NVIDIA PAX-HPC Workshop
17th January 2024

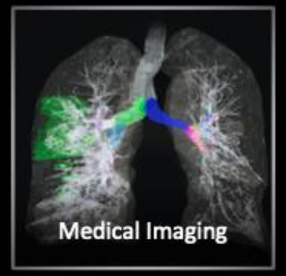
Paul Graham | Senior Solutions Architect

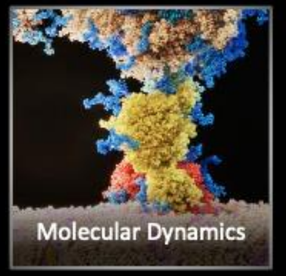
pgraham@nvidia.com

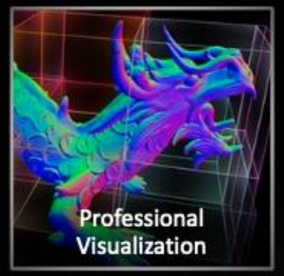
Agenda

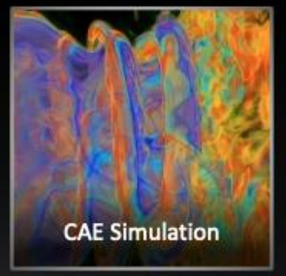
- 10:15-1045 Introduction to Heterogeneous Parallel Computing
- 10:45-11:00 Break
- 11:00-12:00 Key ways to accelerate applications
- 12:00-13:00 Programming for GPUs
- 13:00-13:45 Lunch
- 13:45-17:15 Hands-on practical

