SMB Direct Update

Tom Talpey and Greg Kramer
Microsoft

Outline

- Part I – Ecosystem status and updates
  - SMB 3.02 status
  - SMB Direct applications
  - RDMA protocols and networks
- Part II – SMB Direct details
  - Protocol enhancements
  - Performance results
  - Shock and Awe

Protocols and updates

- SMB 3.02
  - Minor update
    - *SMB Direct remote invalidation*
    - Asymmetric shares
    - Unbuffered read and write operations
  - Documented in MS-SMB2 and MS-SMBD
  - Details yesterday in David Kruse's talk
- SMB Direct
  - Specifies SMB3 RDMA transport
  - Supported by SMB 3.0 and SMB 3.02

Windows Server 2012 R2

- General Availability in October
- Downloadable now for evaluation
- Supports new SMB 3.02 and SMB Direct

Applications using SMB3 and SMB Direct

- 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
    - See Jose Barreto and Matt Kurjanowicz talk Thursday

RDMA transports

- Windows Server 2012 and 2012 R2 support:
  - iWARP (IETF RDMA over TCP)
  - InfiniBand
  - RoCE (RDMA over Converged Ethernet)
- Ethernet:
  - iWARP and RoCE – 10 or 40 Gb
- InfiniBand:
  - Effective – 32 Gb (QDR) or 54 Gb (FDR)

RDMA evolution

- iWARP
  - IETF Standard
  - Extensions currently active in IETF
- RoCE
  - “Routable RoCE” to improve scale
  - DCB deployment (still problematic)
- InfiniBand
  - Roadmap to 100 Gb

RDMA attributes

- iWARP
  - Ethernet, routable
  - No special fabric requirements
  - Up to 40 Gb with good latency and full throughput
- RoCE
  - Ethernet, not routable (but watch this space)
  - Requires Priority Flow Control in fabric (DCB)
  - Up to 40 Gb with good latency and full throughput
- InfiniBand
  - Specialized interconnect, not routable
  - Dedicated fabric and switching
  - Up to 54 Gb with excellent latency and full throughput

SMB3 Services

- Connection management
  - Dialect negotiation, validation
- Authentication
  - Integrity (signing) and/or privacy (encryption)
- Multichannel
  - Provides both trunking/bandwidth and availability
- Resilience and recovery to network failure
- RDMA
- File I/O semantics (Win32, and others)
- Control and Extension semantics
  - Filesystem and IOCTL passthrough
- Remote filesystem access
  - NTFS, VHD, Named Pipes, Memory, …
- RPC

The ISO 7-layer model

- SMB3 presents new value as a **Session layer**
  - Access to:
    - RDMA
    - Multichannel
    - Replay/recovery
- Moving "up the stack" from SMB2
  - Not limited to transport

Layer	Applicability
Application	Hyper-V, SQL, etc.
Presentation	SMB3 File services, RPC, Live Migration, etc.
Session	**SMB3 authentication, multichannel and availability**
Transport	SMB2/3 transport and TCP or RDMA (or NetBIOS ☺)
Network	IP or InfiniBand/RoCE
Link	Ethernet or InfiniBand
Physical	Ethernet or InfiniBand

SMB3 as a Session Layer

- Can see SMB3 as providing multiple services to applications
  - Network transparency
    - Abstracts away the network and lower layers
    - Fully transparent RDMA access!
  - Performance
    - Scalable bandwidth via multichannel and RDMA
  - Recovery
    - Recovers state after network loss
    - Transparently supports clustered servers for higher availability
  - Protection
    - Signing, encryption, integration with Active Directory

