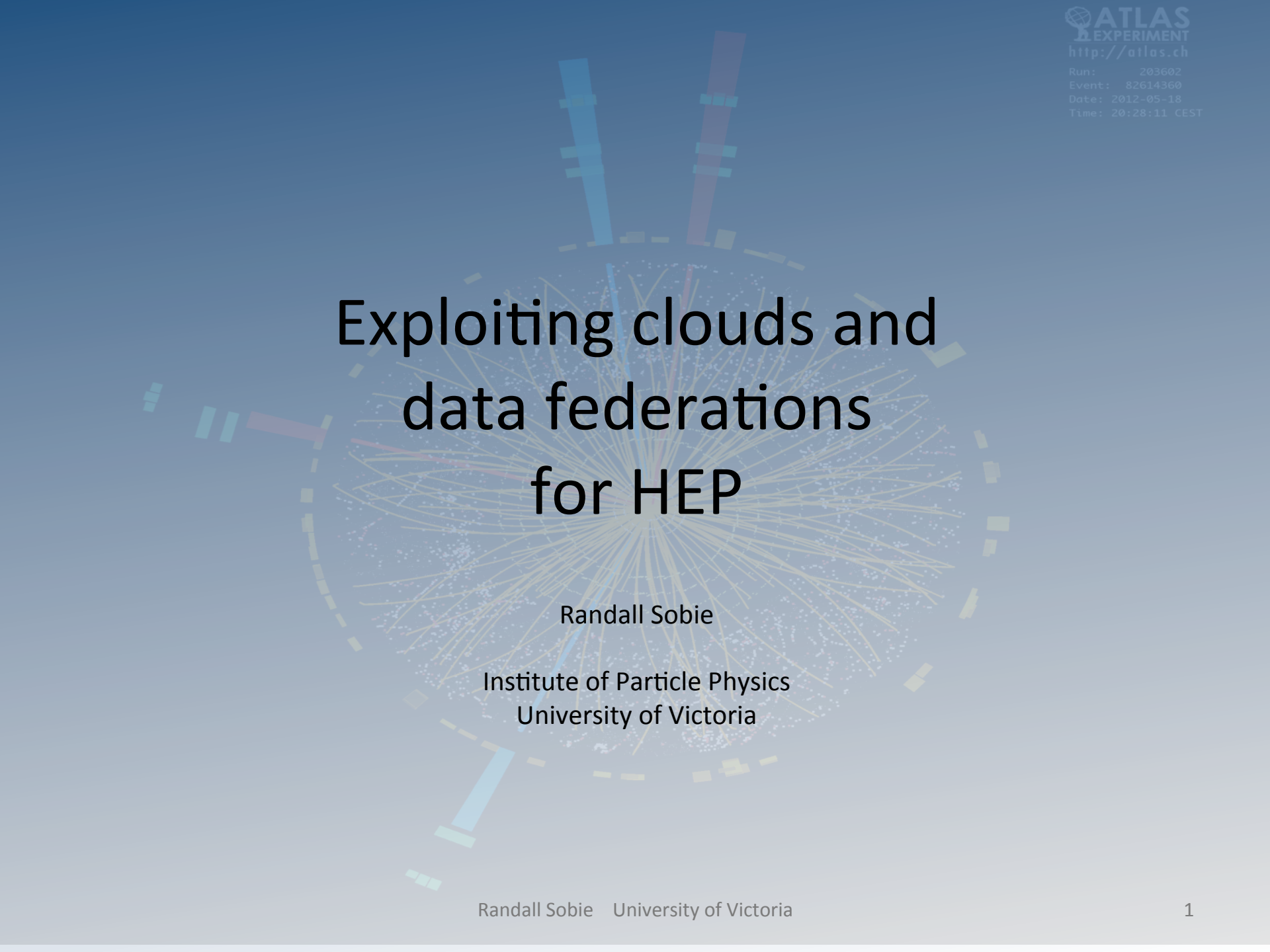
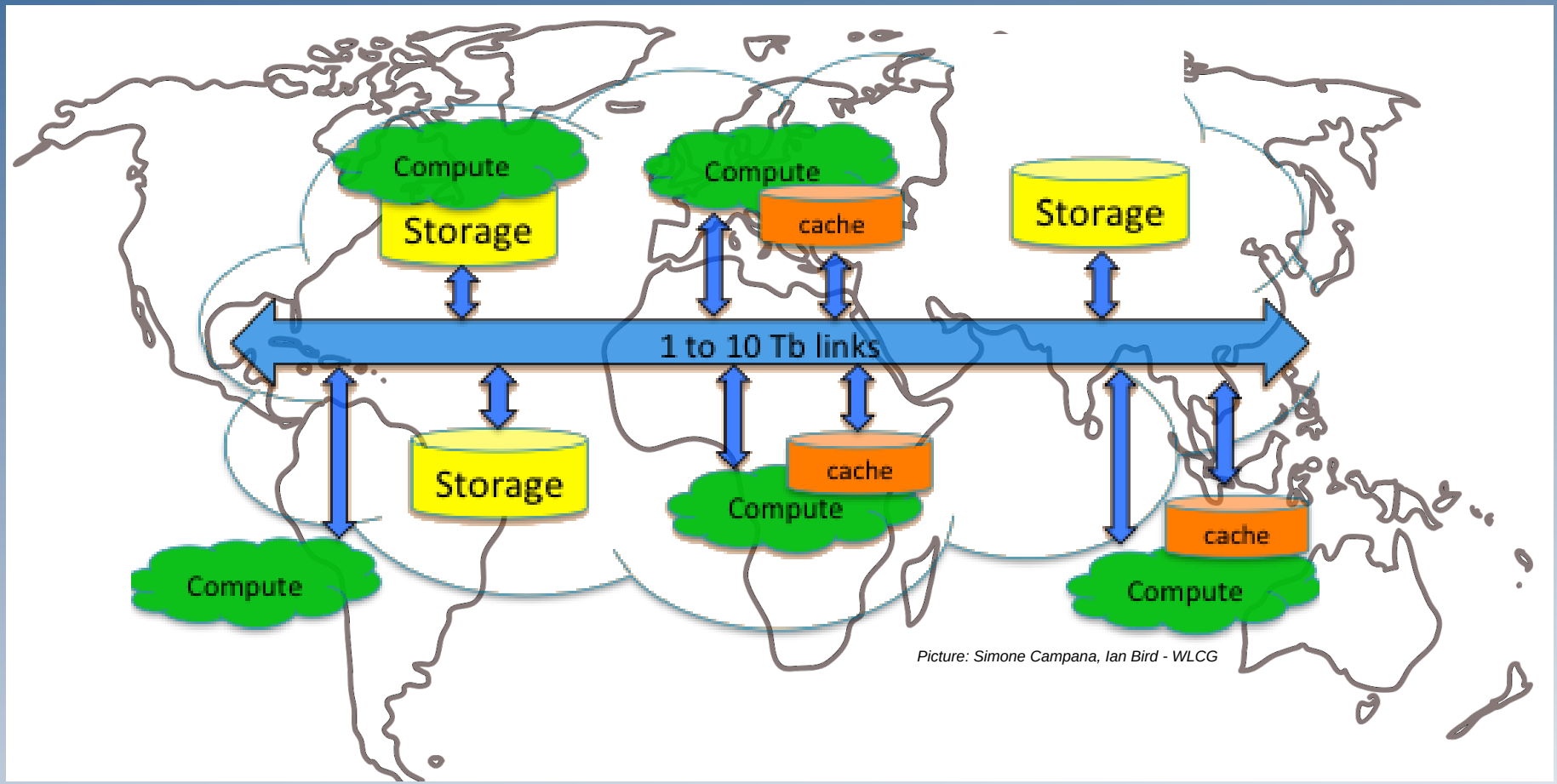


