outthink limits

News from Research

Sven Oehme - oehmes@us.ibm.com

Disclaimer

This is a Research Presentation and doesn’t guarantee any of the demonstrated capabilities, functions or features end up in a GA product

Agenda

- **Performance engineering matters**
  - **Disk drive engineering – how fast is a drive really ?**
  - **How to get data from storage to consumer – Network overhaul**
  - **More than 32 Sub blocks, why and what can we expect from them**
- Spectrum Scale with NVMe
- Spectrum Scale with NVMeoF

Performance engineering matters

Imagine you need to deliver the following goals :

- 2.5 TB/sec single stream IOR as requested from ORNL
- 1 TB/sec 1MB sequential read/write as stated in CORAL RFP
- Single Node 16 GB/sec sequential read/write as requested from ORNL
- 50k creates/sec per shared directory as stated in CORAL RFP
- 2.6 Million 32k file creates/sec as requested from ORNL

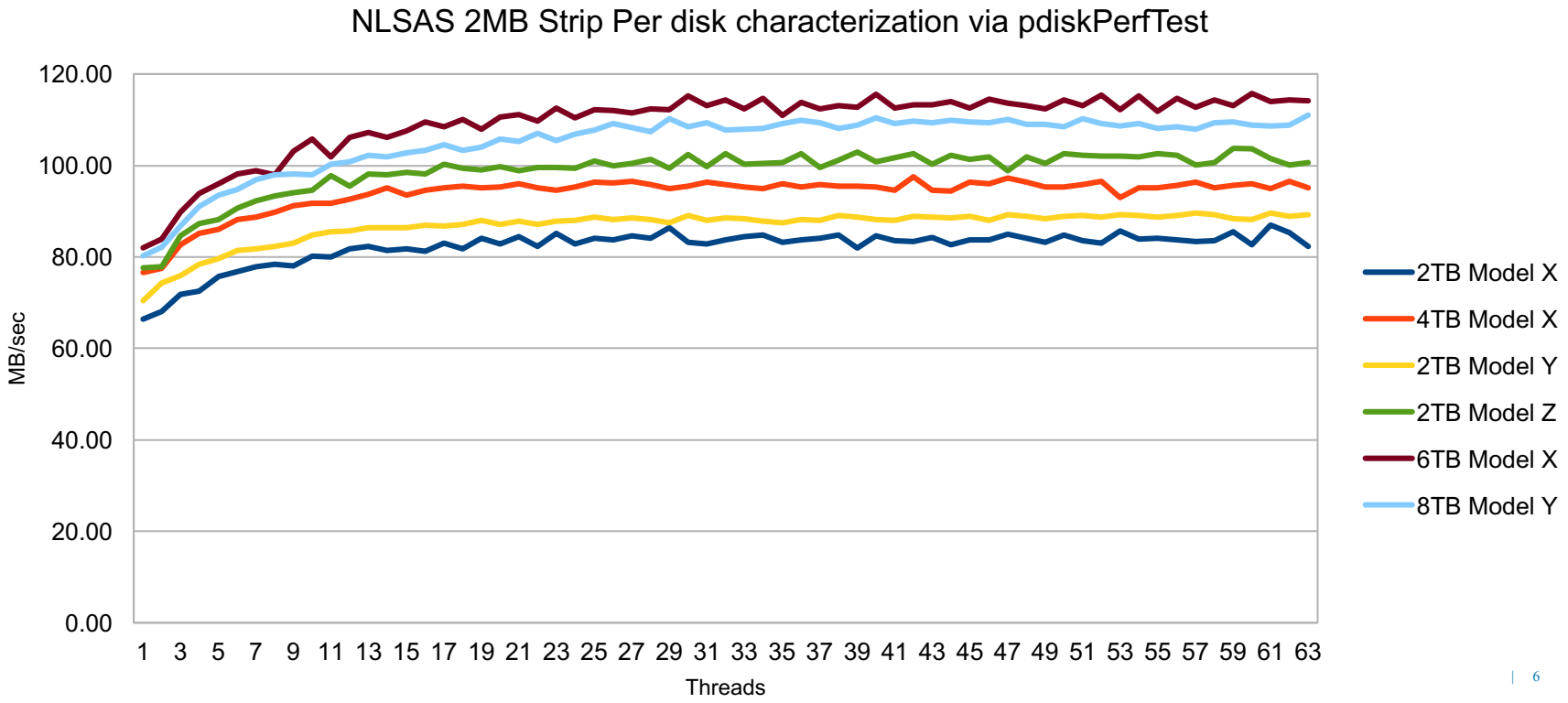
What innovations in Storage would that require ?

