Adventures in AFM

DataDirect Networks UK

Vic Cornell

2015/11/12

Please Ask Questions

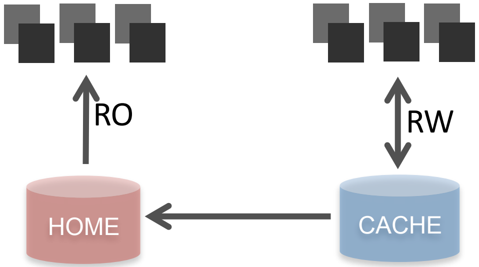
Adventures in AFM

- Still feels like a “new” feature.
- New Features or Tunings Appearing all the time.
- Very much a toolkit which makes it a bit challenging.
- Some sites started with basic modes and are now looking at Async DR as an upgrade.

AFM Modes with GRIDScaler

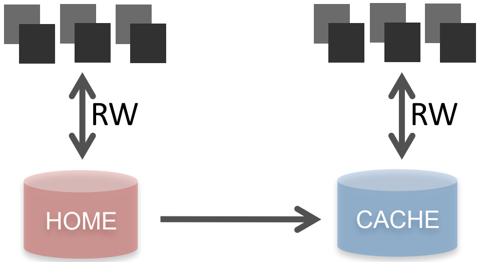
Active File Management

Single Writer



Use Case: Data collection, e.g. from remote sequencer sent to home. Limited scope DR.

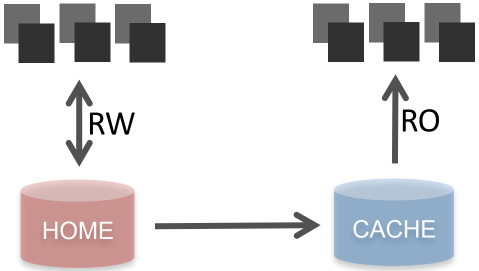
Local Update



Use Case: SW development. After 1st retrieval cache diverges from home. Storage migration.

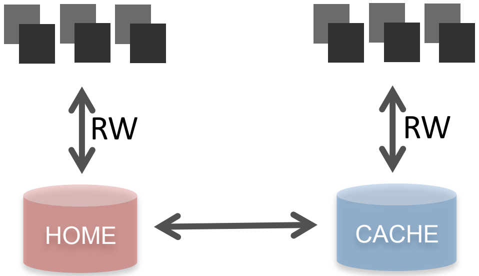
- AFM is an asynchronous, cross cluster utility
- File data is kept consistent in some way between the “cache” and the “home” fileset
- Home does not know cache exists, cache does all the work (checking home for changes, sending updates to home)
- Four Modes:
  - Single Write
  - Read Only
  - Local Update
  - Independent Writer

