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
- IOPS - does it actually mean anything ??

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 ?

Lets start with how fast is a disk drive ?

If you ask the Disk Vendor – 150 MB/sec per drive

If you ask you Block Storage Seller – 100 MB/sec per drive

If you ask an application Person – always to slow

If you ask a HPC admin – it depends

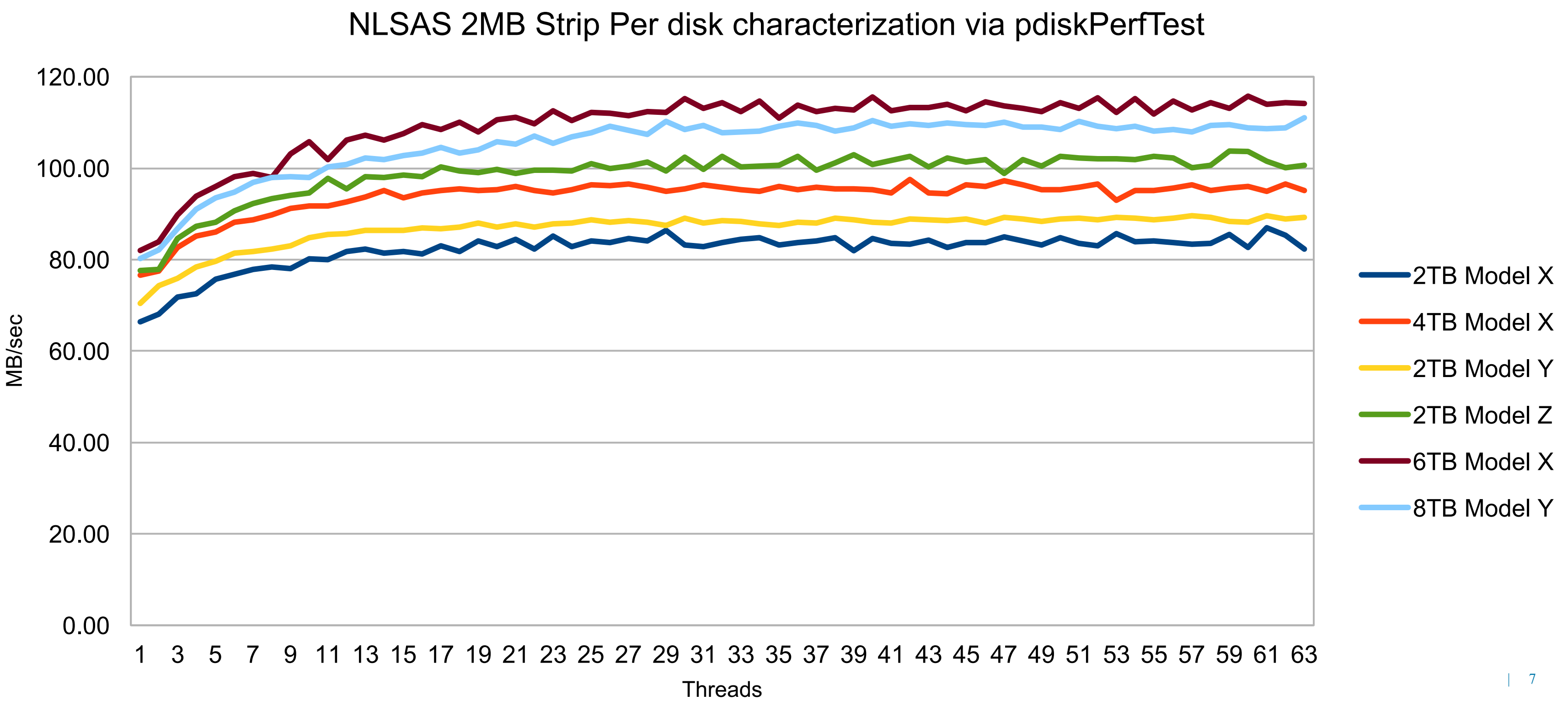
So who is right and how fast are they really ?

