SDC 19 SNIA INDIA

May 23-24, 2019
Bangalore, India

STORAGE DEVELOPER CONFERENCE

Emerging Ethernet standards & their impact on Storage

Anupama B N
NetApp

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

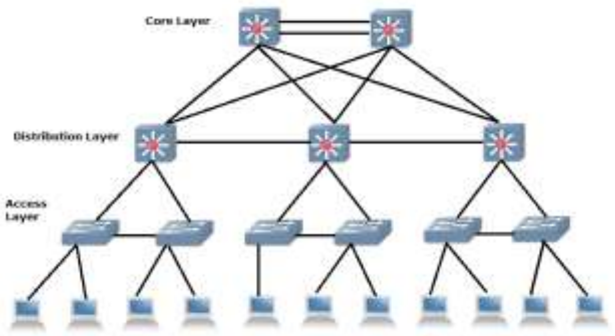
- FCoE
- RDMA
  - iWARP
  - RoCE
- SDN
  - Extension **into** the VM environment vSphere/OpenVswitch/Nexus
  - Provisioning and orchestration tools, focus on **Overlays** – VXLAN, NVGRE, GENEVE, **WAN**
- NVMf
  - NVMe over Fabrics
- iSCSI
- iSER

SDC 19 SNIA INDIA

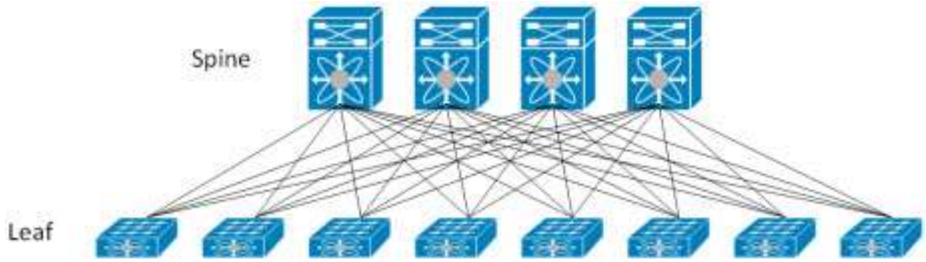
NetApp

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)



