NVIDIA GPU 加速计算和深度学习

长春，17/4/2015

视觉计算的世界领导者

GAMING
GeForce | GRID

ENTERPRISE
Quadro | Tesla | GRID

AUTO
Tegra

目标市场和产品品牌

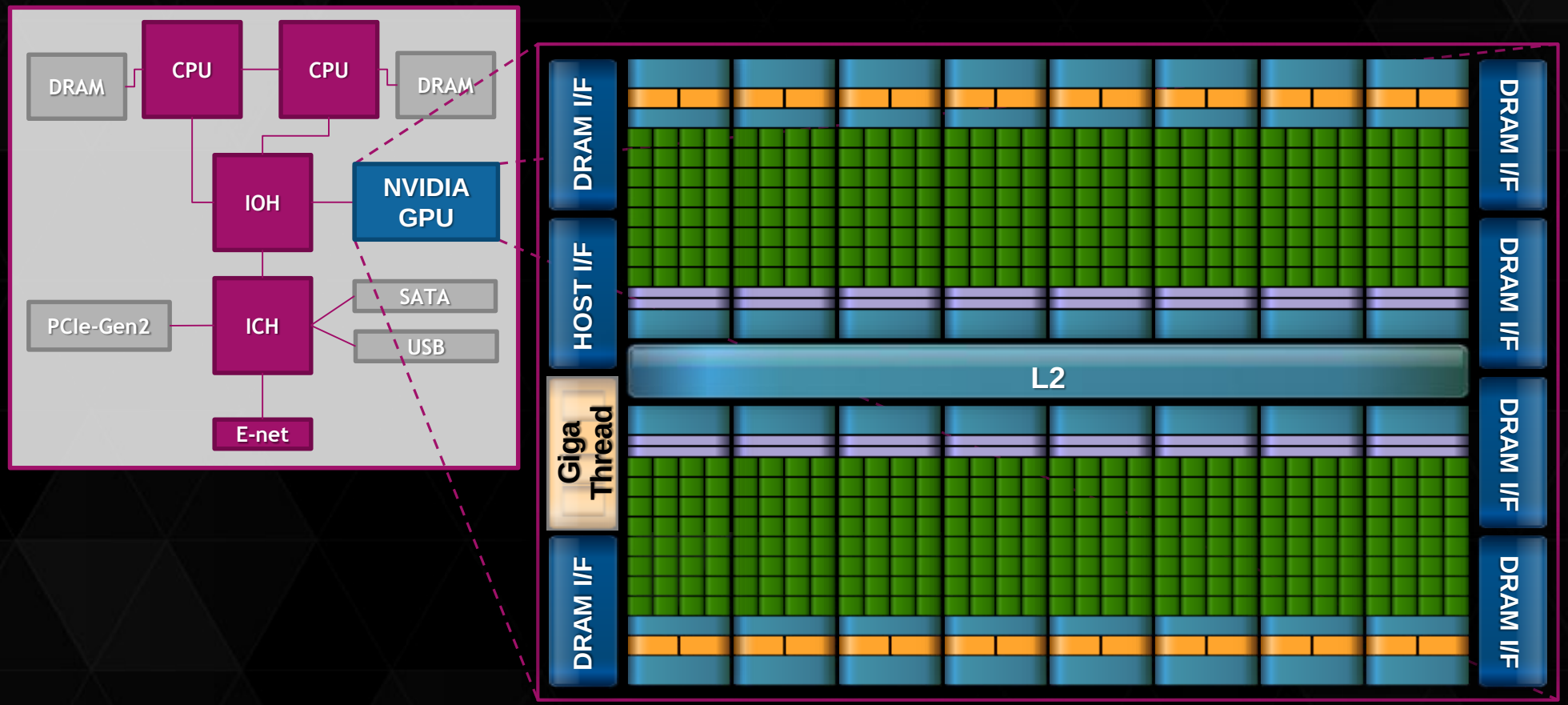
内容

GPU 计算

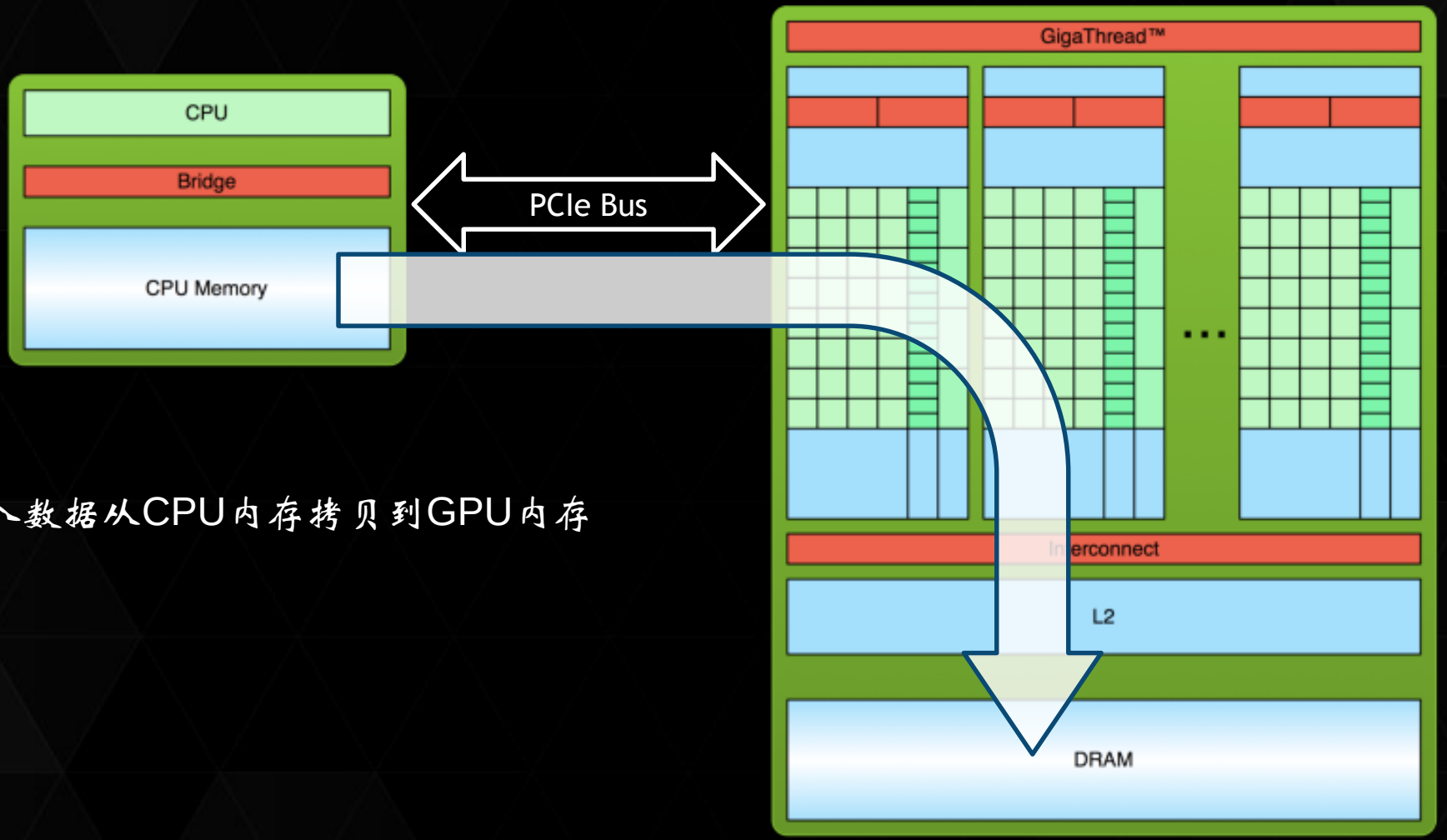
深度学习

什么是 GPU 计算

x86 + GPU 异构计算

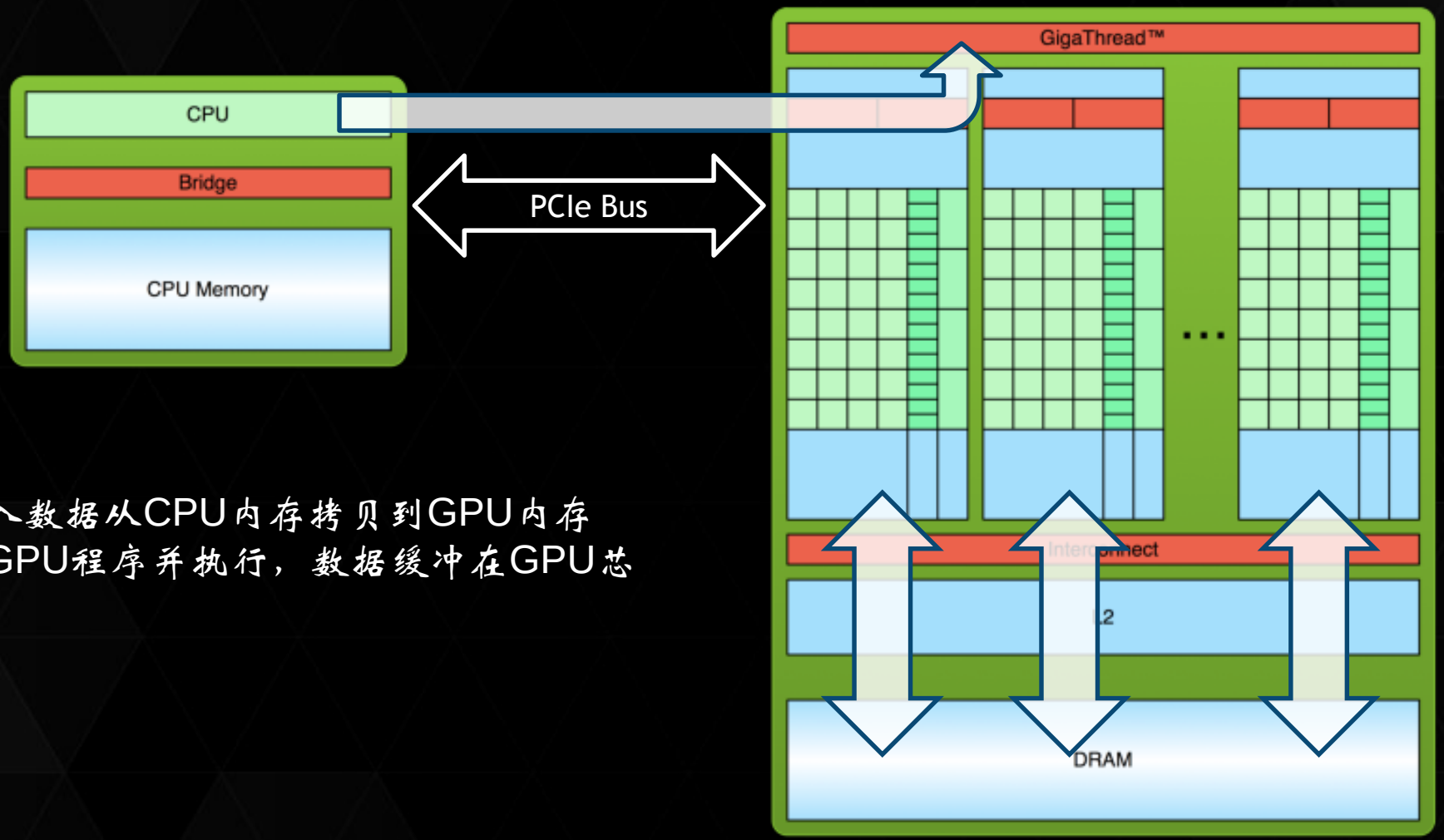


CPU + GPU 处理流程



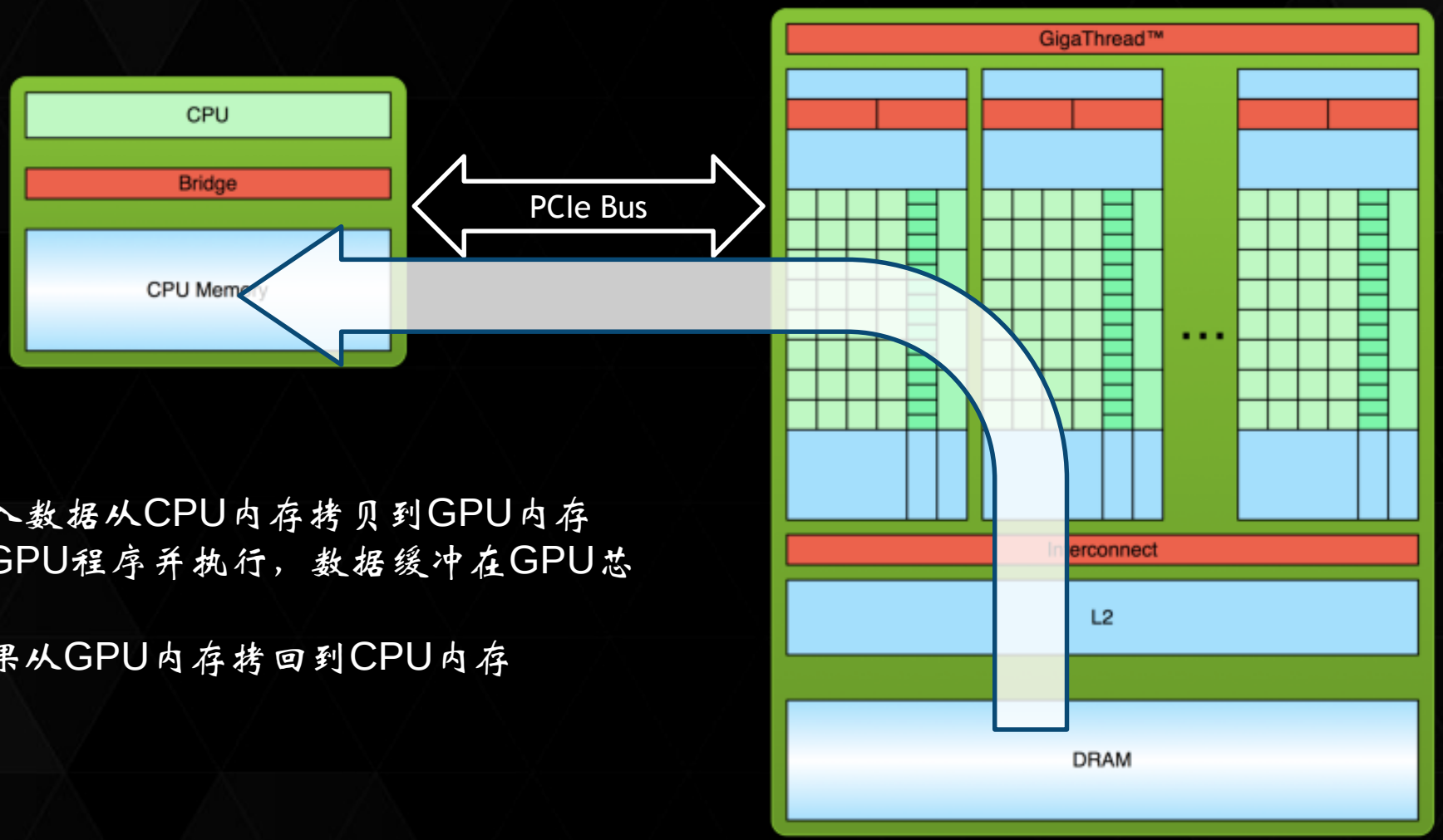
1. 把输入数据从CPU内存拷贝到GPU内存

CPU + GPU处理流程



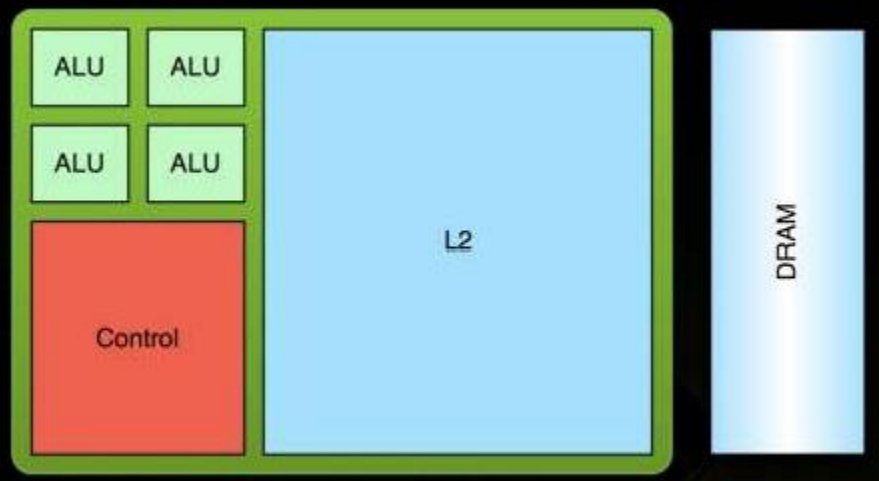
1. 把输入数据从CPU内存拷贝到GPU内存