Other Possible SMB Attributes

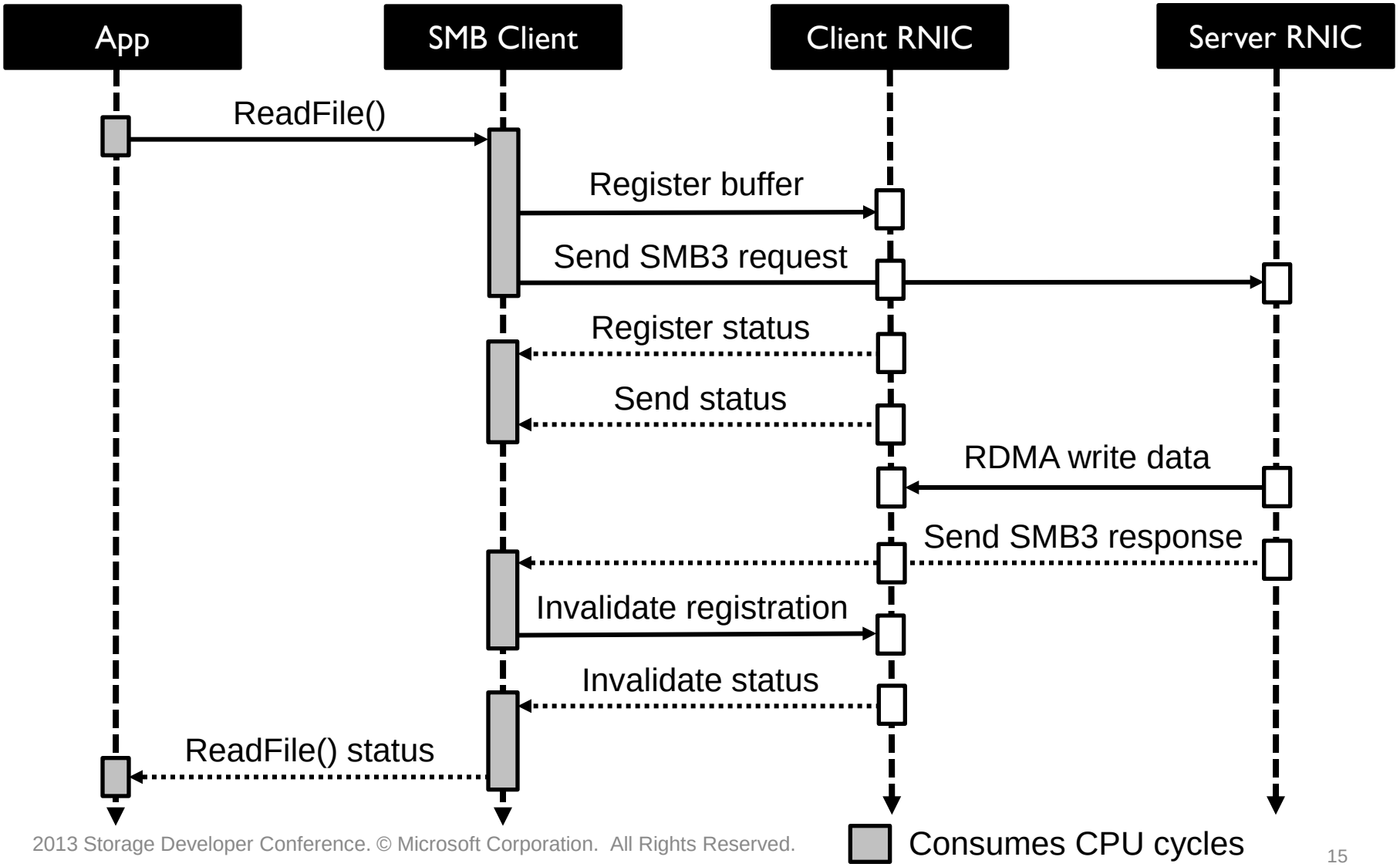
- Greater use by clustering
  - SMB3 for Server-to-Server
- Quality of Service
  - SMB storage traffic limits and rate control
- Cloud deployment

