ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Logic Synthesis in the Twilight of Moore's Law

Near-threshold, Heterogeneous, 3D Design Looking for a New Toolbox

Luca Benini

IIS-ETHZ & DEI-UNIBO

erc European Research Council

SEVENTH FRAMEWORK PROGRAMME

nano-tera.ch

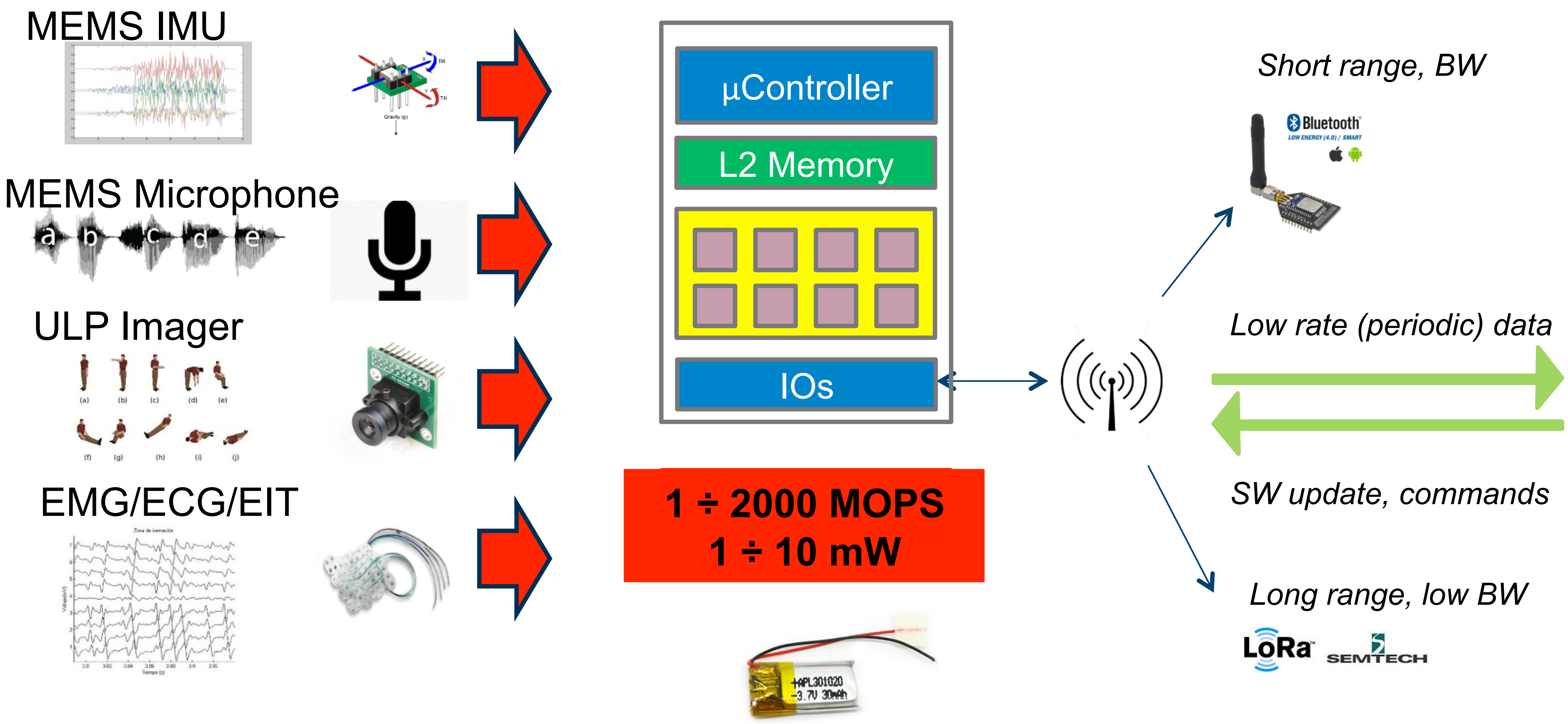
NXP

ST life.augmented

SAMSUNG

IoT: a System View

Sense | **Analyze and Classify** | **Transmit**



MEMS IMU

MEMS Microphone

ULP Imager

EMG/ECG/EIT

µController

L2 Memory

IOs

1 ÷ 2000 MOPS
1 ÷ 10 mW

Short range, BW

Low rate (periodic) data

SW update, commands

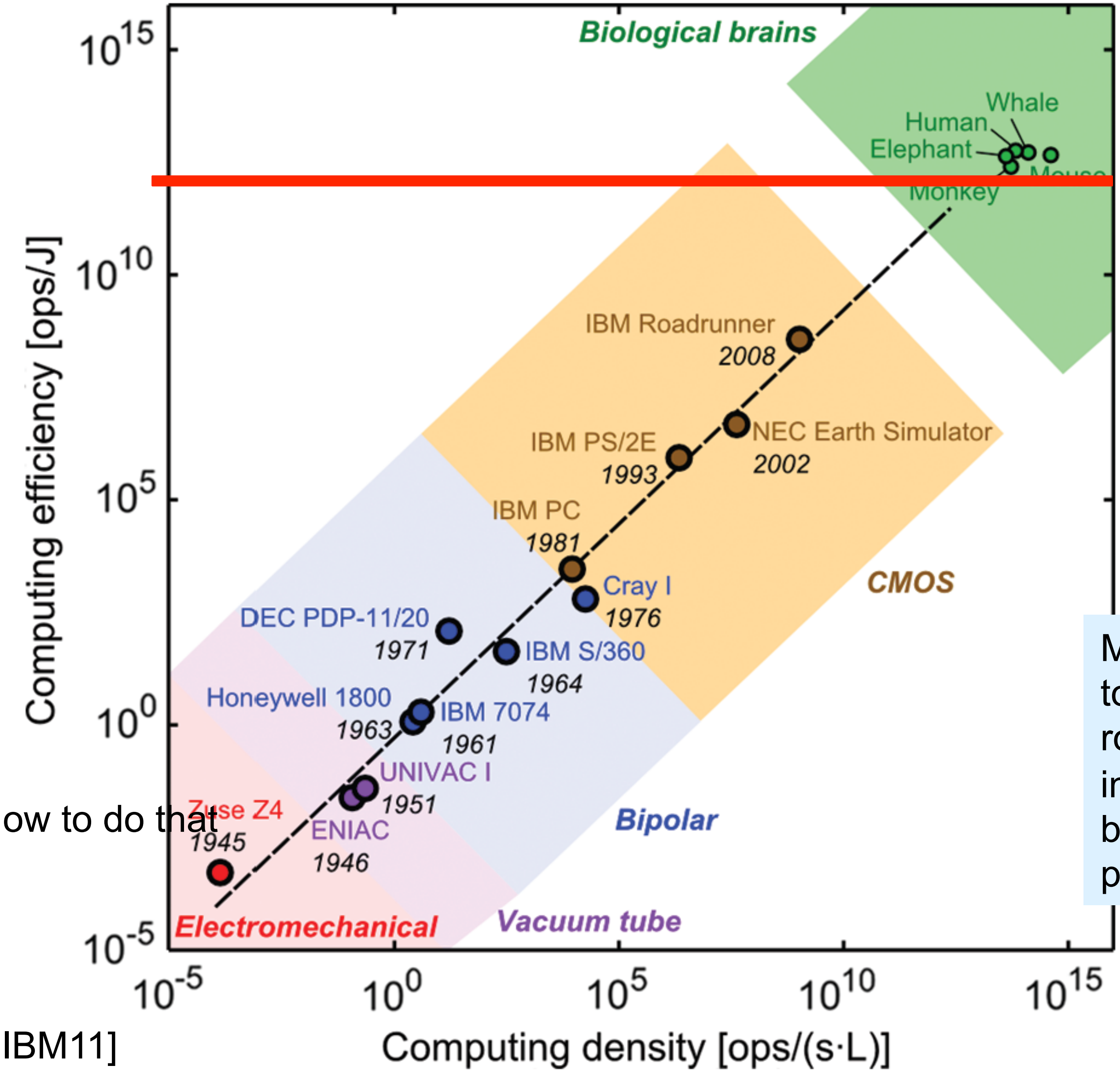
Long range, low BW

100 µW ÷ 2 mW

Battery + Harvesting powered
→ a few mW power envelope

Idle: ~1µW
Active: ~ 50mW

How efficient?



