

21st International Conference on Computing in High Energy and Nuclear Physics **CHEP2015** Okinawa Japan: April 13 - 17, 2015

## Track 4

Middleware,  
software development and tools,  
experiment frameworks,  
tools for distributed computing

Marco Clemencic  
on behalf of the conveners of Track 4

# Disclaimer

I tried to summarize all oral contributions and give a view of all the great work done. I apologize if I have missed something.

I'm very sorry that I did not manage to cover the posters.

# Overview of Track 4

- 42 orals + 48 posters
- Heterogeneous contributions
- Roughly grouped in categories:
  - Middleware
  - Framework
  - Application
  - Software
- Boundaries often fuzzy, so I reorganized them

# Overview of Track 4

- LHC experiments dominating the scene
- Very valuable contributions from
  - Non-LHC/HEP experiments
  - Service providers
  - Computing centers
  - Etc.

# Overview of Track 4

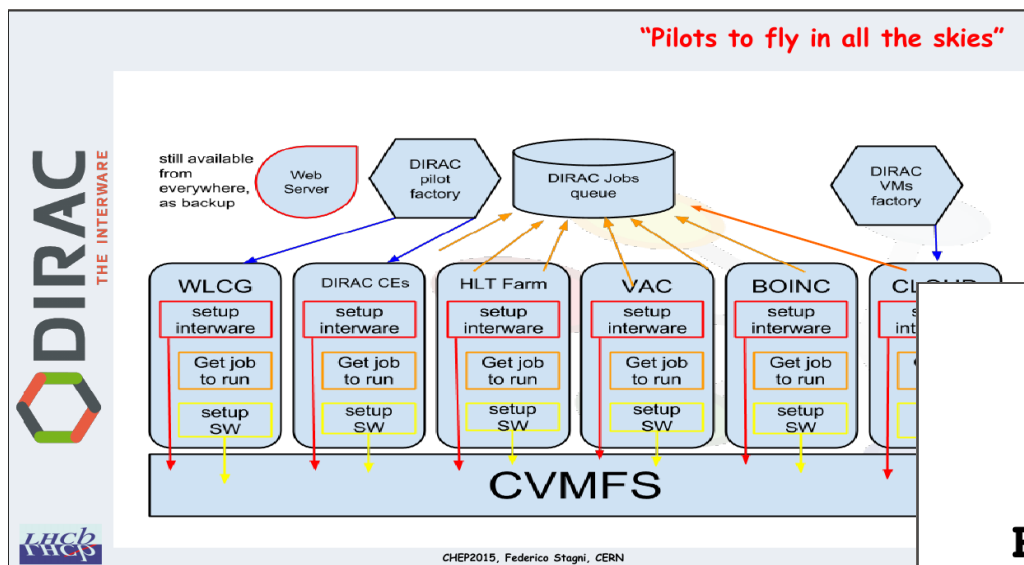
- We discussed mainly about middleware
- But also about frameworks and tools
- A lot of work has been done on improvements
  - “rewrite” is not a bad word, don't be afraid
- Sharing efforts seems the key to success

# Middleware

- Contributions on
  - Job Management/Pilots
  - Data Management
  - Network Awareness
  - Multicore

# Job Management / Pilots

- CMS and LHCb showed how pilots can bring uniformity to the Grid



#289

## Using the glideinWMS System as a Common Resource Provisioning Layer in CMS

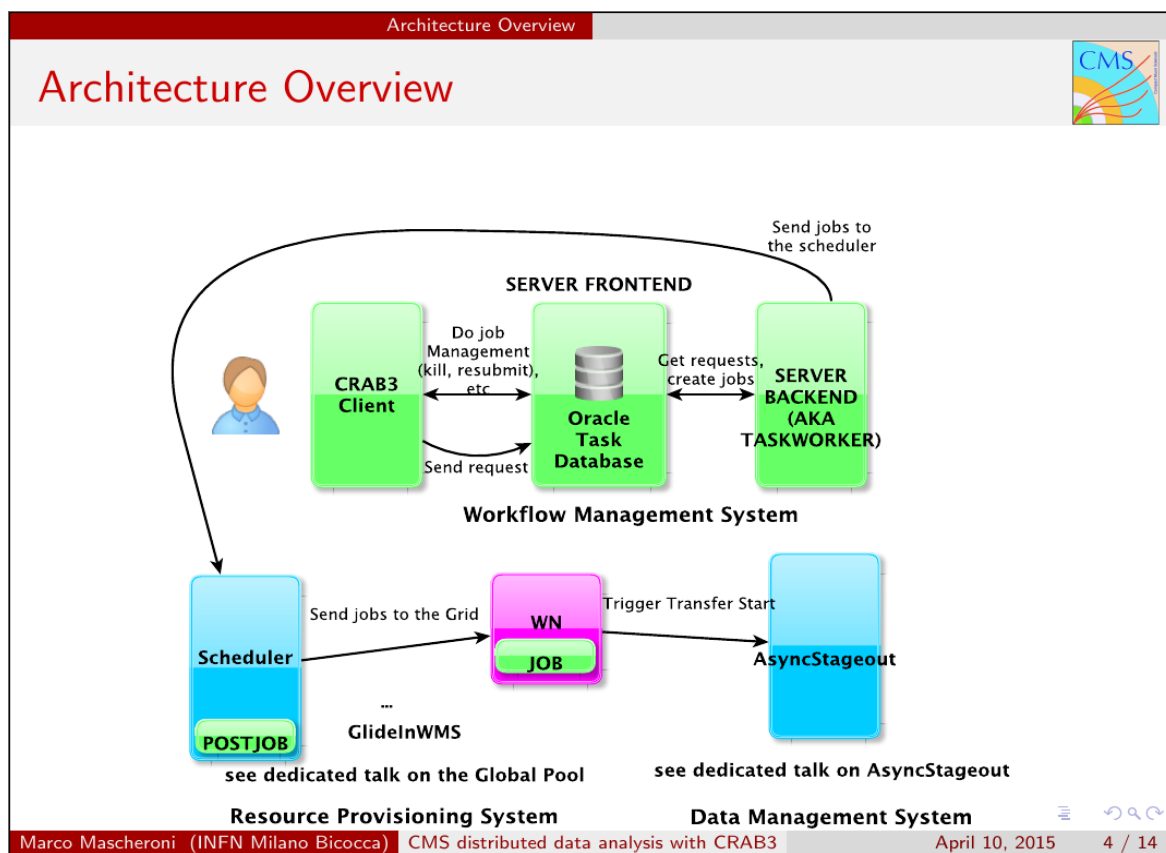
J. Balcas<sup>1</sup>, S. Belforte<sup>2</sup>, B. Boekelman<sup>3</sup>, D. Colling<sup>4</sup>,  
 O. Gutsche<sup>5</sup>, D. Hufnagel<sup>5</sup>, F. Khan<sup>6</sup>, K. Larson<sup>5</sup>, J. Letts<sup>7</sup>,  
 M. Mascheroni<sup>8</sup>, D. Mason<sup>5</sup>, A. McCrea<sup>7</sup>, S. Piperov<sup>9</sup>,  
 M. Saiz-Santos<sup>7</sup>, I. Sfiligoi<sup>7</sup>, C. Wissing<sup>10</sup>

1. Vilnius Univ. (LT) 2. INFN-Trieste (IT) 3. Univ. Nebraska Lincoln (US) 4. Imperial College London (UK) 5. FNAL (US) 6. NCP (PK) 7. UCSD (US) 8. INFN-Milano (IT) 9. Brown (US) 10. DESY (DE)

#113

# Job Management / Pilots

- CMS commissioned CRAB3




- Complete re-implementation
- Integrates with CMS new developments
  - GlideInWMS Global Pool
  - Asynchronous Stage-Out

#345



# Job Management / Pilots

- ATLAS presented an overview of the evolution of PanDA in preparation for RUN2



**The Future of PanDA in ATLAS**  
**Distributed Computing**

K. De, A. Klimentov, T. Maeno, P. Nilsson,  
D. Oleynik, S. Panitkin, A. Petrosyan,  
J. Schovancova, A. Vaniachine, T. Wenaus  
on behalf of ATLAS collaboration

University of Texas at Arlington, USA  
Brookhaven National Laboratory, USA  
Joint Inst. for Nuclear Research, RU  
Argonne National Laboratory, USA

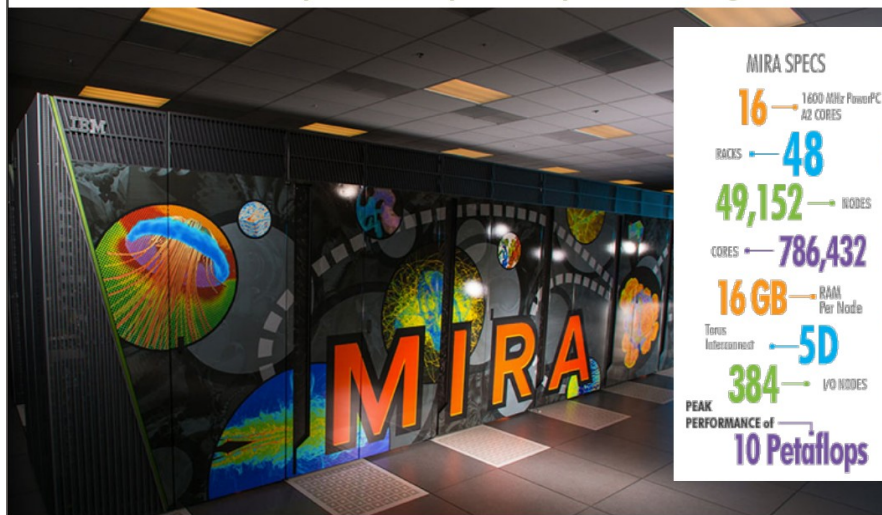
CHEP2015, Okinawa, Japan, April 13-17 2015

- Dynamic Jobs
- Network Awareness
- Event Service
- New Pilot
- Support for HPC
- New Monitoring

# Job Management / Pilots

- We have seen how MIRA became the primary Alpgen event generation site for ATLAS

Mira - Leadership-class Supercomputer at Argonne

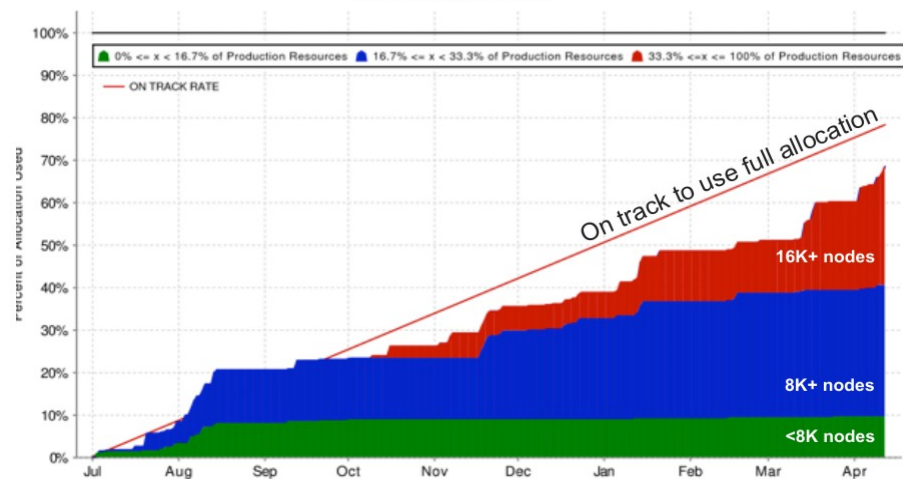


#537

via integration with PanDA

Progress using  
50M hour  
ALCC award

HadronSim  
Machine: MIRA  
Allocation: 50,000,000  
Usage: 34,234,852.67 (68.5%)  
2014-07-01 to 2015-04-12

