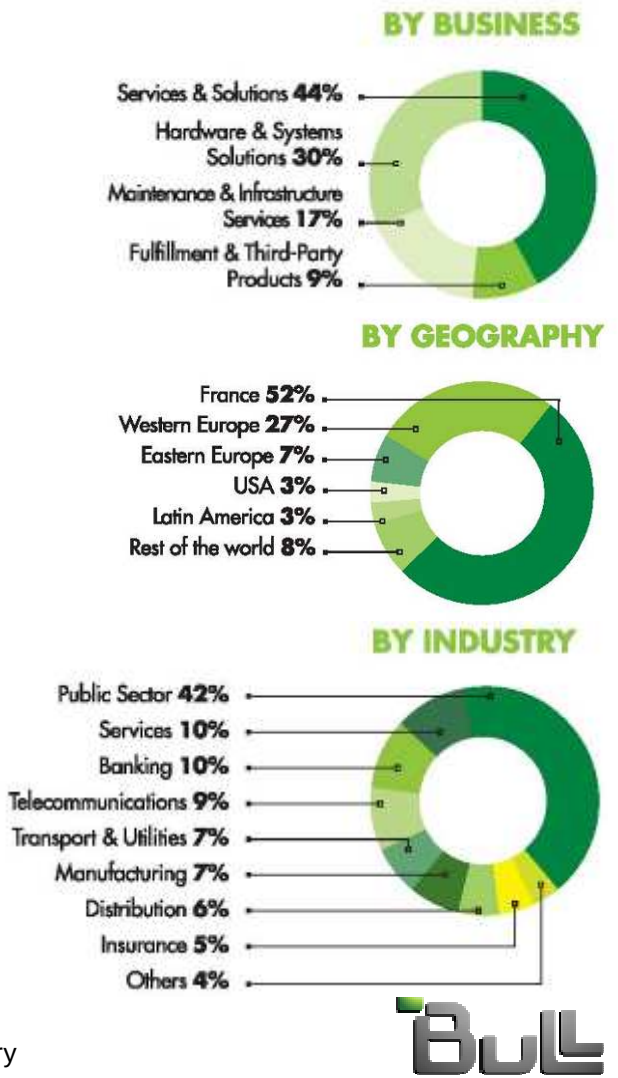

Beyond the Petaflop
Jean-Pierre Panziera, Bull Extreme Computing
BULL
Architect of an Open World

Bull Group

- **A growing and profitable company**
- **A solid customer base**
  - Public sector, Europe
- **Bull, Architect of an Open World™**
  - Our motto, our heritage, our culture
- **Group commitment to become a leading player in Extreme Computing in Europe**
  - The largest HPC R&D effort in Europe
  - 500 Extreme Computing experts - the largest pool in Europe

