Resource Allocation & Scheduling in Moore's Law Twilight Zone

Luca Benini

Università di Bologna & STMicroelectronics

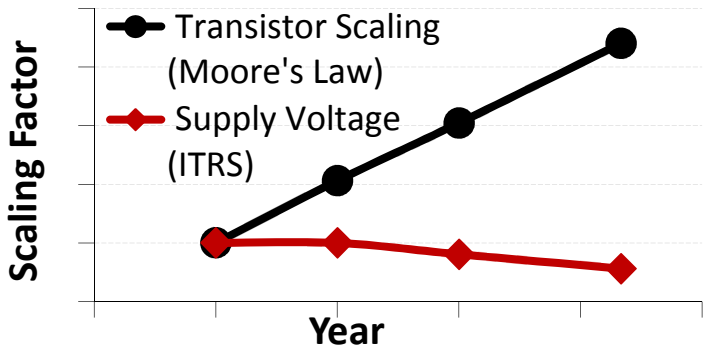
The Twilight of Moore's Law: Economics

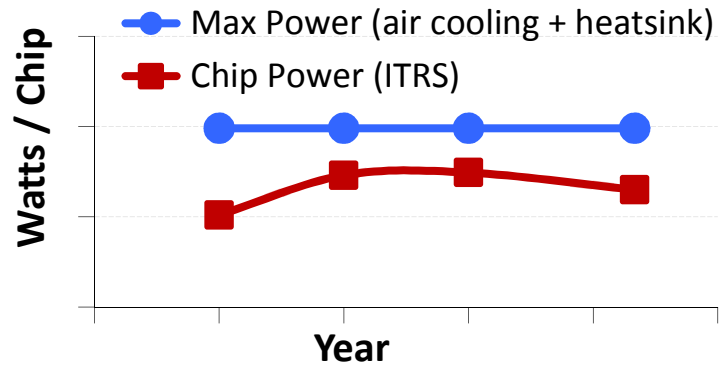
130nm	90nm	65nm/55nm	45/40nm	32/28nm	22/20nm
Altis Semiconductor	Dongbu Hitek	Freescale	Fujitsu	Globalfoundries	Globalfoundries
Dongbu Hitek	Freescale	Fujitsu	Panasonic	Intel	Intel
Freescale	Fujitsu	Globalfoundries	Globalfoundries	Panasonic	Samsung
Fujitsu	Globalfoundries	IBM	IBM	Samsung	ST Microelectronics
Globalfoundries	Grace Semiconductor	Infineon	Intel	ST Microelectronics	TSMC
Grace Semiconductor	IBM	Intel	Renesas (NEC)	Toshiba	UMC
IBM	Infineon	Panasonic	Samsung	TSMC	
Infineon	Intel	Renesas (NEC)	SMIC	UMC	
Intel	Panasonic	Samsung	ST Microelectronics		
Panasonic	Renesas (NEC)	SMIC	Toshiba		
Renesas (NEC)	Samsung	Sony	TSMC		
Samsung	Seiko Epson	ST Microelectronics	UMC		
Seiko Epson	SMIC	Texas Instruments			
SMIC	Sony	Toshiba			
Sony	ST Microelectronics	TSMC			
ST Microelectronics	Texas Instruments	UMC			
Texas Instruments	Toshiba				
Toshiba	TSMC				
TSMC	UMC				
UMC					

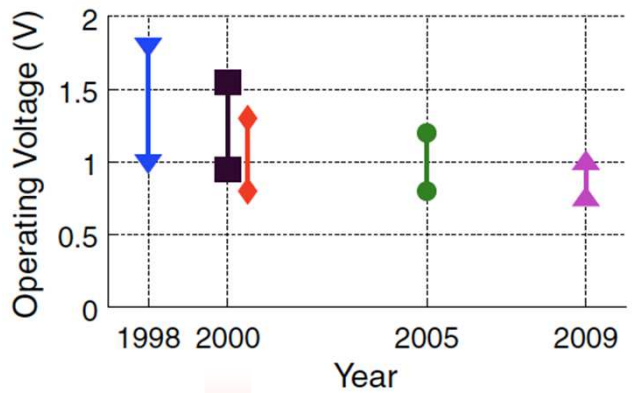
Market volume wall: only the largest volume products will be manufactured with the most advanced technology

life.augmented

The Twilight of Moore's Law: Power







[Watanabe et al., ISCA'10]

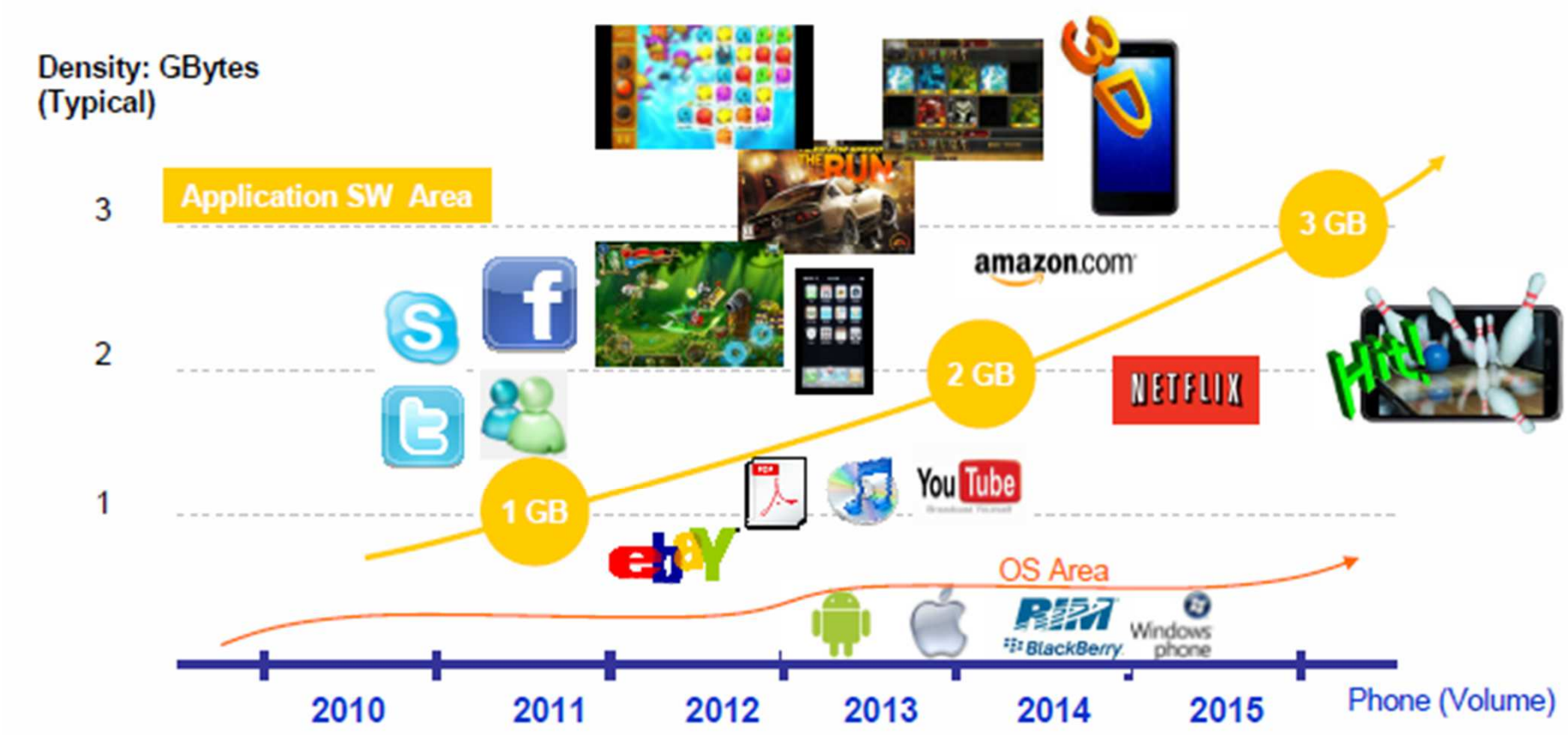
Dark Silicon !!!

Thermal wall: transistor count still increases exponentially but we can no longer power the entire chip (voltages, cooling do not scale)

[Hardavellas11]

life.augmented

The twilight of Moore's Law: IO Bandwidth

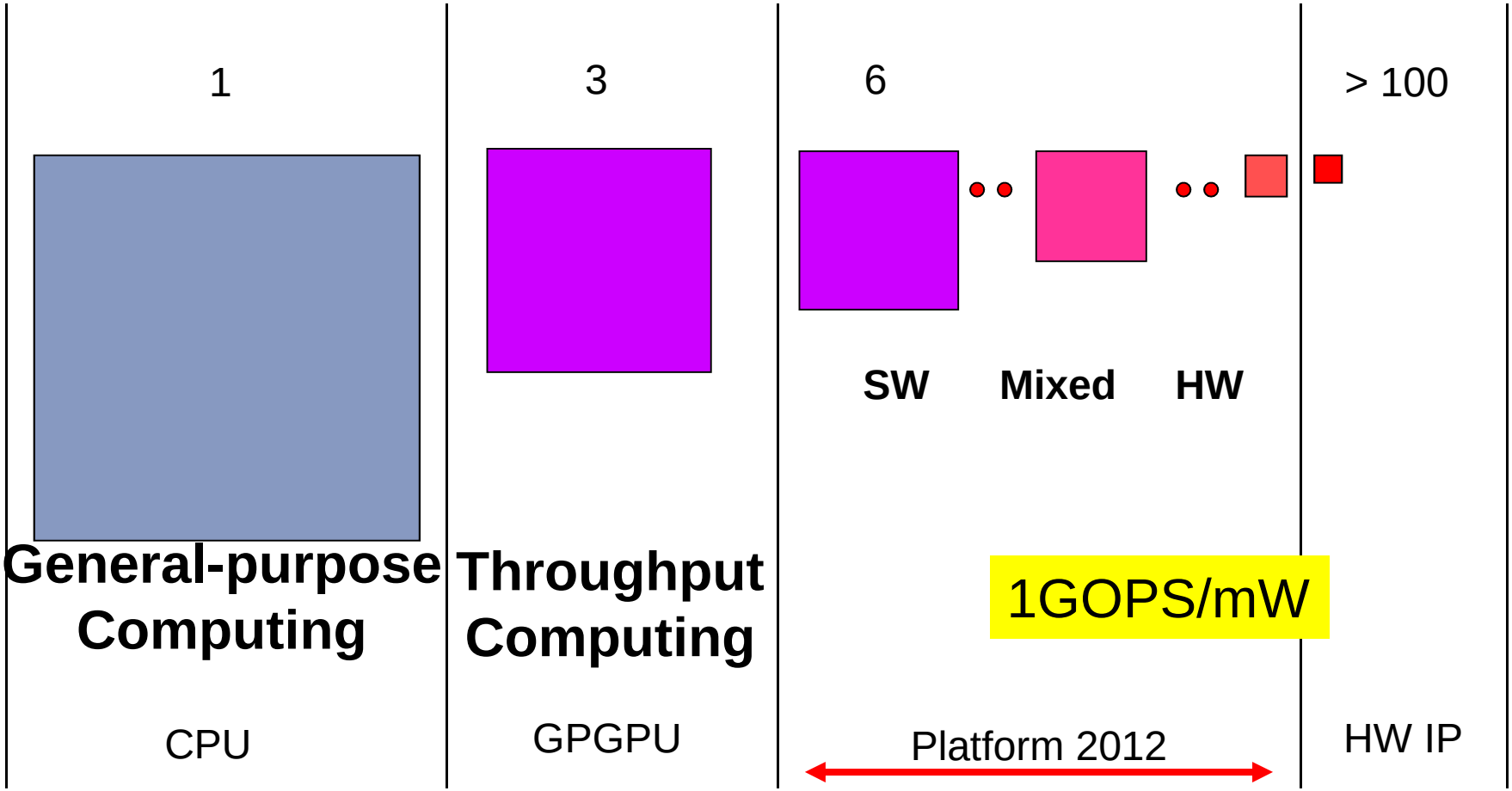


Memory wall: larger datasets and limited bandwidth at high power cost for accessing external memory

life.augmented

STMicroelectronics' Platform 2012

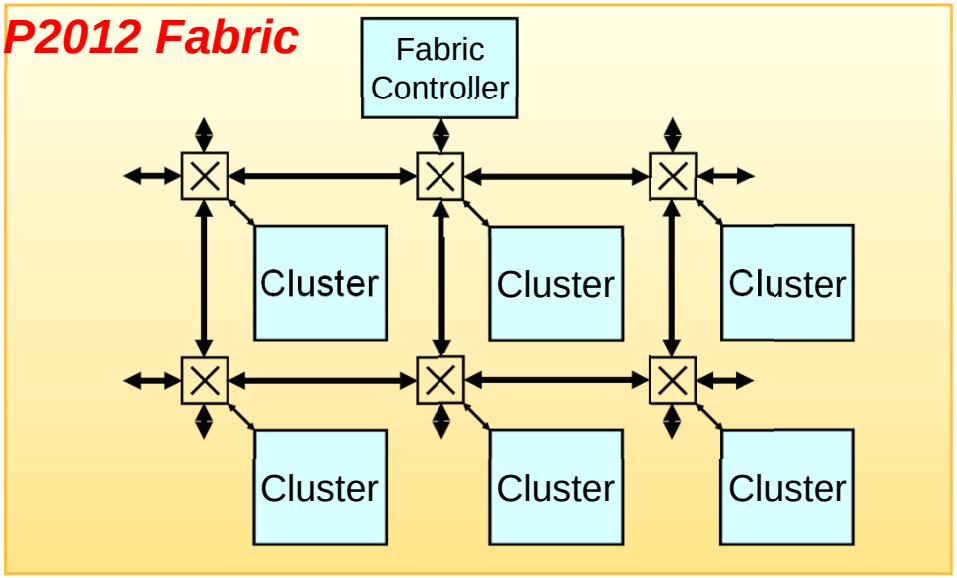
GOPS/mm^2 – GOPS/W

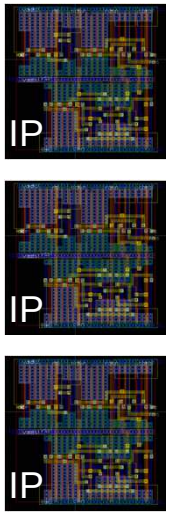


Closing The Accelerator Efficiency Gap

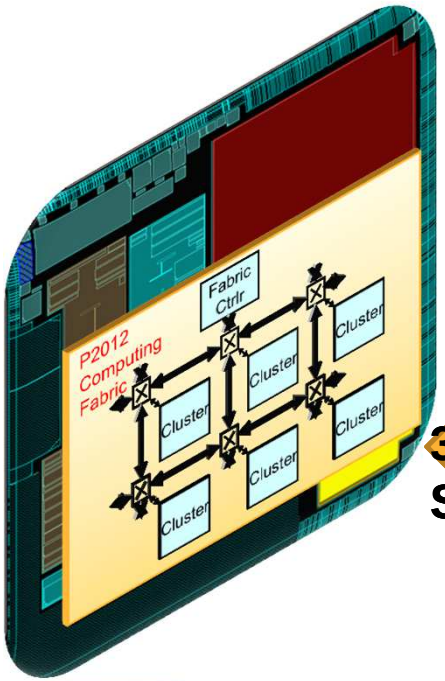
life.augmented

P2012 in a nutshell…



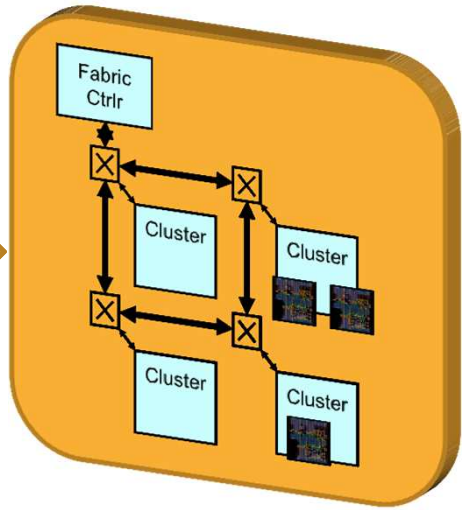


HW Optimized IPs



3D-stackable SW accelerator

Customization design flow



life.augmented

Value Proposition for STM

Heterogeneous Computing

Area/Power/Productivity

IPs for SoCs

Mixing HW and SW

- Video Codecs
- Imaging
- Base Band
- IQI
- …

?

