Emerging Methods of Computing

Luca Benini	ETHZ
Wayne Burleson	University of Massachusetts
Fabien Clermidy	CEA-LETI
Enrico Macii	Politecnico di Torino
Angel Rodriguez-Vazquez	University of Sevilla

Chair: Yusuf Leblebici, EPFL

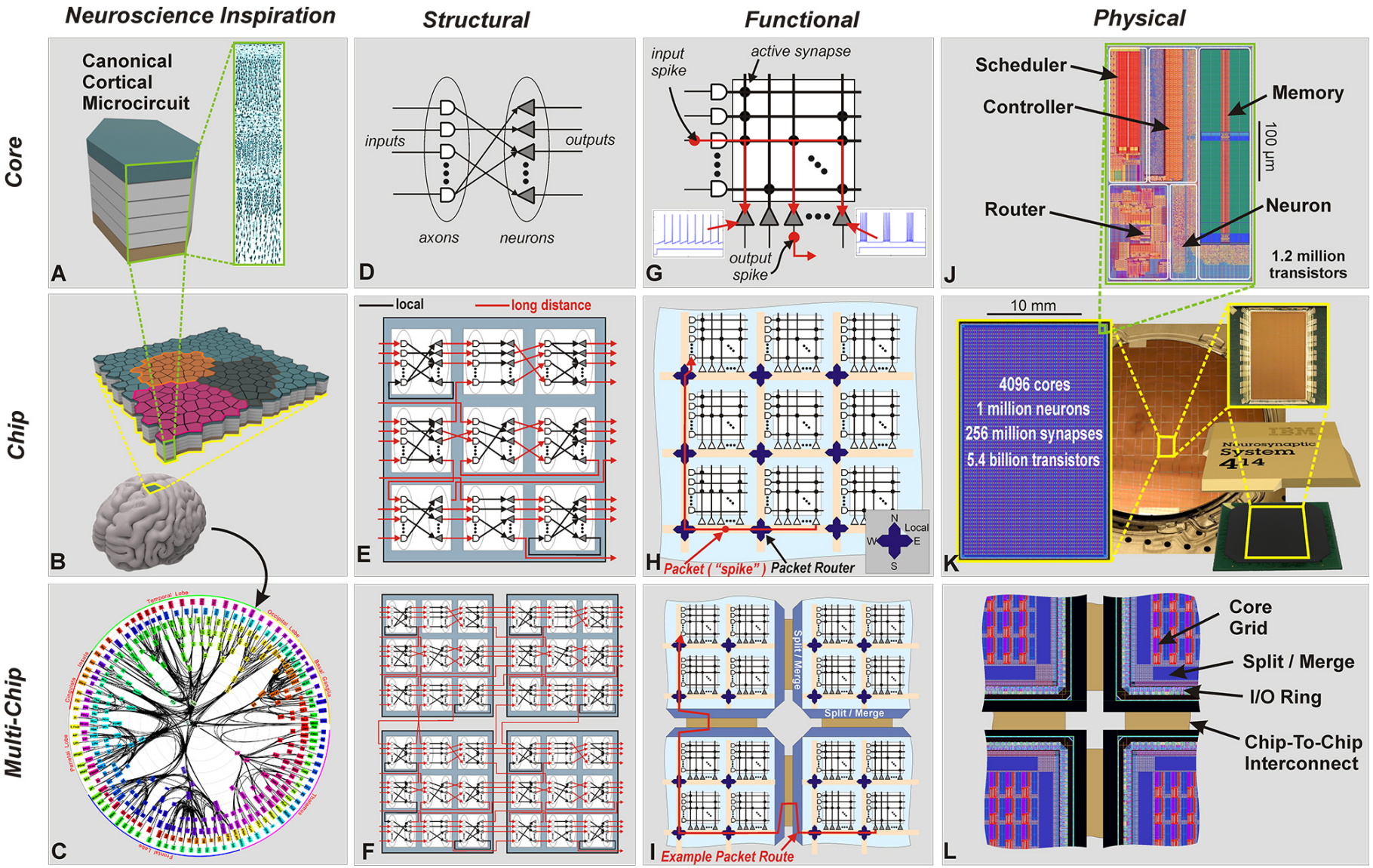
Centre SI
Integrated Systems Centre

EPFL
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

nano-tera.ch

Emerging Methods of Computing

?



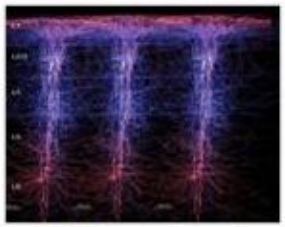
A million spiking-neuron integrated circuit with a scalable communication network and interface

Science 8 August 2014: Vol. 345 no. 6197 pp. 668-673

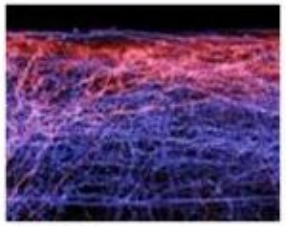
From brains to machines

~10^{10} synapses/cm^2

~10^{6} neurons/cm^2



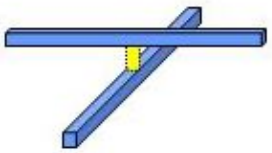
~10^{4} neurons/ cortical column



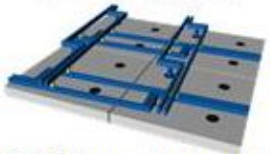
~5X10^{8} long range axons @ 1 Hz

Biological Brain

Synapse	Neurons	Microcircuit	Long-range interconnects	Brain
CROSSBAR JUNCTION	CMOS SUBSTRATE	LAMINAR CIRCUIT	HIGH SPEED BUS	



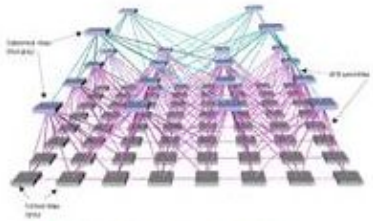
~10^{10} intersections/ cm^2 @ 100 nm pitch



5X10^{8} transistors/ cm^2 @ 500 transistors/ neuron

Layered cortical circuits with ~10^{6} neurons/cm^2

Multi-Gbit/sec digital comms



Electronic Brain

Source: DARPA Synapse project

Architectures

Parallelism

Memory Bandwidth

3D Integration

Heat Management

...

Luca Benini

ETHZ

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Ultra-low power computational sensing for next generation "Internet of Everything" platforms

Luca Benini

IIS-ETHZ & DEI-UNIBO

erc
European Research Council

SEVENTH FRAMEWORK PROGRAMME

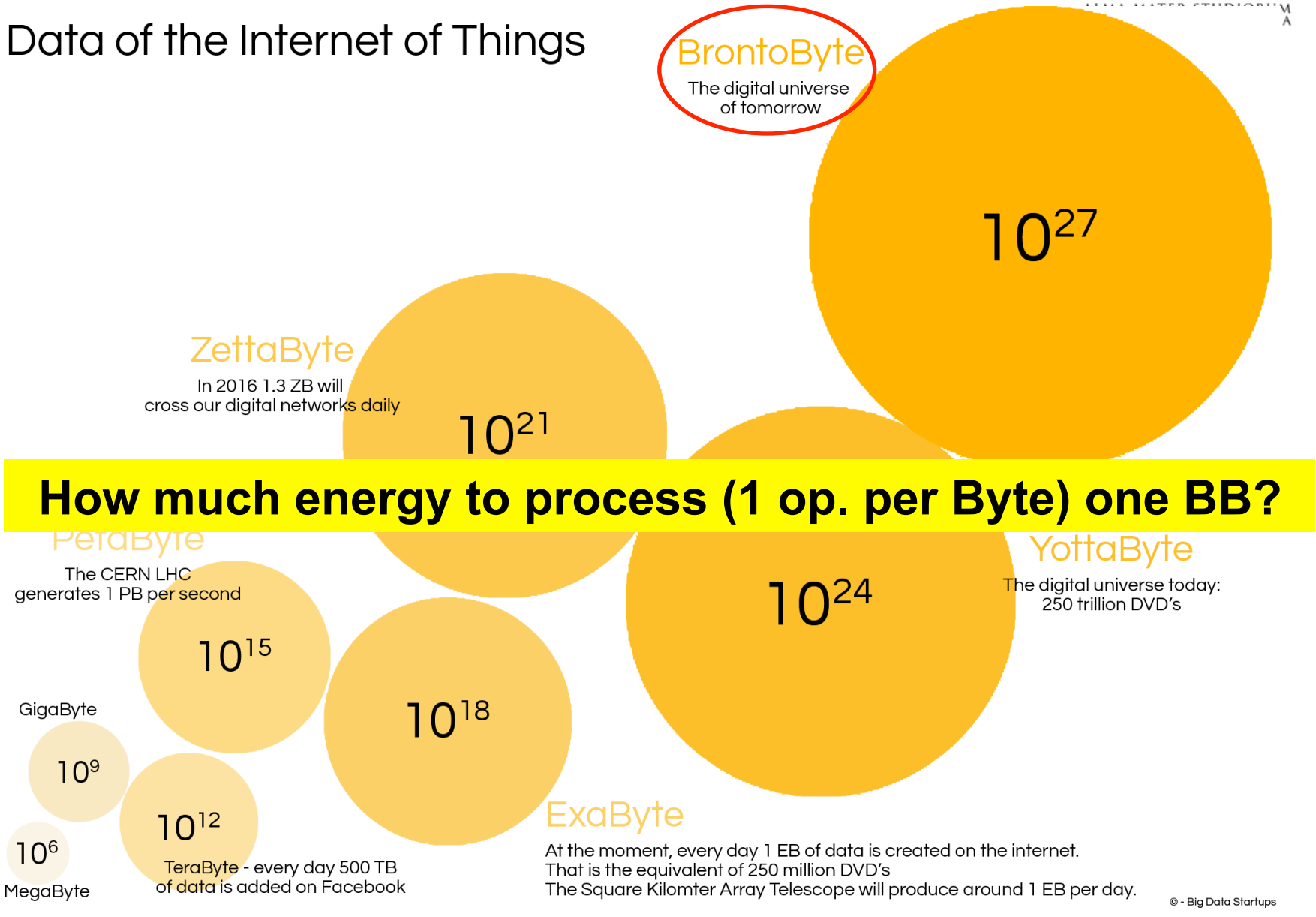
nano-tera.ch

NXP

life.augmented

SAMSUNG

Computing for the IoT

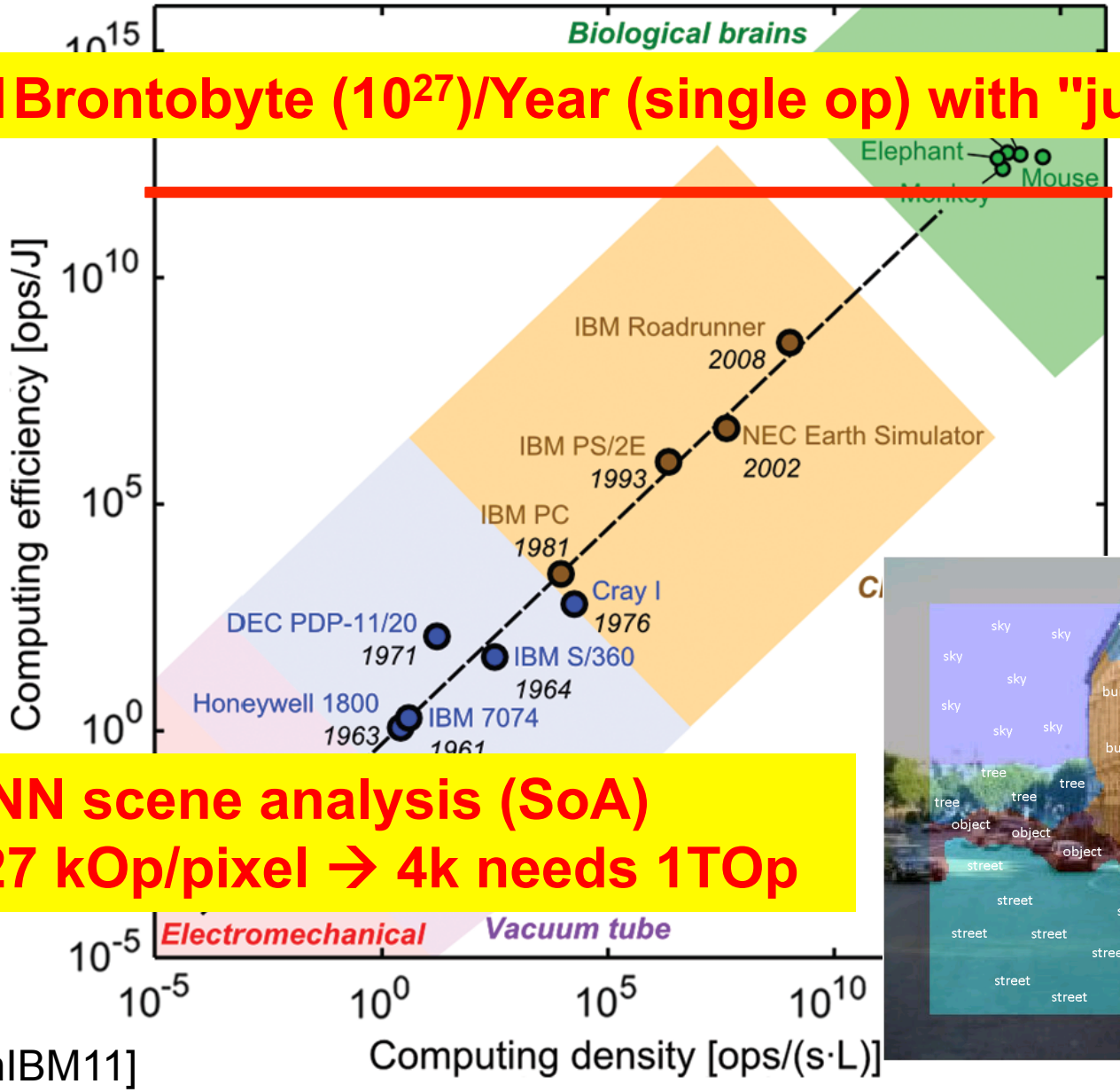


ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

How efficient?

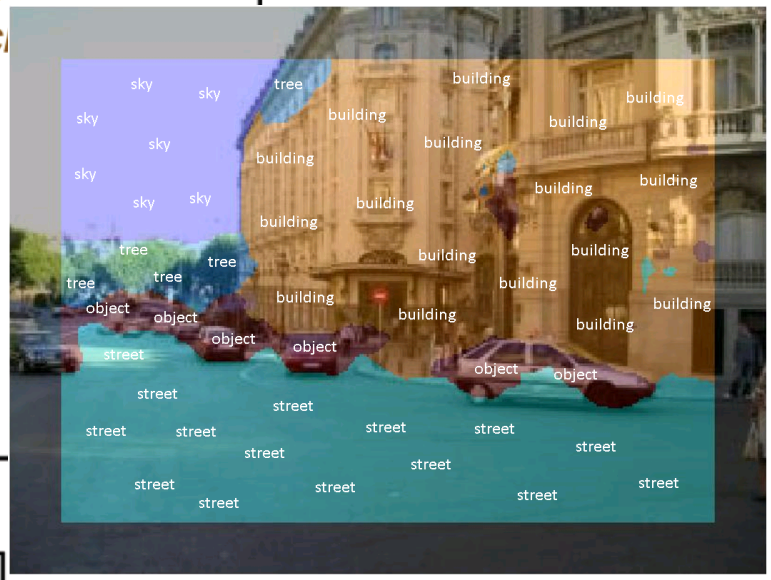
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

1Brontobyte (10^{27})/Year (single op) with "just" ≈ 10MW!



