High-Performance GPU Clustering: GPUDirect RDMA over 40GbE iWARP

Tom Reu
Consulting Applications Engineer
Chelsio Communications
tomreu@chelsio.com
Chelsio Corporate Snapshot

Leader in High Speed Converged Ethernet Adapters

- Leading 10/40GbE adapter solution provider for servers and storage systems
  - ~800K ports shipped
- High performance protocol engine
  - 80MPPS
  - 1.5μsec
  - ~5M+ IOPs
- Feature rich solution
  - Media streaming hardware/software
  - WAN Optimization, Security, etc.
- Company Facts
  - Founded in 2000
  - 150 strong staff
- R&D Offices
  - USA – Sunnyvale
  - India – Bangalore
  - China - Shanghai
RDMA Overview

- Direct memory-to-memory transfer
- All protocol processing handling by the NIC
  - Must be in hardware
- Protection handled by the NIC
  - User space access requires both local and remote enforcement
- Asynchronous communication model
  - Reduced host involvement
- Performance
  - Latency - polling
  - Throughput
- Efficiency
  - Zero copy
  - Kernel bypass (user space I/O)
  - CPU bypass

Performance and efficiency in return for new communication paradigm
iWARP

What is it?

• Provides the ability to do Remote Direct Memory Access over Ethernet using TCP/IP
• Uses Well-Known IB Verbs
• Inboxed in OFED since 2008
• Runs on top of TCP/IP
  • Chelsio implements iWARP/TCP/IP stack in silicon
  • Cut-through send
  • Cut-through receive
• Benefits
  • Engineered to use “typical” Ethernet
    • No need for technologies like DCB or QCN
  • Natively Routable
  • Multi-path support at Layer 3 (and Layer 2)
  • It runs on TCP/IP
    • Mature and Proven
    • Goes where TCP/IP goes (everywhere)
iWARP

- iWARP updates and enhancements are done by the IETF STORM (Storage Maintenance) working group
- RFCs
  - RFC 5041 Direct Data Placement over Reliable Transports
  - RFC 5044 Marker PDU Aligned Framing for TCP Specification
  - RFC 6580 IANA Registries for the RDDP Protocols
  - RFC 6581 Enhanced RDMA Connection Establishment
  - RFC 7306 Remote Direct Memory Access (RDMA) Protocol Extensions
- Support from several vendors, Chelsio, Intel, QLogic
iWARP

Increasing Interest in iWARP as of late

- Some Use Cases
  - High Performance Computing
  - SMB Direct
  - GPUDirect RDMA
  - NFS over RDMA
  - FreeBSD iWARP
  - Hadoop RDMA
  - Lustre RDMA
  - NVMe over RDMA fabrics
iWARP

Advantages over Other RDMA Transports

• It’s Ethernet
  • Well Understood and Administered
  • Uses TCP/IP
    • Mature and Proven
    • Supports rack, cluster, datacenter, LAN/MAN/WAN and wireless
    • Compatible with SSL/TLS
  • Do not need to use any bolt-on technologies like
    • DCB
    • QCN
• Does not require a totally new network infrastructure
  • Reduces TCO and OpEx
# iWARP vs RoCE

<table>
<thead>
<tr>
<th>iWARP</th>
<th>RoCE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Native</strong> TCP/IP over Ethernet, no different from NFS or HTTP</td>
<td>Difficult to install and configure - “needs a team of experts” - Plug-and-Debug</td>
</tr>
<tr>
<td>Works with ANY Ethernet switches</td>
<td>Requires DCB - expensive equipment upgrade</td>
</tr>
<tr>
<td>Works with ALL Ethernet equipment</td>
<td>Poor interoperability - may not work with switches from different vendors</td>
</tr>
<tr>
<td>No need for special QoS or configuration - <strong>TRUE Plug-and-Play</strong></td>
<td>Fixed QoS configuration - DCB must be setup identically across all switches</td>
</tr>
<tr>
<td>No need for special configuration, preserves network robustness</td>
<td>Easy to break - switch configuration can cause performance collapse</td>
</tr>
<tr>
<td>TCP/IP allows reach to <strong>Cloud</strong> scale</td>
<td>Does not scale - requires PFC, limited to single subnet</td>
</tr>
<tr>
<td>No distance limitations. Ideal for remote communication and HA</td>
<td>Short distance - PFC range is limited to few hundred meters maximum</td>
</tr>
<tr>
<td><strong>WAN routable</strong>, uses any IP infrastructure</td>
<td>RoCEv1 not <strong>routable</strong>. RoCE v2 requires lossless IP infrastructure and restricts router configuration</td>
</tr>
<tr>
<td>Standard for whole stack has been <strong>stable</strong> for a decade</td>
<td><strong>ROCEv2 incompatible</strong> with v1. More fixes to missing reliability and scalability layers required and expected</td>
</tr>
<tr>
<td><strong>Transparent and open</strong> IETF standards process</td>
<td>Incomplete specification and <strong>opaque</strong> process</td>
</tr>
</tbody>
</table>

