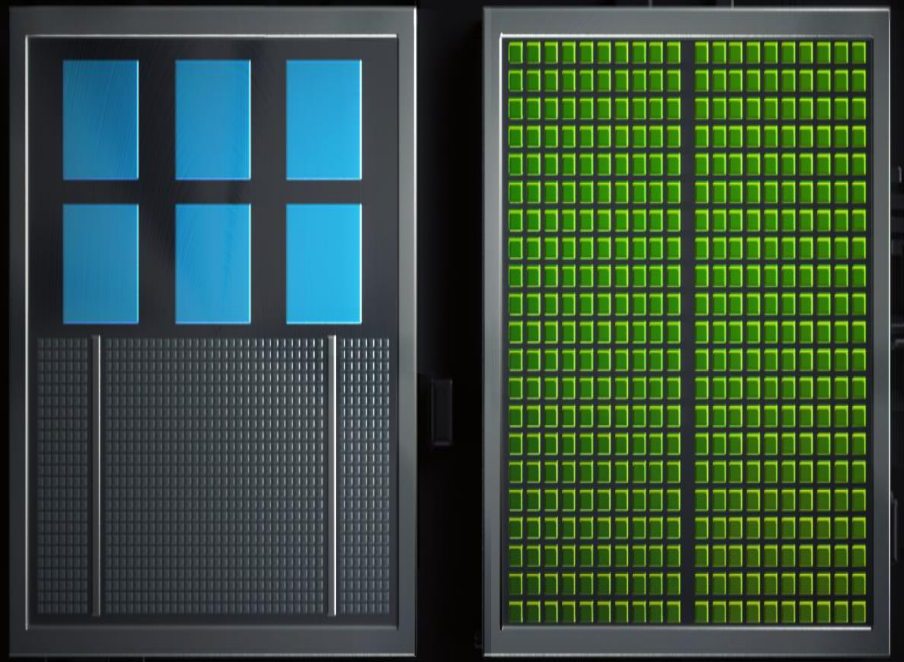
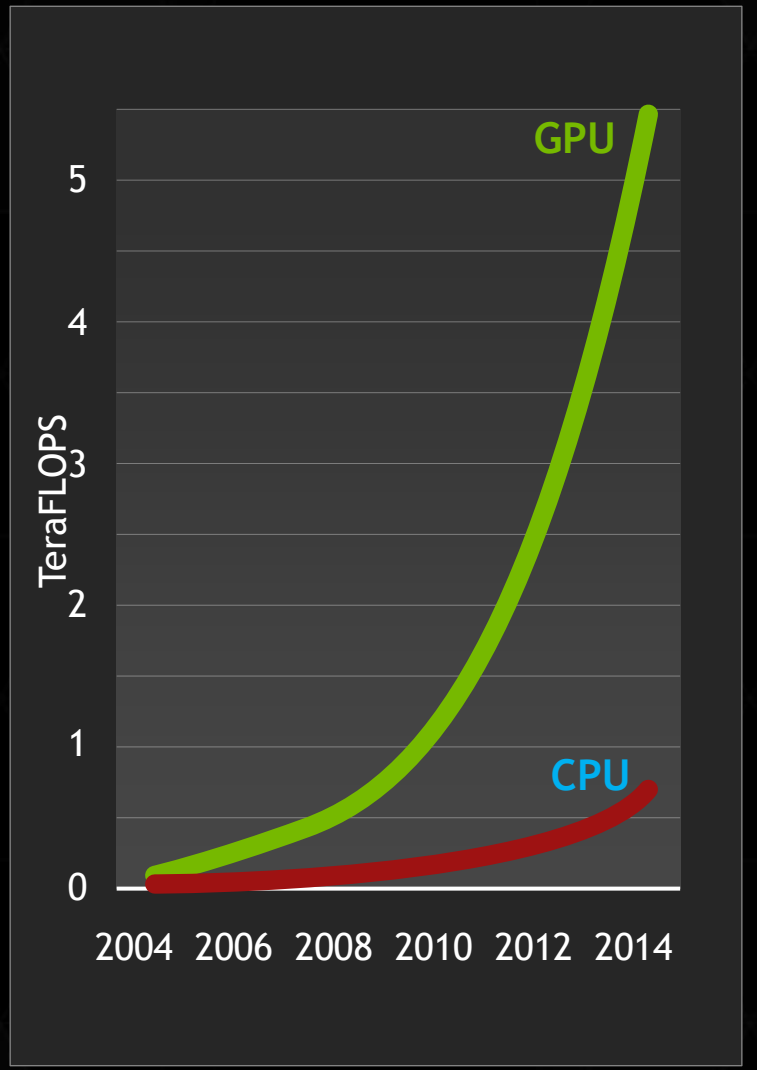
GPU TECHNOLOGY: PAST, PRESENT, FUTURE

Marc Hamilton, Vice President,
Solution Architecture and Engineering





A Decade Of GPU Computing

- From Scientific Computing To Machine Learning
- Mobile Is More Than Just Phones
- GPU Architecture & CUDA Roadmap
- Grid & The Last Mile of Virtualization

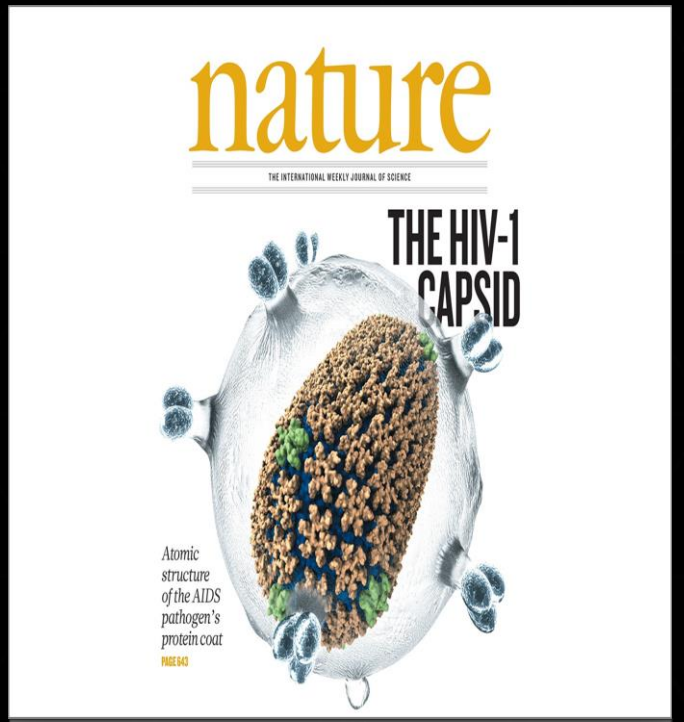
From Scientific Computing To Machine Learning

ResQU

Giving Drones the Vision to Help Fight Fires

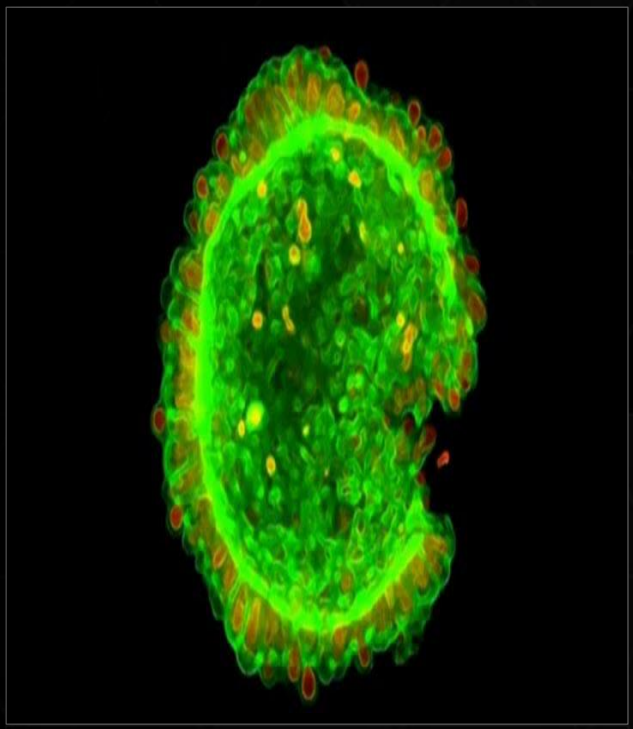
UNIVERSITY OF ILLINOIS

A Breakthrough in HIV Research



HOLOGIC

Early, Accurate Detection of Breast Cancer



The Green500 List

Listed below are the June 2014 The Green500's energy-efficient supercomputers ranked from 1 to 100.