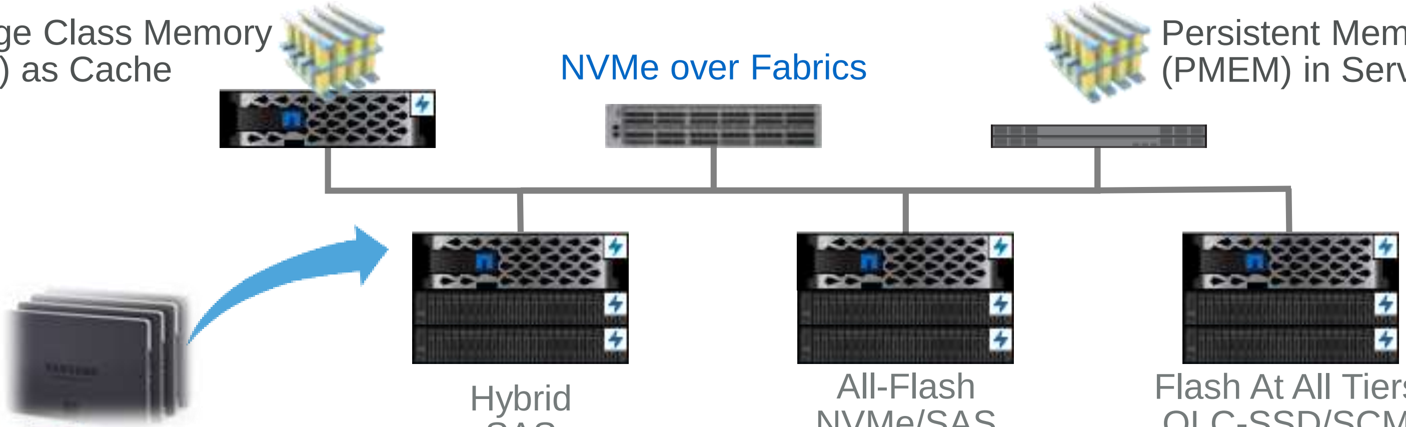
Old



New

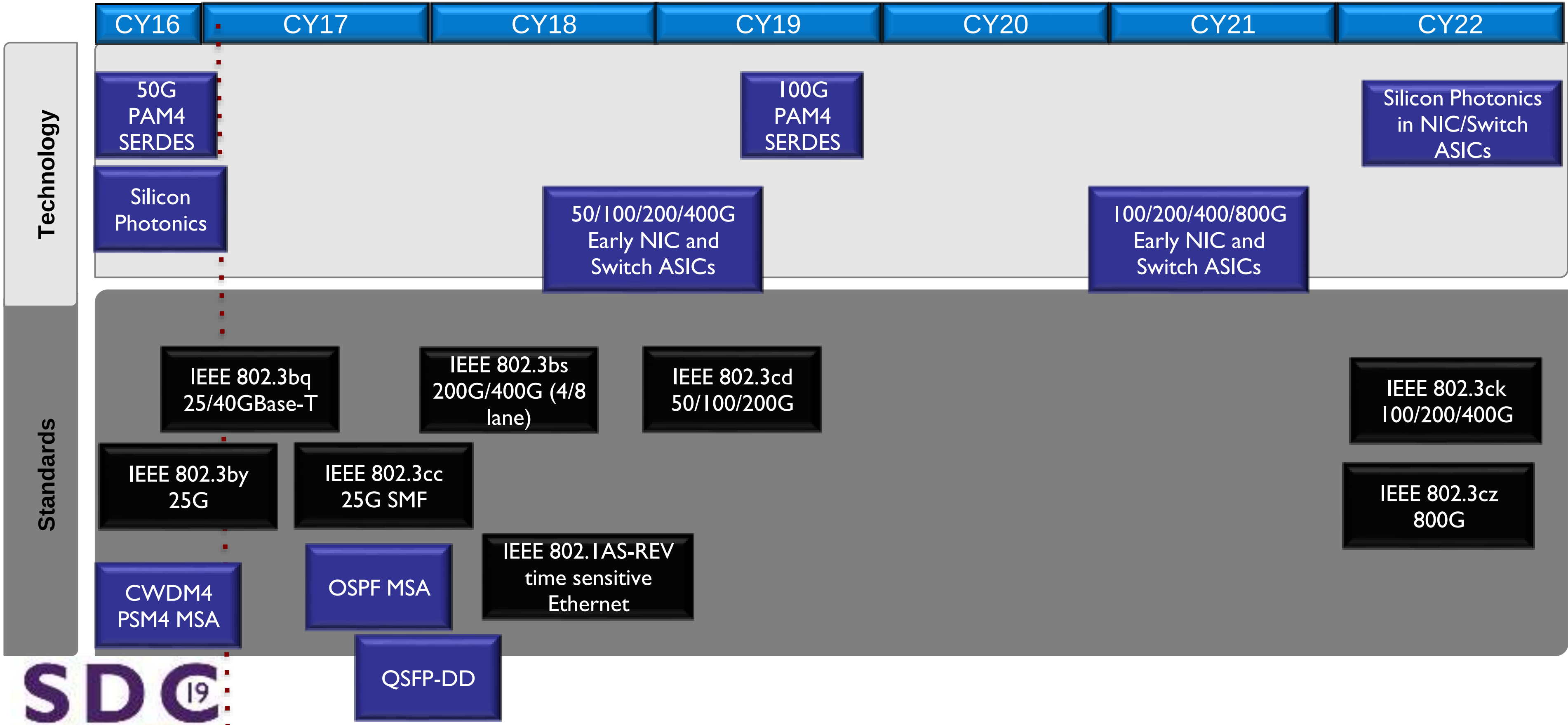
High Speed Interconnects

Low latency access + High speed transports



