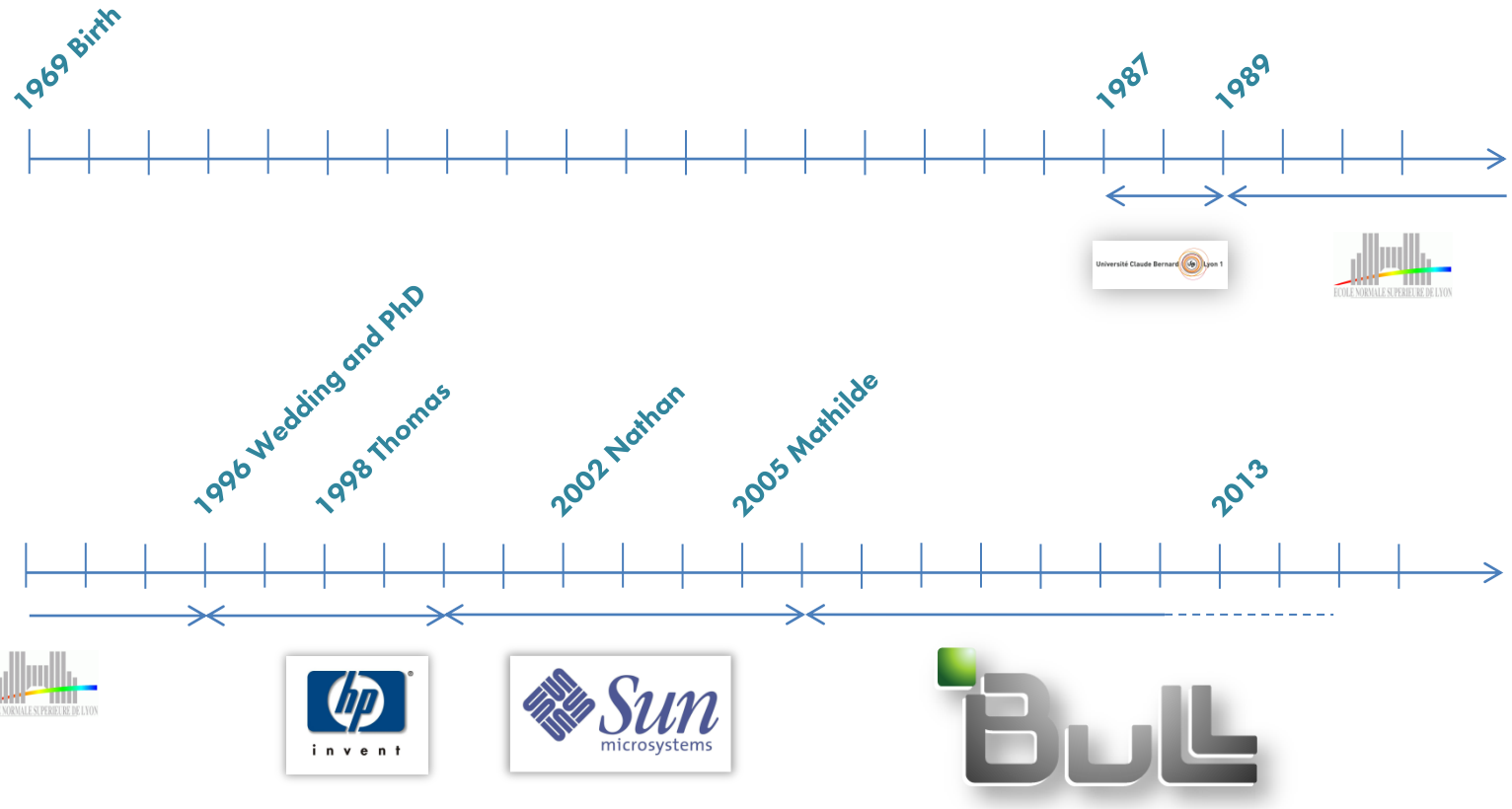
Data Centers

Current trends and impacts

Xavier VIGOUROUX
HOSCAR Workshop
Bordeaux
Sept. 2, 2013

WHO AM I


1969 Birth
1987
1989
1996 Wedding and PhD
1998 Thomas
2002 Nathan
2005 Mathilde
2013
Université Claude Bernard Lyon 1
ECOLE NORMALE SUPERIEURE DE LYON
hp invent
Sun microsystems
BULL

A VERY QUICK SUMMARY

Bef. '95	'95 – '05	Aft. '05
• Many architectures: RISC, MIPS, SIMD, MIMD ... • Many topologies: TORUS, Hypercubes... • Exemples: Cray T3E, Thinking Machine CM-5, Intel Paragon, MasPar MP-1, Parsytec Gcel	• Industrialization • Productivity • Performance gains were seamless	• Power constraint • No more ILP to mine • Complex memory Hierarchy
Hard time for software developers No Software compatibilities (PVM was just appearing)	**Good time for software developers** Better Performance by just waiting	**Hard time for software developers** Getting performance requires work

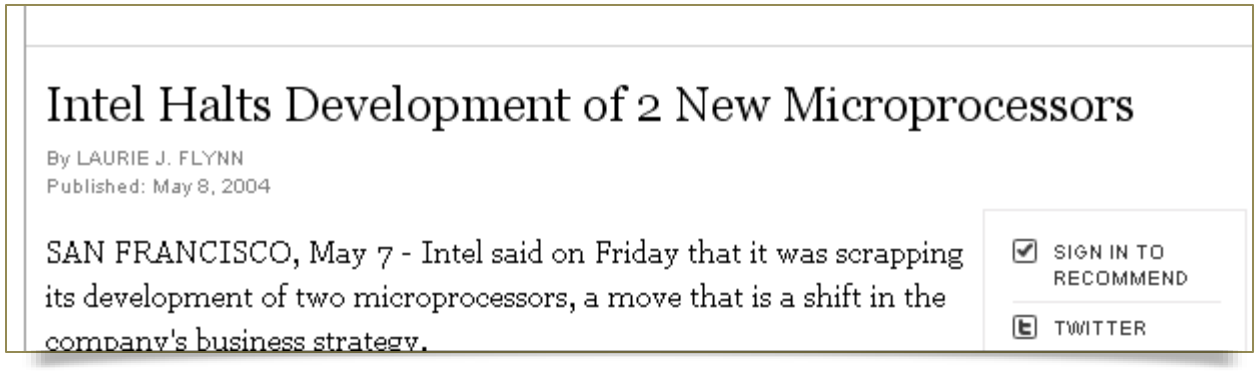
Nutrient Broth

Few dominant species

Welcome to the Jungle

Article suggestion: http://herbsutter.com/welcome-to-the-jungle/

2004 – END OF ROAD FOR MONOCORE



Intel Halts Development of 2 New Microprocessors

By LAURIE J. FLYNN
Published: May 8, 2004

SAN FRANCISCO, May 7 - Intel said on Friday that it was scrapping its development of two microprocessors, a move that is a shift in the company's business strategy.

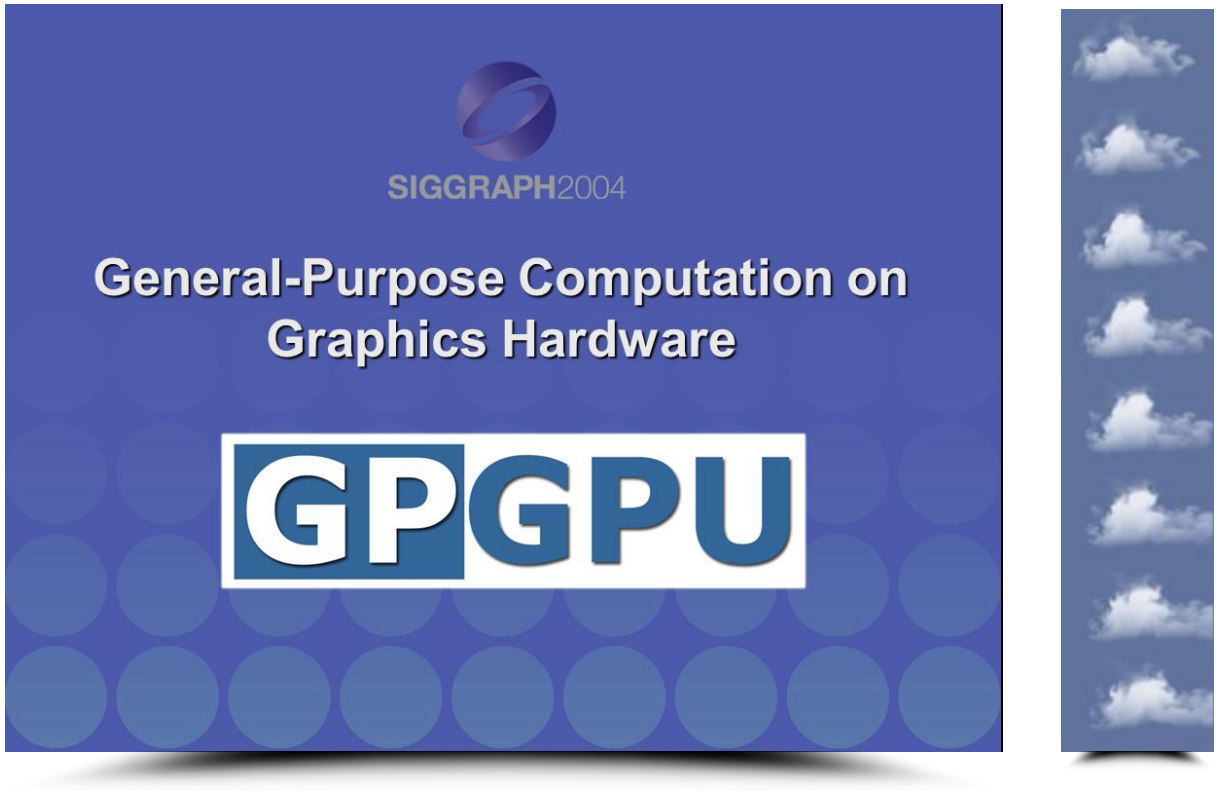
SIGN IN TO RECOMMEND

TWITTER

"THE SHIFT IS ALSO AN INDICATION THAT THE INTEL PENTIUM 4 LINE HAD REACHED ITS PERFORMANCE LIMIT FOR THE AMOUNT OF POWER IT CONSUMED."

