Finis Terrae Architecture

Ignacio López Cabido
Deputy Technical Director
CESGA

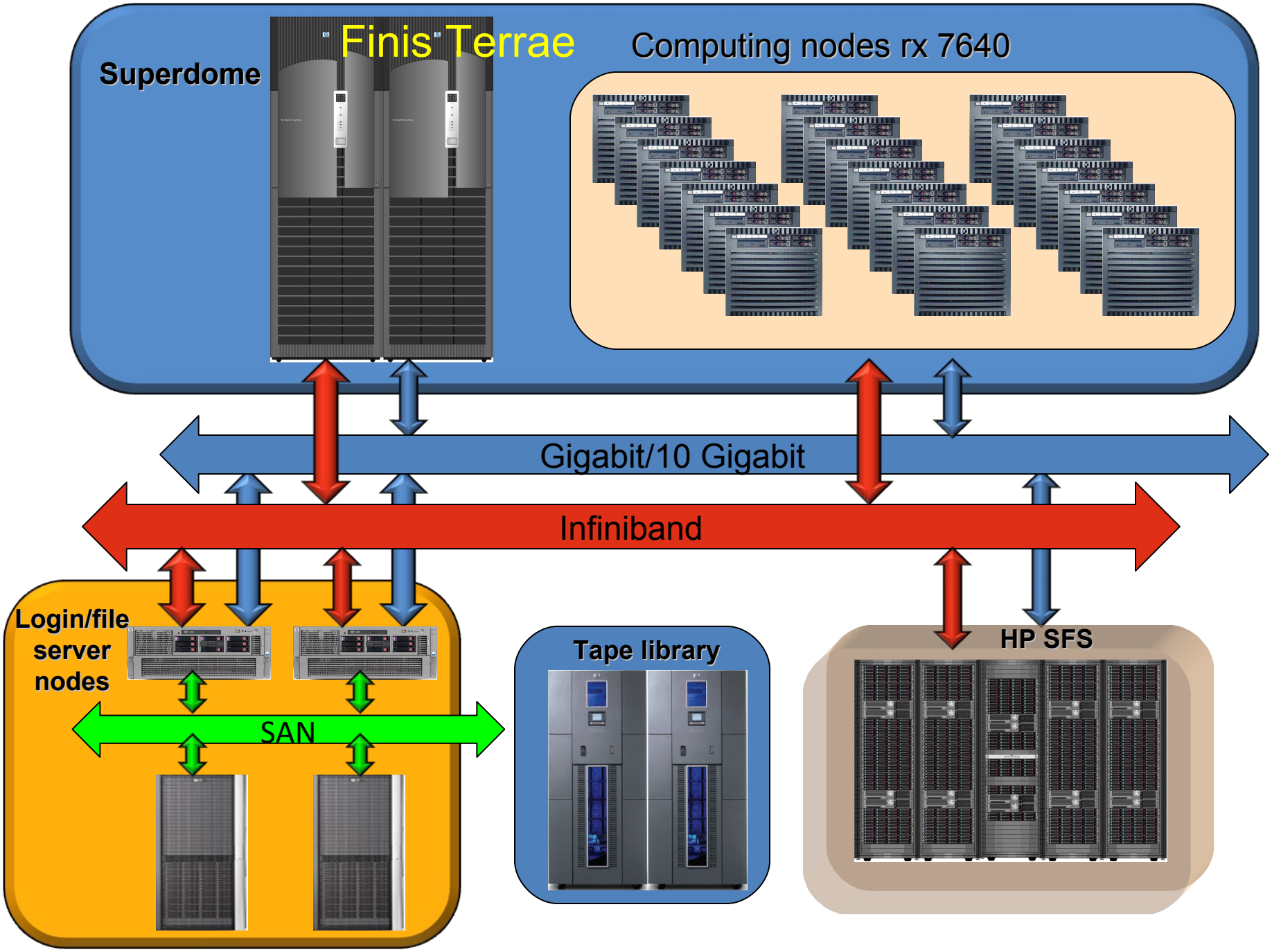
Organized by: CESGA
In collaboration with: XUNTA DE GALICIA, CONSELLERÍA DE INNOVACIÓN E INDUSTRIA, Dirección Xeral de Investigación, Desenvolvemento e Innovación; CSIC
With the financial aid of: GOBIERNO DE ESPAÑA, MINISTERIO DE CIENCIA E INNOVACIÓN; CESGA

