Why A High Performance Parallel Storage Solution

Many of our end users are outgrowing the current storage architecture. Over the years many of our storage environments have consisted of one or more Network Attached Storage (NAS) arrays connected to the High Performance Computing (HPC) cluster. These NAS devices do offer an inexpensive solution for adding large blocks of storage however there are several limitations specifically around performance and scalability.

PSSC Labs is proposing the Parallux high performance parallel storage solution based on the Gluster File System (GlusterFS). Parallux offers significantly greater performance than the current NAS solution as well as the ability to easily scale storage capacity by simply adding more storage nodes. This white paper will outline PSSC Labs performance tests performed in house and theoretical performance based on different architectures.

What is GlusterFS

GlusterFS is an open source software package that aggregates multiple storage servers into a single, large parallel network file system. GlusterFS can be scaled to multiple petabytes (PBs) and thousands of clients. As more storage servers are added to GlusterFS, performance increases and eventually will be limited only by the network backplane.

The architecture of GlusterFS is simple. There is a server and client. Servers are known as storage bricks running a glusterfsd daemon to export a local file system as a volume[1]. The client can mount the created volume. Multiple clients can simultaneously mount the volume, allowing a cluster of high performance computing nodes to simultaneously access the storage servers. This will therefore increase data transfer of large files to a computational cluster.

Gluster Features

- Global name space
- Highly available storage
- Built in replication
- Self-healing
- Ability to re-balance data
- Compatible with POSIX-compatible File System (XFS preferred choice)

Testing GlusterFS Performance

In order to test GlusterFS performance, our engineered configured a basic cluster consisting of a single client and 3 storage servers. The hardware configuration of both the client and server is outlined below.

Client Server (Qty 1)	Storage Server (Qty 3)
• 2 x Xeon E5-2683v4 Processors • 256GB DDR 2400 MHz Memory • 120 GB Solid State SATA III (SSD) Hard Drive • Intel Omnpath 100 Gbps Network Adapter	• 2 x Xeon E5-2630v4 Processors • 256 GB DDR 2400 MHz Memory • 12 x 2 TB SATA III 7200 Hard Drive • LSI 9361-4i RAID Controller • Intel Omnpath 100 Gbps Network Adapter

Installation / Set UP

The Client Server and Storage Server were all connected via an Intel Omnipath 100 Gbps 48 Port Network Switch. Omnipath was selected over other network interconnects (ie 10GigE and Infiniband) because of the relatively low cost for higher bandwidth. Omnipath offers a simple TCP / IP network configuration. CentOS Linux 7.x was installed on the client and server nodes. The hard drives were configured for Raid 0 for maximum performance

Since there was only a single client system and ultimately we wanted to test performance across a "cluster" of client nodes, our engineers created multiple mountpoints on the client server to simulate multiple clients.

Performance Testing

To best test system performance the Flexible IO or "fio" utility was utilized. Fio is a utility we commonly use to compare performance of different configurations as well as stress test our systems. Our engineers primarily use fio to measure two types of storage performance: Throughput and IOPS.

For Throughput performance the following command was used:

```
fio --name=write --iodepth=1 --rw=write --bs=1M --direct=1 --size=1G --numjobs=7 --runtime=999 --group_reporting
```

For IOPS performance the following command was used:

```
fio --name=randwrite --iodepth=1 --rw=randwrite --bs=4k --direct=1 --size=1G --numjobs=8 --runtime=999 --group_reporting
```

Benchmark Results

The goal of the benchmark tests was to show performance of GlusterFS as it scaled from one server node to two and three server nodes.

Initially Fio was run across the client node and single server node. After results were recorded the same test was run across the client node and two server nodes and then three server nodes.

The Fio results are shown below.

