



Education

SMB2 – Big Improvements in the Remote Filesystems Protocol

James Pinkerton, Microsoft

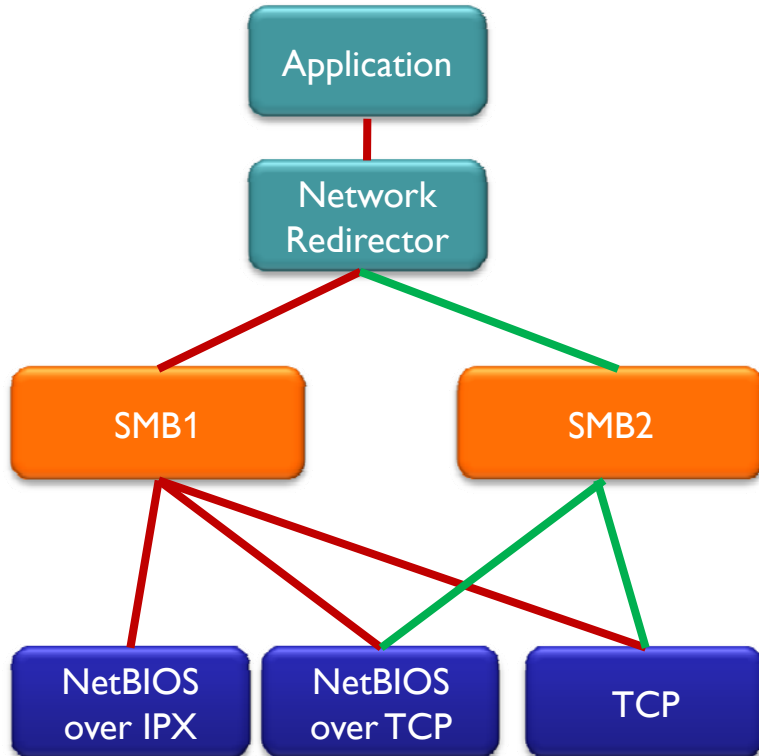
- The material contained in this tutorial is copyrighted by the SNIA.
 - Member companies and individuals may use this material in presentations and literature under the following conditions:
 - ◆ Any slide or slides used must be reproduced without modification
 - ◆ The SNIA must be acknowledged as source of any material used in the body of any document containing material from these presentations.
 - This presentation is a project of the SNIA Education Committee.
 - Neither the Author nor the Presenter is an attorney and nothing in this presentation is intended to be nor should be construed as legal advice or opinion. If you need legal advice or legal opinion please contact an attorney.
 - The information presented herein represents the Author's personal opinion and current understanding of the issues involved. The Author, the Presenter, and the SNIA do not assume any responsibility or liability for damages arising out of any reliance on or use of this information.
- NO WARRANTIES, EXPRESS OR IMPLIED. USE AT YOUR OWN RISK.**

- **SMB2 – A New Remote Filesystem Protocol**
 - ◆ This session will appeal to File System Managers, Developers, IT administrators, and those that are seeking an understanding of how the new SMB2 protocol will behave fundamentally different than the old SMB (a.k.a. CIFS) protocol.
 - ◆ This session will begin by providing a brief overview of the SMB2 protocol, and then walk through a few common scenarios to familiarize the audience with how the protocol is fundamentally different than SMB1.

- The SMB/CIFS/SMB2 protocols are *de facto* standards
 - ◆ Specifications are owned by Microsoft Corporation
 - ◆ Protocol documentation available at:
 - <http://msdn.microsoft.com/en-us/library/cc216517.aspx>
- Multiple interoperable implementations, on different file systems
 - ◆ Samba, Linux, FreeBSD, NetBSD, JCIFS, ...
 - ◆ Network Appliance, EMC, Apple, Sun, ...
 - ◆ And many more
- SMB/CIFS/SMB2 – A stateful remote file access protocol
 - ◆ “Stateful” – the client maintains state of open files to improve performance.
 - NFSv2 and NFSv3 are stateless. NFSv4 is stateful.
 - ◆ “oplocks” – a mechanism to enable the client to cache operations locally for better performance.

