Summary of Track 3B

Experiment Data Processing, Data Handling and Computing Models

Nurcan Ozturk
University of Texas at Arlington

20th International Conference on Computing in High Energy and Nuclear Physics (CHEP2013)
14-18 October 2013, Amsterdam

Overview of Track 3B

- Conveners: Nurcan Ozturk & Robert Illingworth (travel cancelled)
- 27 out of 28 oral presentations have been given in this track. 6 of them were remote, went well w/o much delays (3 talks in a row)
- Thanks to all speakers for their reports and thanks to CERN colleagues for their help with convening the sessions and sending notes
- Main topics of Track 3B:
  - Workload management
  - Data management
  - Evolution of computing models

Workload Management

DIRAC Workload Management System

- In 2009 the core DIRAC development team decided to generalize the software to make it suitable for any user community.
  - Separate LHCb specific functionality into a set of extensions to the generic core libraries
  - Introduce new services to make it a complete solution
  - Support for multiple small groups by a single DIRAC installation
  - General refurbishing of the code, code management, deployment, documentation, etc

Andrei Tsaregorodtsev

DIRAC THE INTERWARE

- LHCb stays the most imp
  - Using DIRAC for all the cc
    - WMS, DMS, Data Productio
  - See presentations here, h
- Belle II
  - Combination of the non-gr grid sites and (commercia clouds is a requirement
  - 2 GB/s, 40 PB in 2019
  - Belle II grid resources
    - WLCG, OSG grids
    - KEK Computing Center
    - Amazon EC2 cloud
  - First production run is don
    - *See T.Kurh's presentation*

DIRAC THE INTERWARE

Community installations

- ILC/CLIC detector Collaboration
  - Base production system on DIRAC
  - MC simulations
  - DIRAC File Catalog was developed to meet the ILC/CLIC requirements
  - See *poster*
- BES III, IHEP, China
  - DIRAC is chosen for the phase III
  - Using DIRAC DMS: File Catalog, Transfer services
- CTA
  - CTA started as FG-DIRAC customer for DIRAC evaluation
  - Now is using a dedicated installation at PIC, Barcelona
  - Using complex workflows
  - See *poster*
- DIRAC evaluations by other experiments
  - Fermi-LAT, LSST, Auger, TREND, Daya Bay, Geant4, …
  - Evaluations can be done with general purpose DIRAC services

9

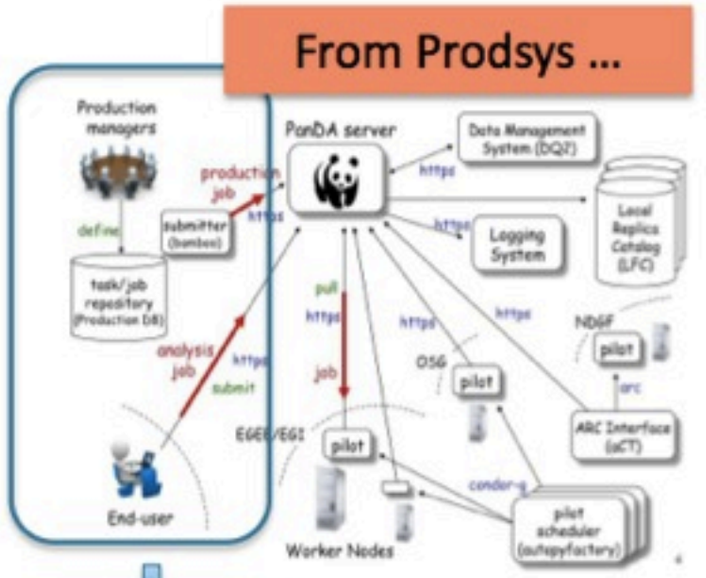
ATLAS Workload Management System

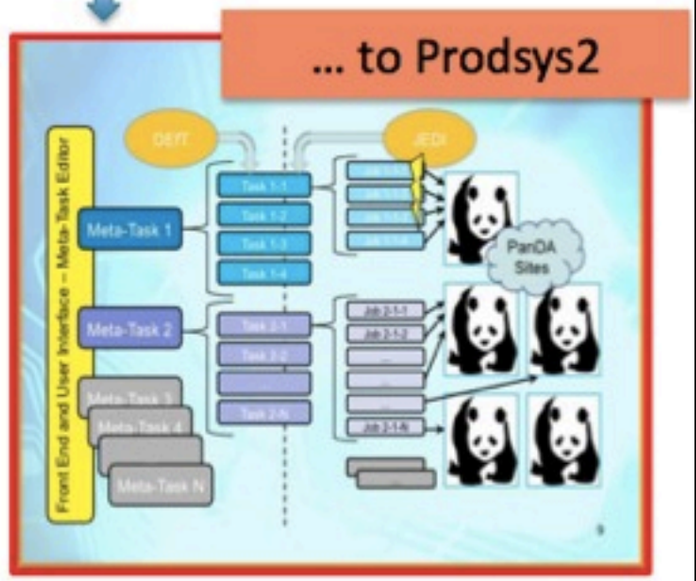
ATLAS distributed computing is evolving to meet the challenges of Run-2; trigger rate, luminosity increase, flat resource budget.

Workload Management in Run-2: Prodsys2

- Prodsys2 core components
  - DEFT: translates user requests into task definitions
  - JEDI: dynamically generates the job definitions
  - PanDA: the job management engine

- Features:
  - Provide a workflow engine for both production and analysis
  - Minimize data traffic (smart merging)
  - Optimized job parameters to available resources





CMS Workload Management System

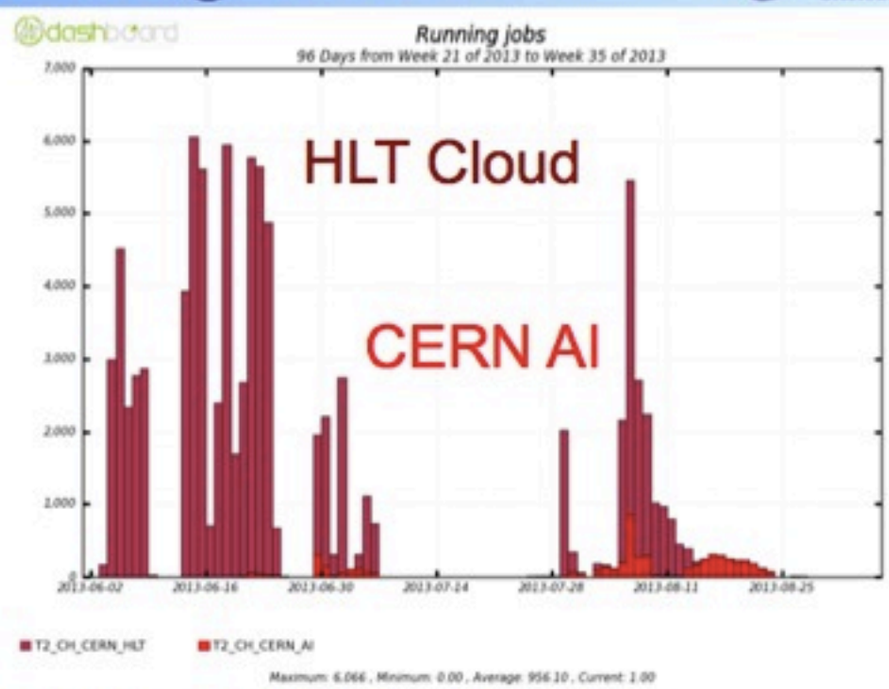
CMS workload management system is evolving to support Clouds and opportunistics resources.