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	4,389.82	GSIC Center, Tokyo Institute of Technology	TSUBAME-KFC - LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x	34.58
2	3,631.70	Cambridge University	Wilkes - Dell T620 Cluster, Intel Xeon E5-2630v2 6C 2.600GHz, Infiniband FDR, NVIDIA K20	52.62
3	3,517.84	Center for Computational Sciences, University of Tsukuba	HA-PACS TCA - Cray 3623G4-SM Cluster, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband QDR, NVIDIA K20x	78.77
4	3,459.46	SURFsara	Cartesius Accelerator Island - Bullx B515 cluster, Intel Xeon E5-2450v2 8C 2.5GHz, InfiniBand 4× FDR, Nvidia K40m	44.40
5	3,185.91	Swiss National Supercomputing Centre (CSCS)	Piz Daint - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Level 3 measurement data available	1,753.66
6	3,131.06	ROMEO HPC Center - Champagne-Ardenne	romeo - Bull R421-E3 Cluster, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR, NVIDIA K20x	81.41
7	3,019.72	CSIRO	CSIRO GPU Cluster - Nitro G16 3GPU, Xeon E5-2650 8C 2GHz, Infiniband FDR, Nvidia K20m	86.20
8	2,951.95	GSIC Center, Tokyo Institute of Technology	TSUBAME 2.5 - Cluster Platform SL390s G7, Xeon X5670 6C 2.93GHz, Infiniband QDR, NVIDIA K20x	927.86
9	2,813.14	Exploration & Production - Eni S.p.A.	HPC2 - iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.8GHz, Infiniband FDR, NVIDIA K20x	1,067.49
10	2,678.41	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	54.60
11	2,629.42	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband FDR, NVIDIA K20x	66.25
12	2,629.42	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband FDR, NVIDIA K20x	66.25
13	2,629.42	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband FDR, NVIDIA K20x	66.25
14	2,629.42	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband FDR, NVIDIA K20x	66.25
15	2,629.10	Max-Planck-Gesellschaft MPI/IPP	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	269.94

