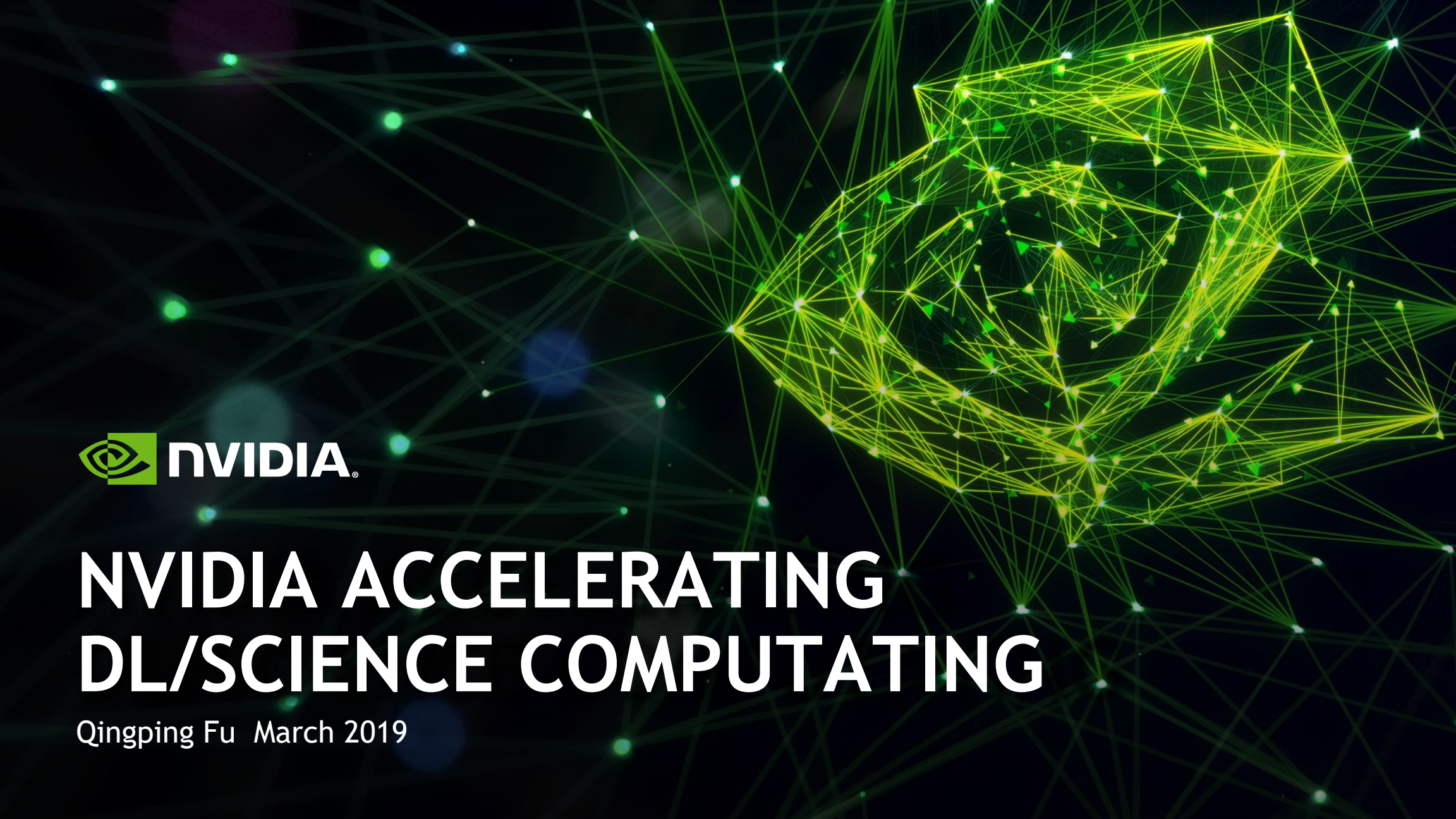


NVIDIA ACCELERATING DL/SCIENCE COMPUTATING

Qingping Fu March 2019

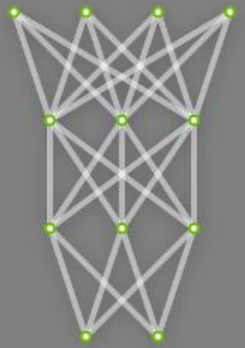
AGENDA

- Deep learning and Scientific Computing
- Why GPU
- Introduction To Volta GPU Platform
- NVIDIA GPU CLOUD
- GPU Accelerated Applications Sharing


DEEP LEARNING AND
SCIENTIFIC COMPUTING

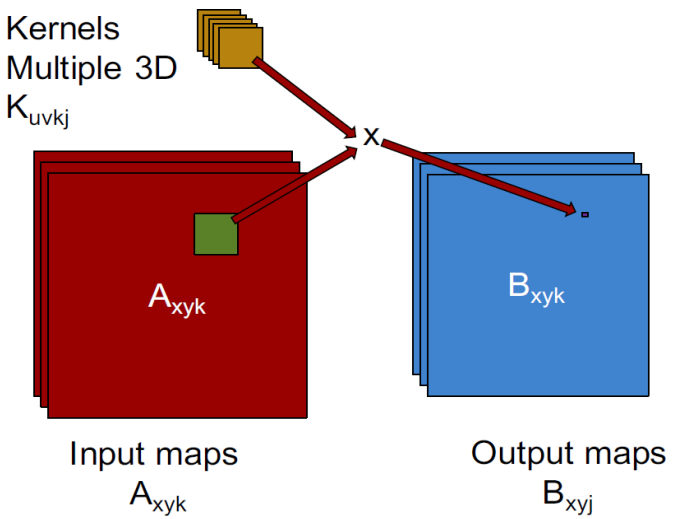
DEEP LEARNING

Untrained
Neural Network
Model



CNN

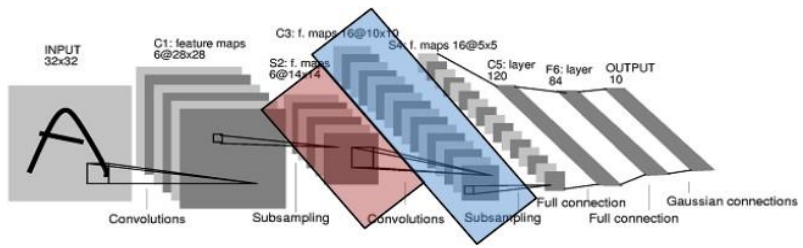
REQUIRES CONVOLUTION AND M X V



Filters conserved through plane

Multiply limited – even without batching.

6D Loop
For each output map j
  For each input map k
    For each pixel x,y
      For each kernel element u,v
        B_{xyj} += $A_{(x-u)(y-v)k}$ x K_{uvkj}

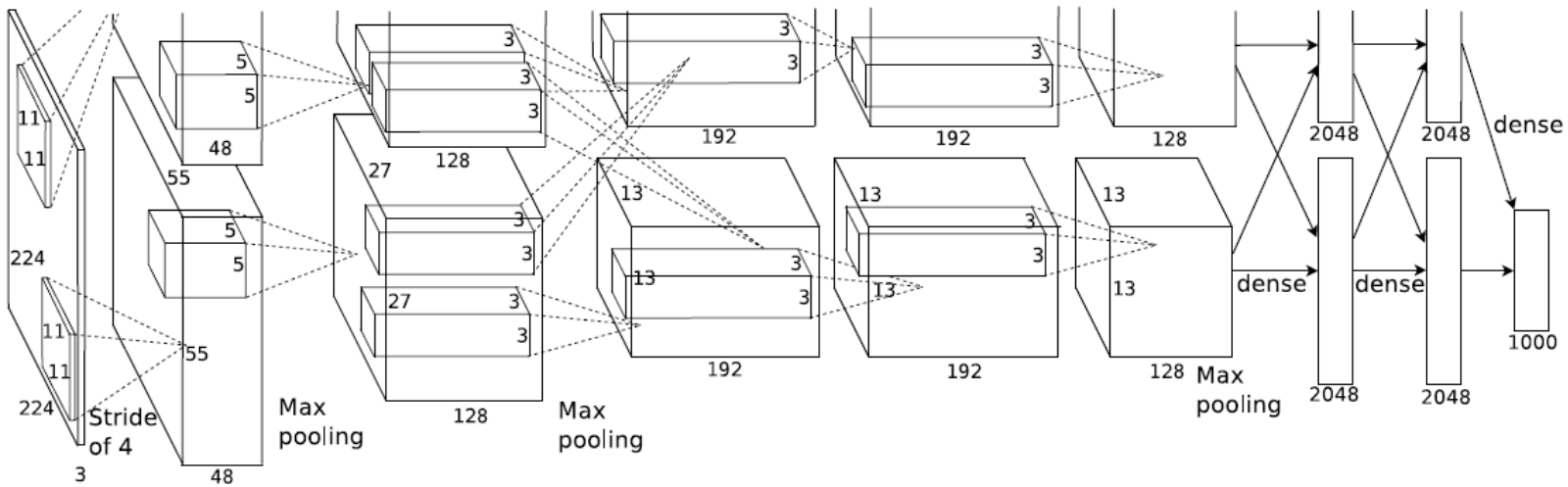


第一个真正的卷积神经网络LeNet
3个卷积层、2个下采样层、1个全连接层、1个输出层
参数量：156+12+1516+32+48120+10164+840=60840

ALEXNET

LSVRC-2010 TOP-1 37.5% TOP-5 17%

1. 使用ReLU激活函数
2. 使用GPU训练
3. 局部响归一化
4. 重叠池化
5. 减少过拟合dropout



包含输入层、5个卷积层（3个进行了最大池化）、3个全连接层。
参数量约为：60million

VGGNET

LSVRC-2014 TOP-1 24.7% TOP-5 7.32%

1. 探究深度对CNN效果的影响
2. 训练了不同深度的6个网络模型
3. 使用小卷积核

参数总量约：130million+

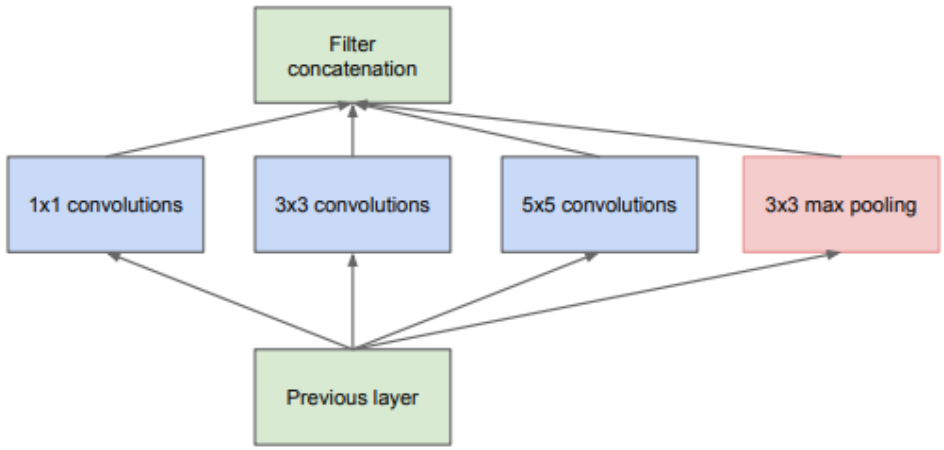
Table 2: **Number of parameters** (in millions).

Network	A,A-LRN	B	C	D	E
Number of parameters	133	133	134	138	144

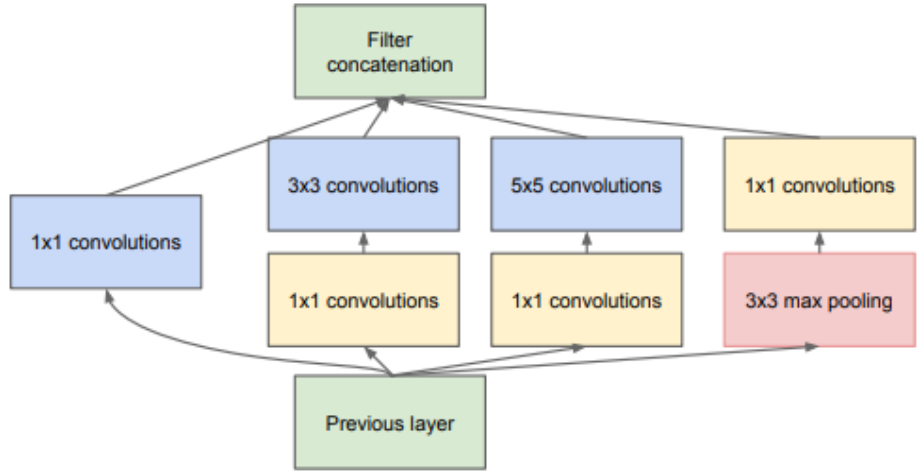
ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

GOOGLENET

LSVRC-2014 TOP-1 23.7% TOP-5 6.67%



(a) Inception module, naïve version



(b) Inception module with dimension reductions

1. 引入Inception module　　2. 1x1卷积降维　　3. 多尺度卷积聚合

RESNET

LSVRC-2015 TOP-1 19.38% TOP-5 3.57%

1. 解决增加网络深度导致性能下降的问题
2. 跨层链接，构造残差模块

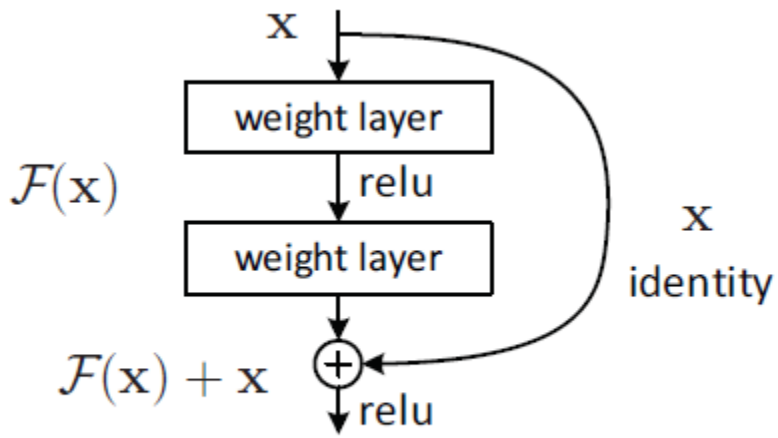
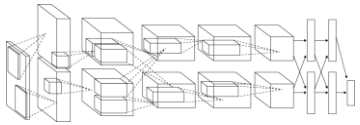


Figure 2. Residual learning: a building block.

CAMBRIAN EXPLOSION

Convolutional Networks



Encoder/Decoder

ReLu

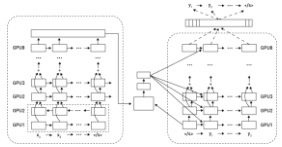
BatchNorm

Concat

Dropout

Pooling

Recurrent Networks



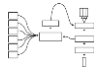
LSTM

GRU

Beam Search

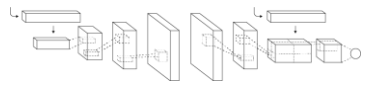
WaveNet

CTC



Attention

Generative Adversarial Networks



3D-GAN

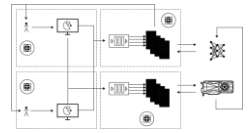
MedGAN

Conditional GAN

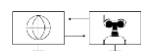
Coupled GAN

Speech Enhancement GAN

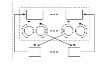
Reinforcement Learning



DQN

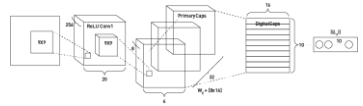


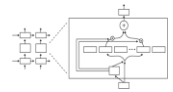
Simulation



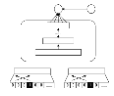
DDPG

New Species





Mixture of Experts



Neural Collaborative Filtering

Block Sparse LSTM

AI - A NEW INSTRUMENT FOR SCIENCE

HPC

> Algorithms based on first principles theory.

> **P**roven models for accurate results

AI

> Neural Networks that learn patterns from large data sets

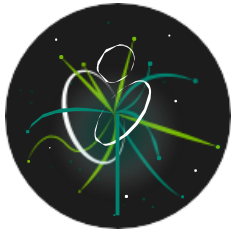
> Improve predictive accuracy and faster response time.

Dramatically Improves Accuracy and Time-to-Solution

Commercially viable fusion energy

Understanding cosmological dark energy and matter

Clinically viable precision medicine

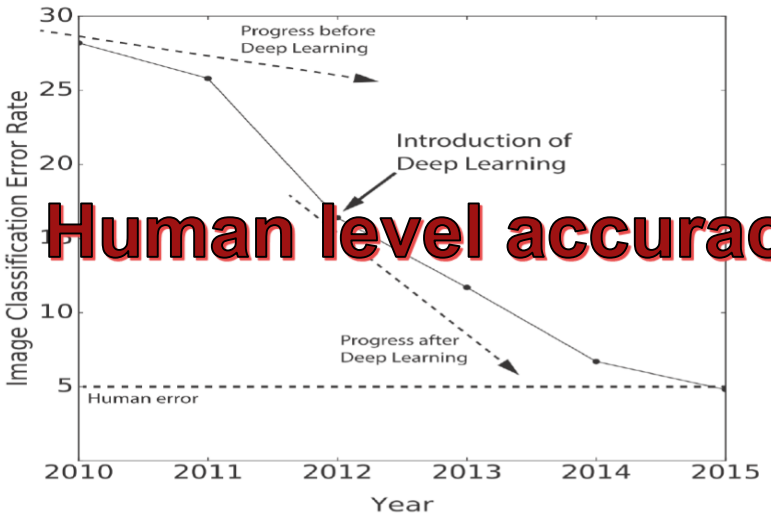


Improvement and validation of the Standard Model of Physics

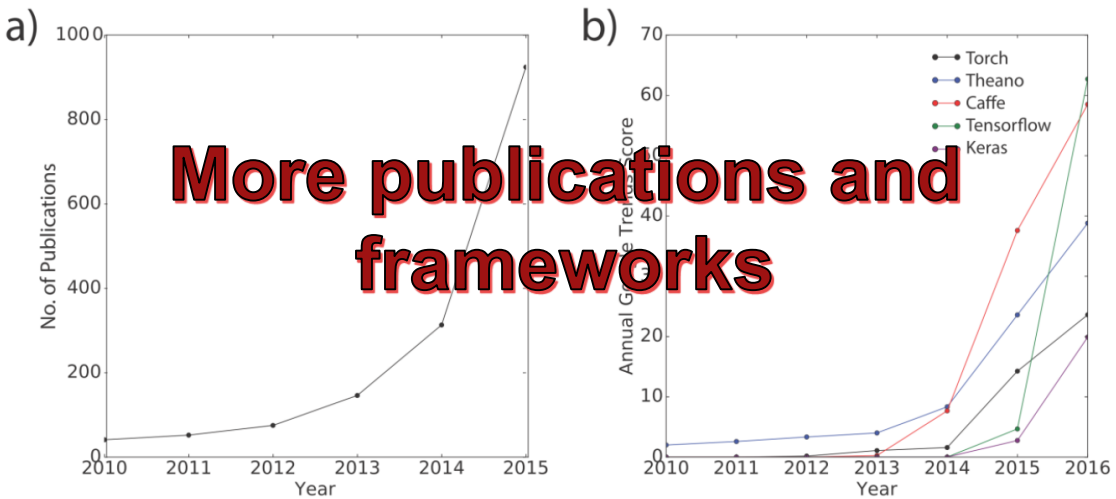
Climate/weather forecasts with ultra-high fidelity

