

Workshop CCR INFN GRID – 16-20 Maggio 2011, Isola d'Elba



**Infrastrutture per il calcolo ad alte prestazioni: dall'elaborazione
allo stoccaggio dei dati**

Simone Tinti

HPC Team leader

simone.tinti@e4company.com



Sommario

- **CPU**
 - performance, TCO
 - RAM bandwidth
- **Storage**
 - soluzioni
 - performance, capacity
- **Interconnessioni**
 - 10GbE Vs IB
 - criticita'
- **GPU**
 - performance



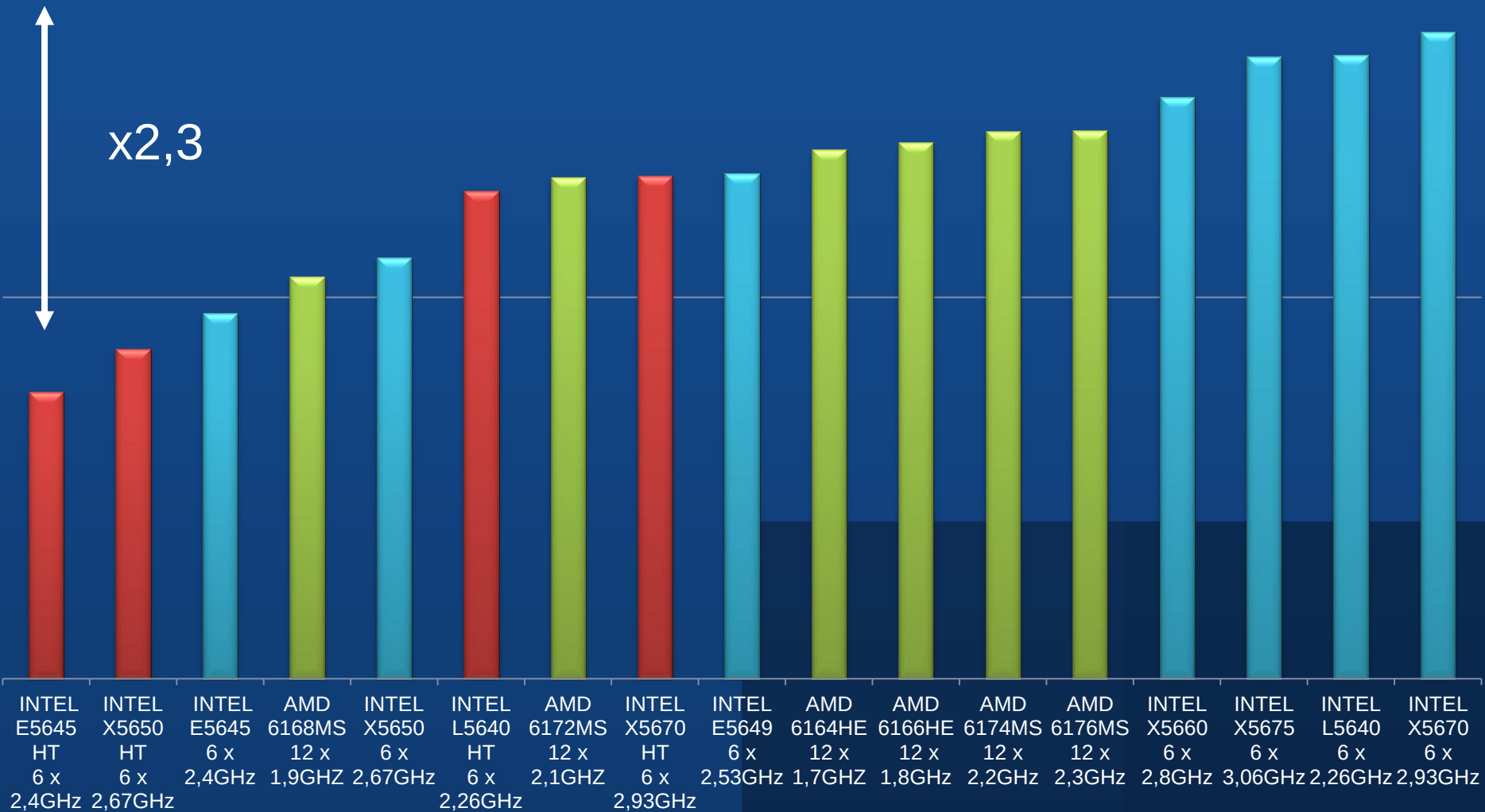
INTEL HT ON

INTEL HT OFF

AMD

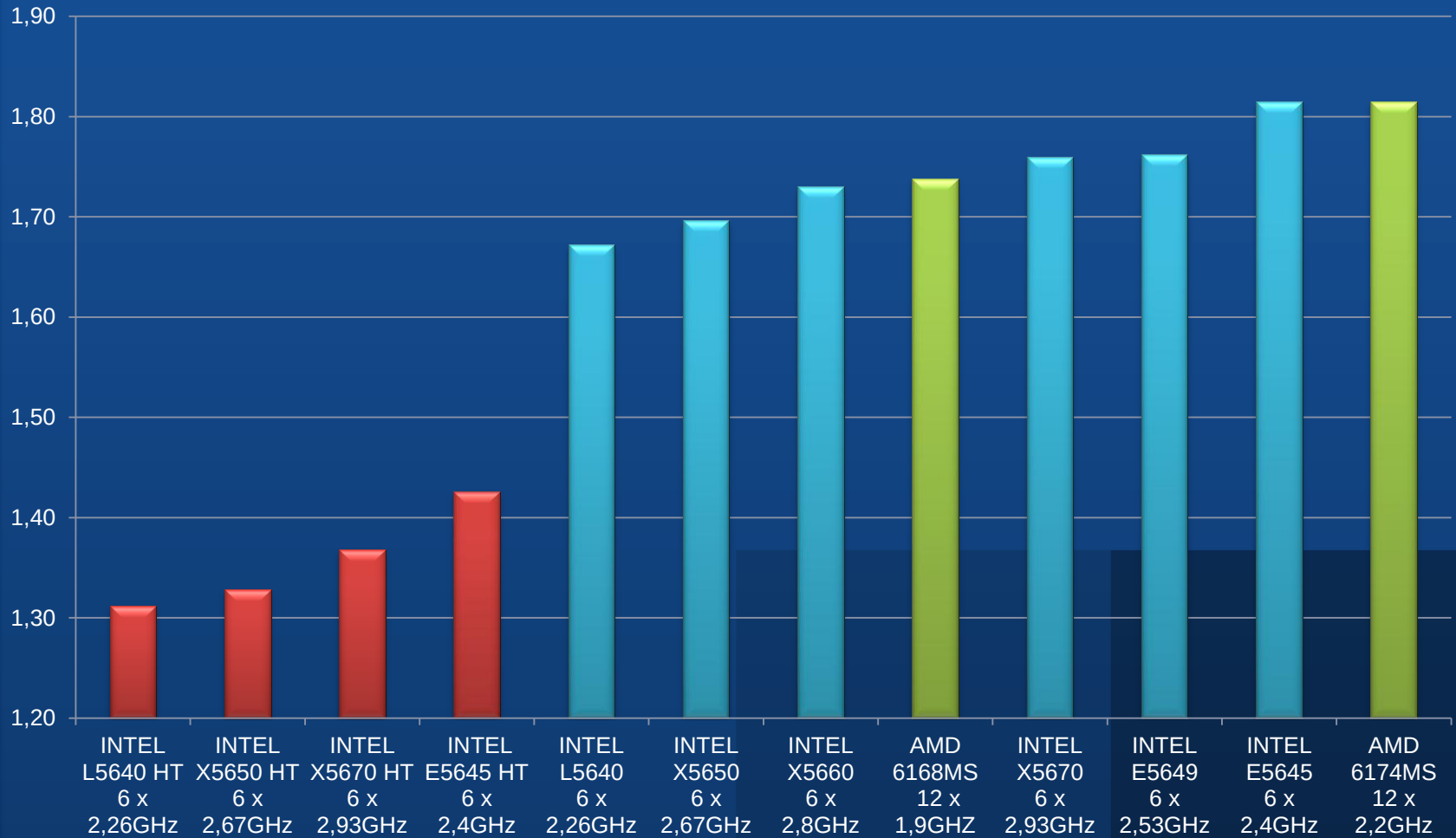
euro/ HEP SPEC2006

x2,3





W / HEP SPEC 2006



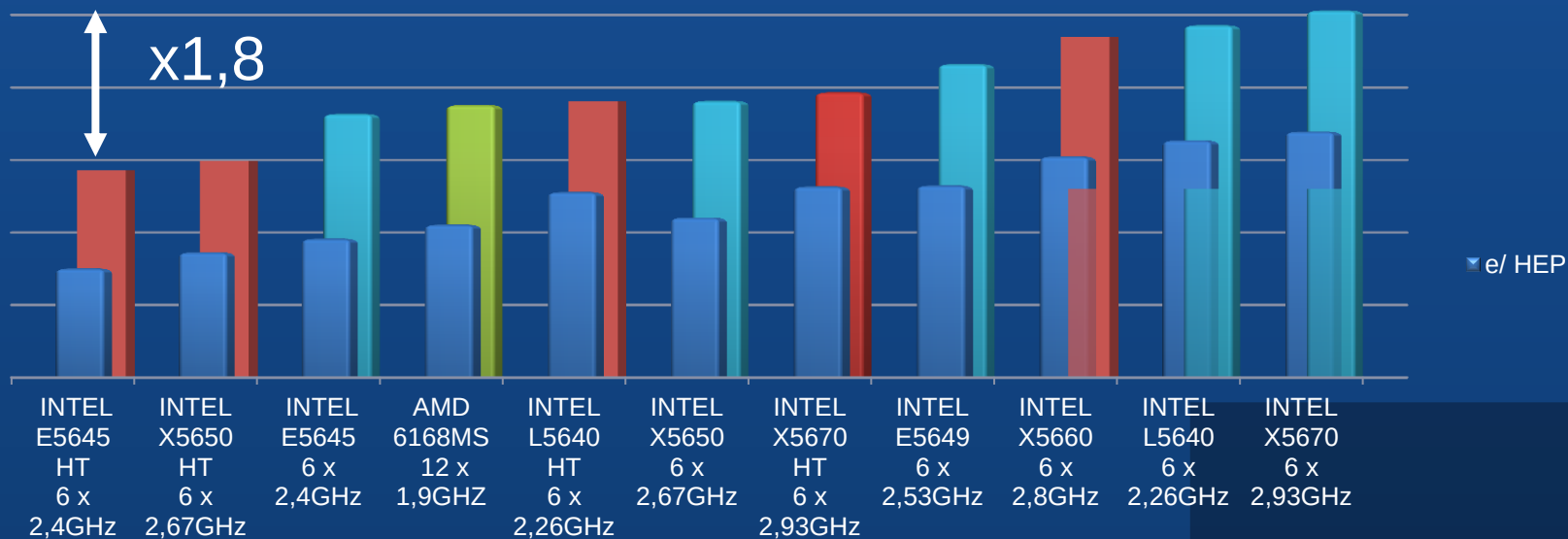
INTEL HT ON

INTEL HT OFF

