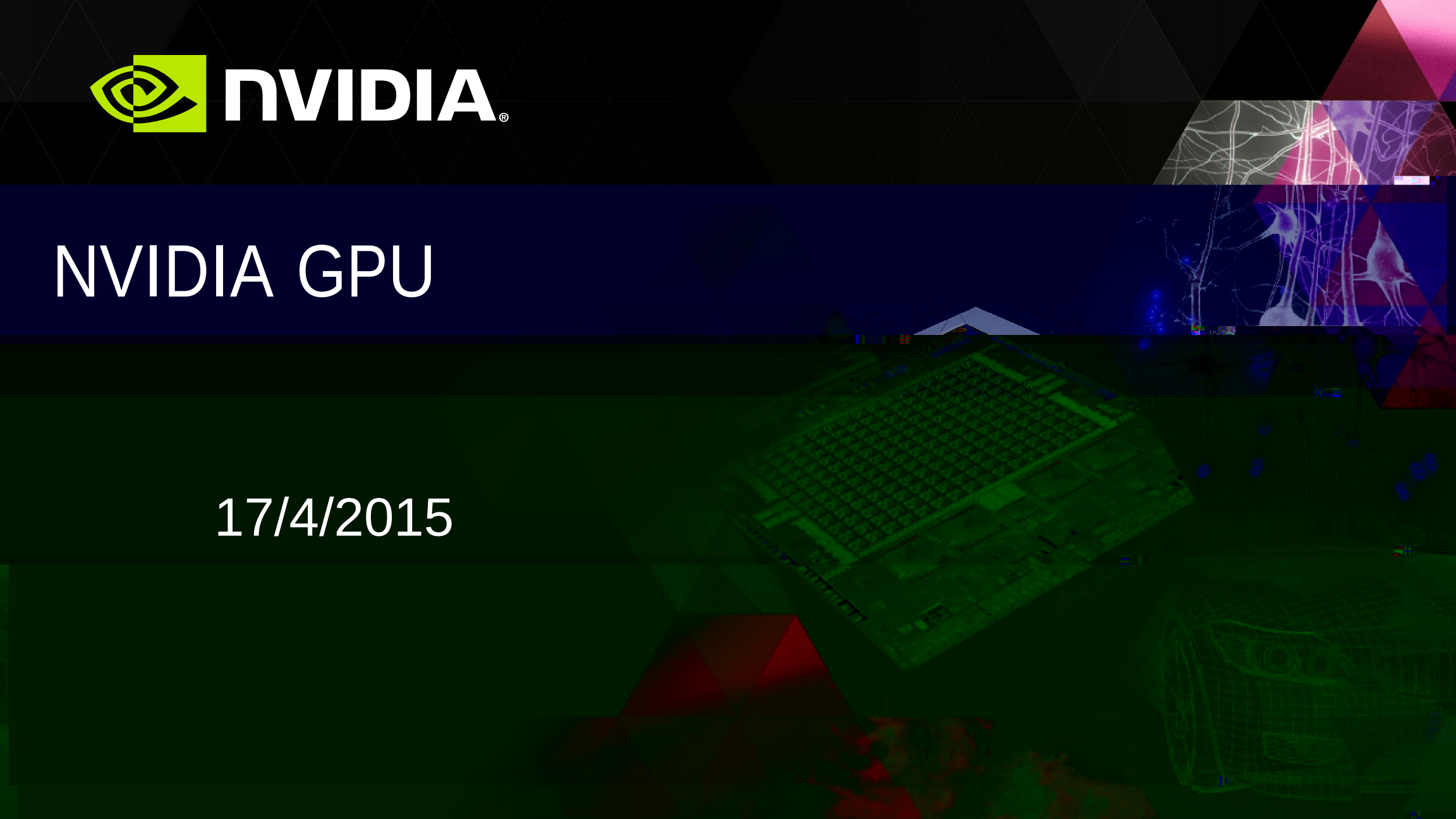
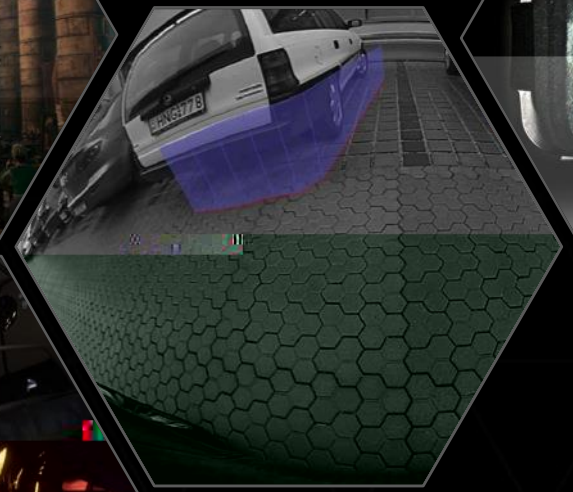
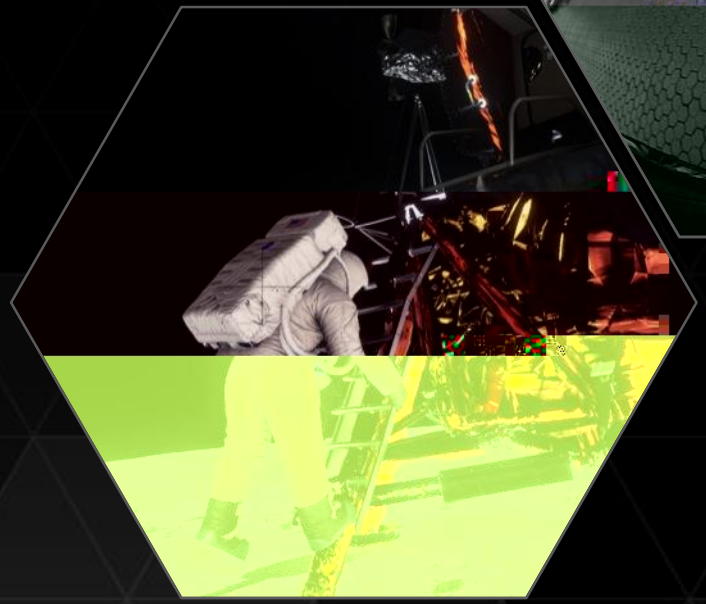


