Team Members:

- ❖ Kuei-Po Huang HADLab@CGU
- ❖ Yun-Ting Zhang HADLab@CGU

Advisor: Prof. Chin-Fu Nien

Mentors:

- ❖ Reese Wang, Pika Wang @Nvidia

NARLabs 國家實驗研究院
國家高速網路與計算中心
National Center for High-performance Computing

OPEN HACKATHONS
OpenACC

Problem trying to solve

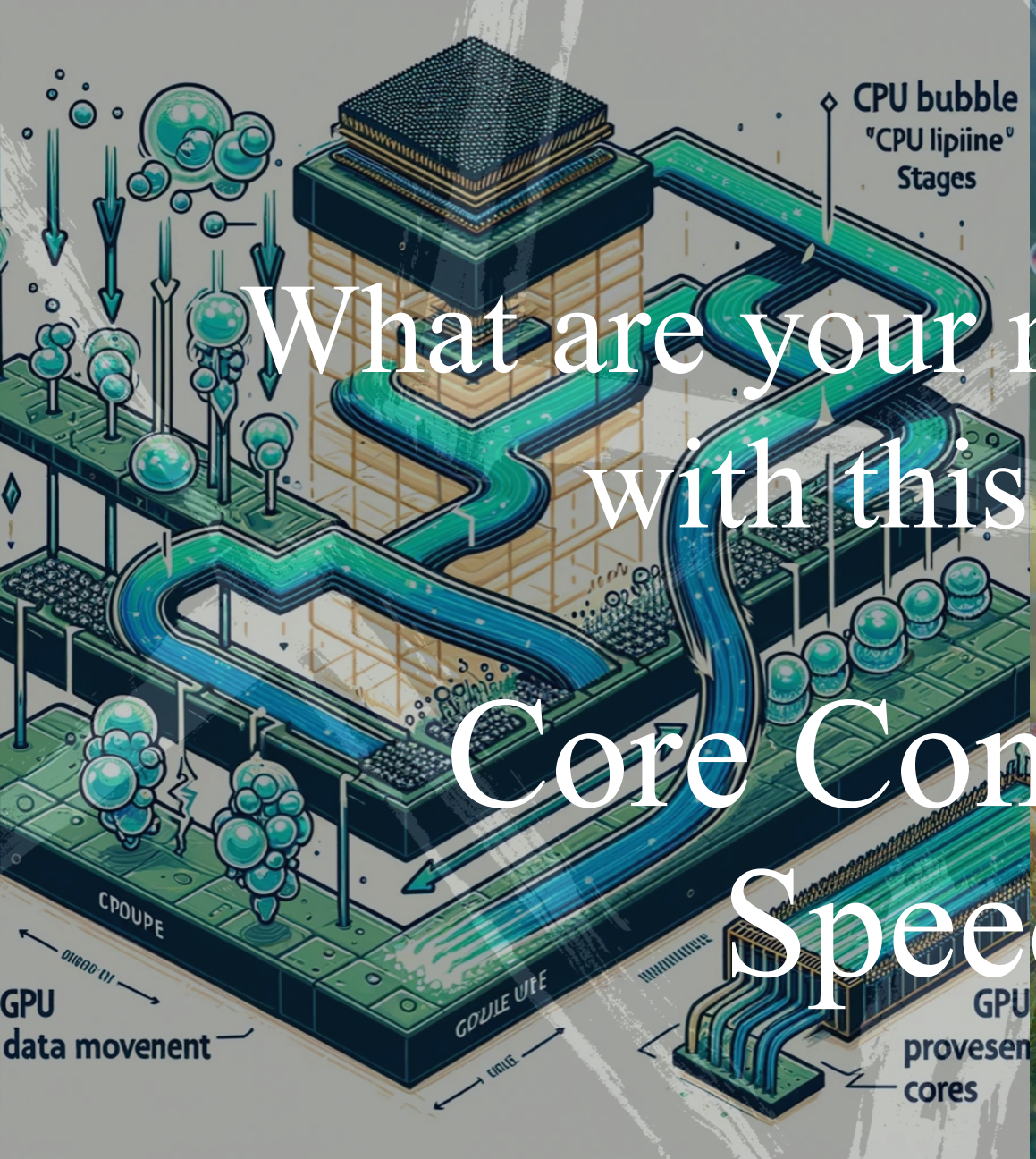
- Combinatorial Optimization Problems.

