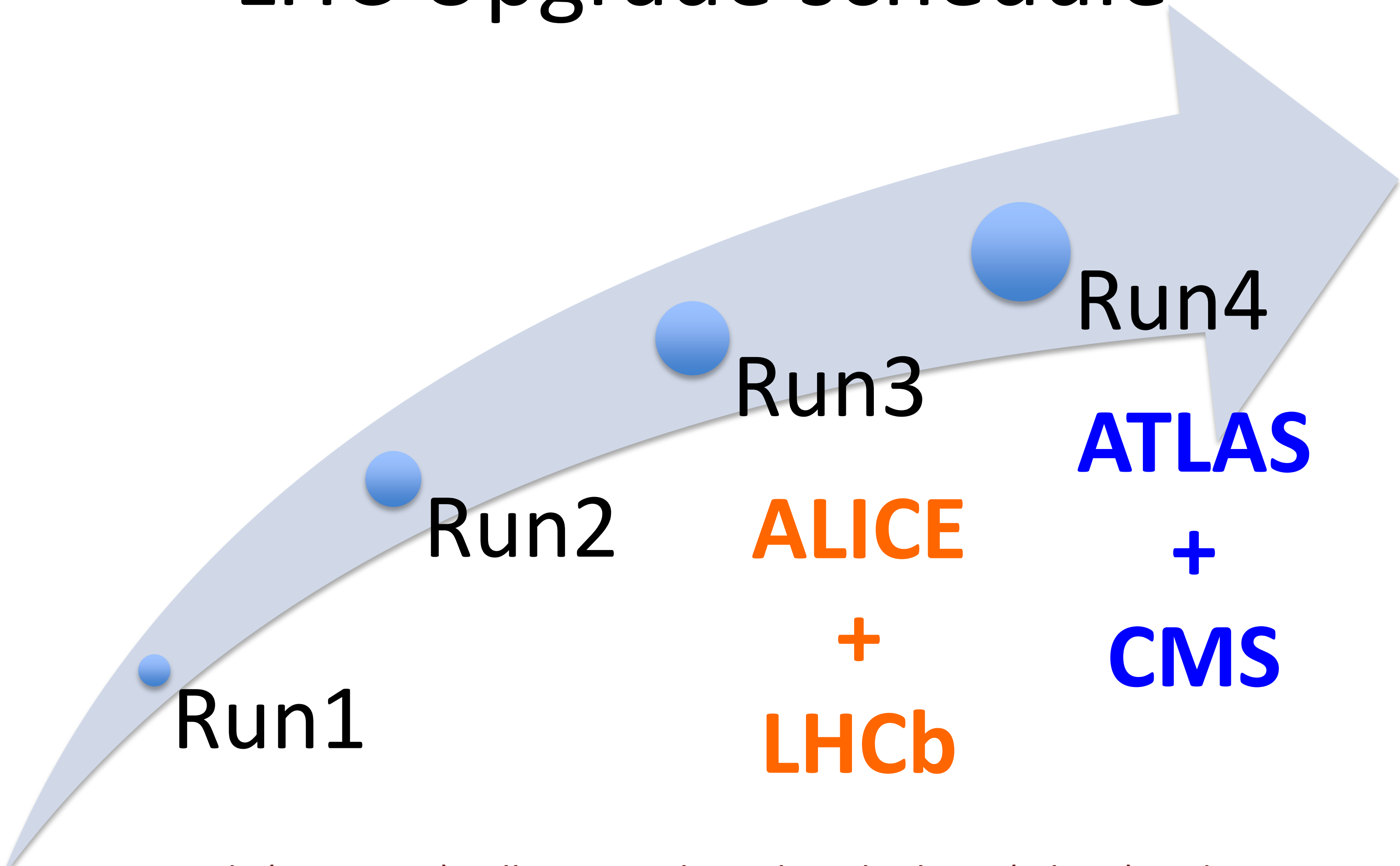
A Large Ion Collider Experiment

Possible common R&D subjects

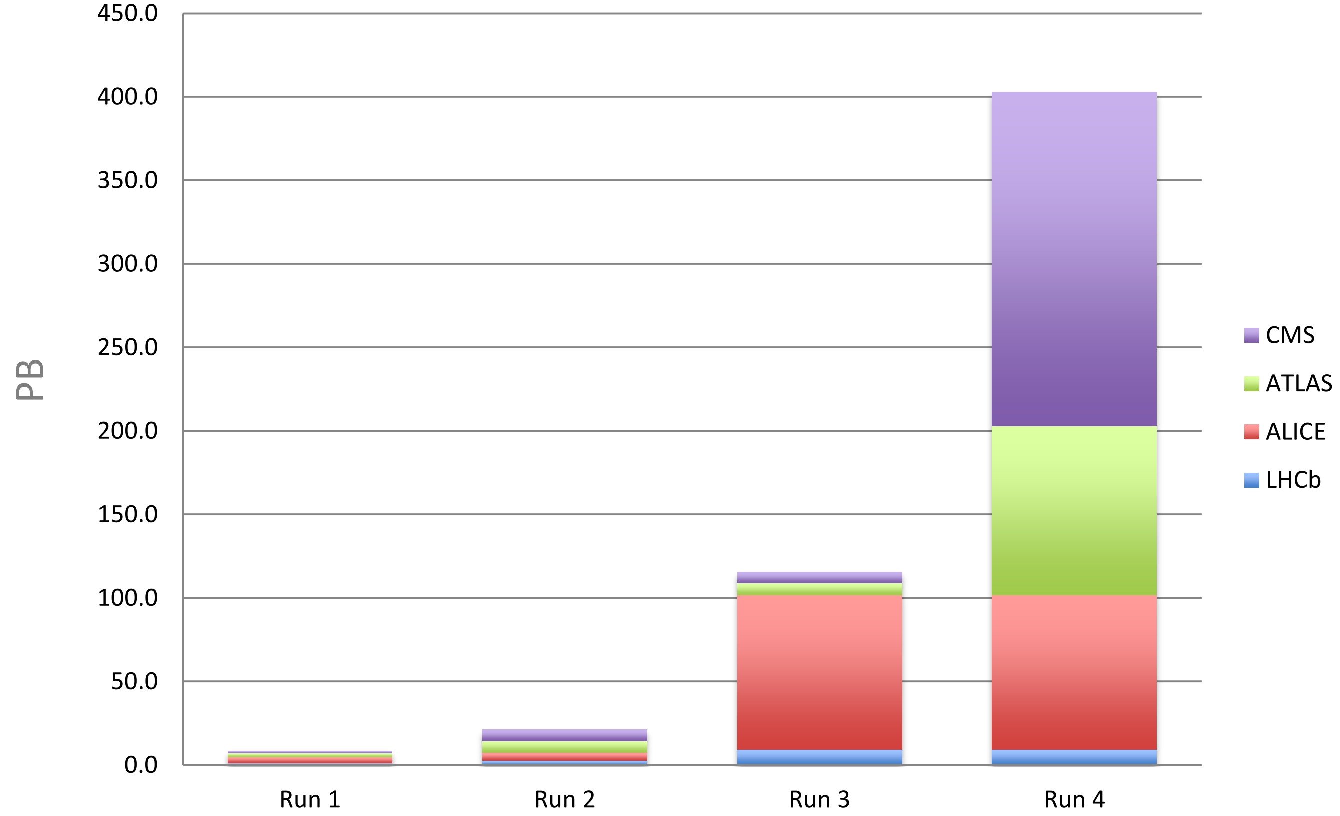
Predrag Buncic

LHC Upgrade schedule



