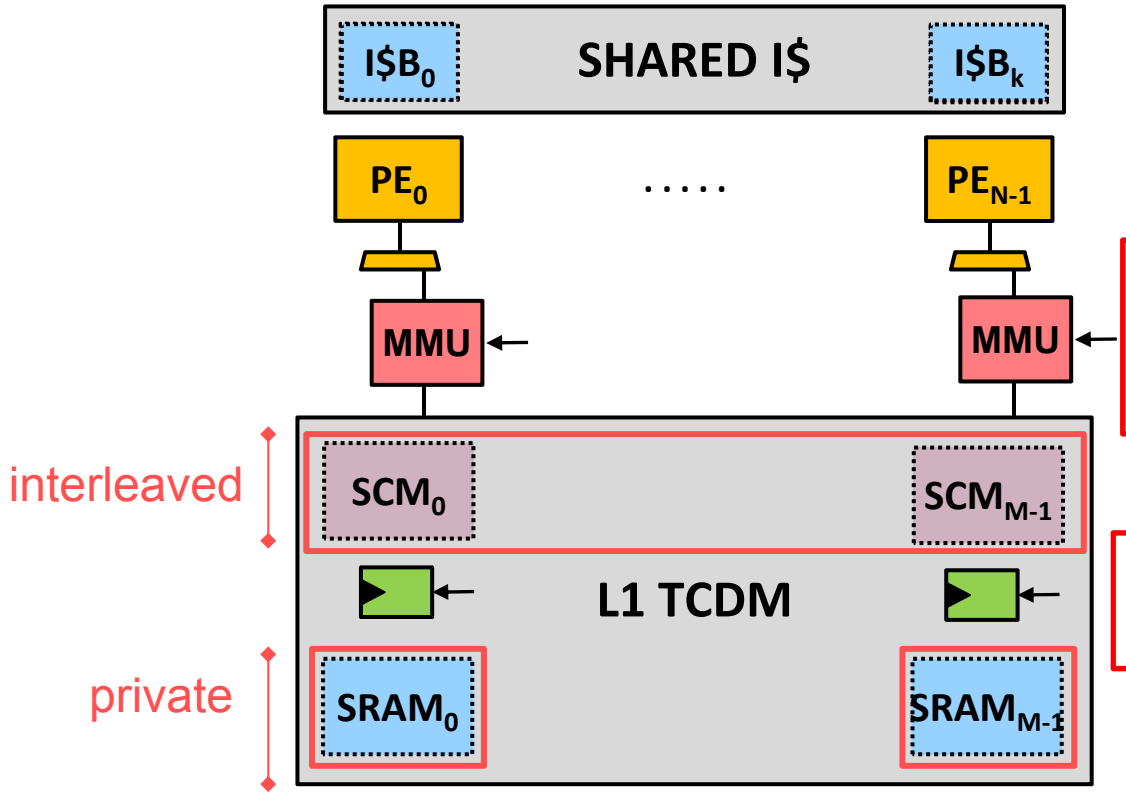


48 maskable events
→General purpose
→DMA
→Timers
→Peripherals (SPI, I2C. GPIO...)

GOALS:
→ Automatically manage shut down of cores during data transfers

Heterogeneous Memory Architecture + Management

Shared I$ to recover SCMs area overhead
Private L0 buffers to reduce pressure on shared I$



MMU (logical/physical add map):
1) Interleaved/private addresses
2) Shutdown of SRAM banks

Reconfigurable Pipeline Stages
for SRAMs degradation@low VDD

SCM on I$ and part of TCDM
to widen the operating range

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

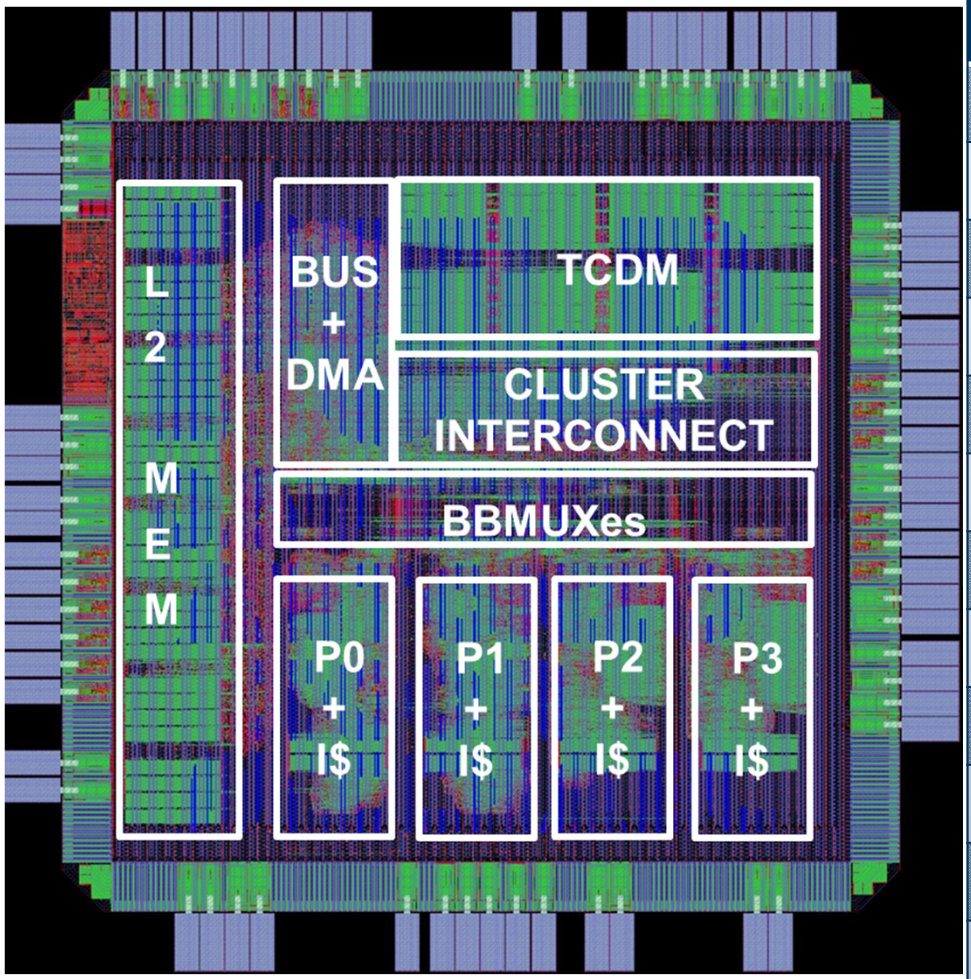
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

The PULP “Family”

