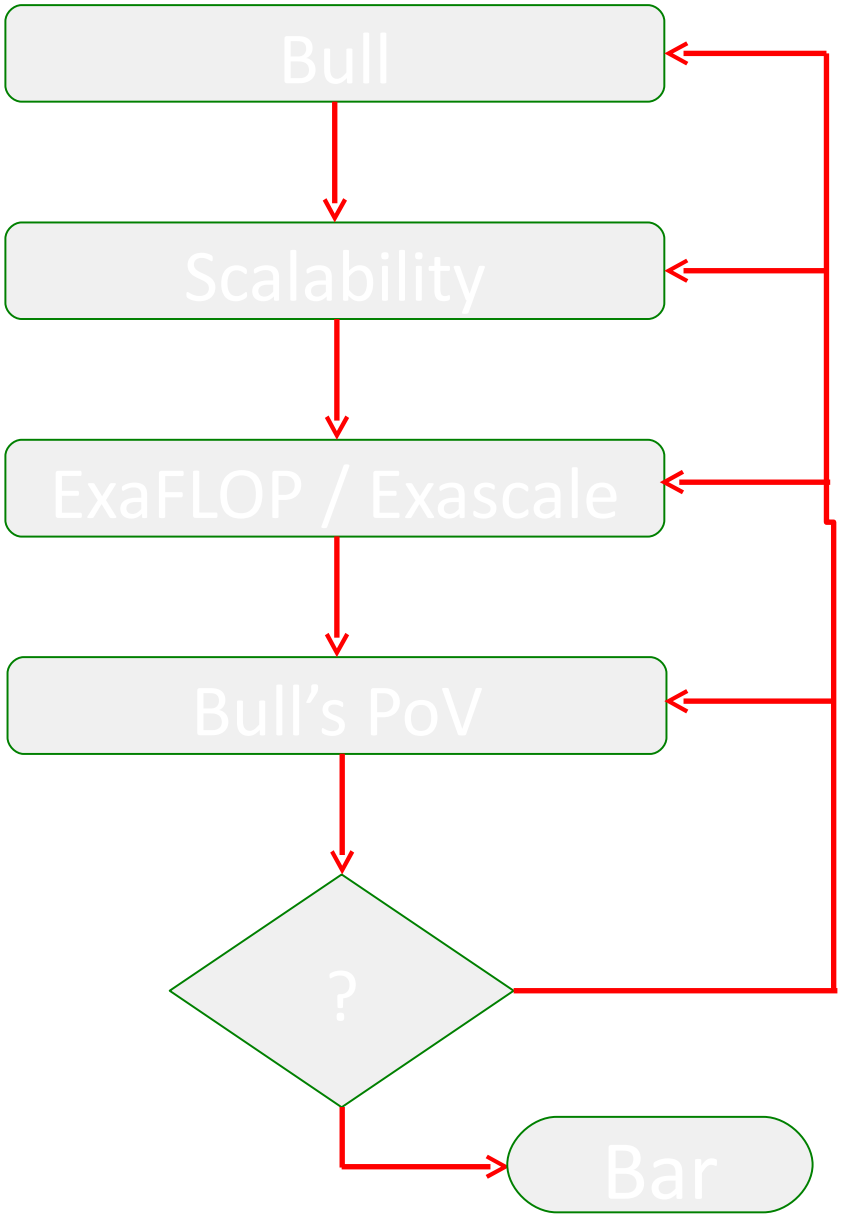
Architect of an Open World™

Defying the Laws of Physics in/with HPC

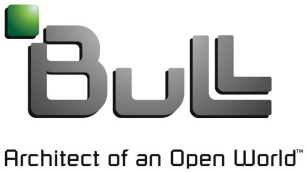
2013 – 11 – 12

Rafa Grimán
HPC Architect

Agenda



Bull



Mastering Value Chain for Critical Processes

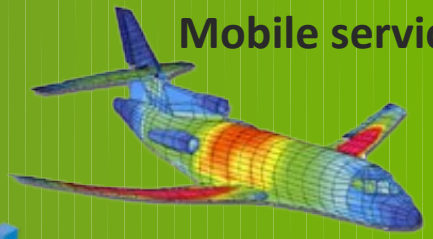
From infrastructures to business applications: design, build and operate powerful and trusted solutions

Cyberdefense

e-Government

e-Banking

Mobile services



Smart Grid

Smart transportation

Computer manufacturer | *Integrator / Outsourcer* | *Service provider* | *Solution provider*

Security provider and operator

TOP500 SUPERCOMPUTER SITES

CODE OF CONDUCT DATA CENTRES

la Région Languedoc Roussillon

SAP Quality Award

SC MAGAZINE AWARDS 2013 EUROPE

- Manage explosion in processing needs **INNOVATE**
- Support shift towards Cloud **OPTIMIZE**
- Drive digital transformation **INTEGRATE**
- Guarantee trust **SECURE**

Bull: from Supercomputers to Cloud Computing

Expertise & services

- HPC Systems Architecture
- Applications & Performance
- Energy Efficiency
- Data Management
- HPC Cloud

extreme factory
stay lean: compute smart

center for excellence in parallel programming

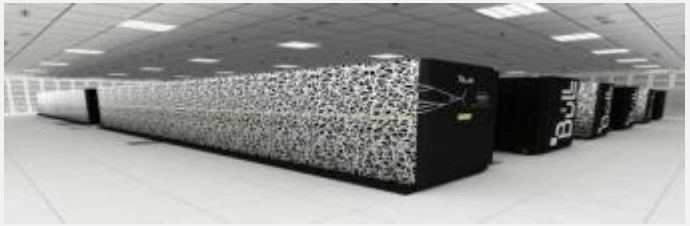
Software

- Open, scalable, reliable SW
- Development Environment
- Linux, OpenMPI, Lustre, Slurm
- Administration & monitoring

bullx supercomputer suite

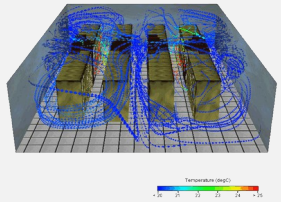
Servers

- Full range development from ASICs to boards, blades, racks
- Support for accelerators



Infrastructure

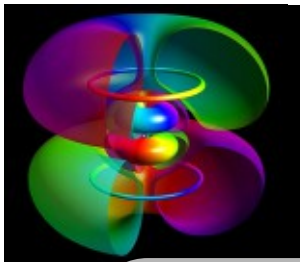
- Data Center design
- Mobile Data Center
- Water-Cooling



