New computing architectures for Green ICT

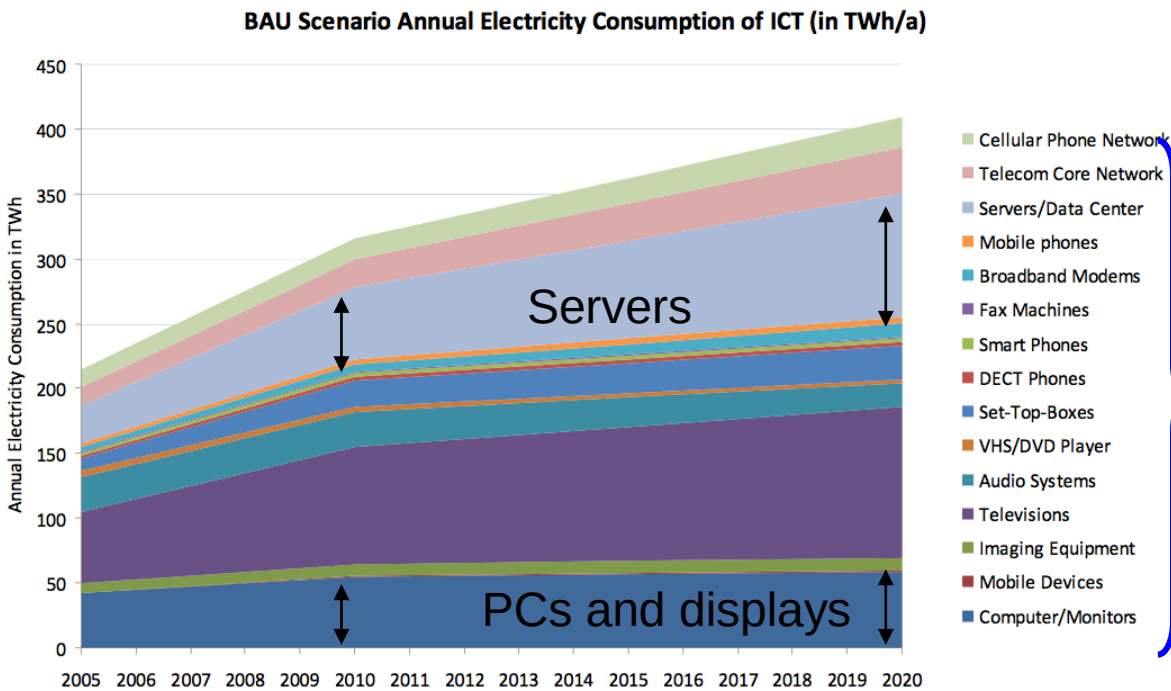
Marc Duranton
Embedded Computing Laboratory (LCE)
CEA LIST DACLE

September 5th, 2011

Impact of ICT on global energy consumption

Energy consumption of ICT

Forecast for all ICT domain by the European Commission



"Business as Usual"
300 TWh in 2010
with 70 TWh
for the ~10 millions Servers
409 TWh in 2020
with100 TWh for Servers

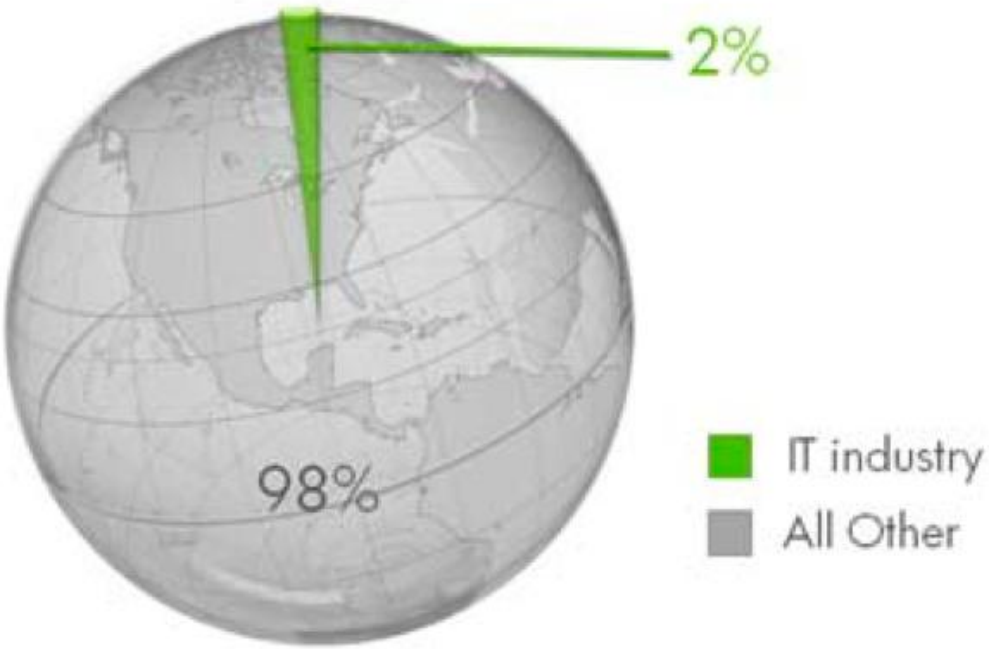
Nota : 1 EPR ~10 TWh

Source: European Commission DG INFSO, Impact of Information and Communication Technologies on Energy Efficiency, final report, 2008
ftp://ftp.cordis.europa.eu/pub/fp7/ict/docs/sustainable-growth/ict4ee-final-report_en.pdf

© CEA. All rights reserved

ICT: 2% of

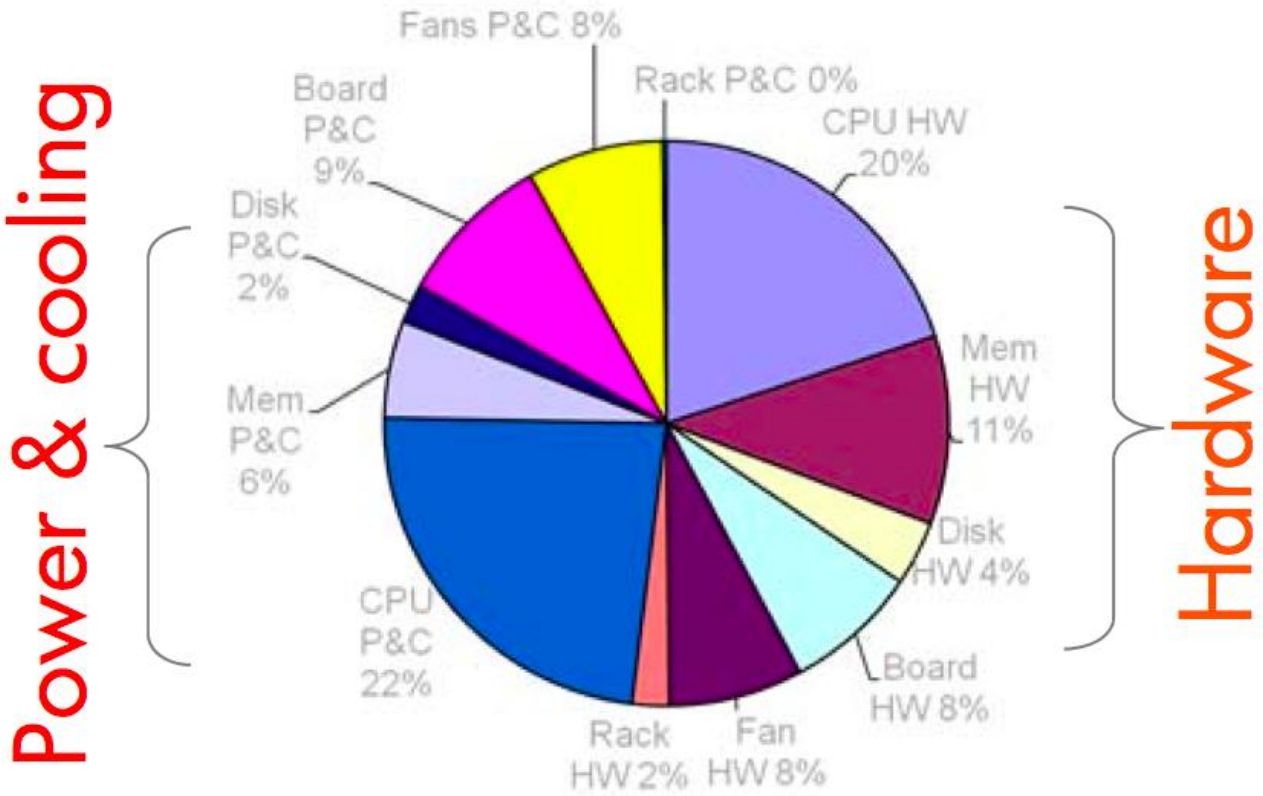
Global carbon emissions



Source: P. Ranganathan, “Saving the world together, one server at a time… » ACACES 2011

© CEA. All rights reserved

Servers: more energy spend in power and cooling parts than in compute?



I will focus only on "compute" engines

Source: P. Ranganathan, "Saving the world together, one server at a time… » ACACES 2011©

© CEA. All rights reserved

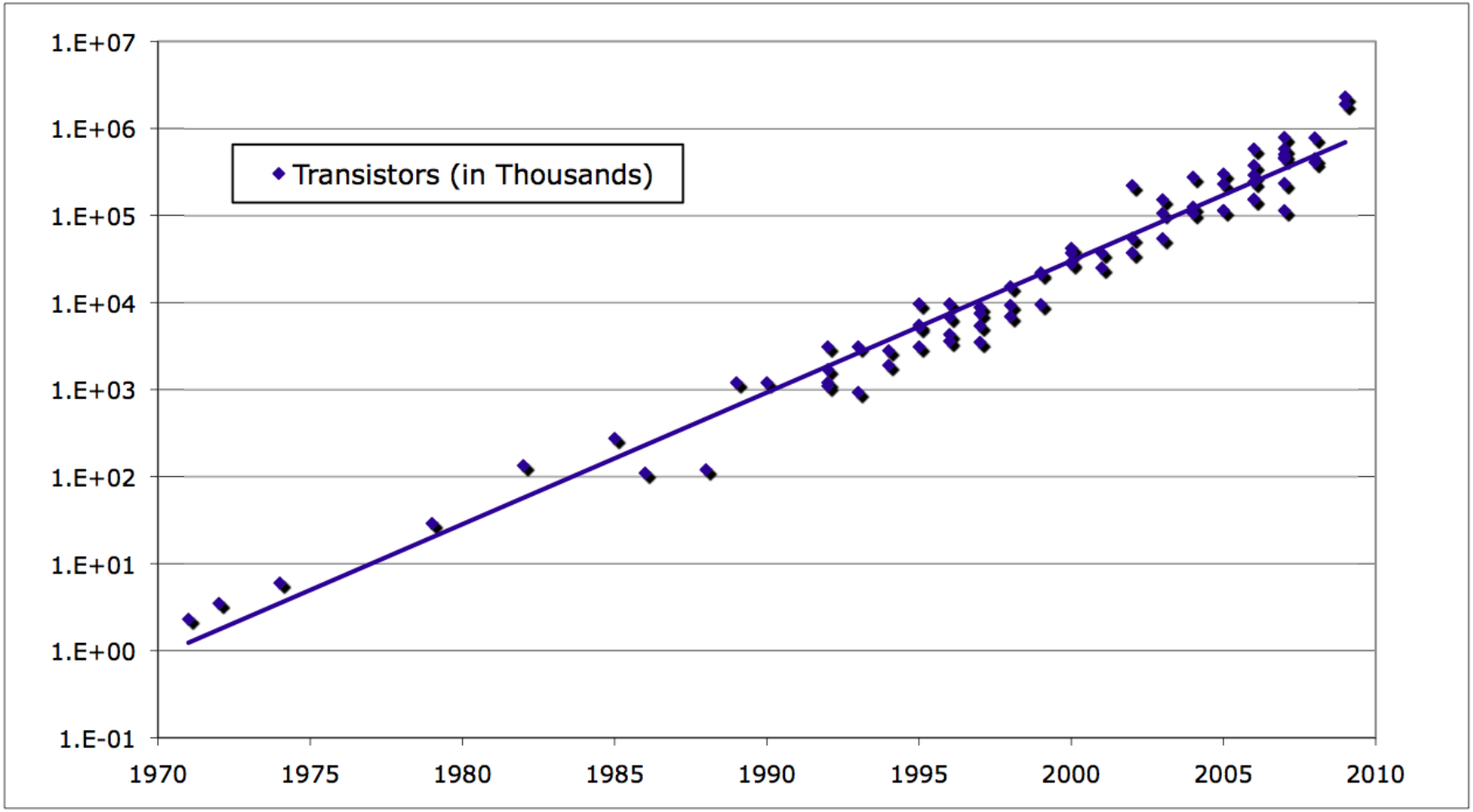
Take home message 1:

- GREEN ICT is important for sustainability
- Servers, PCs and TVs (STB, DVR) are the main contributors
- Portable devices have the highest growth
  - Low power per device
  - Large numbers
- Energy lost in power supply and cooling is also important

© CEA. All rights reserved

Moore’s Law, the fuel of ICT: we are at a turning point!

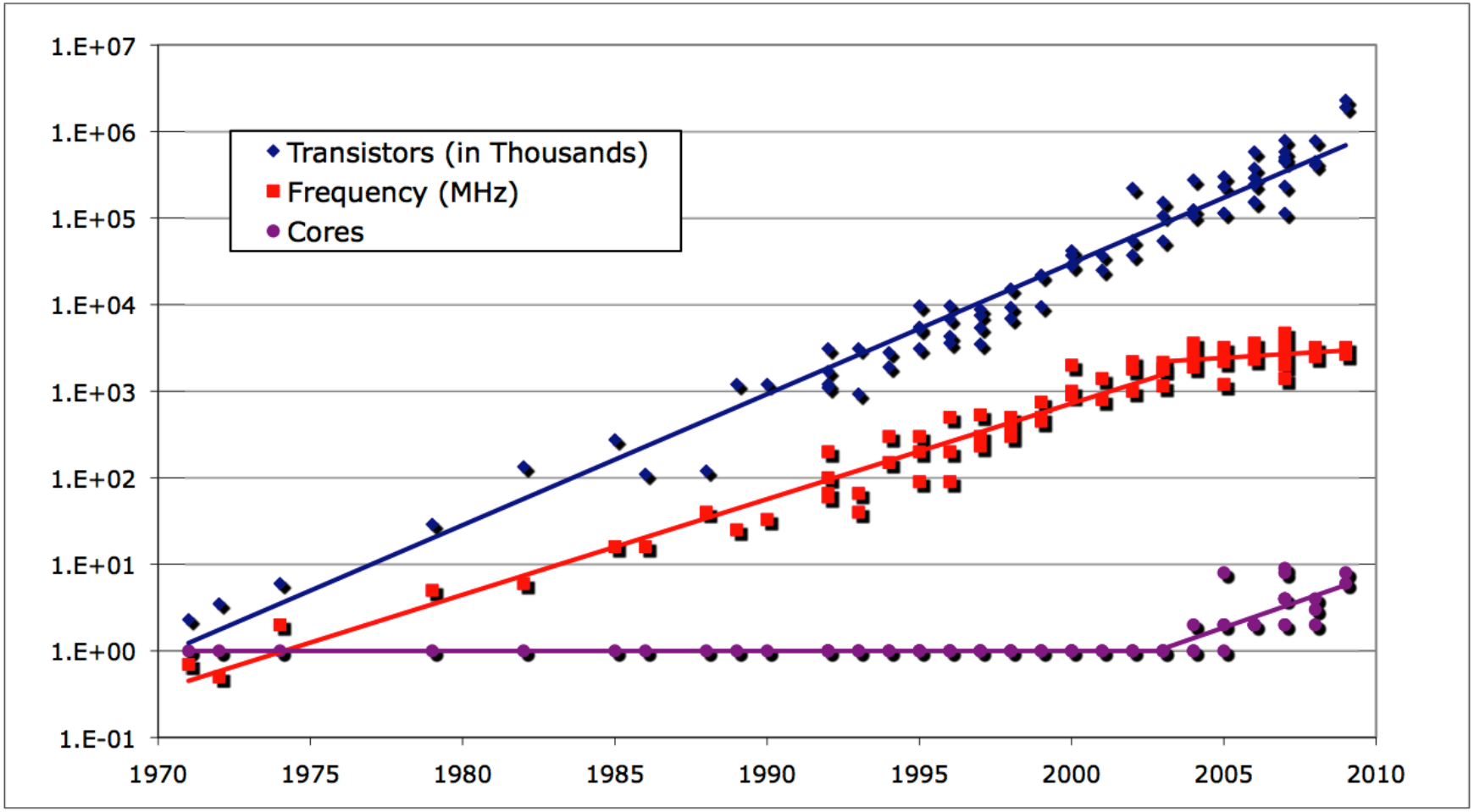
Moore's law: increase in transistor density