Claudio Grandi

Workload Management

Separation of resource allocation and job management

- Via the glidein-WMS
- Support of Clouds and opportunistic resources in addition to Grids are natural extensions



Use of CVMFS and remote data access are key elements for an easy adaptation of the system to Clouds and opportunistic resources

BOSCO is a thin layer that allows submission of glidein through an ssh gateway to opportunistic resources

PanDA Workload Management System for Exascale Computational Science

Tadashi Maeno

BigPanDA project

BROOKHAVEN NATIONAL LABORATORY

Evolving PanDA for Advanced Scientific Computing

- The interest in PanDA by other big data sciences provided the primary motivation to generalize the PanDA system
- A project to extend PanDA as meta application, providing location transparency of processing and data management, for HEP and other data-intensive sciences, and a wider exascale community
- 3 FTE for 3 years from 2012
- Three dimensions to evolution
  - Making PanDA available beyond ATLAS and High Energy Physics
  - Extending beyond Grid (Leadership Computing Facilities, Clouds, University clusters)
  - Integrating network as a resource in workload management

CHEP2013 Amsterdam, Netherlands Tadashi Maeno Oct 14 2013 7

Work Plan

BROOKHAVEN NATIONAL LABORATORY

- 3 year plan
  - Year 1. Setting the collaboration, define algorithms and metrics
    - Hiring process was completed in June 2013
    - Development team is formed (3 FTE)
  - Year 2. Prototyping and implementation
  - Year 3. Production and operations
- 4 work packages
  - WP1 : Factorizing the core
  - WP2 : Extending the scope
  - WP3 : Leveraging intelligent networks
  - WP4 : Usability and monitoring

CHEP2013 Amsterdam, Netherlands Tadashi Maeno Oct 14 2013 8

Common Analysis Framework Project

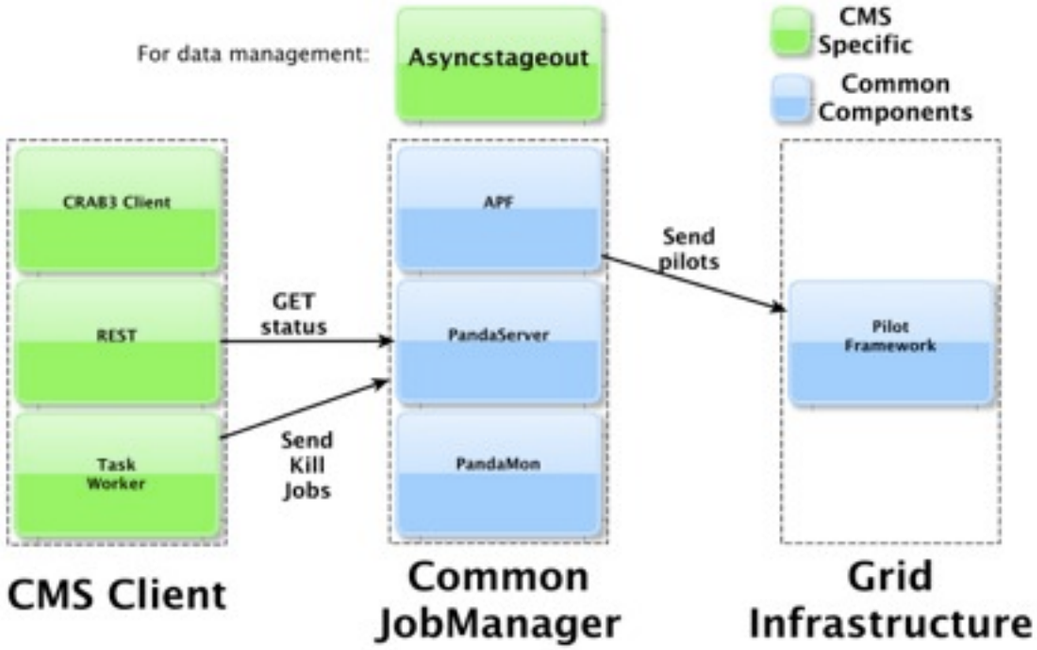
Proof of concept

The objective

To develop a common system for submitting analysis jobs to the distributed infrastructure

Based on the PanDA software

- ATLAS Production and Distributed Analysis job manager
- Able to handle 1M jobs per day (enough for CMS which only require 200k jobs)
- Stable pruduct used from many years by ATLAS



Outcome

December 2012: first basic CMS proof of concept v
system!

Open issues

The testbed showed it was possible to use a common solution between the experiments. However:

- Product still in transition from experiment specific to a service
  - Code installed directly from SVN
  - There are no DB create.sql scripts. Only cloning is possible.
  - Database uses ATLAS naming convention
- No site-level user traceability
  - User code is executed with the pilot credentials, no glExec (although there's work in progress)
- Not a clear separation between scheduling algorithm, and source code
  - Policies cannot be given as an external configuration

Data Management

FTS3 Data Movement Service

Michail Salichos

FTS3: complete rewriting, easy configuration. and represents a very welcome advancement from FTS2. Besides the fact the service is still in its infancy, it has been used for data transfer, in their production systems, by LHC experiments: mainly by ATLAS, but also by CMS and LHCb.

FTS3 – WLCG new data movement service

Many protocols and database back-ends support

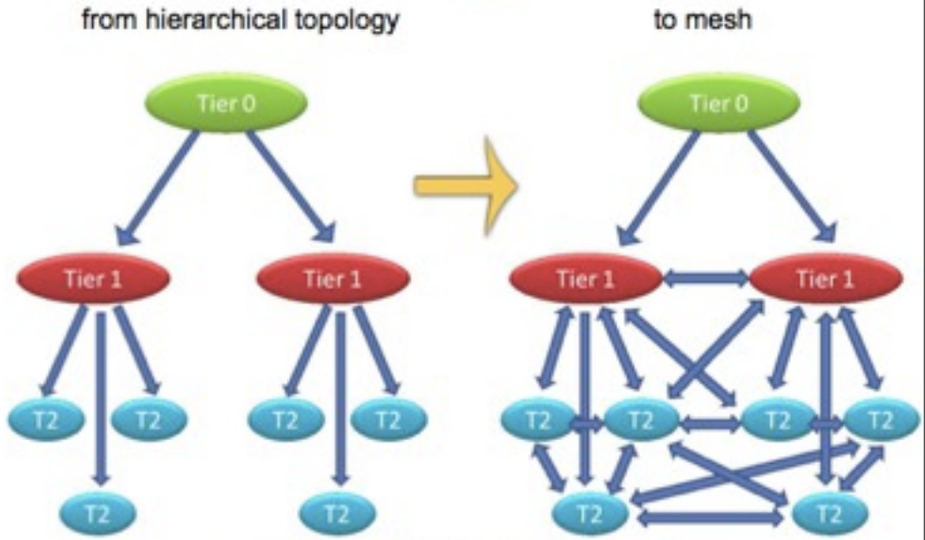
Light-weight service for heavy duty job

FTS3

No configuration needed – all optional

Sophisticated monitoring systems

FTS3 goal



... and move data at a very large scale

ATLAS Data Management (1)