Scientific driver for the chosen algorithm

- Quantum-Inspired Algorithm.

What's the algorithmic motif?

- Providing more diversity solutions.



What are your main objectives with this project?

Core Computation Speed Up

Surpassing Limits: Setting New Milestones in HPC.

What was your goal coming here?

- Porting core computation onto GPU.

What was your initial strategy?

- CPU and GPU Hybrid.

How did this strategy change?

- The strategy improved by using **NVIDIA Nsight** for performance profiling and **JAX** for efficient computation.

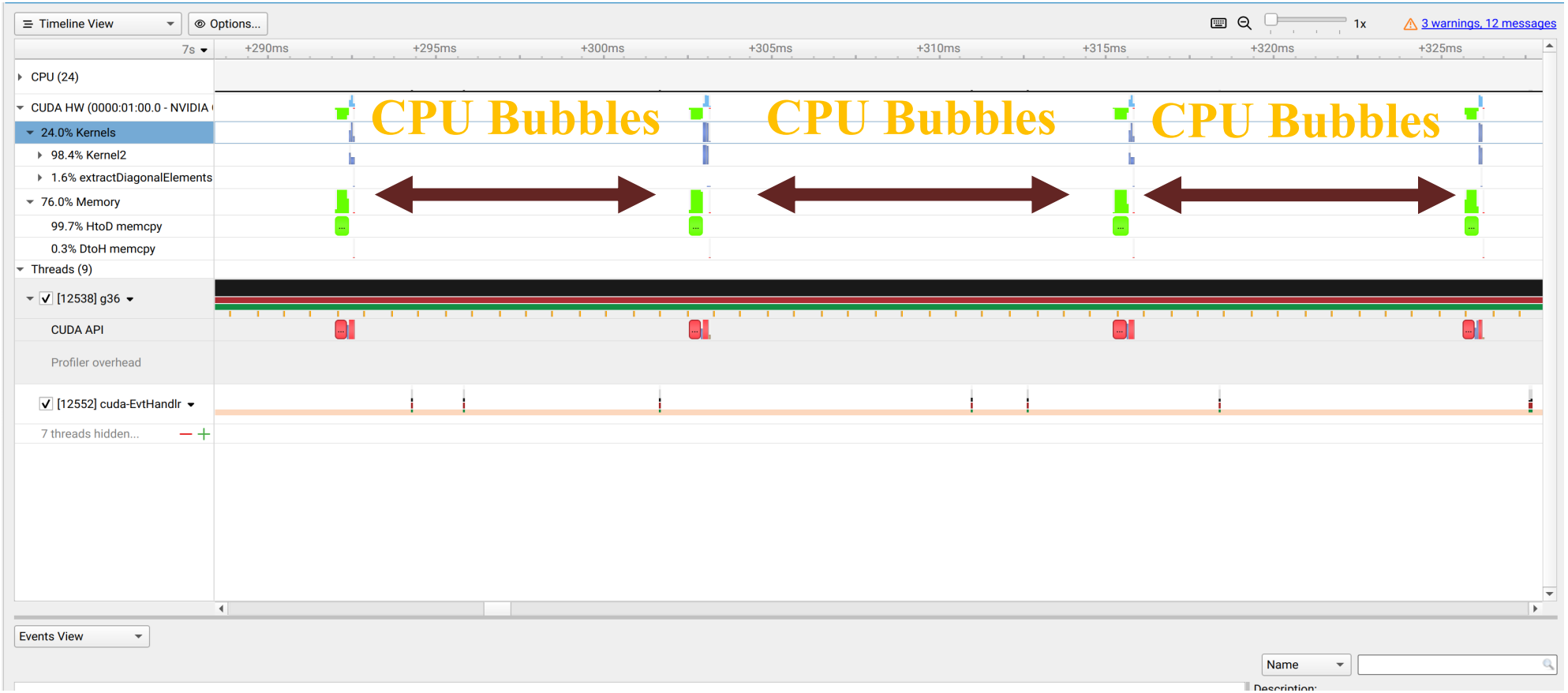
C++ Hybrid Pseudo Code

for iter in iterations:

- CPU pre-processing
- GPU core computing
- CPU post processing

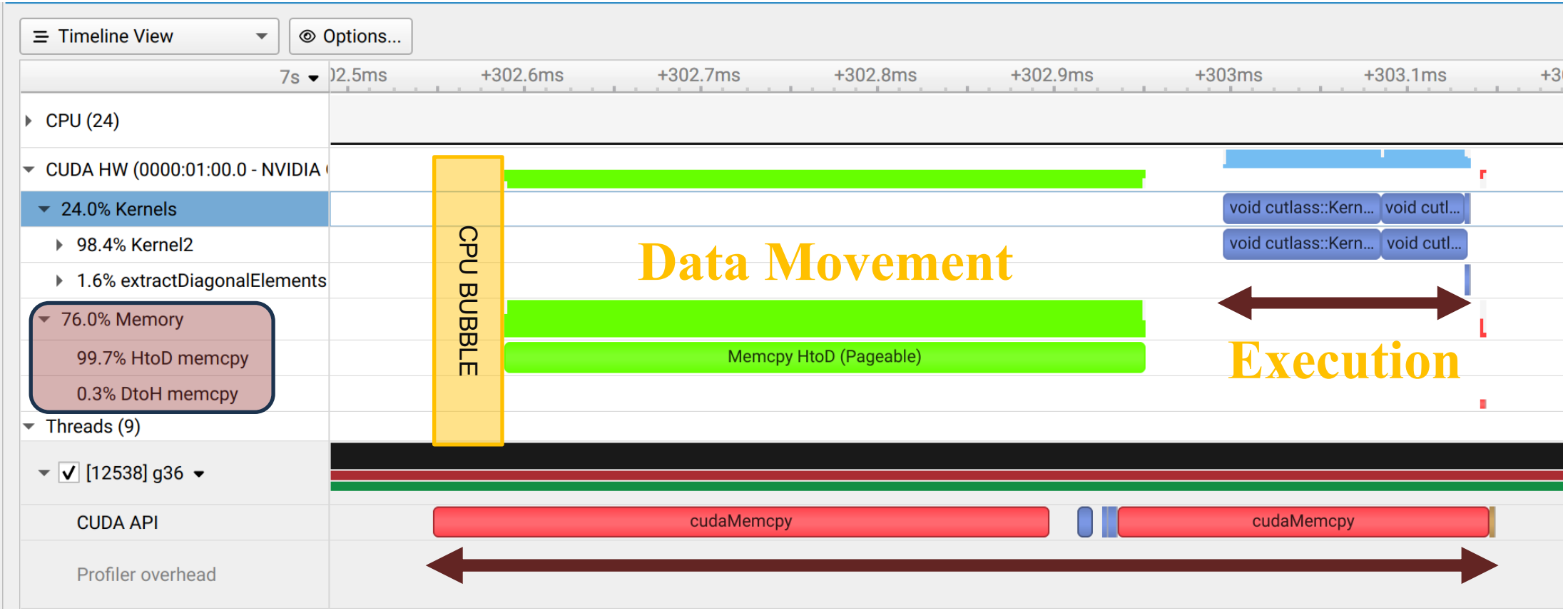
Nsight Profiler

CPU Bubbles



Nsight Profiler

Data Movement