N. BROOKWOOD - ANALYST

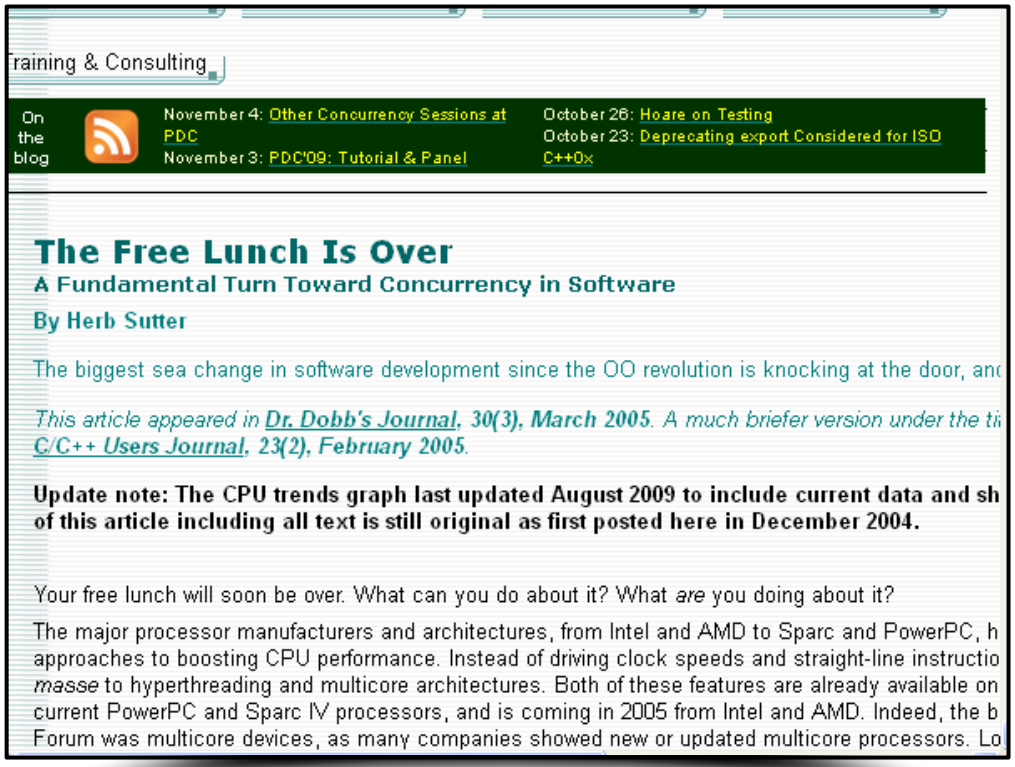
2004 – EARLY ADOPTERS FOR GPGPU



2001 – programmable shader (NVidia)
2002 – name "GPGPU" by Mark Harris
2003 – cloud simulation on GPGPU by Mark Harris

Source: http://gpgpu.org/s2004, http://gpgpu.org/oldsite/data/history.shtml, http://gpgpu.org/about,

2005 – TANSTAAFL* … BUT IT'S OVER



Training & Consulting

On the blog

November 4: Other Concurrency Sessions at PDC
November 3: PDC'09: Tutorial & Panel
October 26: Hoare on Testing
October 23: Deprecating export Considered for ISO C++0x

The Free Lunch Is Over

A Fundamental Turn Toward Concurrency in Software

By Herb Sutter

The biggest sea change in software development since the OO revolution is knocking at the door, and

This article appeared in Dr. Dobb's Journal, 30(3), March 2005. A much briefer version under the ti
C/C++ Users Journal, 23(2), February 2005.

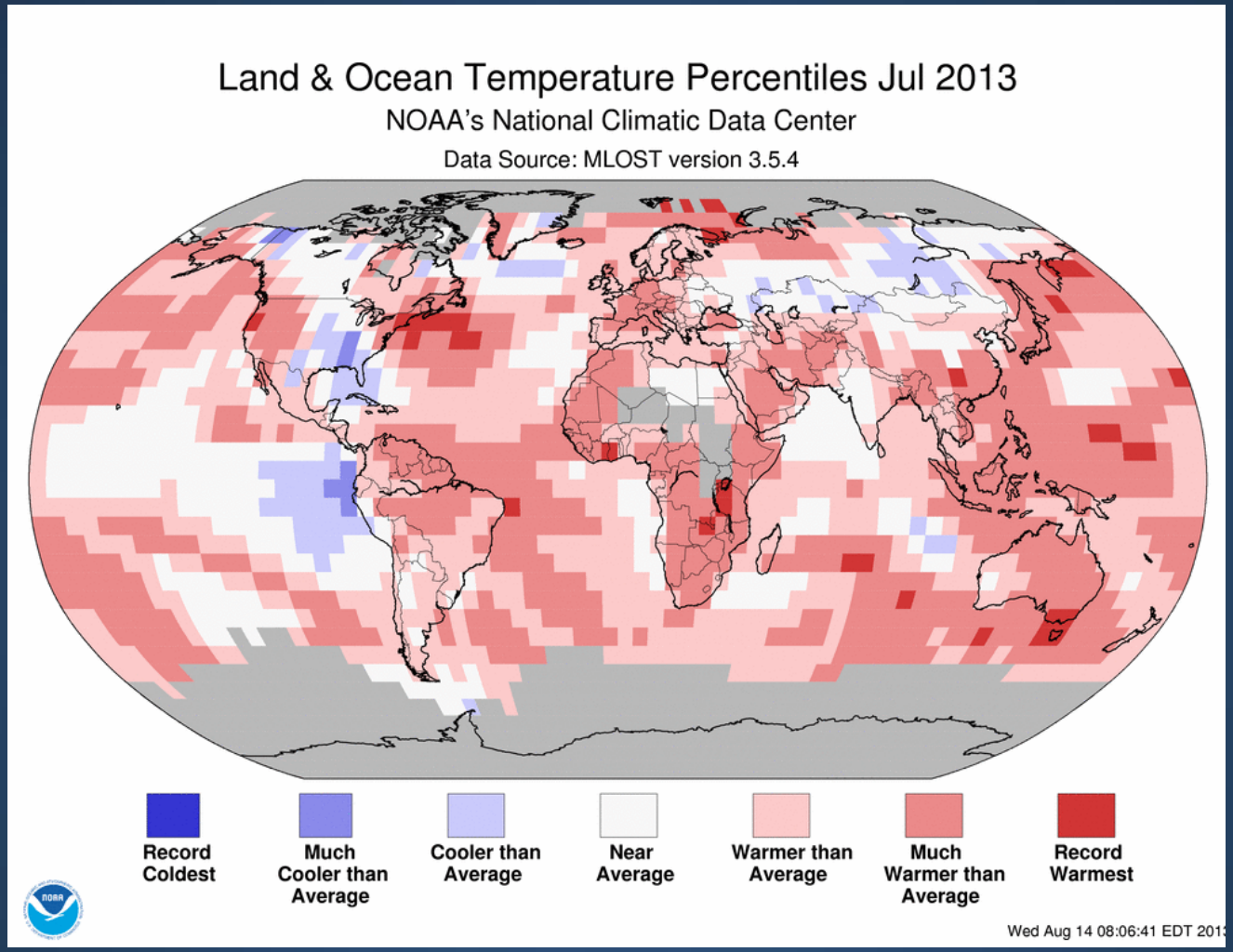
Update note: The CPU trends graph last updated August 2009 to include current data and sh
of this article including all text is still original as first posted here in December 2004.

Your free lunch will soon be over. What can you do about it? What *are* you doing about it?

The major processor manufacturers and architectures, from Intel and AMD to Sparc and PowerPC, h
approaches to boosting CPU performance. Instead of driving clock speeds and straight-line instructio
masse to hyperthreading and multicore architectures. Both of these features are already available on
current PowerPC and Sparc IV processors, and is coming in 2005 from Intel and AMD. Indeed, the b
Forum was multicore devices, as many companies showed new or updated multicore processors. Lo

Ref: "The Free Lunch Is Over - A Fundamental Turn Toward Concurrency in Software" By Herb Sutter in Dr. Dobb's Journal
* "There ain't no such thing as a free lunch"

And now ?