- SMB1/CIFS fundamental design goes back to 1983
 - ◆ 25 years ago...
 - › LAN connectivity was 10 Mbits/sec
 - › Essentially no WAN, no Wireless LAN (WLAN)
 - › Man-in-the-middle attacks non-existent
 - ◆ Over time, common scenarios changed
 - › Incremental features attempted to address issues, but did not keep up
- Enter SMB2
 - ◆ <http://msdn.microsoft.com/en-us/library/cc212614.aspx>
 - ◆ New version of SMB, which substantially simplifies the protocol while also improving security, network fault tolerance, and performance for high speed LAN and WAN

SMB1/SMB2 Transports



- SMB1 today supports 3 primary transports
 - ◆ NetBIOS support is primarily for legacy reasons and interoperation
- SMB2 utilizes the SMB1 connection, and negotiates “up” if both nodes support SMB2
 - ◆ Backwards compatible if server does not support SMB2
- SMB2 does not support NetBIOS over IPX
- Preferred transport for both SMB1 and SMB2 is TCP

SMB1 also supports NetBEUI and NetBIOS over UDP, however they are strongly deprecated (and not shown).

This isn't your father's SMB...

➤ IT Goal: Transfer 10.7 GB over 76 ms, 1 gigabit WAN as fast as possible

◆ First Direction:

- > SMB: it took **5 hours, 40 minutes, 13 seconds**. Rate: 0.56 MB/s
- > SMB2: it took **7 minutes** and 45 seconds. Rate: 25 MB/s

◆ Opposite Direction:

- > SMB: it took **6 hours, 6 minutes, 26 seconds**. Rate: 0.52 MB/s
- > SMB2: it was **8 minutes, 10 seconds**. Rate: 23 MB/s

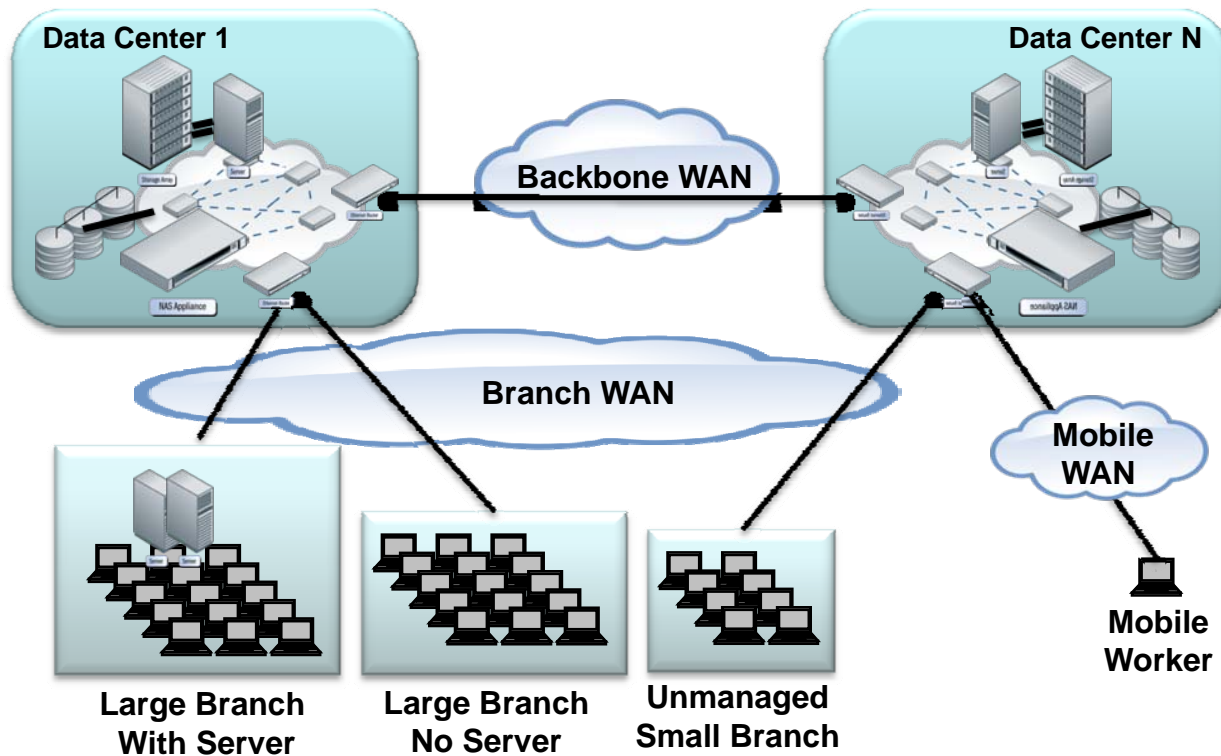
◆ ~**45 times faster** compared to SMB1 over WAN

➤ User Goal: Run robustly on Wireless LAN

- ◆ Using SMB1, long file transfers can abort due to temporary WLAN outage due to interference
- ◆ Using SMB2, temporary network outages are transparently recovered.

