# Devoir No. 6 et Lab. #6

### PRÉ-REQUIS

- Vous devez avoir complété le Pré-Lab 3 et lu les chapitres 2 à 5 de la référence [2]
- À faire avant de commencer le lab (Voir annexe A)
  - 1. Ajouter "cl.exe" aux variables d'environnement du système

# PARTIE 1 - MULTIPLICATION MATRICE-MATRICE

**Objectif**: Écrire une multiplication matricielle C = A x B. Par souci de simplicité, nous allons uniquement utiliser des matrices carrées dont les éléments sont des single-precision floating- point.

#### Méthodologie:

Suivez les instructions ci-dessous :

A. Écrire une fonction "host-stub "qui allouera la mémoire pour les matrices d'entrées et la matrice de sortie, fera le transfert de ces matrices vers le Device, lancera le kernel, fera le transfert de la matrice de sortie vers l'Host et libèrera la mémoire des matrices d'entrées et de sortie sur le Device. Laissez les configurations d'exécution vide pour cette étape.

Cette fonction a quelque peu changé depuis mon laboratoire #5:

```
∃cudaError_t multiplyWithCuda(float* c, const float* a, const float* b, unsigned int size, kernelType type)
     float* dev_a = 0;
     float* dev_b = 0;
    float* dev_c = 0;
cudaError_t cudaStatus;
    cudaStatus = cudaSetDevice(0);
     if (cudaStatus != cudaSuccess) {
        fprintf(stderr, "cudaSetDevice failed! Do you have a CUDA-capable GPU installed?");
        goto Error;
    cudaStatus = cudaMalloc((void**)&dev_c, (size * size) * sizeof(float));
     if (cudaStatus != cudaSuccess) {
        fprintf(stderr, "cudaMalloc failed!");
        goto Error;
    cudaStatus = cudaMalloc((void**)&dev_a, (size * size) * sizeof(float));
     if (cudaStatus != cudaSuccess) {
        fprintf(stderr, "cudaMalloc failed!");
    cudaStatus = cudaMalloc((void**)&dev_b, (size * size) * sizeof(float));
    if (cudaStatus != cudaSuccess) {
        fprintf(stderr, "cudaMalloc failed!");
        goto Error;
    cudaStatus = cudaMemcpy(dev_a, a, (size * size) * sizeof(float), cudaMemcpyHostToDevice);
    if (cudaStatus != cudaSuccess) {
        fprintf(stderr, "cudaMemcpy failed!");
        goto Error;
     cudaStatus = cudaMemcpy(dev_b, b, (size * size) * sizeof(float), cudaMemcpyHostToDevice);
     if (cudaStatus != cudaSuccess) {
        fprintf(stderr, "cudaMemcpy failed!");
         goto Error;
```

```
aunch a kernel on the GPU
   dim3 threadsParBlock(size, size, 1);
   dim3 nombreDeBlock(1, 1, 1);
    switch (type) {
       case KERNEL_ELEMENT:
           multiplyKernelElement << <nombreDeBlock, threadsParBlock >> > (dev_c, dev_a, dev_b, size);
       case KERNEL_ROW:
           multiplyKernelRow << <nombreDeBlock, threadsParBlock >> > (dev_c, dev_a, dev_b, size);
       case KERNEL_COLUMN:
           multiplyKernelColumn << <nombreDeBlock, threadsParBlock >> > (dev_c, dev_a, dev_b, size);
else if (size >= minMatrixWidth && size <= maxMatrixWidth) {</pre>
   dim3 threadsParBlock(size, size, 1);
dim3 nombreDeBlock((size / blockWidth), (size / blockWidth), 1);
    switch (type) {
       case KERNEL ELEMENT:
           multiplyKernelElement << <nombreDeBlock, threadsParBlock >> > (dev_c, dev_a, dev_b, size);
           break:
       case KERNEL ROW:
           multiplyKernelRow << <nombreDeBlock, threadsParBlock >> > (dev_c, dev_a, dev_b, size);
           break:
       case KERNEL COLUMN:
           multiplyKernelColumn << <nombreDeBlock, threadsParBlock >> > (dev_c, dev_a, dev_b, size);
           break:
       default:
           break:
else {
   printf("Undefined behaviour. Please use matrix size between 8 and 64\n");
    return cudaErrorInvalidValue;
  else {
      printf("Undefined behaviour. Please use matrix size between 8 and 64\n");
      return cudaErrorInvalidValue;
 cudaStatus = cudaGetLastError();
  if (cudaStatus != cudaSuccess) {
      fprintf(stderr, "addKernel launch failed: %s\n", cudaGetErrorString(cudaStatus));
      goto Error;
  cudaStatus = cudaDeviceSynchronize();
  if (cudaStatus != cudaSuccess) {
      fprintf(stderr, "cudaDeviceSynchronize returned error code %d after launching addKernel!\n", cudaStatus);
      goto Error;
  cudaStatus = cudaMemcpy(c, dev_c, (size * size) * sizeof(float), cudaMemcpyDeviceToHost);
  if (cudaStatus != cudaSuccess) {
      fprintf(stderr, "cudaMemcpy failed!");
      goto Error;
 cudaFree(dev_c);
 cudaFree(dev_a);
  cudaFree(dev b);
  return cudaStatus;
```

Comme vous pouvez constater, dépendamment de la grosseur de la matrice entrée, le kernell sera calculé différemment. Dans ce programme, je peux faire la multiplication matricielle sur des matrices de 4\*4, et tout autre matrice carrée qui ont une dimension entre 8 et 64. Si une taille plus grande est rentrée, un message est écrit pour indiquer que ce calcul de matrice n'est pas pris en compte dans ce programme.

Pour rendre ce code plus agréable à lire, j'ai créé un enum au début de mon programme pour indiquer les différents types de calcul il est possible de faire avec les threads.

Lorsqu'on appelle la fonction multiplyWithCuda, il est donc important de spécifier quel type d'opération qu'on veut faire tel qu'énoncé dans le cahier dans le laboratoire.

J'ai aussi créé des variables constantes nous permettant de rapidement changer différentes fonctionnalités du programme :

```
const int blockWidth = 8;
const int minMatrixWidth = 8;
const int maxMatrixWidth = 64;
```

J'ai aussi corrigé l'allocation de la mémoire pour correspondre au float.

B. Écrivez un kernel où chaque thread calculera un seul élément de la matrice de sortie. Ajouter les paramètres d'exécution dans la fonction « Host-Stub » nécessaire à ce kernel.

```
global__ void multiplyKernelBow(float* c, const float* a, const float* b, int length)

{
   int row = (blockIdx.y * blockDim.y) + threadIdx.y;
   int column = (blockIdx.x * blockDim.x) + threadIdx.x;
   int indexmatC = (row * length) + column;

   int linearIndexRow;
   int linearIndexColumn;

   float sum=0;
   //Cette condition permet de rester dans les limites
   if (row < length && column < length){
        for (int index = 0; index < length; index++) {
            linearIndexRow = (row * length) + index;
            linearIndexColumn = (index* length) + column;
            sum += a[linearIndexRow] * b[linearIndexColumn];
        }
        c[indexmatC] = sum;
   }
}</pre>
```

Cette fonction permet de calculer un élément de la matrice de sortie. Contrairement au devoir 5, chaque thread doit quand même faire une loop car la matrice de sortie est une somme de toutes les multiplications rangées par colonnes de chaque élément des matrices A et B.

C. Écrivez un kernel où chaque thread calculera tous les éléments d'une ligne de la matrice de sortie. Ajouter les paramètres d'exécution dans la fonction « Host-Stub » nécessaire à ce kernel.

```
global__ void multiplyKernelRow(float* c, const float* a, const float* b, int length)

{
    int row = (blockIdx.y * blockDim.y) + threadIdx.y;
    int indexMatC = (row * length);

    int linearIndexRow;
    int linearIndexColumn;

//Cette condition permet de rester dans les limites

if (row < length) {

    for (int column = 0; column < length; column++) {
        float sum = 0;
        for (int index = 0; index < length; index++) {
            linearIndexRow = (row * length; length) + index;
            linearIndexColumn = (index * length) + column;
            sum += a[linearIndexRow] * b[linearIndexColumn];
        }
        c[(indexMatC + column)] = sum;
}
</pre>
```

Pour faire ceci, il est inévitable de faire une loop imbriqué dans une autre loop, car on doit calculer tous les élélements de la matrice de sortie. En tant que tel, le calcul reste très similaire que la question précédente dans la loop imbriqué, sauf que la première loop va garder la colonne respectif de la matrice de sortie, car on doit calculer une rangée complète dans ce thread.

D. Écrivez un kernel où chaque thread calculera tous les éléments d'une colonne de la matrice de sortie. Ajouter les paramètres d'exécution dans la fonction « Host-Stub » nécessaire à ce kernel.

Cette fois-ci, l'implémentation est très similaire que la question précédente, sauf qu'il faut calculer une colonne pour la matrice de sortie. Pour faire ceci, mathématiquement il faut calculer chaque rangée de la matrice A pour une colonne de la matrice B, et ceci donnera une colonne de la matrice C.

E. Analysez les avantages et inconvénients de chaque kernel ci-dessus.

Comme il est possible de voir dans les captures d'écran plus bas, faire un calcul élément par élément est énormément plus rapide que le calcul colonne ou rangée. Autre que cela, le calcul par colonne versus par rangée ne présente un avantage si significatif en terme de rapidité d'éxecution dans ce contexte

#### Pour A à E,

- A. Validez que le kernel est fonctionnel et effectue le bon calcul en utilisant un simple programme en CUDA C où :
  - a. Les matrices doivent avoir une taille variable 4x4, 8x8, 32x32 et 64x64.

