IBM Spectrum Scale Performance and sizing update

Sven Oehme
Chief Research Strategist Spectrum Scale
IBM Research

ESS Packaging Options

Elastic Storage Server (ESS) is a prepacked solution using on the GNR software.[1]
It comes in various models configured with different HW:

- SSD Models (400/800 GB)
  - GS1, GS2, GS4
  - 2 x High Volume Servers
  - 1/2/4 x JBOD disk enclosures
- 10,000 RPM Models (1.2 TB)
  - GS2, GS4, GS6
  - 2 x High Volume Servers
  - 2/4/6 x JBOD disk enclosures
- NL-SAS Models (2/4/6 TB)
  - GL2, GL4,GL6
  - 2 x High Volume Servers
  - 2/4/6 x JBOD disk enclosures

ESS HW Components

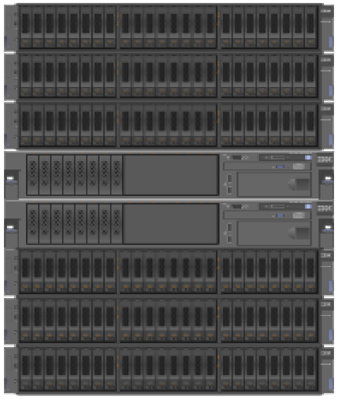
Servers

JBOD Enclosure
2U x 24, 2.5" disks

JBOD Enclosure
4U x 60, 3.5" disks

No storage controller!

GS2



GS6

GL6

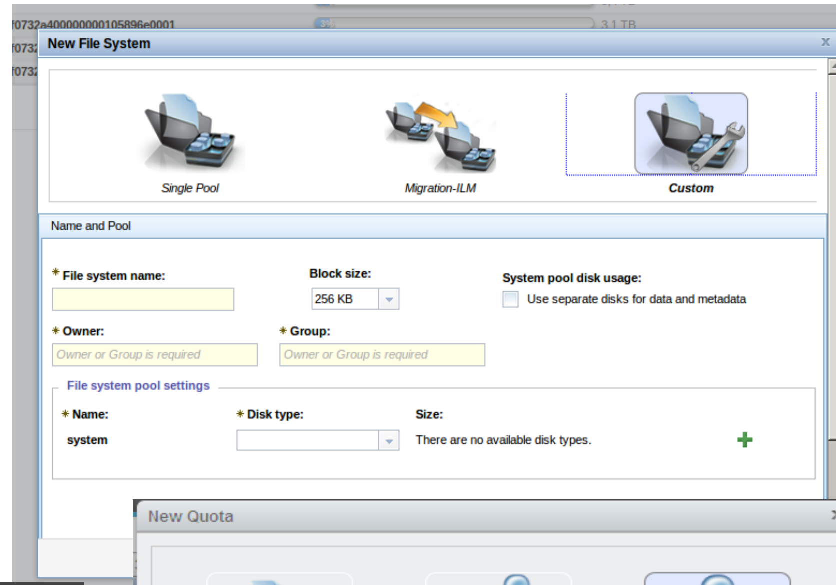
1. Unlike traditional GPFS which communicates with an external block storage controller, GNR is a software storage controller that runs within GPFS, directly managing and communicating with disks.

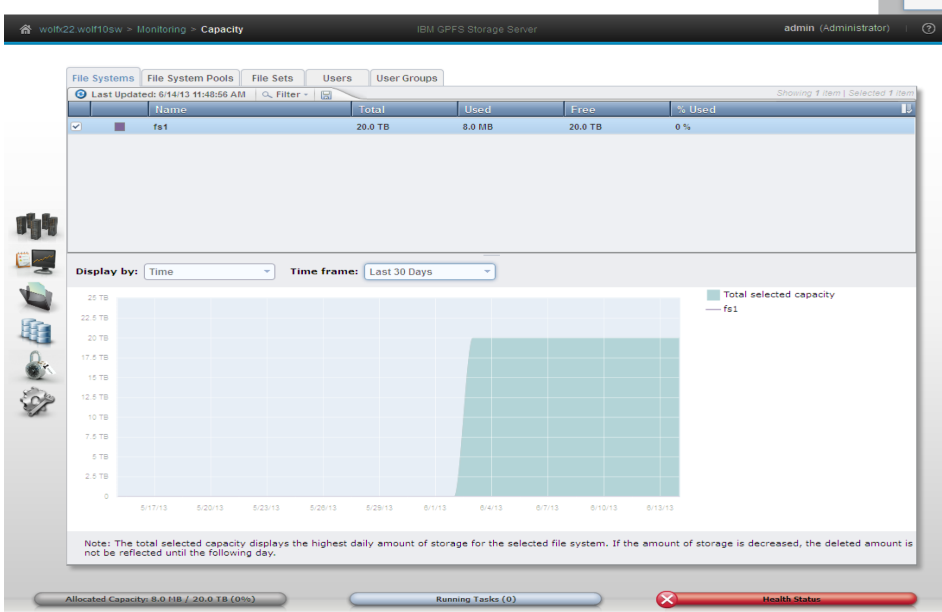
Right Model for the required Size (incomplete List)

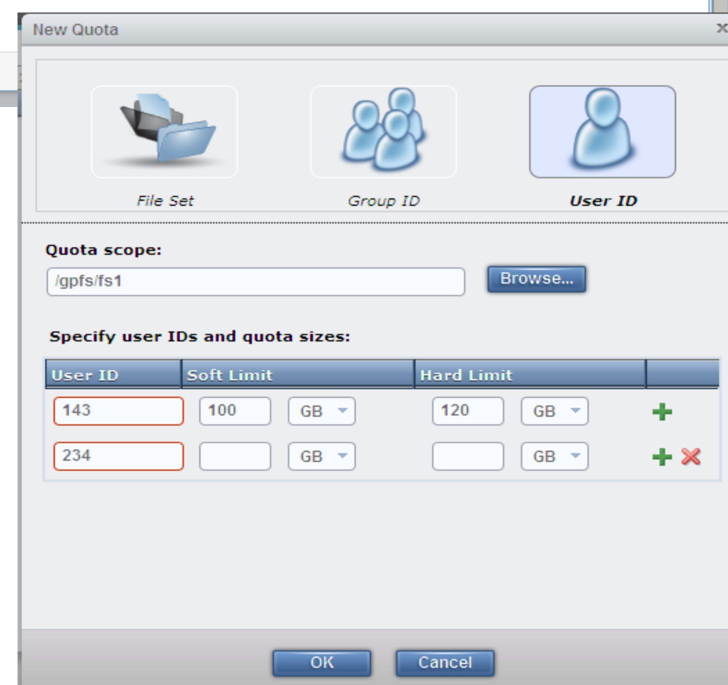
Model	Disk size	Redundancy	Nr. Drives	Raw (TB)	Usable (TB)
GL2	2	8+2P	116	232	170
GL2	2	8+3P	116	232	152
GL2	4	8+2P	116	464	340
GL2	4	8+3P	116	464	305
GL2	6	8+2P	116	696	510
GL2	6	8+3P	116	696	458
GL4	2	8+2P	232	464	340
GL4	2	8+3P	232	464	305
GL4	4	8+2P	232	928	680
GL4	4	8+3P	232	928	610
GL4	6	8+2P	232	1392	1020
GL4	6	8+3P	232	1392	916
GL6	2	8+2P	348	696	510
GL6	2	8+3P	348	696	458
GL6	4	8+2P	348	1392	1020
GL6	4	8+3P	348	1392	916
GL6	6	8+2P	348	2088	1530
GL6	6	8+3P	348	2088	1376

Graphical Management

- Provide an easy-to-use Graphical User Interface for common tasks
  - System Monitoring
  - System Maintenance
  - User Configuration
- Base interface on common IBM Storage Framework
  - Comfortable for users of other IBM technologies







Quick Intro into GPFS Native Raid (GNR)

- **Declustered RAID**
  - Data and parity stripes are uniformly partitioned and distributed across a disk array.
  - Arbitrary number of disks per array (unconstrained to an integral number of RAID stripe widths)
  - All disks used during normal operation (no idle *spares*) and all disks used during rebuild

- **2-fault and 3-fault tolerance (RAID-D2, RAID-D3)**
  - Reed-Solomon parity encoding 2 or 3-fault-tolerant: stripes = 8 data strips + 2 or 3 parity strips
  - 3 or 4-way mirroring

- **End-to-end checksum**
  - Disk surface to Spectrum Scale user/client
  - Detects and corrects off-track and lost/dropped disk writes

- **Asynchronous error diagnosis while affected IOs continue on**
  - If media error: verify and restore if possible
  - If path problem: attempt alternate paths

- **Advanced fault determination**
  - Statistical reliability and SMART monitoring
  - Neighbor check, drive power cycling
  - Media error detection and correction
  - Slow drive detection and handling

- **Supports concurrent disk, enclosure and server firmware updates**

Performance data

None of the following Performance numbers should be reused for sales or contract purposes.