So what Influences total system speed

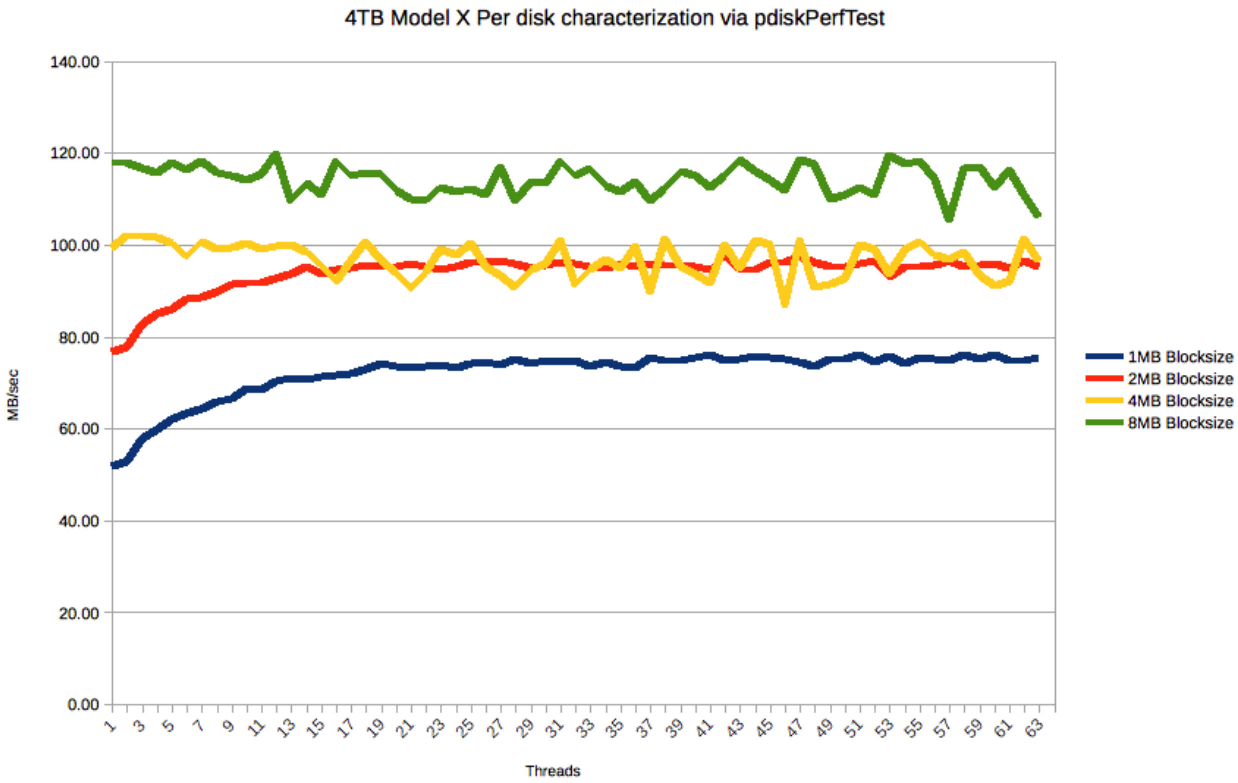
- Filesystem Overhead - who talks to block storage these days ?
- Controller Overhead - SW vs HW and how good is your raid implementation ?
- Raid mode Overhead - that’s a simple math problem 1P vs 2P vs 3P ..
- Cache efficiency - complex , main issue is what context is that i/o performed
- Application 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 ?

