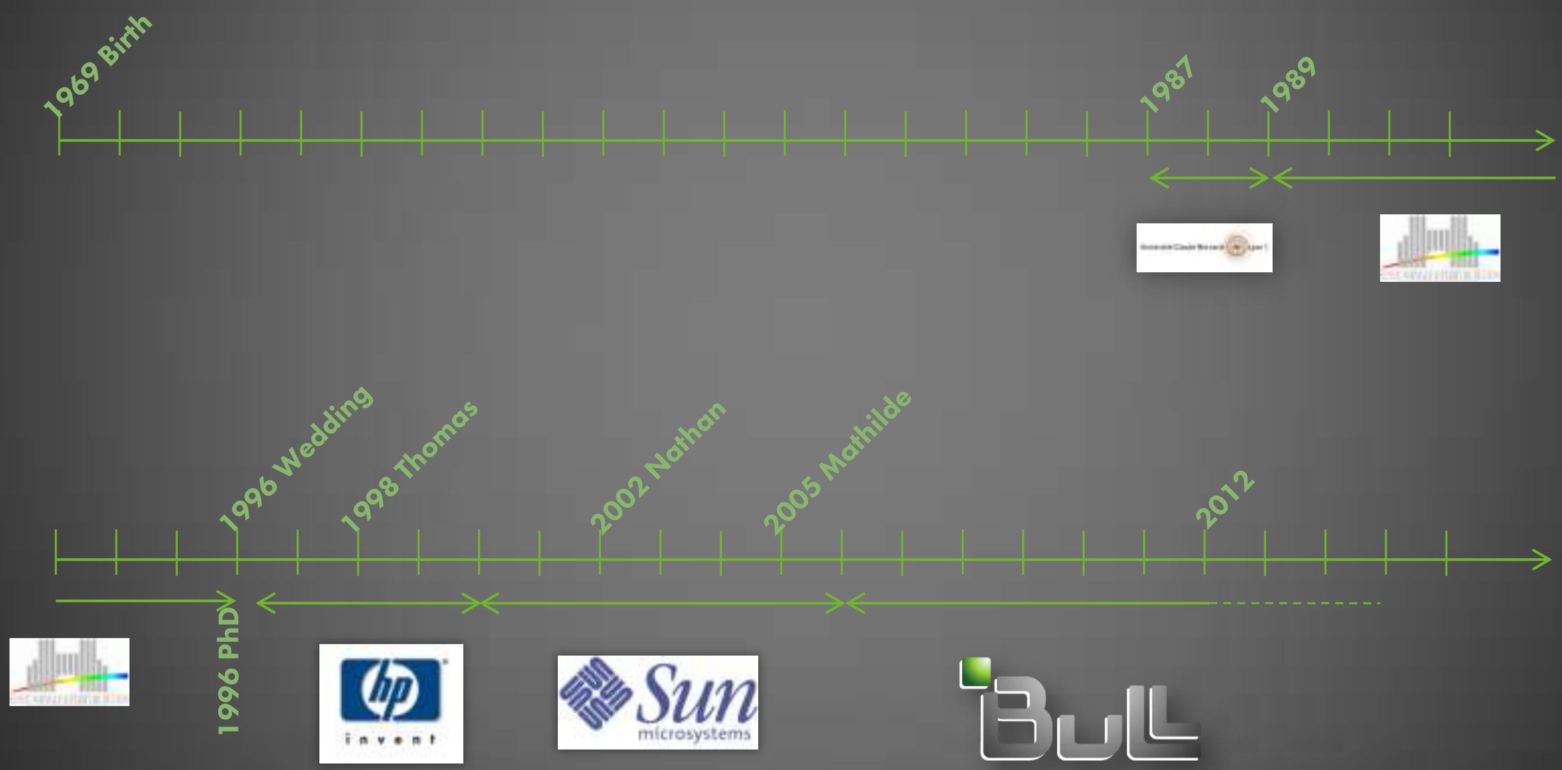
Our Green Future
« Peur bleue » ou « vie en rose » ?

3 visions


Adam Brown © - http://adambrownfolio.wordpress.com/

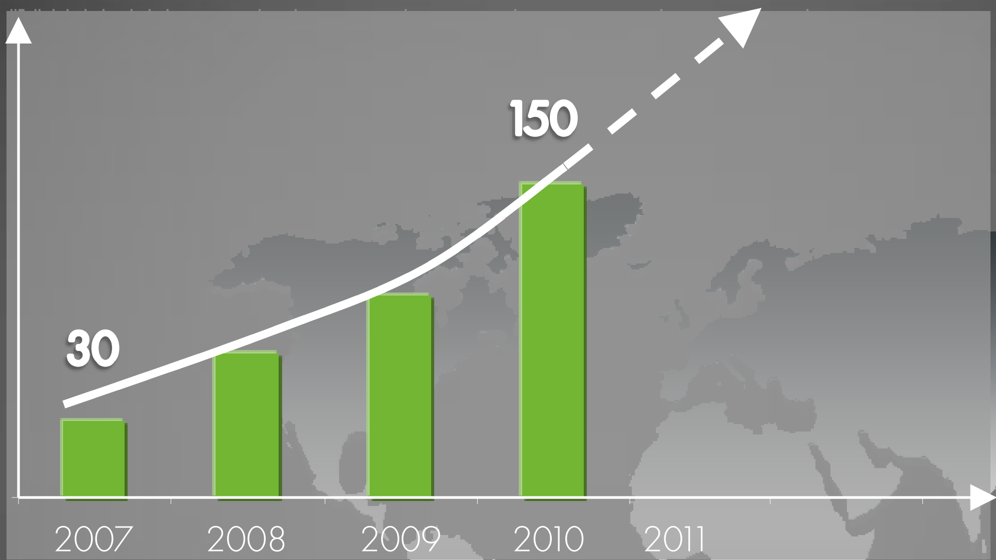
Xavier VIGOUROUX
Responsable « Educ Research » for HPC in Bull

Who am I

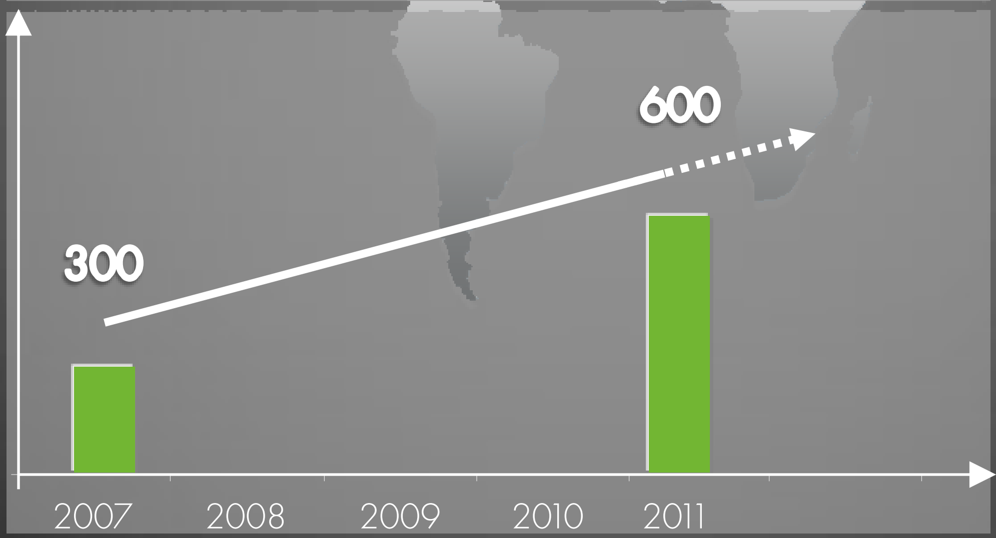

1969 Birth
1987
1989
1996 Wedding
1998 Thomas
2002 Nathan
2005 Mathilde
2012
1996 PhD
invent
Sun
microsystems

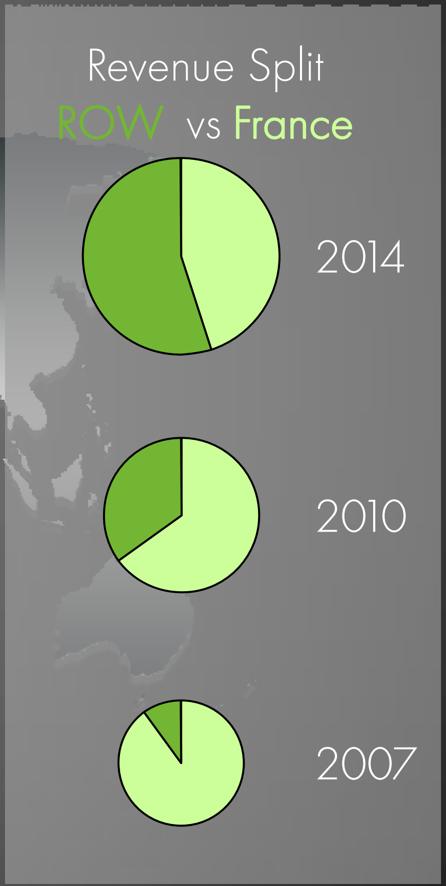
Bull in Extreme Computing

Revenue
Sustained global growth with a strong international focus



HPC experts
The largest group of HPC experts in Europe



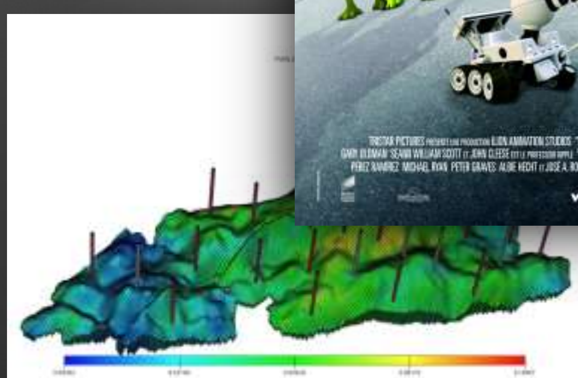


1931 2011

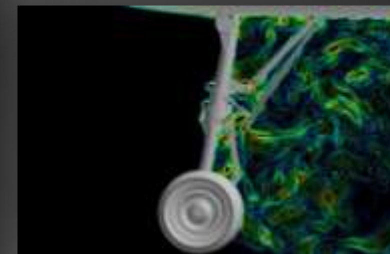
Experimentation

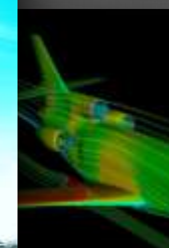
Modelisation

Simulation




Powered by
Bull
PLANETE
51
Avec la voix de
VINCENT
CASSEL
ALIEN
MALGRÉ LUI !
www.planete51-lefilm.com





Top 500

Architect of an Open World

top500.

m.

Linpack

$$Ax = b$$

Flops/s

top3

2 566 Tflops/s

4 701 Tflops/s

186 368 cœurs Xeon +

7 168 accélérateurs

4 040 kW

10 051 Tflops/s

11 280 Tflops/s

705 024 cœurs sparc

12 660 kW

1 759 Tflops

2 331 Tflops

224 162 cœurs AMD

7 000 kW

3 petaflop-scale systems

TERA 100

- 1.25 PetaFlops
  140 000+ Xeon cores
- 256 TB memory
- 30 PB disk storage
- 500 GB/s IO throughput
- 580 m^2 footprint

CURIE

- 2 PetaFlops
  90 000+ Xeon cores
  148 000 GPU cores
- 360 TB memory
- 10 PB disk storage
- 250 GB/s IO throughput
- 200 m^2 footprint

IFERC

- 1,5 PetaFlops
  70 000+ Xeon cores
- 280 TB memory
- 15 PB disk storage
- 120 GB/s IO throughput
- 200 m^2 footprint

3 large GPU based systems

TERA 100

- GPU-based extension
- 198 bullx B505 accelerator blades
- 396 NVIDIA® Tesla™ M2090 GPU processors
- 202,752 GPU cores

CURIE

- GPU-based extension
- 144 bullx B505 accelerator blades
- 288 NVIDIA® Tesla™ M2090 GPU processors
- 147,456 GPU cores

BSC

- GPU-based system
- 126 bullx B505 accelerator blades
- 252 NVIDIA® Tesla™ M2090 GPU processors
- 129,024 GPU cores