So what Influences total system speed

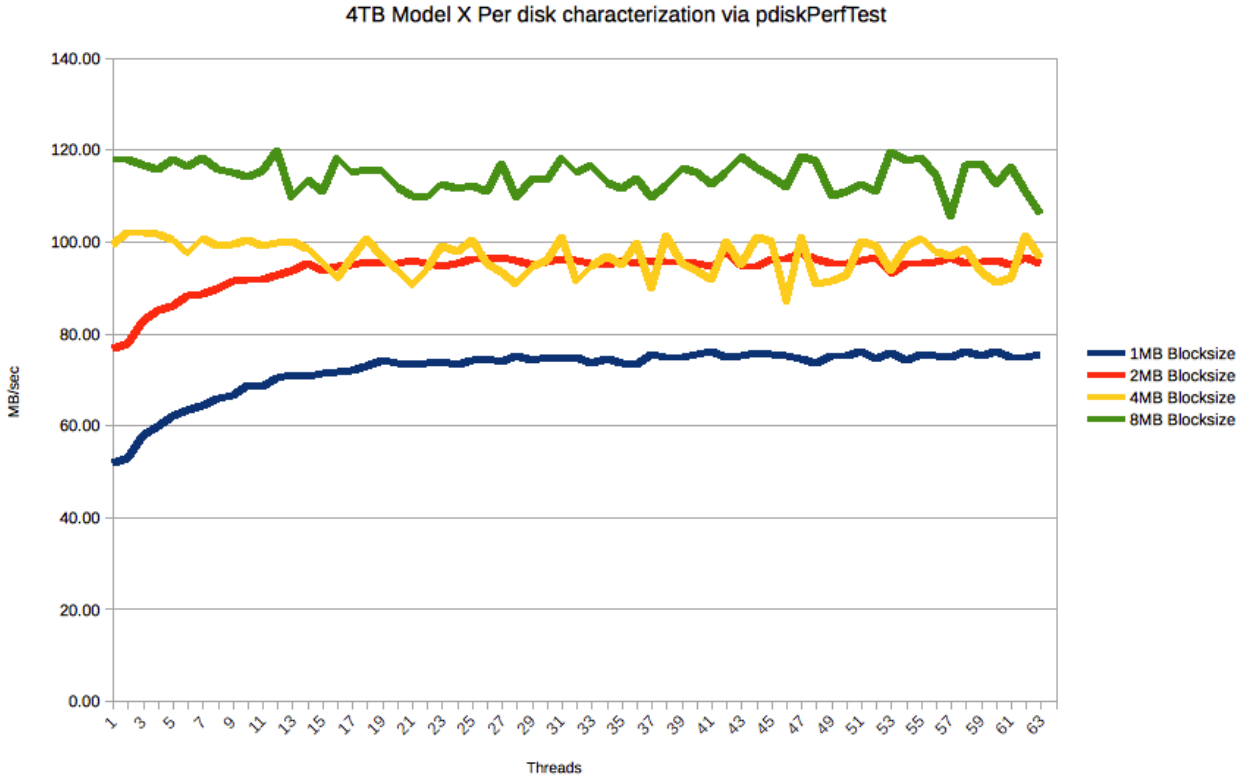
- Filesystem Overhead - who talks to block storage these days ?
- Controller Overhead - SW vs HW and how good is your raid implementation ?
- Raid mode Overhead - that's a simple math problem 1P vs 2P vs 3P ..
- Cache efficiency - complex , main issue is what context is that i/o performed
- Application access Patterns - random vs sequential
- Access Pattern the disks sees - you think its sequential, you are most likely wrong

So how fast perform disks under a large scale filesystem and what can one expect ?

First - all access is random , lets take a look on how different models perform



Closer look at the 4TB Model with different i/o sizes



You need to overhaul the Network communication – done in 4.2.1+

- Why do we need it ?
  - Keep up with the io(not capacity) density of bleeding edge Storage technology (NVMe, etc.)
  - Leverage advances in latest Network Technology (100GE/IB)
  - Single Node NSD Server ‘Scale-up’ limitation
  - NUMA is the norm in modern systems, no longer the exception

- What do we need to do ?
  - Implement an (almost) lock free communication code in all performance critical code path
  - Make communication code as well as other critical areas of the code NUMA aware
  - Add ‘always on’ instrumentation for performance critical data, don’t try to add it later or design for ‘occasional’ collection when needed

That’s what we did in 4.2.1 but there is more to come

Better Network diag tools

```
   fire11                          <c0n29>  192.1.77.11     connected  0    89    0         0          Linux/L
   fire12                          <c0n30>  192.1.77.12     connected  0    79    0         0          Linux/L
Connection details:
  <c0p36> 192.1.20.30/0, 192.167.20.130 (p8n20hyp)
    connection info:
      retry(success): 0(0)
  <c0n0> 192.1.13.5/0, 192.167.13.5 (client05)
    connection info:
      retry(success): 0(0)
      tcp connection state: established     tcp congestion state: open
    packet statistics:
      lost: 0     unacknowledged: 0
      retrans: 0     unrecovered retrans: 0
    network speed(μs):
      rtt(round trip time): 53     medium deviation of rtt: 3
    pending data statistics(byte):
      read/write calls pending: 0
      GPFS Send-Queue: 0     GPFS Recv-Queue: 0
      Socket Send-Queue: 0     Socket Recv-Queue: 0
  <c0n1> 192.1.13.6/0, 192.167.13.6 (client06)
    connection info:
      retry(success): 0(0)
      tcp connection state: established     tcp congestion state: open
    packet statistics:
      lost: 0     unacknowledged: 0
      retrans: 0     unrecovered retrans: 0
    network speed(μs):
      rtt(round trip time): 48     medium deviation of rtt: 4
    pending data statistics(byte):
      read/write calls pending: 0
      GPFS Send-Queue: 0     GPFS Recv-Queue: 0
      Socket Send-Queue: 0     Socket Recv-Queue: 0
```