DEEP LEARNING FOR COMPUTATIONAL CHEMISTRY

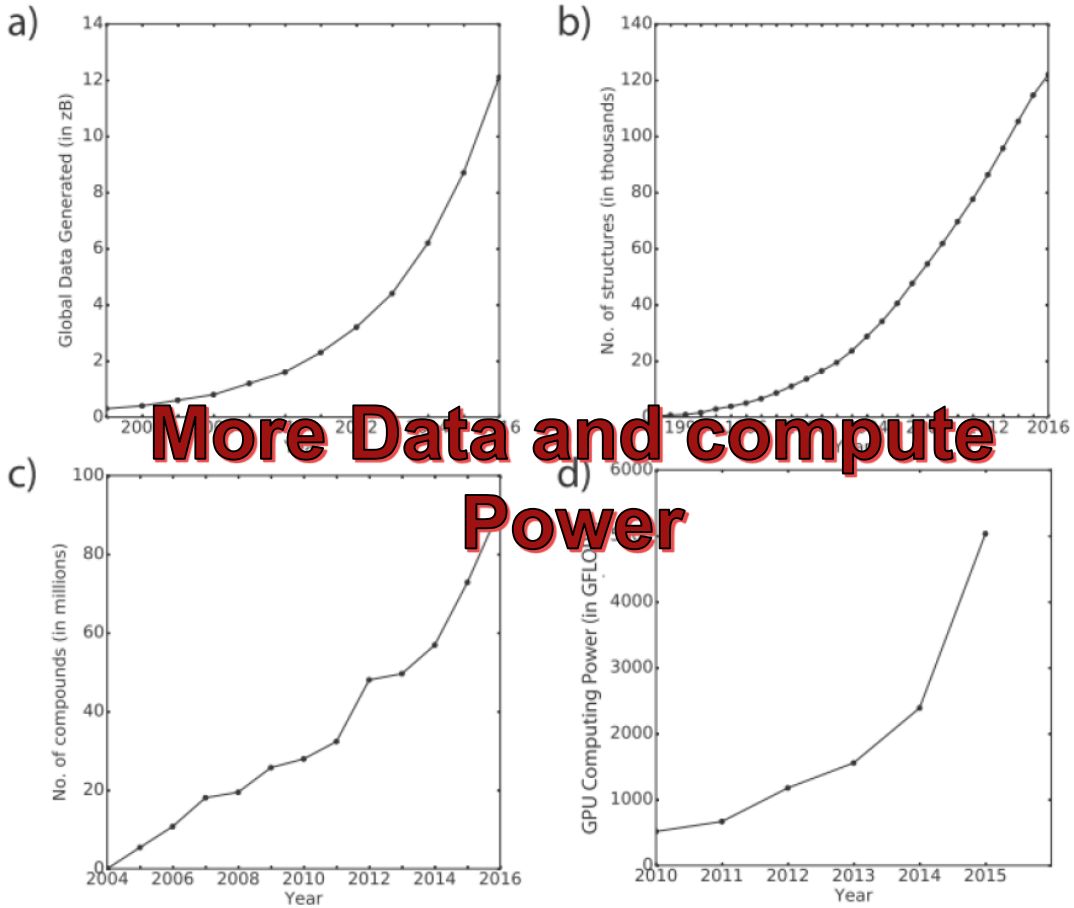
HTTPS://DOI.ORG/10.1002/JCC.24764 PACIFIC NORTHWEST NATIONAL LABORATORY



Human level accuracy



More publications and frameworks



More Data and compute Power

DEEP LEARNING FOR COMPUTATIONAL CHEMISTRY

PACIFIC NORTHWEST NATIONAL LABORATORY

Prediction / Competition	DNN Models	Comments	Non-DNN Models	Comments
Merck Kaggle Challenge (Activity)	0.494 R^2	DNN-based model was the top performing model in the competition.[62]	0.488 R^2	Best non-DNN model in the competition.[140]
	0.465 R^2	Median DNN-based model recreated by Merck post-competition.[66]	0.423 R^2	Best non-DNN model (RF-based) by Merck post-competition.[66]
Activity	0.830 AUC	MT-DNN based model trained on the ChEMBL database.[68]	0.816 AUC	Best non-DNN model (SVM) trained on the ChEMBL database.[68]
	0.873 AUC	MT-DNN based model trained on the PCBA database.[70]	0.800 AUC	Best non-DNN model (RF) based model trained on the PCBA database.[70]
	0.841 AUC	MT-DNN based model trained on the MUV database.[70]	0.774 AUC	Best non-DNN model (RF) based model trained on the MUV database.[70]
NIH Tox21 Challenge (Toxicity)	0.846 AUC	DeepTox (MT-DNN based model) was the top performing model.[86]	0.824 AUC	Best non-DNN model (multi-tree ensemble model) was placed 3rd in the Tox21 challenge.[141]
	0.838 AUC	Runner up in Tox21 challenge was based off associative neural networks (ASNN).[142]		
	0.818 AUC	Post-competition MT-DNN model.[70]	0.790 AUC	Post-competition RF model.[70]
Atom-level Reactivity/ Toxicity	0.949 AUC	DNN-based model that predicts site of epoxidation, a proxy for toxicity.[80]	-	No comparable model in the literature that can identify site of reactivity or toxicity.
	0.898 AUC	DNN-based model that predicts site of reactivity to DNA.[84]		
	0.944 AUC	DNN-based model that predicts site of reactivity to protein.[84]		
Protein Contact	36.0% acc.	CMAPpro (DNN-based model).[106]		
	34.1% acc.	DNCON (DNN-based model).[108]	29.7% acc. 28.5% acc.	Best non-DNN model reported in CASP9, ProC_S3 (RF-based model)[28] and SVMcon (SVM-based model)[27] are listed respectively.

FIRST PRINCIPLES NEURAL NETWORK POTENTIALS FOR REACTIVE SIMULATIONS OF LARGE MOLECULAR AND CONDENSED SYSTEMS

ANGEW. CHEM. INT. ED. 2017, 56, 12828 – 12840

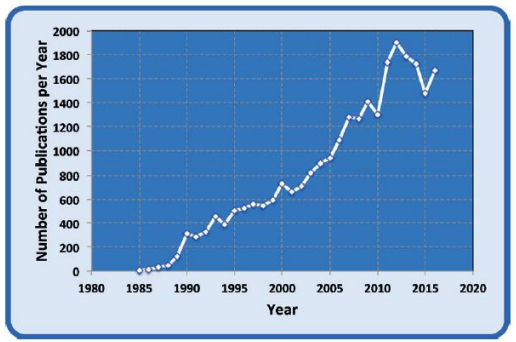


Figure 1. Peer-reviewed publications in chemistry, physics, and materials sciences that made use of artificial neural networks. The data were extracted from the Web of Science[50] in March 2017.

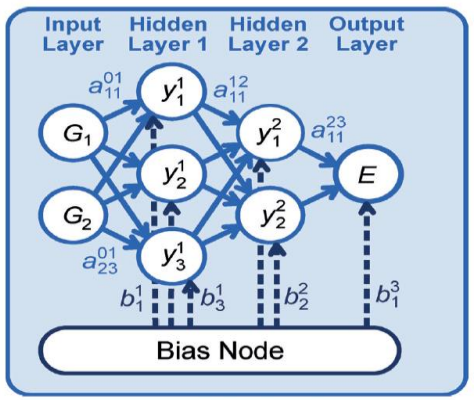


Figure 3. Schematic structure of a feed-forward neural network defining the functional relationship between a two-dimensional input vector $\mathbf{G} = (G_1, G_2)$ that describes the atomic configuration and the potential energy E [Eq. (3)].

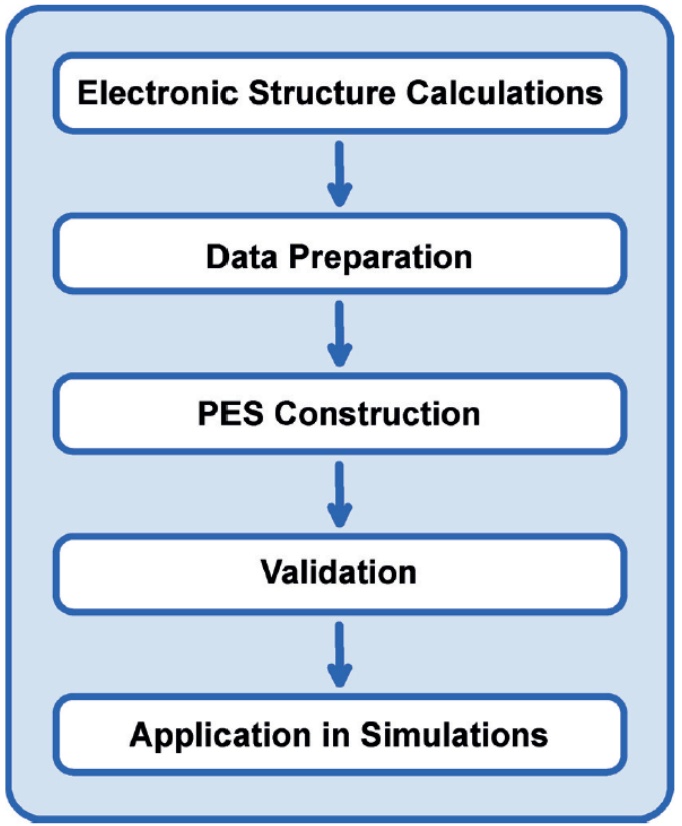


Figure 2. Steps in the construction of machine learning potentials.

Table 1: Applications of high-dimensional neural network potentials.

Year[a]	System	References
2007	bulk silicon	[58, 78, 110]
2010	bulk carbon	[59, 111]
2010	bulk sodium	[112, 113]
2011	ZnO	[82]
2012	bulk GeTe	[114–117]
2012	copper	[56–118]
2012	methanol molecule	[118]
2012	water clusters	[83, 103, 119, 120]
2013	Cu clusters on ZnO	[121]
2014	Au/Cu clusters and water	[122]
2015	Cu-Au nanoalloys	[123]
2015	allyl vinyl ether	[53]
2016	bulk water	[55, 106, 124]
2016	Cu on CeO_2	[125]
2016	water on copper	[108, 126]
2016	*n*-alkanes	[127]
2016	ethane molecule	[128]
2016	Na clusters	[104]
2016	$H_2 + SH$	[129]
2016	$H_2 + H$, $H_2O + H$, $CH_4 + H$	[130]
2016	TiO_2	[131]
2016	bulk gold and surfaces	[132]
2016	NaOH in water	[101, 107]
2017	Cu/Pd/Ag	[133]
2017	HCl at Au(111)	[134]
2017	AuPd(111) surface	[135]
2017	CaF2	[87]
2017	oxygen on Pd surface	[136]
2017	Au clusters	[105]
2017	water on ZnO	[102]

[a] In the case of several publications for a particular system, the year of the first publication is given.

APPLICATIONS OF NEURAL NETWORKS TO THE SIMULATION OF DYNAMICS OF OPEN QUANTUM SYSTEMS

CHEMICAL PHYSICS

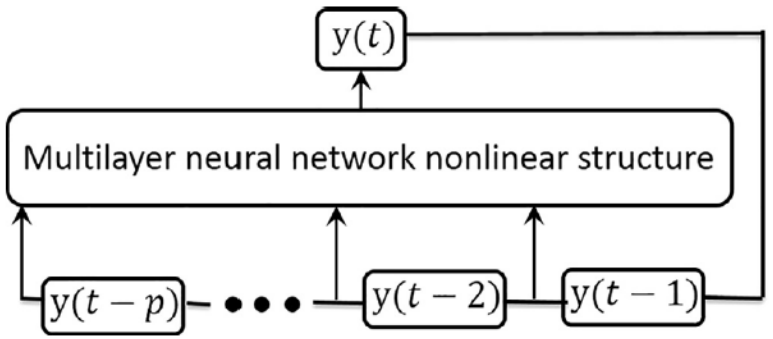


Fig. 2. Non-linear autoregressive neural network is a recurrent, dynamic network with feedback connections. The next term in a sequence is predicted from a fixed number of previous terms using delay taps.

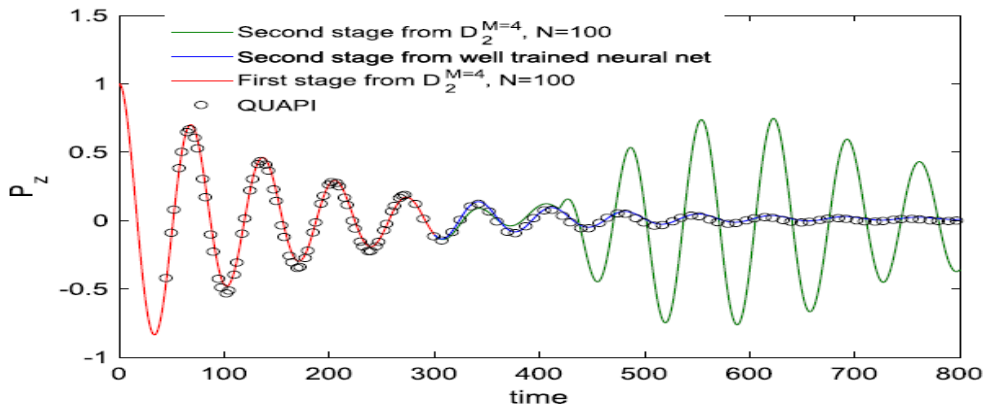


Fig. 7. Dynamics of spin-boson model with multiple boson modes. The parameters used are $\epsilon = 0, \Delta = 0.1\omega_c, \alpha = 0.05, k_BT/\omega_c = 0.01$. In the variational calculation, the number of modes is $N = 100$ and multiplicity of multi-D_2 *Ansatz* is $M = 4$.

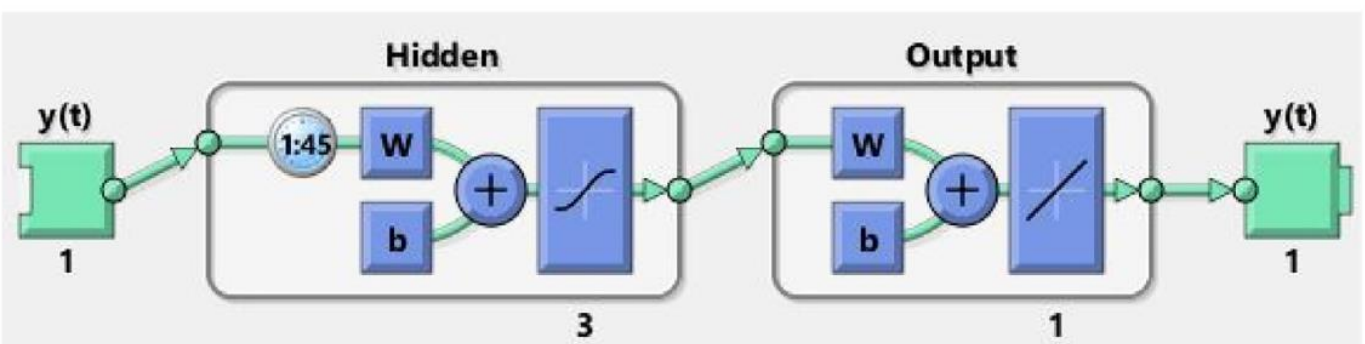
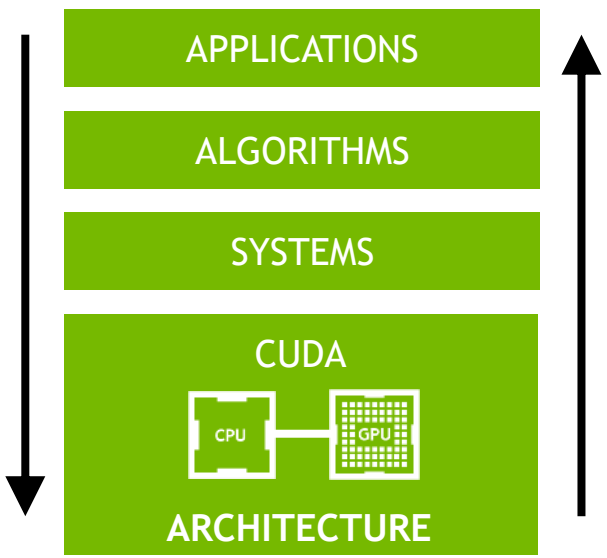
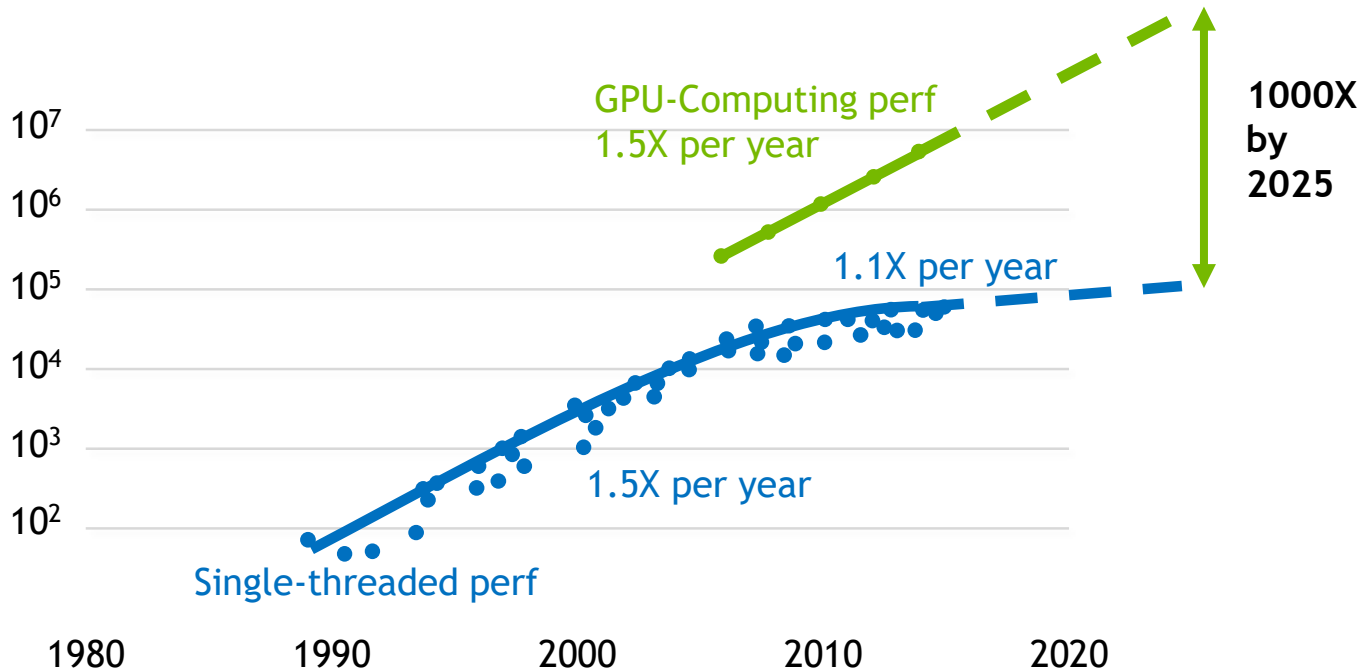


Fig. 3. Structure of the neural network used to estimate time series of observables in open quantum systems discussed in this work. The hyperbolic tangent sigmoid transfer function and the linear transfer function are used in the hidden layer and output layer, respectively.


