

Lessons learned implementing a multi-threaded **SMB2** server in **OneFS**

Aravind Velamur Srinivasan
Isilon Systems, Inc

- Overview
- OneFS Overview
- Multi-threaded SMB Server Overview
- SMB2 Credit Algorithm
- SMB2.1 Leases
- Performance optimizations
- SMB2 vs SMB1 in OneFS

- ❑ SMB2 protocol has its own advantages over SMB1
 - ❑ Less Chatty than SMB1
 - ❑ Inherent performance improvements
- ❑ An SMB2 server poses some challenges to the implementation:
 - ❑ Credit Algorithm
 - ❑ New Lease types in SMB2.1
 - ❑ Performance bottlenecks and optimizations in a multi-threaded SMB server
- ❑ This talk will examine these challenges and lessons learned in OneFS

OneFS Overview

Isilon OneFS Cluster

NAS file server

Scalable

Add more storage in 5 mins

Reliable

8x mirror / +4 parity

Striped across nodes

Single volume file system

3 to 144 nodes

Fully symmetric peers

No metadata servers

Commodity hardware

CPU, Mem, Disks



Isilon OneFS File System



Concurrent access to all files
with all protocols

SMB1/SMB2

NFSv3/NFSv4

SSH

HTTP/FTP

OneFS – Summary

- ❑ OneFS is Isilon's sixth-generation operating system that provides the intelligence behind all Isilon scale-out storage systems.
- ❑ It combines the three layers of traditional storage architectures—file system, volume manager and RAID—into one unified software layer, creating a single intelligent file system that spans all nodes within a cluster.