Data from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanoviç

© CEA. All rights reserved

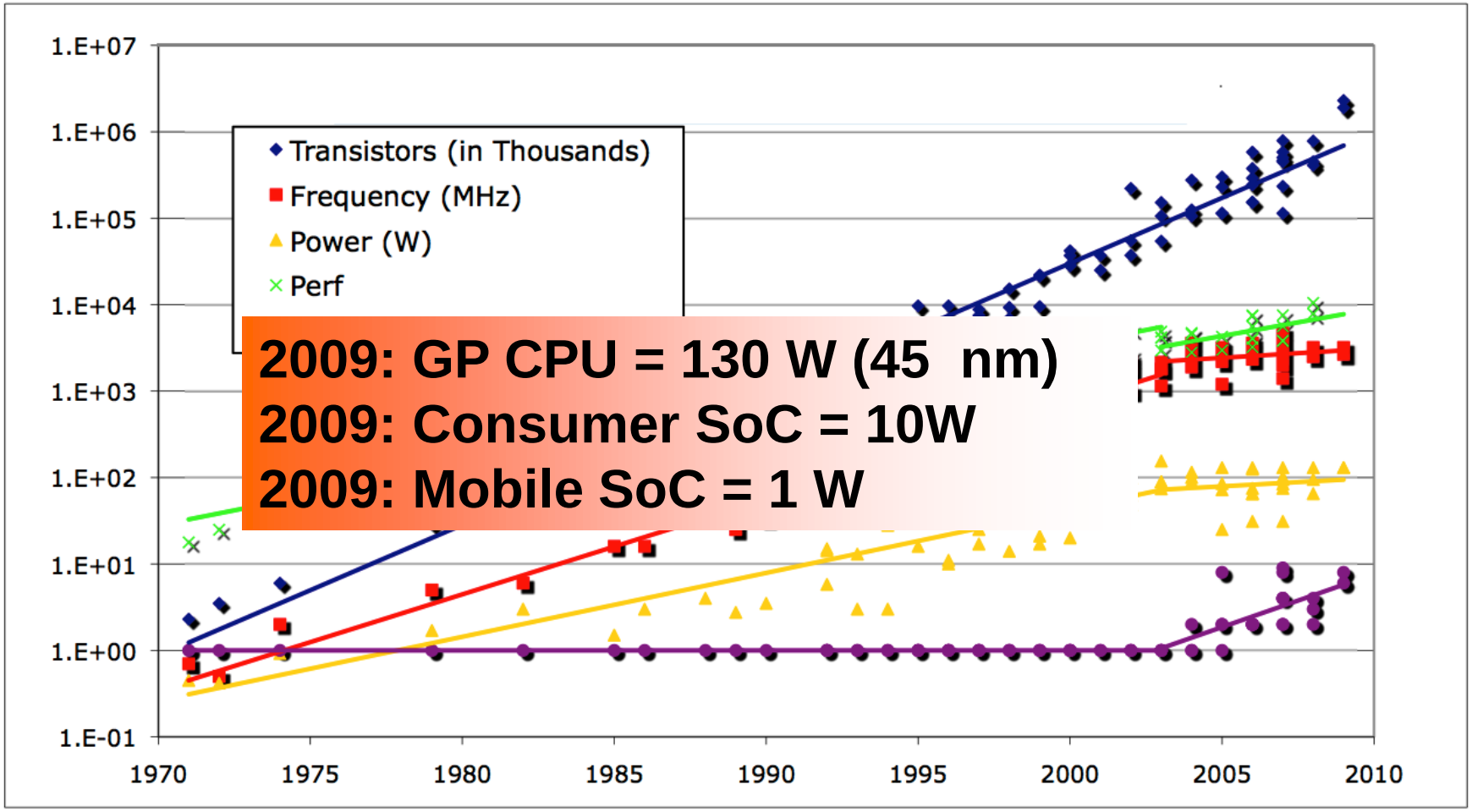
Limited frequency increase => more cores



Data from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanoviç

© CEA. All rights reserved

Power density and dissipation at limit



Data from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanoviç

© CEA. All rights reserved

Take home message 2:

- The era of free frequency scaling is over
- Solution to increase performances: multi-cores
- Limiting factor: power density and dissipation per chip

© CEA. All rights reserved

Many cores: rebirth of simple (and more specialized) cores

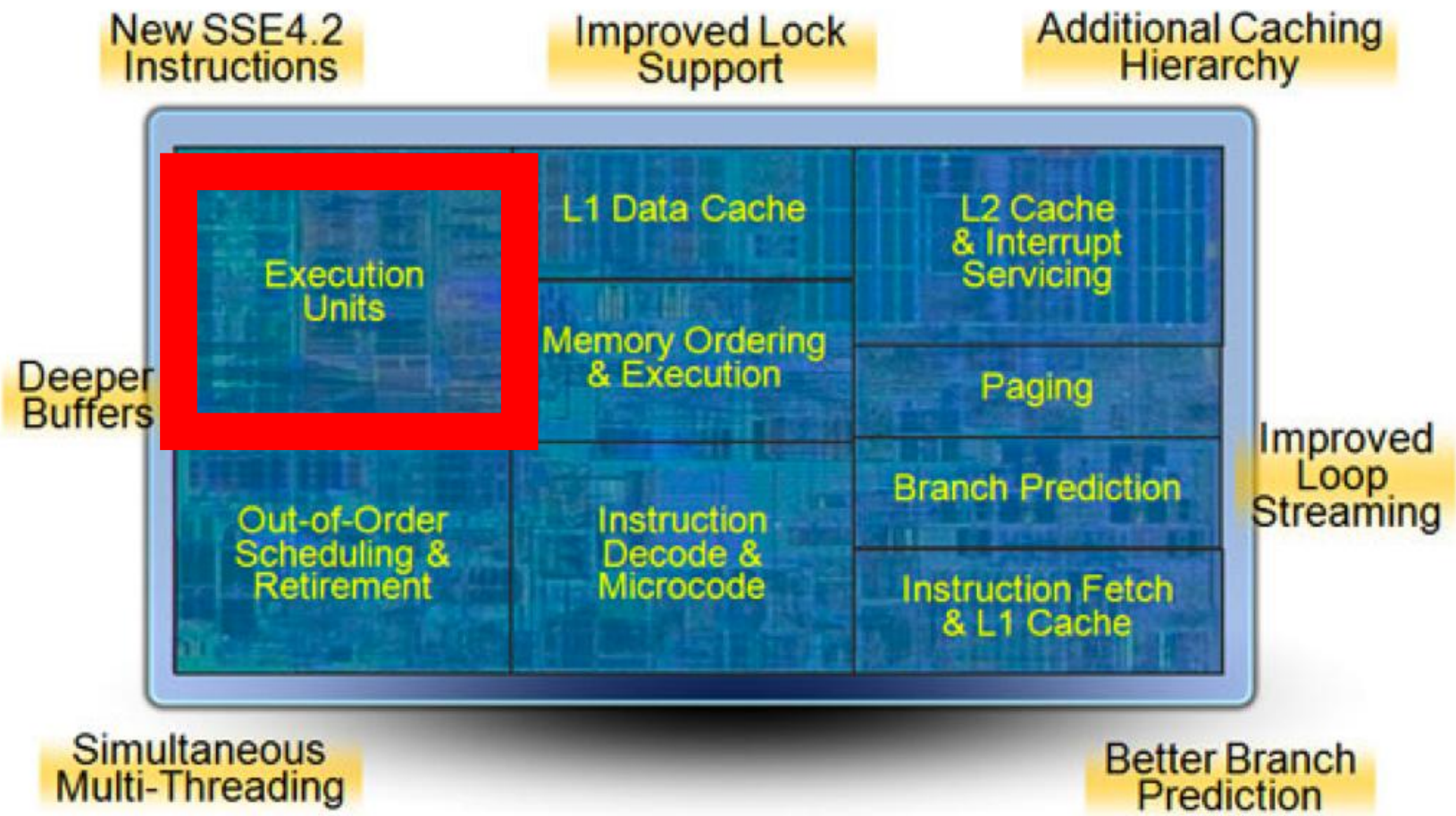
Why using several compute cores?

1. Using several cores is also an answer to the Law of Diminishing Returns [Pollack's Rule] :
   - Effectiveness per transistor decreases when the size of a single core is increased, due to the locality of computation
   - Controlling a larger core and data transport over a single larger core is super-linear
   - **Smaller cores are more efficient** in ops/mm^2/W
2. Large area of today's microprocessors are for best effort processing* and used to cope with unpredictability (branch prediction, reordering buffers, instructions, caches).

* 20+ pipeline stages in high end mono-processors

© CEA. All rights reserved

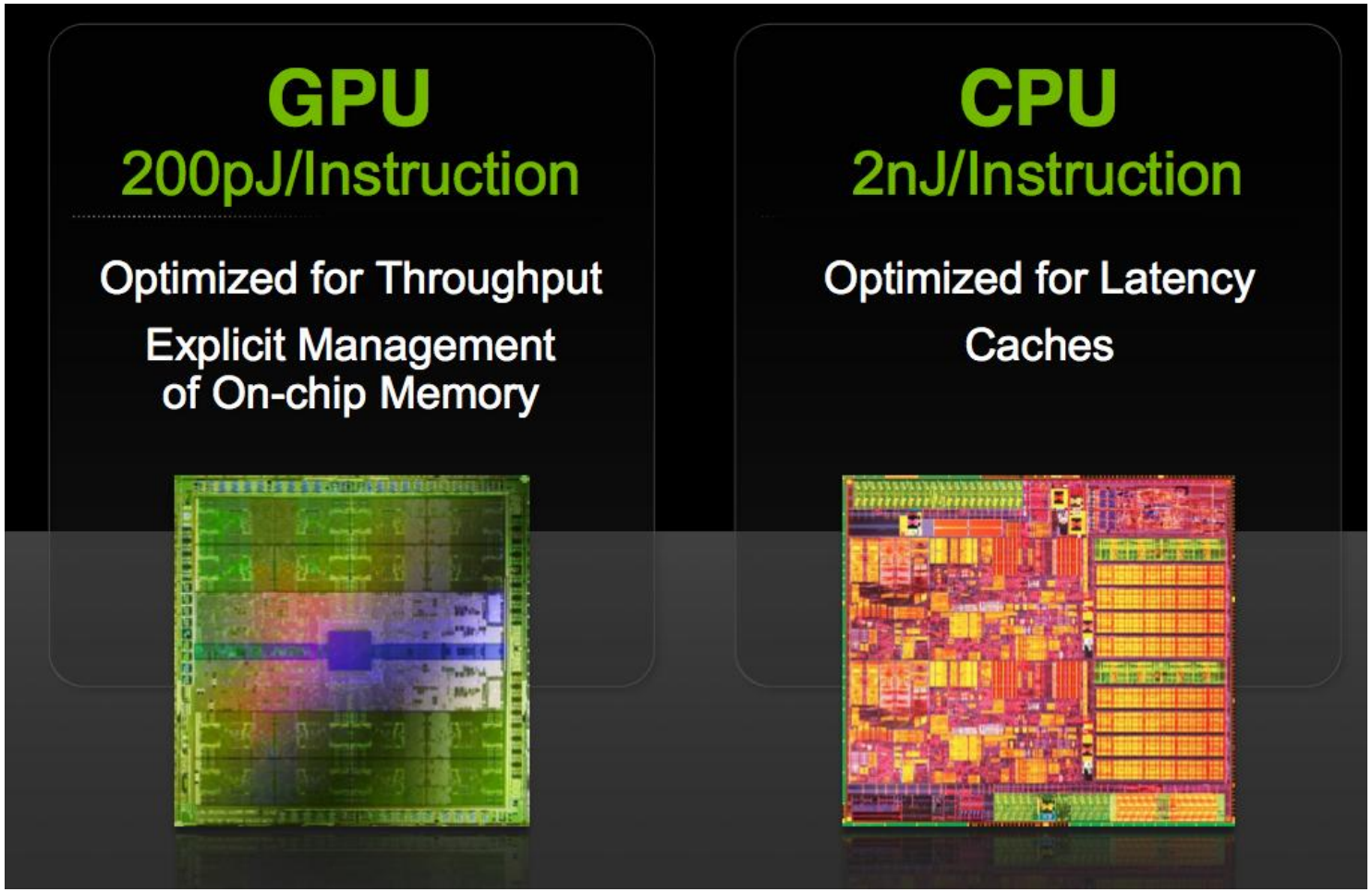
Less than 20% of the area for execution units



Source: Dan Connors, “OpenCL and CUDA Programming for Multicore and GPU Architectures» ACACES 2011

© CEA. All rights reserved

Specialization lead to more efficiency



Source: Bill Dally, « To ExaScale and Beyond »
www.nvidia.com/content/PDF/sc_2010/theater/Dally_SC10.pdf

© CEA. All rights reserved

Take home message 3:

- ***KISS***: Keep It Simple!
- Renew of simple cores, more energy efficient (ops/mm^2/W
- Tune the cores for their primary task: ***specialization*** can give you x10^3 or more
- Forget "as fast as you can" (best effort) and focus on "on time"*

