TP prise en main de CUDA : introduction à la programmation des GPU

P. Kestener

13 décembre 2019

Les TP pourront être faits sur l'une ou l'autre des deux plateformes suivantes :

- **le calculateur ROMEO [1], et son interface web ROMEOLAB (basé sur la technologie de notebook jupyter), accès ouvert jusque fin février 2020,**
- **sur le cloud Amazon (AWS), grâce au programme AWS Educate [2], accès ouvert jusque fin mai 2020.**

On préfèrera utiliser la plateforme ROMEOLAB pendant le cours, et éventuellement AWS après le cours.

1 Prise en main des outils de développement CUDA

1.1 Environnement Unix / Romelab

https ://romeolab.univ-reims.fr est une interface web (basée sur la technologie python / jupyter notebook) qui permet d'utiliser les resources et nœuds de calcul du supercalculateur ROMEO de l'université de Reims.

Activités :

- S'inscrire sur la plateforme RomeoLab.
- L'encadrant vous fournira l'identifiant de la session.
- Vérifier que vous pouvez lancer le notebook intitulé *DocumentationROMEOLAB*. Suivre les indications données pendant le TP.
- Se reporter à la section A.1.

1.2 Environnement Unix / AWS

On utilisera les stations de travail de la salle de TP pour se connecter à distance sur le cloud AWS (les accès nous sont fournis par Nvidia).

Activités :

- Récupérer les identifiants AWS, et se connecter sur le cloud
- S'inscrire sur la plateforme RomeoLab et vérifier que vous pouvez lancer une session Jupyter notebook sur Romeo.

L'objectif du TP est de se familiariser avec le flot de compilation CUDA/C/NVCC.

On pourra consulter les pages de manuel en ligne et/ou utiliser l'aide en ligne de CUDA : http://docs.nvidia.com/cuda/cuda-runtime-api/index.html

1. Merci à Arnaud Renard
2. Merci à Joe Bungo de Nvidia, DLI program manager.

1.3 Compilation du SDK CUDA/C++

Copiez les exemples du SDK CUDA/C++ (SDK = Software Development Kit), i.e. dans votre `HOME`, tapez la commande suivante [3] :

```
cuda-install-samples-10.1.sh .
```

Le SDK contient des exemples d'applications et de programmes en CUDA/C++. Chaque exemple peut être compilé indépendamment des autres en se plaçant dans le sous-répertoire correspondant et en tappant `make`.

1.4 deviceQuery en CUDA/C++ - Know your hardware

Tappez `cd /NVIDIA_CUDA-10.1_Samples/1_Utilities/deviceQuery` pour aller dans le répertoire source de cet exemple et ensuite `make`. Exécutez l'exemple `deviceQuery` qui permet d'avoir toutes les informations sur le matériel disponible [4].

On pourra constater ici que l'on utilise juste l'API CUDA pour interroger le driver de la carte graphique, mais qu'aucun *kernel* CUDA n'est compilé.

1. Exécuter `deviceQuery`.
2. De combien de GPU dispose-t-on sur la machine ?
3. De combien de *streaming multiprocessor* sont faits les GPU ?
4. Utiliser l'information sur le bus mémoire pour en déduire la valeur de la bande passante mémoire crête en GBytes/s.

Pour information, une liste à jour des GPU NVIDIA et leurs caractéristiques :
http://en.wikipedia.org/wiki/List_of_Nvidia_graphics_processing_units

1.5 HelloWorld en CUDA/C++

1.5.1 Utilisation des variables intrinsèques `threadIdx` et `blockIdx`

Activité 1 :

— Récupérer le code source du programme helloworld depuis les planches du cours

— Compiler `nvcc helloworld.cu -o helloworld`

— Quelle est la sortie de ce programme ?

— Que se passe-t-il si on commente la ligne contenant l'appel à `cudaDeviceSynchronize` ? Comment interprète-t-on se résultat ?

