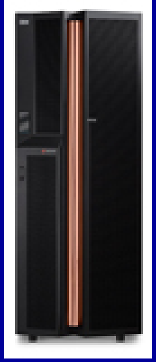
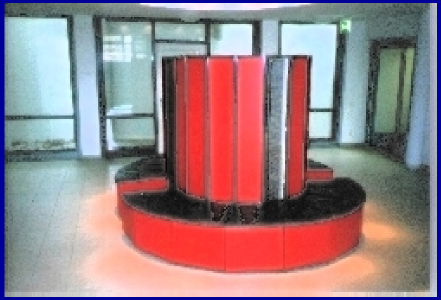


IB usage & perspectives in the Oil&Gas market

S. Requena

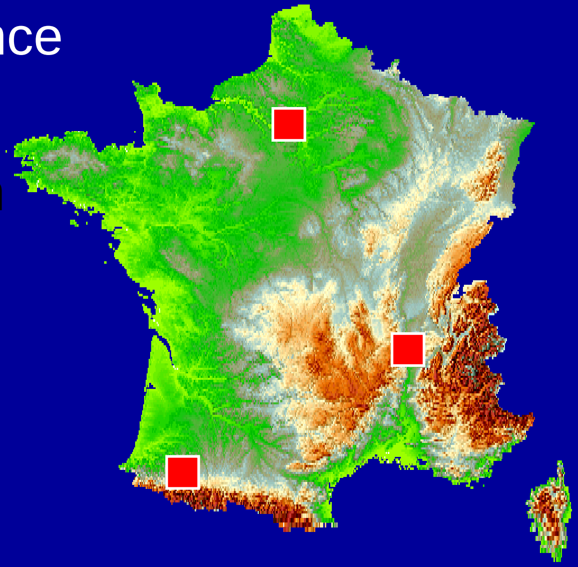


Outline

- Introduction
- IFP's HPC facilities
- IB usage
- IB perspectives and requirements

IFP at a glance

- **IFP (Institut Français du Pétrole)**
  - Independent R&D, Training and Information center for substainable developement in the fields of energy, transport and environment : *www.ifp.fr*
  - Founded in 1944, 3 main locations in France
  - > 1800 people
  - Covered activities : all the Oil & Gas chain
    - Exploration - Reservoir Engineering
    - Driling - Production
    - Refining - Petrochimicals
    - Engines - Fuels



IFP at a glance

- Subsidiaries around the world
  - for performing integrated studies or selling software
  - Beicip Franlab, Axens, RSI, ...

- Some customers and partners
  - Oil & Gas companies (BP, Total, Shell, Aramco, Petrobras, PEMEX, …)
  - Oil & Gas services companies (CGG, Technip-Coflexip, …)
  - Automobile companies (Renault, PSA, Daimler Chrysler,BMW, Ferrari, IVECO, …)

IFP's scientific computing ecosystem

- Personal facilities
  - close to 600 workstations (80% PC, SUN & SGI wks)
  - standardized configurations (HW & SW)
- Shared facilities
  - File servers : NetAPP F940c with 16 TB used for $HOME, shared software and project zones
  - Storage : 2 servers with ADIC tape librairies, Veritas Netbackup (revovery) and SGI DMF (HSM)
  - HPC : 2 clusters (IBM *x*series and *p*series)
  - HPV : a mini 3 nodes IBM DCV cluster for tests

