High-Performance and Embedded Architecture and Compilation

HiPEAC vision *2015*

Ada Europe

Marc Duranton

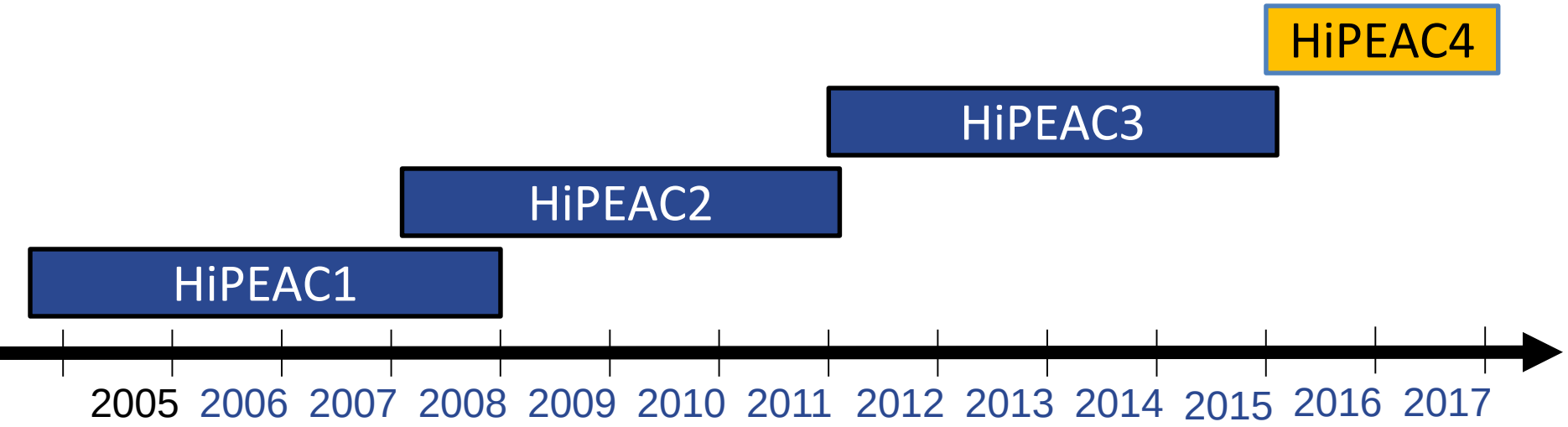
June 16th, 2016

HiPEAC

- HiPEAC was a European Network of Excellence (now a CSA) on **Hi**gh **P**erformance and **E**mbedded **A**rchitecture and **C**ompilation
- Created in 2004, **HiPEAC** gathers over 449 leading European academic and industrial computing system researchers from nearly 320 institutions in one virtual centre of excellence of 1700 researchers.

- Coordinator: Koen De Bosschere (UGent)

HiPEAC history



Grant Agreement no: 687698

449 members, 86 associated members, 379 affiliated members and 803 affiliated PhD students from 318 institutions in 39 countries.

Membership is free of charge.

hipeac.net/members/stats/map

To join, simply email

membership@hipeac.net

HiPEAC mission:

HiPEAC encourages computing innovation in Europe by providing:

- Support for projects (job portal, communication)
- the semi-annual computing systems week,
- The ACACES summer school,

2016 Conference

January 18-20, 2016, Prague, Czech Republic

The 11th HiPEAC conference took place in Prague, Czech Republic, January 18-20, 2016 and gathers more than 650 people

HiPEAC conference

2017 Conference
January 23-25, 2017, Stockholm, Sweden

kdbosche

Home | Call for Papers | Workshops and Tutorials | Venue and travel | Sponsorship

The HiPEAC conference is the premier European forum for experts in computer architecture, programming models, compilers and operating systems for embedded and general-purpose systems.

Computing System Week

HiPEAC

European Network on High Performance and Embedded Architecture and Compilation

kdbosche

Activities | Mobility | Research | Jobs | Industry | The Network | Publications

CSW Porto, April 20-22, 2016

About | Programme | Venues & Accommodation | Attendees

Highly skilled candidates for specialist roles

“If you’re looking for skilled PhD engineers in processor design, system architecture, compilers and tools, look in HiPEAC first, the best ones are there.”

Christian Bertin, STMicroelectronics

- Recruitment portal and events
- Jobs shared via LinkedIn and Twitter
- PhD directory
- Pool of 800+ PhD students
- Internship programme – supporting SMEs and larger businesses

HiPEAC Jobs

Active

Find your ideal computing job in Europe.
There are currently **53 open positions**!

+ ADD A NEW JO

Country

All

Career level

All

Core skills

- [ ] Architecture
- [ ] Compilation
- [ ] Cyber-physical systems
- [ ] Data centers
- [ ] Design space exploration (DSE)
- [ ] Embedded systems
- [ ] Fault tolerance

Newcastle University

PhD Funding in Approximate Computing in Newcastle University (UK) @ *Newcastle University, GB*

Feb. 26, 2016
PhD student
Architecture, Design space exploration (DSE), Embedded systems, Memory, Reconfigurable computing

ARM

R & D Engineer @ *ARM, GB*

Feb. 28, 2016
PhD student, PostDoc, Research Associate, Engineer
Architecture, Simulation

Follow us on LinkedIn

Université de Valenciennes et du Hainaut-Cambrésis

Reliability in Embedded Systems: Simulation and Modeling @ *University of Valenciennes, FR*

Lancaster University

Research Associate in Compiler and Runtime Optimisation @ *Lancaster University, GB*

BSC Barcelona Supercomputing Center

INTERNATIONAL PHD PROGRAM FELLOWSHIPS BSC_LA CAIXA @ *Barcelon Supercomputing Center (BS*

HiPEAC communications

Policy Corner

Sandro D'Elia discussing digital platforms at HiPEAC16

Crossing virtual borders with the Digital Single Market

What is the Digital Single Market and what does it mean for the computing systems sector in Europe? Sandro D'Elia, HiPEAC project officer in the Complex Systems and Advanced Computing Unit at the European Commission, explains.

These are testing times for Europe. Wars are raging on our doorstep, causing numerous refugees to arrive on the continent, and the temptation to build walls and close borders is strong everywhere. Terrorism is becoming a constant fear, and for the first time there is the real possibility that a member state will leave the European Union, while the economy is still plagued with high unemployment in many regions and a fragile banking system. It is not the first crisis, and will not be the last, but things are definitely not easy.

"The digital economy is growing fast and virtual borders are real problems for citizens and businesses"

The European Union is responding to this crisis with many different measures, some of them directly linked to a sector with which we're more familiar: digital technologies. In Euro jargon many of these measures go under the name Digital Single Market, or DSM (in Brussels acronyms grow fast – probably due to the weather).

The main thing that keeps Europe together is the single market: the guarantee that people, goods and money can move around the European Union freely. An interesting point is that the free movement of goods, which is obvious for oranges, furniture or shoes, is not so obvious for digital goods: for example, if you have a contract for British satellite TV in the UK, you cannot watch the Premier League with the same subscription in Greece. Other issues are fragmented data protection rules, differing copyright regimes, lack of access to broadband and geo-blocking of online content. The digital economy is growing very fast and all these virtual borders are real problems for citizens and businesses.

That's why DSM is needed: it is a set of actions which will gradually remove the 'digital borders' which still exist across Europe. It is based on three areas: giving better online access to digital goods and services, creating a regulatory environment where digital networks and services can prosper, and using digital technologies as a driver for economic growth. A specific action will be dedicated to leadership in industrial digital technologies, and this is where you can see the link between high-level European policy and the daily work of the HiPEAC community.

"The European Commission will push for leadership in digital technologies for European industry"

In practical terms, over the next few years the European Commission will push for leadership in digital technologies for European industry, and this will be done by supporting the emergence of digital industrial platforms. All sectors of the economy will be affected: from 'high tech' areas, like aerospace or energy, to some markets which are highly visible to consumers, like the automotive sector, and also sectors which have traditionally been considered 'low tech', such as agriculture and construction.

Policy Corner

For the HiPEAC community there are two main consequences.

- One: funding opportunities in the coming years will be focused on the areas where industrial applications are possible; this means, for example, advanced computing and cyber-physical systems, manufacturing applications, numeric simulation for products and production processes, innovative applications in areas such as food production or civil engineering.
- Two: the creation of platforms will be the preferred approach to support industrial applications (instead of one-shot solutions). And here, it should be clear that a platform, in this context, is not just a piece of software with a public API, but a set of technologies that will become the foundation for a market, where many different actors can provide added value and carry out real business. The aim here is not to create another consumer 'app store', but a market oriented towards industry which can strengthen the leadership of European players in areas of high economic value like transport, critical applications, automotive, aerospace and manufacturing machines.

The overall objective is to strengthen European industry, creating value and jobs. HiPEAC is one element of this strategy: you have the know-how to build digital platforms, you know the technology and you teach digital technologies to the scientists and engineers who will enter the job market in future.

But the task before us is not easy, because thinking in terms of platforms – and not only in terms of technical solutions – requires a different mindset: it requires thinking big, in economics as well as in technological terms, considering issues like standards, certifications, compatibility and of course market evolution. Although challenging, this task is essential because there is no 'plan B': European industry needs digital platforms to stay competitive and to create decent jobs. Otherwise the next digital revolution will be driven by somebody else, and we might not like the results.

Digitising Industry graph

MORE INFORMATION:
https://ec.europa.eu/priorities/digital-single-market_en
https://ec.europa.eu/digital-agenda/en/news/updated-view-european-commission-digitising-european-industry-initiative
https://ec.europa.eu/digital-agenda/en/digitising-european-industry#Latest
https://ec.europa.eu/digital-agenda/en/digitising-european-industry#Article
A video of Sandro D'Elia discussing digital platforms is available on the HiPEAC YouTube channel: http://bit.ly/HiPEAC16_YouTube

"The aim is not to create another consumer 'app store', but a market oriented towards industry which can strengthen the leadership of European players in areas of high economic value"

TWEETS 1.288 | VOLGEND 336 | VOLGERS 825 | VIND-IK-LEUKS 221 | LIJSTEN 8

Volgen

HiPEAC
@hipeac

European network bringing together the cream of the high-performance and embedded architecture and compilation sector. CSA funded by #Horizon2020 / #H2020

Europe
hipeac.net
Geregistreerd in januari 2012

203 foto's en video's

Tweets | Tweets en antwoorden | Media

Vastgemaakte Tweet

HiPEAC @hipeac · 17 mei

'When you have a system which manages your pacemaker, you can't stop it to install a security update' - S D'Elia

> **HiPEAC** @hipeac
> Video: Sandro D'Elia @DigIndEU discusses challenges for #computingsystems youtube.com/watch?v=VWzYat… cc @fet_eu @DSMeu

HiPEAC @hipeac · 6 u

What should the Commission be doing differently? What should they keep the same? asks Paco Ibañez @Electronics_EU

Nieuw op Twitter?

Registreer je nu om je eigen persoonlijke tijdlijn te krijgen!

Registreren

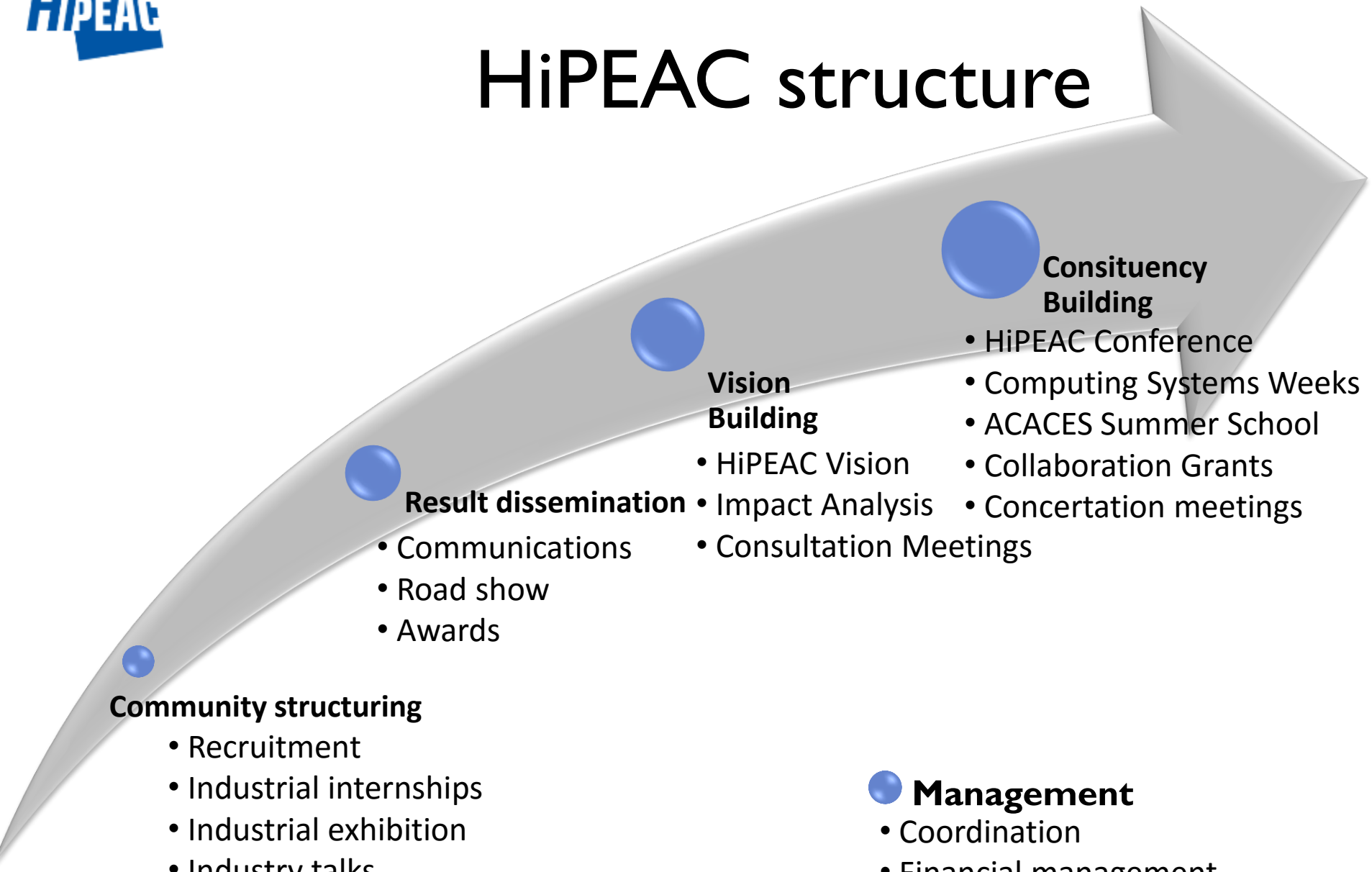
Je bent misschien ook geïnteresseerd in

· Vernieuwen

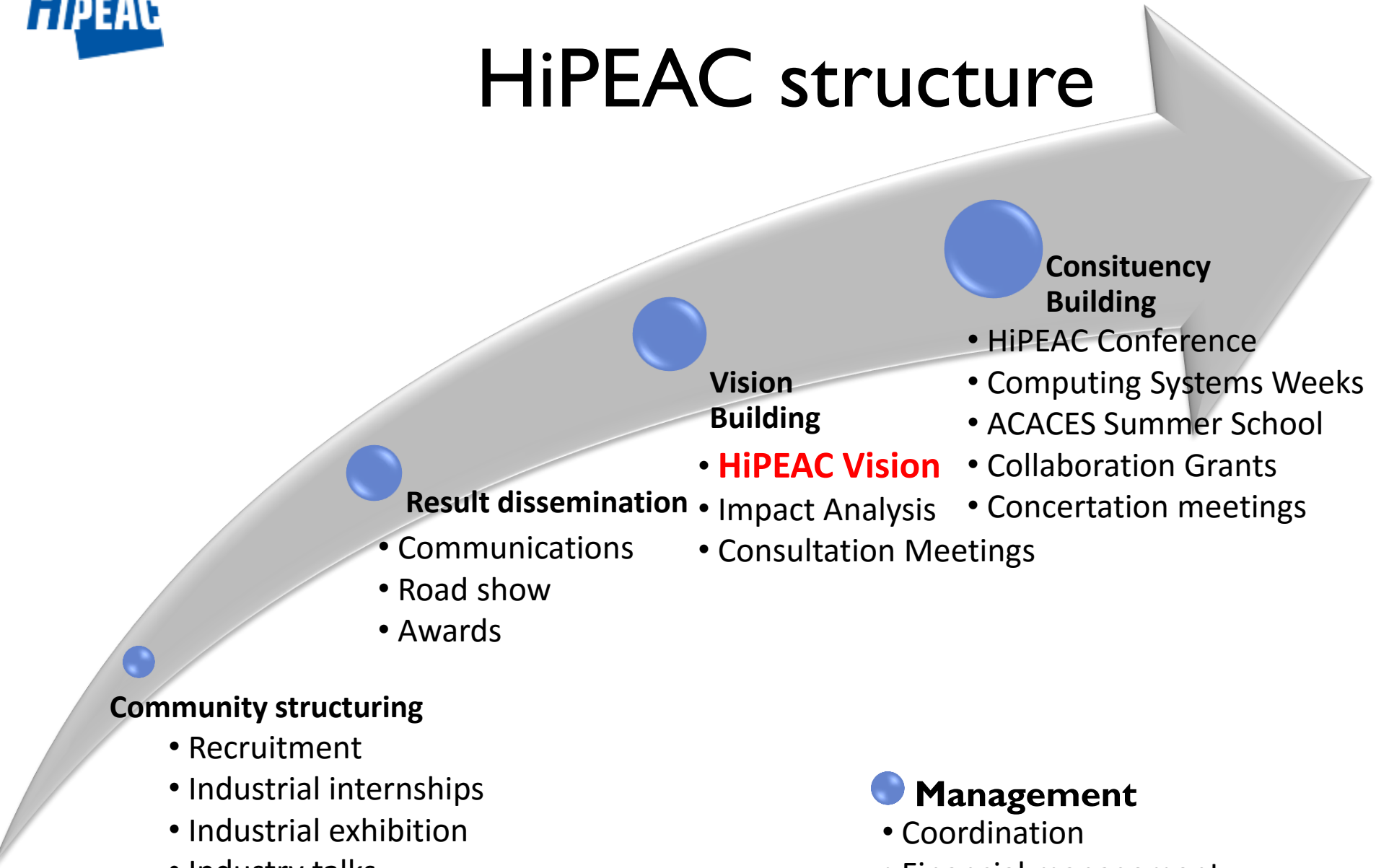
- **Mont-Blanc** @MontBlanc_Eu
- **Adept** @adept_project
- **dividiti** @dividiti
- **REPARA Project** @reparaproject
- **PRACE** @PRACE_RI

@hipeac **hipeac.net/linkedin**

HiPEAC structure



HiPEAC structure



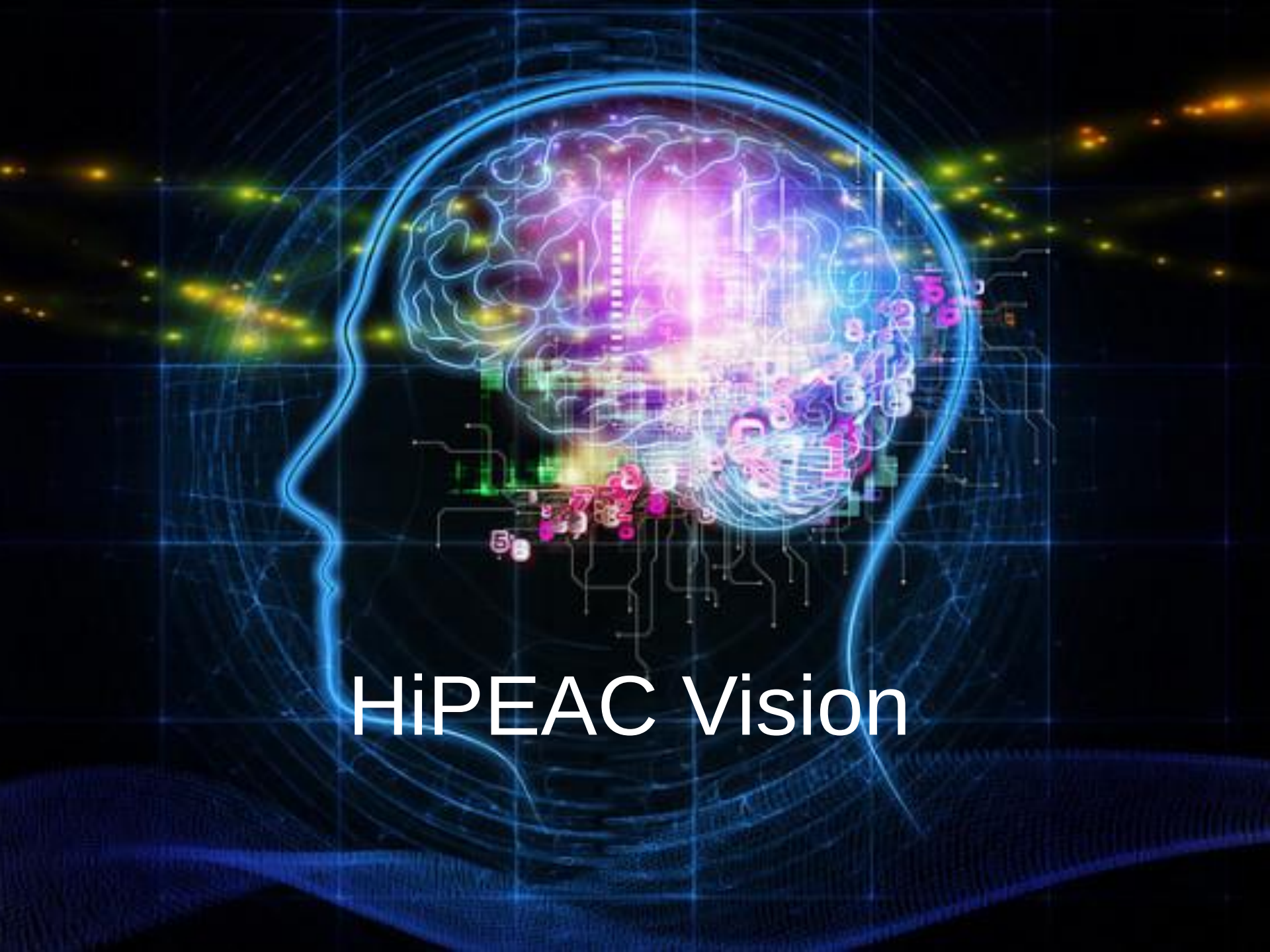
Community structuring
- Recruitment
- Industrial internships
- Industrial exhibition
- Industry talks
- Innovation stimulation

Result dissemination
- Communications
- Road show
- Awards

Vision Building
- **HiPEAC Vision**
- Impact Analysis
- Consultation Meetings

Consituency Building
- HiPEAC Conference
- Computing Systems Weeks
- ACACES Summer School
- Collaboration Grants
- Concertation meetings

Management
- Coordination
- Financial management
- Membership management


HiPEAC Vision

The HiPEAC Vision

The last HiPEAC Vision Document was published in January 2015. ***The next one is scheduled for 2017***

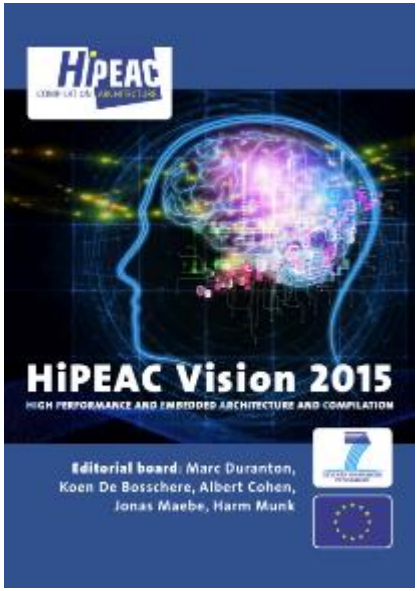
One of its aim is to help defining the next European calls in ICT.

