GPU and server architecture

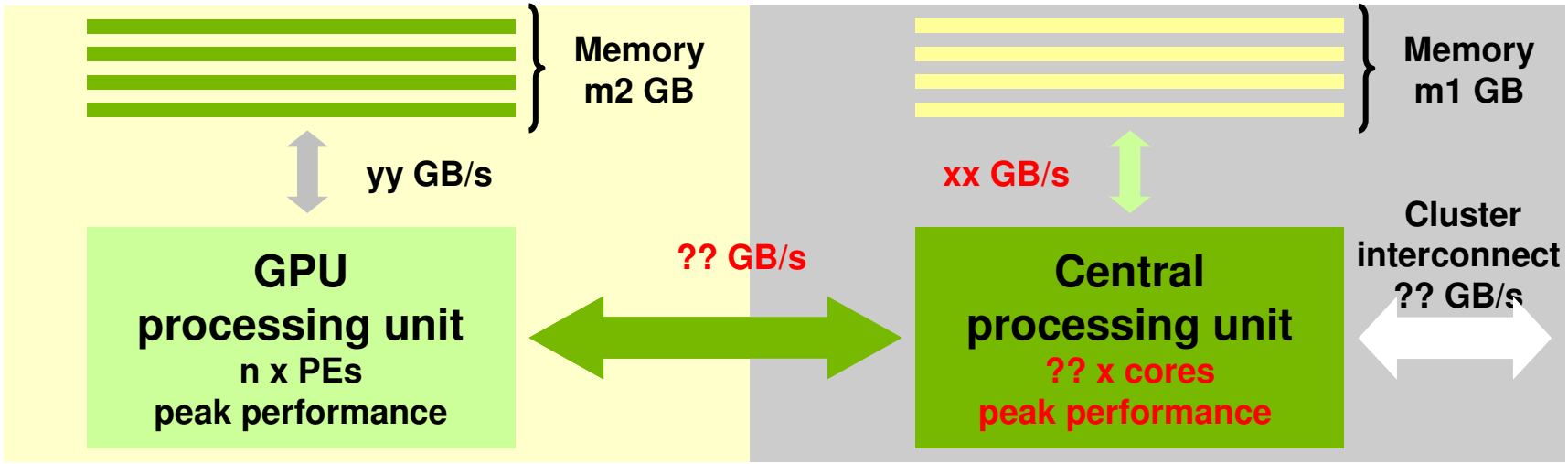
JF Lavignon - 1/07/2009

LIBERATE IT

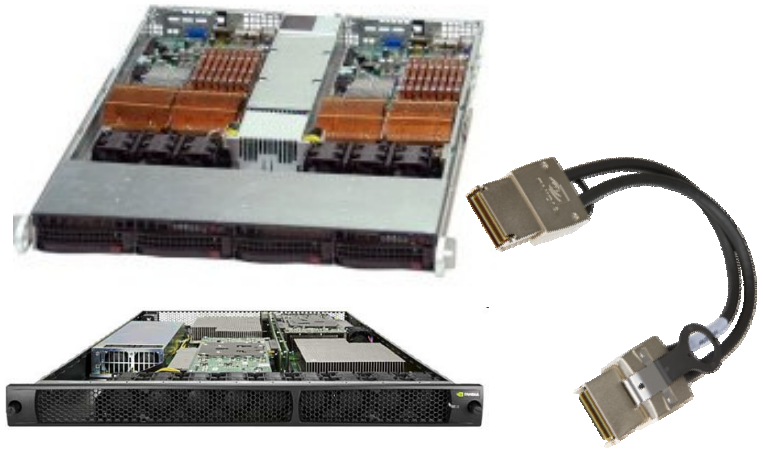
Architecture for GPU system

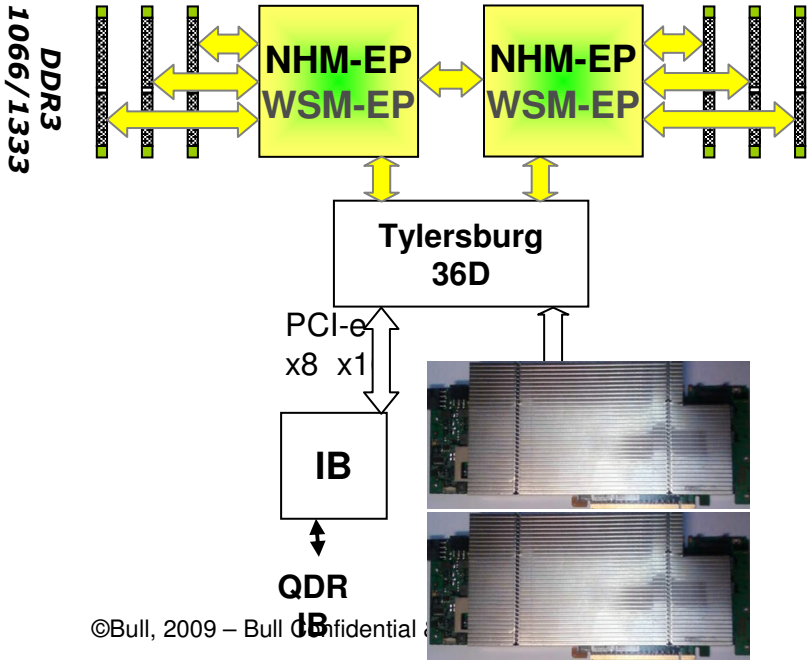
- GPU associated with central system for OS support, network, storage…
- Several parameters to set
- Right balance depends on applications needs

➢ Several options



Simplest architecture : 1x R422-E1 + 1x S1070





# of Tesla Processors	4
# of Computing Cores	960 (240 per processor)
Floating Point Precision	IEEE 754 single & double
Total Dedicated Memory	16 GB (organized as 4.0 GB per GPU)
Memory Interface	4x 512-bit GDDR3 memory interface (organized as a 512-bit interface per GPU)
Memory Bandwidth	408 GB/sec (102 GB/s per GPU to local memory)
Typical Power Consumption	700 W
System Interface	PCIe x16 or x8
Programming environment	CUDA