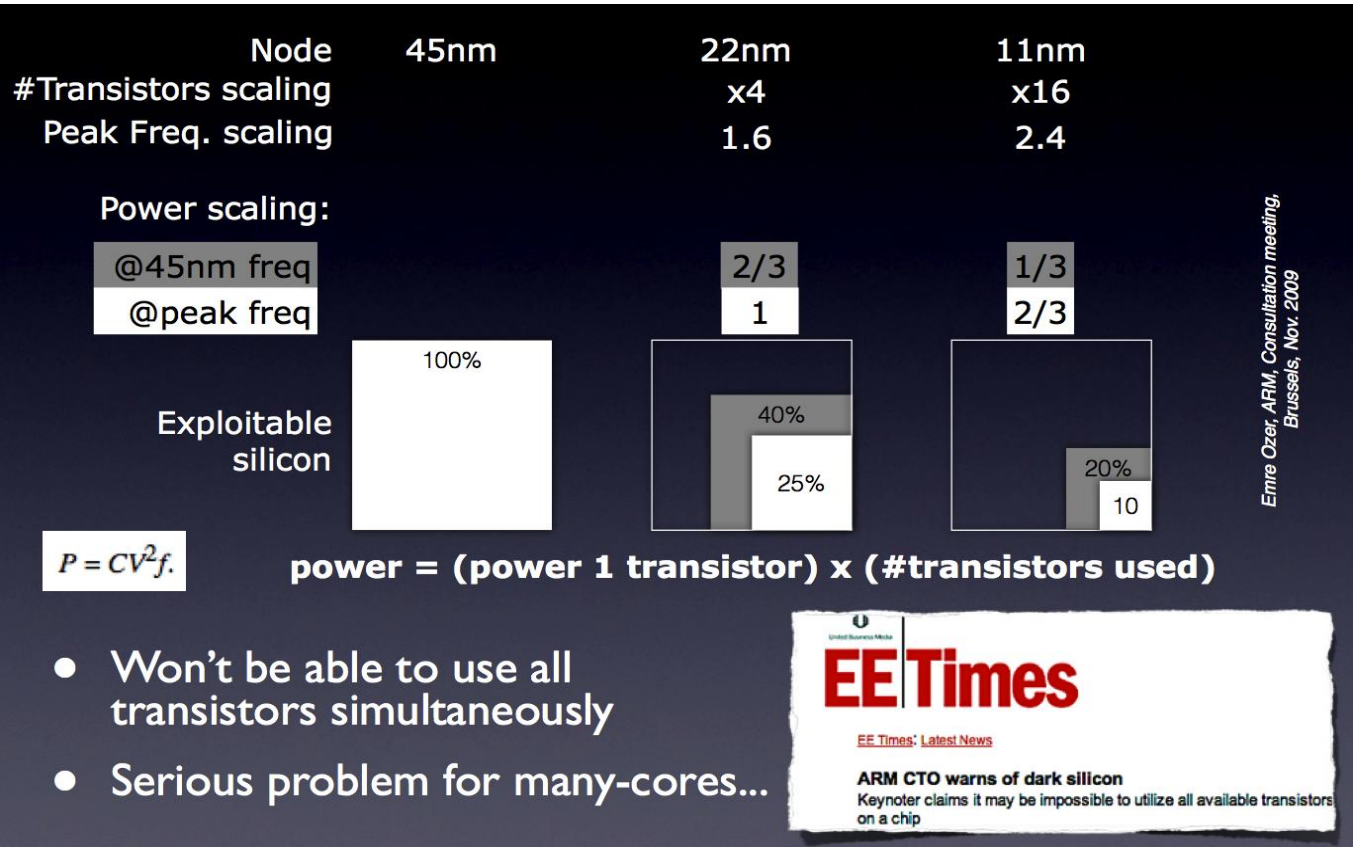
*"The Hare and the Tortoise", La Fontaine, 1668

© CEA. All rights reserved

Dark silicon challenge

« Dark Silicon »

Cf. the talk of B. Falsafi



Parameter (scale factor = a)	Classic Scaling	Current Scaling
Dimensions	1/a	1/a
Voltage	1/a	1
Current	1/a	1/a
Capacitance	1/a	>1/a
Power/Circuit	$1/a^2$	1/a
Power Density	1	a
Delay/Circuit	1/a	~1

Source: Krisztián Flautner « From niche to mainstream:can critical systems make the transition? »

© CEA. All rights reserved

Take home message 4:

- Dark is bad for silicon efficiency
- Use of the (free?) silicon estate for making different specialized cores
- (Higher development cost compensated by longer life time)

=> parallel heterogeneous many-cores

© CEA. All rights reserved

Energy cost of moving data

Energy computation / Data transfer rough calculation

- In 22 nm, swapping 1 bit in a transistor has an energy cost:

  ~ 1 attojoule (10^{-18} J)

- Moving a 1-bit data on the silicon cost:

  ~1 picojoule/mm (10^{-12} j/mm)

- Moving a data 10^9 per second (1 GHz) in silicon has a cost:

  1 pJ/mm x 10^9 s^{-1} = ~1 milliwatt/mm

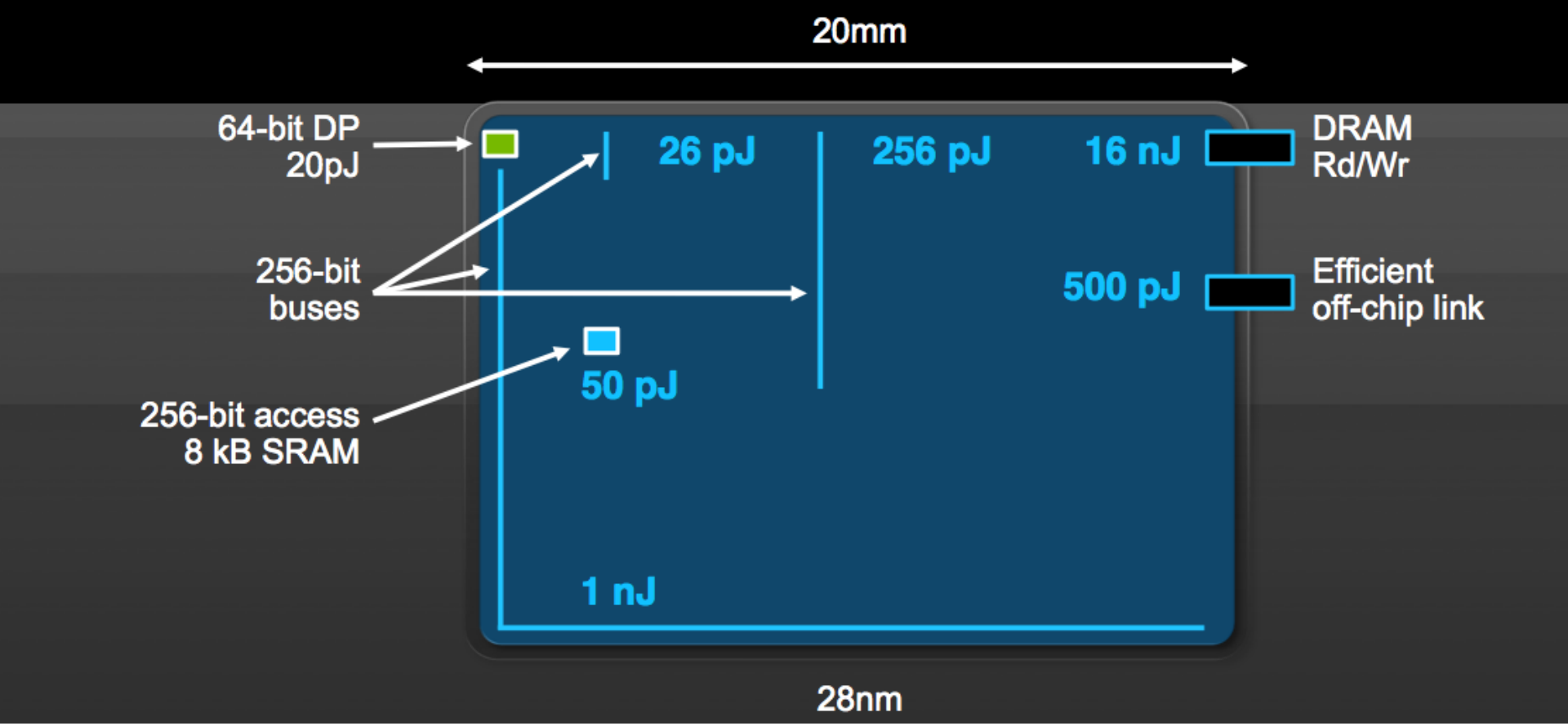
- 64 bit bus @ 1 GHz: ~64 milliwatts/mm (with 100% activity)
- For 1 cm of 64 bit bus @ 1 GHz : 0,64 W/cm

- On modern chips, there are about km of wires on chip, even with low toggle rate, this lead to several Watt/cm^2

© CEA. All rights reserved

The High Cost of Data Movement

Fetching operands costs more than computing on them



Source: Bill Dally, « To ExaScale and Beyond »
www.nvidia.com/content/PDF/sc_2010/theater/Dally_SC10.pdf

© CEA. All rights reserved

Data transfers are the main cause

" The next problem will be in the computer's memory subsystem, which will guzzle more than half of a computer's power by 2008. Computer memory subsystems are pigs"

Greg Papadopoulos, CTO, Sun

© CEA. All rights reserved

Removing the memory hierarchy?

- What if we have small, fast, low power, persistent storage cell?
- This will drive re-thinking the complete memory hierarchy and system architecture
- There are new technologies that have that potential:
  - MRAM, RRAM, Memristors, ….

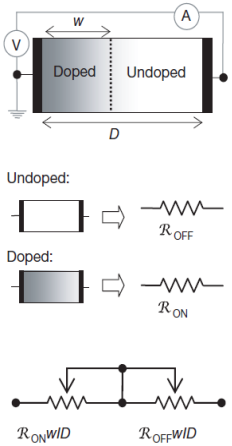
© CEA. All rights reserved

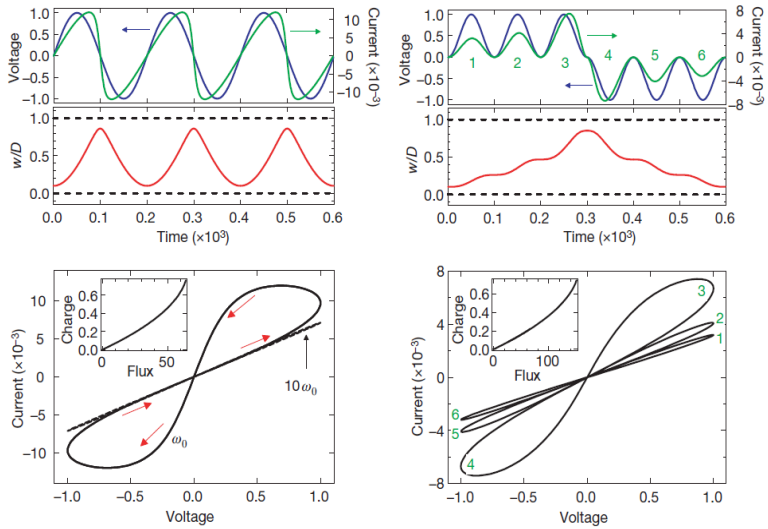
Memristive Devices Principle

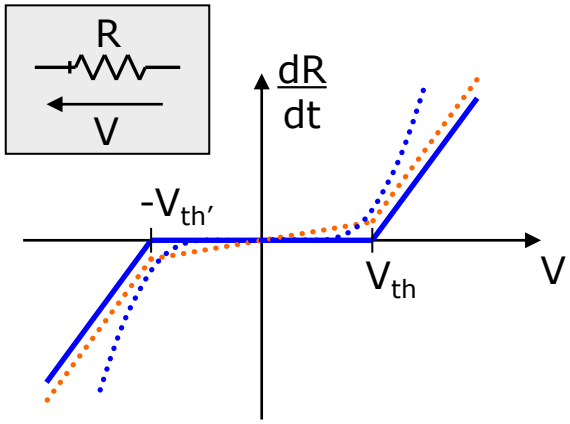
- Important steps
  - First theoretical study[1]

$$v = R(x,i).i \qquad \frac{dx}{dt} = f(x,i)$$

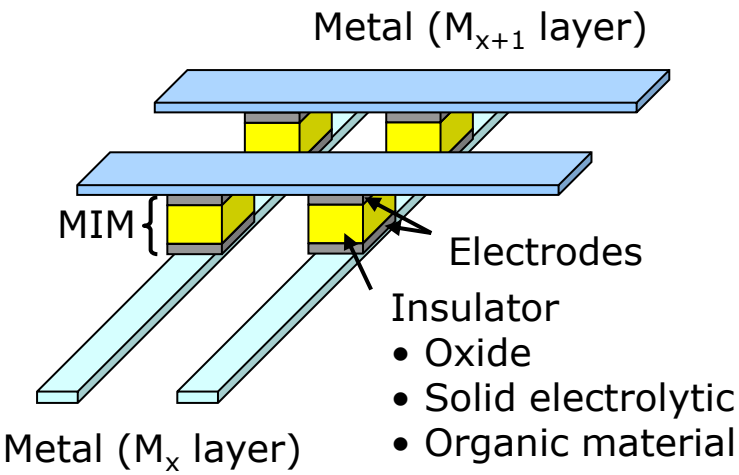
  - First link between a physical device and the theory[2]

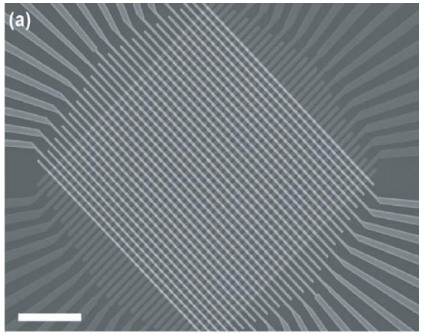






Nonlinear characteristic





Crossbar
(University of Michigan)

Source: CEA, C. Gamrat

[1] L. Chua and S. Kang, *Proceedings of the IEEE*, 1976
[2] D. Strukov et al., *Nature*, 2008

© CEA. All rights reserved

Take home message 5:

- Main power cost is in moving data
- Solution: eliminate data movement!
  - Flatten memory hierarchy
  - Follow evolution of new memory devices
  - In-memory processing
- Co-location of computing and storage
- Replace HDD with new NVRAMS?
- …

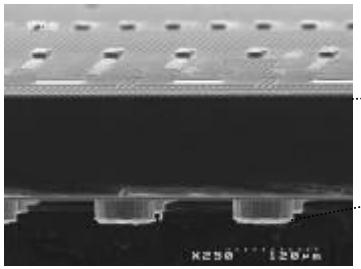
© CEA. All rights reserved

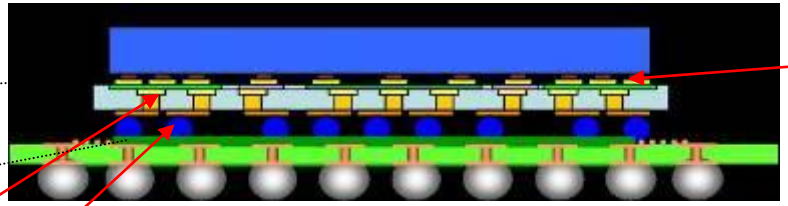
Further decreasing communication cost: 3D stacking, silicon photonics

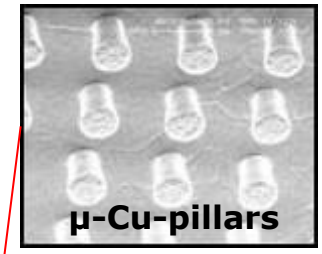
3D stacking: STMicroelectronics & CEA demonstrator

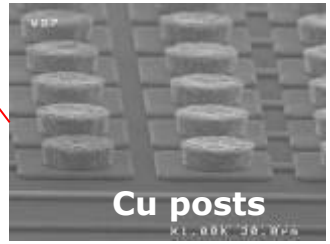
- ☑ **whole chain from design up to packaging exercised**
- ☑ **Functionality demonstrated on 1st samples**