2008

2009

2011

2013



2015

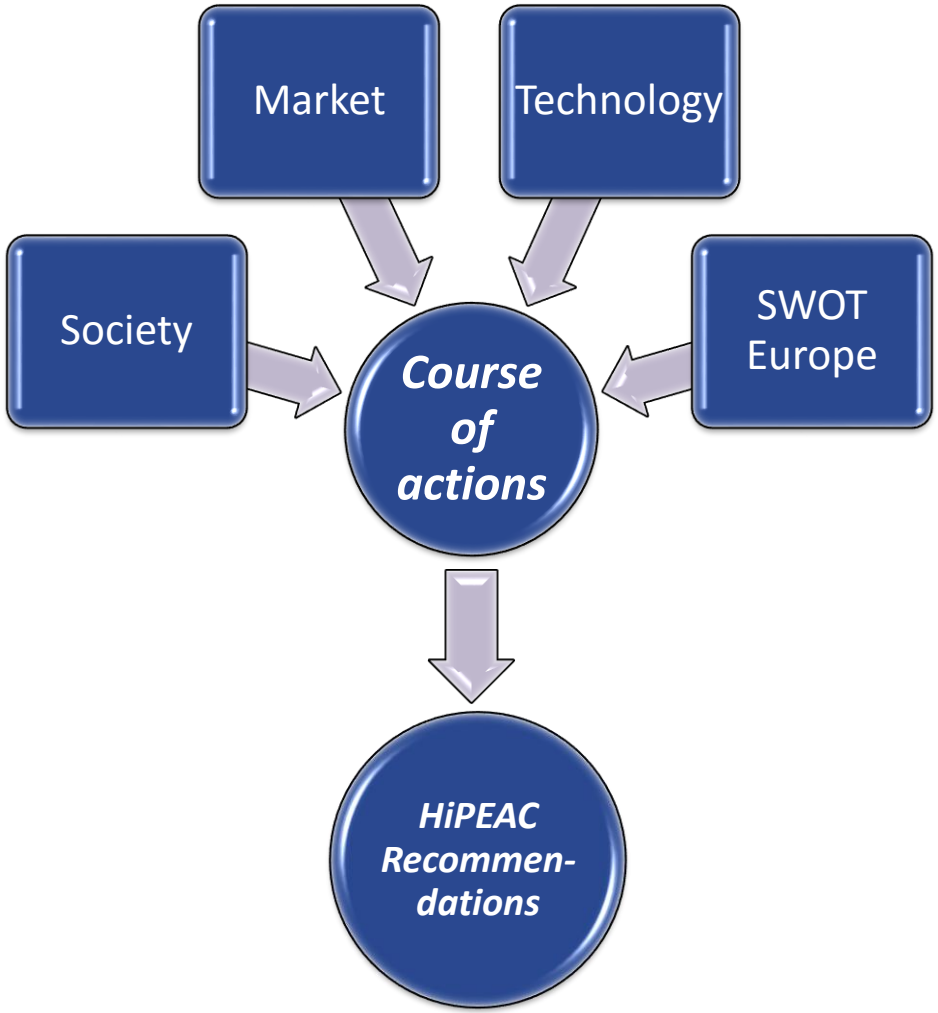
?

http://www.hipeac.org/vision/

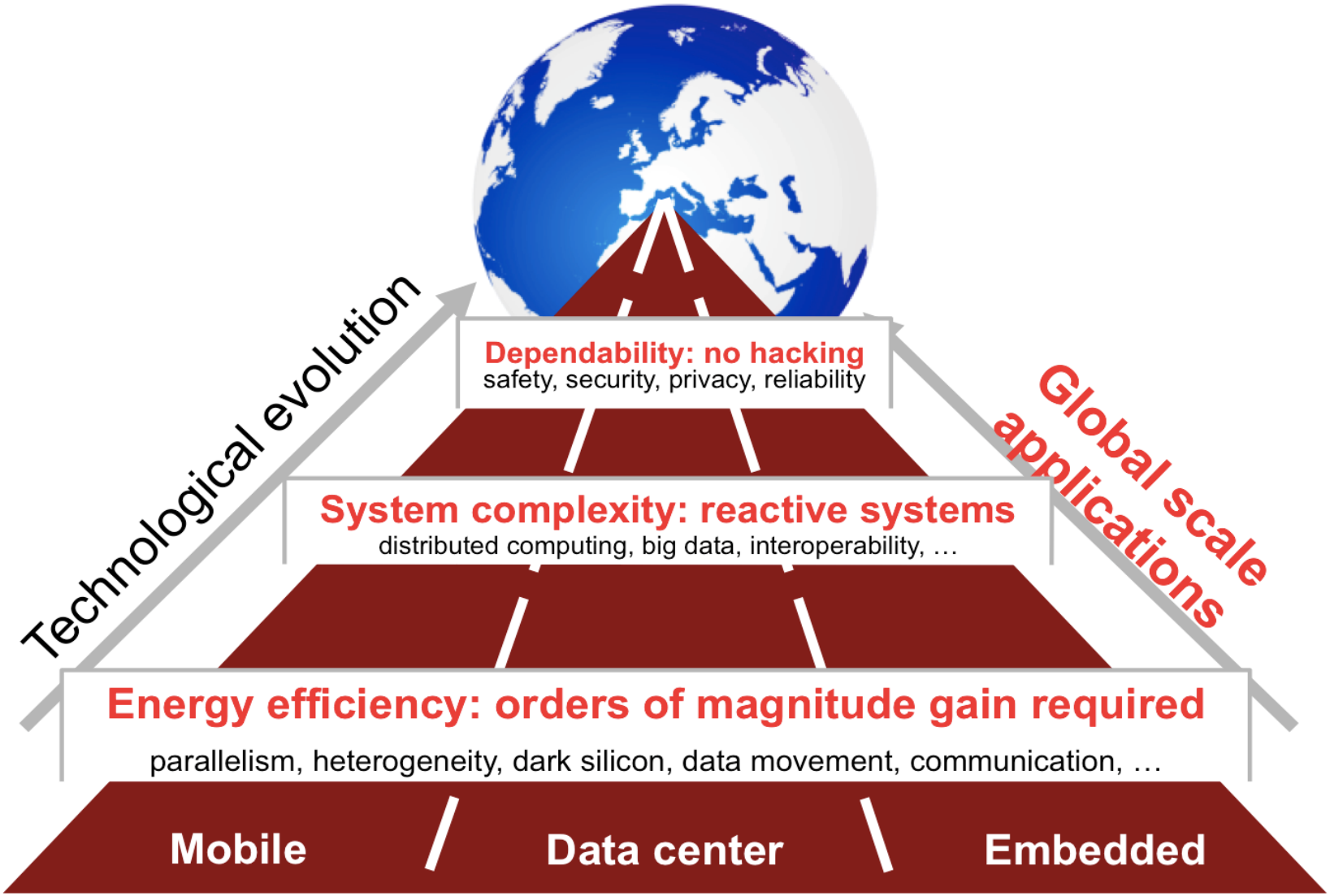
You want to help us?
Fill-in the survey!

https://www.surveymonkey.com/r/hipeacvision2025

Structure of the HiPEAC vision 2015



The HiPEAC vision 2013 is still valid but with even more emphasis



Highlights of the **HiPEAC Vision** 2015

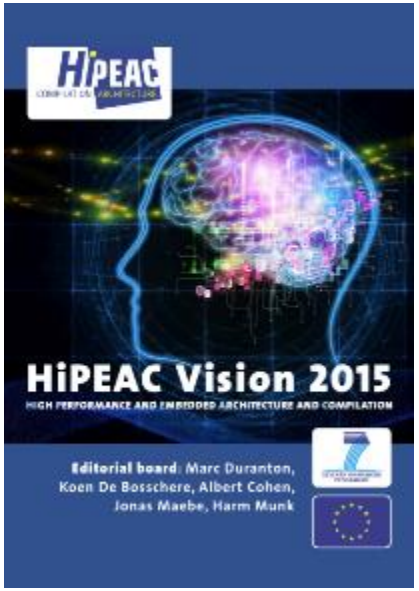
For the first time, we have noticed that the community really ***starts looking for***

disruptive solutions,

and that incrementally improving current technologies is considered inadequate to address the challenges that the computing community faces:

"The End of the World As We Know It"

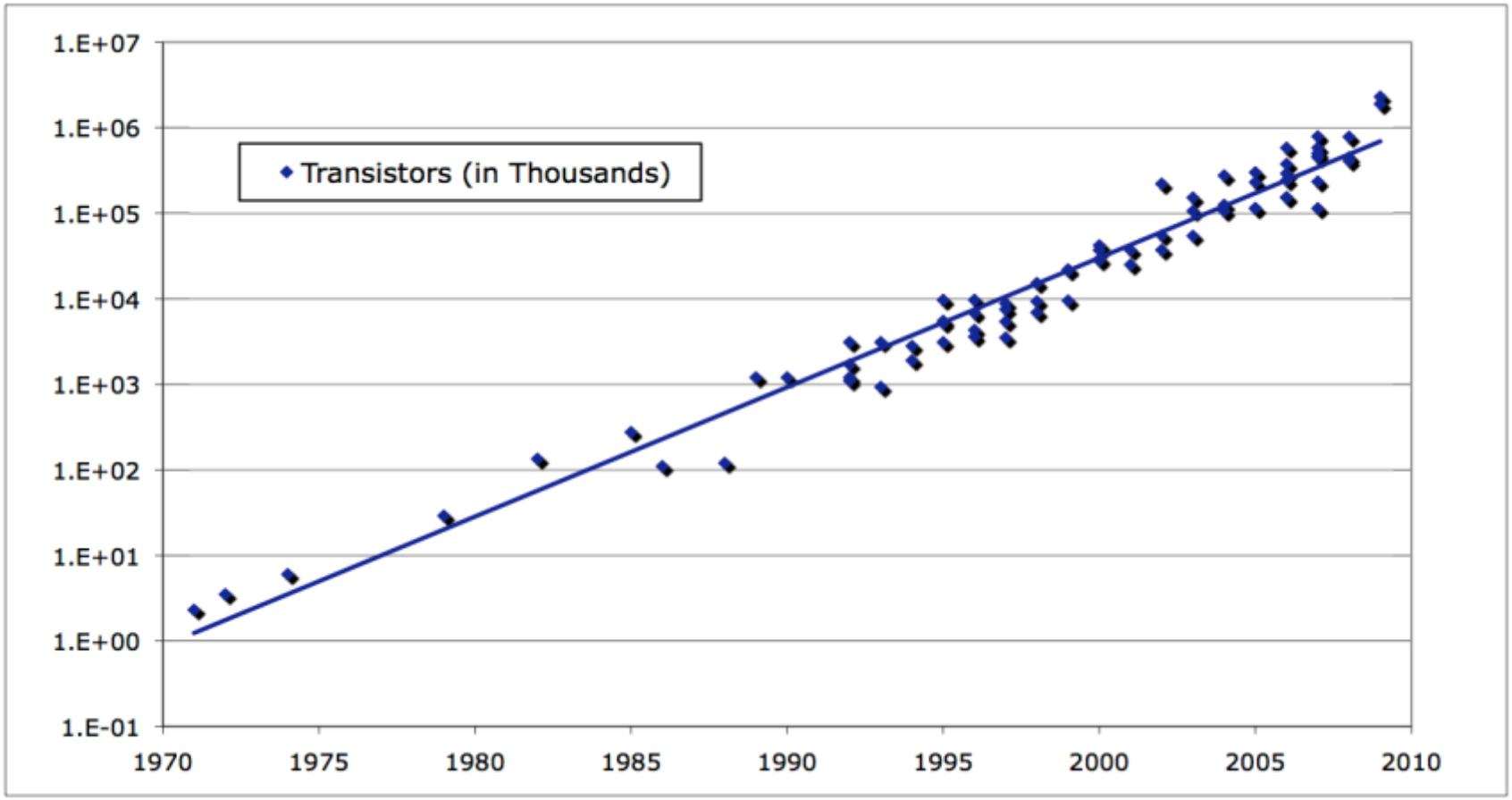
The End of the World As We Know It…



From the technology…

A little bit of history an the impact of technology on software…

Moore's law: increase in transistor density



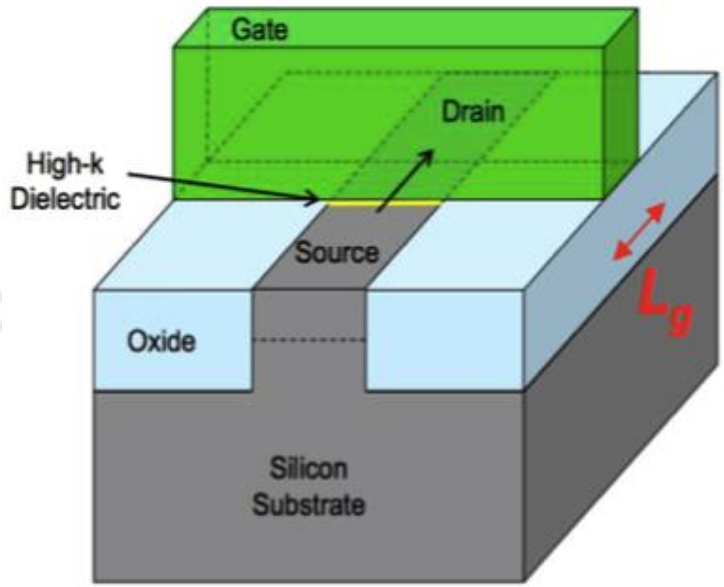
Source from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanoviç

Moore's law still on going…

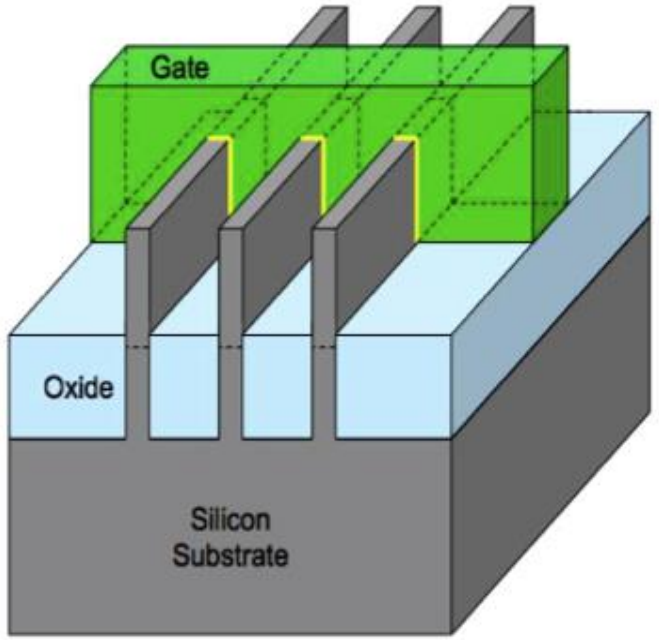
Advanced CMOS roadmap

Evolutionary scaling: technology driven performance improvement

Transistor 2D



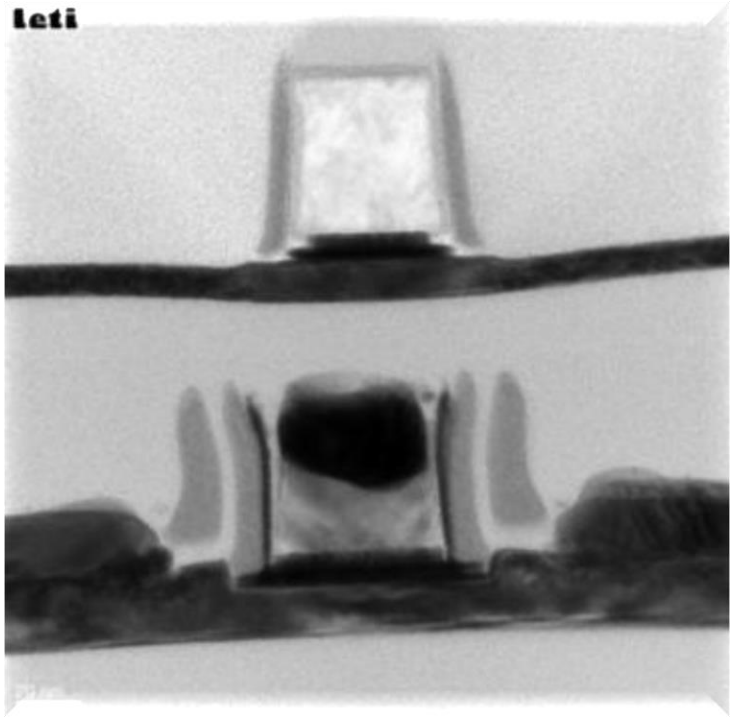
Transistor « Tri-gate » 3D

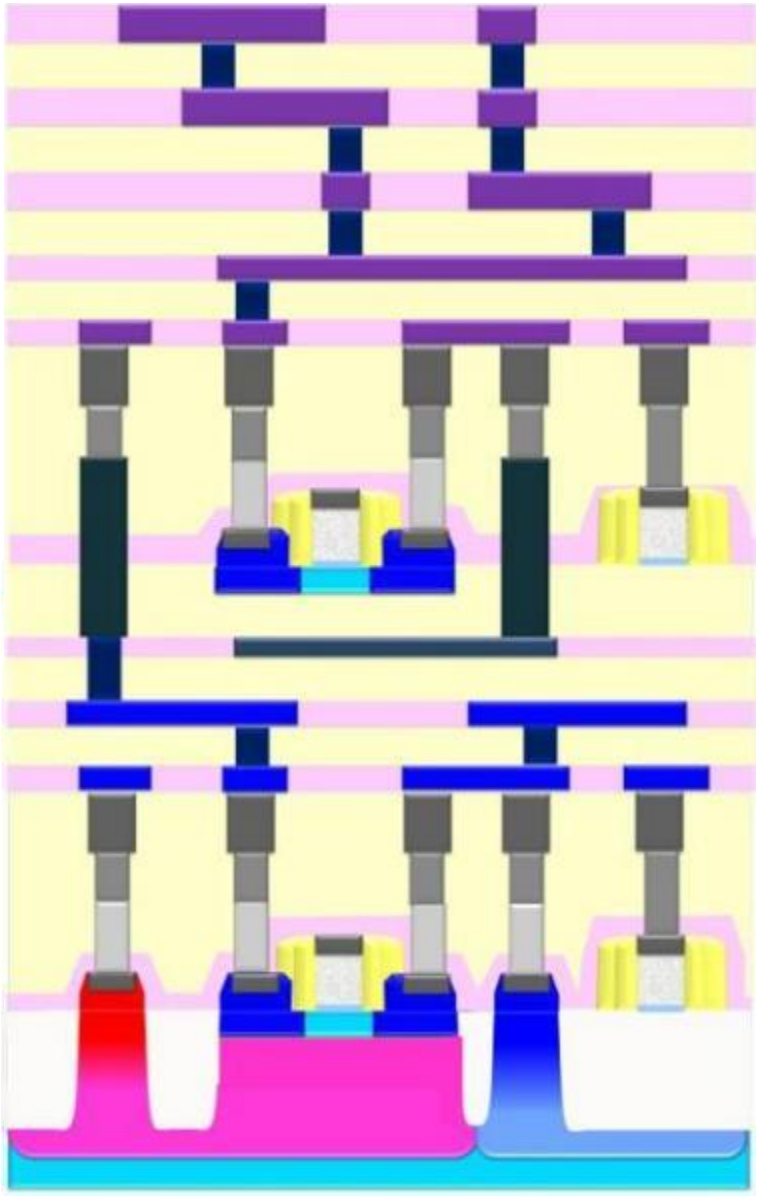


Current control difficult when L_g< 20nm

Better control (2 sides + multiple gates)

M3D principle





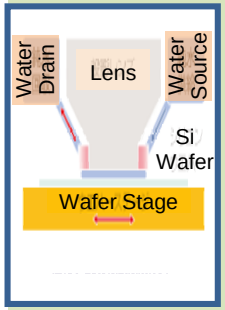
CMOS/CMOS: 14nm vs 2D:
Area gain=55%
Perf gain = 23%
Power gain = 12%

LETI, DAC 2014

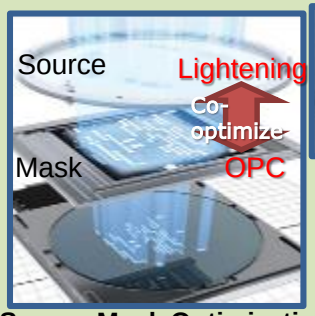
Technology Roadmap

45/40nm	32/28nm	22/20 nm	16/14nm	10nm	7nm	5nm ~Beyond

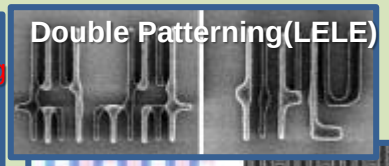
Litho

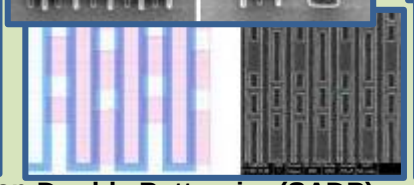


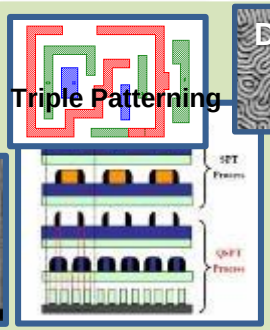
193nm Imm.ArF



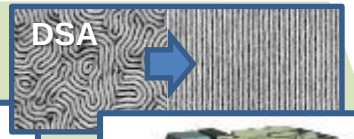
Source Mask Optimization



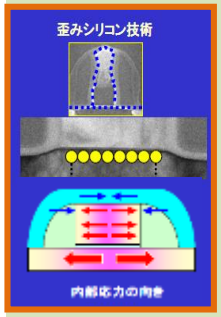

Double Patterning(SADP)



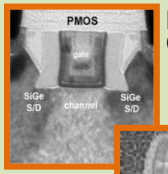
SAQP



Front End

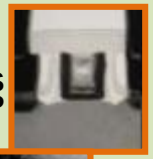


Stress Engineering



HKMG GL

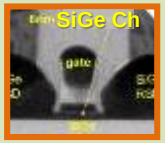
NMOS SiC:P




FDSOI

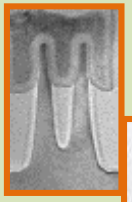
SiGe P-Ch.


