SMB Version 2:

Scaling from Kilobits to Gigabits

Jim Pinkerton

Microsoft Corporation

Data generated by Teresa Yao and Jay Hamidi

9/23/2008

Agenda

- Introduction
- SMB Version 2 design goals
- Characterizing enterprise branch offices
- WAN performance analysis
  - Overview of the software layers & bottlenecks
  - Test environment & results
  - Production environment & results
- Conclusions

Introduction to SMB/CIFS

- SMB history goes back to 1983
  - SMB/CIFS: SMB -> CIFS -> SMB -> SMBv2
    - CIFS = SMB as it shipped in NT4 server
  - Post CIFS (Windows 2000 and later)
    - Kerberos and domains
    - Shadow copy
    - Server – to – Server copy
    - SMB signing
- Recent new industry trends
  - Proliferation of branch offices
  - Server consolidation and virtualization
  - Mobile workers
  - WAN accelerators

SMB Version 2 Design Goals (1/2)

- Reduced complexity
  - Simplified opcodes
    - SMB > 100 vs. SMBv2 = 19
  - Extension mechanism (eg. create context, variable offsets)
- Better WAN throughput, less chattiness
  - Credit based flow control
    - Server can control per client resource consumption
  - More Flexible compounding
    - Parallel or chained - Response for every element in the chain
  - NAT Friendliness - VC count is gone

SMB Version 2 Design Goals (2/2)

- Increased scalability and security
  - Improved scaling:

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

  - Improved Signing security
    - HMAC SHA-256 replaces MD5
- Symbolic Links
- Durable Handles - Reconnect on loss of connection

Network Characteristics

- Network Media link speed spans 10^6
  - Cellular modems
  - Dial-Up networking 9600, 19200 Baud
  - 10Mb/s – 1Gb/s Ethernet
  - Wireless LAN
  - 10 Gb/s Ethernet and 32 Gb/s Infiniband
- Latency
  - <1ms to 1200ms
- WAN deployments
  - Branch to Data Center (low speed, high latency)
  - Data Center to Data Center (high speed, high latency)

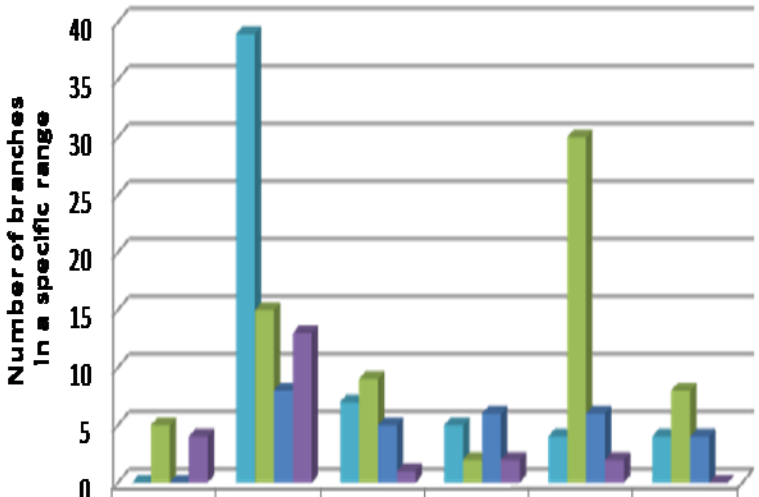
Branch Office Bandwidth

Continental Bandwidth (Mb/s)

	North America	Europe	Asia	South America
Mean	**14**	**33**	**26**	**5.5**
Stddev	28	43	28	7.2
Min	1.3	0.9	3.5	0.5
Max	**155**	**155**	**92**	**25**

- Continental is from any branch in a continent to nearest data center
  - Branches in US, Europe, Asia, South America
  - Data Centers in the US, Europe and 2 in Asia

Branch Bandwidth Distribution (Mb/s)



Bandwidth Range	0-1	2-7	8-10	11-20	21-50	51-200
North America	0	39	7	5	4	4
Europe	5	15	9	2	30	8
Asia	0	8	5	6	6	4
South America	4	13	1	2	2	0

Total sample set = **137 branches**

Branch Office Latencies

- Intercontinental is from any branch in a continent to furthest data center
- All latencies are measured with ICMP ping, on an essentially idle network

Continental Latencies (ms)

	North America	Europe	Asia
Mean	**108**	**109**	**163**
Stddev	69	91	81
Min	20	5	20
Max	**375**	**372**	**460**

Intercontinental Latencies (ms)

	North America	Europe	Asia	South America
Mean	**311**	**466**	**415**	**615**
Stddev	81	83	138	48
Min	110	135	91	188
Max	**655**	**717**	**830**	**1256**

Branch Office BDP

- Bandwidth Delay Product (BDP) = latency * bandwidth
  - Defines how much data must be in flight to achieve maximum bandwidth
- Worst: Denmark to Singapore: 155 Mb/s with 592 ms delay = 11 MB BDP
- Example Data Center to Data Center
  - 1Gbit/s with 76ms delay results in 9.5 MB BDP