Single Storage Server:

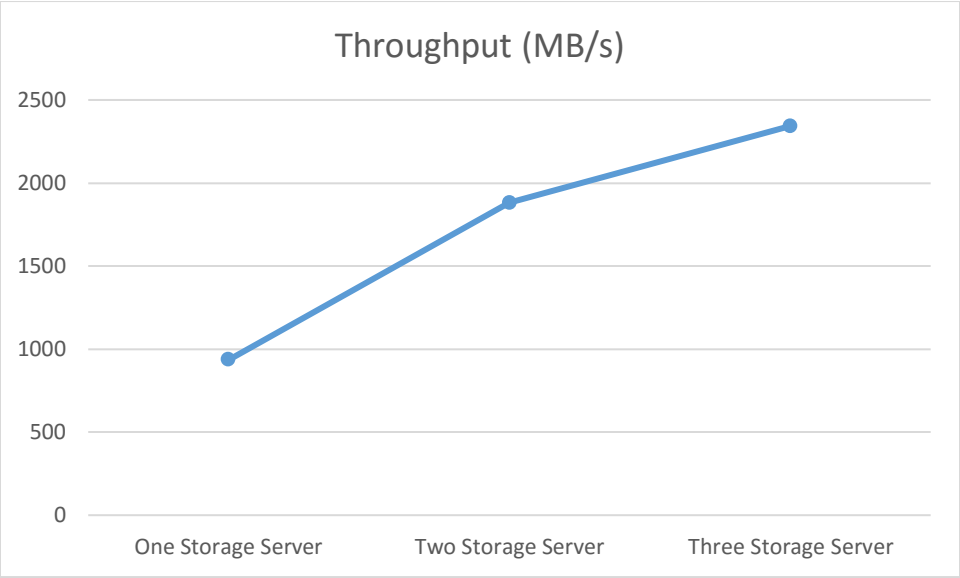
- 935 MB/s throughput
- 26400 IOPS

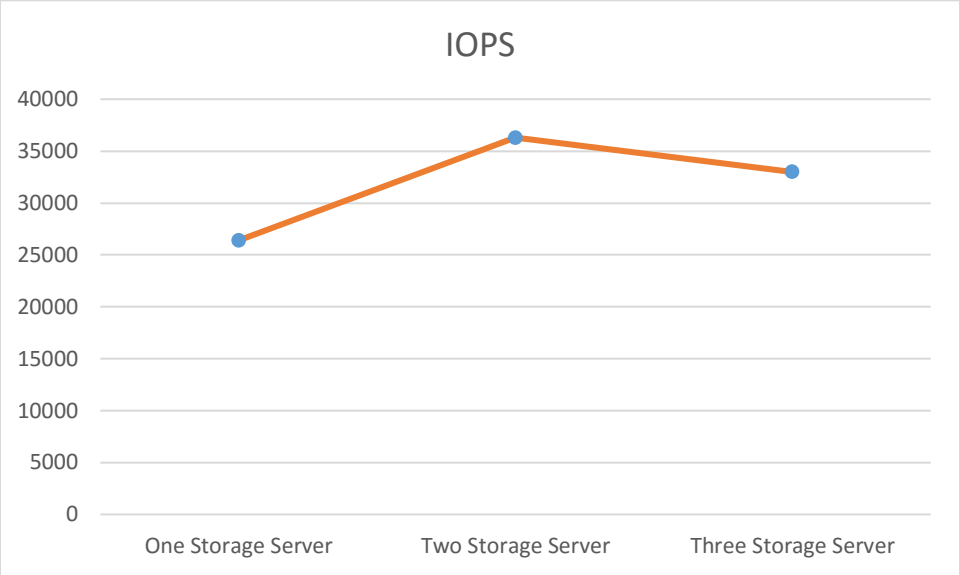
Two Storage Servers:

- 1881 MB/s throughput
- 36300 IOPS

Three Storage Serves:

- 2343 MB/s throughput
- 33000 IOPS





Technical Note: The above results were gathered from non-optimized fio benchmark results. Our engineers believe performance could be slightly improved by testing different system and software configurations. In addition the recorded benchmark numbers are rounded to within less than a 1% margin of error.

