SMB 2.2 Performance

Dan Lovinger
Principal Architect
Microsoft
Overview

- Stats & Methods
- Scenario: OLTP Database
- Scenario: Cluster Motion
- SMB 2.2 Multi Channel

**Agenda**: challenges exposed by new protocol features of SMB 2.2, as experienced by Windows
Then there is the man who drowned crossing a stream with an average depth of six inches.

*W.I.E. Gates*
Starting Point – Metric v. Time

Transaction Rate

- All sorts of interesting things afoot!
Average

- Classic problem – completely different behaviors have the same average
  - understates steady state
  - hides very interesting behavior
- Performance is often in the variation
- How often … how consistently … where are the outliers to analyze …
Distribution: Transaction Rate

- Histogram / Distribution
- How often was the system running this fast?
- Note that data must be resampled into buckets
- In this example
  - buckets are 20 transactions/s wide
  - average located in the [4100 – 4120) transactions/s bucket, with 21 samples
Standard Deviation

- Familiar first approximation – here, +/- 1100 transactions/s
- Strong meaning if the type of distribution is known
  - Gaussian / Normal – the classic Bell Curve - ~34% to each side
  - … which this is not
- Spans >90% of the total distribution, here
Median

- The mid-point of all of the data points
  - 50% higher
  - 50% lower
- 50th Percentile
- Coincidentally matches the peak of this distribution - 4300 transactions/s
Percentiles

- Percentage of the dataset – cut lines
- Relevant to *guarantees* of behavior / performance
- Median locates the center
- 10’s and 90’s
  - No worse than / At least X, Y% of the time
- What makes sense for latency? Bandwidth?
Percentiles

Distribution : Transaction Rate

- Looking back to standard deviation, it really didn’t work.
- … for a 30 minute real workload with a well defined steady state!
Cumulative Distributions

- Visualization of percentiles
Note: for this section, SMB and SMB+ refer to a before and after change, *not a revision of the protocol.*

The issue was *not* in SMB and … very specific to workload. See summary.
OLTP – OnLine Transaction Processing

- Log File – small-midsize sequential IO
- Database File(s) – small random IO

- Fundamental dependency on the SMB 2.2 feature set for continuously available connections to the database content
- New workloads … new problems for an implementation
## Simulated OLTP Hardware Configuration

<table>
<thead>
<tr>
<th></th>
<th>Client</th>
<th>Server</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CPU Type</strong></td>
<td>2 sockets with 6 cores @ 2.66 Ghz (12 cores total)</td>
<td></td>
</tr>
<tr>
<td><strong>Memory Amount</strong></td>
<td></td>
<td>36 GiB</td>
</tr>
<tr>
<td><strong>Network Type</strong></td>
<td></td>
<td>Onboard 1GbE network interface</td>
</tr>
<tr>
<td><strong>Number of Network Links</strong></td>
<td></td>
<td>1 x 1GbE</td>
</tr>
<tr>
<td><strong>Storage Adapter</strong></td>
<td>N/A</td>
<td>1 Fibre Channel Adapter 4Gb/s connectivity</td>
</tr>
<tr>
<td><strong>Storage</strong></td>
<td>N/A</td>
<td>14x 10KRPM HDD RAID0</td>
</tr>
</tbody>
</table>
Simulated OLTP Hardware Configuration

Client

1 Gb Ethernet

SMB

Server

4Gb Fibre Channel

DAS

14 Disk RAID10
17 Years Later

- The Write Bubble
- NT Filesystem Lingo: Valid Data Length
  - High water mark in a file to which user data has been produced
  - Deals with fully allocated files and efficient zeroing
- Code sharing from the FAT implementation
  - May 1994 – clever trick for VDL extension
  - Nov 1994 – ported over with a slight mislocation
- Causal requirements
  - Single-File, Mixed Read/Write, Async, NonCached IO
- No other workload affected!
Simulated OLTP