Continental BDP (MByte)

	North America	Europe	Asia
Mean	**0.21**	**0.25**	**0.53**
Stddev	0.45	0.30	0.46
Min	0.008	0.005	0.083
Max	**2.80**	**1.60**	**1.80**

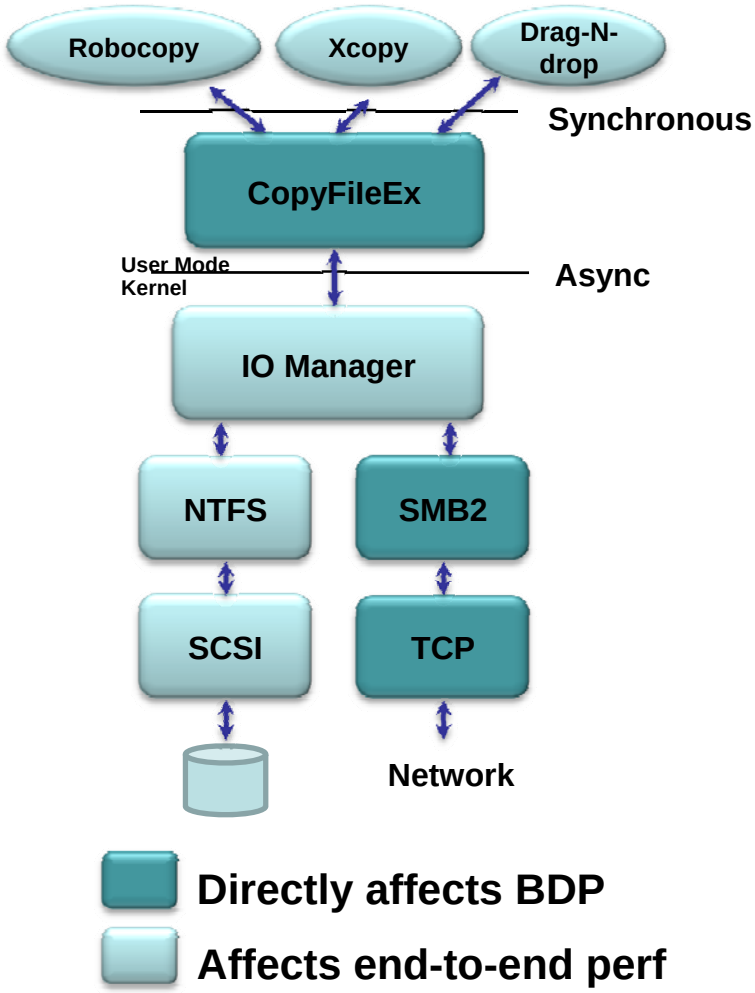
Intercontinental BDP (MByte)

	North America	Europe	Asia	South America
Mean	**0.71**	**2.40**	**1.70**	**0.46**
Stddev	1.40	3.00	1.60	0.58
Min	0.05	0.08	0.27	0.06
Max	**6.60**	**11.00**	**5.90**	**1.90**

Optimizing for WAN

- For low speed WAN links
  - Small SMB PDU
    - Better responsiveness for mix of file transfers and directory enumeration
  - Limit outstanding data
    - Avoid false I/O timeouts due to head-of-queue blocking
- For high speed WAN links
  - Scale the BDP
    - Each layer in the stack has a roll
  - Ensure congestion doesn't ruin performance
- This talk will examine copying a file from a local disk to a remote disk to show all the layers involved

Windows Client Layers



TCP Layer Optimizations for WAN

Feature	Vista RTM	Windows 2008
Receive Window auto-tuning	Yes (up to 16 MB, wininet 256 KB)	Yes (up to 16 MB)
High BDP Congestion Control - CTCP	N (default)	Y
RFC 3782 - NewReno	Y	Y
RFC 2883 – SACK extensions	Y	Y
RFC 3517 – SACK based loss recovery	Y	Y
RFC 4138 – Forward RTO recovery	Y	Y
RFC 3540 - ECN	N (default)	N (default)

Optimization categories:

- Congestion (lots of issues)
- Scale receive window from a small value to something quite large

On Windows, can enabled/disable the following with netsh commands:

- CTCP
- ECN
- Disable receive window autotuning

SMB2 Tuning for low speed links

- Goals:
  - Attempt to maintain application responsiveness
  - Attempt to not cause false timeouts on I/Os
- Algorithms:
  - Timer armed when SMB packet sent – not when application posts I/O.
  - Server ramps from a small number of credits to Smb2MaxCredits
    - Starts at 16, automatically scales to 128, as needed
  - Client throttles credits as a function of bandwidth
  - Dynamically vary PDU size as a function of network bandwidth
    - 0-128 Kbps = 16 KB PDU
    - 128-256 Kbps = 32 KB PDU
    - > 256 Kbps = 64 KB PDU

