

Influence of MPI Process Affinity in NUMA Architectures

ESTABLISHED IN 1993 IN SANTIAGO DE COMPOSTELA (SPAIN)



CESGA



SANTIAGO DE COMPOSTELA



To provide high performance computing and communication resources and services to the scientific community of Galicia and to the National Research Council (CSIC), as well as, to institutions and enterprises with R&D activity.

To promote the use of new information and communication technologies applied to research within the scientific community of Galicia.

MADRID, SPAIN, 2009



- **Galician Universities**
- **Galician Regional Government Research Centres**
- **Spanish National Research Council (CSIC) Centres**
- **Other public or private organizations worldwide**
 - ◉ Hospital R&D Departments
 - ◉ Industries R&D Departments
 - ◉ Technological & Research Centres
 - ◉ Other Universities worldwide
 - ◉ Non-profit R&D organizations

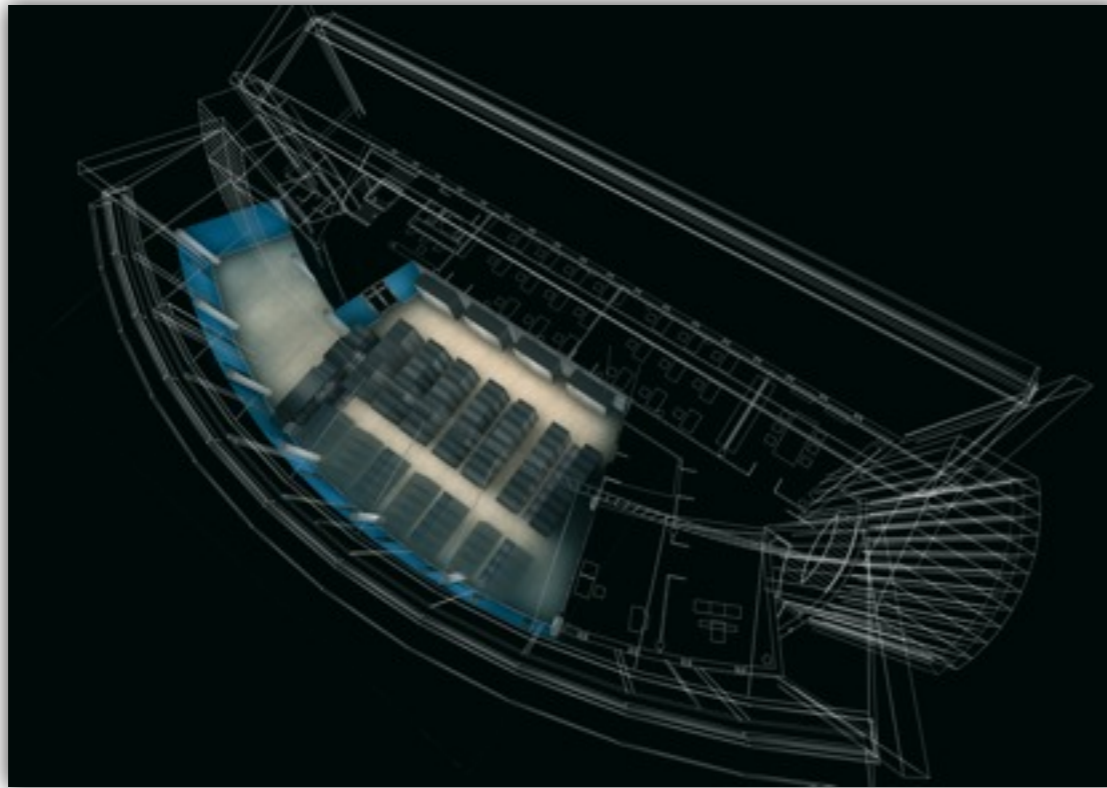
MADRID, SPAIN, 2009



- **HPC, HTC & GRID Computing**
- **User Data Storage**
- **Advanced Communications Network**
- **e-Learning & Collaboration Infrastructures**
- **GIS (Geographical Information Systems)**
- **Transfer to the industry and e-Business Innovation Support**

