Why HSM?
(Hierarchical Storage Management)

Robert C Bell

Technical Services Manager

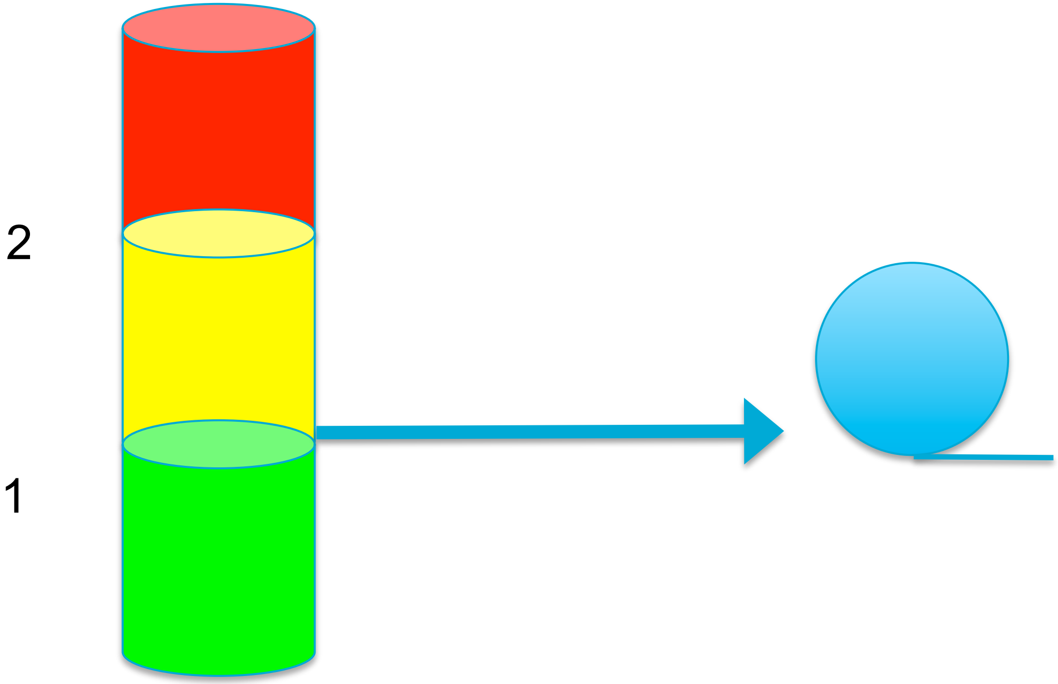
Advanced Scientific Computing

CSIRO

Outline

- CSIRO has been using a hierarchical storage management (HSM) system, the Data Migration Facility, for over 18 years.
- This talk will draw on this experience to give some answers to the question, "Why HSM?”
- Answers to give to Management

HSM: Provides a migrating file system



1. Threshold – start copying to tape
2. Threshold – start deleting data from files
   – typically old or big.

Recall – on access, or by command

Storage – terminology

- **Transfer of data between primary and secondary media**

Process	'Disc'	'Tape'
Backup	data and metadata remain	copy of data and metadata made
Archive	data and metadata removed	copy of data and metadata made
Migration (HSM)	data removed, metadata remains	copy of data made (and sometimes metadata)

Storage – terminology

- **Transfer of data between secondary and primary media**

Process	Action	Reverse of
Reload	Fill entire file system data and metadata from 'Tape' to 'Disc'	Backup
Restore	Copy selected file data and metadata from 'Tape' to 'Disc'	Backup
Retrieve	Copy selected file data and metadata from 'Tape' to 'Disc'	Archive
Recall	Copy selected file data from 'Tape' to 'Disc'. Data is automatically connected to the metadata. Recall can be automatic upon reference to the file.	Migration

HSM

Modes of operation

1. **Traditional supercomputer site: back-end system**
   1. insulates the system – easier for transitions
   2. no direct access – harder for users (archive system)
2. **Direct user access: $HOME or other file system: true HSM**
3. **Hybrid: back-end, plus shared file system access (NFS or better)**