Activité 2 : Récupérez le fichier `helloworld/1/helloworld_block.cu`.
Manipulation :

1. Ouvrir le fichier. Que fait ce programme.
2. Utiliser les informations de l'entête du fichier pour le compiler et l'exécuter (en changant éventuellement le flag d'architecture).
3. Vérifier le bon fonctionnement lorsque vous changer la taille des blocks et la taille de la grille.
4. Dans quel ordre les messages sont affichés ? Peut-on l'expliquer ?
5. Modifier le code de façon à traiter des blocks 2D de taille 4 par 4.

3. Ce script bash est fournit par Nvidia, et disponible après avoir installé les outils CUDA
4. Complément d'information :
http://devblogs.nvidia.com/parallelforall/how-query-device-properties-and-handle-errors-cuda-cc/

1.5.2 Addition de scalaires / Addition de vecteur sur GPU

Le but de l'exercise est d'écrire un premier kernel CUDA et d'apprendre à utiliser les variables intrinsèques du modèle de programmation qui dimensionne la grille de bloc de *threads* et les blocs de *threads* eux-même.

On vous donne le code qui permet d'additionner deux scalaires sur GPU : `helloworld/2/helloworld.cu`.

La première ligne du fichier indique comment compiler le code. Dans cet exemple (le plus simple possible), on additionne deux scalaires sur le GPU.

1. Compiler le code source `helloworld.cu` avec le compilateur `nvcc`. Utiliser l'information donnée par l'exécution du code `deviceQuery` pour déterminer la version d'architecture matérielle du GPU.
   - Utiliser l'option `-arch=...` pour spécifier l'architecture ; e.g. `-arch=sm_30` pour les GPU Kepler
   - À quoi sert l'option `--ptxas-options -v` ?
2. Exécuter le programme avec les paramètres par défaut : les *kernels* `add` et `add2` avec la configuration 1 bloc de *threads*, 1 *thread* par bloc.
3. Manipulation :
   - Utiliser les slides du cours pour modifier le code.
   - Les variables `a` et `b` sont à présent des tableaux qu'il faut allouer et initialiser, de taille N (on pourra fixer $N = 16$ pour test dans un 1er temps). Le tableau `c` doit contenir la somme de `a` et `b`.
   - Écrire le code du nouveau *kernel* en utilisant les variables intrinsèques `threadIdx` et `blockIdx` pour addresser les cases mémoires des tableaux.
   - Configurer l'exécution du *kernel* pour avoir 1 bloc de *threads*, N *threads* par bloc. Afficher le résultat à l'écran.
   - Pour les petites valeurs de N (taille des tableaux), afficher à l'écran le résultat du calcul. Vérifier que le code écrit peut être exécuté de multiple configurations, par exemple en 2 blocs de *threads* avec N/2 *threads* par bloc (en donnant évidemment les mêmes résultats).

1.5.3 Gestion de la mémoire unifiée (facultatif)

Dans les exemples précédents, nous avons géré explicitement les allocations et les transferts mémoires avec respectivement les appels à `cudaMalloc` et `cudeMemcpy`.

Avec l'introduction de la mémoire dite unifiée, on peut délégué les transferts mémoires entre la carte mère et la carte graphique, en utilisant l'appel à `cudaMallocManaged` pour l'allocation mémoire. Avec cet appel on récupère un pointeur qui peut être utilisé à la fois sur le CPU et sur le GPU.

Exercice :

- Lire la page suivante https://devblogs.nvidia.com/unified-memory-cuda-beginners/
- Mettre en œuvre la mémoire unifiée sur le code proposé sur cette page.

1.5.4 Gestion des erreurs et profilage de code

1. - Le code `helloworld/3/helloworld.cu` contient une(des) erreurs. Pouvez-vous les corrigez en vous aidant des messages afficher à l'execution ?
   - Quel était le problème ?
2. - On utilise à présent le code `helloworld/3/helloworld2.cu`. Compiler et visualiser la trace temporelle d'exécution à l'aide de l'outil `nvvp` (Nvidia visual profiler). Quel commentaire peut-on faire ?
   - On apprend ici à utiliser les outils de profiling : `nvvp` (en mode graphique), `nsys` (en mode ligne de commande) et la bibliothèque `nvToolsExt` pour instrumenter le code CPU.