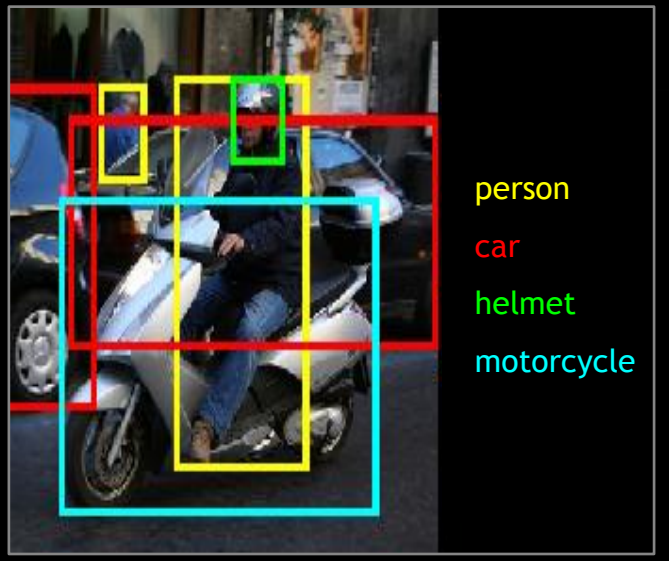
TSUBAME KFC

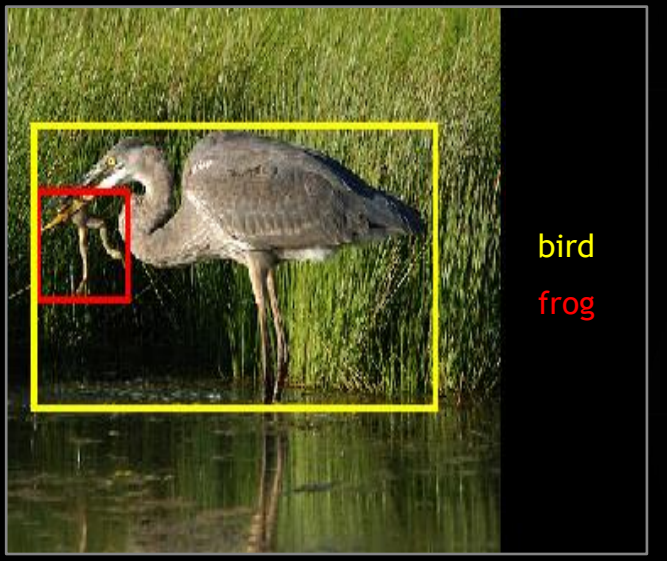
#1 OF "TOP 15" GREEN SUPERCOMPUTERS POWERED BY CUDA GPUS

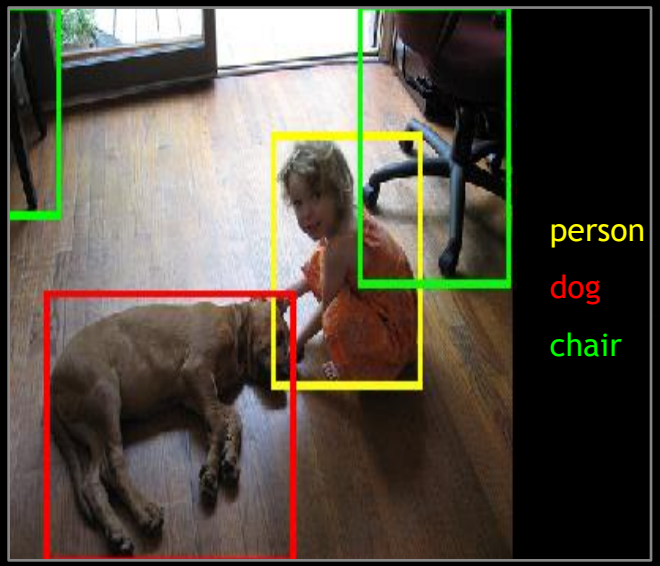
MACHINE LEARNING

Branch of Artificial Intelligence

Computers that learn from data

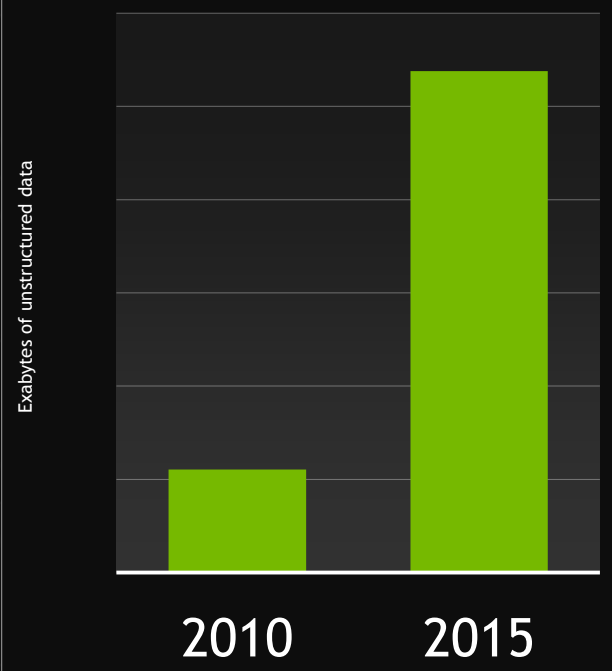






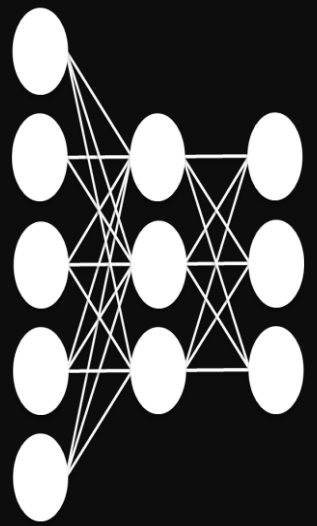
THREE TRENDS CONVERGING

Torrent of Data

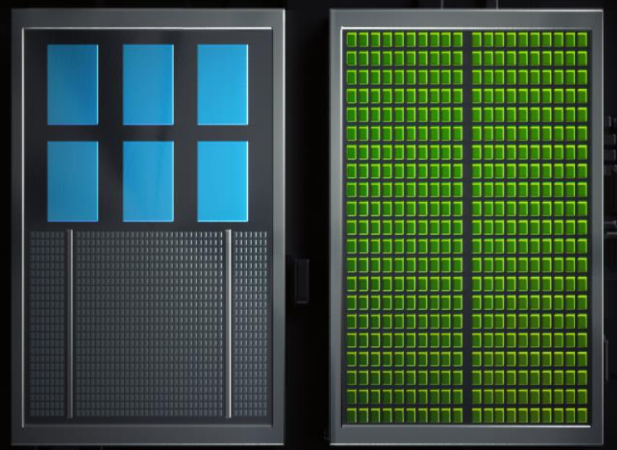


SOURCE: : IDC

Deep Neural Networks



GPU Computing



Deep Learning with COTS HPC Systems

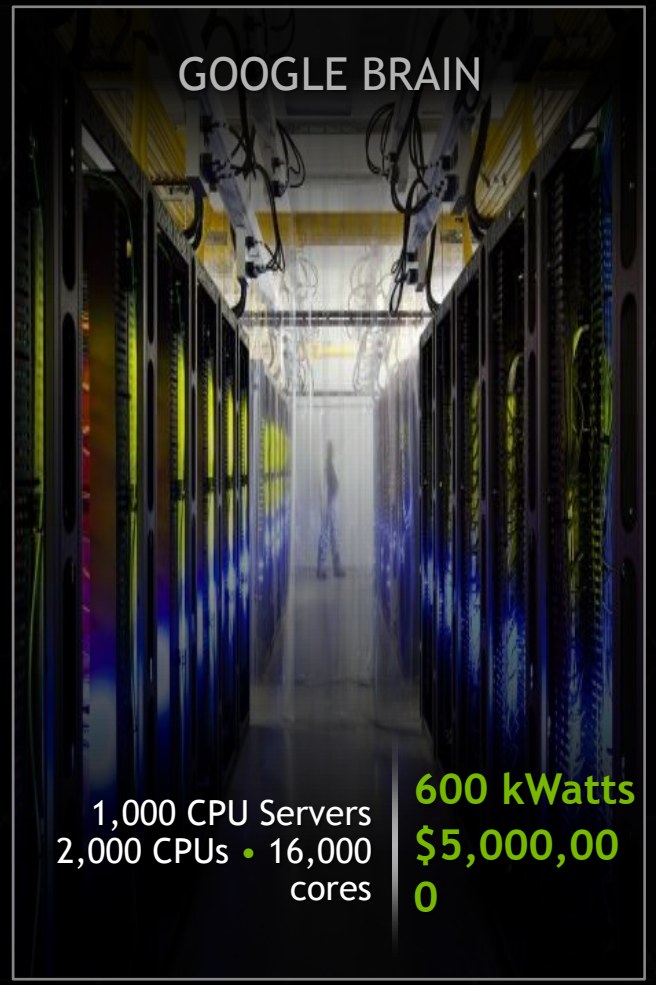
A. Coates, B. Huval, T. Wang, D. Wu, A. Ng, B. Catanzaro

Stanford / NVIDIA • ICML 2013

"Now You Can Build Google's $1M Artificial Brain on the Cheap"

