Mellanox: Optimizing HPC
Enabling the Machine Learning Network
March 2017
The Next-Generation of HPC: Machine Learning, Using Big Data, and the Pending Data Explosion
What Makes Big Data?

VOLUME
Data at Rest
TB to PB of data to process

VELOCITY
Data in Motion
Streaming Data – ms to s to respond

VARIETY
Data in many forms
Semi-Structured, unstructured,

VALUE
Data in $$
Unknown insight from the known data

VERACITY
Data in doubt
Uncertainty due to inconsistency, inaccuracy
Facebook: Pushing the Limits of Data

Cross Platform Chat

Social Network

Instant Messaging

The World’s Most Lucrative Social Network
The Data Divide – Acquisition vs. Analysis

IDC says currently 22% of data useful. By 2020, that number will climb to 37%.

Need intelligent tools to learn from data: supervised, unsupervised, deep, etc.
Understanding Big-Data Learning

Mellanox Efficient Network Solves Two Orthogonal Problems

Predictive Analytics
- Learning insight from data
- Faster Analytics needs RDMA
- Infiniband/Ethernet predominantly used

Big Data
- Handling massive amount of data
- Faster storage needs faster network
- Ethernet predominantly used

Collect
- OLTP, ERP, CRM Systems
- Documents & Emails
- Web, Logs, Click Streams
- Social Networks
- Machine Generated
- Sensor Data
- Geo-Location Data

Predict
- Google
- Caffe
- TensorFlow

Analyze
- Spark
- Hadoop
- Neo4j

Machine Learning/Deep Learning/Neural Networks
- Google
- Caffe
- TensorFlow
- Spark
- MLlib
- Theano
- CNTK
Major Trends Driving Next-Gen Analytics Infrastructure

- Three fundamental changes converging all at once
  1. Ethernet technology transition from 10G -> 25G, 40G -> 50G & 100G
  2. Storage solutions transition from HDD -> SSDs/ NVMe, Storage Arrays -> Scale-Out Storage
  3. Advanced Analytics and Big Data Platforms embracing RDMA (RoCE / Infiniband)

Mellanox is at the Heart of this Transformation
Faster Big-Data Storage is Driving Larger and Faster Scale

- Just three NVMe Flash can saturate 100 Gb/s Link
  - Needs 100GbE ConnectX-4 & RoCE
- Advanced Network Offloads
  - Burn Rubber! Not CPU Cycles.
  - Offload IO processing to network & more analytics on servers

Three NVMe Flash Drives = One 100GbE Links

Fast Storage Needs RDMA

- ½ Bandwidth 10X CPU load
- Full Bandwidth 1/10 CPU load

Without RoCE

With RoCE
Smart Caching at Scale Example: Spark – In-Memory Data Processing Engine

- **Apache Spark**
  - In-Memory, Cluster Computing Framework
  - Originally Developed at AMPLab, Berkley

- **Fits into Hadoop Framework**
  - Replaces MapReduce
  - Builds on top of HDFS

---

>1000 deployments
1000s nodes in large deployments
Deep Learning is the Next-Gen HPC Application

TRAINING

• Billions of TFLOPS per training run
• Years of compute-days on Xeon
• GPUDirect turns years to days

INFERENCING

• Billions of FLOPS per inference
• Seconds for response on Xeon
• GPUDirect turns to instant response
Machine / Deep Learning: Software that Writes Software

* Stanford example of interpreting photos

Little girl is eating piece of cake

Computational Workload Replaces Hand Written Code
Deep Neural Network

Layer 1

Layer 2

Layer 3

Layer 4

output

pixels → edges → object parts (combination of edges) → object models
New, Non-traditional HPC Consumers

**Retail**: SMB, Big Box & Dept. Store Chains:

*Learning / Data inputs >>*

- Buyer Demographics & Characterizations
- Trends & Buying Practices

*Inference outputs <<*

- Data required for targeted promotional campaigns
- Optimized logistics, inventory, and backlog management

**Healthcare Delivery** *(non-research)*

*Learning / Data inputs >>*

- Patient symptom data, Genome data, patient history
- CDC and WHO data
- Demographic history – mortality and morbidity

