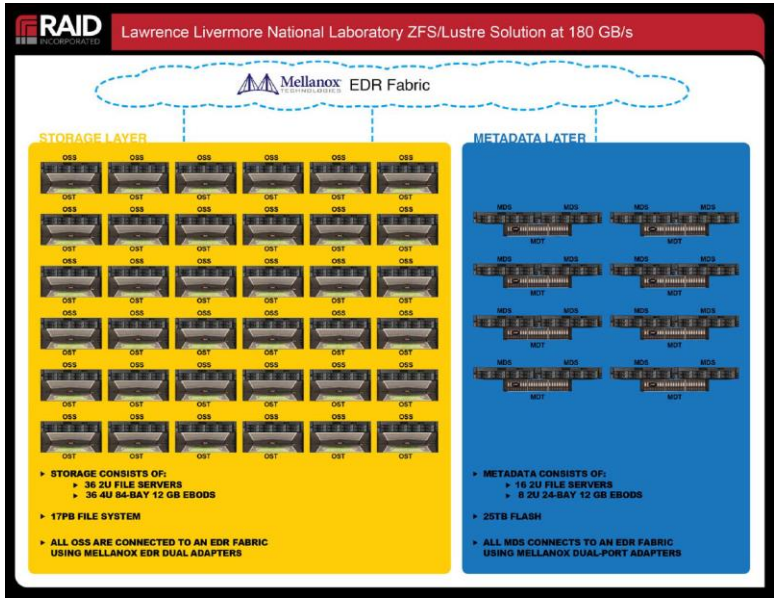
InfiniBand Networked Flash Storage

Superior Performance, Efficiency and Scalability

Motti Beck – Director Enterprise Market Development, Mellanox Technologies

17PB File System with Sustainable180GB/s

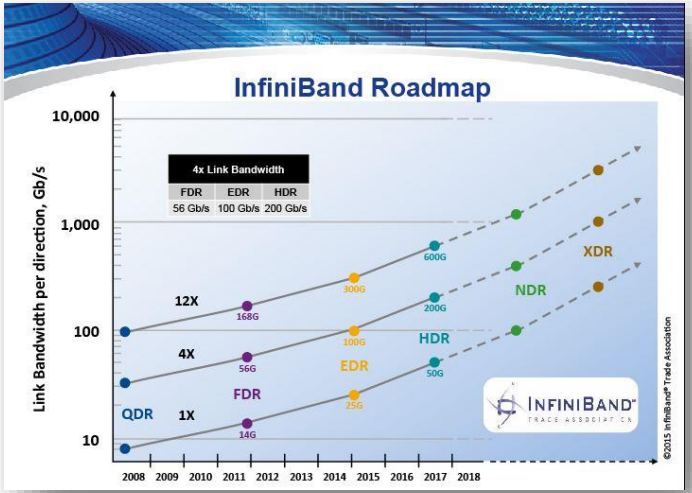
- The parallel file system will run Lustre 2.8 with ZFS OSDs and multiple metadata servers.
- The Lustre file system contains 36 OSS nodes, with each node capable of 5GB/s of sustained data performance, and 16 metadata servers with **25TB of SSD storage** capacity.
- The solution is anchored by enterprise 4U 84 bay 12Gb SAS JBODs, LSI/Avago 12Gb SAS adapters, **Mellanox EDR IB,** HGST 12Gb Enterprise SAS HDDs, and Intel server technologies.
- The file system incorporates 6 scalable storage units each containing six Lustre OSS and six 4U-84Bay **JBODs with 480 8TB SAS drives.** The solution will be employing ZFS on Linux with raidz2 data parity protection. **Resiliency is provided by multipath and HA failover connectivity,** intended to eliminate single points of failure.

