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

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

Chelsio Communications
Accelerate

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

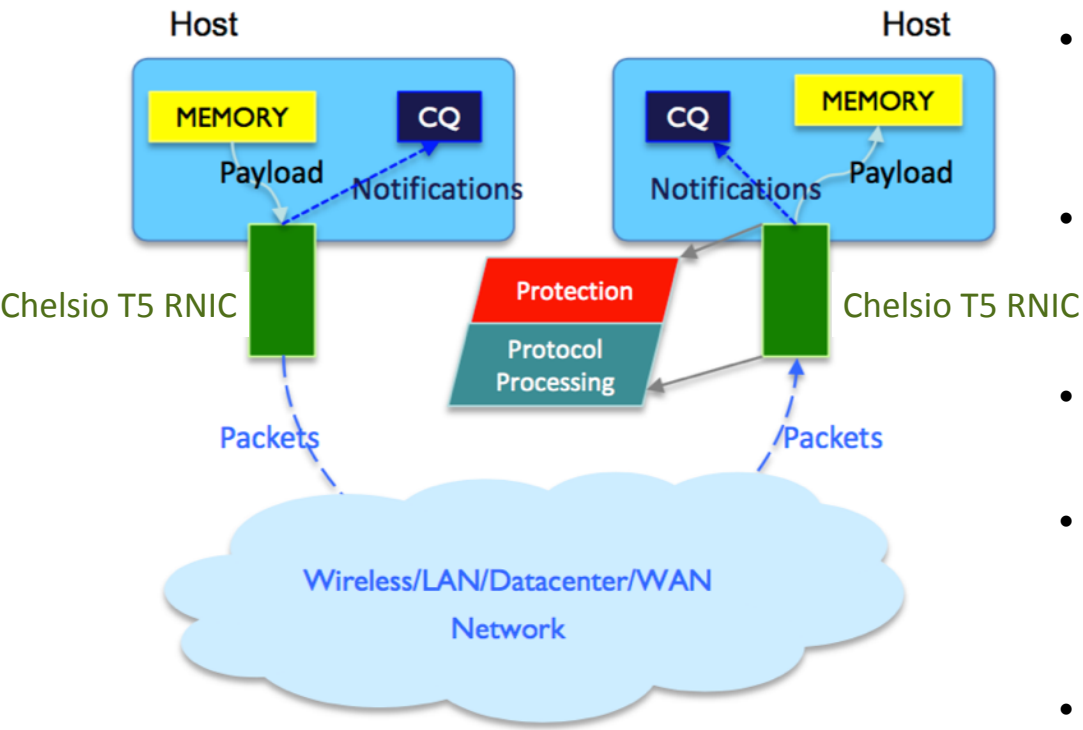
ISO 9001:2000 REGISTERED

CERTIFICATE NO: 164305

OEM Snapshot

NETFLIX inspur 浪潮 IBM ZTE中兴 sgi Microsoft DELL NEC 中科曙光 Sugon HUAWEI FOXCONN Chelsio Communications Accelerate

RDMA Overview



Performance and efficiency in return for new communication paradigm

- 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

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 5040 A Remote Direct Memory Access Protocol Specification
  - 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

