University of Pennsylvania (UPENN) PennMedicine HPC data migration project

Anand Srinivasan
Sr. IT Project Leader

asrini@upenn.edu
or
Anand.Srinivasan@pennmedicine.upenn.edu

www.pennhpc.org

SC17 Spectrum Scale User Group Meeting
Sunday, November 12th, 2017

Acknowledgments

- AdvancedHPC Team
  - Joe Lipman
  - Jeff Tomlinson
  - Toni Falcone
  - Cesar Leal
  - Mark Christe
- General Atomic
  - Martin Margo
- IBM

Who Are We?

- PennMedicine
  - Faculty from the UPENN School of Medicine (PSOM)
  - Clinicians from the UPENN Health System (UPHS/HUP)

- Penn Medicine Academic Computing Services (PMACS)
  - IT group within PennMedicine IS that services mostly PSOM

- PMACS Enterprise Research Applications - HPC (ERA - HPC)
  - An even smaller group within PMACS that services HPC related needs of faculty, researchers and clinicians within PennMedicine.
    - **2 FTE !**

What does our group do?

- Support a group of ~500 named users (faculty, researchers/post-docs, students, staff & clinicians) on our HPC
  - Genomics/Bioinformatics workloads
- Everything HPC related at PennMedicine
  - Systems Administration
    - Install and configure HPC compute nodes
      - Use Chef for configuration management
    - Manage packages
  - HPC Storage Administration
    - SONAS
    - GPFS
    - Archive
  - End User Support
    - Account creation
    - Troubleshoot issues with jobs, work-flows/code
    - End-user training
      - Wiki
  - Plan, architect and procure HPC hardware/solutions
  - HPC billing
  - Occasional AWS support
  - And other duties as assigned! :-)

Our current (older) gear

- 144 IBM iDataPlex compute nodes
  - IBM LSF Job Scheduler
- ~2.7 PB (raw) SONAS
  - 1.7 PB usable
  - IBM's "Scale Out NAS"
  - Runs GPFS
  - 20 nodes (2 mgmt, 10 interface, 8 storage)
- ~3.6 PB Archive
  - Spectralogic Tape Library
  - Filetek (old)
  - Quantum (new) – migration ongoing
- ~1.8 PB misc storage
  - Different data center
  - Used for AFM

Our current (new) gear

- ~4.3 PB (raw) GPFS storage
  - Purchased Summer, 2017
  - 4x Mercury 4U 60-drive controllers
  - 4x Mercury 4U 60-drive expansion units
    - 1x Mercury 4U 60-drive expansion unit (extra)
  - 1x Mercury 2U 24-drive SSD array
  - 6x NSD servers
    - RHEL v7.3
    - IB connected
    - 10GigE for GPFS client traffic
    - 40GigE up-links to our core network
  - GPFS v4.2.3.0

The Problem

- Migrate all data on a nearly 100% full 1.8PB SONAS that is very actively used over to recently purchased GPFS based storage (when it arrives).
  - Maintain close to 100% uptime
  - More importantly, maintain end-user satisfaction
    - Failed jobs due to I/O errors and due to file system reaching capacity
    - Weird NFS issues
      - Played a new game called SONAS Whac-A-Mole!
    - Latency issues due to AFM
      - A failed experiment
  - And do all this before SONAS support runs out!

- Meanwhile back on the SONAS ….
  - With only 82TB left (~23TB in “gold” pool/default write location) users still adding data at the rate of a few TB/day!
  - Wait, didn’t you say you had only 1.7PB of usable space on SONAS?
    - Re-purposed some of the SONAS system pool NSDs to buy us time