-Wired

GOOGLE BRAIN



1,000 CPU Servers
2,000 CPUs • 16,000 cores

600 kWatts
$5,000,000

STANFORD AI LAB

3 GPU-Accelerated Servers
12 GPUs • 18,432 cores

4 kWatts
$33,000

CUDA FOR MACHINE LEARNING

Early Adopters

Image Analytics for Creative Cloud

Speech/Image Recognition

Image Classification

Hadoop

Recommendation

Search Rankings

Use Cases

Image Detection

Face Recognition

Gesture Recognition

Video Search & Analytics

Speech Recognition & Translation

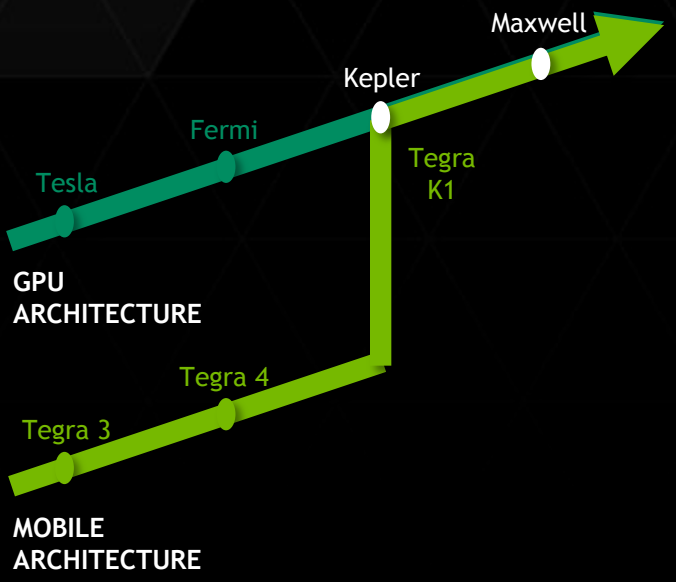
Recommendation Engines

Indexing & Search

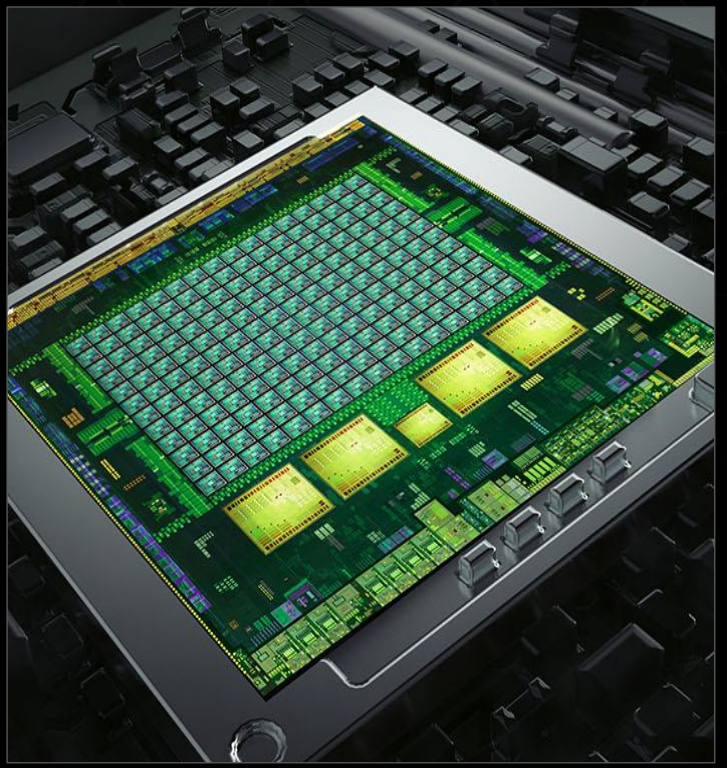
Prominent Research

Mobile - More Than Just Phones

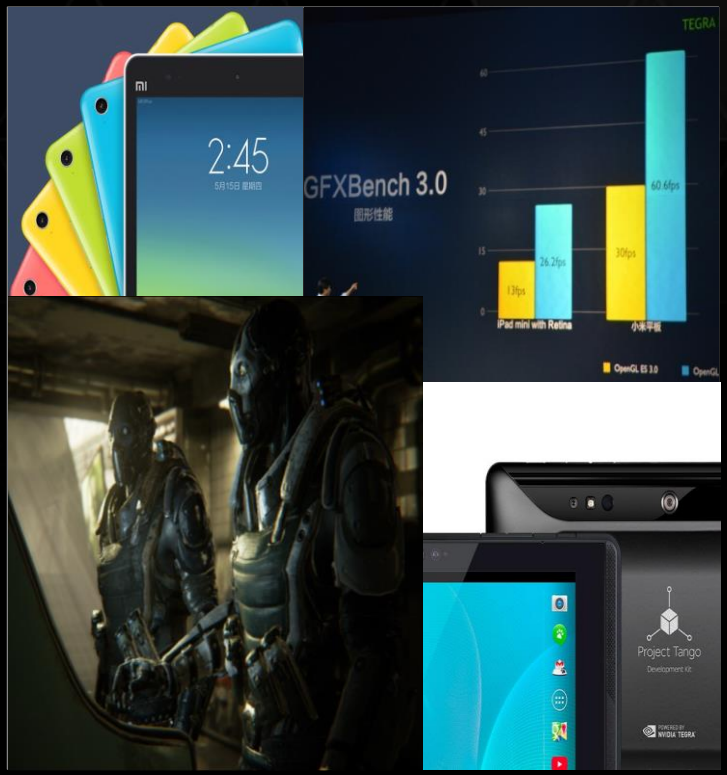
TEGRA TK1



