ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Parallel digital signal processing in a mW power envelope: how and why?

Luca Benini

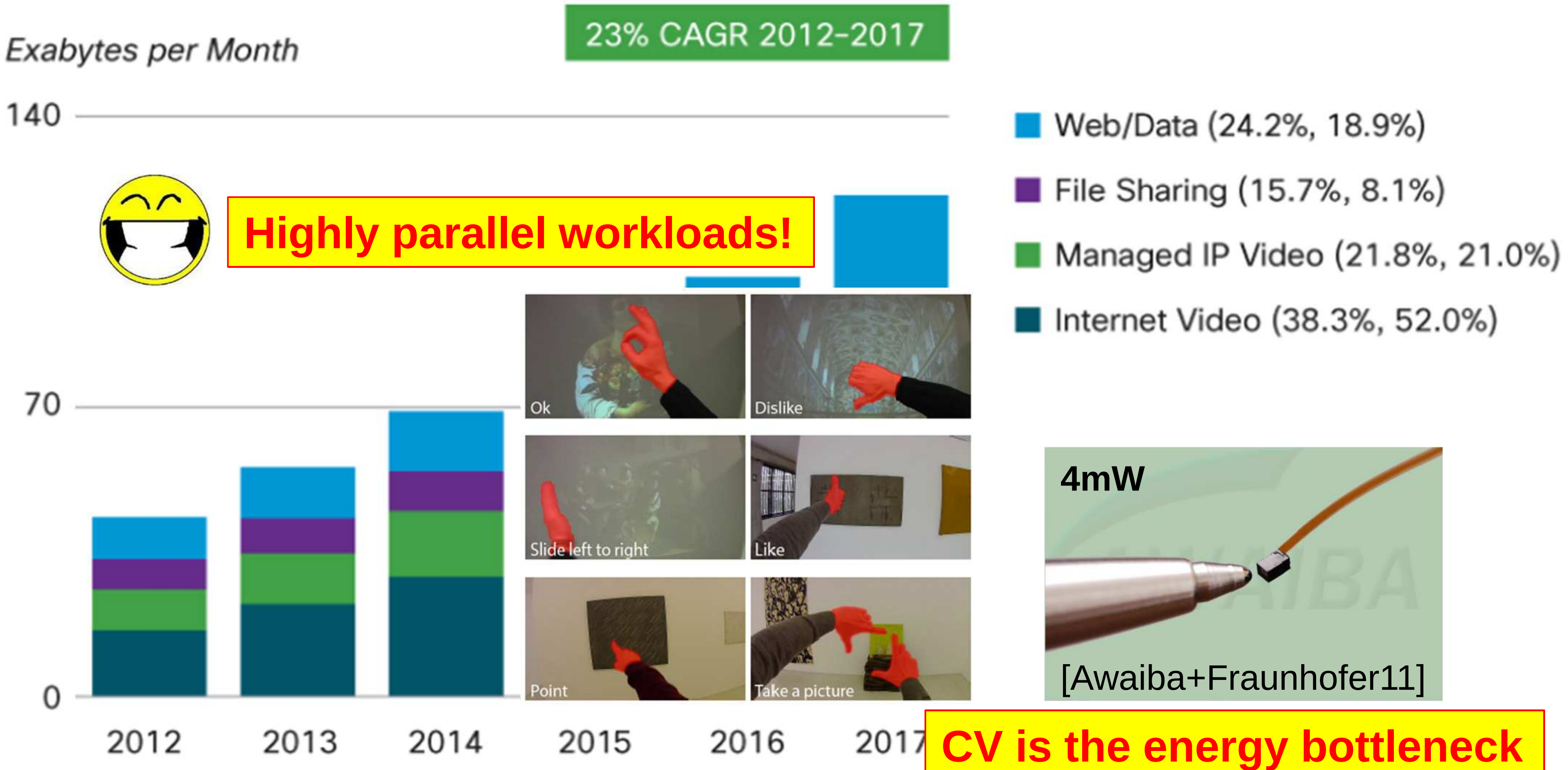
DEI-UNIBO & IIS-ETHZ

Multithermand AdG
Multiscale Thermal Management of Computing Systems

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

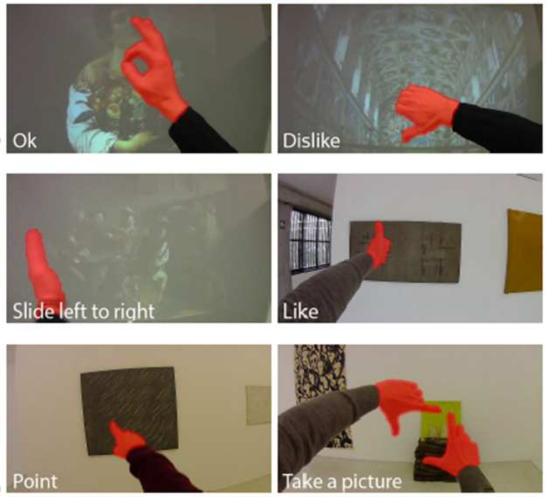
IOT or Data Deluge?

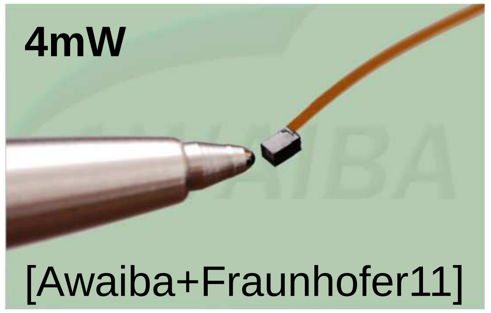
ALMA MATER STUDIORUM



Source: Cisco VNI, 2013
The percentages within parenthesis next to the legend denote the relative traffic shares in 2012 and 2017.

Highly parallel workloads!





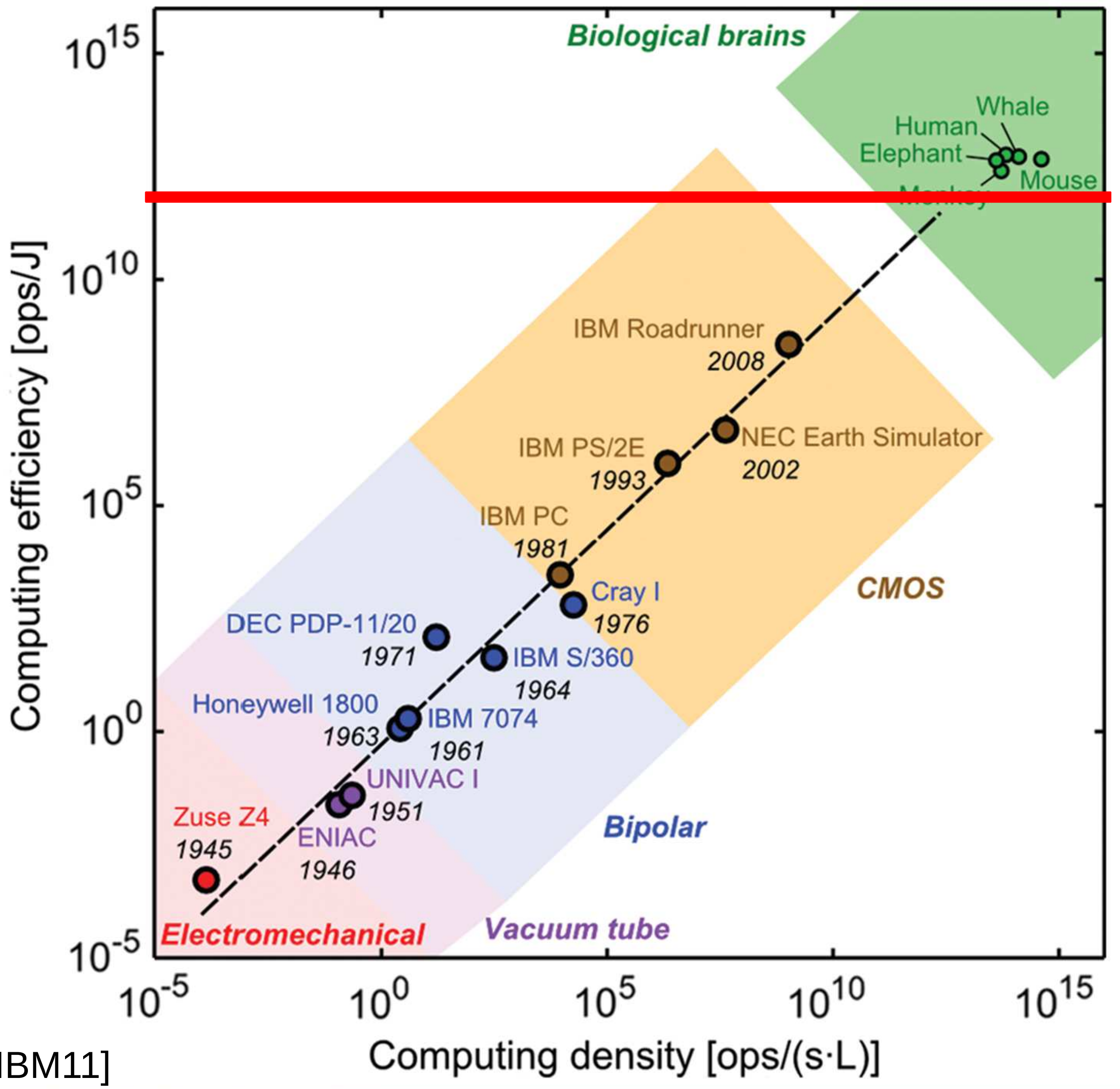
CV is the energy bottleneck

In-situ stream processing & fusion + visual intelligence is a must!
- In a few mW power envelope!!

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

How efficient?

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA



10^{12}ops/J

↓

1pJ/op

↓

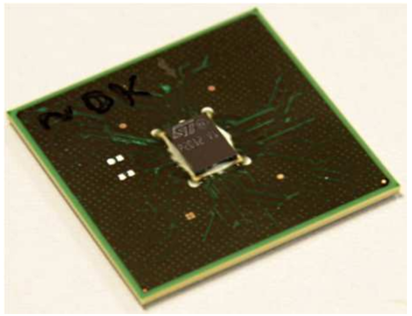
1GOPS/mW

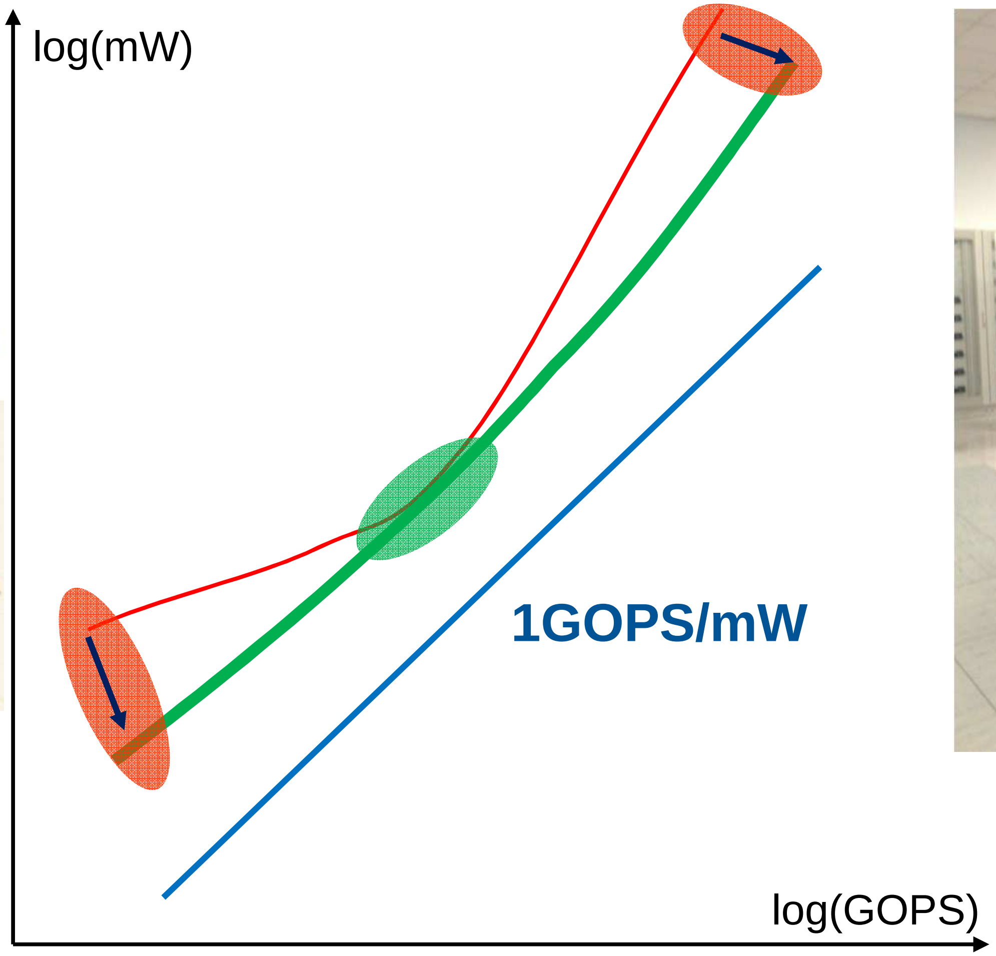
[RuchIBM11]

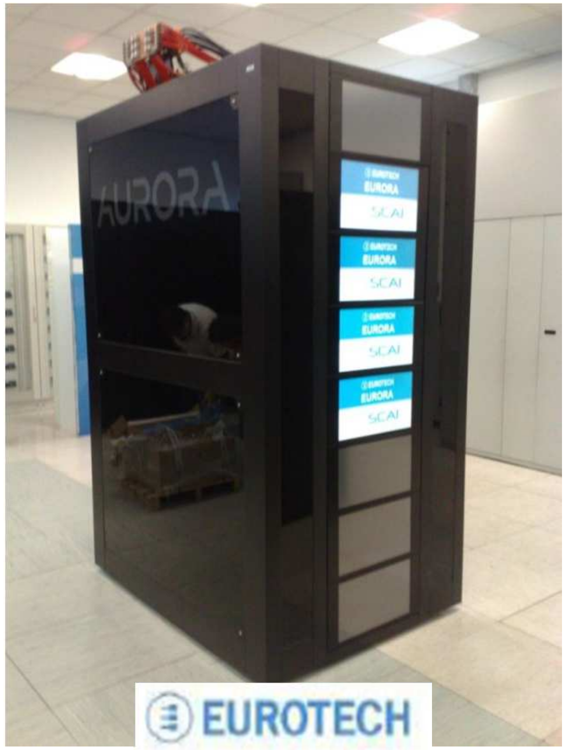
The challenge of Energy Proportionality

0,003GOPS/mW – 30KW

0,04GOPS/mW







From KOPS (10^3) to EOPS (10^{18})!

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

A Short Review on CMOS power
and power minimization

Where is Power Dissipated in CMOS?

- Active (Dynamic) power
  - (Dis)charging capacitors
  - Short-circuit power
    - Both pull-up and pull-down on during transition
- Static (leakage) power
  - Transistors are imperfect switches
- Static currents
  - Biasing currents

Active (or Dynamic) Power

Key property of active power:

$$P_{dyn} \propto f$$

with *f* the switching frequency

Sources:

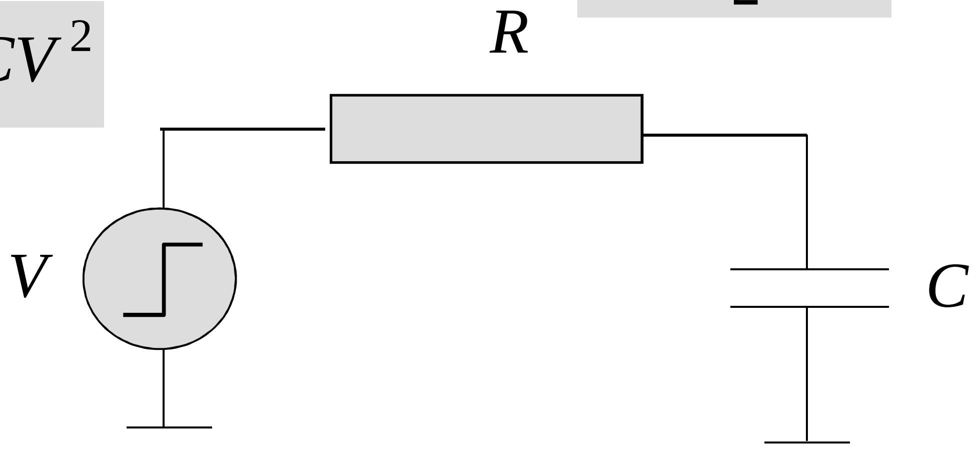
- Charging and discharging capacitors
- Temporary glitches (dynamic hazards)
- Short-circuit currents

Charging Capacitors

Applying a voltage step

$$E_R = \frac{1}{2}CV^2$$

$$E_{0\rightarrow 1} = CV^2$$

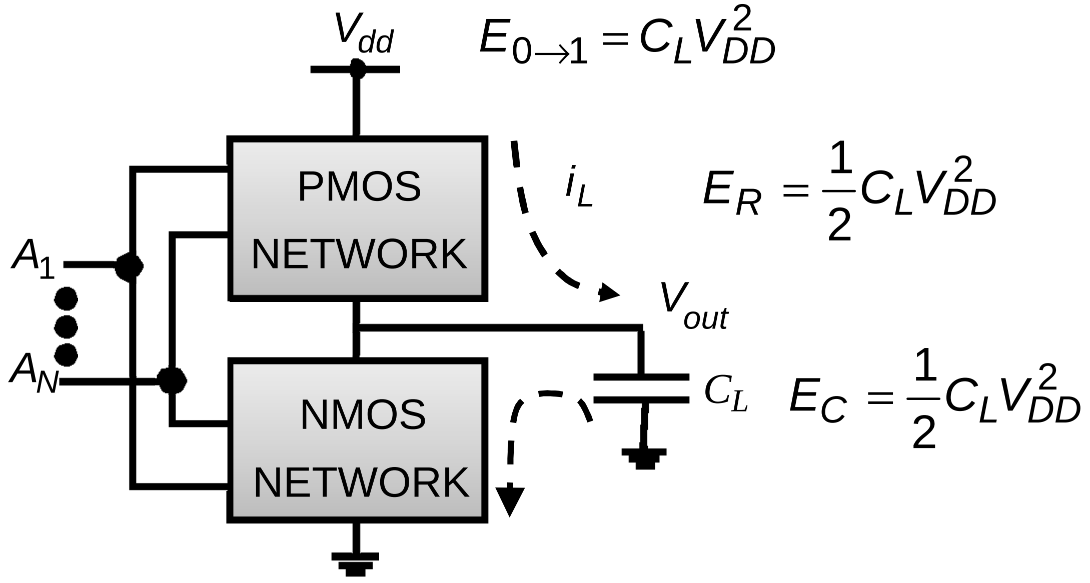


$$E_C = \frac{1}{2}CV^2$$

$$E_{0\rightarrow 1} = \int_0^\infty VC\frac{dV_C}{dt}dt = CV\int_0^V dV_C = CV^2$$

Value of *R* does not impact energy!

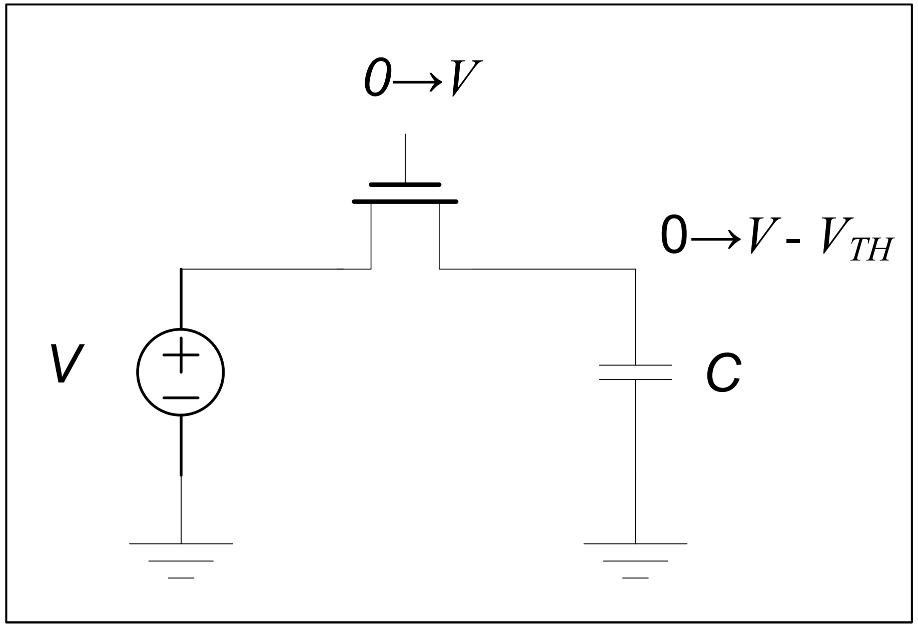
Applied to Complementary CMOS Gate



V_{dd}
$E_{0\rightarrow 1} = C_L V_{DD}^2$
PMOS NETWORK
i_L
$E_R = \frac{1}{2} C_L V_{DD}^2$
A_1
A_N
V_{out}
NMOS NETWORK
C_L
$E_C = \frac{1}{2} C_L V_{DD}^2$

- One half of the power from the supply is consumed in the pull-up network and one half is stored on C_L
- Charge from C_L is dumped during the $1\rightarrow 0$ transition
- Independent of resistance of charging/discharging network

Circuits with Reduced Swing



$$E_{0\to1} = \int_0^{\infty} VC\frac{dV_C}{dt}dt = CV\int_0^{V-V_T} dV_C = CV(V - V_{TH})$$

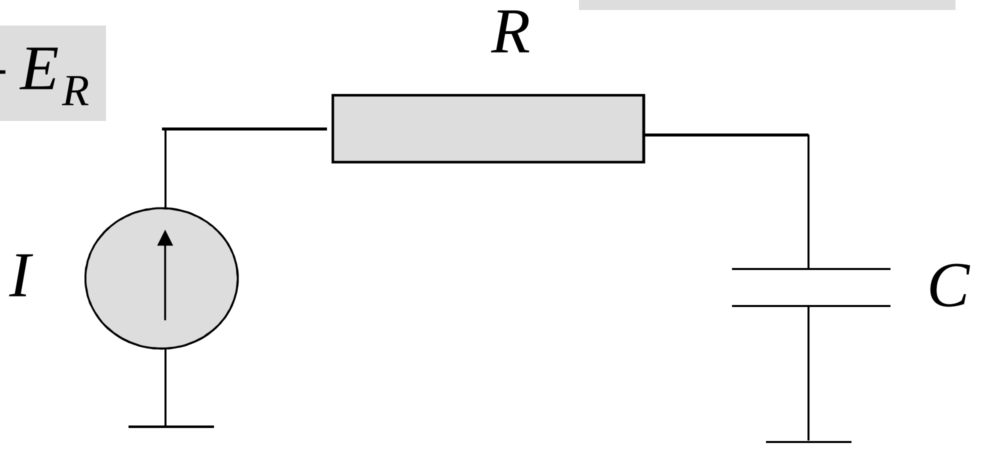
Energy consumed is proportional to output swing

Charging Capacitors - Revisited

Driving from a constant current source

$$E_R = \left(\frac{RC}{T}\right)CV^2$$

$$E_{0\rightarrow 1} = E_C + E_R$$



$$E_C = \frac{1}{2}CV^2$$

$$T = \frac{CV}{I}$$

$$E_R = \int_0^{\infty} I(RI)dt = RI^2T = \left(\frac{RC}{T}\right)CV^2$$

Energy dissipated in resistor can be reduced
by increasing charging time *T* (that is, decreasing *I*)

Charging Capacitors

Using constant voltage or current driver?

$$E_{constant\_current} < E_{constant\_voltage}$$

if

$$T > 2RC$$

Energy dissipated using constant current charging can be made arbitrarily small at the expense of delay: **Adiabatic charging**

