OPENFABRICS ALLIANCE

Microsoft RDMA Update

11th ANNUAL INTERNATIONAL OPENFABRICS SOFTWARE DEVELOPERS' WORKSHOP

Tom Talpey
File Server Architect
Microsoft

#OFADevWorkshop

Outline

- Introduction
  - Microsoft RDMA use (highlights)
- SMB3 and SMB Direct
- Windows Networking
- Azure
- Software Interfaces
  - Network Direct and “Endure”
  - Linux Virtualization

SMB3 and SMB Direct

SMB3

- The primary Windows remote file protocol
- Long SMB history, since 1980's
  - SMB1 since Windows 2000 ("CIFS" before that)
  - SMB2.0 with Vista and Windows Server 2008, 2.1 in 7/2008R2
- Now at dialect 3
  - SMB3.0 with Windows 8/Server 2012, SMB3.02 in 8.1/WS2012R2
  - SMB3.1.1 in Windows 10 and Windows Server Technical Preview
- Supported:
  - Full Windows File API
  - Enterprise applications
  - Hyper-V Virtual Hard Disks
  - SQL Server
  - New in Windows Server 2012 R2:
    - Hyper-V Live Migration
    - Shared VHDX - Remote Shared Virtual Disk MS-RSVD
  - New in Windows 10 and Windows Server Technical Preview
    - SMB3 Encryption hardening and additional cipher negotiation
    - Hyper-V Storage QoS
    - Shared VHDX – RSVD snapshot and backup

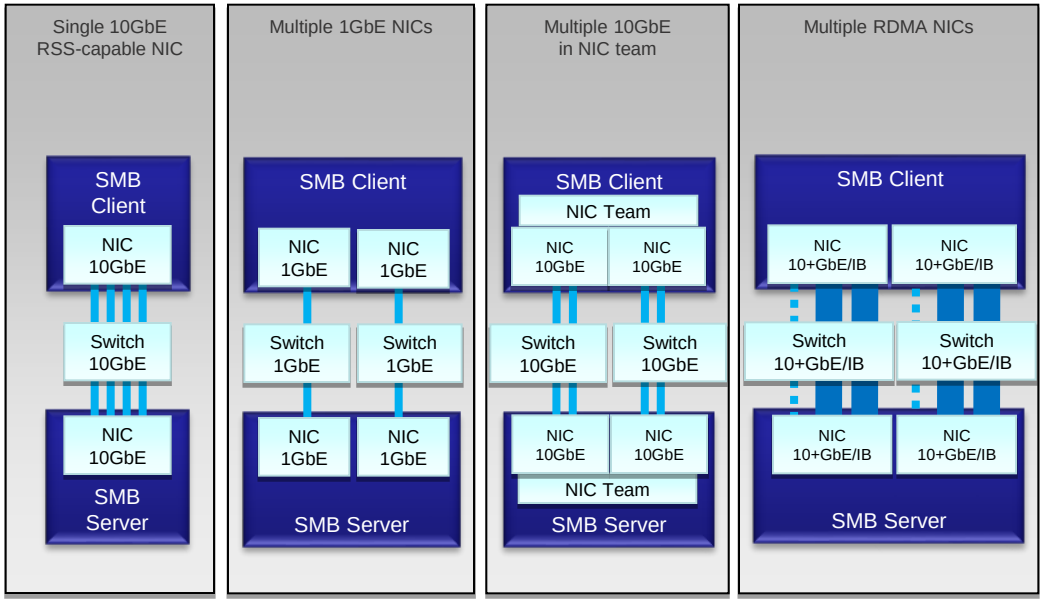
SMB Direct (RDMA)

- Transport layer protocol adapting SMB3 to RDMA
- Fabric agnostic
  - iWARP, InfiniBand, RoCE
  - IP addressing
    - IANA registered (smbdirect 5445)
- Minimal provider requirements
  - Enables greatest compatibility and future adoption
  - Only send/receive/RDMA Write/RDMA Read
  - RC-style, no atomics, no immediate, etc.
- Supported inbox in Windows Server 2012 and 2012R2:
  - iWARP (Intel and Chelsio RNICs at 10 and 40GbE)
  - RoCE (Mellanox HCAs at 10 and 40GbE)
  - InfiniBand (Mellanox HCAs at up to FDR 54Gb)

SMB3 Multichannel

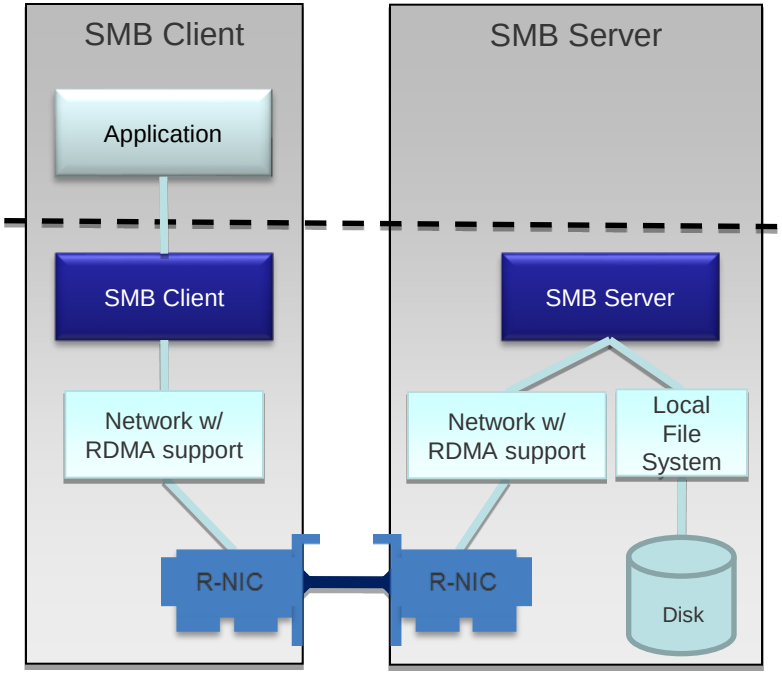
- Full Throughput
  - Bandwidth aggregation with multiple NICs
  - Multiple CPUs cores engaged when NIC offers Receive Side Scaling (RSS) or Remote Direct Memory Access (RDMA)
- Automatic Failover
  - SMB Multichannel implements end-to-end failure detection
  - Leverages NIC teaming if present, but does not require it
- Automatic Configuration
  - SMB detects and uses multiple paths
  - Discovers and “steps up” to RDMA