Fin FET

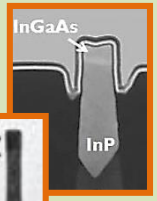


FDSGOI(Pch)

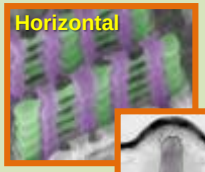
FDSOI(Strain)


Weff boost Fin FET

Ge channel



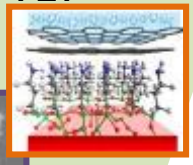
III-V channel



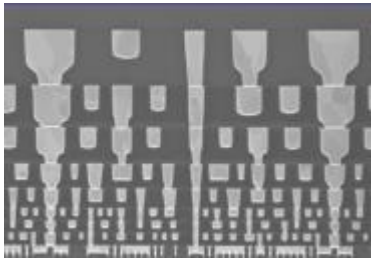
Gate-All-Around

2D Material

Graphene FET

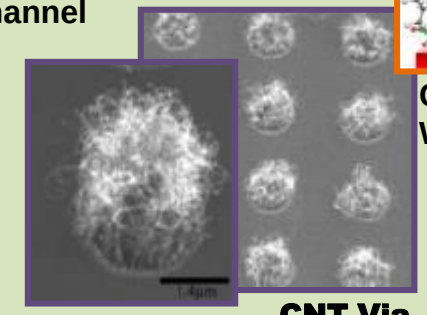

Graphene Wiring

Back End


K=2.4~2.8

Ultra low-k

Air Gap


CNT Via

Courtesy: Yuzo Fukuzaki – JAPAN PIDS

The end of Dennard Scaling

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

Everything was easy:

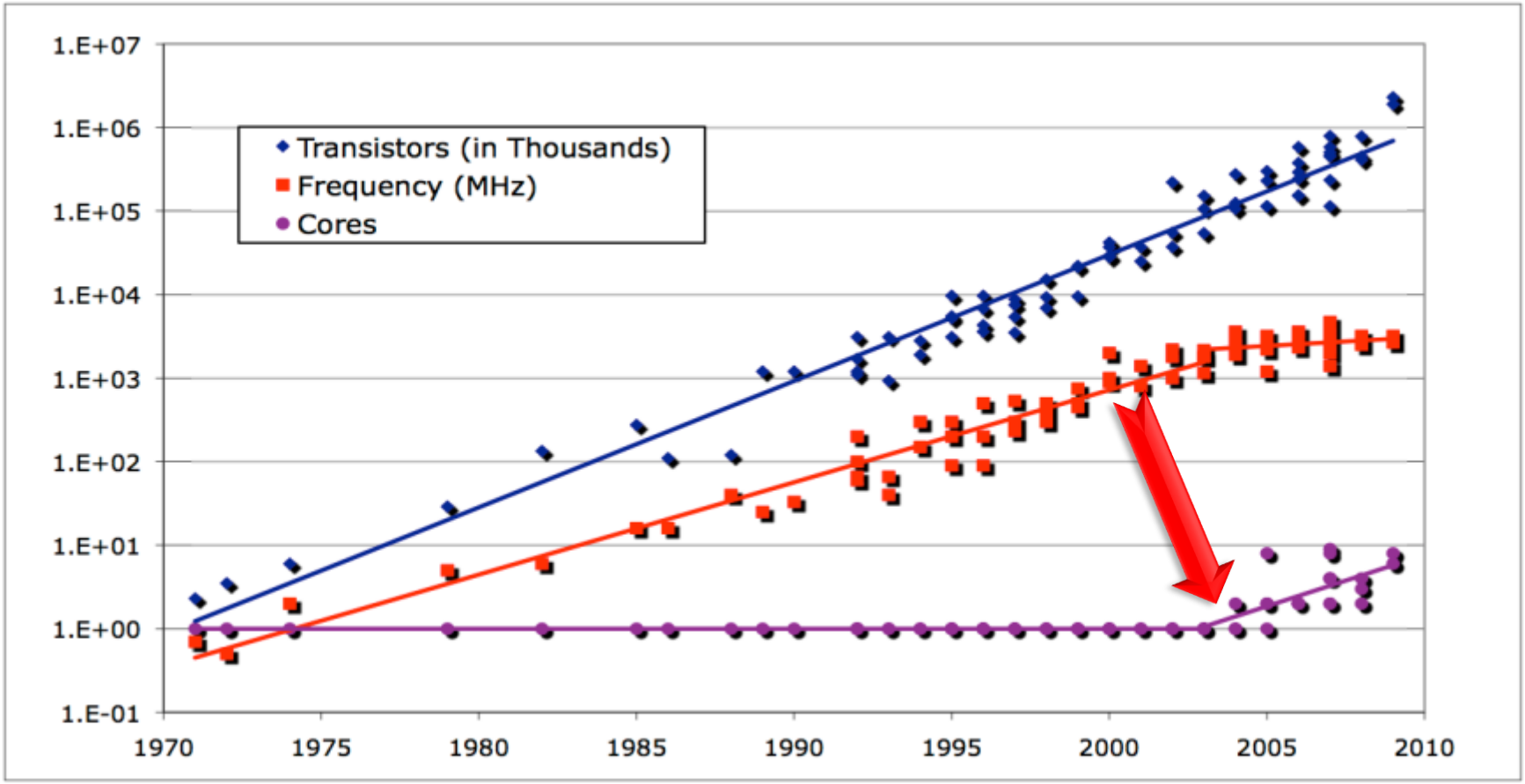
- Wait for the next technology node
- Increase frequency
- Decrease Vdd

->Similar increase of sequential performance

-> No need to recompile (except if architectural improvements)

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

Limited frequency increase -> more cores

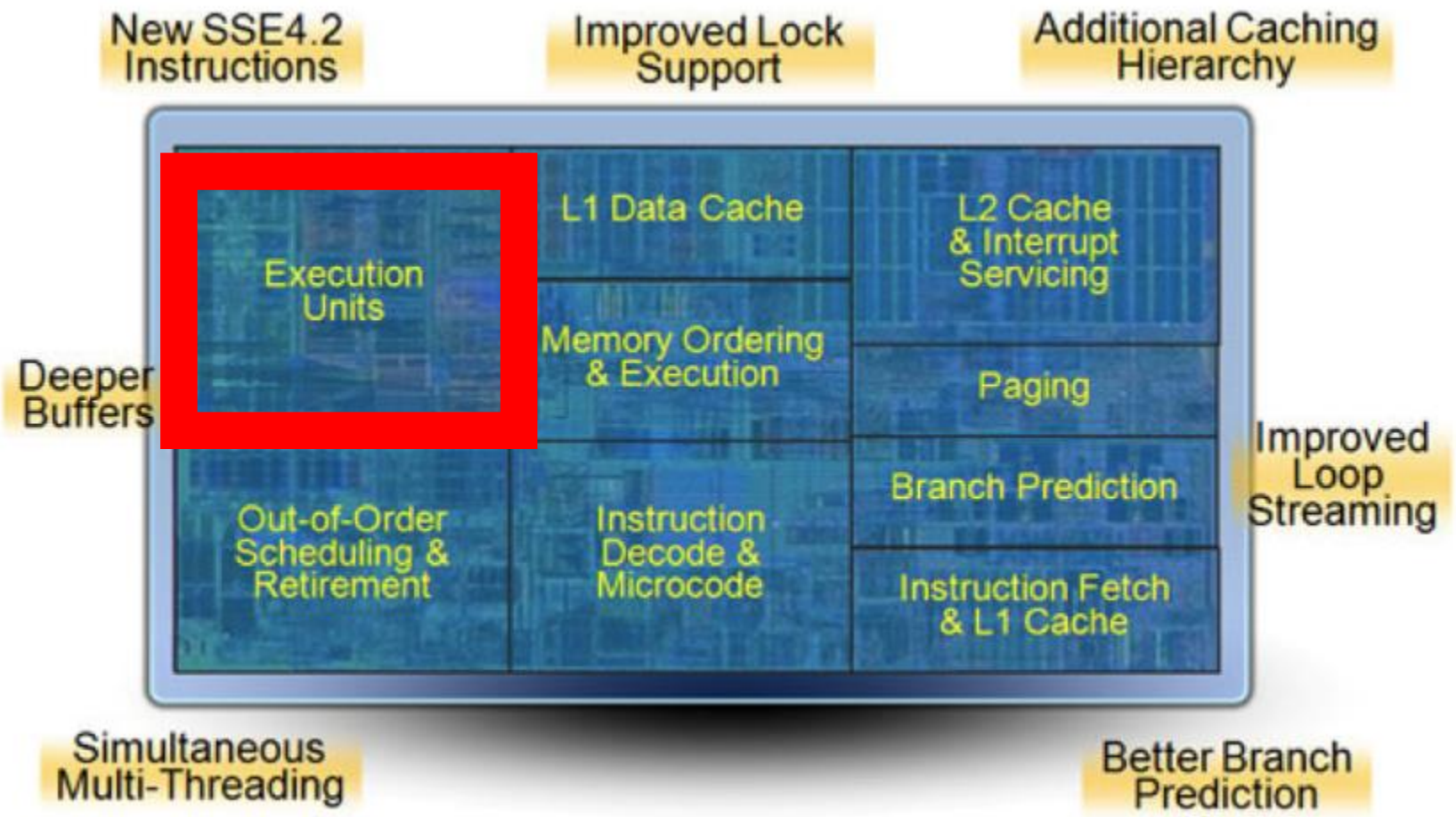


Source from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanoviç

Why using several compute cores?

1. Using several cores is 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).

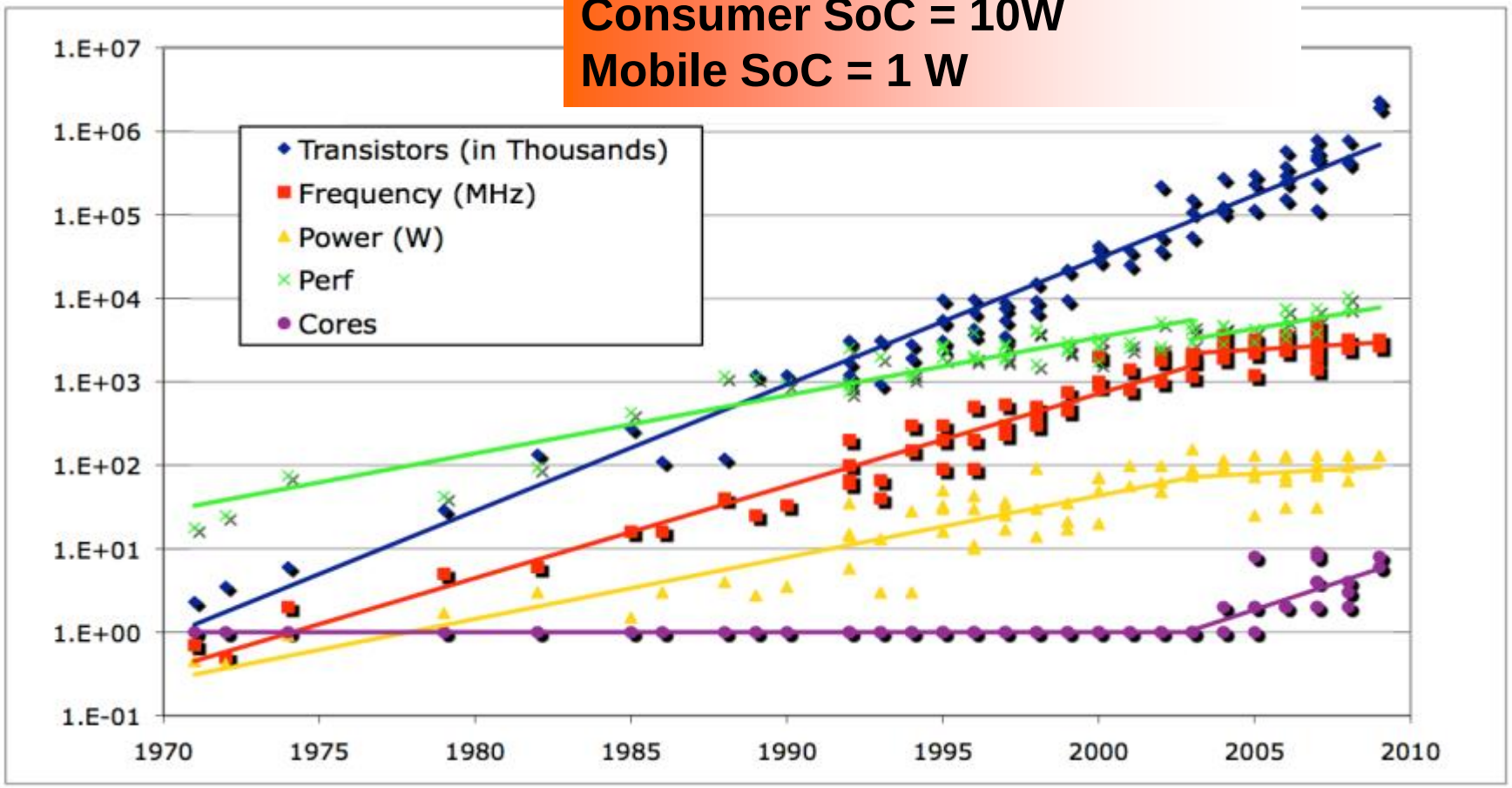
Less than 20% of the area for execution units



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

Limitation by power density and dissipation

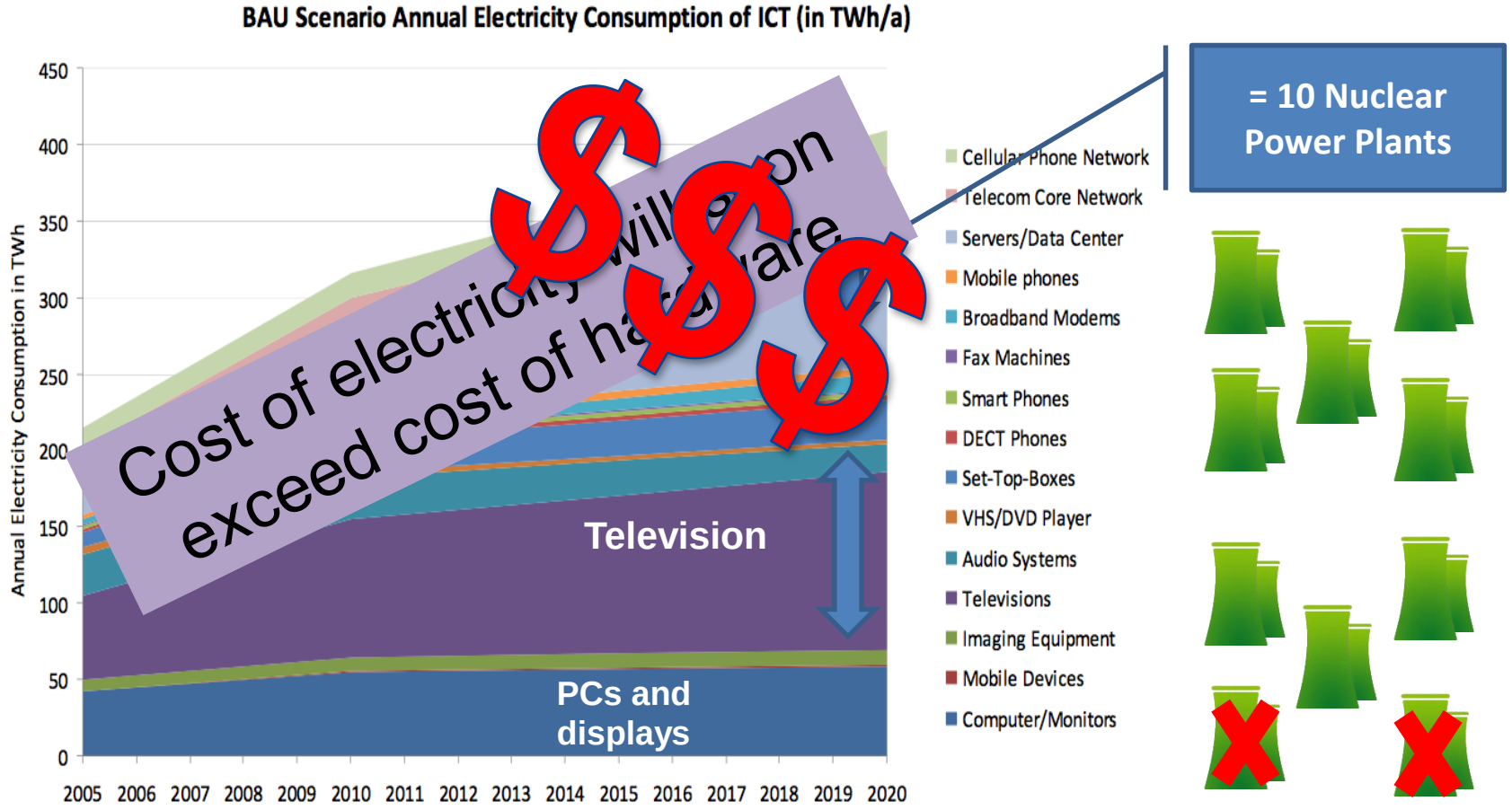
GP CPU = 200 W (45 nm)
Consumer SoC = 10W
Mobile SoC = 1 W



Source from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanovic

Energy consumption of ICT

- Estimated consumption 410 TWh in 2020, 25% for servers

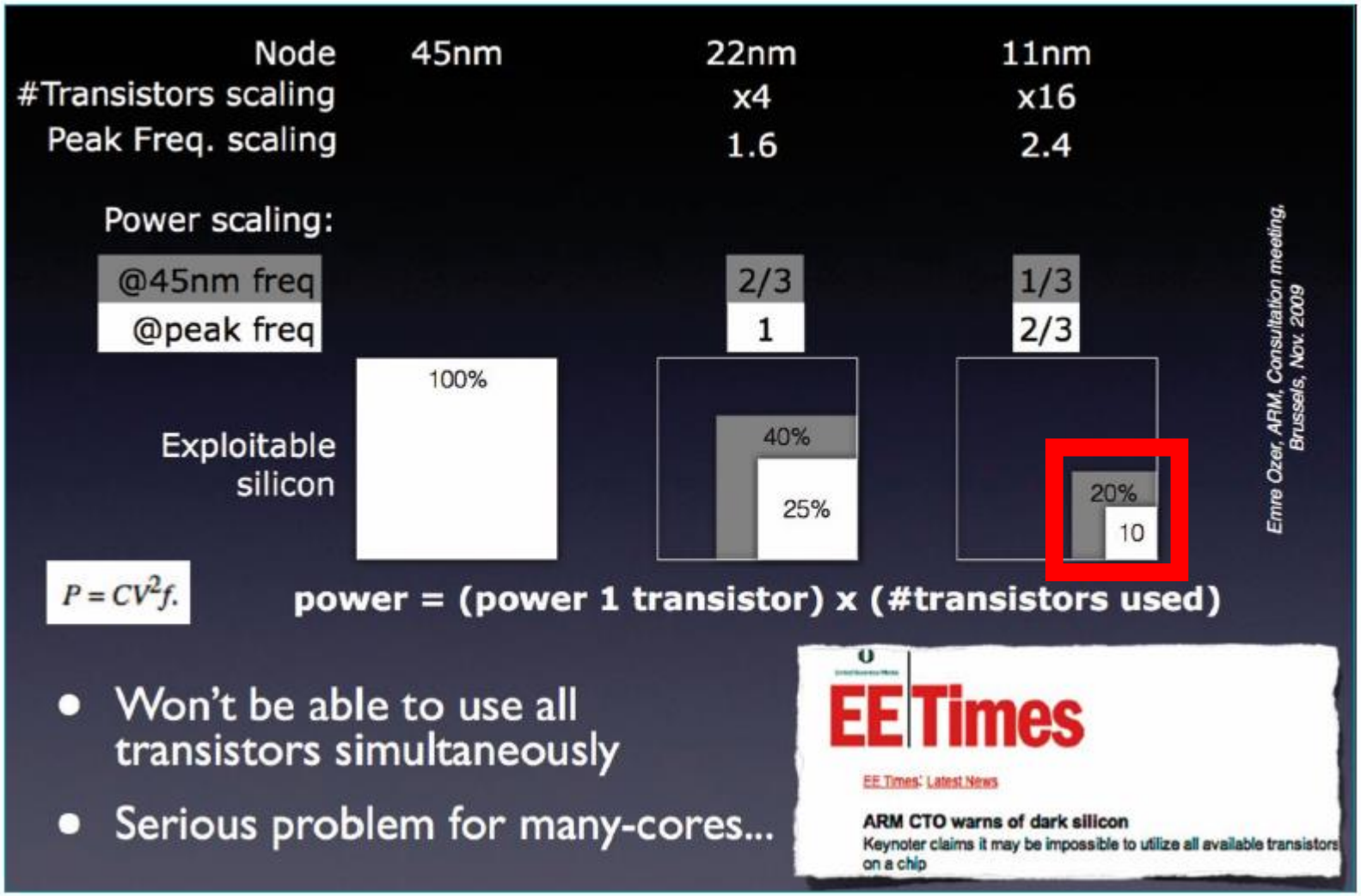


Source: European Commission DG INFSO, Impact of Information and Communication Technologies on Energy Efficiency, final report, 2008

The energy challenge for HPC

Source: Timothy Lanfear, « GPU computing and the future of HPC »

Power limits the active silicon area
=> "Dark silicon"
=> More efficient specialized units



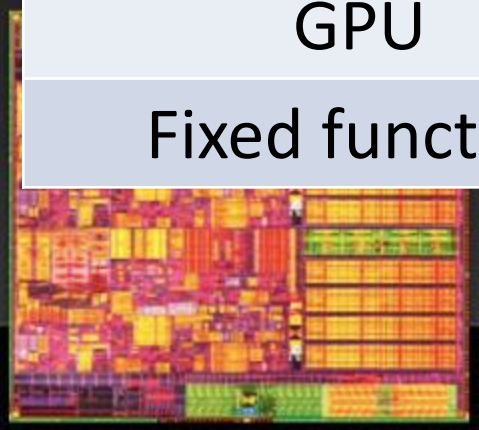
Specialization leads to more efficiency

CPU
1690 pJ/flop

GPU
140 pJ/flop

Type of device	Energy / Operation
CPU	1690 pJ
GPU	140 pJ
Fixed function	

← FPGA with HLS
“software programming space and not only time”