Introduction to Heterogeneous Parallel Computing



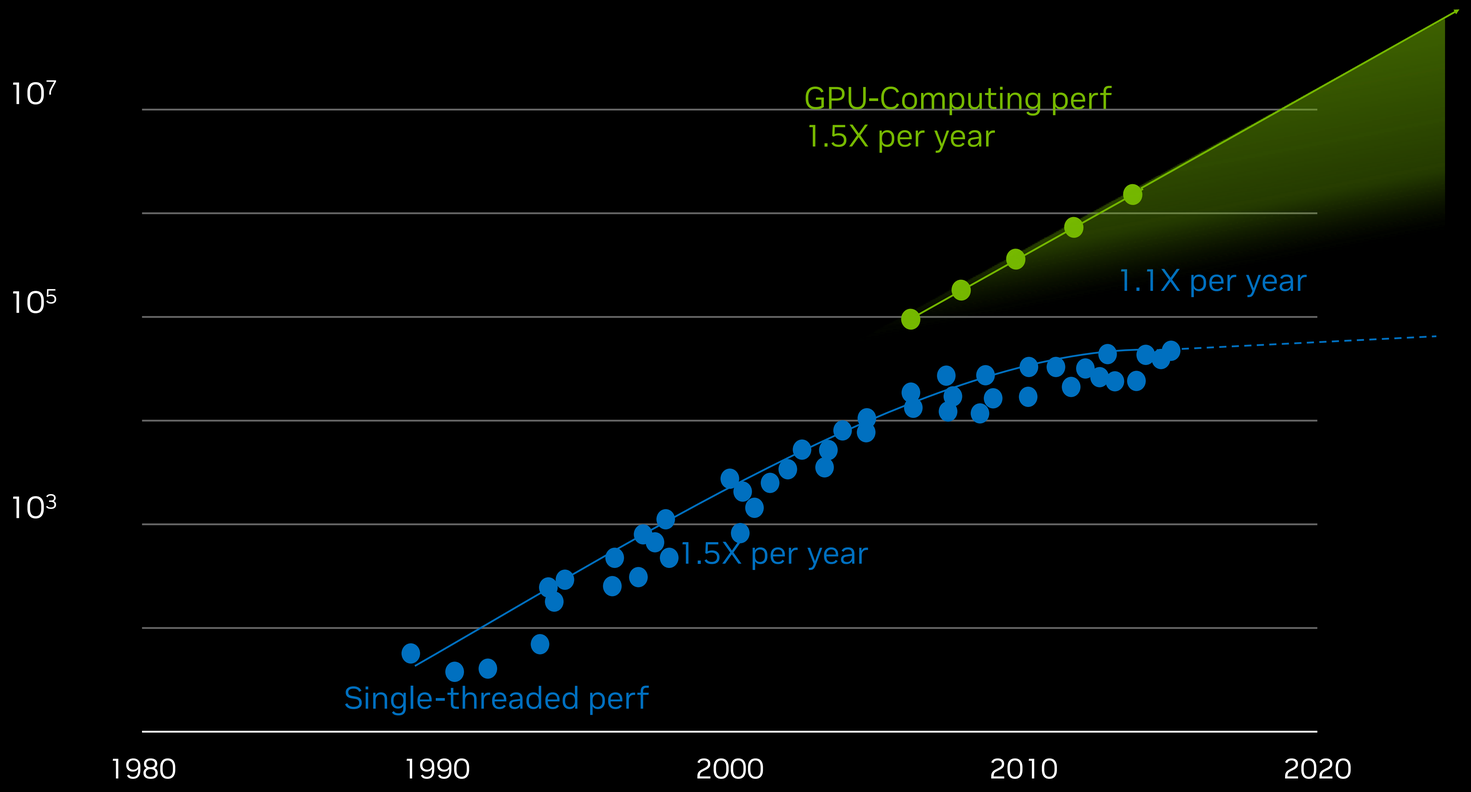




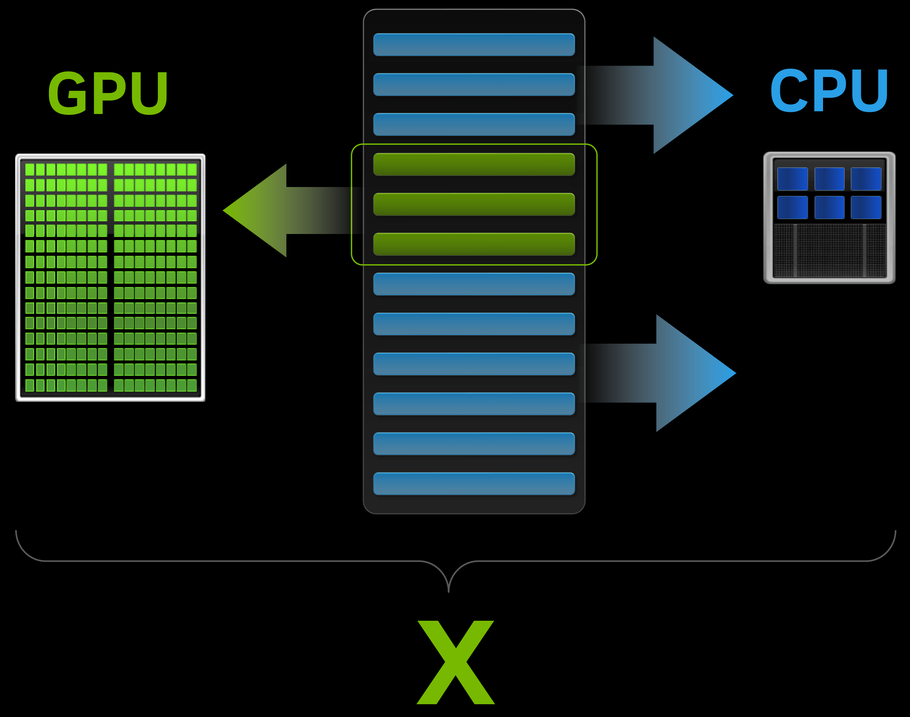


Universe of GPU Computing

Rise of GPU Computing



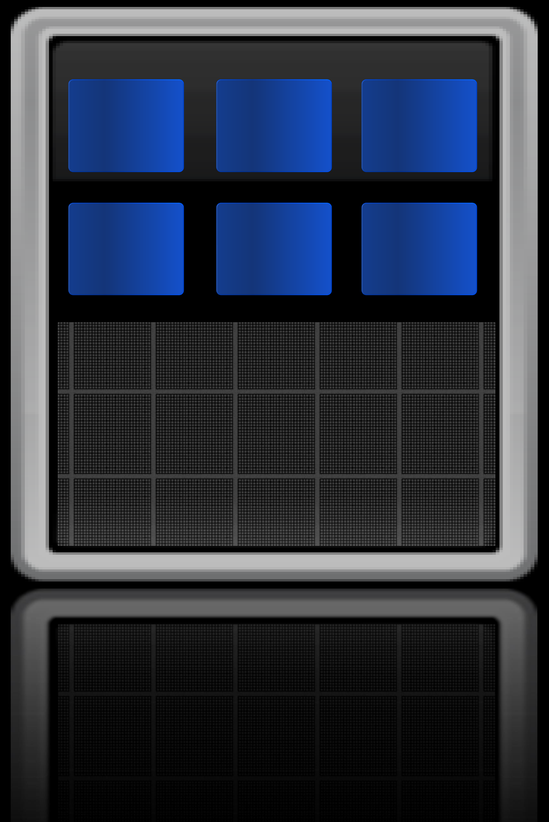
40 Years of CPU Trend Data



Original data up to the year 2010 collected and plotted by M. Horowitz,
F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

CPU is a Latency Reducing Architecture

CPU
Optimized for
Serial Tasks

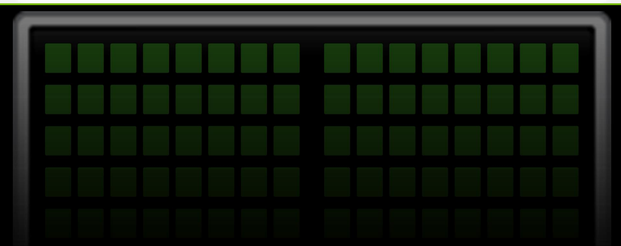


CPU Strengths

- Very large main memory
- Very fast clock speeds
- Latency optimized via large caches
- Small number of threads can run very quickly