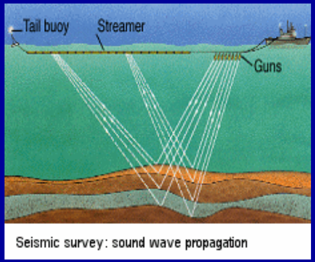
IFP's HPC facilities

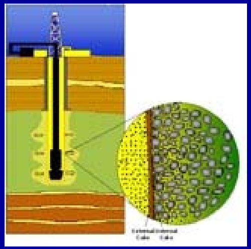
IFP's HPC history

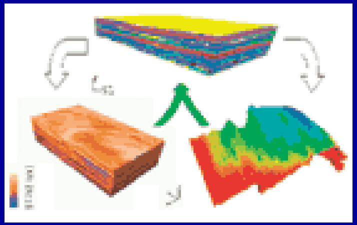
- HPC is mandatory for the Oil & Gas research breakthroughs
  - in the upstream (geophisics, geology and reservoir simulation)
  - in the downstrean (molecular dynamics, car engine simulation)
  - From the sixties to 2003
    - CDC 6600 and 7600
    - Cray XMP 1S and Convex C2
    - Fujitsu VP2400 and VPP500
    - and finaly from 1997 to 2003 : NEC SX5 and SGI O2000

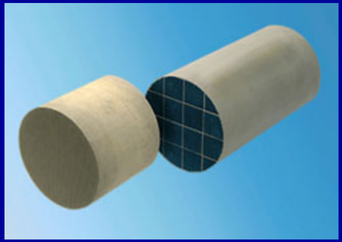
Some oil & gas codes profile

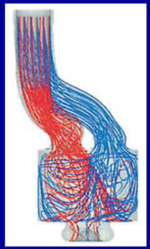
- Upstream market
  - seismic (aquisition, processing, interpolation) : I/O bounded (from 1 to 50 TB of data), big files, parallel I/O, memory bounded
  - geology : memory bounded, MPI bounded
  - reservoir modeling : memory bounded, MPI bounded
- Downstrean market
  - molecular dynamics (catalysts, particulate filters design) : very long simulation time
  - car engine simulation : memory bounded, MPI bounded or OpenMP limited











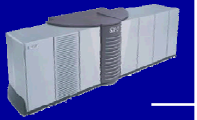
IFP's HPC history

- 80 to 100 people using HPC facilities
- 80% of internal HPC apps coming from all the Oil & Gas chain, with a lot of profiles :
  - Fortran 77, 90, 95, C/C++, Java
  - OpenMP, pthreads, MPI, hybrid
  - memory bound, I/O bound, MPI bound, ...
- Commercial apps used : Fluent, Abaqus, Gaussian, VASP, ...

In 2003
EMC^2 project vision

Now | 1st step (2003-2005) | 2nd step (2006-2008)

2003



Linux Cluster IBM

66 procs Intel 3.06 Ghz

2 GB / node

Myrinet 2000 interconnect

1.5 TB GPFS filesystem

AIX Cluster IBM

16 procs Power4+ 1.7 Ghz

16 to 32 GB / node

Federation interconnect

1.5 TB GPFS filesystem

Single Linux Cluster

- *Thin nodes*
  - 2 procs and 2 to 8 GB mem
  - x86, x86-64 or IA64 based
  - High speed interconnect
- *Wide nodes*
  - 4 to 8 procs and 16 to 64 GB
  - IA64, Powerx based
  - ‘ Medium ’ speed interconnect

Grid Computing

Collaborations

TOTAL

EMC^2 steps

- End 2004 first evolution of the Linux cluster

Linux Cluster IBM

66 procs Intel 3.06 Ghz

2 GB / node

Myrinet 2000 interconnect

1.5 TB GPFS filesystem

AIX Cluster IBM

16 procs Power4+ 1.7 Ghz

16 to 32 GB / node

Federation interconnect

1.5 TB GPFS filesystem

Linux Cluster IBM

66 p Intel 3.06 Ghz, 2 GB / node

104 p AMD Opteron 2.2 Ghz

4 p Intel Itanium2 (SGI Altix 350)

Silverstorm IB 4x 144p

1.5 TB GPFS filesystem

AIX Cluster IBM

16 p Power4+ 1.7 Ghz

16 to 32 GB / node

Federation interconnect

1.5 TB GPFS filesystem

IFP's HPC facilities in 2006

- End 2005 : upgrade & merge

Linux Cluster IBM

66 p Intel 3.06 Ghz, 2 GB / node

4 p Intel Itanium2 (SGI Altix 350)

