CUDA Docs for Humans

modal.com/gpu-glossary

What is this?
Where did it come from?
What does it say?
Where is it going?

I started off at Cal, studying DNN optimization.

Finding Critical and Gradient-Flat Points of Deep Neural Network Loss Functions

by

Charles Gearhart Frye

A dissertation submitted in partial satisfaction of the

requirements for the degree of

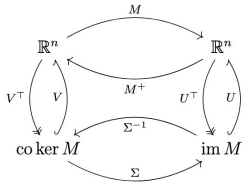
Doctor of Philosophy

Theorem 3.3: Kernel Equals Pseudo-Inverse Kernel for Symmetric M

Let $M \in \mathbb{R}^{n \times n}$ be a symmetric matrix. Then

$$\ker M = \ker M^{+} \tag{3.31}$$

Proof of Theorem 3.3:
We first repeat the commutative diagram relating the SVDs of a matrix and its pseudo-inverse, specialized to a square matrix.



https://charlesfrye.github.io/pdfs/thesis.pdf

Modal

Empirical work done on CPUs with `autograd.numpy` 👴

At W&B, started helping people operationalize research.

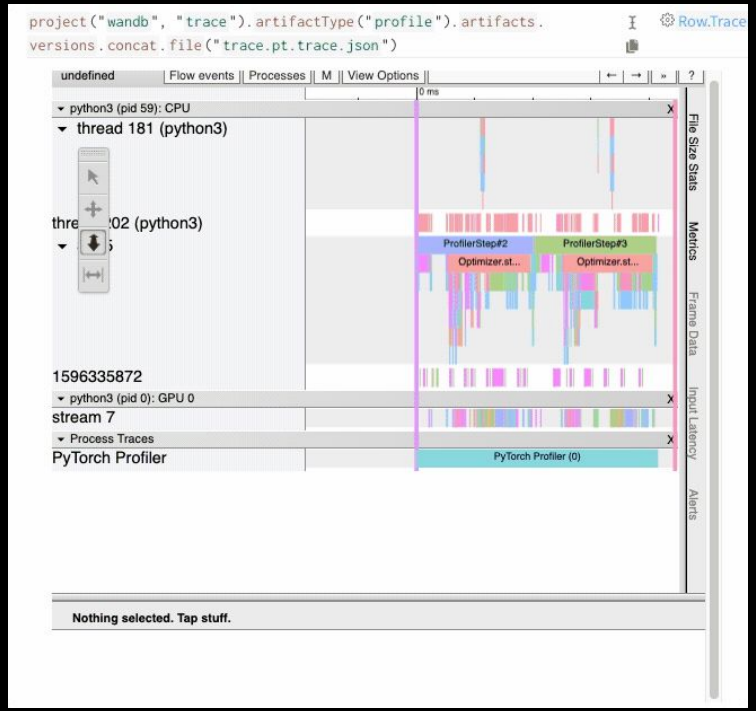
Public Dissection of a PyTorch Training Step

What really happens when you call .forward, .backward, and .step?

Charles Frye Share 9 comments 27 stars

Created on August 2 | Last edited on January 4

Rembrandt, 1632. *The Anatomy Lesson of Dr. Nicolaes Tulp.* wiki



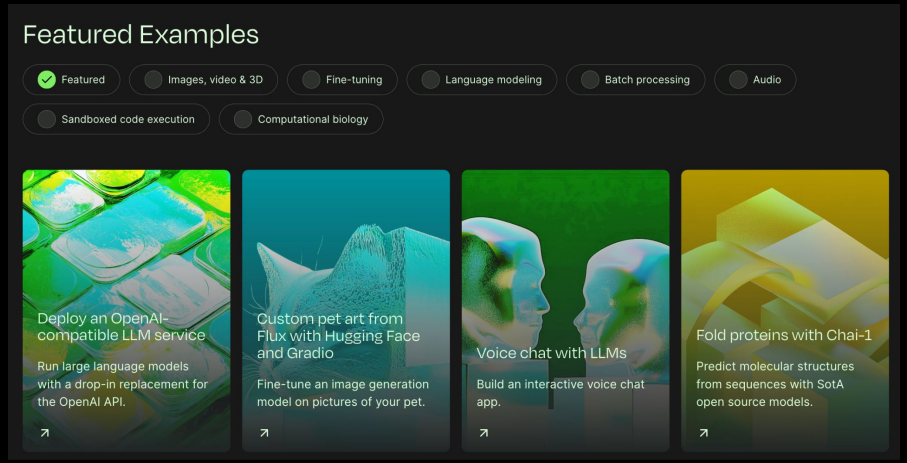
Modal

https://wandb.me/trace-report

Now, at Modal, I'm helping people with deployment!