Read Only



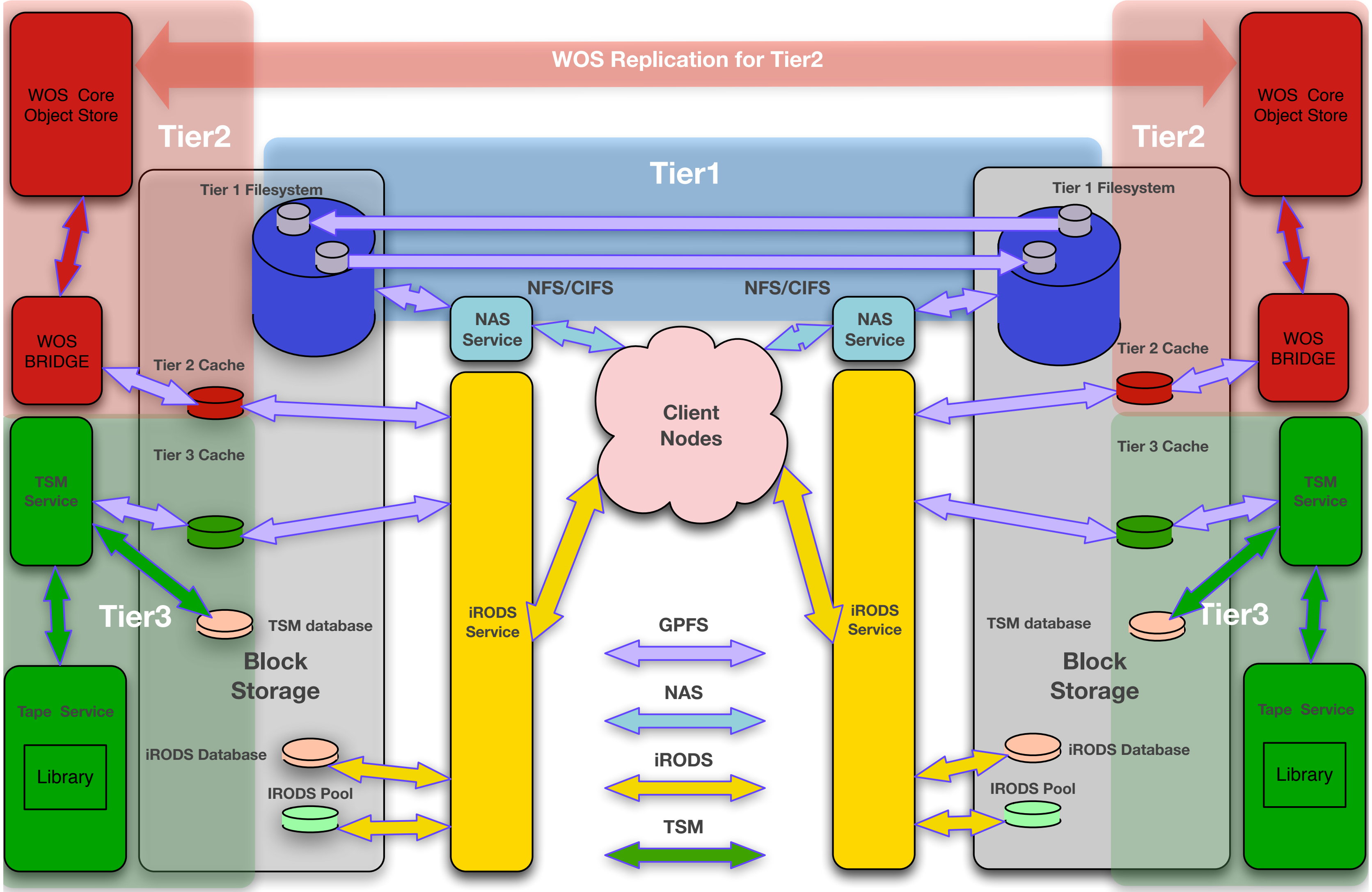
Use Case: Cache accelerates read access to a remote site

Independent Writer



Use Case: Pseudo-shared namespace. Central site data collection.

Imperial College



AFM over NFS

- Two Sites ~30Km apart.
- 10Gb WAN link
- Network latency about 1.3ms
- 4 NFS servers @ Slough DC
- 4 Gateway Servers @ South Kensington
- Single writer Caches on South Ken for "Homes" in Slough
- No NFS tuning as yet

Why Not AFM over GPFS

- NFS is simple and easy to set up.
- Networking is easier to separate.
- We didn’t have 4.1 at the time.
- Want complete independence between sites and Multicluster doesn’t seem to be as “independent” as it might be.
- Don’t need Multicluster parallelism as we will have a large number of AFM relationships so multiple NFS streams will saturate the link.