10^{12}ops/J
↓
1pJ/op
↓
1GOPS/mW

CNN scene analysis (SoA)
127 kOp/pixel → 4k needs 1TOp



[RuchIBM11]

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

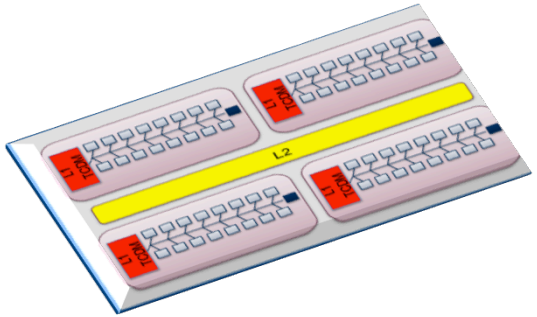
pJ/op Parallel ULP Computing

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

PULP

pJ/op is traditionally the target of ASIC + uCntr

- Scalable: many-core + heterogeneity
- Programmable: OpenMP, OpenCL, OpenVX
- Open: **Software** & **HW** OpenCores www.opencores.org
- **Best-in-class LP silicon technology (partner foundries!)**



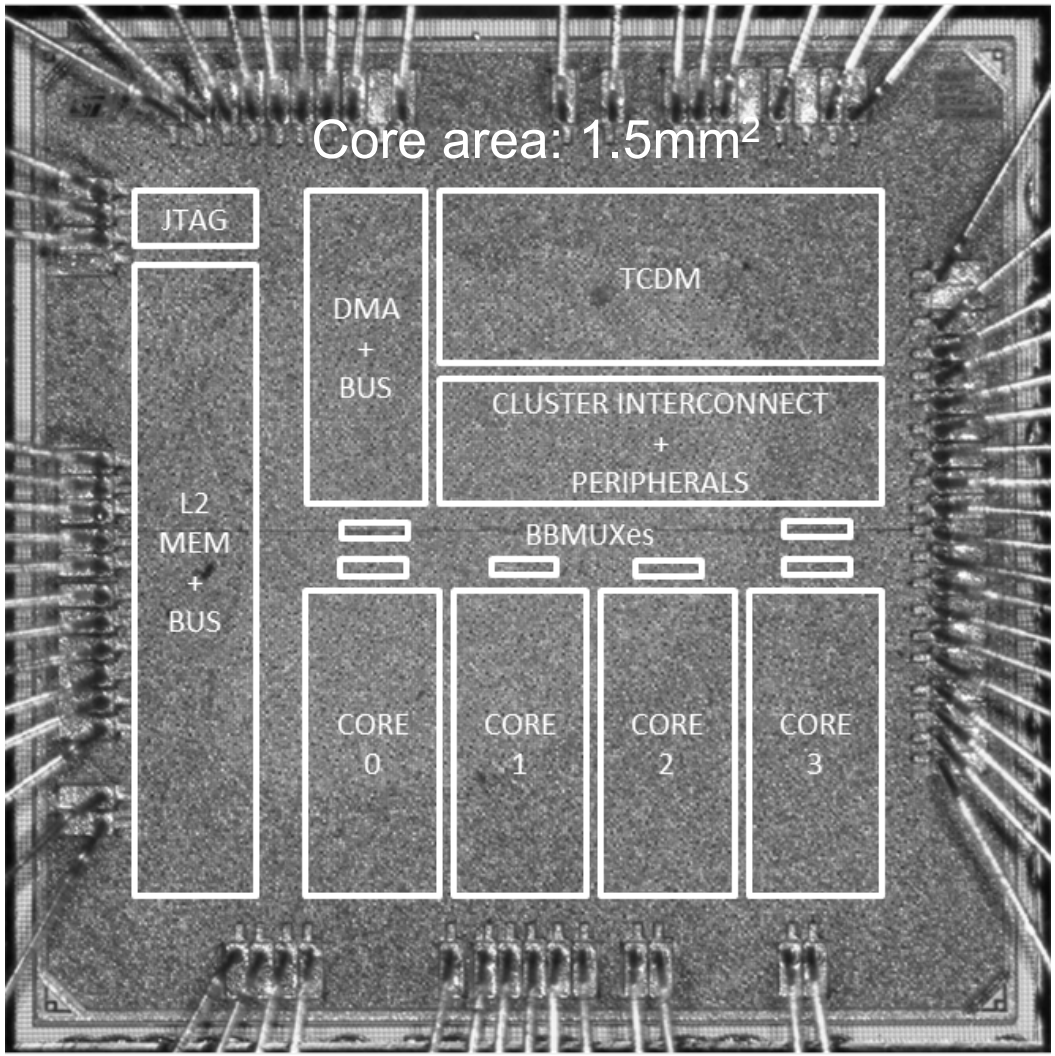
4 Strategic areas

1. **Near-threshold operation**
2. **Instant-on-circuits**
3. **Heterogeneous Architecture**
4. **Volume (3D) integration**

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

PULP v1 – First Demonstrator

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA



First Multi-core chip in 28nm RVT FDSOI (STM free silicon program)

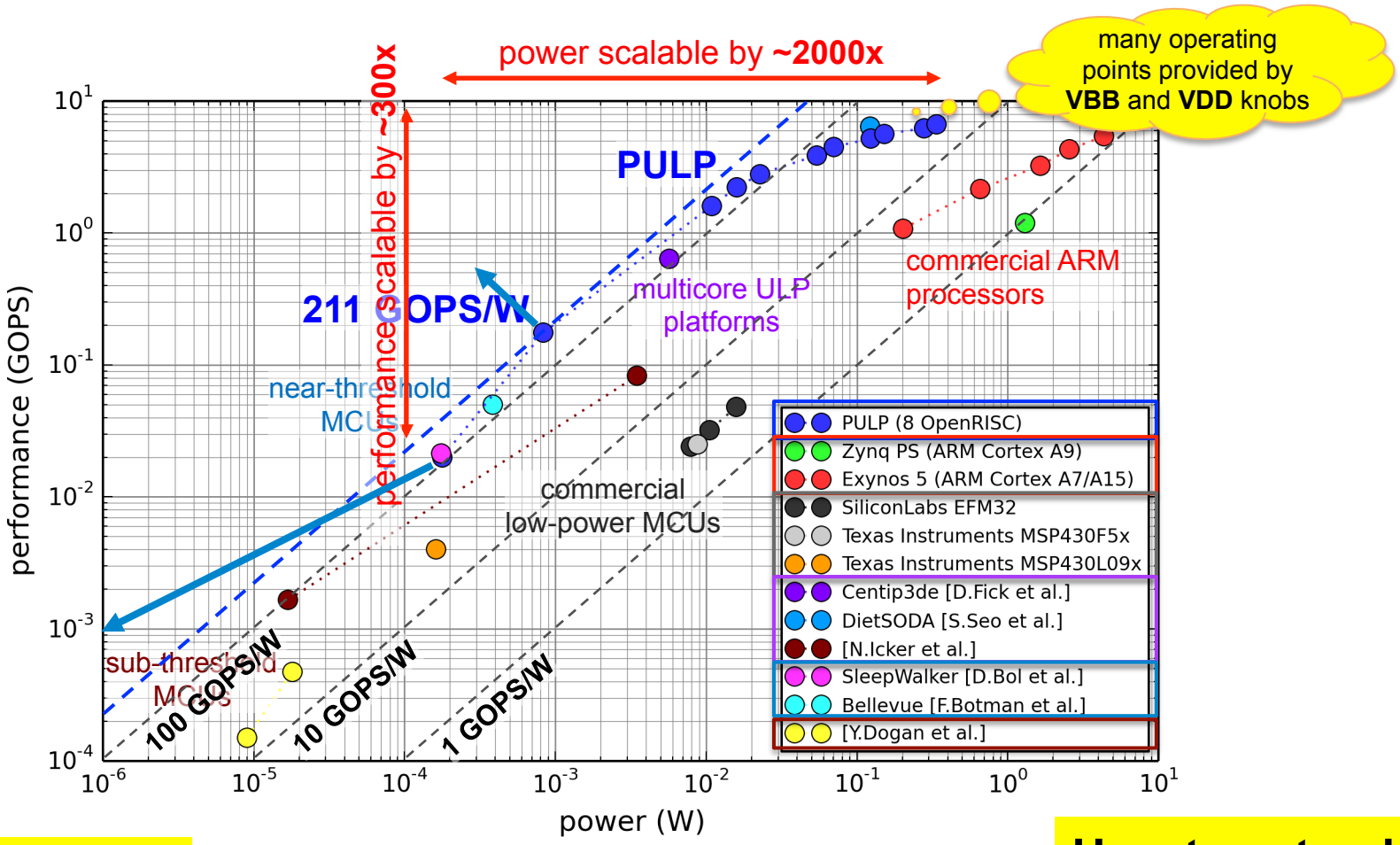
- 4 Open RISC Cores
- 6 Dynamically controllable Body Bias islands.
- Both Reverse and Forward Body Bias
- Functional 0.45V – 1.20V
- Tested and Characterized (July-August 14)

First-time Good silicon

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

PulpV2 - Getting closer

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA



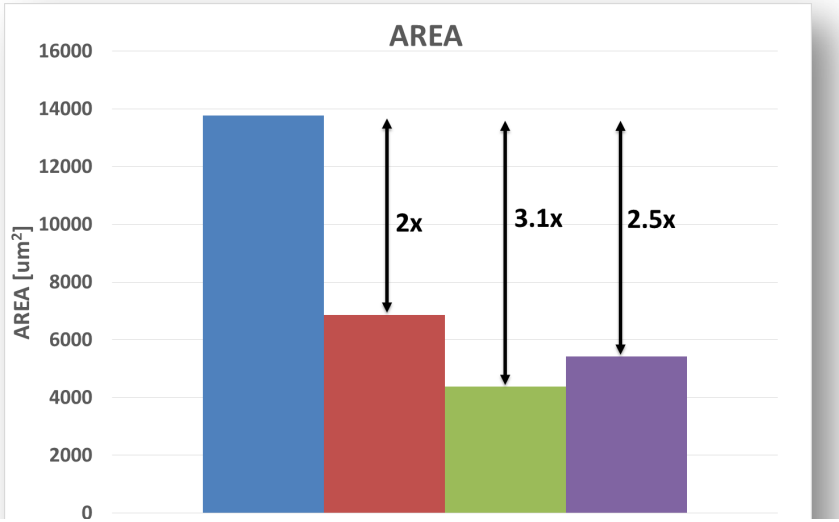
How to close the x5 gap?

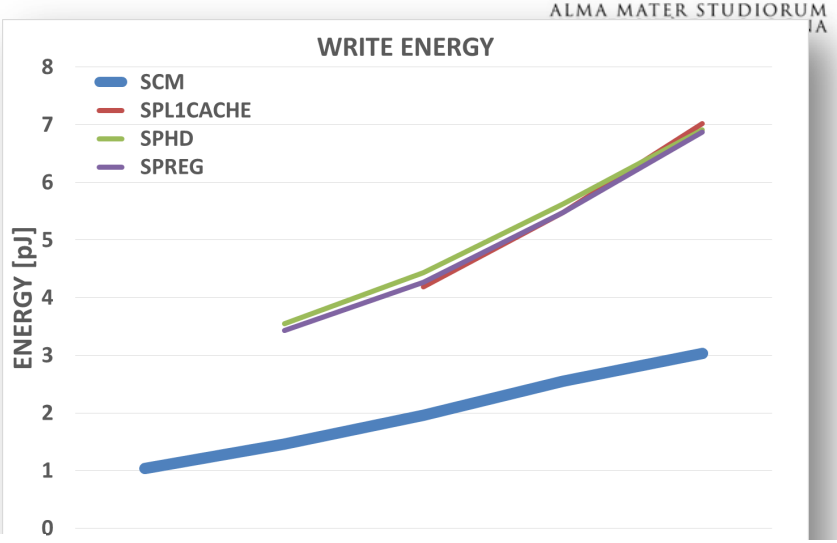
How to extend the range to sub MOPS?

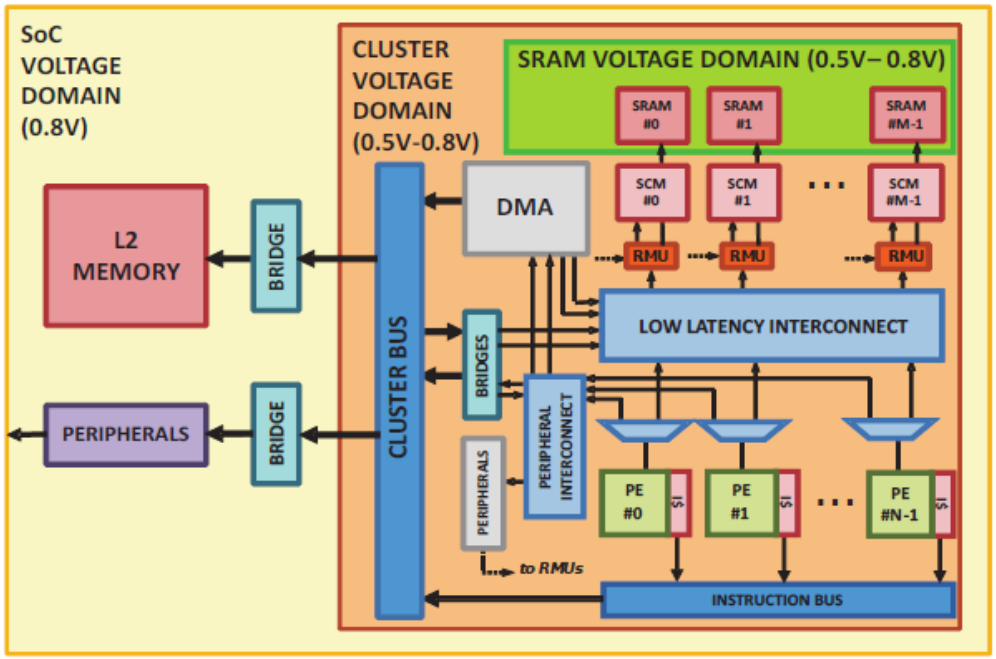
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