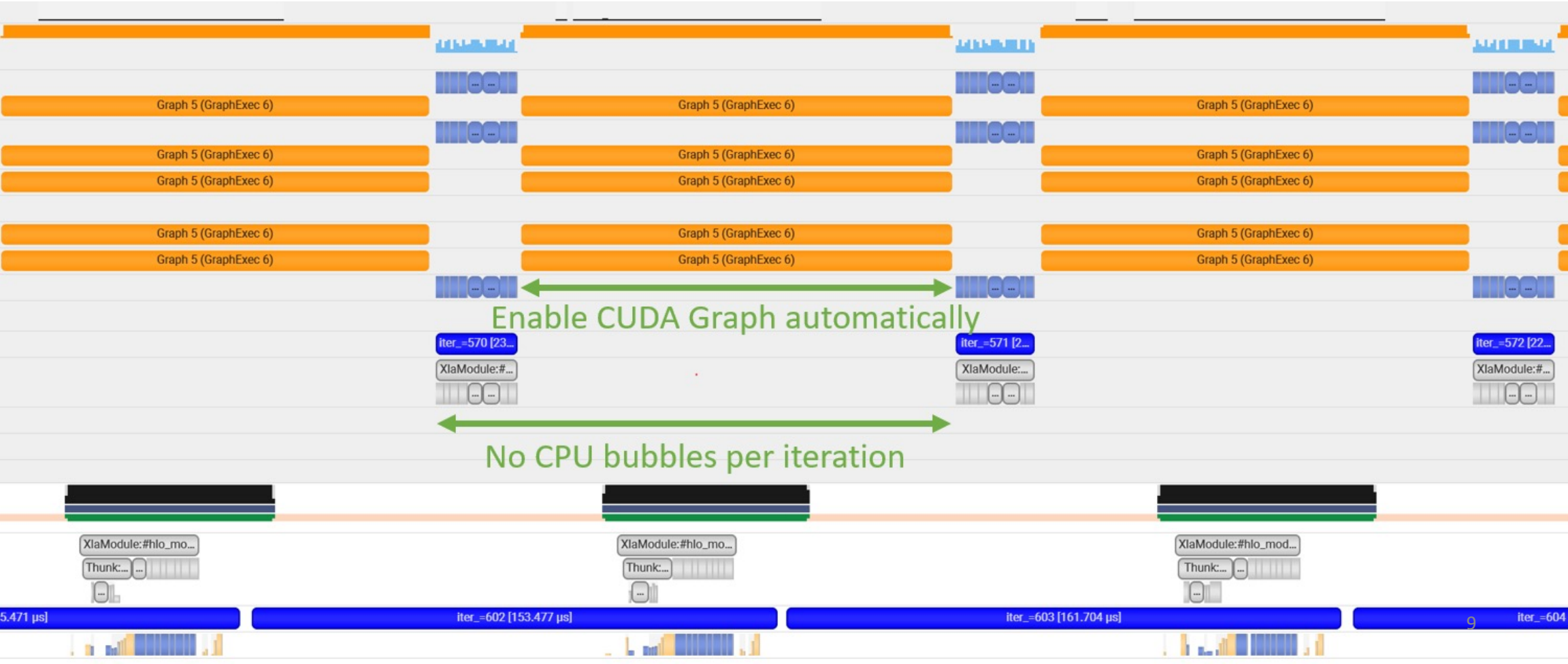
WHY **JAX**

Our mentor suggested us to use **JAX**.

It uses XLA to compile Python functions on GPUs, which means we don’t need to write CUDA kernel.

