Per una ipotesi di Large Prototype

8 novembre ROMA

Alberto Ciampa

INFN-Pisa

Da un diverso punto di vista

- Quindi senza entrare nel merito degli applicativi e degli algoritmi
- Basandoci sulla nostra esperienza di cluster

- Un LP che potremmo progettare, acquisire, installare e manutenere.

Schema generale

- N server uguali connessi via IB QDR
- Abbiamo gia’ lo switch

Un certo numero di macchine fatte così:

- Node-cluster requirements:
  - dual socket CPUs per node
  - 4 (K20x) GPUs per CPU-socket (4 GPU per system it's enough, but attached to the same CPU)
  - 1 Infiniband adapter per CPU-socket (one it's enough in the case of 4 GPU)
  - GPUdirect P2P and GPUdirect RDMA between GPUs attached to the same socket
  - RDMA direct between the IB port and each of the GPU on the system.

Esistono delle macchine fatte così?

- HP dice di sì
- SuperMicro ne ha una che però non va bene… ma ne avrà una nuova che potrebbe essere giusta
- IBM no o meglio no con architettura X86, la avrà con il Power (accordo con Mellanox e NVidia)

E gli altri? Abbiamo parlato con…

- Nvidia perché vorremmo capire da loro quale potrebbe essere il vendor giusto. Li incontriamo il 18/11
- Dell: dobbiamo approfondire (nella stessa settimana, a Denver per SuperComputing)
- AMD: sono perplessi, ma anche con loro appuntamento il 18/11 (anche per capire la loro roadmap)
- INTEL: anche con loro approfondimento in corso

- Ci sarebbero anche altri con i quali stiamo cercando di intavolare contatti e approfondimenti: Tyan, Fujtsu, …

Quello che sappiamo ora

Ci occorrono macchine con

- *motherboard con un processore Intel*
- *4 board NVidia Kepler (K20(x))*
- *una porta IB QDR.*

cinque porte PCI-E gestite dallo stesso processore

- 4 porte PCI-E x 16 (per le Kepler)
- 1 porta PCI-E x 8 (almeno, per la porta IB).

