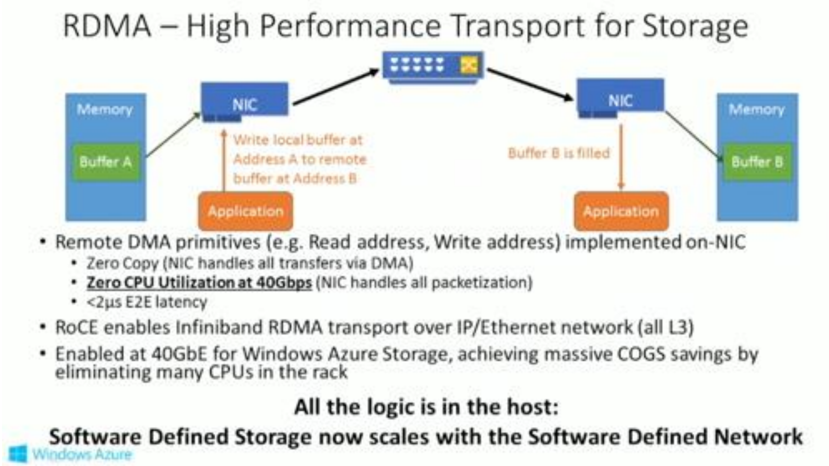
Connecting Flash in Cloud Storage

Kevin Deierling
Vice President Mellanox Technologies
kevind AT mellanox.com

Five Key Requirements for Connecting Flash Storage in the Cloud

1. Economical
2. Massive Scalability & On-Demand Elasticity
3. Converged
4. Fault tolerance & High Availability
5. Virtualization Aware

#1: Cloud Storage Must be Economical



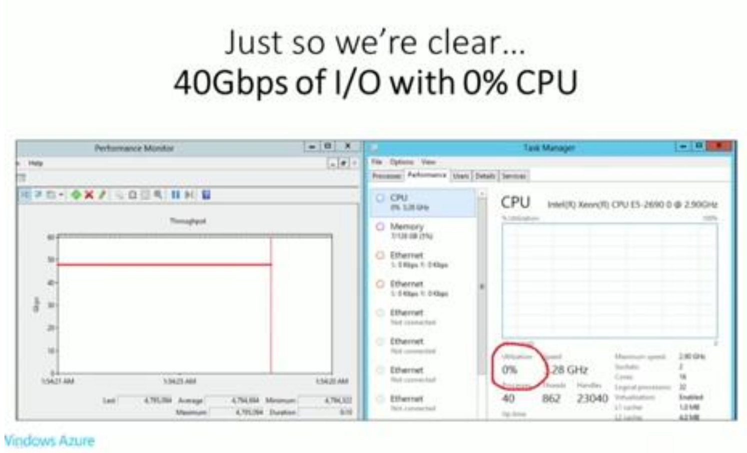
RDMA – High Performance Transport for Storage

- Remote DMA primitives (e.g. Read address, Write address) implemented on-NIC
  - Zero Copy (NIC handles all transfers via DMA)
  - **Zero CPU Utilization at 40Gbps** (NIC handles all packetization)
  - <2µs E2E latency
- RoCE enables Infiniband RDMA transport over IP/Ethernet network (all L3)
- Enabled at 40GbE for Windows Azure Storage, achieving massive COGS savings by eliminating many CPUs in the rack

All the logic is in the host:

Software Defined Storage now scales with the Software Defined Network

Windows Azure



Just so we're clear…

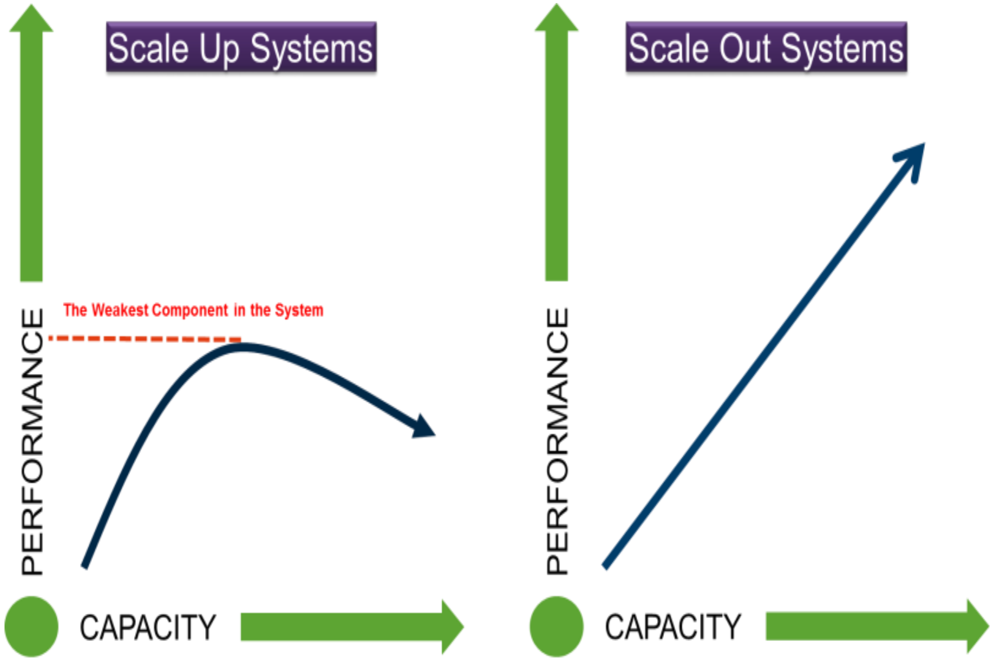
40Gbps of I/O with 0% CPU

Windows Azure

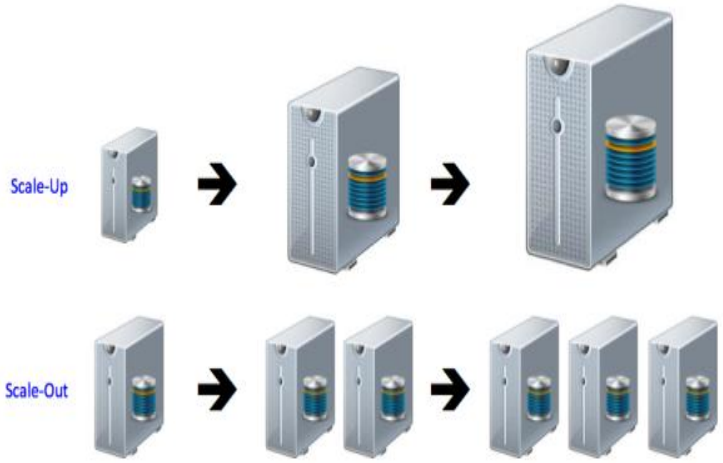
"To make storage cheaper we use lots more network! How do we make Azure Storage scale? RoCE (RDMA over Ethernet) enabled at 40GbE for Windows Azure Storage, achieving massive COGS savings"

