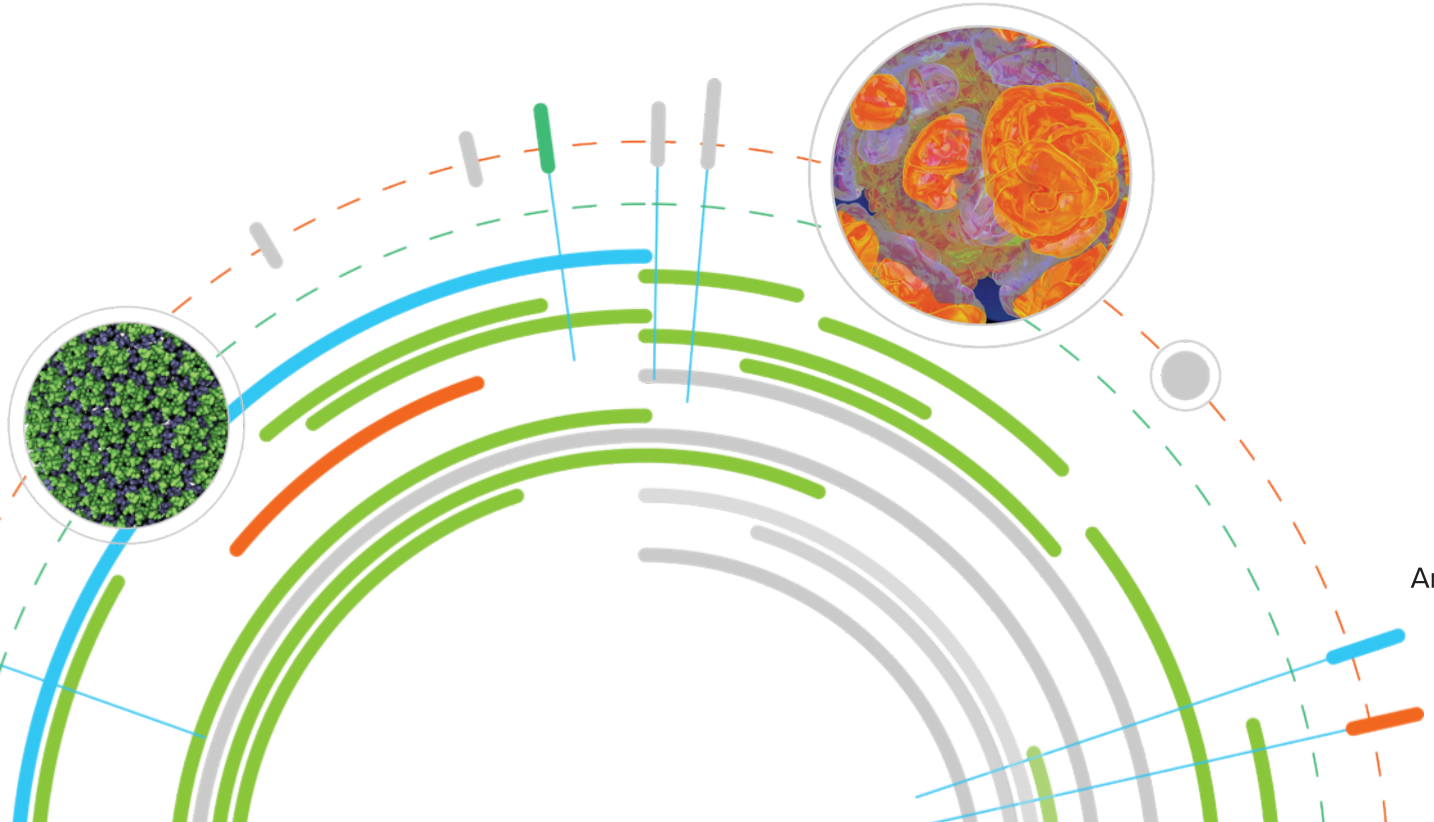
Implementing GPFS AFM As A Burst Buffer



Argonne **Leadership Computing** Facility

Argonne NATIONAL LABORATORY

Agenda

- What is a Burst Buffer vs. tiered storage
- ALCF Operational File System Configuration
- Objectives of the ALCF Burst Buffer Implementation
- Overview of the ALCF Burst Buffer Plan
- Description of the GPFS ESS System
- Overview of GPFS AFM
- State of the Implementation
- Some Implementation Details
- Challenges
- Future Steps

