NVIDIA PASCAL INNOVATIONS

Carlo Nardone, Sen. Solution Architect HPC/DL, Enterprise EMEA

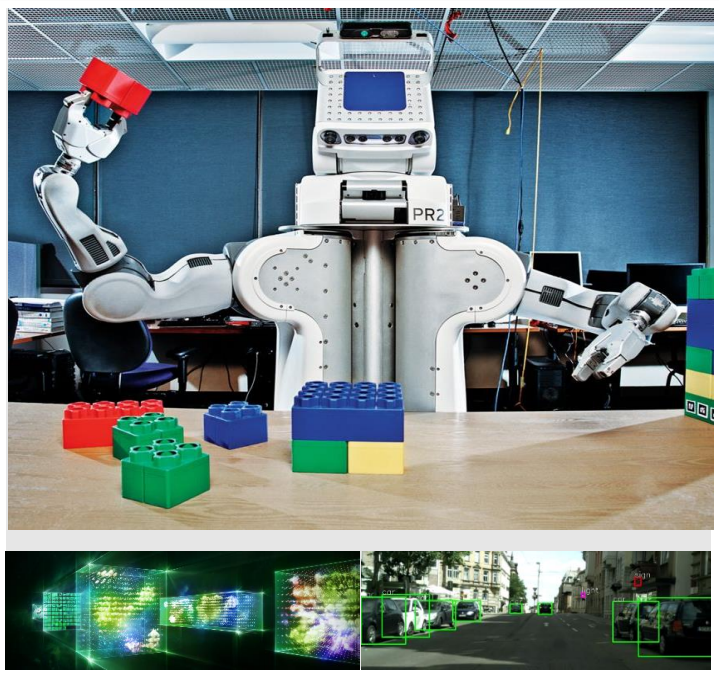
NVIDIA

HPC ADMINTECH 2017

NVIDIA
“THE AI COMPUTING COMPANY”

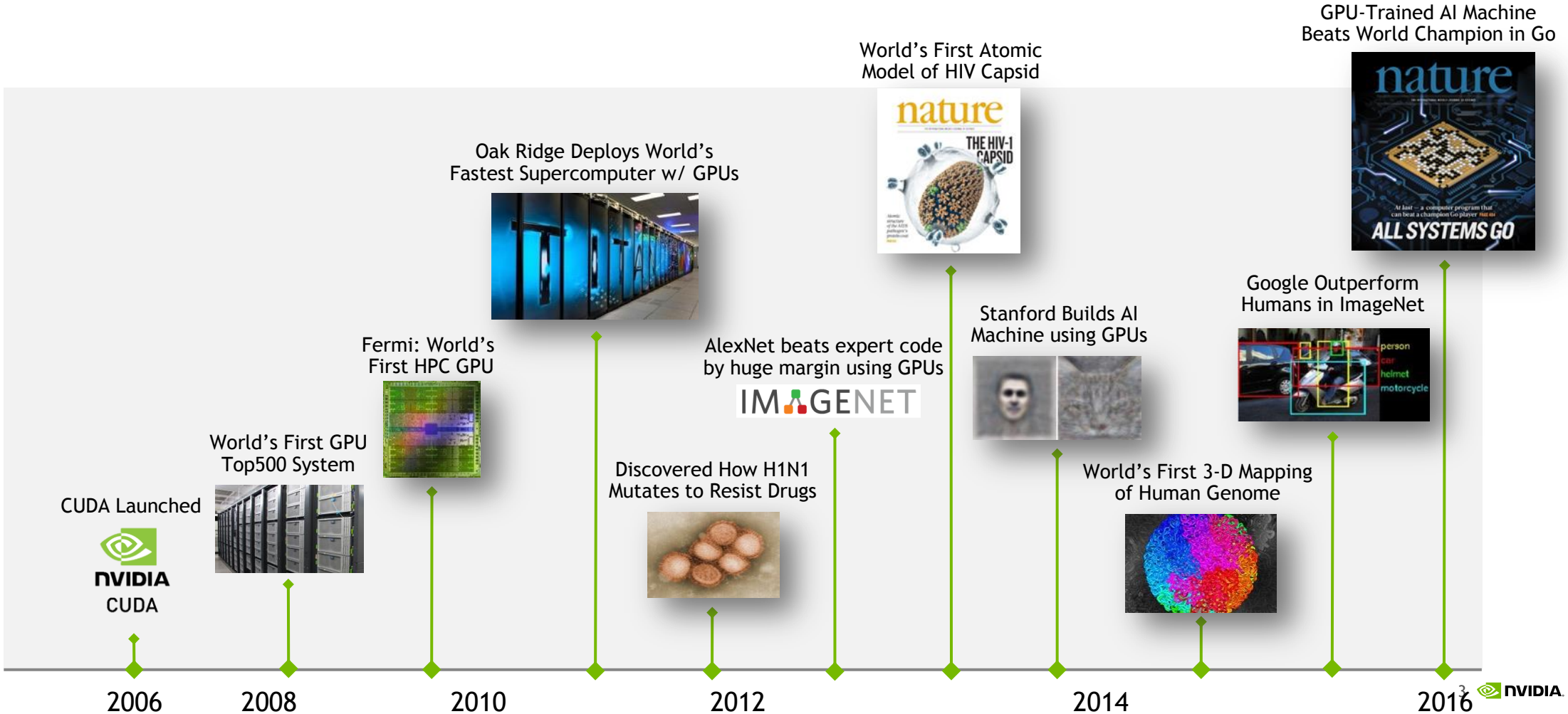
Computer Graphics

GPU Computing



Artificial Intelligence

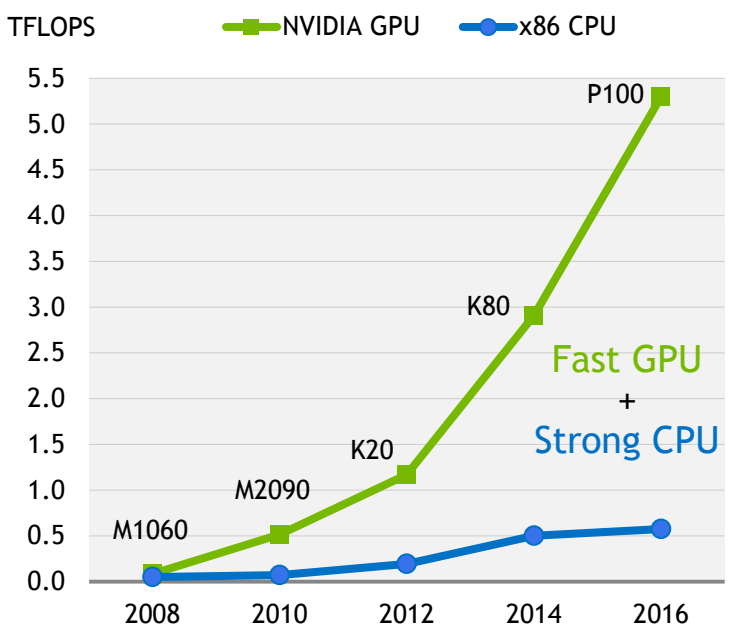
TEN YEARS OF GPU COMPUTING



TESLA ACCELERATED COMPUTING PLATFORM

Focused on Co-Design for Accelerated Data Center

Fast GPU Engineered for High Throughput



Productive Programming Model & Tools

NVIDIA CUDA

OpenACC
Directives for Accelerators

Expert Co-Design

APPLICATION

MIDDLEWARE

SYS SW

LARGE SYSTEMS

PROCESSOR

Accessibility

阿里云计算 Alibaba Cloud Computing

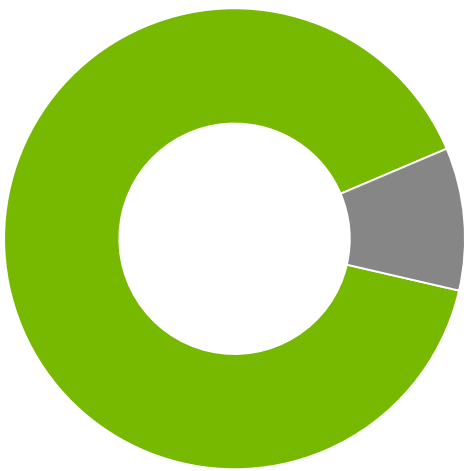
amazon web services

IBM SOFTLAYER

Microsoft Azure

70% OF TOP HPC APPS ACCELERATED

INTERSECT360 SURVEY OF TOP APPS



9 of top 10
Apps Accelerated

35 of top 50
Apps Accelerated

Intersect360, Nov 2015
"HPC Application Support for GPU Computing"

TOP 25 APPS IN SURVEY

- GROMACS
- SIMULIA Abaqus
- NAMD
- AMBER
- ANSYS Mechanical
- Exelis IDL
- MSC NASTRAN

- LAMMPS
- NWChem
- LS-DYNA
- Schrodinger

- Gaussian
- GAMESS

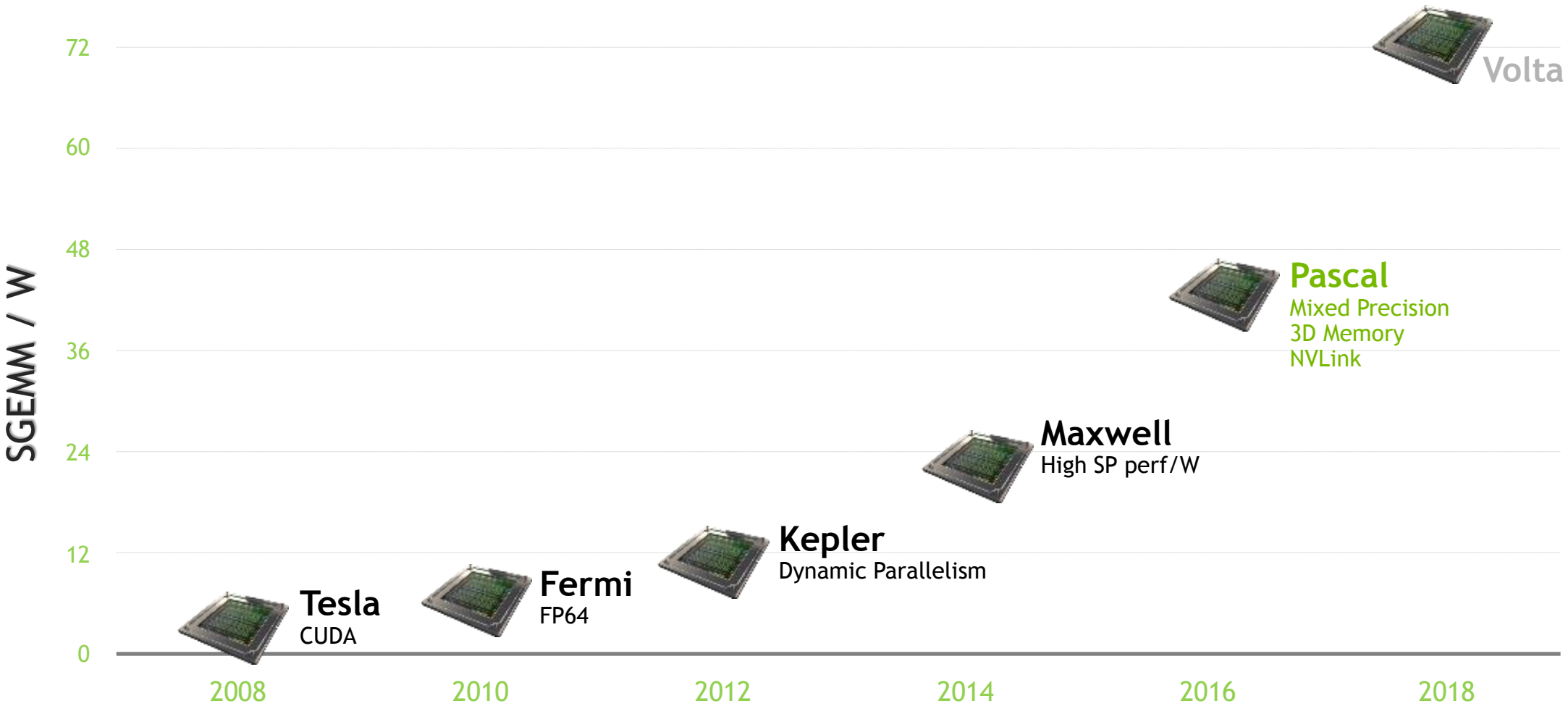
- ANSYS Fluent
- WRF
- VASP
- OpenFOAM
- CHARMM
- Quantum Espresso

- ANSYS CFX
- Star-CD
- CCSM
- COMSOL
- Star-CCM+
- BLAST

= All popular functions accelerated
= Some popular functions accelerated
= In development
= Not supported

PASCAL ARCHITECTURE: TESLA P100

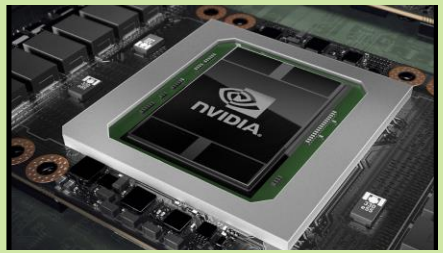
GPU ARCHITECTURES ROADMAP



INTRODUCING TESLA P100

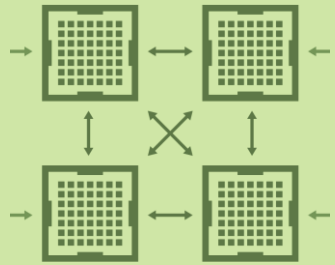
New GPU Architecture to Enable the World's Fastest Compute Node

Pascal Architecture



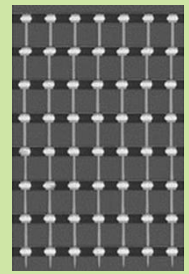
Highest Compute Performance

NVLink



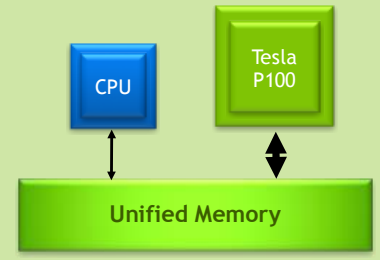
GPU Interconnect for Maximum Scalability