Features

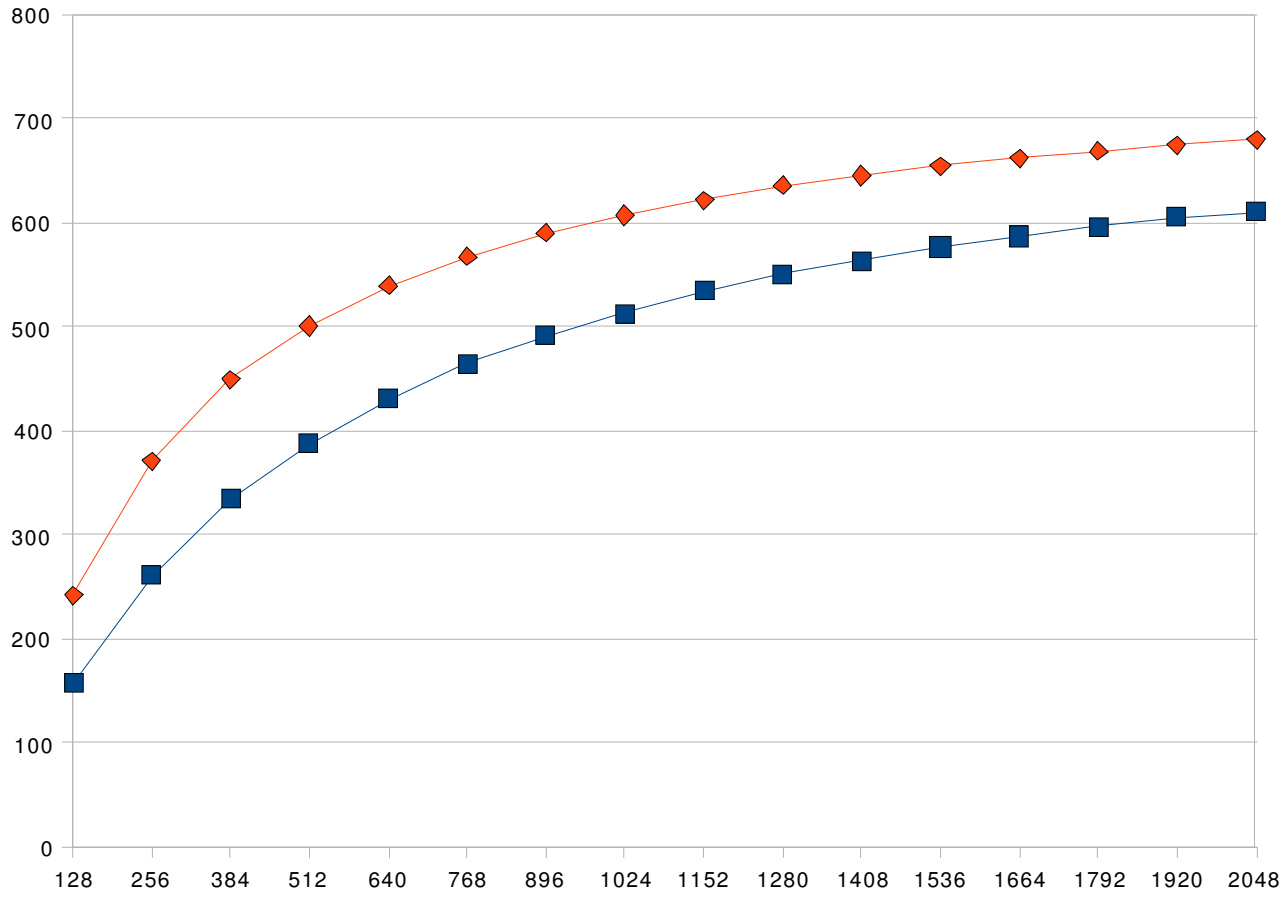
in 2U

- ✓ 4 x GPU T10
- ✓ 2x 2 Xeon 5500
- ✓ 2x 2-Socket
- ✓ 2x 8 DIMMs
- ✓ **2x 1 PCI-Express x16 Gen2**
- ✓ 2x 2 SATA2 Hot-Swap HDD
- ✓ **92% PSU efficiency**

PCIe impact on performance : case SGEMM

- Mesure de la performance d'une multiplication de matrice simple précision dans les deux cas suivants:
  - Un seul port PCI-Express 2.0 partagé par deux GPU.
  - Un port PCI-Express 2.0 par GPU.
- Les paramètres m et n des multiplications de matrices sont fixes pour une occupation de la matrice principale(C) de l'ordre de 1,5 Go.
  - C[m,n] = C[m,n] – A[m,k]*B[k,n]
  - m=n=20000
- On fait varier k de 128 à 2048.
  - Cela modifie le ratio de communications(PCI)/calculs.
  - Les communications sont synchrones

