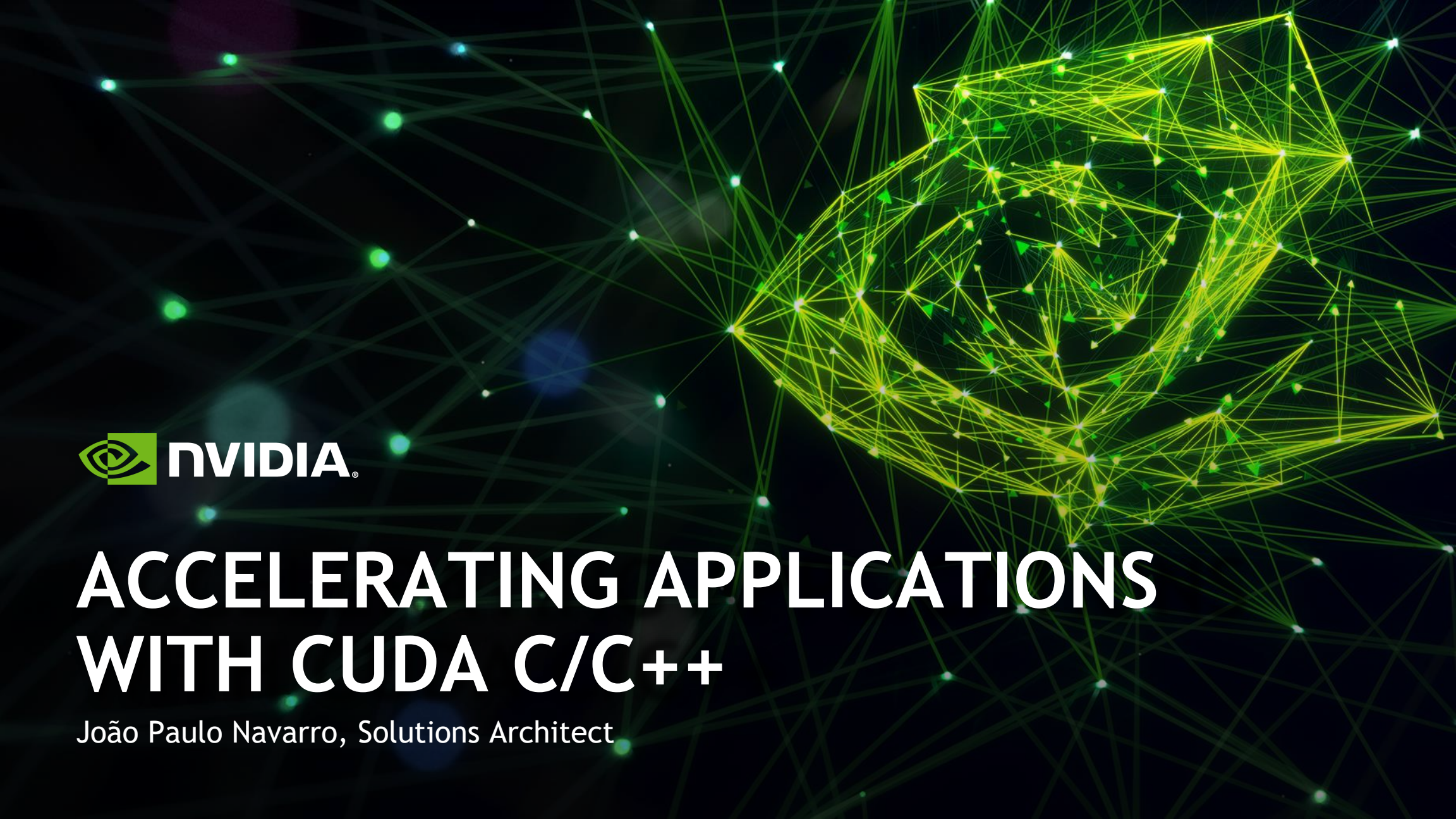


ACCELERATING APPLICATIONS WITH CUDA C/C++

João Paulo Navarro, Solutions Architect

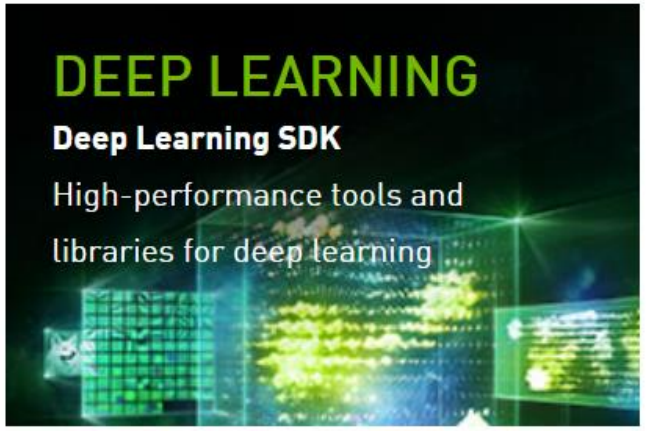
HTTPS://WWW.YOUTUBE.COM/WATCH?V=-P28LKWTZRI

Join the NVIDIA Developer Program

Access everything you need to develop with NVIDIA products.