#### Matrice 4\*4

Élément :

```
// Add vectors in parallel.
cudaError_t cudaStatus = multiplyWithCuda(c, a, b, arraySize, KERNEL_ELEMENT);
Matrix A:
 0.01 2.25 0.77 3.23 }
  2.34 1.92 1.40 3.58
  3.29 2.99 0.70 3.44
  2.84 2.05 1.22 0.06 }
Matrix B:
  0.37 1.46 0.59 0.66 }
  3.95 1.78 0.48 0.02 }
  0.04 1.51 2.13 2.28
  2.41 2.43 0.66 2.65 }
Matrix C:
 16.73 13.05 4.87 10.39 }
 17.12 17.66 7.66 14.30
 [ 21.31 19.52 7.13 12.94 }
 [ 9.35 9.79 5.28 4.86 }
Rangée :
```

```
// Add vectors in parallel.
cudaError_t cudaStatus = multiplyWithCuda(c, a, b, arraySize, KERNEL_ROW);
```

```
Matrix A:
 0.01 2.25 0.77 3.23 }
 2.34 1.92 1.40 3.58 }
 3.29 2.99 0.70 3.44 }
[ 2.84 2.05 1.22 0.06 }
Matrix B:
{ 0.37 1.46 0.59 0.66 }
3.95 1.78 0.48 0.02 }
[ 0.04 1.51 2.13 2.28 }
{ 2.41 2.43 0.66 2.65 }
Matrix C:
[ 16.73 13.05 4.87 10.39 }
{ 17.12 17.66 7.66 14.30 }
{ 21.31 19.52 7.13 12.94 }
{ 9.35 9.79 5.28 4.86 }
```

#### Colonne:

```
// Add vectors in parallel.
cudaError_t cudaStatus = multiplyWithCuda(c, a, b, arraySize, KERNEL_COLUMN);
Matrix A:
{ 0.01 2.25 0.77 3.23 }
[ 2.34 1.92 1.40 3.58 }
3.29 2.99 0.70 3.44 }
2.84 2.05 1.22 0.06 }
Matrix B:
[ 0.37 1.46 0.59 0.66 }
3.95 1.78 0.48 0.02 }
{ 0.04 1.51 2.13 2.28 }
[ 2.41 2.43 0.66 2.65 }
Matrix C:
[ 16.73 13.05 4.87 10.39 }
[ 17.12 17.66 7.66 14.30 }
{ 21.31 19.52 7.13 12.94 }
{ 9.35 9.79 5.28 4.86 }
```

#### Matrice 32\*32

#### Élément:

```
// Add vectors in parallel.
cudaError_t cudaStatus = multiplyWithCuda(c, a, b, arraySize, KERNEL_ELEMENT);
```

#### Matrice A:

```
(8.01) 2.25 6.77 3.23 2.34 1.92 1.48 3.58 3.29 2.99 0.78 3.44 2.84 2.05 1.22 0.06 0.37 1.46 0.59 0.66 3.95 1.78 0.48 0.02 0.04 1.51 2.13 2.28 2.41 2.43 0.66 2.65 }
(8.01) 2.25 6.77 3.23 2.43 1.13 3.21 2.08 1.21 3.50 2.91 3.82 3.70 2.10 0.57 1.85 0.94 3.45 0.84 3.12 3.37 3.94 0.00 2.45 1.57 1.00 1.19 3.36 0.09 1.50 0.57 7.10 0.22 }
(8.04) 3.68 1.21 0.10 2.35 2.76 3.35 2.76 3.35 2.91 1.94 0.82 2.97 1.87 1.83 3.80 2.99 0.43 2.49 1.54 2.94 2.44 2.29 1.45 0.61 0.90 1.70 3.21 2.07 3.96 3.01 1.38 0.68 2.63 }
(9.04) 3.05 1.34 0.14 2.07 2.65 1.70 0.42 3.08 0.57 3.62 2.77 1.21 1.71 0.28 3.80 2.99 0.43 2.99 0.43 2.77 1.90 0.65 2.00 2.02 2.93 1.60 1.12 2.77 2.73 3.02 2.89 1.09 0.40 1.34 1.34 0.14 2.07 1.50 0.45 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.34 0.14 2.05 1.
```

#### Matrice B:

#### Matrice C:

atrix C:
15.31 112.53 89.67 117.14 114.76 107.13 85.61 103.48 118.66 113.65 86.62 106.86 118.85 114.08 111.23 117.48 121.38 137.25 133.06 116.29 117.33 119.75 106.92 114.01 117.46 94.55 117.08 96.14 121.42 114.85 121.04 115.02 )
145.91 125.33 122.53 122.

# Rangée:

```
// Add vectors in parallel.
cudaError_t cudaStatus = multiplyWithCuda(c, a, b, arraySize, KERNEL_ROW);
```

#### Matrice A:

#### Matrice B:

\*\*\*strik\*\*\* B: \*\*\* (2.99 2, 4.09 3, 88 3, 47 3, 36 6, 59 1, 50 1, 14 0, 68 3, 90 2, 97 2, 87 0, 12 2, 58 3, 25 2, 94 3, 27 6, 50 2, 88 1, 50 2, 60 3, 66 3, 56 0, 13 3, 09 1, 80 0, 03 0, 91 1, 67 0, 63 3, 68 } (2.91 0, 97 3, 02 3, 63 0, 85 0, 13 3, 08 0, 91 2, 23 3, 23 3, 41 1, 33 2, 16 3, 97 3, 48 2, 59 0, 91 3, 25 1, 48 1, 47 0, 14 2, 25 1, 59 0, 48 1, 30 3, 60 0, 81 3, 76 0, 59 3, 40 3, 58 1, 78 } (2.00 2, 83 1, 60 3, 29 2, 59 2, 39 1, 60 3, 38 1, 84 2, 99 1, 41 0, 60 1, 91 2, 51 1, 62 1, 53 0, 11 1, 16 2, 39 3, 54 1, 73 3, 73 1, 53 2, 13 0, 40 3, 90 3, 46 1, 75 2, 80 1, 43 1, 63 1, 22 1, 99 0, 32 3, 59 1, 25 3, 81 2, 43 3, 37 3, 73 2, 35 6, 86 1, 12 2, 24 2, 50 6, 11 2, 12 2, 13 0, 40 3, 90 3, 46 1, 75 2, 80 1, 33 2, 41 1, 99 1, 41 1, 90

#### Matrice C:

attis C:
115.31 112.53 89.67 117.14 114.76 107.13 85.01 103.48 118.00 113.05 80.02 106.86 110.85 114.08 111.23 117.48 121.38 137.25 133.00 116.29 117.33 119.75 100.92 114.01 117.40 94.55 117.04 96.14 121.42 114.85 121.94 115.02 }
145.91 125.33 125.59 125.07 125.75 122.83 122.57 110.51 110

#### Colonne:

```
// Add vectors in parallel.
cudaError_t cudaStatus = multiplyWithCuda(c, a, b, arraySize, KERNEL_COLUMN);
```

## Matrice A:

## Matrice B:

#### Matrice C:

```
utrix (c: 115.31 112.85 89.07 117.14 114.76 197.13 85.01 103.48 118.60 113.65 86.02 106.86 110.85 114.08 111.23 117.48 121.38 117.25 133.00 116.29 117.33 119.75 106.92 114.01 117.46 94.55 117.04 96.14 121.42 114.85 121.94 115.02 }
145.91 125.33 123.53 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 113.63 123.13 113.60 123.74 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 123.63 12
```

## b. Les blocs doivent avoir une taille de 8x8 (à l'exception du cas de matrice 4x4)

B. Profilez votre kernel et discuter des résultats

Avec des matrices 32\*32:

#### NvProf:

#### Élément :

```
==21736== Profiling application: a
==21736== Profiling result:
Time
GPU activities: 6.2.4% 6.7520us 1 6.7520us 6.7520us 6.7520us
9.44% 1.2040us 2 1.5360us 1.280us 1.2
```

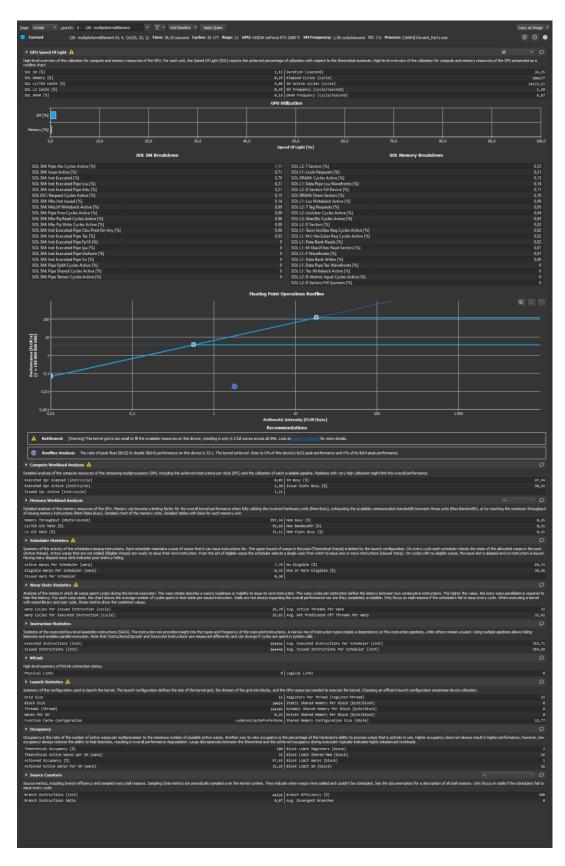
# Rangée:

#### Colonne:

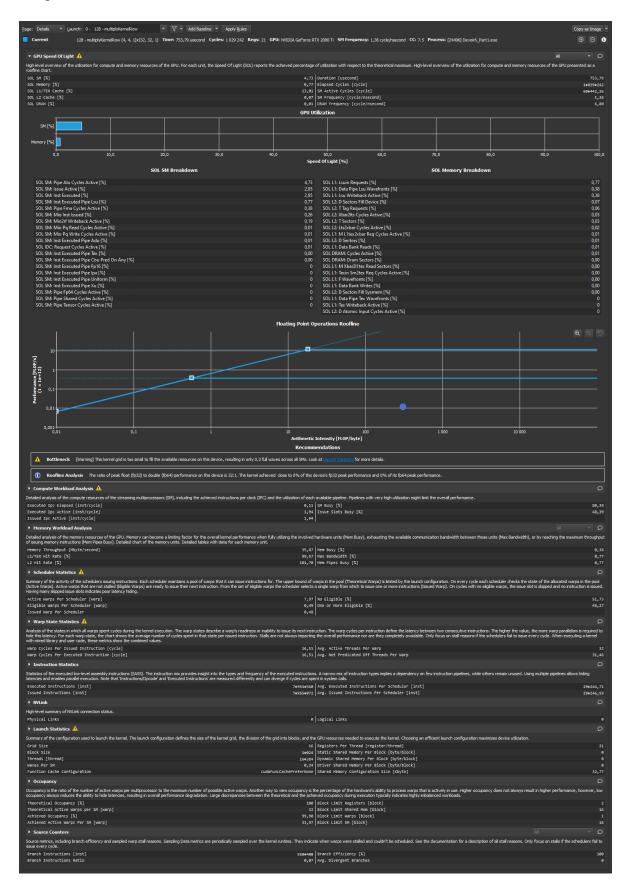
```
=1236== Profiling application: a
=1236== Profiling result:
           Туре
                 Time(%)
                               Time
                                        Calls
                                                    Avg
                                                                         Max
                                                                              Name
GPU activities:
                                                                              multiplyKernelColumn(float*, float const *
                                            1 116.45us
                                                          116.45us
                                                                    116.45us
                 96.60%
                          116.45us
float const *, int)
                   2.57%
                          3.1040us
                                                                    1.8240us
                                               1.5520us
                                                          1.2800us
                                                                               [CUDA memcpy HtoD]
                                                                               [CUDA memcpy DtoH]
                   0.82%
                              992ns
                                                  992ns
                                                             992ns
                                                                       992ns
     API calls:
                                               75.436ms
                  80.31%
                          226.31ms
                                                          2.1000us
                                                                    226.30ms
                                                                               cudaMalloc
                  19.47%
                          54.857ms
                                               54.857ms
                                                          54.857ms
                                                                    54.857ms
                                                                              cudaDeviceReset
                                                                              cudaMemcpy
cudaDeviceSynchronize
                   0.10%
                          274.00us
                                               91.333us
                                                          37.200us
                                                                    184.40us
                   0.05%
                          154.80us
                                               154.80us
                                                          154.80us
                                                                    154.80us
                   0.04%
                          122.40us
                                               40.800us
                                                          1.9000us
                                                                    115.90us
                                                                               cudaFree
                   0.02%
                          56.800us
                                               56.800us
                                                          56.800us
                                                                    56.800us
                                                                              cudaLaunchKernel
                   0.01%
                          14.100us
                                          101
                                                  139ns
                                                             100ns
                                                                       600ns
                                                                              cuDeviceGetAttribute
                   0.00%
                          6.5000us
                                               2.1660us
                                                             200ns
                                                                    5.7000us
                                                                              cuDeviceGetCount
                   0.00%
                          5.8000us
                                               5.8000us
                                                          5.8000us
                                                                    5.8000us
                                                                              cudaSetDevice
                                                                    1.0000us
                   0.00%
                          1.2000us
                                                  600ns
                                                             200ns
                                                                              cuDeviceGet
                   0.00%
                              600ns
                                                  600ns
                                                                       600ns
                                                                              cuDeviceGetName
                                                             600ns
                   0.00%
                              300ns
                                                  300ns
                                                             300ns
                                                                       300ns
                                                                              cuDeviceGetLuid
                   0.00%
                              300ns
                                                   300ns
                                                             300ns
                                                                       300ns
                                                                              cudaGetLastError
                   0.00%
                              200ns
                                                  200ns
                                                             200ns
                                                                              cuDeviceTotalMem
                                                                       200ns
                   0.00%
                              200ns
                                                   200ns
                                                             200ns
                                                                       200ns
                                                                              cuDeviceGetUuid
:\Users\lacommap\Desktop\CudaC\Devoir6_Part1>
```

# NSight:

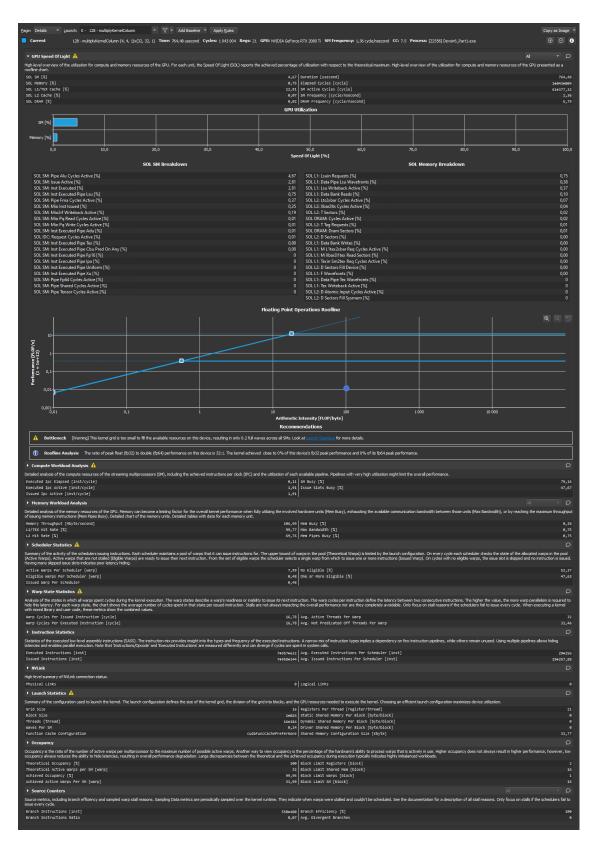
## Élément :



# Rangée:



## Colonne:



# PARTIE 2 – MULTIPLICATION MATRICIELLE OPTIMISÉE

## **Objectifs**

Réimplémentez la multiplication matricielle de la PARTIE 1 en utilisant l'algorithme de tiling [chap. 4, 2]. En programmation paralèlle, le tiling force plusieurs threads à se concentrer sur un sous ensemble de données d'entrée lors de chaque phase de sorte que ce sous ensemble de donnée puisse être stocké dans une mémoire spéciale, la shared memory par exemple, ce qui accélère grandement les accès mémoire. L'étudiant doit démontrer qu'avec des tiles de grande taille, les accès à la mémoire globale n'est plus un principal facteur limitant.

D'autre part, la capacité de raisonner en fonction des limitations matérielles en développant une application est un aspect clé de la pensée computationnelle. Les programmeurs CUDA doivent être conscient des limitations de chaque type de mémoire ainsi que des ressources propres à leur GPU puisqu'elles sont étroitement liées à l'implémentation. Une fois les capacités excédées, le GPU limite le nombre de threads pouvant être exécuter dans chaque SM. Les threads sont assignés aux SM sur une base de bloc par bloc. Chaque device CUDA dispose d'une limitation différente par rapport aux ressources disponible dans les SM, autant une limite de blocs par SM et une limite de threads par SM. La première limitation atteinte deviendra le facteur limitant.

Pour chaque Kernel, l'une ou plusieurs de ces limitations peut devenir le facteur limitant pour le nombre de threads qui sont utilisés simultanément dans le device CUDA.

#### **MÉTHODOLOGIE**

1. Votre implémentation doit être « transparent scalable » (voir dection 3.4 à 3.6 dans [chap.3, 2]). Pour ce faire, vous devez connaître les propriétés de votre device. L'annexe B fournit un exemple d'une « host-stub function » qui permet d'extraire les propriétés.

```
Device 0: NVIDIA GeForce RTX 2080 Ti

Number of multiprocessors: 68
clock rate: 1545000
Compute capability: 7.5
Total amount of global memory: 11533888.00 KB
Total amount of constant memory: 64.00 KB
Total amount of shared memory per block: 48.00 KB
Total amount of shared memory per block: 65536
Warp size: 32
Maximum number of threads per block: 1024
Maximum number of threads per multiprocessor: 1024
Maximum number of warps per multiprocessor: 32
Maximum Grid size: (2147483647,65535,65535)
Maximum block dimension: (1024,1024,64)
Sortie de C:\Users\lacommap\Desktop\CudaC\Devoir6Part2\x64\Debug\Devoir6Part2.exe (processus 7932). Code: 0.
Appuyez sur une touche pour fermer cette fenêtre. . .
```

- 2. L'implémentation doit être limitée à des matrices carrées 64x64. Toutefois, la taille des tuiles doit varier entre 2x2, 4x4, 8x8 et 16x16.
- 3. Validez chaque implementation.

Pour alléger l'affichage je prendrais des matrices 16\*16.