What is a Burst Buffer / Tiered Storage

- Function of a Burst Buffer
  - Absorb spikes in I/O demands
    - Checkpoints one example
  - Delivers a higher level of performance then the underlying PFS
- Possible Locations of Burst Buffers
  - In compute nodes
  - In I/O forwarding agent nodes (example: BG/Q ION nodes)
  - Inserted as a tier between compute and PFS (disk or SSD or RAM based)

There is a bit of a “blurring” between Burst Buffer and tiered storage

The plan at ALCF is to use a “burst buffer” inserted in-between the compute I/O forwarders and the productions PFSs.

This development of this plan was a collaboration between ALCF and IBM GPFS development.

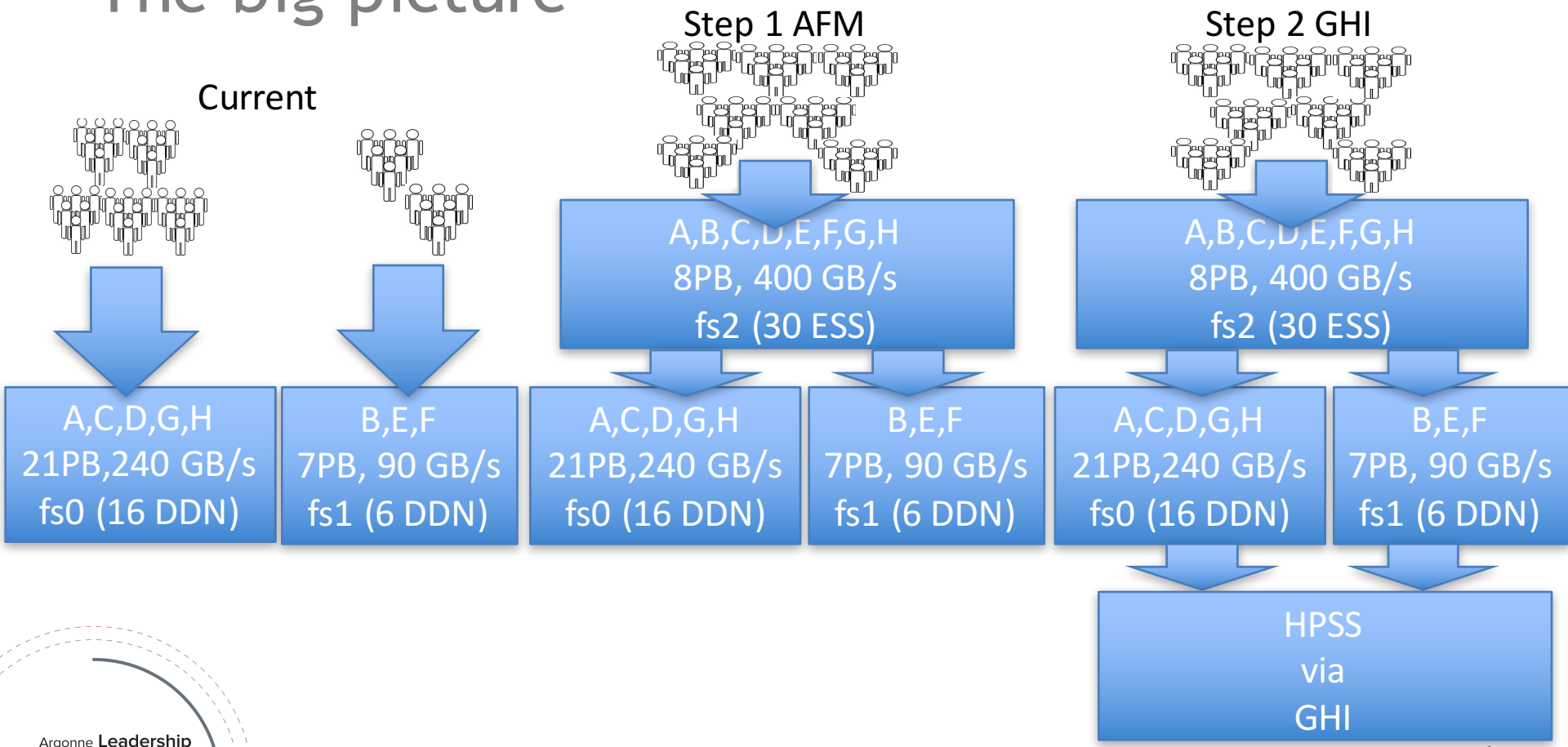
ALCF Operational File System Configuration

