TSM Linux User Experience at CERN

David Asbury, CERN, Geneva, Switzerland

Oxford TSM Symposium, 26 September 2007

Topics

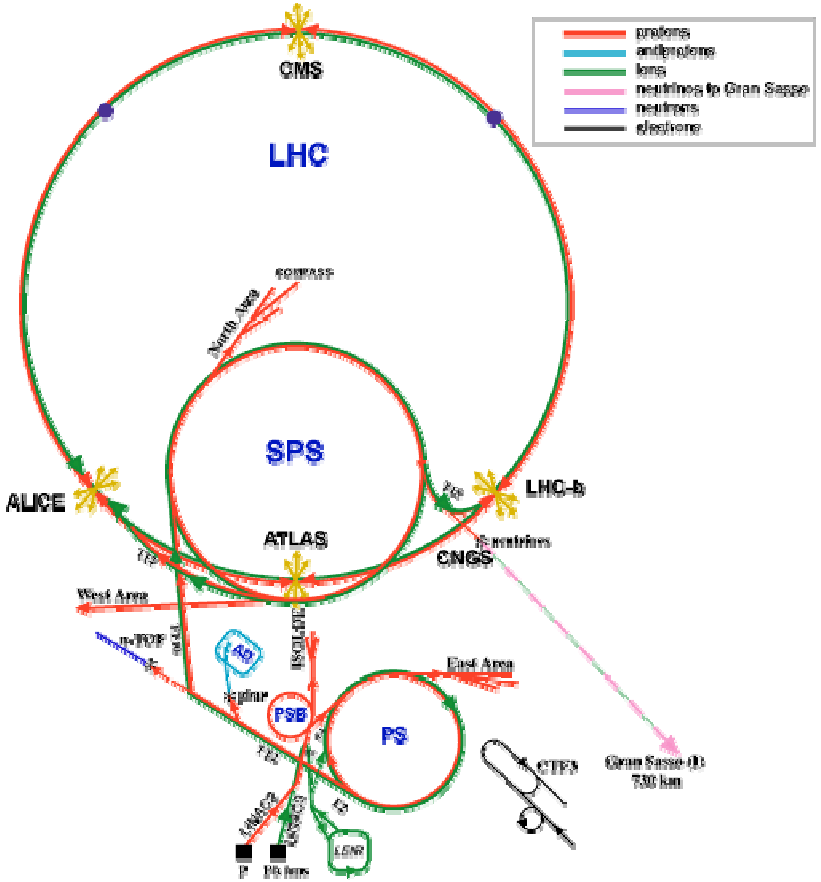
- What is CERN?
- What do we do with all that data?
- How TSM is used in CERN
- Managing the growth of data
- Configuration
- Experience with Linux

What is CERN?