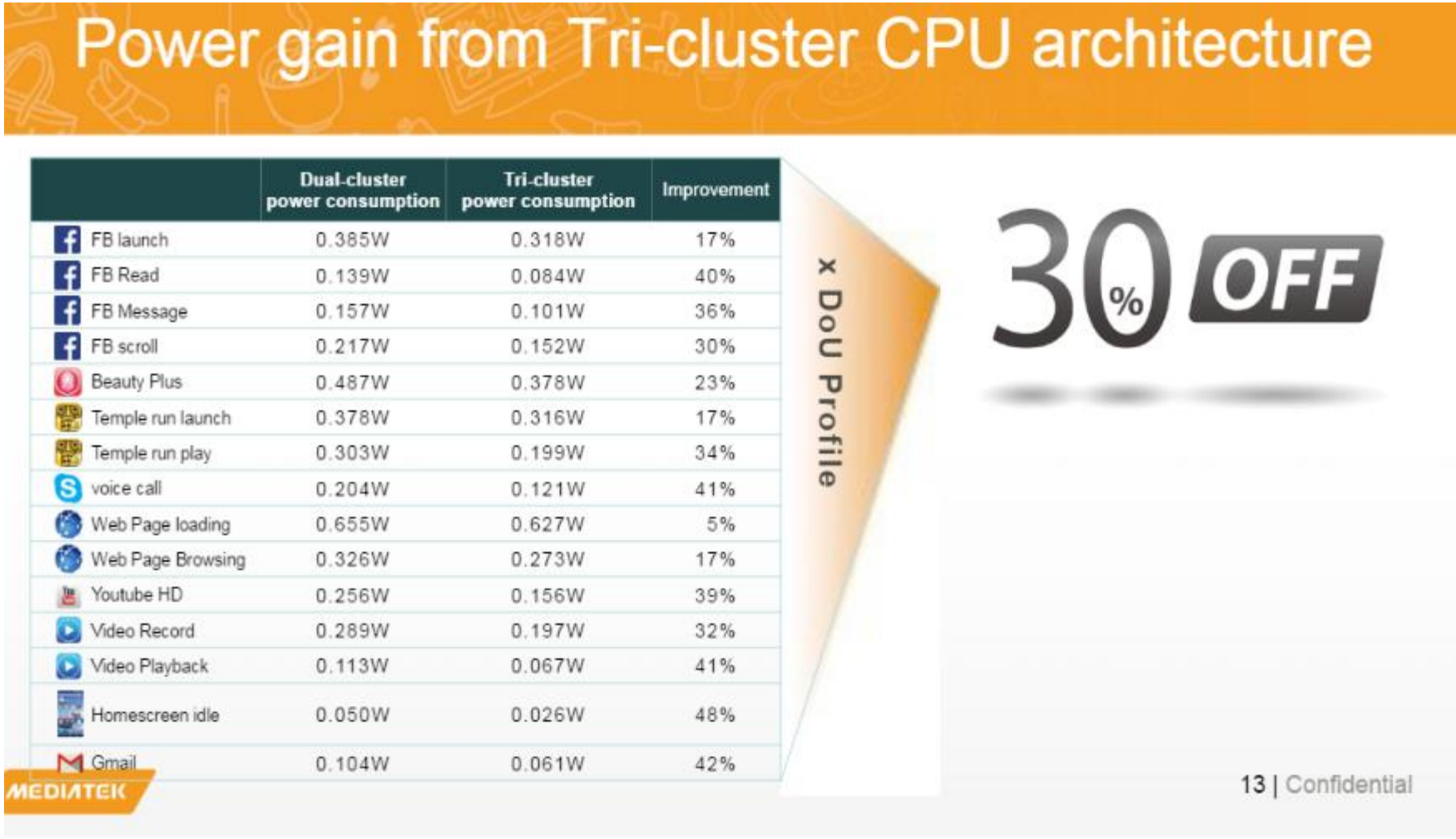


Westmere
32 nm

Kepler
28 nm

Source from Bill Dally (nVidia) « Challenges for Future Computing Systems » HiPEAC conference 2015

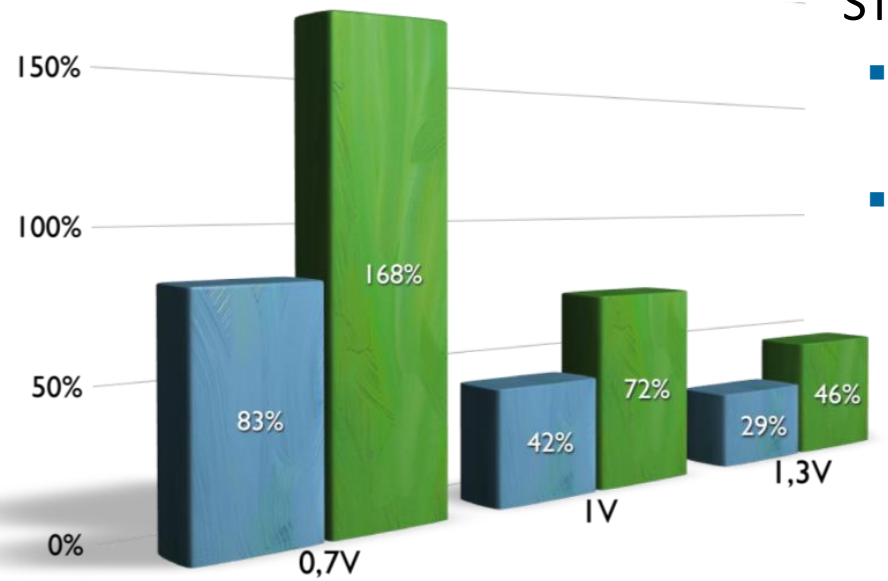
Example of specialization: big-LITTLE architecture from ARM, extended by Mediatek to 3 clusters



	Dual-cluster power consumption	Tri-cluster power consumption	Improvement
FB launch	0.385W	0.318W	17%
FB Read	0.139W	0.084W	40%
FB Message	0.157W	0.101W	36%
FB scroll	0.217W	0.152W	30%
Beauty Plus	0.487W	0.378W	23%
Temple run launch	0.378W	0.316W	17%
Temple run play	0.303W	0.199W	34%
voice call	0.204W	0.121W	41%
Web Page loading	0.655W	0.627W	5%
Web Page Browsing	0.326W	0.273W	17%
Youtube HD	0.256W	0.156W	39%
Video Record	0.289W	0.197W	32%
Video Playback	0.113W	0.067W	41%
Homescreen idle	0.050W	0.026W	48%
Gmail	0.104W	0.061W	42%

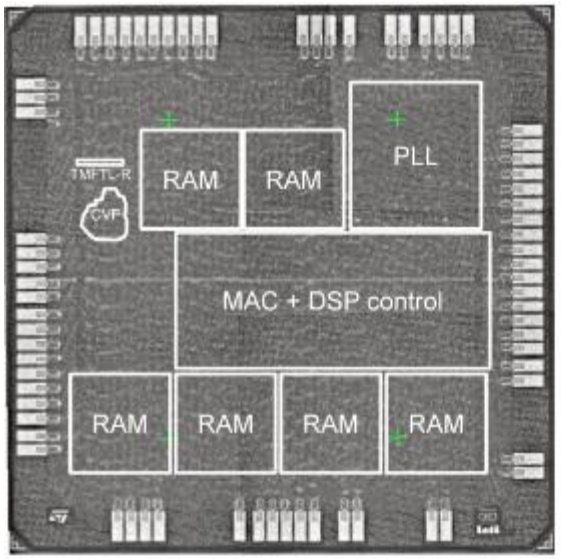
Energy efficient technology: FDSOI

UTBB-FDSOI performance gain versus conventional Bulk CMOS technology.
Blue: no body biasing, Green: FBB = +1V.



- Fully Depleted – Silicon on Insulator
  - **Improved performance-per-watt**
  - **Adaption to variability of loads under software control**

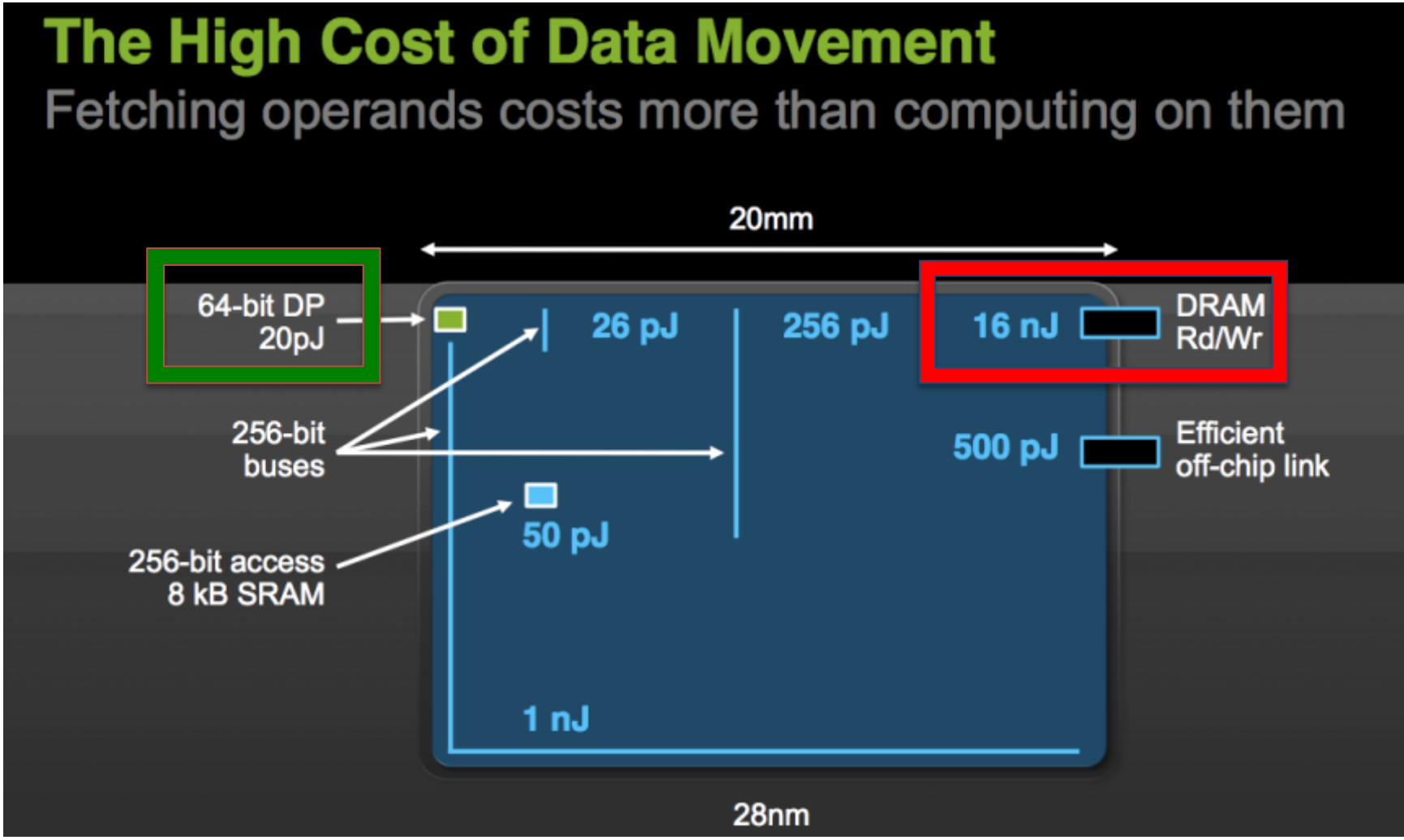
- Demonstrated by CEA tech and STMicroelectronics (ISSCC 2014)
  - Ultra-Wide Voltage Range (UWVR) operations: VDD=[0.39V – 1.3V]
  - High-frequency:
    Fclk > 2.6GHz @ 1.3V
    Fclk > 450MHz @ 0.39V



Technology	UTBB FDSOI 28 nm
Transistors	Flip-Well (LVT) L=24nm
Core area	1 mm²
DSP benchmark	FFT 1024
VDD range	0.397V-1.3V
VBB range	0V/±2V

Prototype chip from Beigné, ISSCC '14

Cost of moving data

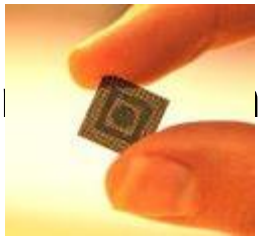


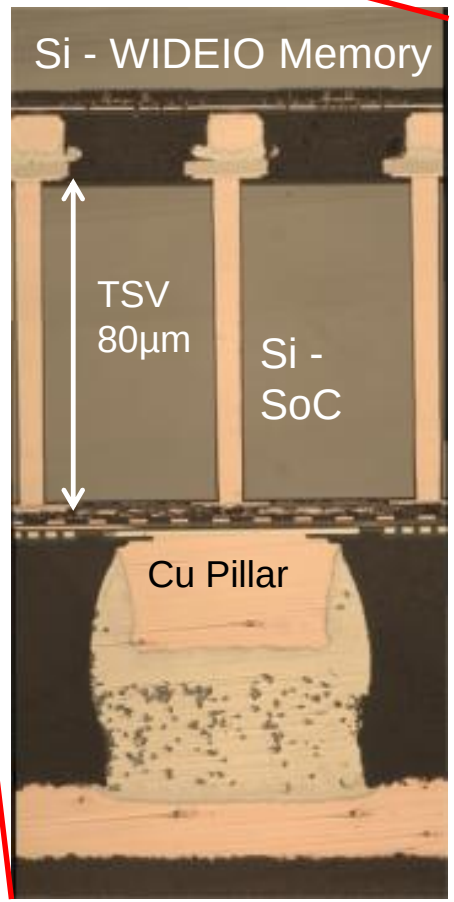
Source: Bill Dally, « To ExaScale and Beyond »
www.nvidia.com/content/PDF/sc_2010/theater/Dally_SC10.pdf

Using the 3rd dimension: 3D stacking

Exemple: WIDEIO memory stacked on top of a MPSoC in the same package

- **Partnership between CEA-LETI, STEricsson, STMicroelectronics and Cadence**
- **High bandwidth: WideIO provides more than 34 Gbytes/s (Currently: 17 GBytes/s)**
- **Low power: 4x power efficiency compared to LPDDR2/3**
- **Compatible with FD-SOI**
- **FBGA Package**
  - **Size 12x12mm, Ball Pitch 0.4m[illegible] thickness**





Source: Denis DUTOIT / CEA

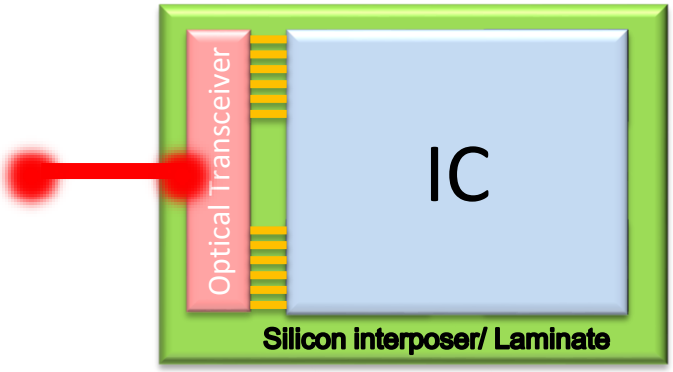
The data transfer challenge

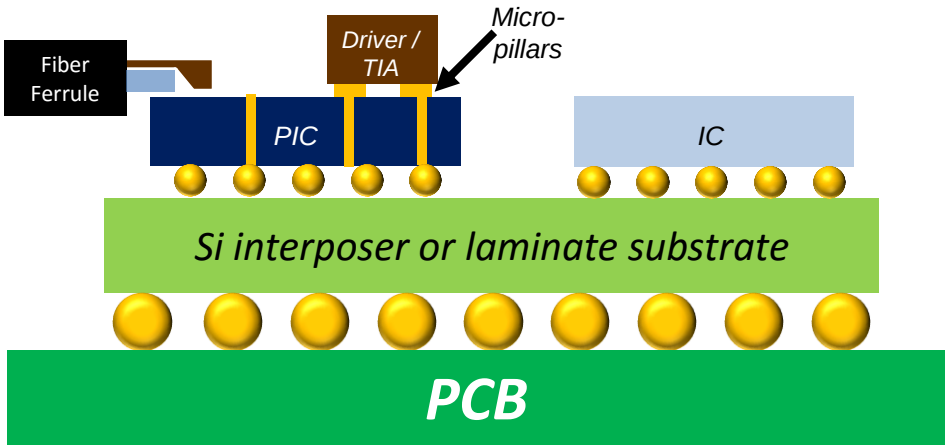
Source: Ahmed Jerraya CEA-Leti

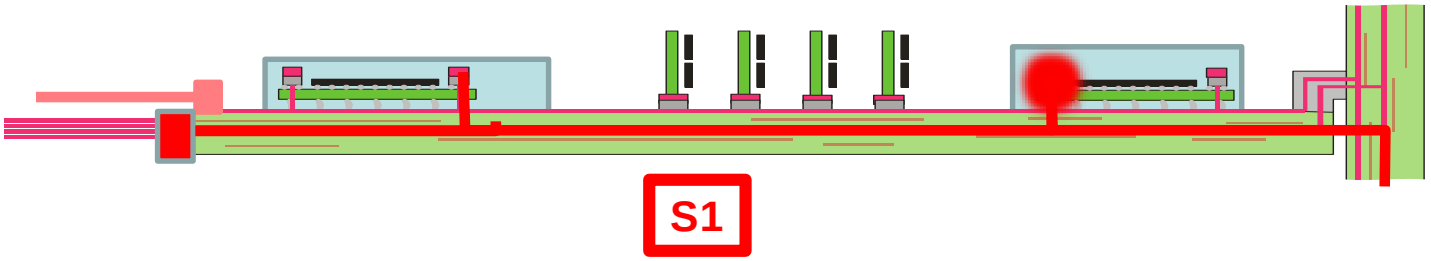
- Memory-interconnect density is becoming the bottleneck
- Bandwidth demand will increase ("**data deluge**")

Memory link, peak bandwidth and power consumption efficiency	Cost for 1TBps memory bandwidth Interface power consumption
 8.532 GBps 30 mW/Gbps 1066 MHz I/O bus clock, 32 bits, 1.5 V, Double Data Rate	240 W
 6.4 GBps 20 mW/Gbps 800 MHz I/O bus clock, 32 bits, 1.2 V, Double Data Rate	160 W
 12.8 GBps 4 mW/Gbps 200 MHz I/O bus clock, 512 bits, 1.2 V, Single Data Rate	32 W
 >TBps 1 mW/Gbps Assume 200 MHz 50K pins connected to SERDES to Photonics	8 W + MUX-DEMUX?