AMD



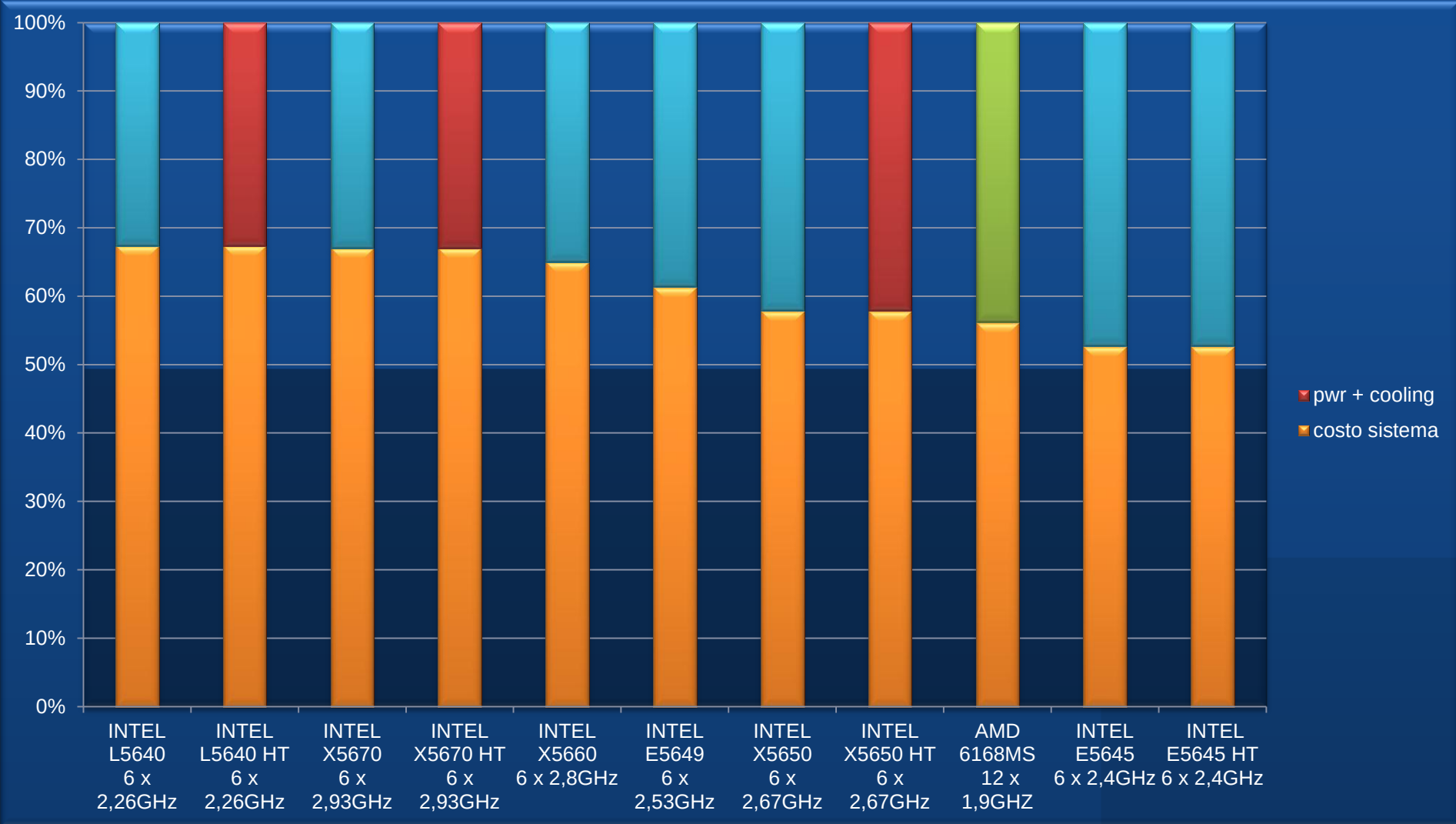
TCO / HEP SPEC 2006



INTEL HT ON

INTEL HT OFF

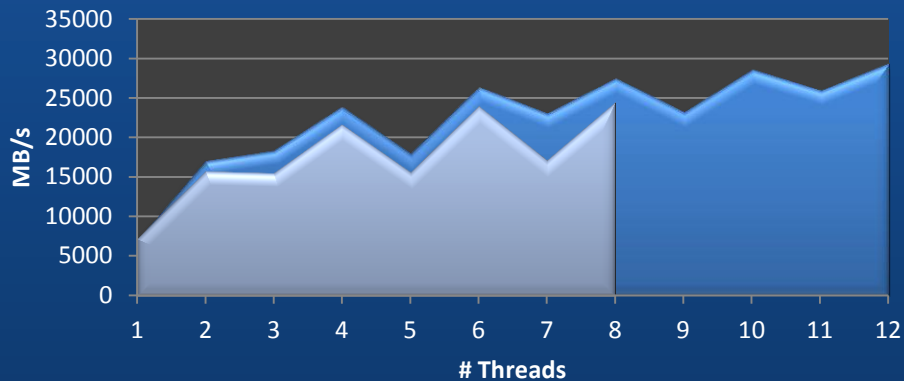
AMD



RAM Bandwidth - STREAM

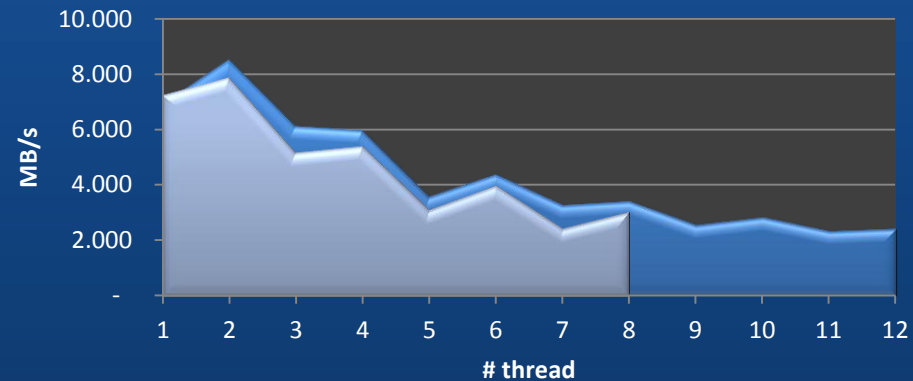
Westmere: bandwidth

INTEL 6 core X5680 INTEL 4 core E5620



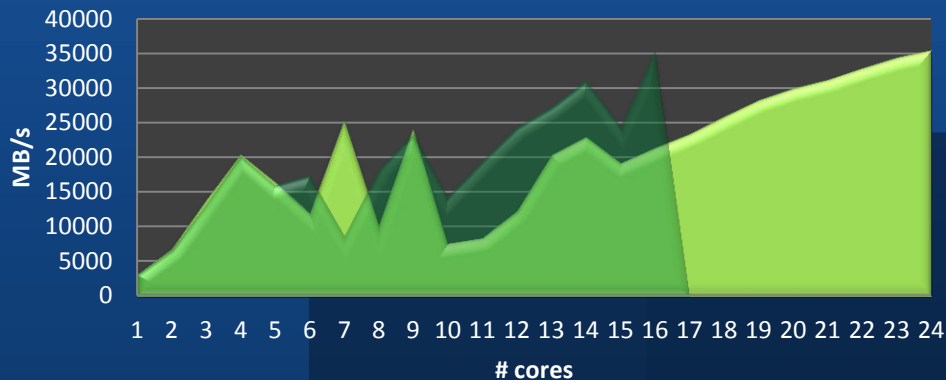
Westmere: Bandwidth / thread

2x INTEL 6 core X5680 2x INTEL 4 core E5620



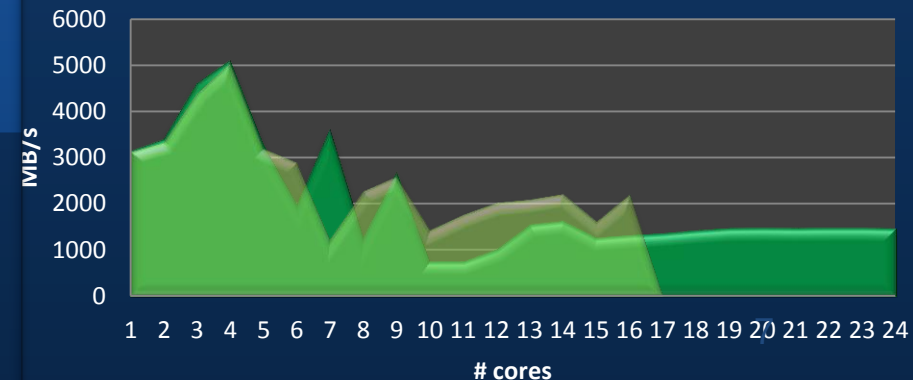
Magny cours: bandwidth

AMD 12 core 6174 AMD 8 core 6128



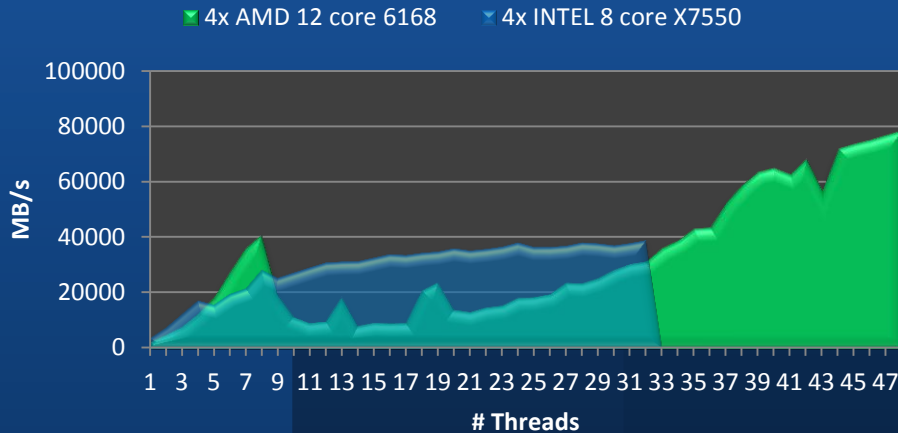
Magny cours: bandwidth / thread

#core AMD 12 core 6174 AMD 8 core 6128



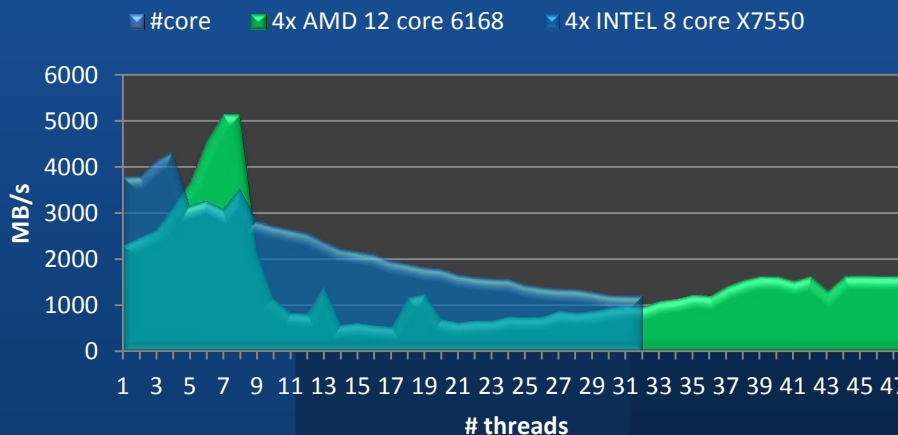