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

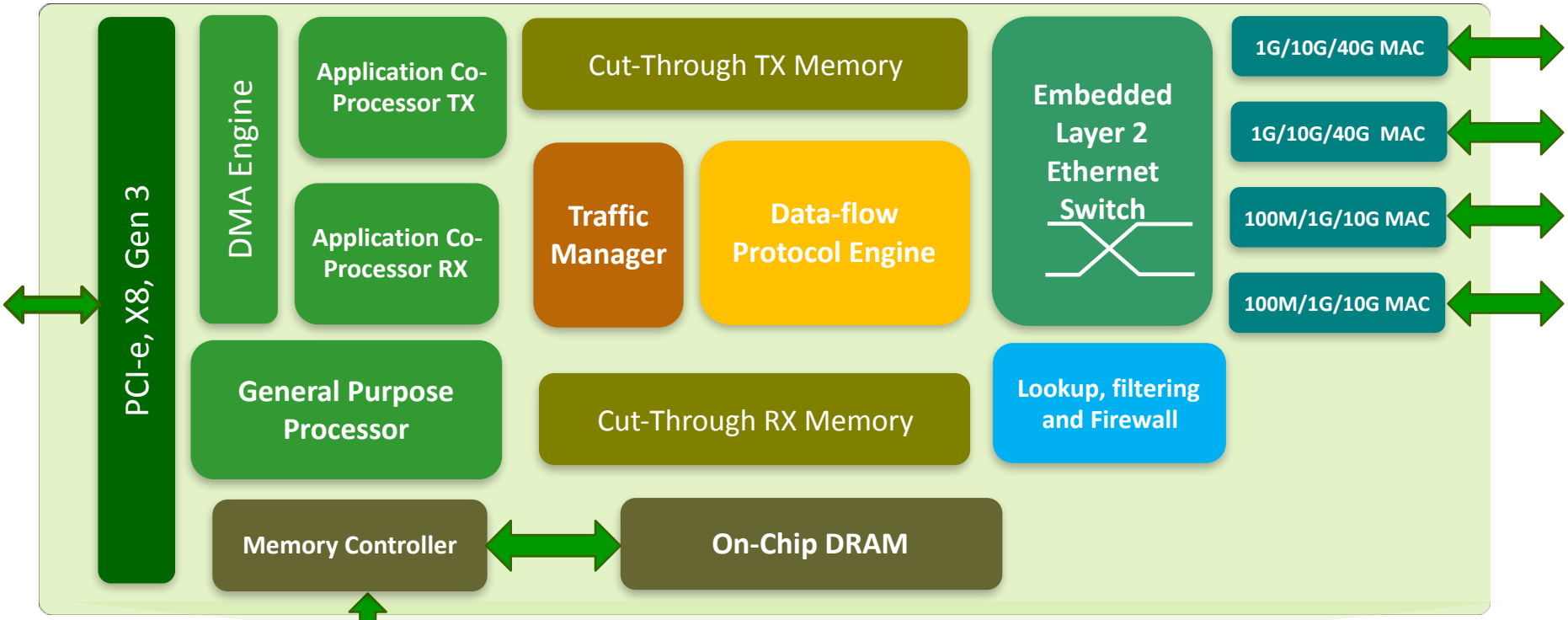
Chelsio's T5

Single ASIC does it all

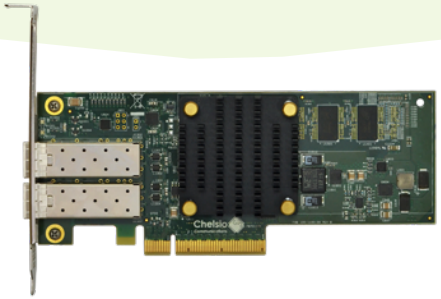
- High Performance Purpose Built Protocol Processor
- Runs multiple protocols
  - TCP with Stateless Offload and Full Offload
  - UDP with Stateless Offload
  - iWARP
  - FCoE with Offload
  - iSCSI with Offload
- All of these protocols run on T5 with a SINGLE FIRMWARE IMAGE
  - No need to reinitialize 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



Optional external DDR3 memory



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

Single connection at 40Gb. Low Latency.