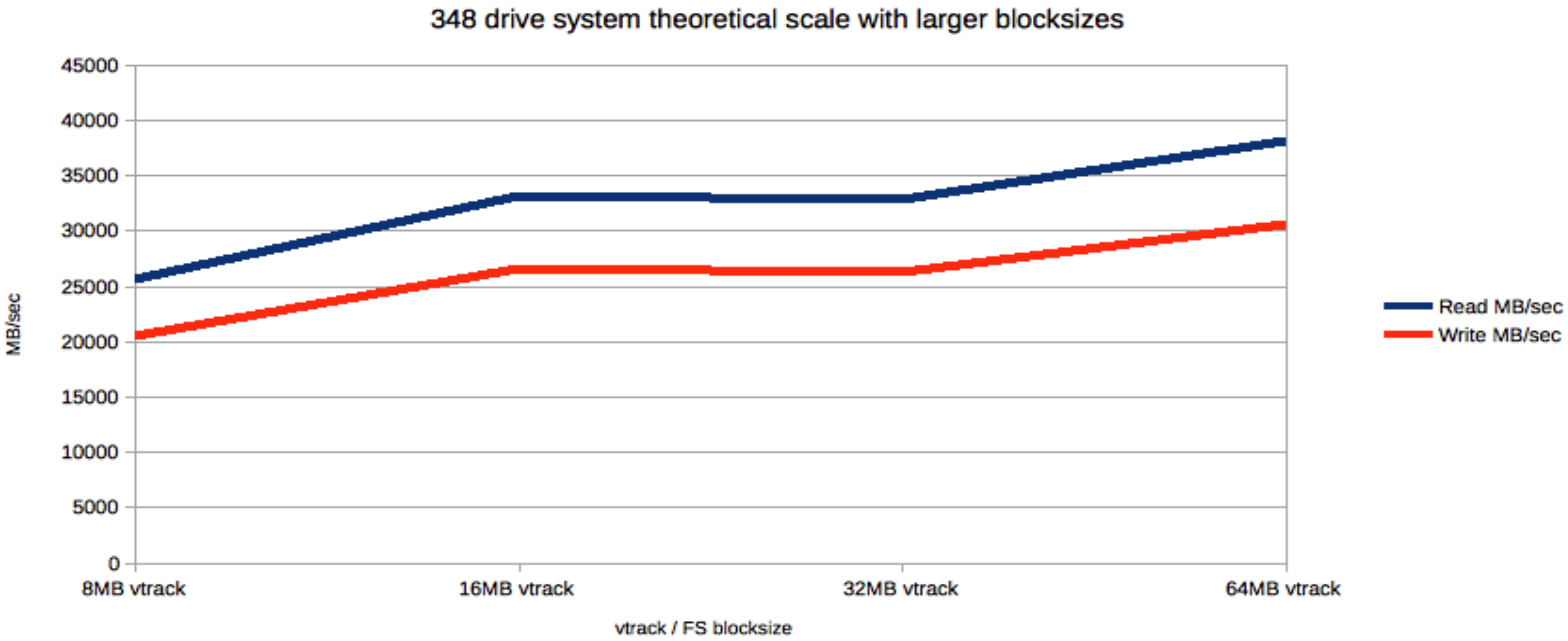
First - all access is random , lets take a look on how different models perform



