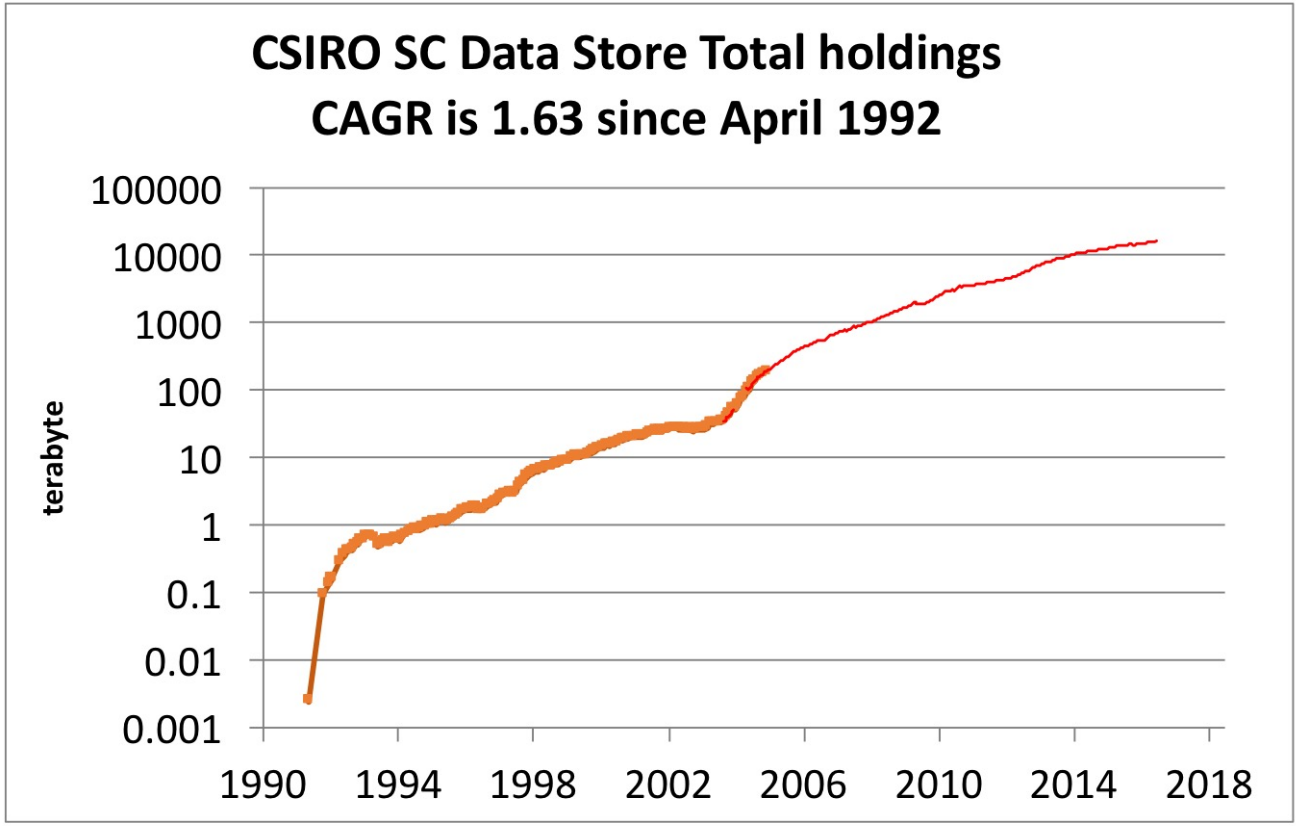
# Silver Anniversary - CSIRO Scientific Computing Data Store

