High Performance Computing

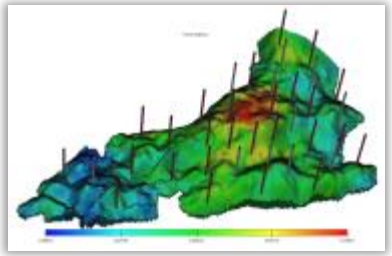
2012/11/19

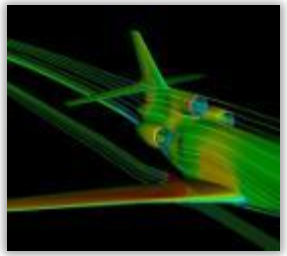
Xavier Vigouroux
Business Development Manager

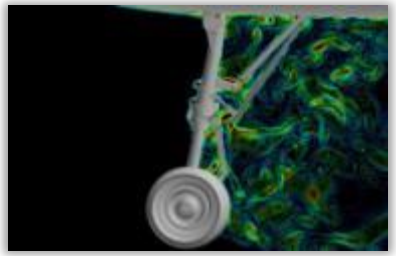
Eco Impact in HPC
Where is it worth to invest ?

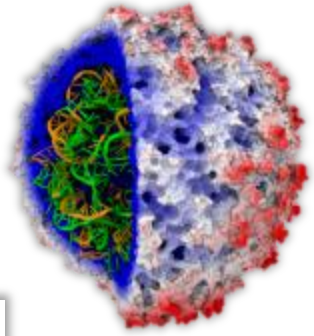
It's worth to invest !

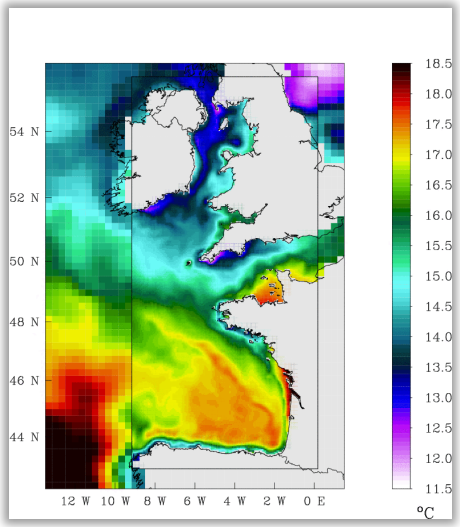
Yes ! It's worth to invest

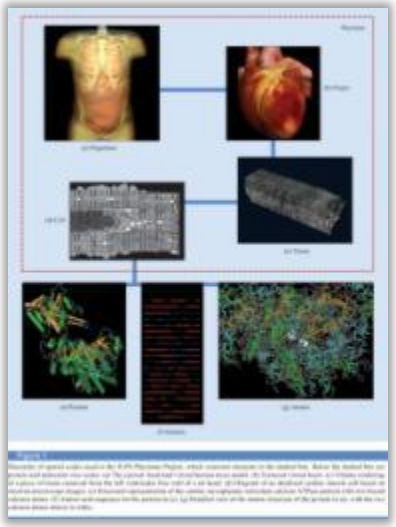






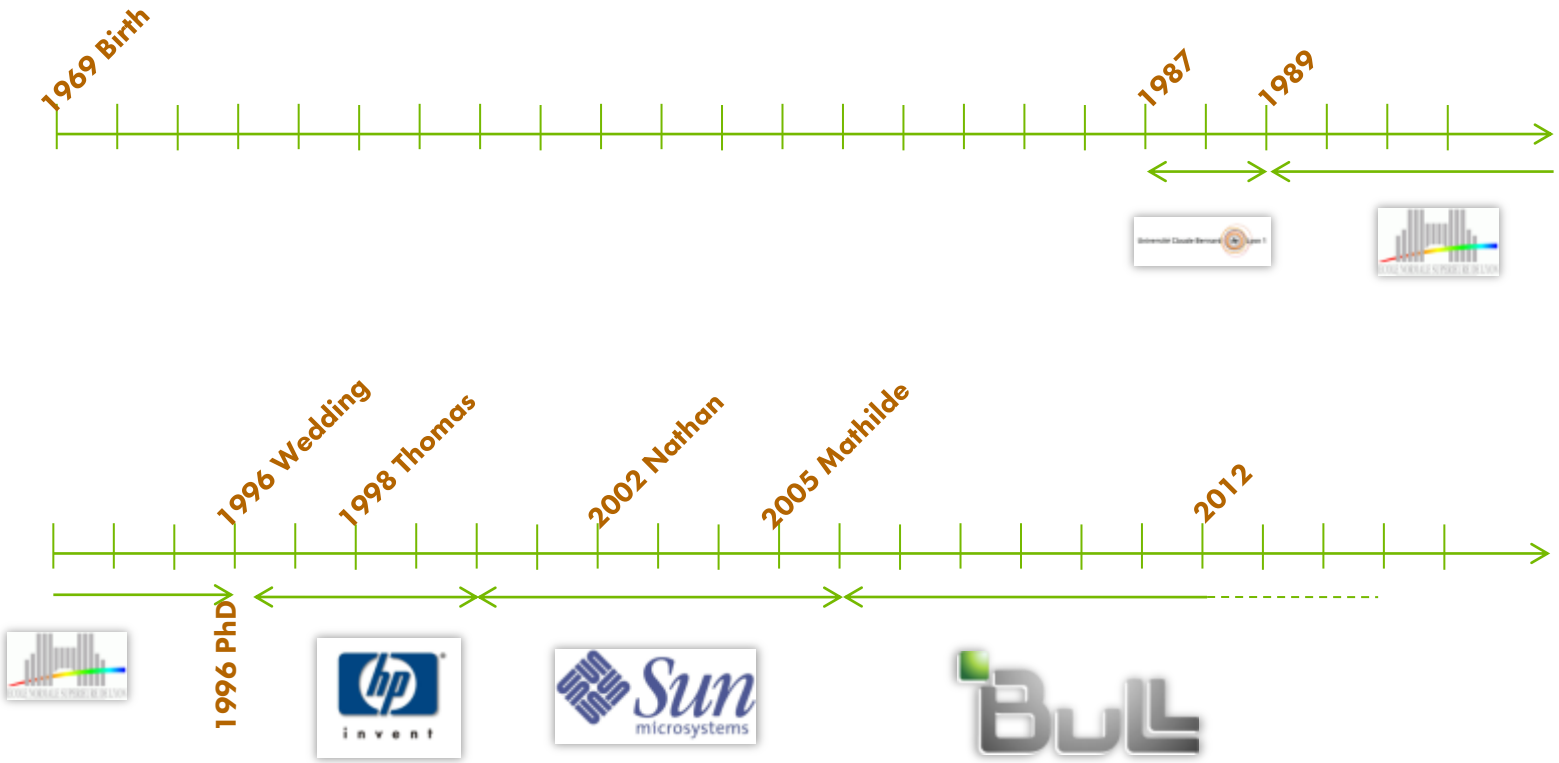






Experimentation
Modelisation
Simulation

Who am I



Bull: European leader in mission-critical digital systems

9,000 EXPERTS

recognized worldwide
in secure systems

OPERATING IN 50 COUNTRIES

Efforts in research in 2011 ***+23%***

€1.3bn REVENUES

+4,6%

growth in 2011

+29%

growth in profitability in 1st quarter 2012

HPC

How large is a machine ?

top500.

m.