CPU Weaknesses

- Relatively low memory bandwidth
- Cache misses very costly
- Low performance/watt



GPU is All About Hiding Latency

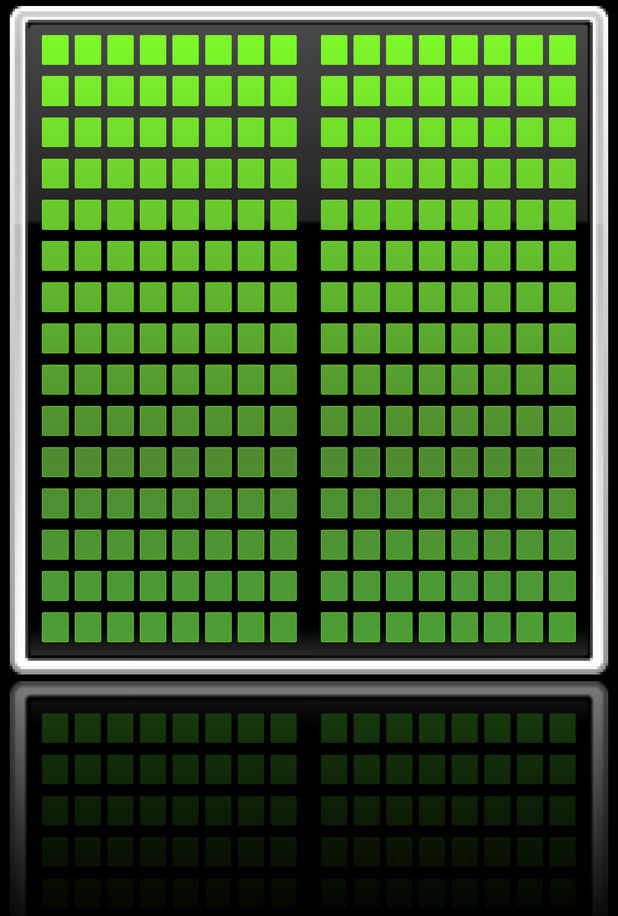
GPU Strengths

- High bandwidth main memory
- Significantly more compute resources
- Latency tolerant via parallelism
- High throughput
- High performance/watt

GPU Weaknesses

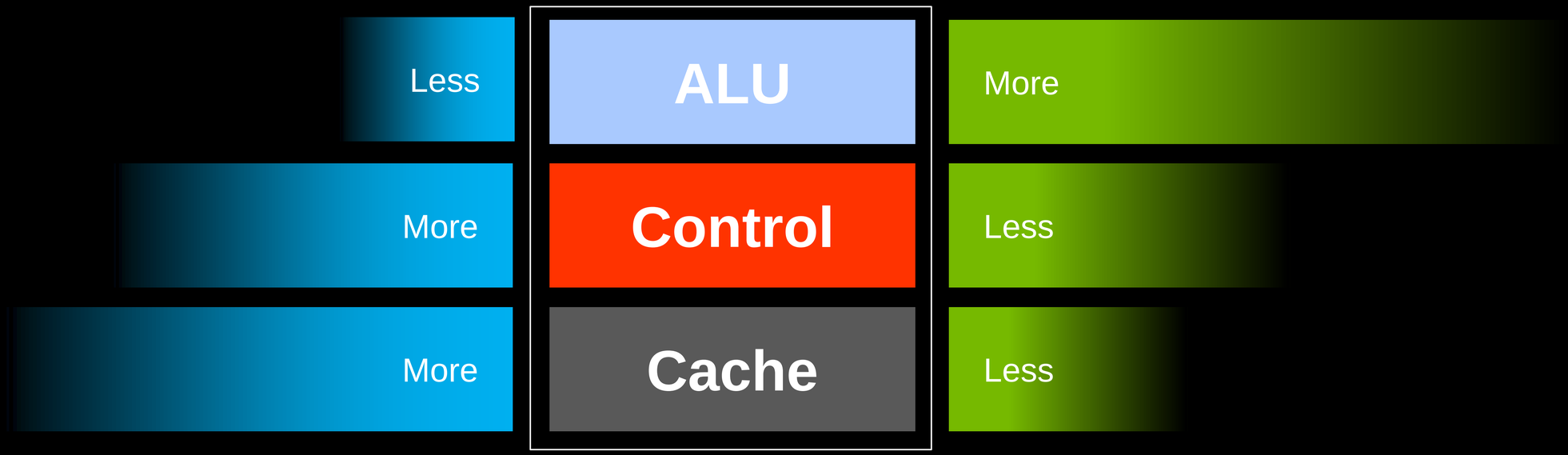
- Relatively low memory capacity
- Low per-thread performance