NVIDIA GPU

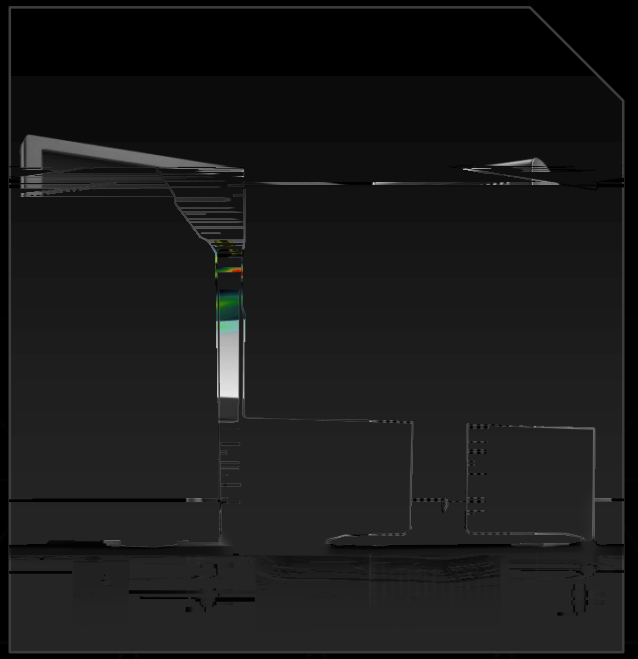
17/4/2015




nature
THE HIV-1 CAPSID
Atomic structure of the AIDS



GeForce | GRID



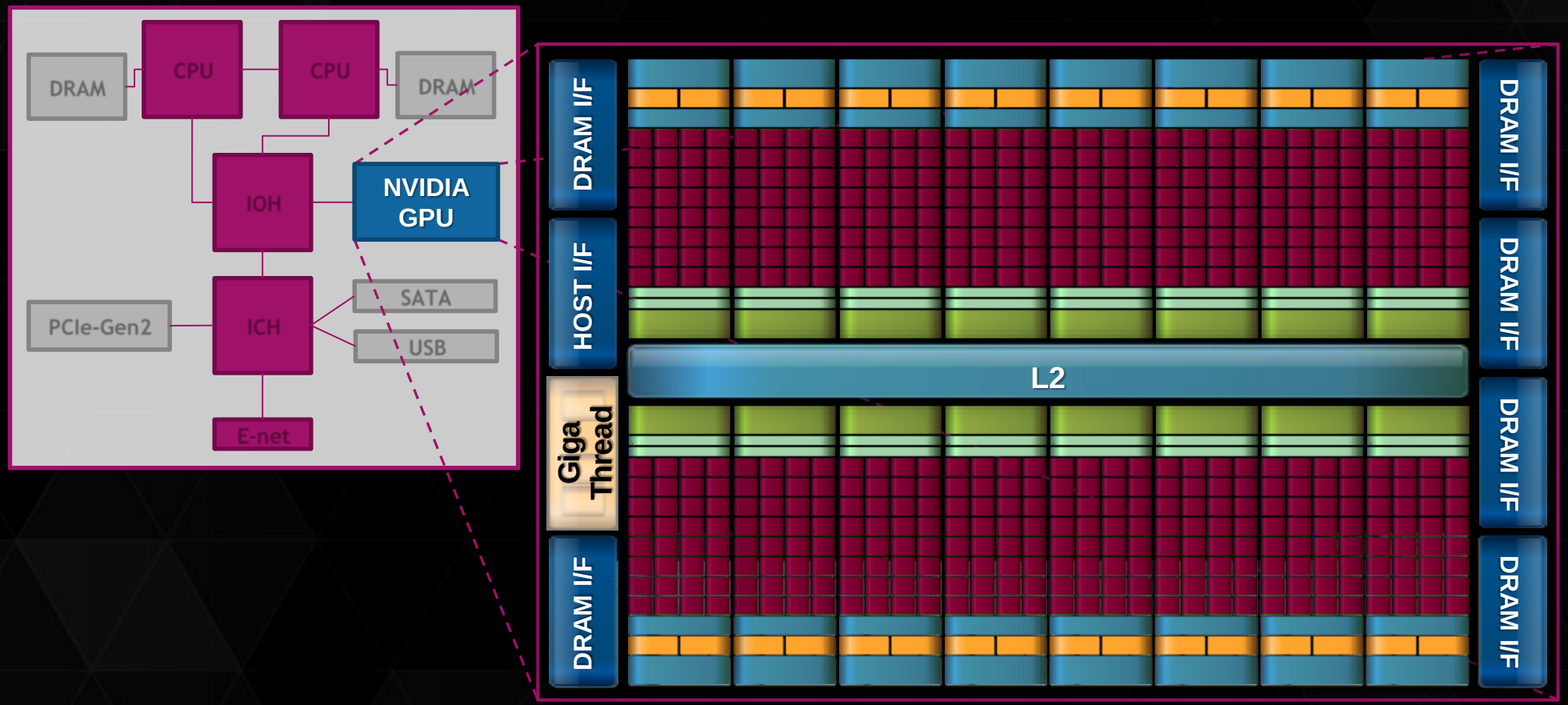
Quadro | Tesla | GRID

AUTO
Tegra

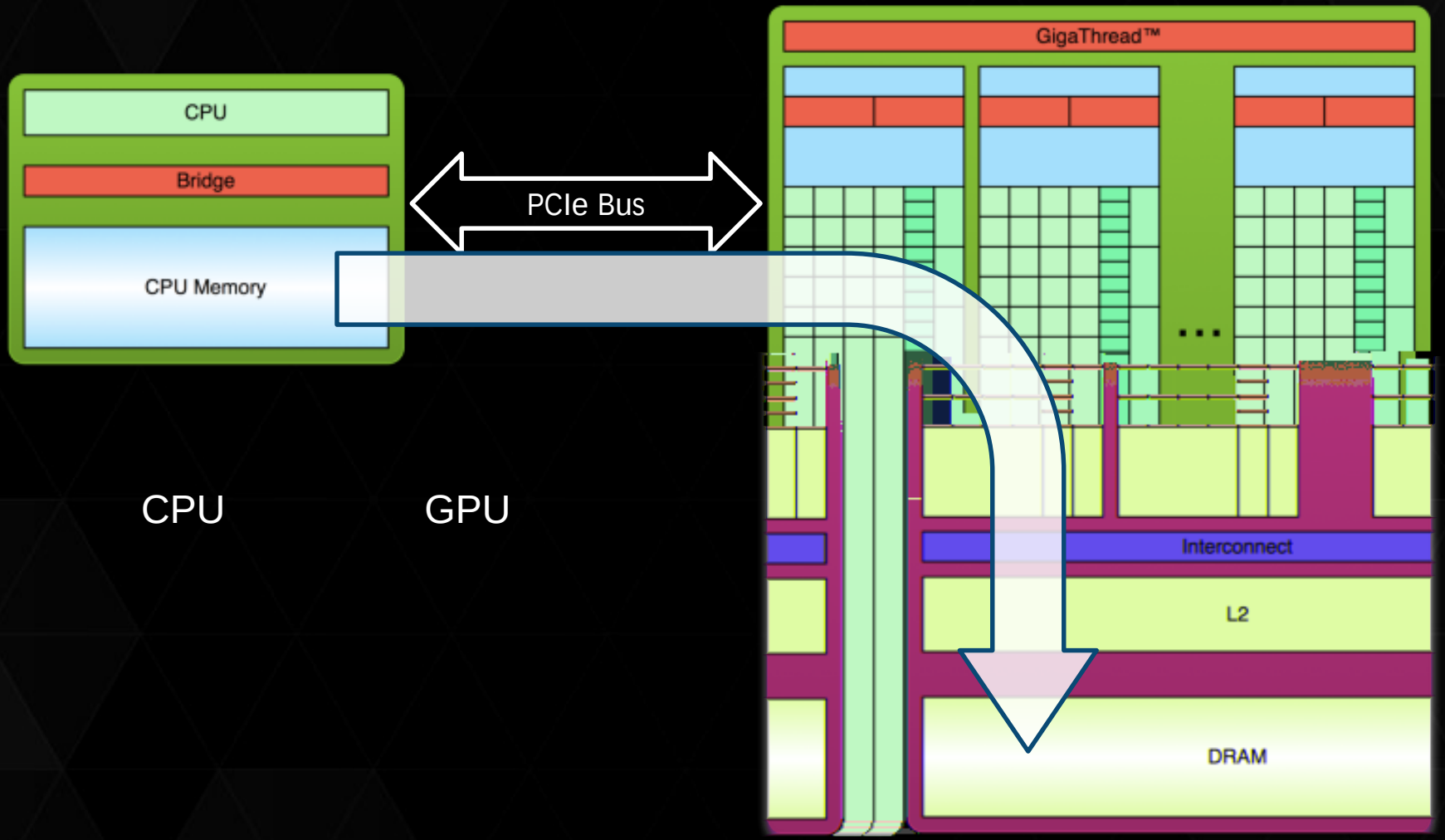
GPU

GPU

x86 + GPU

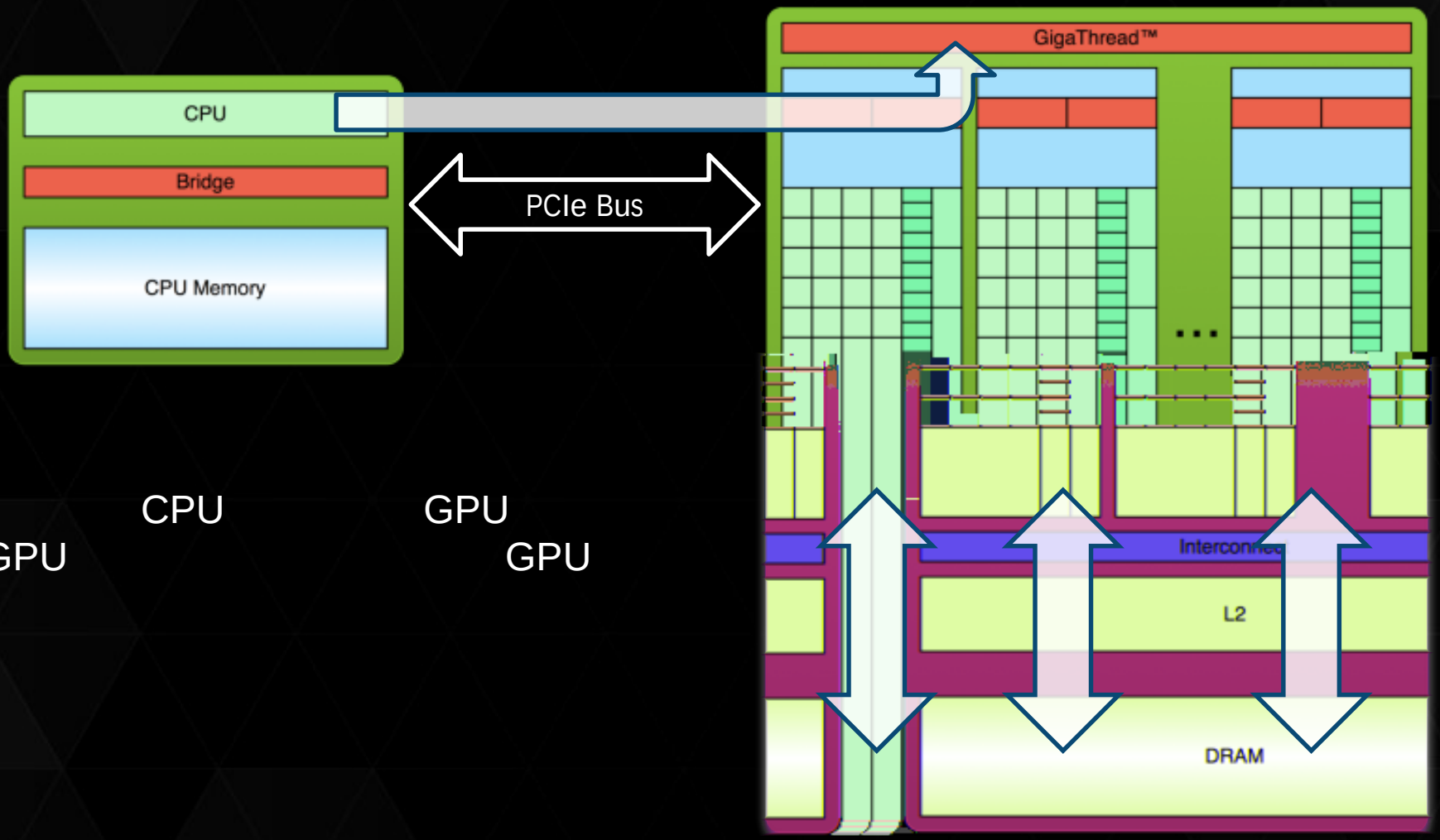


CPU + GPU



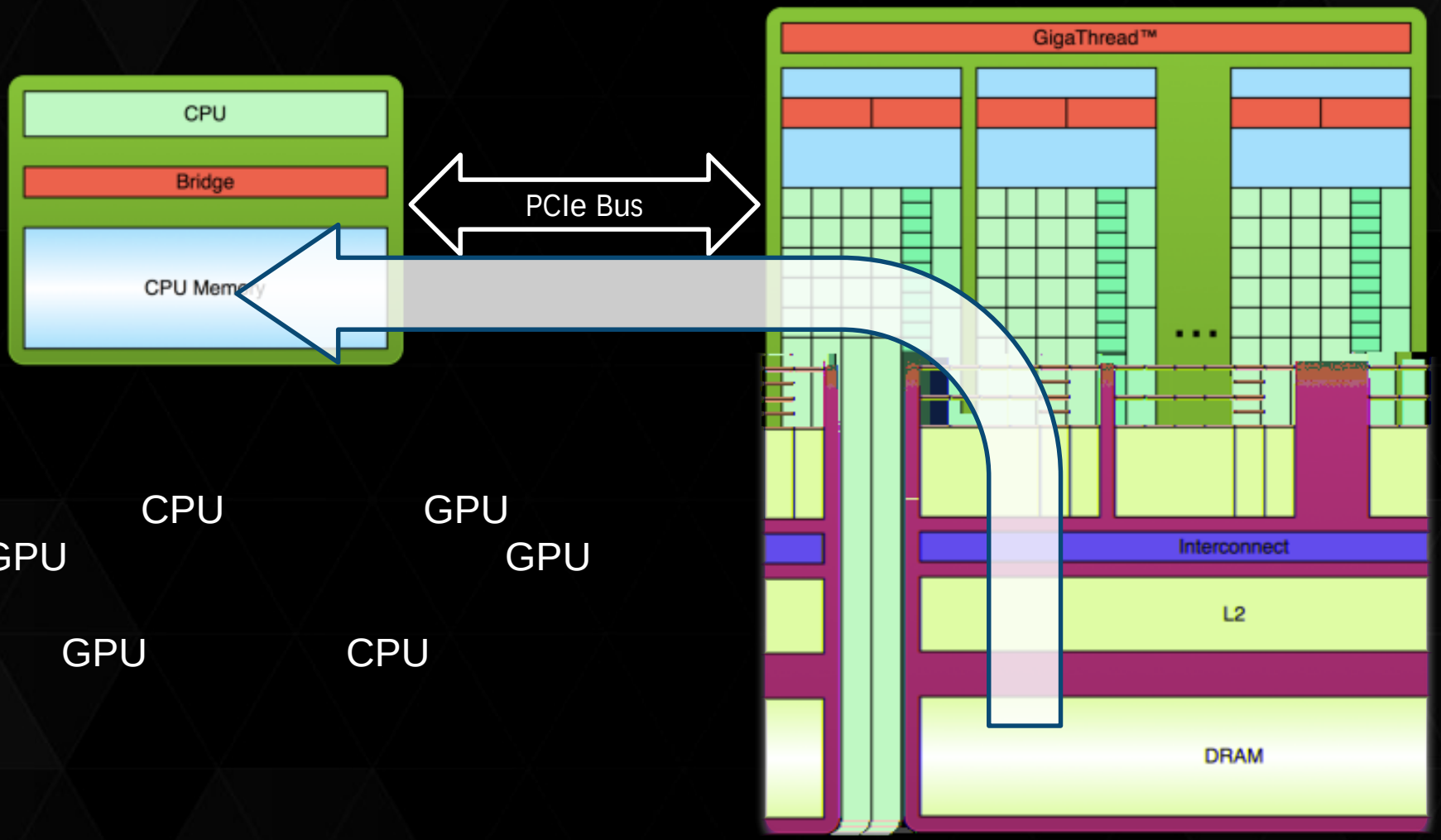
1. CPU GPU

CPU + GPU



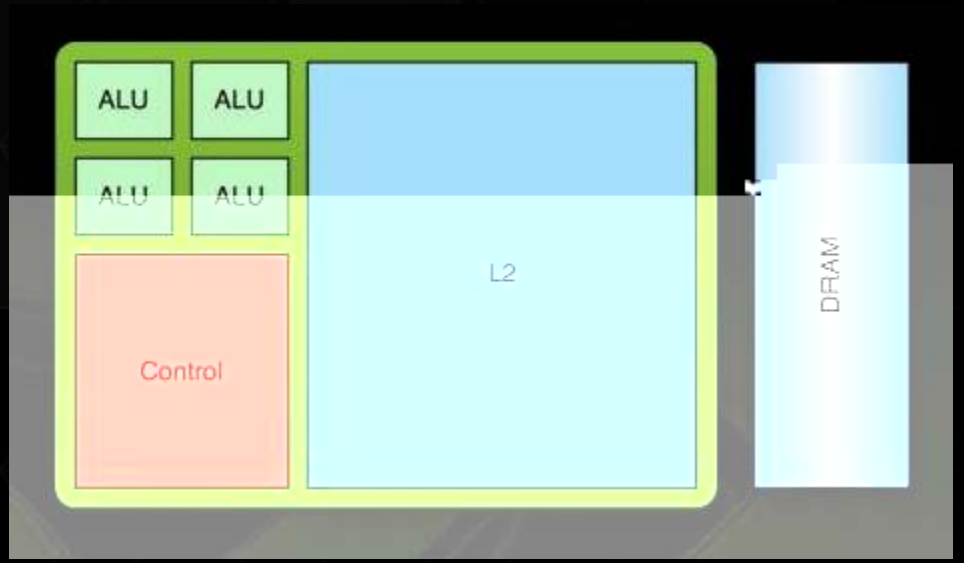
1. CPU GPU
