CERN Lustre Evaluation and Storage Outlook

Tim Bell
Arne Wiebalck

HEPiX, Lisbon
20th April 2010

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Agenda

- Lustre Evaluation Summary
- Storage Outlook
  - Life cycle management
  - Large disk archive
- Conclusions

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Lustre Evaluation Scope

- **HSM System**
  - CERN Advanced STORage Manager (CASTOR)
  - 23 PB, 120 million files, 1’352 servers
- **Analysis Space**
  - Analysis of the experiments’ data
  - 1 PB access with XRootD
- **Project Space**
  - >150 projects
  - Experiments’ code (build infrastructure)
  - CVS/SVN, Indico, Twiki, …
- **User home directories**
  - 20’000 users on AFS
  - 50’000 volumes, 25 TB, 1.5 billion acc/day, 50 servers
  - 400 million files

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Evaluation Criteria

- **Mandatory is support for ...**
  - Life Cycle Management
  - Backup
  - Strong Authentication
  - Fault-tolerance
  - Acceptable performance for small files and random I/O
  - HSM interface
- **Desirable is support for ...**
  - Replication
  - Privilege delegation
  - WAN access
  - Strong administrative control
- **Performance was explicitly excluded**
  - See the results of the HEPiX FSWG

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Compliance (1/3)

- **Life cycle management**
  - **Not OK: no support for live data migration, Lustre or kernel upgrades, monitoring, version compatibility**
- **Backup**
  - **OK: LVM snaphots for MDS plus TSM for files worked w/o problems**
- **Strong Authentication**
  - **Almost OK: Incomplete code in v2.0, full implementation expected Q4/2010 or Q1/2011**

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Compliance (2/3)

- **Fault-tolerance**
  - **OK: MDS and OSS failover (we used a fully redundant multipath iSCSI setup)**
- **Small files**
  - **Almost OK: Problems when mixing small and big files (striping)**
- **HSM interface**
  - **Not OK: Not supported yet, but under active development**

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Compliance (3/3)

- **Replication**
  - **Not OK: not supported (would help with data migration and availability)**
- **Privilege delegation**
  - **Not OK: not supported**
- **WAN access**
  - **Not OK: may become possible once Kerberos is fully implemented (cross-realm setups)**
- **Strong administrative control**
  - **Not OK: pools not mandatory, striping settings cannot be enforced**

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Additional thoughts

- **Lustre comes with (too) strong client/server coupling**
  - Recovery case
- **Moving targets on the roadmap**
  - Some of the requested features are on the roadmap since years, some are simply dropped
- **Lustre aims at extreme HPC rather then a general purpose file system**
  - Most of our requested features are not needed in the primary customers‘ environment

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Lustre Evaluation Conclusion

- **Operational deficiencies do not allow for a Lustre-based storage consolidation at CERN**
- **Lustre still interesting for the analysis use case (but operational issues should be kept in mind here as well)**
- **Many interesting and desired features (still) on the roadmap, so it's worthwhile to keep an eye on it**
- **For details, see write up at https://twiki.cern.ch/twiki/pub/DSSGroup/LustreEvaluation/CERN_Lustre_Evaluation.pdf**

