Setting storage policies

… and implementing them

Robert C. Bell | CSIRO IMT Scientific Computing

12 February 2016

Introduction

1. Policies for storage
   - For users, administrators and managers
   - Why?
     - To maximise productivity (reliability, performance, capacity)
     - To minimise loss, downtime, frustration, wasted effort, expense
2. Implementing policies
   - The management of storage

- Shared systems
- My view, as a service provider and facilitator

Background

- Many of us have come from a background in HPC
  - users, systems administrators
- Assertion: the community is still working with a mind-set that computational resources are scarce, and that users are desperate for *only* these
  - Storage is an after-thought
  - Starting to change with RDSI and similar
- Quote, 2009: "*The facilities and policies for users' data affect the productivity of the users and their perception of the service.*"

May 2005 statement

- "Users typically want every file kept and backed-up, and would be happy to use only one file system, globally visible across all the systems they use, with high-performance everywhere, and infinite capacity!"
  - At zero cost!
- Protection, global, high perf., infinite

Personal devices, desktops

- One file system per device
  - if visible at all
- Protected?
  - only if the user takes action with iTunes, Time Machine, etc
- Performance decays over time
- Cloud for global
- Infinite if you buy another device, or use cloud!

Storage on CSIRO SC shared systems

- $HOME – standard POSIX
- $TMPDIR – standard POSIX
- $FLUSHDIR
  - formerly $PTMPDIR and $WORKDIR
  - now added $FLUSHnDIR
- $DATADIR
- $LOCALDIR
- $STOREDIR
- $MEMDIR
- $OSMs
- Why?

Storage on CSIRO SC shared systems

- $HOME – small, backed-up
- $TMPDIR – job or session temporary
- $FLUSHDIR – longer
  - added $FLUSHnDIR – different perf.
- $DATADIR – ‘project’ area, but no backup
- $LOCALDIR – local on node – performance
- $STOREDIR – DMF
- $MEMDIR – in memory
- $OSMs – area per project
- Why? – getting around limitations

1. Policy for (shared) storage areas

Need:

- To control access and what can be stored there
  - (music or video library? – corporate policies)
- 'quiet enjoyment' for all users
- Performance
- Protection/recovery
- Cover
  - Protection for staff/management from users' misunderstandings
- Controls to stop the FS from filling
  - for every user FS

Policy for (shared) storage areas - space management

- HSM (automatic)
- Quotas (space and inodes)
- Expiry (remove old files)
- Flushing (remove unused files)
- “Name and shame”

CSIRO SC policies

- Had very little, until:
  - 1) Hobart user lost many files (including scripts) when we flushed $WORKDIR
  - 2) Talked with other sites: found one that made users sign a statement about storage policy, to absolve providers of blame!
- Implemented user guide and statement in registration

CSIRO SC registration

…

For further information on CSIRO SC systems please see the SC User Manual at: https://wiki.csiro.au/display/ASC/User+Manual.

In particular, please read the 'File System Conventions' section at: https://wiki.csiro.au/display/ASC/SC+filesystem+conventions

It is imperative that you understand the file systems management policies on SC systems, including:

- automated flushing/removal of files

- backups are limited to a few file systems

- file migration to tape on the CSIRO Data Store

- no guarantee of any file recovery in the event of major disasters.

…

CSIRO SC Users Guide

Please carefully consider how to manage your data when using SC systems. Files can be kept in long-term storage on the datastore (especially large and consolidated files). Home directories are backed up but have limited space available (except for Ruby/datastore). High performance working space for files is available but only for short and medium term use. Copies of critical files should be maintained in multiple geographically separate locations where possible.

The SC storage is about as good as it gets but is still not immune to disaster. In particular most of the hardware is on one site. Only a limited subset of backed-up content gets duplicated to remote sites.

(But now dual-site DMF!)