GPU Accelerator
Optimized for
Parallel Tasks



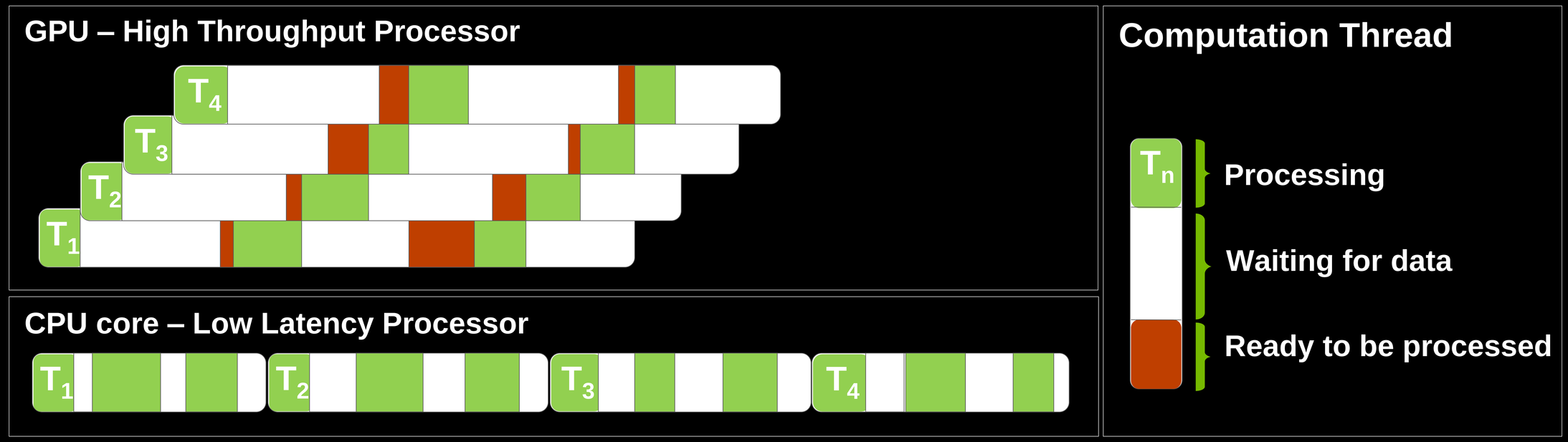
SILICON BUDGET

The three components of any processor



LOW LATENCY VS HIGH THROUGHPUT

- CPU architecture must minimize latency within each thread
- GPU architecture hides latency with computation (data-parallelism, to thousands of threads!)



SPEED VS THROUGHPUT

Speed

Throughput

Which is better depends on your needs...

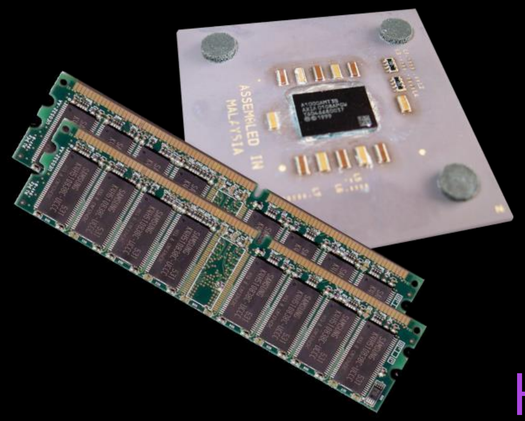
SPEED VS THROUGHPUT

Speed

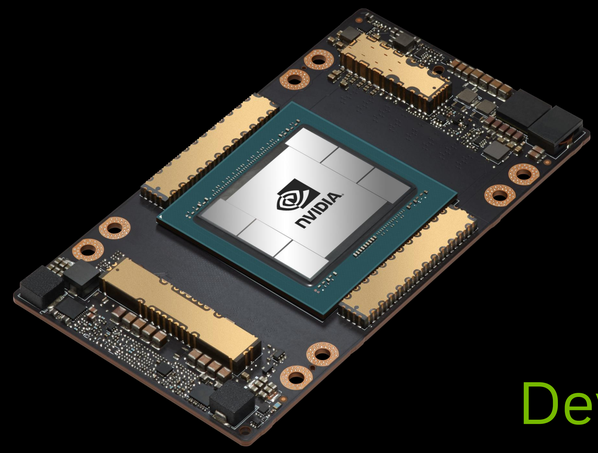
Which is better depends

Heterogeneous Computing

- Terminology:
  - *Host* The CPU and its memory (host memory)
  - *Device* The GPU and its memory (device memory)