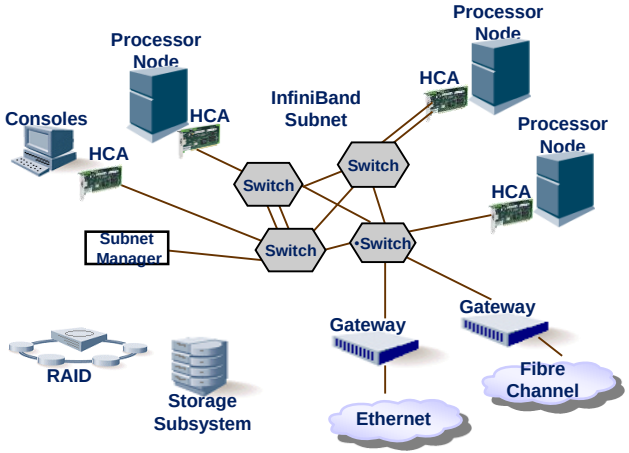


Source: Raid Inc. May 31st, 2016

Why InfiniBand in Storage

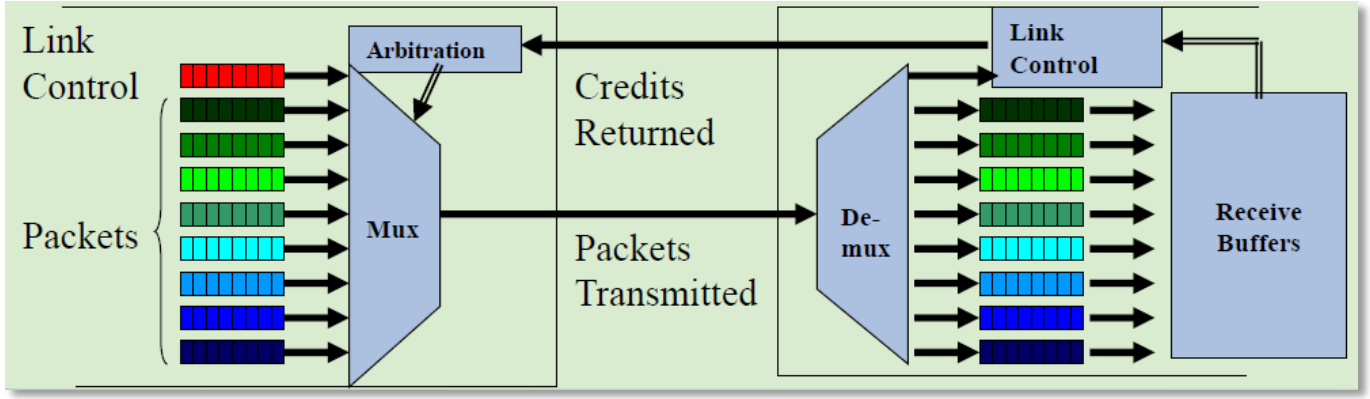
- Industry standard defined by the InfiniBand Trade Association (IBTA)
- Defines System Area Network architecture
  - Comprehensive specification - from physical to applications
- Simplicity drives higher efficiency and resiliency
  - Reliable, lossless, self-managed fabric
  - Transport offload - Remote Direct Memory Access (RDMA)
  - Centralized fabric management – Subnet Manger (SM)