**Robert C. Bell** | CSIRO IMT Scientific Computing

7 February 2017

# Introduction



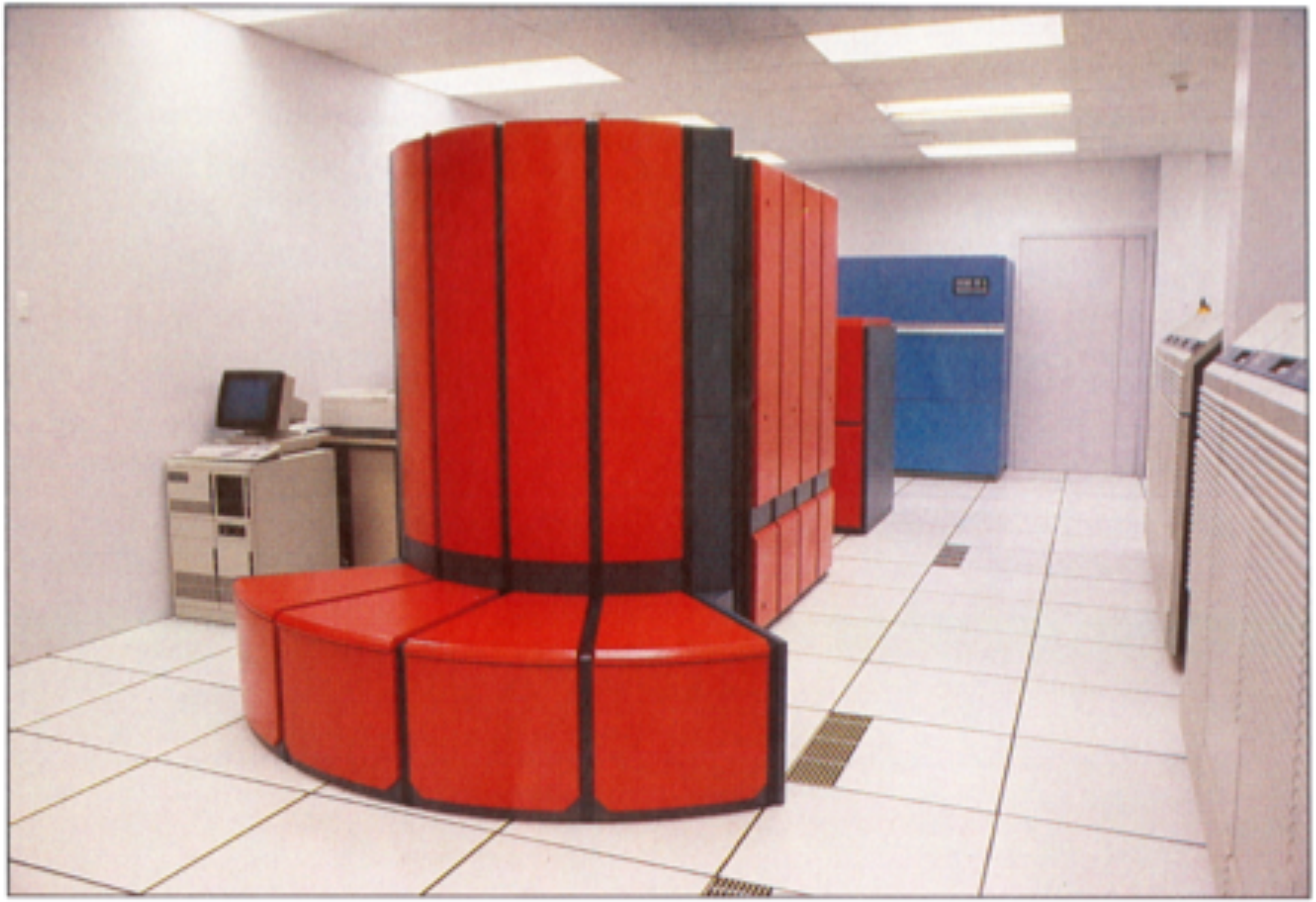
# Introduction

- History
- Technology changes
- Meeting the needs of the users
- Leading the way for other sites
- Continuing value for users
- Lessons learned