Chelsio Communications
Accelerate

Leading Unified Wire™ Architecture

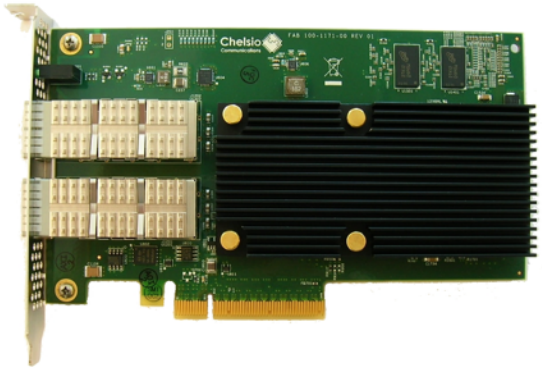
Converged Network Architecture with all-in-one Adapter and Software

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

HFT
- WireDirect technology
- Ultra low latency
- Highest messages/sec
- Wire rate classification

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

Chelsio Communications
Accelerate

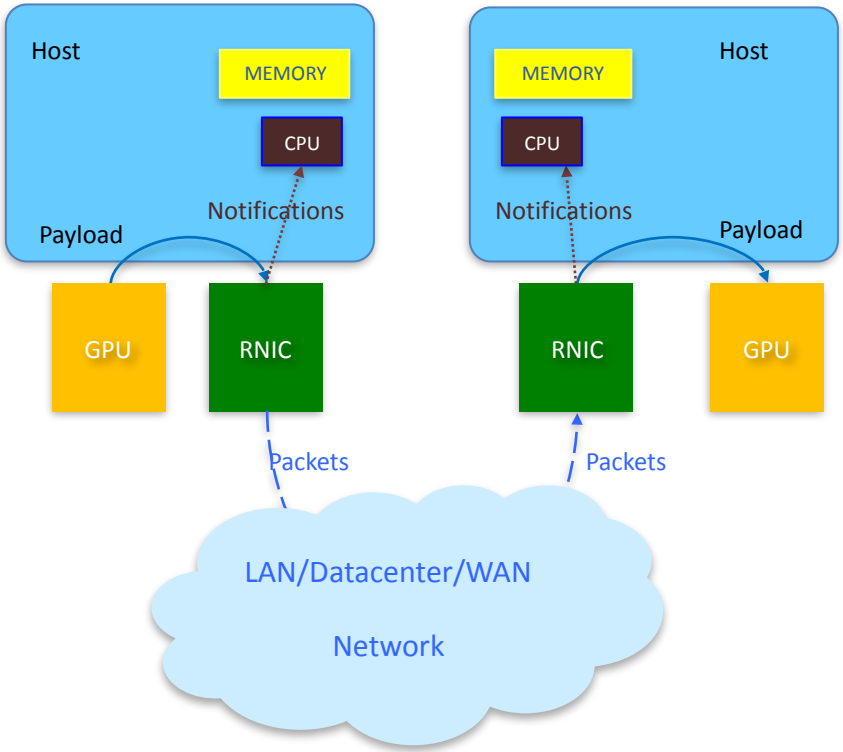
GPUDirect RDMA

- 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

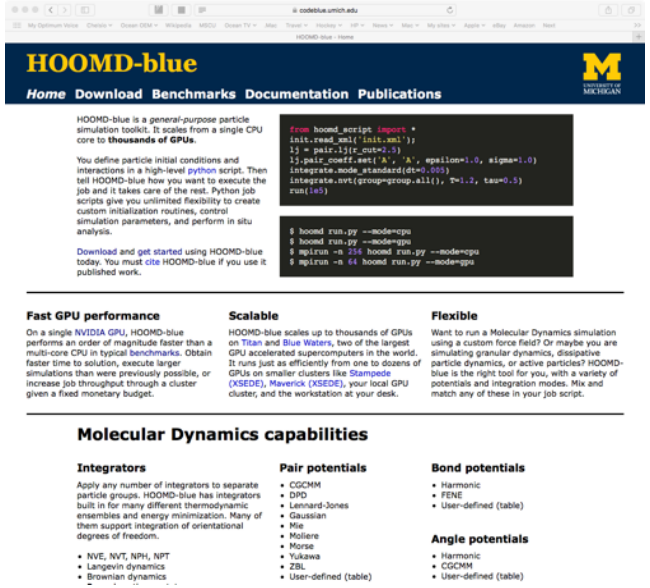
Chelsio Communications

Accelerate

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



HOOMD-blue

Test Configuration

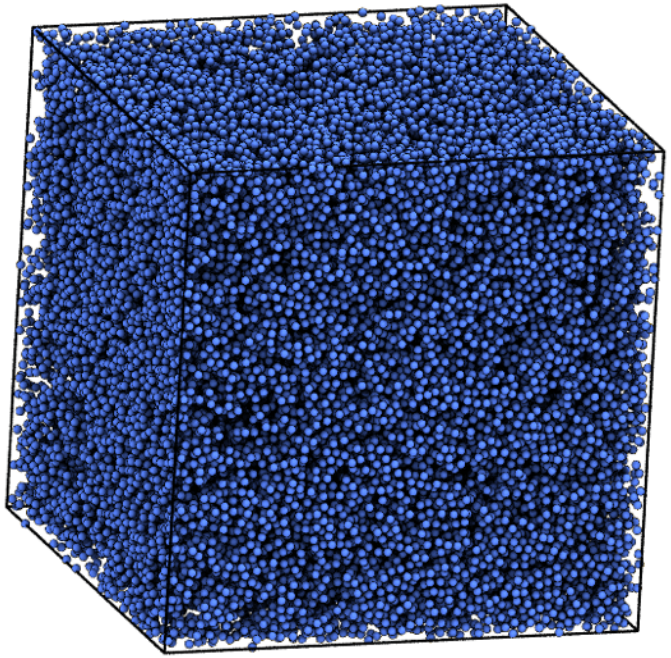
