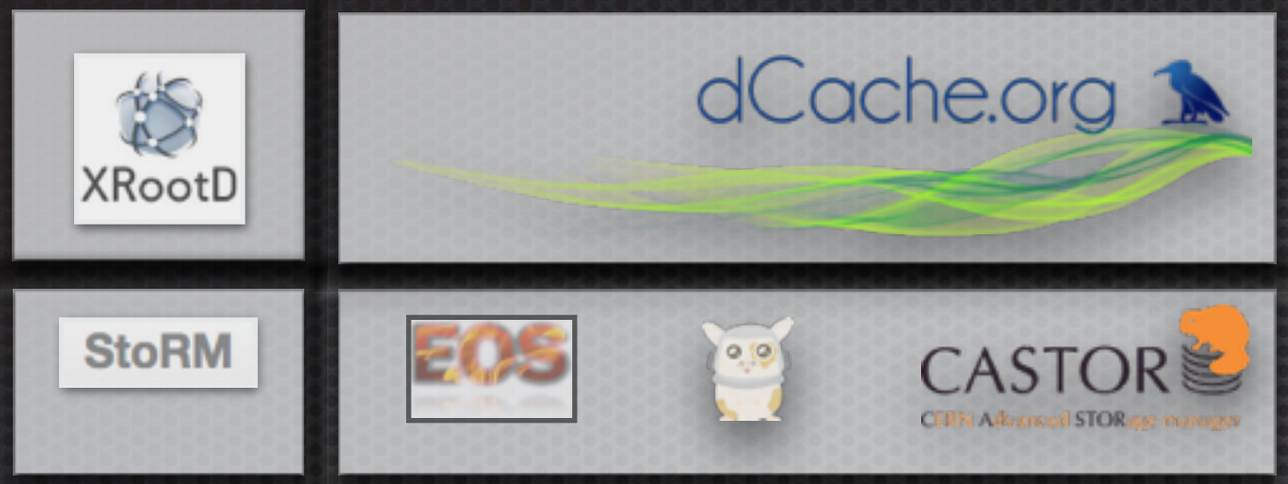


Storage Systems

medium term opportunities

Andreas-Joachim Peters on behalf of WLCG storage developers

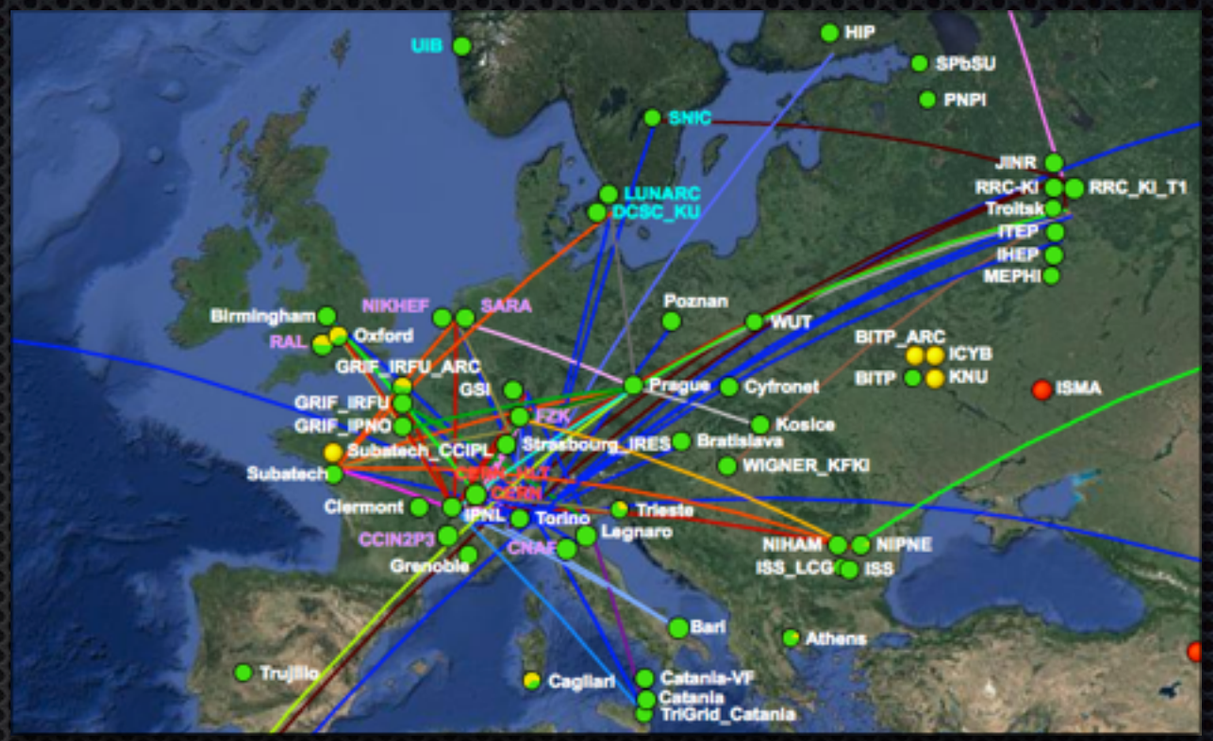
Andrea Ceccanti, Dirk Duellmann, Patrick Fuhrmann, Fabrizio Furano, Andy Hanushevsky, Oliver Keeble