104 p AMD Opteron 2.2 Ghz

Silverstorm IB 4x 144p

1.5 TB GPFS filesystem

AIX Cluster IBM

16 p Power4+ 1.7 Ghz

16 to 32 GB / node

Federation interconnect

1.5 TB GPFS filesystem

> 2 TFlops peak

> 1 TB mem

Single Linux Cluster

66 p Intel 3.06 Ghz, 2GB/node

4 p Inte anium2 I Al 50)

352 cores 2 GB/node

16 p IB

Silverstorm

1.5 TB GPFS filesy em

one of the largest IB switch in the Oil & Gas industry

IB Usage

IB Usage Feedback

- End 2004 : size of the cluster * 2.5
  - current Myrinet 48 ports switch too small !
  - No big motivation from Myricom and delays in 10G availability
  - Need to connect 84 serveurs (compute+storage nodes)
- Investigations on the IB technology
  - need to understand the "IB galaxy" (Mellanox, Agilent, Infinicon, Topspin, Voltaire, Fujitsu, ...) ➔ many players
  - presentation of the roadmap, # stacks, price, ...
  - benchmarks (porting issue, perf., rdma, IPoIB, GPFS, ...)
- Choice of a 144x Silverstorm switch
    - low latency, BW, low cpu overhead, planned evolutions (4x mini to 12x, DDR, QDR)
    - # stack for MPI, GFX, I/O, admin traffic mutualization
  - Silverstorm : software stack stability, commercial presence, $$$