CoWoS HBM2

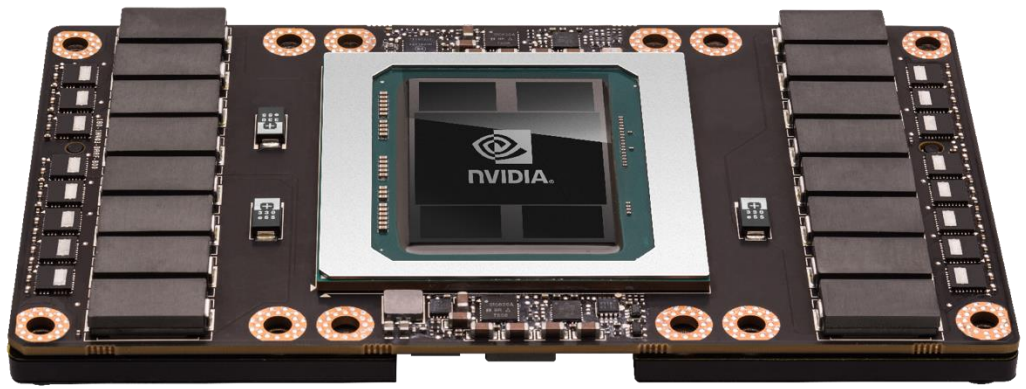


Unifying Compute & Memory in Single Package

Page Migration Engine



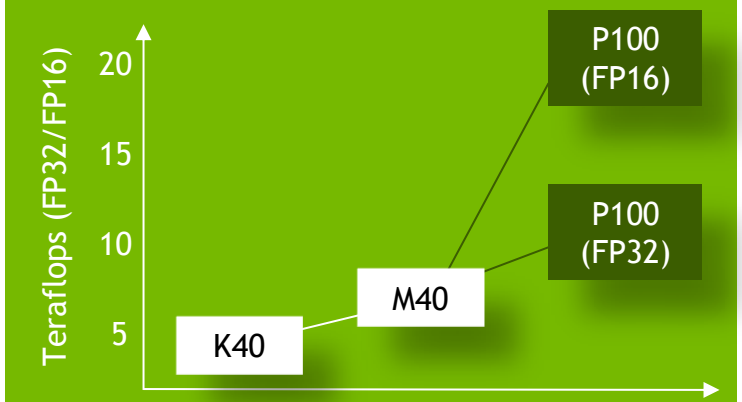
Simple Parallel Programming with Virtually Unlimited Memory Space



GIANT LEAPS IN EVERYTHING

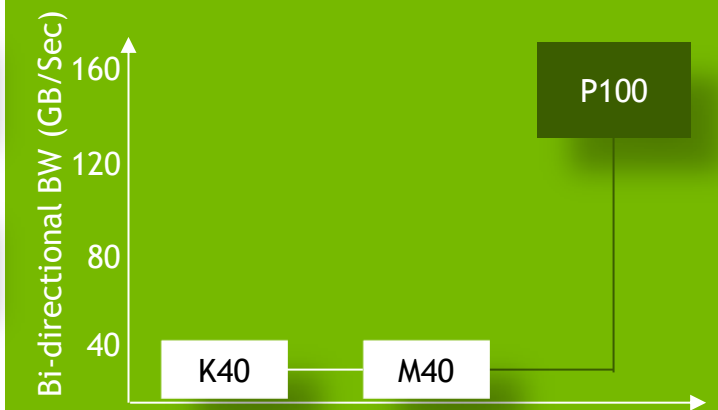
PASCAL ARCHITECTURE

21 Teraflops of FP16 for Deep Learning



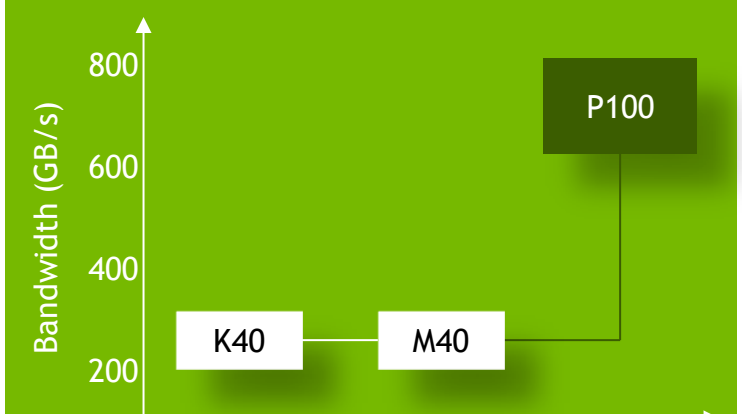
NVLINK

5x GPU-GPU Bandwidth



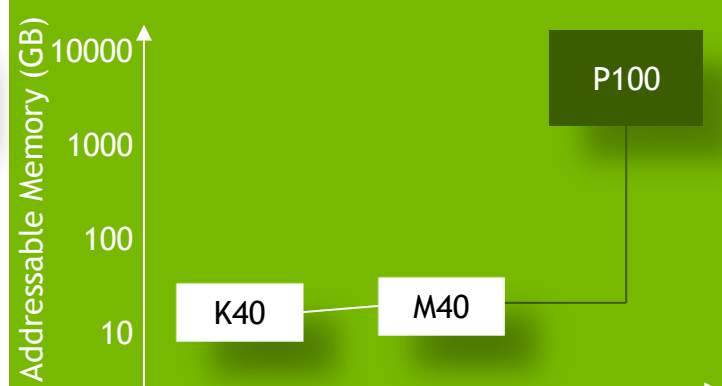
CoWoS HBM2 Stacked Mem

3x Higher for Massive Data Workloads



PAGE MIGRATION ENGINE

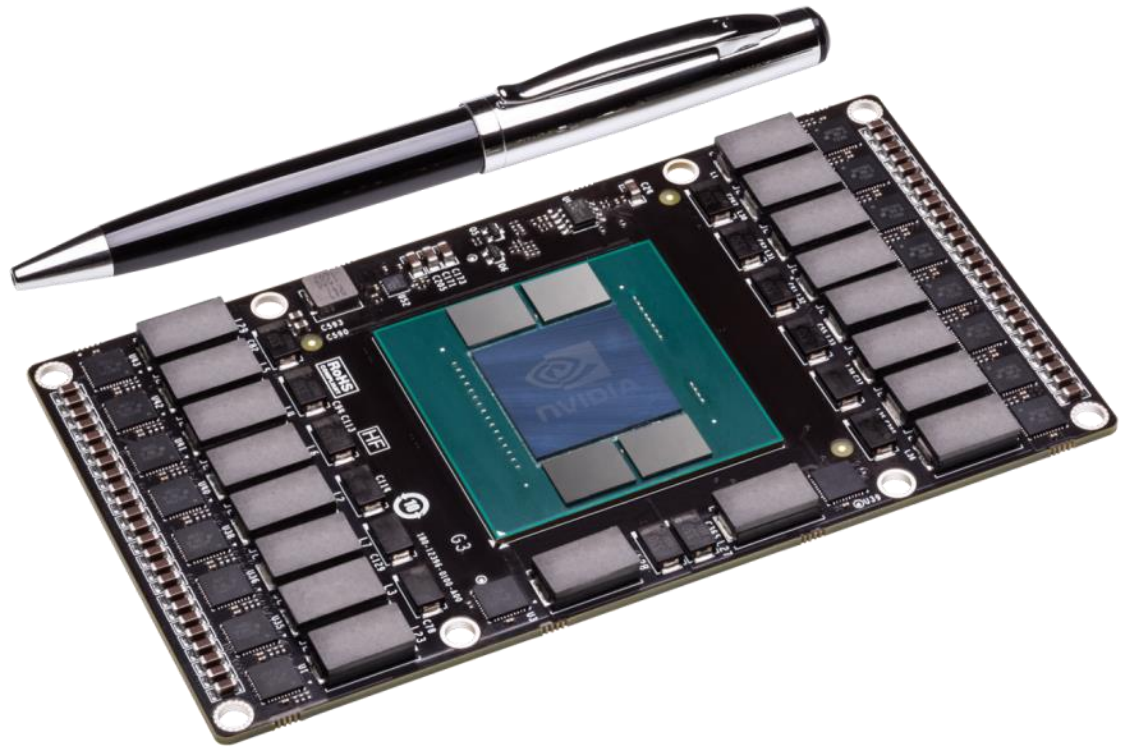
Virtually Unlimited Memory Space



SXM2

3x Performance Density

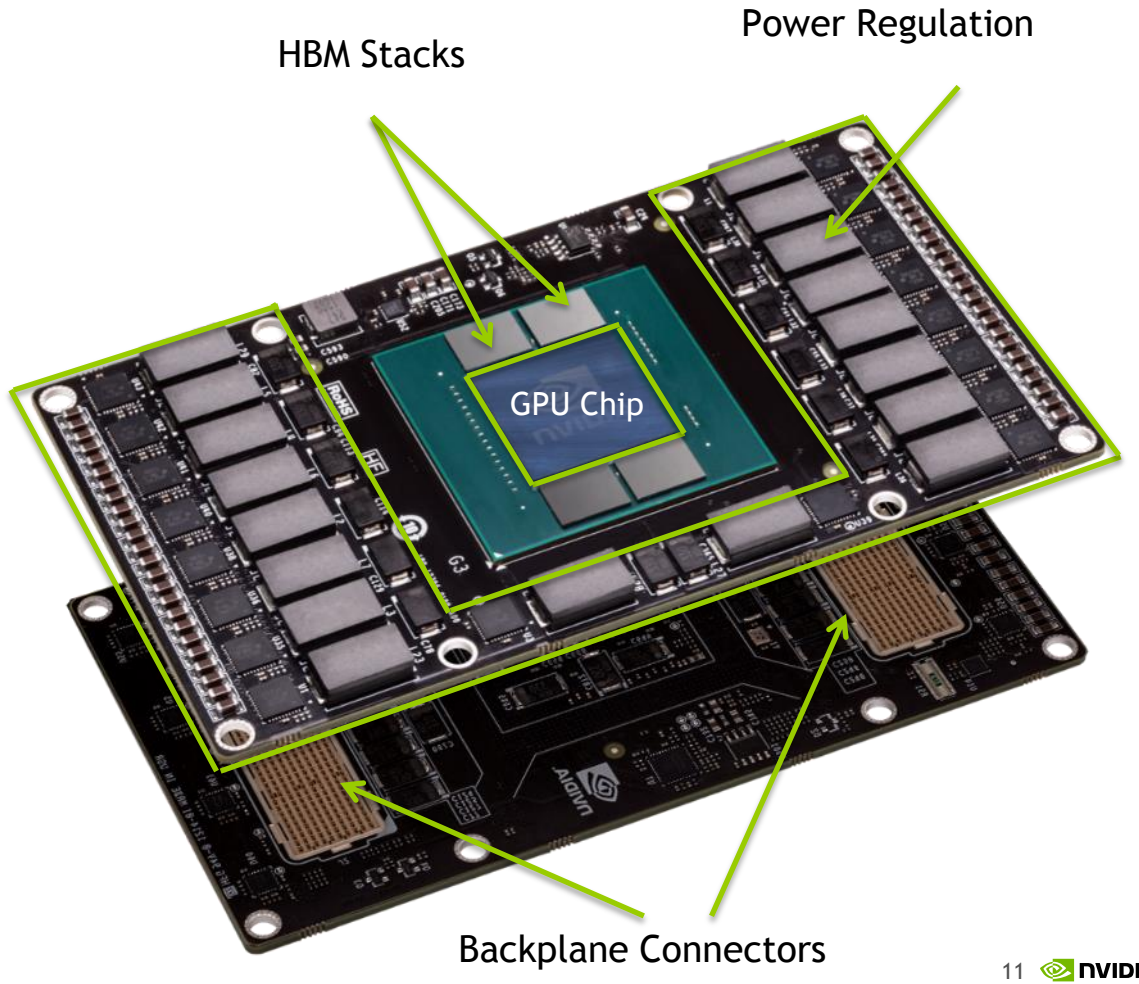
NVLINK	5 to 12X PCIe 3.0
HBM	2x-4X memory BW at same power
Size	1/3 size of PCIe card 80mm x 140mm



SXM2

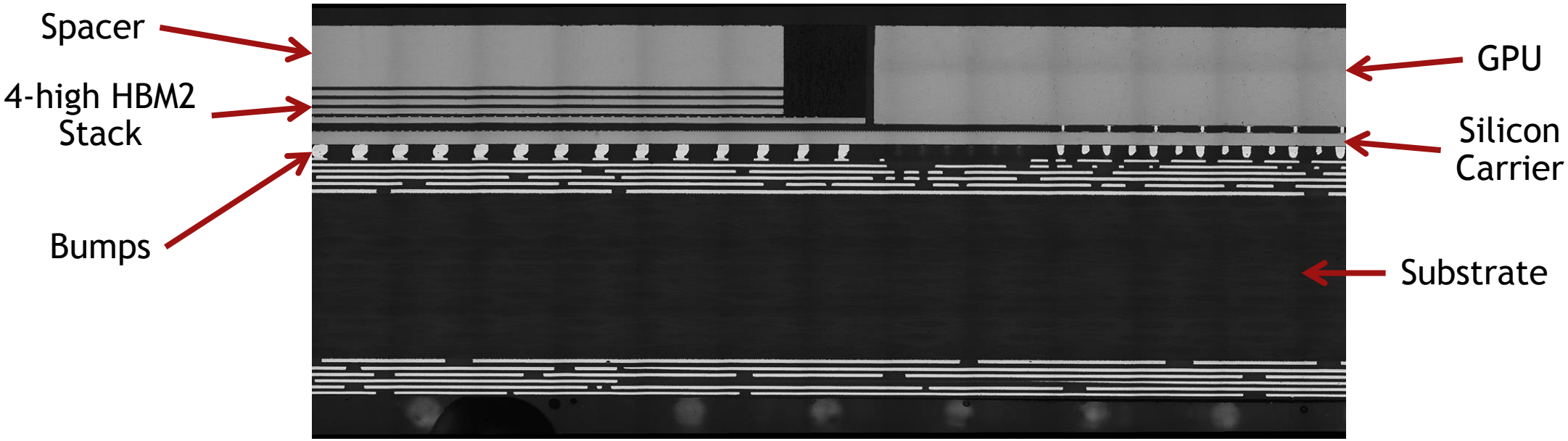
3x Performance Density

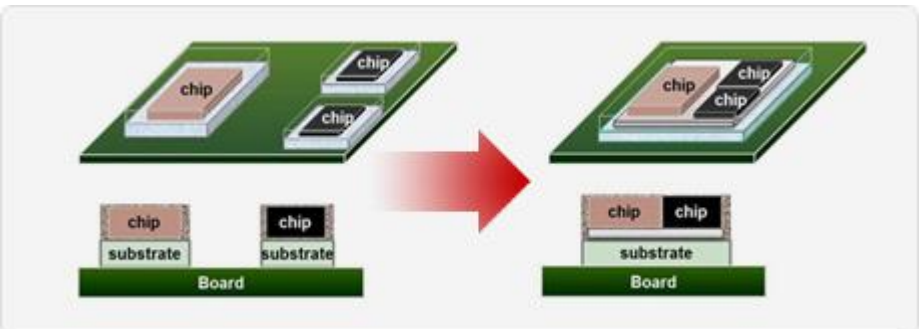
NVLINK	5 to 12X PCIe 3.0
HBM	2x-4X memory BW at same power
Size	1/3 size of PCIe card 80mm x 140mm