Agenda

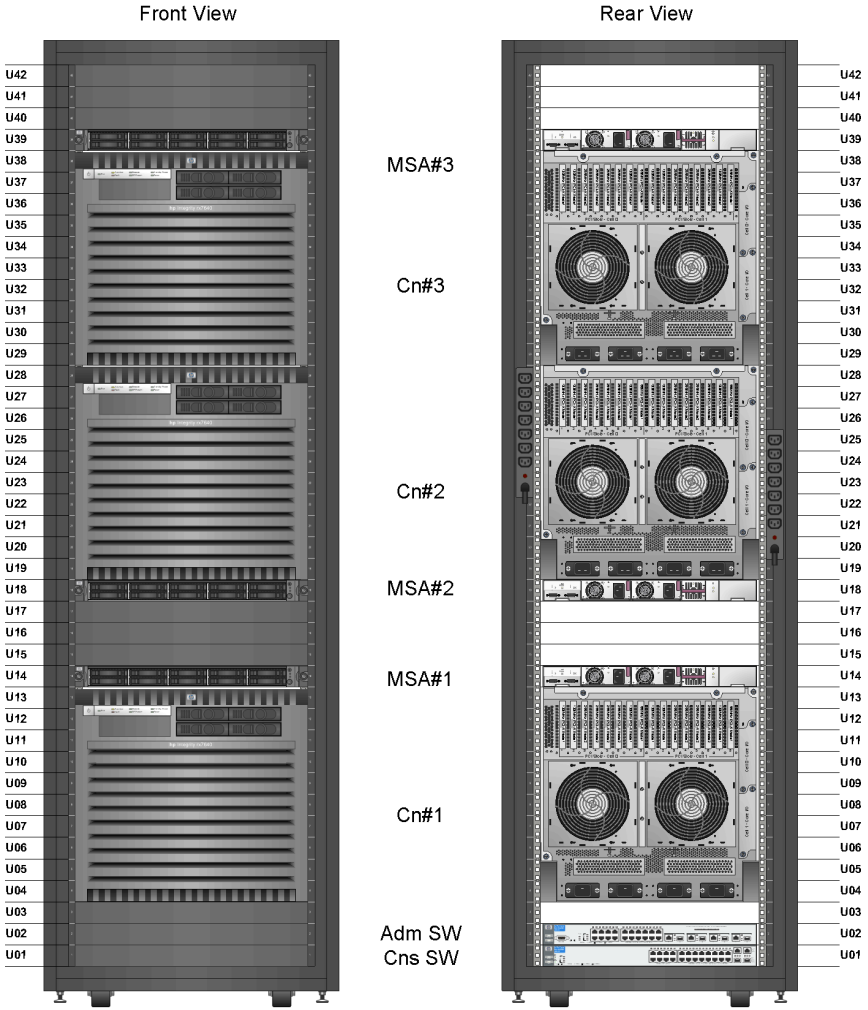
- Finis Terrae architecture overview
  - Computing nodes
  - Superdome
  - Infiniband
  - SFS
- Some results
  - Linpack
  - Stream
  - Pallas MPI
  - Iozone

Finis Terrae: The big numbers

- >2528 processors. 2400 Itanium 2 Montvale processor cores @ 1.6 GHz, 18 MB cache/processor
- 20 TB total memory. Shared memory architecture
- Infiniband interconnect
- SUSE Linux
- 16 TFLOPS peak (≈ 90% actual)
- TOP100 (Nov07 list)