CSIRO SC Users Guide

Understanding file systems and data management

Information on the file system structure of the SC systems is located on the "File systems" section in the System Guides. It is important to read the Data Store section to get an understanding of SC's data store policies and how large amounts of data are managed. In particular, with some directories like $FLUSHDIR and $TMPDIR, where data can be purged regularly, users need to save important data elsewhere.

CSIRO SC Clusters - filesystems

In the table below: 'Properties' denotes the management attributes of the underlying filesystem: back-up (b), quota (q), global (g), local (l), job-temporary (j), flush (f), and/or migrated (m).

Variable name	Properties	purpose
$HOME	q, b	Login settings, scripts, source code and built software A limited amount of space will be available.
$DATADIR	q, g	Persistent files for use in multiple jobs. Ensure that critical files left here are backed up elsewhere.
$FLUSH1DIR, $FLUSH2DIR	q, f, g	Working files semi-persistent between sessions. Ensure that …
$STOREDIR	q, m, b	Long term storage - (nfs mount of) datastore on Ruby.
…		

NCI: file system policies – excerpt

Name[1]	Purpose	Availability	Quota[2]	Timelimit[3]	Backup
/home/unigrp/user	Irreproducible data eg. source code	raijin only	2GB (user)	none	Yes
/short/projectid	Large data IO, data maintained beyond one job	raijin only	72GB (project)	365 days	No

3. Timelimit defines time after which a file is erased on the file system since its most recent access time, as defined by the file access timestamp.

- Not implemented?
  - I have a file there from start of service June 2013
- Still only 49% full
- Exact allocations

Pawsey

- Pawsey Supercomputing Centre Data Storage and Management Policy – 16 pp
  - http://www.pawsey.org.au/wp-content/uploads/2015/01/PawseyDataManagementPolicy20151.pdf
- 1_Data_ownership_legal_ethical_guide.pdf
- 2_Data_documentation_guide.pdf
- 4_Data_Publication_Re-use_Guide.pdf
- 3_Data_storage_sharing_guide.pdf
  - 15 pages total

File system policies - Pawsey

System	Location	Initial Quota	Back-up	Purged	Time limit	Permissions
$HOME	/home/[username]	10 GB	From Q2 2015	No	-	700
$MYGROUP	/group/[project]/[username]	1 TB	No	No	-	750
$MYSCRATCH	/scratch/[project]/[username]	None enforced	No	Yes	30 days	750

- 30 day time limit on /scratch
- https://portal.pawsey.org.au/docs/Supercomputers/Magnus_Purge_Policy

File system purge policy – Pawsey magnus

Motivation

- *On previous supercomputers, the scratch file system was statically allocated to projects for their duration. This led to unintended results:*

1. *Users unable to take advantage of the full scratch system….*
2. *The use of the scratch system for long-term storage….*

- Fixed age – daily scans or Robinhood?

2. Implementation …

of policy for (shared) storage areas

Implementation of policy for (shared) storage areas – space management

- HSM (automatic)
- Quotas (space and inodes)
- Expiry (remove old files)
- Flushing (remove unused files)
- “Name and shame”

Space management – HSM – big winner

- Automatic by DMF when thresholds reached
  - dmmigrate and dmfsfree, site policy
  - Need to augment with dmgets, dmputs
- Need to provide ‘quiet enjoyment’: stop one user having too much impact on others
  - quotas: inodes, and on-line space
  - custom dmget: stops domination
  - new DMF recall queue management
- CSIRO SC Data Store – no limits on data volumes for 24 years

Space management - HSM

- HSM – ‘infinite’ storage, 'elastic' disk
- CSIRO SC one of few sites that provides direct user access with /home on ruby, SGI UV 3000 (UQ?)
    - True meaning of HSM (for users)
    - Otherwise, just an archive store with multiple levels, with users having to move data in and out
  - 10 Tbyte /home quota (on-line) on ruby
    - c.f. 2 Gbyte on raijin, 10 Gbyte on magnus
- Lots of CSIRO enhancements (dmget, etc), lots of guidance for users
- No form-filling; immediate access, low administration
- But: unfamiliar experience for naïve users
  - Users scared off by “df”

Space management - HSM

- User experience:

  The problem with ruby, if its anything like cherax, is that once I get it (the data) there it goes onto tape before I can slice and dice it. This was my experience with my ocean reanalyses and my solution was to do the data analysis locally on a 12TB drive I purchased.