Ethernet Technology and Standards

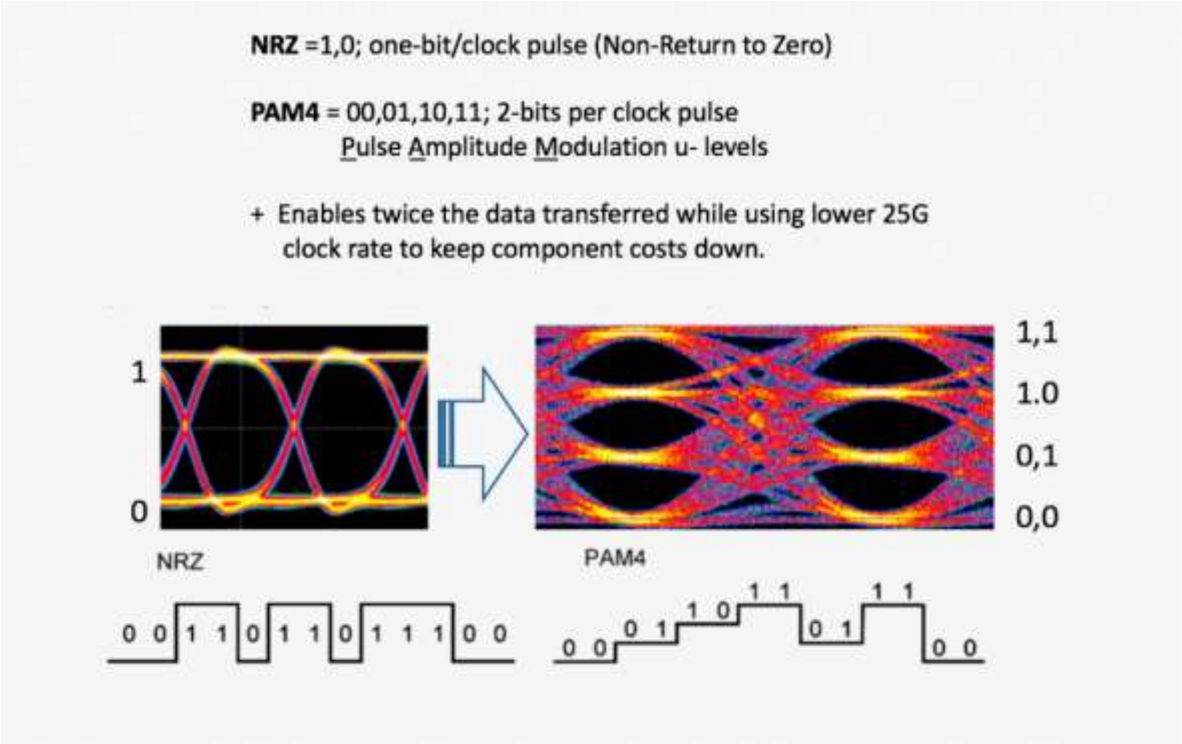
Technology/MSA [Purple]
IEEE Standard [Black]



