Interconnect et MPI

LIBERATE IT

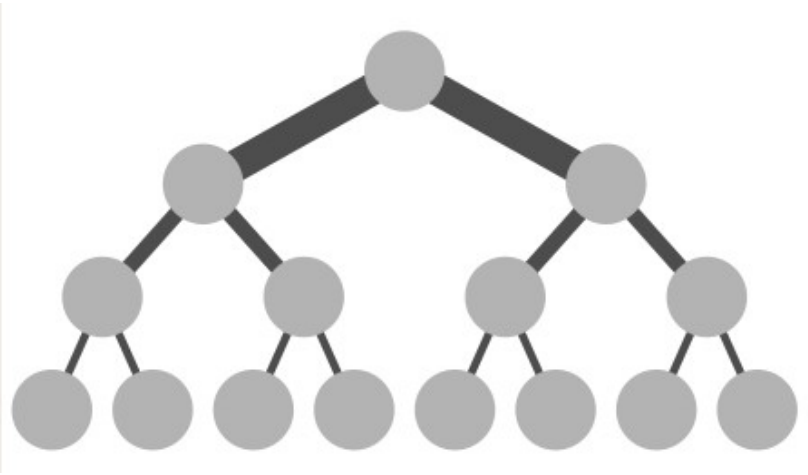
Réseau d'interconnection

- Hiérarchie de switches
  - Chaque passage dans un switch (hop) induit un coup (latence)
  - L'agencement des nœuds et des switches suit une topologie
  - Propriété d'un réseau : diamètre, bande passante de bissection
- Noeuds de calcul équipe de HCA
  - degré (ou *sortance*) des nœuds,
  - Bas niveau : latence, bande-passante, taux de messages (« issue rate »)
- Algorithme de routage
  - Statique, dynamique
  - Nombre de routes par lien ...
- Software
  - Switches/cartes : firmware
  - OS : drivers, bibliothèques de communication bas niveau
  - Applicatifs : MPI, Système de fichier (stockage)

Paramètres important d'un réseau

- **Diamètre** : le maximum des plus petits chemins entre deux noeuds (sur l'ensemble des paires possibles).
- **Largeur de bissection** : nombre de liens que l'on « coupe » en divisant un réseau en deux sous-parties égales.
- **Bande passante totale de bissection** : bande passante minimale soutenue disponible entre chacune des sous-parties.
- **Topologies courantes** : tore, fat-tree, fat-tree « prunée »

  - Ci-contre : fat-tree
    - Diamètre : N/2
    - Largeur : N
    - Bande-passante : 1



Performances bas niveau

Les plus communes :

- Latence
  - Ex. de performances bas niveau (perf. Infiniband QDR)
    - Même noeud, même HCA : 1.09-1.1 µsec
    - 2 noeuds, 1 hop : 1.25-1.26 µsec
    - 2 noeuds, 2 hops : 1.40-1.41 µsec
    - Latence de la carte : 1.1 µsec, latence d'un switch : 1.15 µsec
- Bande passante
  - Exprimée en GB ($8*10^3$ bits) ou GiB ($8*2^{20}$ bits)
  - La bande passante effective est généralement plus petite que la bande passante physique « vraie » du lien et ce à cause de l' « overhead » du protocole (protocole généralement 8b/10b).
    - QDR : 40 Gbits – bande passante effective 32 Gbits
    - D'autre sorte de contention peuvent exister (ex. PCIe)

Débit théorique

- À avoir en tête :
  - Les protocoles PCIe 1.0-2.0 et InfiniBand sont des protocoles 8b/10b
  - HCA Inifniband : PCIe 8x gen 1 ou 2

HCA				débit (MiB/s)	
Type	**bp**	**données**	**Interface PCIe**	**estimé**	**mesuré**
DDR	20 Gbits	1.8 GiB/s	PCIe-1 – 20 Gbits	1525	1525
			PCIe-2 – 40 Gbits	1907	1800
QDR	40 Gbits	3.7 GiB/s		3051	3093

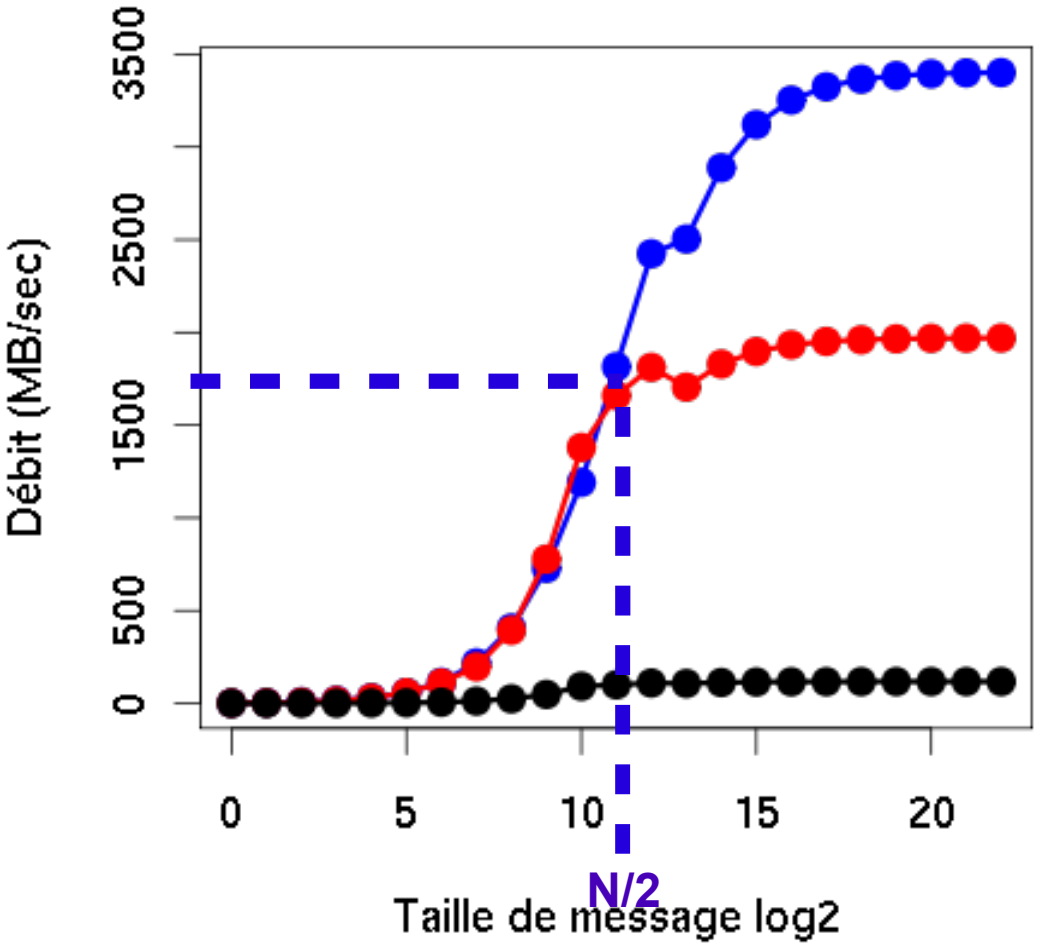
QDR : 40 Gbits soit 5 Gbytes soit 4 Gbytes effectifs = 3814 MiB/sec

