ACCELERATED COMPUTING: THE PATH FORWARD

Jensen Huang, Founder & CEO | SC17 | Nov. 13, 2017
Pioneered CUDA Accelerated Computing

Extending Performance Post-Moore’s Law
Every Computer Maker

Volta in Production

Every Cloud

NVIDIA ACCELERATED COMPUTING
ANNOUNCING
NVIDIA TESLA V100 IN MICROSOFT AZURE

Microsoft Azure
NCv3 Series with NVIDIA Tesla V100
Sign up today at www.aka.ms/ncv3
ANNOUNCING JAPAN’S AIST ADOPTS NVIDIA VOLTA FOR ABCI SUPERCOMPUTER

Most Powerful AI Supercomputer in Japan
4,352 Tesla V100 GPUs
37 PetaFLOPS FP64 HPC Performance
0.55 ExaFLOPS AI Performance
109% Y-Y DATACENTER GROWTH
FYQ3

FY10  FY11  FY12  FY13  FY14  FY15  FY16  FY17  YTD 18

$0  $0.5B  $1B  $1.5B

- Full Fiscal Year
- 1st Three Quarters
### NVIDIA TESLA DATACENTER PLATFORM

<table>
<thead>
<tr>
<th>Market</th>
<th>Segment</th>
<th>Frameworks</th>
</tr>
</thead>
<tbody>
<tr>
<td>$12B</td>
<td>HPC</td>
<td>CUDA, ComputeWorks</td>
</tr>
<tr>
<td>$25B</td>
<td>CSP TRAINING</td>
<td>NVIDIA AI SDK, cuDNN, NCCL, TensorRT, DIGITS, Every Framework</td>
</tr>
<tr>
<td>$3T</td>
<td>ENTERPRISE</td>
<td>NGC, GRID vPC, Quadro vWS</td>
</tr>
</tbody>
</table>

- **80% of Apps by 2020**
- **20M Inference Servers**
- **600M Amazon Packages**
- **20M Inference Servers**
- **600M Amazon Packages**
- **3T IT Industry**

- **$12B Market**
- **$25B Market**
- **$3T IT Industry**
ARCHITECTING MODERN DATACENTERS

DEEP LEARNING COMES TO HPC

OPTIMIZED HPC SOFTWARE
ARCHITECTING MODERN DATACENTERS

- Strong Scaling
- Weak Scaling
- Deep Learning

% Workload

Apps Accelerated
Parallel Speed-Up
% Workload
% Sequential “Amdahl’s Law”

% WORKLOAD

APPS
ARCHITECTING MODERN DATACENTERS

- Strong Core CPU
- Volta 5,120 CUDA Cores
- NVLink for Strong Scaling
- 125 TFLOPS Tensor Core
ARCHITECTING MODERN DATACENTERS

AMBER Simulation of CRISPR

The graph shows the relationship between the number of CPUs and the number of simulations per day. The x-axis represents the number of CPUs ranging from 0 to 100, while the y-axis represents the number of simulations per day ranging from 0 to 80. The data points indicate that as the number of CPUs increases, the number of simulations per day also increases. The dashed line suggests a linear trend.

48 CPU Nodes
Comet Supercomputer
ARCHITECTING MODERN DATACENTERS

The Power of Accelerated Computing
AMBER Simulation of CRISPR

1 Node with 4x V100 GPUs

48 CPU Nodes
Comet Supercomputer
ARCHITECTING MODERN DATACENTERS

Apps Accelerated
Parallel Speed-Up
% Workload
% Sequential “Amdahl’s Law”

% WORKLOAD

% Accelerated

APPS

REACH

APPS
70% OF THE WORLD’S SUPERCOMPUTING WORKLOAD ACCELERATED

Top 15 HPC Applications

- VASP
- AMBER
- NAMD
- GROMACS
- Gaussian
- Simulia Abaqus
- WRF
- OpenFOAM
- ANSYS
- LS-DYNA
- BLAST
- LAMMPS
- ANSYS Fluent
- Quantum Espresso
- GAMESS