- UV3000 (ruby) has bigger disk storage than UV1000 (cherax)
- More help needed … workflows

Space management – quotas

- To control space (most often) & files
- To report on usage quickly
- For all areas that users could fill (/var)
  - Better for a user to hit a quota limit than for a FS to fill –> re-boot…
- Allocation by quotas. Either:
  - Balkanisation
    - sum of all quotas = available space
    - very wasteful
  - Over-commitment – harder to manage

Space management – inode quotas – HSM

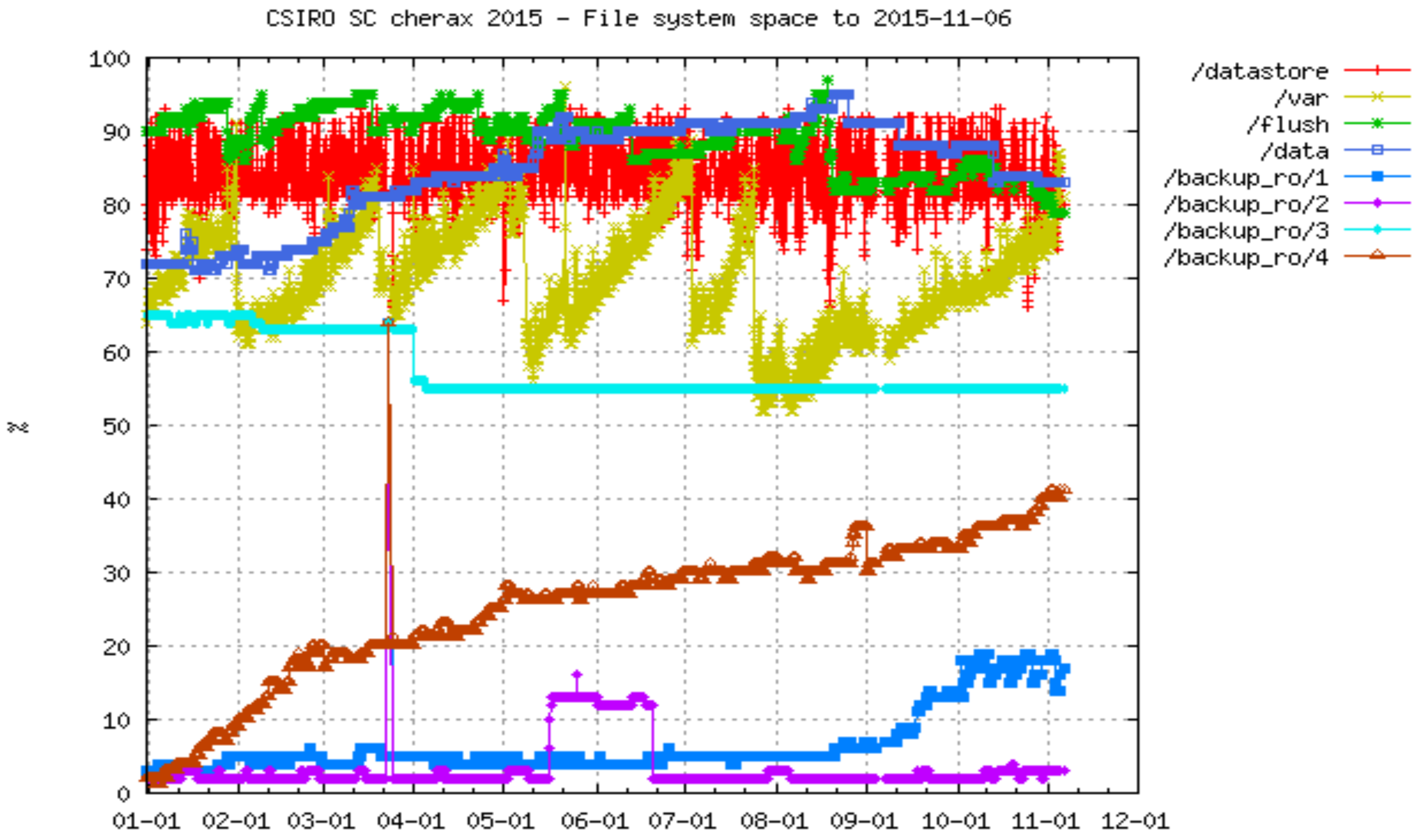
- DMF performance issues when number of files increase
  - Dump time (but Mediaflux..)
  - Default of 150,000 for CSIRO SC DS
- CSIRO SC DS: 15 M small files in 35 M
  - Don’t want to migrate these
    - too much overhead
  - Dump speed on ruby – down from about 3 M files/hour to about 1 M/hour
  - tardir and other encouragement

