Emerging Ethernet standards & their impact on Storage

Anupama B N
NetApp
Agenda

- Ethernet Technology Landscape
- Ethernet Standards and Technology
- Connector Standards and types
- RDMA (RoCE and iWARP)
- RoCE vs iWARP
- RoCE over long distance
- References
Ethernet Technology Landscape

- FCoE
- RDMA
  - iWARP
  - RoCE
- SDN
  - Extension into the VM environment vSphere/OpenVswitch/Nexus
  - Provisioning and orchestration tools, focus on Overlays – VXLAN, NVGRE, GENEVE, WAN
- NVMf
  - NVMe over Fabrics
- iSCSI
- iSER
Data Center

- Bigger – 1km cable runs common
- Fill as you go, leave in place
- Manage via API (remote), ports set up on demand via API
- Leaf/Spine Clos (vs. Tree)
High Speed Interconnects

Low latency + High speed transports

Storage Class Memory (SCM) as Cache

NVMe over Fabrics

Persistent Memory (PMEM) in Server

Hybrid SAS

All-Flash NVMe/SAS

Flash At All Tiers QLC-SSD/SCM
Ethernet Technology and Standards

<table>
<thead>
<tr>
<th>Technology</th>
<th>Standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>CY16</td>
<td>50G PAM4 SERDES</td>
</tr>
<tr>
<td>CY17</td>
<td>Silicon Photonics</td>
</tr>
<tr>
<td>CY18</td>
<td>50/100/200/400G Early NIC and Switch ASICs</td>
</tr>
<tr>
<td>CY19</td>
<td>100G PAM4 SERDES</td>
</tr>
<tr>
<td>CY20</td>
<td>100/200/400/800G Early NIC and Switch ASICs</td>
</tr>
<tr>
<td>CY21</td>
<td>Silicon Photonics in NIC/Switch ASICs</td>
</tr>
<tr>
<td>CY22</td>
<td></td>
</tr>
</tbody>
</table>

- **IEEE 802.3bq**: 25/40GBase-T
- **IEEE 802.3bs**: 200G/400G (4/8 lane)
- **IEEE 802.3cc**: 25G SMF
- **IEEE 802.3by**: 25G
- **IEEE 802.1AS**: time sensitive Ethernet
- **IEEE 802.3ck**: 100/200/400G
- **IEEE 802.3cz**: 800G
- **QSFP-DD**
- **CWDM4 PSM4 MSA**
- **OSPF MSA**
- **IEEE 802.3cd**: 50/100/200G
802.3bs/cd Signaling

NRZ to PAM4

- PAM4
- 1, 2, 4, and 8

Source: Mellanox blog, neophotonics
Connector Types

MSA Mainstream: New Double Density Connectors

Source: SFP-DD consortium, QSFP-DD consortium
What is RDMA

- RDMA – Remote Direct Memory Access
- Benefits
  - Very low latency, very high throughput, ≈ zero CPU
  - Bypasses traditional network stacks (TCP/IP)
  - Provides a Fibre Channel-equivalent solution at a lower cost
- Three hardware technologies
  - RoCE
  - iWARP
  - Infiniband
- Traditional protocols (SMB, NFS, iSCSI) can operate over RDMA
Ethernet RDMA Stack

RDMA Application / ULP

RDMA Software Stack

IB Transport Protocol
IB Network Layer
IB Link Layer

IB Transport Protocol
IB Network Layer
Ethernet Link Layer

IB Transport Protocol
Ethernet Link Layer

iWARP Protocol
TCP
IP
Ethernet Link Layer

InfiniBand
RoCE v1
RoCE v2
iWARP

InfiniBand Management
Ethernet / IP Management
Ethernet / IP Management
Ethernet / IP Management
- Delivers RDMA on top of Pervasive TCP/IP
- Runs over all Ethernet Infrastructure
- TCP provides Flow control and Congestion Management
- Highly routable and scalable Implementation
- Extensions eliminate TCP/IP stack process, mem copies and application contexts switches.
- iWARP addresses n/w bottlenecks of high speed Ethernet and provides high-throughput and low-latency with low-CPU utilization for data communication.
RDMA over Converged Enhanced Ethernet (RoCE)

Same RDMA, different L2 transport

- Remote Direct Memory Access
  - Accelerates data exchange between servers
- Bypass CPU & typical network stack
- Reduced latency
- Converged Enhanced Ethernet
  - Priority Flow Control
  - Enhanced Transmission Selection
  - Lossless Ethernet fabric
RoCE(V2)

- Well known on InfiniBand
- Works well on a lossless network
- Lower latency than alternative Transport protocols (TCP)
- Significantly lower overhead when offloaded to adapter

..BUT

- Ethernet is not lossless by design
- PFC is required to achieve lossless Ethernet fabric
- PFC (Part of DCB) has a high configuration and management overhead – VLANs, Priorities
- PFC is Layer 2 only
## RDMA Pros and Cons