ONF 2014, Microsoft Keynote, Albert Greenberg, SDN in Azure Infrastructure

#2: Cloud Storage Must Scale-Out

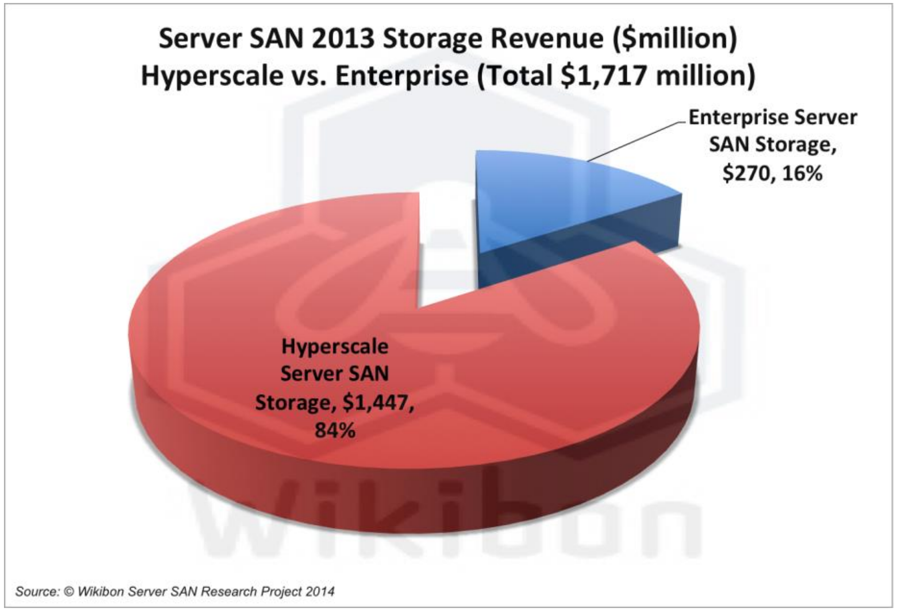


Network Capabilities Determine
Scale Out Performance



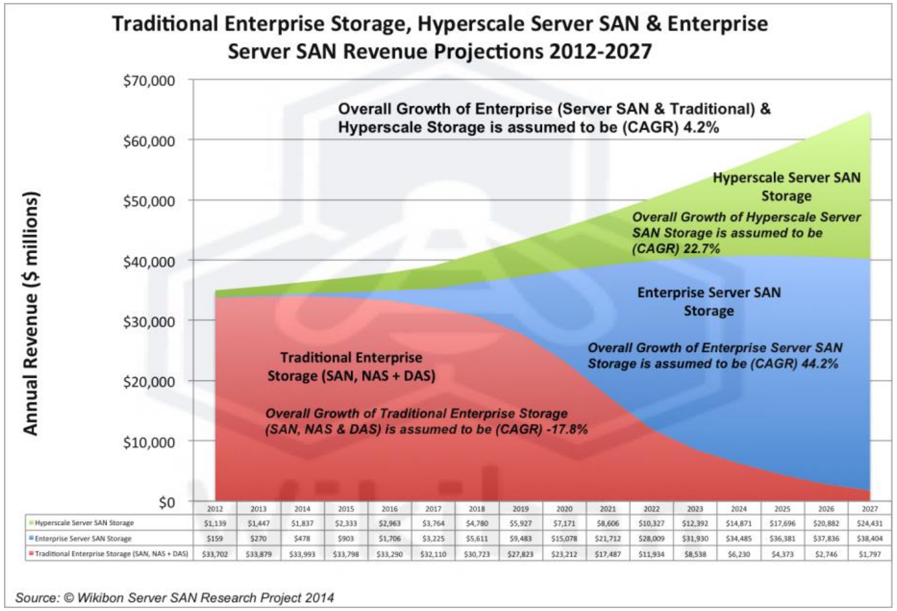
- Scale out required to achieve massive scalability & on-demand elasticity
- Transition from Scale-Up to Scale-Out
  - Only way to support storage capacity growth in a cost-effective manner
  - Accelerated by cloud, big data, HPC
- New scale-out choices

Server SAN: The New Normal in the Cloud


Server SAN 2013 Storage Revenue ($million) Hyperscale vs. Enterprise (Total $1,717 million)

Source: © Wikibon Server SAN Research Project 2014

$1.7B in 2013, 85% is Hyperscale


Traditional Enterprise Storage, Hyperscale Server SAN & Enterprise Server SAN Revenue Projections 2012-2027

	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026	2027
🟩 Hyperscale Server SAN Storage	$1,139	$1,447	$1,837	$2,333	$2,963	$3,764	$4,780	$5,927	$7,171	$8,606	$10,327	$12,392	$14,871	$17,696	$20,882	$24,431
🟦 Enterprise Server SAN Storage	$159	$270	$478	$903	$1,706	$3,225	$5,611	$9,483	$15,078	$21,712	$28,009	$31,930	$34,485	$36,381	$37,836	$38,404
🟥 Traditional Enterprise Storage (SAN, NAS + DAS)	$33,702	$33,879	$33,993	$33,798	$33,290	$32,110	$30,723	$27,823	$23,212	$17,487	$11,934	$8,538	$6,230	$4,373	$2,746	$1,797