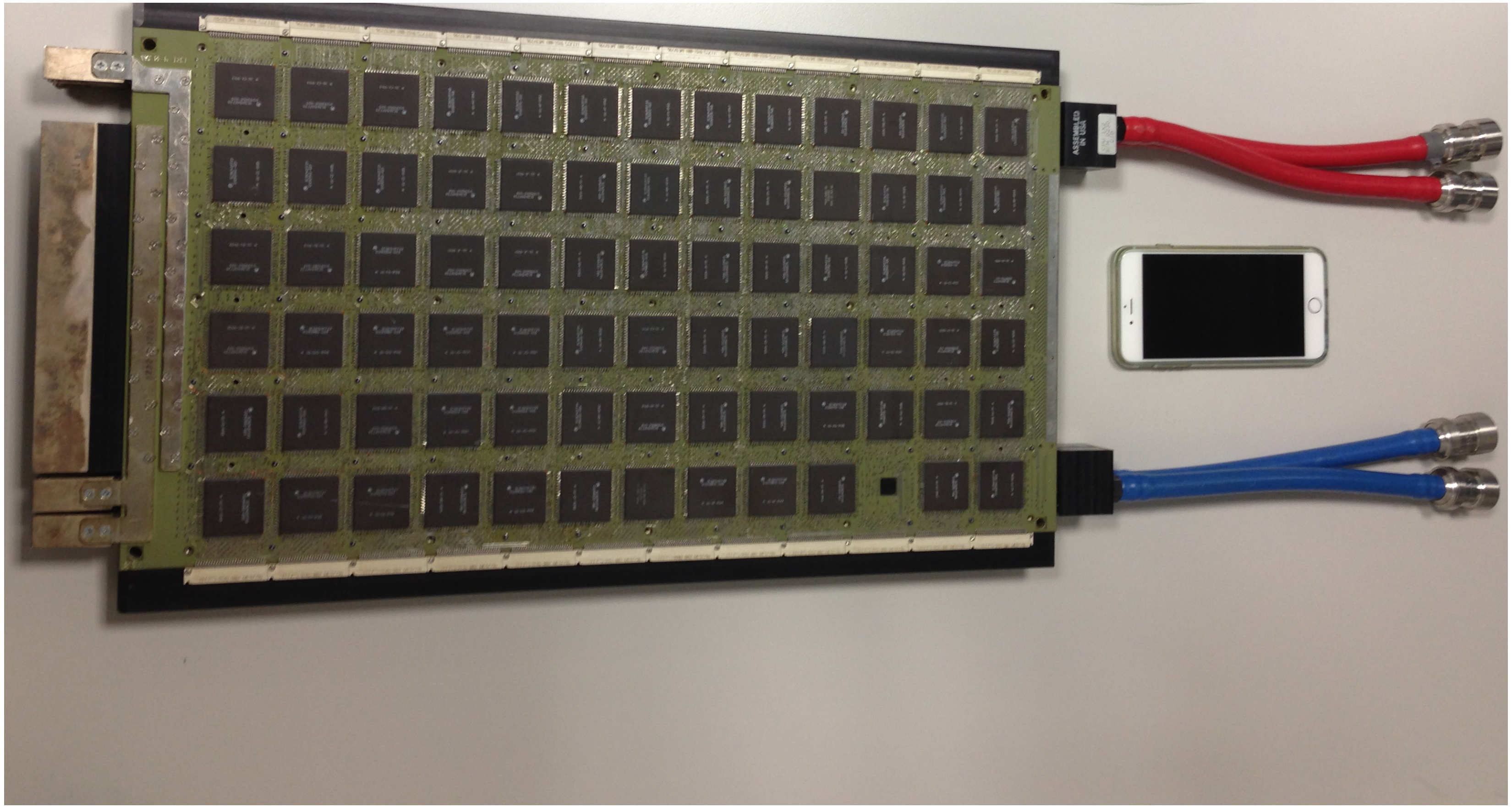
# History

- 14th November 1991
- DMF turned on for the CSIRO home filesystem on the JSF Cray Research Y-MP2/216 (cherax)
- Washing-machine-sized, DD-49 disc units, each holding 1.2 gigabytes with 9.8 megabyte per second transfer rate
- Used manually mounted 3480 tape cartridges (240 Mbyte) using STK tape drives
- Pre-history:

# History



# History



# Pre-History

- CSIRO (Division of Computing Research) had an HSM called the Document Region on its CDC 3600 from the 1960s until it was de-commissioned in 1977: it used operator-mounted 7-track magnetic tape.
- CSIRO (Csironet) had an automated tape store (Braegan/Calcomp ATL) using 6250 bpi 9-track tapes hosted on Fujitsu systems from about 1980 to 1990, with a Terabit File Store built upon this, but without automatic migration.