In Summary

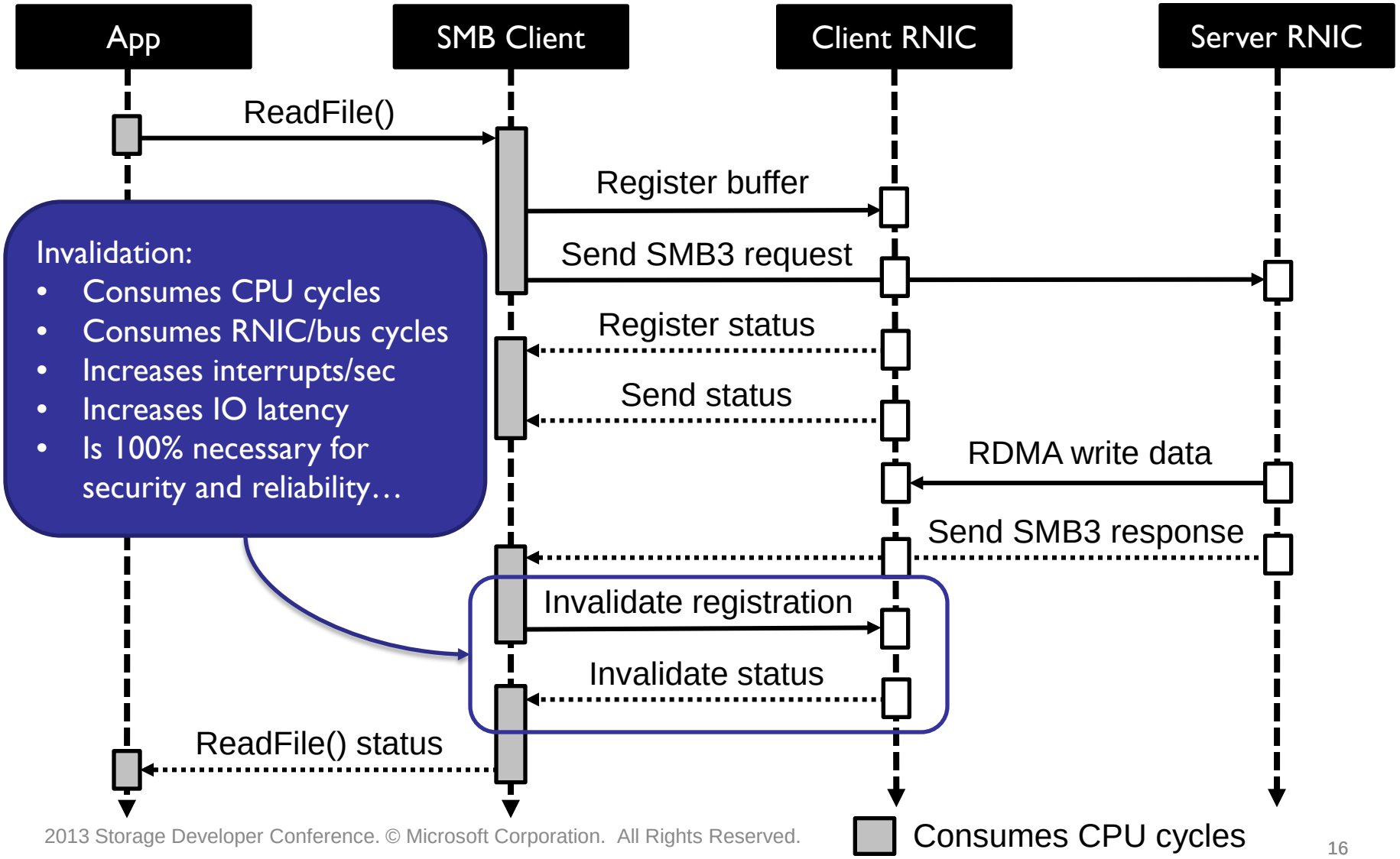
- Look to SMB3 for even broader application
  - Beyond filesystem services
    - For example, WS2012R2 Hyper-V Live Migration
- Look to use of SMB Direct to broaden
  - Transparent SMB3 application use
- Look to see greater application “fidelity”
  - Increasingly sophisticated applications transparently served by SMB3
    - (e.g. clustered, highly available)

Protocol Enhancements and Performance Results

Where can we reduce IO costs?