How to do that

10^{12}ops/J

↓

1pJ/op

↓

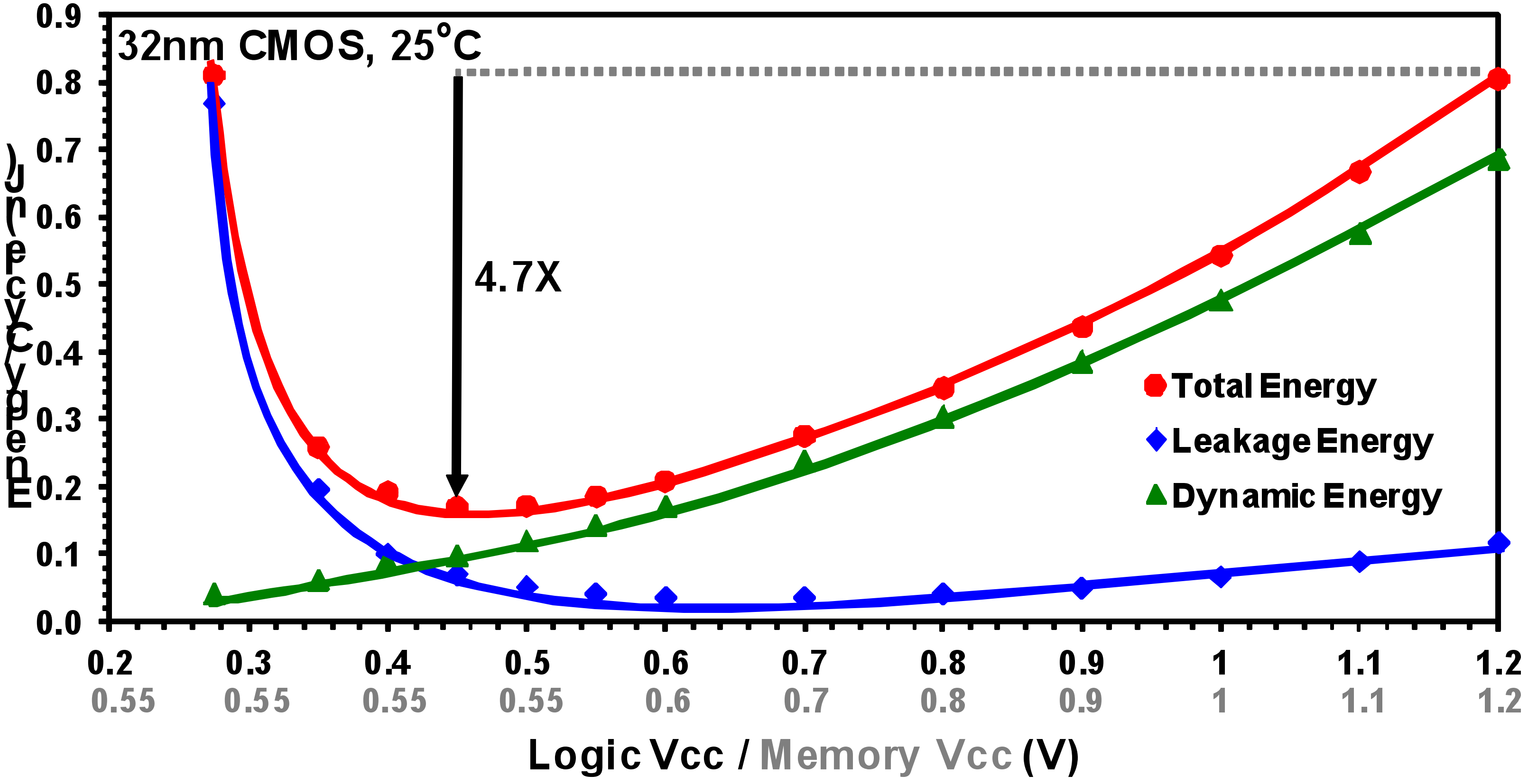
1GOPS/mW

Moore’s law has slowed to roughly 2 ½ years or roughly 30 months (25% increase in the time between semiconductor process nodes)

[RuchIBM11]

Minimum energy operation

Source: Vivek De, INTEL – Date 2013



Near-Threshold Computing (NTC):

1. **Don't waste energy pushing devices in strong inversion**
2. **Recover performance with parallel execution**

ETH

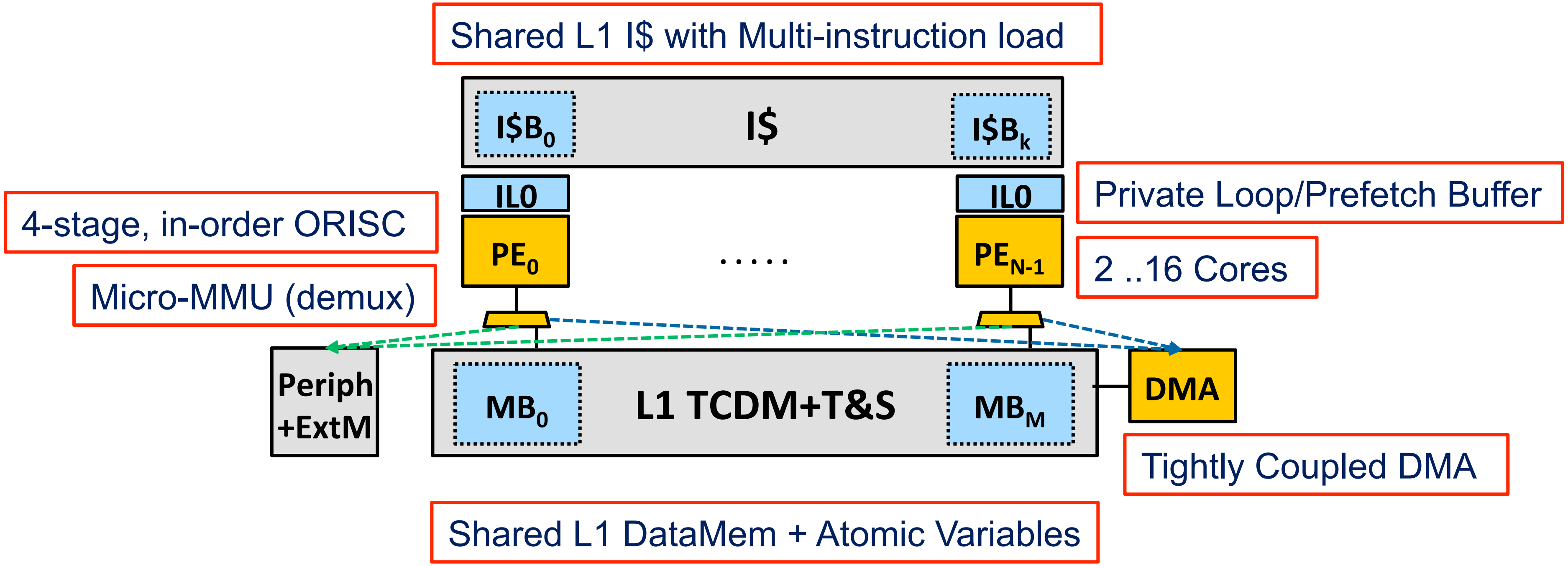
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

PULP – Parallel Ultra Low Power

Near-Threshold Multiprocessing

Open Source Hardware & Software



Shared L1 I$ with Multi-instruction load

Private Loop/Prefetch Buffer

4-stage, in-order ORISC

2 ..16 Cores

Micro-MMU (demux)