WHY GPU

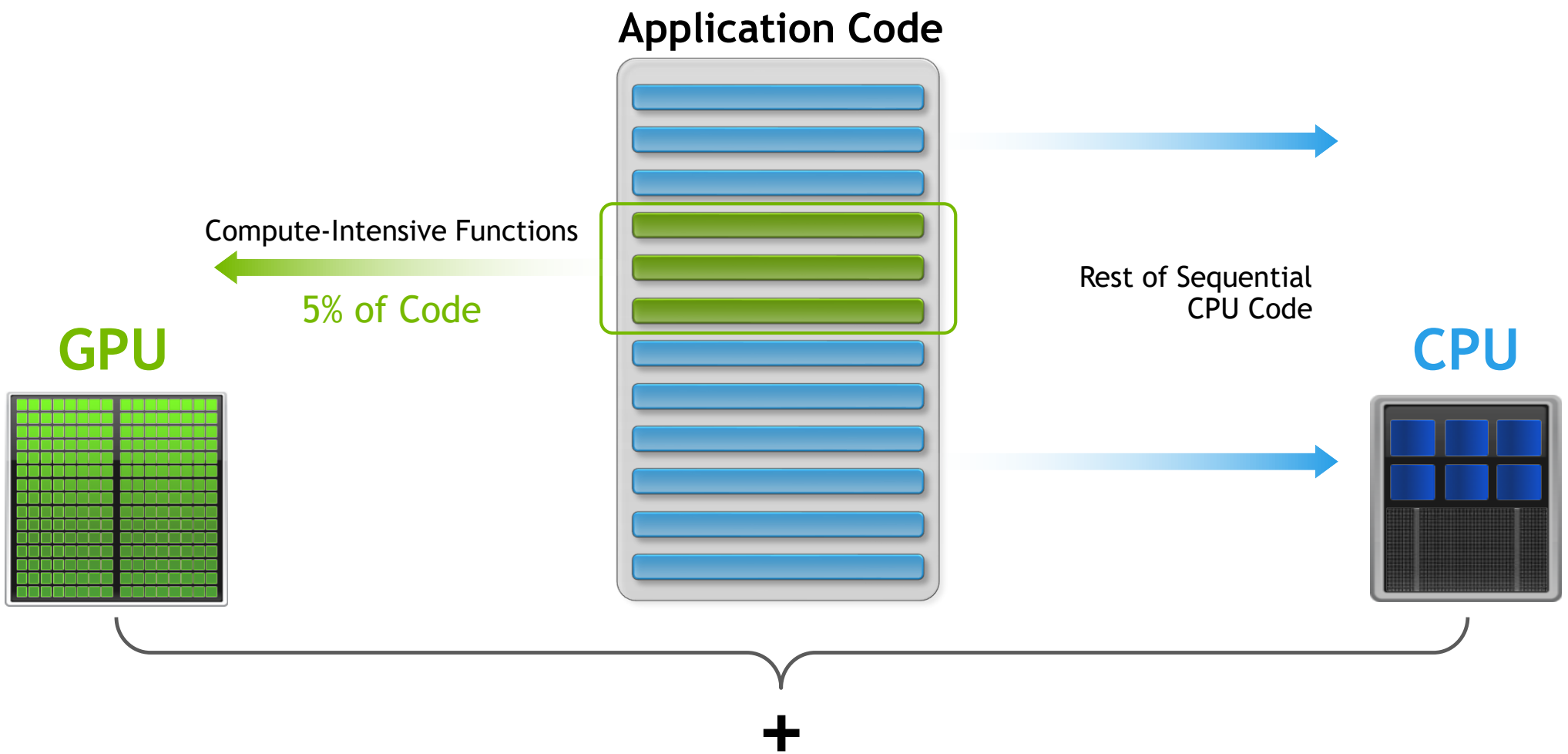
RISE OF GPU COMPUTING



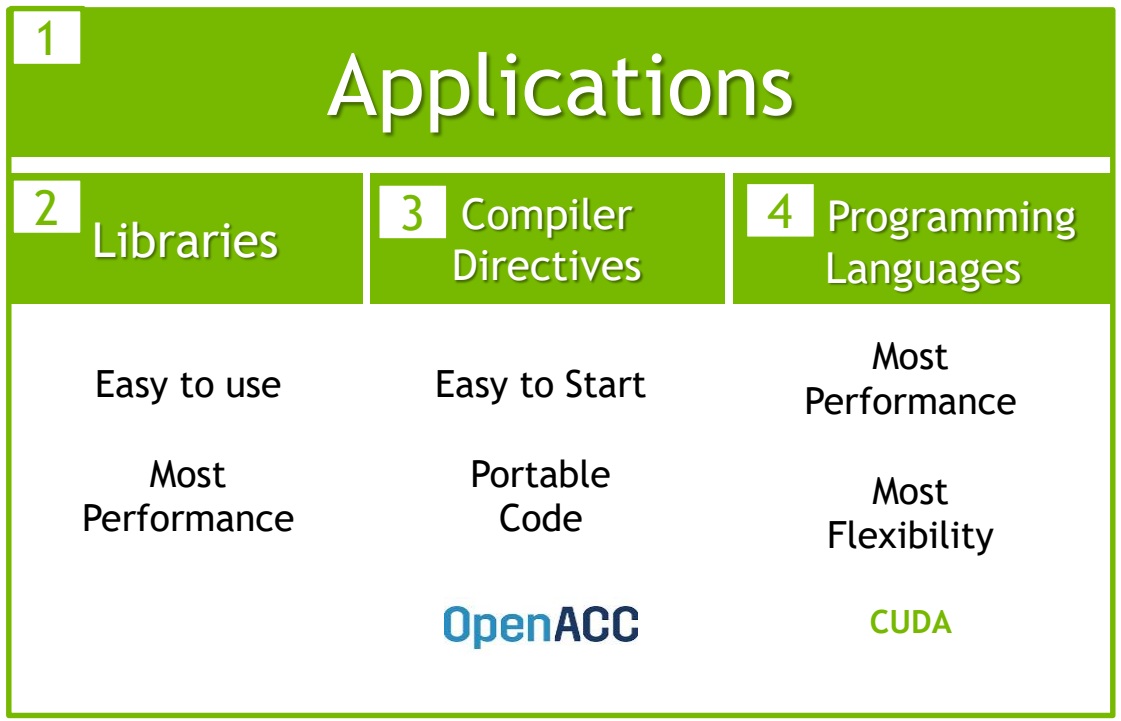


Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

HOW GPU ACCELERATION WORKS



HOW TO START WITH GPUS



1. Review available GPU-accelerated applications
2. Check for GPU-Accelerated applications and libraries
3. Add OpenACC Directives for quick acceleration results and portability
4. Dive into CUDA for highest performance and flexibility

3 STEPS TO CUDA-ACCELERATED APPLICATION

Step 1: Substitute library calls with equivalent CUDA library calls

```
saxpy ( … ) ► cublasSaxpy ( … )
```

Step 2: Manage data locality

- with CUDA: `cudaMalloc(), cudaMemcpy(), etc.`
- with CUBLAS: `cublasAlloc(), cublasSetVector(), etc.`

Step 3: Rebuild and link the CUDA-accelerated library

```
gcc myobj.o –l cublas
```

WHAT IS OPENACC

Programming Model for an Easy Onramp to GPUs

Directives-based programming model for **parallel computing**

Add Simple Compiler Directive

```
main()
 {
  <serial code>
  #pragma acc kernels
  {
    <parallel code>
  }
 }
```

Designed for **performance portability** on CPUs and GPUs

Simple

Powerful & Portable

Read more at www.openacc.org/about

OpenACC is an open specification developed by OpenACC.org consortium

OpenACC DIRECTIVES EXAMPLE

```
!$acc data copy(A,Anew)
 iter=0
 do while ( err > tol .and. iter < iter_max )

  iter = iter +1
  err=0._fp_kind

!$acc kernels
   do j=1,m
    do i=1,n
     Anew(i,j) = .25_fp_kind *( A(i+1,j  ) + A(i-1,j  ) &
                               +A(i  ,j-1) + A(i  ,j+1))
     err = max( err, Anew(i,j)-A(i,j))
    end do
   end do
!$acc end kernels
   IF(mod(iter,100)==0 .or. iter == 1)    print *, iter, err
   A= Anew

 end do
!$acc end data
```

Copy arrays into GPU memory within data region

Parallelize code inside region

Close off parallel region

Close off data region, copy data back

CUDA C SAMPLE

```
void saxpy_serial(int n,
                  float a,
                  float *x,
                  float *y)
{

  for (int i = 0; i < n; ++i)
    y[i] = a*x[i] + y[i];
}

// Perform SAXPY on 1M elements
saxpy_serial(4096*256, 2.0, x, y);
```

```
__global__
void saxpy_parallel(int n,
                    float a,
                    float *x,
                    float *y)
{
  int i = blockIdx.x*blockDim.x +
          threadIdx.x;
  if (i < n) y[i] = a*x[i] + y[i];
}

// Perform SAXPY on 1M elements
saxpy_parallel<<<4096,256>>>(n,2.0,x,y);
```

http://developer.nvidia.com/cuda-toolkit

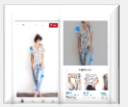
TESLA UNIVERSAL ACCELERATION PLATFORM

Single Platform Drives Utilization and Productivity

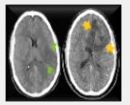
CUSTOMER USECASES


Speech

Translate


Recommender

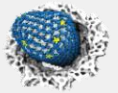
CONSUMER INTERNET

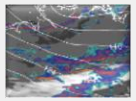

Healthcare

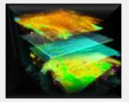

Manufacturing


Finance

INDUSTRIAL APPLICATIONS


Molecular Simulations


Weather Forecasting


Seismic Mapping

SCIENTIFIC APPLICATIONS

APPS & FRAMEWORKS

python™ RAPIDS

TensorFlow PYTORCH mxnet Chainer ONNX

Amber ANSYS NAMD SIMULIA +550 Applications

NVIDIA SDK & LIBRARIES

MACHINE LEARNING/ ANALYTICS

cuDF | cuML | cuGRAPH

DEEP LEARNING

cuDNN | cuBLAS | CUTLASS | NCCL | TensorRT

HPC

CuBLAS | CuFFT | OpenACC

CUDA

TESLA GPUs & SYSTEMS


TESLA GPU


VIRTUAL GPU


NVIDIA DGX FAMILY


NVIDIA HGX

DELL Hewlett Packard Enterprise IBM

SYSTEM OEM

aws Google Cloud Platform Microsoft Azure

CLOUD

NVIDIA DEEP LEARNING SOFTWARE STACK

TRAINING

Training Data

Data Management

Training

Model Assessment

Trained Neural Network

Caffe2 · Chainer · Microsoft Cognitive Toolkit · mxnet · TensorFlow · PYTORCH · theano

INFERENCE

Data center

TensorRT

Embedded

JETPACK SDK

Automotive

DriveWorks SDK

NVIDIA DEEP LEARNING SDK and CUDA

cuDNN

NCCL (GPU0, GPU1, GPU2, GPU3)

cuBLAS

TensorRT

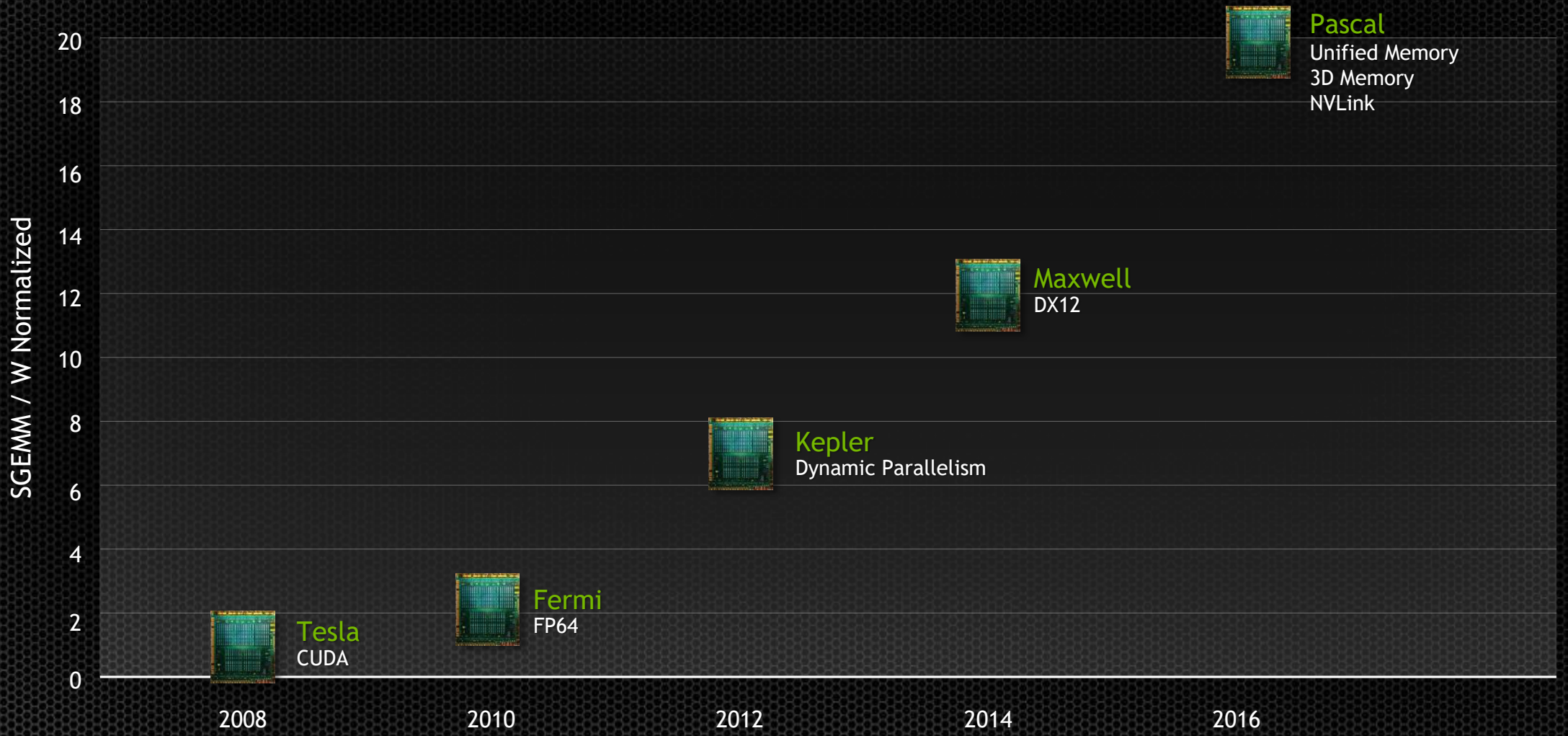
cuSPARSE

DeepStream SDK (NVDEC, Inference, NVENC)

developer.nvidia.com/deep-learning-software


VOLTA GPU

GPU EVOLUTION



TESLA V100 TENSOR CORE GPU

World's Most Advanced Data Center GPU

5,120 CUDA cores
640 NEW Tensor cores
7.8 FP64 TFLOPS | 15.7 FP32 TFLOPS | 125 Tensor TFLOPS
20MB SM RF | 16MB Cache
32 GB HBM2 @ 900GB/s | 300GB/s NVLink

TESLA T4

WORLD'S MOST ADVANCED SCALE-OUT GPU

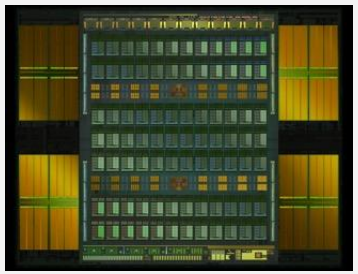
320 Turing Tensor Cores
2,560 CUDA Cores
65 FP16 TFLOPS | 130 INT8 TOPS | 260 INT4 TOPS
16GB | 320GB/s
70 W

Deep Learning Training & Inference
HPC Workloads
Video Transcode
Remote Graphics

TESLA V100

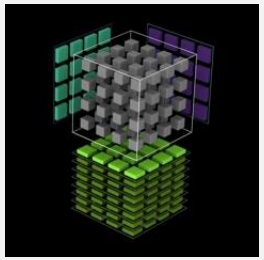
The Fastest and Most Productive GPU for AI and HPC

Volta Architecture



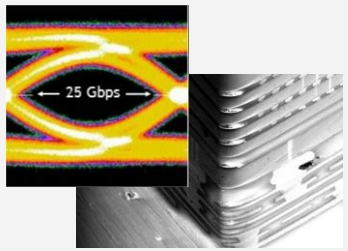
Most Productive GPU

Tensor Core



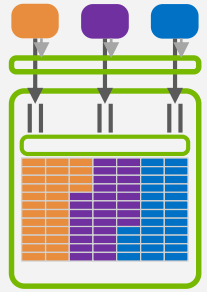
125 Programmable TFLOPS Deep Learning

Improved NVLink & HBM2



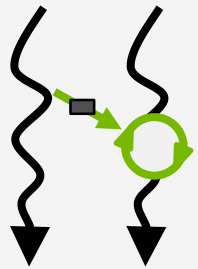
Efficient Bandwidth

Volta MPS

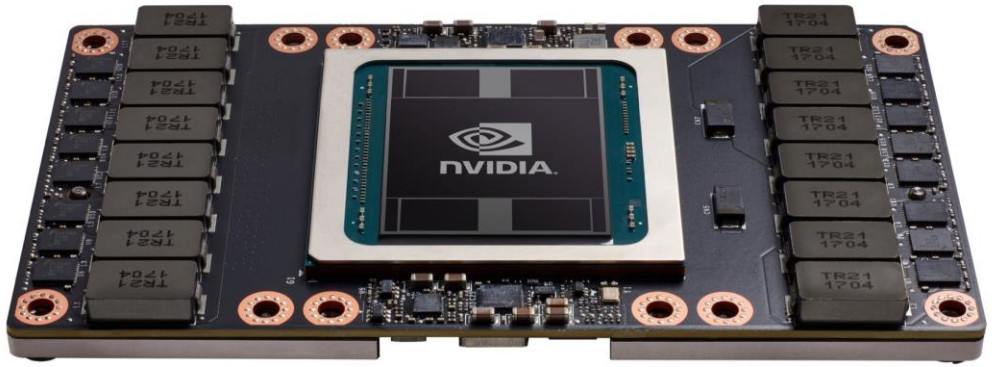


Inference Utilization

Improved SIMT Model



New Algorithms



TESLA V100

21B transistors
815 mm²

80 SM
5120 CUDA Cores
640 Tensor Cores

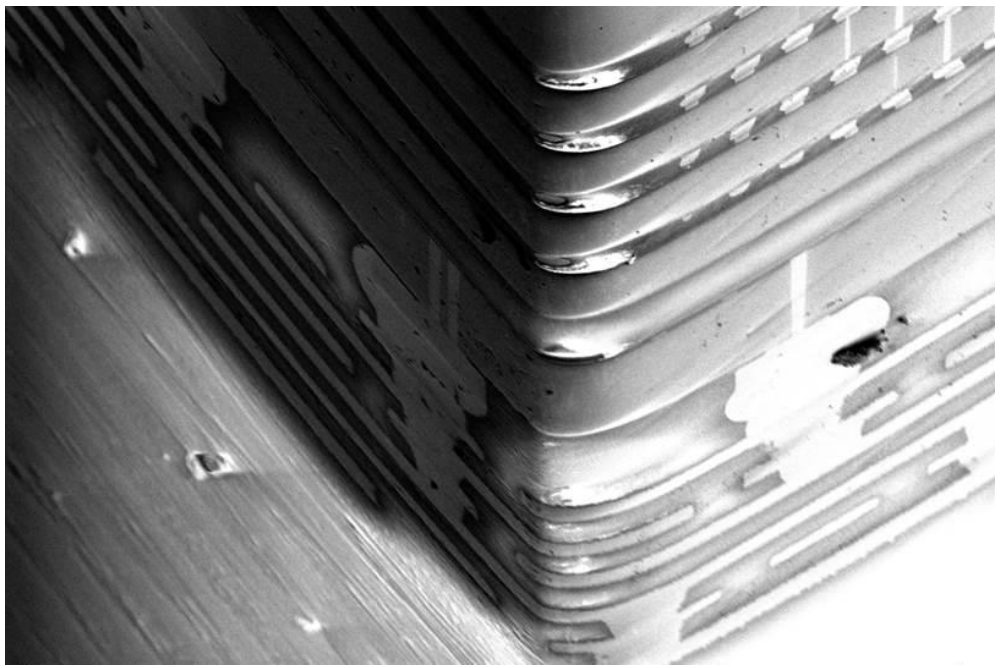
16 GB HBM2
900 GB/s HBM2
300 GB/s NVLink

*full GV100 chip contains 84 SMs

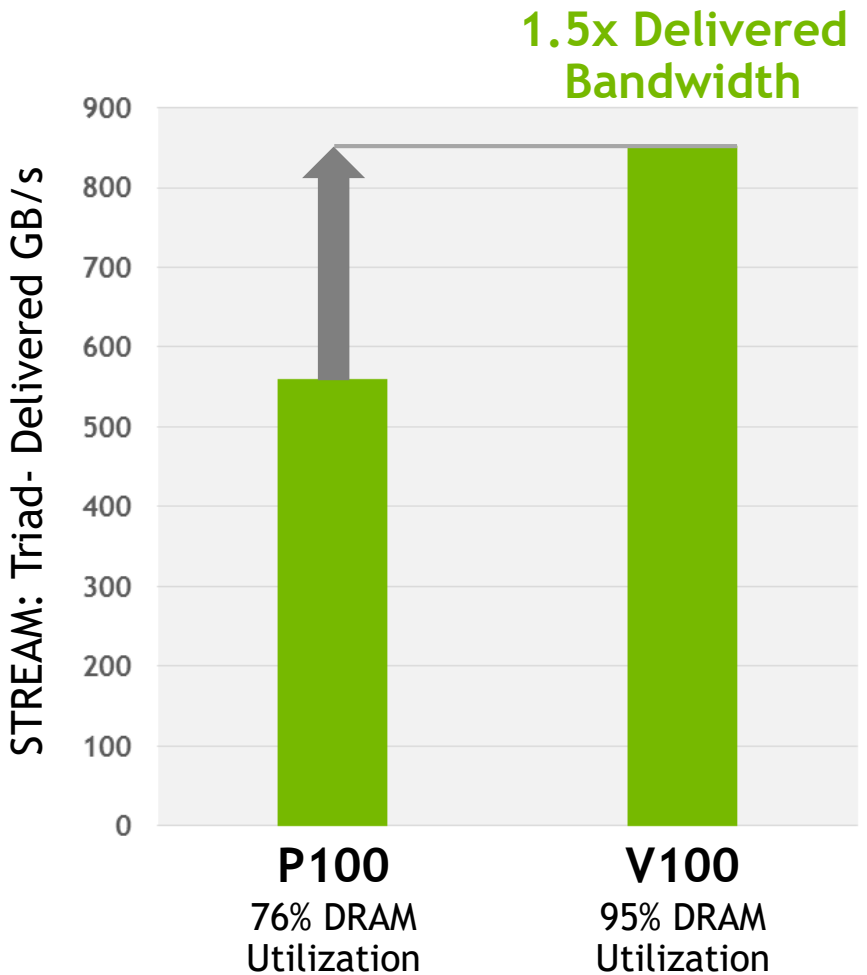
VOLTA GV100 SM

	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

NEW HBM2 MEMORY ARCHITECTURE



HBM2 stack

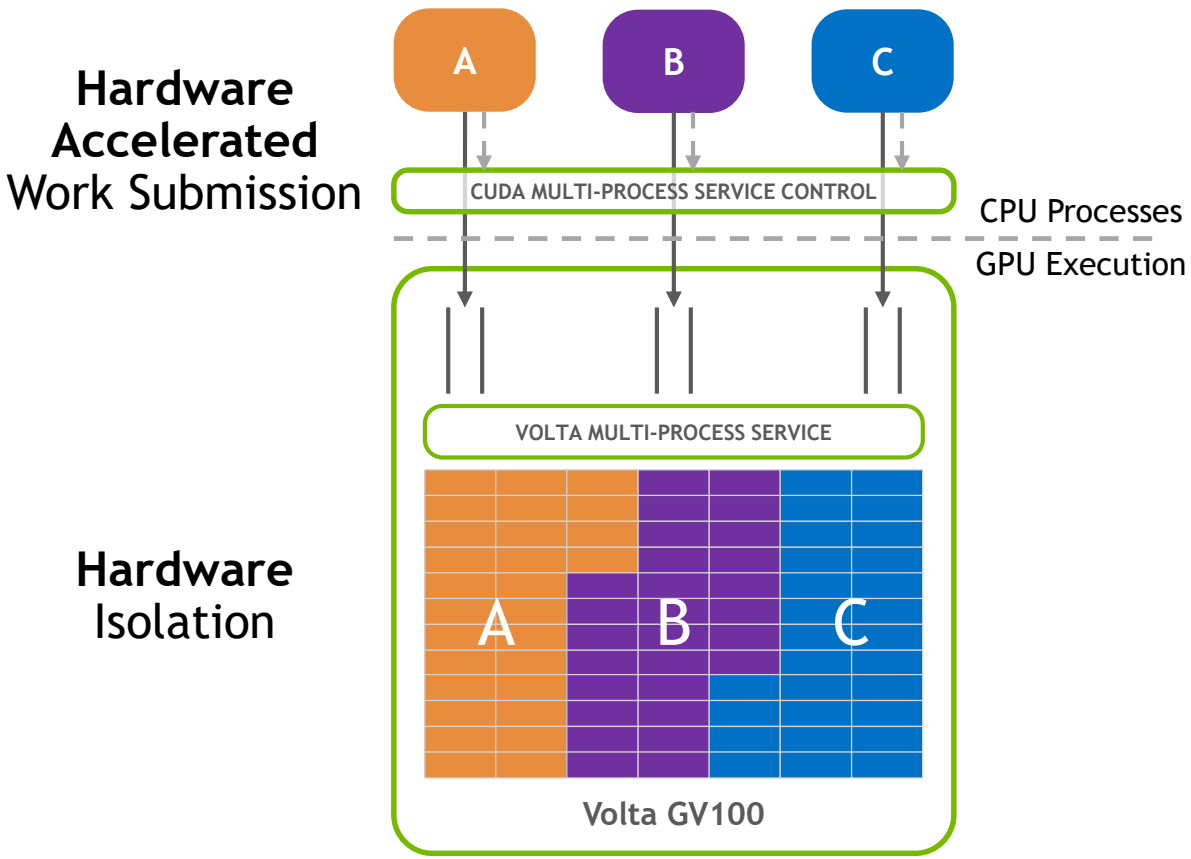


V100 measured on pre-production hardware.

