



“Comparing SMB Direct 3.0 performance over RoCE, InfiniBand and Ethernet”

Anand Rangaswamy | Storage Developer Conference

September 2014

- Leading provider of high-throughput, low-latency server and storage interconnect
 - FDR 56Gb/s InfiniBand
 - 10/40/56GbE Ethernet Switches and Adapters

- Company headquarters:
 - Yokneam, Israel; Sunnyvale, California
 - ~1,432 employees* worldwide



* As of March 2014

Leadership in Strategic Markets with an End to End Portfolio



Big Data



DB/Enterprise



Cloud



Storage



Web 2.0



End-to-End - Cloud, Enterprise, and Storage Optimized - InfiniBand and Ethernet Portfolio

ICs



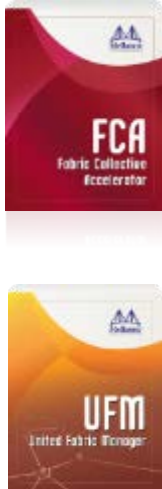
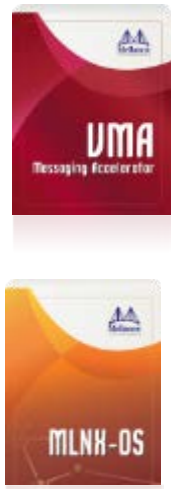
Adapter Cards



Switches/Gateways



Host/Fabric Software



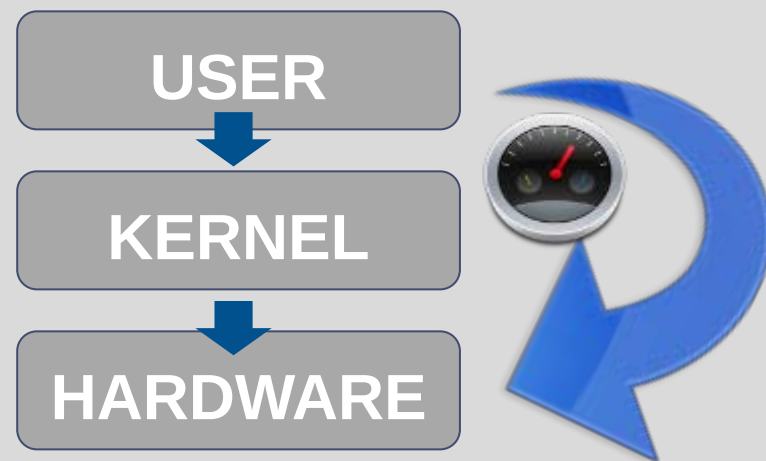
Metro / WAN



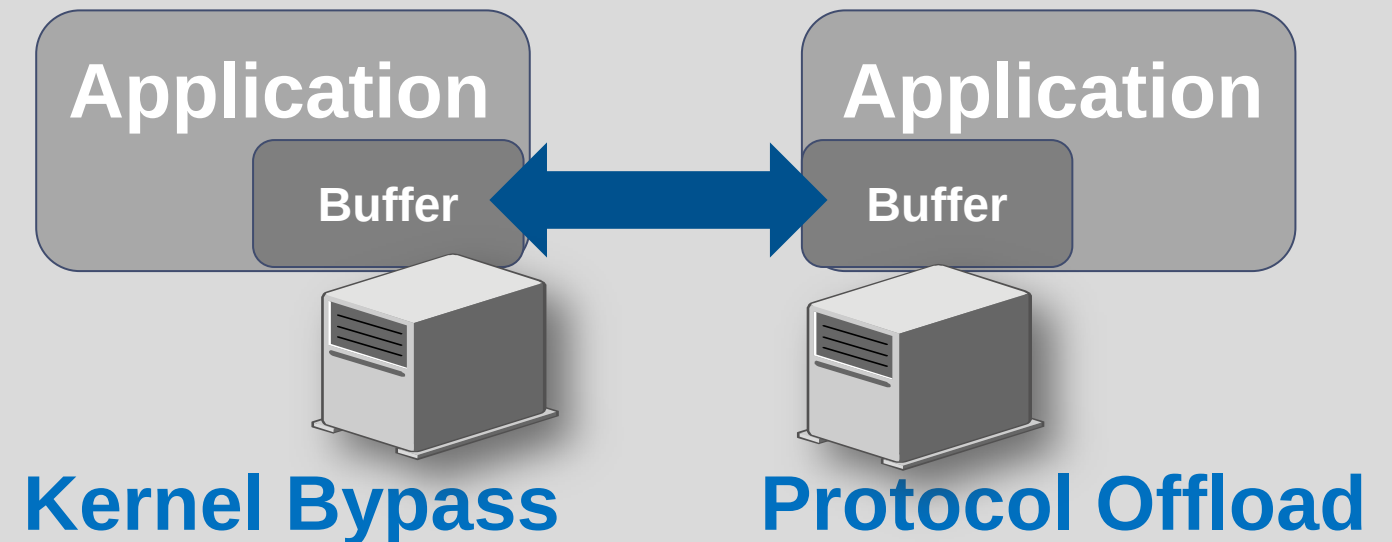
Cables/Modules



ZERO Copy



Remote Data Transfer



Low Latency, High Performance Data Transfers

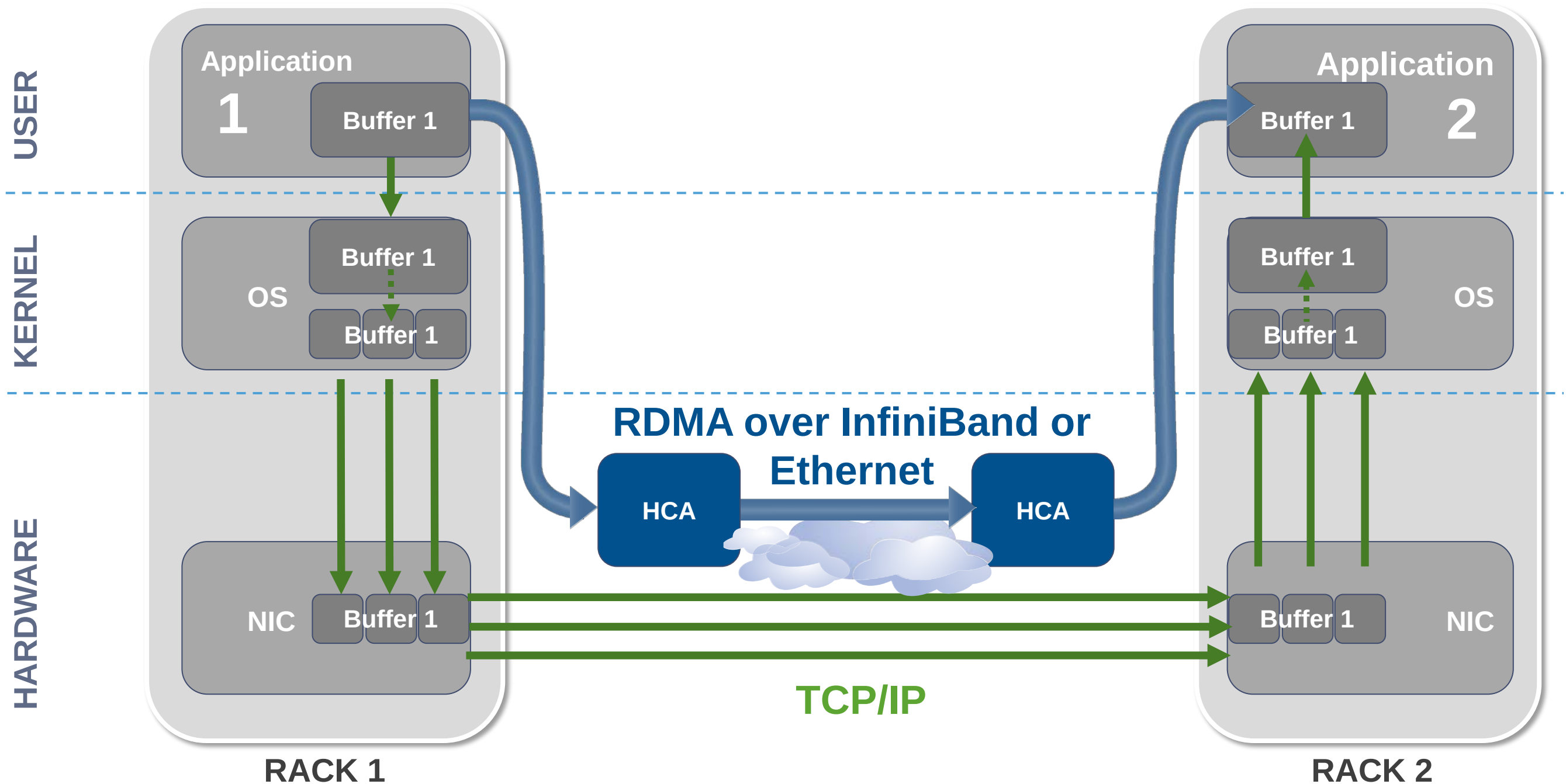


InfiniBand - 56Gb/s

RoCE* – 40Gb/s

* RDMA over Converged Ethernet

RDMA – How it Works



2002

10 Gb/s

SDR

2005

20 Gb/s

DDR

2008

40 Gb/s

QDR

2011

56 Gb/s

FDR
VPI*

2014

100 Gb/s

EDR

FDR InfiniBand New Features and Capabilities

Performance / Scalability

- Switch port bandwidth 56Gb/s
- InfiniBand Router and IB-Eth bridges
- First VPI silicon in the World

