Problèmes Parallelisation

By Professeur
Cédric Bilat
Cedric.Bilat@he-arc.ch

Bandwith

CPU-GPU

Problème

Mesurer la bande pasante en GB/s entre la ram (côté Host) et la gram (coté device). Sur le GPU on s'intéresse ici à la Global Memory.

Observation

Il n'y a pas besoin de kernel ici, mais juste d'effectuer du MM (memory management) entre le host et le device.

Idées

On peut s'intéresser ici à la bandwith

- Min
  - Max
  - Moyenne
  - Mediane

et donner des valeurs selon la direction de transfert

HostToDevice

ou

DeviceToHost

Contraintes

(C1) Utiliser une première fois de la ram *standard*, puis utiliser de la ***page-locked*** ram (coté host). Comparer les résultats !

(C2) Si vous avez plusieurs GPU, effectuer ces mesures pour chacun des GPU. Comparer les résultats !

(C3) Le délivrable est un code, et un document illustrant les résultats.

(C4) Calculer le taux de transfert en faisant varier le volume de données transférés. Etablissez un graphe du taux de transfert en fonction de ce volume. Utilisez l'api « Graphe » fournie pour obtenir de manière dynamique un visuel illustrant la bande passante.

Note

On peut chronométrer avec la classe fournis

ChronoOMPs

ou chronométrer avec une technique

cuda (see *bilat_cuda_practical_guide*).

Y a t-il des différences ?

Questions

Que peut-on dire du bus pci expresse du serveur cuda1 ?

- pci-express 16x16 gen2 ?
- pci-express 16x16 gen3 ?

Même question pour cuda2 !

Note :

Si le bus pci-express est 16x16 gen3, et que la carte n'est pas gen3-ready, alors c'est le plus faible qui dicte le débit !

Rappel

- pci-express 16x16 gen2 donne 8GB/s par sens
- pci-express 16x16 gen2 donne 16GB/s par sens

Variation

(V1) Utilisez deux threads coté host, et effectuez en parallèle une copie

HostToDevice
DeviceToHost

Le GPU accepte-t-il ces copies en parallèles ?
Les **streams** Cuda peuvent-ils venir à la rescousse ?

(V2) On pourrait aussi s'intéresse à la bande passante ram-ram

GPU-GPU

Problème

Même chose que ci-dessus, mais ici on s'intéresse au transfert entre deux GPU, copie de type

peer-to-peer (P2P), see *bilat_cuda_practical_guide*

End