Practical Parallel Computing (実践的並列コンピューティング)

Part 2: GPU
No 4: GPU Architecture

May 20, 2023

Toshio Endo

School of Computing & GSIC

endo@is.titech.ac.jp

Overview of This Course

- Part 0: Introduction
  - 2 classes
- Part 1: OpenMP for shared memory programming
  - 4 classes
- Part 2: GPU programming
  - 4 classes ← We are here (4/4)
  - OpenACC (1.5 classes) and CUDA (2.5 classes)
- Part 3: MPI for distributed memory programming
  - 4 classes

Comparing OpenMP/OpenACC/CUDA

	OpenMP	OpenACC	CUDA
Processors	CPU	CPU+GPU	
File extension	.c, .cc		.cu
To start parallel (GPU) region	#pragma omp parallel	#pragma acc kernels	func<<<…, …>>>()
To specify # of threads	export OMP_NUM _THREADS=…	(num_gangs, vector_length etc)	
Desirable # of threads	# of CPU cores or less	# of GPU cores or “more”	
To get thread ID	omp_thread_num()	-	blockIdx, threadIdx
Parallel for loop	#pragma omp for	#pragma acc loop	-
Task parallel	#pragma omp task	-	-
To allocate device memory	-	#pragma acc data	cudaMalloc()
To copy to/from device memory	-	#pragma acc data #pragma acc update	cudaMemcpy()
Function on GPU	-	#pragma acc routine	__global__, __device__

※ “# of XXX” = “The number of XXX”

Speed of GPU Programs and GPU Architecture

Advanced topic

Case 1: How should block-size be determined?

When creating 1,000,000 threads,

- <<<1, 1000000>>> causes an error
  - blockDim must be <= 1024
- <<<1000000, 1>>> can work, but slow
- <<<1000, 1000>>> is faster → Why?

Case 2: How should each thread access memory?

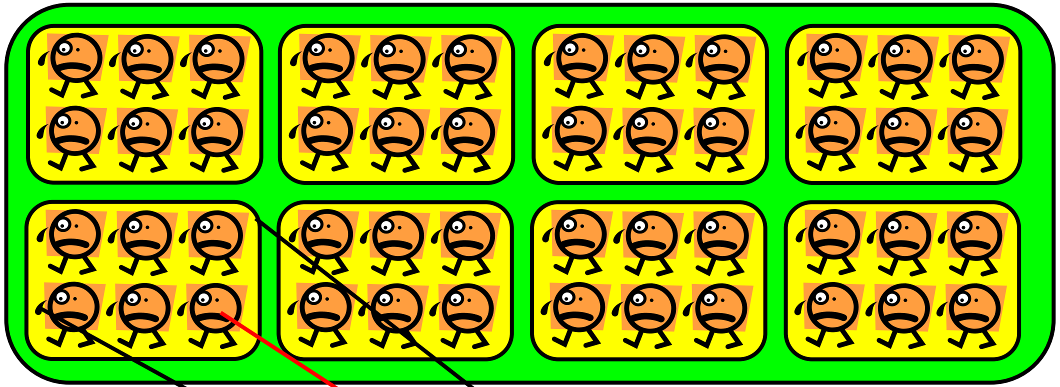
- In mm-cuda, (x = row,y = col) and (x = col, y = row) shows different speed

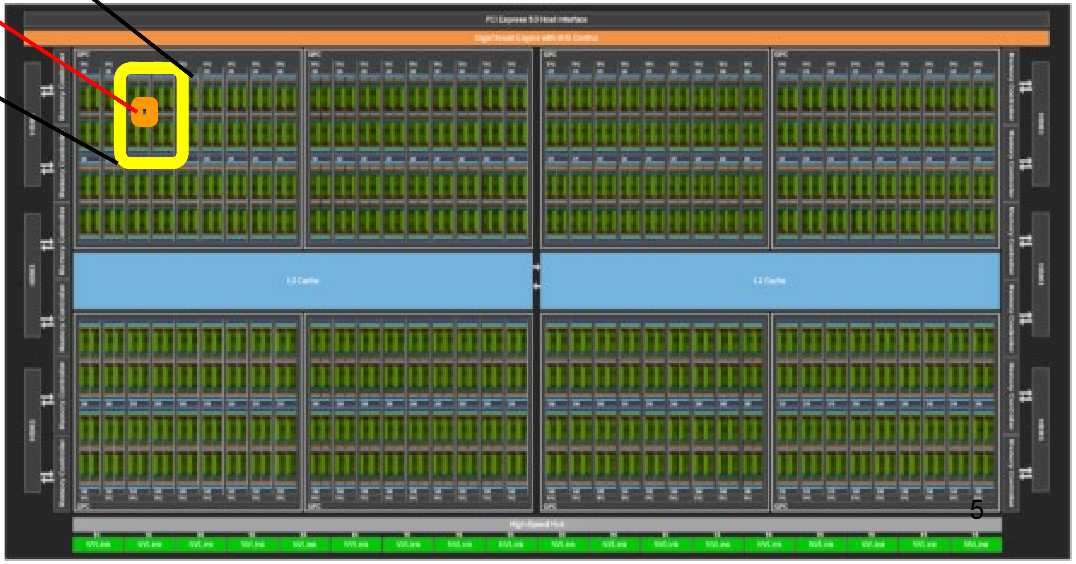
Knowledge of GPU architecture helps understanding of speeds

Why Do We Have to Specify both gridDim and blockDim?