Real time performance analytics - iocounters

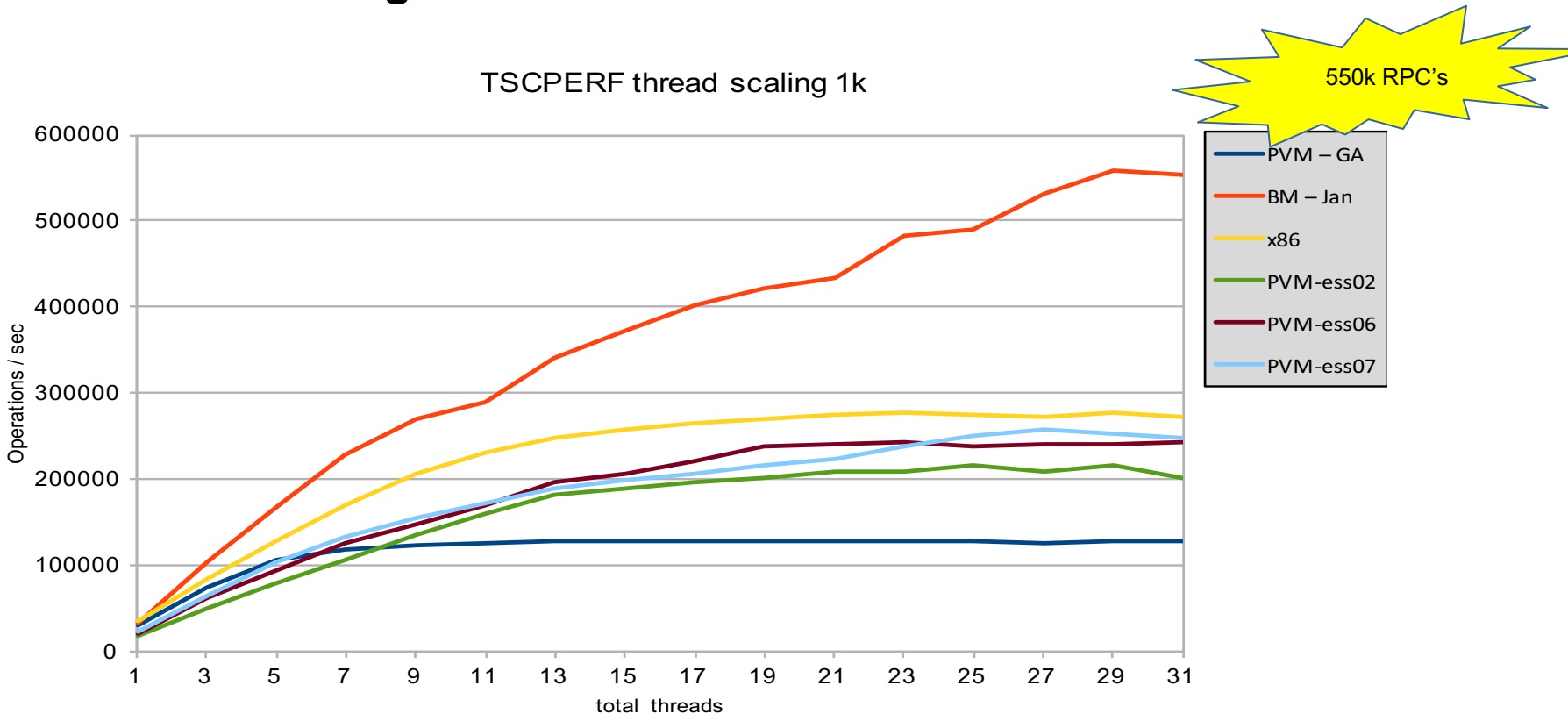
```
iocounters duration 10.108565420 seconds (26282162342 cycles) cycles/seconds 2599989340
reset at Wed May 10 05:37:36 2017
dump  at Wed May 10 05:37:47 2017

                     Call Count Aggregate Seconds   Average Seconds   Maximum Seconds      nvcsw     nivcsw  ID kx call
-------------------- ---------- ------------------ ----------------- ----------------- ---------- ---------- --- ------------------------------
          kx calls:        187       0.000446622       0.000002388       0.000009864          0          0   4 SignalOrBroadcastCondvar
          kx calls:        187      89.085940007       0.476395401      10.090561646        188          0   6 WaitCondvar
          kx calls:        175       0.000379997       0.000002171       0.000007026          0          0  17 ReleaseMutex
          kx calls:          1       0.000000540       0.000000540       0.000000540          0          0  52 ClearPageoutThread
          kx calls:        518       0.000030602       0.000000059       0.000004806          0          0  77 CheckThreadStateCtl
          kx calls:          1       0.000002061       0.000002061       0.000002061          0          0  81 SetDosAttr
          kx calls:          1       0.000003365       0.000003365       0.000003365          0          0 106 AttachSharedMemory
          kx calls:          1       0.000002480       0.000002480       0.000002480          0          0 109 startDMS
          kx calls:         13       0.002199823       0.000169217       0.000691865          0          0 162 GetCounters
          kx calls:         19     255.828182474      13.464641183      40.014821934         19          0 163 WaitFastCondvarForSignal
          kx calls:         13       0.000030259       0.000002328       0.000003059          0          0 164 SignalFastCondvarForSignal
          kx calls:          6       0.000019418       0.000003236       0.000003961          0          0 165 SignalFastCondvarForSignalList

                     Call Count Aggregate Seconds   Average Seconds   Maximum Seconds      nvcsw     nivcsw  ID MutexName
-------------------- ---------- ------------------ ----------------- ----------------- ---------- ---------- --- ------------------------------
    dReleaseMutex:          1       0.000002493       0.000002493       0.000002493          0          0   1 ThreadSuspendResumeMutex
    dReleaseMutex:        132       0.000211378       0.000001601       0.000003227          0          0 120 EEInstanceMutex
    dReleaseMutex:         24       0.000056643       0.000002360       0.000006610          0          0 188 SdrServQueueMutex
    dReleaseMutex:          1       0.000001381       0.000001381       0.000001381          0          0 249 SyncPairMutex
    dReleaseMutex:         12       0.000026644       0.000002220       0.000005623          0          0 443 SyncListMutex
    dReleaseMutex:          3       0.000005133       0.000001711       0.000002072          0          0 446 CCRIOThreadMutex
    dReleaseMutex:          2       0.000003310       0.000001655       0.000001733          0          0 447 CCRIORequestMutex

                     Call Count Aggregate Seconds   Average Seconds   Maximum Seconds      nvcsw     nivcsw  ID CondvarName
-------------------- ---------- ------------------ ----------------- ----------------- ---------- ---------- --- ------------------------------
  dWaitCondvarFast:         13     255.828076094      19.679082776      40.014821686         13          0  18 RcvWorkerCondvar
  dWaitCondvarFast:          6       0.000099896       0.000016649       0.000020294          6          0 350 RdmaSend_Reply
```

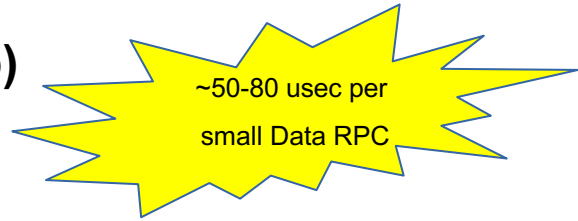
4.2.1 Network Scaling results 1k RPC's between 1 Server and 1 Client



Single client throughput enhancements


```
[root@p8n06 ~]# tsqosperf write seq -n 200g -r 16m -th 16 /ibm/fs2-16m-06/shared/testfile -fsync
tsqosperf write seq /ibm/fs2-16m-06/shared/testfile
  recSize 16M nBytes 200G fileSize 200G
  nProcesses 1 nThreadsPerProcess 16
  file cache flushed before test
  not 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
  fsync at end of test
    Data rate was 16124635.71 Kbytes/sec, thread utilization 0.938, bytesTransferred 214748364800
```

