Université de Pau et des Pays de l'Adour

Exercice 2 Chapitre 13

Master 2 Big Data

GPGPU

Realisé par :

Manon Bédère

Objectif de l'exercice : Parallelize a matrix-matrix multiplication algorithm using OpenACC. Consignes : Ecrire trois versions de code selon les consignes données. Ces codes développer sont à retrouver dans le fichier .zip.

Questions

1) Explain all your choices of optimisation in version 2 and 3.

Dans les versions 2 et 3 , j'ai décidé d'utiliser plusieurs fonctions afin d'optimiser le code de mutliplication de matrices. J'ai utiliseé :

- La clause collapse() car elle est utilisée pour une directive de boucle pour replier les N prochaines boucles en une même boucle.Elle sert dans les cas de boucles imbriquées ou quand les boucles sont très courtes comme dans notre code.
- La clause reduction () est utilisée sur des opérations de type addition, multiplication, max, min ... Dans notre cas on a une addition répétée.
- La clause gang (utilisée dans le code version 3), celle-ci permet le partage des itérations de la boucle ou des boucles à travers les gangs de la région parallèle. Dans notre code, Cela permettra d'utiliser le vecteur pour la boucle de multiplication. Cette clause permet l'optimisation.

2) Compare all the 5 versions of matrix multiplication you wrote so far : 2 CUDA versions from chapter 4 and 3 from this exercice. What is the best version ? EXPLAIN.

Pour la chapitre 4, 2 codes ont été développés pour répondre à cette problématique.

L'exercice 1 du chapitre 4 donne comme temps de calcul :

```
mbedere@scinfe051:/import/etud/26/mbedere/Bureau/gpgpu/Chapitre 4 exercice 1$ nvprof --print-gpu-trace ./1-basicMatMul 1000 1000 1000
Matrix multiplication dimensions: [1000;1000] = [1000;1000] x [1000;1000]
==18077== NVPROF is profiling process 18077, command: ./1-basicMatMul 1000 1000 1000
Ok
==18077== Profiling application: ./1-basicMatMul 1000 1000 1000
==18077== Profiling result:
   Start  Duration            Grid Size      Block Size     Regs*    SSMem*    DSMem*      Size  Throughput  SrcMemType  DstMemType           Device   Context    Stream
  Name
312.55ms  383.12us                    -               -         -         -         -  3.8147MB  9.7235GB/s    Pageable      Device  Quadro RTX 4000         1         7
  [CUDA memcpy HtoD]
313.02ms  370.00us                    -               -         -         -         -  3.8147MB  10.068GB/s    Pageable      Device  Quadro RTX 4000         1         7
  [CUDA memcpy HtoD]
313.40ms  1.2790us                    -               -         -         -         -  3.8147MB  2912.7GB/s      Device           -  Quadro RTX 4000         1         7
  [CUDA memset]
313.41ms  4.1338ms            (32 32 1)       (32 32 1)        48        0B        0B         -           -           -           -  Quadro RTX 4000         1         7
  dgemm(float*, float*, float*, int, int, int, int) [113]
317.55ms  909.70us                    -               -         -         -         -  3.8147MB  4.0951GB/s      Device    Pageable  Quadro RTX 4000         1         7
  [CUDA memcpy DtoH]

Regs: Number of registers used per CUDA thread. This number includes registers used internally by the CUDA driver and/or tools and can be more than what the compiler sh
ows.
SSMem: Static shared memory allocated per CUDA block.
DSMem: Dynamic shared memory allocated per CUDA block.
SrcMemType: The type of source memory accessed by memory operation/copy
DstMemType: The type of destination memory accessed by memory operation/copy
```

L'exercice 2 du chapitre 4 donne comme temps de calcul :

```
mbedere@scinfe051:/import/etud/26/mbedere/Bureau/gpgpu/Chapitre 4 exercice 2$ nvprof --print-gpu-trace ./1-basicMatMul 1000 1000 1000
Matrix multiplication dimensions: [1000;1000] = [1000;1000] x [1000;1000]
==18487== NVPROF is profiling process 18487, command: ./1-basicMatMul 1000 1000 1000
Ok
==18487== Profiling application: ./1-basicMatMul 1000 1000 1000
==18487== Profiling result:
   Start  Duration            Grid Size      Block Size     Regs*    SSMem*    DSMem*      Size  Throughput  SrcMemType  DstMemType           Device   Context    Stream
  Name
321.33ms  382.26us                    -               -         -         -         -  3.8147MB  9.7455GB/s    Pageable      Device  Quadro RTX 4000         1         7
  [CUDA memcpy HtoD]
321.80ms  371.54us                    -               -         -         -         -  3.8147MB  10.027GB/s    Pageable      Device  Quadro RTX 4000         1         7
  [CUDA memcpy HtoD]
322.18ms  1.2800us                    -               -         -         -         -  3.8147MB  2910.4GB/s      Device           -  Quadro RTX 4000         1         7
  [CUDA memset]
322.19ms  4.5851ms           (63 63 1)       (16 16 1)        20  2.0000KB        0B         -           -           -           -  Quadro RTX 4000         1         7
  dgemm(float*, float*, float*, int, int, int, int) [113]
326.78ms  921.25us                    -               -         -         -         -  3.8147MB  4.0437GB/s      Device    Pageable  Quadro RTX 4000         1         7
  [CUDA memcpy DtoH]

Regs: Number of registers used per CUDA thread. This number includes registers used internally by the CUDA driver and/or tools and can be more than what the compiler sh
ows.
SSMem: Static shared memory allocated per CUDA block.
DSMem: Dynamic shared memory allocated per CUDA block.
SrcMemType: The type of source memory accessed by memory operation/copy
DstMemType: The type of destination memory accessed by memory operation/copy
```

