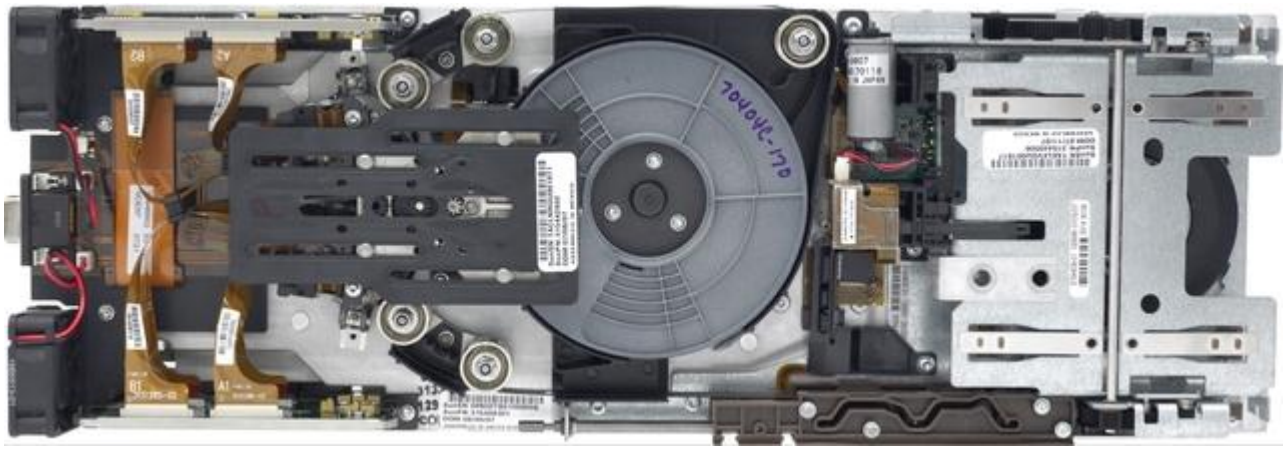
CERN Tape Operations

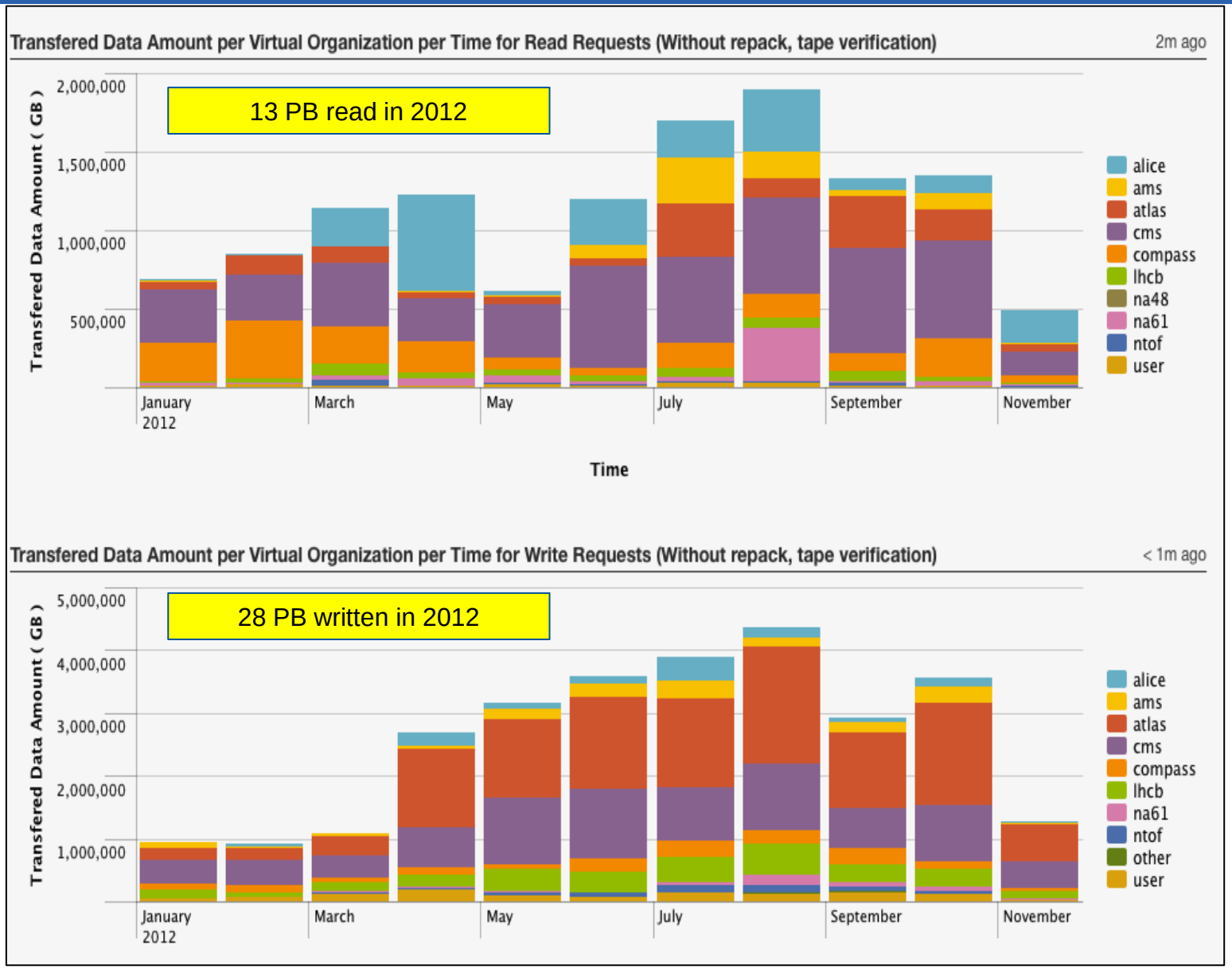


Vladimír Bahyl
Data Storage Services
IT Department

Outline