Data Management in Run-2: Rucio

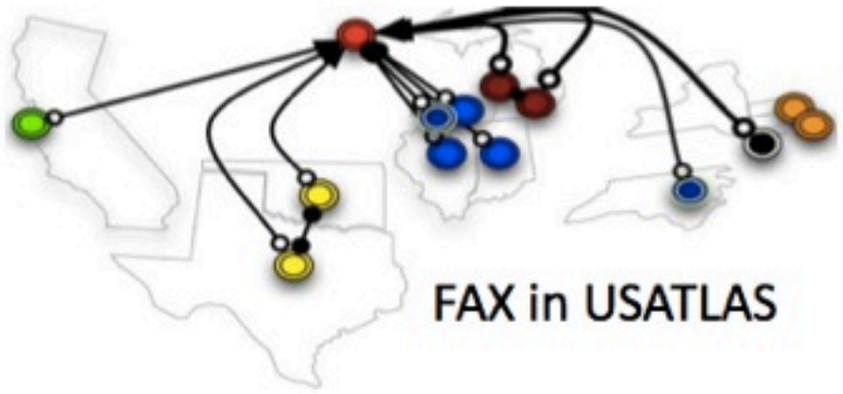
http://rucio.cern.ch/

- Implements a highly evolved Data Management model
  - File (rather than dataset) level granularity
  - Multiple file ownership per user/group/activity
- Features
  - Unified dataset/file catalogue with support for metadata
  - Built-in policy based data replication for space and network optimization
  - Redesign leveraging new middleware capabilities (FTS/GFAL-2)
  - Plug-in based architecture supporting multiple protocols (SRM/gridFTP/xrootd/HTTP…)
  - REST-ful interface

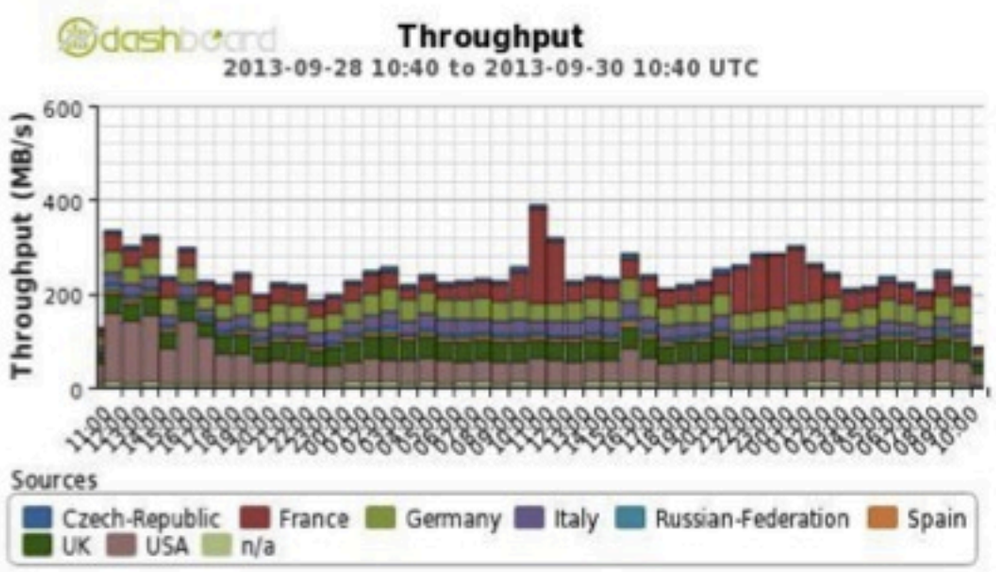
ATLAS Data Management (2)

Data Management in Run-2: FAX

- ATLAS is deploying a federated storage infrastructure based on xrootd
  - Complementary to Rucio and leveraging its new features
  - Offers transparent access to "nearest" available replica
  - The protocol enables remote (WAN) direct data access to the storage
  - Could utilize different protocols (e.g. HTTP) in future



- Scenarios (increasing complexity)
  - Jobs failover to FAX in case of data access failure
    - If the job can not access the file locally, it then tries through FAX
  - Loosening the job-to-data locality in brokering
    - From "jobs-go-to-data" to "jobs-go-as-close-as-possible-to-data"
  - Dynamic data caching based on access
    - File or even event level



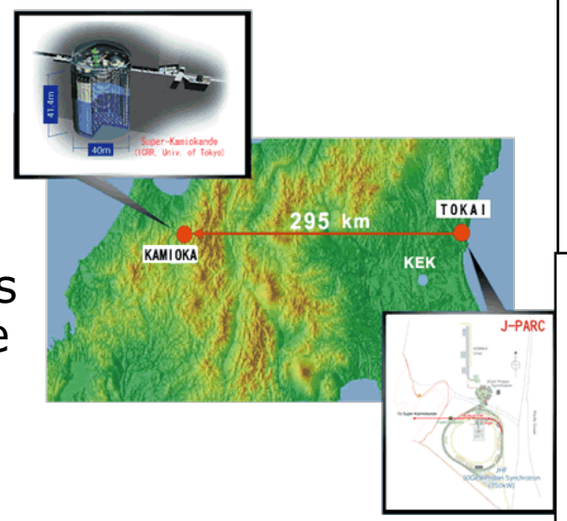
KEK iRODS Data Management System

Wataru Takase

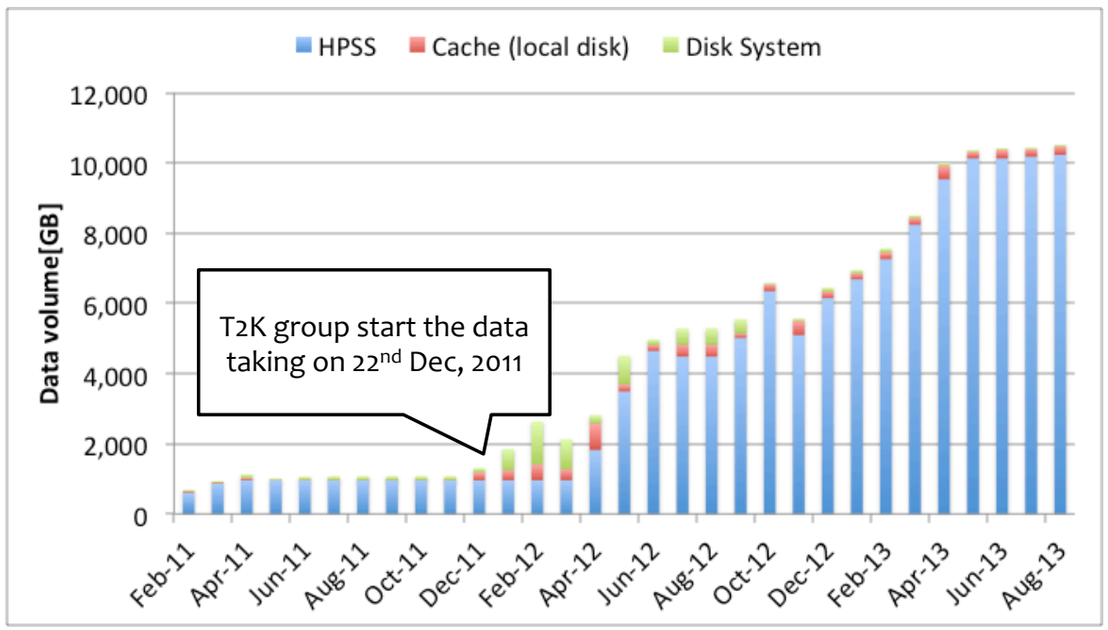
IROD: Integrated Rule-Oriented Data System

Data Management for T2K

- Tokai to Kamioka (T2K) Neutrino experimental group
- The experimental data is stored to KEK storage
- The group needed to provide an easy way to quickly access data collected to evaluate the quality of the data from outside of KEK
- iRODS provided the solution