P2012

Standalone SoC

X Gops
Small
LowPower

Programmable device

- Eco System
- Analytics
- Fragmented Mkts
- …

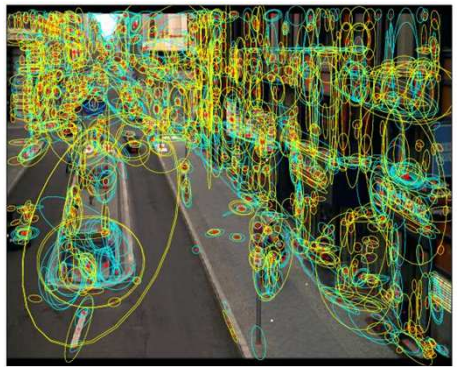
?

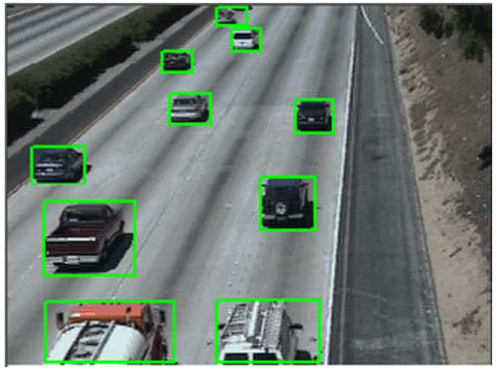
Flexibility/Quick Prototyping

Homogeneous Computing

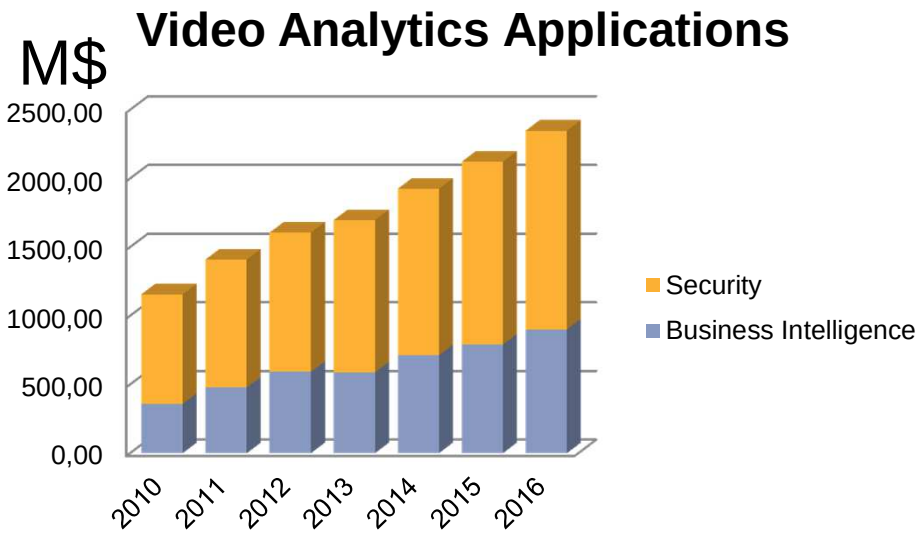
life.augmented

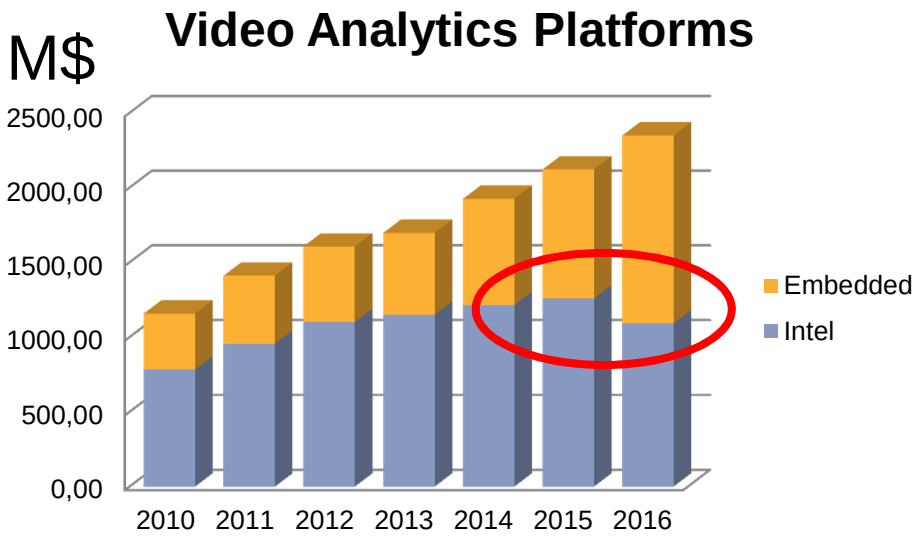
A Killer Application (domain) for P2012



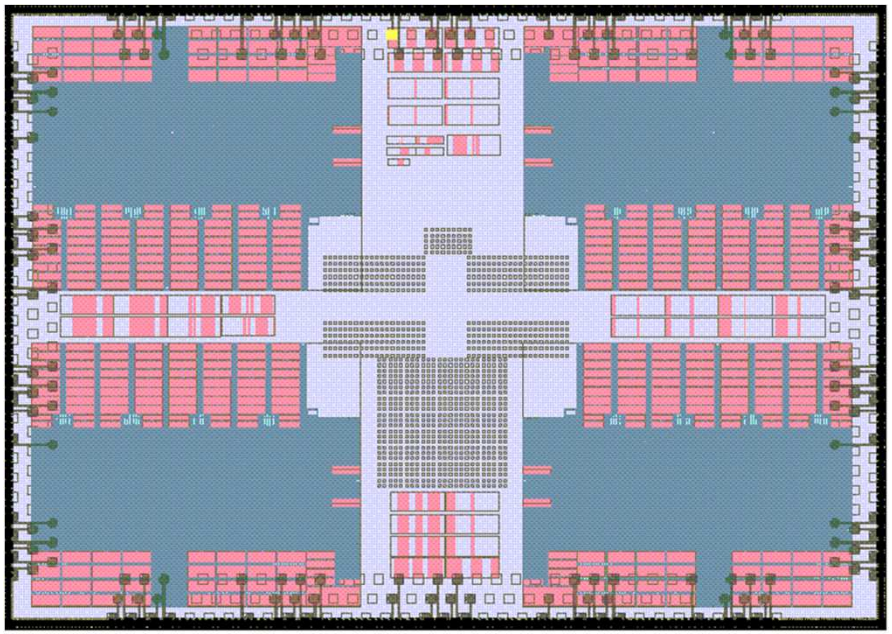
Embedded Visual intelligence





The next killer app: Machines that see (J. Bier)

P2012 SoC in 28nm



Taped out 2/3/12

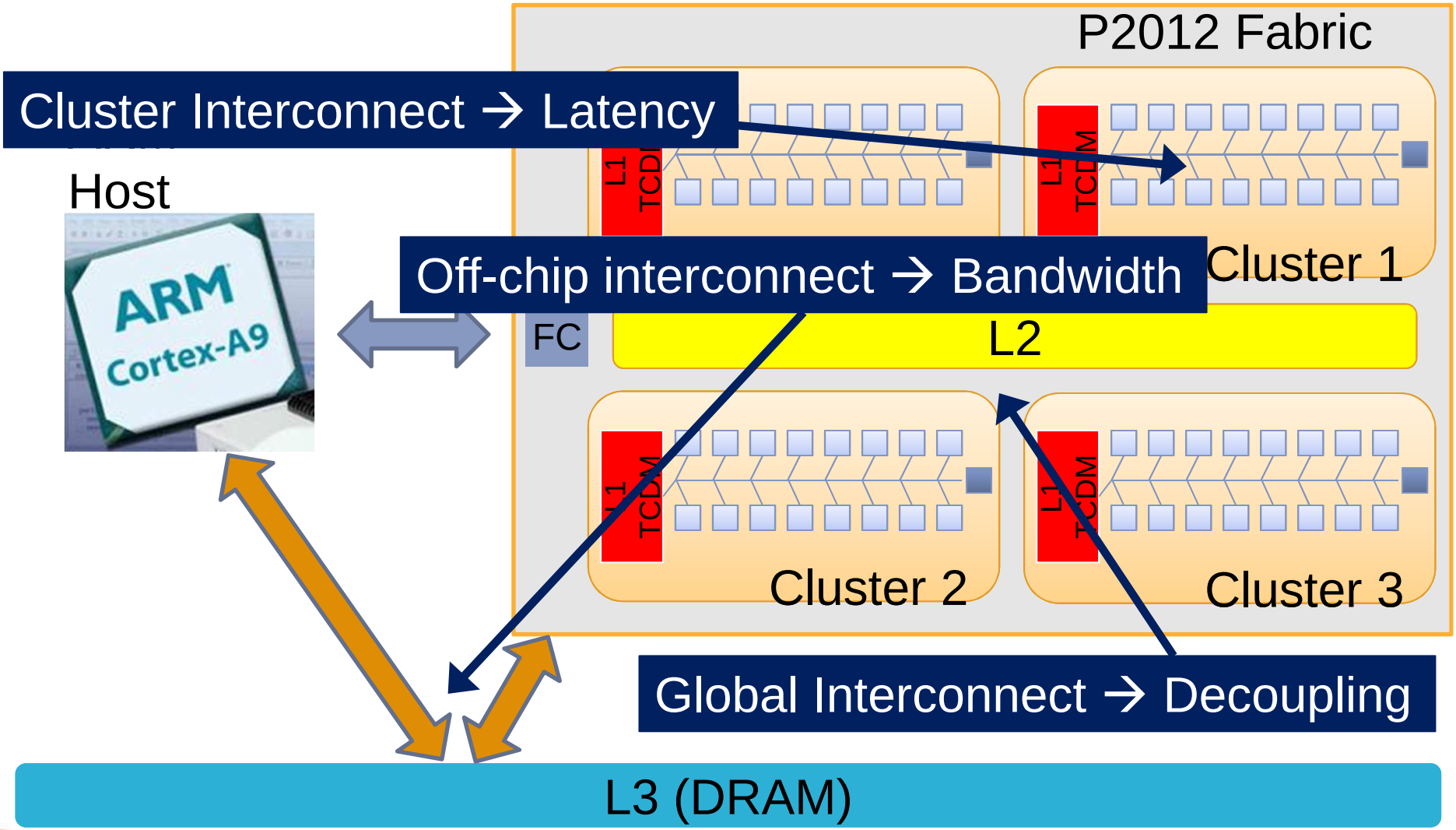
- 4 Clusters, 69 processors
- 80 GFlops
- 1MB L2 mem
- 2D flip chip or 3D stacked
- 600 MHz typ
- < 2 W
- 3.7 mm2 per cluster

Energy efficiency 40GOPS/W →0,04GOPS/mW

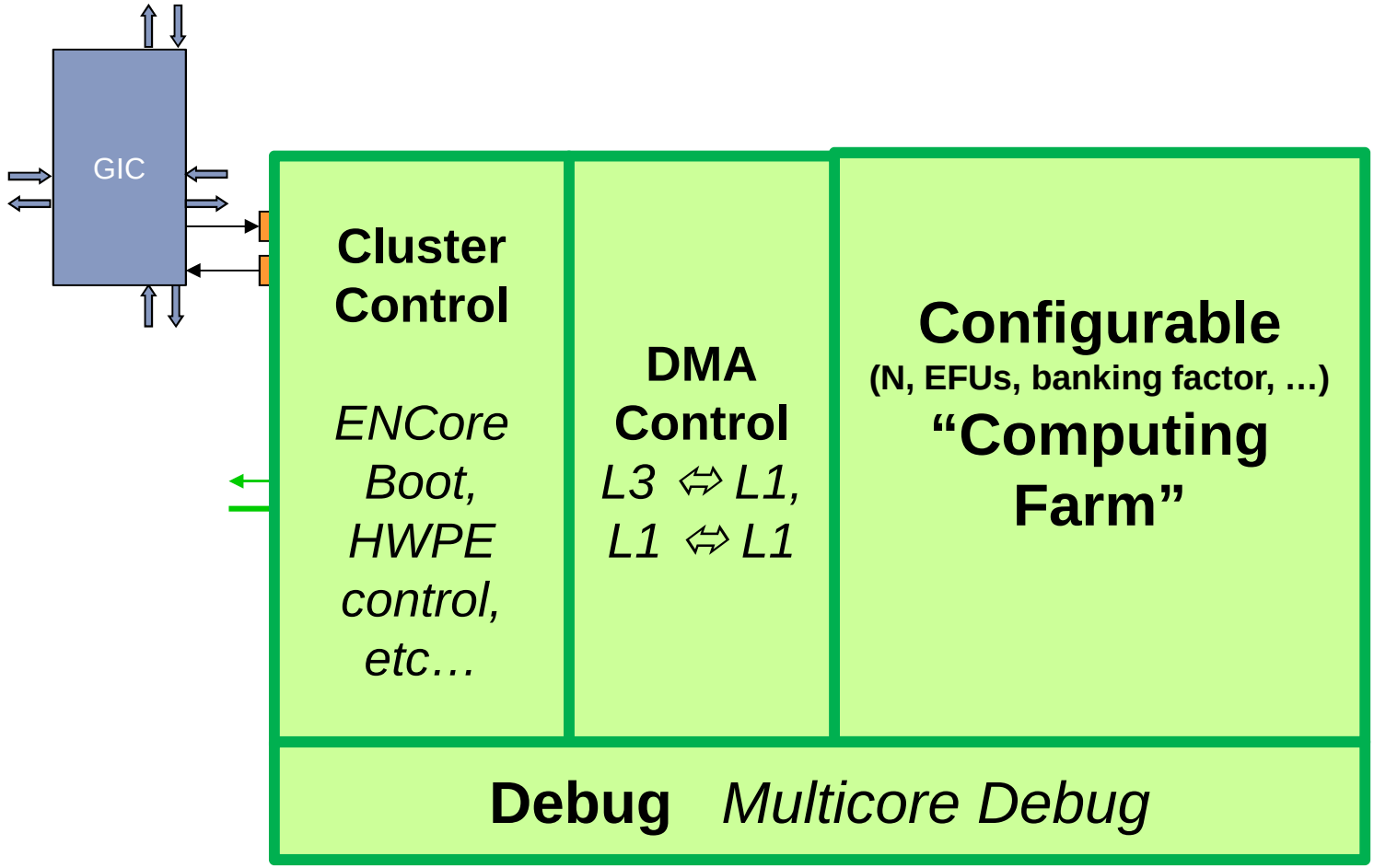
Designing the P2012 Computing SoC

Communication challenge

P2012 as GP Accelerator



The cluster


GIC
Cluster Control
ENCore Boot, HWPE control, etc…
DMA Control
L3 ⇔ L1, L1 ⇔ L1
Configurable
(N, EFUs, banking factor, …)
"Computing Farm"
Debug Multicore Debug