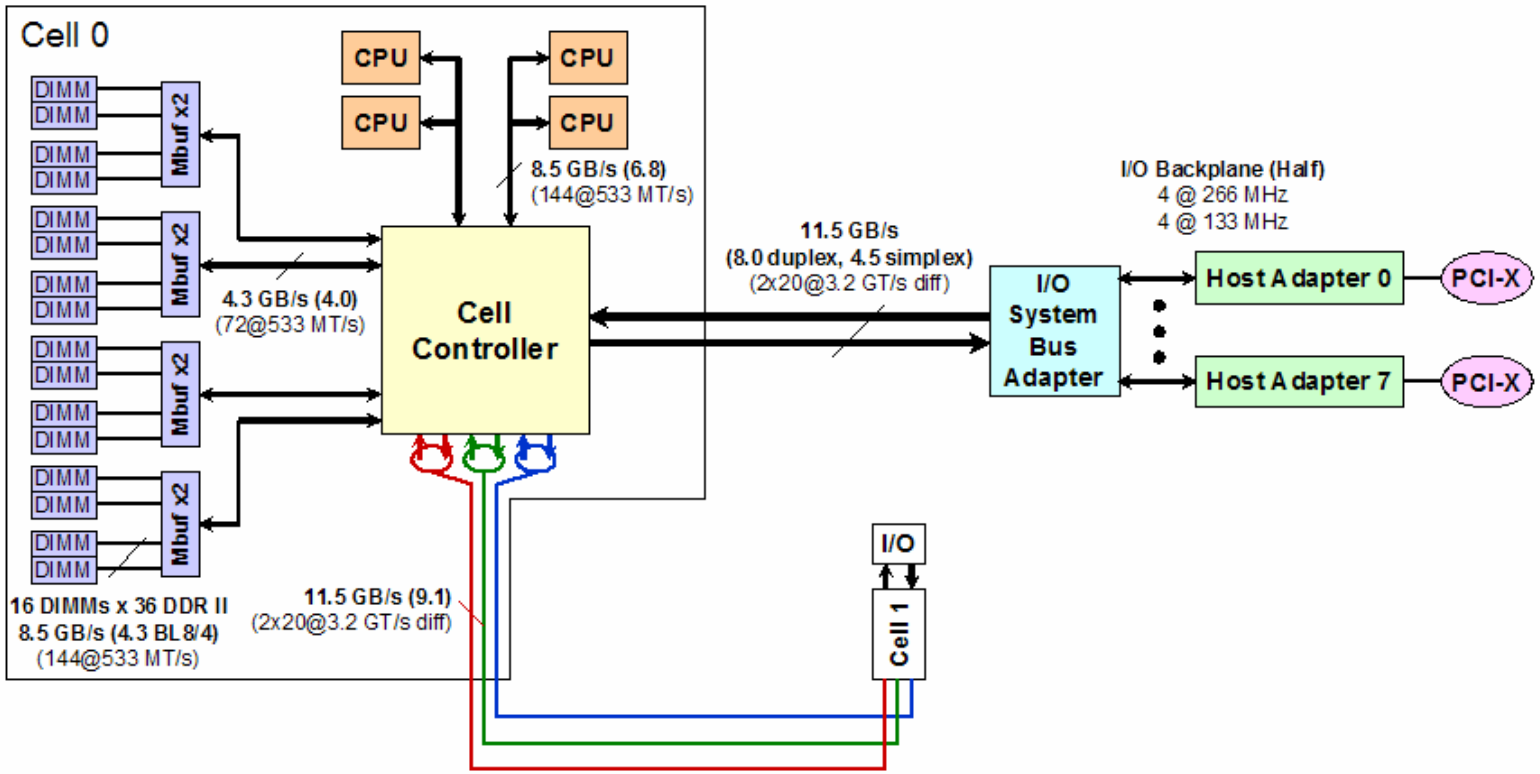

Finis Terrae
Superdome
Computing nodes rx 7640
Gigabit/10 Gigabit
Infiniband
Login/file server nodes
SAN
Tape library
HP SFS