Performances sgemm cumulées sur deux T10

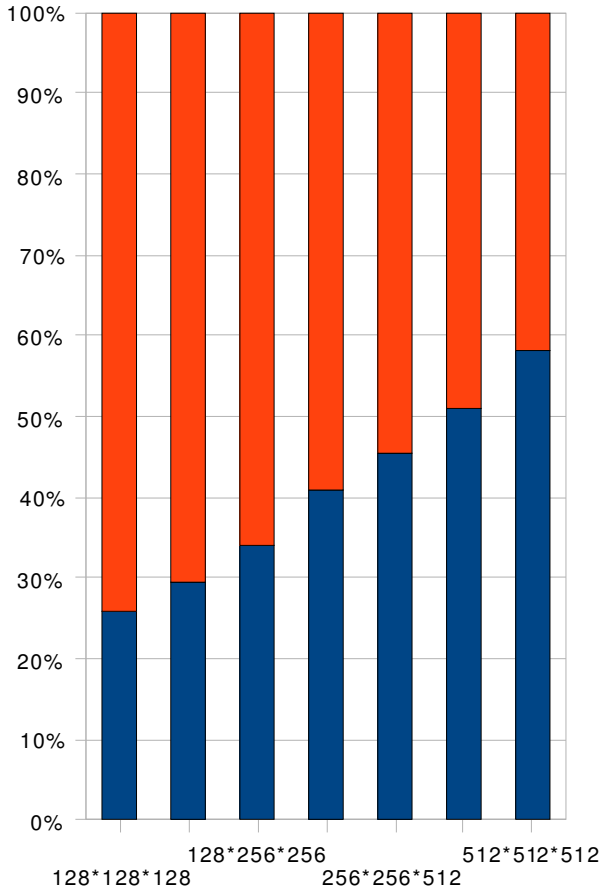


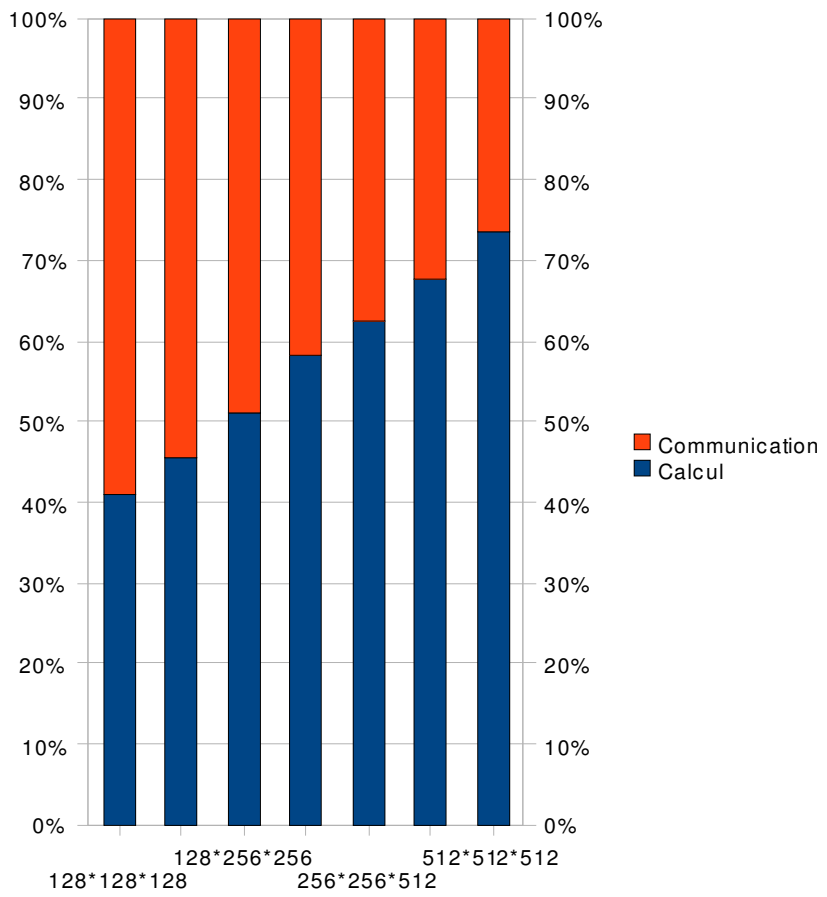
Between 50% and 12% performance increase