cea

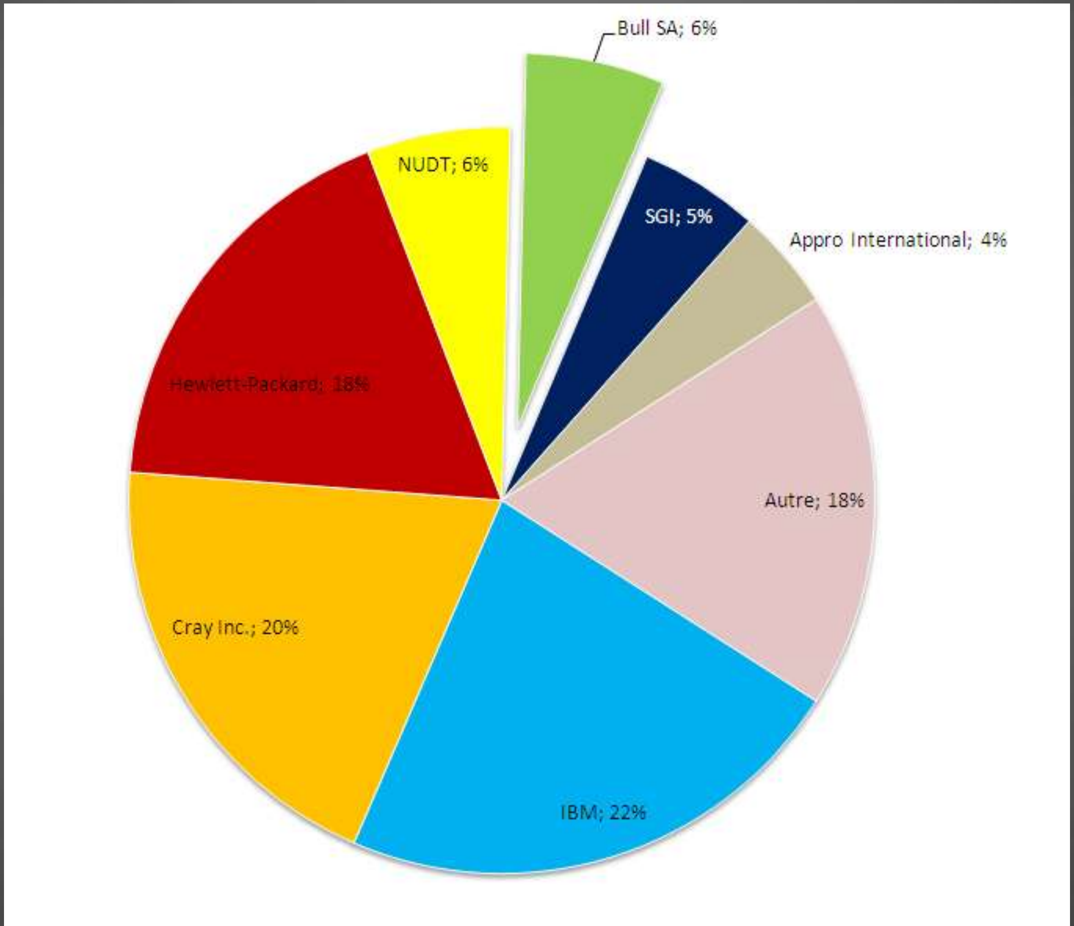
GENCI

PRACE

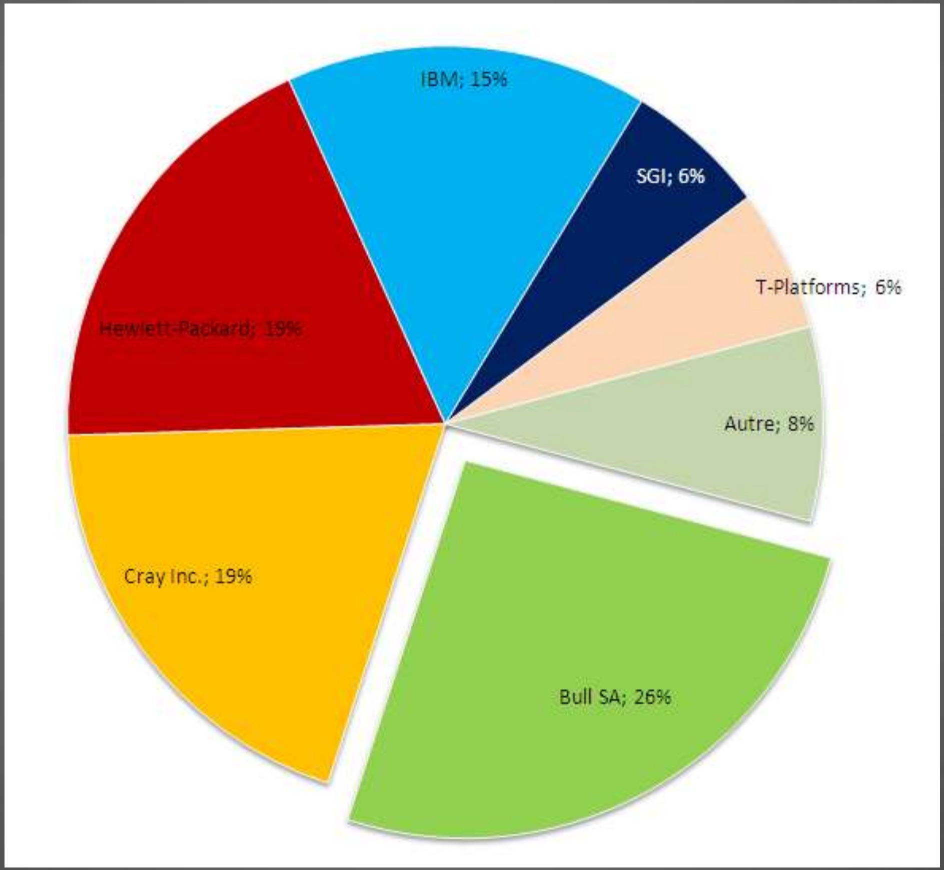
BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

	Rank	Site	Country	Year	Rmax	Rpeak	Effeciency (%)
	9	Commissariat a l'Energie Atomique (CEA)	France	2010	1050000	1254550	83,70%
	27	Bull	France	2011	1,36 Pflops		
	28	International Fusion Energy Research Centre (IFERC), EU(F4E) - Japan Broader Approach collaboration	Japan	2011	1,13 Pflops		
	36	Forschungszentrum Juelich (FZJ)	Germany	2009	274800	308282,88	89,14%
	47	Universitaet Aachen/RWTH	Germany	2011	219838	270538	81,26%
	75	Commissariat a l'Energie Atomique (CEA)	France	2011	154000	274560	56,09%
	93	Atomic Weapons Establishment	United Kingdom	2010	124600	145152	85,84%
	102	Tres Grand Centre de calcul du CEA	France	2011	109900	198161,6	55,46%
	106	Commissariat a l'Energie Atomique (CEA)/CCRT	France	2010	108500	129998	83,46%
	110	Bull	France	2011	106998	124416	86,00%
#1	114	Barcelona Supercomputing Center	Spain	2011	103200	182881,44	56,43%
	148	Bull	France	2010	87470	104602	83,62%
	149	Commissariat a l'Energie Atomique (CEA)	France	2010	87470	104417	83,77%
	154	University of Cologne, Regional Computing Centre	Germany	2010	85900	100171	85,75%
	459	Commissariat a l'Energie Atomique (CEA)	France	2006	52840	63795,2	82,83%

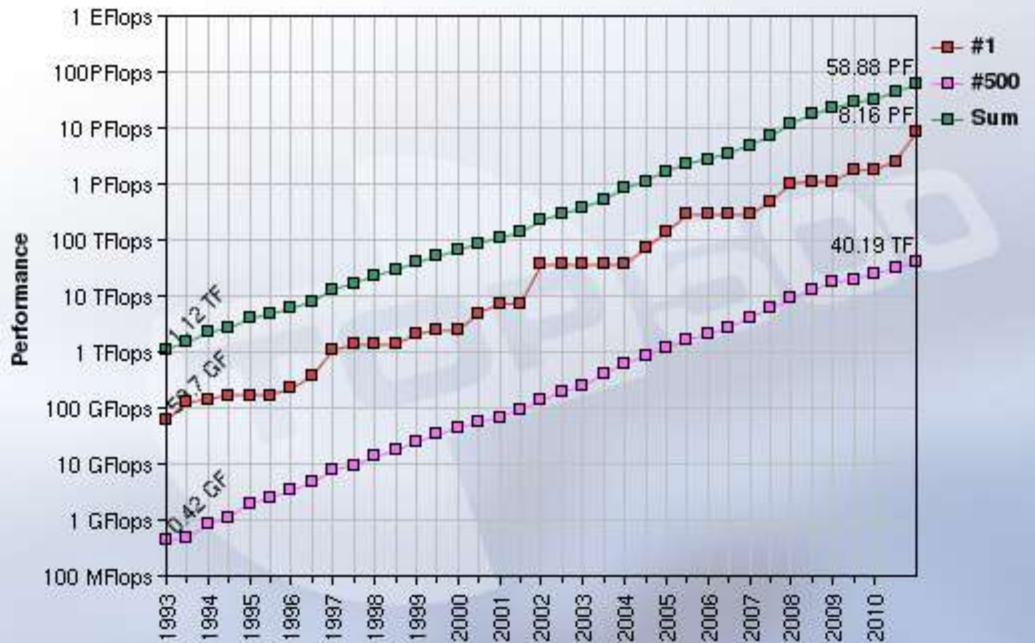
Bull #5 – x86 world wide



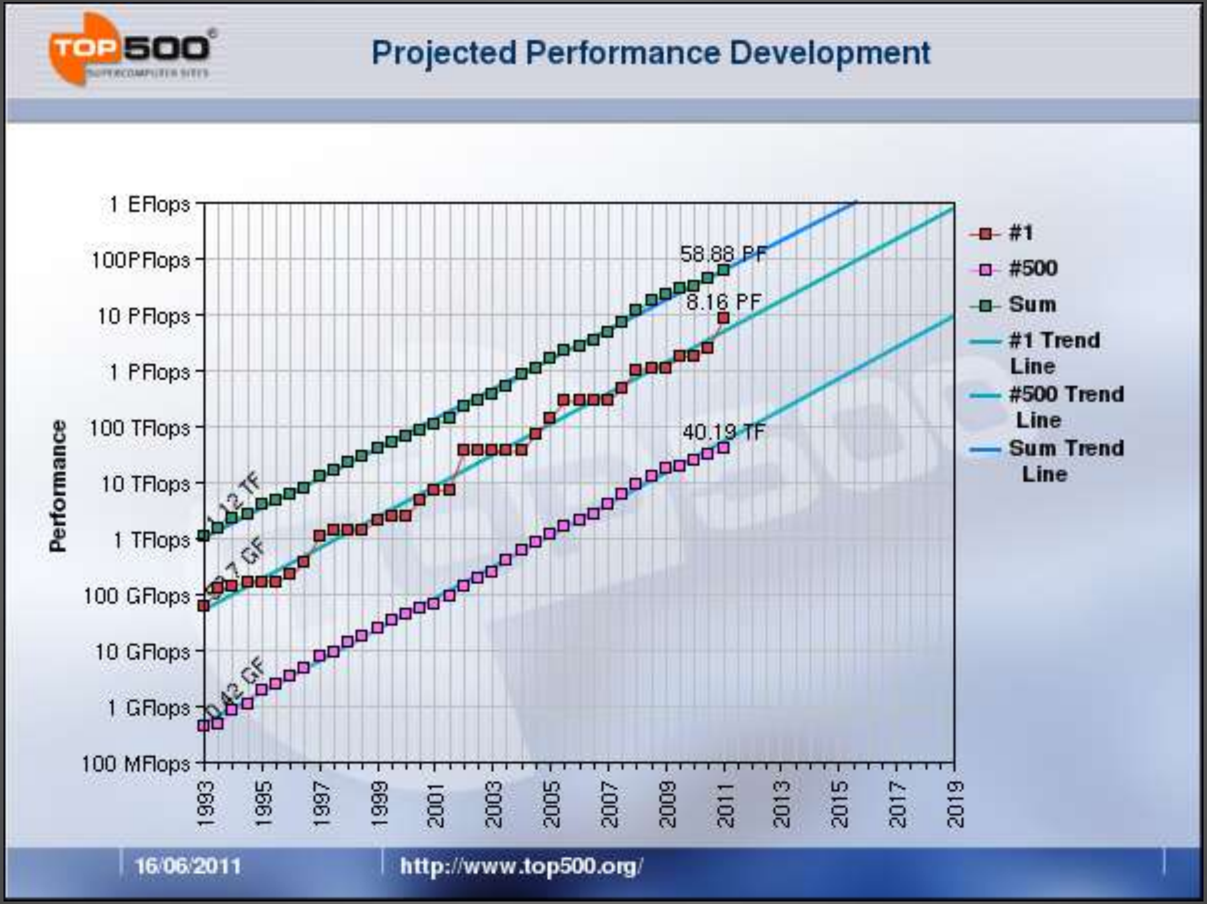
Bull #1 – x86 in Europe



Performance Development



10^{18}


TOP500
Projected Performance Development
1 EFlops
100PFlops
10 PFlops
1 PFlops
100 TFlops
10 TFlops
1 TFlops
100 GFlops
10 GFlops
1 GFlops
100 MFlops
Performance
58.88 PF
8.16 PF
40.19 TF
1.12 TF
59.7 GF
0.42 GF
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013
2015
2017
2019
#1
#500
Sum
#1 Trend Line
#500 Trend Line
Sum Trend Line
16/06/2011
http://www.top500.org/

U.S. DEPARTMENT OF ENERGY

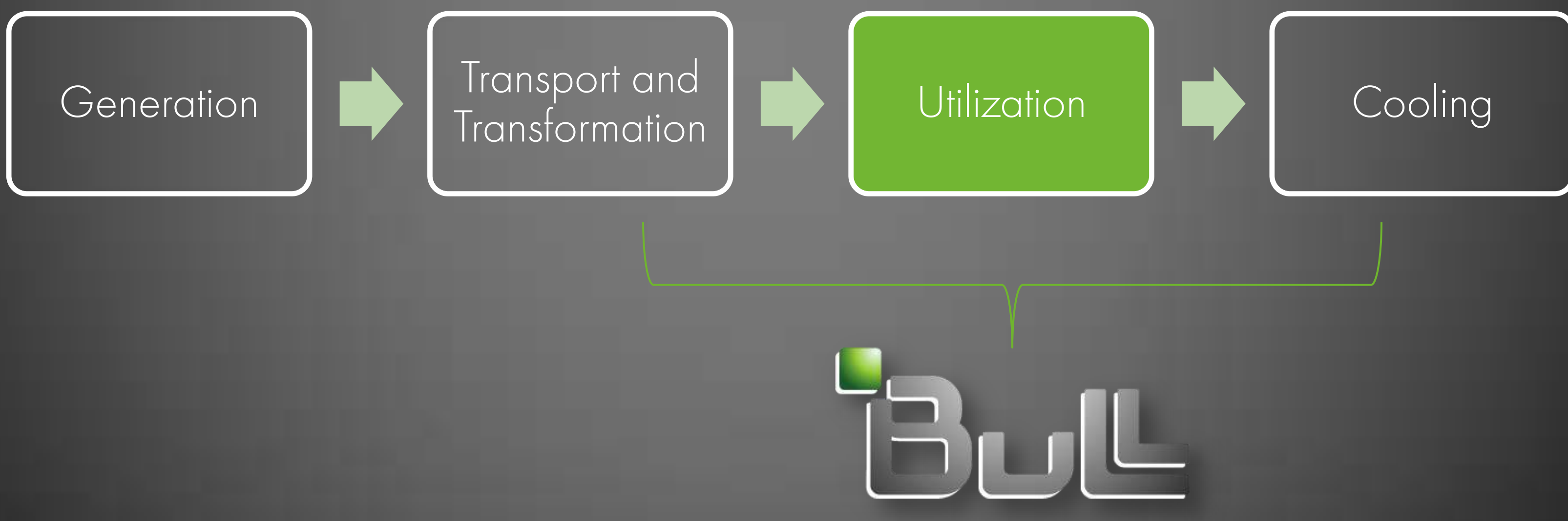
Potential System Architecture

Systems	2009	2018	Difference Today & 2018
System peak	**2 Pflop/s**	**1 Eflop/s**	**O(1000)**
Power	**6 MW**	**~20 MW (goal)**	
System memory	0.3 PB	32 - 64 PB	O(100)
Node performance	125 GF	1,2 or 15TF	O(10) – O(100)
Node memory BW	25 GB/s	2 - 4TB/s	O(100)
Node concurrency	12	O(1k) or 10k	O(100) – O(1000)
Total Node Interconnect BW	3.5 GB/s	200-400GB/s (1:4 or 1:8 from memory BW)	O(100)
System size (nodes)	18,700	O(100,000) or O(1M)	O(10) – O(100)
Total concurrency	225,000	O(billion) [O(10) to O(100) for latency hiding]	O(10,000)
Storage	15 PB	500-1000 PB (>10x system memory is min)	O(10) – O(100)
IO	0.2 TB	60 TB/s	O(100)
MTTI	days	O(1 day)	- O(10)