- and why did NVIDIA decide so?

→ Hierarchical structure of GPU processor is considered




PCI Express 5.0 Host Interface
GigaThread Engine with MIG Control
L2 Cache
L2 Cache
High-Speed Hub
NVLink

Structure of H100 SXM GPU

1 GPU = 132 SMs
1 SM = 128 CUDA cores

→ 1GPU=16,896 CUDA cores

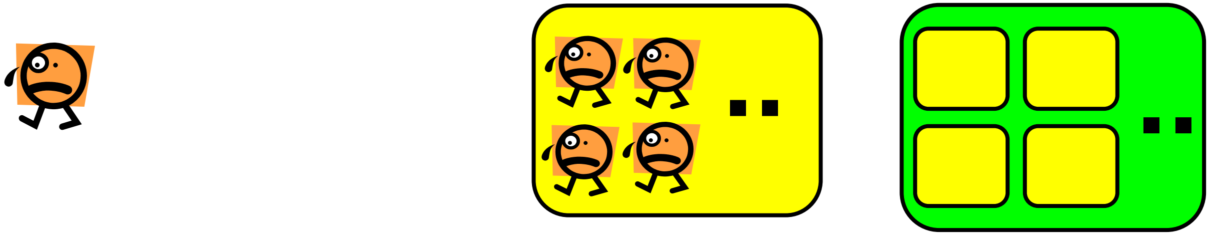
In TSUBAME4 interactive node (1/2 GPU), ~7680 CUDA cores

Mapping between Threads and Cores

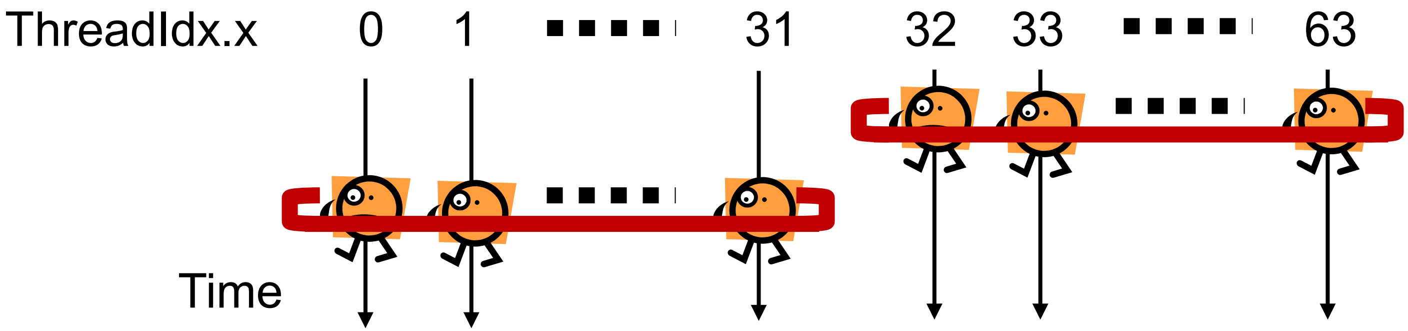
- 1 thread blocks (or more) run on 1 SM
  - At least 132 blocks are needed to use all SMs on H100
  - gridDim (gx*gy*gz) should be ≧132
- 1 thread (or more) run on a CUDA core
  - At least 132*128=16,896 threads in total are needed to use all CUDA cores on H100
  - Total threads (gx*gy*gz * bx*by*bz) should be ≧16896
- 32 consective threads (in a block) are batched (called a *warp*) and scheduled
  - At least 32 threads per block are needed for performance
  - blockDim (bx*by*bz) should be ≧32

Warp: Internal Execution Unit

thread < warp < thread block < grid

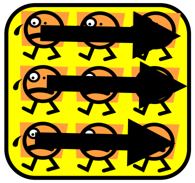


- Threads in a thread block are internally divided into “warp”, a group of contiguous 32 threads
- 32 threads in a warp always are executed synchronously
  - They execute the same instruction simultaneously
  - Only 1 program counter for 32 threads → GPU hardware is simplified
  - Actually 32 threads are executed on 16 CUDA cores



Observations due to Warps

- If number of threads per block (blockDim) is not *32 x n*, it is inefficient
  - Even if blockDim=1, the system creates a warp for it
- Characteristics in memory addresses accessed by threads in a warp affect the performance
  - Coalesced accesses are fast