Note: $t_p(\mathrm{RC}) = 0.69\,RC$

$t_{0\rightarrow 90\%}(\mathrm{RC}) = 2.3\,RC$

Dynamic Power Consumption

Power = Energy/transition • Transition rate

$$= C_L V_{DD}^2 \bullet f_{0\to 1}$$

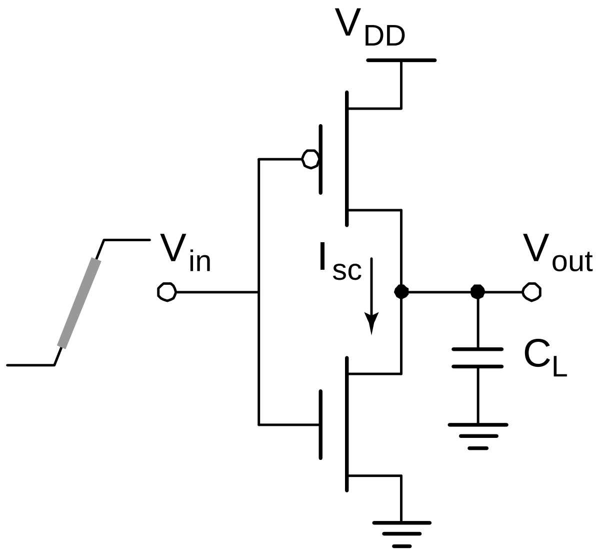
$$= C_L V_{DD}^2 \bullet f \bullet P_{0\to 1}$$

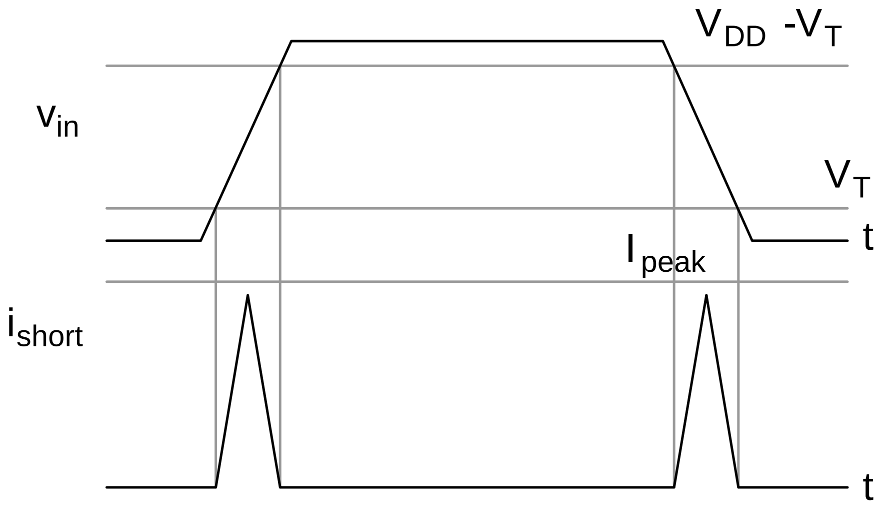
$$= C_{switched} V_{DD}^2 \bullet f$$

- Power dissipation is data dependent – depends on the switching probability
- Switched capacitance $C_{switched} = P_{0\to 1} C_L = \alpha\, C_L$ (α is called the switching activity)

Short-Circuit Currents

(also called crowbar currents)

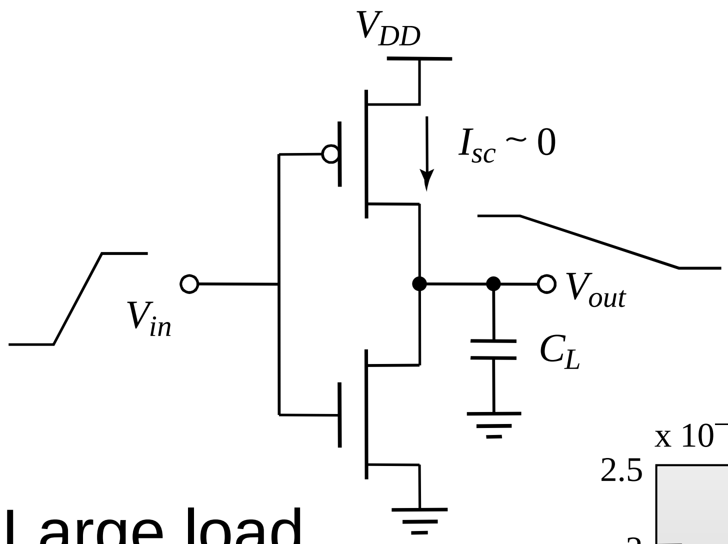




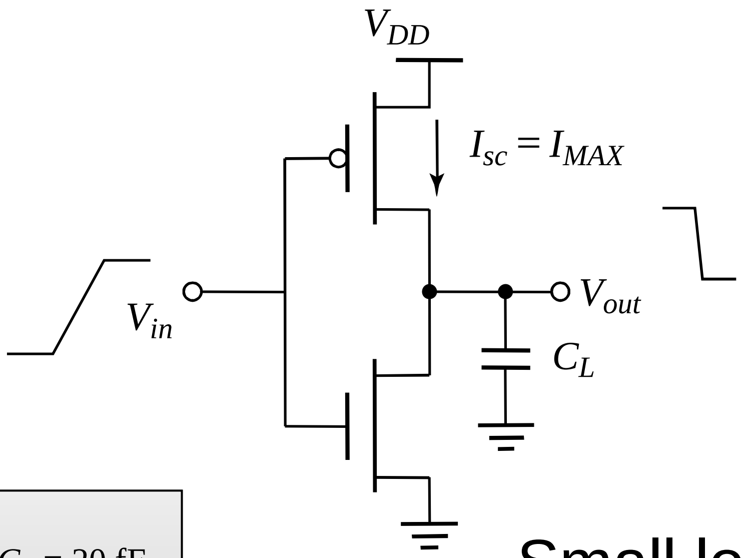
PMOS and NMOS simultaneously on during transition

$$P_{sc} \sim f$$

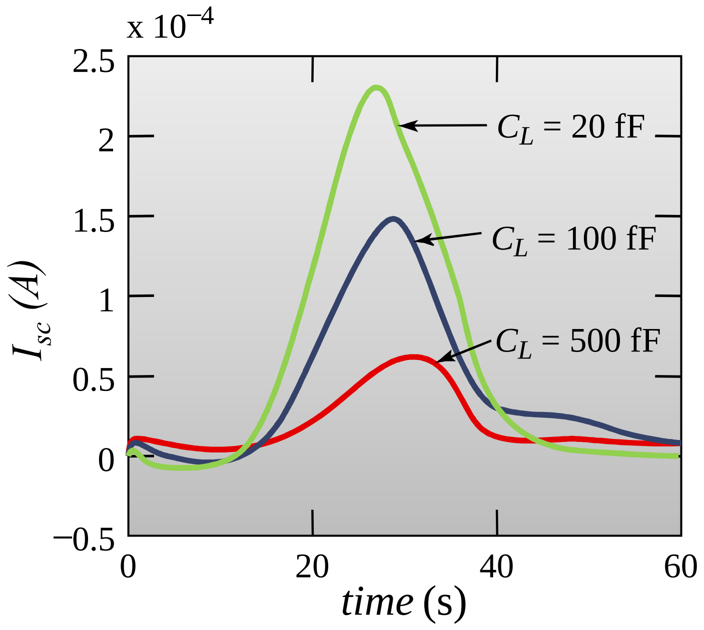
Short-Circuit Currents



Large load



Small load



Equalizing rise/fall times of input and output signals limits P_{sc} to 10-15% of the dynamic dissipation

[Ref: H. Veendrick, JSSC'84]

Modeling Short-Circuit Power

- Can be modeled as capacitor

$$C_{SC} = k\left(a\frac{\tau_{in}}{\tau_{out}} + b\right)$$

a, b: technology parameters

k: function of supply and threshold voltages, and transistor sizes

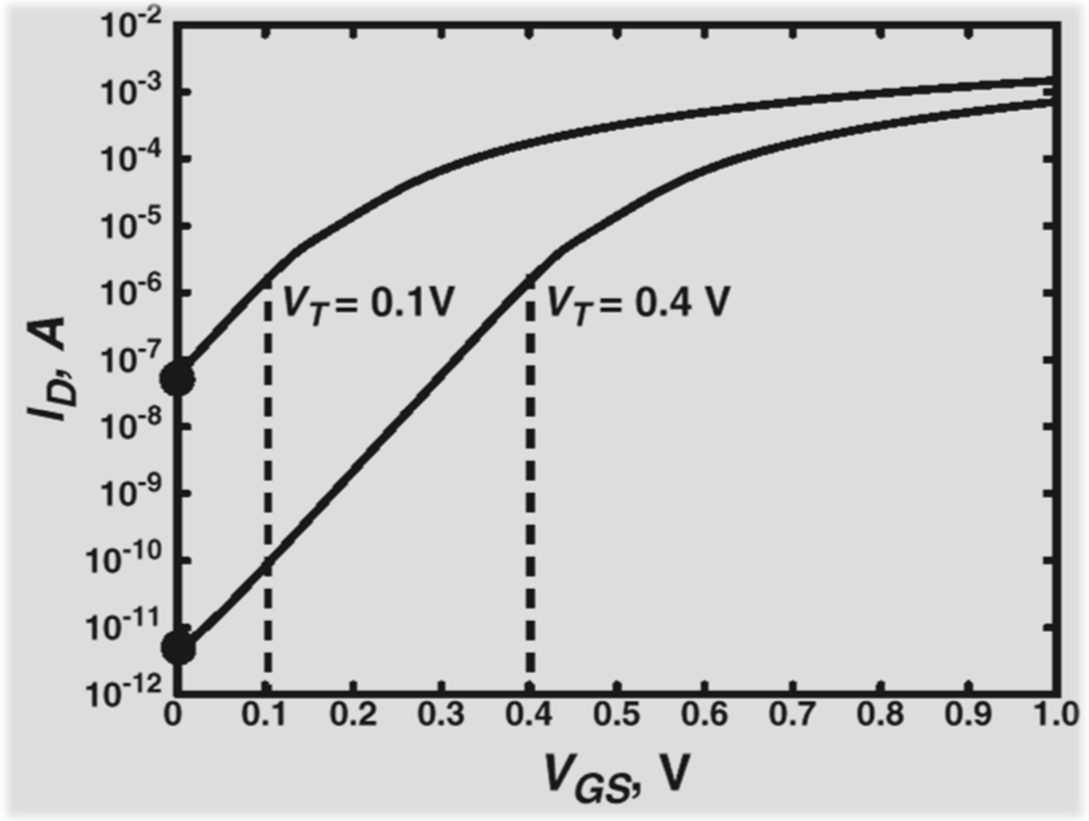
$$E_{SC} = C_{SC} V_{DD}^{\;2}$$

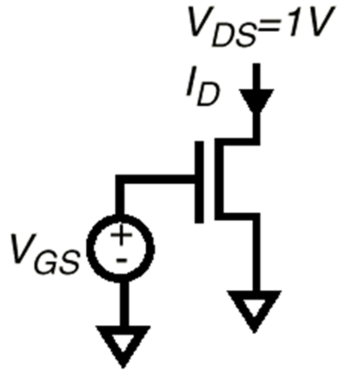
Easily included in timing and power models

Transistors Leak

- Drain leakage
  - – Diffusion currents
  - – Drain-induced barrier lowering (DIBL)
- Junction leakages
  - – Gate-induced drain leakage (GIDL)
- Gate leakage
  - – Tunneling currents through thin oxide

Sub-threshold Leakage





Off-current increases exponentially when reducing V_{TH}

$$I_{leak} = I_0 \frac{W}{W_0} 10^{\frac{-V_{TH}}{S}} \Rightarrow P_{leak} = V_{DD}.I_{leak}$$

Sub-Threshold Leakage

Leakage current increases with drain voltage (mostly due to DIBL)

$$I_{leak} = I_0 \frac{W}{W_0} 10^{\frac{-V_{TH} + \lambda_d V_{DS}}{S}}$$ (for $V_{DS} > 3\ kT/q$)

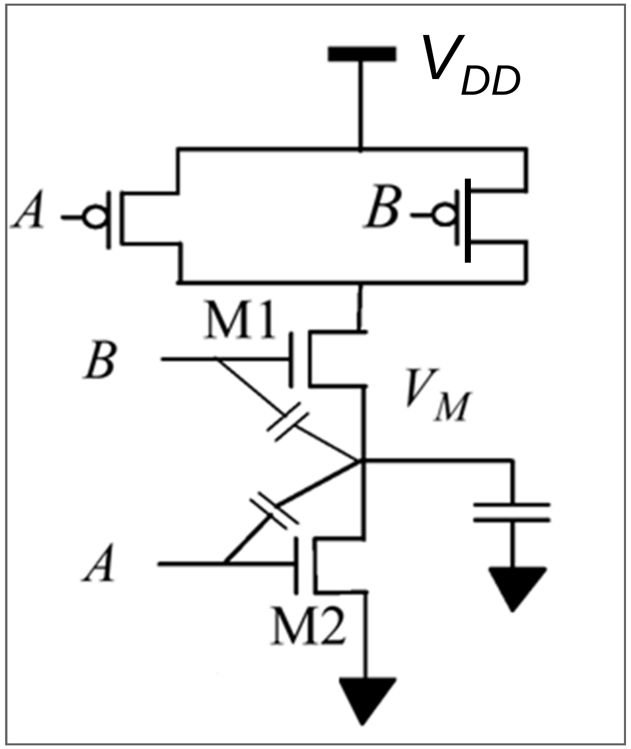
Hence

$$P_{leak} = (I_0 \frac{W}{W_0} 10^{\frac{-V_{TH}}{S}})(V_{DD} 10^{\frac{\lambda_d V_{DD}}{S}})$$

Leakage Power strong function of supply voltage

Stack Effect

NAND gate:



Assume that body effect in short channel transistor is small

$$I_{leak,M1} = I_0' 10^{\frac{-V_M - V_{TH} + \lambda_d (V_{DD} - V_M)}{S}}$$

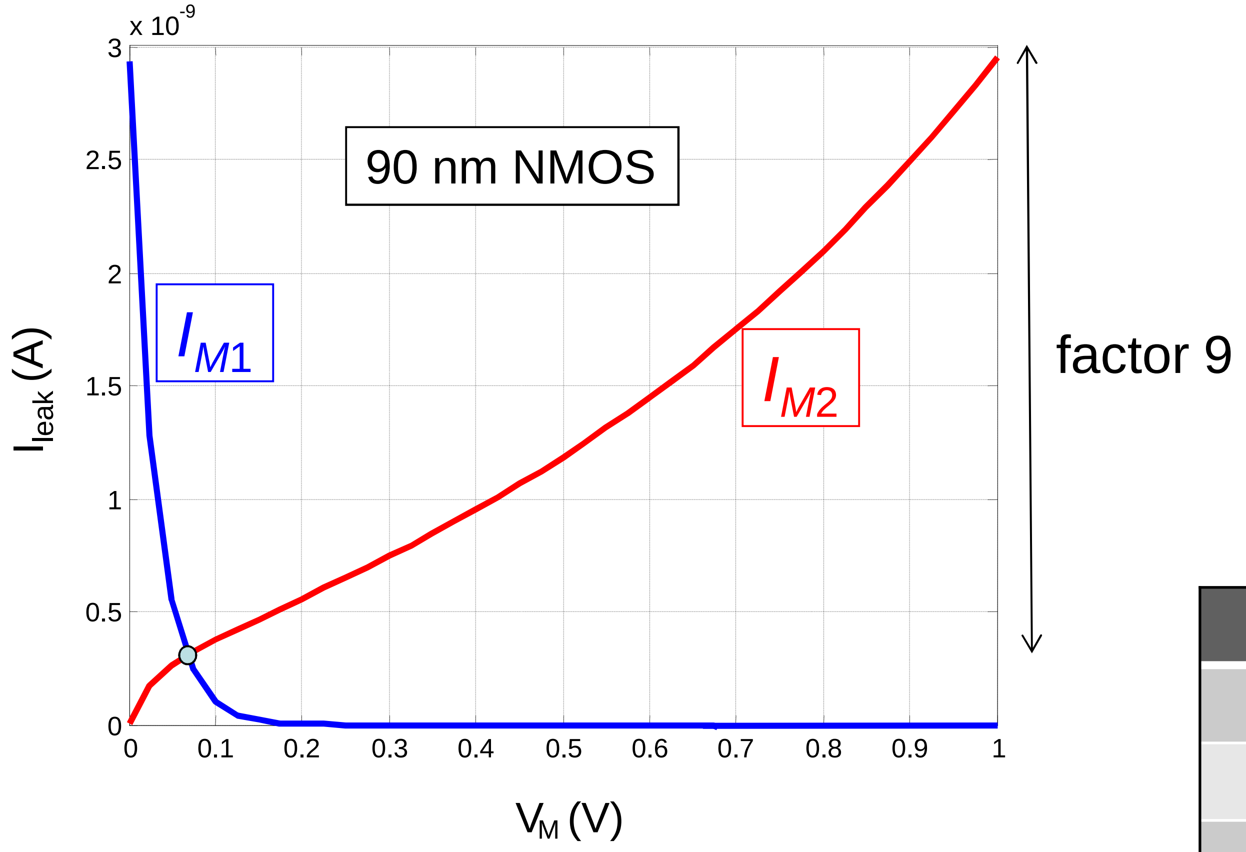
$$I_{leak,M2} = I_0' 10^{\frac{-V_{TH} + \lambda_d V_M}{S}}$$

$$V_M \approx \frac{\lambda_d}{1 + 2\lambda_d} V_{DD}$$

$$\frac{I_{stack}}{I_{inv}} \approx 10^{-\frac{\lambda_d V_{DD}}{S}\left(\frac{1+\lambda_d}{1+2\lambda_d}\right)}$$

(instead of the expected factor of 2)

Stack Effect

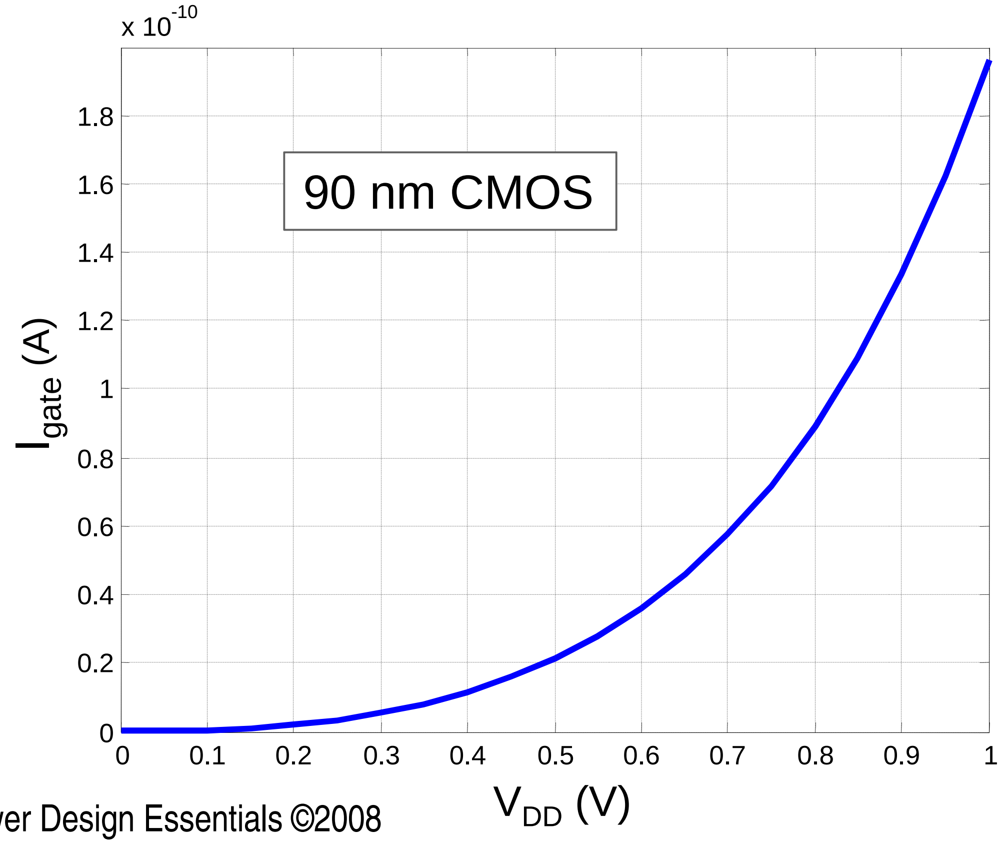


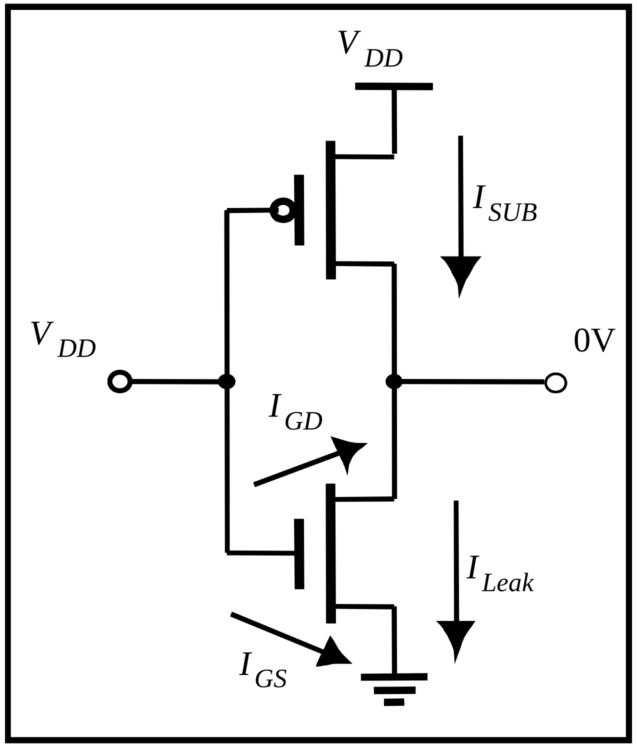
Leakage Reduction	
2 NMOS	9
3 NMOS	17
4 NMOS	24
2 PMOS	8
3 PMOS	12
4 PMOS	16

Gate Tunneling

Exponential function of supply voltage

- $I_{GD} \sim e^{-Tox} e^{VGD}$, $I_{GS} \sim e^{-Tox} e^{VGS}$
- Independent of the sub-threshold leakage





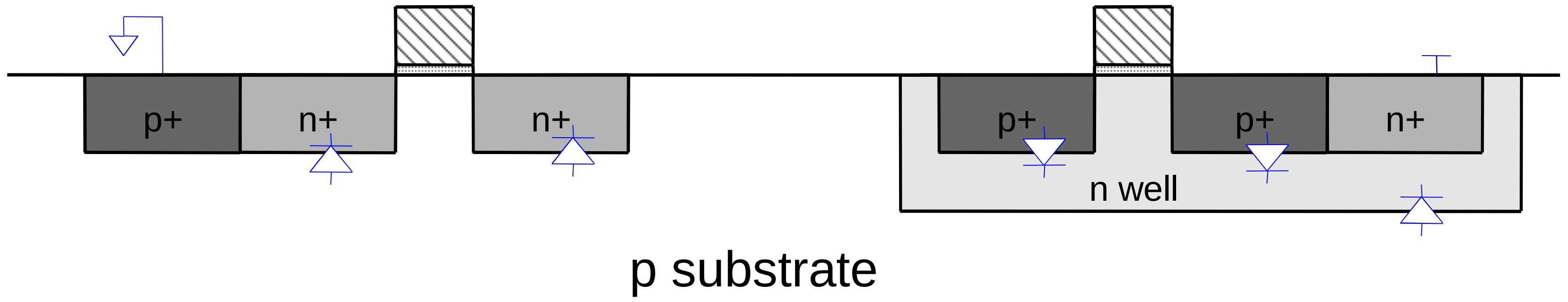
Modeled in BSIM4

Also in BSIM3v3 (but not always included in foundry models)

NMOS gate leakage usually worse than PMOS

Other sources of static power dissipation

- Diode (drain-substrate) reverse bias currents



• Electron-hole pair generation in depletion region of reverse-biased diodes
• Diffusion of minority carriers through junction
• For sub-50nm technologies with highly-doped pn junctions, **tunneling through narrow depletion region** becomes an issue