In the Appendix section the raw individual test results are included. For the throughput tests, reference the column 'maxb' (max bandwidth). For the IOPS tests, reference the the 'iops' column.

Performance on an SSD Based System

For the purpose of evaluating potential performance gains of a Solid State SATA III (SSD) our engineers configured a Surestore U2000SD platform with 24 x 1.0TB Micron M600 SSDs. Using the following command: dd if=/dev/zero of=/dev/sdb bs=1M count=99999, the platform was able to achieve **2.5 GB/s sustained Read and Write** performance. This represents a 2x to 3x performance improvement over the Storage Server based on SATAIII / SAS rotational hard drives.

Conclusion

Working with very large files (1TB+), NRL is most interested in throughput performance. Based on the benchmark results a parallel storage solutions running GlusterFS will offer significantly better perform than the current network attached storage systems. PSSC Labs has proposed two different Parallux configuration options: one based on SATAIII rotational hard drive and the other on SSD hard drives. The two options are briefly outlined below:

Option 1 – Parallux Storage Array (SATA III Hard Drives)
SureStore U4000HD High Performance Storage Array

- 2 x Xeon E5-2643 Processors / 6 Cores / 3.4 GHz
- 256GB System Memory
- 34 x 1 TB SATA III 7200 RPM Hard Drives
- Configured for Raid 60
- LSI 9361-4i High Performance Raid Controller
- Intel 100 Gbps Omnipath Network Backplane

Qty: 12 Storage Nodes
Total Usable Storage Space: Approx 330 TB

Option 2 – Parallux Storage Array (SSD Hard Drives)
SureStore U2000SD High Performance Storage Array

- 2 x Xeon E5-2643 Processors / 6 Cores / 3.4 GHz
- 256GB System Memory
- 24 x 2 TB Solid State SATAIII SSD Hard Drives
- Configured for Raid 60
- LSI 9361-4i High Performance Raid Controller
- Intel 100 Gbps Omnipath Network Backplane

Qty: 10 Storage Nodes
Total Usable Storage Space: Approx 360 TB

It is difficult to accurately forecast throughput performance without constructing the complete system. However based on the benchmark performance and the potential number of nodes for both hardware configuration options, the limiting factor may in fact be the network backplane. Using iperf network benchmark test, the best data transfer rates were measured to be ~5GB/s unidirectional. Bidirectional, that equates to roughly 80Gb/s which essentially maxes out the 100 Gb/s Omnpath network.

References

1 -- https://en.wikipedia.org/wiki/GlusterFS

Appendix A
RAW Fio Results

throughput1: WRITE: io=7168.0MB, aggrb=87091KB/s, minb=87091KB/s, maxb=87091KB/s, mint=84280msec, maxt=84280msec
throughput1: WRITE: io=7168.0MB, aggrb=88105KB/s, minb=88105KB/s, maxb=88105KB/s, mint=83310msec, maxt=83310msec
throughput1: WRITE: io=7168.0MB, aggrb=86848KB/s, minb=86848KB/s, maxb=86848KB/s, mint=84515msec, maxt=84515msec
throughput1: WRITE: io=7168.0MB, aggrb=85532KB/s, minb=85532KB/s, maxb=85532KB/s, mint=85816msec, maxt=85816msec
throughput1: WRITE: io=7168.0MB, aggrb=85409KB/s, minb=85409KB/s, maxb=85409KB/s, mint=85939msec, maxt=85939msec
throughput1: WRITE: io=7168.0MB, aggrb=85727KB/s, minb=85727KB/s, maxb=85727KB/s, mint=85621msec, maxt=85621msec
throughput1: WRITE: io=7168.0MB, aggrb=85459KB/s, minb=85459KB/s, maxb=85459KB/s, mint=85889msec, maxt=85889msec
throughput1: WRITE: io=7168.0MB, aggrb=85124KB/s, minb=85124KB/s, maxb=85124KB/s, mint=86227msec, maxt=86227msec
throughput1: WRITE: io=7168.0MB, aggrb=84081KB/s, minb=84081KB/s, maxb=84081KB/s, mint=87297msec, maxt=87297msec
throughput1: WRITE: io=7168.0MB, aggrb=83775KB/s, minb=83775KB/s, maxb=83775KB/s, mint=87616msec, maxt=87616msec
throughput1: WRITE: io=7168.0MB, aggrb=85501KB/s, minb=85501KB/s, maxb=85501KB/s, mint=85847msec, maxt=85847msec