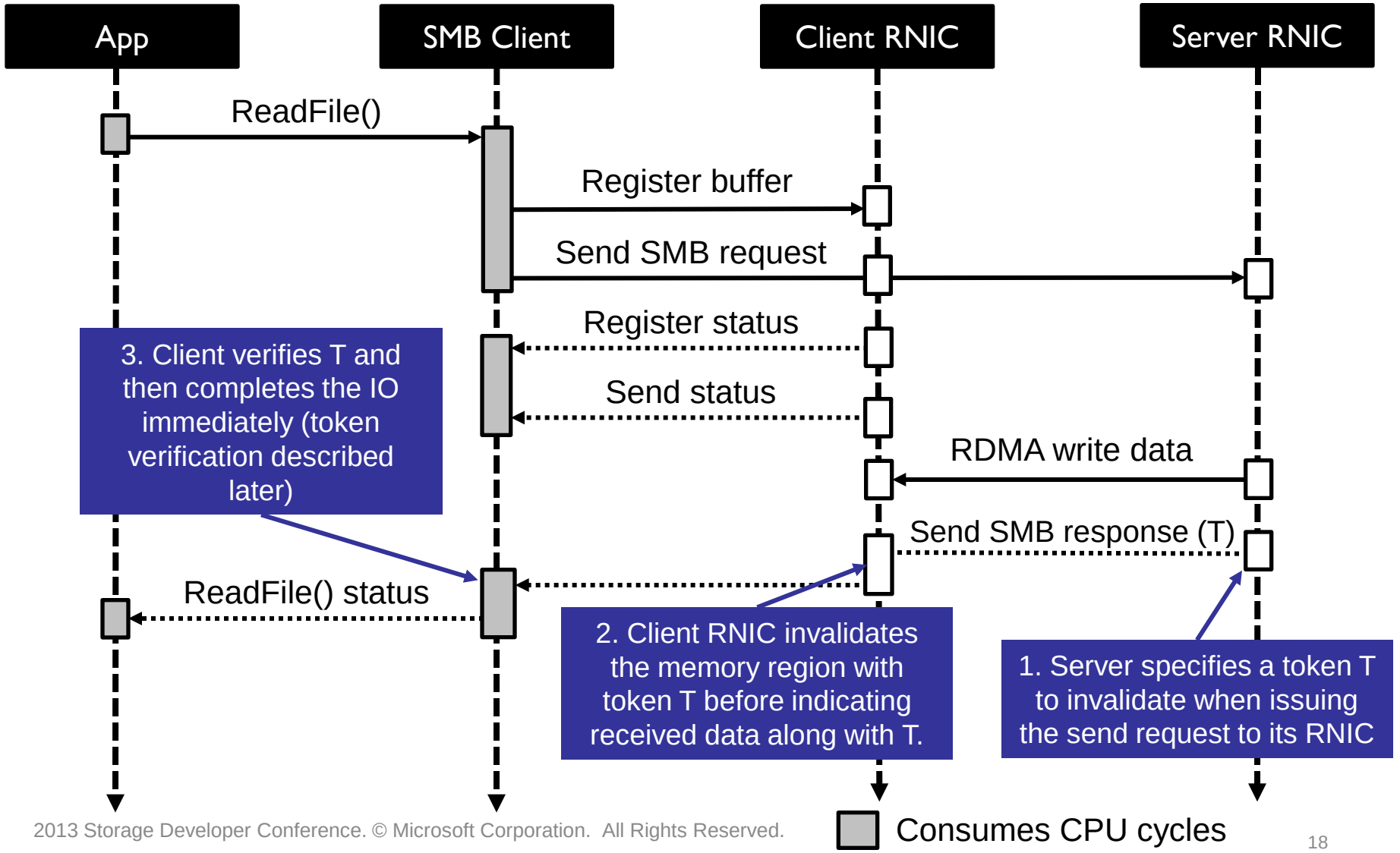
Where can we reduce IO costs?



Why pend IO until invalidation has completed?

- Invalidation guarantees that:
  - Data is in a consistent state following DMA
  - Peer no longer has access to the buffer
    - Registration caches *cannot* provide this guarantee, leading to danger of data corruption / OS crashes due to buggy / malicious peer-initiated RDMA operations!
- Invalidation is necessary, but we can reduce its cost via Send and Invalidate.