Sample Configurations



SMB Direct

- Advantages
  - Scalable, fast and efficient storage access
  - High throughput with low latency
  - Minimal CPU utilization for I/O processing
  - Load balancing, automatic failover and bandwidth aggregation via SMB Multichannel
- Scenario
  - High performance remote file access for application servers like Virtualization and Databases
- Required hardware
  - RDMA-capable network interface (R-NIC)
  - Three types: iWARP, RoCE and InfiniBand



SMB Workload Performance

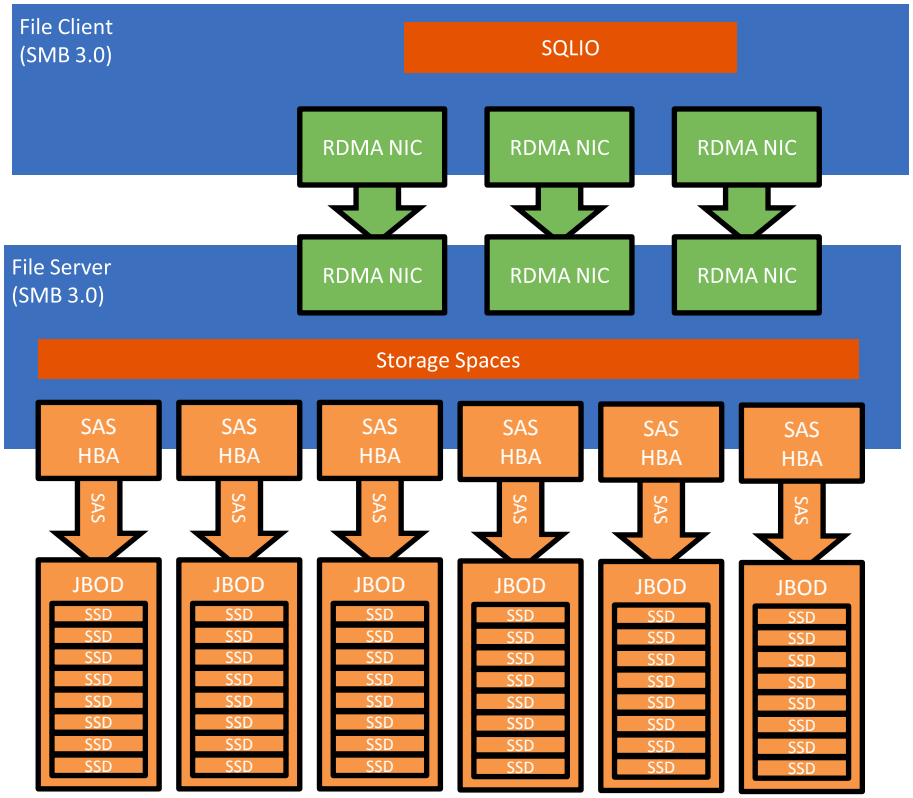
- Key file I/O workloads
  - Small/random – 4KB/8KB, IOPS sensitive
  - Large/sequential – 512KB, bandwidth sensitive
  - Smaller, and larger (up to 8MB), are also relevant
- Multichannel achieves higher scaling
  - Bandwidth and IOPS from additional network interfaces
  - Affinity and parallelism in endnodes
- Unbuffered I/O allows zero-touch to user buffer
- Strict memory register/invalidate per-I/O
  - Key to enterprise application integrity
  - Performance maintained with remote invalidate and careful local behaviors

Other SMB Uses

- SMB as a Transport
  - Provides transparent use of RDMA
- Example: Virtual Machine Live Migration
  - Peer-to-peer memory image transfer
- Bulk data transfer, ultimate throughput
  - Memory-to-memory “unconstrained by disk physics”
  - Very low CPU with RDMA – more CPU to support **live migration**
    - Minimal VM operational blackout

Performance: IOPS and Bandwidth

Single client, single server, 3xIB FDR multichannel connection, to storage and to RAM



Workload	IOPs
8KB reads, mirrored space (disk)	~600,000
8KB reads, from cache (RAM)	**~1,000,000**
32KB reads, mirrored space (disk)	~500,000

Throughput **>16 Gbytes/second**

Larger I/Os (>32KB) – similar results, i.e. larger i/o not needed for achieving full performance!