MADRID, SPAIN, 2009





New HPC Supercomputer 2007

More than: 16,000 GFLOPS 2,580 CPUs 19,000 GB Memory

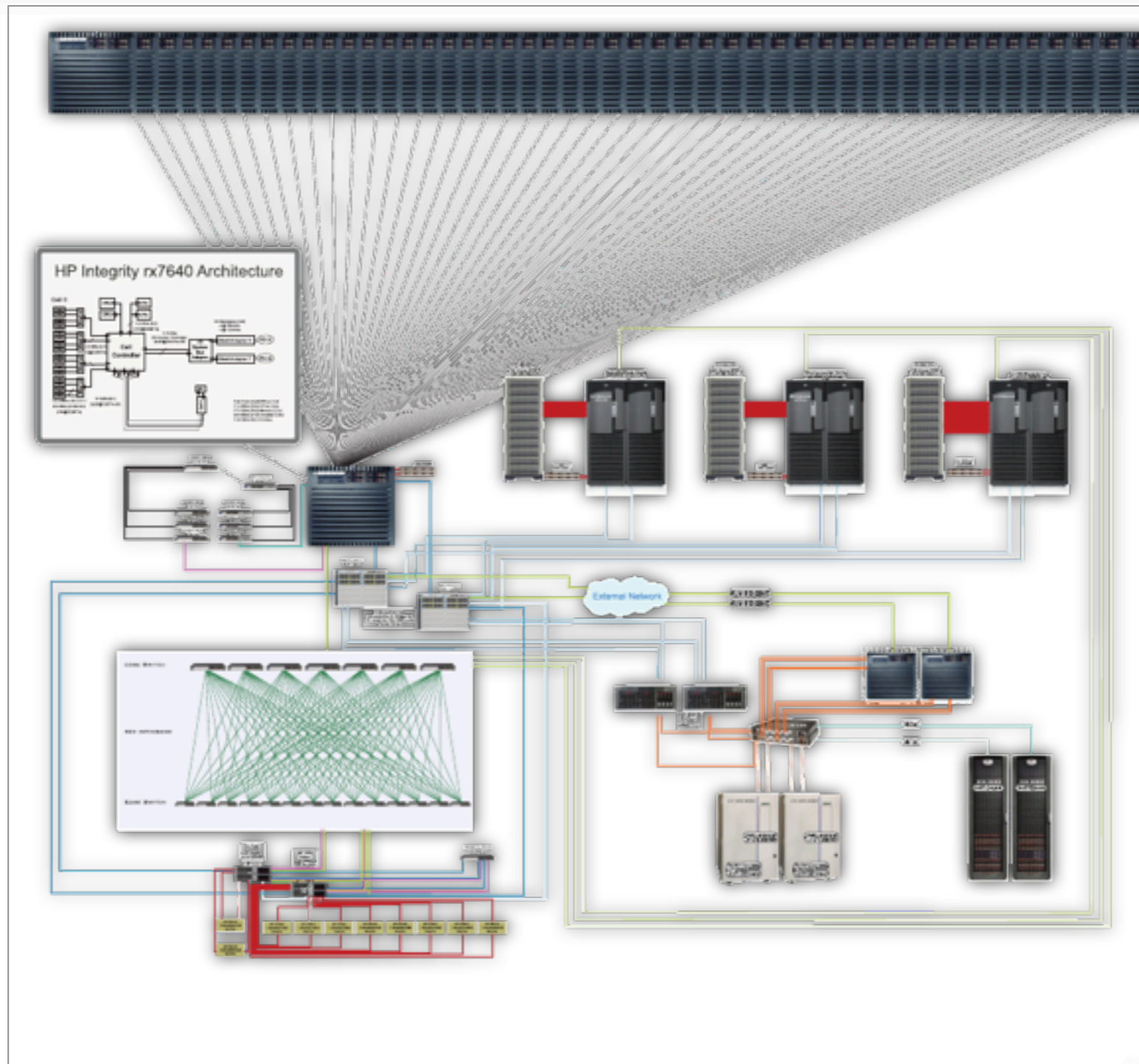
InfiniBand network SuSE Linux

Ranked 100th in the Top500 list of November 2007

MADRID, SPAIN, 2009



Finis Terrae (2007)



Supercomputing Nodes:

147 cc-NUMA Nodes with Itanium CPUs connected through a high performance InfiniBand network (20 Gbps)

- ☒ **1 node: 128 cores, 1024 GB memory**
- ☒ **2 nodes: 64 cores, 128+256 GB memory**
- ☒ **142 nodes: 16 cores, 128 GB memory**
- ☒ **2 testing nodes: 4 cores, 4 GB memory**

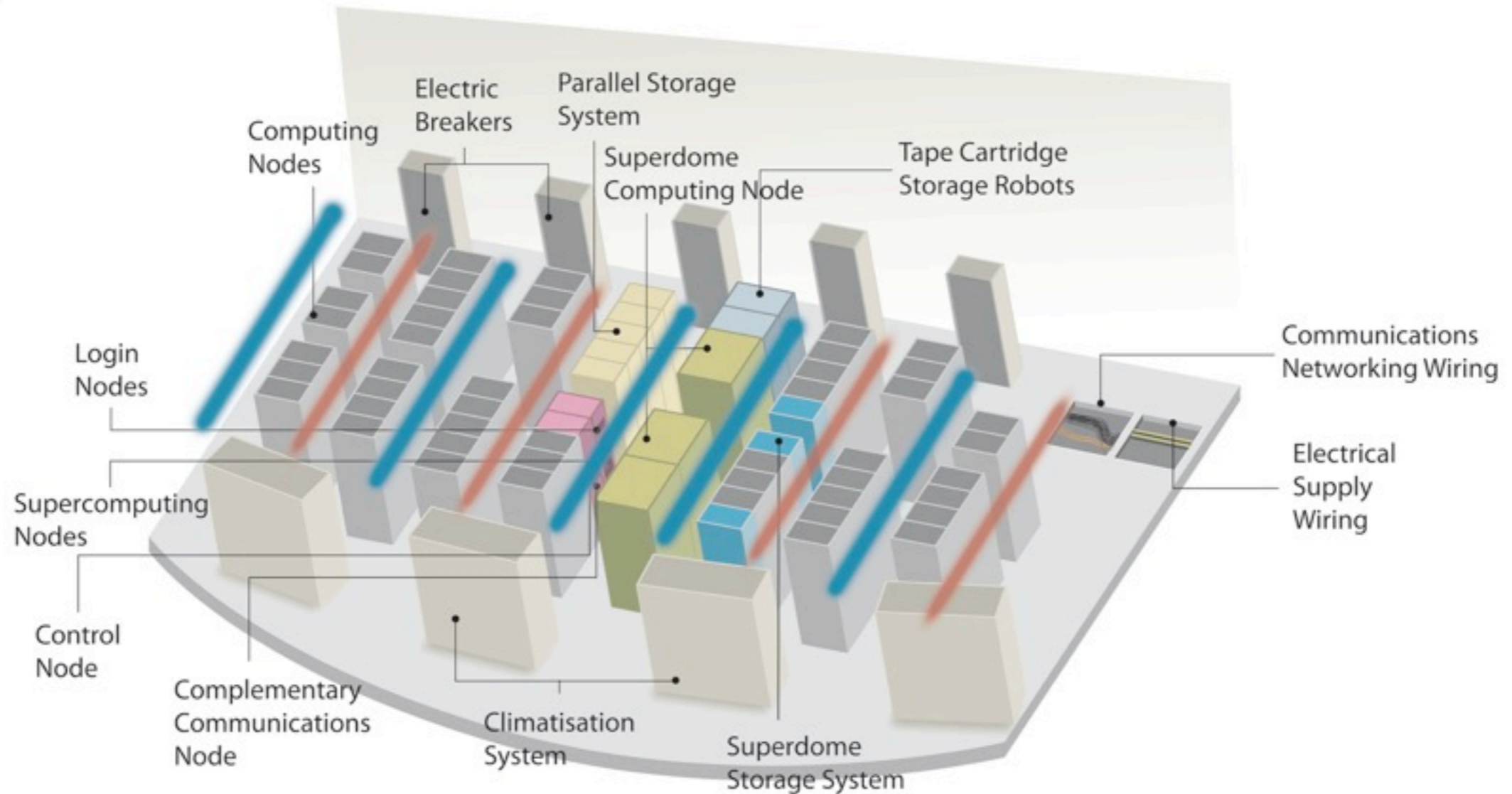
MADRID, SPAIN, 2009






Parallel Filesystem HP-SFS:

- ☑ 20 Nodes (2x Dual-core Intel Xeon 5160 CPUs)
- ☑ 864 Hard Disks
- ☑ 210 TB
- ☑ Based on Lustre
- ☑ Accessed through the InfiniBand network

MADRID, SPAIN, 2009

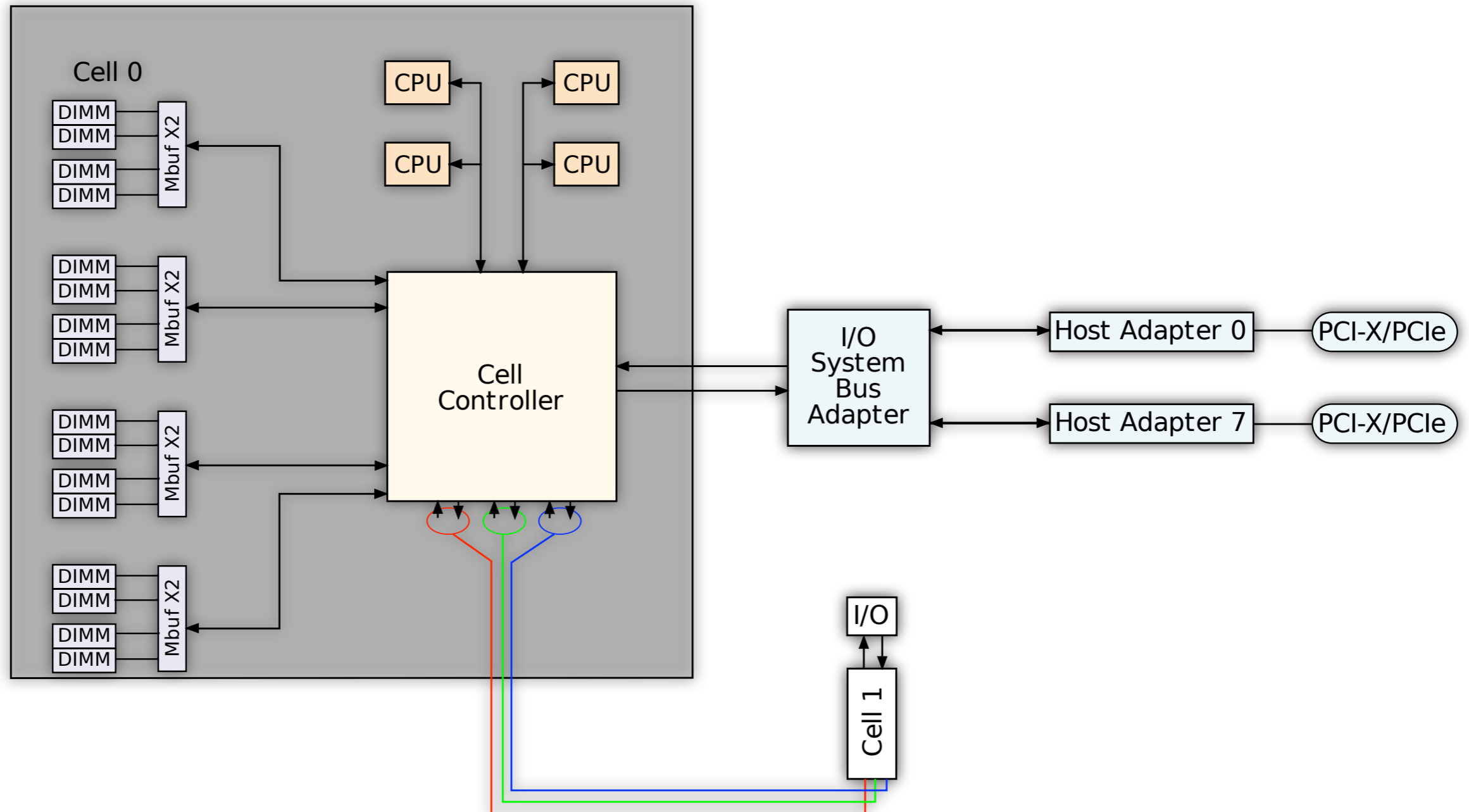


Technical Specs.

Surface Area: 140 m ²	Storage:	Memory:	2.528 Processing Cores	Node Interconnect INFINIBAND
Weight:			142 nodes, each with 16 cores & 128 GB memory	4x DDR at 20 Gbps
 35.000 Kg	2.200.000 GB on tape	19.670 GB	1 node with 128 cores & 1.024 GB memory	85 Km of interconnect cable
	390.000 GB on disk		1 node with 128 cores & 384 GB memory	Open Software: Linux, Lustre, Globus...

