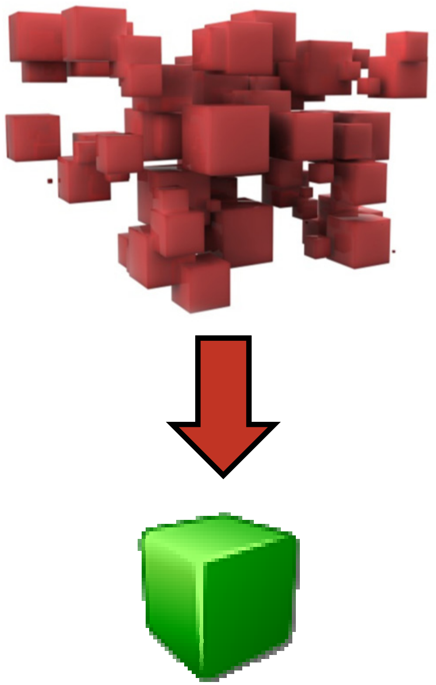
Understanding the EUBOX Service

Towards a trusted ‘Dropbox4Science’

Morris Riedel et al.

Juelich Supercomputing Centre



Track 1 – EUDAT Services
EUDAT 2nd Conference
28.10. – 29.10.2013, Barcelona

SEVENTH FRAMEWORK PROGRAMME

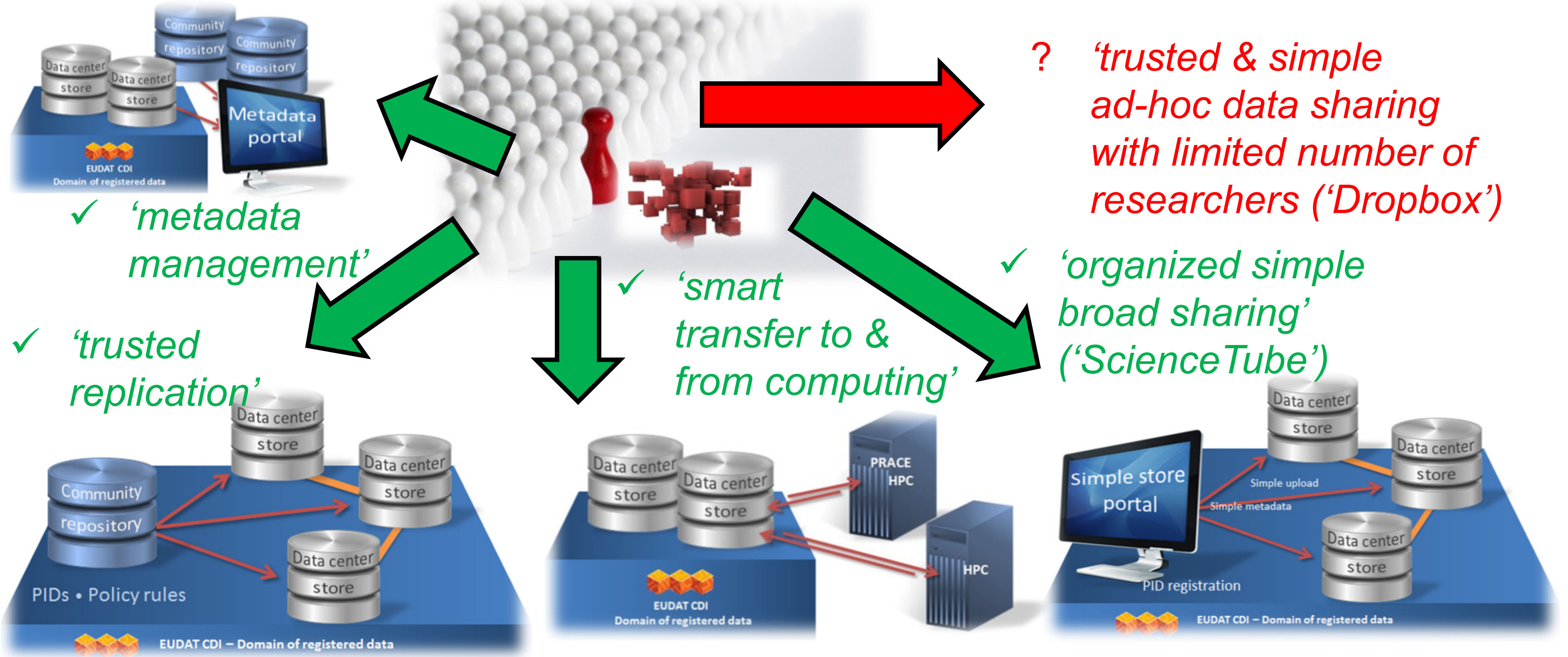
EUDAT

Outline

- Motivation & Goals
- Identified Use Cases
- Analysed Requirements
- Comparison to SimpleStore
- Experimental Setups
- Summary

Motivation (1)

- Establish ‘trusted user experiences’ like a ‘dropbox 4 science’



✓ ‘metadata management’

✓ ‘trusted replication’

✓ ‘smart transfer to & from computing’