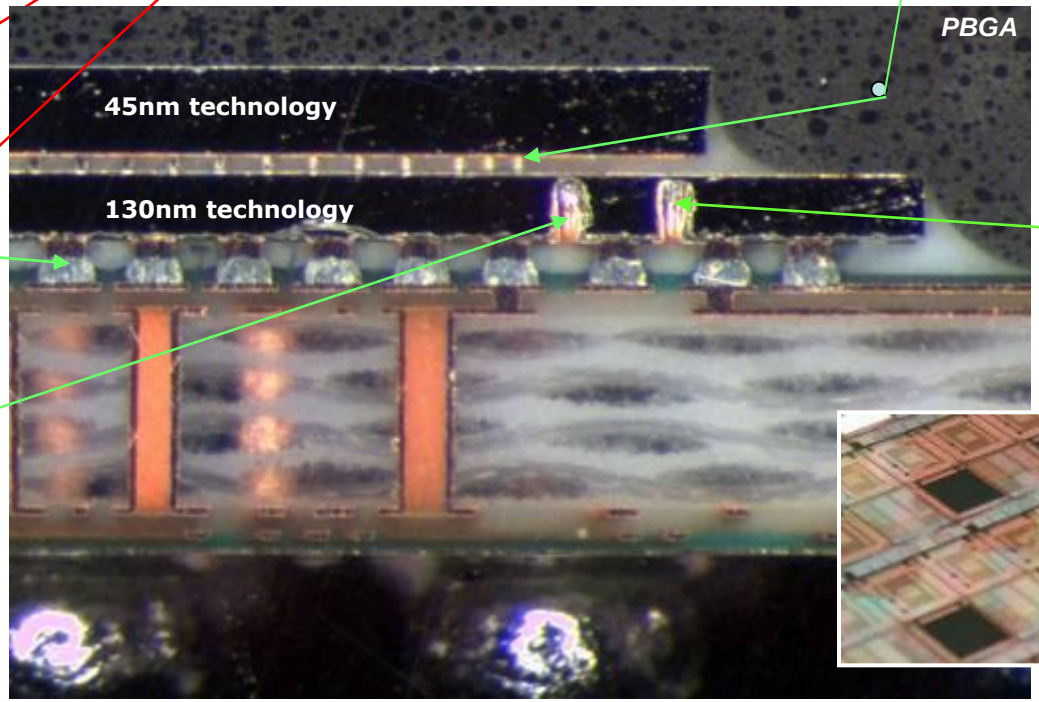
1056 F2F interconnects ;
588 TSV's, 482 bumps ;
864 lead free balls, 1.0 mm pitch

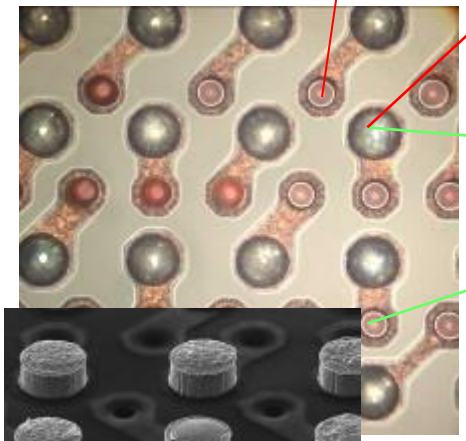




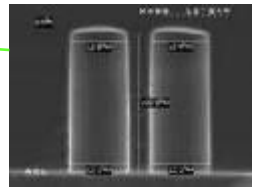




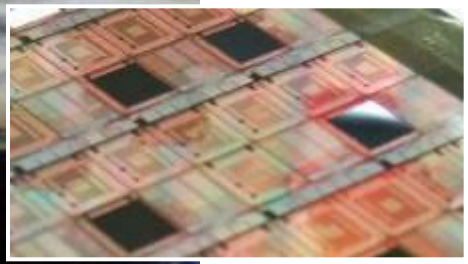




TSVs and Copper-pillars (back side)



TSV's



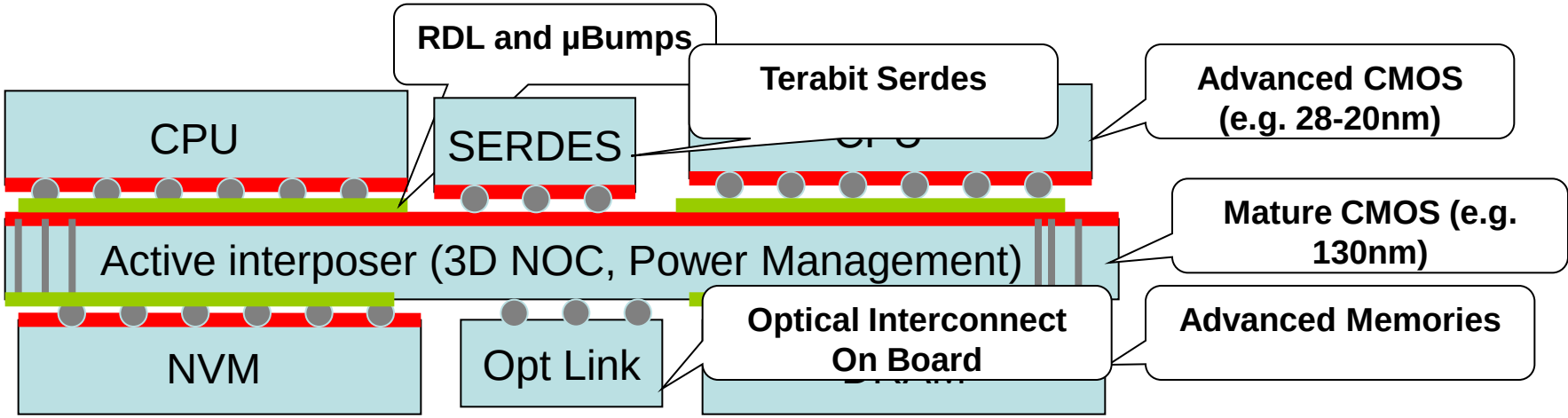
Wafer w. stacked parts

Photos: STMicroelectronics & CEA-LETI

© CEA. All rights reserved

2D and ½ integration: use an interposer

- **Active Interposeur with flipchip on 2 sides**
  - **Integration of several dies CPU/advanced memories**
  - **Innovative architectures: new memory hierarchies (NVM, DRAM)**
  - **3D interconnect with active power management on large size interposer (6cm² today, request of 100cm² on demand)**
  - **Possible use of optical interconnect between dies**



Chip: Multicore CPU: 64 cores, 16 W, 50mm² 20nm.
3D Module: 6 CPU chips , 384 cores, 4cm² 130/65 nm, 700GFLOP, 100W

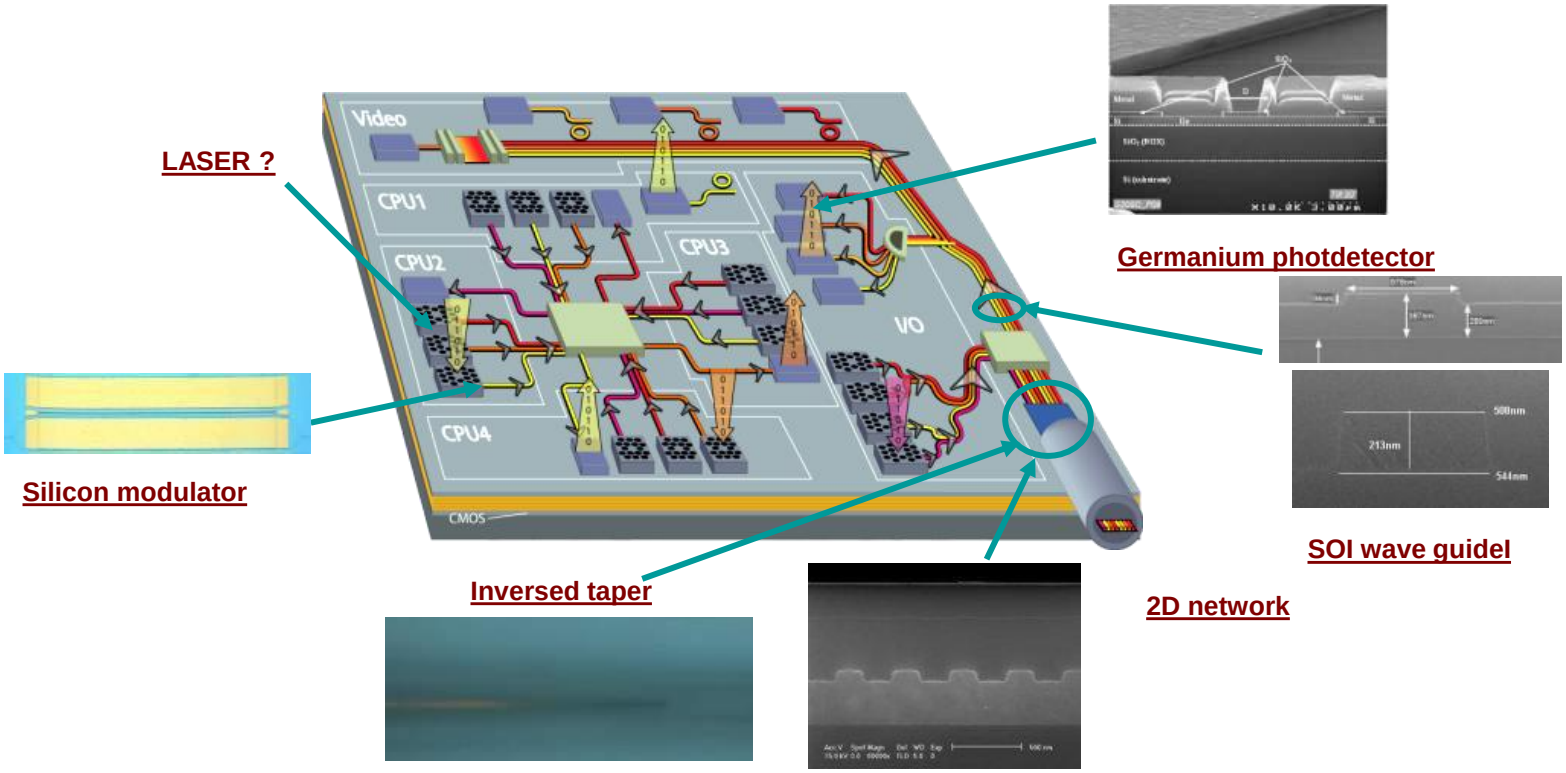
Source: CEA, Ahmed Jerraya

© CEA. All rights reserved

Optical interconnect

CMOS photonic is the integration of a photonic layer with an electronic circuit.

- **Advantages of CMOS photonic are:**
  - Use of standard tools and foundry, wafer scale co-integration
  - **Lower energy** (~100 fJ/bit), (wire: ~1 pJ/mm)
  - High bandwidth (10 Gbps), Low latency (~10 ps/mm)



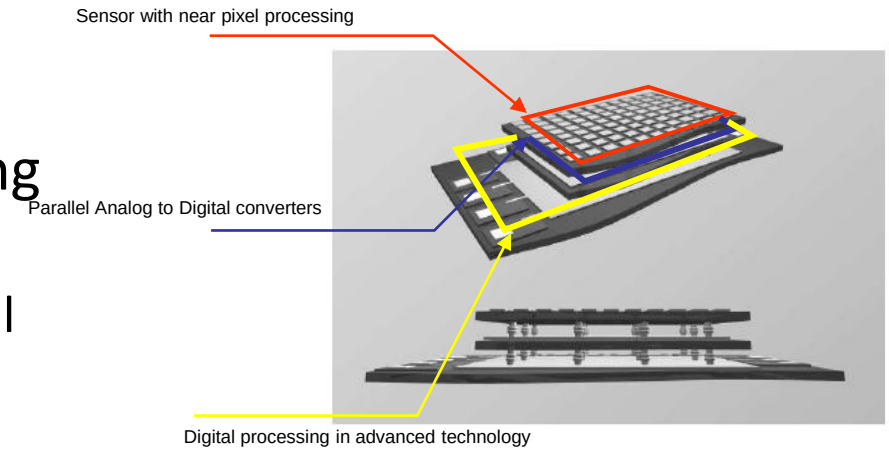
Source: CEA, Ahmed Jerraya

© CEA. All rights reserved

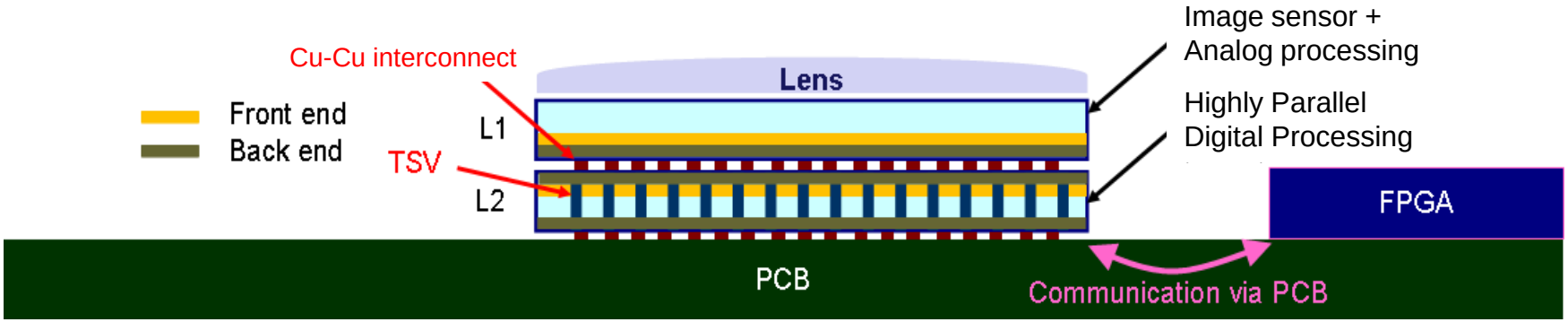
Example of architectures using 3D stacking

Smart retina extracting the information content in a scene, without necessarily reconstructing the video stream

- Ultra low power vs. conventional systems
- Massively parallel processing closed to the sensor
- High level information output (few bytes)



Intelligent Retina in 3D stacking (Near Pixel processing)



© CEA. All rights reserved

Take home message 6:

- Use the 3rd dimension!
- Communicating at speed of light
- Also at board level
  - Silicon Photonics Link aka Light Peak

© CEA. All rights reserved

Energy in an electronic device

Naïve calculation: active power

- For each transition 0 -> 1 et 1-> 0 coded between 0 and V
- Stored and transferred energy: $U = 1/2\ C\ V^2 + 1/2\ C\ (-V)^2 = C\ V^2$
- On 1 –> 0 transition , this energy is not restituted (**irreversible computation**) (Resistance of transmission line R connecting source and capacitance C)
- Dissipated energy $W = 1/2\ R\ Imax^2\ \tau + 1/2\ R\ (-Imax)^2\ \tau = R\ V^2\ (RC)\ (C/RC)^2 = C\ V^2$ because $Imax \sim C\ dV/dt = CV/\tau$ and $\tau = R\ C$
- Therefore dissipated power if running at f: $P = W\ .\ f = C\ V^2\ f$