Source: http://www.ncdc.noaa.gov/sotc/global/

W = $

New constraints
On most consuming parts

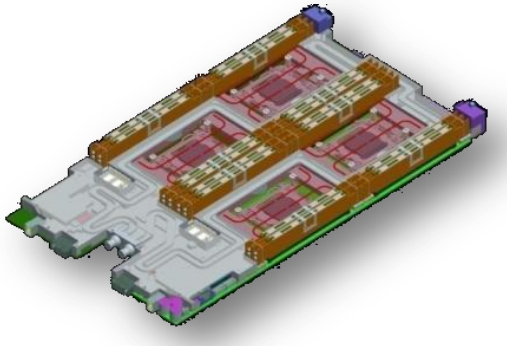
CPU

Cooling

COOLING

12kW/rack

6°C – 30kW

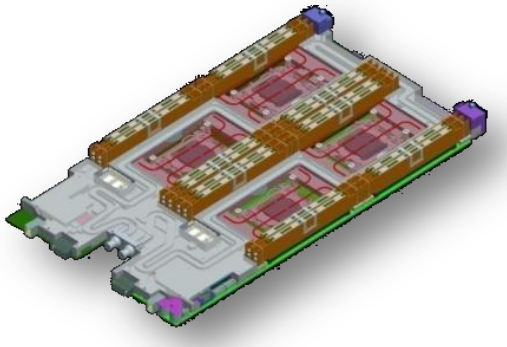


30°C – 80kW/Rack

DIRECT LIQUID COOLING

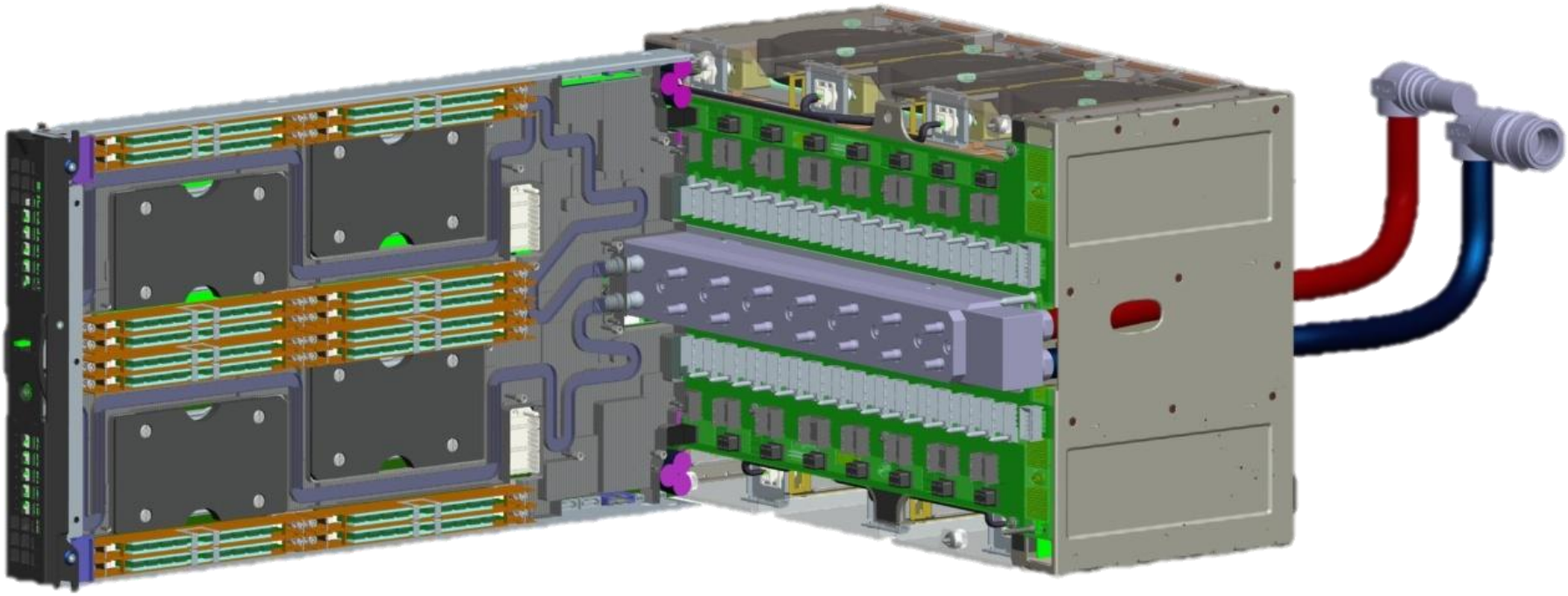
- Water has closed as possible to the heat source (CPU at 70 C)
- Water can be hotter (as ΔT is key)
- Room can be hotter (remove CRAC)
- And without any change in maintenance process
  - CPU can be changed,
  - DIMM can be changed
  - Blades can be removed

PUE < 1.08



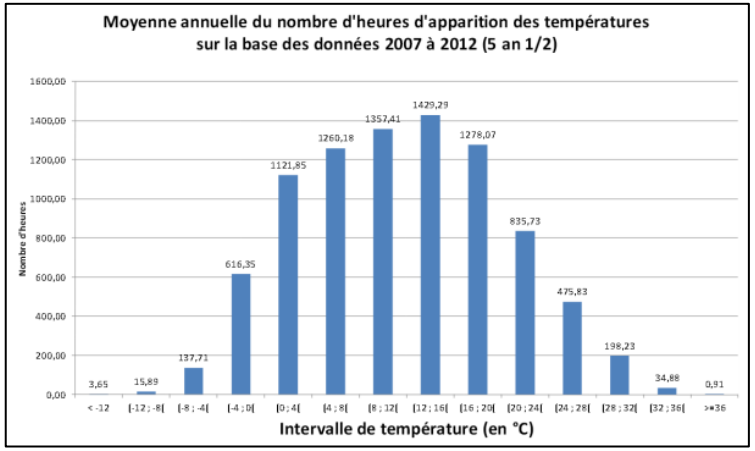
30°C – 80kW/Rack

DIRECT LIQUID COOLING



- Most efficient pure x86 supercomputer in Green 500 june'13 (météo France)
- New version with GPGPU is coming

ADIABATIC COOLING



Cooling capacity	108 kW
Power consumption	2,8 kW
Water used (estim. h/year)	472
Water used (estim. m^3)	53

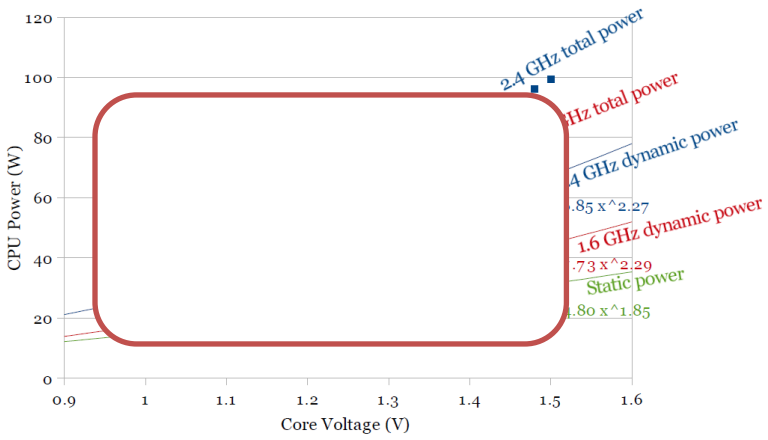
New constraints
On most consuming parts

CPU

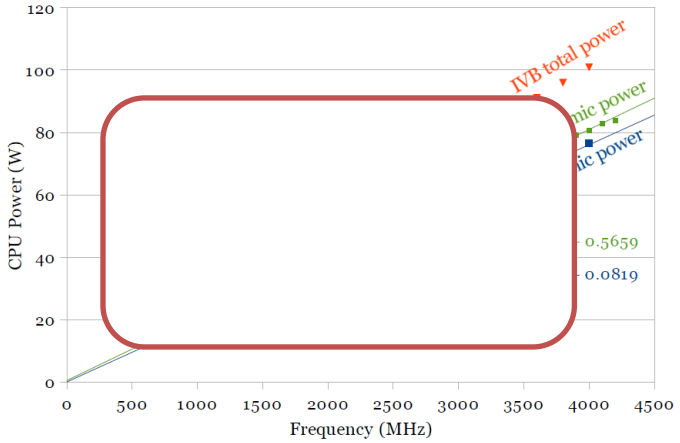
Cooling

FREQUENCY AND POWER CONSUMPTION

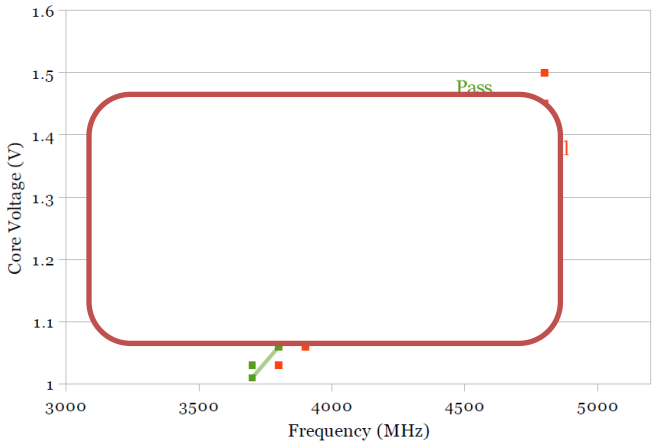
IVB Power vs. Voltage -1.6 GHz and 2.4 GHz, 90°C