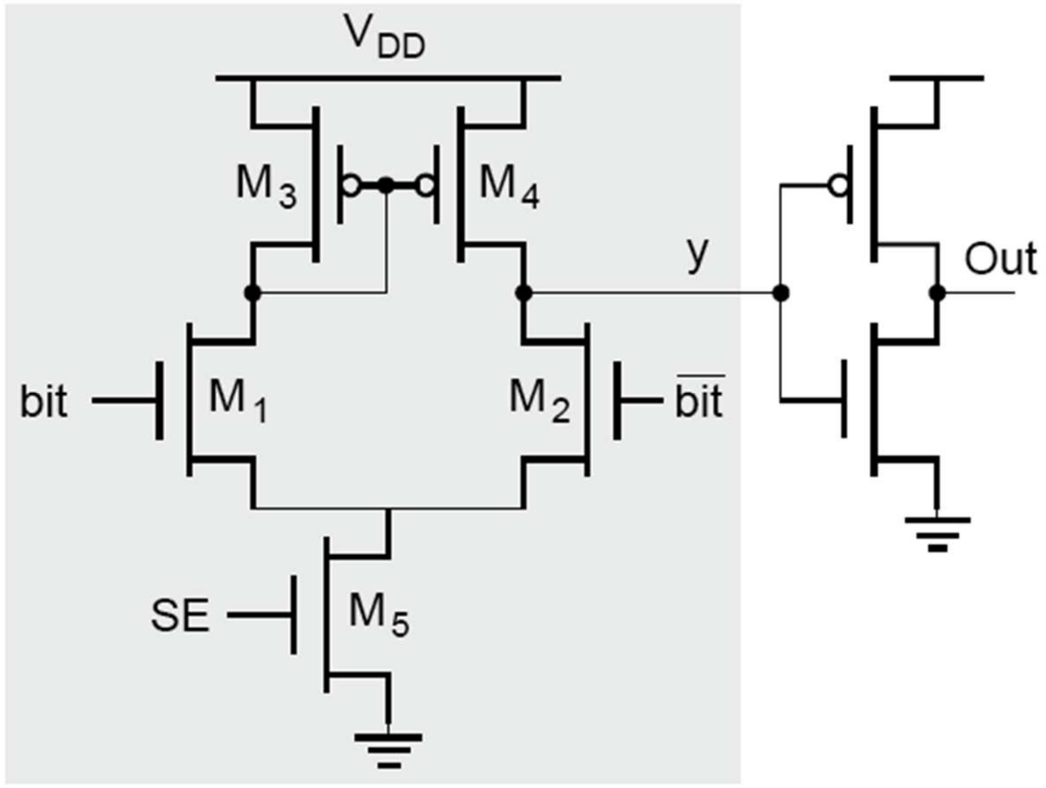
Strong function of temperature

Much smaller than other leakage components in general

Other sources of static power dissipation

- Circuit with dc bias currents:

sense amplifiers, voltage converters and regulators, sensors, mixed-signal components, etc



Should be turned off if not used, or standby current should be minimized

Summary of Power Dissipation Sources

$$P \sim \alpha \cdot \left(C_L + C_{CS}\right) \cdot V_{swing} \cdot V_{DD} \cdot f + \left(I_{DC} + I_{Leak}\right) \cdot V_{DD}$$

- α – switching activity
- C_L – load capacitance
- C_{CS} – short-circuit capacitance
- V_{swing} – voltage swing
- f – frequency

- I_{DC} – static current
- I_{leak} – leakage current

$$P = \frac{energy}{operation} \times rate + static\ \ power$$

The Traditional Design Philosophy

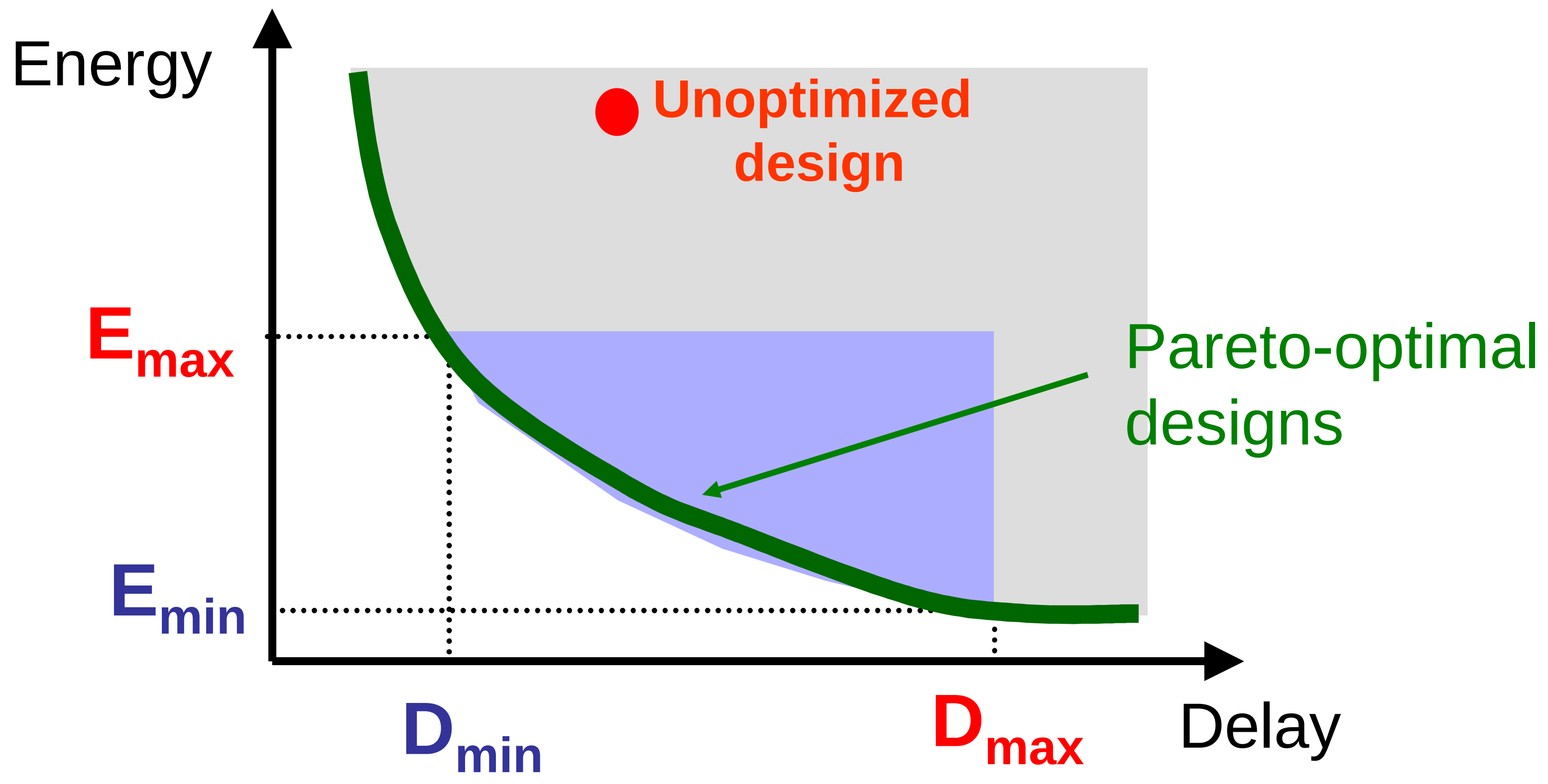
- Maximum performance is primary goal
  - Minimum delay at circuit level
- Architecture implements the required function with target throughput, latency
- Performance achieved through optimum sizing, logic mapping, architectural transformations.
- Supplies, thresholds set to achieve maximum performance, subject to reliability constraints

The New Design Philosophy

- Maximum performance (in terms of propagation delay) is too power-hungry, and/or not even practically achievable
- Many (if not most) applications either can tolerate larger latency, or can live with lower than maximum clock-speeds
- Excess performance (as offered by technology) to be used for energy/power reduction

Trading off speed for power

Exploring the Energy-Delay Space



In energy-constrained world, design is trade-off process

- ♦ Minimize energy for a given performance requirement
- ♦ Maximize performance for given energy budget

[Ref: D. Markovic, JSSC'04]

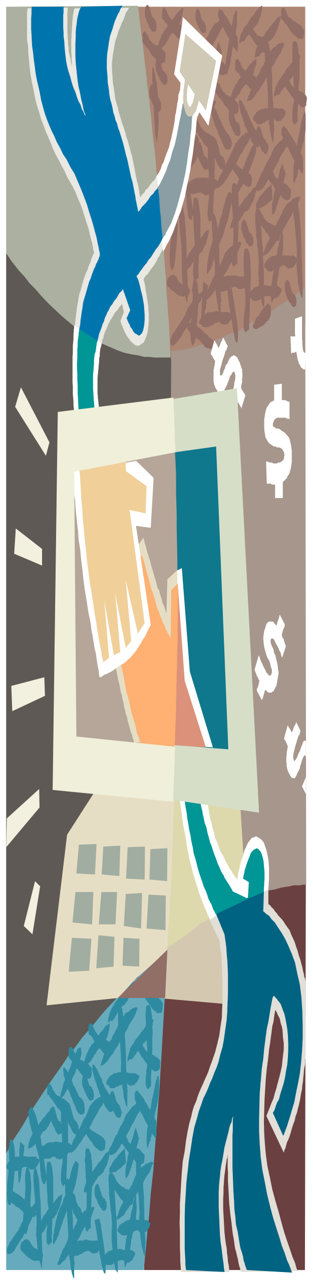
Summary

- Power and energy are now primary design constraints
- Active power still dominating for most applications
  - Supply voltage, activity and capacitance the key parameters
- Leakage becomes major factor in sub-100nm technology nodes
  - Mostly impacted by supply and threshold voltages
- Design has become energy-delay trade-off exercise!

Reducing power @ all design levels

- Algoritmic level
- Compiler level
- Architecture level
- Organization level
- Circuit level
- Silicon level

- Important concepts:
  - Lower Vdd and freq. (even if errors occur) / dynamically adapt Vdd and freq.
  - Reduce circuit
  - Exploit locality
  - Reduce switching activity, glitches, etc.



Algoritmic level

- The best indicator for energy is …..
  …. the number of cycles

- Try alternative algorithms with lower complexity
  - E.g. quick-sort, O(n log n) $\Leftrightarrow$ bubble-sort, O (n^2)
  - … but be aware of the 'constant' : O(n log n) $\Rightarrow$ c*(n log n)

- Heuristic approach
  - Go for a good solution, not the best !!

Biggest gains at this level !!

Compiler level

- Source-to-Source transformations
  - loop trafo's to improve locality
- Strength reduction
  - E.g. replace Const * A with Add's and Shift's
  - Replace Floating point with Fixed point
- Reduce register pressure / number of accesses to register file
  - Use software bypassing
- Scenarios: current workloads are highly dynamic
  - Determine and predict execution modes
  - Group execution modes into scenarios
  - Perform special optimizations per scenario
    - DFVS: Dynamic Voltage and Frequency Scaling
    - More advanced loop optimizations
- Reorder instructions to reduce bit-transistions

Architecture level

- Going parallel
- Going heterogeneous
  - tune your architecture, exploit SFUs (special function units)
  - trade-off between flexibility / programmability / genericity and efficiency
- Add local memories
  - prefer scratchpad i.s.o. cache
- Cluster FUs and register files (see next slide)
- Reduce bit-width
  - sub-word parallelism (SIMD)

Organization (micro-arch.) level

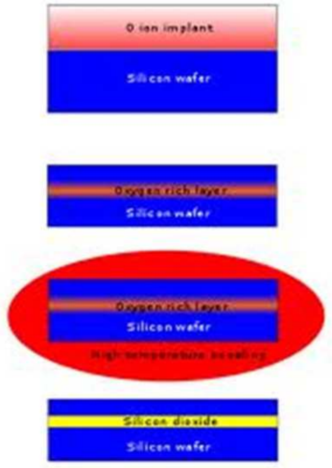
- Enabling Vdd reduction
  - Pipelining
    - cheap way of parallelism
  - Enabling lower freq. $\Rightarrow$ lower V_{dd}
  - Note 1: don't pipeline if you don't need the performance
  - Note 2: don't exaggerate (like the 31-stage Pentium 4)
- Reduce register traffic
  - avoid unnecessary reads and write
  - make bypass registers visible

Circuit level

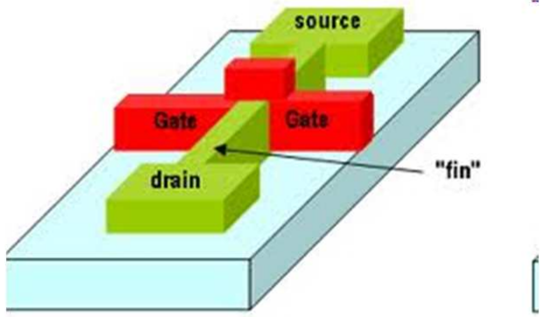
- Clock gating
- Power gating
- Multiple Vdd modes
- Reduce glitches: balancing digital path's
- Exploit Zeros
- Special SRAM cells
  - normal SRAM can not scale below Vdd = 0.7 - 0.8 Volt
- Razor method; replay
- Allow errors and add redundancy to architectural invisible structures
  - branch predictor
  - caches
- .. and many more ..

Silicon level

- Higher V_t (V_threshold)
- Back Biasing control
  - see thesis Maurice Meijer (2011)
- SOI (Silicon on Insulator)
  - silicon junction is above an electr. insulator (silicon dioxide)
  - lowers parasitic device capacitance



- Better transistors: Finfet
  - multi-gate
  - reduce leakage (off-state curent)



- .. and many more

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

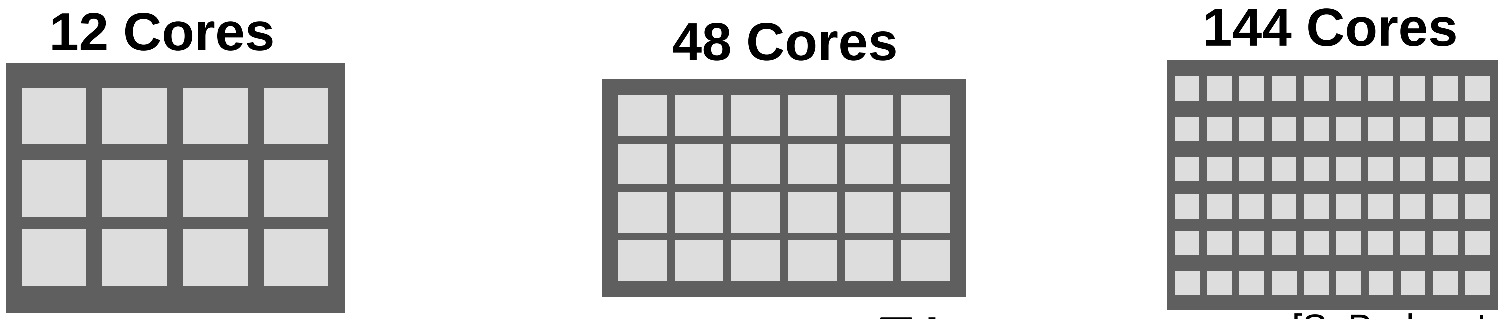
Two Ideas to Remember
…with their caveats

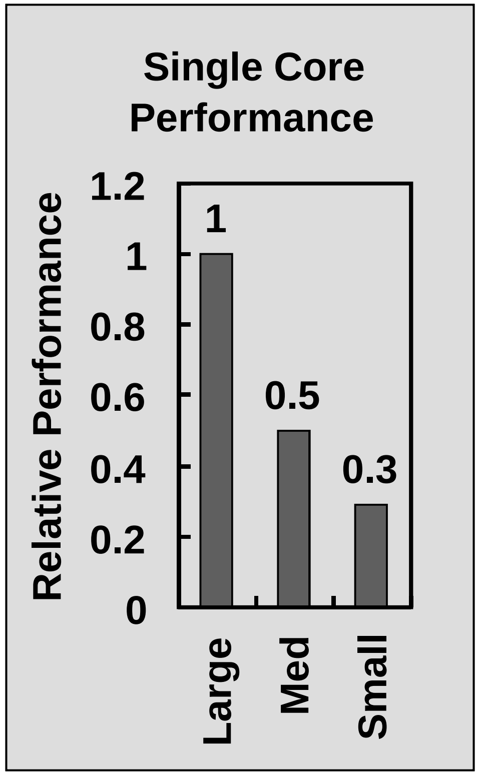
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Go Simple+Parallel

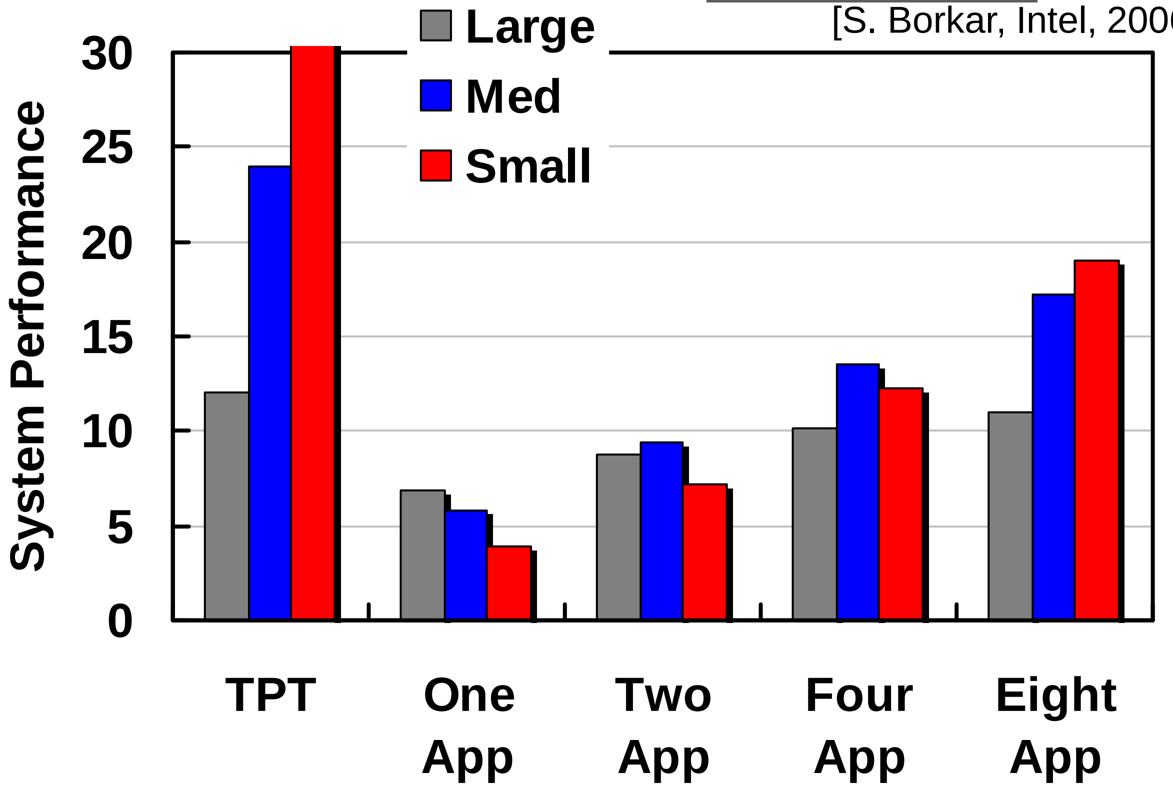
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

13mm, **100W**, 48MB Cache, 4B Transistors, in 22nm





[S. Borkar, Intel, 2006]

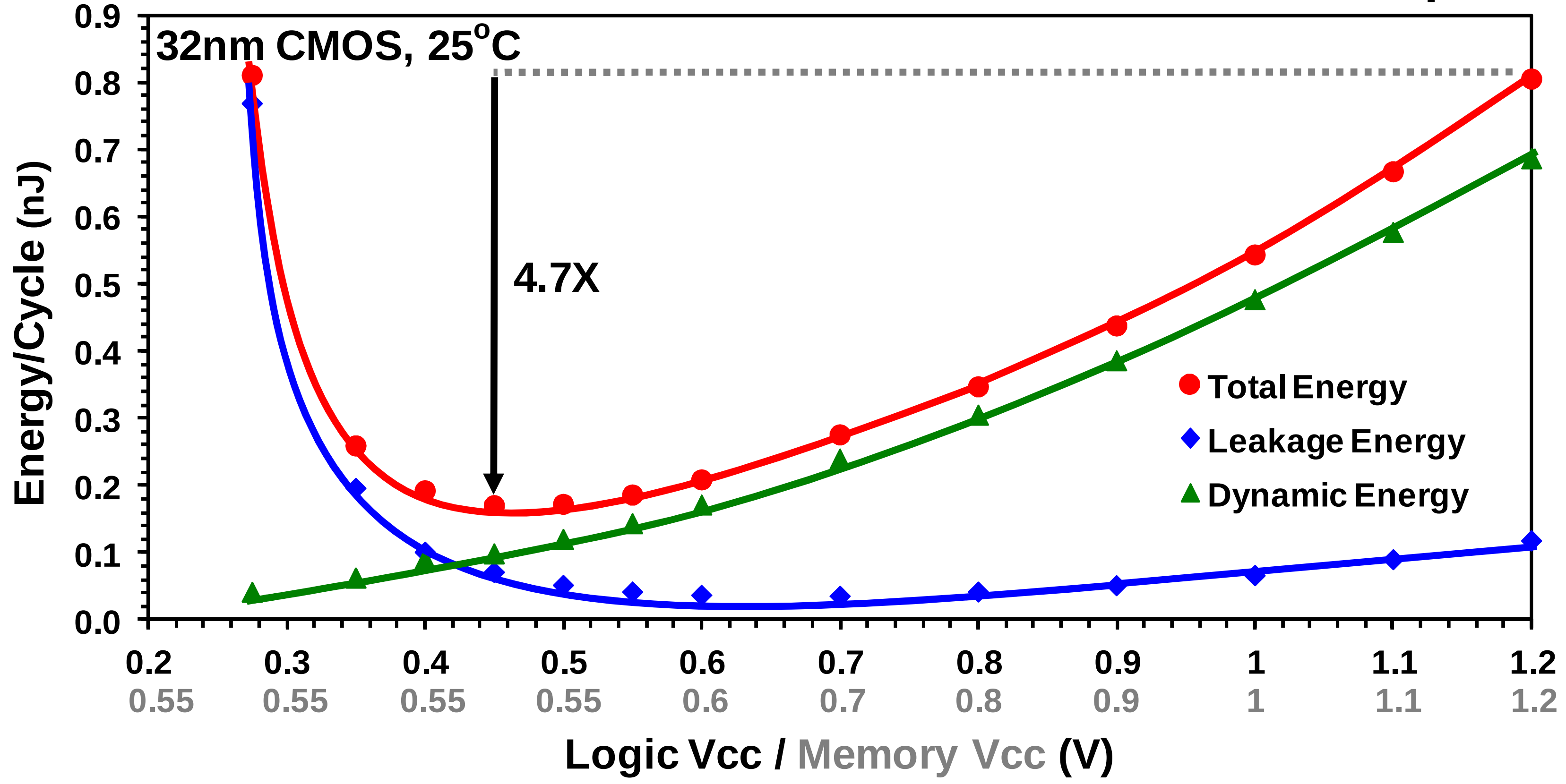


Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Lower Voltage Supply

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

[V. De, Intel, 2013]



…but be careful with Leakage and its variability!