Interconnect impact on performance : 3D code

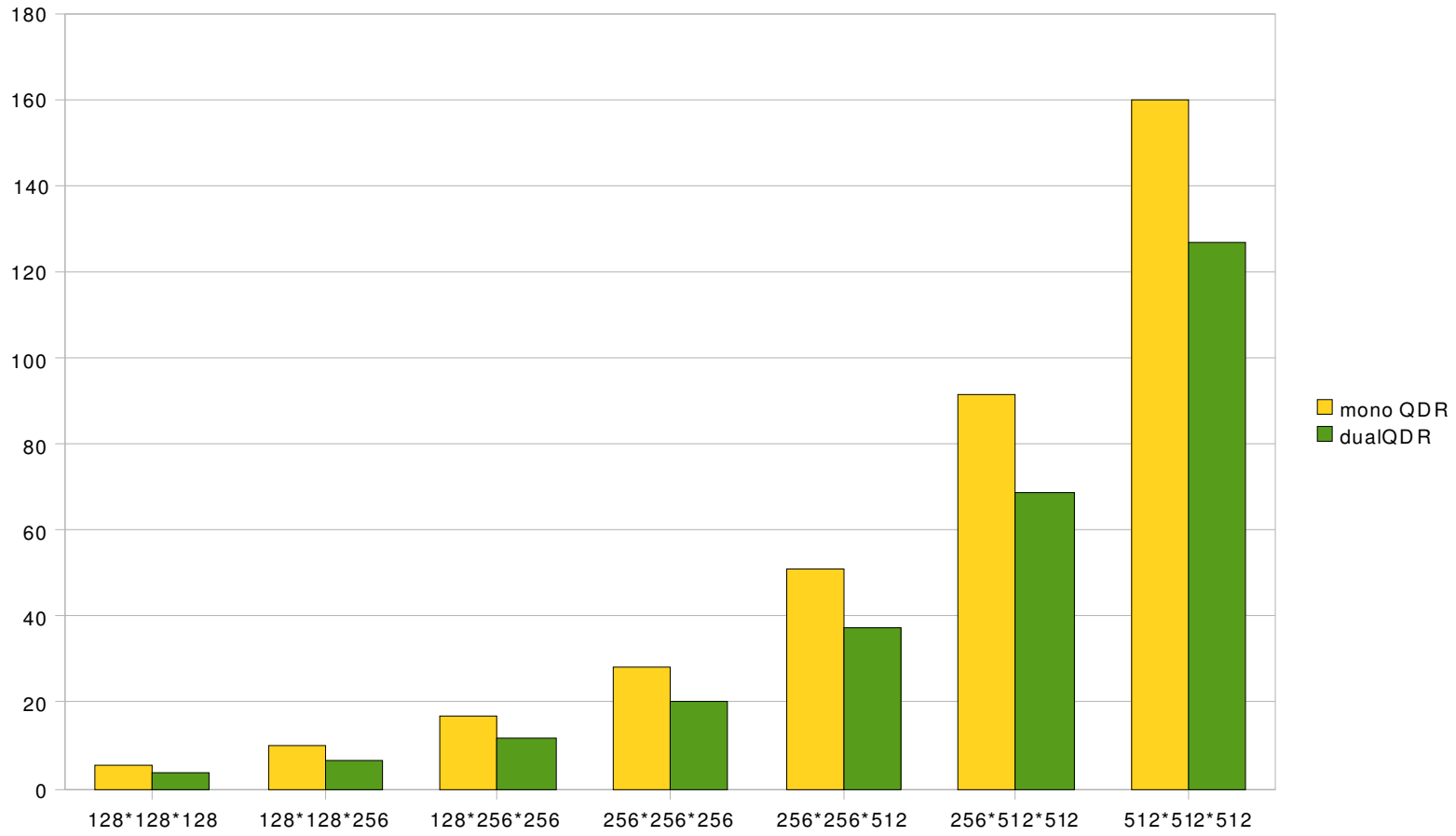
- Décomposition du maillage 3D sur les noeuds
- Le maillage d'un nœud tient dans la mémoire du GPU.
- Performance des calculs double précision est indexée sur la bande passante mémoire du GPU.
  - Chaque variable sera considérée comme étant chargée en moyenne 2 fois en mémoire graphique.
- On fait varier la taille et la forme du maillage par nœuds
- On fixe le nombre de mailles fantômes à 4 dans chaque direction pour chaque variable.
  - Communication des nœuds fantômes dans les axes X, Y et Z.