Reliability / Efficiency

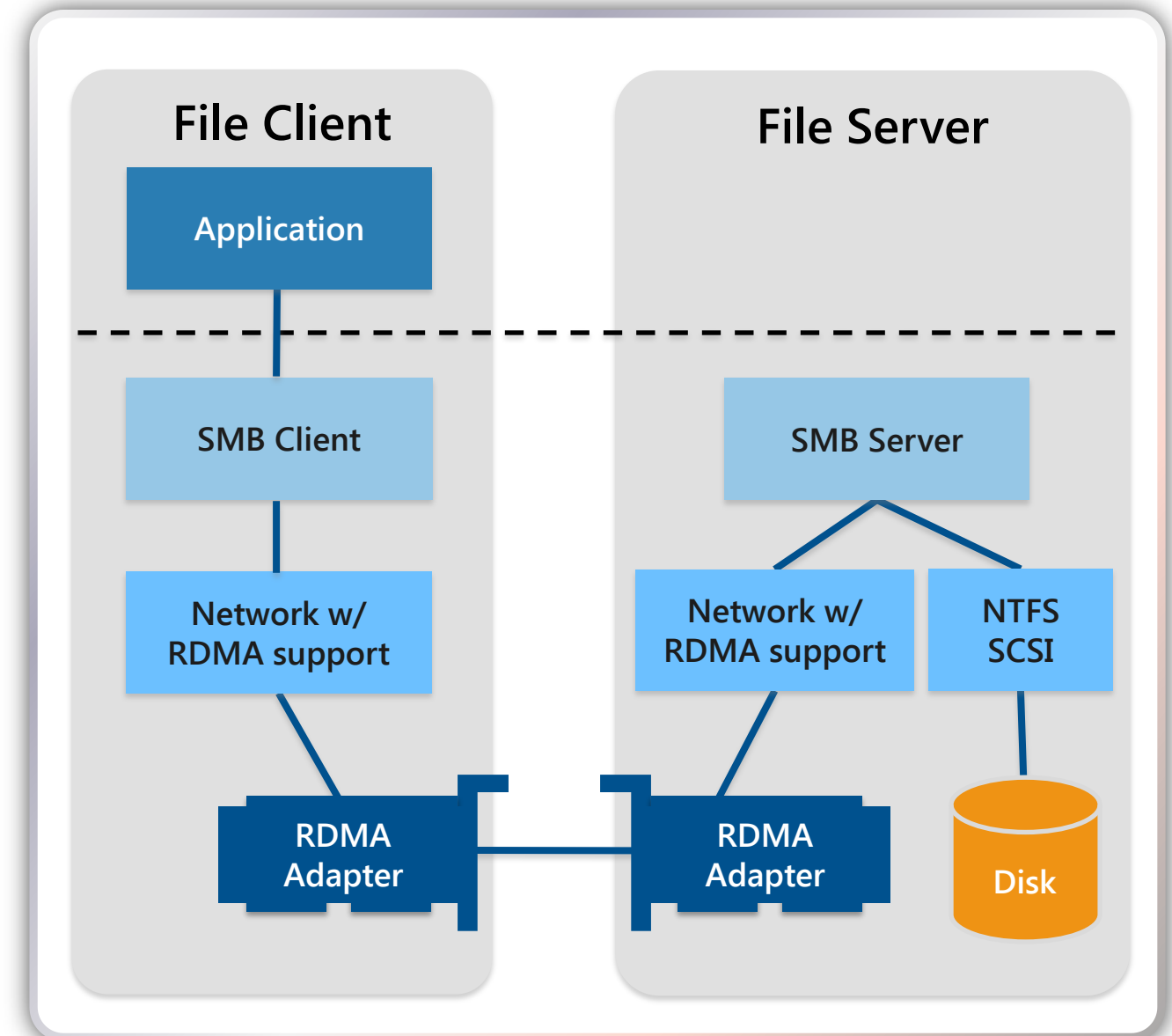
- Link bit encoding – 64/66
- Forward Error Correction
- Low power consumption



Microsoft Windows Server 2012 R2 SMB Direct



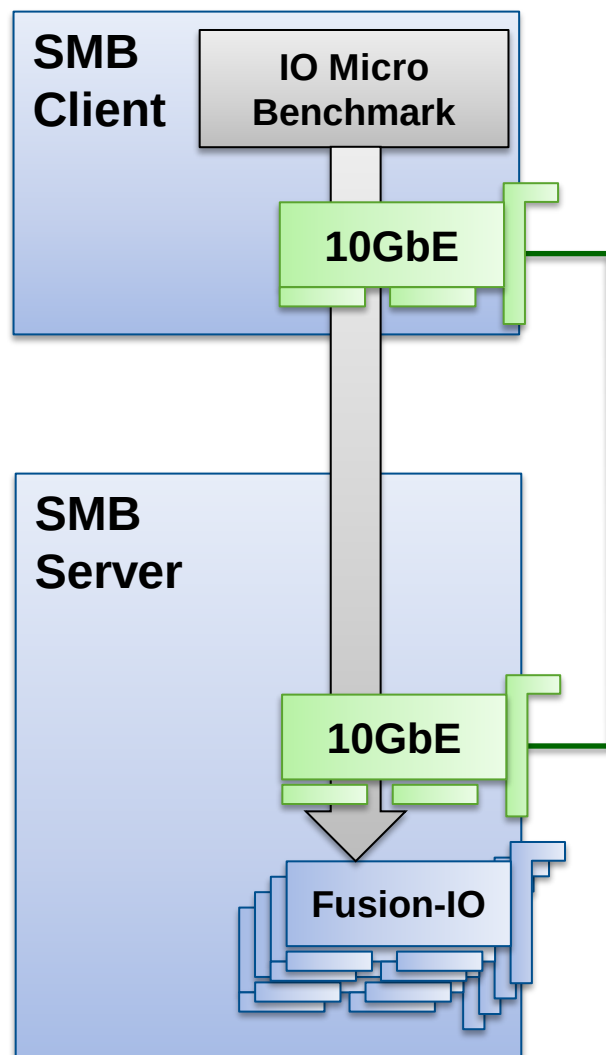
- New class of enterprise file storage with RDMA support
- Lower latency, higher throughput, lower CPU overhead
- Fibre Channel replacement at a lower cost and higher performance
- Leverages Windows Server 2012 R2 Mellanox inbox InfiniBand & Ethernet RDMA drivers
- Accelerates Microsoft Hyper-V over SMB and Microsoft SQL Server over SMB solutions