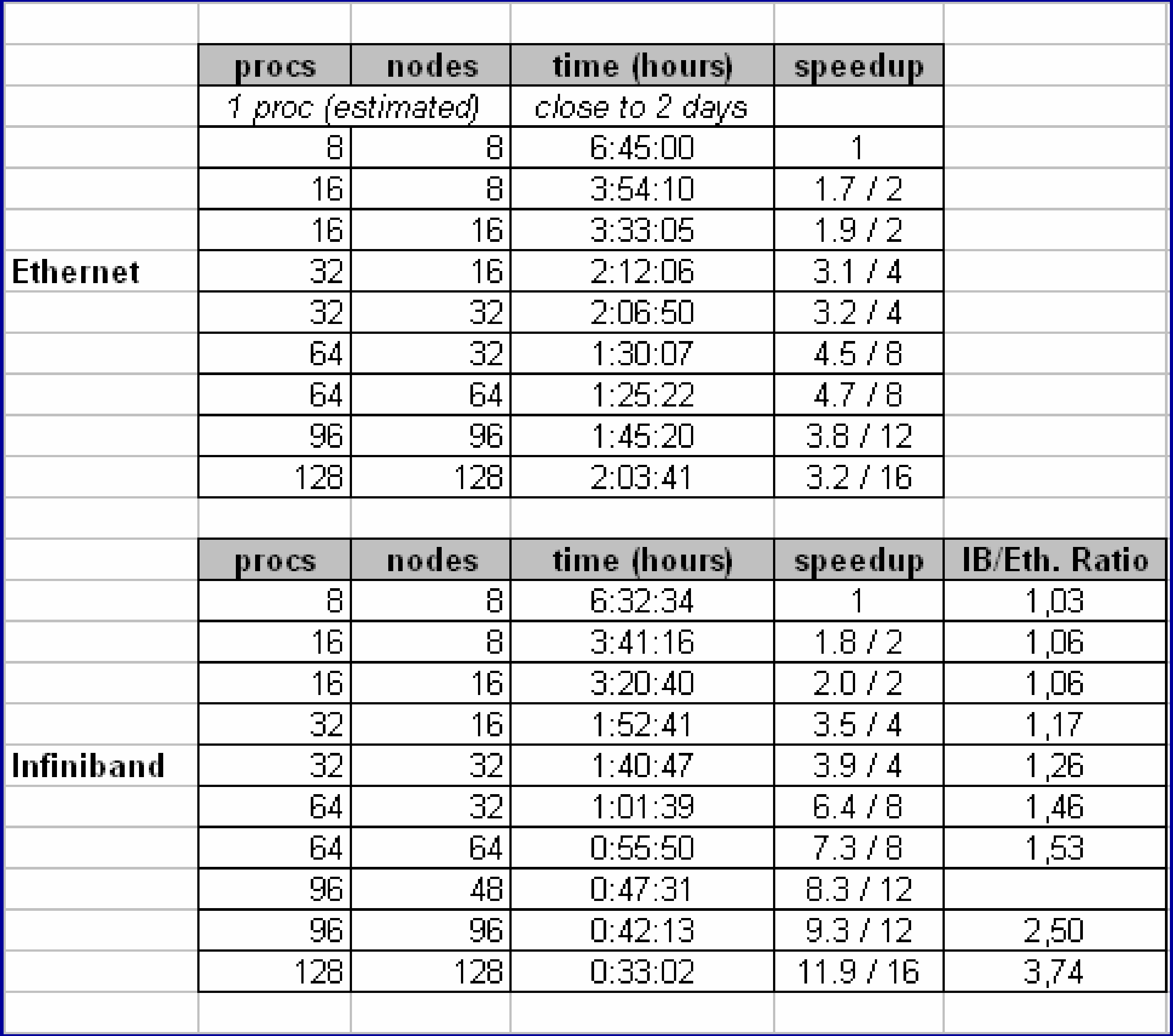
IB Usage Feedback

- Porting issues
  - no big pbs, MVAPICH used then Intel MPI 2.0
    - rdma used for our internal apps, IPoIB for commercial apps
  - GPFS on the IPoIB stack
  - some LSF tuning for the IB support
  - need to play with the mpirun or env variables for
    - memory allocation (RC links) ☹☹☹
    - eager protocol threshold
- Performances
  - On the IB side
    - latency from 5.1 to 5.7 µs, BW from 850 to 1300 MB/s
  - On the IPoIB side
    - latency close to 27 µs, BW : 400 MB/s
    - on our SMC GbE backup network : lat. 51 µs, BW 80 MB/s

IB Usage Feedback

- End 2005 : size of the cluster * 2.5
  - 144x upgraded to 288x ports switch
  - some minor sofware stack instabilites reported
- Since 2 years very good HW and SW stability
  - 70% average load of the cluster
  - all of our internal apps are IB ready
  - Commercial apps ported : Fluent, VASP, Abaqus expected

Parallel upstream code
Performance results



	procs	nodes	time (hours)	speedup
	1 proc (estimated)		*close to 2 days*	
	8	8	6:45:00	1
	16	8	3:54:10	1.7 / 2
	16	16	3:33:05	1.9 / 2
Ethernet	32	16	2:12:06	3.1 / 4
	32	32	2:06:50	3.2 / 4
	64	32	1:30:07	4.5 / 8
	64	64	1:25:22	4.7 / 8
	96	96	1:45:20	3.8 / 12
	128	128	2:03:41	3.2 / 16

	procs	nodes	time (hours)	speedup	IB/Eth. Ratio
	8	8	6:32:34	1	1,03
	16	8	3:41:16	1.8 / 2	1,06
	16	16	3:20:40	2.0 / 2	1,06
	32	16	1:52:41	3.5 / 4	1,17
Infiniband	32	32	1:40:47	3.9 / 4	1,26
	64	32	1:01:39	6.4 / 8	1,46
	64	64	0:55:50	7.3 / 8	1,53
	96	48	0:47:31	8.3 / 12	
	96	96	0:42:13	9.3 / 12	2,50
	128	128	0:33:02	11.9 / 16	3,74

From days to minutes …

basin modeling :

5 M meshes

Opteron 2.2 Ghz

IB 4X SDR

SMC GbE

IB outperforms GbE from 3 to 374%

Parallel upstream code
Customers feedback since 4 years

- Size of the meshes ↗
  - in 2001 : 10k to 30k meshes
  - in 2002 : 300k to 1M meshes
  - in 2005 : 1M to 10M meshes

> 300 x bigger !