- 4 Nodes
- Intel E5-1660 v2 @ 3.7 Ghz
- 64 GB RAM
- Chelsio T580-CR 40Gb Adapter
- NVIDIA Tesla K80 (2 GPUs per card)
- RHEL 6.5
- OpenMPI 1.10.0
- OFED 3.18
- CUDA Toolkit 6.5
- HOOMD-blue v1.3.1-9
- Chelsio-GDR-1.0.0.0
- Command Line:

```
$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

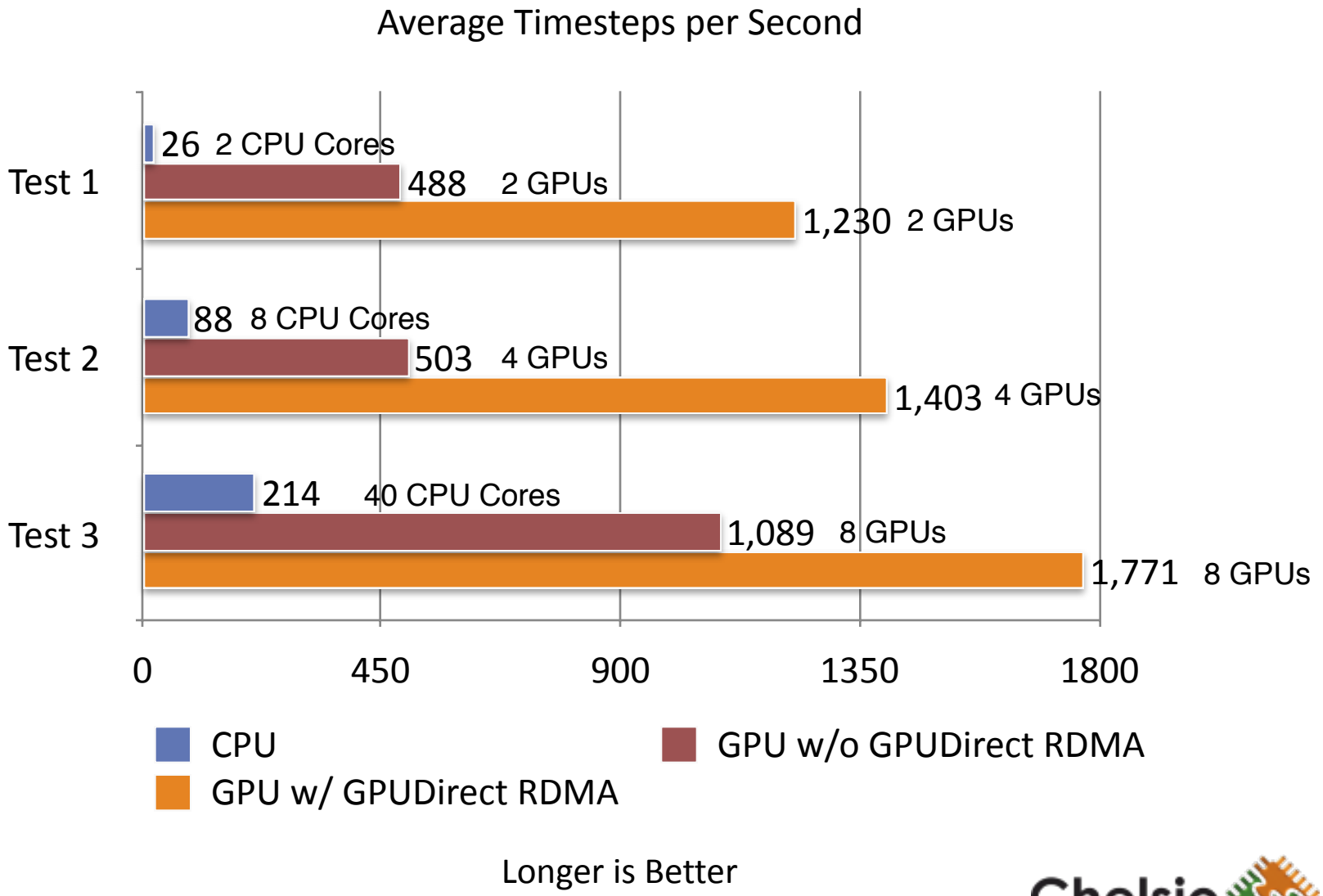