Power vs. Frequency, 1.26 V, variable temperature

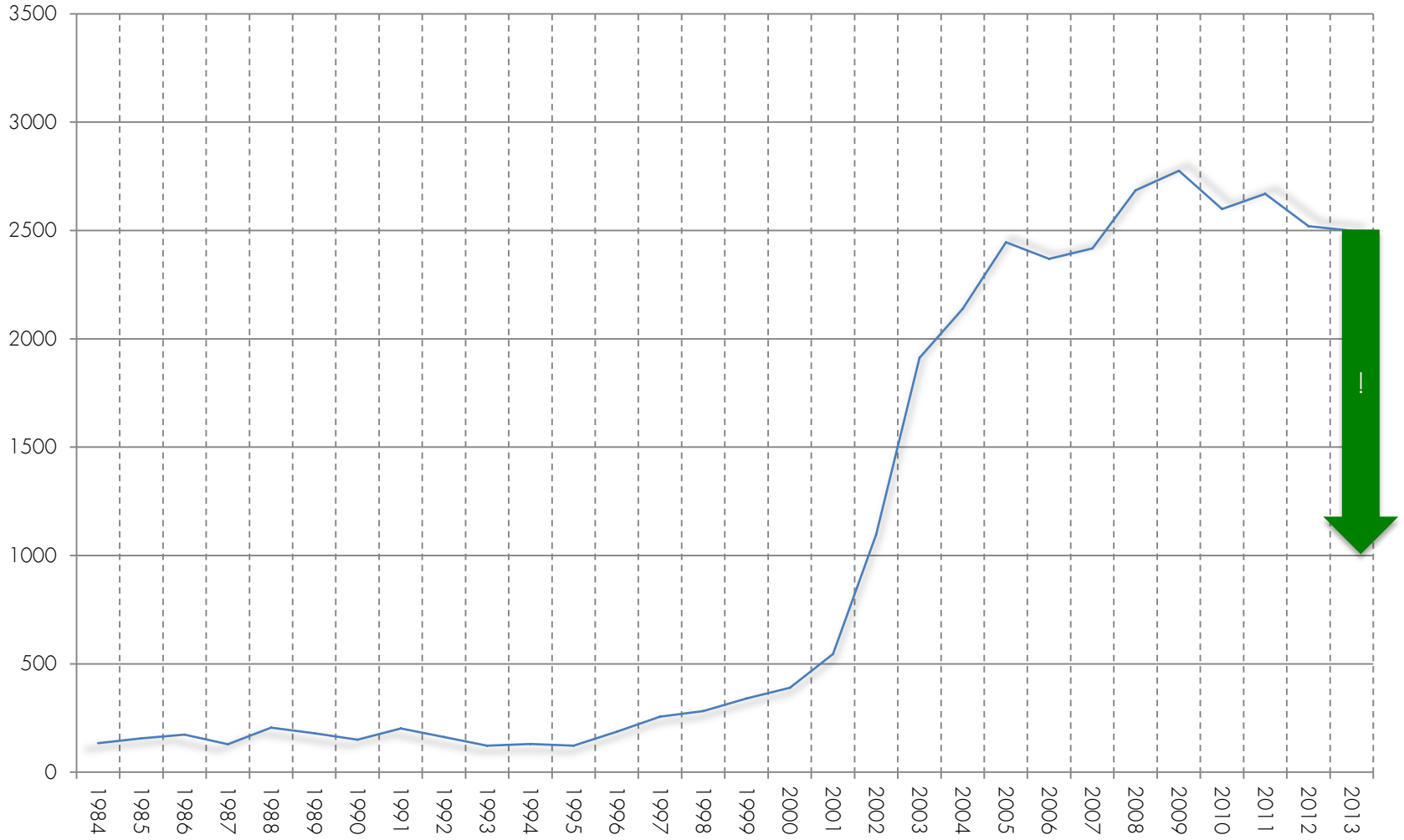


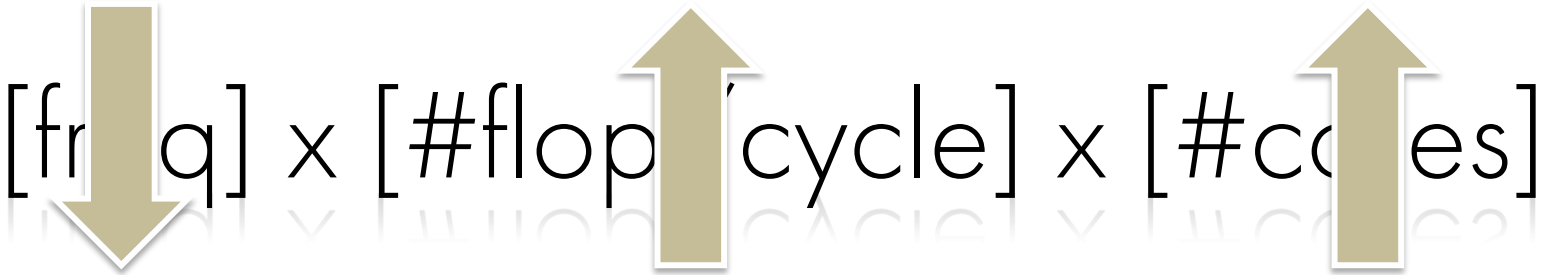
IVB Voltage-Frequency Shmoo plot



Sources: http://blog.stuffedcow.net/2012/10/intel32nm-22nm-core-i5-comparison/

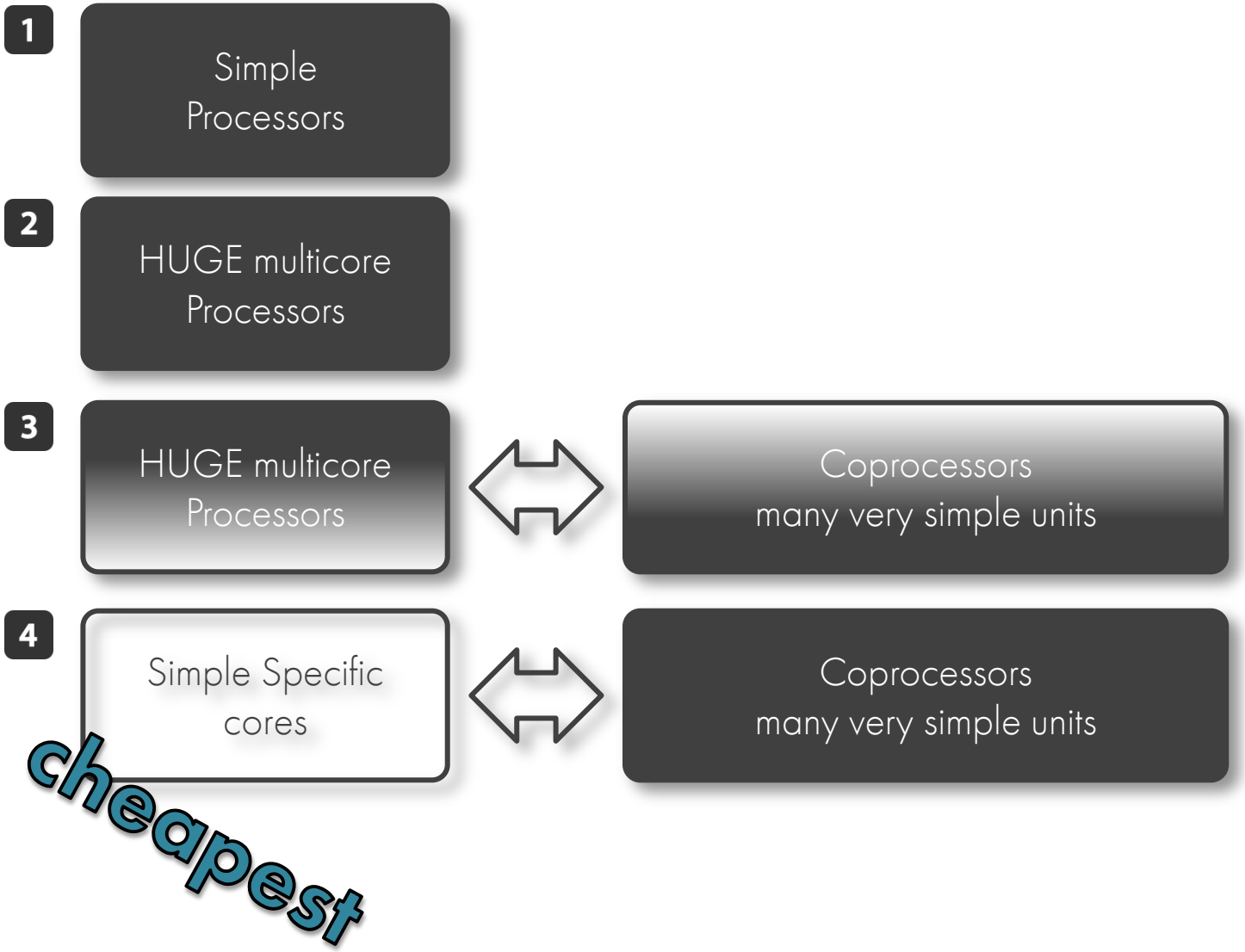
CPU FREQUENCY



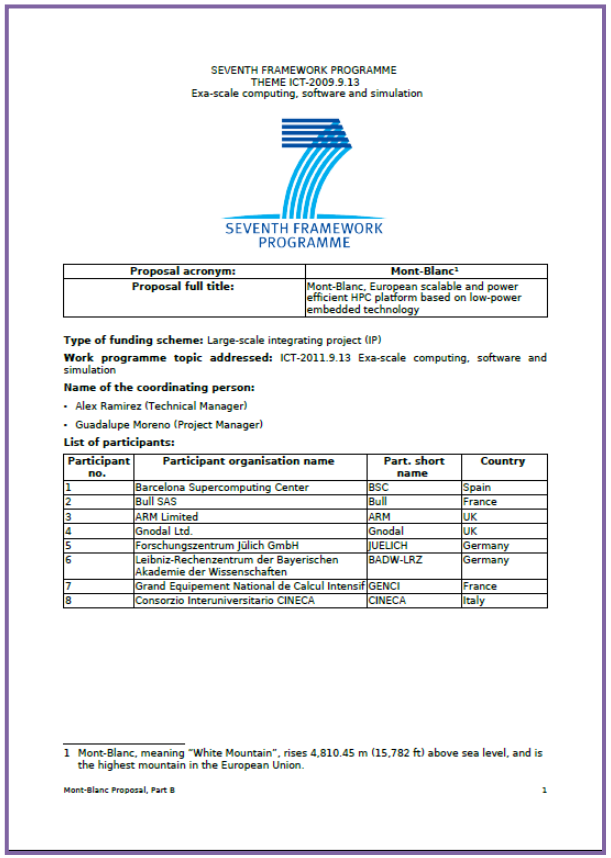

[freq] x [#flop/cycle] x [#cores]