Modal Blog

All Posts | Engineering | Customer Stories | Tutorials | News



https://modal.com/docs/examples

August 5, 2024

Beat GPT-4o at Python by searching with 100 dumb LLaMAs

Scale up smaller open models with search and evaluation to match frontier capabilities.

https://modal.com/blog/llama-human-eval

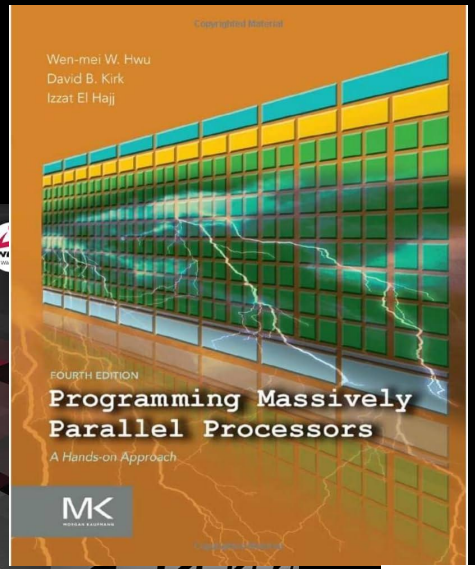
That involved a lot of environment debugging...

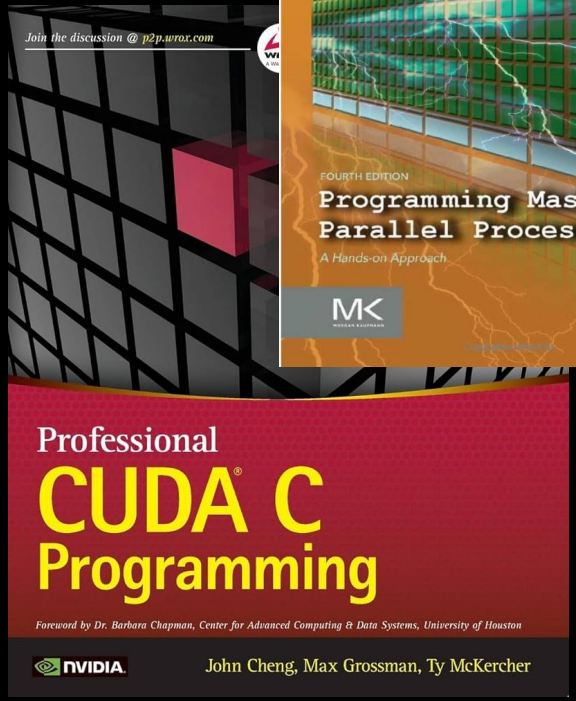
I Am Fucking Done Not Understanding The CUDA Stack

> "The CUDA development environment relies on **tight integration** with the host development environment, including the host compiler and C runtime libraries"
> — sauce, from the horse's mouth

It is unfortunately not possible to develop bleeding-edge applications of GPUs without understanding more about the underlying stack than most would like.

So let's dive in and understand what the layers of that stack are, step-by-step.





CUDA®

CUDA C++ Programming Guide

Release 12.6

NVIDIA Corporation

CUDA®

PTX ISA

Release 8.5

NVIDIA Corporation

CUDA®

NVIDIA CUDA Compiler Driver

Release 12.6

NVIDIA Corporation

RTFM.

Modal

What is this?
Where did it come from?
What does it say?
Where is it going?

There is not one “CUDA”.

/device-hardware

Device Hardware

These terms and technologies are physical
NVIDIA's lingo.

→ CUDA (Device Architecture)

/host-software

Host Software

These terms and technologies are used on the CPU
running GPU programs.