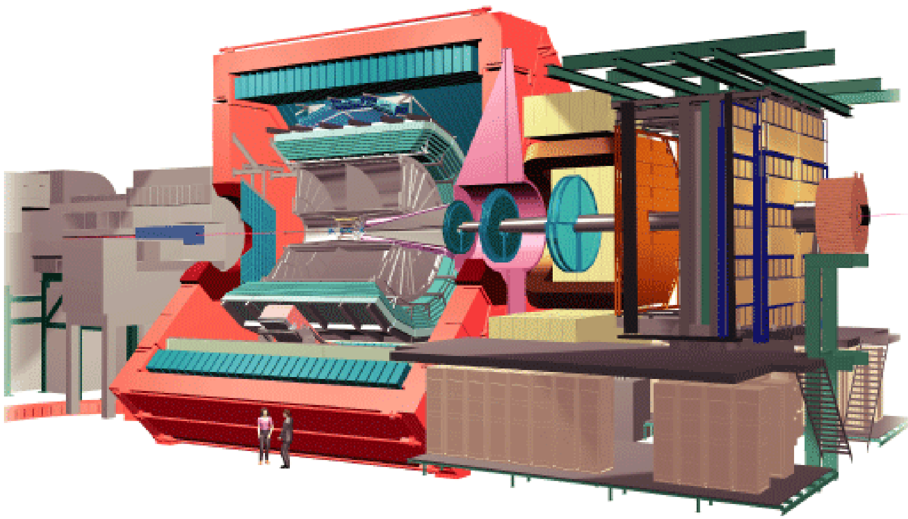
- CPU needs (per event) will grow with track multiplicity (pileup) and energy
- Storage needs are proportional to accumulated luminosity