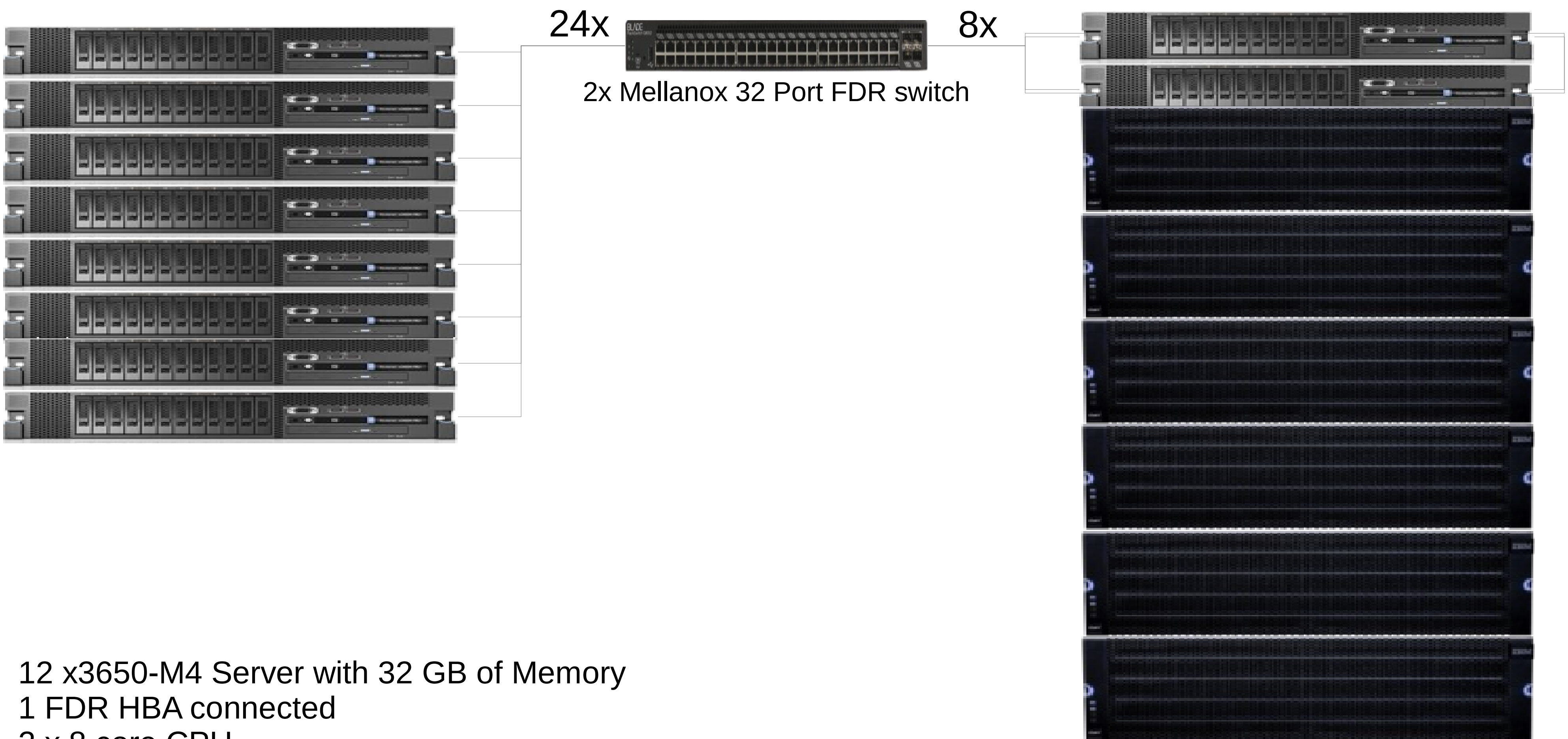
ESS Performance is typically Network bound, therefore the achievable Performance in Production depends heavily on used Network Technology and its scaling capabilities

Typical limits of Infiniband based Systems is ~25 GB/sec
Typical limits of 40GB based Systems is ~14 GB/sec
Typical limits of 10GB based Systems is ~10 GB/sec

Even if the specific ESS device is faster than above Numbers we can't guarantee the achievement of this results

A word of caution : The achieved numbers depends on the right Client configuration and good Interconnect and can vary between environments. They should not be used in RFI's as committed numbers, rather to demonstrate the technical capabilities of the Product in good conditions

Single building block Benchmark Setup



12 x3650-M4 Server with 32 GB of Memory
1 FDR HBA connected
2 x 8 core CPU
RHEL 7.1
GPFS 4.1.0.8
OFED-2.4-1.0.4

1 ESS GL6 System – Version 3.0
2 FDR HBA's connected per Server
GPFS 4.1.0.8 code level
OFED-2.4-1.0.2

Performance data IOR execution command line

```
Summary:
        api                = POSIX
        test filename      = /ibm/fs2-1m-p01/shared/ior//iorfile
        access             = file-per-process
        ordering in a file = sequential offsets
        ordering inter file= no tasks offsets
        clients            = 32 (4 per node)
        repetitions        = 100
        xfersize           = 1 MiB
        blocksize          = 128 GiB
        aggregate filesize = 4096 GiB
```

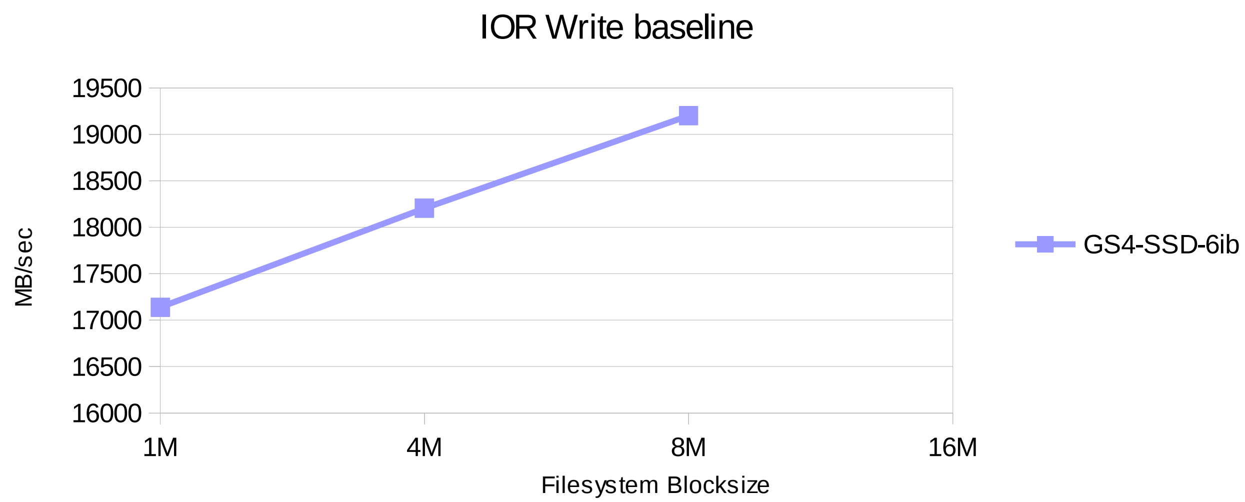
A word of caution : The achieved numbers depends on the right Client configuration and good Interconnect and can vary between environments. They should not be used in RFI's as committed numbers, rather to demonstrate the technical capabilities of the Product in good conditions

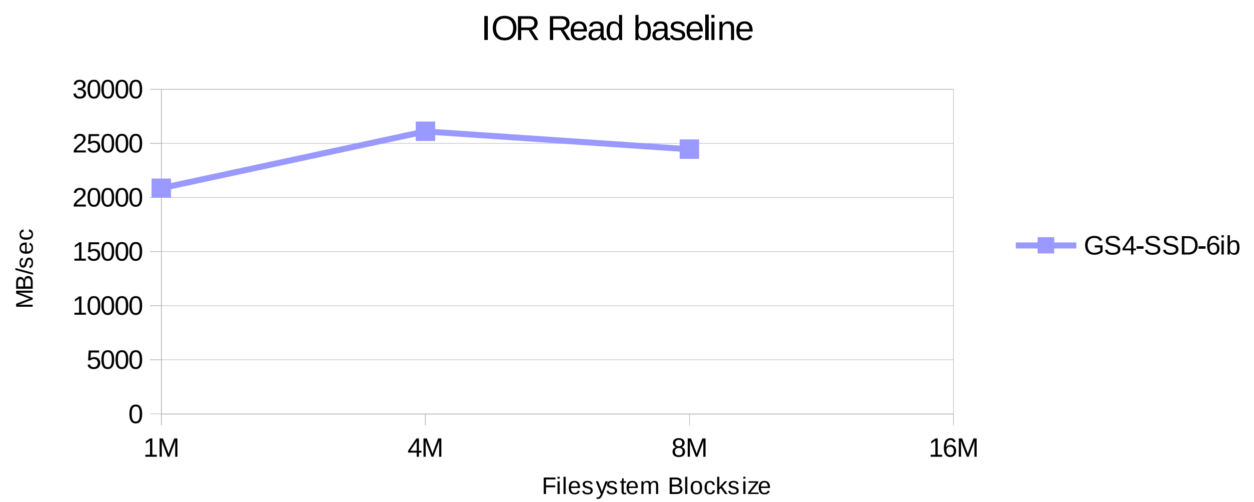
GS4-SSD Benchmark Results – various Blocksizes

Filesystem Blocksize	Write MB/sec	Read MB/sec
1 MB	17139	20858
4 MB	18205	26110
8 MB	19201	24457
16 MB	-	-