UNIFIED ARCHITECTURE



TEGRA K1 – MOBILE SUPER CHIP



BREAKTHROUGH EXPERIENCES

JETSON TK1 DEV KIT

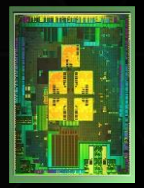
1ST MOBILE SUPERCOMPUTER FOR EMBEDDED SYSTEMS

192 CUDA cores

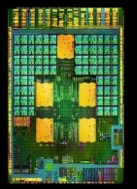
326 GFLOPS

VisionWorks SDK

EVOLUTION OF COMPUTING IN THE CAR

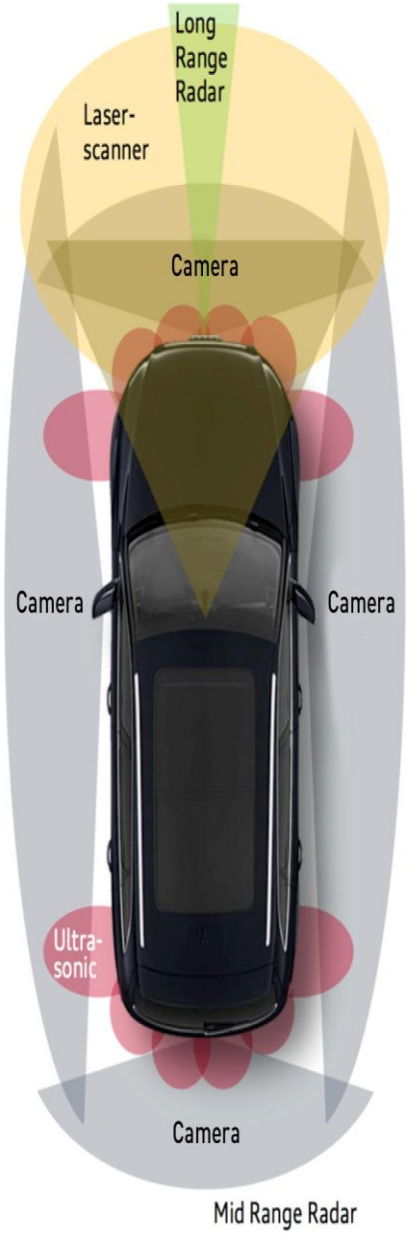


Tegra 3



Tegra 4

Tegra K1

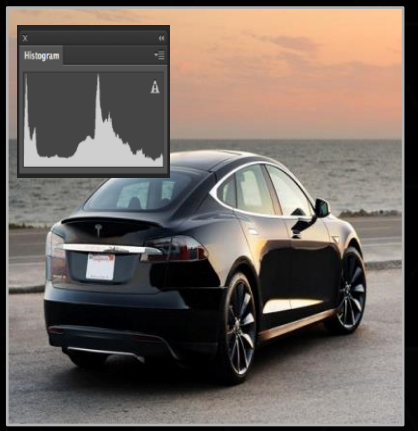


TEGRA TK1 SUPERCOMPUTER FOR DRIVER ASSISTANCE

Pedestrian Detection
Blind Spot Monitoring
Lane Departure Warning

Collision Avoidance
Traffic Sign Recognition
Adaptive Cruise Control

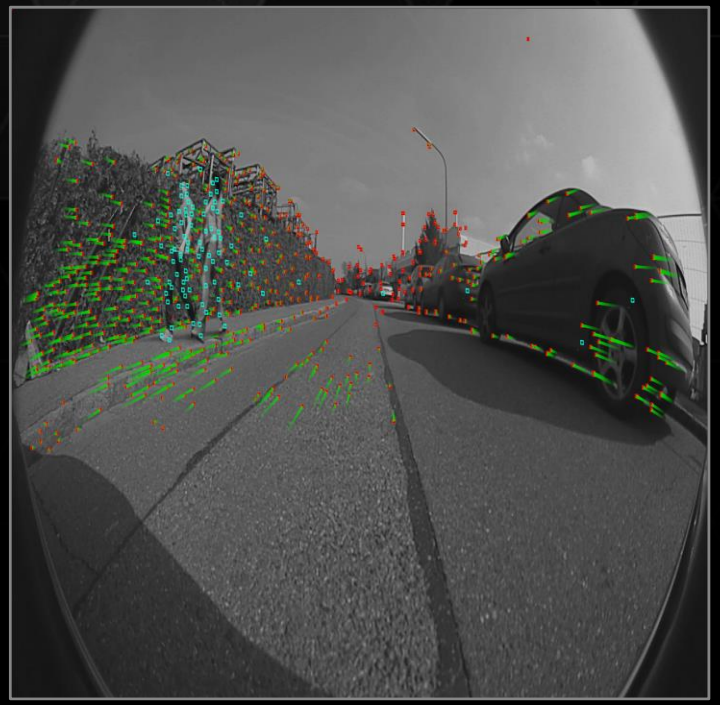
Optical Flow



Histogram

Feature Detection