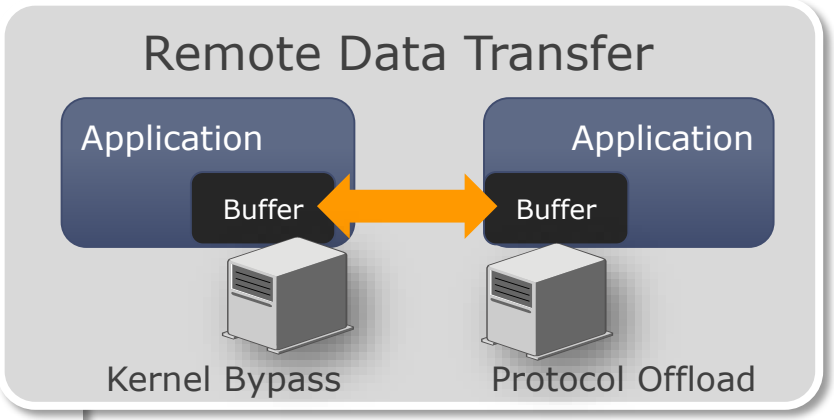
Reliable, Lossless, Self-Managed Fabric

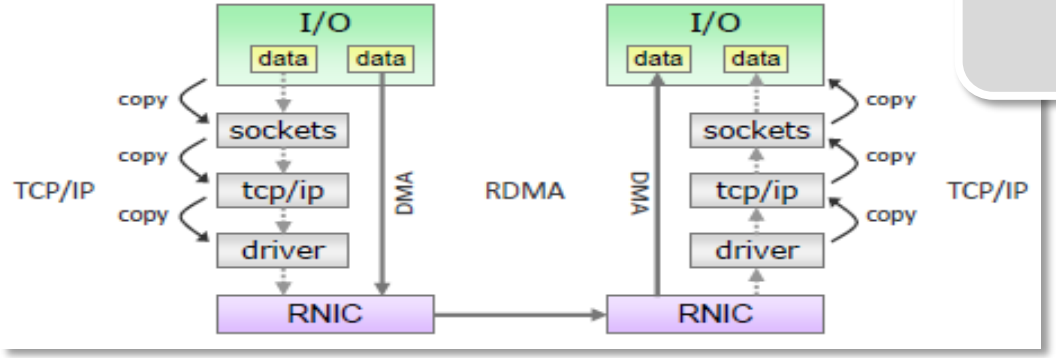
- End to end flow control
  - Credit based flow control for each link
- End to end Congestion management
- Automatic path migration



Remote Direct Memory Access RDMA

- Transport offload
- Kernel bypass

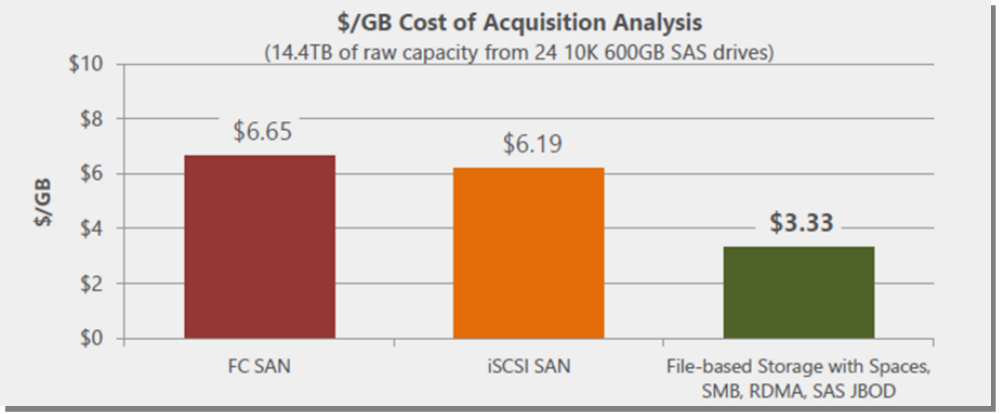


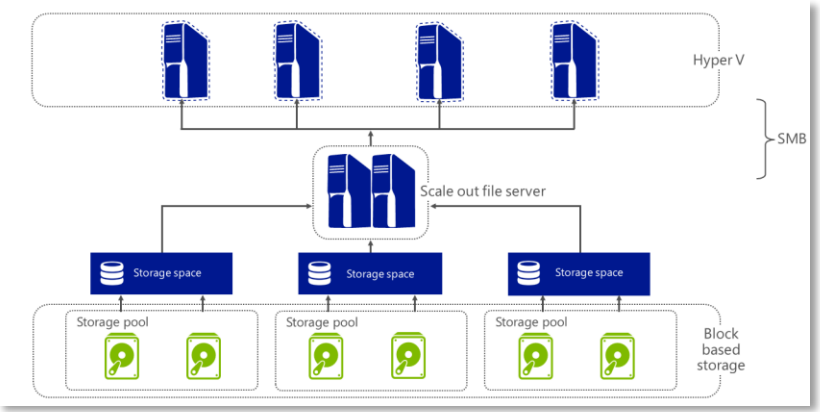


InfiniBand Cuts SAN Cost by 50%

- Delivers SAN-like functionality from the Windows Stack
  - Using SMB Direct (SMB 3.0 over RDMA)