A word of caution : The achieved numbers depends on the right Client configuration and good Interconnect and can vary between environments. They should not be used in RFI's as commited numbers, rather to demonstrate the technical capabilities of the Product in good conditions

GS4-SSD Benchmark Results – various Blocksizes





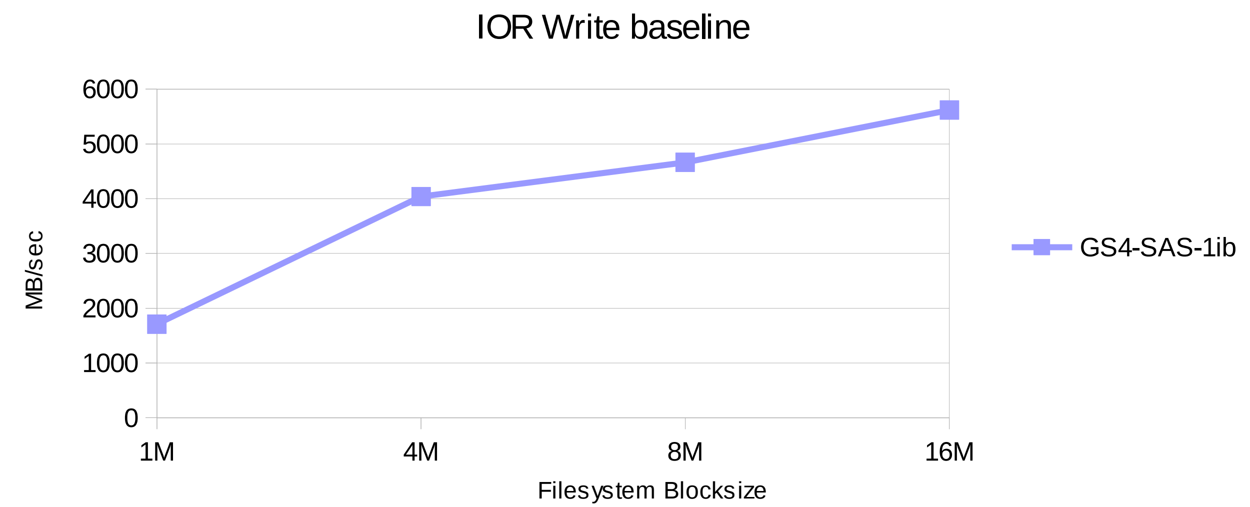
A word of caution : The achieved numbers depends on the right Client configuration and good Interconnect and can vary between environments. They should not be used in RFI's as commited numbers, rather to demonstrate the technical capabilities of the Product in good conditions

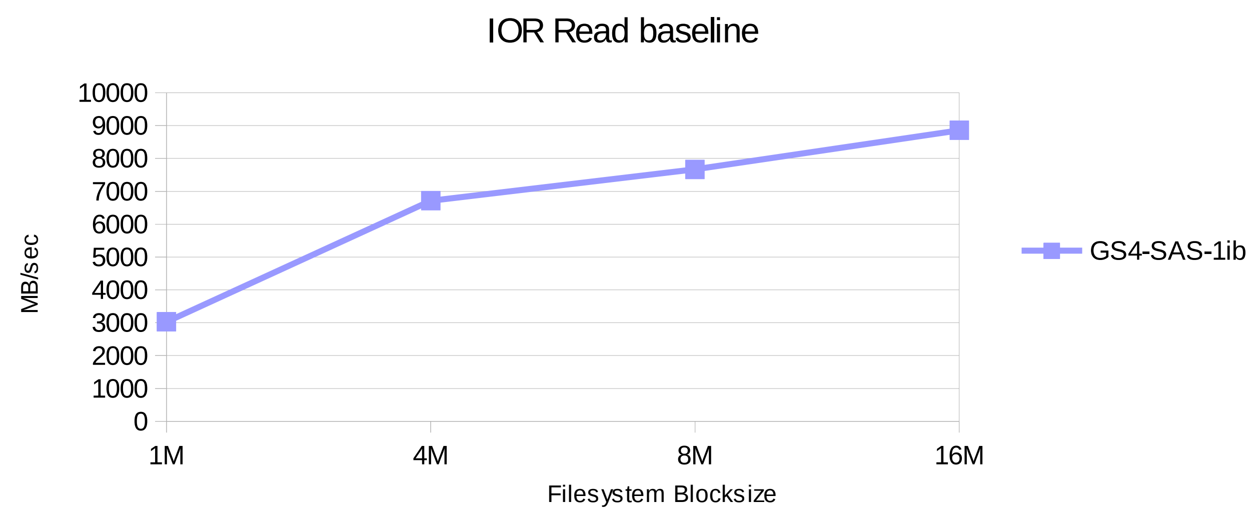
GS4-SAS Benchmark Results – various Blocksizes

Filesystem Blocksize	Write MB/sec	Read MB/sec
1 MB	1709	3029
4 MB	4039	6715
8 MB	4665	7666
16 MB	5619	8858

A word of caution : The achieved numbers depends on the right Client configuration and good Interconnect and can vary between environments. They should not be used in RFI's as commited numbers, rather to demonstrate the technical capabilities of the Product in good conditions

GS4-SAS Benchmark Results – various Blocksizes





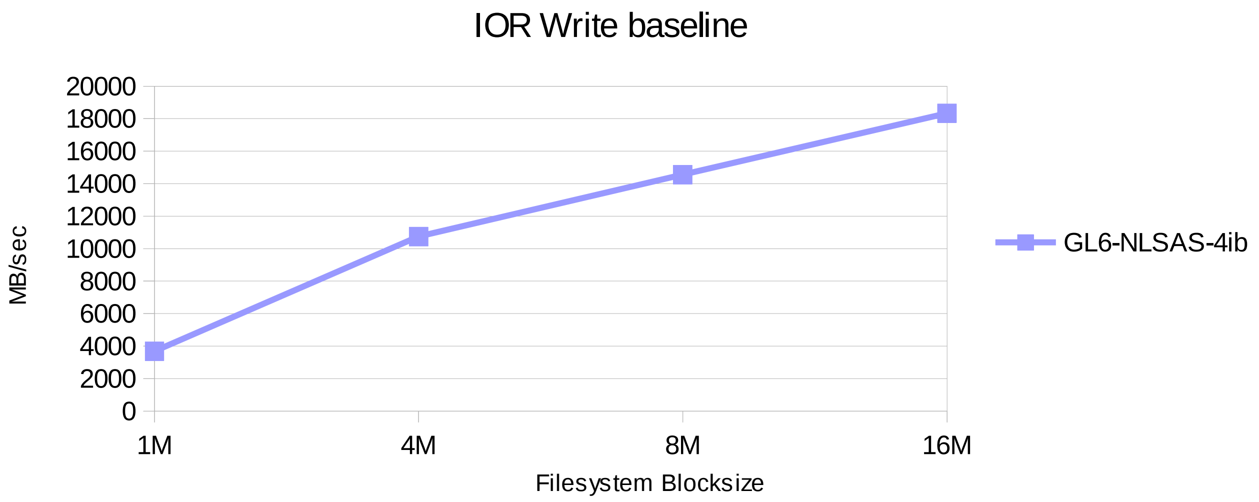
A word of caution : The achieved numbers depends on the right Client configuration and good Interconnect and can vary between environments. They should not be used in RFI's as commited numbers, rather to demonstrate the technical capabilities of the Product in good conditions

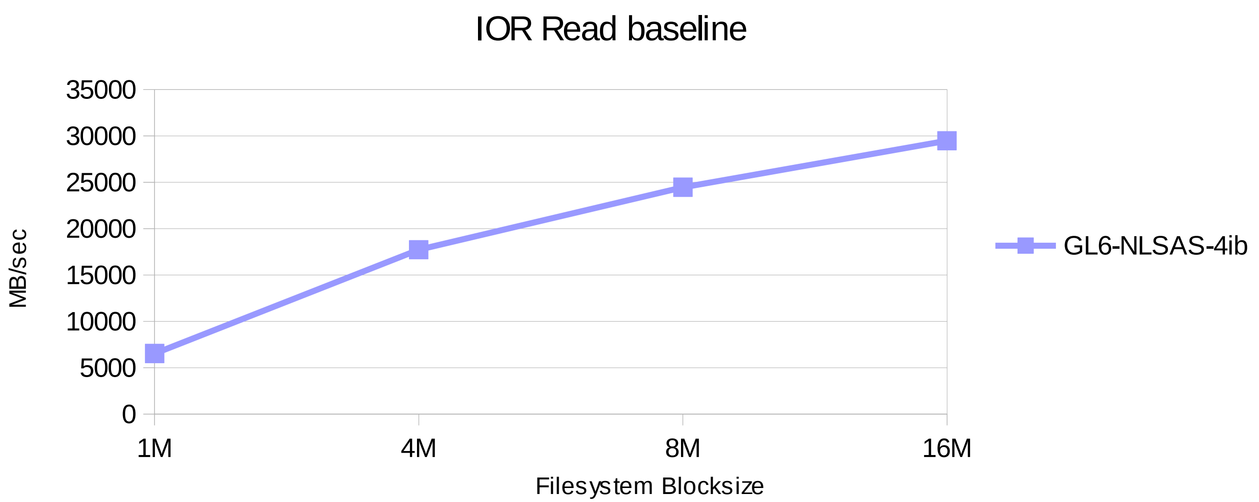
GL6 Benchmark Results – various Blocksizes