Register Now

developer.nvidia.com



Join the NVIDIA Developer Program

Access everything you need to develop with NVIDIA products.

Register Now

developer.nvidia.com

DEEP LEARNING
Deep Learning SDK

High-performance tools and libraries for deep learning

ACCELERATED COMPUTING
NVIDIA ComputeWorks

Everything scientists and engineers need to build GPU-accelerated applications

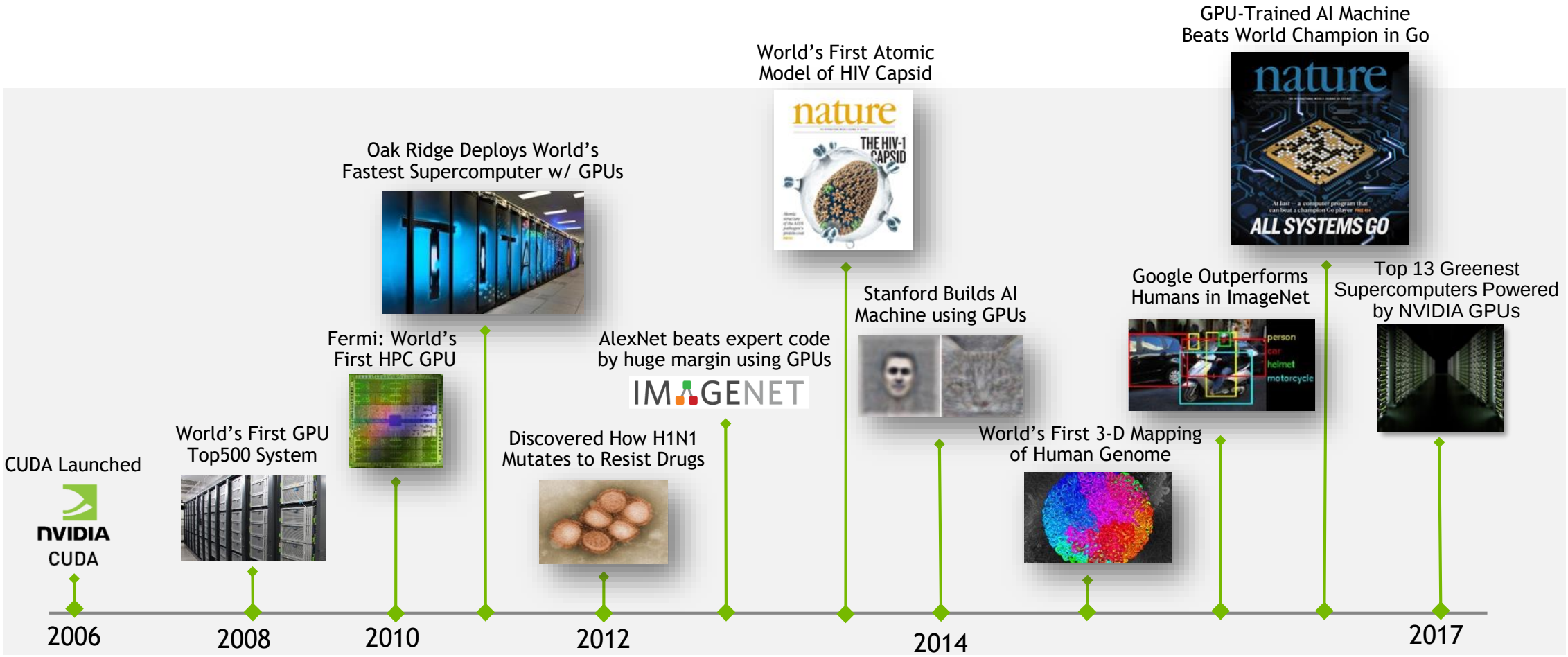
AUTONOMOUS VEHICLES
NVIDIA DRIVE Platform

Deep learning, HD mapping and supercomputing solutions, from ADAS to fully autonomous