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



How would performance change with different Block sizes



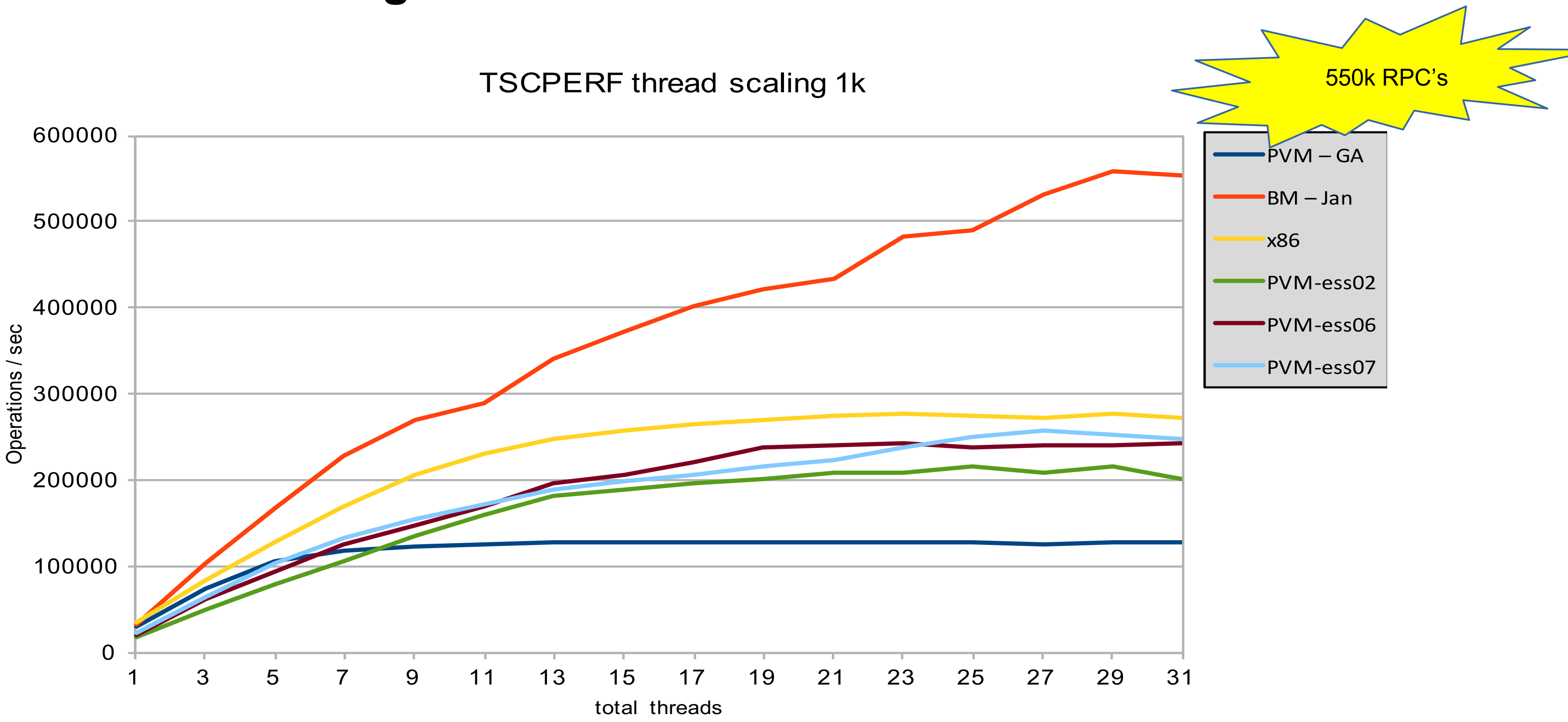
You need to overhaul the Network communication – done in 4.2.1+