CopyFileEx Optimizations for WAN

- **Optimizations are a balance of:**
  - Virtual Address pressure (32 bit OS)
  - Non-paged pool (kernel pinned memory)
  - Filling the BDP
  - For SMB1, keeping under 64 KB PDU for read so don't end up with 2 PDUs

XP(SMB1)

Synchronous 64 KB Writes
Synchronous 60 KB Reads

Vista SP1 (SMB1)

Multiple async 32 KB Writes, 16 chunks
Multiple async 32 KB Reads, 16 chunks

Vista RTM, For SMB2

File Size	Pipeline Depth	Chunk Size
<= 1MB	1	File size rounded to sector size
> 1MB and <= 8 MB	2	1 MB
> 8 MB and <= 256 MB	4	2 MB
> 256 MB	4	8 MB

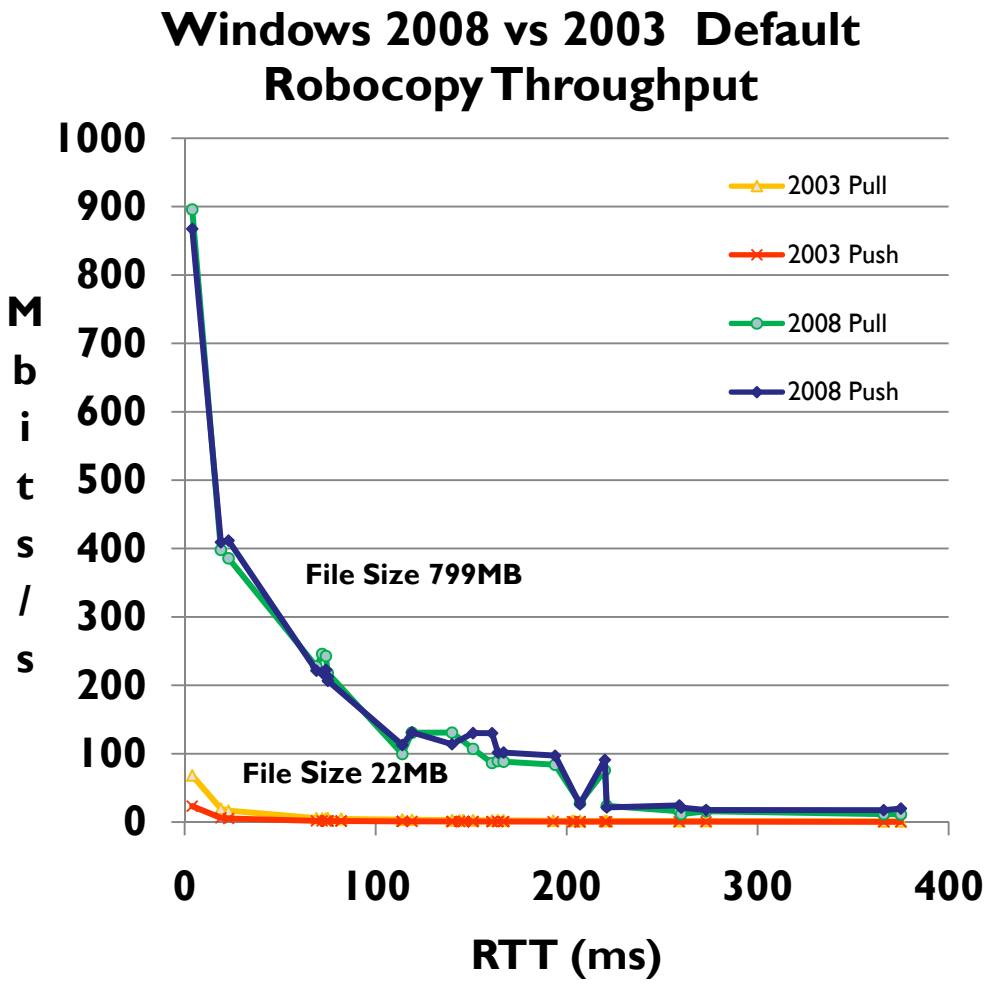
Vista SP1, For SMB2

File Size	Pipeline Depth	Chunk Size
<= 512kB	8	128kB
> 512kB and <= 2 MB	8	256 kB
> 2 MB and <= 8 MB	8	512kB
> 8 MB	8	1 MB

Today's Data for Copying a File: Varying Latency for WAN/MAN

Can we do better?