Source: © Wikibon Server SAN Research Project 2014

Server SAN at 44.2% CAGR over next 15 years

- Server SAN == Scale Out!
- Server SAN: “Direct attached storage (DAS) devices with high speed interconnects and intelligent software …”, David Floyer, Wikibon, Jul 2014

Scale Out Storage: No “Right” Approach

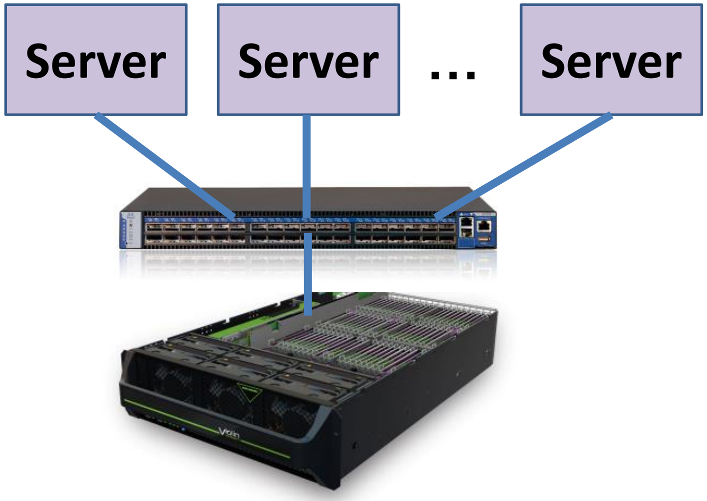
NVMe + Server (Flash-DAS)

Netapp EF540 All Flash Array

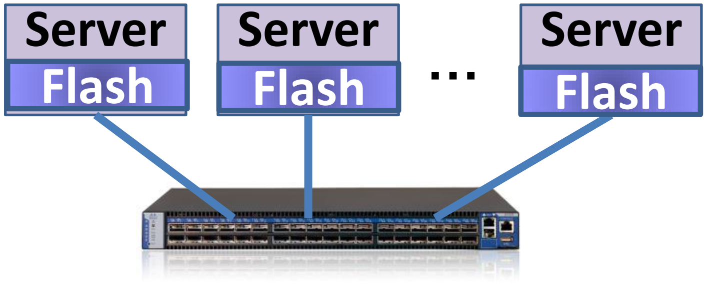
Dell Fluid Cache (Hybrid)

- New Scale Out Options
  - Flash DAS
  - All Flash Arrays
  - Hybrid Cache
- All viable scale out solutions
  - Different trade-offs for different workloads
- Cloud storage needs to be app agnostic
  - Mix of low & high performance apps
  - Requires data to move quickly between nodes

All Flash Arrays vs Flash-DAS



Flash Array (Shared)



Servers with Direct Attached Flash (flash-DAS)