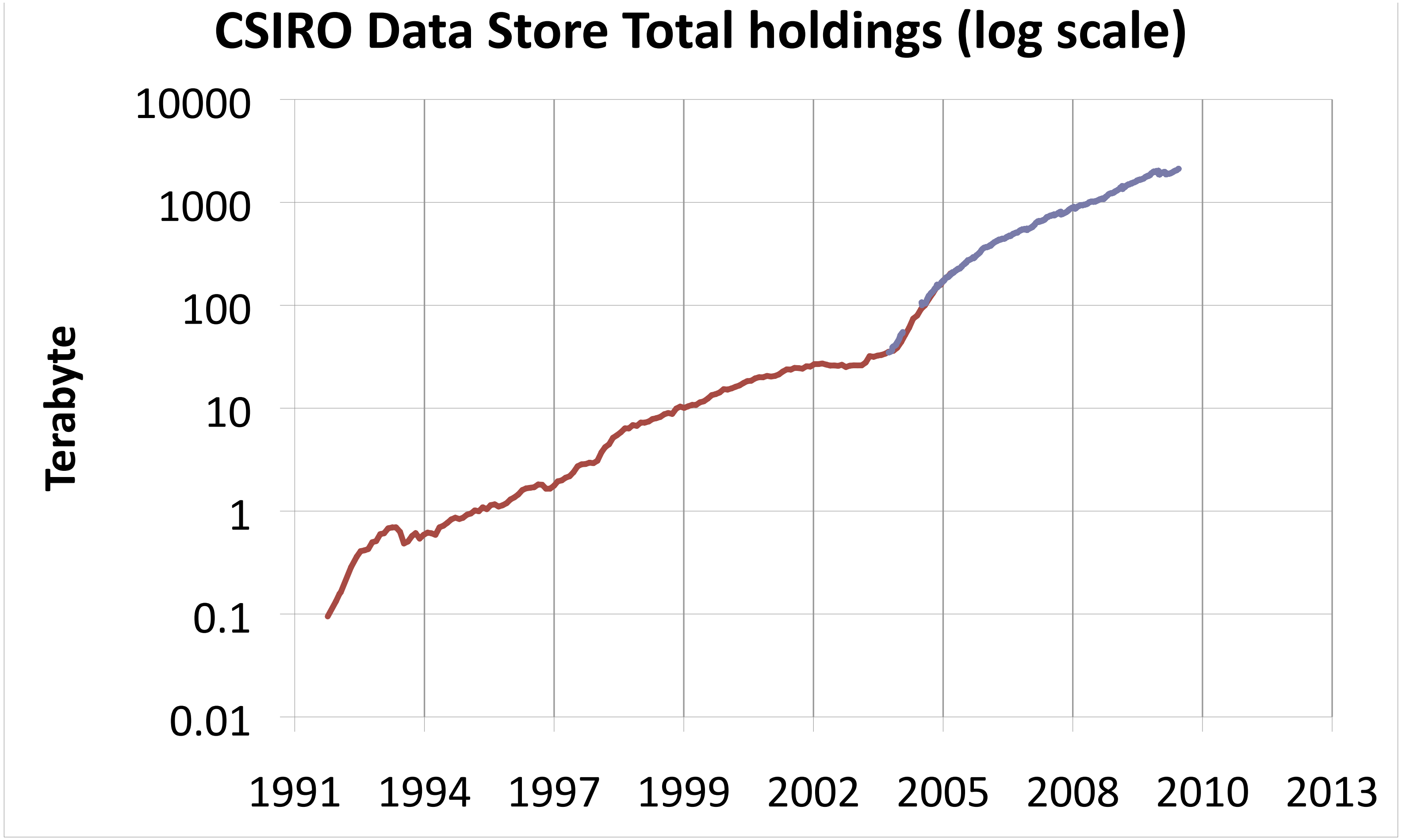
Why an HSM?