- Graph compares Windows 2008 to Windows 2003
  - Latency for just WAN/MAN
    - LAN not shown
  - WAN BW = 1 Gb/s
    - Theoretical is ~900 Mb/s
  - All tests use production servers and networks
    - Congestion is normal
  - All tests are disk to disk - RAID5, 12 disks
  - Push = Write, Pull = Read



Testing WAN using Emulation

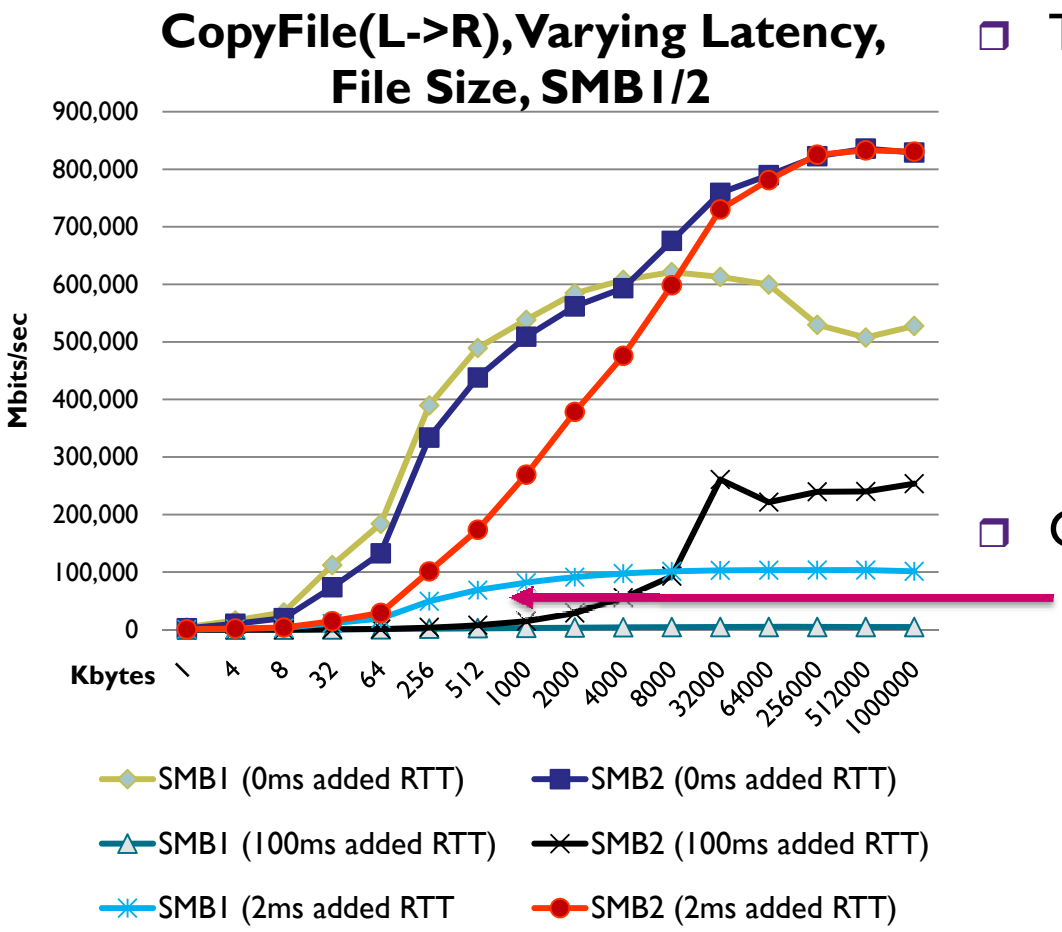
- **Focus**: Create reliable test infrastructure with very tight variance so that small performance regressions are real (and can be automatically detected)
  - Need to take disks out of the equation – use a RAM disk instead
  - Tolerances for even smallish I/Os are pretty tight (<1-2%)

- **Hardware**:
  - 2 hosts
  - 1 gigabit LAN, h/w WAN simulator
  - 64 bit hardware, Intel Xeon, dual core, 1.6 GHz
  - 4 GB RAM (1.2 GB RAM disk)
  - Boot disk on SATA

- **Software**:
  - Windows XP (64 bit) on SMB1 client, Windows Server 2003 (64 bit) on SMB1 server
  - Vista SP1 (64 bit) on SMB2 client, Windows Server 2008 (64 bit) on SMB2 server

- **Qualifiers:**
  - BECAUSE THIS USES A RAM DISK, THEY ARE "BEST CASE" – i.e. never to be seen in the field