Source: "Advanced Scientific Computing Research", Walter M. Polansky, 27 avril 2010

Vision 1
More flop

Decomposition

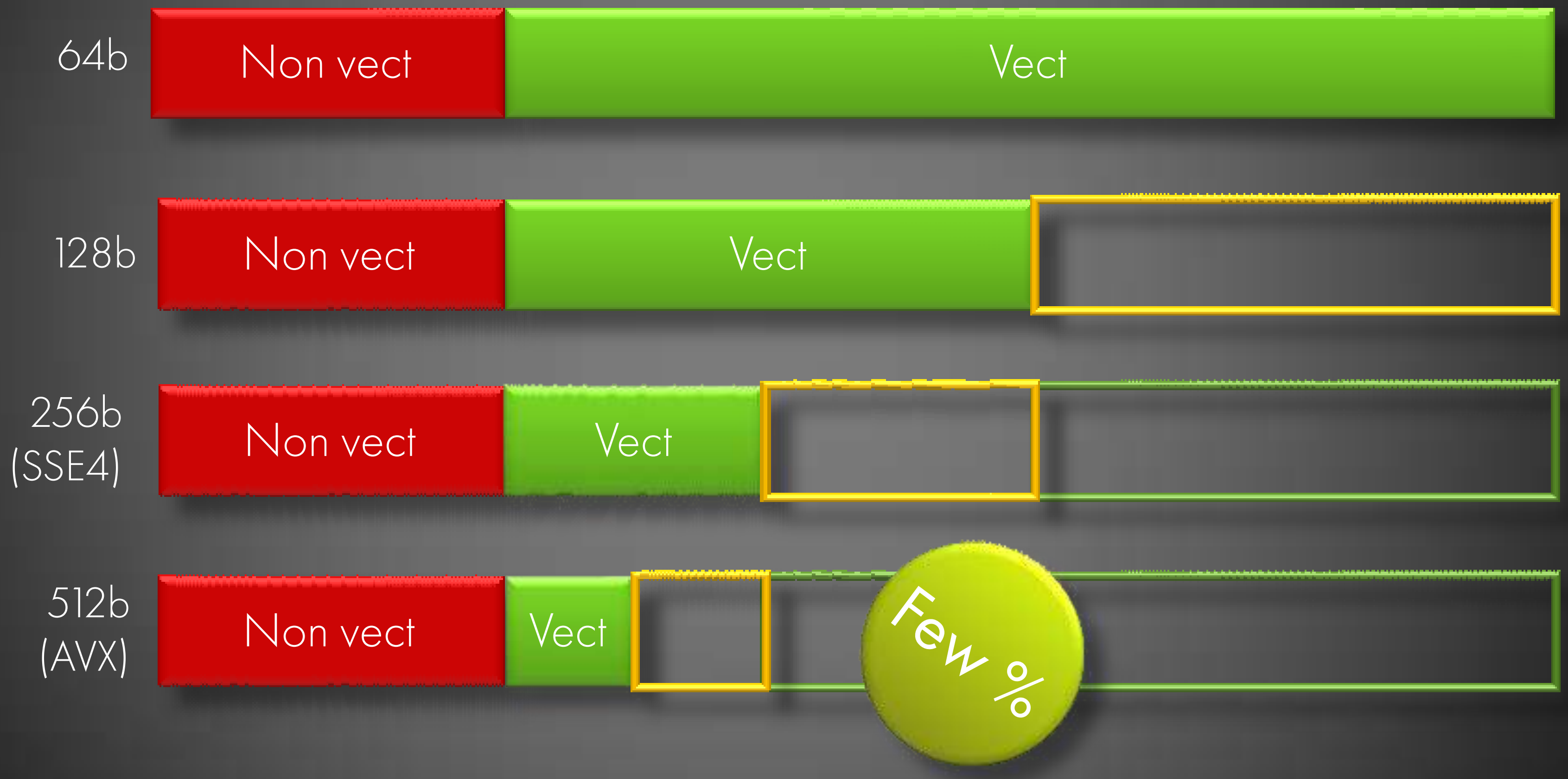


Considering Peak performance
Considering Compute nodes consumption

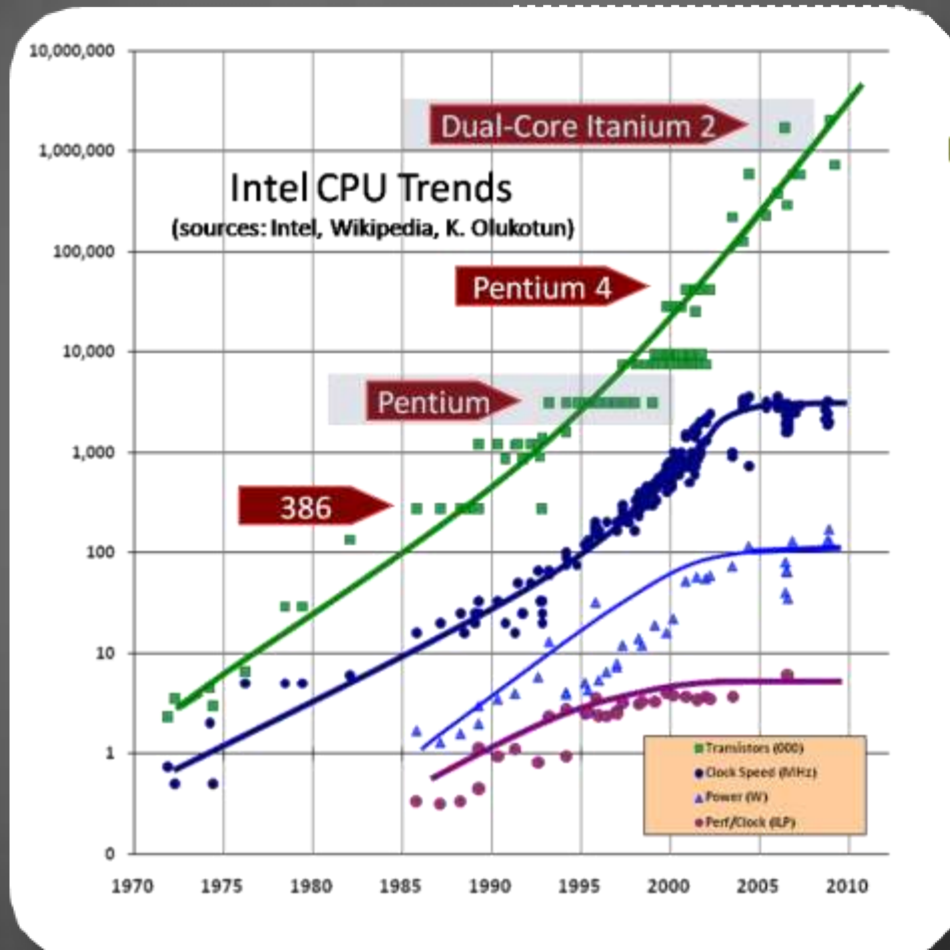
Flop/s/W
x56 kei (sparc)
x43 tianhe (GPU)
x150 Jaguar (AMD)

Where will we find this factor ?

Get more flops with Vectorization



Get more flops with ILP



0%

<0%

Source: ISSCC 2010 TRENDS REPORT

Get more flops with CPU freq

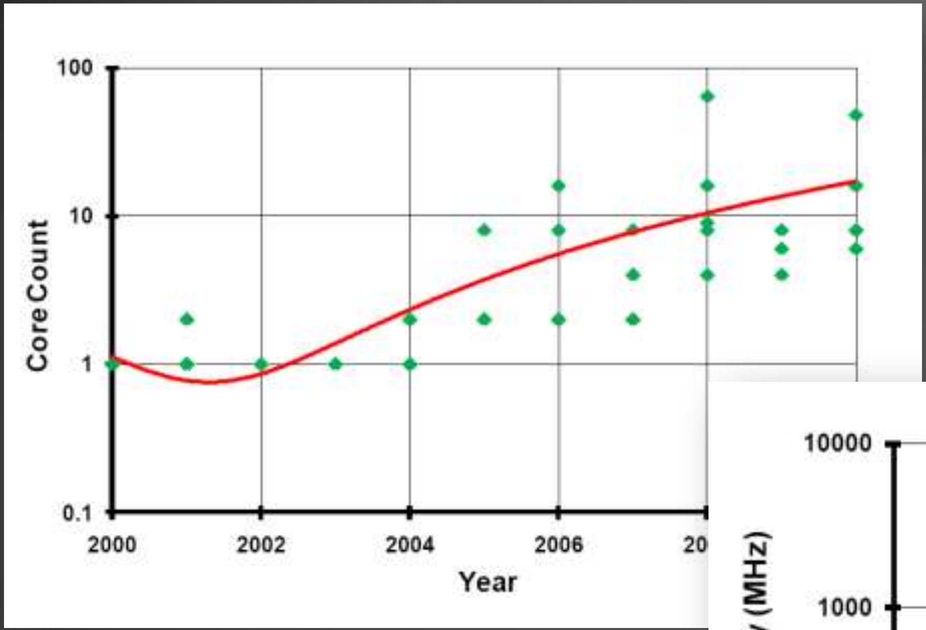
- More GHz, more flops…much More Watt
- Less GHz, less flops … less watt

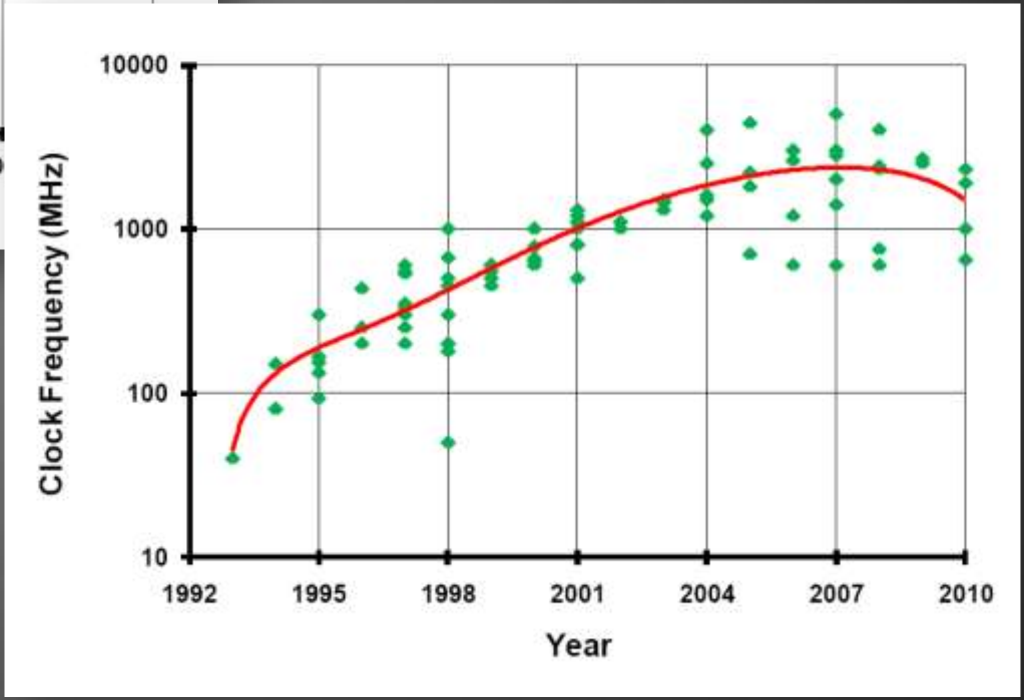
Gain ?

- CPU should be adjustable according to program needs : Why being fast to wqit for others.

Source: ISSCC 2010 TRENDS REPORT

Get more flops with more slower cores





Source: ISSCC 2010 TRENDS REPORT

"A fairly obvious conclusion which can be drawn is that the effort expended on achieving high parallel processing rates is wasted unless it is accompanied by achievements in sequential processing rates of very nearly the same magnitude."

Gene M. Amdahl – 1967

http://www-inst.eecs.berkeley.edu/~n252/paper/Amdahl.pdf

Dark silicon

Currently, the broader computing community is in consensus that we are in "the multicore era." Consensus is often dangerous, however. Given the low performance returns assuming conservative (and to some degree ITRS) scaling, adding more cores will not provide suficient benefit to justify continued process scaling. If multicore scaling ceases to be the primary driver of performance gains at 16nm (in 2014) the "multicore era" will have lasted a mere nine years, a short-lived attempt to defeat the inexorable consequences of Dennard scaling's failure

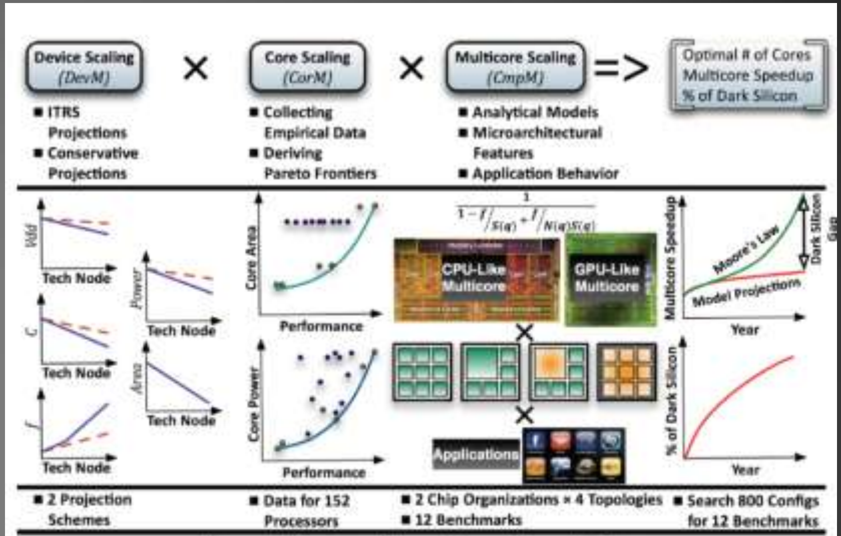


Figure 1: Overview of the models and the methodology

Table 2: Scaling factors for ITRS and Conservative projections.