Single thread small i/o (client – server – device roundtrip)




```
[root@client01 ~]#  tsqosperf read seq -r 4k /ibm/fs2-256k-08/shared/test –dio
tsqosperf read seq /ibm/fs2-256k-08/shared/test
  recSize 4K nBytes 128M fileSize 128M
  nProcesses 1 nThreadsPerProcess 1
  file cache flushed before test
  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 55111.52 Kbytes/sec, Op Rate was 13454.96 Ops/sec, Avg Latency was 0.074 milliseconds, thread utilization 1.000, bytesTransferred
134217728

 [root@client01 mpi]# mmfsadm dump iohist |less

I/O history:

 I/O start time RW    Buf type disk:sectorNum     nSec  time ms      tag1      tag2           Disk UID typ      NSD node context   thread
--------------- -- ----------- ----------------- -----  ------- --------- --------- ------------------ --- --------------- --------- ----------
11:37:54.451846  R        data    4:192933224         8    0.055    284160       504  C0A74D01:58BD6495 cli  192.167.20.127 MBHandler DioHandlerThread
11:37:54.451918  R        data    4:192933232         8    0.055    284160       504  C0A74D01:58BD6495 cli  192.167.20.127 MBHandler DioHandlerThread
11:37:54.451990  R        data    4:192933240         8    0.054    284160       504  C0A74D01:58BD6495 cli  192.167.20.127 MBHandler DioHandlerThread
11:37:54.452061  R        data    4:192933248         8    0.054    284160       504  C0A74D01:58BD6495 cli  192.167.20.127 MBHandler DioHandlerThread
11:37:54.452132  R        data    4:192933256         8    0.055    284160       504  C0A74D01:58BD6495 cli  192.167.20.127 MBHandler DioHandlerThread
11:37:54.452205  R        data    4:192933264         8    0.053    284160       504  C0A74D01:58BD6495 cli  192.167.20.127 MBHandler DioHandlerThread
11:37:54.452275  R        data    4:192933272         8    0.057    284160       504  C0A74D01:58BD6495 cli  192.167.20.127 MBHandler DioHandlerThread
11:37:54.452349  R        data    4:192933280         8    0.056    284160       504  C0A74D01:58BD6495 cli  192.167.20.127 MBHandler DioHandlerThread
```

Shared directory file create – 50k target

```
-- started at 02/28/2017 12:13:13 --

mdtest-1.9.3 was launched with 14 total task(s) on 14 node(s)
Command line used: /ghome/oehmes/mpi/bin/mdtest-pcmpi9131-existingdir -d /gpfs/fs2-1m-
me1/shared/mdtest-ec -i 1 -n 35000 -F -w 0 -Z -p 8
Path: /gpfs/fs2-1m-me1/shared
FS: 17.1 TiB   Used FS: 0.1%   Inodes: 476.8 Mi   Used Inodes: 0.1%

14 tasks, 490000 files

SUMMARY: (of 1 iterations)
   Operation                     Max            Min           Mean        Std Dev
   ---------                     ---            ---           ----        -------
   File creation     :      50032.690      50032.690      50032.690          0.000
   File stat         :    3937604.341    3937604.341    3937604.341          0.000
   File read         :     941193.073     941193.073     941193.073          0.000
   File removal      :     143095.519     143095.519     143095.519          0.000
   Tree creation     :      77672.296      77672.296      77672.296          0.000
   Tree removal      :          0.239          0.239          0.239          0.000

-- finished at 02/28/2017 12:13:39 --
```

More than 32 Sub blocks - why and what to expect ?

- Why do we have Sub blocks ?
  - Allow finer grained allocation – no space wasted
  - Allows coalescing of small files in larger blocks – raid friendly
- What Options do we have today ?
  - We can store data in inode (default <4k)
  - We can allocate a Sub block (1/32th of a Full block)
  - We support 64 KB, 128 KB, 256 KB, 512 KB, 1 MB, 2 MB, 4 MB, 8 MB and 16 MB block size today
- What's wrong with it ?
  - You have to choose between waste space for small files (>4k and <1/32th of block size) or bandwidth
  - You can never ever change it online, filesystem migration required
  - It has a significant performance penalty for small files in large block size filesystems
- So how do we fix it and what will it change ?

Best way to find out – measure it – 16 MB block size filesystem - mdtest

4.2.1 base code - SUMMARY: (of 3 iterations)

Operation		Max	Min	Mean	Std Dev
File creation	:	2296.791	2197.553	2237.644	42.695
File stat	:	3402913.848	3383139.838	3390622.546	8759.559
File read	:	452144.282	383467.565	426670.673	30712.367
File removal	:	202219.699	88486.720	160499.542	51134.019
Tree creation	:	9425.078	3138.312	6945.652	2732.932
Tree removal	:	6710.394	3063.299	5196.237	1551.879

zero-end-of-file-padding (4.2.2 + ifdef for zero padding): SUMMARY: (of 3 iterations)

Operation		Max	Min	Mean	Std Dev
File creation	:	13053.701	12570.060	12866.842	212.194
File stat	:	4077992.847	3291830.765	3600173.039	342592.742
File read	:	450592.091	408552.363	424759.494	18462.970
File removal	:	105876.511	93884.369	99224.908	4982.772
Tree creation	:	8451.948	1936.832	4123.063	3061.035
Tree removal	:	535.050	154.181	363.642	157.800