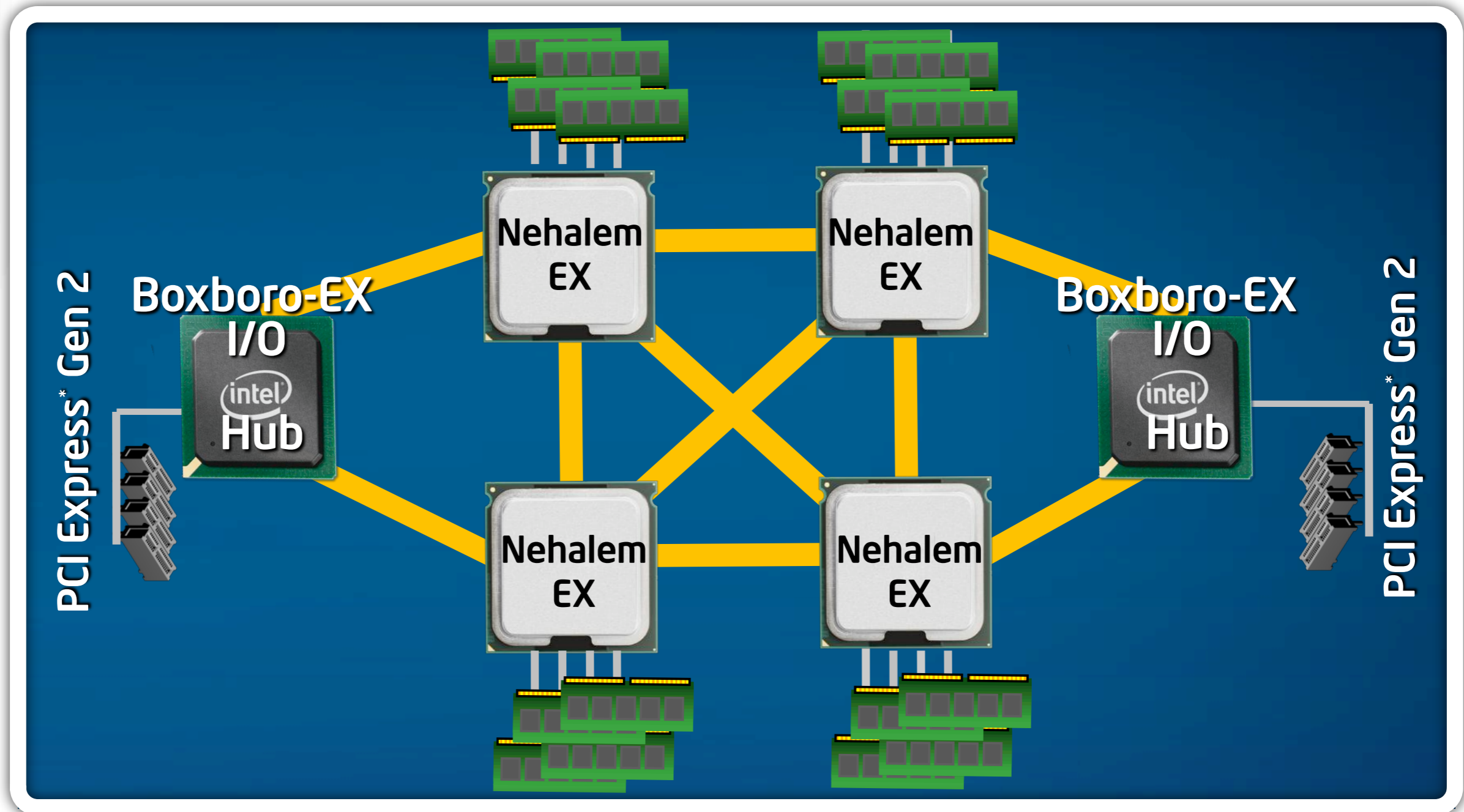
MADRID, SPAIN, 2009

Finis Terrae Node Architecture



MADRID, SPAIN, 2009

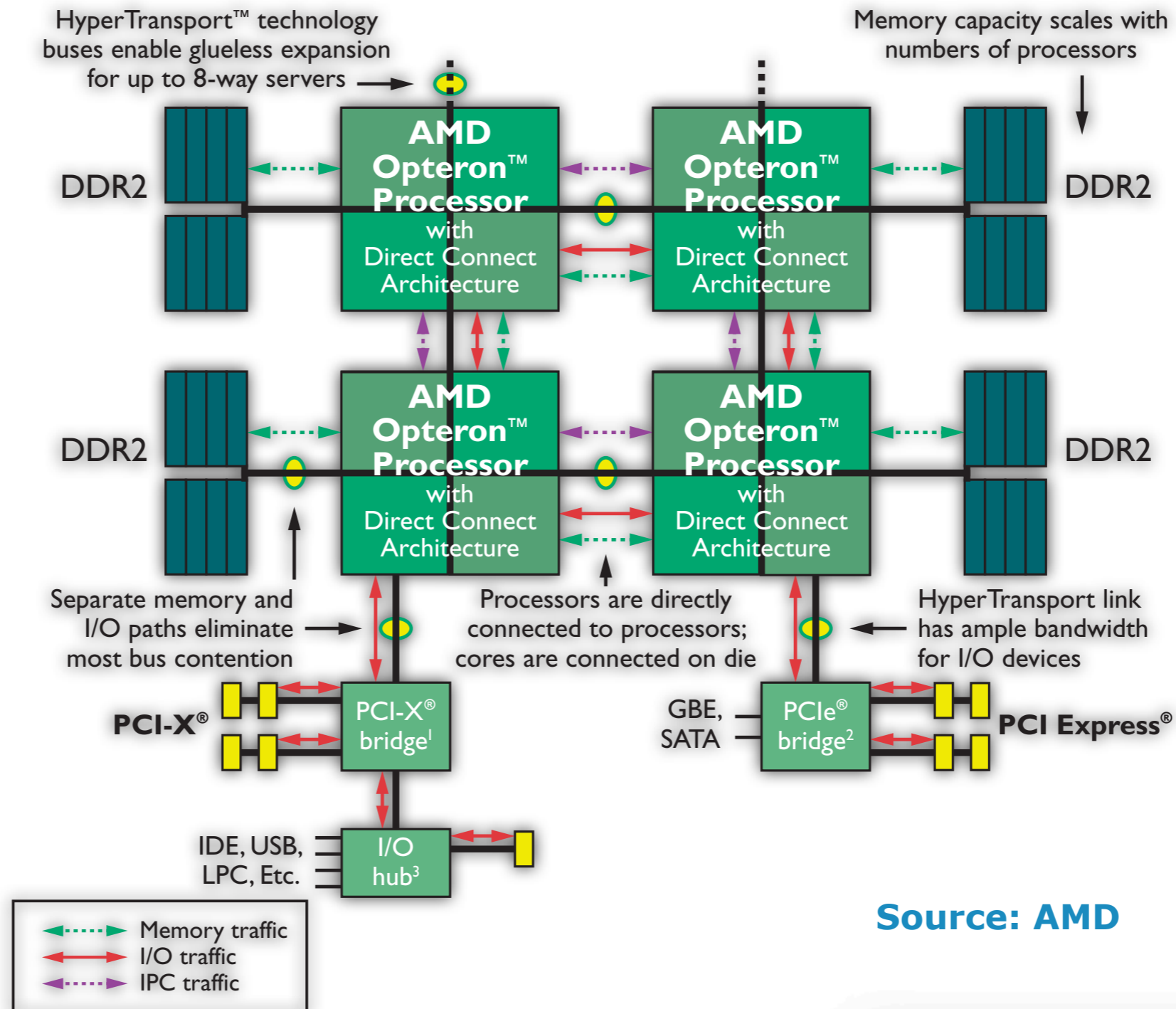
Other examples: Nehalem-EX



Source: Intel

MADRID, SPAIN, 2009

Other examples: Opteron



Source: AMD

MADRID, SPAIN, 2009

HP Integrity RX 7640:

- **16 Cores**
- **2 Cells**
- **1 InfiniBand HCA**

MADRID, SPAIN, 2009

