Present & future solution for data storage at CERN

Andreas-Joachim Peters

for the EOS project team

IT Data & Storage Service Group

CHEP Okinawa 13.-17.4.2015

XRootD
HTTPS
SRM
WebDAV
HTTP OwnCloud
FUSE gridFTP

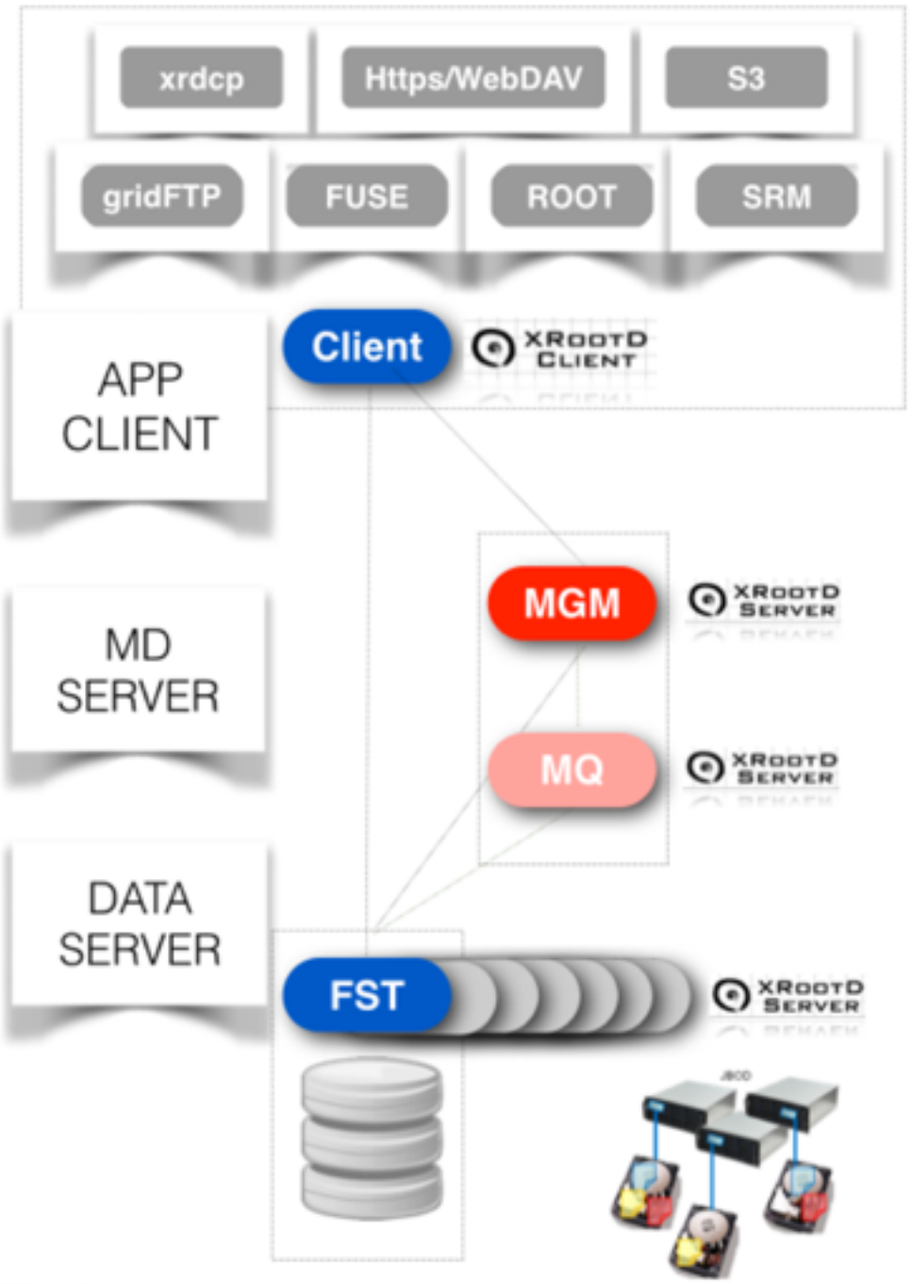
EOS

Disk Storage @ CERN

Overview

- EOS architecture & evolution
- releases
- recent enhancements
- roadmap 2015
- general direction
- summary