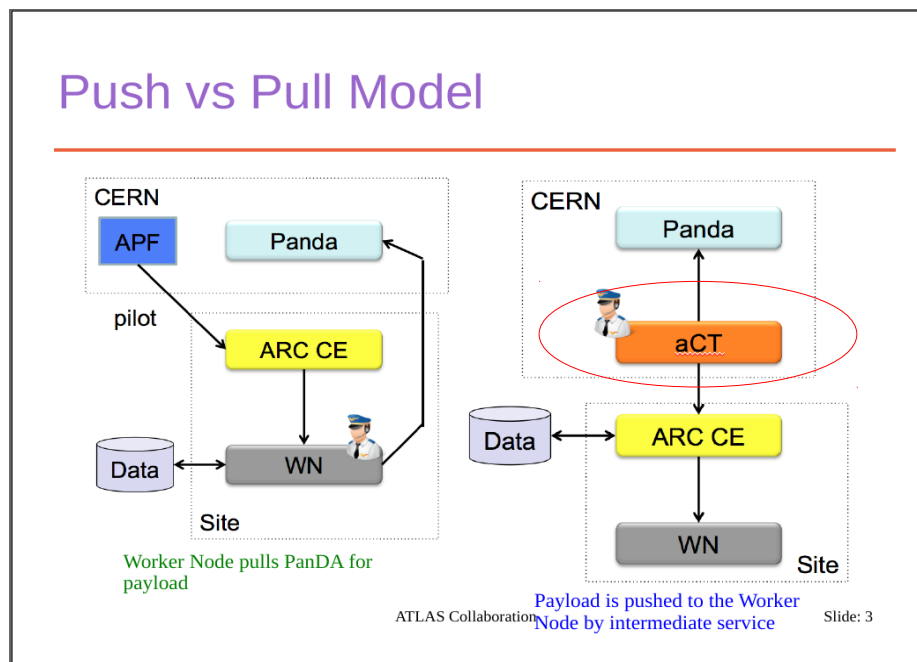


Thomas D. Uram, Argonne Leadership Computing Facility

Computing in High Energy Physics 2015

# Job Management / Pilots

- Many contributions on ARC Control Tower
  - Job Management Layer in front of ARC-CE



#145

norden NordForsk nreic Nordic e-Infrastructure Collaboration

### Example 3: Connecting Norwegian HPC sites for Life Sciences

- Problem: Galaxy only supports one cluster per instance -> no load balancing between clusters
- Solution:
  - Each site installs ARC CE
  - Galaxy pushes jobs to aCT
  - aCT takes care of load balancing as part of job management
- Requires developing an aCT plugin in Galaxy

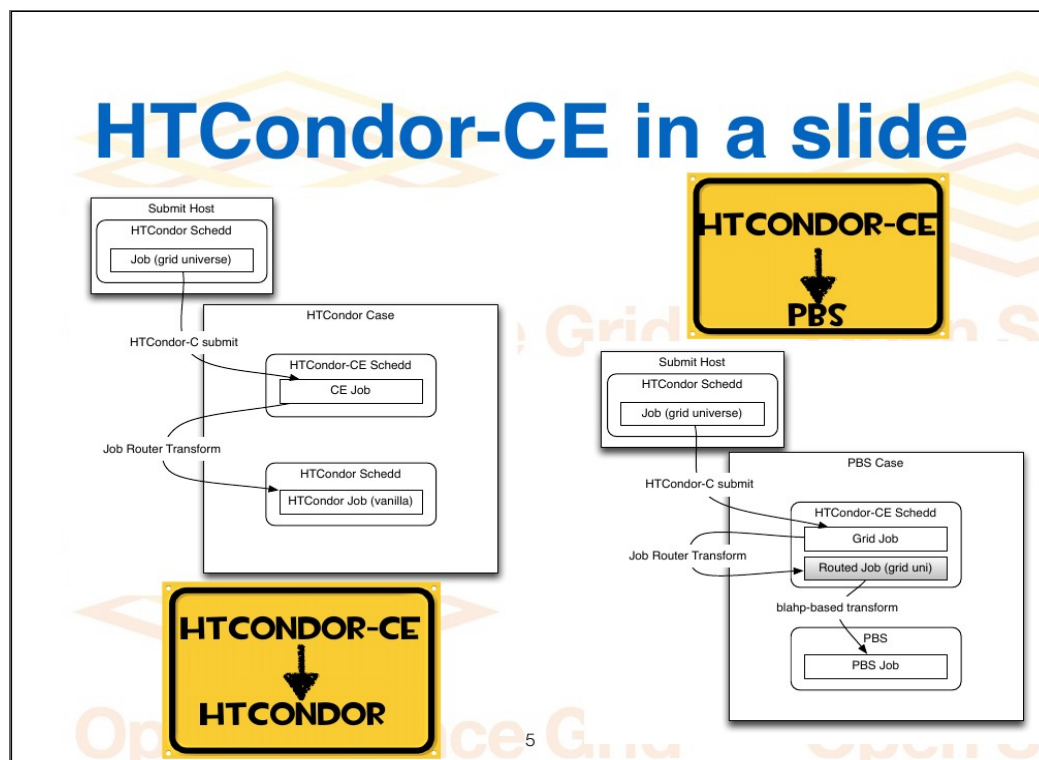
The map shows several Galaxy instances across Norway, each connected to an ARC Control Tower. These towers are connected to a central EGIIS system. Logos for NTNU and The Research Council of Norway are also present.

Figures from Abdulrahman Azab (NeLS) 10

#263

# Job Management / Pilots

- HTCondor-CE
  - use HTCondor to provide a CE interface

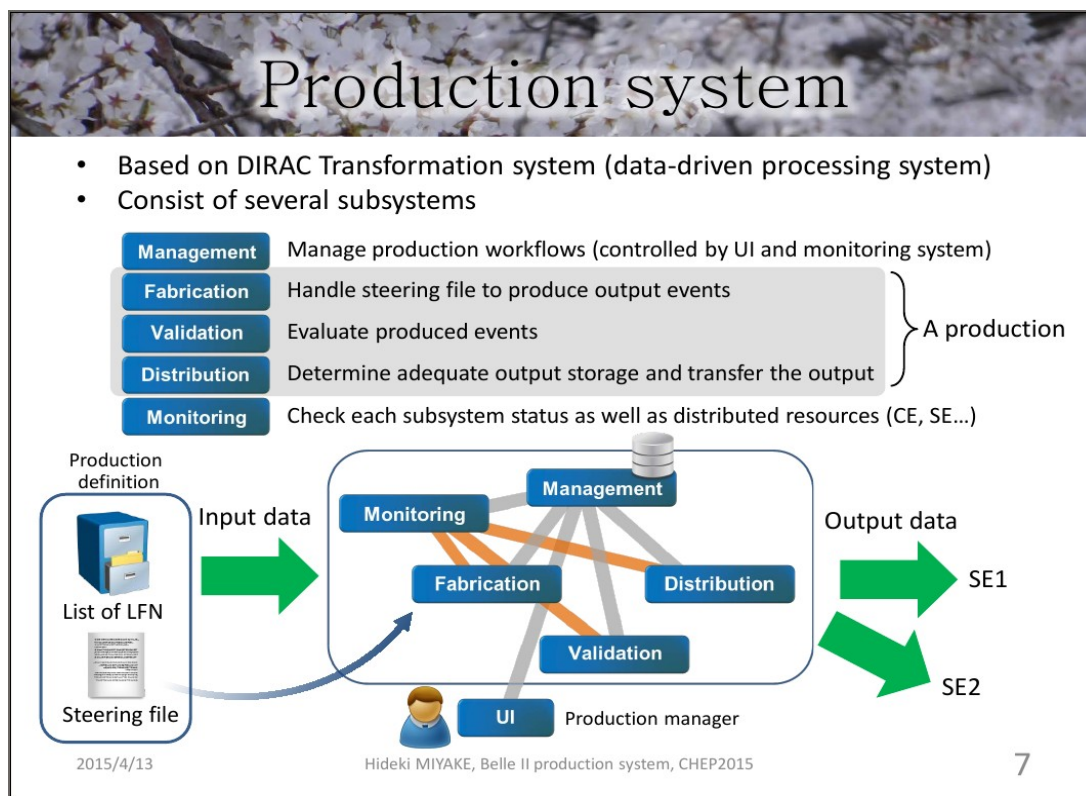


- It's a special configuration of HTCondor
- Choice strategic and technical

#519

# Job Management / Pilots

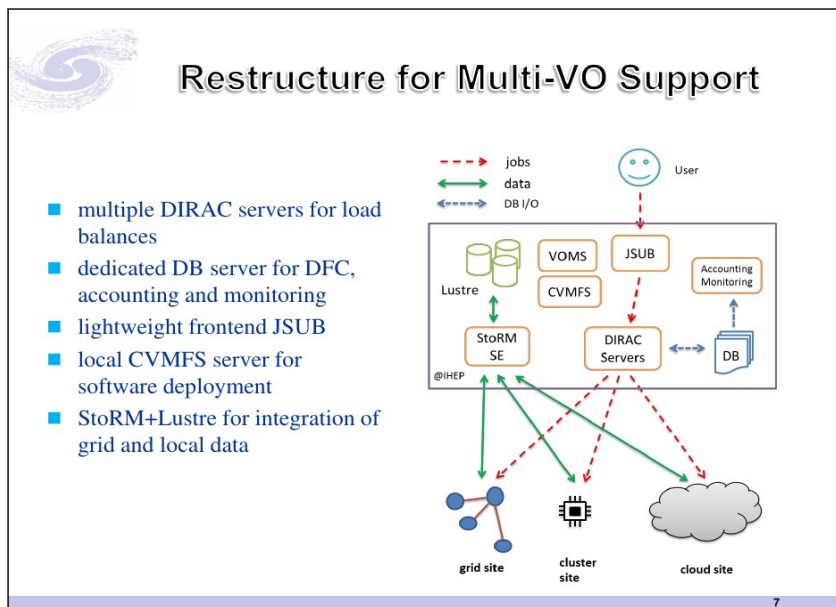
- BelleII adopted DIRAC for their Production System