→ CUDA (Software Platform)

→ CUDA C++ (programming language)

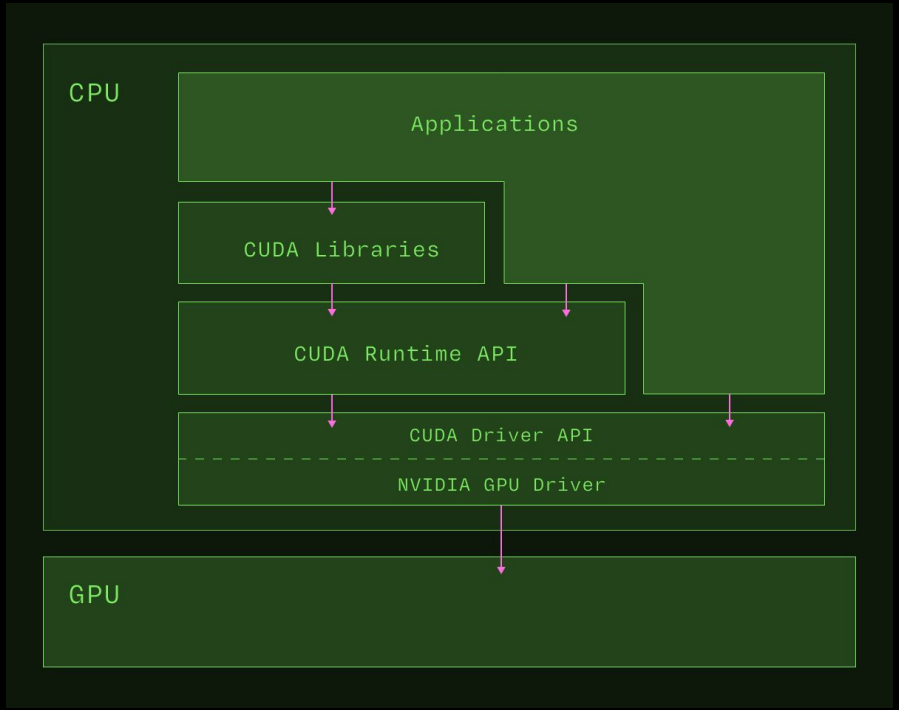
/device-software

Device Software

These terms and technologies are used
NVIDIA's lingo.

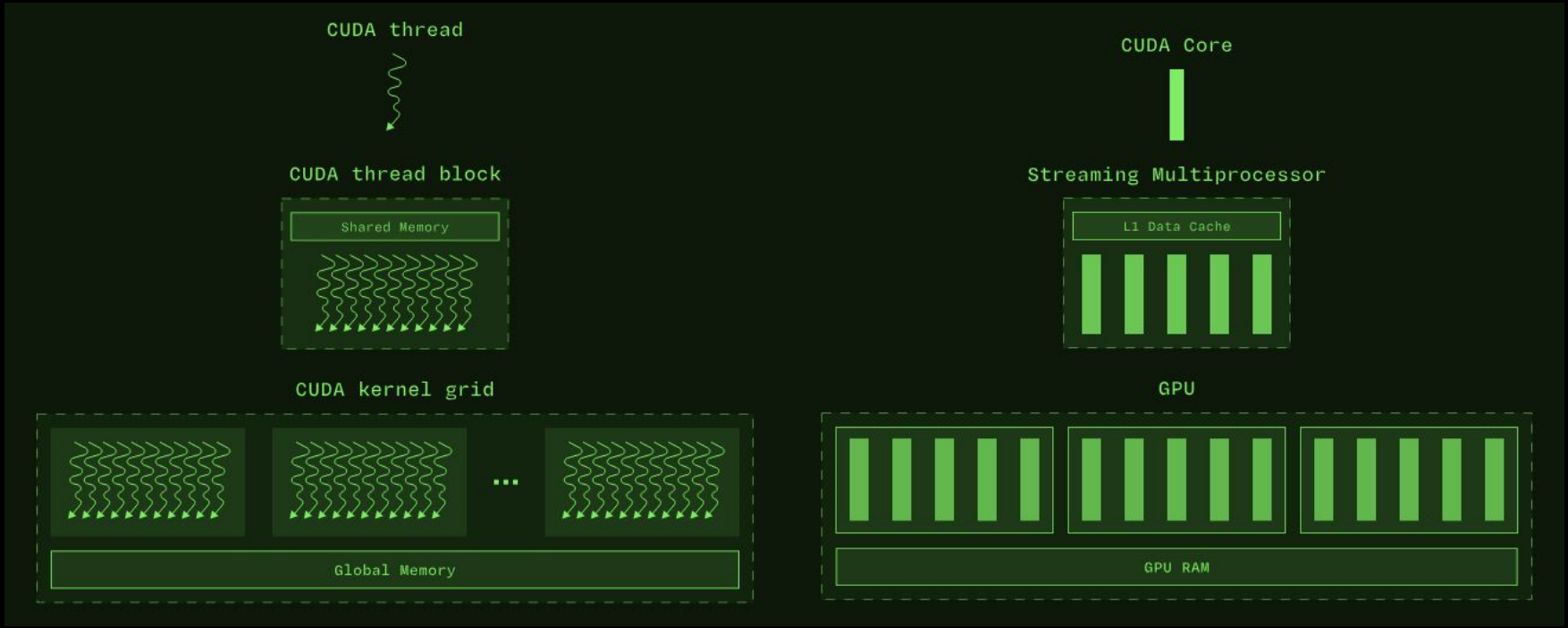
→ CUDA (Programming Model)