Off-chip photonics







Off board: AOC, optical modules

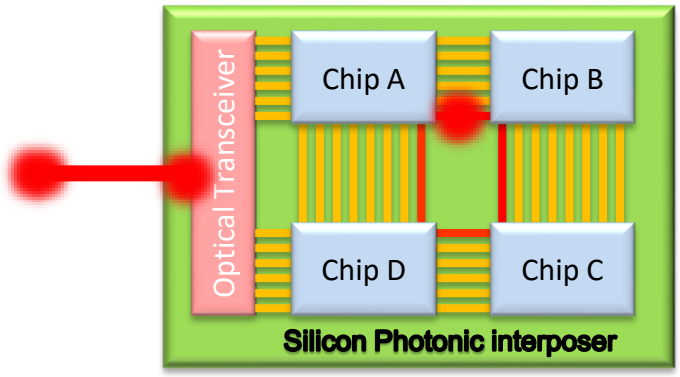
Off chip: Optical I/O

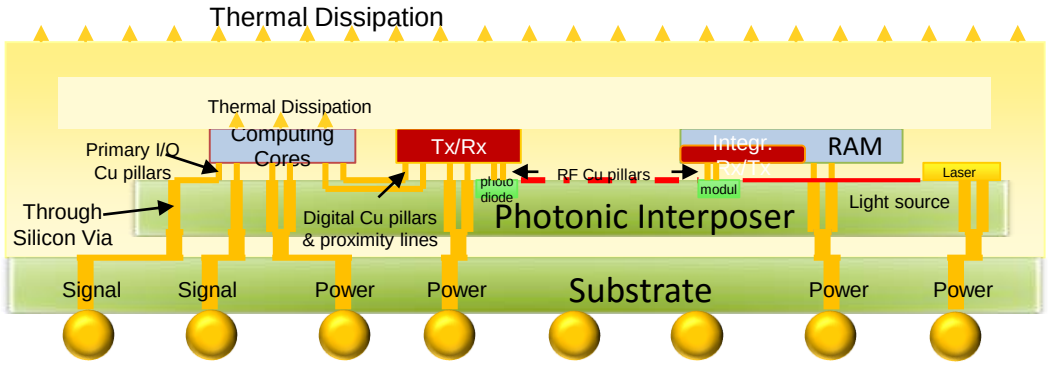
Time

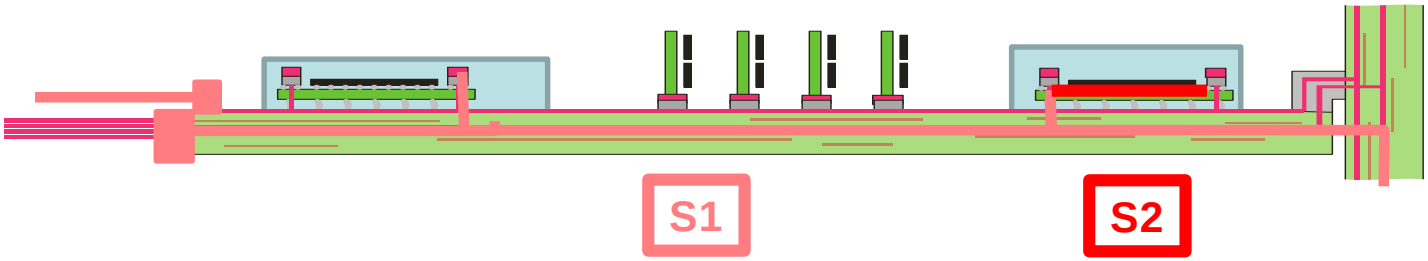
Now

Step 1

In-package photonics







Off board: AOC, optical modules

Off chip: Optical I/O

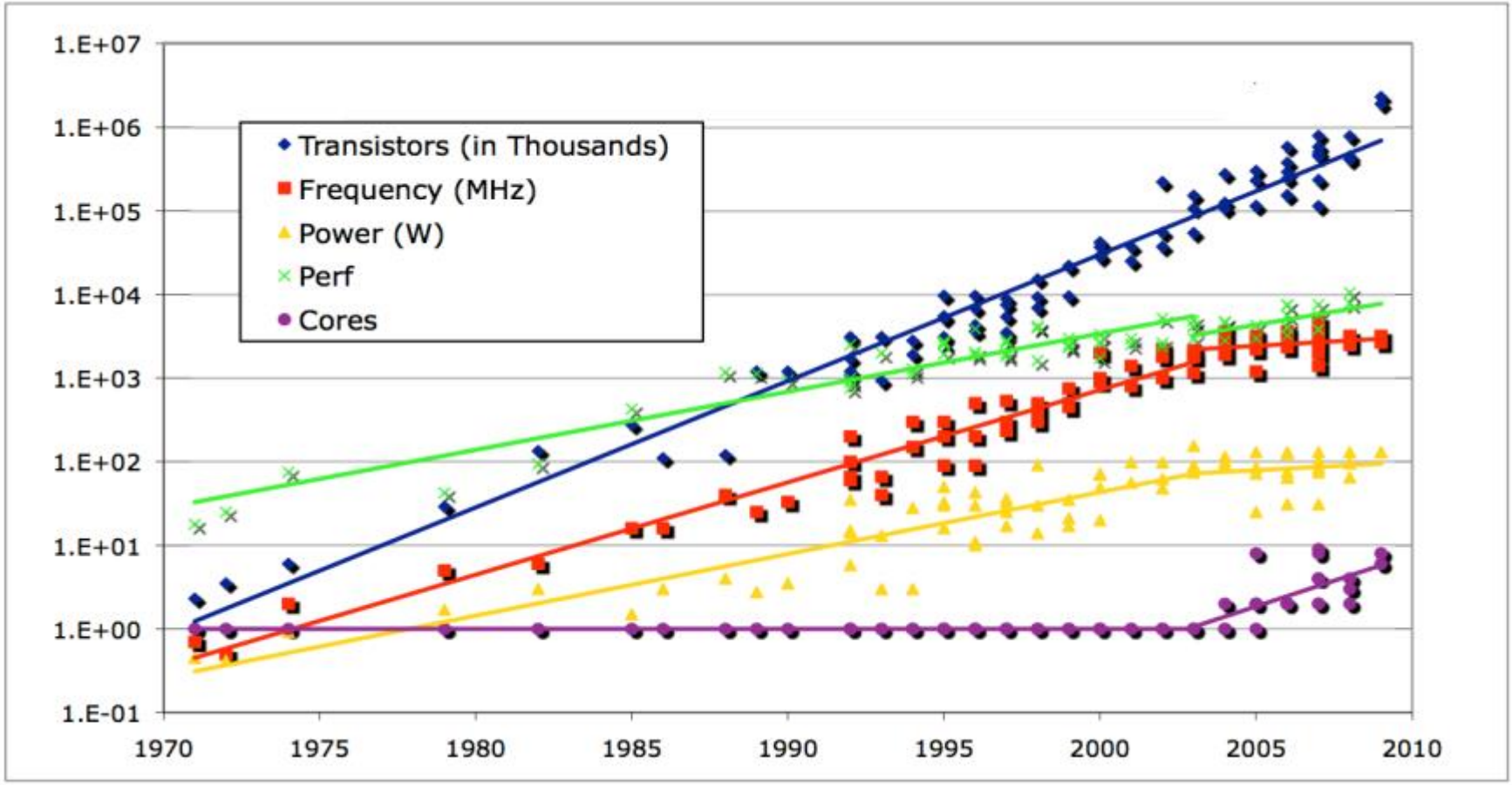
Optical network in package

Now

Step 1

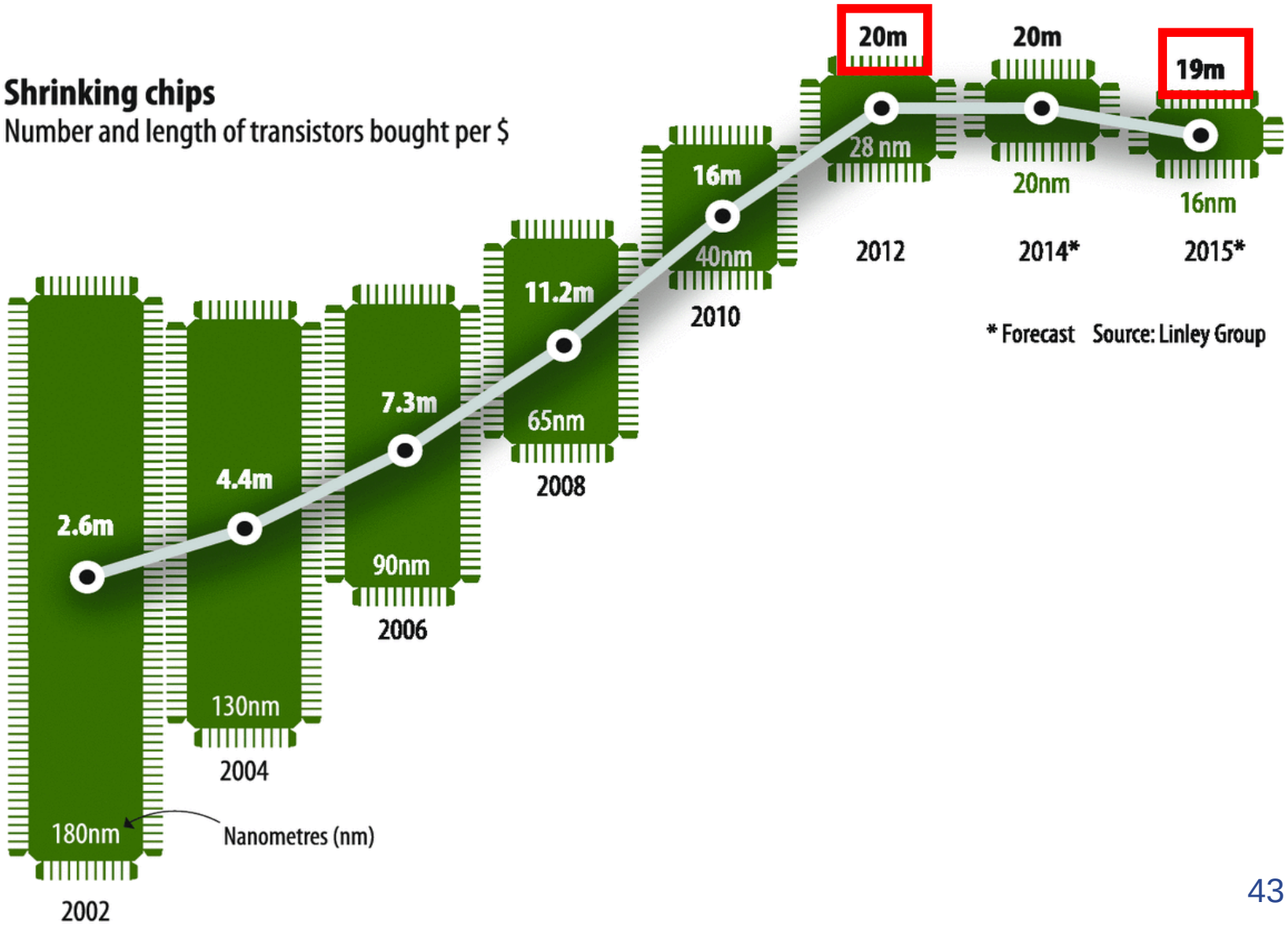
Step 2

Limitation by power density and dissipation … but not only

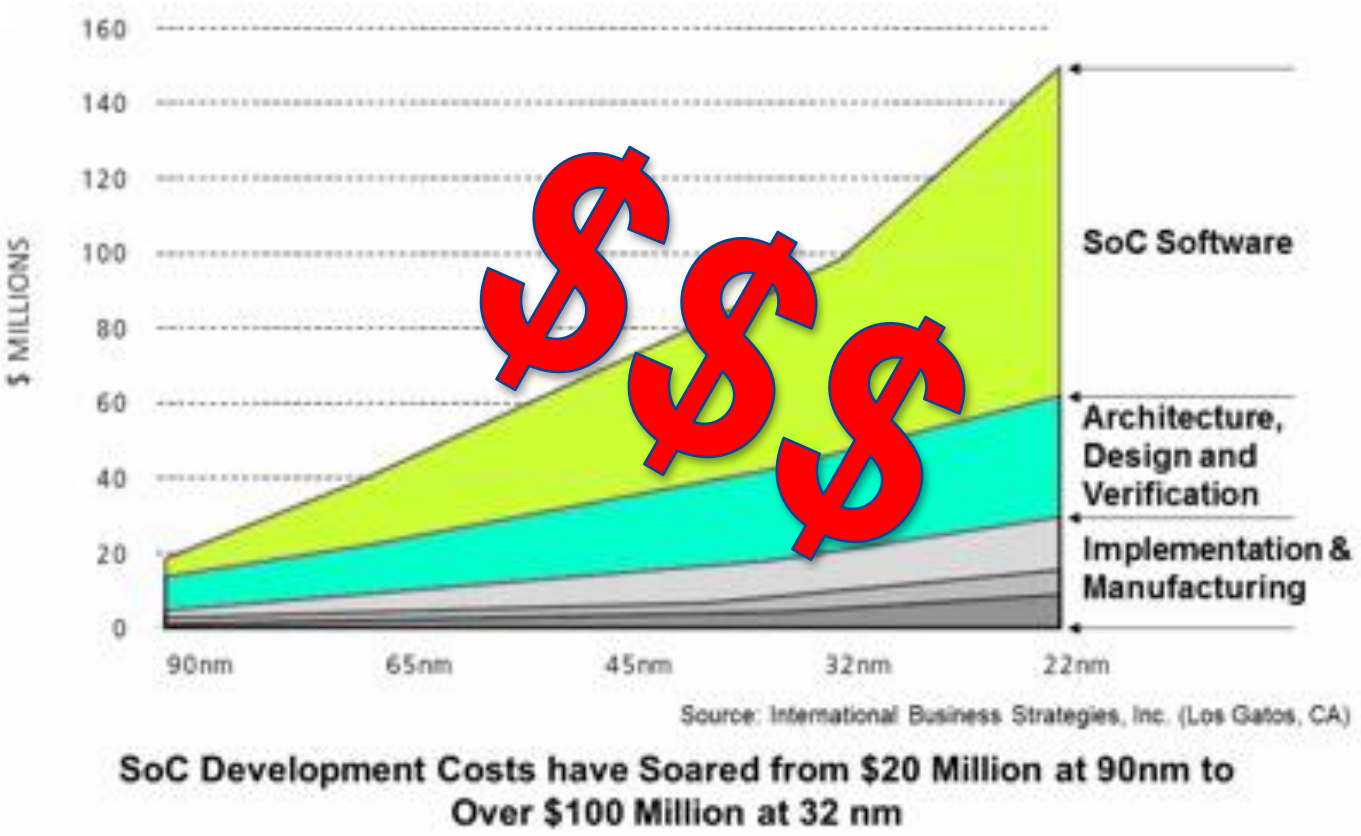


Source from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanovic

The cost per transistors is not decreasing anymore



And the development cost is increasing



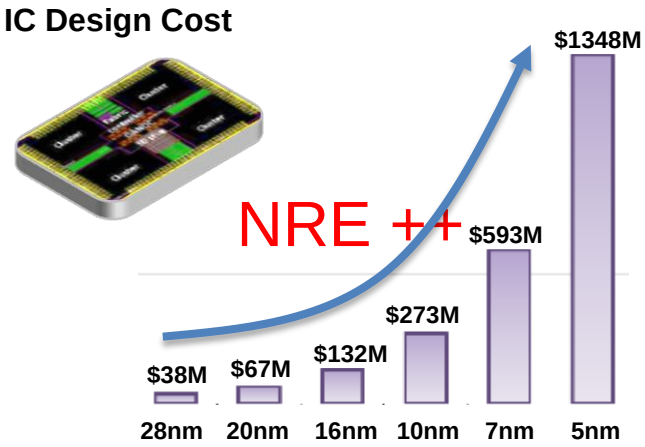
Rock's law: cost of IC plant doubles every 4 years

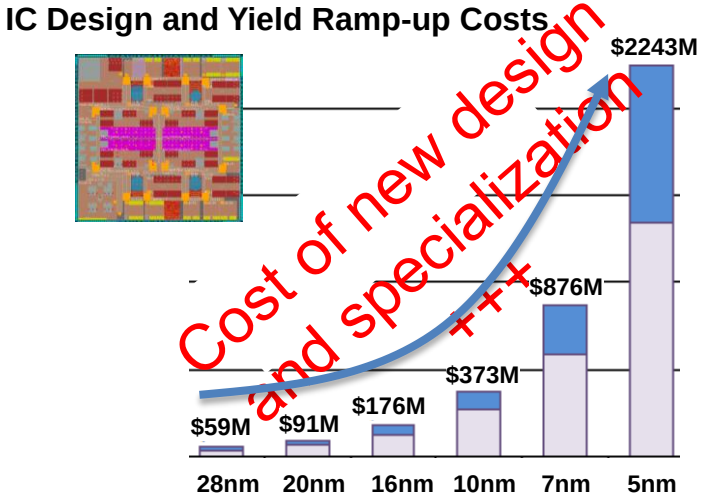
Reaching 10th of $ Billions…

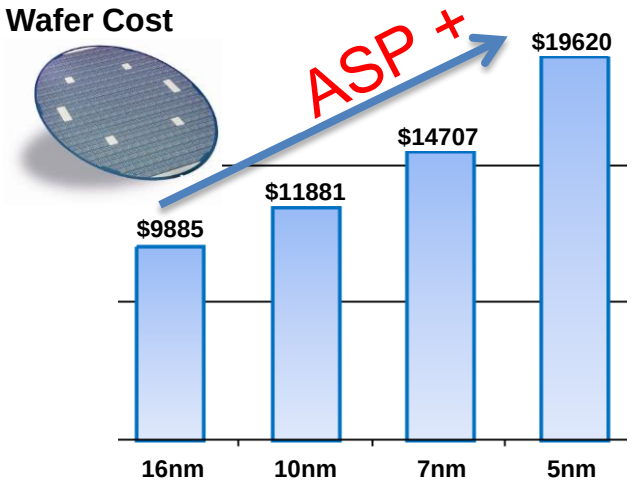
(Samsung will spend about $15 billion to open a new semiconductor factory in South Korea by 2017).

Increased Complexity and Cost

Source IBS, Aug. 2014



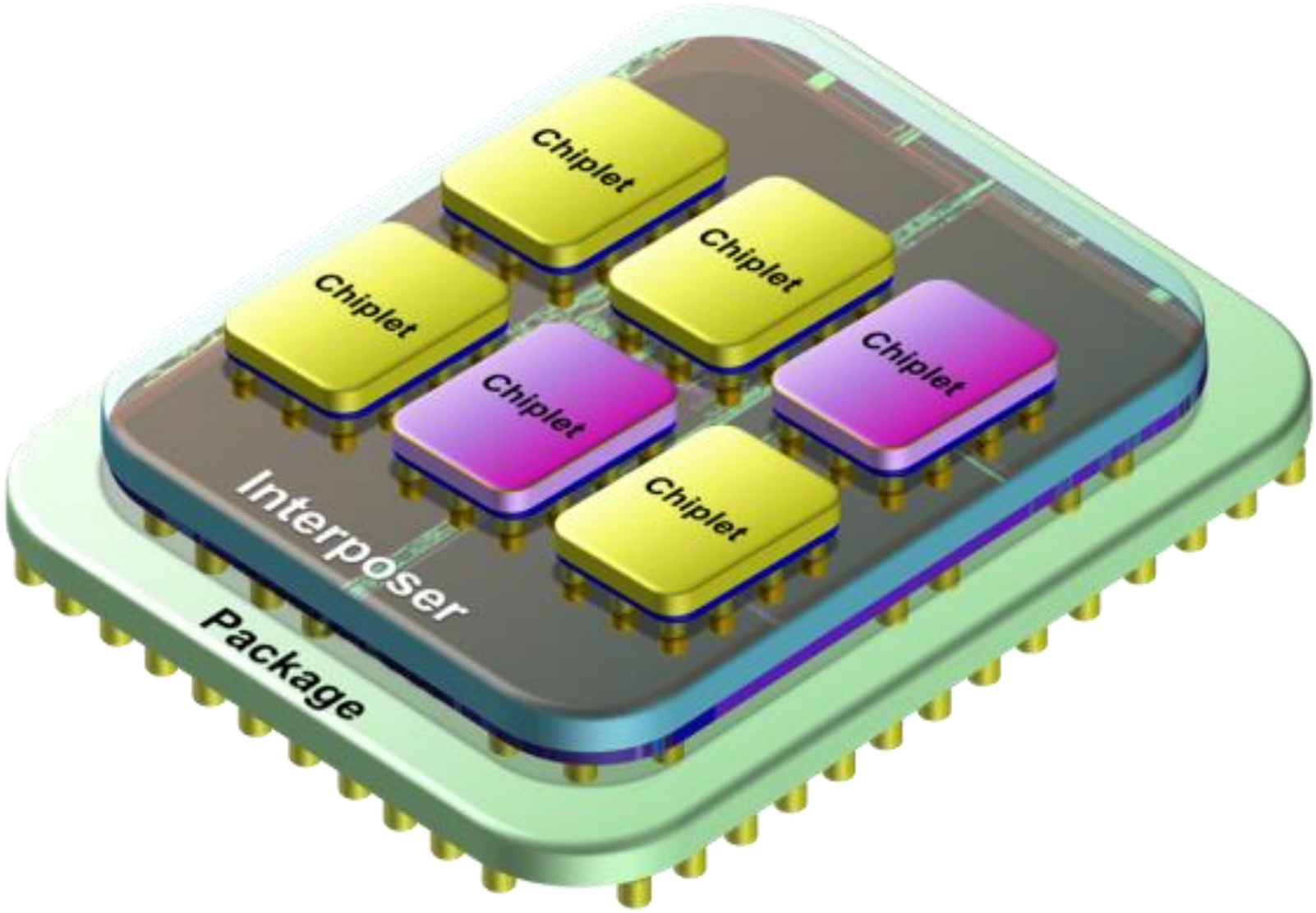




The initial product designs will need to generate high revenues to provide good buyback from the design and yield ramp-up costs.

- Barrier for ***specialization*** to computing
- Barrier for ***advanced feature monolithic*** dies

Specialization with interposer



Source: Denis Dutoit, CEA

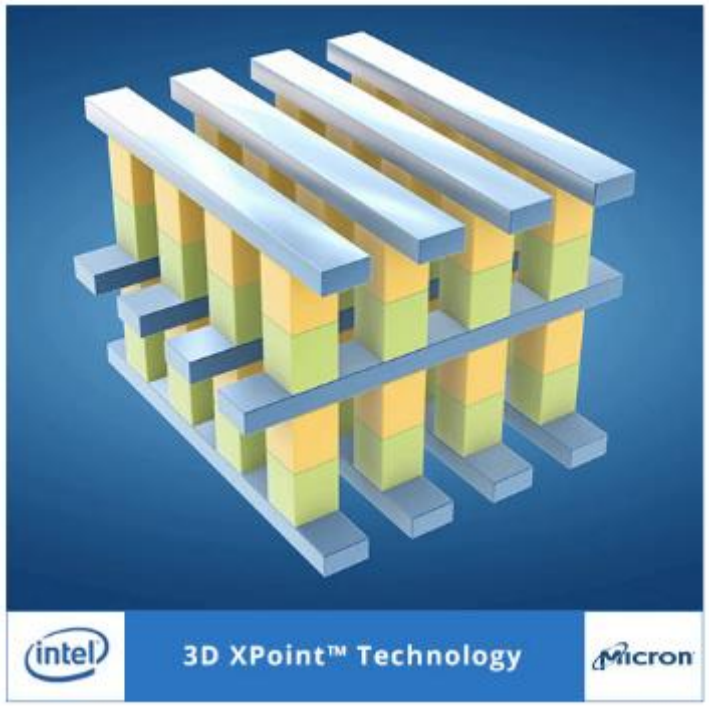
Where are we ?

We can still put more transistors per mm^2 for the coming few years

- But energy is a key limiting factor
  - New technologies (FinFet, FDSOI)
  - 3D stacking
  - More efficient architectures, coprocessors
- SRAM, DRAM didn't scale anymore
- Flash is running out of electrons
- Kryder's law for Hard Disk Drives (forecast 40% increase density per year, reality 15%)
- Non-volatile memories are promising
  - But which technology, at which density and reliability?

Intel and Micron Produce Breakthrough Memory Technology

Intel and Micron begin production on new class of non-volatile memory, creating the first new memory category in more than 25 years.



- New 3D XPoint™ technology brings non-volatile memory speeds up to ***1,000 times faster than NAND***, the most popular non-volatile memory in the marketplace today.
- The companies invented unique material compounds and a cross point architecture for a memory technology that is ***10 times denser than conventional memory***.
- New technology makes new innovations possible in applications ranging from machine learning to real-time tracking of diseases and immersive 8K gaming.