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Introducing PULP

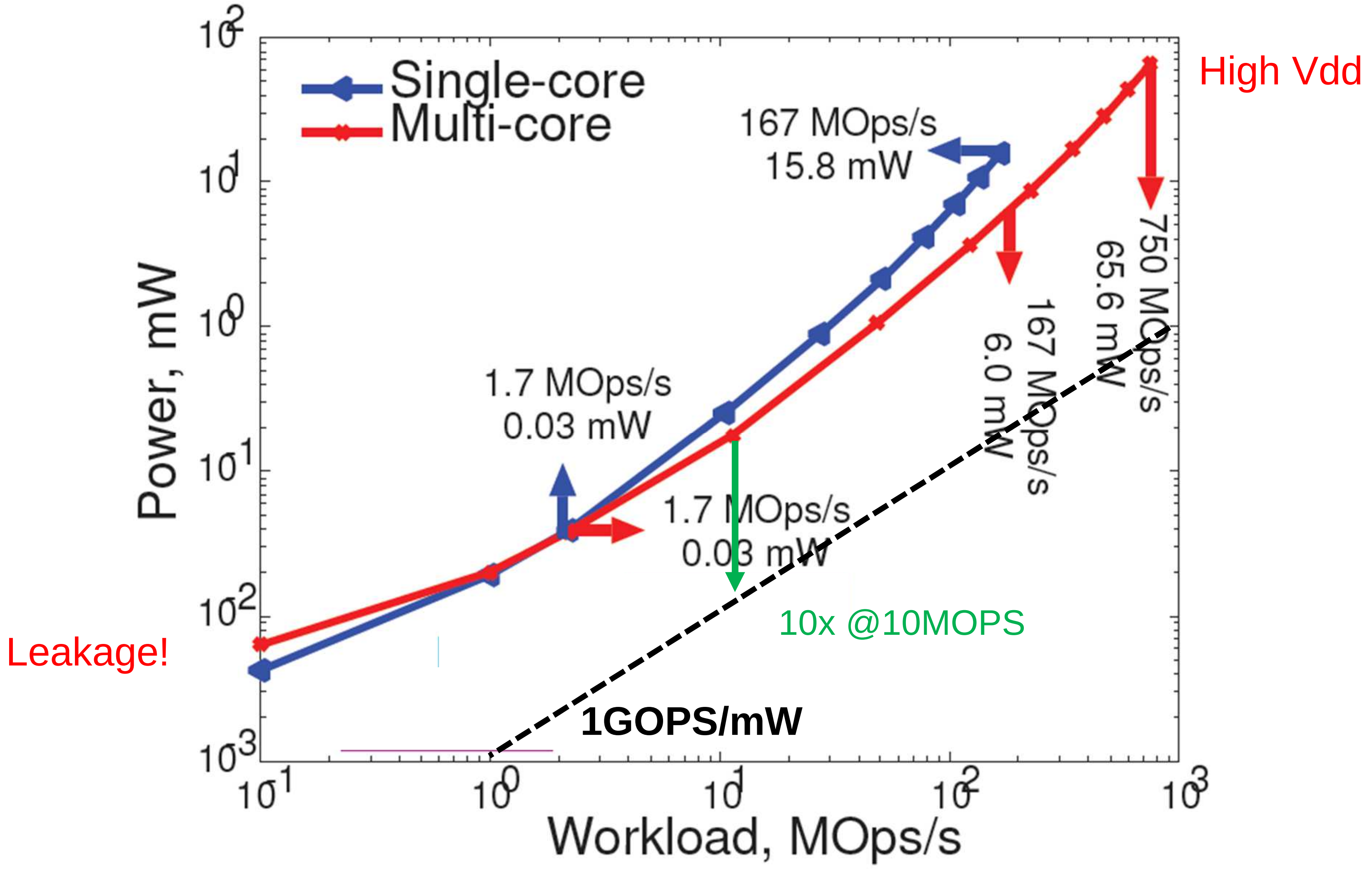
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

NTC Multicore?

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

2010 - With EPFL (Atienza, Burg)

90nm LP-CMOS

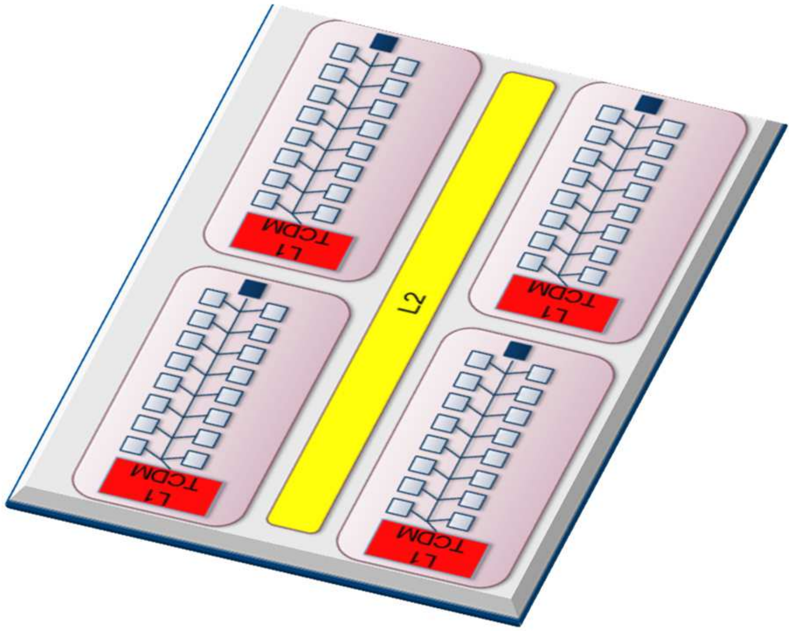


ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

A pJ/OP Parallel ULP Computing Platform

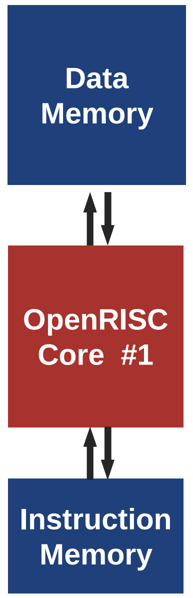
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

- **pJ/OP** is traditionally* the target of ASIC + uCntr
  - Scalable: [KOPS,TOPS], 32bit architecture
  - Flexible: OpenMP, OpenVX
  - Open: Software & HW



***1.57TOPS/W**: Kim et al., "A 1.22TOPS and 1.52mW/MHz Augmented Reality Multi-core Processor with Neural Network NoC for HMD Applications", ISSCC 2014

Making PULP

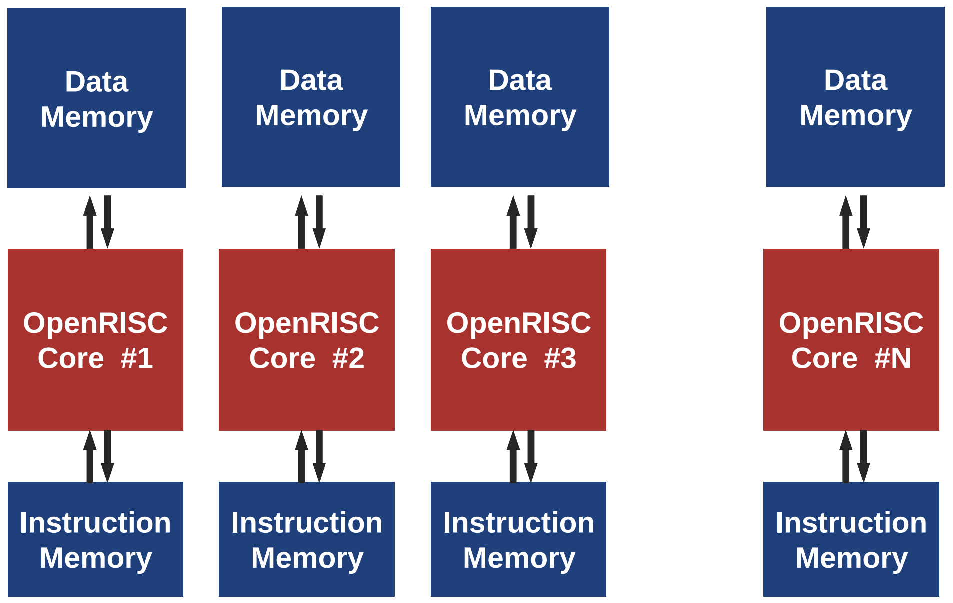


Start with an simple+efficient processor (~1IPC)

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

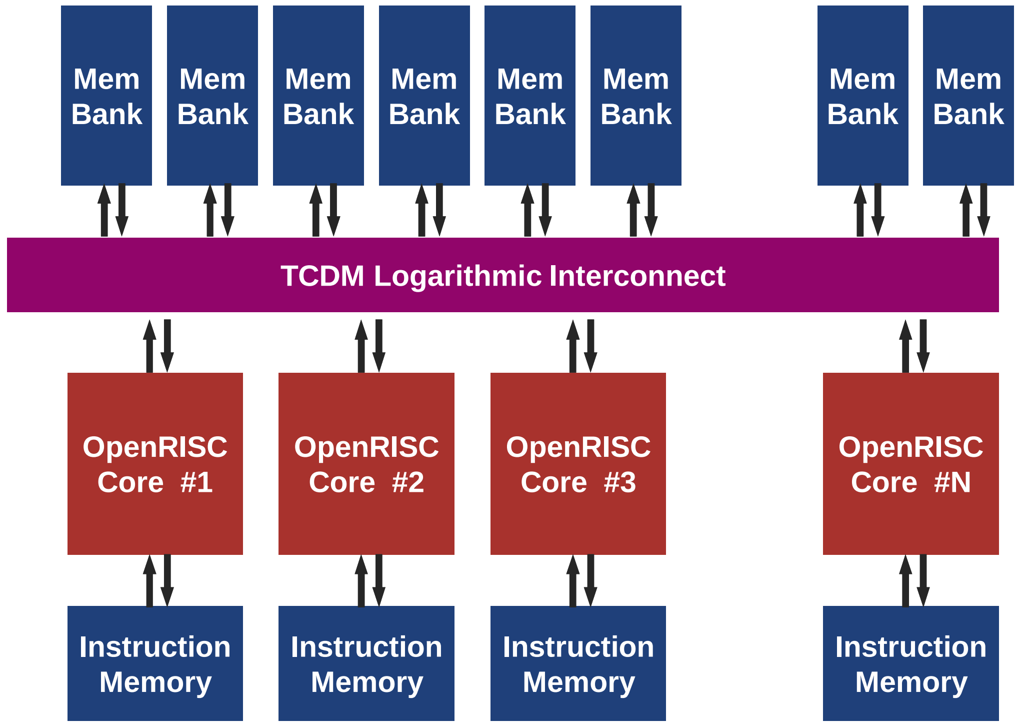
Making PULP

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA



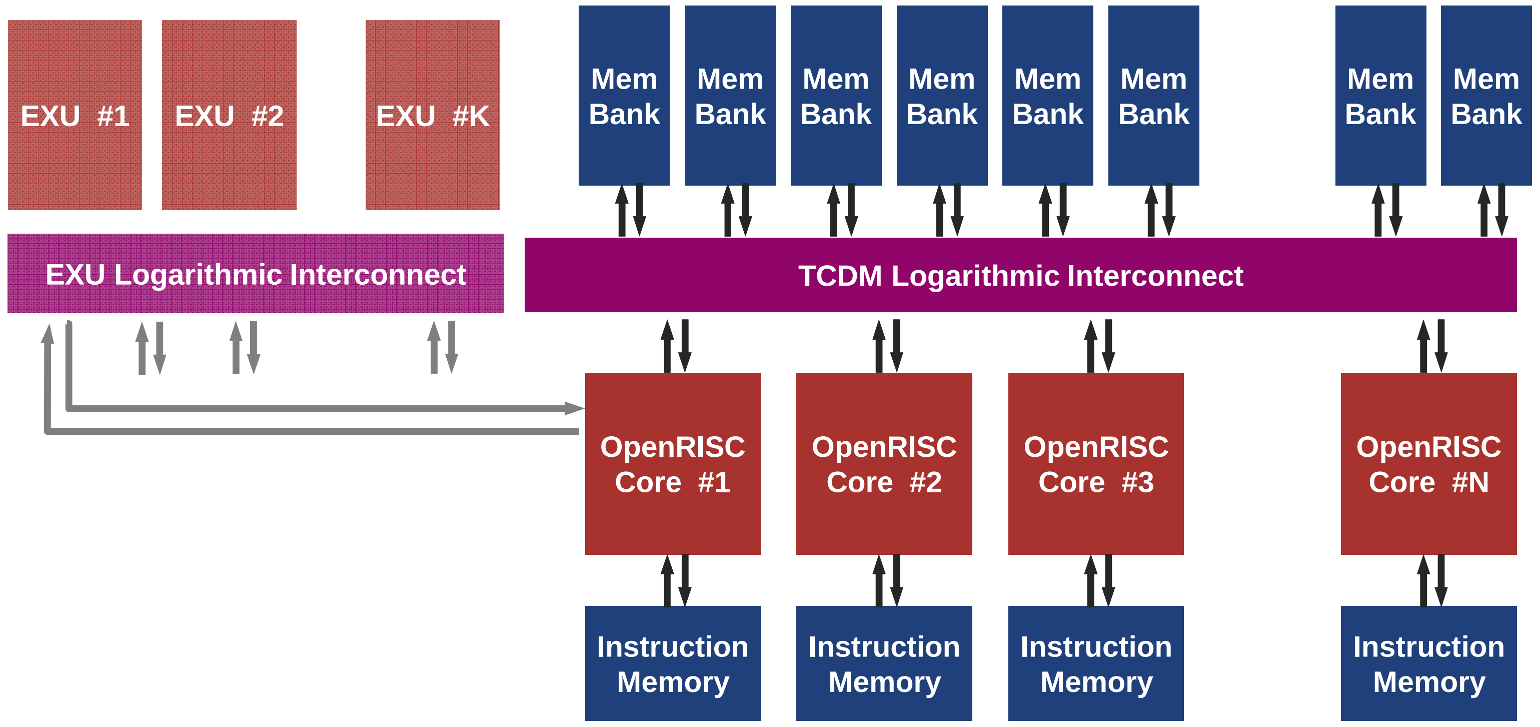
Parallel processors for performance @ NT

Making PULP



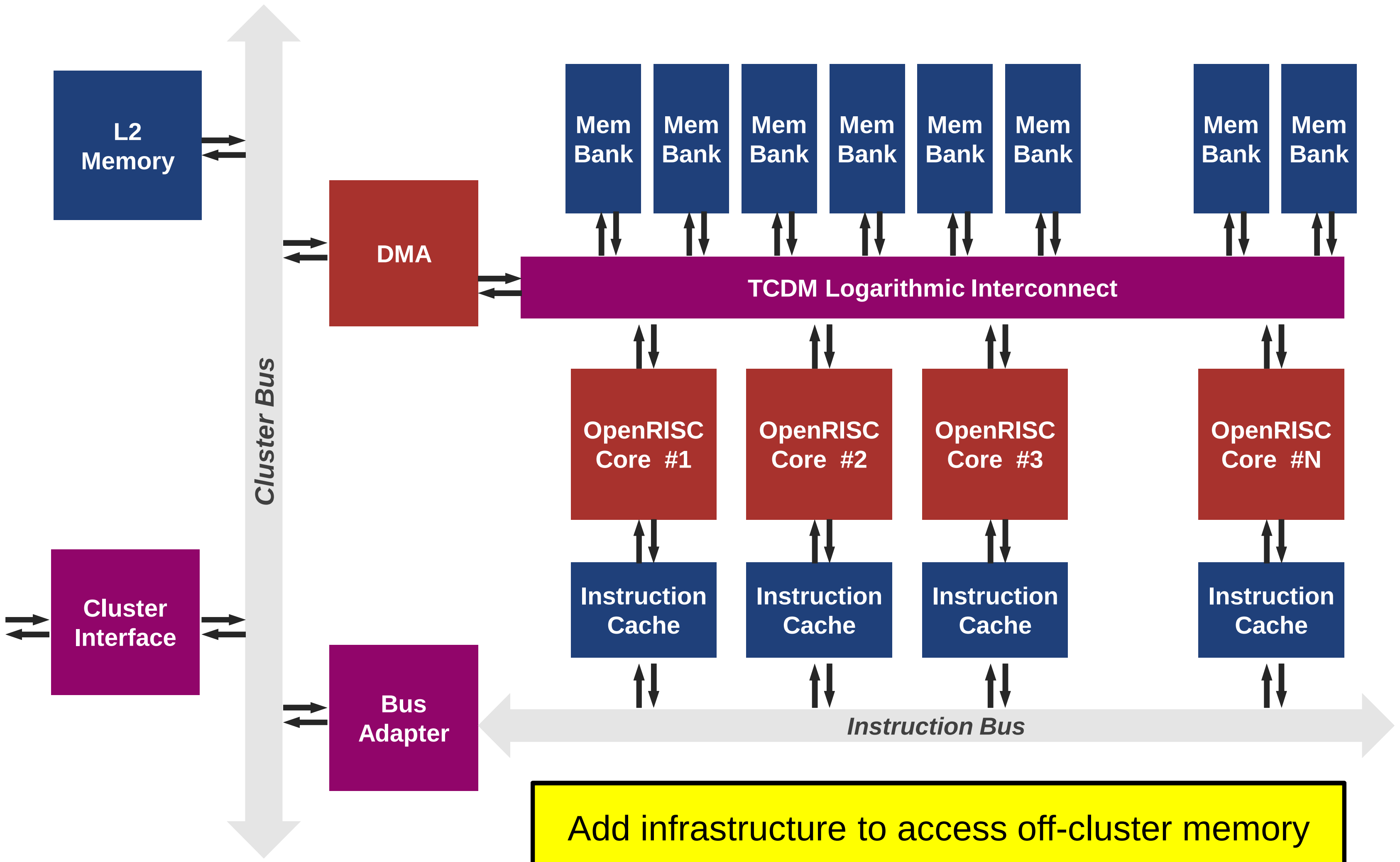
Parallel access to shared memory→Flexibility

Making PULP



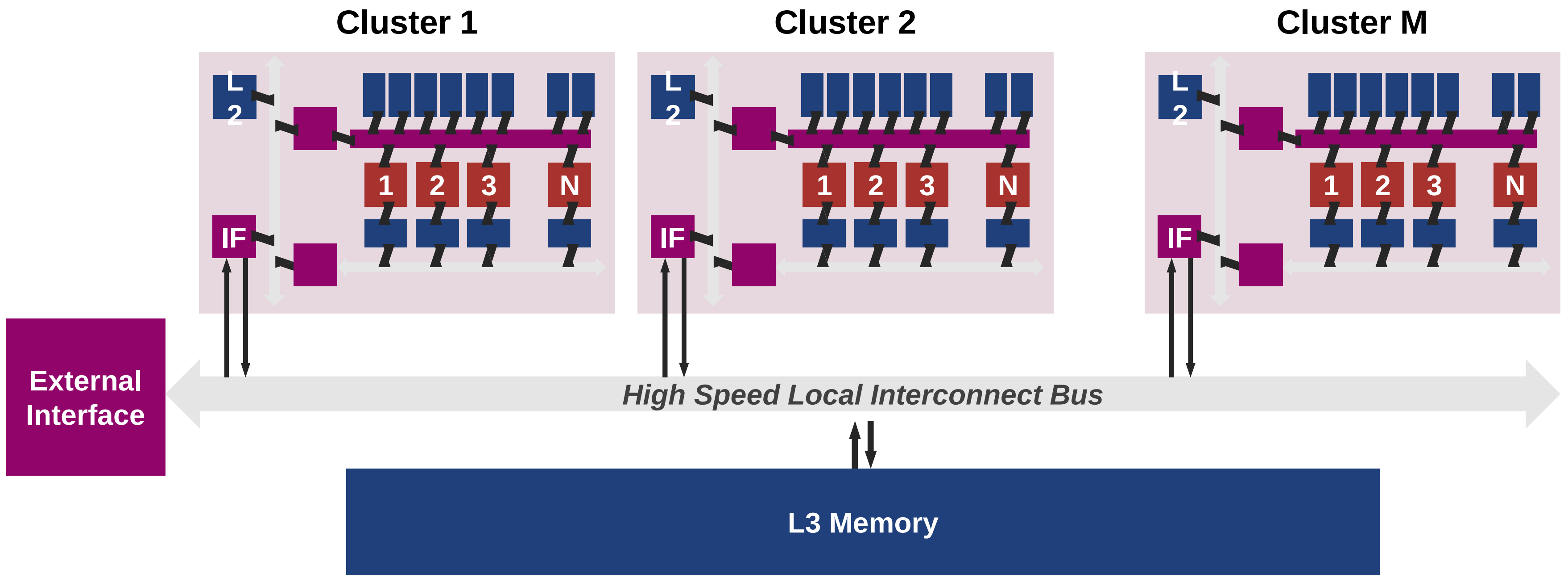
Optional Instruction Extension→Acceleration

Making PULP



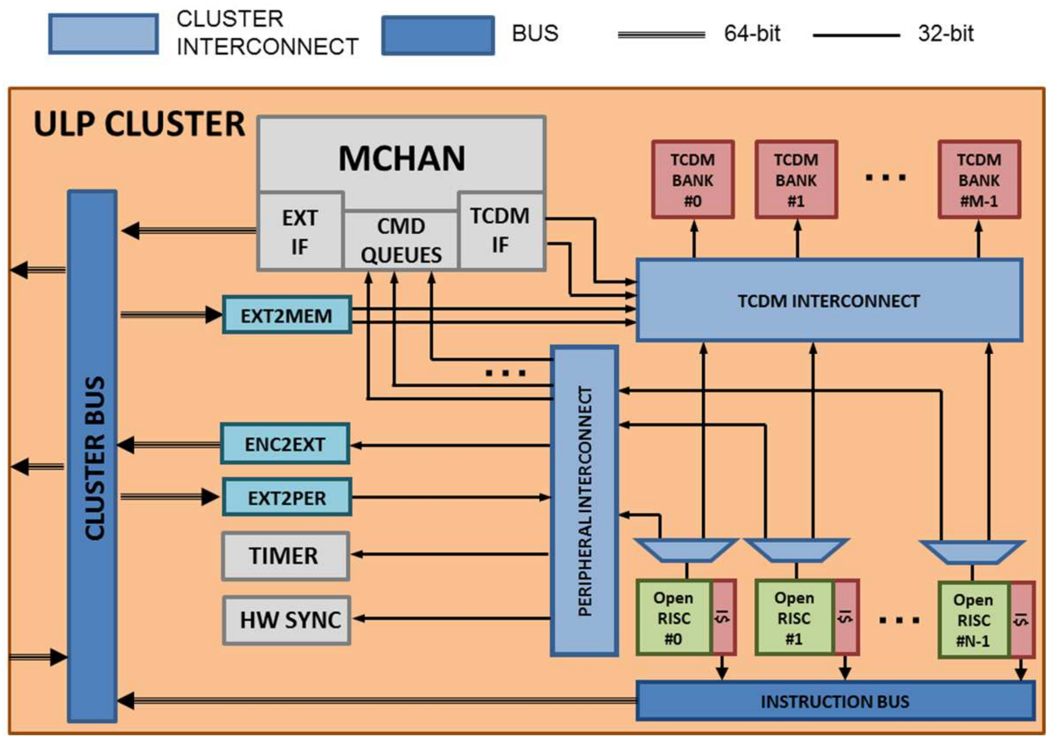
Add infrastructure to access off-cluster memory

Making PULP



Multiple clusters (f,Vdd,Vbb) form a PULP system

PULP Cluster



Design choices

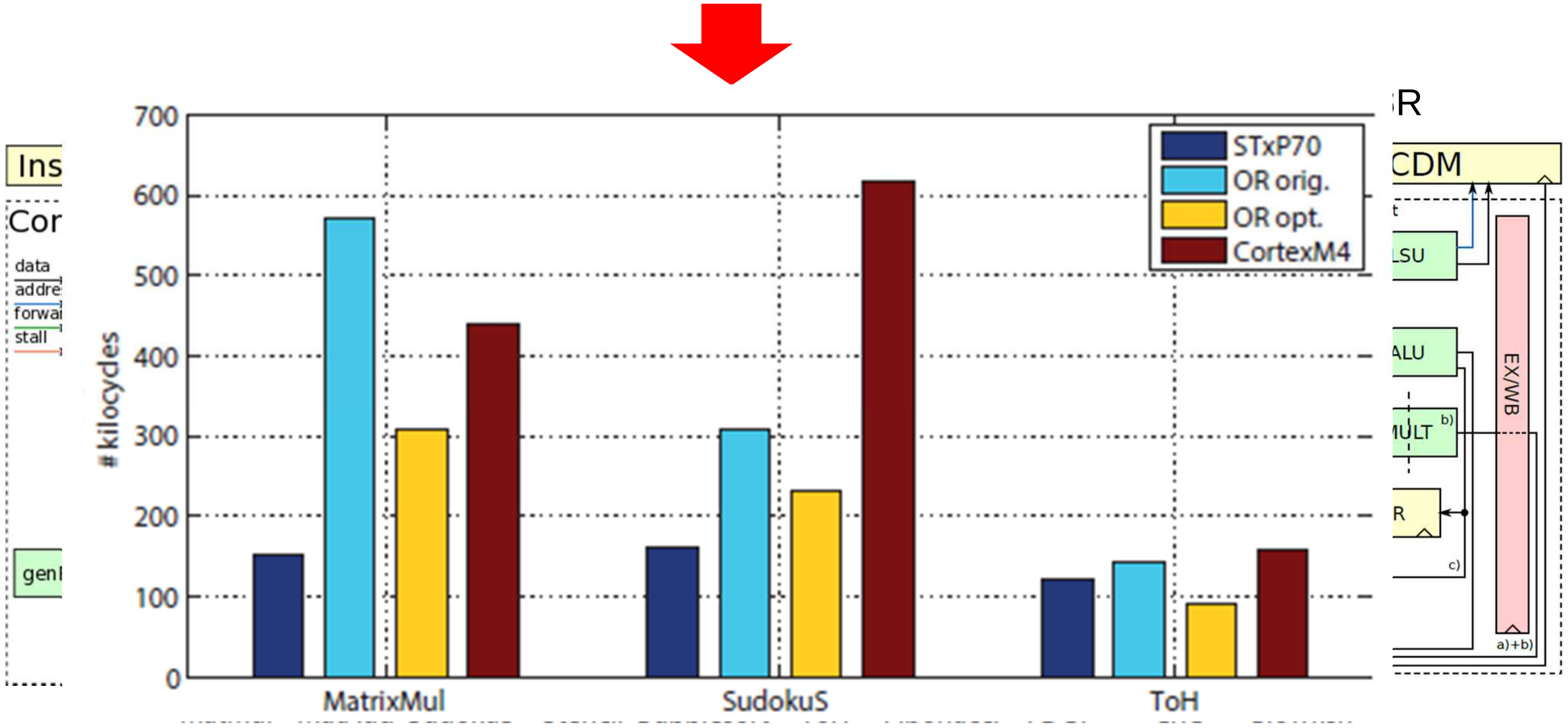
- **I$** → high code locality & simple architecture
- **No D$** → low locality & high complexity: $Bpmm^2_{D\$}/Bpmm^2_{DTCDM}<0{,}4$
- **Sharing L1** → less copies, easy work-balancing, low T_{clk} overhead in NT
- **Multibank** → smaller energy per access, “almost” multiported