P2012 Cluster Main Features

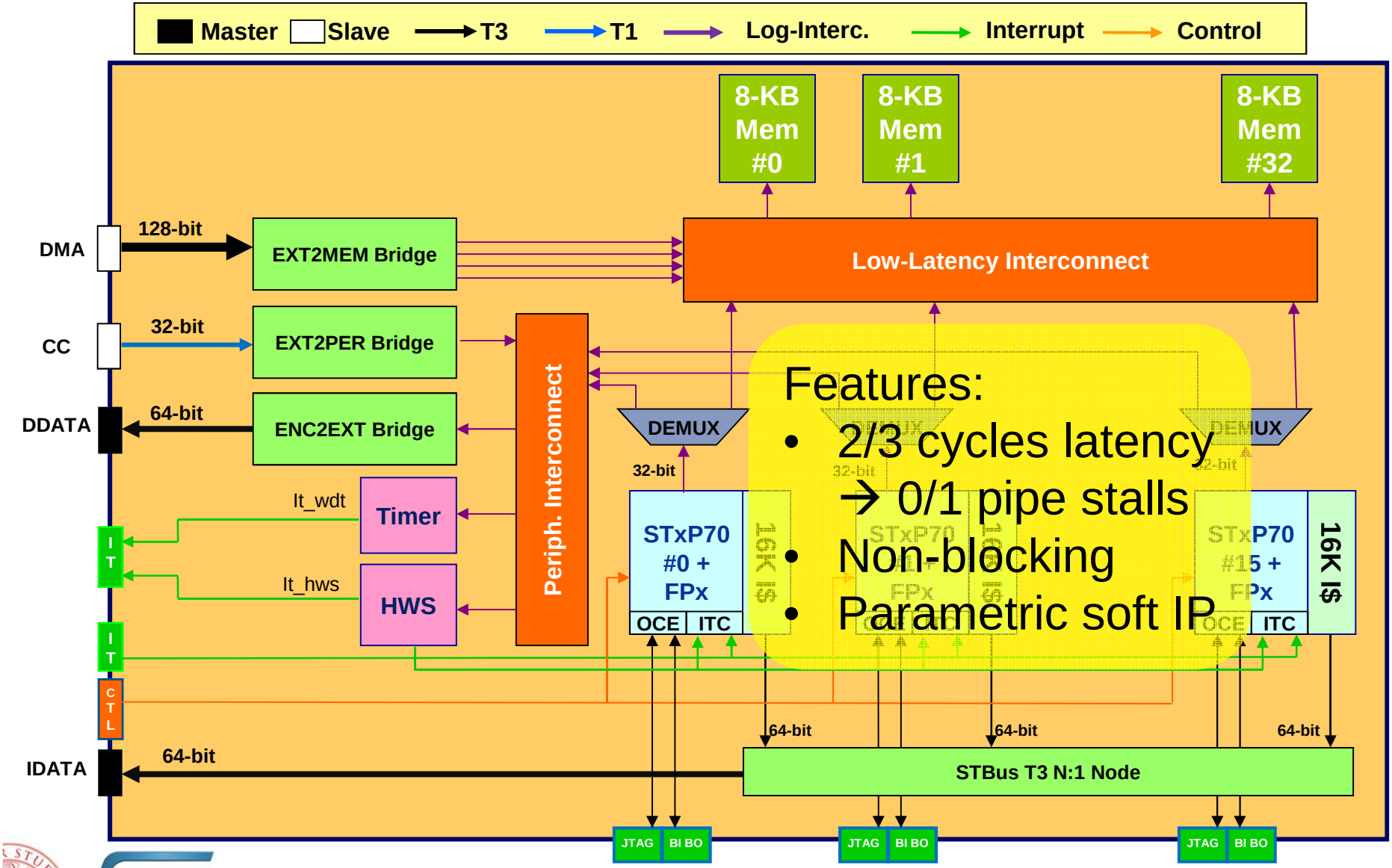
- Symmetric Multi-Processing
- Uniform Memory Access within the cluster
- Non-uniform Memory Access between clusters
- Up to 16 +1 processors per cluster.
- Up to 20.4 GOPS (32 bits) peak per cluster at 600 MHz
- 2 DMA channels allowing up to 6.4 GB/s data transfer

P2012 Cluster Main Features (Cont'd)

- HW Support for synchronization:
  - Fast barrier (within a cluster only) in ~4 Cycles for 16 processors
  - Flexible barrier ~20 cycles for 16 processors
- Seamless combination of non-programmable (HWPEs) and programmable (PEs) processing elements
- High level of customization though:
  - The number of STxP70 processing elements
  - The STxP70 extensions (ISA customization)
  - Up to 32 user-defined H/W PEs
  - Memory sizes
  - Banking factor of the shared memory

Cluster interconnect

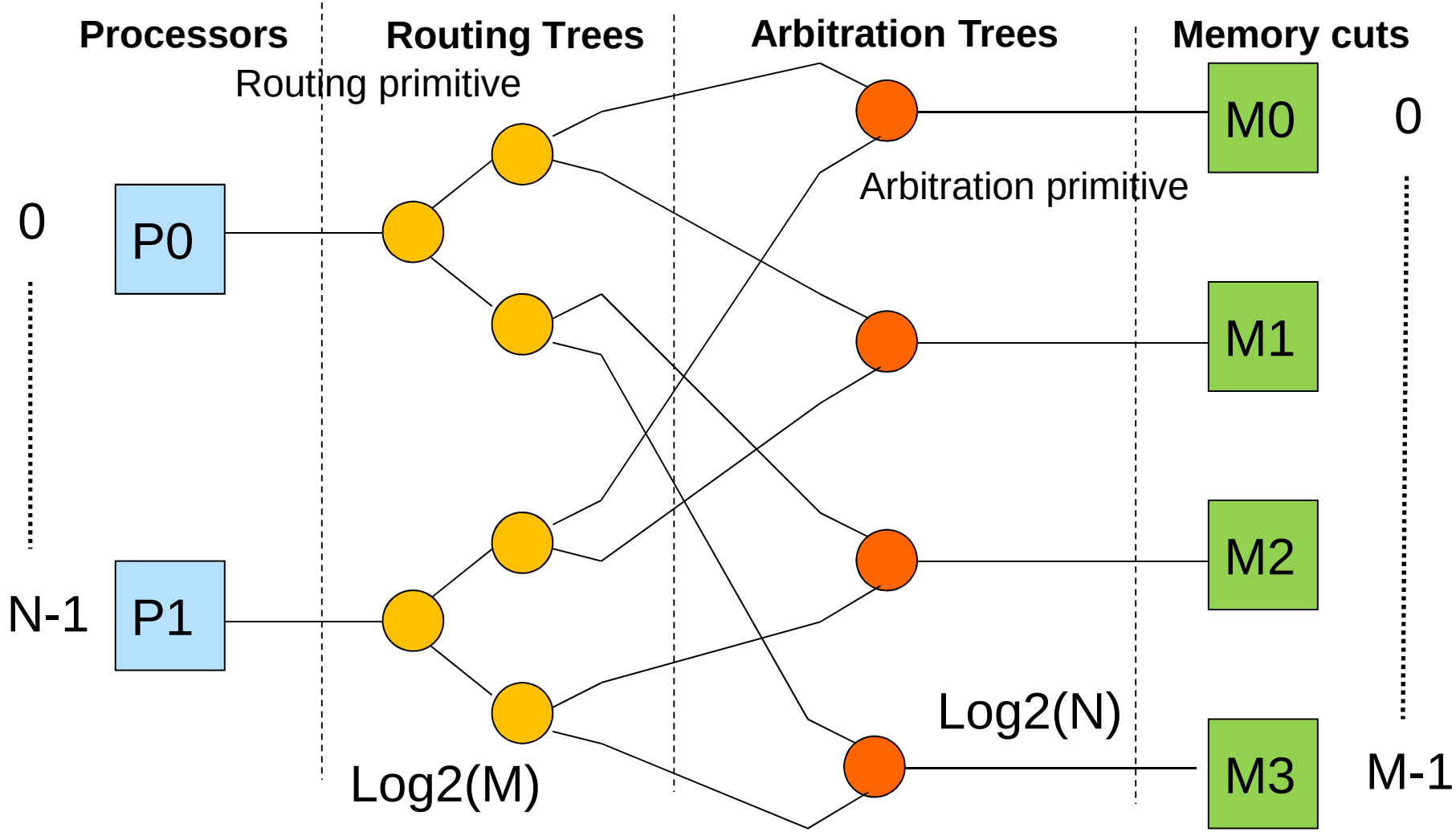
ENCore16



Features:

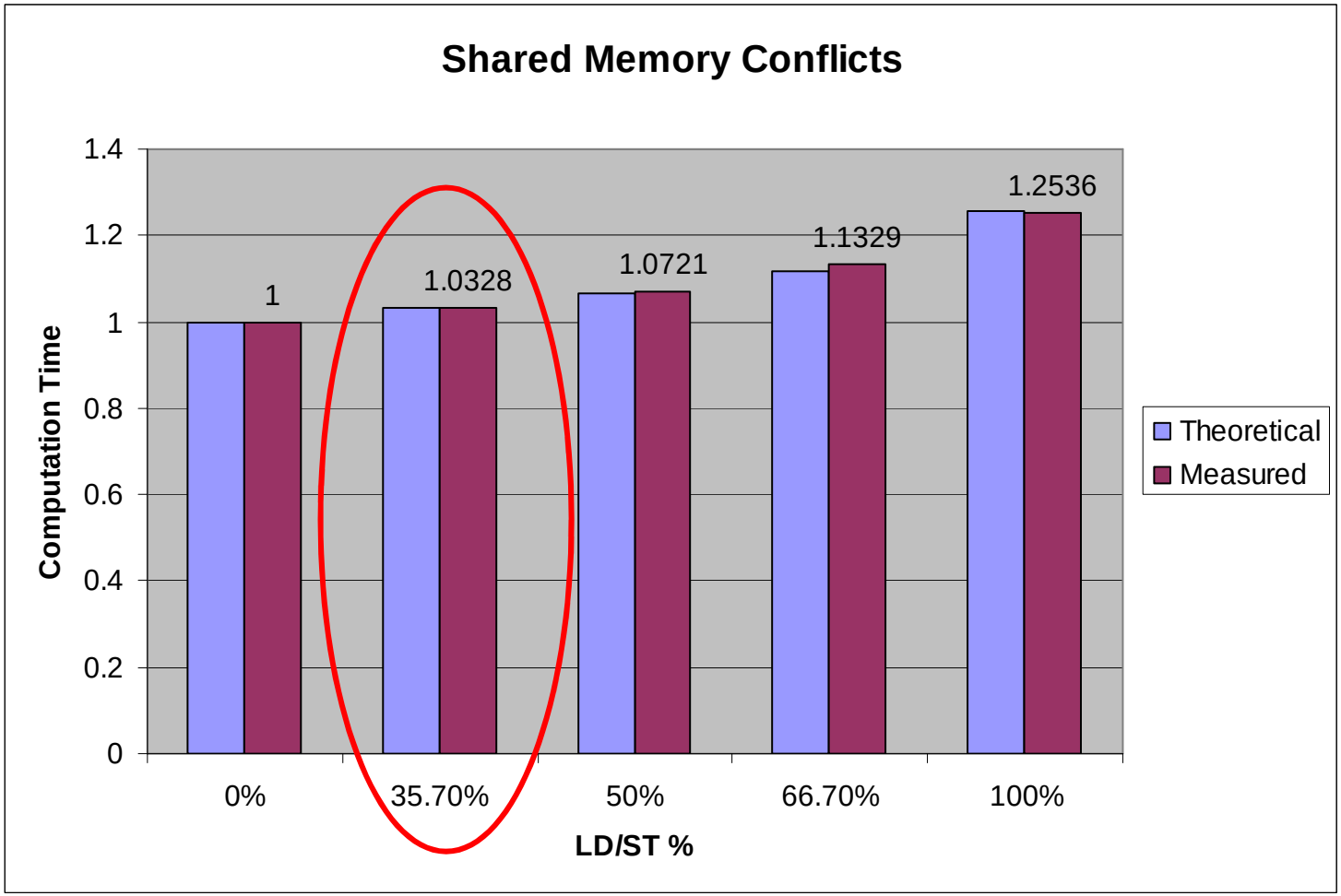
- 2/3 cycles latency
  → 0/1 pipe stalls
- Non-blocking
- Parametric soft IP

Logarithmic Mesh of Trees



Fine-grained Interleaving → routing based on LSBs of addr.

Log ICO Performance

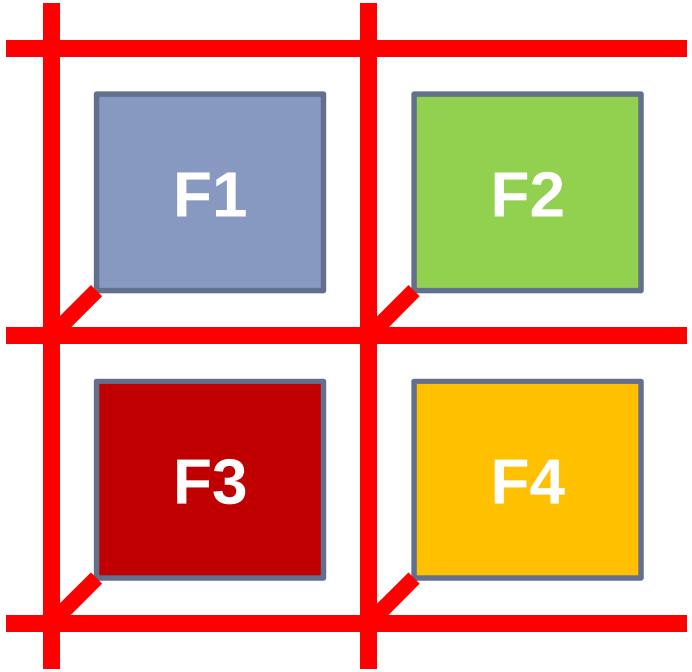


Typical case → worst case?