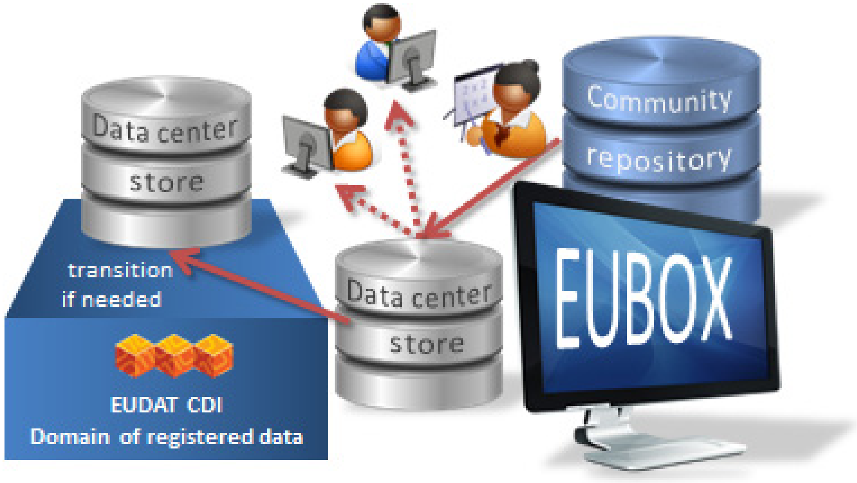
✓ ‘organized simple broad sharing’ (‘ScienceTube’)

? ‘trusted & simple ad-hoc data sharing with limited number of researchers (‘Dropbox’)

EUDAT

Motivation (2)

- Enables easy and ad-hoc (temporary) sharing of research data
  - Circulate data among a couple of research colleagues
  - Access also for non EUDAT
  - Synchronization of data



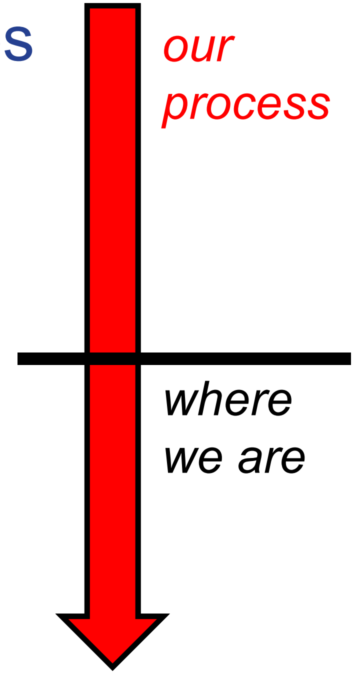
- Offers a seamless transition to the 'EUDAT registered domain' of data
  - Publicly usable open locations
  - Stored only optionally for a long-term period

'Design Phase: Figure not confirmed yet'

Goals of the ‘EUBOX Task Force’