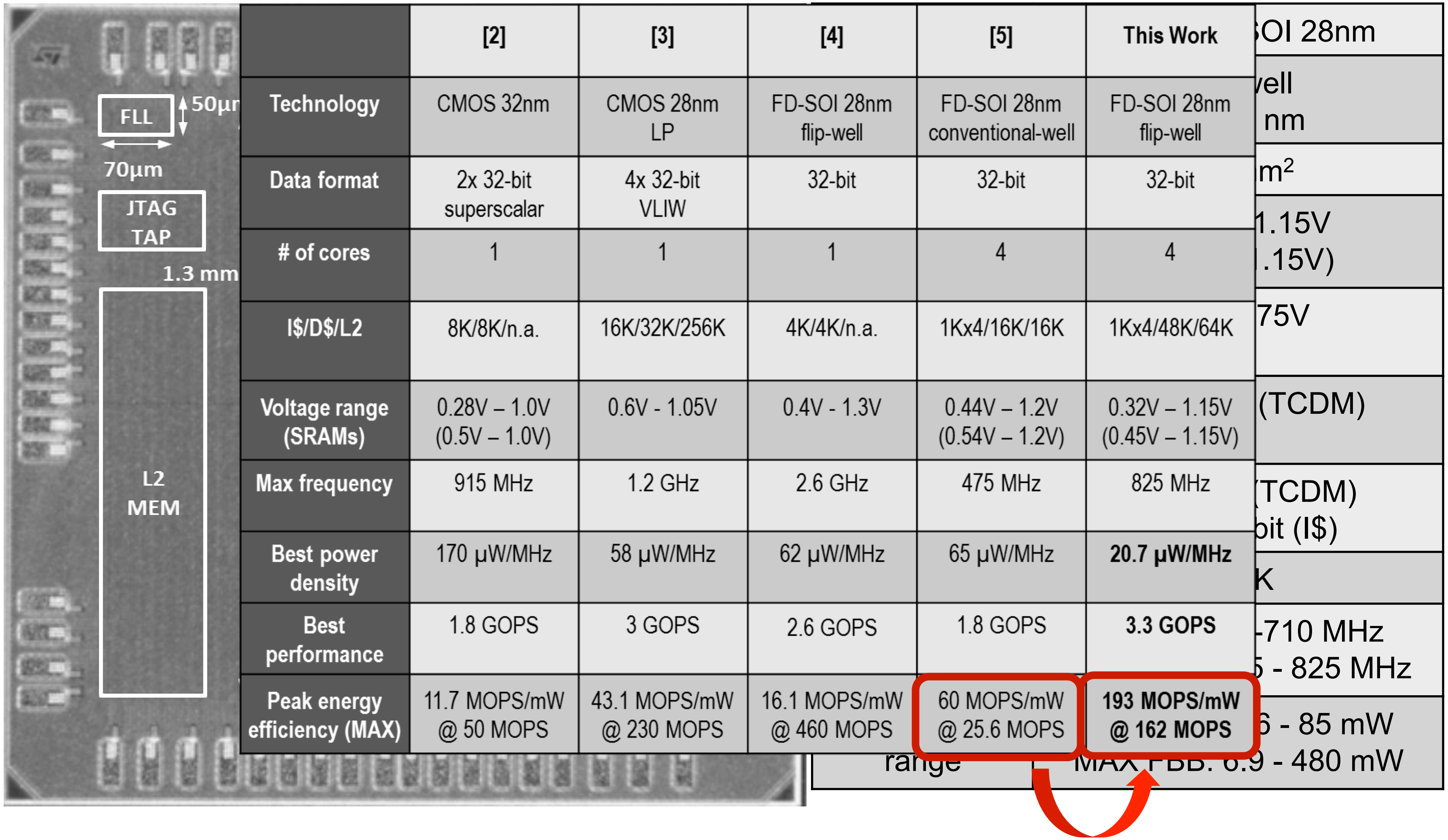
Tightly Coupled DMA

Shared L1 DataMem + Atomic Variables

NT but parallel → Max. Energy efficiency when Active

+ strong PM for (partial) idleness

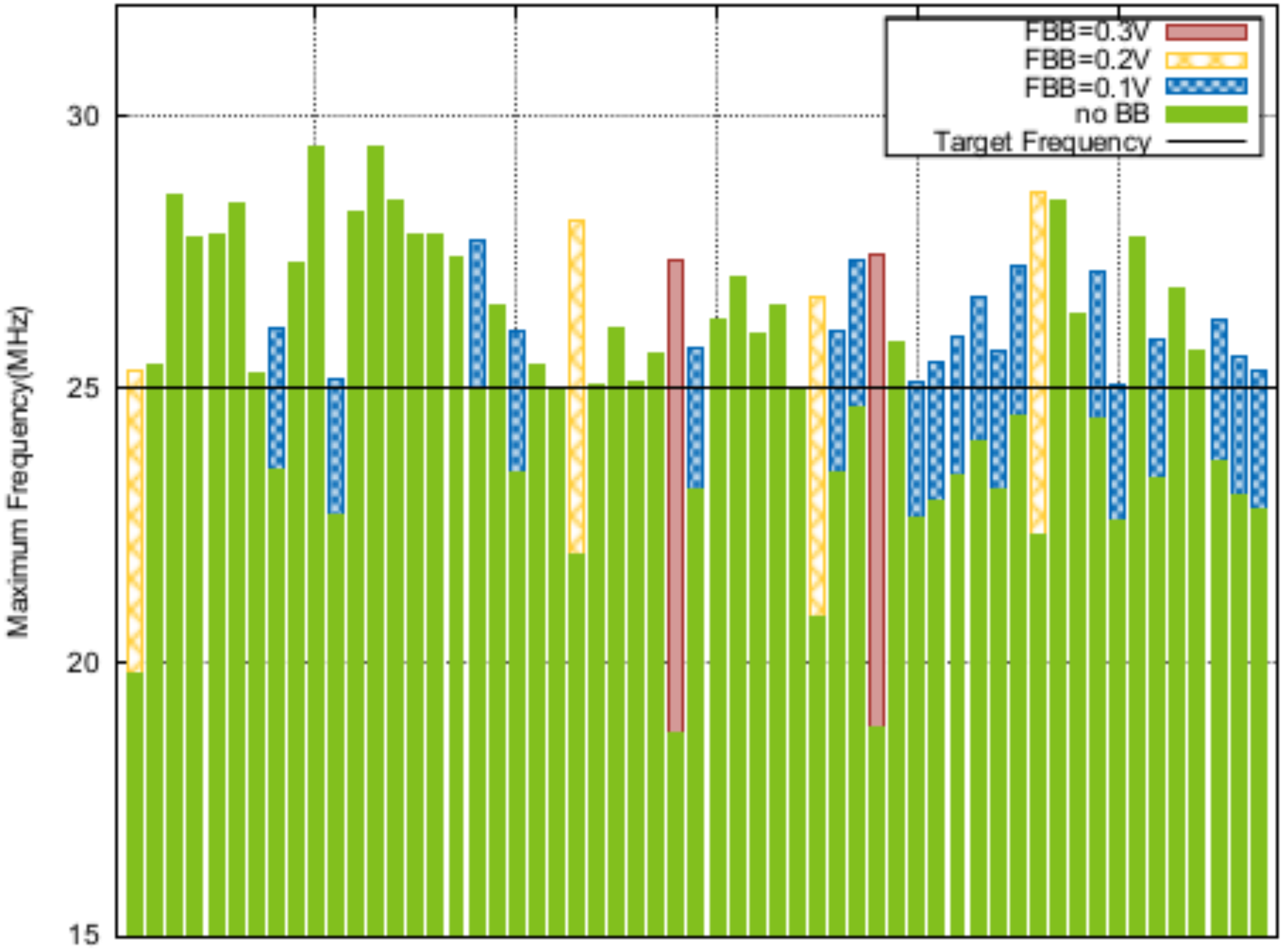
PULP Chips



	[2]	[3]	[4]	[5]	This Work
Technology	CMOS 32nm	CMOS 28nm LP	FD-SOI 28nm flip-well	FD-SOI 28nm conventional-well	FD-SOI 28nm flip-well
Data format	2x 32-bit superscalar	4x 32-bit VLIW	32-bit	32-bit	32-bit
# of cores	1	1	1	4	4
I$/D$/L2	8K/8K/n.a.	16K/32K/256K	4K/4K/n.a.	1Kx4/16K/16K	1Kx4/48K/64K
Voltage range (SRAMs)	0.28V – 1.0V (0.5V – 1.0V)	0.6V - 1.05V	0.4V - 1.3V	0.44V – 1.2V (0.54V – 1.2V)	0.32V – 1.15V (0.45V – 1.15V)
Max frequency	915 MHz	1.2 GHz	2.6 GHz	475 MHz	825 MHz
Best power density	170 µW/MHz	58 µW/MHz	62 µW/MHz	65 µW/MHz	**20.7 µW/MHz**
Best performance	1.8 GOPS	3 GOPS	2.6 GOPS	1.8 GOPS	**3.3 GOPS**
Peak energy efficiency (MAX)	11.7 MOPS/mW @ 50 MOPS	43.1 MOPS/mW @ 230 MOPS	16.1 MOPS/mW @ 460 MOPS	60 MOPS/mW @ 25.6 MOPS	**193 MOPS/mW @ 162 MOPS**

ISSCC15 (student presentations, Hot Chips 15, ISSCC16 (paper+student presentation)

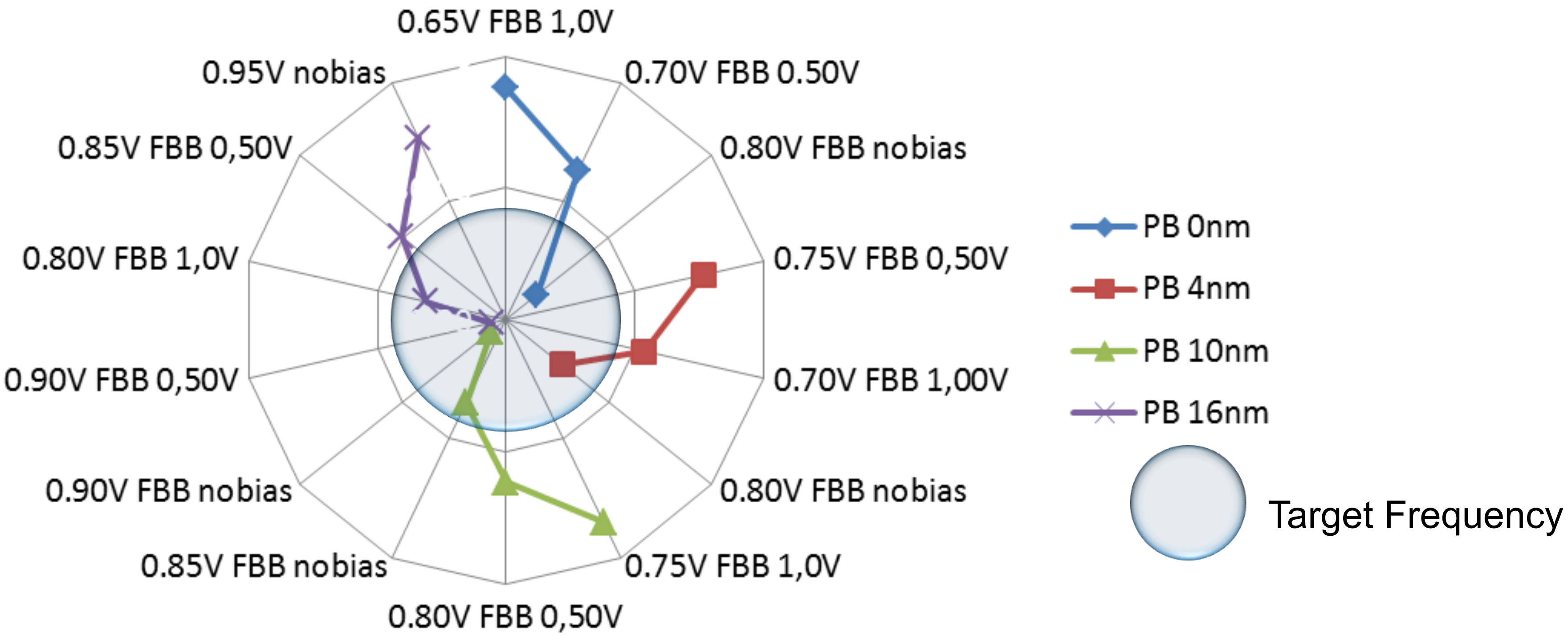
Variability!