CUDA is a software platform.



https://modal.com/gpu-glossary/host-software/cuda-software-platform

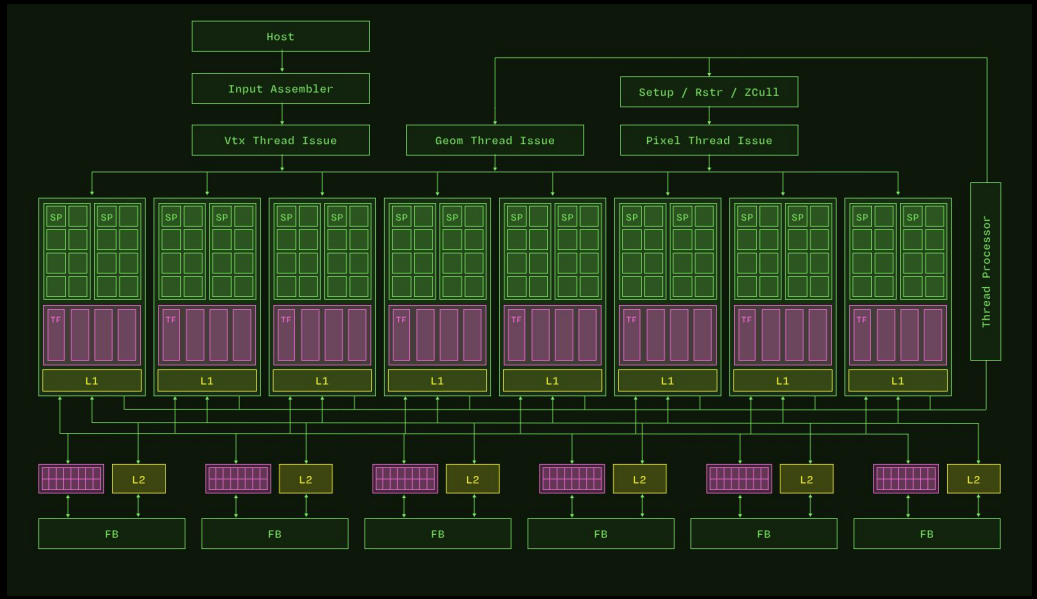
CUDA is a programming model.



https://modal.com/gpu-glossary/device-software/cuda-programming-model

Modal

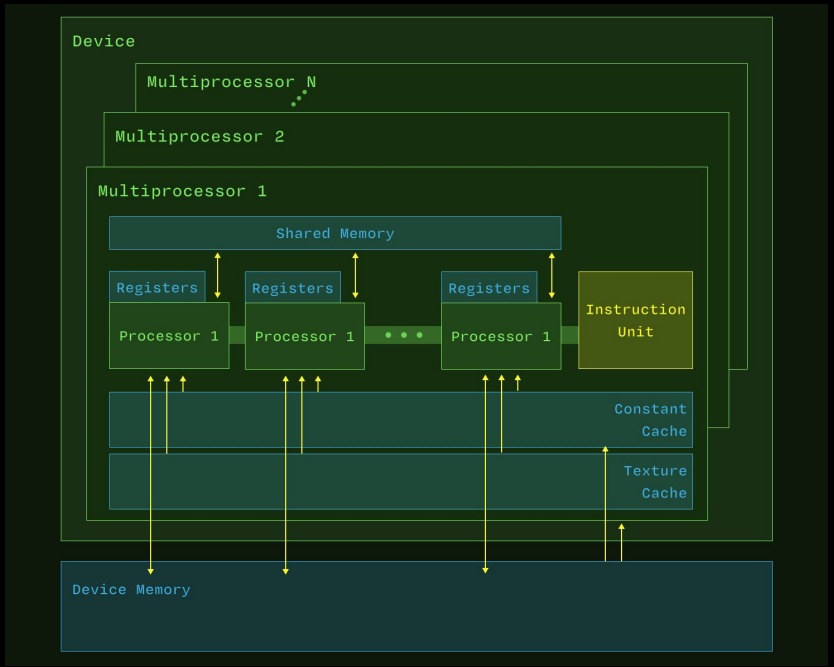
CUDA is a computer architecture principle.