COMPUTER VISION ON CUDA

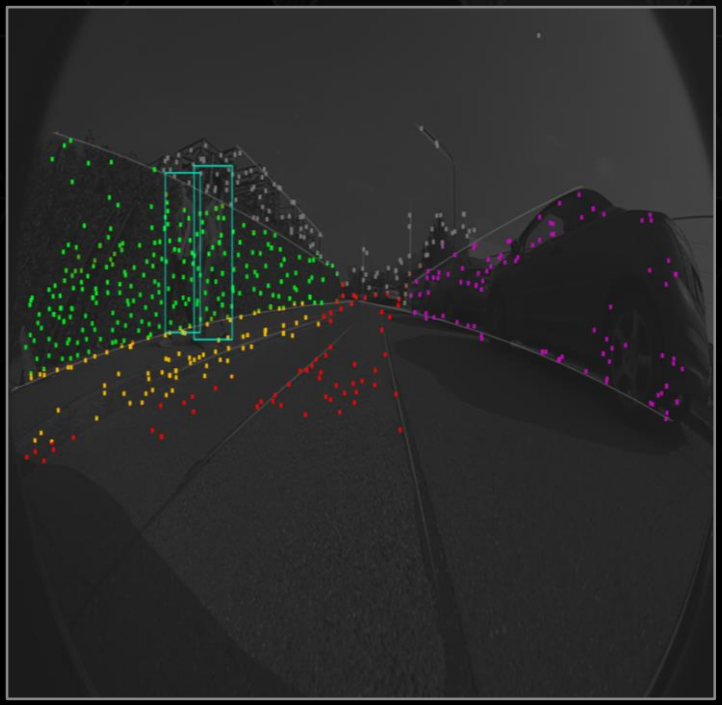


Feature Detection / Tracking

~30 GFLOPS @ 30 Hz

Object Recognition / Tracking

~180 GFLOPS @ 30 Hz

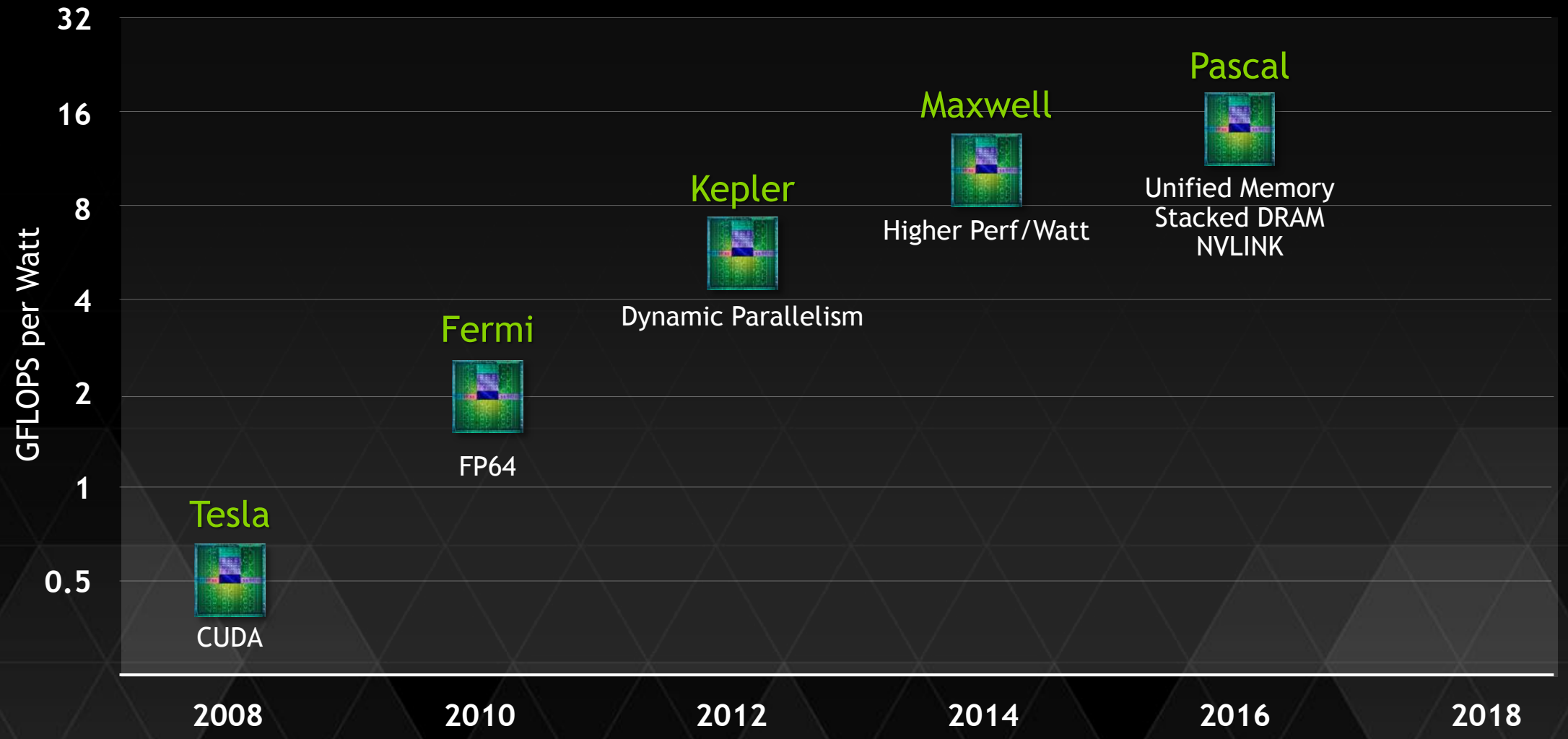


3D Scene Interpretation

~280 GFLOPS @ 30 Hz

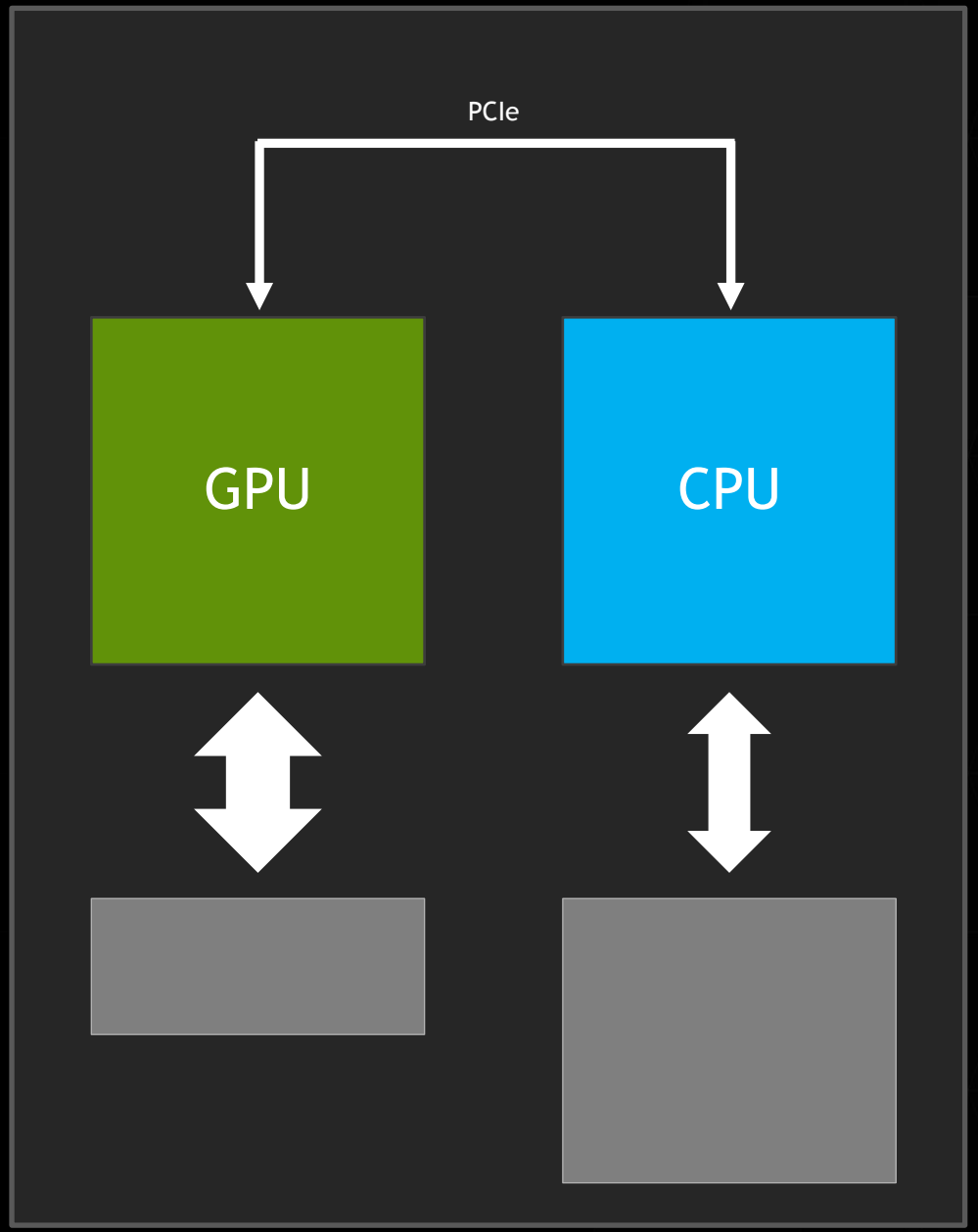
GPU Architecture & CUDA Roadmap

FAST PACED CUDA GPU ROADMAP



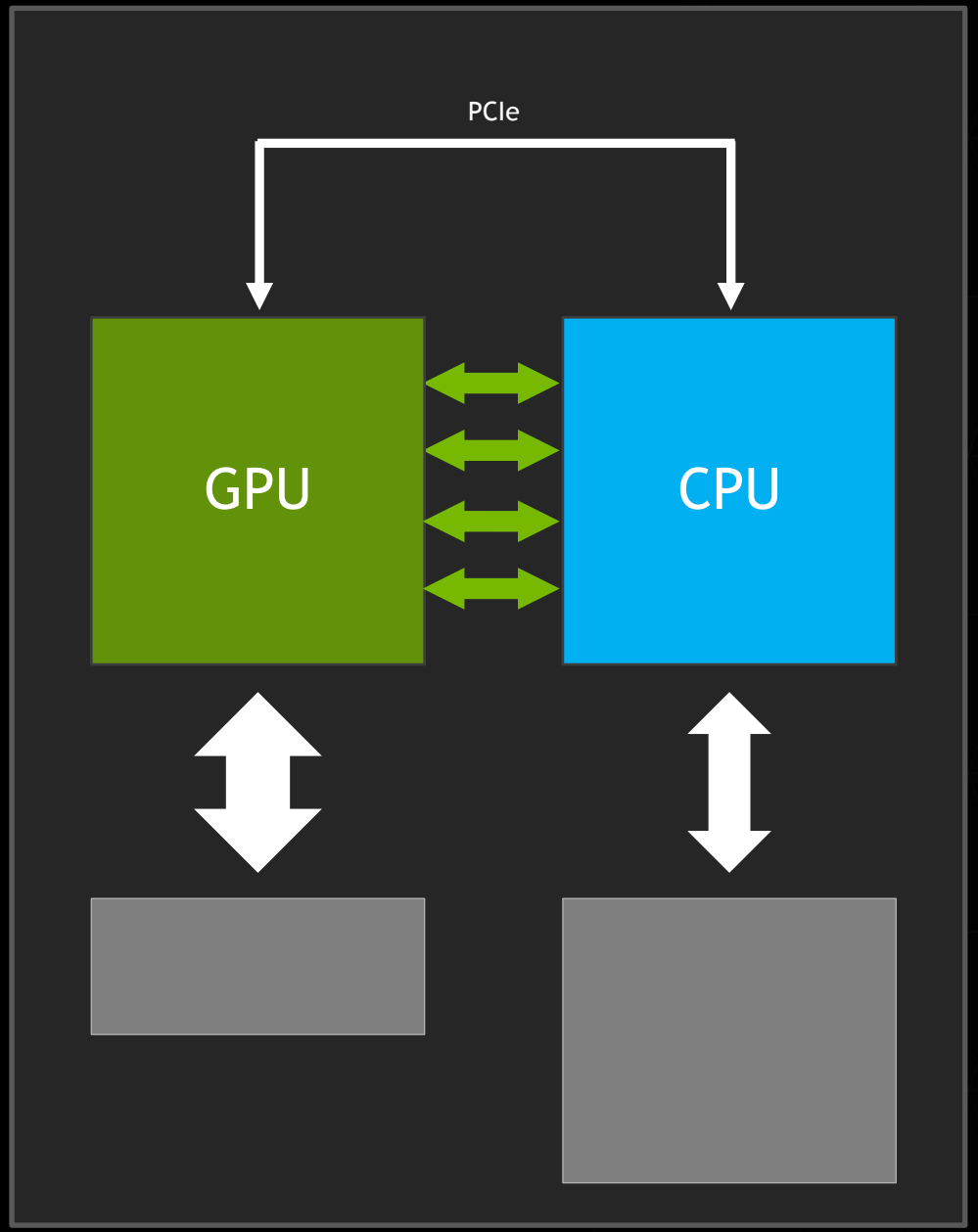
BANDWIDTH BOTTLENECKS

PCI Express	16GB/sec
CPU Memory	60GB/sec
GPU Memory	288GB/sec

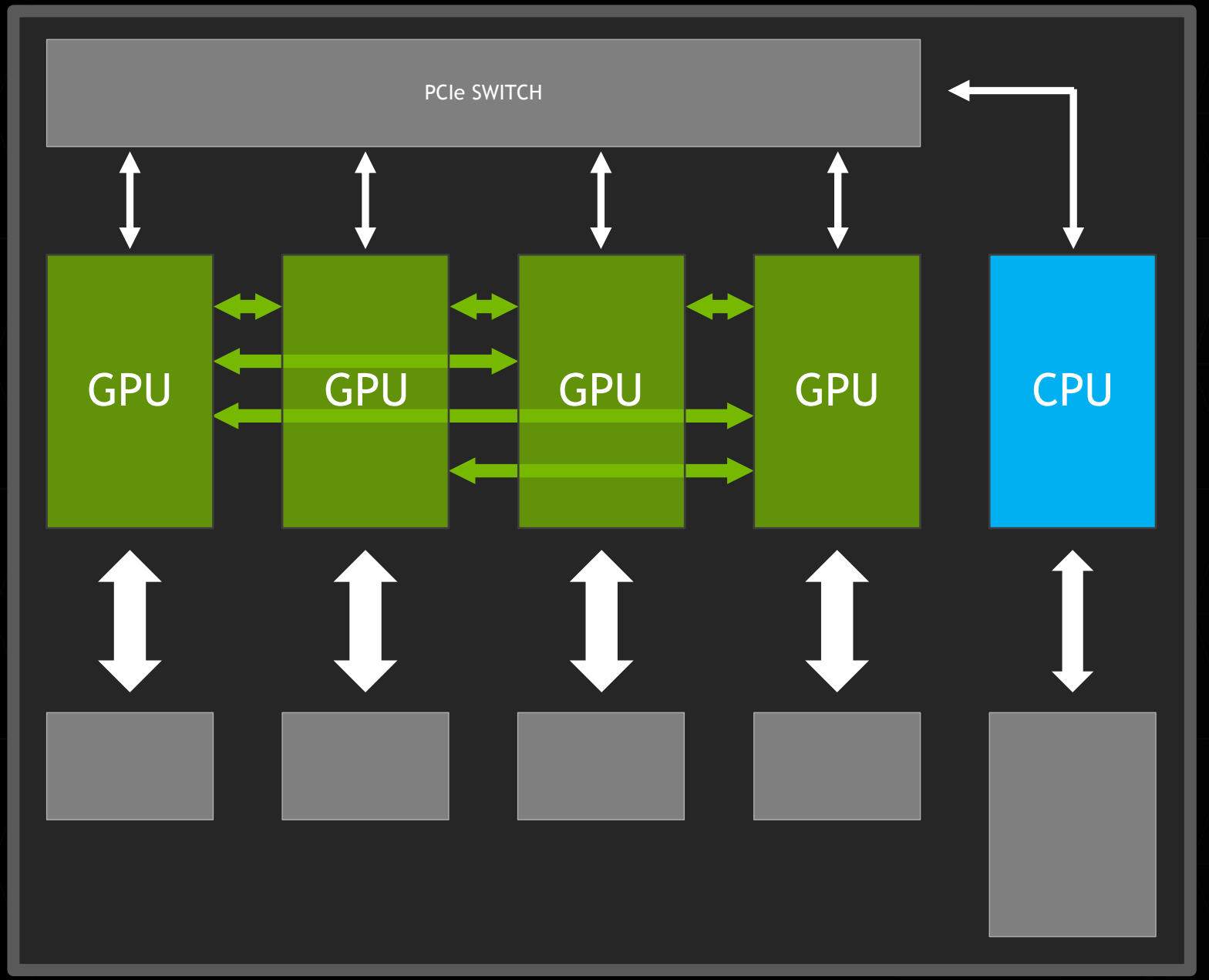


INTRODUCING NVLINK

- Differential with embedded clock
- PCIe programming model (w/ DMA+)
- Unified Memory
- Cache coherency in Gen 2.0
- 5 to 12X PCIe