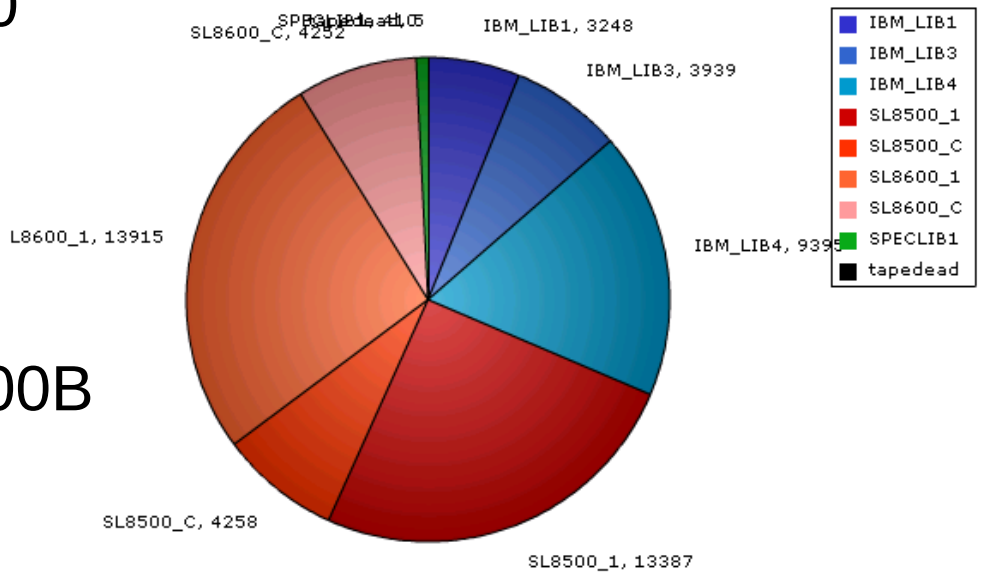
- Where are we?
- Hardware infrastructure
- Tape servers
- Software stack
- Configuration
- Verification
- Monitoring
- Demo

Where are we?