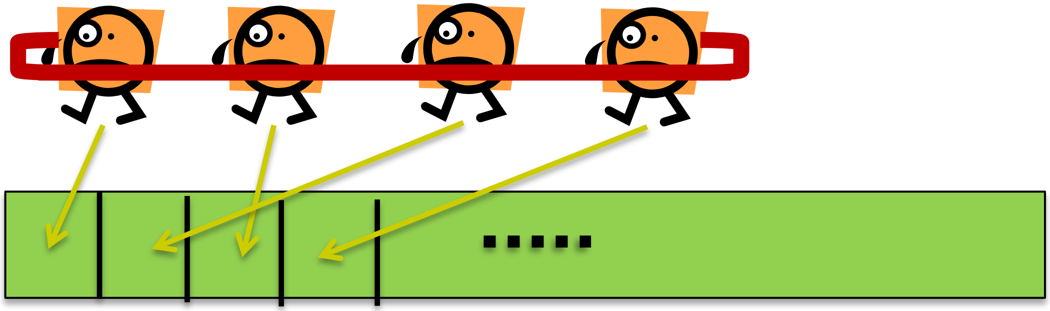


※ In multi-dimensional cases (blockDim.y>1 or blockDim.z>1), “neighborhood” is defined by x-dimension

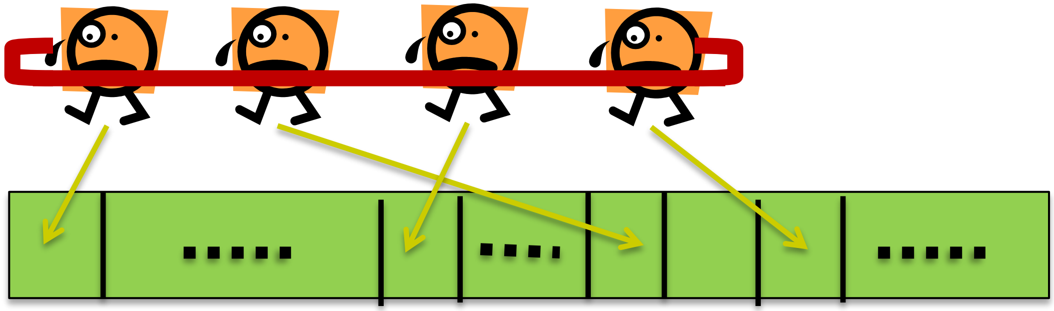
Coalesced Memory Access

- When threads in a warp access "neighbor" address on memory (coalesced access), it is more efficient

Coalesced access
→ **Faster**



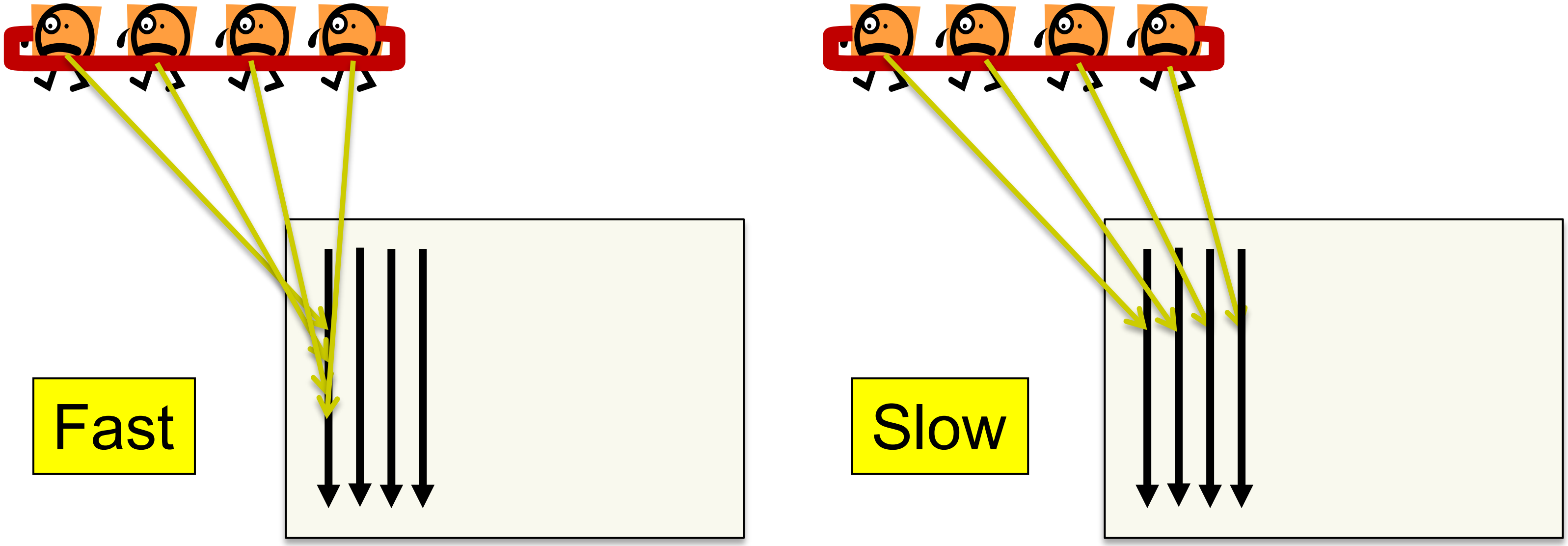
Non-coalesced access
→ **Slower**



Accesses in mm-cuda Sample

- mm-cuda: (x = row,y = col) → coalesced and fast
- mm-nc-cuda: (x = col, y = row) → non-coalesced and slow
  - /gs/bs/tga-ppcomp/24/mm-nc-cuda

We should see “what data are accessed by threads in a warp simultaneously”