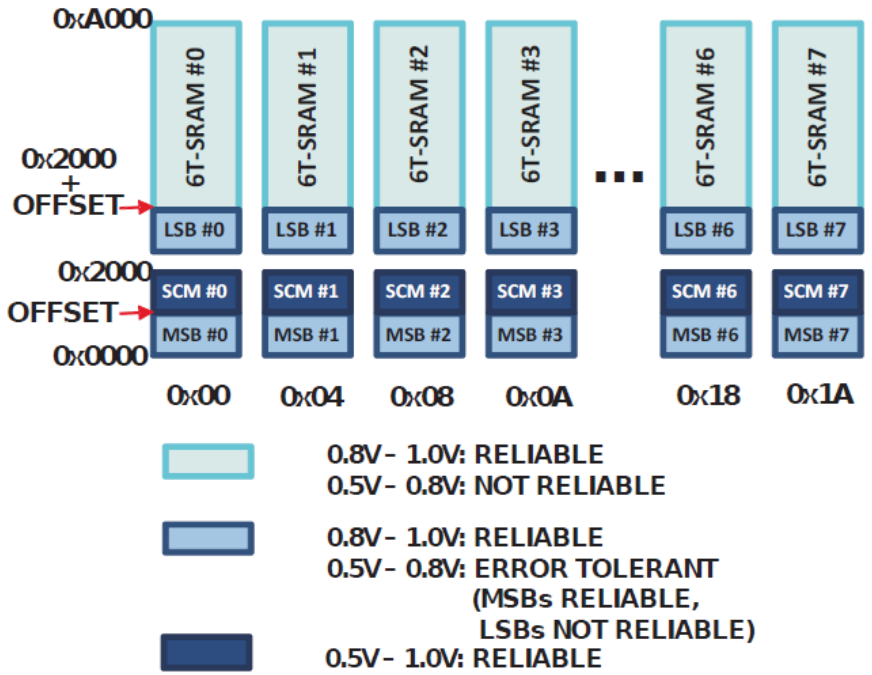
Hybrid + Approximate Memory

ALMA MATER STUDIORUM

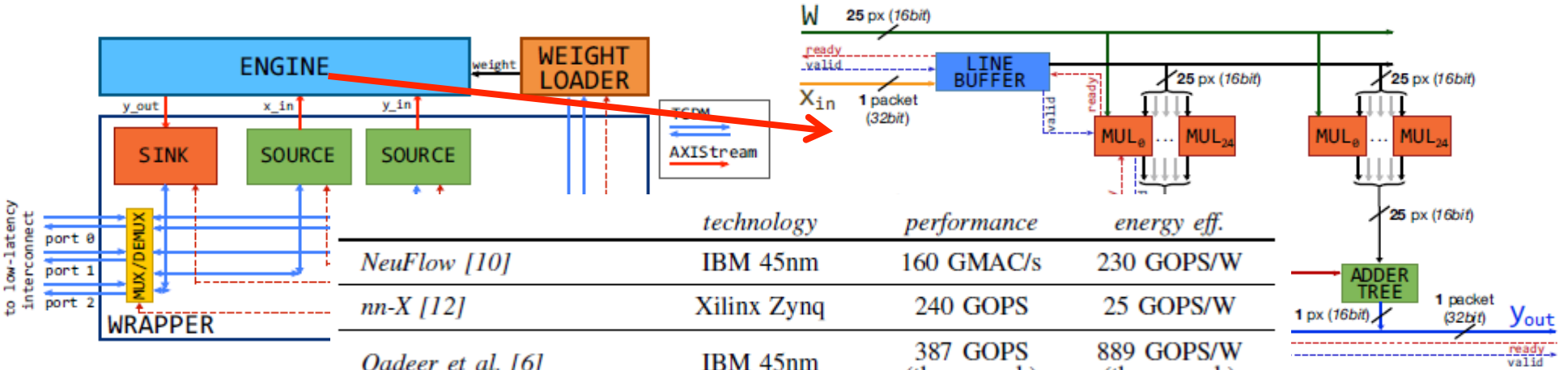




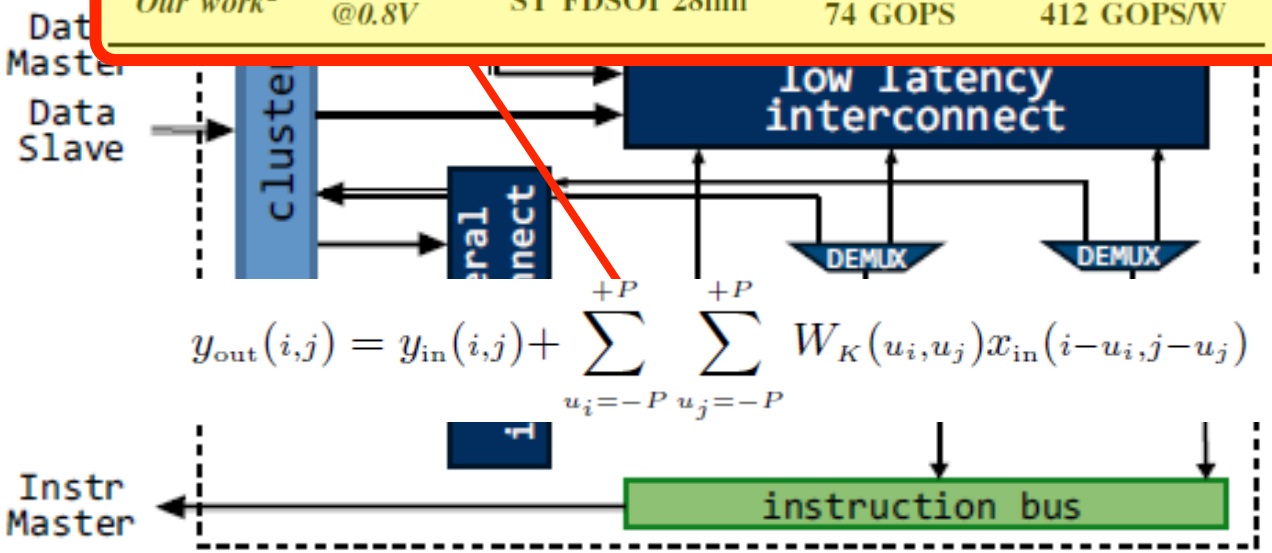




Neural Processing Accelerator



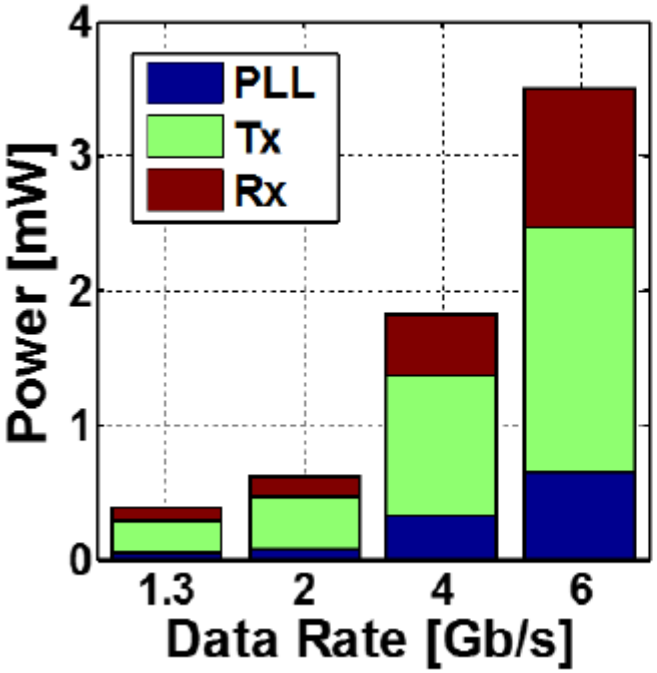
	technology	performance	energy eff.
NeuFlow [10]	IBM 45nm	160 GMAC/s	230 GOPS/W
nn-X [12]	Xilinx Zynq	240 GOPS	25 GOPS/W
Qadeer et al. [6]	IBM 45nm	387 GOPS (theor. peak)	889 GOPS/W (theor. peak)
Camunas-Mesa et al.[7]	350nm	16.6 Mevt/s	185 Mevt/s/W
IBM TrueNorth [26]	Samsung 28nm		46 GSOPS/W
Our work[2]	*@0.4V* *@0.8V*	ST FDSOI 28nm	2.78 GOPS 74 GOPS	2750 GOPS/W 412 GOPS/W

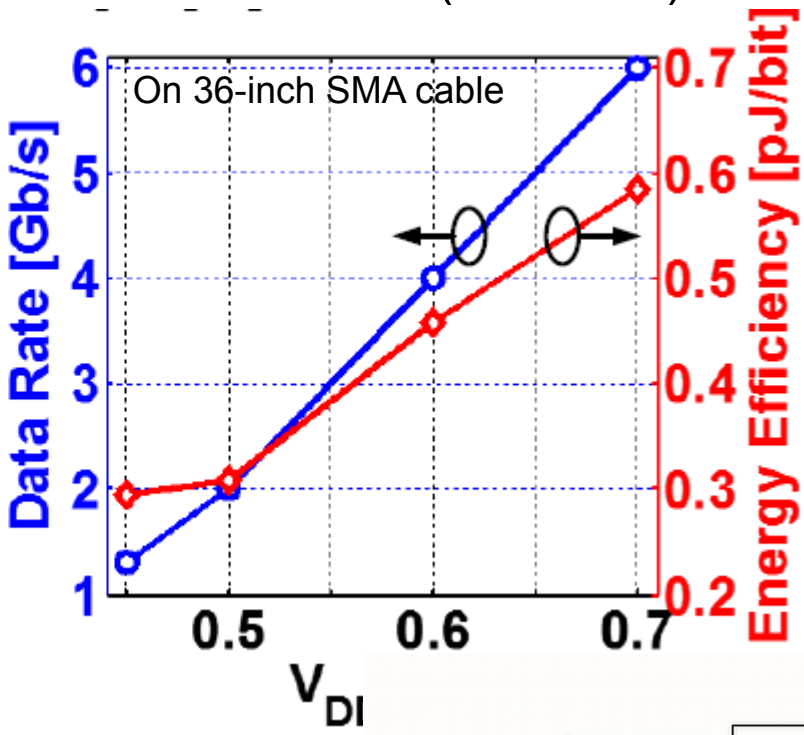


$$y_{\text{out}}(i,j) = y_{\text{in}}(i,j) + \sum_{u_i=-P}^{+P} \sum_{u_j=-P}^{+P} W_K(u_i,u_j)\, x_{\text{in}}(i-u_i, j-u_j)$$

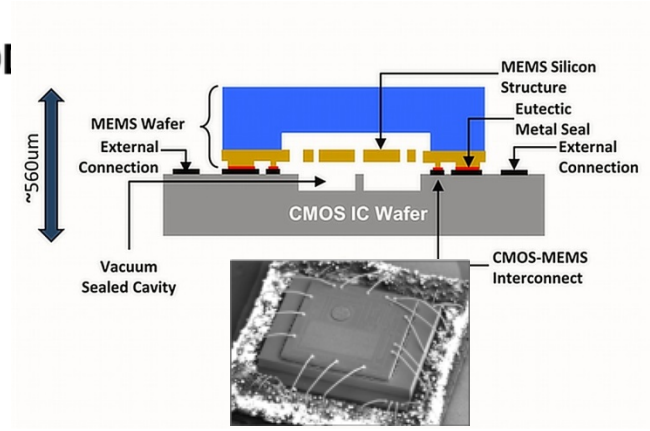
ULP Phy for 3D Integration

- 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)





- 1GBps @ 2mW – one IO word every cycle at 250MHz – SIP+die stacking option for large+low cost **memories + sensors** becomes viable



ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Thank you!

erc
European Research Council

Multithermand AdG
Multiscale Thermal Management of Computing Systems

Wayne Burleson

University of Massachusetts

From Nano to Exa:
Re-thinking
Data Representations and Data Movement

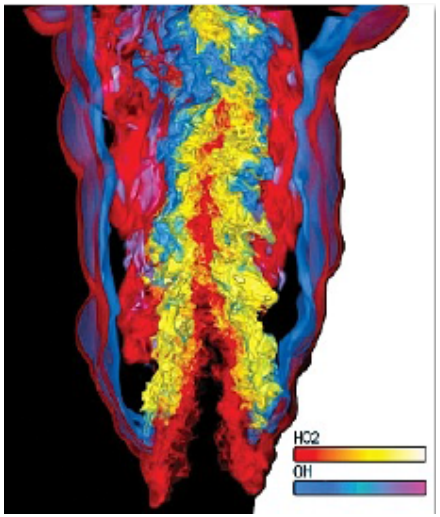


Image courtesy US Dept of Energy

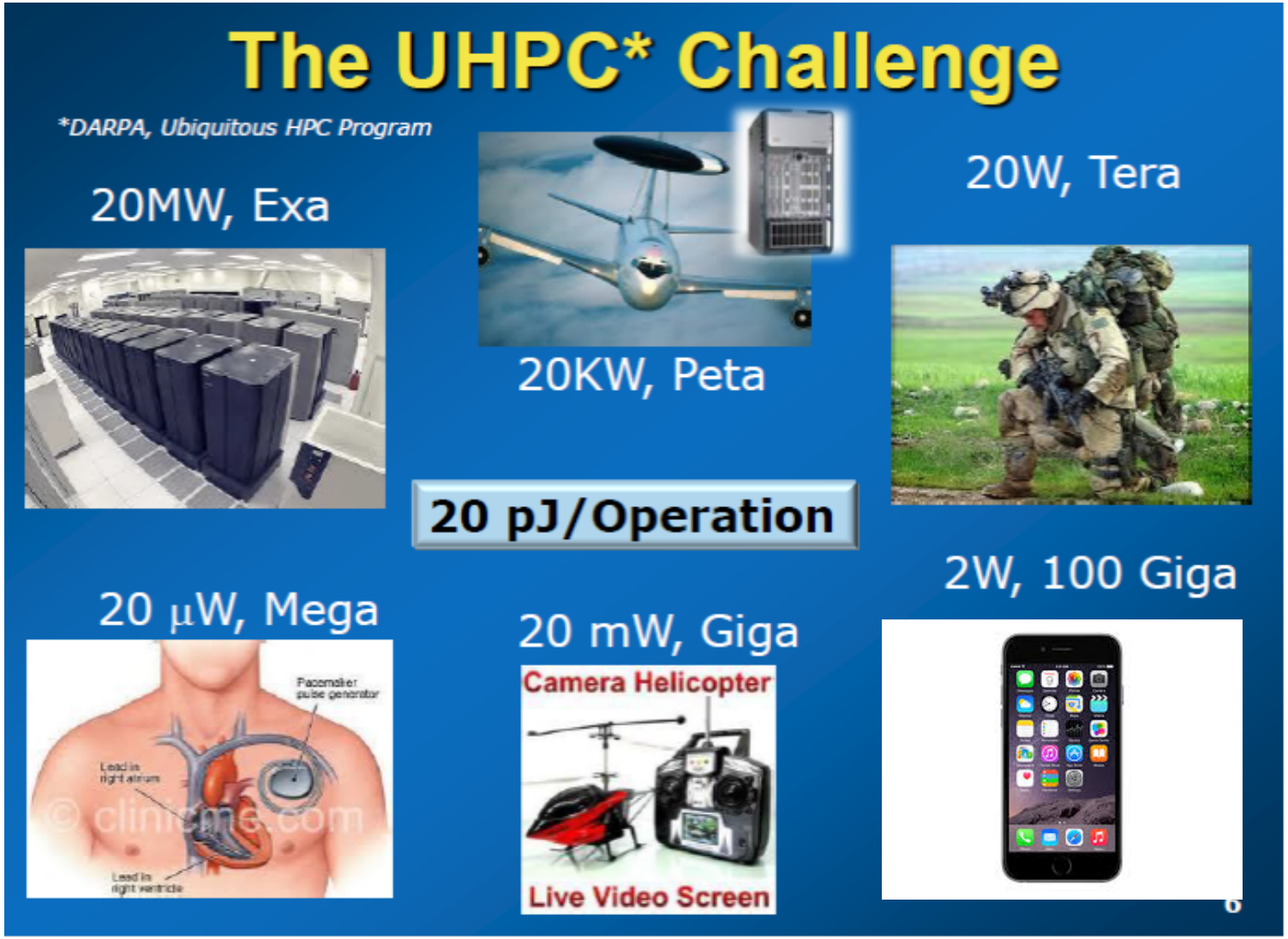
Wayne Burleson

U. Massachusetts, Amherst

AMD Research, Boston

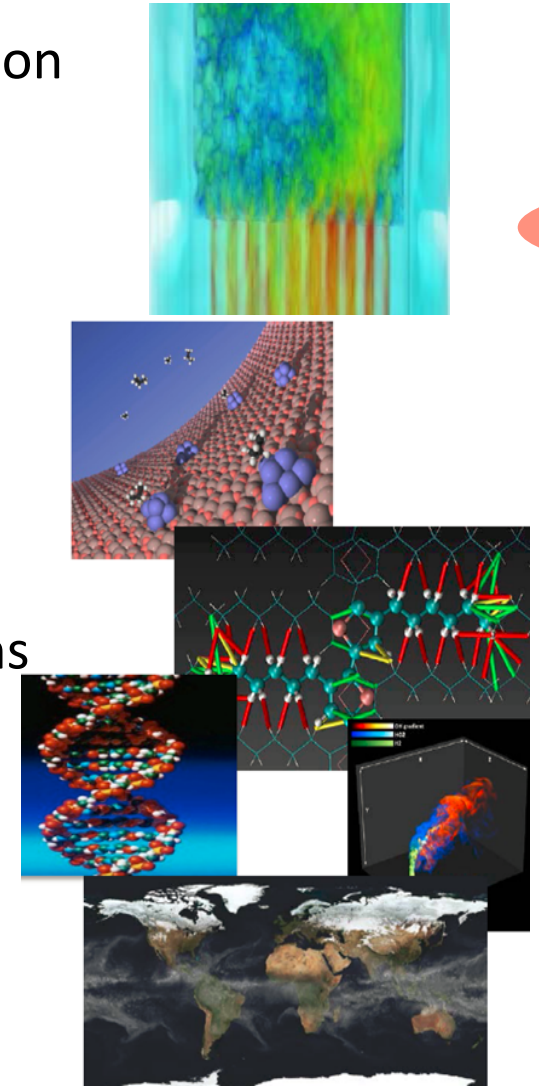
Image courtesy nlp.stanford.edu

Energy-efficient computing across scales, DARPA



Exascale High Performance Computing

- Scientific computation
  - Weather
  - Combustion
  - Materials
  - Energy
  - Genetics
- Evolving codebase
  - Big-data
  - Big-compute
- Future computations
  - Graphs
  - Multimedia
  - Data Analytics