intel
Xeon Phi™ Coprocessor
intel
Xeon Phi™ Coprocessor
NVIDIA

ARCHITECTURE EVOLUTION SUMMARY



ARM®



SEVENTH FRAMEWORK PROGRAMME
THEME ICT-2009.9.13
Exa-scale computing, software and simulation

SEVENTH FRAMEWORK PROGRAMME

Proposal acronym:	Mont-Blanc[1]
Proposal full title:	Mont-Blanc, European scalable and power efficient HPC platform based on low-power embedded technology

Type of funding scheme: Large-scale integrating project (IP)

Work programme topic addressed: ICT-2011.9.13 Exa-scale computing, software and simulation

Name of the coordinating person:

- Alex Ramirez (Technical Manager)
- Guadalupe Moreno (Project Manager)

List of participants:

Participant no.	Participant organisation name	Part. short name	Country
1	Barcelona Supercomputing Center	BSC	Spain
2	Bull SAS	Bull	France
3	ARM Limited	ARM	UK
4	Gnodal Ltd.	Gnodal	UK
5	Forschungszentrum Jülich GmbH	JUELICH	Germany
6	Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften	BADW-LRZ	Germany
7	Grand Equipement National de Calcul Intensif	GENCI	France
8	Consorzio Interuniversitario CINECA	CINECA	Italy

1 Mont-Blanc, meaning "White Mountain", rises 4,810.45 m (15,782 ft) above sea level, and is the highest mountain in the European Union.

Mont-Blanc Proposal, Part B 1

- ARM sells IP not a physical core
- ARM is present in
  - Mobile (phone, tablet, laptop)
  - Embedded (electronics)
  - Enterprise (switch, disk, printers, servers…)
  - Home (TV, camera, games consoles)
- Shipped chips
  - 2012 - 27 billions
  - 2017 - 41 billions

Why ARM cores ?

- Price ($1 - $25)
- Power consumption is key for mobile device (100mW)
- Very versatile

Source: http://ir.arm.com/phoenix.zhtml?c=197211&p=irol-presentations

HOW TO SIZE A MACHINE

[Money] => [CAPEX] + [OPEX]

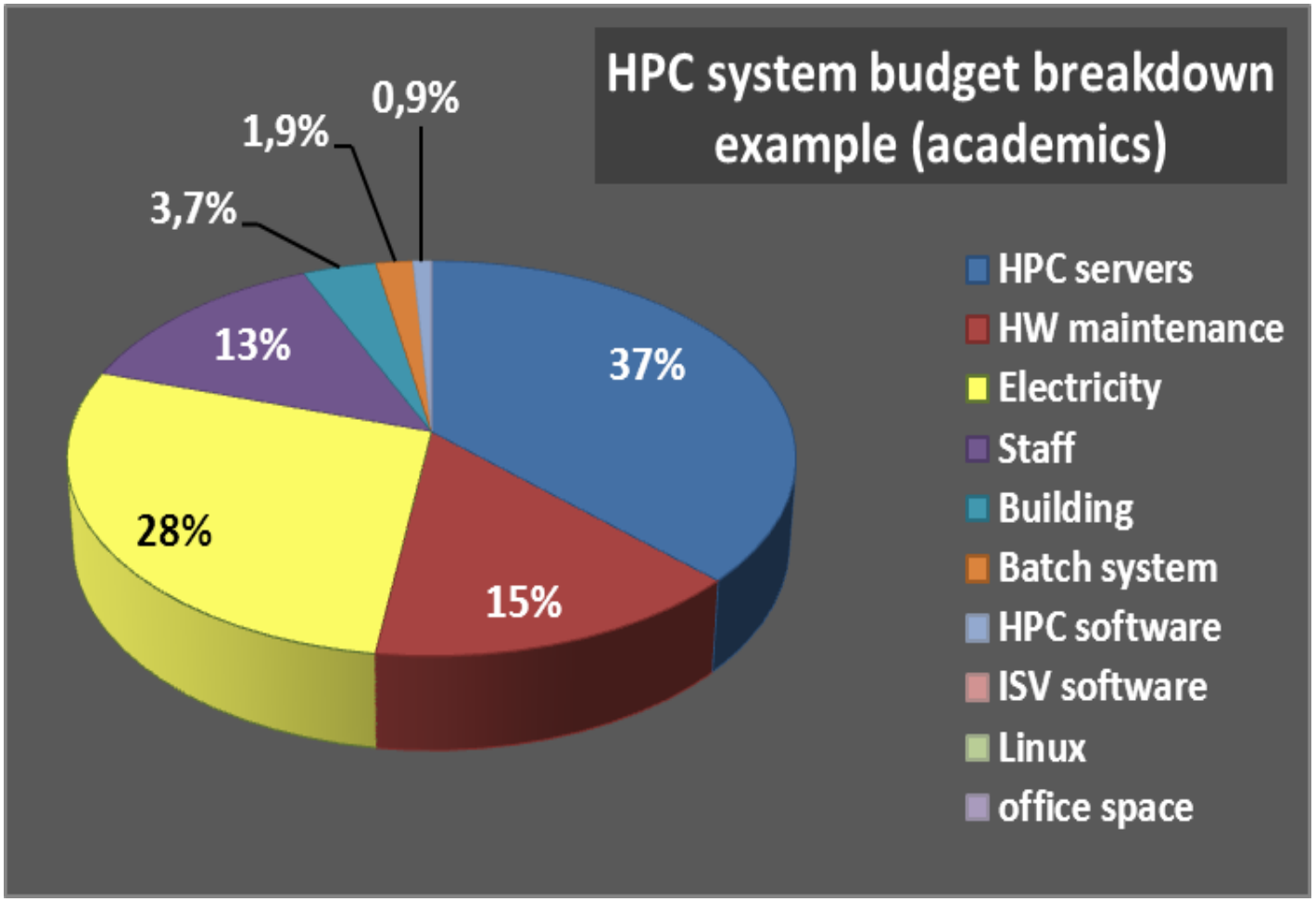
Flops/$ is decreasing
Cheaper
More flops
More Watt

cpu.h

More Watt (from CAPEX and PUE)
$/Watt is increasing
Much More Expensive

W.h

ELECTRICITY, A SIGNIFICANT PART OF HPC BUDGET



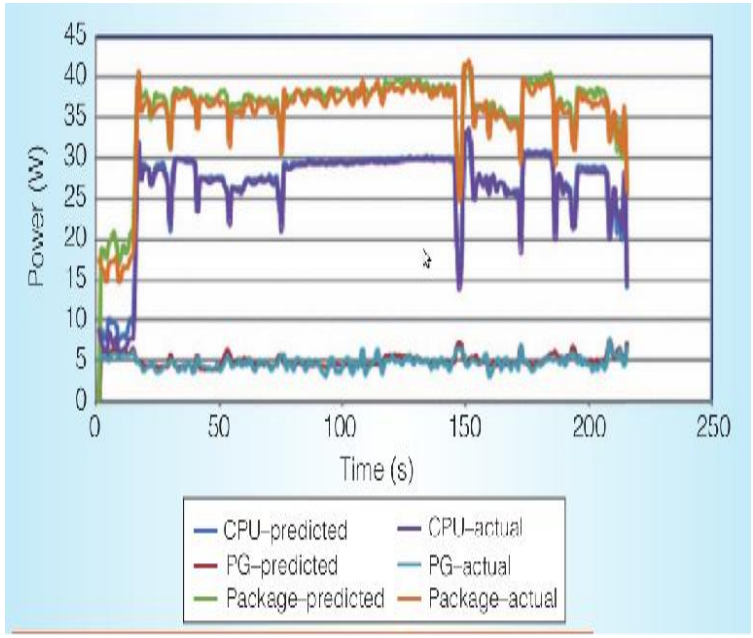
IMPACT OF ELECTRICITY COST INCREASE ON YOU

Soft dev & Users	• Write code taking into account "execution policy" • Execute your job with a choice of policy (time to result, Joule to result, Price to result) • Pay less by a cossing the policy (time to result, Joule to result)
Midleware providers	• You will have to bridge the gap between Hardware and Users • Write code taking into account "execution policy" (upward and downward) • You ~~will~~ face heterogeneity
Cloud And Data Center Director	• You will charge in Watt.hour • You can adjust your batch scheduler according to policy (green energy, power capping, ...) • Efficient hardware will be valuable (DLC, ...) • You will adjust your electricity contract

WHAT IS TO IMPROVE

WHAT DO CONSUME YOUR APPLICATIONS?