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

PULPv1

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

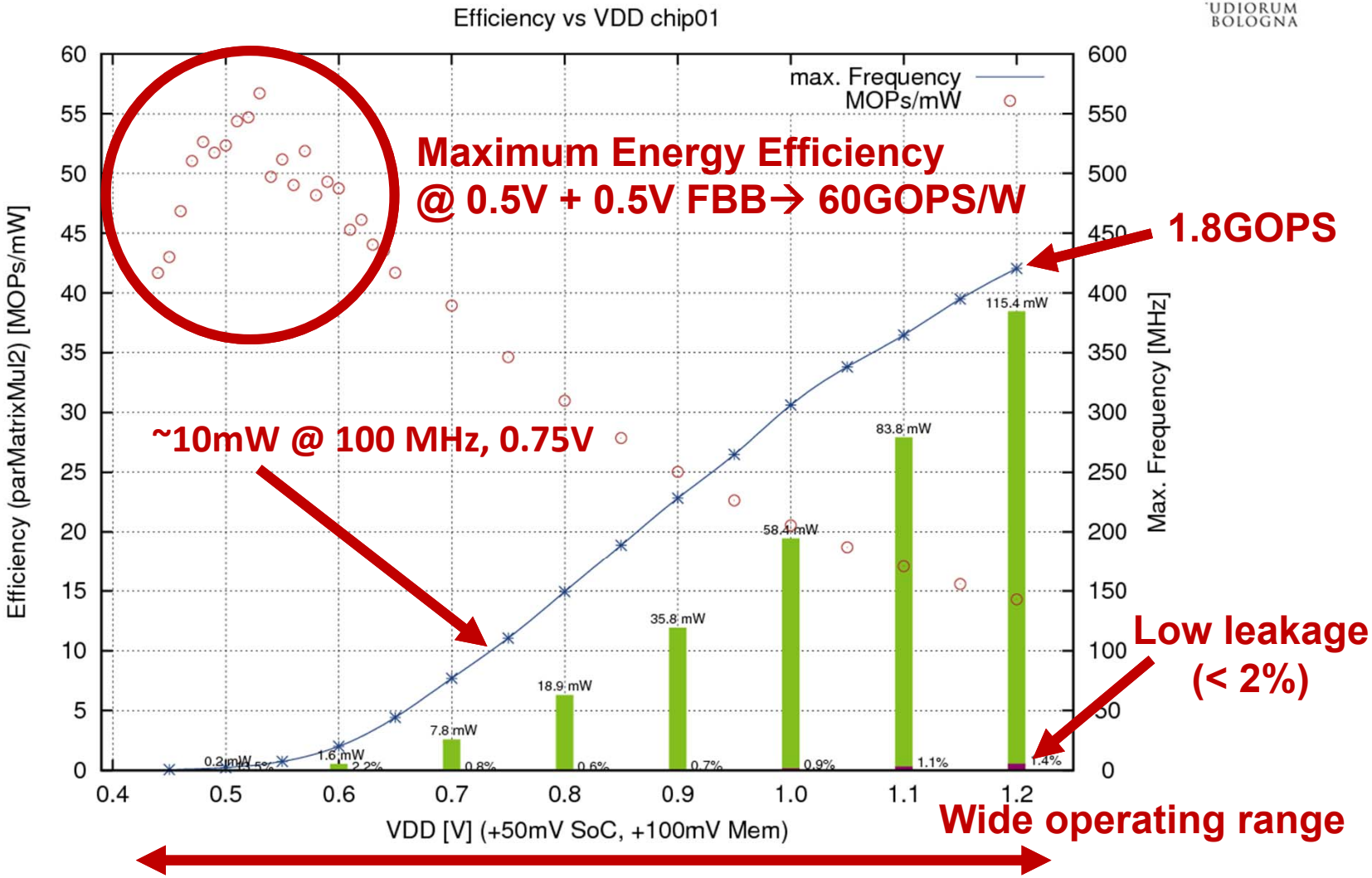


CHIP FEATURES	
Technology	28nm FDSOI (RVT)
Chip Area	3mm^2
# Cores	4xOpenRISC
I$	4x1kbyte (private)
TCDM	16 kbyte
L2	16 kbyte
BB regions	6
VDD range	0.45-1.2V
VBB range	-1.8V - +0.9V
Perf. Range	1 MOPS-1.9GOPS
Power Range	100 µW -127 mW*
Peak Efficiency	60 GOPS/W@0.5V*