COWOS HBM2 STACKED MEMORY

720 GB/s Bandwidth, ECC for free





CoWoS (Chip on Wafer on Substrate) is ® by TSMC

GP100 ARCHITECTURE

56 SMs

3584 CUDA Cores

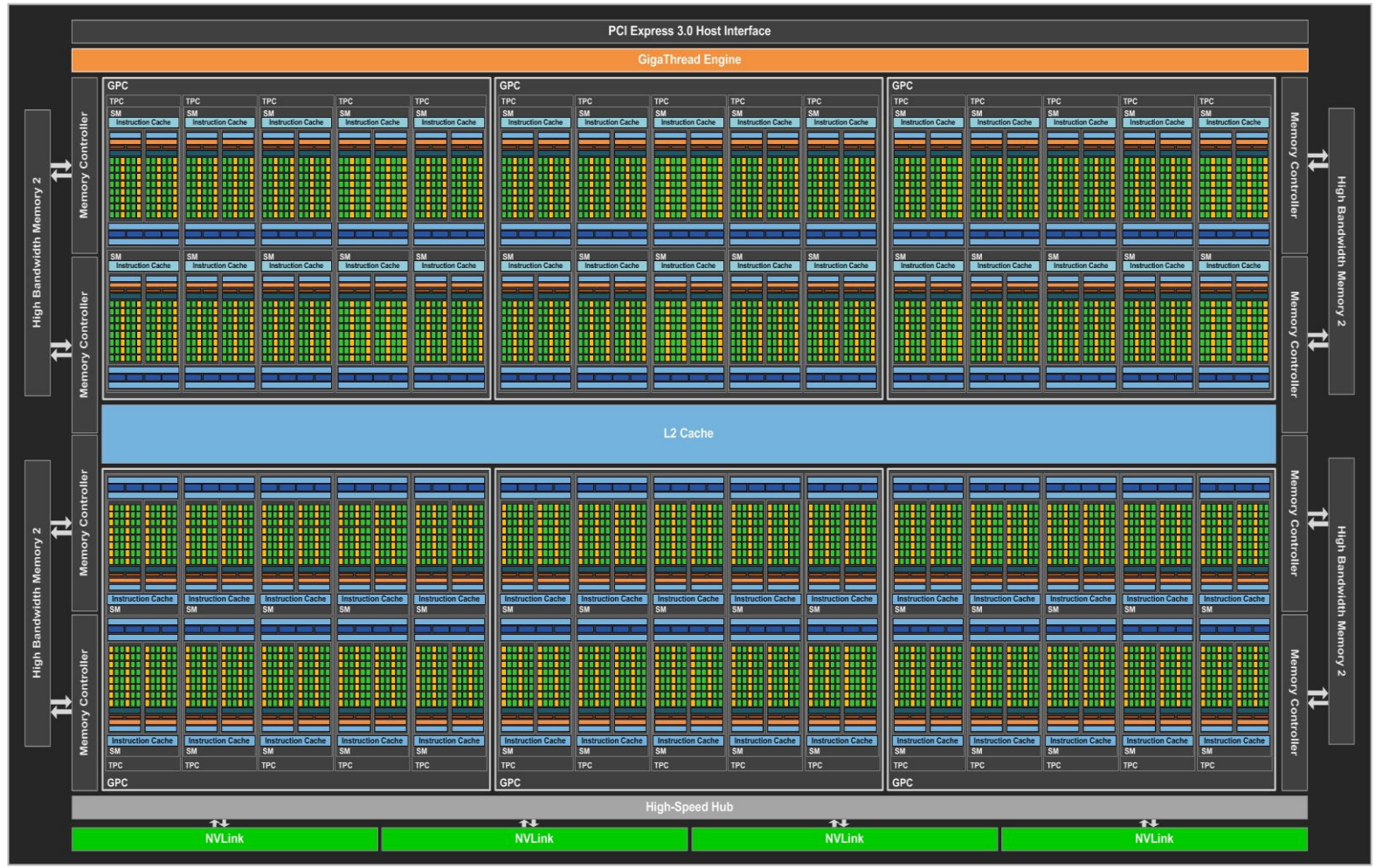
5.3 TF Double Precision

10.6 TF Single Precision

21.2 TF Half Precision

16 GB HBM2

720 GB/s Bandwidth



GP100 SM

	GP100
CUDA Cores	64
Register File	256 KB
Shared Memory	64 KB
Active Threads	2048
Active Blocks	32

IEEE 754 FLOATING POINT ON GP100

3 sizes, 3 speeds, all fast

Feature	New Half precision	Single precision	Double precision
Layout	s5.10	s8.23	s11.52
Issue rate	pair every clock	1 every clock	1 every 2 clocks
Subnormal support	Yes	Yes	Yes
Atomic Addition	Yes	Yes	New Yes

FP16: HALF-PRECISION FLOATING POINT

- 16 bits

  - 1 sign bit, 5 exponent bits, 10 fraction bits
- 2^{40} Dynamic range
  - Normalized values: 1024 values for each power of 2, from 2^{-14} to 2^{15}
  - Subnormals at full speed: 1024 values from 2^{-24} to 2^{-15}
- Special values
  - +- Infinity, Not-a-number

USE CASES
Deep Learning Training
Radio Astronomy
Sensor Data
Image Processing

NVLINK

NVLINK

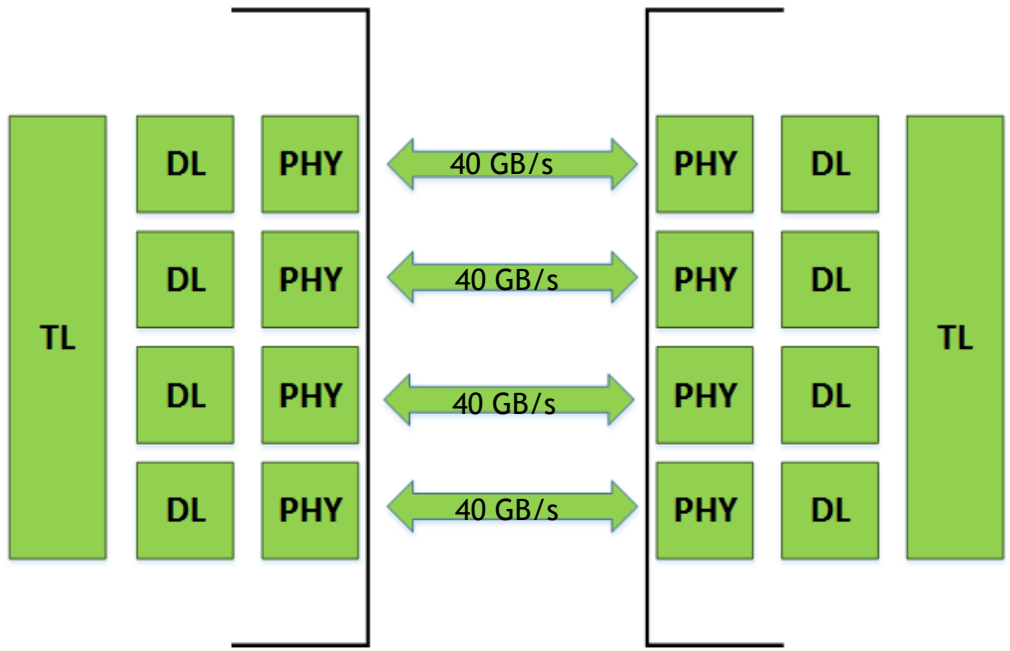
P100 supports 4 NVLinks

Up to 94% bandwidth efficiency

Supports read/writes/atomics to peer GPU

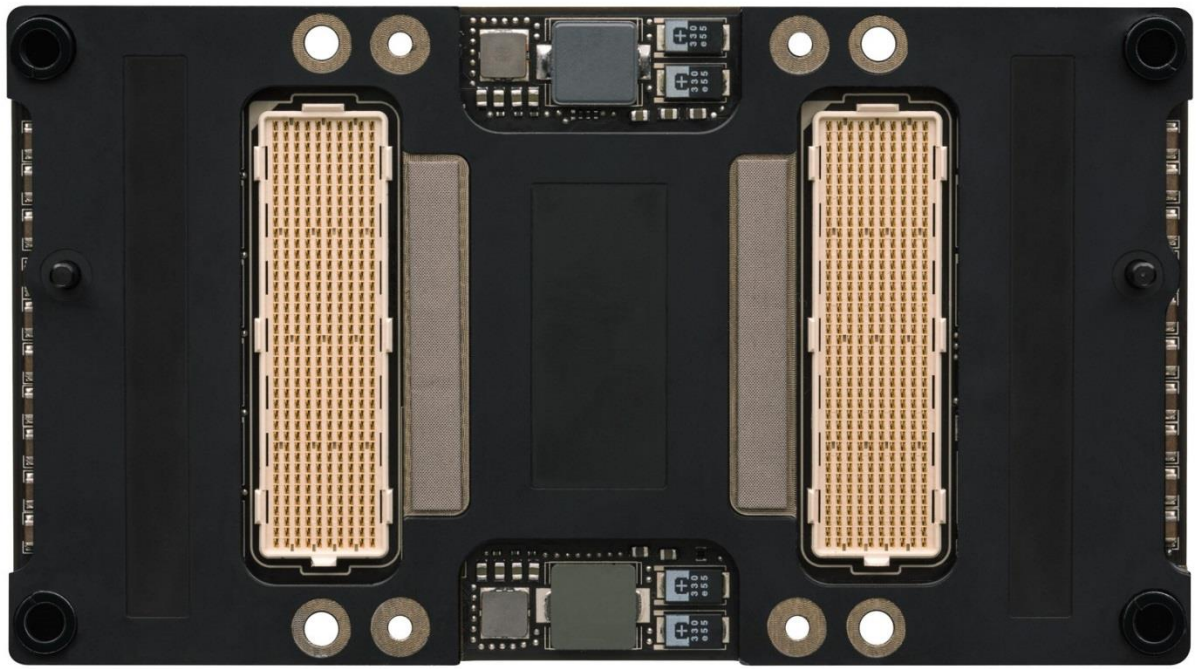
Supports read/write access to NVLink-enabled CPU

