THE NEED FOR ACCELERATED COMPUTING

REYNALDO GOMEZ
NVIDIA ENERGY PARTNER ECOSYSTEM MANAGER
REYNALDO@NVIDIA.COM
Full Stack. Data Center Scale
2,700 Accelerated Applications
450 SDKs, AI Models
30+ Million CUDA Downloads
3 Million Developers
**NVIDIA DATACENTER PLATFORM**

<table>
<thead>
<tr>
<th>BUSINESS APPLICATIONS</th>
<th>Customer Engagement</th>
<th>Seismic Interpretation</th>
<th>Fault Detection</th>
<th>Predictive Maintenance</th>
<th>Field Planning</th>
<th>Precision Marketing</th>
<th>Molecular Simulations</th>
</tr>
</thead>
<tbody>
<tr>
<td>APPLICATION FRAMEWORKS</td>
<td>SMART CITY Metropolis</td>
<td>CONVERSATIONAL AI Riva</td>
<td>AUTONOMOUS VEHICLES Drive</td>
<td>RECOMMENDATION SYSTEMS Merlin</td>
<td>HEALTHCARE Clara</td>
<td><strong>...</strong></td>
<td></td>
</tr>
<tr>
<td>DEVELOPER TOOLKITS</td>
<td>ML &amp; DATA ANALYTICS RAPIDS XGBoost</td>
<td>AI TRAINING &amp; INFERENCE TensorFlow PyTorch</td>
<td>HIGH PERFORMANCE COMPUTING NVIDIA HPC SDK</td>
<td>RENDERING &amp; VISUALIZATION IndeX OptiX CloudXR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACCELERATION LIBRARIES</td>
<td>COMPUTE CUDA-X</td>
<td>NETWORKING, STORAGE &amp; SECURITY DOCA MAGNUM IO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NVIDIA CERTIFIED</td>
<td>VALIDATED SOLUTIONS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SERVERS &amp; CLOUD</td>
<td>DGX HGX Purpose Built</td>
<td>EGX Mainstream &amp; Edge</td>
<td>CSP Instances</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HARDWARE TECHNOLOGIES</td>
<td>GPU NVSwitch BF DPU SMART NIC NVIDIA Switch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
NVIDIA IS A FULL STACK COMPUTING PLATFORM

Amazing Innovation and Expansion of NVIDIA Ecosystem

150 SDKs

RTX  HPC  RAPIDS  AI  CLARA  METRO  DRIVE  ISAAC  AERIAL 5G

CUDA-X-AI

CUDA

MAGNUM IO

ECOSYSTEM

APPLICATIONS

SDK & ENGINES

SYSTEMS

CHIPS

COMPLETE SOFTWARE STACK

FULL STACK INNOVATION

GROWING ECOSYSTEM

1B CUDA GPUs

3M Developers

30M CUDA Downloads

2,500 GPU-Accelerated Applications

9,000 AI Startups

65 updates from last GTC
NEXT WAVE OF AI REQUIRES PERFORMANCE AND SCALABILITY

70% AI Papers In last 2 years discuss Transformer Models

EXPLODING COMPUTATIONAL REQUIREMENTS

Transformer AI Models = 275x / 2yrs
AI Models Excluding Transformers = 8x / 2yrs

HIGHER PERFORMANCE AND SCALABILITY

GPT-3 (175B parameters) 3.5 months to train on 128x A100
MINOR CODE CHANGES FOR MAJOR BENEFITS

Abstracting Accelerated Compute through Familiar Interfaces

**pandas**

```python
In [1]: import pandas as pd
In [2]: df = pd.read_csv('filepath')
```

**CPU Spark**

```python
spark.sql(""
   select
   order
   count(*) as order_count
   from
   orders"")
```

**scikit-learn**

```python
In [1]: from sklearn.ensemble import RandomForestClassifier
In [2]: clf = RandomForestClassifier()
In [3]: clf.fit(x, y)
```

**NetworkX**

```python
In [1]: import networkx as nx
In [2]: page_rank=nx.pagerank(graph)
```

**cuDF**

```python
In [1]: import cudf
In [2]: df = cudf.read_csv('filepath')
```

**GPU Spark**

```python
spark.conf.set("spark.rapid.s.enabled","true")
spark.sql(""
   select
   order
   count(*) as order_count
   from
   orders"")
```

**cuML**

```python
In [1]: from cuml.ensemble import RandomForestClassifier
In [2]: cuclf = RandomForestClassifier()
In [3]: cuclf.fit(x, y)
```