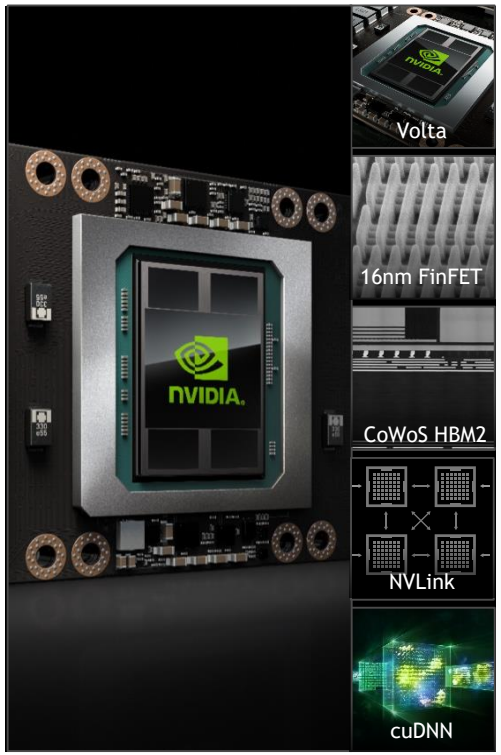
SMART CITIES
NVIDIA Metropolis

Edge-to-cloud development platform for smart cities

ELEVEN YEARS OF GPU COMPUTING

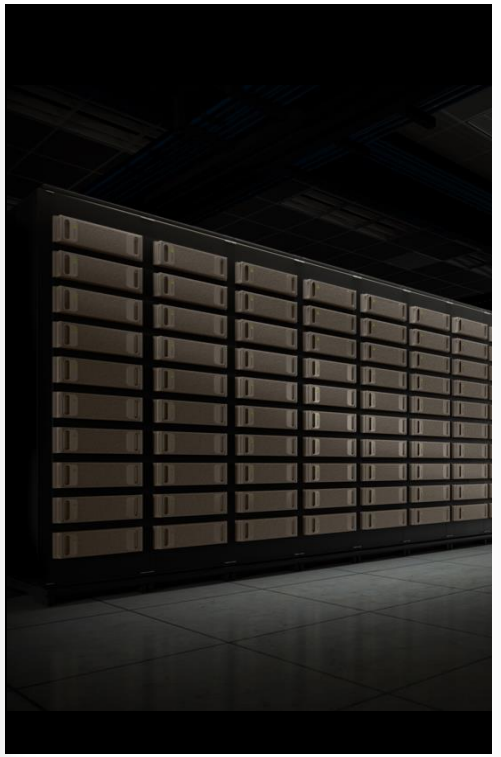


NVIDIA IS DEEPLY INVESTED IN GPU COMPUTING

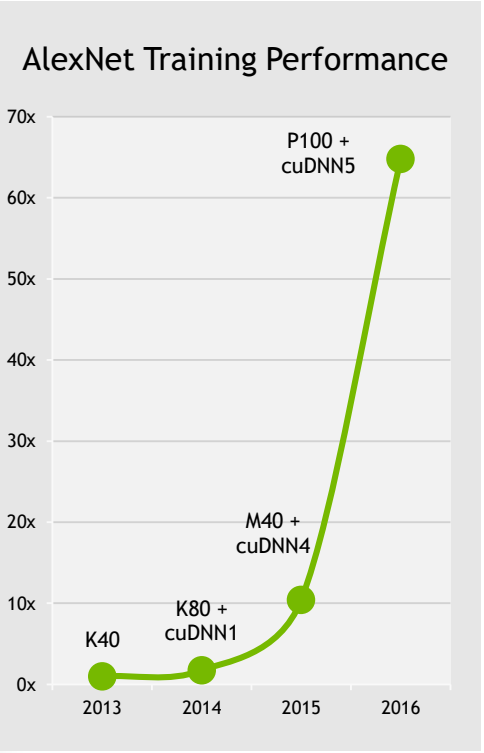