Link to full demo in “Resources” slide below

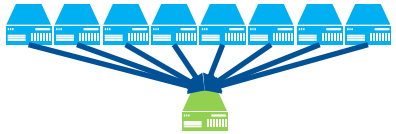
Performance: SMB3 Storage with RDMA on 40GbE iWARP PRELIMINARY

OPENFABRICS ALLIANCE

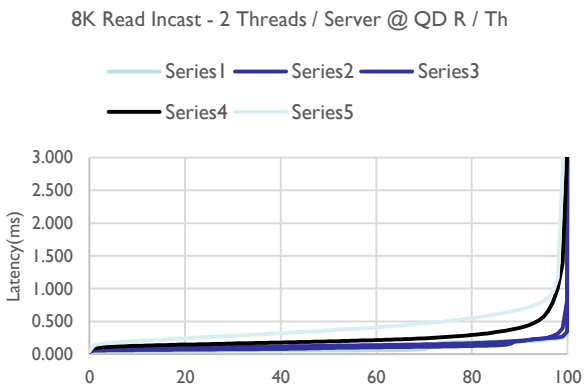
8 KB IOPs* and Latency (ms) Single client, 8 file servers

QD *R* / Thread	Read IOPs	Latency 50th Read	Latency 99th Read
1	163,800	0.056	0.249
2	291,000	0.094	0.270
4	440,900	0.129	0.397
8	492,400	0.195	1.391
16	510,200	0.363	2.810

Incast Performance



Near line-rate with small IOs
- 33.4 Gbps at 8 KB IO
- Excellent latency variance

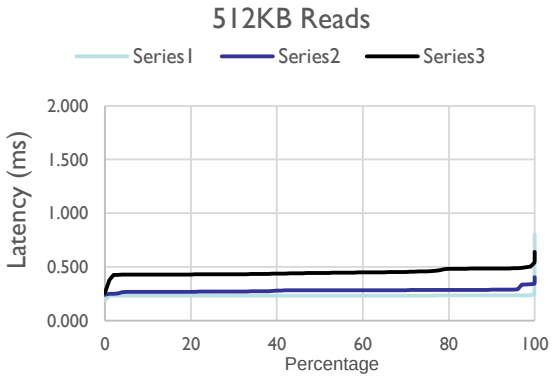


Large IOPs* (512 KB) in Gbps

512KB	Read BW	Write BW
1 Thread		
1 IO	17.82	16.14
2 IO	29.74	23.13
2 Thread		
2 IO/t	37.21	30.61

Client-to-Server Performance

Near line-rate with large IOs
- 37.2 Gbps with 2 threads

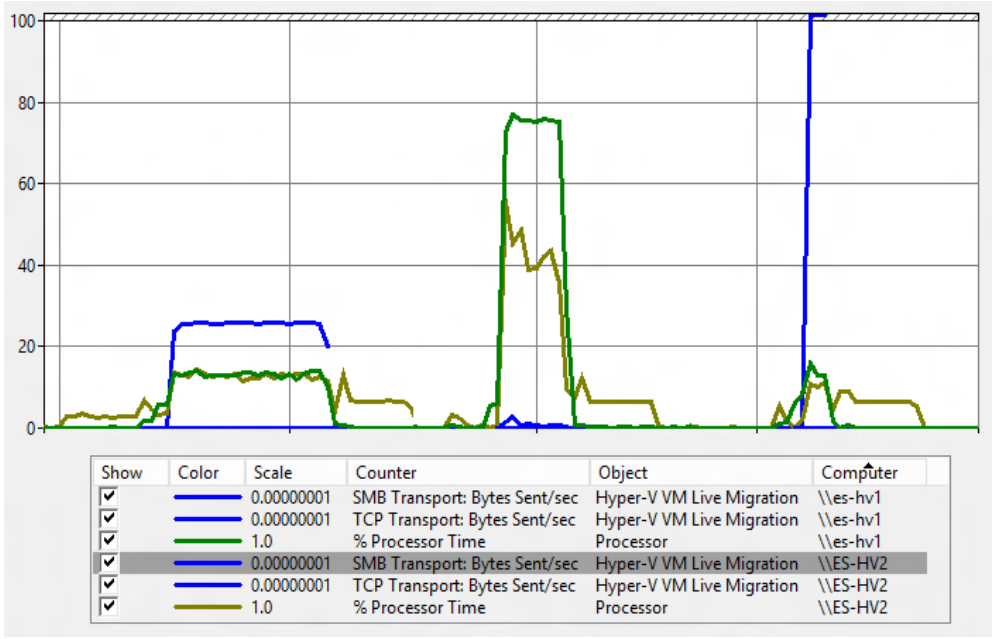


Excellent Log Write Performance
- 7.2M Read IOPs*, 512 Byte, single outstanding IO
- 3.3M Write IOPs*, 512 Byte, single outstanding IO

*IOs are not written to non-volatile storage
iWARP used for these results
Test configuration details in backup

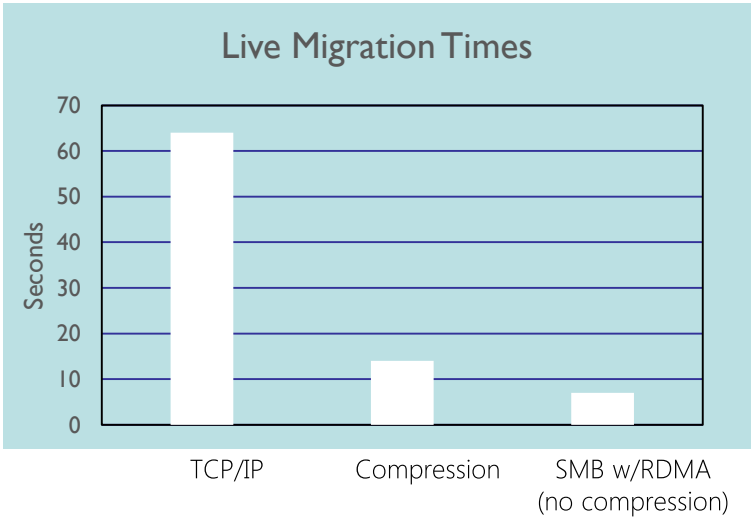
Performance: VM Live Migration on 2x10GbE RoCE

TCP/IP | Compression | SMB+RDMA



- TCP for Live Migration
  - CPU core-limited, long migration time
- TCP with Compression for Live Migration
  - Even more CPU-intensive, reduced migration time
- **SMB Direct for Live Migration**
  - **Minimal CPU, 10x better migration time**
  - **Greatly increased VM live availability**