**cuGraph**

```python
In [1]: import cugraph
In [2]: page_rank=cugraph.pagerank(graph)
```

Average Speed-Ups: 150x

Average Speed-Ups: 10x

Average Speed-Ups: 50x

Average Speed-Ups: 250x
NVIDIA END-TO-END AI SOFTWARE STACK
Deep Learning Streamlined From Conception to Production at Scale

PRE-TRAINED MODELS
- Automatic Speech Recognition
- Image Classification
- Object Detection
- Language Modelling
- Speech Synthesis
- Recommender Systems
- Machine Translation
- Image Segmentation

DATA PREP

TRAIN AT SCALE
- PyTorch
- TensorFlow

OPTIMIZED FOR INFERENCE

DEPLOY AT SCALE

RAPIDS
- Accelerated with NVIDIA

TITRON INFEERENCE SERVER
- NVIDIA
NVIDIA’S HPC PLATFORM

CUDA-X (60+ libraries)
HPC SDK (Fortran, C, C++, Python Dev tools)
Holoscan
cuQuantum
IndeX
Magnum IO
Modulus
NVIDIA AI
Omniverse

NVIDIA HPC
NVIDIA HPC SDK
Available at developer.nvidia.com/hpc-sdk, on NGC, via Spack, and in the Cloud

Develop for the NVIDIA Platform: GPU, CPU and Interconnect
Libraries | Accelerated C++ and Fortran | Directives | CUDA
7-8 Releases Per Year | Freely Available
WORLD’S LEADING ACCELERATED COMPUTING PLATFORM FOR HPC

FULL STACK, DATA CENTER SCALE

- 3 Chips - CPU, GPU, DPU
- 30+ Million CUDA Downloads
- 3+ Million Developers
- 2,700 Applications
- 450 SDKs and AI Models

26X PERFORMANCE IN 6 YEARS

NVIDIA Platform Performance on Top Apps
Moore’s Law

POWERING WORLDS FASTEST SUPERCOMPUTERS

Top500 Platform of Choice
MLPerf HPC & Training
AI Transforming HPC

71%
Use NVIDIA

Accelerated Network
Driving Modern Workloads

#1
Most adopted in Top500

Cloud HPC
Democratizing Access

100%
8 instances In Top 50

Center Panel: geometric mean of application speedups vs. P100 | benchmark applications | Amber [PME-Cellulose_NVE], Chroma [HMC], GROMACS [ADH Dodec], MILC [Apex Medium], NAMD [stmv_nve_cuda], PyTorch (BERT Large Fine Tuner), Quantum Espresso [AUSURF112-jR], Tensorflow [ResNet-50], VASP 6 [Si Huge], GPU node: with dual-socket CPUs with 4x P100, V100, or A100 GPUs. H100 values shown for 2022 projected performance subject to change

The MLPerf name and logo are trademarks of MLCommons Association in the United States and other countries. All rights reserved. Unauthorized use strictly prohibited. See www.mlcommons.org for more information.
**2700+ GPU-ACCELERATED APPLICATIONS**

Transforming Every Industry

<table>
<thead>
<tr>
<th>ARTIFICIAL INTELLIGENCE</th>
<th>CLIMATE &amp; WEATHER</th>
<th>COMPUTATIONAL FINANCE</th>
<th>DATA SCIENCE &amp; ANALYTICS</th>
<th>FEDERAL DEFENSE &amp; OTHER</th>
<th>LIFE SCIENCES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caffe2</td>
<td>Cosmos</td>
<td>O-Quant Options Pricing</td>
<td>Anaconda</td>
<td>ArcGIS Pro</td>
<td>Amber</td>
</tr>
<tr>
<td>MXNet</td>
<td>Gales</td>
<td>MUREX</td>
<td>H2O</td>
<td>EVNI</td>
<td>LAMMPS</td>
</tr>
<tr>
<td>TensorFlow</td>
<td>WRF</td>
<td>MISYS</td>
<td>OmniSci</td>
<td>SocetGXP</td>
<td>GROMACS</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MANUFACTURING, CAD, &amp; CAE</th>
<th>MEDIA &amp; ENTERTAINMENT</th>
<th>MEDICAL IMAGING</th>
<th>OIL &amp; GAS</th>
<th>RETAIL</th>
<th>SUPERCOMPUTING &amp; HER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ansys Fluent</td>
<td>DaVinci Resolve</td>
<td>aidoc</td>
<td>Echelon</td>
<td>Everseen</td>
<td>Chroma</td>
</tr>
<tr>
<td>Abaqus SIMULIA</td>
<td>Premiere Pro CC</td>
<td>PowerGrid</td>
<td>RTM</td>
<td>Deep North</td>
<td>GTC</td>
</tr>
<tr>
<td>AutoCAD</td>
<td>Redshift Renderer</td>
<td>RadiAnt</td>
<td>SPECFEM3D</td>
<td>Third Eye Labs</td>
<td>MILC</td>
</tr>
<tr>
<td>CST Studio Suite</td>
<td></td>
<td></td>
<td></td>
<td>AWM</td>
<td>QUDA</td>
</tr>
</tbody>
</table>