2. 载入GPU程序并执行，数据缓冲在GPU芯片上

CPU + GPU处理流程



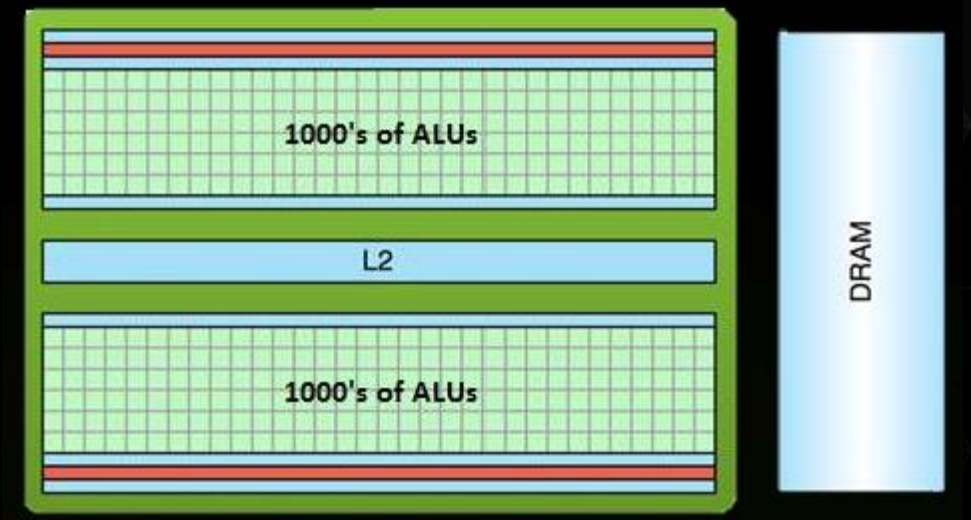
1. 把输入数据从CPU内存拷贝到GPU内存
2. 载入GPU程序并执行，数据缓冲在GPU芯片上
3. 把结果从GPU内存拷回到CPU内存

低延迟 OR 高吞吐量？



CPU

- Optimized for low-latency access to cached data sets
- Control logic for out-of-order and speculative execution

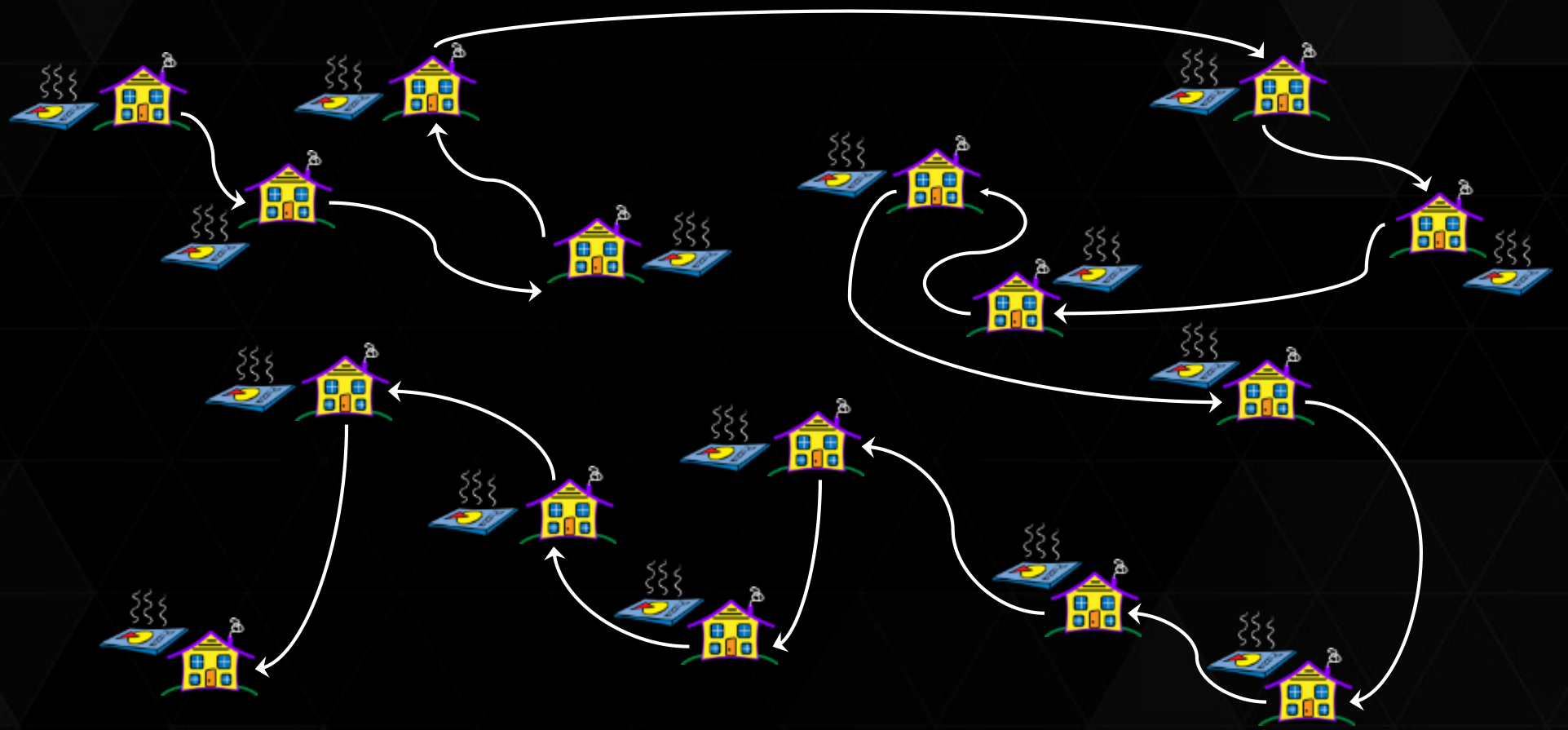


GPU

- Optimized for data-parallel, throughput computation
- Architecture tolerant of memory latency
- More transistors dedicated to computation