Send and Invalidate



Send and Invalidate Benefits

- Reduces RNIC work requests per IO by 1/3rd for high IOP workloads!
- Already supported by major RDMA standards
  - iWARP
  - InfiniBand
  - RoCE
- Requires only minimal protocol changes

SMB Direct Protocol Changes

This page intentionally left blank

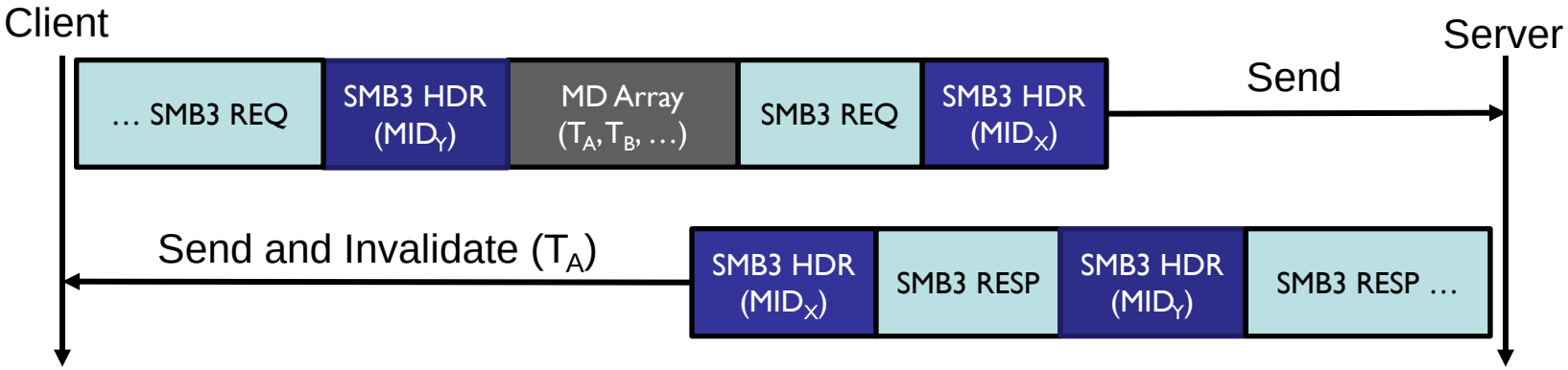
SMB 3.02 Protocol Changes

SMB 3.02 Read and Write requests gain a new Channel value:

Value	Meaning
SMB2_CHANNEL_NONE 0x00000000	No channel information is present in the request. The ReadChannelInfoOffset and ReadChannelInfoLength fields MUST be set to 0 by the client and MUST be ignored by the server.
SMB2_CHANNEL_RDMA_V1 0x00000001	One or more SMB_DIRECT_BUFFER_DESCRIPTOR_V1 structures as specified in [MS-SMBD] section 2.2.3.1 are present in the channel information specified by ReadChannelInfoOffset and ReadChannelInfoLength fields.
SMB2_CHANNEL_RDMA_V1_INVALIDATE 0x00000002	This value is valid only for the SMB 3.02 dialect. One or more SMB_DIRECT_BUFFER_DESCRIPTOR_V1 structures, as specified in [MS-SMBD] section 2.2.3.1, are present in the channel information specified by the ReadChannelInfoOffset and ReadChannelInfoLength fields. The server is requested to perform remote invalidation when responding to the request as specified in [MS-SMBD] section 3.1.4.2.

Using Send and Invalidate (Server)

- The MID of the first SMB3 response in the sent payload must match the MID of the SMB3 request that is associated with the remotely invalidated token.
- Only the first memory descriptor in a SMB3 read / write request's memory descriptor array may be remotely invalidated.