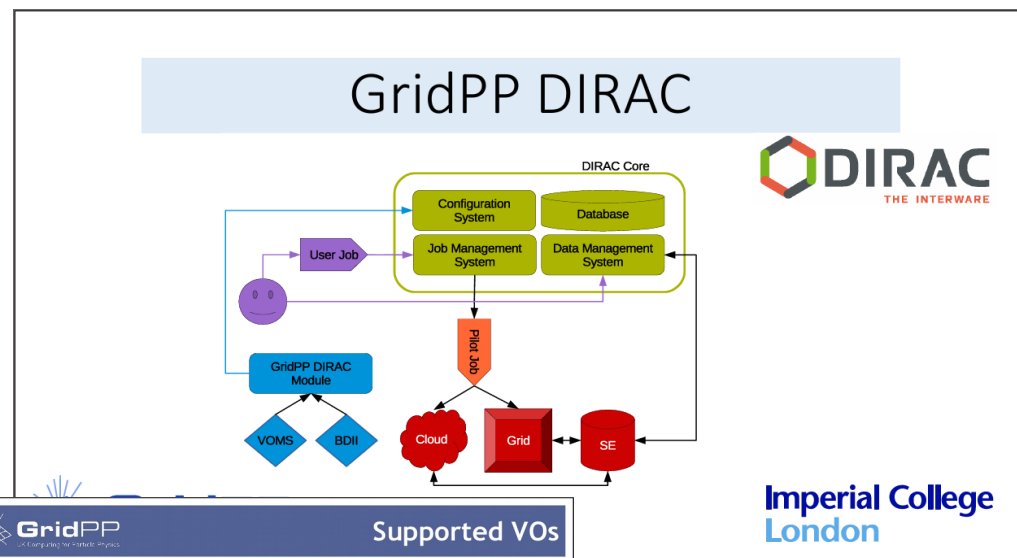
#329

# Job Management / Pilots

- IHEP and GridPP extended DIRAC to support their many (small) VOs



#479



#346

**GridPP**  
A Collaboration for Particle Physics

**Supported VOs**

The first installation was done in April 2013. We are currently supporting following VOs:

- na62.vo.gridpp.ac.uk
- t2k.org
- snoplus
- cernatschool.org
- comet.j-parc.jp
- pheno
- northgrid
- londongrid
- gridpp

13/04/2015 The GridPP DIRAC project 3

#334

# Job Management / Pilots

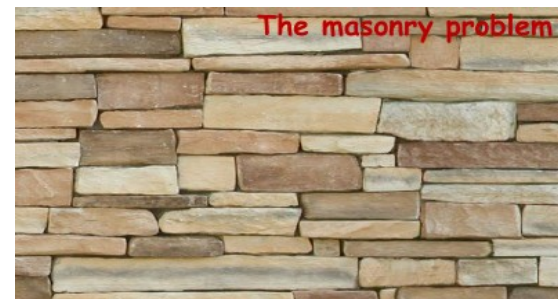
- Fermilab combined existing tools to provide a new Distributed Computing system: FIFE

### FIFE Architecture: Job Management & Software Distribution

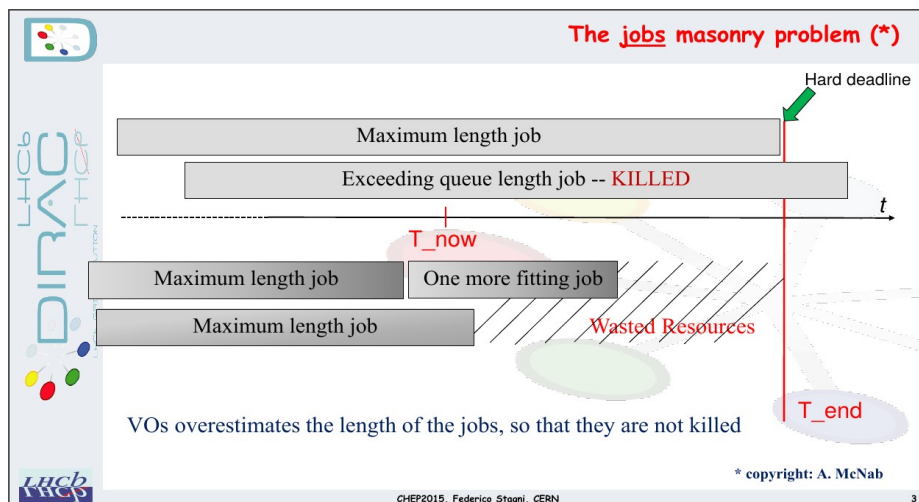
- Job Submission Tools
  - **JobSub**
    - Suite of tools to simplify & manage job submission
    - Provides scalable & Highly Available job submission service
- Resource Provisioning
  - **GlideinWMS**
    - Pilot-based WMS that creates *on demand* a *dynamically-sized overlay condor batch system* on Grid & Cloud resources to address the complex needs of VOs in running application workflows
- Software Distribution
  - **CVMFS**
    - Network file system based on HTTP
    - Optimized for software distribution in a fast, scalable and reliable way
- Data Transfer Client
  - **ifdh**
    - FIFE Data handling tool
    - Provides common interface for transferring data to/from different storage services like file systems, SAMWeb, dCache, Amazon S3

# Job Management / Pilots

- LHCb and ATLAS addressed the “masonry problem”



#112

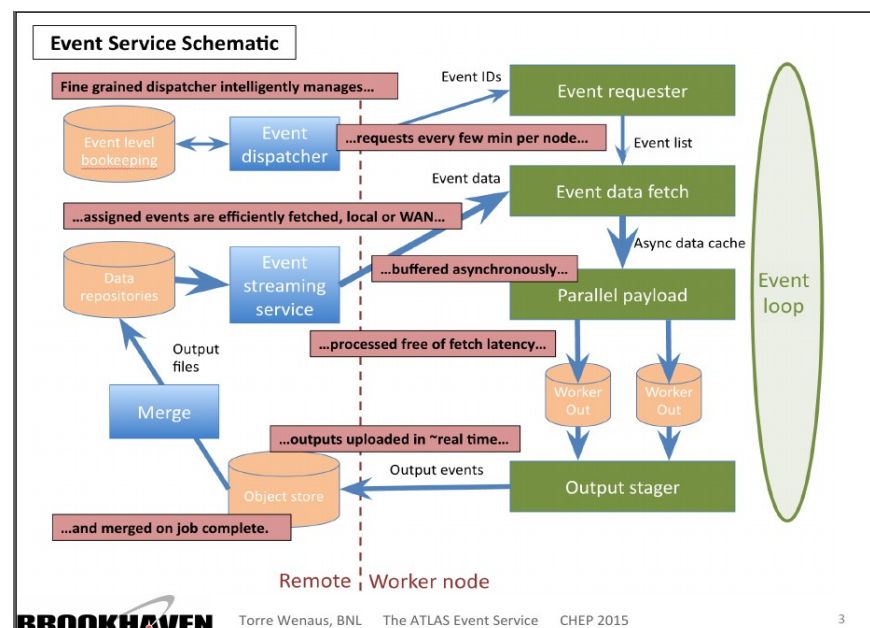


#112

LHCb can gracefully stop simulation jobs just before the allocated time is over.

ATLAS distributes single events to workers.

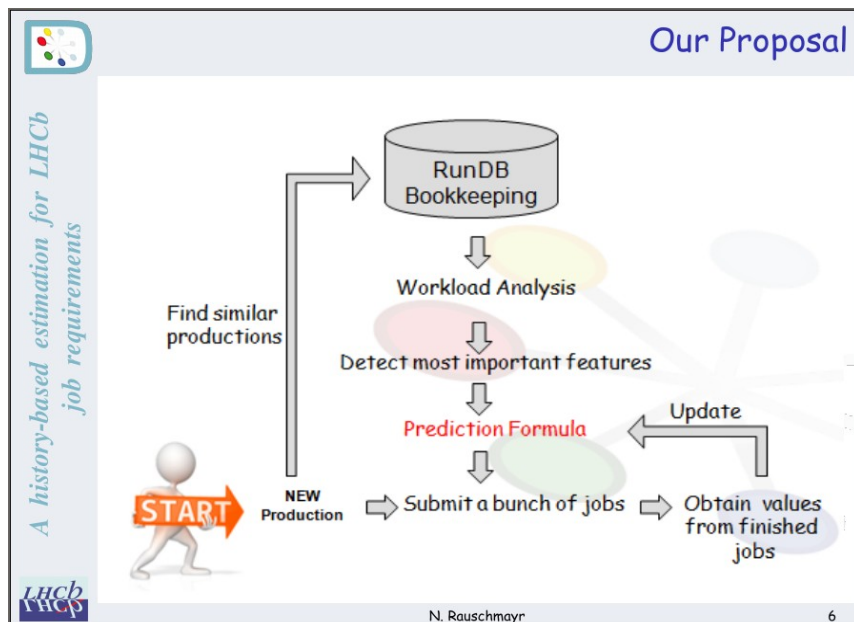
#183





# Job Management / Pilots

- LHCb showed how to predict required resources
- ALICE studied how to increase security on the Grid



#96

Project main goals

Improve computer security in the GRID by:

- > Intrusion prevention
- > Security by isolation
- > Intrusion detection
- > Analysis of Job behavior
- > Machine learning

Andrés Gómez - Frankfurt University, IRI - CHEP 2015, Okinawa Japan  
Slide 4

ALICE  
A JOURNEY OF DISCOVERY

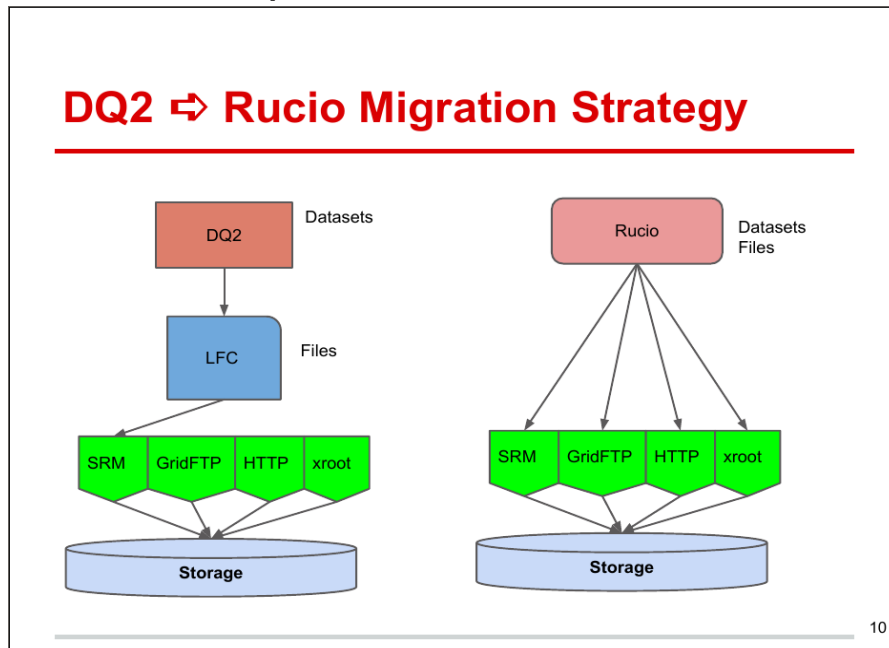
#14

# Middleware

## Data Management

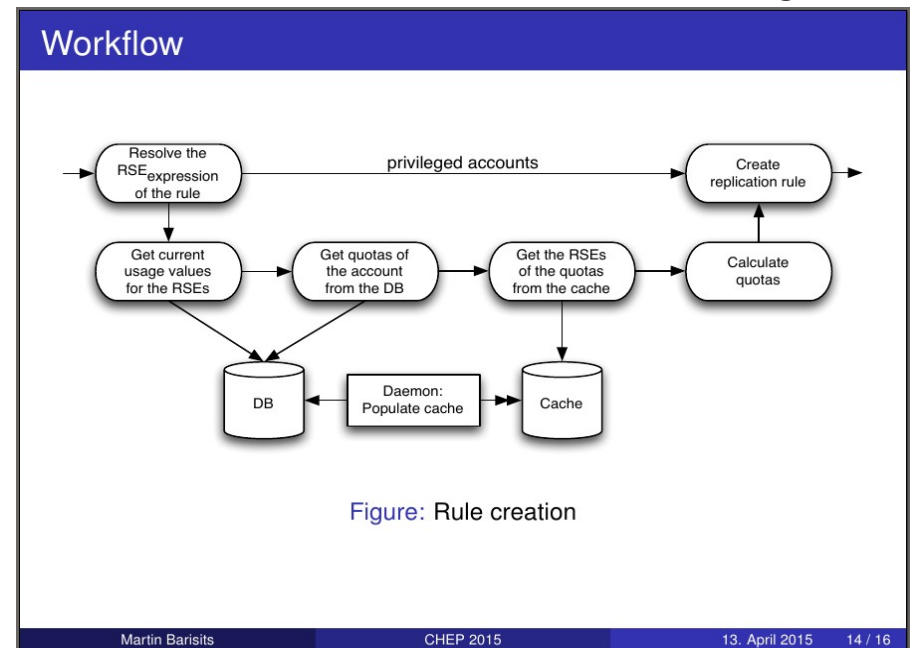
- ATLAS implemented Rucio a new Distributed Data Management tool