Links can be ganged for higher bandwidth



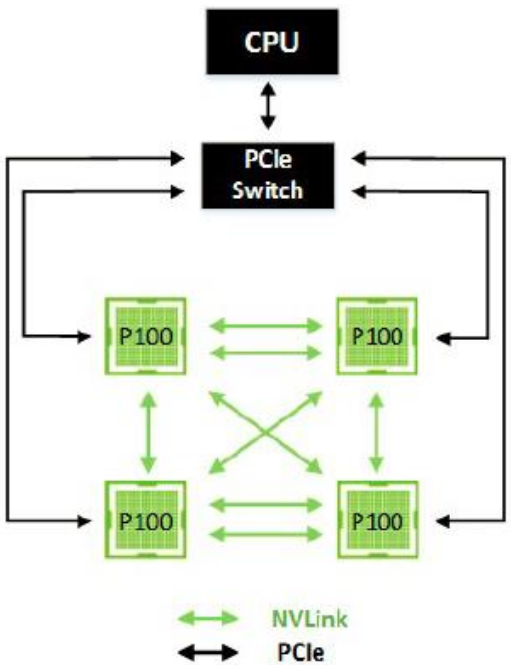
NVLink on Tesla P100

NVLINK CONNECTORS IN TESLA P100



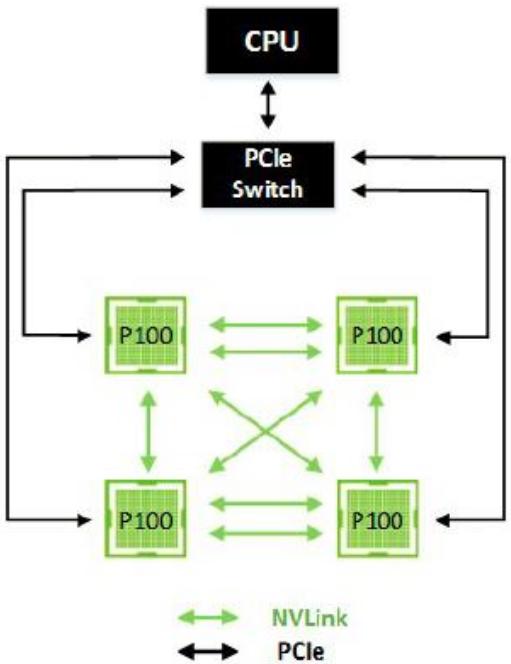
NVLINK TOPOLOGIES

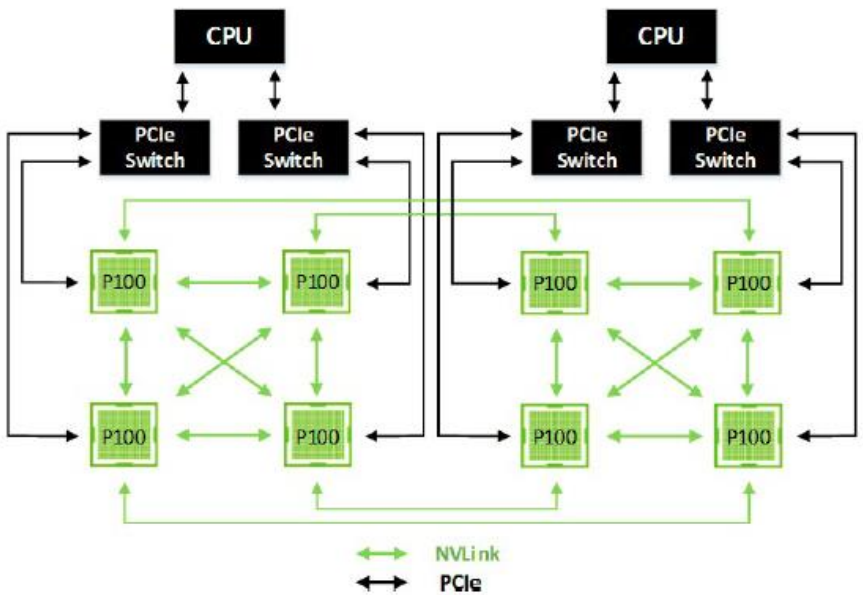
GPU-GPU and GPU-CPU



NVLINK TOPOLOGIES

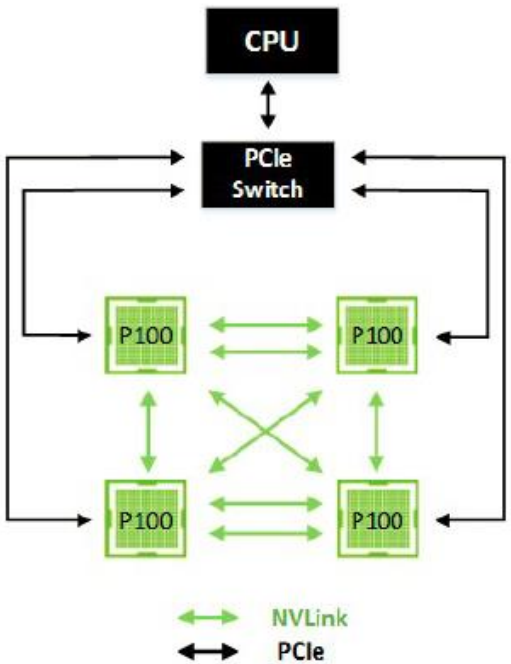
GPU-GPU and GPU-CPU

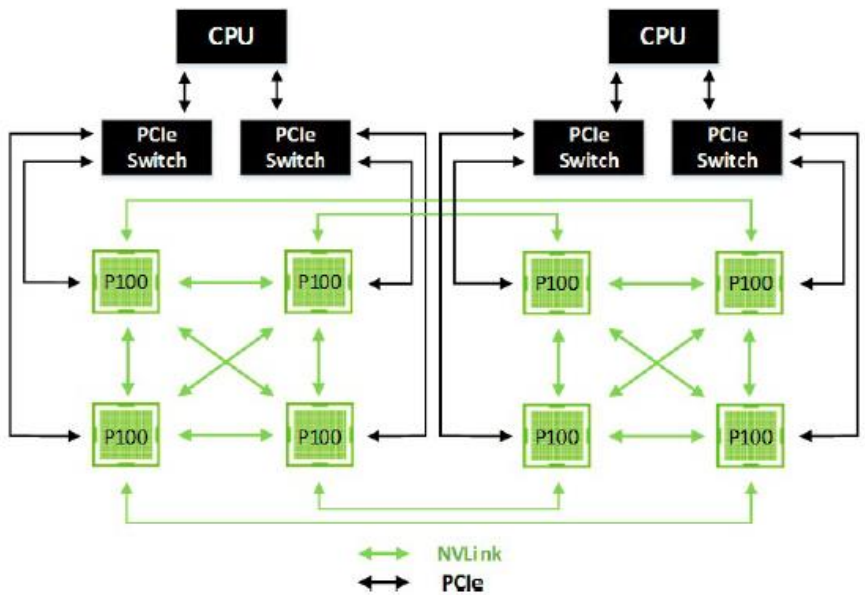


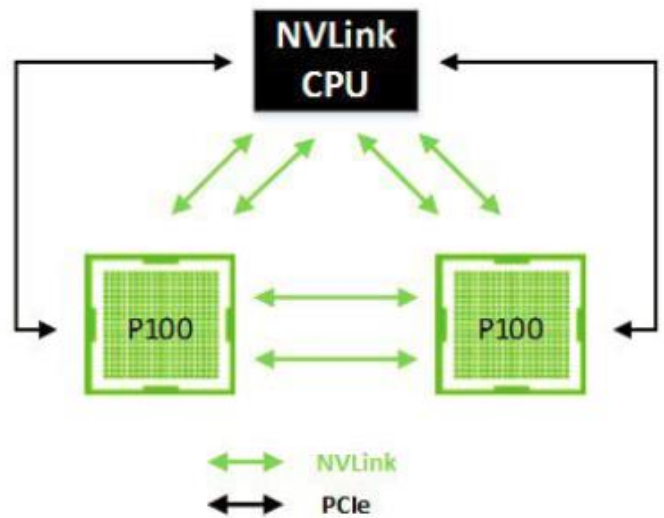


NVLINK TOPOLOGIES

GPU-GPU and GPU-CPU







NVLINK GPU-CPU

IBM Power Systems Server S822LC (codename “Minsky”)

2x IBM Power8+ CPUs and 4x P100 GPUs

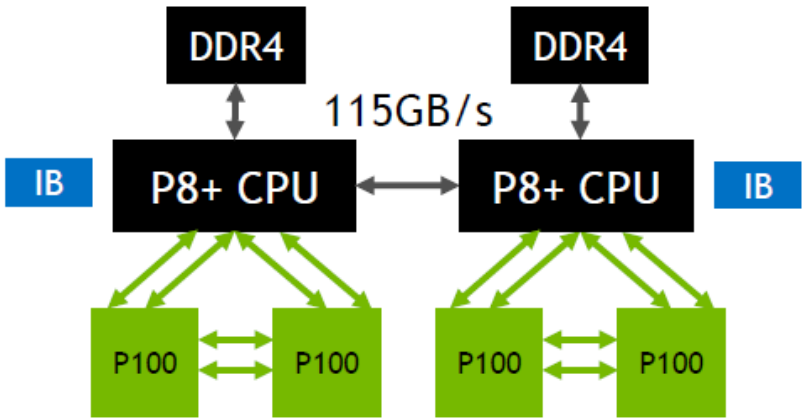
80 GB/s per GPU bidirectional for peer traffic

80 GB/s per GPU bidirectional to CPU

115 GB/s CPU Memory Bandwidth

Direct Load/store access to CPU Memory

High Speed Copy Engines for bulk data movement



UNIFIED MEMORY

PAGE MIGRATION ENGINE

Support Virtual Memory Demand Paging

49-bit Virtual Addresses

- Sufficient to cover 48-bit CPU address + all GPU memory

GPU page faulting capability

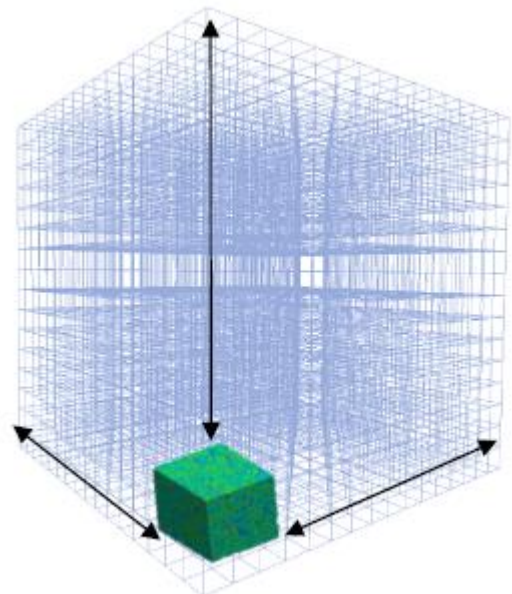
- Can handle thousands of simultaneous page faults

Up to 2 MB page size

- Better TLB coverage of GPU memory

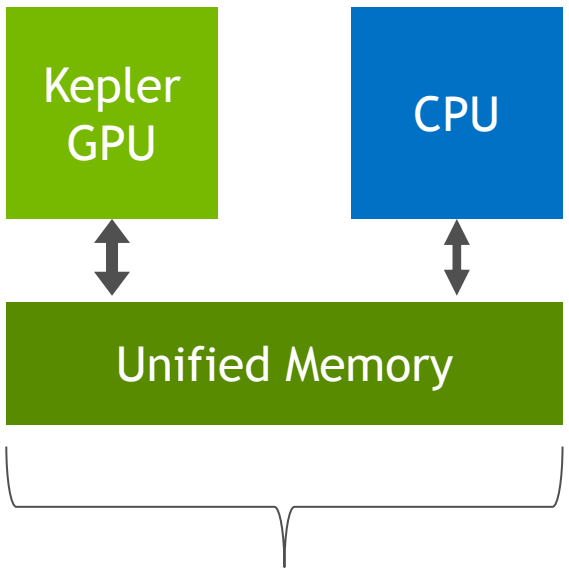
GPU Memory oversubscription now possible with Pascal

- Combustion, Quantum Chemistry, ray tracing, graph analytics …



KEPLER / MAXWELL UNIFIED MEMORY

CUDA 6+



Allocate Up To
GPU Memory Size

Simpler Programming & Memory Model

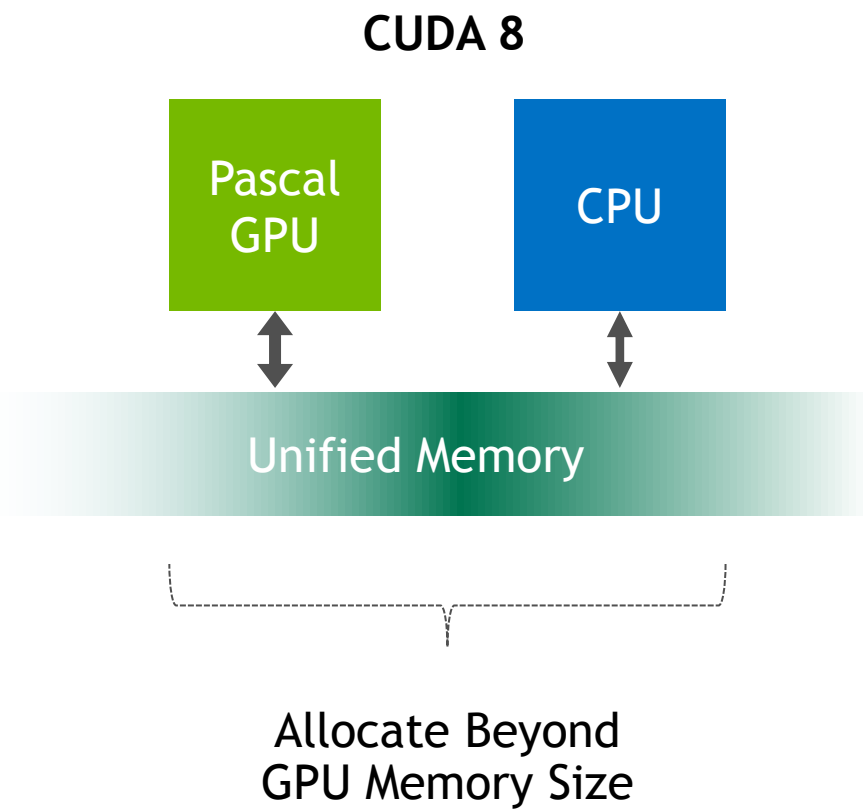
- Single allocation, single pointer, accessible anywhere
- Eliminate need for *explicit copy*
- Greatly simplifies code porting

Performance Through Data Locality

- Migrate data to accessing processor
- Guarantee global coherency
- Still allows explicit hand tuning

PASCAL UNIFIED MEMORY

Large datasets, simple programming, High Performance



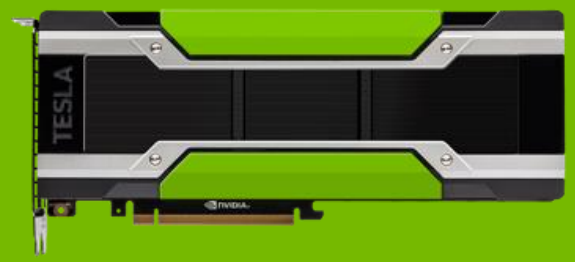
Enable Large Data Models	Oversubscribe GPU memory Allocate up to system memory size
Tune Unified Memory Performance	Usage hints via cudaMemAdvise API Explicit prefetching API
Simpler Data Access	CPU/GPU Data coherence Unified memory atomic operations

NVIDIA TESLA LINEUP

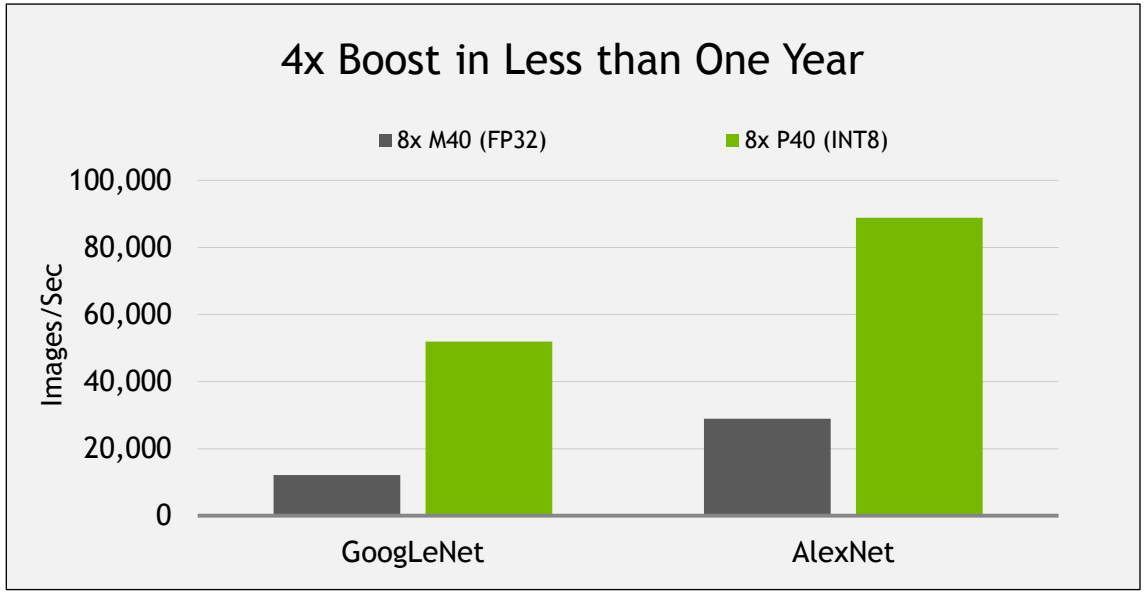
TESLA P100 ACCELERATOR

	
Compute	5.3 TF DP · 10.6 TF SP · 21.2 TF HP
Memory	HBM2: 720 GB/s · 16 GB
Interconnect	NVLink (up to 8 way) + PCIe Gen3
Programmability	Page Migration Engine Unified Memory
Availability	DGX-1: Order Now Cray, Dell, HP, IBM: Q1 2017

