IllinoisHEP Tier3gs Storage Status

David Lesny

Senior Research Physicist

IllinoisHEP T3gs Storage

- Atlas T3 with Grid Services

- Panda site
  - Production queue (IllinoisHEP-condor)
  - Analysis queue (ANALY_IllinoisHEP-condor)

- Software
  - Scientific Linux 5.5 (64 bit)
  - dCache 1.9.5-21 (Chimera) installed via VDT 3.0.3

- Hardware
  - 8 nodes (3 doors, 1 head, pNFS, and 3 pool nodes)
  - Dell R710 (E5540, 24GB) and Intel (E5345, 16GB)
  - H800/MD1200/2TB SAS disks (144TB raw)
  - 12 Drives (1 Tray) per Raid 5 set with 512KB strip size, XFS file system
  - 10Ge network (HP5400, Intel Dual CX4)

IllinoisHEP T3gs Storage

- Good performance
  - Pool nodes are over 1GB/s read, 800MB/s write via dd, 600MB/s Bonnie++
  - FTS transfers over 700MB/s

- Issues
  - dCache 1.9.5-19 and -21 fixed many problems
  - Update using VDT package is very easy

- Network tuning very important
  - 10Ge tuning different than 1Gb
  - Cards need to be in 8x PCIe slots (R710 has both 8x and 4x)
  - Much larger memory needs

- Problems seen with bad tuning
  - Broken network connections
  - Files transferred with errors (bad adler32 checksums)

IllinoisHEP T3gs dCache Tweaks

- Some tweaking of dCache parameters recommended by T2 sites
  - Use 64bit java with memory increase to 2048/4096M
  - gsiftpMaxLogin=1024
  - bufferSize=8388608
  - tcpBufferSize=8388608
  - srmCopyReqThreadPoolSize=2000
  - remoteGsiftpMaxTransfers=2000

- Use Berkley Database for meta data on pool nodes

metaDataRepository=org.dcache.pool.repository.meta.db.BerkeleyDBMetaDataRepository

Network tuning

```
### Additions made by ESnet (http://fasterdata.es.net/TCP-tuning)

# Turn on window scaling
net.ipv4.tcp_window_scaling = 1

# Turn TCP timestamp support on
net.ipv4.tcp_timestamps = 1

# Turn SACK support off
net.ipv4.tcp_sack = 0

# 256 KB default performs well experimentally, and is often recommended by ISVs.
net.core.rmem_default = 262144
net.core.wmem_default = 262144

# Increase TCP max buffer size setable using setsockopt()
net.core.rmem_max = 56623104
net.core.wmem_max = 56623104

# Increase Linux autotuning TCP buffer limits
# min, default, and max number of bytes to use
# Set max to 16MB for 1GE and 32M (33554432) or 54M (56623104) for 10GE
net.ipv4.tcp_rmem = 4096 87380 56623104
net.ipv4.tcp_wmem = 4096 65536 56623104

# Always have enough memory available on a UDP socket for an 8k NFS request,
# plus overhead, to prevent NFS stalling under memory pressure.  16k is still
# low enough that memory fragmentation is unlikely to cause problems.
net.ipv4.udp_rmem_min = 16384
net.ipv4.udp_wmem_min = 16384

# Don't cache ssthresh from previous connection
net.ipv4.tcp_no_metrics_save = 1

# Recommended to increase this for 10G NICS
net.core.netdev_max_backlog = 30000

# Use the Cubic congestion control
net.ipv4.tcp_congestion_control = cubic
```

More network tuning

```
# Increase the transmit buffers
        ifconfig eth0 txqueuelen 5000

# Turn on flow control
        ethtool -A eth0 autoneg on
        ethtool -A eth0 rx on
        ethtool -A eth0 tx on

# Increase the ring buffers
        ethtool -G eth0 rx 4096
        ethtool -G eth0 tx 4096

# Turn on all the assists we can get
        ethtool -K eth0 rx on
        ethtool -K eth0 tx on
        ethtool -K eth0 sg on
        ethtool -K eth0 tso on
        ethtool -K eth0 gro on
        ethtool -K eth0 gso on
```

Dell Open Manage Server Administrator

http://linux.dell.com/repo/hardware/

Installation

wget -q -O - http://linux.dell.com/repo/hardware/latest/bootstrap.cgi | bash

yum install -y srvadmin-all
yum install -y dell_ft_install
yum install -y $(bootstrap_firmware)

Useful commands

inventory_firmware
inventory_firmware_gui
update_firmware

Invoke OMSA

https://node.name:1311/