Tester chip

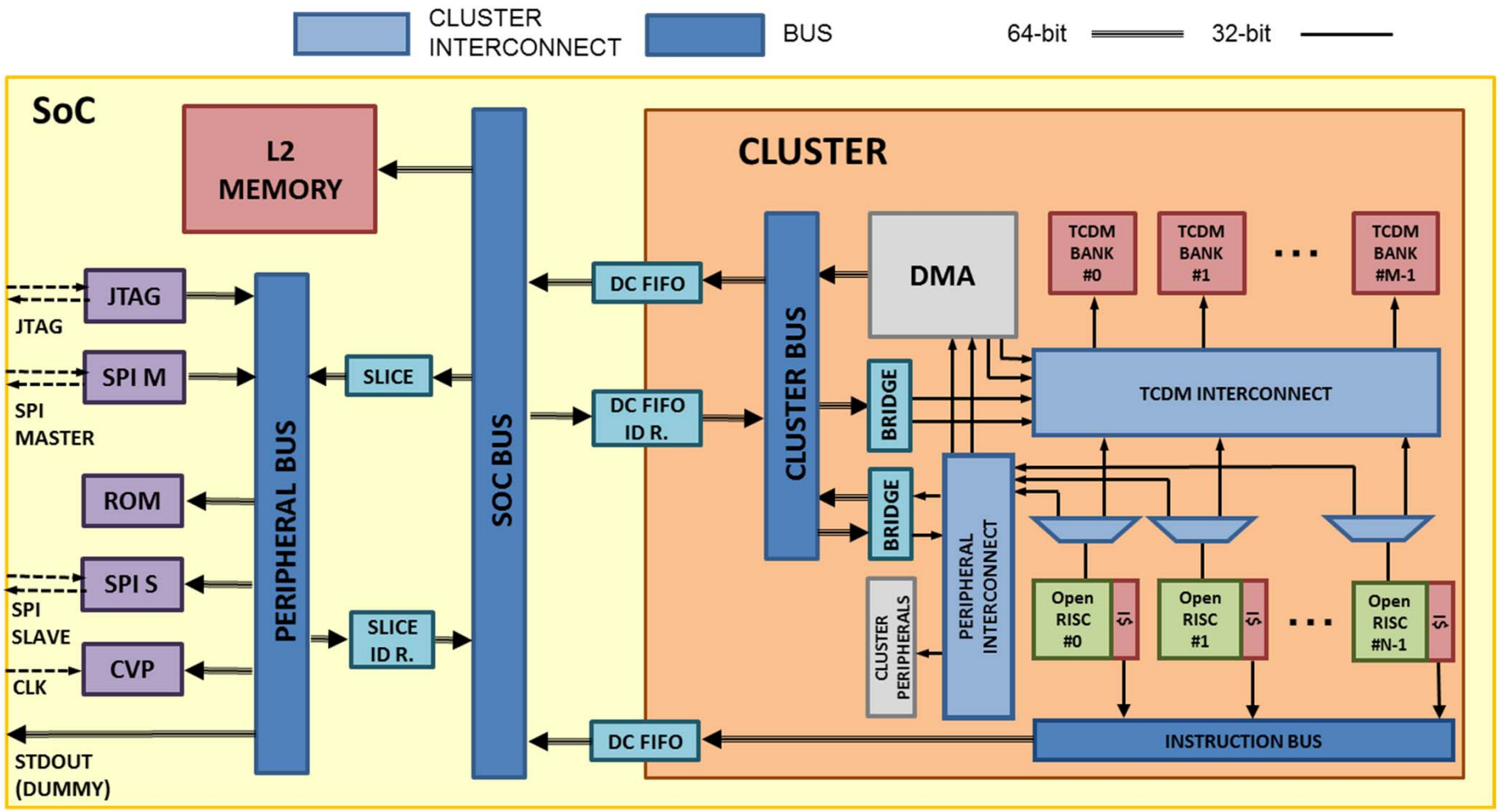
****Does not include IOs***

Measured Results



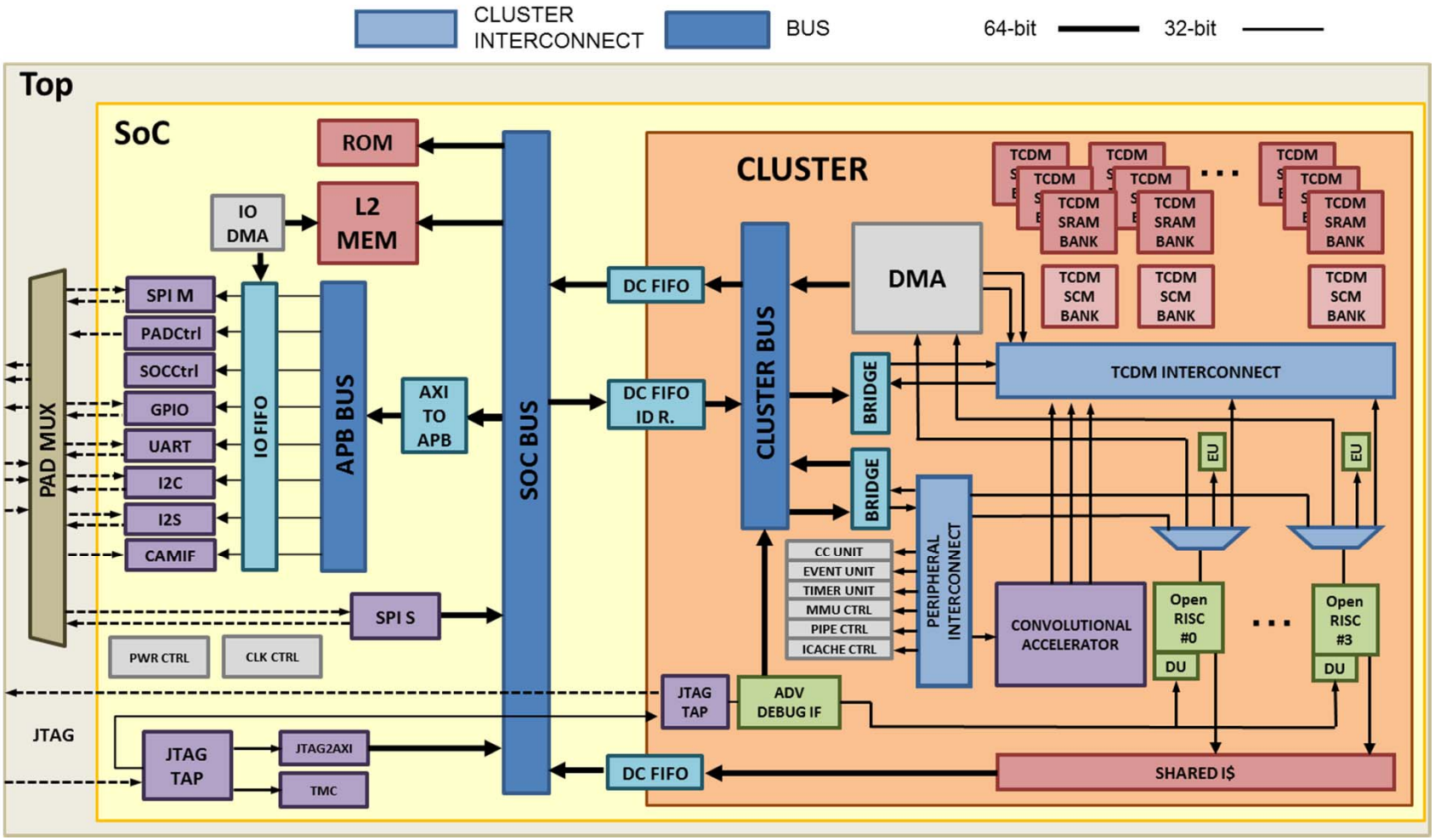
Peak GOPS/W competitive with best-in-class near-threshold (16bit) ULP microcontrollers, plus more than x100 peak GOPS!

PULPv2



= PULPv1 + 2 DVFS regions (SoC + CLUSTER) + Event Unit + Peripherals

PULPv3



= PULPv2 + Extended cores + HW Synch + Shared Cache + HWCE + Shared IOs

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

PULP's Summary

R STUDIORUM
DI BOLOGNA

	PULPv1	*PULPv2*	*PULPv3*
# of cores	4	4	4
L2 memory	16 kB	64 kB	128 kB
TCDM	16kB SRAM	32kB SRAM 8kB SCM	32kB SRAM 16kB SCM
Reconf. pipe. stages	no	yes	yes
I$	4kB SRAM private	4kB SCM private	4kB SCM shared
Body bias regions	yes	yes	yes
DVFS	no	yes	yes
I/O connectivity	JTAG	full	full multiplexed
Extended processor	no	no	Yes
Event unit	no	yes	yes+ HW synchro
Debug unit	no	no	yes

	PULPv1	*PULPv2*	*PULPv3*
Status	silicon proven	post tape out	pre tape out
Technology	FD-SOI 28nm conventional-well	FD-SOI 28nm flip-well	FD-SOI 28nm conventional-well
Voltage range	0.45V - 1.2V	0.3V - 1.2V	0.5V - 0.7V
BB range	-1.8V - 0.9V	0.0V - 1.8V	-1.8V - 0.9V
Max freq.	475 MHz	1 GHz	200 MHz
Max perf.	1.9 GOPS	4 GOPS	1.8 GOPS
Peak en. eff.	60 GOPS/W	135 GOPS/W	385 GOPS/W