Dans la fonction host stub, le nombre de threads par block est équivalent à la taille des tuiles. Par la suite, le nombre de bloc dans un grid est calculé en fonction du nombre de tuilles :

```
⊒cudaError_t <mark>multiplyWithCuda(</mark>float* c, const float* a, const float* b, unsigned int size)
     float* dev_a = 0;
     float* dev_b = 0;
     float* dev_c = 0;
     cudaError_t cudaStatus;
     if (size >64) {
        printf("Taille de la matrice limité à 64 par 64");
         return cudaErrorInvalidValue;
     // Choose which GPU to run on, change this on a multi-GPU system.
     cudaStatus = cudaSetDevice(0);
     if (cudaStatus != cudaSuccess) {
         fprintf(stderr, "cudaSetDevice failed! Do you have a CUDA-capable GPU installed?");
         goto Error;
     // Allocate GPU buffers for three vectors (two input, one output)
     cudaStatus = cudaMalloc((void**)&dev_c, (size * size) * sizeof(float));
     if (cudaStatus != cudaSuccess) {
        fprintf(stderr, "cudaMalloc failed!");
         goto Error;
     cudaStatus = cudaMalloc((void**)&dev_a, (size * size) * sizeof(float));
     if (cudaStatus != cudaSuccess) {
        fprintf(stderr, "cudaMalloc failed!");
         goto Error;
     cudaStatus = cudaMalloc((void**)&dev_b, (size * size) * sizeof(float));
     if (cudaStatus != cudaSuccess) {
         fprintf(stderr, "cudaMalloc failed!");
         goto Error;
```

```
// Copy input vectors from host memory to GPU buffers.
cudaStatus = cudaMemcpy(dev_a, a, (size * size) * sizeof(float), cudaMemcpyHostToDevice);
if (cudaStatus != cudaSuccess) {
    `fprintf(stderr, "cudaMemcpy failed!");
     goto Error;
cudaStatus = cudaMemcpy(dev_b, b, (size * size) * sizeof(float), cudaMemcpyHostToDevice);
 if (cudaStatus != cudaSuccess) {
    fprintf(stderr, "cudaMemcpy failed!");
    goto Error;
 dim3 threadsParBlock(tileLength, tileLength, 1);
 dim3 nombreDeBlock((size + tileLength - 1) / tileLength, (size + tileLength - 1) / tileLength);
 multiplyKernel << <nombreDeBlock, threadsParBlock >> > (dev_c, dev_a, dev_b, size);
 cudaStatus = cudaGetLastError();
if (cudaStatus != cudaSuccess) {
   fprintf(stderr, "addKernel launch failed: %s\n", cudaGetErrorString(cudaStatus));
     goto Error:
 cudaStatus = cudaDeviceSynchronize();
 if (cudaStatus != cudaSuccess) {
    fprintf(stderr, "cudaDeviceSynchronize returned error code %d after launching addKernel!\n", cudaStatus);
    goto Error;
    cudaStatus = cudaMemcpy(c, dev_c, (size * size) * sizeof(float), cudaMemcpyDeviceToHost);
    if (cudaStatus != cudaSuccess) {
         fprintf(stderr, "cudaMemcpy failed!");
         goto Error;
Error:
    cudaFree(dev_c);
    cudaFree(dev_a);
    cudaFree(dev_b);
    return cudaStatus;
```

Pour le kernel, il existe plusieurs implémentation open-source qui fait ce genre de calcul matricielle, je me suis alors inspiré de ce répertoire sur Github :

https://github.com/tgautam03/CUDA-C/blob/master/05 tiled mat mul/tiled mat mul gpu.cu

```
global__ void multiplyKernel(float* c, const float* a, const float* b, int length) {
  int threadY = threadIdx.y;
  int threadX = threadIdx.x;
  int row = (tileLength * blockIdx.y) + threadY;
  int column = (tileLength * blockIdx.x) + threadX;
   _shared__ float tileA[tileLength][tileLength];
  __shared__ float tileB[tileLength][tileLength];
  int numTiles = (length + tileLength - 1) / tileLength;
  for (int t = 0; t < numTiles; t++) {</pre>
      if (row < length && (t * tileLength + threadX) < length) {
   tileA[threadY][threadX] = a[row * length + (t * tileLength + threadX)];</pre>
      else {
          tileA[threadY][threadX] = 0.0f;
      if (column < length && (t * tileLength + threadY) < length) {</pre>
          tileB[threadY][threadX] = b[(t * tileLength + threadY) * length + column];
      else {
          tileB[threadY][threadX] = 0.0f;
      __syncthreads();
      for (int k = 0; k < tileLength; k++) {
          sum += tileA[threadY][k] * tileB[k][threadX];
      __syncthreads();
  if (row < length && column < length) {
        c[(row * length) + column] = sum;
```

Essentiellement, cette technique permet d'exploiter la mémoire partagée de chaque bloc, ce qui réduit considérablement les accès à la mémoire globale et améliore les performances. Les variables « row » et « column » servent à calculer les indices de la matrice C finale que chaque thread est chargé de calculer.

Dans chaque itération, les données nécessaires sont d'abord exportées depuis la mémoire globale vers la mémoire partagée pour chaque tuile. Une fois toutes les données chargées, on s'assure de leur disponibilité grâce à la fonction \_\_syncthreads(), qui synchronise tous les threads du bloc. Ensuite, les threads effectuent les calculs sur ces données en mémoire partagée. Avant de passer à la tuile suivante, on utilise à nouveau \_\_syncthreads() pour garantir que tous les threads ont terminé leurs calculs.

Finalement, les résultats calculés sont écrits dans la mémoire globale pour remplir les éléments correspondants de la matrice C.

#### Tuile 2\*2:

```
0.01 2.25 0.77 3.23 2.34 1.92 1.40 3.58 3.29 2.99 0.70 3.44 2.84 2.05 1.22 0.06
 0.37 1.46 0.59 0.66 3.95 1.78 0.48 0.02 0.04 1.51 2.13 2.28 2.41 2.43 0.66 2.65
 1.80 1.41 0.23 2.43 3.13 3.21 2.08 1.21 3.50 2.91 3.82 3.70 2.16 0.57 1.85 0.94
 3.45 0.84 3.12 3.37 3.99 4.00 2.45 1.57 1.06 1.19 3.36 0.09 1.50 0.37 2.71 0.22
 0.04 3.68 1.10 1.09 2.35 2.76 3.35 2.91 1.94 0.82 2.97 1.87 1.83 3.80 2.98 0.43
 2.40 1.54 2.94 2.44 2.29 1.45 0.61 0.90 1.70 3.21 2.07 3.96 3.01 1.38 0.68 2.63
 1.97 0.25 2.80 2.02 0.59 3.80 0.57 3.62 2.77 1.21 1.71 0.28 3.87 2.73 0.61 3.51
 3.29\ 2.33\ 0.77\ 0.71\ 3.27\ 1.90\ 0.62\ 2.02\ 2.93\ 1.62\ 1.12\ 2.27\ 2.73\ 3.02\ 2.89\ 1.90
 0.49 1.47 3.34 0.14 2.07 2.65 1.70 0.42 3.80 3.69 2.20 1.38 1.89 1.50 3.39 1.27
 1.82 1.09 3.93 1.19 2.96 2.27 0.78 3.05 3.36 1.59 2.00 3.56 0.11 3.98 2.29 0.20
 2.13 0.78 3.37 2.51 2.63 0.79 3.37 0.49 0.44 2.97 1.26 3.76 1.14 1.35 0.56 2.93
 3.34 2.83 2.40 2.99 1.01 0.58 0.01 0.24 3.22 3.41 0.84 0.46 2.21 0.06 0.46 1.82
 3.01 2.74 2.17 0.30 1.75 0.81 2.78 1.16 1.75 0.93 2.31 2.13 2.51 0.64 2.02 3.85
 2.78 3.70 0.76 1.34 0.71 3.98 1.83 3.99 0.39 2.50 0.38 1.75 3.73 0.19 3.58 1.16
 0.91 3.08 1.64 0.81 2.51 2.42 1.81 1.87 2.39 2.54 3.42 3.32 2.50 2.88 2.26 1.50
 0.74 2.95 2.22 3.62 0.97 0.76 2.42 2.79 2.34 1.41 1.98 0.32 2.96 2.45 2.48 2.76
Matrix B:
 3.22 0.60 2.30 3.47 3.65 2.46 2.91 0.17 2.67 3.91 1.26 2.28 1.22 0.70 0.43 3.48
 3.40 2.98 0.62 1.31 0.32 0.31 2.56 3.28 2.18 1.79 1.64 1.19 1.86 2.00 0.61 1.29
 2.95 1.26 3.31 3.84 3.49 2.90 1.20 3.78 0.51 0.26 3.14 2.10 2.44 3.82 0.29 3.50
 2.62 1.29 0.42 2.02 0.91 1.16 3.68 2.20 2.65 0.46 1.97 1.52 1.99 3.17 2.04 1.53
 2.75 2.13 2.43 1.58 0.02 2.83 0.40 2.49 3.45 1.97 2.99 1.99 1.52 3.14 2.21 1.43
 3.82 2.52 0.71 1.50 0.53 2.97 3.81 2.45 0.11 1.32 0.22 2.56 0.53 3.39
                                                                         3.46 2.39
 2.89 3.42 0.06 0.51 2.83 2.47 0.87 0.26 0.68 2.50 1.36 1.28 1.47 2.64 3.21 3.23
 2.11 2.44 3.19 3.60 0.58 2.52 1.61 1.01 0.55 3.42 0.26 1.71 2.29 1.21 2.19 0.90
 1.25 0.44 3.23 0.54 1.14 3.15 3.58 3.16 2.98 2.46 1.44 3.43 0.91 3.45 0.92 1.00
 2.17\  \, 3.94\  \, 0.22\  \, 0.33\  \, 2.10\  \, 1.71\  \, 0.38\  \, 1.04\  \, 3.57\  \, 0.93\  \, 0.59\  \, 0.50\  \, 3.73\  \, 0.32\  \, 0.19\  \, 0.23
 1.35 \ 3.66 \ 1.59 \ 1.73 \ 3.78 \ 3.35 \ 2.14 \ 3.37 \ 2.77 \ 1.59 \ 1.04 \ 0.02 \ 2.10 \ 3.82 \ 1.59 \ 0.96
 2.34 1.02 2.74 3.78 1.74 3.56 0.03 3.76 2.41 3.14 2.31 0.57 0.89 1.53 0.02 1.67
 0.33 2.64 3.42 0.26 3.24 2.65 2.77 3.21 2.12 2.74 0.57 2.76 2.91 3.11 0.12 3.47
 2.58 2.83 0.34 2.21 3.79 0.24 1.10 0.58 3.93 2.48 1.17 3.69 1.47 2.78 0.87 0.62
 0.96 2.09 3.61 0.43 3.61 1.77 0.32 3.13 0.69 3.90 3.10 3.48 0.84 1.83 0.02 3.00
 0.46 1.62 1.24 3.97 0.15 1.01 0.76 0.99 0.61 2.48 3.54 3.76 0.78 3.12 2.58 2.63
 70.80 70.44 60.55 55.63 53.52 71.79 58.92 74.93 70.70 69.21 45.94 61.33 58.39 76.84 39.43 51.63
 49.17 55.72 39.75 45.64 41.92 50.29 33.63 54.18 54.02 49.74 42.96 47.97 38.99 63.23 32.08 42.80
 75.95 76.97 64.24 58.84 65.36 87.00 65.14 83.46 76.66 76.39 53.56 62.77 57.28 88.26 47.16 62.48
 81.42 72.16 62.48 61.42 69.52 80.93 66.88 74.54 61.90 67.46 55.16 64.62 55.39 89.55 51.41 72.63
 77.62 83.65 58.76 56.78 68.43 71.75 58.12 76.98 66.41 77.38 51.61 68.67 54.89 88.44 48.11 60.74
 71.03 68.16 63.49 70.94 66.00 76.27 55.70 78.00 72.34 68.76 58.16 63.07 59.81 82.01 35.69 65.14
 65.69 67.09 63.95 69.06 62.88 71.89 68.78 67.00 56.89 69.80 46.87 77.88 55.18 88.18 47.56 66.43
 71.27 68.53 68.56 63.22 66.02 72.17 58.20 72.37 73.04 83.53 55.71 76.59 52.55 79.42 37.49 64.46
 65.70 70.98 62.60 48.89 67.59 73.68 51.46 77.55 62.28 64.62 53.95 66.86 54.37 81.99 35.07 60.77
 81.04 68.29 71.31 74.53 71.75 80.50 55.19 80.01 72.40 75.56 58.60 71.48 54.67 86.42 39.28 60.14
 70.25 64.43 50.70 67.98 61.70 68.88 41.78 63.04 61.80 61.66 58.96 55.29 53.30 74.84 39.85 63.85
 56.59 49.83 47.54 48.05 48.80 53.69 56.22 58.36 59.48 50.12 43.27 52.23 49.73 61.82 24.73 50.78
 63.07 65.14 60.88 64.05 63.59 68.71 50.39 67.48 56.20 74.25 57.66 64.72 49.31 77.56 38.70 69.60
 74.40 75.85 63.54 58.34 60.90 70.58 62.65 70.61 52.94 80.12 45.19 66.55 56.93 70.47 41.53 70.33
 77.63 83.83 66.00 64.84 72.50 80.28 59.14 85.22 76.03 79.55 57.21 69.55 60.86 90.57 43.11 64.04
 67.19 74.27 59.60 60.77 65.49 64.51 62.49 72.07 63.26 70.87 56.01 72.25 59.15 87.55 44.04 64.86
```