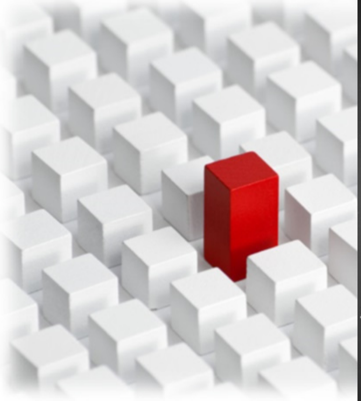
- Exploring solutions towards an EUDAT EUBOX service
  - Major goal ‘User experience is key to the acceptance of the service’

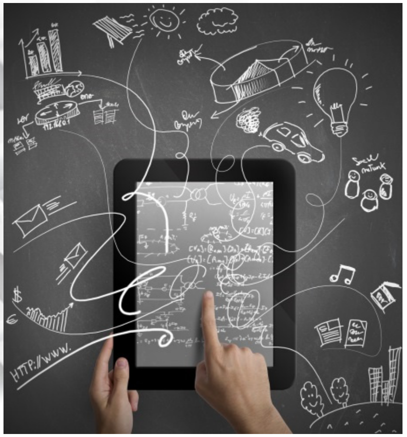
- ✓ Documenting use cases from user communities
  - ✓ Identify derived requirements and constraints
- Analysing various existing programs
  - Gather lessons learned from (test) deployments
  - Comparison matrix with required product features
- Choosing technology/technologies
  - Recommendations that fit user communities best



Selected Community Requirements

- ‘Some research organizations do not allow dropbox’ (e.g. German research organization Max Planck)
  - Trust issues, but an alternative must be as mature as dropbox

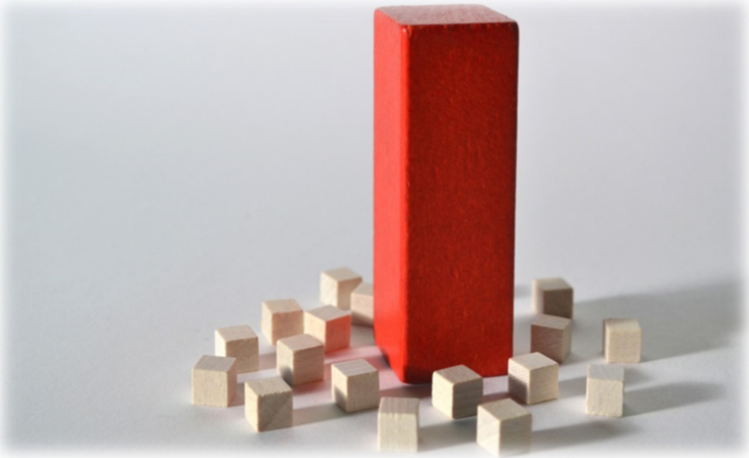


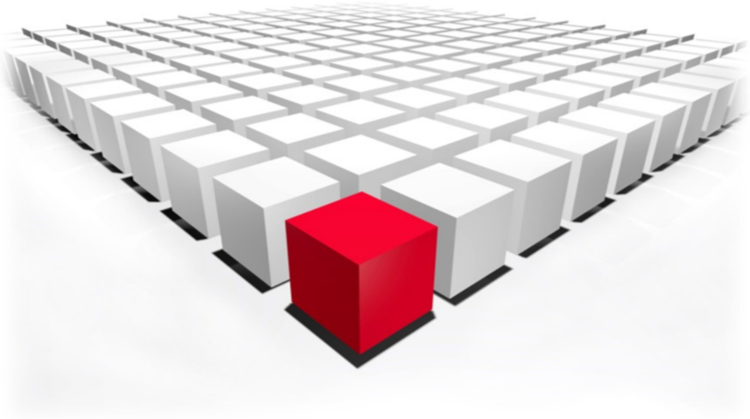


- ‘Simple, secure, and sound’
  - Usable also with mobile devices
- ‘Trusted Access & Sharing’
  - Bi-drectional data synchronization

Selected Service Provider Requirements

- ‘load balancing meaning distributed instances can be load balanced across centers’
- ‘scalable meaning additional nodes with backend storage can be added after time’





- ‘we need to be able to make a EUDAT or even user community branding of the visible service elements’