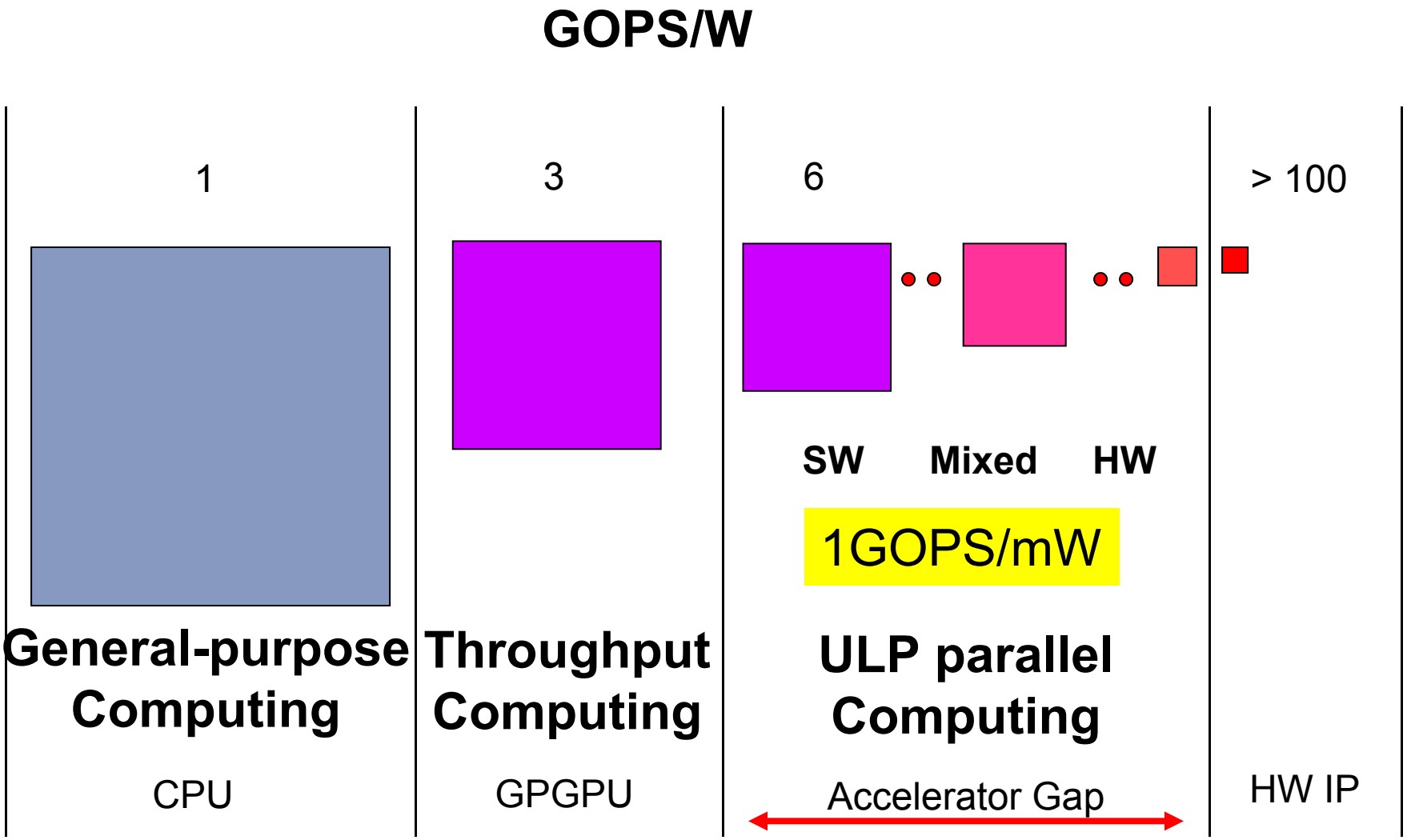
***equivalent 32-bit RISC operations**

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Breaking the GOPS/mW wall

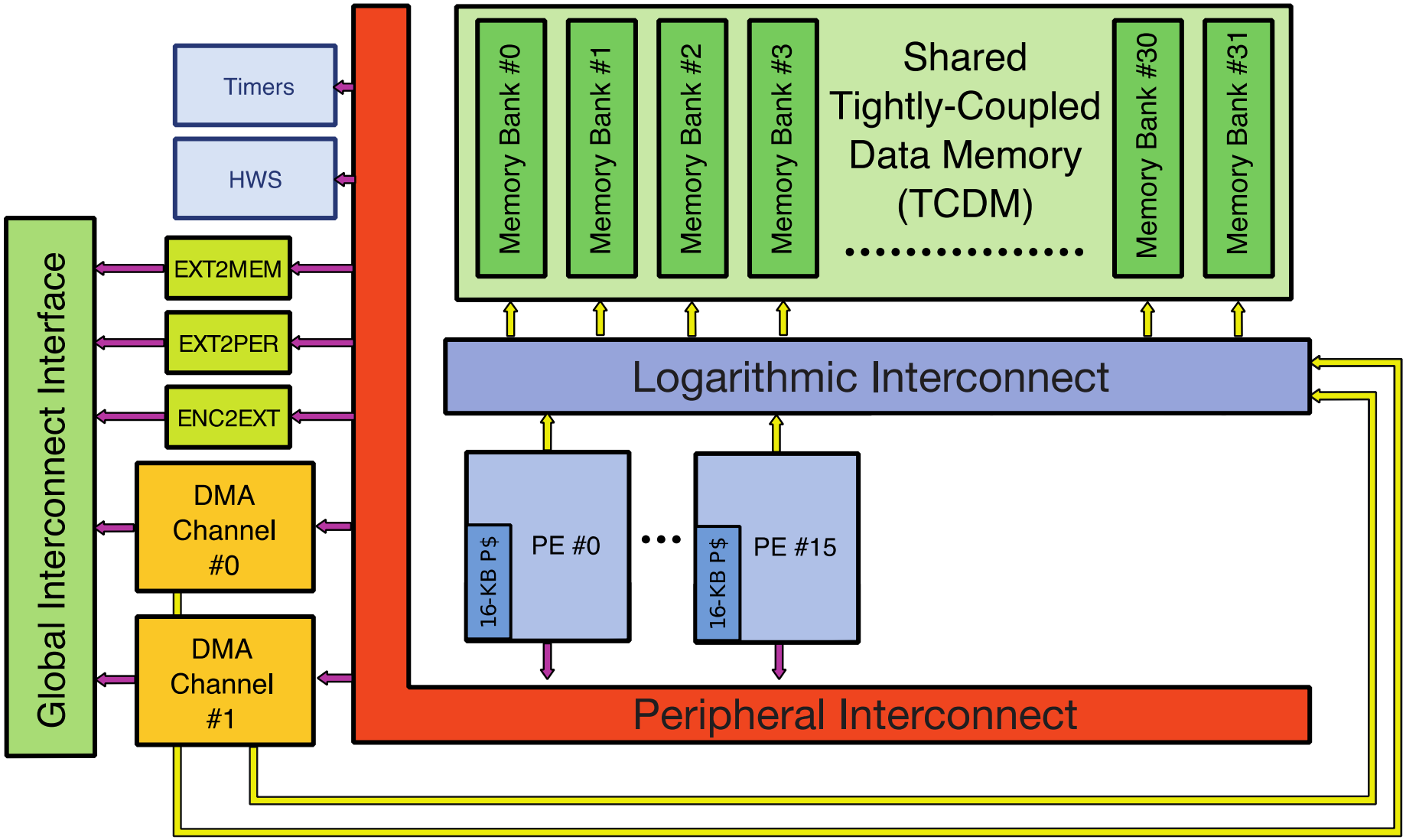
Recovering more silicon efficiency



Closing The Accelerator Efficiency Gap with Agile Customization

Fractal Heterogeneity

Fixed function accelerators have limited reuse… how to limit proliferation?