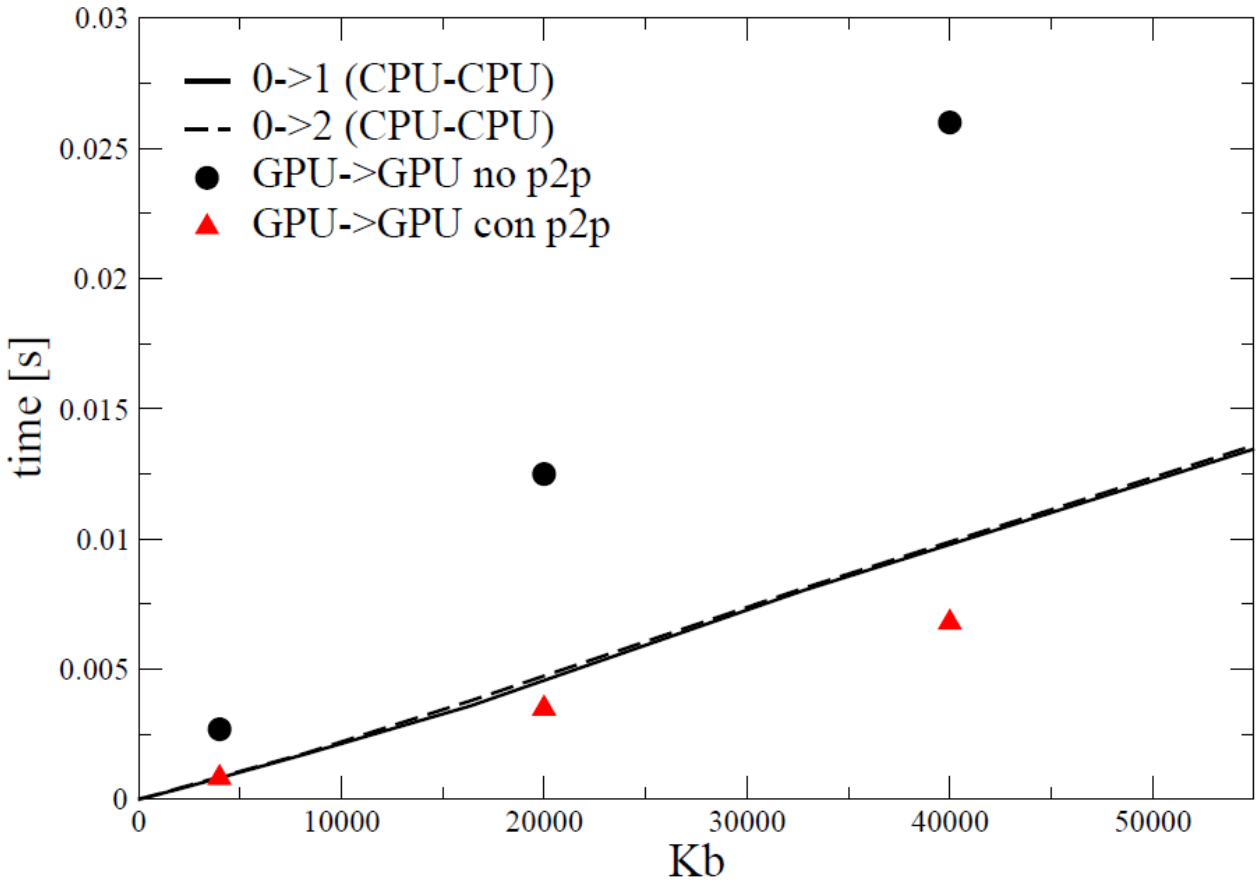
Quello che abbiamo ora…

macchine con 4 Kepler + IB

- ma per il bus PCI-E occorrono due processori (E5)