Global interconnect

Cluster Decoupling

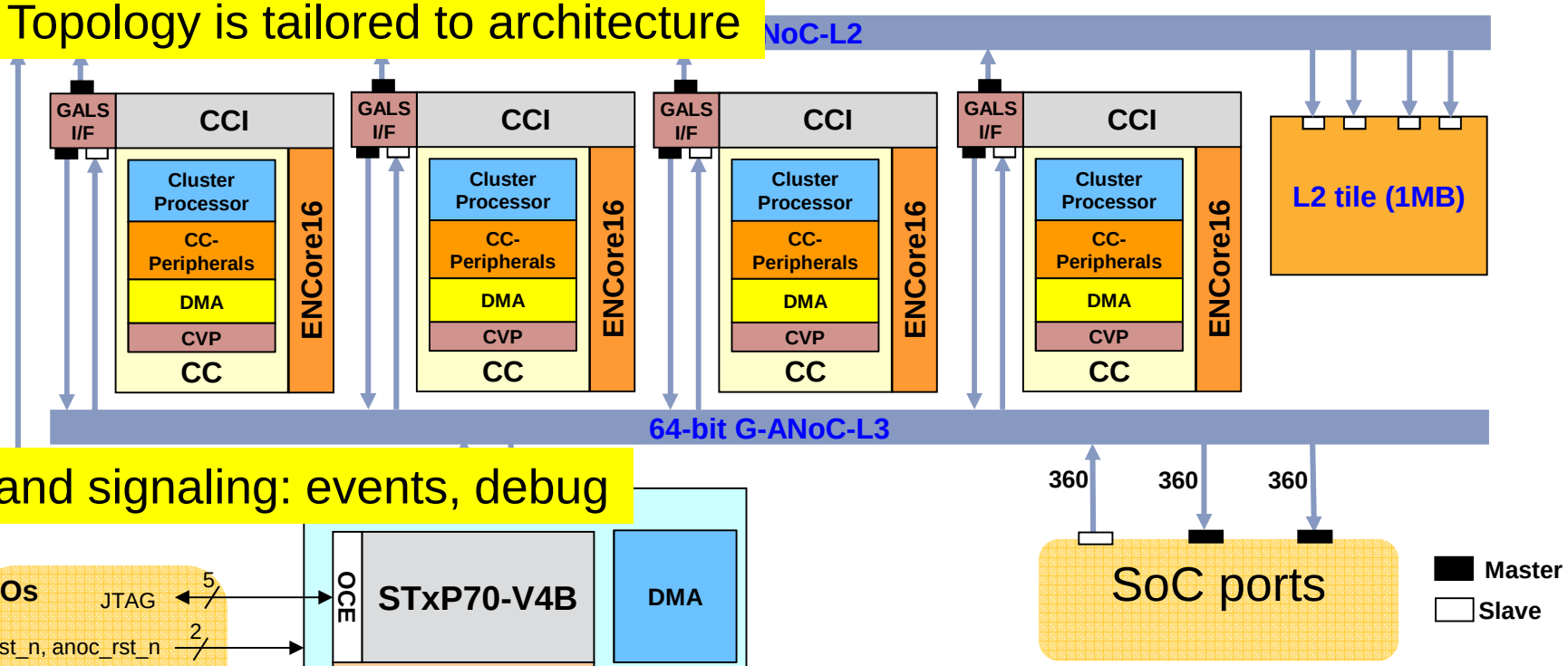
- **Each cluster has its own operating point tunable to the right energy budget**



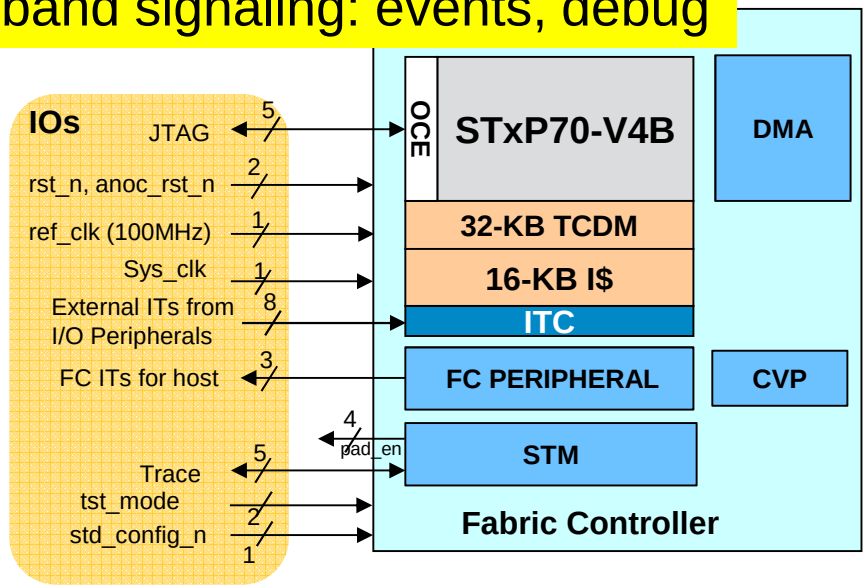
- **Inter cluster communication is handled by a fully asynchronous network on chip**

P2012 SoC GANoC

Topology is tailored to architecture



Sideband signaling: events, debug



System plug: asymmetric bandwidth

Off-chip interconnect

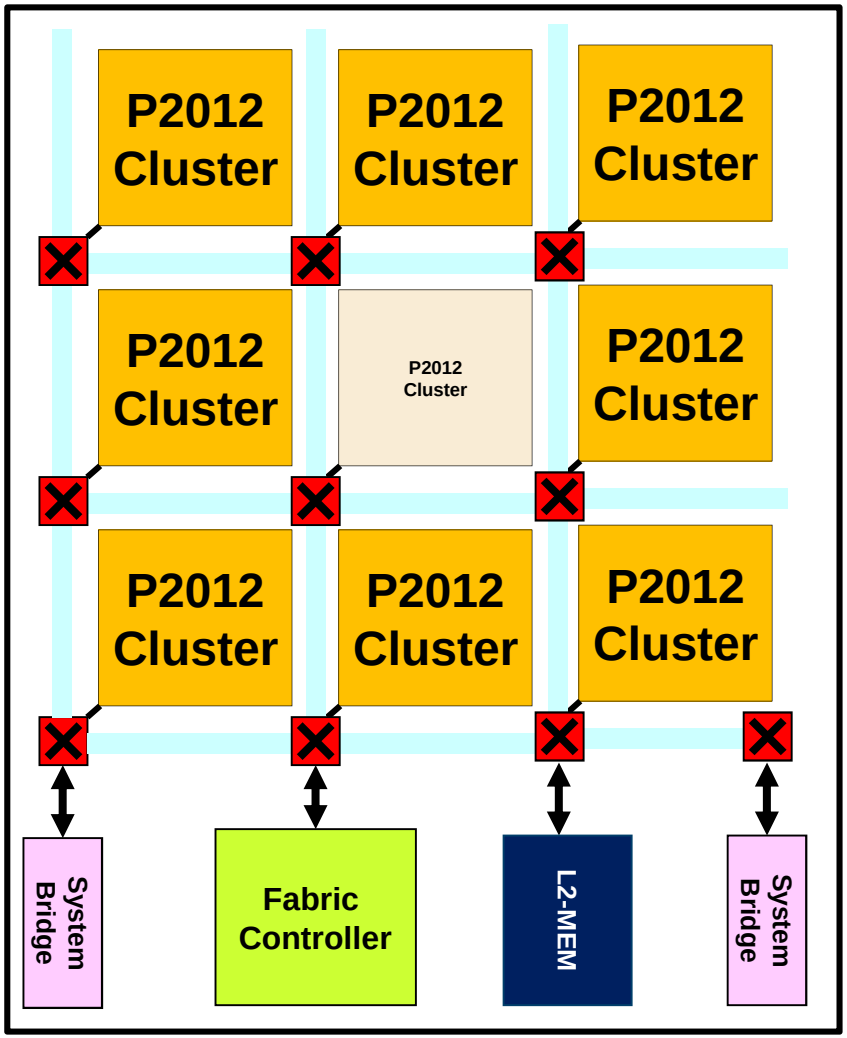
LD/ST and DMA memory transfers

LOG-ICO

GANOC

??

- Intra-Cluster:
  - LD/ST (UMA)
  - DMA: From/to TCDM to/from HWPE
- Inter-Cluster:
  - LD/ST (NUMA)
  - DMA: L1-to/from-L1
- Cluster to/from L2-Mem:
  - LD/ST (NUMA)
  - DMA: L1 to/from L2
- Cluster to/from L3-Mem (though the system bridge):
  - LD/ST (NUMA)
  - DMA: L1 to/from L3



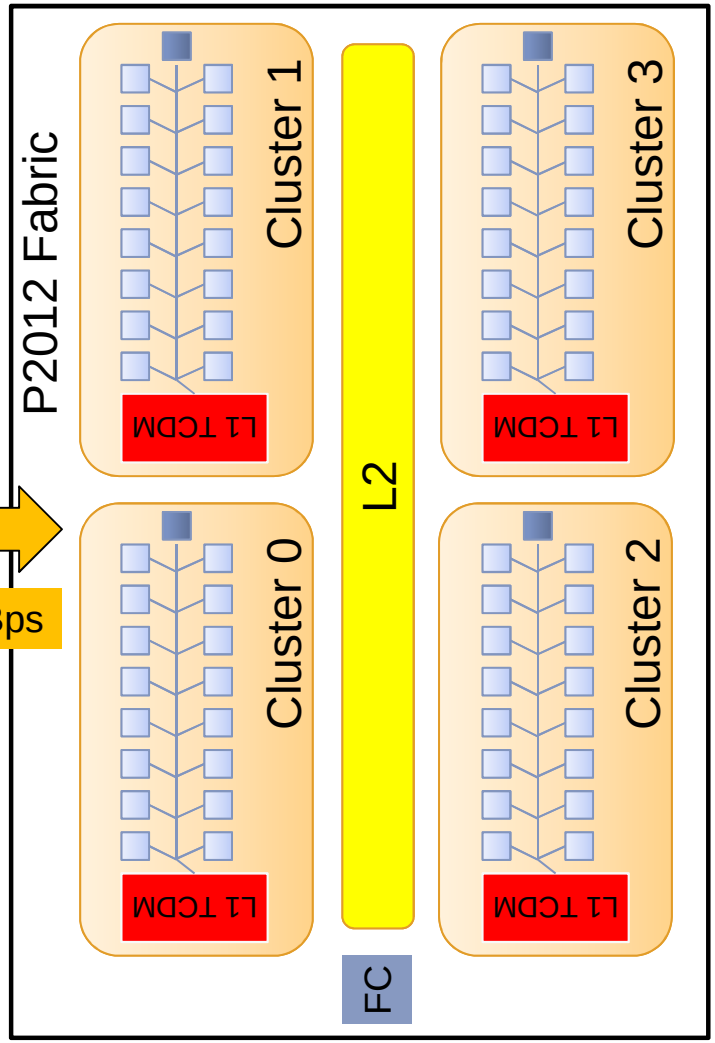
DMA BW 6.4GBps/Cluster → external memory bandwidth?

Short-term 2D strategy: FPGA SoC host

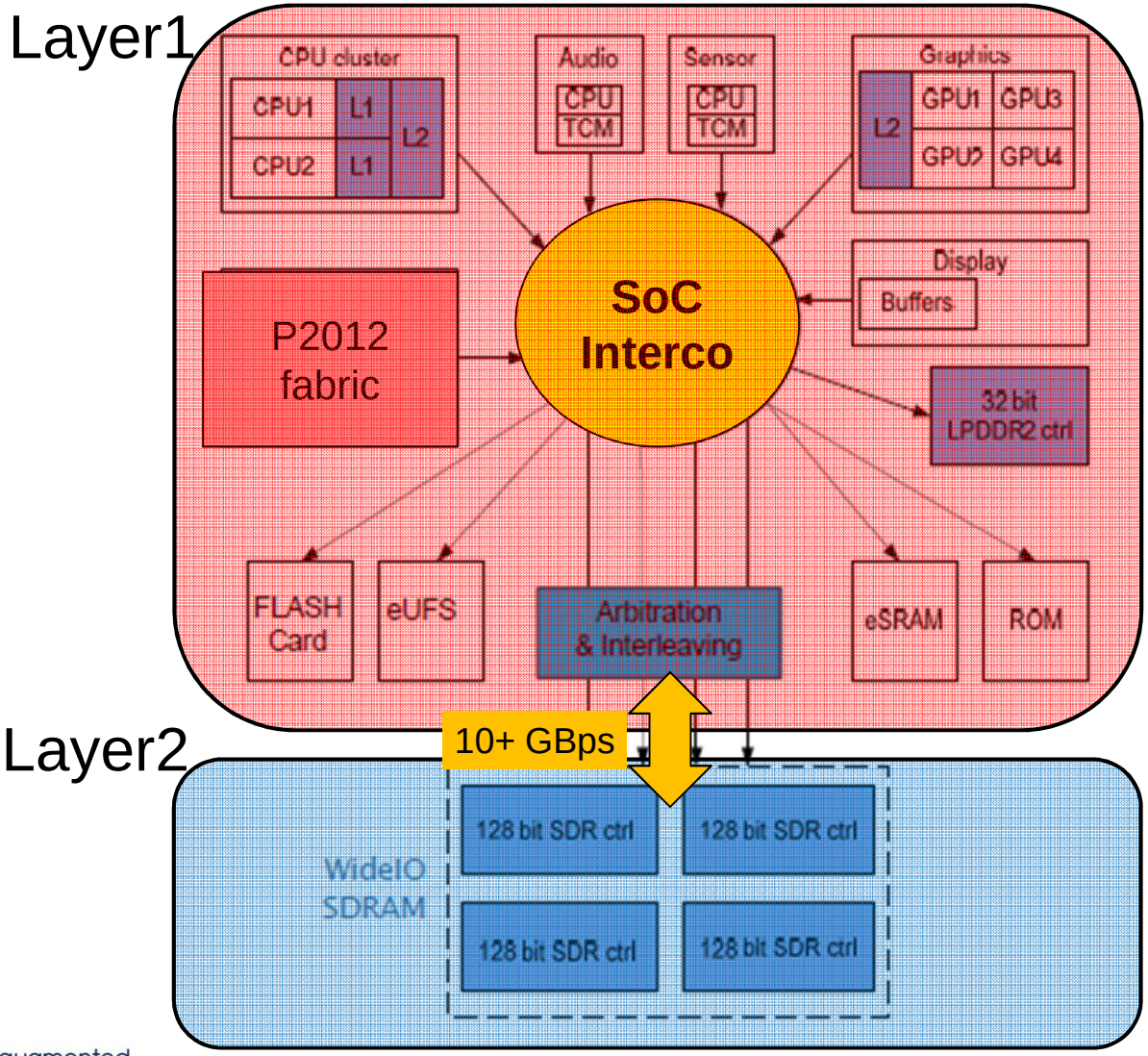
L3 (DRAM)

- Dual ARM Cortex™-A9 MPCore (800MHz, NEON, 32KL1-I$, 32KL1-D$, 512K SH-L2)
- DDR3, DDR2 and LPDDR2 CTRL
- 2x QSPI, NAND Flash and NOR Flash Memory Controller
- 2x USB2.0 (OTG), 2x GbE, 2x CAN2,0B 2x SD/SDIO, 2x UART, 2x SPI, 2x I2C, 4x 32b GPIO
- Advanced Low Power 28nm Programmable Logic:
  - 28k to 350k Logic Cells (approximately 430k to 5.2M of equivalent ASIC Gates)
  - 240KB to 2180KB of Extensible Block RAM
  - 80 to 900 18x25 DSP Slices (58 to 1080 GMACS peak DSP performance)
- PCI Express® Gen2x8 (in largest devices)
- 154 to 404 User IOs (Multiplexed + SelectIO™)
- 4 to 16 12.5Gbps Transceivers (in largest devices)