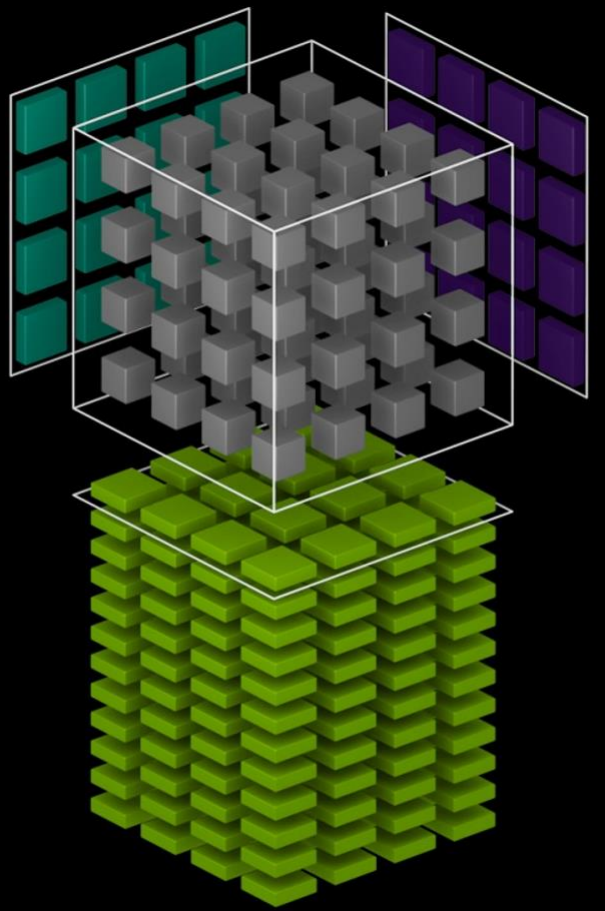
VOLTA MULTI-PROCESS SERVICE

Volta MPS Enhancements:

- Reduced launch latency
- Improved launch throughput
- Improved quality of service with scheduler partitioning
  - More reliable performance
- 3x more clients than Pascal

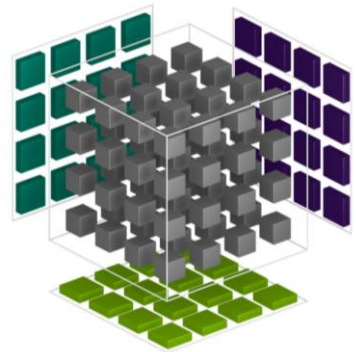


VOLTA TENSOR CORE



TENSOR CORE

Mixed Precision Matrix Math
4x4 matrices



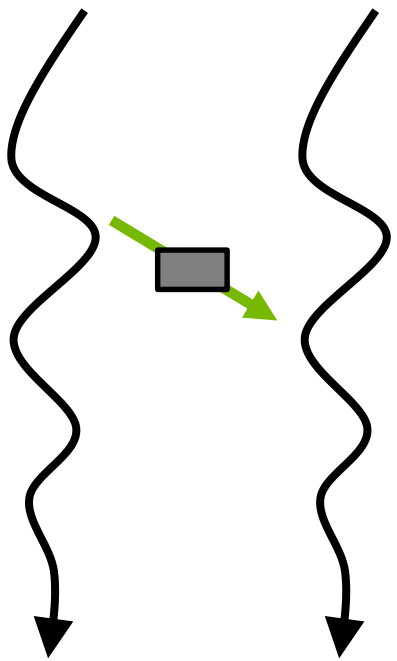
$$D = \begin{pmatrix} A_{0,0} & A_{0,1} & A_{0,2} & A_{0,3} \\ A_{1,0} & A_{1,1} & A_{1,2} & A_{1,3} \\ A_{2,0} & A_{2,1} & A_{2,2} & A_{2,3} \\ A_{3,0} & A_{3,1} & A_{3,2} & A_{3,3} \end{pmatrix} \begin{pmatrix} B_{0,0} & B_{0,1} & B_{0,2} & B_{0,3} \\ B_{1,0} & B_{1,1} & B_{1,2} & B_{1,3} \\ B_{2,0} & B_{2,1} & B_{2,2} & B_{2,3} \\ B_{3,0} & B_{3,1} & B_{3,2} & B_{3,3} \end{pmatrix} + \begin{pmatrix} C_{0,0} & C_{0,1} & C_{0,2} & C_{0,3} \\ C_{1,0} & C_{1,1} & C_{1,2} & C_{1,3} \\ C_{2,0} & C_{2,1} & C_{2,2} & C_{2,3} \\ C_{3,0} & C_{3,1} & C_{3,2} & C_{3,3} \end{pmatrix}$$

FP16 or FP32 | FP16 | FP16 | FP16 or FP32

$$D = AB + C$$

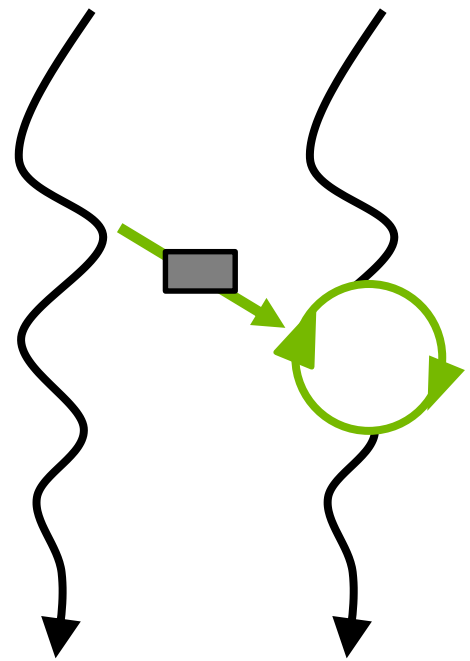
VOLTA: INDEPENDENT THREAD SCHEDULING

Communicating Algorithms



Pascal: Lock-Free Algorithms

Threads cannot wait for messages

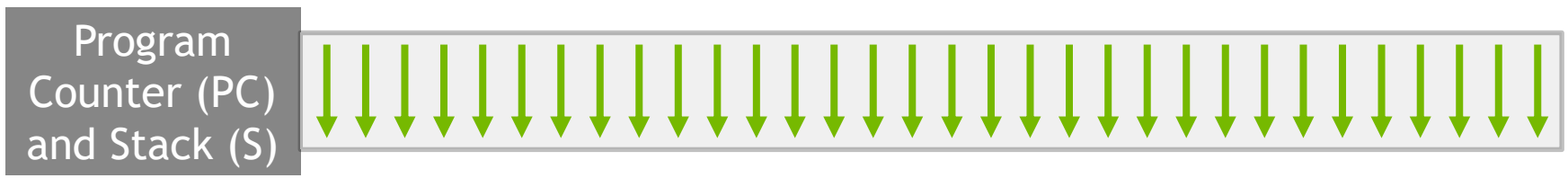


Volta: Starvation Free Algorithms

Threads may wait for messages

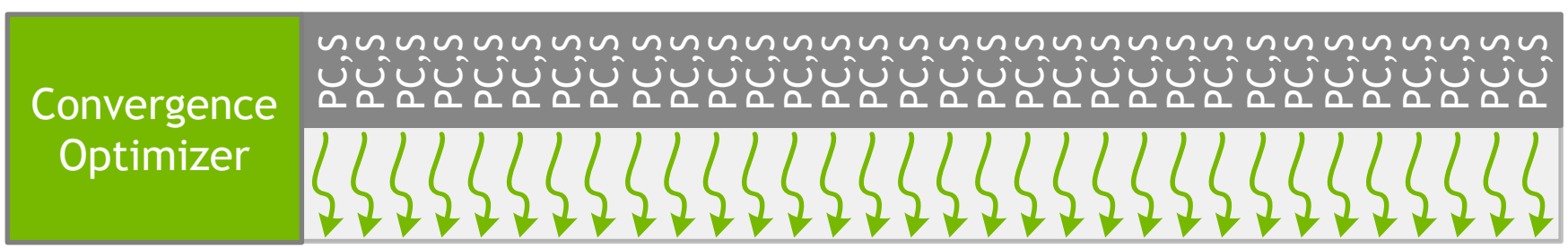
WARP IMPLEMENTATION

Pre-Volta



32 thread warp

Volta



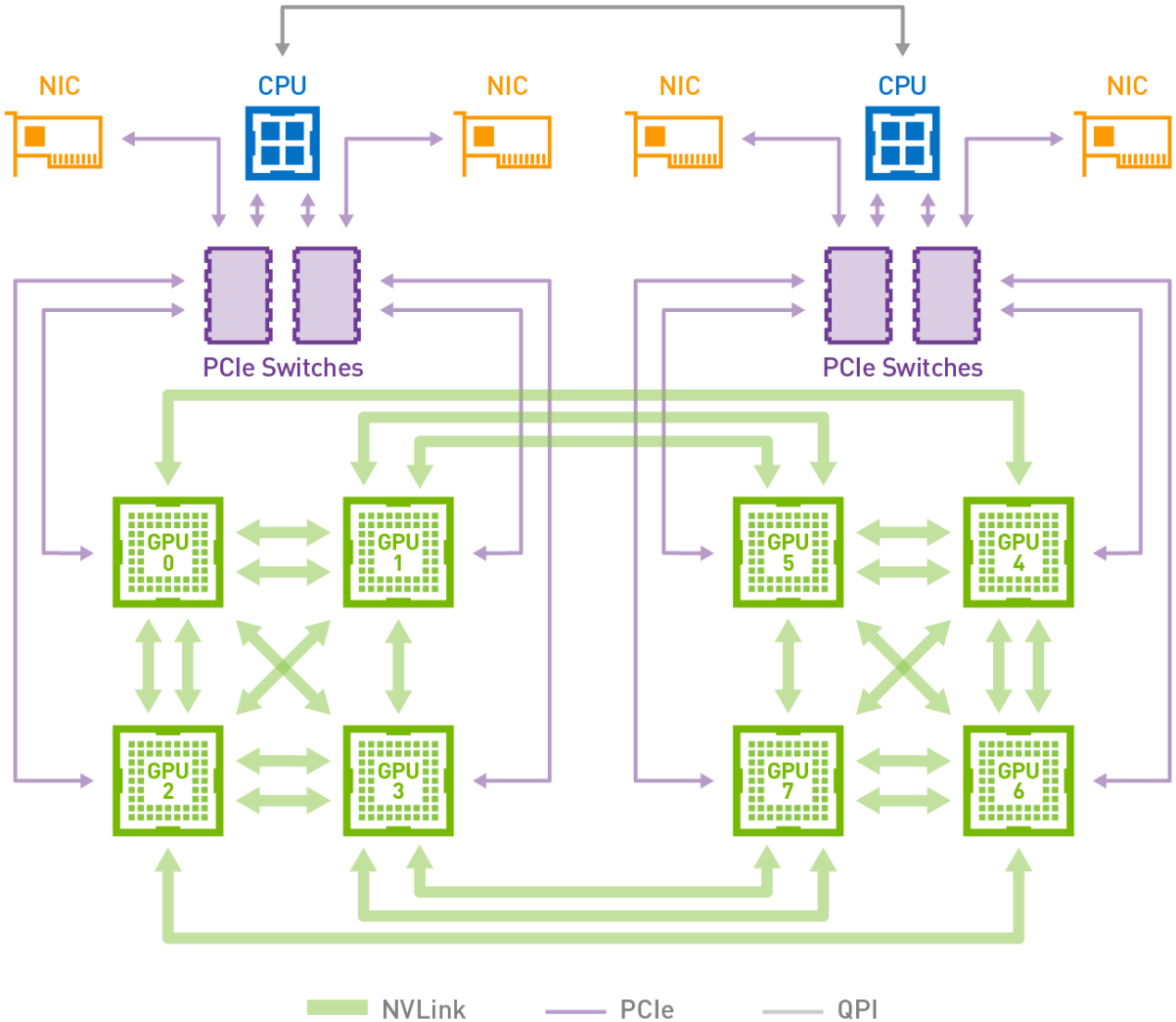
32 thread warp with independent scheduling