	Year	Tech Node (nm)	Frequency Scaling Factor (/45nm)	Vdd Scaling Factor (/45nm)	Capacitance Scaling Factor (/45nm)	Power Scaling Factor (/45nm)
ITRS	2010	45*	1.00	1.00	1.00	1.00
	2012	32*	1.09	0.93	0.7	0.66
	2015	22†	2.38	0.84	0.33	0.54
	2018	16†	3.21	0.75	0.21	0.38
	2021	11†	4.17	0.68	0.13	0.25
	2024	8†	3.85	0.62	0.08	0.12
	31% frequency increase and 35% power reduction per node					
Conservative	2008	45	1.00	1.00	1.00	1.00
	2010	32	1.10	0.93	0.75	0.71
	2012	22	1.19	0.88	0.56	0.52
	2014	16	1.25	0.86	0.42	0.39
	2016	11	1.30	0.84	0.32	0.29
	2018	8	1.34	0.84	0.24	0.22
	6% frequency increase and 23% power reduction per node					

*: Extended Planar Bulk Transistors, †:Multi-Gate Transistors

"Dark Silicon and the End of Multicore Scaling" (ISCA'11), Hadi Esmaeilzadehy Emily Blem, Renée St. Amant, Karthikeyan Sankaralingam, Doug Burger

Manage the memory hierarchy

- Caches
- More Caches (non volatile ?)
- Local memory
- More local memories (non volatile)
- Same node memory
- Remote node memory
- Fast Storage SSD
- Fast Storage HDDs
- Permanent Disks Storage (HDDs)
- … tapes …

« Free lunch is over » (2005)

The Free Lunch Is Over: A Fundamental Turn Toward Concurrency in Software

Home | Blog | Talks | Books & Articles | Training & Consulting

On the blog
November 4: Other Concurrency Sessions at PDC
November 3: PDC09: 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 its name is Concurrency.

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

Update note: The CPU trends graph last updated August 2009 to include current data and show the trend continues as predicted. The rest 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, have run out of room with most of their traditional approaches to boosting CPU performance. Instead of driving clock speeds and straight-line instruction throughput ever higher, they are instead turning *en masse* to hyperthreading and multicore architectures. Both of these features are already available on chips today; in particular, multicore is available on current PowerPC and Sparc IV processors, and is coming in 2005 from Intel and AMD. Indeed, the big theme of the 2004 In-Stat/MDR Fall Processor Forum was multicore devices, as many companies showed new or updated multicore processors. Looking back, it's not much of a stretch to call 2004 the year of multicore.

And that puts us at a fundamental turning point in software development, at least for the next few years and for applications targeting general-purpose

Source: http://www.gotw.ca/publications/concurrency-ddj.htm vi main.c

Nb Noeuds

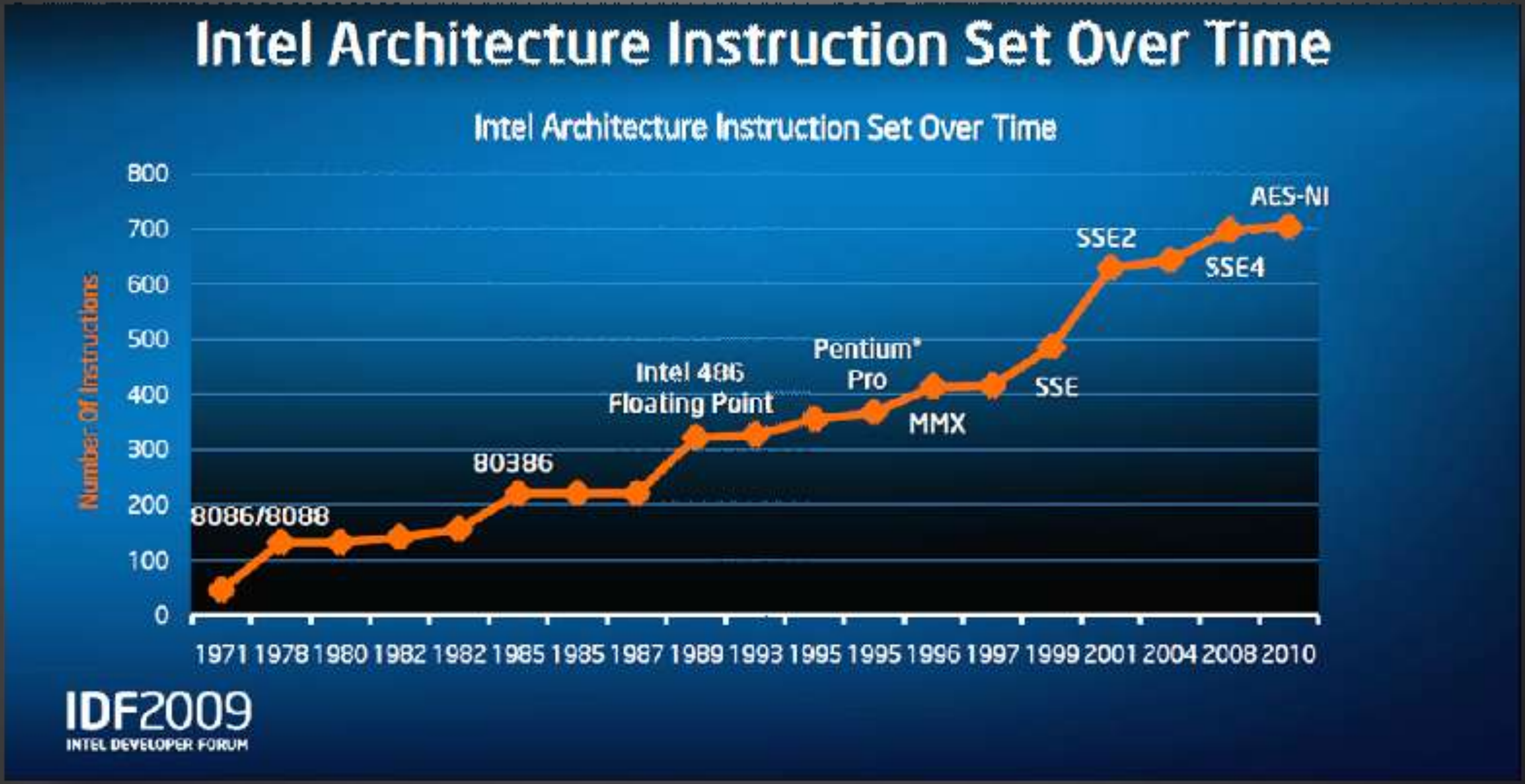
Parallélisme inter-nœud
(mémoire distribuée)

Nb Coeurs

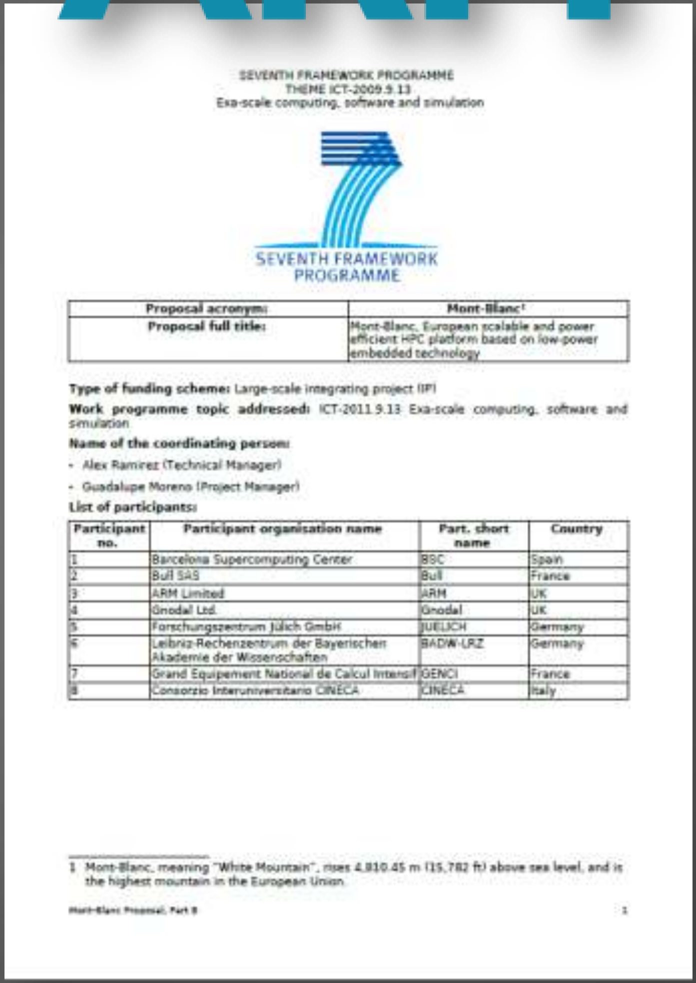
Parallélisme intra-nœud
(mémoire partagée)

Parallélisme au niveau instruction

Get less Watt by simplifying



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:

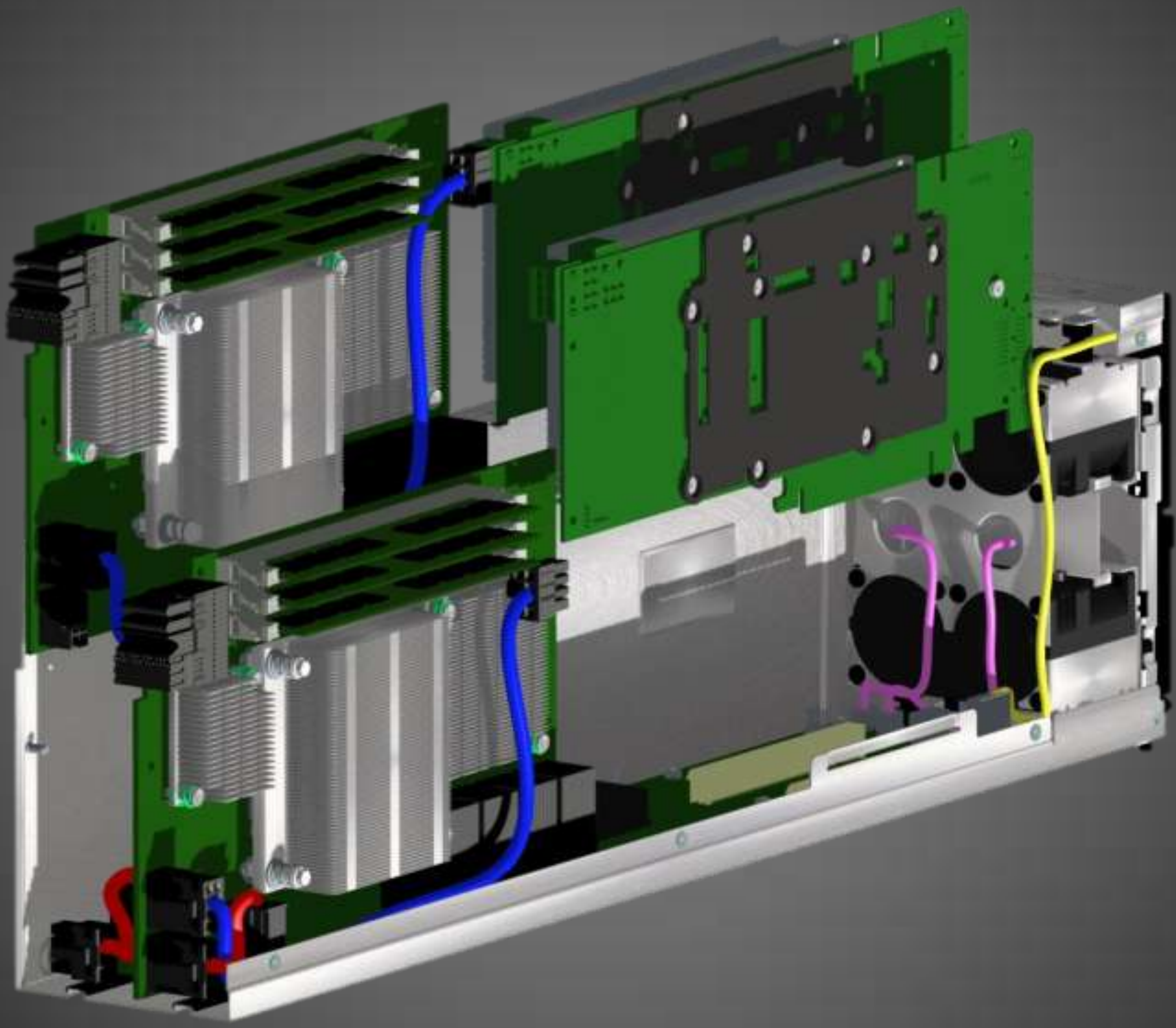
Participant no.	Participant organisation name	Part. short name	Country
1	Barcelona Supercomputing Center	BSC	Spain
2	Bull SAS	Bull	France
3	ARM Limited	ARM	UK
4	Gnodal Ltd.	Gnodal	UK
5	Forschungszentrum Jülich GmbH	JUELICH	Germany
6	Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften	BADW-LRZ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