Bandwidth MP



RAM Bandwidth sistemi MP

sistemi MP: bandwidth / #threads





CPU – lesson learned

- Multithreading?
- Low voltage: una scelta etica, purtroppo antieconomica

Storage



E7412 disk server

HD: SATA 2TB

Host Interface: 10GbE <-> clients

DDN S2A900

SAS + SATA 2TB

host Interface: FC 4/8 <-> 10GbE file server <-> clients



SAS JBOD + Nexenta

HD: SATA 2TB

Host Interface: 10GbE <-> clients



E6500 storage

HD: SATA 3TB

Host Interface: FC 8 <-> 10GbE file server <-> clients

PAS11

SATA 3TB

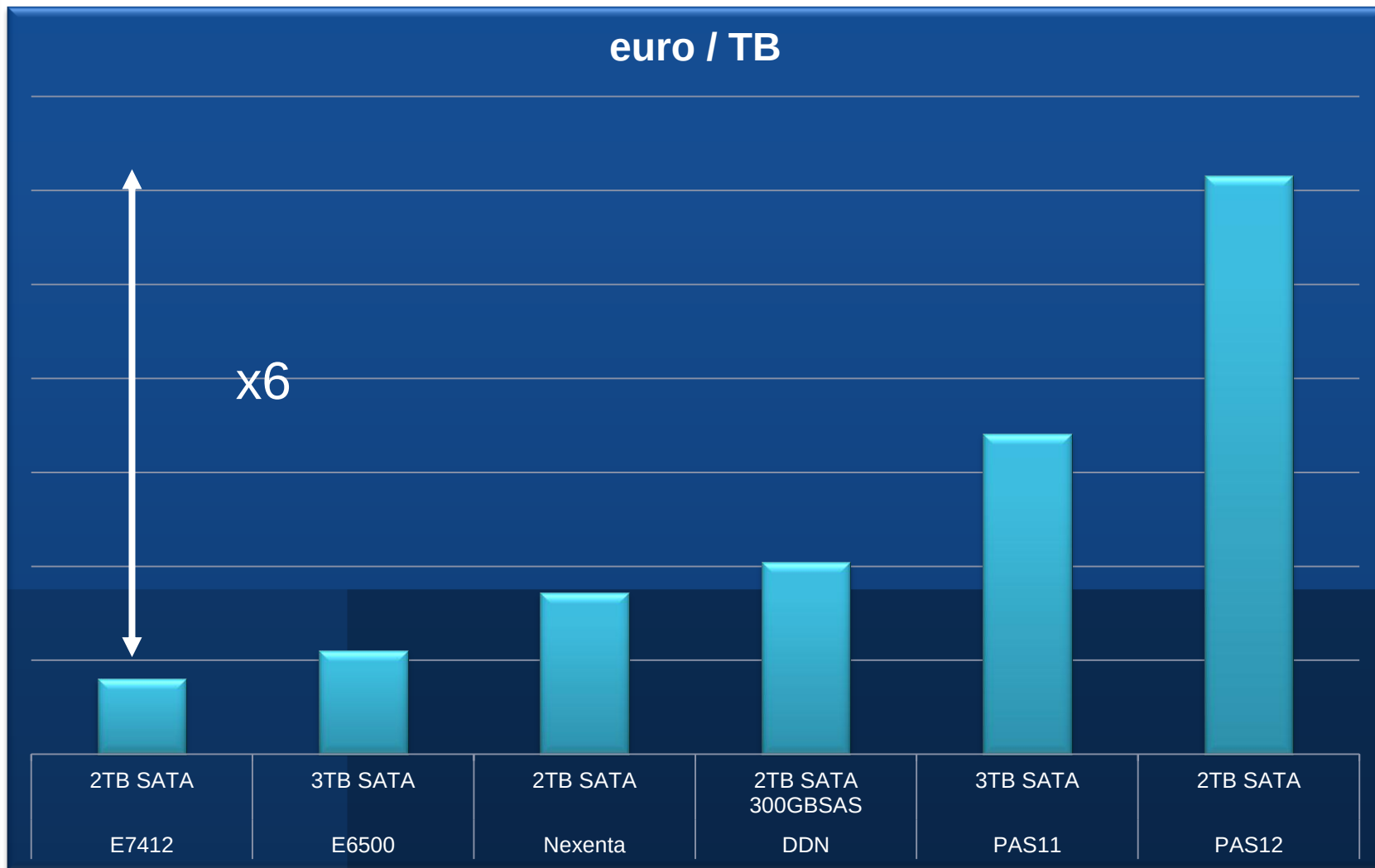
host Interface: 10GbE <-> IB QDR file server <-> clients