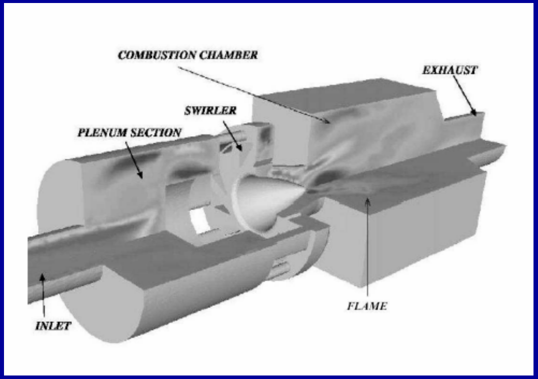
- Performance of visco3d calculator and platforms ↗
  - On a typical 20th century pressure run (30x60x20 meshes)
    - in 2000 on a SGI Origin 2000 system
      - 5550s on 1 proc before optimization and //
      - 170s on 8 procs
    - in 2005 on a Linux Cluster (3.06 Ghz Intel Xeon procs)
      - 21s on 8 procs

> 265 x faster !

- Price/performance of the platforms ↘
  - in 2000 : a SGI Origin 2000 with 16 procs : ~700 k$
  - in 2005 : a Linux Cluster with 16 procs : ~ 35 k$

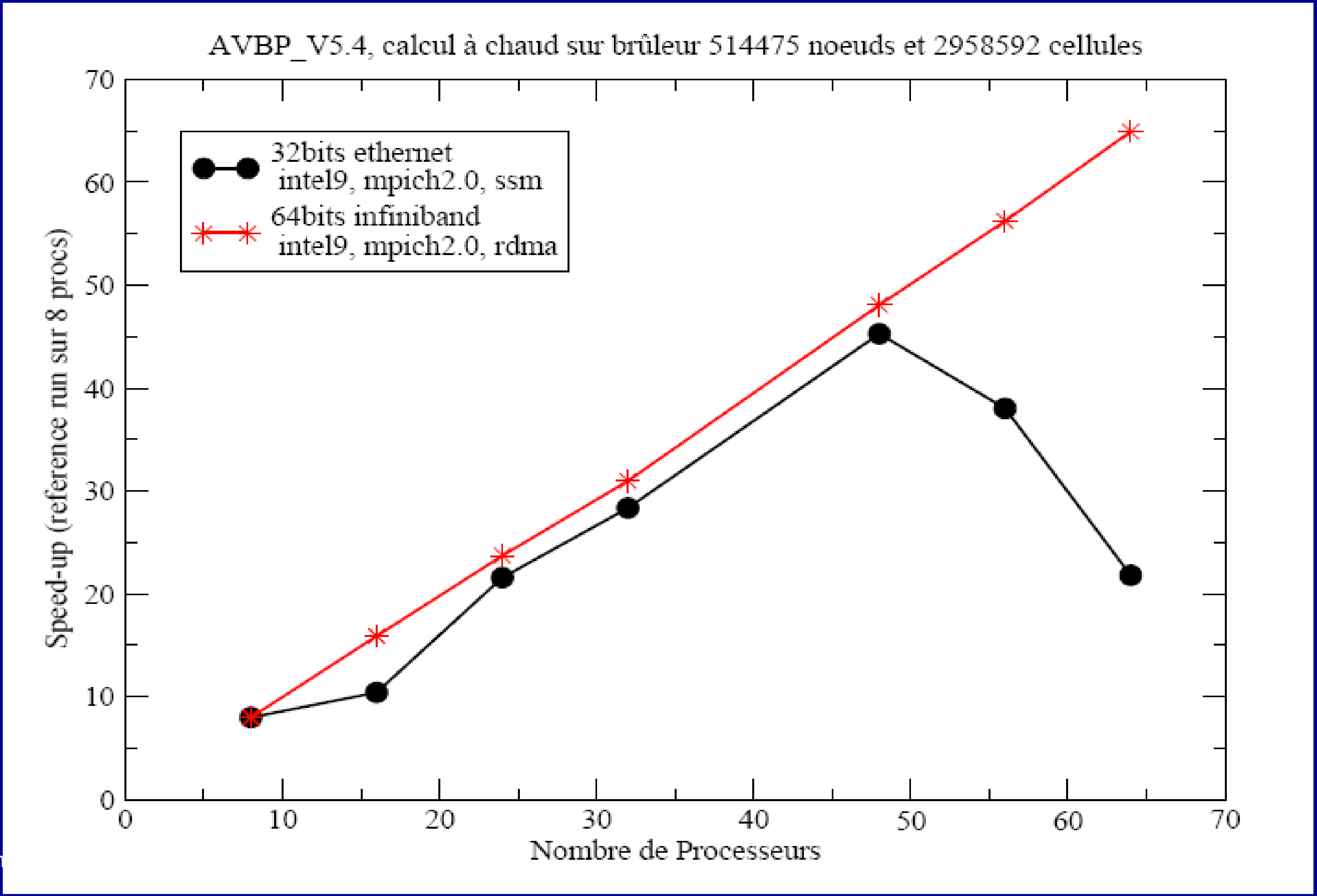
> 20 x cheaper !