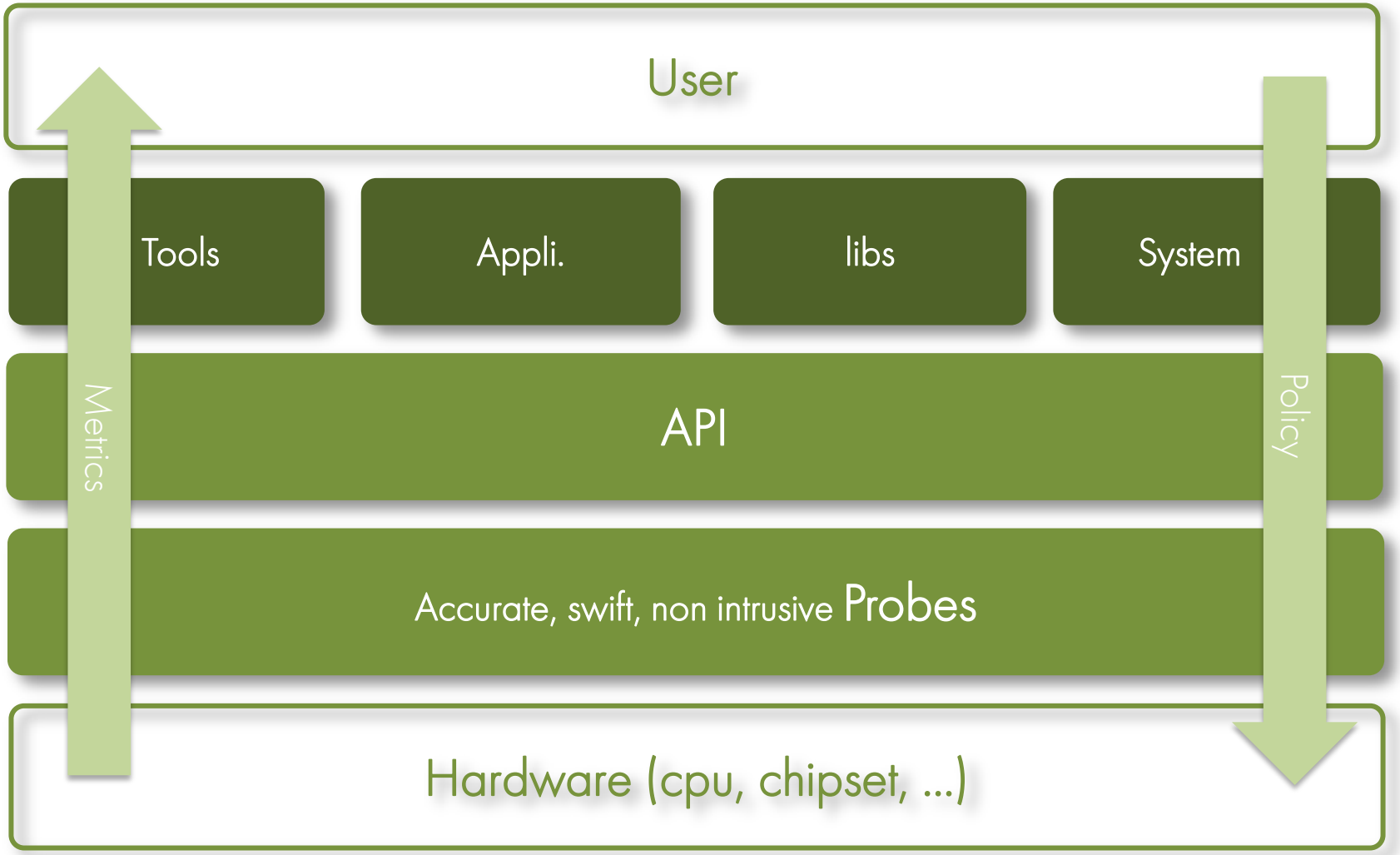
- Pluggins: RAPL, IPMI (OS) and RRD
- Per job (global value & time slice)
- Per node
- Per user
- New srun parameter to allow CPU frequency scaling for job execution



HDEEM

High Definition Energy Efficiency Monitoring

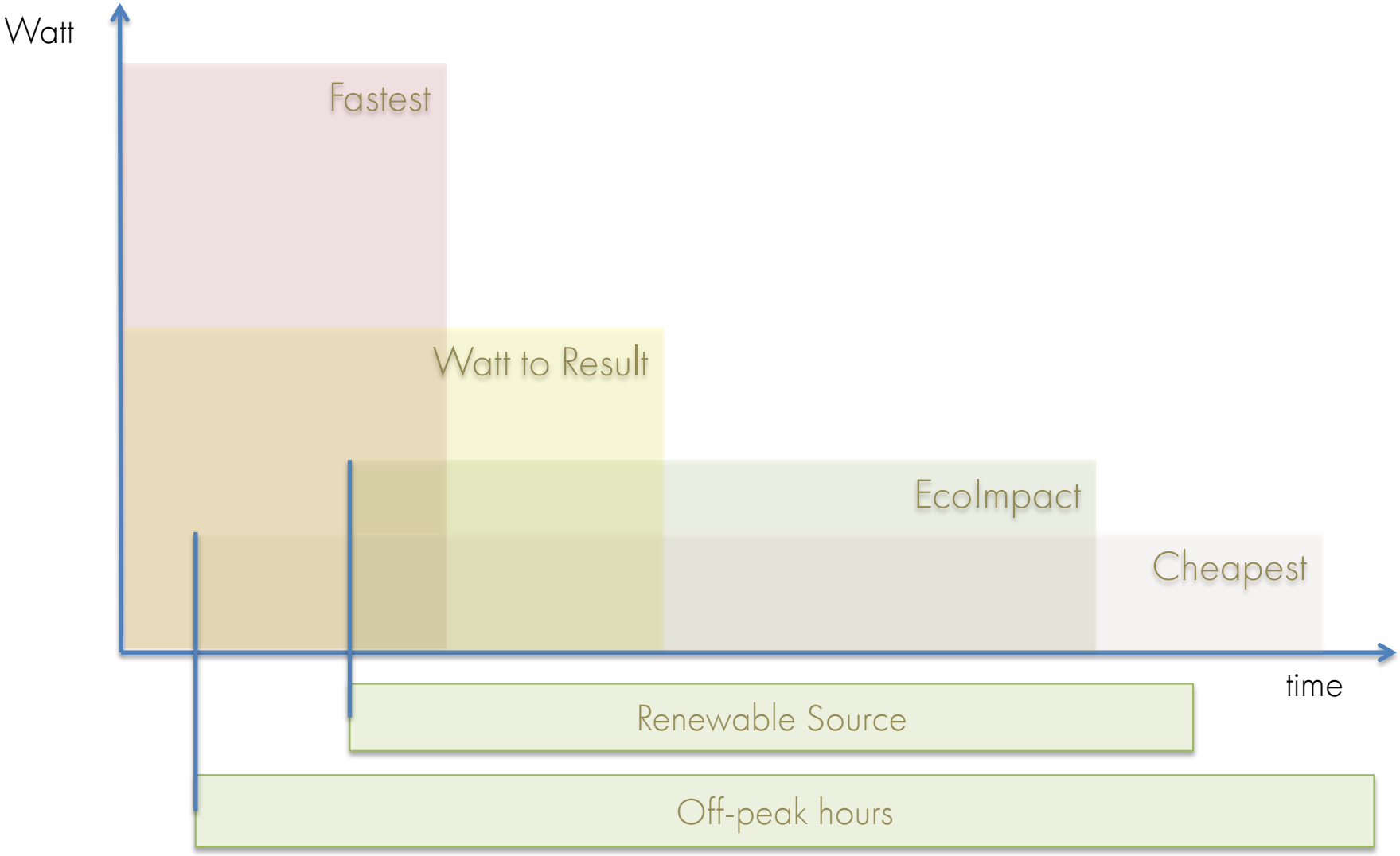
TOOL CHAIN


User
Tools
Appli.
libs
System
API
Metrics
Policy
Accurate, swift, non intrusive Probes
Hardware (cpu, chipset, ...)