500+ Accelerated Applications
4X BETTER HPC SYSTEM TCO

Mixed Workload:
- Materials Science (VASP)
- Life Sciences (AMBER)
- Physics (MILC)
- Deep Learning (ResNet-50)

160 Self-hosted Servers
96 KWatts
Mixed Workload:
Materials Science (VASP)
Life Sciences (AMBER)
Physics (MILC)
Deep Learning (ResNet-50)

12 Accelerated Servers w/ 4 V100 GPUs
20 KWatts

1/3 the Cost
1/4 the Space
1/5 the Power

4X BETTER HPC SYSTEM TCO
NVIDIA POWERS WORLD’S FASTEST DEEP LEARNING PERFORMANCE

ResNet-50

Time to Train

Preferred Networks Feb '17
128 TitanX (Maxwell)

Facebook June '17
256 Tesla P100

IBM Aug '17
256 Tesla P100

Preferred Networks Nov '17
1024 Tesla P100

4.4 Hours

60 Mins

48 Mins

15 Mins

ResNet-50 | Dataset: Imagenet | Trained for 90 Epochs
DEEP LEARNING COMES TO HPC

NEW DATA

TRAINING SET

REGRESSION SET

NEW DATA

SIMULATION (FP64/FP32)

TRAINING (FP32/FP16)

REGRESSION TESTING (FP16/INT8)

INFERENACE (FP16/INT8)

ERRORS
DEEP LEARNING COMES TO HPC

UNSW: PHYSICS
14X Faster Bose-Einstein Condensate Creation

U. FLORIDA & UNC: DRUG DISCOVERY
300,000X Molecular Energetics Prediction

SLAC: ASTROPHYSICS
Gravitational Lensing: From Weeks to 10ms

PRINCETON & ITER: PARTICLE PHYSICS
90% Accuracy for Fusion Sustainment

U.S. DoE: CLEAN ENERGY
33% More Accurate Neutrino Detection

U. PITT: DRUG DISCOVERY
35% Higher Accuracy for Protein Scoring
NVIDIA SATURNV WITH VOLTA

40 PetaFLOPS Peak FP64 Performance | 660 PetaFLOPS DL FP16 Performance | 660 NVIDIA DGX-1 Server Nodes
NVIDIA GPU CLOUD

CLOUD CONTAINER REGISTRY

Containerized in NV Docker
Optimization Across the Full Stack
Always Up-to-Date
Fully Tested and Maintained by NVIDIA
ANNOUNCING
NVIDIA GPU CLOUD FOR HPC

CLOUD CONTAINER REGISTRY
FOR ACCELERATED HPC APPS

Containerized in NVDesk
Optimized for GPU-accelerated Systems
Up-to-Date Containers
Available NOW
Sign up at nvidia.com/gpu-cloud
HPC APPS COMING TO NGC

NOW

NEXT
VISUALIZATION IS VITAL TO SCIENCE
VISUALIZATION IS VITAL TO SCIENCE
UNIFIED VISUALIZATION FOR LARGE DATA SETS

Large-scale Volumetric Rendering
Physically Accurate Ray Tracing
Production-quality Images
Seamless integration with ParaView
Early Access NOW
Signup now at nvidia.com/gpu-cloud

ANNOUNCING NVIDIA GPU CLOUD FOR HPC VISUALIZATION

ParaView with NVIDIA IndeX
ParaView with NVIDIA OptiX
ParaView with NVIDIA Holodeck
NVIDIA GPU CLOUD
OPTIMIZED HPC SOFTWARE

DEEP LEARNING
HPC APPS
HPC VIS
Volta in Every Cloud, Every Computer Maker
Accelerated Computing’s Time Has Come

Deep Learning Comes to HPC

500+ Accelerated Apps
NVIDIA GPU Cloud for HPC

NVIDIA GPU Cloud for HPC Vis

NVIDIA ACCELERATED COMPUTING