# History – DMF

- Saved users from having to set up jobs to copy files to tapes – tar commands, pool of tapes, etc.
- Virtually infinite storage
- inode quotas soon applied
- Resilient – two copies, and off-site backup started in early 1992
- Service provider crisis

# Technology changes

- Hosts/sites (7)
  - 1990-1992: Cray Y-MP2/216: LET, Port Melbourne
  - 1992-1997: Cray Y-MP4/364: University of Melbourne, Carlton
  - 1997-2004: Cray J90se: Bureau of Meteorology, 150 Lonsdale St
  - 2004-2008: SGI Altix 3700: Bureau of Meteorology, 700 Collins St
  - 2008-2012: SGI Altix 4700: Bureau of Meteorology, 700 Collins St
  - 2012-2015: SGI UV1000: Bureau of Meteorology, 700 Collins St
  - 2015-: SGI UV3000: Canberra Data Centre
- Discs/space (7)
- DD-49, DD-60, ?, TP9300, TP9700, is4600, ?
  - 1.2 Gbyte to 113 Tbyte (incl. SSD, plus 377 Tbyte cache and 1.8 Pbyte MAID)

# Technology changes

- Tape (11)
  - STK 3840 - 240 Mbyte
  - STK 3490 Timberline - ~ 500 Mbyte
  - STK Redwood - 50 Gbyte
  - STK T9840A - 10 Gbyte
  - STK T9900A - 60 Gbyte
  - STK T9840C - 40 Gbyte
  - STK T9900B - 200 Gbyte
  - STK T10000A - 500 Gbyte
  - STK T10000B - 1000 Gbyte
  - STK T10000C - 5000 Gbyte
  - STK T10000D - 8500 Gbyte

# Technology changes

- Tape libraries (6)
  - 1991-1993: nil
  - 1993-1997: STK 4400
  - 1997-2004: STK Powderhorn
  - 2004-2008: STK Powderhorn
  - 2008-2015: STK SL8500
  - 2015-: two SL8500s

# Technology changes

- Administrators (4)
- 1991-1997: Peter Edwards, Cray Research
- 1997-1999: Virginia Norling, CSIRO
- 1999-2007: Jeroen van den Muyzenberg, CSIRO
- 2008-: Peter Edwards, CSIRO