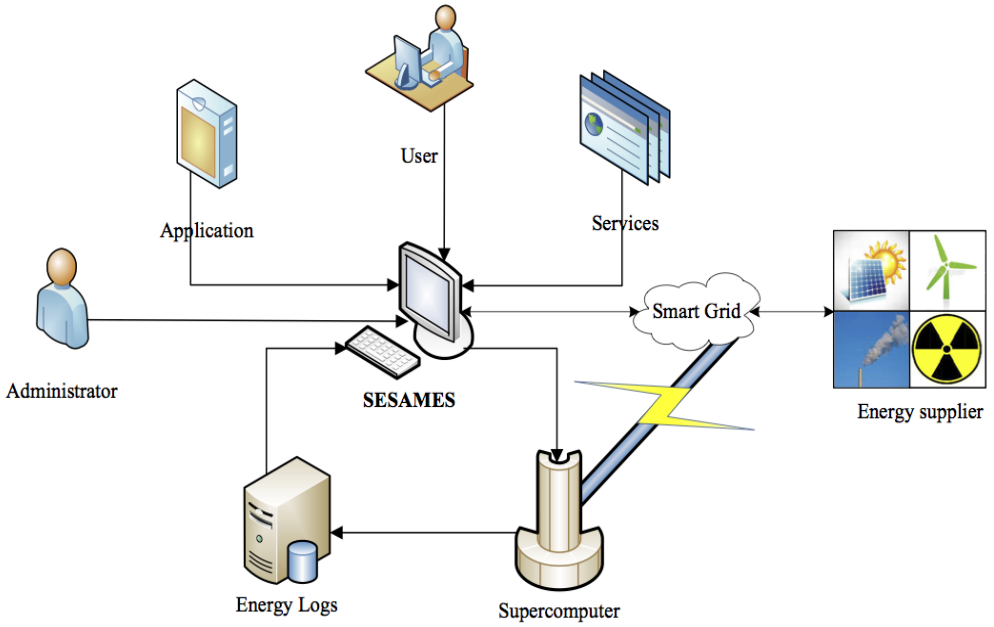
EXECUTION POLICY

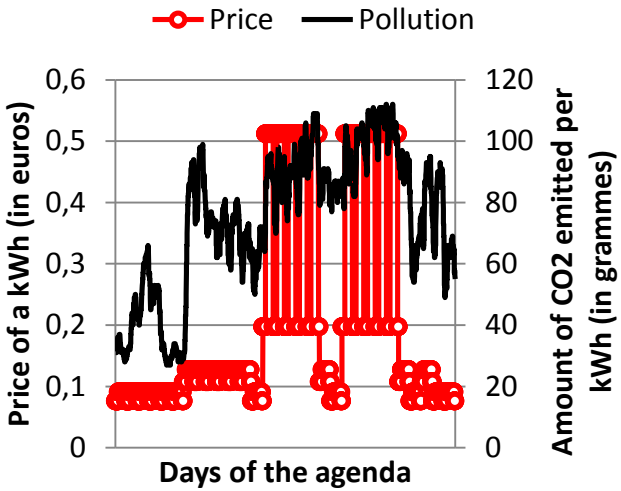
Time to result	Watt to Result	Price to Result	Ecoimpact to Result
As fast as possible	Minimal power consump-tion	Cheapeast Execution	Green Energy
s.	W.h	$	CO_2

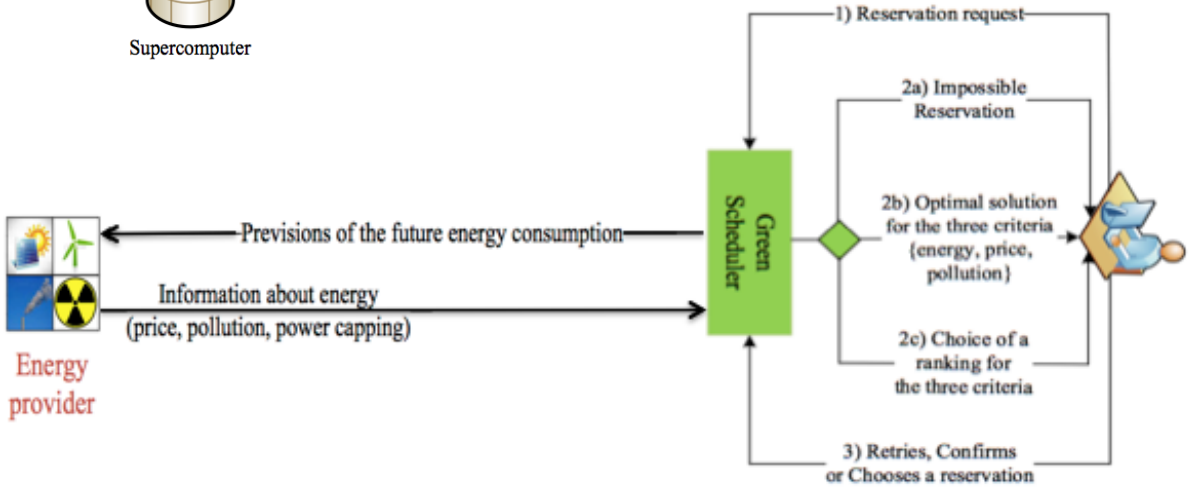
POLICIES COMPARISON



SESAMES FRAMEWORK: A MULTI-CRITERIA GREEN SCHEDULER FOR CONSUMING LESS AND BETTER







M.E.M. Diouri, O. Glück and L. Lefèvre,
SESAMES: A Smart Grid based framework for consuming less and better in extreme scale infrastructures, GreenCom 2013, 20-23 August, Beijing 2013.

TOOLS

COMPUTE PERFORMANCE (flops/s)

- Applications are written for compute performance
- Tools are seeking performance issues (host spot, idle time)
- Cost of a run is expressed in core.hour (because it's easy to measure)

POWER CONSUMPTION (W.h)

- Becomes an increasing budget
- Will be charged on end users
- Is a strict constraint for large systems
- flops/Watt is becoming a metric (see green500).

HDEEM MERGES THE BOTH WORLDS

Precise metrics in performance AND power consumption

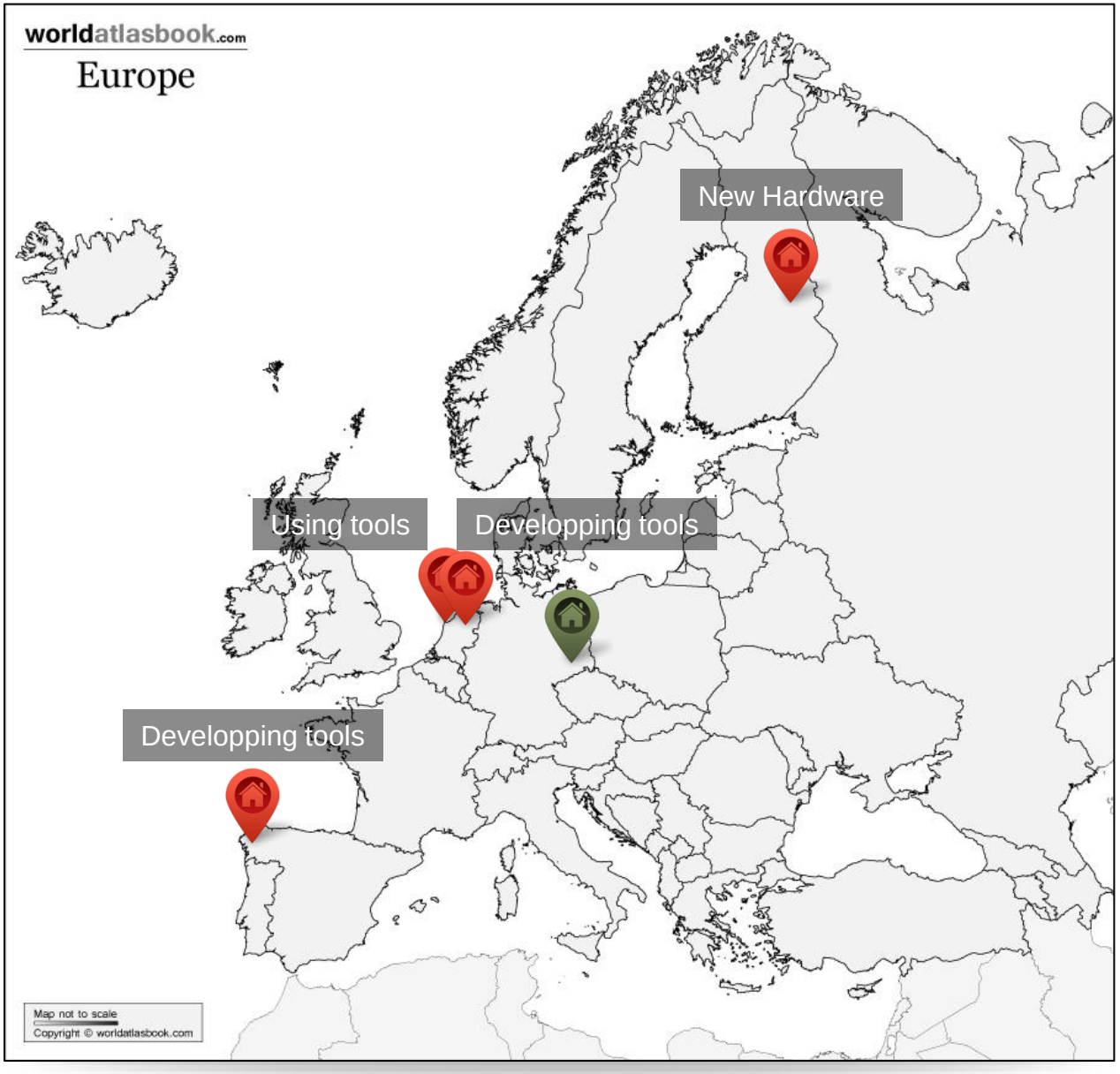
TU DRESDEN IN A NUTSHELL

TECHNISCHE UNIVERSITÄT DRESDEN

The technische universität dresden is the largest institute of higher education in the city of dresden, the largest university in saxony and one of the 10 largest universities in germany.