VOLTA NVLINK

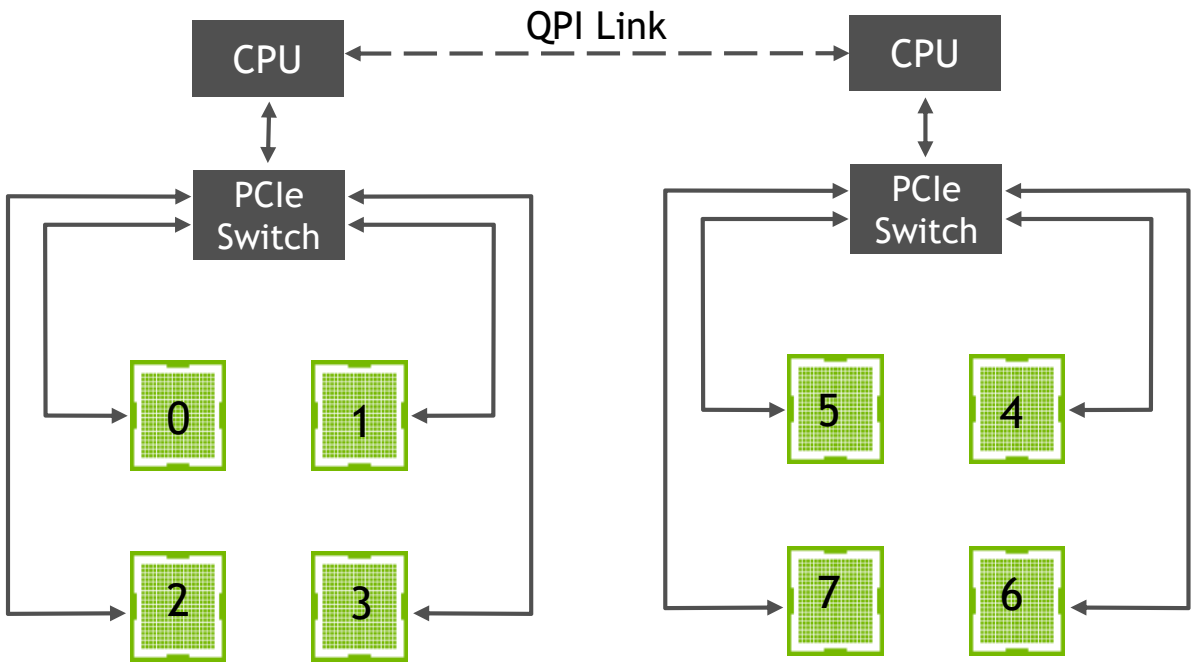
300GB/sec

50% more links

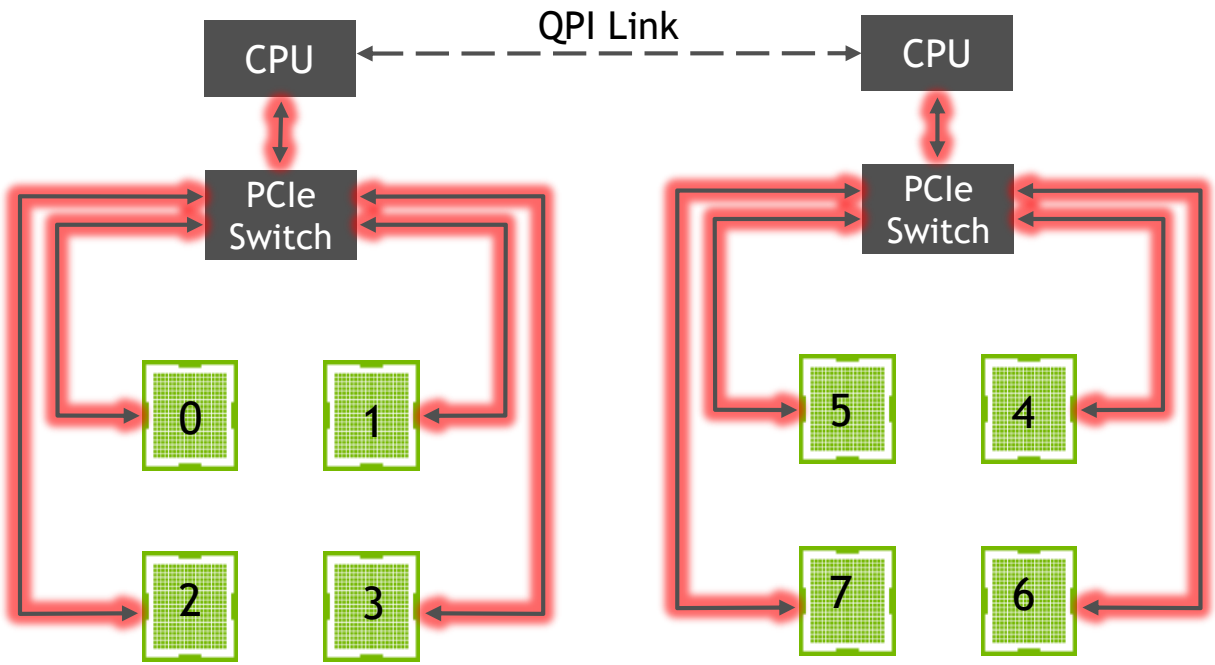
28% faster signaling



DL DATA PARALLELISM - PCIE BASED

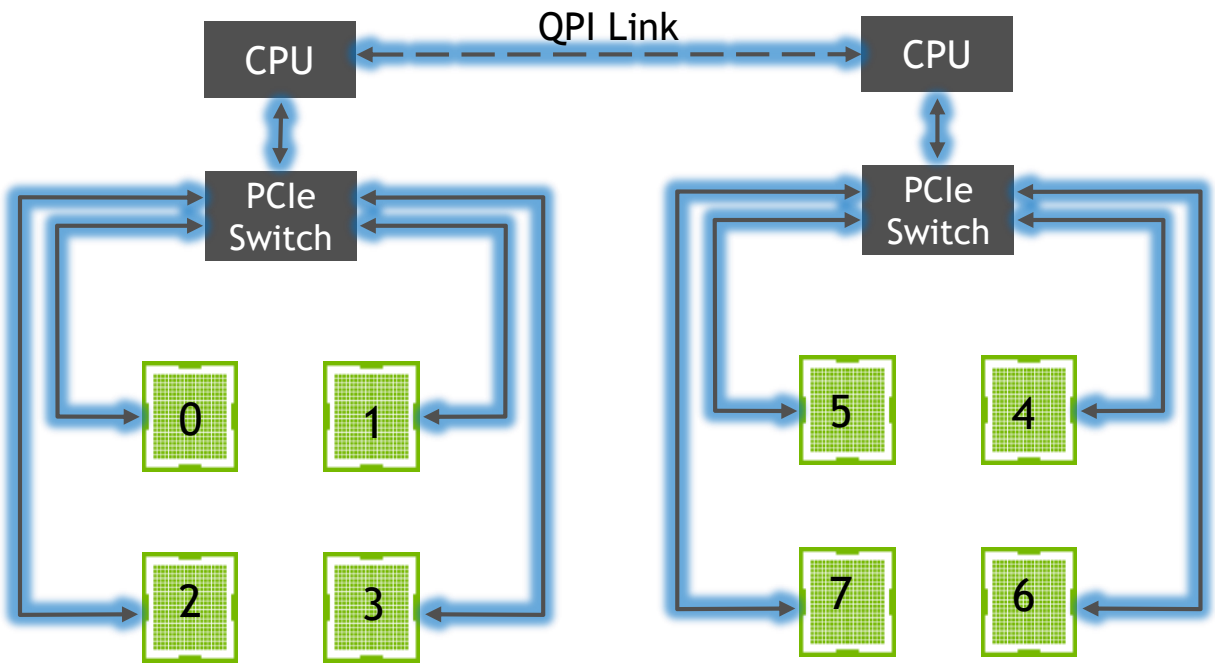


DL DATA PARALLELISM - PCIE BASED



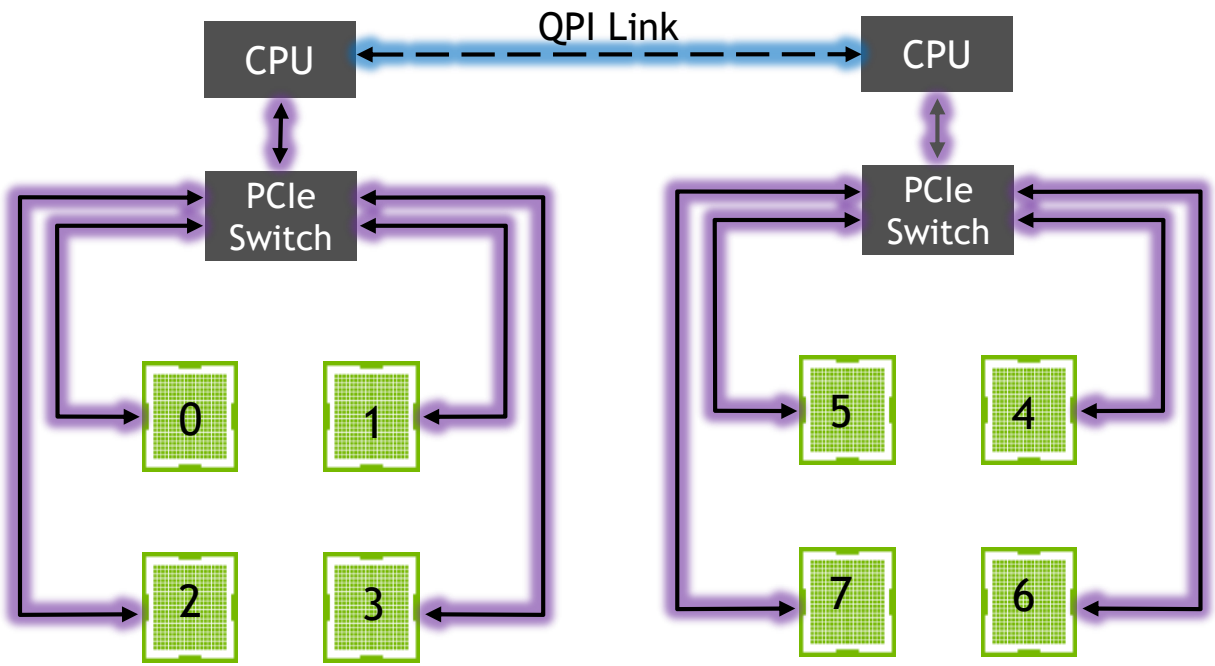
Data loading over PCIe

DL DATA PARALLELISM - PCIE BASED



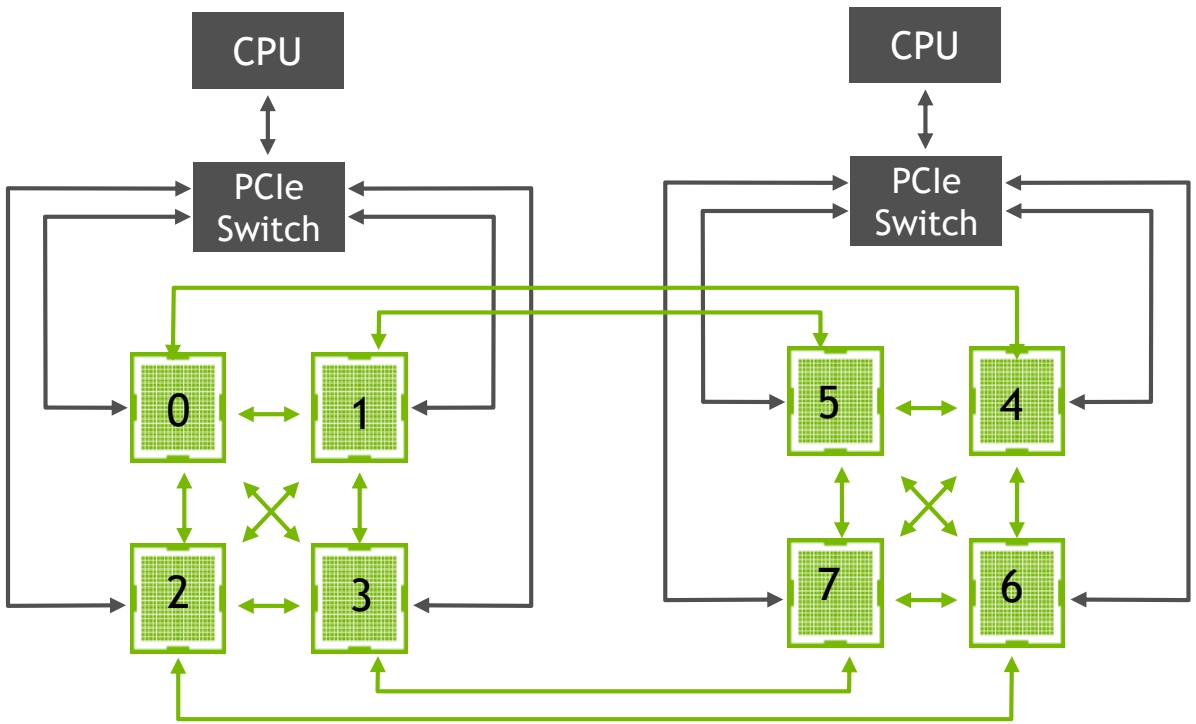
Gradient averaging over PCIe and QPI

DL DATA PARALLELISM - PCIE BASED

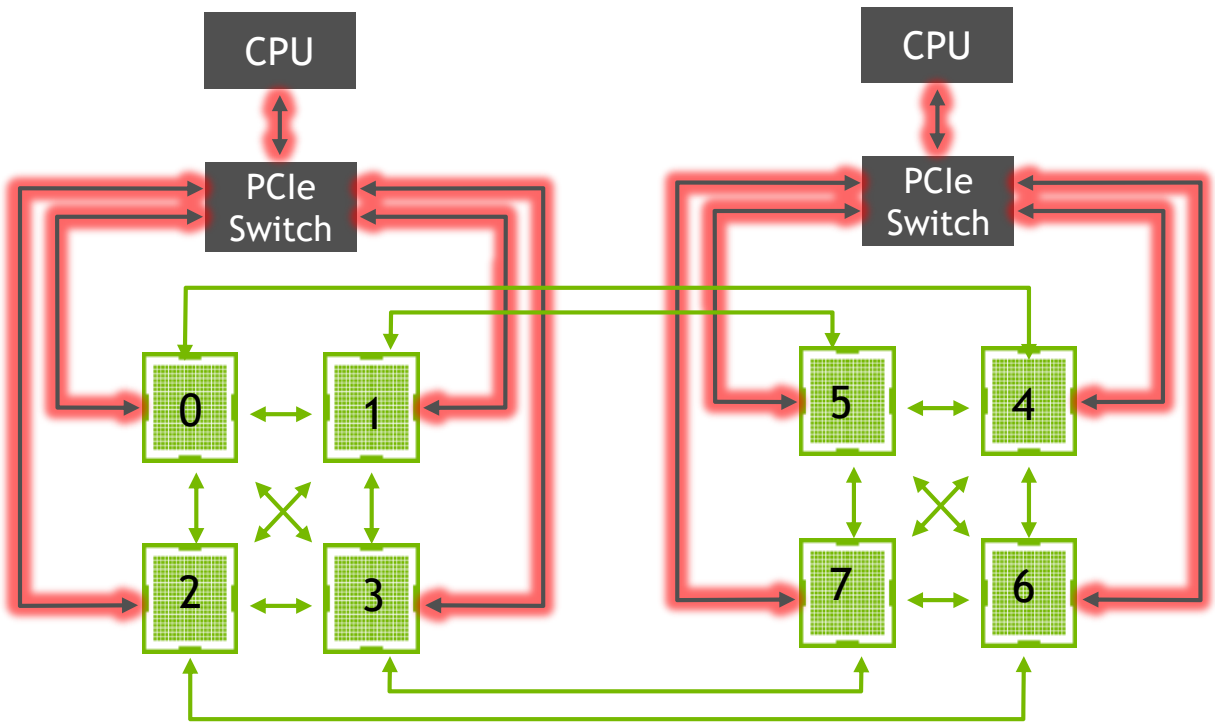


Data loading and gradient averaging share communication resources: Congestion

DL DATA PARALLELISM - NVLINK

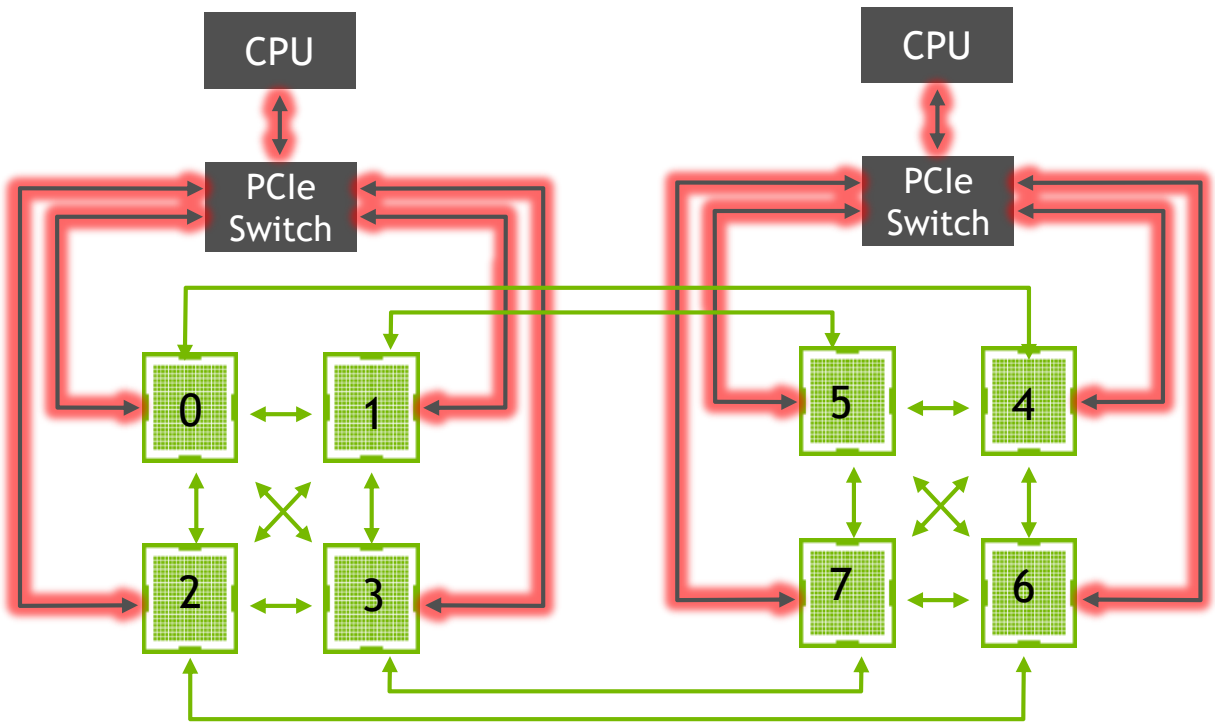


DL DATA PARALLELISM - NVLINK



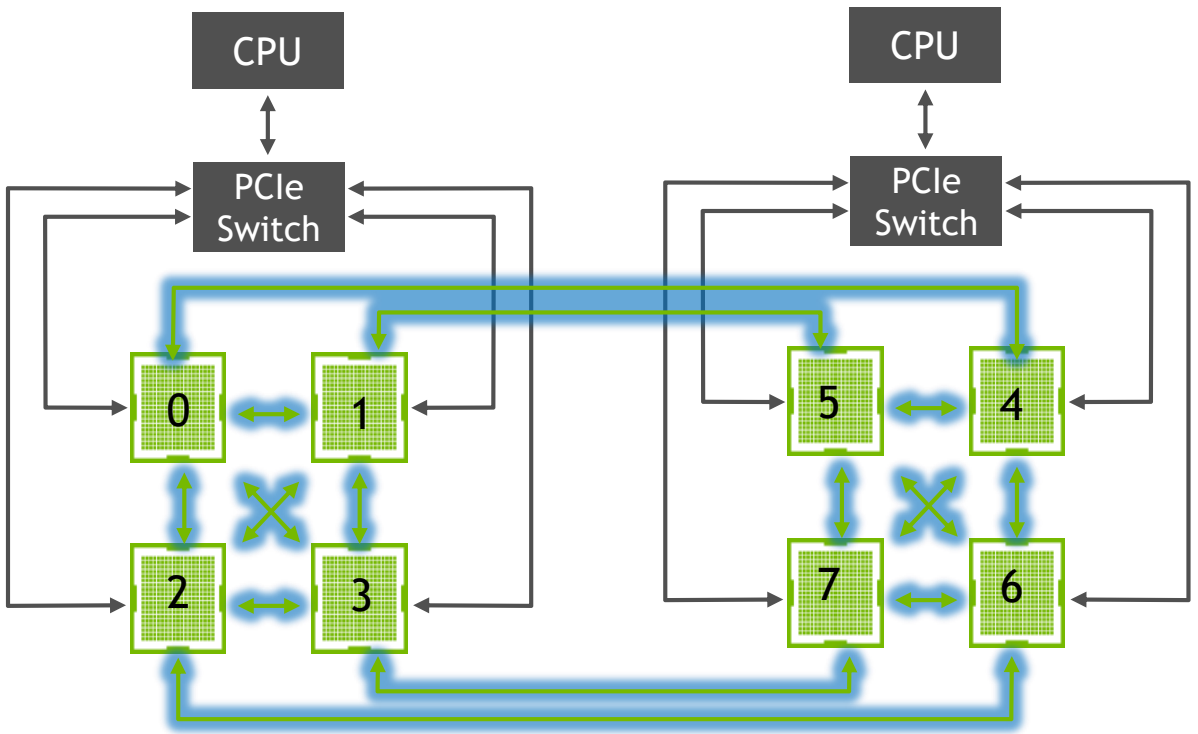
Data loading over PCIe

DL DATA PARALLELISM - NVLINK



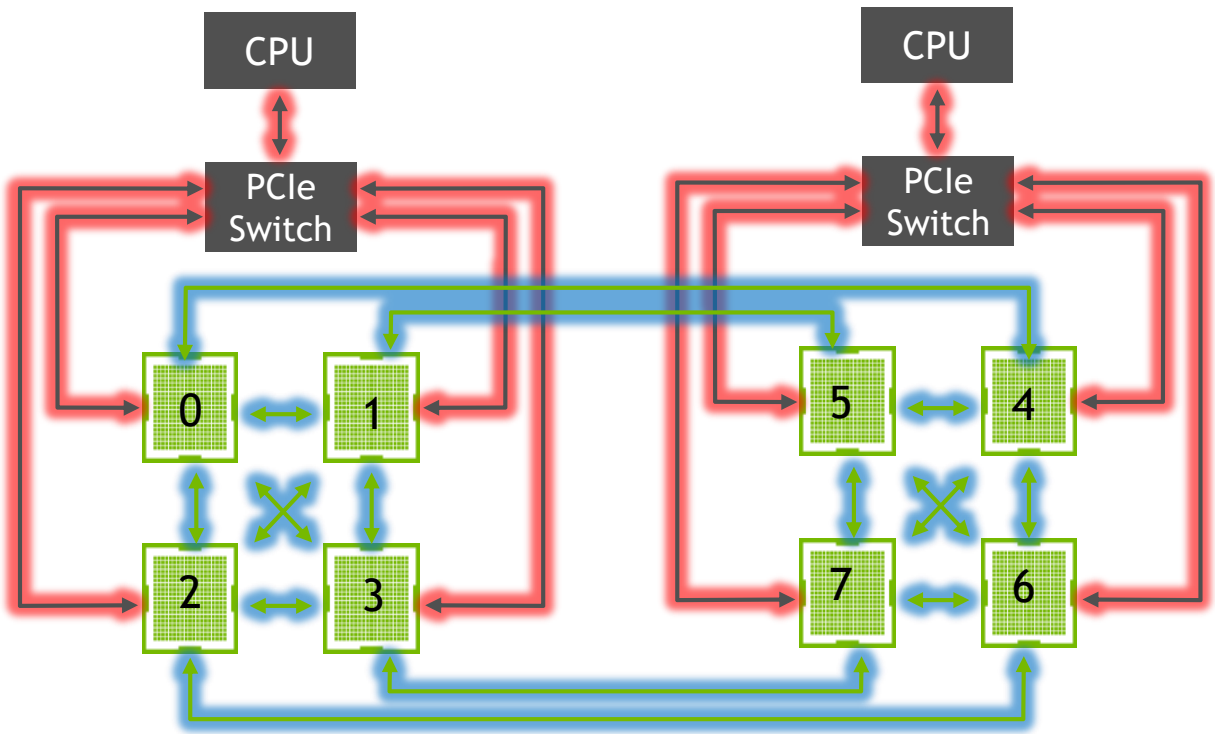
Data loading over PCIe

DL DATA PARALLELISM - NVLINK



Gradient averaging over NVLink

DL DATA PARALLELISM - NVLINK

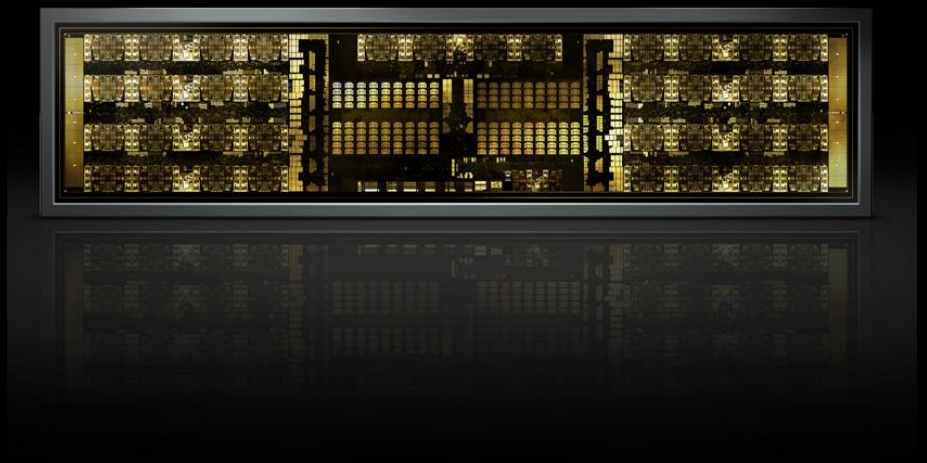


No sharing of communication resources: No congestion

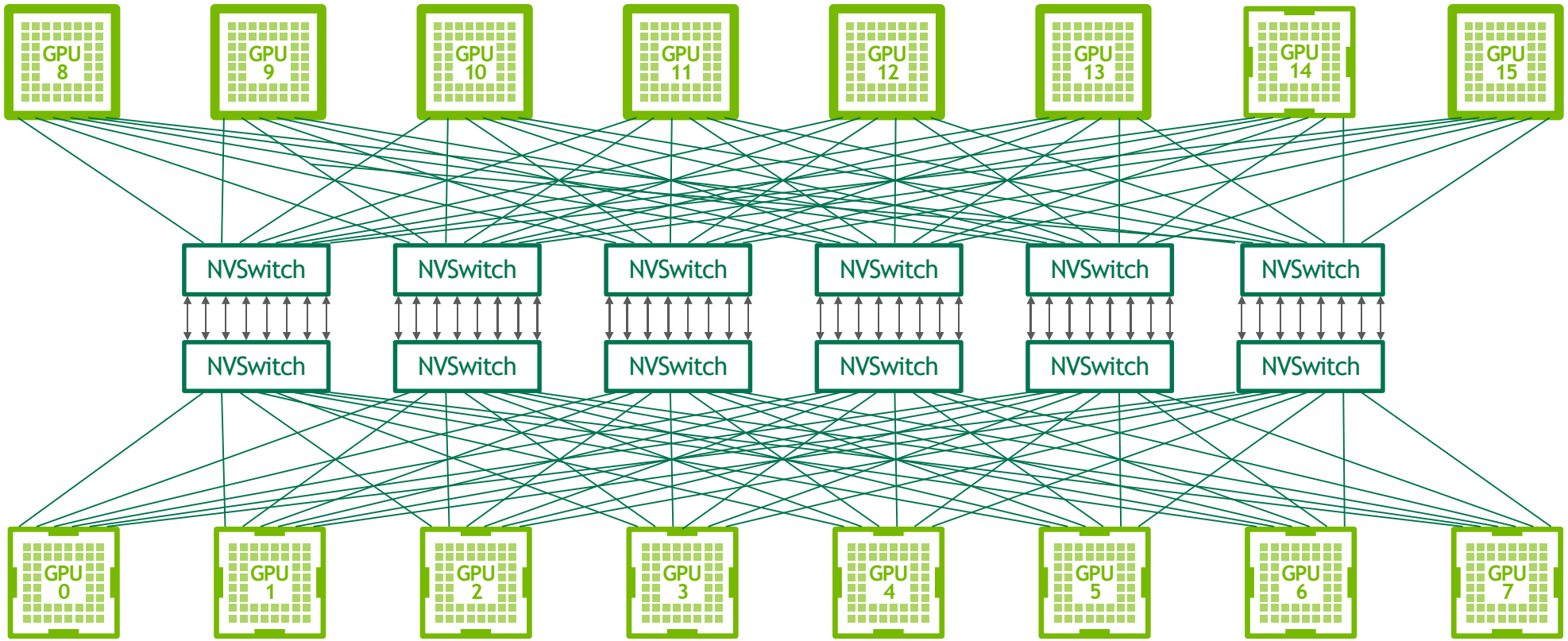
NVSWITCH

WORLD'S HIGHEST BANDWIDTH ON-NODE SWITCH

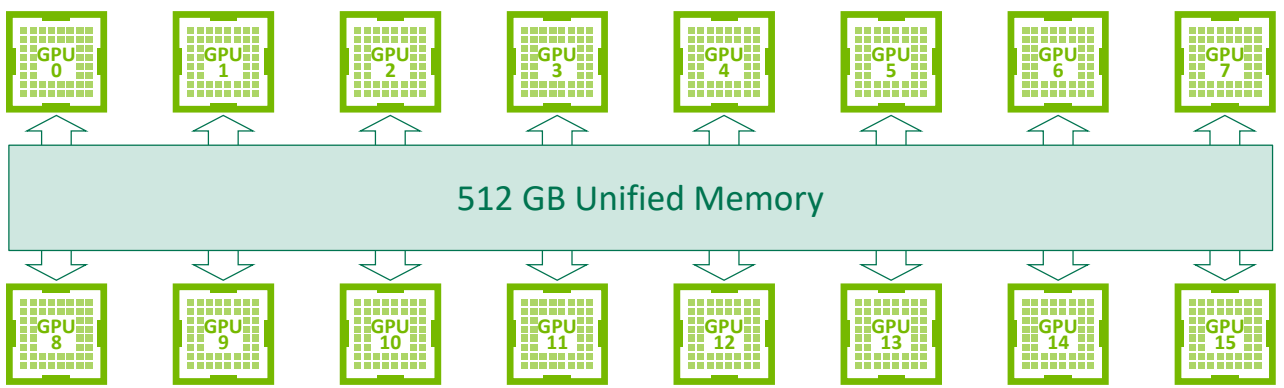
7.2 Terabits/sec or 900 GB/sec
18 NVLINK ports | 50GB/s per port bi-directional
Fully-connected crossbar
2 billion transistors | 47.5mm x 47.5mm package



FULL NON-BLOCKING BANDWIDTH



UNIFIED MEMORY + DGX-2



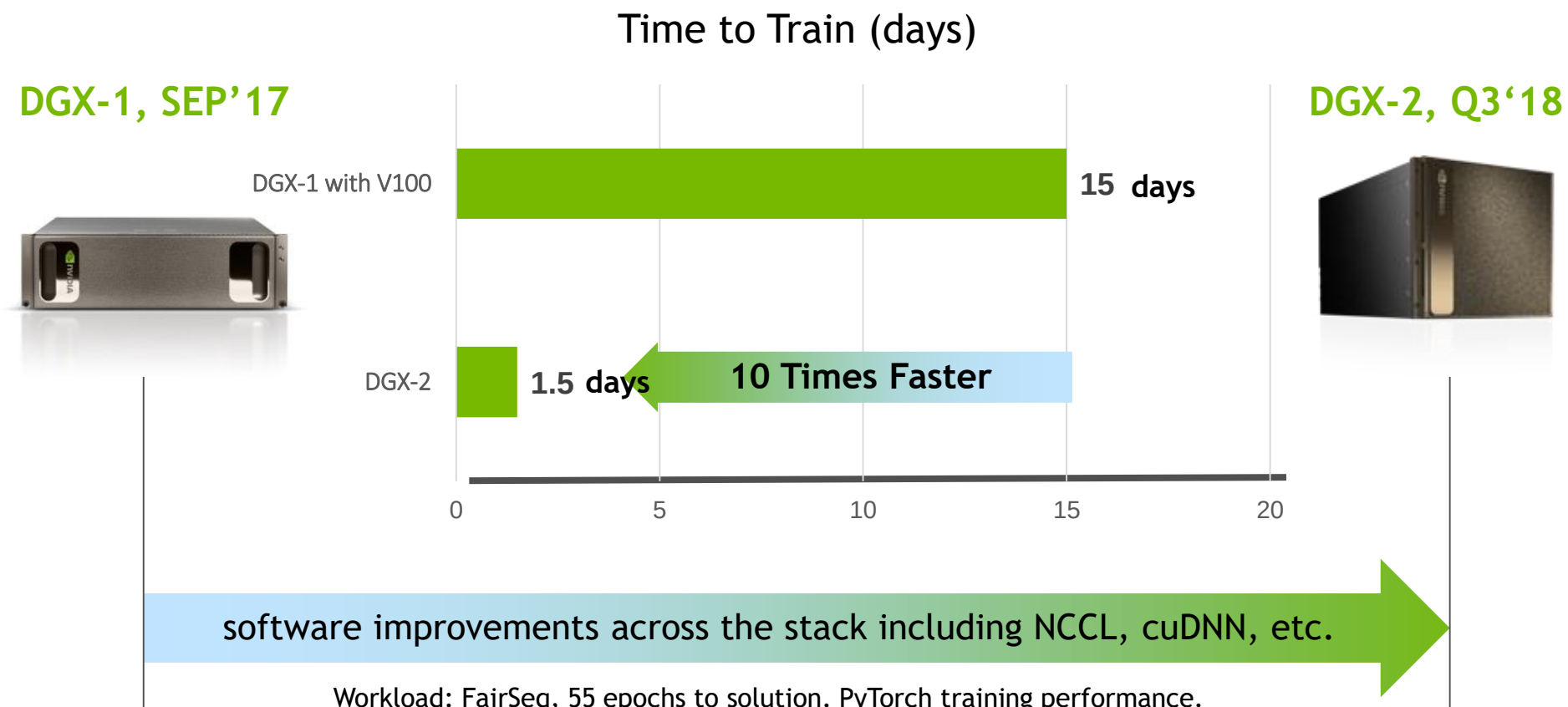
UNIFIED MEMORY PROVIDES

Single memory view shared by all GPUs

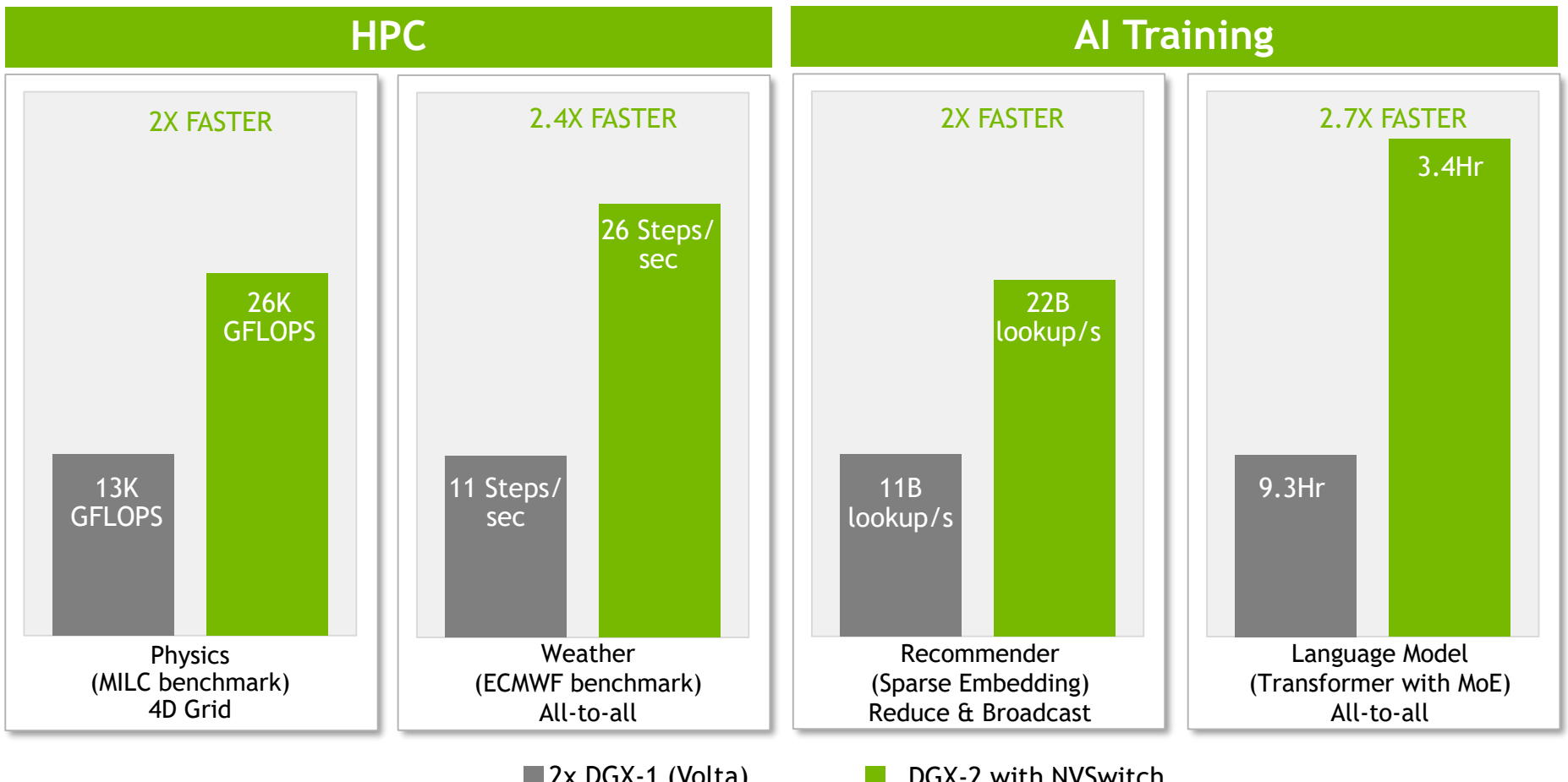