TESLA P40



Highest Throughput for Scale-up Servers

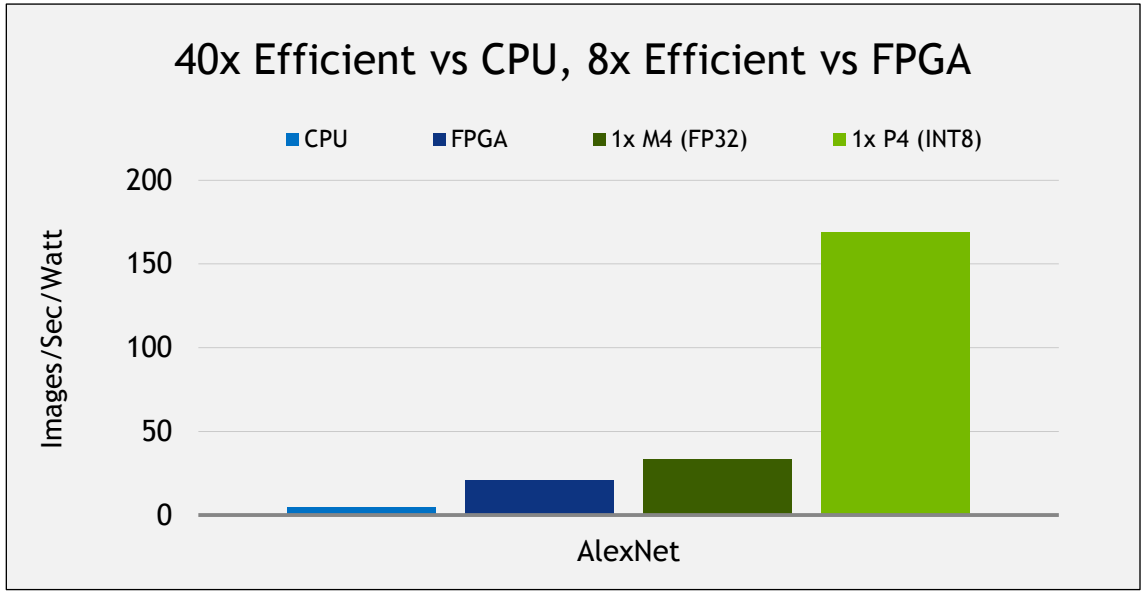


	P40
# of CUDA Cores	3840
Peak Single Precision	12 TeraFLOPS
Peak INT8	47 TOPS
Low Precision	4x 8-bit vector dot product with 32-bit accumulate
Video Engines	1x decode engine, 2x encode engines
GDDR5 Memory	24 GB @ 346 GB/s
Power	250W

GoogLeNet, AlexNet, batch size = 128, CPU: Dual Socket Intel E5-2697v4

TESLA P4

Maximum Efficiency for Scale-out Servers

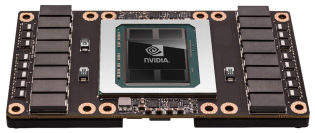


P4	
# of CUDA Cores	2560
Peak Single Precision	5.5 TeraFLOPS
Peak INT8	22 TOPS
Low Precision	4x 8-bit vector dot product with 32-bit accumulate
Video Engines	1x decode engine, 2x encode engine
GDDR5 Memory	8 GB @ 192 GB/s
Power	50W & 75 W

AlexNet, batch size = 128, CPU: Intel E5-2690v4 using Intel MKL 2017, FPGA is Arria10-115
1x M4/P4 in node, P4 board power at 56W, P4 GPU power at 36W, M4 board power at 57W, M4 GPU power at 39W, Perf/W chart using GPU power

END-TO-END ENTERPRISE PRODUCT FAMILY

HYPERSCALE HPC

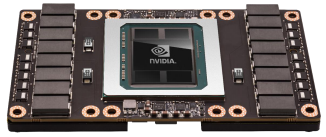


Training - Tesla P100

Inference - Tesla P40 & P4

Hyperscale deployment for deep learning training & inference

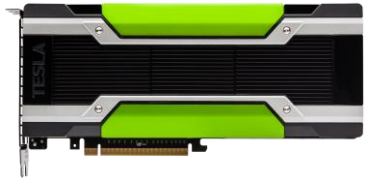
STRONG-SCALE HPC



Tesla P100 with NVLink

Data centers running HPC and DL apps scaling to multiple GPUs

MIXED-APPS HPC



Tesla P100 with PCI-E

HPC data centers running mix of CPU and GPU workloads

FULLY INTEGRATED DL SUPERCOMPUTER

DGX-1

For customers who need to get going now with fully integrated solution

TESLA PRODUCTS DECODER

	K80	M40	M4	P100 (SXM2)	P100 (PCIE)	P40	P4
GPU	2x GK210	GM200	GM206	GP100	GP100	GP102	GP104
PEAK FP64 (TFLOPs)	2.9	NA	NA	5.3	4.7	NA	NA
PEAK FP32 (TFLOPs)	8.7	7	2.2	10.6	9.3	12	5.5
PEAK FP16 (TFLOPs)	NA	NA	NA	21.2	18.7	NA	NA
PEAK TIOPs	NA	NA	NA	NA	NA	47	22
Memory Size	2x 12GB GDDR5	24 GB GDDR5	4 GB GDDR5	16 GB HBM2	16/12 GB HBM2	24 GB GDDR5	8 GB GDDR5
Memory BW	480 GB/s	288 GB/s	80 GB/s	732 GB/s	732/549 GB/s	346 GB/s	192 GB/s
Interconnect	PCIe Gen3	PCIe Gen3	PCIe Gen3	NVLINK + PCIe Gen3	PCIe Gen3	PCIe Gen3	PCIe Gen3
ECC	Internal + GDDR5	GDDR5	GDDR5	Internal + HBM2	Internal + HBM2	GDDR5	GDDR5
Form Factor	PCIE Dual Slot	PCIE Dual Slot	PCIE LP	SXM2	PCIE Dual Slot	PCIE Dual Slot	PCIE LP
Power	300 W	250 W	50-75 W	300 W	250 W	250 W	50-75 W