Big Data Outlook



- Very rough estimate of a new RAW data per year of running using a simple extrapolation of current data volume scaled by the output rates.
  - To be added: derived data (ESD, AOD), simulation, user data…

ALICE @ Run 3



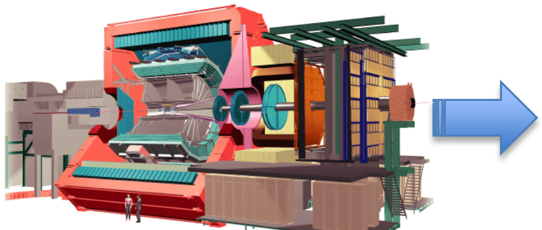
50 kHz

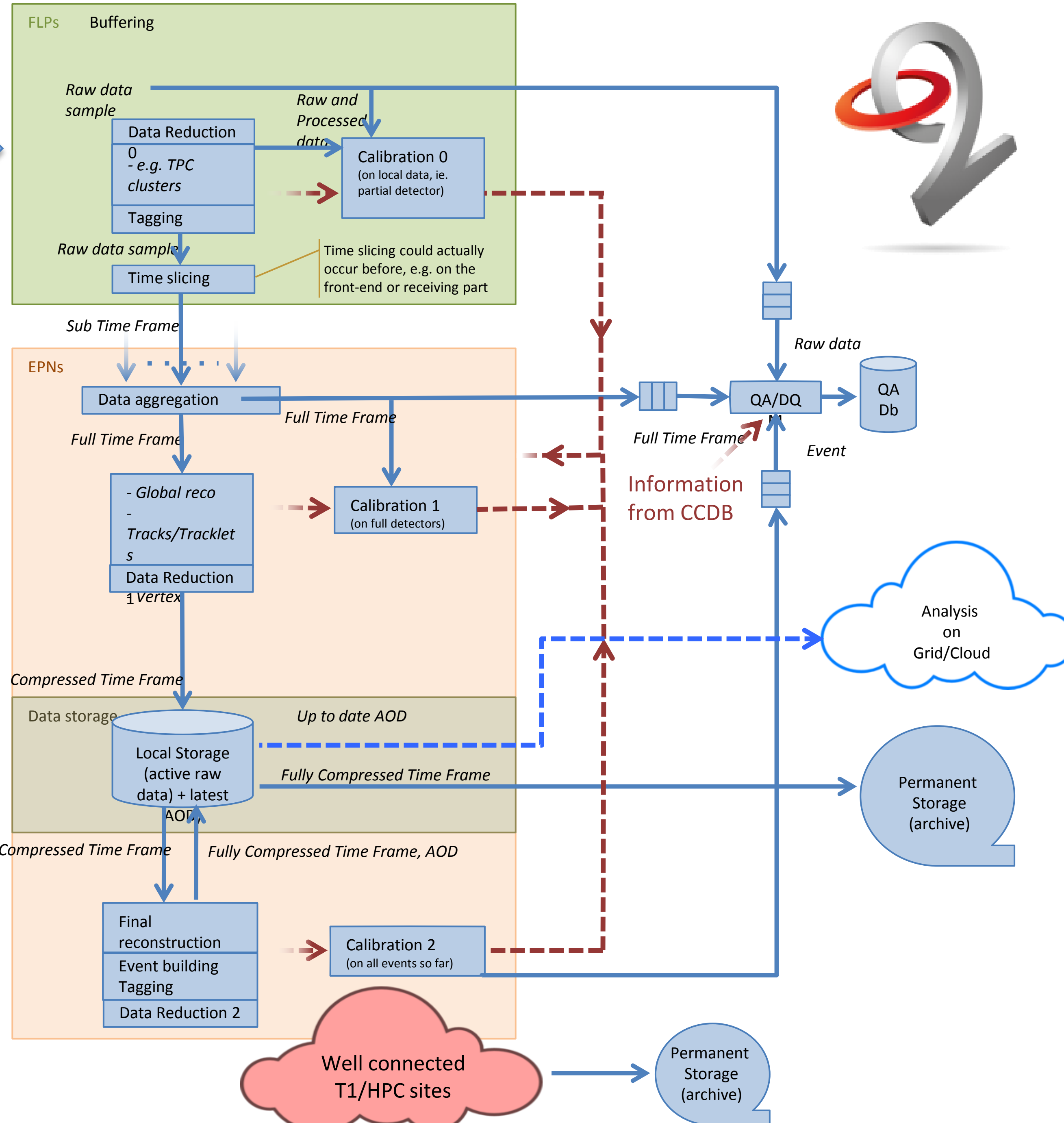
Online/Offline Facility

50 kHz (1.5 MB/event)