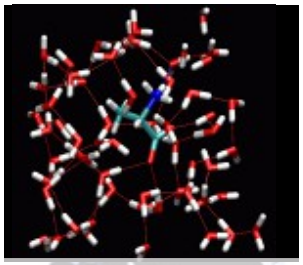
Extreme Computing applications

Electro-Magnetics

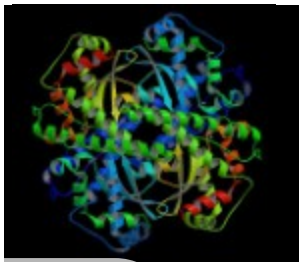
Computational Chemistry
Quantum Mechanics



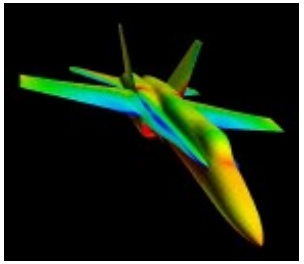
Computational Chemistry
Molecular Dynamics



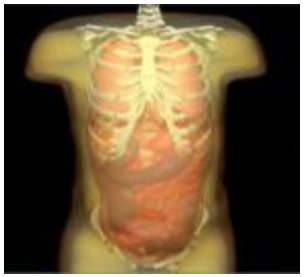
Computational Biology



Structural Mechanics
Implicit

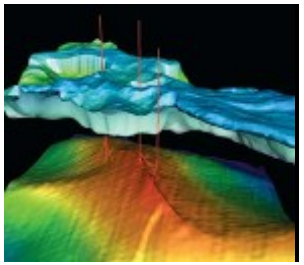


Life science

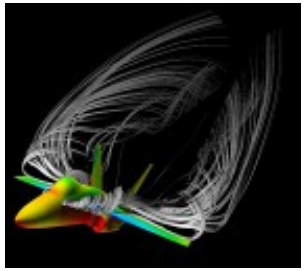


A dedicated team of experts in application performance

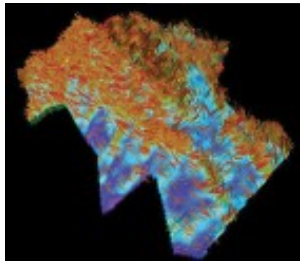
Seismic Processing



Computational Fluid
Dynamics



Reservoir Simulation



Rendering / Ray Tracing

Climate / Weather
Ocean Simulation

Data Analytics

Key R&D cooperation projects

Energy-efficient HPC solution projects
- Mont Blanc, SHARP

Cloud Computing projects for HPC
- XLcloud, EASI CLOUDS, PerfCloud

HPC technologies for Big Data
- TIMCO

Software stack and tools
- H4H, Newcastle, EnergeTIC

Co-design for application sectors
- H4H, Pulsation

Cooperation with customers

Leading HPC technology with Bull

TERA100 – 2010

1st European PetaFlop-scale System

1.25 PFLOPs

CURIE – 2011

1st PRACE PetaFlop-scale System

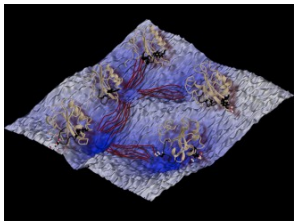
2 PFLOPs

BEAUFIX – 2013

1st Intel Xeon E5-2600 v2 System

Direct Liquid Cooling Technology

More FLOPs ;)



HELIOS – 2012

3rd PETAFLOP system

1.5 PFLOPs

Aimed at successfully controlling nuclear fusion and harnessing it as a future source of energy

SARA – 2013

4th PETAFLOP system

1.3 PFLOPs

U. Dresden – 2013

750 TFLOPs

BSC Minotauro

BSC. Minotauro

BSC is the biggest supercomputing center in Spain and a Tier-0 in PRACE.
BSC supports Spanish scientists in alll reasearch areas

MinoTauro is the most powerful system in Spain

- #114 in Top500.org
- 186 TFLOPs peak performance
- Efficiency of 1,26 TFLOPs/KW:
  - #1 in Europe
  - #7 in the World
- Among the best in GFLOPs/m2 with 25,83 TFLOPs/m2
- Liquid cooling
- 128 blades: 256 M2090 & 128 SSDs
- Agreement BSC-BULL-NVIDIA
- Doubles the computation of

They have put their trust in Bull

Bull User group for eXtreme computing

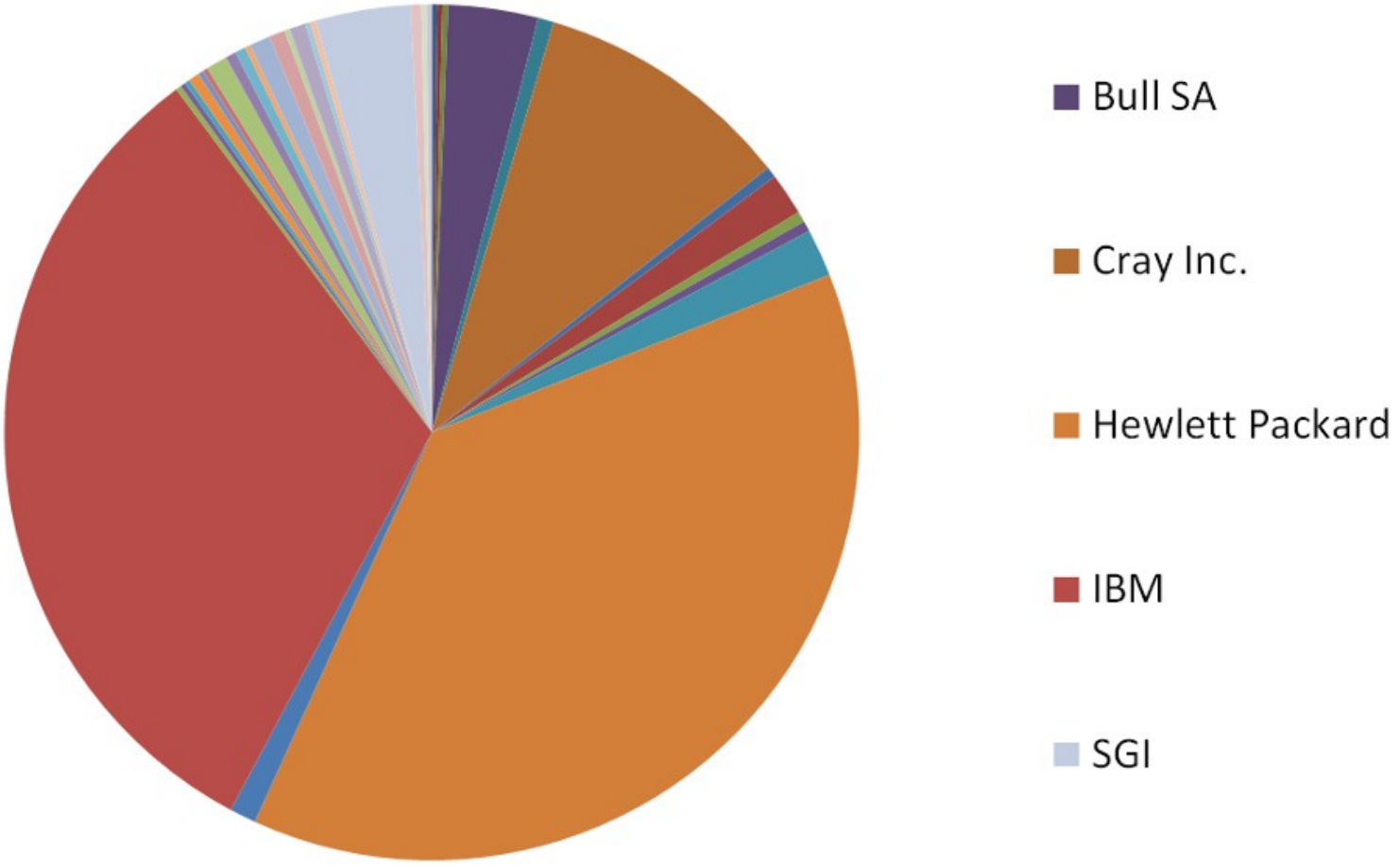
An independent worldwide group of users to:

- Share experience between members and with Bull
- Give Bull direct feedback and input
- Network with Bull HPC experts

Affiliates

Join us at www.bux-org.com

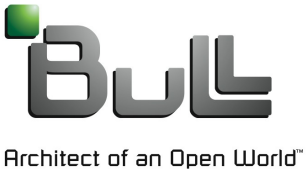
#5 in the World



80.24 % Efficiency

Source: top500 – june 2013 – # of systems

Expect the Unexpected



What is scalability

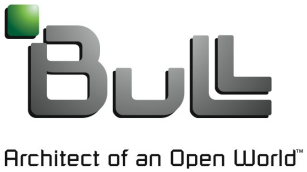
We often see scalability referred to as core count

Scalability should take into account other parameters:

- RAM
- Storage
- I/O
- Software

Having a balanced system is very important

ExaFLOP / Exa-Scale

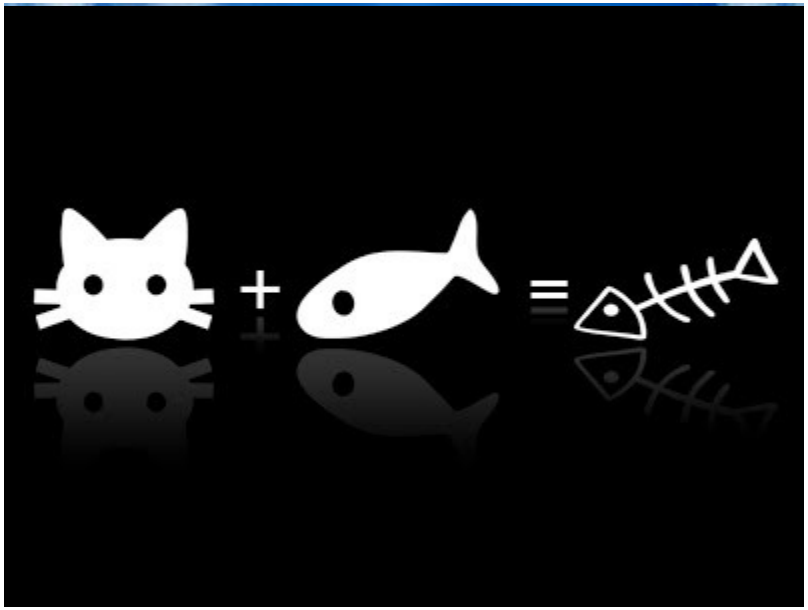


What is HPC?

HPC:

The use of super computers and parallel processing techniques for solving complex computational problems.

A branch of computer science that concentrates on developing supercomputers and software to run on supercomputers.

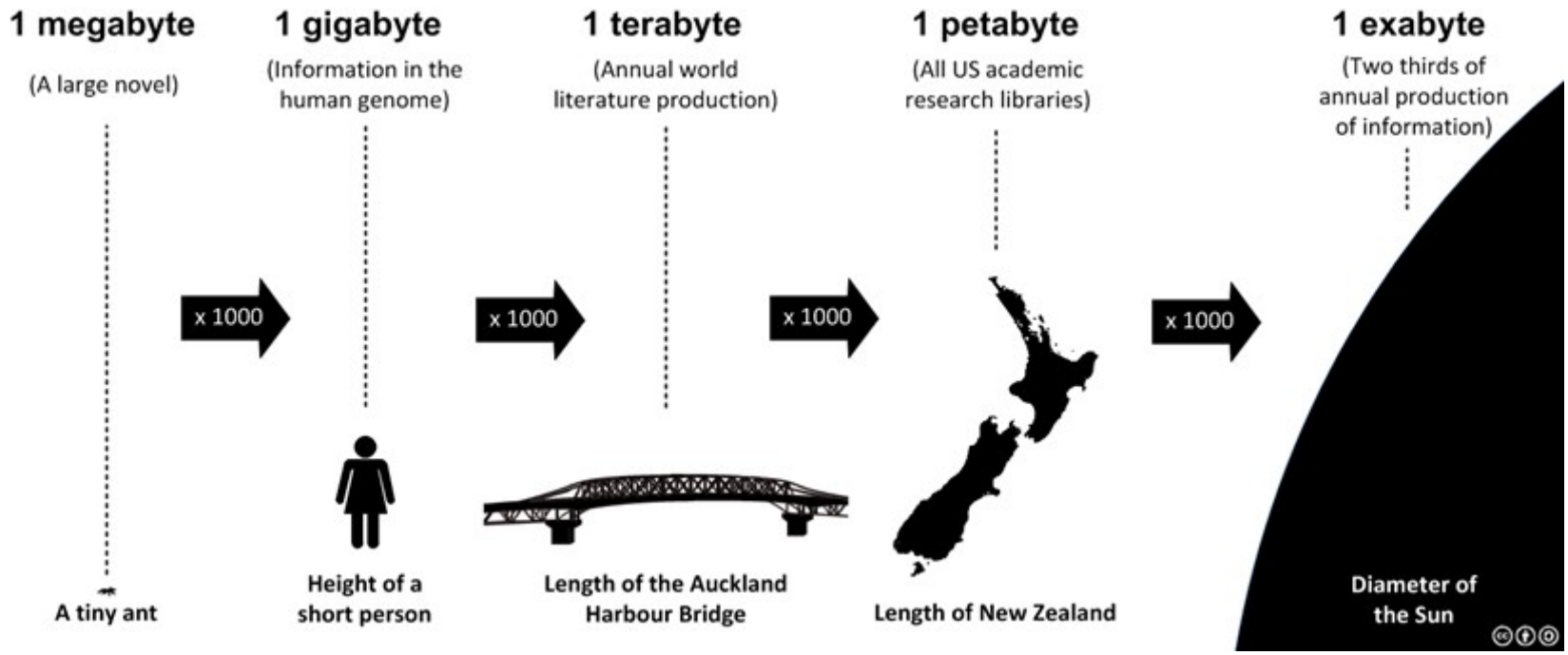


Supercomputer:

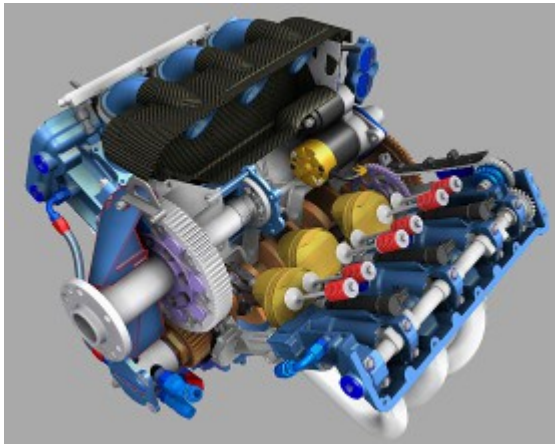
Supercomputers are very expensive and are employed for specialized applications that require immense amounts of mathematical calculations.

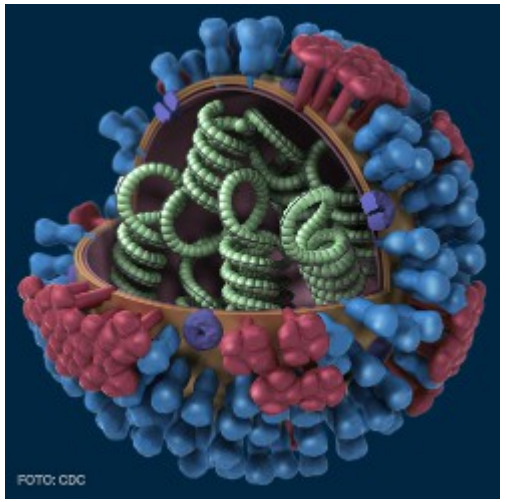
What is an Exa-FLOP?

10^18 FLOPs

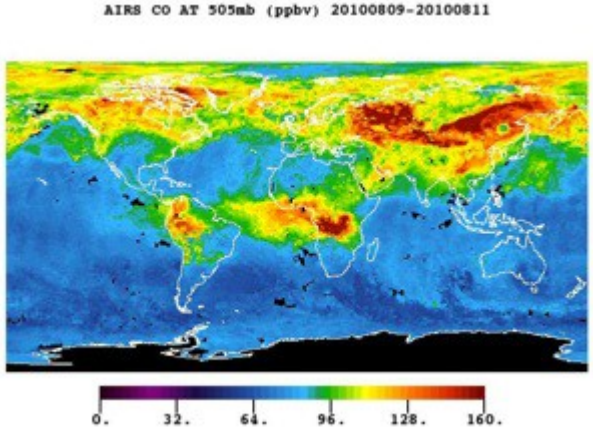


Why?

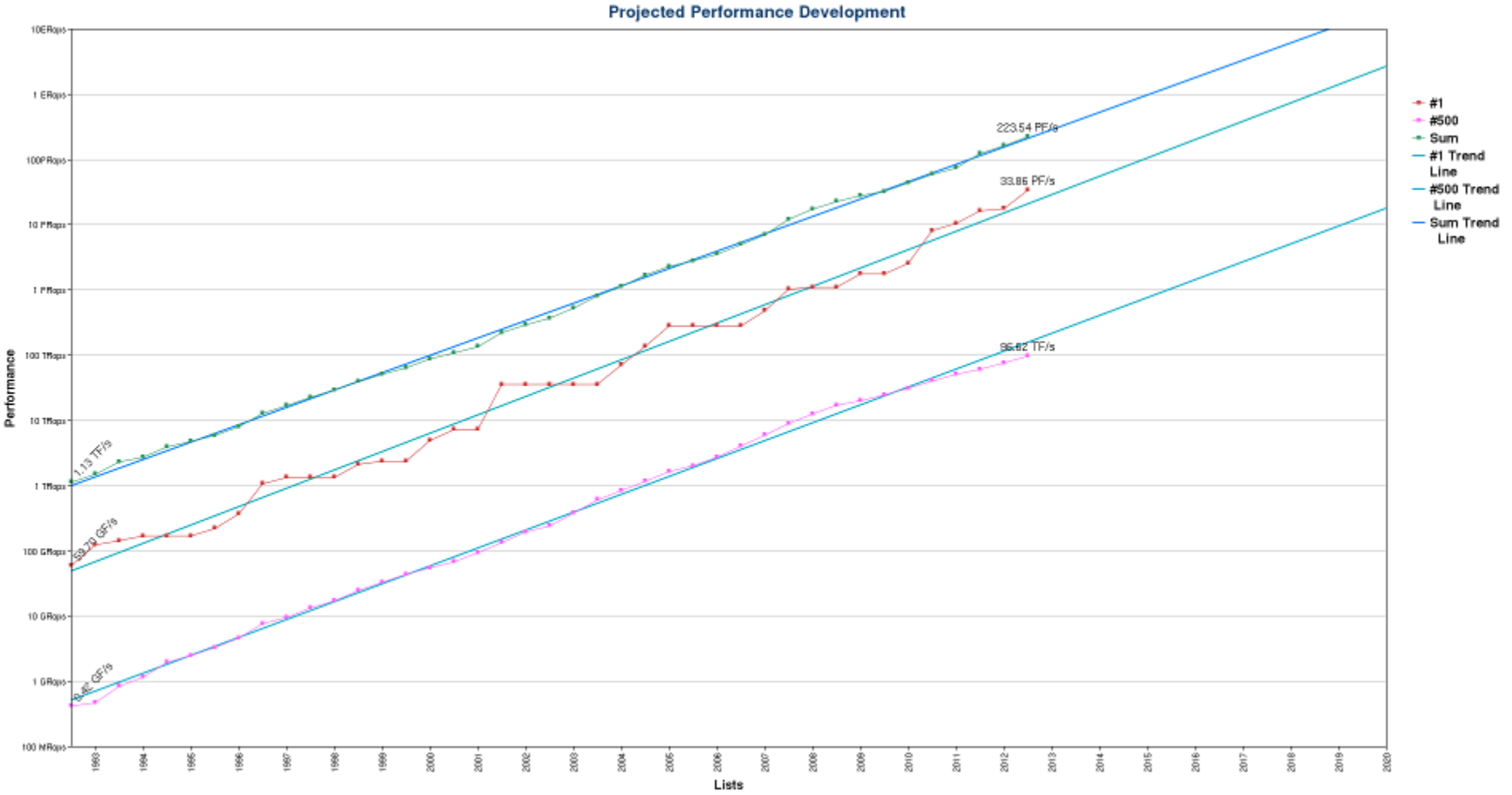




Blue Brain Project



When?



