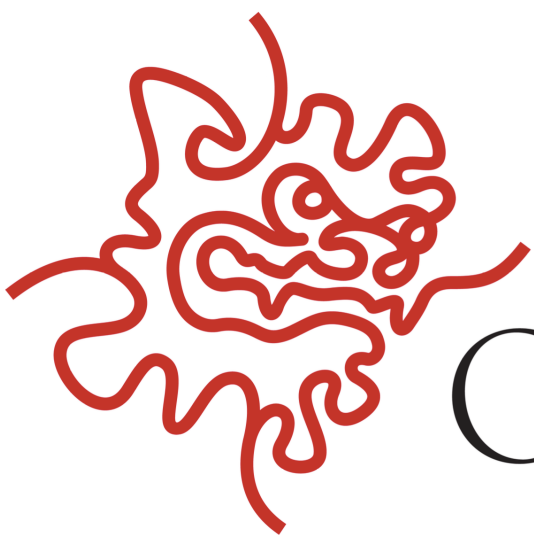
GPU computing with Julia

James Schloss



OIST

August 13, 2021

Overview

Overview

- What is a GPU?

Overview

- What is a GPU?
- What is Julia?

Overview

- What is a GPU?
- What is Julia?
- What is CUDA?

Overview

OIST

- What is a GPU?
- What is Julia?
- What is CUDA?
- What is Kernel Abstractions?

Overview

OIST

- What is a GPU?
- What is Julia?
- What is CUDA?
- What is Kernel Abstractions?
- What is a more complicated example?

Overview

OIST

- What is a GPU?
- What is Julia?
- What is CUDA?
- What is Kernel Abstractions?
- What is a more complicated example?
- What is the meaning of life?

What is a GPU?

What is a GPU?

This thing

Yeah, but what does it do?

Yeah, but what does it do?

Graphics

Yeah, but what does it do?

Graphics

(In parallel, though)

Ok, but why should I care?

Ok, but why should I care?

Dunno, dawg. You are the one that decided to attend this mini-course

Ok, but why should I care?

OIST

If used correctly, GPUs can be much faster for certain tasks

Ok, but why should I care?

OIST

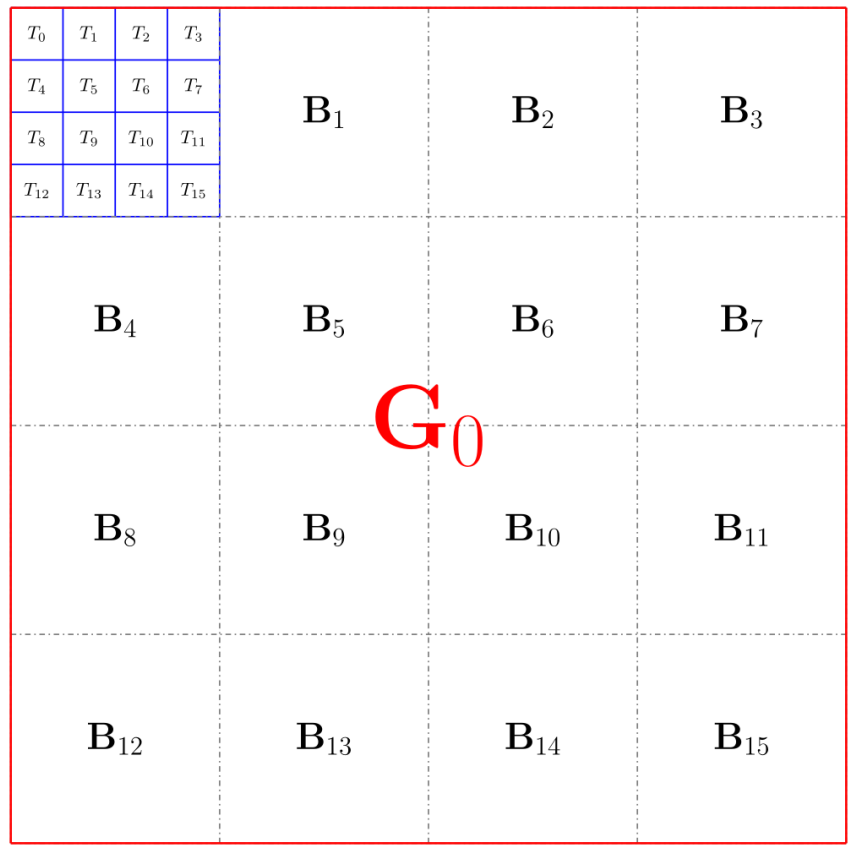
If used correctly, GPUs can be much faster for certain tasks