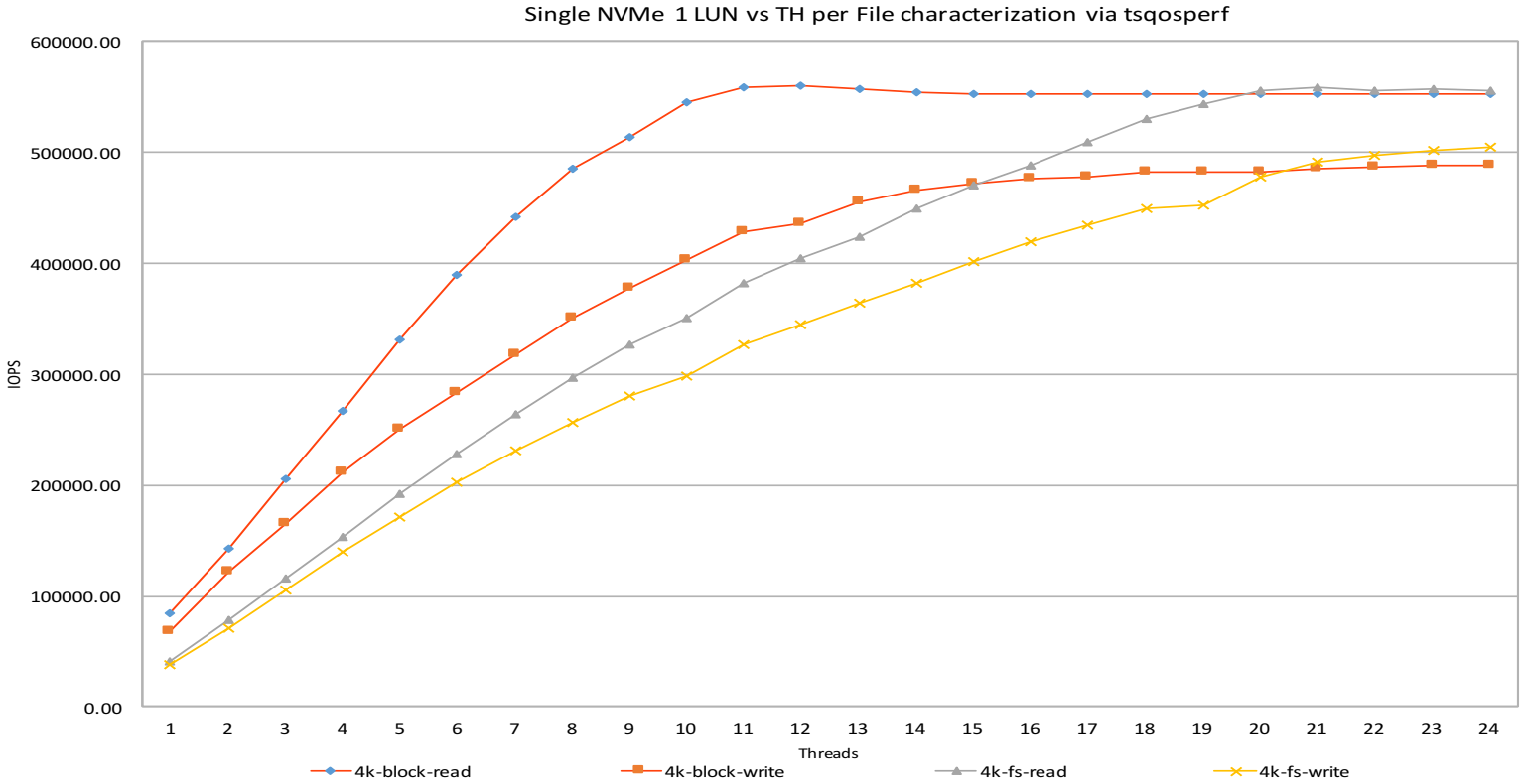
more sub blocks per block (4.2.2 + morethan32subblock code):
SUMMARY: (of 3 iterations)

Operation		Max	Min	Mean	Std Dev
File creation	:	51397.549	33005.542	40316.721	7967.608
File stat	:	3326016.821	3195765.701	3277674.290	58231.427
File read	:	616434.716	543430.803	568013.424	34240.371
File removal	:	134732.546	48867.351	86175.005	35945.588
Tree creation	:	7771.893	1039.578	3648.852	2949.535
Tree removal	:	2879.694	550.493	1859.348	972.530

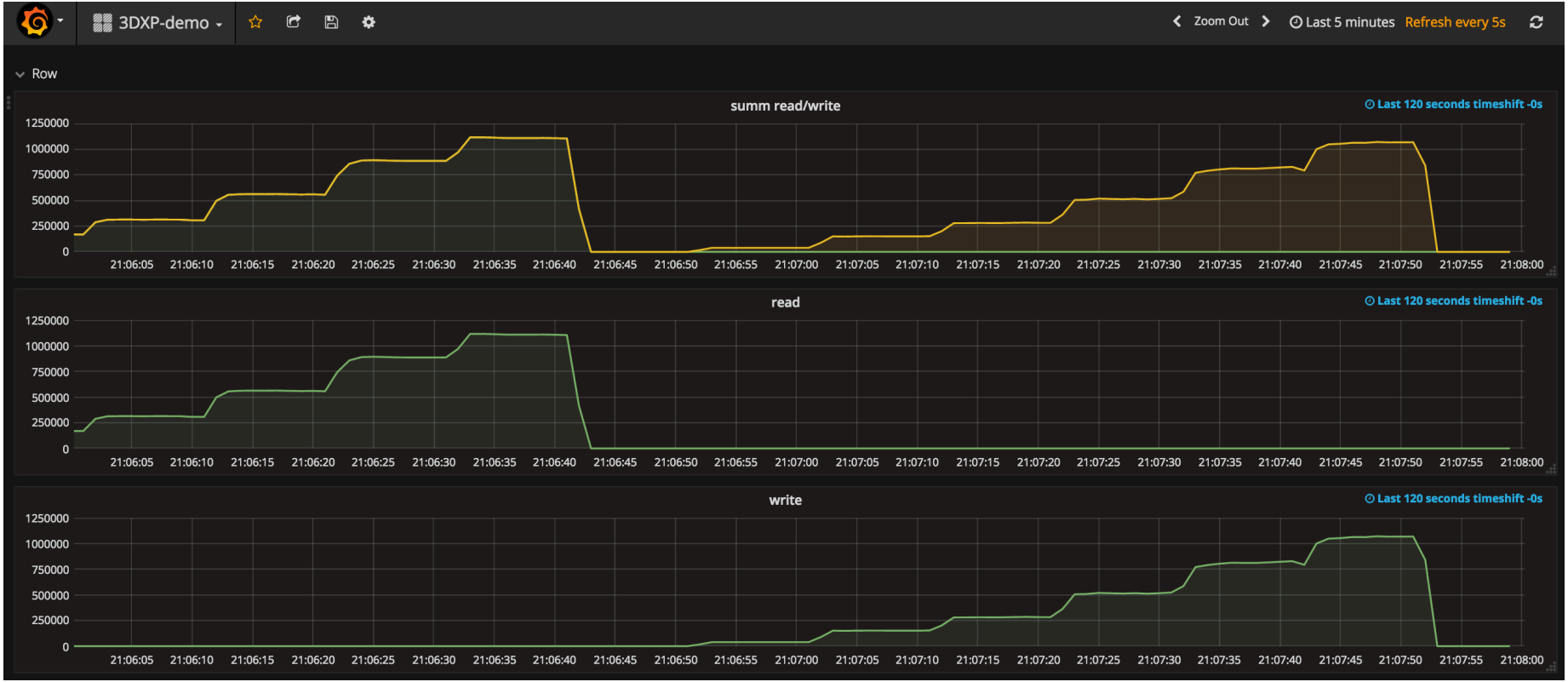
Agenda

- Performance engineering matters
  - Disk drive engineering – how fast is a drive really ?
  - How to get data from storage to consumer – Network overhaul
  - More than 32 Sub blocks, why and what can we expect from them
- **Spectrum Scale with NVMe**
- Spectrum Scale with NVMeoF