1.6 Bande passante mémoire : GPU-GPU et CPU-GPU

Dans un très grande classe de problèmes, le facteur limitant les performances d'un code GPU est l'utilisation de la bande passante mémoire, soit du bus Pci-Express (CPU/GPU) ou du bus mémoire entre le GPU et la mémore de la carte graphique.

On se concentre dans un premier temps sur la bande passante GPU-GPU [5] (i.e. entre le GPU et la mémoire de la carte graphique).

On pourra aussi consulter la documentation effective bandwidth calculation (CUDA best practice guide, section 8) qui explique clairement comment calculer la bande passante effective associée à un noyau CUDA donné.

1. Utiliser le code situé dans le répertoire `code/bandwidth`. Ce code contient plusieurs façons de copier un tableau dans un autre.
2. Ouvrir le fichier `bandwith.cu` et remplir les TODO.
   —
   — Dans un premier temps, on adopte la stratégie naïve : 1 *thread* par élement du tableau à copier. Ecrire le code du kernel `copy`.
   — Dans un deuxième temps, comment modifier le kernel CUDA si on considère que la taille de la grille de *thread* est de taille fixe (indépendante de la taille des tableaux) ? Ecrire le code du kernel `copy2` correspondant. On définira le nombre de blocs comme un multiple du nombre de *streaming multiprocessor*.

   ```
   // utiliser cudaGetDeviceProperties
   cudaDeviceProp prop;
   cudaGetDeviceProperties(&prop, 0);  // pour le device 0
   // utiliser prop.multiProcessorCount
   ```

   — Faites en sorte que le programme prenne en argument de la ligne de commande le nombre d'éléments des tableaux à allouer [6].
3. On utilise les *timer* pour mesurer les temps d'exécution et afficher la bande passante mémoire en GBytes par seconde.

   ```
   // exemple d'utilisation
   #include "CudaTimer.h"
   ...
   CudaTimer timer;
   timer.start();
   // do something
   timer.stop();
   // use timer.elapsed_in_second() to get elapsed time
   ```

4. Exécuter le code plusieurs fois pour différentes tailles de tableau et pour les 2 versions du kernel copy. Que constatez-vous sur la valeur de la bande passante mémoire ?
5. Comparer avec la bande passante maximale possible :

   ```
   // utiliser cudaGetDeviceProperties
   cudaDeviceProp prop;
   cudaGetDeviceProperties(&prop, 0);  // pour le device 0
   printf("  Peak Memory Bandwidth (GB/s): %f\n",
   2.0*prop.memoryClockRate*(prop.memoryBusWidth/8)/1.0e6);
   ```

On revisite l'étude de la bande passante mémoire GPU-GPU et CPU-GPU (via le bus Pci-Express).

1. Compiler l'exemple `bandwidthTest` du SDK CUDA/C et lire l'aide (`./bandwidthTest --help`)
2. Exécuter la version *release* avec l'option `--mode=quick`

5. A titre complémentaire, on pourra consulter le blog https://devblogs.nvidia.com/how-implement-performance-metrics-cuda-cc/
6. Utiliser la routine `atoi` pour convertir une chaîne de caractère en entier.

3. Vérifier les ordres de grandeur discutés en cours des 3 différentes bandes passantes mémoire.
4. Exécuter l'exemple avec l'option *range* pour des tailles de transfert de données comprises entre 0 et 100kB par pas de 10kB. Que constatez-vous ?

1.7 SAXPY en CUDA/C

SAXPY est une des fonctions de base que l'on trouve dans les librairies d'algèbre linéaire de type BLAS [7] qui réalise l'opération $\mathbf{y} = \alpha\mathbf{x} + \mathbf{y}$, où $\mathbf{x}$ et $\mathbf{y}$ sont des vecteurs 1D de réels simple précision.

Copier le répertoire `/opt/local/TP_CUDA/saxpy_cuda_c`. Utiliser le Makefile pour compiler l'exemple. La compilation fournit un exécutable `saxpy` qui calcule la fonction *saxpy* de 4 manières :

1. version séquentielle sur le CPU
2. version parallèle OpenMP sur le CPU
3. version parallèle sur le GPU avec kernel CUDA écrit *à la main*
4. version parallèle sur le GPU en utilisant les routines de la librairie cuBlas [8]

Comparer [9] les performances des 4 versions en faisant varier la taille du tableau d'entrée (paramètre `N` en début du fichier).