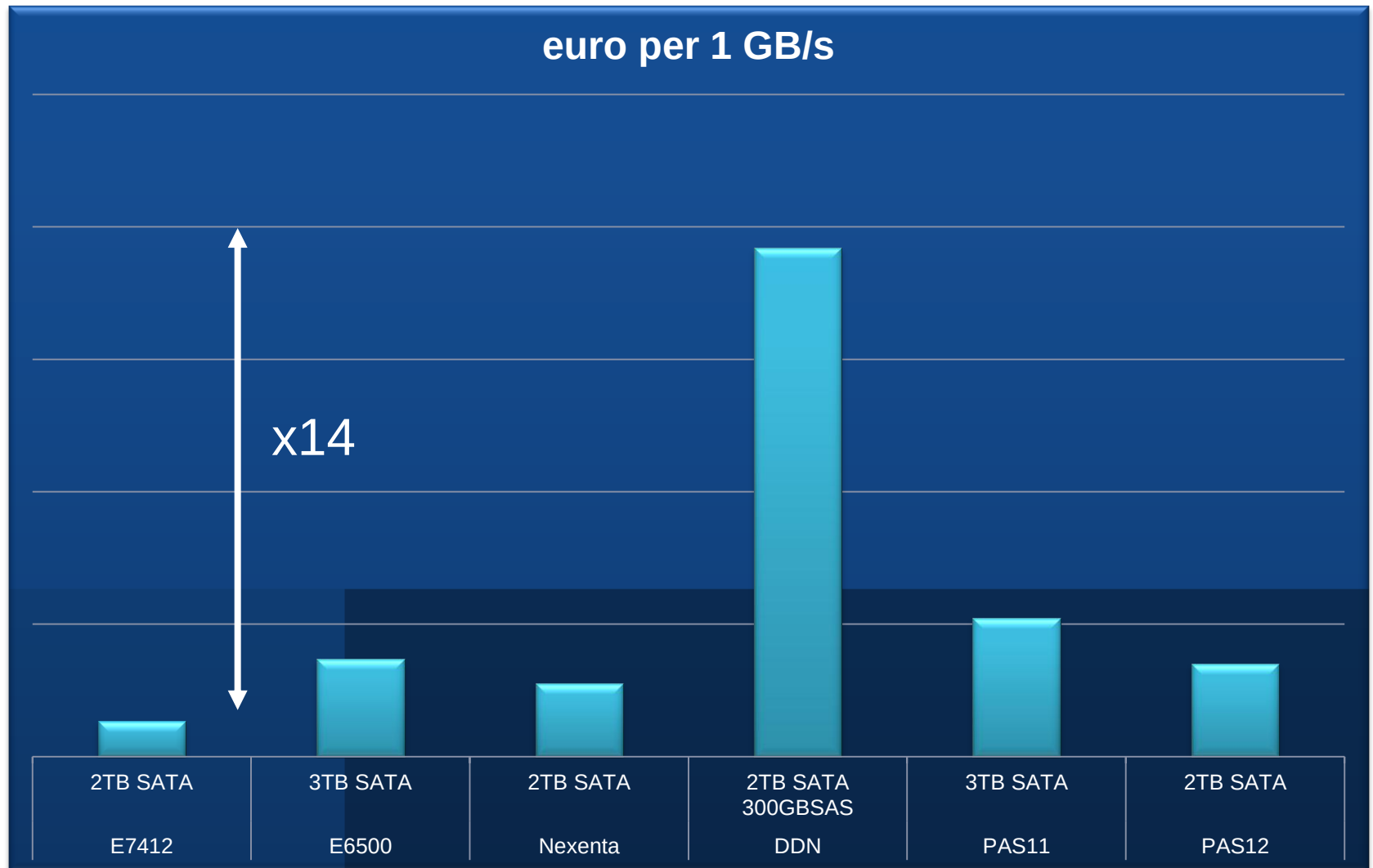
PAS12

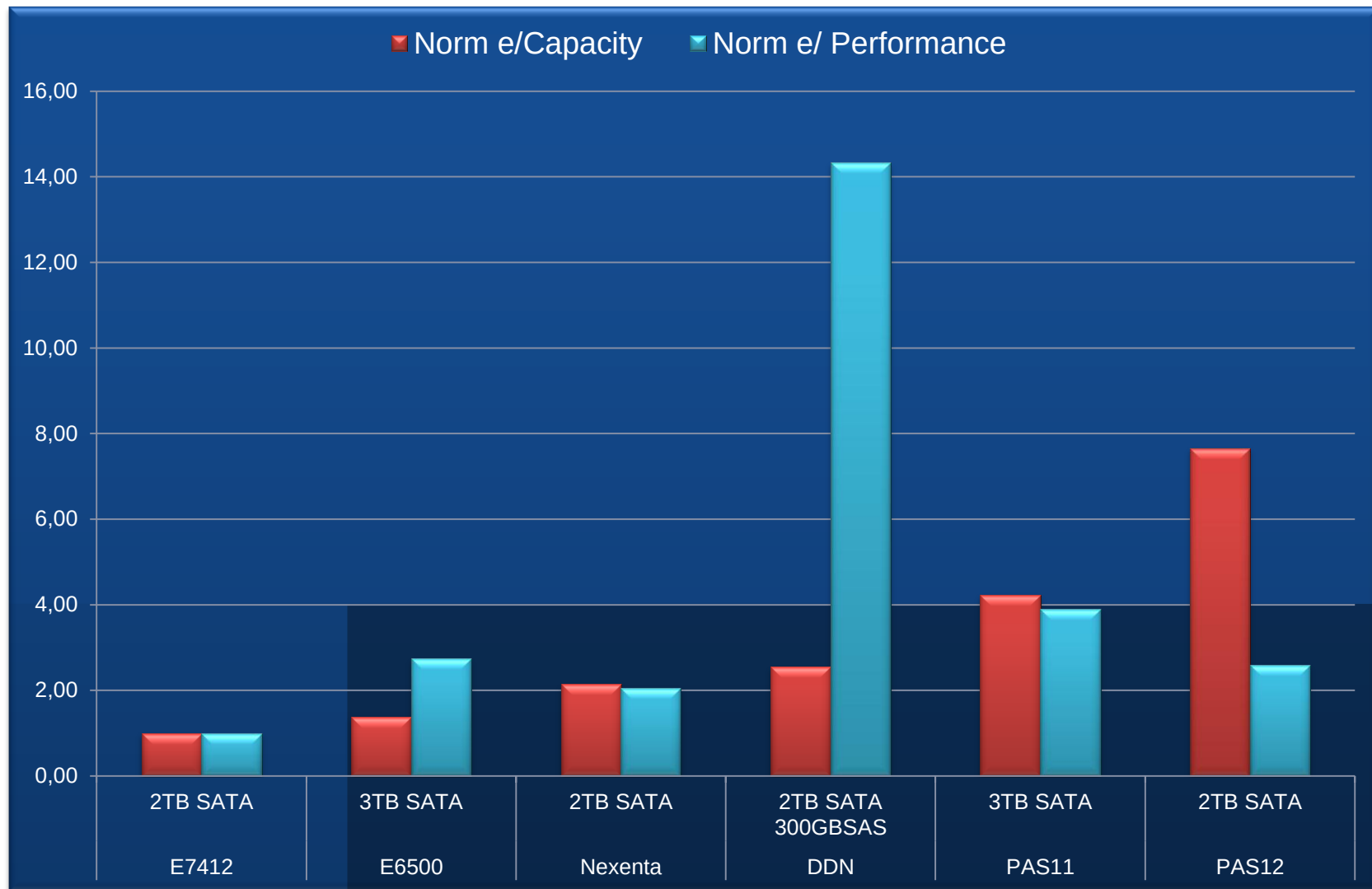
SATA 3TB

host Interface: 10+10GbE <-> IB QDR file server <-> clients



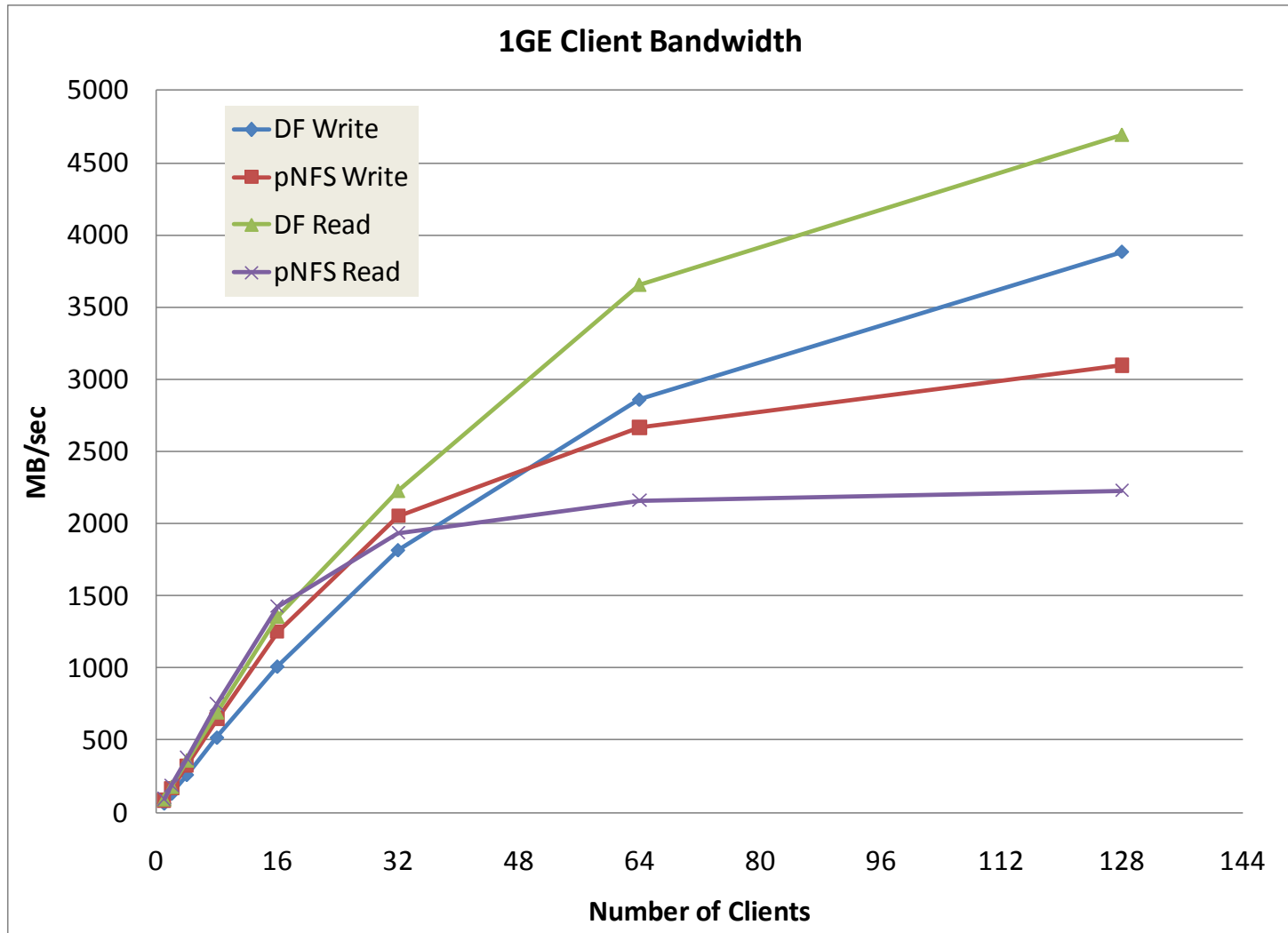








pNFS



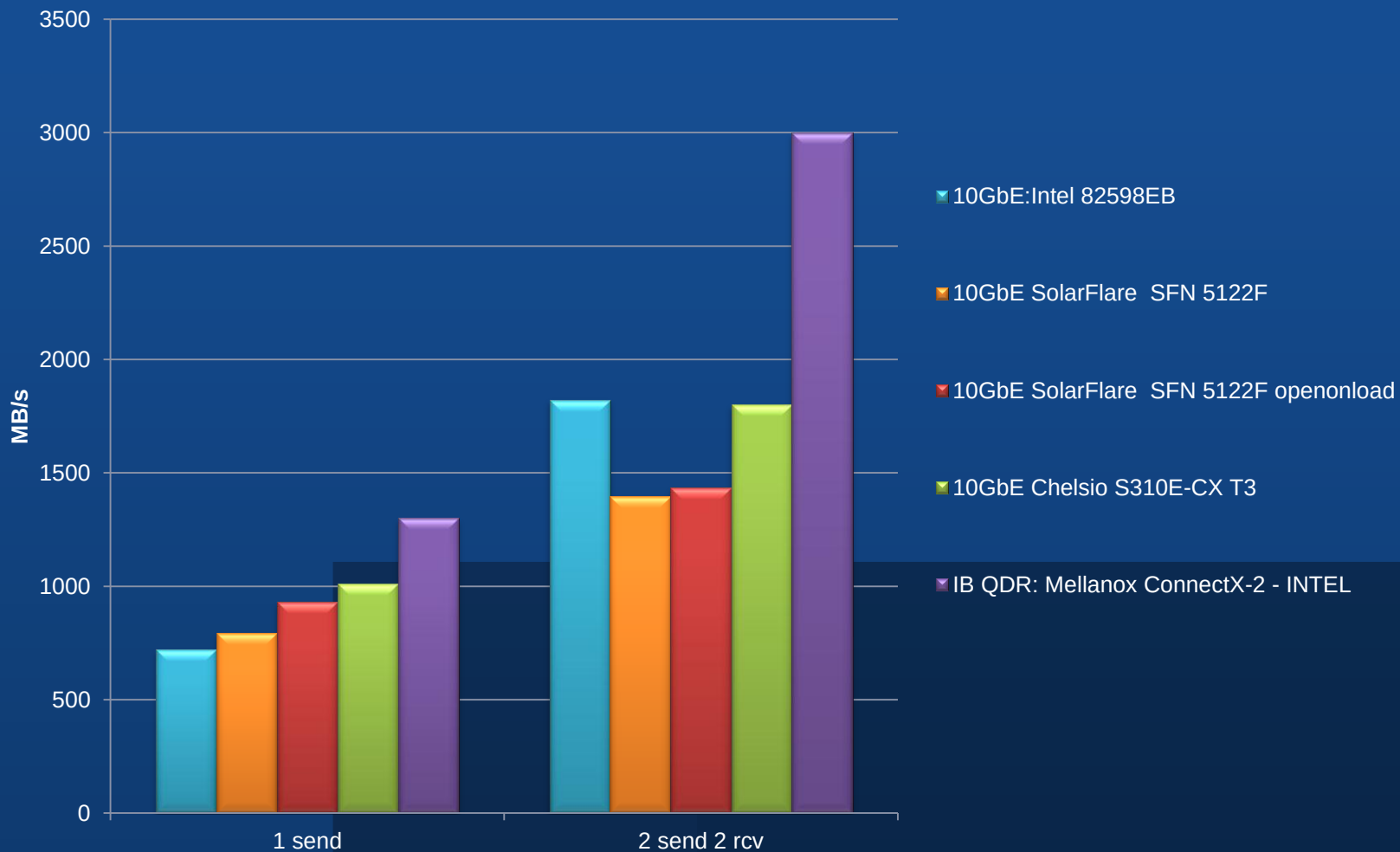


Interconnessioni

- 10 GbE
- 40 GbE
- Infiniband
 - FDR 56Gb/s (encoding 64/66)

