Profiling des transferts de données en mémoire

Bande passante : quantité de données transférées par unité de temps.

Manque : Actual data throughput vs. effective bandwidth

Types de transferts de données

Entre RAM et GPU, interne au GPU (shared, globale, locale) et entre GPU et externe (GPU Directe).

Théorique : avec largeur du bus mémoire et fréquence du bus. Accessible via cudaGetDeviceProperties() (propriétés memoryClockRate et memoryBusWidth).

```
2.0 * memoryClockRate * (memoryBusWidth / 8) * 1.e-6
```

en GB/s (Go/s). Facteur 2.0 = double data rate par clock.

Influence de nombre de kernels, taille blocs et tableau, code de correction des erreurs, coallescence des données.

Comment mesurer la bande passante

Avec données de taille connu, mesure du temps total de transfert. Formule (Go/s) (R_B et W_B en octets par kernel) :

$$BW_{\text{Effective}} = \frac{(R_B + W_B)/10^9}{t}$$

Formule de calcul de la bande passante
Mesurer débit théorique et effectif et ratio.

Trouver le bottleneck mémoire

Création de 3 kernels : normal, math et mémoire. Si mémoire proche de 100% et supér.3.ieur à math, alors limitation à cause de la bande passante.

Vérifier les accès aux DRAM Banks

Coalescence des données, chargement par bloc.