HUF2019 – SciNet Site Presentation
15-18 October 2019, - Indiana University.
by Jaime Pinto
Agenda for this HPSS report

• Last year’s agenda summary
• Last year’s “what is next” summary
• HPSS Overview at SciNet (2019)
• Remote site considerations: cache, network, etc
• Repack utilities – “parallel schedule”
• Pending SciNet BI’s and CR’s (concerns)

What is next?
Last year’s agenda summary

We were testing/investigating HPSS on 5 fronts:

• Upgrade to 7.5.2.u1 and 7.5.2.u2 series: premature EOM
  o bug 6919 was fixed with version 7.5.2.u3, we are now running 7.5.3.u6 on test and production systems

• Comparing Dell MD3460 vs. Dell DSS7500 commodity appliance for disk cache
  o 2x MD3460 (640TB): $89K (storage only)
  o 2x Dell R730xd movers: $22K
  o $156/TB of usable cache ($125/TB without the movers)
  o 3x DSS7500 (1650TB): $132K
  o $83/TB of usable cache (movers included)
  o ~140TB/day sustained ingestion/recalls in “cruise mode”

• Hierarchy exercise: primary and secondary cache for remote site over 1G network (100G now).
  o More comments in the current overview (next few slides)

• Access via VFS-fuse: single mount point for 2 COS’s plus ‘HSI chcos auto’
  o This has become part of regular operation in the production system via a crontab, until this functionality is built into HPSS (CR 547 - bug # 6402) [caveat]

• SwiftOnHPSS with the latest releases of HPSS and Openstack/Keystone (Platform 13 – Queens)
  o This initiative morphed through Tree Frog into the (Ex-Cloud) now Storage Broker project (ongoing). More comments by Jim Gerry
Last year’s “what is next?”

- Re-deploy the 3x DSS7500’s disk-movers as part of the production system, in addition to the existing 2x MD3460, for a total of 1588TB large-files cache and 710TB never-purge small-files cache. Actually 1650TB large-files.
- We are placing an order for 8x LTO8 tape drives and 1000x LTO8 tapes. Will repack all 1st copy material from LTO7 to LTO8, and sometime in 2019 ship all LTO7 gear and tapes to CAC, so the 2nd copy will be hosted there. Actually 3000x LTO-M tapes (IBM couldn’t deliver LTO8’s)
- Deploy /archive junction as a single mount point for small/large files COS in production.
- Possibly use VFS-fuse as a replacement for Globus-DSI, since this project has gone nowhere in 2 years. Actually we readopted Globus-DSI 2.7, in addition to VFS-fuse.
- Further investigate SwiftOnHPSS as a replacement for WOS from DDN.
- Perhaps collaborate with DKRZ on some SwiftOnHPSS development
- RDM - Research Data Management (archival, repository, curation, preservation, citation, metadata)
  - Integrate Globus Data Publication with HPSS (this effort wasn’t leveraged – Globus dropped Publication)
  - Integrate existing Openstack-Swift infrastructure from another department at the University of Toronto – Scholarsportal (this effort wasn’t leveraged – HPSS dropped openstack)
  - We now planning to integrate with FRDR (Canadian Federated Research Data Repository): Globus+DSI+HPSS
- We also experimented with CRAY G200 GPFS for cache + StarFish for discovery
- And we did some testing with HSI/HTAR quotas (Gleicher & Jaime pet project)
HPSS Overview at SciNet

**HPSS**
(7.5.3.u6)

**GPFS**
10PB

**Test system**
SciNet - UofT

- core03
- mover04

**Small Files**
1.8PB

**Large Files**
40TB

**Production system**
SciNet - UofT

- core04
- mover05
- mover06
- mover07
- mover08
- mover09

**Small Files**
690TB

(**never purge**) **Large Files**
1650TB

**Test system 2**
SciNet - UofT

- core05
- mover10

**TS3500**
1st copy tapes
27 PB / 16x LTO8 drives

**TS4500**
2nd copy tapes
18 PB / 12x LTO7 drives

**Production system**
CAC – Queens University

- mover11
- mover12

**Production system**
CAC – Queens University

- mover13

**260 km**
140TB/day

**globus**

**hsi/htar**

**archive02**

**vfs-fuse**

**vfs02**

**dsi02**

**2x40G**

2x40G

2x40G

100G

140TB/day
Last year: Hierarchy Editor (during the pilot)

Very slow direct dump to tape
Last year: Hierarchy Editor (during the pilot) + remote cache

Still very slow transfers, but …

Fast dump to tape from cache
Today (no remote cache yet). But remote drives are always “busy” …

Avg 170-200MB/s per drive for 8 drives

Avg 240-270MB/s per drive for 12 drives
Houston, we have a problem!!!

Why can’t I just insert a (empty!!!) cache node here?

I have to create a new cache+tape branch (new COS) and move same content...
Houston, more problems!!!

- Limitation on what/how to change/edit existing hierarchy.
- Can not just add (cache) nodes in the middle of hierarchy, only below.
- Otherwise I have to create a whole new branch, as a new COS.
- **[caveat]** chcos currently means moving data already on tapes (1\(^{st}\) and 2\(^{nd}\) copies) to a new set of tapes, through the primary cache (impractical at scale, until possibly the next tape technology upgrade).
  Not sure yet the role that "COS Copy to Disk" flag could play in alleviating this situation.
- This should be just a change in DB2 entries, as in the change in PVR GUI tool from local to remote.
- “parallel” remote branches mean 2x migration, or retention on primary cache indefinitely (undesirable).

We are in a limbo for now ...