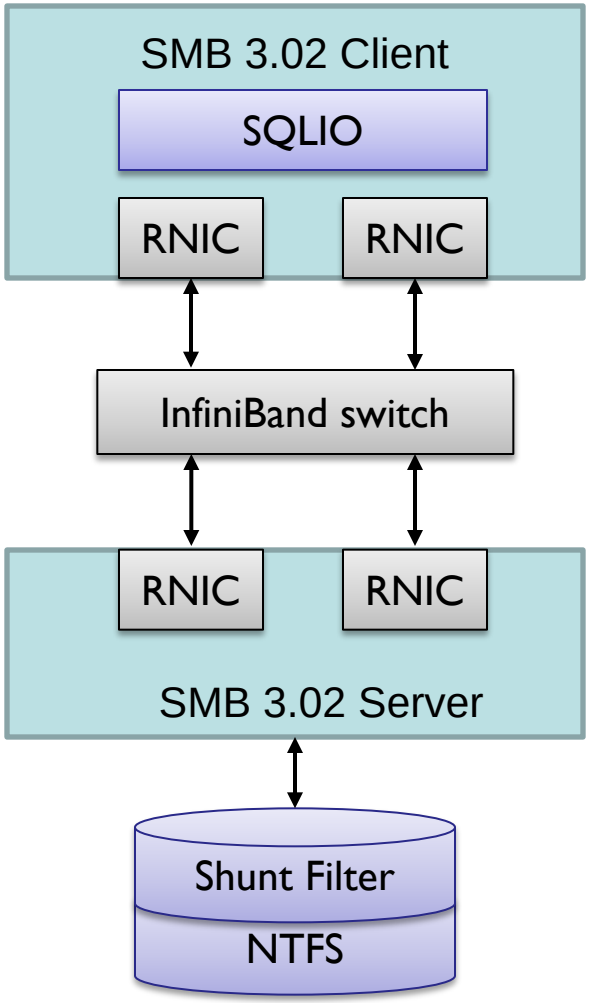
Using Send and Invalidate (Client)

- Client must manually invalidate all memory regions that are not remotely invalidated.
  - The server is not obligated to remotely invalidate and can only invalidate one MR per request.
- Client must verify that a remotely invalidated token for a response with MID=x belongs to the request with MID=x.
  - Required to meet invalidation guarantees about DMA data consistency and peer's inability to further access the memory.

Performance Results

I feel the need for speed…

Benchmark Configuration



Client / Server

2 x Intel Xeon E5-2660 CPUs @ 2.2 GHz – 16 total logical processors (HT disabled)

2 x Mellanox ConnectX-3 56 Gbps InfiniBand RNICs (one active port per RNIC)

Storage / Workload

Server exports a shunted NTFS volume via an SMB3 share. ShuntFilter is an MS-internal file system filter driver that completes read/write IO immediately to simulate fast storage (let's see how fast SMB3/SMBDirect is when storage costs aren't the bottleneck)

SQLIO (http://www.microsoft.com/en-us/download/details.aspx?displaylang=en&id=20163) used to generate IOs with one thread per logical CPU performing async IO against a file on the server.

1 KiB Random IO*

Reads (unbuffered)

IOs per thread	WS 2012 IO/sec	WS 2012 R2 IO/sec	Δ IO/sec	Δ Client CPU/IO	Δ Server CPU/IO
64	783,656	881,499	+12.5%	-17.3%	-36.7%

Writes (unbuffered + write-through)

IOs per thread	WS 2012 IO/sec	WS 2012 R2 IO/sec	Δ IO/sec	Δ Client CPU/IO	Δ Server CPU/IO
64	711,970	808,248	+13.5%	-16.0%	-32.7%

* send/recv

8 KiB Random IO

Reads (unbuffered)

IOs per thread	WS 2012 IO/sec	WS 2012 R2 IO/sec	Δ IO/sec	Δ Client CPU/IO	Δ Server CPU/IO
64	583,108	835,688*	+43.3%	-37.1%	-33.2%

Writes (unbuffered + write-through)

IOs per thread	WS 2012 IO/sec	WS 2012 R2 IO/sec	Δ IO/sec	Δ Client CPU/IO	Δ Server CPU/IO
64	547,224	712,578	+30.2%	-26.0%	-14.9%

* > 1 million 8KiB random read IOPs when client/server upgraded to faster 2.7 GHz CPUs.

