Moving a Datacentre Interstate

Moving a DMF server with 5 PB of data interstate with minimal downtime

Peter Edwards | Sr Systems Administrator

DMF Users Group | February 2016

Old Configuration

In Docklands, Vic

- SGI UV1000
- NetApp RAID (IS4600 SSD, FC & SATA, 2 x IS5500 SAS)
- Copan MAID
- SL8500 with T10kB & C drives
- 5 PB data (x2) on tape

New Configuration

In Hume, ACT

- SGI UV3000 (new)
  - 2 x 56 Gb/s FDR IB and 2 x 40Gb/s ethernet
- Shared DDN RAID (SAS & NL-SAS) (new)

In Black Mountain (Acton), ACT

- Copan MAID (retained)
- Shared SL8500 (new)
- T10kB & C drives (retained)
- T10kD drives(new)
- 4000 tape volumes (retained)

In Clayton, Vic

- Shared SL8500 (new)
- T10kC drives (retained)
- 1300 tape volumes (retained)

Challenges

- Minimise down-time
- Maintain at least 2 copies of migrated files at all times.
- Previous transitions in same building, or (3 times) in same city. This one is interstate, spread over 3 locations.
- Two usable(!) FC connections connecting the 3 sites
- Indeterminate delays while waiting for UV3000 product release
- UV3000 to be installed in APC rack

Challenges (cont'd)

- Sharing non-DMF filesystems via NFS using IB
  - atime not properly supported, so no flushing
- Delegation of tasks to other parts of CSIRO:
  - RAID
  - SL8500's
  - Networking
  - Facilities (power etc)
  - Applications
- Contractors to relocate to ACT the equipment being retained:
  - Copan MAID (SGI & Cope)
  - 12 Tape drives (Oracle & Cope)
  - 4000 Tape volumes (Iron Mountain & Cope)

Fibre Channel

Hume internal (RAID)

8 x 16Gb multi-mode links within the data centre

Hume -> Clayton (tape)

FCIP over 2 x 10GbE single-mode links

Hume -> Black Mountain (tape & MAID)

8 CWDM colours to 8 x 16Gb single-mode links (2 fabrics)

MAID can't be in Hume due to non-standard rack size – hopefully latency won't be an issue

Migration Steps (1)

1. Early preparation
   - Install new T10kD drives in Black Mountain
   - Test tape access to there and to existing drives in Clayton via FC using test system *(see appendix).*
   - Have partitions on shared FC-connected DDN RAID allocated and build XVM filesystems *(see appendix).*

2. Installation
   - Install and configure UV3000 in Hume
   - Rezone RAID partitions and tape drives to it and re-test
   - Set up network access (1 week)
   - Build application packages. (3 weeks)

3. Reconfigure Copan MAID to have a copy of all newly migrated files.

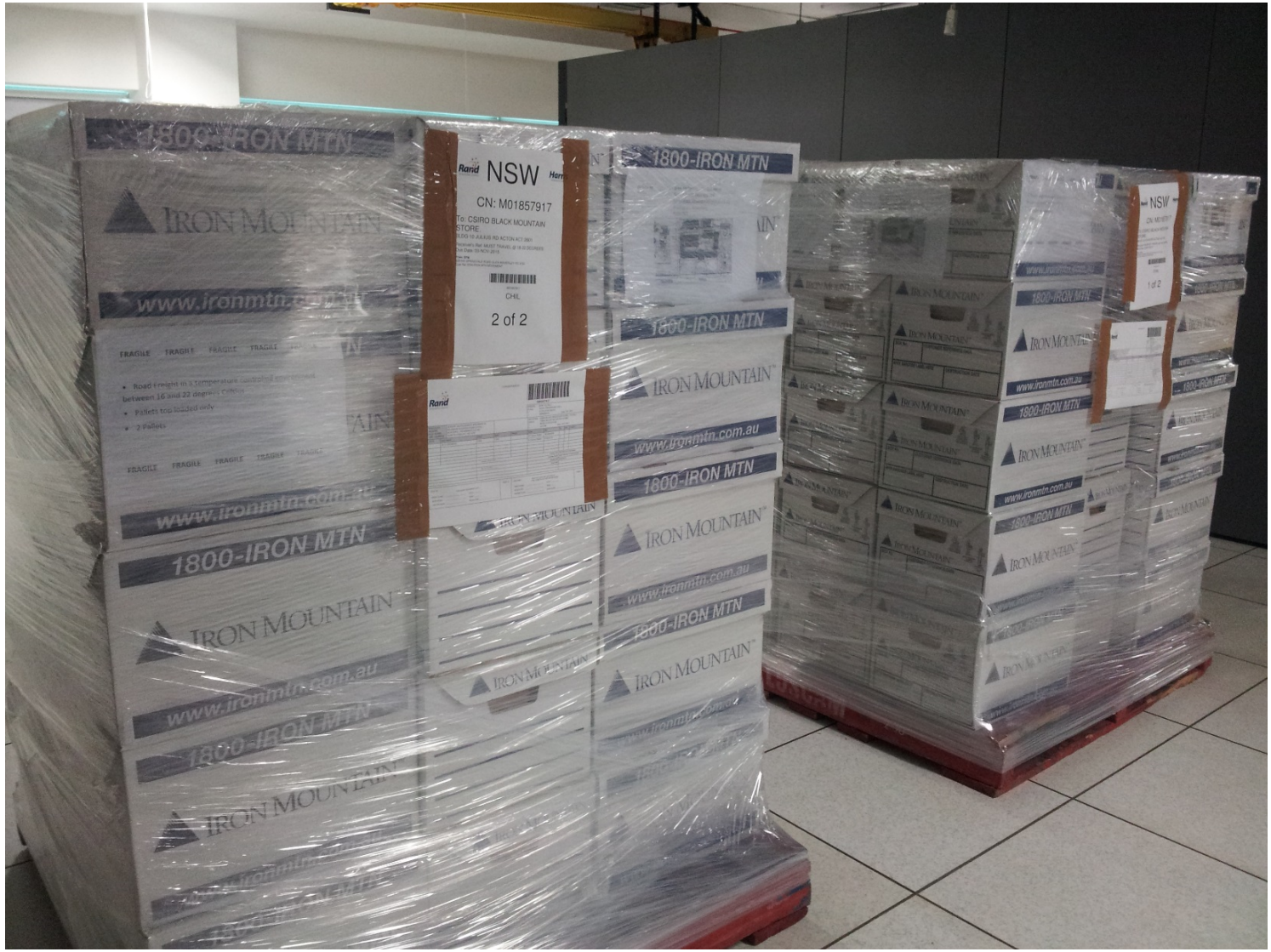
Migration Steps (2)