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Agenda

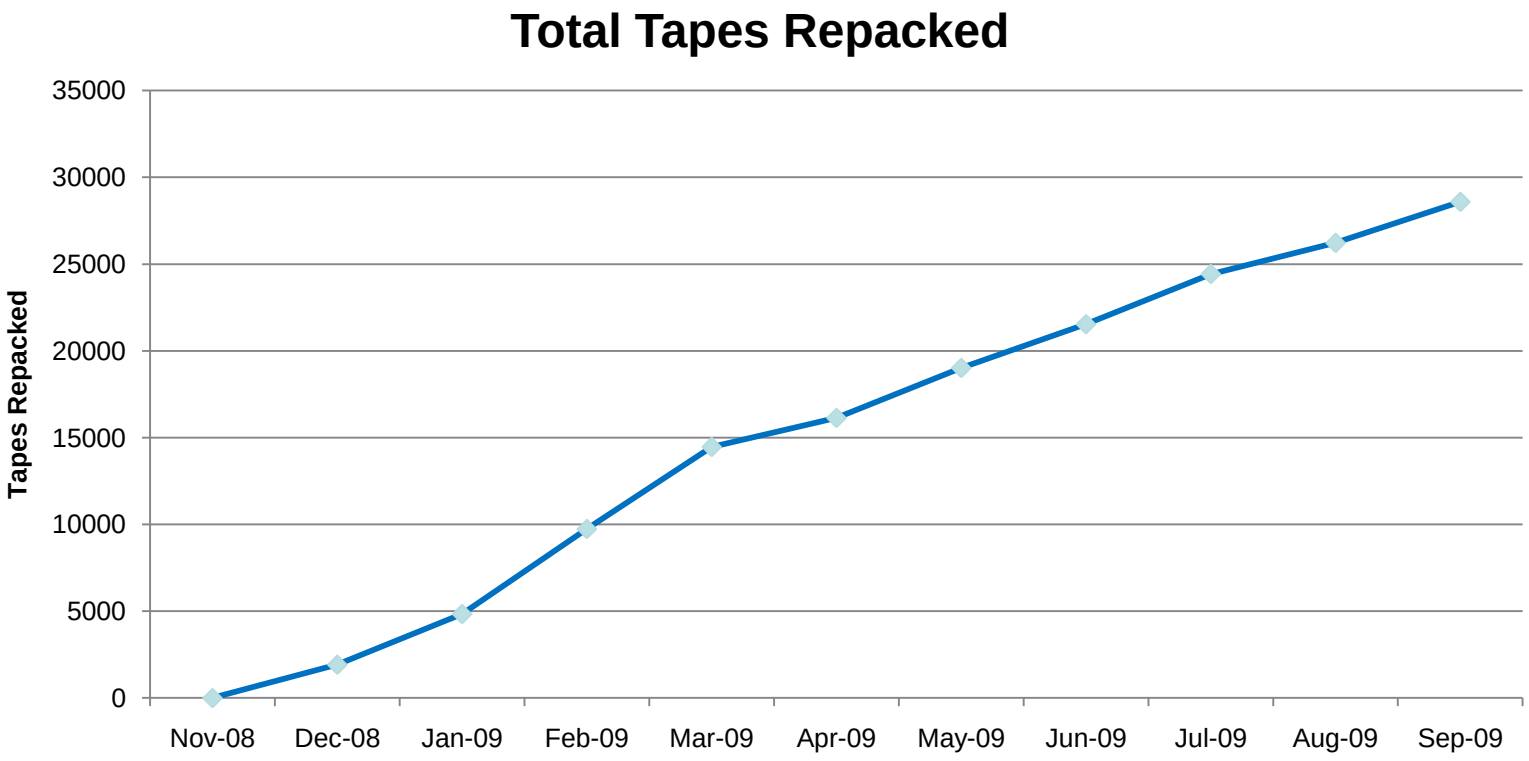
- Lustre Evaluation Summary
- **Storage Outlook**
  - **Life cycle management**
  - **Large disk archive**
- Conclusions

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Life Cycle Management

- Archive sizes continue to grow
  - 28 PB tape used currently at CERN
  - 20 PB/year expected
- Media refresh every 2-3 years
  - Warranty expiry on disk servers
  - Tape drive repacking to new densities
- Time taken is related to
  - **TOTAL** space, not new data volume recorded
  - Interconnect between source and target
  - Metadata handling overheads per file
- Must be performed during online periods
  - Conflicts between user data serving and refresh