- Le prove di Massimo

Ultimo test di Massimo D' Elia



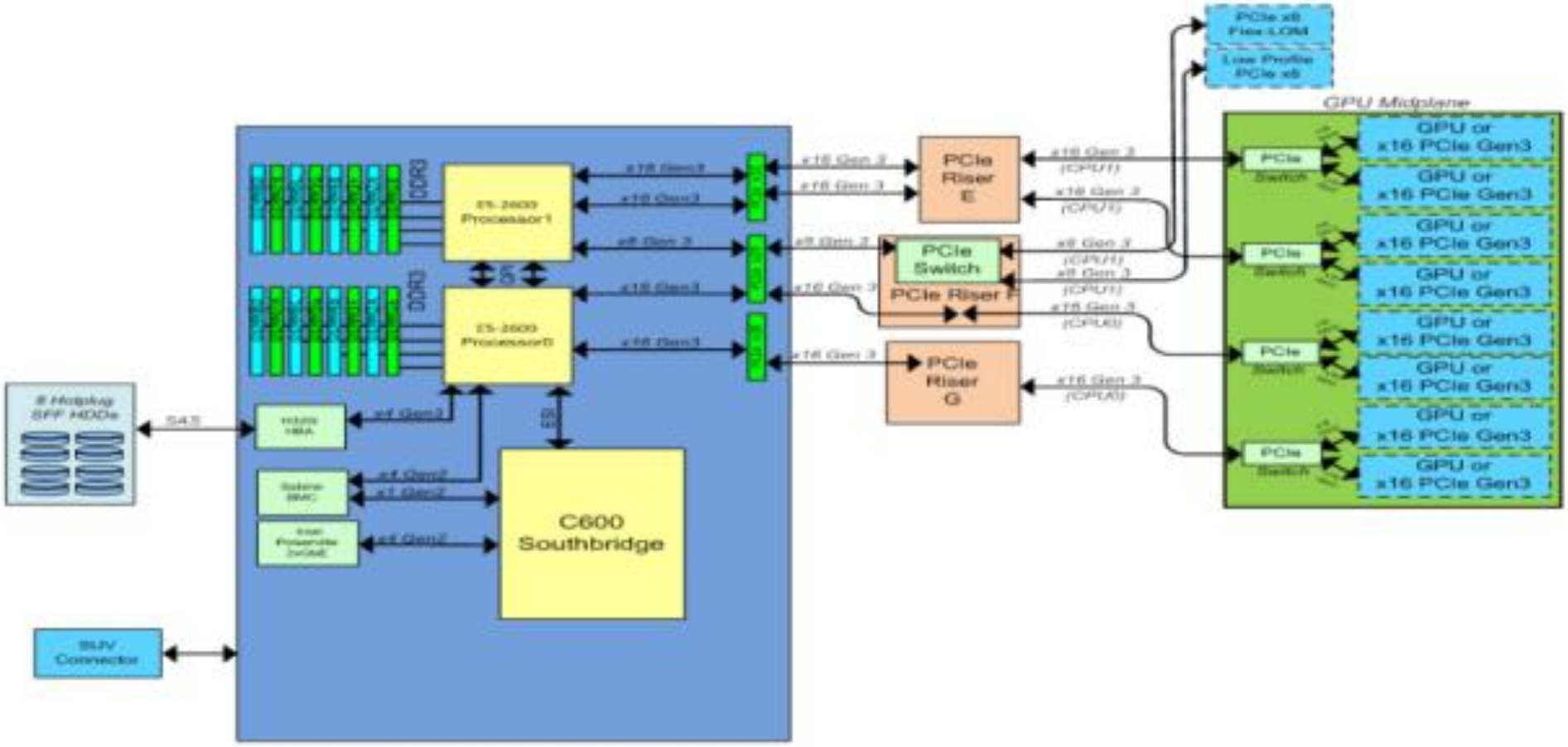
processore

- INTEL

  i processori XEON E5 Westmere forniscono al max 40 lane PCI-e e quindi non si possano collegare 4 pci-e x16 + 1x8 . Ci si deve indirizzare necessariamente alle soluzioni Sandy Bridge / Ivy Bridge che hanno 96 lane.

- AMD ?

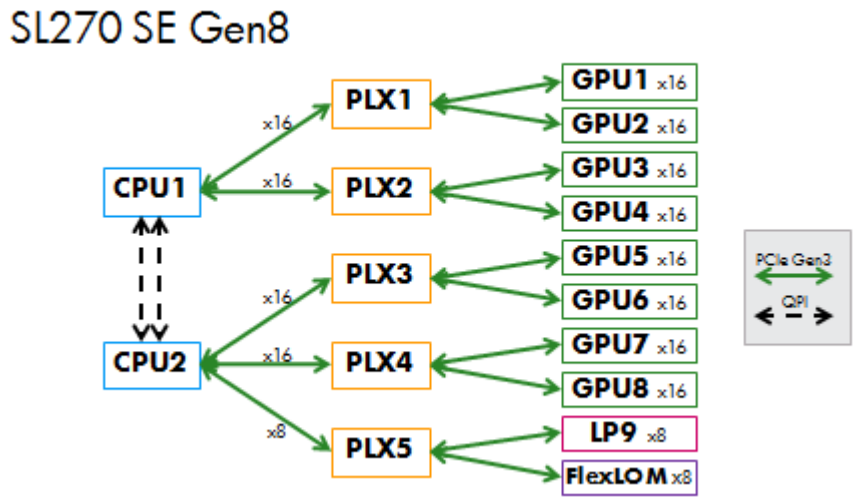
HP SL270


PCIe x8 Flex-LOM
Low Profile PCIe x8
GPU Midplane
PCIe Riser E
PCIe Switch
PCIe Riser F
PCIe Riser G
GPU or x16 PCIe Gen3
E5-2600 Processor1
E5-2600 Processor0
DDR3
QPI
C600 Southbridge
8 Hotplug SFF HDDs
SAS
SUV Connector