4. Mark DCM read-only and xfsdump it to multi-reel tape
   - (5 tapes, 48 hours)

5. DMF changes
   - Stop migrating to primary tapes.
   - Mark existing primary tapes to be unavailable and ship them (and DCM dump) to ACT, together with 8 B and 2 C drives, and reinstall. (11 days)
   - Copan change above helps mitigate performance degradation.

6. Running in degraded mode
   - Continue running at Docklands while drives are installed and tested in ACT.
   - Reload DCM from tapes (10 hrs).
   - Transfer non-user filesystems to ACT over the network (6 hrs).

4000 tapes ready to go



Migration Steps (3)

7. Outage starts 18:00 Friday.
   - Workload already drained
   - Transfer DMF DB and user filesystem xfsdumps over network to ACT and reload there. (15 hrs)
   - Change DNS entries.

8. Outage over 22:00 Saturday.
   - New system ready for use after 28 hour outage.

9. Oops, no it's not! Those files migrated during step 6 do have 2 copies but they're both in Docklands and no longer accessible!
   - Visit Docklands on Sunday to collect half a dozen tapes
   - Can’t deliver to Clayton until Monday

Migration Steps (4)

10. Going live
    - Monday 9:00 – the 6 tapes were delivered to Clayton.
    - Ready to go live, when system stops responding.
      Panic, until we find there's been a power failure in ACT.
    - Go live at 11:45
      – total outage 66 hours.

11. Tidy up
    - Secondary tapes moved from Docklands to Clayton.
    - Last 2 T10kC drives moved to ACT.
    - Generate second tape copies for those files which only have one.
    - Ship MAID to ACT.
    - Deinstall remaining Docklands gear.

12. After 10 years of promises, and some failed attempts, we now have migrated data held in two different locations!

Lessons Learned

- Other people's priorities may not align with yours
- Don’t assume that their understanding of a delegated task matches your own
- Interstate transport takes longer then you think
- Semi-trailers can’t do three point turns
- Arrange weekend access to sites
- The unexpected still happens, though planning helps

Appendix - Performance

- NFS
  - The shared ext4 filesystems have been tested 3.9 GB/s over NFS! (and 4.7 GB/s native)
- Tape throughput
  - Interstate: 1 stream - 248 MB/s read and 231 MB/s write
  - Interstate: 2 streams - 198 MB/s read and 228 MB/s write
  - Suburban: 1 stream - 250 MB/s read and 195 MB/s write
  - Suburban: 7 streams - 250 MB/s read and 193 MB/s write

Appendix – XVM filesystems

- Replaced previous bespoke scripts with a general table-driven one supporting:
  - several common XVM topologies:
    - a single slice
    - a stripe of slices
    - a concat of slices
    - a concat of stripes of slices
  - all with internal or external logs
  - all with or without a SSD slice to contain metadata and directories
  - generation of fstab entries, including ibound values for hybrid filesystems

Appendix – XVM filesystems (example)

XVM topology

```
vol/datastore
   subvol/datastore/data
            concat/concat10
               slice/lun105s0
               stripe/stripe_11
                        slice/lun11s0
                        slice/lun12s0
               stripe/stripe_13
                        slice/lun13s0
                        slice/lun14s0
               stripe/stripe_15
                        slice/lun15s0
                        slice/lun16s0
   subvol/datastore/log
            slice/lun115s0
```

Stanza embedded in shell script:

```
m=/datastore
volname[$m]=

ssd_lun[$m]=105
ssd_nags[$m]=32

hdd_luns[$m]='11 - 16'
stripe_width[$m]=2

log_lun[$m]=115

mnt_opt[$m]=ikeep,dmapi,mtpt=$m
```

Thank you

CSIRO IMT Scientific Computing
Peter Edwards
Sr Systems Administrator

t +61 3 9545 2377
e peter.edwards@csiro.au
w https://wiki.csiro.au/display/ASC/Scientific+Computing+Homepage