iops1: write: io=2255.2MB, bw=9583.2KB/s, iops=2395, runt=240976msec
iops1: write: io=2327.8MB, bw=9914.9KB/s, iops=2478, runt=240402msec
iops1: write: io=2286.3MB, bw=9728.5KB/s, iops=2432, runt=240647msec
iops1: write: io=2303.7MB, bw=9815.6KB/s, iops=2453, runt=240324msec
iops1: write: io=2354.4MB, bw=10018KB/s, iops=2504, runt=240666msec
iops1: write: io=2276.5MB, bw=9686.9KB/s, iops=2421, runt=240639msec
iops1: write: io=2250.2MB, bw=9651.4KB/s, iops=2412, runt=238742msec
iops1: write: io=2325.9MB, bw=9906.1KB/s, iops=2476, runt=240406msec
iops1: write: io=2361.7MB, bw=10054KB/s, iops=2513, runt=240536msec
iops1: write: io=2283.4MB, bw=9719.2KB/s, iops=2429, runt=240570msec
iops1: write: io=2260.2MB, bw=9627.3KB/s, iops=2406, runt=240402msec

throughput2: WRITE: io=7168.0MB, aggrb=173355KB/s, minb=173355KB/s, maxb=173355KB/s, mint=42341msec, maxt=42341msec
throughput2: WRITE: io=7168.0MB, aggrb=176176KB/s, minb=176176KB/s, maxb=176176KB/s, mint=41663msec, maxt=41663msec
throughput2: WRITE: io=7168.0MB, aggrb=173044KB/s, minb=173044KB/s, maxb=173044KB/s, mint=42417msec, maxt=42417msec
throughput2: WRITE: io=7168.0MB, aggrb=170690KB/s, minb=170690KB/s, maxb=170690KB/s, mint=43002msec, maxt=43002msec
throughput2: WRITE: io=7168.0MB, aggrb=169136KB/s, minb=169136KB/s, maxb=169136KB/s, mint=43397msec, maxt=43397msec
throughput2: WRITE: io=7168.0MB, aggrb=170211KB/s, minb=170211KB/s, maxb=170211KB/s, mint=43123msec, maxt=43123msec
throughput2: WRITE: io=7168.0MB, aggrb=171528KB/s, minb=171528KB/s, maxb=171528KB/s, mint=42792msec, maxt=42792msec
throughput2: WRITE: io=7168.0MB, aggrb=170555KB/s, minb=170555KB/s, maxb=170555KB/s, mint=43036msec, maxt=43036msec
throughput2: WRITE: io=7168.0MB, aggrb=185583KB/s, minb=185583KB/s, maxb=185583KB/s, mint=39551msec, maxt=39551msec
throughput2: WRITE: io=7168.0MB, aggrb=170921KB/s, minb=170921KB/s, maxb=170921KB/s, mint=42944msec, maxt=42944msec
throughput2: WRITE: io=7168.0MB, aggrb=161315KB/s, minb=161315KB/s, maxb=161315KB/s, mint=45501msec, maxt=45501msec