From: http://newsroom.intel.com/community/intel_newsroom/blog/2015/07/28/intel-and-micron-produce-breakthrough-memory-technology

Together…

Electrons for compute

Electrons like to interact; easily moved; interaction needed for compute

+ Ions for storage

Ions like to interact; stay put; good for storage

+ Photons to communicate

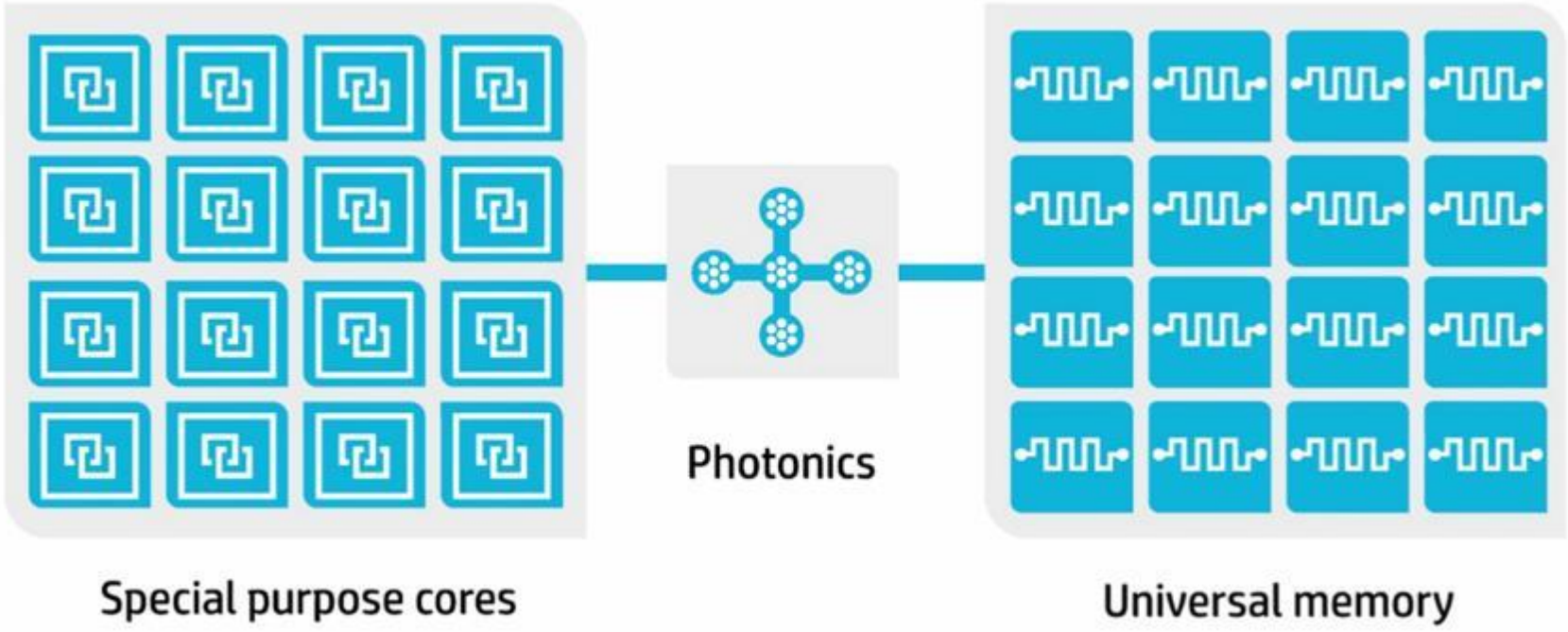
Photons don't like to interact or stay put; good for long-distances

See the presentation on "The Machine" from HP:
https://www.youtube.com/watch?v=JzbMSR9vA-c

Courtesy: Jouppi2011

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

HP and "The Machine" announced in June 2014



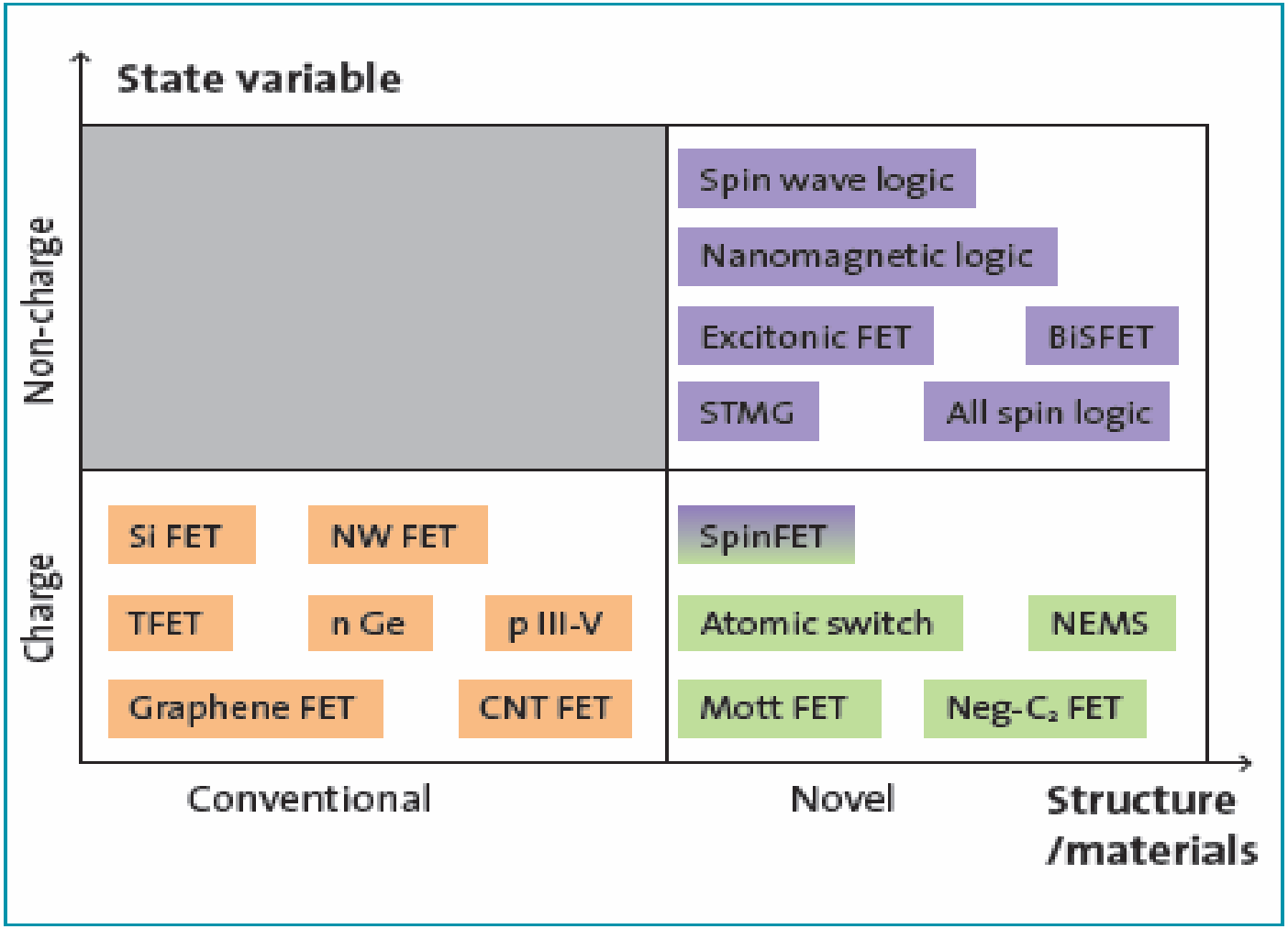
New memories could have a drastic impact on computing:

- **Memory hierarchy revisited…**
- **Files systems revisited…or *disappearing***
- **Pentabyte of storage in portable format…**

Open research areas…

- "Computing in memory"
- Streaming (processing while communicating)
  - Stream analytics
- Reconfigurable computing (Intel + Altera, cf. Microsoft and accelerating "Bing" searches")
- New computing paradigms…
  - Non-Von Neumann
  - Adapted to application domains:
    - *Natural signal processing*: Neural Networks
    - *Optimization*: "Quantum computer" à la D-Wave
- And not only in silicon…

New emerging technologies for computing

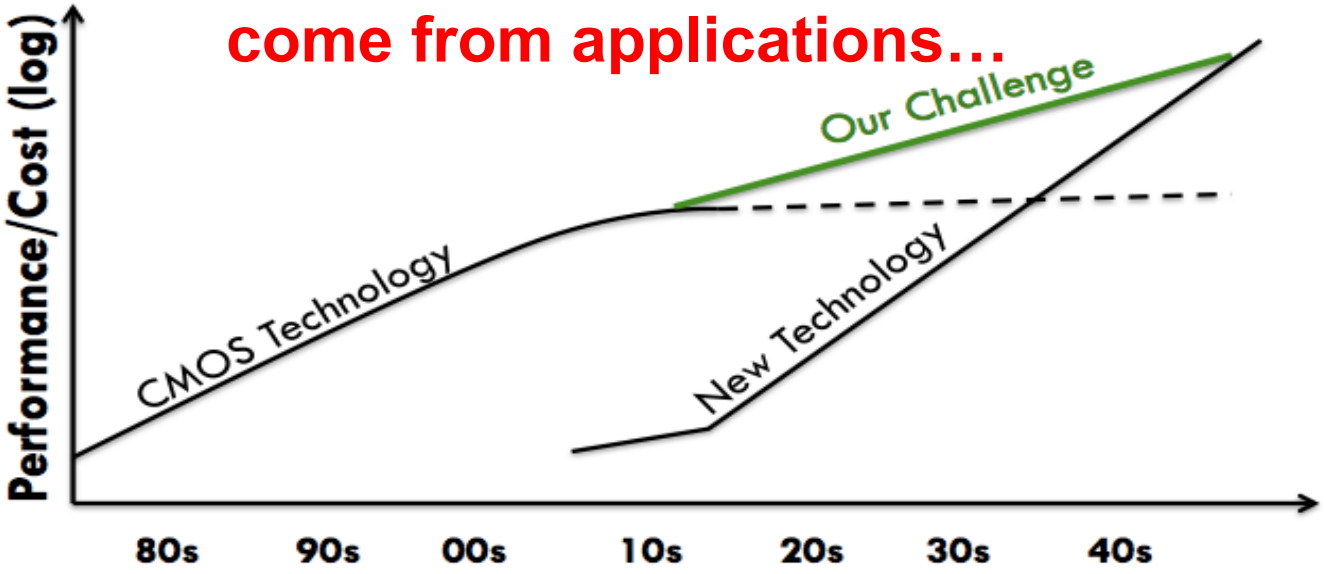


Taxonomy for emerging information processing devices (from [ITRS2013]).

We are entering into a transition period…

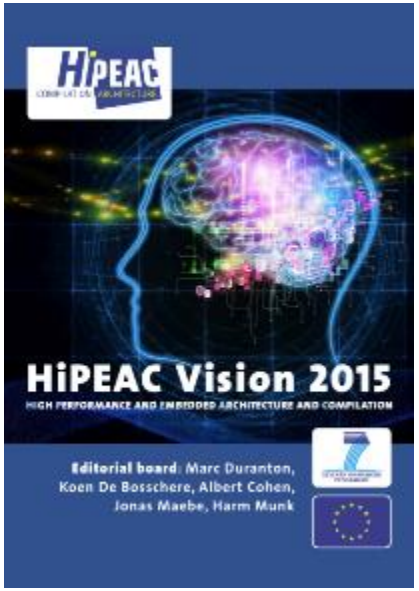
Scaling without Technology Help

But even more challenges come from applications…



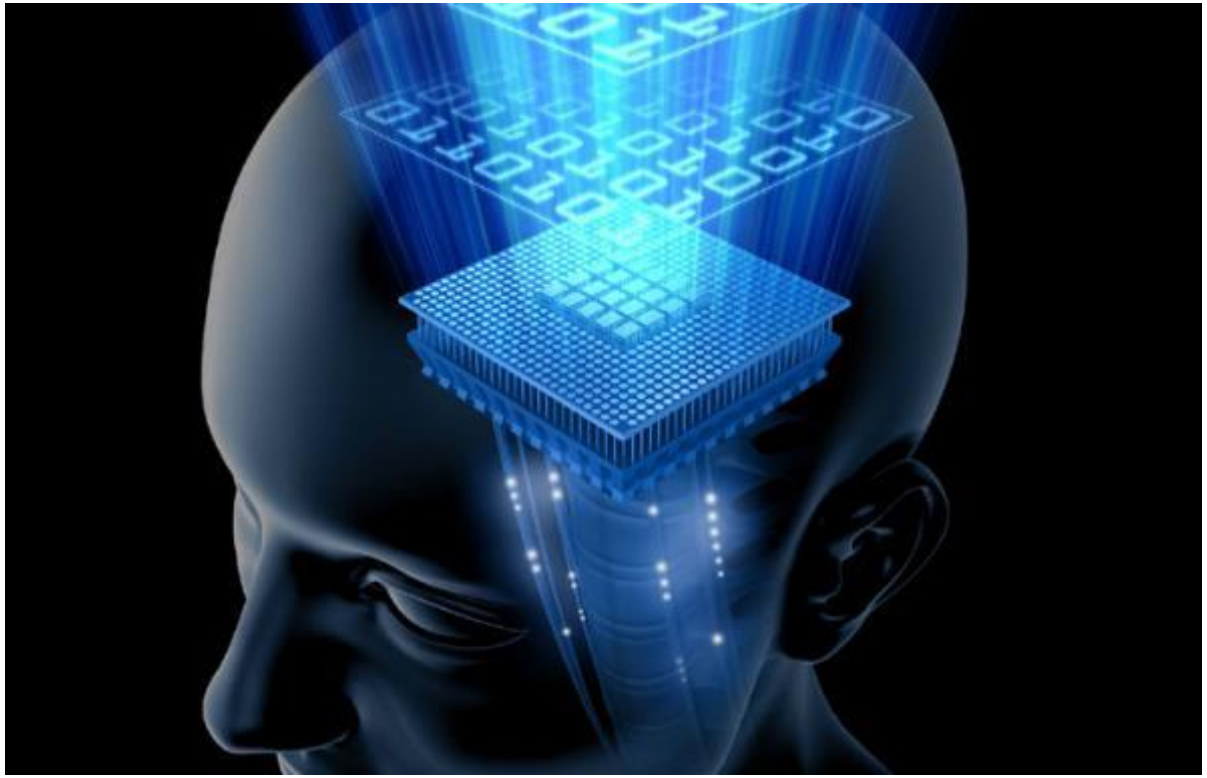
[Hill & Kozyrakis'12]

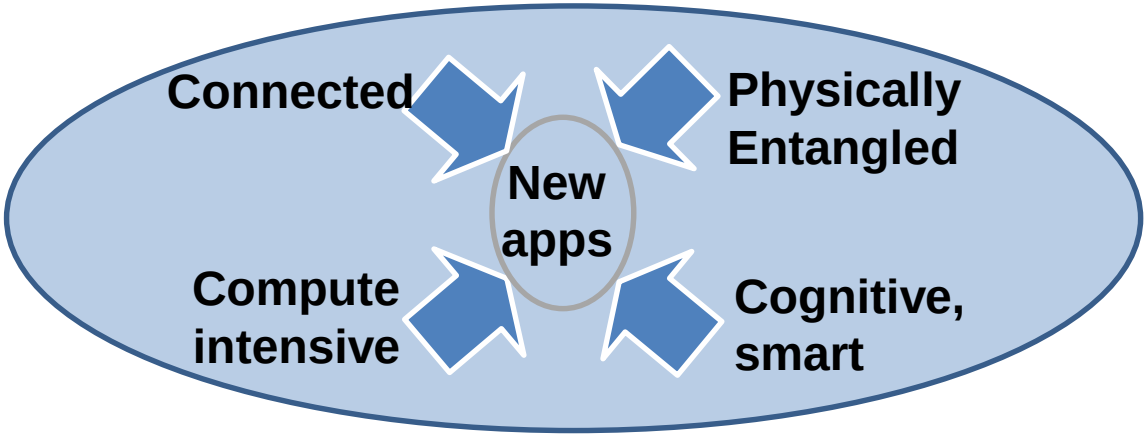
The End of the World As We Know It…



From the applications ecosystem…

New apps will be…





Global integration of communication, computation and reaction

Computer

"Classical" computing

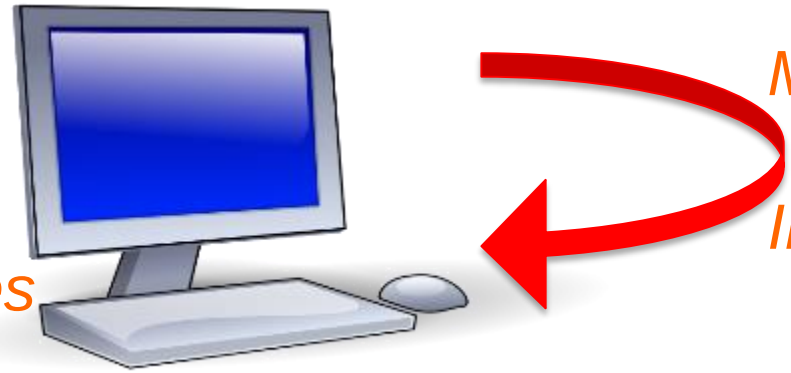
CyberPhysical Systems

Human ⟷ **World**

Before computing era

Global integration of communication, computation and reaction

Applications are delocalized, distributed on collaborating devices



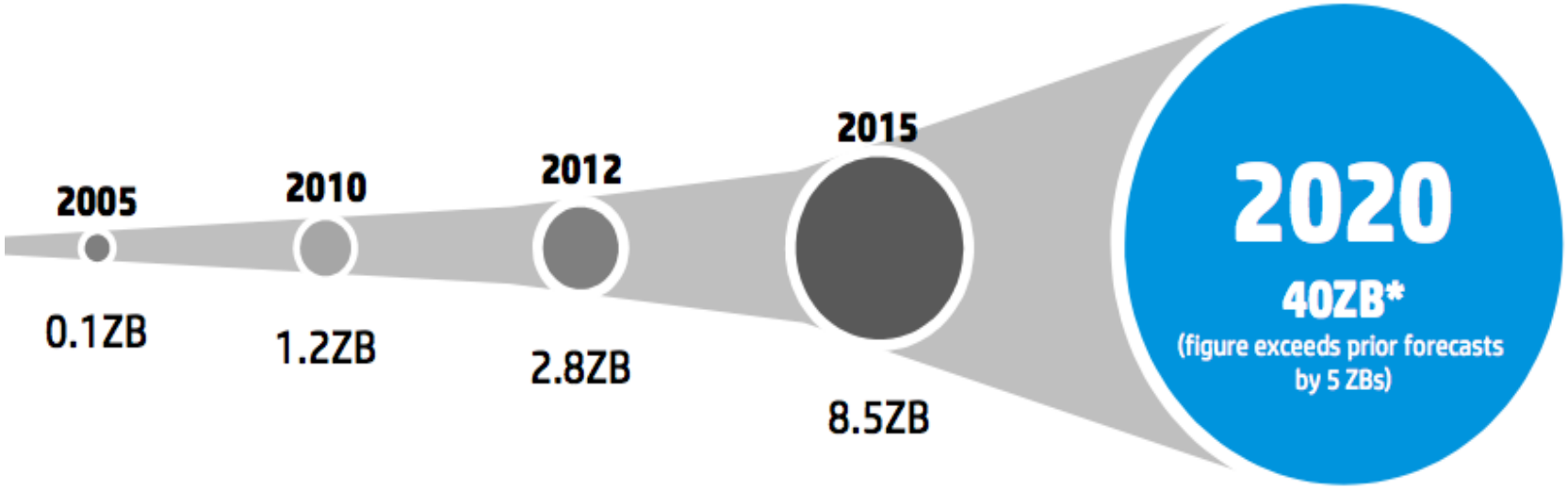
Machine to Machine Interactions

Computer

Human ⟷ **World**

Constraints of the real world e.g. time, …

The data deluge challenge



1 ZB = 10^{21} bytes

40 ZB is equal to 57 times the amount of all the grains of sand on all the beaches on earth.

Source: Paolo Faraboschi, HP, and IDC

IoT: the Internet of Threats

- Today security / privacy issues make the newspaper headlines

Massive **adoption of IoT by citizens** relies on confidence in terms of **security and privacy**

Misuse of information technology might destroy our privacy

Snowden effect

Heartbleed bug - OpenSSL

Internet of things

- Consumers give away private information for free services
- Companies do so for free software (e.g. Android)

Global integration of communication, computation and reaction






Smart sensors

Physical Systems

Internet of Things

Processing, Abstracting Understanding *as early as possible*

Cloud / HPC

Big Data

Cognitive computing / Data Analytics



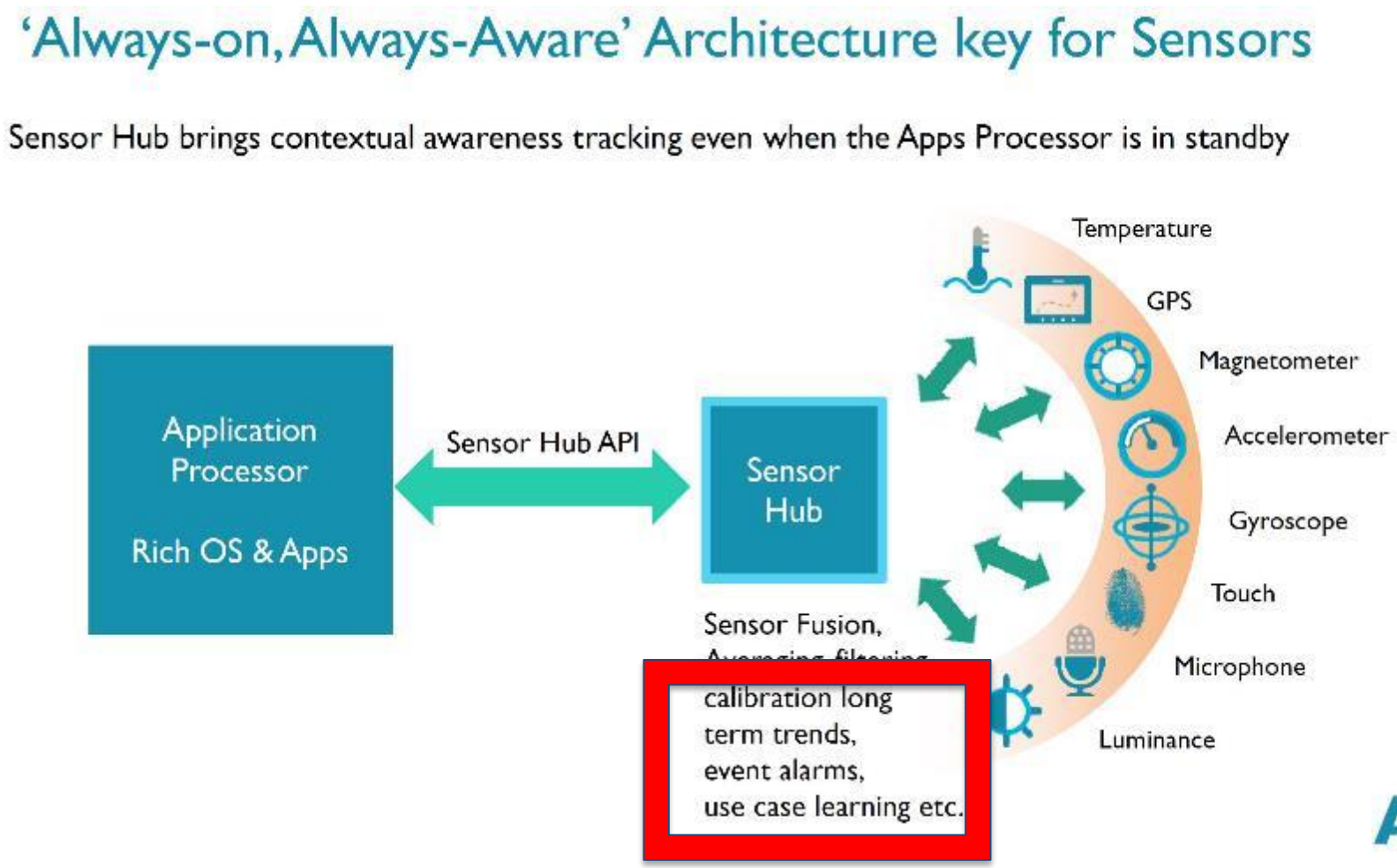


New services

Fog computing
Edge computing
Stream analytics
Fast data…
by real-time
micro-servers
and even
Nano-servers
(concentrator, fusion of several sensors)

Transforming **data** into ***information***

Exemple of architecture for end-nodes



Source: http://www.anandtech.com/show/8542/cortexm7-launches-embedded-iot-and-wearables

Computing becomes increasingly cognitive

- Cognitive computing (IBM)
  - Artificial intelligence meets business intelligence
  - Systems with domain expertise
  - Humans and machines working together
- Deep Learning Systems
  - Google, Facebook, Baidu, etc
  - Use for image recognition, voice…

- Application examples
  - Self-driving car
  - Automatic translation
  - Natural language understanding & reasoning (Watson)

- New workload -> new computing platforms (new accelerators, reconfigurable computing, bio-inspired, …)
- How to “program” it?