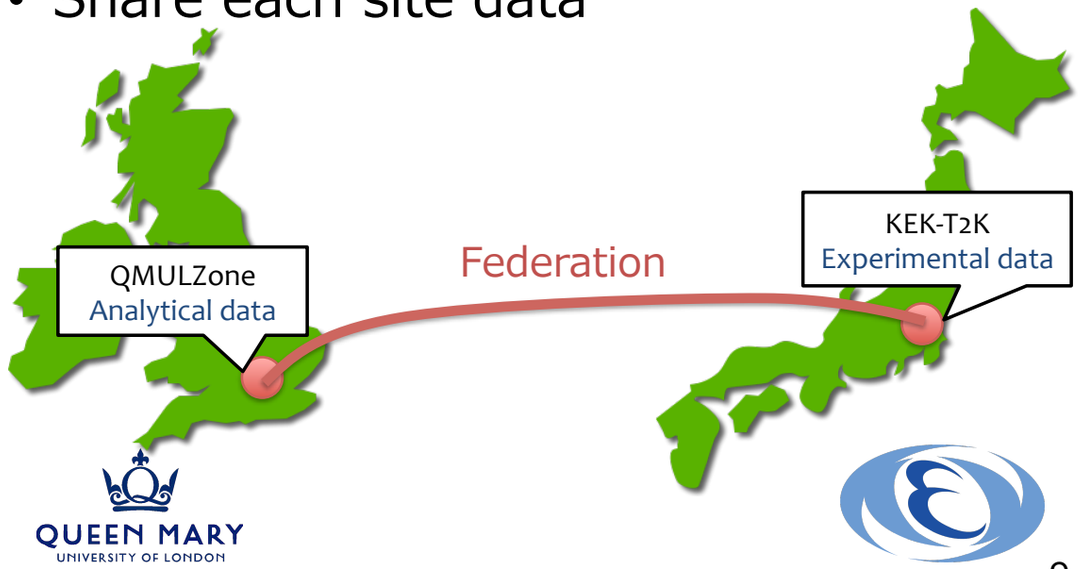
Amount of data in KEK-T2K



10 13

Federation with QMUL

- Data replication among 2 sites
- Share each site data

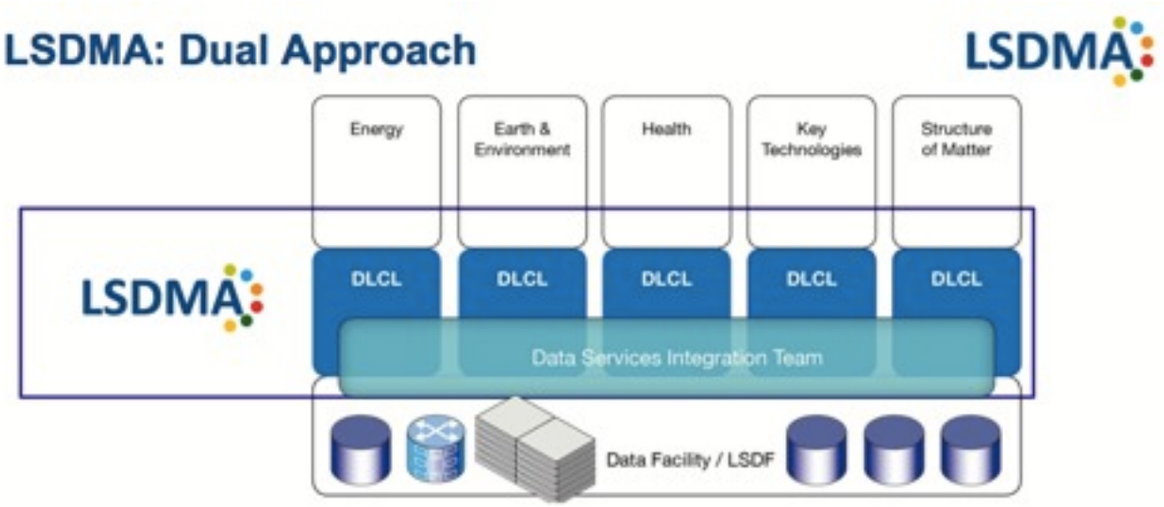


9

LMSDA Project

The Large Scale Data Facility (LSDF) aims to address the data management and processing requirements of several forthcoming data intensive experiments in Germany. LSDMA (Large Scale Data Management & Analysis) builds on services and infrastructure the LSDF provides.

Christopher Jung



Data Life Cycle Labs

Joint R&D with scientific user communities

- Optimization of the data life cycle
- Community-specific data analysis tools and services

Data Services Integration Team

Generic methods R&D

- Data analysis tools and services common to several DLCLs
- Interface between federated data infrastructures and DLCLs/communities

LSDMA Facts & Figures

LSDMA

- Initial duration: 2012-2016
  - Project is a Helmholtz portfolio extension → inclusion of activities into Helmholtz program-oriented funding in 2015, cross-program initiative
- Partners:
  - Helmholtz Association: KIT, DESY, FZJ, GSI
  - External: DKRZ, U-Heidelberg, U-Ulm, TU-Dresden, U-Hamburg, HTW-Berlin, U-Frankfurt
- Coordination: KIT

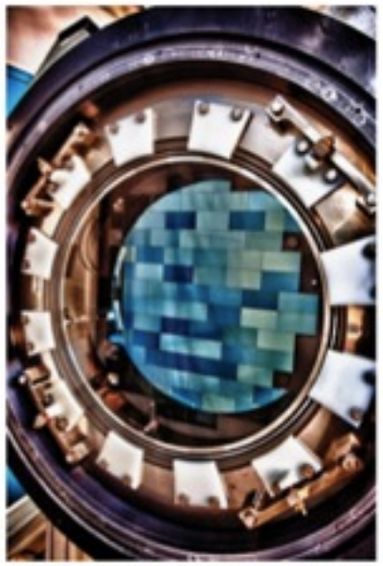
HELMHOLTZ ASSOCIATION

Data Management for Dark Energy Survey Project

Donald Petravick

DECAM and the 4m Blanco Telescope

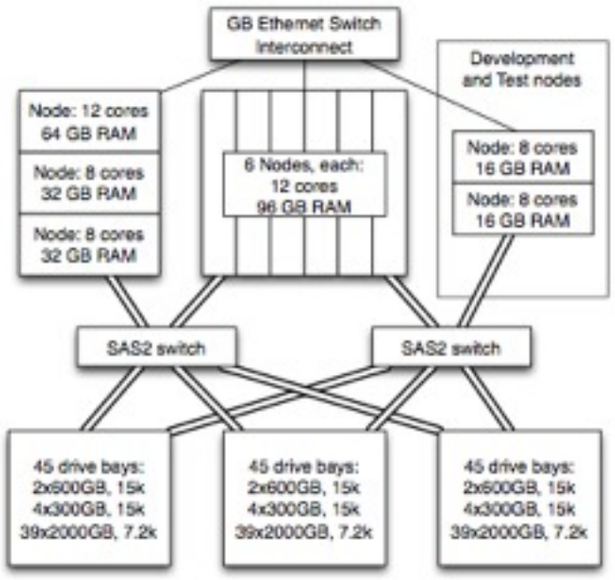
DARK ENERGY SURVEY



Database Cluster

DARK ENERGY SURVEY

- 11 Nodes
- 128 cores
- 736GB RAM
- 135 drives
- 241TB storage
- 6GB SAS2 storage fabric



Data Distibution Mechanisms

DARK ENERGY SURVEY

- Philosophy – simple tools anyone can use.
- DB –
  - User accounts in Oracle DB.
  - Each user has own schema, can see common schema.
  - Operations and Science Schemas.
    - Science schema are optimized for read access, queries are transparently parallelized.
- Files –
  - The main access is https
  - Have already been thanked for the slowdown due to the (s).
  - Have prototype code to redirect to http via temporary links.