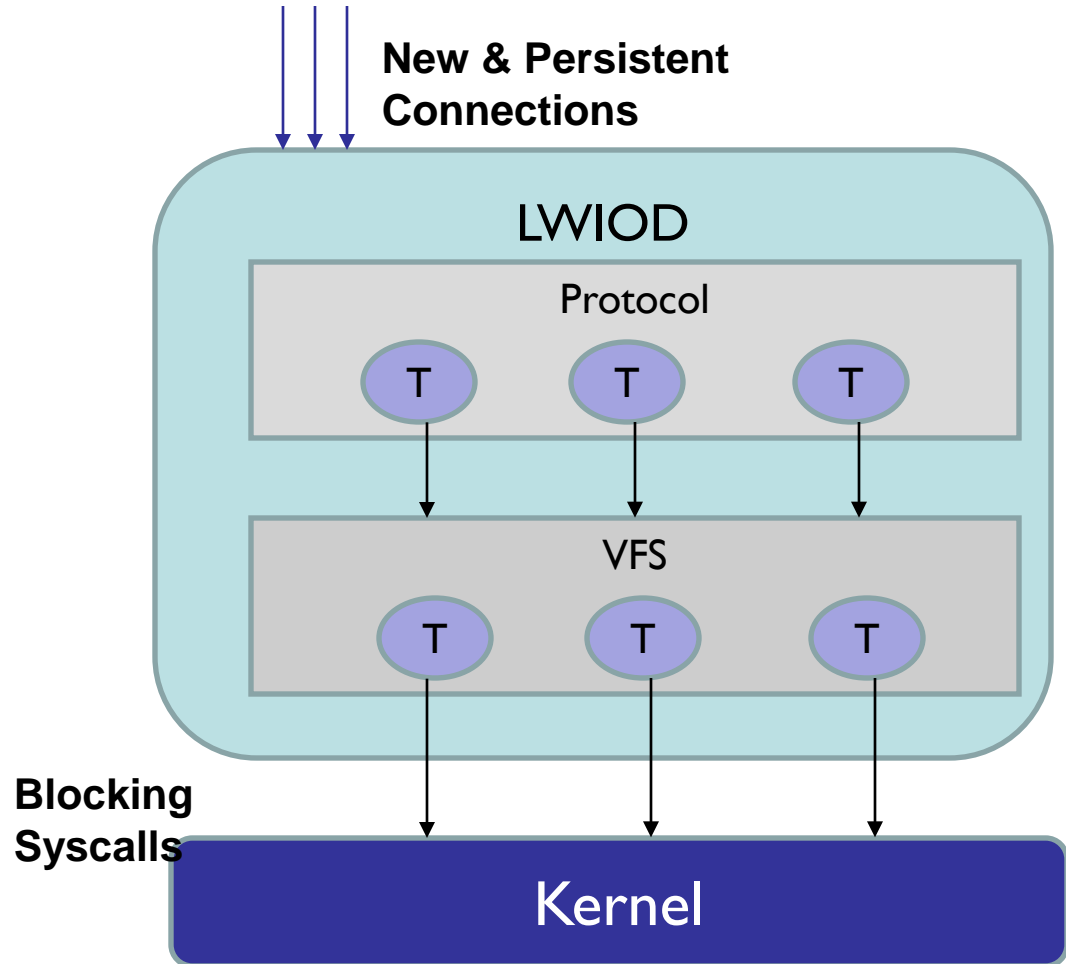
OneFS – Summary

- ❑ OneFS works exclusively with the Isilon scale-out storage system, referred to as a “cluster”.
- ❑ Isilon's OneFS enables:
 - ❑ Independent or linear scalability of performance and capacity
 - ❑ A single point of management for large and rapidly growing repositories of data
 - ❑ Mission-critical reliability and high availability with state-of-the-art data protection

Multithreaded SMB Server

LWIOD Architecture

- ❑ Single Process
- ❑ Multi-threaded



Multithreaded SMB Server Pros

- ❑ Pipelined Network I/O written in parallel
- ❑ Parallel syscalls: network I/O not blocked by file system
- ❑ Idle connections consume very little resources
- ❑ New connections limited by same threadpool as all other operations

Steven's talk at 2010 SDC:

http://www.snia.org/sites/default/files2/sdc_archives/2010_presentations/monday/StevenDanne-man_Comparison_Between_Samba-rev.pdf

SMB2 Credit Algorithm

SMB2 Credit Algorithm

- ❑ SMB2 offers a server-side credit mechanism to throttle greedy clients
- ❑ Server Credit Algorithm, if not implemented properly, can result in weird behavior from the clients

SMB2 Credit Algorithm

- ❑ SMB2 Credits allows the client to send requests in parallel
- ❑ Allows the client to build a pipeline of requests instead of waiting for a response before sending the next request.
- ❑ Especially relevant when using a high latency network.

SMB2 Credit Algorithm Pitfalls

- ❑ The server should gradually scale the credits starting from a low value.
- ❑ If the server grants all available credits at once to the client, the client sometimes uses those credits just for one operation without performing any other operation

SMB2 Credit Algorithm Pitfalls

Example

- ❑ A server which grants all available credits to the client at once (leaving just 1 credit)
- ❑ Try to copy a huge file (around 1 GB) from a Windows 7 client.

SMB2 Credit Algorithm Pitfalls

Example (Contd)

- ❑ Try to create a file using the explorer window, while the copy is in progress
- ❑ The explorer window just hangs until the Write operation completes successfully
- ❑ The client uses all the credits just for the Write operation without using it for other things

SMB2 Credit Algorithm – Lessons Learned

- ❑ Server side credit algorithm can affect the client behavior in some scenarios
- ❑ Match windows server behavior as much as possible
- ❑ Ramp up the credits starting from a smaller value up to a certain limit
- ❑ Can be further improved by using statistics of the existing connection as well as the load on the server

SMB2 Credit Algorithm – Windows Behavior

- ❑ Windows servers gradually ramp up the credits starting from a value of 16, allowing the client to accumulate up to 128 outstanding credits
- ❑ Never grants more than 33 credits per round (using a IGE link).