CPU 送外卖

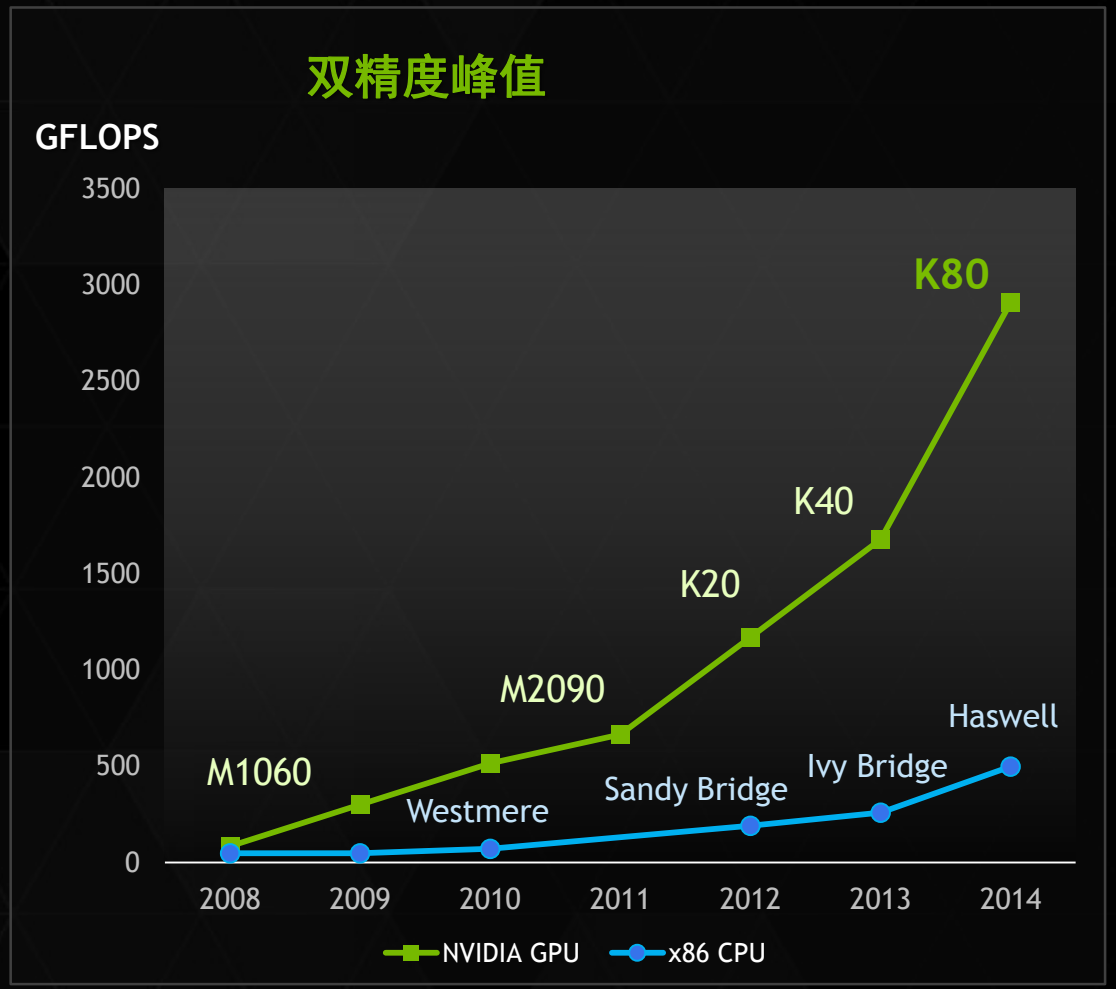
PROCESS
Delivery truck delivers one pizza and then moves to next house