Today's Large Enterprise

- ▶ LAN connectivity changed to between 100 and 10,000 Mbits/sec
- ▶ WAN connectivity within enterprises is common
 - ◆ Huge growth in branch offices (medium bandwidth, high latency)
 - ◆ Huge growth in the number of data centers (high bandwidth, high latency)
- ▶ WLAN connectivity is common (and intermittent!)



➤ A Branch Office:

- ◆ 155Mb/s bandwidth
- ◆ 50ms link to datacenter
- ◆ 300 client machines



➤ A Branch Office:

- ◆ 2 Mb/s bandwidth
- ◆ 1200 ms latency link to datacenter
- ◆ 1 client machine



➤ Conclusions:

- ◆ Branch network connectivity varies substantially
- ◆ Need an automated mechanisms to adjust the protocol

➤ SMB1 Limitations:

- ◆ Designed to be “chatty”, with later “compounding” of multiple requests into a single higher level command
 - › “compounding” – putting multiple commands in a single network packet
- ◆ Limited request pipelining led to poor MAN / WAN performance
- ◆ Limited number of open files, number of shares etc.
- ◆ Difficult to extend, maintain and secure due to protocol complexity

➤ SMB2 Design Goals:

- ◆ Simplification of the protocol combined with general mechanism for compounding to decrease the number of round trips
- ◆ Enable extremely deep pipelining of data transfer for WAN, high speed LAN, without causing errant timeouts or unresponsiveness
- ◆ Enable multiple users traffic to queue data independently of each other
- ◆ Build a solid foundation for continued innovation

- Two general compounding mechanisms can collapse multiple requests into a single packet (and single round trip)
 - ◆ Group commands that must be executed in order
 - ◆ Group unrelated commands (no ordering requirements)
- Designed for highly parallel data transfer, which can automatically scale to meet the needs of the application
 - ◆ Important to balance trade-offs between memory consumption, speed of disk, speed of network, desired transfer rate, application sophistication
 - ◆ Optimized for high speed LAN or WAN
- Intended to be a single connection per user which scales amount of data subject to timeouts as a function of the network performance

➤ Secure & Robust

- ◆ Durable handles to reconnect on temporary loss of network connectivity
 - Allows application handle to survive a network disconnect/reconnect
- ◆ Message signing improved & simplified
 - All implementations must support signing - If client & server settings differ, signing is used by default
 - Moved from MD-5 signing algorithm to more robust SHA-256
 - Signing is per user, not per client machine
- ◆ Reduced attack surface and implementation complexity due to a smaller command set

➤ Miscellaneous

- ◆ Symbolic link support
- ◆ Support for larger reads/writes

➤ SMB1 – history goes back to 1983

- ◆ SMB/CIFS Name Exodus: SMB -> CIFS -> SMB -> SMBv2
- ◆ CIFS = SMB as shipped by Microsoft in NT4 server
 - › Created new name for version of protocol submitted to IETF
 - › Widely adopted by file server vendors
 - › Extended in Unix community for Unix specific requirements (security)
- ◆ Post CIFS improvements
 - › Kerberos and domains, Shadow copy, Server-to-Server copy, SMB signing

➤ SMB2 Versions

- ◆ SMB 2.000 – Only shipped in beta form
- ◆ SMB 2.001 – Initial implementation of SMB2, deprecated
- ◆ SMB 2.002 – Current version of SMB2

Scaling for WAN or high speed LAN

**but first,
a primer on Bandwidth Delay Product**

- TCP has shown that a reasonable rule of thumb for “filling the pipe” is that all layers must support posting enough data to fill the network for the full amount of time to send the data and receive an acknowledgement