(But they can also be way slower for other tasks)

Right. How do they work?

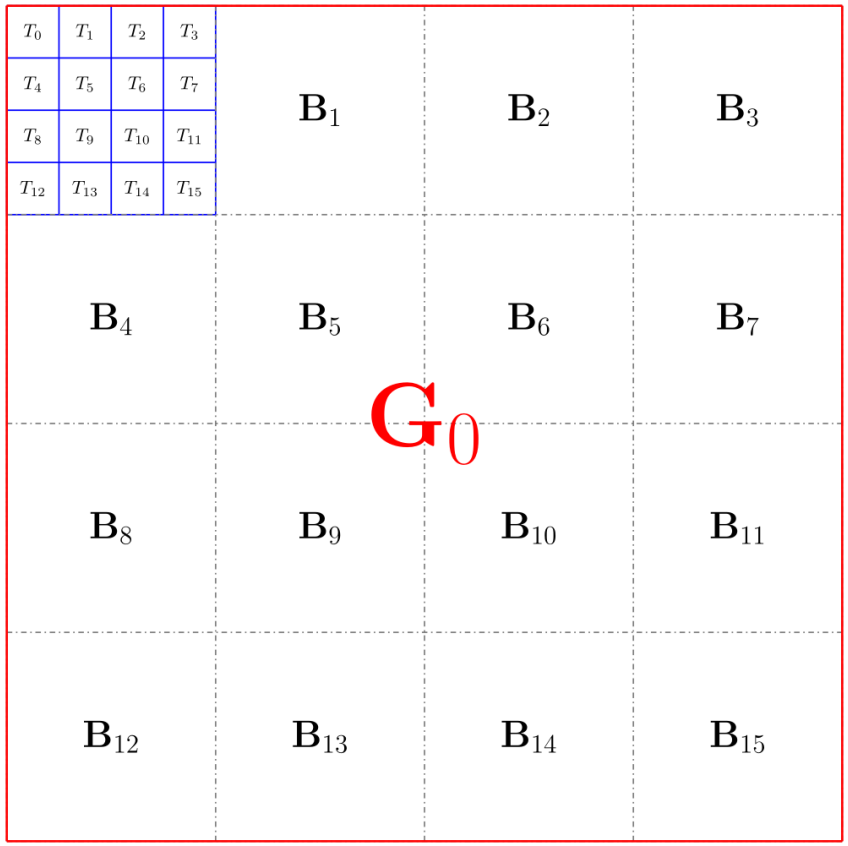
Right. How do they work?