SMB2.1 Leases in OneFS

SMB2.1 Leases in OneFS

- ❑ SMB2.1 Leases provide more fine grained and flexible caching for clients and allow upgrades in addition to the oplock breaks we have in the legacy oplocks.
- ❑ Extending the oplock system to support leases is straightforward for most of the scenarios with a few exceptions

SMB2.1 Leases – Lessons Learned

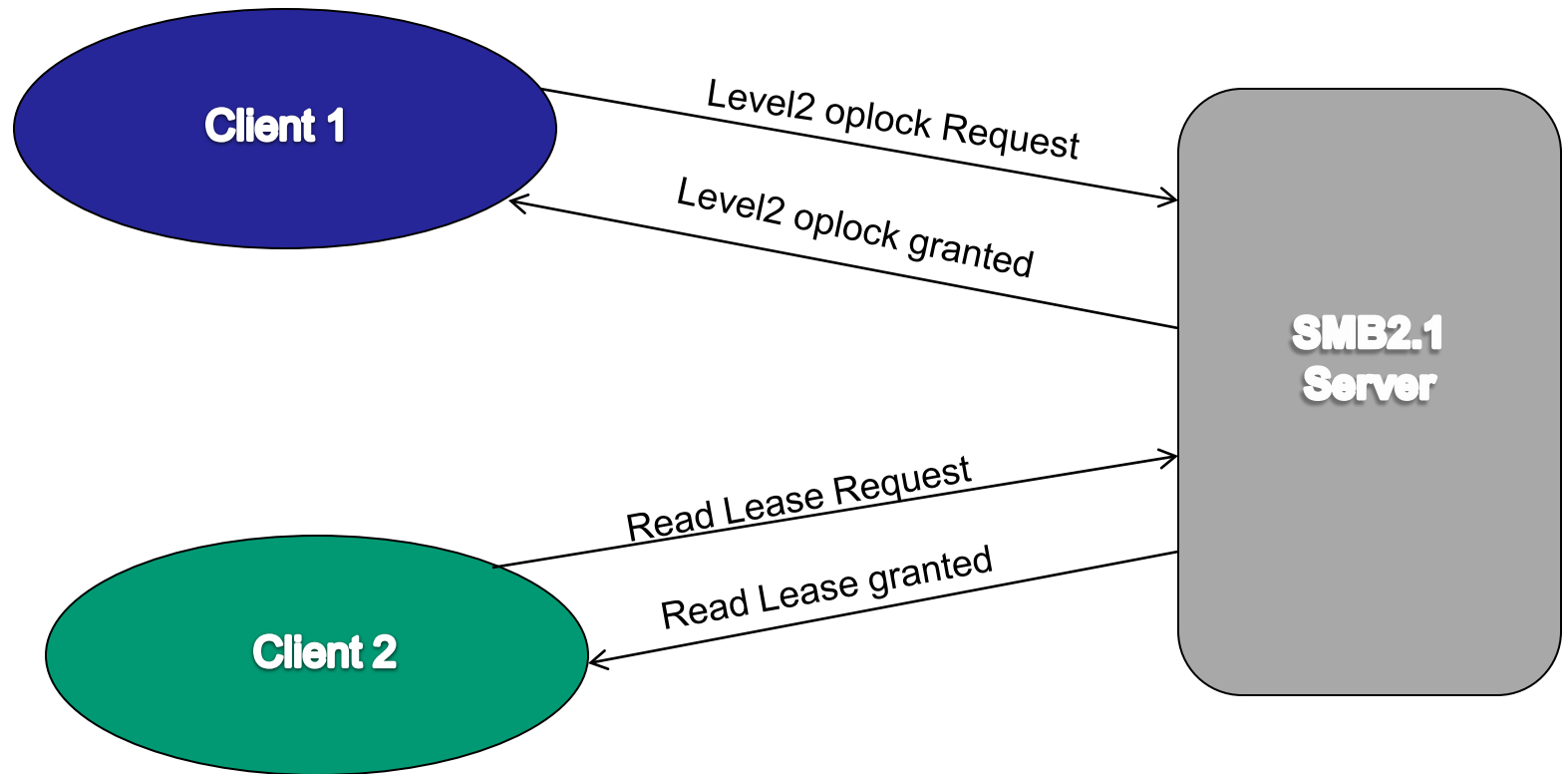
- ❑ The unique key used to identify the lock context has to be changed to accommodate the 128-bit client key generated for leases
- ❑ The interaction between the legacy oplocks and the new lease types are not completely straightforward