Live Migration
2x10 GbE RoCE

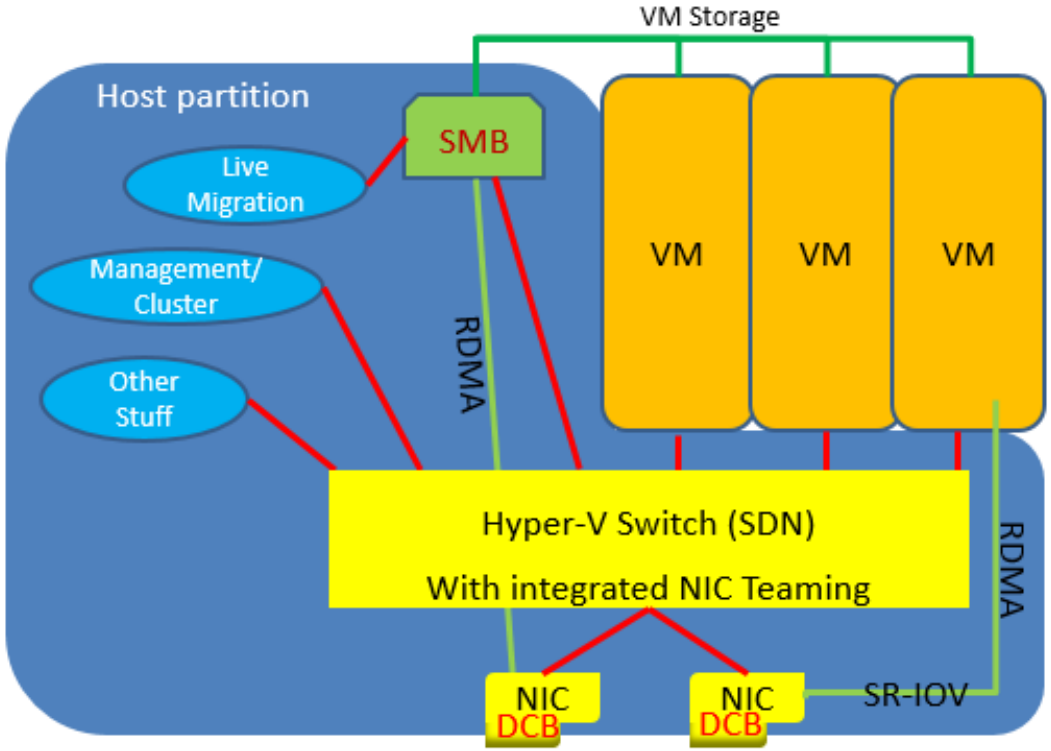


Windows Networking

Converged Fabric

Overview

- Converged fabric supporting virtualized tenant *and* RDMA-enabled disaggregated storage traffic, with quality-of-service guarantees
  - Integrated with SDN-vSwitch
  - Teaming of NICs supported



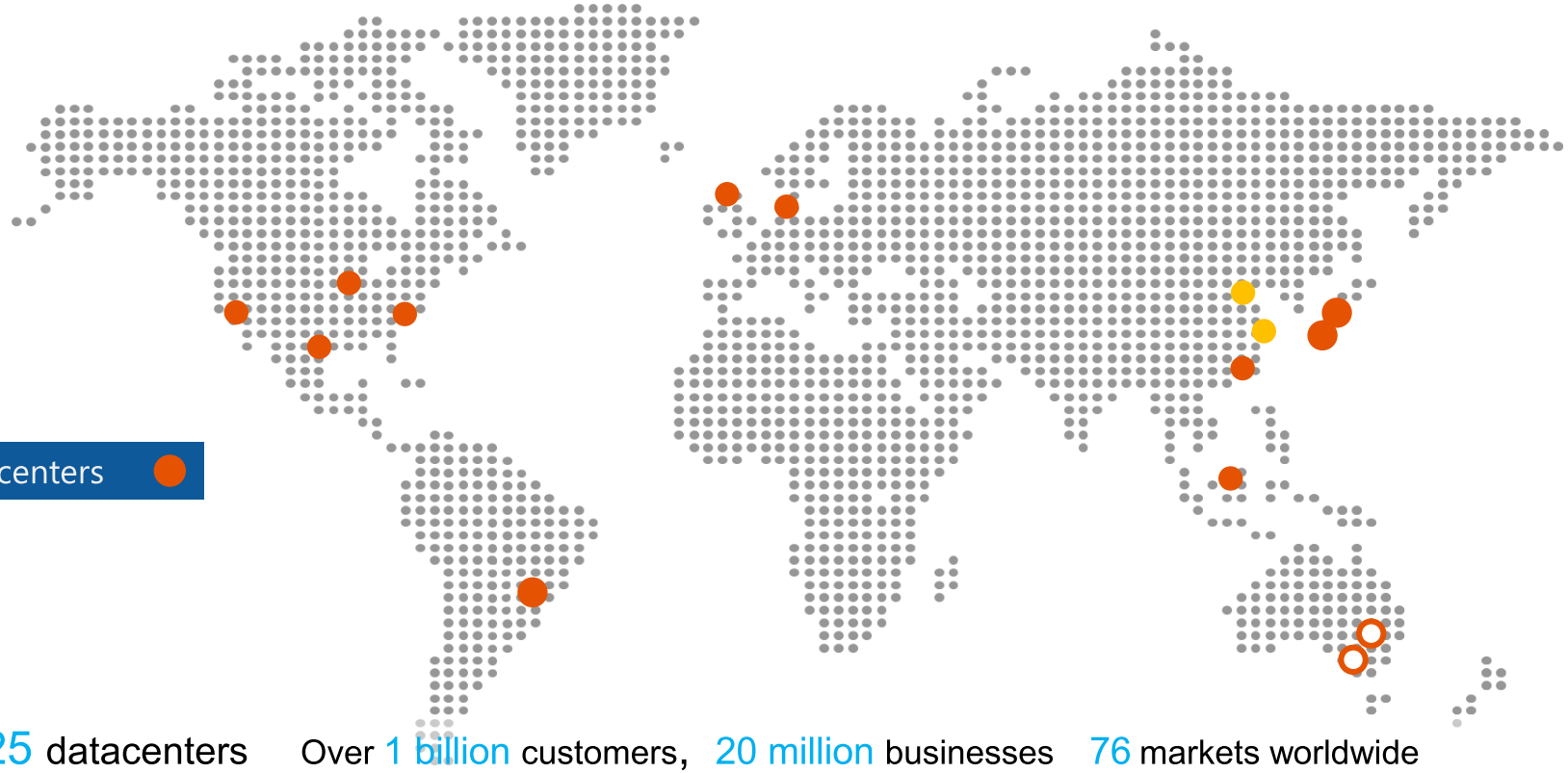
http://channel9.msdn.com/Events/TechEd/Europe/2014/CDP-B248