Transfered Data Amount per Virtual Organization per Time for Read Requests (Without repack, tape verification)
2m ago
13 PB read in 2012
Transfered Data Amount (GB)
2,000,000
1,500,000
1,000,000
500,000
January 2012
March
May
July
September
November
Time
alice
ams
atlas
cms
compass
lhcb
na48
na61
ntof
user
Transfered Data Amount per Virtual Organization per Time for Write Requests (Without repack, tape verification)
< 1m ago
28 PB written in 2012
Transfered Data Amount (GB)
5,000,000
4,000,000
3,000,000
2,000,000
1,000,000
January 2012
March
May
July
September
November
alice
ams
atlas
cms
compass
lhcb
na61
ntof
other
user

Hardware Infrastructure

- IBM
  - 3 x TS3500
  - 44 x TS1140, 18 x TS1130
  - 6000 x JC, 12000 x JB
- Oracle
  - 4 x SL8500
  - 40 x T10000C, 20 x T10000B
  - 8500 x T2, 27000 x T1
- Spectra Logic
  - 1 x T-Finity
  - 5 x LTO5
  - 400 x LTO5



Tape Servers

- Software RAID1 saves money and is working fine
- DELL blades (2010)
  - RAM: 12 GB; CPU: Intel L5520 (4 cores) 2.3 GHz; Disk: 2 x 150 GB; NIC: 10-Gigabit; HBA: Qlogic QME2572
  - Good density on installation
  - 2 interventions required total shutdown ☹
- 2U Broadberry (2011)
  - RAM: 16 GB; CPU: Intel E31260L (4 cores) 2.4 GHz; Disk: 2 x 250 GB; NIC: 10-Gigabit; HBA: Qlogic QLE2562
- Make sure all components can be replaced independently
- We manage the fibre infrastructure

- Experimented with 2 tape drives / server
  - Data transfer is working fine
  - Operational challenges
    - Interventions interfere and take longer; Some daemons shared; Procedures incorrect
  - Not worth it

- QUATTORized today, move to Agile Infrastructure (Puppet & friends) planned for 2013

Software stack

- CASTOR 2.1.13-7
  - `tapebridged` in front of `rtcpd`
  - `rtcpd` connections limited to `localhost`
  - Many bugs fixed – see Steven's talk
- "Buffered tape marks" in production
  - Significantly improves writing speed for small files
  - Require some specific RPMs (kernel module) + new options
  - **You need to have this!**
- Writing until physical EOT
  - Adds 10% extra capacity with Oracle T10000C (5.5 TB)
  - **You should have this (if you have enterprise drives)!**