Parallel dowstream code
3D LES car engine combustion



- AVBP code
  - co-developed by Cerfacs and IFP
- A world class simulation tool
  - unsteady, fully compressible reactive solver with higher order FV & FE schemes
  - proven very high efficiency on massively parallel machines
    - to date the only CFD code having achieved a nearly linear speed-up on 5000 processors of an IBM BlueGene
  - unstructured meshes
- Special features available for piston engines
  - moving meshes with ITC mesh management
  - CFM-LES for premixed, spark ignited combustion

Parallel dowstream code RDMA vs IPoIB performance



IB perspectives and requirements

IFP IB perspectives

- On the compute side
  - Tests and validation of DDR HCAs
    - low latency and bandwith for multicore based nodes
  - Direct connexion of HCAs on memory (HTX, IBM Galaxy)
    - adoption of the HTX adapter by vendors (PathScale, IBM, ?)
  - Validation of the software stack
    - Intel MPI, OpenMPI
  - Windows CCS 2003 tests planned
- On the graphics side
  - Integration of graphics blades on the cluster
  - No data transfert to the client wks, // rendering done on the cluster and compressed output sent thru the LAN

IFP IB perspectives

- On the storage side
  - Evolution of our GPFS infrastructure
  - native IB of GPFS expected (on SRP, iSER, ???)
  - tests of alternative solution : Lustre
  - I/O bays directly connected on the IB switch, no more Brocade !
    - GPFS SAN but also our NetAPP filers

- Need for bandwith, QoS and interoperability !!!

Whishlist on some IB requirements

- single open source SW stack : OpenIB
  - allows HW interoperability, SW stability, integrated on Linux kernel
- multiples HCAs per node
- performance improuvements
  - IPoIB
  - memory allocation (RC)
- storage protocols simplifications : SRP, iSER, NFS/RDMA
- Traffic counters for profiling networks performance
- Integration of IB monitoring tools on CSM or Ganglia
- Dedicated HW for MPI operations (FPGA ?)
  - optimized collectives communications
  - hardware barriers

Thanks you !!!

Questions ?

More info : stephane.requena@ifp.fr