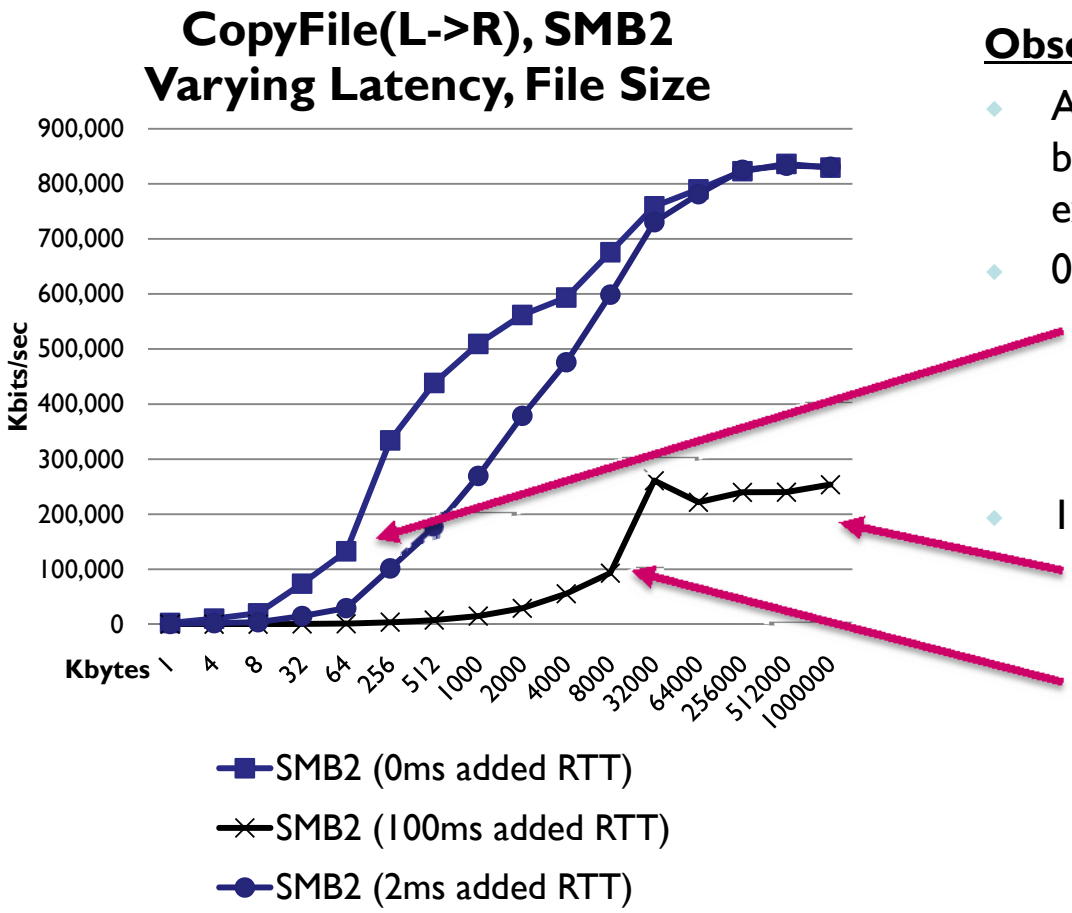
Comparing SMB1 and SMB2



- Test details:
  - SMB1 is Windows XP
  - SMB2 is Vista SP1
  - WAN emulator for 2 ms, 100 ms
  - Direct connect for 0 ms
  - Single outstanding filecopy (i.e. not multiple files)
- Observations:
  - SMB1 2ms filecopy on XP had serious issues (let alone 100 ms)
    - Filecopy has single ~60KB buffer outstanding