— Que constatez-vous lorsque la taille du vecteur est de l'ordre de quelques dizaines de milliers ? Pouvez-vous l'expliquer ? [10]

— Que constatez-vous lorsque la taille du vecteur est de l'ordre de quelques millions à dizaines de millions ? Pouvez-vous l'expliquer ?

— On pourra essayer de tracer l'allure grossière de l'évolution des performances en fonctions de la taille du vecteur pour les quatres variantes.

Complément sur les métriques de mesure de performance :
https://devblogs.nvidia.com/how-implement-performance-metrics-cuda-cc/

Exercice :

— Lire la page :
https://devblogs.nvidia.com/cuda-pro-tip-write-flexible-kernels-grid-stride-loops/

— Modifier le code source du kernel `saxpy_parallel` pour implanter la variante dite *grid-stride-loop*.

— Quels sont les avantages de cette variante par rapport à la version dite monolithique ?

1.8 Manipulation en mémoire partagée

On utilisera le code du répertoire `code/transposition`

Des explications sur les notions d'accès coalescent à la mémoire et de conflit de banc mémoire seront données pendant le TP.

On pourra utiliser cet exemple pour s'initier aux outils développeur `nsight-sys`. On adaptera les *vieilles planches* suivantes :
https://www.olcf.ornl.gov/wp-content/uploads/2013/01/Hands-On-CUDA-Optimization1.pdf

1.9 Algorithmes de reduction

Les algorithmes de reduction (e.g. somme des éléments d'un tableau) sont au cœur de la plupart des applications de calcul scientifiques. La mise en œuvre d'une implémentation parallèle sur GPU n'est du tout triviale.

On va commencer par revisiter les planches de Mark Harris. Voir le répertoire `code/reduction`

Ensuite, placez-vous dans le répertoire des sources de l'exemple `reduction` du SDK CUDA/C++ (`6_Advanced/reduction`).

7. http://en.wikipedia.org/wiki/SAXPY
8. cuBlas est fournie par NVIDIA en installant le toolkit CUDA. Cf /usr/local/cuda-10.1/doc/pdf/CUDA_CUBLAS_Users_Guide.pdf
9. Attention cette étude dépendant TRÈS fortement de la plateforme matérielle utilisée (laptop, desktop or supercalculateur).
10. Utiliser les informations de la commande lstopo pour vous aider à interpréter les résultats sur CPU.

1. Éditez le code source pour comprendre comment appeller les différentes versions de `kernel`.
2. Ouvrez le document PDF (sous répertoire `doc` dans les sources) pour avoir des explications claires sur les différents `kernel`.
3. Le document PDF nous informe que la bande passante mémoire maximale entre le GPU et sa SDRAM externe est de 86.4 GBytes/s pour le GPU FeForce GTX 8800 (toute première version de l'architecture CUDA, fin 2006). Quelle la valeur correspondante pour notre GPU (sur AWS) ? Utiliser les informations de l'exemple `deviceQuery` (celui du SDK) pour connaître la banda passante théorique maximale.
4. Executer l'exemple réduction pour les différents *kernels*, retrouve-t-on les chiffres indiqués dans le document ?

On pourra compléter cet exercice par la lecture du chapitre 12 du livre de Nicholas Wilt : http ://www.cudahandbook.com/.

2 Calcul de type stencil / Équation de la chaleur

On se propose de résoudre l'équation de la chaleur (cf annexe D) par la méthode des différences finies sur une grille cartésienne en 2D puis 3D en explorant plusieurs variantes d'implantations avec CUDA.

L'équation de la chaleur représente l'archétype du problème parallélisable [11] par décomposition de domaine. Le travail proposé permet d'explorer les diverses façons d'implanter cette décomposition dans le modèle de programmation CUDA.