Storage

75 GB/s





1 G2014

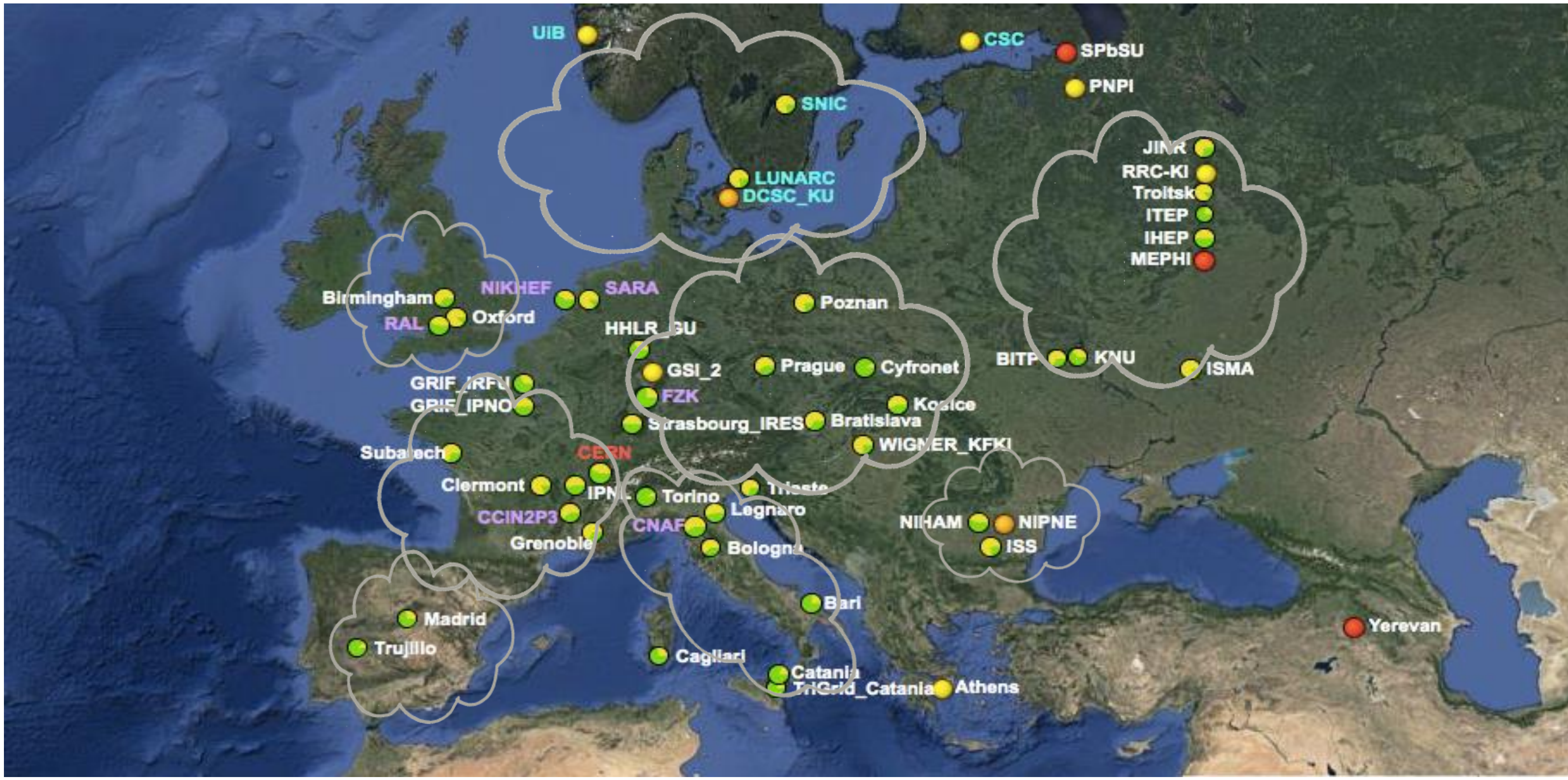
50 PB disk buffer

4 G2014

(Big Data) Storage for O2

- Fast
- Reliable
  - Failure of a single disk or server should not degrade system availability
- Scalable
  - Seamless growth by adding more resources
- Cheap
- High level tools for data management
- Should integrate with Offline applications and provide a common API

From Grid to Cloud(s)



- In order to reduce complexity national or regional T1/T2 centers could transform themselves into Cloud regions
  - Providing IaaS and reliable data services with very good network between the sites, dedicated links to T0

Reducing the complexity

- Deal with handful of clouds/regions instead of individual sites