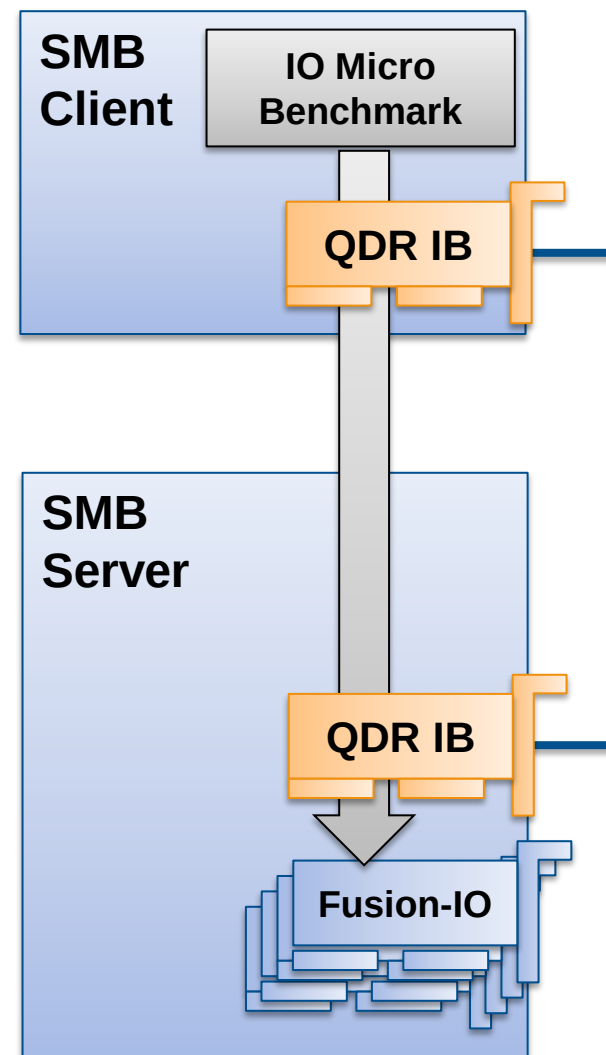
5X Performance Improvement Over 10GbE

Measuring SMB Direct Performance

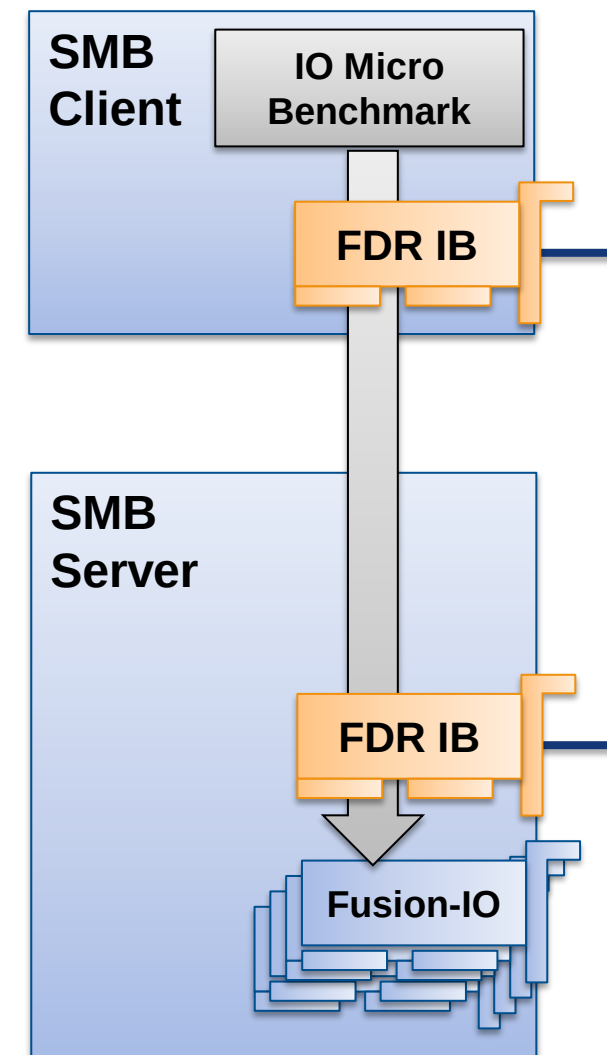
**SMB 3.0 + 10GbE
(non-RDMA)**



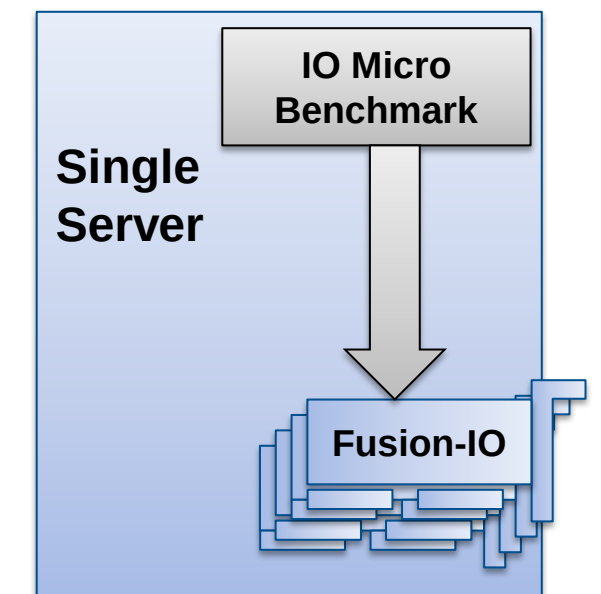
**SMB 3.0 + RDMA
(QDR InfiniBand)**



**SMB 3.0 + RDMA
(FDR InfiniBand)**



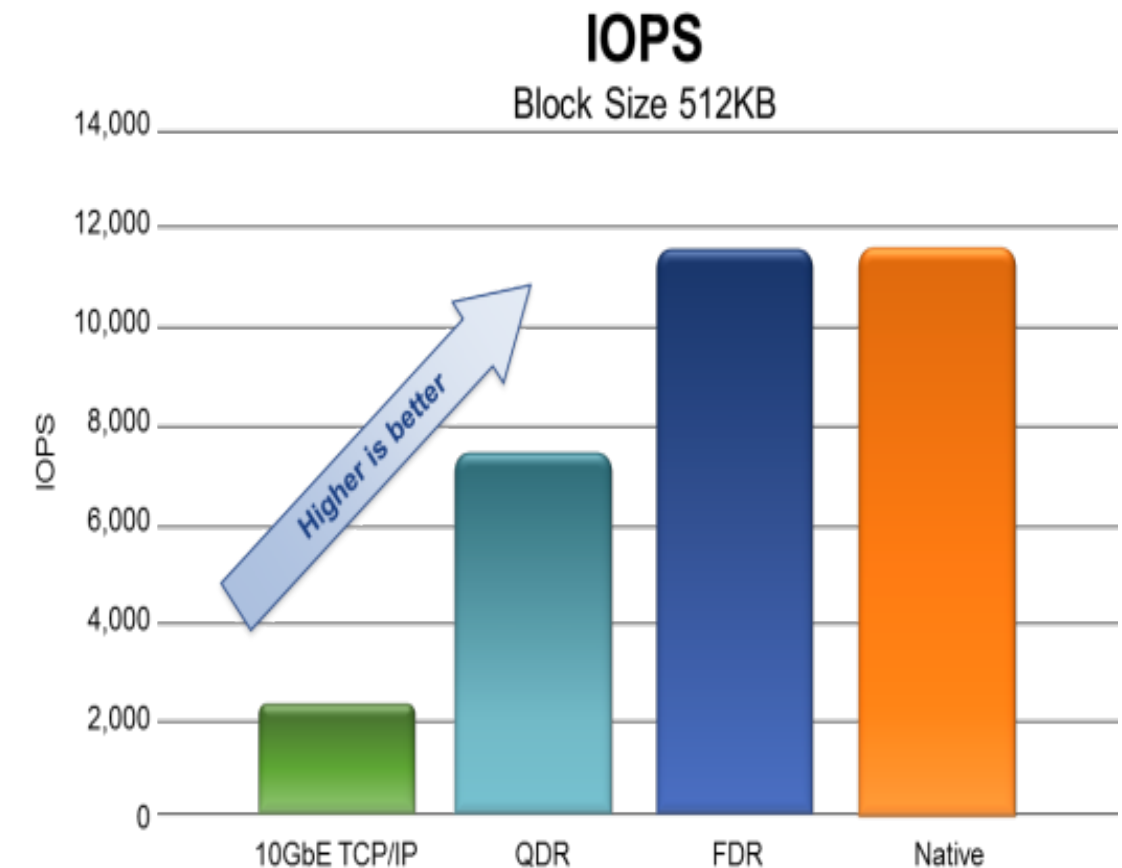
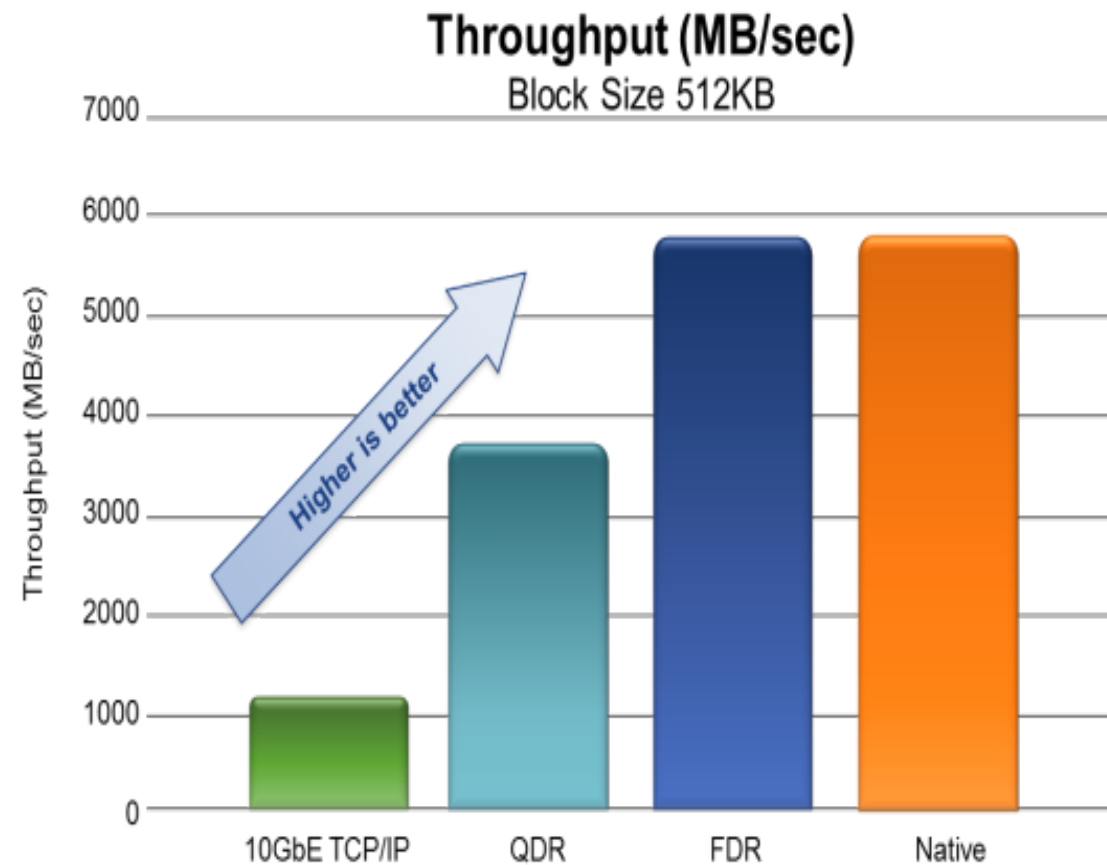
Native