### Replacement for DQ2



#205

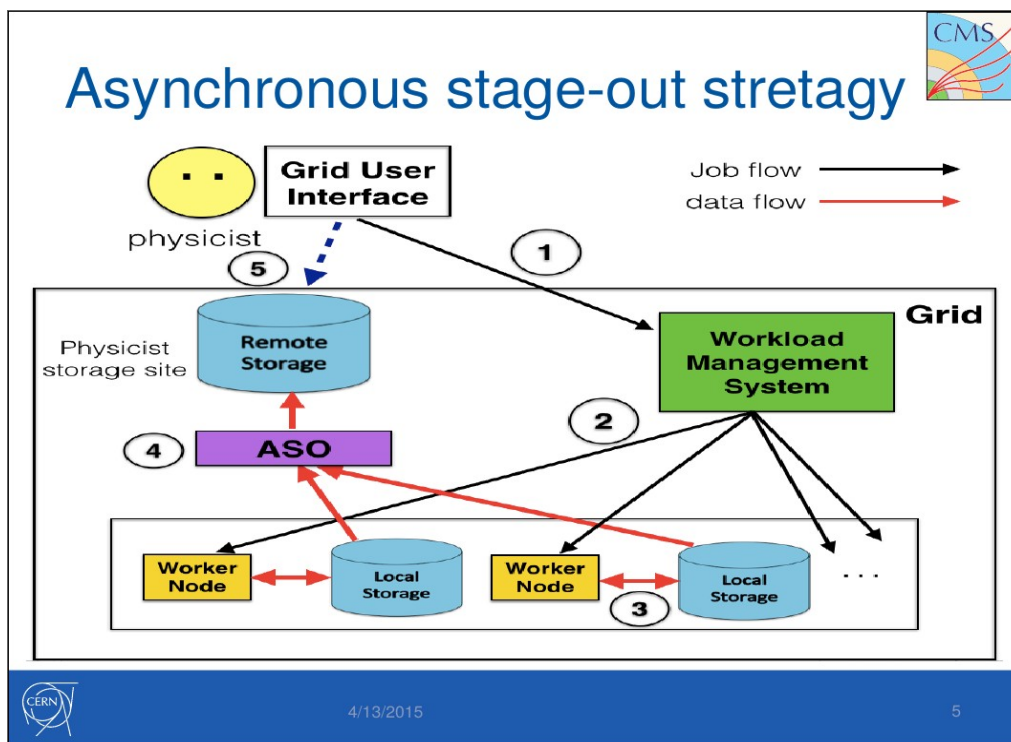
### Flexible Quotas and Accounting



#207

# Data Management

- CMS implemented Asynchronous stage-out to avoid that jobs fail during data transfer



4/13/2015

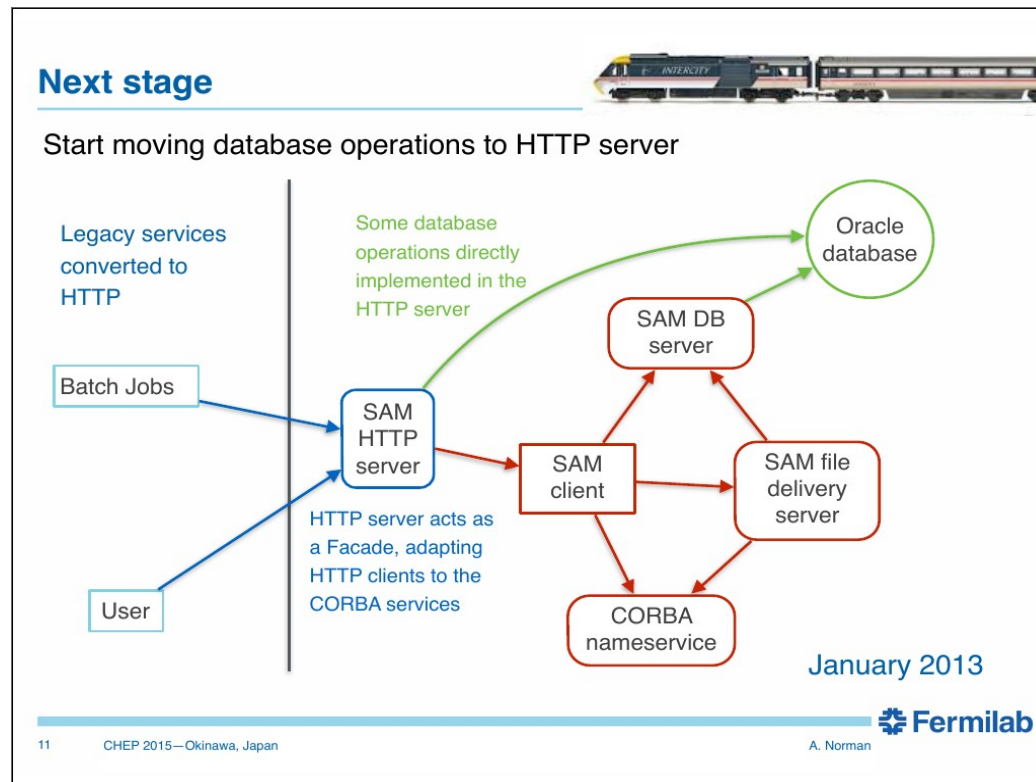
5

#225

# Middleware

## Data Management

- Fermilab re-engineered SAM with new interface while maintaining operations



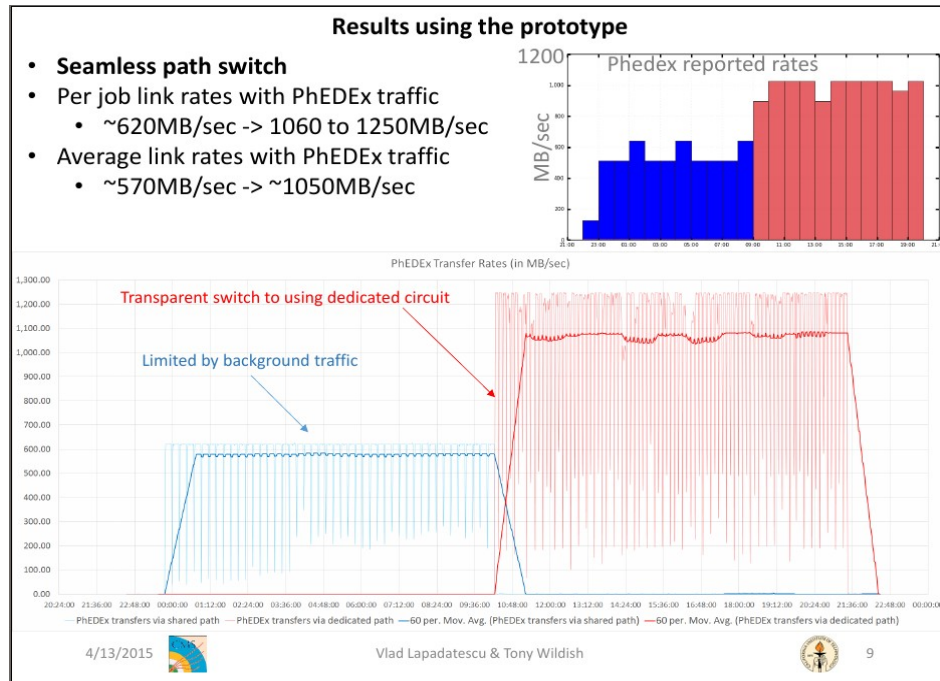
#463

# Middleware

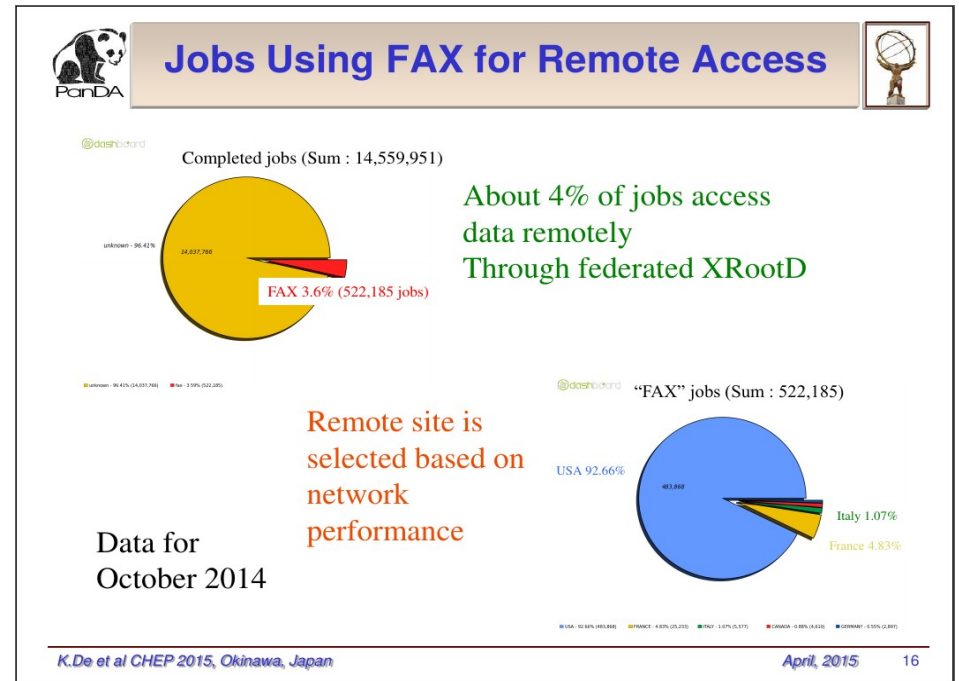
# Network Awareness

- CMS and ATLAS showed uses of Network Awareness

## CMS improves data transfers



## ATLAS improves job submission



#191

#237

# Middleware Multicore

- Report from WLCG Multicore Task Force
  - ATLAS and CMS cases

MANCHESTER  
1824

## Experiments submission

- CMS move the scheduling within the pilot
  - Predictability
  - Shared sites still have single core to handle
- ATLAS: mcore and score in parallel with 1 payload per pilot and let the scheduler do the job.
  - Entropy
  - Predictability still helps
    - Backfilling not an option yet

The diagram illustrates scheduling strategies for ATLAS and CMS. On the left, 'Inside a scheduler' shows a CPU timeline with a 'scheduler/collector gap' and a 'V02 job' block. On the right, 'ATLAS' shows a 'Multi Core Payload' and 'Single Core Payload' blocks. Below it, 'CMS' shows a 'Multi Core Payload' and 'Single Core Payload' blocks. Arrows indicate the flow of information or data between the scheduler and the pilot.