- › Bandwidth-Delay Product (BDP) = Bandwidth of the link * RTT
- › Percent Network Utilization = Amount of Outstanding Data / BDP

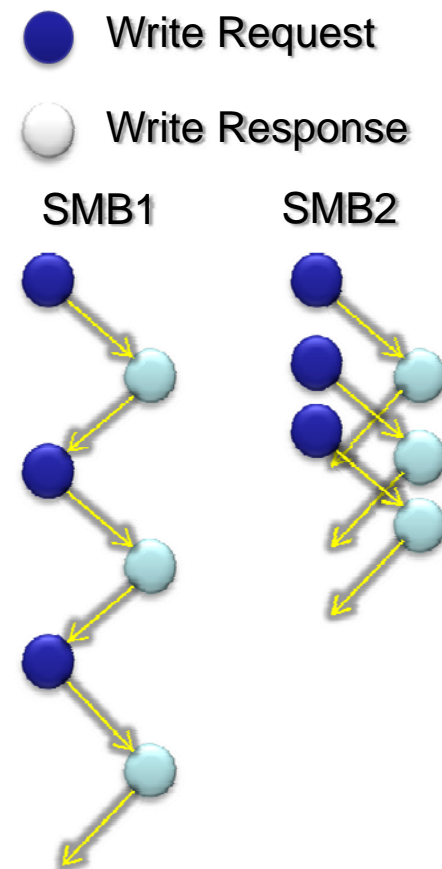
➤ Examples:

- ◆ Branch: 100 Mb/s, 100 ms RTT
 - › $BDP = 10^8/8 * 0.1 = \mathbf{1.25 MB}$
- ◆ Branch: 100 Mb/s, 500 ms RTT
 - › $BDP = 10^8/8 * 0.5 = \mathbf{6.25 MB}$
- ◆ Backbone: 622 Mb/s, 100 ms RTT
 - › $BDP = 6.22 * 10^8/8 * 0.1 = \mathbf{7.7 MB}$
- ◆ LAN: 1000 Mb/s, 1 ms RTT
 - › $BDP = 10^9/8 * 0.001 = \mathbf{125 KB}$

- BDP must grow for all layers – TCP, SMB, Application
 - ◆ TCP - Some network stacks by default allow 64 KB BDP, some allow scaling to 16 MB (or more)
 - ◆ SMB – scales in units of “Protocol Data Units” – often 64 KB
 - › SMB Versions:
 - SMB1 – Some implementations only allow small number of PDUs (single digit)
 - SMB2 – Designed to scale to extremely large number of PDUs (hundreds)
 - › Examples:
 - If BDP = 64 KB, need at least one SMB PDU outstanding at a time
 - If BDP = 8 MB, need at least 128 SMB PDUs outstanding at a time
 - ◆ Application – some applications are not optimized for high BDP networks – they don’t post enough data
 - › Example: A File Copy tool posts one buffer at a time, of size 64 KB

SMB 2 – File Copy Performance

- File Copy performance seen in the real world much faster than SMB1
 - ◆ Up to ~45x throughput for WAN
 - ◆ Up to 2-10x throughput for LAN
- Benefits due to:
 - ◆ SMB2 request pipelining
 - ◆ SMB2 larger request support
 - ◆ TCP stack improvements
 - ◆ Copy file library improvements
 - > Large buffers
 - > Asynchronous, non-cached IO



Reduction in “Chattiness”

(Simplification of the protocol enables sophisticated compounding of operations to reduce round trips)

SMB1 Complexity vs. SMB2

- SMB1 contained over 100 commands (including subcommands)
- SMB2 contains 19 commands
- Example: Contrasting mechanisms to open a file