SMB2.1 Leases – Interaction between Read, RH and Level2 Oplocks

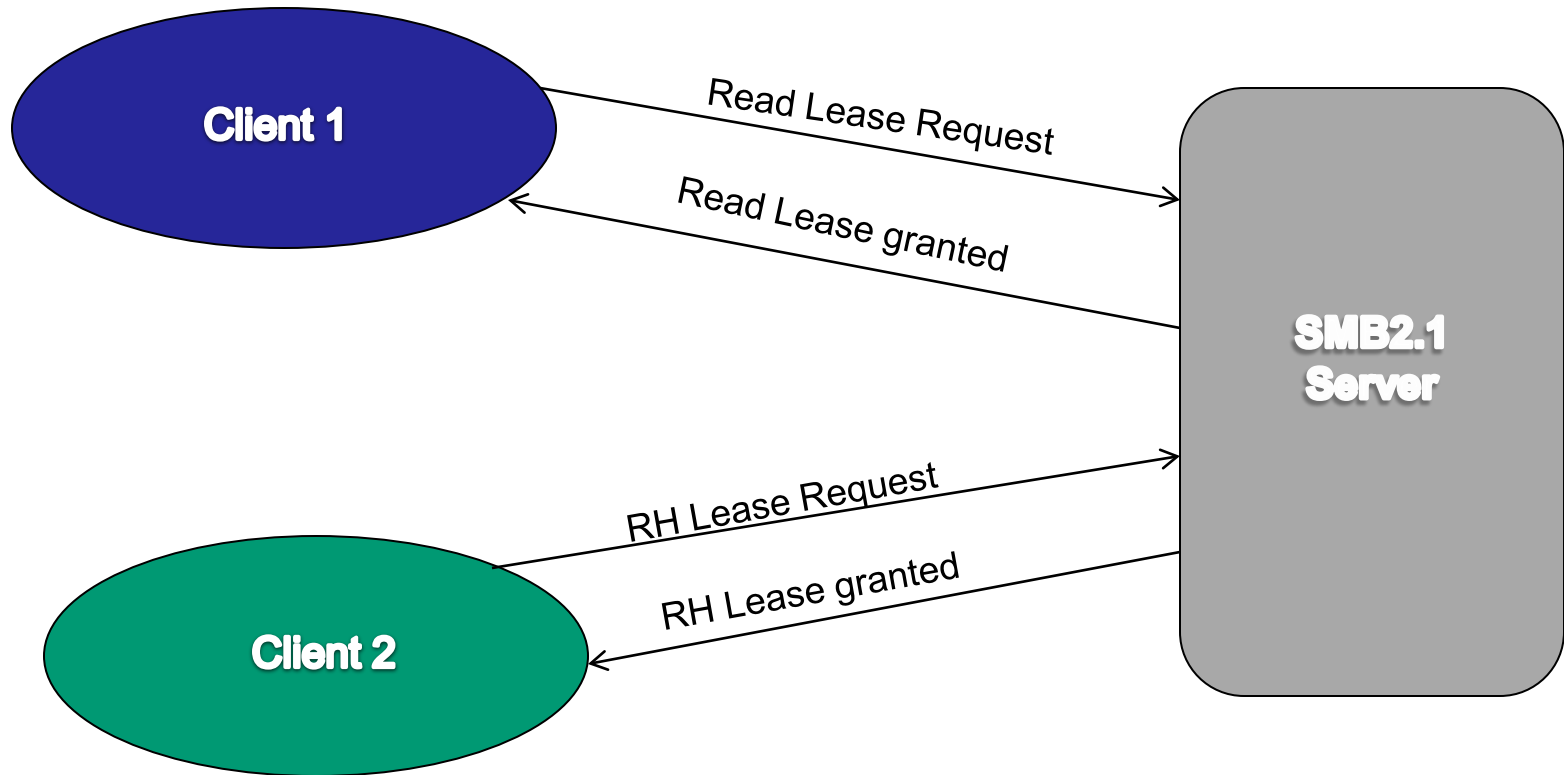
- ❑ The interaction between Read lease, RH lease and Level2oplock is complicated
- ❑ Although Level2 and Read types look similar, they have a different relationship with RH lease
- ❑ Multiple Read and RH oplocks can coexist on the same stream
- ❑ Multiple Read and Level2 oplocks can coexist but Level2 and RH **cannot coexist**