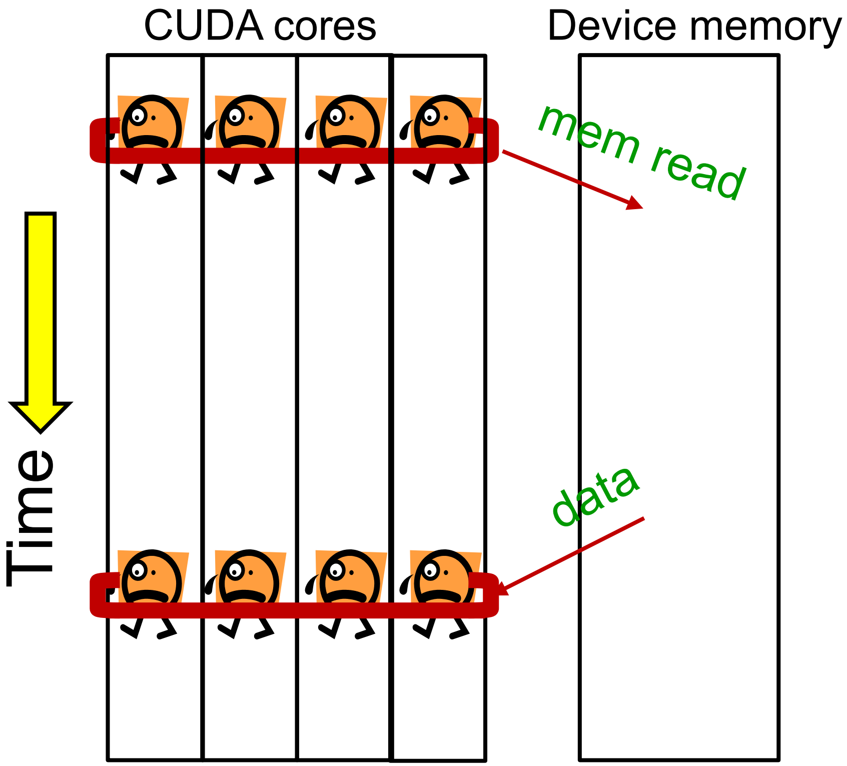
matrices in column-major format

Why #threads >> #cores Works Well on GPUs?

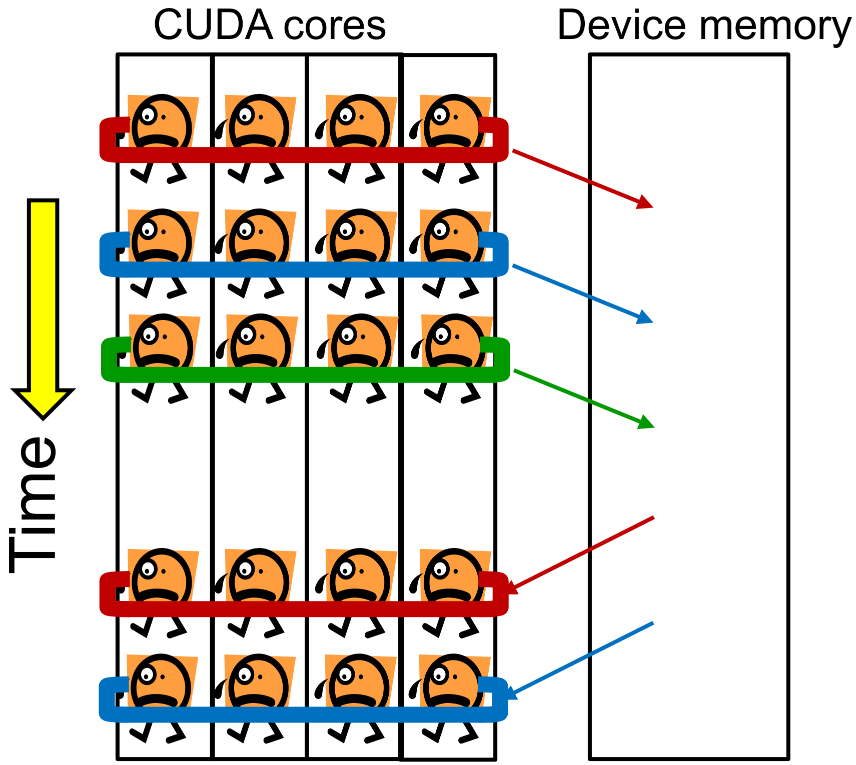
- GPU supports very fast (~1 clock) context switches