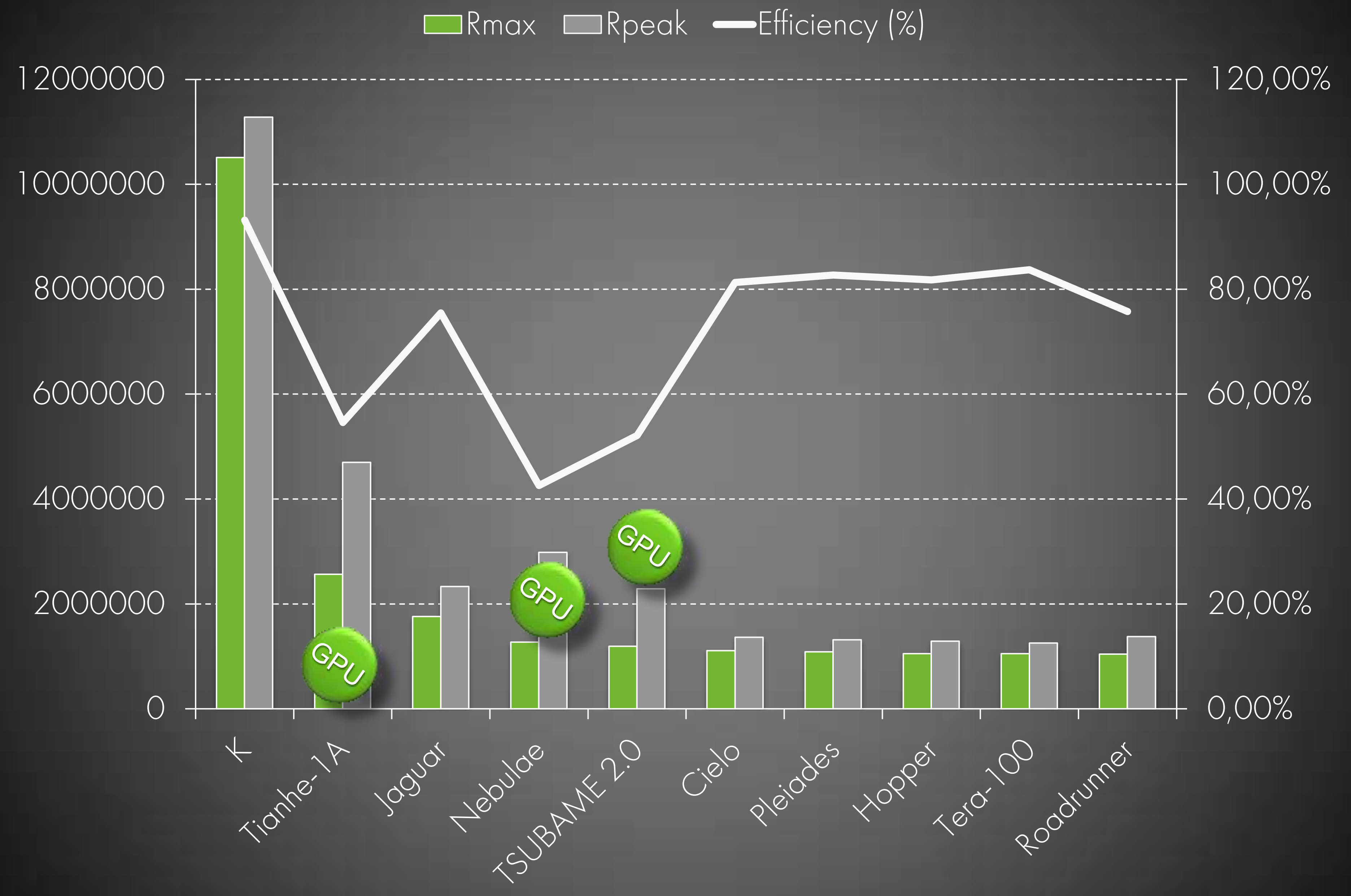
1 Mont-Blanc, meaning "White Mountain", rises 4,810.45 m (15,782 ft) above sea level, and is the highest mountain in the European Union.

GPU throuput cores

512 coeurs à 2,1 GHz (1 flop/cycle)

x86 many cores





Rmax
Rpeak
Efficiency (%)
12000000
10000000
8000000
6000000
4000000
2000000
0
120,00%
100,00%
80,00%
60,00%
40,00%
20,00%
0,00%
GPU
GPU
GPU
K
Tianhe-1A
Jaguar
Nebulae
TSUBAME 2.0
Cielo
Pleiades
Hopper
Tera-100
Roadrunner

vi main.c

Vision 2
Less watt

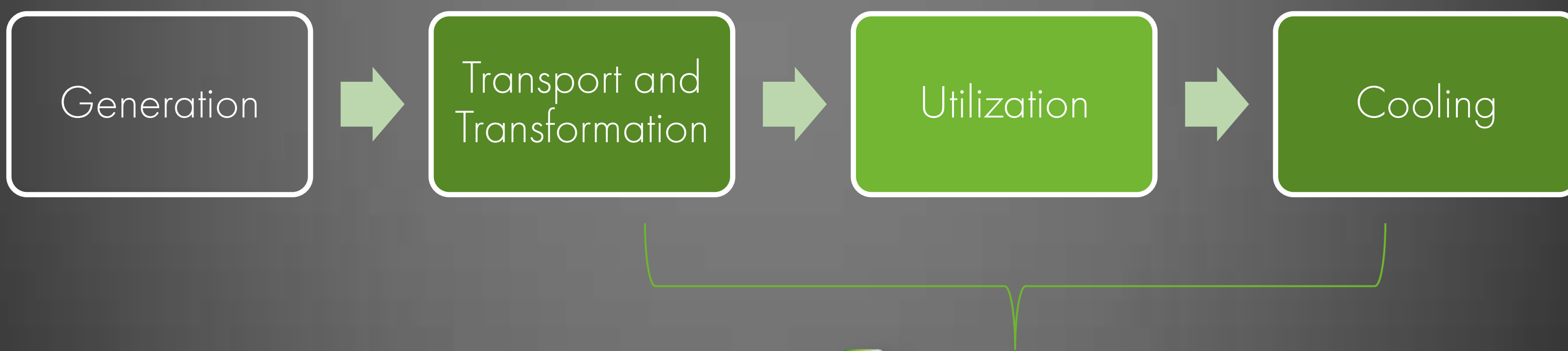
Considering peak performance

Considering Data Center Cost

~~Compute nodes consumption~~

What can we optimize ?

Decomposition




ISSUING COUNTRY
2002 2012

Get more flops with optimized load

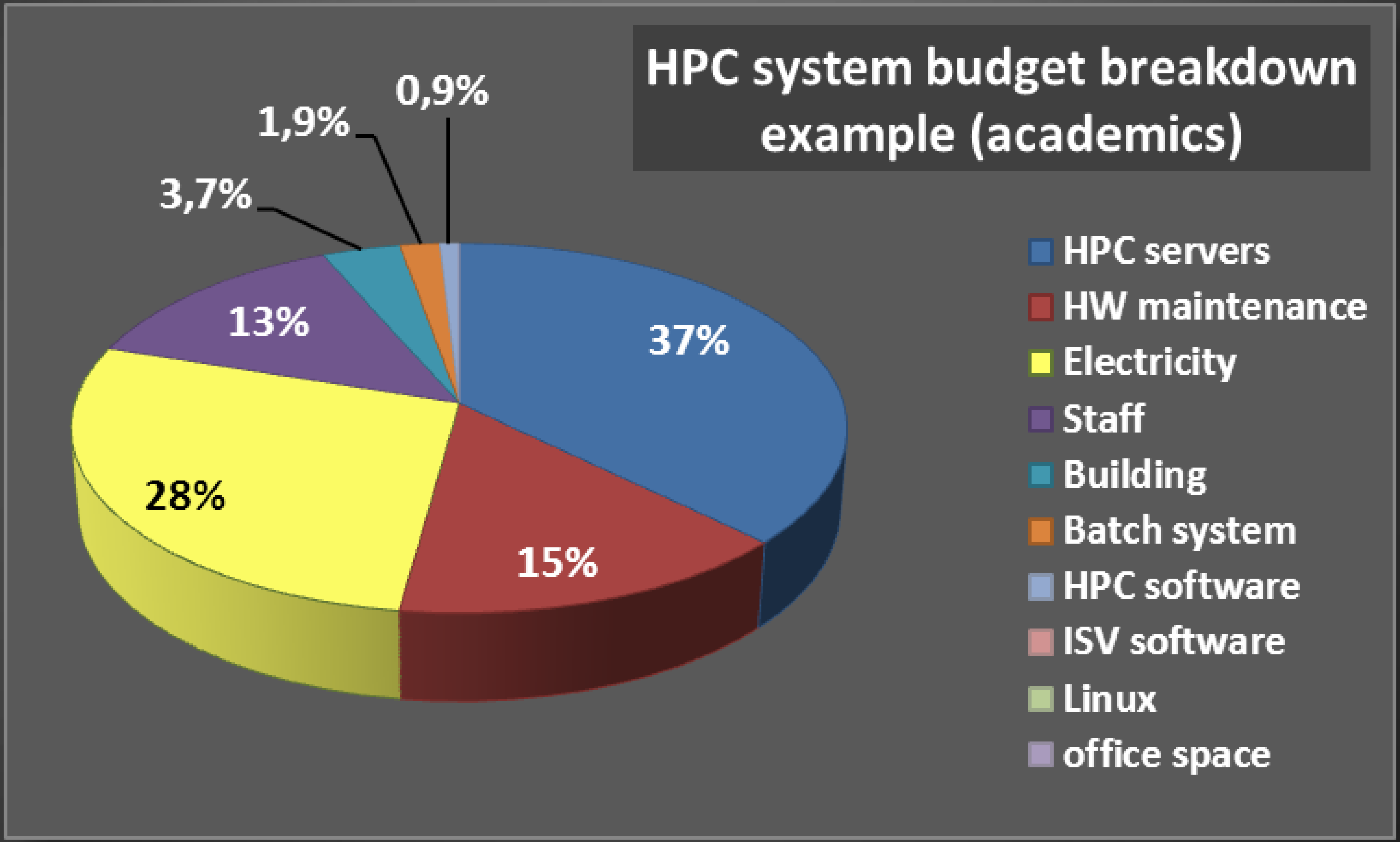
Few %

- HPC machine are heavily used,
- Batch scheduler are efficient
- Shutdown/Restart brings small gains even with NVRAM

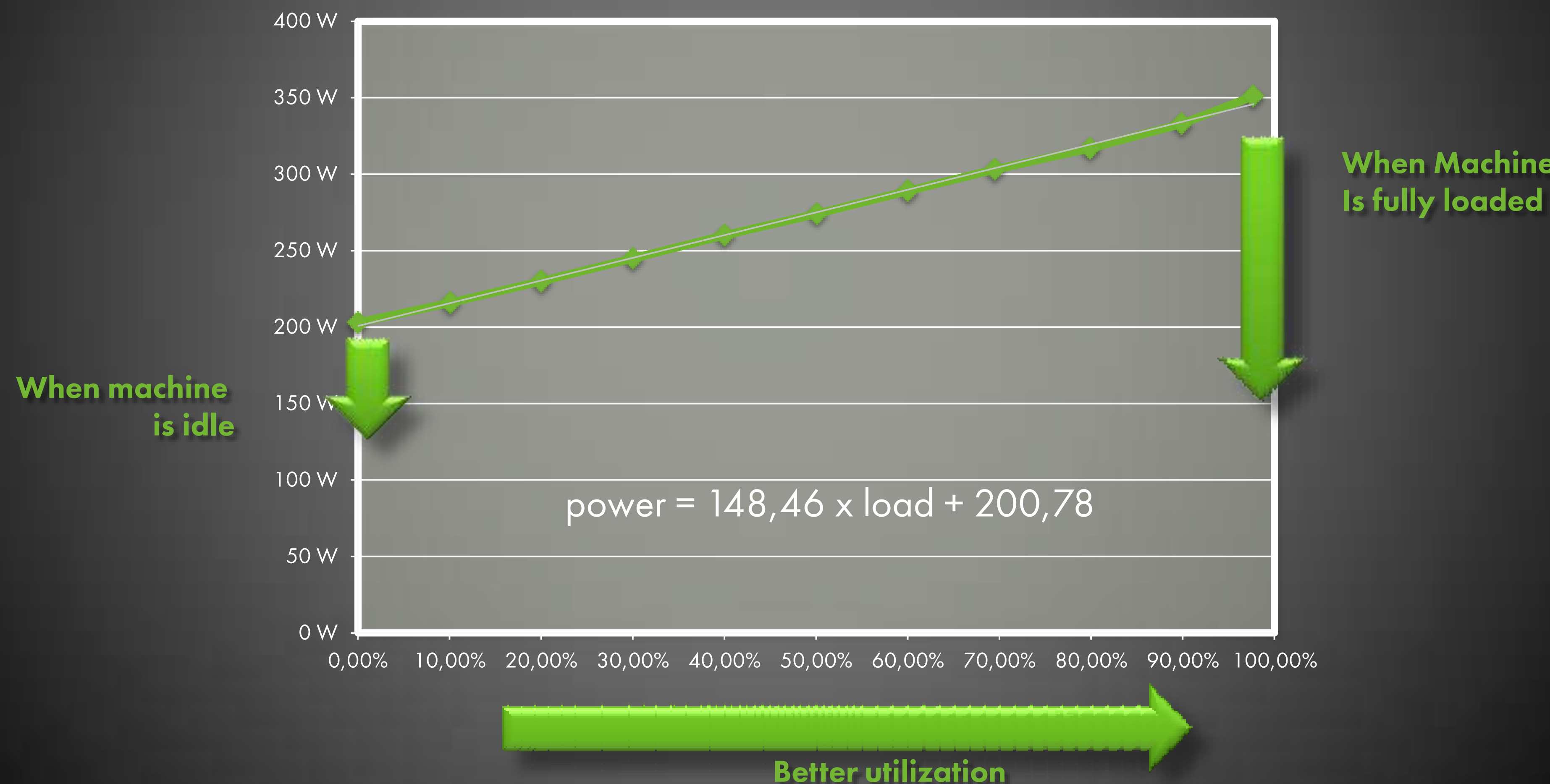
But,

- Fault tolerance may become an issue and waste energy

Jonathan Koomey. 2011. Growth in data center electricity use 2005 to 2010. Oakland, CA: Analytics Press. July