Exploiting clouds and data federations for HEP

Randall Sobie

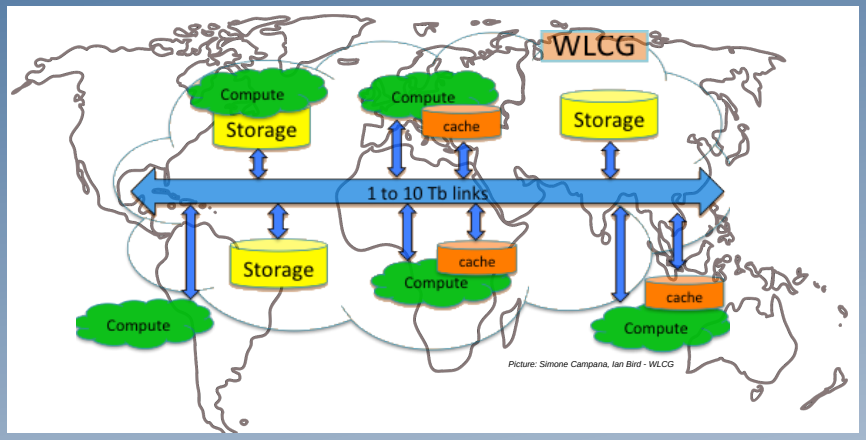
Institute of Particle Physics
University of Victoria

Future of HEP computing



Large, independent centres or federations of compute and storage distributed around the world linked by terabit/second networks

Motivation for the new model



Currently ATLAS (and other HEP projects) operate a grid of ~100 computing centres
(including clouds, HPCs and volunteer computing)

Linked with 10-100G networks

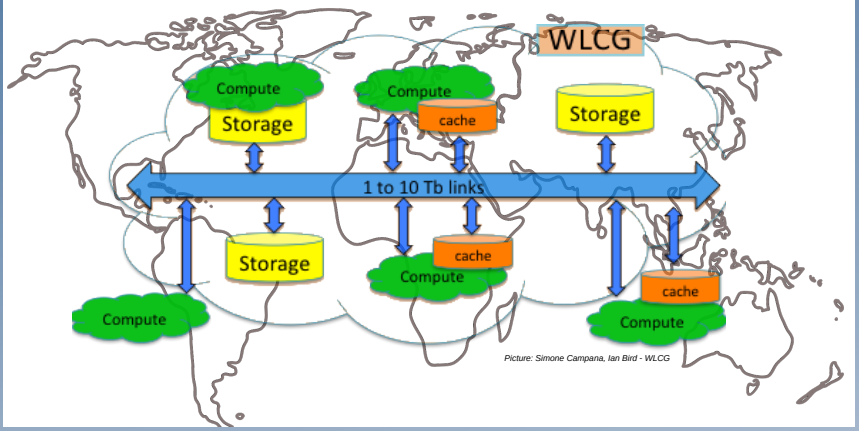
Successfully operating for LHC experiments
(multi 100K jobs; 0.5 Exabyte data samples)

Manpower intensive

Grid technology is not widely adopted, with little new development

Expensive
(operating and development costs)

Data intensive applications



Data intensive applications are run on facilities with a tight coupling between compute and storage

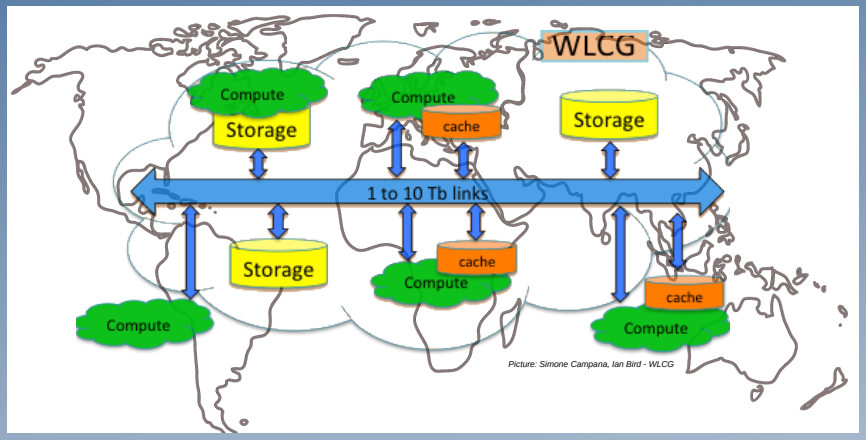
<u>Challenges</u>

Slow to respond to user demands
(difficult to predict the interesting data set)

Inefficient use of expensive storage
(many duplicate copies of data)

Difficult to use non-HEP (opportunistic) resources with data
(e.g. non-HEP clouds or HPCs)

Goals of the new model



Federate regional resources (compute and storage)

