

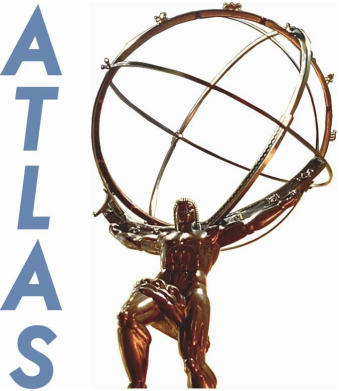
Distributed Analysis in ATLAS

A. Dewhurst (RAL), F. Legger (LMU)

on behalf of the ATLAS collaboration

April 13th, 2015

21th International Conference on Computing in High
Energy and Nuclear Physics (CHEP), Okinawa

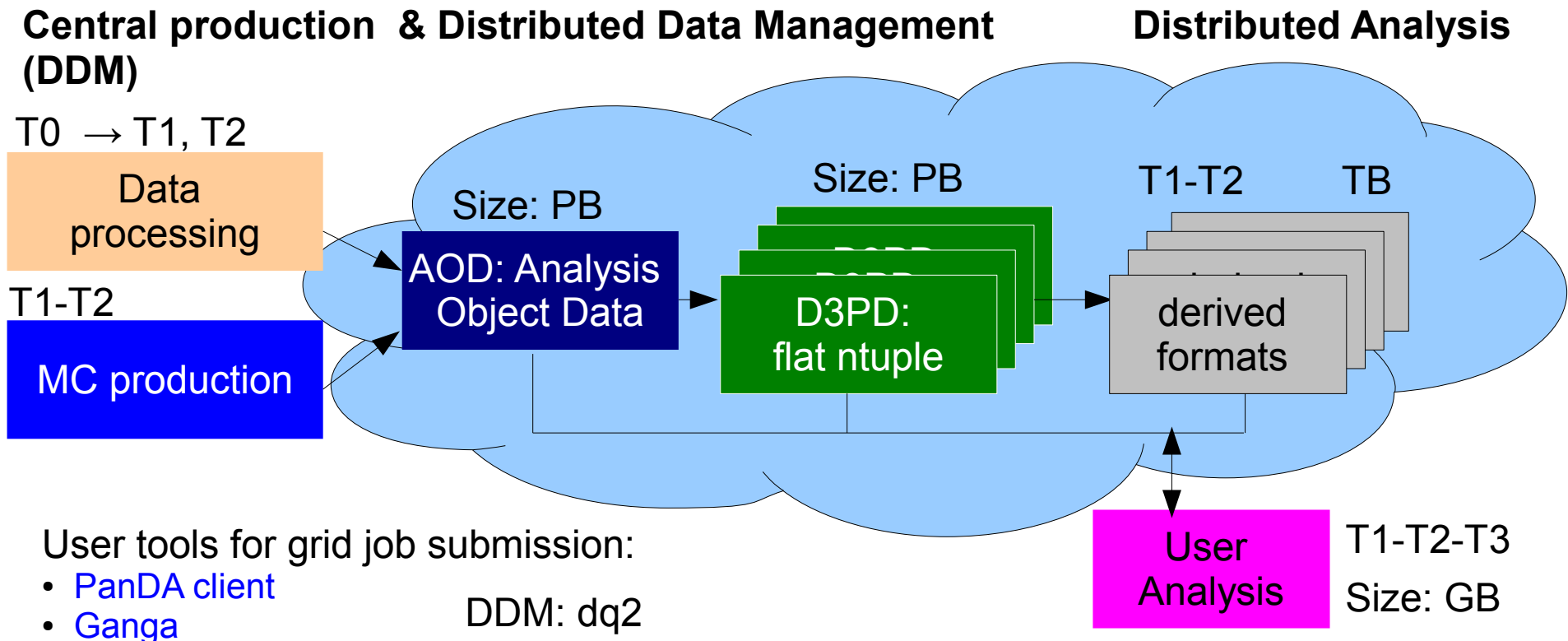


Outline

- Distributed Analysis in ATLAS for **LHC run 2**:
 - Infrastructure:
 - Many **new** developments in event processing and distributed computing:
 - Data analysis model,
 - Derivation framework, **PB** → **TB**
 - Data Format for analysis: **xAOD**
 - Common Analysis Framework
 - Job workload management: **PanDA DEFT/JEDI**
 - Data management system, **Rucio**
 - Already introduced for run 1, improved and consolidated for run 2:
 - Automated testing framework: **HammerCloud**
 - 24x7 user support: **DAST**
 - Putting it all together: performances
 - **Statistics, efficiency, failures**