1GBps



SoC integration: Logical view

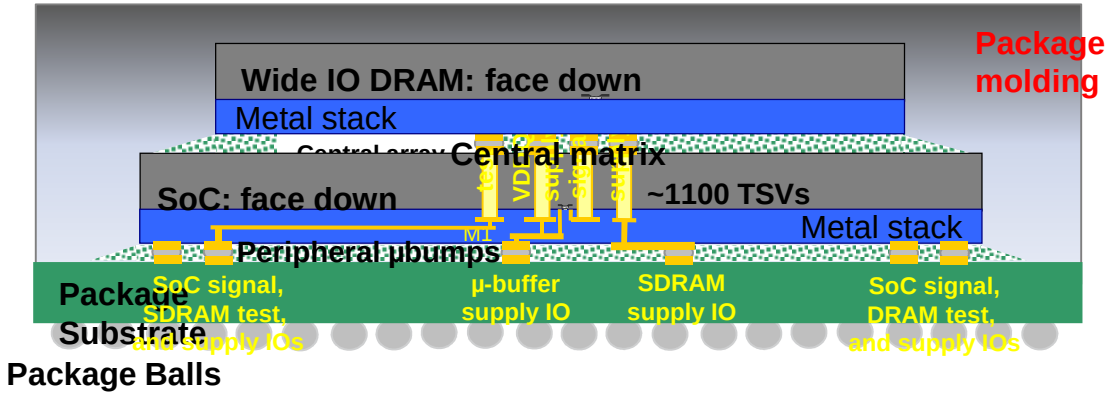


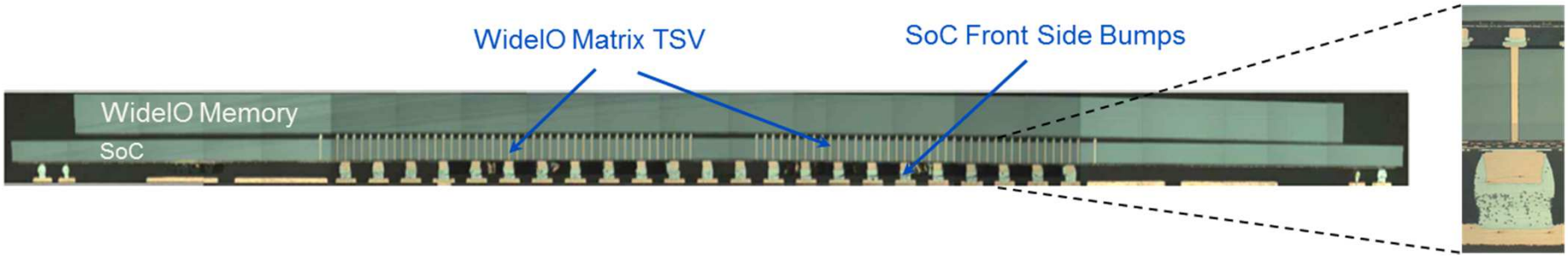
life.augmented

P2012 – 3D Wioming test chip

June 2012 ….

- **Fully functional**
- **High yield**
- **Actual perf higher than expected**





life.augmented

SW & Tools

SW & Tools: Key challenges

- SW Ecosystem: Sticking to SW and SW integration standards throughout the development chain.

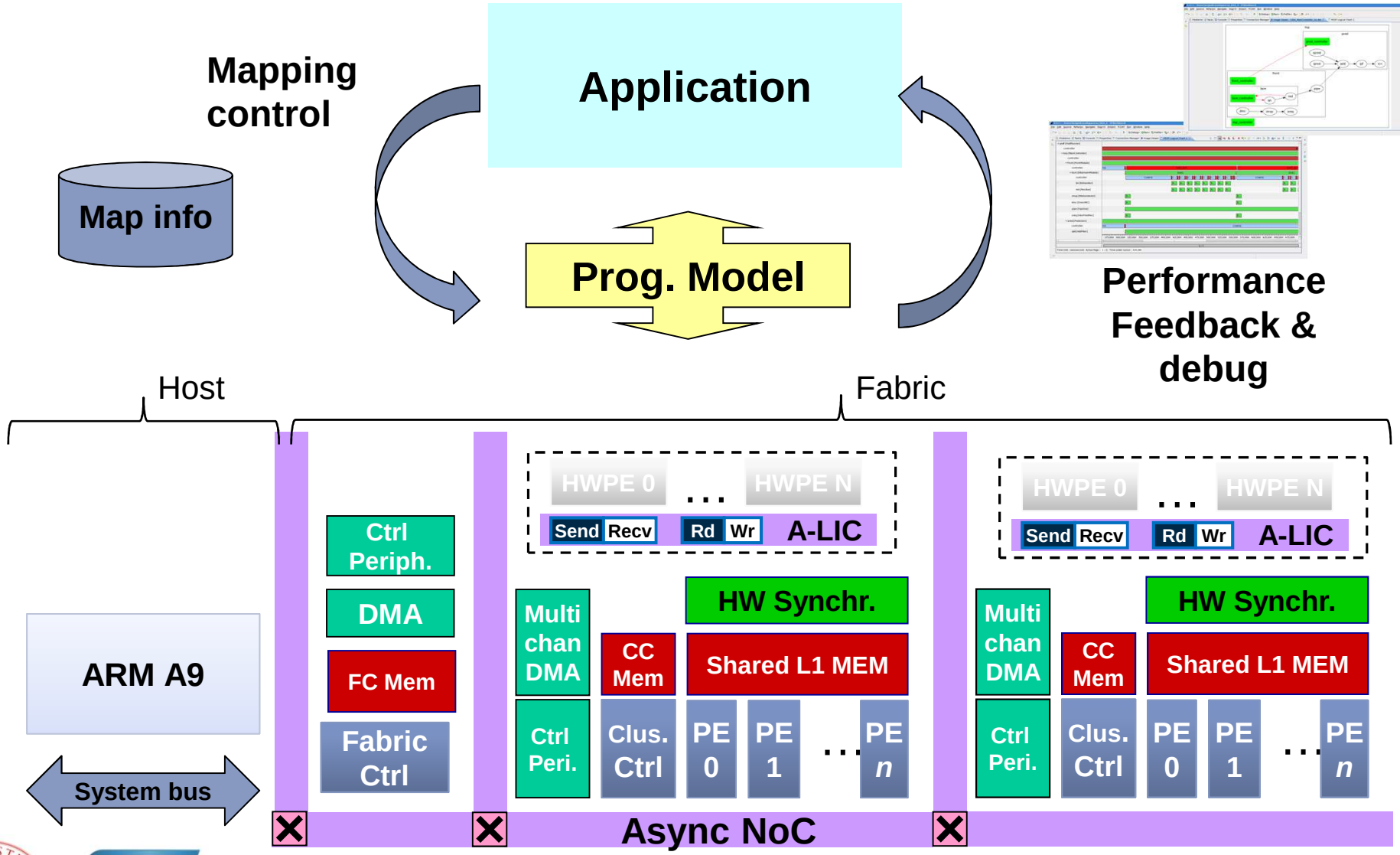
- App Ecosystem: Enabling SW development and application content well ahead of time of Silicon

- Platform Ecosystem: Enabling short iteration loop with all the actors from academic R&D up to product owner

And be efficient (GOPS/W$)!

life.augmented

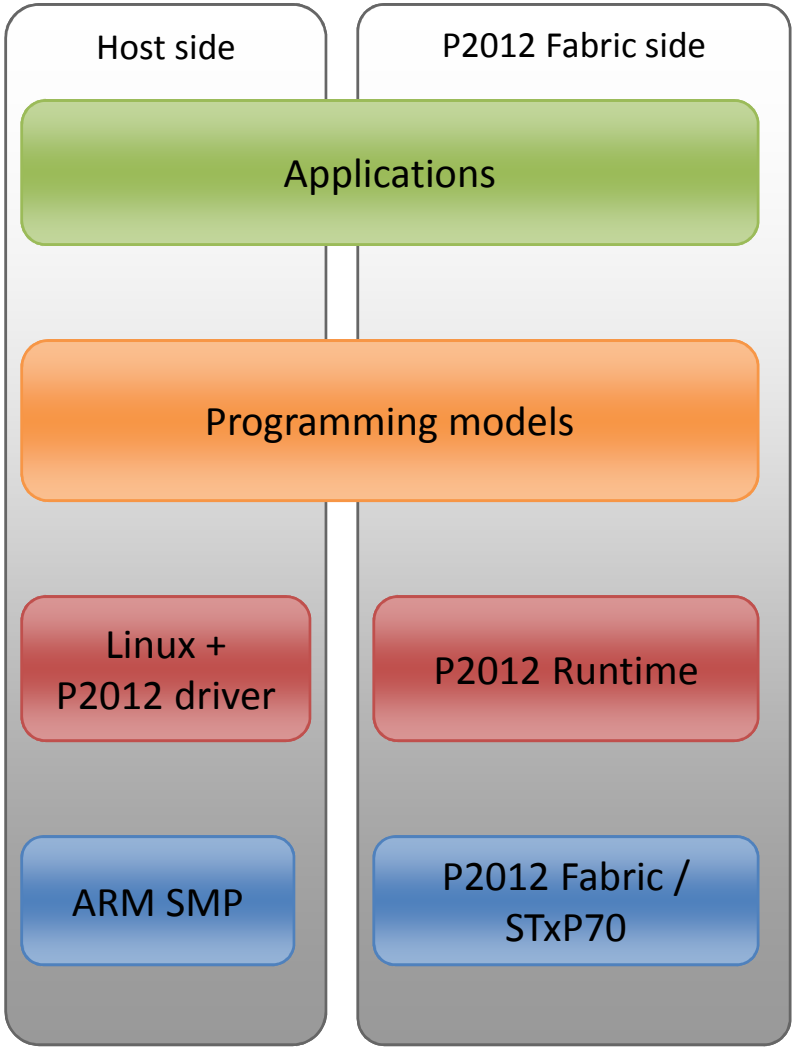
Programming P2012



P2012 Programming Models

- S/W-based platform variant
  - OpenCL
    - CLAM compiler (CL Above Many-Core)
  - NPM (Native Programming Model)
    - Based on MIND components for code partitioning
    - Communication components pulled from a provided library
    - Fine grain parallelism through runtime C-API
- Mixed H/W-S/W platform variant
  - PEDF (Predicated Execution Data Flow)
    - Dataflow based programming
    - Support of H/W PEs accessed via streaming interfaces
    - Not currently part of the public SDK

Software stack (for S/W-based platform)



Focus on augmented-reality applications (based on OpenCV).
Note that P2012 is a GP platform

Focus on OpenCL and dedicated programming model (NPM)

Assume the ARM is running a Linux system

Assume P2012 acts as an accelerator to an Android ARM-SMP system

Programming Heterogeneous Parallelism

Task Parallelism
(run-to-completion)

OpenCL

Data Parallelism
(with some-synchro)