Pros

- Better Flash utilization
- Storage level RAID/HA
- Better Tiering, Balancing

Cons

- Increased Latency

Pros

Flash close to CPU

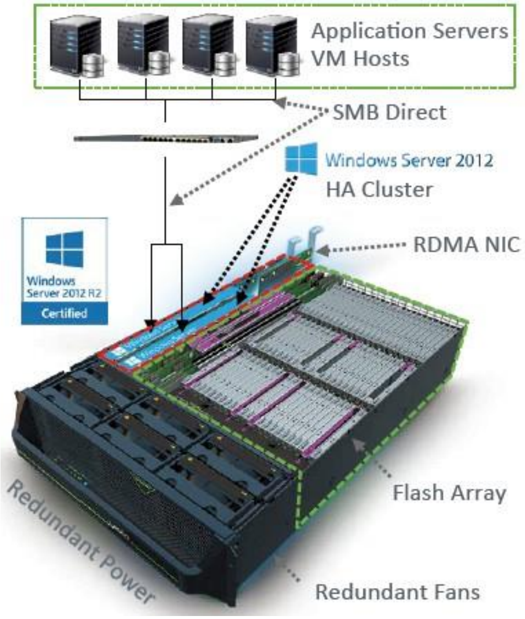
Server level Erasure Coding

Cons

Potentially poor flash utilization

Erasure Coding Consumes Network

All Flash Arrays vs Flash-DAS

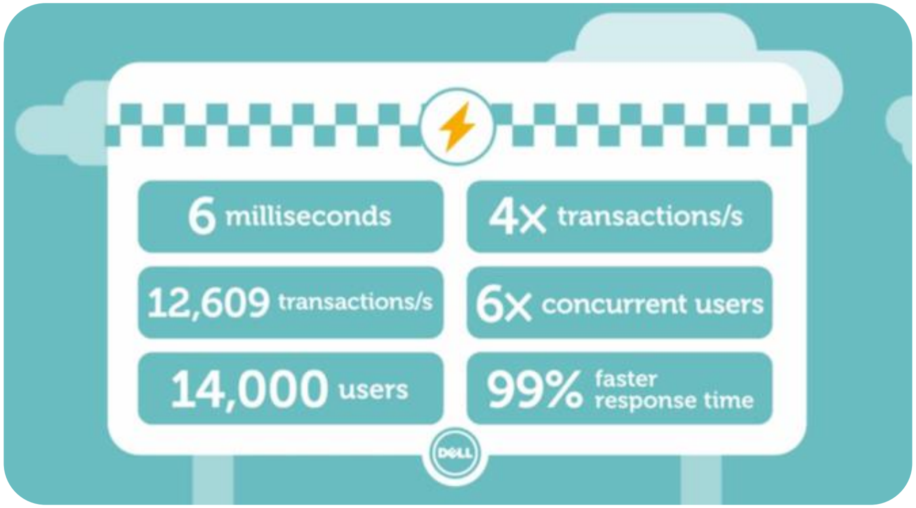


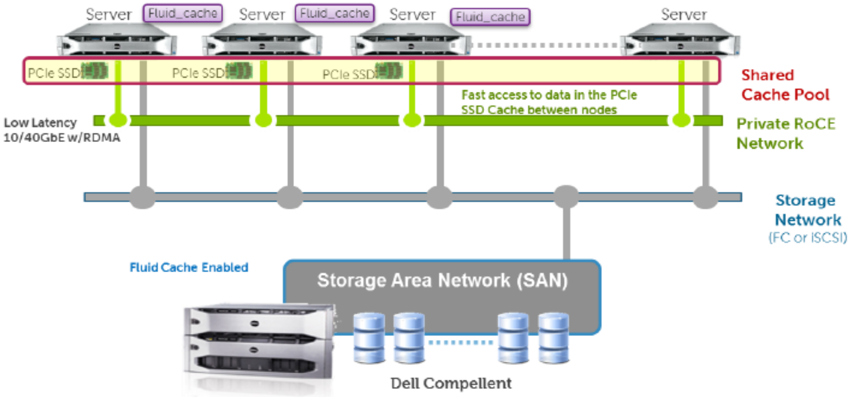
RDMA Enabled Windows Flash Storage Array

Servers with nVME Attached Flash (flash-DAS)

High performance networks with RDMA needed to overcome the limitations of either solution: (AFA or Flash-DAS)

Dell Fluid Cache: Hybrid Approach





Dell Fluid Cache uses low latency RDMA to create a Shared Cache Pool

- Uses iSER over RoCE to create "Shared Cache Pool"
- 4X transactions, 6X Users, & 99% faster response