Average Disk Queue Length Distribution

- SMB
- SMB+
- DAS

Seconds @ Average Queue Length vs Disk Queue Length
Simulated OLTP

8KiB Random IO/s Cumulative Distribution SMB v. DAS
Windows 2008 R2 RTM

IO Operations per Second

Percentile

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

0 200 400 600 800 1000 1200 1400 1600 1800 2000

SMB
SMB+
DAS
## Simulated OLTP

### SMB+ Client v. Server DAS Latency

**Windows 2008 R2 RTM**

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Read Shift (%)</th>
<th>Write Shift (%)</th>
<th>Read Shift (us)</th>
<th>Write Shift (us)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>13%</td>
<td>7%</td>
<td>489</td>
<td>495</td>
</tr>
<tr>
<td>50%</td>
<td>6%</td>
<td>4%</td>
<td>506</td>
<td>521</td>
</tr>
<tr>
<td>90%</td>
<td>3%</td>
<td>2%</td>
<td>709</td>
<td>931</td>
</tr>
<tr>
<td>95%</td>
<td>2%</td>
<td>1%</td>
<td>978</td>
<td>832</td>
</tr>
<tr>
<td>99%</td>
<td>1%</td>
<td>1%</td>
<td>982</td>
<td>881</td>
</tr>
</tbody>
</table>
Simulated OLTP

8KiB Random IO/s Cumulative Distribution SMB v. DAS
Windows 8

Methodology Effect

Percentile

IO Operations per Second

0 200 400 600 800 1000 1200 1400 1600 1800 2000

0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
Simulated OLTP

**SMB+ Client v. Server DAS Latency**

*Windows 8*

### Percentile

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Read Shift (%)</th>
<th>Write Shift (%)</th>
<th>Read Shift (us)</th>
<th>Write Shift (us)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>12%</td>
<td>6%</td>
<td>414</td>
<td>411</td>
</tr>
<tr>
<td>50%</td>
<td>5%</td>
<td>4%</td>
<td>415</td>
<td>416</td>
</tr>
<tr>
<td>90%</td>
<td>2%</td>
<td>1%</td>
<td>425</td>
<td>430</td>
</tr>
<tr>
<td>95%</td>
<td>1%</td>
<td>0%</td>
<td>425</td>
<td>421</td>
</tr>
<tr>
<td>99%</td>
<td>0%</td>
<td>0%</td>
<td>439</td>
<td>461</td>
</tr>
</tbody>
</table>
Simulated OLTP

SMB Server Thread Pool
Sync v. Async Threadpool

- SMB Thread Pool Size
- Time (seconds)

- Default: 32 threads
- Async: 8 threads

- Time range: 0 to 300 seconds
<table>
<thead>
<tr>
<th></th>
<th>Client</th>
<th>Server</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CPU Type</strong></td>
<td>2 sockets with 4 cores @ 2.26 Ghz (4 cores total)</td>
<td></td>
</tr>
<tr>
<td><strong>Memory Amount</strong></td>
<td></td>
<td>24 GiB</td>
</tr>
<tr>
<td><strong>Network Type</strong></td>
<td></td>
<td>Onboard 1GbE network interface</td>
</tr>
<tr>
<td><strong>Number of Network Links</strong></td>
<td></td>
<td>1 x 1GbE</td>
</tr>
<tr>
<td><strong>Storage Adapter</strong></td>
<td>N/A</td>
<td>1 Fibre Channel Adapter 4Gb/s connectivity</td>
</tr>
<tr>
<td><strong>Storage</strong></td>
<td>N/A</td>
<td>24x 10KRPM HDD: 2x 10 HDD RAID0 (DB) 2 HDD RAID0 (LOG)</td>
</tr>
</tbody>
</table>
OLTP Hardware Configuration

Client

1 Gb Ethernet

Server

4Gb Fibre Channel

2x 10 Disk RAID0 (Data)
2 Disk RAID0 (Log)

SMB

DAS
• Full result is intuitively close
  • Median 4270 transactions/s
  • Mean 3920 transactions/s
• 4Gb v. 1Gb effect?
OLTP – First Correction

SMB v. DAS Transaction Rate - Ingest

SMB+ Tps
DAS Tps
OLTP – First Correction

SMB v. DAS Bandwidth - Ingest

SMB+ Read Bytes
DAS Read Bytes
• Far closer!
OLTP – Second Correction

Transaction Rate
To Equal Work Steady State - 5.62 GB OLTP Log

Time (seconds)
0 60 120 180 240 300 360 420 480 540

Transactions/s
0 1000 2000 3000 4000 5000 6000 7000

SMB+
DAS
OLTP – Second Correction

SMB v. DAS Transition Read Bandwidth
To Equal Work Steady State - 5.62 GB OLTP Log

Million Bytes/s

Time (seconds)

- 1Gb
- SMB+
- DAS
OLTP – After Second Correction

- Again intuitive – random IO bandwidth << 1Gb
- Only -3.0% performance relative to DAS over final 19 minutes of the run (+33s)
- 1Gb Ethernet very nearly meets 4Gb FC

