High Performance Interconnect Enabling Deep Learning Efficiency

Mellanox Technologies

Deep learning become a hot topic in the industry
Deep learning need powerful computing system
Interconnect challenge for Deep Learning

- Low latency
- High bandwidth
- Direct communication between accelerator like GPU
- Offload
- Open Standard
High-Performance Designed 100Gb/s Interconnect Solutions

Adapters

ConnectX-4

- 100Gb/s Adapter, 0.7us latency
- 150 million messages per second
- (10 / 25 / 40 / 50 / 56 / 100Gb/s)

Switch

SwitchIB™

- 36 EDR (100Gb/s) Ports, <90ns Latency
- Throughput of 7.2Tb/s

Switch

Spectrum™

- 32 100GbE Ports, 64 25/50GbE Ports
  - (10 / 25 / 40 / 50 / 100GbE)
- Throughput of 6.4Tb/s

Interconnect

LinkX™

- Copper (Passive, Active)
- Optical Cables (VCSEL)
- Silicon Photonics
ConnectX-4 Adapter IC

- Single/dual port VPI adapter
  - InfiniBand: EDR / FDR / QDR / DDR / SDR
  - Ethernet: 100 / 56 / 40 / 50 / 25 / 10GbE

- Highest performance
  - 108Gb/s total throughput with Gen3 x16
  - 0.7us latency
  - 150 million messages per second

- Ethernet offloads
  - HDS, RSS, TSS, LRO, LSOv2
  - Advanced steering

- Multi-Host Technology
  - Up to 4 independent hosts
High Performance Computing Leadership

- GPUDirect Technology
  - Accelerating GPU communications
  - PeerDirect RDMA

- CORE-Direct technology
  - Acceleration engines for collective operations
  - Vector reductions

- Enhance all-to-all communication
  - Reduce CPU overhead
  - Congestion avoidance

- Performance and scalability
  - Dynamically Connected Transport (DCT)
  - Improved caching architecture
# Shattering The World of Interconnect Performance!

## ConnectX-4 EDR 100G InfiniBand

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>InfiniBand Throughput</td>
<td>100 Gb/s</td>
</tr>
<tr>
<td>InfiniBand Bi-Directional Throughput</td>
<td>195 Gb/s</td>
</tr>
<tr>
<td>InfiniBand Latency</td>
<td>0.61 us</td>
</tr>
<tr>
<td>InfiniBand Message Rate</td>
<td>149.5 Mpps</td>
</tr>
<tr>
<td>HPC-X MPI Bi-Directional Throughput</td>
<td>193.1 Gb/s</td>
</tr>
</tbody>
</table>

*First results, optimizations in progress*
Next Generation Switch – Unmatched Capacity & Scalability

Switch-IB EDR InfiniBand

- 36 x 100Gb/s EDR Ports
- 130ns latency
- Systems Availability Now
- 28nm technology

- Congestion Control
- Adaptive Routing
- Multiple topologies (Fat-Tree, DragonFly, Torus, etc.)
- InfiniBand specification 1.3
- Port mirroring
- Low power
Mellanox Adaptive Routing Notification (ARN)

- Internet switch to switch communications
- Faster convergence after routing changes
- Fast notification to decision point
- Fully configurable (topology agnostic)
- Adapt even when congestion is far away!

Faster Routing Modifications, Resilient Network
New Capabilities: Adaptive Routing & InfiniBand Router

- **Adaptive Routing**
  - ARN – Adaptive Routing Notification
    - Internal switch \(\rightarrow\) switch communications
    - Fast convergence after routing changes
    - Fast notification to decision point
  - Run over UD + MXM3.0

- **InfiniBand Router**
  - Segment large networks into smaller subnets \(\rightarrow\) scale!
    - Scaling to more than 100k nodes with cross bisectional bandwidth
  - Isolation
  - Implementation on EDR 1U based platforms
Shattering The World of Interconnect Performance!