5

- Good progress
- It works already
  - Fine tuning needed

#225

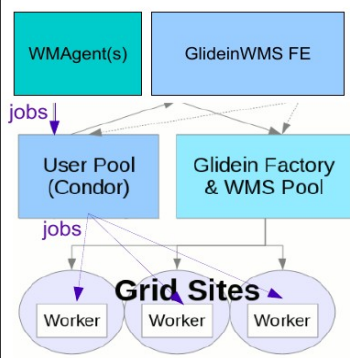
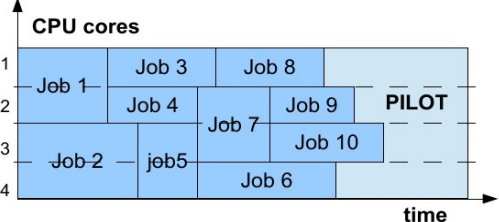
# Middleware Multicore

- CMS reported on their successful use of multicore jobs on the Grid
- Ready for RUN2



## CMS WM and SI

- CMS workload management and submission infrastructure is based on:
  - **WMAgents**: manage centralized workflows populating job queues, assigning job priorities, handling errors and job retrials, etc.
  - **GlideinWMS**: matches jobs to resources managing a transient pool of computing resources controlled by **pilot jobs**
- **Main tool: multicore pilots** with internal dynamic partitioning of resources



See talk "Using the glideinWMS System as a Common Resource Provisioning Layer in CMS" for more details  
4-13-2015 Evolution of CMS WM for multicore support - Antonio Pérez-Calero Yzquierdo 4

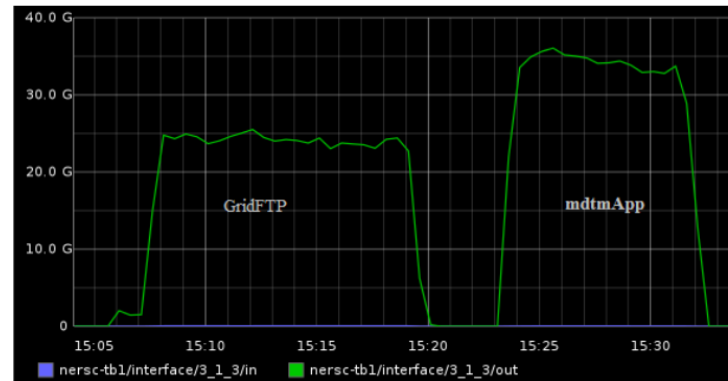
#409

# Middleware Multicore

- Interesting report of the importance of multicore awareness in data transfers

## MDTM Early Test Evaluation

- Wide-area network links, end-to-end tests
- Performance comparison with GridFTP, bandwidth captured at ESnet's edge router



Parallel large file transfers(16 streams, 2TB, 8MB blocks), from SSDs to /dev/null, with 40Gbps links and 50Gbps aggregate disk bandwidth

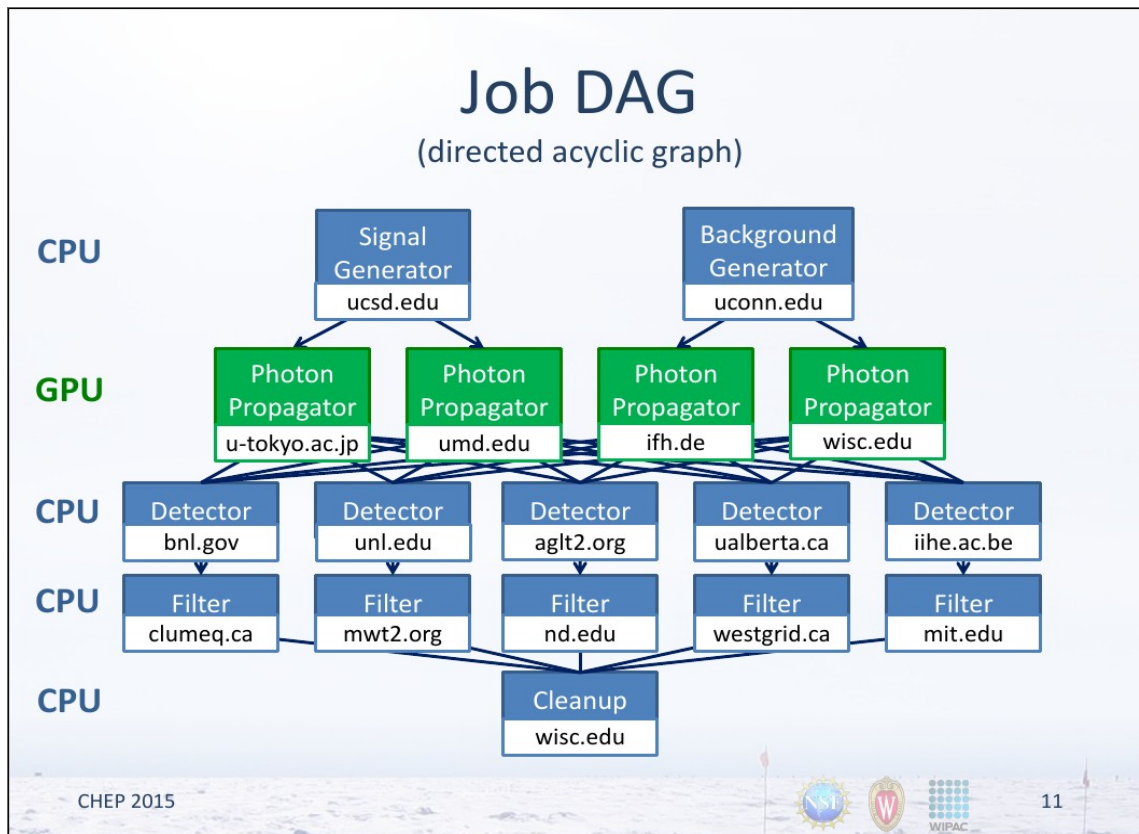


# Frameworks

- Contributions on
  - Experiment frameworks
  - Monitoring frameworks
  - Validation frameworks
  - Analysis frameworks

# Experiment Frameworks

- IceCube presented their new IceProd2



- Complete rewrite on
  - Python
  - SQLite
  - CVMFS
  - Web API
- Pilot jobs
- User permissions

#496

# Experiment Frameworks

- Reports on ROOT 6 and beyond
  - Impossible to summarize all the changes, see #441
  - Impressive work on optimization and validation

## Conclusion

- **ROOT** Modernization underway
  - Starting to add **new** API that will overtime replace then deprecated historical API
  - Making writing [physics [analysis]] code even simpler, more intuitive and more robust
- Main Driving Principles
  - Simplicity
  - Robustness
  - Performance
    - Embrace multi-tasking and vectorization
  - Provide even better features
  - Continue our many collaborations (e.g. **Python, R, I/O**)

Philippe CANAL  
root.cern.ch

CHEP 2015 - Okinawa  
2015

13 April

19

## Start up Time

- **Very first feature seen by the user**
  - Baseline: ROOT5, ~100 ms (Python 2.7 ~20 ms)

Solution:

- Leverage PCH to store I/O information of ROOT most used classes (Hist, RooFit, ...)
- Optimise data structures and algorithms holding/manipulating autoloading info: e.g. use STL!
- Optimise reading of ROOTmap files

Version	time [ms]
6.00.00	275
6.02.06	240
6.03.03	200

Strive for technical excellence in all corners

16/4/2015

CHEP2015, Okinawa - Track 4

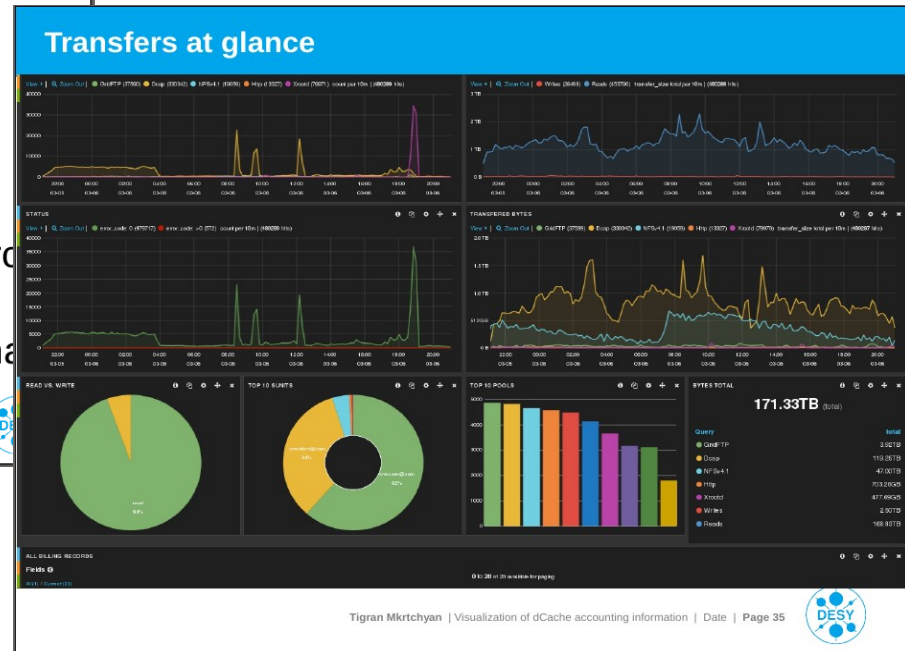
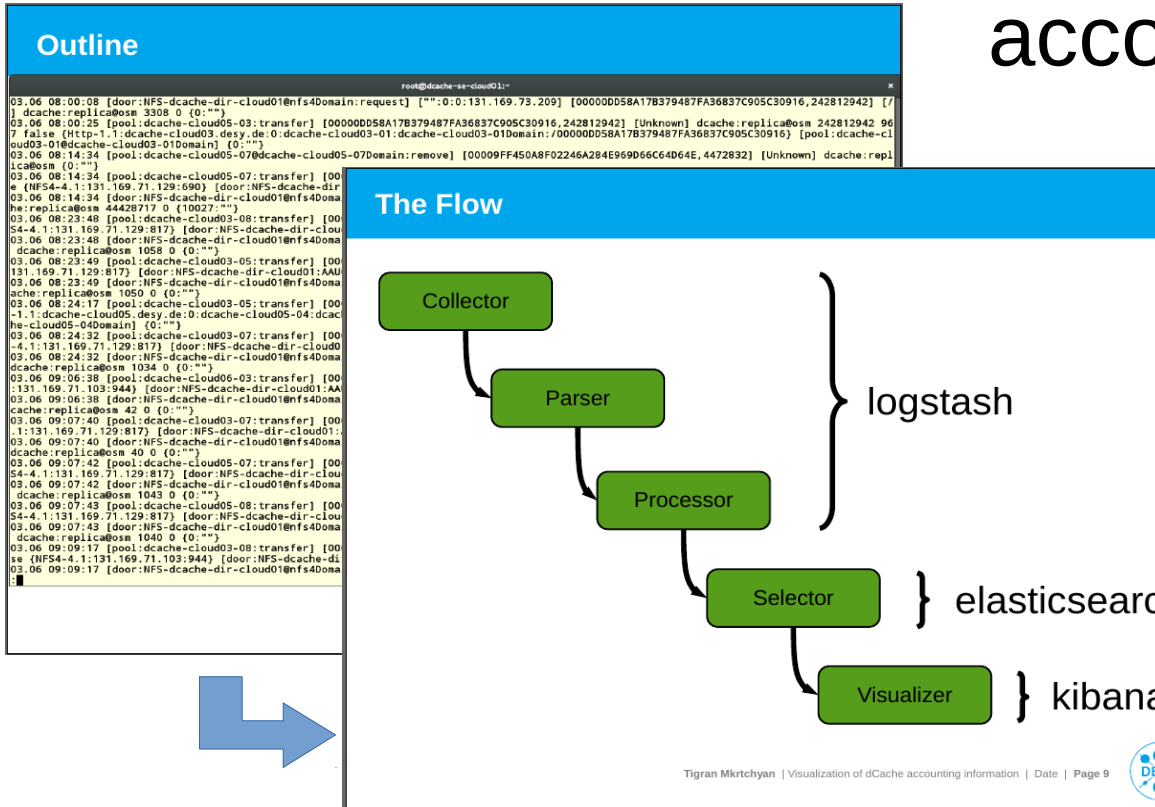
16