The Top Ten Exascale Challenges, with Technical Approaches

1. Energy efficiency
2. Interconnect technology
3. Memory Technology
4. Scalable System Software
5. Programming systems
6. Data management
7. Exascale Algorithms
8. Algorithms for discovery, design, and decision
9. Resilience and correctness
10. Scientific productivity

Images courtesy US Dept of Energy, Advanced Scientific Computing

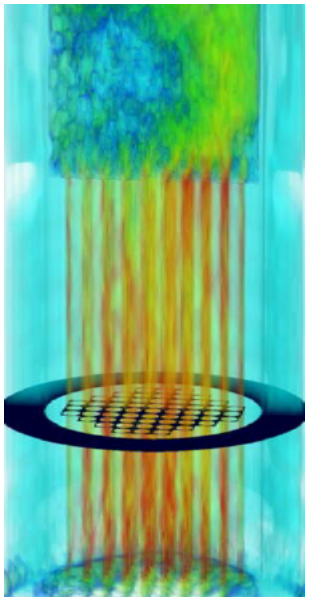
Public-private partnerships for Exascale Research (2014)

- **AMD** will conduct research for an integrated exascale node architecture. Particular areas of emphasis include near-threshold-voltage logic and other low-power computing technologies. AMD will investigate a new standardized memory interface
- **Cray Inc.** will explore alternative processor design points, including ARM microprocessor designs.
- **Intel** will use this award to continue to advance research in energy efficient node and system architectures, including software targeted at developing extreme-scale systems.
- **NVIDIA** will build on its work in FastForward 1, with a strong focus on energy efficiency, programmability, and resilience.
- **IBM** will investigate next-generation standardized memory interface.

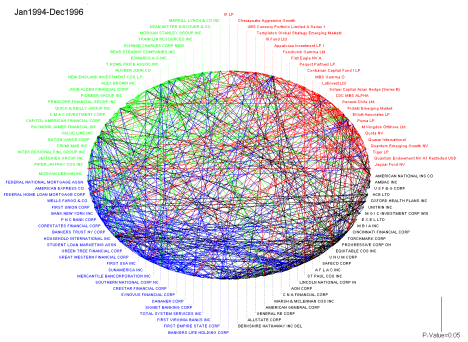
Data representations

- Modeling the **analog** world
  - Physics
  - Chemistry
  - Biology
- Modeling the **virtua**l world
  - Graphs, Relationships,
  - Social, Behavioral, Economic
  - Metrics, Costs, Risk
- Modeling the **observed** world
  - Media and Signal Processing
  - Neuromorphic
  - Approximate

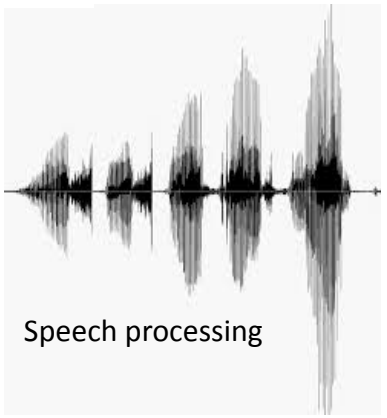
- Data representations
  - Data types, Floating point,… Graphics, Integer,
  - Standards: IEEE 754
  - Sampling: Adaptive Mesh, Compressive sensing
  - Arithmetic and Numerical effects
  - Software development and validation
  - Libraries: BLAS, LINPACK, MATLAB,



Combustion turbulence

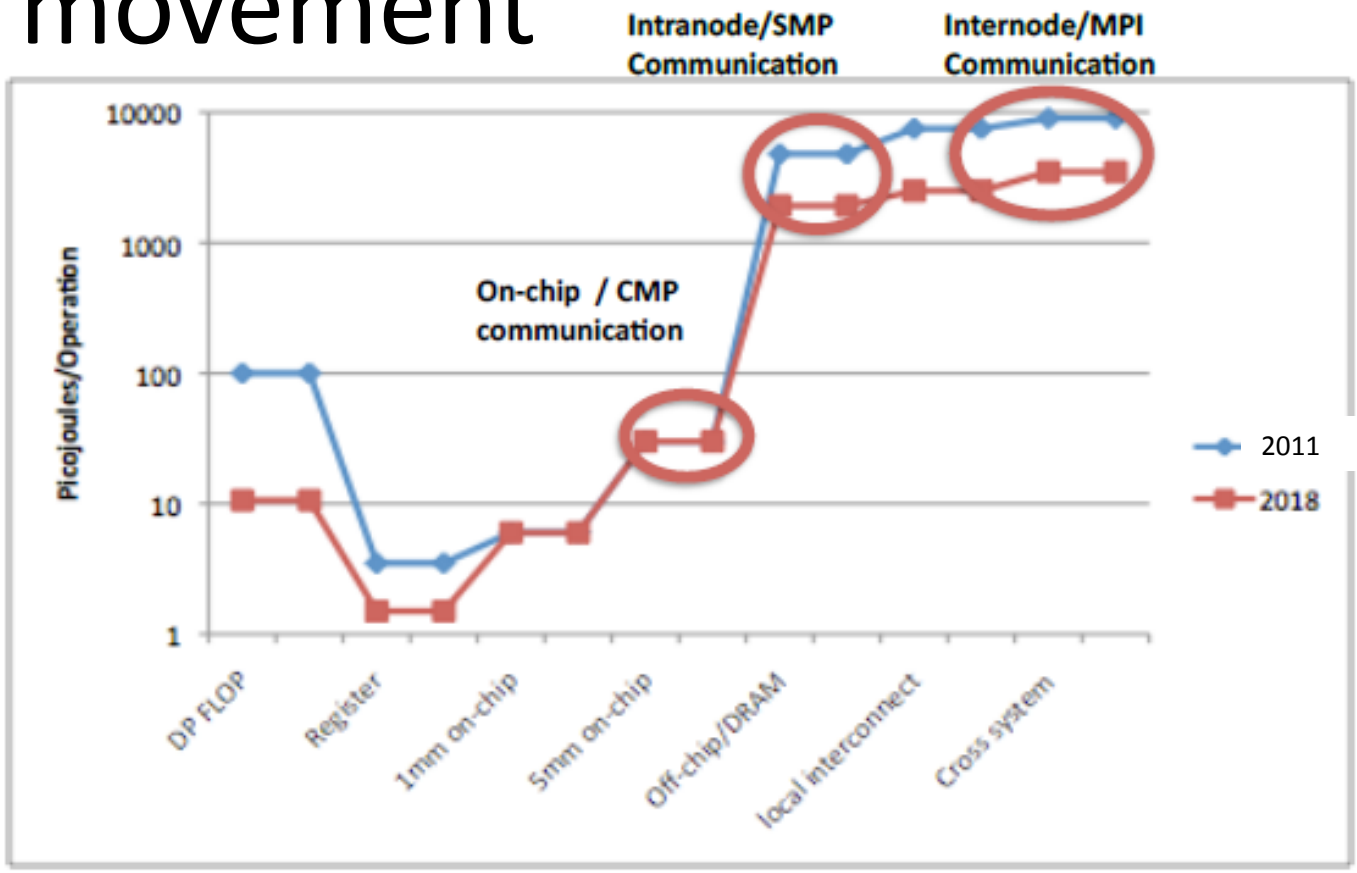


Financial Risks (2008)



Speech processing

Data movement



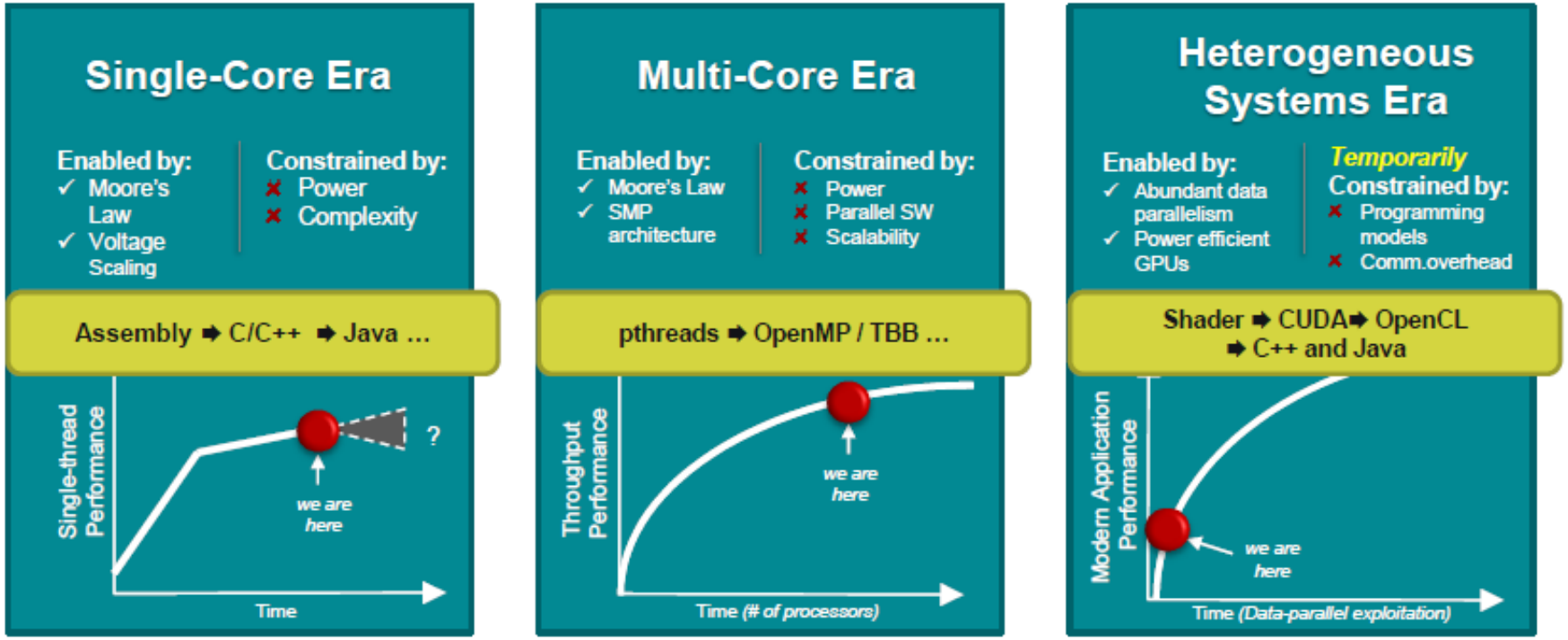
ASCR Exascale Programming Challenges Workshop , 2011

Figure 5: With new scaling rules and massive growth in parallelism, data locality is increasingly important. This diagram shows the cost of a double-precision multiply-add at different levels of the memory hierarchy (from the cost of performing the flop to the movement of data operands from registers, different distances on-chip, and distances off-chip.) *The model or FPU and register access is based on the Tensilica LX2 core energy model, the energy consumed for cross-chip wires uses the Orion2 power model which is based on Balfour's model [6], memory access is based on the JEDEC DDRx memory roadmap, and cross-system is based on projections for VCSEL-based optical transceivers.*

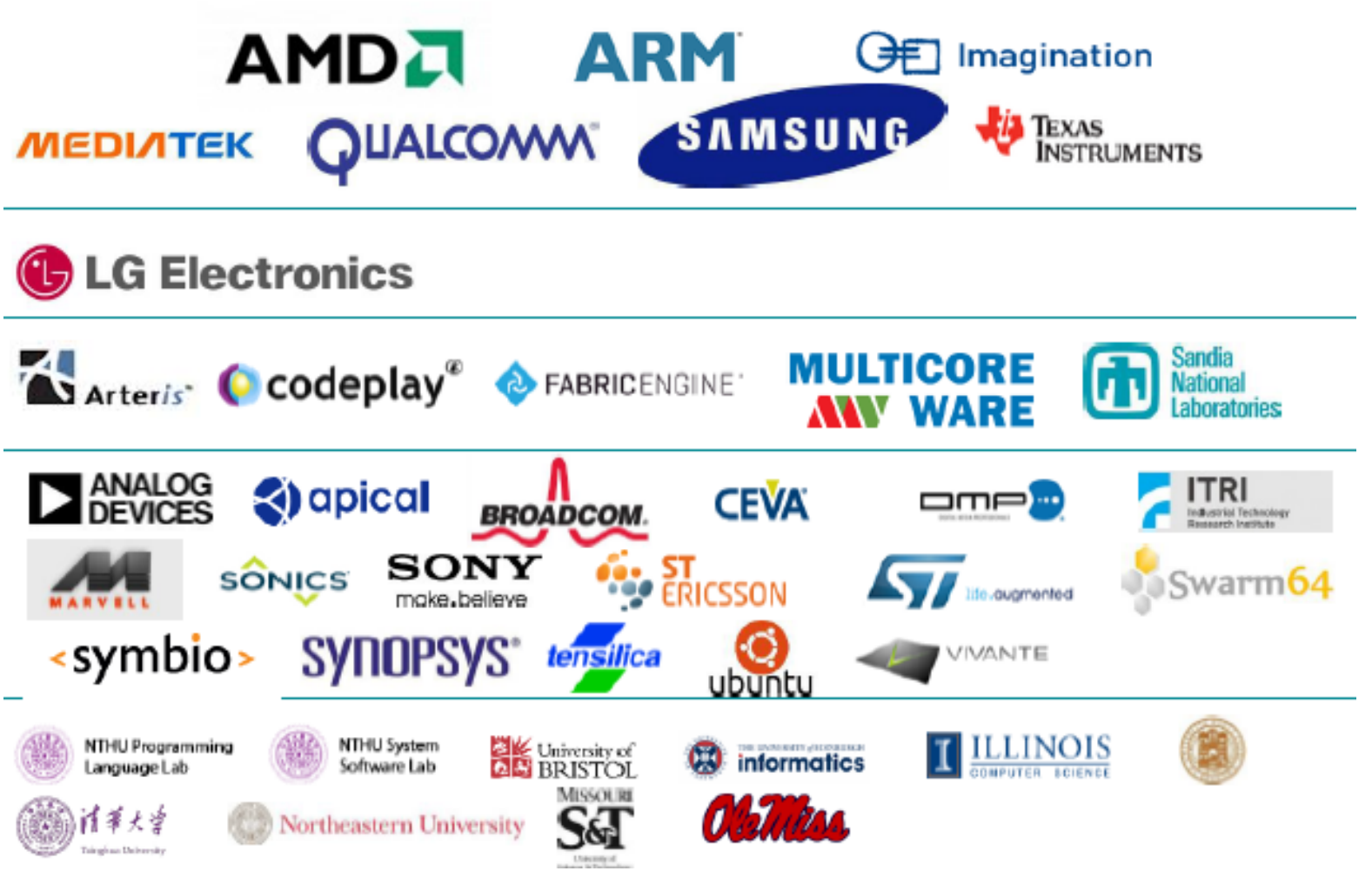
Data movement

- On-chip interconnect
  - Caches: L2 and L3 on-die, CPU-GPU shared memory, coherence,
  - Networks on Chip (circuits, architecture, protocols)
- Off-chip
  - Fast interfaces: electrical/optical
  - Non-volatile memory, Multi-level memory
  - Die-stacked memory, 2.5D, 3D
  - Compression, encryption, coding,…
- Memory allocation, management
  - Interprocessor communication - MPI
- Work-flow management, avoid disk/file system where possible…
  - eg HPC: combine simulation, analysis, visualization

Vision: Heterogeneous Systems Era



Heterogeneous System Architecture (HSA) **Open** Ecosystem!