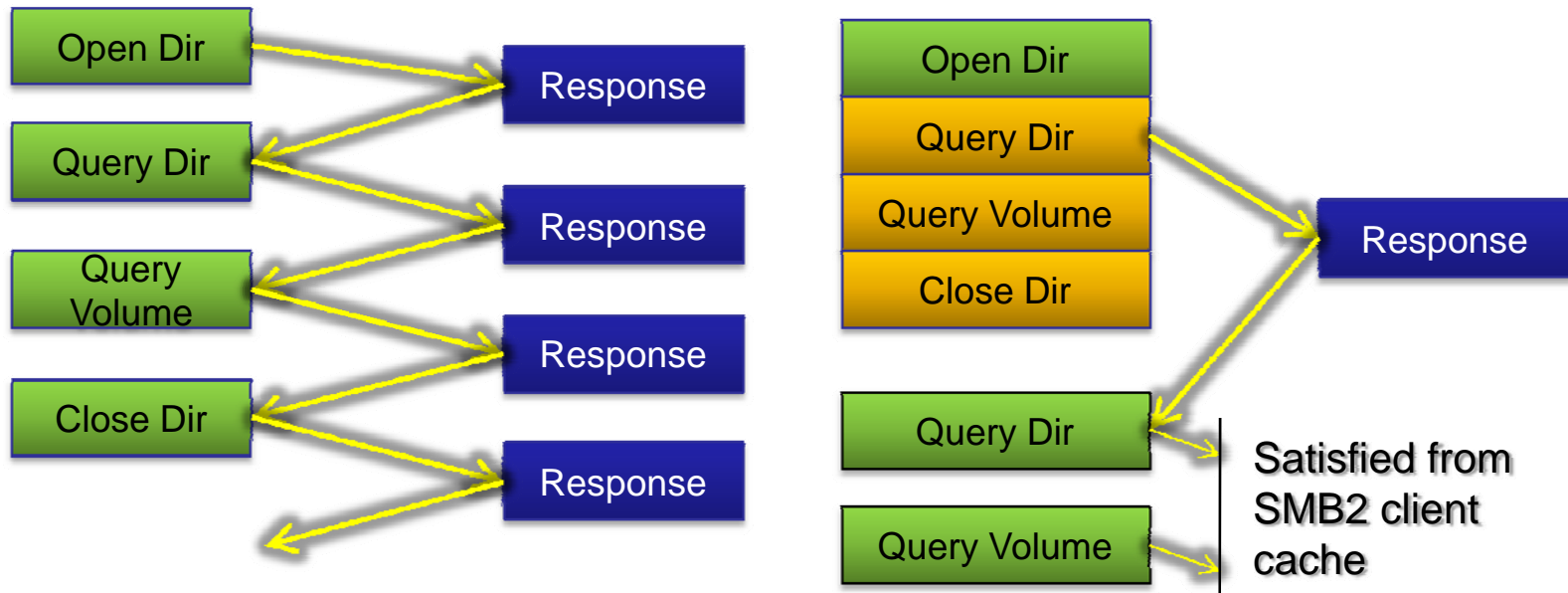
SMB1 Open Requests	
Explicit Opens	Implicit Opens
NT_CREATE_ANDX	RENAME
NT_TRANSACT_CREATE	MOVE
CREATE_TEMPORARY	COPY
OPEN_PRINT_FILE	QUERY_PATH_INFO
CREATE	SET_PATH_INFO
CREATE_DIRECTORY	DELETE_DIRECTORY
CREATE_NEW	CHECK_DIRECTORY
OPEN	TRANS2_FIND_FIRST2
OPEN_ANDX	SEARCH
TRANS_OPEN2	

SMB2 Open Requests
Create

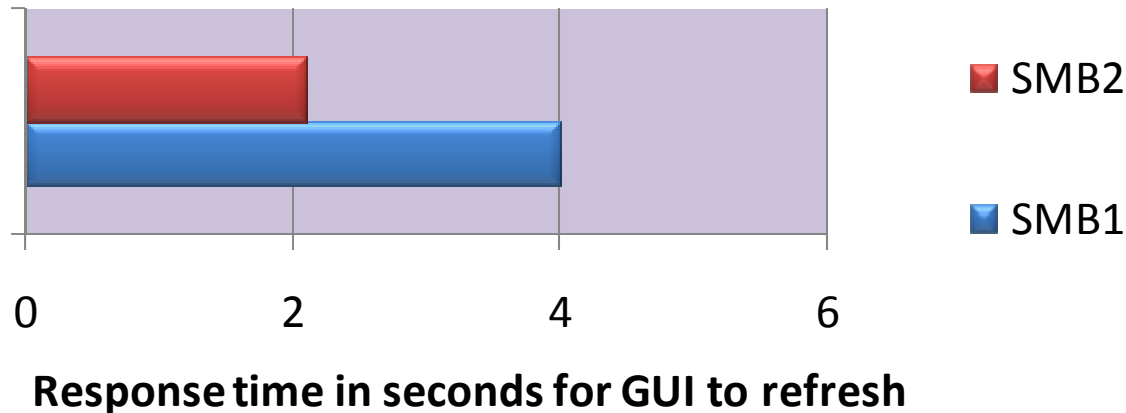
+ Compounding of additional operations into a single packet