Needs for Exascale

x 1000 user requirements: **Extreme User** ;)

x 1000 memory capacity: **Extreme Memory**

x 1000 communications bandwith: **Extreme Networking**

x 1 / 1000 Latency: **Extreme Latency**

x 1000 [illegible]

x 1000 [illegible]

x 1000 [illegible]

x 1000 in scalability: **Extreme Application Scalability**

x 1000 [illegible]

x 1000 in visualization: **Extreme Visualization**

x 1000 in system management: **Extreme Management**

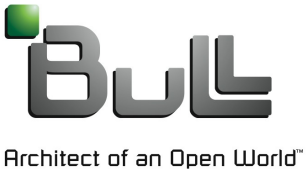
… but…

x <1 in cost. **Extreme Cost Savings** (even more so during a crisis)

x <1 electricity consumption: **Extreme Power Consumption**

All targeted for 2020!
(USA: DARPA / DoD / DoE projects)
(Europe: Mont-Blanc, DEEP, Cresta)

Bull's PoV



Extreme Reliability and Management

Tens/hundreds of thousands of nodes

- Millions of cores & DIMMS
- Tens of thousands of disks
- Kilometers of cables.

All working without disruption

Checkpoint/restart SW for millions of threads

Power consumption management

Incident reporting

Alarm management

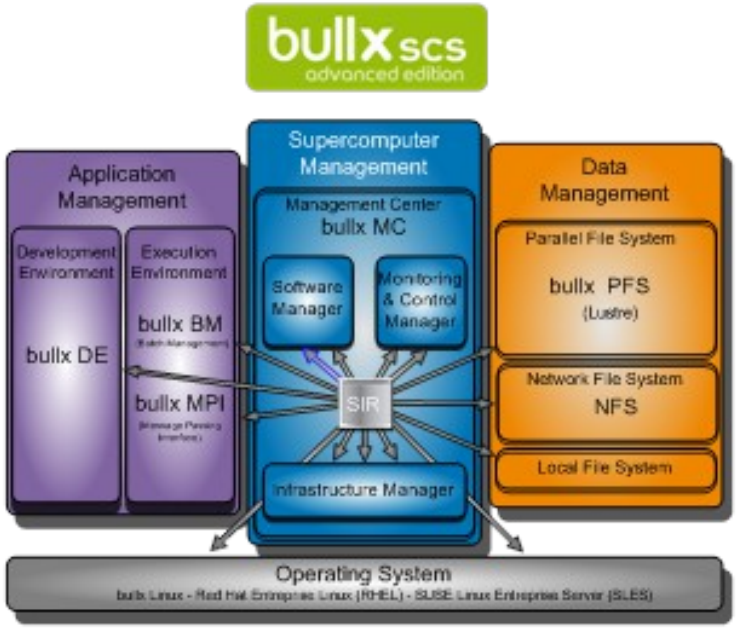
bscs4 – Management

bscs 4 : bullx supercomputer suite 4

bscs4 – Management

Complete HPC software suite

- Addresses supercomputer lifecycle needs
- Modular, flexible and extensible
- Installation / Configuration
- Monitoring: health & power management
- Distributed management: 1000s nodes



Error management

- Based on SEC: Simple Event Correlator
- Error detection and correlation

ARGOS:

- Incident processing, management & follow-ups
- Availability measurements
- GUI & CLI & multiuser environment

Extreme Networking and Topologies

bscs4 – Networking

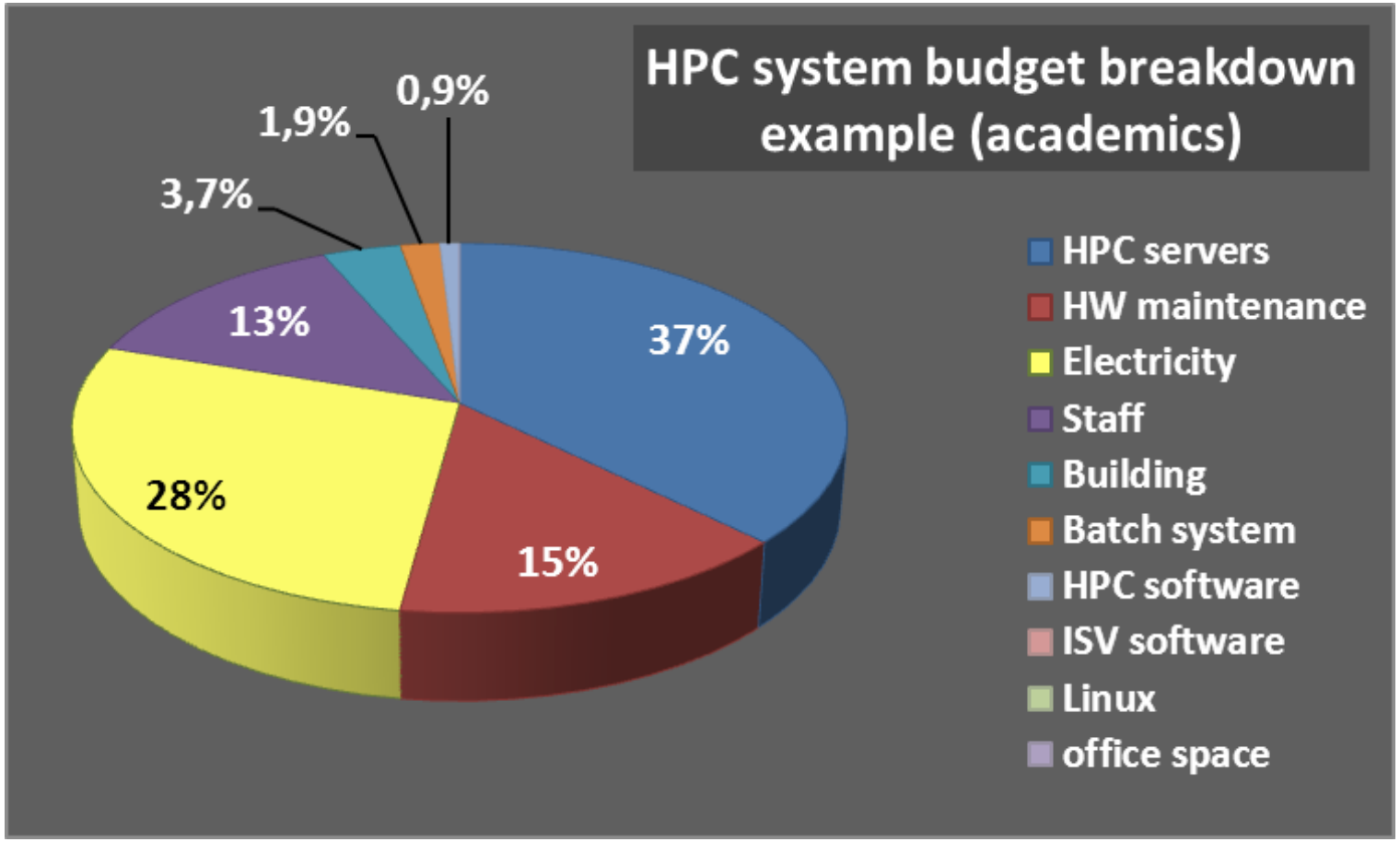
Design correct routing algorithms:

- Use all primary routes equally
- Always use shortest route
- Avoid additional switches in HA

Diagnostic tools that display:

- Network topology
  - Faulty/sub-optimal links in red
- Bandwidths
- Interconnect availability
  - Between switches
  - Inside switch
- Route(s) used by the subnet manager
- Basic IB information: LID, Location

Extreme Computing Power Consumption



ExaFLOP system

Exaflop Today

500+MW?

	40MW	Other misc. power consumption: … Power supply losses Cooling … etc
Storage	**10MW**	10EB disk @ 10TB/disk @10W
Comms	**100MW**	100pJ com per FLOP
Memory	**150MW**	0.1B/FLOP @ 1.5nJ per Byte
Compute	**200MW**	~400W / Socket

Exaflop Target

<20MW

	3MW	Other misc. power consumption: … Power supply losses Cooling … etc
Storage	**~2MW**	10EB SSD @ 5TB/SSD
Comms	**~1MW**	9pJ per FLOP
Memory	**2MW**	0.1B/FLOP @ 150pJ per Byte
Compute	**12MW**	50K Sockets @~200W each

Source: Intel – Intel User Forum, April 2009

Extreme Computing: Processors's Future

Systems with std processors: thin nodes & SMP nodes

- 2014 single node with 2 sockets achieving 1 TFLOP
- 2020:
  - hundreds of cores per socket
  - node with many sockets and thousands of cores
- Low power processors (ARM, Intel, etc)

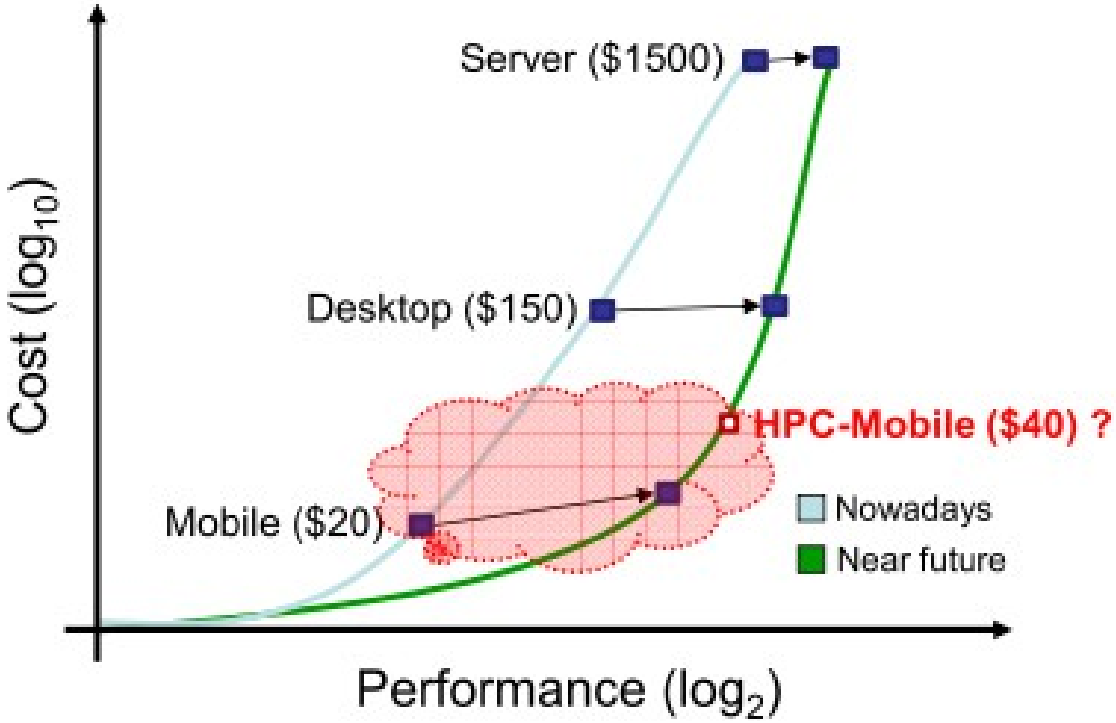
Hybrid systems based on accelerators (GPU, Intel MIC)

- NVIDIA K20x: 1.31 TFLOPs
- Intel MIC SE10P: 1.073 TFLOPs
- FPGA?
- Vector capabilities again?

Xeon vs Atom/ARM

Swap out Xeons for embedded processors?

- Reduced price
- Low wattage and low heat production
- Support for double precision
- Support for Linux



Mont Blanc

Develop future European Exascale systems

Based on power-efficient technology

4 Tier-0 hosting partners in PRACE

Financed under the Objetivo FP7 ICT-2011.9.13 Exa-scale computing, software and simulation:

- 3 year project (Octobre 2011 – September 2014)
- Total budget: 14.5 M€ (8.1 M€ EC)

MontBlanc Targets

Target 1: Develop a prototype based on energy-efficient embedded processors

- Scalable to 50 PFLOPS at 7 MWatt
- Competitive with leaders on the Green500 list in 2014
- Develop a complete HPC SW system