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

OpenRisc Optimization

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

- Superpipelining harmful for energy efficiency
  - Focused speed optimization on the critical path dominated by MEM
- Low pipeline depth → high IPC with simple microarchitecture

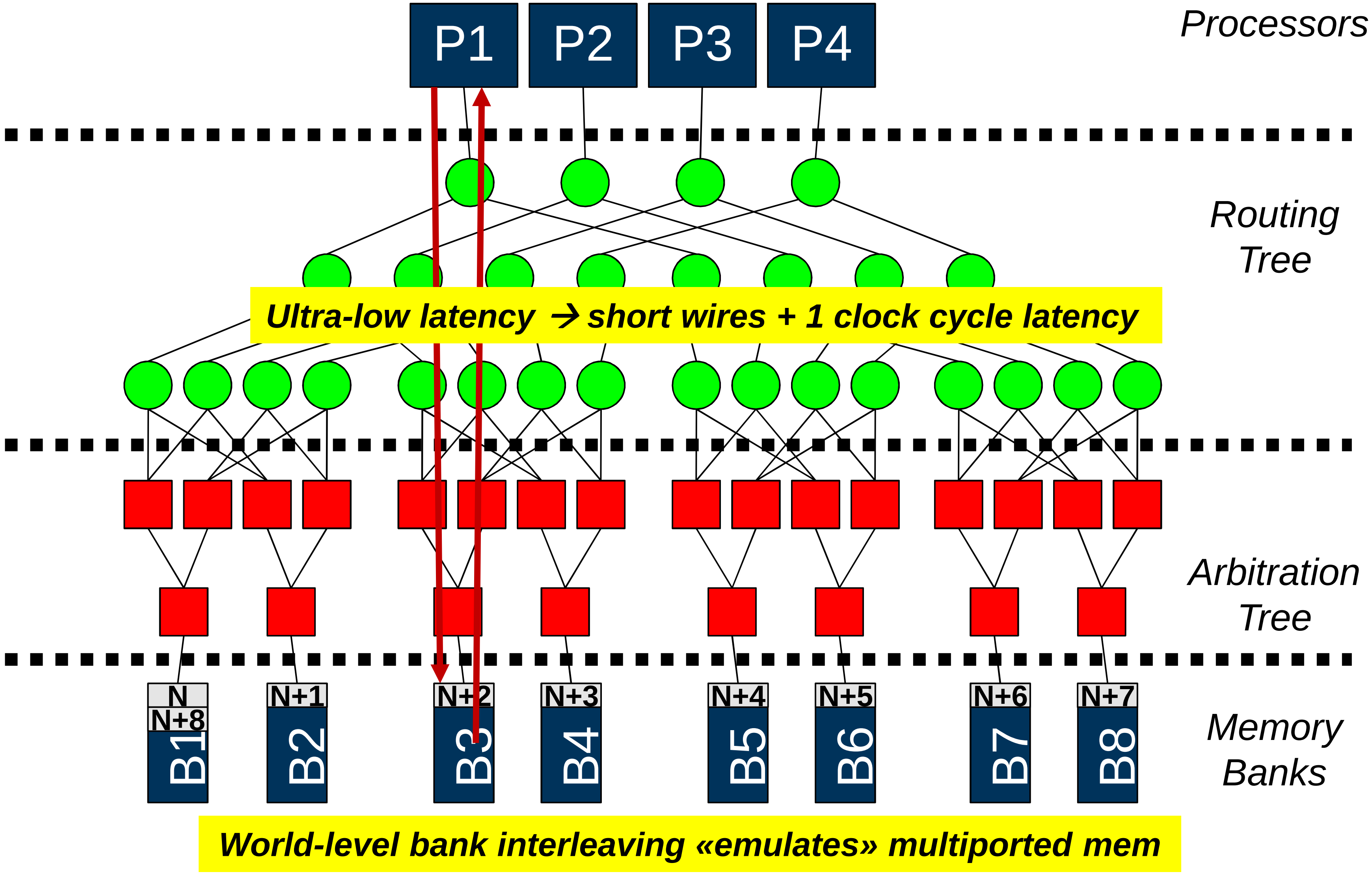


50% less energy per operation on average, 5% more area

ETH Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zurich

Logarithmic Interconnect

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA



P1
P2
P3
P4
Processors
Routing Tree
Ultra-low latency → short wires + 1 clock cycle latency
Arbitration Tree
N
N+8
N+1
N+2
N+3
N+4
N+5
N+6
N+7
B1
B2
B3
B4
B5
B6
B7
B8
Memory Banks
World-level bank interleaving «emulates» multiported mem

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

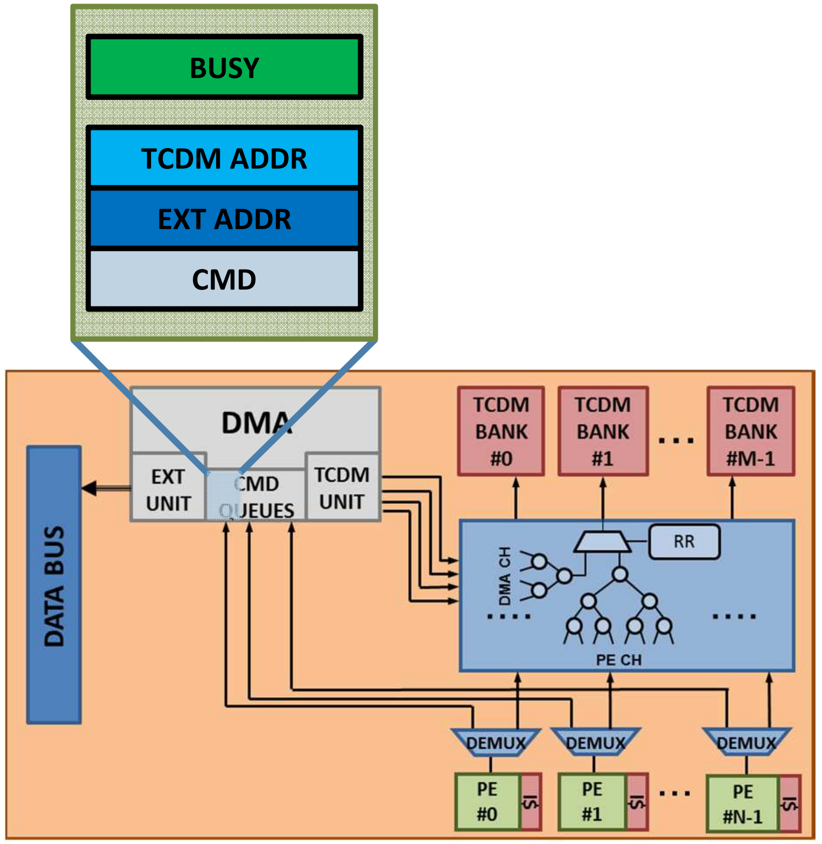
Low latency programming Interface

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA

- Each command queue is dedicated to a core of the cluster: arbitration is made in hardware
  → No need to reserve (lock/unlock) the programming channel

- COREs program DMA through register mapped on the DEMUX
  → *The registers belongs to aliases, no need for the processors to calculate (per-core) offsets*

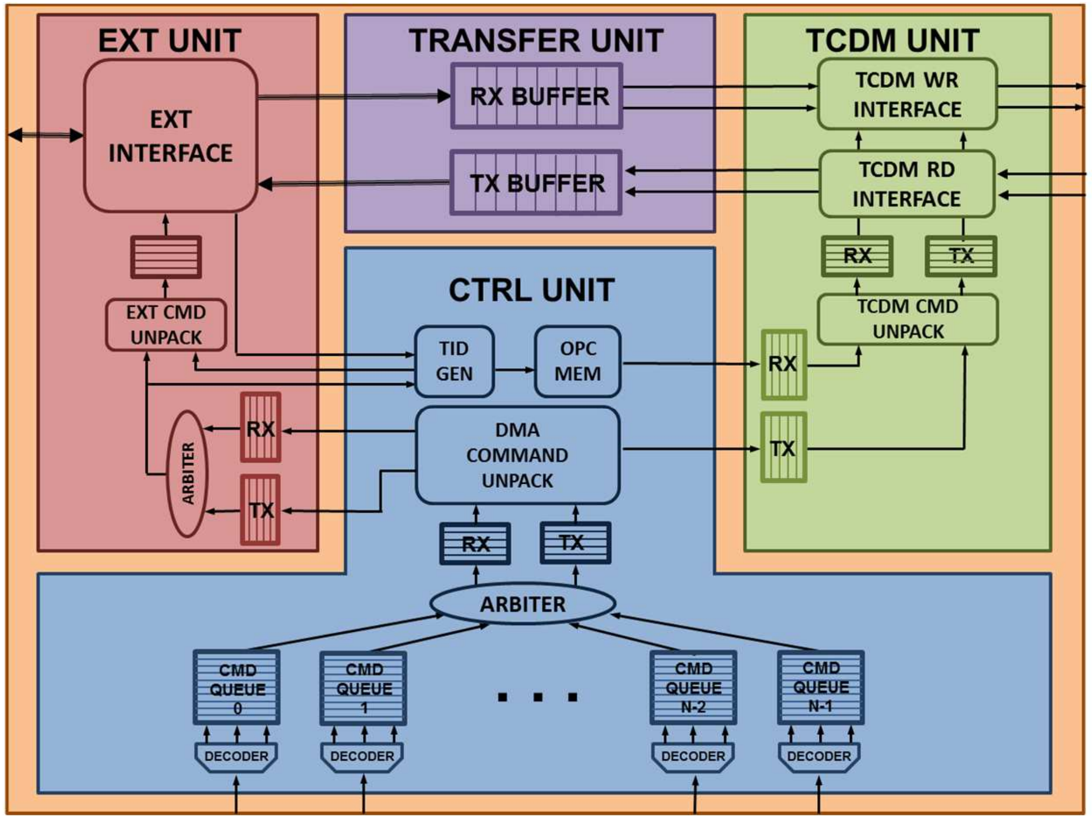
- A programming sequence requires
  1. *check a free command queue*
  2. *write address of buffer in TCDM*
  3. *write address of buffer in L2 memory*
  4. *Trigger data transfer*
  5. *Synchronization*



Programming Latency: ~10 CLOCK CYCLES!!!

DMA Architecture Overview

- **CTRL UNIT:**
  - Arbitration – forwarding and synchonization of incoming requests
- **TRANSFER UNIT:**
  - FIFO Buffers tor TX and RX channnels
- **TCDM UNIT:**
  - Bridge to TCDM protocol – 4 channels (2 RD and 2 WR) 32 bit each
- **EXT UNIT:**
  - Bridge to AXI, 64 bit

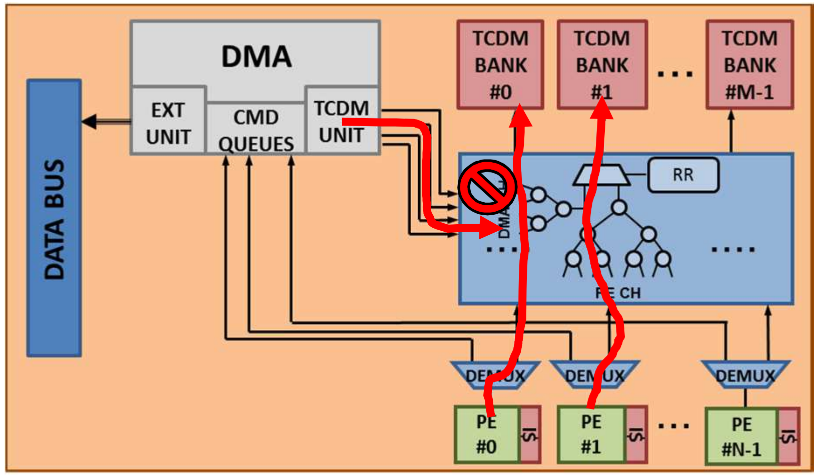


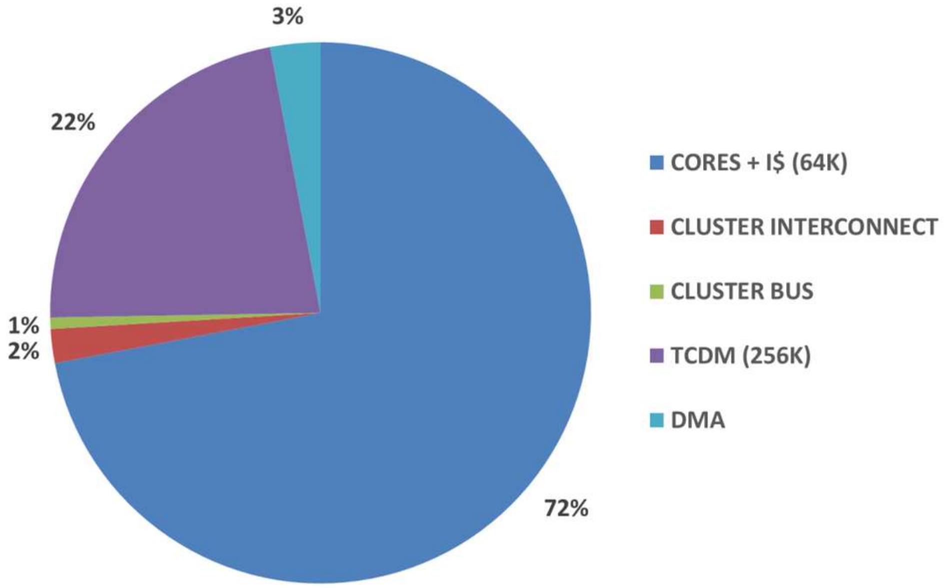
Key idea: only channel packets buffered internally – no DMA transfers!

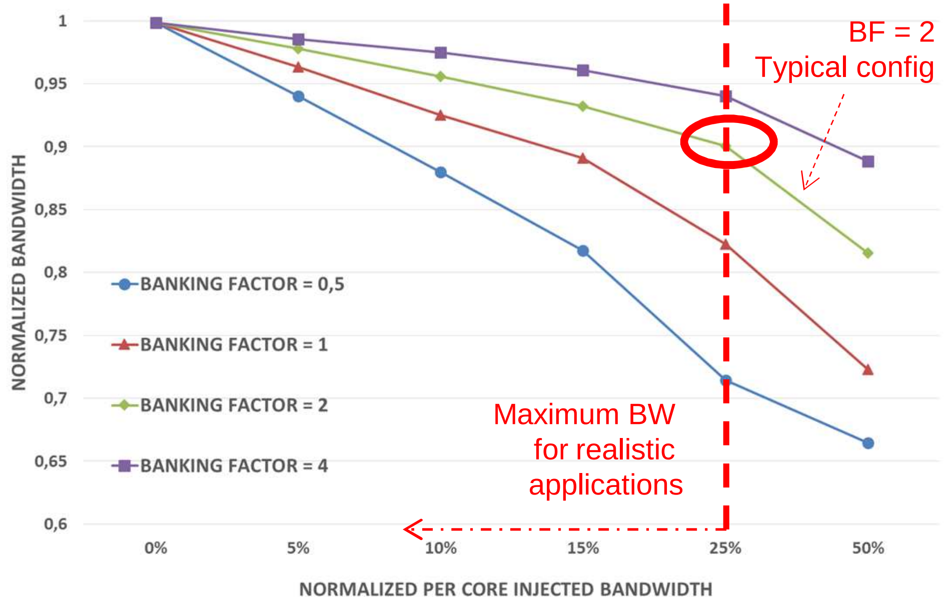
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

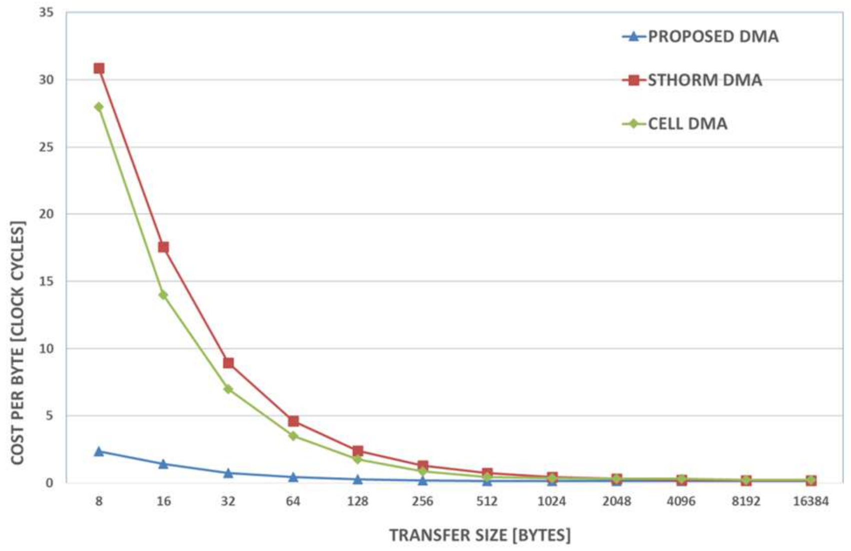
Cluster DMA

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA





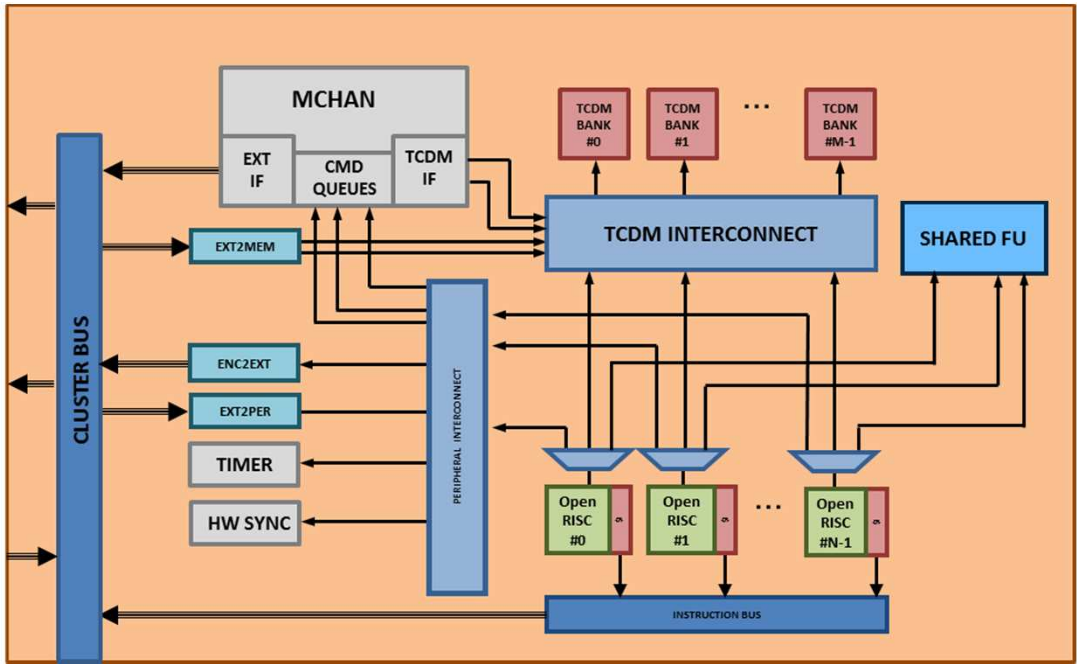


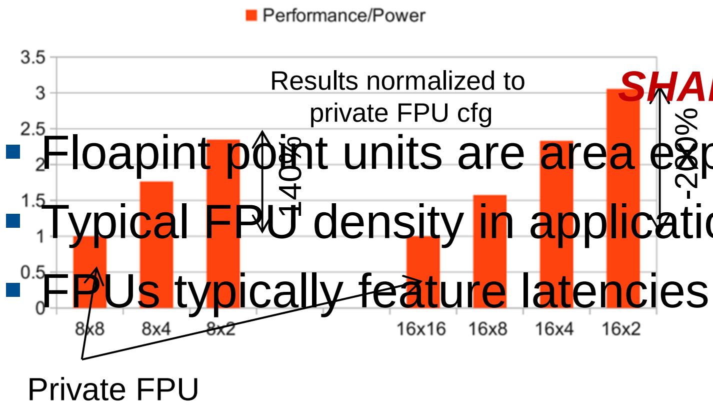


Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

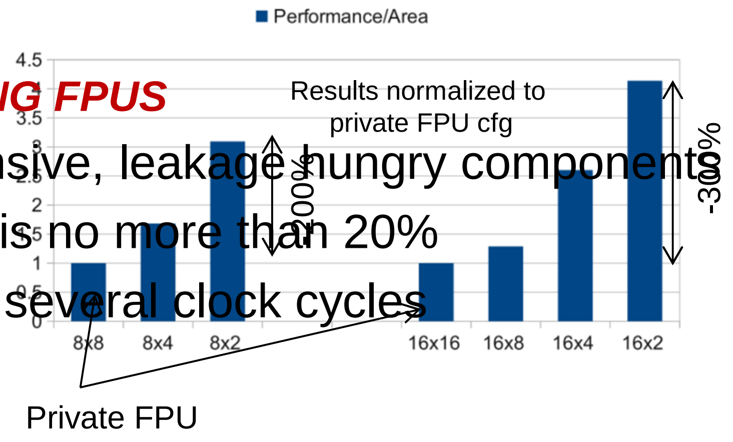
Sharing fucntional units (instruction set extensions)

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA





SHARING FPUS



- Floapint point units are area expensive, leakage hungry components
- Typical FPU density in application is no more than 20%
- FPUs typically feature latencies of several clock cycles