Learn to Accelerate

- Brain-inspired (**deep convolutional networks**) systems are high performers in many tasks over *many domains*

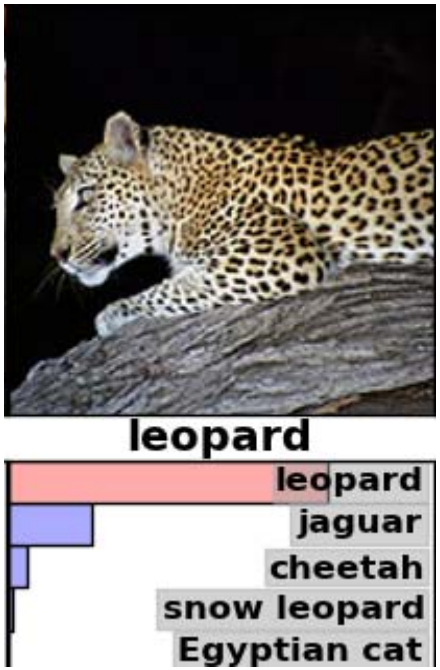
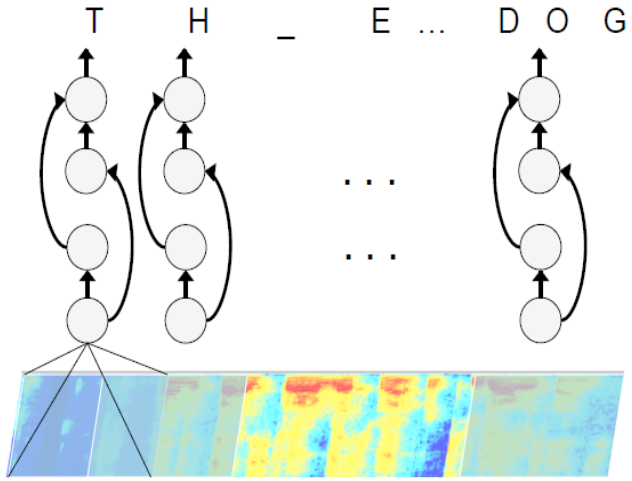


Image recognition
[RussakovskyIMAGENET2014]

- Human:
  85% (untrained),
  94.9% (trained)
- CNN:
  93.4% accuracy



Speech recognition
[HannunARXIV2014]

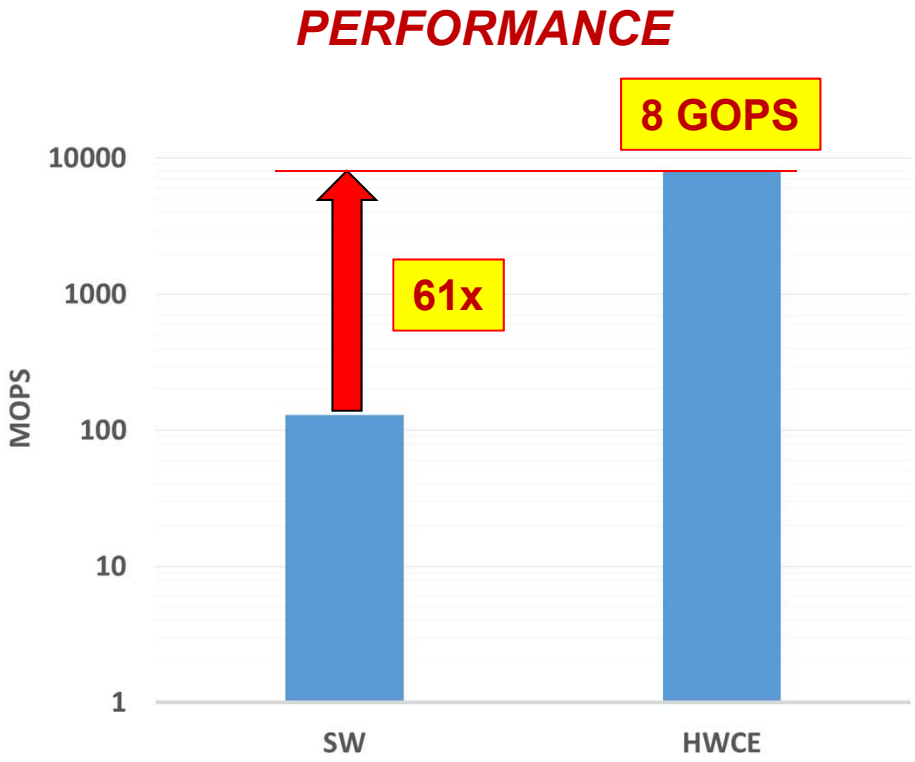
Flexible acceleration: learned CNN weights are “the program”

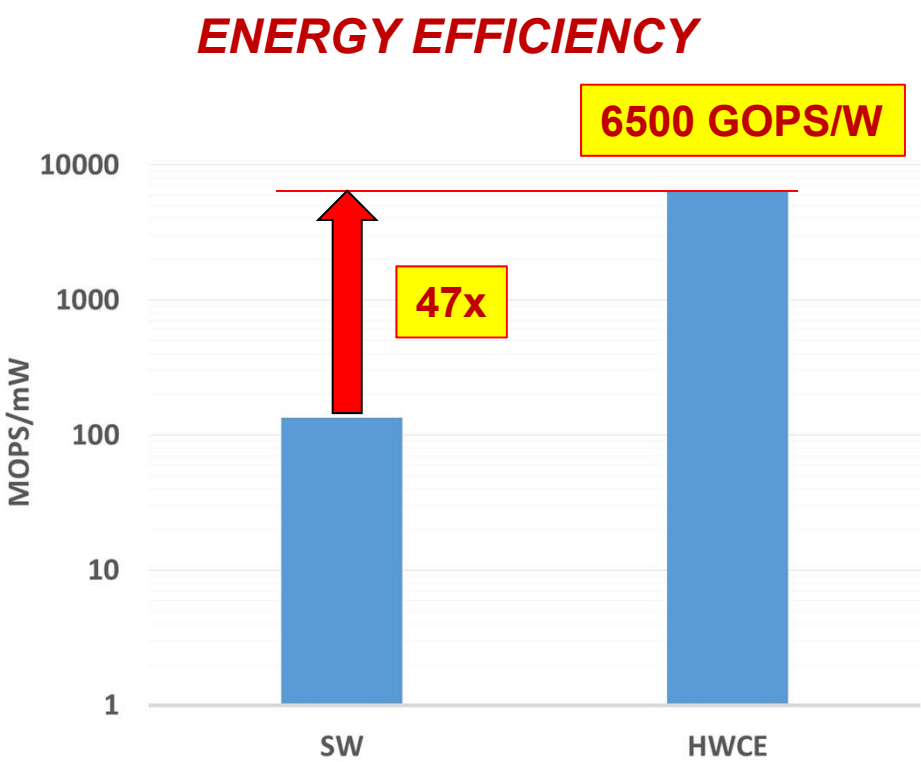
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

PULP CNN Performance

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Average performance and energy efficiency on a 32x16 CNN frame





PULPv3 ARCHITECTURE, CORNER: tt28, 25° C, VDD= 0.5V, FBB = 0.5V

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

ALMA MATER STUDIORUM
UNIVERSITA DI BOLOGNA

Thanks for your attention!!!

cea leti

POLITECNICO MILANO

LELAND STANFORD JUNIOR UNIVERSITY

EPFL ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

life.augmented

UNIVERSITY OF ILLINOIS

www-micrel.deis.unibo.it/pulp-project

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

References

[RuchIBM11] Ruch, P., "Toward five-dimensional scaling: How density improves efficiency in future computers," *IBM Journal of Research and Development* , vol.55, no.5, pp.1-13, 2011.