SMB2.1 Leases – Interaction between SDC

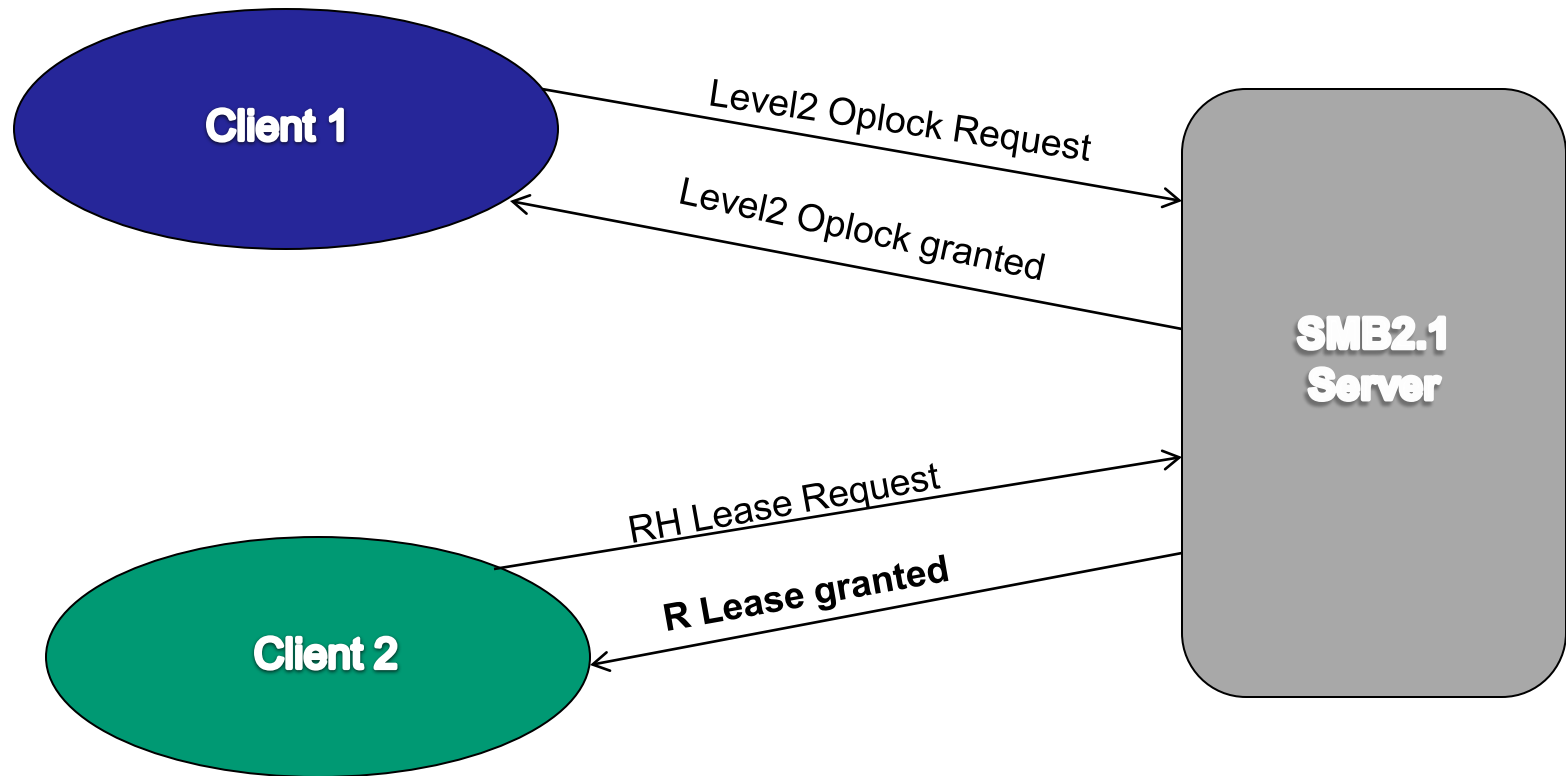
Read, RH and Level2 Oplocks