# Meeting the needs of the users

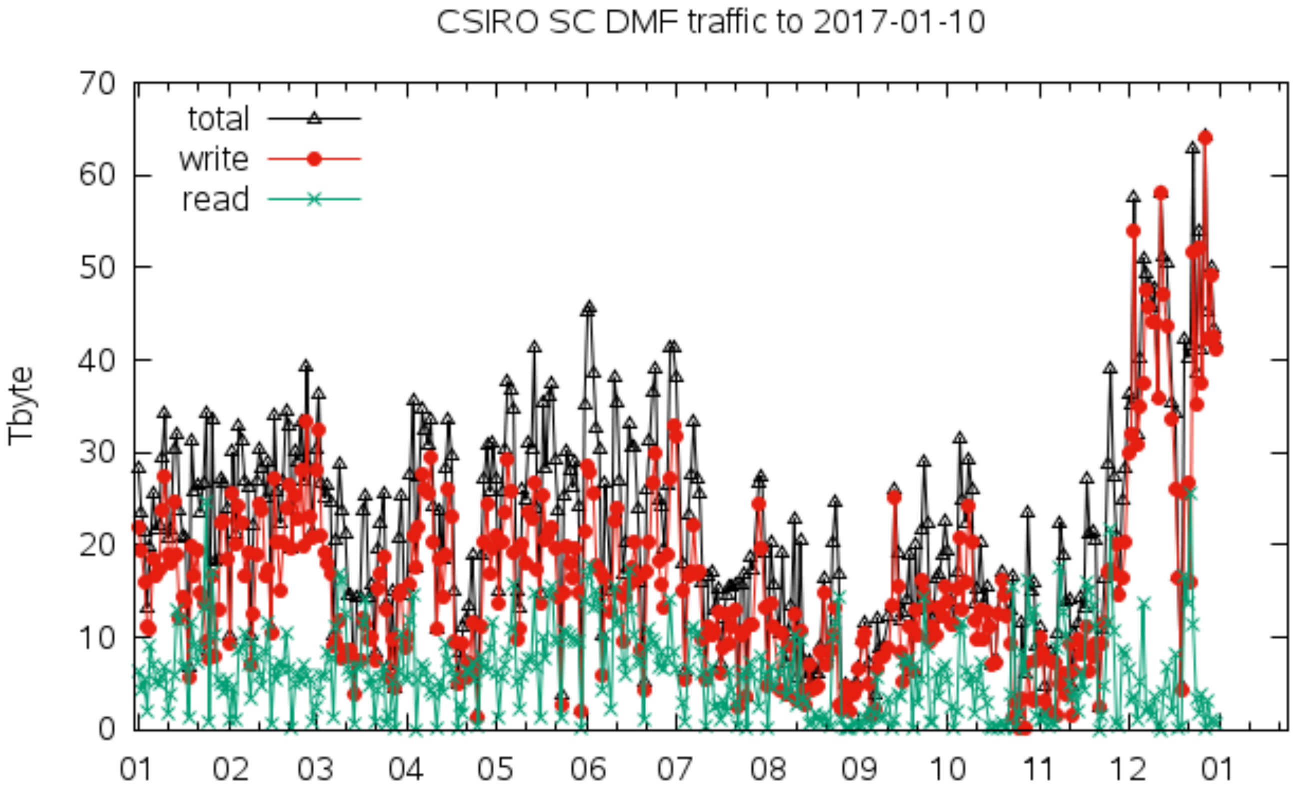
- Direct access to migrating filesystem
  - not an archive: true HSM – unlike most other sites
  - saves a lot of file transfers, copying files from archive to working area and back
  - no longer the CSIRO HPC system: refocussed the host for data-intensive, data-centric computing.
- November 2016:

“The Data Migration facility has been essential to the climate extremes, variability and regional modelling teams, as it underpins our ability to pursue new science and our ability to service client needs.

It is the primary system that we access that can reliably cope with the sheer volume of data that needs to be produced and analysed on a regular basis.

The HPC staff have worked to continually improve the system that supports the high-quality science that CSIRO is known for.  Your efforts are deeply appreciated.“