[AziziISCA10] O. Azizi, et. al., "Energy-Performance Tradeoffs in Processor Architecture and Circuit Design: A Marginal Cost Analysis" *Proceedings of the 37th annual international symposium on Computer architecture, ISCA 2010, pp.* 26-36, June 19–23, 2010.

[Nilsson2014] John-Olof Nilsson et.al., "Foot-mounted inertial navigation made easy", 2014 International Conference on Indoor Positioning and Indoor Navigation, 27-30 October 2014.

[Benatti2014] S .Benatti et. al., "EMG-based hand gesture recognition with flexible analog front end," IEEE Biomedical Circuits and Systems Conference (BioCAS), pp.57,60, Oct. 2014.

[Lagorce2014] Lagorce et. al., "Asynchronous Event-Based Multikernel Algorithm for High-Speed Visual Features Tracking", IEEE Trans Neural Netw Learn Syst. 2014 Sep 16.

[VoiceControl] TrulyHandsfree™Voice Control, available: http://www.sensory.com/wp-content/uploads/80-0342-A.pdf

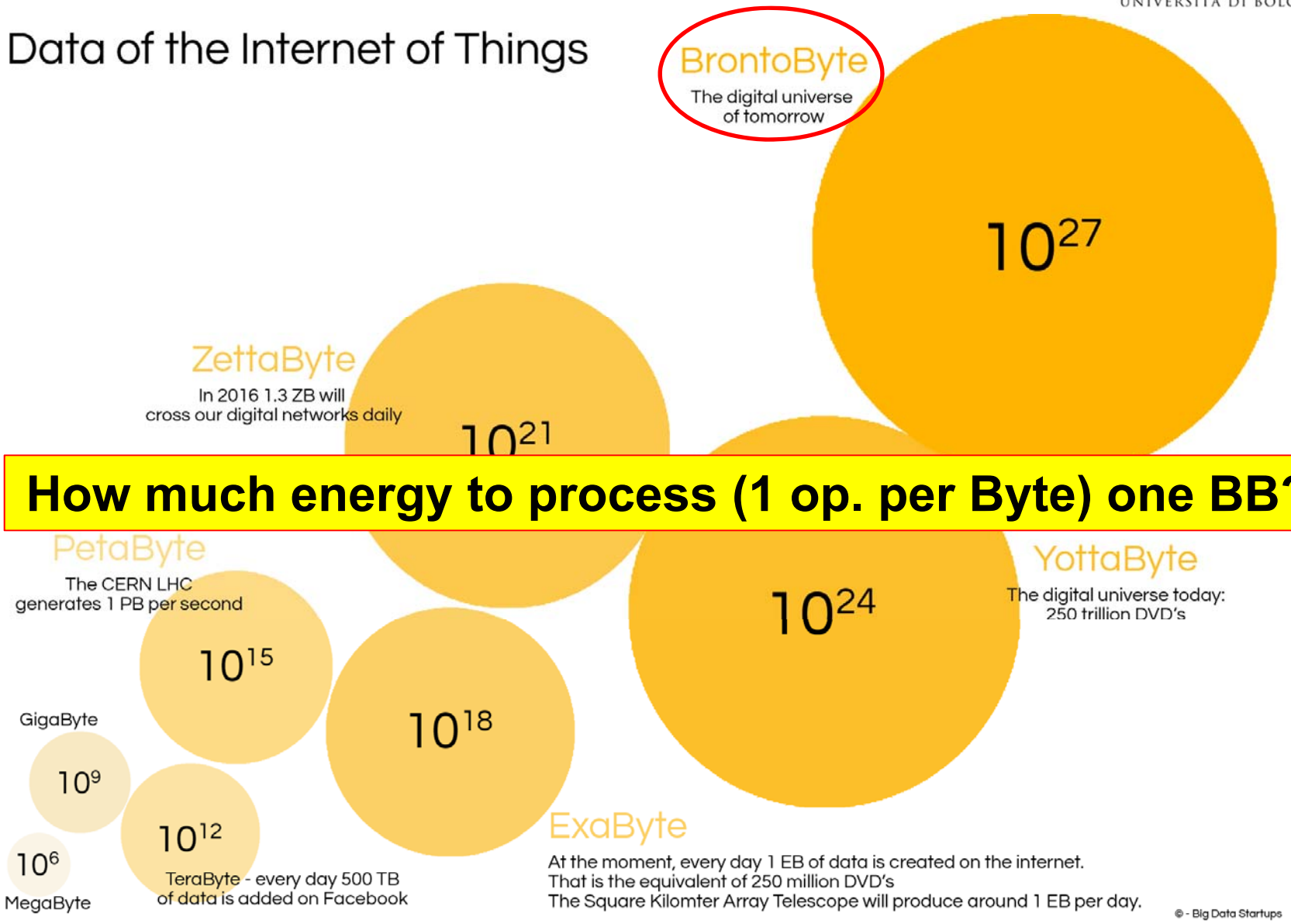
[VivekDeDATE13] De, Vivek, "Near-Threshold Voltage design in nanoscale CMOS," *Design, Automation & Test in Europe Conference & Exhibition DATE, 2013.*

[DoganICSDPTMO2011] Dogan, A. Y., et al., "Power/performance exploration of single-core and multi-core processor approaches for biomedical signal processing," Integrated Circuit and System Design, Power and Timing Modeling, Optimization, and Simulation, pp. 102-11, 2011.

[RussakovskyIMAGENET2014] O. **Russakovsky**, "ImageNet Large Scale Visual Recognition Challenge", International Journal of Computer Vision, 2014.