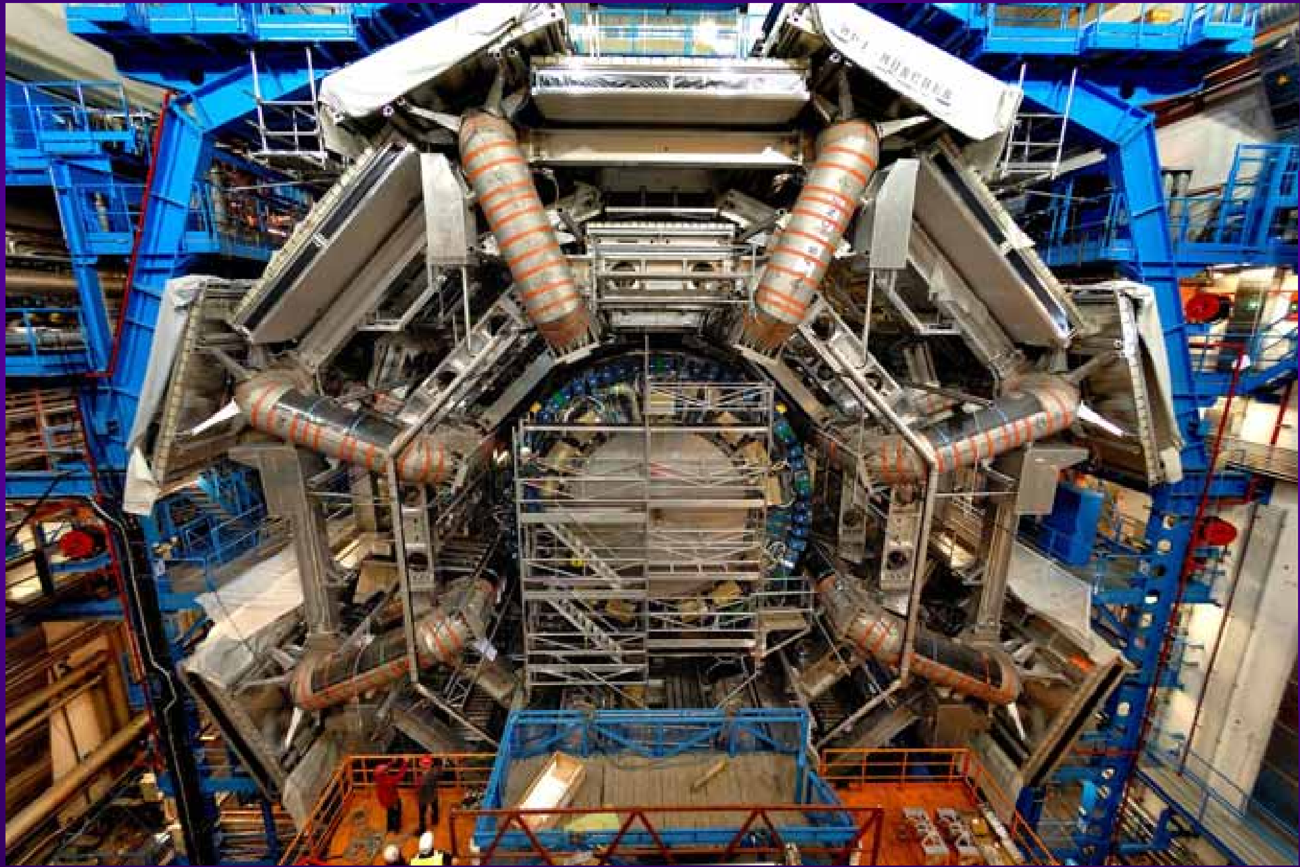
- European Laboratory for Particle Physics
- French-Swiss border near Geneva
- 20 member states, ~3000 staff
- ~6500 visiting scientists from ~500 institutes, ~80 nationalities
- Large Hadron Collider (LHC) to open 2008

Large Hadron Collider

Accelerator Complex



Atlas Experiment



Data Pyramid

Derived data,
Physics dbs

Mail, Home directories
Databases, systems etc.

Raw Data from
experiments is
distributed among
10 other Grid sites.
~15PB per year