Discussion and demo begins at minute 40:00

Azure

Windows Azure

OPENFABRICS ALLIANCE



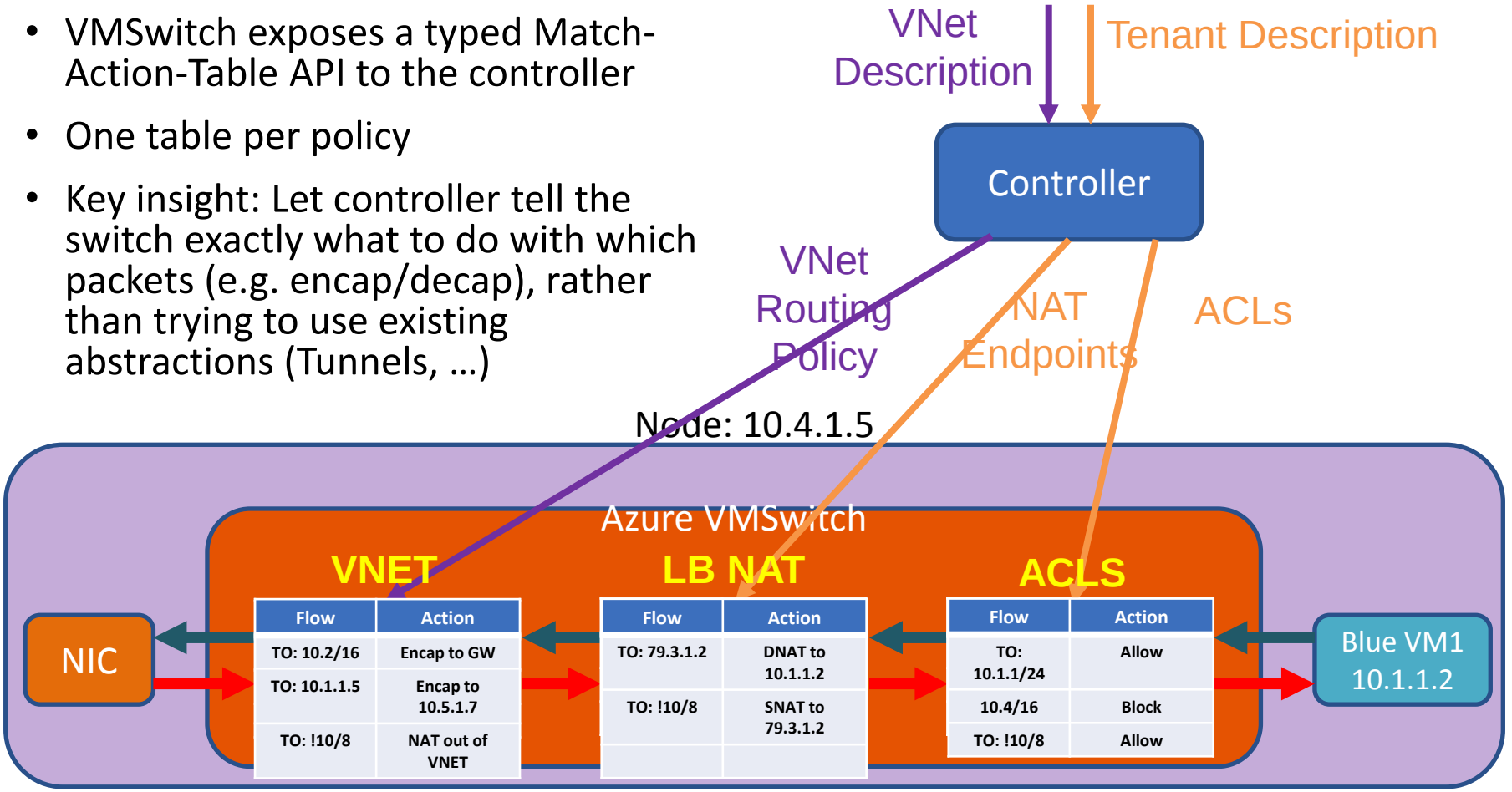
Over 25 datacenters Over 1 billion customers, 20 million businesses 76 markets worldwide

Summary

- **Scenario**: BYO Virtual Network to the Cloud
  - Per customer, with capabilities equivalent to on premise counterpart
- **Challenge**: How do we scale virtual networks across millions of servers?
- **Solution**: Host SDN solves it: scale, flexibility, timely feature rollout, debuggabililty
  - Virtual networks, software load balancing, …
- **How:** Scaling flow processing to millions of nodes
  - Flow tables on the host, with on-demand rule dissemination
  - RDMA to storage
- **Demo**: **ExpressRoute** to the Cloud (Bing it!)