- The Grid is split into Blocks
- Blocks are split into Threads



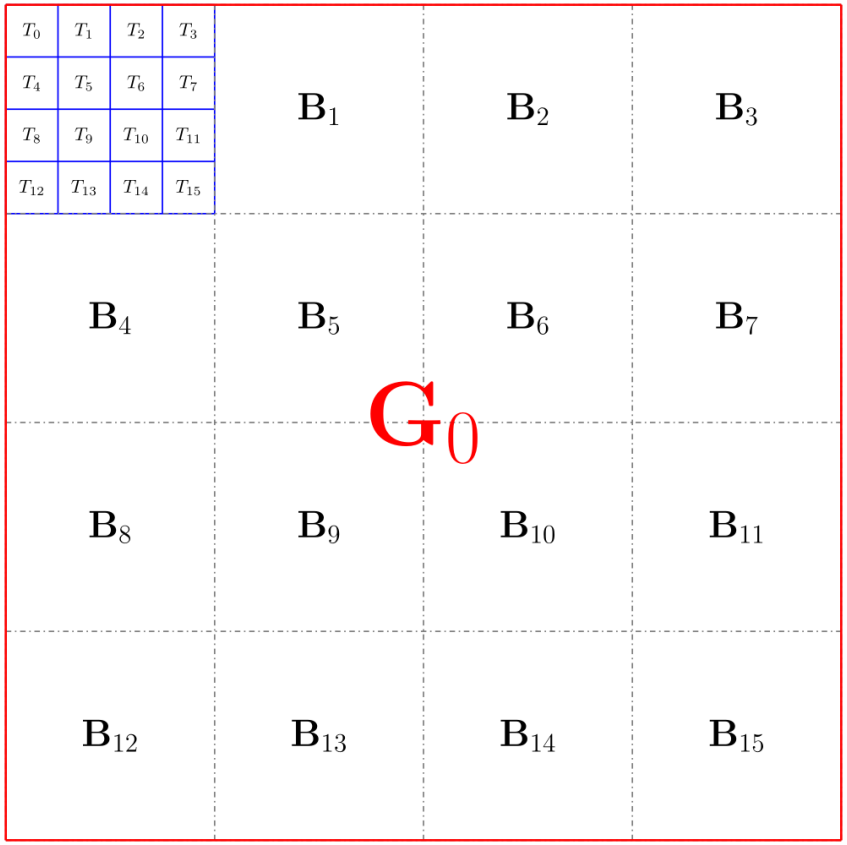
Right. How do they work?

- The Grid is split into Blocks
- Blocks are split into Threads
- Threads have local memory
- Everything has global memory



Right. How do they work?

- The Grid is split into Blocks
- Blocks are split into Threads
- Threads have local memory
- Everything has global memory
- Shared mem is faster than global memory, but a pain to use



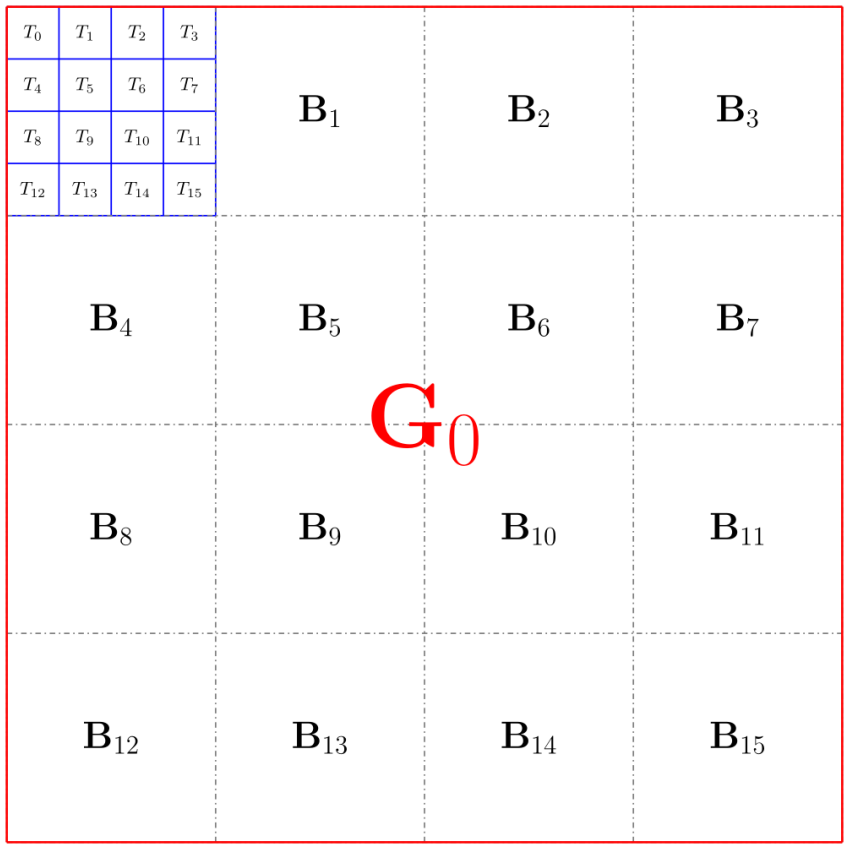
What are they bad at?