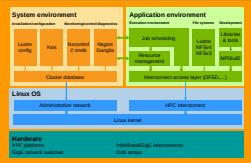
REVENUE BREAKDOWN



©Bull, 2010
Bull confidential and proprietary

bullx hardware for peta-scalability

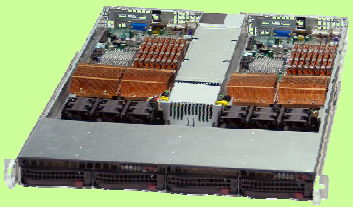
bullx cluster suite



bullx supernodes

bullx blade system

bullx rack-mounted servers



HPC wire Editors'Choice Awards 2009

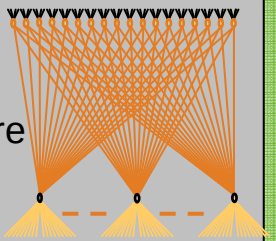
HPC wire Readers'Choice Awards 2009

ACCELERATORS

Water cooling

Storage

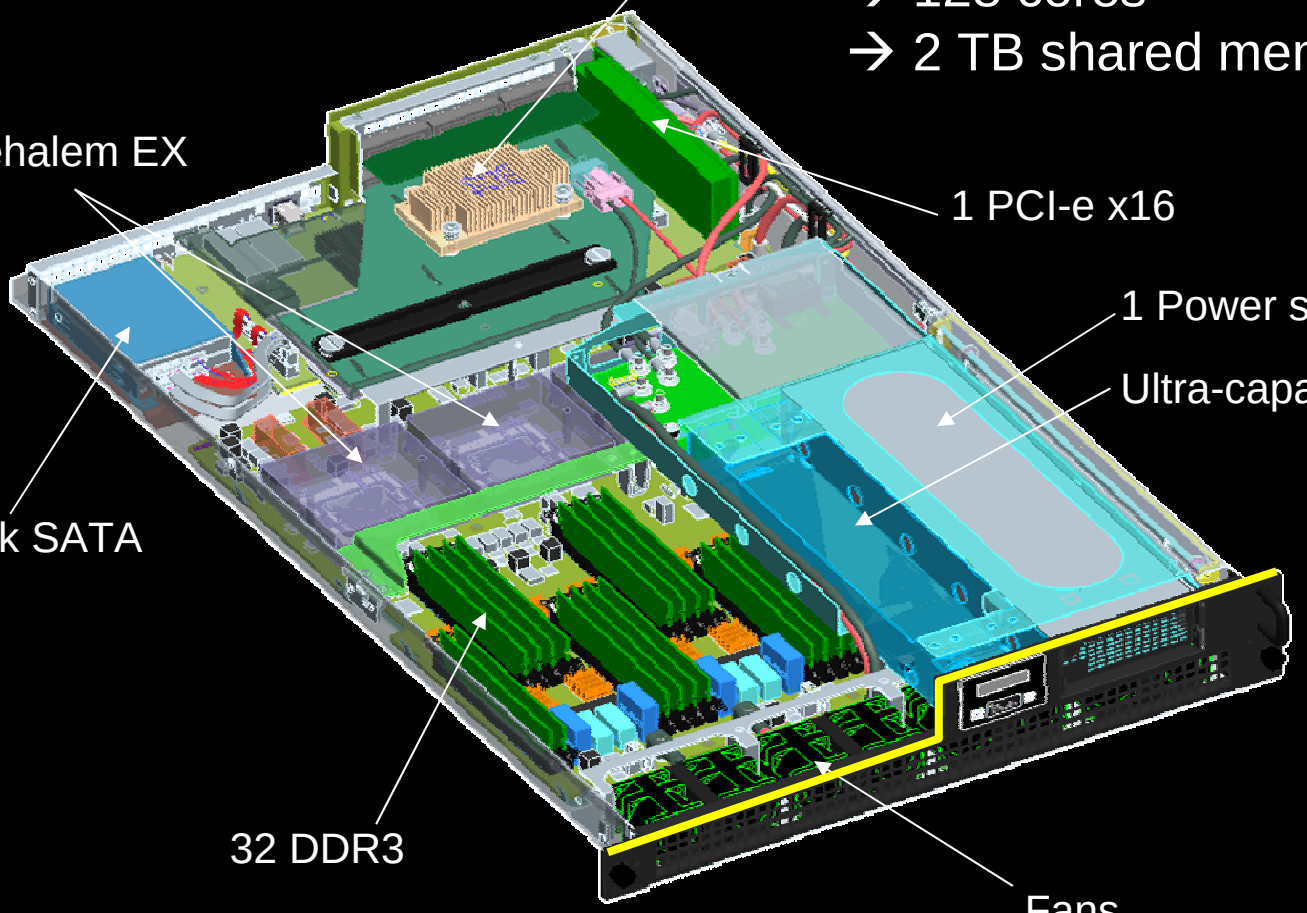
Architecture



©Bull, 2010
Bull confidential and proprietary

bullx-S6010 Compute node; x4 → SMP supernode

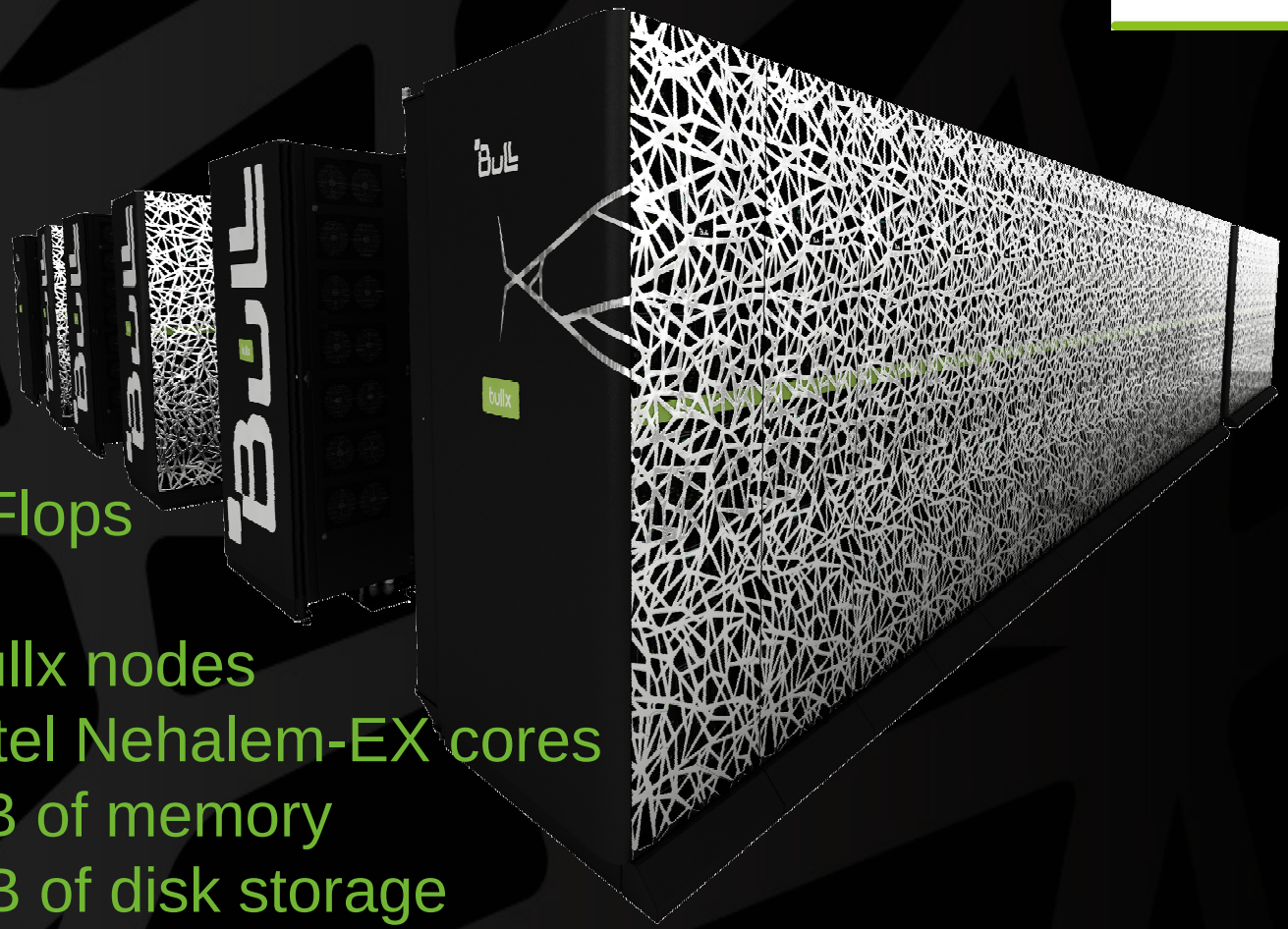
BCS cc-NUMA:
→ 128 cores
→ 2 TB shared memory