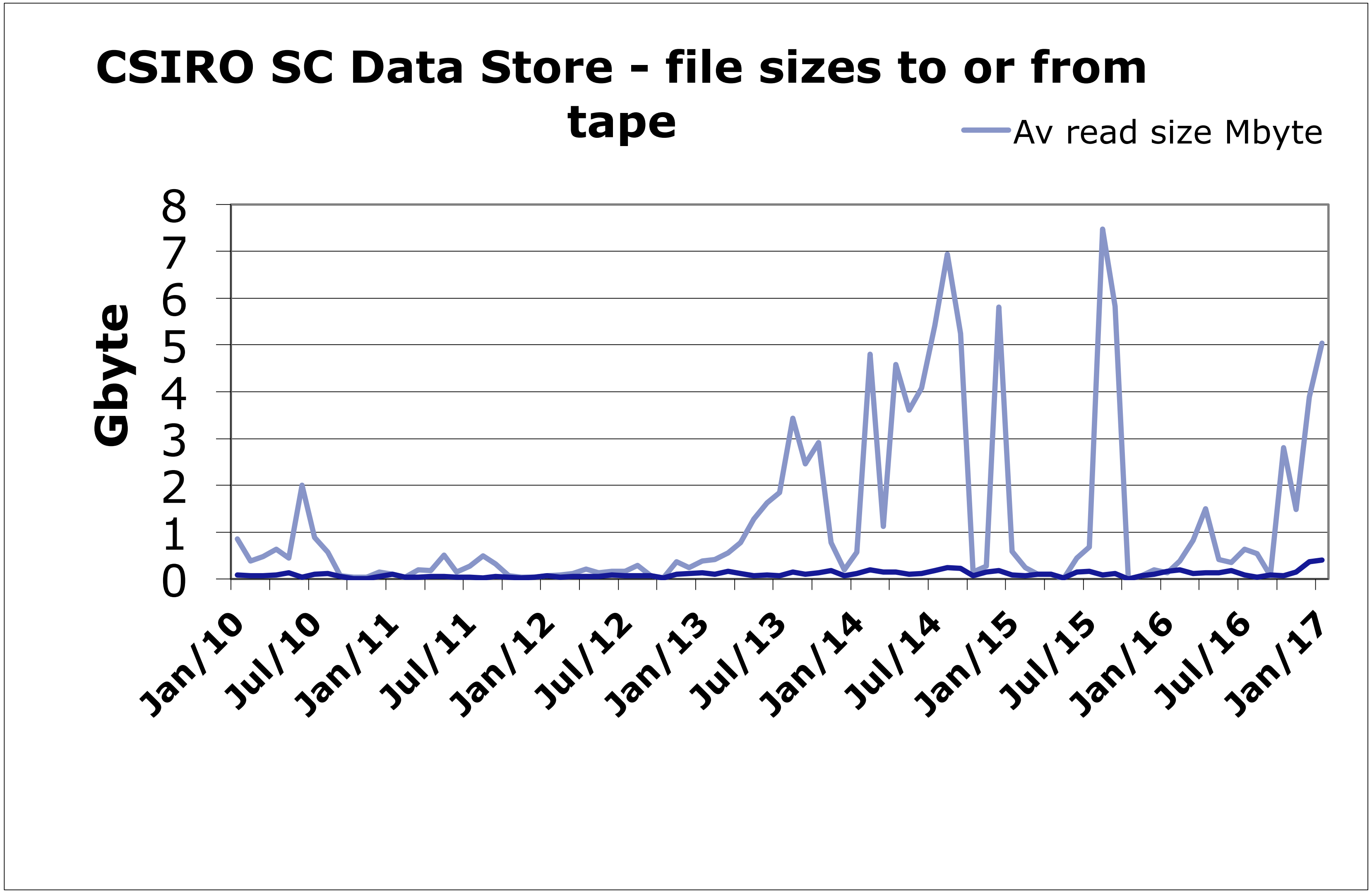
# Meeting the needs of the users



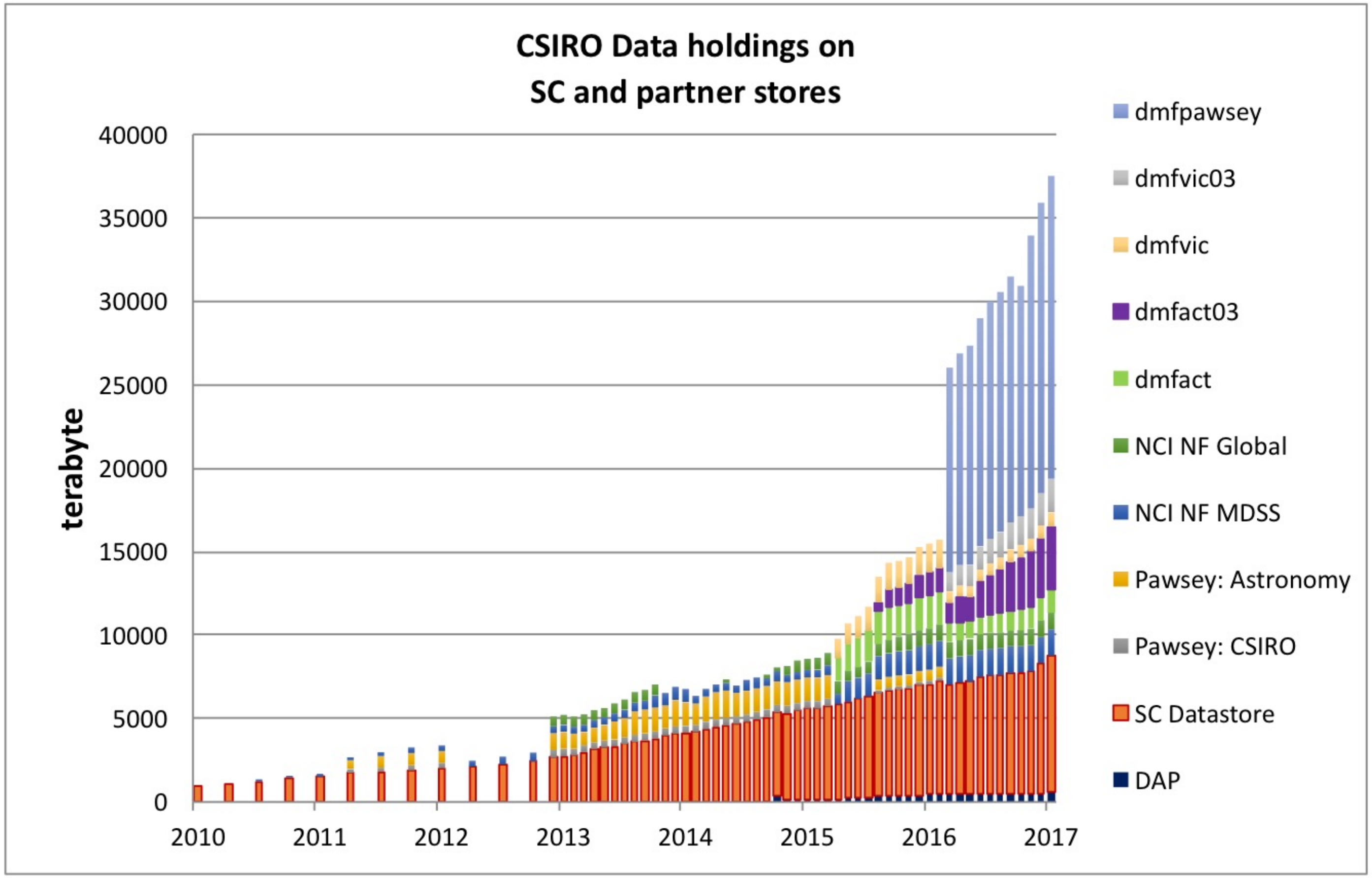
# Meeting the needs of the users

- One user added 350 Tbyte in a month, 636 Tbyte in 2 months
- Other users hate it? (DMF = Don’t Migrate ***MY*** Files!)
  - Always the file they want seems to be off-line, and even after recall, files are returned to off-line before processing.
- Vast improvements over time
  - Better off-line to on-line space ratio – was up to 100:1, now 68:1, but with cache is 16:1, with MAID 3.3:1
  - Cache (0.5 s) and MAID (~ 20 s)
  - Local dmget wrapper – optimised recalls in volume batches, provides –a flag
  - Predictive recalls – Peter’s talk
  - Average size of file recalled from tape

# Meeting the needs of the users



# CSIRO SC Data



# Leading the way for other sites

- Many sites in Australia have followed CSIRO’s use of DMF
- Early ones:
  - UTas
  - UQ
  - QUT
  - JCU
- Later:
  - NCI
  - Pawsey
  - RDS sites
- Founding of DMF User Group meeting in 2009

# Continuing value for users

- Continuing value for users
  - Quote from earlier
  - Safe place to put data: inside CSIRO, long-term
  - Insulates users from media obsolescence (floppies, Exabytes, CDs, etc)
  - Dual site
  - Staff that care
  - Recovery in the event of mistakes – restores from backups
  - Host for backups of other systems: now cross-mounted:

# Continuing value for users

Cluster: In /home/myuser/.Snapshots on 10th January (cut-down)

```
20141002 -> .../home.20141002.seq.0/myuser
20160321 -> .../home.20160321.seq.512/myuser
20160727 -> .../home.20160727.seq.640/myuser
20160930 -> .../home.20160930.seq.704/myuser
20161127 -> .../home.20161127.seq.768/myuser
20161213 -> .../home.20161213.seq.784/myuser
20161229 -> .../home.20161229.seq.800/myuser
20170102 -> .../home.20170102.seq.804/myuser
20170105 -> .../home.20170105.seq.807.recycle/myuser
20170106 -> .../home.20170106.seq.808/myuser
20170107 -> .../home.20170107.seq.809/myuser
20170108 -> .../home.20170108.seq.810/myuser
20170109 -> .../home.20170109.seq.811/myuser
README
```

# Continuing value for users

=================
pearcey snapshots
=================

Users of this system have access to snapshots made of files in their home directory on pearcey.

Each user now has a ~/.Snapshots subdirectory which contains symbolic links to the directories in the backup system which hold copies of their files.

This is not just to allow self-serve file restorations, though this is certainly possible.  Rather it's a debugging aid, allowing users to see when their files were changed using "ls -l" and what the changes were, using "diff".

# Continuing value for users

For example,

```
pearcey$   cd ~/.Snapshots
pearcey$   dmls -l */.bash_history
<output deleted>
-rw------- 4 myuser users 19935 2014-06-12 13:51 (REG) 20140614/.bash_history
-rw------- 4 myuser users 19935 2014-06-12 13:51 (REG) 20140615/.bash_history
-rw------- 4 myuser users 19935 2014-06-12 13:51 (REG) 20140616/.bash_history
-rw------- 4 myuser users 19935 2014-06-12 13:51 (REG) 20140617/.bash_history
-rw------- 1 myuser users 21504 2014-06-18 10:59 (REG) 20140618/.bash_history
```

You can see from the time-stamps that ~/.bash_history was unchanged between 12th and 17th of June.  The differences between then and the next day could be seen with

```
   diff 20140617/.bash_history 20140618/.bash_history
```

# Lessons learned

- Started out providing HPC – quickly turned into a data problem
- Constant education and help for users, continual monitorin
  - More items in the CSFbulls and HPCbulls about using DMF than any other single issue
- Need to continue to enhance the service
- Continual media transition – have to have enough hardware to cope
- Vital to limit the number of files from the start! Consolidation – tardir
- Problems with scalability of xfsdumps and FSes
- DMF saves disc space, but needs high-performance disc
- SSD can provide dramatic performance improvements: X 25 in one case.
- DMF is vital for getting around the backup problem
- DMF provides management of disc space – many sites struggle with management of FSes – restrictive quotas, balkanisation, poor or absent flushing, etc.  With DMF, almost always there will be enough disc space for the user needs
- Automatic – no need for users to fill in forms, send e-mails: expands on demand

# Summary

- 25 years of expansion to meet user needs
- Supported by knowledgeable users
- Vital to have good systems administrators
- DMF as HSM provides best experience for data-intensive computing
- Mantras:
  - "off-line data is dead data"
  - "an HSM is for life, not just for Christmas"
- Interface between CSIRO Data Portal and SC Data Store being built
- Thanks to those who made it happen

# Thank you

**CSIRO IMT Scientific Computing**
Robert C. Bell
CSIRO HPC National Partnerships

**t** +61 3 9545 2368

**E** Robert.Bell@csiro.au

**w** www.csiro.au

# Past talks

- Why HSM?
- Why is storing data at multiple sites so hard?
- Strategies for workflows on a migrating filesystem
- Using DMF as target for backups
- Ways to lose data
- Setting storage policies, and implementing them
- Monitoring DMF usage and reporting

- Coming: Scratch management, and a new CSIRO scalable flushing algorithm