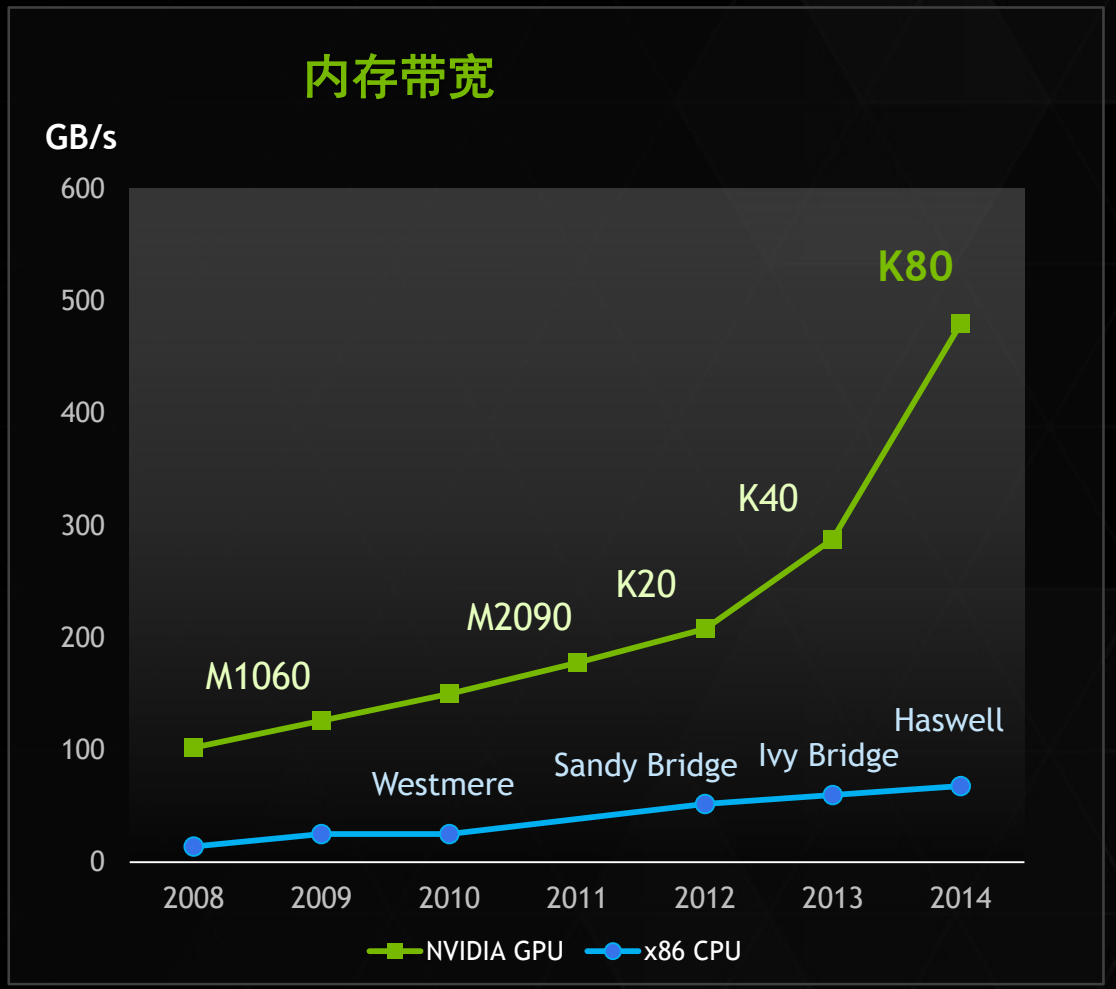


NVIDIA GPU 送外卖

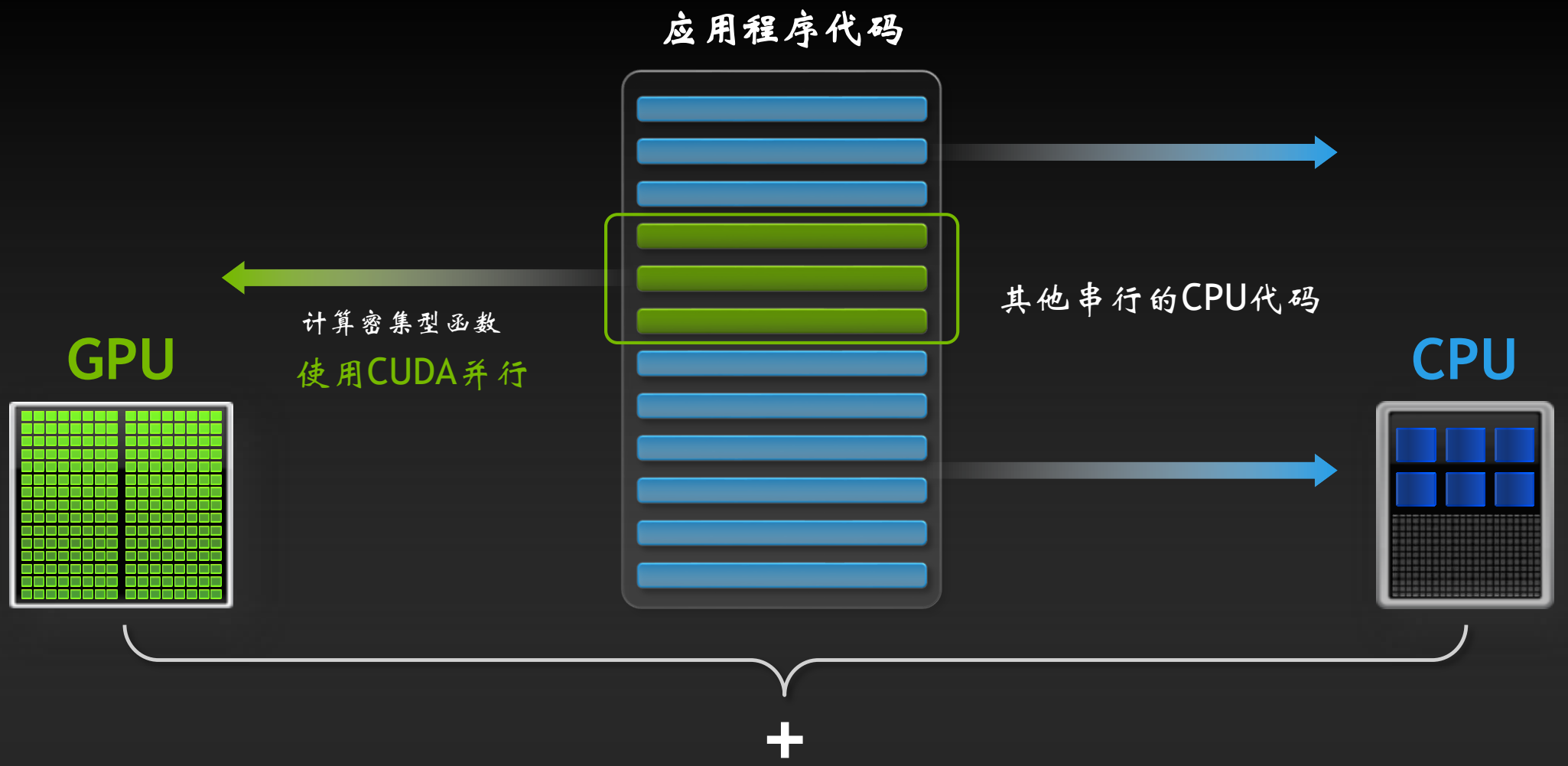
PROCESS
Many deliveries
to many houses

GPU 的性能大大领先于CPU

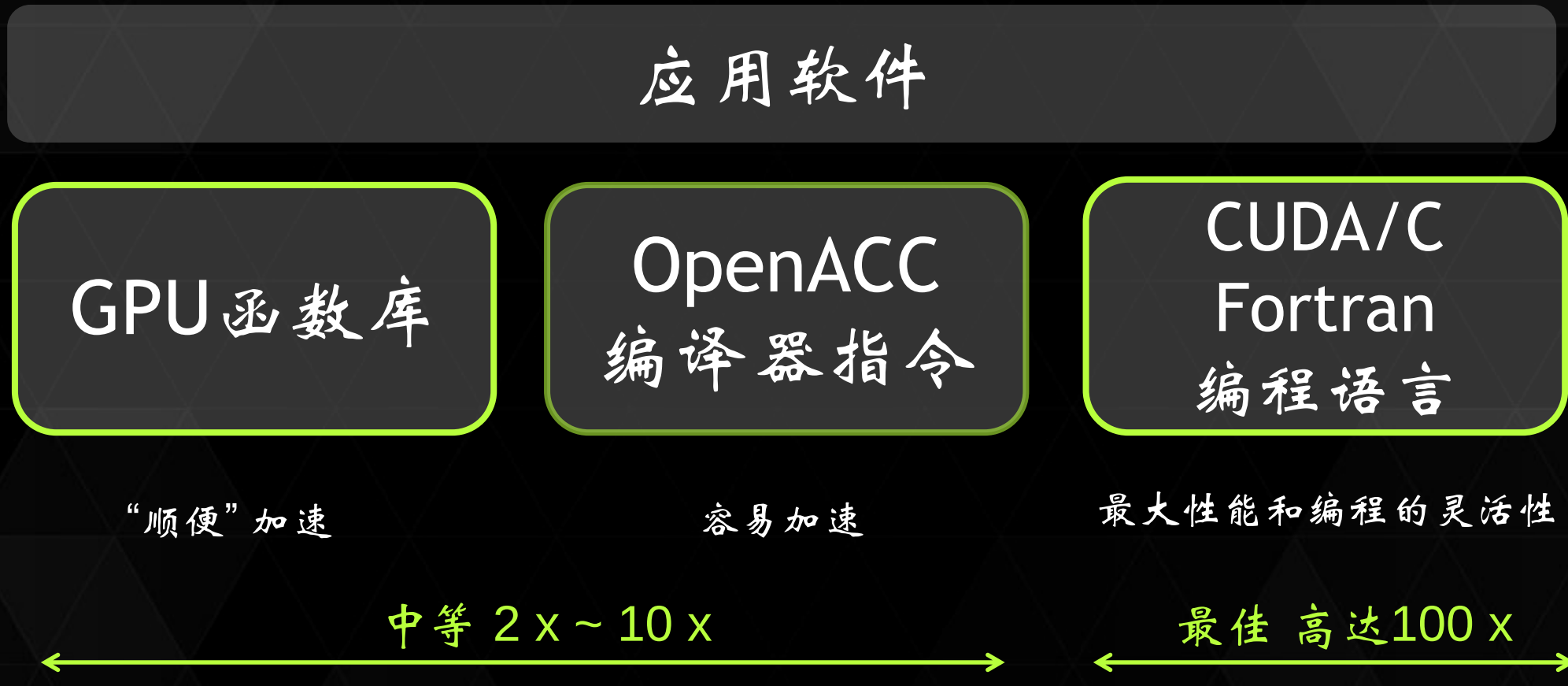




CPU + GPU = 加速应用



3 种方法加速应用



CPU + GPU 异构计算的优点

- 高性能
- 高能效
- 节省空间
- 高性价比

NVIDIA GPU 完善的开发者生态系统

Numerical Packages

MATLAB
Mathematica
NI LabView
pyCUDA

Debuggers & Profilers

cuda-gdb
NV Visual Profiler
Parallel Nsight
Visual Studio
Allinea
TotalView

GPU Compilers

C
C++
Fortran
OpenCL
DirectCompute
Java
Python

Parallelizing Compilers

PGI Accelerator
CAPS HMPP
mCUDA
OpenMP

Libraries

BLAS
FFT
LAPACK
NPP
Video
Imaging
GPULib

GPGPU Consultants & Training

acceleware · GASS · STONE RIDGE TECHNOLOGY · ANEO · GPU Tech · TECH-X · CAPS · HPC · EM Photonics · SCALABLE GRAPHICS · WIPRO

Microsoft

OEM Solution Providers

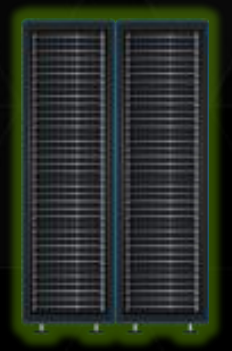
DELL · hp · IBM · CRAY · ASUS · SUPERMICRO · sgi · FUJITSU · BULL · APPRO · lenovo 联想 · NEC

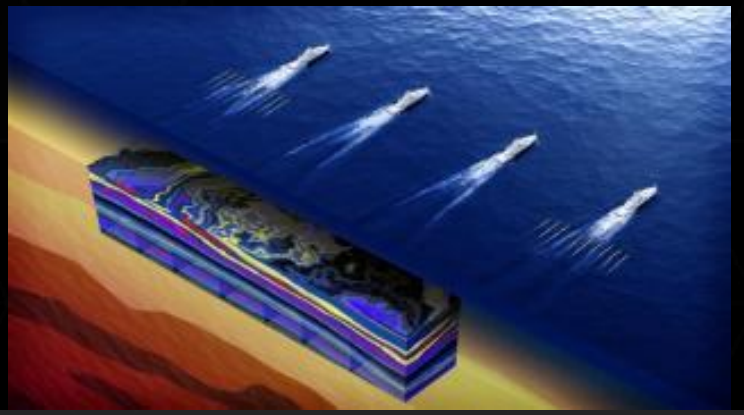
NVIDIA

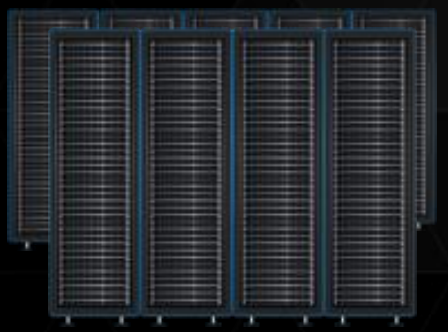
CUDA 的几个数字：

>487,000,000	CUDA GPUs
>2,000,000	CUDA 工具箱下载
>140,000	活跃的开发者
>700	所大学教授 CUDA

GPU的应用领域： 石油天然气







速度提高 4x-6x

时间：28 天 → 7 天

能源效率：4x

电力和制冷的费用：¼

总体费用：减少12%

改进运营成本/资本支出

相同的成本，
更多的销售额

逆时偏移
RTM

更好的图像，
更高的吞吐量

NVIDIA

GPU的应用领域：国防及公共安全

- 视频分析
  - 视频增强
  - 视频稳定
- 人像识别
- 指纹识别
- 雷达成像
- 卫星图形处理

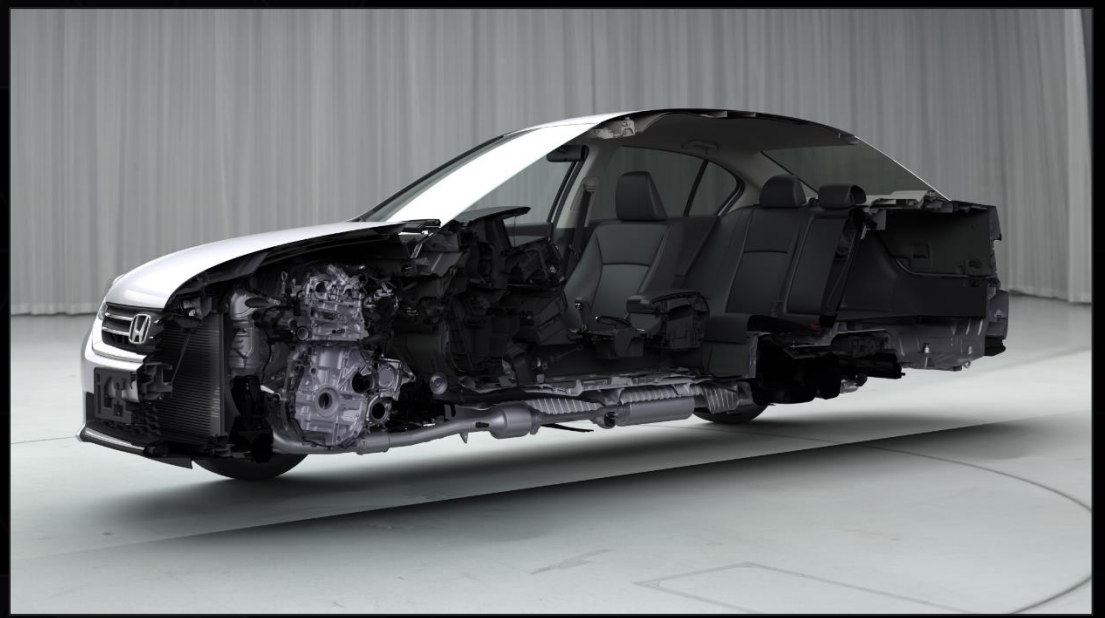
GPU的应用领域：医疗

- 医疗成像
  - 超声波
  - CT
  - 核磁共振
  - X-RAY

SIEMENS medical

GE Healthcare

GPU的应用领域：工业设计



GPU的应用领域：流体分析

GPU的应用领域：电影制作与动画

GPU的应用领域：生命科学/新药开发

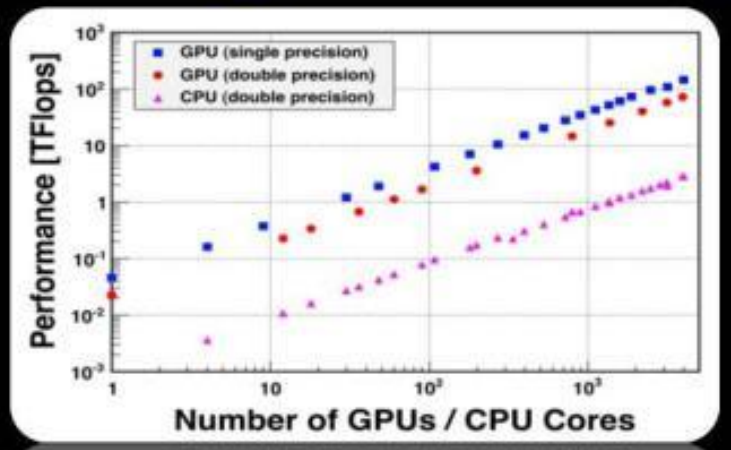
GPU的应用领域：天气及海洋预报

ASUCA TeraFlop Scaling (Weather Modeling)