Bull confidential and proprietary

Architect of an Open World

TERA 100

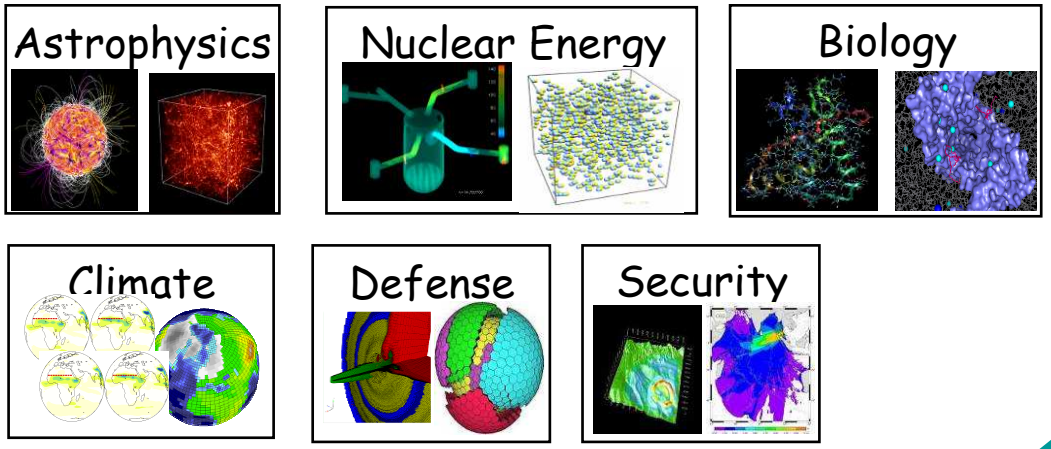


1.25	PFlops
4 300	bullx nodes
140 000	Intel Nehalem-EX cores
300	TB of memory
20	PB of disk storage
	QDR InfiniBand interconnect
500	GB/s bandwidth to the global file system

Architect of an Open World™

TERA 100 a step on the CEA roadmap

CEA : a major actor in the HPC field



Numerical simulation is an essential tool

2001

N°5 Monde
N°1 EU
4,8 TF
TERA 1

x10

N°5 Monde
N°1 EU
63 TF
TERA 10

x20

N° ? Monde
N°? EU
➢**1Pf**
➢**TERA 100**

x30

N° ? Monde
N°? EU
➢**30Pf**
➢**TERA 1000**

x30

N° ? Monde
N°? EU
➢ **1 Ef**
➢**EXA 1**

2019

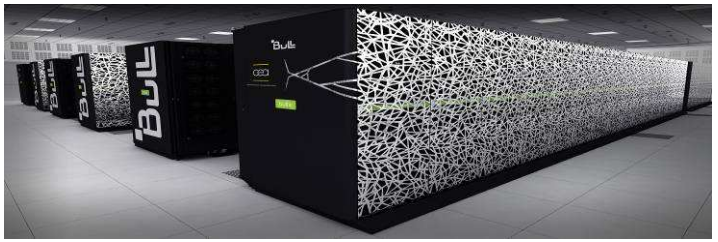
TGCC

PRACE

GENCI

An ambitious roadmap for a strategic goal:

Maintaining the capacity of designing and building very large computing systems in Europe



CEA open computing center

CEA classified computing center

CEA/DAM has the operational responsibility of implementing this roadmap

Beyond the Petaflop

Users' needs keep increasing by orderS of magnitude

- Higher Resolution
  - → much larger models
  - → more computations / element
- Better numerical approximations
  - → more computations / element
- Faster

- ~same budgets
  - → ~same number of "cheap" components
  - → ~same consumption

x 10

x 1000 (3D)

x 10 (time steps)

x ...

x ...

x >1000

Bull
Architect of an Open World™

Beyond the Petaflop; opportunities & challenges

Users' needs keep increasing by orderS of magnitude

- Higher Resolution
  - → much larger models
  - → more computations / element
- Better numerical approximations
  - → more computations / element
- Faster

weak scaling ; MTBF

strong scaling, MTBF

compute intense, strong scaling

strong scaling

- ~same budgets
  - → ~same number of "cheap" components
  - → ~same consumption

Bull confidential and proprietary

Bull
Architect of an Open World™

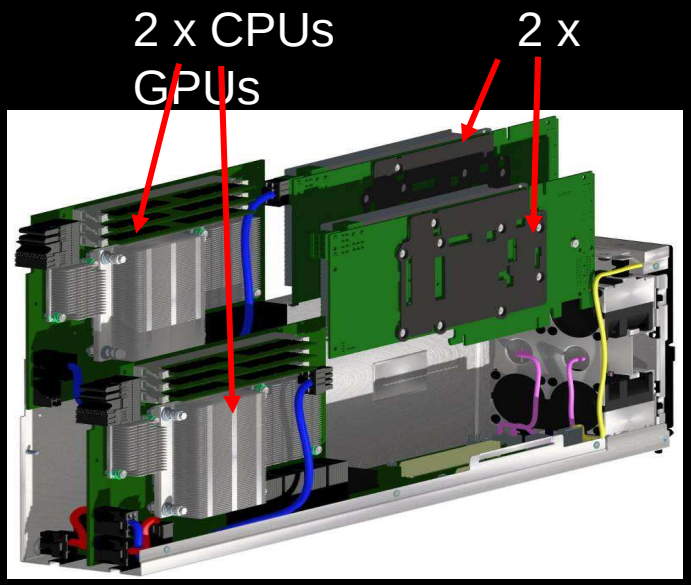
1000x factor

A 1000x improvement requires:

- if relying on technology ***evolution***
  10 hops (2^10 ≈ 1000) → 15-20 years
- a dramatic performance improvement → accelerators
- a significant power consumption reduction → super-green
- a significant performance improvement *and* a dramatic power reduction