DGX-1 SUPERCOMPUTER

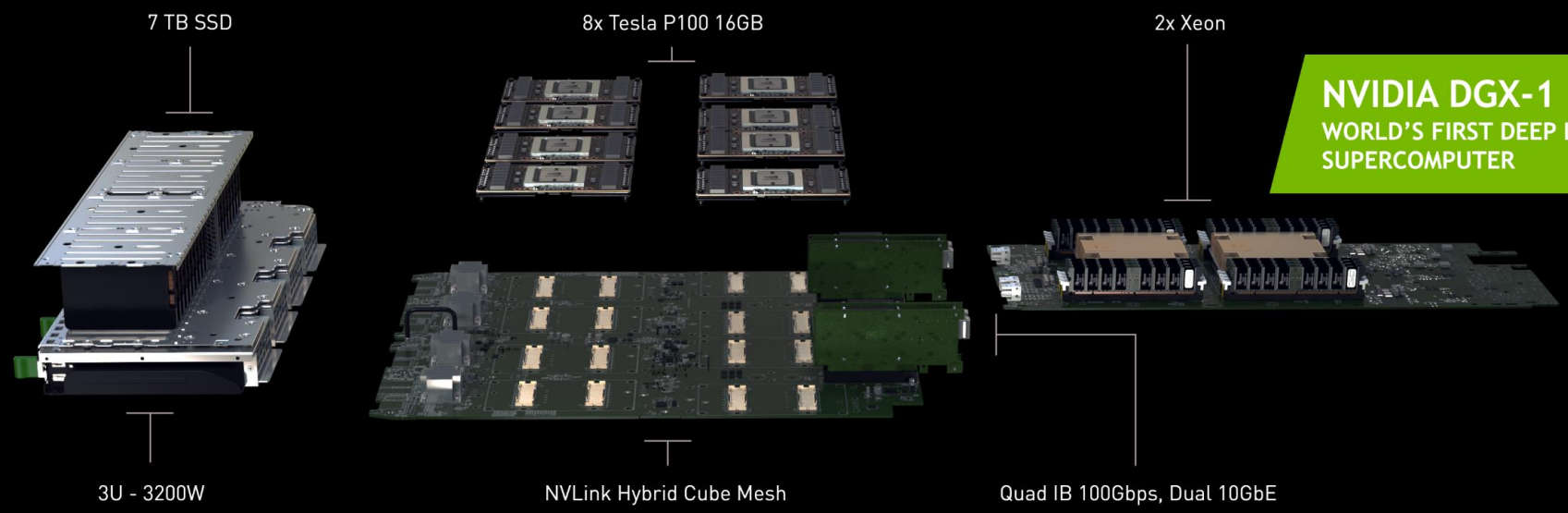
NVIDIA DGX-1

AI Supercomputer-in-a-Box designed for the Data Center

170 TFLOPS | 8x Tesla P100 16GB | NVLink Hybrid Cube Mesh

2x Xeon | 8 TB SSD RAID 0 | Quad IB 100Gbps, Dual 10GbE | 3U — 3200W

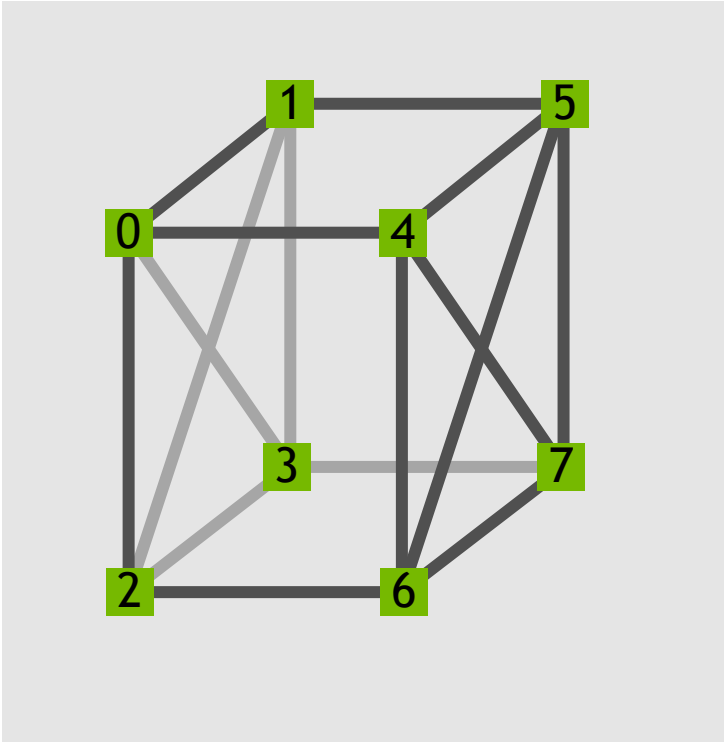
NVIDIA DGX-1 DEEP LEARNING SYSTEM

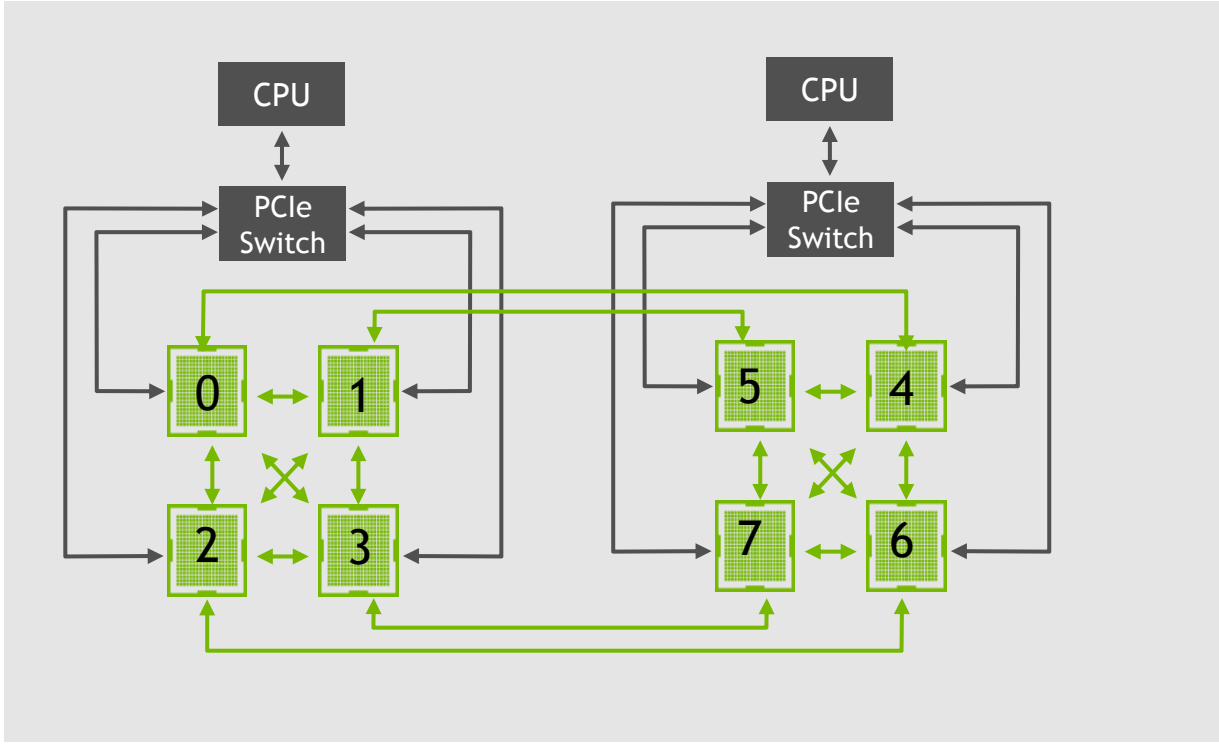


NVIDIA DGX-1
WORLD'S FIRST DEEP LEARNING SUPERCOMPUTER

170 TFLOPS

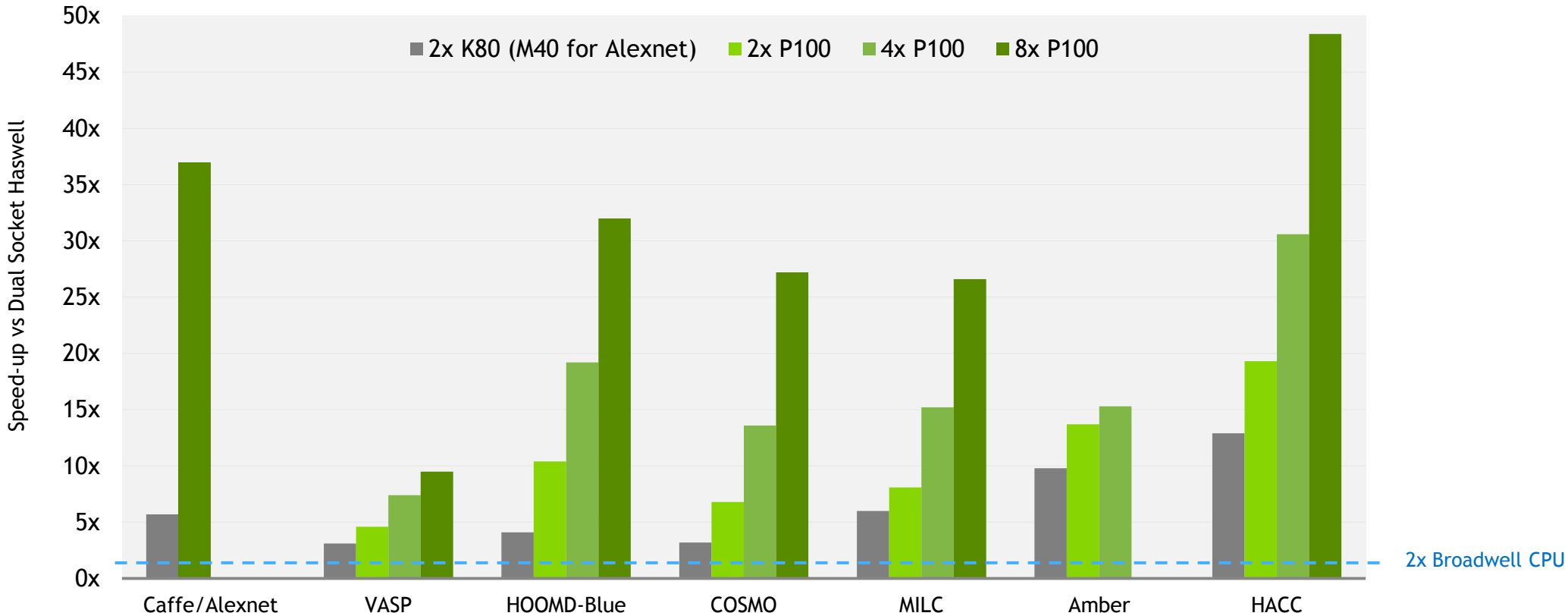
8 GPU CUBE MESH



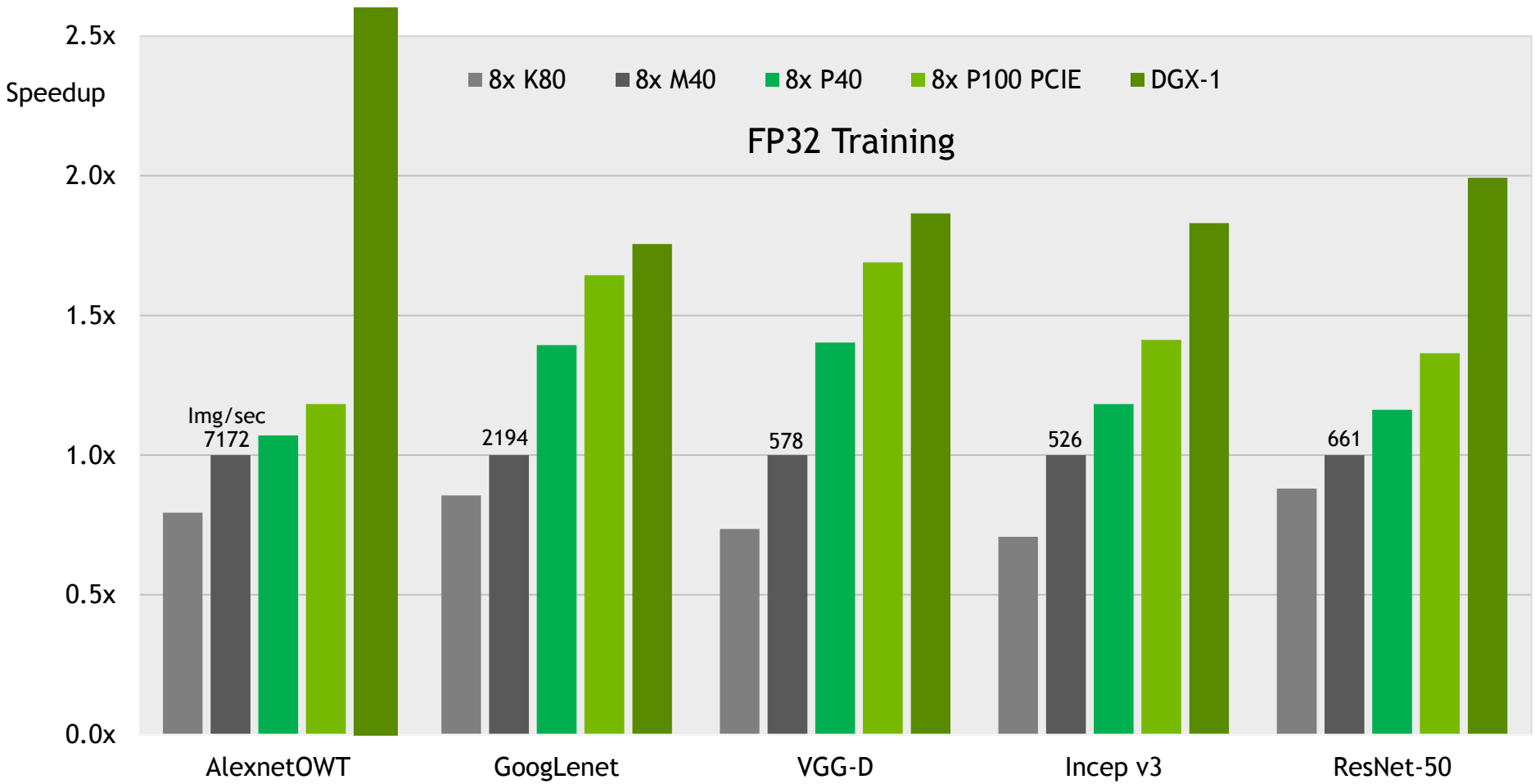


HIGHEST ABSOLUTE PERFORMANCE DELIVERED

NVLink for Max Scalability, More than 45x Faster with 8x P100

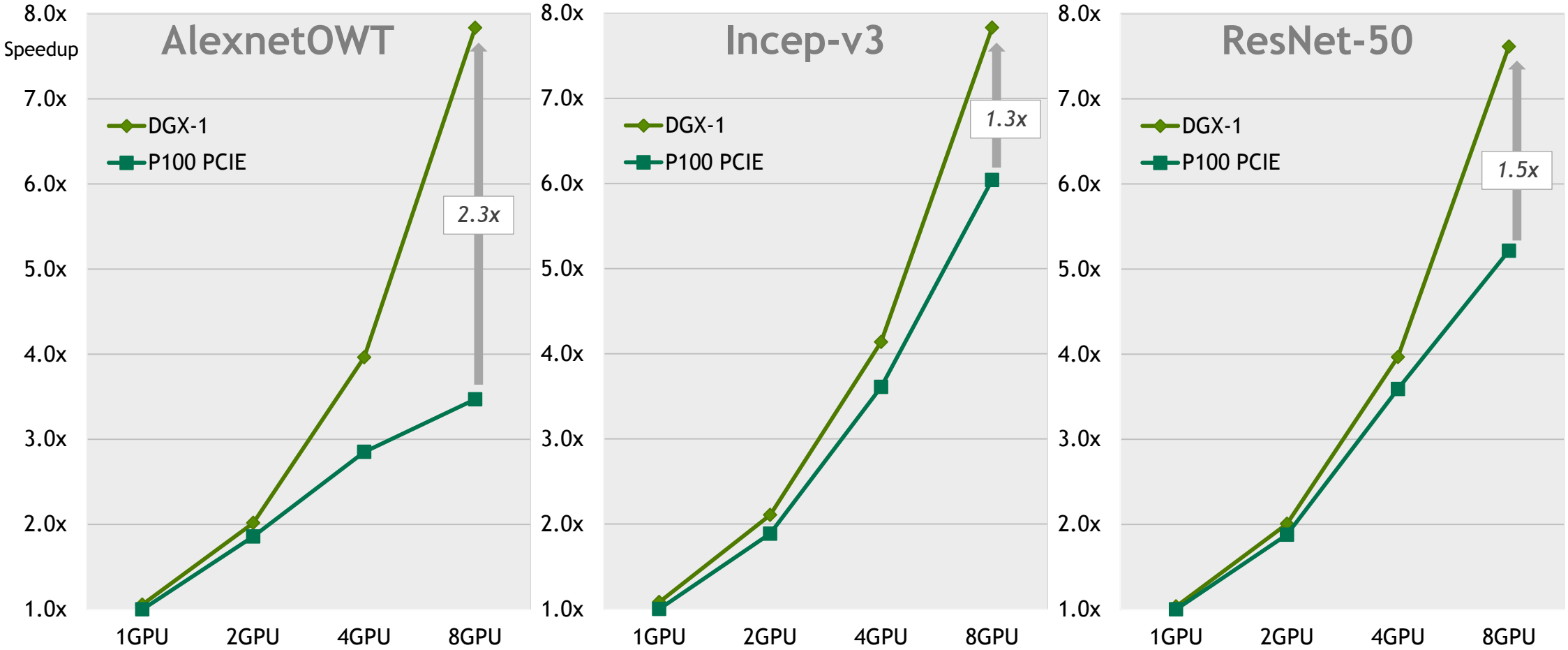


P100 FOR FASTEST TRAINING



Deepmark test with NVCaffe. AlexnetOWT/GoogLenet use batch 128, VGG-D uses batch 64, Incep-v3/ResNet-50 use batch 32, weak scaling
K80/M40/P100/DGX-1 are measured, P40 is projected, software optimization in progress, CUDA8/cuDNN5.1, Ubuntu 14.04

NVLINK ENABLES LINEAR MULTI-GPU SCALING



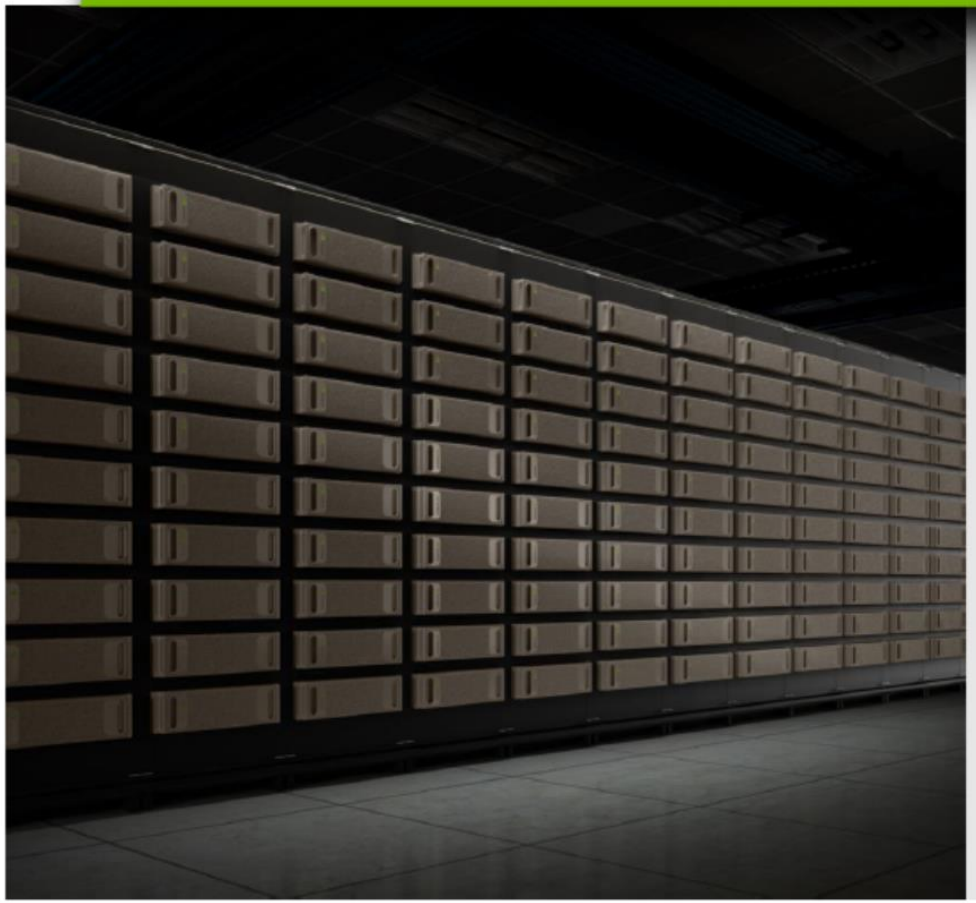
Deepmark test with NVCaffe. AlexnetOWT use batch 128, Incep-v3/ResNet-50 use batch 32, weak scaling,
P100 and DGX-1 are measured, FP32 training, software optimization in progress, CUDA8/cuDNN5.1, Ubuntu 14.04

DETAILED SPECIFICATION

Component	Description
Base Server	Dual Intel® Xeon® CPU motherboard with x2 9.6 GT/s QPI 8 Channel with 2 DPC DDR4, Intel®C610 Chipset, AST2400 BMC
	GPU baseboard supporting 8 SXM2 modules and 4 PCIE x16 slots for IB NICs
	Chassis with 3+1 1600W Power supply and support for up to 12 2.5 inch drives
	x1 GbE Management Port
	x1 COM port
	x2 USB 3.0 Ports (Rear)
CPU	x2 E5-2698 v4, 20-core, 2.2GHz, 135W
GPU	x8 Tesla P100 16GB
System Memory	512 GB using x16 2133 32GB DDR4 LRDIMM
SAS Raid Controller	8 port LSI SAS 3108 RAID Mezzanine
10 GbE NIC	10GbE dual port X540 Mezzanine
IB EDR NICs	x4 Mellanox ConnectX-4 VPI MCX455A-ECAT, Single port, x16 PCIE
SSD RAID Array	x4 1.92TB, 6 GB/s, Raid 0 Configuration, Samsung PM863 6 Gb/s SATA 3.0 SSD
SSD OS Drive	x1 480 GB, 6 GB/s, Intel S3610 6 Gb/s SATA 3.0 SSD