<http://smb3.info>

Performance Comparison

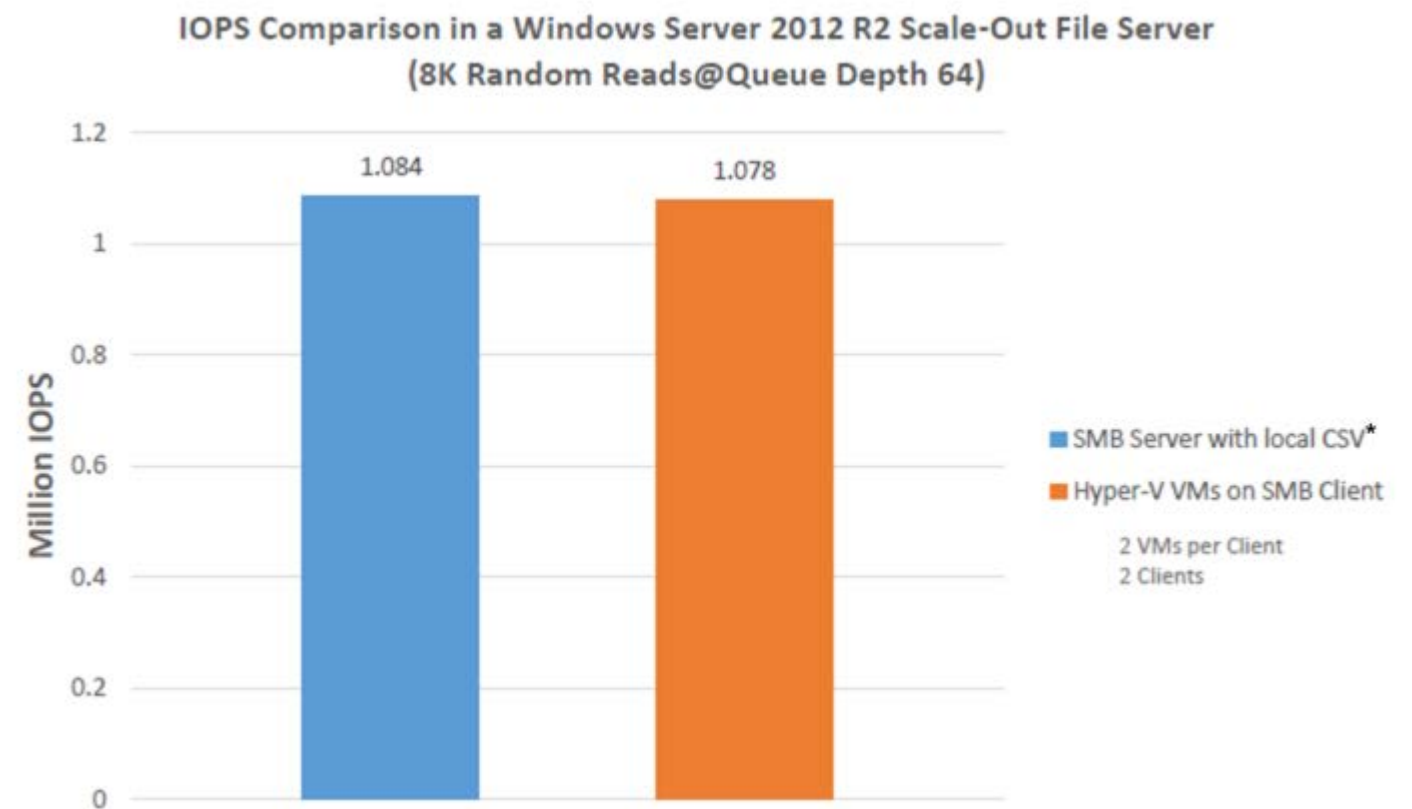
FDR 56Gb/s InfiniBand delivers 5X higher throughput with 50% less CPU overhead vs. 10GbE



Native Throughput Performance over FDR InfiniBand

Scale-Out File Server Cluster using Windows Server 2012 R2 SMB Direct

- 1.1 Million IOPS for 8KB random reads
- 700K IOPS for 8KB random writes
- 100Gbps aggregate read bandwidth
- 92Gbps aggregate write bandwidth



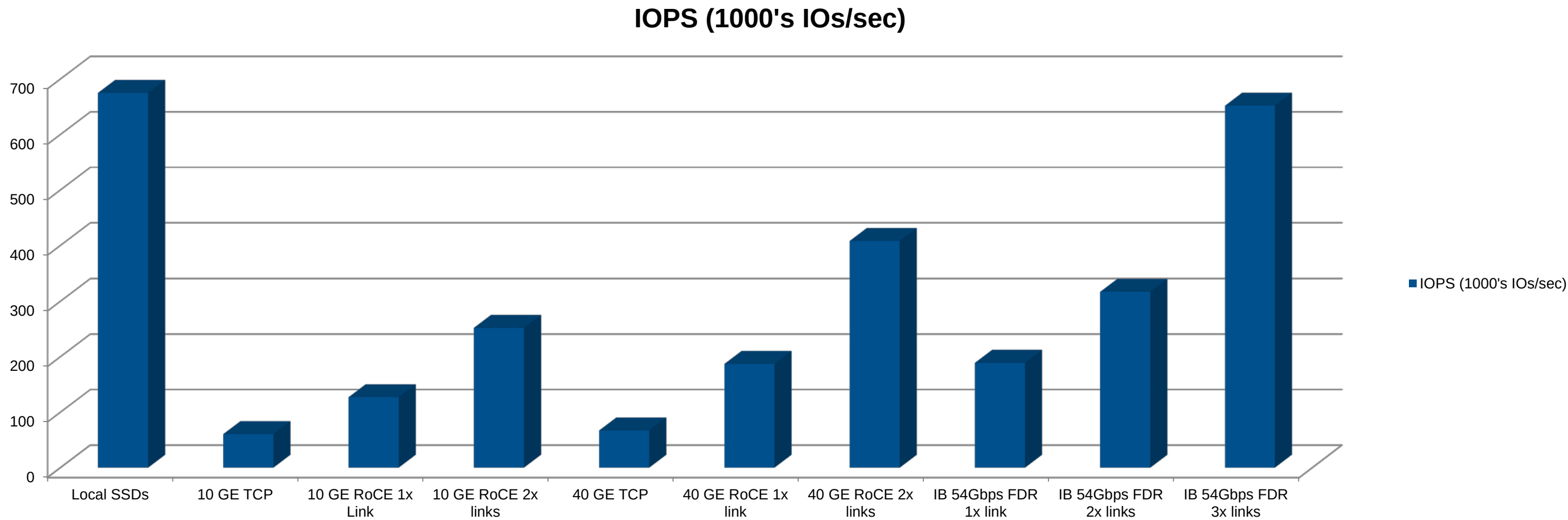
Source: <http://www.mellanox.com/related-docs/applications/Achieving-1.1M-IOPS-over-SMB-Direct.pdf>