Linpack

$$Ax = b$$

Flops/s

Top 3

16 325 Tflops/s

20 132 Tflops/s

1 572 864 PowerPC cores

7.9 MW

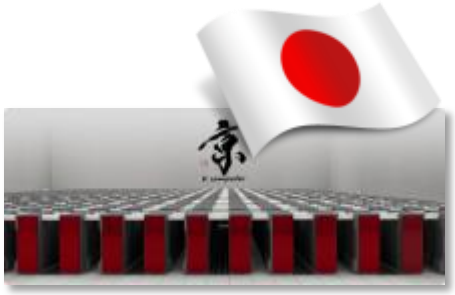
17 590 Tflops/s

27 112 Tflops/s

299 008 cores AMD

261 632 cores GPU

8.2 MW



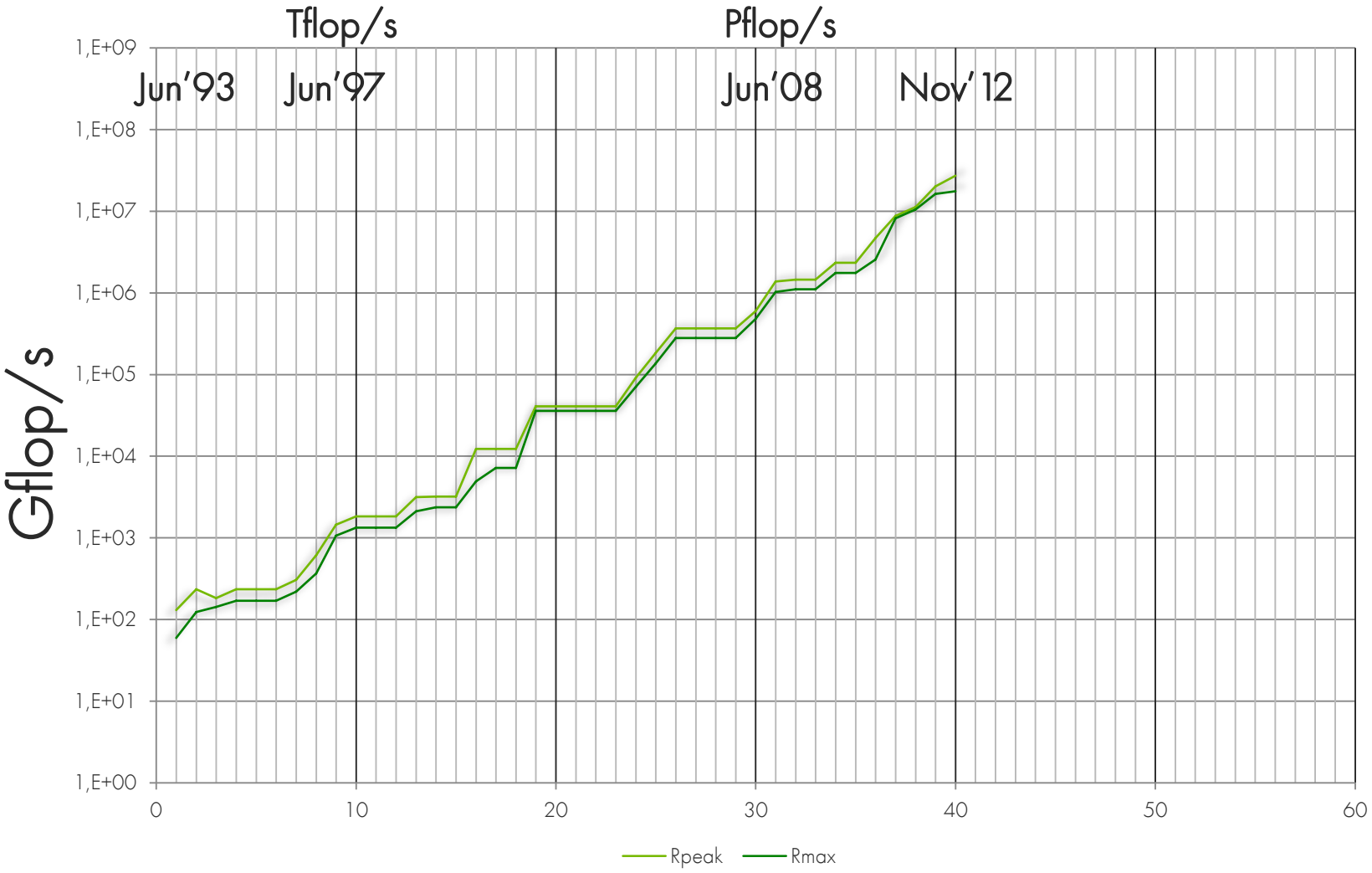
10 051 Tflops/s

11 280 Tflops/s

705 024 sparc cores

12.7 MW

TOP500



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 m² 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 m² footprint

IFERC

- 1,5 PetaFlops
  70 000+ Xeon cores
- 280 TB memory
- 15 PB disk storage
- 120 GB/s IO throughput
- 200 m² footprint