<table>
<thead>
<tr>
<th></th>
<th>Log Volume</th>
<th>Data Volumes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DAS</td>
<td>SMB2</td>
</tr>
<tr>
<td>Mean IO Operations/s</td>
<td>2125</td>
<td>2010 (-5.4%)</td>
</tr>
<tr>
<td>Median</td>
<td>2135</td>
<td>2055 (-3.7%)</td>
</tr>
<tr>
<td>80th</td>
<td>2195</td>
<td>2130 (-3.1%)</td>
</tr>
<tr>
<td>90th</td>
<td>2225</td>
<td>2155 (-3.0%)</td>
</tr>
<tr>
<td>95th</td>
<td>2240</td>
<td>2175 (-3.1%)</td>
</tr>
</tbody>
</table>
OLTP – More than 4/8 KiB

Database File IO Size

Database File IO - Bytes Transferred by IO

Note: 8iKB @ 49%
OLTP – Log Matters

Log Write Size

# IOs (Thousands)

IO Size (KiBytes)

0 4 8 12 16 20 24 28 32 36 40 44 48 52 56 60

0.1 1 10 100 1000
Understanding workload is critical for performance

Enabling new workloads is both a protocol and platform effort

Windows 7 / 2008 R2 KB 2536493

- [x] Single-File
- [x] Mixed Read/Write
- [x] Async
- [ ] NonCached IO

... otherwise, **no issue**. It took a highly scaled **database** to hit it.

- No other workload affected!
Cluster Motion

- SMB 2.2 Continuous Availability $\Rightarrow$ Cluster
- Implementations should look at performance of essential cluster management operations
- Bet: scale of clustered resources will rise significantly, relative to pre-SMB 2.2 services
- Planned v. Unplanned Motion
- **Goal**: define the time budget.
Cluster Motion

- Resource Group
- Container of objects representing the file service
- High scale on disks depending on implementation
  - HW/SW Volume
- Smaller scale on total number of groups
Cluster Motion

- Resource Control Manager handles the process for Windows, generalizable to other platforms.
- Cluster log follows the state changes, similar to:

```
```
Cluster Motion

Offline -> Online (Disks) - Initial
Offline -> Online (NonDisk) Initial
Cluster Motion

- Crossover Ethernet cluster network
- Simultaneous disconnect of crossover network and loss of node (e.g., node power loss)
- Unplanned motion – 80s delay
- Windows 2008 R2 KB 2575625
  - [http://support.microsoft.com/kb/2575625](http://support.microsoft.com/kb/2575625)
Cluster Motion - Disks

Hard Disk Offline - Initial

Clock Time (seconds)

Disks - 120 in 3 resource groups
Cluster Motion - Disks

- Hard disk resource internals
- Removal of FSCTL_LOCK_VOLUME prior to dismount
- Removal of cluster disk timer dependency to remove PR reservation
- Removal of soft guard between IOCTL_OFFLINE_VOLUME and the online process
- Reduce conservative polling interval
Cluster Motion - Disks

Hard Disk Offline - Final

Clock Time (seconds)

Disks - 120 in 3 resource groups
Cluster Motion - Final

Offline -> Online (Disks) - Final

Clock Time (seconds)

Disks - 120 in 3 resource groups
Cluster Motion

- Summary
  - Large outlier in unplanned loss removed
  - Relative: ~80% faster at scale
  - Absolute: sub-10s motion … at scale
  - Larger or smaller than possible reality?

- Implementers
  - Look at your platform
  - … repeat under loads …
Multichannel – Goals

- **Availability / Network fault tolerance**
  - Make the SMB2 protocol resilient to interface, link or switch failures.
  - Move “link awareness” higher up the stack to enable more intelligent decision making.
    - Augment NIC teaming at the network layer.
    - Keep fallback paths ready, prioritize available links.
    - React quickly to changes to network availability.

- **Performance – utilize the available resource**
Multichannel - Terminology

- **Channel**: underlying transport connection
- **Session**: authenticated user context.
- **Session Binding**: map Session ↔ Channel
  - N:N relationship

![Diagram showing channels, sessions, and bindings]
## Test Hardware Configuration