Flow Tables

- VMSwitch exposes a typed Match-Action-Table API to the controller
- One table per policy
- Key insight: Let controller tell the switch exactly what to do with which packets (e.g. encap/decap), rather than trying to use existing abstractions (Tunnels, …)

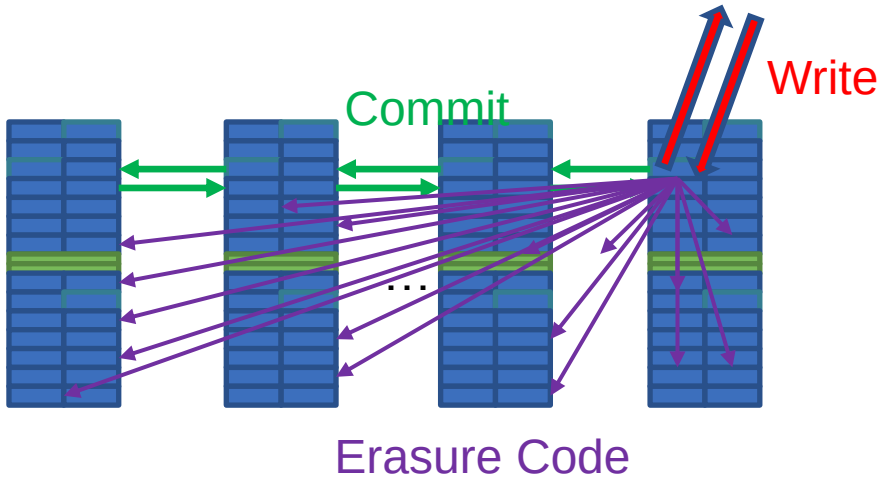


A complete virtual network needs storage as well as compute!

How do we make Azure Storage scale?

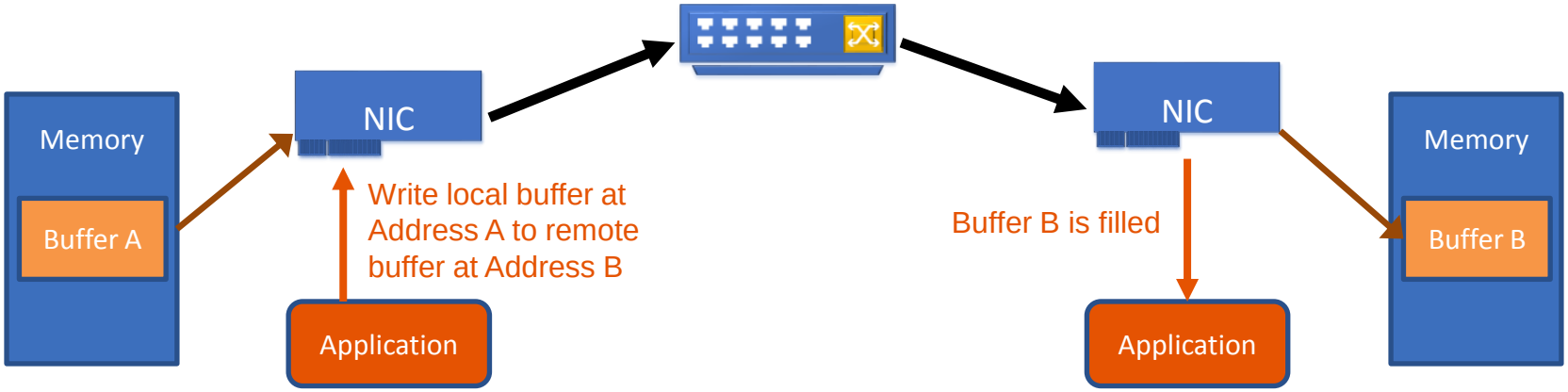
Azure Storage is Software Defined, Too

- 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
- **Requires amplified network I/O for each storage I/O**



To make storage cheaper, we use lots more network!

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

- RoCEv2 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

Software Interfaces

Azure RDMA Software

- Network Direct – high performance networking abstraction layer for IHVs
- Endure – Enlightened RDMA on Azure
- Linux Guest RDMA

What is Network Direct?

- Microsoft defined interfaces for RDMA
- Transparent support of IB, iWARP, and RoCE
- IP-addressing based