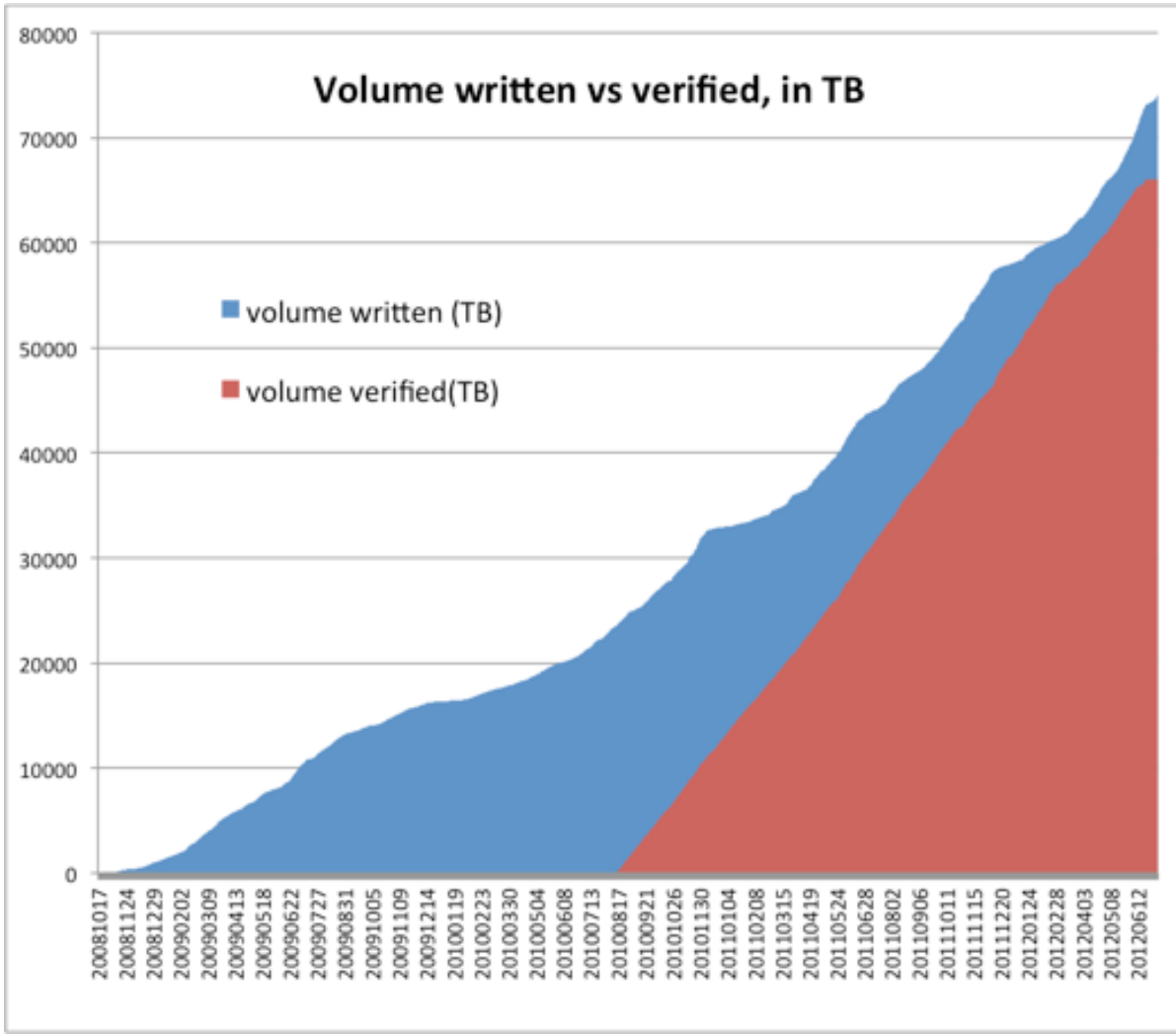
Configuration

- Drive names explained
  - Format: TechnologyLibraryDrive@Tapeserver
  - Examples:
    `T10C5415@tpsrv054`, `T10B651F@tpsrv958`
    `35921012@tpsrv035`, `S1LTO504@tpsrv029`
- Device groups (DGN) names explained
  - Format: TechnologyLibraryBuilding
  - Examples: `T10KC5`, `T10KB6`, `3592B1`, `S1LTO5`

- Symbolic links
  - `/dev/tape`
    - By default, UDEV creates `/dev/tape` as a directory with ID of the tape device
    - We overrule that to have `/dev/tape` a symlink to `/dev/nst1`
    - Very handy with the `mt` command
  - `/dev/smc`
    - On IBM robots, SCSI media changer device may vary from server to server
    - We define a link on each server to simplify `/etc/castor/TPCONFIG`