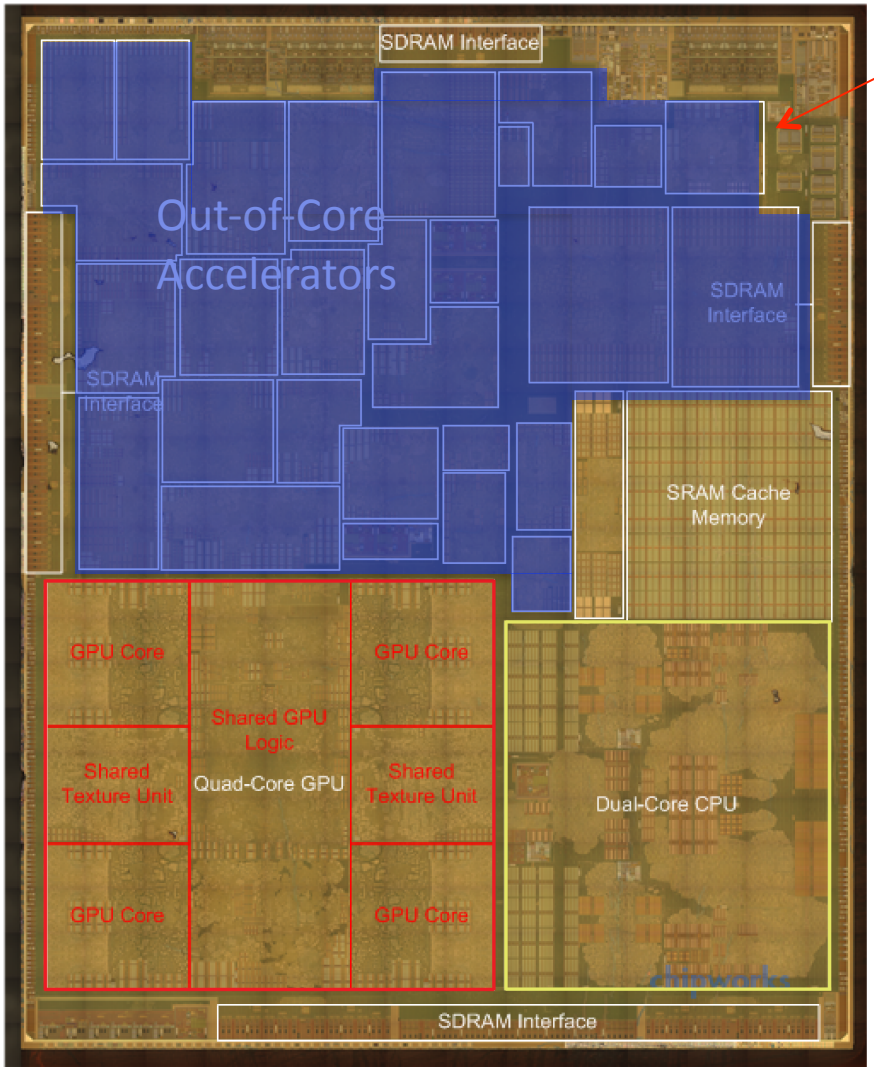
The HSA Vision…

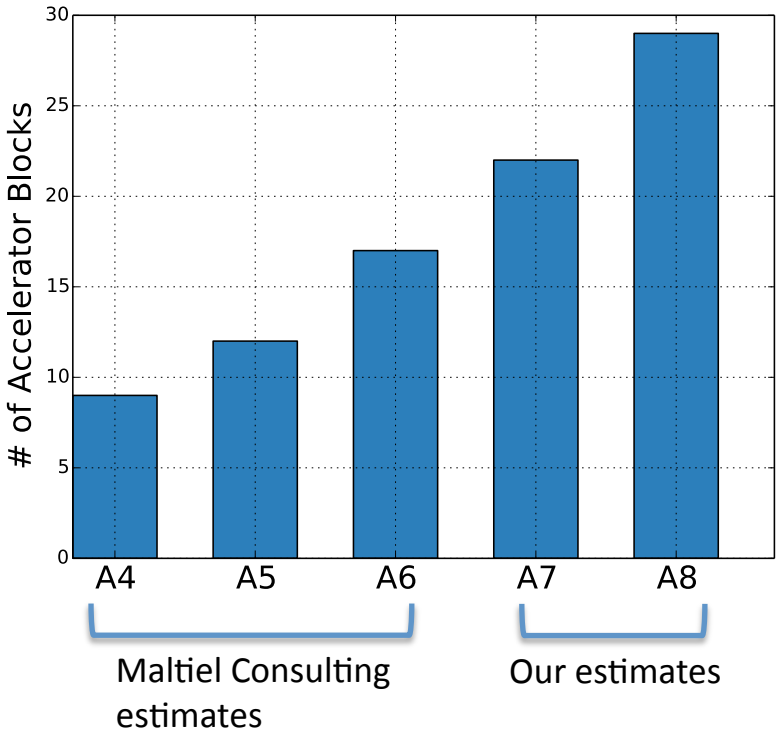
THE HSA FUTURE

- Architected heterogeneous processing on the SOC
- Programming of accelerators becomes **much easier**
- Accelerated software that runs across multiple hardware vendors
- Scalability from smart phones to super computers on a common architecture
- GPU acceleration of parallel processing is the initial target, with DSPs and other accelerators coming to the HSA system architecture model
- Heterogeneous software ecosystem evolves at a much faster pace
- Lower power, more capable devices in your hand, on the wall, in the cloud

CPUs, GPUs, and Accelerators:

Apple A8 SoC





From David Brooks, Harvard, 2014

Lessons Learned

- From Exa to Nano
  - Heterogeneous Compute, GPU-compute, Co-design
  - Data Movement: New Memory archs, NoC, Resiliency
  - Open SW Systems, MPI + X,
- From Nano to Exa
  - Dedicated DSP, GPU and other accelerators
  - Customized data representations
  - Low-power design: Near-threshold, Leakage control, DVFS
- Software development (HW/SW “contracts”)
    - Data representations: types/arithmetic
    - Memory models, Parallel programming

Fabien Clermidy

CEA-LETI

FROM RESEARCH TO INDUSTRY

cea tech

Computing goes 3D

Fabien Clermidy, PhD

Head of Digital Architecture & Design Lab

www.cea.fr

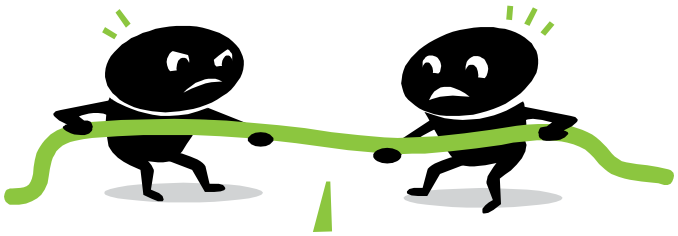
leti & list

Technology versus application

- Scalable
- Flexible
- Adaptable
- Low NRE cost

- Energy efficient
- High performance
- Small area

General purpose



Dedicated Units

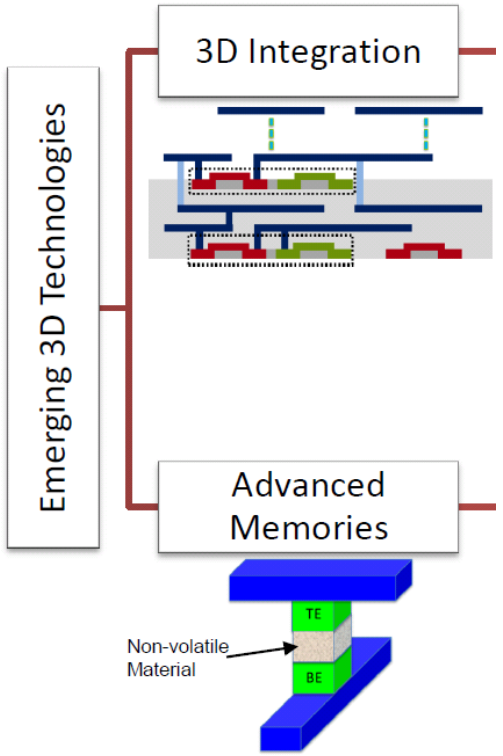
3D Technologies

- 3D stacking

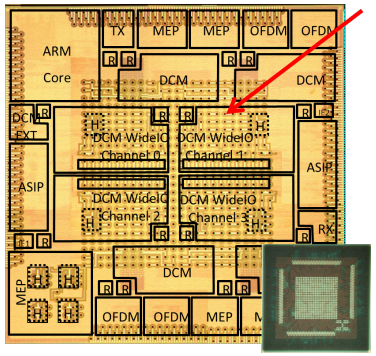
  - Heterogeneous integration
  - Yield improvement
  - Servers, big data

- Monolithic 3D Integration
  - Reduce Silicon footprint
  - Reduce routing wirelength
  - Low-power & high performance

- BEOL NVM
  - Logic in memory designs
  - Non-volatility for normally-off IoT



3D Stacking



Cost

High Yield, time-to-market, new functionalities

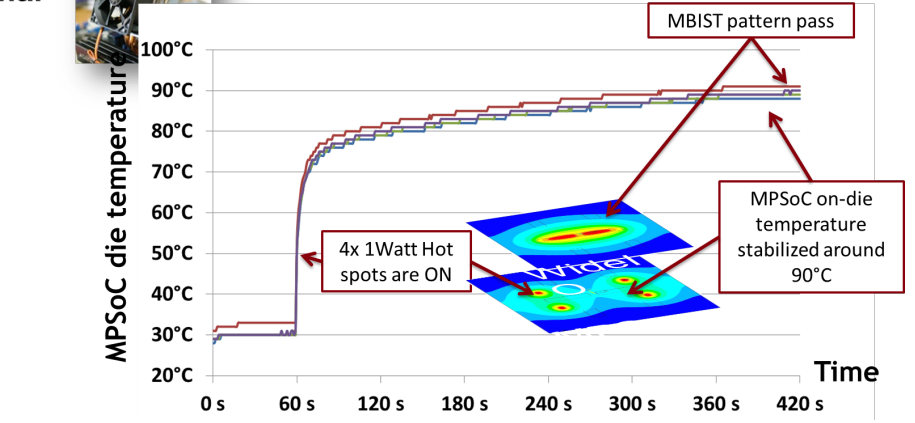
3D integration

High inter-die bandwidth

Ultra-short reach interconnects

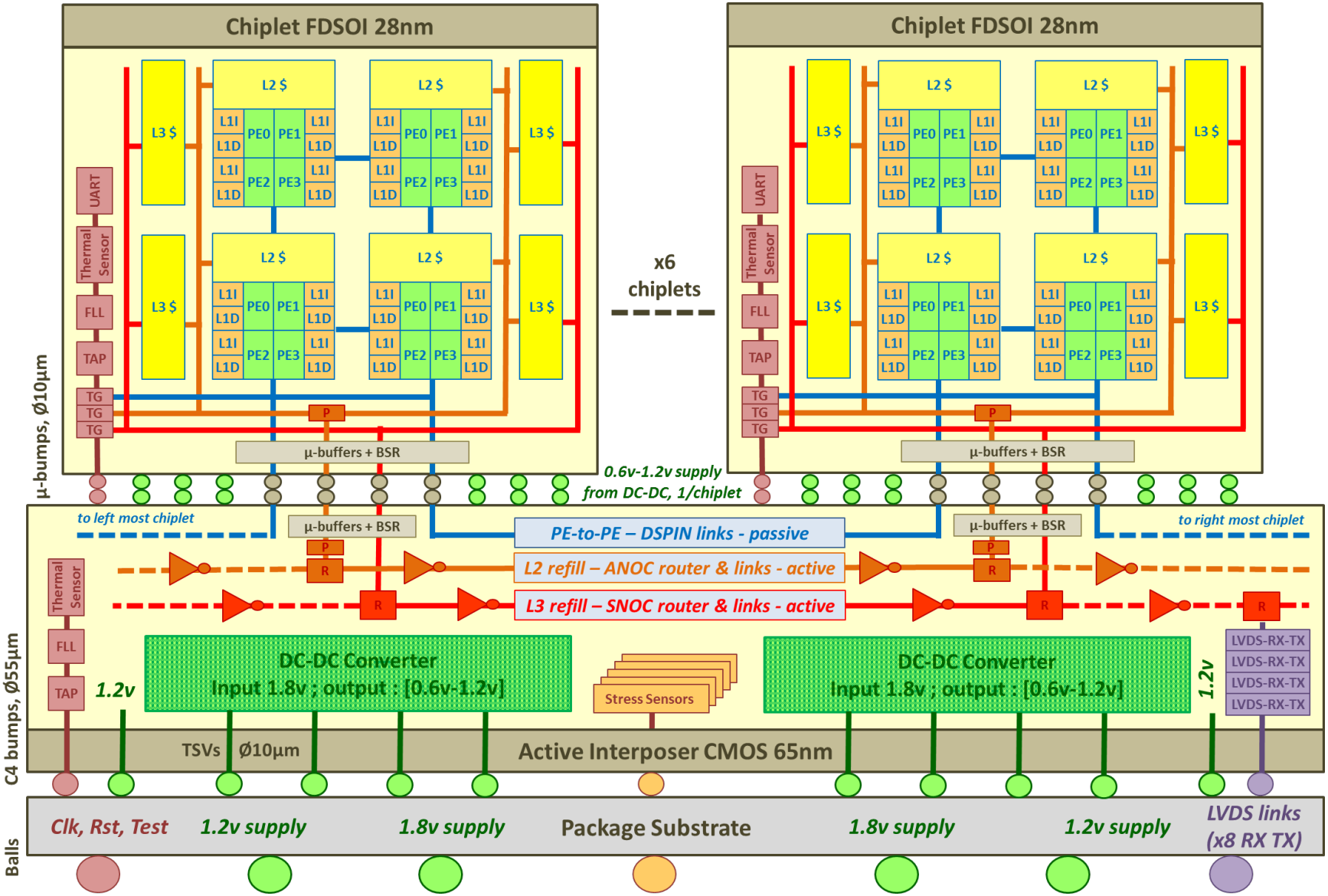
Memory and I/O bandwidth

Energy and thermal

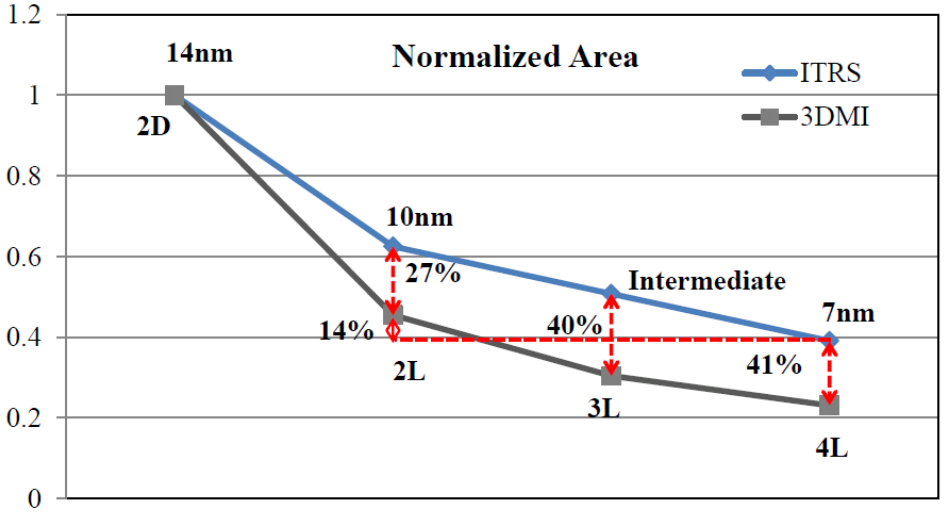


Memory Type	LPDDR3 - [1]	WideIO - This work
Package	PoP / Discrete	3D-IC
BW (Gbyte/s)	6.4 GB/s	12.8 GB/s
Total power		293 mW*
MPSoC power		121 mW*
Memory Power		81 mW*
I/O power		91 mW*
I/O power efficiency	3.7 pJ/bit**	0.9 pJ/bit*