<table>
<thead>
<tr>
<th>SwitchIB EDR 100G InfiniBand</th>
<th>Latency [ns]</th>
<th>BW[Gb/s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>@EDR Speed*</td>
<td>90</td>
<td>~100</td>
</tr>
<tr>
<td>(Strong FEC enabled)</td>
<td>(current release 140)</td>
<td>(current release 94.5)</td>
</tr>
<tr>
<td>@FDR Speed</td>
<td>130</td>
<td>52.5</td>
</tr>
<tr>
<td>(LLR enabled, requires FDR systems FW upgrade)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>@QDR Speed</td>
<td>114</td>
<td>31.3</td>
</tr>
<tr>
<td>InfiniBand Message Rate</td>
<td>149.5 Million/sec</td>
<td></td>
</tr>
</tbody>
</table>

*First results, optimizations part of next releases*
Excellent Cost Performance ratio for Datacenter Applications

Integrating the Best of Breed Physical Technologies

- Feature Rich
  - Low Power
  - 25Gb/s SerDes
- CMOS Switch and NIC

- Fastest, >25Gb/s
  - Optical Modulator
  - and Photo Diode
- Silicon Photonics

- Laser & Modulator Drivers
  - Trans-Impedance Amplifier
  - I/O Booster for Copper
- SiGe BiCMOS Chips

- High-speed
  - Electrical
  - and Optical
- Package & Assembly

Delivering Best in Class LinkX® Products
Mellanox Leadership in Silicon Photonics

- Data center standard 100G QSFP package
- Data center standard single mode fiber
- Mellanox leadership
  - First to demo in a public forum like OFC
  - 100 Gb/s links in QSFP
  - 100 Gb/s with 4dB link budget
  - 100 Gb/s transmission with BER better than $10^{-15}$
  - Both WDM and parallel solutions in QSFP form factor
  - Power consumption < 3 watts

Making 100Gb/s Deployments as Easy as 10Gb/s
EDR InfiniBand Performance – Commercial Applications

**OptiStruct Performance**  
(Engine_Assy.fem)

- Altair
- 40% improvement

**RADIOSS 13.0 Performance**  
(NEON1M11, MPP)

- Altair
- 25% improvement

**LS-DYNA Performance**  
(neon_refined_revised)

- FDR InfiniBand
- EDR InfiniBand

Performance Rating vs. Number of Nodes
- Weather Research and Forecasting Model
- Optimization effort with the HPCAC
- EDR InfiniBand delivers 28% higher performance
  - 32-node cluster
  - Performance advantage increase with system size
PeerDirect Technology

- Based on Peer-to-Peer capability of PCIe
- Support for any PCIe peer which can provide access to its memory
  - NVIDIA GPU, Xeon Phi, AMD, FPGA etc.

![PeerDirect RDMA Diagram](image-url)
- Eliminates CPU bandwidth and latency bottlenecks
- Uses remote direct memory access (RDMA) transfers between GPUs
- Resulting in significantly improved MPI SendRecv efficiency between GPUs in remote nodes
- Based on PeerDirect technology
GPUDirect Sync (GPUDirect 4.0)

- GPUDirect RDMA (3.0) – direct data path between the GPU and Mellanox interconnect
  - Control path still uses the CPU
    - CPU prepares and queues communication tasks on GPU
    - GPU triggers communication on HCA
    - Mellanox HCA directly accesses GPU memory

- GPUDirect Sync (GPUDirect 4.0)
  - Both data path and control path go directly between the GPU and the Mellanox interconnect

Maximum Performance For GPU Clusters
• HOOMD-blue is a general-purpose Molecular Dynamics simulation code accelerated on GPUs
• GPUDirect RDMA allows direct peer to peer GPU communications over InfiniBand
  • Unlocks performance between GPU and InfiniBand
  • This provides a significant decrease in GPU-GPU communication latency
  • Provides complete CPU offload from all GPU communications across the network
• Demonstrated up to 102% performance improvement with large number of particles
Leading Supplier of End-to-End Interconnect Solutions

Enabling the Use of Data

Comprehensive End-to-End InfiniBand and Ethernet Portfolio

<table>
<thead>
<tr>
<th>ICs</th>
<th>Adapter Cards</th>
<th>Switches/Gateways</th>
<th>Software and Services</th>
<th>Metro / WAN</th>
<th>Cables/Modules</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="ICs" /></td>
<td><img src="image" alt="Adapter Cards" /></td>
<td><img src="image" alt="Switches/Gateways" /></td>
<td><img src="image" alt="Software and Services" /></td>
<td><img src="image" alt="Metro / WAN" /></td>
<td><img src="image" alt="Cables/Modules" /></td>
</tr>
</tbody>
</table>

At the Speeds of 10, 25, 40, 50, 56 and 100 Gigabit per Second
Thank You