#441

#381

# Monitoring Frameworks

- From DESY we saw how to display dCache accounting informations

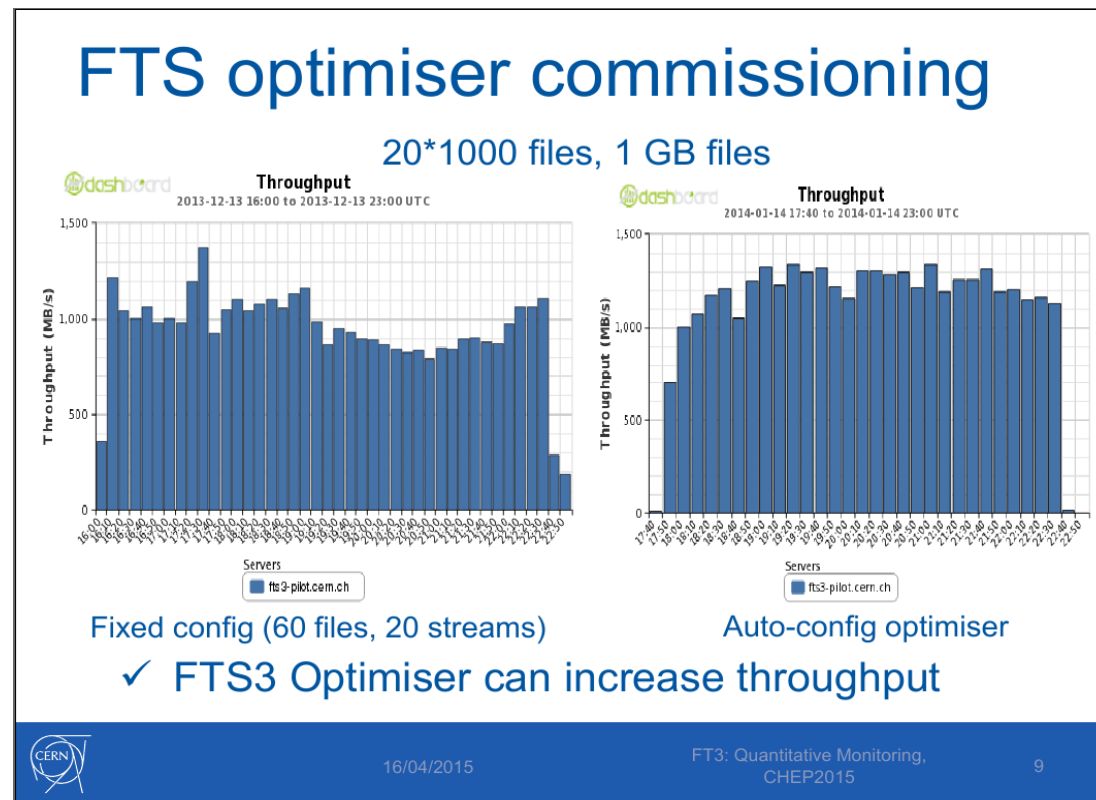


#45



# Monitoring Frameworks

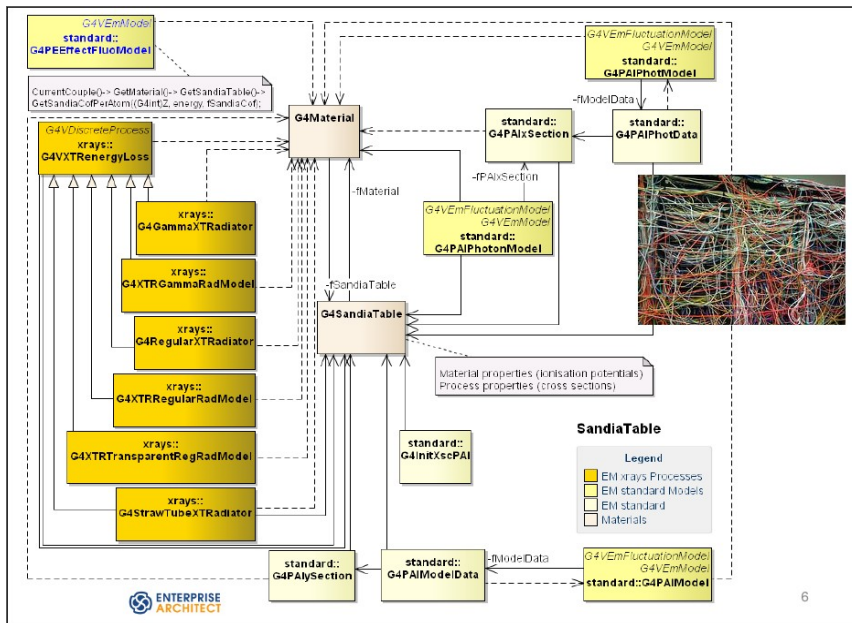
- The quantitative monitoring of FTS3 has been crucial for commissioning and production



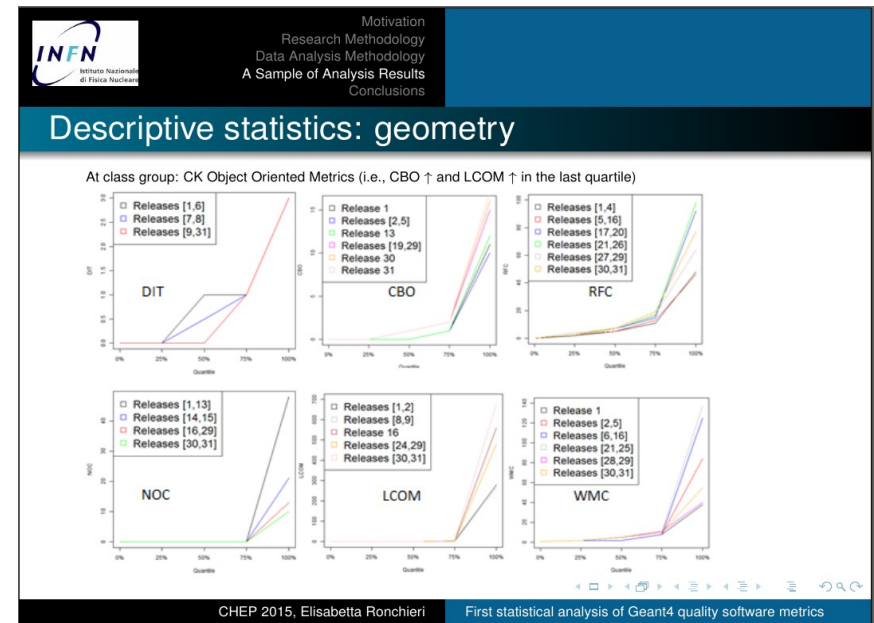
#232

# Validation Frameworks

- Geant4 has been used as a test bench to study
  - Testability
  - Statistical analysis of software quality