cea

GENCI

PRACE

FUSION FOR ENERGY

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

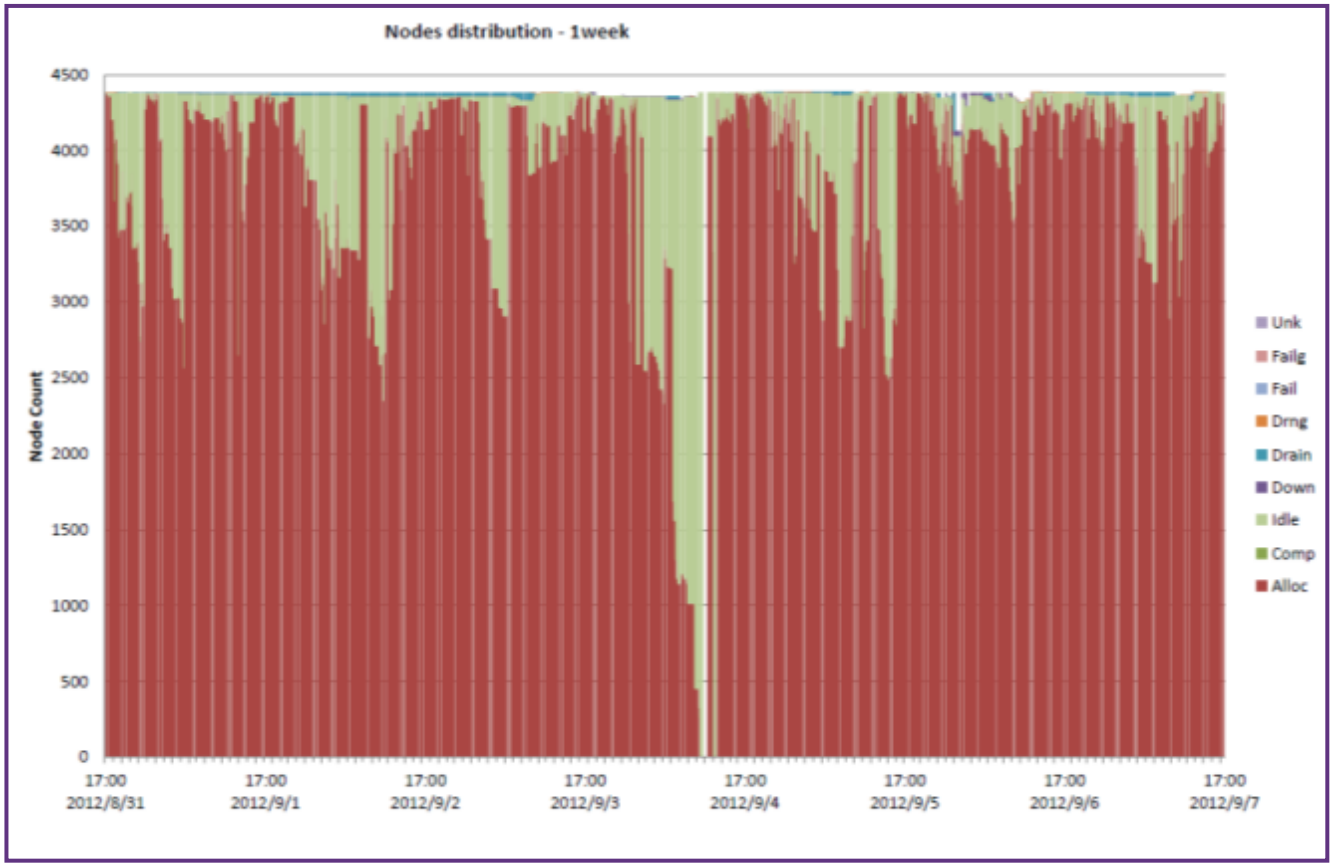
#5

#2

Source: http://www.top500.org – list nov'12

Efficiency today

Load the supercomputer



Ramp up the users

Tier 0

- 1 Petaflop/s – 50000 cores
- European, PRACE
- TGCC, IFERC

Tier 1

- 100 Teraflop/s – 5000 cores
- National

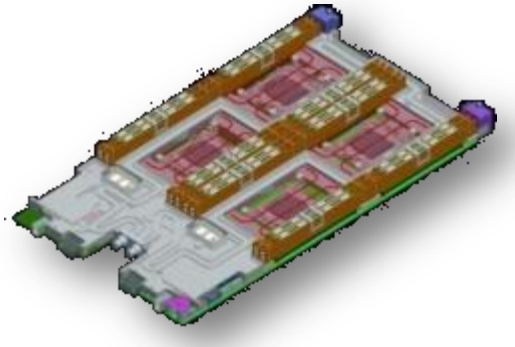
Meso Centres

- 10 Teraflop/s – 500 cores
- Regional

Cooling

12kW/rack

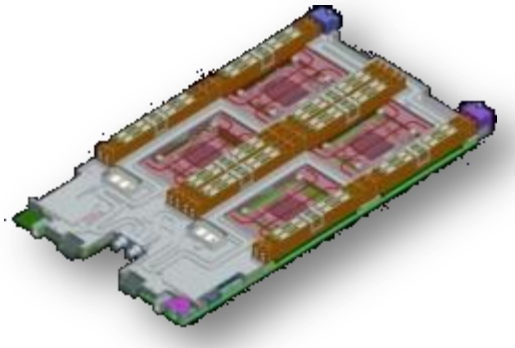
6°C – 30KW



30°C – 80KW

Cooling

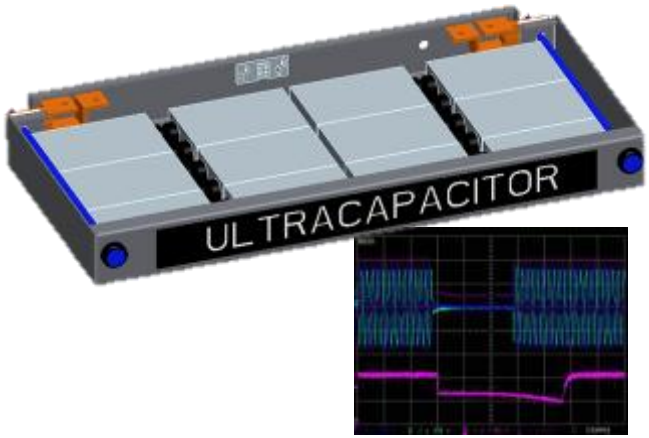
- Water has closed as possible to the heat source
- Water can be hotter (as delta T is key)
- Room can be hotter (remove CRAC)
- But, maintainability is key ! No change in maintenance process
  - CPU can be changed,
  - DIMM can be changed
  - Blades can be removed



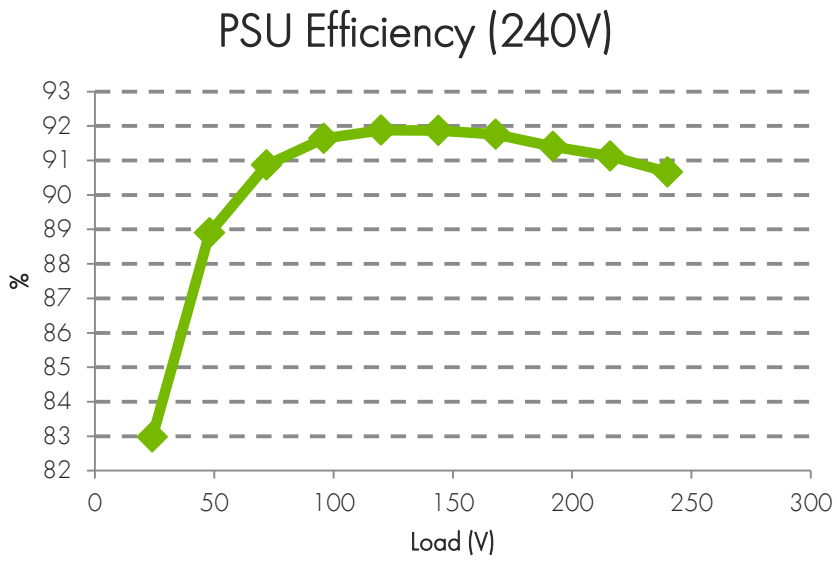
30°C – 80KW

Power Feeding

15%

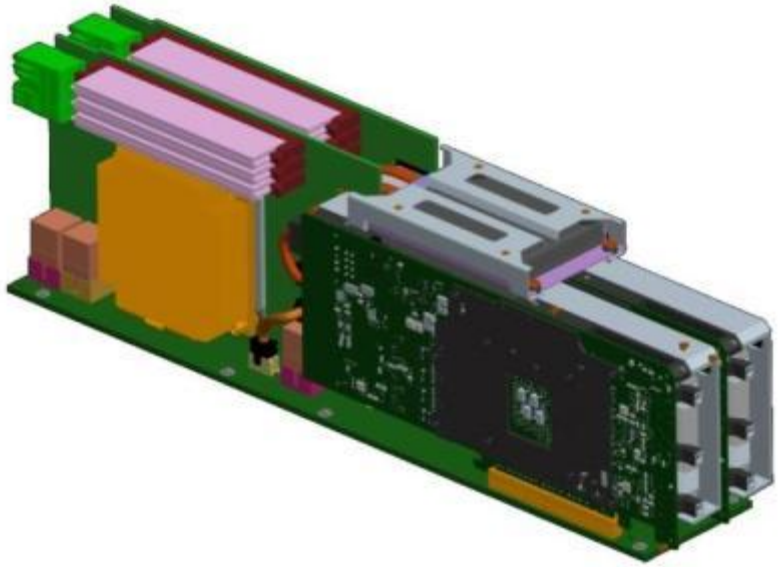


10%



Efficient Hardware

2 x CPUs 2x GPUs/Xeon Phis



Bull - Energy aware batch scheduler

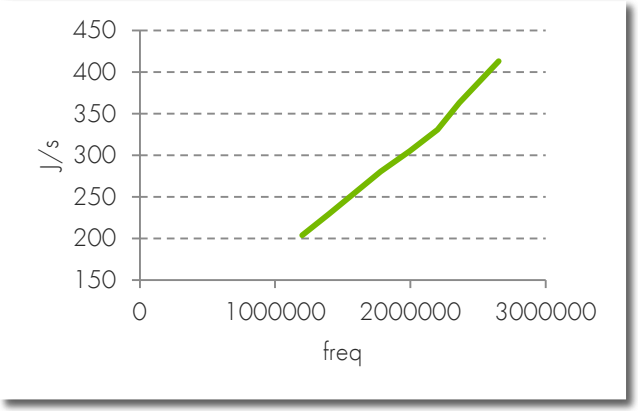
Fix the CPU frequency

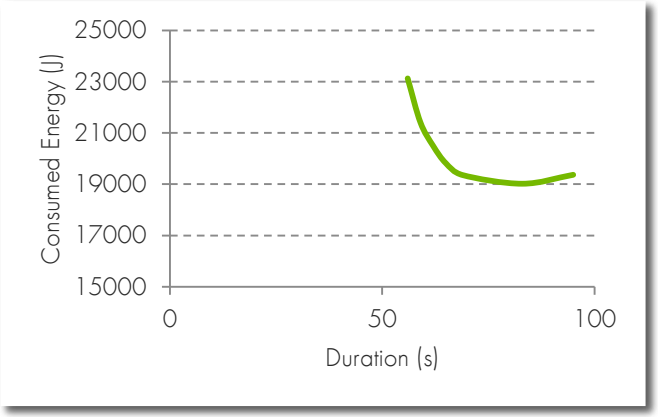
```
$# srun --cpu-freq=2700000 --resv-ports -N2 -n64 ./cg.C.64&
$# sacct -j 58 -format=jobid,elapsed,aveCPUFreq,consumedenergy

       JobID    Elapsed AveCPUFreq ConsumedEnergy
------------ ---------- ---------- --------------
          66   00:00:49    2640340          19668
```