2. GPU GPU

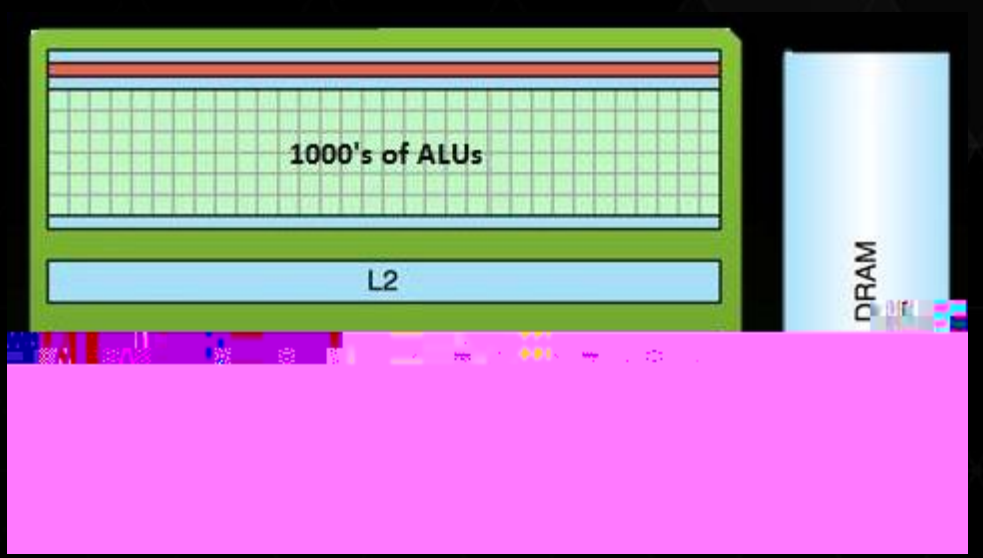
CPU + GPU



1. CPU GPU
2. GPU GPU
3. GPU CPU

OR





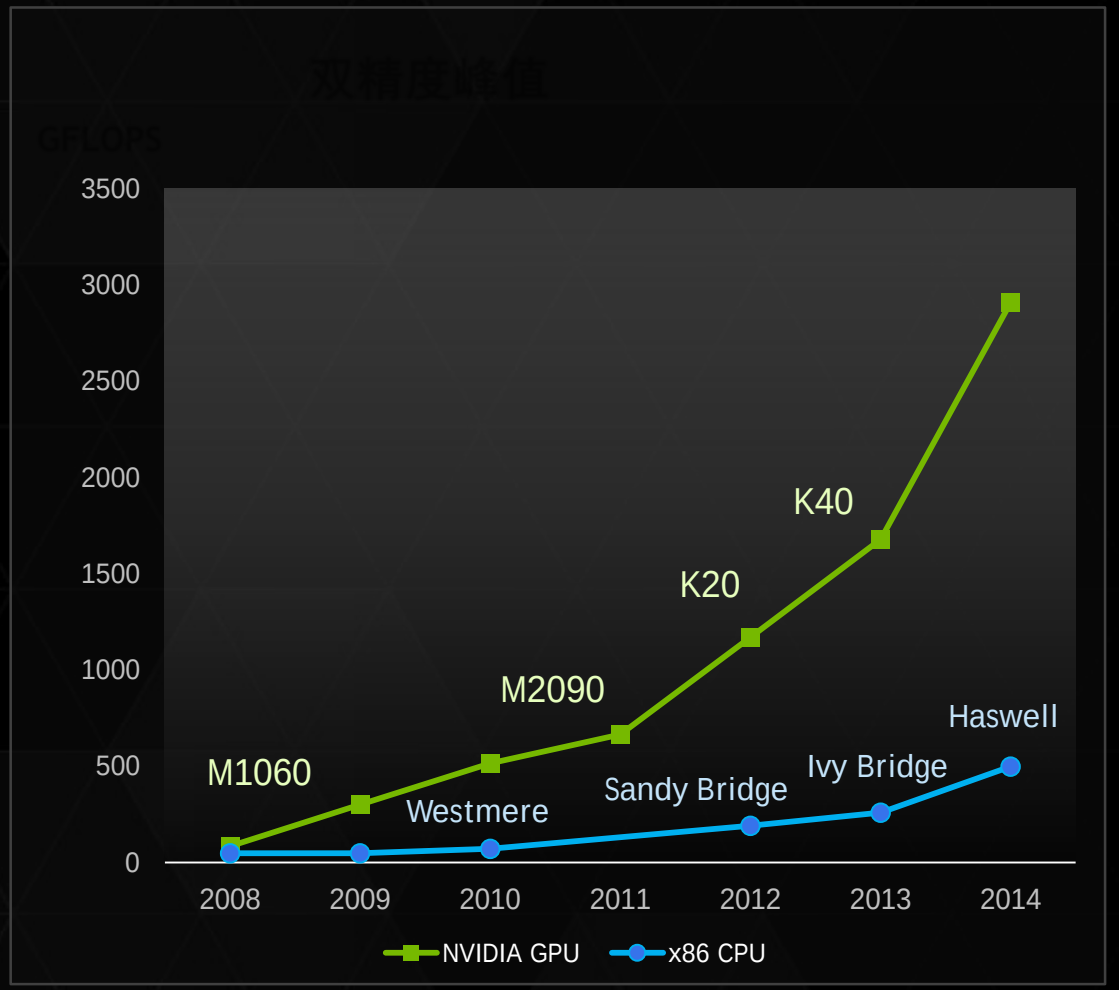
CPU

Delivery truck delivers one pizza and then moves to next house

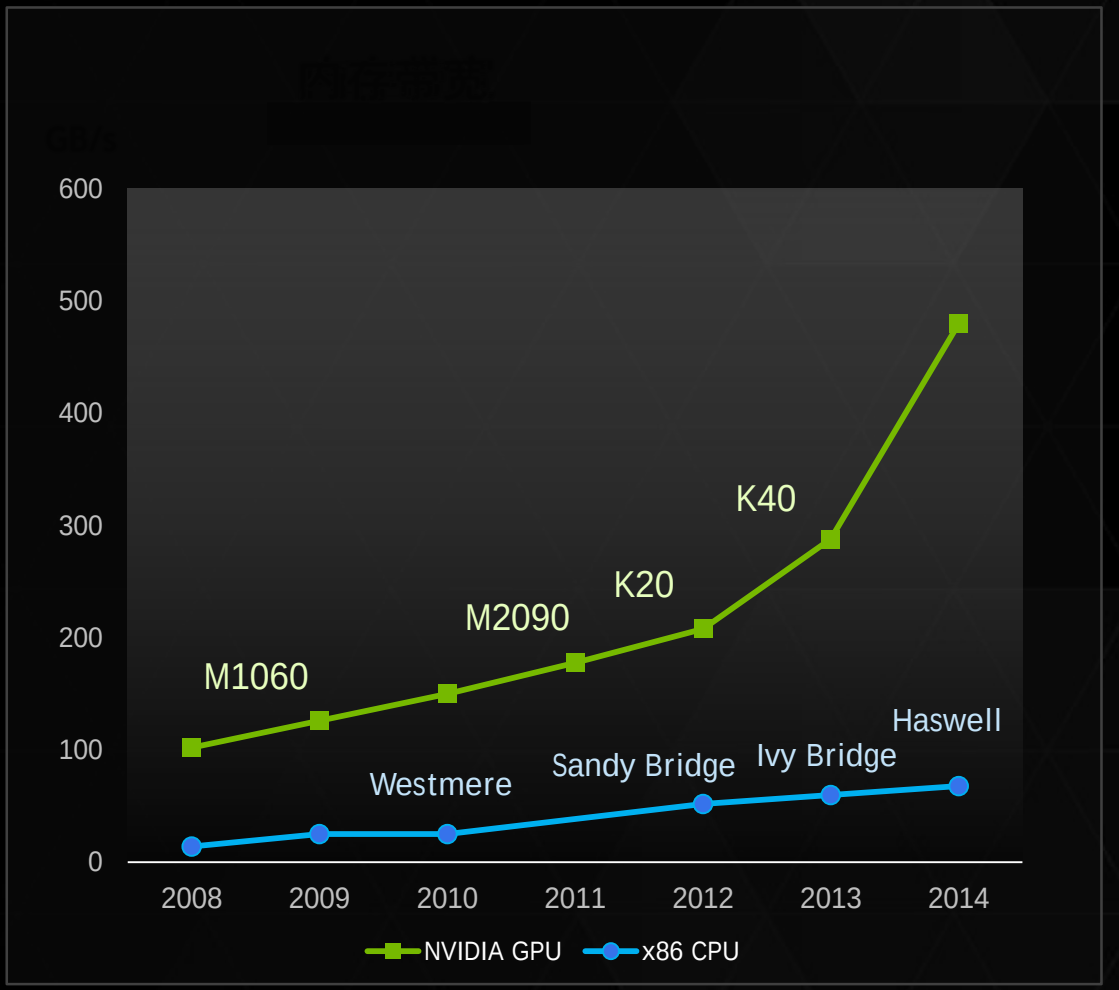
NVIDIA GPU

Many deliveries
to many houses

GPU

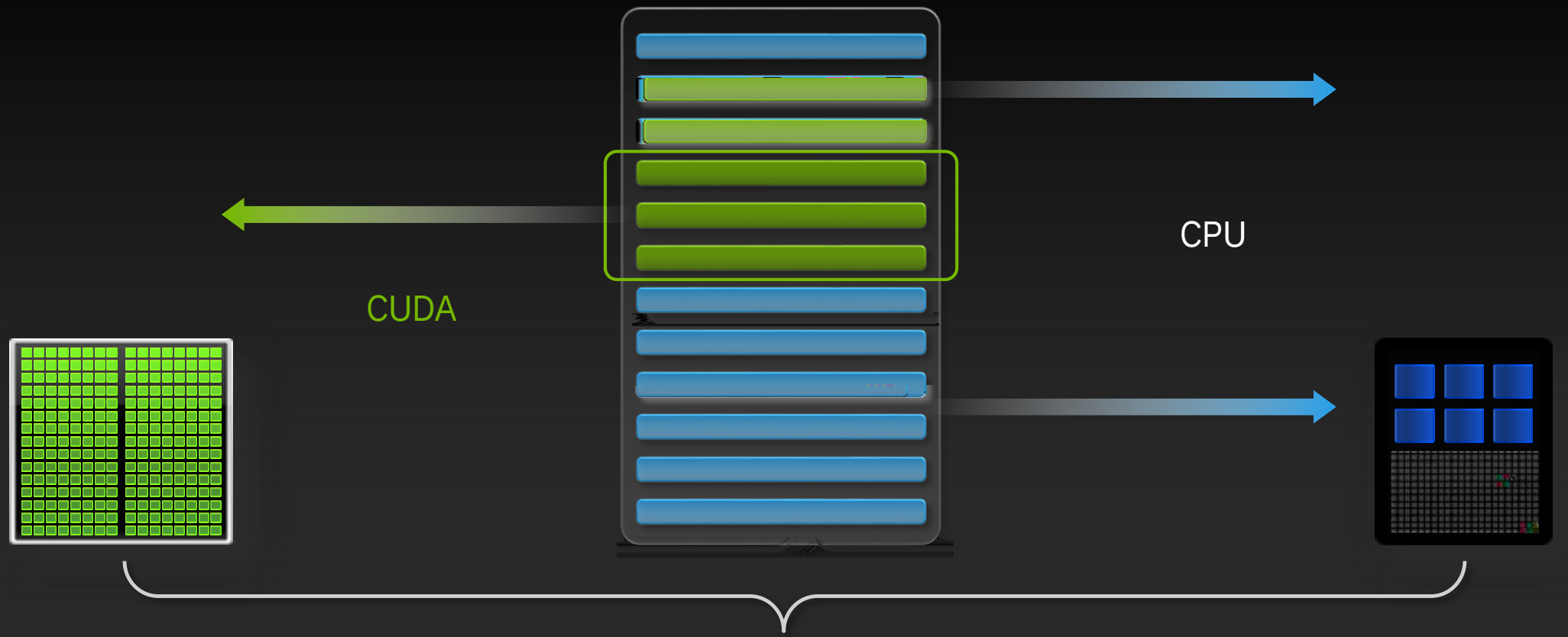


CPU

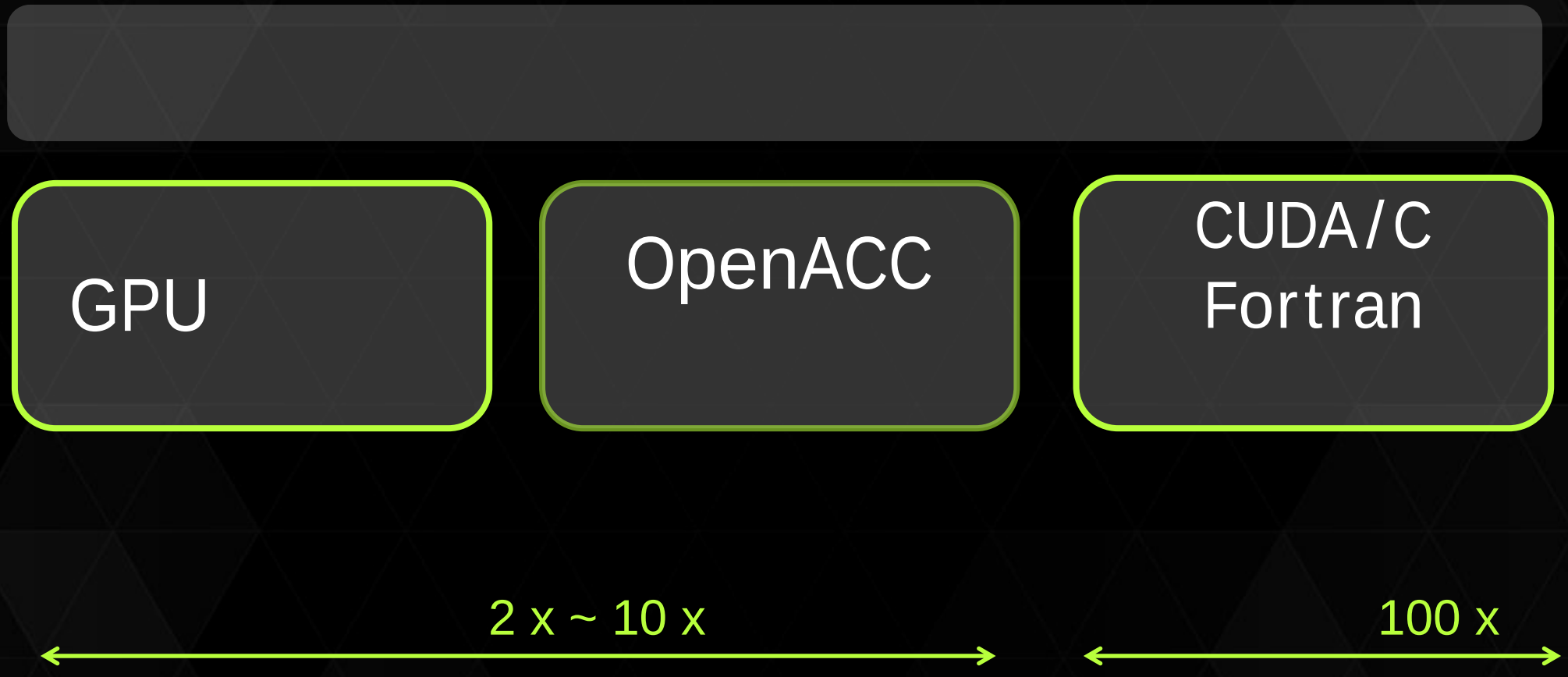


NVIDIA

CPU + GPU =



3



CPU + GPU

-
-
-
-

NVIDIA GPU

i □ □□□□
l □ □□□□

h PpgPS
h □□□□ □□□□
i dg□ s□□□
□Tr UP

U□ □□□□□□
& l □□□□□□□

□□□-□□
i s s□□□□ l □□□□□□
l □□□ □ i □□□□□
s □□□□ o□□□□□
P □□□□
p□□□ s□□□

b l r T□ □□□□

T
T++
a□□□□□□
k □□Tg
U□□□ □T□ □□□
e□□□
l □□□□□

l □□□ □ □□□□
T□ □□□□

l bdP □ □□□□□□
TPl o c h l l
Tr UP
k □□h l

g□ □□□□□

SgPo
aap
gPl PTf
i l l
s □□□□
d □□□□□
b l r g□

bl bl r T□□□□ □□□□□ & p□□□□□□

acceleware GASS STONE RIDGE
ANEO GPU Tech
CAPS
EM Photonics SCALABLE GRAPHICS
WIPRO