ARM Coretex-A15

P2012

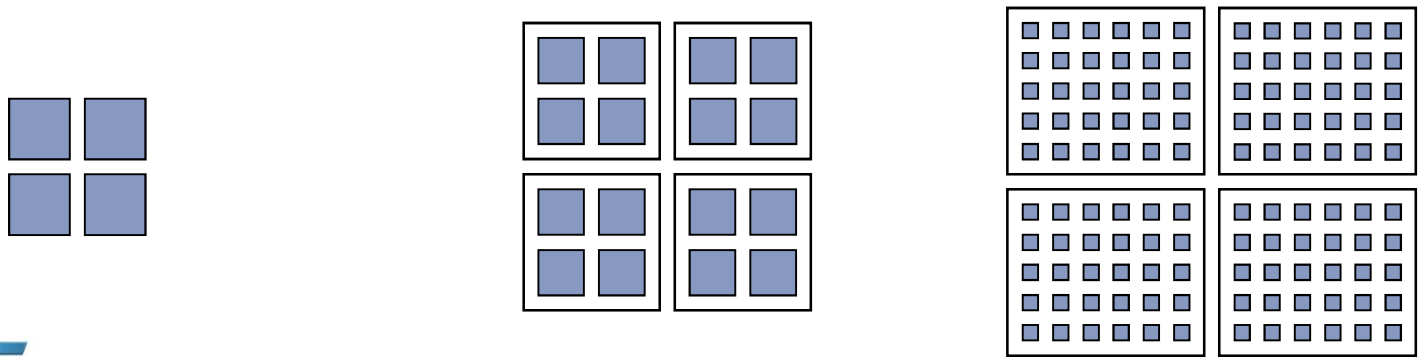
Rogue

Imagination

More parallelism

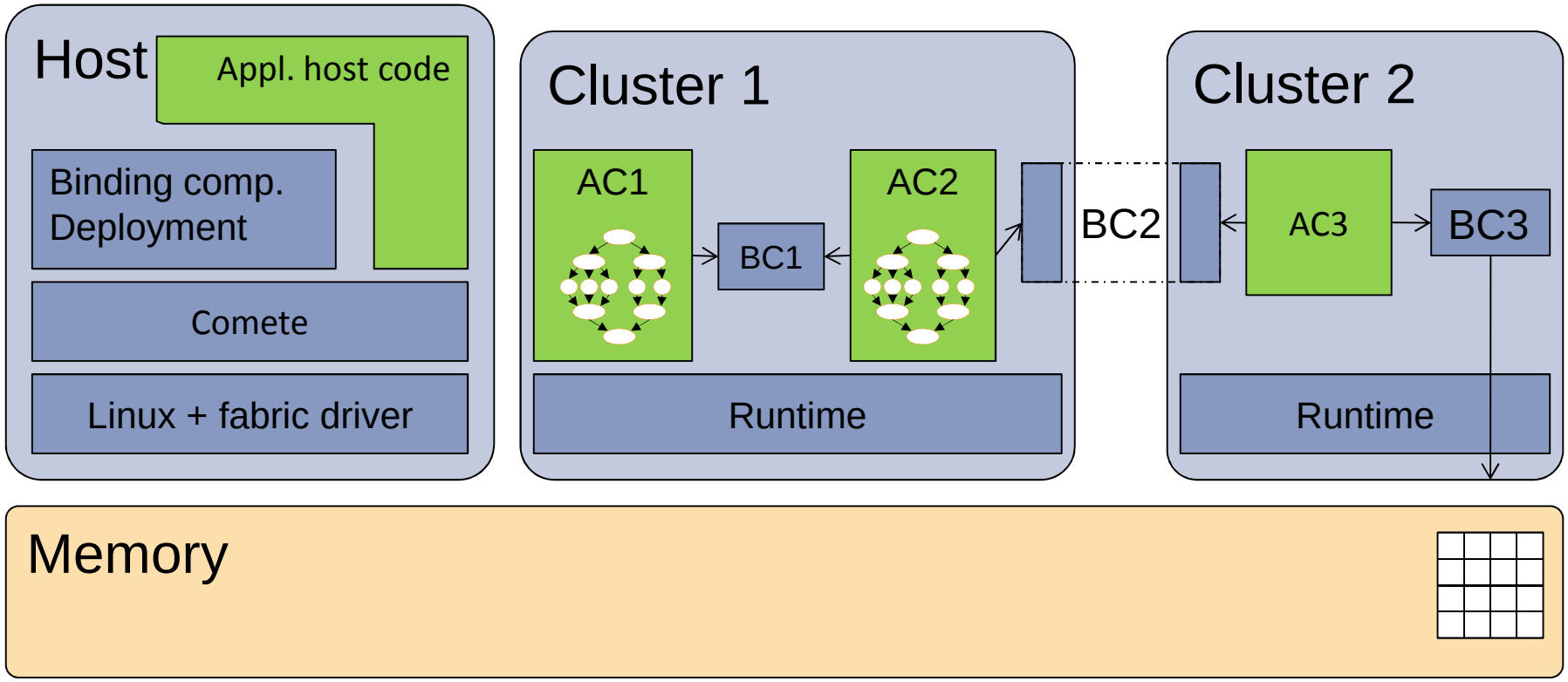
More programmability

Multi-core CPU | Many-core | GPU



life.augmented

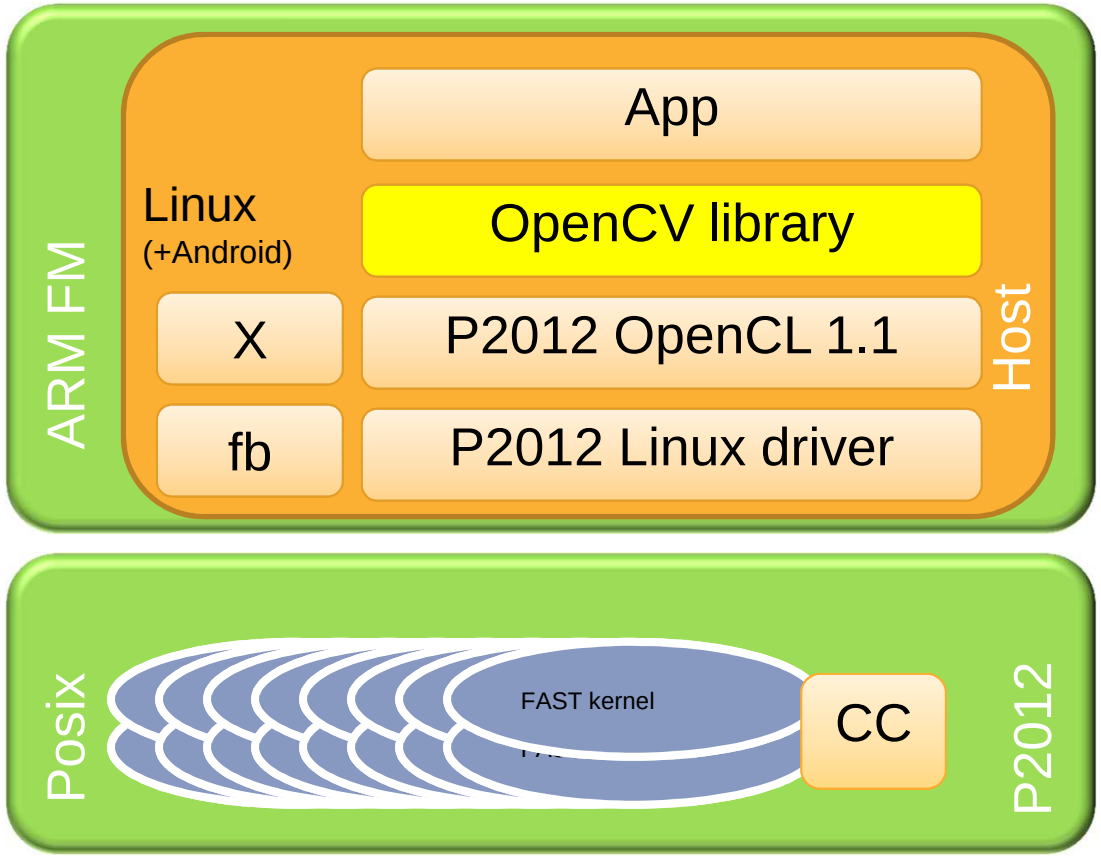
Native Programing Model



- Application Component (computation → actor) + internal parallelism
- Binding Component, from a library (communication)
- Building blocks for data-flow with HW & SW filters → heterogeneous computing (e.g. based RVC-CAL MPEG standard)

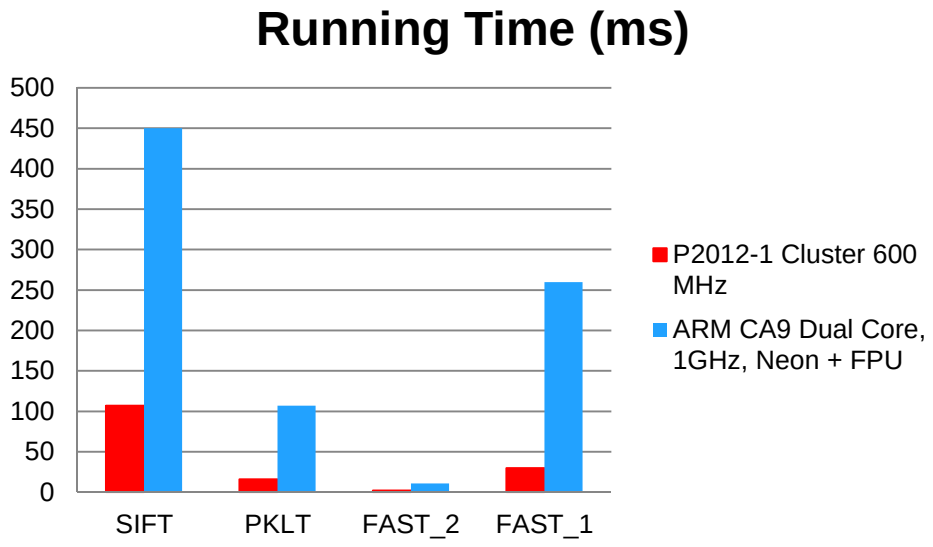
Image Understanding: OpenCV on P2012

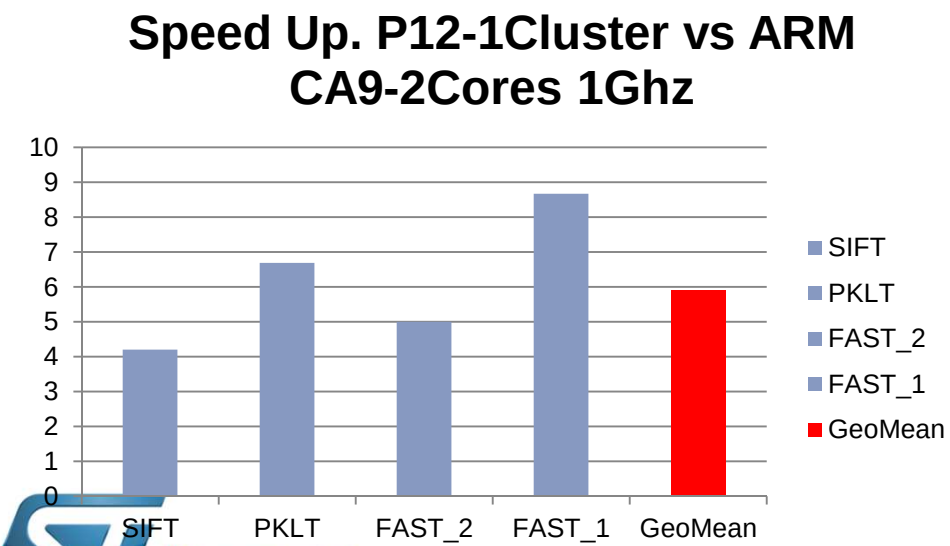
OCV functions accelerated with P2012 (transparently for the App programmer) → Standard domain specific APIs

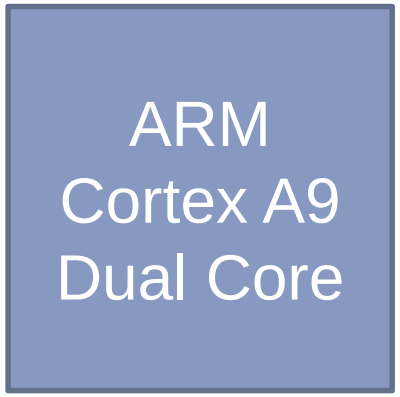


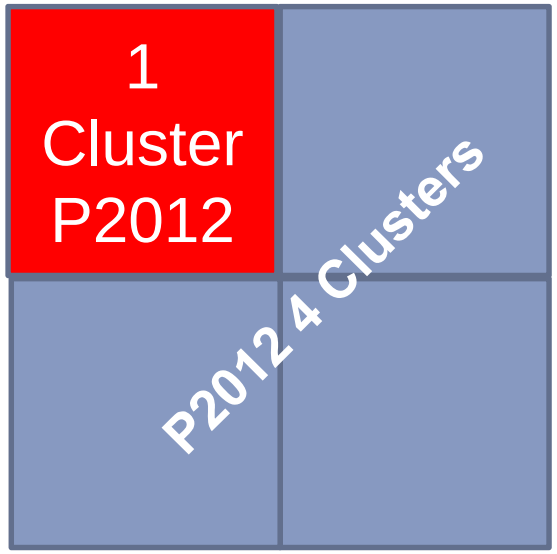
FAST: key point detection

Visual Analytics Benchmarks



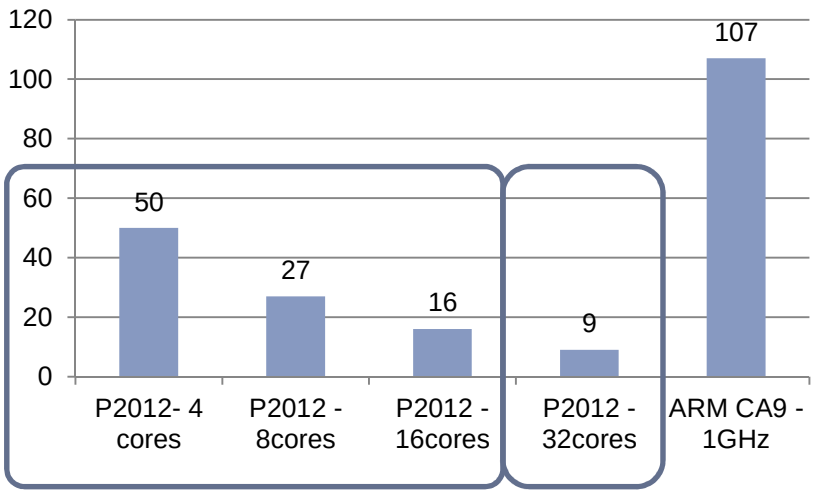






Scalability (PKLT)

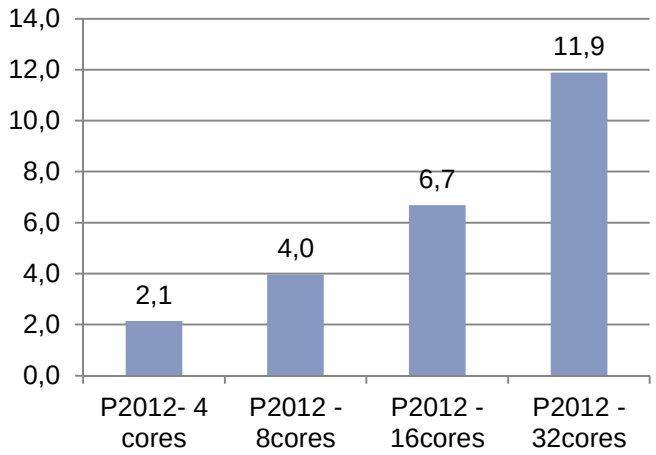
Running time [ms]



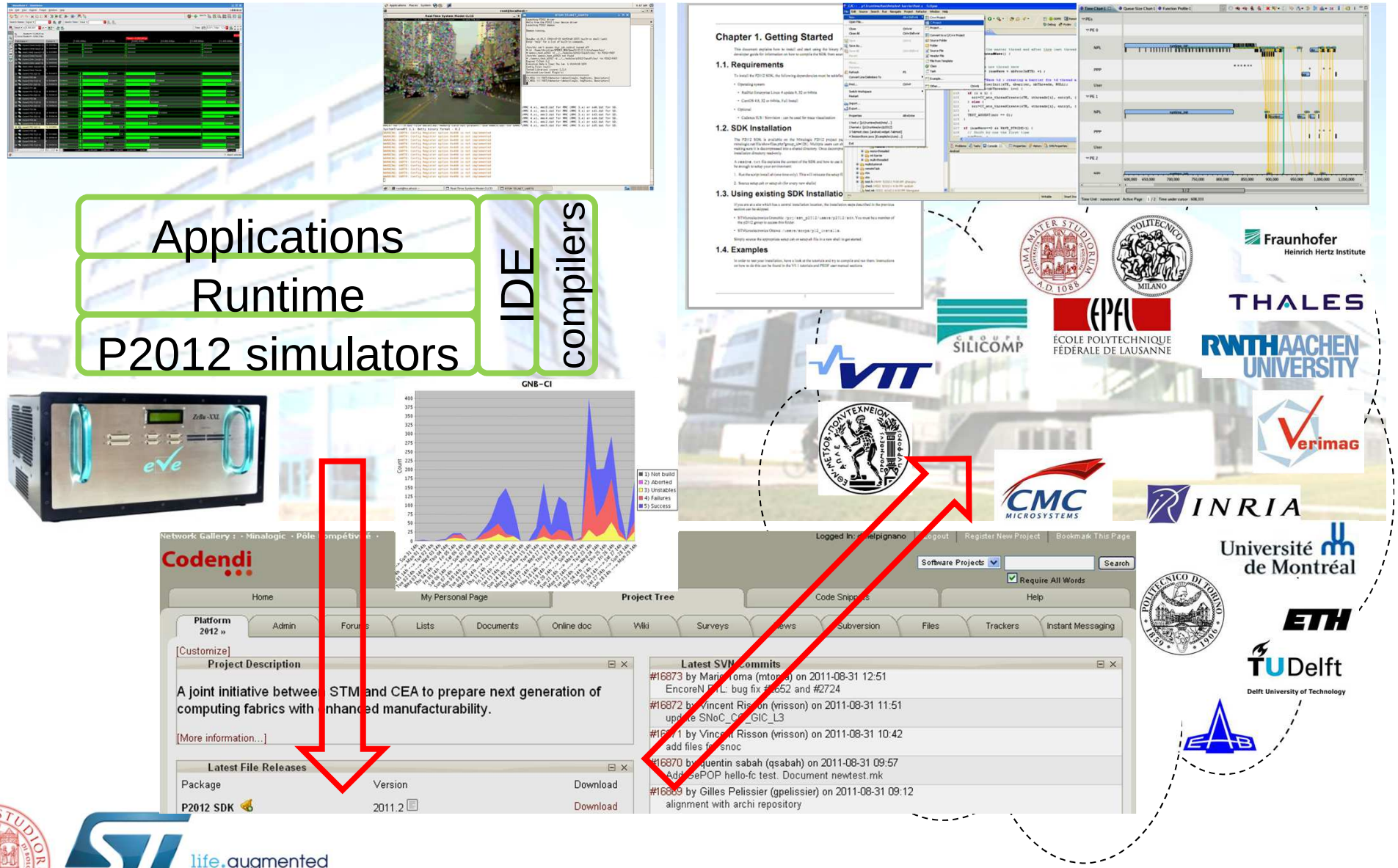
1 cluster: varying ND range

2 clusters

Speedup vs ARM CA9



P2012: SDK / Software Ecosystem



Allocation and scheduling for P2012

Managing memory

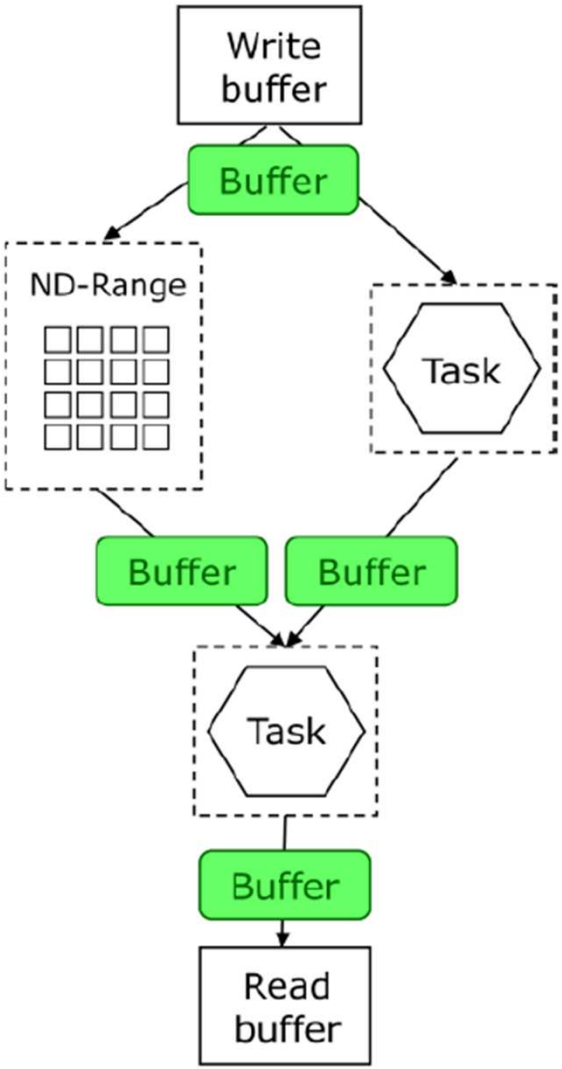
P2012 ≠ GPU: Kernel level parallelism

Independent Clusters + Independent processor IF

- Work-item divergence is not an issue
- P2012 supports more complex OpenCL task graph than GPUs. Both **task-level** and **data-level (ND-Range)** are possible