Architect of an Open World™

performance: GPU accelerators



	GPU / CPU ratio
GFlops (DP)	7
Memory BW	4.5
consumption	2
Memory Size	1 / 8

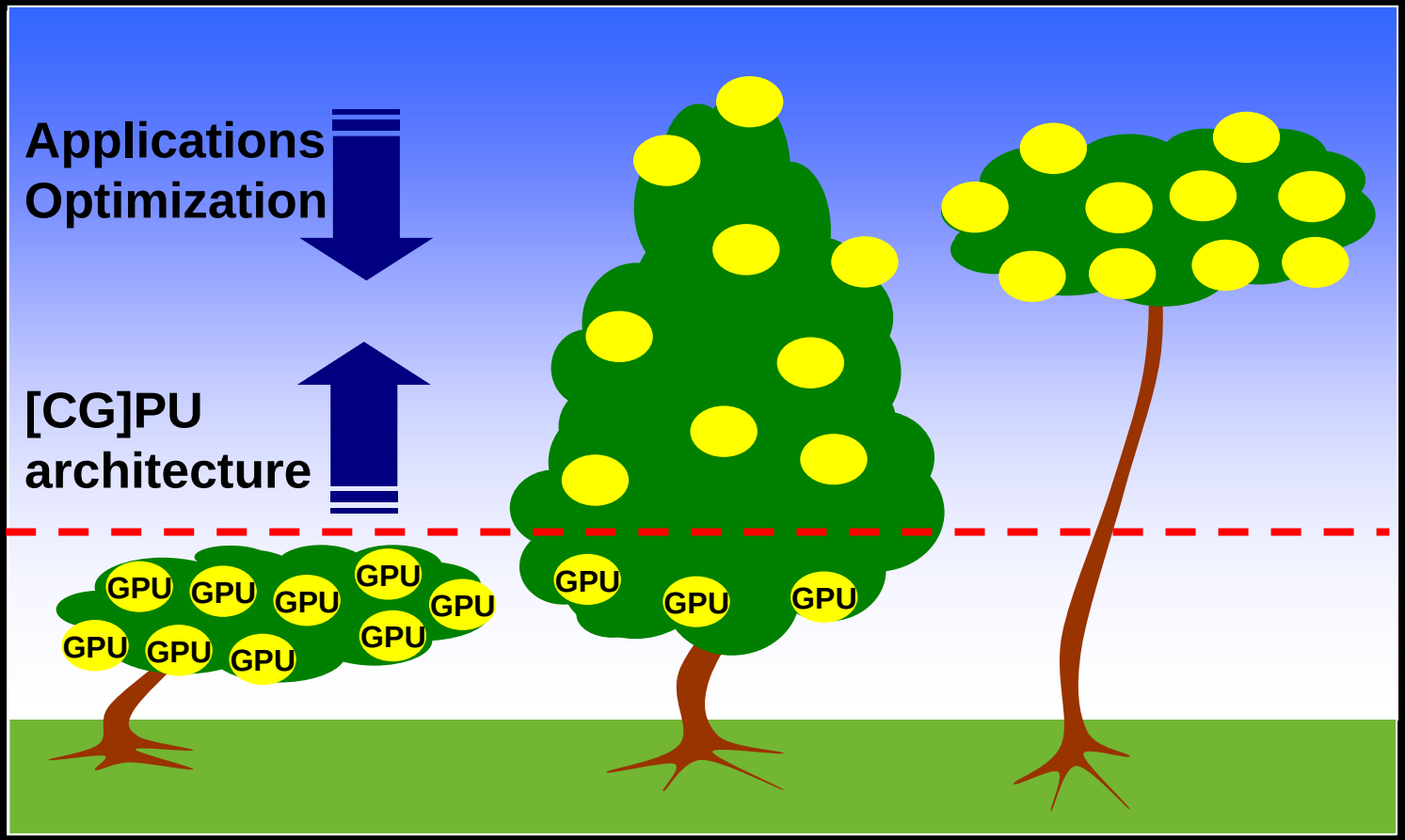
Successfully Accelerated Applications have:

- small kernels
- moderate size datasets, or good data locality
- moderate communications

- *Graphics rendering*
- *Seismic modeling and imaging*
- *Molecular Dynamics, Astrophysics*
- *Financial simulations*
- *Electromagnetism*
- *Genomics*
- *... more ...*