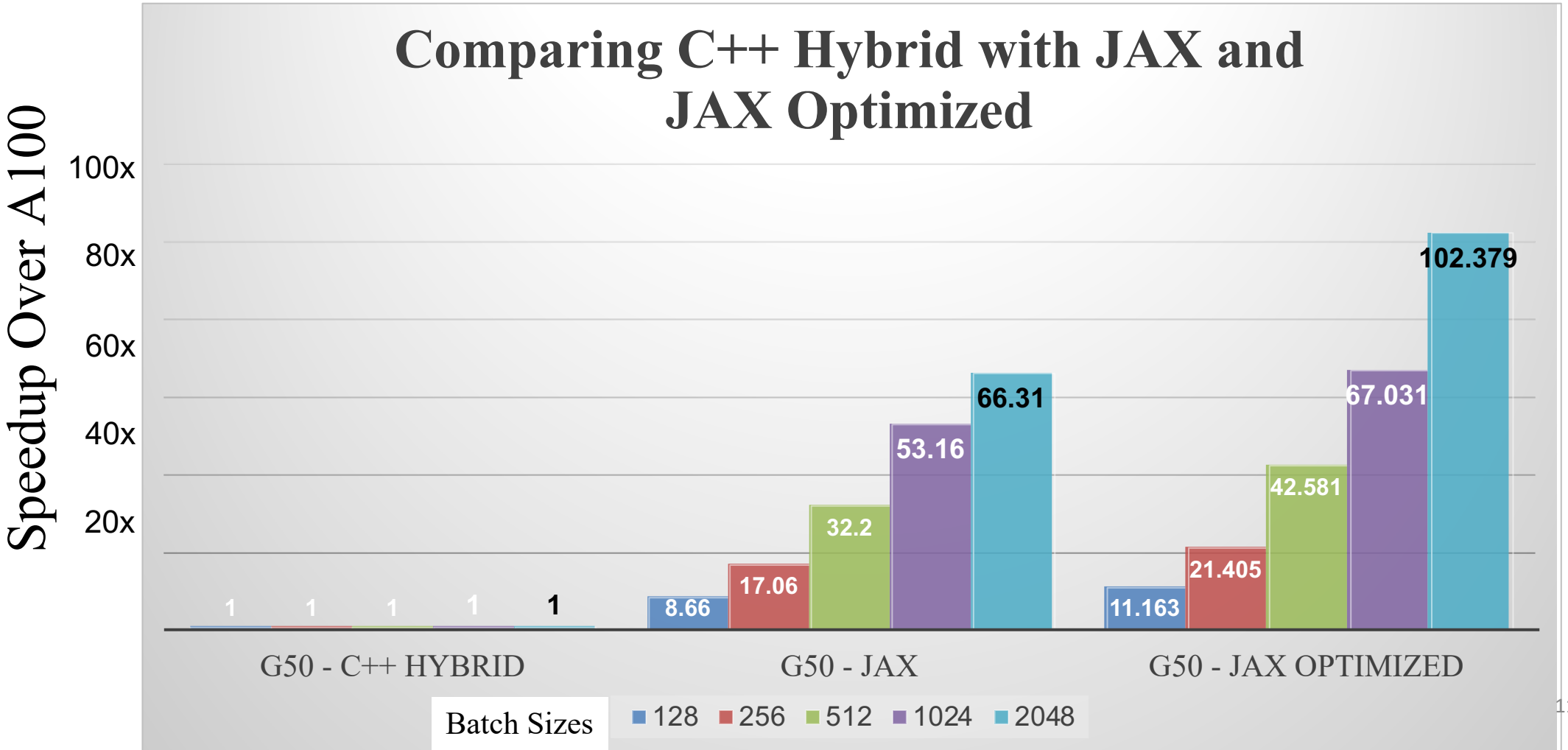
JAX can automatically generate the fusion kernels, which eliminates the frequently global memory access and reduce the launched kernels to achieve high GPU utilization.