A Web Based Data Catalog for Data Access and Analysis

Brian Van Klaveren

Data Catalog Overview

SLAC

- Initially developed for Fermi Gamma-Ray Space Telescope (Fermi-LAT)
- Designed to be experiment independent
- Development started 2007, in use since launch 2008
- Written in Java (JSP for web)
- Now in use by other experiments (EXO, LSST)

What is it?

SLAC

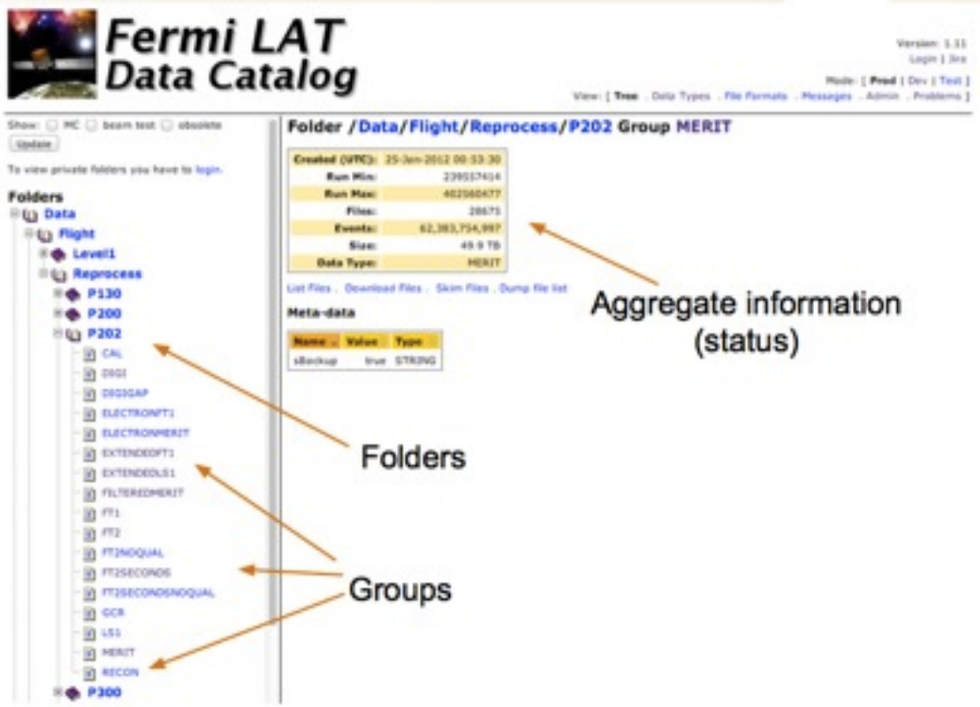
It is a metadata database for your files.

And that metadata is stored in a virtual hierarchy.

It is not tied to any processing system.
It is not tied to any file system or protocol.
It needs only an RDBMS and servlet container.

Current web application

SLAC



Conclusion

SLAC

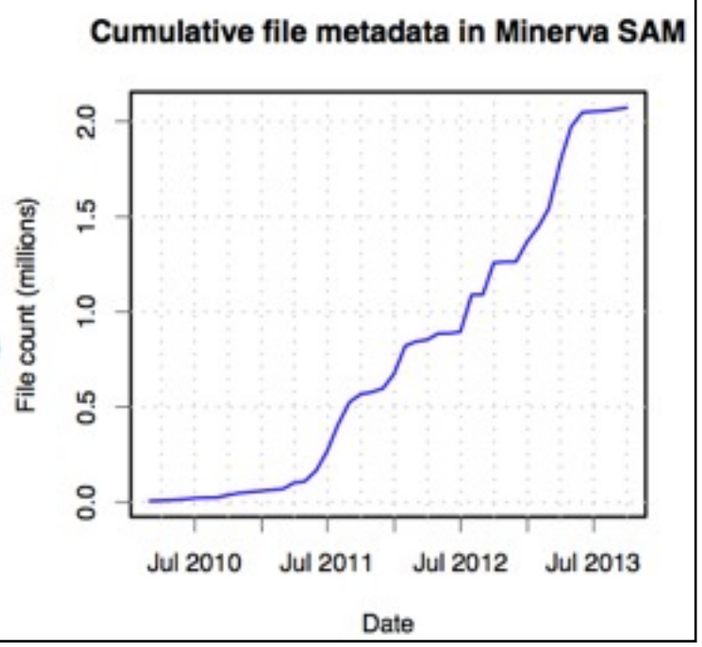
- The Data Catalog we've made for Fermi-LAT is very successful and used heavily across several experiments
- Generalizing the core and adding a plugin architecture allows us to support the needs of different experiments going forward
- RESTful interfaces are a good way to support many languages without repeating yourself, and a good way to support off-site use of our Data Catalog
- RESTful interfaces also make it easy to create responsive and adaptable human interfaces for data browsing and retrieval
- HTML5 technologies coupled with RESTful interfaces can enable us to create responsive web pages that scale

SAM Data Management System

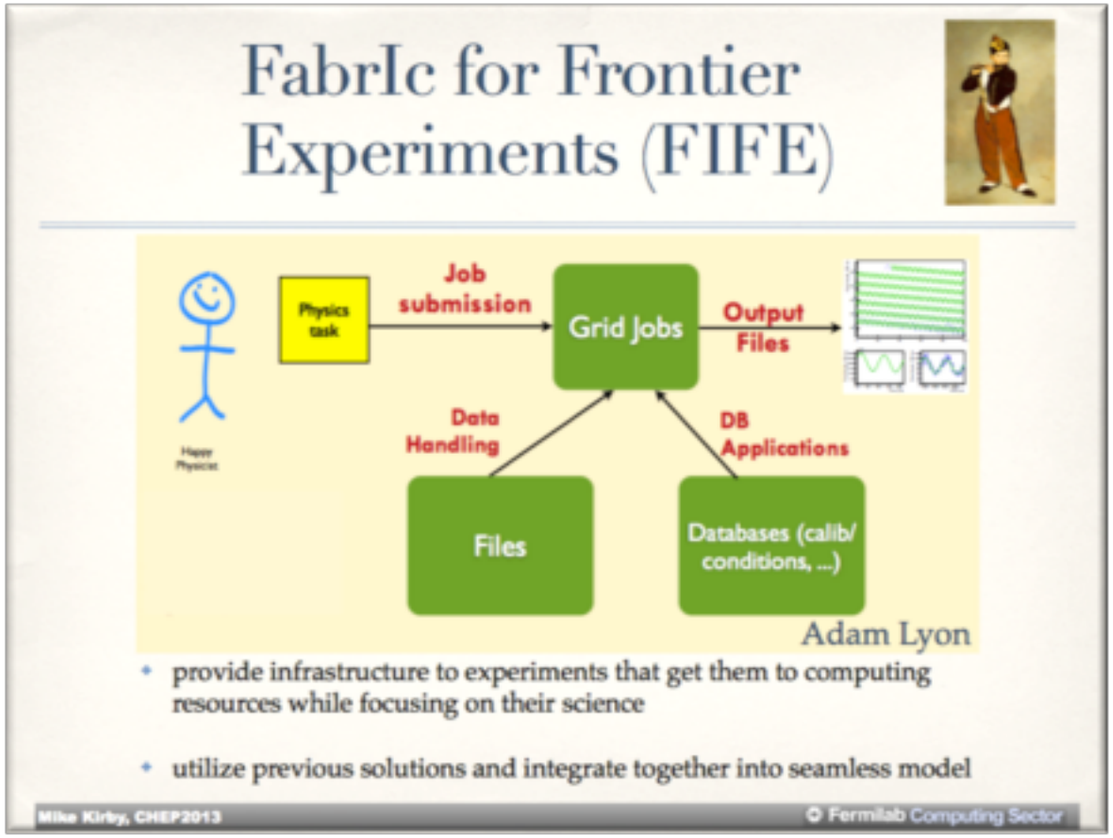
Robert Illingworth

- SAM (Sequential Access via Metadata) was originally begun as a data handling system for Run II of the Tevatron
- Used very successfully by CDF and D0 (metadata for over 207 million files; over 600 physics papers published)
- Now modernizing the architecture for the next decade of operations
  - build on extensive past experience
  - make use of modern technologies

- In full use
  - Minos (~100 TB/yr)
  - Minerva (~10 TB/yr)
  - NOvA (~PB/yr)
- Deploying
  - MicroBoone (~100 TB/yr)
  - Darkside (~100 TB/yr)
  - LBNE (many PB/yr)
- Planned
  - g-2 (~1 PB total)



SAM as part of FIFE Project at Fermilab



Public Storage for OSG

Tanya Levshina

Motivation for the OSG Public Storage

Goals:

- Manage opportunistic storage provided by OSG sites.
- Help small Virtual Organizations with grid jobs data handling.