RAM DISK USED – DO NOT QUOTE THESE NUMBERS

SMB2 Issue with 12 MB BDP

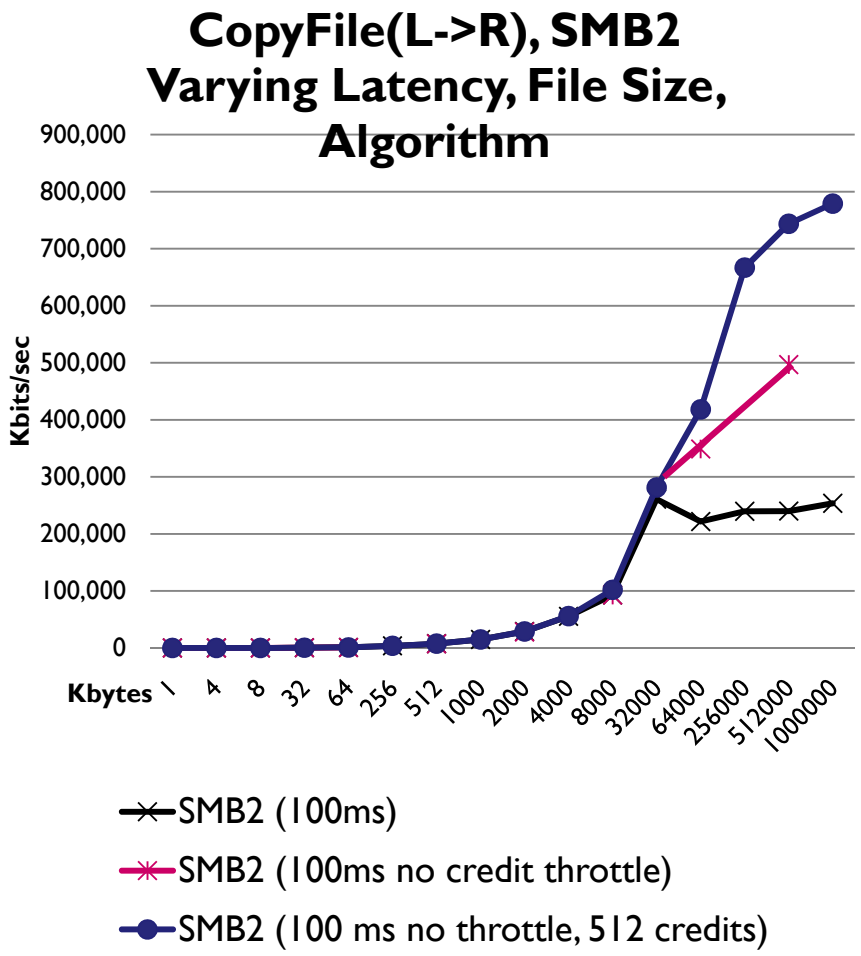


Observations

- Additional latency requires larger files before start of climbing BW curve (as expected)
- 0 ms:
  - Slope changes substantially at 64 KB
    - Prefetch/full disk reads kick in
    - SMB2 pipelining kicks in
- 100 ms:
  - Bandwidth tops out due to unknown issues
  - Slope changes substantially at 8 MB
    - CopyFileEx uses 8 buffers one MB in size

RAM DISK USED – DO NOT QUOTE THESE NUMBERS

SMB2 Bottlenecks Solved



Observations

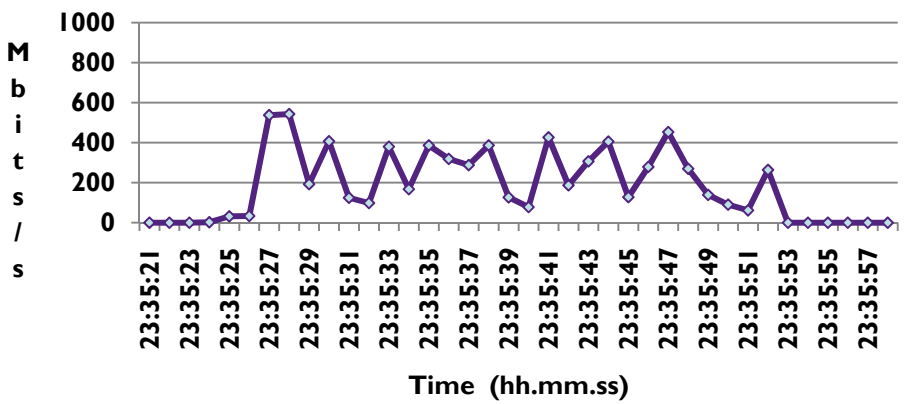
- At 100 ms latency, Windows 2008 RTM tops out at ~250 mb/s
- Experiments:
  - Removing credit throttling due to BW estimation issue moves BW to ~500 mb/s
  - Removing credit throttling and increasing credits to 512 moves to ~800 mb/s

RAM DISK USED – DO NOT QUOTE THESE NUMBERS

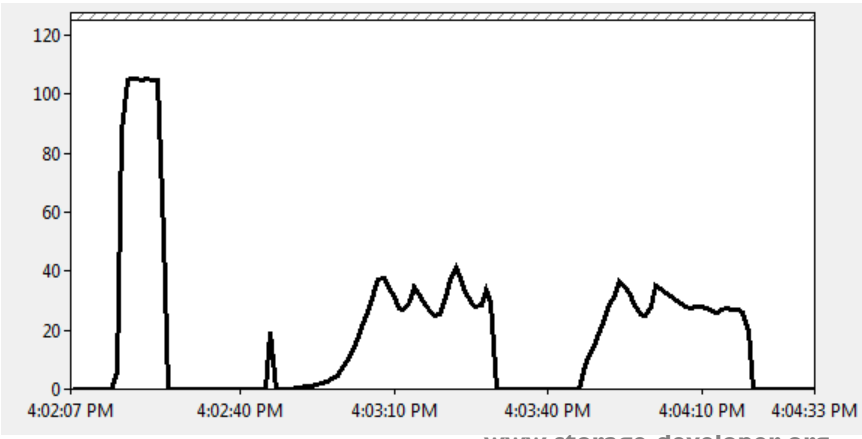
Analyzing Real Networks

- Disk-to-disk transfers
  - Raid5, 12 disks
- WAN Characteristics
  - 70 ms, 1 Gb/s WAN = 9.5 MB BDP
  - Production network (losses are normal)
- Top graph: File Copy
  - Poor performance was due to wide BW variance over single robocopy transfer (average ~250 Mb/s)
  - Poor interactions between SMB bandwidth throttling, TCP congestion window
- Bottom graph: NT-TTCP
  - Three NTttcp transfers in a row
  - TCP doesn't recover well under some loss events combined with high BDP