Microsoft

k Vh o□ □□□□ l □□□□□□□

DELL hp IBM
CRAY ASUS SUPERMICRO
sgi FUJITSU BULL
lenovo 联想 NEC
NVIDIA

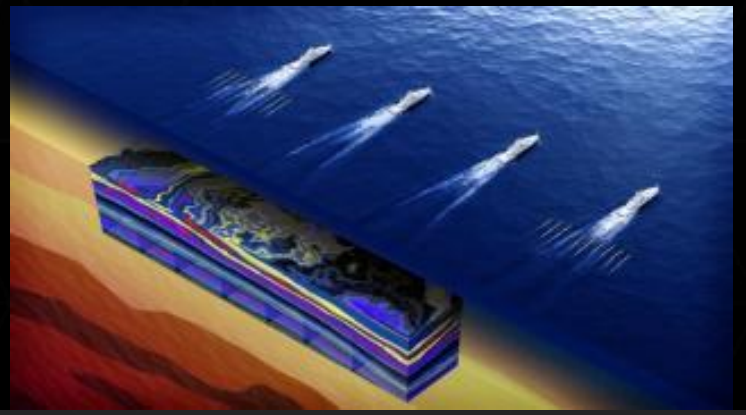
CUDA

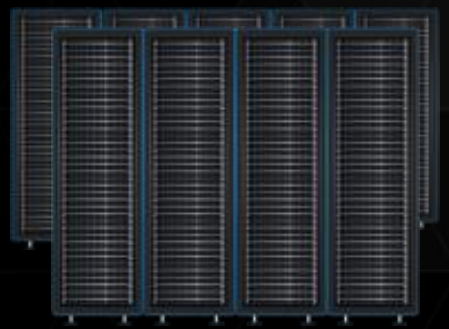
CUDA GPUs

CUDA

CUDA

GPU





4x-6x →

4x

12% /

RTM

NVIDIA

GPU

GPU

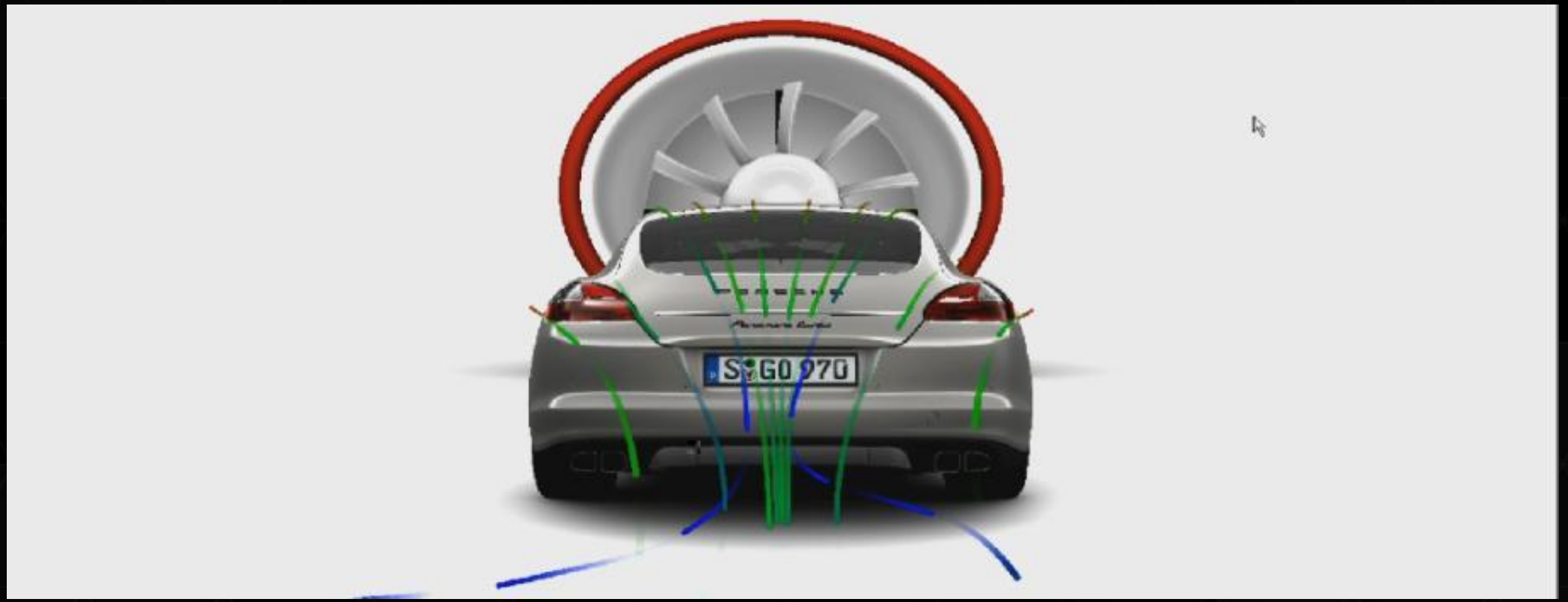
- CT
- X-RAY

SIEMENS medical

GE Healthcare

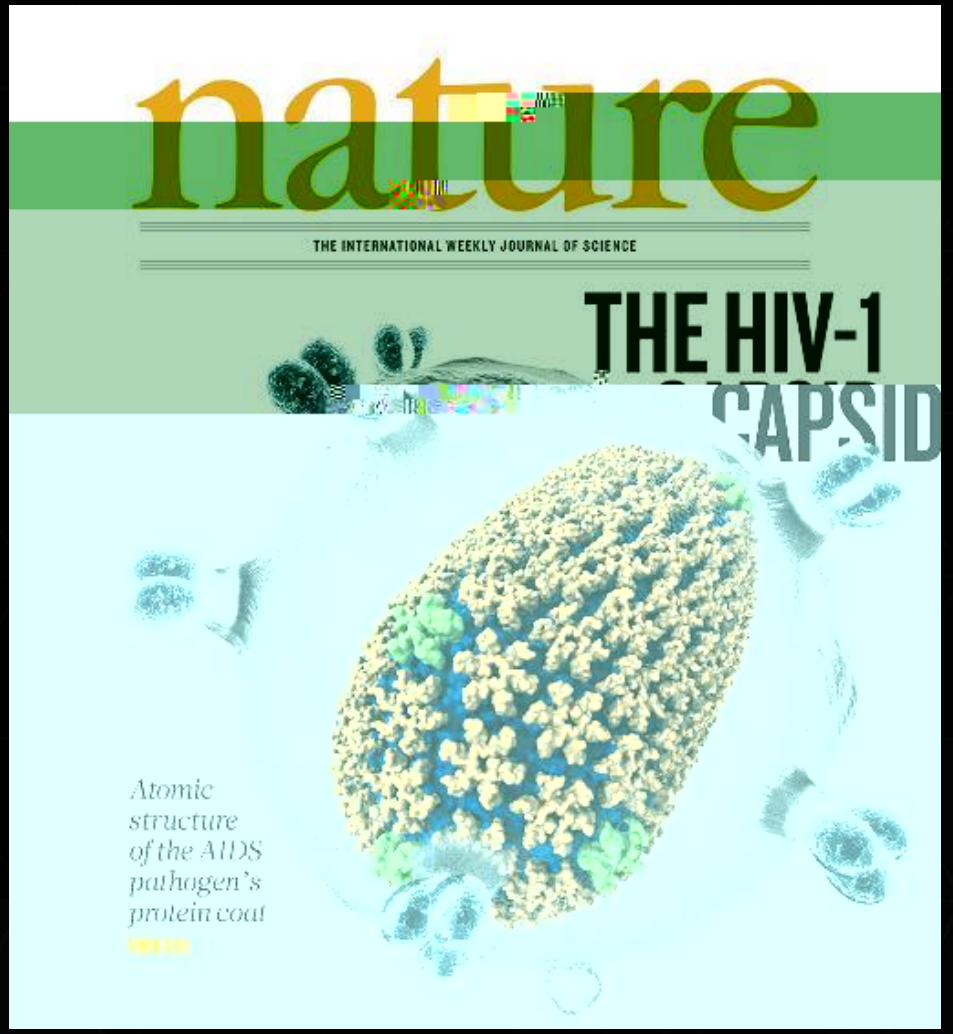
GPU

GPU



GPU

GPU /


nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
THE HIV-1
CAPSID
Atomic structure of the AIDS pathogen's protein coat

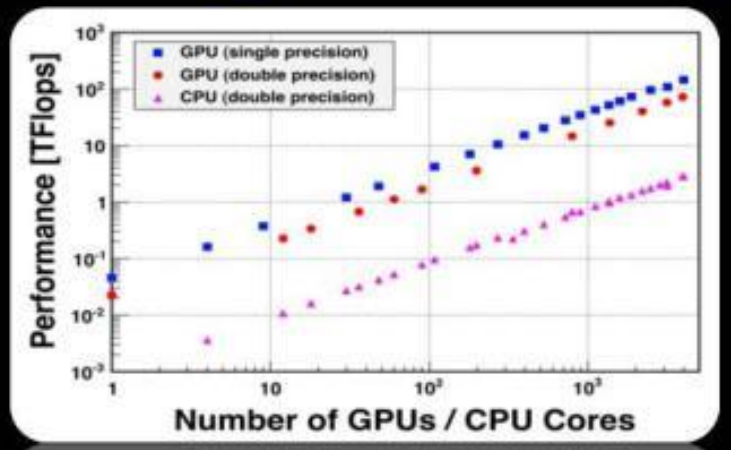
GPU

ASUCA TeraFlop Scaling (Weather Modeling)

3990 Tesla M2050s

145.0 Tflops SP

76.1 Tflops DP



Before GPUs

After GPUs

Simulation on Tsubame 2.0, TITech Supercomputer

NVIDIA

- 40x application performance increase
- 80% lower costs
- Calculation results in minutes vs. overnight

Understanding market risk is key to avoiding surprises that impact profitability. J.P. Morgan depends on Tesla GPUs to analyze risk faster than their competitors and to arrive at better decisions through more frequent, more complex calculations. GPUs are

JPMorgan

GPU

Typical scale of training data

Baidu 百度

Datasets

- Image recognition: 100 millions
- OCR: 100 millions
- Speech: 10 billions
- CTR: 100 billions

Training time:

Weeks to Months
on GPU clusters

Big data + Deep learning + HPC = Success

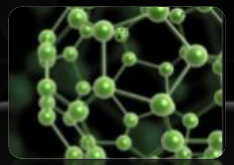
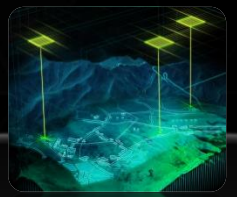
Projected training data to grow 10x each year

“

.”