1. Provides' infinite' capacity
2. Copes with change
3. Provides storage protection for the long-term
4. Gets around the backup problem
5. Green: can save energy – disc and tape, versus disc only.
6. Improves users productivity
7. Can lower the cost of storage
8. Allows users to have direct access to vastly more data than otherwise.
9. Allows users to ‘see’ all of their data holdings where they work
10. Solves the space management problem.
11. Host for data-intensive computing
12. Target for backups of other systems
13. Better for users than archive system
14. Avoids the bit-rot problem

A1: Provides ‘infinite’ capacity

- **Satisfies users’ and science needs**
- **Enhances their productivity**
- **Easy to add capacity to meet need**
  - **every day (no reconfiguration of file systems, disc arrays)**
- **No user space quotas!**
- **No filling of file systems**
- **CSIRO Data Store: CAGR of 1.8**

CSIRO Data Store growth



A2. Copes with change

- **CSIRO ASC Data Store:**
  - 4th site, 5th host, 9th tape format
- **Avoids “Off-line data is dead data”**
- **System carries out the transitions: not dependent on users copying data from old media: cards, Exabytes, floppies, CDs, etc.**
- **Shifts responsibility from users to systems – economies of scale**

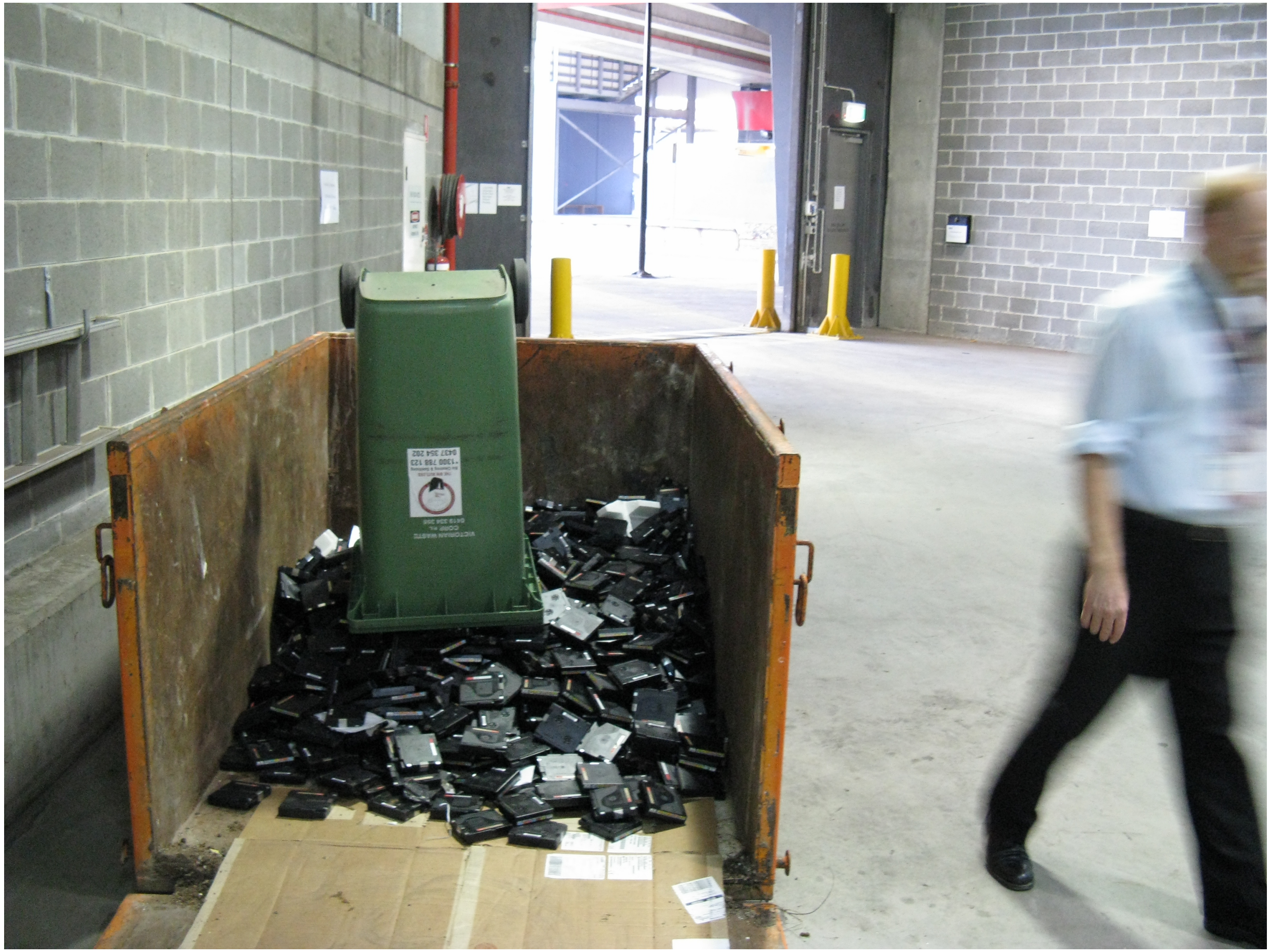
A3. Provides storage protection for the long-term

