HEPiX Prague 2012

CESNET

## HEPiX Spring 2012

# Building the Czech national storage facility

Jiří Horký
(jiri.horky@cesnet.cz)

MINISTRY OF EDUCATION, YOUTH AND SPORTS

EUROPEAN UNION
EUROPEAN REGIONAL DEVELOPMENT FUND

**24/4/2012**

2007–13
OP Research and Development for Innovation

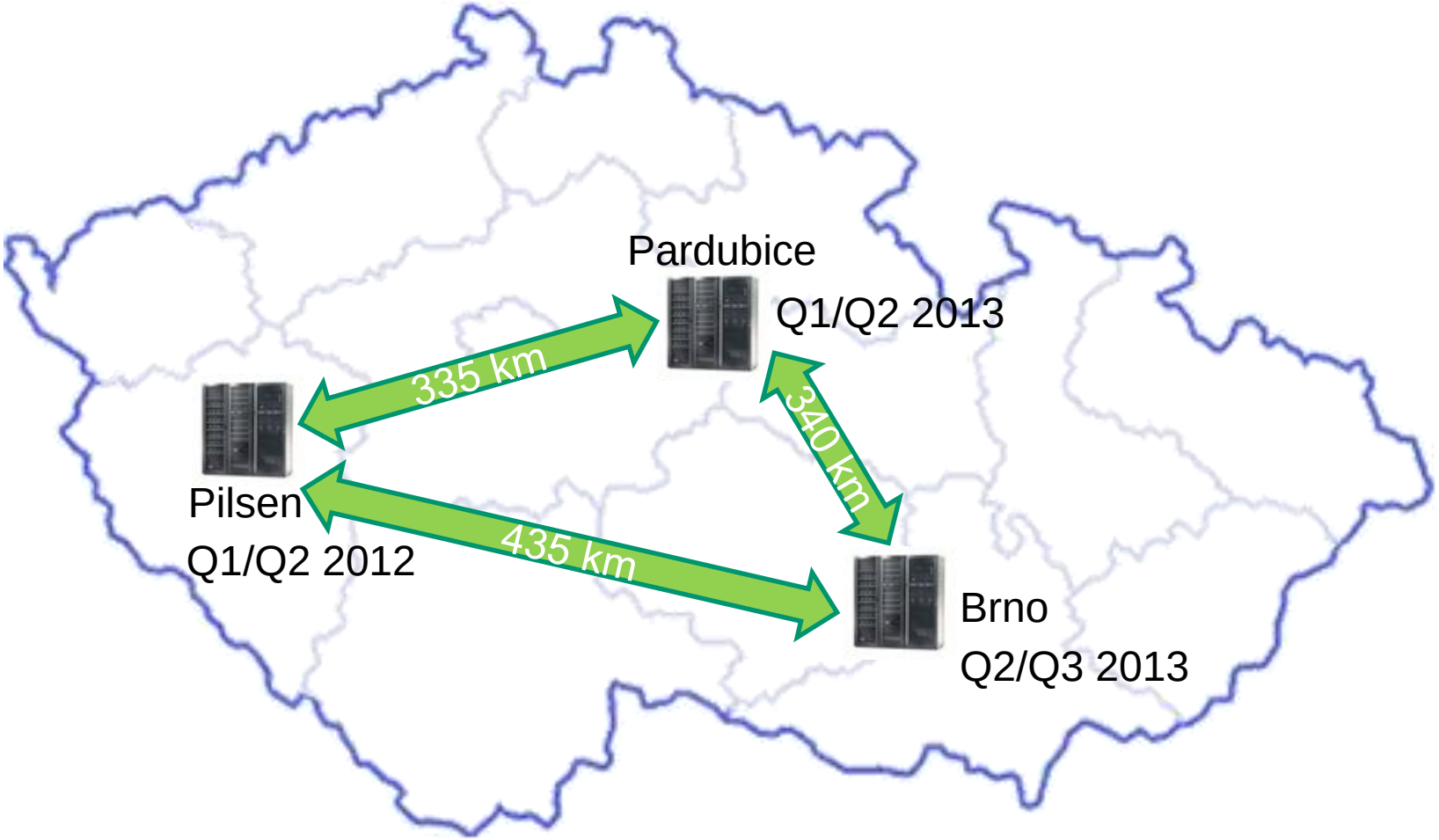
# CESNET

- CESNET is the Czech NREN (Research and Education Network provider)
- Not only a plain network provider, it coordinates and builds e-infrastructure services nowadays:
  - Metacentrum – NGI_CZ, active in EGI, EMI
  - AAI infrastructure – PKI, federalized authentication system (eduID.cz), ....
  - collaborative tools (videoconferencing, VOIP)
  - Data storage facility – this presentation
  - and others...

# Data storage facility

- Three geographically separated storage locations (Pilsen, Pardubice, Brno)
- Total capacity 15-20 PB
- Designed for science and research community
  - Large research projects
  - Public universities
  - Academy of Sciences
  - Public Libraries
  - Digitalization of books
  - ....