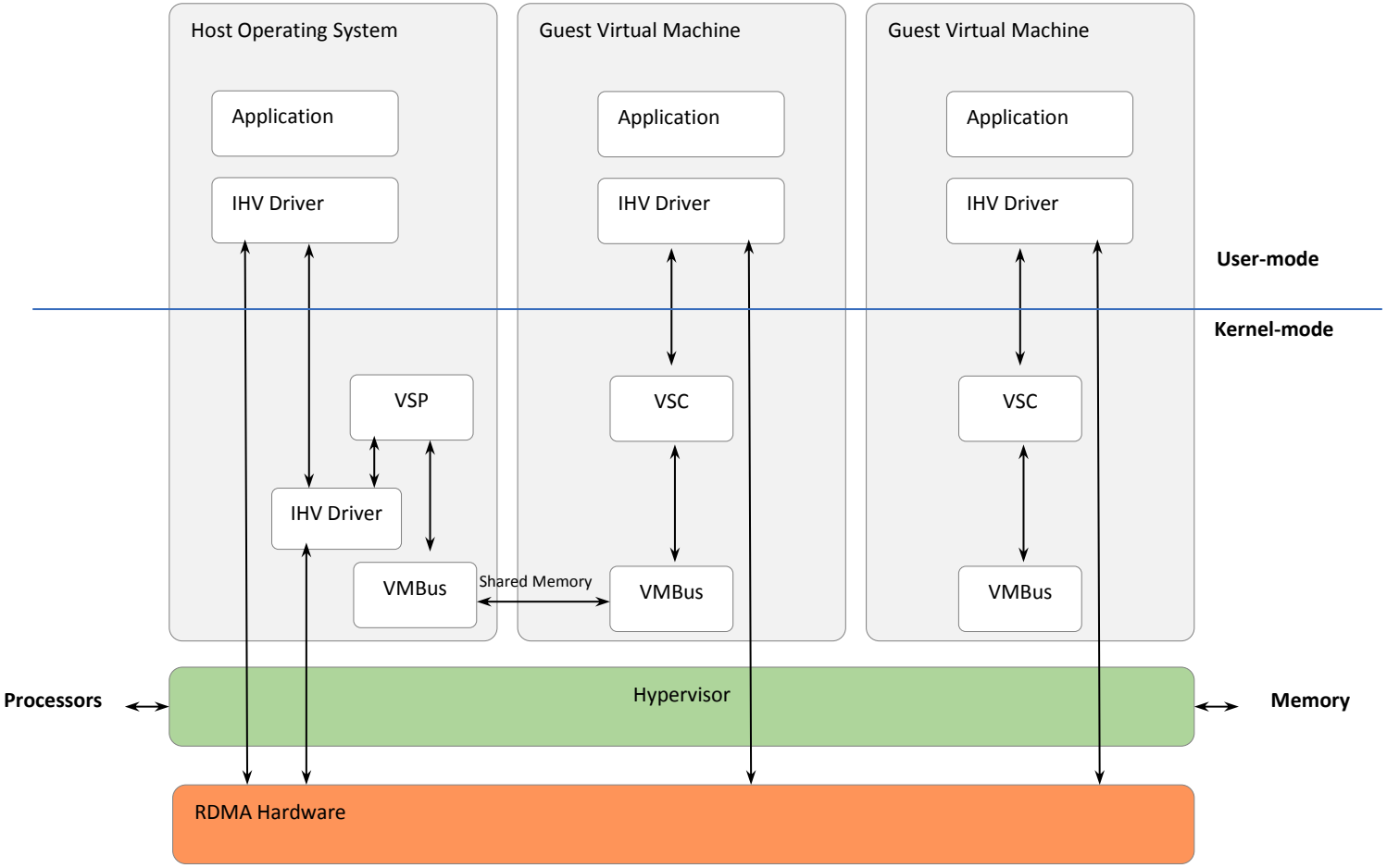
Why Network Direct?

- Designed for Windows
- Higher level of abstraction
- Fabric agnostic
- Stable ABI
- No Callbacks
- Dynamic provider discovery
- Extensible
- Easier to understand
- Easier to develop for

What is Endure

- Virtualization layer for RDMA
- Enables full bypass, with zero overhead to the VM
- Targets Azure-supported RDMA fabric(s)
- In production on Azure
  - Scalable to 1000's of cores
  - Now supporting Intel MPI
- Future – support Linux guest
  - Under development

Enlightened Network Direct On Azure



Linux Guest RDMA On Azure

- Based on Network Direct Technology
- All Control-plane operations over VMBUS
  - Connection establishment/teardown
  - Creation/management of all resources: PD, QP, CQ etc.
- Data-plane operations bypass the guest and host kernels
- RDMA NIC directly operates on pinned application memory
- New Linux driver bridges the gap between the Linux programming model and the Windows programming model
- No change to the Linux user-level environment

Network Direct Kernel Programming Interface (NDKPI)

- Derived from Network Direct
  - Specifically for kernel consumers
- Used by SMB Direct
- Versions
  - NDKPI 1.1 – Windows Server 2012
  - NDKPI 1.2 – Windows Server 2012R2
  - NDKPI 2.0 – Under development
- NDKPI Reference (resources, below)

Thank You

OPENFABRICS ALLIANCE

ANNUAL INTERNATIONAL OPENFABRICS SOFTWARE DEVELOPERS' WORKSHOP

Backup

SMB implementers (alphabetical order)

- Apple
  - MacOS X 10.2 Jaguar – CIFS/SMB 1.x (via Samba)
  - MacOS X 10.7 Lion – SMB 1.x (via Apple's SMBX)
  - MacOS X 10.9 Mavericks – SMB 2.1 (default file protocol)
  - MacOS X 10.10 Yosemite – SMB 3.0 (default file protocol)
- EMC
  - Older versions – CIFS/SMB 1.x
  - VNX – SMB 3.0
  - Isilon OneFS 6.5 – SMB 2
  - Isilon OneFS 7.0 – SMB 2.1
  - Isilon OneFS 7.1.1 – SMB 3.0
- Microsoft
  - Microsoft LAN Manager – SMB
  - Windows NT 4.0 – CIFS
  - Windows 2000, Server 2003 or Windows XP – SMB 1.x
  - Windows Server 2008 or Windows Vista – SMB 2
  - Windows Server 2008 R2 or Windows 7 – SMB 2.1
  - Windows Server 2012 or Windows 8 – SMB 3.0
  - Windows Server 2012 R2 or Windows 8.1 – SMB 3.0.2
  - Windows Technical Preview – SMB 3.1.1
- NetApp
  - Older versions – CIFS/SMB 1.x
  - Data ONTAP 7.3.1 – SMB 2
  - Data ONTAP 8.1 – SMB 2.1
  - Data ONTAP 8.2 – SMB 3.0
- Samba (Linux or others)
  - Older versions – CIFS/SMB 1.x
  - Samba 3.6 – SMB 2 (some SMB 2.1)
  - Samba 4.1 – SMB 3.0
- And many others…
  - Most widely implemented remote file protocol in the world, available in ~every NAS and File Server
  - See the SDC participants on slide 30

Information on this slide gathered from publicly available information as of February 2015.