Automatic migration of data between GPUs

User control of data locality

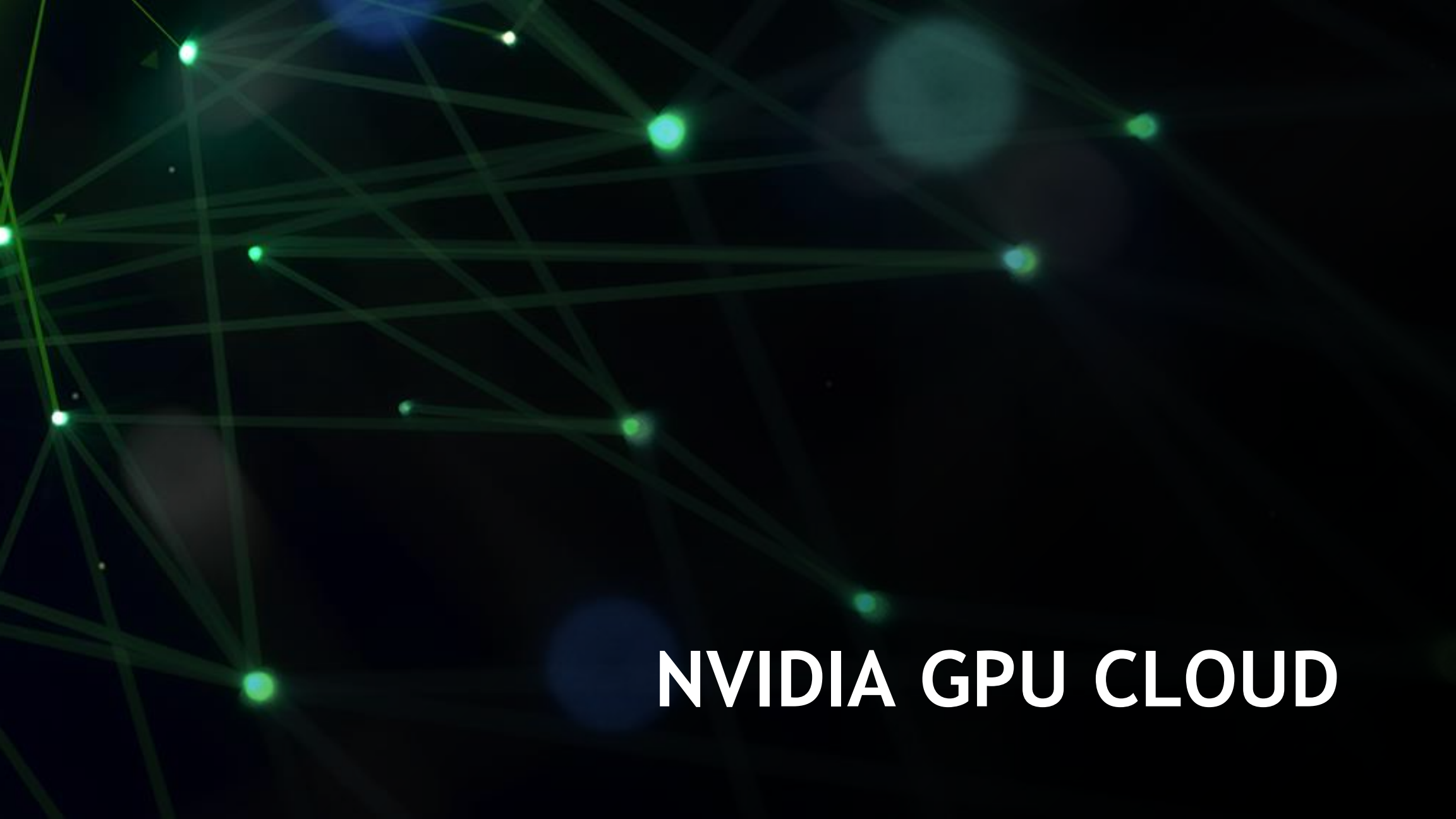
10X PERFORMANCE GAIN IN LESS THAN A YEAR



UP TO 3X HIGHER PERFORMANCE WITH NVSWITCH



2 HGX-1V servers have dual socket Xeon E5 2698v4 Processor. 8 x V100 GPUs. Servers connected via 4XIB ports | HGX-2 server has dual-socket Xeon Platinum 8168 Processor. 16 V100 GPUs

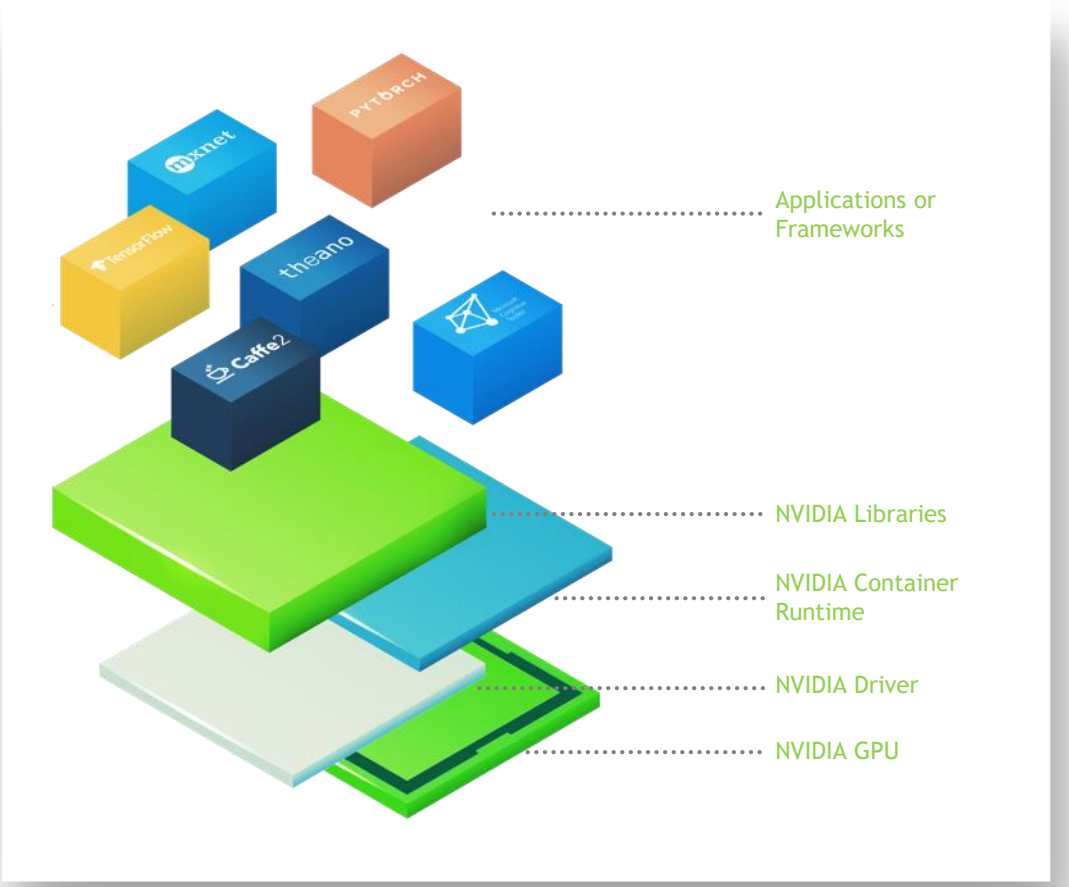

NVIDIA GPU CLOUD

CHALLENGES WITH COMPLEX SOFTWARE

Current DIY GPU-accelerated AI and HPC deployments are **complex** and **time consuming** to build, test and maintain

Development of software frameworks by the community is moving **very fast**

Requires high level of **expertise** to manage driver, library, framework dependencies



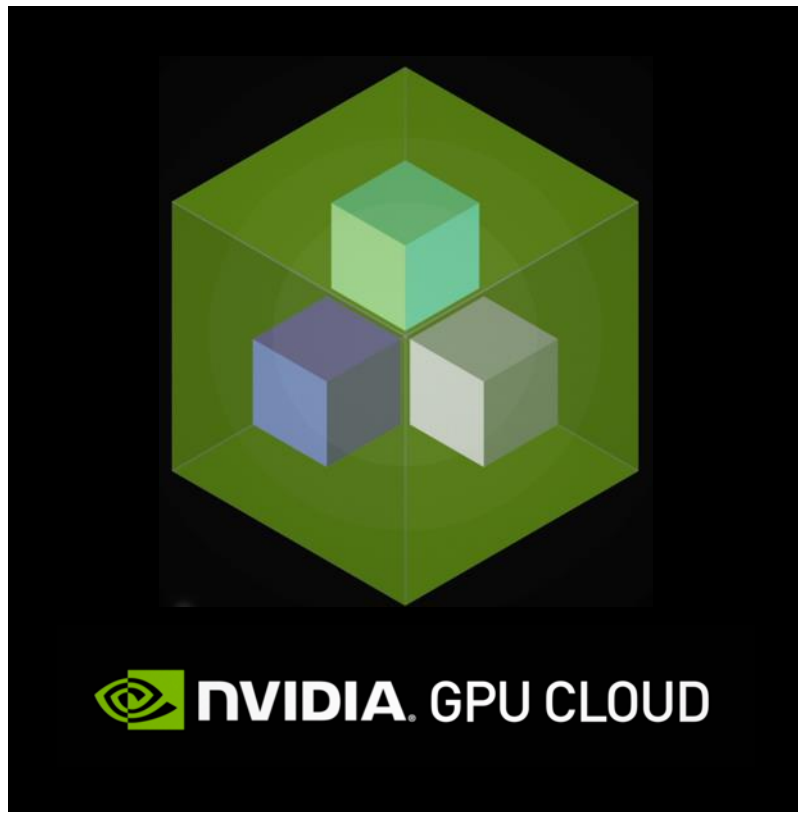
NGC

Accelerated Stacks for AI, Machine Learning, and HPC

Comprehensive Library of GPU-Accelerated Containers

Innovate In Minutes, Not Weeks

Run Anywhere



THE DESTINATION FOR GPU-ACCELERATED SOFTWARE

HPC	Deep Learning	Machine Learning	Inference	Visualization	Infrastructure
BigDFT	Caffe2	H2O Driverless AI	DeepStream	Index	Kubernetes on NVIDIA GPUs
CANDLE	Chainer	Kinetica	DeepStream 360d	ParaView	
CHROMA	CUDA	MATLAB	TensorRT	ParaView Holodeck	
GAMESS	Deep Cognition Studio	OmniSci (MapD)	TensorRT Inference Server	ParaView Index	
GROMACS	DIGITS	RAPIDS		ParaView Optix	
LAMMPS	Microsoft Cognitive Toolkit				
Lattice Microbes	MXNet				
MILC	NVCaffe				
NAMD	PaddlePaddle				
PGI Compilers	PyTorch				
PicOnGPU	TensorFlow				
QMCPACK	Theano				
RELION	Torch				
vmd					