Eliminate coupling between compute and storage

Utilize opportunistic resources for data intensive applications

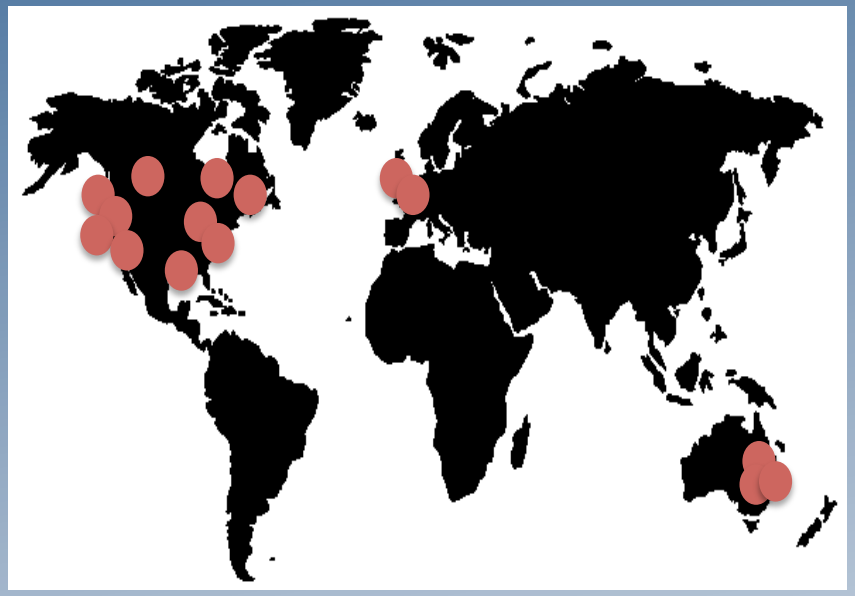
<u>Transitioning to the new model</u>

Using federated cloud computing systems
(e.g. UVIC distributed cloud for ATLAS and Belle II projects)

Funded by CFI Cyberinfrastructure Program to build data federation
(*For both dedicated and opportunistic resources)*

UVIC, TRIUMF, ATLAS-CERN, CERN-IT Project started in July 2016

Distributed cloud computing system



Dedicated and opportunistic resources
(ATLAS and BelleII)

Production system for many years

Typically 5000 cores (peak ~9000 cores)
Primarily low I/O but high I/O on selected sites

Use clouds in North America and Europe

OpenStack (private and commercial)
OpenNebula (private)
Amazon EC2
Microsoft Azure

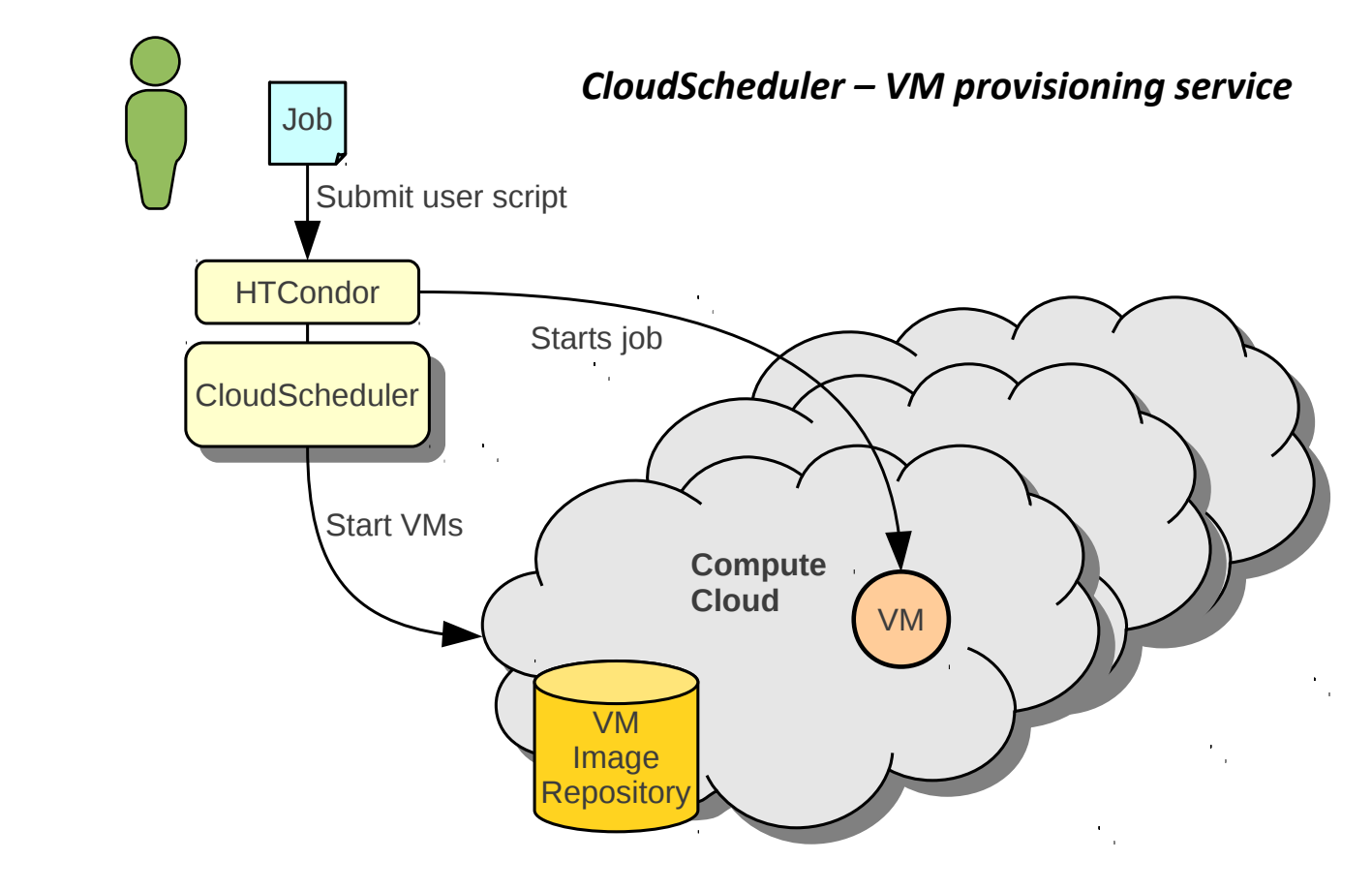
Overview of system given at HPCS-2016:

http://heprc.phys.uvic.ca/sites/heprc.phys.uvic.ca/files/Sobie-HPCS.pdf

Overview of HEP cloud use

http://heprc.phys.uvic.ca/sites/heprc.phys.uvic.ca/files/Sobie-Cloud-CHEP.pdf

Distributed batch cloud computing



Design conceived 2008 and CloudScheduler first deployed in 2009

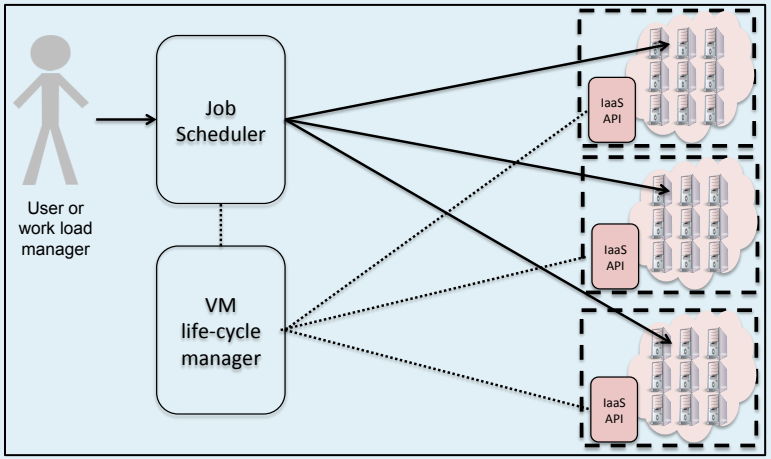
Software and services

Integration of many existing, open-source components
(Only develop missing elements)

Panda, DIRAC, HTCondor-client
Client job submission

HTCondor
Batch job system

CloudScheduler
VM provisioning and management



OpenStack
Amazon EC2
Microsoft Azure
(GCE)

microCernVM (cloud_init)

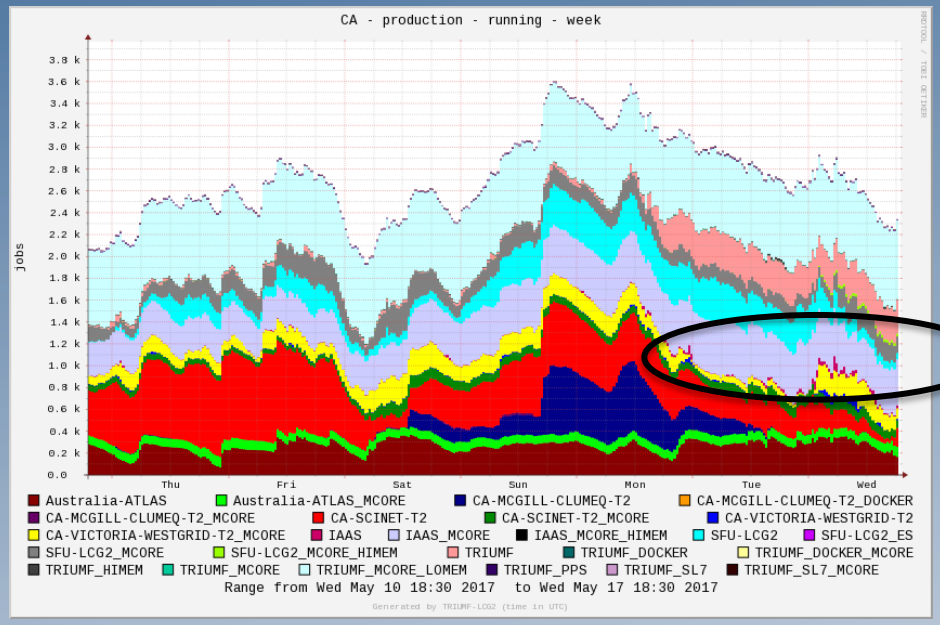
Glint
VM distribution over remote clouds

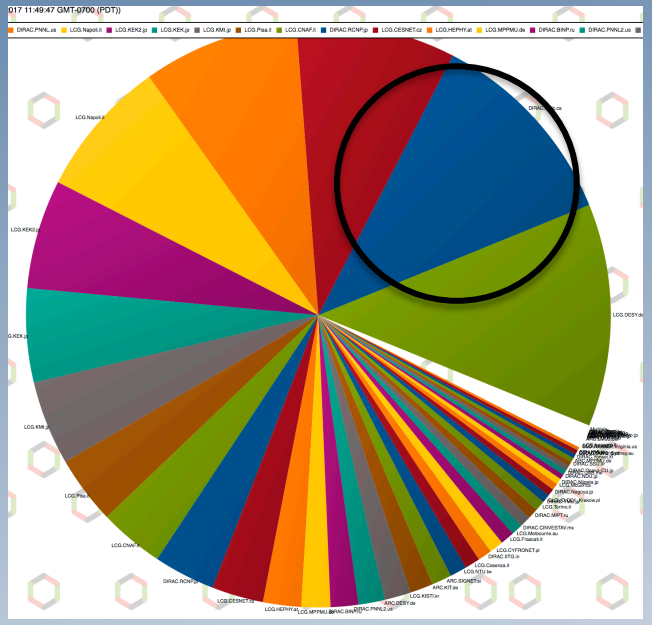
Shoal/Squids/CVMFS
Squid cache discovery service

Munin/Ganglia/Grafana/…
Monitoring systems

Production system for many years
On-going development to manage technology changes, improve reliability and adding new capabilities

ATLAS and Belle II





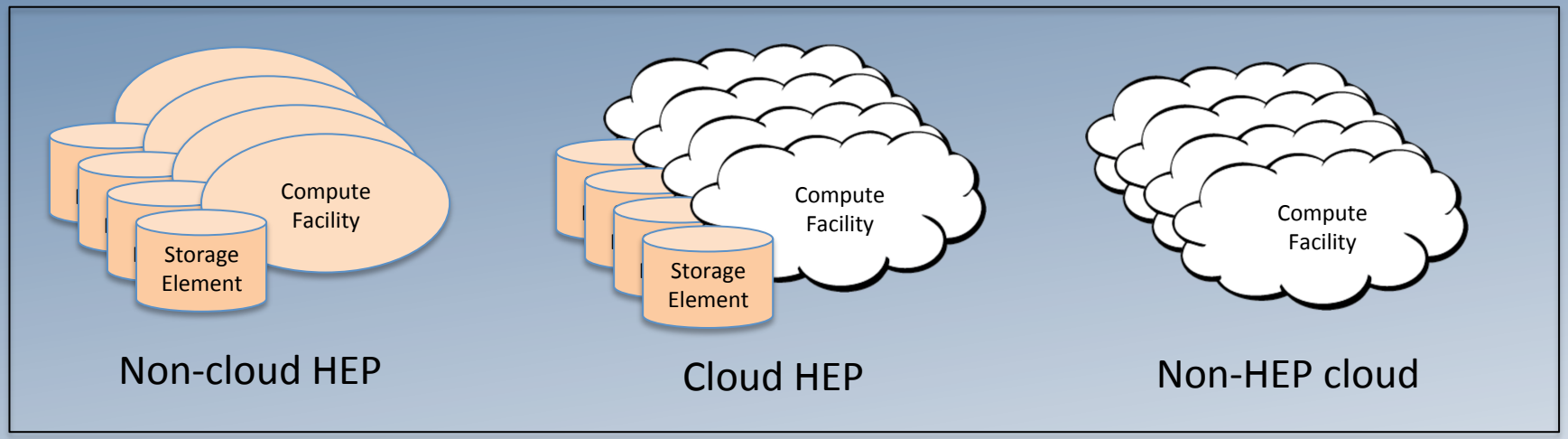
ATLAS
Account for 25% of Tier-2 production in Canada
Also use clouds at CERN, Munich (and UK)