*Inference outputs <<*

- Diagnosis, diagnostic recommendations
- Proposed courses of treatment (covering multiple spectrums)
IOT: Driving the Data Explosion

Service Providers (Sprint, AT&T, Verizon, Spectrum, Amazon, Google, GM [On-Star])

Learning / Data inputs >>
- Buyer Demographics and Behaviors
- Product utilization (GIS, Usage history, Patterns, Content consumed)

Inference outputs <<
- Targeted ad criteria, Location-based content, Pattern and actuarial categorization

Automotive & Transportation Industry (Trucks, Airlines & Ships)

Learning / Data inputs >>
- GIS and driver categorization
- Maintenance, product behavior, and traffic / travel patterns
- Weather and climactic data

Inference outputs <<
- Location-based content, Traffic analysis, and Route optimizations
- Logistics, flow, and content management

Smart Devices / Appliances (Samsung, LG, Sony, Nest, Whirlpool)

Learning / Data inputs >>
- GIS and climactic data
- Maintenance, product behavior, and customer usage patterns
- Usage pattern data

Inference outputs <<
- Optimized device / appliance behavior tailoring
- Maintenance, break-fix, and parts ordering
Helping Accelerate Big-Data Learning: The Compute – Enabled Network

In-Network Computing Key for Highest Return on Investment
Example: Accelerating TensorFlow with gRPC over RDMA

- **Open Source Machine Learning from Google**
  - **Hottest ML Software**
  - Powers 40 Different Google Products

- **Distributed Training With gRPC Framework**
  - Google’s Optimized RPC for Distributed Network

- **RDMA Acceleration over UCX**
  - Unified Communication X (UCX)
    - Initiative by ORNL, Mellanox, IBM, NVidia, etc.
  - Integration With Upstream gRPC/TensorFlow

- **>2X Performance Gain With RDMA**
Example: Accelerating PaddlePaddle with RDMA

- ML Software from Baidu
  - Use Case: Self Driving Car, Word Prediction, Image Processing
  - **Hottest ML in China**

- Demands Faster Network: Ethernet/Infiniband
  - Downside: CPU/GPU Bottleneck; Slower Training

- RDMA (GPUDirect) Speeds Training
  - Lowers Latency, Increases Throughput
  - More Cores for Training
  - Much Better Results With Optimized RDMA

- Current Status:
  - Open Sourced with TCP
  - RDMA Release Soon

![Graph showing training time comparison between TCP, RDMA, and Optimized RDMA. Lower is better.](image)

- ~2x Faster on Paddle Training With Mellanox RoCE

---

**Press Release:** Mellanox Ethernet Accelerates Baidu’s Machine Learning Platforms

Testing by Baidu on ConnectX-4 40GbE and Spectrum SN2700
Thank You
Questions?

- **Mellanox Contacts:**
  - Glen Sheets, Sr. Director, Dell OEM Sales  
    glens@mellanox.com
  - Pat Kelly, Sr. Director, EMC OEM Sales  
    pat@mellanox.com
  - Denise Cowart, Sr. Sales Account Mgr., Dell Team  
    denisec@mellanox.com
  - Will Stepanov, Sales Manager, Dell Team  
    will@mellanox.com
  - Tina Davis, Sr. OEM Marketing Mgr., Dell Team  
    tina@mellanox.com
  - Ramnath Sai Sagar, Cloud Technical Marketing Mgr.  
    ramnath@mellanox.com
  - Jon Moran, Manager, Field Application Engineer, Dell Team  
    jonmoran@mellanox.com
  - Branko Cenanovic, Staff Field Application Engineer, EMC Team  
    branko@mellanox.com
  - Ron Gonzalez, Sr. Customer Project Manager, Dell Team  
    rong@mellanox.com
  - Bryan Varble, Staff Architect, Dell Team  
    bryan@mellanox.com
Backup Material:
In-Network Computing and Acceleration Engines

**RDMA GPUDirect**
Most Efficient Data Access and Data Movement for Compute and Storage platforms, SRIOV for HPC Clouds
200G with <1%CPU Utilization
10X Performance Improvement with GPUDirect