Filesystem Blocksize	Write MB/sec	Read MB/sec
1 MB	3681	6516
4 MB	10748	17725
8 MB	14552	24458
16 MB	18337	29481

A word of caution : The achieved numbers depends on the right Client configuration and good Interconnect and can vary between environments. They should not be used in RFI's as commited numbers, rather to demonstrate the technical capabilities of the Product in good conditions

GL6 Benchmark Results – various Blocksizes





A word of caution : The achieved numbers depends on the right Client configuration and good Interconnect and can vary between environments. They should not be used in RFI's as commited numbers, rather to demonstrate the technical capabilities of the Product in good conditions

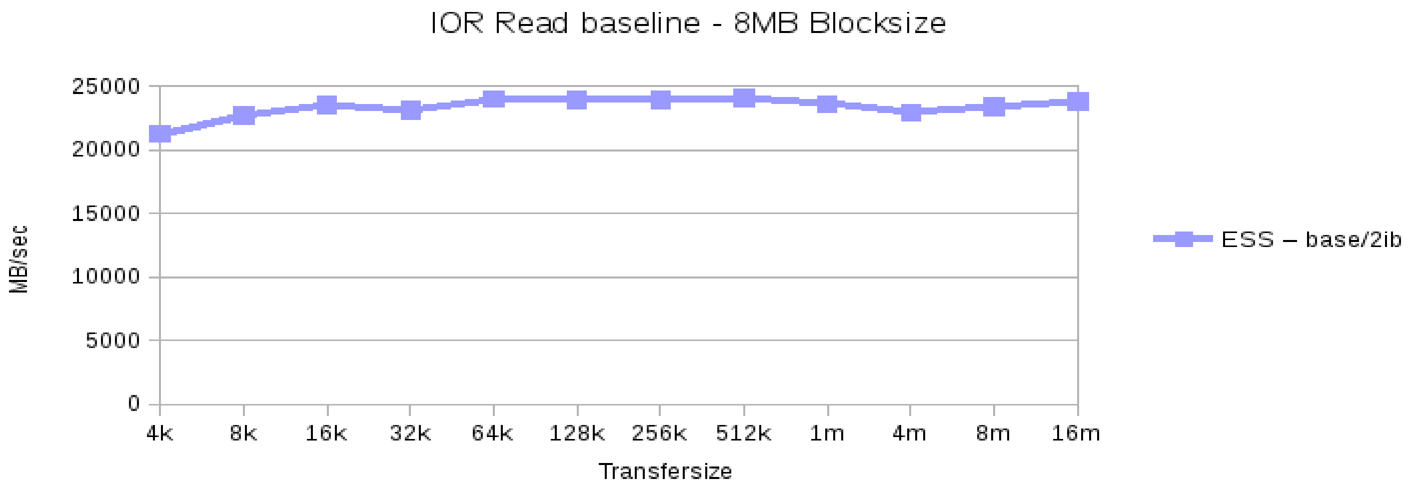
GL6 Benchmark Results – various Transfersizes – 8MB Blocksize

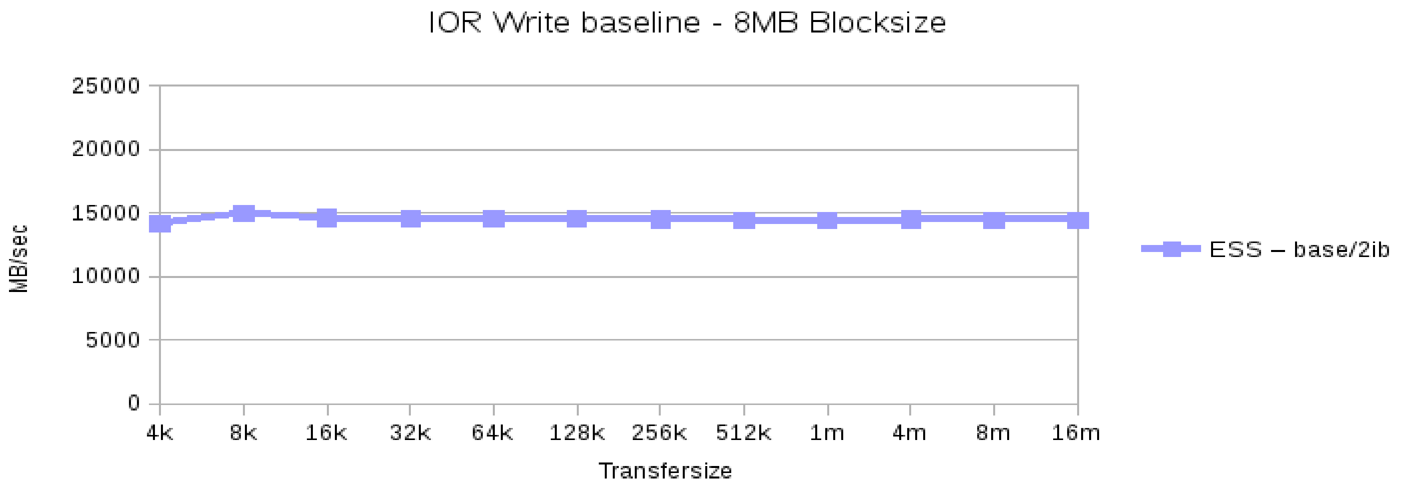
Transfersize	Write MB/sec	Read MB/sec
1 MB	14304.18	23049.39
4 MB	14439.93	23156.19
8 MB	14804.21	23297.08
16 MB	14583.56	21324.18

As one can see from the data, the transfersize has minimal impact on the overall throughput

A word of caution : The achieved numbers depends on the right Client configuration and good Interconnect and can vary between environments. They should not be used in RFI's as commited numbers, rather to demonstrate the technical capabilities of the Product in good conditions

GL6 Benchmark Results – various Transfersizes – 8MB Blocksize

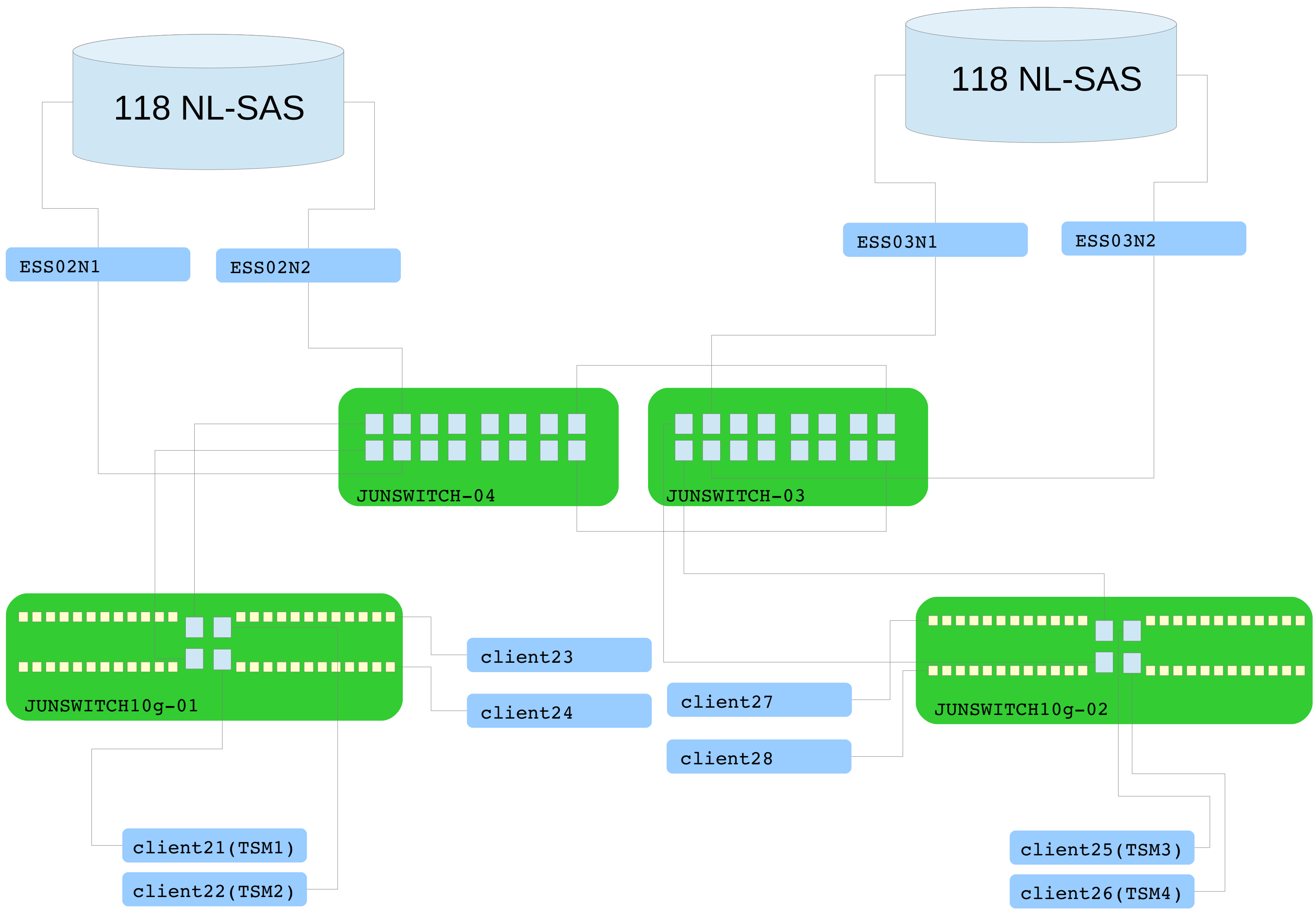




A word of caution : The achieved numbers depends on the right Client configuration and good Interconnect and can vary between environments. They should not be used in RFI's as commited numbers, rather to demonstrate the technical capabilities of the Product in good conditions