1. Décompresser l'archive `TP_CUDA/chaleur/heat2d3d_ensta2013_sujet.tar.gz` sur votre compte local.
2. Tout au long de l'exercice, il faudra éditer le `Makefile` pour permettre la compilation des différentes variantes.
3. Une version de référence en C++ est fournie. Elle est constituée de 4 fichiers :
   (a) `heat_solver_cpu.cpp` : contient le `main`, les allocations mémoire et les sorties dans des fichiers pour visualisation,
   (b) `heat_kernel_cpu.cpp` : les routines de résolution du schéma numérique (à l'ordre 2 et à l'ordre 4),
   (c) `param.cpp` : definition des structures de données pour paramètres du problème (taille des tableaux, nombre de pas de temps, sortie graphique, etc...),
   (d) `heatEqSolver.par` : exemple de fichier de paramètres (utilisant le format GetPot)
   (e) `misc.cpp` : les routines utiles annexes (initialisation des tableaux).
4. Compiler, exécuter la version de référence avec les valeurs par défaut des paramètres et de la condition initiale.
5. Reprendre la question précédente en modifiant la condition initiale/condition de bord (fichier `misc.cpp`, routine `initCondition2D`, mettre par exemple tous les bords à 0 sauf un bord à 1) et en calculant 1000 pas de temps avec une sortie graphique tous les 100 pas de temps sur un domaine 2D de taille 256×256. Visualiser les résultats (images PNG ou fichiers VTK) en vous aidant des informaions contenues dans le sous-répertoire `visu`.

On se propose de porter cet algorithme sur GPU graduellement en terme d'optimisation.

version naïve 2D Dans un premier temps, on n'utilisera pas la mémoire partagée du GPU, tous les accès mémoire se feront à partir des tableaux situés en mémoire globale.

11. Voir le site `http://www.cs.uiuc.edu/homes/snir/PPP/` pour avoir plus d'informations sur les différents archétypes de problèmes parallèles

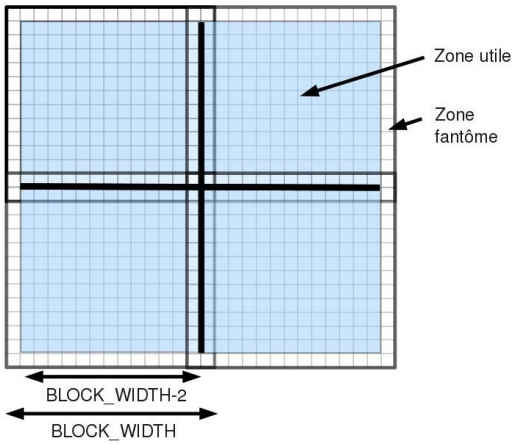

FIGURE 1 – Schéma représentant la grille de blocs de threads utilisée dans le kernel défini dans le fichier `heat2d_kernel_gpu_shmem1.cu` ; les blocs de thread adjacent se recouvre sur une zone de largeur 2.

1. Editer le fichier `heat2d_solver_gpu_naive.cu`, et remplir de façon appropriée les endroits signalés par `TODO`. Consulter la documentation en ligne de CUDA pour savoir comment utiliser les routines `cudaMalloc` et `cudaMemcpy` [12].
2. Editer de même le fichier `heat2d_kernel_gpu_naive.cu` qui contient le code du *kernel* exécuté sur le GPU
3. Editer le `Makefile` et décommenter les lignes correspondantes à la compilation de cette version.
4. Vérifier que le code est fonctionnel (en comparant aux résultats de la version de référence).
5. Comparer les performances de cette première version sur des tableaux qui ont des tailles en puissance de 2 et non-puissance de 2.

version simple avec mémoire partagée 2D Reprendre les questions précédentes en utilisant les fichiers `heat2d_solver_gpu_shmem1.cu` et `heat2d_kernel_gpu_shmem1.cu`. La mémoire coté GPU est à présent allouée avec les routines `cudaMallocPitch` (voir la documentation de l'API CUDA : `/usr/local/cuda/doc/html/group__CUDART__MEMORY.html`) pour respecter les alignements mémoire. On se propose dans cette version d'utiliser la mémoire partagée (meilleurs temps d'accès) ; en revanche la mémoire partagée étant privée à un bloc de *threads*, il est nécessaire d'utiliser un découpage du domaine en blocs qui se chevauchent (voir la figure 1) pour assurer la continuité de l'accès aux données (voir les explications données pendant le TP). On pourra s'aider de ce schéma pour déterminer quel *thread* accède à quelle case mémoire.