MPI

LIBERATE IT

Performances point-à-point

- Indicateurs :
  - Latence
    - Petits messages
  - bande passante
    - Gros messages
  - N/2
  - taux de messages
- Permet de comparer les technos entre elles.

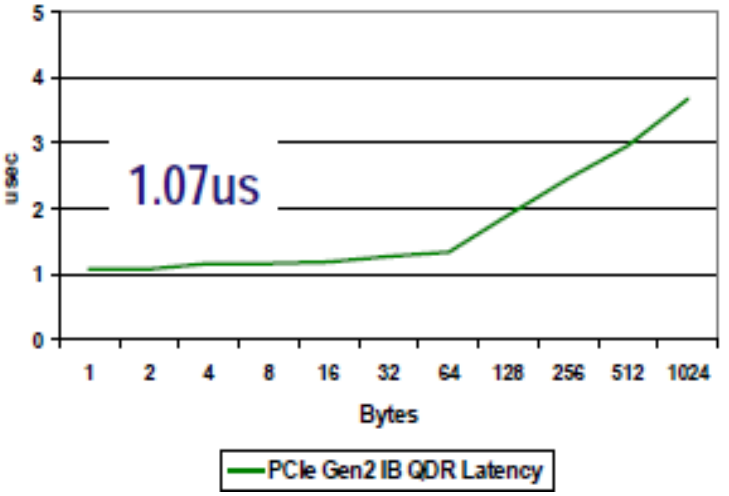


La bande passante ...

- D'un point de vue utilisateur applicatif : bande passante MPI
  - Les perf. bas niveaux sont intéressantes pour le stockage (SAN)
- Benchmarks élémentaires : PingPong, PingPing, SendRecv
  - Plusieurs implem. Disponibles : IMB (ex. pallas), OMB (OSU MPI Benchmarks) et les vôtres !
  - Bande passante d'un lien point-à-point (mono- et bidirectionnelle).
    - → Ne détecte pas de possible congestions : « pruning », routage ...
- Opérations point-à-point simultanées entre paires multiples
  - Montrent les contentions
- Attention : sur les SMP (!) un test sur l'ensemble des noeuds donne un résultat « moyenné » entre internoeud (interconnect) et intranoeud (« shared memory »).
  - Encore une fois, tout dépend ce que l'on souhaite mesurer
    - → 1 processus par coeur : c'est ce que va faire (en général) une appli. Utilisateur
    - → 1 processus par noeud : test l'interconnect eau travers de MPI.

… Mais pas que la bande passante

- La latence est une quantité primordiale pour les petits messages
  - Elle permet aussi de définir le taux de messages (~2 bytes)
- Les tests sur les opérations collectives sont intéressantes pour l'applicatif
  - Sensible à l'engorgement
  - MPI_Alltoall(v)
    - Code à base de FFT, par ex. les codes de chimie (cpmd)
- La taux de messages est une quantité intéressante : il fournit une indication sur la capacité de traitement de la carte.



IMB

- http://software.intel.com/en-us/articles/intel-mpi-benchmarks/
- Intel MPI Benchmarks (ex-PALLAS)
- Portage
  - C
  - Immédiat
- Exécution
  - Simple

OMB

- http://mvapich.cse.ohio-state.edu/benchmarks/
- OSU MPI Benchmarks
- Bien que livré par défaut avec les implémentations MVAPICH, il fonctionne avec toutes les souches
  - Complémentaire d'IMB
    - Certains test intéressant : message_rating ...
- Portage
  - C
  - Immédiat
- Exécution
  - Simple

Contenu de la boîte à outil « interconnect-MPI »

- Des tests interconnect bas niveau
  - Stack soft de l'interconnect
    - Débit, latence, trace_route et autres outils de diagnostique
- Des tests MPI
  - OMB
  - IMB
  - Tests
    - Point-à-point
    - Opérations Collectives et « message-rating »
    - Test fonctionnel de certaines « features » MPI
  - Le benchmarks MPI sont un domaine de prédilection pour les benchmarks maison !