Data To Network: building balanced throughput storage in a world of increasing disk sizes

> Tristan Suerink
IT Architect
Facts of storage at this moment

> We all want/need fast, cheap, reliable, bulk storage
> SSD’s are nice but still expensive
> Hard drives are getting a lot bigger without getting faster
> Host RAID controllers are lagging behind
> Everybody is focusing on object/http/s3/ceph
> How to combine all of this with our needs?
Storage requirements

- 8MiB/s per TiB read
- 4MiB/s per TiB write
- 2 reads and 1 write per 100TiB at the same time
- RAID6-like redundancy with maximum of 15 drives
- GridFTP/xrootd/webdav (you all know them ;-)
- Running CentOS
- Redundant connections to external enclosures
- Manageable for a real Sysadmin
Some History

- SUN X4500 with multiple software RAID on CentOS
- DDN SA9900 with Dell servers using Infiniband
- Dell R710 + MD3200 + MD1200 using SAS2
- Dell R515 with internal 3TB drives
- Supermicro 36x 4TB drives in 4U
Solution 2017

2GB/s Bottleneck!

HOST

Enclosure RAID

Enclosure RAID

Enclosure RAID

Enclosure RAID

Dell MD3460

Dell MD3460

DISK

DISK

DISK

DISK

DISK

DISK

DISK

DISK

cache mirroring

cache mirroring

cache mirroring

cache mirroring

zaterdag 12 mei 18
V Solution 2017

> 1x Dell R730 with 2x E5-2620v4 and 40Gbit/s NIC
> 2x Dell MD3460
  > 2x hardware raid controllers with SAS3 connectors
  > 60x 4TB NLSAS drives
> ±376TiB netto storage
> Uses 10U cabinet space
> ±36Gbit/s throughput measured at NIC
> Maximum cache mirroring 2GB/s
> Using 4TB drives, 1.8GB/s is needed
> So we’re stuck with 4TB drives due to performance constrains
Ok, what else? Options:

> DIY
  > JBOD with software raid (x86 and ppc64le)
  > JBOD with ZFS (x86)
  > JBOD with erasure coding (x86)
  > Object storage (white paper study)

> Vendor supported
  > Host hardware raid
  > Enclosure hardware raid
The JBODs

> JBOD test setup
  > Test one: HGST 4U60 60x 8TB helium drives
  > Test two: Sandisk Infiniflash 16x 8TB SSD’s

> Connected one after the other to:
  > One x86 server with dual socket E5-2690v3
  > two LSI SAS3008 with 4 paths in total

> Base line setup: Dell R730 + MD3460 raid6 13+2p

> Software:
  > Ubuntu 16.04 with mdadm
  > CentOS 7 with mdadm
  > FreeBSD 11 with ZFS
The JBODs, results:

<table>
<thead>
<tr>
<th></th>
<th>Ubuntu</th>
<th>CentOS</th>
<th>FreeBSD ZFS</th>
</tr>
</thead>
<tbody>
<tr>
<td>MiB/s</td>
<td>1.306</td>
<td>1.298</td>
<td>2.495</td>
</tr>
<tr>
<td></td>
<td>1.604</td>
<td>1.350</td>
<td>4.138</td>
</tr>
</tbody>
</table>

MD3460 - 4500 MiB/s

HGST 4U60

SanDisk Infiniflash

zaterdag 12 mei 18
The JBODs, pro and con:

> Pro’s:
  > It’s cheap
  > It can be fast
  > Maximum in control

> Cons:
  > How the hell do you correctly monitor your disks?!?
  > Less plug and play
  > Still a lot of work needed to be done on the software side
  > Linux mdadm raid6 implementation is a bottleneck
  > Very skilled sysadmin needed
The JBODs, erasure coding

> Mellanox cards have erasure coding offloading
> Ceph doesn’t work well on just one node as a “software raid”
> Other Reed Solomon based solutions weren’t usable at the time
> So, It’s not a suitable comparison system at this point...
V Latest solution

HOST

Enclosure RAID

cache mirroring

Enclosure RAID

DISK DISK DISK DISK

6GB/s Bottleneck!