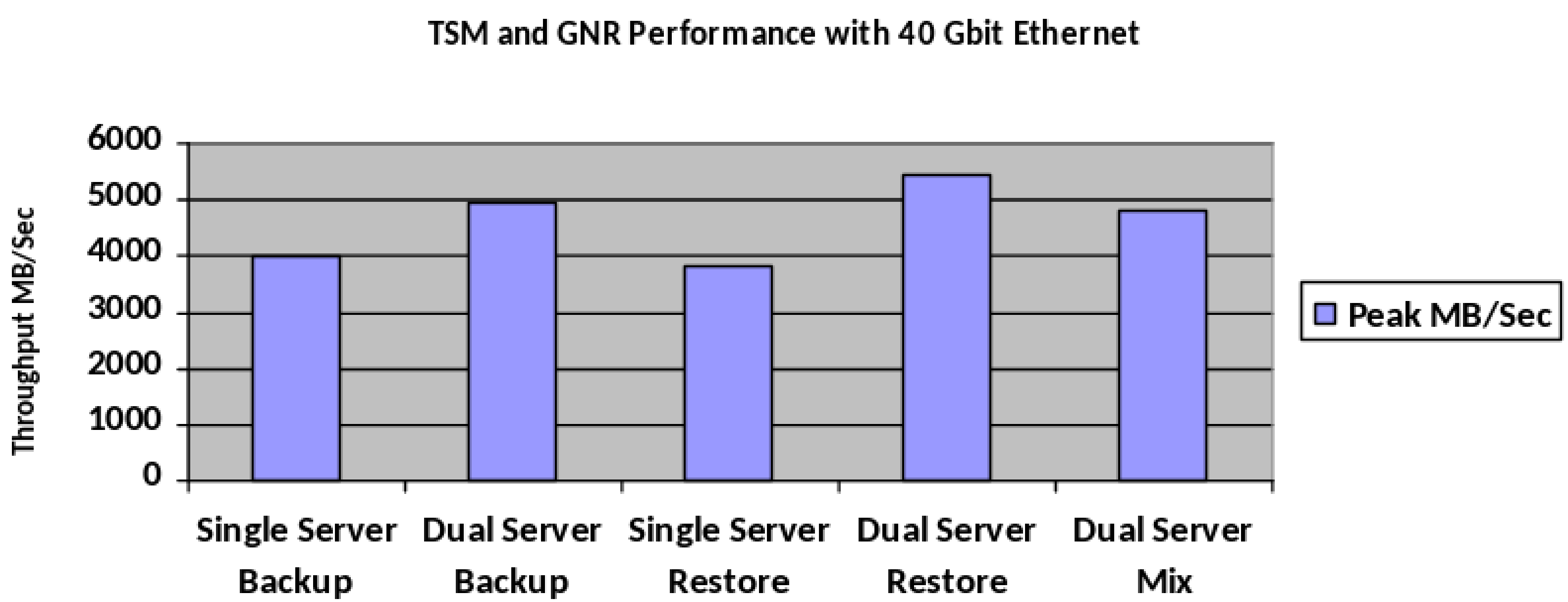
Application Specific Simulations

TSM performance testing with GL2(118 NL-SAS) *



TSM performance testing with GL2(118 NL-SAS) *

```
Peak backup performance with multiple sessions from a single TSM server:      4017 MB/sec
Peak backup performance with multiple sessions from two TSM servers is:       4981 MB/sec
Peak restore performance with multiple sessions for a single TSM server is:   3834 MB/sec
Peak restore performance with multiple session from two TSM server is:        5424 MB/sec
Peak mixed workload performance from two TSM servers is.           4821 MB/sec
```

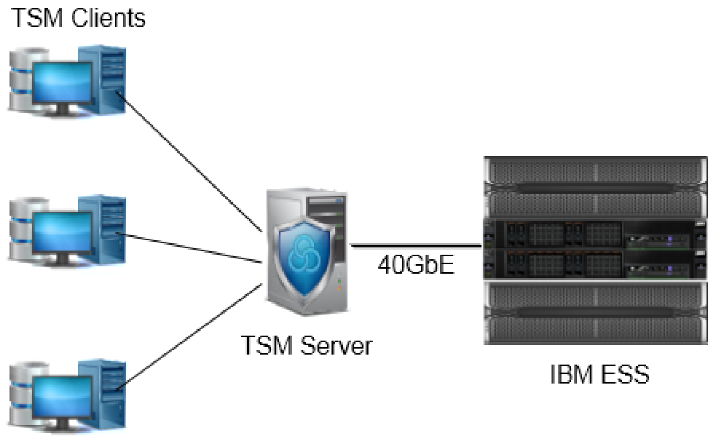


* Performance was limited by drive count in single GL2 device

TSM Blueprint: Spectrum Protect with Elastic Storage Server GL2

Enhancements planned for 2Q15

- Support for IBM Elastic Storage Server (ESS)
  - Configuration instructions for large TSM server with ESS GL-2
  - Configuration script support for automating TSM server setup with ESS
  - Initially published for Linux x86_64



- Check https://ibm.biz/TivoliStorageManagerBlueprints for availability of the TSM Blueprint

Details of the Test Results available

Storageneers

[Scale out backup with TSM and GSS: Performance test results]

[Elastic Storage with GPFS Native RAID performance test results with Tivoli Storage Manager over 40 GBit Ethernet]

The Register®

The peak TSM/Isilon throughput was 800MB/sec while the TSM/GPFS throughput was 5.4GB/sec (5,400MB/sec) – almost seven times faster. It's not an apples for apples comparison, but it clearly shows that Isilon is not the only fruit and GPFS could be a more flavoursome fruitstuff.
With these results, acronymically TSM could stand for The Speed Machine. ®

[Mirror, mirror on the wall, who has the best TSM backend of all?]

[Big Blue stuffs data into backup at GIGABYTES/sec]

Digital Video Simulation with Specsfs2014 VDA Benchmark

Business Metric	Requested Op Rate	Achieved Op Rate	Avg Lat (ms)	Total KBps	Read KBps	Write KBps	Run Sec	# Cl	Cl Proc	Avg File Size KB	Cl Data Set MiB	Start Data Set MiB	Init File Set MiB	Max File Space MiB
200	2000.00	2000.44	9.55	925240.69	79274.40	845966.29	300	10	40	1048576	450560	4505600	4505600	4915200
400	4000.00	4000.78	11.87	1843411.66	157015.82	1686395.84	300	10	80	1048576	901120	9011200	9011200	9830400
600	6000.00	6001.13	14.26	2760527.32	234360.60	2526166.72	300	10	120	1048576	1351680	13516800	13516800	14745600
800	8000.00	8001.69	19.44	3693372.00	313248.66	3380123.34	300	10	160	1048576	1802240	18022400	18022400	19660800
1000	10000.00	10001.38	22.63	4613519.11	394198.15	4219320.96	300	10	200	1048576	2252800	22528000	22528000	24576000
1200	12000.00	12001.93	29.69	5533122.30	471924.10	5061198.21	300	10	240	1048576	2703360	27033600	27033600	29491200
1400	14000.00	14001.86	38.63	6457107.05	550226.58	5906880.46	300	10	280	1048576	3153920	31539200	31539200	34406400
1600	16000.00	16000.77	47.30	7376301.04	630555.06	6745745.97	300	10	320	1048576	3604480	36044800	36044800	39321600
1800	18000.00	17999.80	70.28	8300488.87	709605.89	7590882.98	300	10	360	1048576	4055040	40550400	40550400	44236800
2000	20000.00	19517.72	125.49	9004564.53	789208.01	8215356.52	300	10	400	1048576	4505600	45056000	45056000	49152000