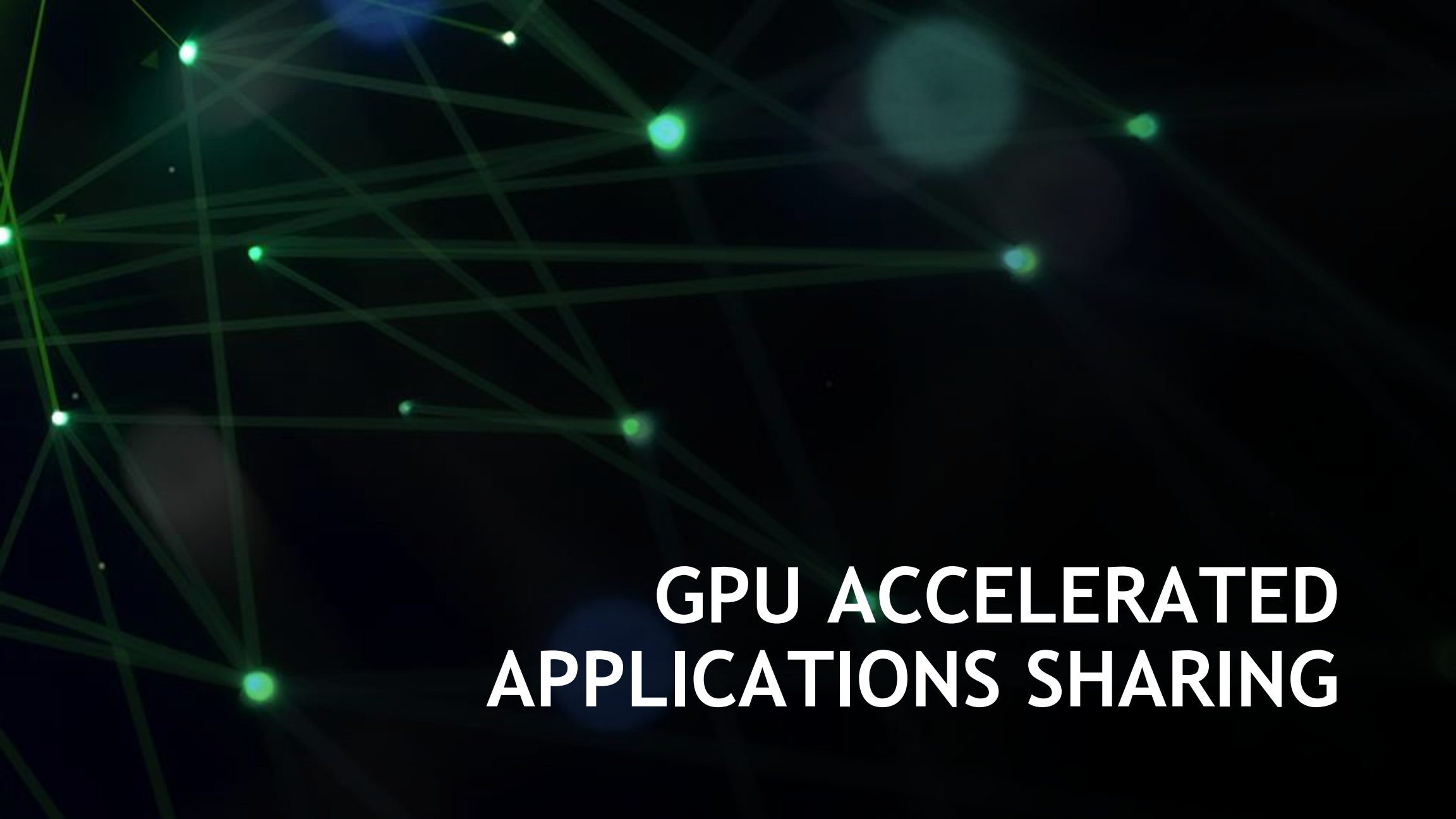
10 containers

SOFTWARE ON THE NGC CONTAINER REGISTRY

42 containers

October 2017

November 2018


GPU ACCELERATED
APPLICATIONS SHARING

GPU-ACCELERATED HPC APPLICATIONS

580+ Applications

LIFE SCIENCES

50+ app

Including:
- Gaussian
- VASP
- AMBER
- HOOMD-Blue
- GAMESS

MFG, CAD, & CAE

111 apps

Including:
- Ansys Fluent
- Abaqus SIMULIA
- AutoCAD
- CST Studio Suite

PHYSICS

25 apps

Including:
- QUDA
- MILC
- GTC-P

OIL & GAS

18 apps

Including:
- RTM
- SPECFEM 3D

CLIMATE & WEATHER

3 apps

Including:
- Cosmos
- Gales
- WRF

DEEP LEARNING

38 apps

Including:
- Caffe2
- MXNet
- Tensorflow

MEDIA & ENT.

142 apps

Including:
- DaVinci Resolve
- Premiere Pro CC
- Redshift Renderer

FEDERAL & DEFENSE

14 apps

Including:
- ArcGIS Pro
- EVNI
- SocetGXP

DATA SCI. & ANALYTICS

23 apps

Including:
- MapD
- Kinetica
- Graphistry

SAFETY & SECURITY

19 apps

Including:
- Cyllance
- FaceControl
- Syndex Pro

COMP. FINANCE

16 apps

Including:
- O-Quant Options Pricing
- MUREX
- MISYS

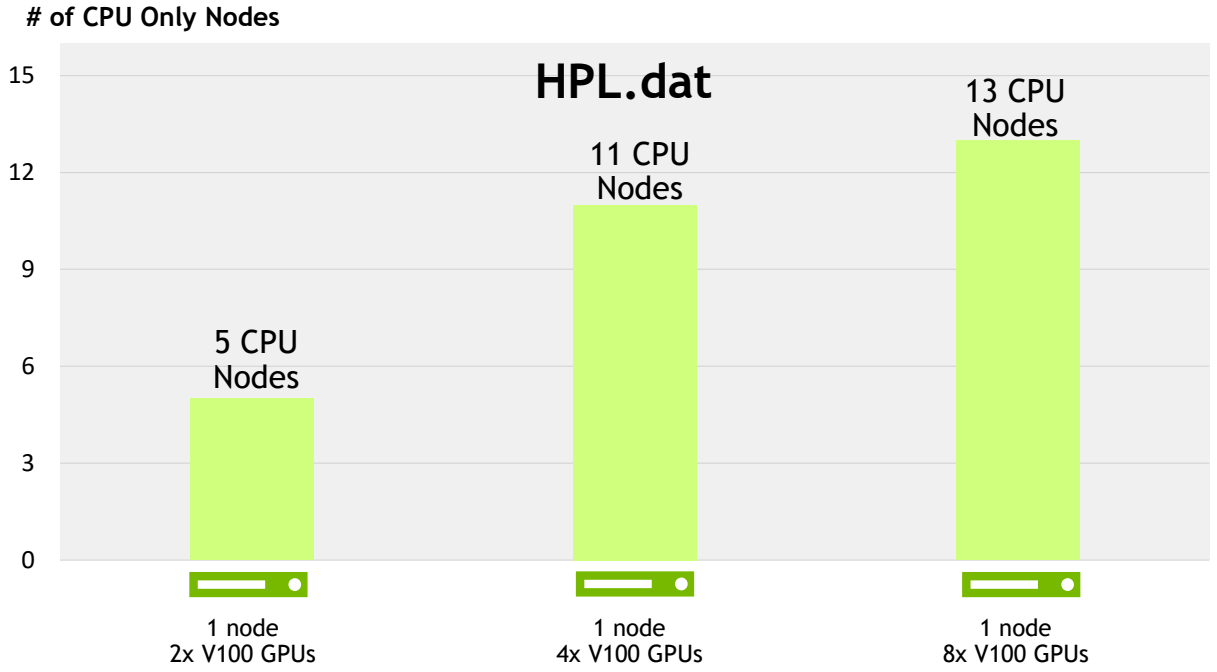
TOOLS & MGMT.

16 apps

Including:
- Bright Cluster Manager
- HPCtoolkit
- Vampir

Linpack Performance Equivalency

Single GPU Node vs Multiple Skylake CPU-Only Nodes



CPU Server: Dual Xeon Gold 6140@2.30GHz, GPU Servers: same CPU server w/ V100 PCIe
CUDA Version: CUDA 9.0.103; Dataset: HPL.dat
To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

Linpack

Benchmark

Measures floating point computing power

The LINPACK Benchmarks are a measure of a system's floating point computing power. Introduced by Jack Dongarra, they measure how fast a computer solves a dense n by n system of linear equations Ax = b, which is a common task in engineering.

VERSION
2.1

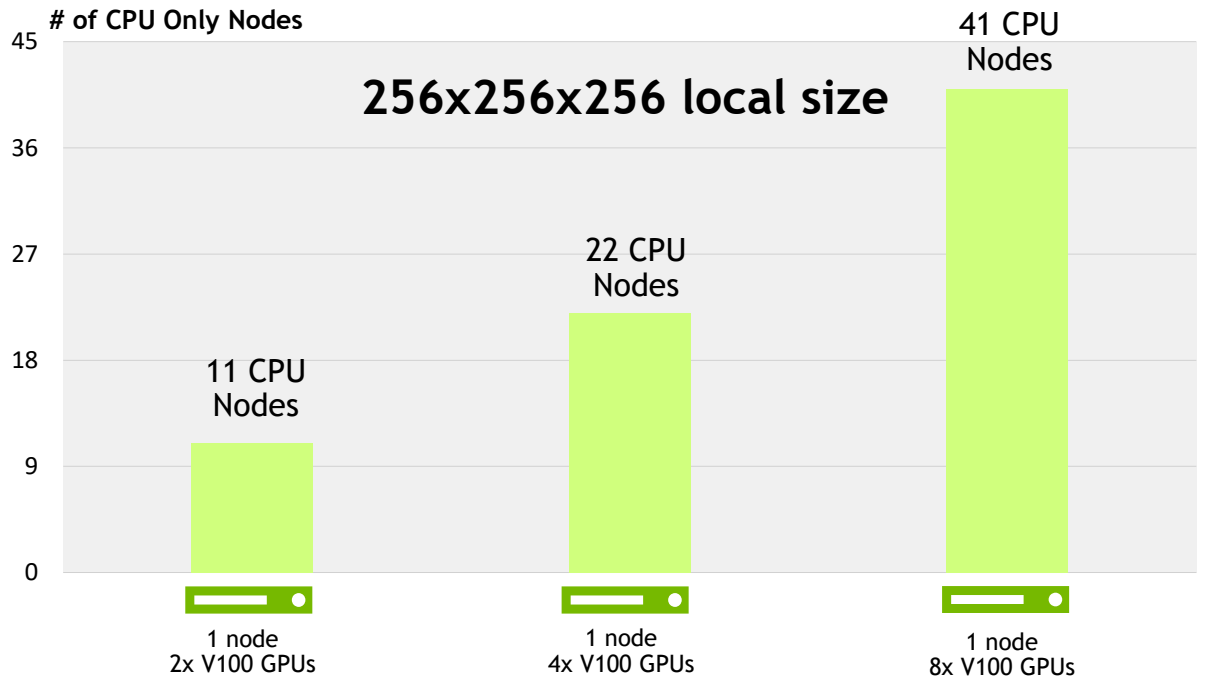
ACCELERATED FEATURES
All

SCALABILITY
Multi-GPU and Multi-Node

MORE INFORMATION
https://www.top500.org/project/**linpack**/

HPCG Performance Equivalency

Single GPU Node vs Multiple Skylake CPU-Only Nodes



Speed up vs CPU server	11x	22x	41x

CPU Server: Dual Xeon Gold 6140@2.30GHz, GPU Servers: same CPU server w/ V100 PCIe
CUDA Version: CUDA 9.0.103; Dataset: 256x256x256 local size
To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

HPCG

Benchmark

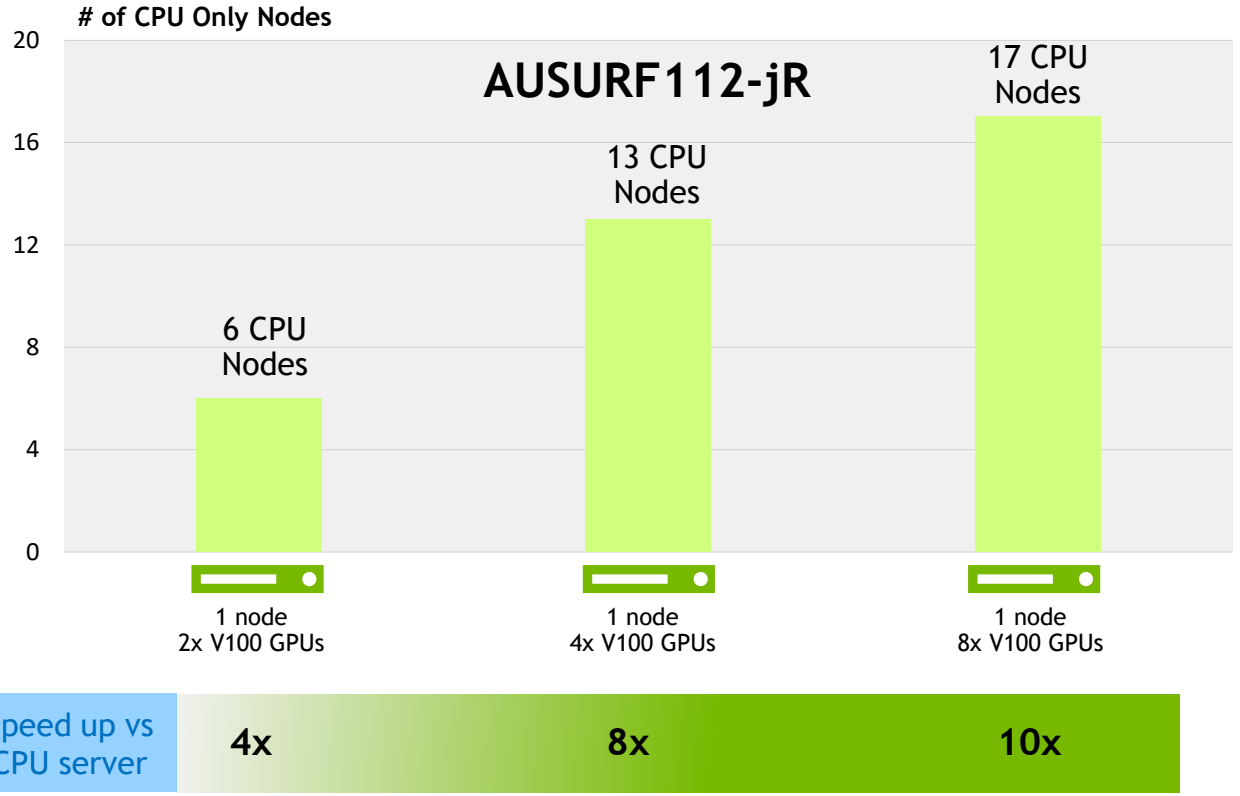
HPCG performance can be impacted by many system capabilities, each of which has strong correlation to performance requirements of real applications

- **Sparse Matrix Vector Multiplication (SpMV)**
- **Symmetric Gauss-Seidel smoother (SymGS)**
- **Global Dot Product**
- **Vector Update**
- **Multigrid preconditioner**

VERSION
3
ACCELERATED FEATURES
All
SCALABILITY
Multi-GPU and Multi-Node
MORE INFORMATION
http://www.hpcg-benchmark.org/index.html

Quantum Espresso Performance Equivalency

Single GPU Node vs Multiple Skylake CPU-Only Nodes



CPU Server: Dual Xeon Gold 6140@2.30GHz, GPU Servers: same CPU server w/ V100 PCIe or V100 SXM2 on 8X V100 config
CUDA Version: CUDA 9.2.88; Dataset: AUSURF112-jR
To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

Quantum Espresso

Material Science (Quantum Chemistry)

An Open-source suite of computer codes for electronic structure calculations and materials modeling at the nanoscale

VERSION
6.1

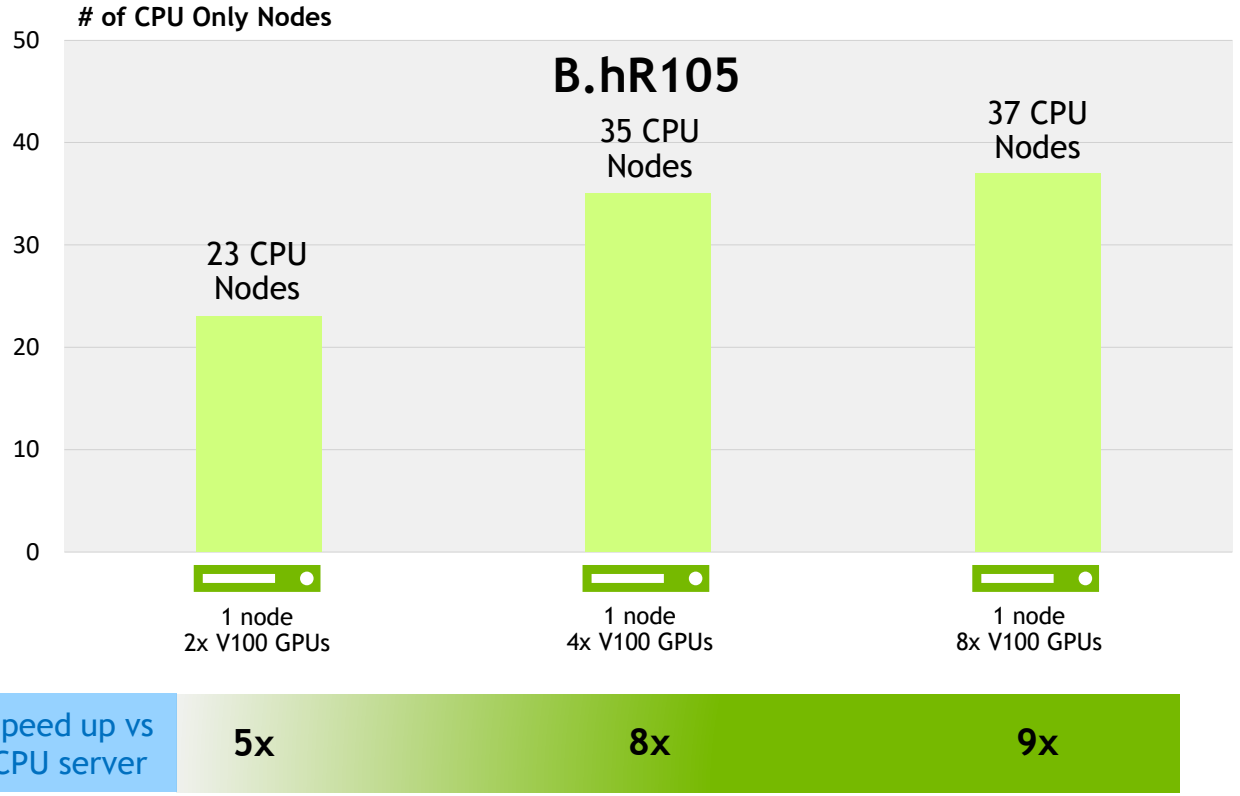
ACCELERATED FEATURES
Linear algebra (matix multiply), explicit computational kernels, 3D FFTs

SCALABILITY
Multi-GPU and Multi-Node

MORE INFORMATION
http://www.quantum-espresso.org

VASP Performance Equivalency

Single GPU Node vs Multiple Skylake CPU-Only Nodes



CPU Server: Dual Xeon Gold 6140@2.30GHz, GPU Servers: same CPU server w/ V100 PCIe or V100 SXM2 on 8X V100 config
CUDA Version: CUDA 9.0.176; Dataset: B.hR105
To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

VASP

Material Science (Quantum Chemistry)

Complex package for performing ab-initio quantum-mechanical molecular dynamics (MD) simulations using pseudopotentials or the projector-augmented wave method and a plane wave basis set