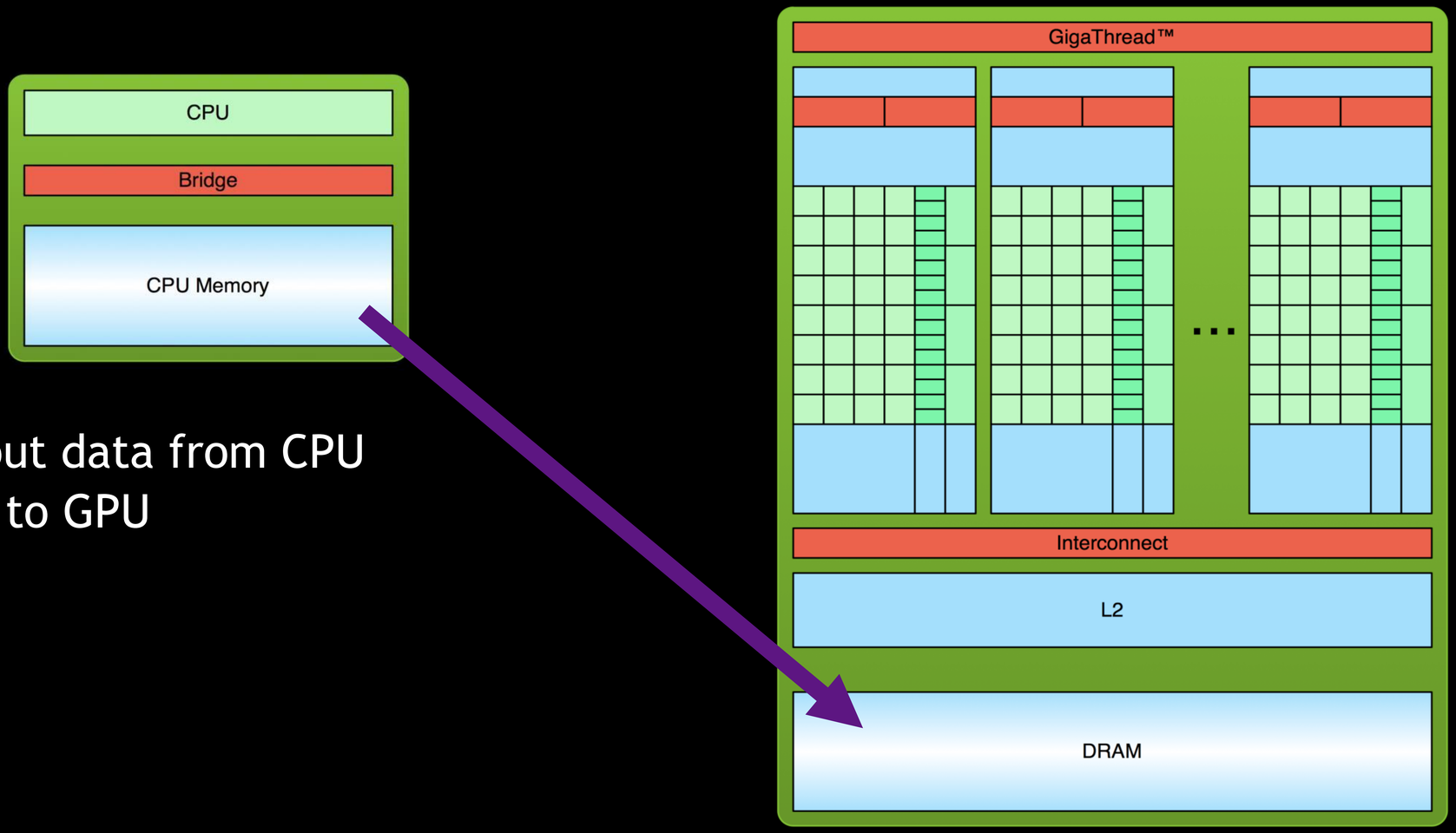


Host



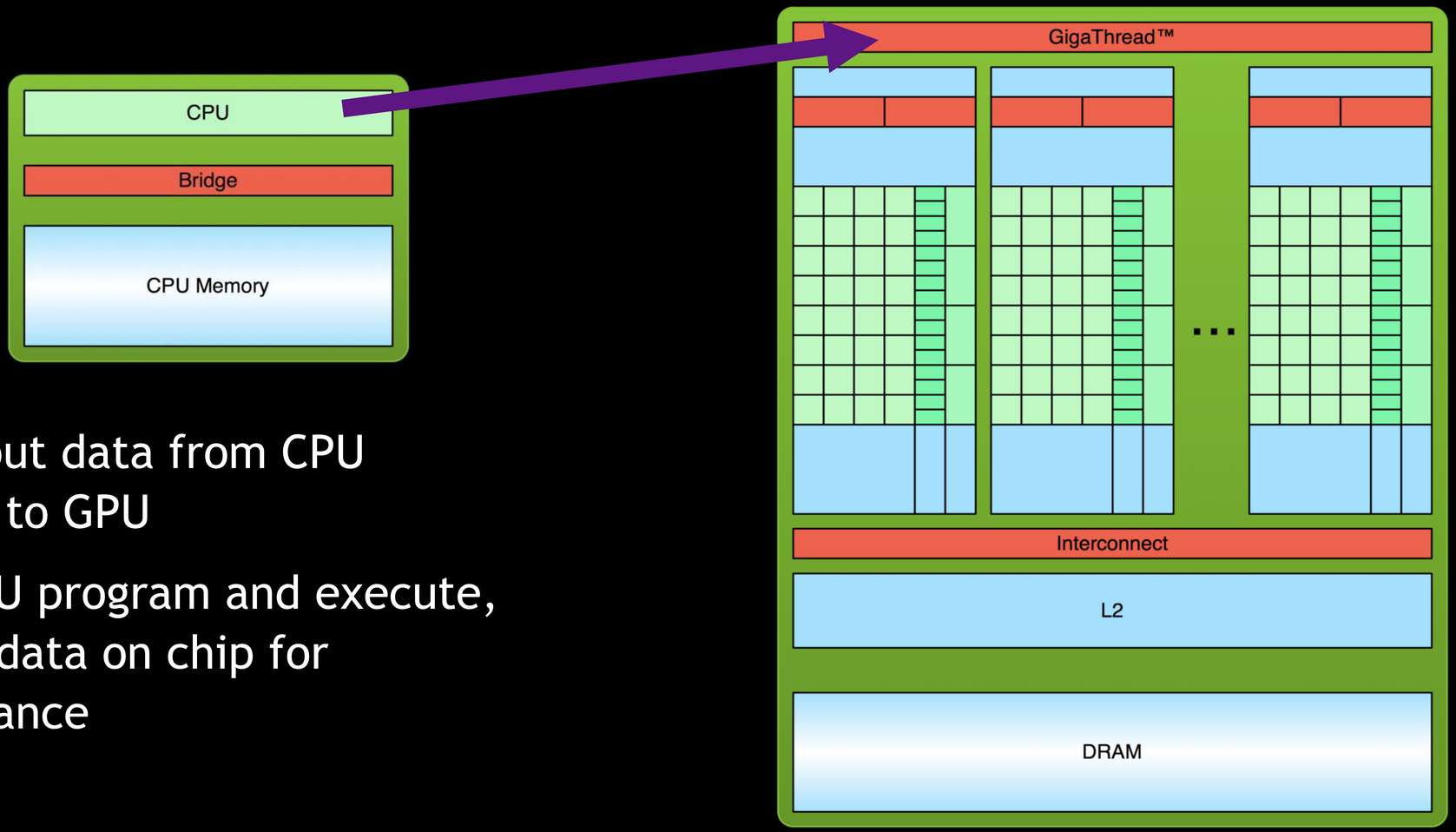
Device