Object Storage leveraging ESS GL6 - Setup

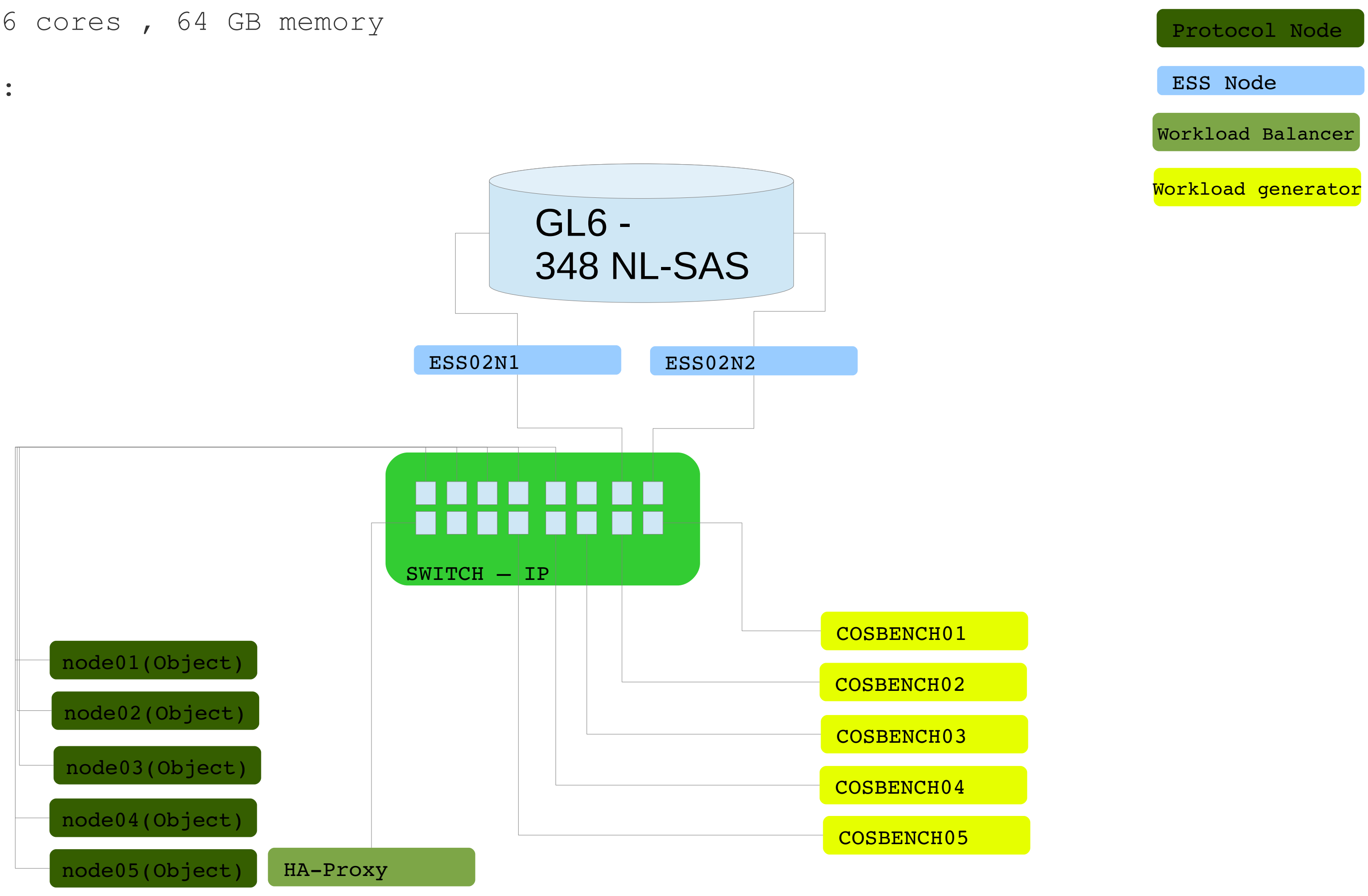
HW Setup :

Spectrum Scale Object Protocol Nodes:

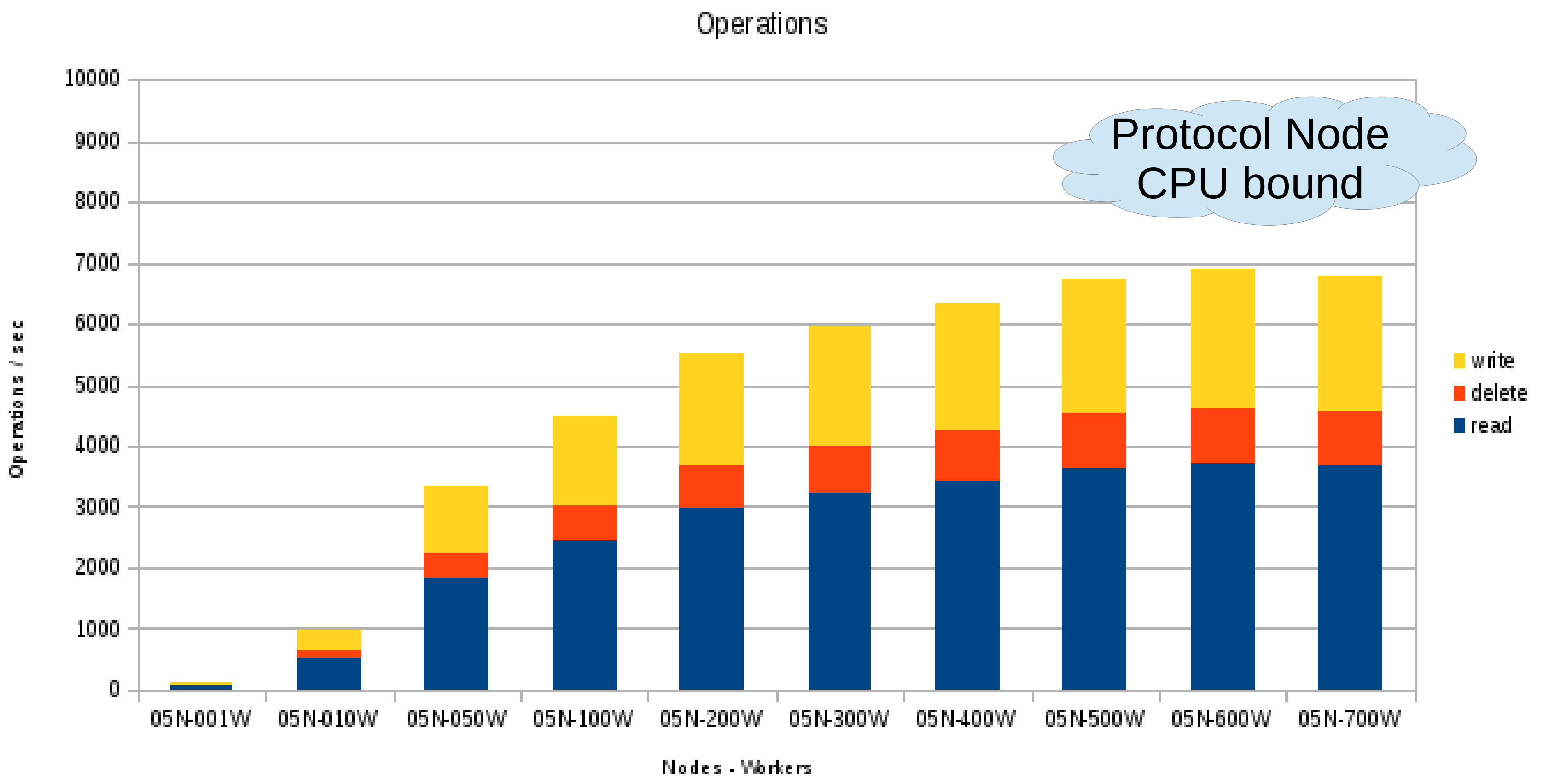
5 x IBM X3650 M4 with 16 cores , 64 GB memory

Spectrum Scale Storage :

1 x ESS GL6



5 Nodes, 4 Containers, 15K Object Size with 1 – 700 Workers



****This are preliminary numbers with the upcoming 4.1.1 release without any significant tuning effort**

50 Worker Read/Write 10MB files – 1.5 GB/sec

HA Proxy Limited

```
ID: w320 Name: T0-W050-C10-O1K-S10M-baseline Current State: finished

Final Result

General Report

  Op-Type       Op-Count   Byte-Count Avg-ResTime Avg-ProcTime  Throughput   Bandwidth  Succ-Ratio

read           29.52 kops 295.24 GB  337.39 ms   47.28 ms     98.49 op/s   984.94 MB/S 100%
delete         7.1 kops   0 B        37.81 ms    37.81 ms     23.68 op/s   0 B/S       100%
write          18.12 kops 181.25 GB  262.42 ms   153 ms       60.47 op/s   604.66 MB/S 100
```

****This are preliminary numbers with the upcoming 4.1.1 release without any significant tuning effort**

Latency Tests GS4-SSD

Random 4k Read (cache Miss)

```
[root@client01 ~]# /perform/gpfsperf-mpi read rand -r 4k -n 1g -dio  /ibm/fs2-1m-p01/shared/random/test-large-client01-01

/perform/gpfsperf-mpi read rand /ibm/fs2-1m-p01/shared/random/test-large-client01-01

  recSize 4K nBytes 1G fileSize 50G

  nProcesses 1 nThreadsPerProcess 1

  file cache flushed before test

  not using data shipping

  using direct I/O

  offsets accessed will cycle through the same file segment

  not using shared memory buffer

  not releasing byte-range token after open

    Data rate was 6910.39 Kbytes/sec, iops was 1727.60, thread utilization 1.000

    Record size: 4096 bytes, 1073741824 bytes to transfer, 1073741824 bytes transferred

    CPU utilization: user 1.42%, sys 1.04%, idle 97.41%, wait 0.12%
```