NVIDIA

HPC


Schlumberger

BR PETROBRAS

Eni

Chevron

Statoil

HARVARD School of Engineering and Applied Sciences

STANFORD UNIVERSITY

Georgia Tech

ETH Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zurich

UNIVERSITY OF CAMBRIDGE

Air Force Research Laboratory

Raytheon

NASA

Naval Research Laboratory

CSCS

NCSA

Tokyo Institute of Technology

OAK RIDGE National Laboratory

Lawrence Livermore National Laboratory

J.P.Morgan

BARCLAYS

BNP PARIBAS

MUREX

Baidu 百度

salesforce SOFTWARE

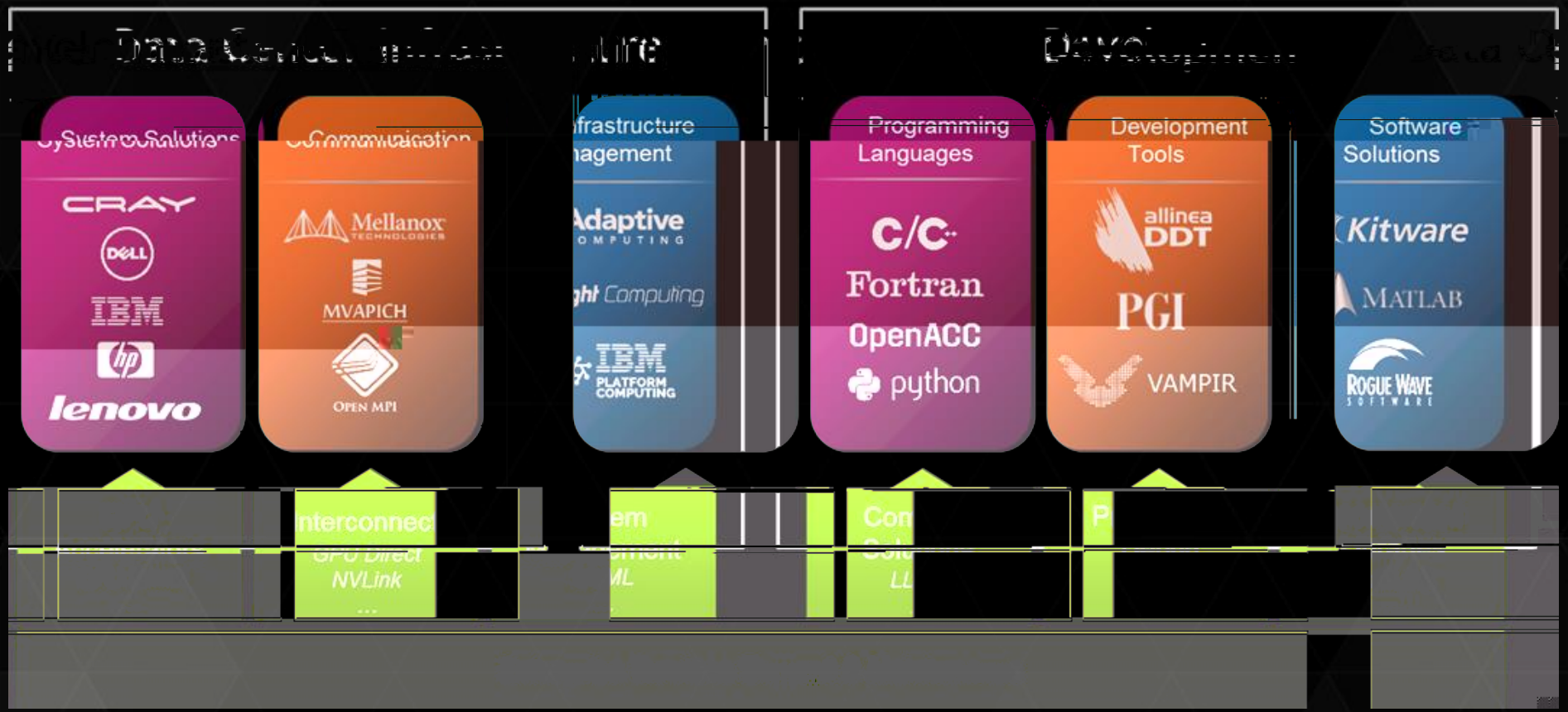
SHAZAM

amazon.com

Yandex

NVIDIA

Tesla



2014,

%

.