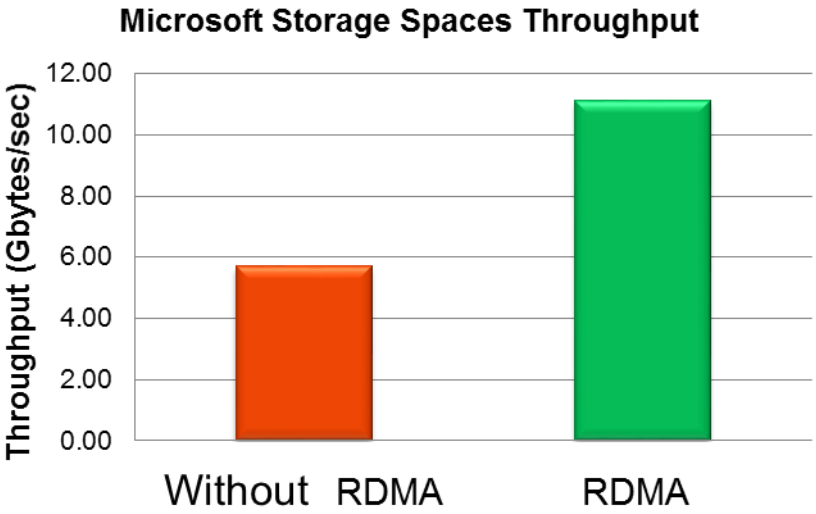
- Utilize inexpensive, industry-standard, commodity hardware
  - Eliminate the cost of proprietary hardware and software from SAN solutions