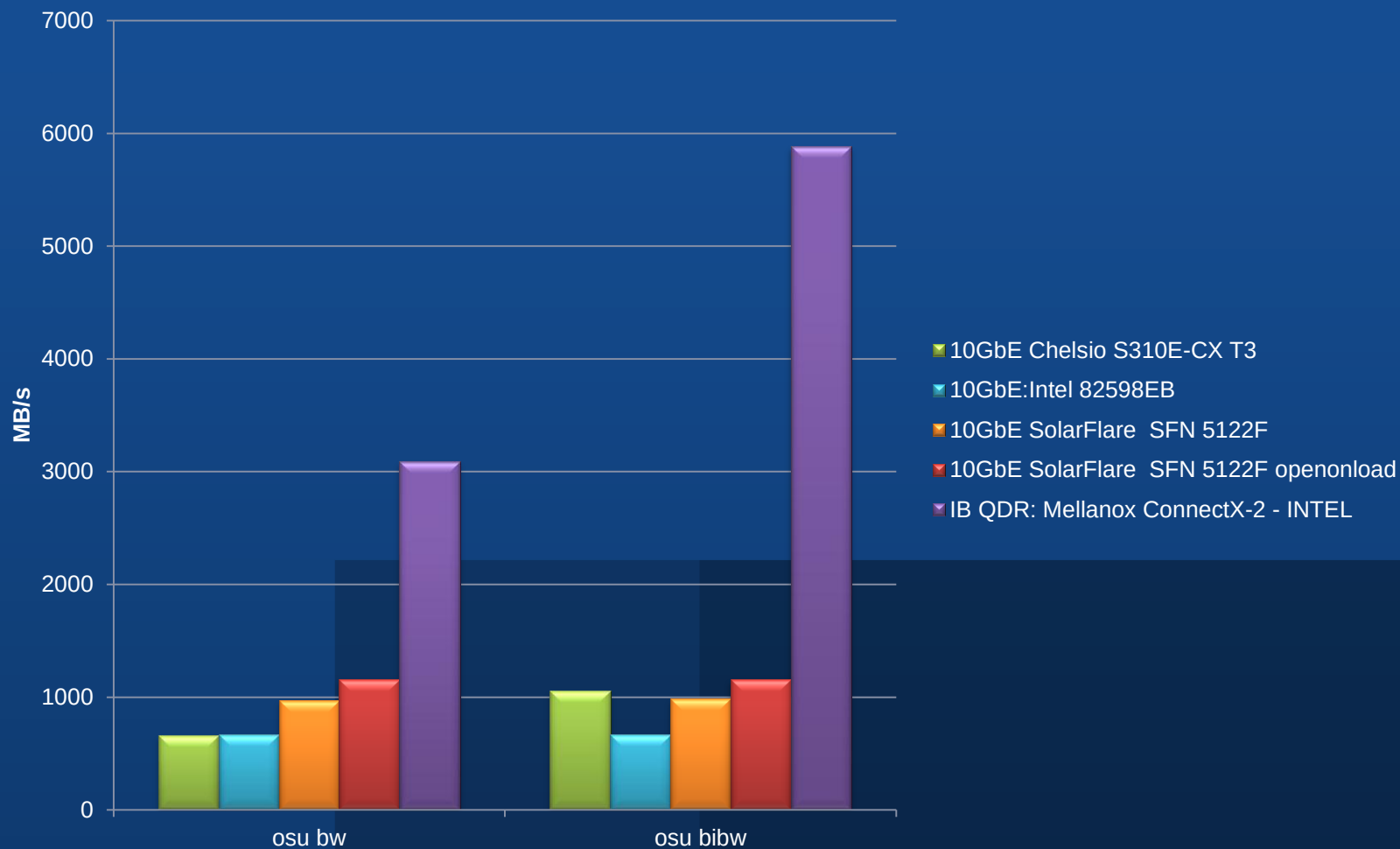


TCP performance - Iperf



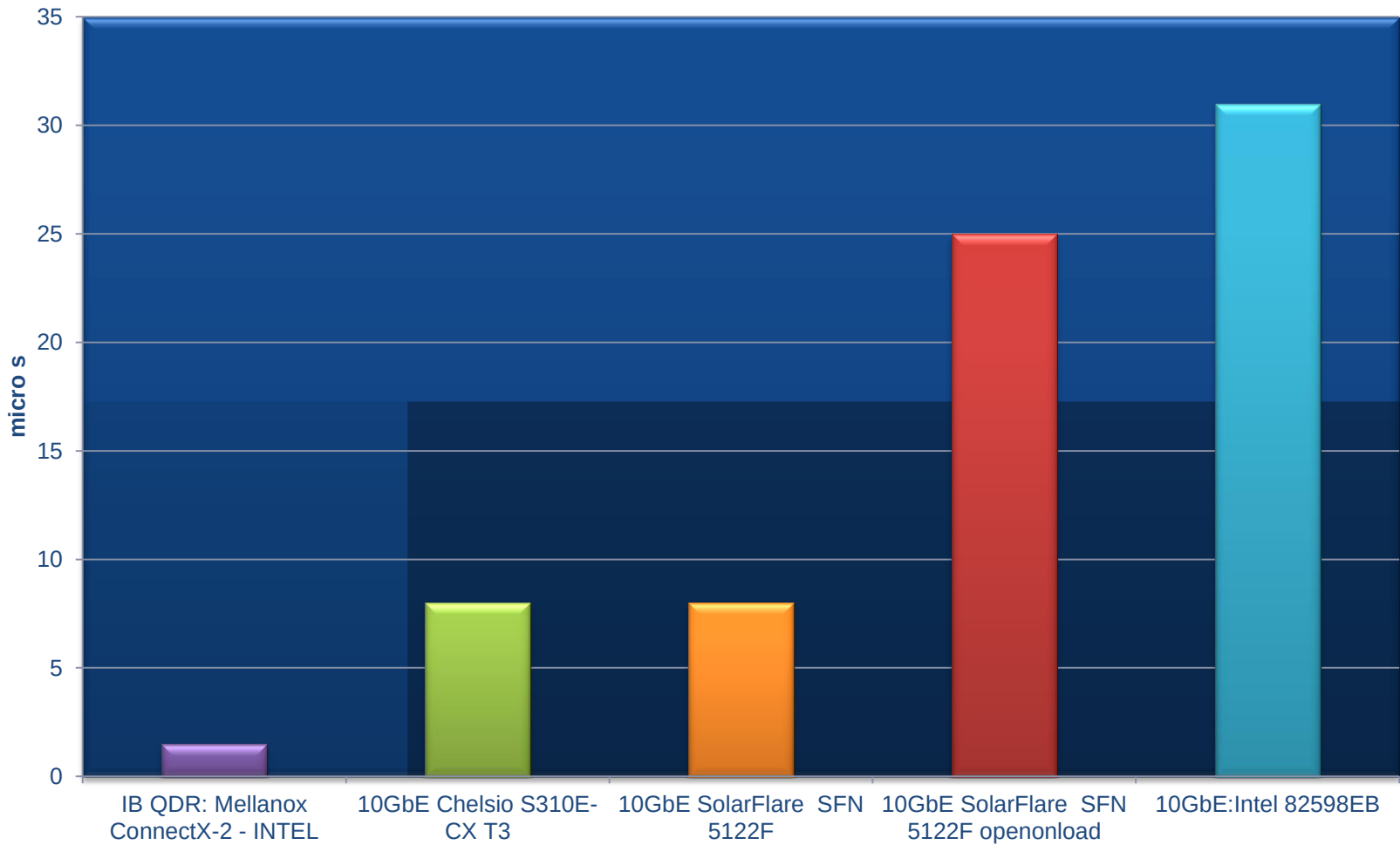


MPI max bandwidth 1: 10GbE Vs IB QDR



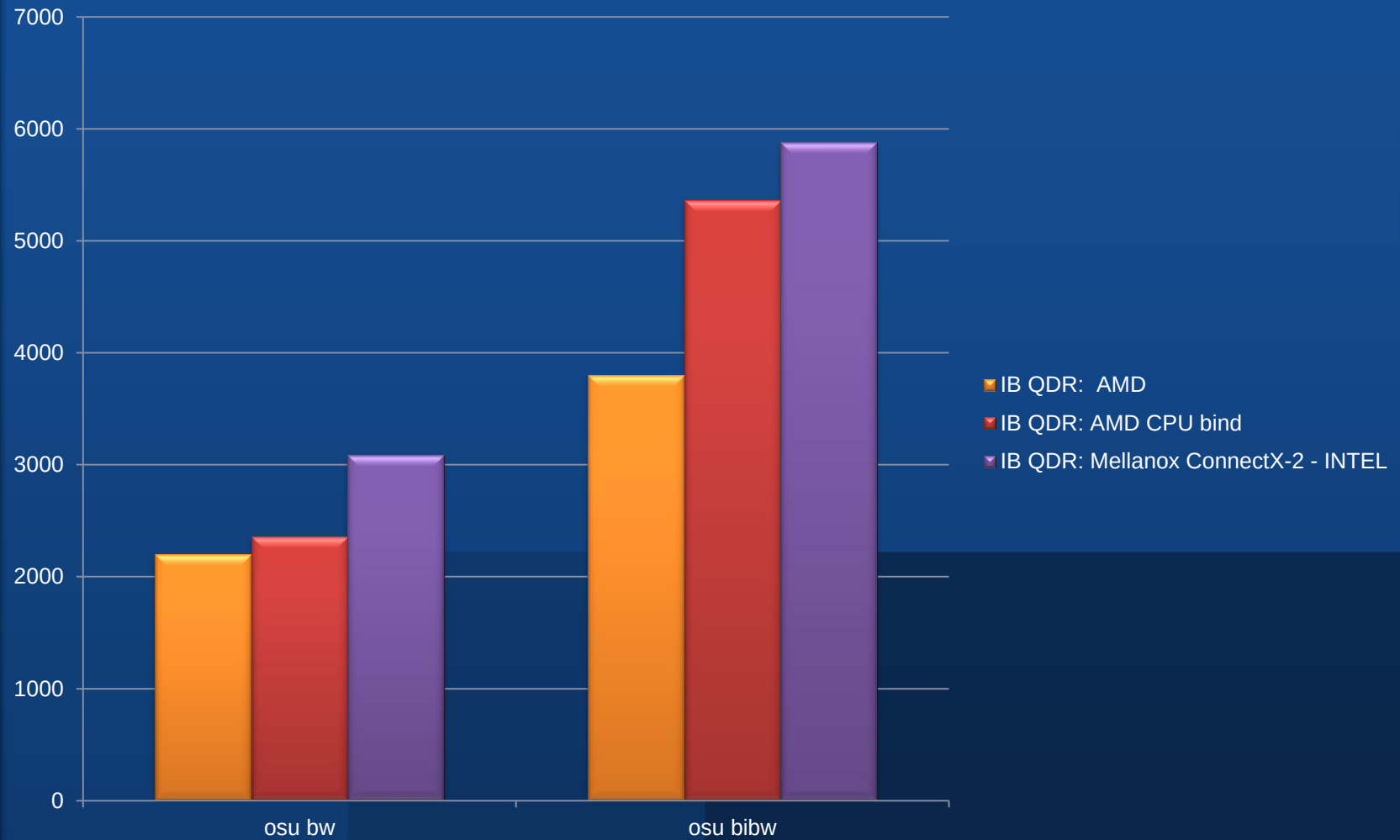


MPI min latency (OSU_lat)