Watson from IBM, “cognitive computer”

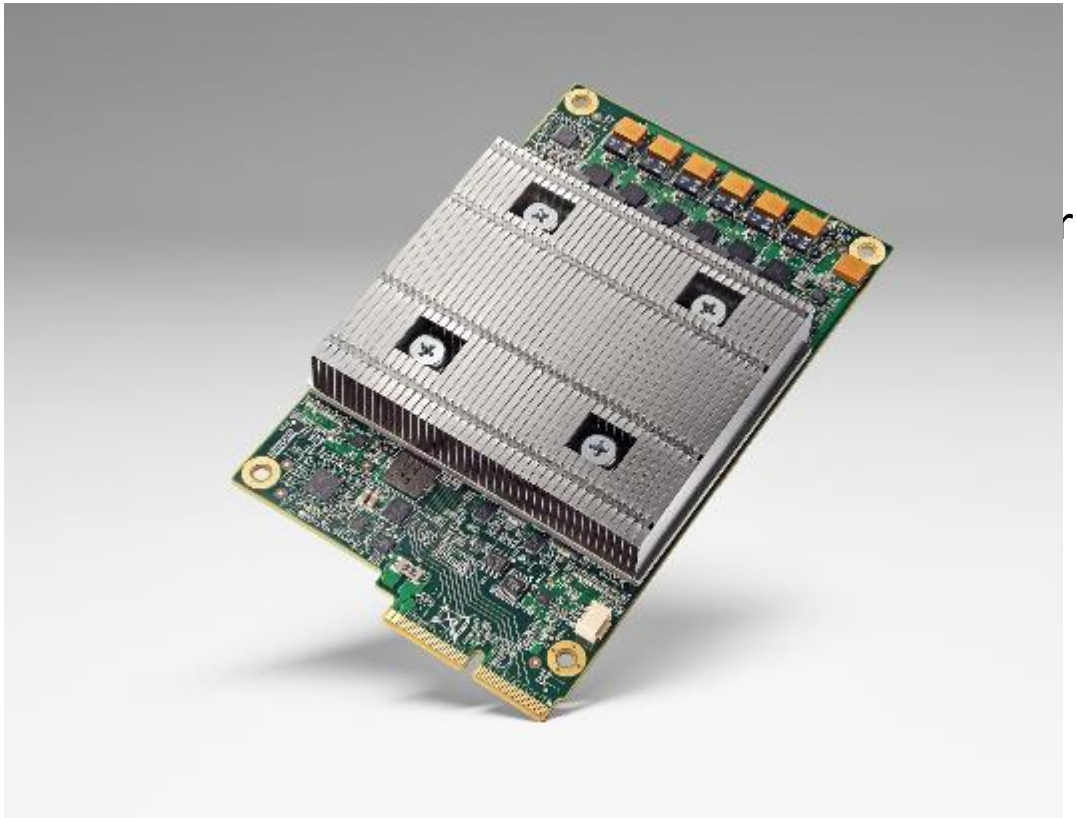
Deep Learning is Everywhere (ConvNets are Everywhere)

Y LeCun

- **Lots of applications at Facebook, Google, Microsoft, Baidu, Twitter, IBM…**
  - Image recognition for photo collection search
  - Image/Video Content filtering: spam, nudity, violence.
  - Search, Newsfeed ranking
- **People upload 800 million photos on Facebook every day**
  - (2 billion photos per day if we count Instagram, Messenger and Whatsapp)
- **Each photo on Facebook goes through two ConvNets within 2 seconds**
  - One for image recognition/tagging
  - One for face recognition (not activated in Europe).
- **Soon ConvNets will really be everywhere:**
  - self-driving cars, medical imaging, augemnted reality, mobile devices, smart cameras, robots, toys…..

Google's Tensor Processing Unit (TPU)

- AlphaGo was powered by TPUs in the matches against Go world champion, Lee Sedol.



Economical drive for data analytics

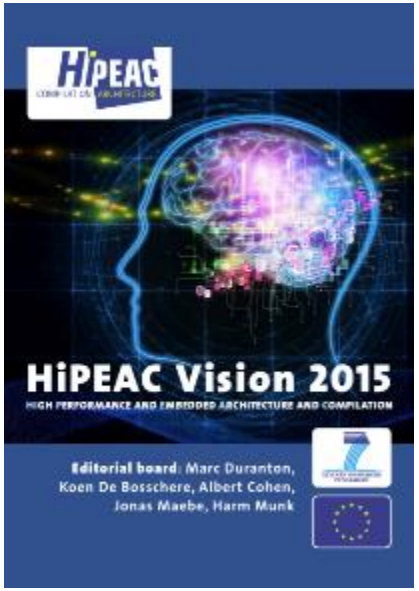
"The Power of 1 Percent"

What if... Potential Performance Gains in Key Sectors

Industry	Segment	Type of Savings	Estimated Value over 15 Years (Billion nominal US dollars)
Aviation	Commercial	1% Fuel Savings	$30B
Power	Gas-fired Generation	1% Fuel Savings	$66B
Healthcare	System-wide	1% Reduction in System Inefficiency	$63B
Rail	Freight	1% Reduction in System Inefficiency	$27B
Oil & Gas	Exploration & Development	1% Reduction in Capital Expenditures	$90B

Note: Illustrative examples based on potential one percent savings applied across specific global industry sectors.
Source: GE estimates

The End of the World As We Know It…



The "software crisis"…

Software crisis

- The productivity challenge
  - Better tools and languages… also supporting legacy
- The correctness challenge (non-functional requirements)
  - portability, time (for CPS systems), accuracy
- The performance challenge
  - Modern abstractions prohibit performance optimizations
- The data challenge
  - Size (big data), security, integrity
- The holistic challenge
  - Global optimizations

Goal: dependable or trustable software

How to ensure software (and systems) that are:

- **Safe:** system operating without causing unacceptable risk of physical injury or damage to the health of people, either directly, or indirectly as a result of damage to property or to the environment.
- **Secure:** system keeping integrity, availability, confidentiality and privacy.
- **Reliable**: ensure good behavior under variable conditions, including ageing

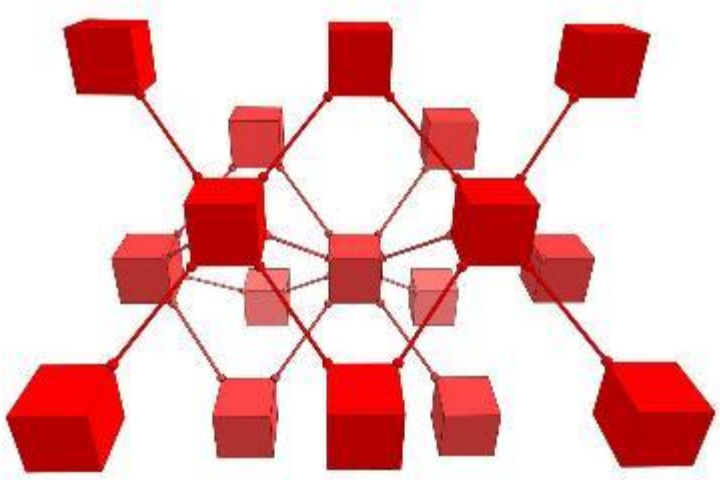
- How to ensure these properties, and correctness of the results for **reactive systems, distributed systems**, etc…

We need you

to find solutions!

Managing complexity….

"Nontrivial software written with threads, semaphore, and mutexes is ***incomprehensible*** *by humans"*



Edward A. Lee

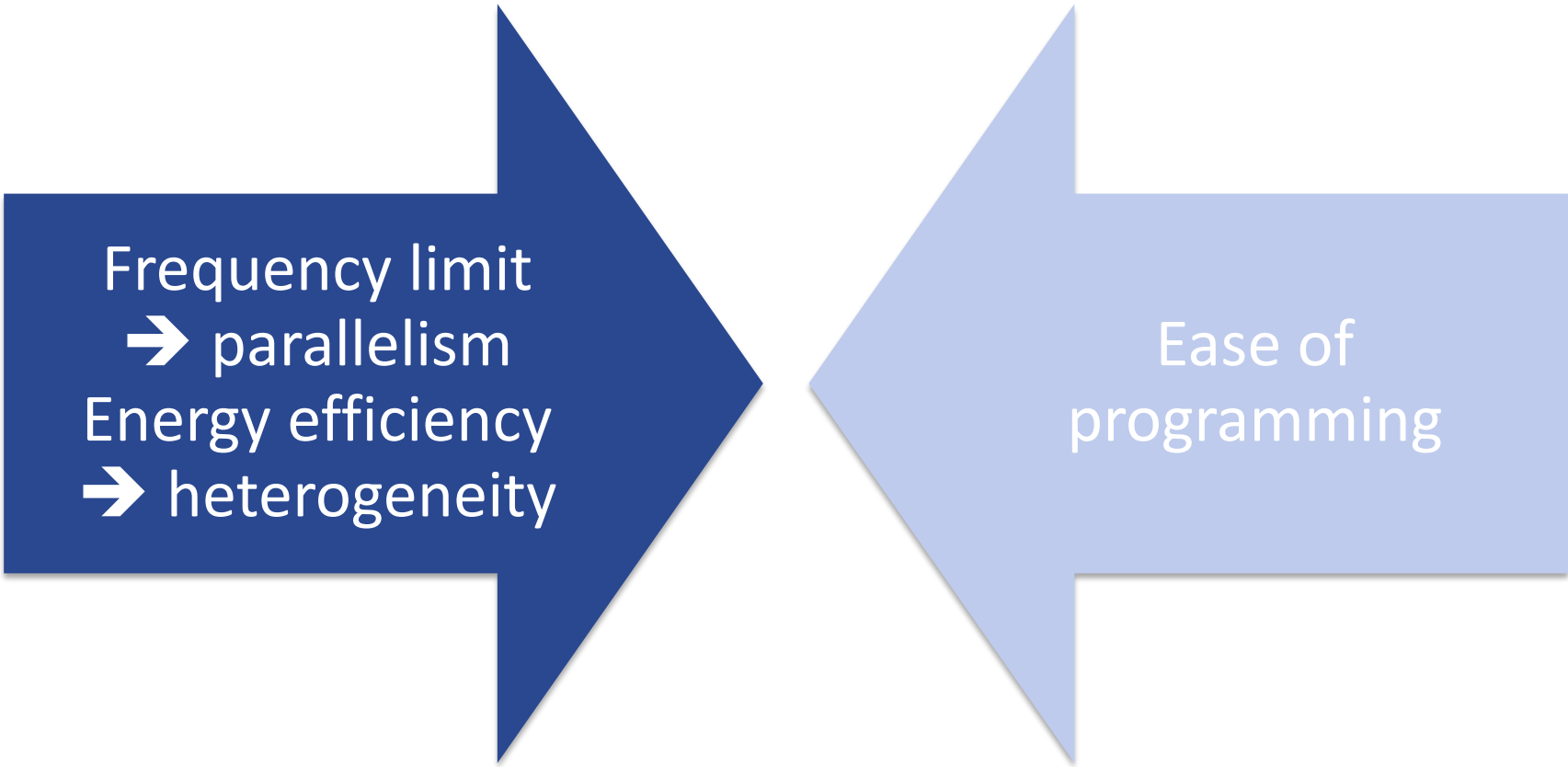
The future of embedded software
ARTEMIS 2006

Parallelism, multi-cores, heterogeneity, distributed computing, seems to be too complex for humans ?

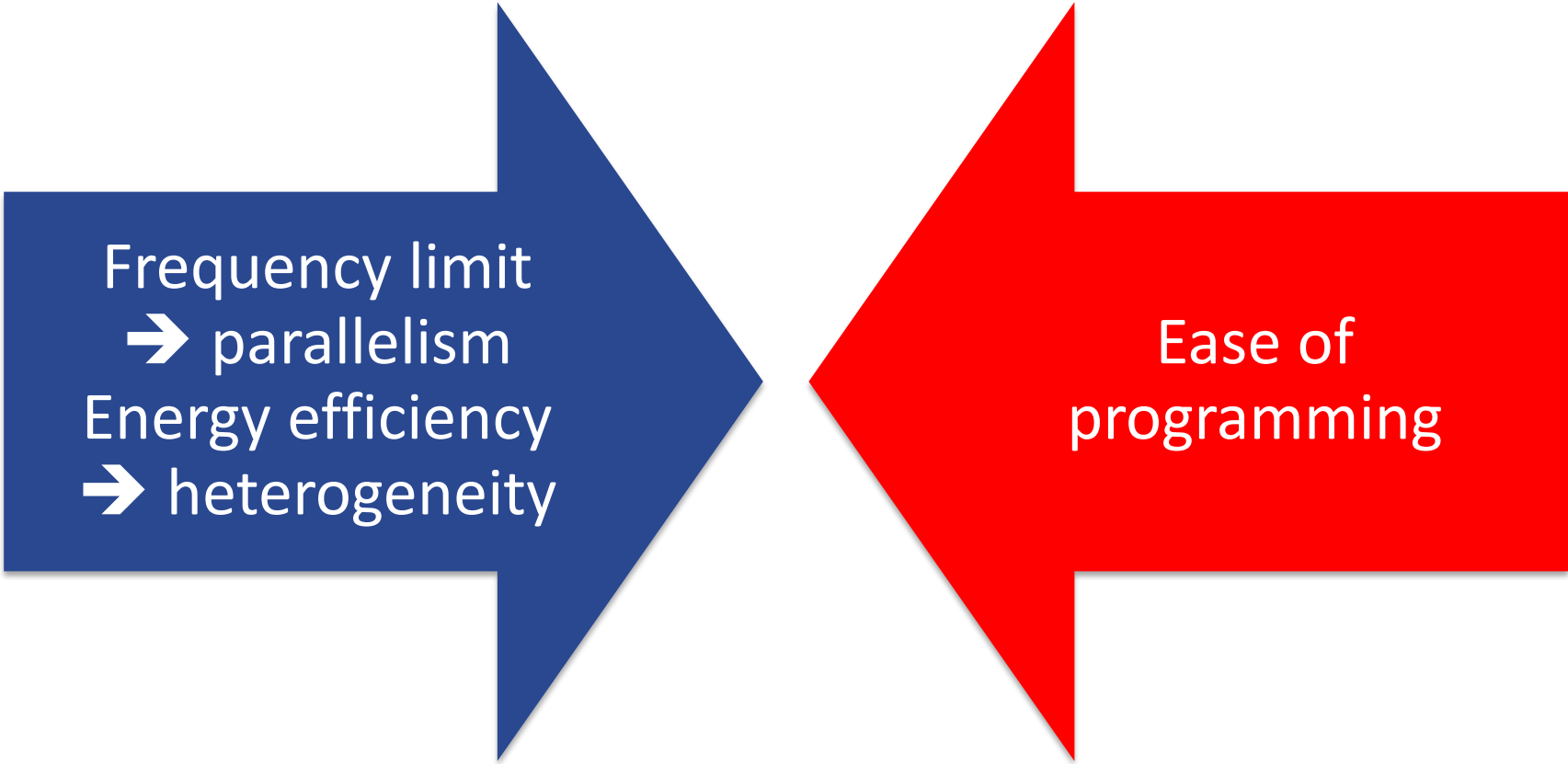
```
public synchronized void addChangeListener(ChangeListener listener) {
    NamedObj container = (NamedObj) getContainer();
    if (container != null) {
        container.addChangeListener(listener);
    } else {
        if (_changeListeners == null) {
            _changeListeners = new LinkedList();
            _changeListeners.add(0, listener);
        } else if (!_changeListeners.contains(listener)) {
            _changeListeners.add(0, listener);
        }
    }
}
```

A Story: Ptolemy Project Code Review Introduced Deadlock

Parallelism and specialization are not for free…



Parallelism and specialization are not for free…



More and more black/grey boxes

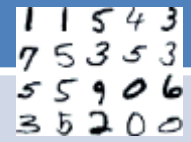
- Complete applications are distributed onto different (distant) hardware
- Only part of the software is available in source form for the developer
- Programming through API or binary libraries
  - Success of Python, interpreted shell, GUI, etc
- Everything as a service…
- More and more assembling high level functions which source code is unavailable
- Problem of validation and test…

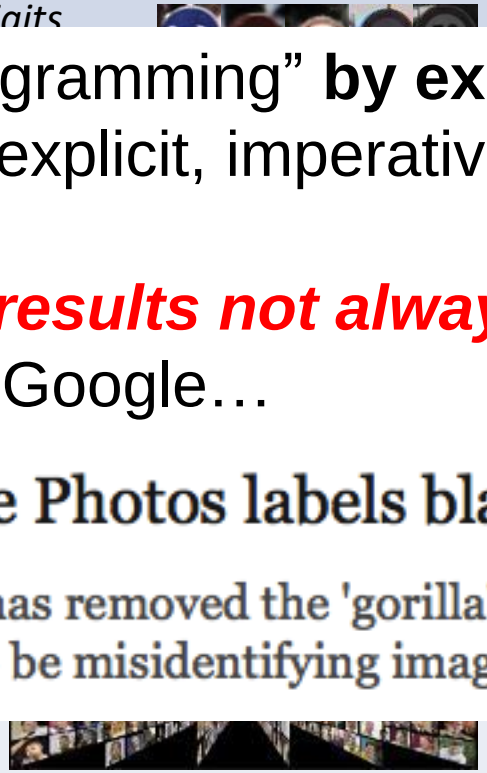
Quality of experience is key

- Software often over-constraint: e.g. highest precision is not always required
- By lowering the precision requirements, power can be saved.
- Challenges
  - How to specify the precision requirements?
  - How to specify a HW/SW interface to control the precision
  - New algorithms?
  - How to ensure the correctness for the application?

Deep Neural Networks: state-of-the-art in image recognition…But

Database	# Images	# Classes	Best score
MNSIT *Handwritten digits*	60,000 +	10	99.79% [3]
GTSRE *Traffic*		43	99.46% [4]
CIFAR- *airplan* *deer, d*		10	91.2% [5]
Caltec		101	86.5%
Im			
De			[2]





"Programming" **by example**,
Not explicit, imperative programming

But ***results not always guaranteed***…
e.g. Google…

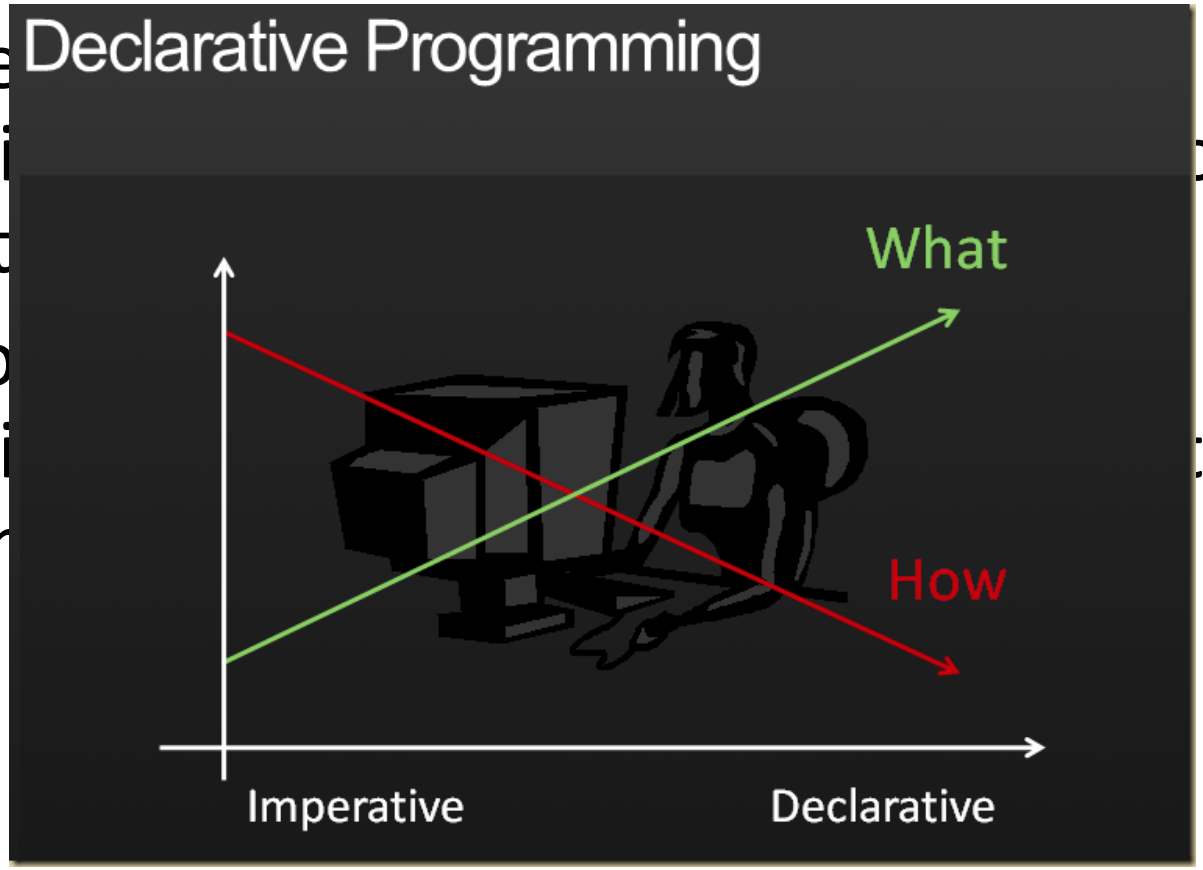
Google Photos labels black people as 'gorillas'

Google has removed the 'gorilla' tag from its new Photos app, after it was found to be misidentifying images of black people

- **State-of-the-art are Deep Neural Networks *every time***

Let the computer do the job:

- Describing ***what*** the program should accomplish, rather than describing ***how*** to accomplish it as a sequence of the programming language primitives.
- For e of an appli code for it
- (Goo tter at archi optir



Hardware design is also software…

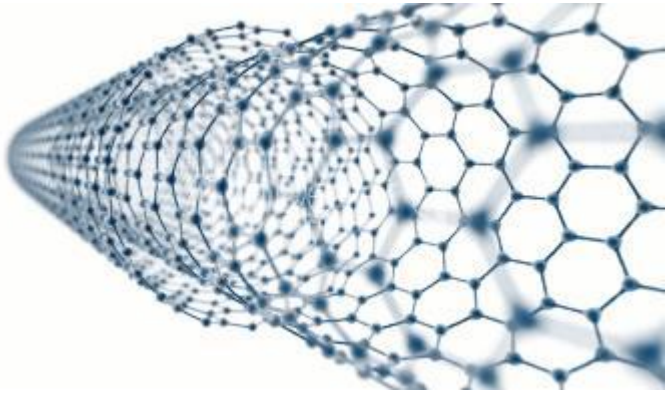
Formal specifications, model-driven design
Stateflow, StateCharts, LUSTRE, ….

Intermediate code
(C, C++, SystemC…)

Computer processing
(HLS, core generators
…), …)

Conclusion: *What should we do* (as HiPEAC)?