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

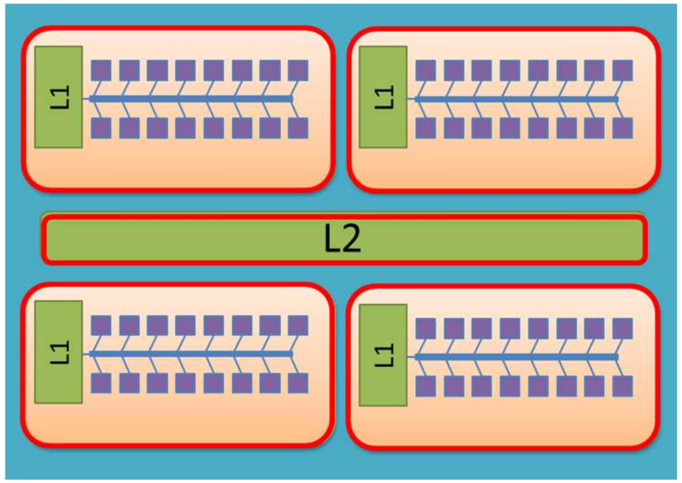
Scaling up

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

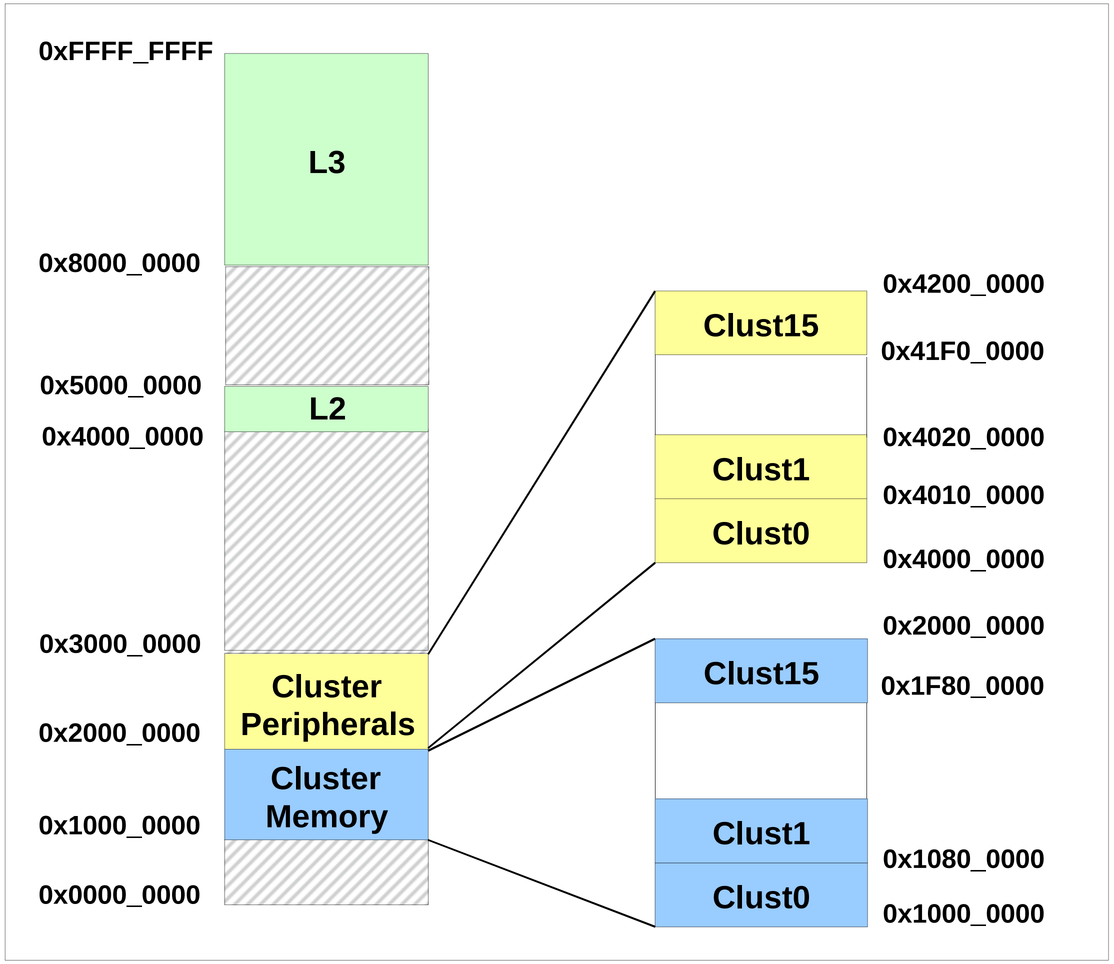
- 32 bit architecture

$\Rightarrow$ 4 GB of memory

- Clusters in Vdd, CLK domains
- L2 (2D & 3D) ready
- Host VM IF for Heterogeneous Computing



Memory Map:



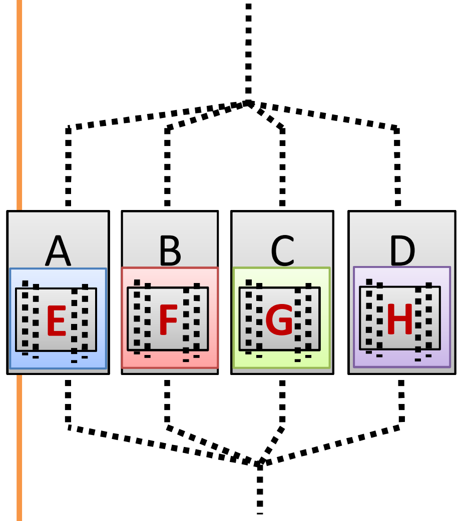
Programmability: OpenMP

```
while(1)
{
    #pragma omp parallel num_threads(4)
    {
        #pragma sections
        {
            #pragma section
            {
A.              #pragma omp parallel num_threads(16)
                ColorScaleConv();
            }
B
            }
            #pragma section
            {
C.              #pragma omp parallel num_threads(16)
                cvMoments();
            }
            #pragma section
            {
D.              #pragma omp parallel num_threads(16)
                cvAdd();
            }
        }
    }
}
```

```
void ColorScaleConv()
{
    #pragma omp for
    for(i = 0; i < FRAME_SIZE; i++)
    {
        [ALGORITH]
    }
}
```

A powerful abstraction for specifying structured parallelism

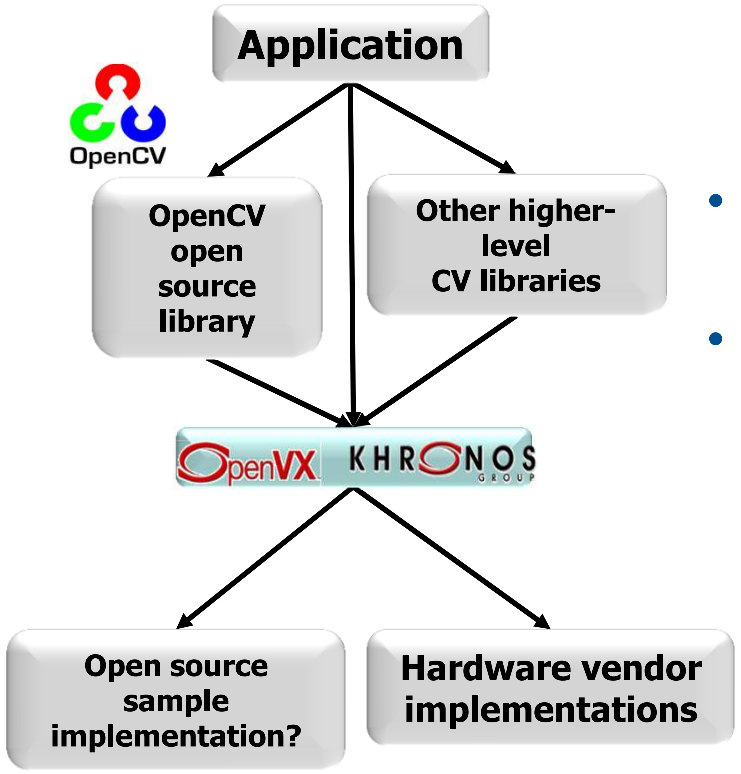
And very suitable for NUMA (cluster-based) systems



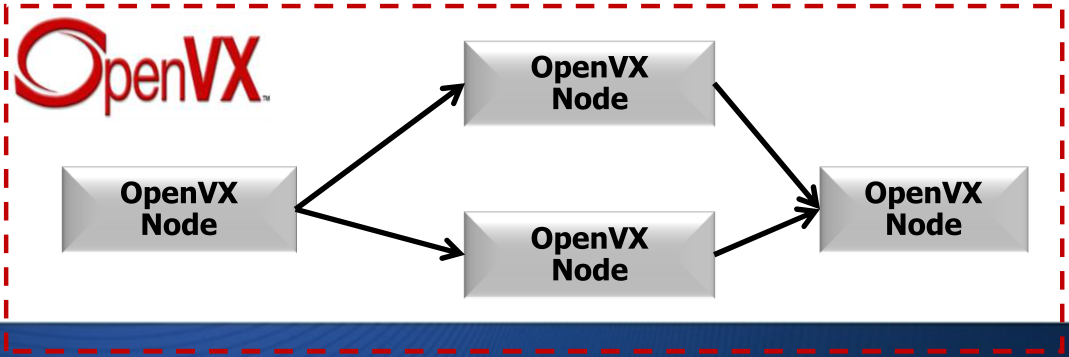
LLVM + OpenCL

+OpenVX →domain specific language

Programming: OpenVX

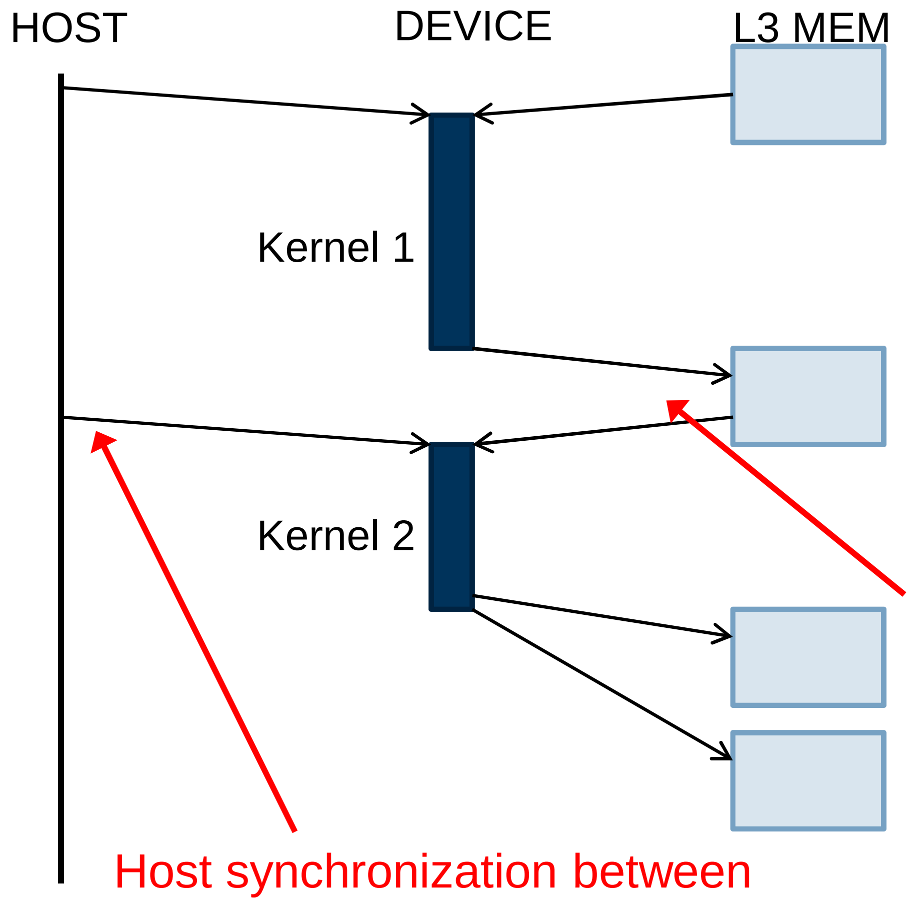

Application
OpenCV
OpenCV open source library
Other higher-level CV libraries
OpenVX KHRONOS GROUP
Open source sample implementation?
Hardware vendor implementations

- C-based standard API for vision kernel
  - Defines a set of **standard kernels**
  - Enables hardware vendors to ***implement accelerated imaging and vision algorithms***
- Focus on enabling real-time vision
  - On mobile and embedded systems
- **Graph execution model**
  - Each node can be implemented in software or accelerated hardware
  - **Data transfer between nodes may be optimized**



Vanilla OVX Runtime

Leverages OpenCL runtime with OVX nodes treated as OCL kernels

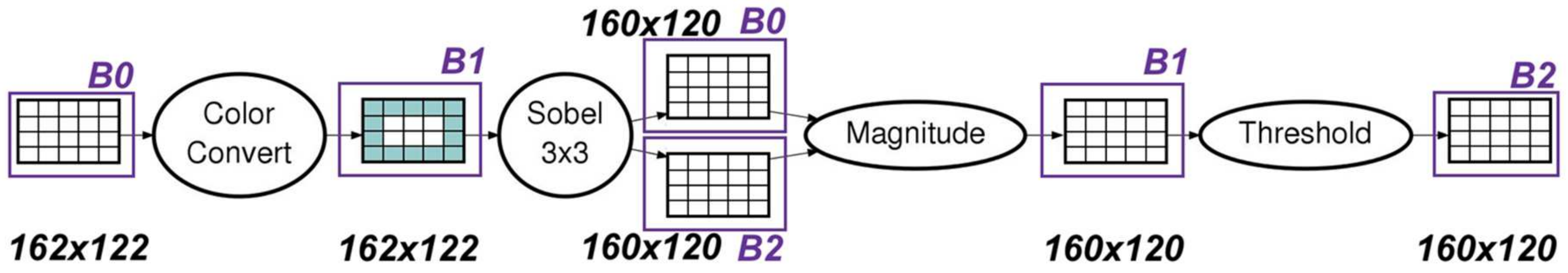


Intermediate results stored in L3 mem → each kernel boundary implies two accesses to all image data (write + read)

Host synchronization between successive kernels is imposed

Localized Execution

- Smaller image sub-regions (**tiles**) totally reside in TCDM
- All kernels are computed at tile level, no more at image level → *intermediate results are also allocated in TCDM*



✓ TCDM is partitioned into 3 buffers (B0, B1, B2)

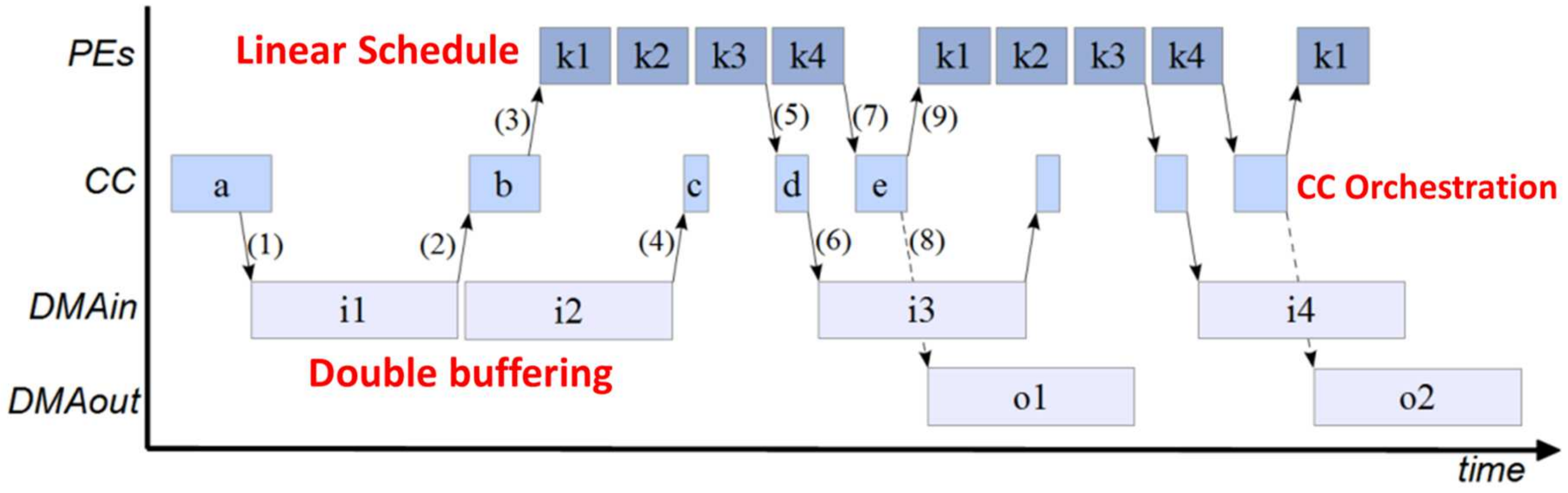
✓ Output tile size is 160x120

✓ Sobel3x3 requires image overlapping (1 pixel for each direction), and so the tile size is 162x122

✓ ColorConvert does not require tile overlapping, but must provide a 162x122 result tile for the next stage → **tile size propagation**

Localized Execution

- Smaller image sub-regions (**tiles**) totally reside in TCDM
- All kernels are computed at tile level, no more at image level → *intermediate results are also allocated in TCDM*



✓ TCDM is partitioned into 3 buffers (B0, B1, B2)

✓ Output tile size is 160x120

✓ Sobel3x3 requires image overlapping (1 pixel for each direction), and so the tile size is 162x122

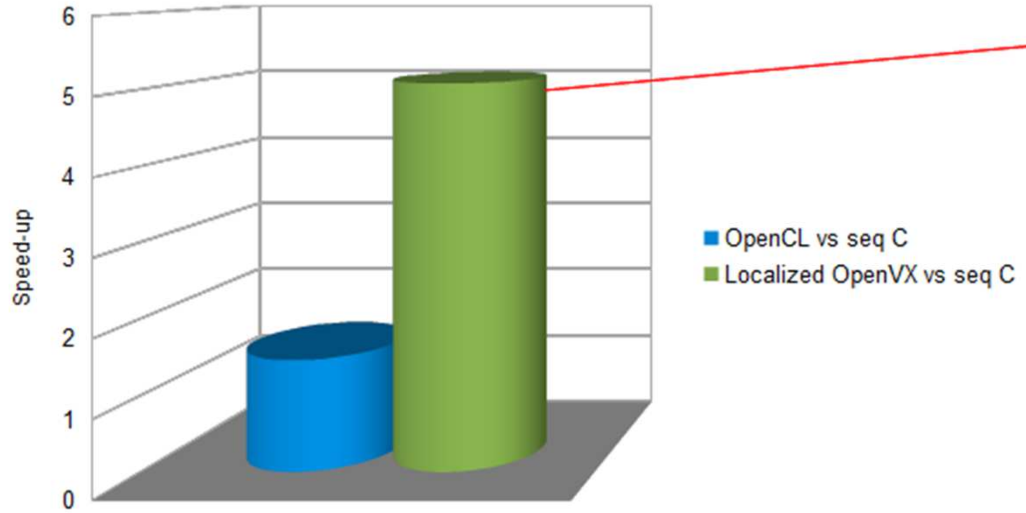
✓ ColorConvert does not require tile overlapping, but must provide a 162x122 result tile for the next stage → **tile size propagation**

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

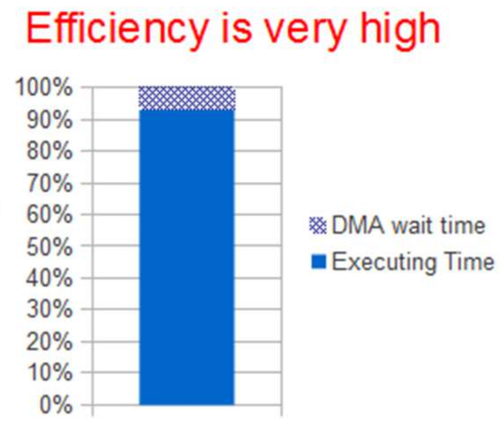
Localized Execution Results

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

- Framework prototype
  - OpenVX support
  - A limited subset of kernel has been implemented
  - Polynomial time (i.e. fast) heuristics suitable for ***just-in-time execution***



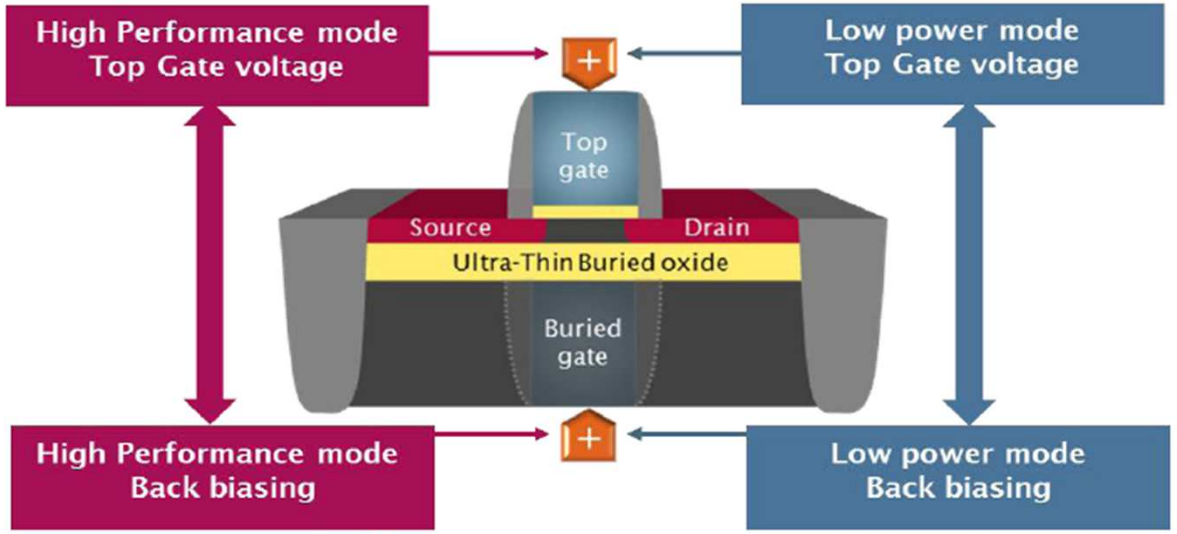
Speed-up (single cluster)

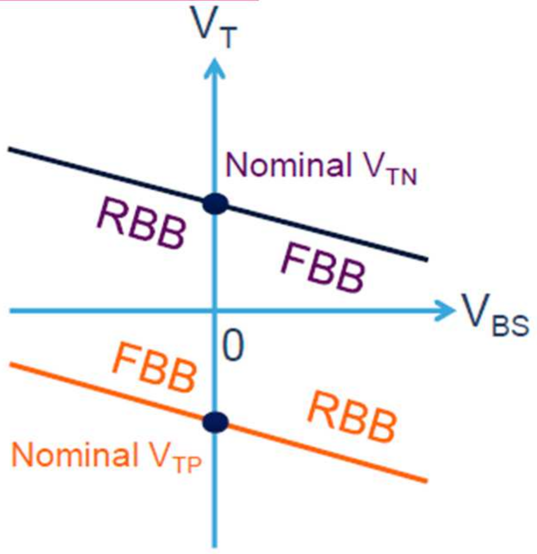


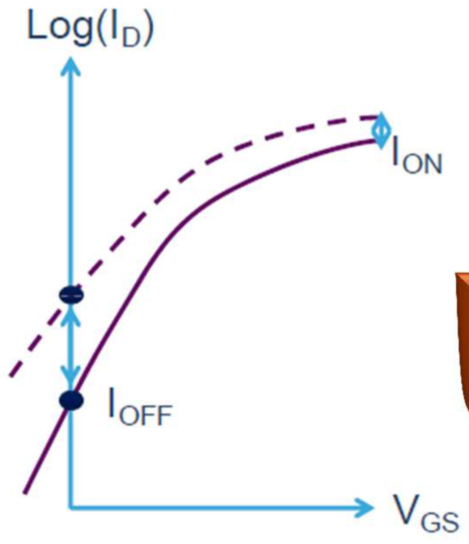
Cluster PE Efficiency

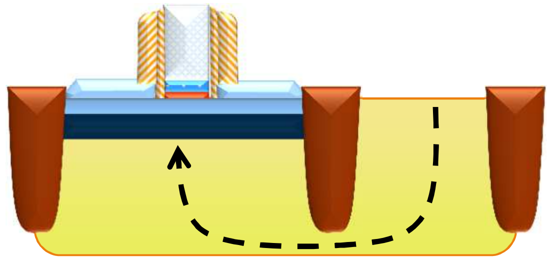
And what about Technology?