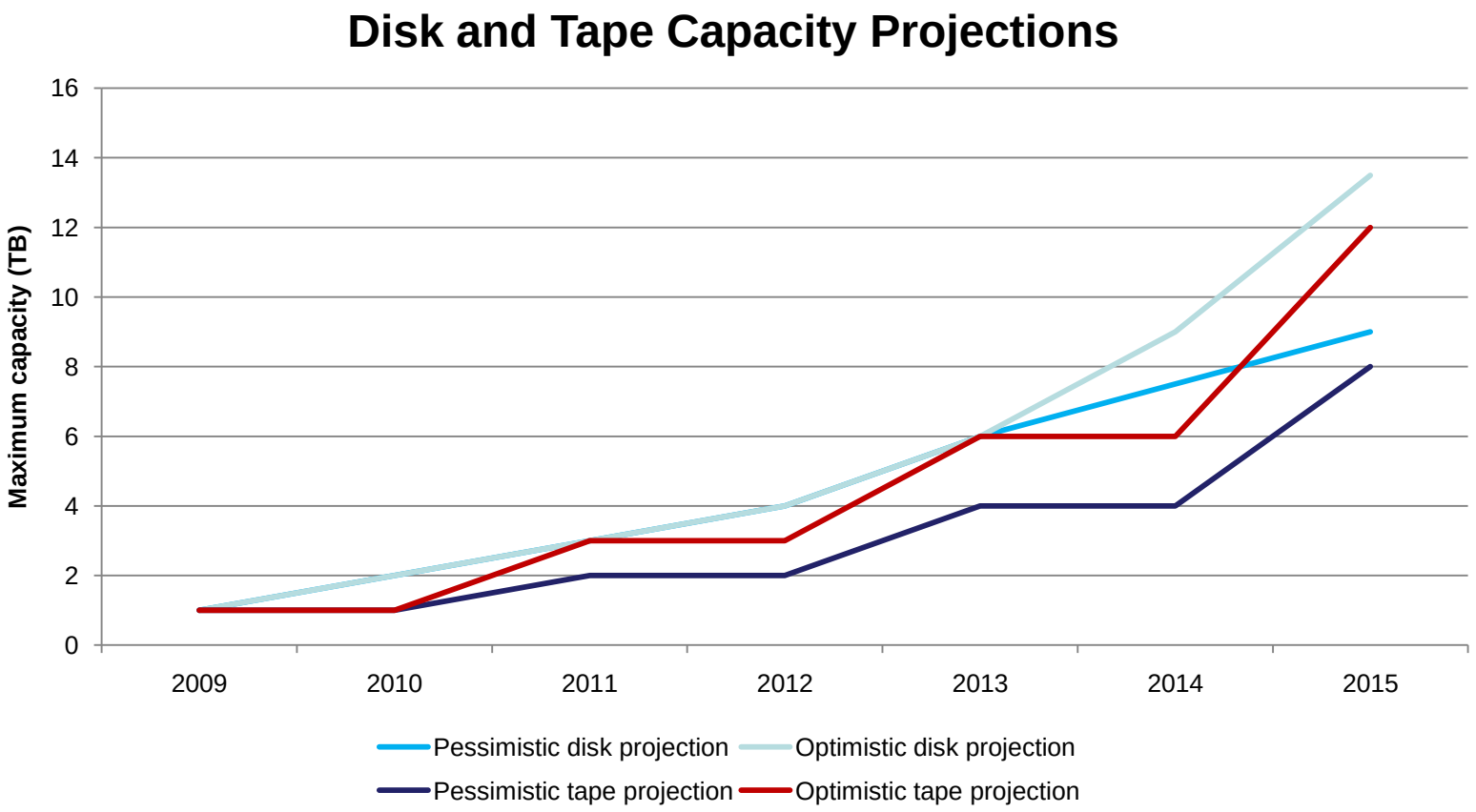
Repack Campaign



- Last repack campaign took 12 months to copy 15PB of data
- When next drives are available, there will be around 35PB of data
- To complete repack in 1 year, data refresh will require as much resources as LHC data recording
- This I/O capacity needs to be reserved in the disk and tape planning for sites with large archives

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

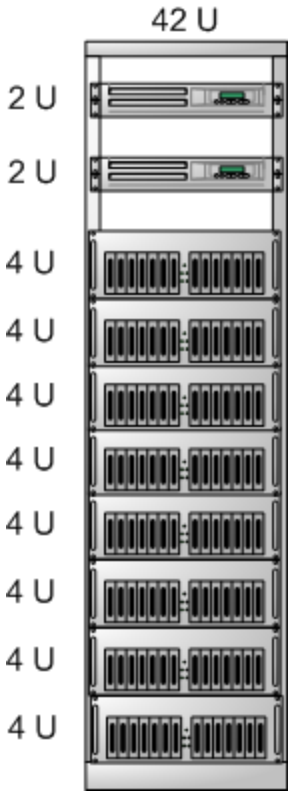
Disk Based Archive?



•Can we build a disk based archive at reasonable cost compared to a tape based solution?