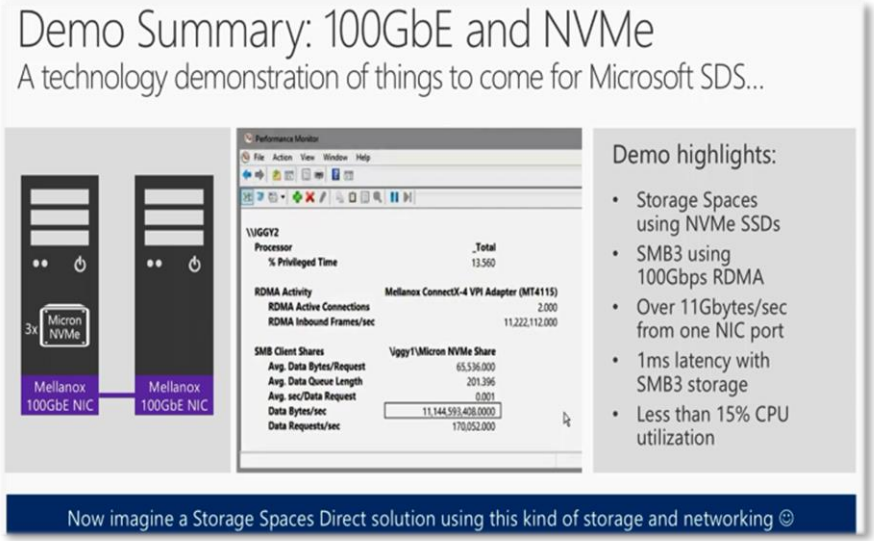




Source: Microsoft

RDMA Delivers Higher Cloud Efficiency

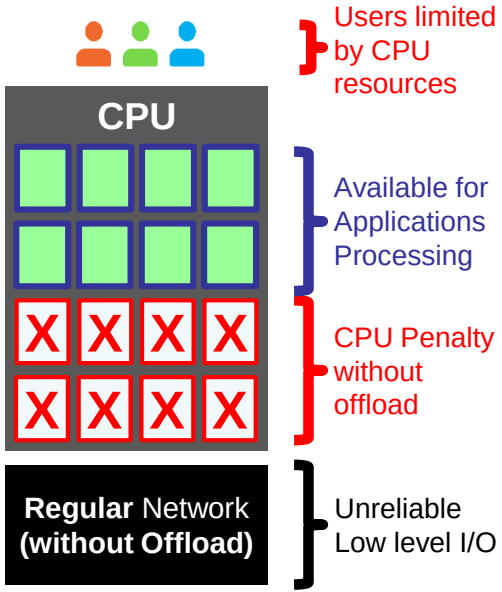


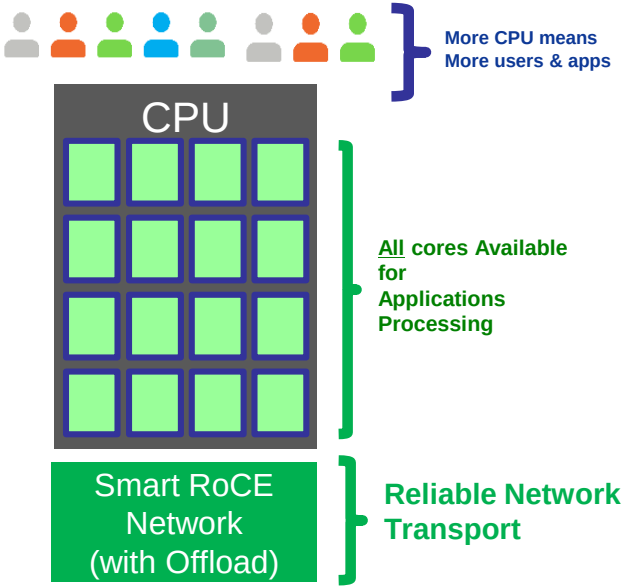


- 2X better performance with RDMA
  - 2X higher bandwidth & 2X better CPU efficiency
- RDMA achieves full Flash storage bandwidth
  - Remote storage without compromises

RDMA Frees Up CPU for Application Processing

54Gb/s → **90Gb/s**



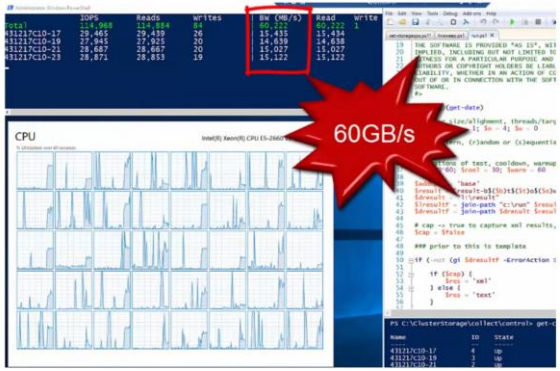


- Half of CPU capacity consumed moving data
  - Even though achieving only half throughput
- Cores unavailable for application processing

- All CPU available to run applications
  - Better efficiency = more users
- Smart network delivers better TCO

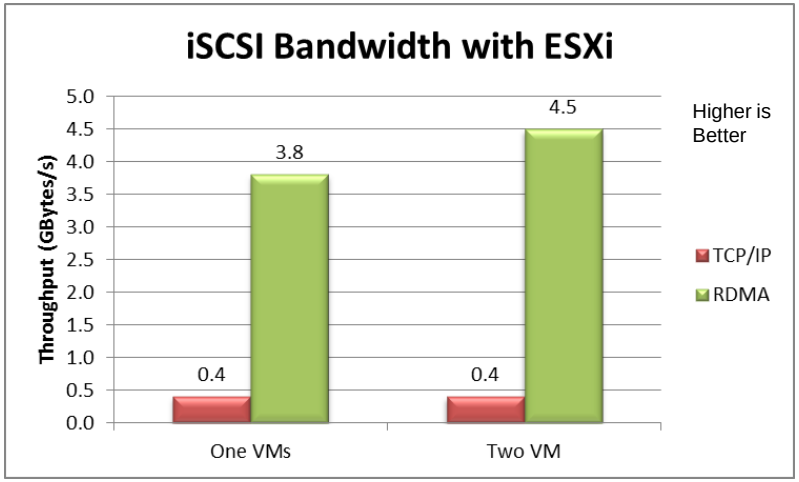
RDMA Removes Storage Bottlenecks in Enterprise