Near threshold FDSOI technology


High Performance mode Top Gate voltage
Low power mode Top Gate voltage
Top gate
Source
Drain
Ultra-Thin Buried oxide
Buried gate
High Performance mode Back biasing
Low power mode Back biasing

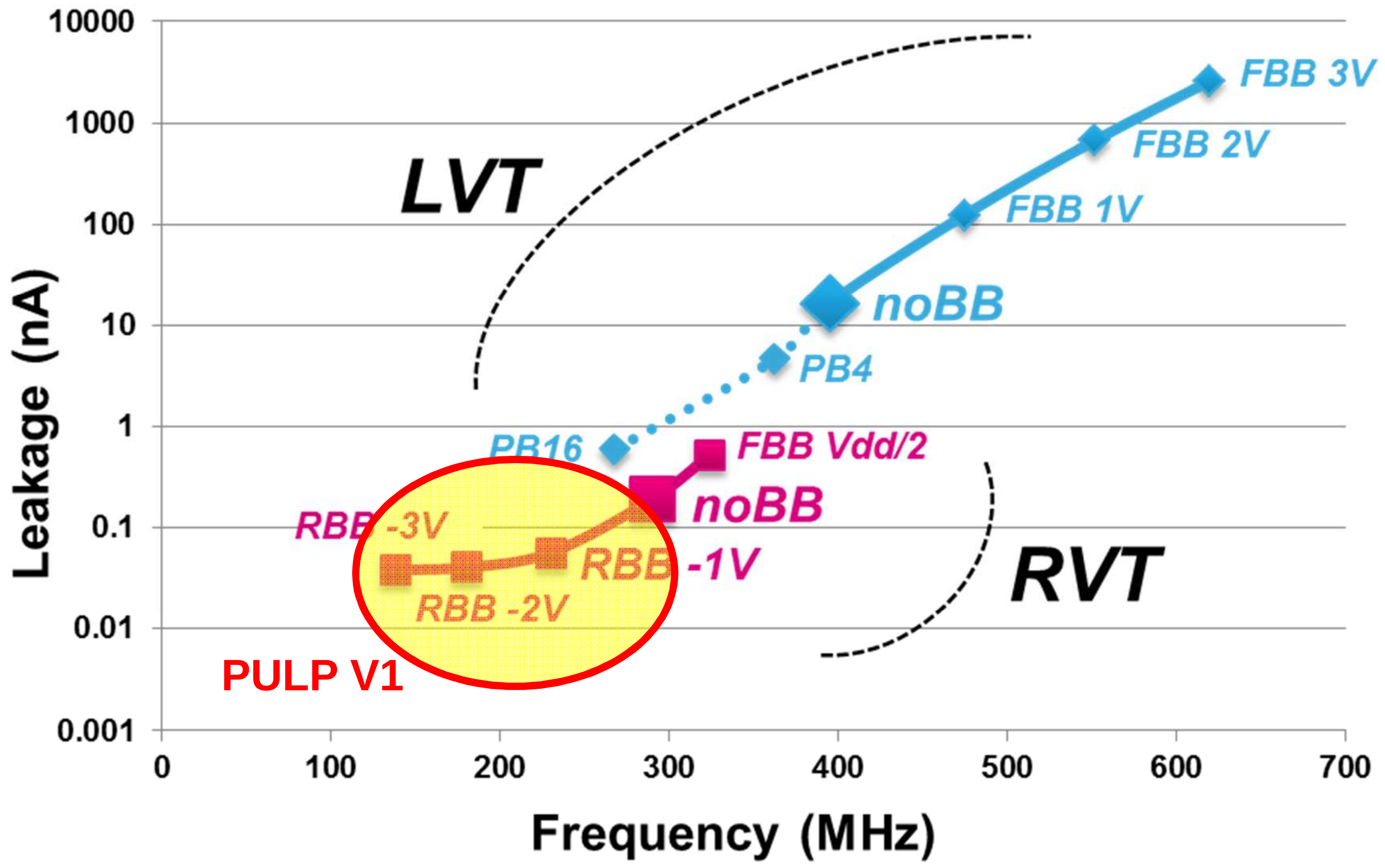




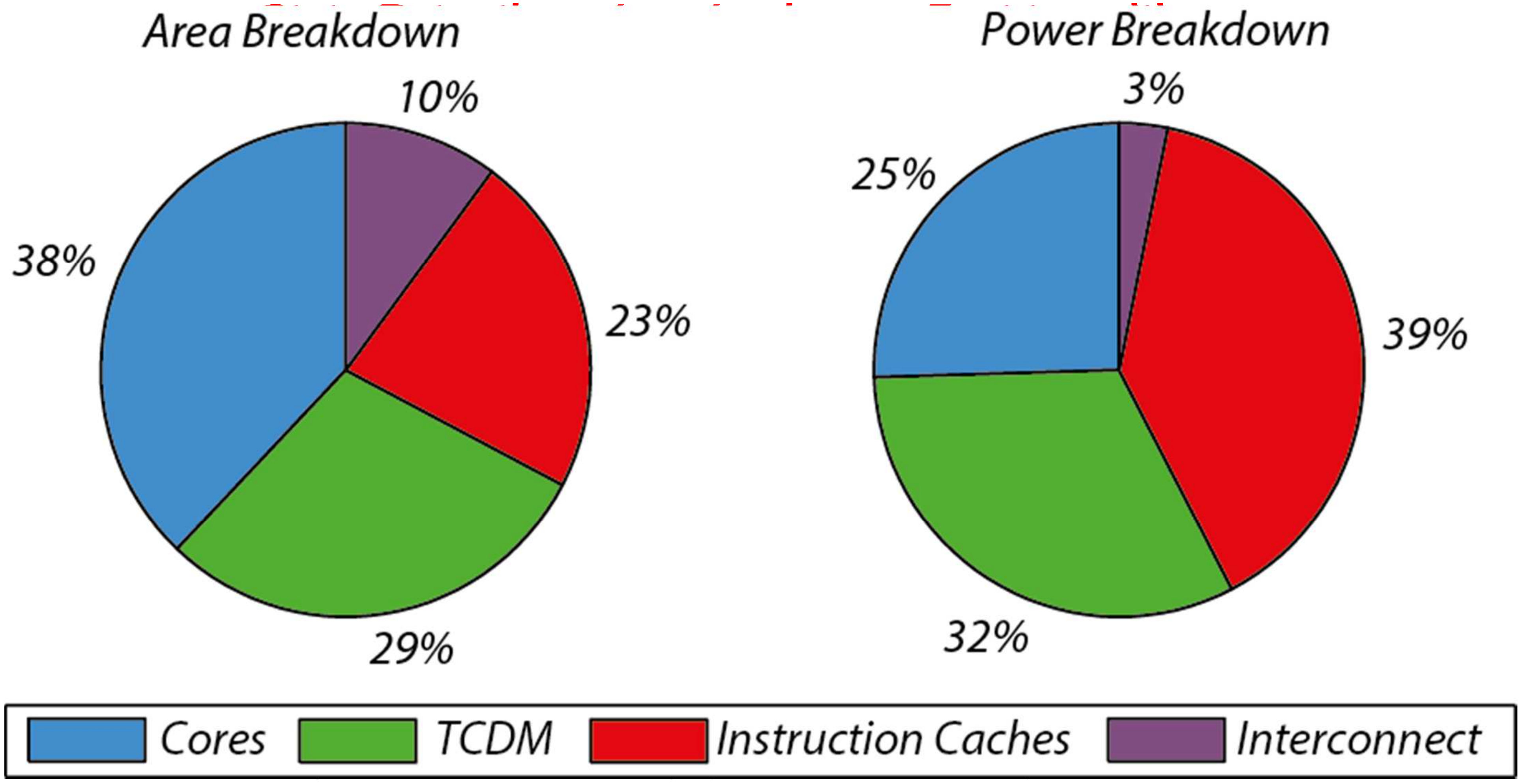


Body bias: Highly effective knob for leakage control!

Near threshold FDSOI technology



PULP V1: Doing nothing well (with RBB)



More than 50% of power into memories… this is the next focus area!

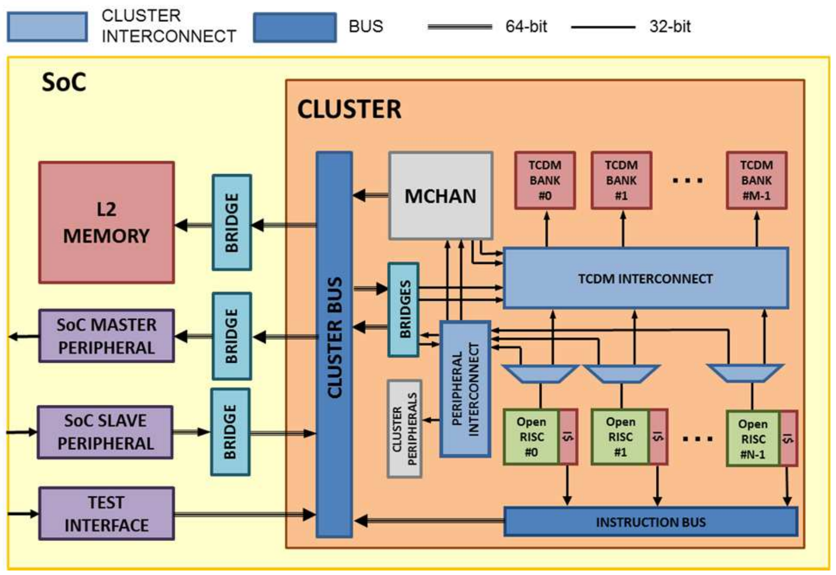
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

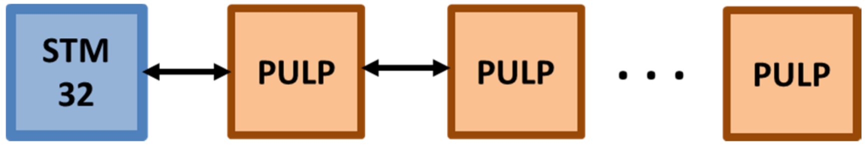
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Introducing PULP V2

Board/Application-ready chip

- Implementation of a master and a slave peripheral
  - Standard peripheral (e.g. SPI)
  - Integration with FPGAs or standard low power external memories
- PULP as multi-core accelerator for micro-controller host
  - STM32 host core
  - PULP muti-core accelerator
- Daisy chain of PULP chips
  - Pipeline of parallel processing units
  - Each core perfroms a stage of computation and forword temporary data to another stage

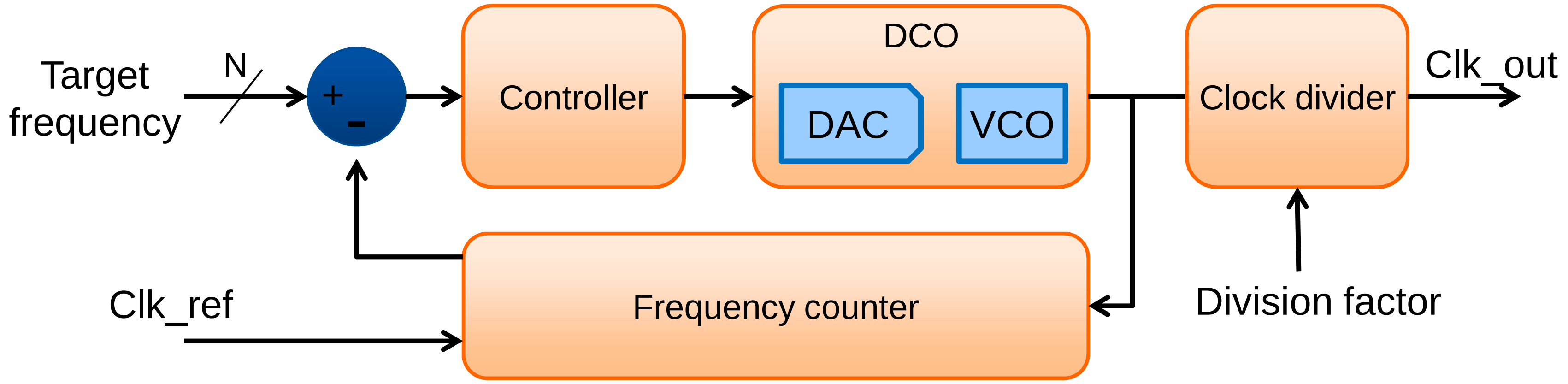




Standalone mode is also supported!

Clock generator: FLL

- All-digital clock generation based on a Frequency Lock Loop (FLL)
- From 2GHz down to 15KHz (through clock division)
- Frequency step 10MHz (at lowest division factor)
- Small area 3300µm² (50 times smaller than classic PLL)
- Suited to fine-grain GALS architectures
- Frequency reprograming in less than 180ns
- No frequency overshoot
- 15ps jitter

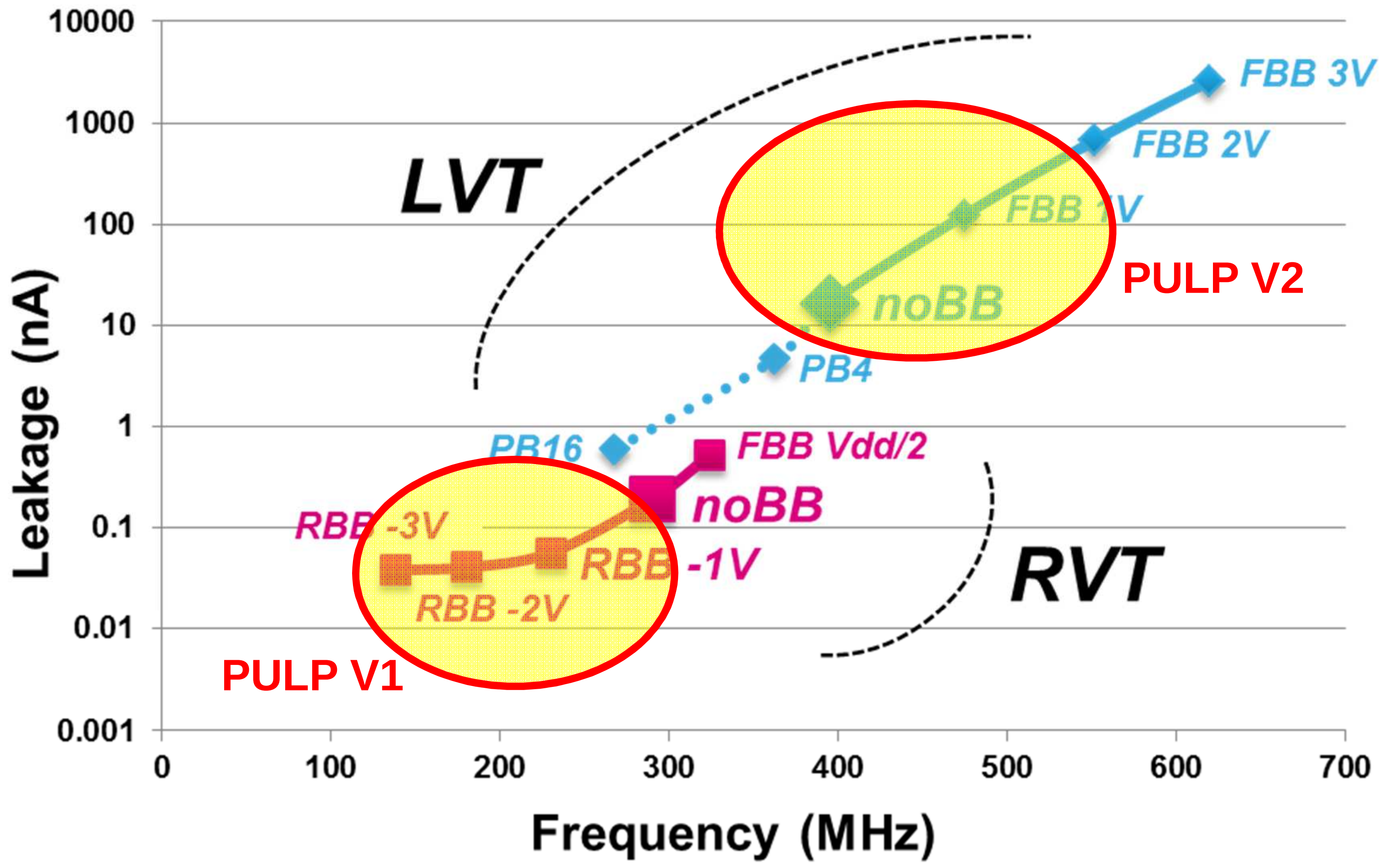


S. Lesecq, D. Puschini, E. Beigne, P. Vivet, and Y. Akgul, “Low-Cost and Robust Control of a DFLL for Multi-Processor System-on-Chip,” IFAC Proceedings Volumes, vol. 18, pp. 1940–1945, 2011

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Near threshold FDSOI technology

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

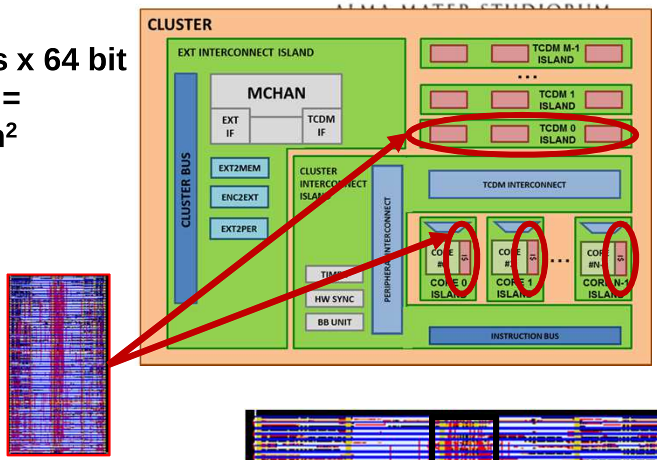


Low-leakage vs. Low voltage (0.3V) → **reactive** or proactive?

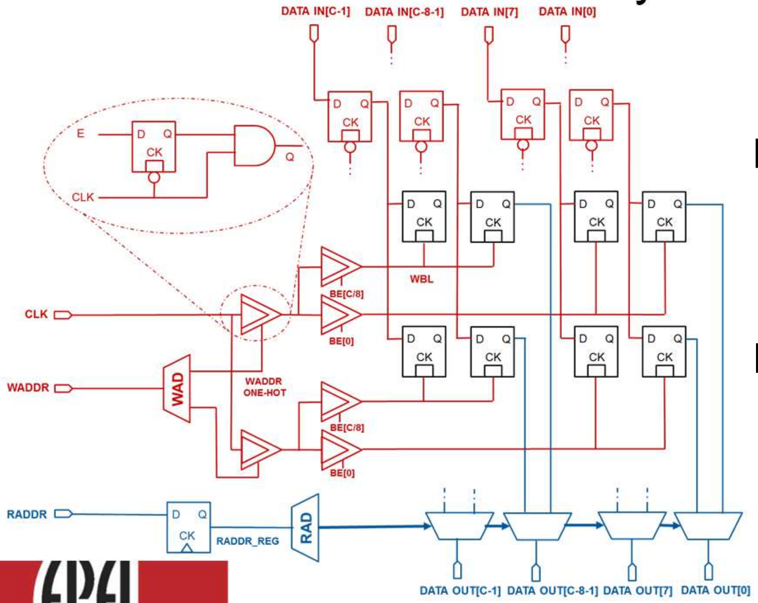
ULP Latch-based SCM

- Based exclusively on standard cells
- Voltage range identical to the core
  - Only static logic
- Layout based on guided P&R
  - Close to 100% density

64 words x 64 bit
162 x 85 =
13.7k um^2



SCMs

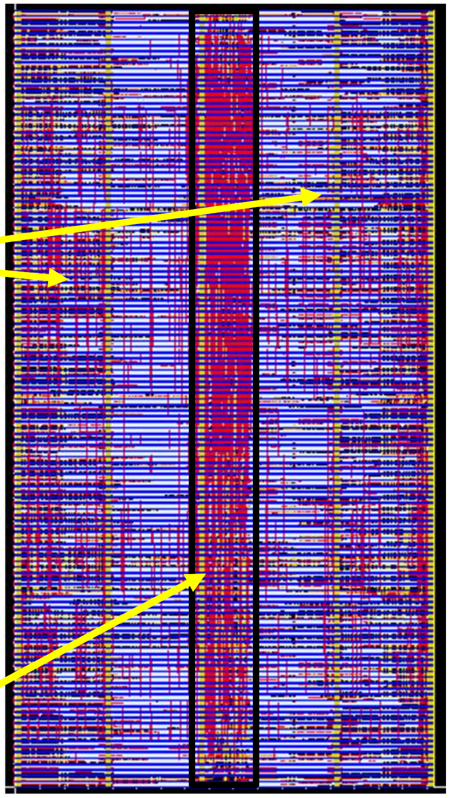


Macro Size:
128 x 32 (4 kb)
86µm x 160µm

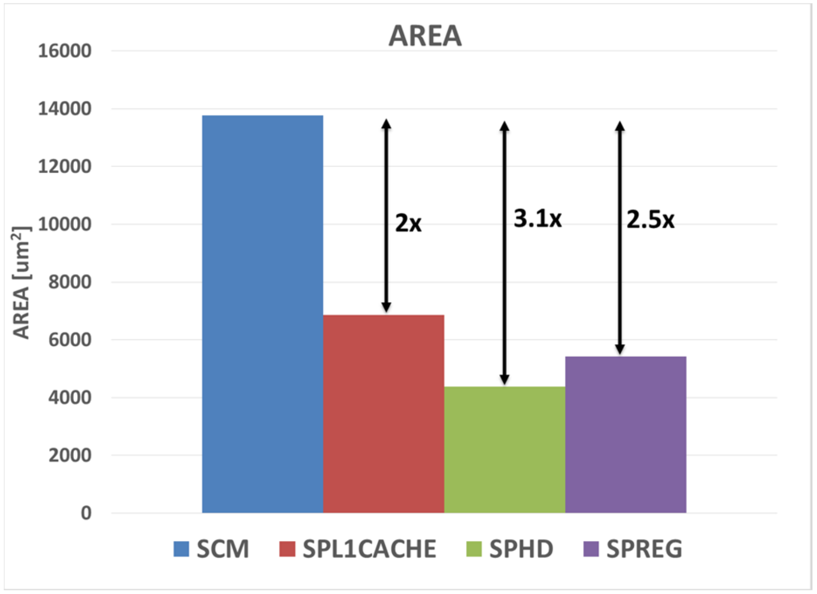
Input/Output Delay:
0.3ns/0.7ns @ss0.9V125°
2ns/3.3ns @tt0.3V25° (FBB)