V100 Miracles

NVIDIA DGX-1



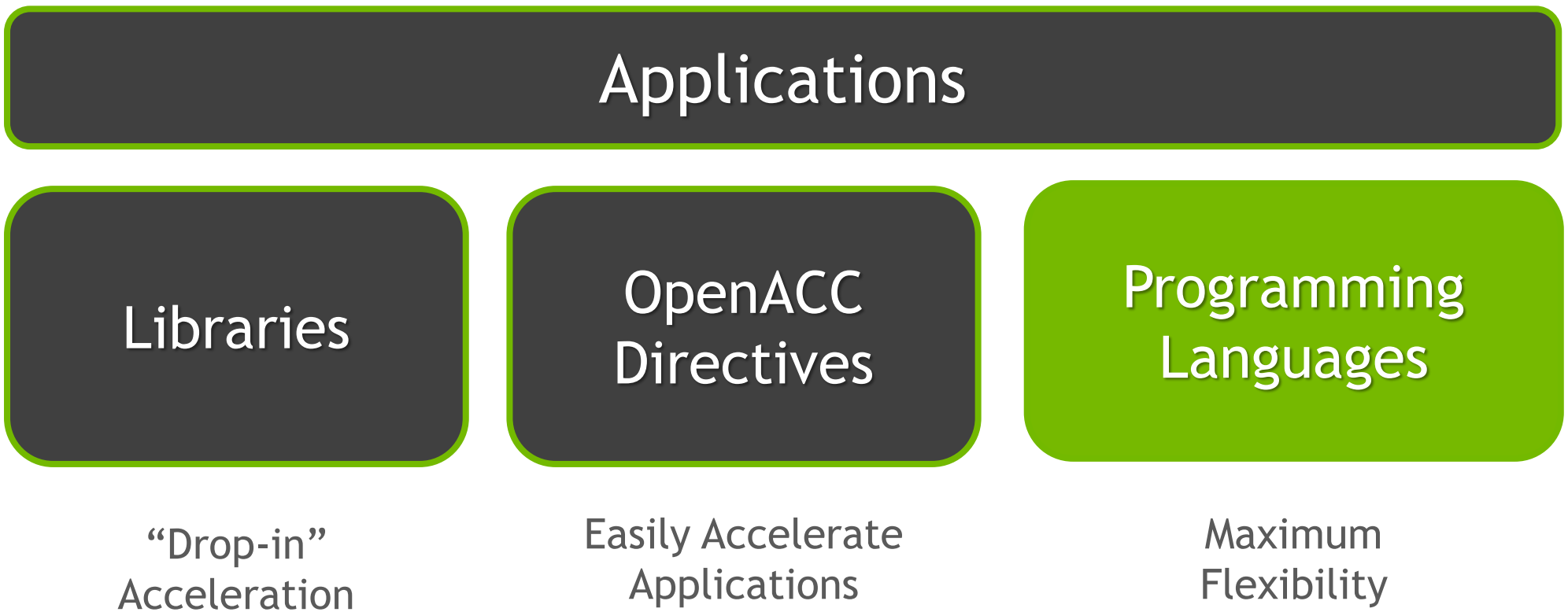
NVIDIA DGX SATURNV



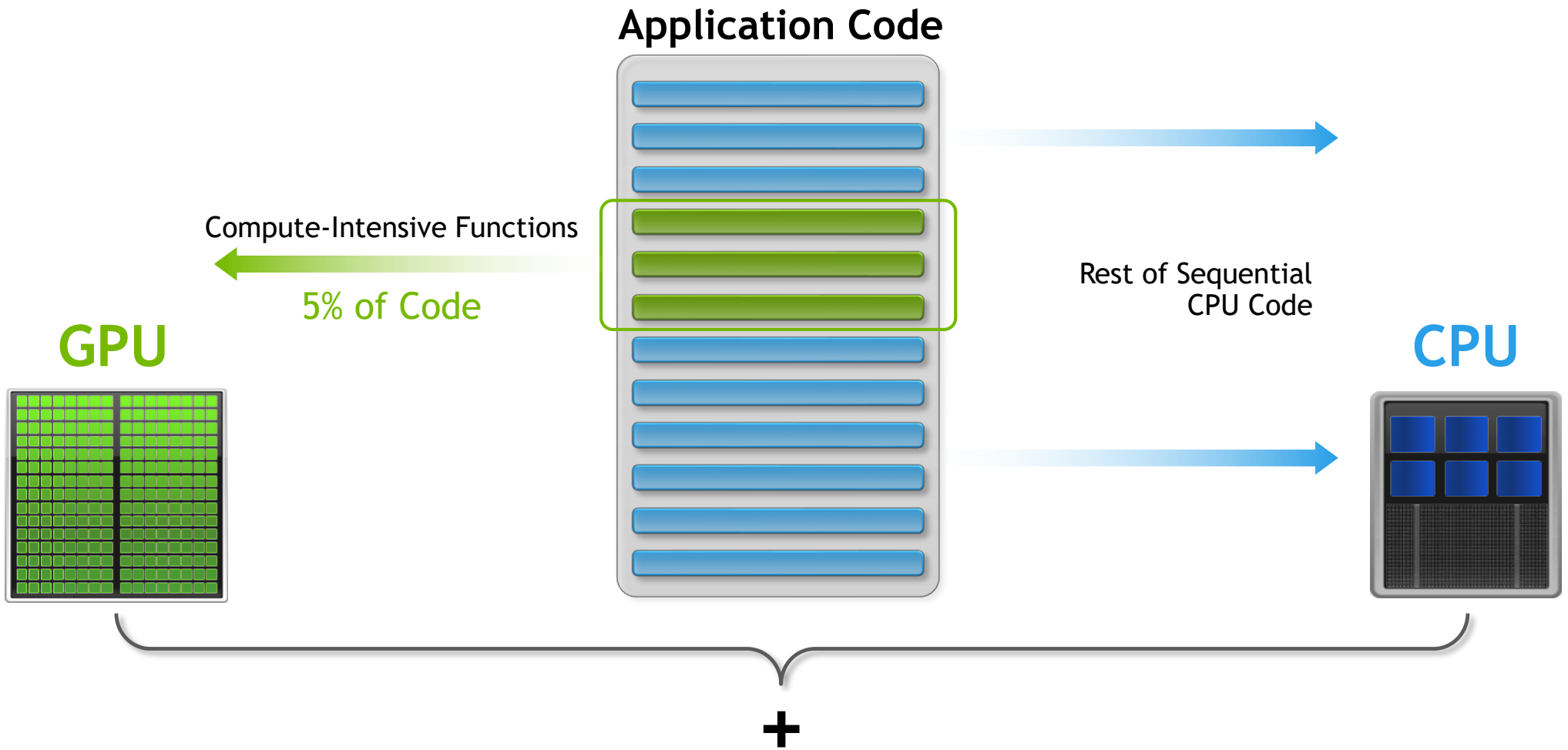
65x in 3 Years


GPU PROGRAMMING

3 WAYS TO ACCELERATE APPLICATIONS



HOW GPU ACCELERATION WORKS

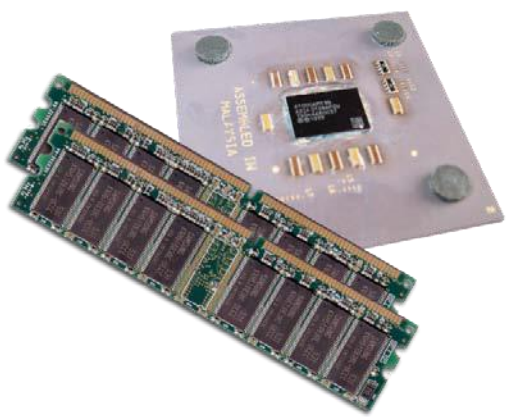