3990 Tesla M2050s

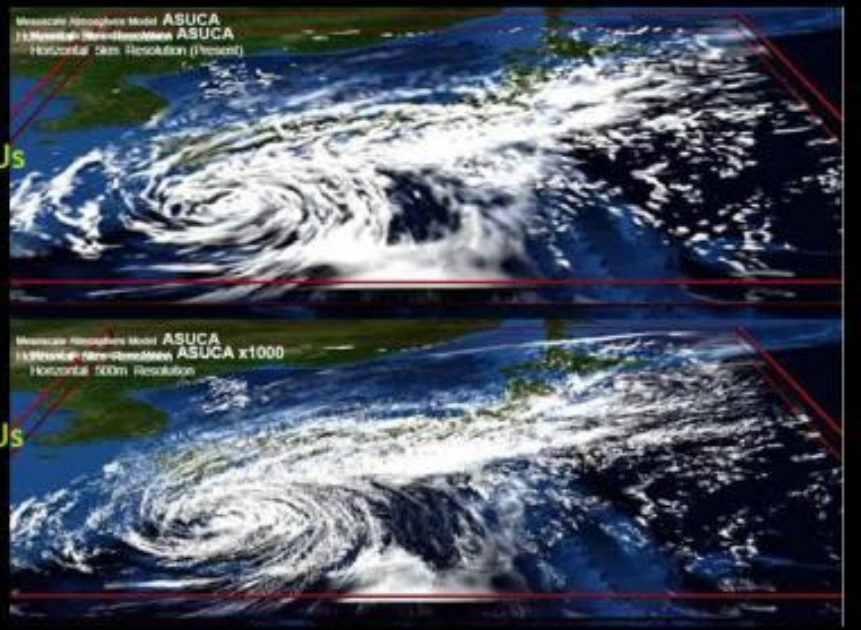
145.0 Tflops SP

76.1 Tflops DP



Before GPUs

After GPUs



Simulation on Tsubame 2.0, TITech Supercomputer

NVIDIA

GPU的应用领域：金融

近实时的定价和风险计算

- 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 enabling calculation of risk across various investment products in a matter of minutes instead of what was previously an overnight processing task.

JPMorgan

GPU的应用领域：机器学习

吴韧博士

杰出科学家，深度学习研究院

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

“ CUDA as a programming model has become very mature over the years and is perfect for training deep neural networks, especially for large models. ”

NVIDIA

从HPC到企业数据中心

石油天然气

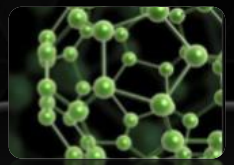
Schlumberger

BR PETROBRAS

Eni

Chevron

Statoil



高等教育

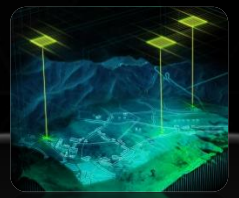
HARVARD School of Engineering and Applied Sciences

STANFORD UNIVERSITY

Georgia Tech

ETH Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zürich

UNIVERSITY OF CAMBRIDGE



政府

Air Force Research Laboratory

Raytheon

NASA

Naval Research Laboratory

超算

CSCS

NCSA

Tokyo Institute of Technology

OAK RIDGE National Laboratory

Lawrence Livermore National Laboratory

金融

J.P.Morgan

BARCLAYS

STANDARD LIFE

BNP PARIBAS

MUREX

互联网

Baidu 百度

salesforce

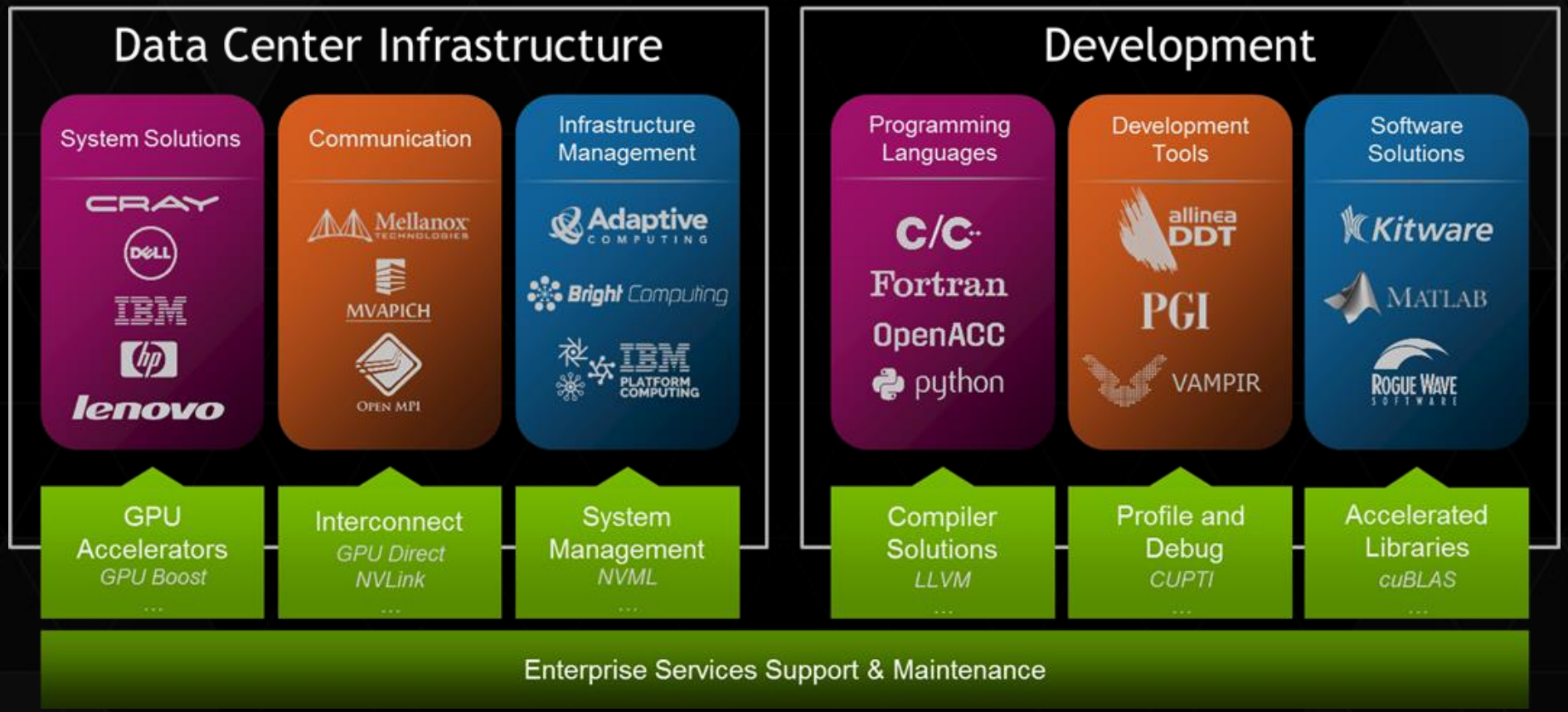
SHAZAM

amazon.com

Yandex

NVIDIA

Tesla 加速平台



“ Accelerators Will Be Installed in More than Half of New Systems ”

“In 2014, NVIDIA enjoyed a dominant market share with 85% of the accelerator market.”