Temperature awareness BB/leakage management is essential

Synthesis Challenge

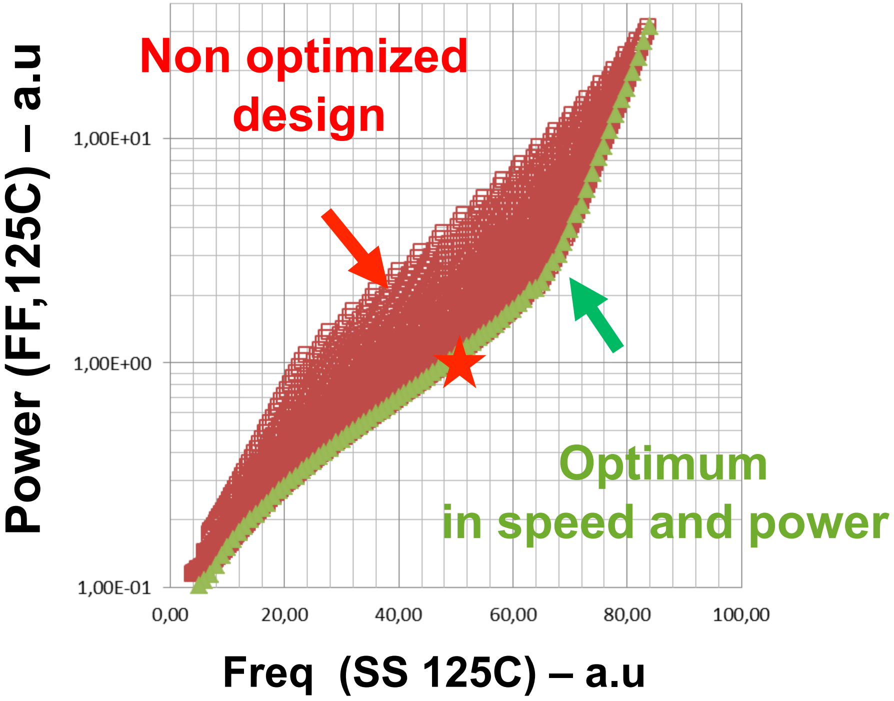
- An extensive set of parameters to consider:
  - Supplies, Poly biasing, Body biasing, Gate sizing
  - Subject to temperature, reliability, mission profile constraints



(Vdd, Pb, BB) choice becomes a power-delay trade off exercise

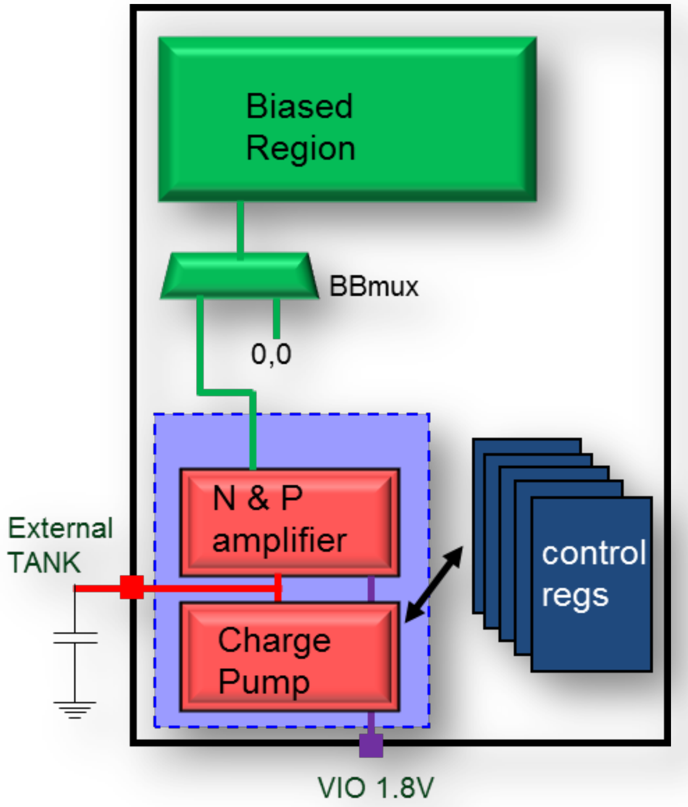
Optimization and Trade-off

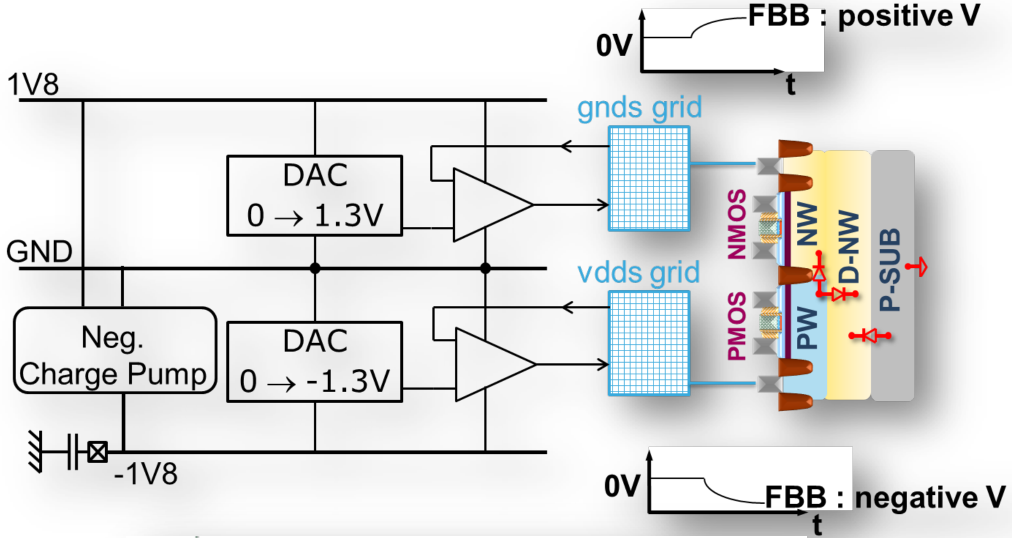
- Conditions
  - 28nm UTBB FDSOI
  - V_{DD}^{min} (0.5V) < V_{DD} < V_{DD}^{max} (1.3V)
  - P_b^{min} (0) < P_b < P_b^{max} (16nm)
  - B_b^{min} (0) < B_b < B_b^{max} (2.0V)
  - Pdyn/Pstat ratio = 50%
  - Power,Perf corners

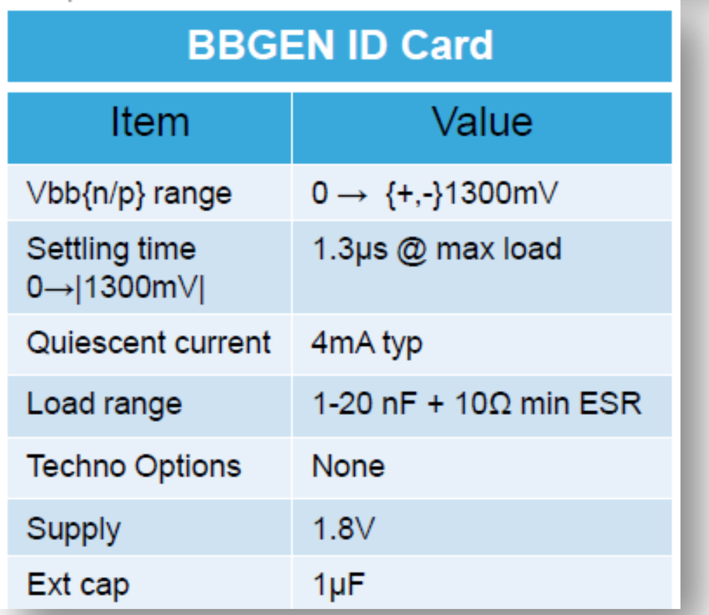


- An optimized design means:
  - Maximize performance for given power
  - Minimize power for given performance
  - Area constraint
- The optimum vector is a function (Vdd, Pb, BB)
  - Strongly dependent on chosen corners
  - Static + **Dynamic**

Dynamic Body Bias







BBGEN ID Card	
Item	Value
Vbb{n/p} range	0 → {+,-}1300mV
Settling time 0→|1300mV|	1.3µs @ max load
Quiescent current	4mA typ
Load range	1-20 nF + 10Ω min ESR
Techno Options	None
Supply	1.8V
Ext cap	1µF