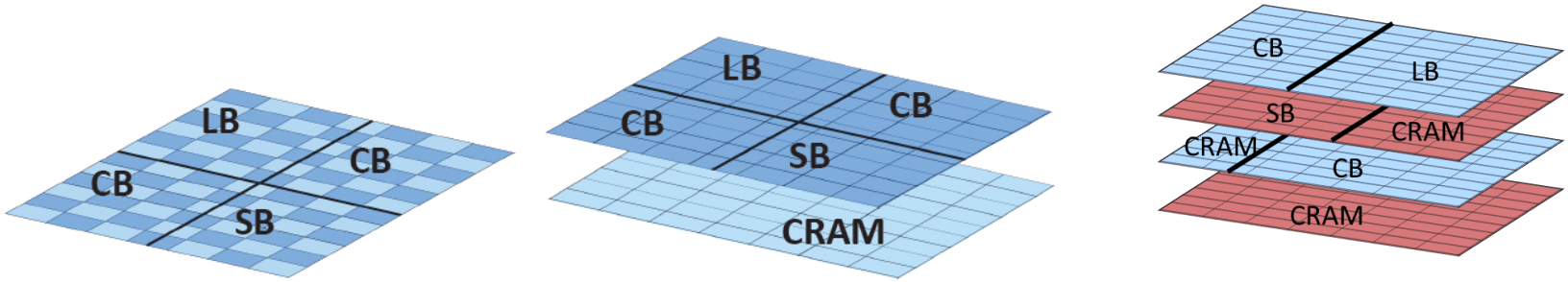
3D many-core interposer



3DMI vs. Traditional Scaling

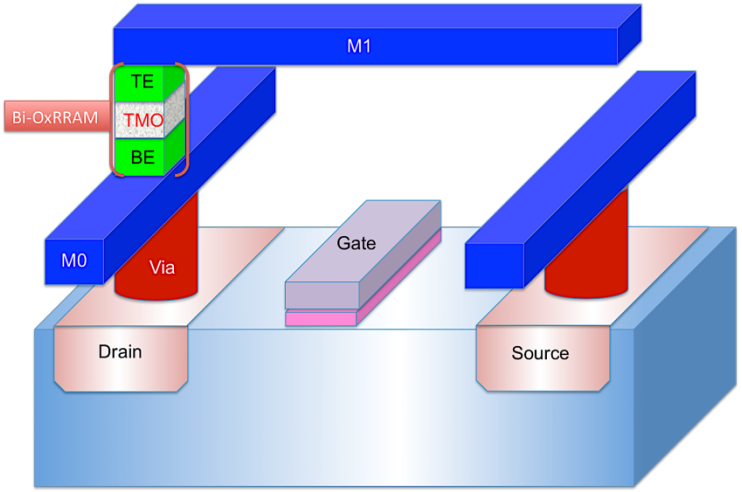


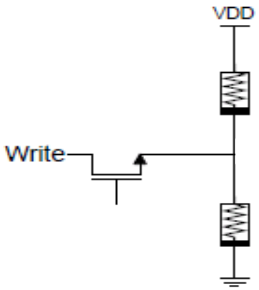
- Case study: FPGA partitioning on 2/3/4 layers
- Comparison between 3DMI FDSOI 14nm (available) and ITRS FinFET 10 & 7 nm (predictive)



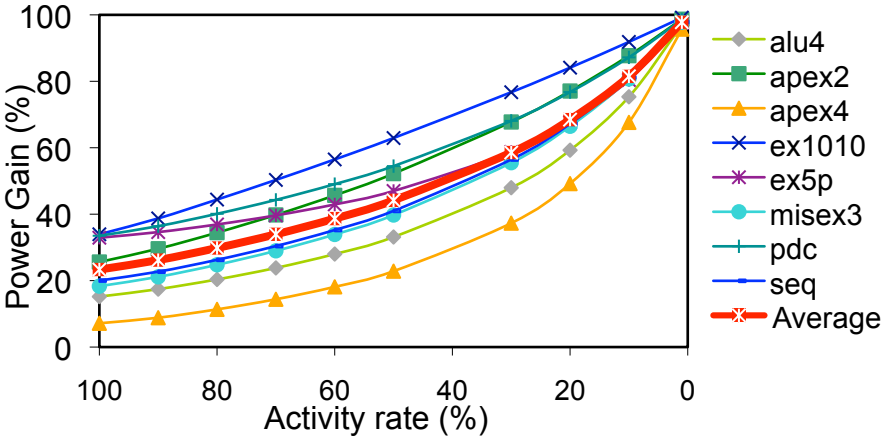
RRAM

- Features:
  - BEOL technology => low footprint
  - Non-volatile => no leakage
- Applications
  - Memory-in-logic => IoT
  - High density crossbars => Big data





	OxRAM [Chen.IEDM 2009]	PCRAM [Servalli.IEDM 2009]	CBRAM [Palma.TED 2014]
R_{on}	1kΩ	10kΩ	4.2kΩ
R_{off}	100kΩ	2MΩ	10GΩ
$I_{leakage}$	15µA	0.75µA	0.15nA



contact.dacle@cea.fr

INSTITUT CARNOT CEA LETI

INSTITUT CARNOT CEA LIST

MINATEC®

digiteo

leti

Centre de Grenoble
17 rue des Martyrs
38054 Grenoble Cedex

list

Centre de Saclay
Nano-Innov PC 172
91191 Gif sur Yvette Cedex

Enrico Macii

Politecnico di Torino

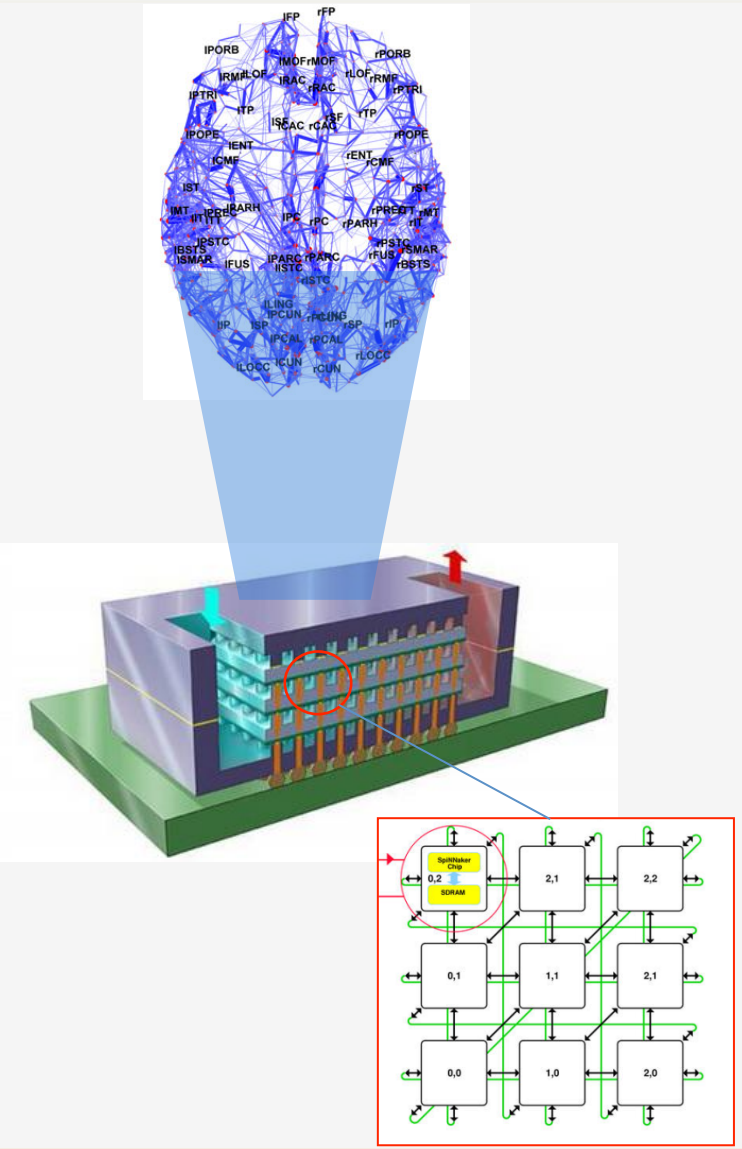
Neuromorphic Computing Models:
Optimization of Multicore Neuromorphic Computing Platforms

POLITECNICO
DI TORINO

Enrico Macii
Dip. di Automatica e Informatica

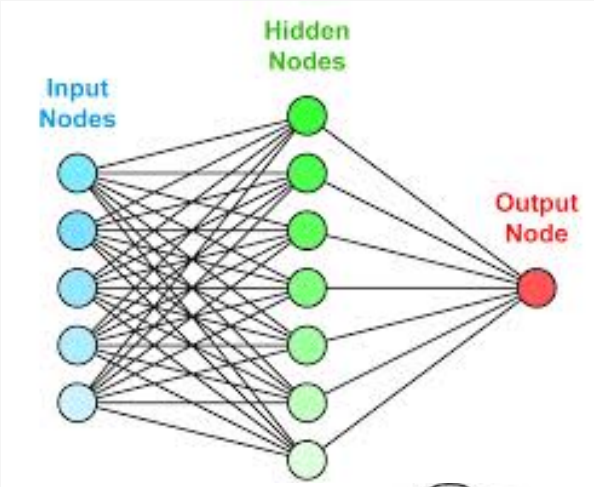
Brain-like Computing Systems

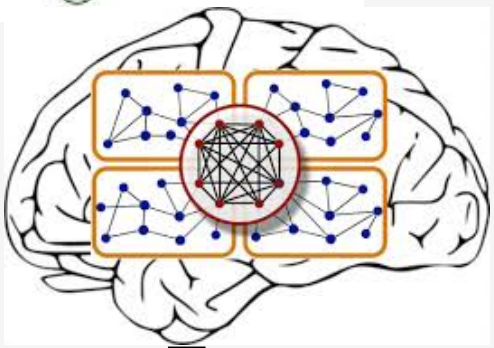
- Neuromorphic HW devices/platforms ("artificial brains"):
  - Densely interconnected multicores – brain emulation (NM-MC)
  - New devices implementing HW neuron models (e.g., memristors)
- Potential:
  - Innovative brain-inspired computing methods
  - Real-time emulation of (part of) the human brain (not achievable with SW neuron simulators)
- Challenges:
  - Brain simulation/emulation capability is currently limited to a few thousands of neurons (fly's brain)
  - Support neural network features (e.g., plasticity)
- Focus on NeuroMorphic-MultiCores (NM-MC)
  - NM HW @ HBP:
    - BrainScaleS (U. Heidelberg)
    - SpiNNaker (U. Manchester)

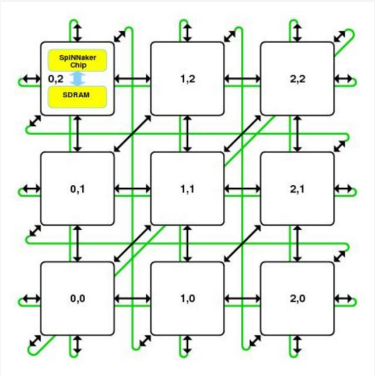


Define neuron models

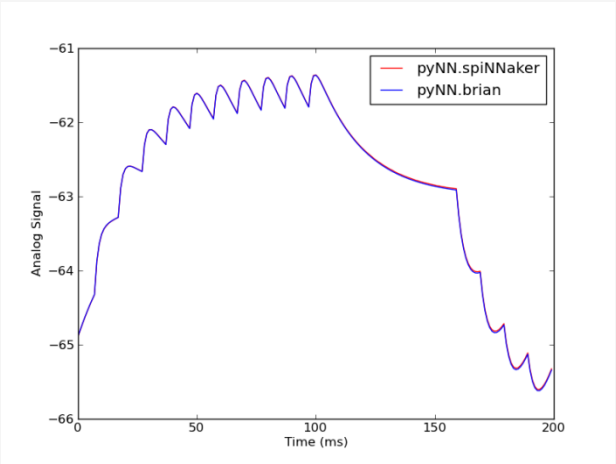
Define neuron networks and populations





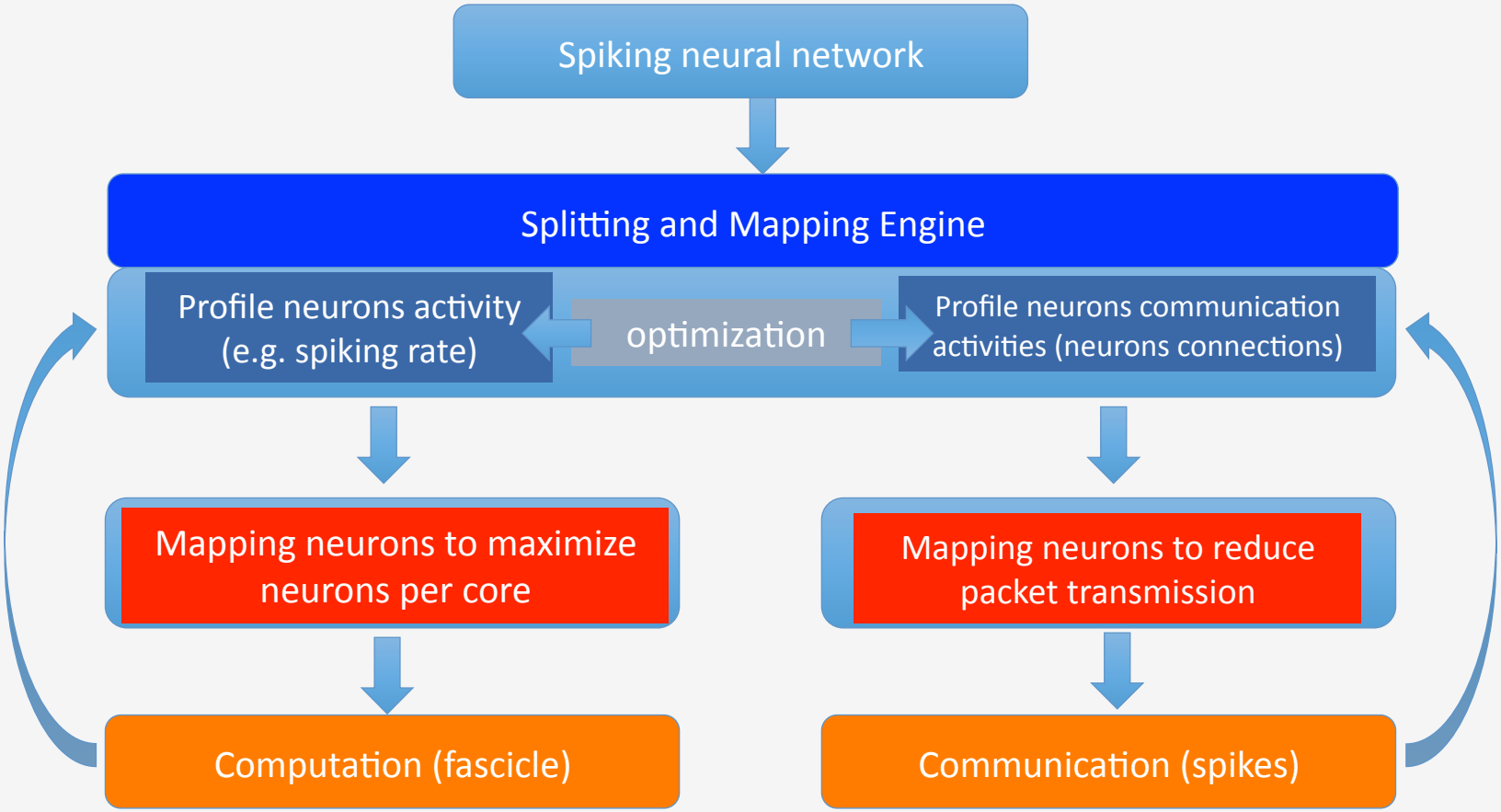


Allocate populations to cores
(max ~100 per core)



Perform neuron activity simulation

Conceptual Scheme



We are currently looking at optimizing:

- **Mapping of neurons on NM-cores depending on their activity (spiking rate)**
- **Network usage by communication–aware neuron population splitting and allocation**

Discussion

Existing knowledge of brain functions is used to design an affordable supercomputer that can itself serve as a tool to investigate brain behavior…

…and that it can contribute to a fundamental, biological understanding of how the brain works.