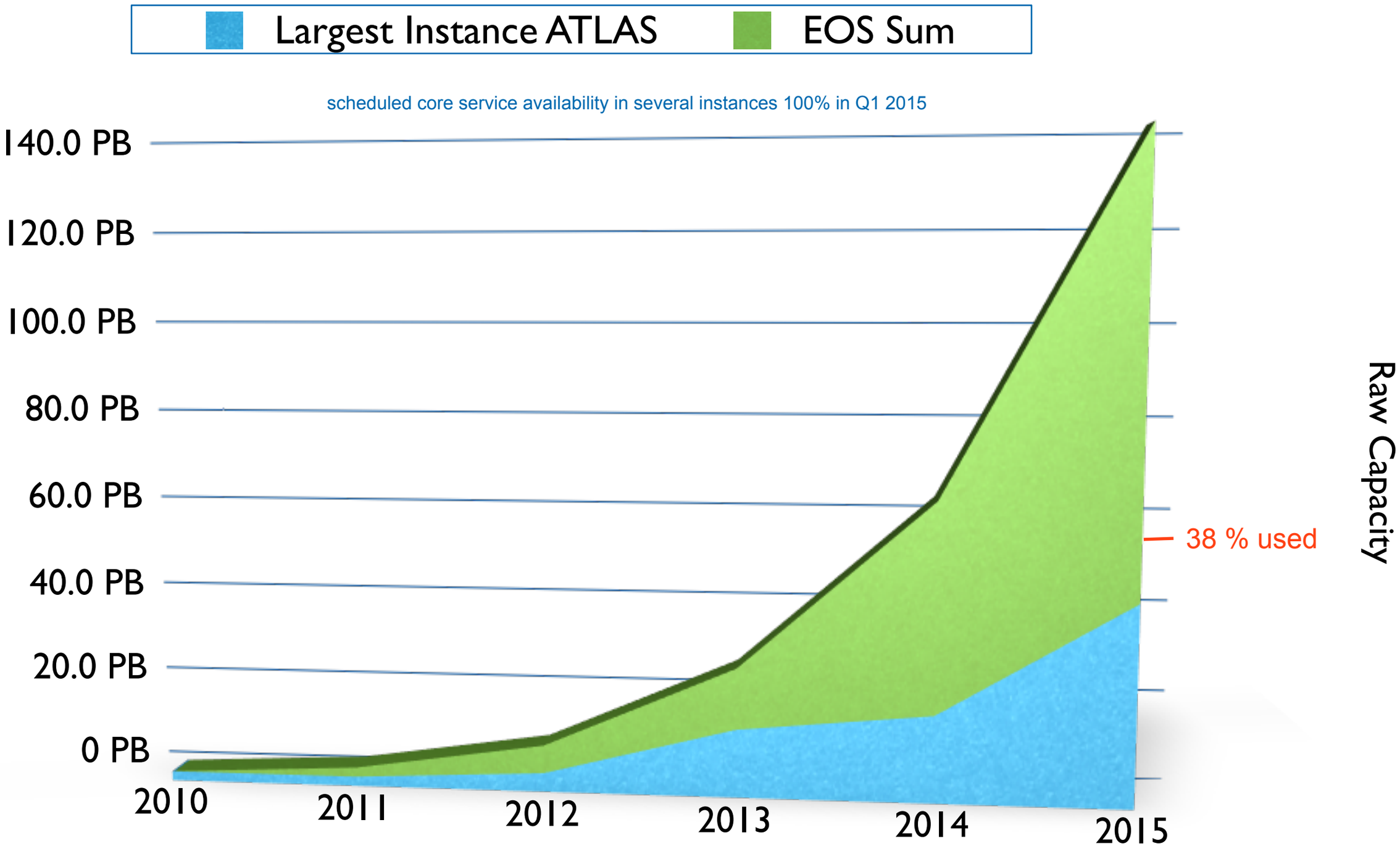
Architecture



- project since 2010
- production since 2011
- simple - license-free - JBOD hardware
- in-memory namespace
- strong security