Computing racks: rx 7640



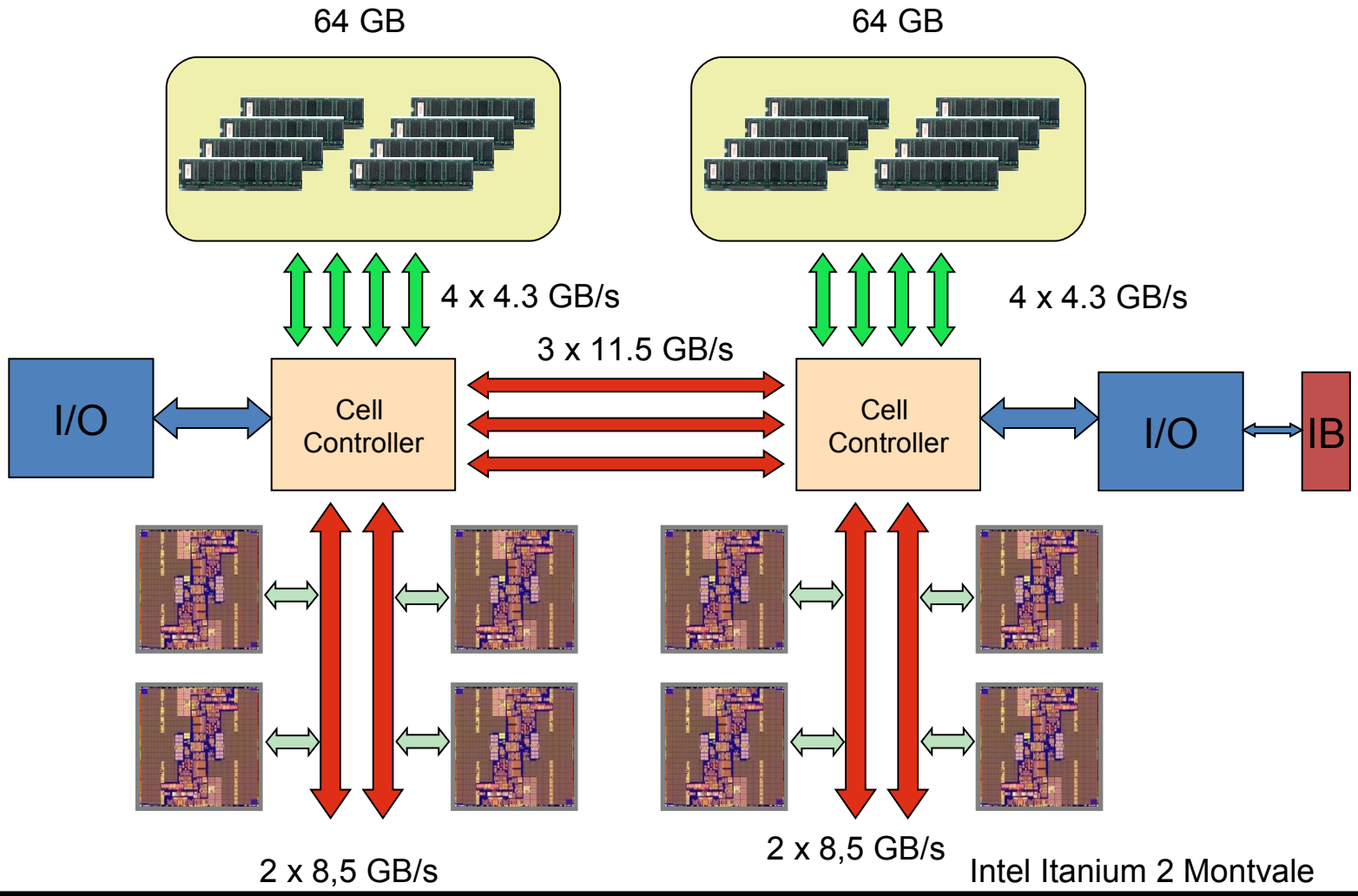
- 48 racks
- Each rack 3 HP Rx 7640 servers + 3 disk cabinets
- 142 servers total, with:
  - 16 cores of Itanium 2 Montvale @ 1.6 GHz , 18 MB cache
  - 128 GB memory
  - 4 or 6 SAS 146 GB SAS disks
  - Suse Linux

HP RX7640: Architecture



Total Peak (Sust) BW per Cell
17.0 GB/s (13.6) CPUs (1.3x)
17.0 GB/s (16.0) Memory (2.1x)
34.6 GB/s (27.3) Crossbar (4.2x)
11.5 GB/s (8.0) I/O (4.4x)