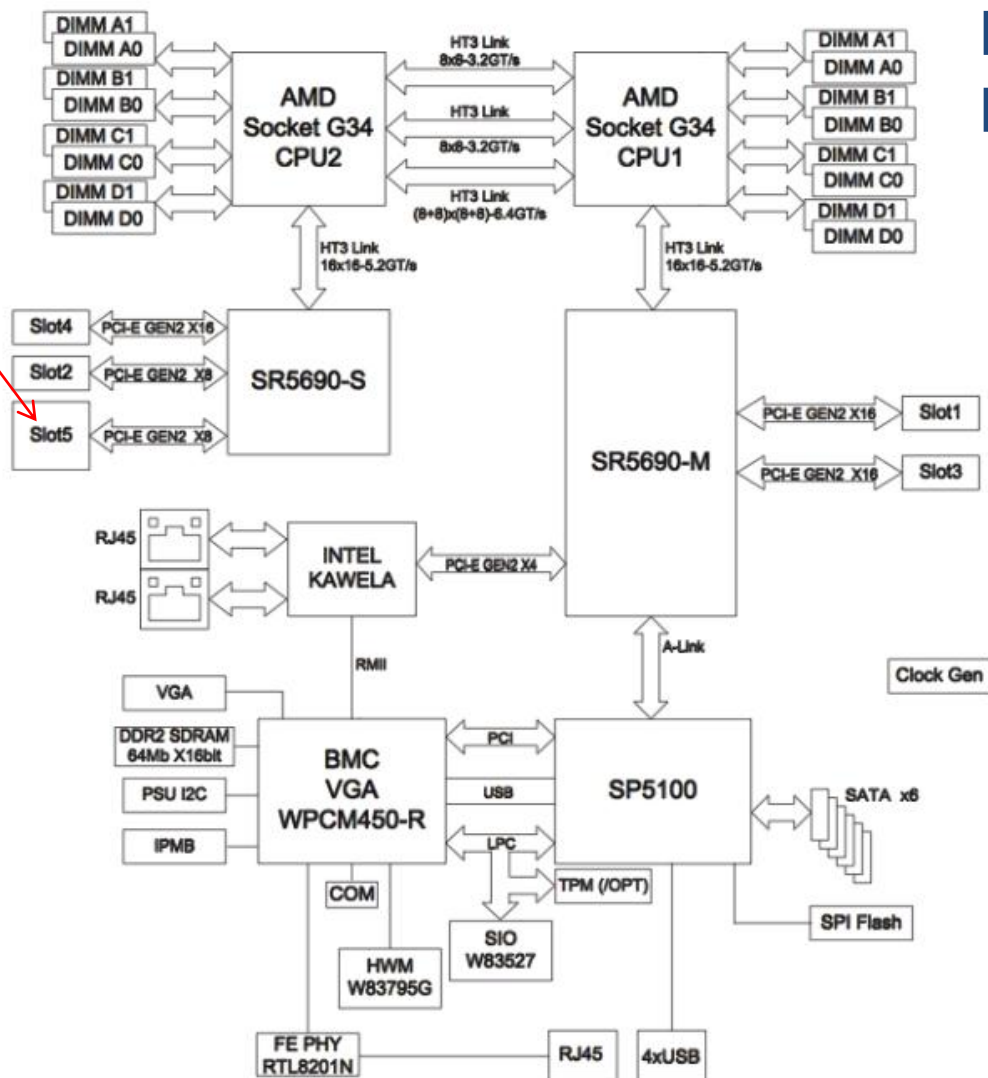


MPI max bandwidth 2: Inter-nodo





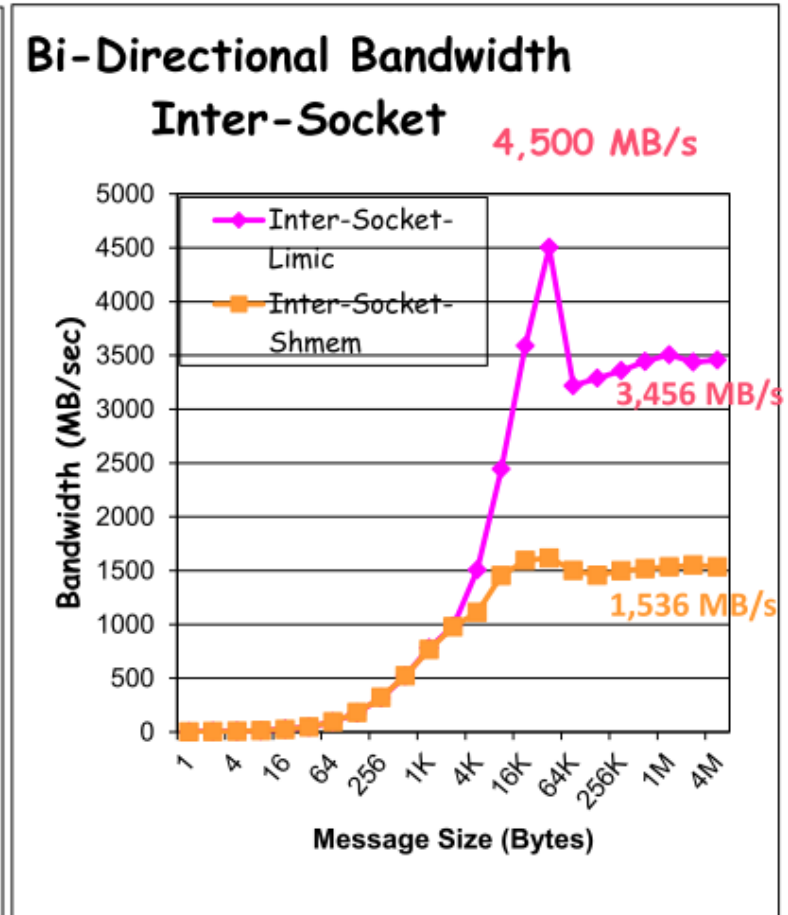
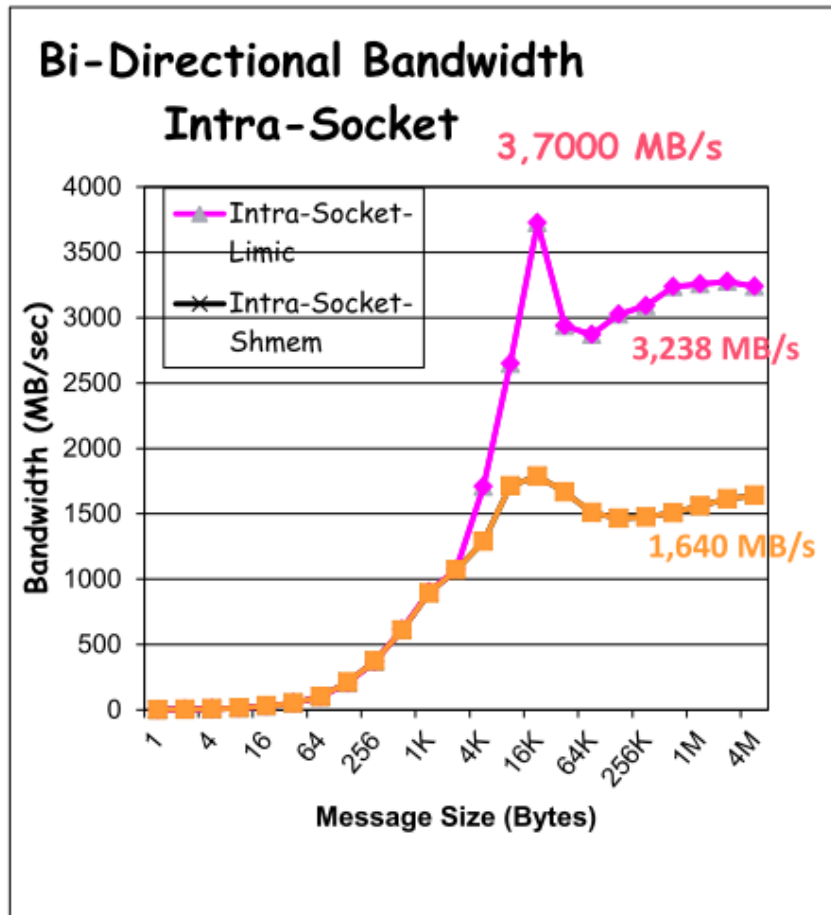
IB HCA



Multi chipset
Motherboard



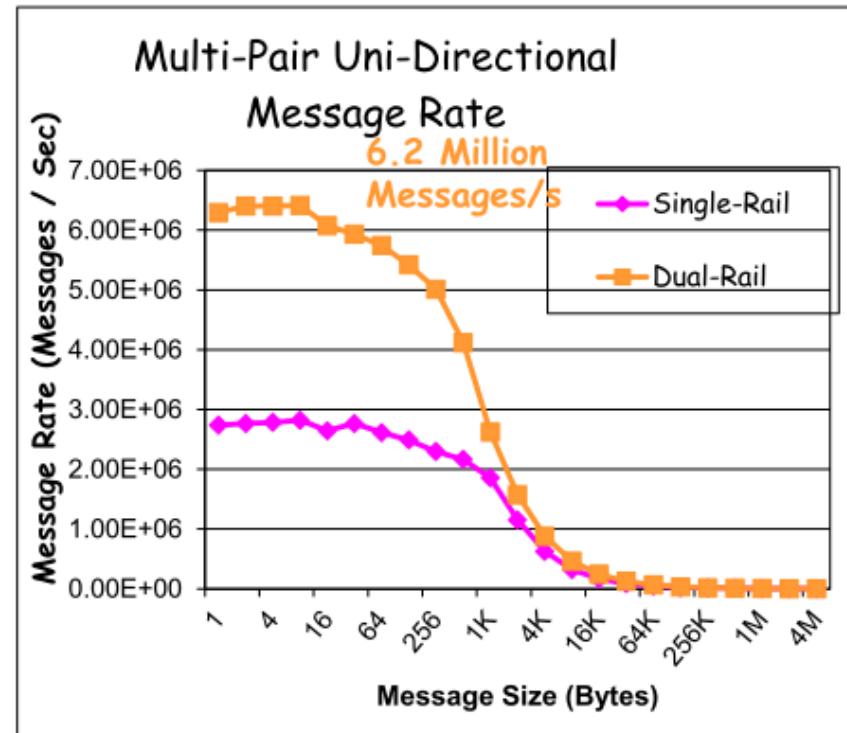
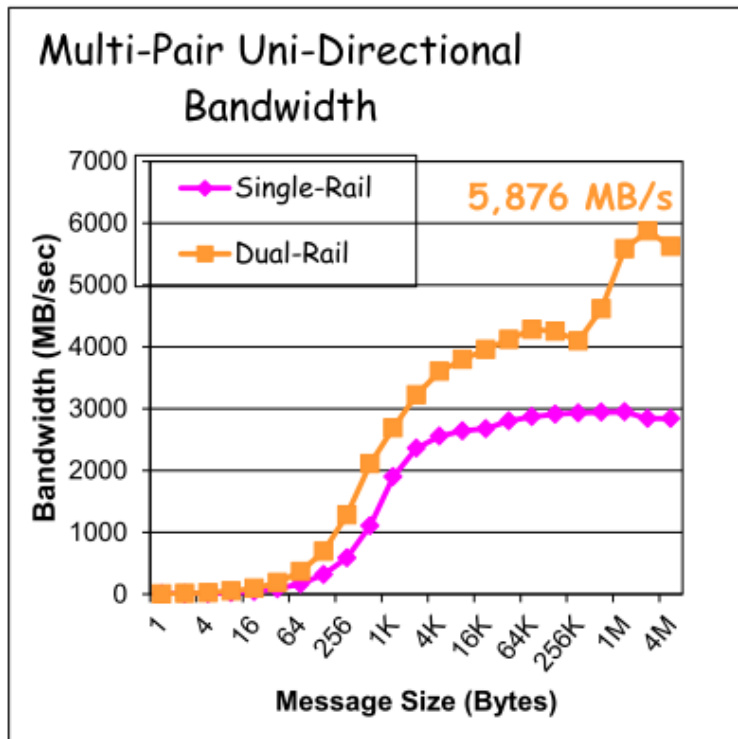