INTRODUCING DGX SATURNV

124 NVIDIA DGX-1 "Rocket for Cancer Moonshot"



Fastest AI Supercomputer in TOP500

4.9 Petaflops Peak FP64
19.6 Petaflops Peak FP16

Most Energy Efficient Supercomputer

#1 Green500
9.5 GFLOPS per Watt

Rocket for Cancer Moonshot

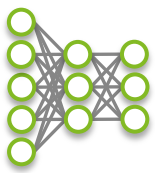
CANDLE Development Platform
Common platform with DOE labs — ANL, LLNL, ORNL, LANL

DGX-1 SOFTWARE

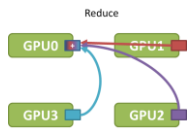
NVIDIA DGX-1 SOFTWARE STACK

Optimized for Deep Learning Performance

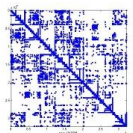
Accelerated Deep Learning



cuDNN

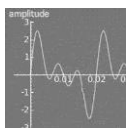


NCCL



cuSPARSE

cuBLAS



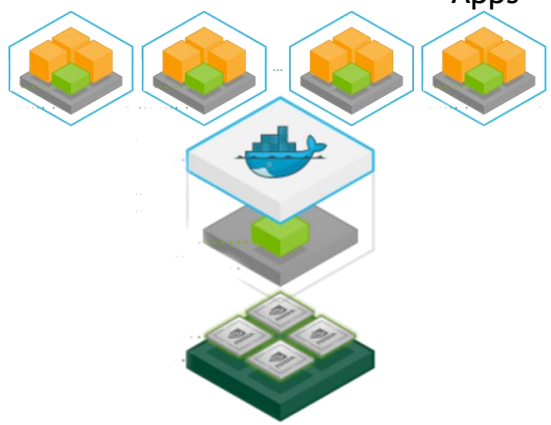
cuFFT

Container Based Applications

Digits

DL Frameworks

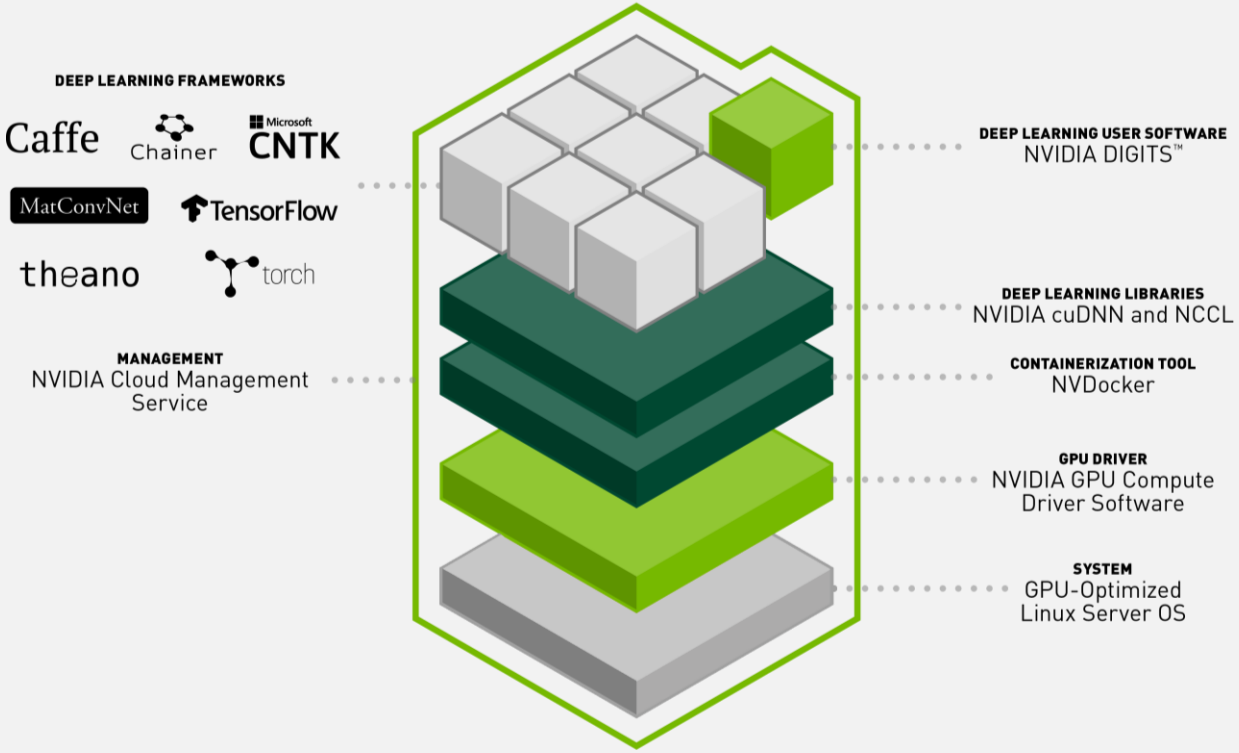
GPU Apps



NVIDIA Cloud Management

DGX STACK

Fully integrated Deep Learning platform



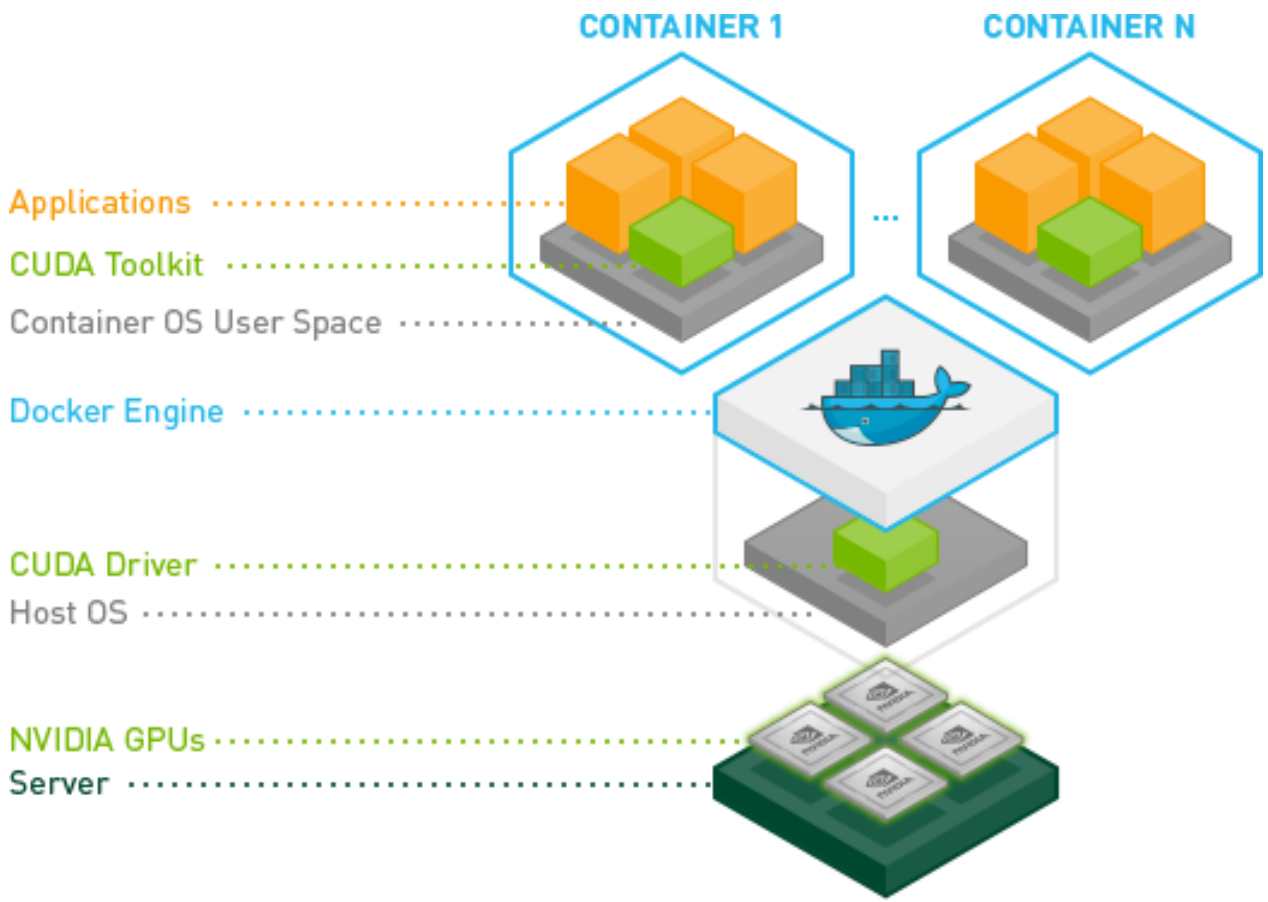
Instant productivity — plug-and-play, supports every AI framework

Performance optimized across the entire stack

Always up-to-date via the cloud

Mixed framework environments —containerized

Direct access to NVIDIA experts


CONTAINER 1
CONTAINER N
Applications
CUDA Toolkit
Container OS User Space
Docker Engine
CUDA Driver
Host OS
NVIDIA GPUs
Server

NCCL

Accelerating multi-GPU collective communications

GOAL:

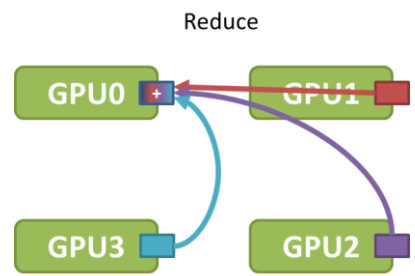
A library of accelerated collectives that is easily integrated and topology-aware so as to improve the scalability of multi-GPU applications

APPROACH:

Pattern the library after MPI's collectives

Handle the intra-node communication in an optimal way

Allow for both multi-threaded and multi-process approaches to parallelization



NCCL FEATURES

Green = Available in Github version – Black: DGX-1 only

Collectives:

Broadcast
All-Gather
Reduce
All-Reduce
Reduce-Scatter
Scatter
Gather
All-To-All
Neighborhood

Key Features:

Single-node, any number of GPUs
Host-side API
Asynchronous/non-blocking interface
Multi-thread, multi-process support
In-place and out-of-place operation
PCIe/QPI support
NVLink support

COMPUTE.NVIDIA.COM

GPU Accelerated Cluster Management Portal

GPU Container Hosting

Repository of pre-configured GPU Apps

NVDocker GPU Accelerated Containers

Cloud Scheduling

Point and click App deployment UI

Mesos Scheduler Connected to Cloud

Cloud Cluster Management

Cluster-wide telemetry & analysis

NVIDIA HW health monitoring

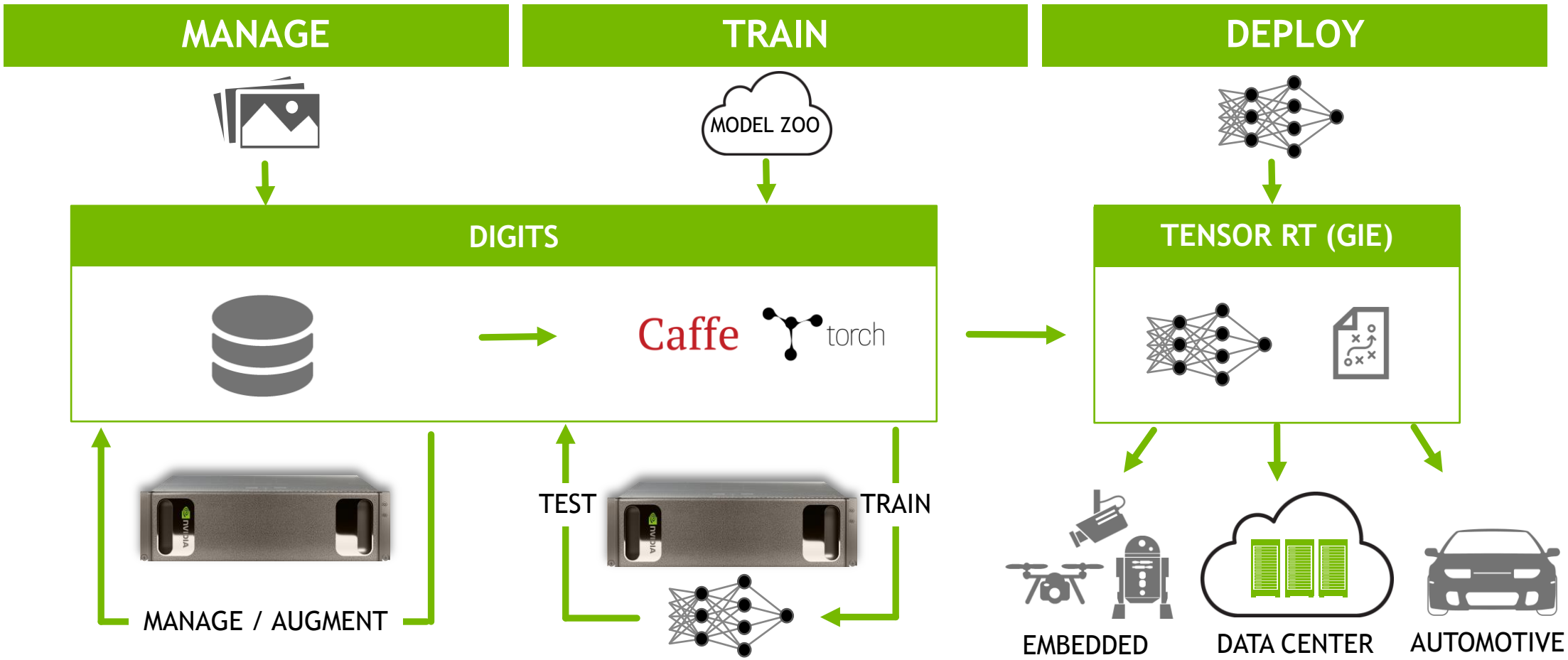
Updates & Security

Customer data stay on-premises

Notifications and rapid deployment of updates

DGX-1 IN THE WORKFLOW

A complete GPU-accelerated deep learning workflow



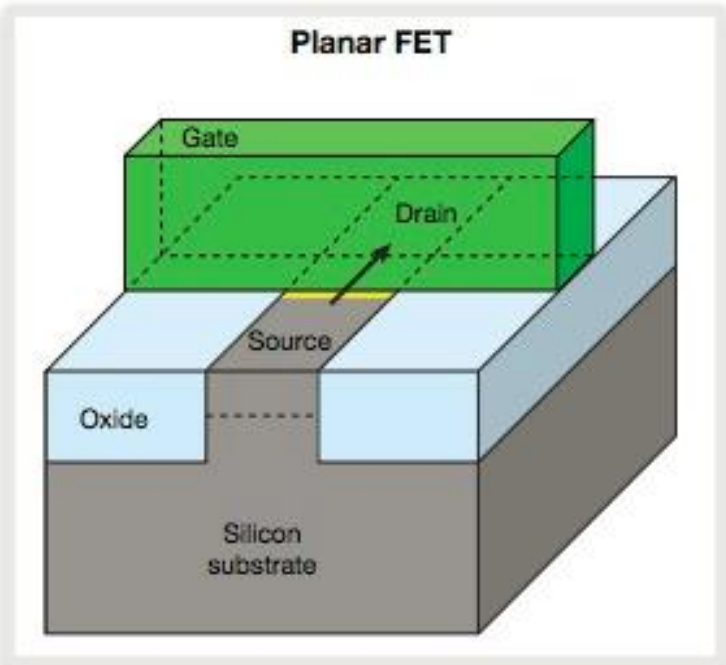
THANK YOU!