802.3bs/cd Signaling

NRZ to PAM4

- **PAM4**
- **1, 2, 4, and 8**



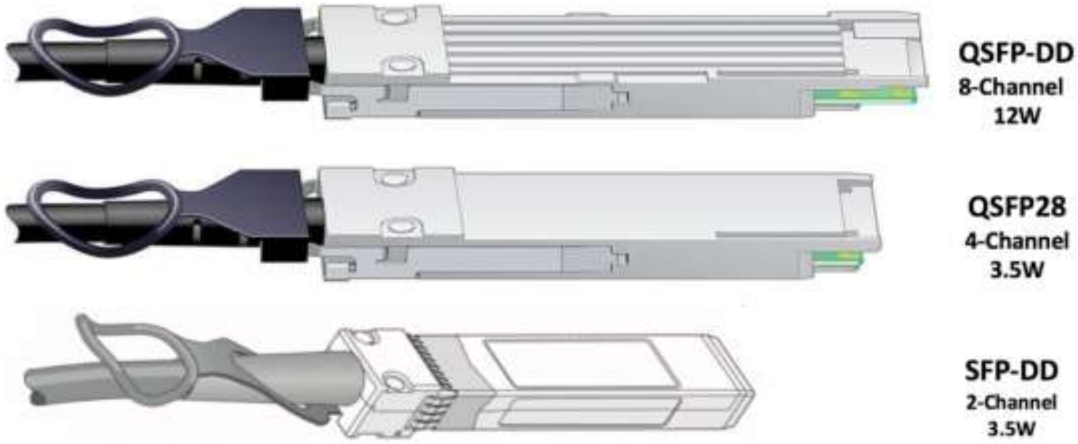
Source: Mellanox blog, neophotonics

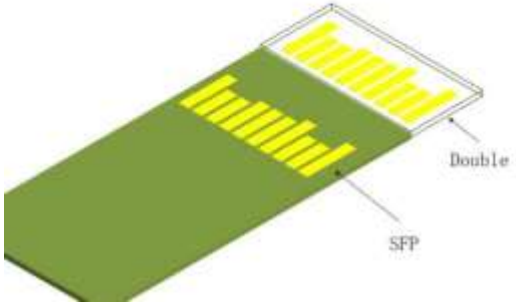
SDC 19 SNIA INDIA

NetApp

Connector Types

MSA Mainstream: New Double Density Connectors





Source SFP-DD consortium, QSFP-DD consortium

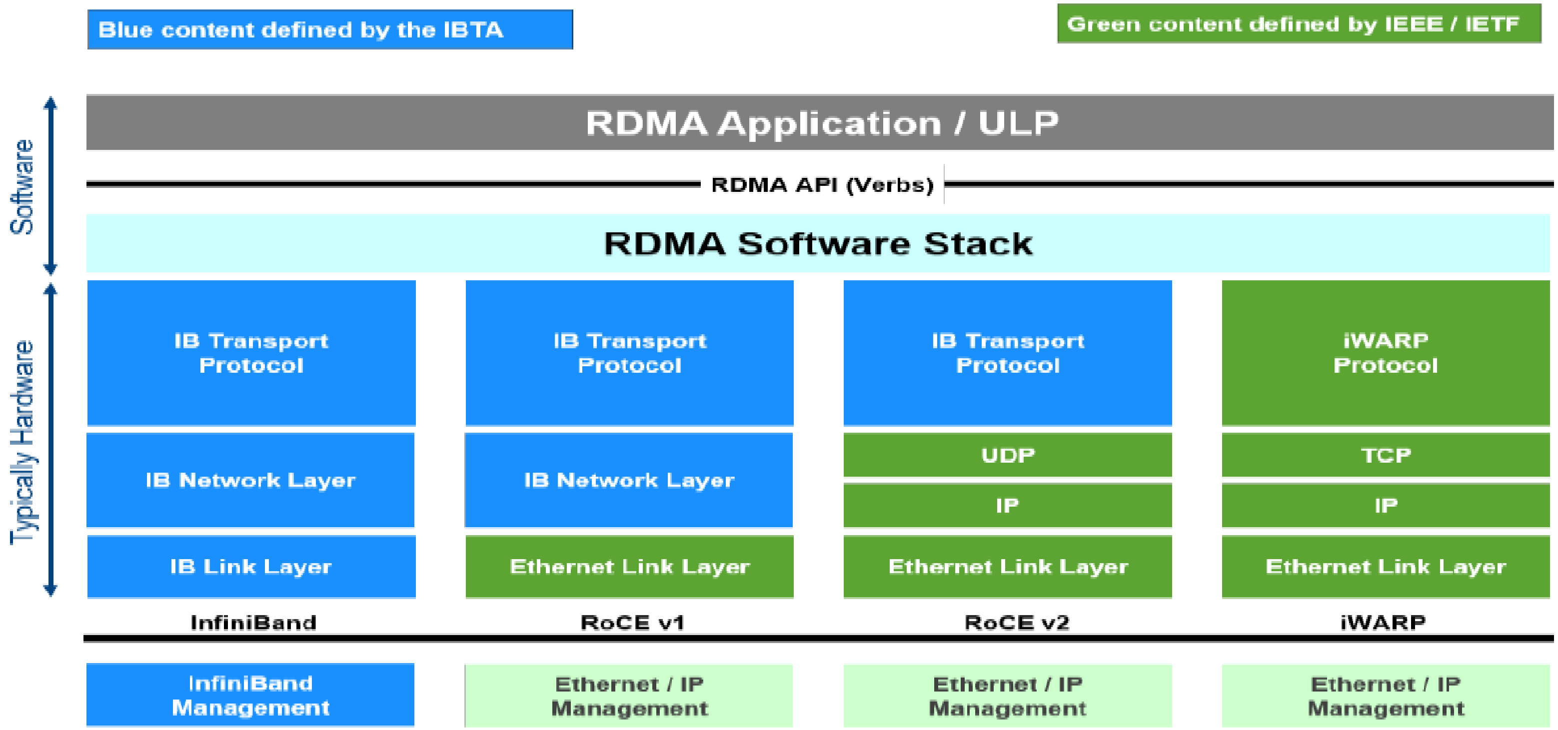
SDC 19 SNIA INDIA