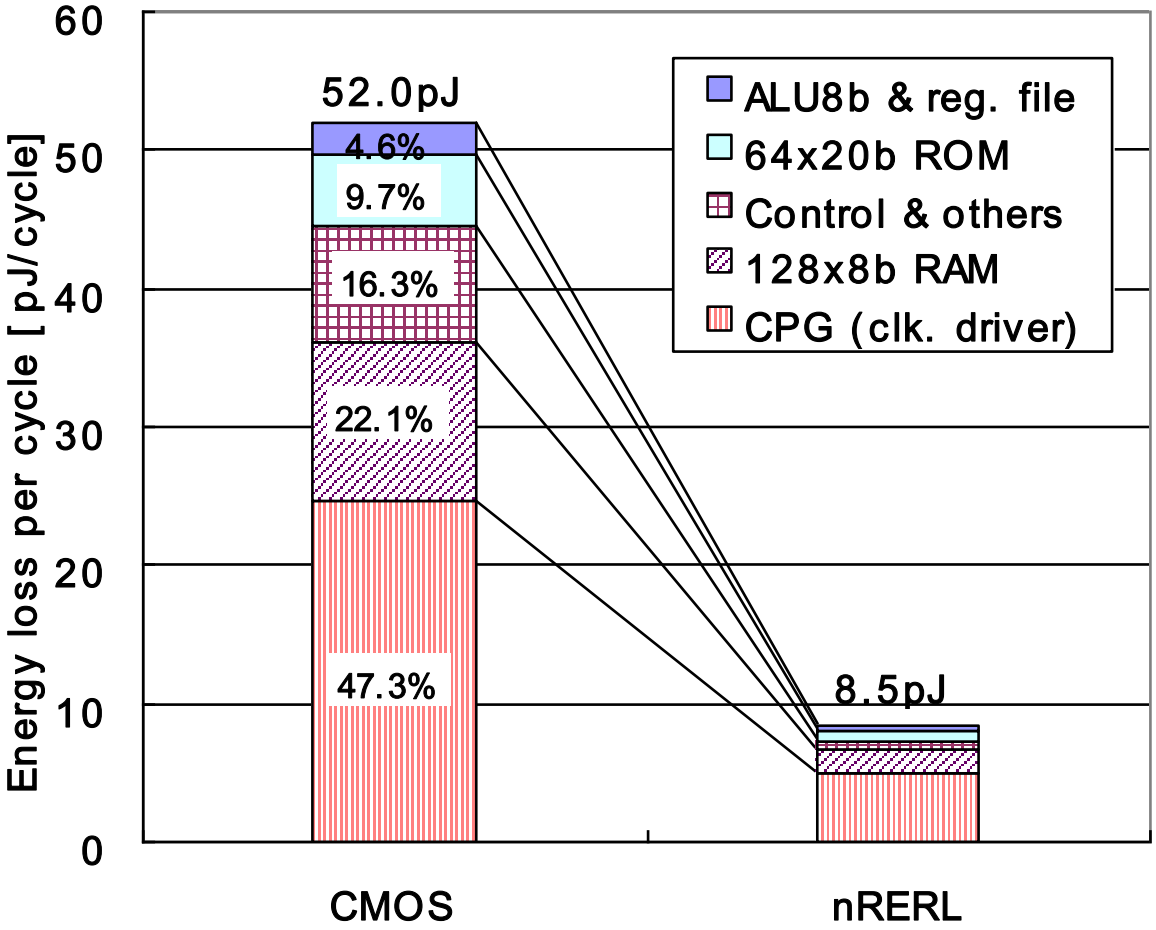
- Always more performances (Teraflops - 10^{12} - Petaflops, Exaflops, ….)

  =>Increase of the number of transistors (n) per die
- Stagnation of increase of clock frequency (f) -> use of parallelism
- Not anymore decrease of power supply (V) -> *"Dissipation wall"*

$$P = n.C.V^2.f$$

© CEA. All rights reserved

Examples using energy recovery logic



Source: S. Kim and S. Chae « Implementation of a Simple 8-bit Microprocessor with Reversible Energy Recovery Logic » RC'05: The First Int'l. Workshop on Reversible Computing

© CEA. All rights reserved

(von Neumann?)-Landauer (VNL) Bound

A rigorous result first stated clearly by Rolf Landauer, IBM, 1961
(von Neumann had suggested something similar in 1949 but did not publish details)

- Invertibility implies physical information can't be destroyed!
  - Only reversibly (i.e., mathematically invertibly) transformed!
- When we lose or discard a bit's worth of logical information,
  - e.g., by erasing or destructively overwriting a bit storage location…

  the 'lost' information must actually remain in existence,
- If the logical bit was <u>originally</u> known (then destroyed), entropy has increased in this process by $\Delta S = 1$ bit $= k \ln 2$.
  - The energy in the heat reservoir must be increased by an amount

$$\Delta S \cdot T_{env} = kT_{env} \ln 2 \quad \approx 3 \text{ zJ} = 3\ 10^{-6} \text{ fJ } @T_{room} = 300 \text{ K}$$

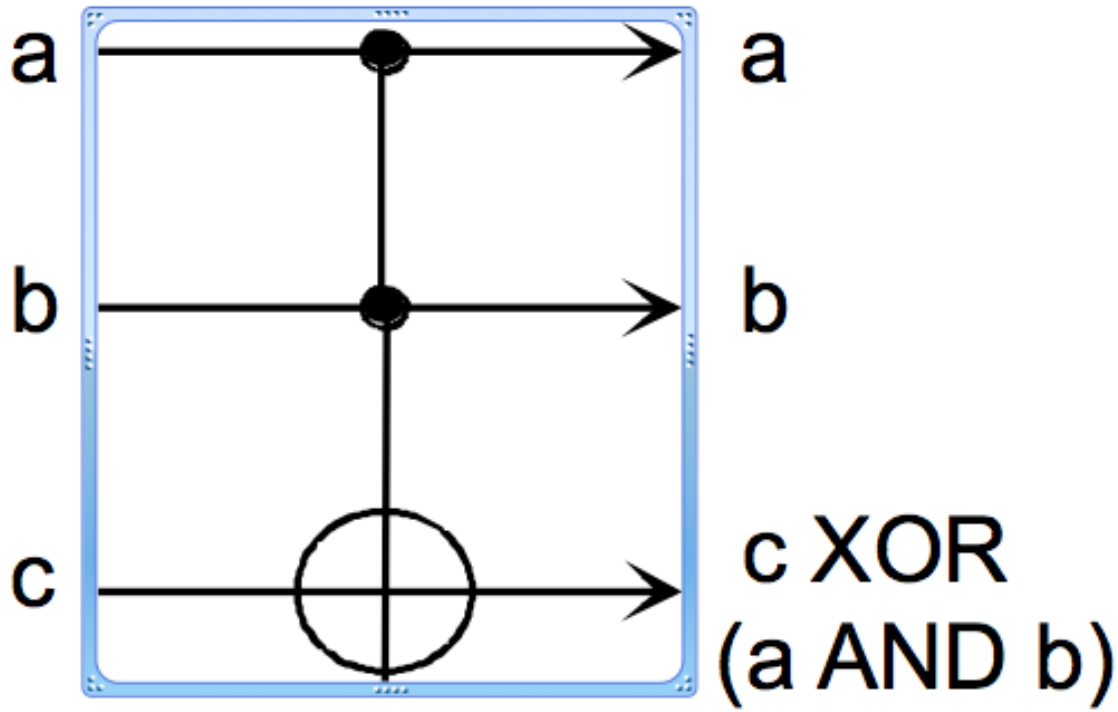
- Conventional computer logic uses irreversible gates, but these can be simulated by reversible gates: e.g. Toffoli gate is universal.

Note: $K = 1.380\ 6488 \times 10^{-23}$ JK^{-1}, 300 K Ln 2 = $2.87\ 10^{-21}$ J

Source: M. Frank, "Introduction to Reversible Computing"

© CEA. All rights reserved

Reversible logic: Toffoli gate



Toffoli gate

© CEA. All rights reserved

Naïve calculation: the rest

- Leakage current (static)
  - Increase with smaller technology nodes
  - Increase with lower V
  - Increase with temperature (positive feedback ☹)

Power: $P_{leak} = V\ I_{leak}$

- Short circuit current
  - P and N transistors simultaneously « on » when switching, creating short-circuits
  - Depends on the transistor size and command signal

Power: $P_{sc} = T_{sc}\ V\ I_{peak}$

  - **Total :** $P_{per\ unit} = C\ V^2\ f + T_{sc}\ V\ I_{peak} + V\ I_{leak} + \ldots$

© CEA. All rights reserved

Potential optimizations

$$P_{per\ unit} = C\,V^2 f + T_{sc}\,V\,I_{peak} + V\,I_{leak}$$

Average power, peak power, power density, energy-delay, ...

CIRCUITS

- Voltage scaling/islands
- Clock gating/routing
  Clock-tree distribution, half-swing clocks
- Redesigned latches/flip-flops
  pin-ordering, gate restructuring, topology restructuring, balanced delay paths, optimized bit transactions
- Redesigned memory cells
  Low-power SRAM cells, reduced bit-line swing, multi-Vt, bit line/word line isolation/segmentation
- Other optimizations
  Transistor resizing, GALS, low-power logic

ARCHITECTURE

- Voltage/freq scaling
- Gating
  Pipeline, clock, functional units, branch prediction, data path
- Split instrucn windows
- SMT thread throttling

- Bank partitioning
- Cache redesign
  Sequential, MRU, hash-rehash, column-associative, filter cache, sub-banking, divided word line, block buffers, multi-divided module, scratch
- Low-power states
- DRAM refresh-control

- Switching control
  Gray, bus-invert, address-increment
- Code compression
- Data packing/buffering

COMPILER, OS, APPLICATION

- Switching control
  Register relabeling, operand swapping, instruction scheduling
- Memory access reduce
  Locality optimizations, register allocation
- Power-mode-control

- CPU/resource schedule
- Memory/disk control
  Disk spinning, page allocation, memory mapping, memory bank control
- Networking
  Power-aware routing, proximity-based routing, balancing hop count, …
- Distributed computing
  Mobile agents placement, network-driven computation

- Fidelity control
- Dynamic data types
- Power API

Source: P. Ranganathan, “System architectures for servers and datacenters » ACACES 2011

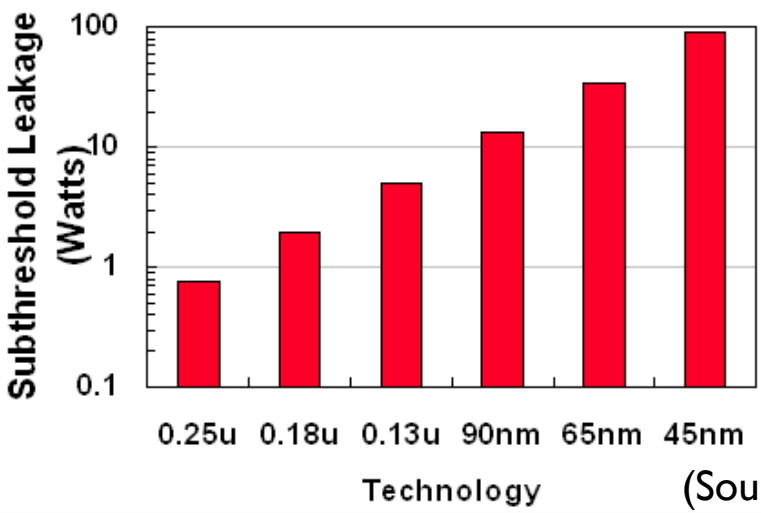
© CEA. All rights reserved

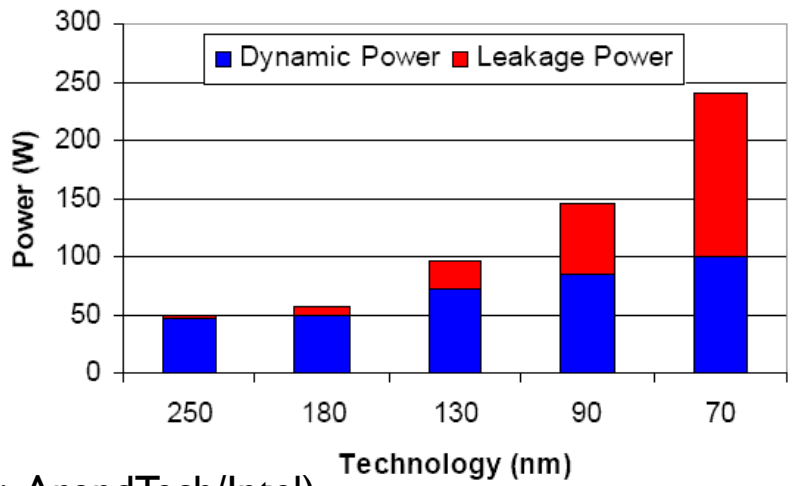
Vdd and clock frequency control

- DVS: Dynamic Voltage Scaling
- DFS: Dynamic Frequency Scaling
- Main ideas:
  1. Set the **lowest frequency possible** for performing the task
  2. Then decrease power voltage to its minimum for that frequency

$$P = n.C.V^2.f$$

- Shutdown (isolate) unused blocks (to fight leakage)



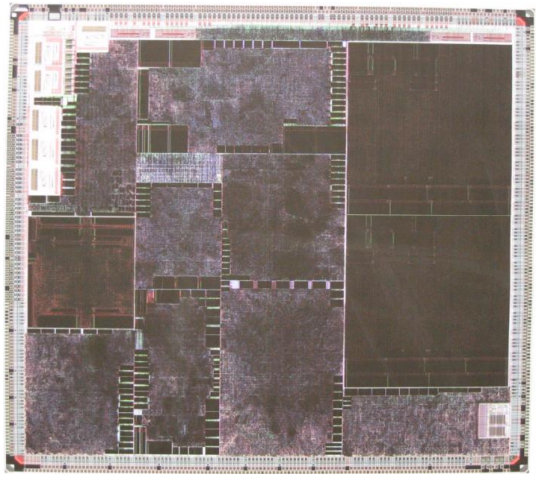


(Source: AnandTech/Intel)

© CEA. All rights reserved

This is not totally new: 11 years ago…

- HDTV SoC (HDTV compute requirement: few 100 Gop/s)
  - New HD TV with image enhancement:~ 1 Tops
- Low power (< 4W) and energy management
- Multiple programmable processing units (multi-cores)
- Heterogeneity of processing units: processors, DSP, accelerators, GPU, …
- Each IP is running at the **minimum clock frequency to perform its task** (multiple clock domains, dynamic load management)
- Everything integrated on one die (SoC – System on Chip -)

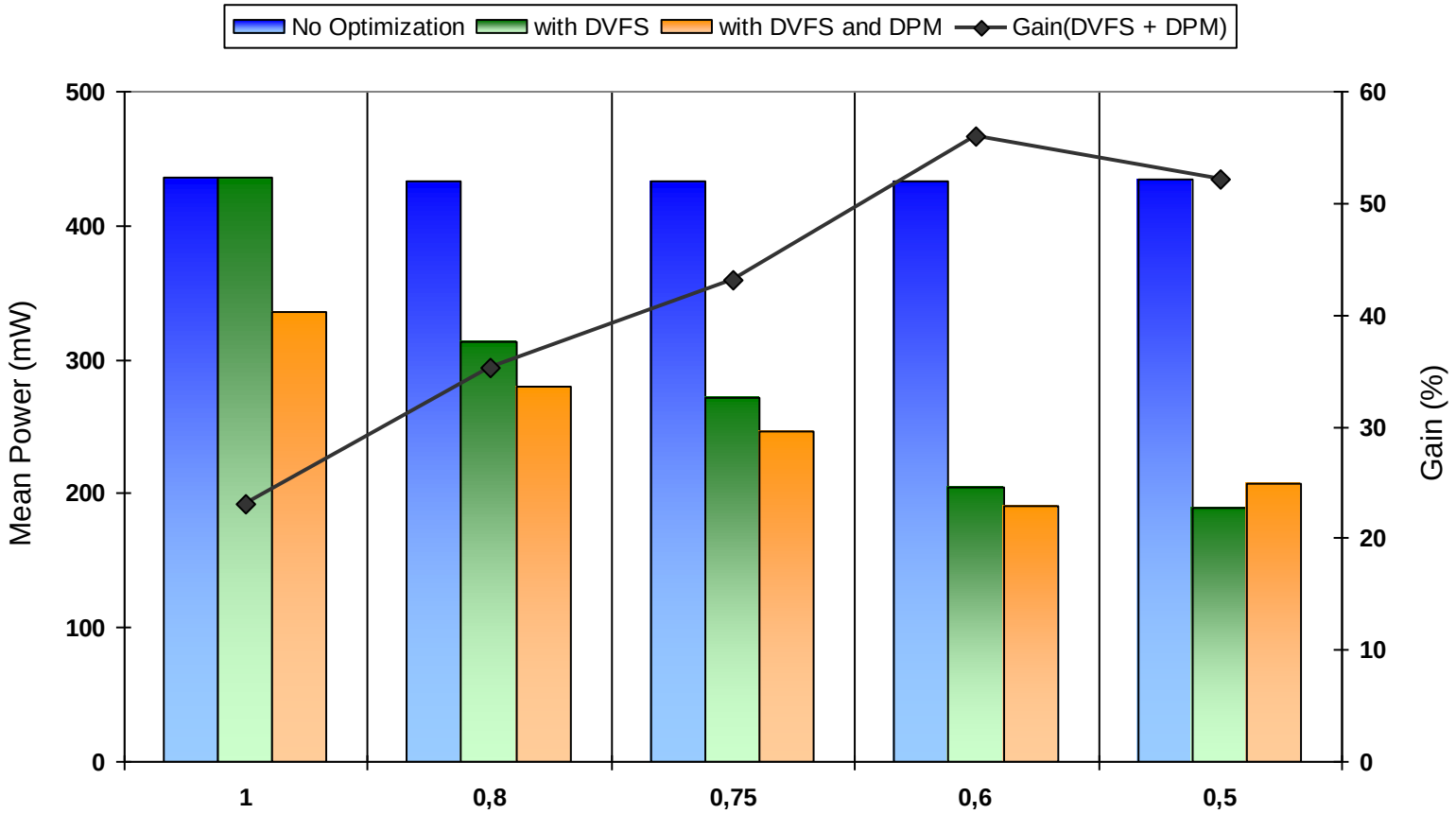


SoC for Digital TV
Philips Semiconductors – 2000 -
• 3 processor cores (Mips + TriMedia)
• > 100 Gops
• 4 W
• 60 accelerators
• 250 RAMs
• **100 clock domains**