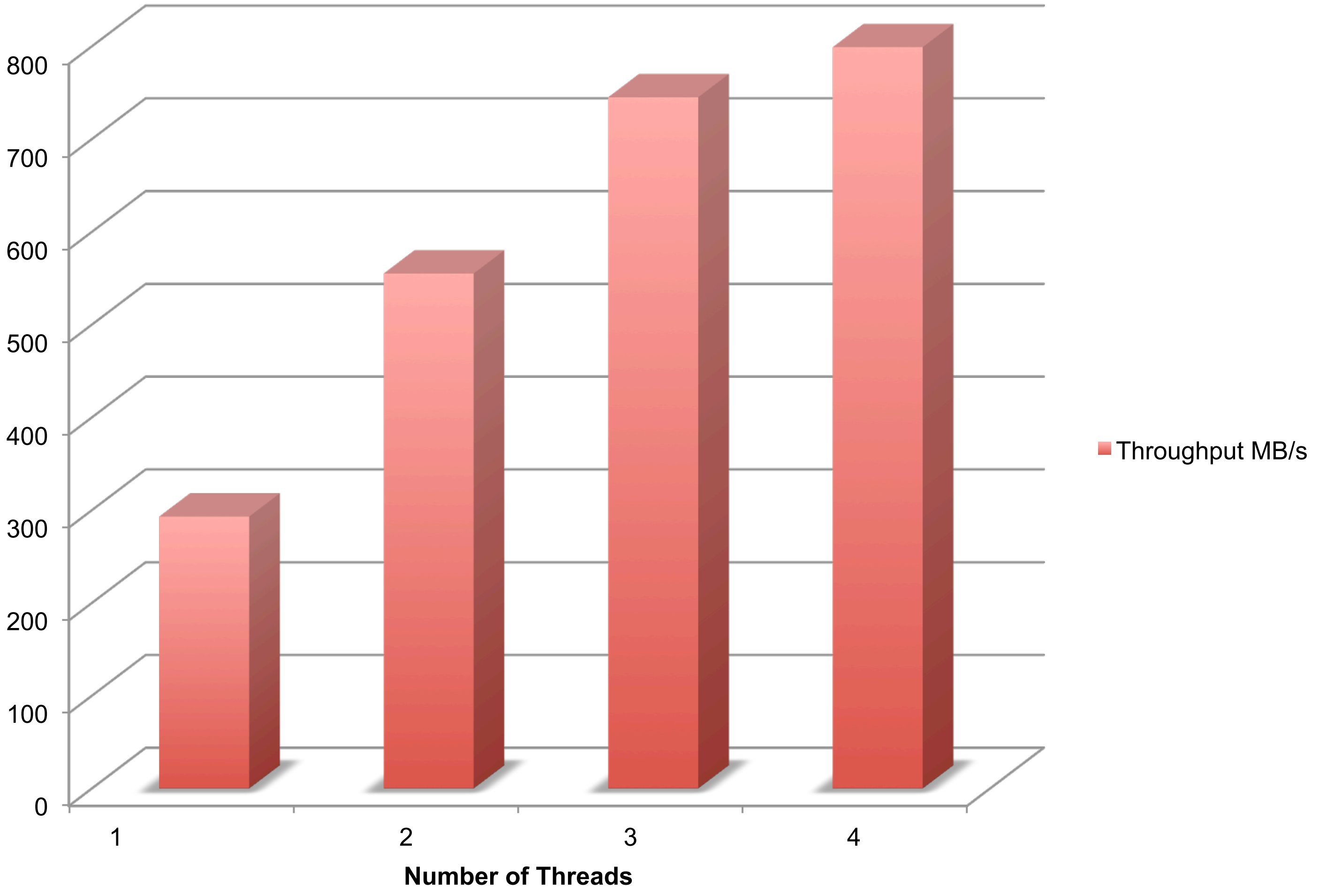
AFM over NFS

- Starting to hit the link throughput limit
- Not quite as fast as possible as assignment to gateway nodes is a bit arbitrary.
- Doesn’t seem to like 2 of the Gateway servers.

```
Fileset Name    Fileset Target                                Cache State     Gateway Node    Queue Length   Queue numExec
------------    --------------                                -------------   ------------    ------------   -------------
lollipop        nfs://mbslafm/SL-Tier1/lollipop               Active          mbsknas02-ib    0              1680053
BSS             nfs://mbslafm/SL-Tier1/BSS                    Active          mbsknas04-ib    0              728071
test1           nfs://mbslafm/SL-Tier1/test1                  Active          mbsknas02-ib    0              90021
NPC             nfs://mbslafm/SL-Tier1/NPC-SL                 Active          mbsknas02-ib    0              542825694
```