Ah, like everything

What are they bad at?

Ah, like everything

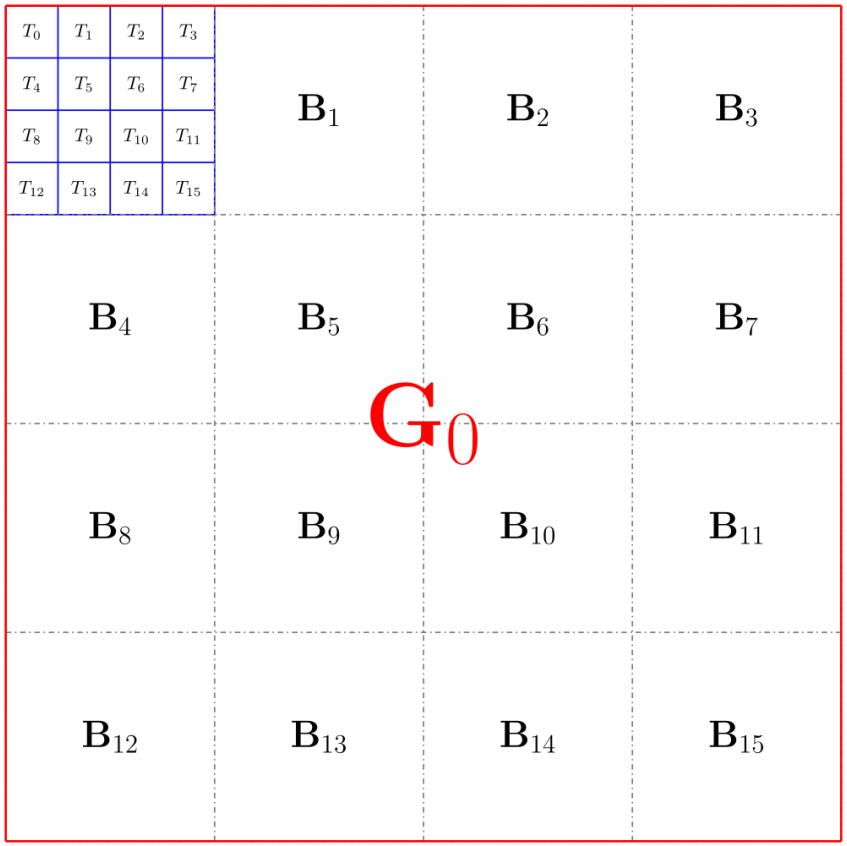
- Iterative things are bad because threads are weak



What are they bad at?