SIMPLE PROCESSING FLOW



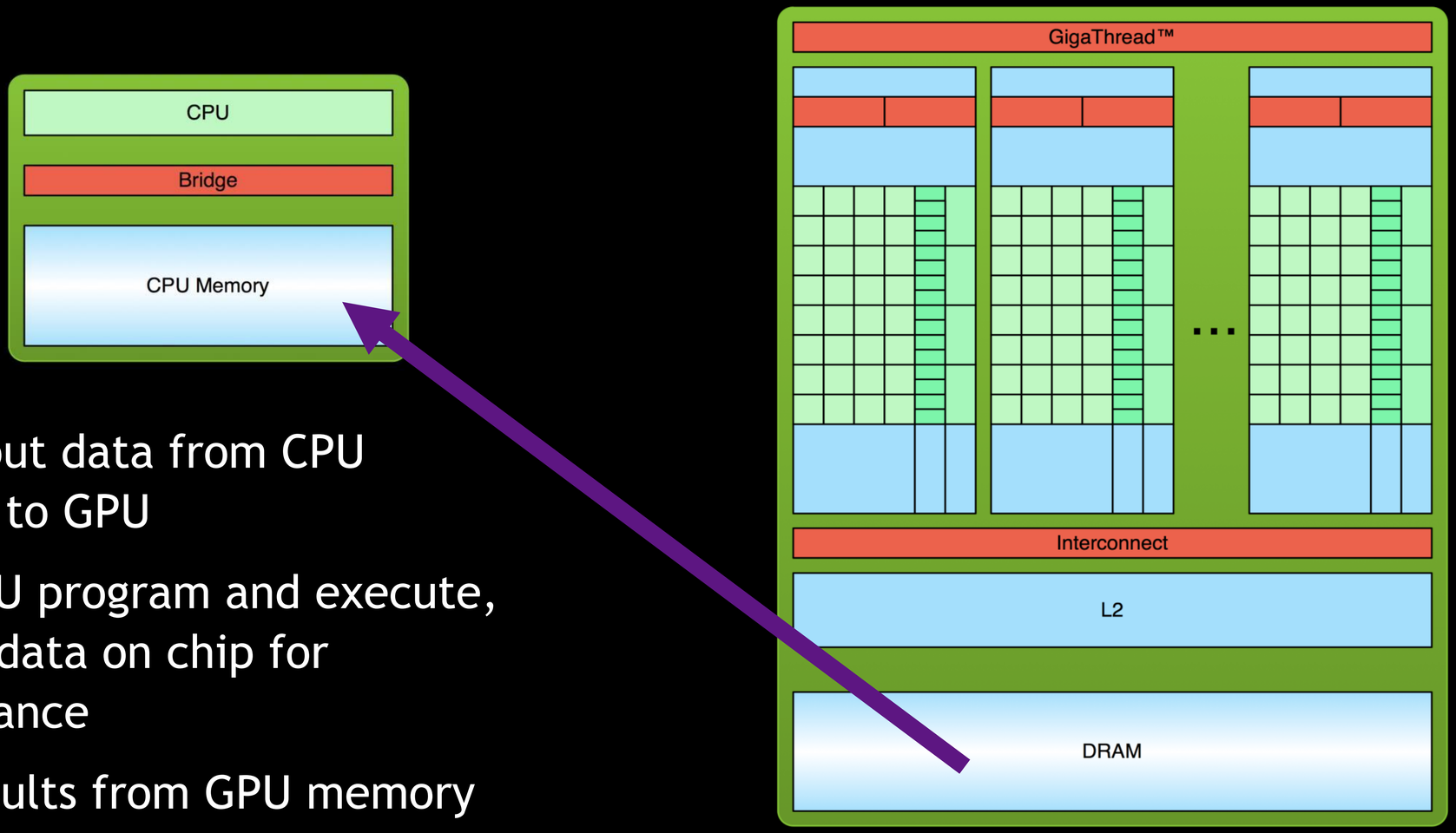
1. Copy input data from CPU memory to GPU