Source: Top 6 predictions for HPC in 2015 Intersect360 RESEARCH

NVIDIA

Libraries

AmgX

cuDNN

cuBLAS

OpenCV

Thrust

OpenACC

Programming Languages

C/C++

Fortran

python

Java

Power

ARM POWERED

x86

2015 Tesla GPU

Tesla K40

Best Single GPU Performance

Server, Workstation, Liquid Cooled

Higher Ed, Data Analytics, HPC Labs, Defense

Double Precision Workloads

Tesla K80

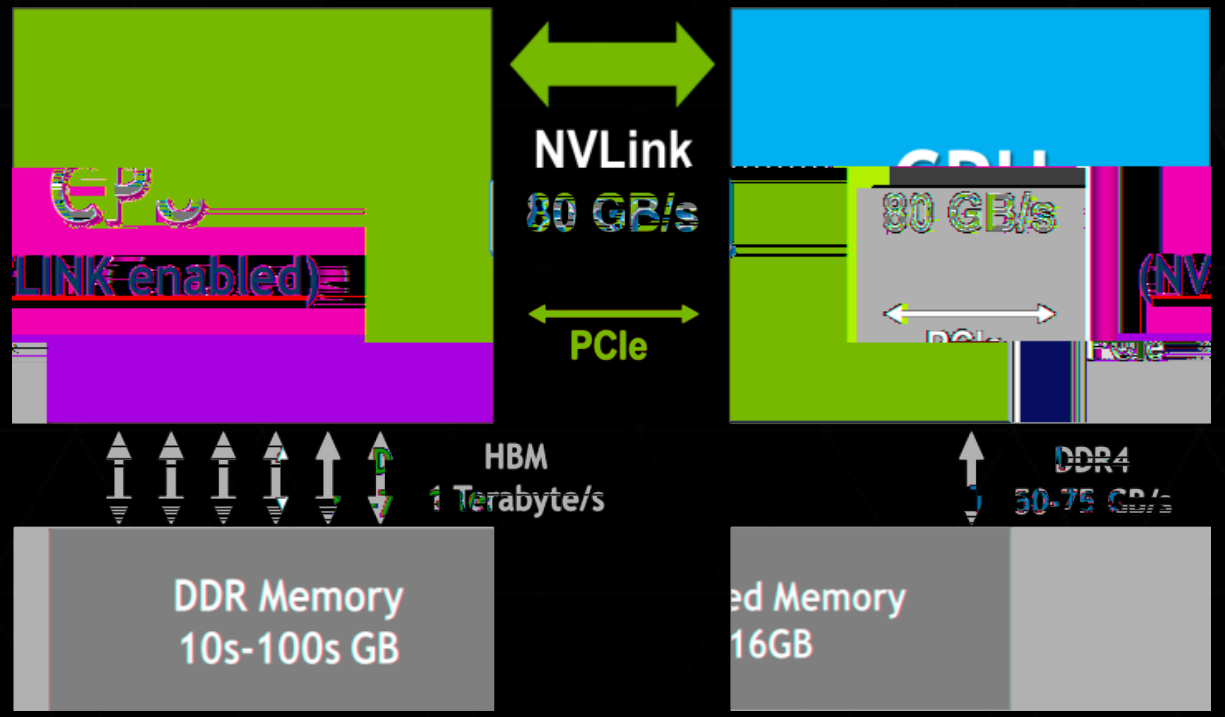
Maximize Throughput within a Server

Seismic, Data Analytics, HPC Labs, Defense

Multi-GPU Accelerated Apps

Single and Double Precision Workloads

NVLINK CPU GPU

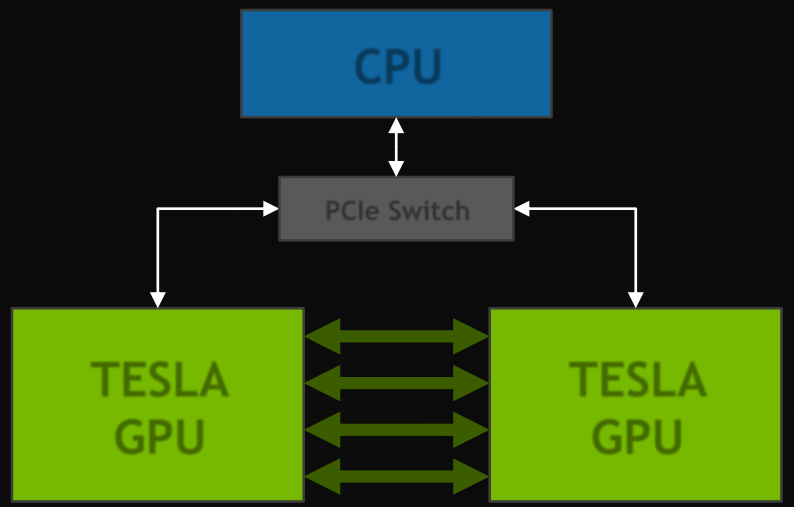


NVLINK 1.0

4 data links with 20GB/s each

X86 GPU NVLINK

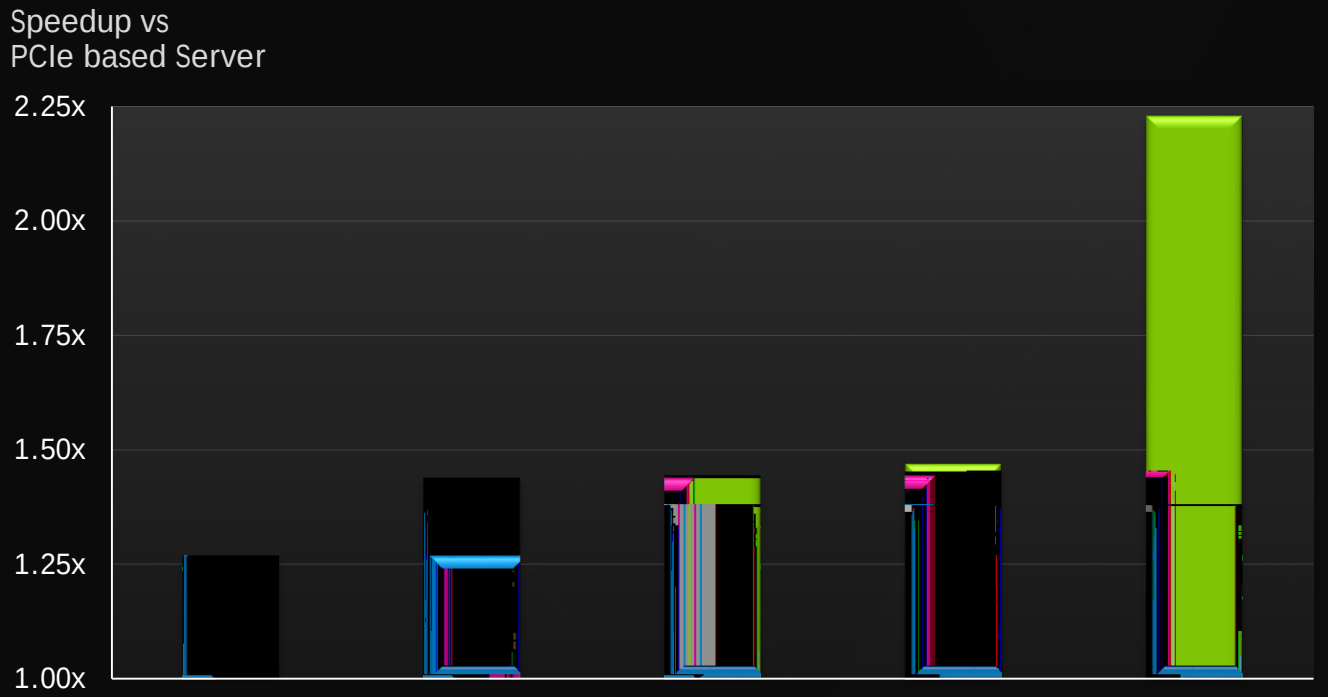
GPUs Interconnected with NVLink



5x Faster than
PCIe Gen3 x16

Over 2x Application Performance Speedup

When Next-Gen GPUs Connect via NVLink Versus PCIe



To learn more: http://www.nvidia.com/object/nvlink.html

US to Build Two Flagship Supercomputers

Powered by the Tesla Platform

OAK RIDGE National Laboratory

Lawrence Livermore National Laboratory

IBM

NVIDIA

100-300 PFLOPS Peak

10x in Scientific App Performance

IBM POWER9 CPU + NVIDIA Volta GPU

NVLink High Speed Interconnect

40 TFLOPS per Node, >3,400 Nodes

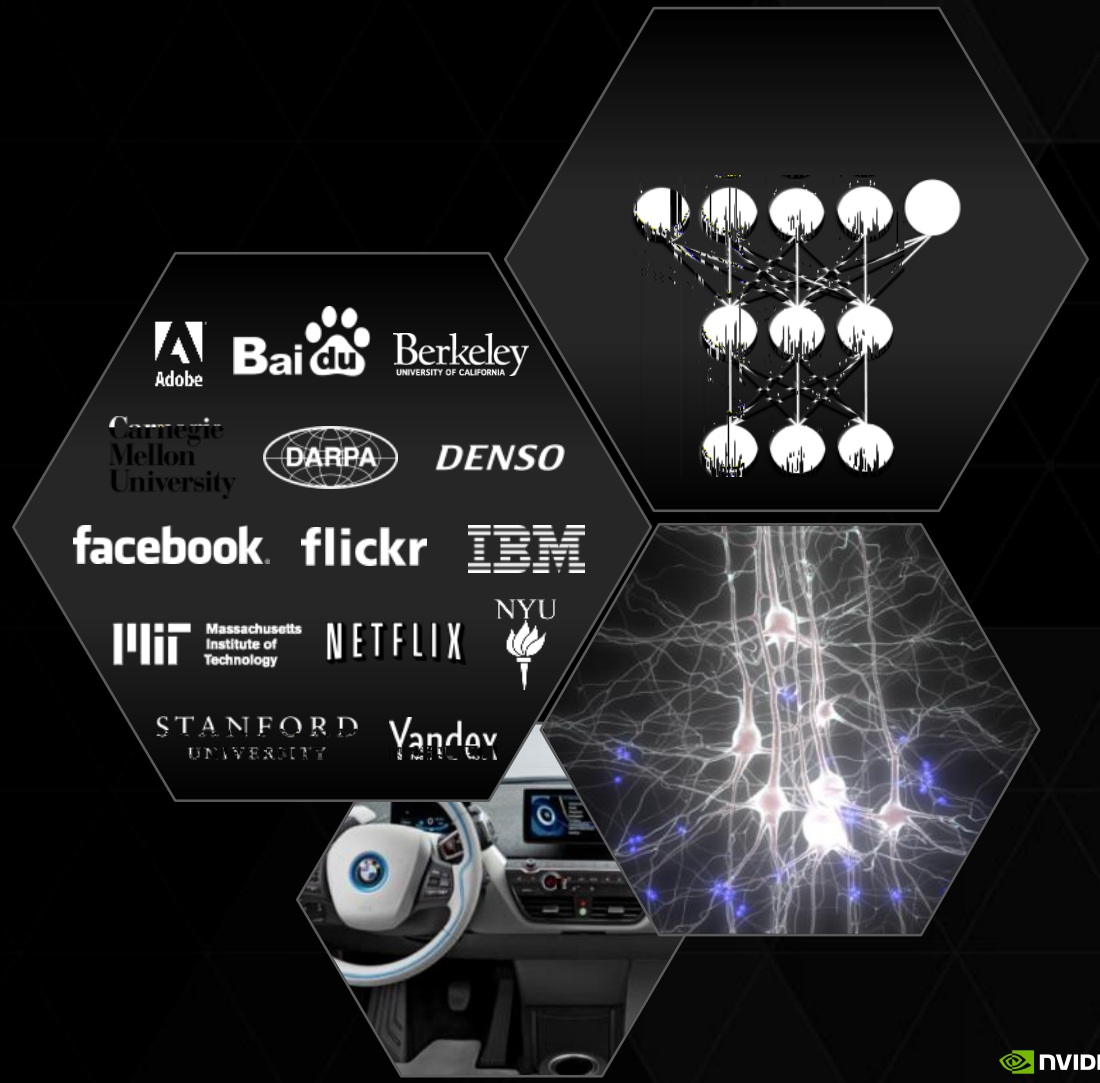
2017

Major Step Forward on the Path to Exascale

"Machine Learning" is in some sense a rebranding of AI.

The focus is now on more specific, often perceptual tasks, and there are many successes.

Today, some of the world's largest internet companies, as well as the foremost research institutions, are using GPUs for machine learning.