SONAS Whac-A-Mole

- NFS clients lose connectivity randomly to NFS (SONAS) interface nodes
  - Fix: force unmount and remount the filesystem

```
[asrini@consign ~]$ bjobs -u all -m node005
No unfinished job found on host/group <node005>
[asrini@consign ~]$ ssh node005
Last login: Fri Nov 10 21:35:30 2017 from 172.16.105.100
[sysadmin@node005 ~]$ date
Fri Nov 10 21:38:49 EST 2017
[sysadmin@node005 ~]$ uptime
 21:38:51 up 2 days, 14:06,  1 user,  load average: 6.39, 4.63, 2.65
[sysadmin@node005 ~]$ ps aux|grep -v root
USER       PID %CPU %MEM    VSZ   RSS TTY      STAT START   TIME COMMAND
rpc       2340  0.0  0.0  18976   912 ?        Ss   Nov08   0:00 rpcbind
rpcuser   2360  0.0  0.0  23348  1332 ?        Ss   Nov08   0:00 rpc.statd
ntp       3314  0.0  0.0  28588  2028 ?        Ss   Nov08   0:00 ntpd -u ntp:ntp -p /var/run/ntpd.pid -g
postfix   3412  0.0  0.0  83264  3496 ?        S    Nov08   0:00 qmgr -l -t fifo -u
postfix  19035  0.0  0.0  83092  3456 ?        S    21:16   0:00 pickup -l -t fifo -u
sysadmin 25420  0.0  0.0 102084  1956 ?        S    21:38   0:00 sshd: sysadmin@pts/0
sysadmin 25421  0.1  0.0 108336  1788 pts/0    Ss   21:38   0:00 -bash
sysadmin 25513  1.0  0.0 110228  1144 pts/0    R+   21:38   0:00 ps aux
[sysadmin@node005 ~]$ top -b -n1 | head
top - 21:40:01 up 2 days, 14:07,  1 user,  load average: 6.12, 4.91, 2.89
Tasks: 638 total,   1 running, 637 sleeping,   0 stopped,   0 zombie
Cpu(s):  0.3%us,  0.1%sy,  0.0%ni, 99.6%id,  0.0%wa,  0.0%hi,  0.0%si,  0.0%st
Mem:  264504388k total,  7895212k used, 256609176k free,   496040k buffers
Swap: 10485756k total,        0k used, 10485756k free,   945404k cached

  PID USER      PR  NI  VIRT  RES  SHR S %CPU %MEM    TIME+  COMMAND
 3963 root       0 -20 42392 4780 1776 S  3.9  0.0   7:47.59 isflim
25804 sysadmin  20   0 15428 1540  824 R  2.0  0.0   0:00.01 top
    1 root      20   0 23376 1580 1272 S  0.0  0.0   0:05.06 init
```

Other NFS issues?

- Random I/O errors

```
[root@fs02nsd04 ~]# date; ls /mnt/[redacted]/Unaligned/Basecall_Stats_COHP8ACXX/Phasing/ |wc -l
Thu Oct 12 10:43:10 EDT 2017
ls: reading directory /mnt/[redacted]/Unaligned/Basecall_Stats_COHP8ACXX/Phasing/: Input/output error
814
```

```
[root@fs02nsd04 ~]# date; ls /mnt/[redacted]/Unaligned/Basecall_Stats_COHP8ACXX/Phasing/
Thu Oct 12 10:43:18 EDT 2017
s_1_02_phasing.xml      s_2_1_1301_cycle.txt    s_3_1_2202_phasing.txt  s_4_2_1106_phasing.txt  s_5_2_2302_phasing.txt  s_6_3_1207_cycle.txt    s_7_3_2201_cycle.txt
s_1_103_phasing.xml     s_2_1_1301_phasing.txt  s_3_1_2203_cycle.txt    s_4_2_1107_phasing.txt  s_5_2_2303_phasing.txt  s_6_3_1207_phasing.txt  s_7_3_2201_phasing.txt
s_1_110_phasing.xml     s_2_1_1302_cycle.txt    s_3_1_2203_phasing.txt  s_4_2_1108_phasing.txt  s_5_2_2304_phasing.txt  s_6_3_1208_cycle.txt    s_7_3_2202_cycle.txt
s_1_1_1101_cycle.txt    s_2_1_1302_phasing.txt  s_3_1_2204_cycle.txt    s_4_2_1201_phasing.txt  s_5_2_2305_phasing.txt  s_6_3_1208_phasing.txt  s_7_3_2202_phasing.txt
s_1_1_1101_phasing.txt  s_2_1_1303_cycle.txt    s_3_1_2204_phasing.txt  s_4_2_1202_phasing.txt  s_5_2_2306_phasing.txt  s_6_3_1301_cycle.txt    s_7_3_2203_cycle.txt
s_1_1_1102_cycle.txt    s_2_1_1303_phasing.txt  s_3_1_2205_cycle.txt    s_4_2_1203_phasing.txt  s_5_2_2307_phasing.txt  s_6_3_1301_phasing.txt  s_7_3_2203_phasing.txt
s_1_1_1102_phasing.txt  s_2_1_1304_cycle.txt    s_3_1_2205_phasing.txt  s_4_2_1204_phasing.txt  s_5_2_2308_phasing.txt  s_6_3_1302_cycle.txt    s_7_3_2204_cycle.txt
s_1_1_1103_cycle.txt    s_2_1_1304_phasing.txt  s_3_1_2206_cycle.txt    s_4_2_1205_phasing.txt  s_5_2_phasing.txt       s_6_3_1302_phasing.txt  s_7_3_2204_phasing.txt
s_1_1_1103_phasing.txt  s_2_1_1305_cycle.txt    s_3_1_2206_phasing.txt  s_4_2_1206_phasing.txt  s_5_3_1101_cycle.txt    s_6_3_1303_cycle.txt    s_7_3_2205_cycle.txt
```