Problems:

- The common tools for automatic management of allocated storage do not exist.
- Small VOs have difficulties finding appropriate storage, verifying its availability, and monitoring its utilization.
- The involvement of a Production Manager, site administrators, and VO support personnel is required to allocate or rescind storage space.

IROD: Integrated Rule-Oriented Data System

iRODS integration pros and cons

- Advantages:
  - Allows a user to pre-stage data to OSG_DATA and SRM SEs via iRODS without dealing with sites, gathering scattered information about site resources, worrying about storage location and end path.
  - Provides a global namespace that has information about files location, size, etc.
  - Manages quota per VO/resource.
  - Doesn't impose any burden on the sites
- Disadvantages:
  - File pre-staging/download happens in two hops.
  - One cannot utilize iRODS features fully because of the architecture we are using:
    - We need to write and maintain custom scripts
    - Cannot achieve same performance

Use Cases

- SNOWMASS (Simulate hundreds of millions of high-energy proton-proton collisions, which mimic the collisions expected at future hadron colliders).
  - Need to pre-stage big files (3 – 15 GB) to selected SEs.
  - Need to download these files on a worker node during job execution.
- EIC (Electron Ion Collider at BNL: Modeling the performance and optimizing the design)
  - Pattern A: Pre-stage files (1 GB) to OSG_DATA and copy files from $OSG_DATA to a worker node during job execution.
  - Pattern B: Pre-stage a file to "SRM" SEs then copy file to all worker nodes.
- DetectorDesign (Medical Imaging, University of New Mexico: Investigating how different simulated SPECT system geometries can affect reconstructed images)
  - Upload output files to a local/remote storage from a worker node.
  - Download all the files from various SEs to user's laptop.

Summary

- The OSG still doesn't have a generic approach for public storage. A pressing need to provide a data handling solution for small VOs is mounting.
- Integration with iRODS seems to provide a feasible solution for accessing and managing public storage at the OSG sites.
- The iRODS scalability problems need to be addressed before we can move to production deployment and offer it as a common solution for the OSG small VOs.

Evolution of Computing Models

Multicore Jobs

Going multicore

Advantages for multicore jobs:

- Fully **exploit future CPU capabilities**, adapting code to new architecture designs
- Reduced **memory consumption** per core, as memory is shared between threads
- Reduced **number of jobs** to be handled by our Workload Management System
- **Output files** of larger size requiring less managing and merging operations

Multicore jobs will be intensively used in the near future: need scheduling strategies to handle them

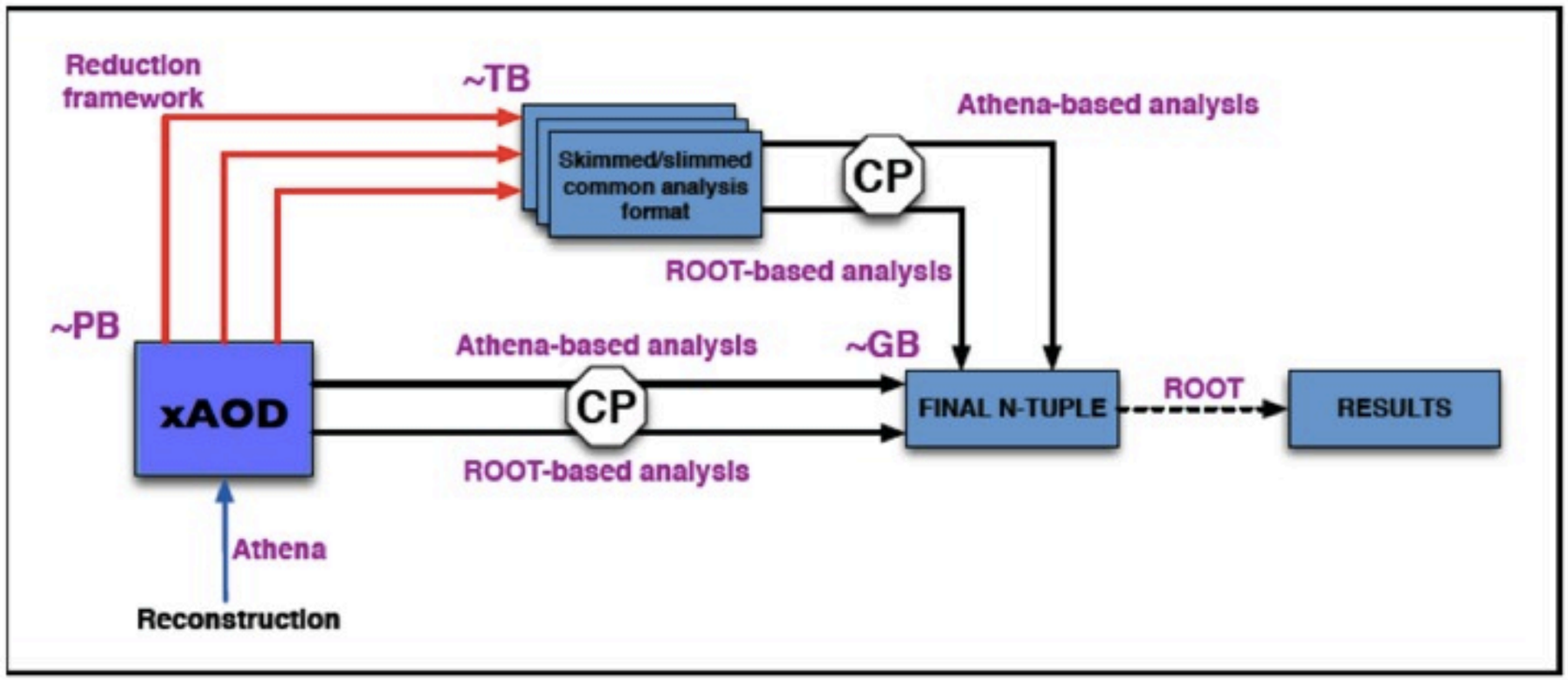
- CMS multicore **scheduling strategy** is under development, based on the idea of **multicore pilot with dynamic allocation** of internal slots
- Principle has been tested and **it works**!
- Several sources of inefficiencies in scheduling identified
- Room for efficiency improvement in terms of both new features and fine tuning
- Test during 2014 first with single core jobs, then multicore application
- **Objective: multicore application and scheduling strategy ready for LHC restart by 2015!** 20

Antonio Maria Perez Calero Yzquierdo

ATLAS Analysis Model Evolution

Paul Laycock

The run 2 analysis model - Trains



The run 2 analysis model will keep the best features of the run 1 model while optimising the computing resource usage