  [server side security]
- many protocols
- quota, tunable QoS
- Dev&Ops @ CERN/IT

EOS DEPLOYMENT



EOS STORAGE IN NUMBERS

April 2015	
Capacity	**140 PB**
Server	**1.400**
Hard Disks	**44k**
Files	**271 M**
Directories	**26 M**
Replicas	**0.5 B**
Connectivity [theor.]	**13 Tbit**
random IOPS	**2.2 M**
Disk BW [theor.]	**3.3 TB/s**
Internal Messaging	**150 kHz**
State Machine	**3M kv pairs**
Users storing data	**~3k**
Quota rules	**9.600**

single thread namespace stat rate
160 kHz

multi threaded namespace stat rate
1 MHz

memory footprint **0.5-1 kb/file**

Releases

- Run 2 Production Version
  BERYL Aquamarine >= 0.3.107
  archiving tools: CHEP talk 298

EOSALICE
EOSATLAS
EOSCMS
EOSLHCB
EOSUSER
EOSLHCB

- XRootD V4 based Version, infrastructure aware
  CITRINE 0.4.x

- Ceph based R&D bundle
  DIAMOND
  storage R&D: CHEP talk 297
  Ceph CERN: CHEP talk 287

recent

Enhancements

- multi-platform **WEBDAV** access
  - Android
  - IOS
  - OSX
  - Windows

CERNBOX: CHEP talk 327

- **secondary group** support in ACLs
- **sticky ownership** - owner/group conservation for *sync&share*
- **symbolic attributes** - inherit extended directory attributes via attribute link
  [lower memory consumption - 100 bytes less per directory]
- **online compactification** for directory change-log files
  [faster boot time]

recent

Enhancements

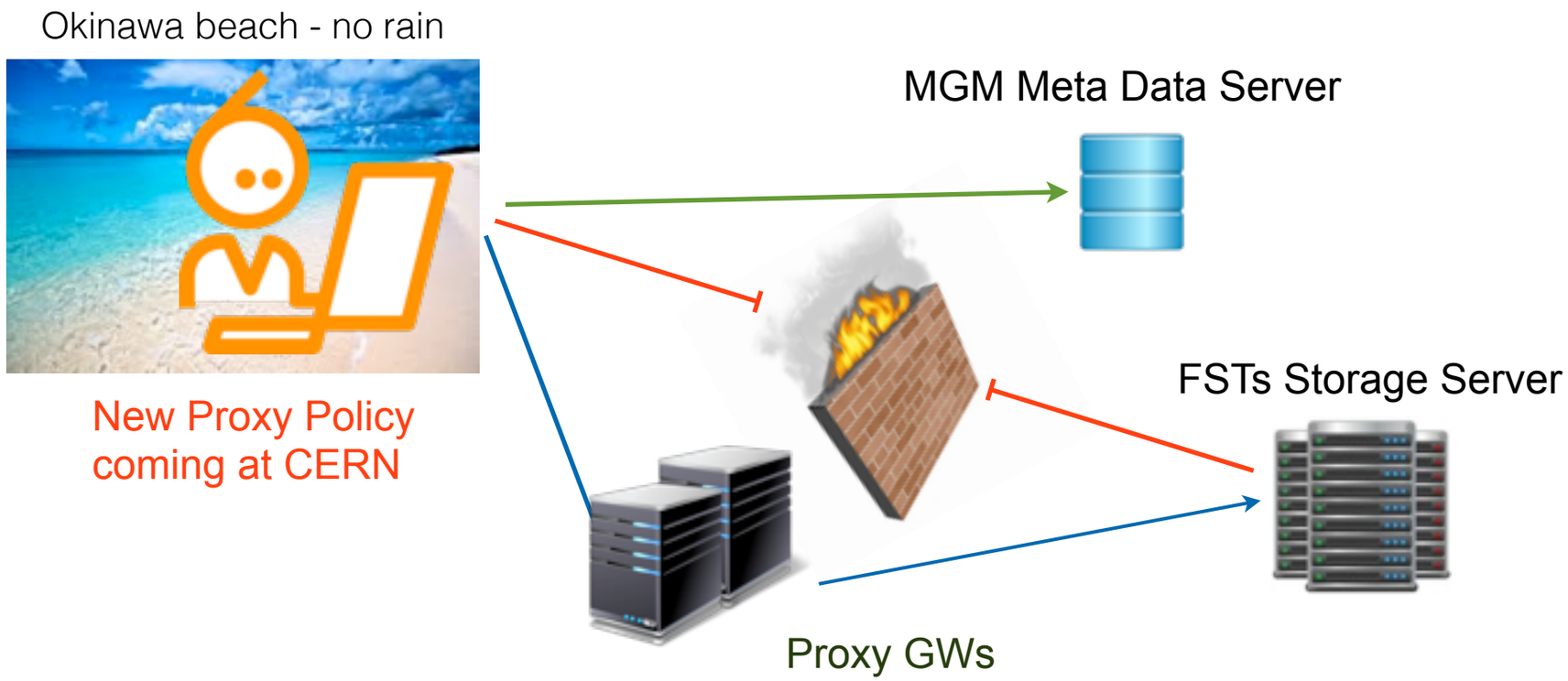
- **FUSE** mount **rm -r** detection and prevention
  - can block recursive deletions issued from a shell if the command is issued too high in the directory tree e.g. block rm -rf /eos/