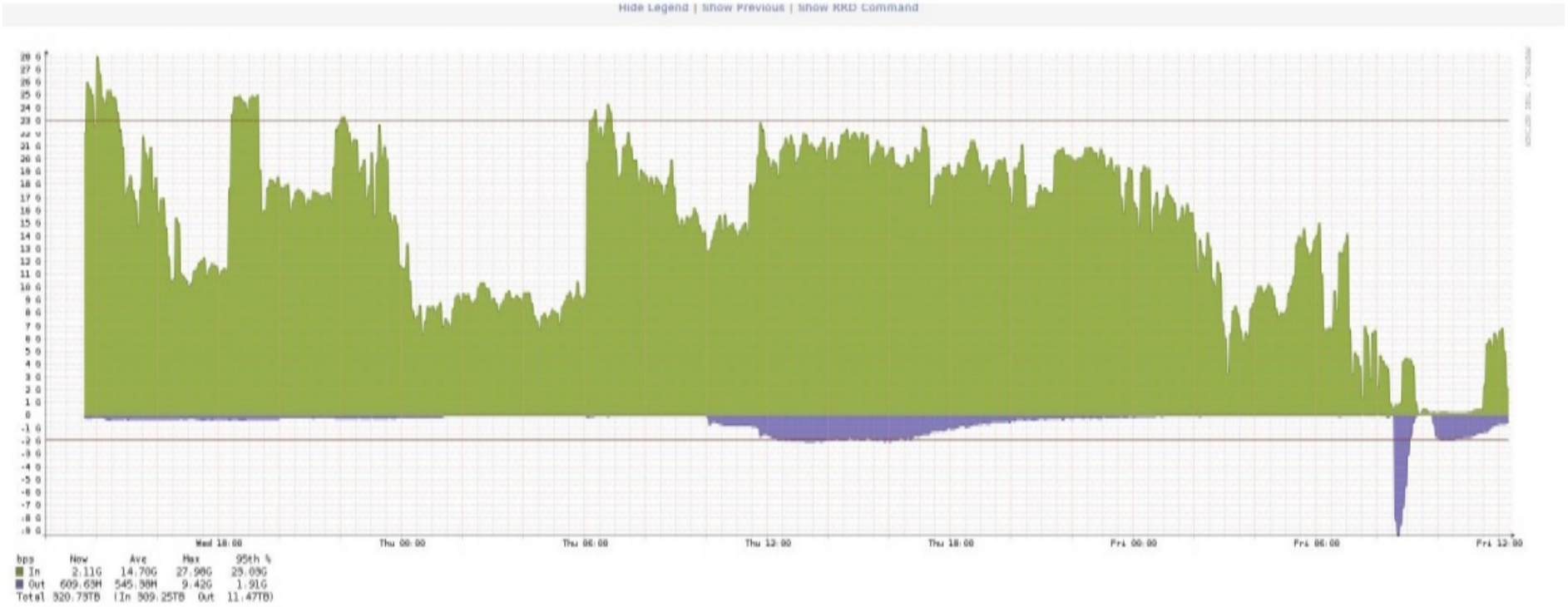
```
[root@fs02nsd04 ~]# date; ls /mnt/[redacted]/Unaligned/Basecall_Stats_COHP8ACXX/Phasing/ |wc -l
Thu Oct 12 10:43:20 EDT 2017
1971
```