NVMe , Block , local Filesystem access – Single node – single device



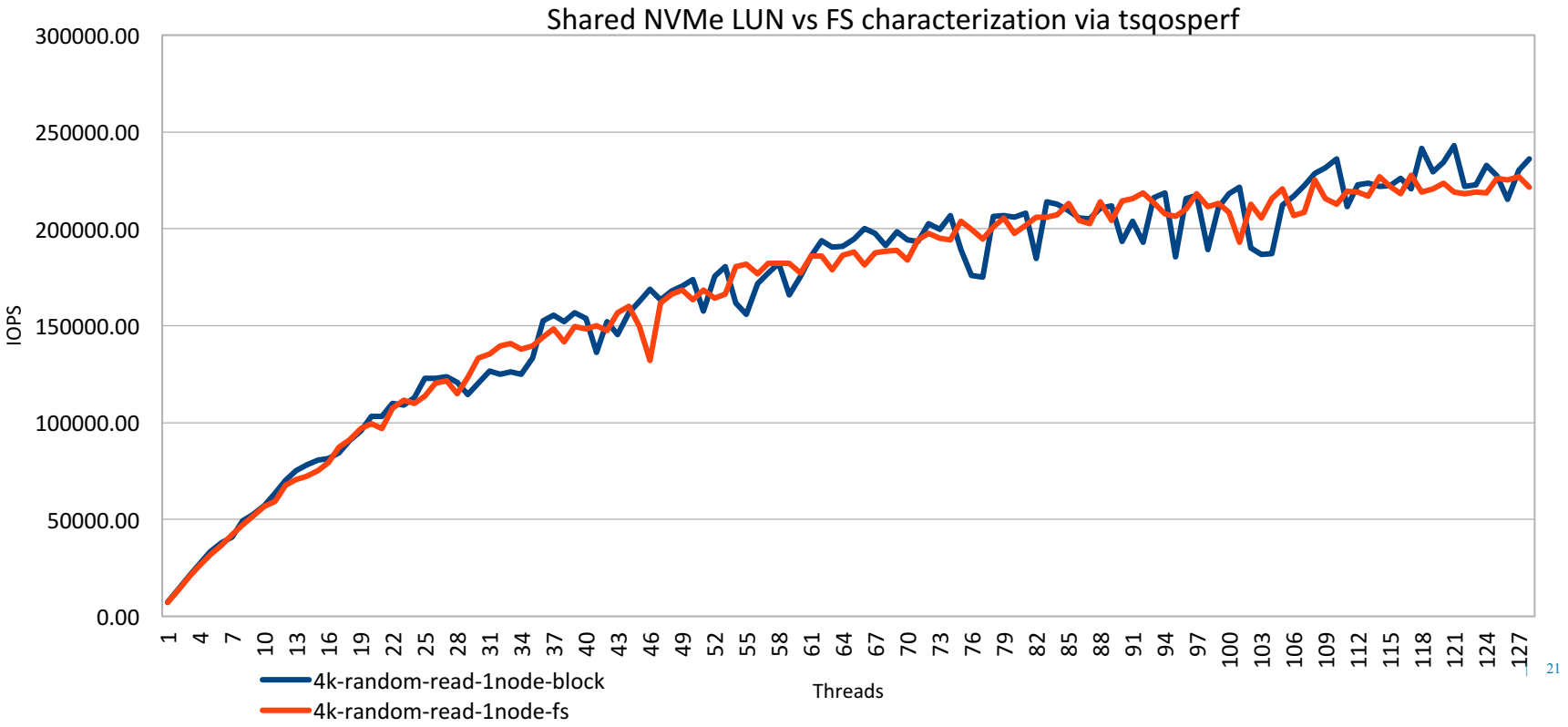
Single Node 1 – 64 Thread tests – 1.2 Milion @ 50 usec Response time



