HPC Interconnect Technology update

Paving the Road to Exascale – HPC User Forum

2017
The Ever-Growing Demand for Higher Performance

Performance Development

Terascale  Petascale  Exascale


LANL  Wuxi

1st  1st

The Interconnect is the Enabling Technology

SMP to Clusters  Single-Core to Many-Core

Co-Design

Application
Software
Hardware

© 2017 Mellanox Technologies
Exponential Data Growth – The Need for Intelligent and Faster Interconnect

**CPU-Centric (Onload)**

- Must Wait for the Data
- Creates Performance Bottlenecks

**Data-Centric (Offload)**

- Analyze Data as it Moves!

Faster Data Speeds and In-Network Computing Enable Higher Performance and Scale
Data Centric Architecture to Overcome Latency Bottlenecks

CPU-Centric (Onload)

- HPC / Machine Learning Communications Latencies of 30-40us

Data-Centric (Offload)

- HPC / Machine Learning Communications Latencies of 3-4us

Intelligent Interconnect Paves the Road to Exascale Performance
In-Network Computing to Enable Data-Centric Data Center

In-Network Computing Key for Highest Return on Investment
In-Network Computing to Enable Data-Centric Data Center

- CPU
- GPU
- CPU
- GPU
- CPU
- GPU

RDMA
CORE-Direct Tag-Matching
In-Network Memory

Programmable (FPGA)

Programmable (ARM)

SHARP v1/v2

BlueField

In-Network Computing Key for Highest Return on Investment
<table>
<thead>
<tr>
<th>In-Network Computing and Acceleration Engines</th>
</tr>
</thead>
</table>

### 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

### 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

### 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

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

### 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)
MPITag-Matching Offload Advantages

- 31% lower latency and 97% lower CPU utilization for MPI operations
- Performance comparisons based on ConnectX-5

Mellanox In-Network Computing Technology Deliver Highest Performance
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.

OpenFOAM : Lid Driven Cavity Flow
icoFoam solver, 2D 1 million cells

Performance Rating (Jobs per Day)

<table>
<thead>
<tr>
<th>Number of Nodes</th>
<th>Intel MPI</th>
<th>Open MPI</th>
<th>HPC-X (SHARP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>200</td>
<td>200</td>
<td>300</td>
</tr>
<tr>
<td>8</td>
<td>400</td>
<td>400</td>
<td>600</td>
</tr>
<tr>
<td>16</td>
<td>600</td>
<td>600</td>
<td>1,200</td>
</tr>
<tr>
<td>32</td>
<td>800</td>
<td>800</td>
<td>1,600</td>
</tr>
<tr>
<td>64</td>
<td>1,000</td>
<td>1,000</td>
<td>2,200</td>
</tr>
</tbody>
</table>

HPC-X with SHARP Delivers **2.2X** Higher Performance over Intel MPI
Scalable, flexible, high performance, high bandwidth, end-to-end connectivity

Standards-based and supported by the largest eco-system

Supports all compute architectures: x86, Power, ARM, GPU, FPGA etc.

Native Offloading architecture

RDMA, GPUDirect, rCUDA, SHARP and other accelerations

Backward and future compatible

Scalable HPC Depends on Mellanox
OrionX Reports Position InfiniBand as the Leading HPI Technology and Mellanox the Leading Vendor

July 7, 2016 by staff

In this special guest feature, Peter Foulkes from OrionX outlines a series of new reports that show how InfiniBand continues to dominate the market for High Performance Interconnects.

The OrionX Constellation reports, published June 29th address the evolution, environment, evaluation and excellence ratings for the High Performance Interconnect (HPI) market. Defined as the very high end of the networking equipment market where high bandwidth and low latency are non-negotiable, HPI technologies support the most demanding workloads that are typical of extreme-scale systems in high performance computing (HPC), artificial intelligence, cloud computing, and web-scale deployments.

[Link to the article]

InfiniBand Enables Intelligent Networks

January 13, 2016 by staff

In this special guest feature, Gilad Shainer from Mellanox writes that the network is the key to future scalable systems.

HPC Frequently Reinvents Itself to Keep Pace

In the world of high-performance computing, there is a constant and ever-growing demand for even higher performance. Technology providers have worked ceaselessly to keep up with that demand, with each new generation of faster, more reliable, and more efficient systems. Ultimately, though, every technology reaches its limits, and progress can therefore stall unless there is a way forward.

[Link to the article]

Slidecast: Advantages of Offloading Architectures for HPC

April 19, 2016 by Eyal Brodsky

Interconnect: Your Future with InfiniBand

Supports up to 6 Different Subnets

[Link to the article]

Media Resources

[Link to the article]

Interview: Why Co-design is the Path Forward for Exascale Computing

March 8, 2016 by Eyal Brodsky

Gild Shainer, Mellanox

[Link to the article]

April 12, 2016

The Ultimate Debate – Interconnect Offloading Versus Onloading

Gild Shainer, Mellanox

The high performance computing market is going through a technology transition—the Co-Design transition. As has already been discussed in many articles, this transition has emerged in order to solve the performance bottlenecks of today’s infrastructures and applications. Performance bottlenecks today are created by multi-core CPUs and the existing CPU-centric system architecture.

How are multi-core CPUs the source for today’s performance bottlenecks? In order to understand that, we need to go back in time to the era of single-core CPUs. Back then, performance gains came from increases in CPU frequency and from the reduction of networking functions (network adapter and switches). Each new generation of product brought faster CPUs and lower-latency network adapters and network interface cards (NICs).

One of the primary conversations these days in the field of networking is whether it is better to offload network functions onto the CPU or better to offload these functions to the interconnect hardware.

Offloading Interconnect technology is easier to build, but the issue becomes the CPU utilization, because the CPU must manage and execute network operations. It has less availability for applications, which is its primary purpose.

Offloading, on the other hand, seeks to overcome performance bottlenecks in the CPU by performing the network functions, as well as complex communications operations.
The insideHPC Guide to Co Design Architecture

The use of Co-Design and offloading are important tools in achieving Exascale computing. Application developers and system designers can take advantage of network offload and emerging co-design protocols to accelerate their current applications. Adopting some basic co-design and offloading methods to smaller scale systems can achieve more performance on less hardware resulting in low cost and higher throughput. Learn more by downloading this guide.

Link to the article

Smart Interconnect: The Next Key Driver of HPC Performance Gains

With the increasing adoption of in-memory architectures and cloud computing, high performance interconnect (HPC) technologies have become a more critical part of IT systems. Today, HPC represents the core market segment of the interconnects of the networking equipment market, supporting applications requiring extremely low latency and exceptionally high bandwidth.

As big data analysis, machine learning, and business optimization applications become more prevalent, HPC technologies are increasingly important for enterprises as well. More demanding enterprise applications, as well as high-performance computing (HPC) applications, are generally characterized with scale-out clusters based on large numbers of “Galexy” nodes. The complementation for large node counts places a heavy

Link to the webinar

Choosing the right interconnect for high-performance compute and storage platforms is critical for achieving the highest possible system performance and overall return on investment.

Over time, interconnect technologies have become more sophisticated, and include more intelligent capabilities (offload engines), which enable the interconnect to do more than just transferring data. Intelligent interconnect can increase system efficiency, interconnect with offload engine (olofmed interconnect) dramatically reduces CPU overhead, allowing more CPU cycles to be dedicated to applications and therefore enabling higher application performance and user productivity.

Link to the article

© 2017 Mellanox Technologies
Thank You