Ah, like everything

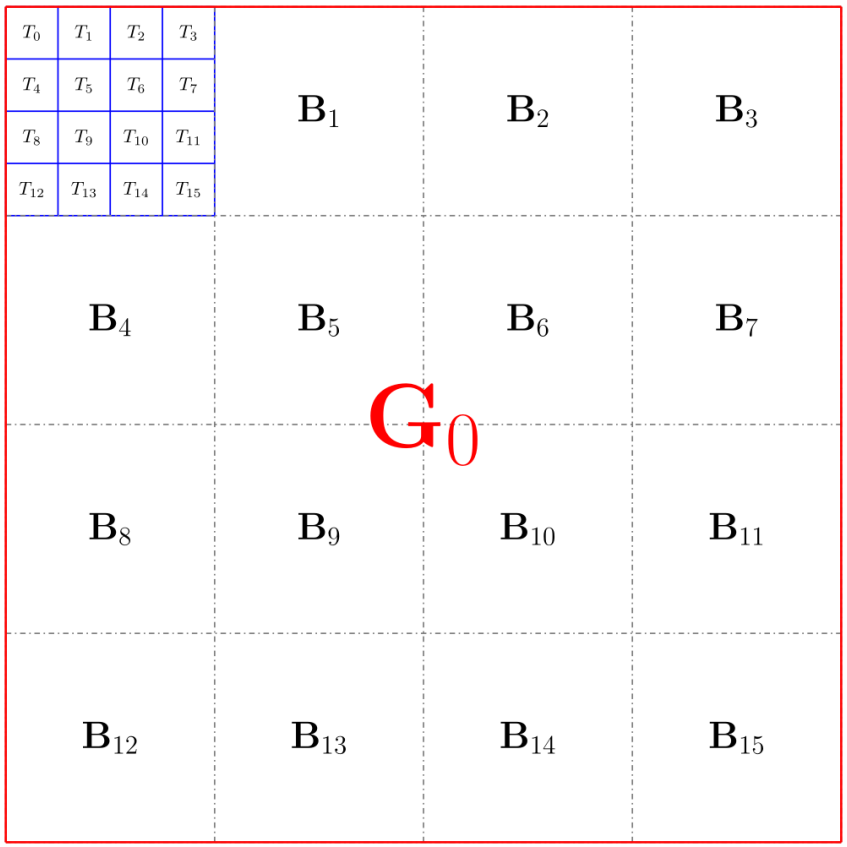
- Iterative things are bad because threads are weak
- Really big problems (> 32 GB) are bad because we have limited memory



What are they bad at?

Ah, like everything

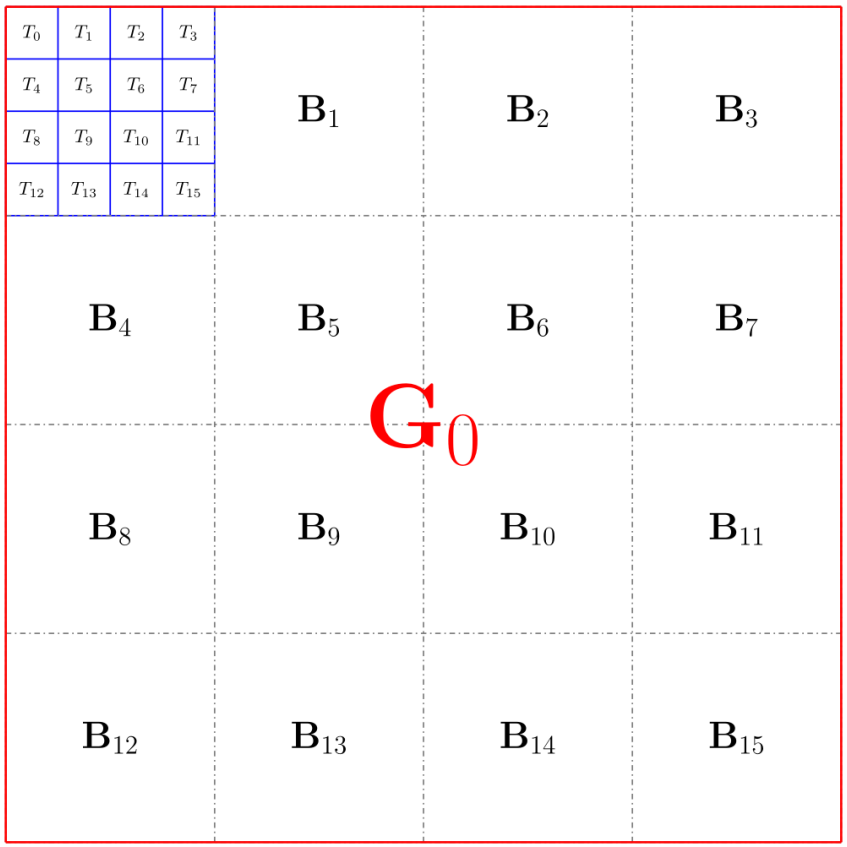
- Iterative things are bad because threads are weak
- Really big problems (> 32 GB) are bad because we have limited memory
- Recursive things are bad because threads are weak and we have limited memory



What are they bad at?