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Storage in a Rack



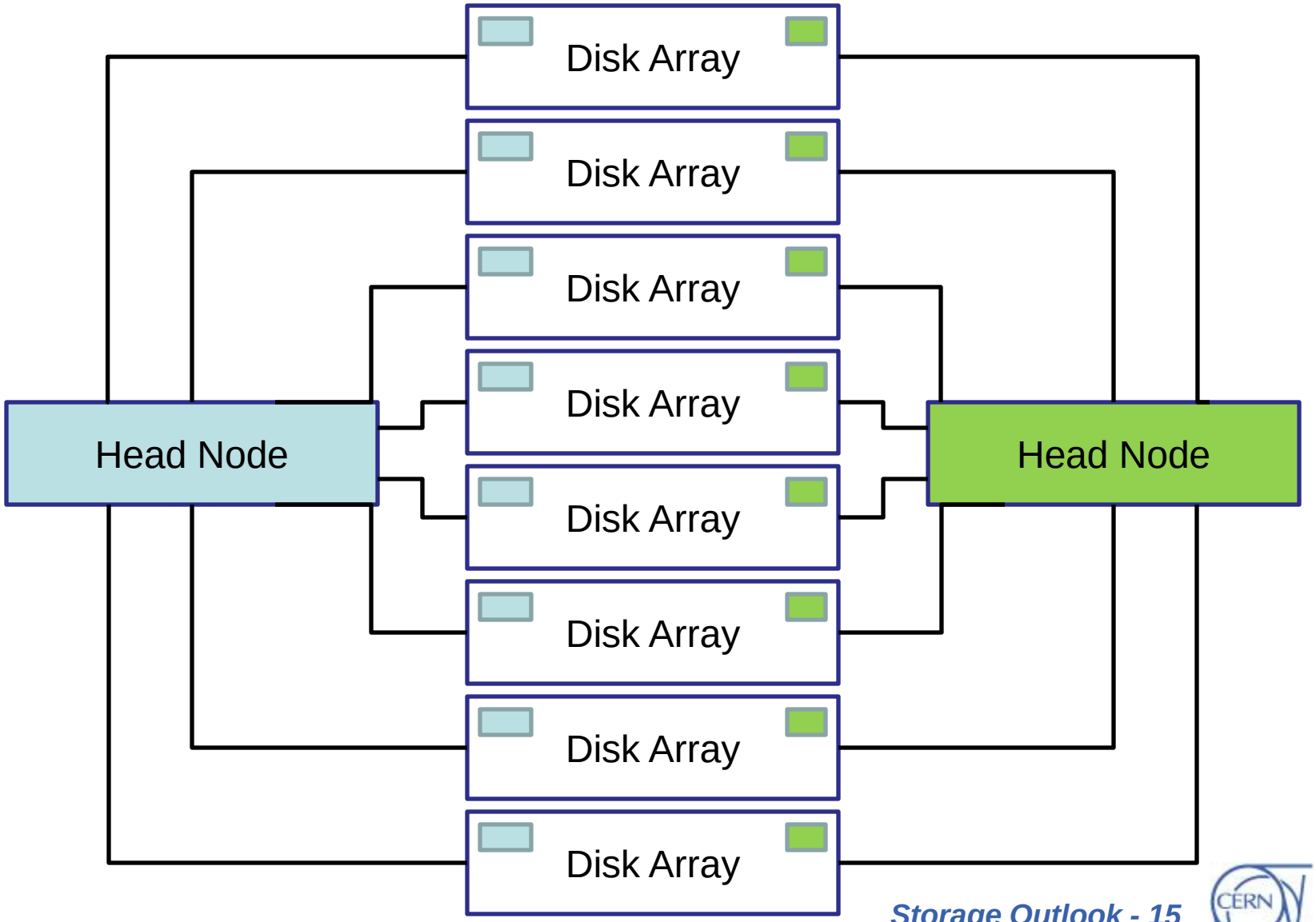
- Tape Storage at CERN
  - 1 drive has 374 TB storage
  - Average rate 25 MB/s
- Disk Server equivalent
  - 2 head nodes
    - 2 x 4 port SAS cards
  - 8 JBOD expansion units
    - 45 x 2 TB disks each
  - Capacities
    - 720 TB per rack
    - 540 TB when RAID-6 of 8 disks
    - 270 TB per head node