Please contact the implementers directly to obtain the accurate, up-to-date information on their SMB implementation.

SMB 3.0.2

- Asymmetric Scale-Out File Server Clusters
  - SMB share ownership which can move within the File Server Cluster
  - Witness protocol enhanced to allow moving client per SMB share
  - In Windows, SMB clients automatically rebalance
- SMB Direct Remote Invalidation
  - Avoids specific invalidation operations, improving RDMA performance
  - Especially important for workloads with high rate of small IOs
- Unbuffered read/write operations
  - Per-request flags for read/write operations
- Remote Shared Virtual Disk Protocol
  - New protocol defines block semantics for shared virtual disk files
  - Implements SCSI over SMB (SMB protocol used as a transport)

SMB 3.1.1 (future)

- Under Development as of this presentation
  - More details in SNIA SDC 2014 presentations, and Microsoft protocol document previews
- Features include:
  - Extensible Negotiation
  - Preauthentication Integrity
    - Increased Man-in-the-Middle protection
  - Encryption improvements
    - Faster AES-128-GCM
  - Cluster improvements
    - Dialect rolling upgrade, Cluster Client Failover v2
- Related new and enhanced protocols in preview:
  - Storage Quality of Service (MS-SQOS)
  - Shared VHDX v2 (supporting virtual disk snapshots) (MS-RSVD)

Links to protocol documentation

Specification	Description
[MS-CIFS]: Common Internet File System (CIFS) Protocol Specification	Specifies the Common Internet File System (CIFS) Protocol, a cross-platform, transport-independent protocol that provides a mechanism for client systems to use file and print services made available by server systems over a network.
[MS-SMB]: Server Message Block (SMB) Protocol Specification	Specifies the Server Message Block (SMB) Protocol, which defines extensions to the existing Common Internet File System (CIFS) specification that have been implemented by Microsoft since the publication of the [CIFS] specification.
[MS-SMB2]: Server Message Block (SMB) Protocol Versions 2 and 3 Specification	Specifies the Server Message Block (SMB) Protocol Versions 2 and 3, which support the sharing of file and print resources between machines and extend the concepts from the Server Message Block Protocol.
[MS-SMBD]: SMB Remote Direct Memory Access (RDMA) Transport Protocol Specification	Specifies the SMB Remote Direct Memory Access (RDMA) Transport Protocol, a wrapper for the existing SMB protocol that allows SMB packets to be delivered over RDMA-capable transports such as iWARP or InfiniBand while utilizing the direct data placement (DDP) capabilities of these transports. Benefits include reduced CPU overhead, lower latency, and improved throughput.
[MS-SWN]: Service Witness Protocol Specification	Specifies the Service Witness Protocol, which enables an SMB clustered file server to notify SMB clients with prompt and explicit notifications about the failure or recovery of a network name and associated services.
[MS-FSRVP]: File Server Remote VSS Provider Protocol Specification	Specifies the File Server Remote VSS Protocol, an RPC-based protocol used for creating shadow copies of file shares on a remote computer, and for facilitating backup applications in performing application-consistent backup and restore of data on SMB shares.
[MS-RSVD]: Remote Shared Virtual Disk Protocol	Specifies the Remote Shared Virtual Disk Protocol, which supports accessing and manipulating virtual disks stored as files on an SMB3 file server. This protocol enables opening, querying, administering, reserving, reading, and writing the virtual disk objects, providing for flexible access by single or multiple consumers. It also provides for forwarding of SCSI operations, to be processed by the virtual disk.

Note: Protocols published by Microsoft, and available to anyone to implement in non-Windows platforms.
http://www.microsoft.com/openspecifications/

SNIA SMB2/SMB3 Plugfest

OPENFABRICS ALLIANCE

- SMB/SMB2/SMB3 Plugfest happens every year side-by-side with the Storage Developer Conference (SNIA SDC) in September
- Intense week of interaction across operating systems and SMB implementations.

Agenda, Calendar and past content at:

http://www.snia.org/events/storage-developer

Participants in the 2014 edition of the
SNIA SMB2 / SMB3 Plugfest
Santa Clara, CA – September 2014

SMB3 / SMB Direct / RDMA Resources

- Microsoft protocol documentation
  - http://www.microsoft.com/openspecifications
- SNIA education SMB3 tutorial
  - http://www.snia.org/education/tutorials/FAST2015
- SNIA SDC 2014
  - http://www.snia.org/events/storage-developer/presentations14#smb

More SMB3 Resources

- Jose Barreto's blog
  - http://smb3.info/
  - The Rosetta Stone: **Updated Links on Windows Server 2012 R2 File Server and SMB 3.02**
    - http://blogs.technet.com/b/josebda/archive/2014/03/30/updated-links-on-windows-server-2012-r2-file-server-and-smb-3-0.aspx
  - Performance Demo
    - http://blogs.technet.com/b/josebda/archive/2014/03/09/smb-direct-and-rdma-performance-demo-from-teched-includes-summary-powershell-scripts-and-links.aspx
- Microsoft Technet
  - Improve Performance of a File Server with SMB Direct
    - http://technet.microsoft.com/en-us/library/jj134210.aspx
- Windows Kernel RDMA interface
  - NDKPI Reference (provider)
    - http://msdn.microsoft.com/en-us/library/windows/hardware/jj206456.aspx

Performance: SMB3 with RDMA on 40GbE
CONFIGURATION

- Arista Switches
  - Interconnect constrained to single 40GbE link
- Chelsio T580 2-port 40GbE iWARP
  - Single Port Connected
- Server Chassis
  - 2x Intel® Xeon processor E5-2660 (2.20 Ghz)
- I/O not written to non-volatile storage