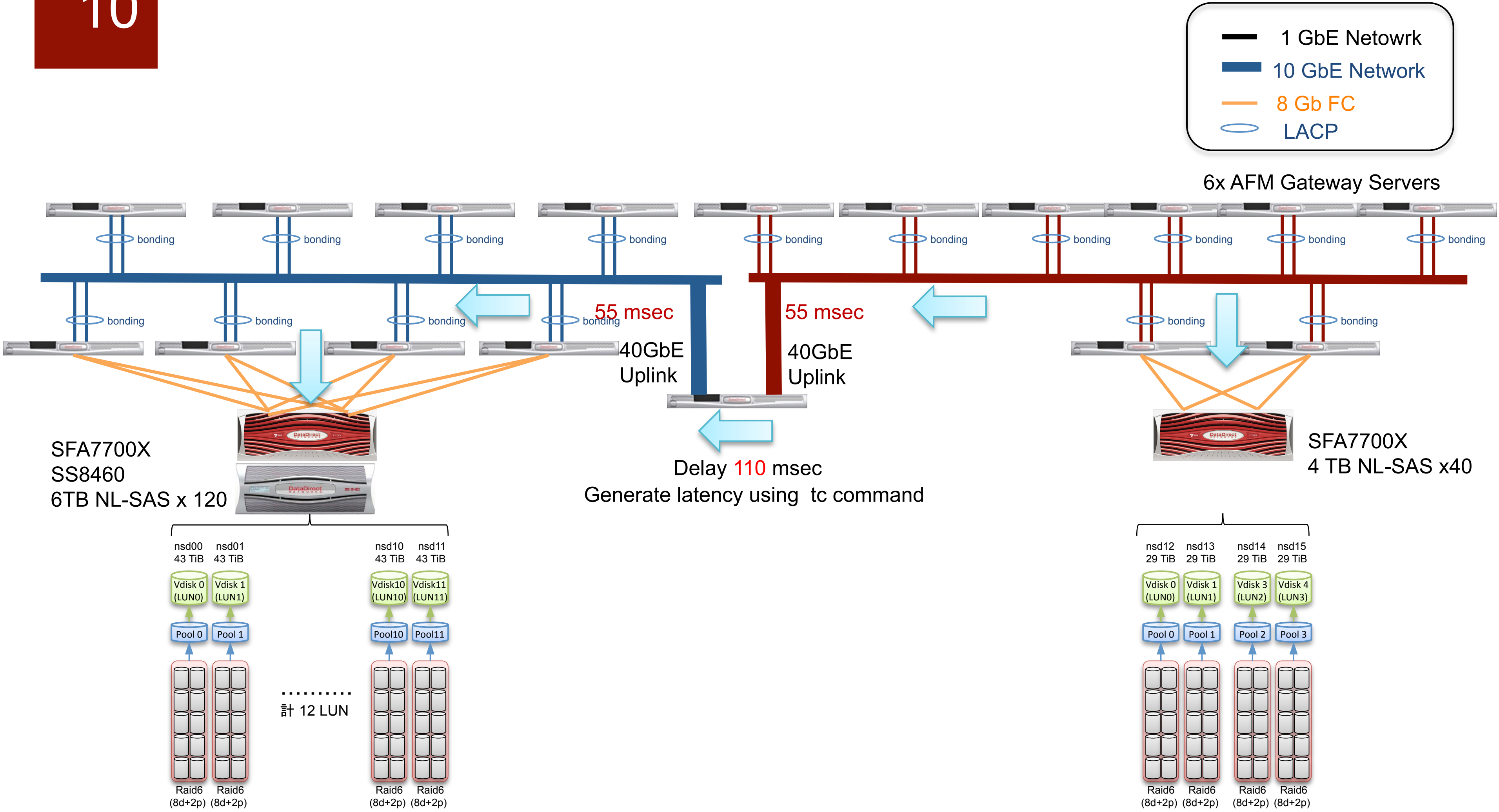
- Recently hit a memory limit with “afmHardMemThreshold”