Moving new 2\(^{nd}\) copy material only through the branch with cache might be an acceptable workaround, while leaving old material where it is, until the next major tape technology upgrade at CAC (LTO7 to LTO8). This would still imply moving data on the 1\(^{st}\) copy as well, but at least would buy us time, until HPSS comes up with a better solution for chcos utility. Or we wait until both SciNet and CAC deploy LTO9 !!!
Cache Considerations

- Always better to have more cache, as much as possible.
- Repurpose/stretch the life of “older” storage appliances/file systems if possible (GPFS, Lustre, etc)
- Consider commodity Storage Building Blocks as well (Dell, HP, Lenovo, etc)
- Keep files on cache for as long as possible, use high watermarks for purging.
- Never purge small files from cache. Always better/faster to recall from cache than from tapes.
- Remote cache keeps remote tape drives from threading water during migrations, in case of network/transfer inefficiencies. Drives don’t have to wait, so are freed up for repacks.
- But apparently, if we don’t setup the cache in the hierarchy from day one, we are in a predicament
Network Considerations

ESnet has a presentation that shows that even a small amount of packet-loss, combined with latency, can have a killer effect. The 100G tuning a slide (near the end) shows 10x decrease in throughput when you get to "regional"-sized distances, even with multiple streams.

https://www.es.net/assets/Uploads/100G-Tuning-TechEx2016.tierney.pdf

SciNet-CAC: RTT 6.3 ms

![Throughput vs. Increasing Latency with .0046% Packet Loss](image-url)
Kernel optimization parameters on all movers

# /etc/sysctl.conf
# kernel semaphore limits for HPSS 7.5
kernel.shmni = 32768
kernel.msgmni = 131072
kernel.shmmmax = 134907760640
kernel.shmall = 65872930
kernel.sem = 4096 2048000 32 32768
kernel.msgmnb = 65536
kernel.msgmax = 65536
kernel.randomize_va_space = 0

# 100Gbps optimization parameters (MTU 9000, qlength 60000)
net.ipv4.tcp_congestion_control=htcp
net.ipv4.tcp_mtu_probing=1
net.core.rmem_max=2147483647
net.core.wmem_max=2147483647
net.ipv4.tcp_rmem=4096 87380 2147483647
net.ipv4.tcp_wmem=4096 65536 2147483647
net.core.default_qdisc=fq

# GPFS/IB optimization parameters
net.ipv4.conf.ib0.arp_filter=1
net.ipv4.conf.ib0.arp_ignore=1
net.ipv4.tcp_low_latency=1
Repack Utilities: “parallel schedule”

- Bash utilities set available as a tarball, upon request
- General idea: leverage any idle drives for as long as possible
  - showNext.sh: tapes with largest deletion holes flagged first (or early EOM)
  - showDriveLoop.sh: keeps track of available drive-pairs on the libraries
  - showActivityLoop.sh: /opt/hpss/bin/rtmu summary
  - repackLOOP.sh: function(showNext, showDriveLoop)
  - killRepackLOOP.sh: function(showActivityLoop, drive waits)
Pending SciNet BI’s

• Quotas (space & files, by users & groups, for large uid# & gid#): since HUF 2011, here at IU
  o SciNet doing some work with Gleicher on HSI/HTAR quotas, but it wouldn’t enforce access via VFS or Globus
• htar limitations (68GB max per file, long path names, file name character set+): since HUF 2012
• These might be the longest running BI’s ever, still without a solution.
  o I understand from Gleicher he won’t do any work for HPSS 8.1+, and he will be out of NERSC by early 2020.

Pending CR’s & concerns

• CR 547 (bug # 6402): Access via VFS-fuse, single mount point for multiple COS’s with file auto-placement in the proper COS.
• Globus-DSI support for HPSS 8.1 may become an issue
• We’ll open a CR (or BI) for introduction of new cache onto existing hierarchy.
• We may open CR to have multiple streams per file/segment for remote migrations and repacks, so to better saturate network connections, similar to GridFTP, bbcp, WDT, iperf3, etc.
• We’d like to see some priority as well for the CR to have fine control over which files to purge first, based on size over age first, for instance.
What is next?

- Revamp the TS3500 at SciNet and turn it into a TS4500 (components delivered)
- Populate the new TS4500 with 3000 new ‘true’ LTO8 tapes (delivered)
- Add 2 more LTO-8 drives at SciNet (delivered).
- Send 4 more LTO-7 drives to CAC, for a total of 16 (will need a new high-density enclosure).
- Deploy CRAY G200 appliance as cache for small files in production: 1.8PB capacity.
- Deploy Dell MD3460 as remote cache at CAC: 690TB capacity.
- Deploy HSI/HTAR based quotas from Gleicher
- Upgrade test system to HPSS 8.1, and later production system to 8.1.u1 or 8.1.u2 (TBD-hsi/htar).
- Start the migration from LTO-M to LTO-8 tapes at SciNet. Total 1st copy capacity 36PB.
- Send 3000x LTO-M’s to CAC. Total 2nd copy capacity 36PB.
- Start Storage Broker testing. Hardware already in place (test system 2).
- RDM - Research Data Management (archival, repository, curation, preservation, citation, metadata, etc), by integrating HPSS with FRDR (Canadian Federated Research Data Repository): Globus+DSI+HPSS (+ StarFish).

- BTW: Niagara is being expanded from 60,000 to 80,000 cores, and doubling storage to 20PB! Maybe we can keep our #69 rank in the Top 500 in 2020, and remain #1 in Canada.
Questions???

Many thanks!
Jaime Pinto  
pinto@scinet.utoronto.ca