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

DSS

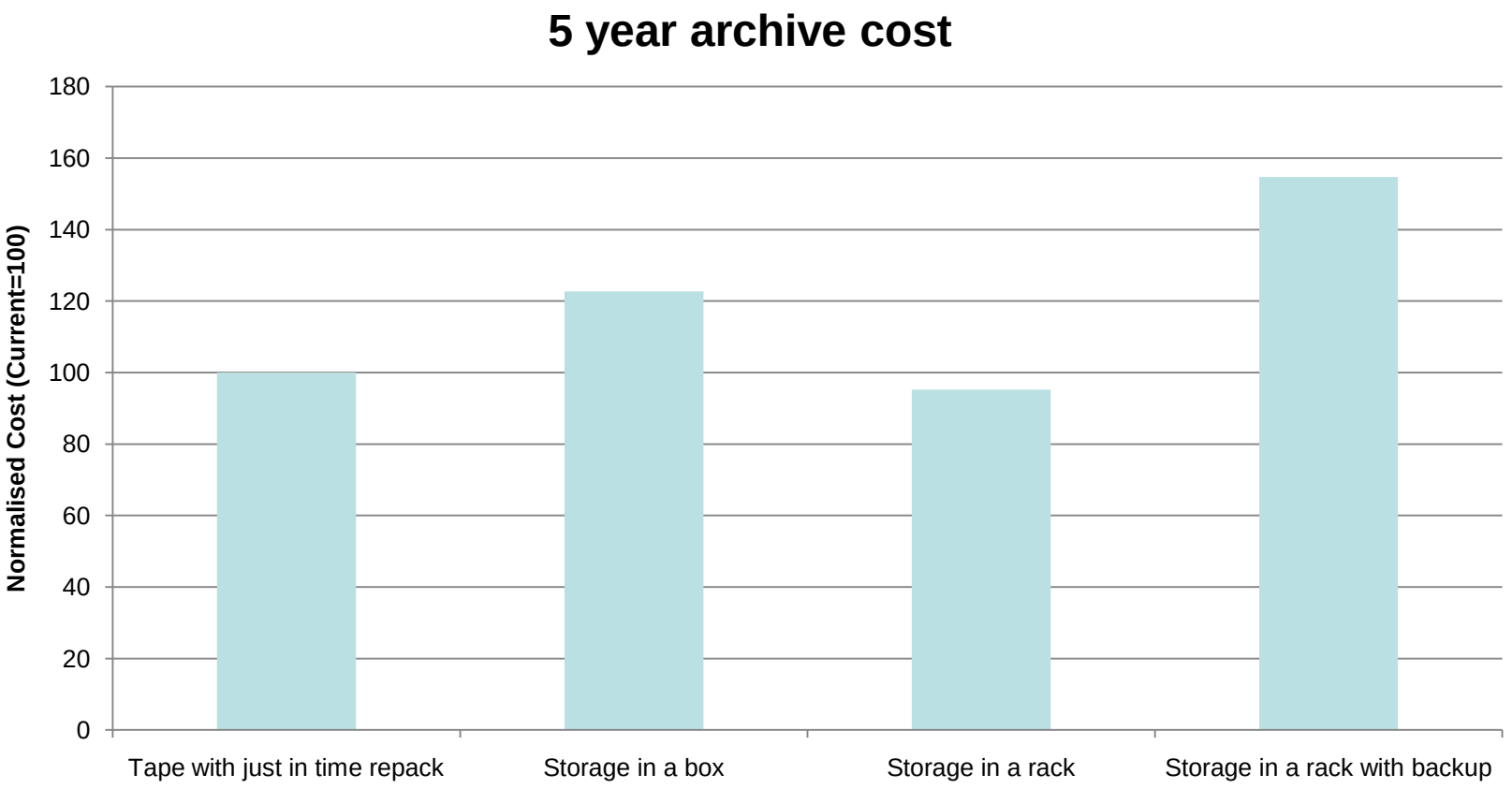
High Availability

CERN IT Department



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

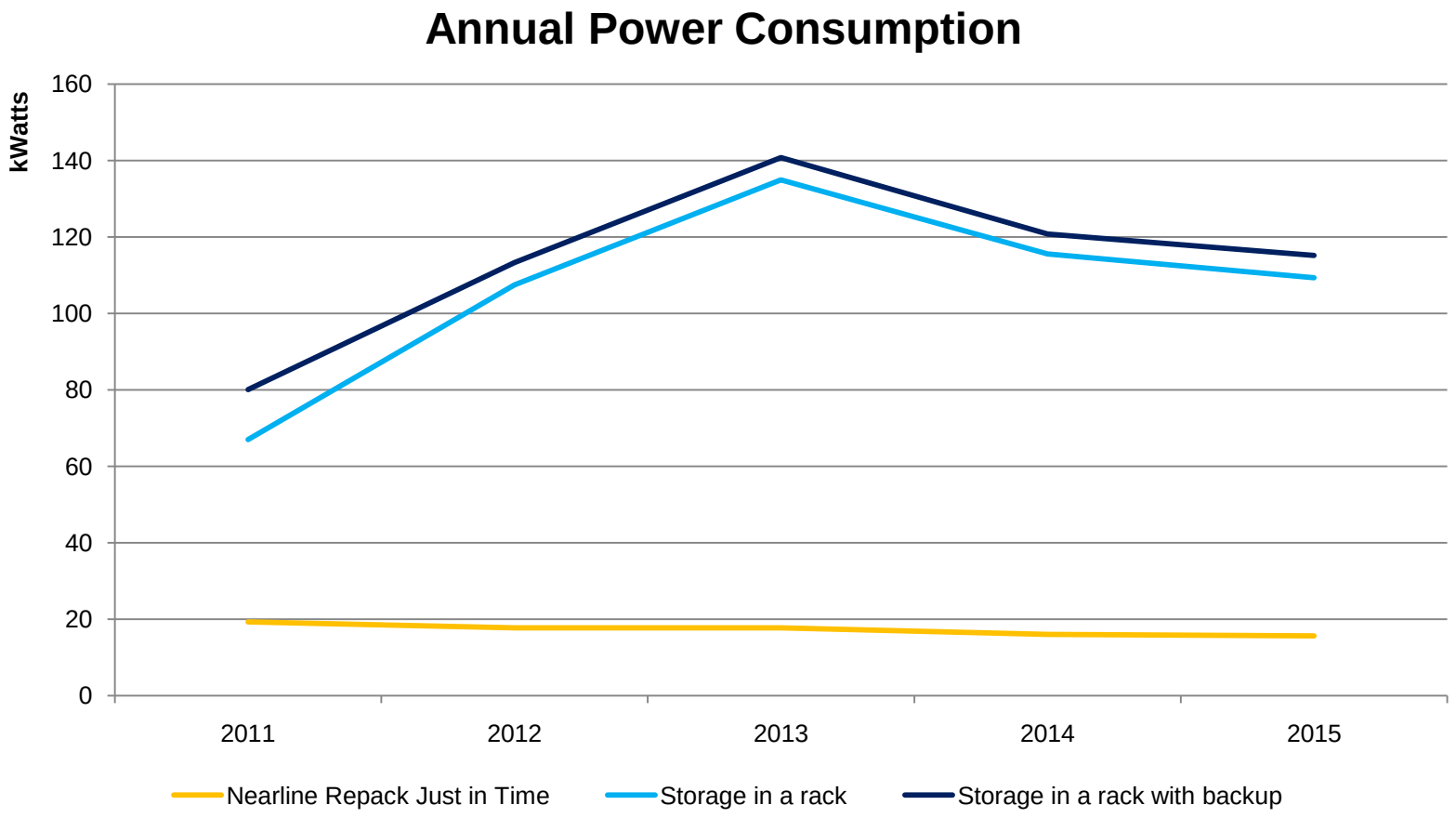
Simulation 20 PB/yr 2011-15



- Costs normalised to tape HSM as 100
- Storage in a rack can be comparable with tape on cost/GB

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Simulation for power



- Additional power consumption of 100 kWatt
- Cost included in the simulation

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Areas to investigate

- Reliability
  - Corruptions
  - Scrubbing
- Availability
  - Fail-over testing
- Power conservation
  - Disk spin down / up
- Lifecycle management
  - 40 days to drain at gigabit ethernet speeds
- Manageability
  - Monitoring, Repair, Install
- Operations cost
  - How much effort is it to run

Conclusions

- Lustre should continue to be watched but currently is not being considered for Tier-0, Analysis or AFS replacement
- Lifecycle management is a major concern for the future as the size of the archive grows
- Disk based archiving may be an option
  - In-depth reliability study before production
  - Watch trends for disk/tape capacities and pricing
  - Adapt software for multiple hierarchies

Backup Slides

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Use Cases

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it