Architect of an Open World

Application tree and GPU low hanging fruits

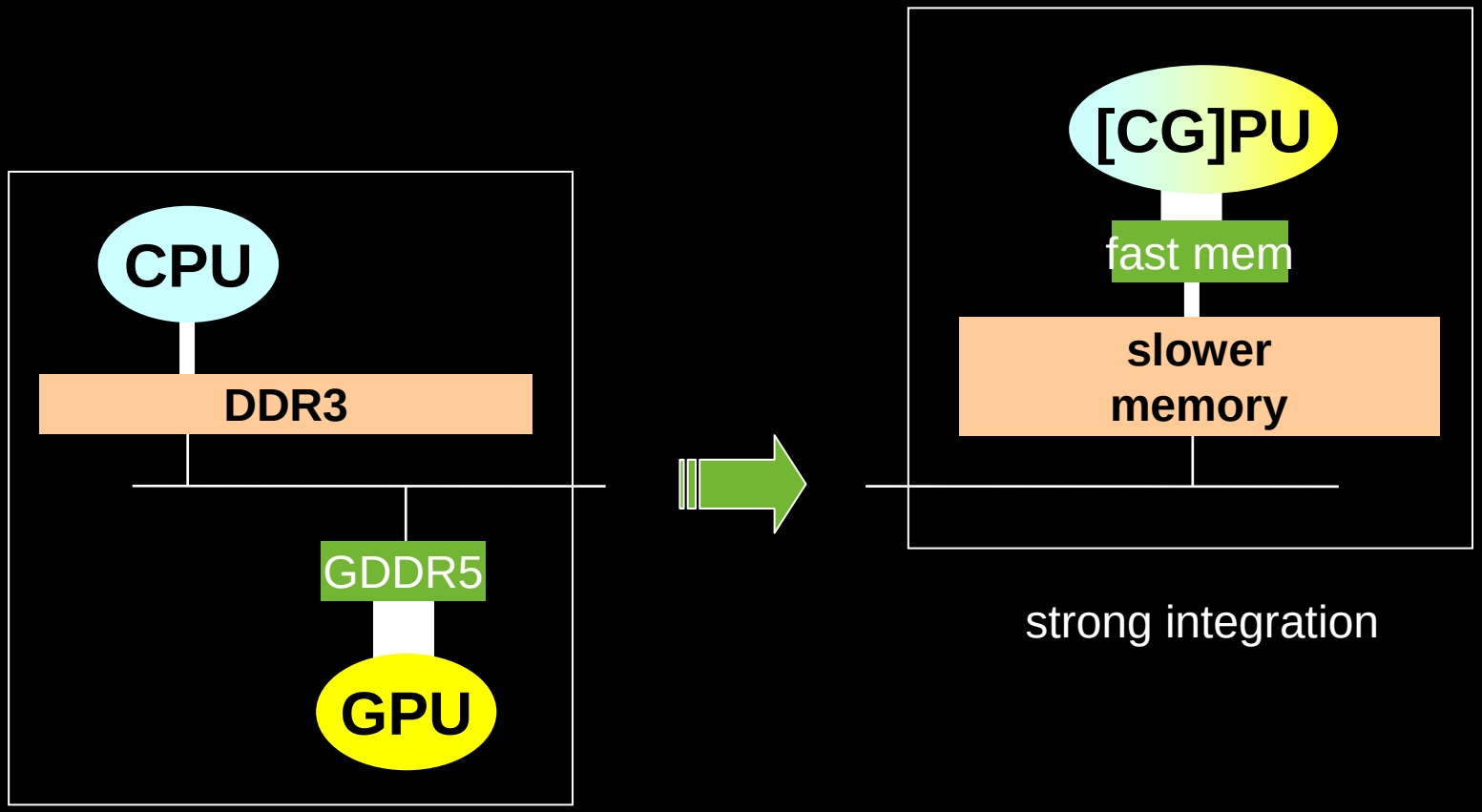


mid-2010: 15-20% applications are accelerated with GPUs

Bull confidential and proprietary

Bull
Architect of an Open World

[CG]PU architecture (r)evolution



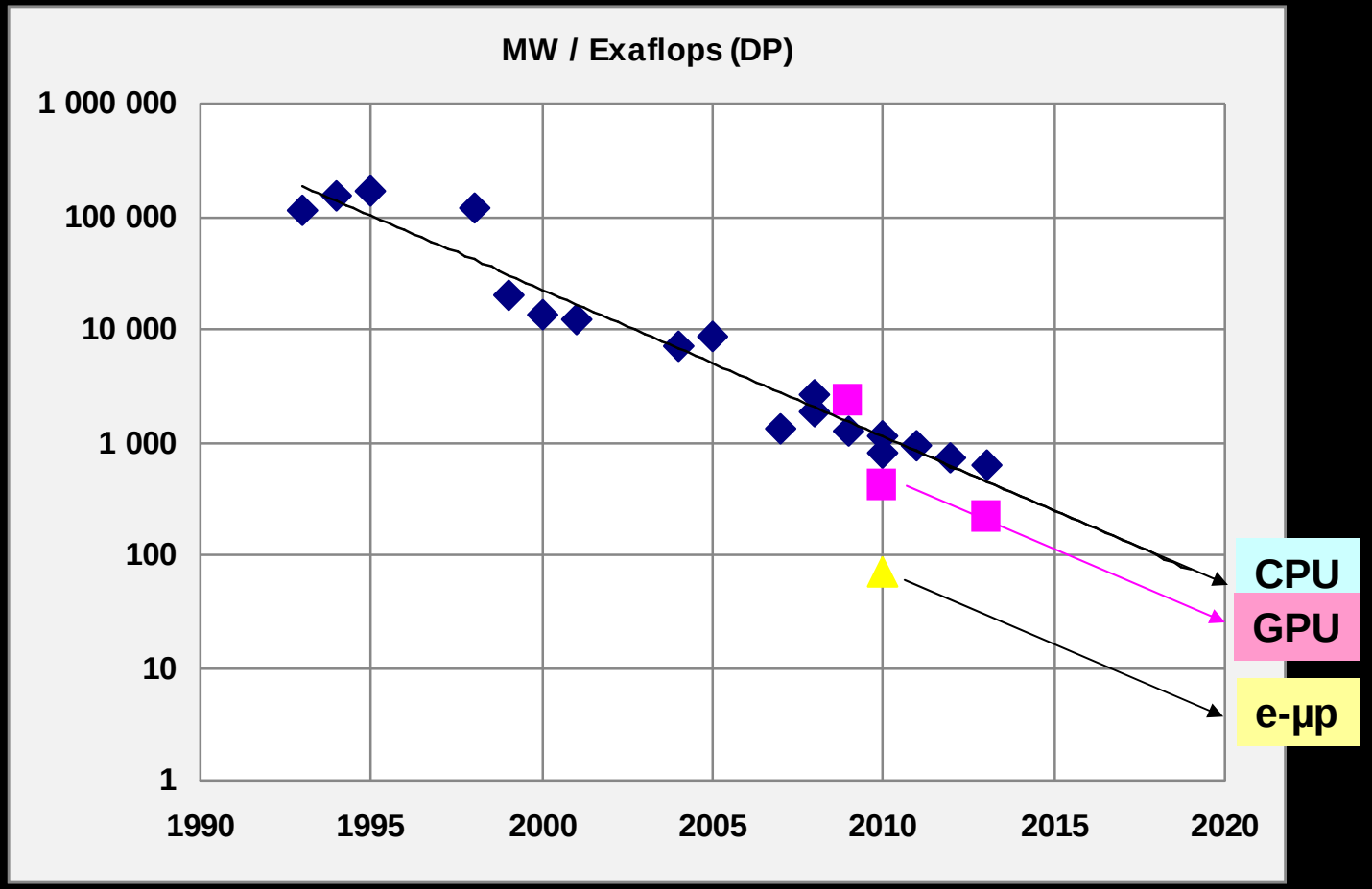
weak integration

strong integration

Bull confidential and proprietary

Architect of an Open World

Power consumption pW/flops (MW / Exaflops)