Comparaison mono QDR / dual QDR: %



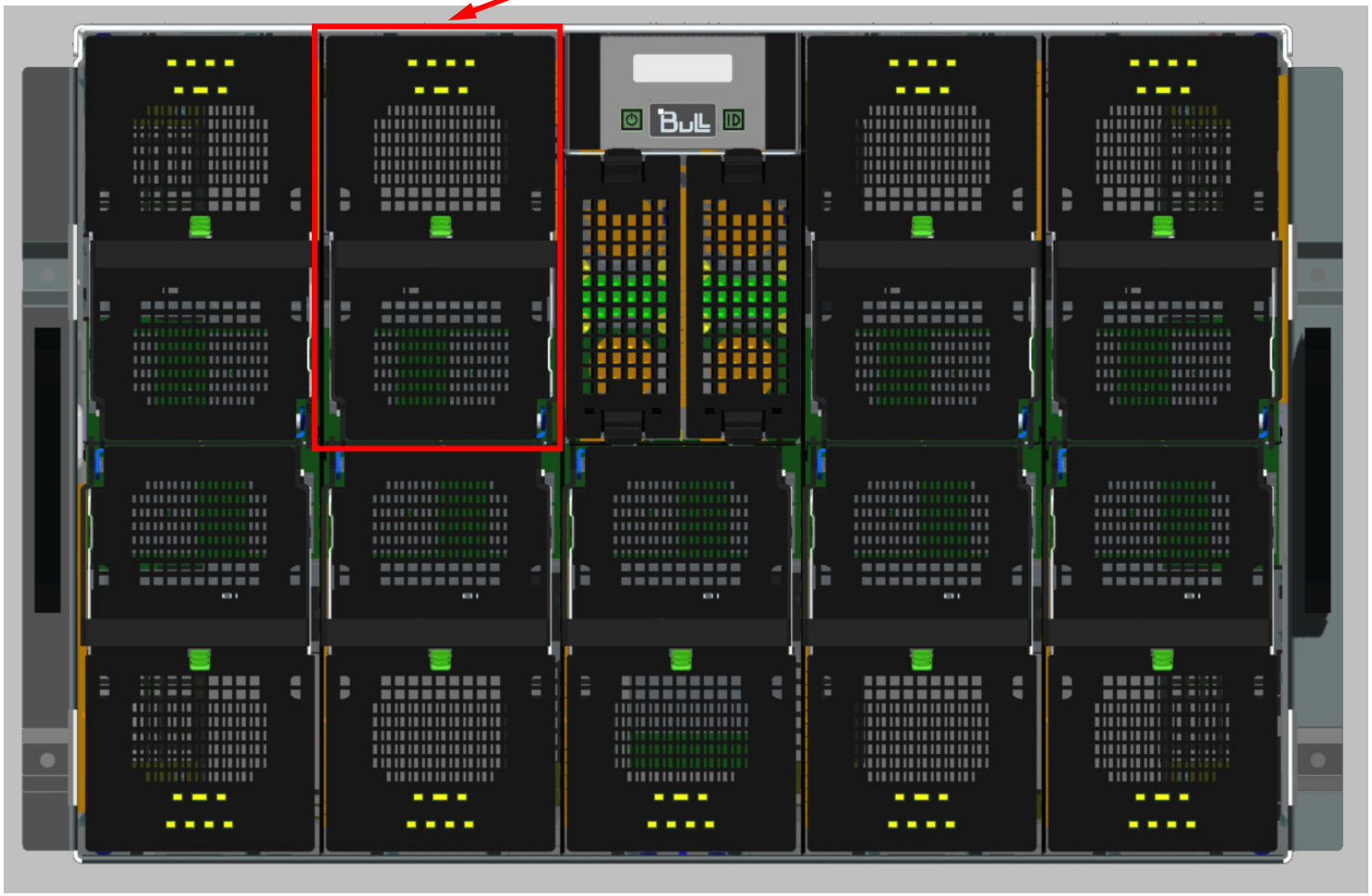


Comparaison: temps par itération en ms



bullx B505 : dual-width dual-GPU blade

One of the 9 dual-width slots in Inca



>>> 18 GPUs in 7 U

GPU hosting : Outline of the architecture

- Blade type integration in bullx B505
  - just an additional blade type (dual-width)
  - Improved density (2.5+ GPUs/U)
  - Fine granularity : 2 GPUs
- Low cost and power overhead
  - Stripped down support server
  - Infrastructure costs shared between blades
- Highest performance
  - Dedicated PCIe (no "oversubscription") and PCI-e/IB balance
    - Positively differentiates offer from S1070
  - bullx blade system cooling capabilities allow hosting of top speed GPUs