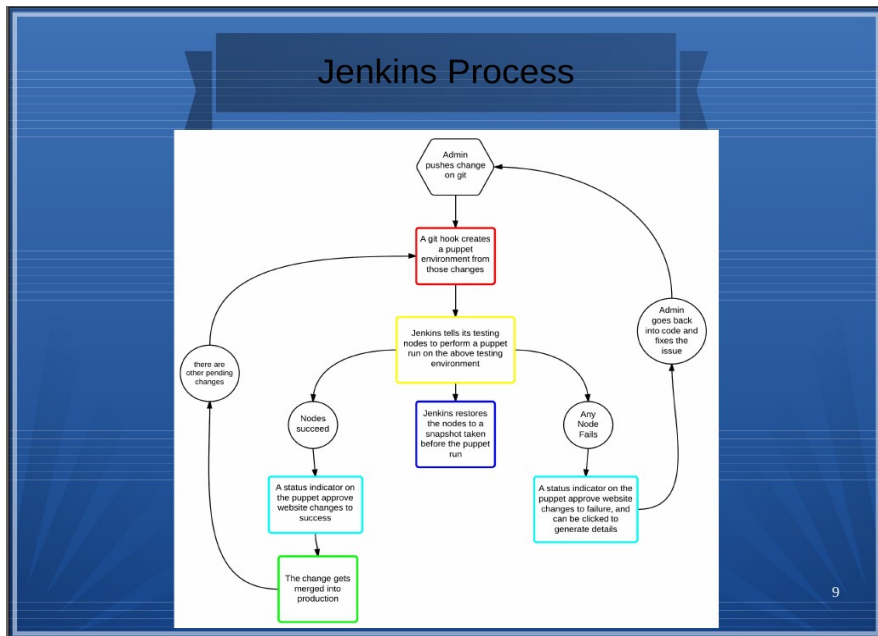
#348



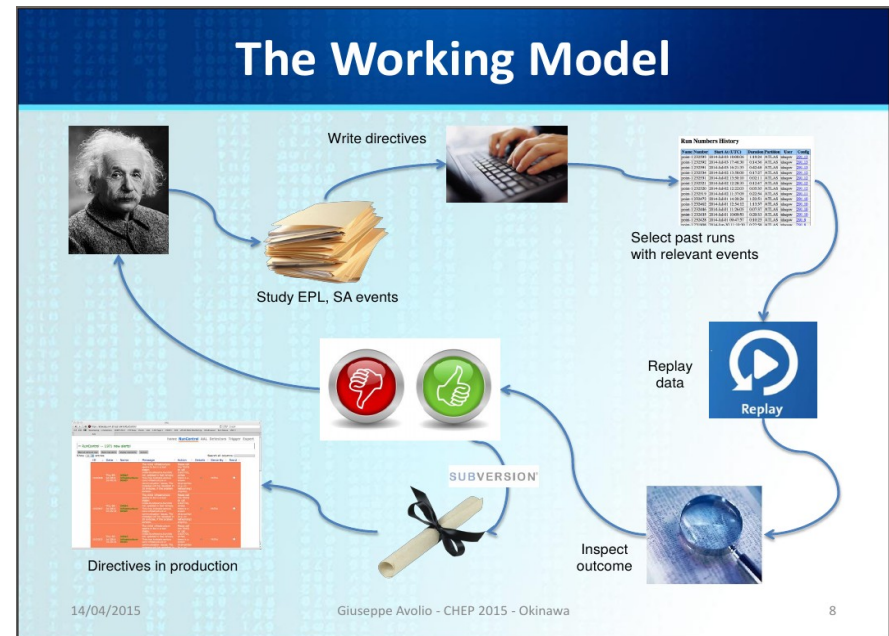
#485

# Validation Frameworks

- Validate Puppet configuration in Jenkins-CI



- Validate ATLAS Shifter Assistant directives

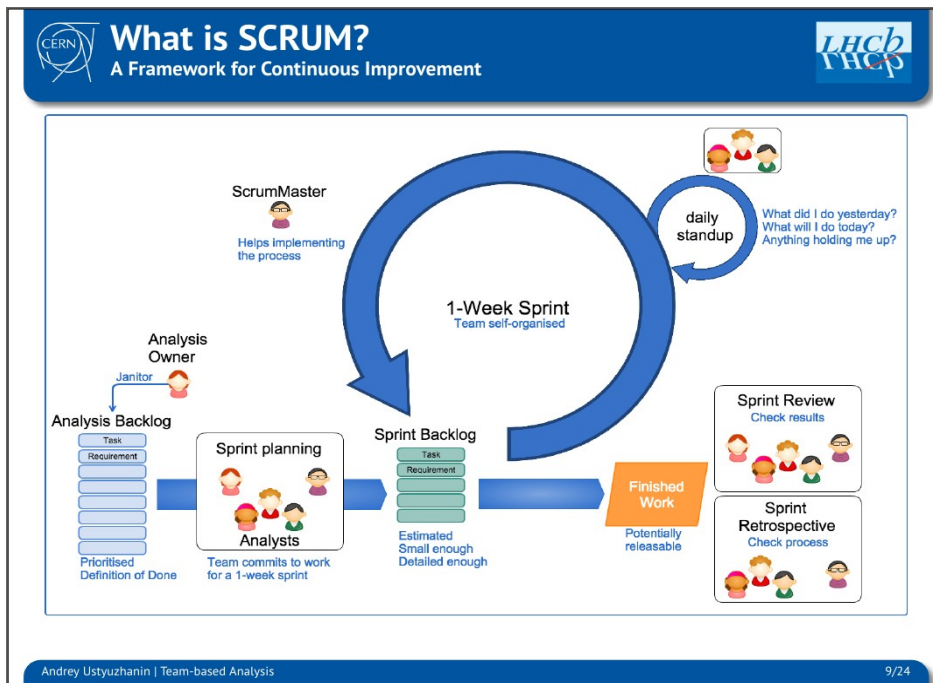


#28

#36

# Analysis Frameworks

- LHCb showed that it is possible to apply the SCRUM agile methodology to physics analysis



**A Virtual Pinboard**  
Visual organisation of tasks with Trello

The screenshot shows a Trello board titled "SumoProject: First Observations at LHCb". It features a **Backlog** of tasks, a **Sprint Backlog** with tasks like "check data / MC agreement" and "compute PID efficiency", and a **Checklist** with items like "missed beam pipe peak in fit" and "fix ratio of signal/signal". The board also displays **Activity** logs, **Members**, and **Attachments** including plots and documents.

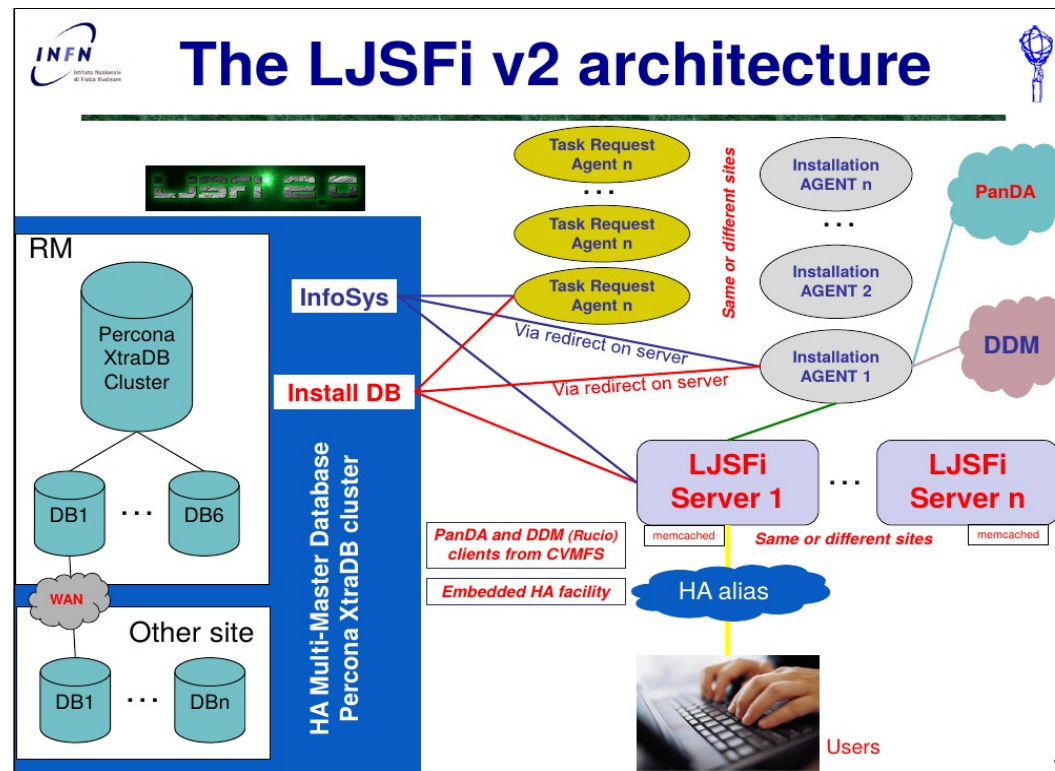
Andrey Ustyuzhanin | Team-based Analysis 11/24



# Tools

# Tools

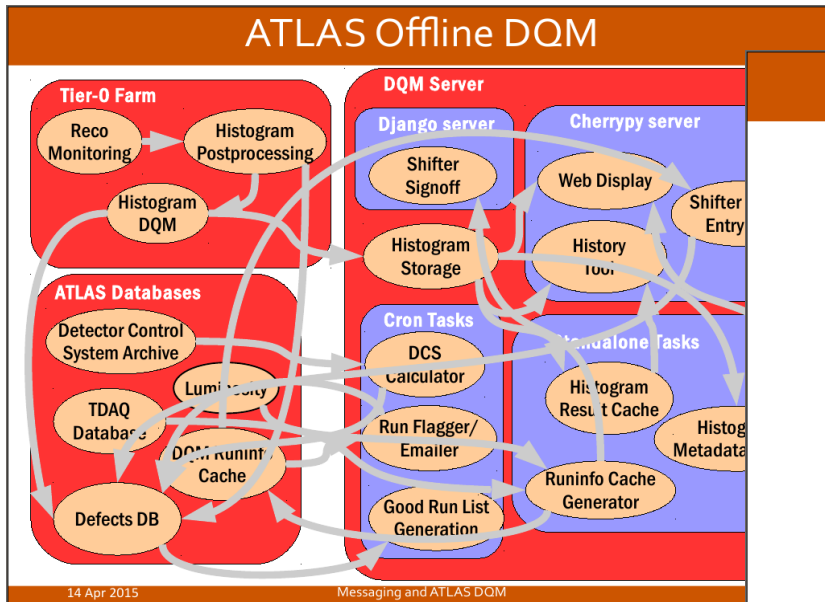
- ATLAS presented the new incarnation of their Software Installation System



#204

# Tools

- ATLAS decided to leverage on standard Message Queue technologies to synchronize Data Quality Monitoring tasks



#176

## Messaging Queues at CERN

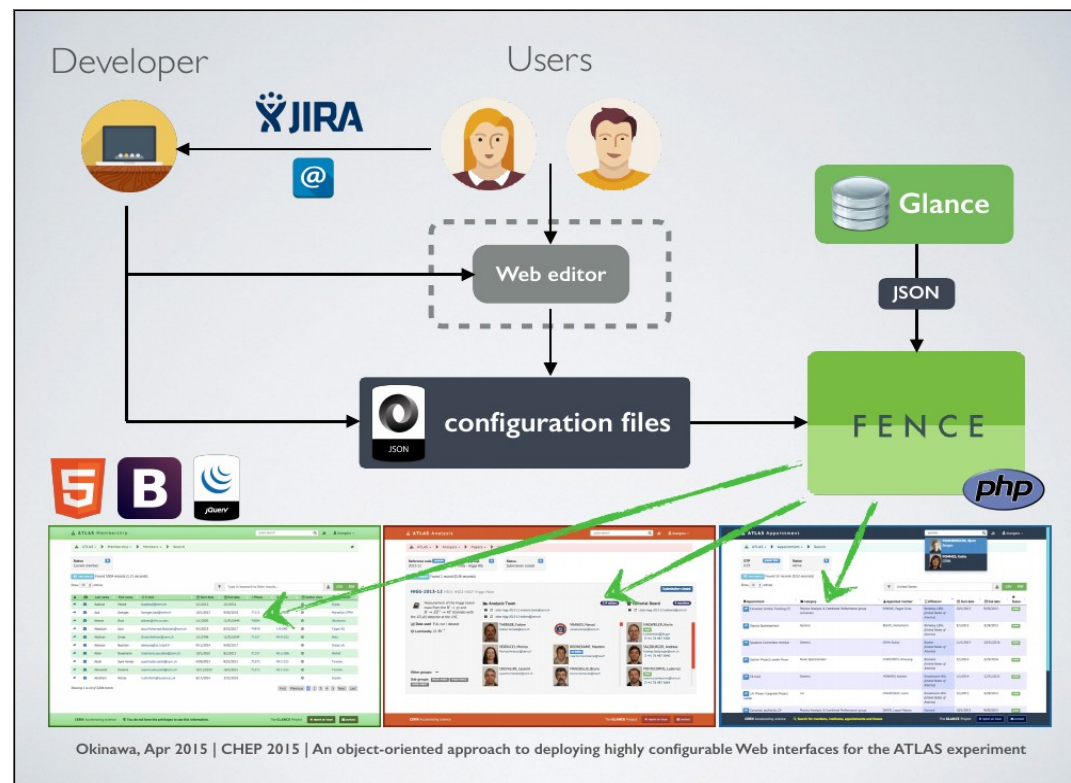
- Many options for messaging brokers
- CERN has standardized on **ActiveMQ**
  - Wrinkle: doesn't integrate at all with standard CERN auth mechanisms (must use app-specific passwords or certificates)
  - for security reasons, creating queues requires coordination with CERN IT
- Use **STOMP** protocol
  - near-universal availability of client libraries for different languages
- Piggybacking on servers set up for ATLAS Event Server project
  - message rate of few/minute is negligible perturbation

ØMQ in bad location on complexity/benefit curve  
Would need to support RabbitMQ ourselves