---

Efficient Performance™
Chelsio’s T5

Single ASIC does it all

• High Performance Purpose Built Protocol Processor
• Runs multiple protocols
  • TCP with Stateless Offload and Full Offload
  • UDP with Stateless Offload
  • iWARP
  • FCoE with Offload
  • iSCSI with Offload
• All of these protocols run on T5 with a SINGLE FIRMWARE IMAGE
  • No need to reinitialize the card for different uses
  • Future proof e.g. support for NVMf yet preserves today’s investment in iSCSI
T5 ASIC Architecture

High Performance Purpose Built Protocol Processor

- Single processor data-flow pipelined architecture
- Up to 1M connections
- Concurrent Multi-Protocol Operation

- Single connection at 40Gb. Low Latency.
Leading Unified Wire™ Architecture

Converged Network Architecture with all-in-one Adapter and Software

- **Storage**
  - NVMe/Fabrics
  - SMB Direct
  - iSCSI and FCoE with T10-DIX
  - iSER and NFS over RDMA
  - pNFS (NFS 4.1) and Lustre
  - NAS Offload
  - Diskless boot
  - Replication and failover

- **Virtualization & Cloud**
  - Hypervisor offload
  - SR-IOV with embedded VEB
  - VEPA, VN-TAGs
  - VXLAN/NVGRE
  - NFV and SDN
  - OpenStack storage
  - Hadoop RDMA

- **HPC**
  - iWARP RDMA over Ethernet
  - GPUDirect RDMA
  - Lustre RDMA
  - pNFS (NFS 4.1)
  - OpenMPI
  - MVAPICH

- **Networking**
  - 4x10GbE/2x40GbE NIC
  - Full Protocol Offload
  - Data Center Bridging
  - Hardware firewall
  - Wire Analytics
  - DPDK/netmap

- **Media Streaming**
  - Traffic Management
  - Video segmentation Offload
  - Large stream capacity

**Single Qualification — Single SKU**
**Concurrent Multi-Protocol Operation**
GPUDirect RDMA

- Introduced by NVIDIA with the Kepler Class GPUs. Available today on Tesla and Quadro GPUs as well.
- Enables Multiple GPUs, 3rd party network adapters, SSDs and other devices to read and write CUDA host and device memory.
- Avoids unnecessary system memory copies and associated CPU overhead by copying data directly to and from pinned GPU memory.
- One hardware limitation
  - The GPU and the Network device MUST share the same upstream PCIe root complex.
- Available with Infiniband, RoCE, and now iWARP.
GPUDirect RDMA

T5 iWARP RDMA over Ethernet certified with NVIDIA GPUDirect

- Read/write GPU memory directly from network adapter
  - Peer-to-peer PCIe communication
  - Bypass host CPU
  - Bypass host memory
- Zero copy
- Ultra low latency
- Very high performance
- Scalable GPU pooling
  - Any Ethernet networks
Modules required for GPUDirect RMDA with iWARP

- Chelsio Modules
  - cxgb4 - Chelsio adapter driver
  - iw_cxgb4 - Chelsio iWARP driver
  - rdma_ucm - RDMA User Space Connection Manager
- NVIDIA Modules
  - nvidia - NVIDIA driver
  - nvidia_uvm - NVIDIA Unified Memory
  - nv_peer_mem - NVIDIA Peer Memory
Case Studies
HOOMD-blue

- General Purpose Particle simulation toolkit
- Stands for: **Highly Optimized** **Object-oriented** **Many-particle Dynamics - Blue Edition**
- Running on GPUDirect RDMA - **WITH NO CHANGES TO THE CODE - AT ALL!**
- More Info: [www.codeblue.umich.edu/hoomd-blue](http://www.codeblue.umich.edu/hoomd-blue)
HOOMD-blue

Test Configuration

- 4 Nodes
- Intel E5-1660 v2 @ 3.7 Ghz
- 64 GB RAM
- Chelsio T580-CR 40Gb Adapter
- NVIDIA Tesla K80 (2 GPUs per card)
- RHEL 6.5
- OpenMPI 1.10.0
- OFED 3.18
- CUDA Toolkit 6.5
- HOOMD-blue v1.3.1-9
- Chelsio-GDR-1.0.0.0
- Command Line:

```
$MPI_HOME/bin/mpirun --allow-run-as-root -mca btl_openib.want_cuda_gdr 1 -np X -hostfile /root/hosts -mca btl_openib,sm,self -mca btl_openib_if_include cxgb4:0:1 --mca btl_openib_cuda_rdma_limit 65538 -mca btl_openib_receive_queues P,131072,64 -x CUDA_VISIBILE_DEVICES=0,1 /root/hoomd-install/bin/hoomd ./bmark.py --mode=gpu|cpu
```
HOOMD-blue