- There are three main production level GPFS file systems:
  - mira-fs0 – 19 PiB
  - mira-fs1 - 7 PiB
  - mira-home – 1.1 PiB
- These are based on DDN 12K-E (Embedded NSD servers) storage running GPFS 3.5.0.31.
- These file systems are mounted to all the BG IO servers, a Cray visualization cluster, Globus DTN nodes and HPSS data movers.
- symlinks to the project filesets are used to mask the underlying file system from the users. /project contains links to mira-fs0 and mira-fs1 filesets.

Objectives of the ALCF Burst Buffer

- Relives the asymmetry of performance between mira-fs0 and mira-fs1 for users.
  - Cache misses might cause some run to run variations but all users are equally subject to this and given our cache size our assumption is that this will be fairly rare.
- Minimal user / admin intervention required
  - Ideally policy will do the right things at the right times.
  - Should eliminate any manual copying of files as the cache->home synchronization is all handled by the file system
- Better overall experience for both users and admins

The big picture



What is GPFS AFM

- Active File Management
- In our configuration both the cache and home file systems are GPFS.
- AFM creates a relationship between an Independent Writer (IW) fileset in cache and a project fileset in a home cluster.
- When data is written to cache fileset it is sync'ed to its home fileset
- The cache fileset shows all files in the home fileset even if the home file is not resident in cache.
- If a file not in cache is opened and read from cache (beyond it's first few blocks) it is copied into the cache and becomes cache resident.
- Files may be evicted from cache as they also exist in the home file system.
- Fundamental assumption: Our file systems are big enough that recalls will be rare.

ESS Cluster

- 30 ESS p-Series S822L node pairs. Each node pair has access to 4 disk enclosures GL-4 (DCS3700 - 4U-60 drives capacity).
- GPFS Native RAID (GNR) is used to manage disks. Recovery groups are defined, NSDs are defined on GNR virtual disks.
- Total of ~8.5 PiB in cache file system.
- Based on ESS 3.5.x (service pack 3.5.3 / GPFS 4.1.1-7 planned soon).
- Each server has a pair of IB QDR links.
- mira-fs0/mira-fs1 file systems are mounted to all nodes in the ESS cluster using the GPFS remote mount feature.
- Only the ESS file system fs2 is mounted to the BG ION nodes.

Implementation Overview

- Current State: Internal testing
  - 2 ESS node pairs deleted from ESS and used to form a test “home” cluster. This file system is mounted to all remaining ESS nodes.
- Using AFM mode that allows for a GPFS home file system.
- mira-fs0 and mira-fs1 will be remotely mounted to ESS system
  - mira-home will not be AFM managed
- Home file system defined as: gpfs:///
  - Null server list which signifies remotely mounted to cache cluster
- Cache File sets created with --afm flag.
- Home filesets need to have mmafmconfig run against them to create the expected AFM directory structures for recovery processing. Failure to do so resulted in AFM gateway node panics.

Commands to create home fileset

- Home:

```
[root@ess29-srv1 ~]# mmcrfileset fm-fs1 DemoFileSet
Fileset DemoFileSet created with id 21 root inode 319360.

 [root@ess29-srv1 ~]# mmlinkfileset fm-fs1 DemoFileSet -J /gpfs/fm-fs1/projects/DemoFileSet
Fileset DemoFileSet linked at /gpfs/fm-fs1/projects/DemoFileSet

[root@ess29-srv1 ~]# mmafmconfig enable /gpfs/fm-fs1/projects/DemoFileSet/

[root@ess29-srv1 ~]# ls -la /gpfs/fm-fs1/projects/DemoFileSet/
total 64
drwx------.  3 root root    512 Jun   9 23:26 .
drwxr-xr-x. 24 root root 32768 Jun   9 23:23 ..
drwxr-xr-x.  3 root root    512 Jun   9 23:26 .afm
[root@ess29-srv1 ~]#
```