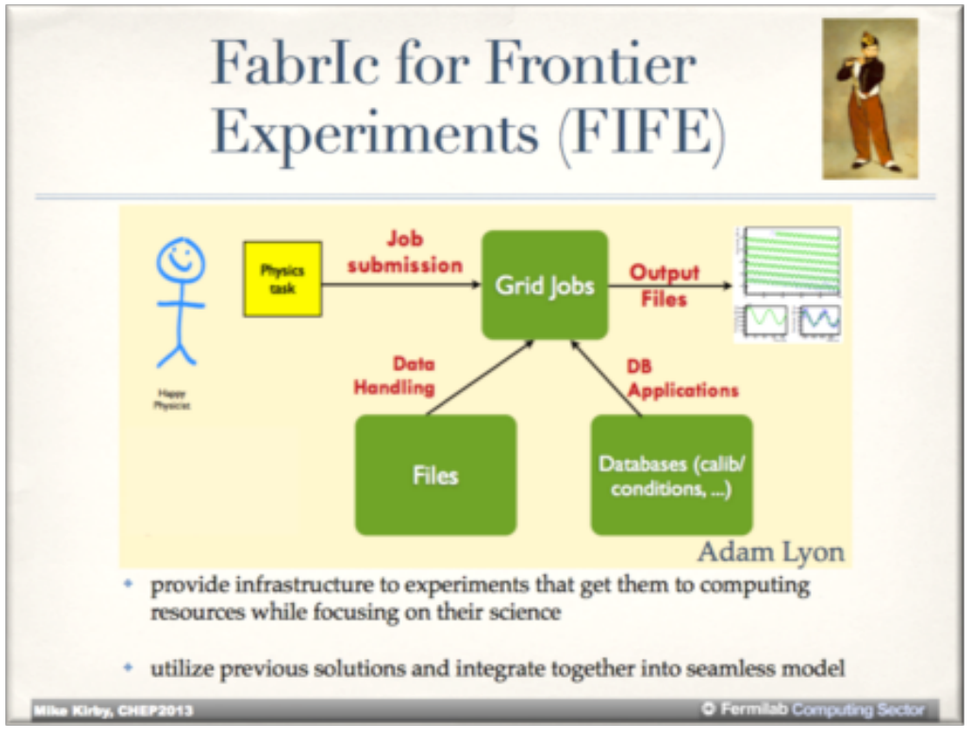
- Reconstruction output will be directly ROOT-readable
- The Train model will be used in DPD production
- The AOD software will be allowed to be updated

FIFE Project at Fermilab

Fabric for Frontier Experiments project at FermiLab

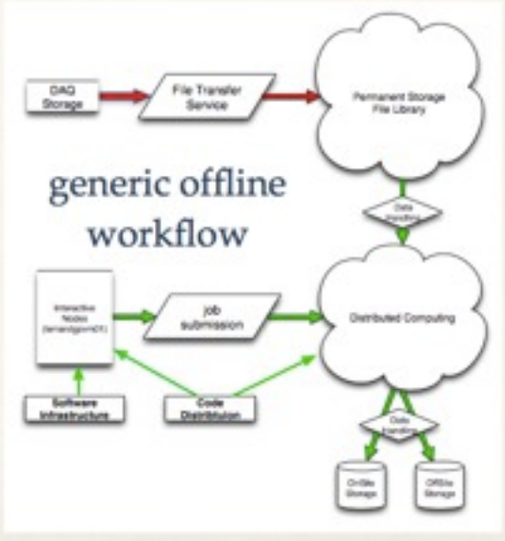
Mike Kirby

- M. Kirby explained the Physics program in the coming decade at FNAL
  - Exposed which are the computing needs
  - Strategy for providing computing resources
  - Ease the transition to OSG for small projects
  - Introduced the FabrIc for Frontier Experiments (FIFE) model and strategy and how current services are integrated
  - This offers data handling, job submission, build and code distribution, among other features
- This new approach helps 'small' VOs to easily integrate in OSG, hence increasing the the number of 'Happy Physicists' around



FIFE Strategy

- address all of the computing needs for experiments
- modular enough so that experiments can take what they need
- well enough designed so that while underlying solution may change, interface will be consistent
- provide mechanism for feedback from experiments to incorporate their tools and solutions
- help experiments utilize computing beyond the Fermilab campus
- integrate new tools and resources from outside Fermilab and other communities as they develop



Plans for improvements

- FIFE architecture is currently undergoing re-evaluation and re-architecture process
- job submission infrastructure modify to client-server model
- local storage element making transition to shared dCache pools
- data handling project continues to integrate new resources without any change in user interface
- starting new integration push for several experiments in both the Intensity Frontier and Cosmic Frontier

LHCb Computing Model Evolution

Marco Cattaneo

2015: suppression of reprocessing

- During LS1, major redesign of LHCb HLT system
  - Poster "The LHCb Trigger Architecture beyond LS1"
  - HLT1 (displaced vertices) will run in real time
  - HLT2 (physics selections) deferred by several hours
    - Run continuous calibration in the Online farm to allow use of calibrated PID information in HLT2 selections
    - HLT2 reconstruction becomes very similar to offline
- Automated validation of online calibration for use offline
  - Includes validation of alignment
  - Removes need for "first pass" reconstruction
- Green light from validation triggers 'final' reconstruction
  - Foresee up to two weeks' delay to allow correction of any problems flagged by automatic validation
  - No end of year reprocessing
    - Just restripping
- If insufficient resources, foresee to 'park' a fraction of the data for processing after the run
  - Unlikely to be needed before 2017 but commissioned from the start

8

Going beyond the Grid paradigm

- Distinction between Tiers for different types of processing activities becoming blurred
  - Currently, production managers manually attach/detach sites to different production activities in DIRAC configuration system
    - In the future sites declare their availability for a given activity and provide the corresponding computing resources
- DIRAC allows easy integration of non WLCG resources
  - In 2013, ~20% of CPU resources from LHCb HLT farm
    - 6.5% from Yandex
  - Vac infrastructure — Talk 119, 11:22 Thursday, Distr. Proc. and Data Handling A
    - Virtual machines created and contextualised for virtual organisations by remote resource providers
  - Clouds — Talk 31, 13:30 Tuesday, Distr. Proc. and Data Handling A
    - Virtual machines running on cloud infrastructures collecting jobs from the LHCb central task queue
  - Volunteer computing
    - Use the BOINC infrastructure to enable payload execution on arbitrary compute resources

9

Conclusions

- The LHCb computing model has evolved to accommodate within a constant budget for computing resources the expanding physics programme of the experiment
- The model has evolved from the hierarchical model of the TDR to a model based on the capabilities of different sites
- Further adaptations are planned for 2015. We do not foresee the need for any revolutionary changes to the model or to our frameworks (Gaudi, Dirac) to accommodate the computing requirements of LHCb during Run 2

18

Belle II Computing Model

Thomas Khur

Distributed Computing System

- Based on existing, well-proven solutions plus extensions for Belle II
  - DIRAC for job management
  - AMGA for metadata

DIRAC
THE INTERWARE

ARDA METADATA CATALOGUE PROJECT

CernVM File system

  - CVMFS for software distribution (thanks to CERN and Steve Traylen for providing the Stratum-0 server, and to GridKa for the stratum-1 server)

root/svn/trunk/grid/BelleDIRAC

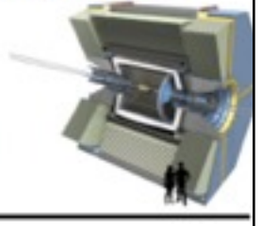
FrameworkSystem/	4326 (9 months ago) by myco: basic sites management service for BelleDIRAC
Web/	5519 (4 months ago) by hideki: remove unused AMGA API
WorkloadManageme...	6098 (2 months ago) by hideki: fix a bug
gbasf2/	6647 (2 weeks ago) by hideki: fix unnecessary AMGA initialization
README	4325 (9 months ago) by myco: init files for BelleDIRAC distribution
__init__.py	6348 (6 weeks ago) by hideki: release for 2nd MC campaign