Data analysis model – run 1

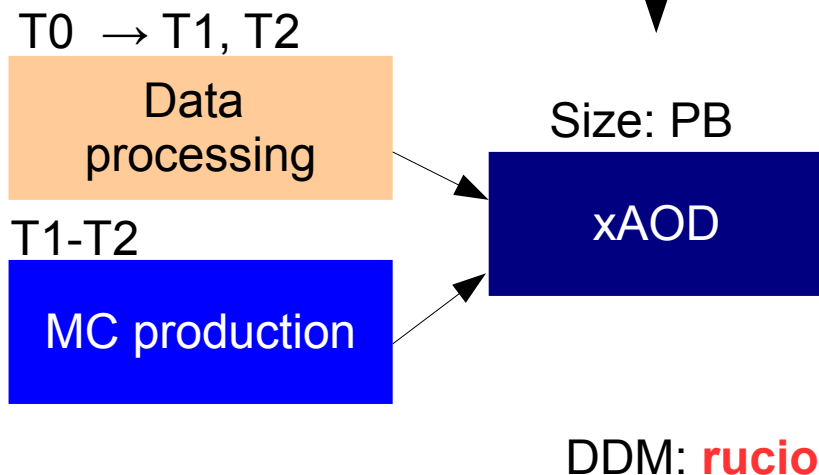
- Run 1 experience:
 - Physicists used many data formats, from class based AODs (Analysis Object Data) to flat ROOT ntuples (D3PDs) → **duplication of data, user jobs reading in average only 10% of input files, running less than 1 hour, producing many output files**
 - No common framework for analysis → from Athena (ATLAS reconstruction software) analysis to standalone ROOT → no common way of doing common tasks, hard to instrument user jobs



Data analysis model – run 2

- During LHC shutdown, huge efforts to improve:
 - New **data format**, **xAOD**: readable in both Athena and ROOT
 - Optimize read/write access both locally and remotely

See talk 171



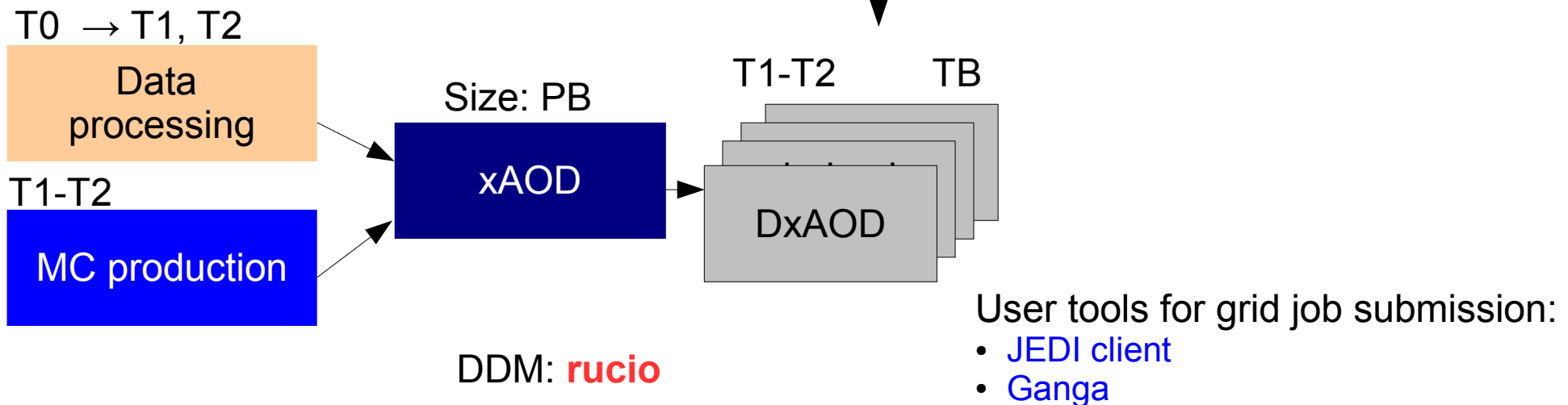
User tools for grid job submission:

- JEDI client
- Ganga

Data analysis model – run 2

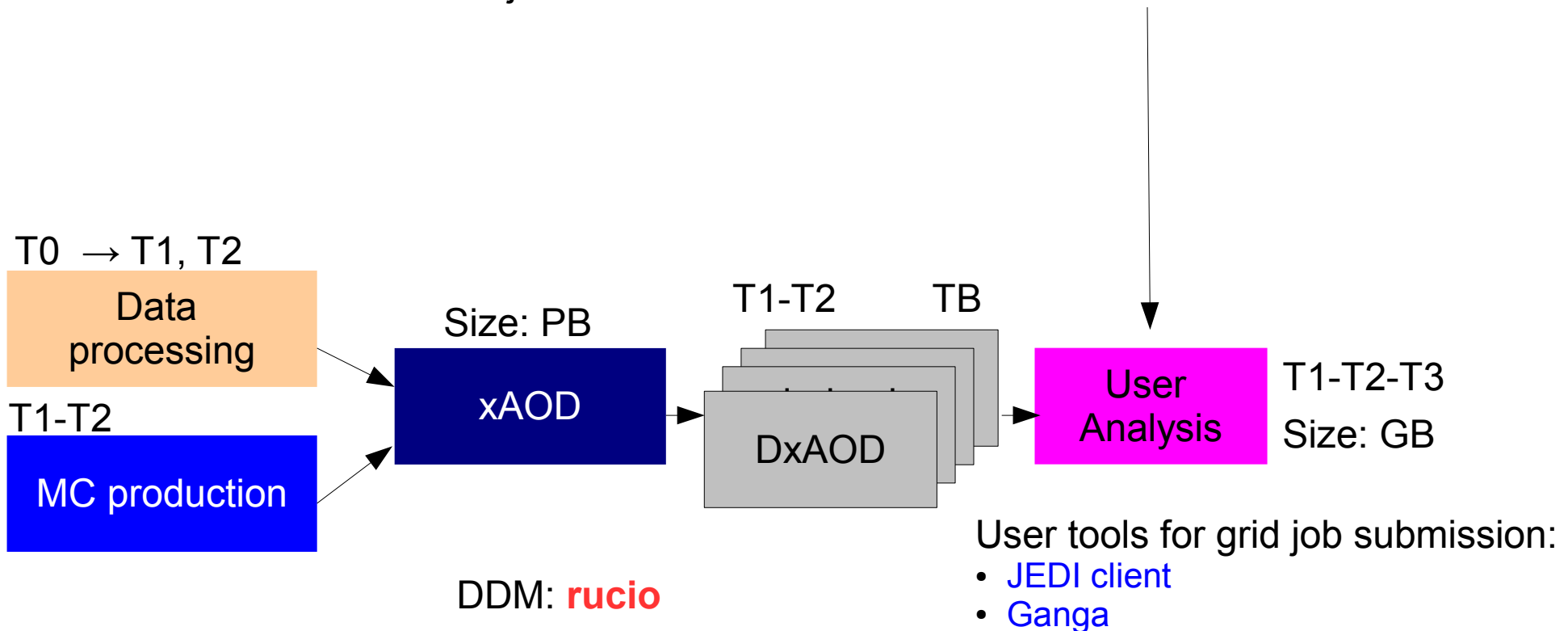
- During LHC shutdown, huge efforts to improve:
 - New data format, **xAOD**: readable in both Athena and ROOT
 - Common **derivation framework** to reduce data size from **PB to TB (train model)**
 - Centrally managed by physics groups, better integration with ATLAS distributed computing activities
 - Reduced to **1-2%** (**5-8%**) of original size on data (MC)
 - currently **~60** derivations (mostly small)