# Distributed storage locations



# Financing and scale

EU structural funds + projects of Ministry of Education, Youth and Sports

- VaVpI - project “eIGeR”
- 100 mil. CZK ~ 4 mil. EUR (infrastructure only)
- May 2011 – October 2013

- Project “Large infrastructure”
  - 2011 - 2015
  - operational costs

- 5 FTEs in total (including tender preparation, user interaction...)

# Use cases

- Long term storage with high redundancy (geographical replicas)
- Individual as well as institutional backups and/or archive
- Input/output data for batch jobs
- Exchange point of scientific data among collaborations using standard file protocols
- Grid Storage element
  - Part of FZU Tier 2 or a separate Tier 3 site

→ Very broad possible usage patterns

# Technical assumptions

- Assumptions:
  - Emphasis on economical aspects + transparent behavior to users -> HSM system
  - Backup and archival demands foreseen -> requirement for a tape library
  - Difference in usage patterns will be covered by migration policies, e.g.:
    - archival – move everything from disk to tape almost immediately
    - input files for batch jobs – migrate only files touched > one month ago

# The Pilsen storage site

- The tender won by SGI
- Multi-tier, managed by DMF (Data Migration Facility), SLES OS, CXFS file system
- Disk systems in two IS4600 arrays:
  - Tier1: 50 TB FC drives
  - Tier2: 450 TB SATA drives
- Tier3: Spectra T-Finity tape library with dual robotics
  - 2200 slots with LTO5 (3.3PB in total), 3600 slots licensed, 4500 slots physically installed, upgradable to much more
- 8Gbit redundant FC SAN, 10 Gbit Ethernet
- 2x10 Gbit connection to CESNET2 network
- 6 frontends managed by Pacemaker in HA cluster
- 2 HSM servers in HA mode,one  system for administration

# The Pilsen storage site

# Access to storage services

- File access – access to the same namespace
  - NFSv4 with strong Kerberos authentication
  - SAMBA v2
  - SCP, SFTP (sshfs)
  - FTPS – vsftpd 2.3.2
  - rsync

# Access to storage services

- Grid Storage Element
  - dCache selected as the best candidate
- Block access
  - iSCSI directly from arrays or reexported from FEs
  - FC possible, but not recommended
  - latency issues
- Future plans
  - central backup service
  - right now, our storage system only used as a backend of existing software of the users

# User management

- One user identity in the whole CESNET e-Infrastructure (data storage group, NGI_CZ, video services...)
- Divide and conquer in practice:
  - users arranged in groups with an admin from the group
  - resource owners negotiate on service quantity and quality (quotas, migration policies...) with the group admin
  - the group admin decides who can join the group and what quality of service each particular user receives

# User management

- We require users to be well verified
  - "somebody we trust has seen user's ID"
  - managed by eduID.cz federation of identities (Shibboleth)
  - then we create a Kerberos principal & password which forms his e-Infrastructure identity

- Grid Storage Element is a special case

# Storage for the physics community



- FZU one of the largest customers
  - actually, they would probably use everything if allowed!
- **dCache** selected as a preferred SE implementation
  - implementation with DMF already on some sites
  - will be used for ATLAS as well as for AUGER
- Part of the Tier-2 site or a separated storage?
  - reliability concerns *vs* easier access to the capacity
  - decision yet to be made

# Storage for the physics community



- xrootd for ALICE
  - just another xrootd server for clusters in Prague?
    - concerns about job effectiveness due to latency (Pilsen <-> Prague = 3.5ms)
  - how to efficiently use it with tape?

# Customers interaction

- Goal: define the customer's needs and provide the proper solution
  - very important!
  - but only very few user groups who expect to have non trivial amount of data have clear idea of what they really need
    - HEP community is quite exceptional in this!
  - often very overestimated initial requirements
- Policy: allocate just a portion of demanded space, transparently grow it if needed (xfs_grow works great!)

# Open questions

- Inter site replication
  - file system based, file triggered (i-notify), maybe even block based?
  - or just as an FTP backend for HSM in each site?
- efficient use of the purchased storage systems for ALICE
- stick with tape libraries or try MAID technologies?
- Long term goal: One namespace across all three sites accessible by standard file protocols
  - is it achievable by the current technology?
  - and with independent tenders for each location?

# Conclusion

- Challenging and unique project in Czech Rep.
- 3.8PB available now, 15PB+ expected by the end of the next year
- Broad range of user demands -> broad range of technical challenges
- First experimental users are coming!

**Thank you for your attention.**
**Questions?**

Jiří Horký
(jiri.horky@cesnet.cz)

MINISTRY OF EDUCATION, YOUTH AND SPORTS

EUROPEAN UNION
EUROPEAN REGIONAL DEVELOPMENT FUND

2007–13
OP Research and Development for Innovation

# Technical challenges

- HA cluster deployment
  - 6 FEs with static IPs + 6 floating IP addresses
  - the setup should survive outage of 5 servers
- Floating IP problems
  - Kerberos keytab lookup for new hostnames
    - NFSv4 – “hidden” option of svcgssd daemon
    - sshd – option to make this work in newer versions (not in SLES ☹)
  - Same multi-host certificate on FEs – seems to work fine so far ☺
- True active/active NFSv4 solution?
- CXFS Metadata performance