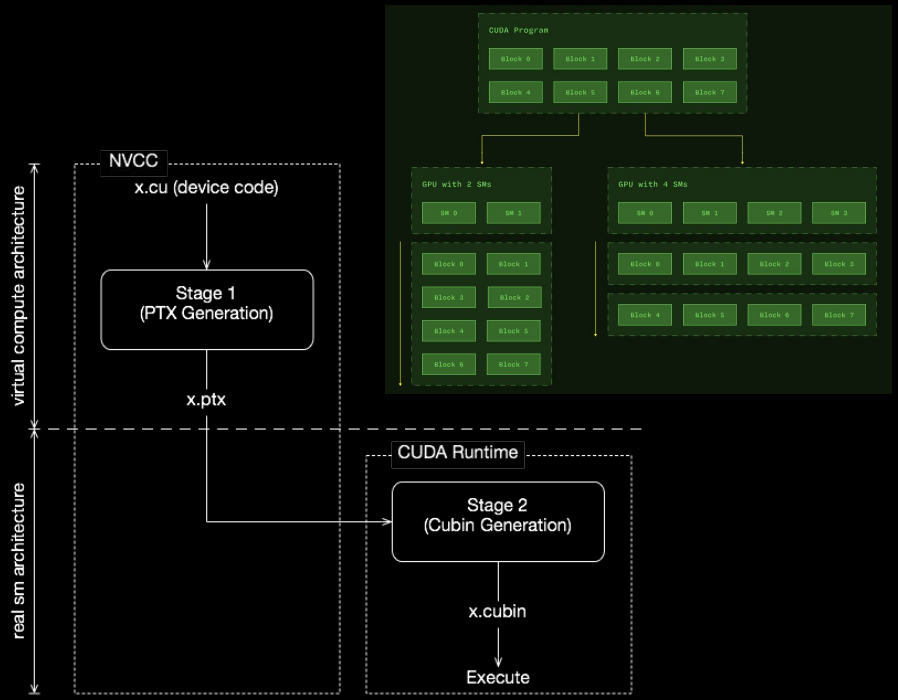


https://modal.com/gpu-glossary/device-hardware/cuda-device-architecture

Modal

The most important part of the CUDA stack isn't called CUDA.





https://modal.com/gpu-glossary/device-software/parallel-thread-execution

Modal

What is this?
Where did it come from?
What does it say?
Where is it going?

Short-term goals. Watch this space!

- ChatGPU
  - How do we make this as easy and extensible as possible? `llms.txt`?
- Interactive code snippets
  - Inspired by Rust By Example et al.
  - Will require a Modal acc't, but will fit in our free tier.
- Interactive diagrams
- Better content on synchronization
  - Atomics vs barriers
- Better content on warpgroups/thread block clusters

Mid-term goals. Looking for collaborators. We have the GPUs.

- Performance debugging
  - New terms: bank conflict, occupancy, coarsening
- GPU fleet mgmt
  - New terms: dcgm, thermal design power
- Multi-GPU hardware & programming
  - New terms: PCIe, SXM, NVLink, NCCL

Speculative goals. Can/should we do this?

- Multi-node hardware & programming
    - New terms: NVLink Switch, NIC, Ethernet, TCP, IP, Infiniband
- Triton?
    - We have even less experience here than in CUDA C++
- Open up the material on GitHub?
    - Open source succeeds when it deduplicates non-differentiating labor
- Online course? Partner with a university?

we're hiring btw :)

email charles@modal.com

if you want to go CUDA MODE