Belle II
Account for 10% of global production

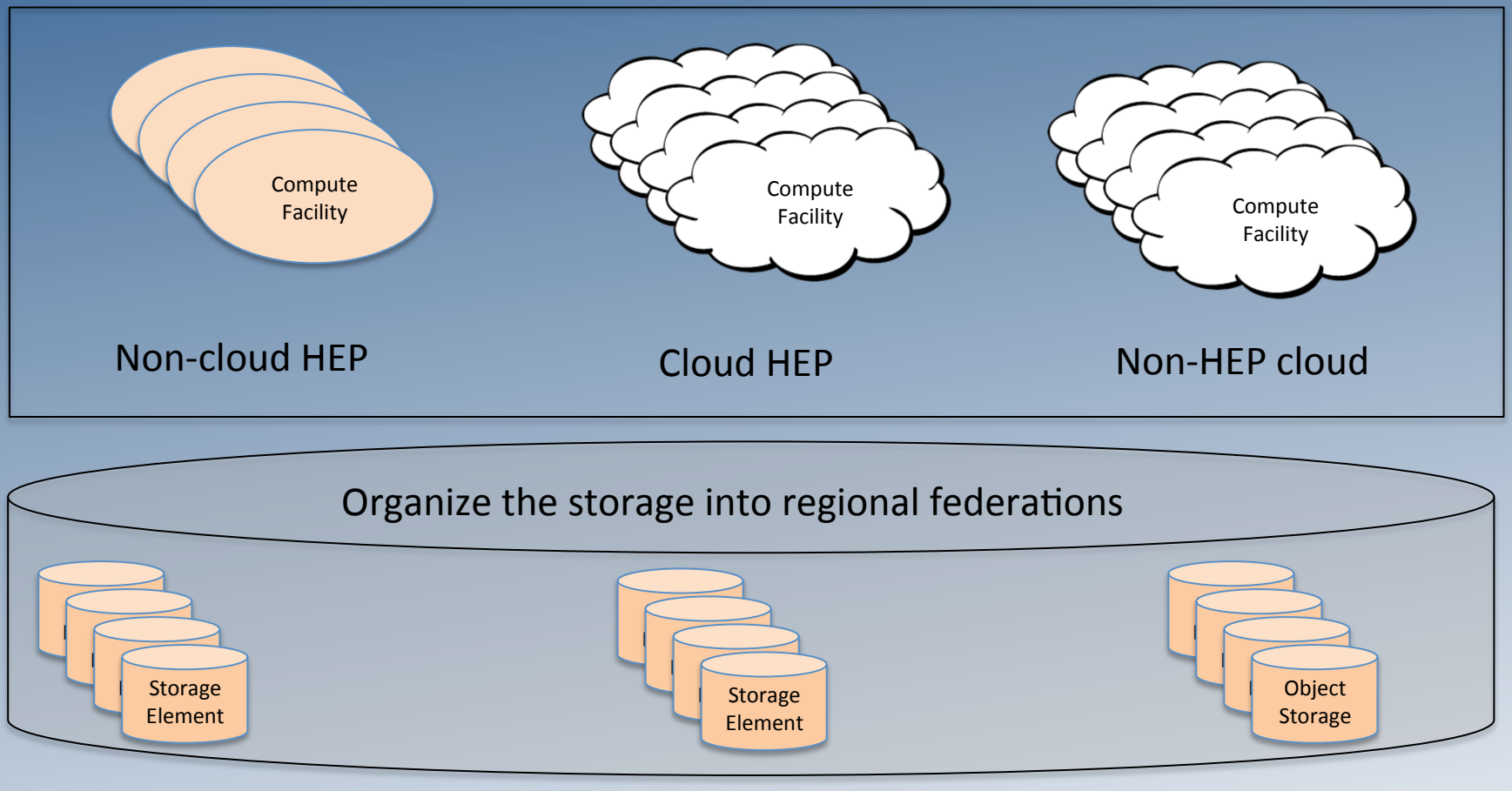
Current data management strategy

Data is distributed across many sites (some duplication)