- 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

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



Single client throughput enhancements

16 GB/sec single Node !

```
[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)

~50-80 usec per small Data RPC

```
[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 --
```

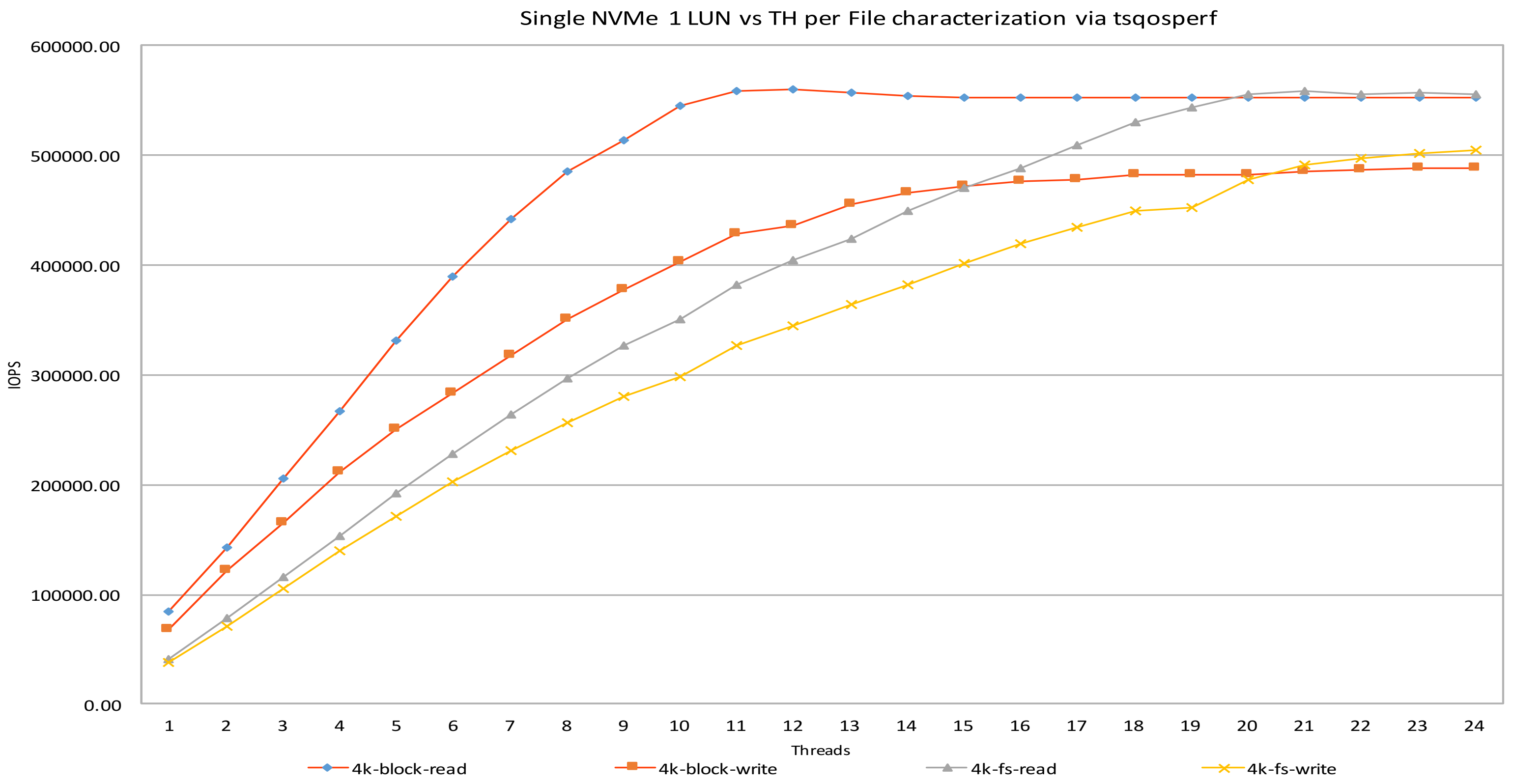
So what does all this mean to me ?

- Ask your Storage vendor for 100% random full block performance numbers, there is no sequential access, if you see 40-50 MB/sec/disk you should look for a different vendor ☺
- Don’t get confused from statements like ‘HW raid is superior’ , as you have seen on the previous slides SW raid can get the spec'd performance numbers out of a drive
- Cache matters, the i/o context matters, in order to get most efficiency you need to know the i/o context , you can only get this with embedded raid code in the filesystem