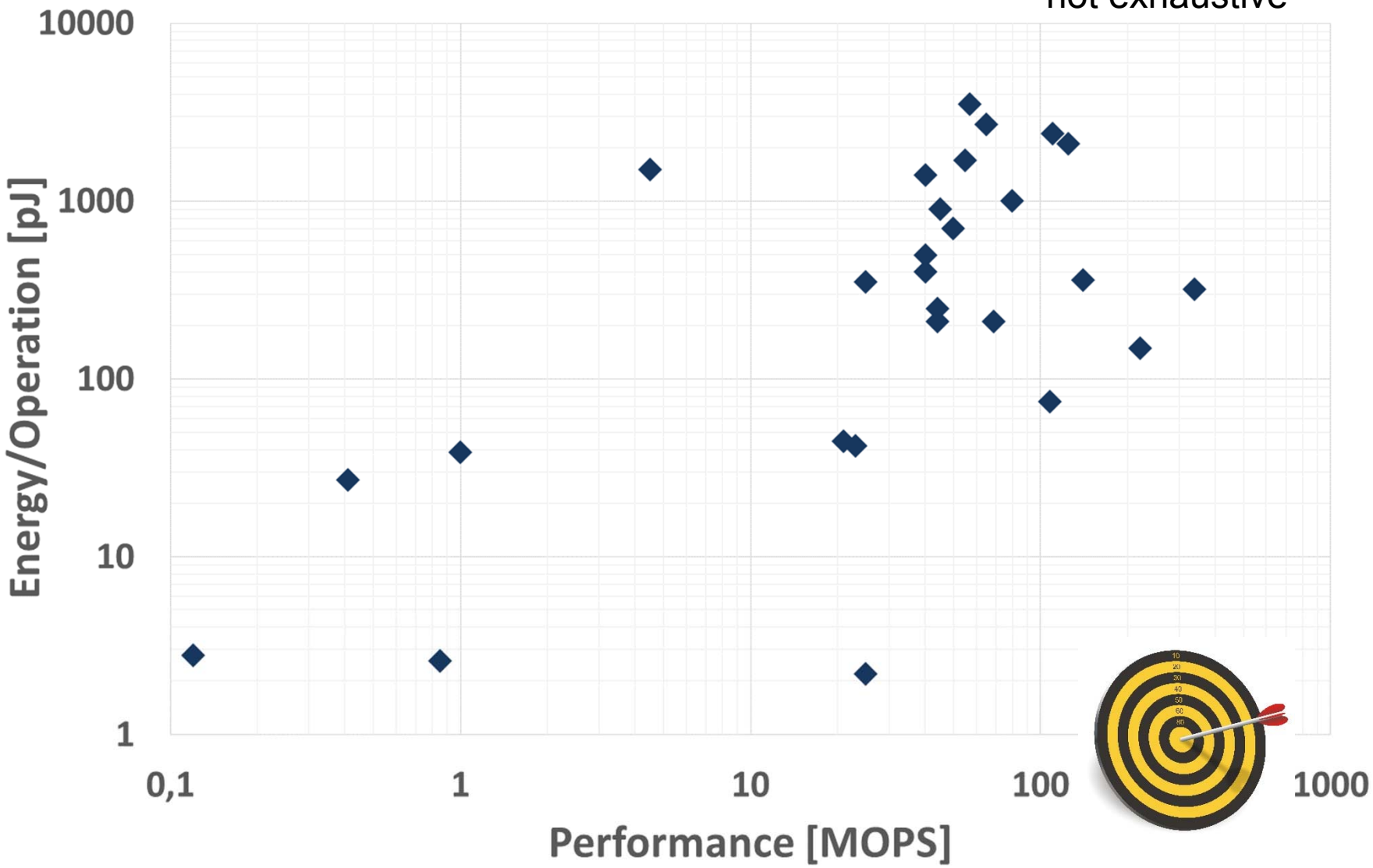
[HannunARXIV2014] A. Hannun " Deep Speech: Scaling up end-to-end speech recognition", arXiv, 2014.

How Big is the IoT?