See talk 164



Data analysis model – run 2

- During LHC shutdown, huge efforts to improve:
 - New data format, **xAOD**: readable in both Athena and ROOT
 - Common derivation framework to reduce data size from **PB to TB (train model)**
 - Common **analysis framework**, “customizable” by the various physics groups
 - integration of grid clients for data management and job submission
 - monitor what user jobs do



Prodsys2, **new** production system

- Higher scalability, flexibility, user-friendliness with respect to Prodsys1
 - Components of interest to users:
 - **DEFT**, database engine for tasks:
 - Concept now based on **tasks** rather than individual jobs
 - Allows for more complex work-flows, such as chaining jobs
 - **JEDI**, job execution and definition interface:
 - Brokering and task/job management moved **server-side**
 - **Scout jobs** to estimate needed grid resources
 - If all scout jobs fail → task is stopped
 - simplification of client tools
 - **faster job submission times!**
 - better **retrial** mechanism for failed jobs
 - New job and task monitoring
- **JEDI** in use for analysis since August 14th, 2014
- Both PanDA and Ganga clients

See poster 100

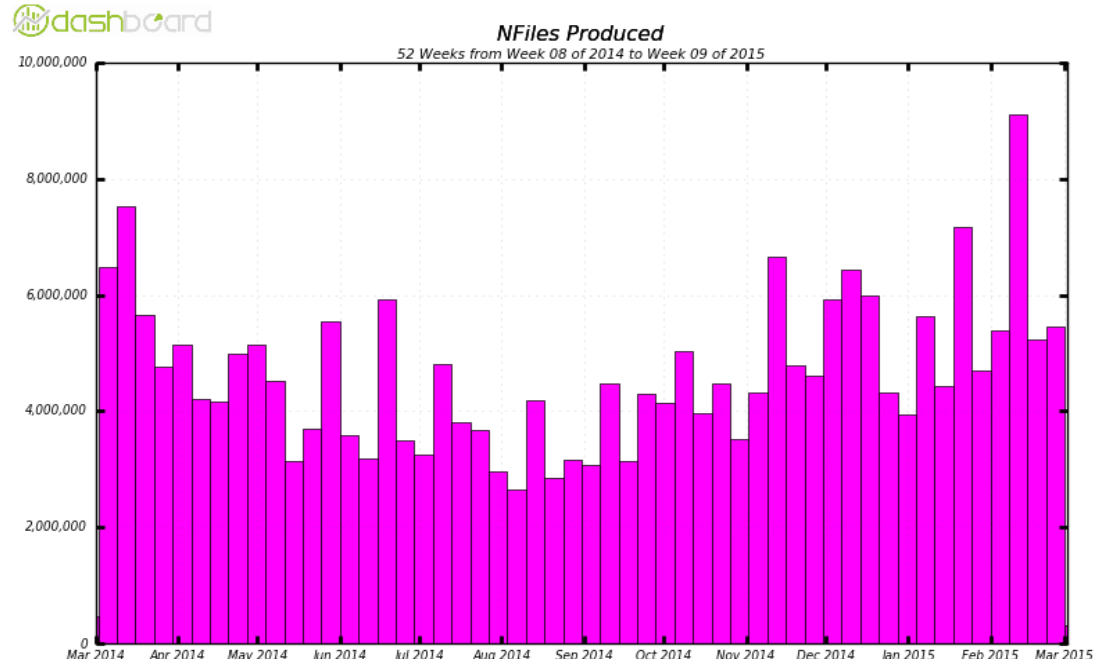
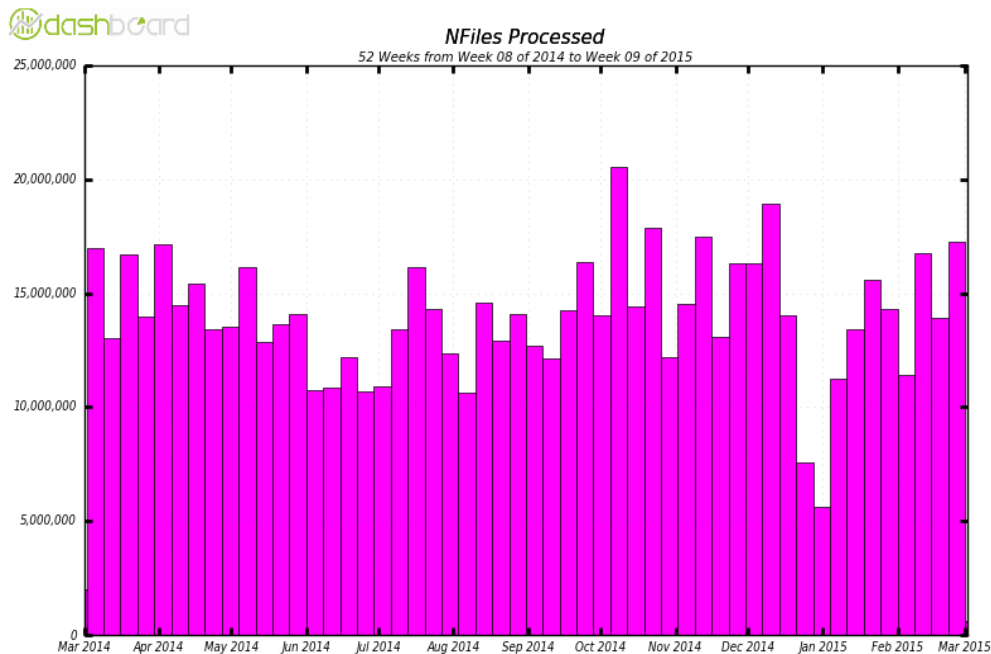