- Network matters, Bandwidth is not everything you need to be able to use it, you can’t beat RDMA today
- Latency matters, more layers don’t help, you need to condense layers
- Data distribution is important today, we will solve this soon

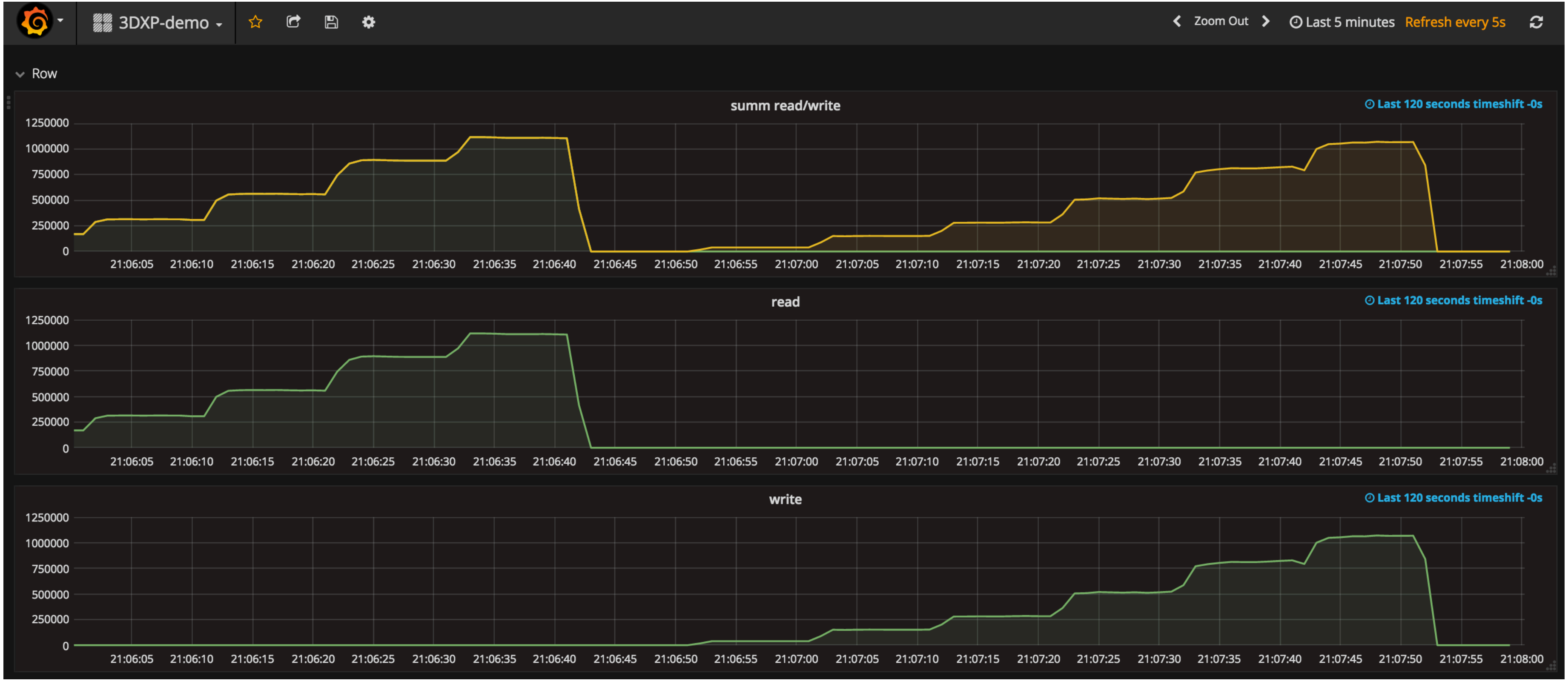
Agenda

- 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**
- IOPS - does it actually mean anything ??

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



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



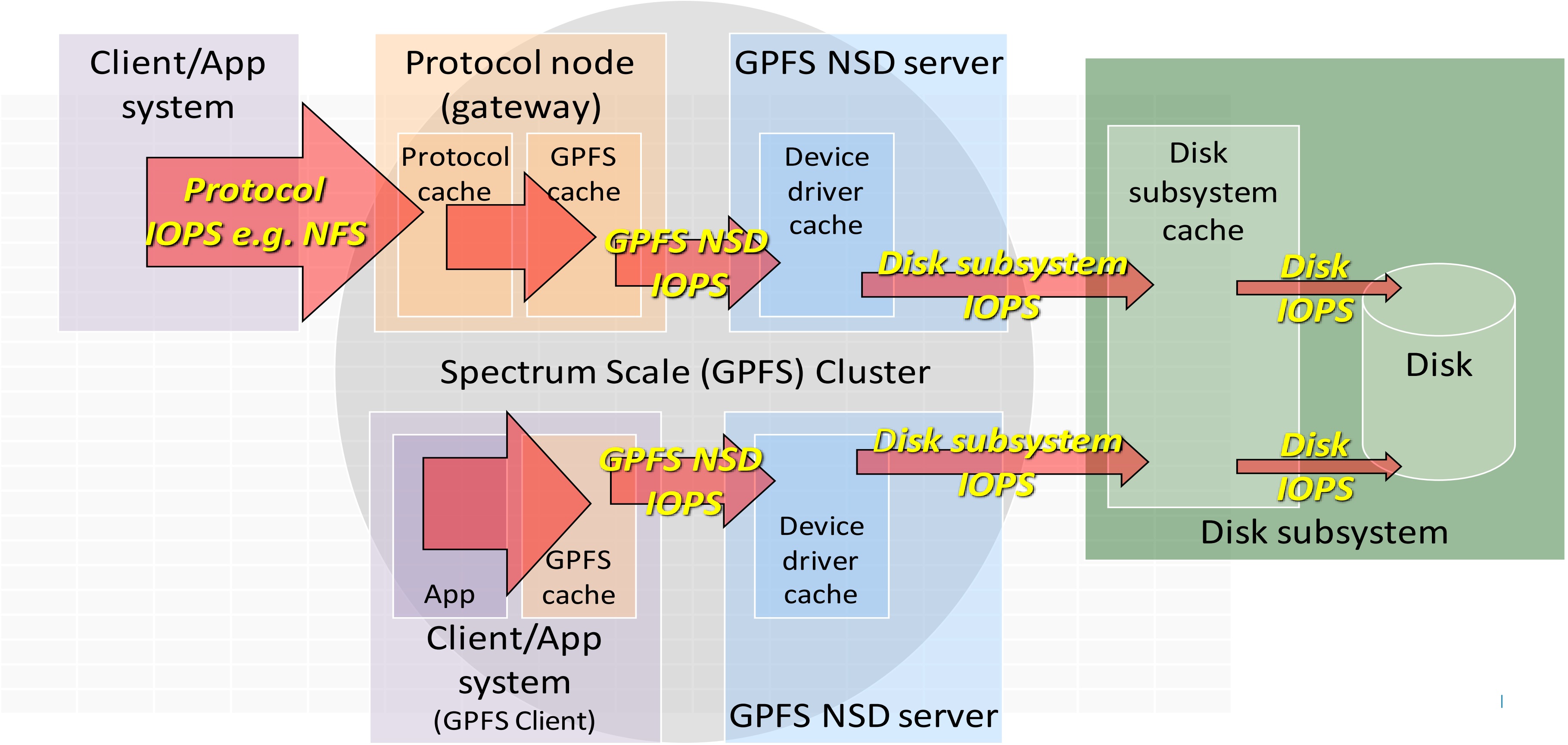
Agenda

- 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

- **IOPS - does it actually mean anything ??**

What are IOPS and why there is no answer to the "how many IOPS does a ESS GX system" question



“IOPS” run with IOZONE across 19 nodes