- Each cloud/region would provide reliable data management and sufficient processing capability
    - What gets created in a given cloud, stays in that cloud
- This could dramatically simplify scheduling and high level data management

- Again, data management is the key

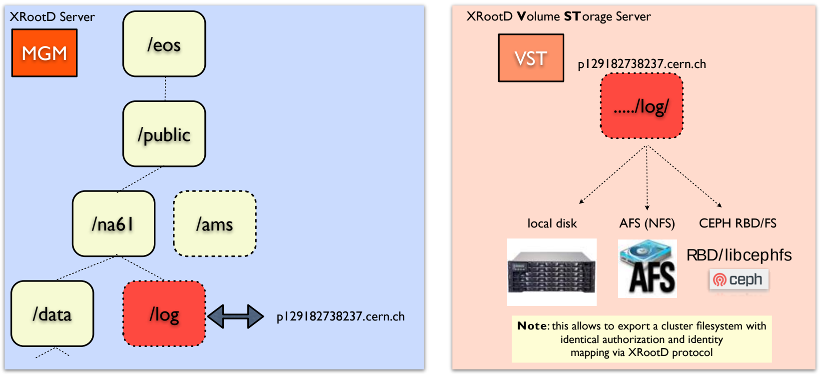
CITRINE VST

PRIVATE PROPERTY
Program of Work Proposal
No agreed IT strategy!

Infinity ∞

- EOS *Infinity*

AFS-like attached volumes hosting data+meta data of a subtree
- small/many file use cases
- allows to attach any mountable FS tree into EOS namespace
- allows to have extended attributes on file and directory level for meta data tagging



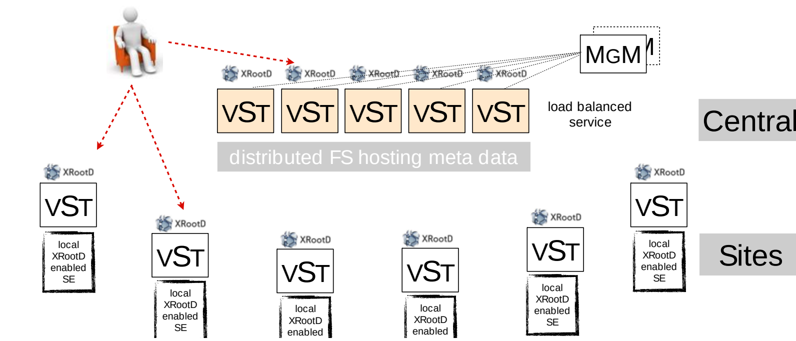
CITRINE VST

PRIVATE PROPERTY
Program of Work Proposal
No agreed IT strategy!