➔ With many threads, memory access latency can be hidden

#threads == #cores



#threads > #cores

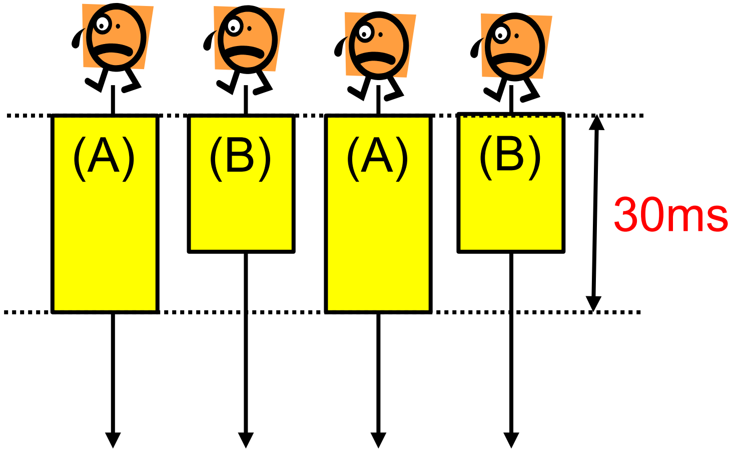


Considering Branches in Parallel Programs

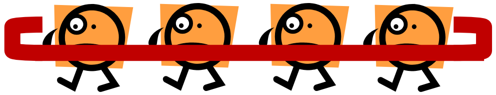
Consider this code. How long is execution time?

```
if (thread-id % 2 == 0) {
        :  // (A) 30msec
} else {
        :  // (B) 20msec
}
```

On CPU (OpenMP)

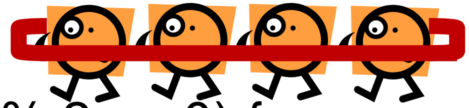


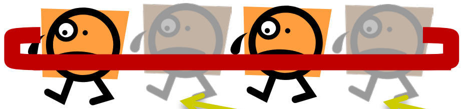
On GPU, threads in a warp must execute the same instruction. What happens?

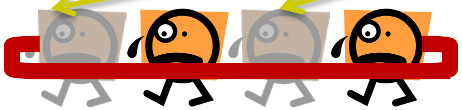


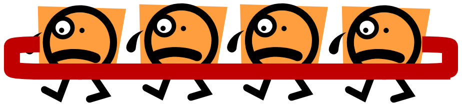
Branches on GPU (1)

```
     :
     :
if (thread-id % 2 == 0) {
     :
     :
     :
} else {
     :
     :
     :
}
     :
     :
```









Some threads are made sleep

Both “then” and “else” are executed!

→ Answer to previous question is 50ms !

※ Similar cases happen in for, while…

Branches on GPU (2)

- As exceptional cases, if threads in a warp “agree” in branch condition, either “then” part or “else” part is executed → Efficient!
- If there is difference of opinion (previous page), it is called a divergent branch

→ Agreement among buddies (threads in a warp) is important for speed

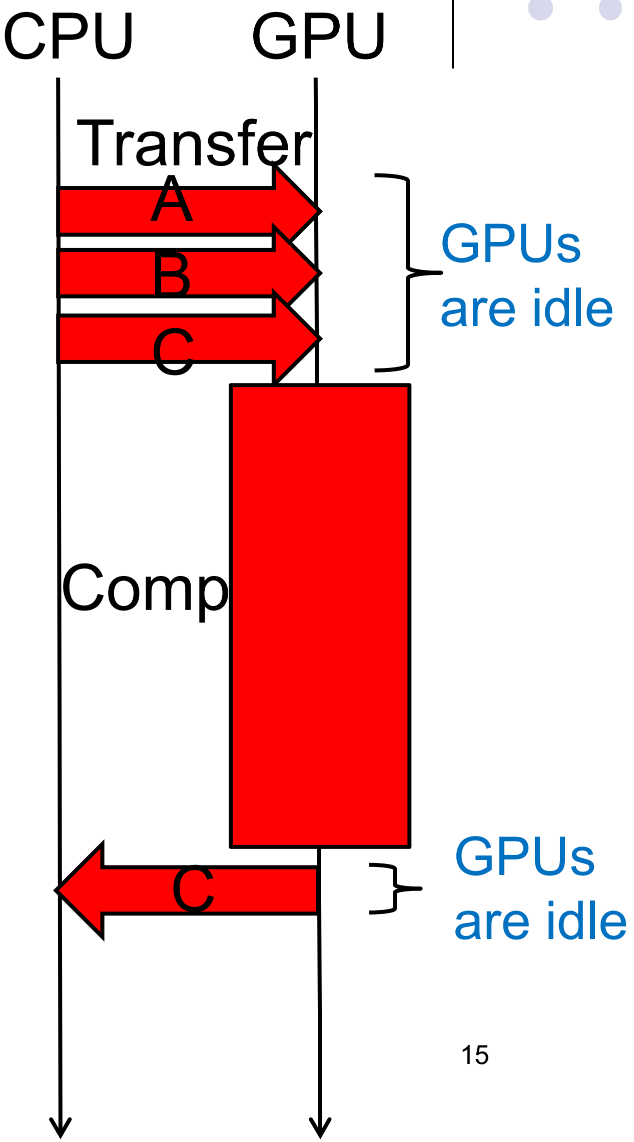
Considering Data Transfer Costs of mm Sample

- In mm sample, the speed is degraded by data transfer costs ☹

- This can be improved by combination of:

1. Split computation
2. Using CUDA streams

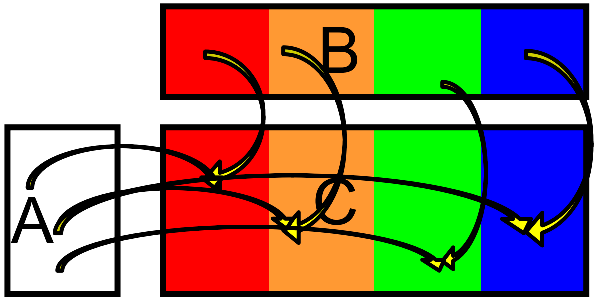
→ The faster sample is at /gs/bs/tga-ppcomp/24/mm-str-cuda/



Split mm Computation (1)

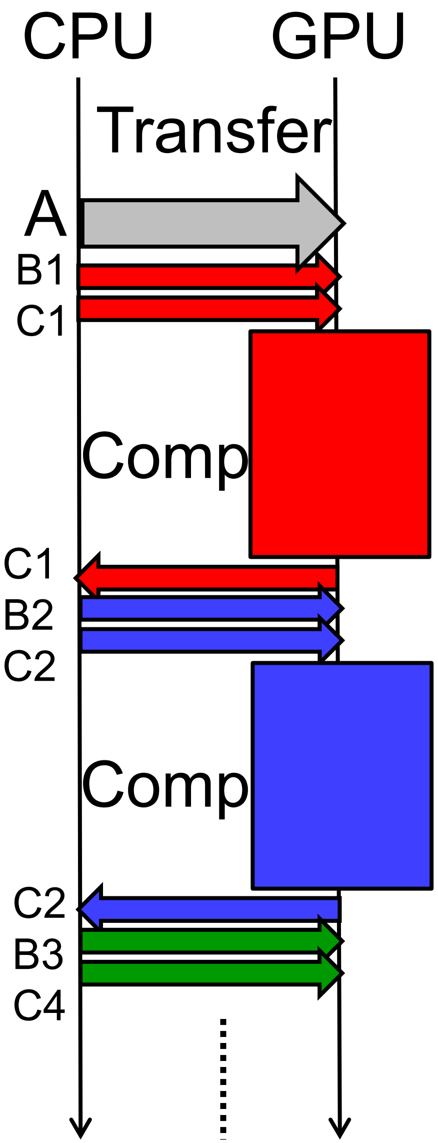
- Computation of “C 🡐 A×B” is split by splitting B and C vertically
  - C_1 🡐 A × B_1, C_2 🡐 A × B_2 , … , C_n 🡐 A × B_n

The n computations are independent each other



A is reused for all computations

Split mm Computation (2)



Algorithm:

(1) Copy A from CPU to GPU

(2) For each partition i (sequentially)

  (1) Copy Bi and Ci to GPU

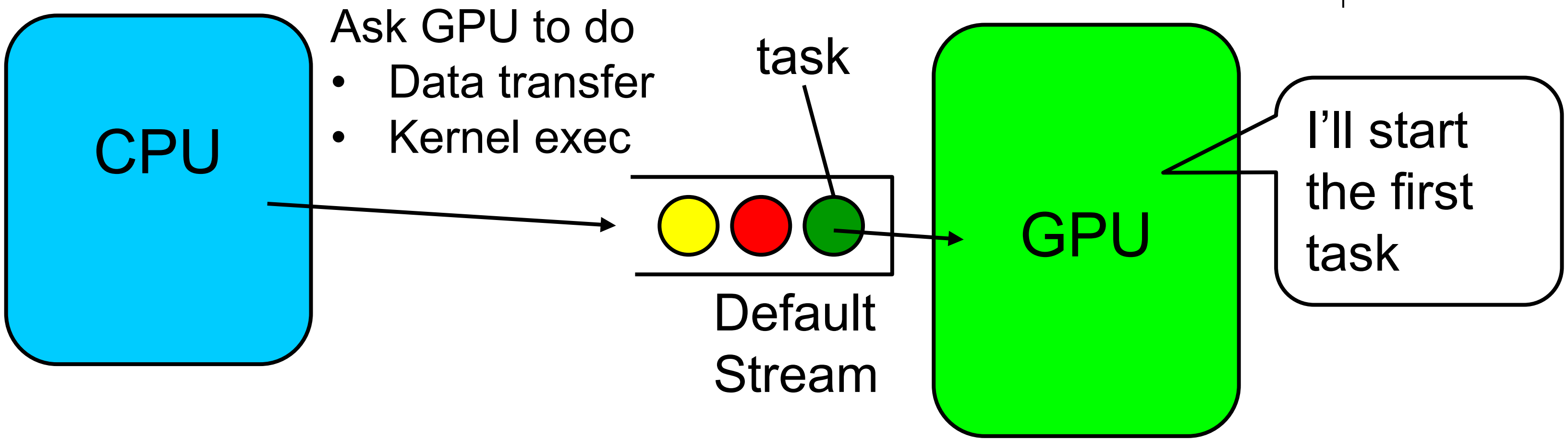
  (2) Compute $C_i \leftarrow A \times B_i$

  (3) Copy back Ci to GPU

This does NOT improve speed yet, since neither total computation costs nor total transfer costs change

→ cudaStream is useful for hiding transfer costs

How GPU Executes Tasks
(Without multiple streams)