**Collectives**
CORE-Direct and SHArP Technologies
Executes and Manages Data Aggregation and Reduction Algorithms
Accelerates MPI, PGAS/SHMEM and UPC Communication Performance, Accelerates Machine Learning Training Algorithms

**Network Transport**
All Communications Managed and Operated by the Network Hardware; Adaptive Routing and Congestion Management, Dynamic Connected Transport (DCT)
Maximizes CPU Availability for Applications, increases Network Efficiency and Scalability

**Tag Matching**
MPI Tag-Matching Offload
MPI Rendezvous Protocol Offload
Accelerates MPI Application Performance

**Storage**
NVMe over Fabrics Offloads, T10-DIF and Erasure Coding offloads
Efficient End-to-End Data Protection, Background Check-Pointing (burst-buffer) and More. Increase System Performance and CPU Availability

**Security**
Data Encryption / Decryption (IEEE XTS standard) and Key Management; Federal Information Processing Standards (FIPS) Compliant
Enhances Data Security Options, Enables Protection Between Users Sharing the Same Resources (Different Keys)
Highest-Performance 100/200Gb/s Interconnect Solutions

**Adapters**
200Gb/s Adapter, 0.6us latency
200 million messages per second
(10 / 25 / 40 / 50 / 56 / 100 / 200Gb/s)

**Switch**
40 HDR (200Gb/s) InfiniBand Ports
80 HDR100 InfiniBand Ports
Throughput of 16Tb/s, <90ns Latency

**Switch**
32 100GbE Ports, 64 25/50GbE Ports
(10 / 25 / 40 / 50 / 100GbE)
Throughput of 3.2Tb/s

**Interconnect**
Transceivers
Active Optical and Copper Cables
(10 / 25 / 40 / 50 / 56 / 100 / 200Gb/s)
VCSELs, Silicon Photonics and Copper

**Software**
MPI, SHMEM/Pgas, UPC
For Commercial and Open Source Applications
Leverages Hardware Accelerations
Quantum 200G HDR InfiniBand Smart Switch

- 40 Ports of 200G HDR InfiniBand
- 80 Ports of 100G HDR100 InfiniBand
- Switch System 800 Ports 200G, 1600 Ports 100G

- 16Tb/s Switch Capacity
- Extremely Low Latency of 90ns
- 15.6 Billion Messages per Second

- In-Network Computing (SHARP Technology)
- Flexible Topologies (Fat-Tree, Torus, Dragonfly, etc.)
- Advanced Adaptive Routing
ConnectX-6 200G HDR InfiniBand and Ethernet Smart Adapter

- 100/200Gb/s Throughput
- 0.6usec End-to-End Latency
- 175/200M Messages per Second

- PCIe Gen3 and Gen4
- Integrated PCIe Switch and Multi-Host Technology
- Advanced Adaptive Routing

- In-Network Computing (Collectives, Tag Matching)
- In-Network Memory
- Storage (NVMe), Security and Network Offloads
SHArP Performance Advantage: The Computational Network

- MiniFE is a Finite Element mini-application
  - Implements kernels that represent implicit finite-element applications

10X to 25X Performance Improvement!
OpenFOAM is a popular computational fluid dynamics application.

HPC-X with SHArP Technology

HPC-X with SHArP Delivers 2.2X Higher Performance over Intel MPI
Safe Harbor Statement

These slides and the accompanying oral presentation contain forward-looking statements and information. The use of words such as “may”, “might”, “will”, “should”, “expect”, “plan”, “anticipate”, “believe”, “estimate”, “project”, “intend”, “future”, “potential” or “continued”, and other similar expressions are intended to identify forward-looking statements. All of these forward-looking statements are based on estimates and assumptions by our management that, although we believe to be reasonable, are inherently uncertain. Forward-looking statements involve risks and uncertainties, including, but not limited to, economic, competitive, governmental and technological factors outside of our control, that may cause our business, industry, strategy or actual results to differ materially from the forward-looking statement. These risks and uncertainties may include those discussed under the heading “Risk Factors” in the Company’s most recent 10K and 10Qs on file with the Securities and Exchange Commission, and other factors which may not be known to us. Any forward-looking statement speaks only as of its date. We undertake no obligation to publicly update or revise any forward-looking statement, whether as a result of new information, future events or otherwise, except as required by law.