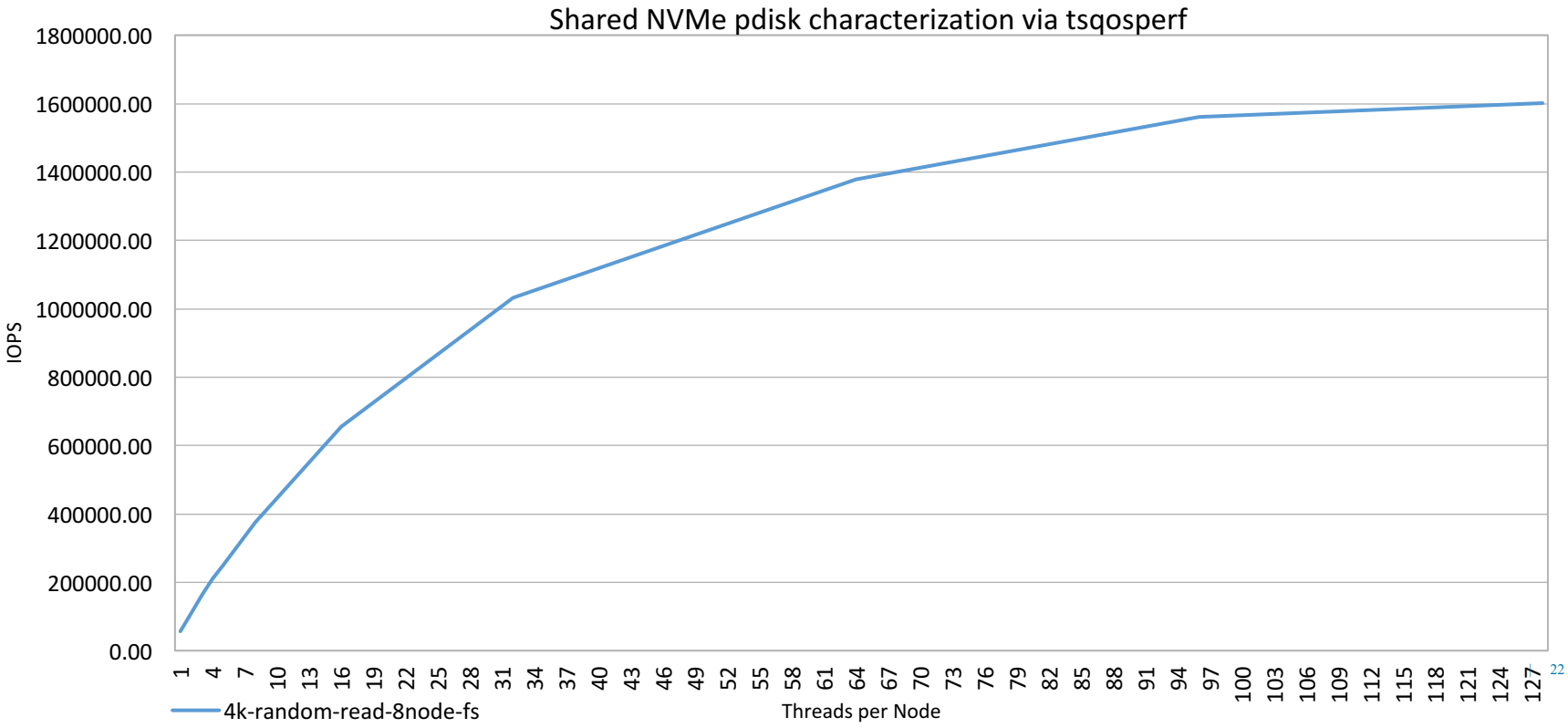
Agenda

- Performance engineering matters
  - Disk drive engineering – how fast is a drive really ?
  - How to get data from storage to consumer – Network overhaul
  - More than 32 Sub blocks, why and what can we expect from them
- Spectrum Scale with NVMe
- **Spectrum Scale with NVMeoF**

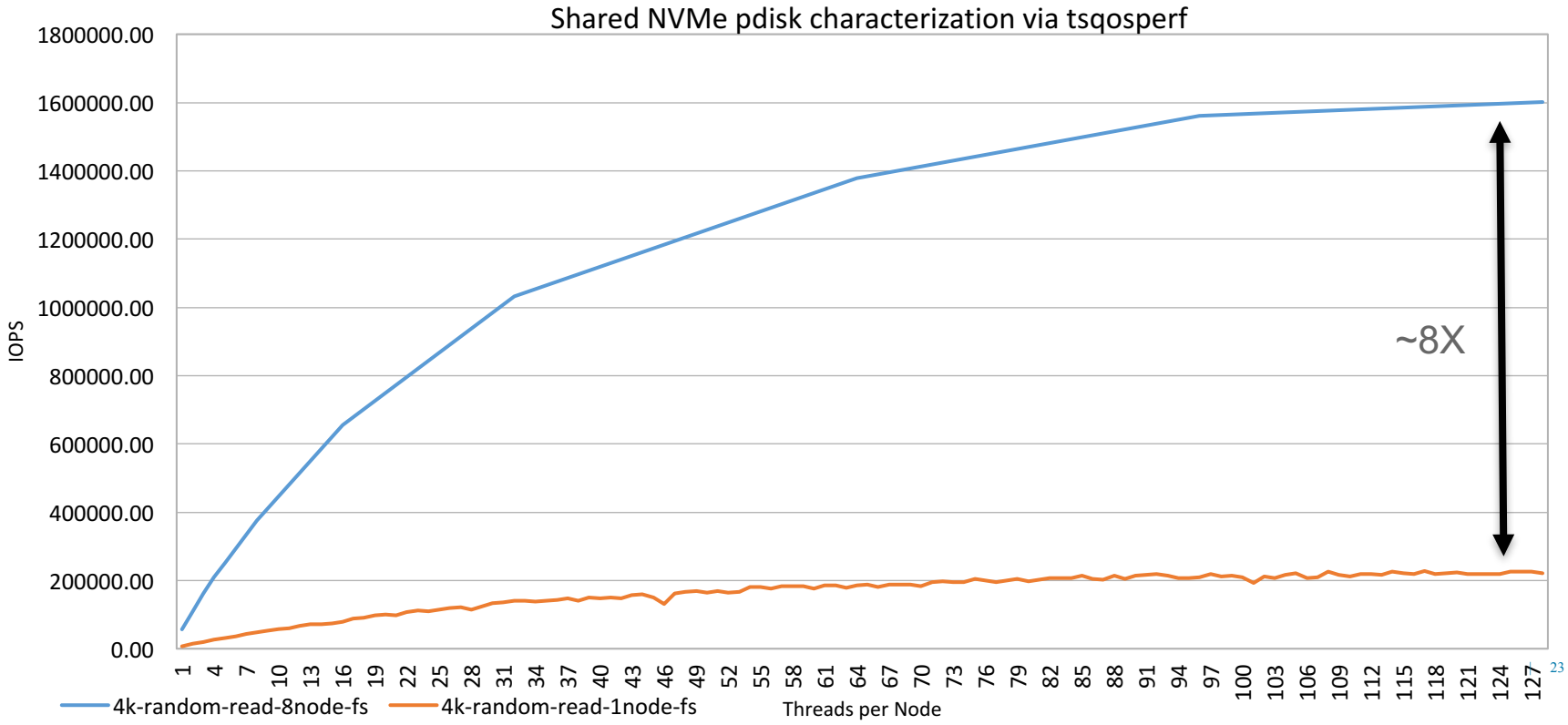
NVMeoF device – single Node – block vs filesystem – 4k RR DIO