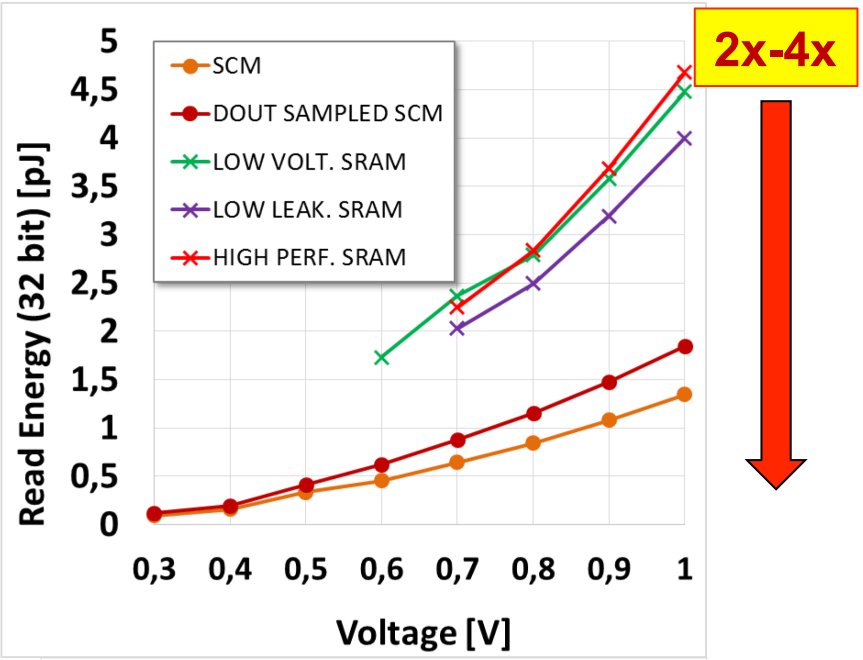
Dynamic adaptation can also be used to «remove» extremely adverse corners and ease MC-MM optimization

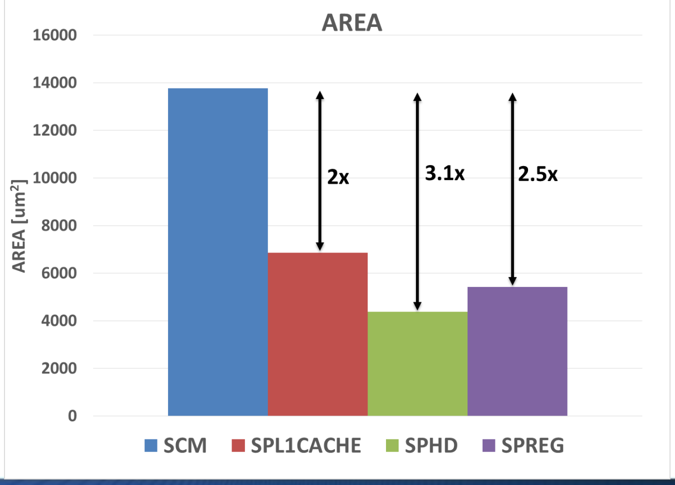
ULP Bottleneck: Memory

- "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
  - Major area overhead (2x)

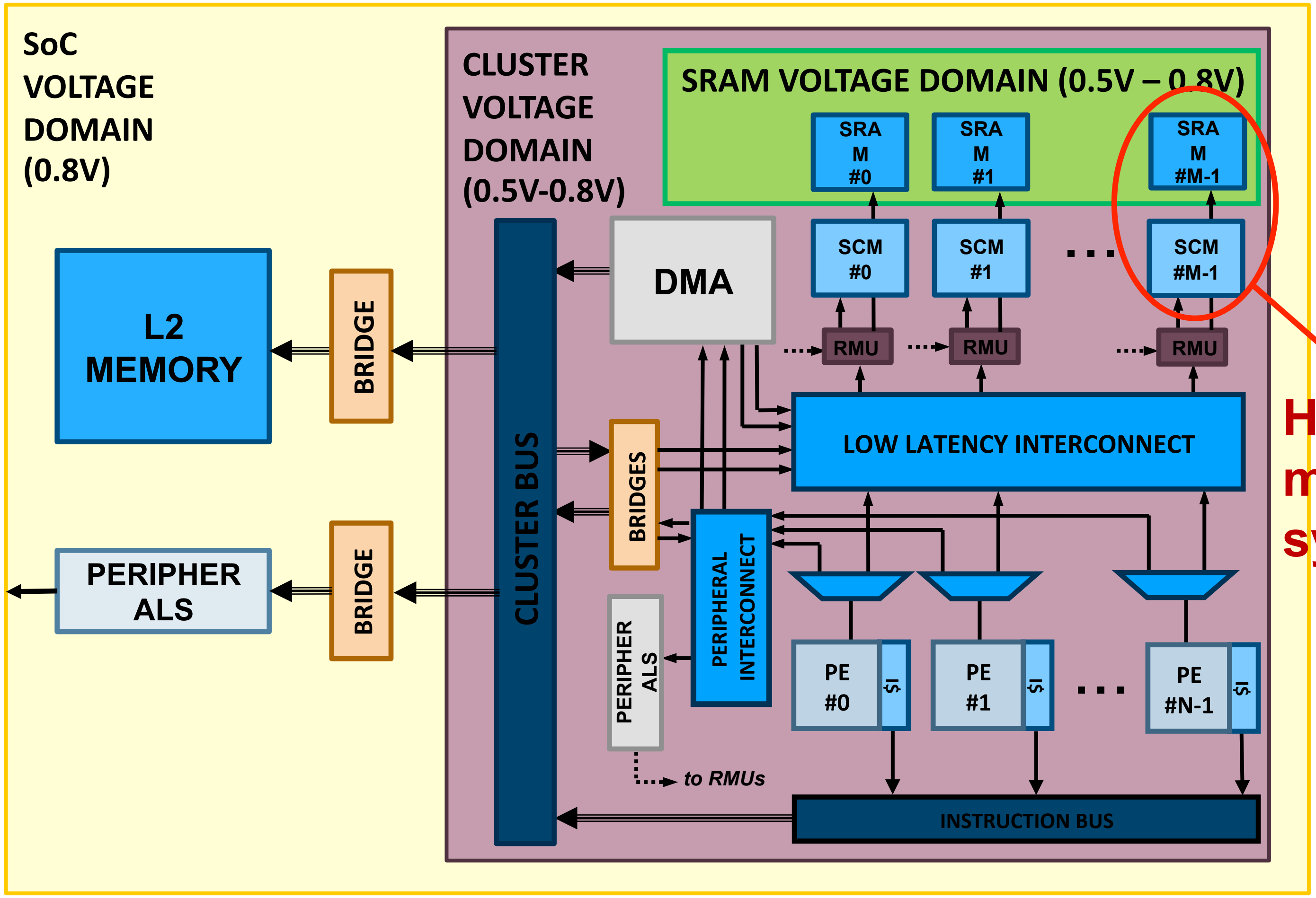
Need help exploring memory tradeoffs!

256x32 6T SRAMS vs. SCM





Static vs. Dynamic again…



Hybrid memory system

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Approximate Computing to the Rescue

Approximate → Adequate

Less-than-perfect results perceived as correct by the users
e.g. image processing (filtering)

RGB to GRAYSCALE

RGB to GRAYSCALE (+ 10% error)

Approximation is not always acceptable
→ **Application and program phase dependent!**

Approximate Storage?

- Retention voltage

	Retention
SCM	0.25V
6T-SRAM	0.29V

- Probability of **flip-bit error** on a single bit during read/ write operations

Voltage (V)	0.50	0.55	0.60	0.65	0.70	0.75	0.80
P(flip-bit) SCM	0.0	0.0	0.0	0.0	0.0	0.0	0.0
P(flip-bit) 6T	**0.0037**	**0.0012**	**0.0003**	**5.24e-5**	**4.35e-6**	**4.16e-8**	0.0

Energy vs. Precision tradeoff → big range!

