Paving The Road to Exascale Computing

Gilad Shainer, Dror Goldenberg

HPC@mellanox.com

IEEE Cluster 2010
IEEE International Conference on Cluster Computing 2010
20 - 24 September 2010 - Heraklion, Crete, Greece
Connectivity Solutions for Efficient Computing

Mellanox Interconnect Networking Solutions

<table>
<thead>
<tr>
<th>ICs</th>
<th>Adapter Cards</th>
<th>Host/Fabric Software</th>
<th>Switches/Gateways</th>
<th>Cables</th>
</tr>
</thead>
<tbody>
<tr>
<td>![ICs Image]</td>
<td>![Adapter Cards Image]</td>
<td>![Host/Fabric Software Image]</td>
<td>![Switches/Gateways Image]</td>
<td>![Cables Image]</td>
</tr>
</tbody>
</table>

© 2010 MELLANOX TECHNOLOGIES - CONFIDENTIAL -
Mellanox InfiniBand builds the most powerful clusters
- 5 of the Top10 (#2, #3, #6, #7, #10) and 64 of the Top100

InfiniBand represents 42% of the Top500 (208 systems)
- All InfiniBand clusters use Mellanox solutions
InfiniBand Unsurpassed System Efficiency

**World Leading Compute Systems Efficiency Comparison**

- Top500 systems listed according to their efficiency
- InfiniBand is the key element responsible for the highest systems efficiency
Highest Performance

- **Highest throughput**
  - 40Gb/s node to node and 120Gb/s switch to switch
  - Nearly 90M MPI messages per second
  - Send/receive and RDMA operations with zero-copy

- **Lowest latency**
  - 1-1.3usec MPI end-to-end
  - 0.9-1us InfiniBand latency for RDMA operations
  - Lowest latency 648-port switch – 25% to 45% faster vs other solutions

- **Lowest CPU overhead**
  - Full transport offload maximizes CPU availability for user applications
Advanced HPC Capabilities

- **CORE-Direct (Collectives Offload Resource Engine)**
  - Collectives communication are communications used for system synchronizations, data broadcast or data gathering
  - Eliminates system noise and jitter issue
  - Increases the CPU cycles available for applications
  - Allows the processes communications to progress asynchronously

- **GPUDirect**
  - Enables fastest GPU-to-GPU communications
  - Reduces 30% of the GPU-to-GPU communication time

- **Congestion control**
  - Eliminates network congestions (hot-spots) related to many senders and a single receiver

- **Adaptive routing**
  - Eliminated networks congestions related to point to point communications sharing the same network path
Mellanox – NVIDIA GPUDirect Technology

- Allows Mellanox InfiniBand and NVIDIA GPU to communicate faster
  - Eliminates memory copies between InfiniBand and GPU

Mellanox-NVIDIA GPUDirect Enables Fastest GPU-to-GPU Communications
Accelerating GPU Based Supercomputing

- Fast GPU to GPU communications
- Native RDMA for efficient data transfer
- Reduces latency by 30% for GPUs communication
33% performance increase with GPUDirect

Performance benefit increases with cluster size
### InfiniBand Link Speed Roadmap

<table>
<thead>
<tr>
<th>Lanes per direction</th>
<th>5G-IB DDR</th>
<th>10G-IB QDR</th>
<th>14G-IB-FDR</th>
<th>26G-IB-EDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>60+60</td>
<td>120+120</td>
<td>168+168</td>
<td>300+300</td>
</tr>
<tr>
<td>8</td>
<td>40+40</td>
<td>80+80</td>
<td>112+112</td>
<td>200+200</td>
</tr>
<tr>
<td>4</td>
<td>20+20</td>
<td>40+40</td>
<td>56+56</td>
<td>100+100</td>
</tr>
<tr>
<td>1</td>
<td>5+5</td>
<td>10+10</td>
<td>14+14</td>
<td>25+25</td>
</tr>
</tbody>
</table>

NDR = Next Data Rate  
HDR = High Data Rate  
EDR = Enhanced Data Rate  
FDR = Fourteen Data Rate  
QDR = Quad (4xSDR) Data Rate  
DDR = Double (2xSDR) Data Rate  
SDR = Single Data Rate (not shown)
Bringing HPC to the Exascale

The Only Scalable Networking for Exascale HPC
Thank You

HPC@mellanox.com