KIT Thomas Kuhr CHEP 14.10.2013

Summary

Belle II

- Belle II will search for New Physics with O(50) times more data than current B factories
- Huge data volume is a challenge for the computing
  - Distributed computing system based on existing technologies and infrastructures
  - Workflow abstraction with projects and datasets
- First two MC production campaigns this year
  - ✔ Belle II distributed computing system works!
  - ✔ Bottlenecks and issues identified
  - Many thanks to technology and resource providers!
- Next steps:
  - MC campaign with more (cloud) sites
  - Further automatize and harden the system
  - Exercise user analysis on the grid



KIT Thomas Kuhr CHEP 14.10.2013 Page 19

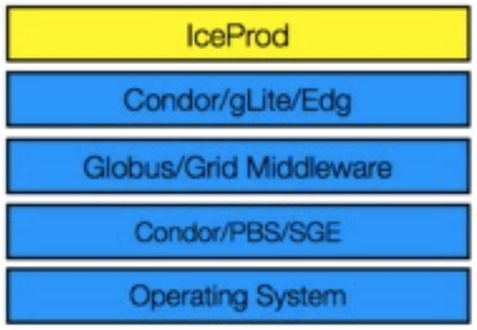
IceCube Computing Model

IceProd

IceProd is a software package based on Python, XMLRPC and GridFTP. It is driven by a central database in order to coordinate, administer and drive production of simulations and processing of data.

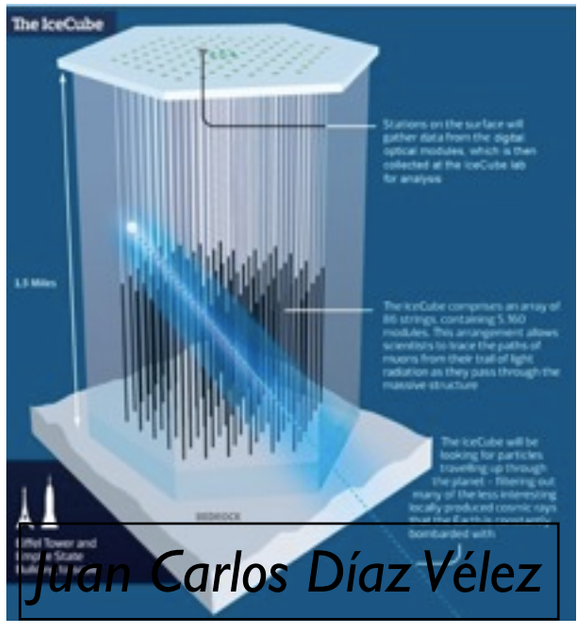
It is not a replacement for batch queuing systems or grid middleware.

IceProd runs as a separate layer on top of other middleware and can take advantage of a variety of computing resources including grids and batch systems such as CREAM, Condor, NorduGrid, PBS and SGE



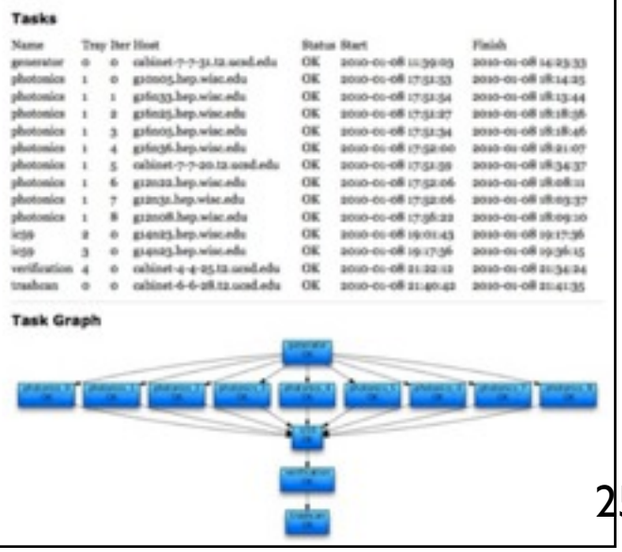
ICECUBE

Juan Carlos Díaz Vélez



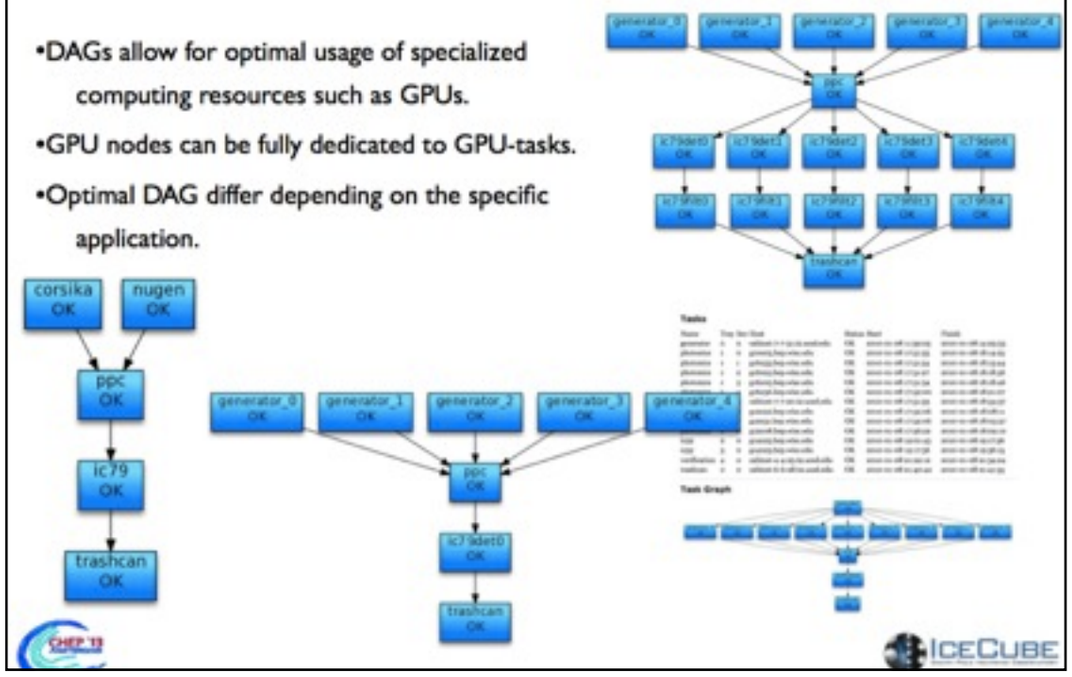
DAG (Directed Acyclical Graph) -based simulation

- Separate simulation segments into tasks
- Assign task to a node in DAG
- Different tasks can have specific hardware or software requirements



GPU-based Production

- DAGs allow for optimal usage of specialized computing resources such as GPUs.
- GPU nodes can be fully dedicated to GPU-tasks.
- Optimal DAG differ depending on the specific application.



ICECUBE

Summary

- Data handling and processing are becoming more and more challenging with increasing complexity of scientific experiments.
- Interest is growing among other big data scientific communities to use HEP-driven workload management systems. Efforts on making these systems generic (DIRAC, PanDA) for any user community.
- LHC experiments are facing the challenges of the upcoming run. Computing models are evolving to meet the requirements; new data and workload management systems are being developed. Adaptations to computing models to support use of multicore jobs, federated storage infrastructure, cloud computing and opportunistics resources.