Lennard-Jones Liquid 64K Particles Benchmark

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



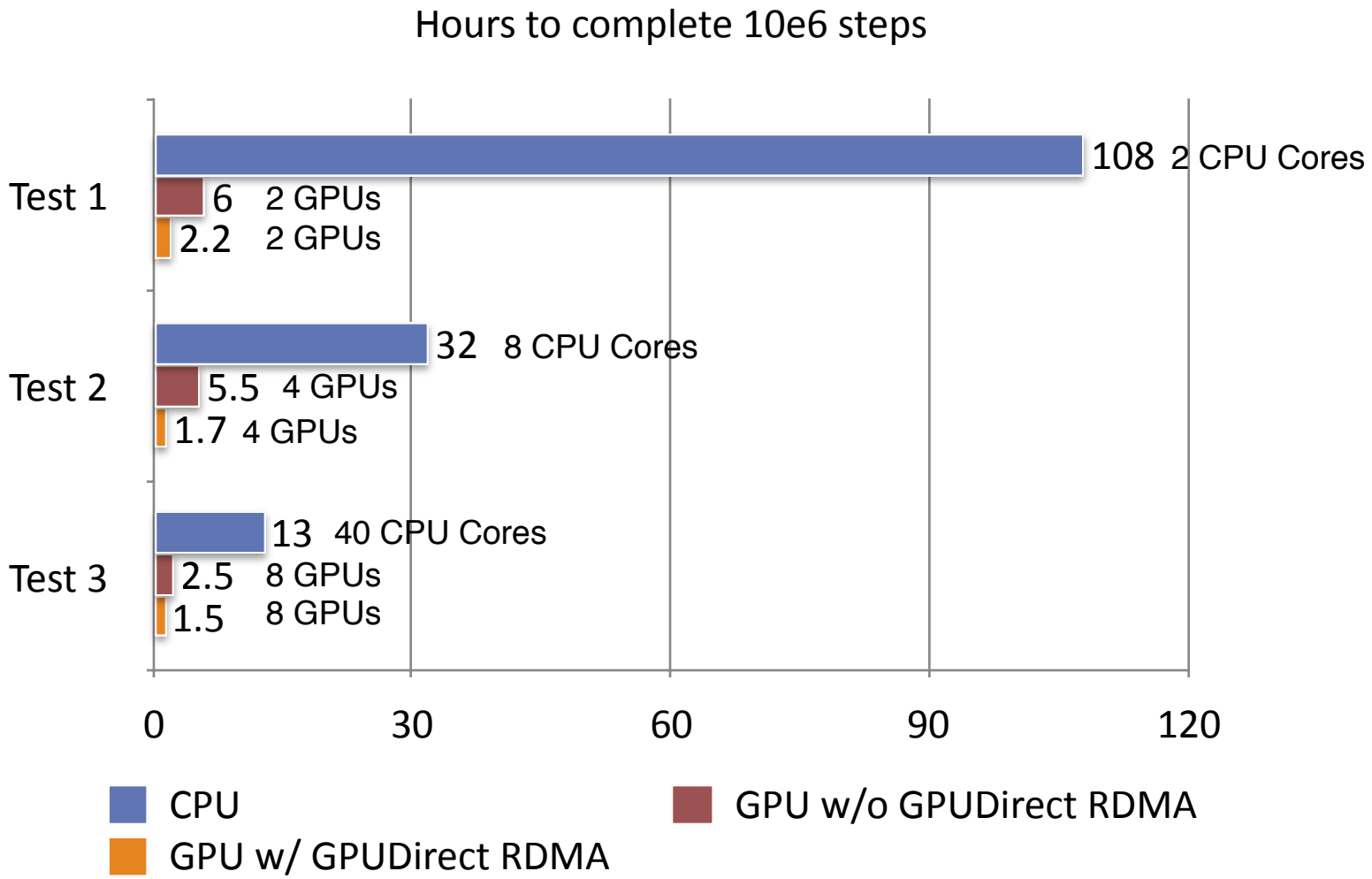
HOOMD-blue

Lennard-Jones Liquid 64K Particles Benchmark Results



HOOMD-blue

Lennard-Jones Liquid 64K Particles Benchmark Results

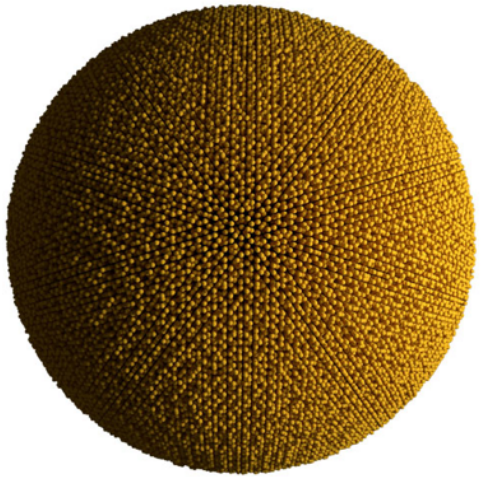


Shorter is Better

HOOMD-blue

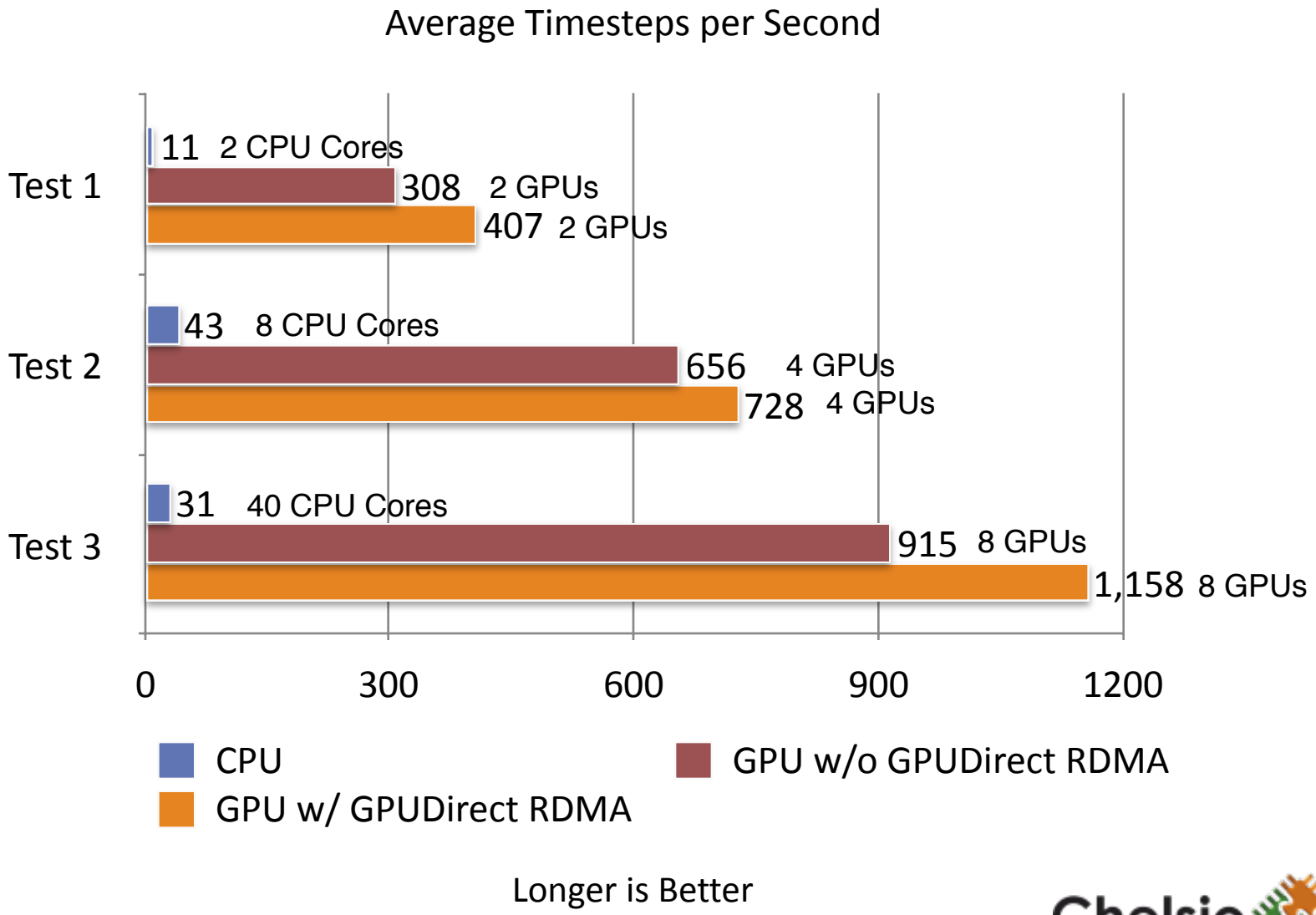