Pour le chapitre 13, 3 codes ont été développés pour cet exercice. Les temps observés pour ces 3 différents codes sont les suivants :

Temps obtenu pour le premier code :

```
mbedere@scinfe051:/import/etud/26/mbedere/Bureau/gpgpu/Chapitre13Exercice2/Question1$ nvprof --print-gpu-trace ./1-basicMatMul 1000 1000 1000
Matrix multiplication dimensions : [1000;1000] = [1000;1000] * [1000;1000]
==7632== NVPROF is profiling process 7632, command: ./1-basicMatMul 1000 1000 1000
Ok
==7632== Profiling application: ./1-basicMatMul 1000 1000 1000
==7632== Profiling result:
   Start  Duration            Grid Size      Block Size     Regs*    SSMem*    DSMem*      Size  Throughput  SrcMemType  DstMemType           Device   Context    Stream
  Name
300.59ms  312.89us                    -               -         -         -         -  3.8147MB  11.906GB/s      Pinned      Device  Quadro RTX 4000         1        14
  [CUDA memcpy HtoD]
301.29ms  312.12us                    -               -         -         -         -  3.8147MB  11.936GB/s      Pinned      Device  Quadro RTX 4000         1        14
  [CUDA memcpy HtoD]
301.68ms  56.775ms           (1000 1 1)       (128 1 1)        35        0B      512B         -           -           -           -  Quadro RTX 4000         1        14
  main_32_gpu [34]
358.48ms  316.25us                    -               -         -         -         -  3.8147MB  11.780GB/s      Device      Pinned  Quadro RTX 4000         1        14
  [CUDA memcpy DtoH]

Regs: Number of registers used per CUDA thread. This number includes registers used internally by the CUDA driver and/or tools and can be more than what the compiler sh
ows.
SSMem: Static shared memory allocated per CUDA block.
DSMem: Dynamic shared memory allocated per CUDA block.
SrcMemType: The type of source memory accessed by memory operation/copy
```

Temps obtenu pour le deuxième code :

```
mbedere@scinfe051:/import/etud/26/mbedere/Bureau/gpgpu/Chapitre13Exercice2/Question2$ nvprof --print-gpu-trace ./1-basicMatMul 1000 1000 1000
Matrix multiplication dimensions : [1000;1000] = [1000;1000] * [1000;1000]
==9663== NVPROF is profiling process 9663, command: ./1-basicMatMul 1000 1000 1000
Ok
==9663== Profiling application: ./1-basicMatMul 1000 1000 1000
==9663== Profiling result:
   Start  Duration            Grid Size      Block Size     Regs*    SSMem*    DSMem*      Size  Throughput  SrcMemType  DstMemType           Device   Context    Stream
  Name
344.17ms  313.33us                    -               -         -         -         -  3.8147MB  11.889GB/s      Pinned      Device  Quadro RTX 4000         1        14
  [CUDA memcpy HtoD]
344.86ms  312.15us                    -               -         -         -         -  3.8147MB  11.934GB/s      Pinned      Device  Quadro RTX 4000         1        14
  [CUDA memcpy HtoD]
345.25ms  56.888ms          (65535 1 1)       (128 1 1)        32        0B      512B         -           -           -           -  Quadro RTX 4000         1        14
  main_32_gpu [34]
402.17ms  316.53us                    -               -         -         -         -  3.8147MB  11.769GB/s      Device      Pinned  Quadro RTX 4000         1        14
  [CUDA memcpy DtoH]

Regs: Number of registers used per CUDA thread. This number includes registers used internally by the CUDA driver and/or tools and can be more than what the compiler sh
ows.
SSMem: Static shared memory allocated per CUDA block.
DSMem: Dynamic shared memory allocated per CUDA block.
SrcMemType: The type of source memory accessed by memory operation/copy
```

Temps obtenu pour le dernier code :

```
mbedere@scinfe051:/import/etud/26/mbedere/Bureau/gpgpu/Chapitre13Exercice2/Question3$ nvprof --print-gpu-trace ./1-basicMatMul 1000 1000 1000
Matrix multiplication dimensions : [1000;1000] = [1000;1000] * [1000;1000]
==10210== NVPROF is profiling process 10210, command: ./1-basicMatMul 1000 1000 1000
Ok
==10210== Profiling application: ./1-basicMatMul 1000 1000 1000
==10210== Profiling result:
   Start  Duration            Grid Size      Block Size     Regs*    SSMem*    DSMem*      Size  Throughput  SrcMemType  DstMemType           Device   Context    Stre
  Name
305.42ms  312.98us                    -               -         -         -         -  3.8147MB  11.903GB/s      Pinned      Device  Quadro RTX 4000         1
  [CUDA memcpy HtoD]
306.11ms  312.37us                    -               -         -         -         -  3.8147MB  11.926GB/s      Pinned      Device  Quadro RTX 4000         1
  [CUDA memcpy HtoD]
306.50ms  56.886ms          (65535 1 1)       (128 1 1)        32        0B      512B         -           -           -           -  Quadro RTX 4000         1
  main_32_gpu [34]
363.43ms  316.53us                    -               -         -         -         -  3.8147MB  11.769GB/s      Device      Pinned  Quadro RTX 4000         1
  [CUDA memcpy DtoH]

Regs: Number of registers used per CUDA thread. This number includes registers used internally by the CUDA driver and/or tools and can be more than what the compiler
ows.
SSMem: Static shared memory allocated per CUDA block.
DSMem: Dynamic shared memory allocated per CUDA block.
SrcMemType: The type of source memory accessed by memory operation/copy
```

Les meilleurs temps obtenus sont obtenus avec les codes du chapitre 13 en utilisant OpenACC. Cela peut provenir du fait que OpenACC est spécifique pour les accélérateurs.