01.02.2016 WLCG Collaboration Workshop
ISCTE-IUL - University Institute of Lisbon

Current State, Platform & Goals

- positive: no new problem to invent - opportunities to take
  - we are successfully operating a globally accessible distributed storage systems in the framework of WLCG
  - agree & converge towards a homogenised toolset and mode of operation to maximise efficiency for foreseen budgets and prepare for upcoming future scale challenges
  - aim to offer a forum where **sites & experiments** are invited to contribute requirements and **technology providers** get to agreements & solutions

Today we operate on this scale & complexity



167 storage elements

to be compared with …

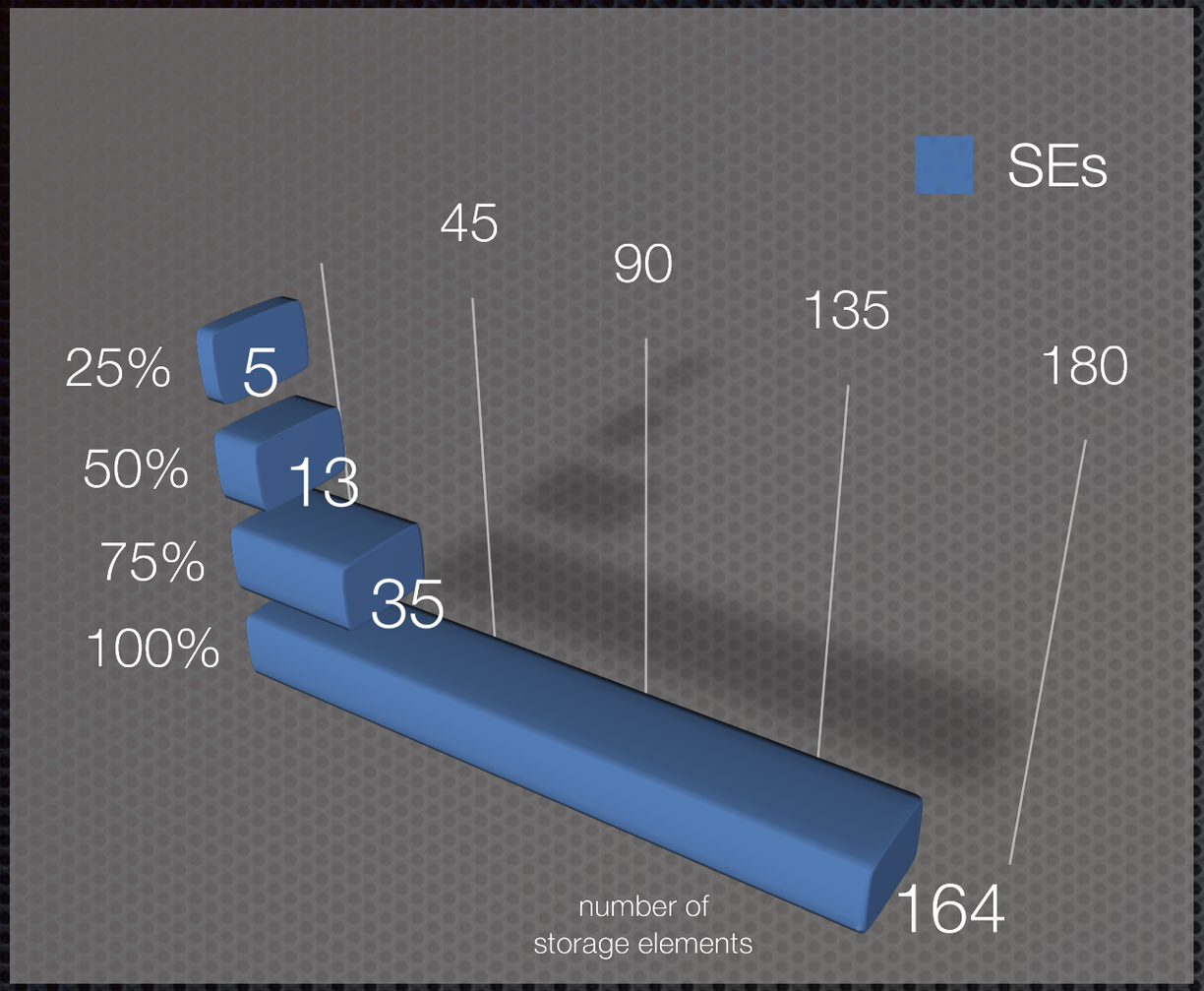
Google

13 data centres

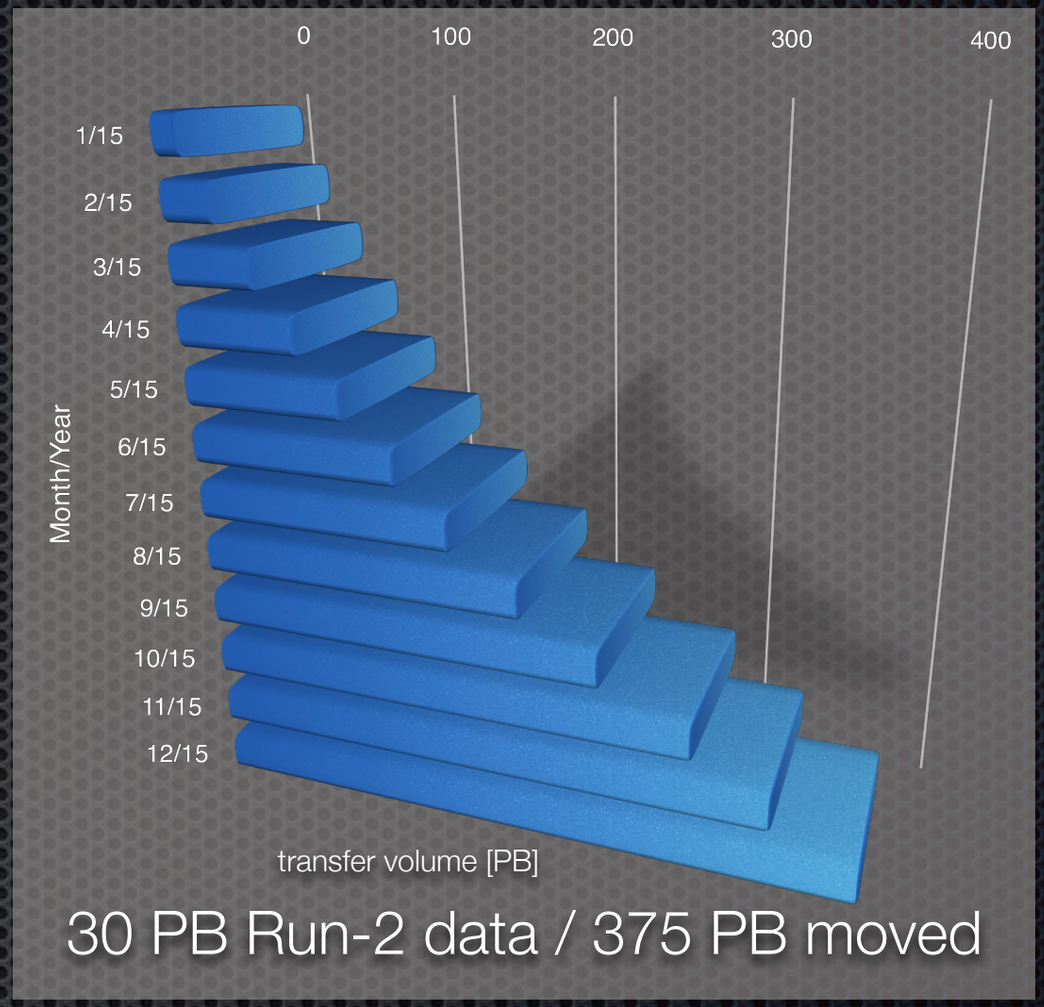
World