1. Après avoir testé cette version, que se passe-t-il en terme de performance si on échange les rôles des indices de *thread* `tx` et `ty` ? Expliquez.

version simple avec mémoire partagée 2D - variante Il s'agit d'une légère variante de la version précédente. A présent, on alloue un tableau en mémoire partagée plus grand que la taille des blocs de threads pour tenir compte des cellules *fantômes* qui permettent aux blocs de threads de travailler sur des blocs complètement indépendants. Ne pas hésiter à demander des explications pendant le TP.

version optimisée avec mémoire partagée 2D (facultatif) Afin de profiter au maximun de la copie en mémoire partagée, on demande à chaque *threads* de calculer plusieurs cellules du tableau de sortie. Le *kernel* est divisé en 2 *sous-kernels* travaillant respectivement sur les lignes puis les colonnes.

12. Voir la doc en ligne CUDA : `/usr/local/cuda-8.0/doc/html/index.html` sur la machine `rhum`

version 3D naïve Les limitations de CUDA font que l'on ne peut pas créer des grilles de bloc de threads de n'importe quelle dimension dans la direction `z`. On se propose de reprendre la version 2D naïve et de l'étendre en 3D, chaque *thread* s'occupant de toute une colonne suivant `z`

1. Ecrire le code du *kernel* `heat3d_ftcs_naive_kernel` dans le fichier `heat3d_kernel_gpu_naive.cu`
2. Vérifier qu'il est fonctionnel en comparant les résultats avec ceux de la version de référence.

version 3D optimisée On reprend la version 2D qui utilise la mémoire partagée. La partagée ayant une taille maximale qui ne permet pas d'allouer dans la direction `z` la même taille que dans les directions `x` et `y`, on se contente d'allouer le tableau suivant :

```
__shared__ float shmem[3][BLOCK_HEIGHT][BLOCK_WIDTH];
```

qui contient les données de 3 plans en `z` consécutifs. On peut ainsi accéder à toutes les cases mémoires voisines nécessaires à la mise à jour de $\phi_{i,j,k}^{n+1}$ (cf Eq. (4)). A chaque fois qu'un plan est calculé complètement (dans un bloc), on avance en permuttant les indexes des plans (`shmem[z]`) et en chargeant les données du plan suivant dans `shmem[3]`.

1. Remplir les trous laissés dans le *kernel* `heat3d_ftcs_sharedmem_kernel` dans le fichier `heat3d_kernel_gpu_shmem1.cu`; vérifier la fonctionalité du programme et tester ses performances.
2. On pourra également développer une version où seul le plan médian est stocké en mémoire partagée, les données des plans $z-1$ et $z+1$ étant mise dans des registres (variables locales du thread courant). Cette version présente l'avantage de mieux utiliser les resources matérielles (**équilibrer registres et mémoire partagée**).

Pour aller plus loin...

Etude de l'influence de certains paramètres sur les performances CPU/GPU.

1. impact de l'ordre du schéma numérique (ordre 2 ou 4).
2. impact de la double précision (ajouter le symbol `USE_DOUBLE` aux flags de compilation, voir le Makefile en tête de fichier)
3. impact de la dimension des tableaux de simulation
   - Essayer la version 3D naïve avec des tailles de tableaux en puissance de 2 et non-puissance de 2.
   - Faîtes la même chose avec la version 3D en mémoire partagée. Que constatez-vous ?
4. impact de la dimension des blocs de *threads* et de leur forme (bloc carré ou alongé suivant x)

3 Tutoriel OpenACC

Utilisation du matériel pédagogique :
https://github.com/eth-cscs/SummerSchool2019/tree/master/topics/openacc
Sur RomeoLab, on pourra consulter le lab `GPUBootCampGTC2019OpenACC`

4 Activités complémentaires

4.1 Initiation à cmake pour un projet CUDA

- Voir le project template https://github.com/pkestene/cuda-proj-tmpl
- Voir le code source du projet LBM / C++