THE BASICS

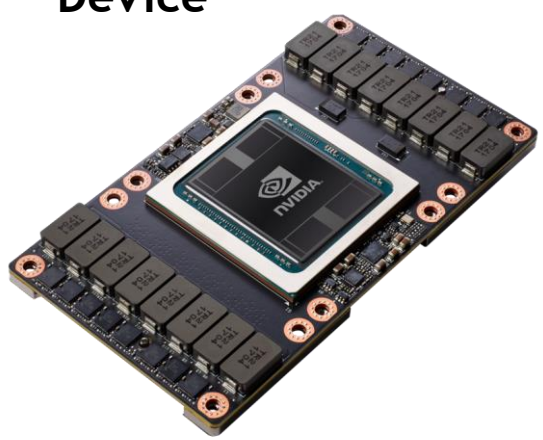
Heterogenous Computing

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

Host



Device




ACCELERATING C/C++
CODE WITH CUDA ON
GPUS

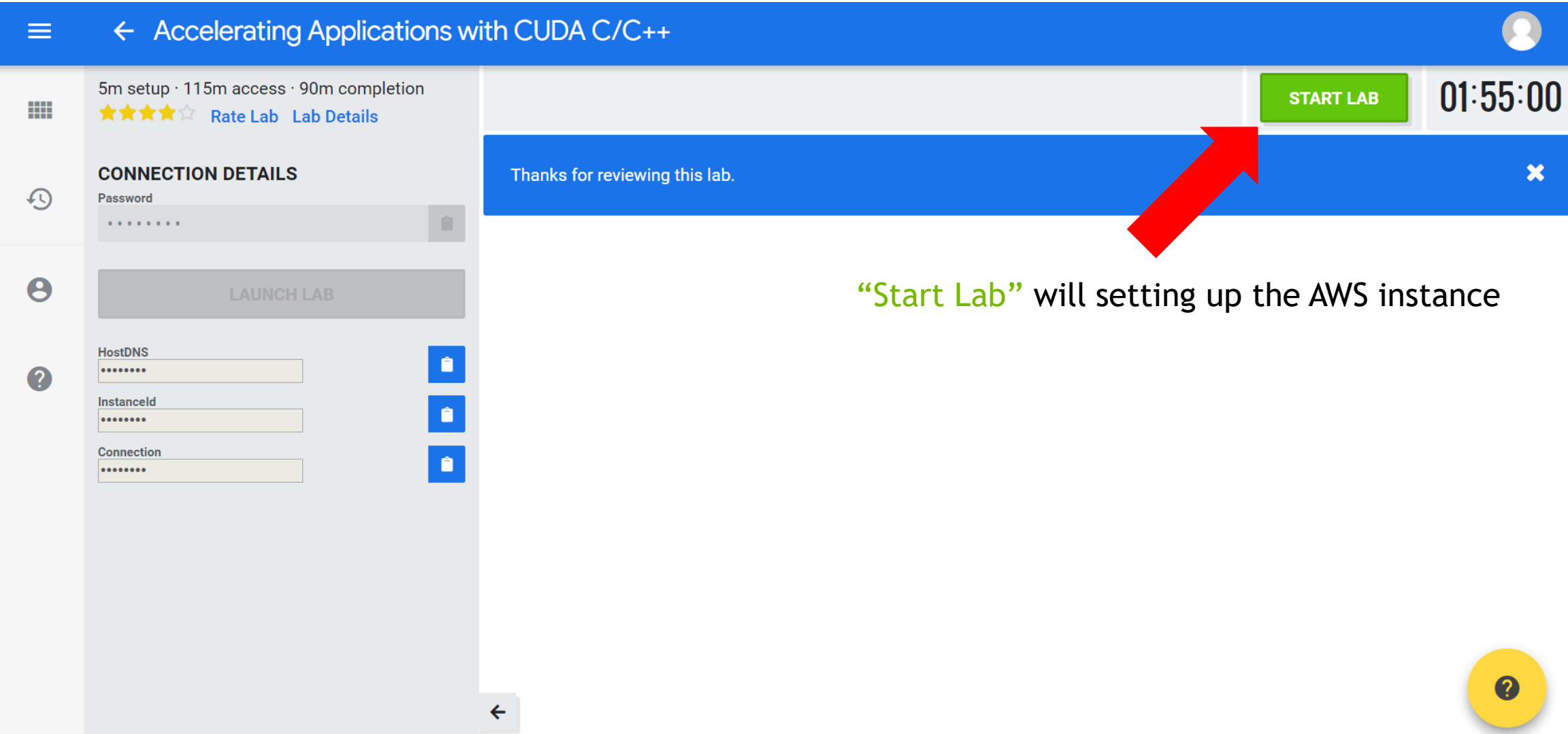
ACCELERATING APPLICATIONS WITH CUDA C/C++