7 continents

Online Diskspace Contributions



Global Transfer Volume 2015



30 PB Run-2 data / 375 PB moved

Some Examples from 2015

ALICE WAN/LAN 13/228 PB/year = **5.7%** 20% CPU@CERN 50% IO@CERN

storage space CMS disk/ATLAS disk = **0.7** CMS tape/ATLAS tape = **1.1**

Files never read (CERN): LHCb **7.5%** 0.17 PB - ALICE **15%** 1.5 PB

… there is room to optimise DM efficiencies …

Storage opportunities

Efficiency & Cost

Data Distribution Models

Cloud Resources

Commercial Cloud
Storage & CPU

Protocols

Convergence

Federations

Convergence

New Technologies

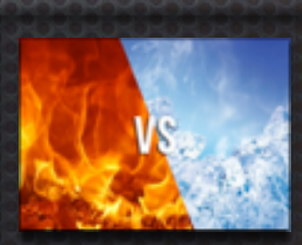
Convergence
Open Source Trends

Storage Topology

Operation & Data Distribution Models

- evolution: more dynamic (work-load defined) data placement with less required online space and CDN

- more tape than disk - aggressive archiving
- from active pre-placement towards read-through caching



  - few large managed (distributed) storage systems providing long-term persistency online/offline data : ”custodial”
    - a single storage system can be geographically distributed within acceptable latencies
  - many smaller cache storage systems with less operational effort: ”volatile” CDN components:
    - **xrootd**: XrdFileCache
    - **http**: varnish, nginx/cache, squid …

New Directions

Cloud Resources

Commercial Cloud
Storage & CPU

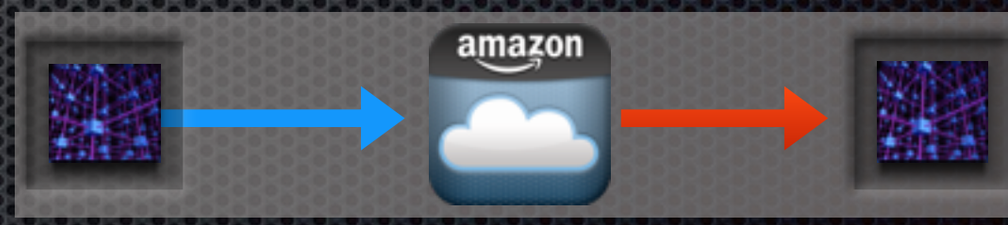
Commercial Cloud

- Cloud Services are becoming competitive
  "buy the service" or "run the service"
  - cost / unit GB/$ HEPSPEC/$
  - reliability / availability / efficiency
  - integration & operational costs