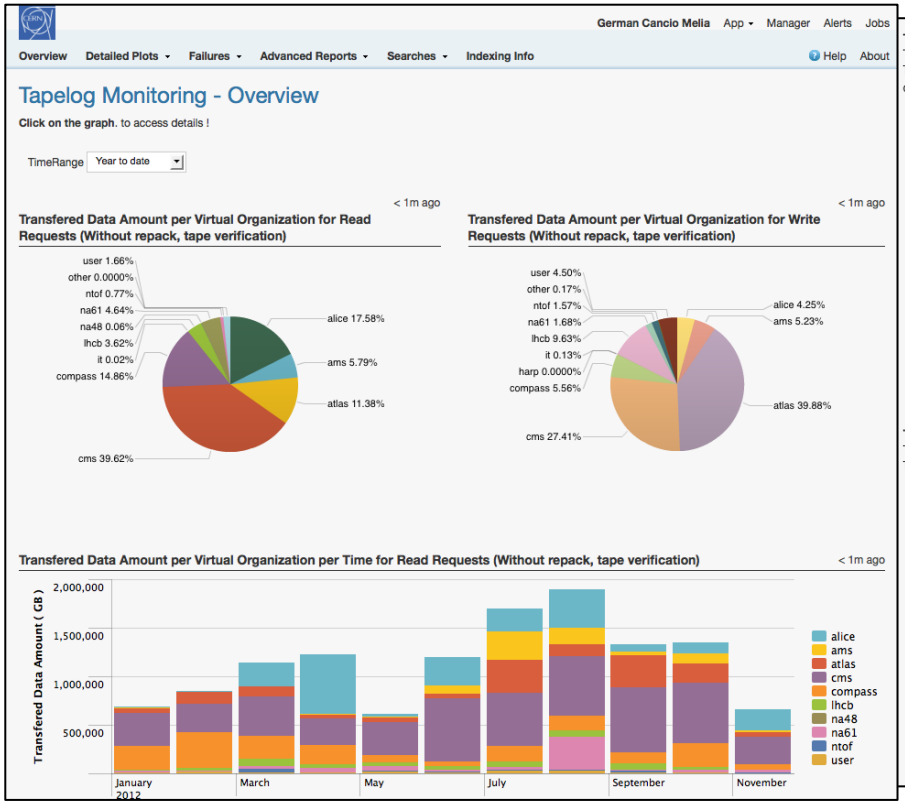
Verification

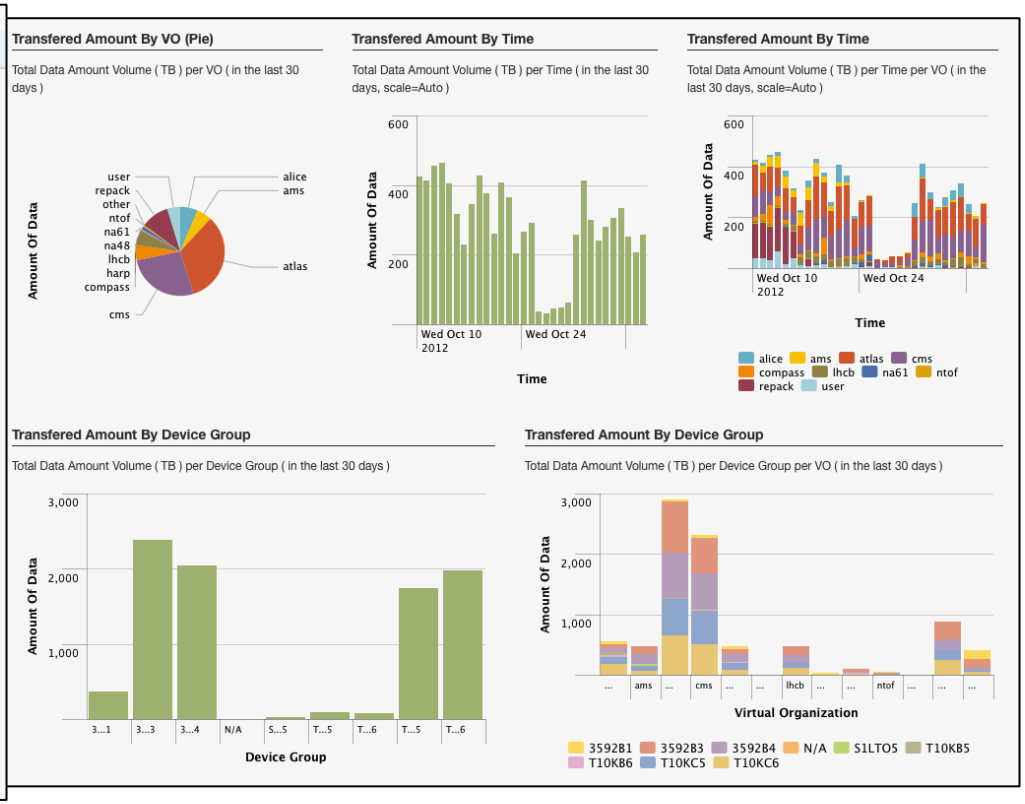
- `tape-scrub`
  - Engine submitting tapes for verification and checking results
  - Uses `readtp -n`
  - Follows the (queued) load
- Checked integrity
  - Raw data
  - CASTOR metadata
- CERN specific but not difficult to port
- Identify problems before the users = minimize data unavailability



Monitoring

- All tape related logs centralized since long time
  - Tape Log APEX application
- New reporting GUI based on Splunk





Demo

- Tape drive for anybody
- Tape Library for VIPs

Conclusion

- Tape Archive capacity will cross 0.1 EB in 2013
- We have recently
  - Improved writing
  - Keep an eye on reading
  - Implemented permanent data verification
  - Enhanced monitoring charts
- Spectra Logic T-Finity certified with CASTOR
- We are ready for growth