Selected Use Case

EPOS
EUROPEANPLATEOBSERVINGSYSTEM

- Earth Plate Observatory System (EPOS)

one process of many

1. INGV center records data (seimic, GPS, etc.)
2. Real-time seismological data is gathered in different data centers (e.g., Rome, Ancona, Grottaminarda,...)
3. Data gathering for different parts of the Italian peninsula in parallel
4. The data acquired by the different data centers are basically different
5. But there is some 'seismic station redundancy' among data centers
6. The archive is centralised in Rome@INGV and the data gathered in the other centers must be replicated here for the long-term
7. For one common/overlap area where all the data are stored temporarily for a buffer of say 1 to 10 days it would enable to make all the quality checks before final archiving with all the data handy

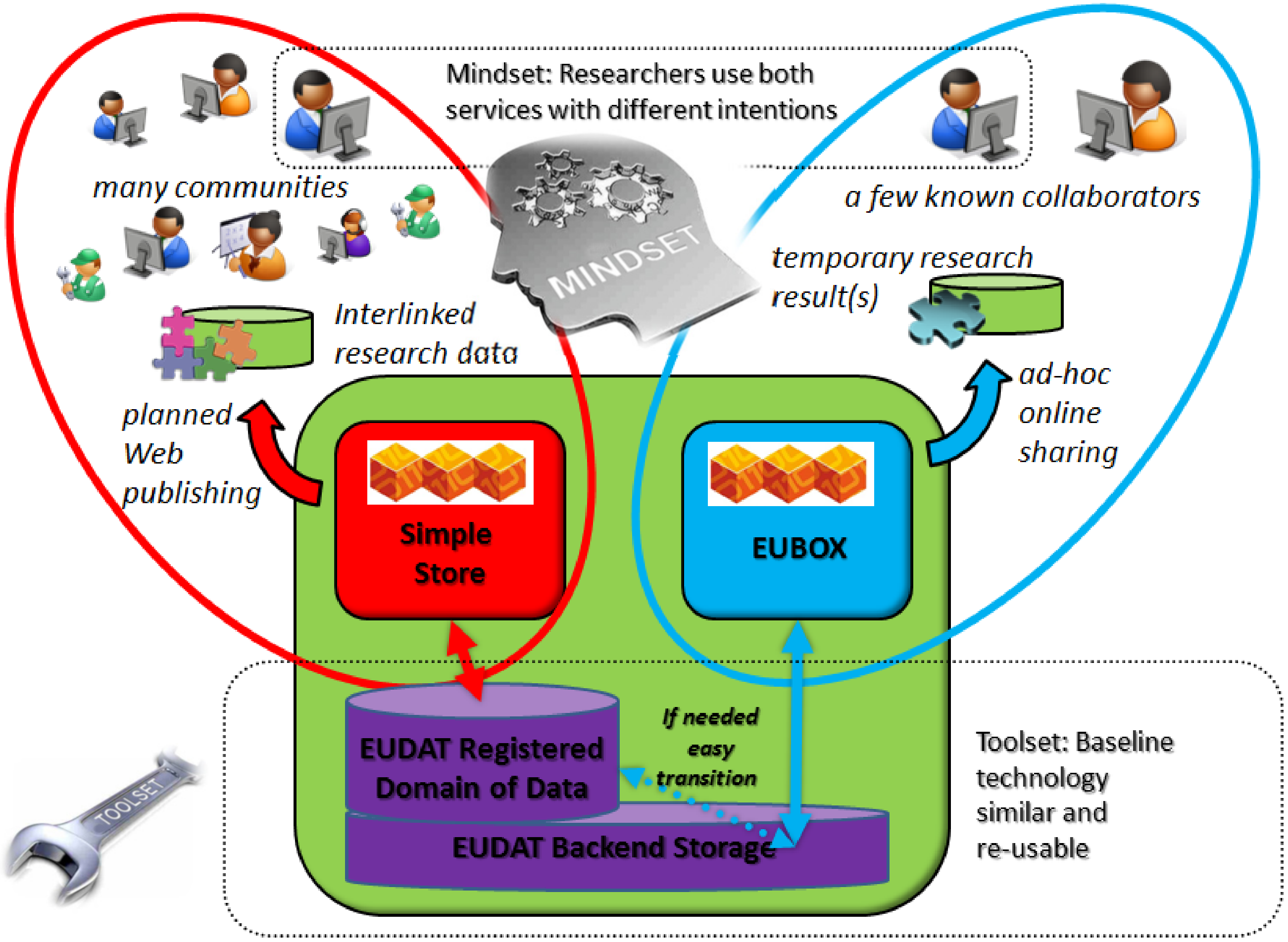
➢ Temporary storing Research data for quality checks (e.g. ingest gap data)