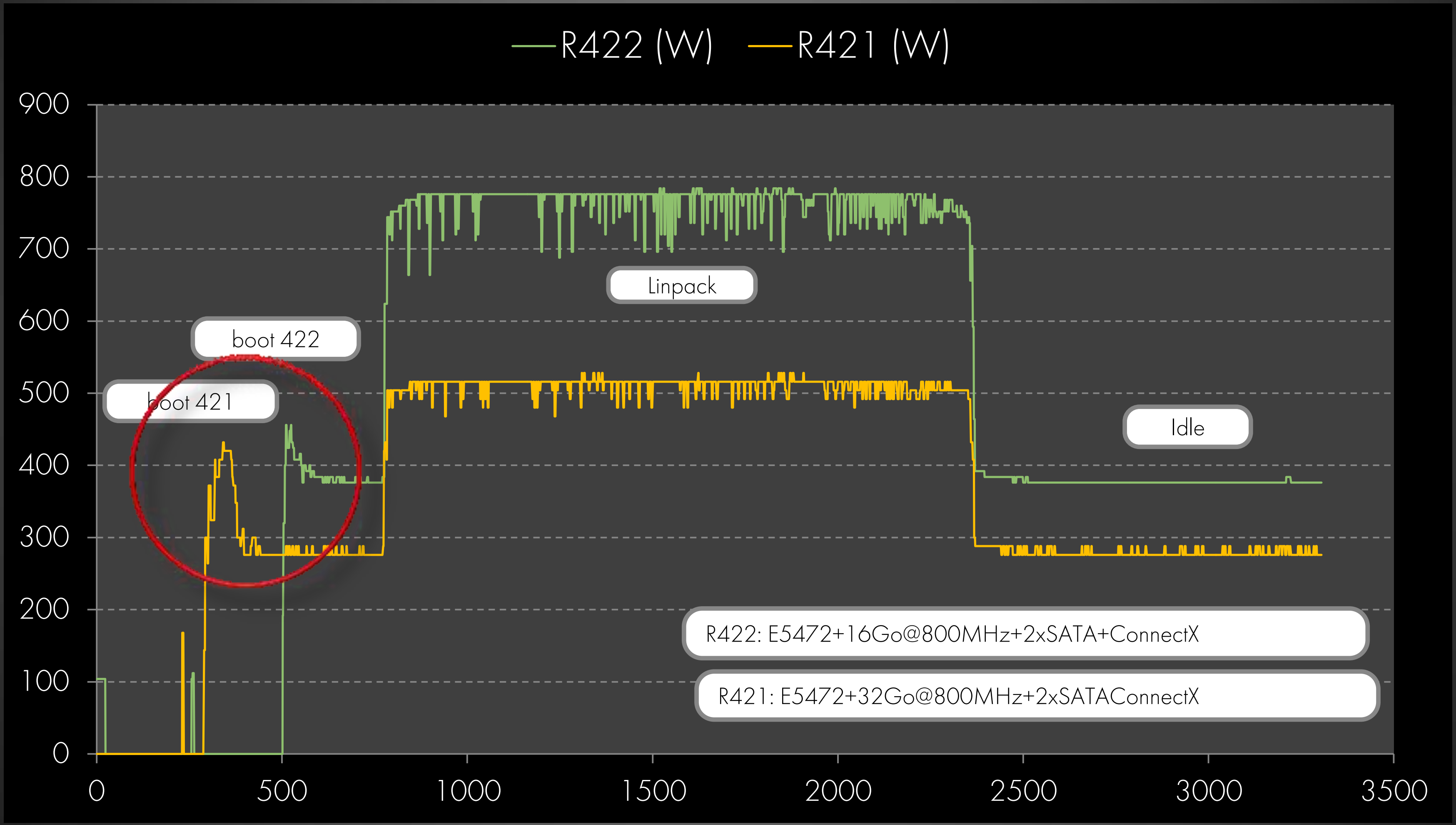

HPC system budget breakdown example (academics)
0,9%
1,9%
3,7%
13%
37%
28%
15%
HPC servers
HW maintenance
Electricity
Staff
Building
Batch system
HPC software
ISV software
Linux
office space

Compute nodes



Westmere

Get less Watt per concentrating



Joule Effect

- Power consumption can be split in two parts:
  - 2/3 are dynamic (when gates are chinging their states) and

$$P_{switching} = Capacitance.Voltage^2.frequency$$

  - 1/3 is static due to leaks

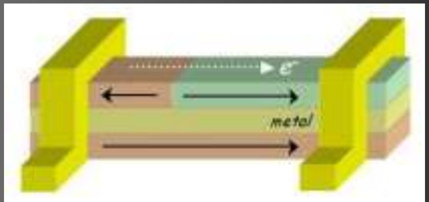
$$P_{stand\text{-}by} = I_{leak} \, . \, Voltage$$

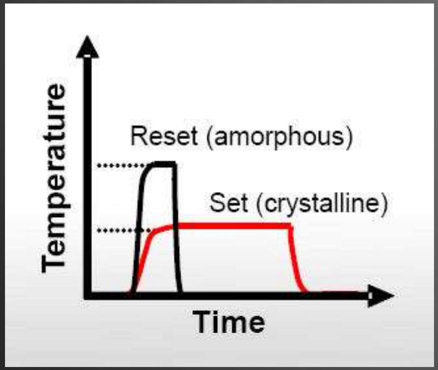
  - Voltage has to be high with high frequency (to have clear rising edges)

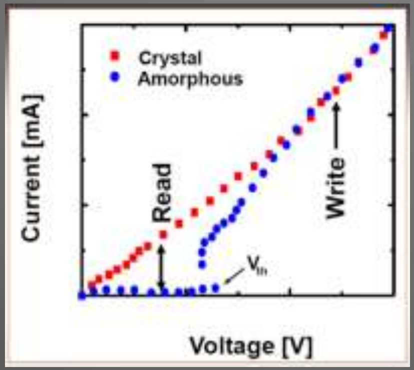
Sources
http://en.wikipedia.org/wiki/CMOS
http://en.wikipedia.org/wiki/Dynamic_frequency_scaling
« Low voltage, low power VLSI subsystems », Kiat Seng Yeo, Kaushik Roy

Get less watt with Memory

- Memory is becoming resistive
- Several technologies : PCRAM, memristor (2008)
- High density, very good cyclability
- Side effect,
  - memory becomes non-volatile.
  - Shudown becomes more efficient (in time and power)

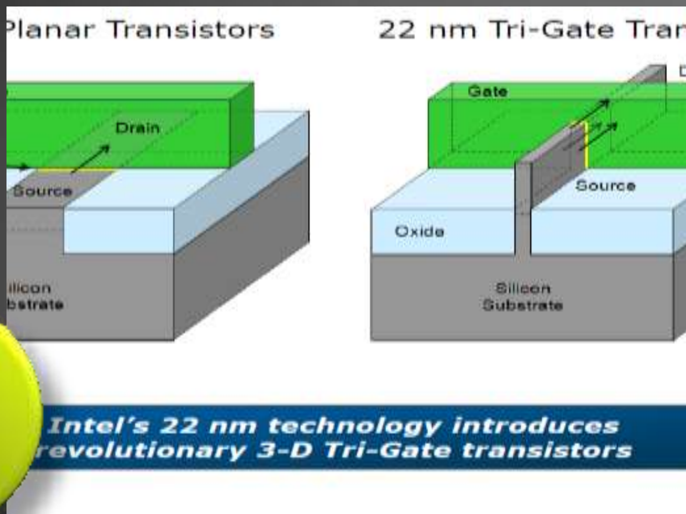






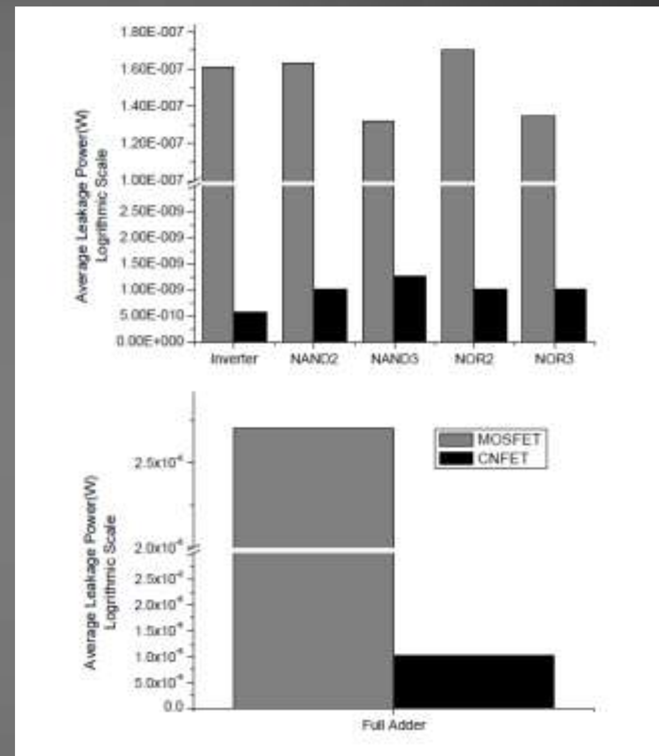
http://agigatech.com/blog/pcm-phase-change-memory-basics-and-technology-advances/
http://julie.grollier.free.fr/memristors.htm

Get less watt with Transistor innovation



x2

- faster (+37%)
- Much less leak
- Lower Voltage
- 50% less power consumption

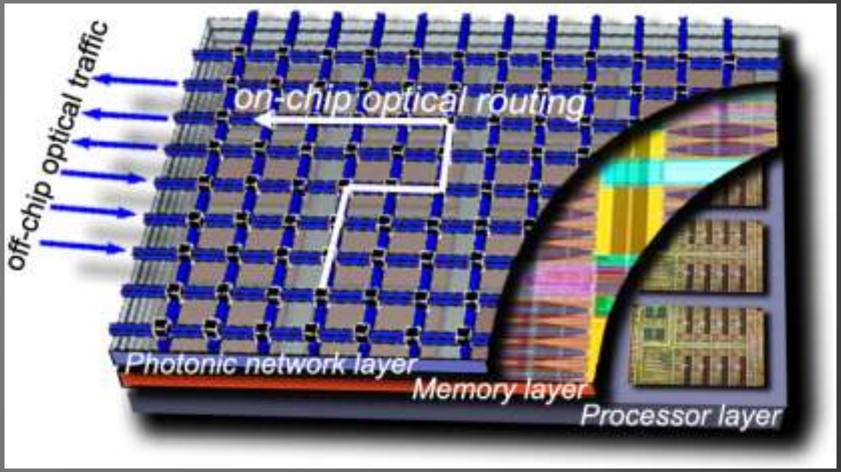


Carbon Nanotubes

Source: Intel Developper forum 2011
http://www.ece.neu.edu/groups/hpvlsi/publication/2009IMTC_CHO.pdf

Get more perf Stacking and photonic

- Higher Bandwidth,
- Lower latency
- Efficient data Moving
- More performance

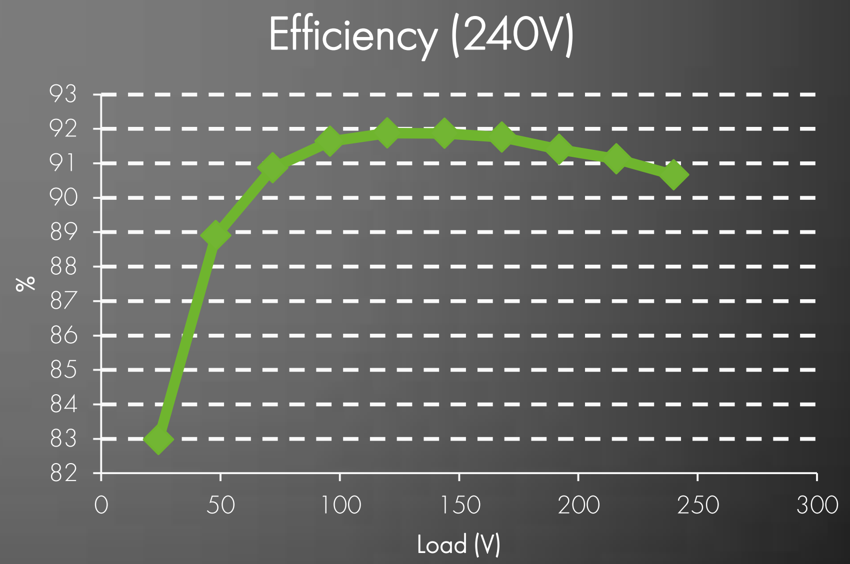


http://www.research.ibm.com/photonics/publications/SEMICON_Tokyo_12_1_2010.pdf

Get less watt with balancing

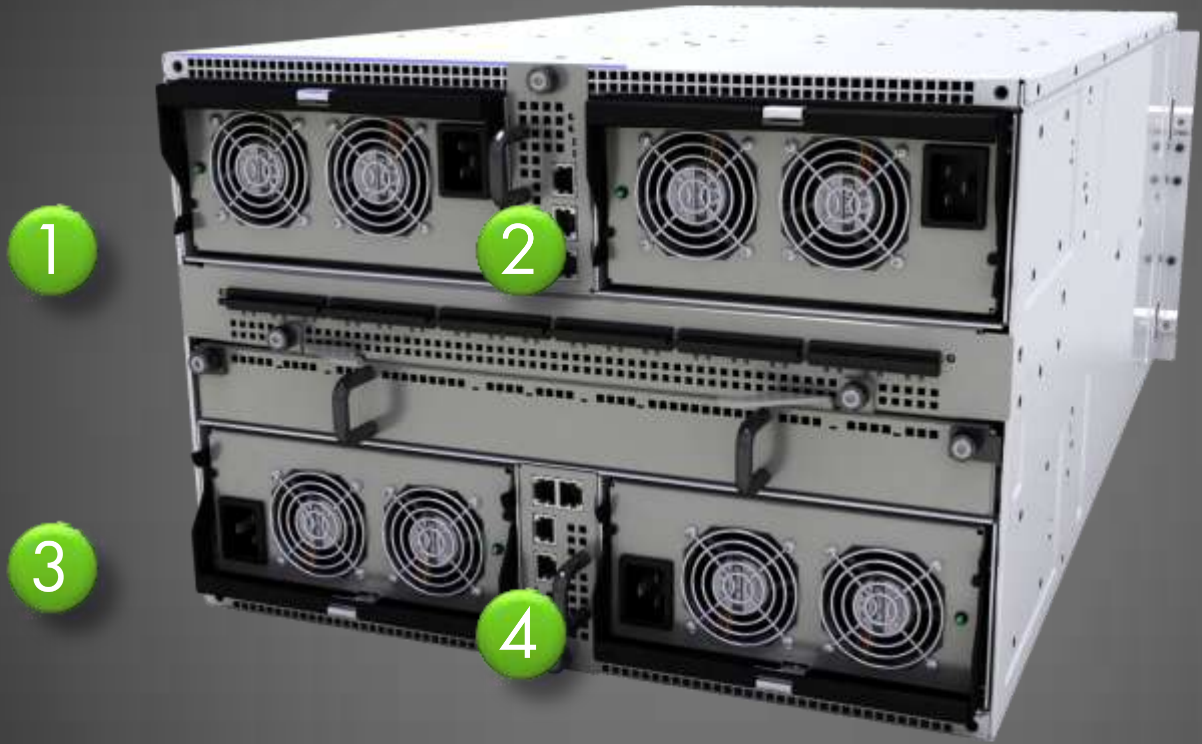
Few %

- 15% of the energy is lost in UPS
- Power supply unit has an efficiency depending on the power
- Many transformations ([1] is great)



[1] http://www.thegreengrid.org/~/media/WhitePapers/IssuesRelatingtotheAdoptionofHVdcPowerintheDataCenter_v1.ashx?lang=en

Get less watt with balancing



Get less watt by removing UPS

15%

Less Watt with power capping

- power consumption for each rack and the whole cluster, with GUI
- action launched when power consumption reaches warning/critical threshold (default action is log + mail only, power capping possible).
  - log + mail when N temperature upper non critical alerts (WARNING) from distinct hosts are catched by the powerManager in less than P seconds (N=3, P=5 by default).
  - log + mail when N temperature upper critical alerts (CRITICAL) are catched by the powerManager in less than P seconds (N=1, P=1 by default).
  - a mail is sent when the consolidated value of a rack reaches a customizable threshold (default is 50 °C).