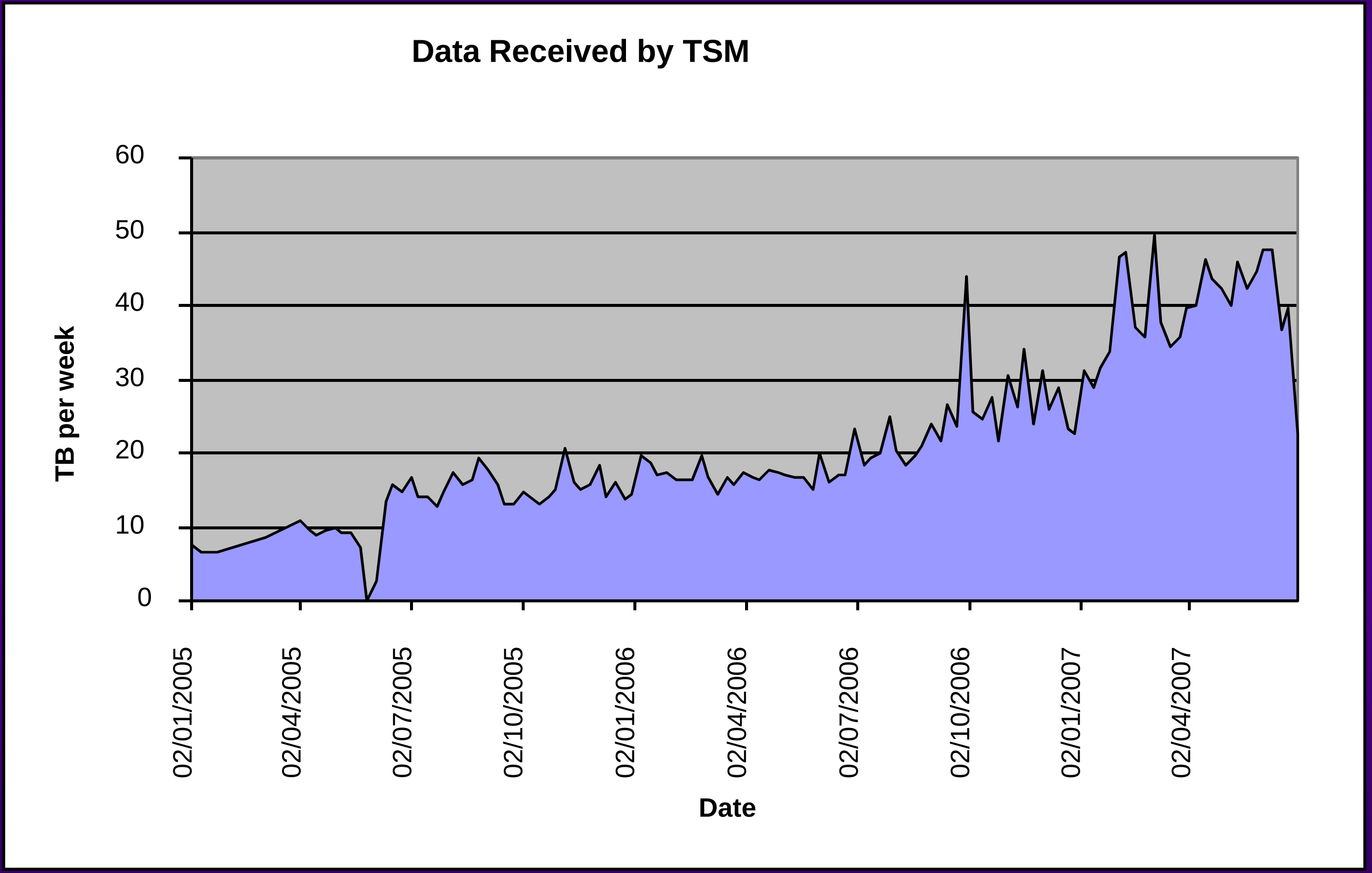
CERN Policy on Backup

- Home Directories
  - AFS — AFS volume backup -> Castor
  - Windows DFS — TSM
- Mail
  - Microsoft Exchange — TSM
- Databases — TSM
- Unix group & project servers — TSM
- Experimental Data — Castor

Only Servers

Castor: CERN Advanced Storage Manager (local)

Growth!



Data Received by TSM
TB per week
60
50
40
30
20
10
0
02/01/2005
02/04/2005
02/07/2005
02/10/2005
02/01/2006
02/04/2006
02/07/2006
02/10/2006
02/01/2007
02/04/2007
Date