14 Apr 2015      Messaging and ATLAS DQM      9

# Tools

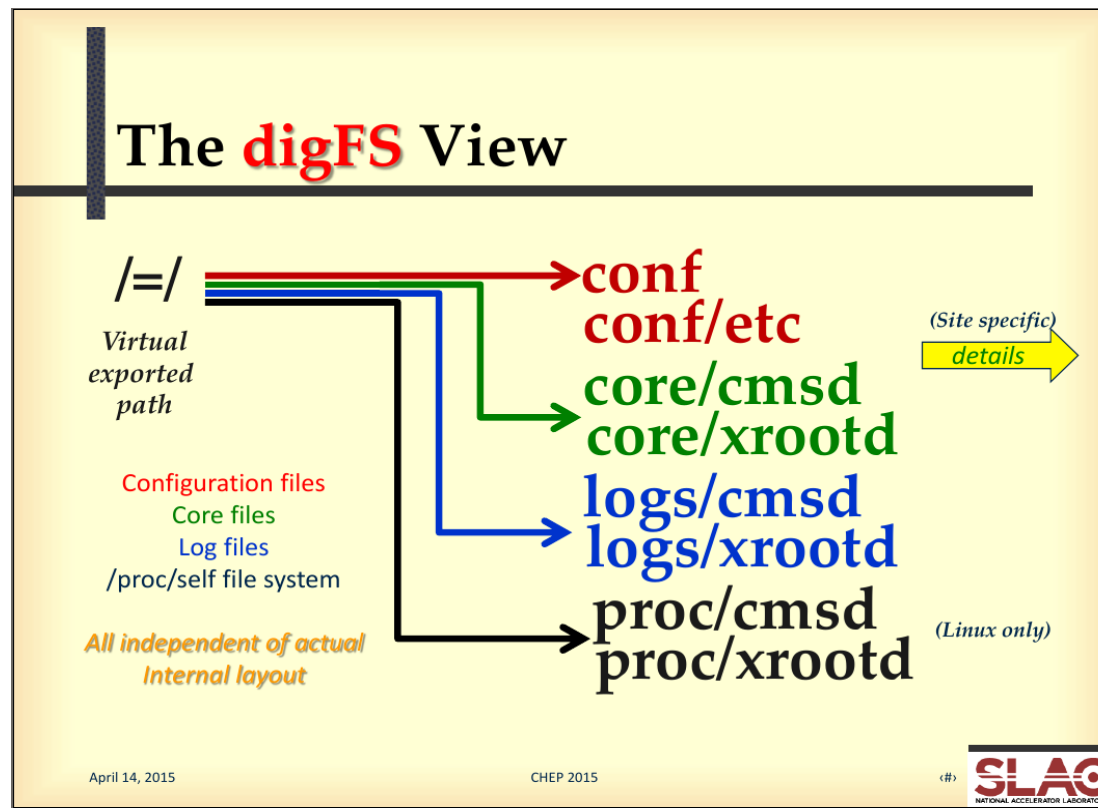
- From ATLAS we saw an interesting new way of developing web interfaces



#167

# Tools

- From SLAC a great contribution for debugging Grid jobs



#310

# Tools

- Deep insight on the features of IgProf
  - a feature rich profiler for HEP
  - including power monitoring for energy efficient code



...any reference to real facts or persons is purely coincidental...

```
...  
std::vector<int> foo;  
for (int i = 0; i < 1000000; ++i)  
    foo.push_back(0); // unneeded memory churn!  
...
```

3

#478

## IgProf web

igprof\_pp\_25202.0\_step3 - x86\_64, igprof-navigator

[Back to profiles index](#)

Counter: PERF\_TICKS, first 1000 entries

Sorted by cumulative cost

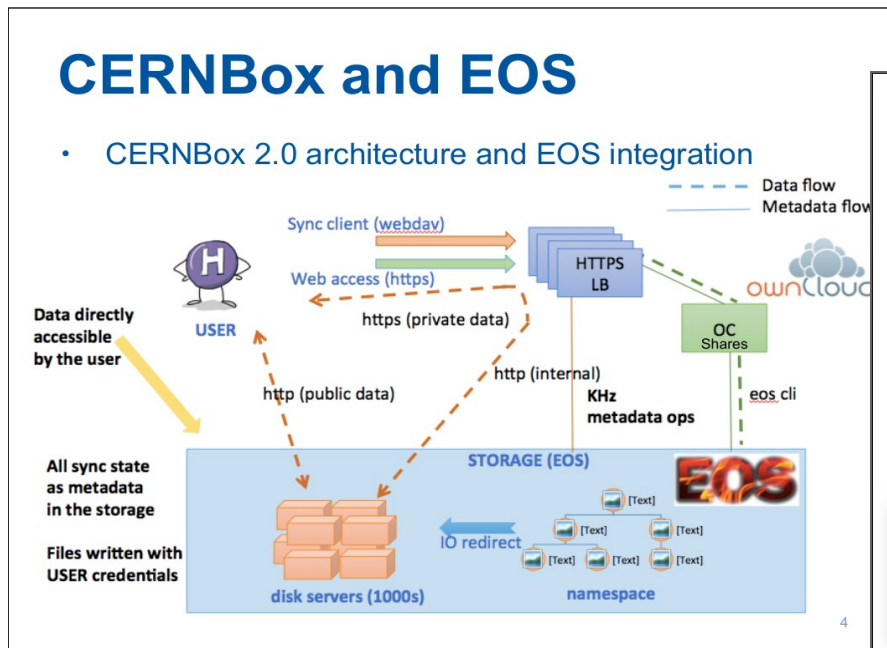
(Sort by self cost)

Rank	Total %	Cumulative	Symbol name
1	100.00	2,130.54	<spontaneous>
3	97.40	2,075.14	__libc_start_main
2	97.40	2,075.14	_start
4	97.36	2,074.30	main
5	97.31	2,073.21	main::lambda(#1)::operator()() const
6	94.11	2,005.03	edm::EventProcessor::runToCompletion()
7	94.11	2,005.02	boost::statechart::state_machine<state_machine::Machine, state_machine::Starting, str
12	92.60	1,972.82	edm::EventProcessor::readAndProcessEvent()
11	92.60	1,972.82	state_machine::HandleEvent::readAndProcessEvent()
10	92.60	1,972.82	state_machine::HandleEvent::HandleEvent(boost::statechart::state_machine::Handl
9	92.60	1,972.82	boost::statechart::state_machine::HandleEvent::state_machine::HandleEvent::boos
8	92.60	1,972.82	boost::statechart::simple_state_machine::FirstLumi_state_machine::HandleLumis
15	92.60	1,972.81	edm::EventProcessor::processEventsForStreamAsync(unsigned int, std::atomic<bool>*)
14	92.60	1,972.81	edm::StreamProcessingTask::execute()
13	92.60	1,972.81	tbb::internal::custom_scheduler<tbb::internal::intel_scheduler_traits>::local_wait_fc
16	92.59	1,972.62	edm::EventProcessor::processEvent(unsigned int)

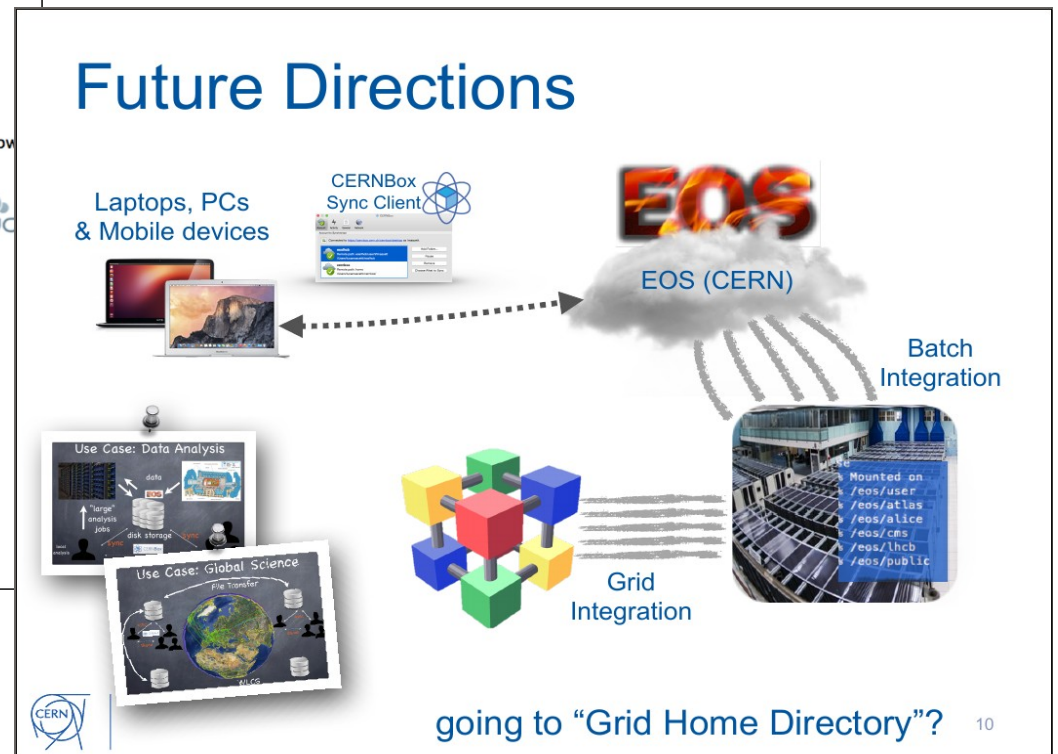
14

# Tools

- CERN developed an EOS-based Dropbox alternative: CERNBox



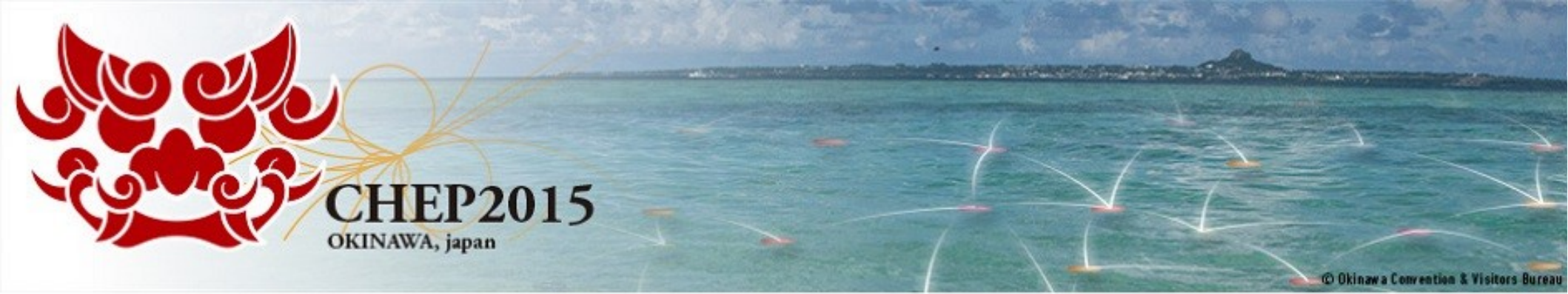
#327



# Conclusions

- We are working for a better (computing) world
  - Continuous efforts towards improvements
  - Sometimes “rewrite” is good
- Common solutions are beneficial to many
- Many are beneficial to common solutions





© Okinawa Convention & Visitors Bureau

21st International Conference on Computing in High Energy and Nuclear Physics **CHEP2015** Okinawa Japan: April 13 - 17, 2015

I want to thank the organizers  
for the great work they have done  
to make CHEP 2015 a success

Thank you!