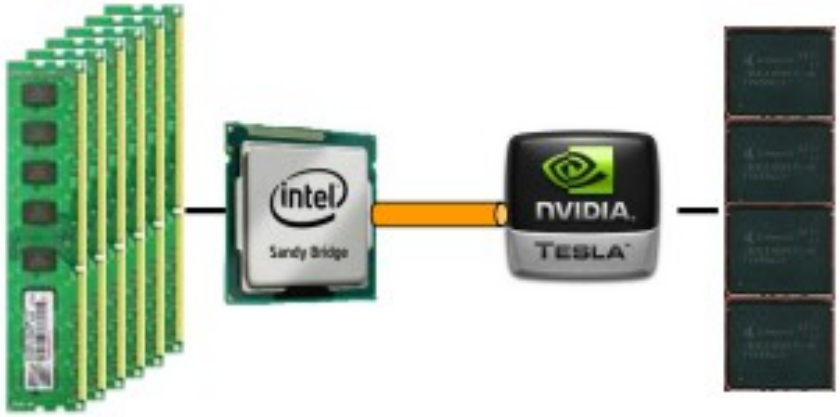
Target 2: Design the next generation HPC systems and embedded technology solving major limitations found in the prototype

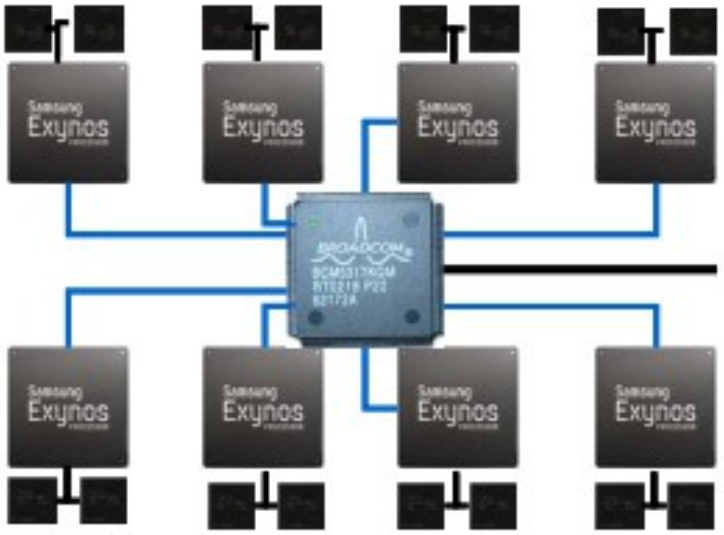
- Scalable to 200 PFLOPS at 10 MWatt
- Competitive with the Top500 leaders in 2017
- Scalable to 1 EFLOPS at 20 MWatt
- Compete with Top500 leaders in 2020

Target 3: Port and optimize representative HPC

- 11 applications

MontBlanc: Intel vs ARM





Sandy Bridge + NVIDIA K20

- 1600 GFLOPS
- 2 address spaces
- 32 GB/s CPU-GPU
- 68 + 192 GB/s
- > $3000
- > 400 Watt

8-socket Exynos 5450

- 1600 GFLOPS
- 16 address spaces
- 12.8 GB/s CPU-GPU
- 102 GB/s
- < $200
- < 100 Watt

Direct Liquid Cooling

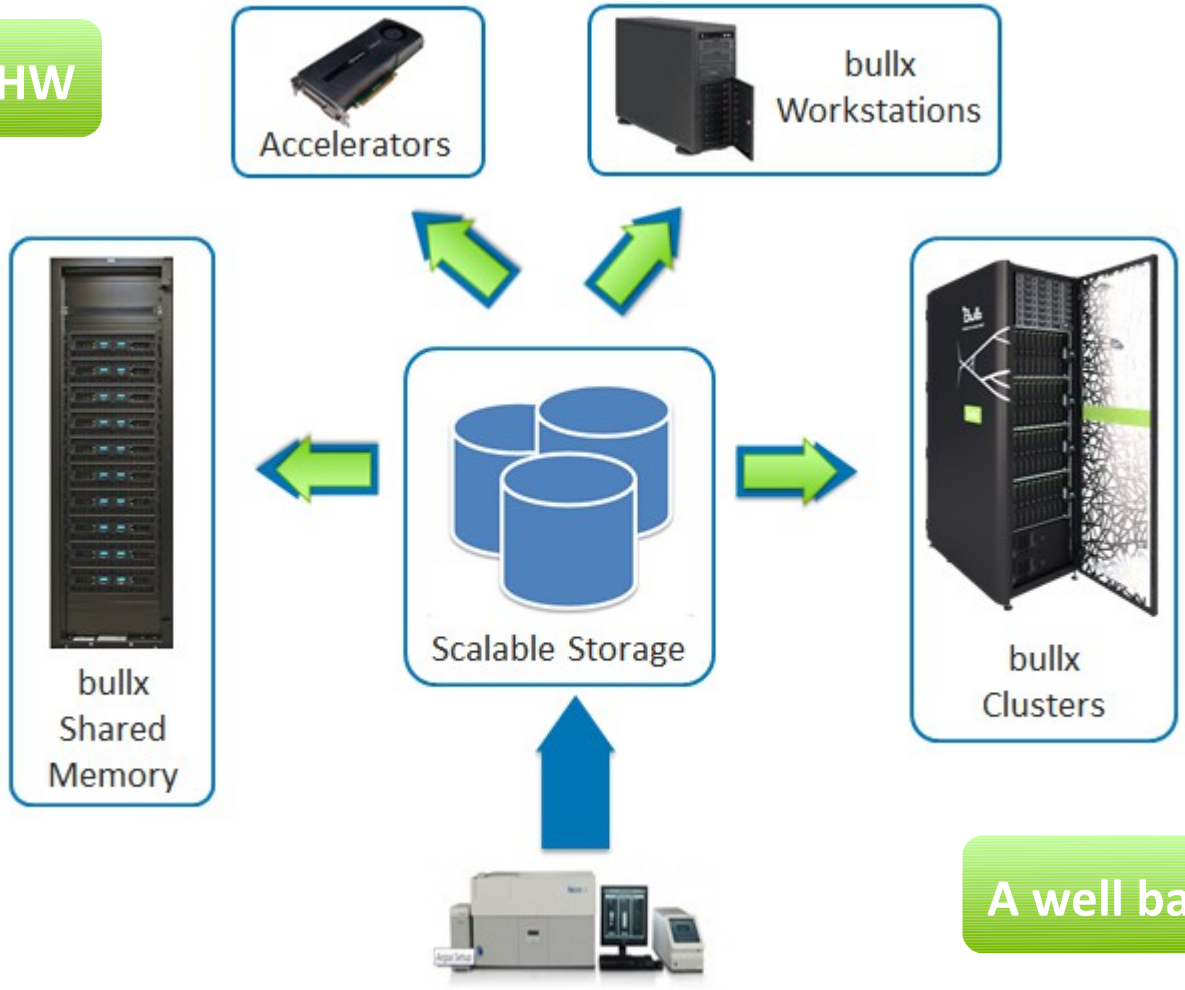
Direct Liquid Cooling DLC rack:

- Dual-pump unit (80 kW cooling capacity)
- Warm water: avoid chillers
- Rack includes 5 chassis, each with:
  - 18 dual-processor nodes
  - Embedded 1st level InfiniBand switch
  - Extra Embedded Gigabit Ethernet switch
  - Optional: Ultracapacitor
- Extremely easy to maintain

Max config	2013	2014
Processors	6 PF	12 PF
Accelerators	20 PF	28 PF

Why all the Fuss?

