RDMA/IP “Mini BOF” STORM WG IETF-88

Background

Tom Talpey

Agenda

- Introduction/Background – 20 min
  - Tom Talpey
- iWARP – 15 min
  - Brian Hausauer
- RoCE – 15 min
  - Diego Crupnicoff
- Data Center Ethernet – 10 min
  - Pat Thaler
- Discussion – 1 hr
  - All

Goals

- Assess state of RDMA
- Interest in continuing IETF RDMA work
- Explore cross-standards-org liaison(s)
- Discussion of possible future WG activity

What is RDMA

1. Secure and efficient sharing and transfer of memory directly to/from network
2. Messaging paradigm for low-latency

- Protocols:
  - "iWARP" MPA/TCP|SCTP – DDP – RDMAP
    - Typically Ethernet 10-40Gb
  - InfiniBand (InfiniBand Trade Association (IBTA))
    - Specialized link layer 40-56Gb, moving higher
  - RoCE (also IBTA)
    - RDMA over Converged Ethernet (InfiniBand messages)
    - Datacenter Ethernet 10-40Gb
- All currently shipping from multiple vendors and supported by major operating systems

Previous IETF Work

- RDMA Consortium 2002-2003
  - External stds org submitted specs to IETF (2002-2003)
  - DDP, RDMAP, MPA, iSER/DA
  - Also: Verbs (RDMA pseudo-API) and SDP (Sockets Direct) not adopted by IETF
- RDDP 2002-2007
  - RFC4296 Architecture, RFC4297 Problem statement (2004-2005)
  - RFC5040 RDMAP, 5041 DDP, 5042 Security, 5043-5044 MPA/TCP, SCTP (2006)
- IPS 2001-2007
  - RFC5046 iSER, 5047 Datamover (2006)
- STORM 2009-present
  - RFC6581 MPA peer connect (2011)
  - RFC6580 RDDP Registries (2012)
  - TBD RDMAP extensions, iSER (active)
- NFSv4 (in perpetuity ☺)
  - RFC5532 NFS/RDMA problem statement (2008)
  - RFC5666-5667 NFS/RDMA protocol (2008)

Upper Layers using RDMA

- Storage
  - NFSv2/v3/v4
  - iSER
  - SMB3 (Microsoft)
  - SRP (SCSI RDMA Protocol) (ANSI T10)
- “High Performance Computing”
  - MPI
  - Financial
  - Scientific/HPC
- Virtualization
  - E.g. migration, backup/cloning
- Differing fabric use and requirements
  - Storage: send/receive/read/write: efficiency, IOPS
  - HPC: +atomics/immediate: latency
  - Others: +bulk transfer: bandwidth

Lower Layers Used by RDMA

- Ethernet
- Data Center Ethernet
  - DCB, PFC, QCN
- InfiniBand
- Other

RDMA Trends

- Hardware (NIC device) offload
- TCP/iWARP
  - Perceived device complexity
  - Routable, scalable on standard networks
- RoCE
  - Perceived device simplicity/efficiency, complexity in network
  - Not routable, help!
- Scaleout
  - Datacenter, cloud deployment
- Congestion management
- Workloads (goals)
  - Storage! (IOPS)
  - Low-latency messaging (scientific, clusters, etc)
  - Network shared memory (latency, signaling, active/active)
  - Bulk transfer (bandwidth)

Virtualization

- Increasing use of RDMA in virtualized environments
  - Storage access (small IOPS at low overhead)
  - Migration (memory-to-memory at high bandwidth and low overhead)
  - Storage management (drive cloning, transfer)
  - RDMA access directly from guest VMs
- Encapsulation typical
  - Implies IP addressing and endpoint management
  - Device virtualization (e.g. SRIOV)
- Standards/BCPs for RDMA encapsulation needed?
  - Protocol implications?

Other related work

- Verbs?
- Richer messaging interface?
- Encapsulation requirements and interface?
- Transport layer e.g. congestion/slowstart?
- Related external standards organizations
  - IBTA
  - ANSI T10
  - IEEE
  - Other
- Related Working Groups
  - NFSv4
  - NVO3?
  - TSV/TCPM