SIMPLE PROCESSING FLOW



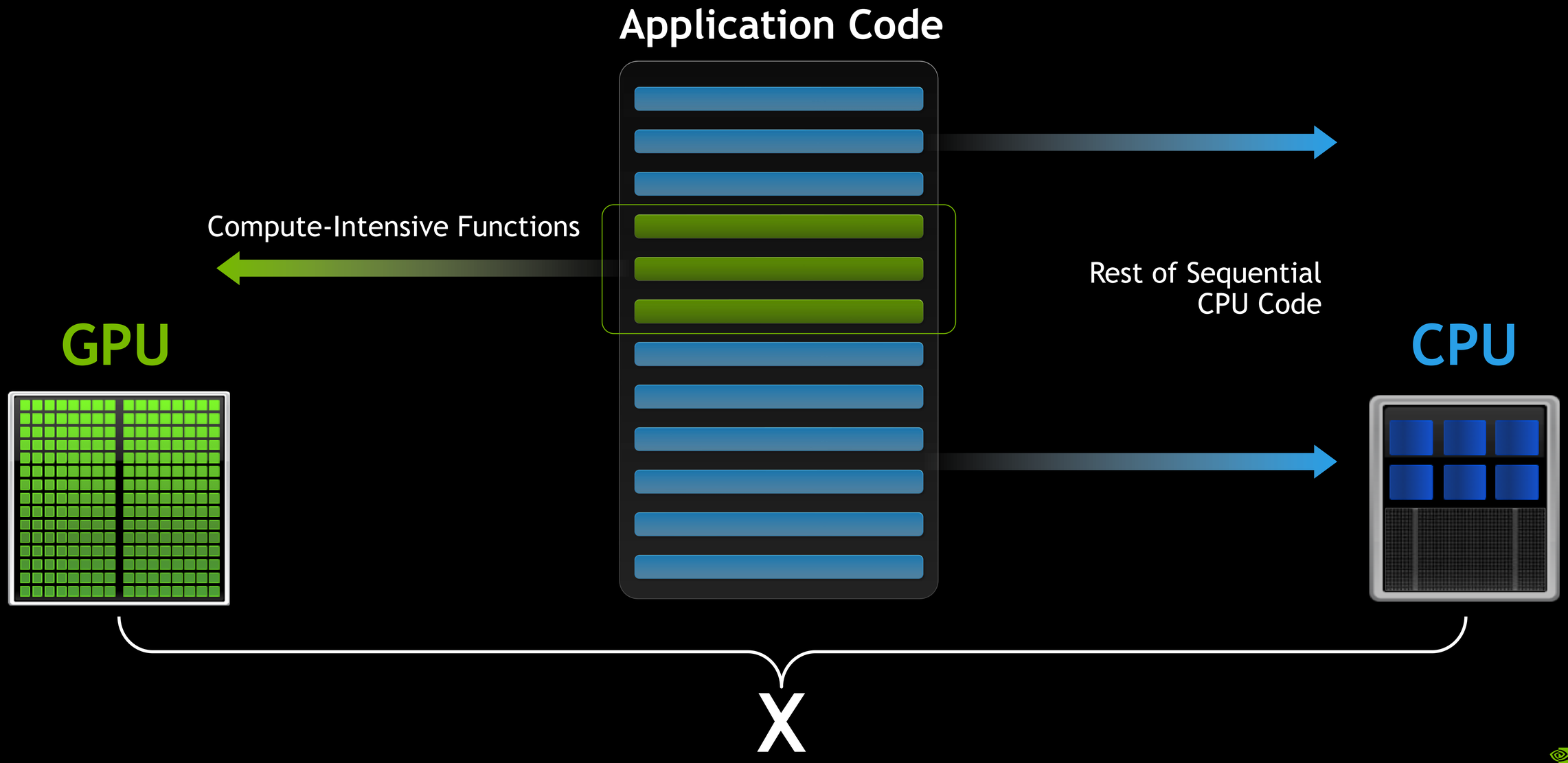
1. Copy input data from CPU memory to GPU
2. Load GPU program and execute, caching data on chip for performance