4.2 Initiation à python/cuda

Deux situations pratiques :

— si on a une application existante écrite en python, que l'on souhaite accélérer en portant quelques noyaux de calcul en CUDA, on préfèrera utiliser numba, CuPy ou pycuda

— si on a une grande quantité de code existante, écrite en CUDA/C++ et que l'on veut les utiliser depuis python, on préfèrera utiliser une bibliothèque de bindings comme pybind11, cython ou SWIG

4.2.1 Tutoriel Numba

Cf https://github.com/ContinuumIO/gtc2019-numba

Sur Romeo, ces notebook sont déjà installés, le lab s'appelle `gtc2019-numba`. Attention, CuPy n'est pas installé, le 2eme notebook ne fonctionnera pas.

Sur votre laptop, assurez-vous d'avoir les prérequis suivants :

— Cuda

— Miniconda3

— Jupyter : `conda install jupyter`

— CuPy : `pip install cupy-cuda101` (using pip from Miniconda)

```
# Download notebooks
git clone https://github.com/ContinuumIO/gtc2019-numba.git
cd gtc2019-numba
# Start jupyter notebook
jupyter notebook
```

4.2.2 CMake / Cuda/C++ / pybind11

Exemple de projet *template* : https://github.com/pkestene/pybind11-cuda

4.3 Cuda/C++ / cython / swig

Exemple de projet *template* : https://github.com/pkestene/npcuda-example

5 Exécuter le tutoriel Python/Numba sur une instance AWS

Voici les informations minimale pour exécuter le tutoriel Python/Numba sur une instance AWS

1. Lancer une image sur une instance EC2 avec resources GPU, e.g.
   — image : `Deep Learning Base AMI (Ubuntu 18.04) Version 21.0`
   — instance : `g4dn.xlarge`
2. Se connecter sur la machine (avec votre clé ssh sauvegardée préalablement) :
   — `ssh -o "IdentitiesOnly=yes" -i  /.ssh/myEC2key.pem ubuntu@XXX.YYY.ZZZ.KKK`
3. Récupérer les sources du tutoriel
   — `git clone https://github.com/ContinuumIO/gtc2019-numba.git`
4. Lancer le serveur Jupyter (sans interface web locale)
   — `cd gtc2019-numba`
   — `jupyter notebook --no-browser`
5. Sur votre machine locale (laptop ou autre), sous linux, créer un tunnel ssh vers l'instance EC2 (cela redirige le port local 9999 vers le port 8888 de la machine distante, où s'exécute le notebook) :
   — `ssh -i  /.ssh/myEC2key.pem -NfL 9999:localhost:8888 XXX.YYY.ZZZ.KKK`
6. Ouvrir le notebook et suivre le tutoriel
   — firefox http ://localhost :9999

6 Compléments

6.1 Programmation multi GPU

On pourra utiliser le code suivant https://github.com/NVIDIA/multi-gpu-programming-models

A Accès aux machines

A.1 ROMEOLAB

https://romeolab.univ-reims.fr/
But : lancer un bureau virtuel sur un nœud Romeo :

— Se connecter à romeolab : https ://romeolab.univ-reims.fr

— Ouvrir par exemple le lab `Atelier Romeo - Introduction à Romeo`

— Exécuter la commande `display_vnc()` ; au bout de quelques secondes apparaît en dessous de la cellule un lien hypertexte intitulé `Click here to open VNC access in new frame.` ; cliquer sur ce lien, cela ouvre un bureau virtuel dans un nouvel onglet de votre navigateur.

— Ouvrir un terminal : cliquer sur le bouton `Applications Menu -> Terminal Emulator`

— Vous êtes prêt pour le TP.

A.2 Amazon webservice (AWS)

Pendant le cours, on vous donnera les identifiants des comptes étudiants AWS ; ils ont été crédités de 125 dollars par NVidia. **Faites bien attention à ne pas dépasser ce plafond ! ! !**
On fera un tutoriel au début du cours pour lancer une machine virtuelle avec un GPU activé.
Référence complémentaire : http://cs231n.github.io/aws-tutorial/

A.3 Autres resources en ligne pour le calcul sur GPU

— Kaggle (https://www.kaggle.com/) : vous pouvez créer des notebooks jupyter/python et accéder à des GPU de type P100, 30 heures de calcul gratuites par semaine.