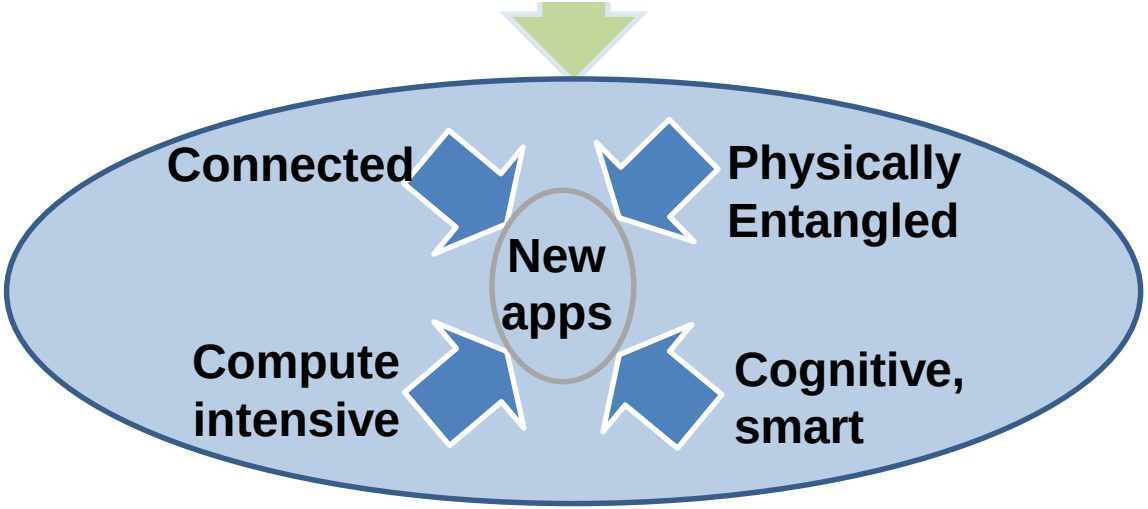
(From the HiPEAC vision 2015)

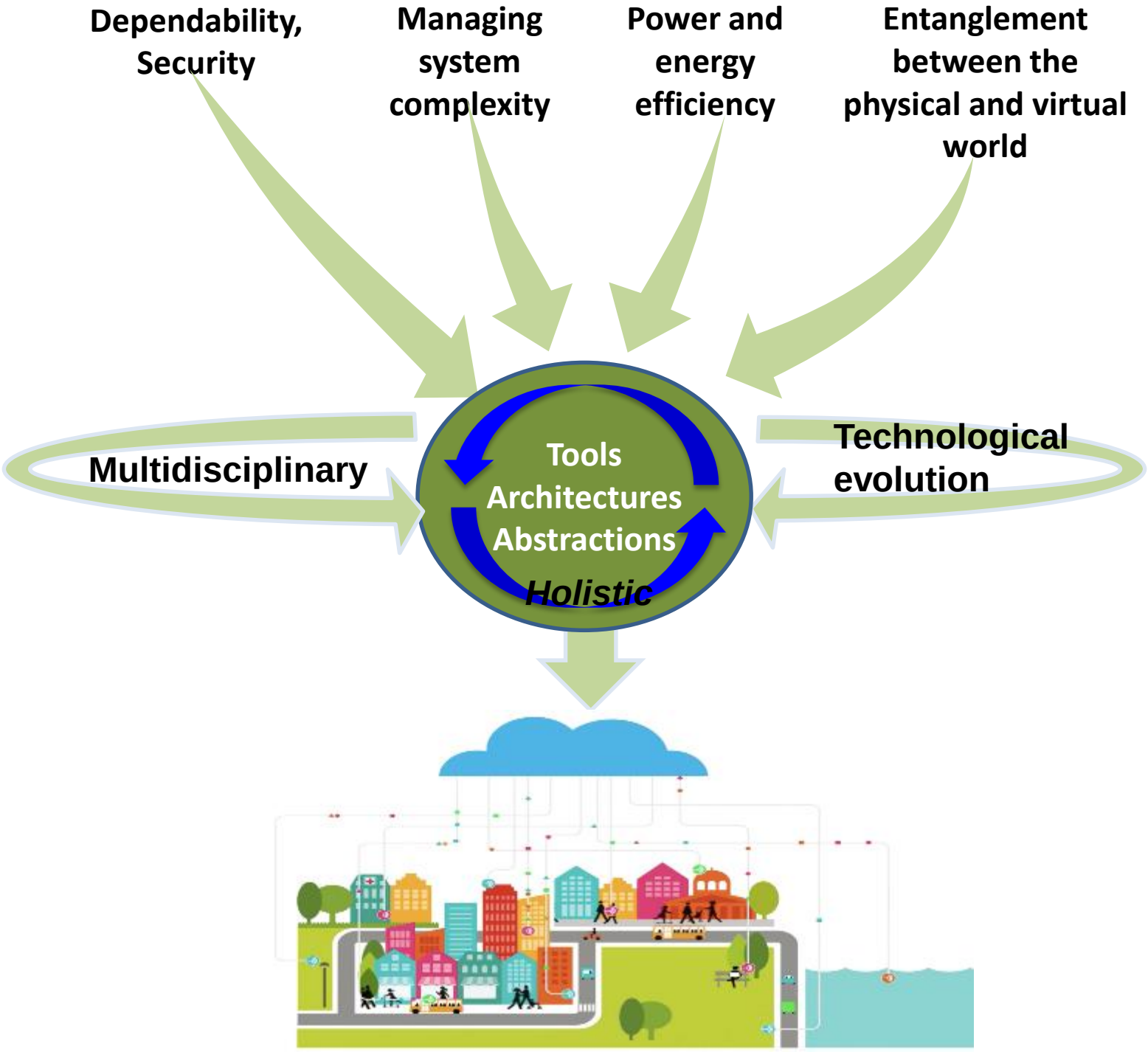


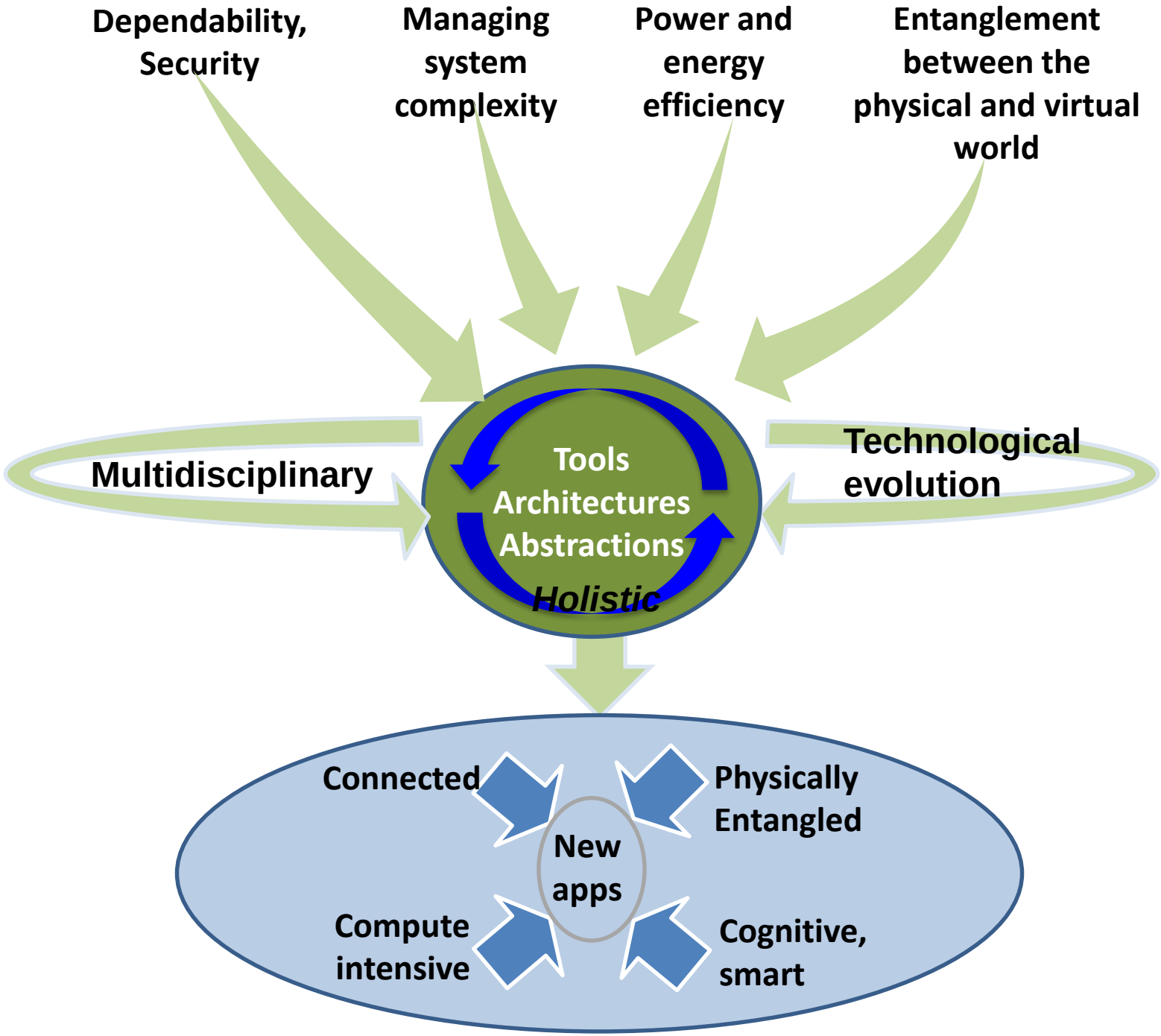
Multidisciplinary

Technological evolution

HiPEAC COMPILATION ARCHITECTURE




Dependability, Security
Managing system complexity
Power and energy efficiency
Entanglement between the physical and virtual world
Multidisciplinary
Technological evolution
Tools
Architectures
Abstractions
Holistic

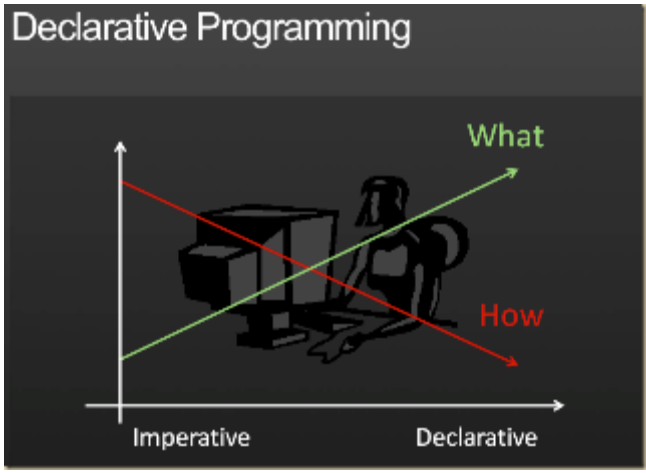

Dependability, Security
Managing system complexity
Power and energy efficiency
Entanglement between the physical and virtual world
Multidisciplinary
Tools
Architectures
Abstractions
Holistic
Technological evolution
Connected
Physically Entangled
New apps
Compute intensive
Cognitive, smart

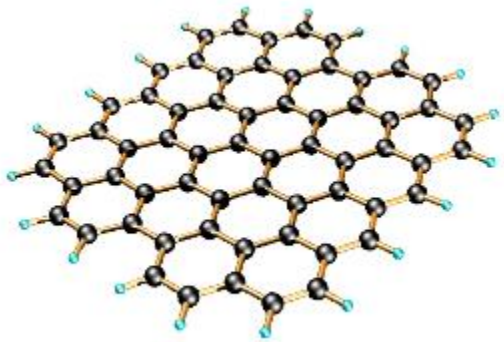
Highlights of the **HiPEAC Vision** 2015

Video available at
https://www.hipeac.net/publications/vision/

Time to think differently?

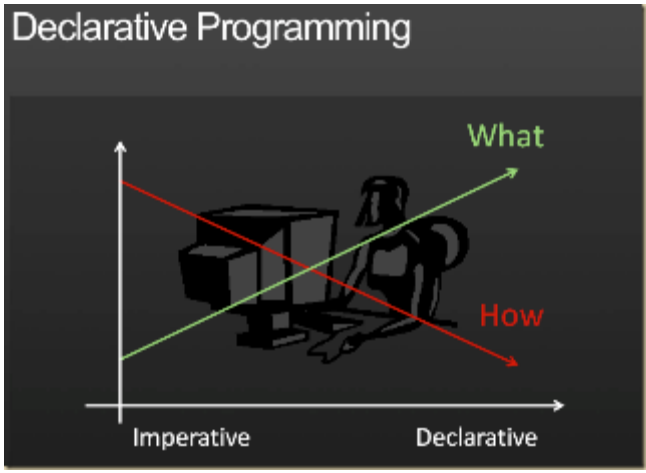
- Approximate computing
- Cognitive computing
- Neuromorphic computing
- Declarative programming
- New computing technologies
  - Graphene
  - Spintronic
  - Quantum…

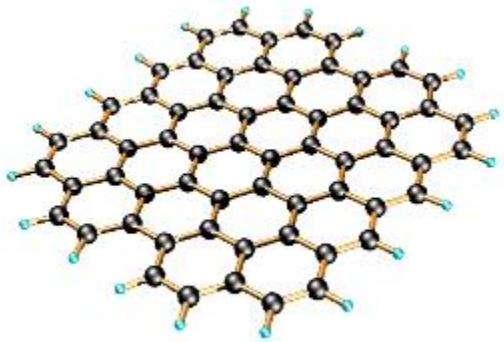




Time to think differently?

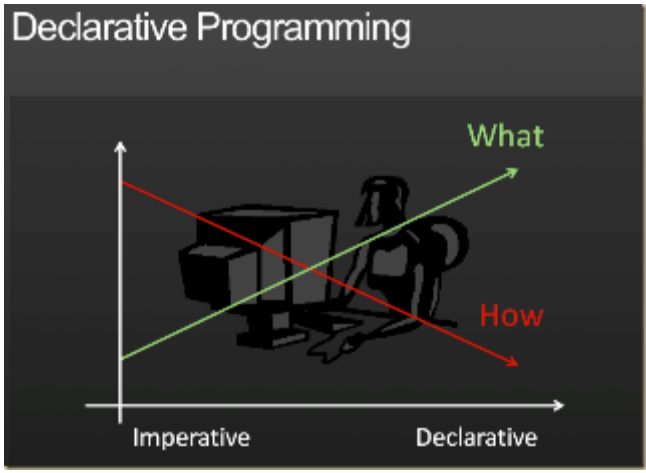
- ~~Approximate~~ computing
- Cognitive computing
- Neuromorphic computing
- Declarative programming
- New computing technologies
  - Graphene
  - Spintronic
  - Quantum…

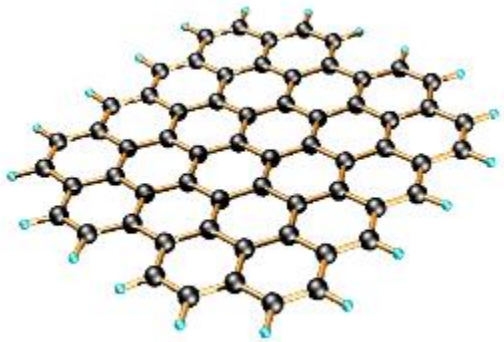




Time to think differently?

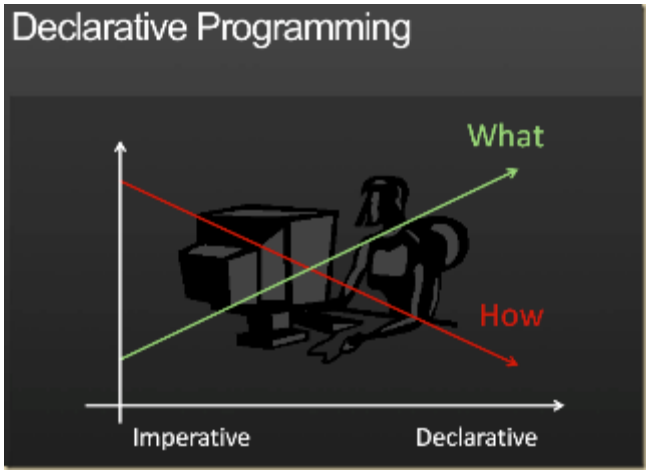
- Adequate computing
- Cognitive computing
- Neuromorphic computing
- Declarative programming
- New computing technologies
  - Graphene
  - Spintronic
  - Quantum…

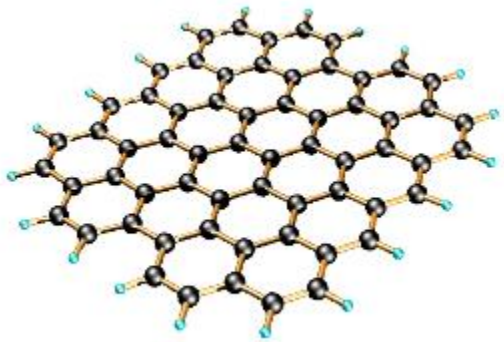


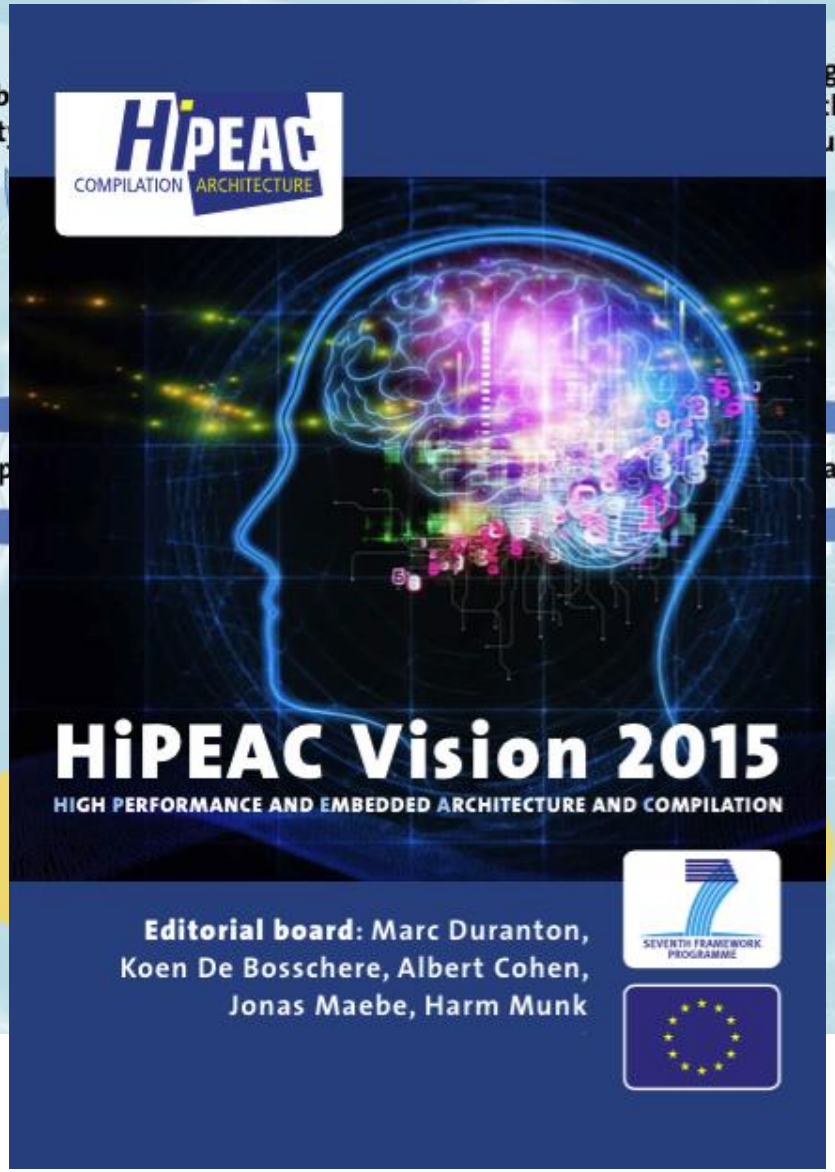


Time to think differently?

- Adequate computing
- Cognitive computing
- Neuromorphic computing
- Declarative programming
- New computing technologies
  - Graphene
  - Spintronic
  - Quantum…

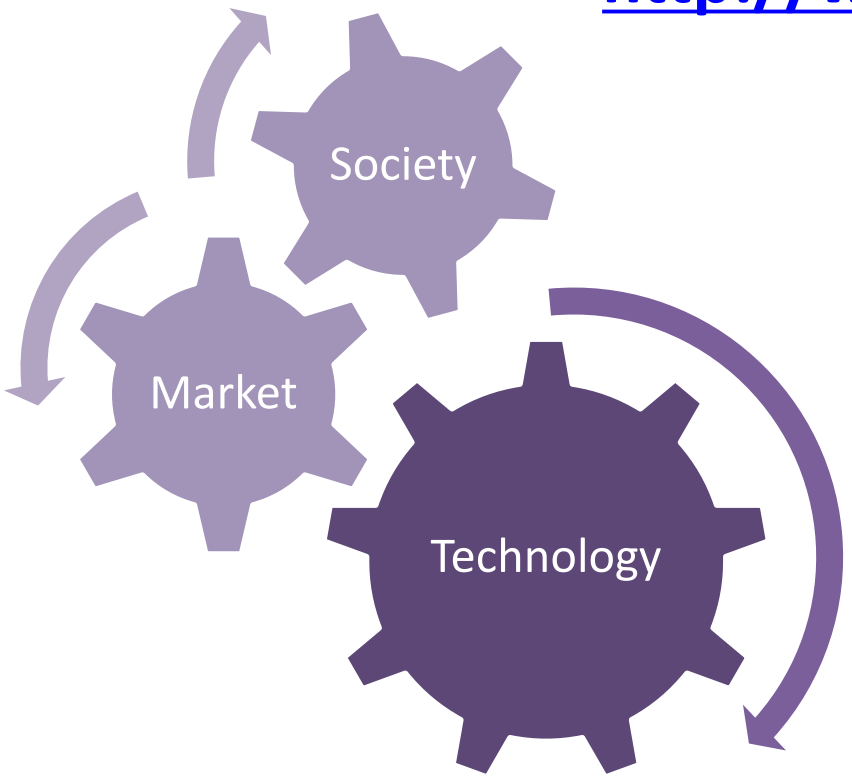


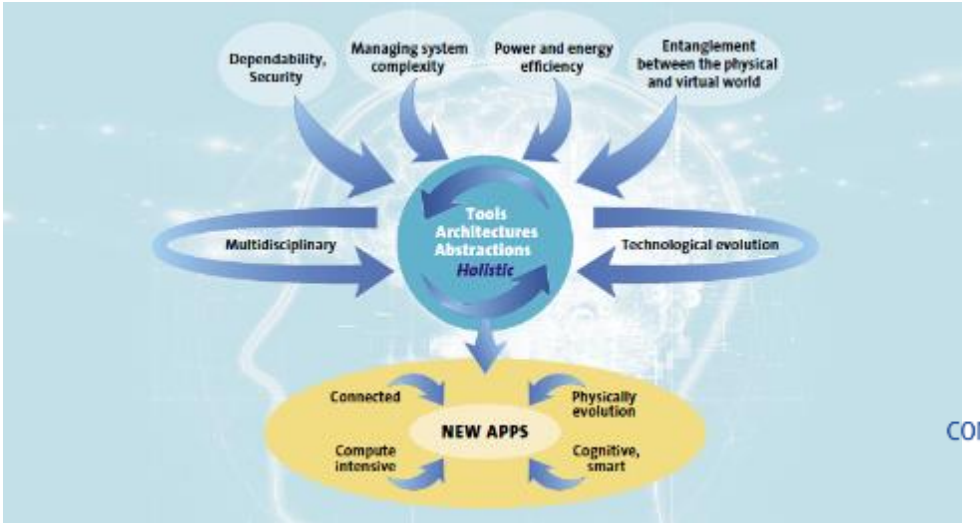




http://www.hipeac.org/vision

http://www.hipeac.org/vision





COMPILATION ARCHITECTURE