NetApp

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



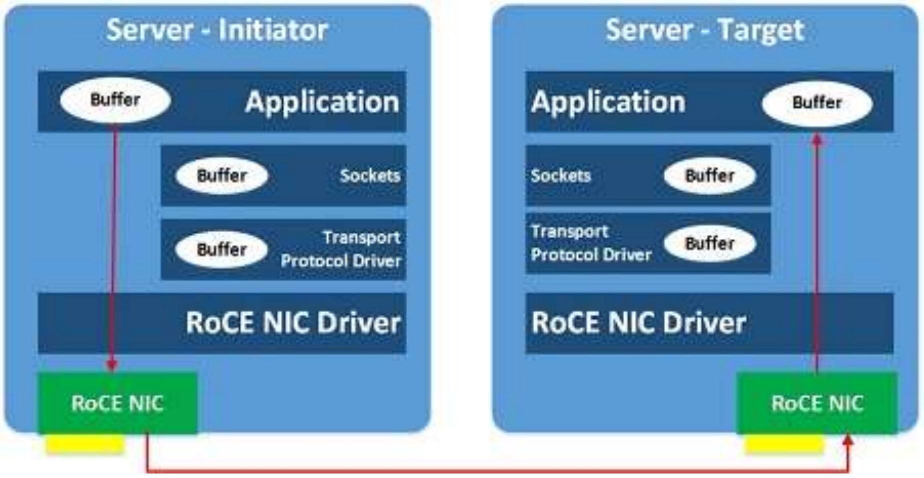
iWARP

- 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



SDC 19 SNIA INDIA

NetApp

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

SDC 19 SNIA INDIA

NetApp

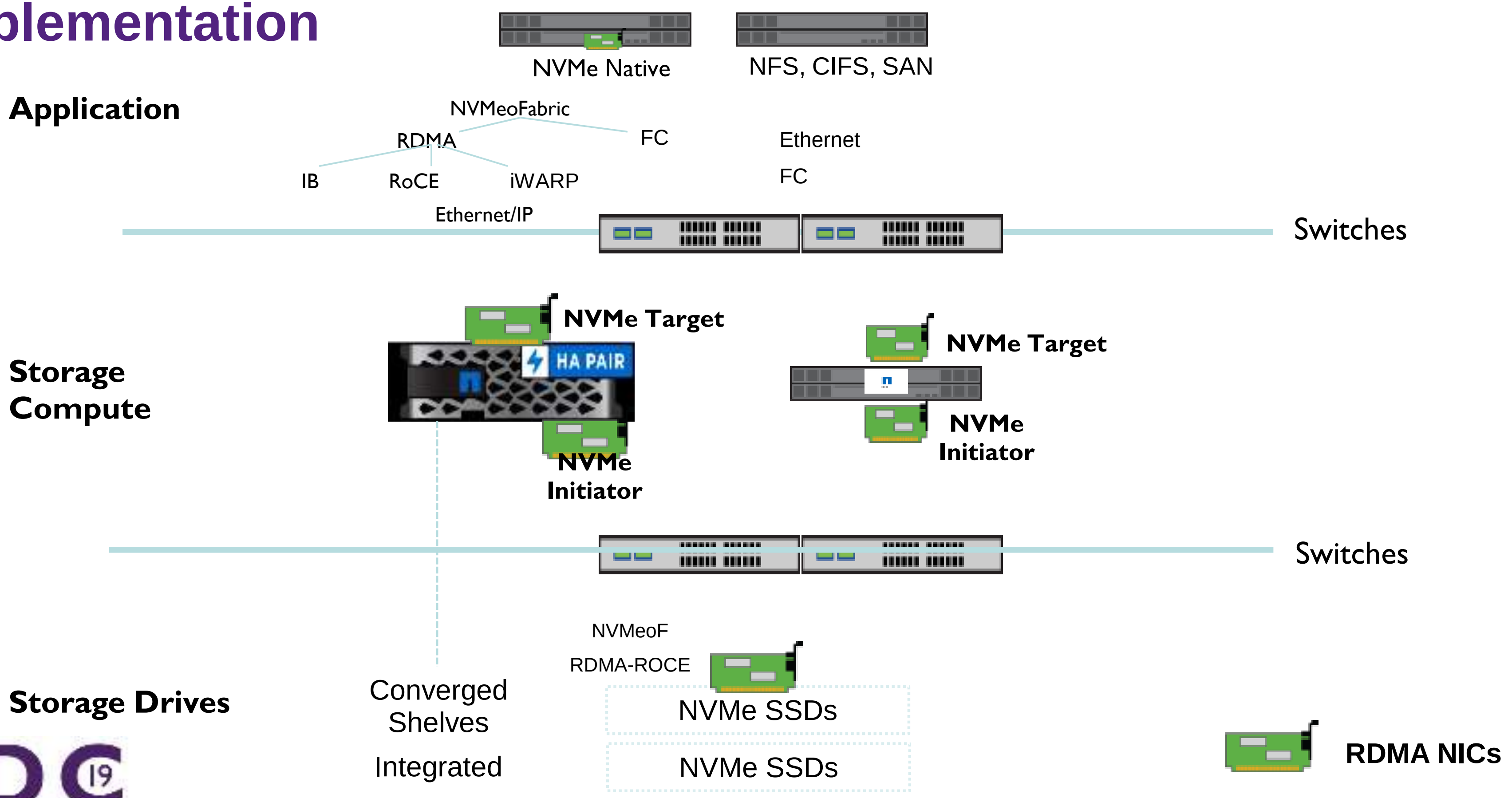
RDMA Pros and Cons

Transport	Pros		Cons
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	Low CPU Utilization under load Low latency	• 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

	RoCE	iWARP
Underlying Network	UDP	TCP
Congestion Management	Rely on DCB	TCP handles with flow control
Adapter Offload	Full DMA	Full DMA w/TCP/IP
Routability	Yes	Yes
Cost	Need DCB enabled Switch Infra	Depends on the deployment , no requirement of Switch conf