5X MORE BANDWIDTH FOR MULTI-GPU SCALING



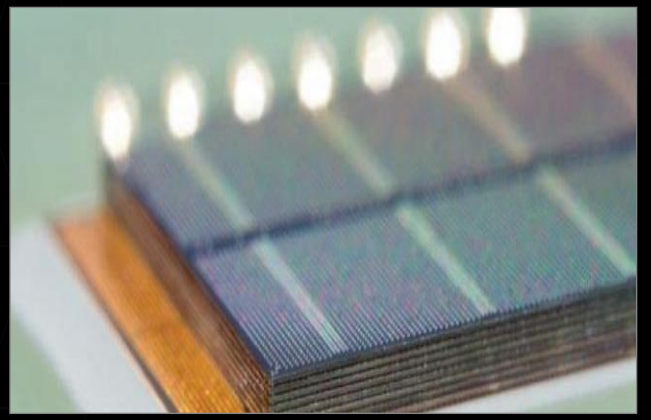
3D MEMORY

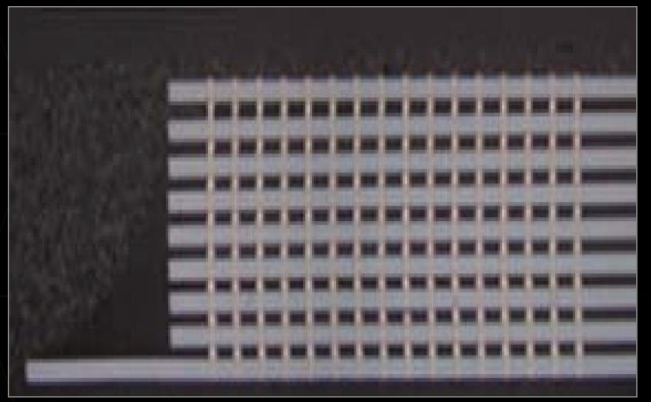
3D Chip-on-Wafer integration

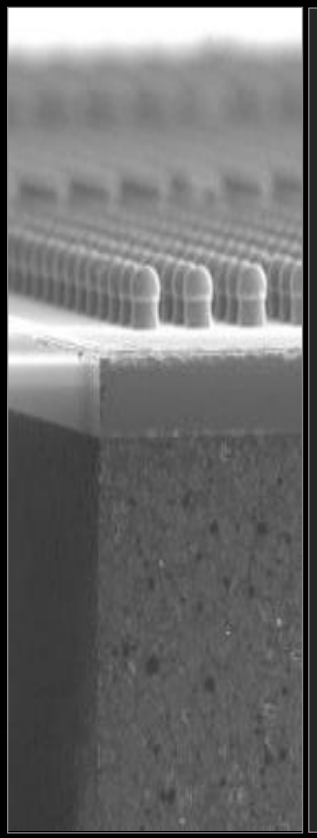
Many X bandwidth

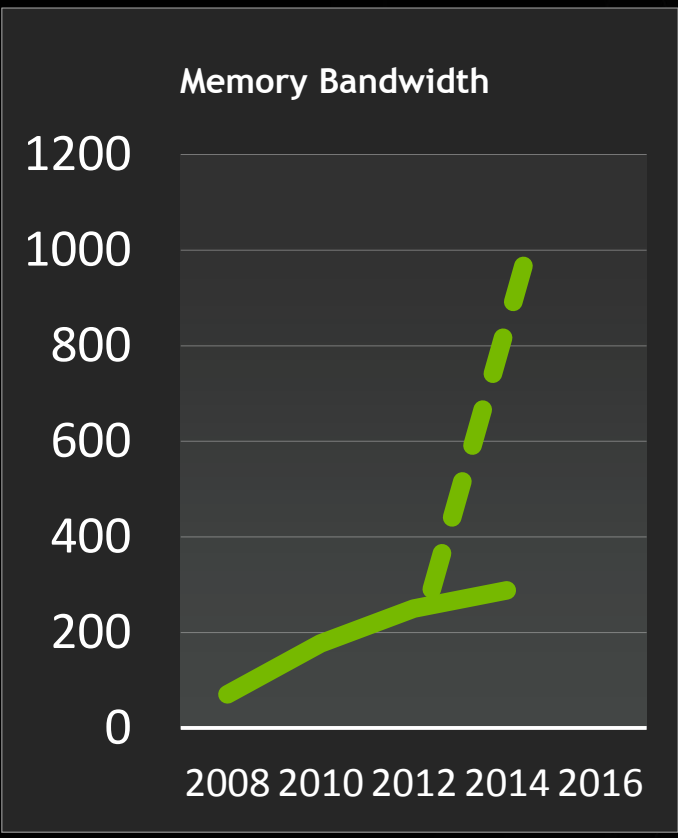
2.5X capacity

4X energy efficiency



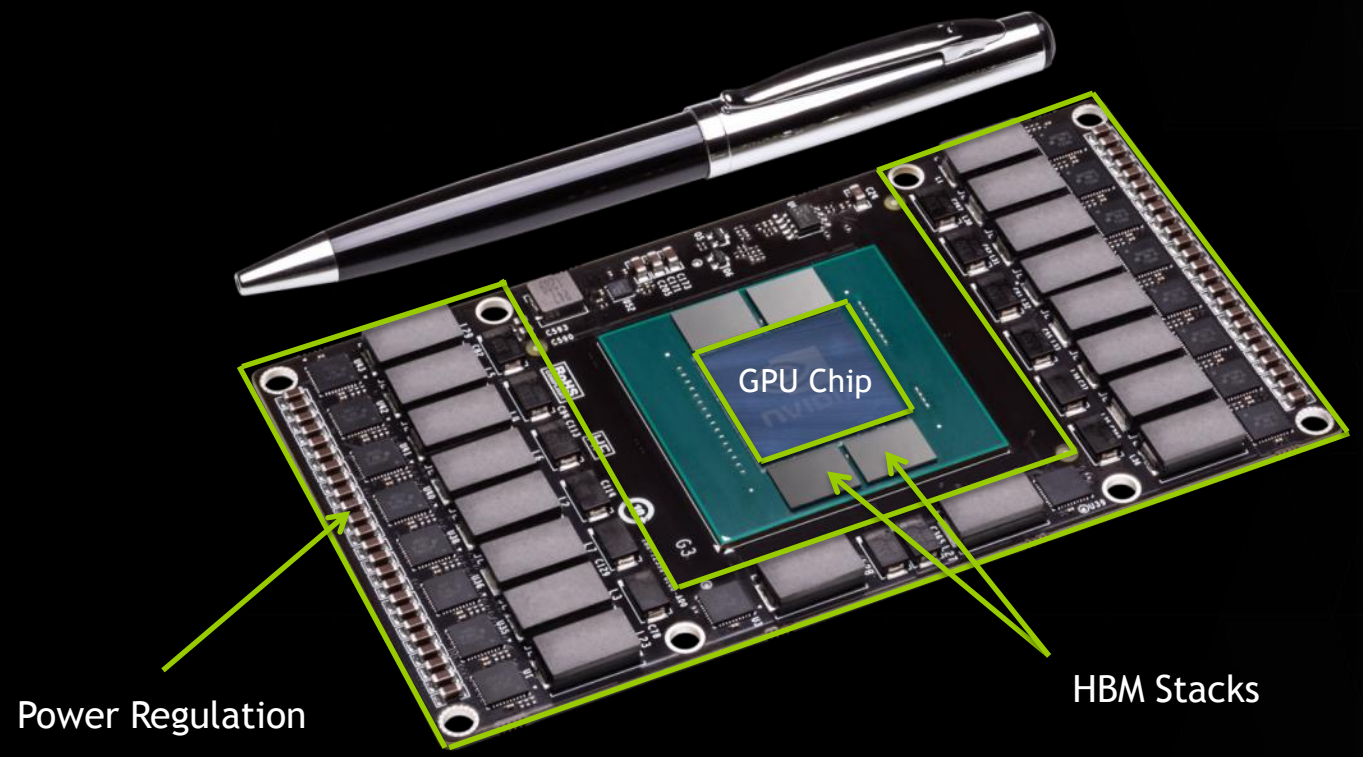






PASCAL

NVLink	5 to 12X PCIe 3.0
3D Memory	2 to 4X memory BW & size
Module	1/3 size of PCIe card



CUDA-ENABLED GPUS	CUDA DOWNLOADS	ACADEMIC PAPERS	UNIVERSITY COURSES
522M	2.5M	58K	770

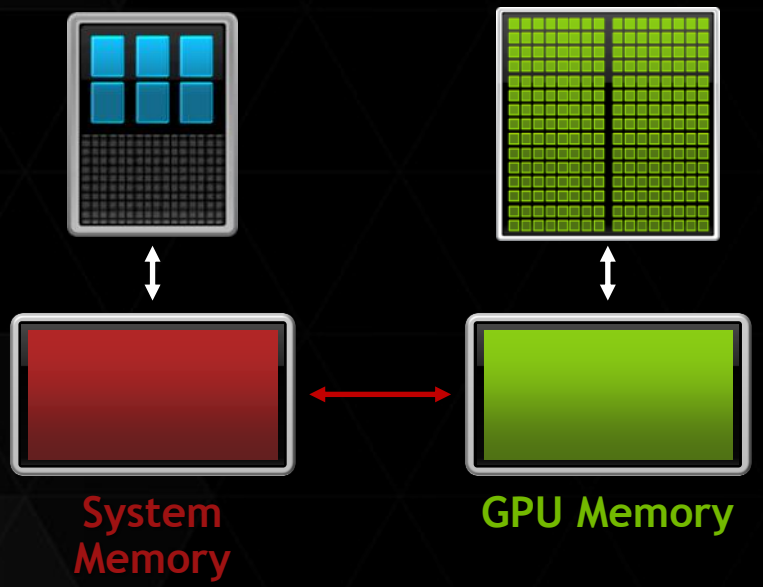
CUDA EVERWHERE

GOALS FOR THE CUDA PLATFORM

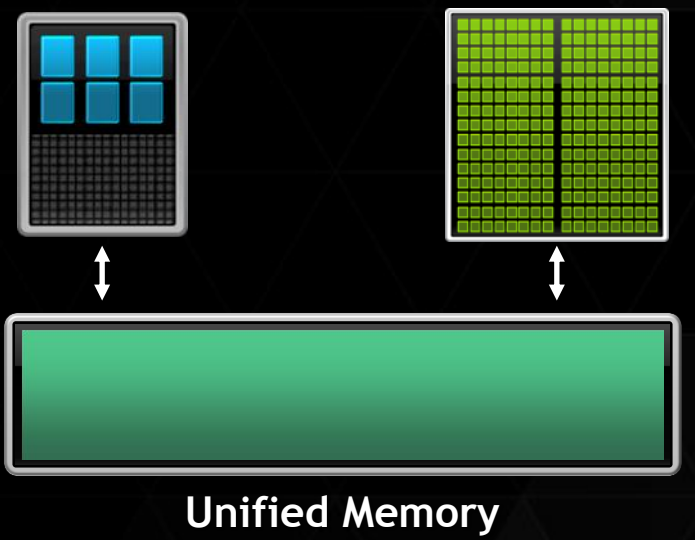
Simplicity	• Learn, adopt, & use parallelism with ease
Productivity	• Quickly achieve feature & performance goals
Portability	• Write code that can execute on all targets
Performance	• High absolute performance and scalability