AFM Performance for NFS over 10Gb WAN



10

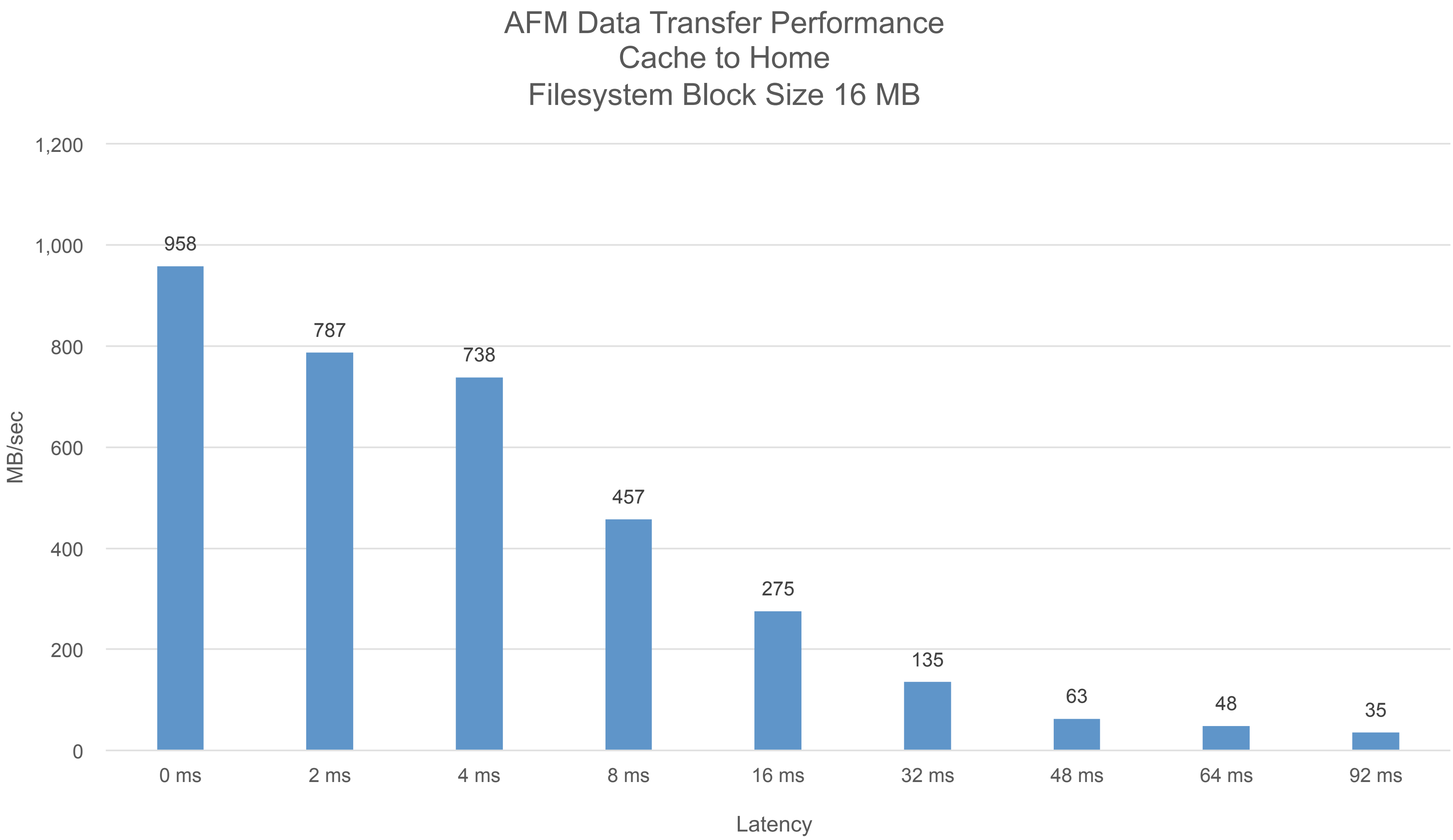
AFM – GPFS Multi-cluster Testing



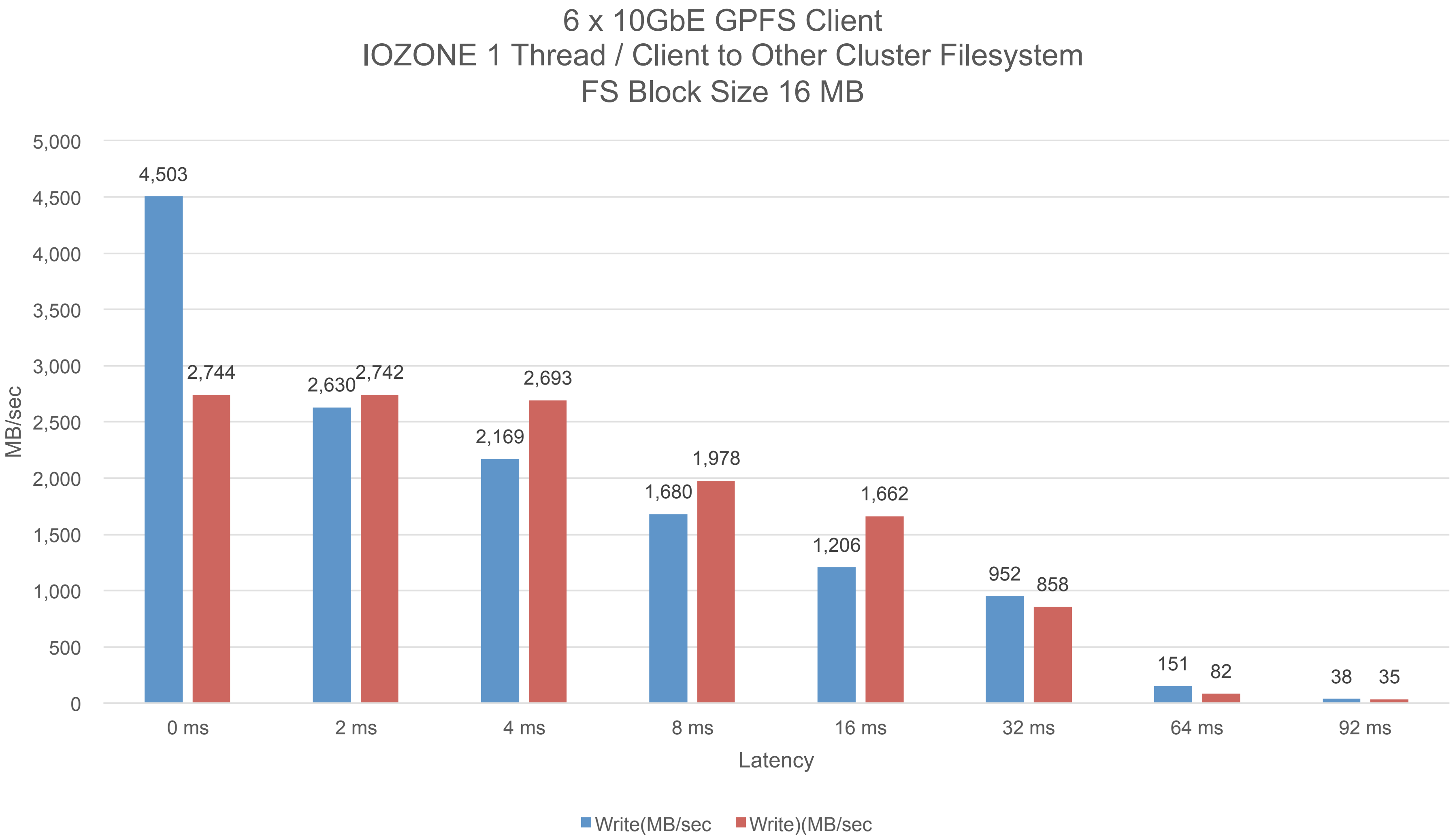
Home

Cache

AFM data transfer performance Cache to Home (40GbE Uplink)



Multi-Cluster Iozone performance (40GbE Uplink)



DDN STORAGE

AFM - No buffer space available

AFM recovery failing with E_NOBUFS usually means that gateway node memory usage crossed the **afmHardMemThreshold** config value.

Run the following command to know the current AFM queue memory usage on gateway node:

```
mmfsadm dump afm | grep QMem
```

Try increasing the memory and verify if recovery progresses.

```
mmchconfig afmHardMemThreshold=10G -i
```

AFM - No buffer space available

"On average each message queued at gateway node takes around 350 bytes of memory.

The mmafmctl getstate command provides the number of messages in queue.

When queue memory usage is approaches afmHardMemThreshold, the gateway node starts flushing the queue without waiting for async delay. (default 15 seconds)

Current queue memory usage (approximately) = number of messages in queue * 350 bytes. – for 5GB that's ~16 million files.

Sometimes queue memory usage keeps growing because replication has stopped (Unmounted or disconnected)

When afmHardMemThreshold is reached, queues are dropped.

In this case AFM will run recovery on next fileset access.

Questions?

vcornell@ddn.com

Thank You!