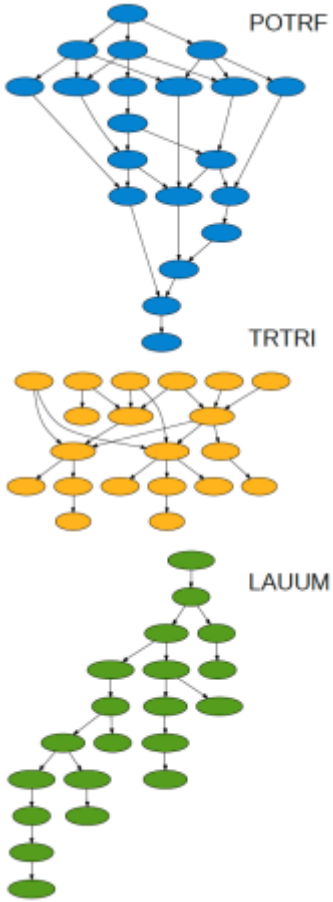
Effective CPU Frequency

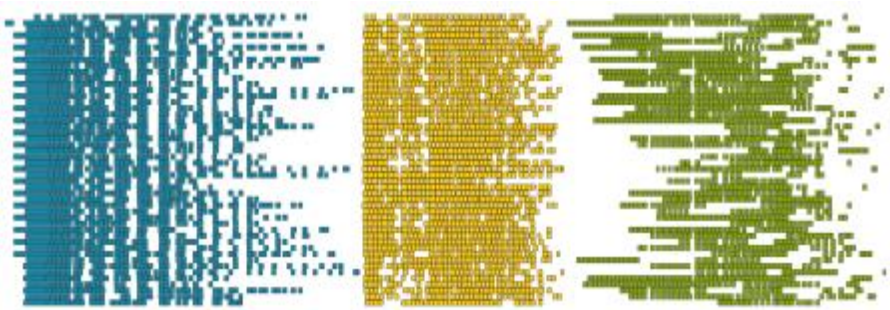
Job Power consumption

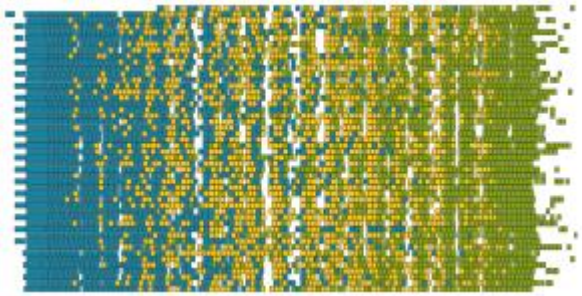


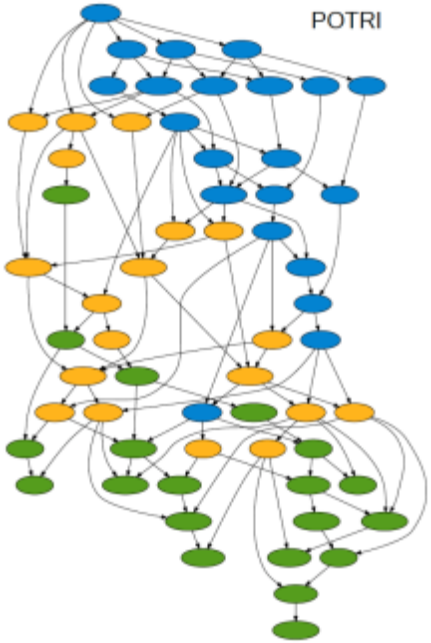


Use efficient Library







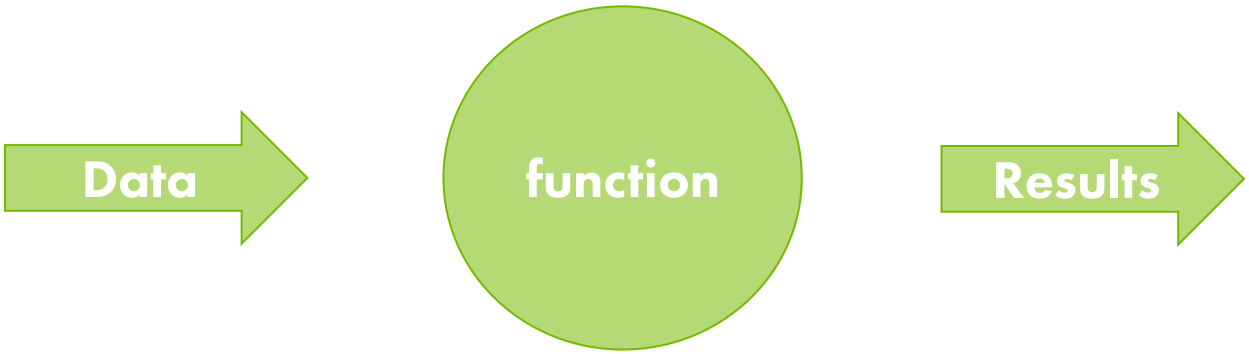


Courtesyof Jack DONGARRA – MAGAM + starPUi

Future Efficiency

2 Key elements

- [ ] Computing the function
- [ ] Managing Data



Data Movement and Floating point

- FPS-164 and VAX (1976)
  - 11 Mflop/s; transfer rate 44 MB/s
  - Ratio of flops to bytes of data movement: 1 flop per 4 bytes transferred
- Nvidia Fermi and PCI-X to host
  - 500 Gflop/s; transfer rate 8 GB/s
  - Ratio of flops to bytes of data movement: 62 flops per 1 byte transferred
- Flop/s are cheap, so are provisioned in excess

[DATA] Non Volatile RAM

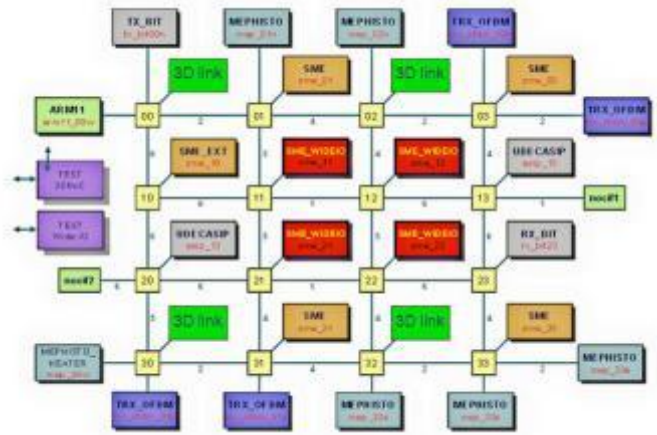
Table ERD3 Current Baseline and Prototypical Memory Technologies