- **Secondary copies on tape can provide a greater level of protection.**
  - Two or more copies can be easily kept (copies could go off-site)
  - No system administrator action can remove data from an off-line tape. The same is not true for on-line disc.
  - Could keep almost all copies ever made (no hard-deletes)
  - See A2 – copes with transitions

A3. Provides storage protection for the long-term

- **Can keep everything (almost)!**
- **No point in removing old files**
  - Data more than 3 years old is < 10%
- **CSIRO ASC Data Store:**
  - 18 years of DMF
  - 20 years of ‘same’ file system
  - files go back to mid-1960s

Your data, our responsibility



A4: Gets around the backup problem

- **User file systems need backup!**
  - protection from mistakes, etc: another talk!
- **Backups: need ~10 X space**
- **With HSM, data from files is trickled out**
  - on demand, or at set times – multiple copies
- **Backups have to deal only with**
  - metadata
  - new files
  - small files (at CSIRO, not migrated).

A4: Gets around the backup problem

- **CSIRO Data Store user file system**
  - 1.1 Pbyte total (x 2 copies)
  - 6.6 Tbyte primary disc
  - 17 million inodes
  - Full dump 90 Gbyte (done weekly)
  - Incremental ~ 3 Gbyte (done daily)
- **Dumps can share tape infrastructure with HSM**

A5: Green: can save energy

- **Disc and tape, versus disc only**
  - 1 Pbyte of disc – 12 kW + cooling costs?
  - 1 Pbyte of tape – 0 kW + air-cond costs?
  - Tape drives – 100 W each
  - Tape library – a few kW

A6: Improves users' productivity

- **Users should not be restricted in their science by the capacity of the storage system, or by the policies of their systems administrators!**
- **Direct access to all files**
- **Reduces the need for users to have and manage files in multiple places**
- **Reduces the need for file transfers**
  - the most-error prone part of many workflows

A7: Can lower the cost of storage

- **Often used to justify HSM**
  - Move old data from disc to (cheaper) tape
- **But commodity disc costs ~ tape costs**
- **Still justified if using enterprise high-performance and cost disc arrays**
  - (However, need such disc arrays to drive tape at full speed)
- **Need to show user productivity gains:**
  - e.g., only place to store large data sets
- **Energy savings => cost savings**

A8: Allows users to have direct access to vastly more data than otherwise

- **All the users' permanent file holdings can be in one place: backed-up**
- **Users can see all the metadata, can recall any file within minutes, have small files always on line.**
- **CSIRO Data Store: single users with 100+ Tbyte**

A9: Allows users to ‘see’ all of their data holdings where they work

- **With direct access to HSM, users don’t have to run special commands or login to other systems, to see their holdings**
- **Gets away from old ‘compute-centric’ HPC centre, where data always had to be staged from a separate system**
- **Standard UNIX/Linux commands and access**
- **With good HSM (e.g. DMF), augmented commands to deal with file residence, e.g. dmls, dmget**

A10: Solves the space management problem.