Source: Top 6 predictions for HPC in 2015 Intersect360 RESEARCH

NVIDIA

开放的生态系统
加速应用支持多平台

Libraries

AmgX

cuDNN

cuBLAS

OpenCV

Thrust

Compiler Directives

OpenACC

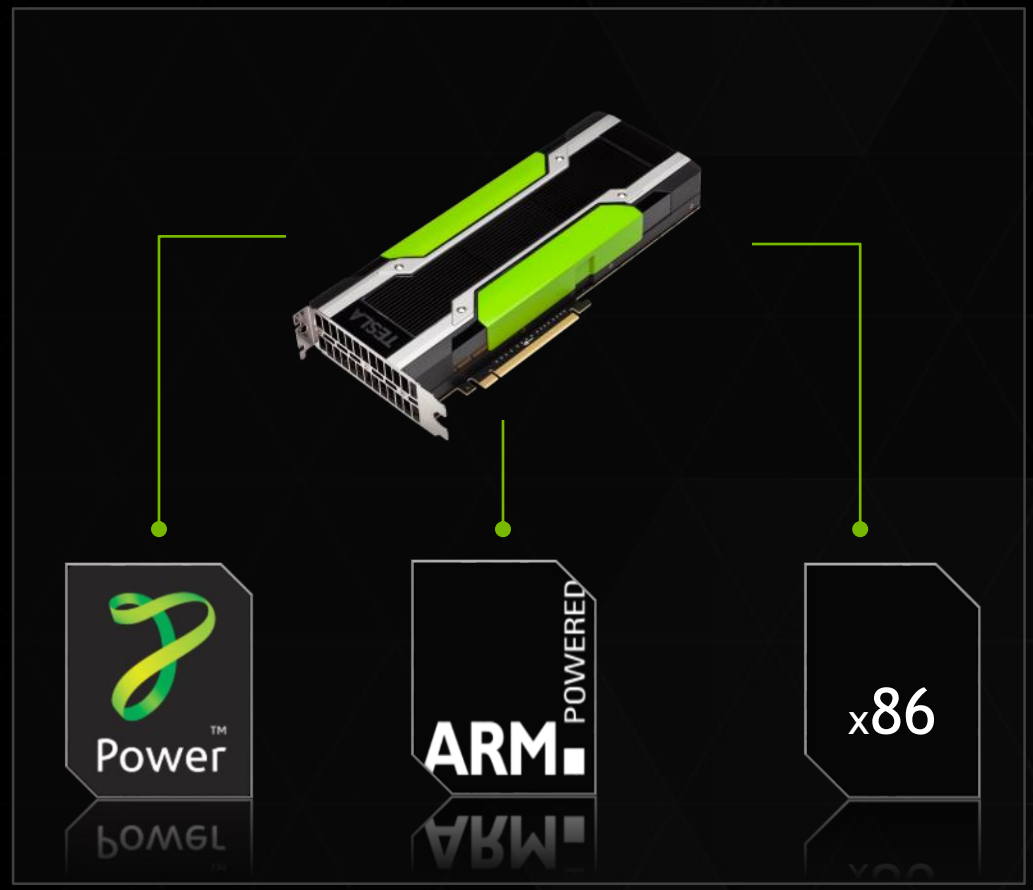
Programming Languages

C/C++

Fortran

python

Java



NVIDIA

2015 年 Tesla GPU 加速器产品

Tesla K40

Best Single GPU Performance

Server, Workstation, Liquid Cooled

Higher Ed, Data Analytics, HPC Labs, Defense

Double Precision Workloads

Tesla K80

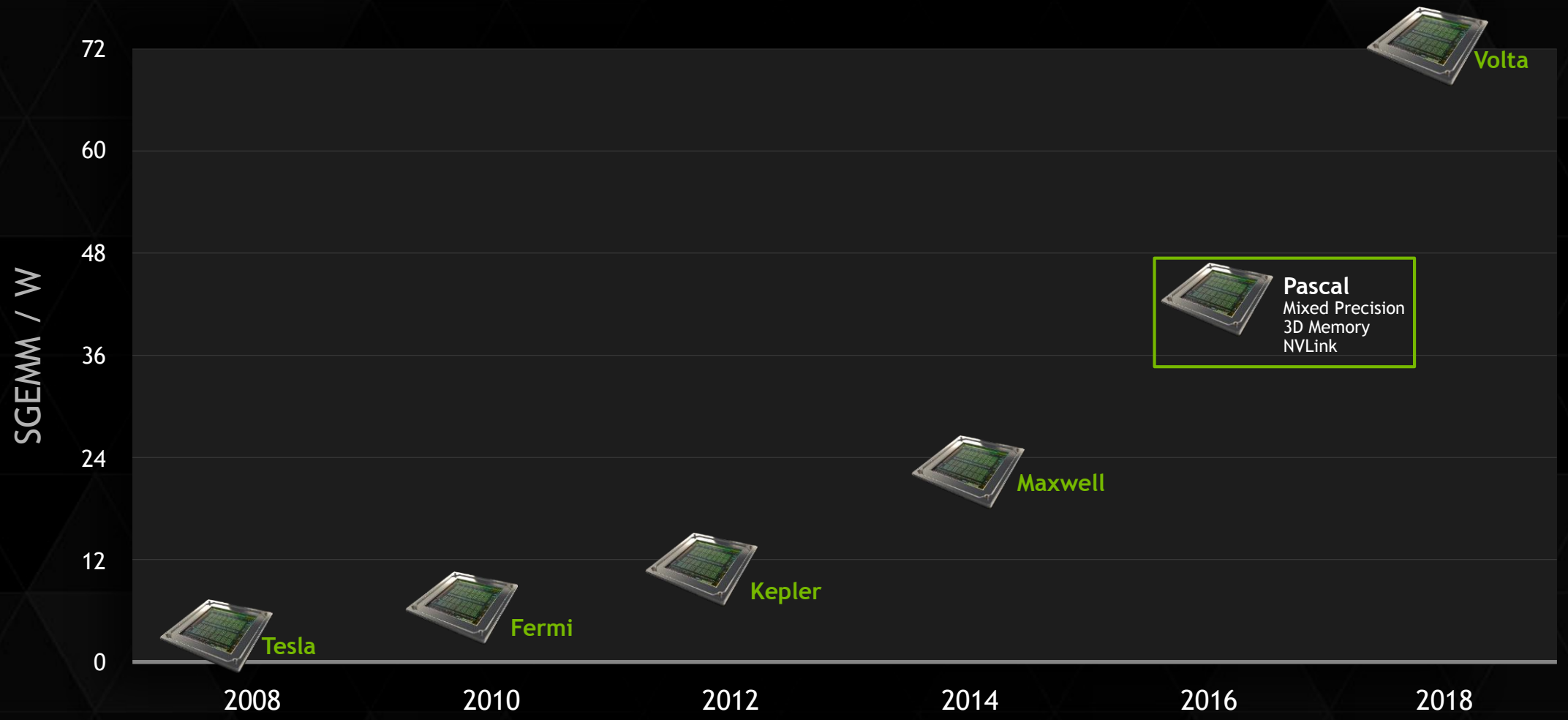
Maximize Throughput within a Server

Seismic, Data Analytics, HPC Labs, Defense

Multi-GPU Accelerated Apps

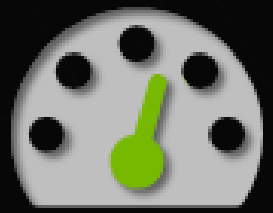
Single and Double Precision Workloads

GPU 发展路线图



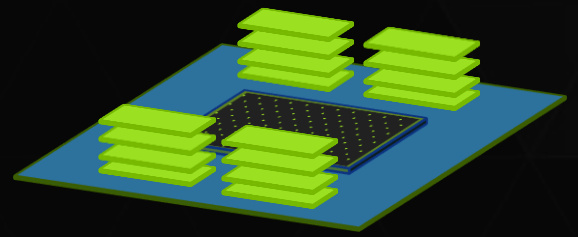
PASCAL: 下一代 TESLA GPU

Peak Performance



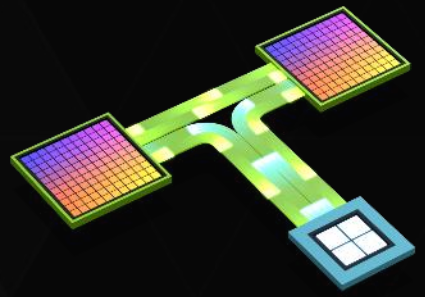
>3 TeraFLOPS

Stacked Memory



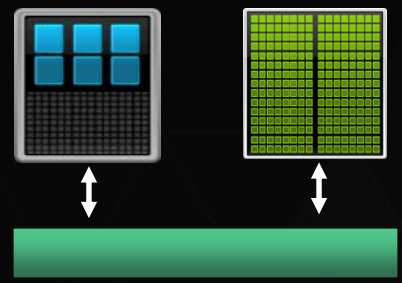
4x Higher Bandwidth (~1 TB/s)
Larger Capacity (16 GB)

NVLink High-Speed Interconnect



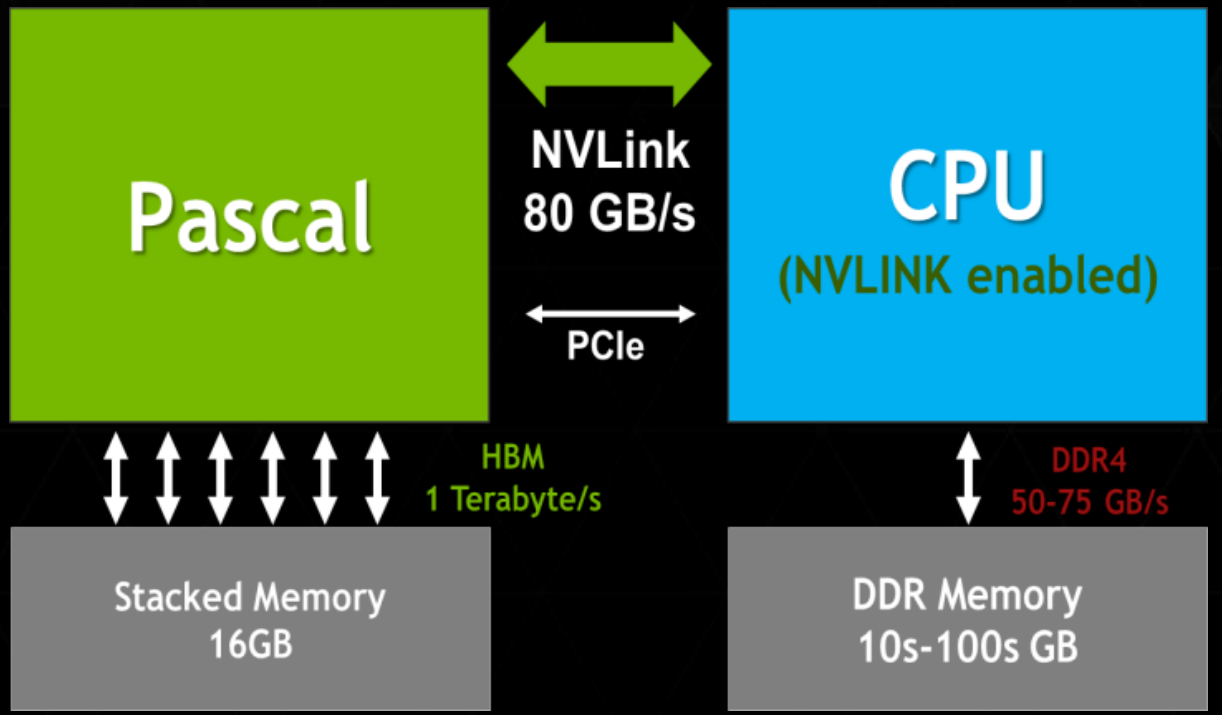
80 GB/sec
POWER CPU & GPU-to-GPU Interconnect