Data-intensive jobs are sent to the site or federation with the data



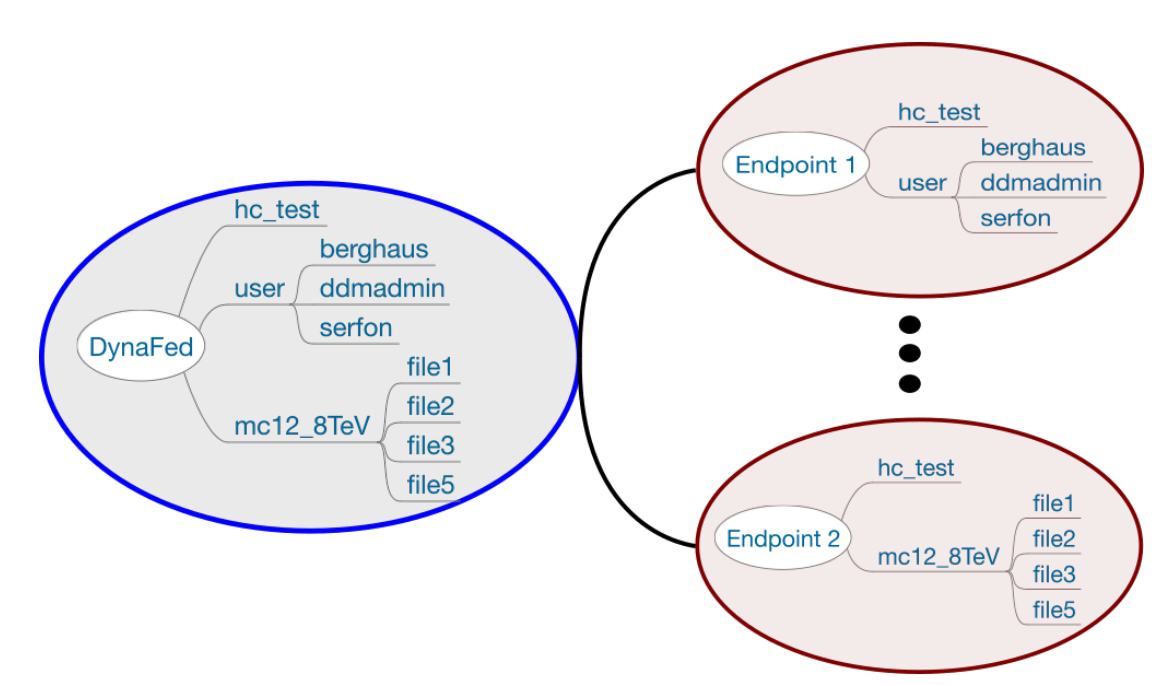
Federate existing storage



Retrieve the data from the optimal site based on location, load and network

Dynamic data federation

CERN-IT group has developed a “dynamic data federation” system (Dynafed)
http://lcgdm.web.cern.ch/dynafed-dynamic-federation-project



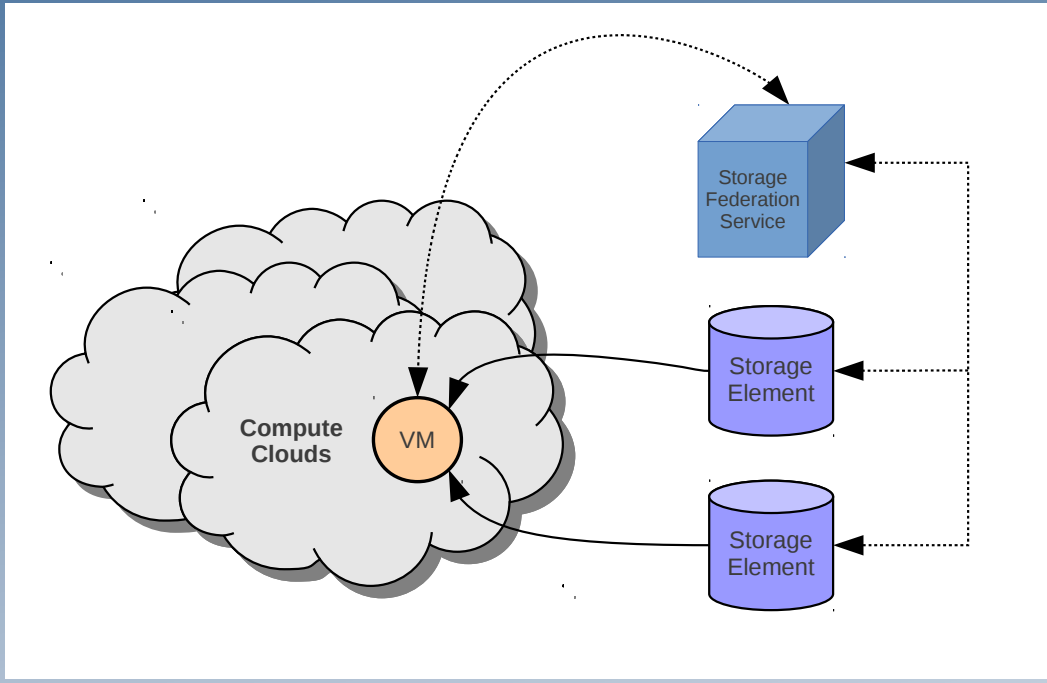
Aggregates storage and metadata farms on-the-fly

Creates (the illusion of) a unique namespace from a set of distinct storage or metadata endpoints

Exposes standard protocols that support redirections and WAN data access

Read and write support

Using Dynafed with clouds



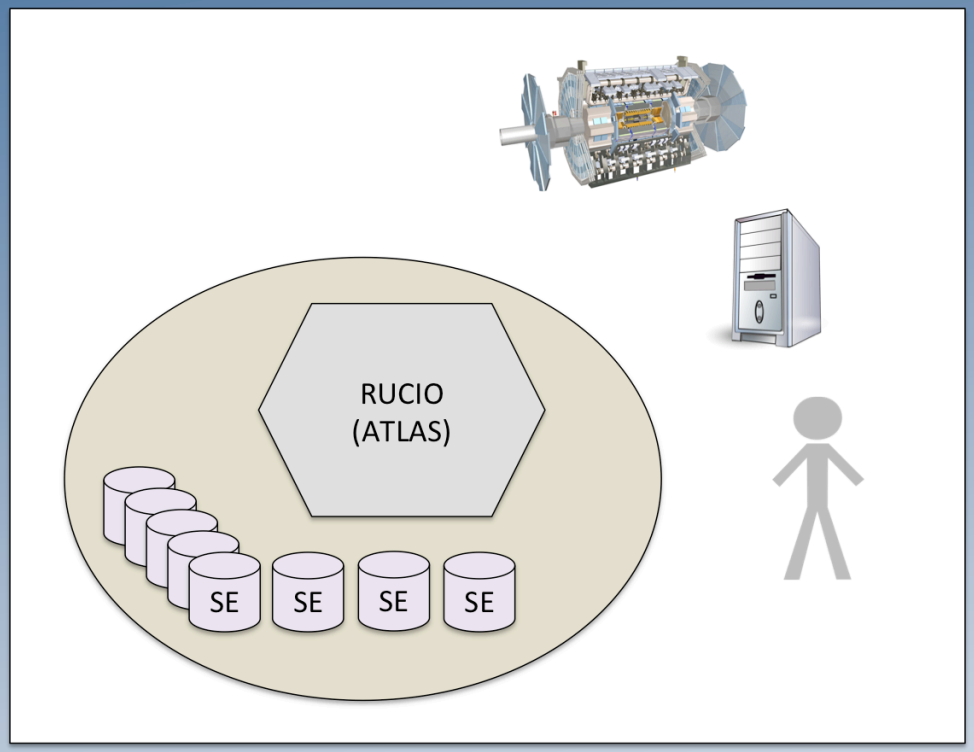
Established federations of existing “Storage Elements” (SE)

Directs the VM to the nearest storage system with the input data

GeoIP used to select the nearest site
(other information can be also used, e.g. load)