NVIDIA MODULUS
Physics Machine Learning Platform

**TRAINING NEURAL NETWORKS USING BOTH DATA AND THE GOVERNING EQUATIONS**

\[
\frac{\partial p}{\partial t} + \nabla \cdot (p \mathbf{u}) = 0
\]

\[
p \frac{\partial \mathbf{u}}{\partial t} + \nabla \cdot (\mathbf{u} \otimes \mathbf{u}) = -\nabla p + \nabla \times (\mathbf{v} \times \mathbf{B}) + \mathbf{f}
\]

**ADVANCING SCIENTIFIC DISCOVERY WITH MODULUS**

**RENEWABLE ENERGY**
Siemens Gamesa: Up to 4000X Speedup of Wind Turbine Wake Optimization

**CLIMATE CHANGE**
45,000X Speedup of Extreme weather Prediction with FourCastNet

**INDUSTRIAL HPC**
NETL: 10,000X Faster Build Of high-fidelity surrogate models

**HEALTHCARE**
Achieve high-fidelity results faster for blood flow in inter-cranial aneurysm

**DIGITAL TWINS**
Kinetic Vision: Design Optimization Using parameterized models

Get started today with NVIDIA Modulus
RISE OF HPC AT THE EDGE
Posing a New Set of Challenges for HPC

10X - 100X MORE DATA
50+ GIANT SCALE Instruments WW

AI SUPERCOMPUTING AT THE EDGE
ENABLES REAL-TIME INSIGHTS AND CONTROL

STREAMING DEPLOYMENT IS HARD
FOR DATA SCIENTISTS, RESEARCHERS AND DEVOPS

ELT ESO
LIGO
SKA

ALS @LBNL
APS @ANL
Diamond, UK

Real-Time Streaming AI Processing
100x Data Collected
1x Data Transfer
ML/AI Model
Edge
Data Center

Streaming Data Performance
Developer Ease-of Use
Easily Scale Implementation
Combining multiple data streams
NVIDIA A100 80GB
Supercharging The World’s Highest Performing AI Supercomputing GPU

80GB HBM2e
For largest datasets and models

2TB/s +
World’s highest memory bandwidth to feed the world’s fastest GPU

3rd Gen Tensor Core

Multi-Instance GPU

3rd Gen NVLink
NVIDIA A100 GPU
Greatest Generational Leap - 20X Volta

<table>
<thead>
<tr>
<th>Feature</th>
<th>Peak</th>
<th>Vs Volta</th>
</tr>
</thead>
<tbody>
<tr>
<td>FP32 TRAINING</td>
<td>312 TFLOPS TF32</td>
<td>20X</td>
</tr>
<tr>
<td>INT8 INFEERENCE</td>
<td>1,248 TOPS</td>
<td>20X</td>
</tr>
<tr>
<td>FP64 HPC</td>
<td>19.5 TFLOPS</td>
<td>2.5X</td>
</tr>
<tr>
<td>MULTI INSTANCE GPU</td>
<td></td>
<td>7X GPUs</td>
</tr>
</tbody>
</table>
DGX A100 640GB SYSTEM

For the Largest AI Workloads

640 GB of GPU memory per system to increase model accuracy and reduce-time-to-solution

Up to 3X higher throughput for large-scale workloads

Double the GPU memory for MIG for more flexible AI development, analytics, and inference

Available individually, or part of DGX SuperPOD Solution for Enterprise
SERVER-CLASS SOLUTION IN A WORKSTATION PACKAGE

Data Center Technology Outside the Data Center

First and only workstation with 4-way NVIDIA HGX A100

Four A100 Tensor Core GPUs, up to 320GB total HBM2E
3rd generation NVLink
200GB/s bi-directional bandwidth between any GPU pair, almost 3x compared to PCIe Gen4