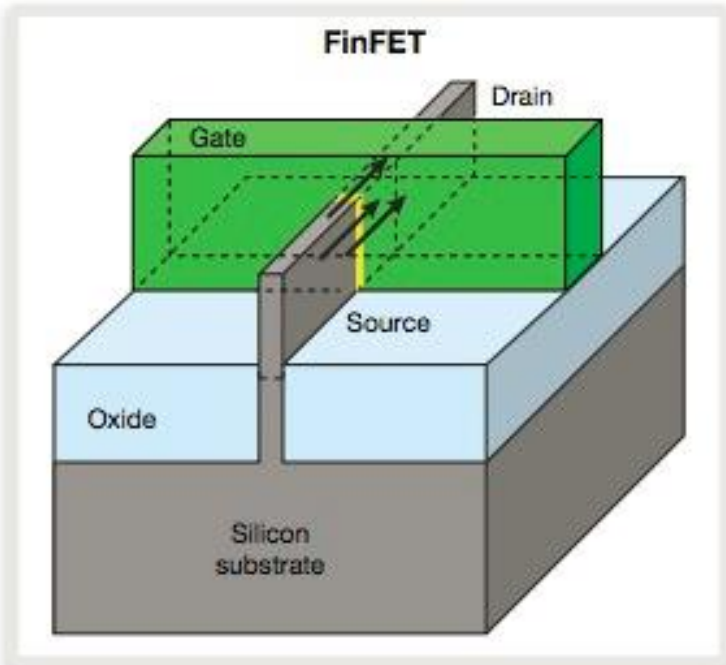
cnardone@nvidia.com

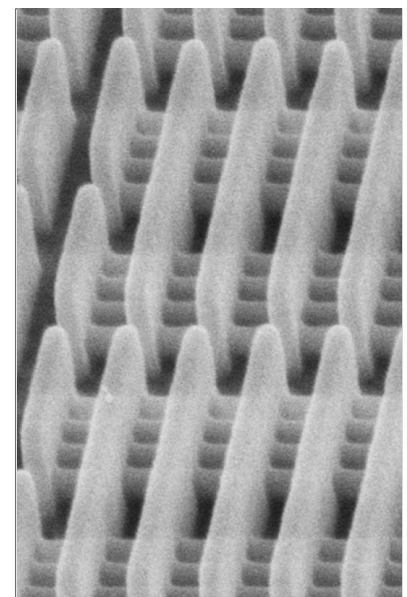
BACKUP SLIDES

FINFET TECHNOLOGY

16 nm



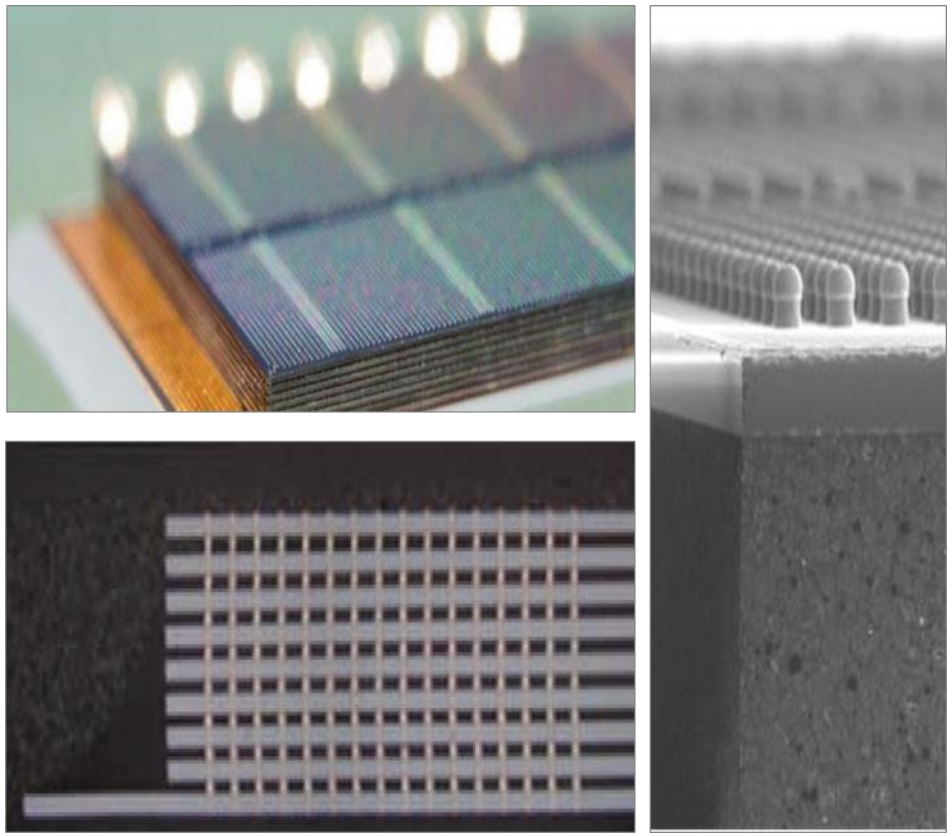


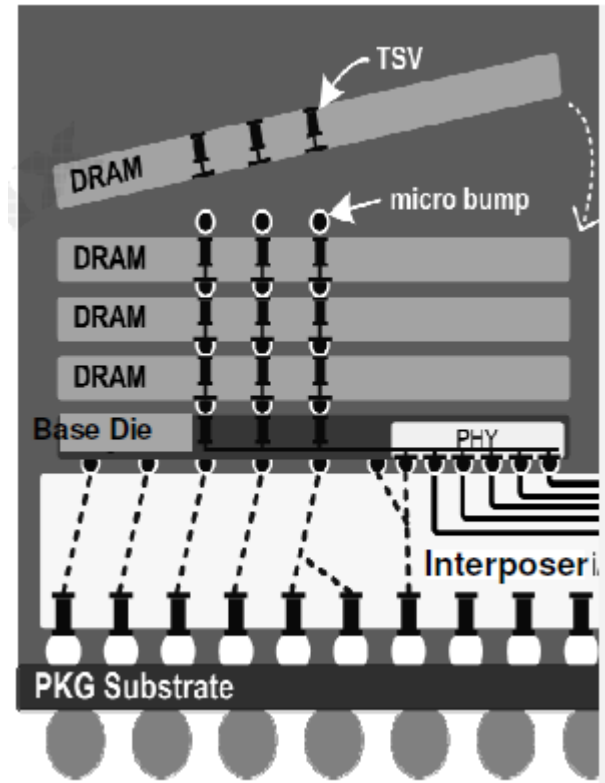


16nm
FinFET

Source: SemiWiki.com,Synopsys

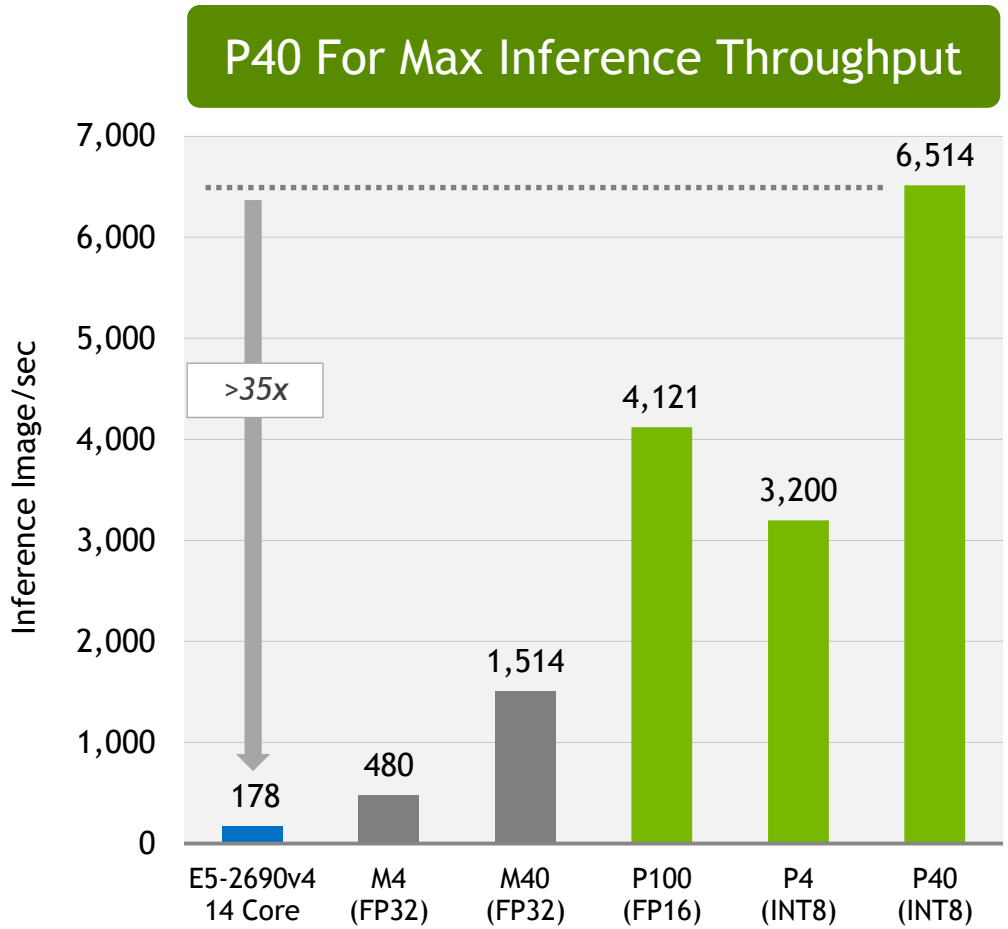
HBM 3D STACKED MEMORY

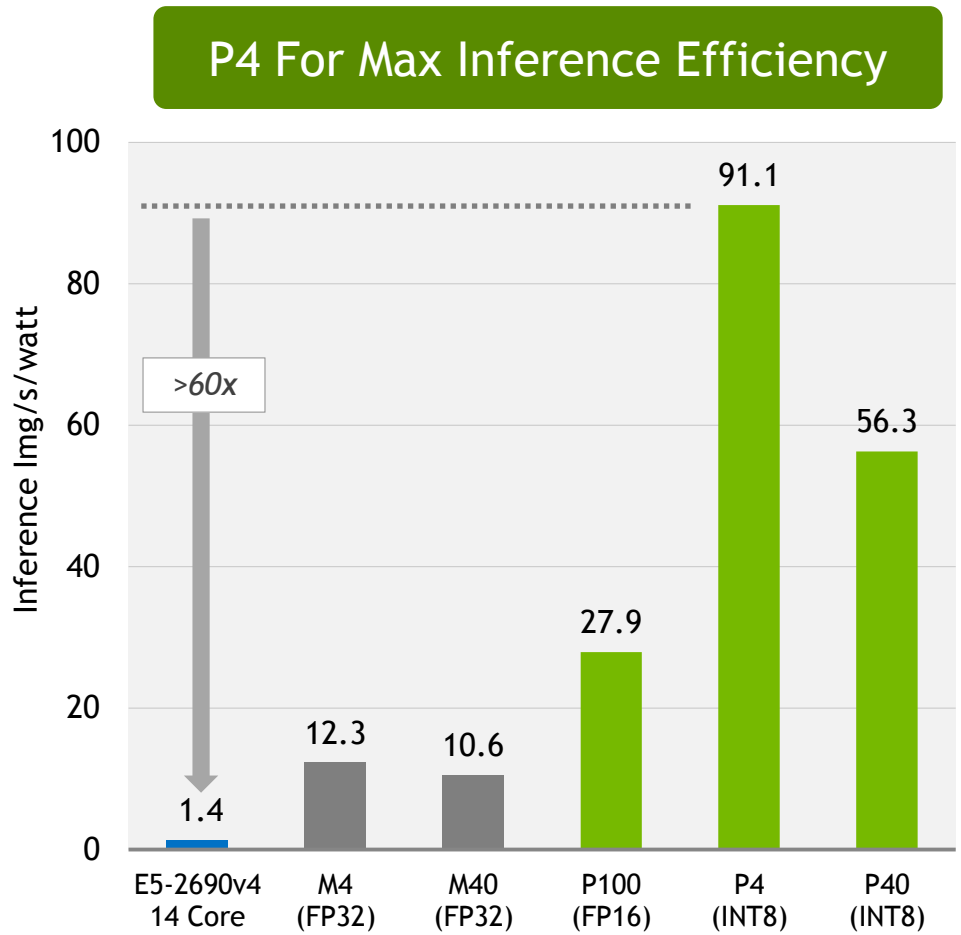




Source: Hynix

P40 & P4 DELIVER MAX INFERENCE PERFORMANCE





All results are measured, based on GoogLenet with batch size 128
Xeon uses MKL 2017 GOLD with FP32, GPU uses TensorRT internal development ver.

P100 FOR FASTEST TRAINING

	M40 MAXWELL	P40 PASCAL	P100 PASCAL
FP16 / FP32 (TFLOPs)	NA / 7	NA / 12	21.2 / 10.6
Register File	6 MB	7.5 MB	14 MB
Memory BW	288 GB/s	346 GB/s	732 GB/s
Chip-Chip BW	32 GB/s (PCIE)	32 GB/s (PCIE)	160 GB/s (NVLINK) + 32 GB/s (PCIE)
Mem Size (Max DL model size)	24 GB	24 GB	16GB x 8 (Model Parallel)