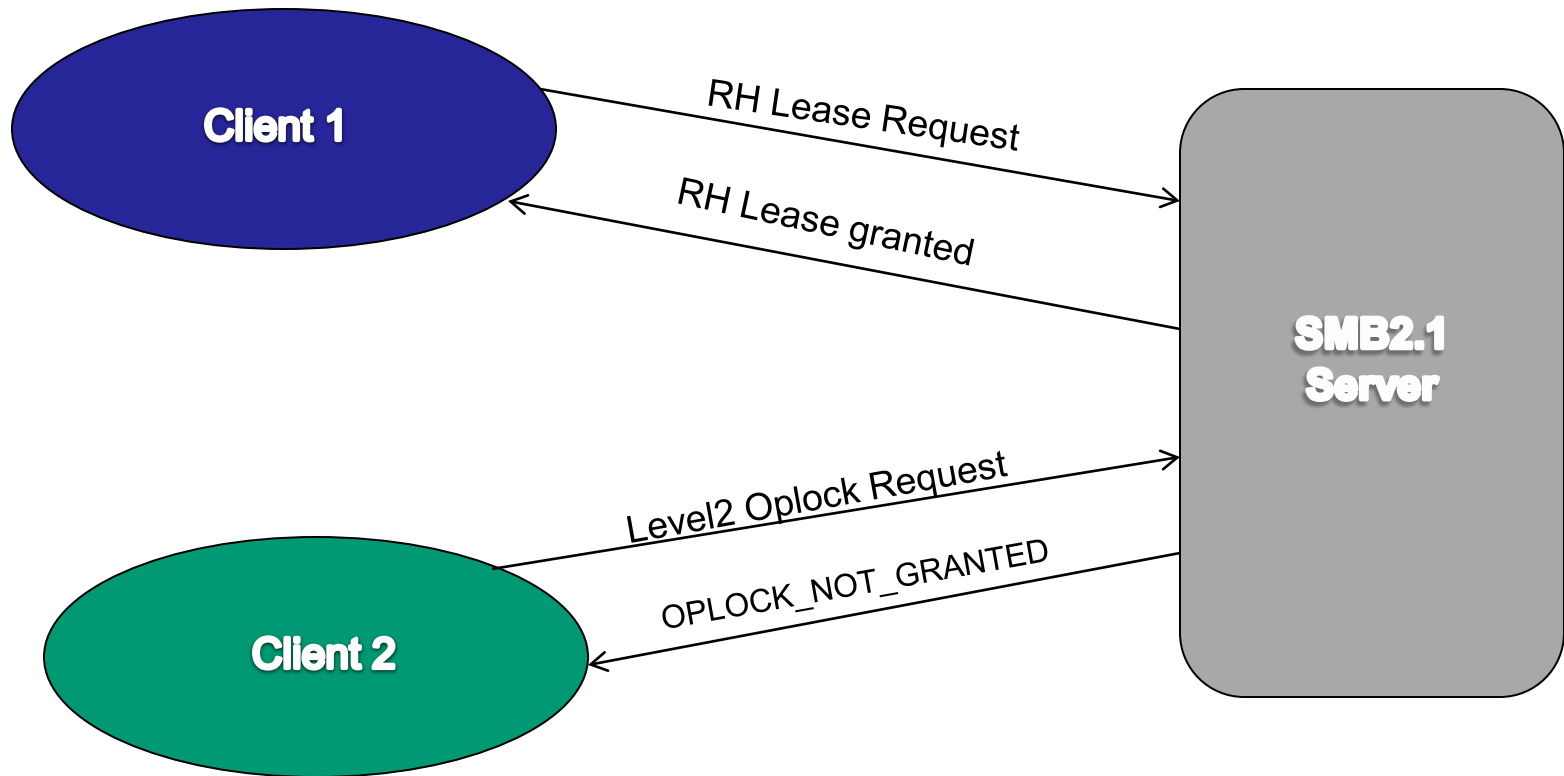
SMB2.1 Leases – Interaction between Read, RH and Level2 Oplocks