#### Tuile 4\*4:

```
0.01 2.25 0.77 3.23 2.34 1.92 1.40 3.58 3.29 2.99 0.70 3.44 2.84 2.05 1.22 0.06 }
 0.37 1.46 0.59 0.66 3.95 1.78 0.48 0.02 0.04 1.51 2.13 2.28 2.41 2.43 0.66 2.65
 1.80 1.41 0.23 2.43 3.13 3.21 2.08 1.21 3.50 2.91 3.82 3.70 2.16 0.57 1.85 0.94
 3.45 0.84 3.12 3.37 3.99 4.00 2.45 1.57 1.06 1.19 3.36 0.09 1.50 0.37 2.71 0.22 0.04 3.68 1.10 1.09 2.35 2.76 3.35 2.91 1.94 0.82 2.97 1.87 1.83 3.80 2.98 0.43
 2.40 1.54 2.94 2.44 2.29 1.45 0.61 0.90 1.70 3.21 2.07 3.96 3.01 1.38 0.68 2.63
 1.97 0.25 2.80 2.02 0.59 3.80 0.57 3.62 2.77 1.21 1.71 0.28 3.87 2.73 0.61 3.51
 3.29 2.33 0.77 0.71 3.27 1.90 0.62 2.02 2.93 1.62 1.12 2.27 2.73 3.02 2.89 1.90
 0.49 1.47 3.34 0.14 2.07 2.65 1.70 0.42 3.80 3.69 2.20 1.38 1.89 1.50 3.39 1.27
 1.82\ 1.09\ 3.93\ 1.19\ 2.96\ 2.27\ 0.78\ 3.05\ 3.36\ 1.59\ 2.00\ 3.56\ 0.11\ 3.98\ 2.29\ 0.20
 2.13 0.78 3.37 2.51 2.63 0.79 3.37 0.49 0.44 2.97 1.26 3.76 1.14 1.35 0.56 2.93
 3.34 2.83 2.40 2.99 1.01 0.58 0.01 0.24 3.22 3.41 0.84 0.46 2.21 0.06 0.46 1.82
 3.01 2.74 2.17 0.30 1.75 0.81 2.78 1.16 1.75 0.93 2.31 2.13 2.51 0.64 2.02 3.85
 2.78 3.70 0.76 1.34 0.71 3.98 1.83 3.99 0.39 2.50 0.38 1.75 3.73 0.19 3.58 1.16
 0.91 3.08 1.64 0.81 2.51 2.42 1.81 1.87 2.39 2.54 3.42 3.32 2.50 2.88 2.26 1.50
 0.74 2.95 2.22 3.62 0.97 0.76 2.42 2.79 2.34 1.41 1.98 0.32 2.96 2.45 2.48 2.76
Matrix B:
 3.22 0.60 2.30 3.47 3.65 2.46 2.91 0.17 2.67 3.91 1.26 2.28 1.22 0.70 0.43 3.48
 3.40 2.98 0.62 1.31 0.32 0.31 2.56 3.28 2.18 1.79 1.64 1.19 1.86 2.00 0.61 1.29
 2.95 1.26 3.31 3.84 3.49 2.90 1.20 3.78 0.51 0.26 3.14 2.10 2.44 3.82 0.29 3.50
 2.62 1.29 0.42 2.02 0.91 1.16 3.68 2.20 2.65 0.46 1.97 1.52 1.99
                                                                       3.17 2.04 1.53
 2.75 2.13 2.43 1.58 0.02 2.83 0.40 2.49 3.45 1.97 2.99 1.99 1.52 3.14 2.21 1.43
 3.82 2.52 0.71 1.50 0.53 2.97 3.81 2.45 0.11 1.32 0.22 2.56 0.53 3.39 3.46 2.39
 2.89 3.42 0.06 0.51 2.83 2.47 0.87 0.26 0.68 2.50 1.36 1.28 1.47 2.64 3.21 3.23
 2.11 2.44 3.19 3.60 0.58 2.52 1.61 1.01 0.55 3.42 0.26 1.71 2.29 1.21 2.19 0.90
 1.25 0.44 3.23 0.54 1.14 3.15 3.58 3.16 2.98 2.46 1.44 3.43 0.91 3.45 0.92 1.00
 2.17 3.94 0.22 0.33 2.10 1.71 0.38 1.04 3.57 0.93 0.59 0.50 3.73 0.32 0.19 0.23
 1.35 3.66 1.59 1.73 3.78 3.35 2.14 3.37 2.77 1.59 1.04 0.02 2.10 3.82 1.59 0.96
 2.34 1.02 2.74 3.78 1.74 3.56 0.03 3.76 2.41 3.14 2.31 0.57 0.89 1.53 0.02 1.67
 0.33 2.64 3.42 0.26 3.24 2.65 2.77 3.21 2.12 2.74 0.57 2.76 2.91 3.11 0.12 3.47
 2.58 2.83 0.34 2.21 3.79 0.24 1.10 0.58 3.93 2.48 1.17 3.69 1.47 2.78 0.87 0.62
 0.96 2.09 3.61 0.43 3.61 1.77 0.32 3.13 0.69 3.90 3.10 3.48 0.84 1.83 0.02 3.00
 0.46 1.62 1.24 3.97 0.15 1.01 0.76 0.99 0.61 2.48 3.54 3.76 0.78 3.12 2.58 2.63
 70.80 70.44 60.55 55.63 53.52 71.79 58.92 74.93 70.70 69.21 45.94 61.33 58.39 76.84 39.43 51.63
 49.17 55.72 39.75 45.64 41.92 50.29 33.63 54.18 54.02 49.74 42.96 47.97 38.99 63.23 32.08 42.80
 75.95 76.97 64.24 58.84 65.36 87.00 65.14 83.46 76.66 76.39 53.56 62.77 57.28 88.26 47.16 62.48
 81.42 72.16 62.48 61.42 69.52 80.93 66.88 74.54 61.90 67.46 55.16 64.62 55.39 89.55 51.41 72.63
 77.62 83.65 58.76 56.78 68.43 71.75 58.12 76.98 66.41 77.38 51.61 68.67 54.89 88.44 48.11 60.74
 71.03 68.16 63.49 70.94 66.00 76.27 55.70 78.00 72.34 68.76 58.16 63.07 59.81 82.01 35.69 65.14
 65.69 67.09 63.95 69.06 62.88 71.89 68.78 67.00 56.89 69.80 46.87 77.88 55.18 88.18 47.56 66.43
 71.27 68.53 68.56 63.22 66.02 72.17 58.20 72.37 73.04 83.53 55.71 76.59 52.55 79.42 37.49 64.46 65.70 70.98 62.60 48.89 67.59 73.68 51.46 77.55 62.28 64.62 53.95 66.86 54.37 81.99 35.07 60.77
 81.04 68.29 71.31 74.53 71.75 80.50 55.19 80.01 72.40 75.56 58.60 71.48 54.67 86.42 39.28 60.14
 70.25 64.43 50.70 67.98 61.70 68.88 41.78 63.04 61.80 61.66 58.96 55.29 53.30 74.84 39.85 63.85
 56.59 49.83 47.54 48.05 48.80 53.69 56.22 58.36 59.48 50.12 43.27 52.23 49.73 61.82 24.73 50.78
 63.07 65.14 60.88 64.05 63.59 68.71 50.39 67.48 56.20 74.25 57.66 64.72 49.31 77.56 38.70 69.60
 74.40 75.85 63.54 58.34 60.90 70.58 62.65 70.61 52.94 80.12 45.19 66.55 56.93 70.47 41.53 70.33 77.63 83.83 66.00 64.84 72.50 80.28 59.14 85.22 76.03 79.55 57.21 69.55 60.86 90.57 43.11 64.04
 67.19 74.27 59.60 60.77 65.49 64.51 62.49 72.07 63.26 70.87 56.01 72.25 59.15 87.55 44.04 64.86
```

