

# Lập trình song song trên GPU

## BT02: Cách thực thi song song trong CUDA



---

Nên nhớ mục tiêu chính ở đây là **học, học một cách chân thật**. Bạn có thể thảo luận ý tưởng với bạn khác, nhưng **bài làm phải là của bạn, dựa trên sự hiểu thật sự của bạn**. **Nếu vi phạm thì sẽ bị 0 điểm cho toàn bộ môn học**.

---

### Đề bài

Trong bài này, bạn sẽ áp dụng những hiểu biết về cách thực thi song song trong CUDA để viết chương trình giải quyết bài toán “reduction” (cụ thể là, ta sẽ tính tổng của một mảng số nguyên). Mình đã viết sẵn cho bạn khung chương trình trong file “bt02.cu” đính kèm; bạn chỉ viết code ở những chỗ có từ “// TODO”:

- Hàm kernel 1, 2, và 3 (bạn xem nội dung cụ thể của các hàm kernel này trong slide “06-CachThucThiTrongCUDA\_P3.pdf”; ở đây, để đơn giản, ta giả định  $2 * \text{kích-thước-block} = 2^k$  với  $k$  là một số nguyên dương nào đó). Hàm kernel sẽ tính giá trị tổng cục bộ trong từng block; sau đó, host sẽ chép các giá trị tổng cục bộ này về bộ nhớ của mình và cộng hết lại để ra giá trị tổng toàn cục.
- Tính “gridSize” (từ “blockSize” và “n”) khi gọi hàm kernel.

Với mỗi hàm kernel, chương trình sẽ in ra:


- Kích thước grid và kích thước block.
- Thời gian chạy của hàm kernel (“kernel time”) và thời gian host thực hiện cộng các giá trị tổng cục bộ của các block (“post-kernel time”).
- “CORRECT” nếu kết quả tính được giống với kết quả đúng, “INCORRECT” nếu ngược lại.

Hướng dẫn về các câu lệnh (phần hướng dẫn dưới đây là cho Linux, nếu bạn nào dùng GPU cá nhân trên Windows thì cũng tương tự, chỉ khác là: file chạy sau khi biên dịch có đuôi exe, và khi chạy file này thì dùng `.\` thay vì `./`):

- Biên dịch file “bt02.cu”: `nvcc bt02.cu -o bt02`
- Chạy file “bt02”: `./bt02`  
Mặc định thì sẽ dùng block có kích thước 512; nếu bạn muốn dùng block có kích thước khác, chẳng hạn 256, thì bạn truyền thêm tham số dòng lệnh: `./bt02 256`

Cụ thể, các yêu cầu của bài tập này như sau:

### Câu 1 (1.5đ)

Hoàn thành hàm kernel 1 và phần tính “gridSize” (khi gọi hàm kernel) trong file “bt3.cu”. Trong file bài làm, bạn ghi nhận lại kết quả chạy bằng cách chụp lại màn hình (để chụp lại một phần màn hình, trong Windows 10, bạn có thể ấn +Shift+S, rồi ấn giữ chuột trái và kéo chọn vùng cần chụp, rồi Ctrl+V vào file word là xong.)

### Câu 2 (2đ)

Chạy chương trình ở câu 1 (ta đang chỉ tập trung vào hàm kernel 1, hàm kernel 2 và 3 thì tạm chưa đụng đến) với các kích thước block khác nhau: 1024, 512, 256, 128. Với mỗi kích thước block, bạn điền các kết quả theo mẫu bảng biểu bên dưới (trong đó, Total time = Kernel time + Post-kernel time). Với số block / SM và occupancy, ngoài việc điền kết quả vào bảng thì bạn cũng cần giải thích thêm là tại sao lại tính ra được như vậy. Khi tính số block / SM và occupancy, tạm thời ta chỉ xét 2 ràng buộc của SM là số block tối đa và số thread tối đa; bạn có thể tra cứu 2 ràng buộc này ở [document của CUDA](#), mục “Programming Guide” (“Guide” **không có s**), mục “H. Compute Capabilities”, bảng “Table 14. Technical Specifications per Compute Capability”, 2 ràng buộc đó là “Maximum number of resident blocks per multiprocessor” và “Maximum number of resident threads per multiprocessor”.

Block size	Grid size	Num blocks / SM	Occupancy (%)	Kernel time (ms)	Post-kernel time (ms)	Total time (ms)
1024	8193					
512	16385					
256	32769					
128	65537					

Giải thích tại sao khi thay đổi block size thì “kernel time” và “post-kernel time” lại thay đổi như vậy?

### Câu 3 (3đ)

Hoàn thành hàm kernel 2 và 3 trong file “bt02.cu”. Trong file bài làm, bạn ghi nhận lại kết quả chạy (tương tự câu 1; để kích thước block là 512).

### Câu 4 (1.5đ)

Giả sử block có kích thước là 128. Với mỗi hàm kernel: với mỗi giá trị “stride”, cho biết trong mỗi block có những warp nào bị phân kỳ (không xét block cuối).

### Câu 5 (2đ)

Hoàn thành file “bt02\_p2.cu” (những chỗ mình để “// TODO”) để thực hiện “reduce” hoàn toàn bằng device (cách biên dịch và chạy file này tương tự như “bt02.cu”). Trong file bài làm, ghi nhận lại kết quả chạy bằng cách chụp lại màn hình.

## Nộp bài

Bạn tổ chức thư mục bài nộp như sau:

- Thư mục <MSSV> (vd, nếu bạn có MSSV là 1234567 thì bạn đặt tên thư mục là 1234567)
  - File code “bt02.cu” ứng với code của câu 1 và 3
  - File code “bt02\_p2.cu” ứng với code của câu 5
  - File bài làm “bt02.pdf” (ở đầu file bạn ghi họ tên và MSSV)

Sau đó, bạn nén thư mục <MSSV> này lại và nộp ở link trên moodle.