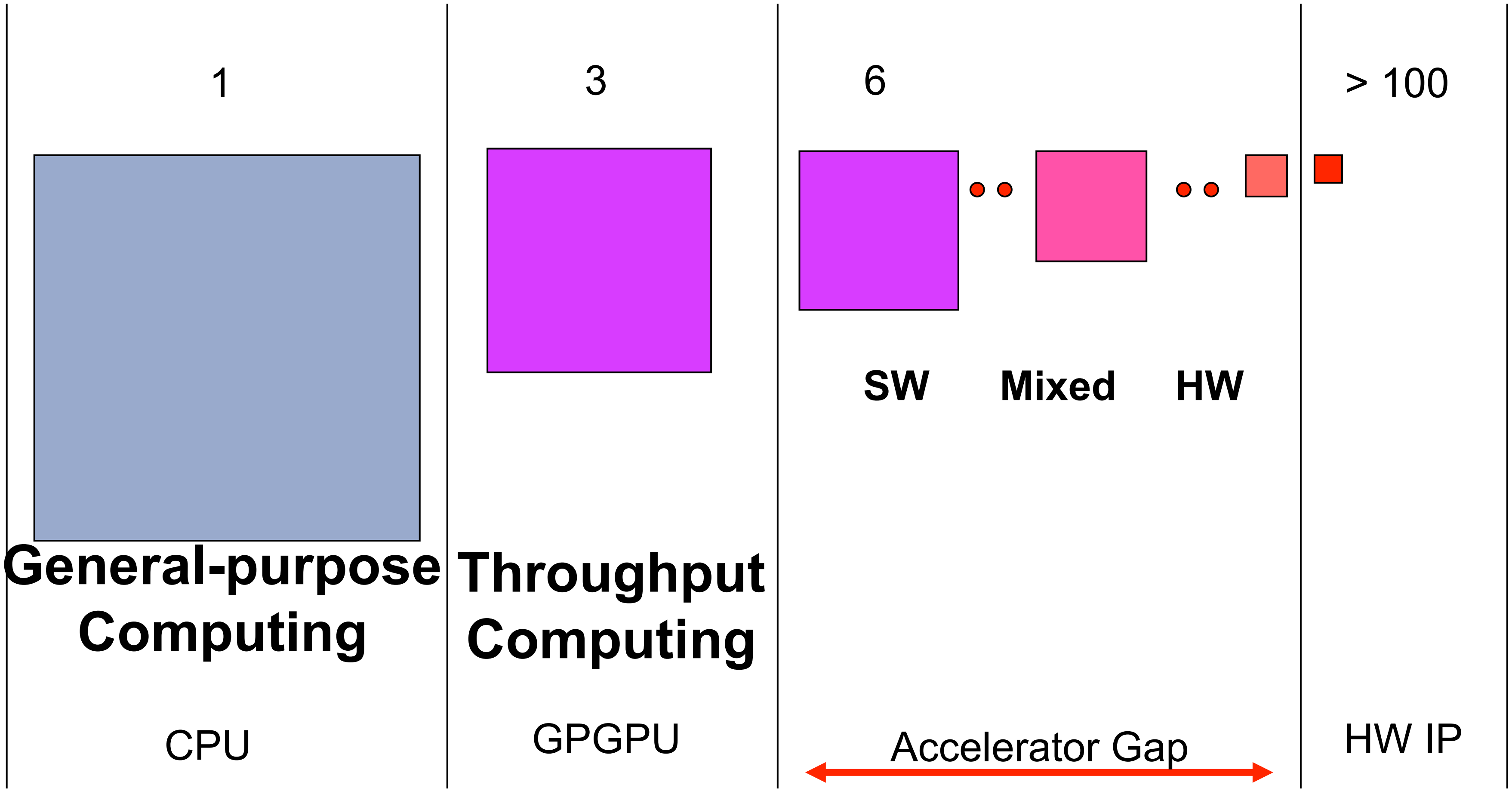
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Acceleration

Recovering more silicon efficiency

GOPS/W



1

General-purpose Computing

CPU

3

Throughput Computing

GPGPU

6

SW Mixed HW

Accelerator Gap

> 100

HW IP

Closing The Accelerator Efficiency Gap with Agile Customization

Learn to Accelerate

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

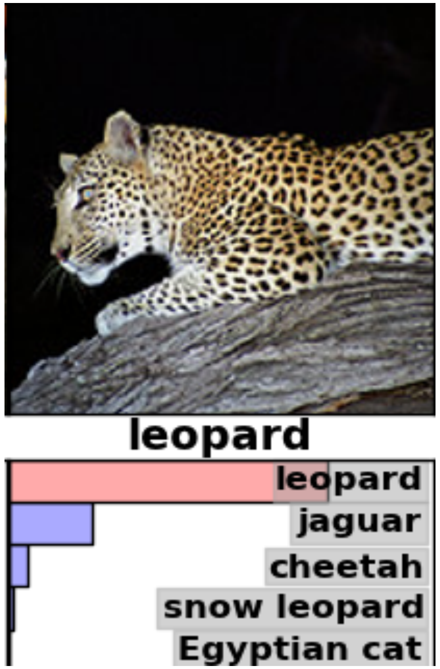
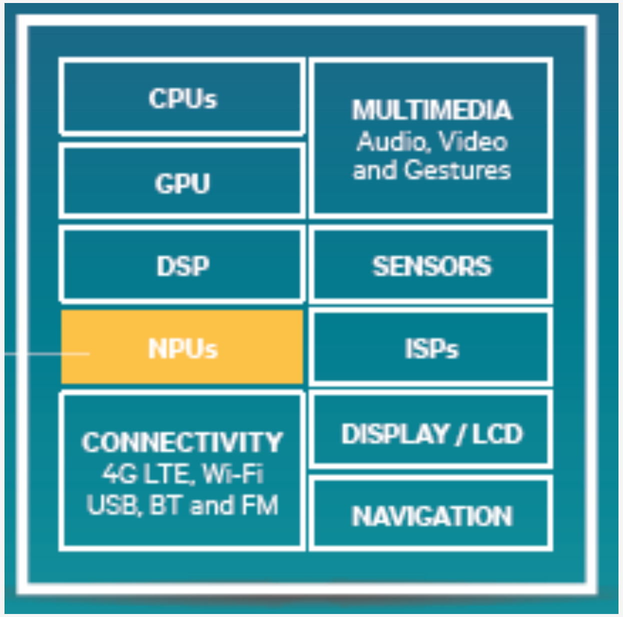


Image recognition
[Russakovsky et al., 2014]



CPUs

GPU

DSP

NPUs

MULTIMEDIA
Audio, Video
and Gestures

SENSORS

ISPs

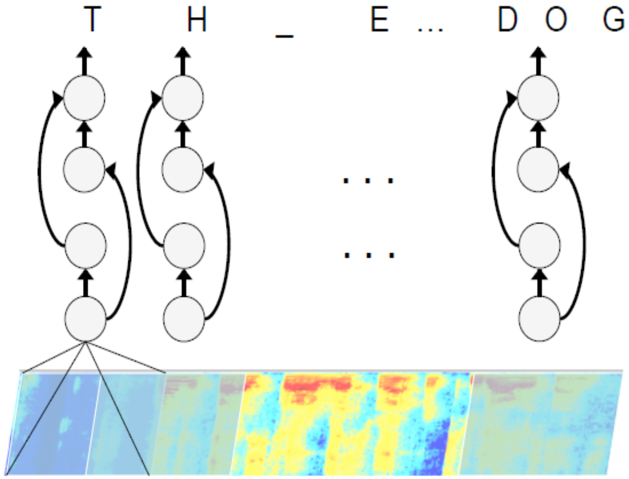
CONNECTIVITY
4G LTE, Wi-Fi
USB, BT and FM

DISPLAY / LCD

NAVIGATION

Spiking NN Accelerator

QUALCOMM

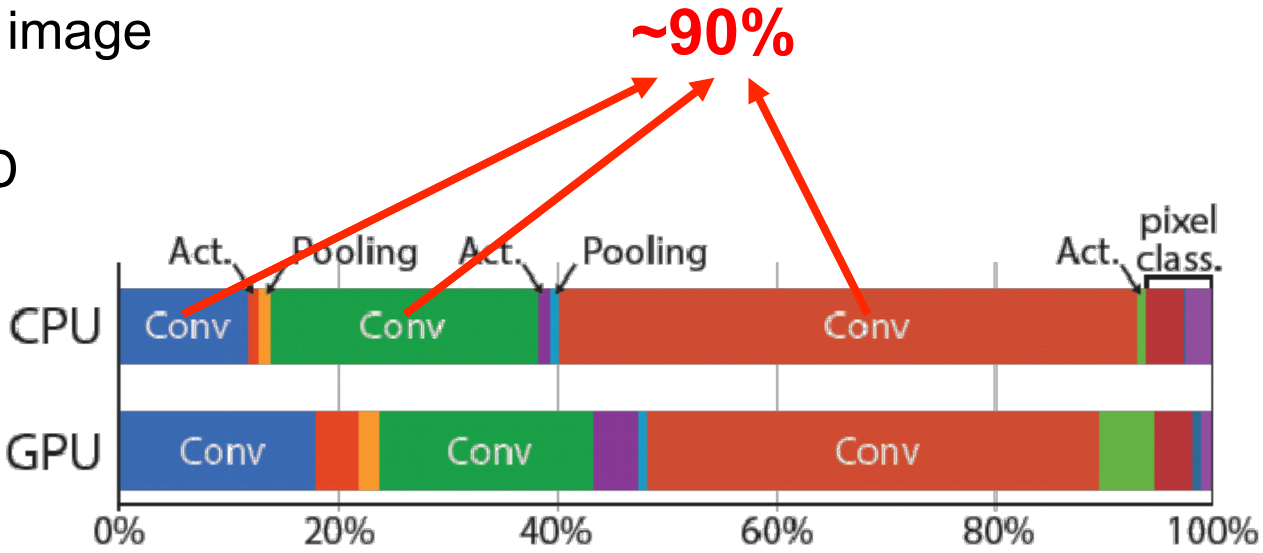


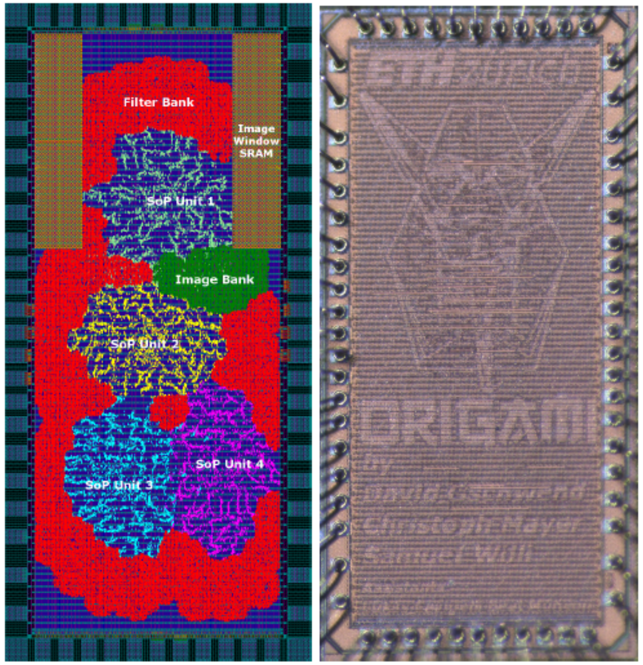
Speech recognition
[Hannun et al., 2014]