Ethernet RDMA NIC Implementation



16

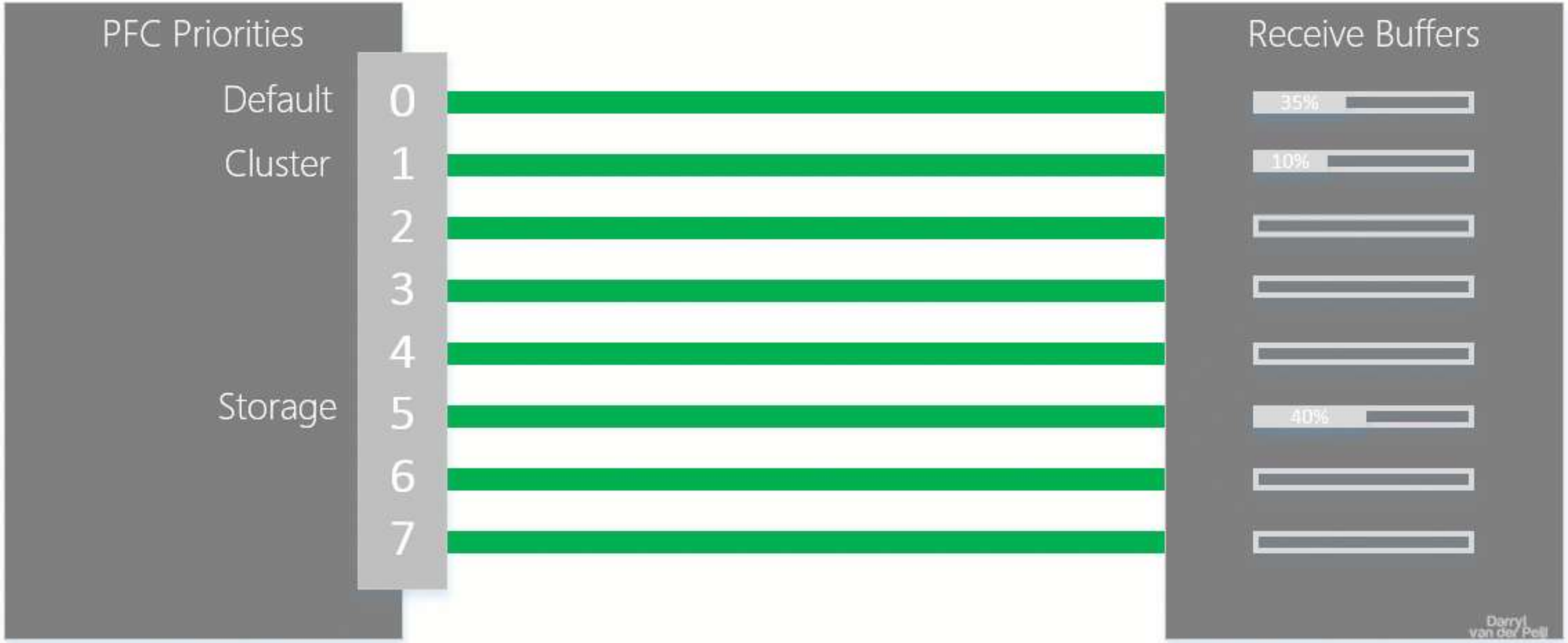
PFC –Priority Flow Control

- 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

SDC 19 SNIA INDIA

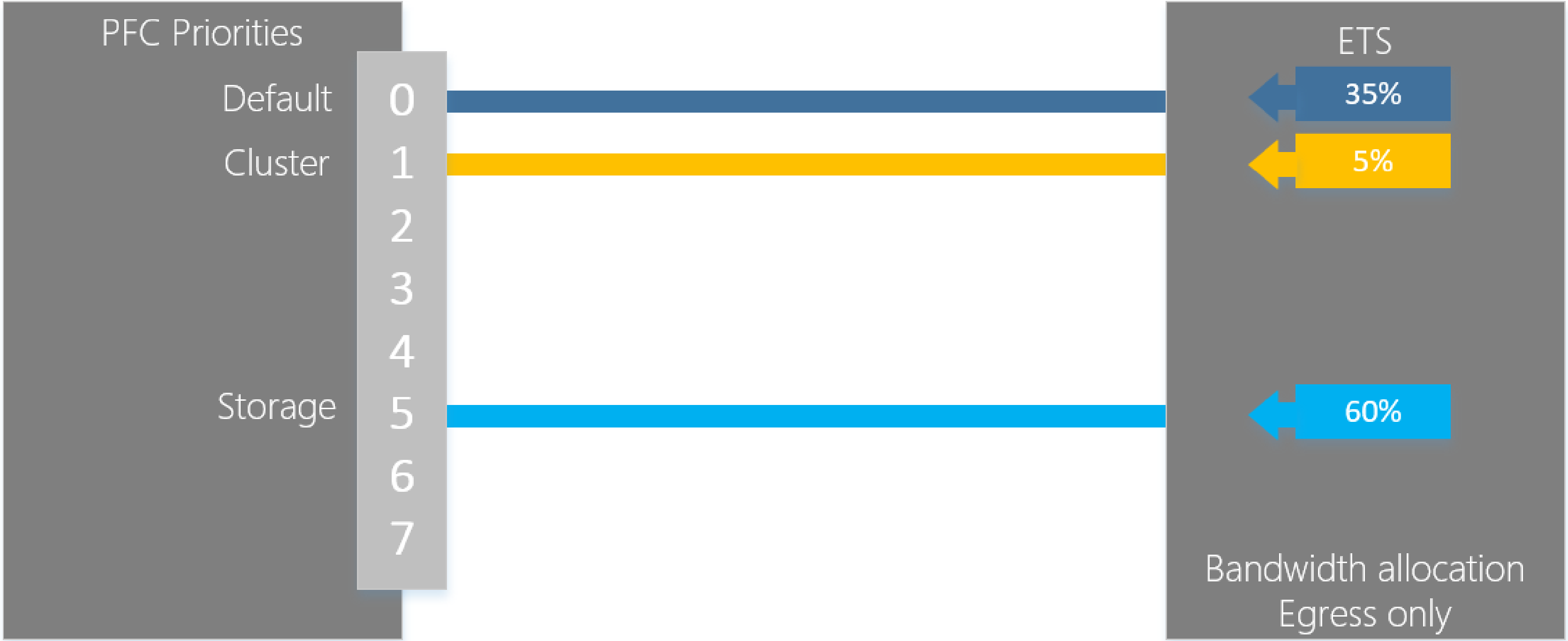
NetApp

PFC contd..



https://www.darrylvanderpeijl.com/wp-content/uploads/part1_PFC.gif

PFC-ETS


PFC Priorities
Default
Cluster
Storage
0
1
2
3
4
5
6
7
ETS
35%
5%
60%
Bandwidth allocation
Egress only

RoCE for Long Distance

- Minimize the recovery impact from lost packets
  - Congestion, faulty networking components, alpha particles, etc.
- Congestion

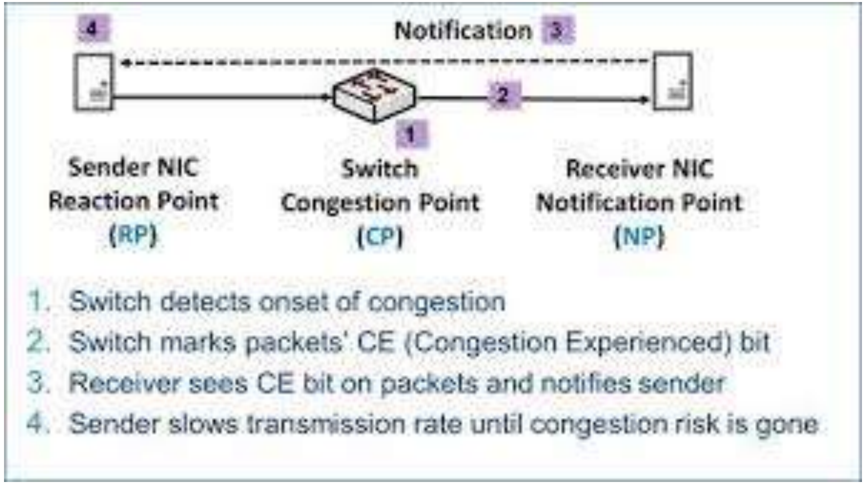
Can not use normal congestion control

- 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

SDC 19 SNIA INDIA

NetApp

References

- http://conferences.sigcomm.org/sigcomm/2015/pdf/papers/p523.pdf
- https://community.mellanox.com/s/article/understanding-qos-configuration-for-roce