iops2: write: io=1761.9MB, bw=13231KB/s, iops=3307, runt=136350msec
iops2: write: io=1788.6MB, bw=13442KB/s, iops=3360, runt=136211msec
iops2: write: io=1753.4MB, bw=13232KB/s, iops=3307, runt=135693msec
iops2: write: io=1794.5MB, bw=13755KB/s, iops=3438, runt=133584msec
iops2: write: io=1687.5MB, bw=12726KB/s, iops=3181, runt=135779msec
iops2: write: io=1754.8MB, bw=13227KB/s, iops=3306, runt=135843msec
iops2: write: io=1758.4MB, bw=13202KB/s, iops=3300, runt=136358msec
iops2: write: io=1741.5MB, bw=13212KB/s, iops=3303, runt=134966msec
iops2: write: io=1768.2MB, bw=13526KB/s, iops=3381, runt=133858msec
iops2: write: io=1799.6MB, bw=13638KB/s, iops=3409, runt=135116msec
iops2: write: io=1698.5MB, bw=12923KB/s, iops=3230, runt=134582msec

throughput3: WRITE: io=7168.0MB, aggrb=214846KB/s, minb=214846KB/s, maxb=214846KB/s, mint=34164msec, maxt=34164msec
throughput3: WRITE: io=7168.0MB, aggrb=211729KB/s, minb=211729KB/s, maxb=211729KB/s, mint=34667msec, maxt=34667msec
throughput3: WRITE: io=7168.0MB, aggrb=211796KB/s, minb=211796KB/s, maxb=211796KB/s, mint=34656msec, maxt=34656msec
throughput3: WRITE: io=7168.0MB, aggrb=213441KB/s, minb=213441KB/s, maxb=213441KB/s, mint=34389msec, maxt=34389msec
throughput3: WRITE: io=7168.0MB, aggrb=211066KB/s, minb=211066KB/s, maxb=211066KB/s, mint=34776msec, maxt=34776msec
throughput3: WRITE: io=7168.0MB, aggrb=212933KB/s, minb=212933KB/s, maxb=212933KB/s, mint=34471msec, maxt=34471msec
throughput3: WRITE: io=7168.0MB, aggrb=211503KB/s, minb=211503KB/s, maxb=211503KB/s, mint=34704msec, maxt=34704msec
throughput3: WRITE: io=7168.0MB, aggrb=211662KB/s, minb=211662KB/s, maxb=211662KB/s, mint=34678msec, maxt=34678msec
throughput3: WRITE: io=7168.0MB, aggrb=215363KB/s, minb=215363KB/s, maxb=215363KB/s, mint=34082msec, maxt=34082msec
throughput3: WRITE: io=7168.0MB, aggrb=208719KB/s, minb=208719KB/s, maxb=208719KB/s, mint=35167msec, maxt=35167msec
throughput3: WRITE: io=7168.0MB, aggrb=216641KB/s, minb=216641KB/s, maxb=216641KB/s, mint=33881msec, maxt=33881msec

iops3: write: io=1612.7MB, bw=9875.3KB/s, iops=2468, runt=167160msec
iops3: write: io=1724.5MB, bw=12260KB/s, iops=3065, runt=144029msec
iops3: write: io=1715.2MB, bw=10558KB/s, iops=2639, runt=166328msec
iops3: write: io=1679.6MB, bw=10273KB/s, iops=2568, runt=167415msec
iops3: write: io=1656.4MB, bw=10280KB/s, iops=2570, runt=164992msec
iops3: write: io=1634.5MB, bw=10025KB/s, iops=2506, runt=166943msec
iops3: write: io=1622.6MB, bw=9957.6KB/s, iops=2489, runt=166863msec
iops3: write: io=1742.2MB, bw=10720KB/s, iops=2679, runt=166420msec
iops3: write: io=1704.8MB, bw=11886KB/s, iops=2971, runt=146804msec
iops3: write: io=1738.9MB, bw=10903KB/s, iops=2725, runt=163309msec
iops3: write: io=1651.3MB, bw=10151KB/s, iops=2537, runt=166569msec
iops3: write: io=1621.6MB, bw=9993.3KB/s, iops=2498, runt=166158msec
iops3: write: io=1730.6MB, bw=10636KB/s, iops=2658, runt=166615msec