- **Good systems managers prevent file systems filling (or jobs fail)**
  - quotas
    - not complete solution, since we over-allocate
    - “a resource divided is a resource diminished”
  - flushing – not for permanent areas
  - ‘name and shame’ big users – tiresome
- **HSM – automatic space management**
  - thresholds or time based
    - CSIRO Data Store – disc 50% empty each morning, ready for the users

A11: Host for data-intensive computing

Misquoting Tennyson: *The Brook*:

“Flops may come and flops may go, but bytes go on forever”

Hamming: “The object of computing is insight, not numbers”

- Increasingly, computing is becoming ubiquitous, but the big problems are around big data: astronomy, high-energy physics, remote sensing, climate model output, etc.

A11: Host for data-intensive computing

- **Move from ‘compute-centric’ to ‘data-centric’ facilities**
- **Direct user access to vast quantities of data is key – HSM allows this**
  - otherwise, users have to manage the residency – ugly!

A12: Target for backups of other systems

- **Most centres have many systems and servers – VMs coming too!**
- **These need backups:**
  - user and system areas
- **Don’t want tape drives on each system**
- **Don’t want to backup over network to tape drives on a server**
- **Don’t want HSM on all systems**
- **Want disc to disc to tape backup**

A12: Target for backups of other systems

- **Use an HSM-managed disc area as a target for backups**
- **Can cope with large volumes**
- **Get tape handling for nothing**
- **Can keep most on tape**
  - restores are rare
- **On disc, so have individual file addressibilty – fast restores**

A12: Target for backups of other systems

- **CSIRO ASC experience:**
    - Can use rsync to get full backup for the cost of incrementals every time
    - Can use rsync --link-dest to reduce ratio of backup/source data to < 2 compared with typically 10
    - Can use Tower of Hanoi scheme to optimise the keeping of backup sets

A13: Better for users than archive system

- **Good storage systems have to be more than an archive**
- **An archive system requires explicit actions by users to receive and recall data.**
  - Our experience is that archive systems don't work without making life difficult for users on the HPC system.
  - Users are reluctant to copy their files away from where they work, and even more reluctant to remove files from their primary disc area.
  - Archive systems do not provide management of the disc space area where users do their work.

A14: Avoids the bit-rot problem

- **Large disc-based archives are subject to the bit-rot problem**
  - Dormant data is not read, and could be corrupt
  - Increasing problem with larger discs
  - Need to have program of checking discs
    - eats into bandwidth
  - Even with RAID, an increasing problem
  - RAID re-build times are unacceptable, and data at much higher risk during re-build

A14: Avoids the bit-rot problem

- **HSM**
  - All inodes scanned every day for dumps
  - New small files read every day until full dump
  - Weekly full dump reads and writes all small files (and all directories)
  - Large files read and written to secondary media (DMF)
  - Large files have checksumming, and are re-read and written at least every few years. At least two copies.
  - Recalled files are checked upon recall, and are essentially only temporarily in the store.

HSM Drawbacks?

- Files off-line – delays
  - User education (on-going)
- Overhead for each file
  - Keep small files on-line, use cache disc
  - Restrict numbers of files => consolidation
- Backup doesn’t go far back in time
  - Currently 35 days or more at CSIRO
- Need automation, fast disc, tapes
- Need good systems staff!

Conclusion – Why HSM?

- Capability: enables users to do large data-intensive computing
- Improves users’ productivity
- HSM is the only way to continually provide large and growing capacity for user needs
- Supports change
- Provides greater protection for data
- Can get around the backup problem

- For users, can be strange initially

CSIRO ASC
Dr Robert Bell

Phone: +61 3 9669 8102
Email: Robert.Bell@csiro.au
Web: http://hpsc.csiro.au/

Thank you

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

CSIRO