© CEA. All rights reserved

Intelligent software to adapt power management strategies to application needs

- With timing guarantee for Real Time applications
- Best effort in dataflow applications



Source: CEA, R. David, MPSoC 2010

© CEA. All rights reserved

Take home message 7:

- Optimizing $n.C.V^2.f$ should be a whole system concern
- Fight leakage current
  - Isolate non running circuits
- Design with energy recovering logic, adiabatic logic, …
- Dynamic adaptation of the physical parameters of the chip (Vdd, Vbias, isolation) depending on activity
- Again: “On-time” instead of “best effort”*

*E. A. Lee, "Computing Needs Time," Communication of the ACM, vol. 52, no. 5, May 2009

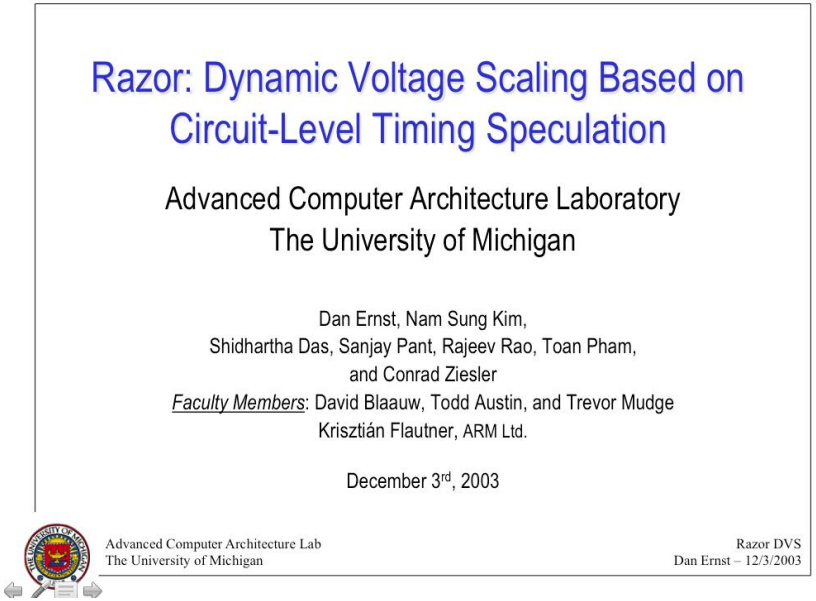
© CEA. All rights reserved

Pushing further, allowing some errors: Razor (University of Michigan, ARM)

Razor: allow few errors in decreasing V

Energy-Optimal Characteristics

Program	Optimal V_{dd}	Error Rate	% Energy Reduced	% IPC Reduced
bzip	1.1	0.31%	57.6%	0.70%
crafty	1.175	0.41%	50.5%	0.60%
eon	1.3	1.21%	34.4%	1.24%
gap	1.275	1.15%	30.1%	2.49%
gcc	1.375	1.62%	23.7%	1.47%
gzip	1.3	1.03%	35.6%	0.41%
mcf	1.175	0.67%	48.7%	0.00%
parser	1.2	0.61%	47.9%	0.29%
twolf	1.275	2.67%	30.7%	0.31%
vortex	1.3	0.53%	42.8%	0.14%
vpr	1.075	0.01%	64.2%	0.00%
Average			*42.4%*	



Source: « Razor: A Low-Power Pipeline Based on Circuit-Level Timing Speculation », Dan Ernst et all. Proceedings of the 36th International Symposium on Microarchitecture (MICRO-36'03)

© CEA. All rights reserved

Take home message 8:

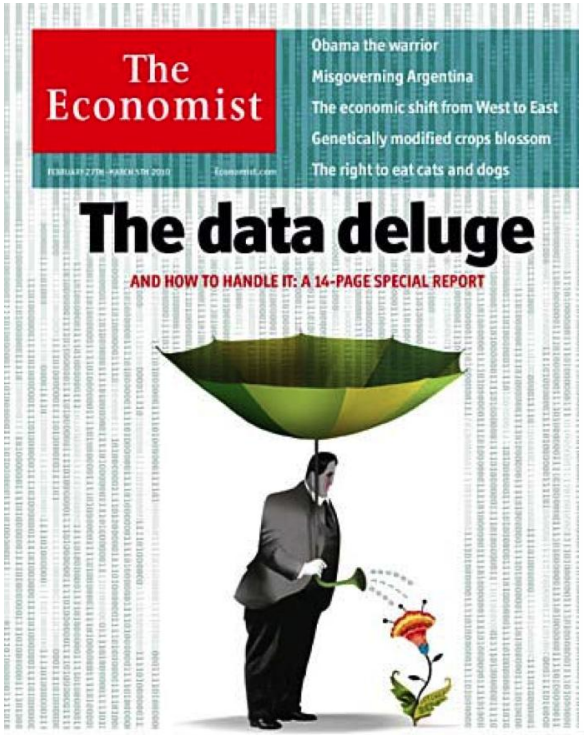
- Forget worst case design, allow errors to happen (and anyway they will happen with new technology nodes)
- Detect and correct errors at architecture level

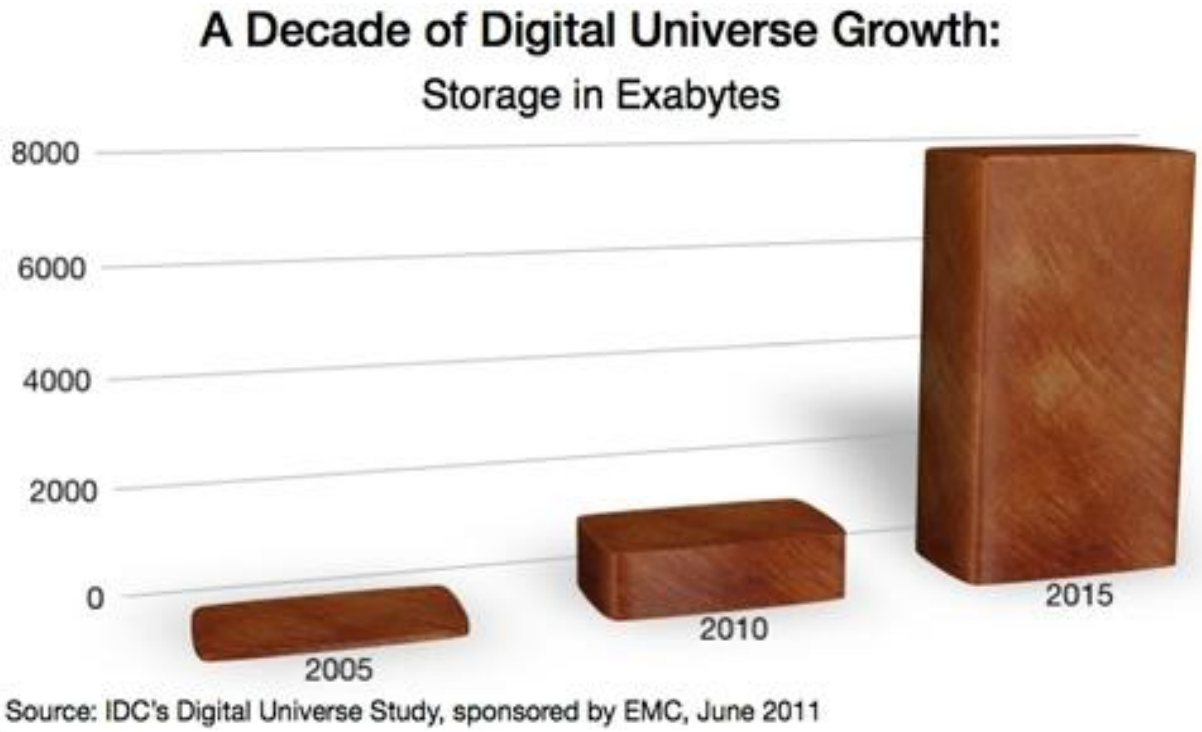
© CEA. All rights reserved

Interlude: Evolution of applications - More and more natural data processing

Data Deluge: flooding by information

Our world is creating more bits than ever: in 2011, more than 1.8 zettabytes (1.8×10^{21}) , growing by a factor of 9 in 5 years

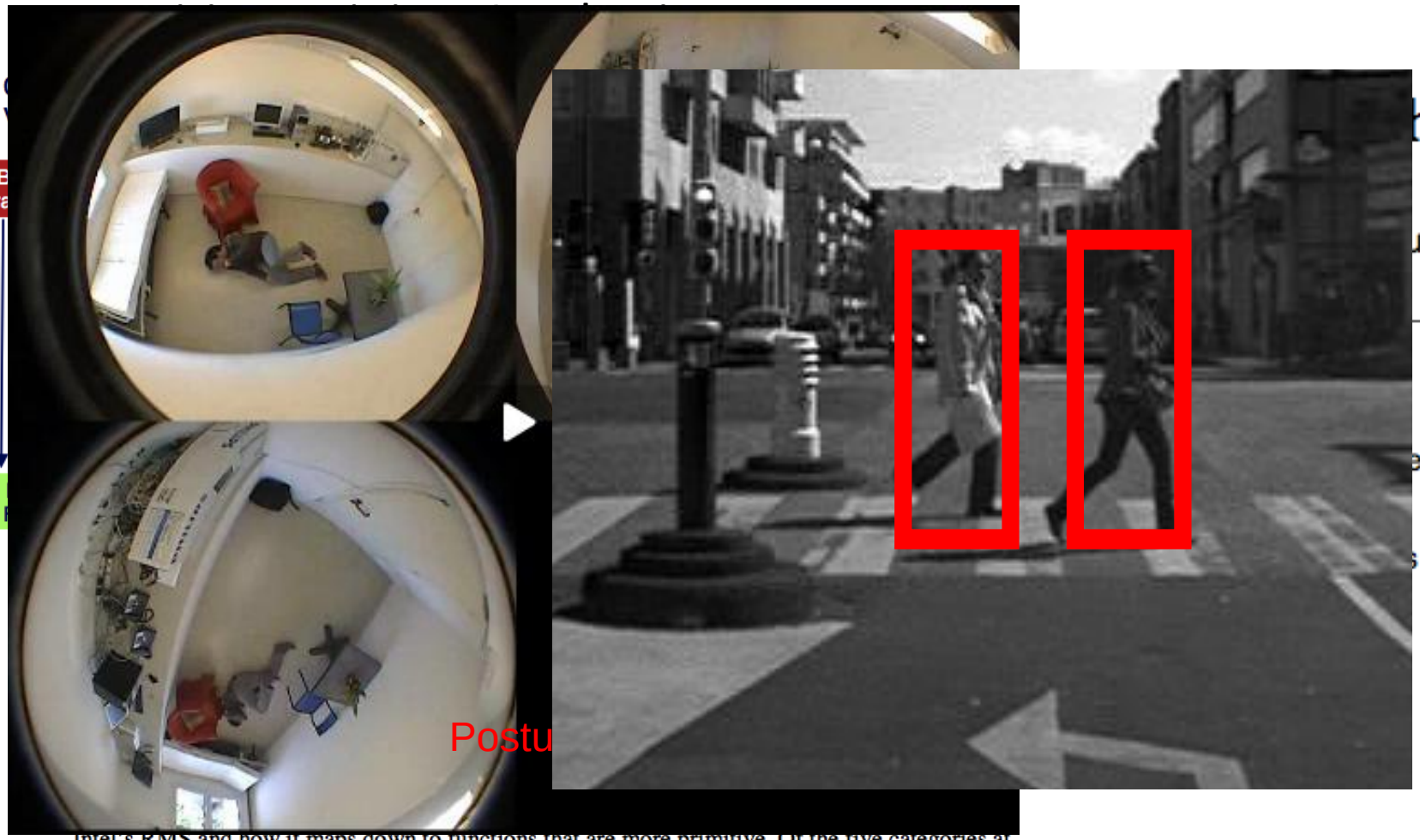




© CEA. All rights reserved

RMS and Intel vision for future applications

More and more applications are not only « number crunching » :



Intel's RMS and how it maps down to functions that are more primitive. Of the five categories at the top of the figure, Computer Vision is classified as Recognition, Data Mining is Mining, and Rendering, Physical Simulation, and Financial Analytics are Synthesis. [Chen 2006]

Source: « The Landscape of Parallel Computing Research: A View from Berkeley », Krste Asanovic et all.