#### Tuile 8\*8:

```
0.01 2.25 0.77 3.23 2.34 1.92 1.40 3.58 3.29 2.99 0.70 3.44 2.84 2.05 1.22 0.06
 0.37 1.46 0.59 0.66 3.95 1.78 0.48 0.02 0.04 1.51 2.13 2.28 2.41 2.43 0.66 2.65 1.80 1.41 0.23 2.43 3.13 3.21 2.08 1.21 3.50 2.91 3.82 3.70 2.16 0.57 1.85 0.94
 3.45 0.84 3.12 3.37 3.99 4.00 2.45 1.57 1.06 1.19 3.36 0.09 1.50 0.37 2.71 0.22
 0.04 3.68 1.10 1.09 2.35 2.76 3.35 2.91 1.94 0.82 2.97 1.87 1.83 3.80 2.98 0.43
 2.40 1.54 2.94 2.44 2.29 1.45 0.61 0.90 1.70 3.21 2.07 3.96 3.01 1.38 0.68 2.63
 1.97 0.25 2.80 2.02 0.59 3.80 0.57 3.62 2.77 1.21 1.71 0.28 3.87 2.73 0.61 3.51
 3.29 2.33 0.77 0.71 3.27 1.90 0.62 2.02 2.93 1.62 1.12 2.27 2.73 3.02 2.89 1.90
 0.49 1.47 3.34 0.14 2.07 2.65 1.70 0.42 3.80 3.69 2.20 1.38 1.89 1.50 3.39 1.27
 1.82 1.09 3.93 1.19 2.96 2.27 0.78 3.05 3.36 1.59 2.00 3.56 0.11 3.98 2.29 0.20
 2.13 0.78 3.37 2.51 2.63 0.79 3.37 0.49 0.44 2.97 1.26 3.76 1.14 1.35 0.56 2.93
 3.34 2.83 2.40 2.99 1.01 0.58 0.01 0.24 3.22 3.41 0.84 0.46 2.21 0.06 0.46 1.82
 3.01 2.74 2.17 0.30 1.75 0.81 2.78 1.16 1.75 0.93 2.31 2.13 2.51 0.64 2.02 3.85
 2.78 3.70 0.76 1.34 0.71 3.98 1.83 3.99 0.39 2.50 0.38 1.75 3.73 0.19 3.58 1.16
 0.91 \ 3.08 \ 1.64 \ 0.81 \ 2.51 \ 2.42 \ 1.81 \ 1.87 \ 2.39 \ 2.54 \ 3.42 \ 3.32 \ 2.50 \ 2.88 \ 2.26 \ 1.50
 0.74 2.95 2.22 3.62 0.97 0.76 2.42 2.79 2.34 1.41 1.98 0.32 2.96 2.45 2.48 2.76
Matrix B:
 3.22 0.60 2.30 3.47 3.65 2.46 2.91 0.17 2.67 3.91 1.26 2.28 1.22 0.70 0.43 3.48 3.40 2.98 0.62 1.31 0.32 0.31 2.56 3.28 2.18 1.79 1.64 1.19 1.86 2.00 0.61 1.29
 2.95 1.26 3.31 3.84 3.49 2.90 1.20 3.78 0.51 0.26 3.14 2.10 2.44 3.82 0.29 3.50
 2.62 1.29 0.42 2.02 0.91 1.16 3.68 2.20 2.65 0.46 1.97 1.52 1.99 3.17 2.04 1.53
 2.75 2.13 2.43 1.58 0.02 2.83 0.40 2.49 3.45 1.97 2.99 1.99 1.52 3.14 2.21 1.43
 3.82\ 2.52\ 0.71\ 1.50\ 0.53\ 2.97\ 3.81\ 2.45\ 0.11\ 1.32\ 0.22\ 2.56\ 0.53\ 3.39\ 3.46
 2.89 3.42 0.06 0.51 2.83 2.47 0.87 0.26 0.68 2.50 1.36 1.28 1.47 2.64 3.21 3.23 2.11 2.44 3.19 3.60 0.58 2.52 1.61 1.01 0.55 3.42 0.26 1.71 2.29 1.21 2.19 0.90
 1.25 0.44 3.23 0.54 1.14 3.15 3.58 3.16 2.98 2.46 1.44 3.43 0.91 3.45 0.92 1.00
  2.17 3.94 0.22 0.33 2.10 1.71 0.38 1.04 3.57 0.93 0.59 0.50 3.73 0.32 0.19 0.23
  1.35 3.66 1.59 1.73 3.78 3.35 2.14 3.37 2.77 1.59 1.04 0.02 2.10 3.82 1.59 0.96
 2.34 1.02 2.74 3.78 1.74 3.56 0.03 3.76 2.41 3.14 2.31 0.57 0.89 1.53 0.02 1.67
 0.33 2.64 3.42 0.26 3.24 2.65 2.77 3.21 2.12 2.74 0.57 2.76 2.91 3.11 0.12 3.47
  2.58 2.83 0.34 2.21 3.79 0.24 1.10 0.58 3.93 2.48 1.17 3.69 1.47 2.78 0.87 0.62
 0.96 2.09 3.61 0.43 3.61 1.77 0.32 3.13 0.69 3.90 3.10 3.48 0.84 1.83 0.02 3.00
 0.46 1.62 1.24 3.97 0.15 1.01 0.76 0.99 0.61 2.48 3.54 3.76 0.78 3.12 2.58 2.63
 70.80 70.44 60.55 55.63 53.52 71.79 58.92 74.93 70.70 69.21 45.94 61.33 58.39 76.84 39.43 51.63
 49.17 55.72 39.75 45.64 41.92 50.29 33.63 54.18 54.02 49.74 42.96 47.97 38.99 63.23 32.08 42.80
 75.95 76.97 64.24 58.84 65.36 87.00 65.14 83.46 76.66 76.39 53.56 62.77 57.28 88.26 47.16 62.48
 81.42 72.16 62.48 61.42 69.52 80.93 66.88 74.54 61.90 67.46 55.16 64.62 55.39 89.55 51.41 72.63
 77.62 83.65 58.76 56.78 68.43 71.75 58.12 76.98 66.41 77.38 51.61 68.67 54.89 88.44 48.11 60.74
 71.03 68.16 63.49 70.94 66.00 76.27 55.70 78.00 72.34 68.76 58.16 63.07 59.81 82.01 35.69 65.14 65.69 67.09 63.95 69.06 62.88 71.89 68.78 67.00 56.89 69.80 46.87 77.88 55.18 88.18 47.56 66.43
 71.27 68.53 68.56 63.22 66.02 72.17 58.20 72.37 73.04 83.53 55.71 76.59 52.55 79.42 37.49 64.46 65.70 70.98 62.60 48.89 67.59 73.68 51.46 77.55 62.28 64.62 53.95 66.86 54.37 81.99 35.07 60.77
 81.04 68.29 71.31 74.53 71.75 80.50 55.19 80.01 72.40 75.56 58.60 71.48 54.67 86.42 39.28 60.14
 70.25 64.43 50.70 67.98 61.70 68.88 41.78 63.04 61.80 61.66 58.96 55.29 53.30 74.84 39.85 63.85
 56.59 49.83 47.54 48.05 48.80 53.69 56.22 58.36 59.48 50.12 43.27 52.23 49.73 61.82 24.73 50.78
 63.07 65.14 60.88 64.05 63.59 68.71 50.39 67.48 56.20 74.25 57.66 64.72 49.31 77.56 38.70 69.60
 74.40 75.85 63.54 58.34 60.90 70.58 62.65 70.61 52.94 80.12 45.19 66.55 56.93 70.47 41.53 70.33 77.63 83.83 66.00 64.84 72.50 80.28 59.14 85.22 76.03 79.55 57.21 69.55 60.86 90.57 43.11 64.04
 67.19 74.27 59.60 60.77 65.49 64.51 62.49 72.07 63.26 70.87 56.01 72.25 59.15 87.55 44.04 64.86 }
```