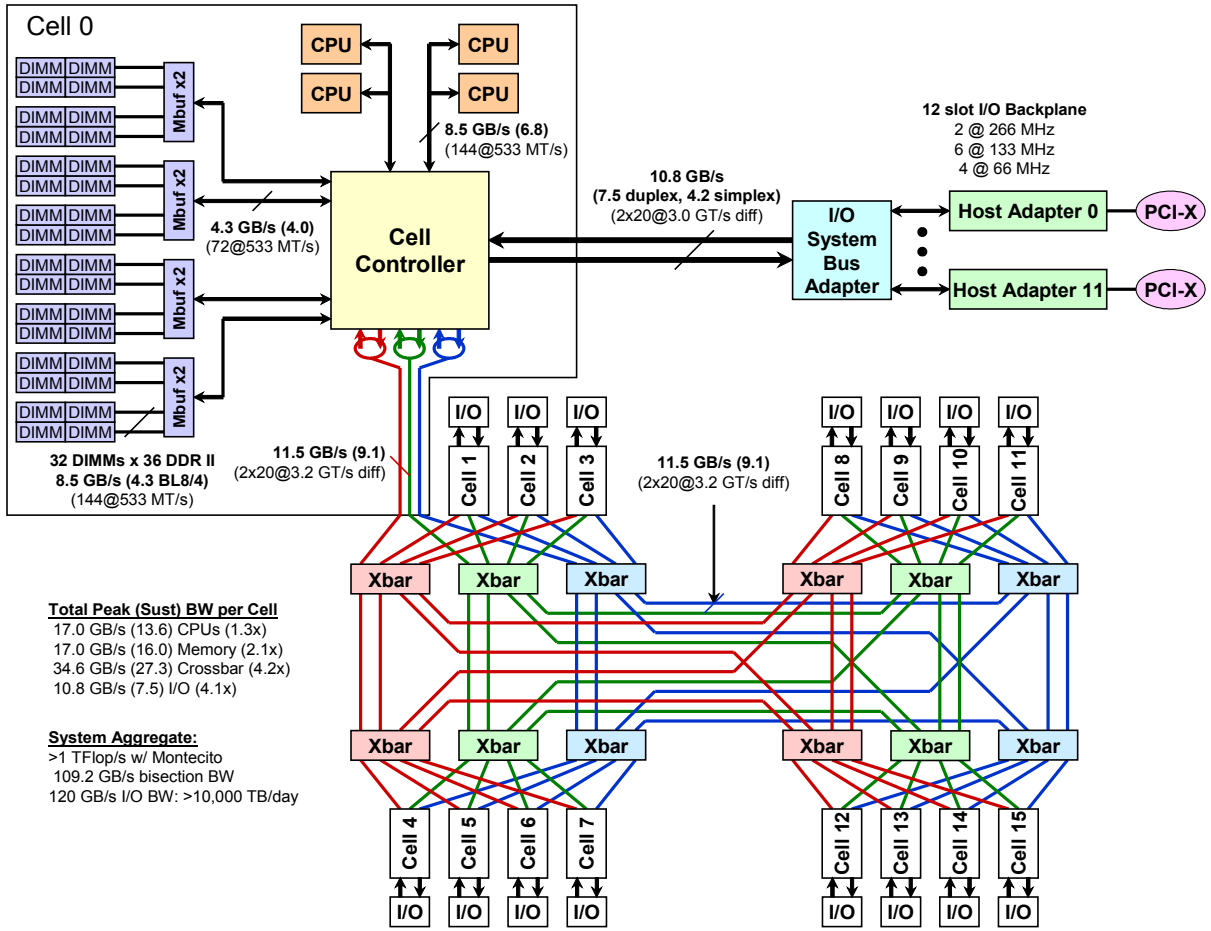
HP RX7640: Architecture (II)



New Superdome

- 64 Processors/128 cores of Itanium 2 Montvale @ 1.6 GHz, 18 MB cache
- 1 TB shared memory
- 128 x 72 GB SAS disks (9.2 TB) for scratch use
- Suse Linux SLES 10