- multi-user **FUSE** mount supporting user private kerberos and X509 **authenticated connections**
  - user mapping independent of uid/gid of calling process
  - planning to do secure cern-wide EOS mount

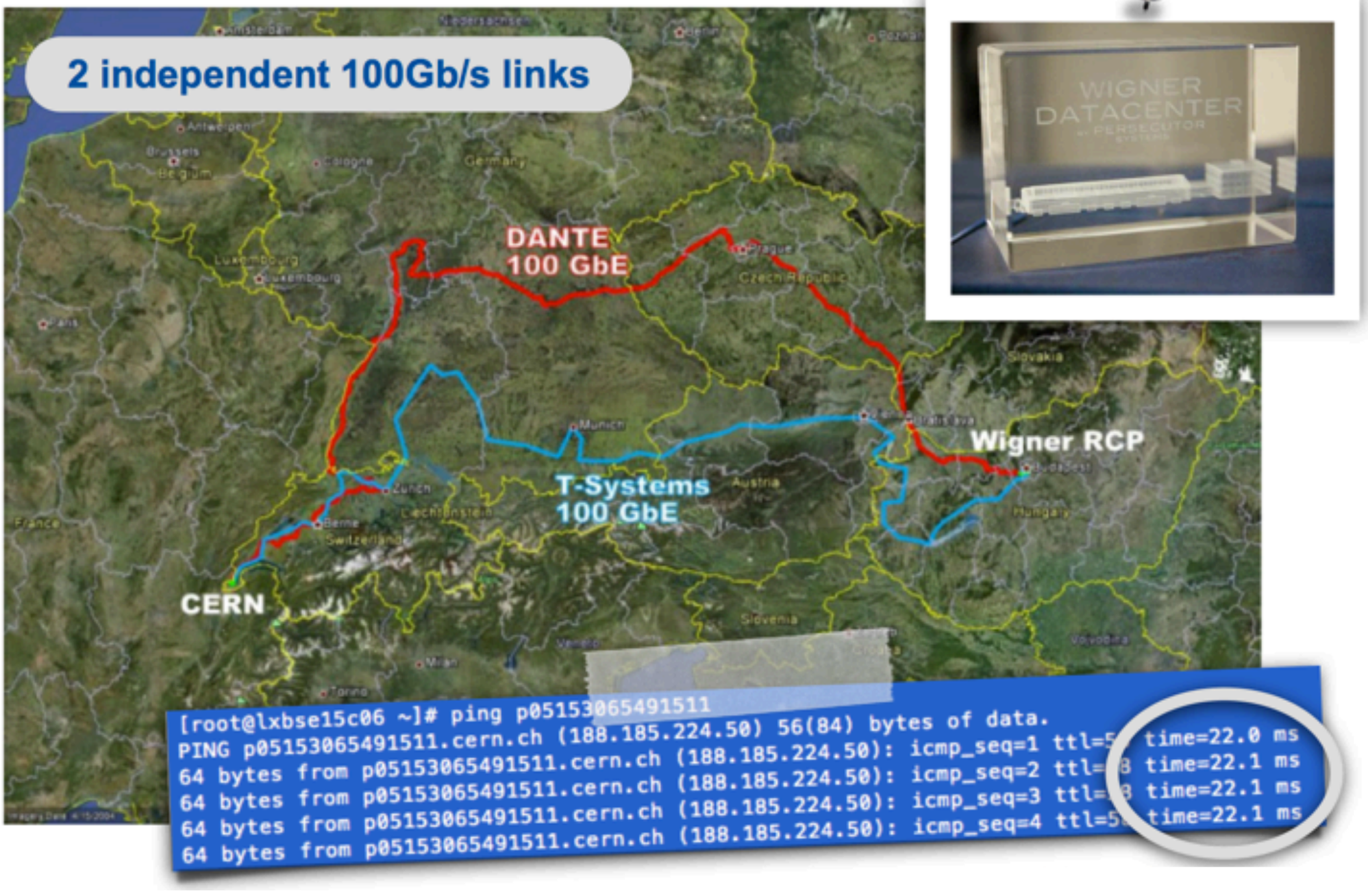
recent

Enhancements

- location defined **IO proxy**
  - re-route all non-metadata traffic coming to/from