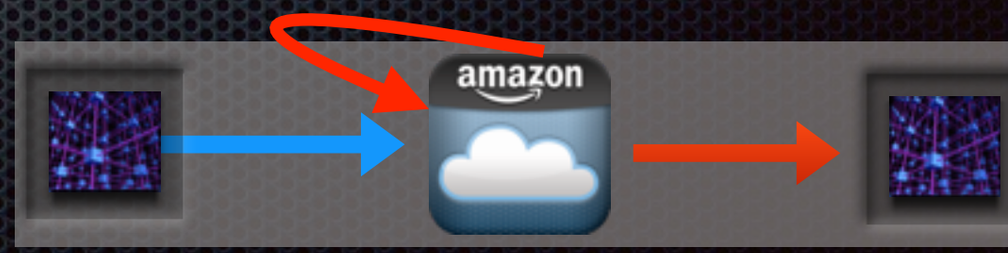
Accelerating Scientific Discovery in the Cloud - 25th May 2015

the US ATLAS team, led by Michael Ernst turned to AWS to ensure that the experiment always has access to the massive computational resources they require.

- Implication of cloud CPU resources
  - requires remote access optimisations for efficiency
  - result in increase of remote IO in current storage systems



- Implication of cloud storage resources
  - how & where are these resources attached?
    - temporary storage & remote data access
    - data cache
    - data storage
  - what is the lifetime of resources and data hosted?
  - impact due to simplifications in commercial cloud protocols [no S3 multi-range request]

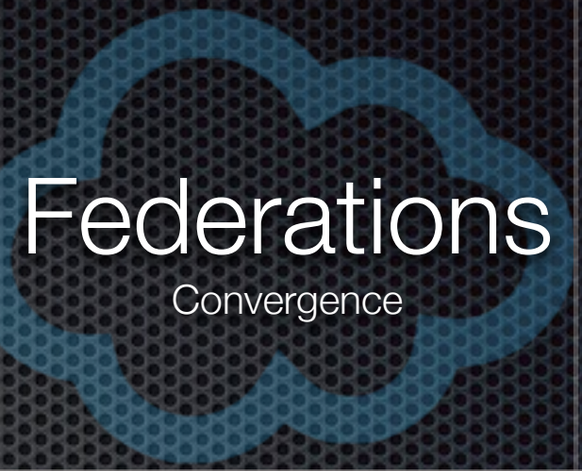


Protocols 

… we use many for different use cases …

- today: **srm gsiftp xroot http(s) dav s3 file**[nfs]
- opportunity to focus and converge on long-term stable interfaces
  - **srm** [control protocol]
    - disk
      - remove SRM from disk only storage system
      - we agreed on DAV specification to replace SRM space reporting
    - tape
      - in the future extract minimal SRM interface for tape storage systems and we will agree on a specification for the tape staging interface
  - **gsiftp**
    - replace with **xroot/http**-falvoured protocols
      - **http** is world-wide most frequently used protocol - native protocol in many (commercial) storage systems - natural candidate to replace **gsiftp** - can we avoid complication of credential delegation via unified storage tokens for third-party copy?
      - **xroot** already in production for ALICE instead of gsiftp since years
  - **file**
    - keep as most stable interface
  - **analysis**
    - requires additional client/server support
      - **xrootd** enabled storage
      - davix + specialised HTTP storage (not provided today by commercial and open source cloud storage)

Global Storage



Complement of Federations, Storage & Data Management Systems

- today
  - federations **xrootd-based** (FAX/AAA) & **http-based** (DynaFED)
    - catalog-driven federation to be compared to real-time federation
    - real-time XRootD federation today in ATLAS/CMS, ALICE since few years only catalog-driven
      - distributed effort & support
    - HTTP real-time federation = **DynaFED**
      - test setups (LHCb, ATLAS, belle, CANARIE) - prod setups CCC, CMS@HOME, BNL
      - deployment non-intrusive - apt solution to integrate object/cloud storage (S3 data bridge BNL)
    - real-time federation independent of central data management - can release central load
  - federations on storage software level
    - dCache distributed setup (Nordic T1)
    - XRootD: technical possible, currently no setup
    - EOS distributed deployment (CERN, Hungary, Taiwan, Australia)
    - geo-replication (limited) federation support in CEPH S3
    - multi-FS mounts over WAN (GPFS, NFS etc.) as federation

new possibilities with improved networks

Global Storage

Unified Solution for Storage Access Tokens & Identities

- today
  - **gridmap/voms/gums** complex in a distributed environment
  - ALICE special - token based authentication

- future
  - **signed URLs** as defined by **AWS** (S3) are similiar to ALICE tokens. They allow a very simple mechanism to generate access tokens to a globally distributed storage system and are supported by commercial clouds
  - try to get an agreement under storage providers to support signed URLs as defined by **AWS** without the need to implement the full S3 protocol or similar decentralized storage tokens like **Google Macaroons**
  - **federated identity** support (social logins etc.)