Superdome: Architecture



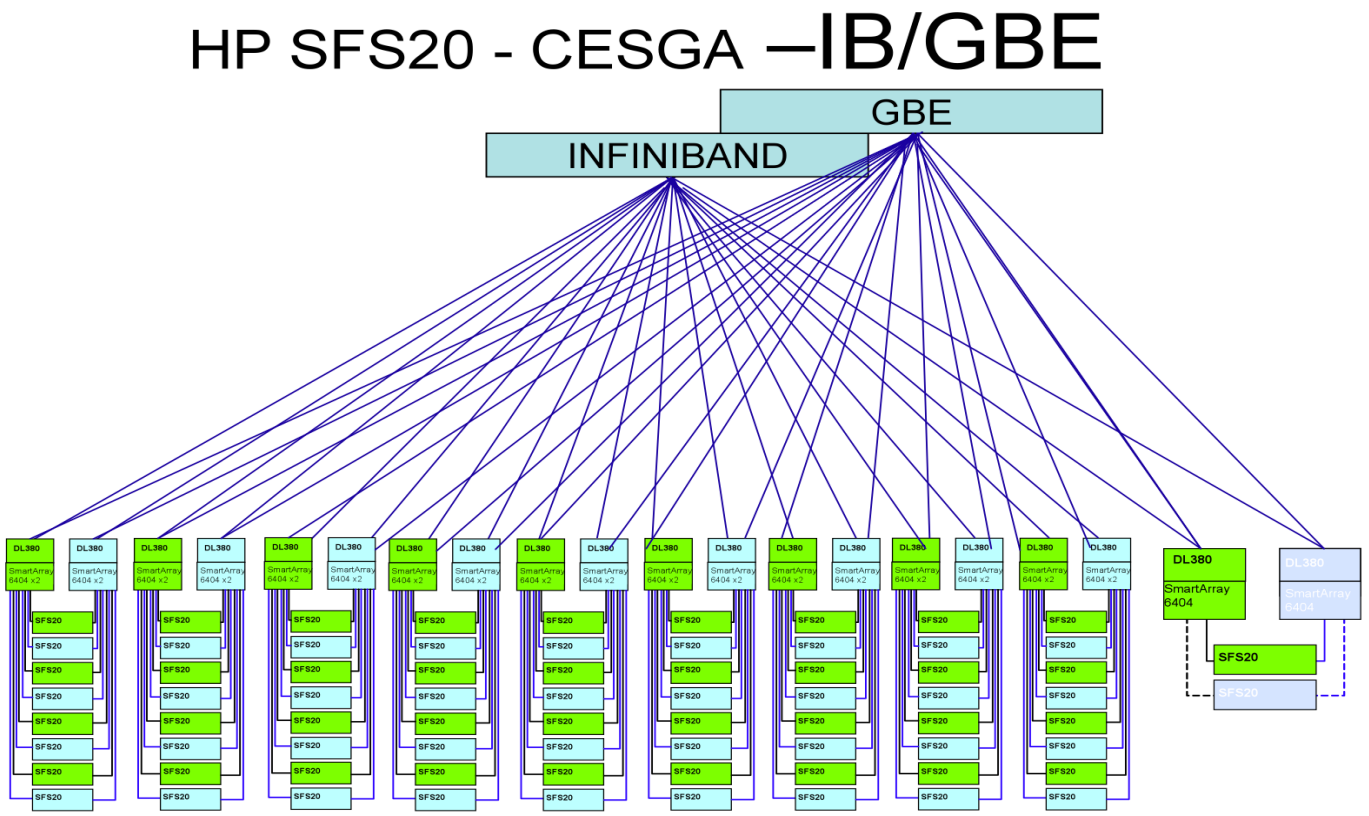
Infiniband Network

- Low latency standard interconnect
- Switch Voltaire ISR 9288
- 4X DDR, 20 (16) Gbps non blocking.
- Low MPI latency (≈7μs real)
- Central director switch connecting computing nodes, SFS and login/storage nodes.