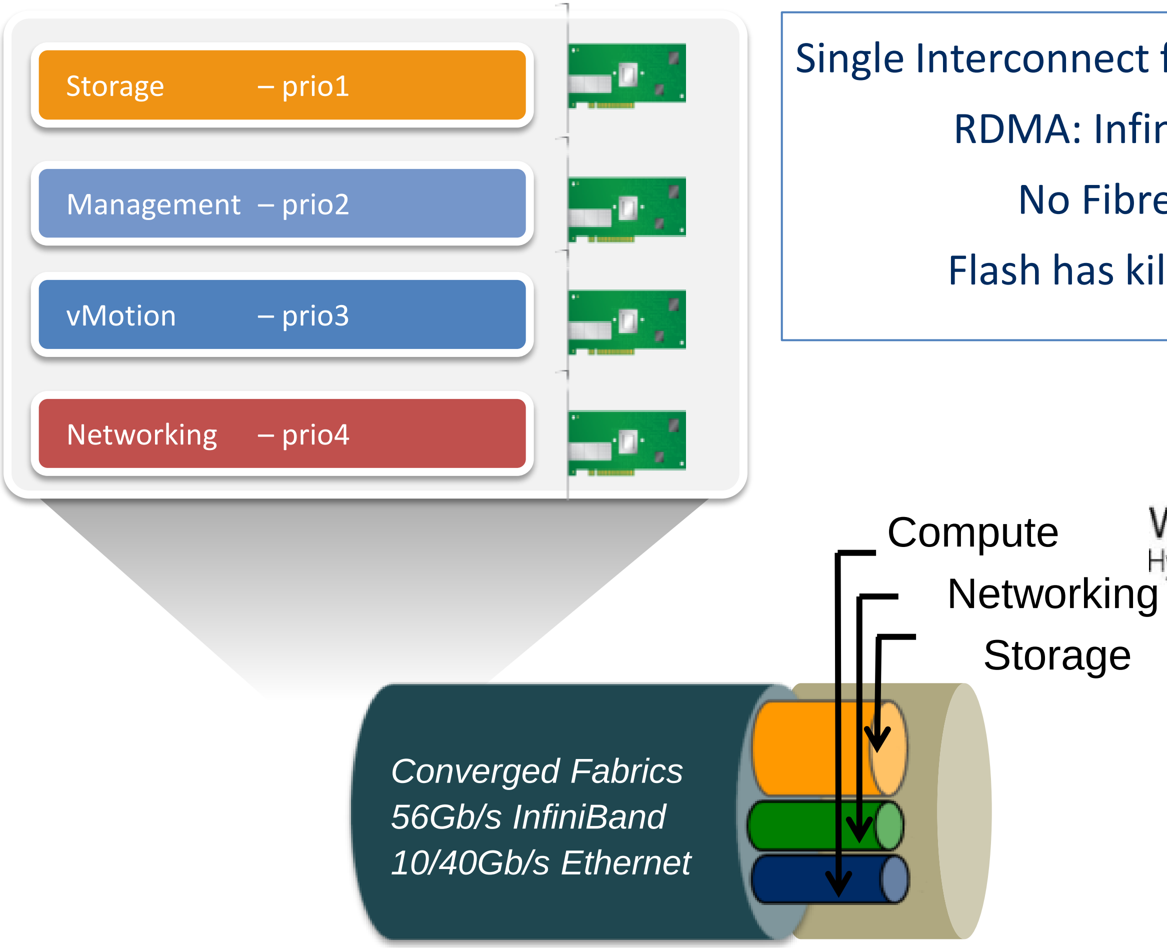
#3: Cloud Storage Must be Converged

Single Interconnect for Compute, Networking, Storage

RDMA: InfiniBand & Ethernet (RoCE*)

No Fibre Channel in the Cloud!