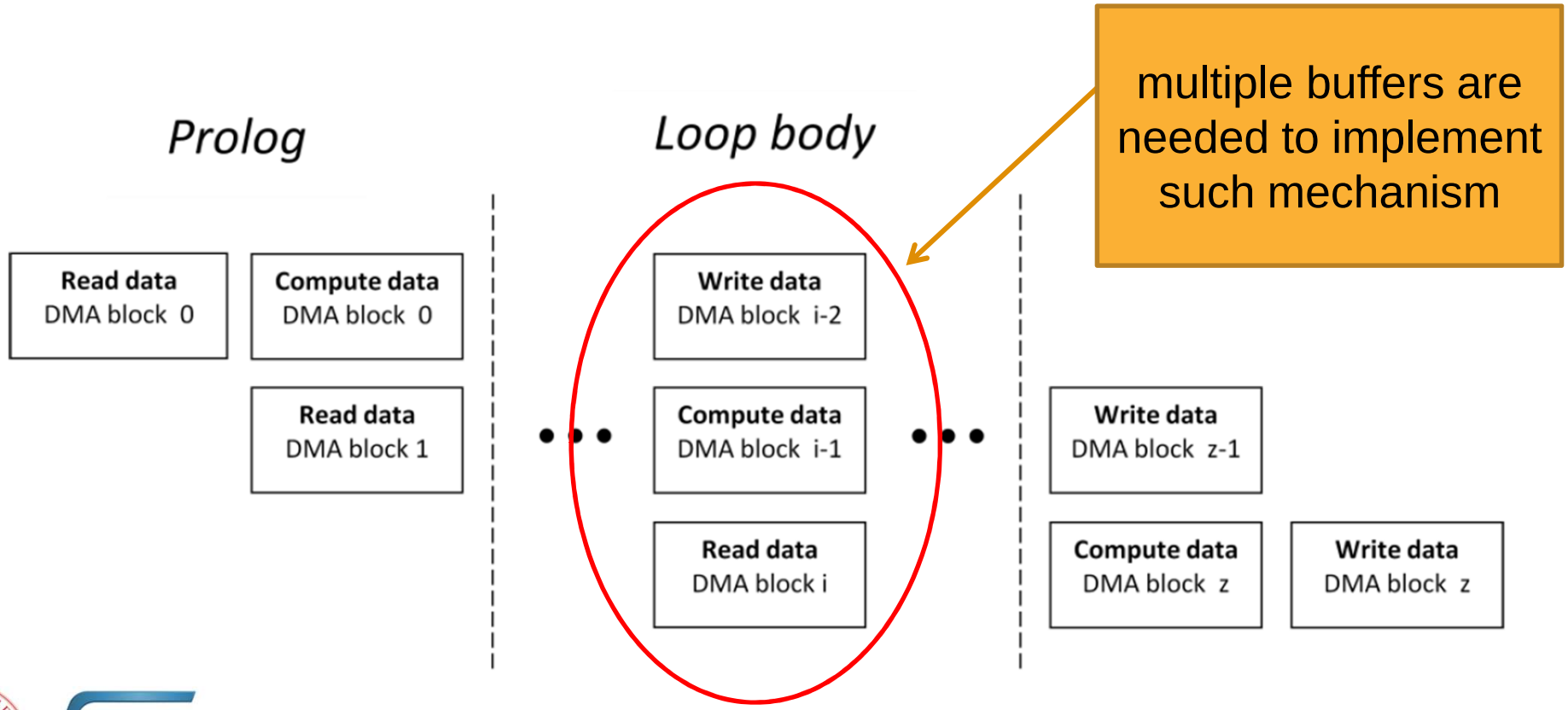
P2012 cores are not HW-multithreaded

- P2012 OCL runtime does not accept more than 16 work-items per work group when creating an ND-Range.
- But you can chose which work to do (e.g. using “case”) in each WI



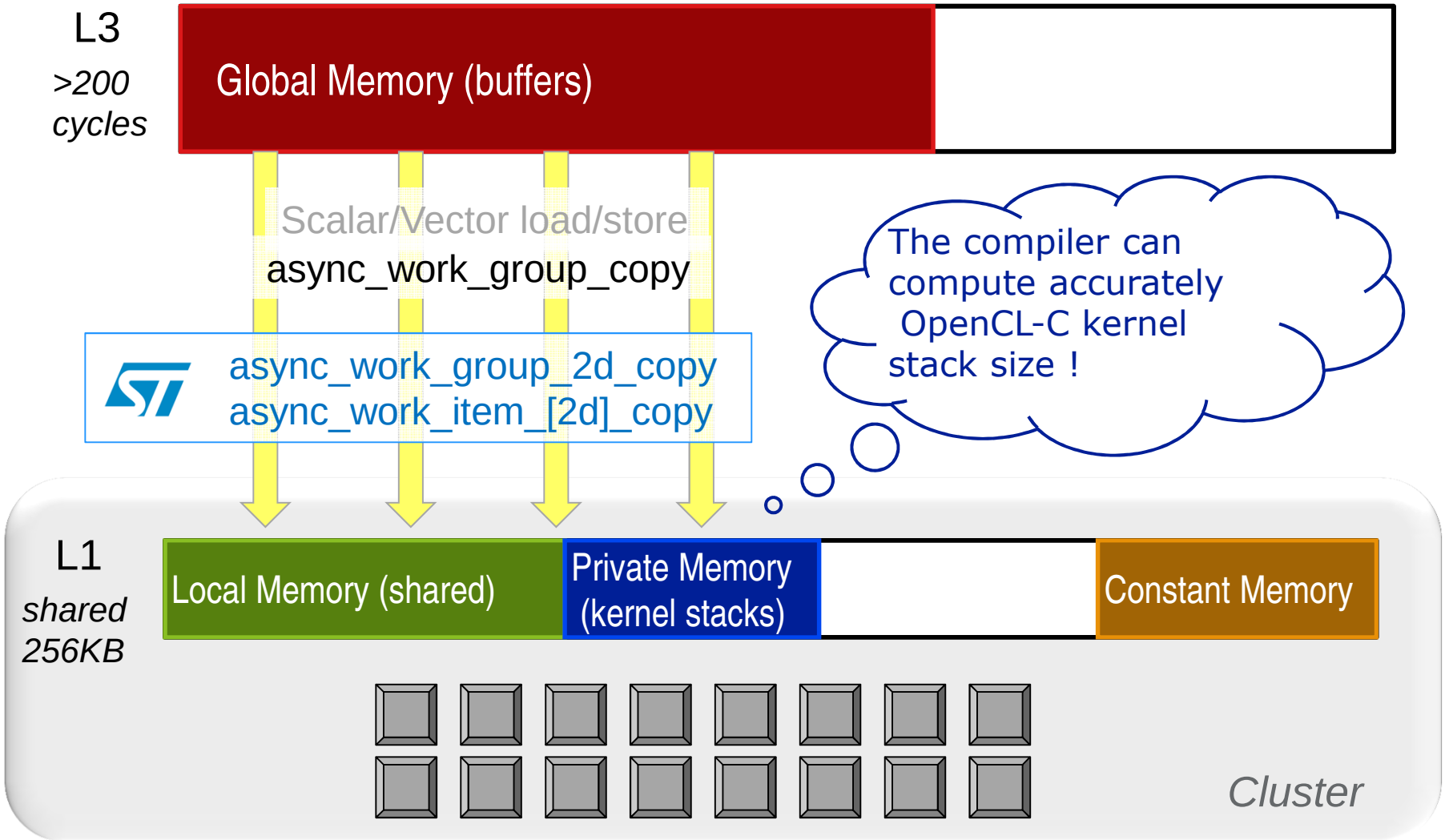
P2012 ≠ GPU: Managing vs. hiding Mem Latency

The best way to hide memory transfer latencies when programming for P2012 is to overlap computation with DMA transfers. This technique is based on software pipelining and double buffering



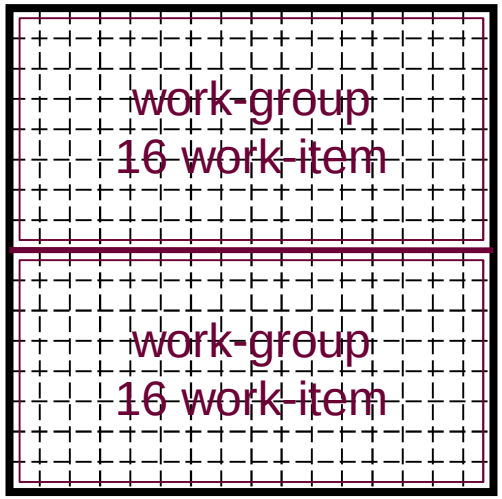
life.augmented

Memory Mapping and Data Movements

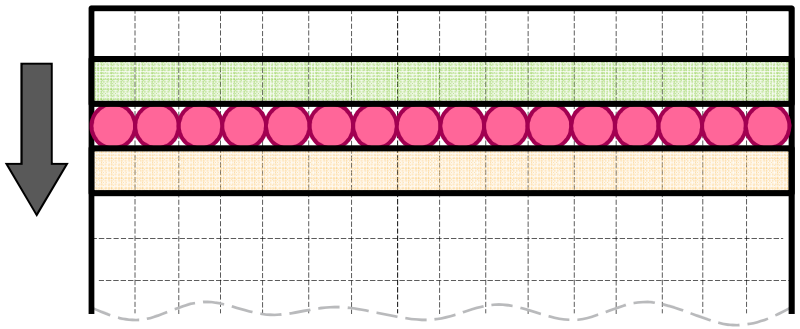


P2012 OpenCL Programming Style

Data



Double buffering
Software pipelining



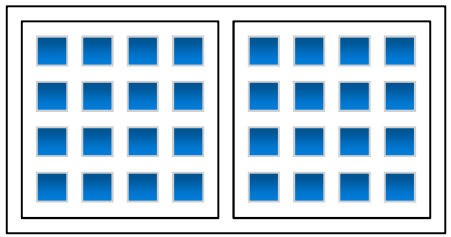
Write block N-1
`local` → `global`

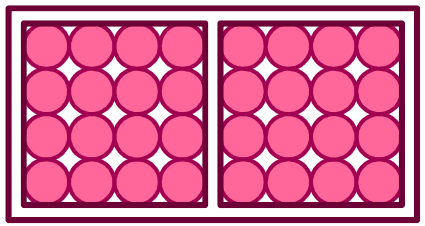
Process block N
`local` → `local`

Read block N+1
`global` → `local`

P2012
2 clusters
16 PE per clusters

ND-Range
2 work-groups
16 work-items per WG





P2012 OpenCL Programming Challenge

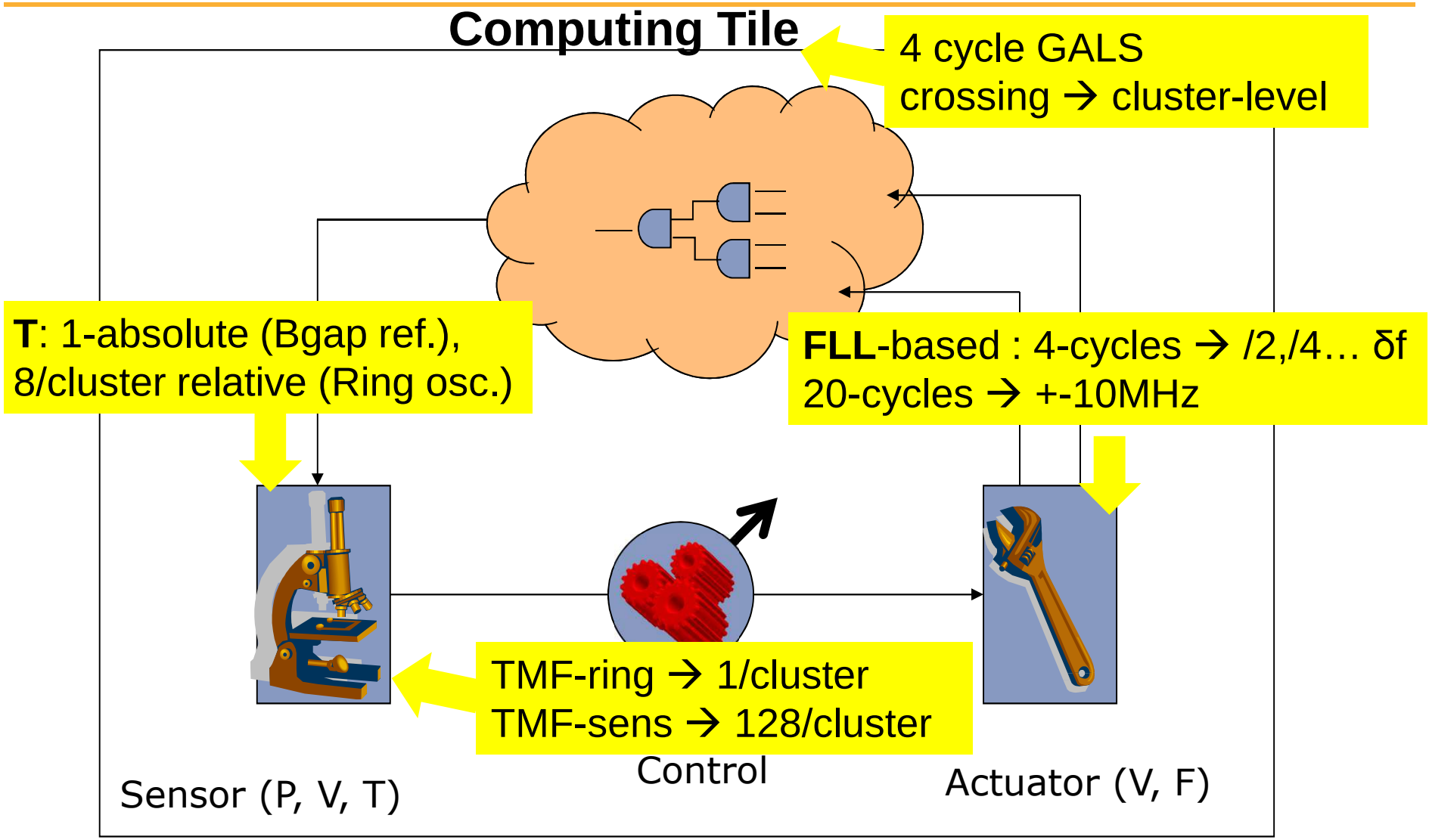
1. Fill processors/clusters with processing
   - Fully with *data parallelism* when available
   - *Task parallelism* otherwise (mid-coarse grain)
2. Optimize the data locality
   - Use `local` & `private` as *user managed cache*
   - Minimize `global` ↔ `local`/`private`
3. Parallelize memory transfers & computation
   - Use asynchronous copies (DMA)
   - Satisfy memory size constraints

Can we automate this?