Quasicrystal Benchmark

- runs a system of particles with an oscillatory pair potential that forms a icosahedral quasicrystal
- This model is used in the research article: Engel M, et. al. (2015) Computational self-assembly of a one-component icosahedral quasicrystal, Nature materials 14(January), p. 109-116.



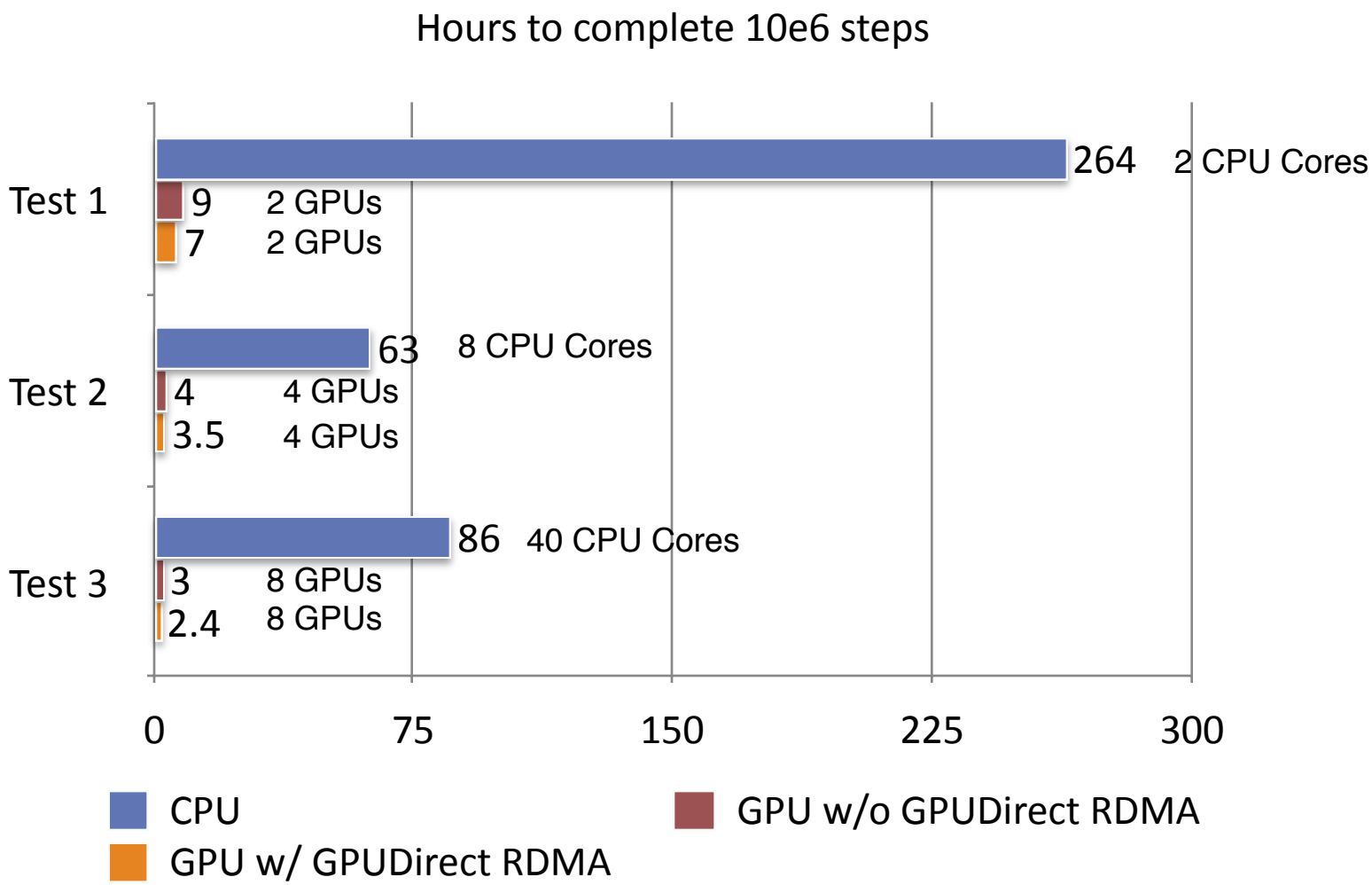
HOOMD-blue

Quasicrystal results



HOOMD-blue

Quasicrystal results



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.
- GPUDirect RDMA White Paper: http://www.chelsio.com/wp-content/uploads/resources/T5-40Gb-Linux-GPUDirect.pdf
- 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?

Chelsio Communications

Accelerate

Thank You

Chelsio
Communications
Accelerate