*Cluster Shared Volumes

FDR InfiniBand enables Native Performance

Mellanox Testing 8KB I/O

Echostreams flash array with 24 SSDs

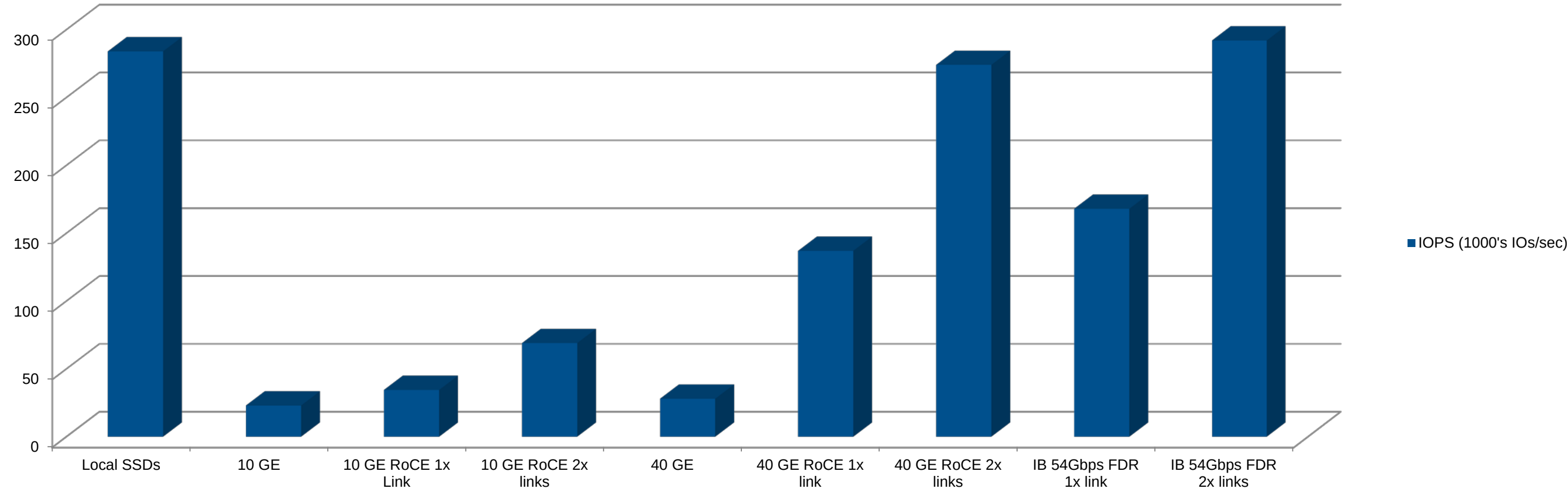


Mellanox Testing 32KB I/O

Echostreams flash array with 24 SSD



IOPS (1000's IOs/sec)

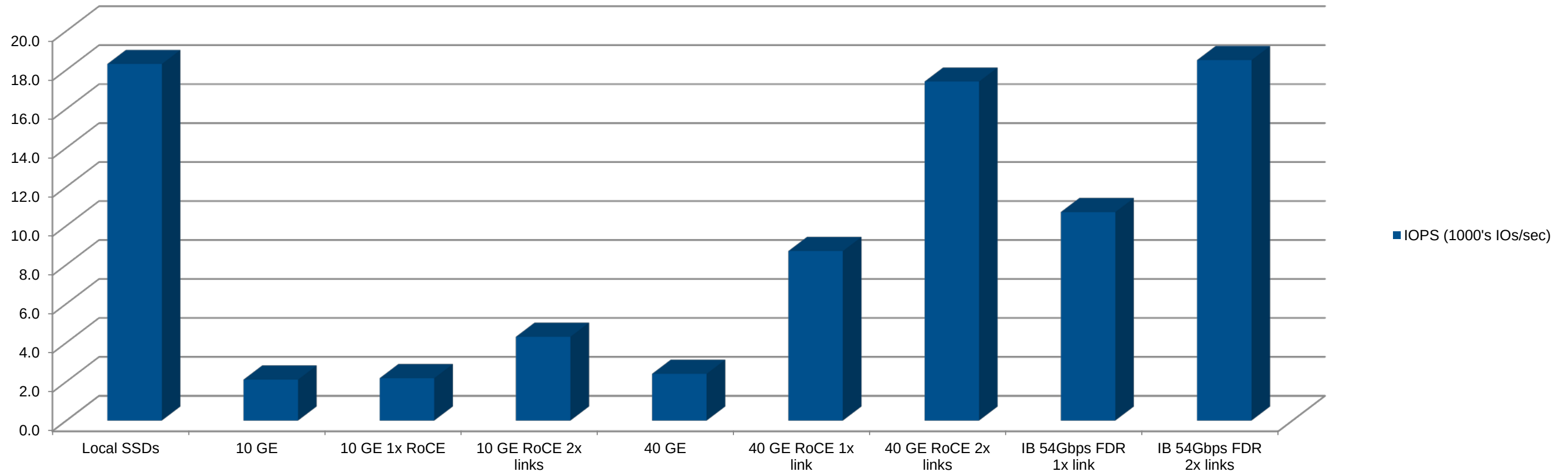


Mellanox Testing 512KB I/O

Echostreams flash array with 24 SSDs



IOPS (1000's I/Os/sec)

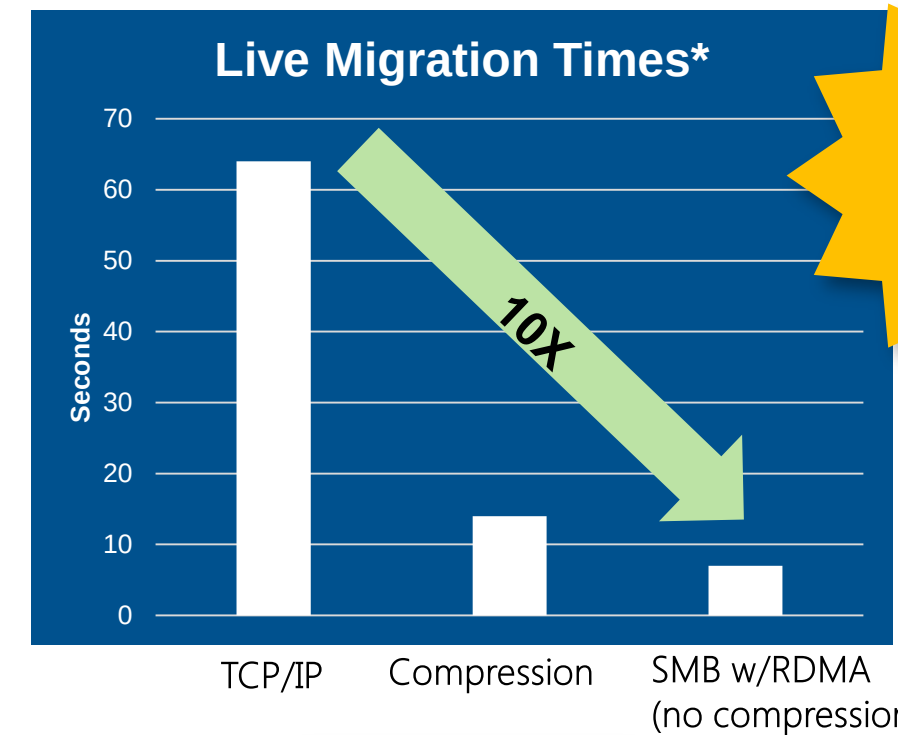
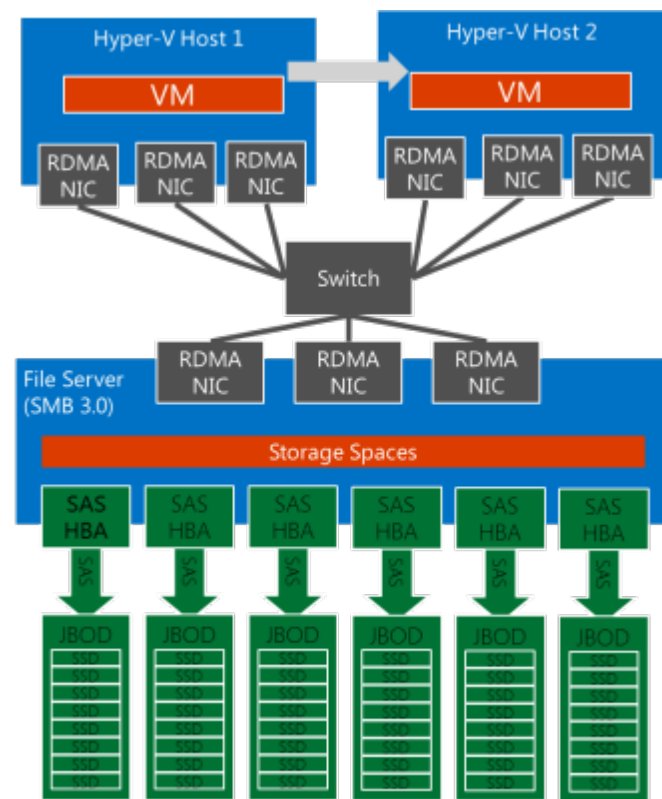