1727 IOPS translates to 0.579 ms / request

Seq 4k Read (cache Miss)

```
[root@client01 ~]# /perform/gpfsperf-mpi read seq -r 4k -n 1g -dio  /ibm/fs2-1m-p01/shared/random/test-large-$HOSTNAME-02

/perform/gpfsperf-mpi read seq /ibm/fs2-1m-p01/shared/random/test-large-client01-02

  recSize 4K nBytes 1G fileSize 50G

  nProcesses 1 nThreadsPerProcess 1

  file cache flushed before test

  not using data shipping

  using direct I/O

  offsets accessed will cycle through the same file segment

  not using shared memory buffer

  not releasing byte-range token after open

    Data rate was 22977.24 Kbytes/sec, iops was 5744.31, thread utilization 1.000

    Record size: 4096 bytes, 1073741824 bytes to transfer, 1073741824 bytes transferred

    CPU utilization: user 1.29%, sys 1.43%, idle 97.16%, wait 0.12%
```

5744 IOPS translates to 0.174 ms / request

Seq 4k Read (cache hit)

```
[root@client01 ~]# /perform/gpfsperf-mpi read seq -r 4k -n 1g -dio  /ibm/fs2-1m-p01/shared/random/test-large-$HOSTNAME-02

/perform/gpfsperf-mpi read seq /ibm/fs2-1m-p01/shared/random/test-large-client01-02

  recSize 4K nBytes 1G fileSize 50G

  nProcesses 1 nThreadsPerProcess 1

  file cache flushed before test

  not using data shipping

  using direct I/O

  offsets accessed will cycle through the same file segment

  not using shared memory buffer

  not releasing byte-range token after open

    Data rate was 39279.57 Kbytes/sec, iops was 9819.89, thread utilization 1.000

    Record size: 4096 bytes, 1073741824 bytes to transfer, 1073741824 bytes transferred

    CPU utilization: user 1.65%, sys 1.93%, idle 96.36%, wait 0.06%
```

9819 IOPS translates to 0.101 ms / request

Random 4k Write

```
[root@client01 ~]# /perform/gpfsperf-mpi write rand -r 4k -n 1g -dio  /ibm/fs2-1m-p01/shared/random/test-large-$HOSTNAME-02

/perform/gpfsperf-mpi write rand /ibm/fs2-1m-p01/shared/random/test-large-client01-02

  recSize 4K nBytes 1G fileSize 50G

  nProcesses 1 nThreadsPerProcess 1

  file cache flushed before test

  not using data shipping

  using direct I/O

  offsets accessed will cycle through the same file segment

  not using shared memory buffer

  not releasing byte-range token after open

  no fsync at end of test

    Data rate was 14174.92 Kbytes/sec, iops was 3543.73, thread utilization 1.000

    Record size: 4096 bytes, 1073741824 bytes to transfer, 1073741824 bytes transferred

    CPU utilization: user 1.80%, sys 1.42%, idle 96.67%, wait 0.11%
```

3543 IOPS translates to 0.282 ms / request

Node Sizing

Protocol	Min Memory Recommendation	Min CPU Socket Recommendation
NFS 4000 connections per node <= 128K / cluster (32 NFS nodes * 4000)	base of >= 64GB	1 CPU socket
SMB 3000 connections per node / <= 20K / cluster	x2 memory from a base of 64GB >= 128GB	2 CPU socket
Object 2000 connections per node <= 32K / cluster (16 object nodes * 2000)	x2 memory from a base of 64GB >= 128GB	2 CPU socket
Combination of multiple protocols 2000 Object + 3000 SMB + 4000 NFS per node	Any mix of protocols x2 memory from a base of 64GB >= 128GB	Any mix of protocols 2 CPU socket

ESS is now SAP certified

IBM Elastic Storage Server, International Business Machines Corporation

13 / 22

Contact

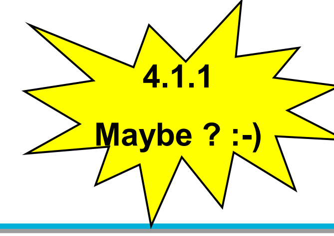
Website	www.ibm.com
Phone	+1 520-574-4600

Details

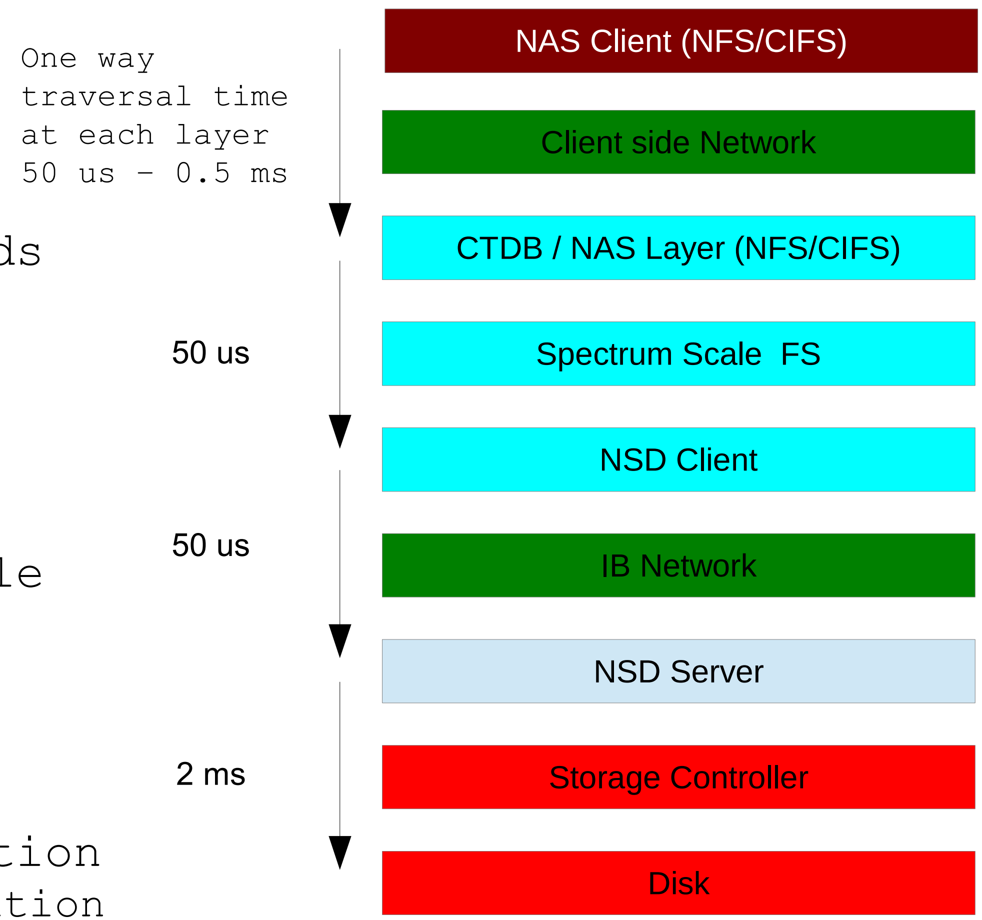
Certification Date	2015-04-01
Certified Until	2018-03-31
Server System	IBM Elastic Storage Server
Model	GS2
Storage Connector	NAS - Distributed file system
Restrictions & Comments	Models included into Storage Family: IBM ESS GS2, GS4, GS6, GL2, GL4, GL6; Scale out up to 8 SAP HANA nodes (1TB) per GS2 (SSD); up to 16 SAP HANA nodes (1TB) per GL6 (HDD); Connection options: 10/40 GBit Ethernet, 56 GBit Infiniband; See SAP Notes 784391 and 1084263 for GPFS support on SAP HANA nodes.

http://global.sap.com/community/ebook/2014-09-02-hana-hardware/enEN/enterprise-storage.html

Pain Point: Small and Synchronous Write Performance



- Common issue in
  - Small and medium-sized workloads
  - EDA workload
  - Virtual Machine Solutions
- Issue across wide range of workloads
  - VMs
  - Databases
  - Windows home directory
  - Logging
  - ISSM (ECM, Websphere, etc)
- Require low-latency and non-volatile memory
  - Flash-backed DIMMs
  - Large batteries
  - Fast SSDs (Fusion-IO, etc)
- Cannot optimize data path in isolation
  - Recovery log updates occur on application writes to sparse files, e.g., VM disk images