- **Flexible** acceleration: learned CNN weights are “the program”

Computational Effort

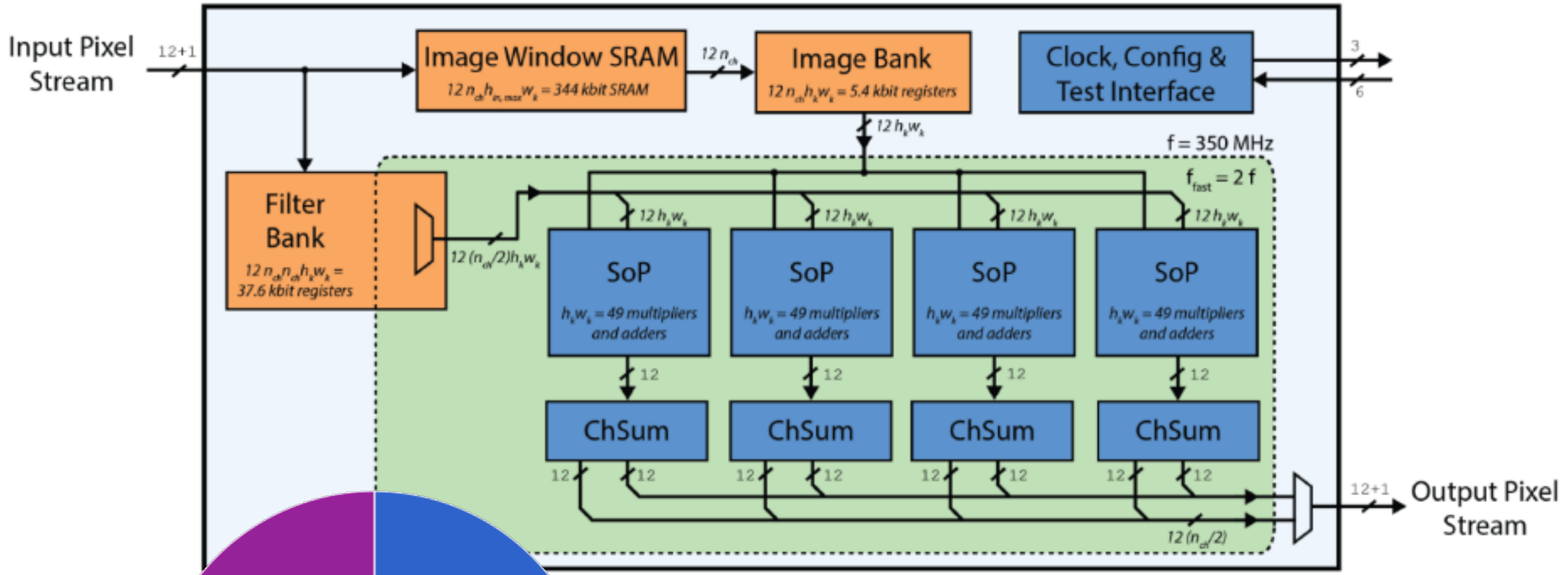
- Computational effort
  - 7.5 GOp for 320x240 image
  - 260 GOp for FHD
  - 1050 GOp for 4k UHD

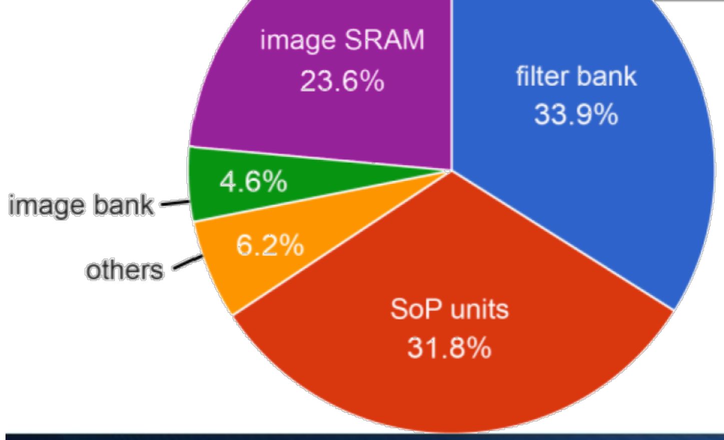




Origami a CNN accelerator

Origami: The Architecture


Input Pixel Stream
12+1
Image Window SRAM
$12 n_{ch} h_{in,max} w_k$ = 344 kbit SRAM
$12 n_{ch}$
Image Bank
$12 n_{ch} h_k w_k$ = 5.4 kbit registers
Clock, Config & Test Interface
3
6
$12 h_k w_k$
f = 350 MHz
f_{fast} = 2 f
Filter Bank
$12 n_{ch} n_{ch} h_k w_k$ = 37.6 kbit registers
$12 (n_{ch}/2) h_k w_k$
$12 h_k w_k$
SoP
$h_k w_k$ = 49 multipliers and adders
12
ChSum
12
12+1
Output Pixel Stream
$12 (n_{ch}/2)$

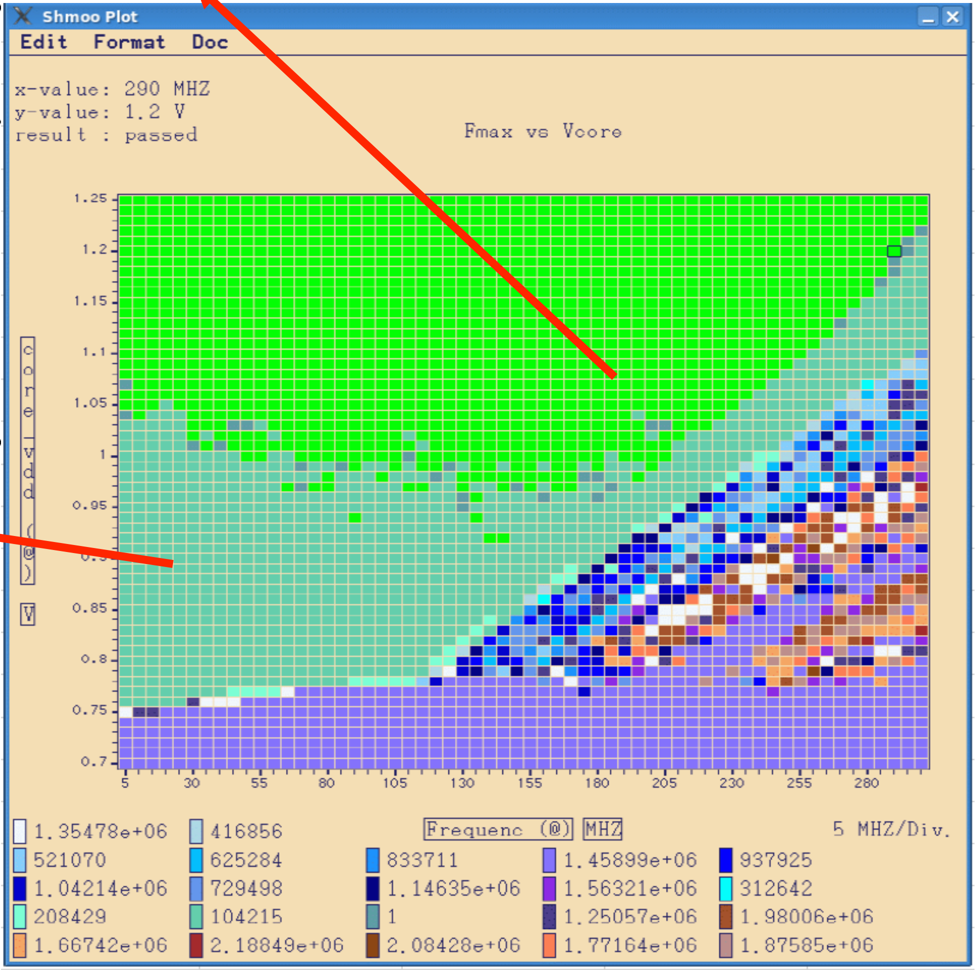


- FP needed?
  - 12-bit signals sufficient
  - Input to classification double-vs-12-bit accuracy loss < 0.5% (80.6% to 80.1%)

Smooth Degradation with Vdd↓

publication	V_{core} V	power mW	efficiency GOp/s/W
ConvEngine [46]	0.72	398	1030
ShiDianNao [44]	0.8	61.3	2098
NeuFlow [24]	0.8	239	1339
HWCE [43]	0.8	180	260
HWCE [43]	0.4	0.73	1375
this work	0.8	86.1	2276
this work	0.53	7.81	9475