14

Hyper-V Live Migration on SMB Direct with Mellanox Interconnect



- SMB as a transport for Live Migration of VMs
- Delivers the power of SMB to provide:
 - RDMA (SMB Direct)
 - Streaming over multiple NICs (SMB Multichannel)
- Provides highest bandwidth and lowest latency



New in
Windows
Server
2012 R2

*Live migration
can stream over
multiple
networks for
improved
bandwidth*

*Live Migration
can take
advantage of
high speed
networking*

*RDMA enables
offloading
CPU resources
to NIC during
live migration*

Source: TechED'13 Opening Keynote Session with Satya Nadella

Success Story: Edgenet

■ Reduced complexity vs. FC storage management

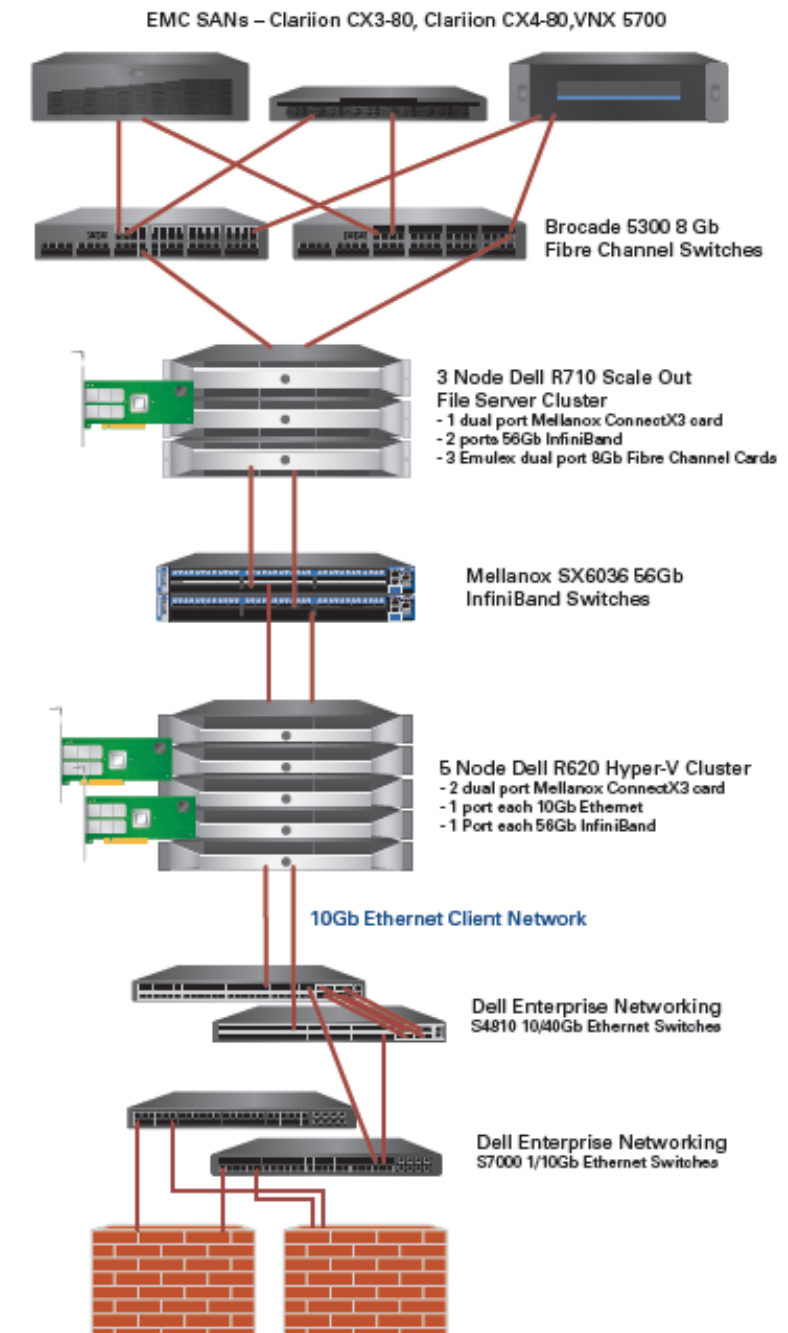
- Creates a more manageable storage subsystem
 - SMB easier to provision than Fibre Channel devices
- The storage server manages all the connections
 - Client Servers only need to know one path via SMB

■ 10x Acceleration vs FC

- Allows consolidation of many FC arrays over SMB
- Live migration between servers at record speeds

■ Mellanox VPI solutions

- Provides user access to the Hyper-V hosted application



Source: http://www.mellanox.com/related-docs/case_studies/CS_Edgenet_Boost_Performance_Lower_Cost.pdf

SMB Direct Storage Solutions

Introducing The Windows Flash Array

violin
MEMORY



Windows in a Flash!

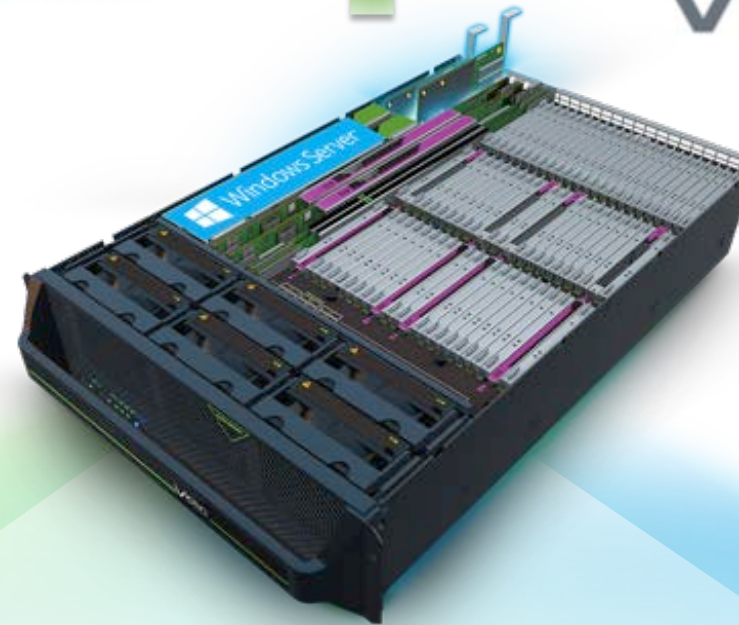
>1 year of Joint Violin - Microsoft Development