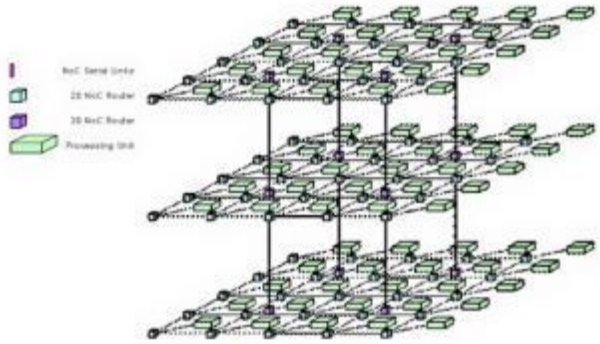
		Baseline Technologies					Prototypical technologies [A]		
		DRAM		SRAM [C]	Flash		FeRAM	STT-MRAM	PCM
		Stand-alone [A]	Embedded [C]		NOR Embedded [C]	NAND Stand-alone [A]			
Storage Mechanism		Charge on a capacitor		Inter-locked state of logic gates	Charge trapped in floating gate or in gate insulator		Remnant polarization on a ferroelectric capacitor	Magnetization of ferromagnetic layer	Reversibly changing amorphous and crystalline phases
Cell Elements		1T1C		6T	1T		1T1C	1(2)T1R	1T(D)1R
Feature size F, nm	**2011**	36	65	45	90	22	180	65	45
	2024	9	20	10	25	8	65	16	8
Cell Area	**2011**	$6F^2$	$(12\text{-}30)F^2$	140 F^2	10 F^2	4 F^2	$22F^2$	$20F^2$	$4F^2$
	2024	$4F^2$	$(12\text{-}50)F^2$	140 F^2	10 F^2	4 F^2	$12F^2$	$8F^2$	$4F^2$
Read Time	**2011**	<10 ns	2 ns	0.2 ns	15 ns	0.1ms	40 ns [G]	35 ns [J]	12 ns [K]
	2024	<10 ns	1 ns	70 ps	8 ns	0.1ms	<20 ns [H]	<10 ns	< 10 ns
W/E Time	**2011**	<10 ns	2 ns	0.2 ns	1µs/10ms	1/0.1 ms	65 ns [G]	35 ns {J]	100 ns [K]
	2024	<10 ns	1 ns	70 ps	1µs/10ms	1/0.1 ms	<10 ns[H]	<1 ns	<50 ns
Retention Time	**2011**	64 ms	4 ms	[D]	10 y	10 y	10 y	>10 y	>10 y
	2024	64 ms	1 ms	[D]	10 y	10 y	10 y	>10 y	>10 y
Write Cycles	**2011**	>1E16	>1E16	>1E16	1E5	1E4	1E14	>1E12	1E9
	2024	>1E16	>1E16	>1E16	1E5	5E3	>1E15	>1E15	1E9
Write Operating Voltage (V)	**2011**	2,5	2,5	1	10	15	1.3-3.3	1,8	3 [K]
	2024	1,5	1,5	0,7	9	15	0.7–1.5	<1	<3
Read Operating Voltage (V)	**2011**	1,8	1,7	1	1,8	1,8	1.3–3.3	1,8	1,2
	2024	1,5	1,5	0,7	1	1	0.7–1.5	<1	<1
Write Energy (J/bit)	**2011**	4E-15 [B]	5,00E-15	5,00E-16	1E-10 [E]	>2E-16 [F]	3E-14 [I]	2.5E-12 [A]	6E-12 [L]
	2024	2E-15 [B]	2,00E-15	3,00E-17	1E-11 [E]	>2E-17 [F]	7E-15 [I]	1.5E-13 [A]	~1E-15 [M]

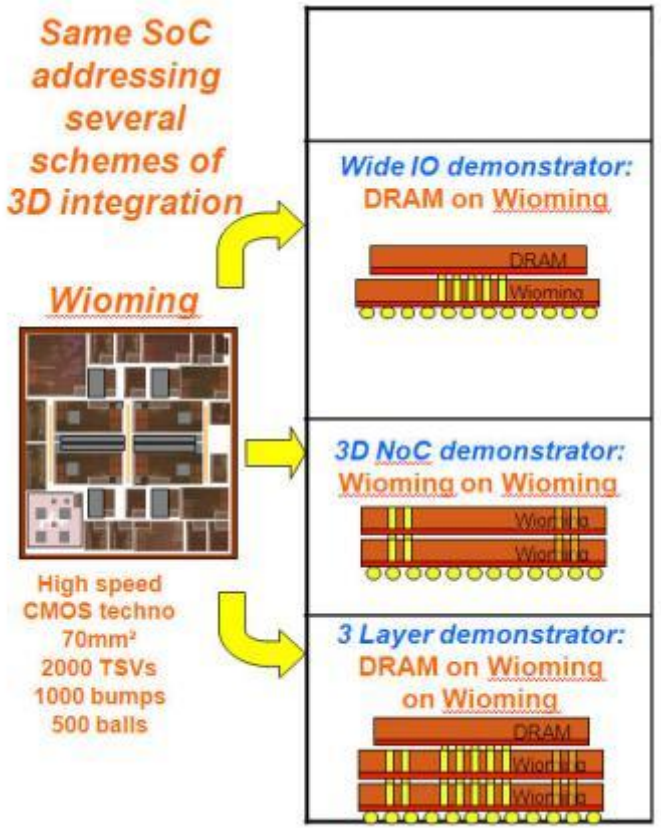
Source: ITRS

[DATA] 3D Wioming from CEA leti

Wide IO Memory Interface Next Generation



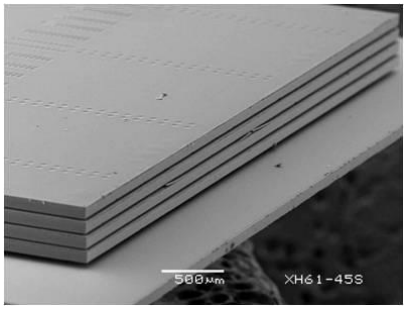




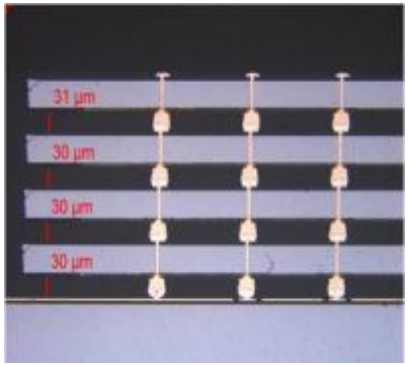
Courtesy of CEA leti

[DATA] Hybrid Memory Cube

- Wide I/O standard comes from mobile world
- True « 3D » (with Through Silicon Vias »)



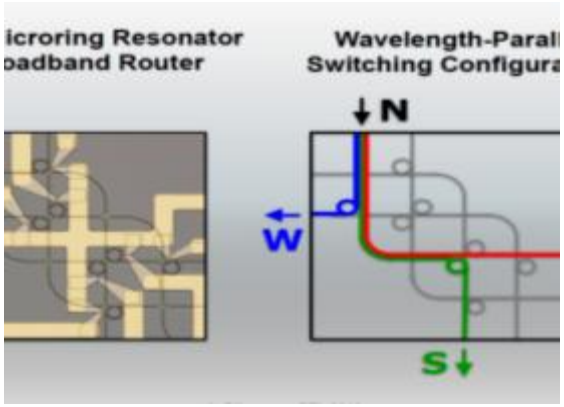
	Bandwidth	Energy
DDR-3	10,66 GB/s	50-70pJ/bit
HMC	128GB/s	8 pJ/bit

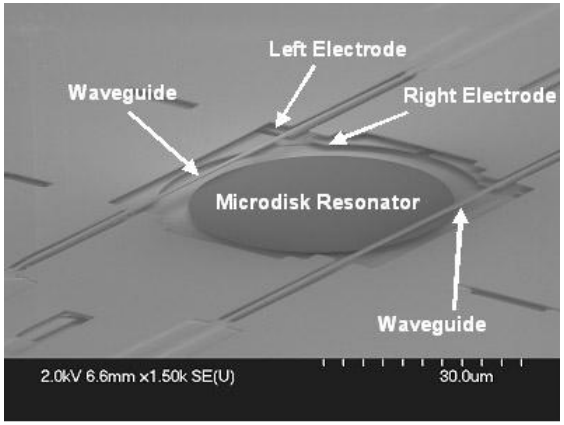


Source: http://www.synopsys.com,
http://blogs.intel.com/research/2011/09/15/hmc/,
http://www.semiconwest.org/sites/semiconwest.org/files/docs/Shekhar%20Borkar_Intel.pdf

[DATA] Photonics

- Photonic is getting closer to the CPU
- Basic blocks are in place:
  - Modulator
  - Multiplexer
  - Routing
  - Receiver