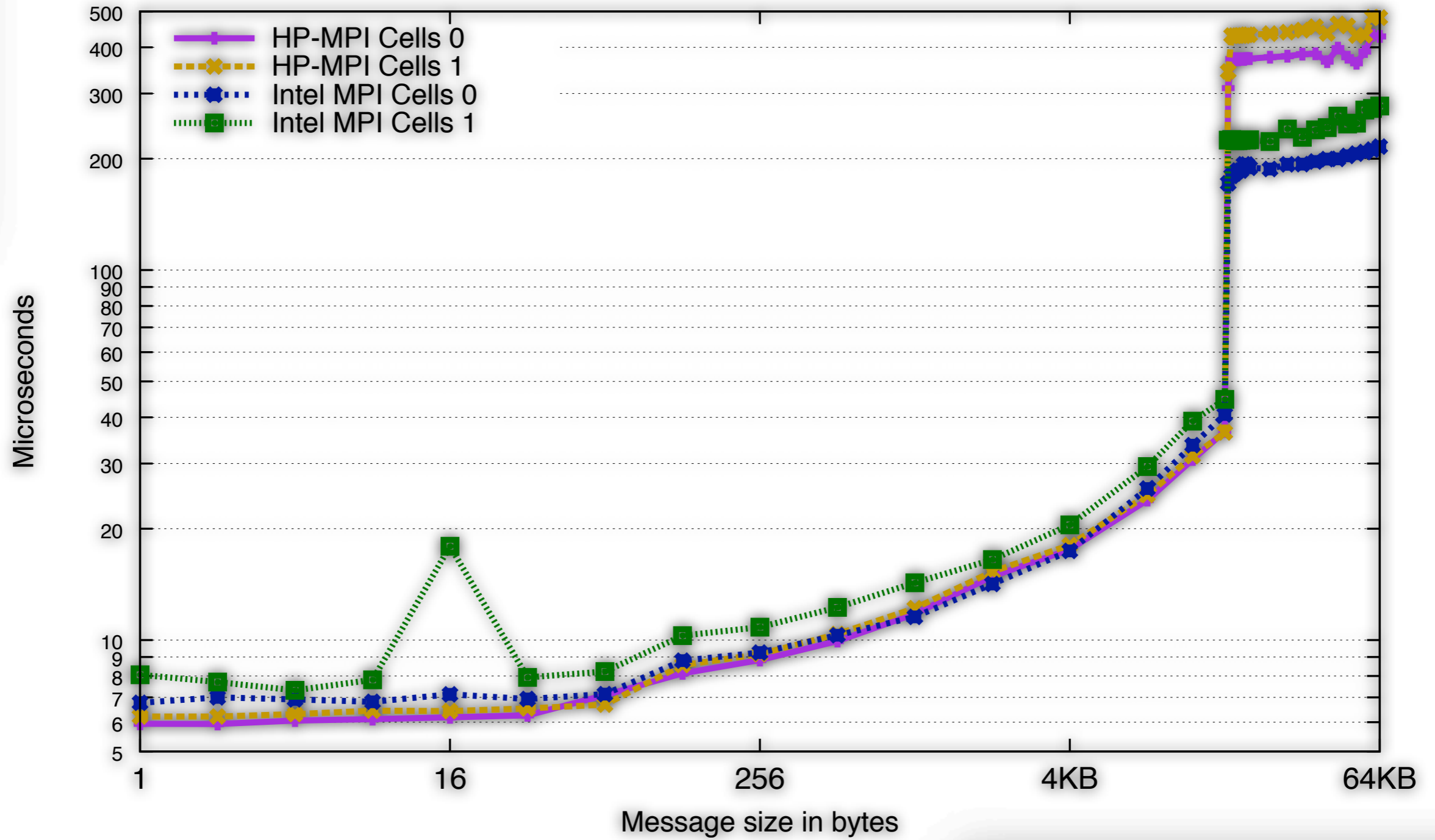


HP Integrity RX 7640:

- **16 Cores**
- **2 Cells**
- **1 InfiniBand HCA**
 - ⊙ **It is not only a bottleneck (it is placed on one cell, what about the other one?)**
 - ⊙ **In Nehalem/Opteron machines the problem is quite similar (there is no cells, but they are NUMA machines with I/O interfaces associated to one/two processors)**