Global Storage

Data Management

Convergence

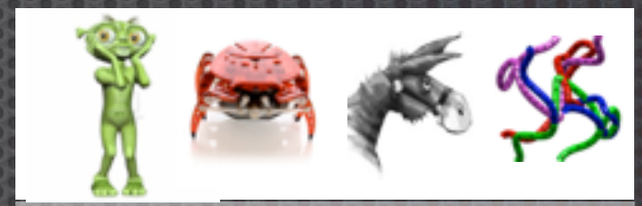
Key to efficient storage usage - intelligent data management

- efficiency and long-term scalability & maintenance of global data management solutions?

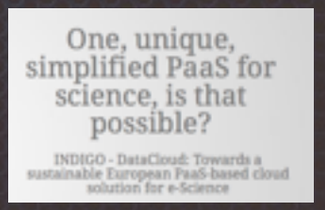
  four different data management systems for four experiments

  AliEN, DBS, Dirac, Rucio +++

  - are there fundamental differences in the data distribution, data access model and efficiency?
  - is there a possibility and/or an interest for convergence?
    - standard to build interfaces to WMS e.g. **meta link files** to allow separation of WMS & DM
    - **standalone** DM modules - **VO agnostic**
  - all implementations have a DB centric model managing storage systems in a flat hierarchy



- Can DM complexity be reduced on the storage provider level and via federation with a **thin homogeneous/ shared middlleware** layer?
  - (SW) **OneData** Indigo project - (HW) **OSiRIS**, Open Storage Research InfraStructure
  - Functional extension of existing federations



Object Storage

New Technologies
Convergence
Open Source Trends

… a standard for scale-out storage … but not only one …

discussion topic

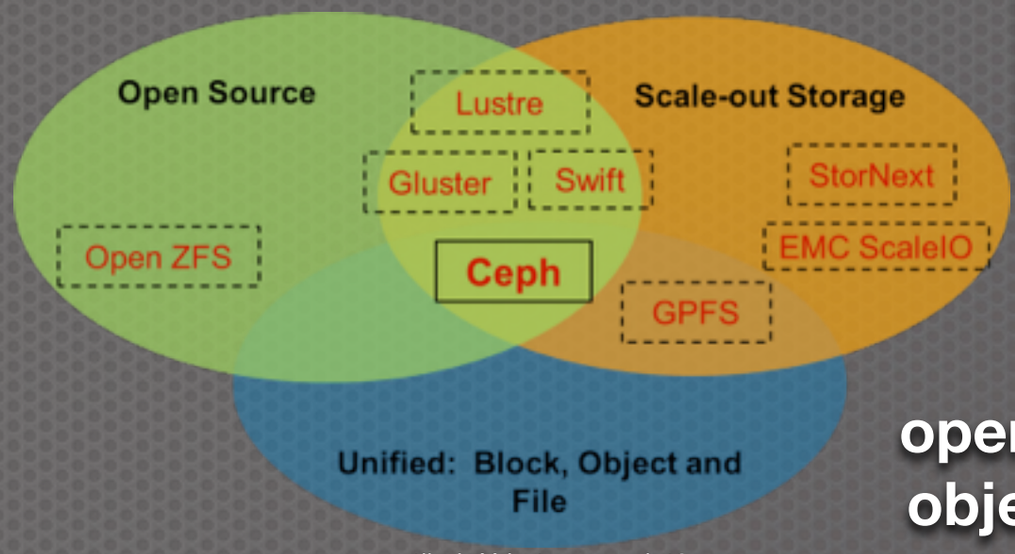
Unified Storage one fits all | Specialized Storage

compatibilty of ‘cloud’

jclouds

addressing cloud compatibility issues

“The storage unicorn” - what is special about **CEPH**?