© CEA. All rights reserved

Take home message 9:

- Always more and more data to process
- Data coming more and more from the "physical" world
  - Indexing of images, music,…,
  - Content analysis and understanding
  - **No "exact" processing required**

Challenge:

Data Deluge meet the Energy Wall in a Physical and Connected World

(HiPEAC Roadmap 2011)

© CEA. All rights reserved

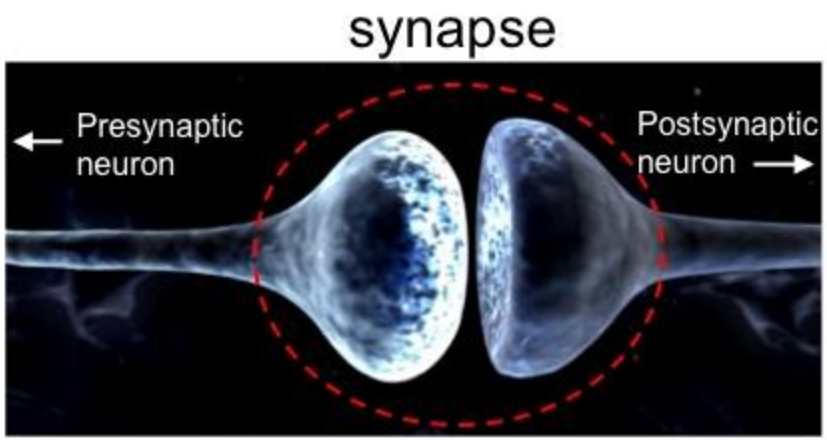
New computing devices and New computing paradigms for the “physical” data deluge

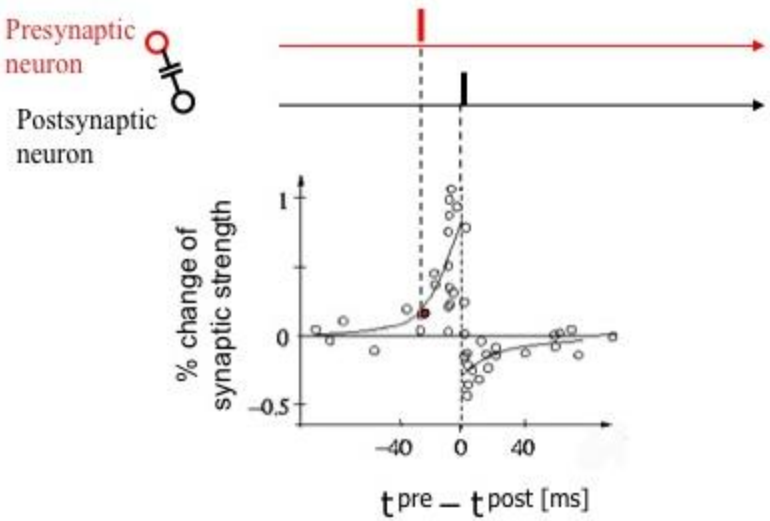
For example, take inspiration from biology

Example: STDP (Spike-Timing-Dependent Plasticity)

- Local learning rule
- Energy efficient information coding based on the time of arrival
- Can be replicated with simple physical devices
  - Ultra low power???
- Applications in pattern recognition, clustering, classification

Source: CEA, R. Heliot



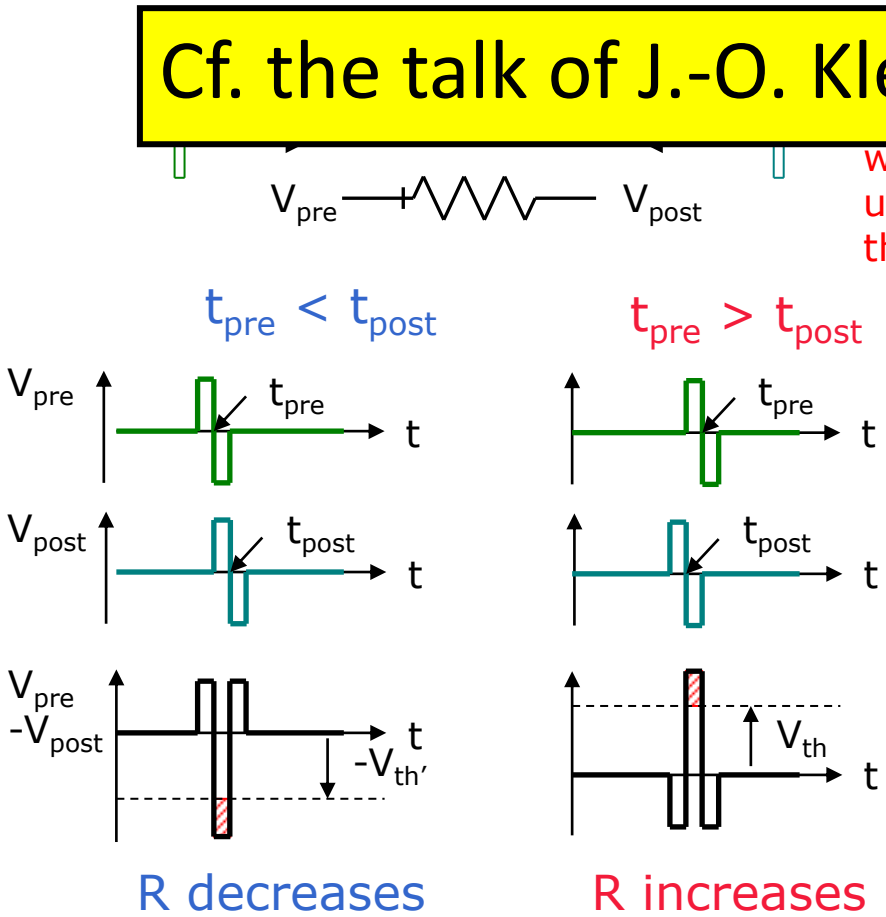


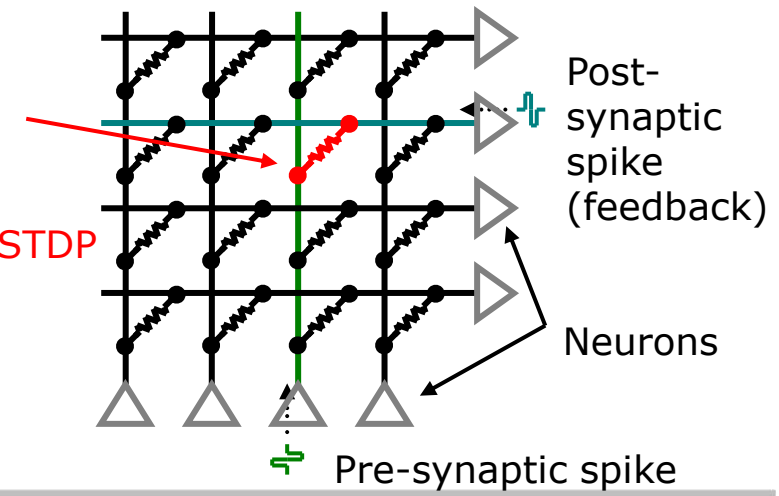
Song et. al (2000)
Pfister et. al (2006)

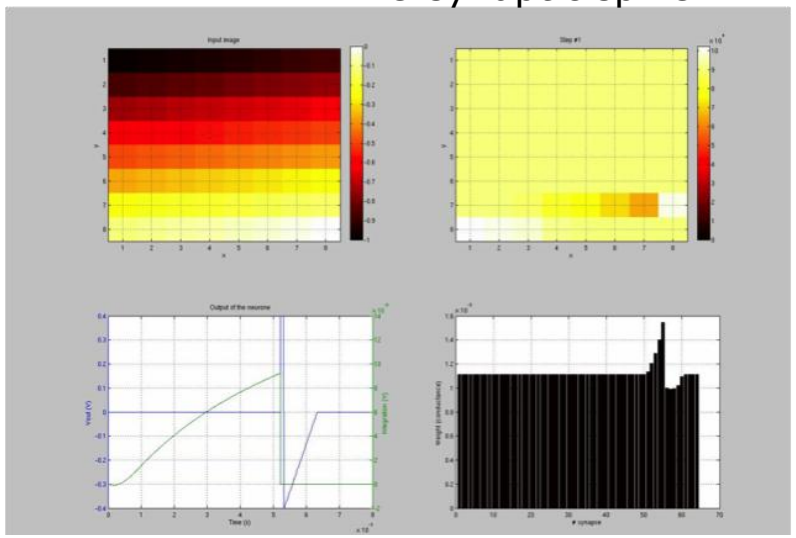
© CEA. All rights reserved

Memristive device and STDP: paving the way to unsupervised learning= *no programming!*

Cf. the talk of J.-O. Klein





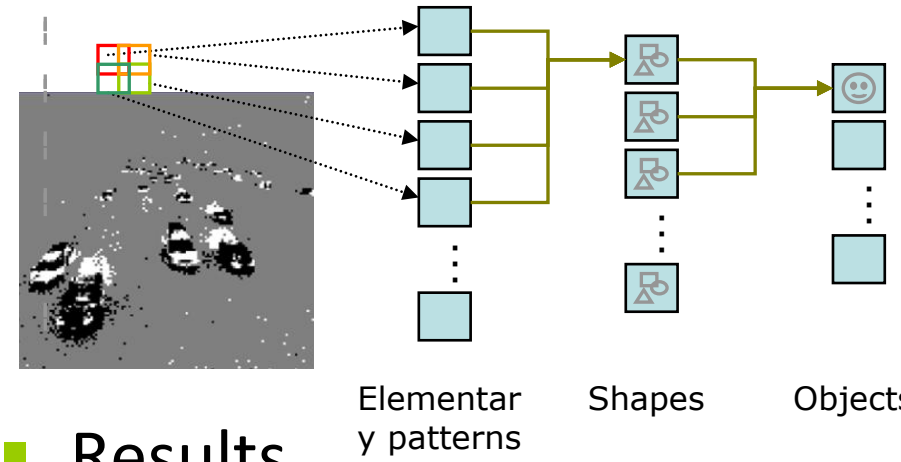


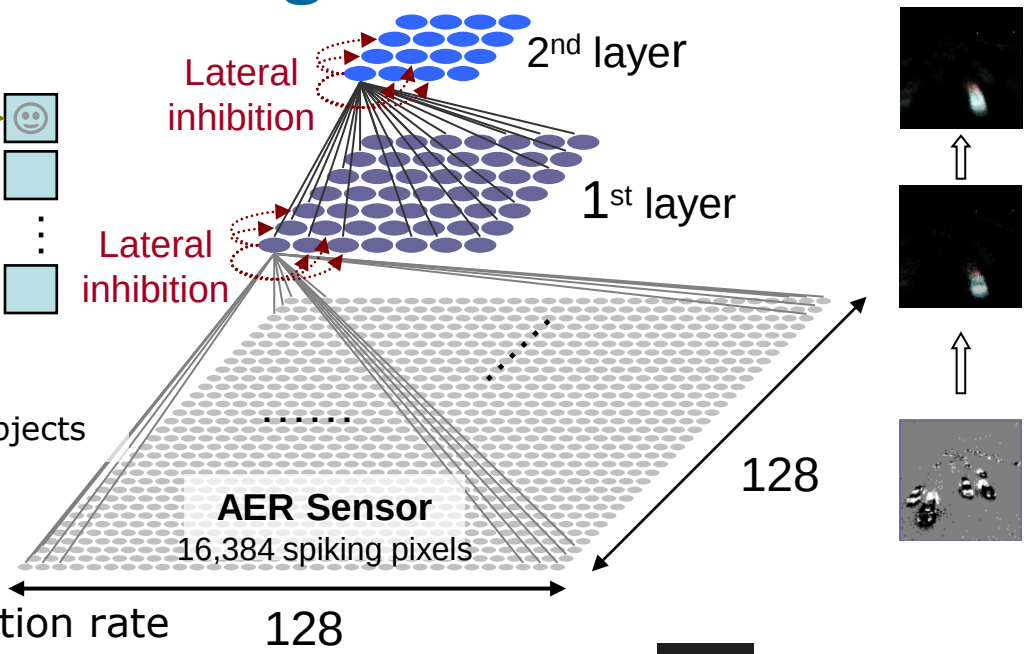
Simulation with AMS Designer (Cadence)
Functional models in VHDL-AMS

1. G. Snider, *Nanoscale Architectures*, 2008
2. B. Linares-Barranco and T. Serrano-Gotarredona, *Nature Precedings*, 2009

© CEA. All rights reserved

A Real life Vision Application using memristive devices as synapstors: counting cars

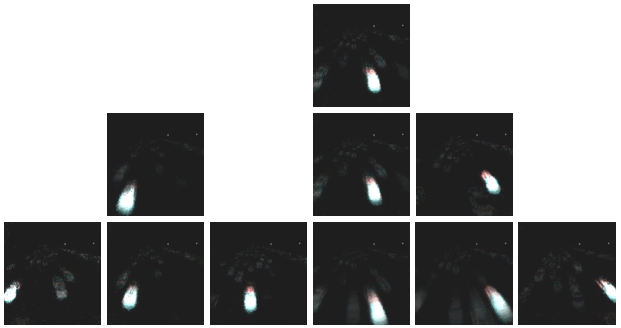




- Results
  - 98% cars detection rate
  - High noise tolerance: same detection rate with 50% added noise
  - Very **High variability immunity**: good performances with 20% synaptic variability (> 95% on learned traffic lanes)

O. Bichler, D. Querlioz, S. J. Thorpe, J.-P. Bourgoin and C. Gamrat, *"Unsupervised Features Extraction from Asynchronous Silicon Retina through Spike-Timing-Dependent Plasticity"*, International Joint Conference on Neural Networks (IJCNN) 2011 (submitted)