Unified Memory



Single Memory Space
Lower Developer Effort

NVLINK 连接 CPU 和GPU

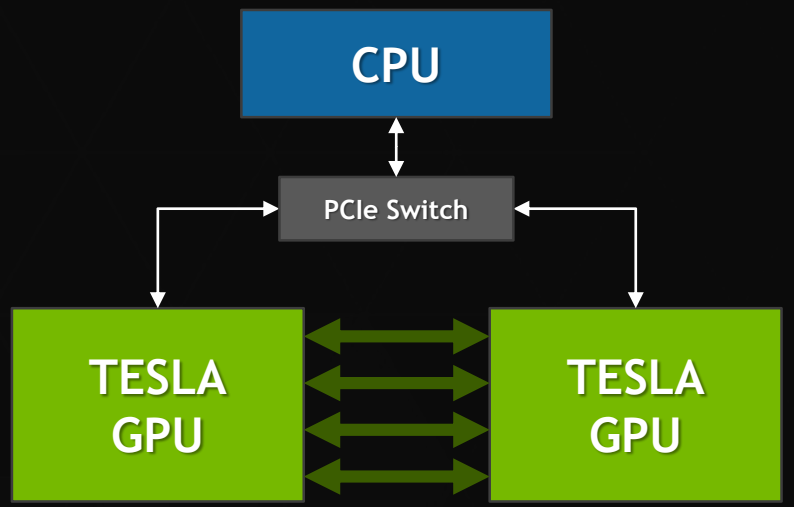


NVLINK 1.0

4 data links with 20GB/s each

X86平台多GPU NVLINK 的性能

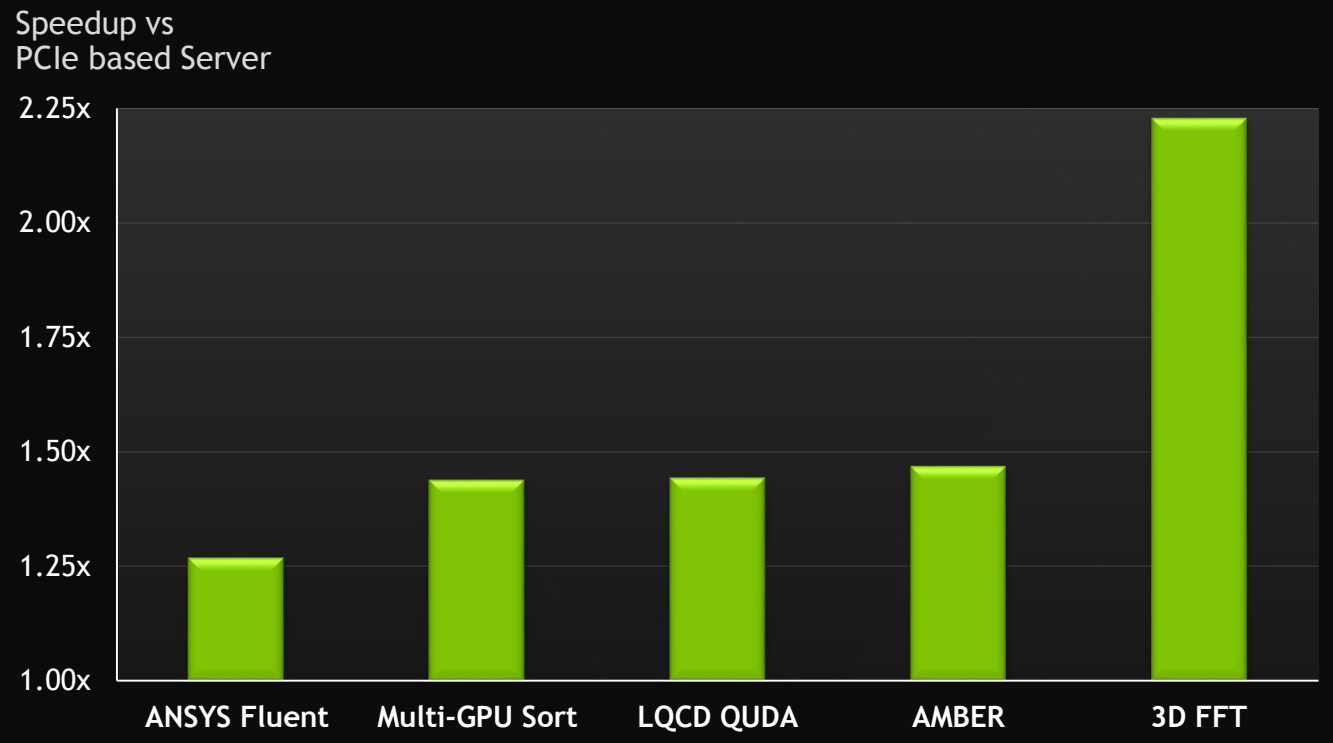
GPUs Interconnected with NVLink



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

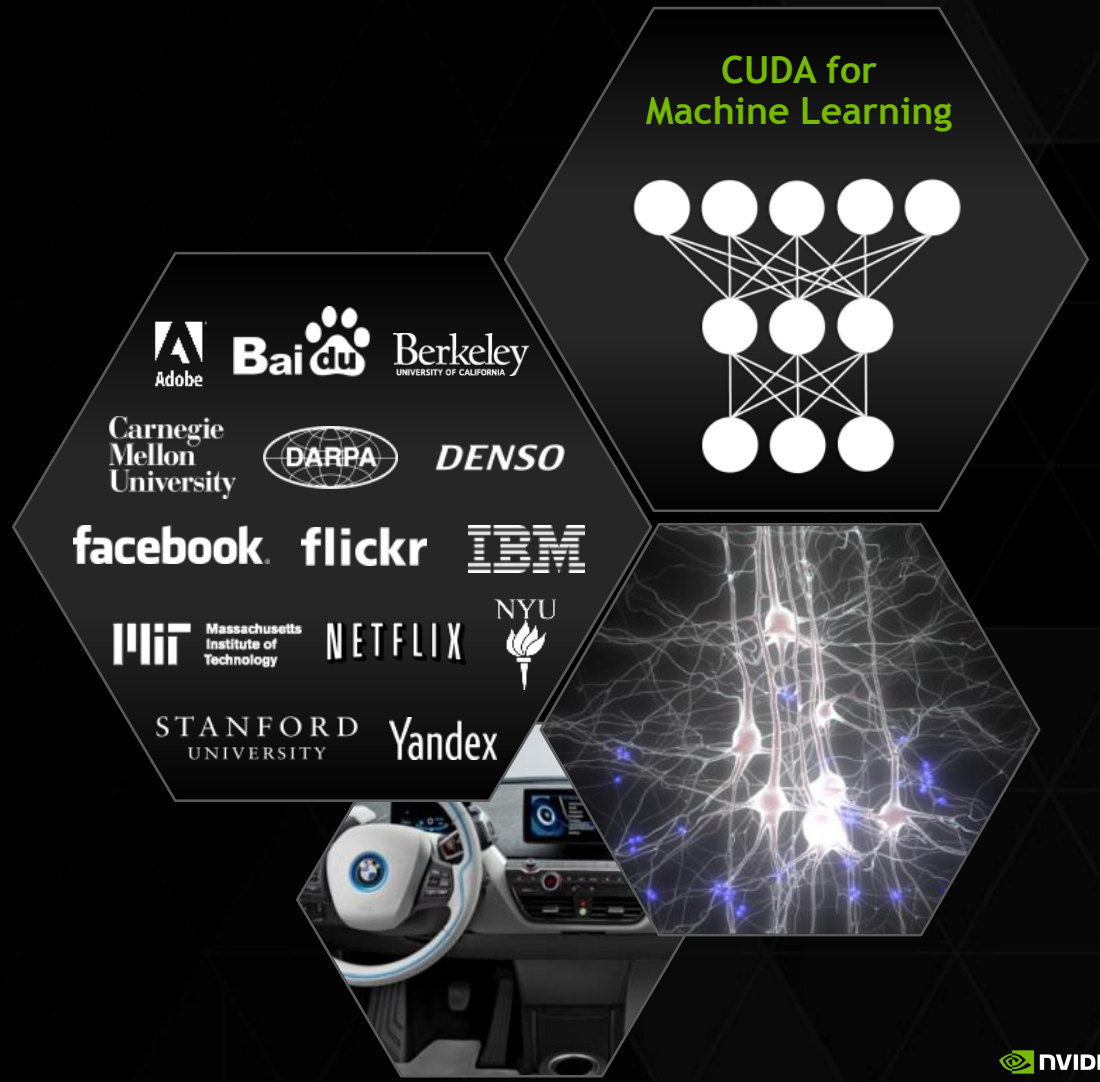
深度学习

加速机器学习

“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.



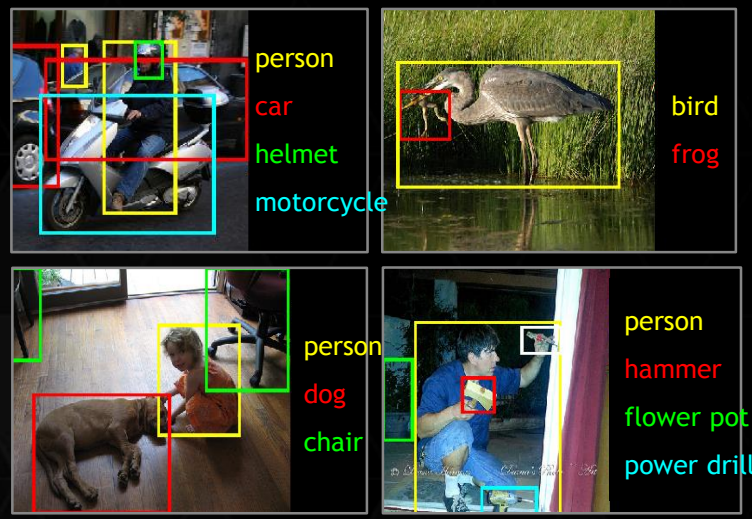
GPU – 机器学习的平台

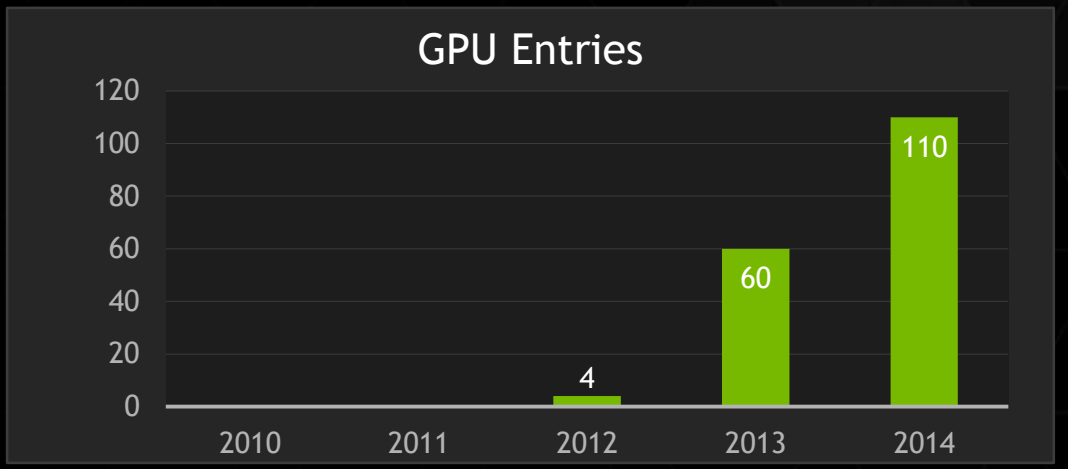
Image Recognition Challenge

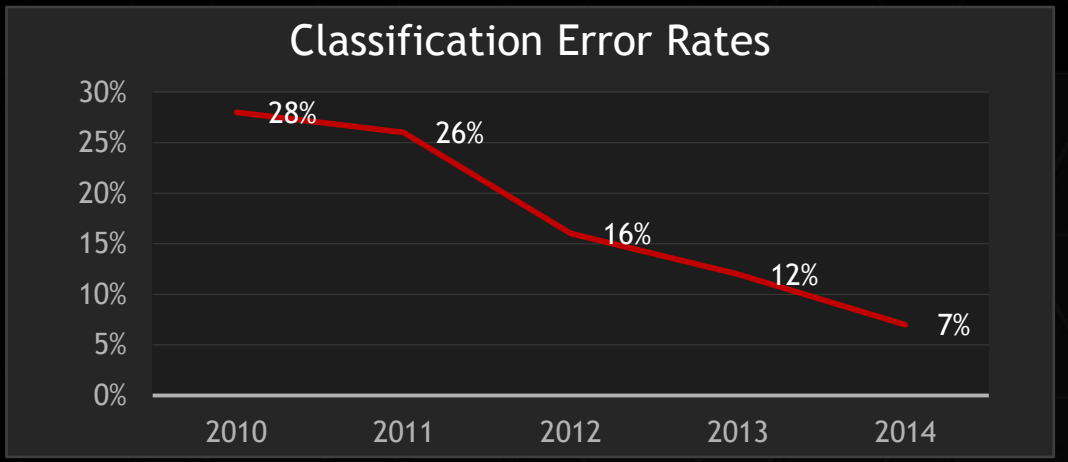
1.2M training images • 1000 object categories