➢ After (manual) checks the research data can be stored permanently

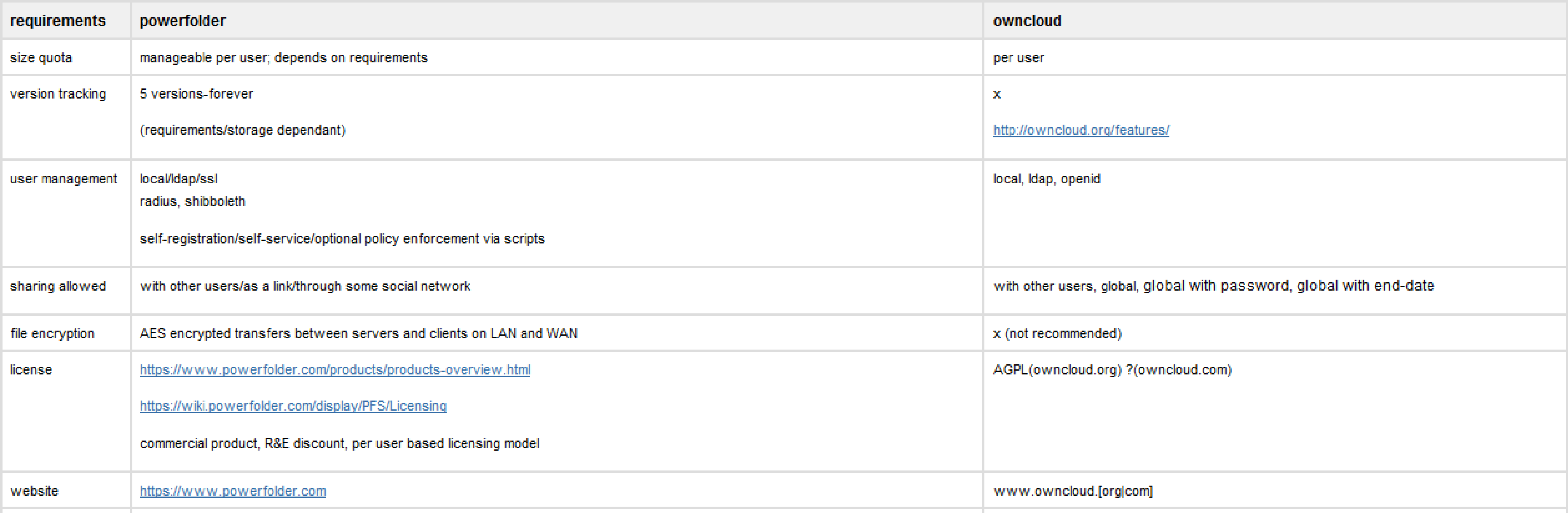
With thanks to Alberto Michelini (INGV, EPOS community)

EUDAT

Comparisons to SimpleStore Service


Mindset: Researchers use both services with different intentions
many communities
a few known collaborators
MINDSET
Interlinked research data
temporary research result(s)
planned Web publishing
ad-hoc online sharing
Simple Store
EUBOX
EUDAT Registered Domain of Data
If needed easy transition
EUDAT Backend Storage
TOOLSET
Toolset: Baseline technology similar and re-usable

Candidate Technologies & Evaluations (1)