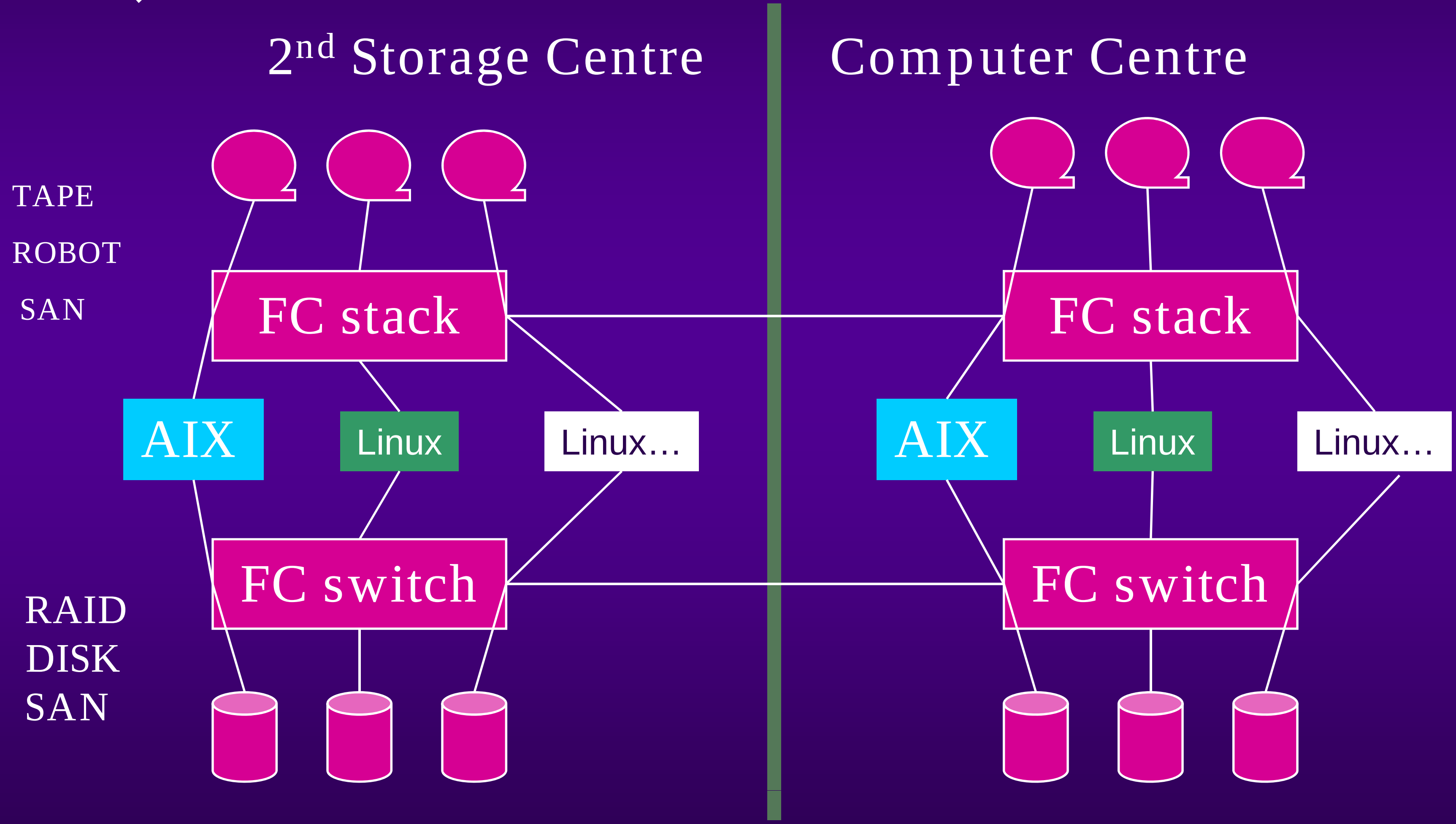
Managing growth

- **Ask the major clients for forecasts**
- **Monitoring everything they do too!**
  - **Servergraph, moving to home-grown TSMMS**
- **Want a repeatable "unit" of TSM**
  - **Can add when needed to avoid performance problems**
  - **Use existing TSM FC infrastructure**
  - **Profit from local Linux expertise and installation**
  - **Make use of physics robotic tape infrastructure**

A Unit of TSM Capacity

- PC running standard RHEL4 64-bit Linux
  - 4 cpus, 8GB memory, 2 Qlogic HBAs for FC
- System disks mirrored by 3ware card
- Disks for TSM db & log mirrored by TSM
- RAID6 disks for staging areas
- Use physics robot tape infrastructure

TSM Configuration



Setting up the Linux etc.