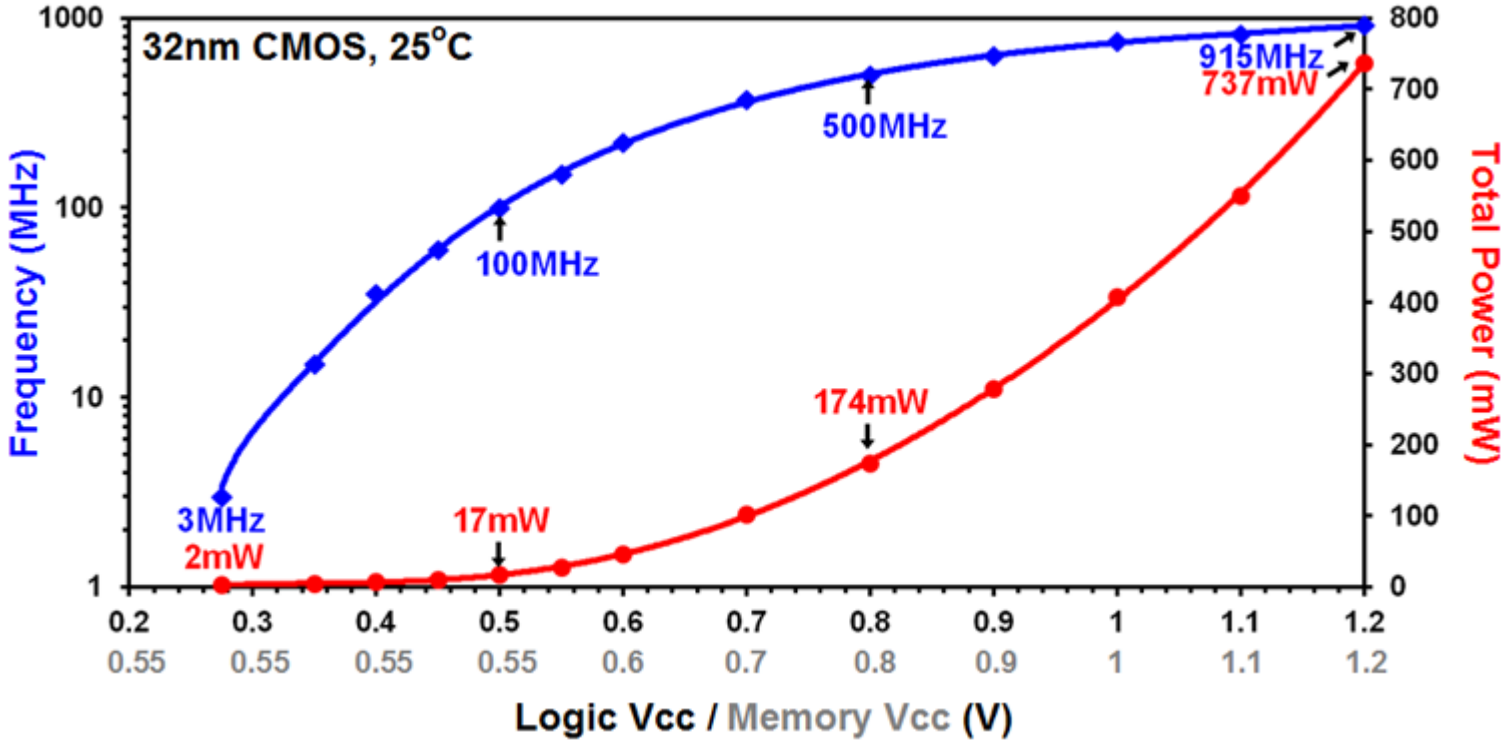




- Off chip Energy cost very close to On chip Energy cost

To go deeper: "Silicon Photonics for Exascale Computing" Keren Bergman, Columbia universisty Lightwave Research Laboratory