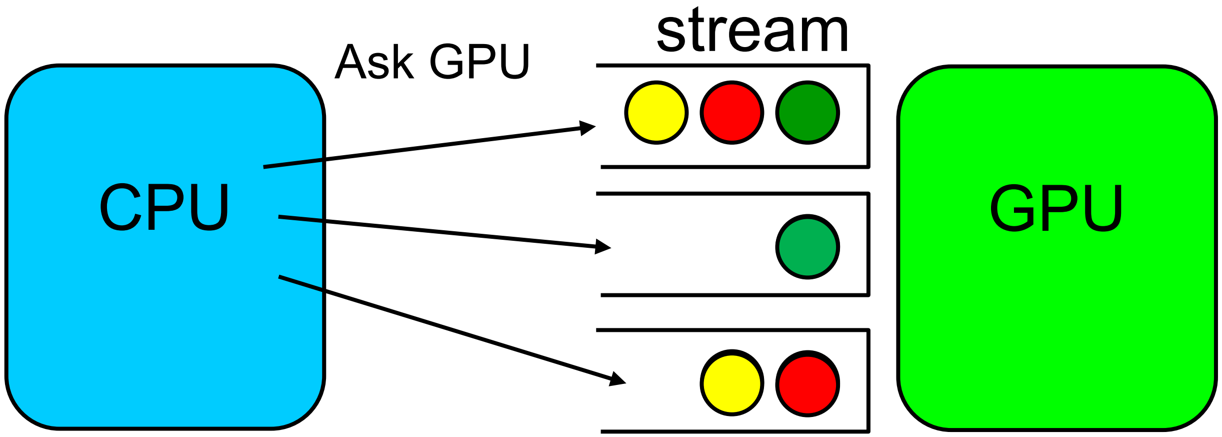


- A GPU is idle until asked to do something by CPU
- CPU asks the GPU to do one of followings (called tasks here)
  - Data transfer (Host → Device) or
  - GPU Kernel function execution or
  - Data transfer (Device → Host)
- Then the task is put on a FIFO queue, called *default stream*
- GPU takes a task from the stream and executes it in FIFO

Asynchronous Executions with cudaStream (1)

What are streams?

- GPU's "service counters" that accept tasks from CPU
  - In addition to default stream user program can create streams,
  - Each stream looks like a FIFO queue



All of CUDA sample programs, except mm-str-cuda, are using the single "default stream"

Asynchronous Executions with cudaStream (2)

Create a stream

```
cudaStream_t str;
cudaStreamCreate(&str); // Create a stream
```

Data transfer using a specific stream

```
cudaMemcpyAsync(dst, src, size, type, str);
```

Call GPU kernel function using a stream

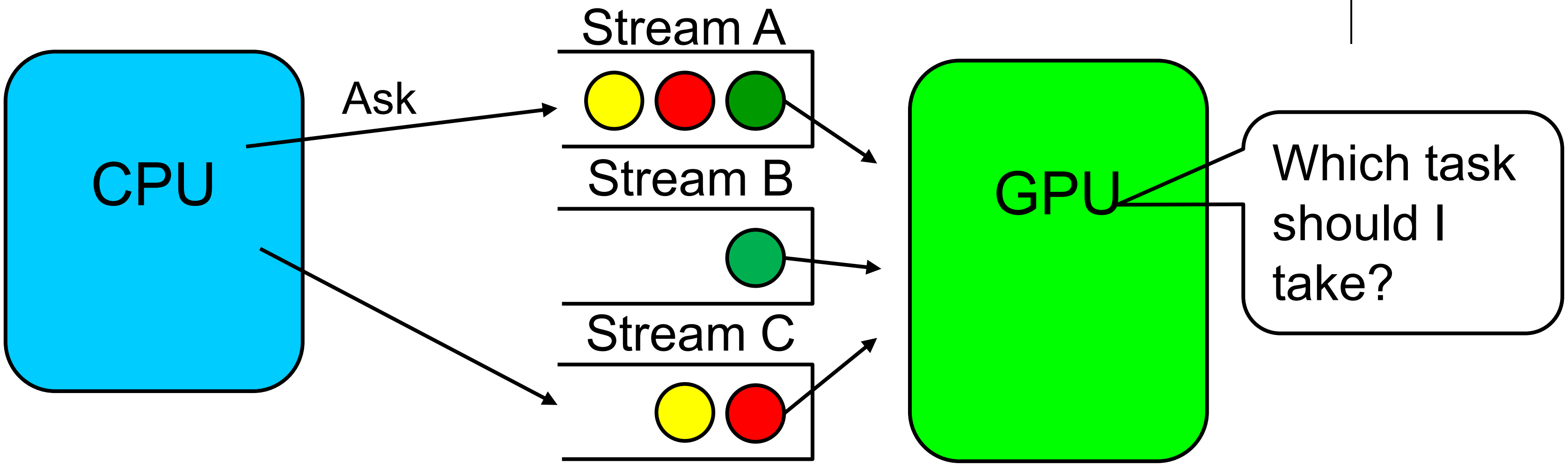
```
func<<<gs, bs, 0, str>>>( … );
// 3rd parameter is related to for “shared memory”
```

Wait until all tasks on a stream are finished

```
cudaStreamSynchronize(str);
```

※ The default stream is expressed as (cudaStream_t)0

How GPU Executes Tasks with Multiple Streams



- Rule: Tasks on the same stream are done in FIFO
  - The GPU considers that “tasks on one stream have dependencies, so I’ll do them in the order”
- If tasks are in different streams, and have different kinds, they may be done simultaneously
  - Kinds: Host→Device, kernel, Device→Host
  - Note: If tasks are in the same kind, no speed up

mm-str-cuda sample: Overlapping Computation and Transfer

n streams can be used for n independent "task sets"

- C1 🡐 A × B1 (includes H->D, Calc, D->H)
- C2 🡐 A × B2

  ....