- All is configurable

More power with power capping (2)

- Data center has an overall Energy Budget
- Energy is spread over components (compute, UPS, cooling, ...)

- The power capping must ensure the consumption is always less than the budget.
- The power capping must ensure the service is as high as possible.

- Winter: less cooling, more compute power
- Summer: more cooling, less compute power.

Froid

Steve Iunker

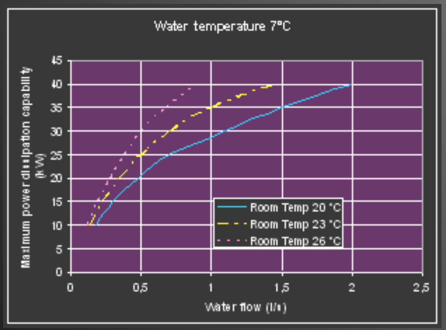
Less watt with better cooling

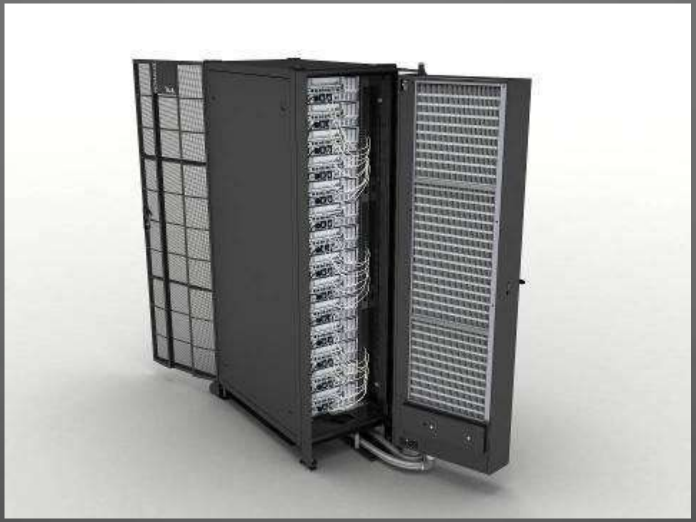
- "Pollution, we concentrated it or spread it"
- We spread it, melt it for a long time.

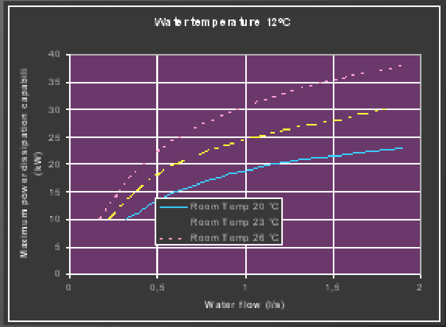
- However, we can have
  - Larger fans
  - Managed fans

Less watt with water cooling

- T° air front == T° air back
- 40 kW can be absorbed in stead of 12kW (air)
- Many, many, sensors…T° air, T° inlet, T° outlet, pressure, speed, … can be used for « event correlator ».
- Gain depends of the context



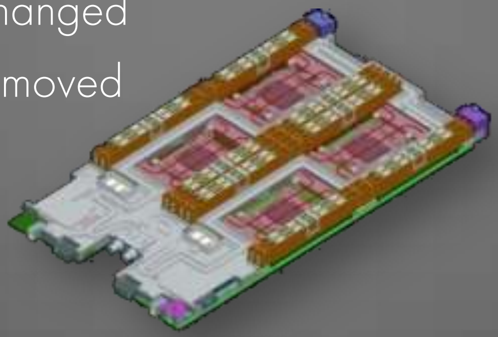




Less € with Direct liquid cooling

30%

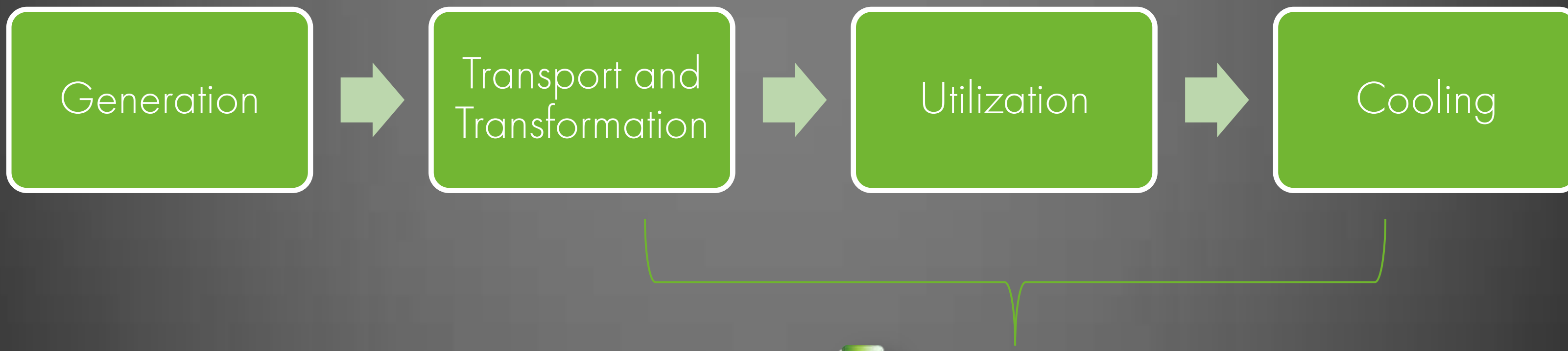
- Water has closed as possible to the heat source
- Water can be hotter (as delta T is key)
- Room can be hotter (remove CRAC)
- But, maintenability is key !vNo change in maintenance process
  - CPU can be changed,
  - DIMM can be changed
  - Blades can be removed



Vision 3
Less Impact

Architect of an Open World

Decomposition

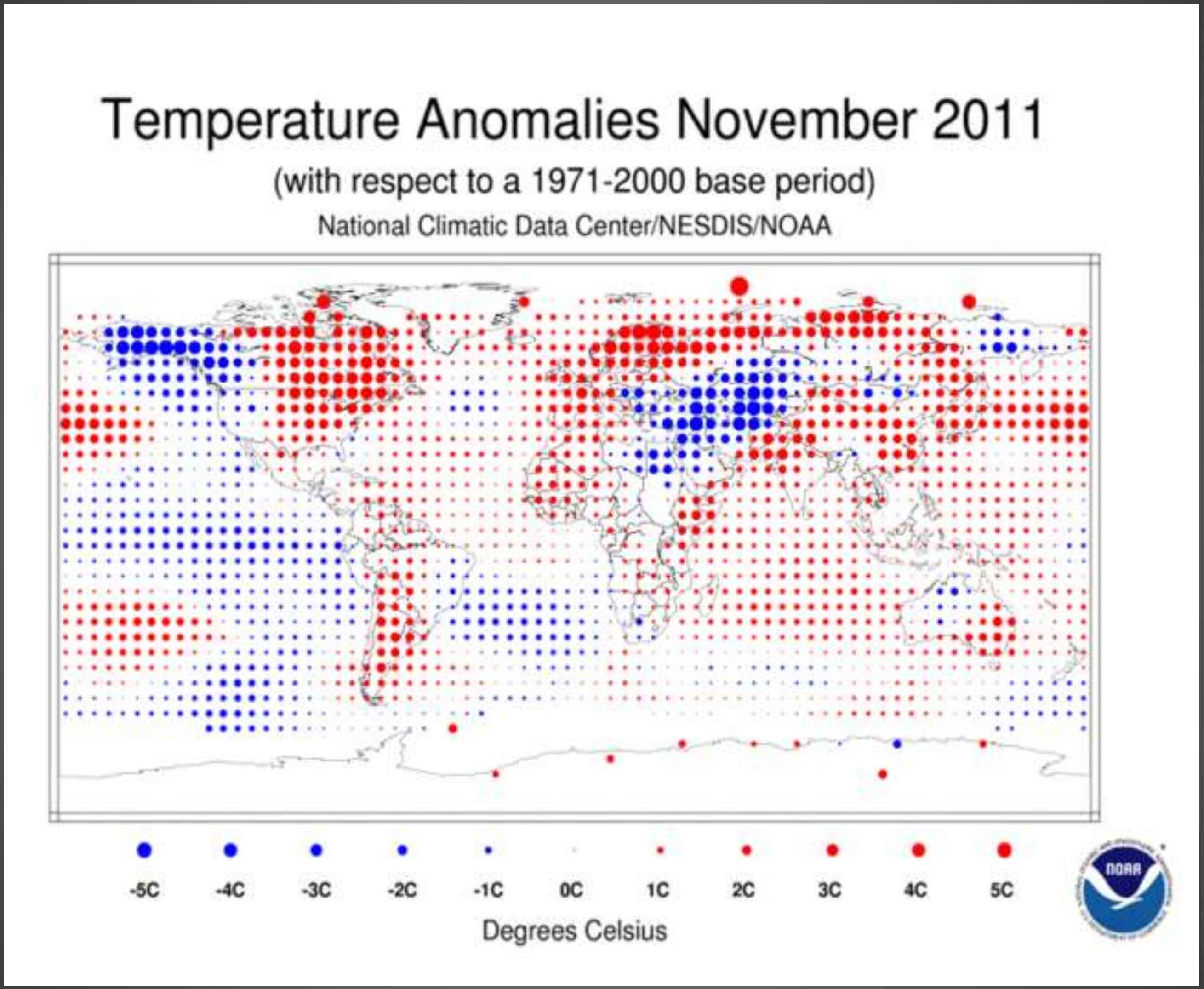


Considering Real ~~peak~~ performance

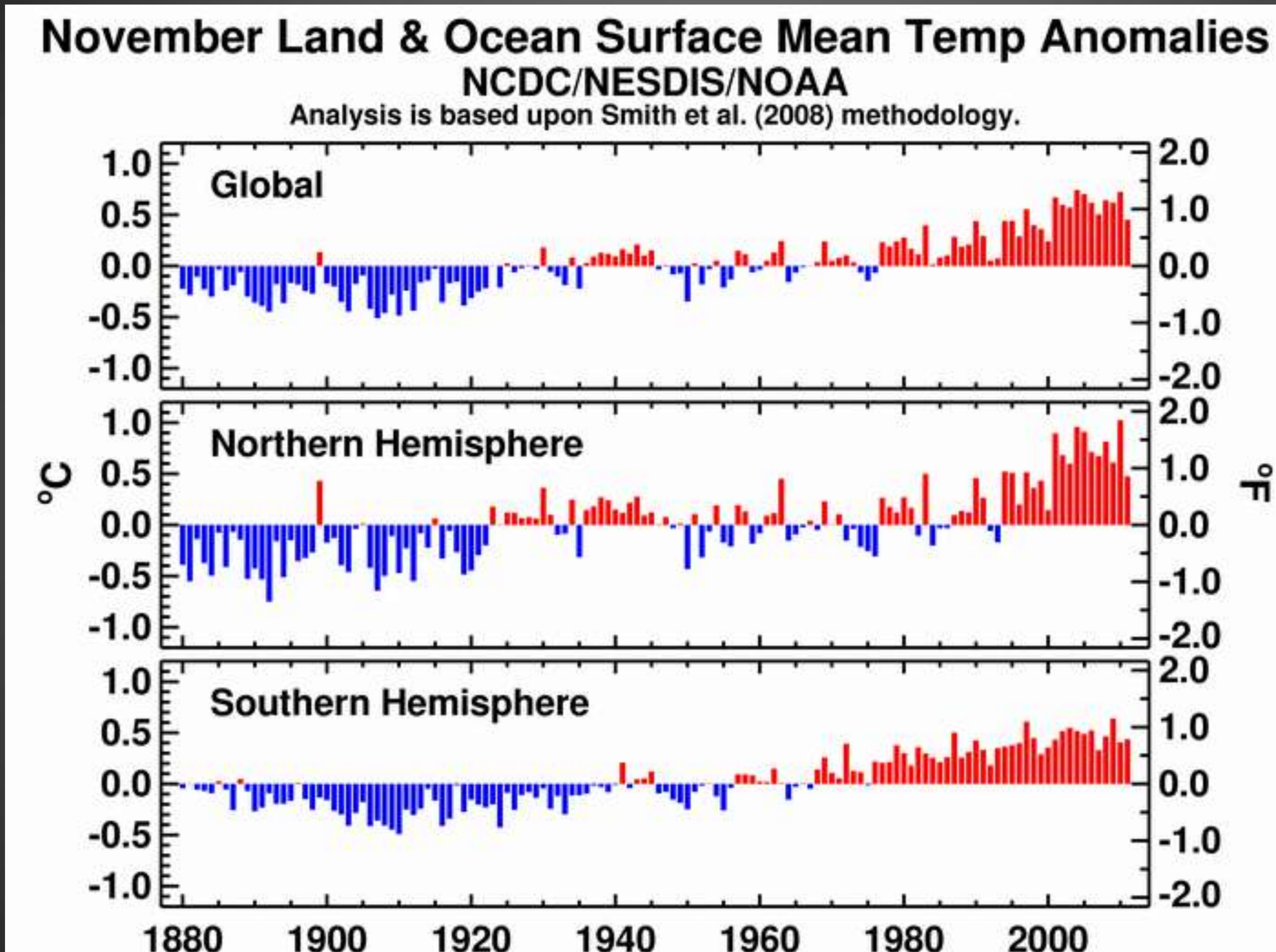
Considering power production

~~Compute nodes consumption~~

What can we optimize ?