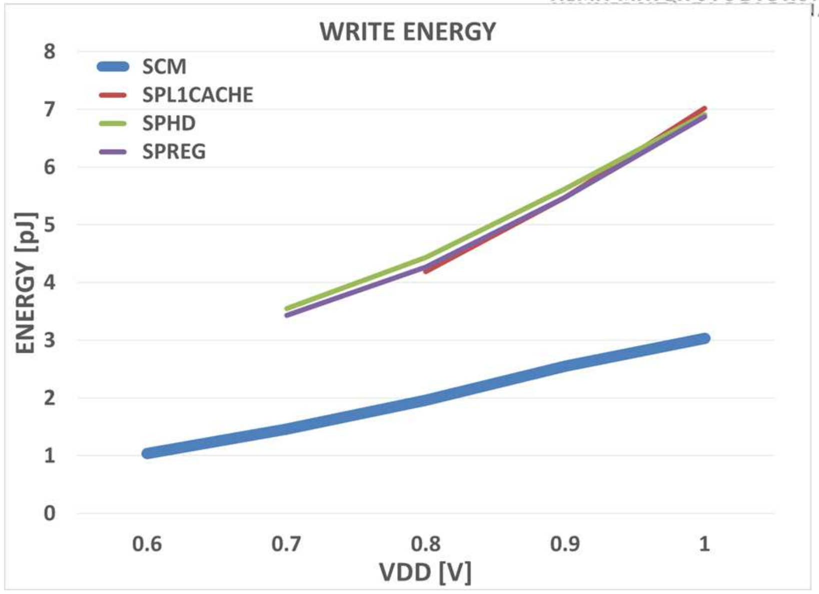
Storage Array + output mux

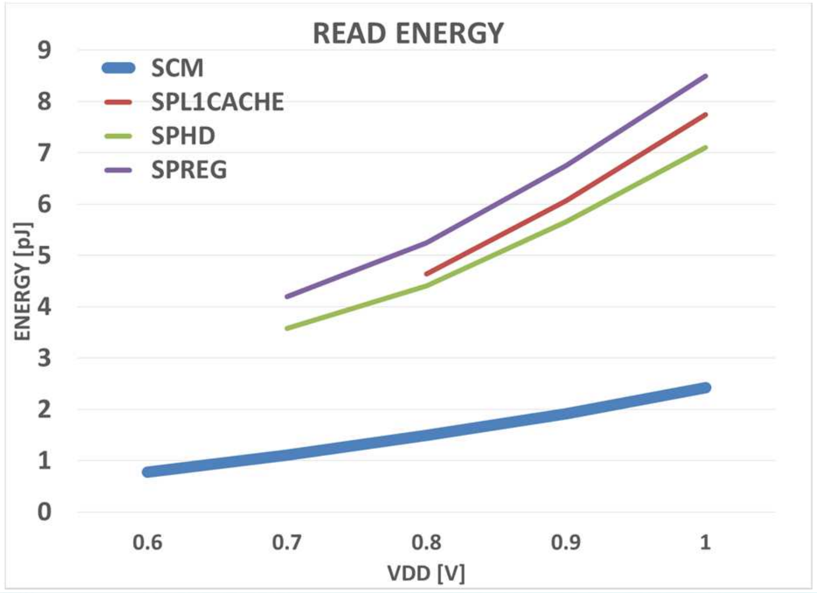
Decoders

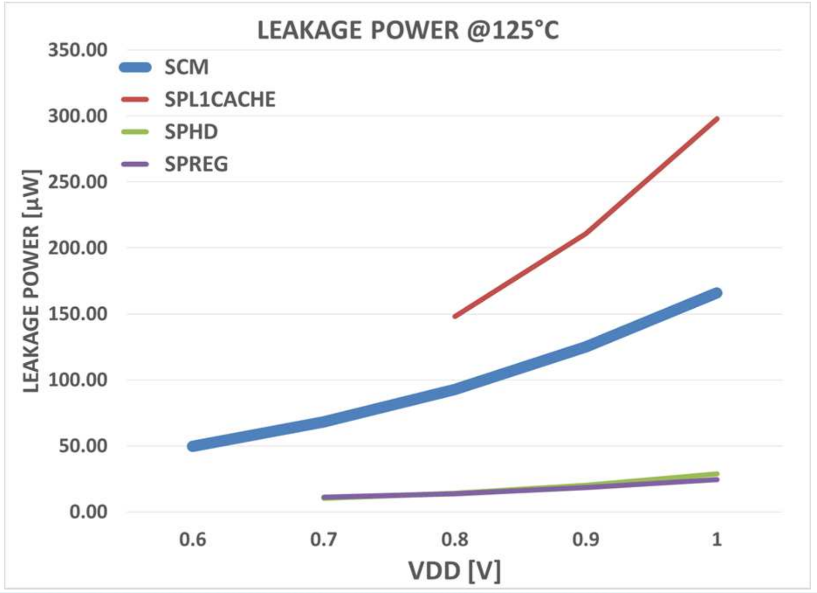


Comparison to 64x64 SRAM





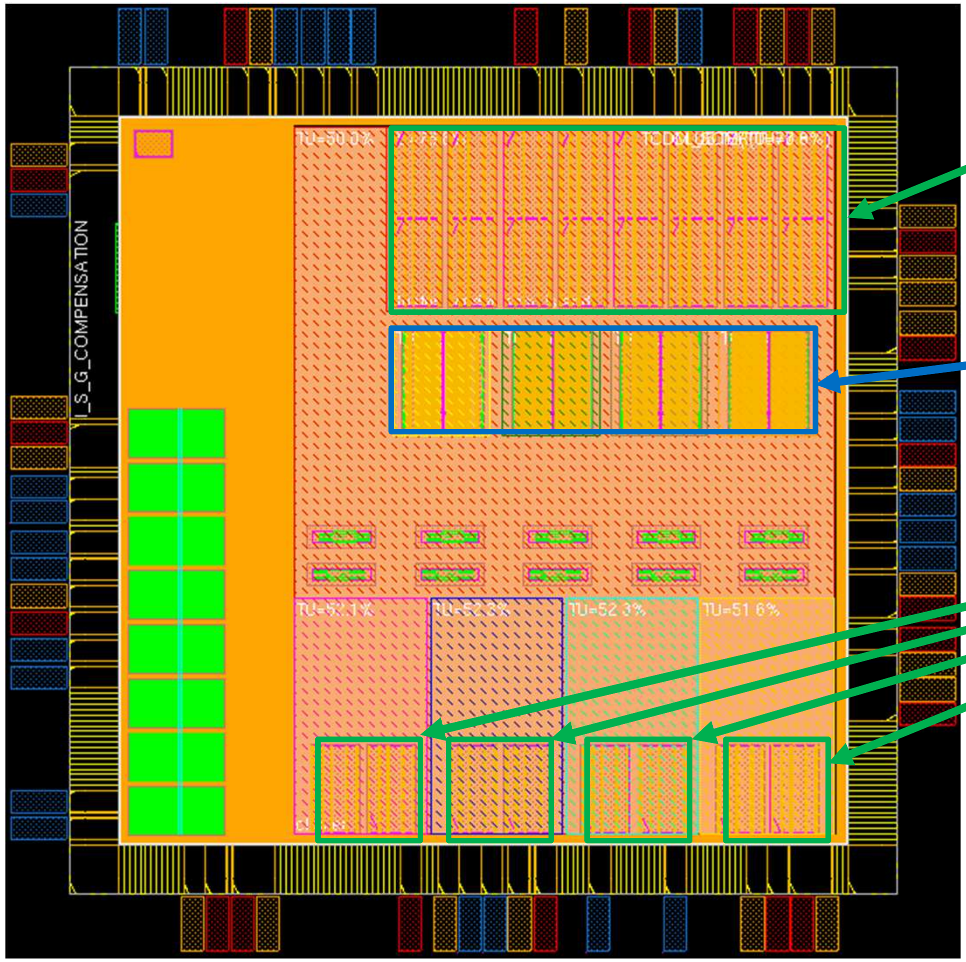




Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

SCM Integration into PULP2

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA



16 banks of
128x32 bit SCM
(8 kbyte)

8 banks of
512x32 bit SRAM
(16 kbyte)

2 banks of
128x32 bit SCM
(2 kbyte)
Instruction
cache per core

PULP V2 is taping out this Week, PULP V3 is on the drawing board…

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

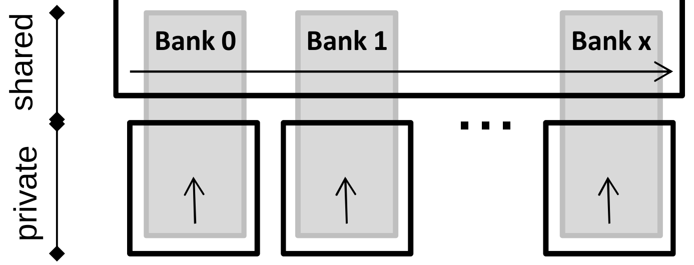
Reconfigurable address mapping

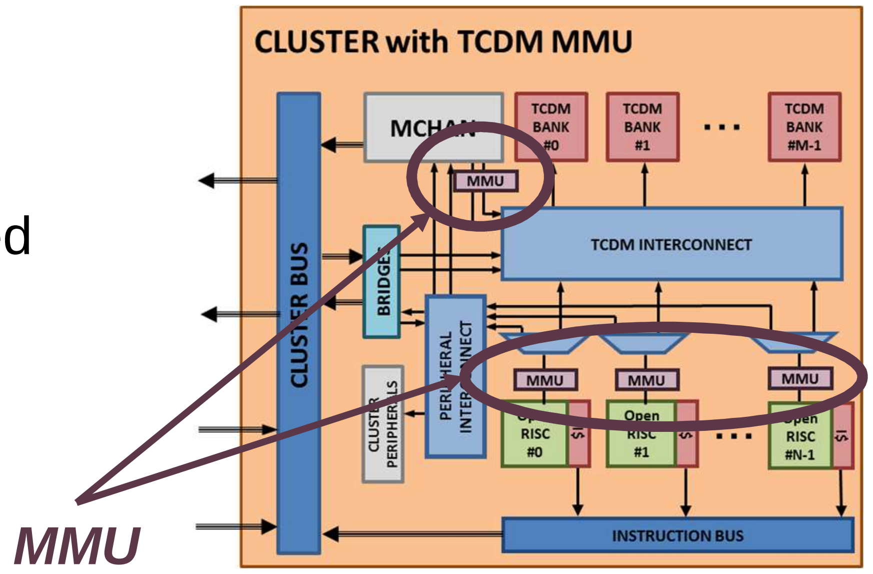
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

- Support for different address mappings:
  - Interleaved: horizontal shutdown and reduces conflicts in shared segments
  - Non-interleaved mapping: private memory avoids conflicts

- Basic MMU
  - Coexistence of both shared and private memory segments with different address mappings
  - Adapt address mapping is adapted to accommodate partial memory shut-down

Mapping between logical-physical addresses

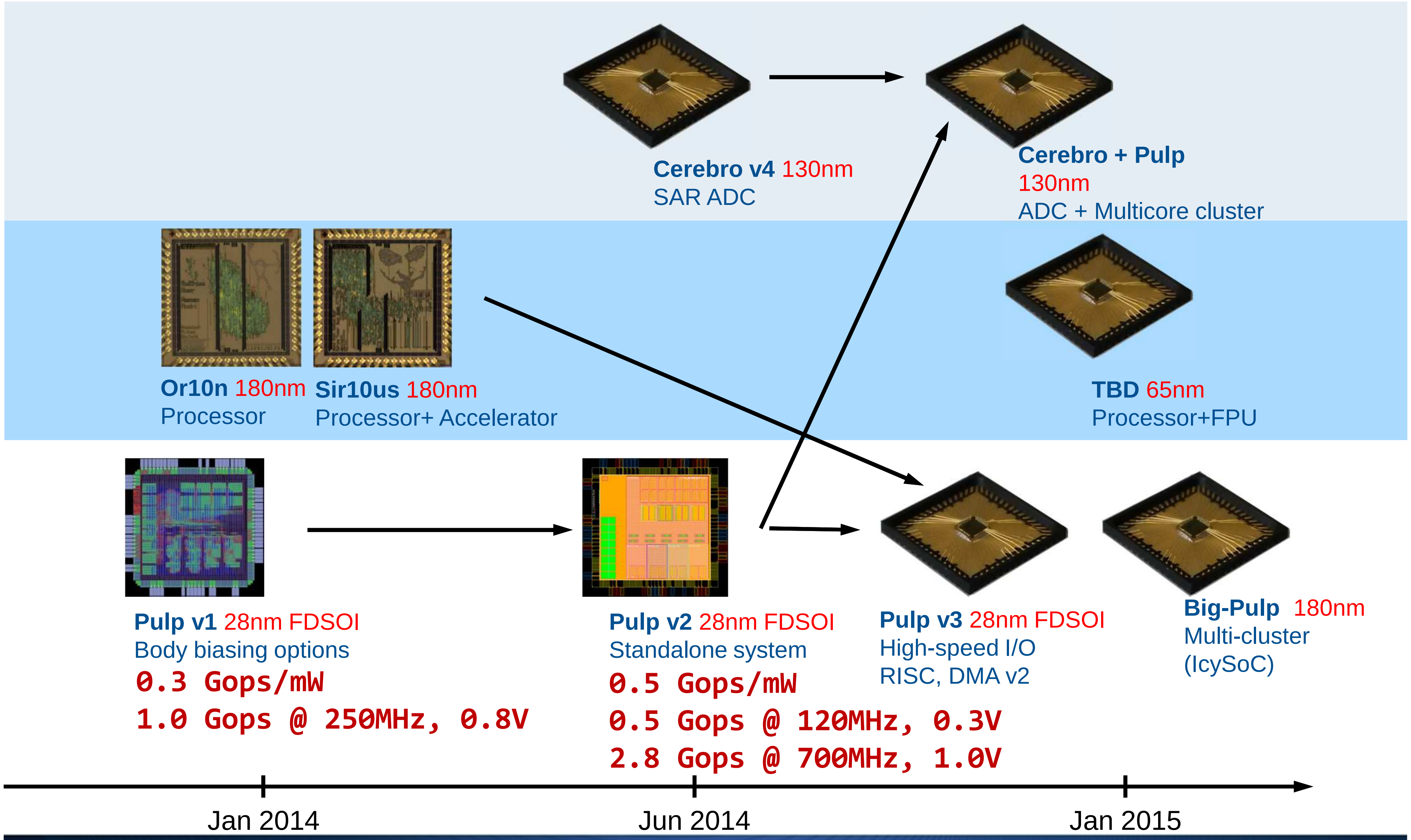




ETH Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

PULP Family Development

ALMA MATER STUDIORUM UNIVERSITÀ DI BOLOGNA



Cerebro v4 130nm
SAR ADC

Cerebro + Pulp
130nm
ADC + Multicore cluster

Or10n 180nm
Processor

Sir10us 180nm
Processor+ Accelerator

TBD 65nm
Processor+FPU

Pulp v1 28nm FDSOI
Body biasing options
0.3 Gops/mW
1.0 Gops @ 250MHz, 0.8V

Pulp v2 28nm FDSOI
Standalone system
0.5 Gops/mW
0.5 Gops @ 120MHz, 0.3V
2.8 Gops @ 700MHz, 1.0V

Pulp v3 28nm FDSOI
High-speed I/O
RISC, DMA v2

Big-Pulp 180nm
Multi-cluster
(IcySoC)

Jan 2014
Jun 2014
Jan 2015

An ULP Computing Ecosystem

HARDWARE IPs

- PROCESSOR
- INTERCONNECT (LOCA
- MEMORY HIERARCHY MEMO
- HARDW
- …

SOFTWARE

- PILER/TOOLCHAIN
- GRAMMING MODELS
- IME

Building an open-source ecosystem for exploring (<u>with silicon!</u>) next-generation parallel computing platforms

SILICON

- OPTIMIZATION FLO
- IMPLEMENTATION F
- VERIFICATION FLOW
- FULL CUSTOM IPS
- …

BUT …

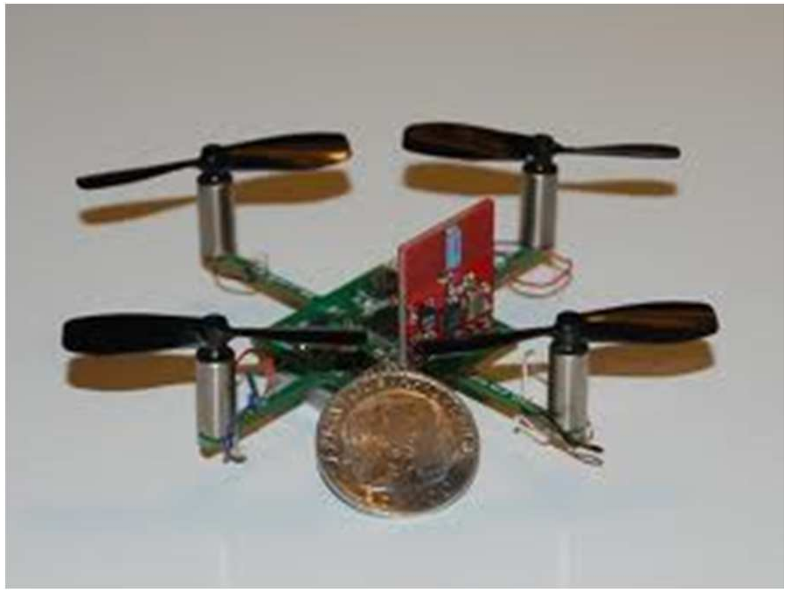
- SUPPORT FOR DEBUG
- SUPPORT FOR PROFILING
- DESIGN FOR TESTABILITY
- …..

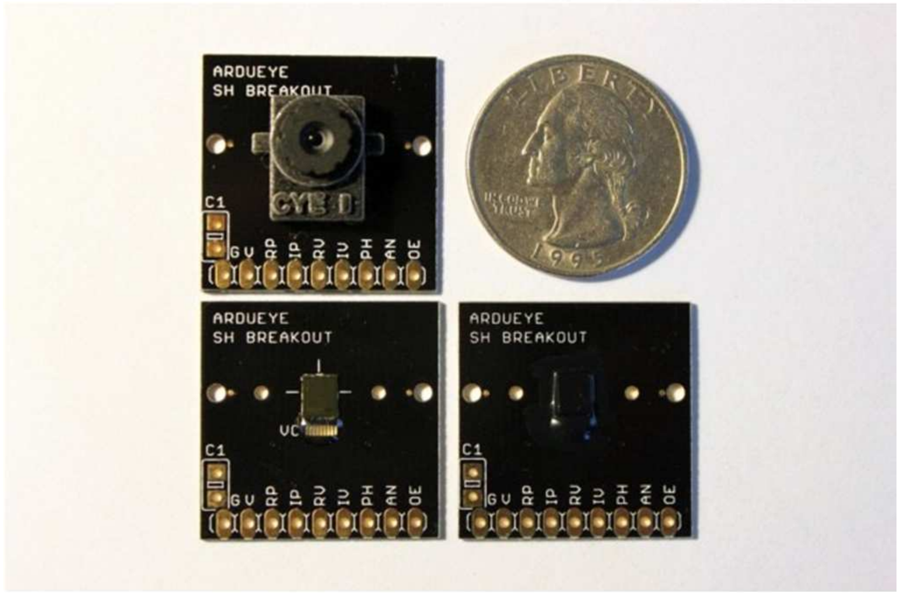
VALIDATION

- L PLATFORM
- EMULATION PLATFORM (FPGA)
- BENCHMARKS
- REGRESSION TESTS
- …

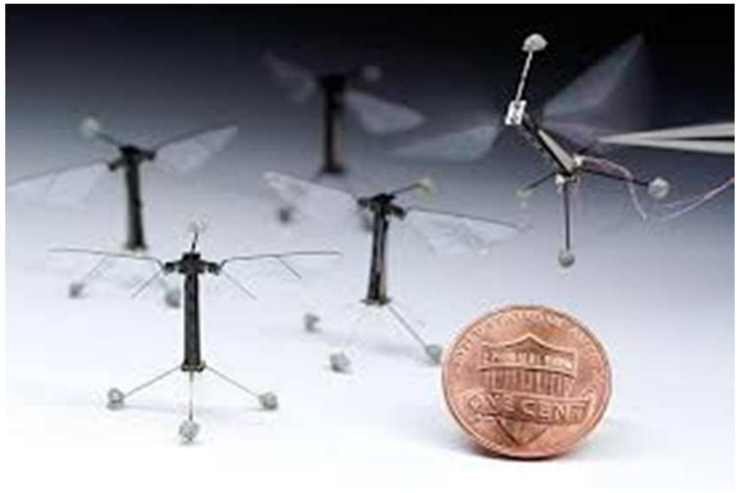
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

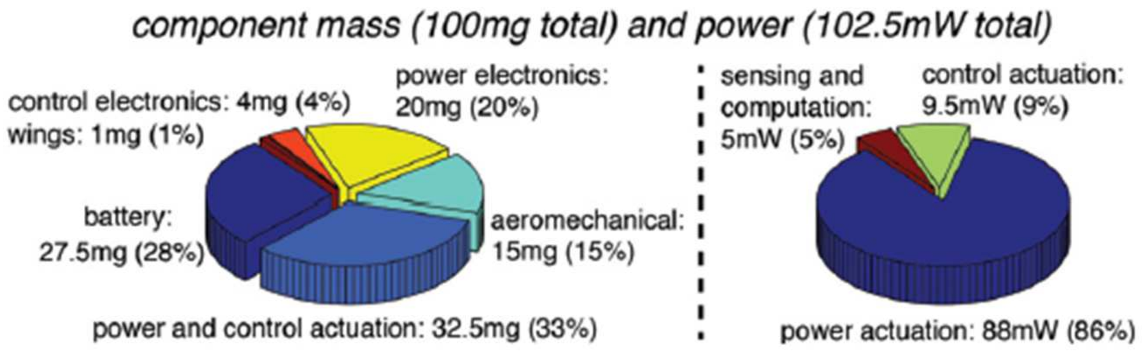
To do what?





112x112pixel (300uW)




component mass (100mg total) and power (102.5mW total)
control electronics: 4mg (4%)
wings: 1mg (1%)
power electronics: 20mg (20%)
battery: 27.5mg (28%)
aeromechanical: 15mg (15%)
power and control actuation: 32.5mg (33%)
sensing and computation: 5mW (5%)
control actuation: 9.5mW (9%)
power actuation: 88mW (86%)

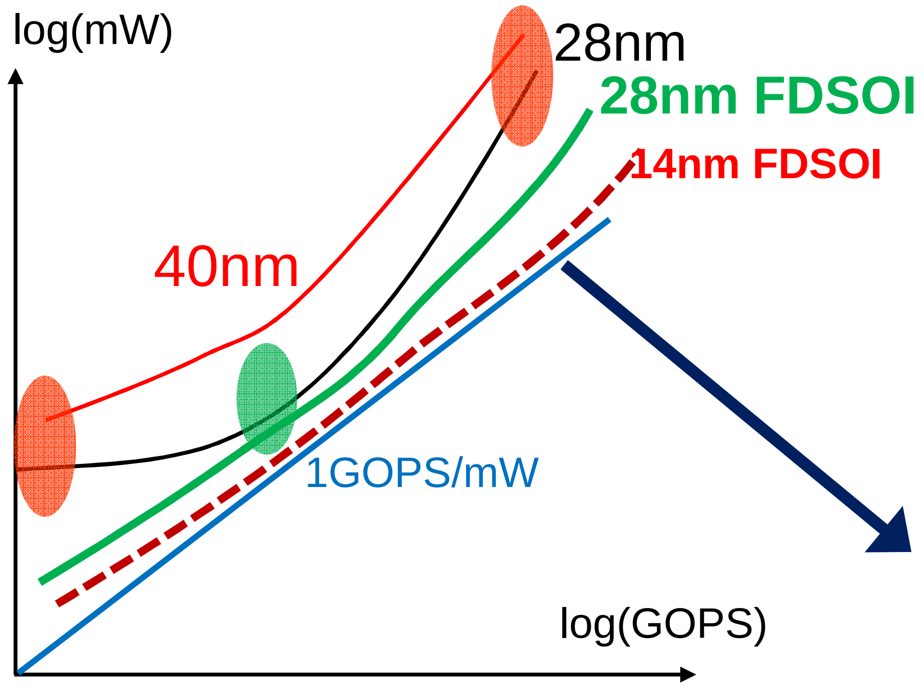
[Wood13]

@60fps 0.76MPx/s → with 1KOPS/pixel we need 0.75Gops!

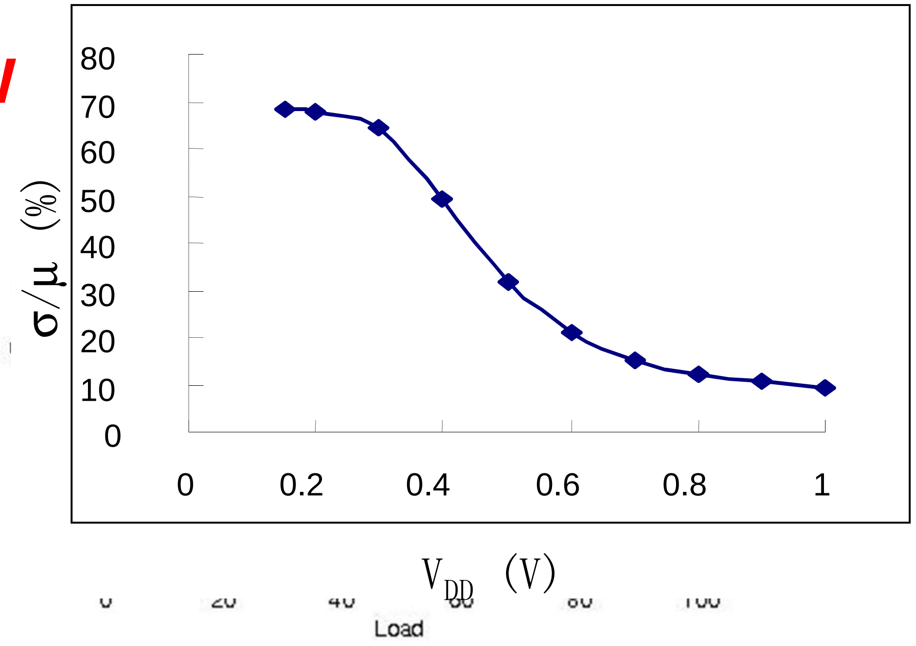
The Grand Challenge: Energy proportionality

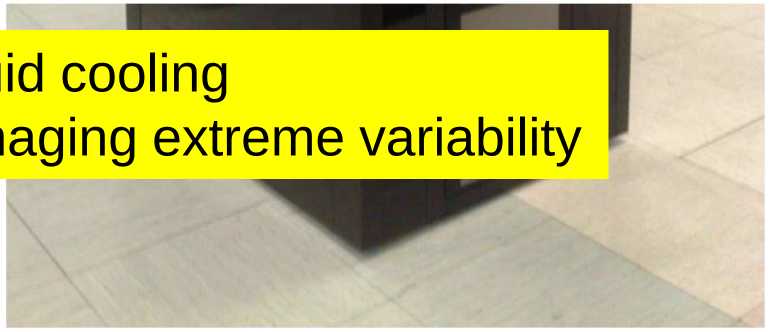
3,200MOPS/W – 30KW

PULP1-2 32b, Kops-Tops @ pJ/W



Goal: reduce «bending up» of energy curve at low & **high** perf!





+ Liquid cooling
+ Managing extreme variability

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Thank you!

European Research Council
Multithermand AdG
Multiscale Thermal Management of Computing Systems