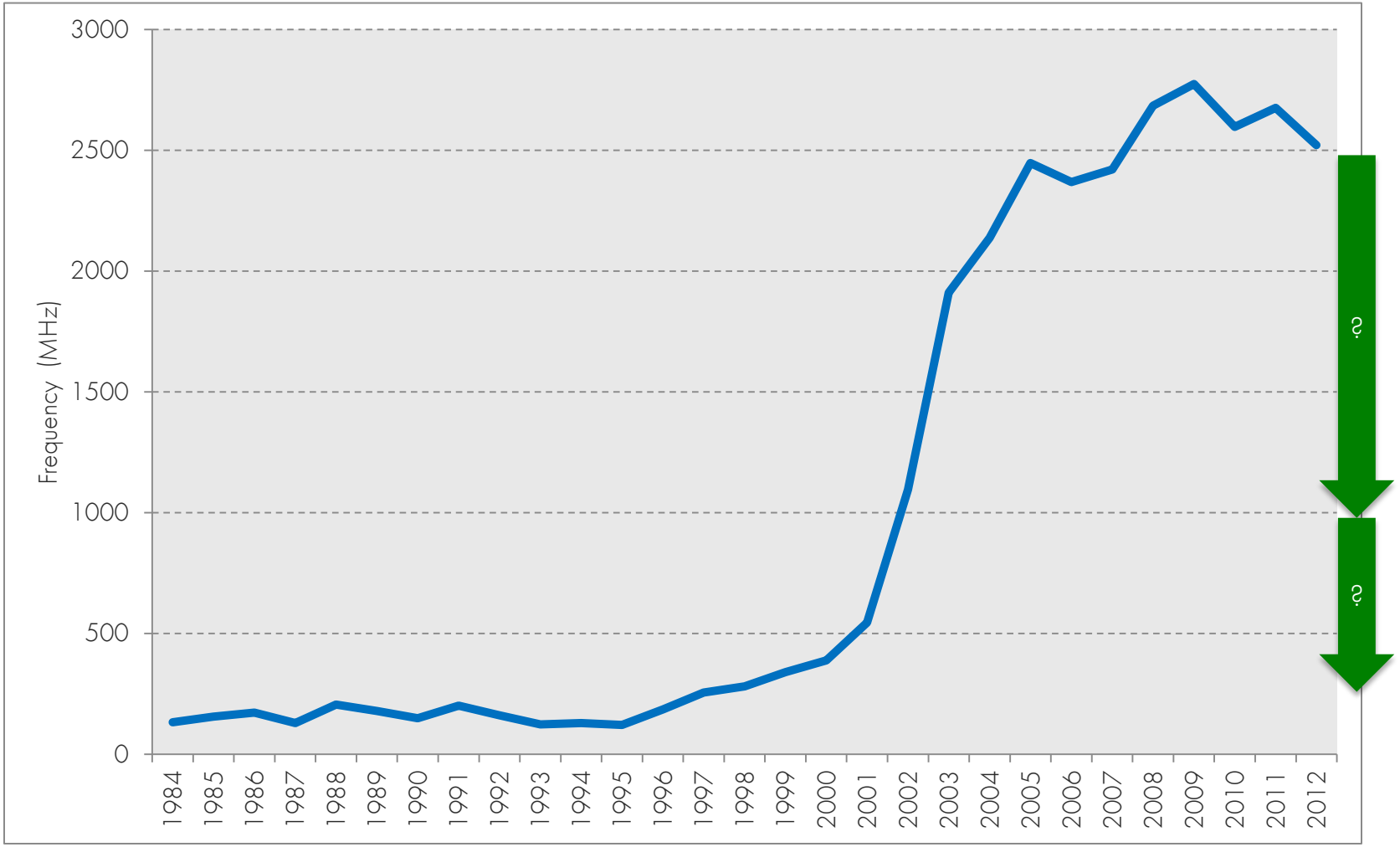
[FUNCTION] Near threshold Voltage



- Drawback: surface is increasing

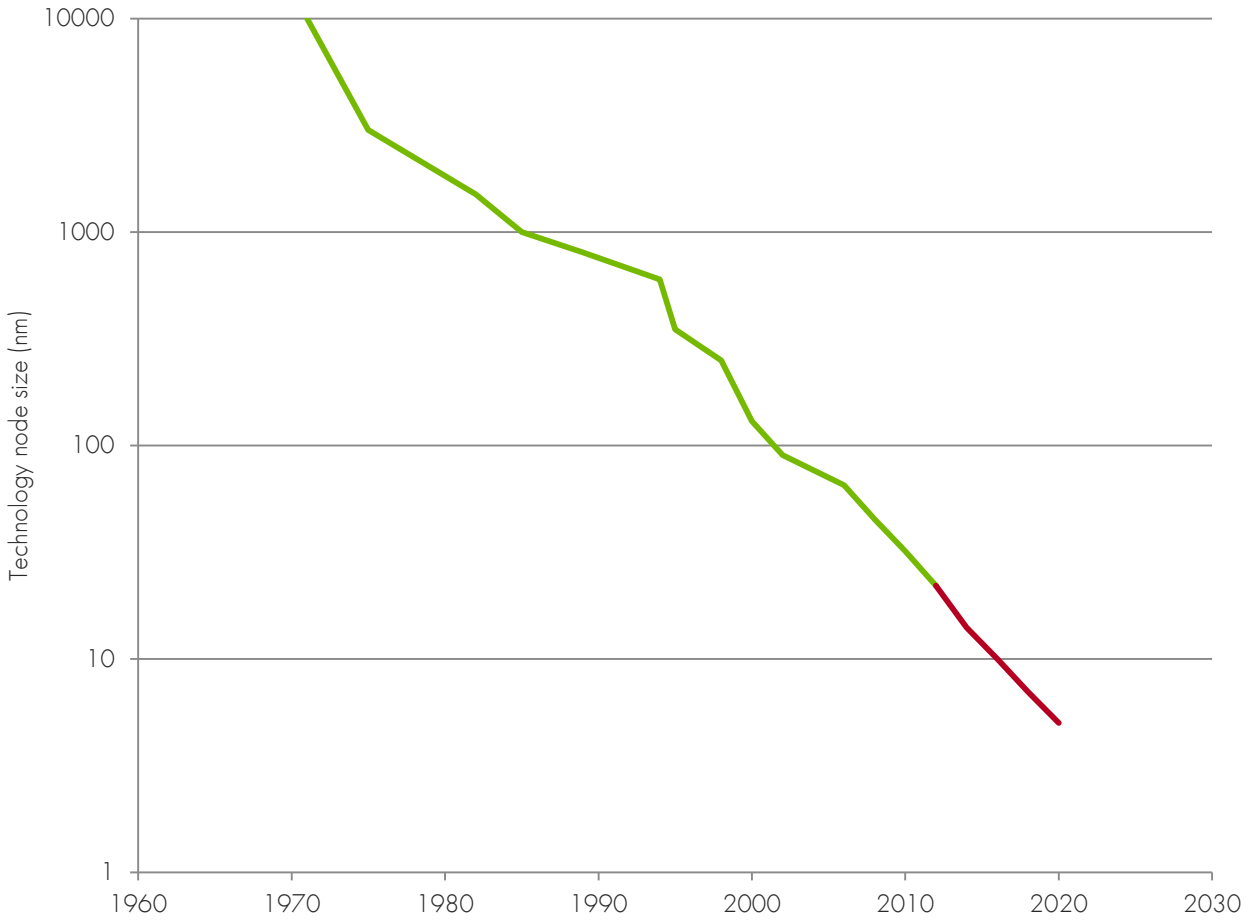
http://www.realworldtech.com/near-threshold-voltage/

[FUNCTION] CPU frequency



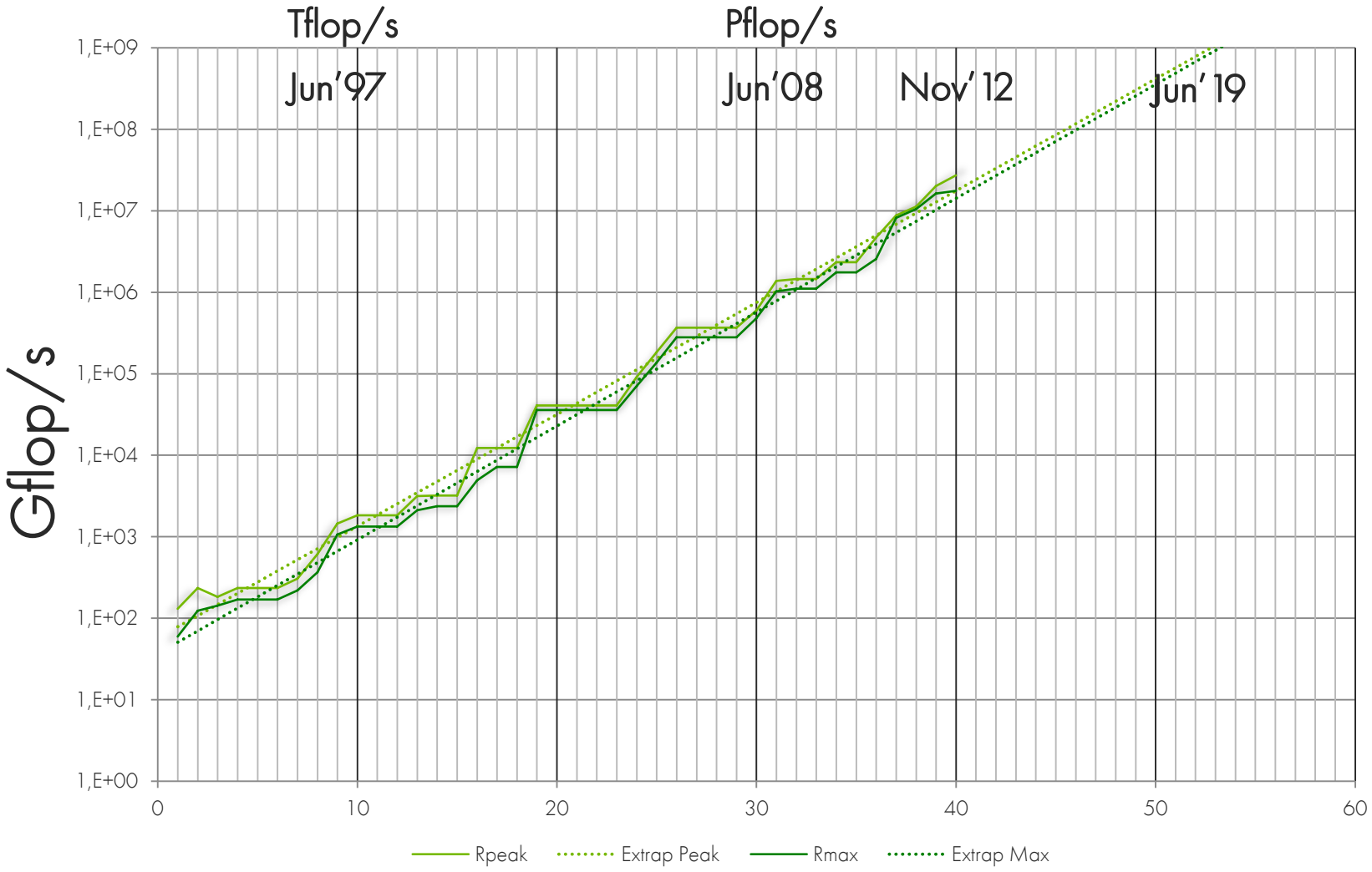
Source: top500 - mean freq

[FUNCTION] Mask reduction



The exascale

TOP500

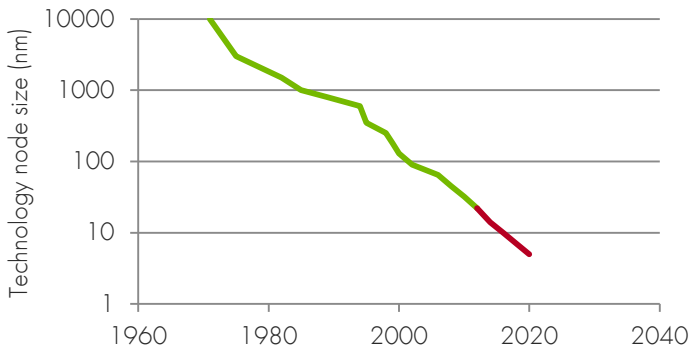


Is it a long way to Exascale

	Jun '97		Nov ' 08		Nov '12		Jun ' 19
	asci red		roadrunner		Titan		Eflop
	Sandia		ORNL		ORNL		???
MW	0,50		2,48		8,3		20
Eflop max	1,45E-06		0,001		0,027		1
nodes	7264		6480		18688		64000
racks	104		296		200		300
U/rack	30		42		42		42
Concur.	9152		129'600		50 M		O(10^9)
		Year factor		Year factor		Year factor	
Gflop/W	2,91E-03	1,549	0,4	1,651	3,3	1,519	50,0
Tflop/node	2,00E-04	1,798	0,2	1,708	1,5	1,441	15,6
W/node	68,8	1,161	383,3	1,035	439,3	0,949	312,5
nodes/rack	69,8	0,904	21,9	1,437	93,4	1,135	213,3
W/rack	4807,7	1,050	8390,1	1,487	41045,0	1,077	66666,7
Tflop/U	4,66E-04	1,579	0,09	2,455	3,23	1,637	79,37
W/U	160,3	1,019	199,8	1,487	977,3	1,077	1587,3

What's a « node » ?

