M2 - Architecture et Programmation d'accélérateurs Matériels.

(APM 2016-2017)

TP3

Stream et Asynchronisme

julien.jaeger@cea.fr
patrick.carribault@cea.fr

Les objectifs de ce TP sont :
— Asynchronisme
— Exécutions concurrentes de kernels, notion de streams et d'events
— Copies asynchrones
— Opérations collectives
— Introduction à l'optimisation

I Streams, Events

Ouvrir le fichier *stream_ex1.cu*.

Q.1: Combien de streams sont utilisés dans ce programme ?

Q.2: Combien y a-t-il de flux d'exécution différents ?

Q.3: Classer les fonctions et les appels de kernels en deux catégories : les synchrones et les asynchrones.

Q.4: L'exécution des différentes fonctions CUDA au sein d'un stream est-elle ordonnée (i.e. séquentielle) ou bien désordonnée (Out-of-Order) ?

Q.5: Dessiner un schéma représentant les différentes tâches sur les flux d'exécution ainsi que leur dépendances.
Nous prendrons comme notation :
A $\rightarrow$ **B** : l'exécution de B ne peut commencer tant que celle de A n'est pas terminée.
A $\Rightarrow$ **B** : l'exécution de B ne peut se terminer tant que celle de A n'est pas terminée.

Liste des tâches :

Création S0	Création S1	Malloc Host	Thread Sync
requête cpy H→D [0,N[	requête cpy H→D [N,2N[	requête cpy D→H [0,N[	requête cpy D→H [N,2N[
cpy H→D [0,N[	cpy H→D [N,2N[	cpy D→H [0,N[	cpy D→H [N,2N[
requête K1 [0,N[	requête K1 [N,2N[	exécution K1 [0,N[	exécution K1 [N,2N[

Q.6: A quoi sert la commande `cudaMallocHost` ? Pourquoi ne pas utiliser la fonction `malloc` ?

Q.7: Utiliser la fonction système `gettimeofday` pour mesurer le temps passé dans l'exécution du programme CUDA.

Q.8: Avec les fonctions `gettimeofday` et `cudaThreadSynchronize` mesurer le temps d'exécution des kernels uniquement (sans compter le temps des transferts mémoire etc...).

Q.9: Quel est l'impact (inconvénient) sur l'exécution générale du programme de cette dernière façon de mesurer ? Vous pouvez vous aider du graphe de dépendance précédent et l'enrichir.

Q.10: Comment mesurer le temps de chaque kernel avec la technique précédente ?

Q.11: Comment utiliser les événements (fonctions `cudaEvent*`) pour mesurer le temps de chaque kernel et le temps total de tous les kernels ? Commentaire par rapport à la façon précédente ?