New Cooling System, Pump Refrigerant 2-Phase Cooling

Maintenance-free, sealed system
No need to check, or refill, water-level
Single loop for CPU and four GPUs
Non-toxic, non-flammable, non-condensing
<table>
<thead>
<tr>
<th></th>
<th>DGX Station A100 320GB</th>
<th>DGX Station A100 160GB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GPUs</strong></td>
<td>4x NVIDIA A100 Tensor Core GPUs</td>
<td></td>
</tr>
<tr>
<td><strong>GPU Memory (total)</strong></td>
<td>320GB</td>
<td>160GB</td>
</tr>
<tr>
<td><strong>Performance</strong></td>
<td>2.5 petaFLOPS AI; 5 petaOPS INT8</td>
<td></td>
</tr>
<tr>
<td><strong>System Memory</strong></td>
<td>512GB DDR4 RDIMM, 3200MT/s</td>
<td></td>
</tr>
<tr>
<td><strong>Storage</strong></td>
<td>OS: 1 x 1.92TB M.2 NVME</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Data: 1 x 7.68TB U.2 NVME</td>
<td></td>
</tr>
<tr>
<td><strong>CPU</strong></td>
<td>AMD® Epyc® CPU 7742, 2.25GHz to 3.4GHz, 64 cores/128 threads, PCIe Gen4</td>
<td></td>
</tr>
<tr>
<td><strong>Networking</strong></td>
<td>Dual 10GBASE-T (RJ45)</td>
<td></td>
</tr>
<tr>
<td><strong>Display GPU</strong></td>
<td>4GB, 4x Mini DisplayPort</td>
<td></td>
</tr>
<tr>
<td><strong>Acoustics</strong></td>
<td>&lt;37dB</td>
<td></td>
</tr>
<tr>
<td><strong>Cooling</strong></td>
<td>Custom refrigerant cooling system for GPUs and CPU</td>
<td></td>
</tr>
<tr>
<td><strong>System Power (max)</strong></td>
<td>1,5kW</td>
<td></td>
</tr>
<tr>
<td><strong>Management</strong></td>
<td>AST2500, IPMI, Redfish</td>
<td></td>
</tr>
<tr>
<td><strong>System Dimensions</strong></td>
<td>518 D x 256 W x 639 H (mm)</td>
<td></td>
</tr>
<tr>
<td><strong>Operating Temp.</strong></td>
<td>5°C to 35°C (41°F to 95°F)</td>
<td></td>
</tr>
</tbody>
</table>
ADOPTED BY LEADING COMPANIES ACROSS INDUSTRIES
DGX Station Delivers AI Supercomputing to More Teams, From Anywhere

6 Of the Top 10 US Government Institutions
6 Of the Top 10 Global Car Manufacturers
7 Of the Top 10 US Hospitals
10 Of the Top 10 Aerospace & Defense Companies
NVIDIA A100 80GB PCIE
Supercharging The World’s Highest Performing AI Supercomputing GPU

Flexible Deployment Option for Mainstream OEM Servers
Excellent Upgrade Path for V100 32GB PCIE Customers

A100 80GB PCIE
Throughput vs V100
32GB PCIE

3X
Simulation
Chroma

5X
AI Training
DLRM Recommender

5X
BERT Large Inference
Natural Language Processing

7X
NVLINK Bridge
Fast Fourier Transforms
NVIDIA LIQUID-COOLLED GPUS FOR MAINSTREAM SERVERS

- **HGX A100**
  - Shipping: Q3 2022

- **A100 PCIe**
  - Q4 2022

- **HGX H100**
  - Q3 2022

- **H100 PCIe**
  - Early 2023

- **30%**: Lower power consumption
- **66%**: Less rack space
- **<1.2 PUE**: Meet efficiency targets

*Energy & Space Efficiency of A100 PCIe Liquid Cooled vs Air Cooled*
GRACE CPU SUPERCHIP
The CPU for AI and HPC Infrastructure

HIGHEST CPU PERFORMANCE
Superchip Design with 144 high-performance Armv9 Cores
Estimated Specrate2017_int_base of over 740

HIGHEST MEMORY BANDWIDTH
World’s first LPDDR5x memory with ECC, 1TB/s Memory Bandwidth

HIGHEST ENERGY EFFICIENCY
2X Perf/Watt, CPU Cores + Memory in 500W

2X PACKING DENSITY
2x density of DIMM based designs

RUNS FULL NVIDIA COMPUTING STACKS
RTX, HPC, AI, Omniverse

AVAILABLE 1H 2023
HIGHEST ACCELERATED PERFORMANCE
Grace CPU plus Hopper GPU Acceleration

~600GB MEMORY AVAILABLE TO GPU
Enables Giant AI Models for Training & Inference

HIGHEST MEMORY BANDWIDTH 3.5GB/s
LPDDR5x and HBM3

NEW 900GB/S COHERENT INTERFACE
NVLink-C2C connecting Grace to Hopper

15X HIGHER SYSTEM MEMORY BANDWIDTH TO GPU
NVLink-C2C vs PCIe

RUNS FULL NVIDIA COMPUTING STACKS
RTX, HPC, AI, Omniverse

AVAILABLE 1H 2023