NVMeoF device – 8 Nodes – just filesystem access – 4k RR DIO



NVMeoF device – 1 vs 8 Nodes – just filesystem access – 4k RR DIO



Backup

Some nice LROC stats

```
node06.gpfs:  Total objects stored 60014924 (5659216 MB) recalled 167229775 (34068168 MB)
node07.gpfs:        Data objects stored 26751870 (6338382 MB) recalled 132025766 (32850787 MB) = 493.52 %
node07.gpfs:        Directory objects stored 13237336 (413632 MB) recalled 3940102 (123098 MB) = 29.77 %
node07.gpfs:        Inode objects stored 38047588 (148623 MB) recalled 38059597 (148668 MB) = 100.03 %
node07.gpfs:  Total objects stored 78044689 (6900725 MB) recalled 174027235 (33122854 MB)
node08.gpfs:        Data objects stored 12310895 (2907511 MB) recalled 95516173 (23974064 MB) = 775.87 %
node08.gpfs:        Directory objects stored 8248666 (257766 MB) recalled 1812308 (56636 MB) = 21.97 %
node08.gpfs:        Inode objects stored 28387781 (110890 MB) recalled 28400811 (110939 MB) = 100.05 %
node08.gpfs:  Total objects stored 48953206 (3276199 MB) recalled 125730596 (24141710 MB)
node09.gpfs:        Data objects stored 21393146 (5431197 MB) recalled 142377365 (35532150 MB) = 665.53 %
node09.gpfs:        Directory objects stored 10078094 (314914 MB) recalled 2521433 (78770 MB) = 25.02 %
node09.gpfs:        Inode objects stored 22836203 (89204 MB) recalled 22849021 (89252 MB) = 100.06 %
node09.gpfs:  Total objects stored 54315320 (5835390 MB) recalled 167748551 (35700309 MB)
node10.gpfs:        Data objects stored 19392503 (4905809 MB) recalled 138941859 (34607929 MB) = 716.47 %
node10.gpfs:        Directory objects stored 9482116 (296286 MB) recalled 2409845 (75282 MB) = 25.41 %
node10.gpfs:        Inode objects stored 19553122 (76379 MB) recalled 19557308 (76395 MB) = 100.02 %
node10.gpfs:  Total objects stored 48431155 (5278549 MB) recalled 160908778 (34759543 MB)
node11.gpfs:        Data objects stored 19757211 (5537707 MB) recalled 134703417 (33621805 MB) = 681.79 %
node11.gpfs:        Directory objects stored 6776553 (211739 MB) recalled 1695805 (52962 MB) = 25.02 %
node11.gpfs:        Inode objects stored 14563441 (56888 MB) recalled 14563796 (56889 MB) = 100.00 %
node11.gpfs:  Total objects stored 41098960 (5806395 MB) recalled 150964218 (33731910 MB)
node12.gpfs:        Data objects stored 13659294 (3400875 MB) recalled 112260794 (28042845 MB) = 821.86 %
node12.gpfs:        Directory objects stored 9137807 (285536 MB) recalled 2253142 (70399 MB) = 24.66 %
node12.gpfs:        Inode objects stored 19255316 (75216 MB) recalled 19257266 (75222 MB) = 100.01 %
```

Legal notices

Copyright © 2017 by International Business Machines Corporation. All rights reserved.

No part of this document may be reproduced or transmitted in any form without written permission from IBM Corporation.

Product data has been reviewed for accuracy as of the date of initial publication. Product data is subject to change without notice. This document could include technical inaccuracies or typographical errors. IBM may make improvements and/or changes in the product(s) and/or program(s) described herein at any time without notice. Any statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business. Any reference to an IBM Program Product in this document is not intended to state or imply that only that program product may be used. Any functionally equivalent program, that does not infringe IBM's intellectually property rights, may be used instead.

THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER OR IMPLIED. IBM LY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NONINFRINGEMENT. IBM shall have no responsibility to update this information. IBM products are warranted, if at all, according to the terms and conditions of the agreements (e.g., IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided. Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products in connection with this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. IBM makes no representations or warranties, ed or implied, regarding non-IBM products and services.

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents or copyrights. Inquiries regarding patent or copyright licenses should be made, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 1 0504- 785
U.S.A.

Information and trademarks

IBM, the IBM logo, ibm.com, IBM System Storage, IBM Spectrum Storage, IBM Spectrum Control, IBM Spectrum Protect, IBM Spectrum Archive, IBM Spectrum Virtualize, IBM Spectrum Scale, IBM Spectrum Accelerate, Softlayer, and XIV are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at http://www.ibm.com/legal/copytrade.shtml

The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.

IT Infrastructure Library is a Registered Trade Mark of AXELOS Limited.

Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.

Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.

ITIL is a Registered Trade Mark of AXELOS Limited.

UNIX is a registered trademark of The Open Group in the United States and other countries.

* All other products may be trademarks or registered trademarks of their respective companies.

Notes:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

This presentation and the claims outlined in it were reviewed for compliance with US law. Adaptations of these claims for use in other geographies must be reviewed by the local country counsel for compliance with local laws.

Special notices

This document was developed for IBM offerings in the United States as of the date of publication. IBM may not make these offerings available in other countries, and the information is subject to change without notice. Consult your local IBM business contact for information on the IBM offerings available in your area.

Information in this document concerning non-IBM products was obtained from the suppliers of these products or other public sources. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. Send license inquires, in writing, to IBM Director of Licensing, IBM Corporation, New Castle Drive, Armonk, NY 10504-1785 USA.

All statements regarding IBM future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

The information contained in this document has not been submitted to any formal IBM test and is provided "AS IS" with no warranties or guarantees either expressed or implied.

All examples cited or described in this document are presented as illustrations of the manner in which some IBM products can be used and the results that may be achieved. Actual environmental costs and performance characteristics will vary depending on individual client configurations and conditions.

IBM Global Financing offerings are provided through IBM Credit Corporation in the United States and other IBM subsidiaries and divisions worldwide to qualified commercial and government clients. Rates are based on a client's credit rating, financing terms, offering type, equipment type and options, and may vary by country. Other restrictions may apply. Rates and offerings are subject to change, extension or withdrawal without notice.

IBM is not responsible for printing errors in this document that result in pricing or information inaccuracies.

All prices shown are IBM's United States suggested list prices and are subject to change without notice; reseller prices may vary.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

Any performance data contained in this document was determined in a controlled environment. Actual results may vary significantly and are dependent on many factors including system hardware configuration and software design and configuration. Some measurements quoted in this document may have been made on development-level systems. There is no guarantee these measurements will be the same on generally-available systems. Some measurements quoted in this document may have been estimated through extrapolation. Users of this document should verify the applicable data for their specific environment.