```
Iozone: Performance Test of File I/O
Version $Revision: 3.414 $
Compiled for 64 bit mode.
Build: linux-AMD64
Run began: Sat Mar 30 02:17:43 2013
OPS Mode. Output is in operations per second.
Record Size 4 KB
File size set to 33554432 KB
Network distribution mode enabled.
Command line used: /usr/local/bin/iozone -i 0 -i 1 -O -t 76 -r 4k -s 32g -+m
/ghome/oehmes/mpi/19.clients.76.iozone
Time Resolution = 0.000001 seconds.
Processor cache size set to 1024 Kbytes.
Processor cache line size set to 32 bytes.
File stride size set to 17 * record size.
Throughput test with 76 processes
Each process writes a 33554432 Kbyte file in 4 Kbyte records
```

IOZONE run results

Test running:
Children see throughput for 76 initial writers = 5859945.57 ops/sec
Min throughput per process = 67595.85 ops/sec
Max throughput per process = 87392.73 ops/sec
Avg throughput per process = 77104.55 ops/sec
Min xfer = 6488345.00 ops
Test running:
Children see throughput for 76 rewriters = 5925287.16 ops/sec
Min throughput per process = 67195.85 ops/sec
Max throughput per process = 88241.34 ops/sec
Avg throughput per process = 77964.30 ops/sec
Min xfer = 6388664.00 ops
Test running:
Children see throughput for 76 readers = 7193675.62 ops/sec
Min throughput per process = 79806.80 ops/sec
Max throughput per process = 112921.42 ops/sec
Avg throughput per process = 94653.63 ops/sec
Min xfer = 5929575.00 ops
Test running:
Children see throughput for 76 re-readers = 7195287.09 ops/sec
Min throughput per process = 76148.84 ops/sec
Max throughput per process = 121510.56 ops/sec
Avg throughput per process = 94674.83 ops/sec
Min xfer = 5257923.00 ops

So what does all this mean ?

The test was performing simultaneous sequential 4k buffered read and write I/O from 19 clients connected via 2 independent Infiniband fabrics to 2 fully populated GL6 systems.

The overall working set size was 2.4 Terrabyte while the combined cache size of all components was only 5% of the working set, this was chosen to eliminate the large influence of the cache to show the performance capabilities of the I/O path.

The write test performed **5.9 million 4k iops** or 23.7GB/sec throughput on average

The read test performed **7.2 million 4k iops** or 28.7GB/sec throughput on average

But a word of caution : This numbers are unlikely to be achieved with real customer workloads, like other Storage Systems provides peak numbers and were only collected to demonstrate the technical capabilities of the System

Legal notices

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

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.