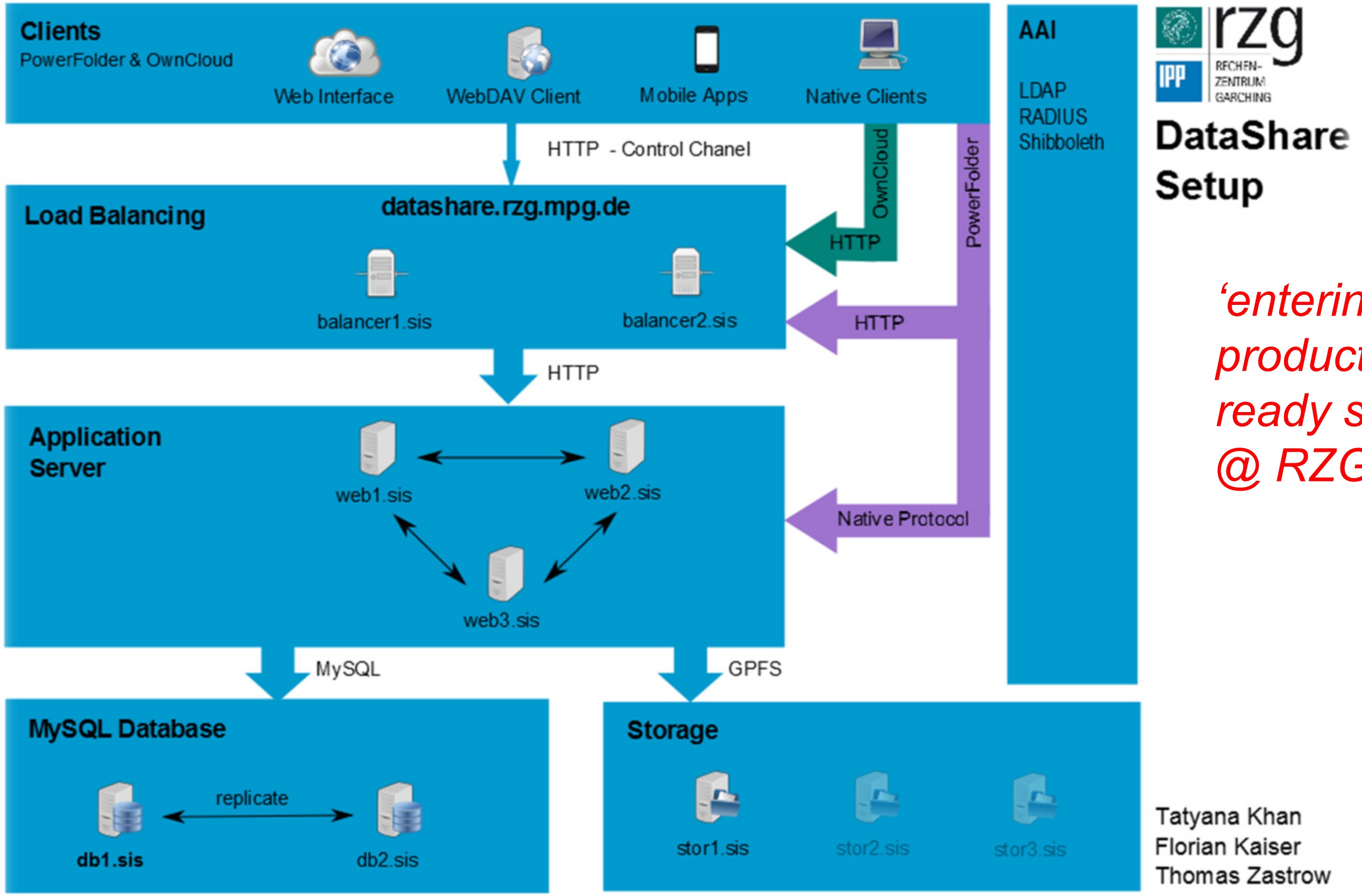
requirements	powerfolder	owncloud
size quota	manageable per user; depends on requirements	per user
version tracking	5 versions-forever (requirements/storage dependant)	x http://owncloud.org/features/
user management	local/ldap/ssl radius, shibboleth self-registration/self-service/optional policy enforcement via scripts	local, ldap, openid
sharing allowed	with other users/as a link/through some social network	with other users, global, global with password, global with end-date
file encryption	AES encrypted transfers between servers and clients on LAN and WAN	x (not recommended)
license	https://www.powerfolder.com/products/products-overview.html https://wiki.powerfolder.com/display/PFS/Licensing commercial product, R&E discount, per user based licensing model	AGPL(owncloud.org) ?(owncloud.com)
website	https://www.powerfolder.com	www.owncloud.[org|com]