SMB2.1 Leases – Interaction between Read, RH and Level2 Oplocks



SMB2.1 Leases – Interaction between Read, RH and Level2 Oplocks



SMB2.1 Leases – Level2 and RH

- ❑ If we already have a RH oplock and a request for a Level2 oplock comes in for the same stream, the request is not granted.
- ❑ On the other hand, if we already have a Level2 oplock and a request for a RH oplock comes in for the same stream, a Read oplock is granted instead of RH
- ❑ Verified using SMB2-LEASE torture test against Windows server

Multithreaded SMB Server Improvements

Multithreaded Server Bottlenecks

- ❑ Thread synchronization
- ❑ Thread context-switch overhead

Thread Synchronization Pitfalls

- ❑ Lock Order Reversal
 - ❑ Especially between Connection and Session objects
- ❑ Starvation due to excess synchronization
 - ❑ Example locking the connection socket unnecessarily

Thread Context Switch

- ❑ The overhead of a context switch between threads is extremely high
- ❑ Internal testing revealed that a context switch can take upto 20us
- ❑ Try to minimize context switching as much as possible by performing direct execution of work instead of handing off to another thread

SMB2 vs SMB1 in OneFS

SMB2 vs SMB1

- ❑ SMB2 offers inherent performance benefits over SMB1
- ❑ We ran IOZone to measure the performance of SMB1 vs SMB2 in OneFS.

- ❑ IOZone is a filesystem benchmark tool. The benchmark generates and measures a variety of file operations.
- ❑ The important metrics for IOZone are the single and multi-client numbers for the following operations:
 - ❑ write
 - ❑ random-read (rread)
 - ❑ random-write (rwrite)
 - ❑ read

SMB2 vs SMB1 - IOZone

- ❑ Each SKU in the result table has two entries: one for a single thread, and one for peak numbers.
- ❑ Each entry specifies the number of threads used to drive the specified load.
- ❑ For Single lines, these will always be 1. For Peak lines the number of threads will vary depending on how many threads it took to drive that particular test to peak.
- ❑ IOZone block size used for these tests were 512 KB and the results are in MBps

IOZone Test Results – SMB1 vs SMB2

SMB1 Results

Nodes	Disk	S/P	Write/ Threads	Rread/ Threads	Rwrite/ Threads	Read/T
3	24	Single	141.56/1	17.70/1	108.16/1	195.66/1
3	24	1:1	402.91/3	43.95/3	213.80/3	551.95/3
3	24	Peak	893.46/57	245.10/57	730.74/57	1312.91/21

SMB2 Results

Nodes	Disk	S/P	Write/ Threads	Rread/ Threads	Rwrite/ Threads	Read/T
3	24	Single	227.13/1	27.80/1	134.06/1	313.90/1
3	24	1:1	640.08/3	57.97/3	266.22/3	917.12/3
3	24	Peak	1051.75/30	256.50/30	791.93/51	1513.79/12

Note: The results are not official numbers but rather experimental numbers to show the relative performance benefit of SMB2 over SMB1

SMB2 vs SMB1 - Conclusion

- ❑ SMB2 offers significant performance benefits over SMB1
- ❑ SMB2 clients lot better than SMB1 clients
- ❑ Our test results further highlight the fact that SMB2 performs better than SMB1, independent of the underlying operating system.

Questions?

Contact

Aravind Srinivasan

asrinivasan@isilon.com