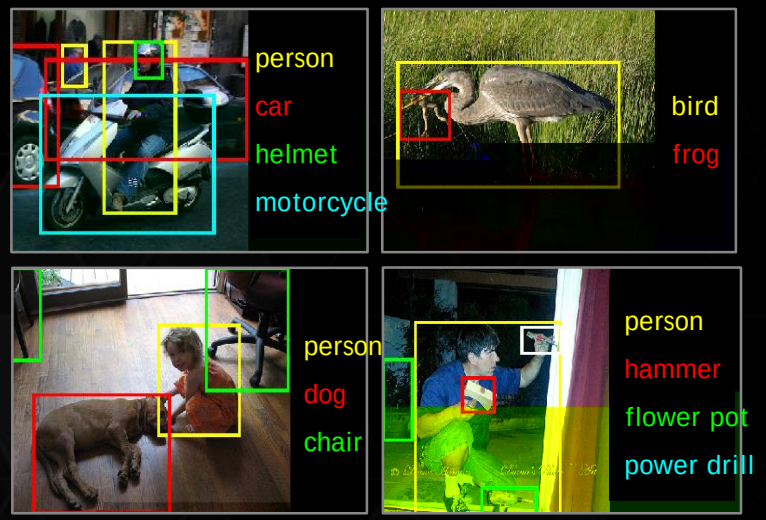


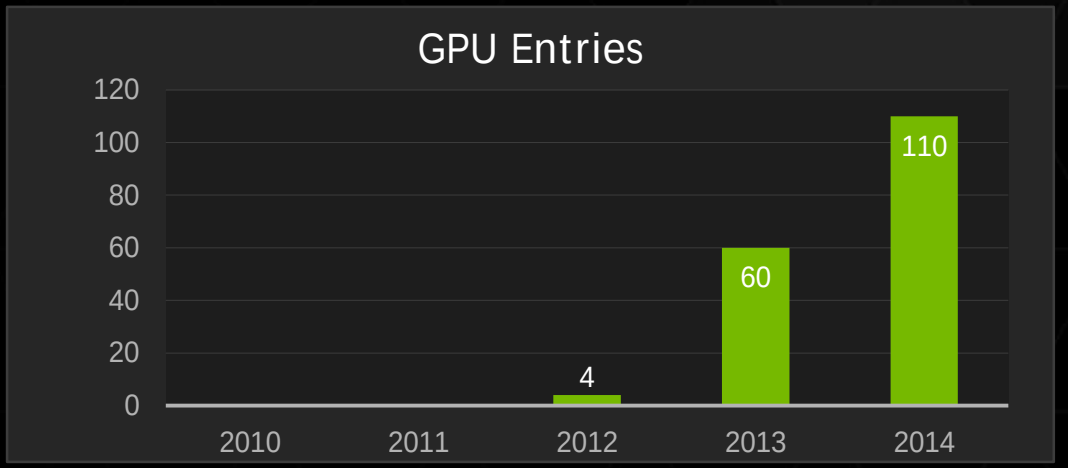
GPU –

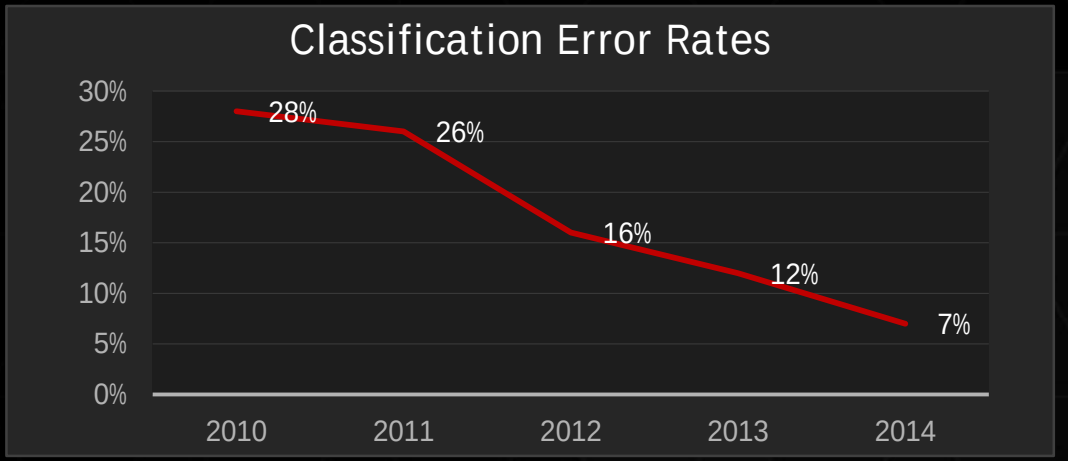
1.2 *1000*

Hosted by

IMAGENET





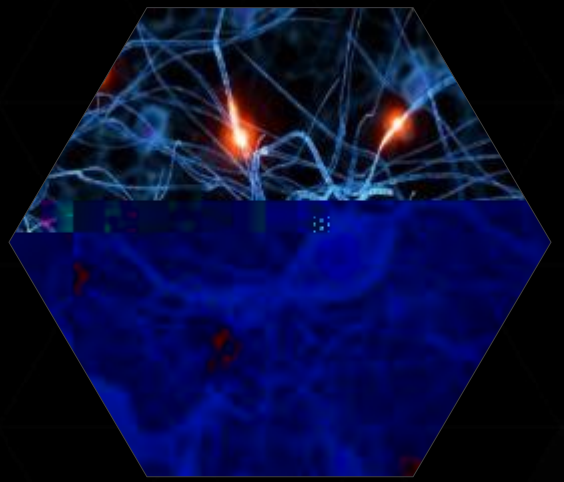


NVIDIA

3

大数据

模型



GPU加速器

GPU

Early Adopters

Adobe
Image Analytics for Creative Cloud

Baidu
Speech/Image Recognition

flickr
Image Classification

IBM
Hadoop

NETFLIX
Recommendation

Yandex
Search Rankings

Use Cases

&

&

&

Talks @ GTC

facebook

NYU

STANFORD UNIVERSITY

DARPA

DENSO

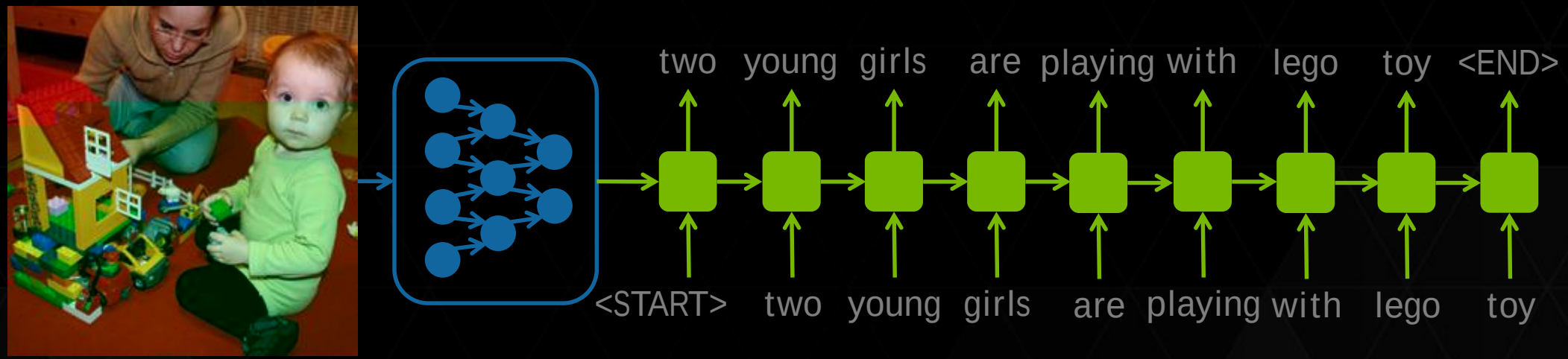
Carnegie Mellon University

MIT Massachusetts Institute of Technology

Berkeley UNIVERSITY OF CALIFORNIA

ConvNets + RNN/LSTM

- Multiple papers (Stanford, Google, UC Berkley, and others) in Nov 14



http://cs.stanford.edu/people/karpathy/deepimagesent/ http://googleresearch.blogspot.co.uk/2014/11/a-picture-is-worth-thousand-coherent.html
http://arxiv-web3.library.cornell.edu/pdf/1411.4389v1.pdf

Car and Lane detection via CNN from Baidu, Stanford, Twitter, TI

Fig. 1: Sample output from our neural network capable of lane and vehicle detection.

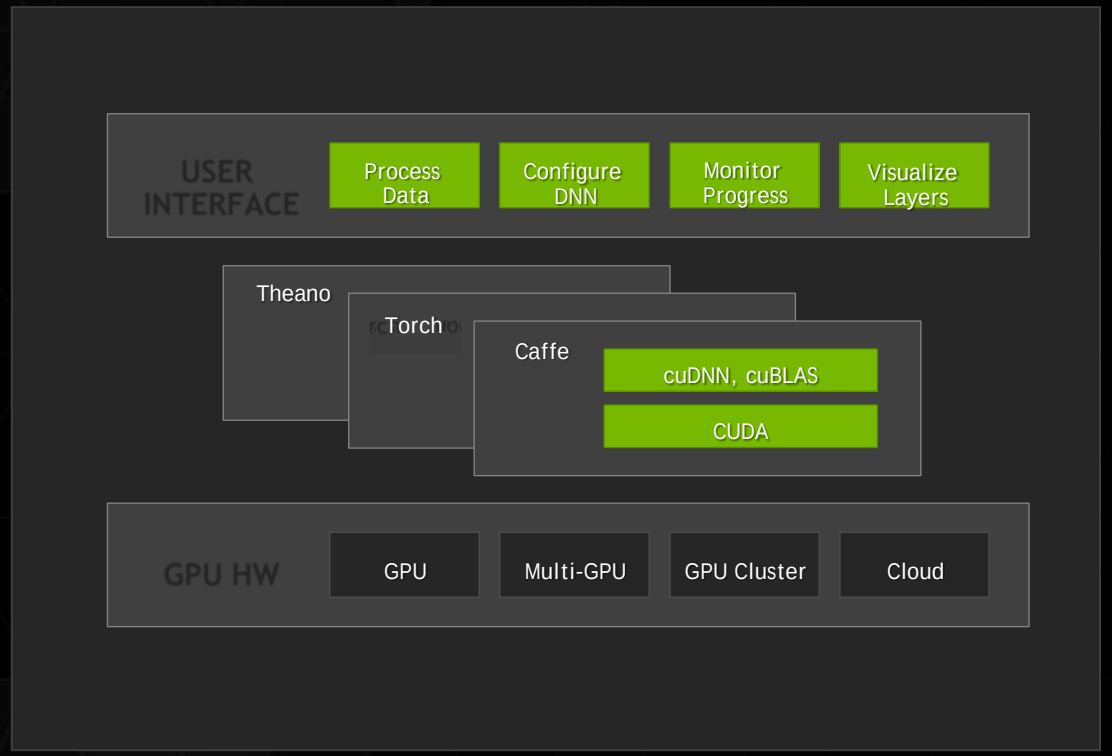
50+
Deep Learning Sessions

Adobe	Google
Alibaba	iFlytek, Ltd
Baidu	NUANCE
Carnegie Mellon	Stanford Univ
Facebook	UC Berkeley
Flickr / Yahoo	Univ of Toronto

Developer Labs

Caffe
Torch
Theano

DiGITS – GPU



DEEP GPU TRAINING SYSTEM FOR DATA SCIENTISTS

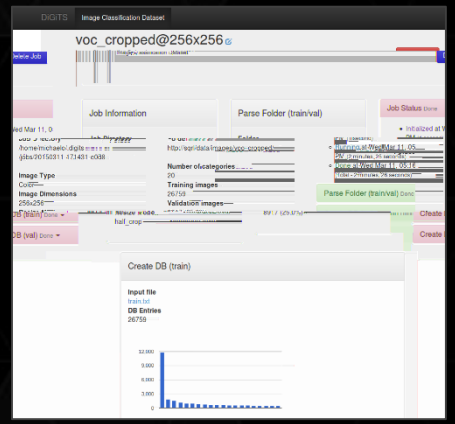
Design DNNs

Visualize activations

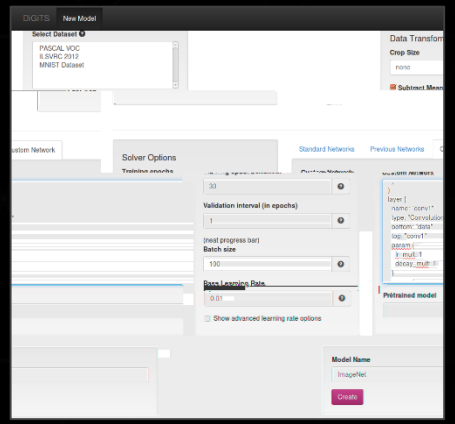
Manage multiple trainings

DIGITS

Process Data

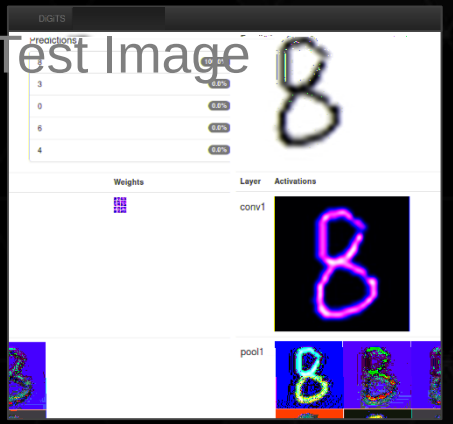


Configure DNN



Monitor Progress

Visualize Layers



NVIDIA

HLUO@NVIDIA.COM