Challenge is to integrate Dynafed with the existing data management systems

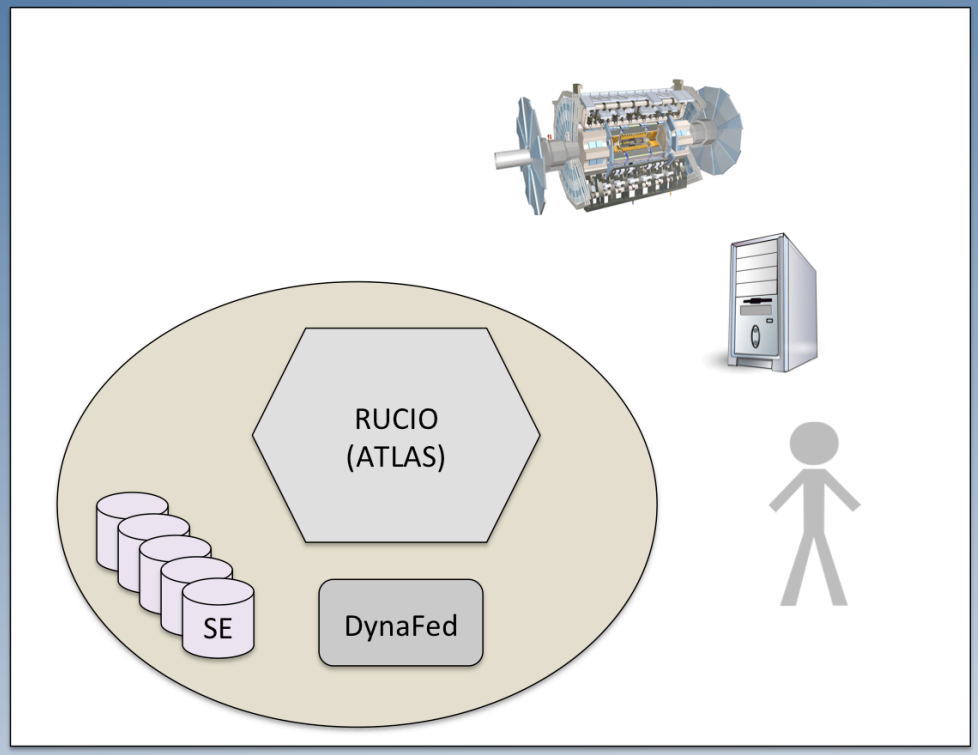
ATLAS data management



RUCIO
ATLAS data management system
(1G files, 230 PB on 130 sites)

Discover data
Transfer data
Delete data
Ensure consistency

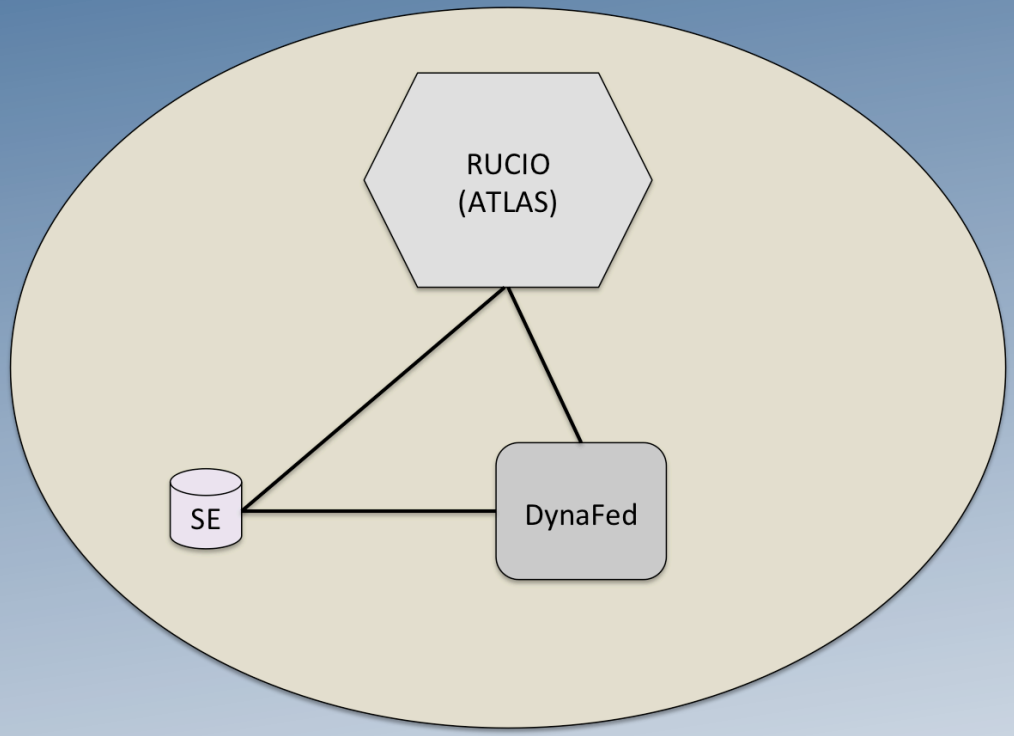
Application interface to data federation



ATLAS would like to see a consolidation of smaller or national sites into data federations

Help reduce the complexity, make more efficient use of the storage, and enable use opportunist computing