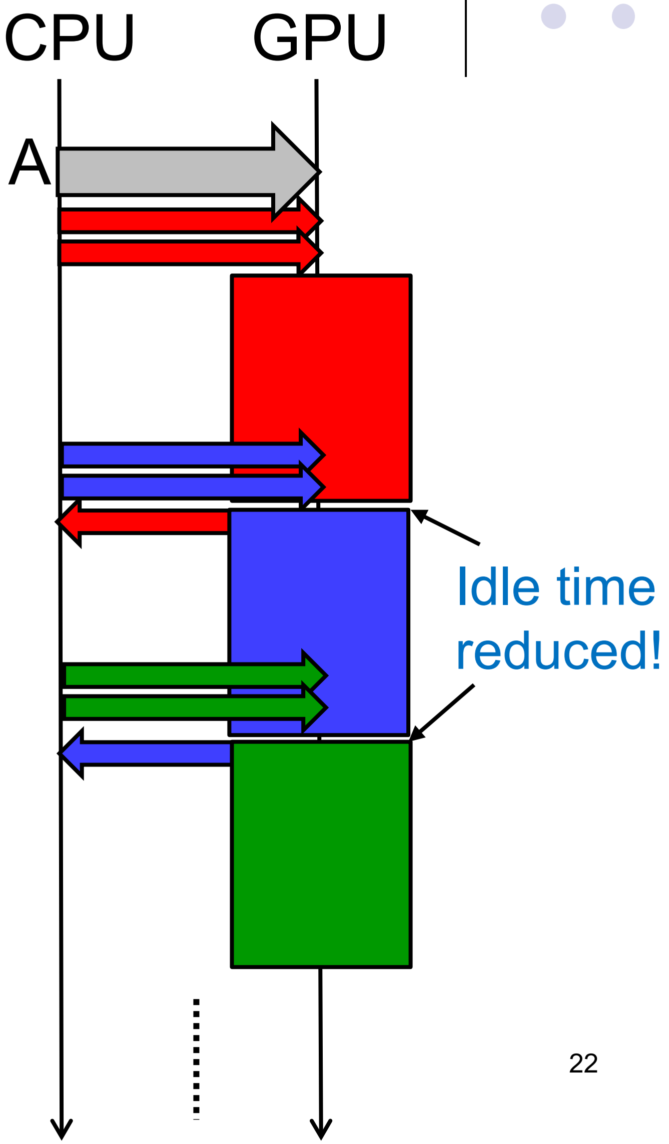
- Cn 🡐 A × Bn

→ We will see speed up since

(Total comp time + Total trans time)

is improved to roughly

max(Total comp time, Total trans time)



Notes on mm-str-cuda

- In allocation of A, B, C on host, cudaHostAlloc() is used instead of malloc()
  - cudaHostAlloc(&A, sizeof(double)*m*k, cudaHostAllocMapped);
  - This allocates “pinned” memory ➔ cudaMemcpyAsync() gets faster
- Currently, mm-str-cuda uses NDIV streams, but there are other ways
  - Using 2 (or 3) streams and 2 (or 3) smaller buffers repeatedly

➔ We can save GPU memory and exceed GPU memory capacity!

Improvement of mm-acc
(Matrix multiplication in OpenACC)

- Sample /gs/bs/tga-ppcomp/24/mm-v2-acc/
  - Loop order: JLI ➔ JIL
  - Access to Cij is reduced
  - J, I loop is configured with “gang, vector” option
    - gang ⬄ thread block, vector ⬄ thread
- The idea of overlapping is also applicable to OpenACC
  - Sample /gs/bs/tga-ppcomp/24/mm-str-v2-acc/
  - ... async option is used

More Things to Study

- Using CUDA shared memory
  - fast and small memory than device memory
- Unified memory in recent CUDA
  - cudaMemcpy can be omitted for automatic data transfer
  - Google with “cudaMallocManaged”
- Using specialized hardware to accelerate deep learning
  - Tensor core, Transformer engine...
- Using multiple GPUs towards petascale computation
  - MPI+CUDA, MPI+OpenACC
- More and more…

Assignments in GPU Part (Abstract)

Choose one of [G1]—[G3], and submit a report

Due date: May 30 (Thursday)

[G1] Parallelize “diffusion” sample program by OpenACC or CUDA

[G2] Evaluate speed of “mm-acc” or “mm-cuda” in detail

※ In OpenACC case, mm-meas-acc sample is useful

[G3] (Freestyle) Parallelize *any* program by OpenACC or CUDA.

For more detail, please see ppcomp-2-1 slides

Next Class:

- Part 3: MPI Programming (1)
  - Introduction to distributed memory parallel programming