<table>
<thead>
<tr>
<th>Transport</th>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
</table>
| Non-RDMA Ethernet | • TCP/IP-based protocol  
• Works with any Ethernet switch  
• Wide variety of vendors and models  
• Support for in-box NIC teaming  | • High CPU Utilization under load  
• High latency  |
| iWARP           | • TCP/IP-based protocol  
• Works with any Ethernet switch  
• RDMA traffic routable  
• Offers up to 100 Gbps per NIC port today*  | • Requires enabling firewall rules  |
| RoCE            | • Ethernet-based protocol  
• Works with Ethernet switches  
• Offers up to 100 Gbps per NIC port today*  
• Routable with RoCEv2  | • Requires DCB switch with Priority Flow Control (PFC)  |
| InfiniBand      | • Switches typically less expensive per port*  
• Switches offer high speed Ethernet uplinks  
• Commonly used in HPC environments  
• Offers up to 54Gbps per NIC port today*  | • Not an Ethernet-based protocol  
• RDMA traffic not routable via IP infrastructure  
• Requires InfiniBand switches  
• Requires a subnet manager (typically on the switch)  |
## RoCE vs IWARP differences

<table>
<thead>
<tr>
<th></th>
<th>RoCE</th>
<th>IWARP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Underlying Network</td>
<td>UDP</td>
<td>TCP</td>
</tr>
<tr>
<td>Congestion Management</td>
<td>Rely on DCB</td>
<td>TCP handles with flow control</td>
</tr>
<tr>
<td>Adapter Offload</td>
<td>Full DMA</td>
<td>Full DMA w/TCP/IP</td>
</tr>
<tr>
<td>Routability</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Cost</td>
<td>Need DCB enabled Switch Infra</td>
<td>Depends on the deployment, no requirement of Switch conf</td>
</tr>
</tbody>
</table>
Ethernet RDMA NIC Implementation

Application

NVMe Native

NFS, CIFS, SAN

NVMeoFabric

Ethernet

FC

RDMA

IB

RoCE

iWARP

Ethernet/IP

Storage Compute

NVMe Target

NVMe Target

NVMe Initiator

NVMeoF

RDMA-ROCE

NVMe SSDs

NVMe SSDs

Converged Shelves

Integrated

Storage Drives

Switches

Switches

2019 Storage Developer Conference India © All Rights Reserved.
PFC – Priority Flow Control

- By nature Ethernet is a lossy network
- Ethernet provides flow control mechanism which makes it lossless – 2 options:
  - Applied FC over the whole port (Priority Flow Control - 802.3x)
  - Applied FC over specific priority (Priority Flow Control - 802.1Qbb)
- PFC negotiation between switch-host can be done by DCB (Data Center Bridging)
  - Using Data Center Bridging Exchange (DCBX) negotiation
  - End points (switch & host) exchange information about their capabilities
  - If PFC is supported, it will be used
  - If PFC is not supported, Global FC will be used
  - If DCBX is not supported or the PFC capability is not supported, manual configuration is required
- Routers rebuild the layer 2 header
  - Among it the routers rebuild the PCP filed using a DSCP to PCP mapping
PFC contd..

![Diagram of PFC Priorities and Receive Buffers]

RoCE for Long Distance

- Minimize the recovery impact from lost packets
  - Congestion, faulty networking components, alpha particles, etc.
- Congestion
  Can not use normal congestion control
  - PFC and ECN latency is too great because of distance
  - Solution options - Packet Pacing (NIC and Application) - Prioritize flows (QPs) through local networks (NIC and Switches)
    - ECN, PFC, other QOS
- Enhance recovery for lost packet
  - Resilient RoCE
  - Create a lot of small flows (Application)
  - Minimize the latency of retry
Routable RoCE

- Routable RoCE requires a higher level congestion mechanism
  - ECN – Explicit Congestion Notification
- ECN can slow down traffic to prevent congestion
- ECN configuration overhead is lower than PFC, simple and easy

Source: Mellanox web
Resilient RoCE

- Resilient RoCE can cope with packet loss and Out of Order packets
- ECN is suggested but not required
- Out of Order packets are held in buffer to fill the gaps. Re-ordered packets are then written to memory
- Missing packets are requested from the sender

So..

- No loss – everything is fast
- Some loss – slows down, but stays in working order
- Still significantly better than TCP/IP
References

- https://community.mellanox.com/s/article/understanding-qos-configuration-for-roce
CONCLUSION

- High-Speed Ethernet is the new back-bone which could replace FC
- Different media/storage via network require reliable connectivity with High throughput and low latency.
- High Availability and Disaster Recovery solutions are On-Demand with high data re-locational capabilities across geographies.
- Transports for NVMe over Fabric with Ethernet is gaining momentum.
Questions ?