0% bit flips



```
x-value: 290 MHZ
y-value: 1.2 V
result : passed
```

1% bit flips

67% energy improvement

Really needing synthesis tools for exploring the approximation space for these «arithmetically dense» architectures

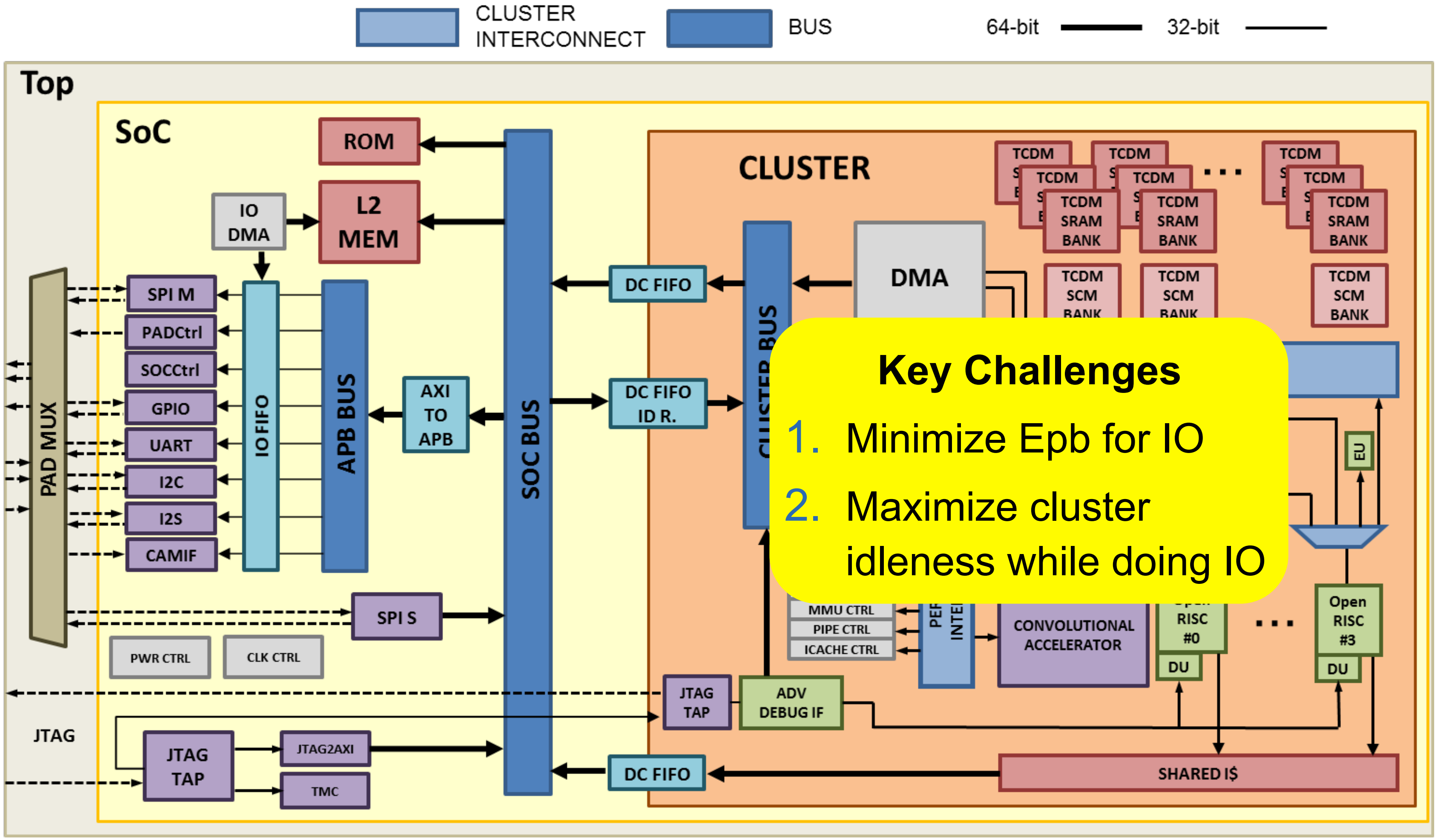
1. Numerical precision
2. Controlled error tolerance

Conclusions

- ioT Energy efficiency requirements are super-tight
  - Technology scaling alone is not doing the job for us
  - Ultra-low power "traditional computing" architecture and circuits are needed, but not sufficient in the long run
- Approximation for energy efficiency is apromising direction
  - SW and SW-abstractions are key
- Need synthesis tools more than ever!

Thank you!

Next bottleneck - IO



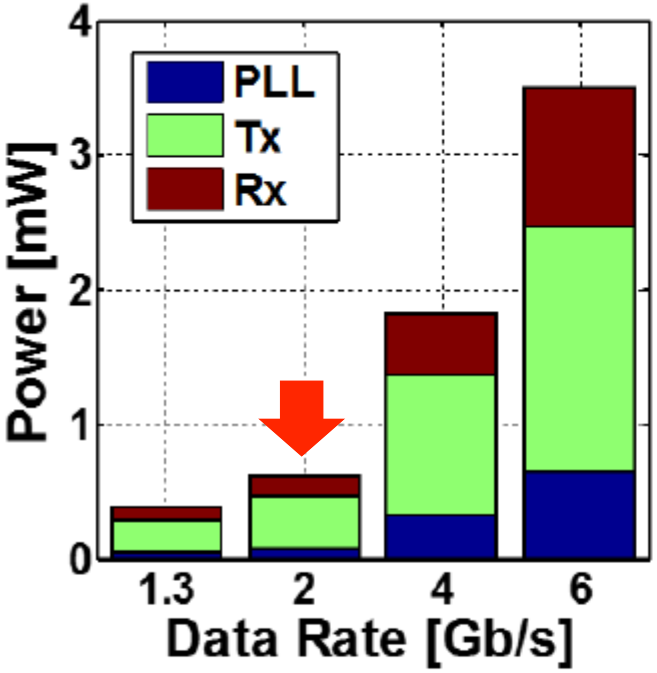
Key Challenges

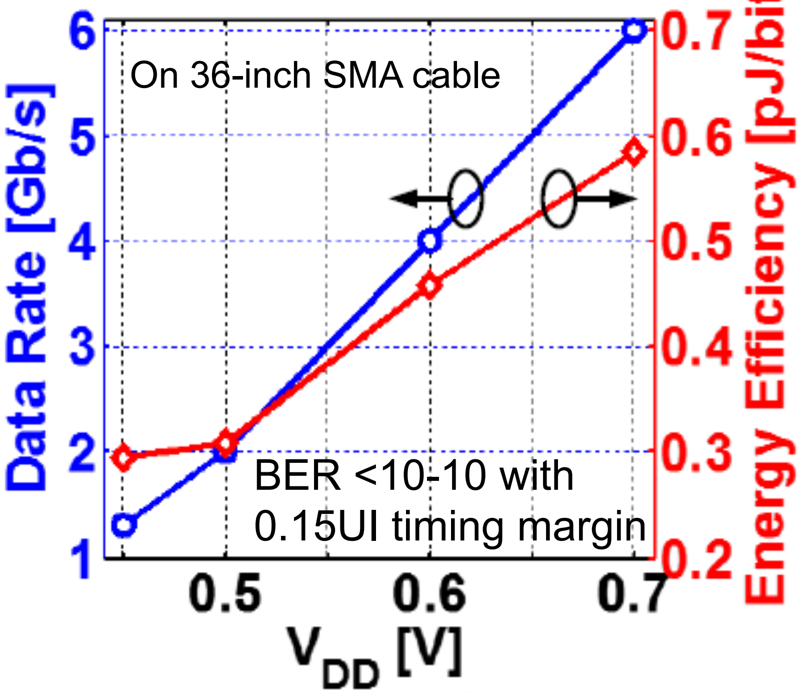
1. Minimize Epb for IO
2. Maximize cluster idleness while doing IO

Flexible and low-pin count interface layer – (Quasi)-Serial is better

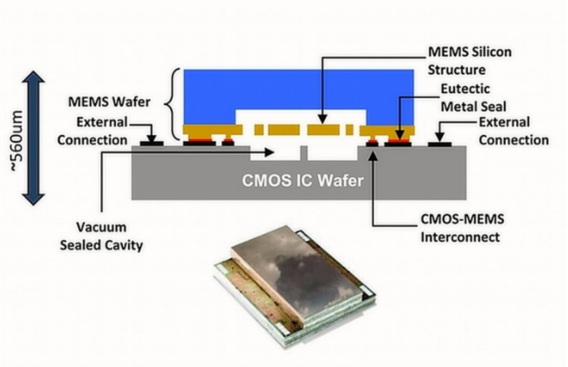
ULP Serial Phy

- A 0.45-0.7V 1-6Gb/s 0.29-0.58pJ/bit Source Synchronous Transceiver Using Automatic Phase Calibration in 65nm CMOS (0.15mm^2)





- Source-synchronous, pseudo-differential, unterminated, Voltage Mode, 200mVpp, 1/8 rate CLK, self-calibrating PLL-based phase generator
- Low-cost SIP+die stacking option for **processor + memories + sensors** becomes viable


MEMS Silicon Structure
Eutectic Metal Seal
External Connection
MEMS Wafer
External Connection
~560um
CMOS IC Wafer
Vacuum Sealed Cavity
CMOS-MEMS Interconnect