SMB Direct for Microsoft applications

RDMA as a Storage Superhighway

Enables file server functions



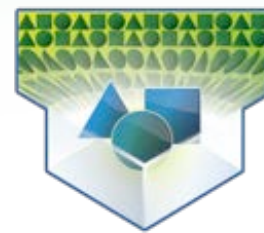
Transformative Speed in Storage

Consolidation and cost savings

Increases server utilization



Encryption



Dedupe



Live Migration



Mirroring



Scale-Out



Beta Site Performance: SQL* and Hyper-V



2x Increase
SQL Writes



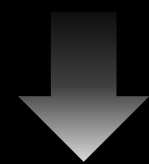
54% Increase
SQL Reads



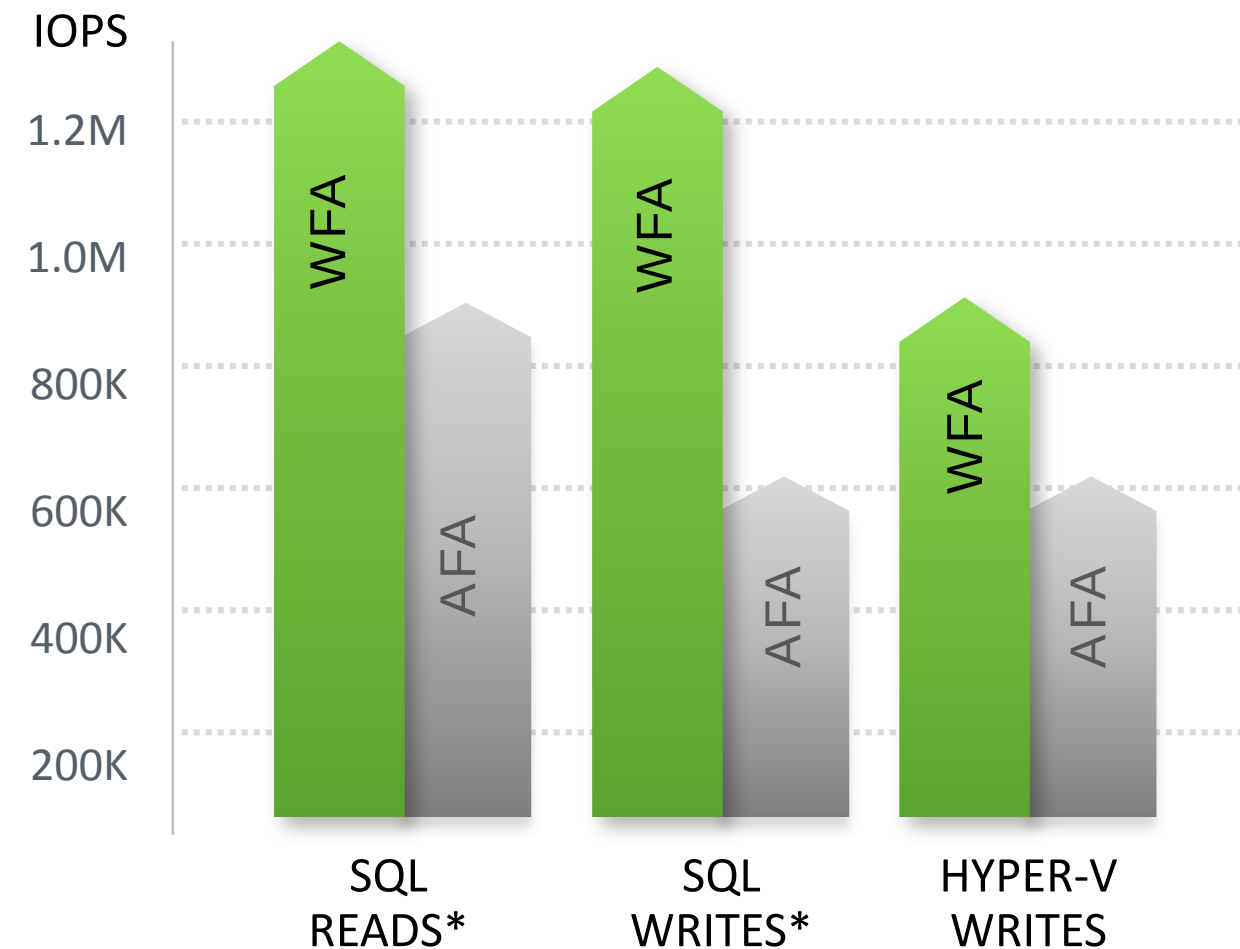
41% Increase
Hyper-V Writes



30% Decrease
Application Server CPU Utilization



50% Decrease
Latency at Load



- SMB Direct reduces CPU load up to 30%
- Latency averages 500 microseconds at Load

*All Microsoft SQL results based on SQLIO load test with Infiniband connectivity

DataOn Cluster-in-a-Box over RDMA

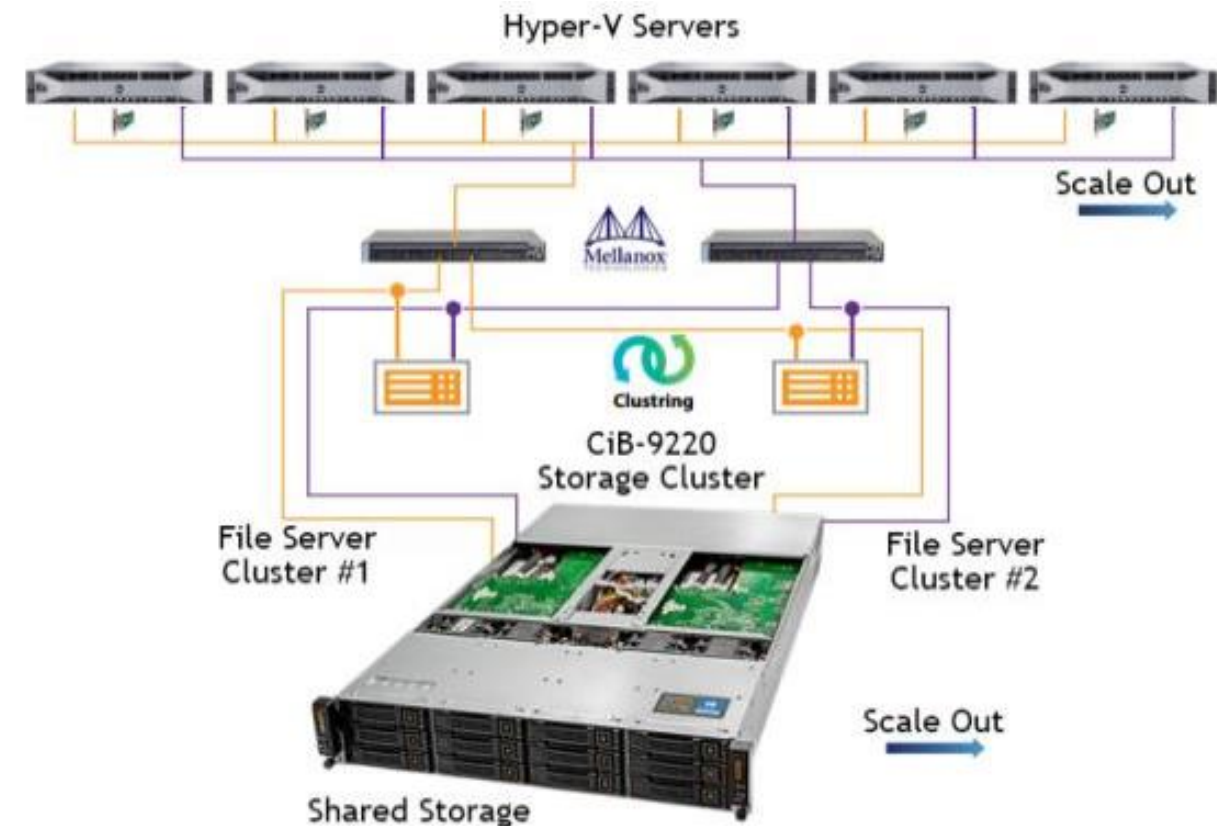
- Extremely economical
- Simplified architecture
- Easy to install appliance
- CiB interconnects:
 - ConnectX-3 40GbE
 - ConnectX-3 56Gbs InfiniBand

- Deployed at:

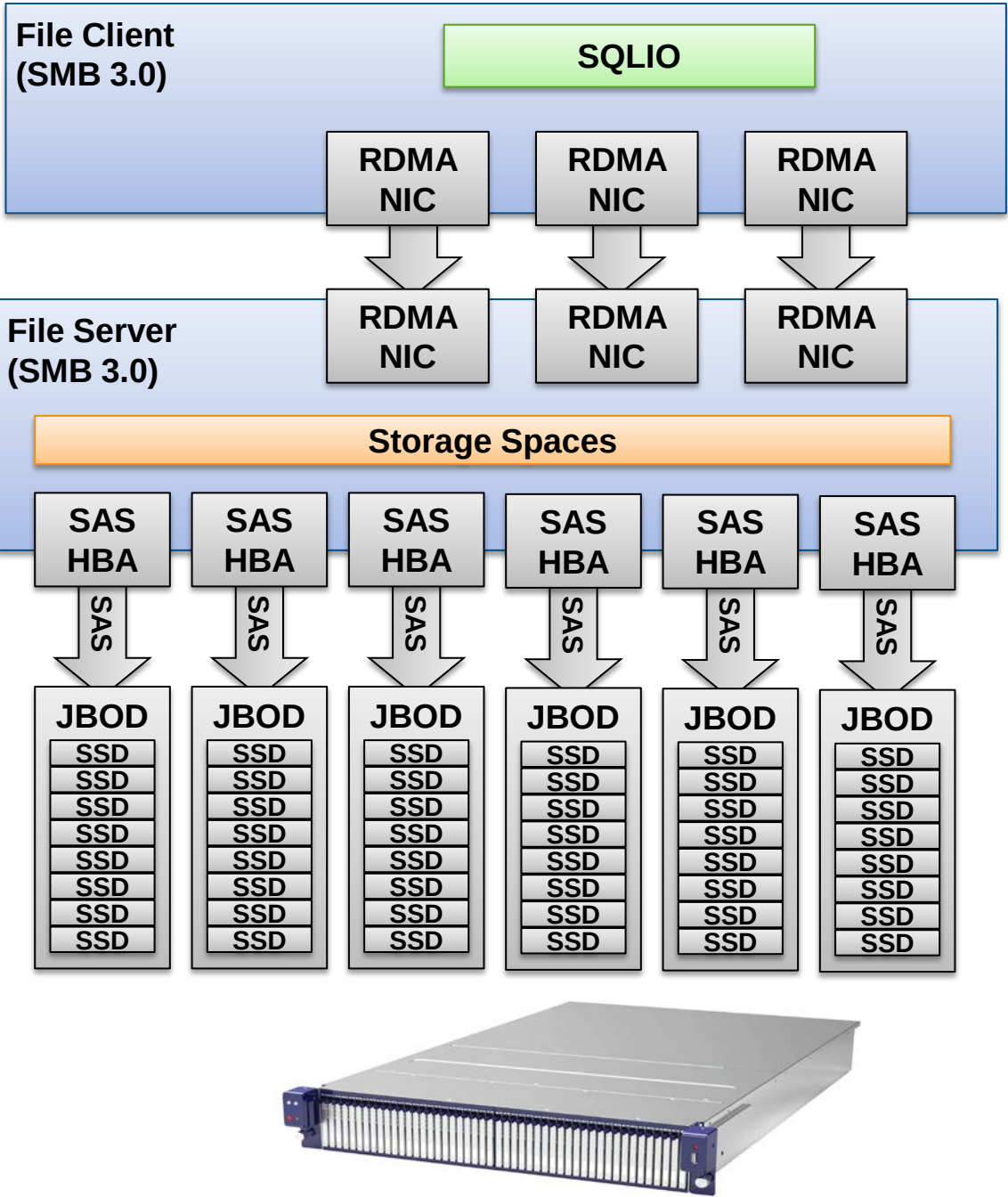
GRACOM

Insight

INTELLIGENT
DECISIONS



EchoStreams: InfiniBand Enables Near Linear Scalability



**8KB random reads
from a mirrored space
(disk)
~600,000 IOPS**

Processor	_Total
% Processor Time	62.698
SMB Client Shares	_Total
Avg. Data Bytes/Request	8,191.986
Avg. Data Queue Length	235.712
Avg. sec/Data Request	0.000
Data Bytes/sec	4,930,617,864
Data Requests/sec	601,881.064

Processor	_Total
% Processor Time	98.242
SMB Client Shares	_Total
Avg. Data Bytes/Request	8,191.985
Avg. Data Queue Length	291.933
Avg. sec/Data Request	0.000
Data Bytes/sec	9,193,607,926
Data Requests/sec	1,122,268.592

**8KB random reads
from cache (RAM)
~1,000,000 IOPS**

**32KB random reads
from a mirrored space
(disk)
~500,000 IOPS
~16.5 GBytes/sec**

Processor	_Total
% Privileged Time	64.560
SMB Client Shares	_Total
Avg. Data Bytes/Request	32,767.935
Avg. Data Queue Length	482.211
Avg. sec/Data Request	0.001
Data Bytes/sec	16,473,350,042.2151
Data Requests/sec	502,725.747

- Preconfigured Microsoft Hyper-V cloud IaaS hardware platform
 - Full integrated converged data center solution
- Iron Pod Description
 - Fault tolerant Mellanox 40/56GbE server and storage network
 - Fully redundant network for high availability
 - ConnectX-3 Pro adapters for RDMA and NVGRE acceleration
 - Dual 40/56GbE network adapters
- More info about Iron Networks:
 - http://www.mellanox.com/related-docs/solutions/Iron_Networks_Microsoft_Fast_Track_Architecture.pdf
 - <http://ironnetworks.com/Microsoft-Converged-Infrastructure/IronPOD-Cloud-Building-Blocks>
 - <http://ironnetworks.com/Microsoft-Converged-Infrastructure/IronPOD-System-500>



Questions?



Thank You

Analyze 1 Petabyte of Data in 1 Second

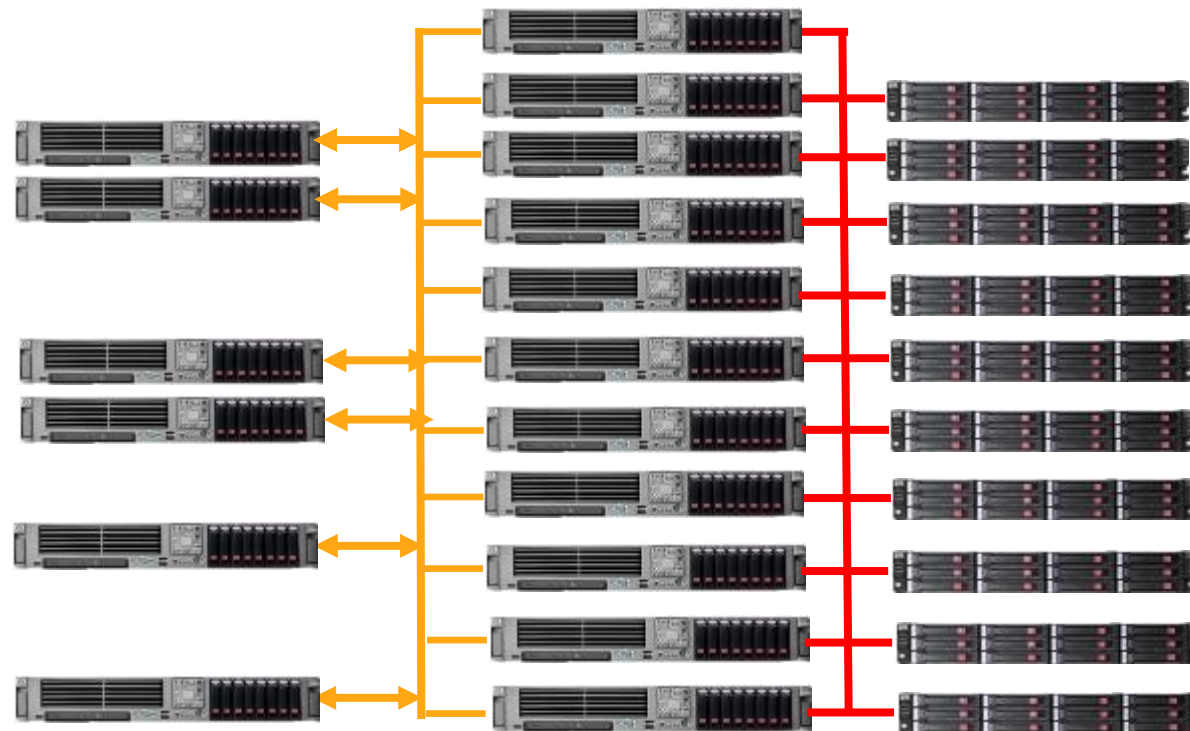
- Up to 100X faster performance than legacy data warehouse queries
- Up to 50X faster data query, up to 2X the data loading rate
- Unlimited storage scalability for future proofing
- Accelerated by Mellanox FDR 56Gb/s InfiniBand end-to-end solutions



Microsoft SQL Server Appliance – 10X Faster & 50% Lower Capital Cost



Parallel Data Warehouse Appliance V1



Ethernet, InfiniBand & Fibre Channel

CONTROL RACK

DATA RACK

- 160 cores on 10 compute nodes
- 1.28 TB of RAM on compute
- Up to 30 TB of temp DB
- Up to 150 TB of user data

\$ Estimated total HW component list price: \$1M

Parallel Data Warehouse Appliance V2



InfiniBand

RACK 1

- 128 cores on 8 compute nodes
- 2TB of RAM on compute
- Up to 168 TB of temp DB
- Up to 1PB of user data

\$ Estimated total HW component list price: \$500K