Integration challenges



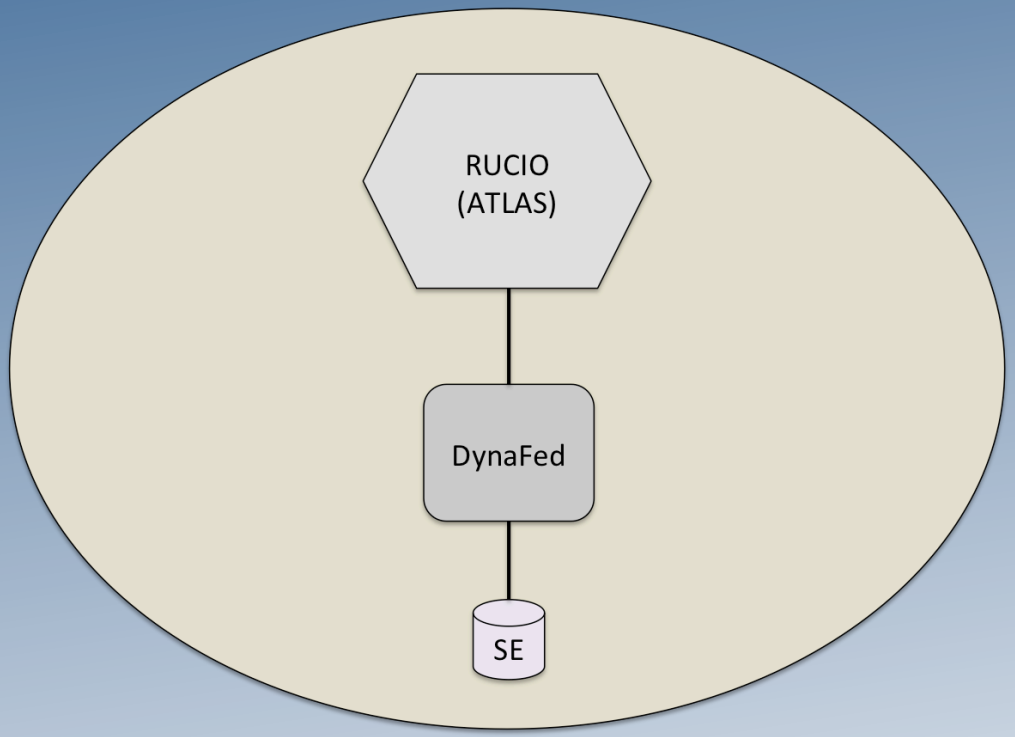
Making a data federation is easy

Integrating the federation with the existing data management is complex

A single file will appear twice in this view

Making it easy to delete it or confuse RUCIO

Initial deployment



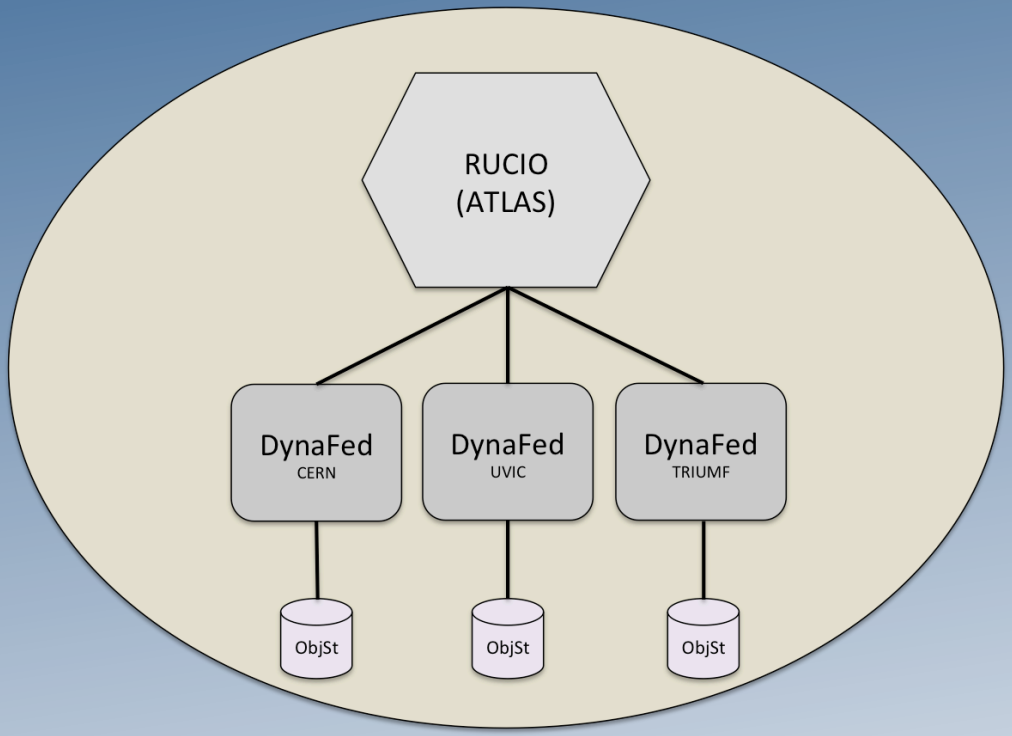
RUCIO must remain the Master

RUCIO manages the Dynafed instance
(no ambiguities)

Long term – we may be identify sites as “read-only” (volatile storage)

Then Dynafed has access but no control of data

Current status



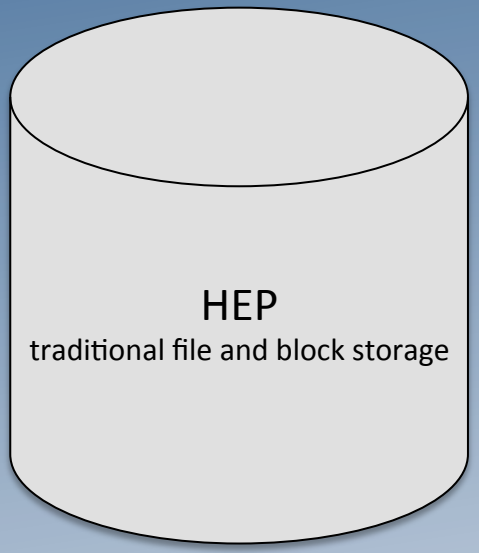
Using Ceph object storage for data
(not yet used by ATLAS for data storage)

Operational system at CERN

Nearly operational at Victoria

Starting installation at TRIUMF

Object storage



Arguments for using object storage:

Scalable – easy to add new storage

Reliable – able to replicate data

Simple (minimal) – store, copy, get, delete

Configurable - REST API's

Fast, easy access – HTTP interface

Requires changes to our data management methods
(e.g. files cannot be renamed)

Project status

- CFI funding started July 2016 (3 years)
  - 8 staff (developers and computer-HEP-physicists)
  - 1 at CERN and 1 at TRIUMF
  - 2 FTE CERN (in-kind) contribution

- Activities
  - Rewriting cloud provisioning software (10K+ cores and improved reliability)
  - Established data federations in Canada and CERN
  - Building and testing storage systems with Object Storage (CEPH)
  - Working with ATLAS data team to sort out interaction with data federation
  - Investigating how Belle II can use a data federation
  - Goal is to have a small ATLAS production system before end-2017

Summary

- Our distributed cloud computing platform is running well for the ATLAS and Belle II experiments
  - Making significant contributions to both projects
  - Ongoing developments and improvements on the core infrastructure

- CFI-funded project is completing its 1st year of the 3-year project
  - Established a strong team of developers and physicists
  - Initially focusing on the ATLAS requirements
  - Parallel development for the Belle II experiment