Allocation and scheduling for P2012

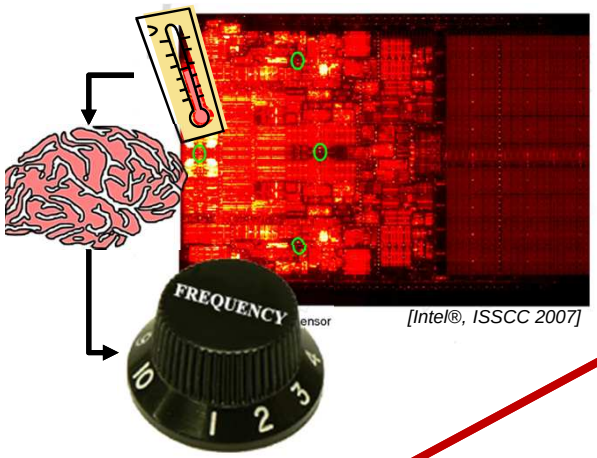
Managing Power & Temperature

Closed Loop Control



Thermal Controller

GOAL: track perf. Request @ safe T



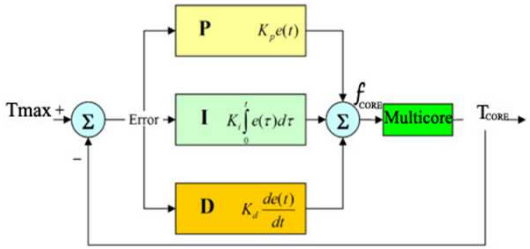
COMPLEXITY

Threshold based controller

- T > Tmax → low freq
- T < Tmin → high freq
- cannot prevent overshoot
- thermal cycle

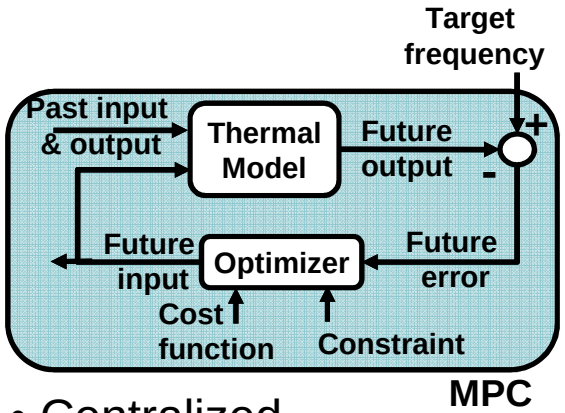
Classical feed-back controller

- PID controllers
- Better than threshold based approach
- Cannot prevent overshoot



Model Predictive Controller

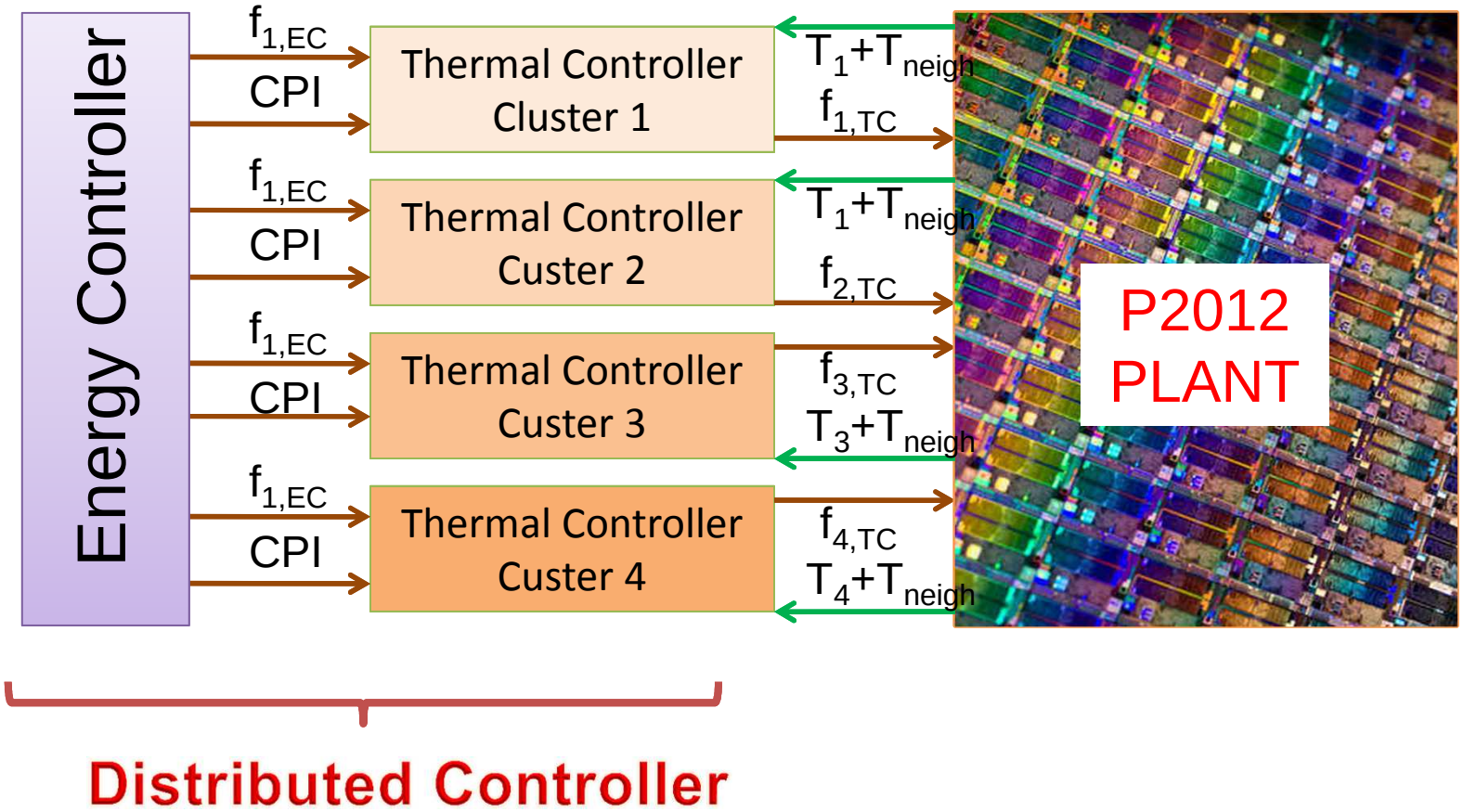
- Internal prediction: avoid overshoot
- Optimization: maximizes performance



- Centralized
  - aware of neighbor cores thermal influence
  - All at once – MIMO controller
- **Complexity !!!**

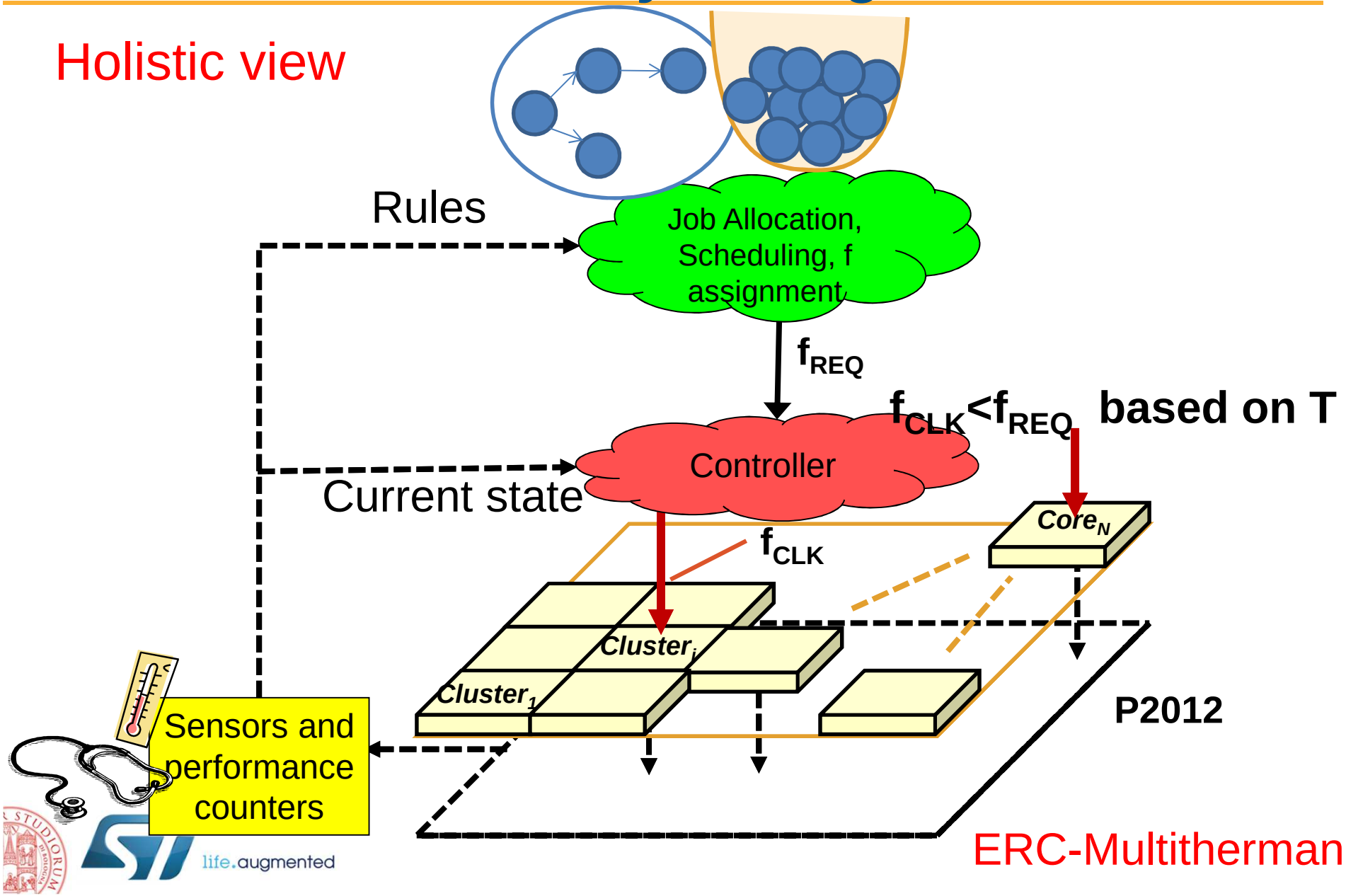
life.augmented

High Level Architecture



Energy Proportionality + Thermal & Variability management

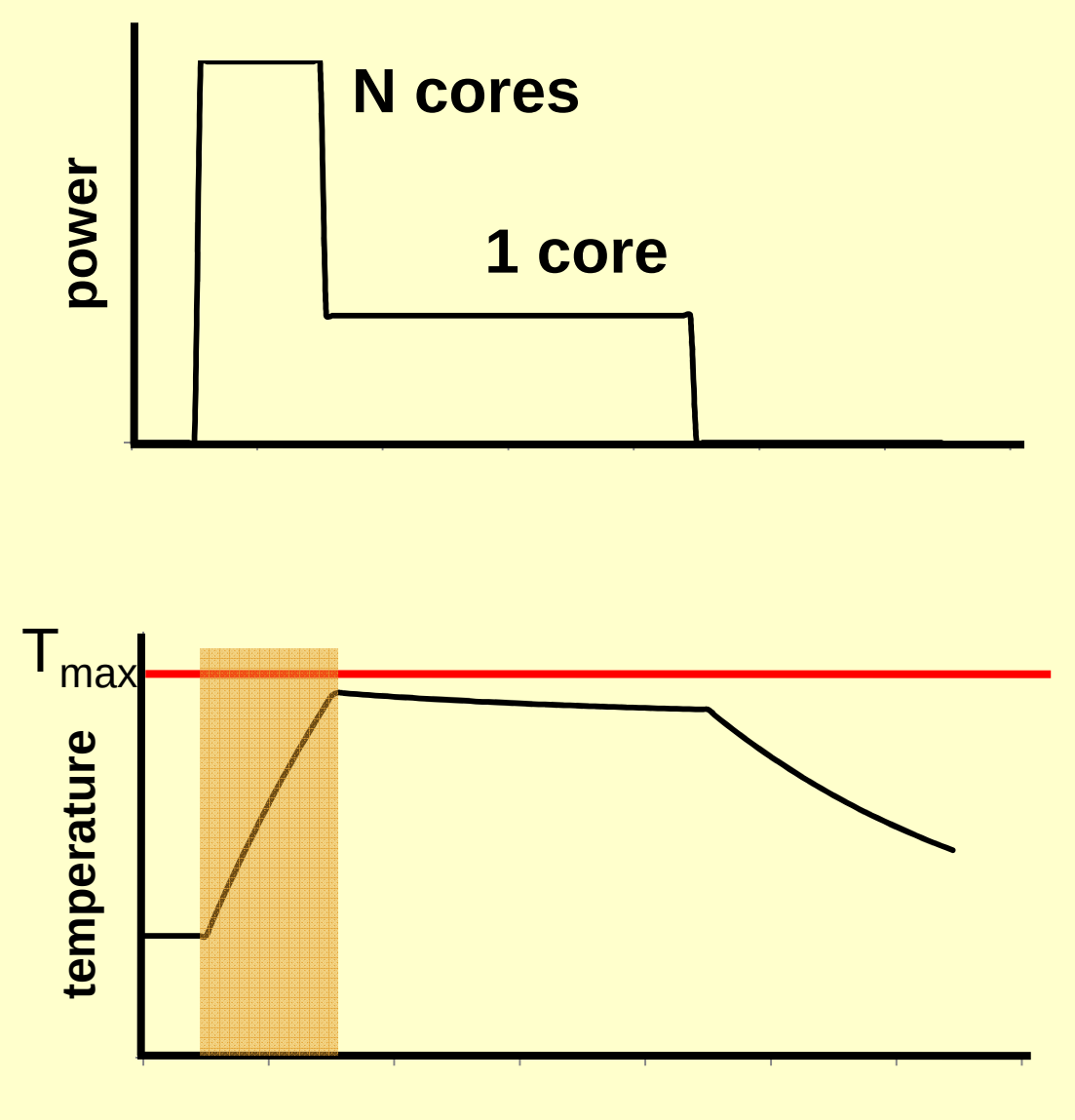
Holistic view



ERC-Multitherman

MPC: Preliminary Results

- Trace Driven Simulation (Matlab)
- Power Model: Nonlinear vs. linear
- Thermal Model: first vs. second order
- Centralized vs. Distributed



MPC enables exploitation of thermal capacitance (and PCM enhancers) for computational sprinting

Thank you!

life.augmented