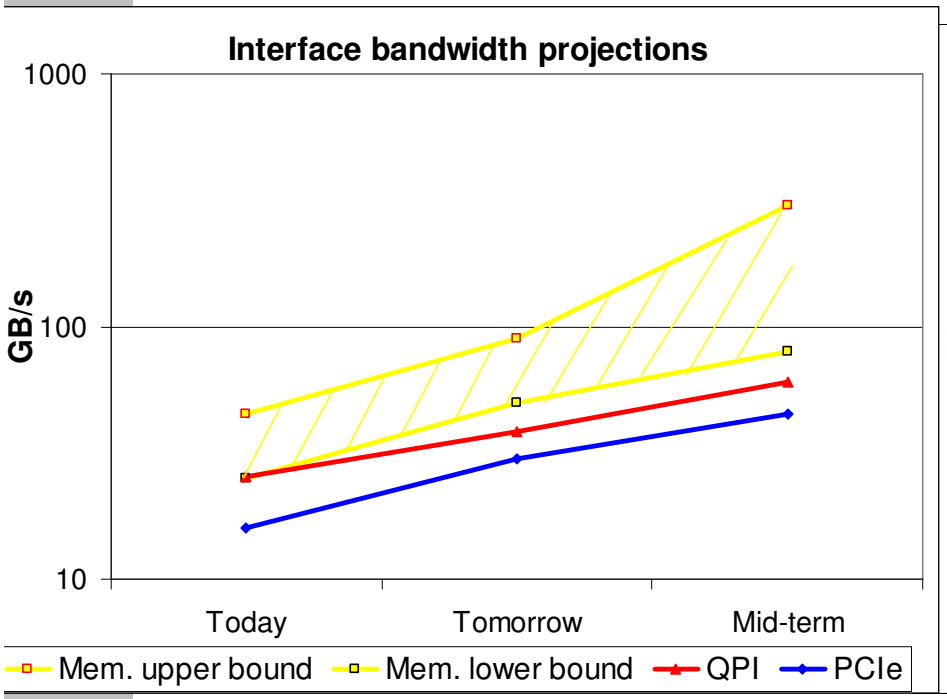
Block diagram of support server

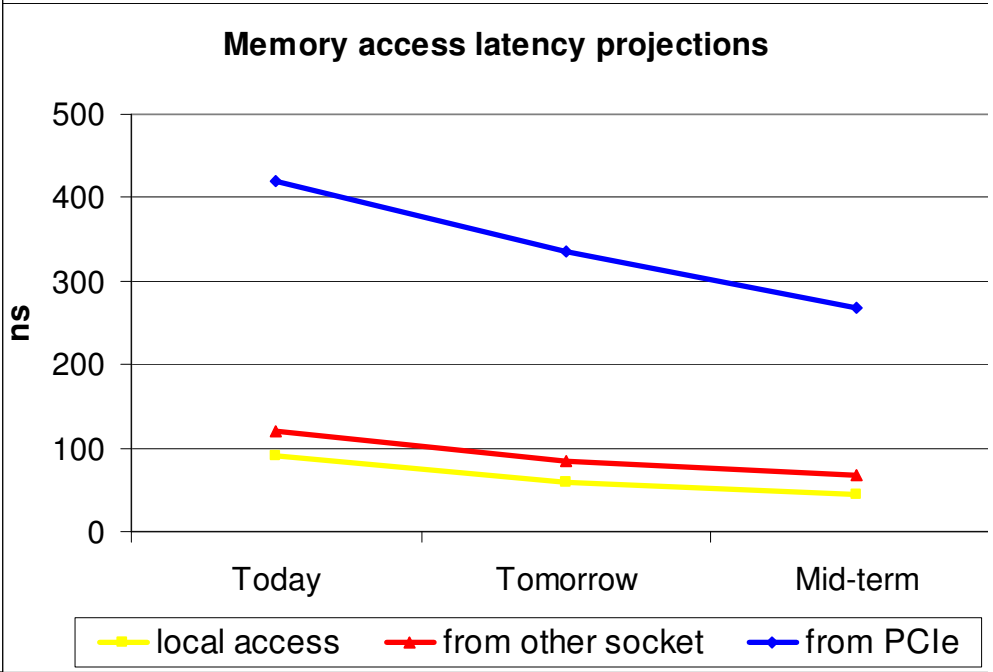
- No PCIe oversubscription and QDR IB availability
- Server target contents :
  - 2 processor sockets
    - 4 cores or less,
    - entry/medium/high SKU
  - Memory (up to 6 slots)
    - 3 slots per processor socket
  - BMC, …
  - 2 Tylersburg I/O hubs (36 PCIe lanes)
    - Two x16 PCIe interfaces
    - One or two x8 PCIe for Infiniband connection (on board ConnectX, at QDR rate)

Improving the current architecture

- How to improve memory access ?

		PCIe (x16)	Socket to socket	Socket to local memory
Today (2008-2009)	Bandwidth	16 GB/s	25.6 GB/s	25-45 GB/s
	Latency	400 + ns	100 + ns	100 - ns
Tomorrow (2010-2011)	Bandwidth	30 GB/s	38 GB/s	50-90 GB/s
	Latency	300 + ns	100 - ns	50 + ns
Mid-term (from 2012)	Bandwidth	45 GB/s ??	50-60 GB/s	80-300 GB/s
	Latency	300 - ns	50 + ns	50 - ns





Bull's position

- From a hardware viewpoint, Bull believes that current connection architecture for GPGPUs (i.e. the PCI-express) is both
  - A valid trade-off for the 2 years, or so, to come
  - An interim approach before solutions with a better architected memory connection appear

- PCI express connection (even with Gen3) creates a memory access bottleneck resulting either in
  - Limitation in the ability to take advantage of available processing performance (performance impact and market reach impact)
  - Duplication of central memory capabilities (cost impact)

We recommend to view the initial investment in the current technology as foundation work to reap full benefits with new architecture

Architect of an Open World™

LIBERATE IT