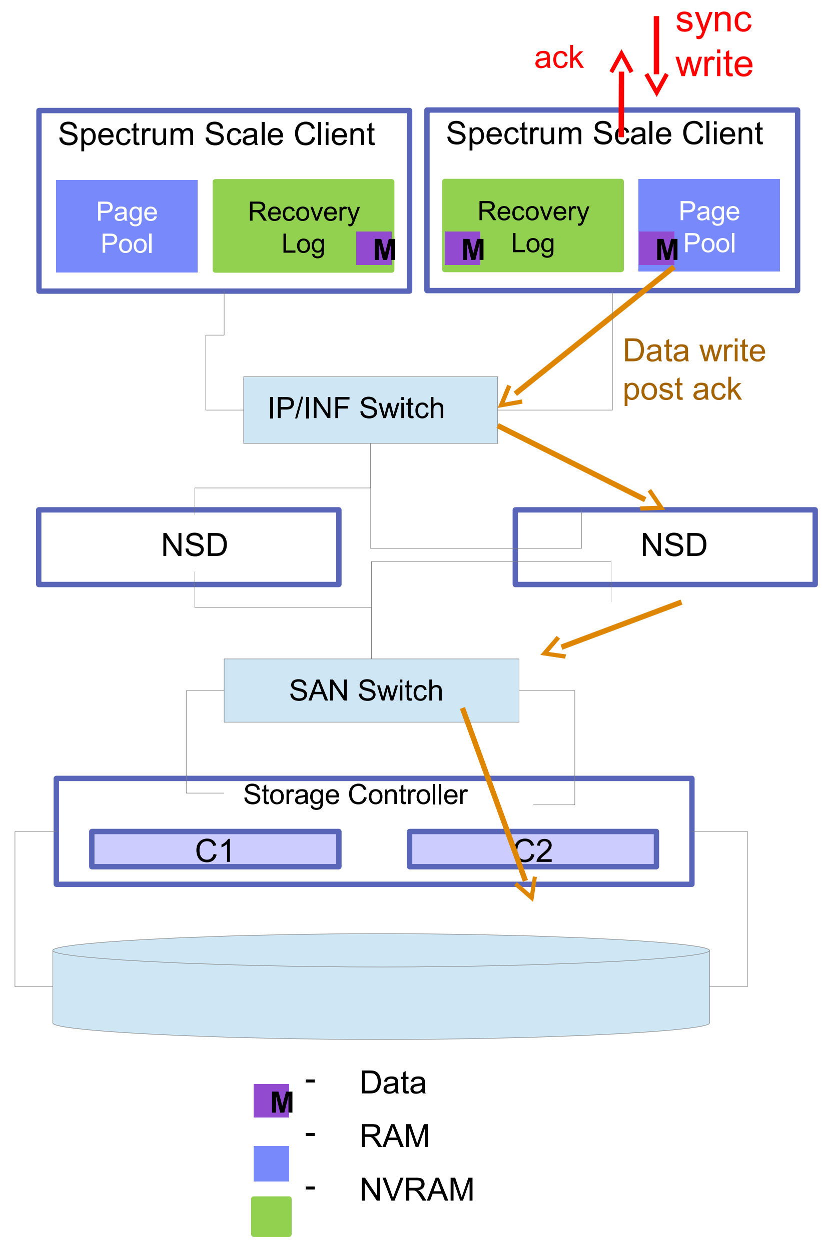


4KB Total Round Trip Time = ~5 ms

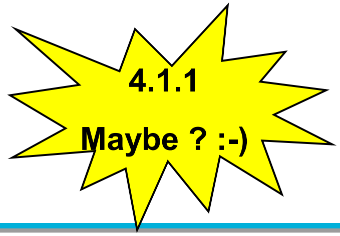
Solution : HAWC – Highly available Write Cache

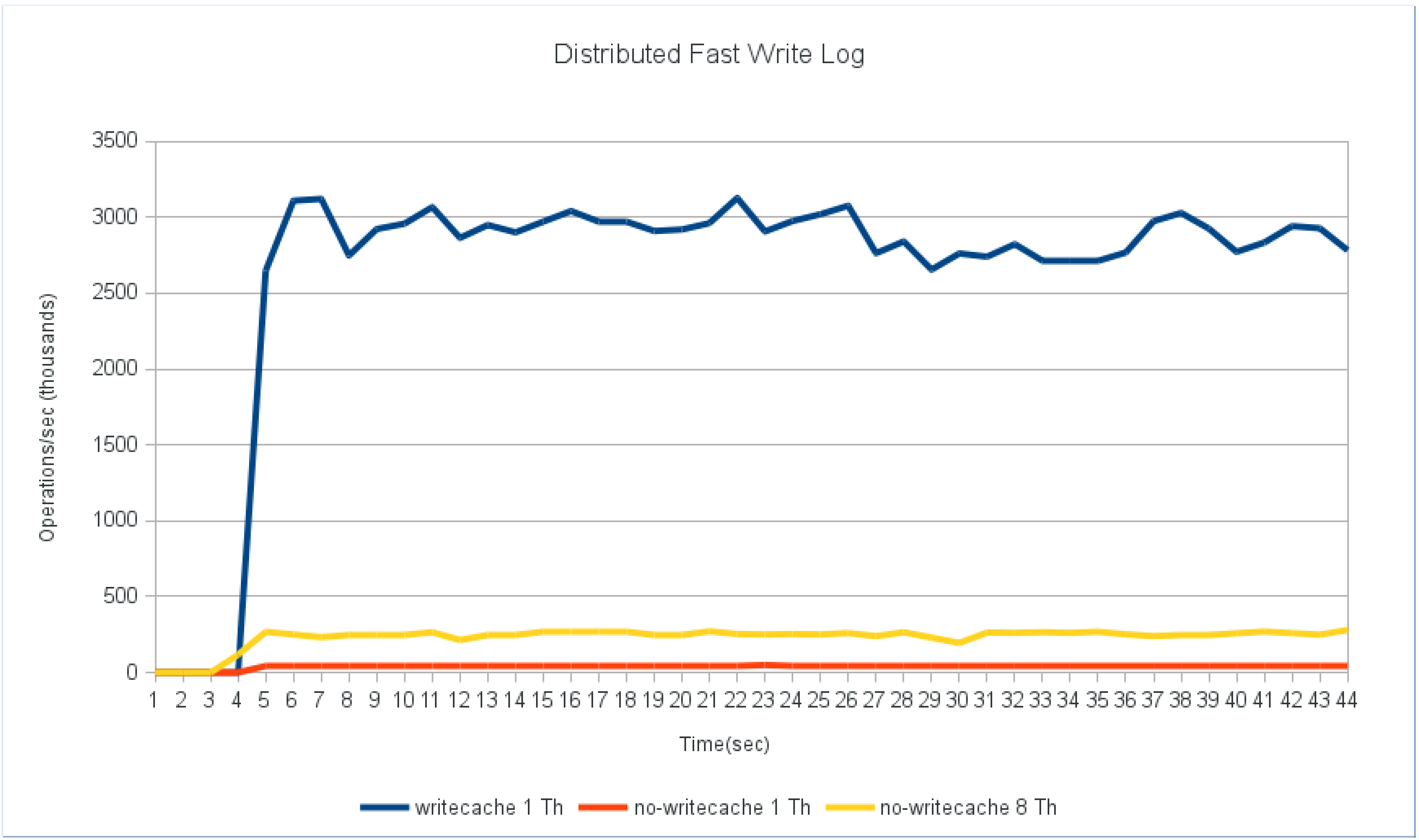
4.1.1
Maybe ? :-)

- HAWC (Log writes)
  - Store recovery log in client NVRAM
  - Either replicate in pairs or store on shared storage
  - Log writes in recovery log
  - Log small writes and send large writes directly to disk
  - Logging data only hardens it
  - Data remains in pagepool and is sent to disk post-logging
    - Leverages write gathering
    - Fast read-cache performance
  - On node failure, run recovery log to place data on disk



HAWC potentials performance compare





Disclaimer

- The information in this document is IBM CONFIDENTIAL.
- This information is provided on an "AS IS" basis without warranty of any kind, express or implied, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. Some jurisdictions do not allow disclaimers of express or implied warranties in certain transactions; therefore, this statement may not apply to you.
- This information is provided for information purposes only as a high level overview of possible future products. PRODUCT SPECIFICATIONS, ANNOUNCE DATES, AND OTHER INOFORMATION CONTAINED HEREIN ARE SUBJECT TO CHANGE AND WITHDRAWAL WITHOUT NOTICE.
- USE OF THIS DOCUMENT IS LIMITED TO SELECT IBM PERSONNEL AND TO BUSINESS PARTNERS WHO HAVE A CURRENT SIGNED NONDISCLUSURE AGREEMENT ON FILE WITH IBM. THIS INFORMAITON CAN ALSO BE SHARED WITH CUSTOMERS WHO HAVE A CURRENT SIGNED NONDISCLOSURE AGREEMENT ON FILE WITH IBM, BUT THIS DOCUMENT SHOULD NOT BE GIVEN TO A CUSTOMER EITHER IN HARDCOPY OR ELECTRONIC FORMAT.

Important notes:

- IBM reserves the right to change product specifications and offerings at any time without notice. This publication could include technical inaccuracies or typographical errors. References herein to IBM products and services do not imply that IBM intends to make them available in all countries.
- IBM makes no warranties, express or implied, regarding non-IBM products and services, including but not limited to Year 2000 readiness and any implied warranties of merchantability and fitness for a particular purpose. IBM makes no representations or warranties with respect to non-IBM products. Warranty, service and support for non-IBM products is provided directly to you by the third party, not IBM.
- All part numbers referenced in this publication are product part numbers and not service part numbers. Other part numbers in addition to those listed in this document may be required to support a specific device or function.
- MHz / GHz only measures microprocessor internal clock speed; many factors may affect application performance. When referring to storage capacity, GB stands for one billion bytes; accessible capacity may be less. Maximum internal hard disk drive capacities assume the replacement of any standard hard disk drives and the population of all hard disk drive bays with the largest currently supported drives available from IBM.

IBM Information and Trademarks

- The following terms are trademarks or registered trademarks of the IBM Corporation in the United States or other countries or both: the e-business logo, IBM, xSeries, pSeries, zSeries, iSeries.
- Intel, Pentium 4 and Xeon are trademarks or registered trademarks of Intel Corporation. Microsoft Windows is a trademark or registered trademark of Microsoft Corporation. Linux is a registered trademark of Linus Torvalds. Other company, product, and service names may be trademarks or service marks of others.