- https://www.snia.org/sites/default/files/ESF/RoCE-vs.-iWARP-Final.pdf
- http://files.gpfsug.org/presentations/2017/Manchester/04_Mellanox.pdf

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 .

Ethernet NIC

NIC ASICs [Red]
SOC/Netwk processor ASICs [Blue]

CY17 | CY18 | CY19 | CY20 | CY21 | CY22

Broadcom
- Cumulus 2x25G gen3
- Stratus 1x100G gen3
- Thor 1x200G gen4
- Stingray 1x100G gen3
- Stingray 2 1x200G gen4

Mellanox
- ConnectX-5 2x100G gen4
- ConnectX-6 2x200G gen4
- Bluefield 2x100G gen4 x32
- Bluefield 2Lx 2x100G gen4 x16

Chelsio
- T6 2x100G gen3
- T7 2x100G gen4 RoCE + iWARP
- T8 2x200G gen4 RoCE + iWARP

Intel
- Columbiaville 2 x 100G gen4 RoCE + iWARP
- Mt. Stellar 2x200G 2x gen4 x16

Cavium
- Arrowhead 2/4x25G gen3
- Elbrus (E5) 2x100G (50G serdes) 1x200G (50G serdes) gen4 RoCE + iWARP
- F1 2x200G (50G serdes) 1x100G (25G serdes) gen4 RoCE + iWARP
- FastLinQ
- Liquid I/O III 1x100G gen4

NetApp

Questions ?