Commands to create cache/home relationship

- On the cache cluster:

```
[root@ess-mgmt1 gmcpheet]# mmcrfileset ess-fs0 DemoFileSet -p afmMode=iw -p afmTarget=gpfs:///gpfs/fm-fs1/projects/DemoFileSet --inode-space=new
Fileset DemoFileSet created with id 41 root inode 44040195.

[root@ess-mgmt1 gmcpheet]# mmlinkfileset ess-fs0 DemoFileSet -J /gpfs/ess-fs0/projects/DemoFileSet
Fileset DemoFileSet linked at /gpfs/ess-fs0/projects/DemoFileSet

[root@ess-mgmt1 DemoFileSet]# df -h|grep gpfs
/dev/fm-fs1                577T  7.1G  577T   1% /gpfs/fm-fs1
/dev/ess-fs0               7.9P  873G  7.9P   1% /gpfs/ess-fs0
```

Add data to cache fileset

```
[root@ess-mgmt1 DemoFileSet]# pwd
/gpfs/ess-fs0/projects/DemoFileSet

[root@ess-mgmt1 DemoFileSet]# dd if=/dev/zero of=foo count=100 bs=1048576
100+0 records in
100+0 records out
104857600 bytes (105 MB) copied, 0.0646135 s, 1.6 GB/s

[root@ess-mgmt1 DemoFileSet]# ls -l foo
-rw-r--r--. 1 root root 104857600 Jun  9 23:35 foo
[root@ess-mgmt1 DemoFileSet]#
```

View data at home

```
[root@ess29-srv1 ~]# cd /gpfs/fm-fs1/projects/DemoFileSet/
[root@ess29-srv1 DemoFileSet]# ls -l
total 102400
-rw-r--r--. 1 root root 104857600 Jun  9 23:36 foo
[root@ess29-srv1 DemoFileSet]#
```

View State of filesets on Cache

```
[root@ess-mgmt1 DemoFileSet]# mmafmctl ess-fs0 getstate
Fileset Name    Fileset Target                              Cache State    Gateway Node    Queue Length   Queue numExec
------------    --------------                              -------------  ------------    ------------   -------------
DemoFileSet     gpfs:///gpfs/fm-fs1/projects/DemoFileSet    Active         ess10-srv1             0          109

[root@ess-mgmt1 DemoFileSet]# mmlsfileset ess-fs0 DemoFileSet -L --afm
Filesets in file system 'ess-fs0':

Attributes for fileset DemoFileSet:
===================================
Status                          Linked
Path                            /gpfs/ess-fs0/projects/DemoFileSet
Id                              41
Root inode                      44040195
Parent Id                       0
Created                         Thu Jun  9 23:31:07 2016
Inode space                     21
Maximum number of inodes        100096
Allocated inodes                100096
Permission change flag          chmodAndSetacl
afm-associated                  Yes
Target                          gpfs:///gpfs/fm-fs1/projects/DemoFileSet
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 Gateway Flush Threads 4
Prefetch Threshold              0 (default)
Eviction Enabled                yes (default)
[root@ess-mgmt1 DemoFileSet]#
```

Implementation Challenges

- This cluster was originally based on the GSS product which was xSeries (Intel) based.
  - Cluster was migrated from xSeries to pSeries (ESS) to address support concerns.
- Kernel Panics in mpt2sas kernel module
  - Redhat bug: 1259907 (bug) /1318560 (back port to 7.1 zStream)
  - Difficult/elongated PD/recreate path. Up to 3 weeks between recreates.
  - Fixed in RH 7.2 stream but we needed to ask Red Hat for a port to RH 7.1 zStream. This fix was very recently released by Red Hat.
  - As soon as available this will be incorporated in our ESS cluster.
  - Fixed in: kernel-3.10.0-229.33.1.el7

Implementation Challenges

- Quotas:
  - No communication between home and cache WRT to quota settings.
  - Data is sync'ed to home as root and root does not error on over quota
  - Found best to turn off auto-migration on cache filesets.
    - Prevents over-running home cache hard limits

- Bug found after file evicted and subsequently re-read from cache cluster it was not becoming resident again in cache. Efix has been delivered and will be installed as part of the service pack 3.5.3 upgrade.

Future Steps

- Once Blue Gene System is supported on GPFS 4.1.1 then the AFM cluster can move to GPFS 4.2 (ESS 4.x)
- Allow the Cray Theta system access to Mira project data in GPFS via DVS mounts of ESS file system.

Questions

- Any questions, otherwise thank you for your attention.

- Gordon McPheeters
  - gmcpheeters@anl.gov