Keep in touch with us

Team-jpsales@ddn.com

@ddn_limitless

company/datadirect-networks

102-0081
東京都千代田区四番町6-2
東急番町ビル 8F

TEL:03-3261-9101
FAX：03-3261-9140

DDN STORAGE

Imperial Config

```
[57]root@mbskwb01 /root>
mmlsconfig
Configuration data for cluster MedBio-SKen.mbskgs01-ib:
-----------------------------------------------------
clusterName MedBio-SKen.mbskgs01-ib
clusterId 2175721959885733063
dmapiFileHandleSize 32
verbsRdma enable
maxMBpS 22400
healthCheckInterval 20
afmHashVersion 1
minReleaseLevel 4.1.0.4
cipherList AUTHONLY
subnets 10.0.0.0
verbsPorts mlx4_1/1 mlx4_1/2
[mbskicat01-ib,mbskirods01-ib]
verbsPorts mlx4_0/1
[common]
worker1Threads 512
maxFilesToCache 64000
maxStatCache 128000
nsdMinWorkerThreads 64
nsdMaxWorkerThreads 1280
nsdThreadsPerDisk 16
nsdThreadsPerQueue 6
nsdSmallThreadRatio 3
maxInodeDeallocPrefetch 32
stealAggressiveThreshold 6
prefetchThreads 200
flushedDataTarget 1000
flushedInodeTarget 1000
maxblocksize 16M
autoload yes
pagepool 8G
[nas]
pagepool 32G
[common]
adminMode central

File systems in cluster MedBio-SKen.mbskgs01-ib:
----------------------------------------------
/dev/SK-Tier1
/dev/SK-Tier3
/dev/test
```

```
# mmlscluster

GPFS cluster information
========================
  GPFS cluster name:         Home.gs01
  GPFS cluster id:           7808808277142161757
  GPFS UID domain:           Home.gs01
  Remote shell command:      /usr/bin/ssh
  Remote file copy command:  /usr/bin/scp
  Repository type:           CCR

 Node  Daemon node name  IP address    Admin node name  Designation
--------------------------------------------------------------------
   1   gs01              10.10.10.141  gs01             quorum-manager-gateway
   2   gs02              10.10.10.142  gs02             quorum-manager-gateway
   5   gs05              10.10.10.145  gs05             quorum-manager-gateway
   6   gs06              10.10.10.146  gs06             quorum-manager-gateway
   7   gs07              10.10.10.147  gs07             gateway
  10   gs08              10.10.10.148  gs08             gateway
  11   gs09              10.10.10.149  gs09             gateway
  12   gs10              10.10.10.150  gs10             gateway
```

```
# mmlscluster

GPFS cluster information
========================
  GPFS cluster name:         Cache.gs03
  GPFS cluster id:           10590991901404442788
  GPFS UID domain:           Cache.gs03
  Remote shell command:      /usr/bin/ssh
  Remote file copy command:  /usr/bin/scp
  Repository type:           CCR

 Node  Daemon node name  IP address    Admin node name  Designation
--------------------------------------------------------------------
   1   gs03              10.10.10.143  gs03             quorum-manager-gateway
   2   gs04              10.10.10.144  gs04             quorum-manager-gateway
   3   r21-gs            10.10.10.196  r21-gs           gateway
   4   r22-gs            10.10.10.197  r22-gs           gateway
   5   r23-gs            10.10.10.198  r23-gs           gateway
   6   r24-gs            10.10.10.199  r24-gs           gateway
   7   r25-gs            10.10.10.200  r25-gs           gateway
   8   r26-gs            10.10.10.201  r26-gs           gateway
```

```
Configuration data for cluster Home.gs01:
-----------------------------------------
dmapiFileHandleSize 32
minReleaseLevel 4.1.1.0
ccrEnabled yes
autoload yes
cnfsNFSDprocs 256
flushedDataTarget 1024
flushedInodeTarget 1024
logBufferCount 20
logWrapThreads 16
maxBufferCleaners 1024
maxFileCleaners 1024
maxFilesToCache 12000
maxGeneralThreads 1280
maxInodeDeallocPrefetch 128
maxMBpS 10800
maxStatCache 512
maxReceiverThreads 32
nsdbufspace 50
nsdMaxWorkerThreads 1024
nsdMinWorkerThreads 1024
nsdThreadsPerDisk 16
nsdThreadsPerQueue 16
prefetchPct 60
prefetchThreads 288
scatterBufferSize 262144
worker1Threads 1024
worker3Threads 32
tiebreakerDisks nsd00
maxblocksize 16384K
pagepool 32G
clusterName Home.gs01
clusterId 7808808277142161757
cipherList AUTHONLY
adminMode central
```

```
Configuration data for cluster Cache.gs03:
------------------------------------------
dmapiFileHandleSize 32
minReleaseLevel 4.1.1.0
ccrEnabled yes
autoload yes
cnfsNFSDprocs 256
flushedDataTarget 1024
flushedInodeTarget 1024
logBufferCount 20
logWrapThreads 16
maxBufferCleaners 1024
maxFileCleaners 1024
maxFilesToCache 12000
maxGeneralThreads 1280
maxInodeDeallocPrefetch 128
maxMBpS 10800
maxStatCache 512
maxReceiverThreads 32
nsdbufspace 50
nsdMaxWorkerThreads 1024
nsdMinWorkerThreads 1024
nsdThreadsPerDisk 16
nsdThreadsPerQueue 16
pagepool 32G
prefetchPct 60
prefetchThreads 288
scatterBufferSize 262144
worker1Threads 1024
worker3Threads 32
maxblocksize 16M
tiebreakerDisks nsd12
clusterName Cache.gs03
clusterId 10590991901404442788
cipherList AUTHONLY
adminMode central
```

```
# mmlsfs gpfs
flag                value                    description
------------------- ------------------------
-----------------------------------
 -f                 524288                   Minimum fragment size in bytes
 -i                 4096                     Inode size in bytes
 -I                 32768                    Indirect block size in bytes
 -m                 1                        Default number of metadata replicas
 -M                 2                        Maximum number of metadata replicas
 -r                 1                        Default number of data replicas
 -R                 2                        Maximum number of data replicas
 -j                 cluster                  Block allocation type
 -D                 nfs4                     File locking semantics in effect
 -k                 all                      ACL semantics in effect
 -n                 8                        Estimated number of nodes that will mount file system
 -B                 16777216                 Block size
 -Q                 user;group;fileset       Quotas accounting enabled
                    user;group;fileset       Quotas enforced
                    none                     Default quotas enabled
 --perfileset-quota Yes                      Per-fileset quota enforcement
 --filesetdf        Yes                      Fileset df enabled?
 -V                 14.23 (4.1.1.0)          File system version