VERSION

5.4.4

ACCELERATED FEATURES

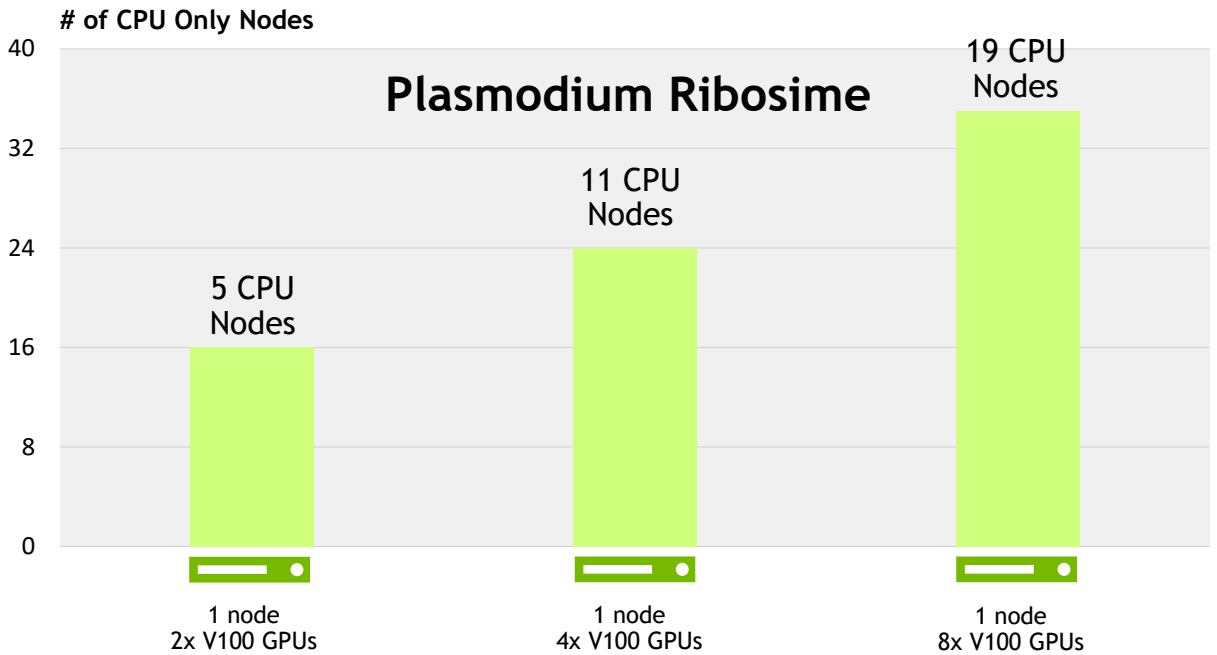
Blocked Davidson (ALGO = NORMAL & FAST), RMM-DIIS (ALGO = VERYFAST & FAST), K-Points and optimization

SCALABILITY

Multi-GPU and Single Node

Relion Performance Equivalency

Single GPU Node vs Multiple Skylake CPU-Only Nodes



Speed up vs CPU server	13x	20x	28x

CPU Server: Dual Xeon Gold 6140@2.30GHz, GPU Servers: same CPU server w/ V100 PCIe
CUDA Version: CUDA 9.0.176; Dataset: Plasmodium Ribosime
To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

Relion

Microscopy

Stand-alone computer program that employs an empirical Bayesian approach to refinement of (multiple) 3Dreconstructions or 2D class averages in electron cryo-microscopy (cryo-EM)

VERSION

2.0.3

ACCELERATED FEATURES

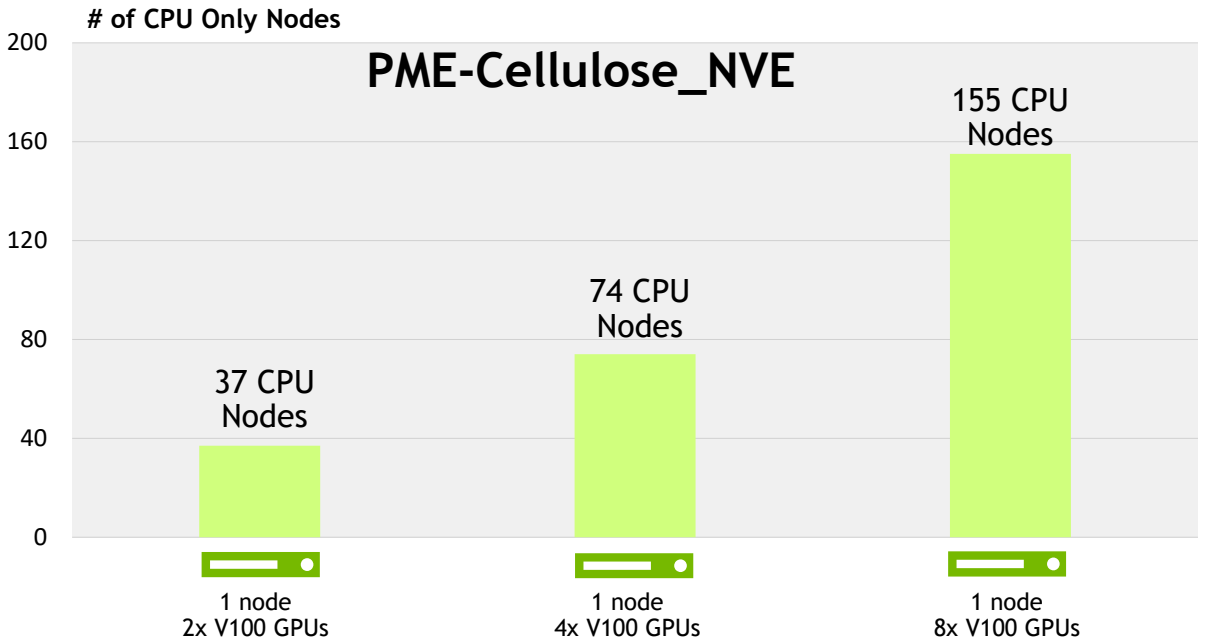
Reduced memory requirements; high-resolution cryo-EM structure determination in a matter of day on a single workstation

SCALABILITY

Multi-GPU and Single Node

AMBER Performance Equivalency

Single GPU Node vs Multiple Skylake CPU-Only Nodes



Speed up vs CPU server	37x	74x	155x

CPU Server: Dual Xeon Gold 6140@2.30GHz, GPU Servers: same CPU server w/ V100 PCIe or V100 SXM2 on 8X V100 config
CUDA Version: CUDA 10.0.130; Dataset: PME-Cellulose_NVE
To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

AMBER

Molecular Dynamics

Suite of programs to simulate molecular dynamics on biomolecule

VERSION
18.6

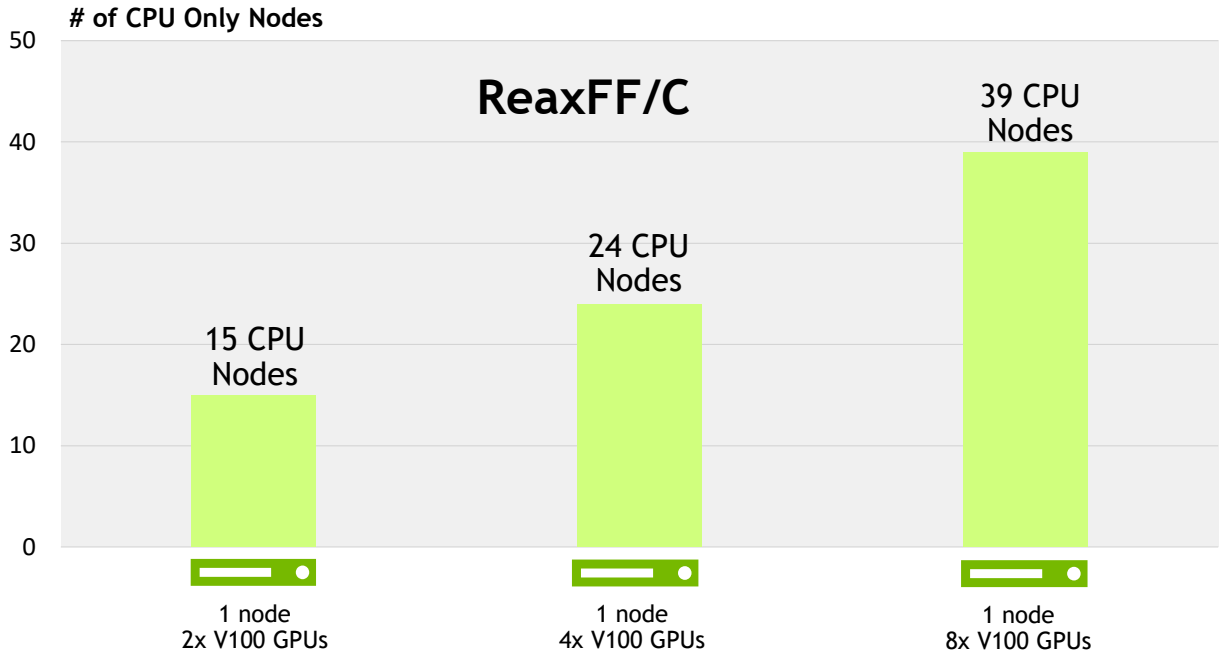
ACCELERATED FEATURES
PMEMD Explicit Solvent and GB Implicit Solvent

SCALABILITY
Multi-GPU and Single Node

MORE INFORMATION
http://ambermd.org/gpus

LAMMPS Performance Equivalency

Single GPU Node vs Multiple Skylake CPU-Only Nodes



CPU Server: Dual Xeon Gold 6140@2.30GHz, GPU Servers: same CPU server w/ V100 PCIe or V100 SXM2 on 8X V100 config
CUDA Version: CUDA 10.0.130, Dataset: ReaxFF/C
To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

LAMMPS

Molecular Dynamics

Classical molecular dynamics package

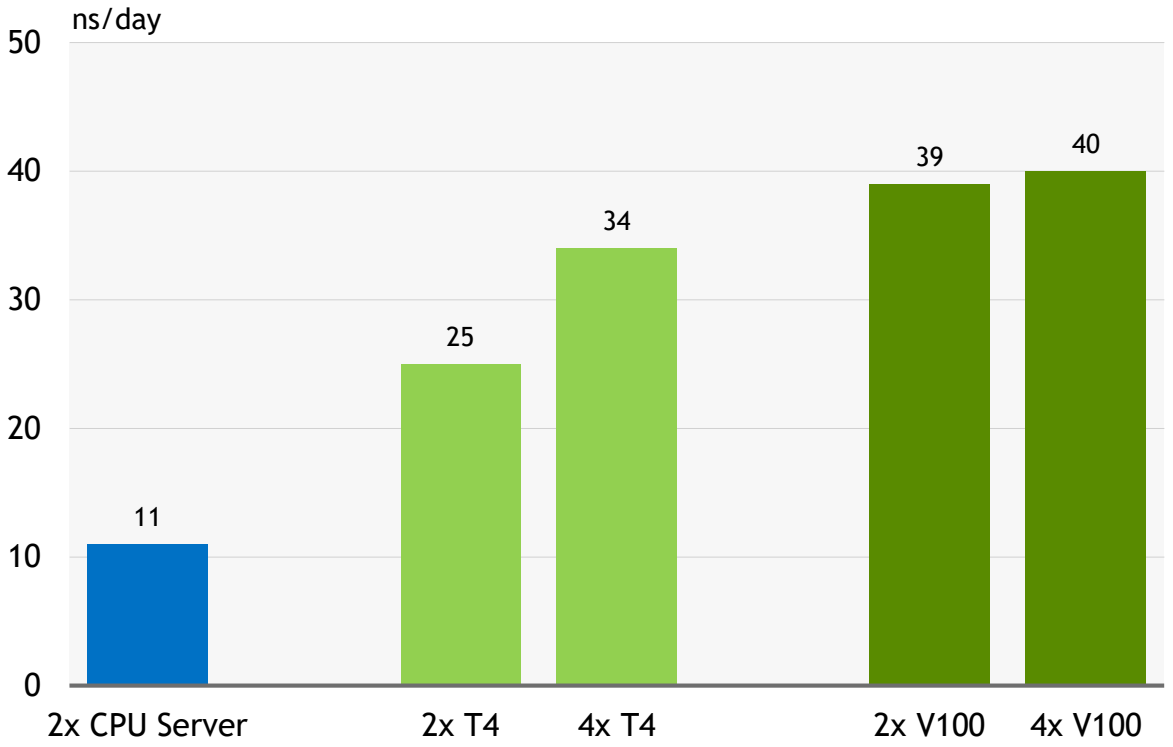
VERSION
2018

ACCELERATED FEATURES
Lennard-Jones, Gay-Berne, Tersoff, many more potentials

SCALABILITY
Multi-GPU and Multi-Node

More Information
http://lammps.sandia.gov/index.html

GROMACS Raw Performance



CPU: Dual Xeon Gold 6140@2.30GHz, GPU: Dual Xeon Gold 6140@2.30GHz with GPU servers as shown | CUDA 10.0.130 GROMACS v2018, Dataset: Cellulose

GROMACS

Molecular Dynamics

Simulation of biochemical molecules with complicated bond interactions

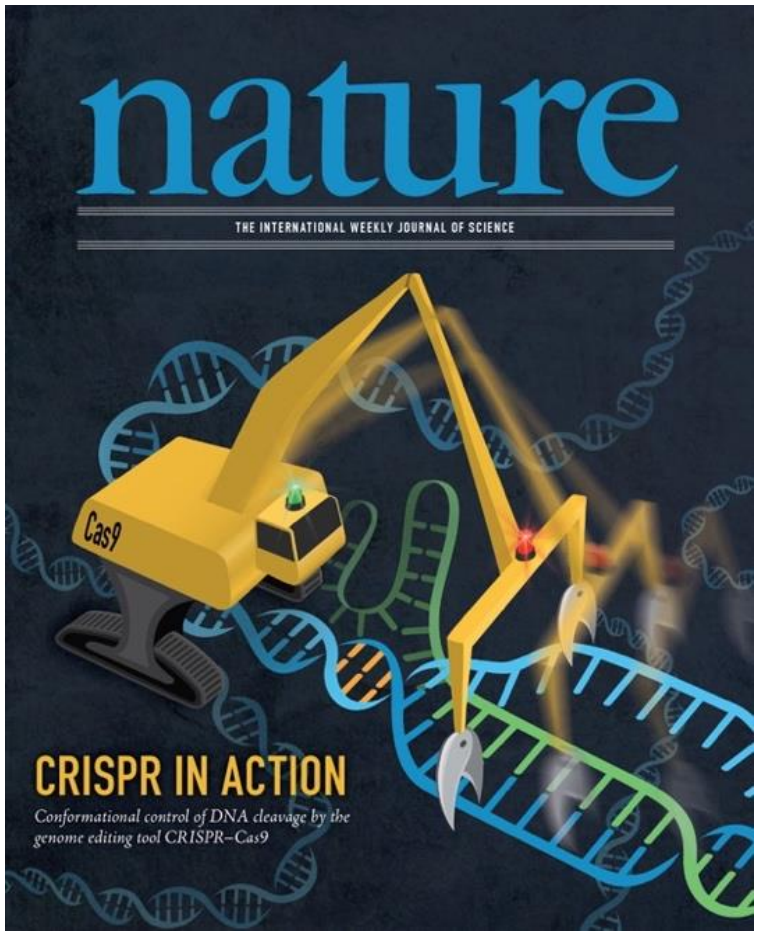
VERSION
2018

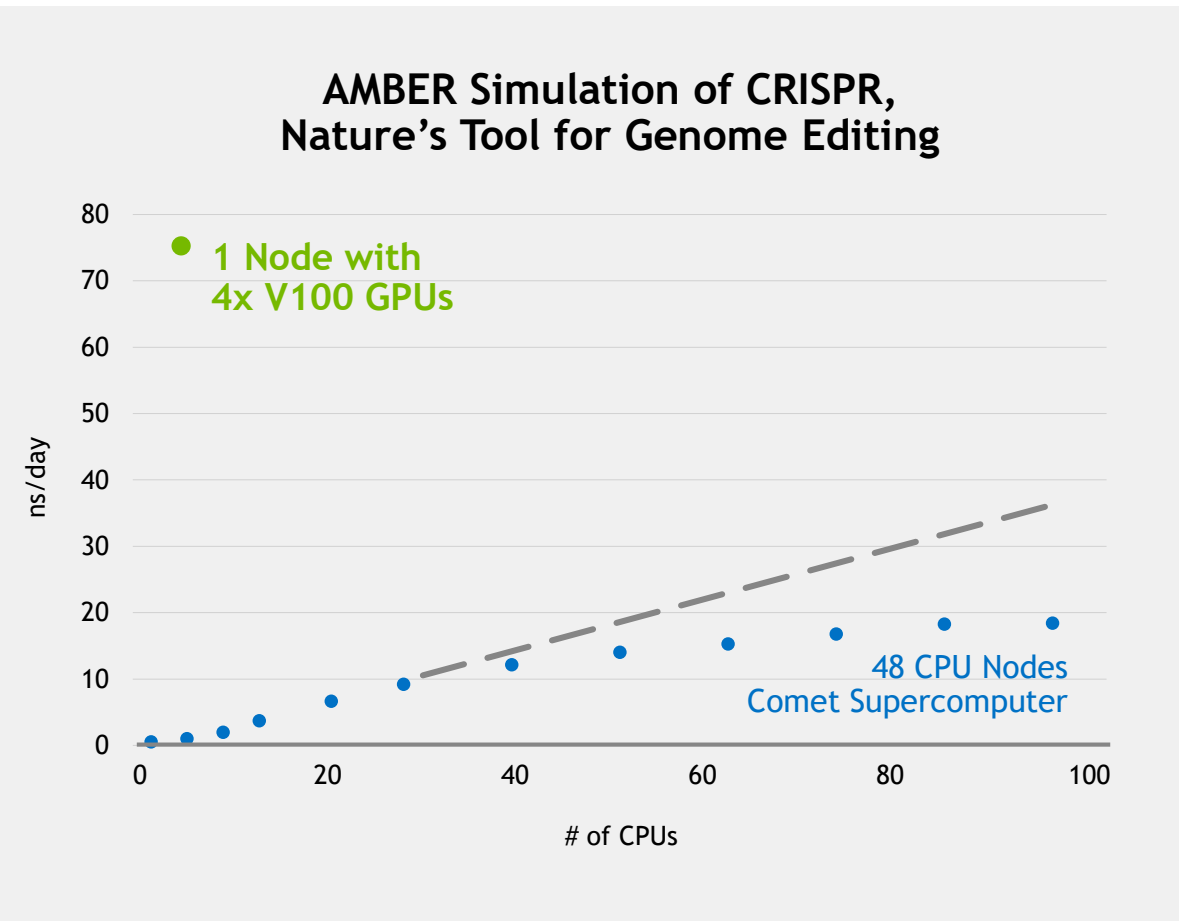
ACCELERATED FEATURES
Implicit (5x), Explicit (2x) Solvent

SCALABILITY
Multi-GPU, Single Node

BIG INEFFICIENCIES WITH CPU NODES

Single GPU Server 3.5x Faster than 50 CPU-only servers

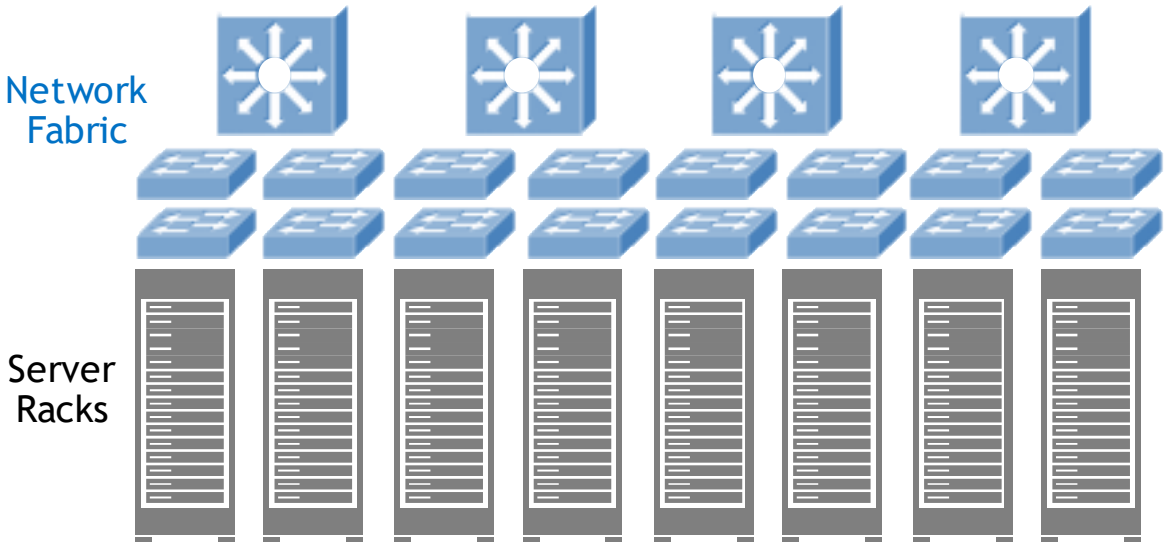




AMBER 16 Pre-release, CRSPR based on PDB ID 5f9r, 336,898 atoms
CPU: Dual Socket Intel E5-2680v3 12 cores, 128 GB DDR4 per node, FDR IB

WEAK NODES

Lots of Nodes Interconnected with
Vast Network Overhead



STRONG NODES

Few Lightning-Fast Nodes with
Performance of Hundreds of Weak Nodes


NVIDIA