Nsight: No CPU Bubbles with JAX

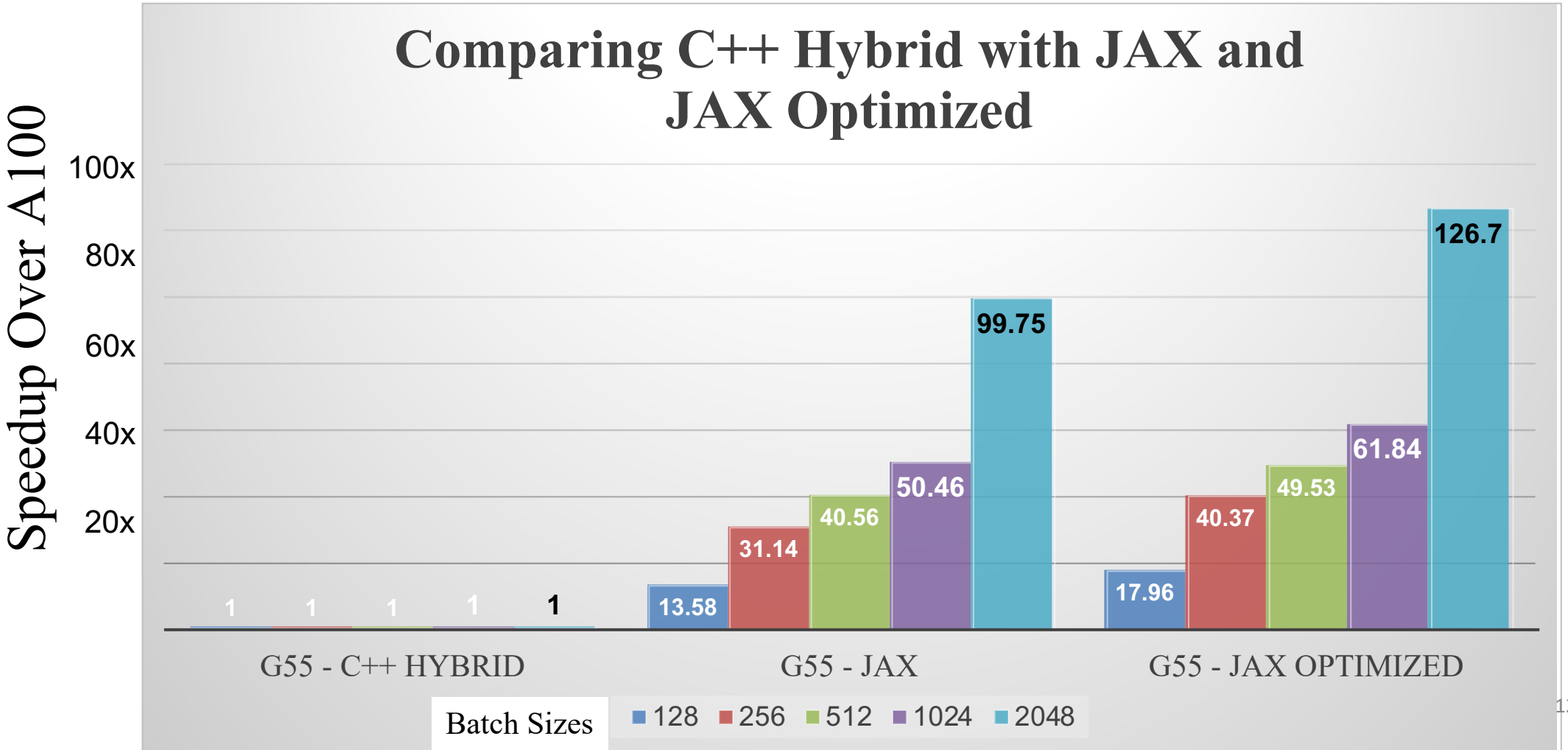


What is the before-after Speedup?