Rucio, **new** data management system

- Needed to cope with large data volume in run 2
- User/group **quotas**
- All datasets have a **lifetime**
- Transfers from disk to TAPE automatically managed
- Full integration with prodsys2
 - Users can now automatically transfer job outputs to local disk space
 - Metadata (such as number of events) can be directly queried

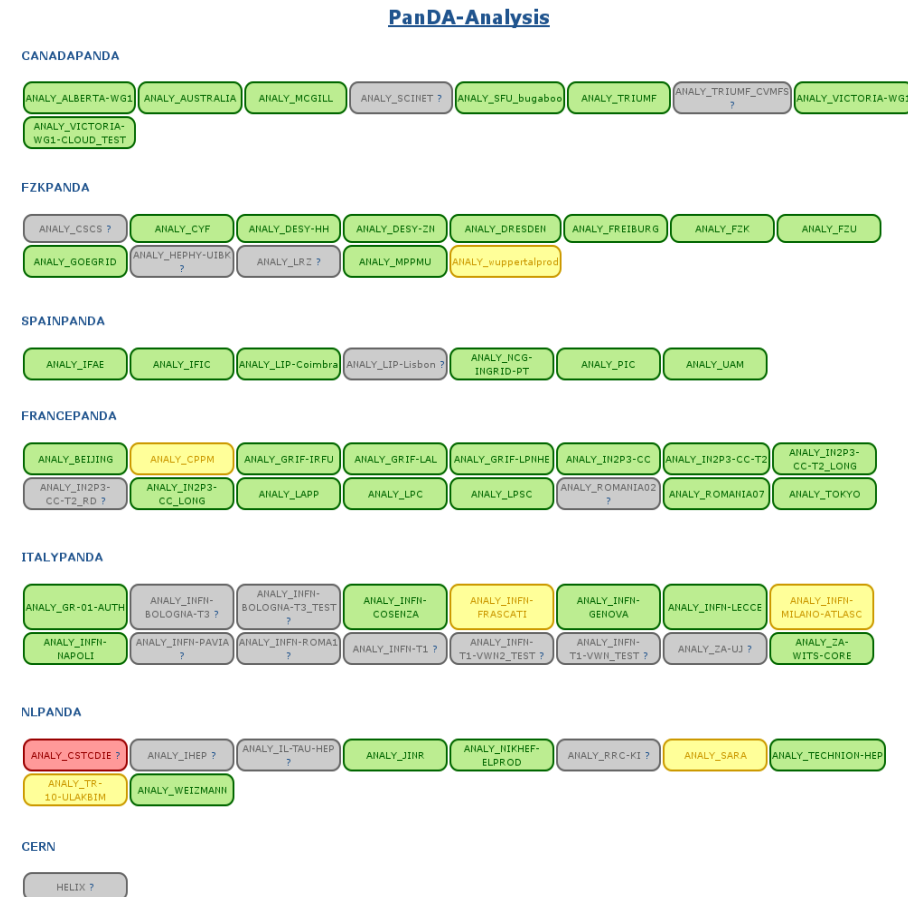
In production since December 1st, 2014

See talk 205



HammerCloud

- Tool for automatic site testing
 - both **functional** and **stress** tests
 - Used by ATLAS, CMS, LHCb
 - Crucial tool to test **new deployments** (JEDI, Rucio) before going to production/exposing changes to users
 - Used also for **R&D** of new data access technologies (Federated Xrootd Access FAX, http with Webdav/Davix)
 - Fully integrated in ATLAS Grid Information System (AGIS)
- Suite of 3 **AFTs**, Analysis Functional Tests, mimicking typical user analysis are used for **automatic exclusion** of sites failing the tests from brokering
 - Typical efficiency of analysis functional tests: **95%** → constant over time, clouds, ...

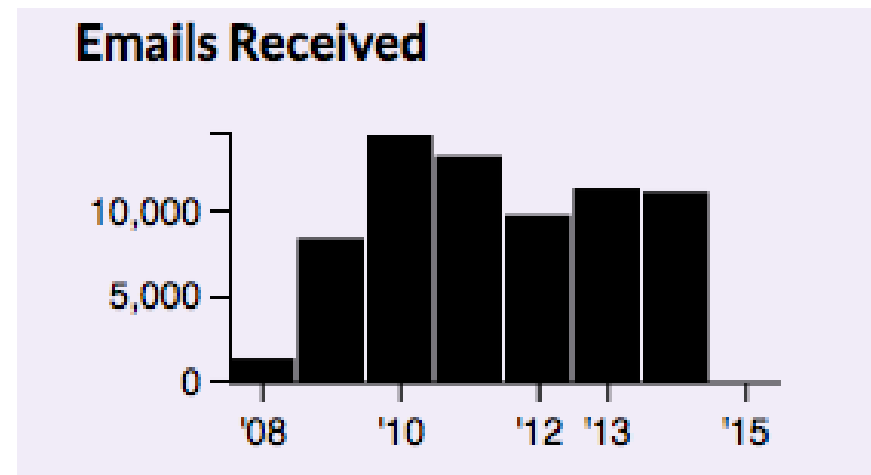


See poster 159

Distributed Analysis Support Team **DAST**

- User support with dedicated mailing list
 - expert shifters covering **16 hours/day** (American and European time-zones)
 - Critical to help users to solve grid issues fast
- Covered by DAST:
 - Rucio and Jedi clients
 - Site services/issues
 - Physics analysis tools
 - Monitoring systems

Since Oct 2008: 1,032 users; 89,478 emails exchanged (more than 10,000 a year!)