machines outside of Meyrin CC through a generic proxy service



Wigner Computer Centre



22ms latency

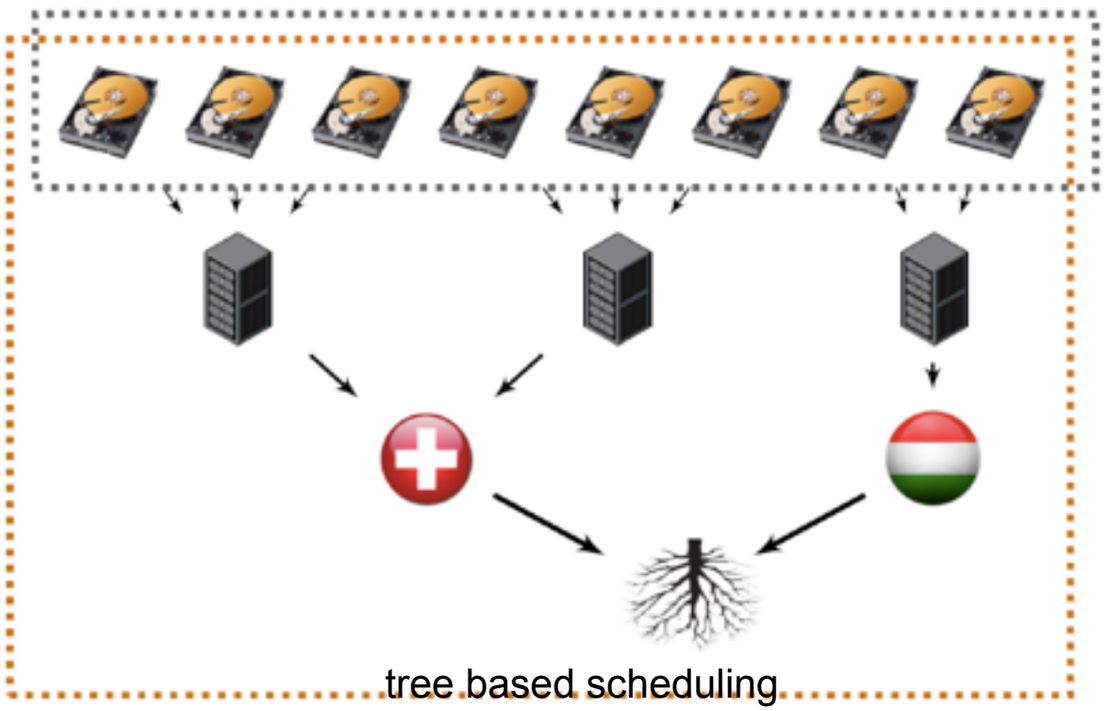
Infrastructure Aware Scheduling

- Meyrin & Wigner CC 22ms network latency
  - main strategy placing one replica in each CC
  - batch jobs are tagged with their GEO location and access is preferring local replica