Flash has killed the Fibre Channel HDD



Windows Server Hyper-V

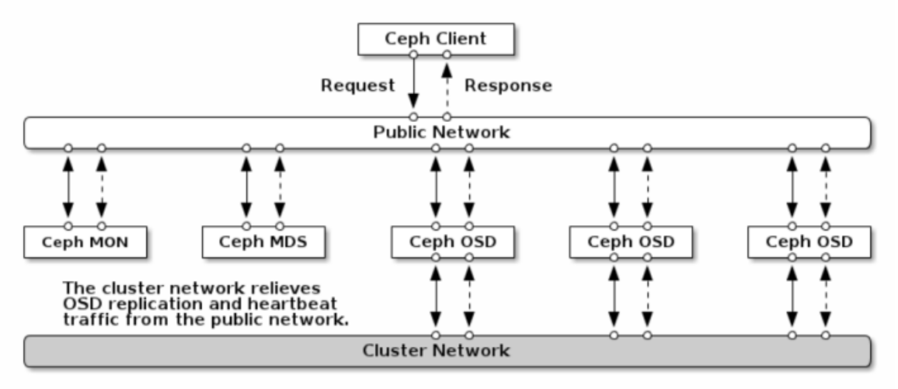
vmware

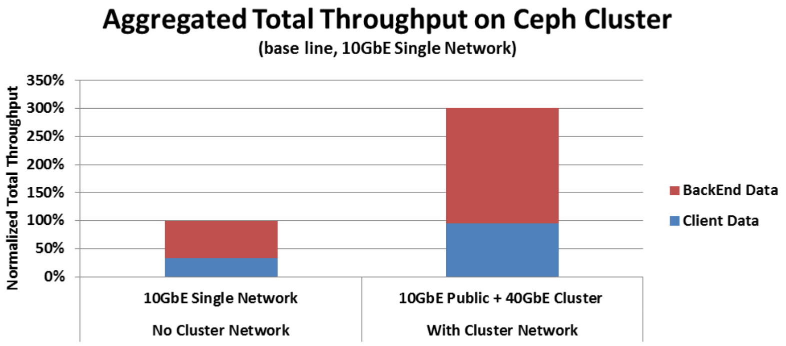
KVM

* RoCE: RDMA over Converged Ethernet

Public & Private Clouds Converging on Fast RDMA Interconnects

Front & Back End Converging Too!





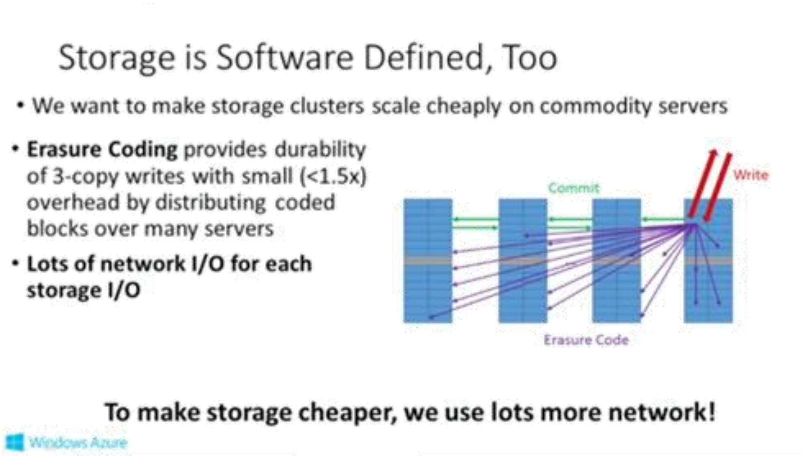
Example: Ceph Back-End Cluster Network Demands High Throughput Interconnect

- Traditional Scale-Up Storage
  - Front-end connectivity comes out-of-the box
  - Back-end connectivity hidden inside the box
    - Higher performance needed due to write-multiplaction (RAID, Mirroring, Caching, Journaling, etc)
- Cloud storage can converge front & back end!