#### Tuile 16\*16:

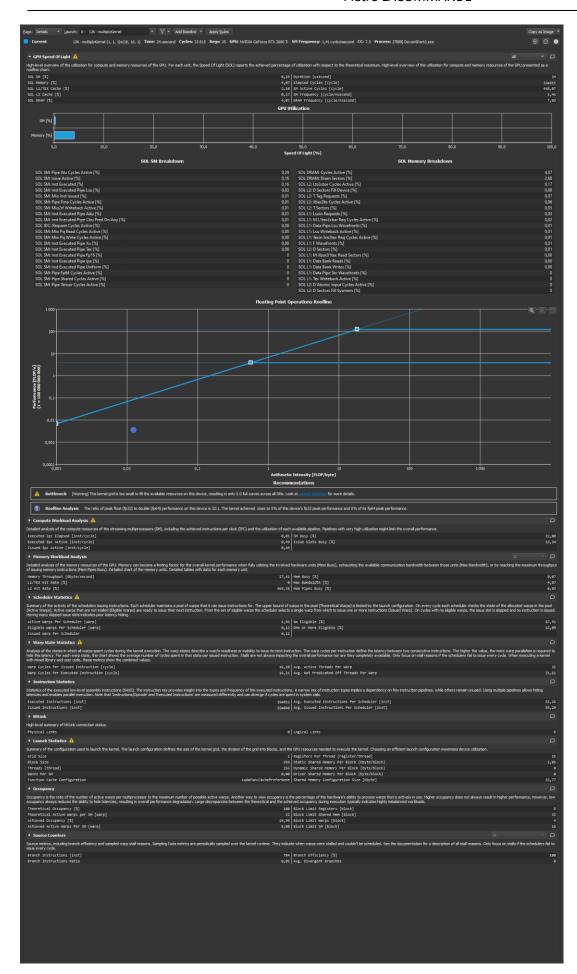
```
0.01 2.25 0.77 3.23 2.34 1.92 1.40 3.58 3.29 2.99 0.70 3.44 2.84 2.05 1.22 0.06 0.37 1.46 0.59 0.66 3.95 1.78 0.48 0.02 0.04 1.51 2.13 2.28 2.41 2.43 0.66 2.65 1.80 1.41 0.23 2.43 3.13 3.21 2.08 1.21 3.50 2.91 3.82 3.70 2.16 0.57 1.85 0.94
 3.45 0.84 3.12 3.37 3.99 4.00 2.45 1.57 1.06 1.19 3.36 0.09 1.50 0.37 2.71 0.22
 0.04 3.68 1.10 1.09 2.35 2.76 3.35 2.91 1.94 0.82 2.97 1.87 1.83 3.80 2.98 0.43
2.40 1.54 2.94 2.44 2.29 1.45 0.61 0.90 1.70 3.21 2.07 3.96 3.01 1.38 0.68 2.63 1.97 0.25 2.80 2.02 0.59 3.80 0.57 3.62 2.77 1.21 1.71 0.28 3.87 2.73 0.61 3.51 3.29 2.33 0.77 0.71 3.27 1.90 0.62 2.02 2.93 1.62 1.12 2.27 2.73 3.02 2.89 1.90 0.49 1.47 3.34 0.14 2.07 2.65 1.70 0.42 3.80 3.69 2.20 1.38 1.89 1.50 3.39 1.27
 1.82 1.09 3.93 1.19 2.96 2.27 0.78 3.05 3.36 1.59 2.00 3.56 0.11 3.98 2.29 0.20
 2.13 0.78 3.37 2.51 2.63 0.79 3.37 0.49 0.44 2.97 1.26 3.76 1.14 1.35 0.56 2.93
 3.34 2.83 2.40 2.99 1.01 0.58 0.01 0.24 3.22 3.41 0.84 0.46 2.21 0.06 0.46 1.82 3.01 2.74 2.17 0.30 1.75 0.81 2.78 1.16 1.75 0.93 2.31 2.13 2.51 0.64 2.02 3.85
 2.78 3.70 0.76 1.34 0.71 3.98 1.83 3.99 0.39 2.50 0.38 1.75 3.73 0.19 3.58 1.16 0.91 3.08 1.64 0.81 2.51 2.42 1.81 1.87 2.39 2.54 3.42 3.32 2.50 2.88 2.26 1.50
 0.74 2.95 2.22 3.62 0.97 0.76 2.42 2.79 2.34 1.41 1.98 0.32 2.96 2.45 2.48 2.76
latrix B:
2.62 1.29 0.42 2.02 0.91 1.16 3.68 2.20 2.65 0.46 1.97 1.52 1.99 3.17 2.04 1.53
 2.75 2.13 2.43 1.58 0.02 2.83 0.40 2.49 3.45 1.97 2.99 1.99 1.52 3.14 2.21 1.43
 3.82 2.52 0.71 1.50 0.53 2.97 3.81 2.45 0.11 1.32 0.22 2.56 0.53 3.39 3.46 2.39
 2.89 3.42 0.06 0.51 2.83 2.47 0.87 0.26 0.68 2.50 1.36 1.28 1.47 2.64 3.21 3.23
 2.11 2.44 3.19 3.60 0.58 2.52 1.61 1.01 0.55 3.42 0.26 1.71 2.29 1.21 2.19 0.90 1.25 0.44 3.23 0.54 1.14 3.15 3.58 3.16 2.98 2.46 1.44 3.43 0.91 3.45 0.92 1.00 2.17 3.94 0.22 0.33 2.10 1.71 0.38 1.04 3.57 0.93 0.59 0.50 3.73 0.32 0.19 0.23
 1.35 3.66 1.59 1.73 3.78 3.35 2.14 3.37 2.77 1.59 1.04 0.02 2.10 3.82 1.59 0.96
 2.34 1.02 2.74 3.78 1.74 3.56 0.03 3.76 2.41 3.14 2.31 0.57 0.89 1.53 0.02 1.67
 0.33 2.64 3.42 0.26 3.24 2.65 2.77 3.21 2.12 2.74 0.57 2.76 2.91 3.11 0.12 3.47
 2.58 2.83 0.34 2.21 3.79 0.24 1.10 0.58 3.93 2.48 1.17 3.69 1.47 2.78 0.87 0.62 0.96 2.09 3.61 0.43 3.61 1.77 0.32 3.13 0.69 3.90 3.10 3.48 0.84 1.83 0.02 3.00
 0.46 1.62 1.24 3.97 0.15 1.01 0.76 0.99 0.61 2.48 3.54 3.76 0.78 3.12 2.58 2.63
70.80 70.44 60.55 55.63 53.52 71.79 58.92 74.93 70.70 69.21 45.94 61.33 58.39 76.84 39.43 51.63 49.17 55.72 39.75 45.64 41.92 50.29 33.63 54.18 54.02 49.74 42.96 47.97 38.99 63.23 32.08 42.80 75.95 76.97 64.24 58.84 65.36 87.00 65.14 83.46 76.66 76.39 53.56 62.77 57.28 88.26 47.16 62.48 81.42 72.16 62.48 61.42 69.52 80.93 66.88 74.54 61.90 67.46 55.16 64.62 55.39 89.55 51.41 72.63
 77.62 83.65 58.76 56.78 68.43 71.75 58.12 76.98 66.41 77.38 51.61 68.67 54.89 88.44 48.11 60.74
 71.03 68.16 63.49 70.94 66.00 76.27 55.70 78.00 72.34 68.76 58.16 63.07 59.81 82.01 35.69 65.14
 65.69 67.09 63.95 69.06 62.88 71.89 68.78 67.00 56.89 69.80 46.87 77.88 55.18 88.18 47.56 66.43
71.27 68.53 68.56 63.22 66.02 72.17 58.20 72.37 73.04 83.55 55.71 76.59 52.55 79.42 37.49 64.46 65.70 70.98 62.60 48.89 67.59 73.68 51.46 77.55 62.28 64.62 53.95 66.86 54.37 81.99 35.07 60.77 81.04 68.29 71.31 74.53 71.75 80.50 55.19 80.01 72.40 75.56 58.60 71.48 54.67 86.42 39.28 60.14 70.25 64.43 50.70 67.98 61.70 68.88 41.78 63.04 61.80 61.66 58.96 55.29 53.30 74.84 39.85 63.85
 56.59 49.83 47.54 48.85 48.80 53.69 56.22 58.36 59.48 50.12 43.27 52.23 49.73 61.82 24.73 50.78 63.07 65.14 60.88 64.05 63.59 68.71 50.39 67.48 56.20 74.25 57.66 64.72 49.31 77.56 38.70 69.60
 74.40 75.85 63.54 58.34 60.90 70.58 62.65 70.61 52.94 80.12 45.19 66.55 56.93 70.47 41.53 70.33
 77.63 83.83 66.00 64.84 72.50 80.28 59.14 85.22 76.03 79.55 57.21 69.55 60.86 90.57 43.11 64.04 67.19 74.27 59.60 60.77 65.49 64.51 62.49 72.07 63.26 70.87 56.01 72.25 59.15 87.55 44.04 64.86
```

4. Profilez chaque implementation.

## Tuile 2\*2:

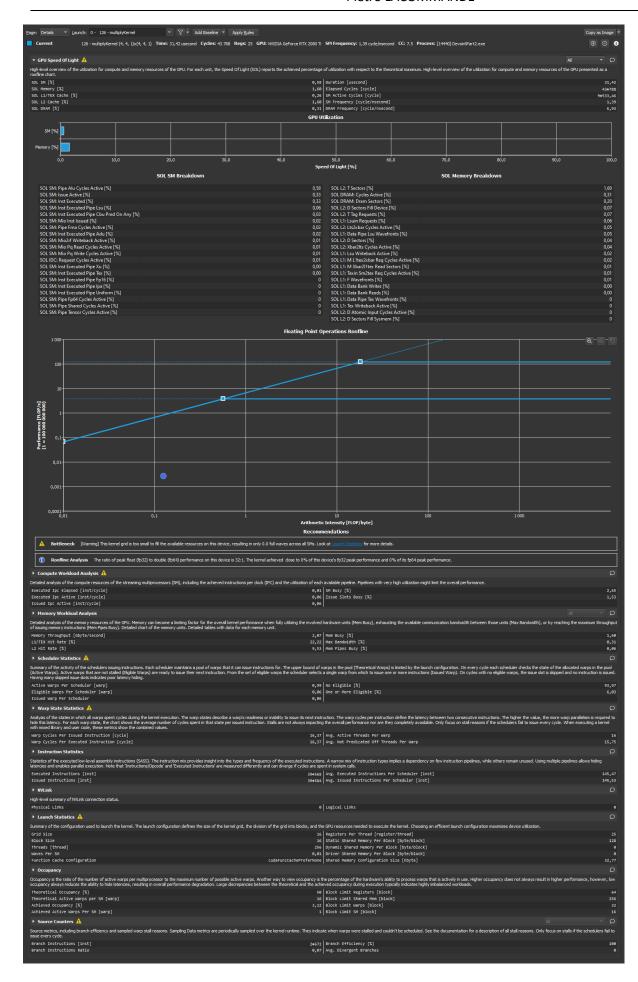
#### Nvprof:

```
Max Name
960us multiplyKernel(float*, float const *, float const *, int)
240us [CUDA memcpy HtoD]
960ns [CUDA memcpy DtoH]
.93ms cudaMalloc
                                                                                             Avg
4.8960us
1.4880us
                                                                                                                 4.8960us 4.8960us
1.1520us 1.8240us
                                                                                              1.4880us
960ns
960ns
50.980ms
110.17us
39.733us
68.100us
43.000us
154ns
4.7330us
6.1000us
                                                                                                                  960ns
5.7000us
50.947ms
51.500us
1.9000us
68.100us
                                                                                                                                      960ns
200.93ms
                                                                                                                                      200.93ms
50.947ms
218.00us
112.50us
68.100us
43.000us
1.2000us
1.5000us
1.1000us
                                                     200.94ms
50.947ms
330.50us
119.20us
68.100us
43.000us
15.600us
14.200us
6 1000us
                                                                                                                                                           cudaDeviceReset
cudaMemcpy
                                                                                                                                                           cudaFree
cudaLaunchKernel
                                       0.02%
0.01%
                                                                                                                  43.000us
100ns
                                                                                                                                                           cudaDeviceSynchronize
cuDeviceGetAttribute
                                                                                    101
3
1
                                       0.01%
0.00%
0.00%
0.00%
                                                     6.1000us
1.7000us
                                                                                                                  6.1000us
                                                                                                                                                           cudaSetDevice
                                                                                               6.1000us
                                                                                                                  200ns
1.1000us
400ns
                                                                                                                                                           cuDeviceGetName
cuDeviceGetLuid
                                                                                               1.1000us
                                                     1.1000us
                                                                                                                         300ns
200ns
                                                            300ns
200ns
                                                                                                                                             300ns
200ns
                                                                                                                                                           cuDeviceTotalMem
cuDeviceGetUuid
                                                                                                     300ns
                                                                                                     200ns
                                                            200ns
                                                                                                     200ns
                                                                                                                         200ns
                                                                                                                                              200ns
                                                                                                                                                           cudaGetLastError
   :\Users\lacommap\Desktop\CudaC\Devoir6Part2>
```