Lennard-Jones Liquid 64K Particles Benchmark

• Classic benchmark for general purpose MD simulations.
• Representative of the performance HOOMD-blue achieves for straight pair potential simulations
HOOMD-blue

Lennard-Jones Liquid 64K Particles Benchmark Results

Average Timesteps per Second

Test 1
- 26 CPU Cores
- 488 timesteps
- 2 GPUs
- 1,230 timesteps
- 2 GPUs

Test 2
- 88 CPU Cores
- 503 timesteps
- 4 GPUs
- 1,403 timesteps
- 4 GPUs

Test 3
- 214 CPU Cores
- 1,089 timesteps
- 8 GPUs
- 1,771 timesteps
- 8 GPUs

CPU
GPU w/o GPUDirect RDMA
GPU w/ GPUDirect RDMA

Longer is Better

Chelsio Communications Accelerate
HOOMD-blue

Lennard-Jones Liquid 64K Particles Benchmark Results

Hours to complete 10e6 steps

Test 1
- 6 hours with 2 GPUs
- 2.2 hours with 2 GPUs

Test 2
- 5.5 hours with 4 GPUs
- 1.7 hours with 4 GPUs

Test 3
- 13 hours with 40 CPUs
- 2.5 hours with 8 GPUs
- 1.5 hours with 8 GPUs

CPU
- 108 hours with 2 CPUs

GPU w/o GPUDirect RDMA
- 32 hours with 8 CPUs

GPU w/ GPUDirect RDMA
- Shorter is Better

Chelsio Communications Accelerate
• runs a system of particles with an oscillatory pair potential that forms a icosahedral quasicrystal
**HOOMD-blue**

**Quasicrystal results**

Average Timesteps per Second

- **Test 1**
  - 2 CPU Cores: 11
  - 2 GPUs: 308
  - 407 (2 GPUs)

- **Test 2**
  - 8 CPU Cores: 43
  - 4 GPUs: 656
  - 728 (4 GPUs)

- **Test 3**
  - 40 CPU Cores: 31
  - 8 GPUs: 915
  - 1,158 (8 GPUs)

**Legend**
- **CPU**
- **GPU w/o GPUDirect RDMA**
- **GPU w/ GPUDirect RDMA**

*Longer is Better*
HOOMD-blue

Quasicrystal results

Hours to complete 10e6 steps

Test 1
- CPU: 264 hours
- GPU w/o GPUDirect RDMA: 2 hours
- GPU w/ GPUDirect RDMA: 9 hours

Test 2
- CPU: 63 hours
- GPU w/o GPUDirect RDMA: 4 hours
- GPU w/ GPUDirect RDMA: 4 hours

Test 3
- CPU: 86 hours
- GPU w/o GPUDirect RDMA: 3 hours
- GPU w/ GPUDirect RDMA: 2.4 hours

Shorter is Better
Caffe
Deep Learning Framework

• Open source Deep Learning software from Berkeley Vision and Learning Center
• Updated to include CUDA support to utilize GPUs
• Standard version does NOT include MPI support
• MPI implementations
  • mpi-caffe
    • Used to train a large network across a cluster of machines
    • model-parallel distributed approach.
  • caffe-parallel
    • Faster framework for deep learning.
    • data-parallel via MPI, splits the training data across nodes
Summary

GPUDirect RDMA over 40GbE iWARP

- iWARP provides RDMA Capabilities to a Ethernet network
- iWARP uses tried and true TCP/IP as its underlying transport mechanism
- Using iWARP does not require a whole new network infrastructure and the management requirements that come along with it
- iWARP can be used with existing software running on GPUDirect RDMA which NO CHANGES required to the code
- Applications that use GPUDirect RDMA will see huge performance improvements
- Chelsio provides 10/40Gb iWARP TODAY with 25/50/100 Gb on the horizon
More information

GPUDirect RDMA over 40GbE iWARP

- Visit our website, www.chelsio.com, for more White Papers, Benchmarks, etc.
- Webinar: https://www.brighttalk.com/webcast/13671/189427
- Beta code for GPUDirect RDMA is available TODAY from our download site at service.chelsio.com
- Sales questions - sales@chelsio.com
- Support questions - support@chelsio.com
Questions?
Thank You