Accelerated Computing Infrastructure with NVIDIA

Dr. Pallavi Mohan
Senior Scientist & Solutions Architect, NVIDIA
Computer Industry Fundamental Transitions

ACCELERATED COMPUTING

GENERATIVE AI
Modern Data Centers are Becoming AI Factories

Producing Intelligence from Data
Key Data Center Trends

Demand for compute grows as data centers become power limited

Massive AI Models Drive New Use Cases
- LLMs and GenAI Driving an Inflection Point

Data Centers are Power Limited
- Need to Become More Efficient

- >200 TWh/year
- 2%
- 5% by 2030

Data Center Electricity Usage
Share of Global Energy Usage
Forecasted Share of Energy Usage
Accelerated Computing is the Path Forward
NVIDIA Accelerated Computing for Modern Data Centers

End of Moore’s Law

Acceleration Takes a Full Stack

APPLICATION

ACCELERATION LIBRARY

SYSTEM

CPU

DPU

GPU

1.5X per year

10^3

10^4

10^5

10^6

10^7

10^8

10^9

10^10

ACCELERATED COMPUTING

SCALE UP & OUT

Al

AI

1.1X per year

1.5X per year

CPU

Lower is better

4K A100

1 month to 1 week

Days to hours

4K H100

Training Time (Weeks)

Time-to-Train by LLM Size
(Billion parameters)

H100 Supercharges AI

4K H100

1000

530

175

70

Days

to hours
NVIDIA DGX Platform

Cloud

DGX Cloud
Multi-node AI training software as a service solution.

Software

Base Command Platform
Centralized control of AI training projects across the DGX platform.

Clusters and Systems

DGX SuperPOD
Leadership-class AI infrastructure for on-premises and hybrid deployments.

DGX BasePOD
Proven reference architectures for AI infrastructure delivered with leading storage providers.

DGX H100
AI supercomputer optimized for large generative AI and other transformer-based workloads.

DGX A100
AI supercomputer delivering world-class performance for mainstream AI workloads.
The DGX Platform Powers Your AI Journey From End-to-End

Delivering incremental value for your AI initiatives, as your needs grow

<table>
<thead>
<tr>
<th>Day One</th>
<th>Scaled Infrastructure</th>
<th>AI Center of Excellence</th>
</tr>
</thead>
<tbody>
<tr>
<td>your 1st DGX systems</td>
<td>DGX BasePOD</td>
<td>DGX SuperPOD</td>
</tr>
</tbody>
</table>

DGX Systems powers every step in your AI journey

Integrated software that powers AI innovation

- NVIDIA AI Enterprise:
  - Pre-trained models, optimized frameworks
  - Customize/fine-tune pre-trained models
  - Optimize/accelerate inference
  - Kubernetes or Slurm scheduling
  - Add/manage DGX within your existing compute infrastructure (cloud, non-GPU)
  - Accelerate storage & network IO
  - Fully optimized OS stack

NVIDIA Base Command

- NVIDIA Base Command Premium
  - In addition to features on the left:
    - NVIDIA Base Command Platform:
      - Simplify developer workflow
      - Dataset management
      - Batch processing
      - Monitoring

Continuous roadmap of innovative features delivered to customers

DGX Cloud - The Cloud-First Way to Get Access to Your Own DGX AI Supercomputer
NVIDIA DGX GH200: The Trillion Parameter Instrument of AI

Massive memory supercomputing for emerging AI

World’s first system built with NVIDIA NVLink Switch System

- Nearly **500X** more system memory
- **48X** GPU-to-GPU bandwidth
- **7X** CPU-to-GPU bandwidth
- **5X** interconnect power efficiency

256 Grace Hopper Superchips | **1 EFLOPS** AI Performance | **144 TB** unified fast memory
36 L2 NVLink switches | **900 GB/s** GPU-to-GPU bandwidth | **128 TB/s** bisection bandwidth

Available year-end 2023
NVIDIA BlueField DPU Platform
Software-Defined, Hardware-Accelerated Infrastructure Compute Platform

Accelerated Performance
Meet the most stringent performance requirements, run the most demanding workloads

Cloud-Scale Efficiency
Free up x86 cores to business apps, achieve unprecedented scale and efficiency levels

Robust Zero-Trust Security
Ensure comprehensive data center security without compromising performance

Programmable Infrastructure
Develop and run applications consistently with maximum performance

DPU ACCELERATED SERVER

NVIDIA BlueField DPU with Arm Cores & Accelerators
Infrastructure Management
Software-defined Security
Software-defined Storage
Software-defined Networking
Acceleration Engines
Offload | Accelerate | Isolate
**NVIDIA BlueField-3 Overview**

Massive Advancements, Built for Cloud Scale

- **400Gb Networking**
  - 2X Network Bandwidth
  - 2X Network Pipeline
  - 4X Host Bandwidth

- **Programmable Engines**
  - 4X Arm Compute
  - 5X Memory
  - New Datapath Accelerator

- **Zero-Trust Security**
  - 4X IPsec Acceleration
  - 2X TLS Acceleration
  - New MACsec Acceleration

- **Composable Storage**
  - 2X Storage IOPs
  - 2X Storage Encryption
  - New NVMe/TCP Acceleration

*Compared to previous BlueField generation*
NVIDIA DOCA
Comprehensive Acceleration SDK for BlueField DPUs

- Unified software framework for BlueField DPUs
- Offload, accelerate, and isolate infrastructure processing
- Support for hyperscale, enterprise, supercomputing and hyperconverged infrastructure
- Software compatibility for generations of BlueField DPUs
- Rich partner ecosystem
BlueField Powers NVIDIA-Accelerated Computing Systems

Full-Stack, Data Center-Scale, Multi-Domain Acceleration

Generative AI | Scientific Computing | 5G Networks | Distributed Database | Internet Services

DOCA

BlueField-3 DPU

CUDA

H100 GPU
Accelerating Scientific Computing Workloads
Ignite High-Performance Computing with NVIDIA BlueField and Quantum InfiniBand

- Unleash application performance and system efficiency
- MPI performance acceleration
- Computational storage and advanced workloads
- Adaptive performance isolation

Application Performance Improvement

- Mathematical Modeling: 26%
- Molecular Dynamics: 20%
- Weather and Climate: 20%
Accelerated Computing is Sustainable Computing

BlueField-3 Enables Power-Efficient Cloud Data Centers
fin