Hands-On Lab

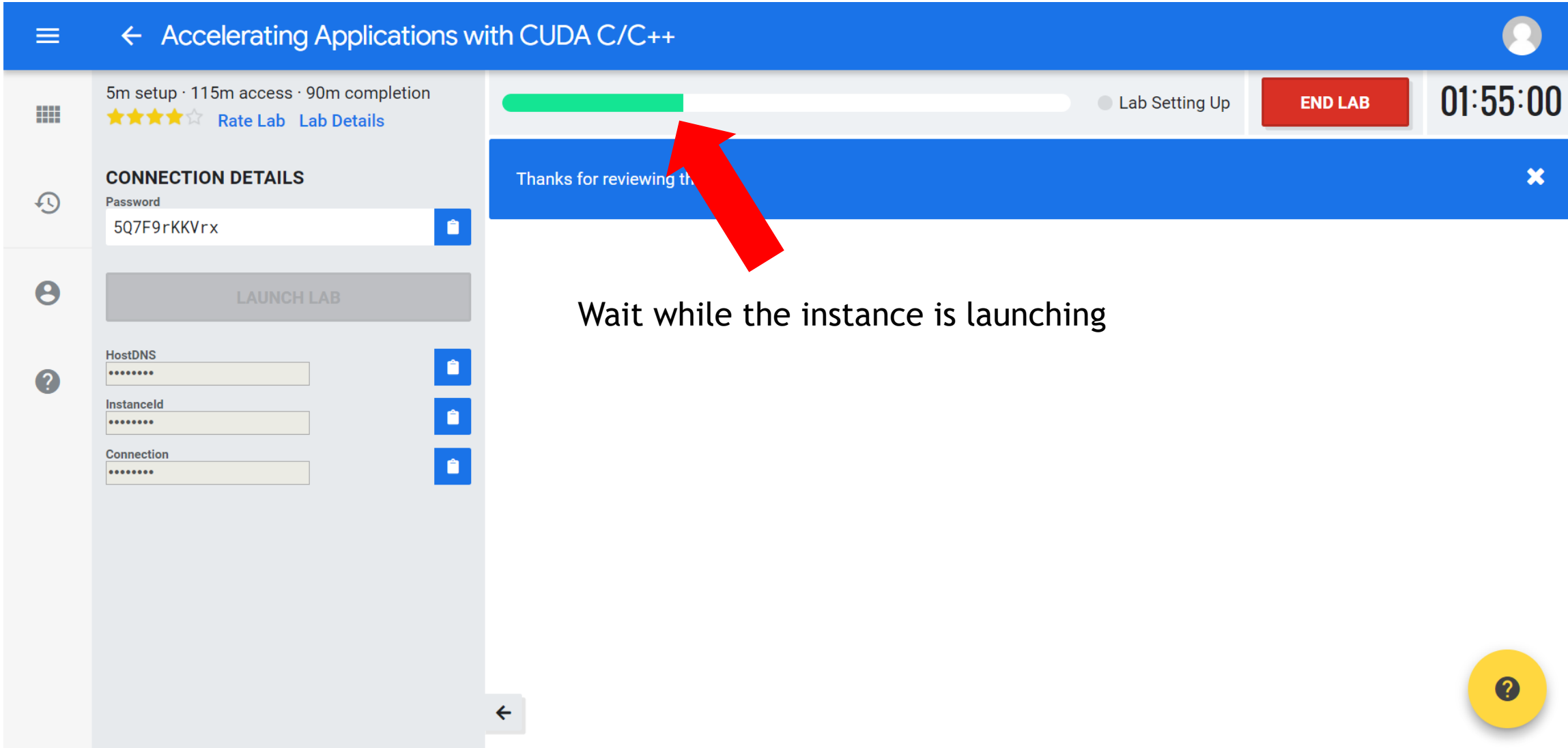
Register at

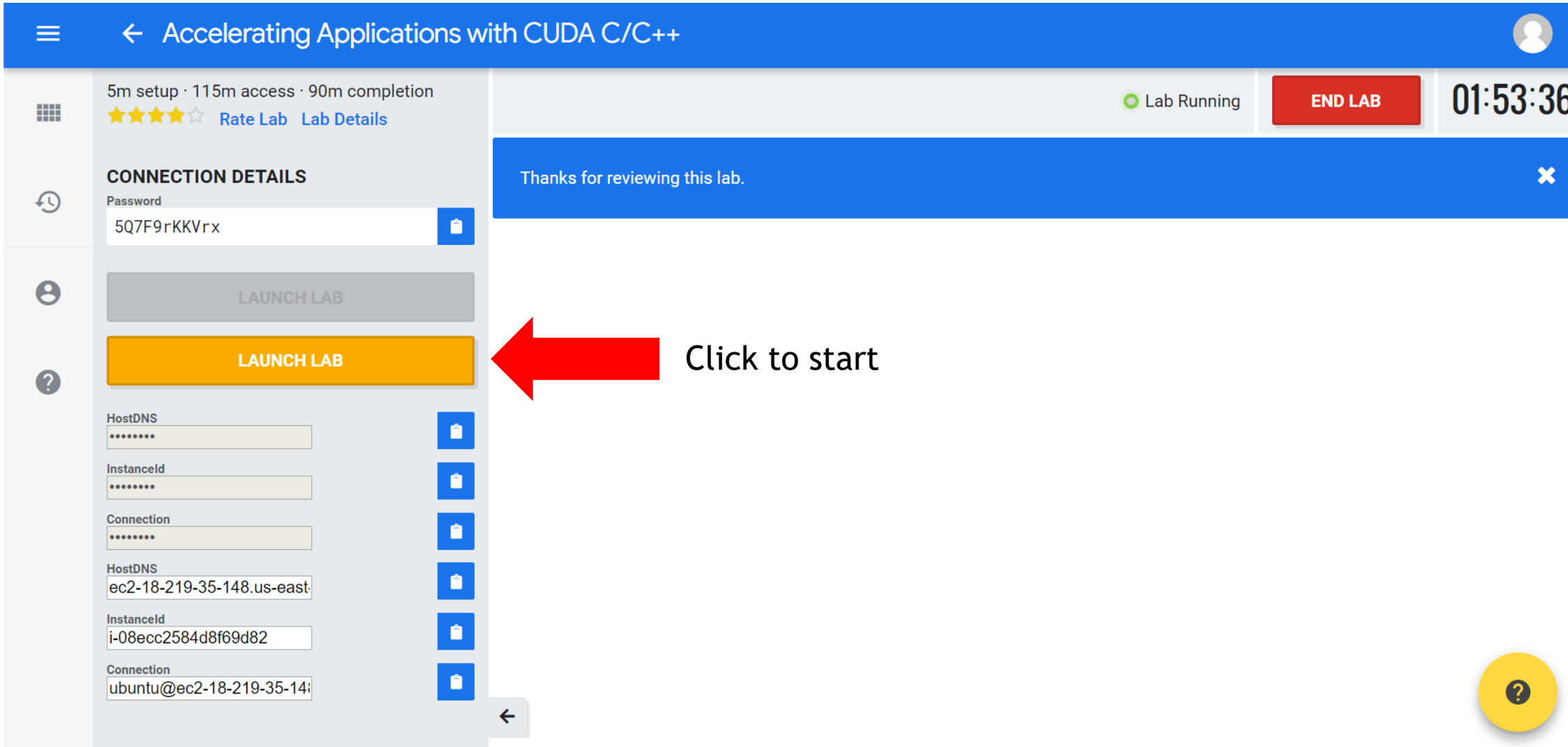
https://developer.nvidia.com/

"NVIDIA QwikLabs"
https://nvlabs.qwiklab.com



"Start Lab" will setting up the AWS instance

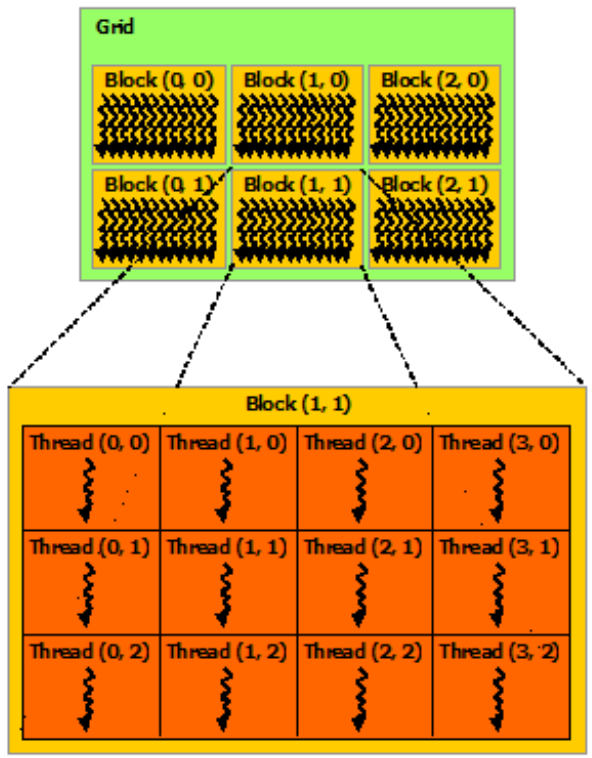

Accelerating Applications with CUDA C/C++
5m setup · 115m access · 90m completion
Rate Lab
Lab Details
CONNECTION DETAILS
Password
5Q7F9rKKVrx
LAUNCH LAB
HostDNS
InstanceId
Connection
Lab Setting Up
END LAB
01:55:00
Thanks for reviewing th
Wait while the instance is launching


