
CSIRO

Hybrid SSD-HDD Filesystems

Peter Edwards

The original idea

1. Build an XVM volume from concatenated subvolumes where the first and only initial concat would be comprised of RAID protected SSD disks. The mkfs *inode32* option would be used to keep inodes at the front of the filesystem.

2. The 21 million inodes for the main DMF filesystem would be restored with xfsrestore.

3. Extra concats of HDD disks would then be added to the filesystem via "*xvm attach*" and *xfs_growfs*. This adds extra allocation groups, requiring the use of *maxpct* to prevent them from being used for new inodes.

Refinements

- Use of xfsrestore only allocates inodes for currently existing files; new inodes could go anywhere in the 32 bit part of the filesystem, and file data could end up in SSD.

  Instead, a large number (hundred million?) of empty files would be created and later deleted, thereby preallocating most of the SSD space as inodes. This depends on the *ikeep* mount flag being set, which is the default for DMAPI-mounted filesystems.

- SGI suggested the option of being able to change the position of the inode32 boundary.

- The original plan was abandoned in favour of this approach as the requirement that the SSD be at least 1% (the minimum *maxpct* value) of the filesystem was too wasteful.

Benchmarks

Attempted to measure read and write transfer rates with different block-sizes and different number of concurrent streams. Also speed of various types of metadata operations.

Remarkably difficult due to multiple levels of caching. And even if you disable them all, is what you're then measuring relevant to the real world?

The SSD performance (measured separately from the HDDs) was also poor. SGI found that we had a suboptimal volume layout for the SSD (4+2P RAID6 over 5 shelves) and suggested 4+1P+1HS until we got a 6th shelf.

Also, SGI advised that use of an external log made a noticeable difference, whether on SSD or not.

More hardware

It was realised that a much better configuration could be achieved if we bought 8 more FC disks, and an extra shelf. The final configuration is:

Type	Drives	Configuration	Equiv JBOD Capacity
SSD	6 x 73GB	4+2P, divided into many LUNs	4 x 73GB
FC	82 x 450GB	8(8+2P) + 2HS	64 x 450GB
SATA	60 x 1000GB	Originally 4 LUNs of 2(6+1P)+4HS Later changed to 6(8+2P)+0HS	48 x 1000GB

SSD is used for logs and inode concats on 4 filesystems, and just for logs on 4 more.

And less hardware

We were going to connect our Altix 4700 with a limited number of 4Gb/s FC HBAs to the IS4600 via 4Gb/s FC switches.

But with the replacement of the Altix by a UV1000, we realised that if 8 ports on the disk controller were connected directly to the UV, it could all run at 8Gb/s. And free up ports for tapes so we don't need to buy another FC switch.

Setting *ibound*

```
thorax# xvm show -topology vol/datastore
vol/datastore                 0 online,open,accessible
   subvol/datastore/data    56329075712 online,open,accessible
      concat/concat0        56329075712 online,tempname,open,accessible
         slice/meta7s0       134200320 online,open,accessible
         stripe/fc          56194875392 online,open,accessible
            slice/is4600_fc_40s0     7024359680 online,open,accessible
            ...
            slice/is4600_fc_47s0     7024359680 online,open,accessible
   subvol/datastore/log        244480 online,open,accessible
      slice/log7s0             244480 online,open,accessible
```

In the highlighted line above, the 134200320 is the value to be used for mkfs.xfs and /etc/fstab:

- mkfs.xfs -d agsize=134200320s -i size=512 -l logdev=/dev/lxvm/datastore_log /dev/lxvm/datastore
- /dev/lxvm/datastore /hybrid xfs ibound=134200320,logdev=/dev/lxvm/datastore_log 1 0

  (omitting the DMAPI flags)

Performance

Approximate speed-up due to hybrid filesystem:

- dmaudit 6x
- dmscanfs 2-4x
- xfsdump level 0 2.5x
- xfsdump level 9 10x
- xfsdump level 9 (just phase 1&2) 20x

This is not just due to the SSD. We now access the disk array through 8 x 8Gb/s FC ports instead of 4 x 4, the array has more trays, the disks are faster and there are slightly more spindles in use. So we're comparing a system crippled by slow disk with one with multiple speed-ups.

(This slide was added after the original presentation.)

DCM reliability

Initially we were going to set up the SATA as a single RAID6 LUN and use it as our new DCM.

But the change to use 8 interfaces rather than 6 made the use of RAID5 more sensible.

But this left us vulnerable to a second drive failure happening while rebuilding from a first, causing all the DCM files to be lost.

No loss of user data, but a big performance hit for weeks/months while the DCM repopulated.

We now plan on rotating around 4 DCMs each based on a 2(6+1P) LUN. Two failures could still destroy a DCM, but would be less likely, and we would only lose ¼ of our "virtual DCM".

Virtual DCM

We do populate the DCM during migration, but mainly after recalls, which is a bit unusual.

To rotate around the four genuine DCMs, we will:

- write a site-defined migration policy (based on the sample shipped with DMF) to handle population during migration.
- add similar logic to our existing local script which populates on recalls.

A Migration Group is an easier solution, but appears to be inefficient when used in this way. Site policies can give more flexibility, at the cost of extra implementation effort.

Back to a conventional DCM

In the end, we abandoned the “virtual DCM” idea, and implemented a single DCM with a 6(8+2P) + 0 HS configuration, using an external SSD-resident log.

With RAID6, we felt the risk of having no hot spares was acceptable for a filesystem which was only to used as a cache.

(This slide was added after the original presentation.)

References

- An introduction to SSDs, presented by Jeremy Higdon (SGI) to the SGI Users Group: http://hpsc.csiro.au/users/dmfug/shared/SGI_USER_Group/Jeremy_Higdon-SSD.ppt

Thanks

- Initial idea

  David Honey SGI NZ

- Answering many questions, and implementing *ibound*

  Geoffrey Wehrman SGI US

- Assisting with XVM performance problems over many months

  Jeremy Higdon SGI US

- Helping an XVM newbie

  Susheel Gokhale SGI Aust

- Thanks also to Coca-Cola Australia and Youtube for the Mother Energy Drink image used on the cover slide.

Advanced Scientific Computing

Peter Edwards
Systems Support Manager

Email: peter.edwards@csiro.au

www.csiro.au

Thank you

Contact Us
Phone: 1300 363 400 or +61 3 9545 2176
Email: enquiries@csiro.au Web: www.csiro.au

CSIRO