512 KiB Sequential IO

Reads (unbuffered)

IOs per thread	WS 2012 MBytes/sec	WS 2012 R2 MBytes/sec	Δ IO/sec	Δ Client CPU/IO	Δ Server CPU/IO
12	10,702	11,366	+6.2%	-9.3%	-14.3%

Writes (unbuffered + write-through)

IOs per thread	WS 2012 MBytes/sec	WS 2012 R2 MBytes/sec	Δ IO/sec	Δ Client CPU/IO	Δ Server CPU/IO
12	10,769	11,412	+6.0%	-12.2%	-10.3%

Performance Results Summary

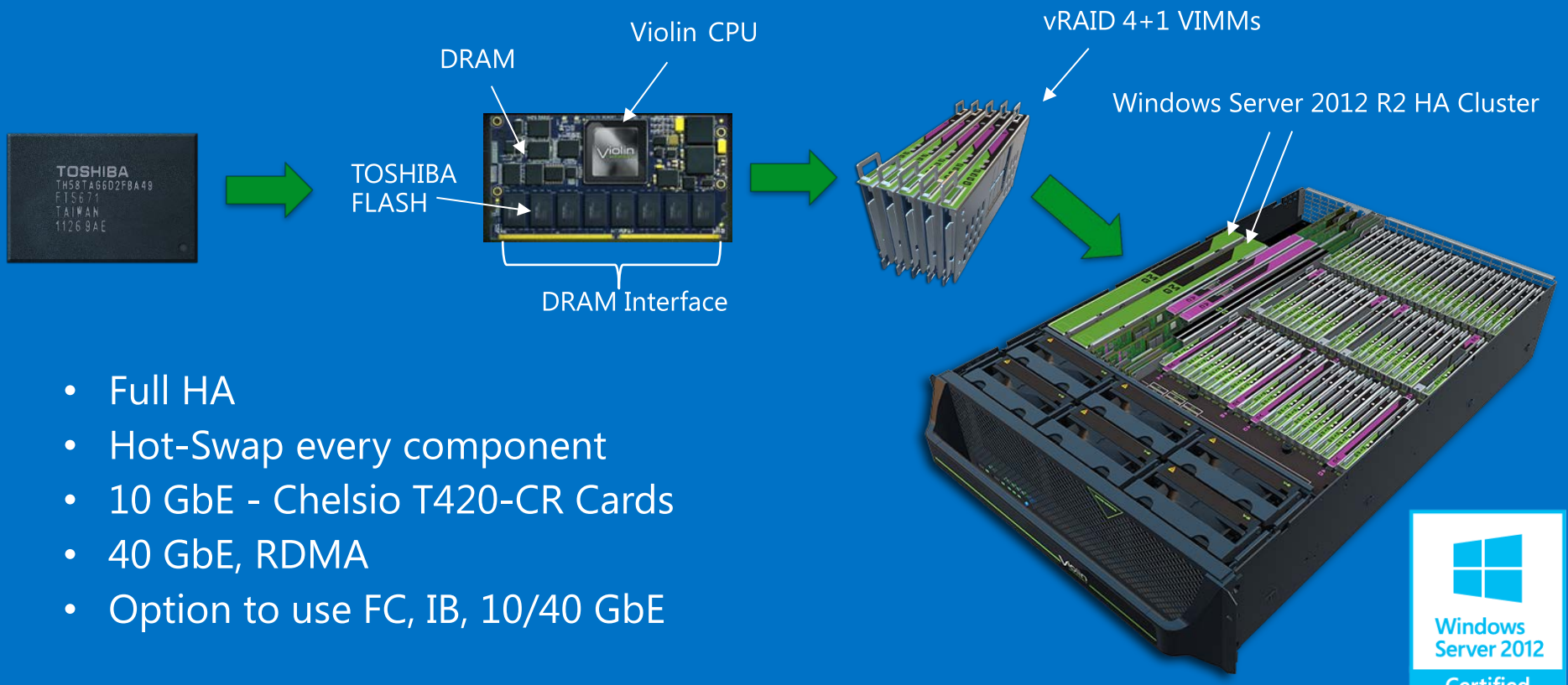
- Increased IOPS (12% - 43%) and higher bandwidth
- Decreased CPU per IO (15% - 36%)
  - Client has more CPU for applications
  - Server scales to more clients
- No increase in RDMA resources consumed per connection
- No new hardware required (send and invalidate is already supported by existing RNICs)