New (GPFS) storage to the rescue!

- Pre-“burn-in” by AdvancedHPC/GA
  - NSD server/cluster setup
  - LUNs setup
- ~12-13 hrs install/setup time after arriving at our datacenter
- SONAS to new GPFS migration kicked off within a couple of days of install
- Used a combination of
  - tar
    - (cd /<src> && tar cf -.)|pv|(cd /<dest> && tar xf -)
  - rsync
    - Catch stuff that was missed
      - Remember those weird NFS IO errors?
  - ncat
  - cron
    - Periodically do directory listing on NFS mount points
      - Remember those weird NFS IO errors?

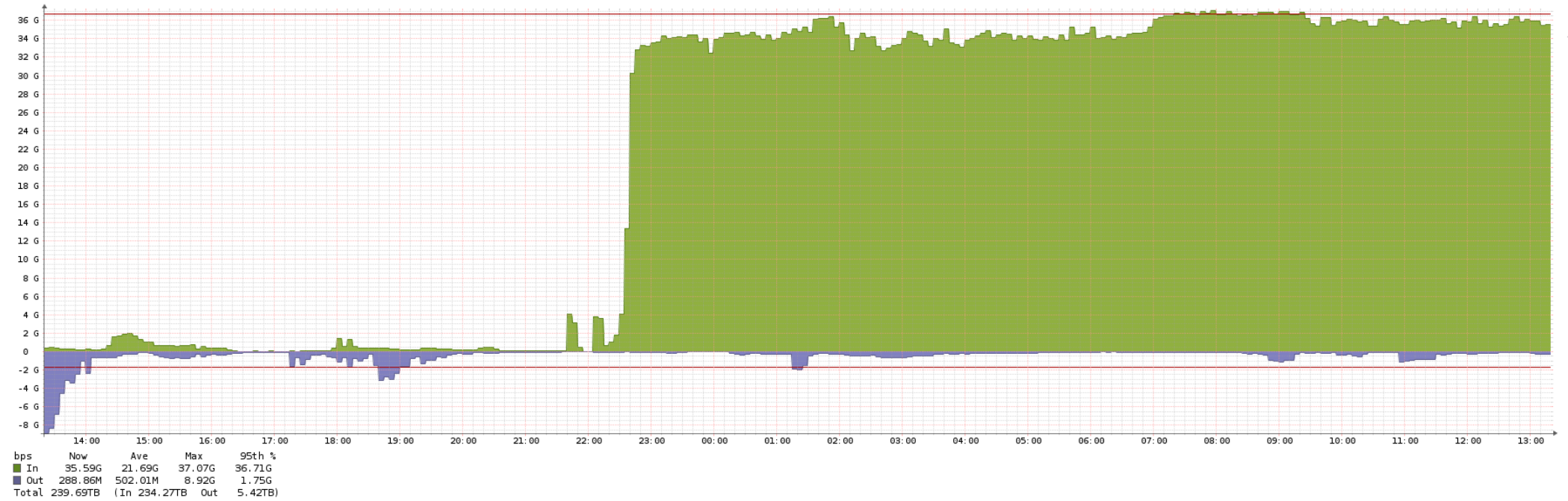
Results

- ~58% utilization of 40GigE link
  - ~320TiB in 48hrs



More results...

- ~86% utilization of 40GigE link
  - ~240TiB in ~13 hrs!



Questions?
or
Suggestions?

Thank you!