Source JM Denis

Bull confidential and proprietary

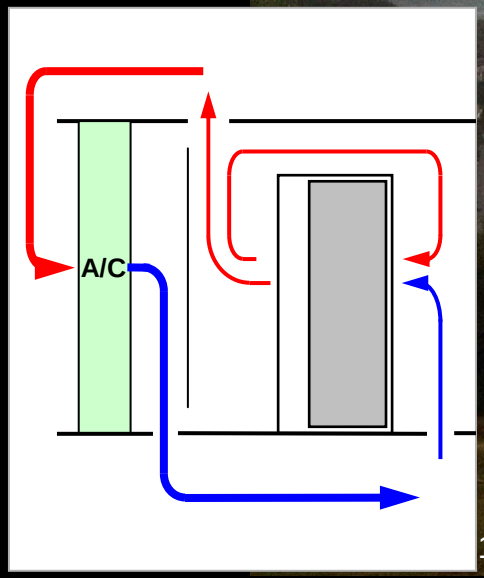
Architect of an Open World

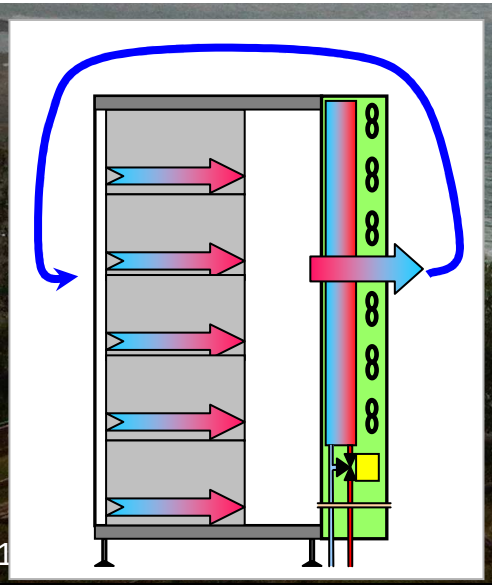
Improving Power Usage Effectiveness → free cooling

Air-cooled		
20 kW/rack		
Room	20°C	
A/C water	7-12°C	
PUE	≥ 1.7	

Water-cooled doors		
40 kW/rack		
Room	20°C	27°C
Water	7-12°C	14-19°C
PUE	1.6-1.7	1.4-1.5

Direct-Liquid-cooling	
60 kW/rack	
Room	27°C
Water	>30°C
PUE	1.1-1.2





1, 201

Architect of an Open World™

bullx cluster suite components

bullx cluster suite

System environment

Installation/configuration

Monitoring/control/diagnostics

- Lustre config
- Ksis
- Nscontrol // cmds
- Nagios Ganglia

Bull System Manager cluster database

Application environment

Execution environment

- Job scheduling
- Resource management

Development

- Libraries & tools
- bullx MPI

File systems

- Lustre NFSv4 NFSv3

Interconnect access layer (OFED,…)

Linux OS

- Administration network
- HPC interconnect
- Linux kernel

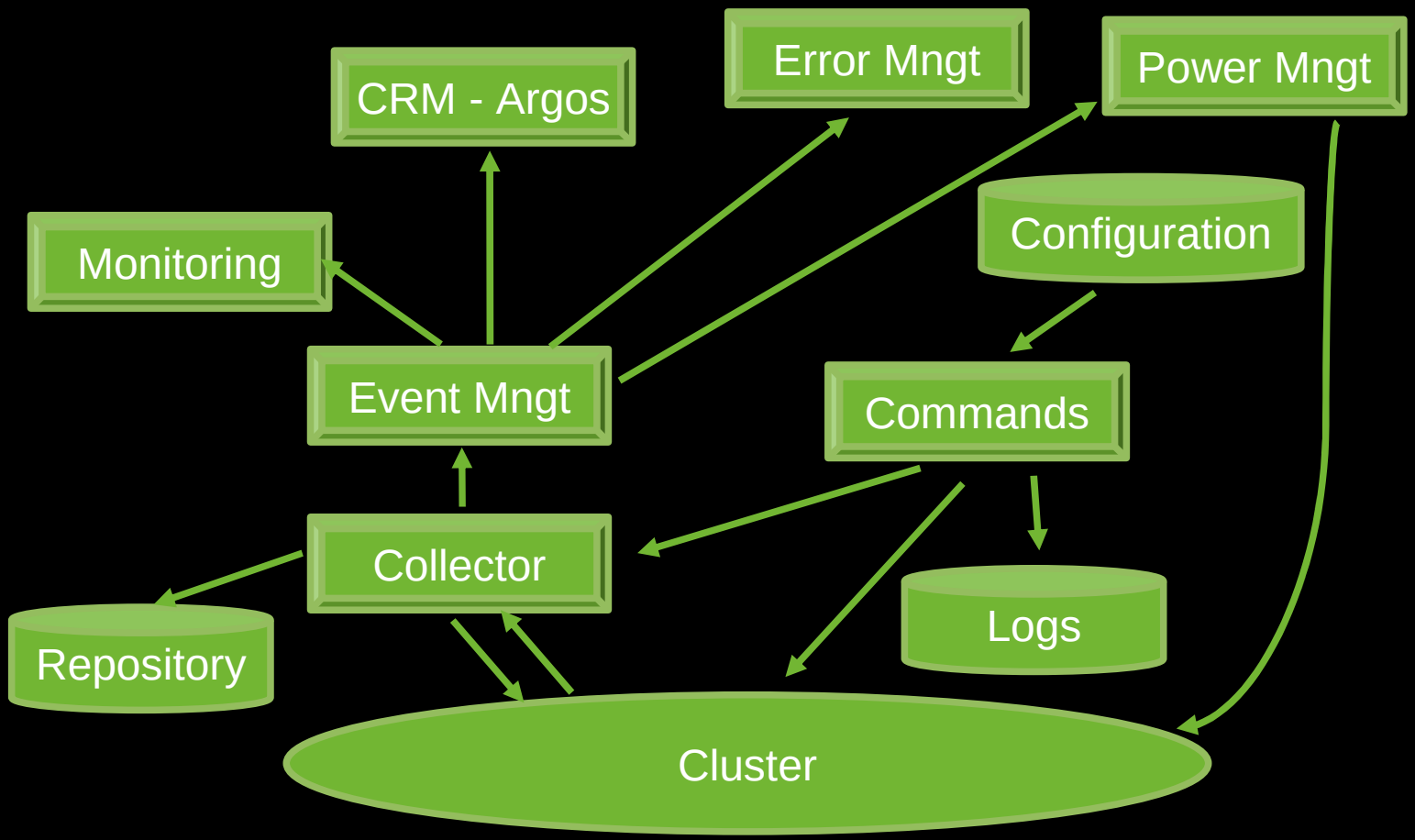
Hardware

XPF SMP platforms
GigE network switches
InfiniBand/GigE interconnects
Bull StoreWay/disk arrays