Unity

- EOS *Unity*

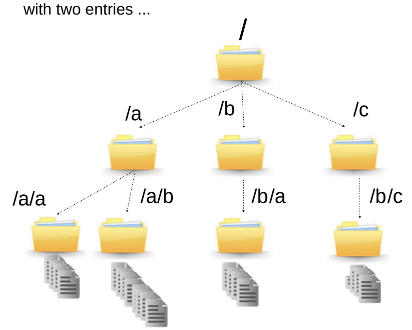
Today's Federations provide a redundant functionality via a read-only overlay network. A complete storage federation should have also placement capabilities, honor replication policies and a global reliable namespace. We can use a group of VSTs to host the global logical namespace redirecting read and write requests to VSTs hosting a logical or physical namespace (sites). A site VST is just a redirection and report gateway to any regular XRootD enabled SE or a local EOS setup. For placement and file access we can extend the already existing geo placement/scheduling capabilities of EOS used for the CERN/Wigner CC setup.

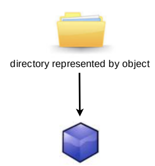


Diamond R&D

PRIVATE PROPERTY
Program of Work Proposal
No agreed IT strategy!

- trivial idea: store a namespace in a scalable object store
  - we can represent data in a *hierarchical structure* using directories and files and we *don't need* to group an infinite amount of files into a single directory
  - each *file* is a *list entry* with meta data in a directory
  - each *directory* is represented as an *object* in an object store
  - to circumvent central locking we can allow a conflict if two files get created with the same name and different contents and make it visible in the namespace like a conflict in DropBox with two entries ...





dir.attributes

owner	acl	xattr
root root	xyz	user.x sys.y

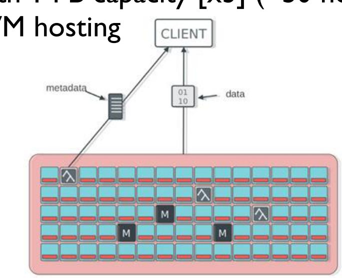
file table

Name	Size	Cks	Locatio	UUID
a	1	0xa	1:2	A
b	2	0xb	2:3	B
c	3	0xc	3:4	C
d	2	0x4	4:5	D
e	1	0x5	5:6	E

Diamond R&D Scalable Object Store/Namespace using CEPH

PRIVATE PROPERTY
Program of Work Proposal
No agreed IT strategy!

- ceph is an open source implementation of an object store providing features like *dynamic resizing, self-healing, guaranteed consistency, low read latency, async object IO, extended attributes + key-value map per object, object notifications*
- IT-DSS provides now a ceph (rados) object store **service** with 1 PB capacity [x3] (~50 nodes) - initially for VM hosting

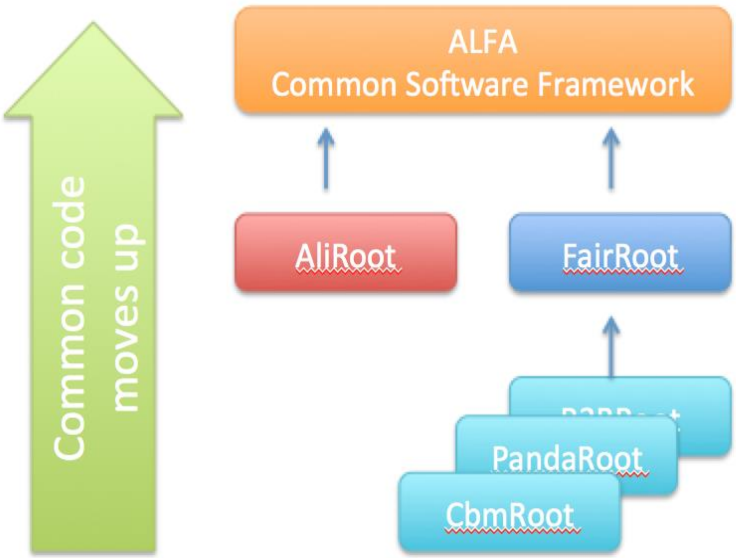


Andreas Peters

EOS+

- EOS is already tested to the scale required for O2 internal buffer some extras might be needed
  - Media aware caching (SSD, fast disk, shingled disk…)
  - Sophisticated disk pool monitoring, visualization
- Scalable global name space
  - Replacement for file catalog
- Storage federation
- Integration of foreign file systems for specific purposes