<table>
<thead>
<tr>
<th></th>
<th>Client</th>
<th>Server</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CPU Type</strong></td>
<td>2 sockets with 6 cores @ 2.66 Ghz (12 cores total)</td>
<td></td>
</tr>
<tr>
<td><strong>Memory Amount</strong></td>
<td></td>
<td>48 GiB</td>
</tr>
<tr>
<td><strong>Network Type</strong></td>
<td></td>
<td>2 network interface adapter cards</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Each card has 2 10GbE interfaces</td>
</tr>
<tr>
<td><strong>Number of Network Links</strong></td>
<td></td>
<td>4 x 10GbE</td>
</tr>
<tr>
<td><strong>Storage Adapter</strong></td>
<td>N/A</td>
<td>2 RAID Host Bus Adapters – 6Gb/s SAS connectivity</td>
</tr>
<tr>
<td><strong>Storage</strong></td>
<td>N/A</td>
<td>12 3G SATA SSD</td>
</tr>
</tbody>
</table>
Test Hardware Configuration

RAID0 – 12 SSDs

Server

Client

10GbE

RAID0 – 12 SSDs

Server

Client

10GbE
## IOMETER Configuration

<table>
<thead>
<tr>
<th>I/O Size</th>
<th>Number of Workers</th>
<th>Queue Depth</th>
<th>Total Queued</th>
</tr>
</thead>
<tbody>
<tr>
<td>512</td>
<td>32</td>
<td>8</td>
<td>256</td>
</tr>
<tr>
<td>1024</td>
<td>32</td>
<td>8</td>
<td>256</td>
</tr>
<tr>
<td>4096</td>
<td>32</td>
<td>8</td>
<td>256</td>
</tr>
<tr>
<td>8192</td>
<td>16</td>
<td>32</td>
<td>512</td>
</tr>
<tr>
<td>16384</td>
<td>16</td>
<td>16</td>
<td>256</td>
</tr>
<tr>
<td>32768</td>
<td>16</td>
<td>8</td>
<td>128</td>
</tr>
<tr>
<td>65536</td>
<td>16</td>
<td>8</td>
<td>128</td>
</tr>
<tr>
<td>131072</td>
<td>16</td>
<td>8</td>
<td>128</td>
</tr>
<tr>
<td>262144</td>
<td>16</td>
<td>8</td>
<td>128</td>
</tr>
<tr>
<td>524288</td>
<td>16</td>
<td>8</td>
<td>128</td>
</tr>
<tr>
<td>1048576</td>
<td>16</td>
<td>8</td>
<td>128</td>
</tr>
</tbody>
</table>

- **Constant #Op**
- **Constant Bytes**
- **Scaled Bytes**
Server (Local) Baseline

Server/Local Read Throughput

- Read Sequential
- Read Random

MB/sec vs IO Size (bytes)

512 | 1024 | 4096 | 8192 | 16384 | 32768 | 65536 | 131072 | 262144 | 524288 | 1048576

Preliminary
SMB 2.2 Client 10GbE Interface Scaling

SMB 2.2 Client Interface Scaling - Throughput

- 1 x 10GbE
- 2 x 10GbE
- 3 x 10GbE
- 4 x 10GbE

IO Size (bytes)

MB/sec

512 1024 2048 4096 8192 16384 32768 65536 131072 262144 524288 1048576

0 500 1000 1500 2000 2500 3000 3500 4000 4500
Server to SMB 2.2 Client Comparison

Server/Local vs. Client Throughput

- Server/Local Throughput
- Client (4 x 10GbE) Throughput

MB/sec vs. IO Size (bytes)
SMB 2.2 Client Interface Scaling - IOps

- 1 x 10GbE
- 2 x 10GbE
- 3 x 10GbE
- 4 x 10GbE

IO Size (bytes)

IOPS

0 50000 100000 150000 200000 250000 300000 350000
512 1024 4096 8192 16384 32768 65536 131072 262144 524288 1048576
Server to SMB 2.2 Client IOps

Server/Local vs. Client IOps with CPU %

- Server/Local IOps
- Client (4 x 10GbE) IOps
- Client (4 x 10GbE) Priv CPU
- Server/Local Priv CPU

I/Os per second vs. IO Size

Priv CPU % vs. IO Size
Multi Channel

- Implementation Challenges
  - Nodes with varying interface counts (1, 2, 3 NIC)
  - Operation distribution across interfaces
- Performance goal: scale
Summary

- SMB 2.2 enables new workloads
- Workload demands will expose new issues in platforms and implementations – be prepared
- New protocol features create new performance expectations
- See Also
  - George/Kruse (Tuesday 1:00pm & 2:00pm)
  - Kramer/Talpey (Tuesday 4:05pm)
- Existing platform improvements
  - Windows 2008 R2 KB 2575625 – Cluster Unplanned with Crossover Ethernet
  - Windows 7 / 2008 R2 KB 2536493 – Single-File Mixed Read/Write Async NonCached IO