#4: Cloud Storage Needs Fault Tolerance

Traditional RAID
at Disk Level

Erasure Coding
at Server Level



Storage is Software Defined, Too

- We want to make storage clusters scale cheaply on commodity servers
- **Erasure Coding** provides durability of 3-copy writes with small (<1.5x) overhead by distributing coded blocks over many servers
- **Lots of network I/O for each storage I/O**

Commit
Write
Erasure Code

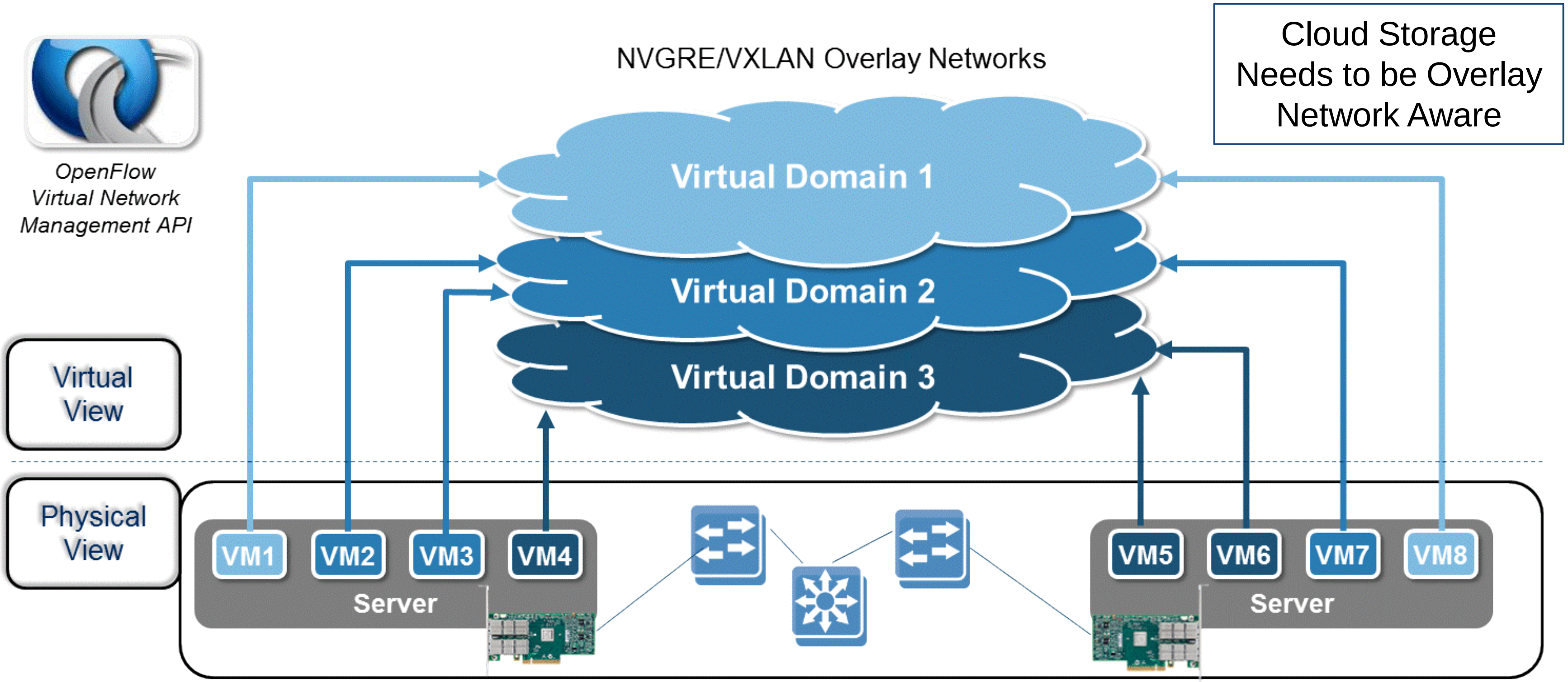
To make storage cheaper, we use lots more network!

Windows Azure

ONF 2014, Microsoft Keynote, Albert Greenberg, SDN in Azure Infrastructure

- But in the Cloud the Fault Domain has changed!
  - Extend beyond RAID to just correct disk-level failures
  - Erasure coding performs error correction at the level of the server-storage unit
- Erasure coding is effective but uses more network

#5: Needs Virtual-Network Aware



- Clouds exploiting overlay network virtualization
  - Multi-tenancy & isolation
  - Virtual network extending to storage
- Virtual overlay networks need hardware enforcement & acceleration

Thanks!
Questions