- CITRINE defines three placement strategies
  - scattered 

e.g max. distance regarding <site>
  - hybrid 
e.g min. distance regarding <site>, but max. distance in <rack>
  - co-located 

e.g min. distance regarding <site>

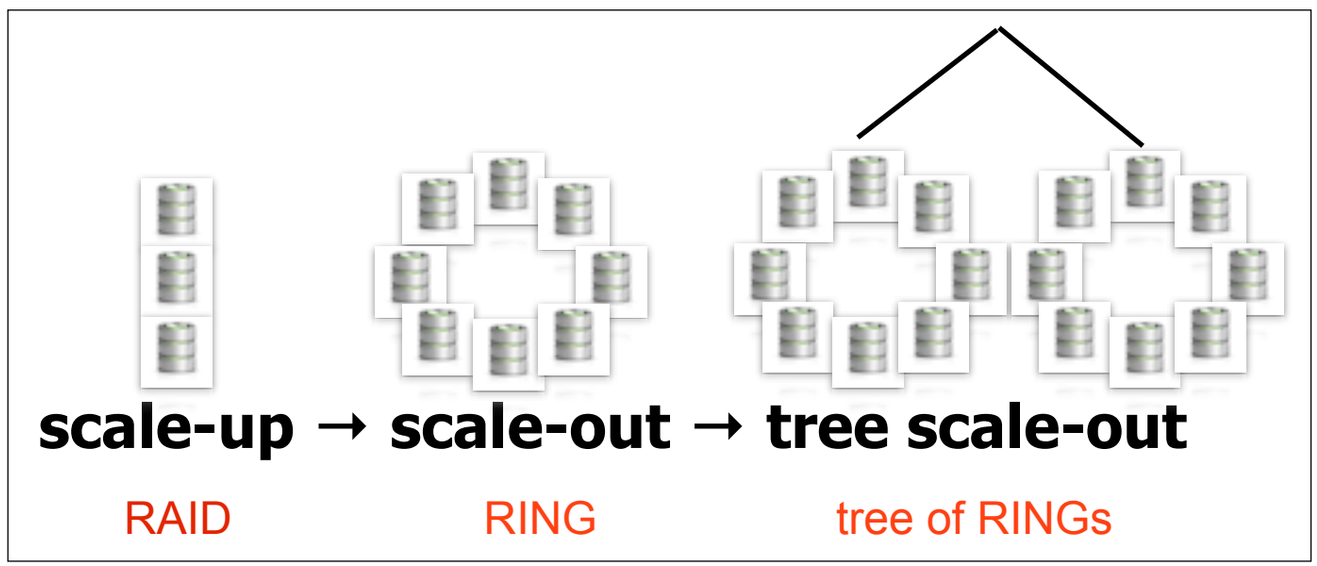


Roadmap 2015

- eliminate namespace boot time
  - **approach 1** → move in-memory namespace to **in-memory maps/vectors** with real-time persistence [`mmap index +WAL`]
  - **approach 2** → cache part of namespace in-memory and persist in external storage - **kv DB** [ArDB/ForestDB] or **object store** [rados]
- IO plug-ins
  - **kinetic** → ethernet drives - libkinect_client : Openlab project
  - **rados** → CEPH pools - using object striping/chunking libraries (libradosstriper & libradosfs)
  - **xrootd** → remote XRootD cluster - using XRootD client
  - **...**
- mount external trees into EOS namespace [e.g. S3 buckets]

General Direction

- **integration of open solutions**
  - **kinetic, rados**
  - **kv stores**
  - more **HTTP, WebDAV, S3, *Sync&Share***
- rely on **trivial scalability** for extremely large clusters
  - don’t scale too big instances
  - deploy several storage instances in a flat tree



Summary

- EOS provides a very flexible **disk storage** platform for a **large user community**
  - about 3k users storing data today
  - integrated in T0 workflow by ATLAS & CMS
- demonstrated **unprecedented scalability**
  - largest low-cost HEP storage installation site today with 140 PB and 44k disks
- **strategic direction** for CERN based disk storage
  - for physics data (user/group/grid)
  - as 'new-style' home directory via CERNBOX
- current & future **challenges**
  - guarantee smooth & stable Run-2 operation
  - prepare for future scale and use-cases

Thank you!

Questions or Comments?