However, this research has possible impact outside brain simulation:

- Design brain interfaces for robotics
- Provide insights into specific properties of the different hardware architectures
- Explore non-von Neumann computing outside the realm of brain-science
- Develop resource optimization techniques (efficient on-chip/off-chip, energy efficient computation) for densely interconnected multi-core systems

Angel Rodriguez-Vazquez

University of Sevilla

Centre SI
Integrated Systems Centre

EPFL
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

nano-tera.ch

VISION BEYOND IMAGING

Hybrid Cellular Architectures

for

High-Sensitivity, High-Speed, High-Resolution

Vision Systems

with

Reduced SWaP

Angel Rodríguez-Vázquez

Institute of Microelectronics of Seville-SPAIN

University of Seville

Spanish Council of Research (CSIC)

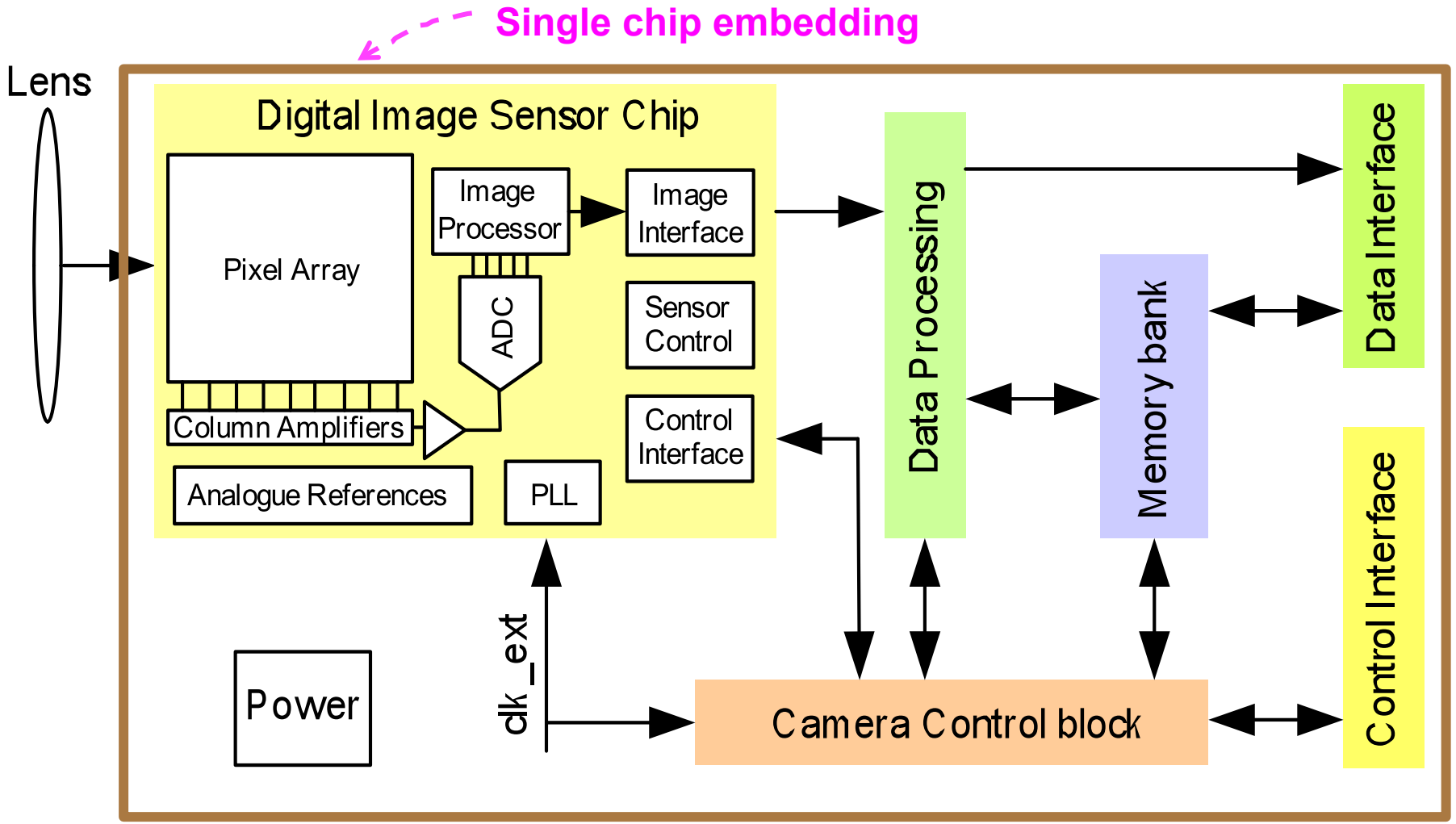
arodri-vazquez@us.es, angel@imse-cnm.csic.es

Trends of MicroElectronic Imaging Systems

- Reducing the PIXEL SIZE
  - even despite diffraction limits
  - < 1mm @ 4T-pixels with transistor sharing
- Increasing the pixel FILL FACTOR
  - Use of Back Side Illumination technologies
- Increasing the PIXEL COUNT
  - > 10Mpixels
- Increasing READOUT SPEED AND DATA TRANSFER
  - > 10Gpixel/sec
- Increasing READ-OUT ACCURACY
  - < 1e- noise > 150dB DR Image acquisition
- Increasing SYSTEM EMBEDDING
  - Single Chip "Camera"

HUGHE NUMBER OF DATA AND INTENSIVE DATA MANIPULATIONS !!

Typical CMOS Image Sensor Architecture

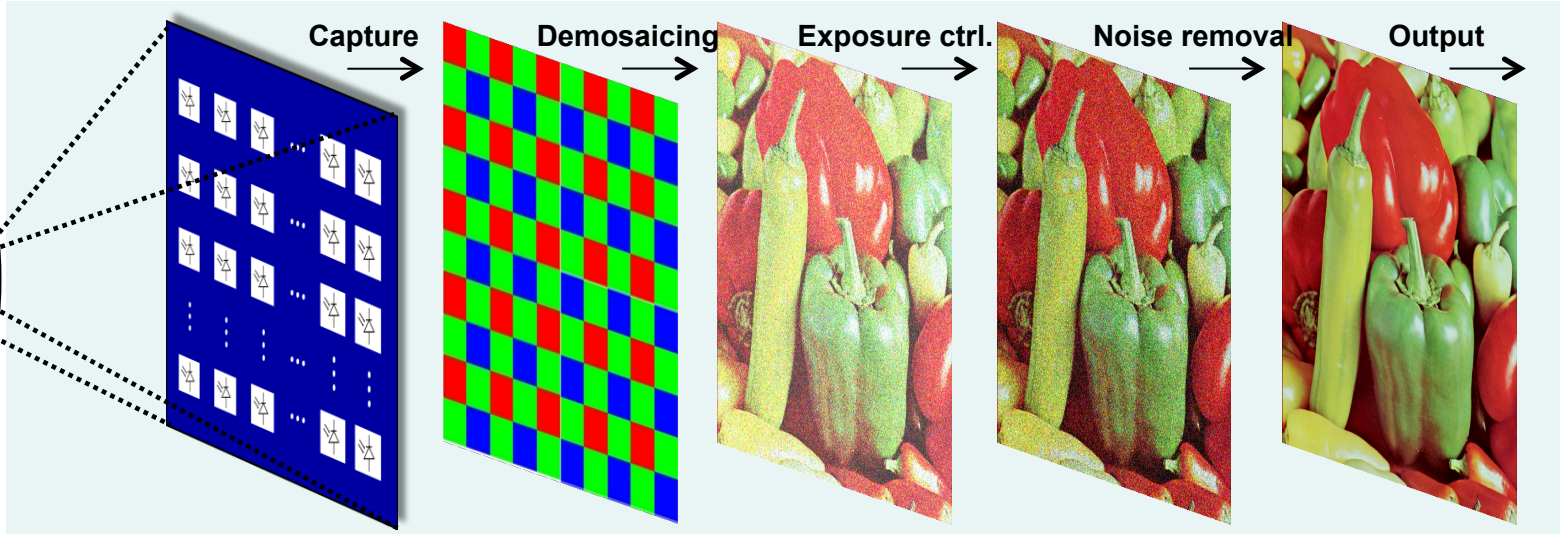


- **FRAME-BASED CONCEPT; clear border between sensing and processing**

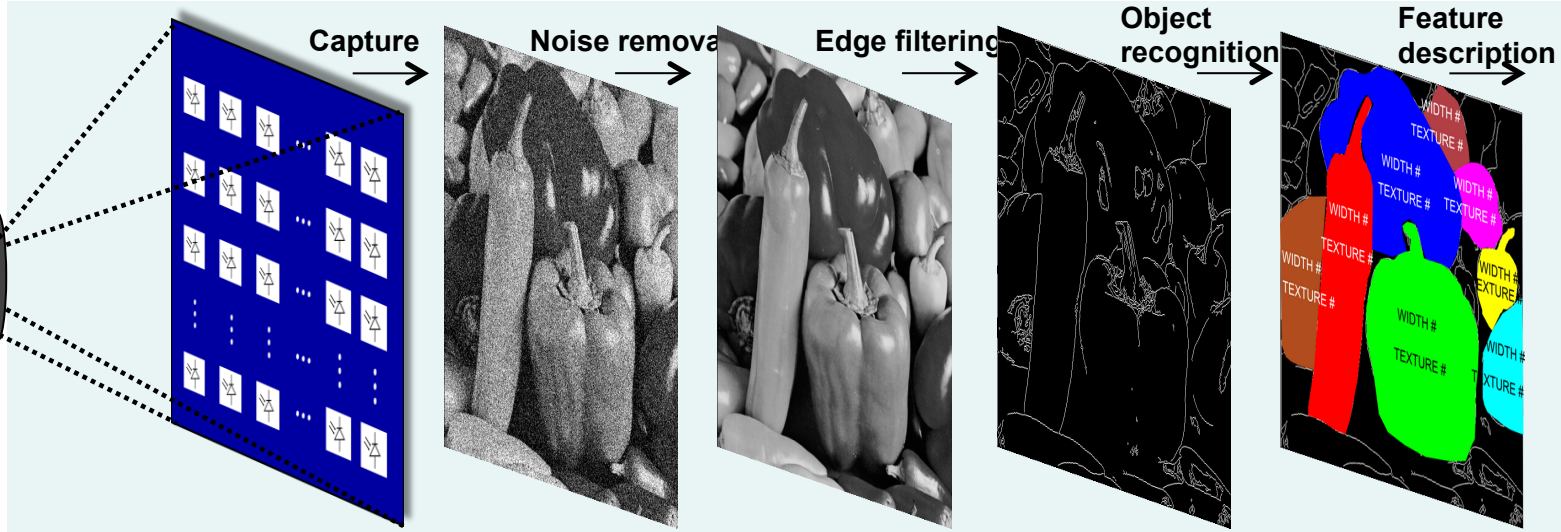
- **IS THIS ADEQUATE FOR VISION ??**

Vision and Imaging have Different Goals !!