UNIFIED MEMORY
DRAMATICALLY LOWER DEVELOPER EFFORT

Developer View Today

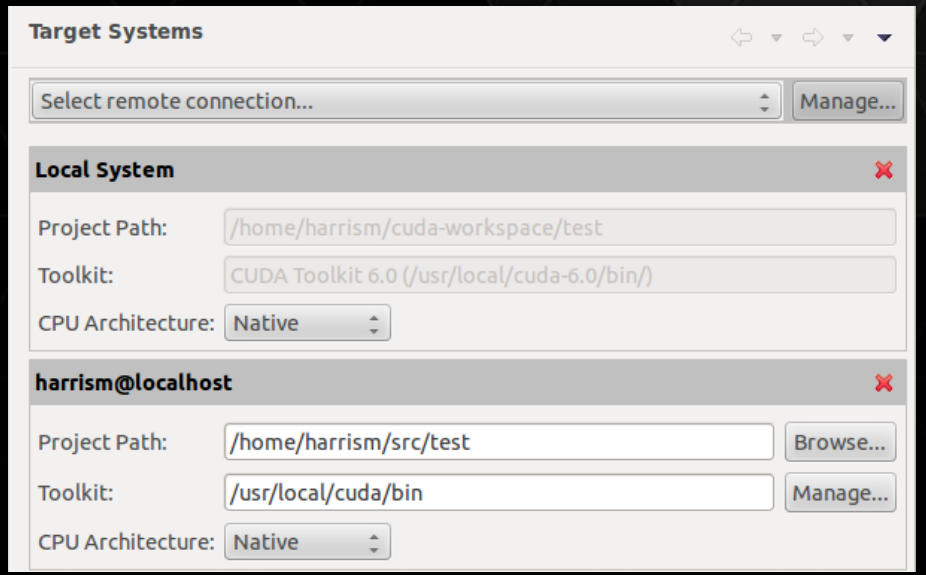


Developer View With Unified Memory

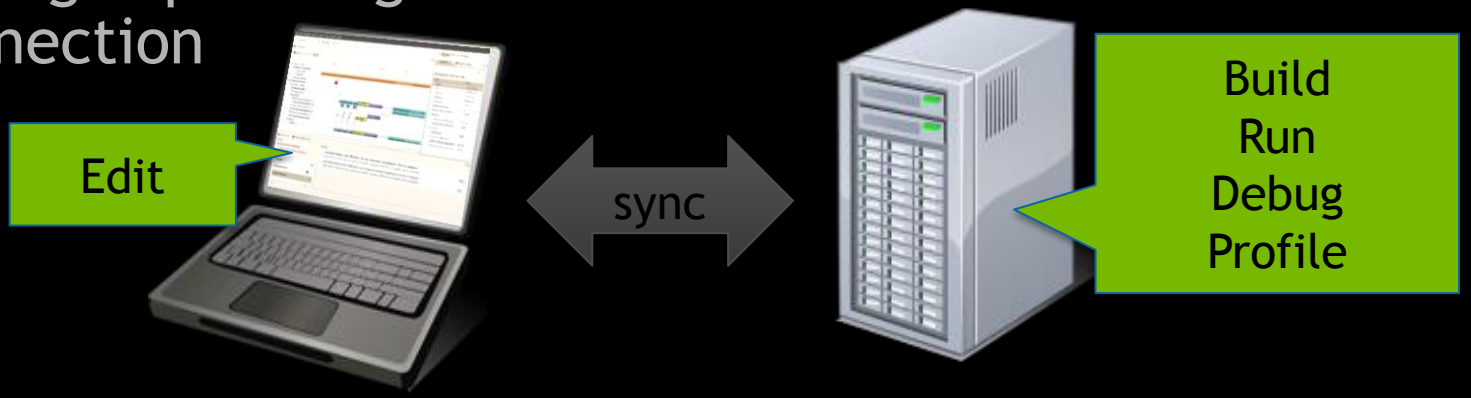


REMOTE DEVELOPMENT TOOLS

- Local IDE, remote application
  - Edit locally, build & run remotely
  - Automatic sync via ssh
  - Cross-compilation to ARM



- Full debugging & profiling via remote connection



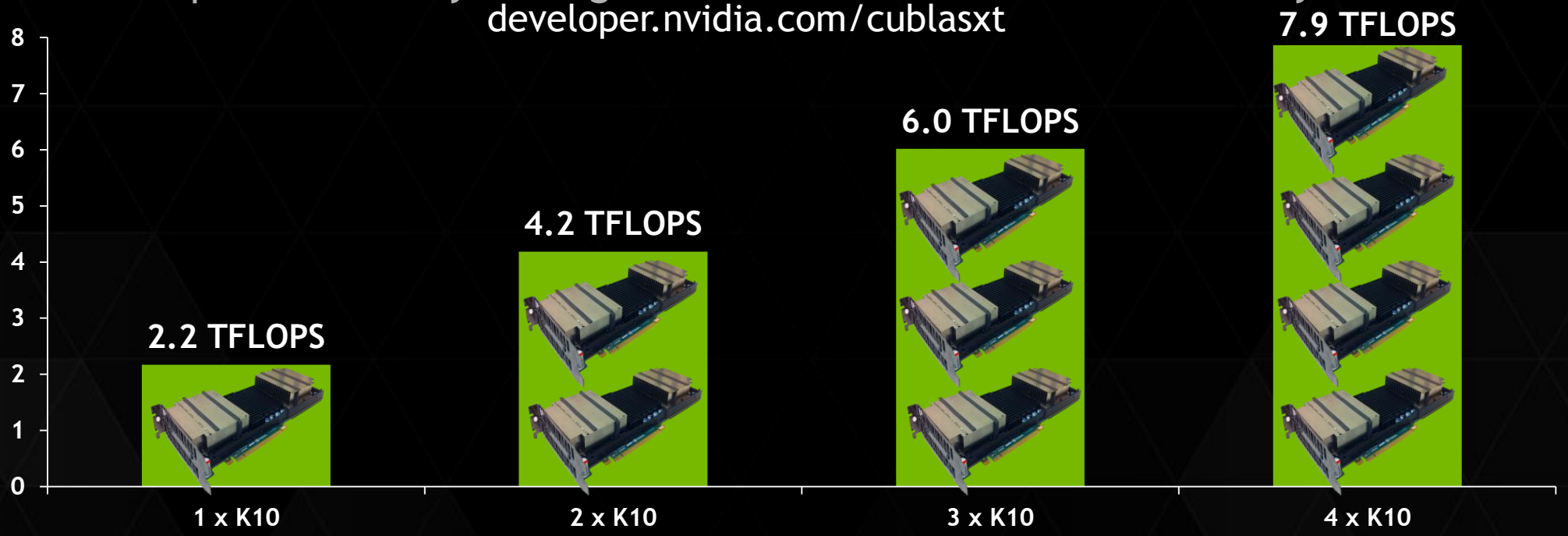
EXTENDED (XT) LIBRARY INTERFACES

Automatic Scaling to multiple GPUs per node

cuFFT 2D/3D & cuBLAS level 3

Operate directly on large datasets that reside in CPU memory

developer.nvidia.com/cublasxt



16K x 16K SGEMM on Tesla K10

GRID Graphics Accelerated VDI

The Original Graphics GPU Returns To The Data Center

NVIDIA Grid GPUs Power Enterprise Virtualization 2.0

IMPORTANCE OF A GPU

COMMERCIAL MARKETS

NEW OPPORTUNITY

25M

200M

400M

MARKET SERVED TODAY

100M

MUST HAVE

3D Engineering & Design Apps

AUTODESK.
SOLIDWORKS
PTC

PLM & Volume Design

Ps
AutoCAD
TEAMCENTER
AUTODESK.

Media Rich Web

Fl
HTML 5

INCREASINGLY NICE TO HAVE

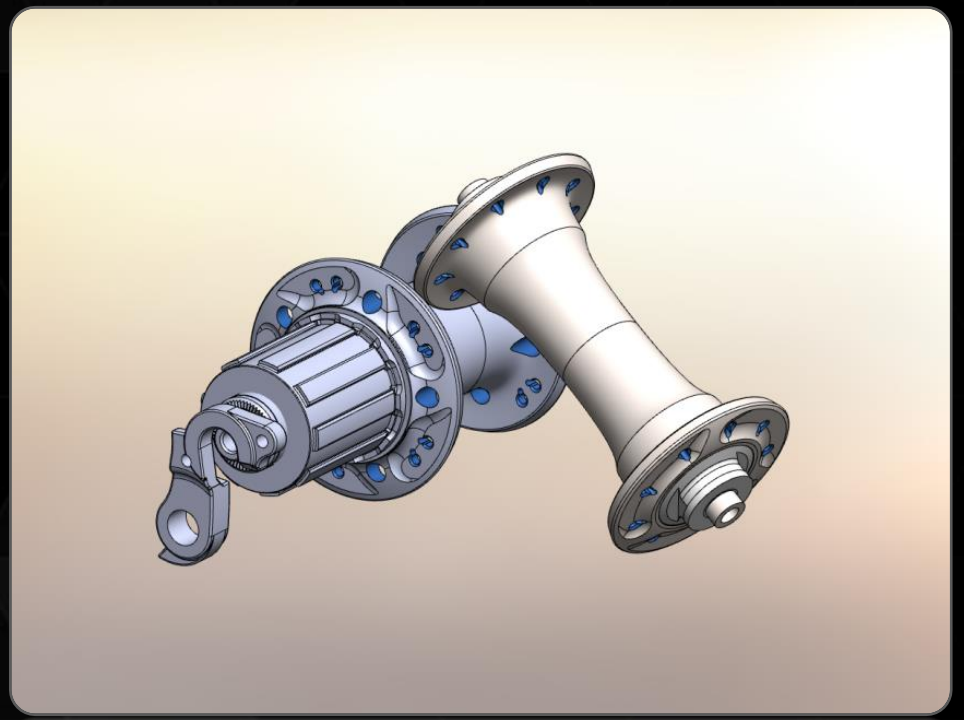
Office Productivity

Medical Records

Cerner
Epic
McKESSON
SIEMENS

NIGHT AND DAY DIFFERENCE

Without GPU



With GPU

GRID ACCELERATED GRAPHICS

Thank You