Performance Results Summary

- WS 2012 R2 results reflect the untuned, out-of-the-box customer experience.
  - Tuning for specific workloads / hardware increases performance.
  - WS 2012 results reflect hand-tuning to improve performance vs. WS 2012 R2 (double the number of memory regions per connection)

One more thing…

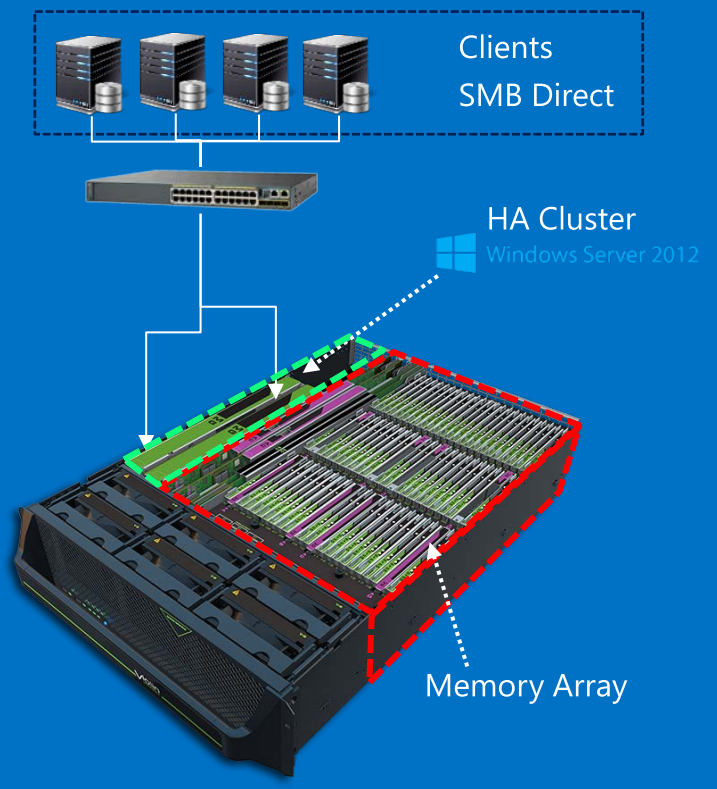
Violin Memory Windows Storage Server Array



- Full HA
- Hot-Swap every component
- 10 GbE - Chelsio T420-CR Cards
- 40 GbE, RDMA
- Option to use FC, IB, 10/40 GbE

Windows Server 2012 Certified

Violin Memory Array Configuration



Performance:

- 100% Reads – 4KiB: >1Million IOPS
- 100% Reads – 8KiB: >500K IOPS
- 100% Writes – 4KiB: >600K IOPS
- 100% Writes – 8KiB: >300K IOPS

Configuration as tested:

V6616 - SLC
- 4 x Dual Port Mellanox ConnectX-3 Cards
- 2 x Internal Gateways
  - 8-Core Sandy Bridge at 1.8 GHz
  - Windows Server 2012 R2
  - 48GB DRAM
- Failover Cluster
- 8 1TB Shares exported – 2 Per Client

Interconnect
- 40 GbE – RoCE RDMA
- SMB Direct

4 External Clients
- Quad Core Xeon – 2.53 GHz
- 24 GB DRAM
- 2 x Dual Port Mellanox ConnectX-3
- Windows Server 2012 R2, SQLIO

Planned configuration for GA:
- MLC: 64TB, 32TB, 12TB
- SLC: 16TB

*Samples and POC gear available immediately

Questions?