Bull confidential and proprietary

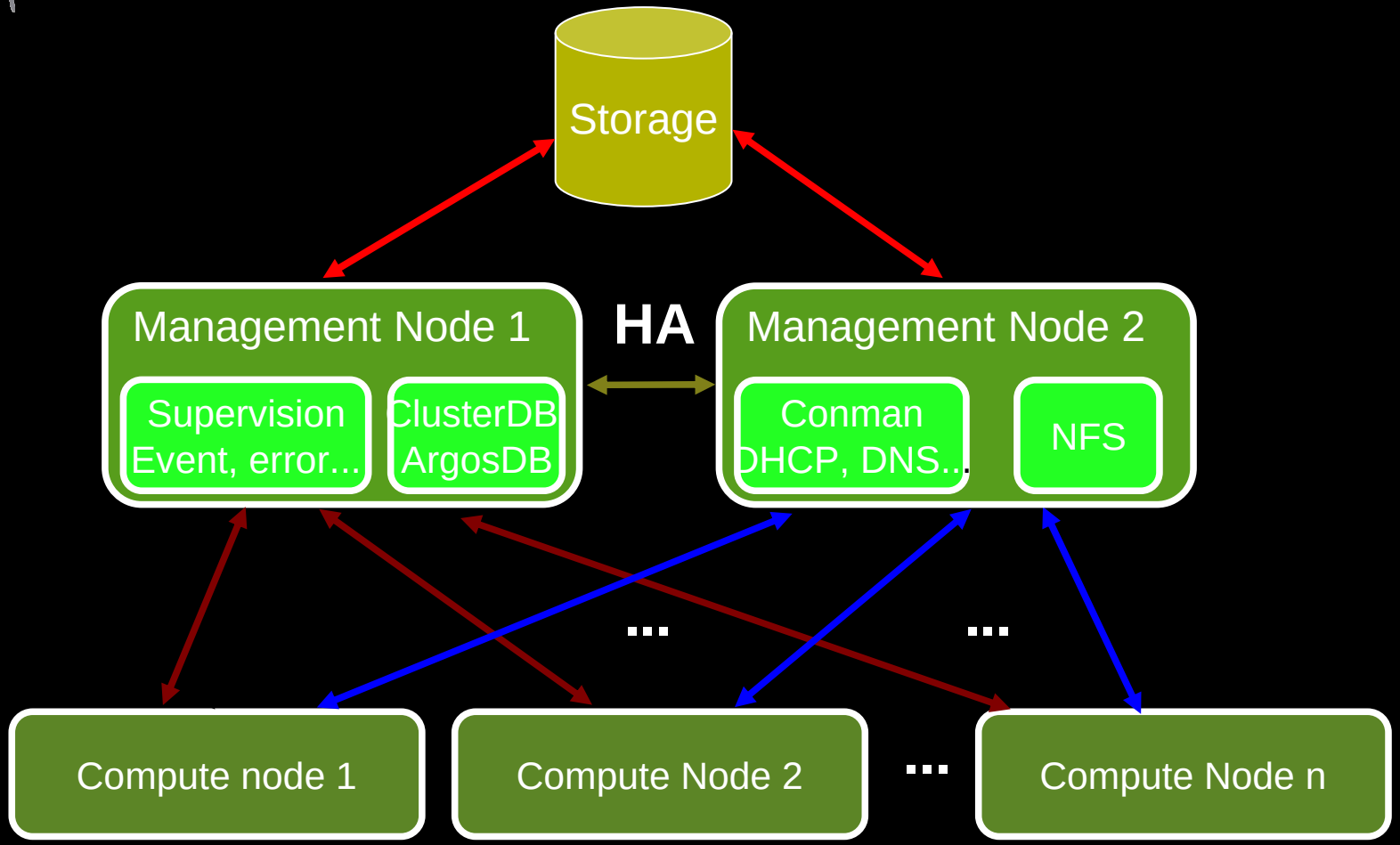
Bull
Architect of an Open World

Cluster Management : functional view



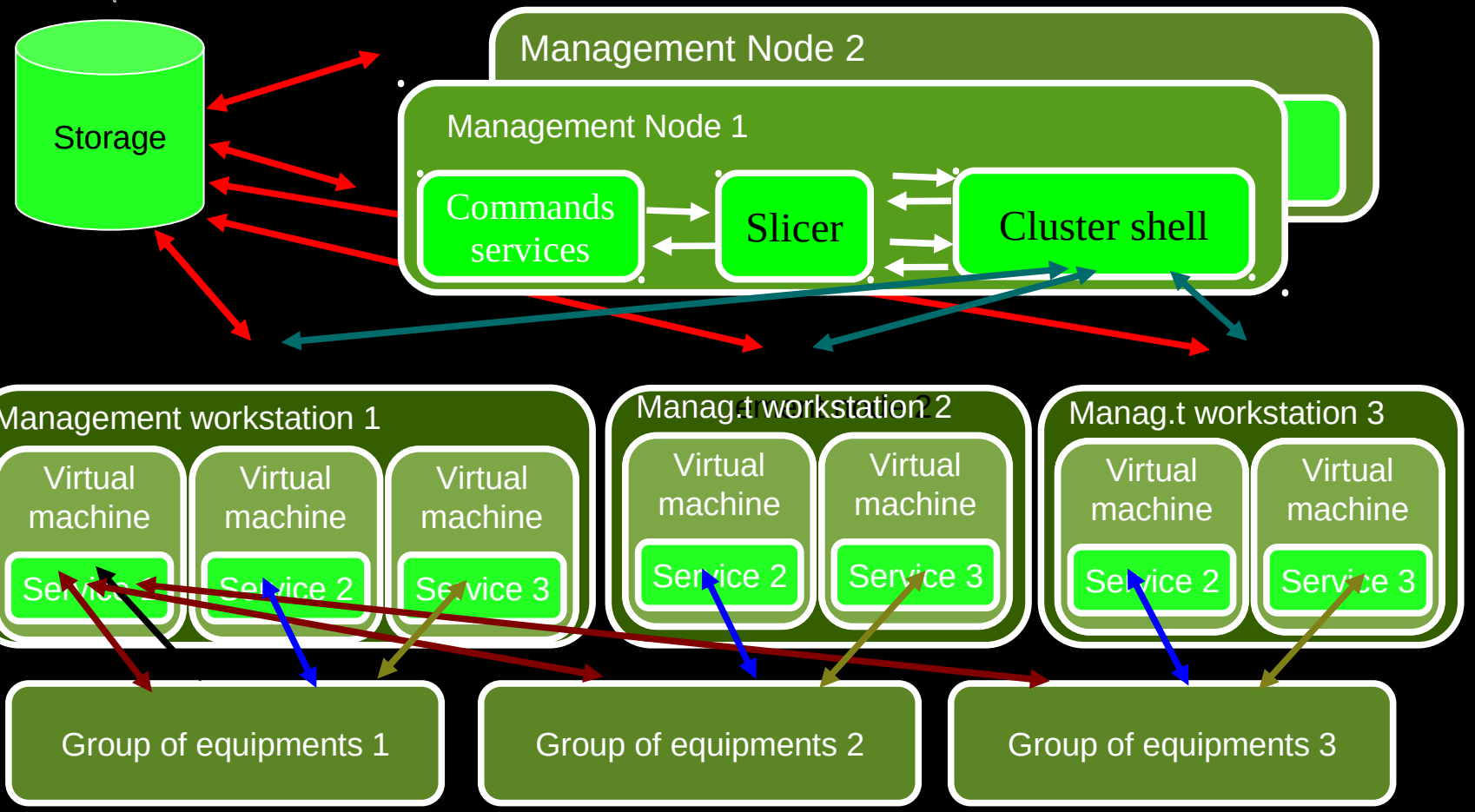
Architect of an Open World

Cluster Management for up to 2000 nodes



Bull
Architect of an Open World

Distributed Cluster Management



Bull confidential and proprietary

Architect of an Open World

