Lập trình song song trên GPU

Hoàng Minh Thanh (18424062)

BT2 : Nhân hai ma trận

Bộ môn Công nghệ phần mềm

Khoa Công nghệ thông tin

Đại học Khoa học tự nhiên TP HCM

Contents

1 Quá trình cài đặt

Vì máy tính cả nhân của em không có GPU nên bắt buộc em phải sử dụng Google Colab : https://colab.research.google.com/drive/17TOLU6_vHtm0BI_joh1VszJj1GNcZ3oJ#scrollTo=gn39NrTLwFUI

(Thầy có thể vào link Online để xem luồng chạy dễ hơn báo cáo)

a) Chương trình CUDA nhân hai ma trận :

Cài đặt ở file **multipleMatrixCuda.cu**

Đây là kết quả cộng 2 ma trận với kích thước 3x3

```
Input 1 :
0.278576 0.404004 0.852235
0.761552 0.698785 0.339177
0.074193 0.923960 0.743110
Input 2 :
0.848834 0.937748 0.271307
0.785023 0.377963 0.901248
0.066271 0.237196 0.872775
Time host : 0 milliseconds
Time device : 53 milliseconds
Output :
1.127410 1.341752 1.123541
1.546574 1.076748 1.240425
0.140463 1.161156 1.615885
```

b) Chương trình đo thời gian chạy trên device ở với nhiều kích thước khác nhau :

Cài đặt ở file **calcTimeOnDevice.cu**

Đây là phần cài đặt đo thời gian khi tính toán trên nhiều gridSize và blockSize khác nhau:

```
Function        |Block size     |Grid size      |Time (ms)
addMatOnHost    |               |               |407174
addMatOnDevice2D|32 x 32        |257 x 257      |25871 ms
addMatOnDevice2D|16 x 32        |513 x 257      |25672 ms
addMatOnDevice2D|32 x 16        |257 x 513      |26548 ms
addMatOnDevice2D|16 x 16        |513 x 513      |26206 ms
addMatOnDevice1D|32 x 1 |257 x 1        |28852 ms
addMatOnDevice1D|64 x 1 |129 x 1        |30367 ms
addMatOnDevice1D|128 x 1        |65 x 1 |29085 ms
addMatOnDevice2DNotMix|32 x 1   |257 x 8193     |27747 ms
addMatOnDevice2DNotMix|64 x 1   |129 x 8193     |25729 ms
addMatOnDevice2DNotMix|128 x 1  |65 x 8193      |26264 ms
addMatOnDevice2DNotMix|256 x 1  |33 x 8193      |25777 ms
addMatOnDevice2DNotMix|512 x 1  |17 x 8193      |25388 ms
addMatOnDevice2DNotMix|1024 x 1 |9 x 8193       |26081 ms
addMatOnDevice2DNotMix|2048 x 1 |5 x 8193       |24975 ms
addMatOnDeviceMix|32 x 1        |257 x 8193     |27975 ms
addMatOnDeviceMix|64 x 1        |129 x 8193     |27432 ms
addMatOnDeviceMix|128 x 1       |65 x 8193      |26223 ms
addMatOnDeviceMix|256 x 1       |33 x 8193      |26115 ms
addMatOnDeviceMix|512 x 1       |17 x 8193      |25985 ms
addMatOnDeviceMix|1024 x 1      |9 x 8193       |26380 ms
addMatOnDeviceMix|2048 x 1      |5 x 8193       |25187 ms
```

c) Chương trình đo thông tin Device trong CUDA

Cài đặt ở file : **getDeviceInfo.cu**

Kết quả truy vấn thông tin trên Cuda

Có tham khảo cài đặt ở địa chỉ : https://gist.github.com/stevendborrelli/4286842

```
Số lượng CUDA device : 1.
****************** DEVICE 1 ******************
Có một device hỗ trợ CUDA
Tên Device: Tesla P100-PCIE-16GB
Số revision nhiều: 6
Số revision nhỏ: 0
Tổng kích thước bộ nhớ toàn cục : -108134400
Tổng kích thước bộ nhớ chia sẻ trên một block : 49152
Tổng kích thước bộ nhớ hằng : 65536
Kích thước Warp: 32
Kích thước block tối đa: 1024 x 1024 x 64
Kích thước grid tối đa: 2147483647 x 65535 x 65535
Tỉ lệ đồng hồ: 1328500
Số lượng đa xử lý: 56
```

2 Báo cáo và rút ra nhận xét, kết luận

a) Đề cài đặt chương trình chạy trên GPU thì ta cần :

1. Khởi tạo dữ liệu ltrên host (CPU)
2. Khởi tạo bộ nhớ cho các biến tính toán trên device (GPU)
3. Copy dữ liệu từ host sang device
4. Thực hiện goi hàm tính toán trên device
5. Sau khi thực hiện xong thì copy kết quả từ device sang host
6. Xuất kết quả từ host

b) Nhân xét và gridSize và blockSize :

Có 3 hướng cấu hình blockSize và gridSize

- Mở rộng grid2D và block2D
- Mở rộng grid1D và block1D
- Mở rộng grid2D và block1D

Từ kết quả phép tính toán trên có thể thấy :

- Phương pháp mix (kết hợp) là phương pháp có kết quả tốt nhất
- Số nhân càng nhiều thì sức tính toán càng nhanh

Rút ra so sánh và kết luận (**Đối với GPU trên Google Colab**) :

Các phương pháp :

- Mở rộng grid2D và block2D : số nhân càng nhiều càng chạy nhanh (26003 ms -> 25970 ms)
- Mở rộng grid1D và block1D : số grid càng nhiều càng tốt (29156 ms -> 29643 ms)
- Mở rộng grid2D và block1D : số block càng cao và số grid càng giảm thì càng tốt (30085 ms -> 23598 ms)

c) Kết quả thông tin của một device của Google Colab Pro

- Tên Device: Tesla P100-PCIE-16GB
- Số revision nhiều: 6
- Số revision nhỏ: 0
- Tổng kích thước bộ nhớ toàn cục : -108134400
- Tổng kích thước bộ nhớ chia sẻ trên một block : 49152
- Tổng kích thước bộ nhớ hằng : 65536
- Kích thước Warp: 32
- Kích thước block tối đa: 1024 x 1024 x 64
- Kích thước grid tối đa: 2147483647 x 65535 x 65535
- Tỉ lệ đồng hồ: 1328500
- Số lượng đa xử lý: 56