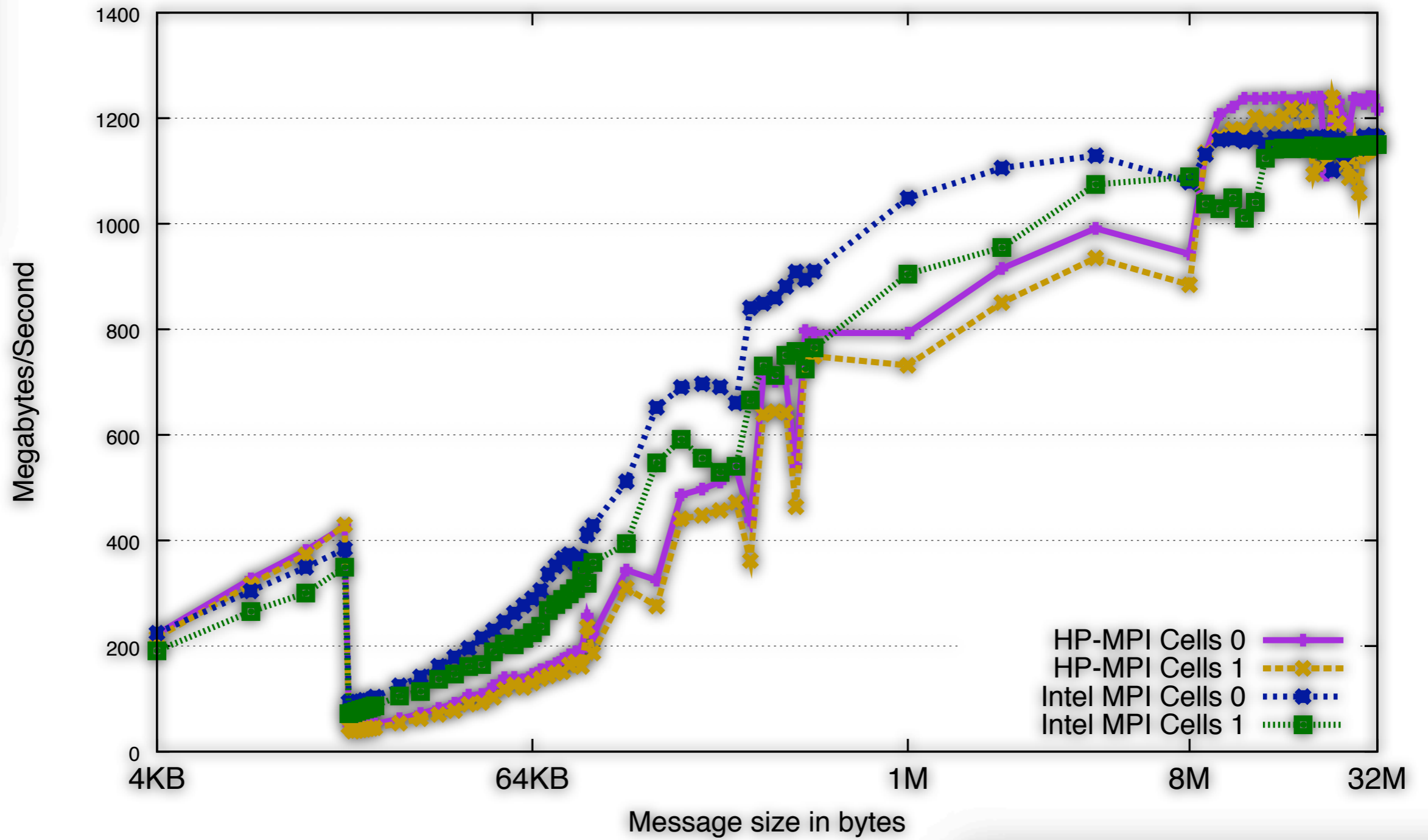
MADRID, SPAIN, 2009

PingPong Latency Between Nodes

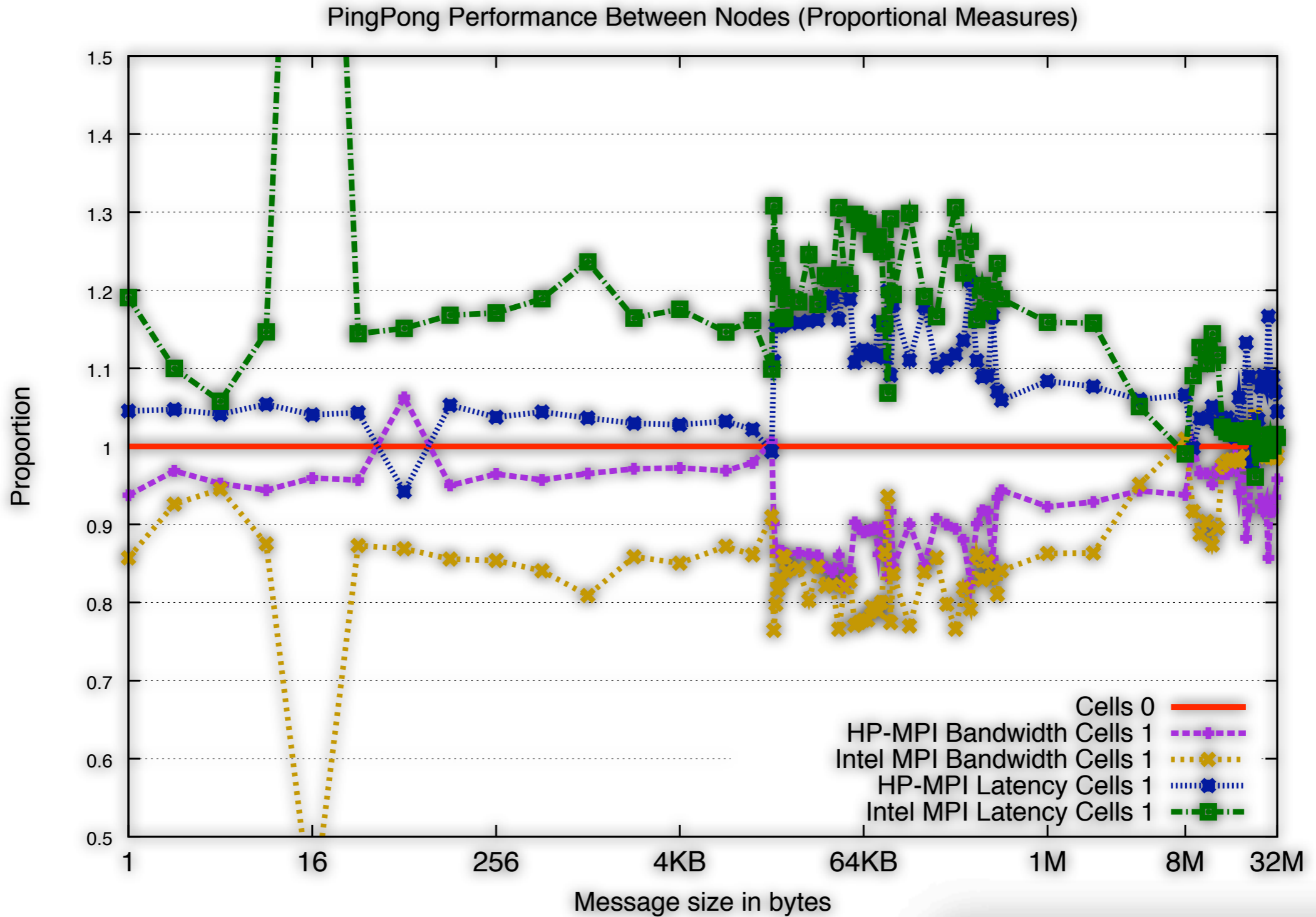


MADRID, SPAIN, 2009

PingPong Bandwidth Between Nodes

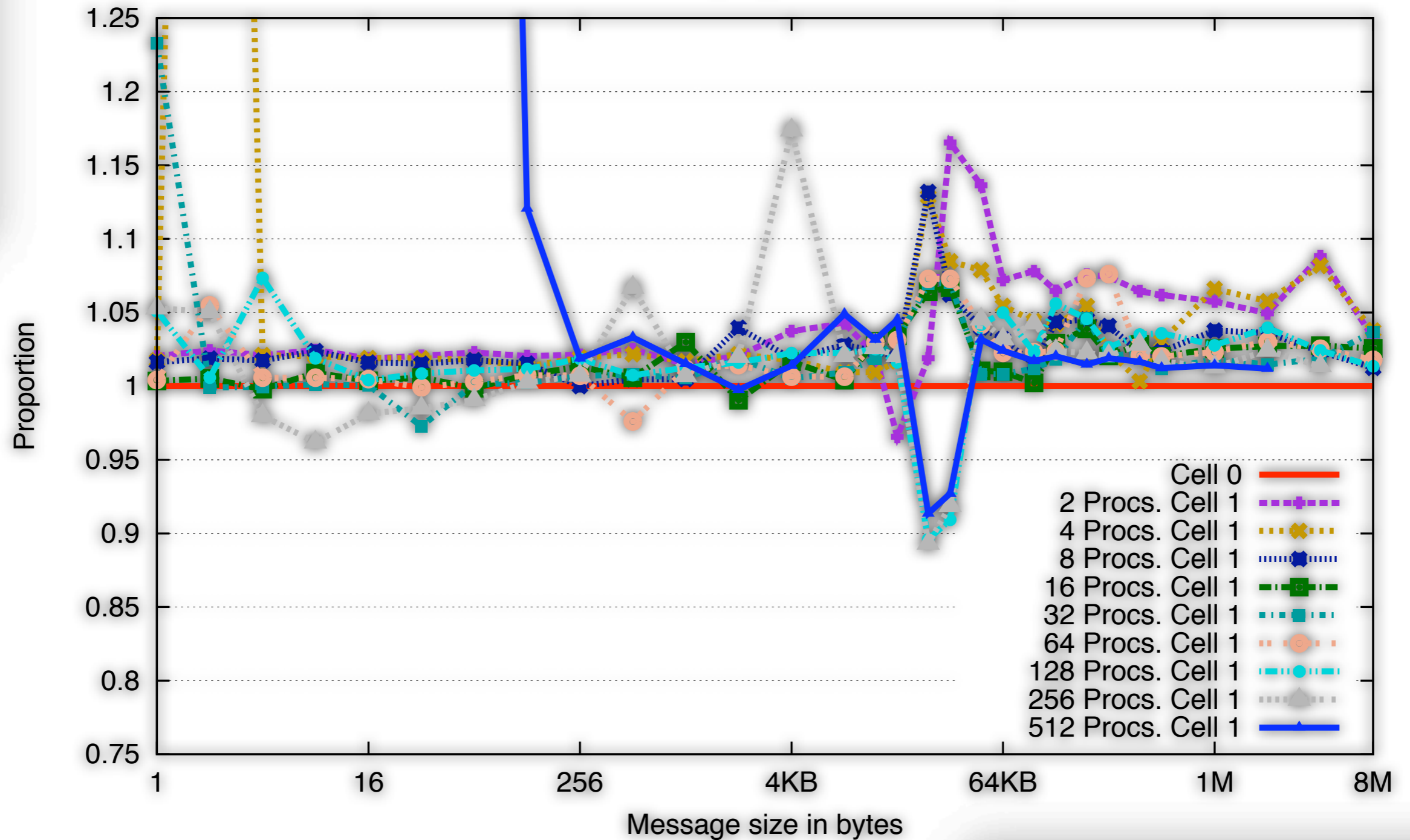


MADRID, SPAIN, 2009



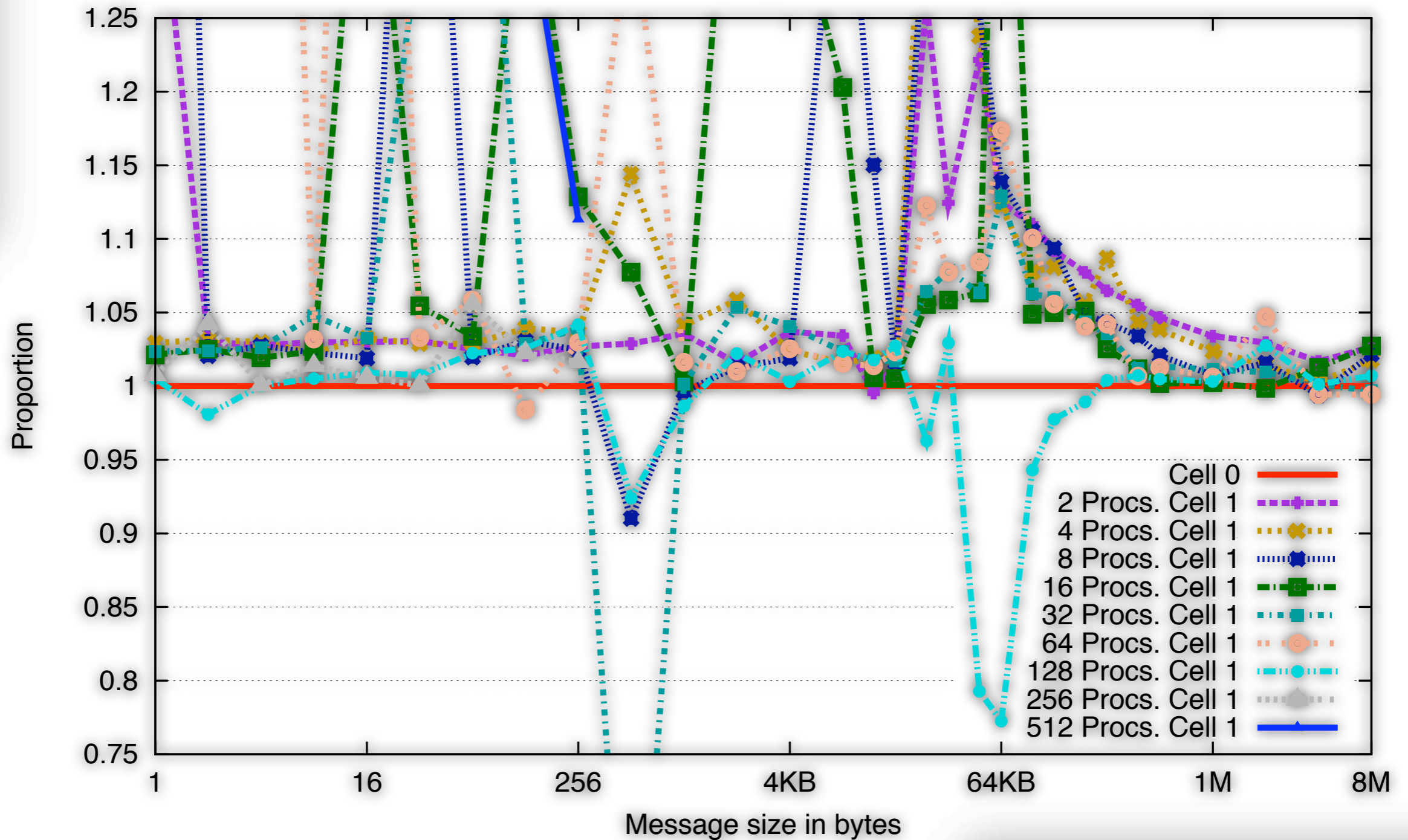
MADRID, SPAIN, 2009

Alltoall HP-MPI (Proportional Measures)



MADRID, SPAIN, 2009

Alltoall Intel MPI (Proportional Measures)



MADRID, SPAIN, 2009

Point to point:

- Roughly 20% in average
- Up to 30%

Collectives (All to all):

- Roughly 3% in average
- Up to 15% (With HP-MPI. With Intel MPI unstable -but consistent- measures)

MADRID, SPAIN, 2009

Systems made of:

- 64? (RX-like) - 512? (Superdome-like) cores
- 2? - 16? cells
- 1 InfiniBand HCA?
 - ⊙ How would it perform in a MPI_Alltoall operation?
 - ⊙ The scenario is quite similar in the x86-64 world.

MADRID, SPAIN, 2009



Current setups:

- **Noticiable differences (not HUGE, but noticeable)**
- **Collective operations are less affected**
 - ⦿ **The time spended due to the algorithm complexity and intranode communications hides the problem**
- **Users can benefit from carefully planned affinity (specially SFS users)**

MADRID, SPAIN, 2009



Future setups:

- **The problem might become:**
 - ⊙ **Bigger (more cores accessing non-local HCA or an HCA accessing more non-local data -RDMA-)**
 - ⊙ **Widespread (all x86-64 systems will be NUMA)**

MADRID, SPAIN, 2009



Solution:

- **Add hardware (HCAs)**
 - ⊙ **But it is expensive. Does it compensate?**
 - ⊙ **And we/runtime/application can not choose the HCA to be used**
 - **What do we do now?**

MADRID, SPAIN, 2009

Solution:

- **Add hardware (HCAs)**
 - ⊙ **But it is expensive. Does it compensate?**
 - ⊙ **And we/runtime/application can not choose the HCA to be used**
 - **What do we do now?**
 - **It is necessary to develop some mechanisms to allow the software stack (driver "routing"?) to choose the "nearest" HCA**

MADRID, SPAIN, 2009



Thank you!

dalvarez@cesga.es

MADRID, SPAIN, 2009