About the software challenge…

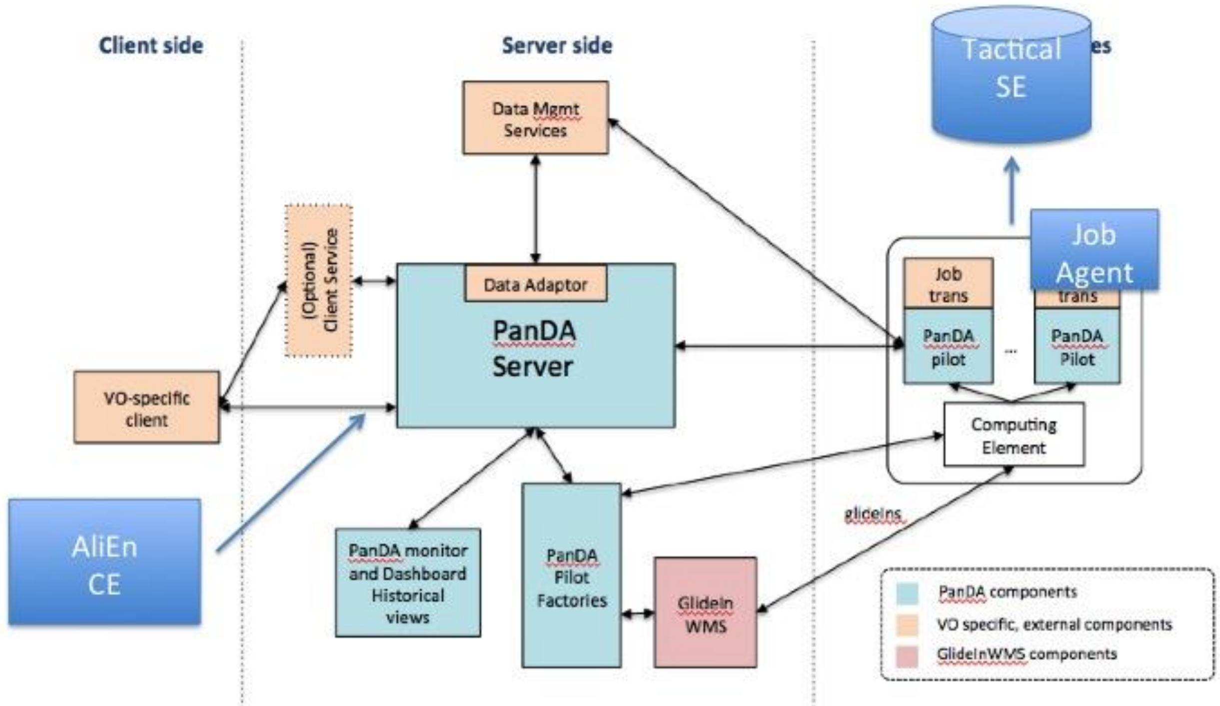


Alice + Fair = α(AlFa)

the basis of AliRoot 6.0

- Equally challenging task on the short time scale
  - Muticore, possibly heterogeneous computing
- Multiprocessing vs multithreading
- Too difficult to do it alone
  - ALICE decided to team up with FAIR and develop its new framework based on distributed, multiprocess, message driven architecture
- Open for collaboration with other experiments at CERN and elsewhere

AliEn, PanDA, HPC and simulation



- 70% of all CPU cycles spent in ALICE are for simulation
- Adapting AliEn to use PanDA was seen as a possible way to use HPC resources
  - Not an easy job, requires application to support multithreading and/or grid framework to support whole node submission

Another idea…

- Do not consider simulation as yet another set of jobs to run on the grid
- Instead, treat the HPC resources as simulation data source
  - Similar to raw data from experiment
  - Simulation (possibly pre-deployed) needs only the configuration parameters
  - The result needs to be registered in common name space and storage pool
- Use PanDA as the simulation data source
  - If the software is distributed/configured via CVMFS and we find a common solution for data store/management this could be a completely generic solution

Summary

- Big Data store/management for Online farms
- Generalization of EOS data management to allow seamless federation and aggregation of different data stores via common interface providing scalable namespace
- Data driven, multi-process software framework development aimed for efficient use of many core platforms ranging from a single node to the entire Online farm
- Simulation as data source/service aimed for opportunistic use of HPC resources