Accelerating Applications with CUDA C/C++
5m setup · 115m access · 90m completion
Rate Lab
Lab Details
CONNECTION DETAILS
Password
5Q7F9rKKVrx
LAUNCH LAB
LAUNCH LAB
HostDNS
InstanceId
Connection
HostDNS
ec2-18-219-35-148.us-east
InstanceId
i-08ecc2584d8f69d82
Connection
ubuntu@ec2-18-219-35-14
Lab Running
END LAB
01:53:36
Thanks for reviewing this lab.
Click to start


V100 ARCHITECTURE

THREAD HIERARCHY

Grid, Block & Threads



TESLA V100

21B transistors
815 mm²

80 SM
5120 CUDA Cores
640 Tensor Cores

16 GB HBM2
900 GB/s HBM2
300 GB/s NVLink

*full GV100 chip contains 84 SMs

VOLTA GV100 SM

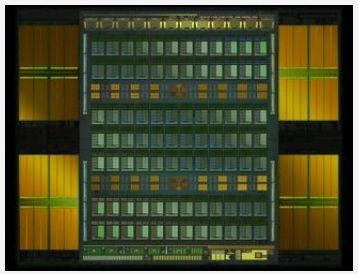
	GV100
FP32 units	64
FP64 units	32
INT32 units	64
Tensor Cores	8
Register File	256 KB
Unified L1/Shared memory	128 KB
Active Threads	2048

NVIDIA

TESLA V100

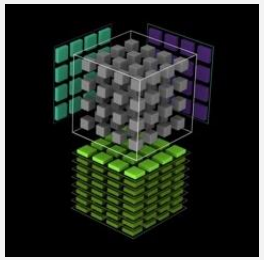
The Fastest and Most Productive GPU for AI and HPC

Volta Architecture



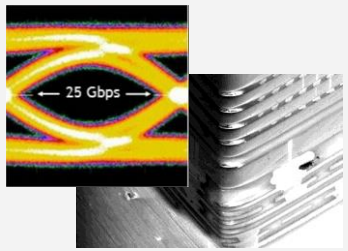
Most Productive GPU

Tensor Core



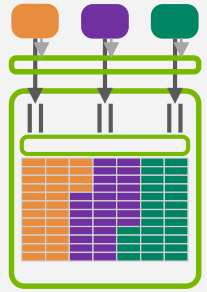
125 Programmable TFLOPS Deep Learning

Improved NVLink & HBM2



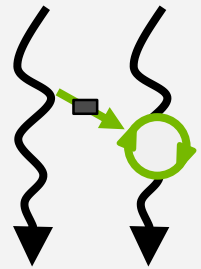
Efficient Bandwidth

Volta MPS

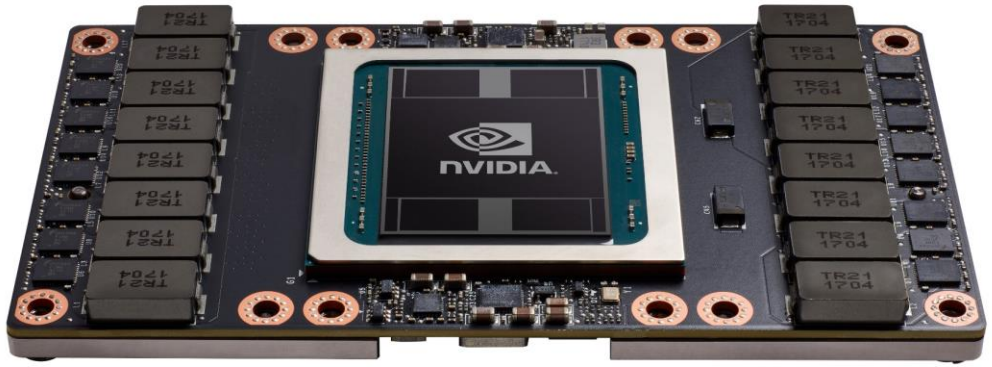


Inference Utilization

Improved SIMT Model



New Algorithms



HOW TO CONTINUING LEARNING?

DLI and Hands-on Labs

https://www.nvidia.com/en-us/deep-learning-ai/education/

- Self-paced labs for all NVIDIA technologies
  - https://nvidia.qwiklab.com

NVIDIA HW GRANT PROGRAM

Titan X Pascal

- Scientific Computing
- HPC
- Deep Learning

Quadro P6000

- Scientific Visualization
- Virtual Reality

Jetson TX2 (Dev Kit)

- Robotics
- Autonomous Machines

https://developer.nvidia.com/academic_gpu_seeding

http://www.nvidia.com/object/inception-program.html


NVIDIA