## Tuile 4\*4:

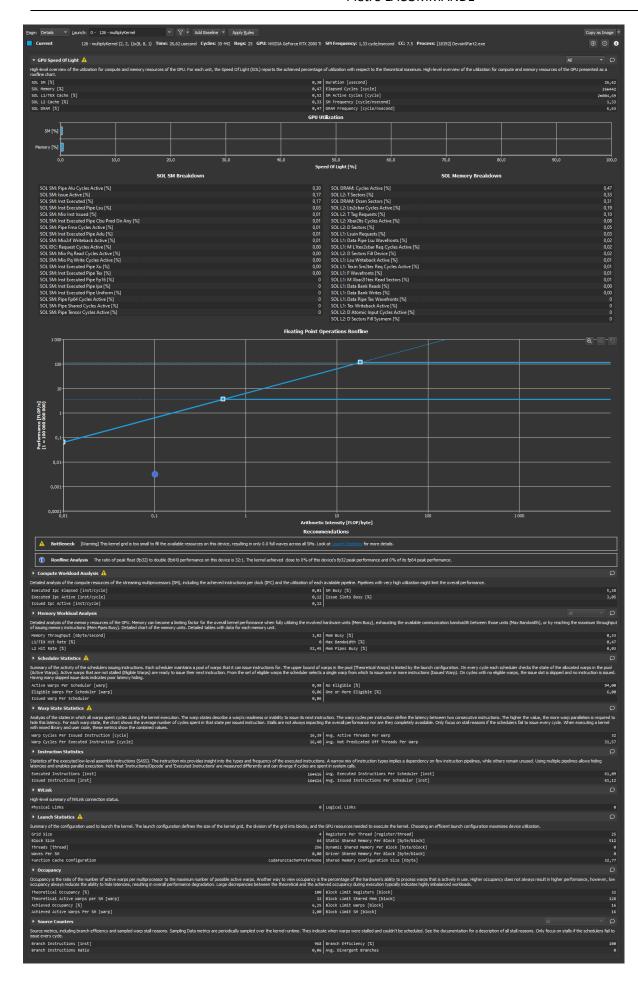
## Nvprof:



## Tuile 8\*8:

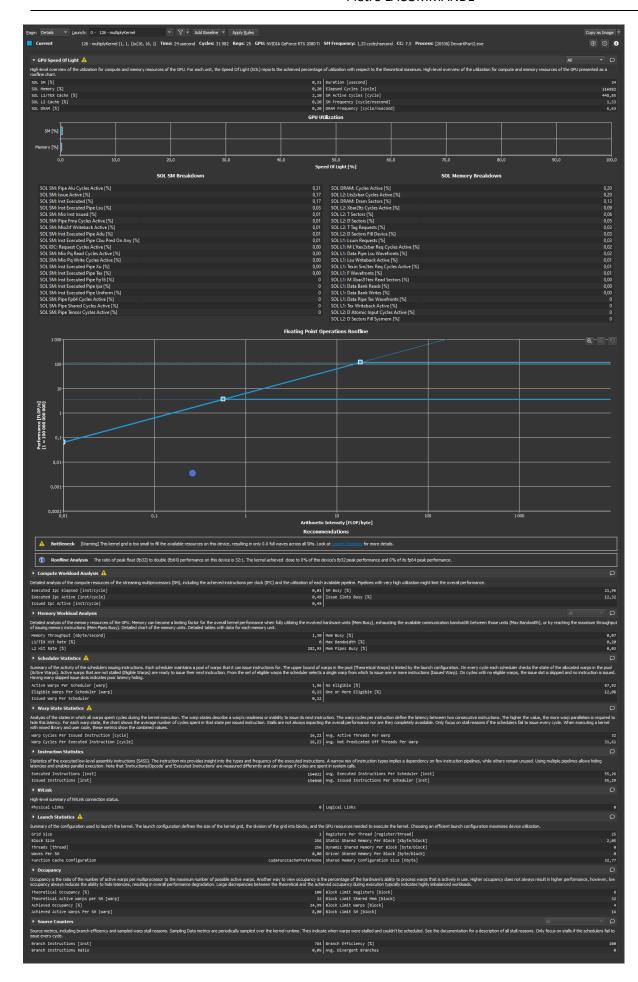
## Nvprof:

```
Avg
3.5840us
1.5520us
                                                                                                                                   name
multiplyKernel(float*, float const *, float const *, int)
[CUDA memcpy HtoD]
[CUDA memcpy DtoH]
cudaMalloc
cudaDeviceReset
                                                                                                 3.5840us
1.1200us
                                                                                                                  3.5840us
1.9840us
                                                                               928ns
66.275ms
52.262ms
95.033us
38.166us
52.800us
40.000us
                              12.18%
79.02%
20.77%
0.11%
0.05%
                                                                                                 928ns
1.8000us
52.262ms
33.900us
1.8000us
                                                                                                                  928ns
198.82ms
52.262ms
                                            928ns
198.83ms
        API calls:
                                             52.262ms
                                                                                                                  201.60us
108.30us
52.800us
40.000us
800ns
                                            285.10us
114.50us
                                                                                                                                   cudaMemcpy
cudaFree
                                0.02%
0.02%
0.01%
0.00%
0.00%
                                            52.800us
40.000us
                                                                                                 52.800us
                                                                                                                                    cudaLaunchKernel
                                                                                                 40.000us
                                                                                                                                   cudaDeviceSynchronize
cuDeviceGetAttribute
                                            14.200us
                                                                                    140ns
                                                                                                      100ns
                                                                                                                  5.3000us
5.5000us
1.3000us
600ns
                                                                               1.9660us
5.5000us
                                                                                                                                   cuDeviceGetCount
cudaSetDevice
                                             5.9000us
                                                                                                      200ns
                                                                                                 5.5000us
                                                                                                      100ns
600ns
300ns
300ns
200ns
                                0.00%
0.00%
                                            1.4000us
600ns
                                                                                    700ns
600ns
                                                                                                                                   cuDeviceGetName
                                                                                                                                    cuDeviceTotalMem
                                 0.00%
                                                  300ns
                                                                                      300ns
                                0.00%
0.00%
                                                  300ns
200ns
                                                                                                                        300ns
200ns
                                                                                                                                   cuDeviceGetLuid
cudaGetLastError
                                                                                     300ns
                                                                                     200ns
                                                                                                      100ns
                                0.00%
                                                  100ns
                                                                                     100ns
                                                                                                                        100ns
                                                                                                                                    cuDeviceGetUuid
 \Users\lacommap\Desktop\CudaC\Devoir6Part2>
```



## Tuile 16\*16:

# Nvprof:



5. Discutez de la vitesse d'exécution du kernel versus les ressources utilisées (shared memory, registres, etc.).

Je constante que lorsque le taille de tuiles augmente, plus que la fonction kernel a un temps d'exécution plus rapide, sauf lorsque le nombre de tuile équivaut à la taille de la matrice. De plus, on peut voir qu'on utilise de plus en plus la mémoire partagé lorsqu'on augmente la taille des tuiles. Cela est normal, car plus que la grosseur des tuiles est élevé, plus que nous devons charger les données dans la mémoire partagé depuis la mémoire globale, ce qui est diminue le nombre d'accès à la mémoire globale et augmente la rapidité d'exécution. Le nombre de registre par thread est constant de 25. Il est évident par contre, que ces implémentations avec plus de tuiles utilise plus de ressources, alors il est important de bien connaître le GPU avant d'exploiter toutes les ressources pour une vraie application.

6. Comparez et discutez des résultats par rapport à ceux obtenus à la PARTIE 1.

La méthode dans la partie 2 va vraiment plus vite que les méthodes dans la partie 1. La partie 1 fonctionne quand même sauf que celle-ci ne profite pas de la puissance d'un GPU. En effet, l'accès à la mémoire globale peut grandement ralentir la rapidité d'une application. Il est important de bien connaître les différents types d'accès mémoire et leur latence pour pouvoir créer des applications optimales.

# **REFERENCES**

[2] D. Kirk and W.-M. Hwu, *Programming Massively Parallel Processors- A Hands-on Approach*, 3rd Edition, Morgan Kaufmann, 2017. (Elsevier Science & Technology, ISBN: 978-0-12-811986-0).

GEI1084-Architecture des ordinateurs et calcul accéléré Jérémy POUPART and Messaoud AHMED OUAMEUR Pietro LACOMMANDE