Networks for AI
Gilad Shainer — SVP, Networking
FORWARD-LOOKING STATEMENTS

Except for the historical information contained herein, certain matters in this presentation including, but not limited to, statements as to: our markets and market opportunity; our growth and growth drivers; and the benefits, impact, and performance of our products and technologies are forward-looking statements. These forward-looking statements and any other forward-looking statements that go beyond historical facts that are made in this presentation are subject to risks and uncertainties that may cause actual results to differ materially. Important factors that could cause actual results to differ materially include: global economic conditions; our reliance on third parties to manufacture, assemble, package and test our products; the impact of technological development and competition; development of new products and technologies or enhancements to our existing product and technologies; market acceptance of our products or our partners’ products; design, manufacturing or software defects; changes in consumer preferences and demands; changes in industry standards and interfaces; unexpected loss of performance of our products or technologies when integrated into systems and other factors.

NVIDIA has based these forward-looking statements largely on its current expectations and projections about future events and trends that it believes may affect its financial condition, results of operations, business strategy, short-term and long-term business operations and objectives, and financial needs. These forward-looking statements are subject to a number of risks and uncertainties, and you should not rely upon the forward-looking statements as predictions of future events. The future events and trends discussed in this presentation may not occur and actual results could differ materially and adversely from those anticipated or implied in the forward-looking statements. Although NVIDIA believes that the expectations reflected in the forward-looking statements are reasonable, the company cannot guarantee that future results, levels of activity, performance, achievements or events and circumstances reflected in the forward-looking statements will occur. Except as required by law, NVIDIA disclaims any obligation to update these forward-looking statements to reflect future events or circumstances. For a complete discussion of factors that could materially affect our financial results and operations, please refer to the reports we file from time to time with the SEC, including our most recent Annual Report on Form 10-K, Quarterly Reports on Form 10-Q, and Current Reports on Form 8-K. Copies of reports we file with the SEC are posted on our website and are available from NVIDIA without charge.
The Data Center is The Computer

The network defines the data center

- **Cloud**
  - Multi-tenant
  - Variety of small-scale workloads
  - Traditional ethernet network can suffice

- **Generative AI Cloud**
  - Multi-tenant
  - Variety of workloads including larger scale Generative AI
  - Traditional ethernet network for North-South traffic
  - NVIDIA Spectrum-X ethernet for AI fabric (East-West)

- **AI Factories**
  - Single or few users
  - Extremely large AI models
  - NVIDIA NVLink and InfiniBand gold standard for AI fabric
The NVIDIA AI Network Advantage
Hardware & software accelerated in-network computing

Software Acceleration
NCCL — NVIDIA Collective Communication Library
The SDK library for AI communications - connects the GPUs and the network for the AI network operations.

Hardware Acceleration
SHARP — NVIDIA Scalable Hierarchical Aggregation and Reduction Protocol Technology
SHARP is part of the InfiniBand and NVLink switch ASICs. It enables the network to perform data reduction operations, an important element of AI workloads. This decreases the amount of data traversing the network and dramatically reduces collective operations time.

NCCL Performance With vs Without SHARP (In-network Computing)

1.7X HIGHER
InfiniBand (with SHARP)

Best Theoretical Ethernet performance

Message Size (MiB)

8 16 32 64 128 256 512 1024 2048 4096 8192 16384
In the Cloud — The Two Worlds of Ethernet

East-West for distributed and disaggregated processing | North-South for user-to-cloud communications

Control / User Access Network (North-South)
Traditional Ethernet

AI Fabric (East-West)
Spectrum-X

Loosely Coupled Applications
Distributed Tightly-Coupled Processing

TCP (Low Bandwidth Flows and Utilization)
RoCE (High Bandwidth Flows and Utilization)

High Jitter Tolerance
Low Jitter Tolerance
(Long Tail Kills Performance)

Heterogeneous Traffic
Average Multi-Pathing
Bursty Network Capacity
Predictable Performance

Data Center

East – West
North – South
NVIDIA Spectrum-X Ethernet AI Platform
The ethernet fabric for AI

**Spectrum-4 Switch**
- 100 billion transistors, TSMC 4N
- 51.2T bandwidth, 100G SerDes
- 64 x 800G Ports, 128 x 400G ports
- 8K GPUs in 2-level, 500K in 3 level

**Architectural Advantages over Traditional Ethernet...**
- End-to-end optimized (DPU to Switch)
- Adaptive routing over lossless ethernet
- Congestion control for multi-tenant traffic isolation

**...Deliver 1.6X Better AI Fabric Performance vs Traditional Ethernet**
- 95% effective bandwidth vs 60% for traditional ethernet
- Maintains NCCL performance in noisy environment
- Secured virtualized network

**New class of ethernet, built from the ground-up for Generative AI cloud workloads**
NVIDIA Quantum InfiniBand Platform
The gold standard for large scale AI

Quantum-2 Switch
57 billion transistors, TSMC 7N
64 X 400G ports, 100G SerDes
64 In-Network Computing AI Engines
65K GPUs in 3-level, 2M in 4-level

InfiniBand’s Inherent Architectural Differentiators...
• Architected for large scale, end-to-end optimized
• In-network computing with SHARP
• Pure software-defined network

...Deliver >2X AI Fabric Performance over Traditional Ethernet
• Limitless GPU scaling (3-level 65K GPUs, 4-level 2M GPUs)
• 1.7X higher NCCL AllReduce bandwidth performance
• Nearly 100% effective bandwidth at scale
• Extremely low latency, 10x better at scale and under load

Hardened over 20+ years for the most extreme-scale workloads in the world’s largest supercomputers
The Network “Pays for Itself”!
Performance gains far outweigh the entire cost of the network

InfiniBand offers the most scalability...

NDR 400G InfiniBand (64-port switches)
3 switch level network topology (up to 65K nodes)
4 switch level network topology (up to 2M nodes)

... and the fastest performance

Relative AI Fabric Performance (NCCL Allreduce)

<table>
<thead>
<tr>
<th></th>
<th>Traditional Ethernet</th>
<th>NVIDIA Spectrum-X Ethernet</th>
<th>NVIDIA InfiniBand</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1.6X</td>
<td>&gt;2X</td>
</tr>
</tbody>
</table>
Strong Growth and Large Opportunity

Data Center Networking Growth Turbocharged by AI

Revenue more than doubled since Mellanox acquisition
NVIDIA InfiniBand revenue more than tripled
Spectrum-X to boost cloud AI ethernet market

NVIDIA Networking Revenue
*Q1 CY23 refers to Q1 of NVIDIA’s FY24, ending April 30, 2023.
The Mellanox acquisition closed on April 27, 2020.

Data Processing Units — In every server
Data Center Switching — InfiniBand & ethernet
High speed optical modules & cables

$60 BILLION

Networking Market Opportunity