MPI bandwidth 3: Intra-nodo



Fonte: Dhabaleswar K. (DK) Panda, Sayantan Sur (The Ohio State University) - MPI Performance and Optimizations, Tutorial at HPC Advisory Council Workshop, Lugano 2011



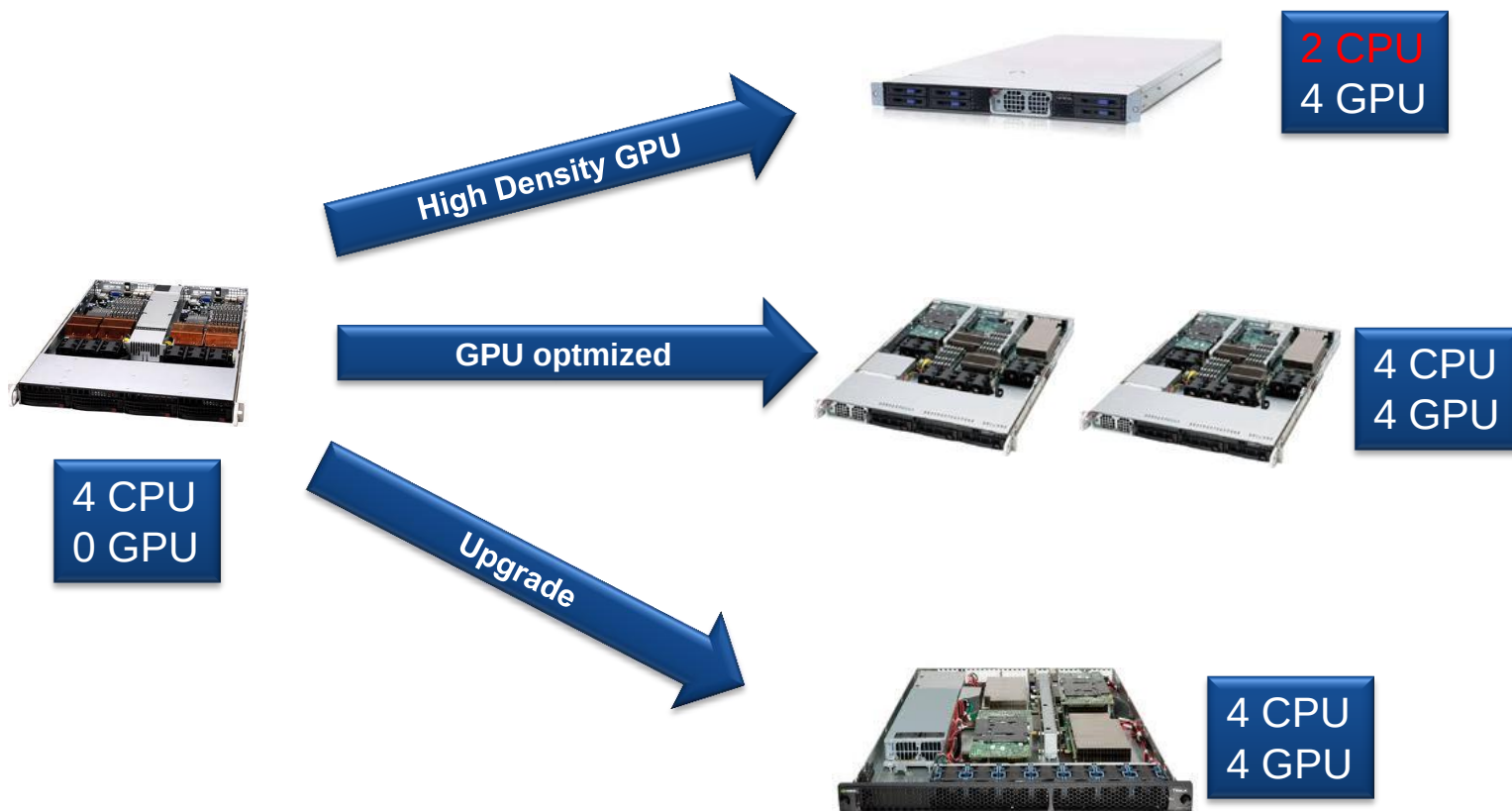
MPI Bidirectional bandwidth 5: Multirail