SMB 2 - Compounding with Caching

- A common request sequence (green boxes)
 - ◆ Left side shows resulting client-server requests without compounding.
 - ◆ Right side shows resulting client-server requests with compounding
 - › SMB2 client speculatively generates the yellow requests and caches the results
 - ◆ Example below collapses 4 round trips into a single round trip
- Higher latency networks see greater benefits from SMB2



- Strong improvements in GUI directory enumeration
- Example below:
 - ◆ Opening a directory with 50 files through a GUI
 - ◆ Network – 1Gb/s, 100ms RTT
 - ◆ Both client and server running same OS



➤ Wireless networks can have spurious disconnects

- ◆ SMB1 – Large file transfers can be difficult on WLAN, WWAN
 - › TCP connection loss aborts transfer. User must restart entire transfer
- ◆ SMB2 – Durable Handles make large file transfers robust
 - › SMB2 can preserve session state even if the TCP connection is lost
 - › Automatic reconnect when network returns, starts where left off

➤ Multi-user clients

- ◆ SMB1 used a single connection per client
 - › All users/applications share a “queue” – thus head-of-line blocking issues, particularly for slow networks
 - › Signing has a security vulnerability – signing certificate for the life of the connection was that of the first user that was authenticated
- ◆ SMB2 uses a single connection per user
 - › Each user’s data on a client machine is enqueued on a different TCP connection
 - › Signing uses the user’s credentials

➤ Scalability for file sharing is increased

Limits	SMB1	SMB2
Number of Users	Max 2^{16}	Max 2^{64}
Number of Open Files	Max 2^{16}	Max 2^{64}
Number of Shares	Max 2^{16}	Max 2^{32}

➤ Typical implementation changes:

Capability	SMB1	SMB2
Typical Largest PDU	16-64 KB	64 KB
Connection	Per Server	Per Server Per User
Typical Maximum BDP	64 – 512 KB	8 MB

Summary of SMB2 vs. SMB1

- SMB2 is a simpler protocol
 - ◆ 19 instructions instead of ~100
- SMB2 is more scalable
 - ◆ Number of open files, users, shares
- SMB2 performs well on WAN/MAN, high speed LAN
 - ◆ Scaling of credits, compounding
- SMB2 more robust on intermittent networks
 - ◆ Durable Handles
- SMB2 is more secure
 - ◆ Signing per user, SHA-256

- Please send any questions or comments on this presentation to SNIA: tracknetworking@snia.org

**Many thanks to the following individuals
for their contributions to this tutorial.**

- SNIA Education Committee

**David Kruse
Tom Jolly
Joe White
SW Worth
Teresa Yao**

Appendix

➤ Industry Abbreviations:

- ◆ WAN – Wide Area Network
- ◆ LAN – Local Area Network
- ◆ WLAN – Wireless Local Area Network
- ◆ RTT – Round Trip Latency (usually in milliseconds)
- ◆ MB – Megabytes = 1,000,000 bytes
- ◆ KB – Kilobytes = 1,000 bytes

➤ Abbreviations specific to this talk:

- ◆ BDP – Bandwidth Delay Product – a measure of how much data must be outstanding to fill a connection
- ◆ RPC – Remote Procedure Call
- ◆ GUI – Graphical User Interface

➤ Branch connectivity rules of thumb:

- ◆ Latency (RTT)
 - > ~10 ms MAN
 - > ~100 ms WANcontinental
 - > ~500 ms WAN Transcontinental
 - > ~1000 ms WAN Satellite
- ◆ Bandwidth
 - > 2-50 Mbit/sec, increasing over time
 - > Developing countries have significant number of analog lines

➤ Data center connectivity rules of thumb:

- ◆ Latency rule of thumb is the same as branch
- ◆ Bandwidth is much higher
 - > 100 – 10,000 Mbits/sec, increasing over time