Microcontrollers Landscape

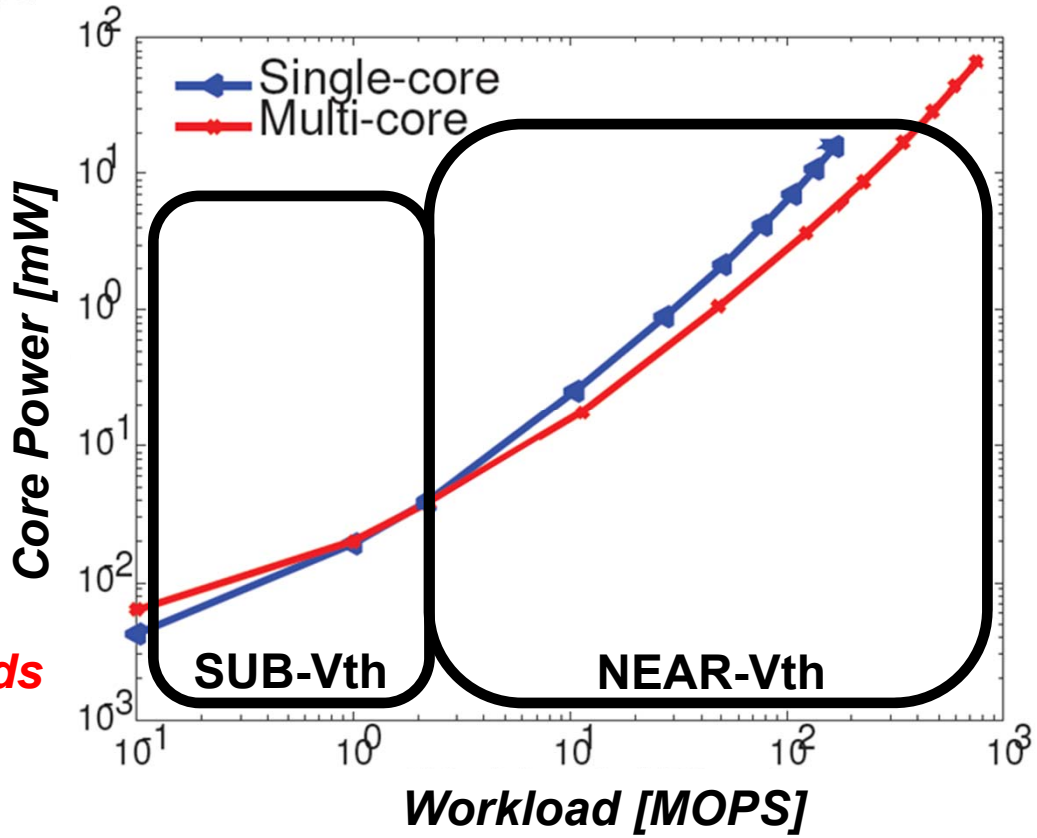
*not exhaustive



ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Parallel NTC



High Workloads

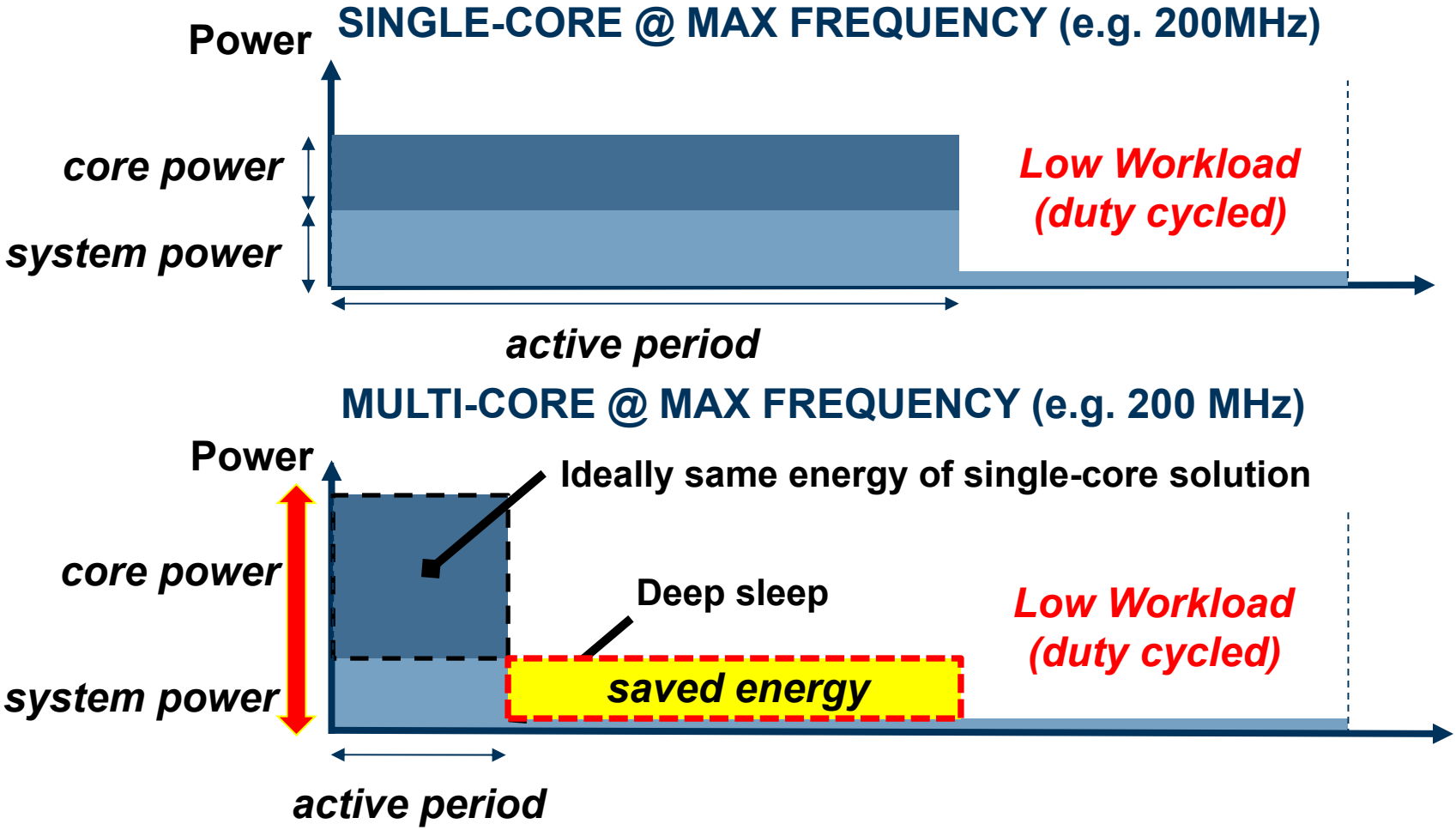
Low Workloads

[DoganICSDPTMO2011]

Target Workload [MOPS]	1-Core Energy Efficiency (ideal) [MOPS/mW]	4-Cores Energy Efficiency (ideal) [MOPS/mW]	Ratio
100	43	55	1.3x
200	33	50	1.5x
400	18	43	2.4x

*Measured on our first prototype

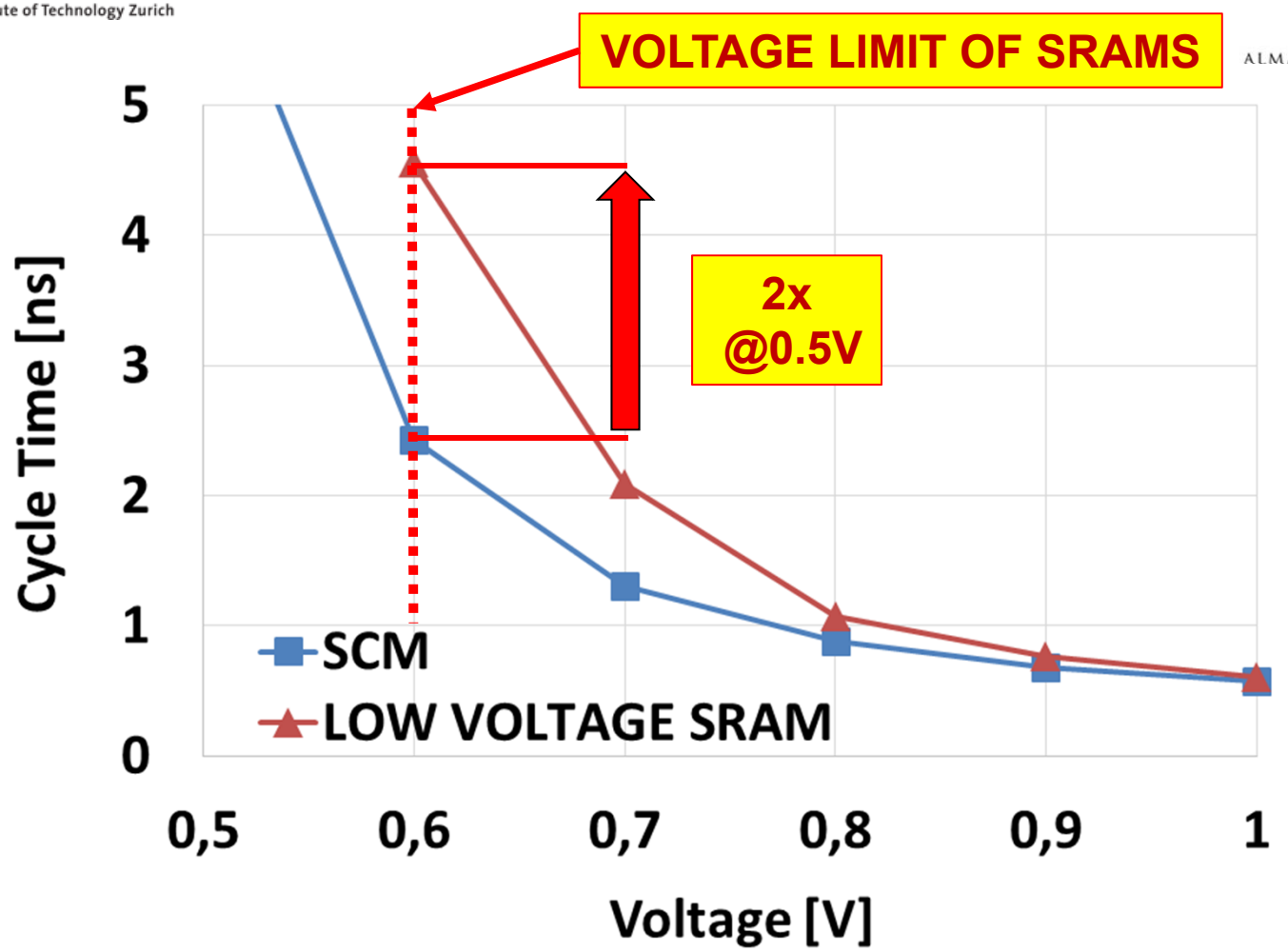
Parallel NTC + Race to Halt



Going faster allows to integrate system power over a smaller period

The main constraint here is the power envelope

Back to SRAMs



SRAM performance rapidly degrades at low voltage

SRAM VDDMIN is higher than logic (and SCM)