- General Evaluations (versions, license, etc.)
- Categories: deployment, access, reliability, additionals

Candidate Technologies & Evaluations (2)

---deployment---		
service-container	ORACLE JDK/JRE only apache in case of clustering solutions as a web backend for load balancing solutions	apache2
persistence	sql server/mysql/system build in DB	mysql/sqlite
storage-backends	local storage or network storage, server nodes should share the same file system	filesystem, s3, swift, external WebDAV
--- access---		
webdav	x	x
browser	x (rudimentary)	x (HTML5 based, good usability)
desktop client	Windows, OS X, Linux	Windows, OSX, Linux
mobile client	Android, iOS, Windows Phone, Blackberry	Android (0,79 €), iOS (0,89 €) Windows phone, Blackberry: no dedicated ownCloud client, generic WebDAV clients available
--- reliability ---		
high availability	HA solution with loadbalancer	HA solution with loadbalancer
scalability	farming: compute capacity can dynamically be expanded by adding further server nodes	farming: compute capacity can dynamically be expanded by adding further server nodes
replication	data is stored at server side on a shared file system, can be replicated in the backend, e.g. optionally backed up on powerfolder.com cloud	filesystem: tape/replikation s3, swift: by architecture
---additional---		calendars(caldav), supports visualization for different format
branding	preconfigured branded apps, installable from app stores	
technology	proprietary, peer to peer based protocol	server based (HTTP/WebDAV)

Experimental Setup Example



'entering production ready status @ RZG'

Tatyana Khan
Florian Kaiser
Thomas Zastrow

Summary

- Requirements for EUBOX service are good understood
  - Stable and mature technology → otherwise no alternative to Dropbox
  - Security is a key issue and the ‘entry barrier’ needs to be low
- Evaluation of candidate technologies takes some time
  - We have to limit the amount of technologies to expertise/partners
  - Expand current candidate technologies: e.g. openstack
- Entering validation phases for choosing technologies
  - E.g. get ‘hundreds of small files’ into the system (‘measurements’)
  - E.g. share a ‘big data file’ among colleagutes in the system
  - Example use cases around document exchanges expected to work very well → but research data different

EUDAT

services
measurement
within
fusion
new
funding
Policy-based
concepts
forms
cross-disciplinary
resources
device
Analysis
Centre
climate
Computer
expertise
Approaches
construction
Fusion
year
like
Computational
Test
E.g
Computing
disciplines
results
Enable
Cross-Disciplinary
PB
community
Environment
Modelling
different
EU
methods
High
big
nodes
Key
scientific
important
Science
centers
storage
basis
Infrastructure
files
etc
DKRZ
Use
model
international
SMAQ
simulations
ENES
technologies
formats
engineering
References
brain
increasing
via
EUDAT
physical
IPCC
Big
Project
Data
processing
Common
TB
Methods
MapReduce
create
computing
Using
European
databases
GB
Supercomputing
Work
images
codes
analysis
stored
Resources
analyze
Large
Summary
data-intensive
Service
Scientific
levels
often
DLCL
Level
external
use
JSC
computational
Research
Matters
Concrete
performance
research
e.g
science
HPC
Juelich
datasets
acquisition
system
compute
Landscape
System
Climate
modelling
AR5
Hardware
large-scale
cores
access
hundreds
Nutshell
Services
management
network
Technologies
Earth
Structure
DLCLs
Understanding
structures
distributed
current
General
Germany
Simulation
directory
project
Processing
day
Provide
NASA
Energy
systems
Health