Windows 2008 x64 RTM pull
RTT = 70ms

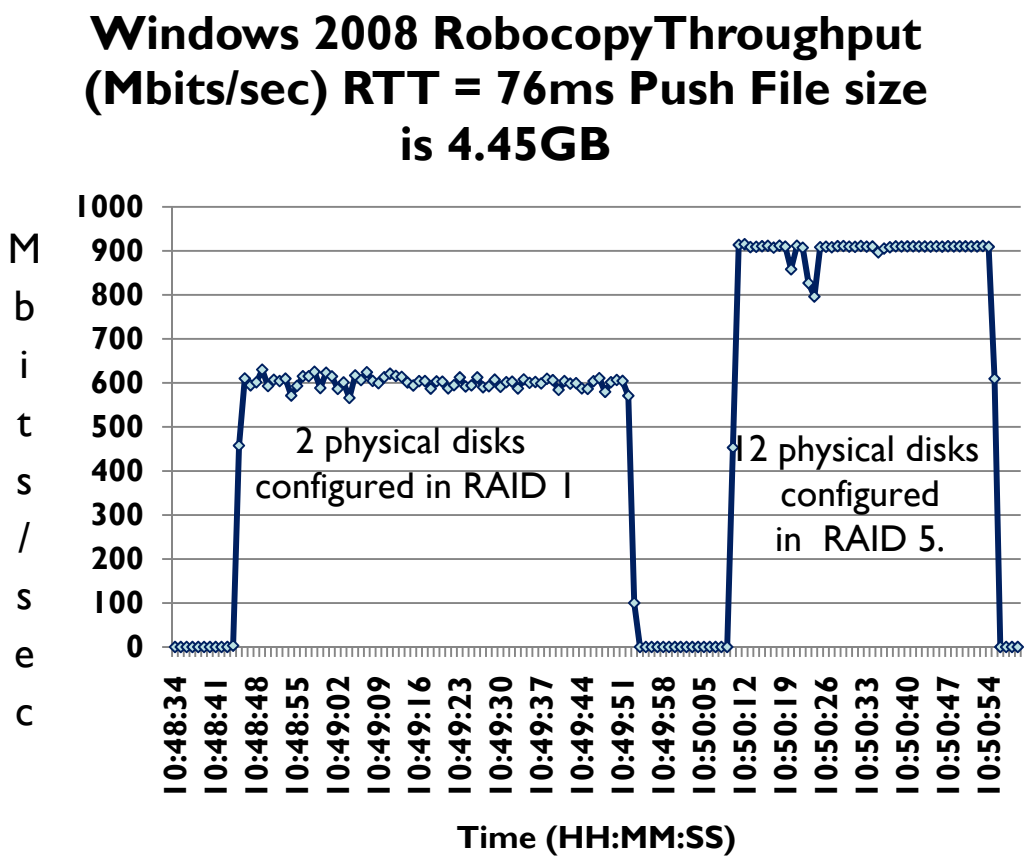


3 different NTttcp runs



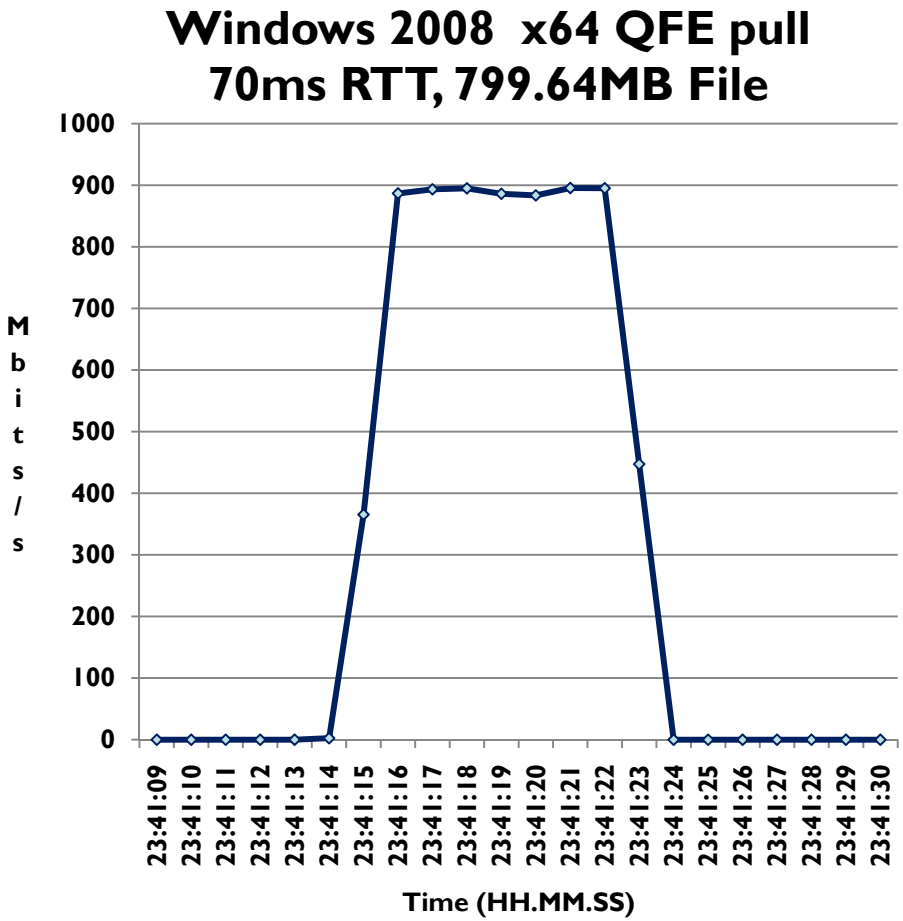
Source Disk Bottleneck

- For SMB2, source disk for filecopy must be able to keep up



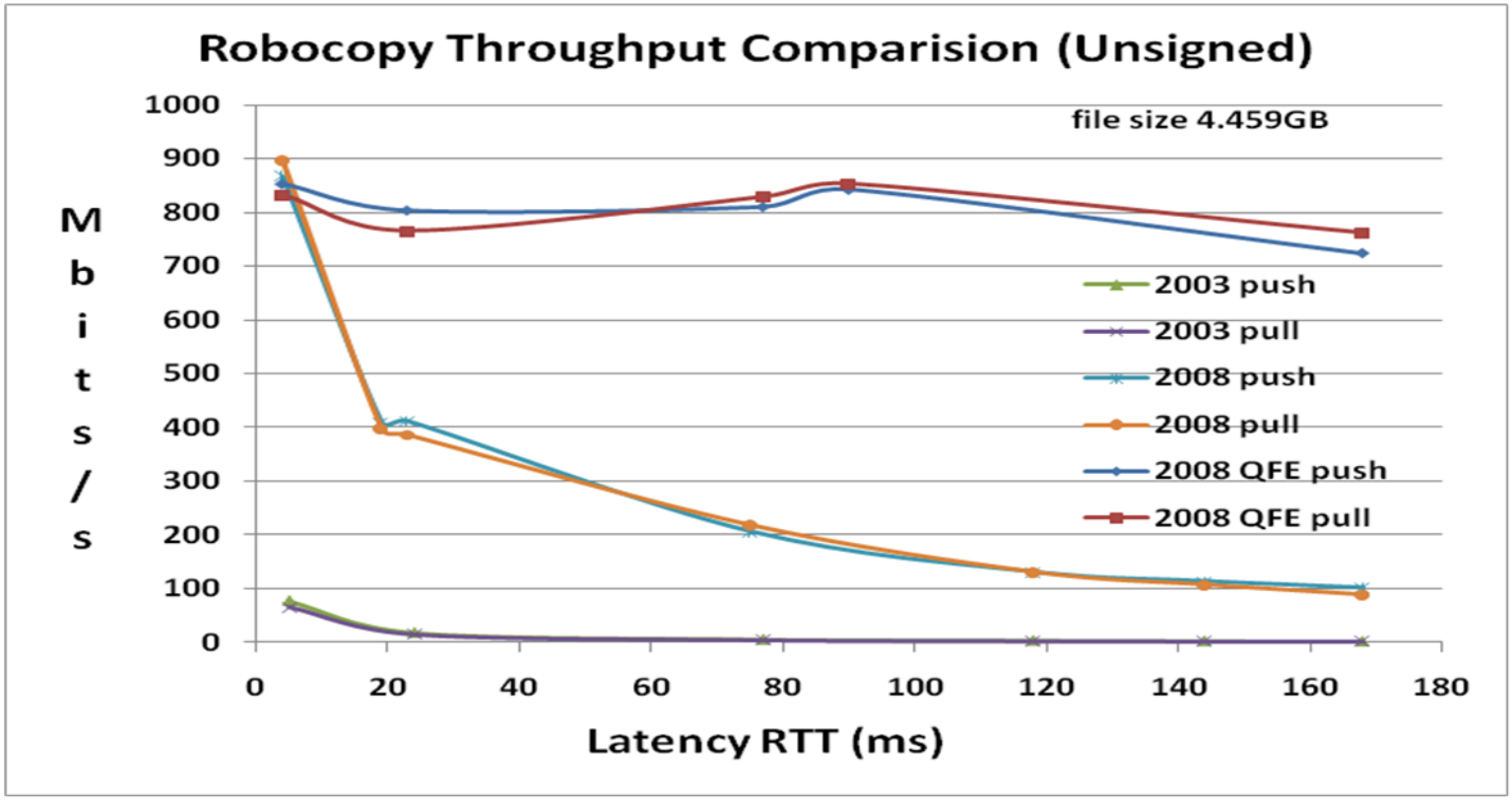
Problems Solved

- Fixed bugs in TCP congestion window
- Fixed bug in SMB2 credit throttling
- Extended maximum credits to 512



Final Production Data with QFE

- Installed SMB2 and TCP QFE and increased SMB2 max credits
- Data collected between many different data centers to test different latencies



Conclusions on WAN Performance

- BDP theory matches practice for finding bottlenecks
  - Credit management and credit throttling issues resolved
  - Each layer of the stack must be optimized for BDP
- TCP congestion control with loss is tricky, and Windows 2008 was not initially optimized for long, fat pipes (100 ms, 1 gigabit)
- Branch analysis
  - Latency rules of thumb
    - Continental: ~100 msec
    - Intercontinental: ~500 msec
  - Average branch BDP today is ~ 1 MB, and growing

Questions?