Node evolution : « à la Phi »



Gflop/W	50,0	
Tflop/node	15,6	1-2 Phi
W/node	312,5	
nodes/rack	213,3	
Tflop/U	79,37	10 Phi
W/U	1587,3	

- 60 cores @22nm => ~500 cores @8nm
- 1 Tflops @22nm => ~8 Tflops @ 8nm

- With µarch improvement, it should be less !

- Phi power consumption will have to be around 150W-300W
- Between 200 and 400 Phi per Rack

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

1050 MHz – 700 MHz

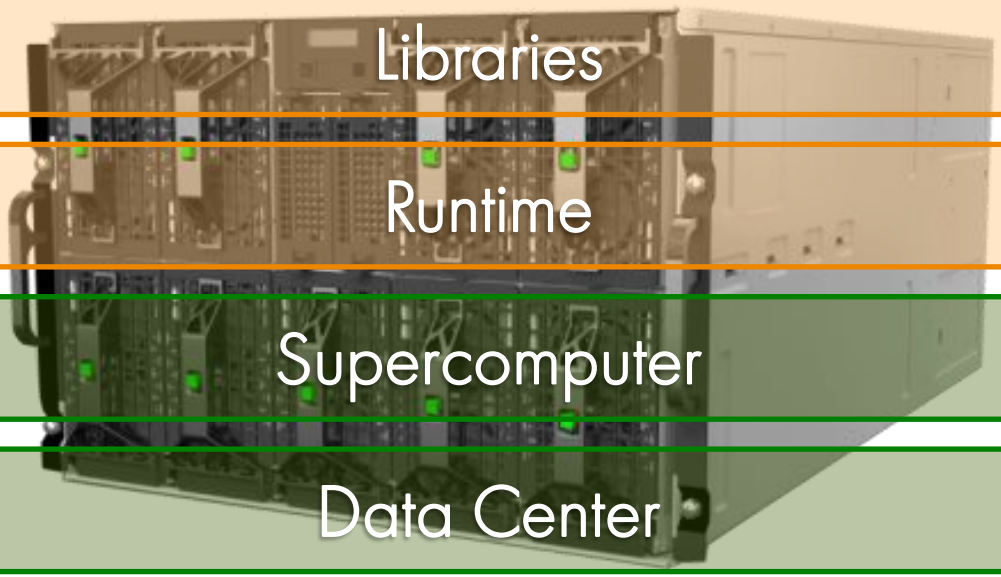
16 - 64

22’000’000

Gene Amdahl

John Gustafson

Stack



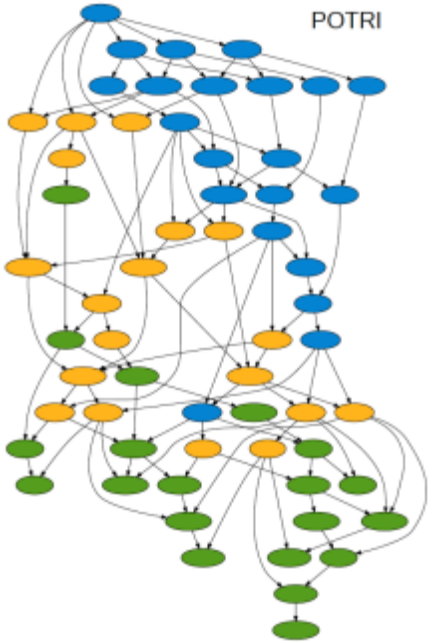
Hurry UP

Speed UP

Keep the Pace

Strong Integration

- Ressource Manager
- Data exchange
- Library
- Code
- Resiliency
- Topology
- ...



- See Fastforward approach
- See MAGMA [UTK] with StarPU [INRIA]

Supercomputer Suite

COMPLETE

All functionalities
- Cluster management
- Development factory
- Execution environment
- Data storage and access

All sizes
- From department
- to Top 10

OPEN

Best of breed
- Linux,
- OpenMPI,
- HPC Toolkit,
- Nagios,
- OFED,
- Slurm,
- Lustre,
- Shine, ...

+ Bull added value

FLEXIBLE

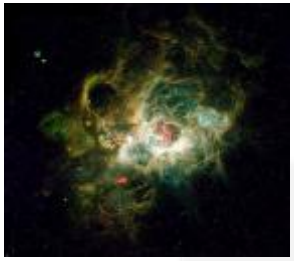
Modular:
Get what you need
when you need

INTEGRATED

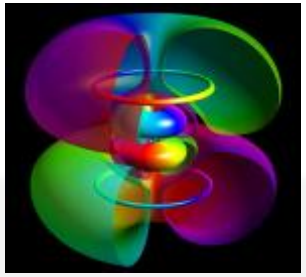
Installed, deployed
and operated as a
single software

Applications and Performances team

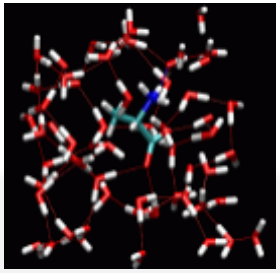
Electro-Magnetics



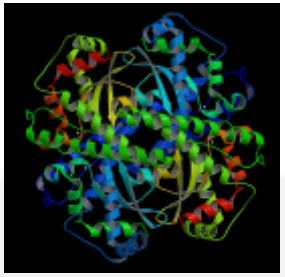
Computational Chemistry
Quantum Mechanics



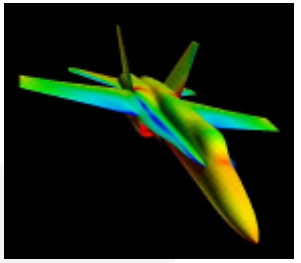
Computational Chemistry
Molecular Dynamics



Computational Biology



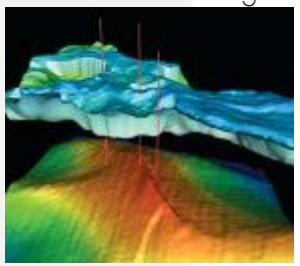
Structural Mechanics
Implicit



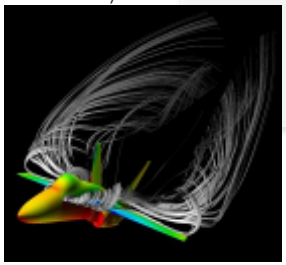
Structural Mechanics
Explicit

One Bull expert (PhD)
for each segment

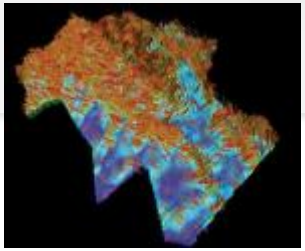
Seismic Processing



Computational Fluid
Dynamics



Reservoir Simulation



Rendering / Ray Tracing

Climate / Weather
Ocean Simulation

Data Analytics

METEO FRANCE

Toujours un temps d'avance

première étape (2013-2014) une puissance de calcul d'environ 1 Petaflops[*], puis à l'horizon 2016, une performance totale dépassant 5 Petaflops.

Cette augmentation des moyens de calcul de Météo-France se traduira par une évolution technologique importante : le passage de la technologie vectorielle à la technologie scalaire qui repose sur les standards du marché et permet de fournir une puissance de calcul parallèle nettement supérieure pour un moindre coût total de possession.
Le choix de Météo-France souligne le savoir-faire développé par Bull en matière de parallélisation des codes applicatifs utilisés dans les domaines de la météorologie et des sciences du climat. Une plus grande parallélisation est essentielle pour l'utilisation optimale des machines. Elle entraîne une évolution indispensable des codes de calcul qui représente en elle-même un grand challenge, auquel sont confrontés tous les instituts météorologiques dans le monde.

« Bull est fier du choix de Météo-France pour nos plus récents et nos plus puissants supercalculateurs bullx. » déclare Philippe Vannier, Président-directeur général de Bull. *« Le choix de Météo-France*

Eco Impact in HPC
Where is it worth to invest ?

Globally !

Applications are already late!

Architect of an Open World™

For any question
xavier.vigouroux@bull.net