Running Grid jobs – March 2014 – 2015



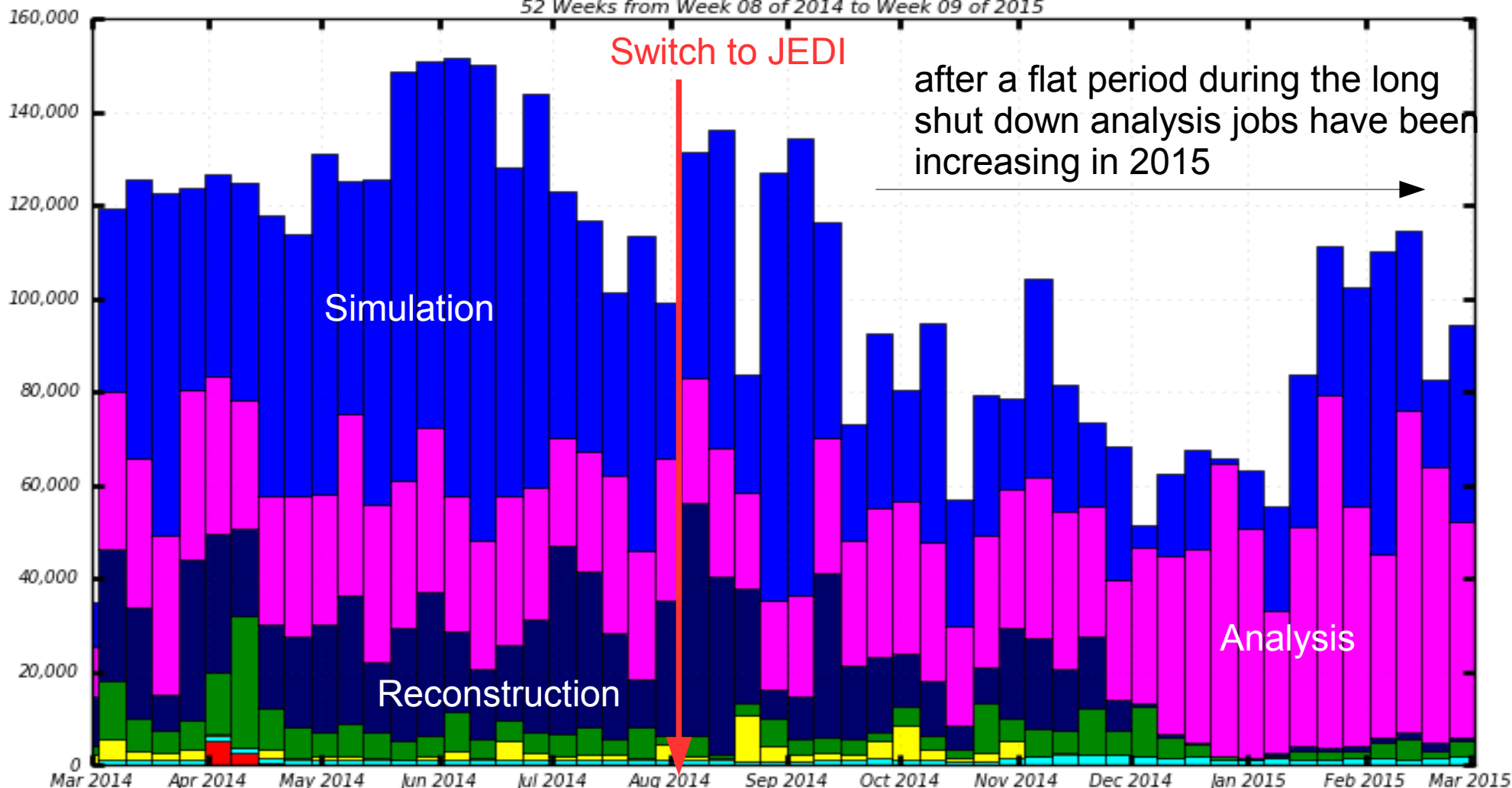
Running jobs

All T1s and T2s

52 Weeks from Week 08 of 2014 to Week 09 of 2015

Switch to JEDI

after a flat period during the long shut down analysis jobs have been increasing in 2015