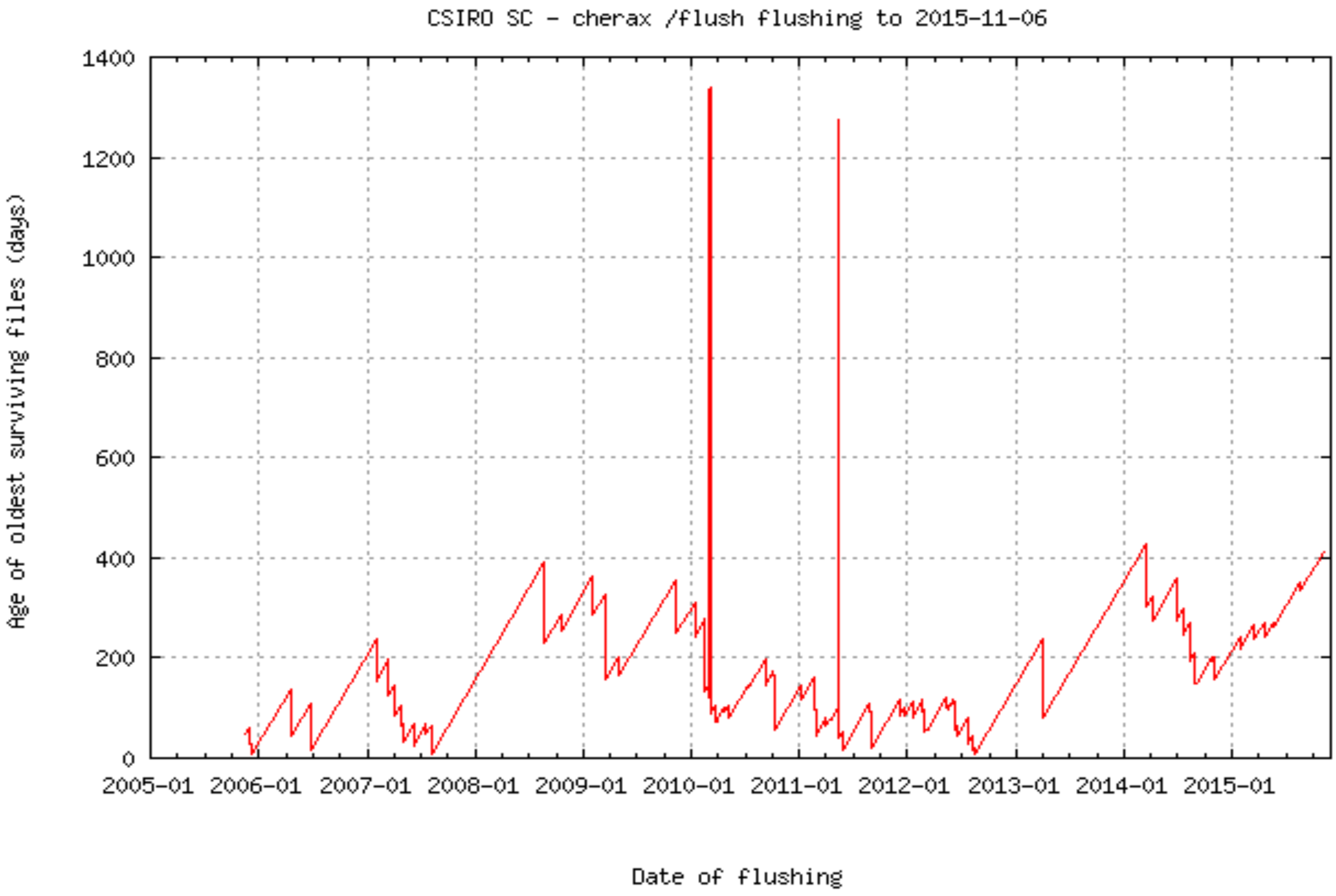
Space management – expiry

- One site used to remove all files from /home more than a year old
- Other site wiped all the user areas apart from /home at the beginning of each new allocation cycle
  - Did not suit continuing projects

Space management - flushing

- Often not implemented
- CSIRO SC – scripts and program to implement policy
  - triggered when usage reaches a threshold (typically 95%)
  - file audit – sorts, and deletes oldest until second threshold reached (typically 90%), or 7 days (rare)
  - uses mtime and atime – problem for FSes that don’t do atime
  - also removes empty directories


CSIRO SC cherax 2015 - File system space to 2015-11-06
%
0
10
20
30
40
50
60
70
80
90
100
01-01
02-01
03-01
04-01
05-01
06-01
07-01
08-01
09-01
10-01
11-01
12-01
/datastore
/var
/flush
/data
/backup_ro/1
/backup_ro/2
/backup_ro/3
/backup_ro/4


CSIRO SC - cherax /flush flushing to 2015-11-06
Age of oldest surviving files (days)
0
200
400
600
800
1000
1200
1400
2005-01
2006-01
2007-01
2008-01
2009-01
2010-01
2011-01
2012-01
2013-01
2014-01
2015-01
Date of flushing