 --create-time      Wed Oct 28 13:24:31 2015 File system creation time
 -z                 No                       Is DMAPI enabled?
 -L                 16777216                 Logfile size
 -E                 Yes                      Exact mtime mount option
 -S                 No                       Suppress atime mount option
 -K                 whenpossible             Strict replica allocation option
 --fastea           Yes                      Fast external attributes enabled?
 --encryption       No                       Encryption enabled?
 --inode-limit      134422528                Maximum number of inodes in all inode spaces
 --log-replicas     0                        Number of log replicas
 --is4KAligned      Yes                      is4KAligned?
 --rapid-repair     Yes                      rapidRepair enabled?
 --write-cache-threshold 0                   HAWC Threshold (max 65536)
 -P                 system                   Disk storage pools in file system
 -d                 nsd00;nsd01;nsd02;nsd03;nsd04;nsd05;nsd06;nsd07;nsd08;nsd09;nsd10;nsd11
Disks in file system
 -A                 yes                      Automatic mount option
 -o                 none                     Additional mount options
 -T                 /gpfs                    Default mount point
 --mount-priority   0                        Mount priority
```

```
# mmlsfs cache
flag                value                    description
------------------- ------------------------
-----------------------------------
 -f                 524288                   Minimum fragment size in bytes
 -i                 4096                     Inode size in bytes
 -I                 32768                    Indirect block size in bytes
 -m                 1                        Default number of metadata replicas
 -M                 2                        Maximum number of metadata replicas
 -r                 1                        Default number of data replicas
 -R                 2                        Maximum number of data replicas
 -j                 cluster                  Block allocation type
 -D                 nfs4                     File locking semantics in effect
 -k                 all                      ACL semantics in effect
 -n                 8                        Estimated number of nodes that will mount file system
 -B                 16777216                 Block size
 -Q                 user;group;fileset       Quotas accounting enabled
                    user;group;fileset       Quotas enforced
                    none                     Default quotas enabled
 --perfileset-quota yes                      Per-fileset quota enforcement
 --filesetdf        yes                      Fileset df enabled?
 -V                 14.23 (4.1.1.0)          File system version
 --create-time      Mon Nov  9 14:43:41 2015 File system creation time
 -z                 no                       Is DMAPI enabled?
 -L                 16777216                 Logfile size
 -E                 yes                      Exact mtime mount option
 -S                 no                       Suppress atime mount option
 -K                 whenpossible             Strict replica allocation option
 --fastea           yes                      Fast external attributes enabled?
 --encryption       no                       Encryption enabled?
 --inode-limit      120078336                Maximum number of inodes in all inode spaces
 --log-replicas     0                        Number of log replicas
 --is4KAligned      yes                      is4KAligned?
 --rapid-repair     yes                      rapidRepair enabled?
 --write-cache-threshold 0                   HAWC Threshold (max 65536)
 -P                 system                   Disk storage pools in file system
 -d                 nsd12;nsd13;nsd14;nsd15  Disks in file system
 -A                 yes                      Automatic mount option
 -o                 none                     Additional mount options
 -T                 /cache                   Default mount point
 --mount-priority   0
```

```
# mmlsfileset cache iw -X
Filesets in file system 'cache':

Attributes for fileset iw:
===========================
Status                                    Linked
Path                                      /cache/iw
Id                                       1
Root inode                                  524291
Parent Id                                  0
Created                                    Mon Nov  9 23:16:32 2015
Comment
Inode space                                  1
Maximum number of inodes                    102400
Allocated inodes                             102400
Permission change flag                       chmodAndSetacl
IAM mode                                     off
afm-associated                               Yes
Target                                     gpfs:///gpfs/iwhome
Mode                                       independent-writer
File Lookup Refresh Interval                30 (default)
File Open Refresh Interval                  30 (default)
Dir Lookup Refresh Interval                 60 (default)
Dir Open Refresh Interval                   60 (default)
Async Delay                                 15 (default)
Last pSnapId                                0
Display Home Snapshots                         no
Number of Read Threads per Gateway      64
Number of Gateway Flush Threads         1024
Prefetch Threshold                          0 (default)
Eviction Enabled                            yes (default)
Number of Write Threads per Gateway     64
```

```
# /etc/sysctl.conf
net.core.netdev_max_backlog = 250000
net.core.rmem_max = 16777216
net.core.wmem_max = 16777216
net.core.rmem_default=16777216
net.core.wmem_default=16777216
net.core.optmem_max=16777216
net.ipv4.tcp_mem=16777216 16777216 16777216
net.ipv4.tcp_rmem = 4096 87380 16777216
net.ipv4.tcp_wmem = 4096 65536 16777216
net.ipv4.tcp_timestamps=0
net.ipv4.tcp_sack=1
net.ipv4.tcp_fack=1
net.ipv4.tcp_window_scaling=1
net.ipv4.tcp_low_latency=0
net.ipv4.tcp_moderate_rcvbuf=0
vm.swappiness=60
vm.dirty_expire_centisecs=1000
vm.dirty_writeback_centisecs=500
vm.dirty_background_ratio=5
vm.dirty_ratio=80
```