HP-SFS

- Lustre based parallel filesystem.
- 216 Terabytes total. 864 x 250 GB SATA disks
- 18 cells (Proliand DL380 servers) and 72 HP SFS-20 disk cabinets.
- Performance > 10 GB/s read, 6 GB/s write
- Accessed through infiniband
- All the nodes see a regular filesystem (/sfs)

HP-SFS Storage subsystem



Linpack and stream benchmarks

- Linpack : 14.10 TFLOPs with 2528 cores

Stream CN RX7640 (MB/s)

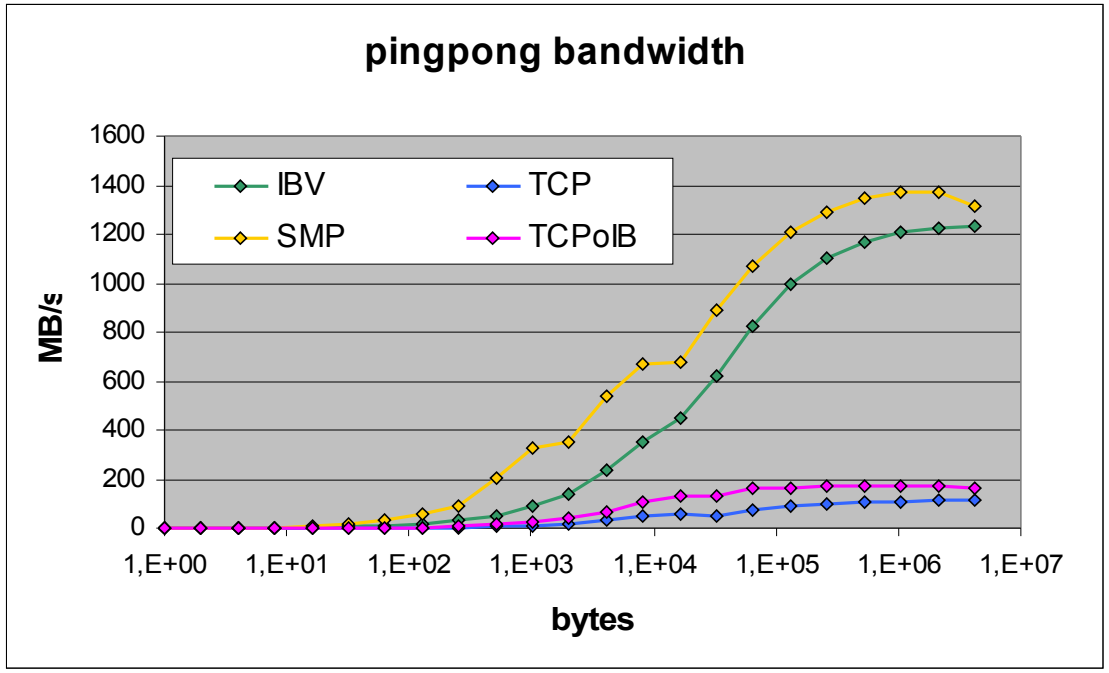
MB/s	16	576	1152
Copy	17836,54	689693,6039	1373937,6619
Scale	17075,78	687337,7336	1369492,2305
Add	18542,2	761919,2969	1519575,3961
Triad	19104,48	763490,9244	1522534,1362

Stream Superdome (MB/s)

Node	Copy	Scale	Add	Triad
Superdome	153224,1003	152744,5704	169354,5667	169613,6162

Pallas MPI Benchmark

	IBV	TCP	SMP	TCPoIB
latency	*5.93 us*	*74.91 us*	*1.93 us*	*50.4 us*



Some results: Iozone on SFS

nodes	File Size	Nb proc	ost	write bdwth KB/s	read bdwth KB/s
A 9nodes ---> cn010-cn018	16G	8 per node	72	**7050130**	66816419
B 9 nodes --> cn010-cn018	16G	8 per node	72	**7076311**	63719567
C 18 nodes --> cn010-cn027	16G	4 per node	72	**7047576**	168841738
D 18 nodes --> cn010-cn027	64G	4 per node	72	**7242699**	**13487185**
E 20 nodes -> cn001 – cn020	1 G	1 per node	72	**6976 MB/s**	**13632 MB/s**
F 70 nodes -> cn001 – cn020	1 G	1 per node	72	**6654 MB/s**	**14603 MB/s**
G 141 nodes -> cn001 – cn142 (- cn141)	1 G	1 per node	72	**6182 MB/s**	**14756 MB/s**

Thanks !

Ignacio López (nlopez@cesga.es)