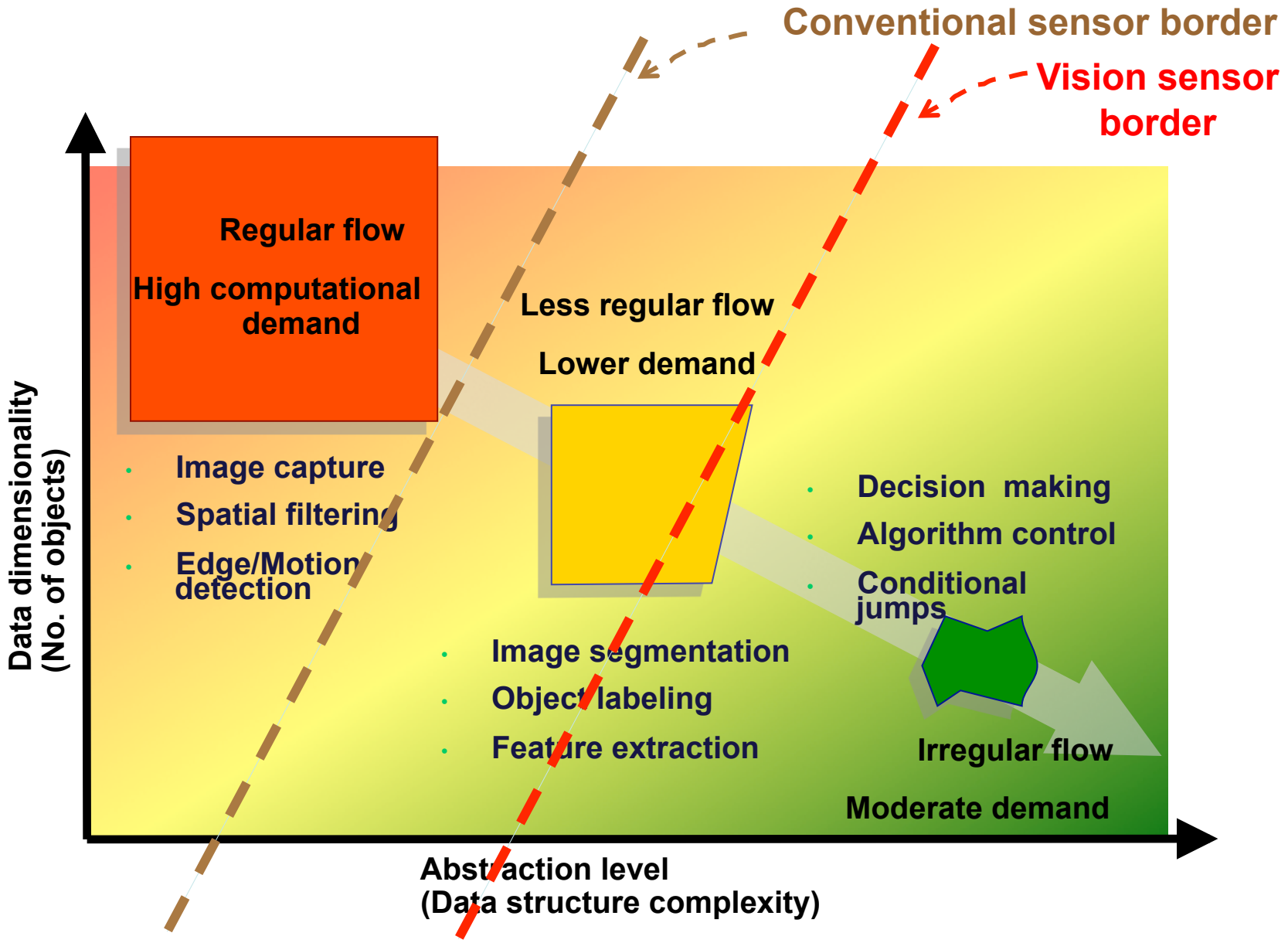
IMAGING MAIN OBJECTIVE: image quality

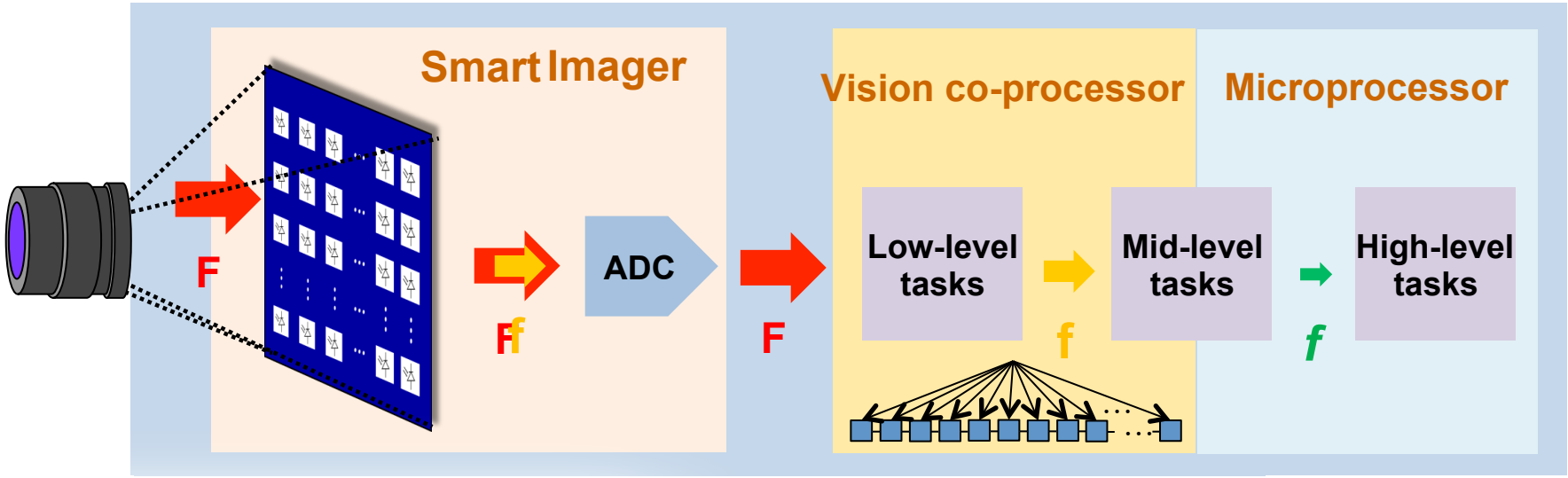


VISION MAIN OBJECTIVE: scene understanding

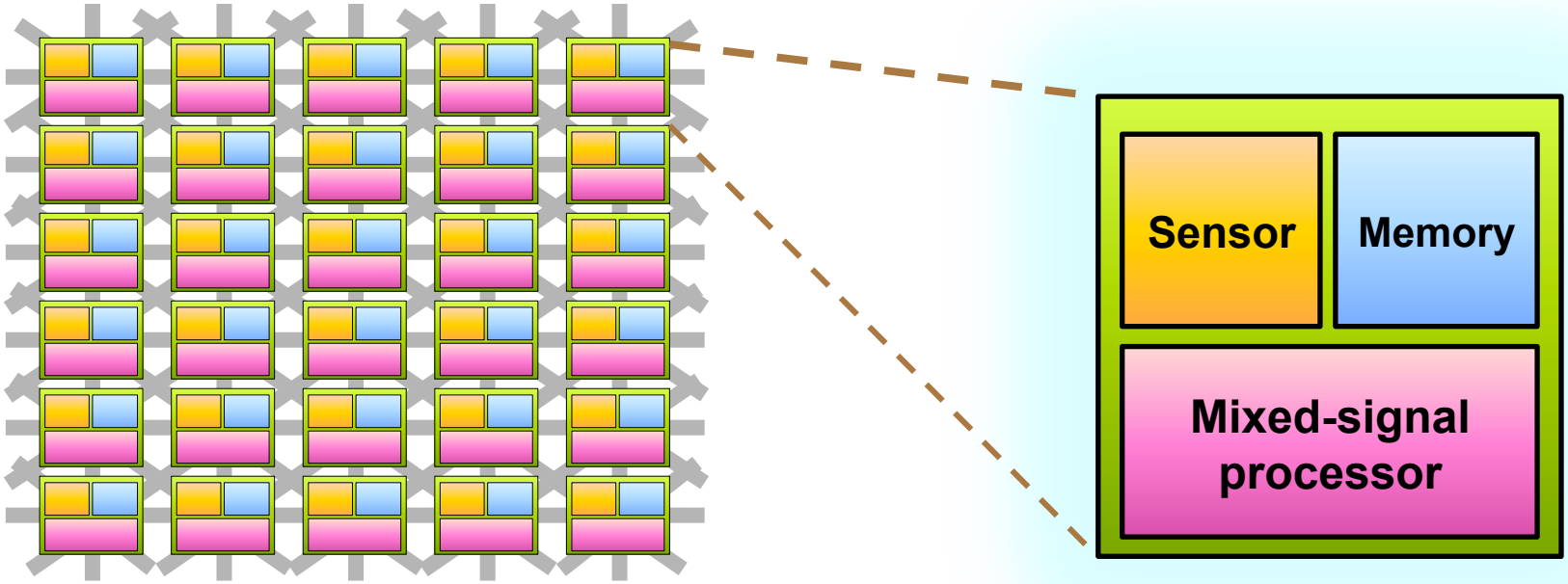


Data Reduction in the Vision Processing Chain



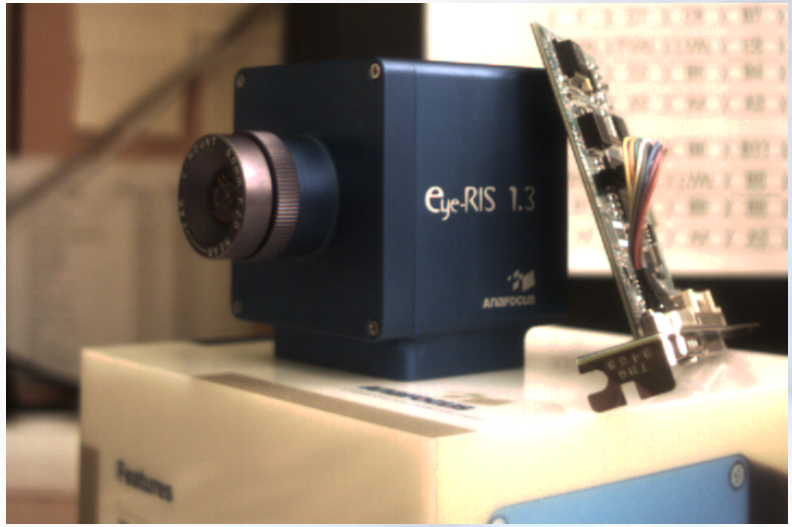


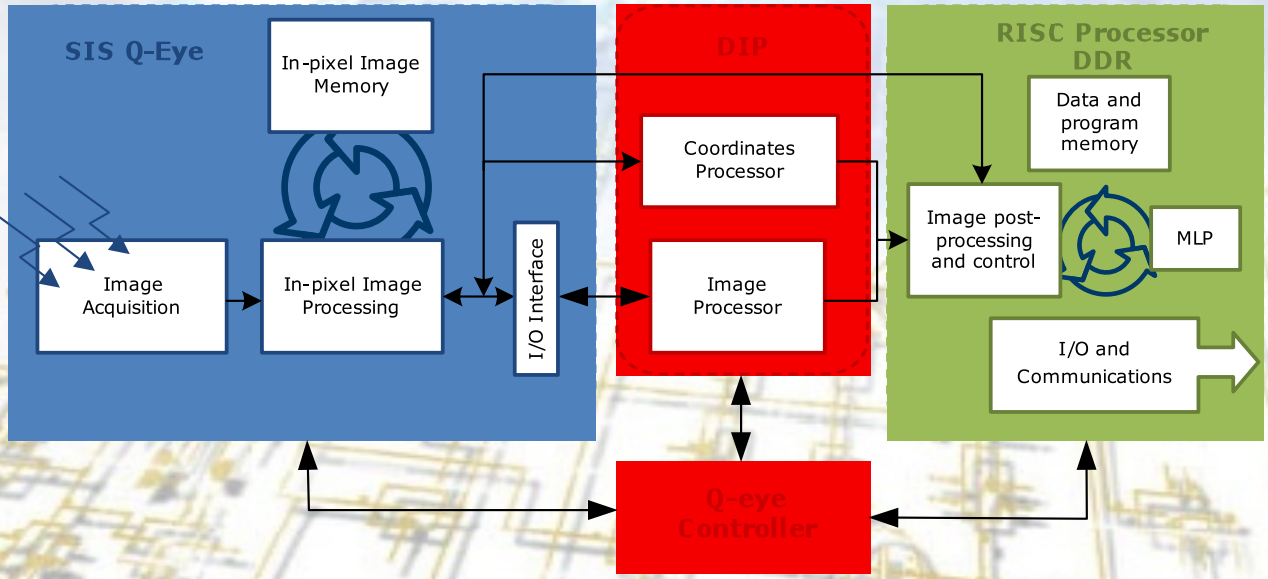
Information Flow: F >> f > *f*



SIMD – Cellular Nonlinear Networks provide a computation framework for that

Eye-RIS: A Representative Industrial SIMD-CNN Vision Systems





AnaFocus

Some Open Points for Discussion

- How to achieve convergence with mainstream industrial approaches ??

  Hardware-software standards

- Which can kind of data encoding will be used for sensor-processor outcomes ??

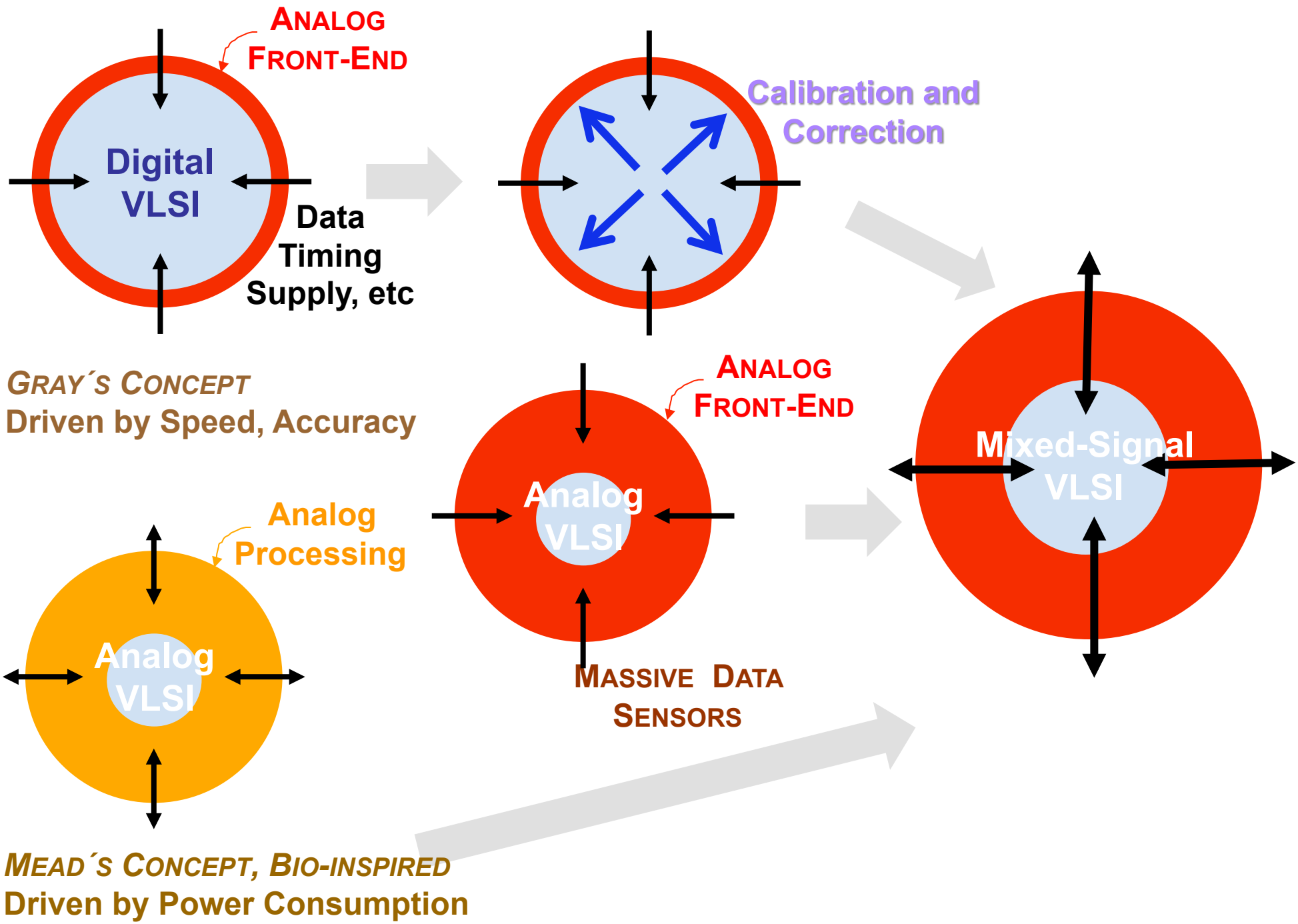
  only spatial-temporal events

  data features

- Is it possible to derive a reduced set of data features which constitute an universal basis for vision tasks ??

  ad-hoc sensor design constraint industrial deployment

- May conventional frame-based architectures confront the challenges of vision ??


ANALOG
FRONT-END
Digital
VLSI
Data
Timing
Supply, etc
Calibration and
Correction
GRAY´S CONCEPT
Driven by Speed, Accuracy
ANALOG
FRONT-END
Analog
VLSI
Mixed-Signal
VLSI
Analog
Processing
Analog
VLSI
MASSIVE DATA
SENSORS
MEAD´S CONCEPT, BIO-INSPIRED
Driven by Power Consumption

Emerging Methods of Computing

Luca Benini	ETHZ
Wayne Burleson	University of Massachusetts
Fabien Clermidy	CEA-LETI
Enrico Macii	Politecnico di Torino
Angel Rodriguez-Vazquez	University of Sevilla

Chair: Yusuf Leblebici, EPFL

Centre SI
Integrated Systems Centre

EPFL
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

nano-tera.ch