Ah, like everything

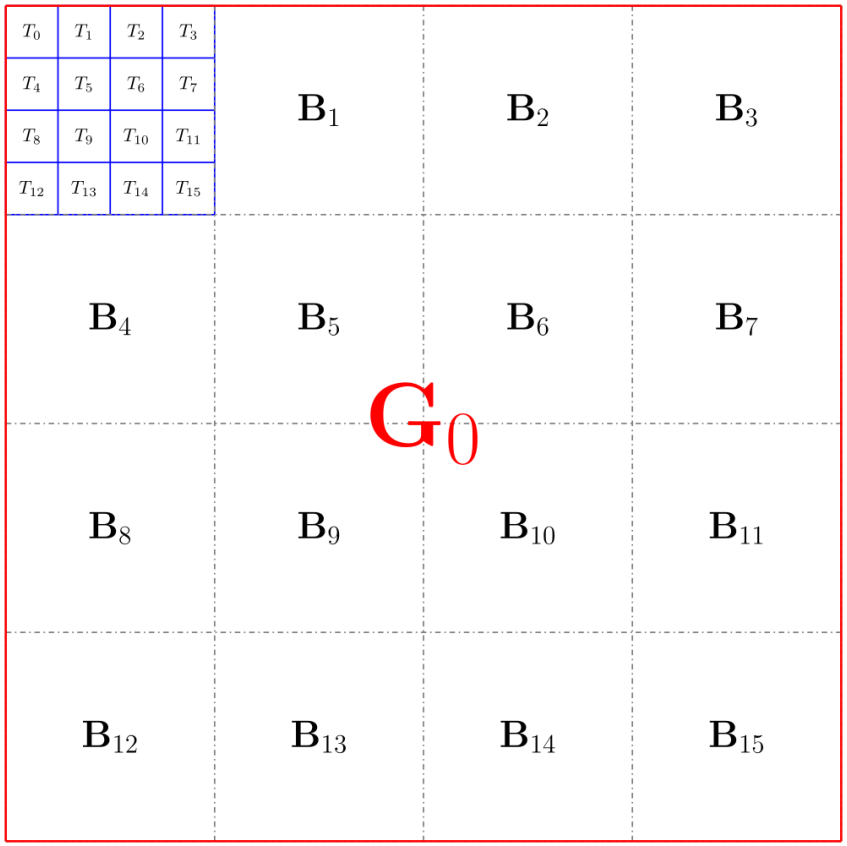
- Iterative things are bad because threads are weak
- Really big problems ($>$ 32 GB) are bad because we have limited memory
- Recursive things are bad because threads are weak and we have limited memory
- Good luck getting data off the GPU and onto your SSD with any reasonable speed



What are they bad at?

Ah, like everything

- Iterative things are bad because threads are weak
- Really big problems (> 32 GB) are bad because we have limited memory
- Recursive things are bad because threads are weak and we have limited memory
- Good luck getting data off the GPU and onto your SSD with any reasonable speed
- Please don't do multi-GPU. Just... Please.



So, uh...

OIST

But, I mean, they can do matrix operations quickly, so there's that.

but...

Pop-quiz: What's the slowest part of any n-dimensional FFT operation?

but...

Pop-quiz: What's the slowest part of any n-dimensional FFT operation?

The transpose

What is Julia?

OIST

Live demo (hopefully)

What is CUDA

Again, live demo

What is CUDA.jl

OIST

We are doing it live!

What is Kernel Abstractions

OIST

Why did you even make these slides?