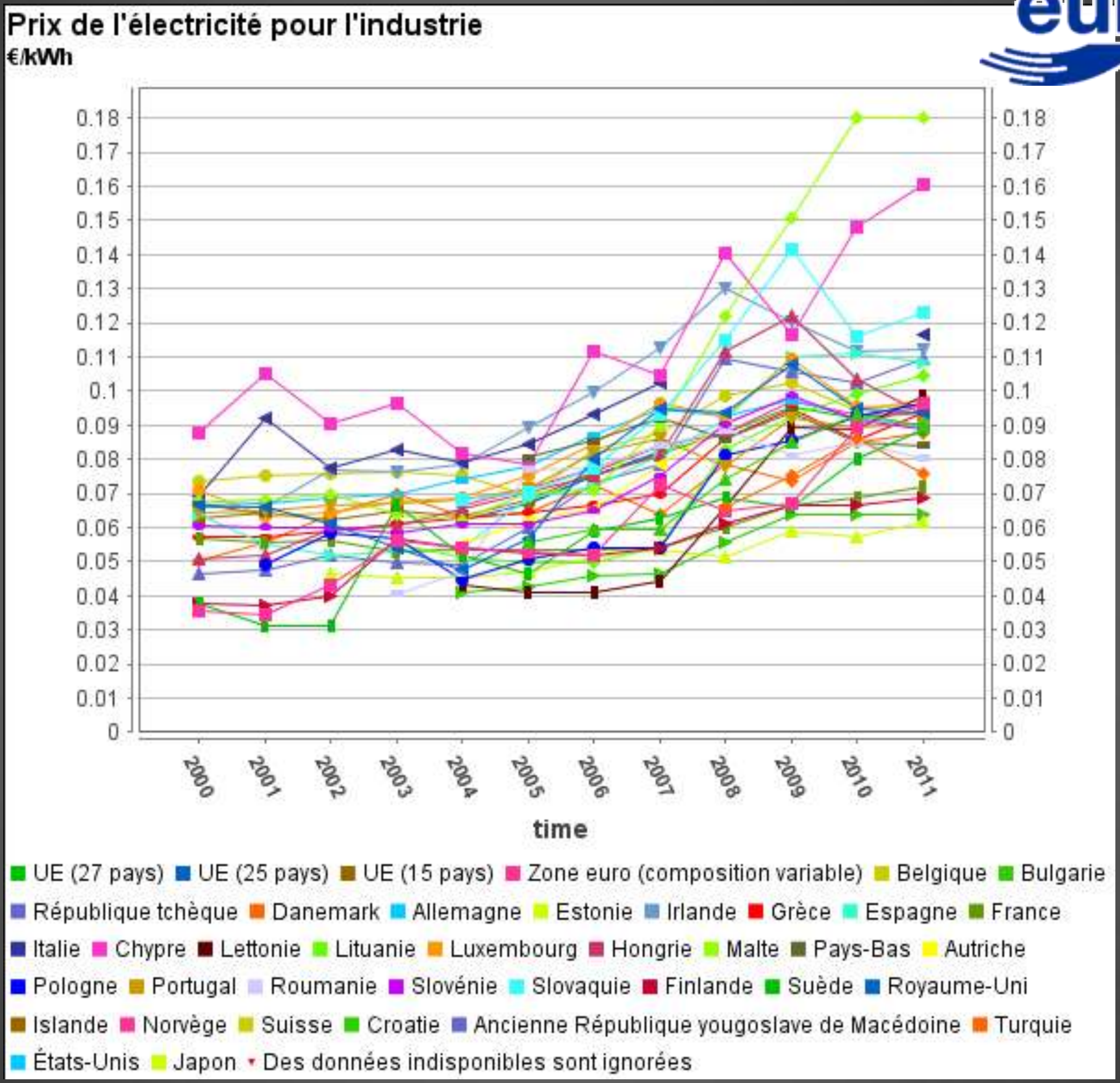


Source: http://www.ncdc.noaa.gov/sotc/global/



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

Generation



Source: http://epp.eurostat.europa.eu

Re-Nuclear or Renewable

- Cost of Energy in France is lower in 2010 than 1995 (in constant Euro) But, increase in the last two years is higher than inflation
- Prediction: +30% in 2016 (for end user)

- Nuclear:
  - Cost of Maintenance
  - Cost of security

- Renewable
  - Cost of dismantling
  - Cost of new plant
  - Need complementary energy (wind may vary), production must be equal to consumption.

http://www.assemblee-nationale.fr/12/rapports/r3277.asp

Nuclear plant efficiency

today	EPR	coal
33%	36%	40%

No cogeneration in France

Transport & Transformation

Leaks and thief

$$P_{joule} = RI^2 = R.P_{elec}{}^2/3U^2$$

- Leaks during transportation is weak, because U is huge
  - in France 12TWh (2,5%) is lost here
- Leaks during distribution, in 2005, in France,
  - 18TWh (5,3%) are lost by distribution
    - 2/3 is technical (P_{joule})
    - 1/3 is non technical ☺

http://fr.wikipedia.org/wiki/R%C3%A9seau_%C3%A9lectrique
Jean-Michel Tesseron, « Les pertes des réseaux électriques: estimations et achats », dans *ACTU SEE*, Société de l'Electricité, de l'Electronique et des Technologies de l'Information et de la Communication, décembre 2006
http://www.rte-france.com/fr/developpement-durable/les-engagements/indicateurs/volet-economique-2/pertes-electriques-du-reseau

One solution: decentralized production

Enfin, il y a une solution radicale au traitement des pertes, c'est la production décentralisée. L'un des avantages de l'éolien, du photovoltaïque ou de la petite hydraulique par exemple est d'être produits sur les lieux de consommation réduisant à zéro la longueur des lignes d'acheminement et donc les pertes créées par celles-ci.

« Perte d'énergie dans les réseaux de distribution d'électricité » Mathias Laffond – Juin 2009
http://mathiaslaffont.files.wordpress.com/2011/01/rdv-tel-27_04_09.pdf

Tri generation

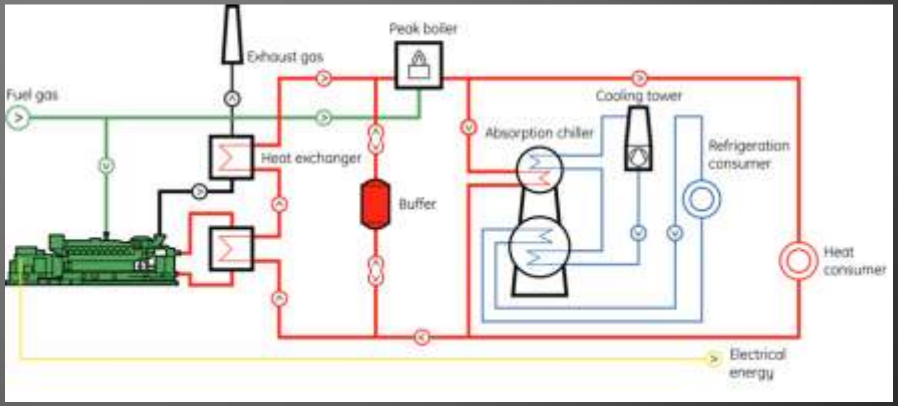
x2,3

CCHP

- One engine produces:
  - Electricity
  - Cold water
  - Hot water

80%

- Efficient
- Local

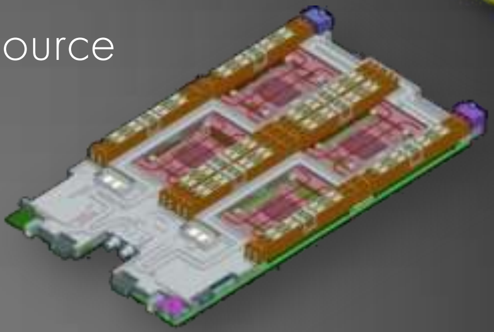


chiller

Less € with Direct liquid cooling

A lot

- Water has closed as possible to the heat source
- Water can be hotter (as delta T is key)
- But, maintenability is key !



- Hot watter can be reused
- I hate free cooling

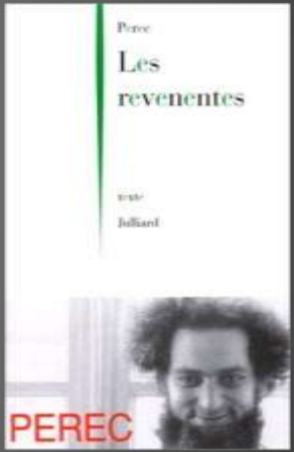
Conclusion

Les contraintes font la créativité

"Constraints Feed Imagination"
George Perec, OULIPO, 1969
(1936 – 1982)

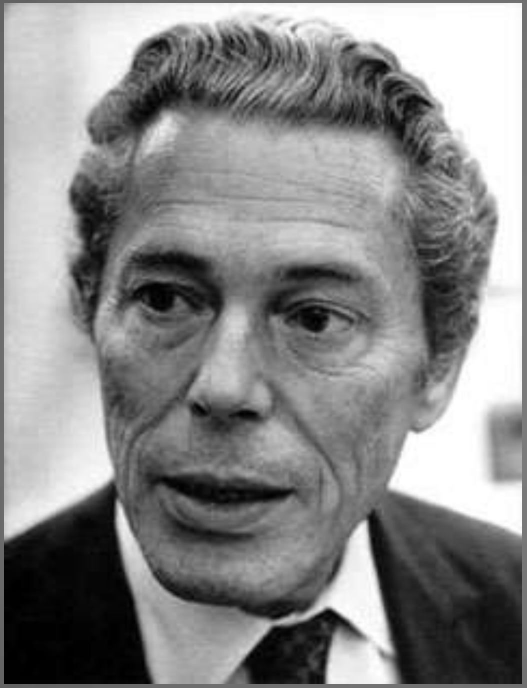
La disparition
315p - 1969

Les revenentes
1972



Jacques MONOD

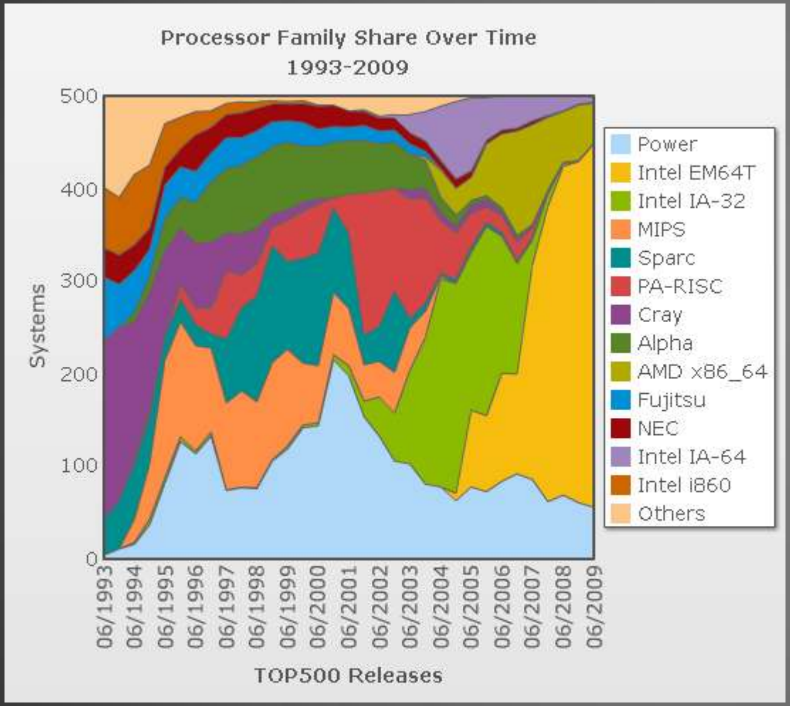
1965: prix nobel de médecine

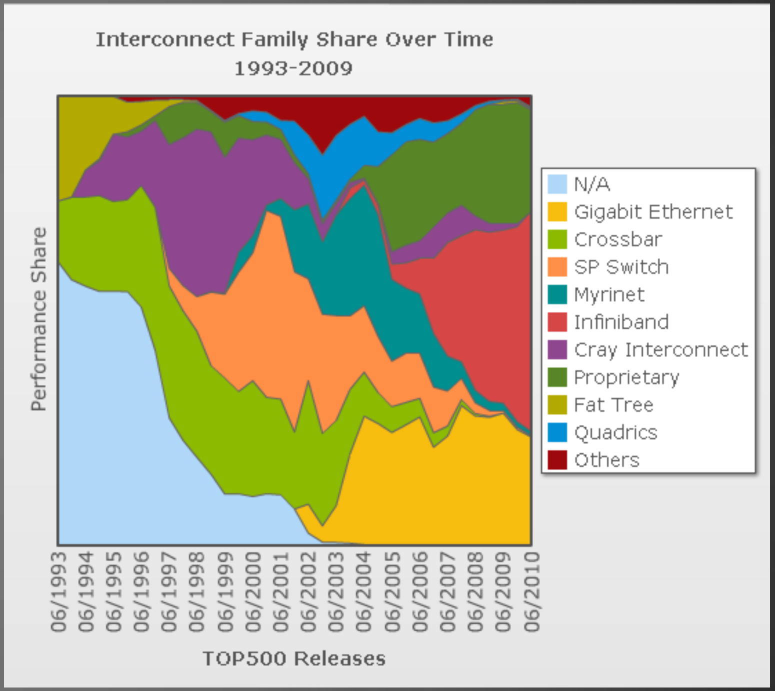


1910 - 1976

1970

C'est pas la première extinction





http://www.bull.com/fr/emploi/recrutement/stages.php

Xavier.vigouroux@bull.net