Hosted by

IMAGENET





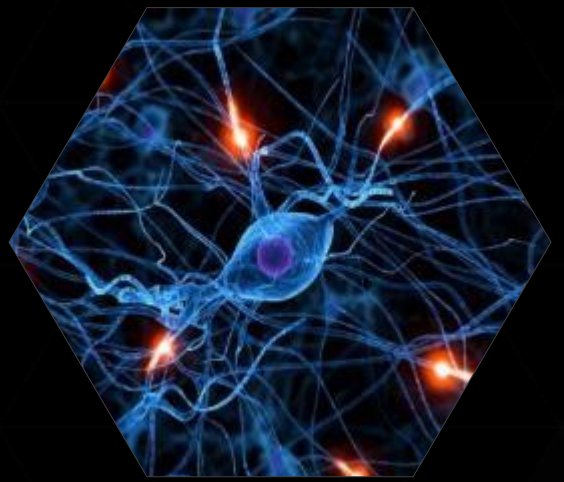


NVIDIA

深度学习的3大驱动力

大数据

模型



GPU加速器

GPU在深度学习的广泛使用

Early Adopters

Adobe
Image Analytics for Creative Cloud

Baidu
Speech/Image Recognition

flickr
Image Classification

IBM
Hadoop

NETFLIX
Recommendation

Yandex
Search Rankings

Use Cases

- Image Detection
- Face Recognition
- Gesture Recognition
- Video Search & Analytics
- Speech Recognition & Translation
- Recommendation Engines
- Indexing & Search

Talks @ GTC

facebook

NYU

STANFORD UNIVERSITY

DARPA

DENSO

Carnegie Mellon University

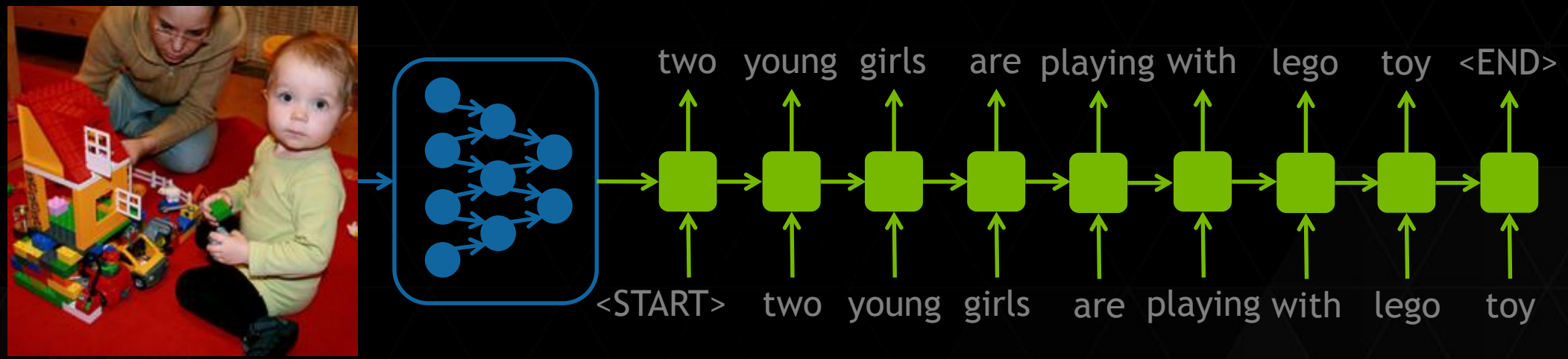
MIT Massachusetts Institute of Technology

Berkeley UNIVERSITY OF CALIFORNIA

应用实例：图像到文字的翻译

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

应用实例：汽车的自动驾驶

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.

http://arxiv.org/pdf/1504.01716v1.pdf

50+
Deep Learning Sessions

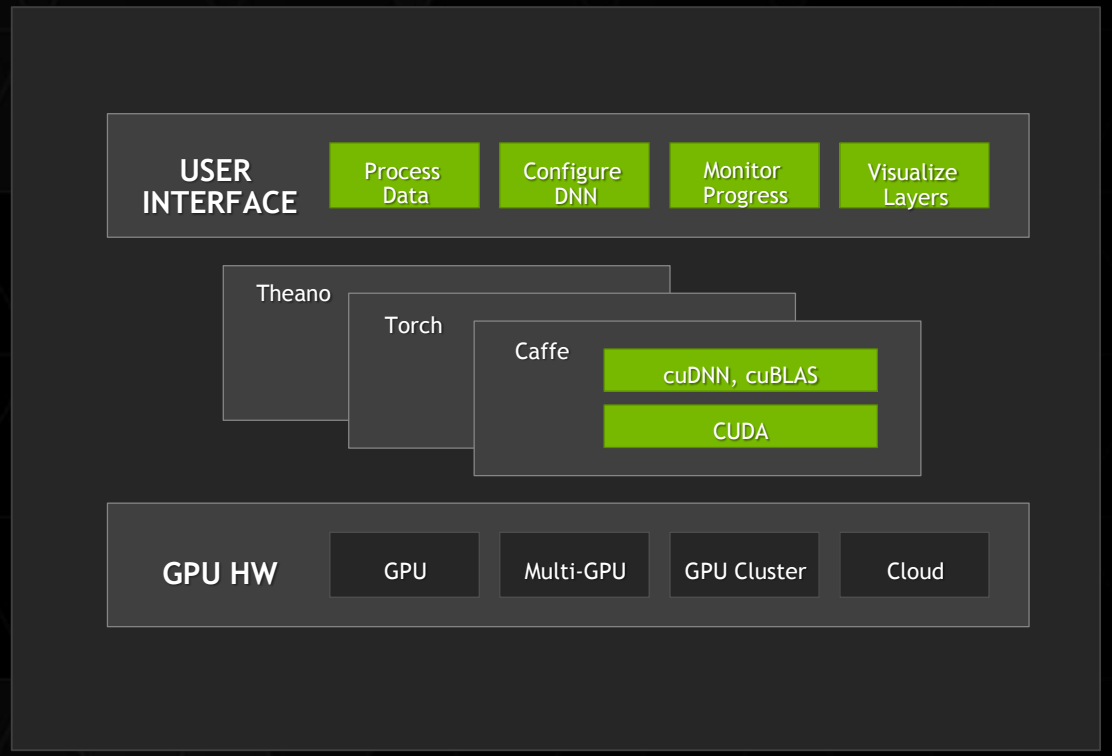
www.gputechconf.com

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 训练系统

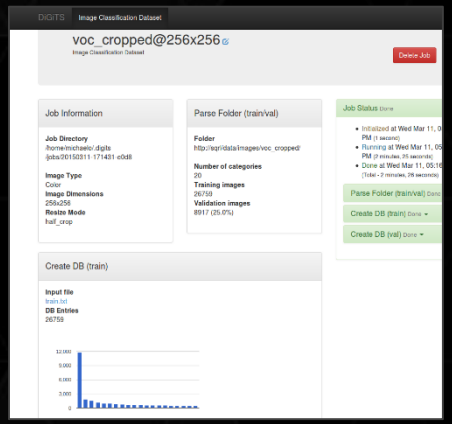


DEEP GPU TRAINING SYSTEM FOR DATA SCIENTISTS

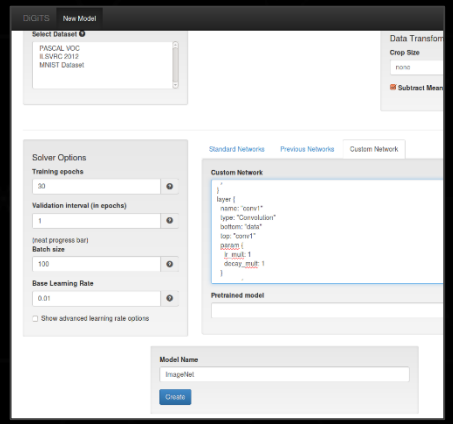
- Design DNNs
- Visualize activations
- Manage multiple trainings

DIGITS

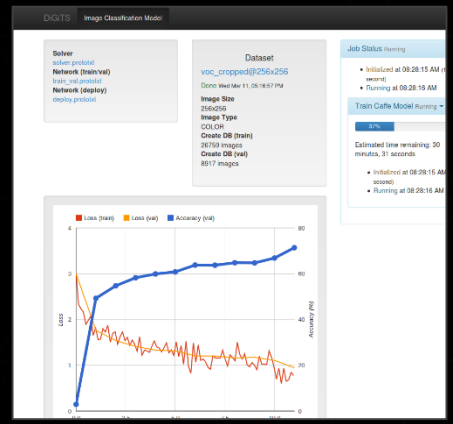
Process Data



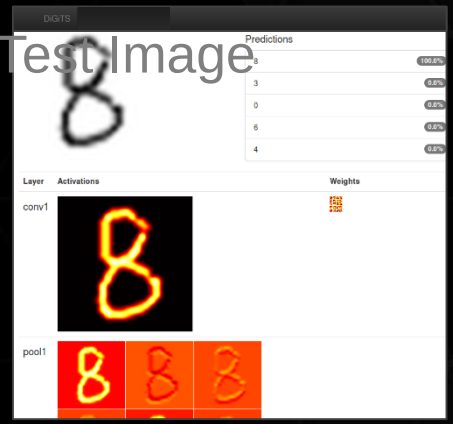
Configure DNN



Monitor Progress



Visualize Layers



谢谢！

HLUO@NVIDIA.COM