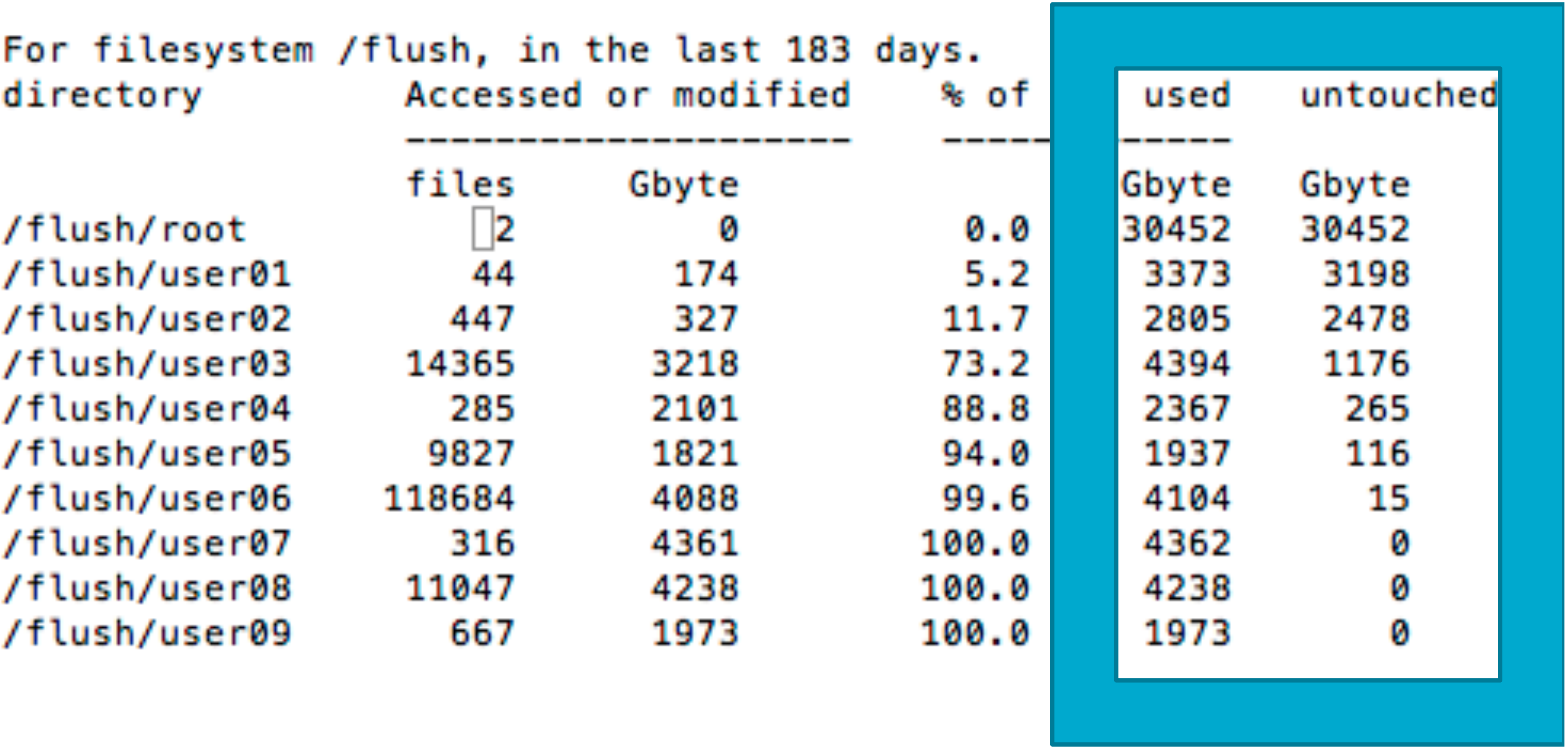
Space management – flushing – why is it hard?

- Need to be careful!
- Access and modify time
  - access time is dodgy with some filesystems (e.g. NFS)
- Sheer amount of work:
  - 283 M, 460 M files seen
- Global FSes – slow metadata operations – typically 1000/s
- Questionable policies – e.g. fixed times
  - wastes space and processing
- Bad practice to have policy and not implement it

Space management – “name and shame”

- Last resort – only thing left for over-quota’ed project areas
- Lists of big users
  - sometimes augmented with measures of ‘waste’
- Ineffective, except for small groups
- Relies on harnessing peer-group pressure

Policy for (shared) storage areas
- space management – "Name and shame"



```
For filesystem /flush, in the last 183 days.
directory           Accessed or modified    % of       used    untouched
                    --------------------    -----      -----
                    files        Gbyte                 Gbyte    Gbyte
/flush/root             2            0       0.0       30452    30452
/flush/user01          44          174       5.2        3373     3198
/flush/user02         447          327      11.7        2805     2478
/flush/user03       14365         3218      73.2        4394     1176
/flush/user04         285         2101      88.8        2367      265
/flush/user05        9827         1821      94.0        1937      116
/flush/user06      118684         4088      99.6        4104       15
/flush/user07         316         4361     100.0        4362        0
/flush/user08       11047         4238     100.0        4238        0
/flush/user09         667         1973     100.0        1973        0
```

Conclusion

- Policies for storage
  - necessary for users, systems staff and management
  - range of options: maximise the value of the resources
  - need to communicate the policies (beforehand!)
- Implementing policies
  - necessary, to avoid disasters and wastage
  - tends to be over-looked
  - disasters in waiting (users' ignorance and complacency), masked by reliable hardware (mostly)
  - mustn't add to the disasters!

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