SIMPLE PROCESSING FLOW



1. Copy input data from CPU memory to GPU
2. Load GPU program and execute, caching data on chip for performance
3. Copy results from GPU memory back to CPU memory

Small Changes, Big Speed-up



Hardware

GH100 GPU architecture

https://resources.nvidia.com/en-us-tensor-core/gtc22-whitepaper-hopper

GH100 GPU architecture

https://resources.nvidia.com/en-us-tensor-core/gtc22-whitepaper-hopper

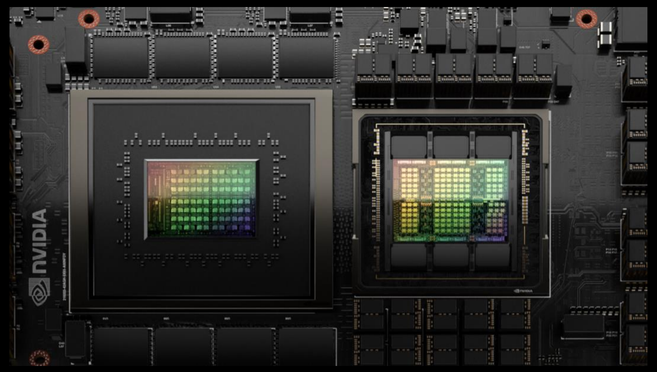
ecture

ht e/gtc22-whitepaper-hopper

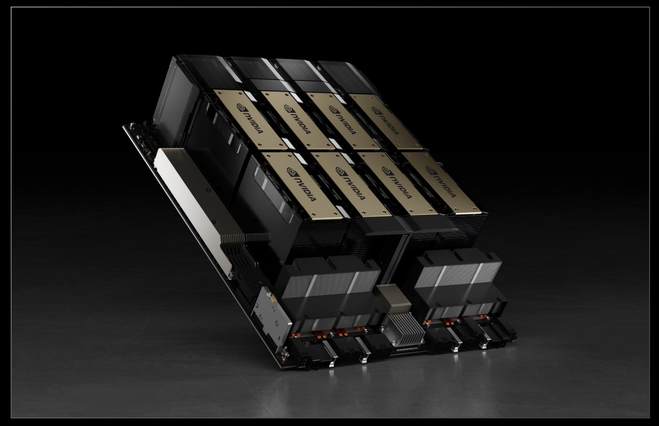
ecture

ht e/gtc22-whitepaper-hopper

- 132 of these in the H100 SXM5 GPU
- > 16000 FP32 cores in total
- We usually want number of threads >> num cores
- So we need **a lot** of threads!



Multi-die



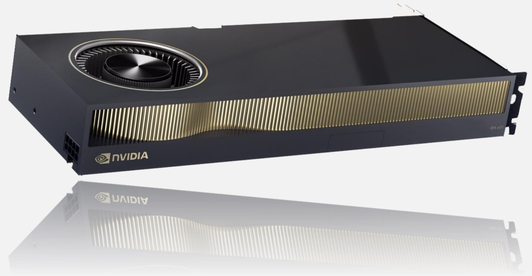
Multi-chip

Multi-node

NVIDIA RTX 6000 Ada Generation

Ada for the Enterprise Desktop

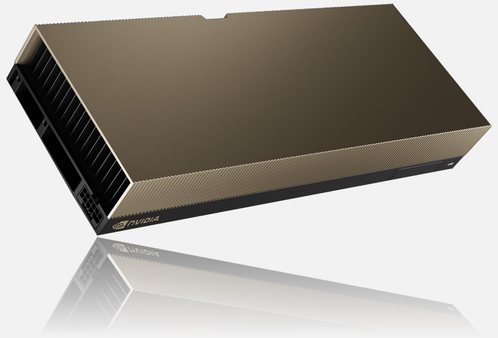
- Up to 2x Faster than RTX A6000
- 48GB GDDR6 w/ECC Memory
- 4x DP1.4 Display Outputs
- PCIe Gen 4
- vGPU Support
- Available from channel partners starting in December 2022, OEM partners early 2023



NVIDIA L40/L40S

Ada for the Data Center

- Up to 2x Faster than A40
- 48GB GDDR6 w/ECC Memory
- 4x DP1.4 Display Outputs
- PCIe Gen 4
- vGPU Support
- Availability starting in December 2022



Ramen Shop Demo

Blog post on making the demo

*Specifications, availability dates subject to change. vGPU support in Q1 2023

Announcing Ada Lovelace Architecture

Next Generation RTX for Enterprise

Agenda

- 10:15-1045 Introduction to Heterogeneous Parallel Computing
- 10:45-11:00 Break
- 11:00-12:00 Key ways to accelerate applications
- 12:00-13:00 Programming for GPUs
- 13:00-13:45 Lunch
- 13:45-17:15 Hands-on practical