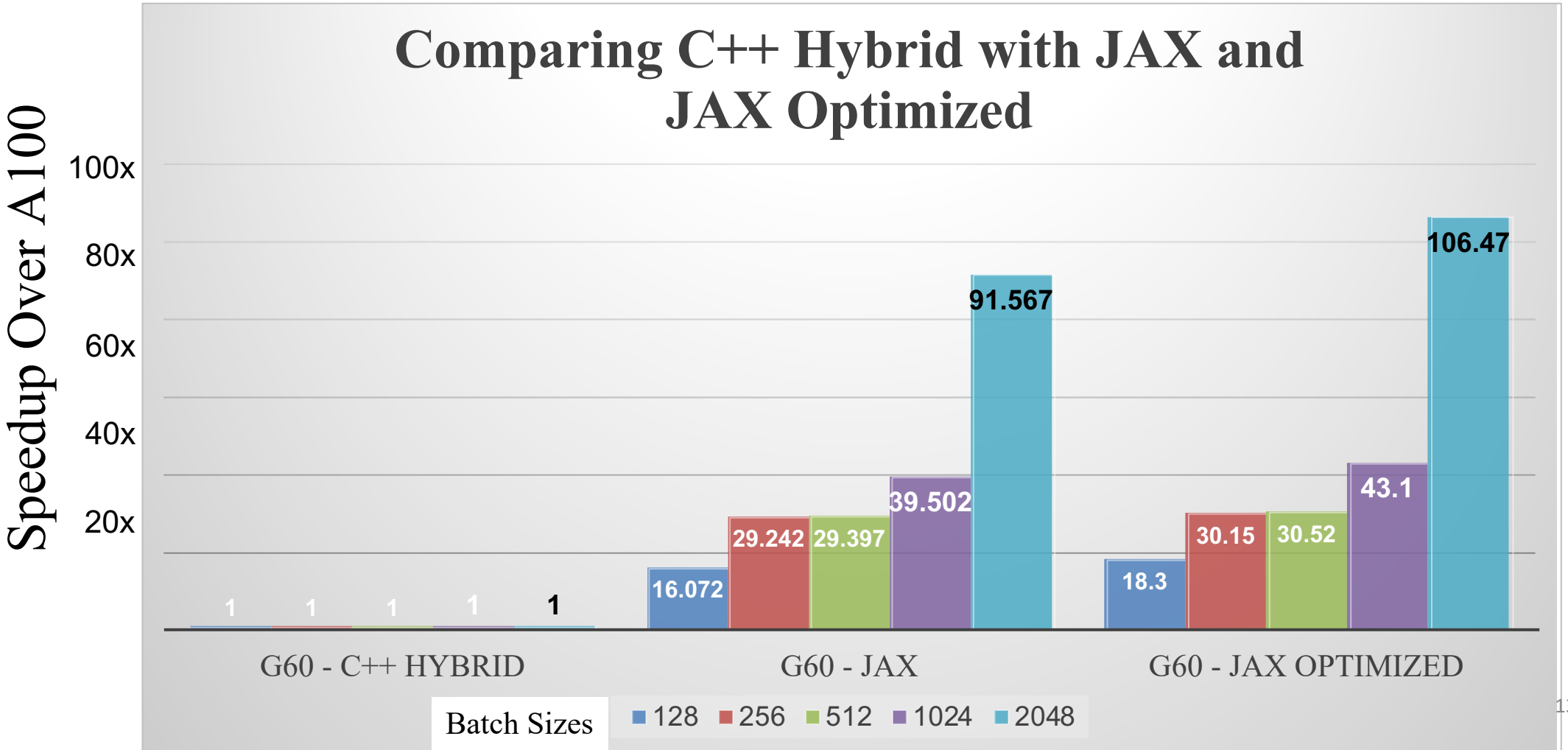
Unlocking the Power of JAX: Accelerating G50 Analysis with 3000-D Capabilities



Unlocking the Power of JAX: Accelerating G55 Analysis with 5000-D Capabilities



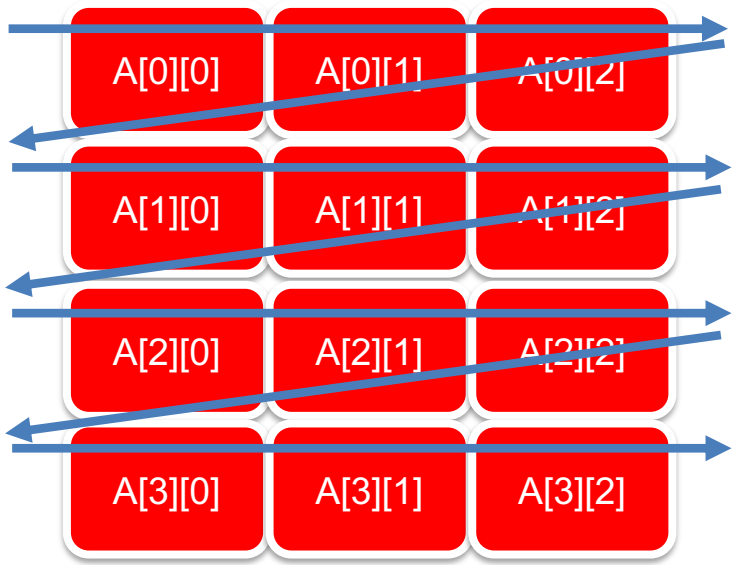
Unlocking the Power of JAX: Accelerating G60 Analysis with 7000-D Capabilities