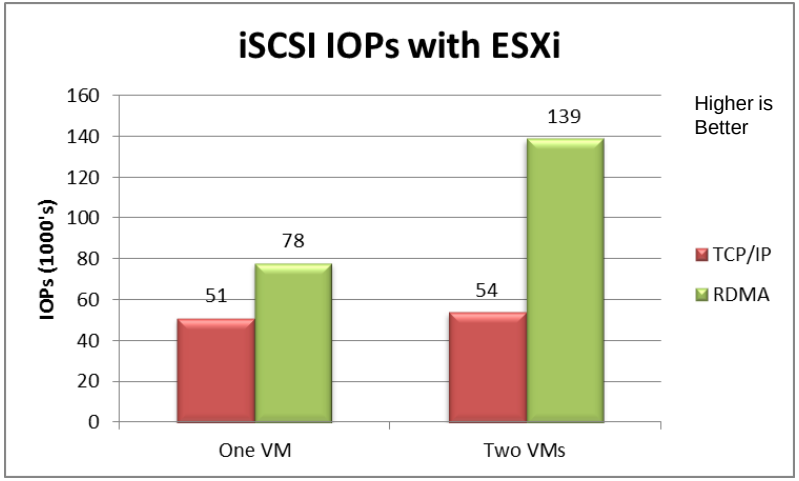
- Stellar Storage Performance with 100Gb/s
- Microsoft labs benchmarked Storage Spaces @ 100Gb/s
  - 4 Node Cluster
  - NVMe Flash & ConnectX-4 100GbE Adapters
- Storage Spaces Direct over RoCE
- Achieved 480Gb/s (60GB/s) throughput
  - Transmit entire content of Wikipedia in 5 seconds
- Storage Spaces Direct will be part of Azure Stack
  - Private cloud package for 2017



https://blogs.technet.microsoft.com/filecab/2016/05/11/s2dthroughputtp5/

Storage Acceleration Running iSER*





* iSCSI over RDMA

- RDMA Superior Across the Board
  - Throughput & IOP's
  - Efficiency & CPU Utilization
  - Scalability

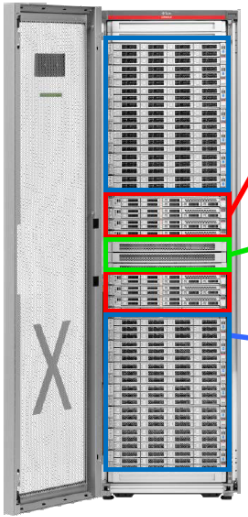
Test Setup: ESXi 5.0, 2 VMs, 2 LUNS per VM

*iSCSI Extensions for RDMA

10x Bandwidth Performance Advantage vs TCP/IP
2.5x IOPS Performance With iSER Initiator

InfiniBand Enables Most Cost Effective Database Storage

Exadata X5-2 Product Components



- **Scale-Out Database Servers**
  - **Two 18-core x86 Processors (36 cores)**
  - Oracle Linux 6
  - Oracle Database Enterprise Edition
  - Oracle VM (optional)
  - Oracle Database options (optional)
- **Fastest Internal Fabric**
  - 40 Gb/s InfiniBand
  - Ethernet External Connectivity
- **Scale-Out Intelligent Storage**
  - **High-Capacity Storage Server**
  - **Extreme Flash Storage Server**
  - **Exadata Storage Server Software**

X5-2 Database Server

36 cores per server
256 – 768 GB DRAM

High-Capacity Storage Server

Extreme Flash Storage Server

Source: Oracle

Dominant in Storage and Database Platforms

Leading in Performance, Efficiency and Scalability