#students	34'993
#foreign students	3'442 (9,8%)
#employees	6'123
#faculties	14
Total Budget	€ 500M (circa)

INTERESTED IN HDEEM



Developping tools

SOLUTION OVERVIEW

Admin (2)

IB Fabric / GigE Fabric

Login (2)

Export (2)

Instrumented blades

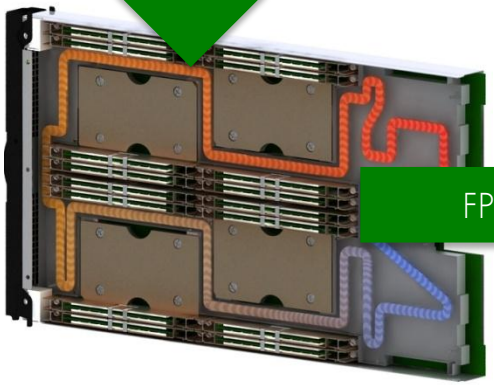
SMP (2)

FASS – Phase 1

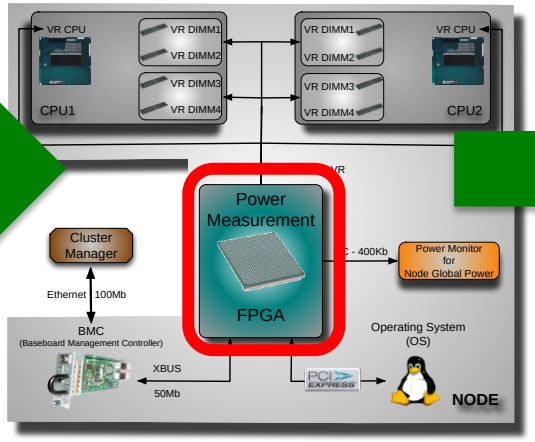
GPU Cluster

SATA / SAS

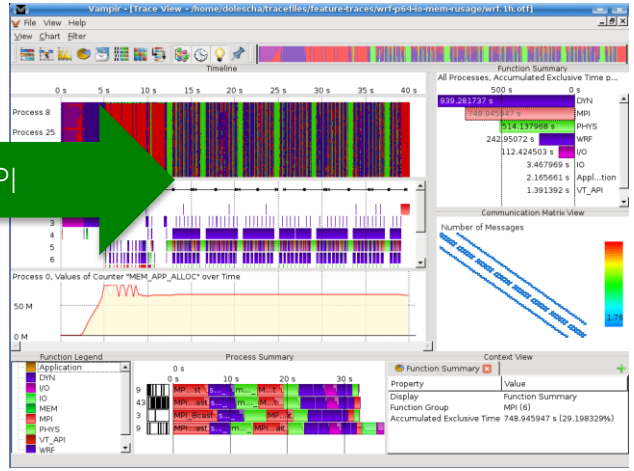
32 Nodes
48x Kepler



FPGA

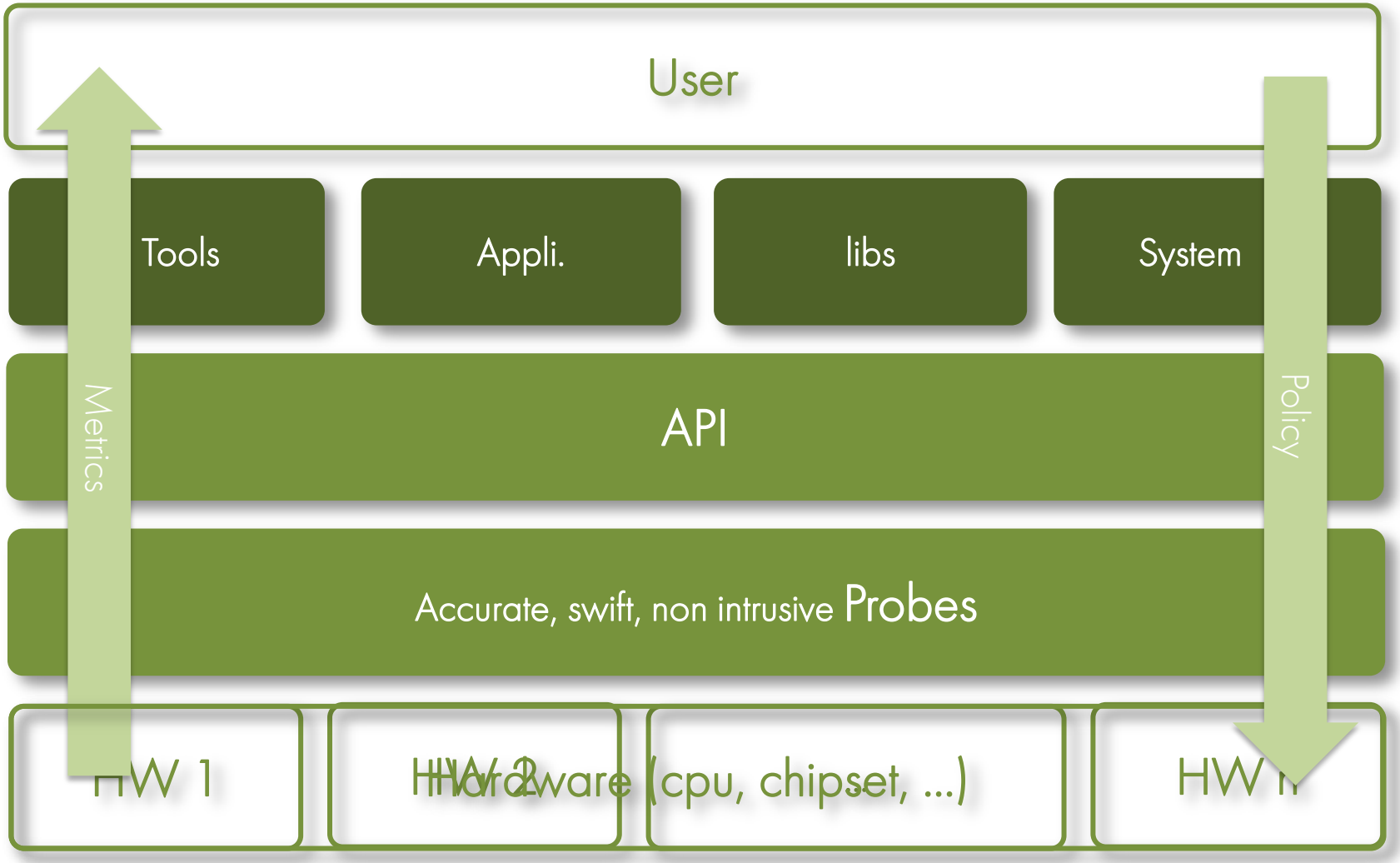


API



Freq: 500Hz - Error: 2%

DON'T FORGET ABOUT THE JUNGLE



allinea

CAPS

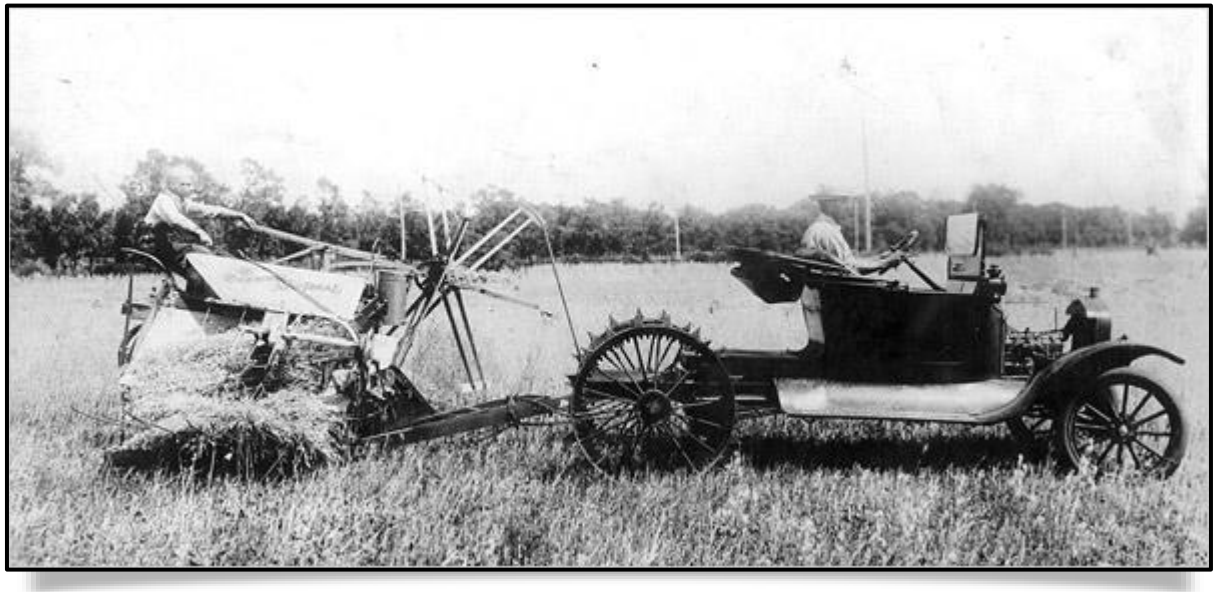
UNIVERSITE JOSEPH FOURIER
SCIENCES.TECHNOLOGIE.SANTÉ

Architect of an Open World™

1914

Production: 1908 – 1927
50% of the world market
Low cost ($850 – 4 months)
Easy Maintenance
93 min to build one

IF ALL YOU HAVE IS A HAMMER...



Frequency

Multicore

Simplification

Diversification

Optimisation