Know your HW



A well balanced system

Why all the Fuss?

DON'T JUDGE SOMEONE JUST BECAUSE THEY SIN DIFFERENTLY THAN YOU.

Why all the fuss?

Know your SW tools

bscs4 – Development

bullx BM: based on SLURM

- Increased scalability and performance up to 65000+ nodes
- Increased robustness
- Energy efficient management techniques
- Preemption: suspending/resuming lower priority jobs
- fine-grain task placement upon specific cores to obtain better application performance

bullx MPI: MPI library based on Open MPI

- Increased scalability (65k nodes in production & 130k nodes in lab)
- Failover improvements
- Nightly regression tests with real applications
- Intel Phi integration
- Diagnostic tools: profiling library and checking library
- MPI-IO: Lustre integration

bullx MPI: main features

- **Fine-grained process affinity**: A group of processes communicating intensively will have the best possible bandwidth and latency.
- **Long lasting communication detection**: bullx MPI detects long lasting communication transfers and stops active waiting on the incoming communication device. This enables the core of the process to change its

Predefined profiles: fast deployment and not have to manage the 400+ MPI options from the beginning.

Kernel-based data mover: Processes on the same compute node will optimize the memory bandwidth for intranode MPI transfers.

Interconnect sub-optimal use detection: detects situations where the interconnect parameters

http://www.bull.com/extreme-computing-services/cepp.html

The Center is sponsored by

intel

and supports

intel inside Xeon

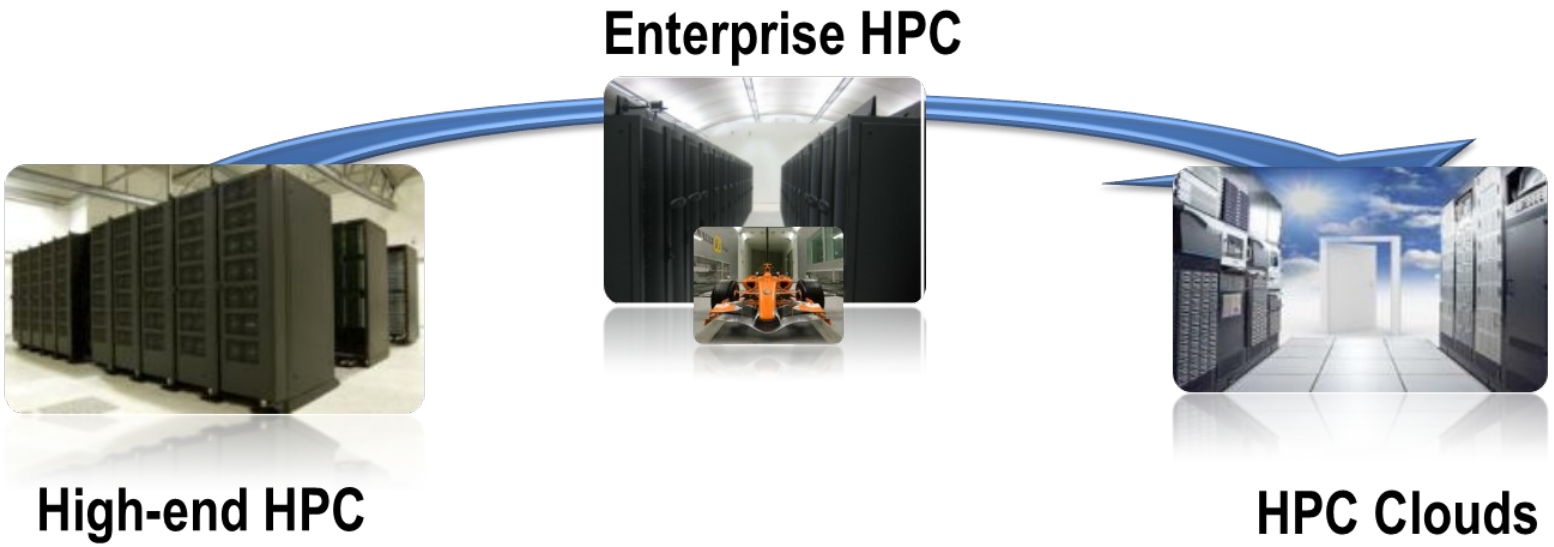
intel inside Xeon Phi

The **first European center of industrial and technical excellence for parallel programming**.

Highest level of expertise and skills : 2000 scientists surveyed:

- 45% spend more time coding than 5 years ago
- 38% spend at least 1/5 of their time coding
- 47% have good understanding of software testing

extreme factory



extreme factory

stay lean: compute smart

Extreme Storage

Need to manage and control deluge in data volume

Concurrent data access from all nodes

- Parallel file systems

1 EByte today is simple:

- With 300M€ you would have 1 EByte with 15 TB/s
  - If you have 1.000.000 cores: 15 MB/sc per core
- Performanace seems OK ... but price doesn't ;)

Disk controllers with Terabytes of cache

¿Will tape survive? ¿Backup to disk directly?

Disk size will surpass 14 TB in 2020:

- Dozens of PetaBytes per rack
- 1 Terabyte in a USB pen drive

SSD evolution: bigger and faster

bullx PFS, Management & Performance

Improved NUMA-IO

Predefined Lustre profiles

- Speed up deployment

HA:

- On each Lustre server node
  - IO backend monitoring through local scripts
  - Errors reported through syslog
- On each Lustre client node
  - Lustre objects connection checking
  - Reading exchange counters in /proc/fs/lustre...
- On a regular basis
  - Write/read/compare test launched on 2 random nodes

- Shine
  - Central management
  - Ease Lustre installation
- Bandwidth aggregation
  - Multi-rail on IB for IO nodes
  - Multi-rail on IB for clients
- Topology Aware:
  - Optimize memory allocation
  - Optimize process placement and utilization
  - Avoid intra-machine data traffic
- bullx MPI (MPI-IO) integration:
  - Applications have an optimized mapping between the MPI parallel I/O functions and Lustre

bscs4 – Power Management

Tool to implement power policies

Predefined rules

If a rack has a very high temperature, PowerManager can tell bullx BM not to send jobs to nodes on that rack

Conclusions

ExaFLOP will solve one of Humanities biggest issues (in CS):

How things are

ExaFLOP will create one of Humanities biggest issues (in CS):

Change

Architect of an Open World™