Beyond the Petaflop

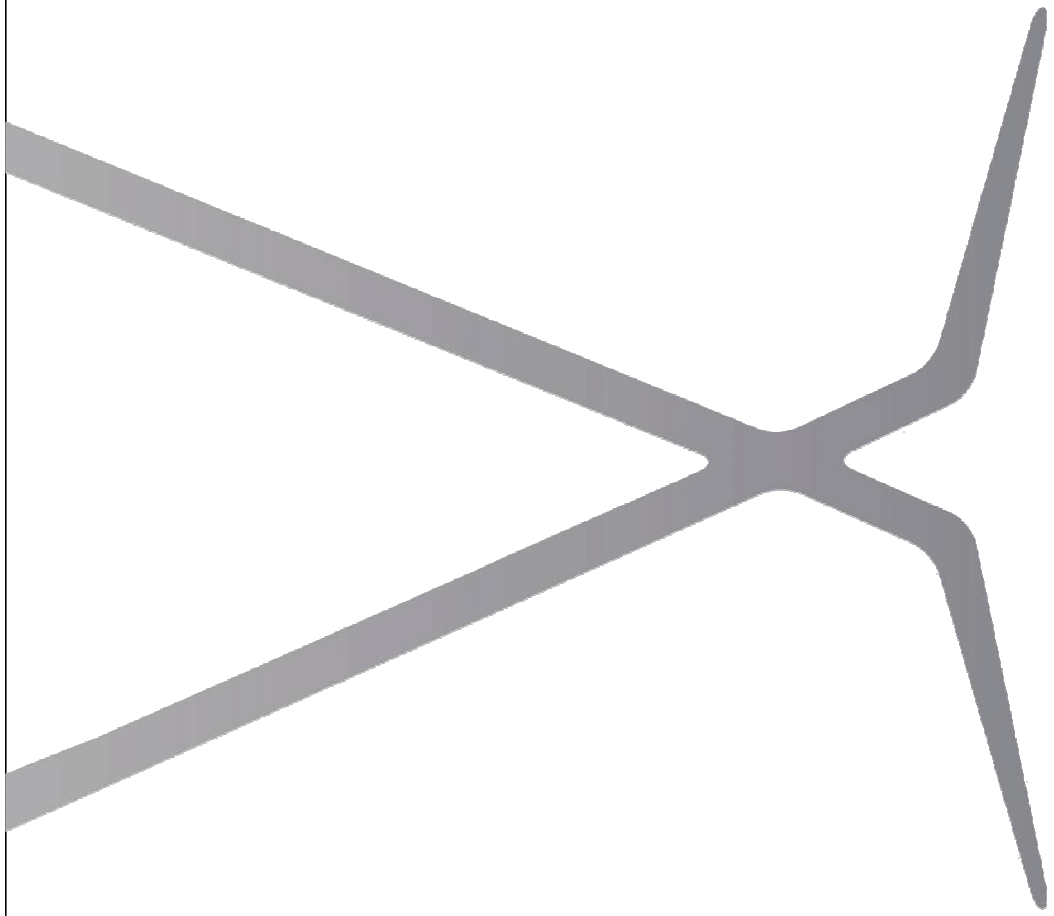
Scaling systems beyond the PetaFlop is a great challenge:

- better components integration
- denser systems
- minimal PUE
- scalable SW management

Scaling applications beyond the PetaFlop is a big challenge

- better development tools
- scalable development tools

Architect of an Open World™



bullx

instruments for innovation