- IBM only supports specific Linux kernels
- IBM tape drives need specific IBM driver
- More restrictive than AIX or Solaris
  - No “smitty” system tool like AIX
- Must reload FC driver to add devices
  - Disks MUST be labelled in /etc/fstab for safety
  - Cannot avoid Unix disk cache with ext3 fs
  - Tape drive devices may change name if add new ones
  - Must change access to tape devices if TSM not run as root
  - Usually have to reboot

Spec of 1st TSM on Linux

- PC Intel Xeon 2x3Ghz cpus, 4GB memory
- System disks mirrored by 3ware card
- Standard RHEL4 64-bit Linux (specified)
- Raptor disks mirrored by TSM for db & log
- SATA RAID6 Infortrend array for staging
- Ext3 file system used (specified)
- 8 IBM 3592J tapes (300GB) in 3584 robot

1st TSM on Linux

- Started well, performance okay
- Functioned normally
- High load (>1 cpu) when doing i/o
- Sometimes does not schedule all TSM processes concurrently?
- Beware of Linux “tools” for devices
  - Rewound tape drives!
- Added 2nd Linux machine …

Spec of 2nd TSM on Linux

- AMD Opteron dual core, 4 cpu, 8GB mem.
- System disks mirrored by 3ware card
- Standard RHEL4 64-bit Linux (specified)
- Raptor disks mirrored by TSM for db & log
- SATA RAID6 Infortrend array for staging
- 6 LTO3 HP drives in STK 8500 robot
- LTO drives mounted via ACSLS
  - Need special script to create device files

2nd TSM on Linux

- Started well, but high cpu with i/o again
- Corrupted file systems with high disk i/o
  - /var/log/messages “trying to seek off end of disk”
  - Reboot stopped - needed manual fsck of file systems
  - System down for some hours to check file systems and ran TSM AUDIT on disks to cleanup
- Upset Backup clients!
  - System not available when needed
  - Backups corrupted? - yes

Tracing the Corruption

- Tried changing RAID arrays, updated kernel and Qlogic FC driver
- Tried single-processor kernel.
  - Better, but still corrupted
- Borrowed RedHat certified PC
  - Still corrupted with memory problems, audit errors
- Eventually moved big clients back to AIX
  - Linux better lightly loaded, but still see audit errors

Corruption

- Needs high disk i/o – only seen with disks connected by FC
- Single processor kernel was better, but too slow (limited cpu for i/o)
- Did not seriously suspect RAID arrays as have worked well with AIX for years
- Difficult to separate Linux fs from FC
- Run TSM AUDIT frequently, but cannot check data (only metadata)

CERN Corruption Survey

- Used fsprobe program in C (not TSM)
- Just reads/writes Unix files and checks
- Run on ~3000 farm PCs in CERN for some weeks
- Variety of silent corruption found:
  - Memory errors, less than expected. 1-bit errors are corrected
  - Sector/page sized regions corrupted
  - Larger blocks of invalid data – ext3 file system?
- All makes of PC eventually showed errors
- Memory is most dangerous place for your data!

Conclusions

- Jury still out. Linux fs or FC-related?
- Linux offers cheaper repeatable unit?
- Problem: no single point of contact
  - No clear line between hardware and software
  - Different PCs show corruption in different ways
  - Extremely time consuming, disruptive to service

Next Steps

- Try IBM configuration certified for TSM
  - PC, Qlogic HBAs, IBM RAID with RHEL4
- Pay IBM to take all problems (Redhat too)
- Hope for clear answer to problem – do not want to repeat all this with new hardware!
- *Talk in TSM Symposium 2009 on results?*

Acknowledgements

- Lio Frost-Ainley
- Gordon Lee
- Tim Bell (boss)
- Charles Silvan (Expert from GATE & IBM)
- Peter Kelemen (Corruption Survey)

Contact Details

- David Asbury, CERN IT Department
- Email: david.asbury@cern.ch
- CERN Website: www.cern.ch

Questions?