— Google Colab (ttps://colab.research.google.com) est un service de Cloud gratuit avec des resources GPU ; voir par exemple https://colab.research.google.com/notebooks/gpu.ipynb ; vous aurez accès sans doute à des GPU de type K80.

B API CUDA

On pourra se servir de la documentation officielle de l'API CUDA :
/usr/local/cuda/doc/html/index.html

C Editeur de texte

Pour avoir la coloration syntaxique du c++ dans emacs sur les fichiers d'extension .cu, tapper : `Echap-x c++-mode` et `entrée`.

D Schéma numérique pour l'équation de la chaleur

Dans la section 2, on utilise la méthode des différences finies et plus précisément un schéma explicite de type FTCS [13] pour résoudre l'équation de la chaleur

$$\partial_t \phi = D \left[\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} \right], \; 0 \leq x \leq L_x, \; 0 \leq y \leq L_y, \; t \geq 0 \tag{1}$$

13. Forward Time Centered Space

sur un domaine rectangulaire muni de conditions de bord et d'une condition initiale.

On étudie deux versions du schéma, qui correspond à deux approximations de la dérivée seconde :

— différence centrée à 3 points, ordre 2 (erreur en h^2)

$$\begin{aligned} D^2\phi(x_0) &= \frac{1}{h^2}\left[\phi(x_0-h) - 2\phi(x_0) + \phi(x_0+h)\right] \\ &= \phi''(x_0) + \frac{1}{12}h^2\phi^{(4)}(x_0) + O(h^4) \end{aligned}$$

— différence centrée à 5 points, ordre 4 (erreur en h^4)

$$\begin{aligned} D^2\phi(x_0) &= \frac{1}{12h^2}\left[-\phi(x_0-2h) + 16\phi(x_0-h) - 30\phi(x_0) + 16\phi(x_0+h) - \phi(x_0+2h)\right] \\ &= \phi''(x_0) + \frac{1}{90}h^4\phi^6(x_0) + O(h^6) \end{aligned}$$

Le schéma FTCS (à 3 points) met à jour le tableau 2D $\phi_{i,j}$ (ou 3D $\phi_{i,j,k}$) au temps $t = t_{n+1} = (n+1)\Delta t$ par la relation :

$$\phi_{i,j}^{n+1} = \phi_{i,j}^n + D\frac{\Delta t}{\Delta x^2}\left[(\phi_{i+1,j}^n - 2\phi_{i,j}^n + \phi_{i-1,j}^n) + (\phi_{i,j+1}^n - 2\phi_{i,j}^n + \phi_{i,j-1}^n)\right] \quad (2)$$

$$= R_2\phi_{i,j}^n + R\left[\phi_{i+1,j}^n + \phi_{i-1,j}^n + \phi_{i,j+1}^n + \phi_{i,j-1}^n\right] \quad (3)$$

où $R = D\Delta t/\Delta x^2$ et $R_2 = 1 - 4R$ (en 2D). De manière similaire en 3D, on trouve

$$\phi_{i,j,k}^{n+1} = R_3\phi_{i,j,k}^n + R\left[\phi_{i+1,j,k}^n + \phi_{i-1,j,k}^n + \phi_{i,j+1,k}^n + \phi_{i,j-1,k}^n + \phi_{i,j,k+1}^n + \phi_{i,j,k-1}^n\right] \quad (4)$$

avec $R_3 = 1 - 6R$ en 3D.

N.B. : En pratique, pour résoudre l'équation de la chaleur, on préfére utiliser un schéma implicite comme celui de Crank-Nicolson, qui conduit à l'inversion d'un système linéaire tridiagonal et qui possède l'avantage d'être inconditionnellement stable.

On pourra consulter la réference suivante :

— *Finite Difference Methods for Ordinary and Partial Differential Equations*, R. J. LeVeque, SIAM, 2007.
http://www.amath.washington.edu/~rjl/booksnotes.html