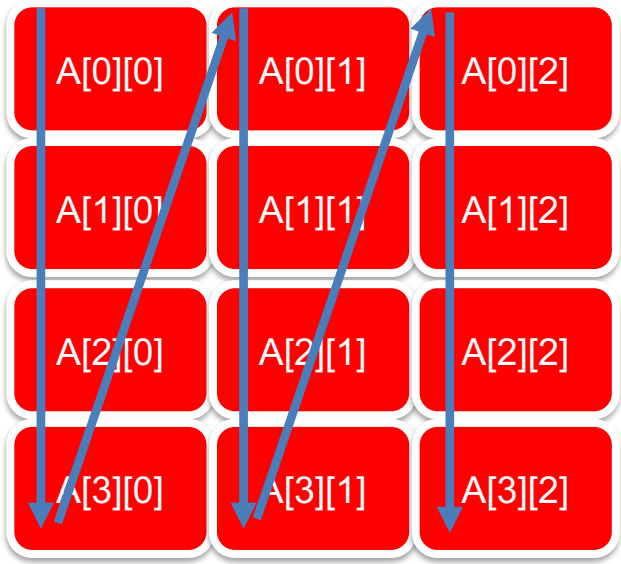


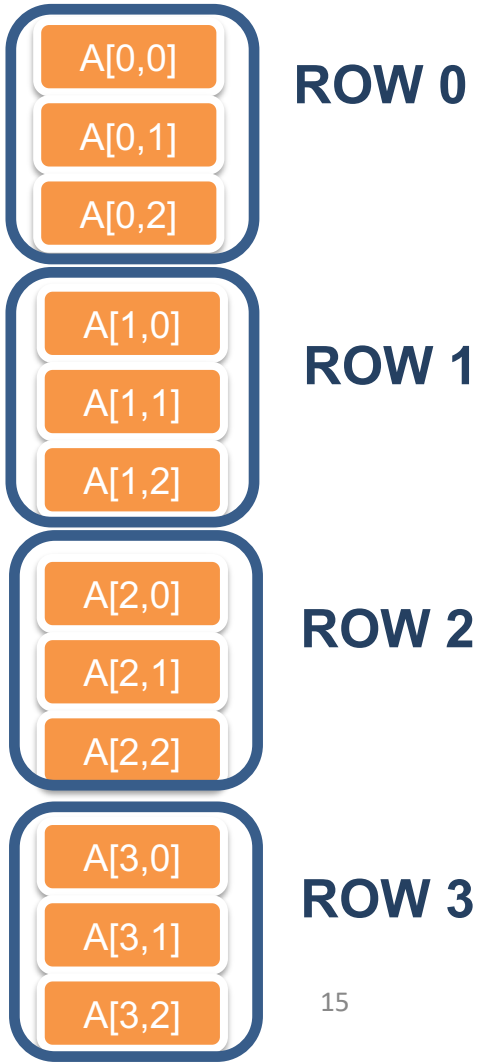
How does acceleration make an impactful difference?

- The **CPU** implementation is significantly **less suitable in applications** where timeliness is critical, such as high-frequency trading or real-time route planning, which demand rapid decision-making within a limited timeframe.

What problems have you encountered?







A product that will increase customer stickiness for Nvidia.

Wishlist

What do you wish existed to make your life easier?

Summary

- Utilizing JAX and **Nsight** system to find the bottleneck and easily port it into GPU.
- Achieving up to **126x** end to end speedup.
- JAX is easily extended on our algorithm comparing to C++ **CUDA kernel**.
- Moreover, the mentor's insights into **Nvidia** tools enabled us to effectively analyze and optimize our algorithms.