Seagate AssuredSAN 4004
Latest solution

> 1x Fujitsu RX2530M4 with 2x Intel 4110 and 40Gbit/s NIC
> 1x Seagate AssuredSAN 4004
  > 2x hardware raid controllers with SAS3 connectors
  > 56x 8TB NLSAS drives
> ±350TiB netto storage
> Uses in total 5U cabinet space
> ±36Gbit/s throughput measured at NIC
> Maximum cache mirroring 6GB/s
> Using 8TB drives, 3.6GB/s is needed

> So we can grow using 12TB drives!
Let’s try something fun!

HOST

6GB/s Bottleneck!

Enclosure RAID ↔ cache mirroring ↔ Enclosure RAID

DISK  DISK  DISK  DISK  
Seagate AssuredSAN 4004

Enclosure RAID ↔ cache mirroring ↔ Enclosure RAID

DISK  DISK  DISK  DISK  
Seagate AssuredSAN 4004
Building a real (DTN) storage system

- 1x IBM S822L - POWER8 - 2x 8 core
- 2x Seagate AssuredSAN 4004
- 2x LSI 9405W-16e
- 2x 350TiB netto storage
- 8x SAS3 connections
- 2x 100Gbit/s ethernet
- No SSD’s!

Let’s compare with:

- Fujitsu RX2530M4 - 2x Intel Xeon 4110 Silver
- Dell R7415 - 1x AMD EPYC 7451
## Results!

<table>
<thead>
<tr>
<th></th>
<th>Intel Xeon</th>
<th>AMD EPYC</th>
<th>POWER8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reads and Writes combined</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R1</td>
<td>2.764 MiB/s</td>
<td>3.260 MiB/s</td>
<td>6.284 MiB/s</td>
</tr>
<tr>
<td>W1</td>
<td>1.777 MiB/s</td>
<td>2.226 MiB/s</td>
<td>3.660 MiB/s</td>
</tr>
<tr>
<td>R1+W1</td>
<td>3.660 MiB/s</td>
<td>5.072 MiB/s</td>
<td>7.040 MiB/s</td>
</tr>
<tr>
<td>R2+W1</td>
<td>4.50 MiB/s</td>
<td>4.02 MiB/s</td>
<td>8.80 MiB/s</td>
</tr>
<tr>
<td>R8+W4</td>
<td>3.90 MiB/s</td>
<td>4.76 MiB/s</td>
<td>10.135 MiB/s</td>
</tr>
<tr>
<td>R16+W8</td>
<td>3.22 MiB/s</td>
<td>3.59 MiB/s</td>
<td>8.945 MiB/s</td>
</tr>
</tbody>
</table>
Reason for IBM POWER

- PCI-e Gen 4 (2x faster than Gen 3)
- Few cores with high clock frequency (>> 3.8 GHz)
- PCI-e lanes can be collapsed on a single core per socket
- Future proof with the OpenCAPI interface
- Capable of handling 1Tbit/s of total throughput
- 70% CPU resources free for network handling
- With one POWER box, I can handle the load of three x86 boxes
The future!
Questions?

Thanks to the following companies for helping out:

- HGST/WD/Sandisk
- Seagate/DotHill
- IBM
- Dell
- Fujitsu
- Intel
- Mellanox

More info: https://wiki.nikhef.nl/grid/SystemDesign
V Test Environment

> OS: Ubuntu 18.04
> Test suite: http://web.grid.sara.nl/acceptance_test/
> Used the dd test from this suite
> 350TiB of testfiles
> Re-used the same HBAs and SAS cables
> All firmwares on the latest stable level
> Hyperthreading turned off
> Optimal memory configuration for every platform
Host hardware RAID

- Going above 3TB drives is not recommended
- 1.8GB/s at max just on 100% read
- Mixed read/write at max ±1.1GB/s combined
- These cards prefer writing and don’t have enough buffer to give reading a chance
- There is also not a lot of development in this area