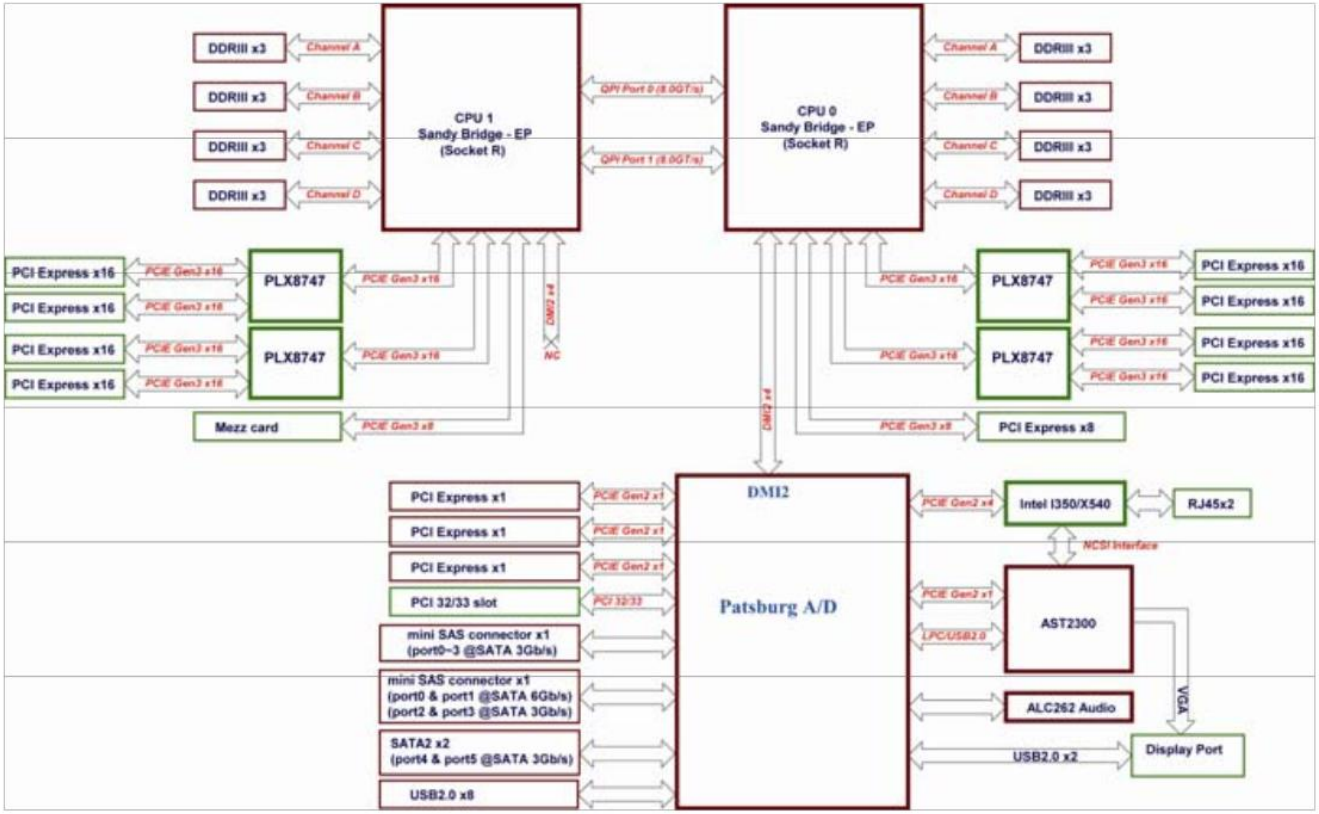
Ancora piu’ nel dettaglio

Ma c’e’ un PLX

Cosa significherà?



PLX anche su Tyan



S7059 Block Diagram

domande

- ogni socket ha 4 GPU connesse a due a due ad uno swtch PLX8746 a 46 linee PCI-express. Gli switch sono a loro volta connessi al root-complex della CPU, a cui e' connesso anche un bus pciE 8X in cui si puo' pluggare una scheda Infiniband
  - in questa configurazione e' possibile abilitare il peer-to-peer tra qualsiasi coppia di schede GPU ? Inoltre e' possibile effettuare il RDMA tra una qualunque delle GPU e la scheda Infiniband pluggata nel bus 8X ?

- la PLX produce switch PCIE a 96 linee. Quindi, esiste un qualche configurazione di un sistema quad-GPU, in cui tutte le GPU e la scheda Infinibad sono connesse direttamente allo switch PLX che a sua volta e' connesso al root-complex della CPU ?

Le macchine HP

- Anche se sembrano quelle giuste (da verificare, comunque)
- … costano (pare) 40.000 euro di listino…

- Vale la pena guardare anche altro: SuperMicro

SuperMicro

- Il sistema attuale:

  SYS-2027GR-TRFH (2U 6GPU)

  Anche qui riser card con PLX

  (RSC-R2UG-A2E16-A e RSC-R2UG-A2E16-B)

- Il sistema che stanno per annunciare:

  SYS-4027GR-TR (4U 8GPU)

Sistema nuovo Supermicro SYS-4027GR-TR (4U 8GPU)

SUPERMICRO

Confidential

X9DRG-O(T)F-CPU

Key Features:

- ✓ Dual SNB EP E5-2600 (Socket R up to 150W)
- ✓ C602 Chipset
- ✓ 24 DIMM, 768GB Reg. ECC DDR3 up to 1600MHz
- ✓ 8 PCI-E 3.0 x16 (double-width)
- ✓ 2 PCI-E 3.0 x8 (in x16 slot)
- ✓ 1 PCI-E 2.0 x4 (in x16 slot)
- ✓ 2 SATA3 + 8 SATA2 ports
- ✓ Intel I350 Dual Gigabit LAN or X540 dual 10GBASE-T (T SKU)
- ✓ USB 2.0 Type A connector
- ✓ 17" x 19" Form Factor

System Model Number:

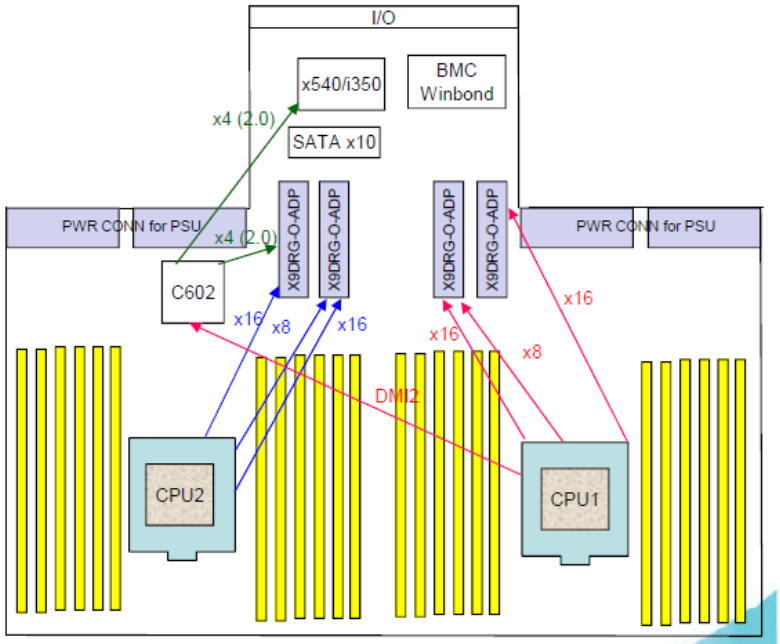
- ✓ SYS-4027GR-TR(T)

Status:

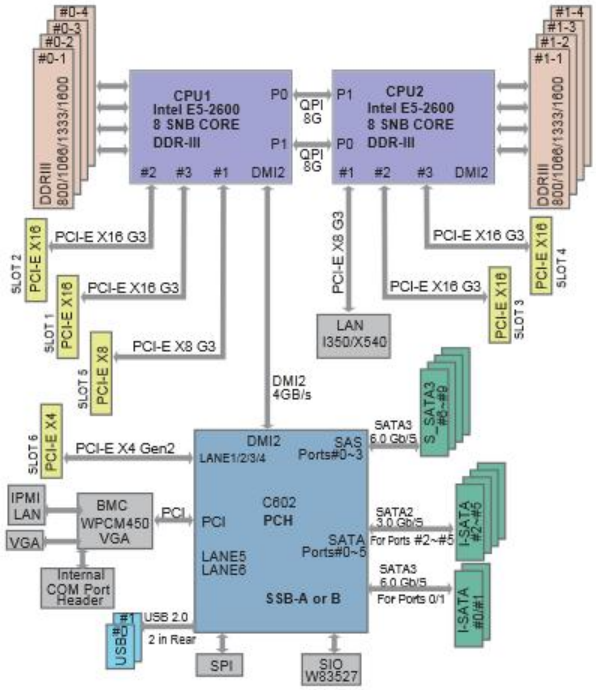
- ✓ Sampling Sept 2013
- ✓ PR Oct 2013

Applications:

- ✓ Grid VDI
- ✓ HPC



SYS-2027GR-TRFH (2U 6GPU)



System Block Diagram

Note: This is a general block diagram and may not exactly represent the features on your motherboard. See the Motherboard Features pages for the actual specifications of each motherboard.

SYS-4027GR-TR

SUPERMICRO

Confidential

X9DRG-O-PCIE, X9DRG-O-ADP