Fonte: Dhabaleswar K. (DK) Panda, Sayantan Sur (The Ohio State University) - MPI Performance and Optimizations, Tutorial at HPC Advisory Council Workshop, Lugano 2011

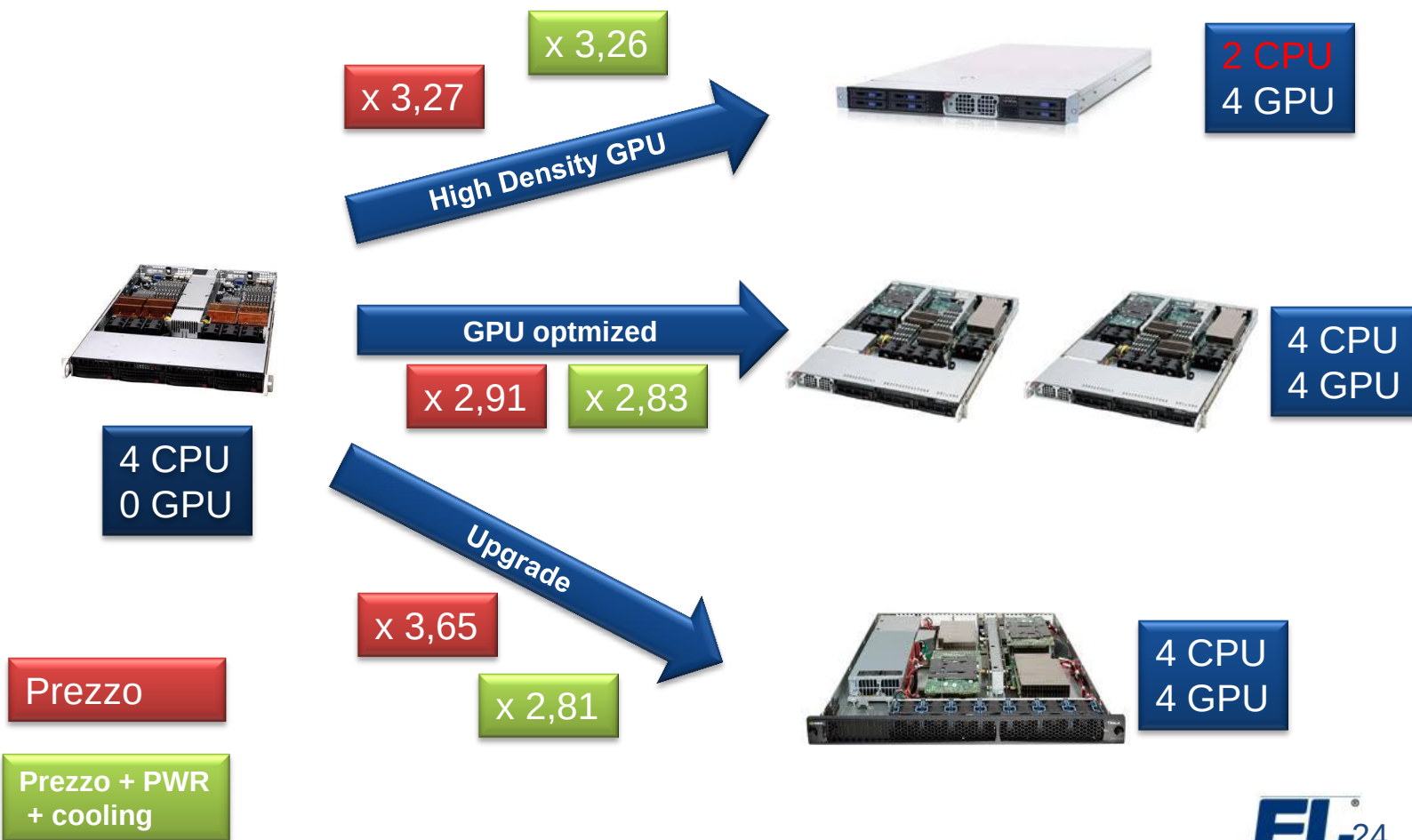


GPU: quando convergono?



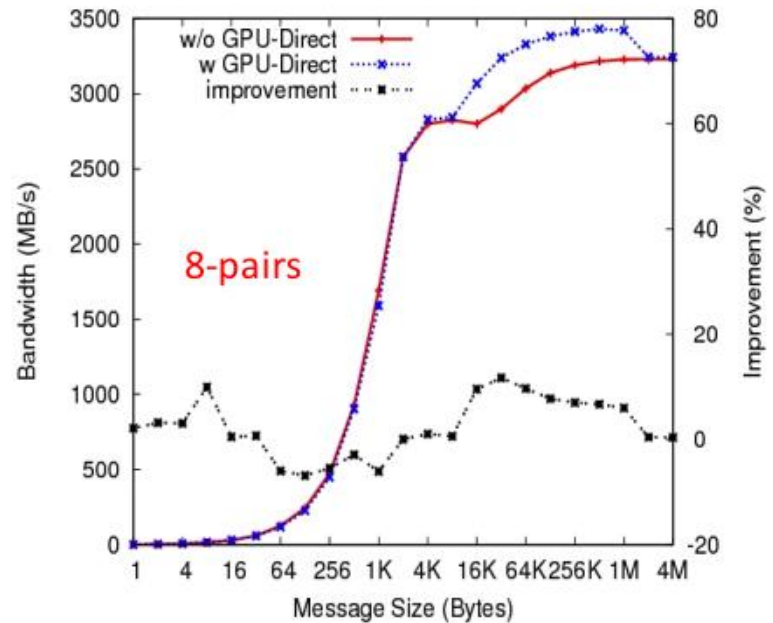
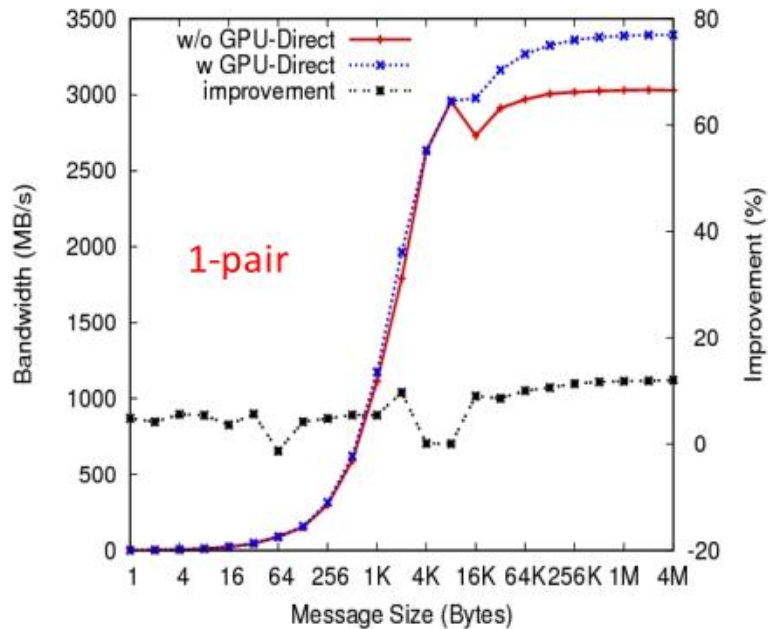


GPU: incremento dei costi



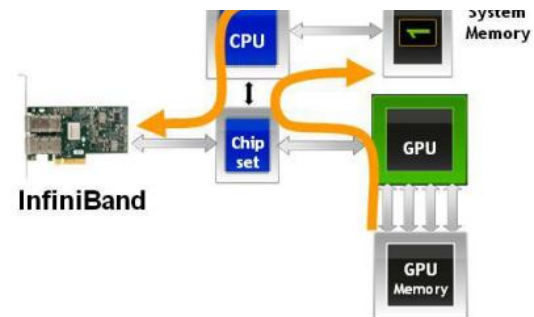


Comunicazione GPU to GPU



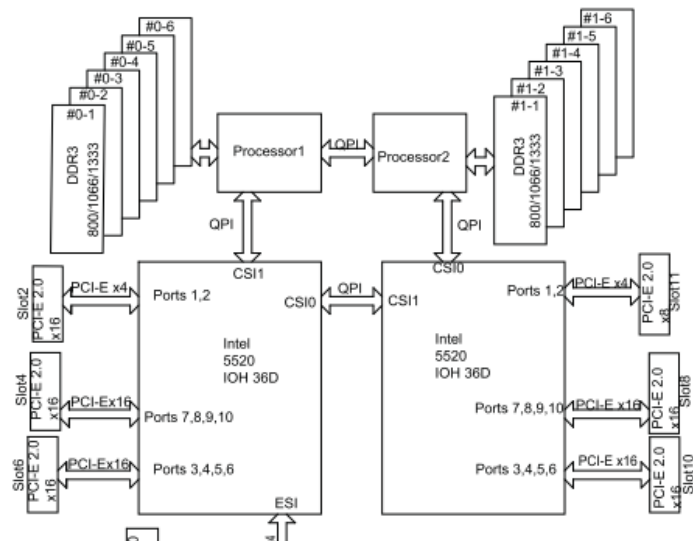
GPU direct: host buffer comune a GPU e HCA IB

Incrementi fino al 12%





CUDA 4.0



GPU peer to peer
su bus PCI-Express

NVIDIA GPU Direct 2,0 Technology

Unified Virtual Addressing (UVA)

Thrust C++ Template Libraries Performance Primitives



Domande ?