Source: CEA, C. Gamrat

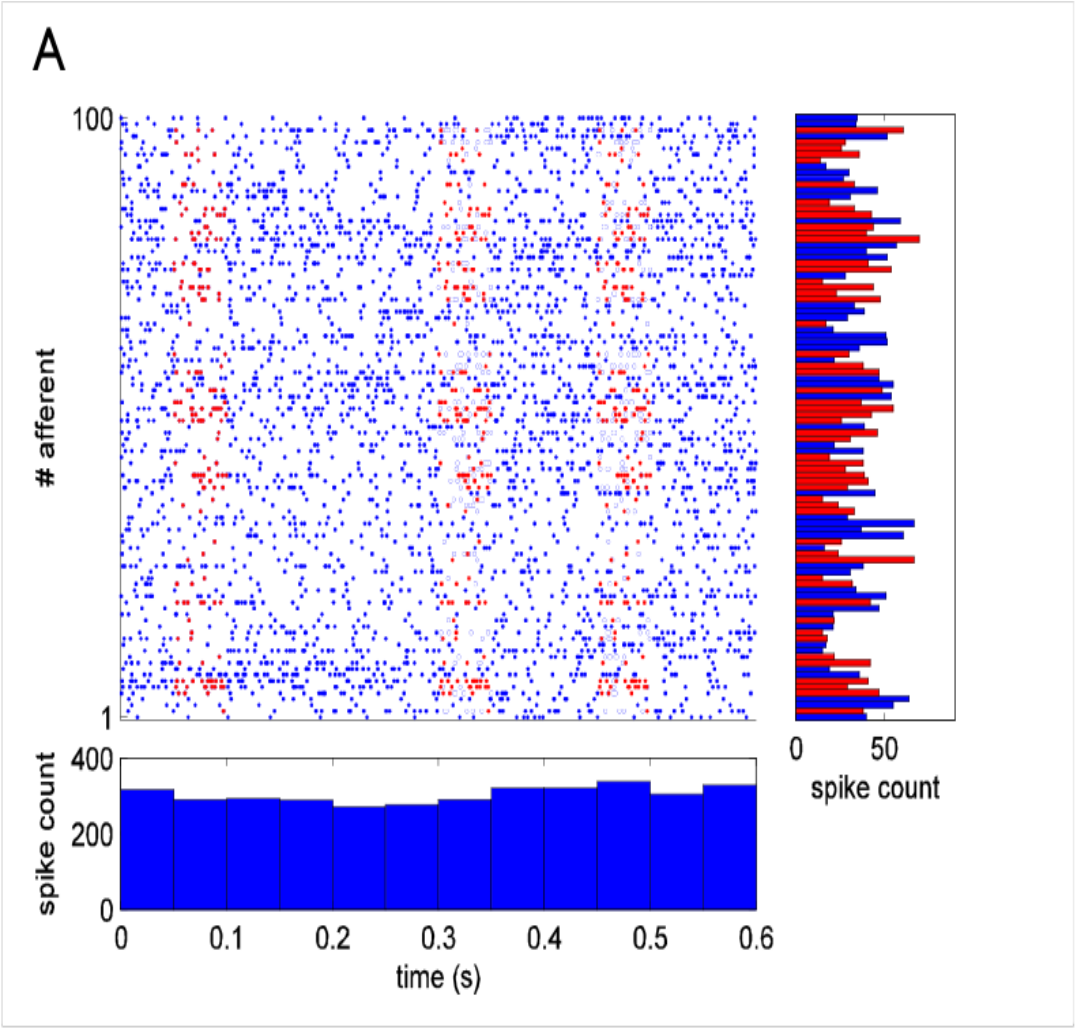


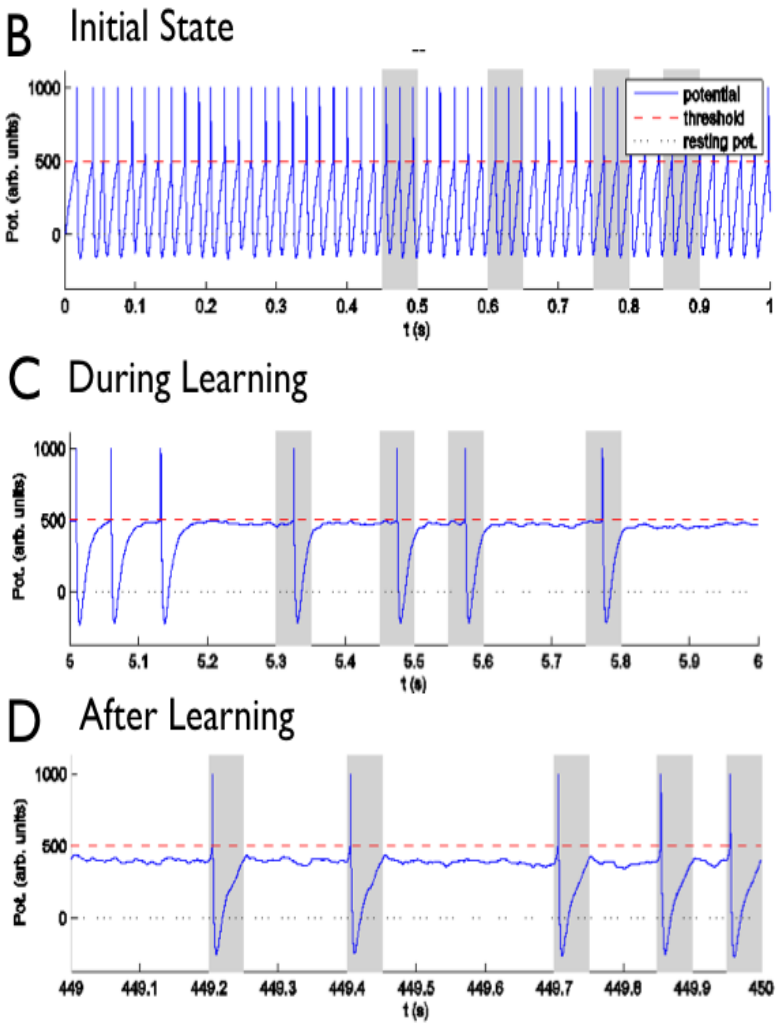
1st lane 2nd lane 3rd lane 4th lane 5th lane 6th lane

Second layer weight reconstruction

© CEA. All rights reserved

Other application :Unsupervised pattern matching in random data

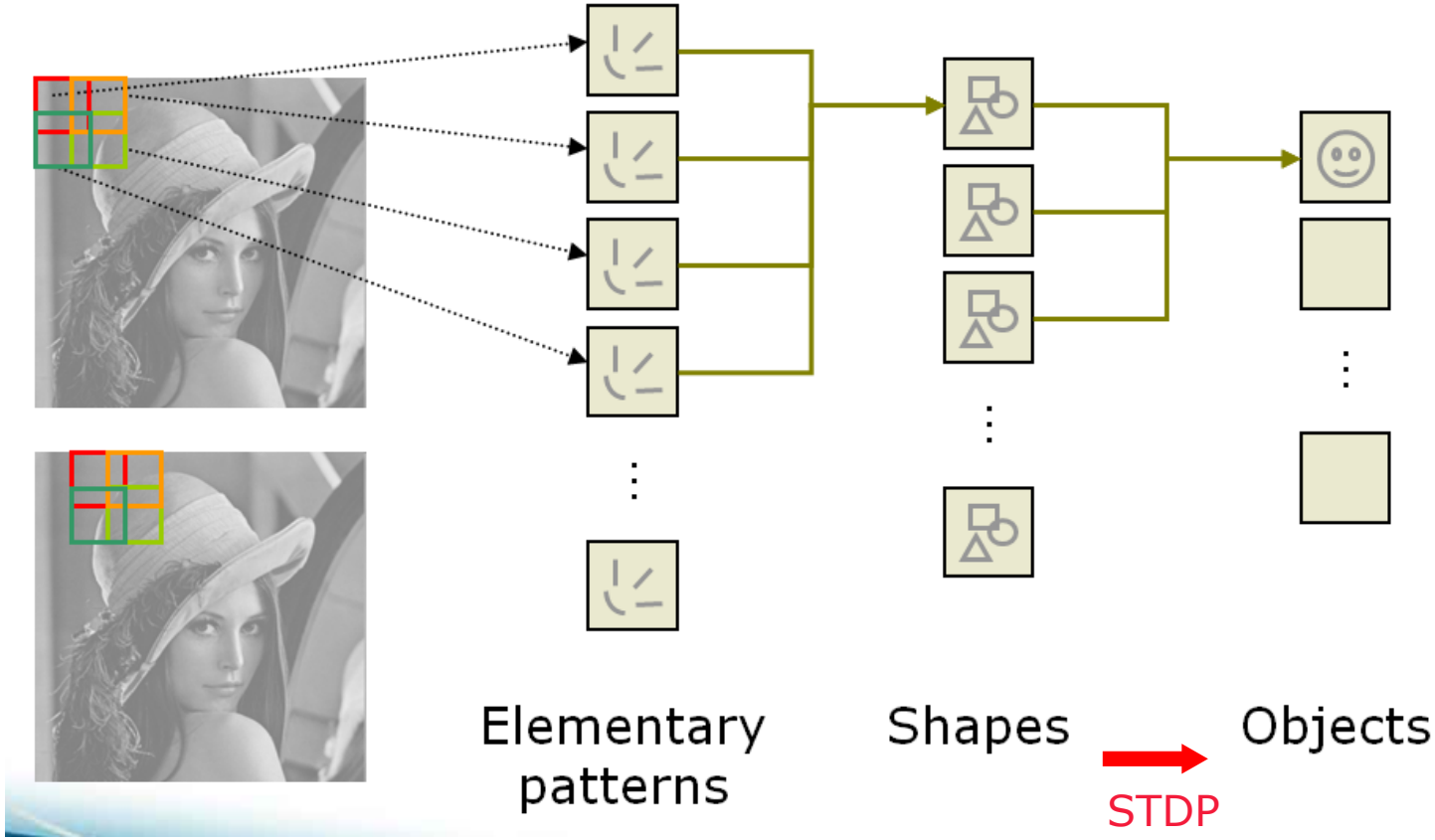




Source: Masquelier T, Guyonneau R, Thorpe SJ. 2008. « Spike timing dependent plasticity finds the start of repeating patterns in continuous spike trains". *PLoS ONE* 3:e1377

© CEA. All rights reserved

Other Application :Unsupervised pattern matching for image sensors

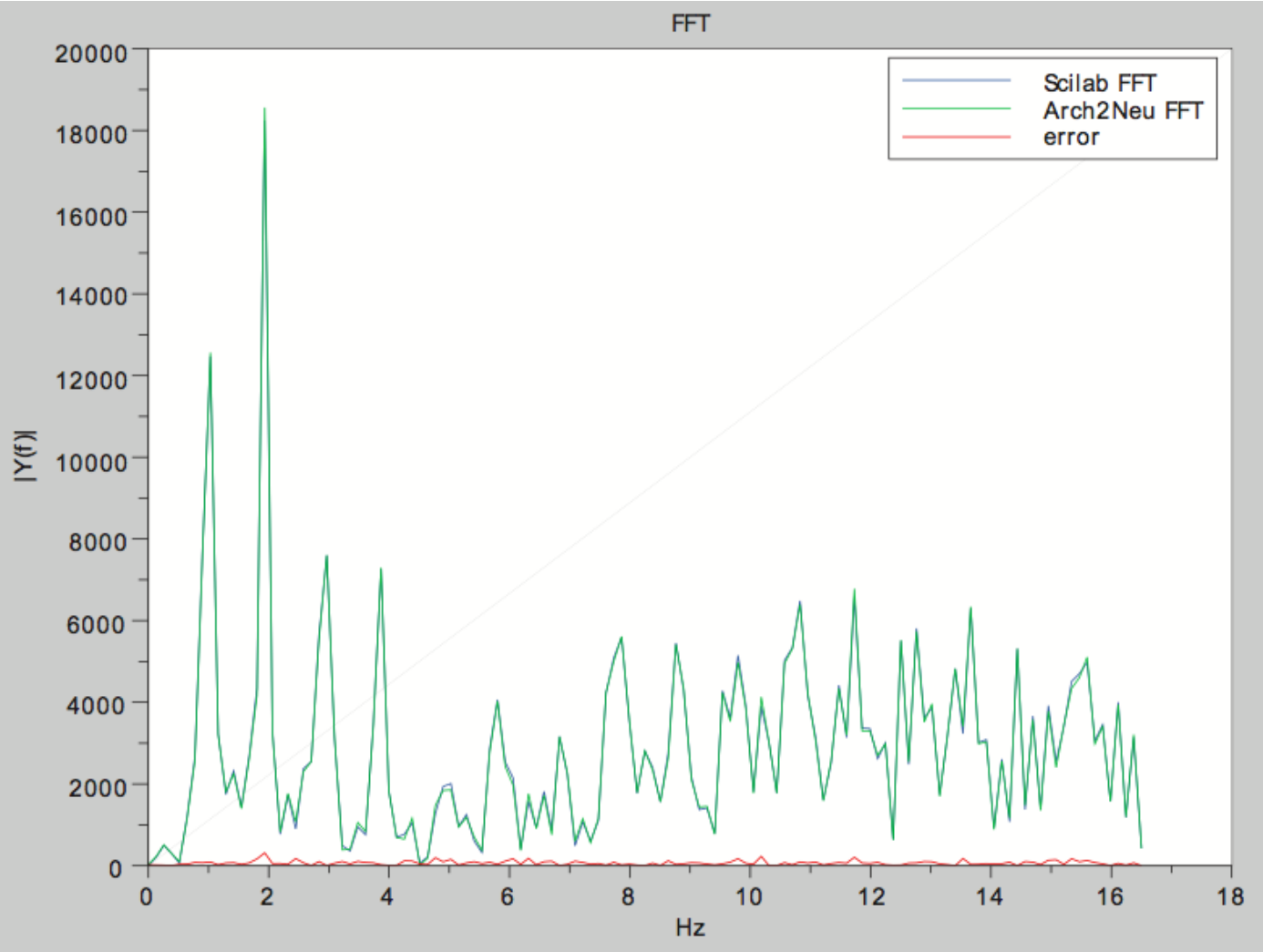


Hierarchical architecture based on memristive devices for unsupervised pattern matching in image sensors.

Source: CEA, O. BichlerO. Bichler, “Learning with Memristive Devices”, 2010 Workshop on Innovative Memory Technologies, Grenoble, June 21, 2010

© CEA. All rights reserved

Also: 256 point FFT realized with spiking neurons



Source: CEA, C. Gasnier « mentation d'une e de Fourier sur une architecture neuromorphique » Arch2Neu project

© CEA. All rights reserved

Take home message 10:

- **New devices will be small, but have large dispersion, low reliability**
- **New computing paradigms required (bio-inspired, probabilistic, stochastic, ...)**
- **New technologies will have impact on architectures**
- **Combining best of both worlds: *Hybrid Systems*, "new paradigms" and "classical"**

© CEA. All rights reserved

Conclusion

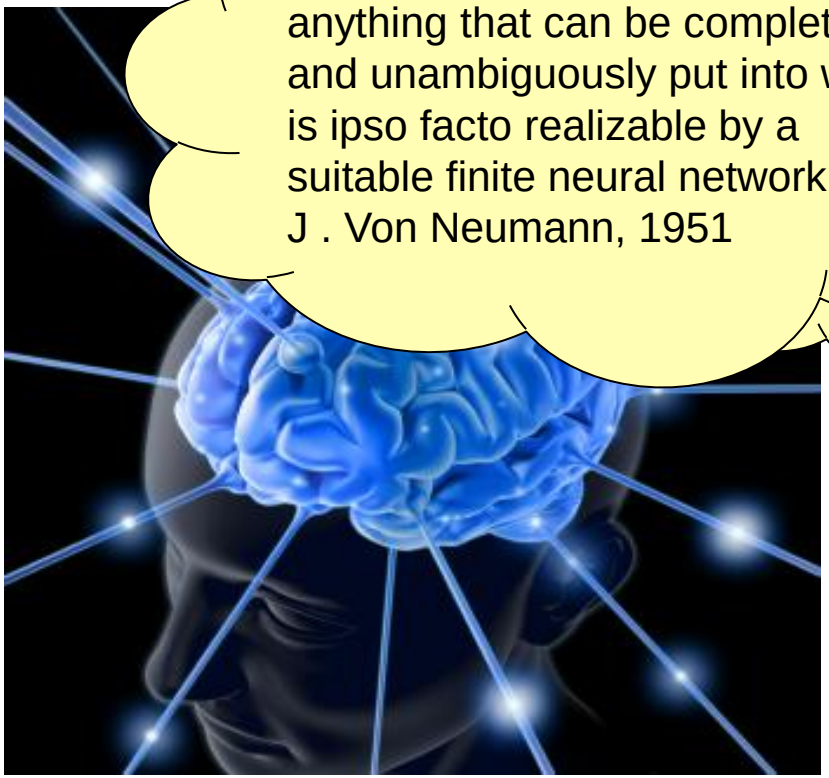
Overview of some trends…

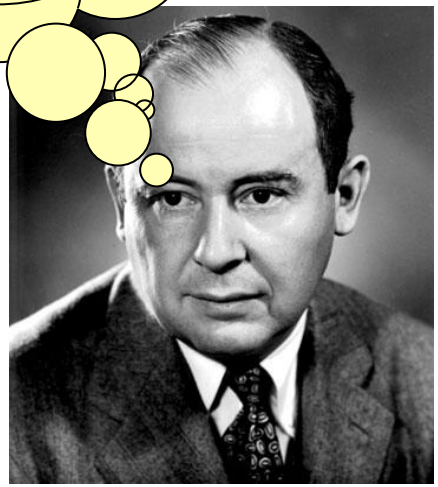
- Limit of frequency scaling -> parallelism, multi-cores
- Limit of energy density -> simple, small, efficient cores
- “Dark Silicon” -> heterogeneity of specialized cores
- New NVRAMS -> Co-locating compute and storage
- Cost of moving data -> prefer locality
  - 3D interconnect
  - Optical interconnect
- More and more “natural data processing” -> new compute paradigms, without explicit programming, in hybrid systems

© CEA. All rights reserved

This seems quit

« The McCulloch-Pitts result puts an end to this. It proves that anything that can be completely and unambiguously put into words is ipso facto realizable by a suitable finite neural network. »
J . Von Neumann, 1951





Evolution: Moore and Darwin, similar result?

Exiting new opportunities are ahead of us!

© CEA. All rights reserved

LABORATOIRE D'ÉLECTRONIQUE
ET DE TECHNOLOGIES
DE L'INFORMATION

LABORATOIRE D'INTÉGRATION
DES SYSTÈMES
ET DES TECHNOLOGIES

Thank you for listening