open source object store

versatile LAN storage platform
limited federation support in S3
distributed RADOS not suited well for setups with latency

a new zoo …

abiquo · amazon web services · CloudSigma · cloudstack (open source cloud computing) · DigitalOcean · docker · Google Cloud Platform · hp Cloud · Joyent · openstack (cloud software) · rackspace (the open cloud company)

Apache jclouds application example: **s3proxy** for Google/Microsoft/Swift **cloud storage**

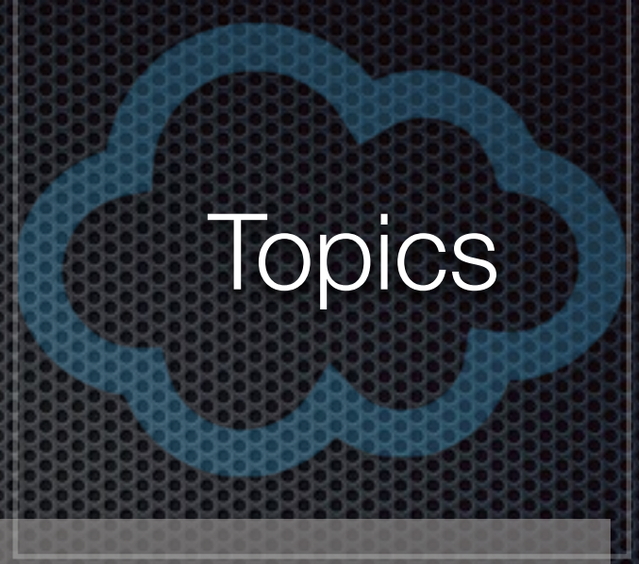
Midterm Opportunities

- **low** growth of online data - **more** archive=cold data
  - changing storage technology might not improve GB/$ - gains by local optimisations like de-duplication/compression/erasure coding are desirable but alone not sufficient
- **reduce** global complexity
  - few custodial, more volatile storage
- **reduce/unify** protocols
  - remove SRM, replace gridftp, unify storage token support
- **converge & contribute** in data management/federation solutions

… a collaborative effort between users & storage providers - let's go forward!


tenki
hvala
ขอขอบคุณคุณ
takk
спасибо
kam sah hamnida
dank je
הדות
dhanyavadagalu
tack
дзякуй
dziękuję
bedankt
blagodaram
mèsi
xièxie
tanemirt
gracias
rahmet
enkosi
mochchakkeram
trugarez
djere dieuf
danke
kop khun krap
faafetai lava
arigatô
akun
shukriya
ありがとう
kia ora
dankon
děkuji
manana
diolch
grazzi
mamnun
gràçie
kiitos
spas
ačiū
thank you
barka
शुक्रिया
sulpāy
tapadh leat
dhanyavad
gratias ago
tau
dankie
gràcies
chnorakaloutioun
bayarlalaa
teşekkür ederim
σας ευχαριστώ
obrigada
sagolun
murakoze
taiku
mahalo
didi madloba
chokrane
rahmat
sukriya
obrigado
dakujem
misaotra
welalin
najis tuke
terima kasih
mercé
ارك شك
謝謝
köszönöm
asante
grazie
nandri
mersi
sobodi
nanni
vinaka
mauruuru
merci
matondo
cảm ơn bạn
go raibh maith agat
paldies
ngiyabonga

Discussion



- **change of storage topology**
  run many simple storage systems and few complex ones - segregate site/task types

- **convergence of protocols & access tokens**
  are there reasons prohibiting changes in this area? can we define a time frame ?

- **ratio between disk/tape storage**
  what is needed in terms of disk resources? is 1:1 necessary?

- **data management & federation**
  is there an interest to move towards shared middleware for data management?
  what is the the future of federations and how will they look like?

- **cloud resources**
  what is still needed in terms of development for integration? what is the operational impact?