MC Simulation
Others

Analysis
Extra Production

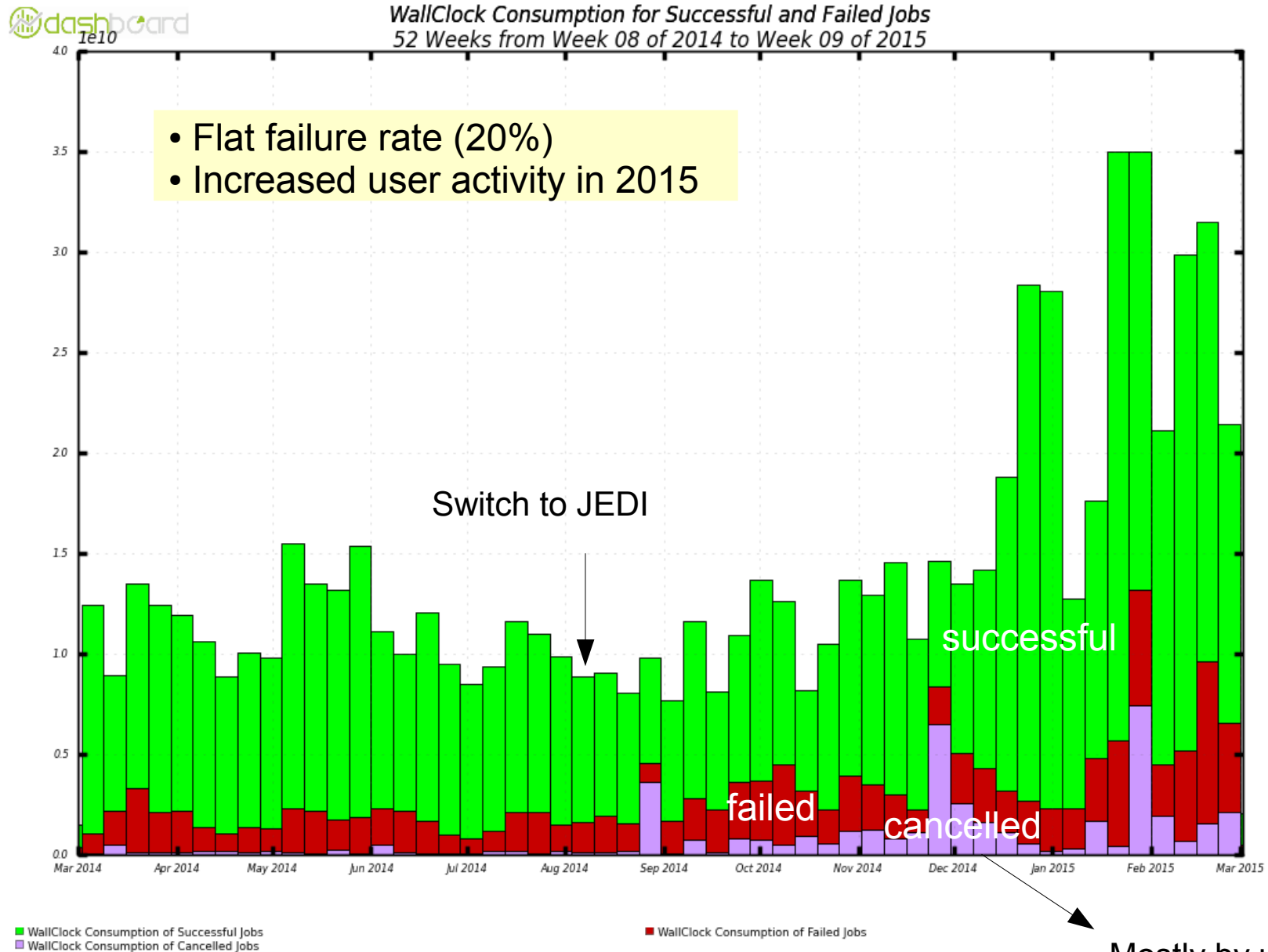
MC Reconstruction
unknown

Group Production

Data Processing

Maximum: 151,671 , Minimum: 0.00 , Average: 101,722 , Current: 94,427

Analysis jobs - March 2014-2015

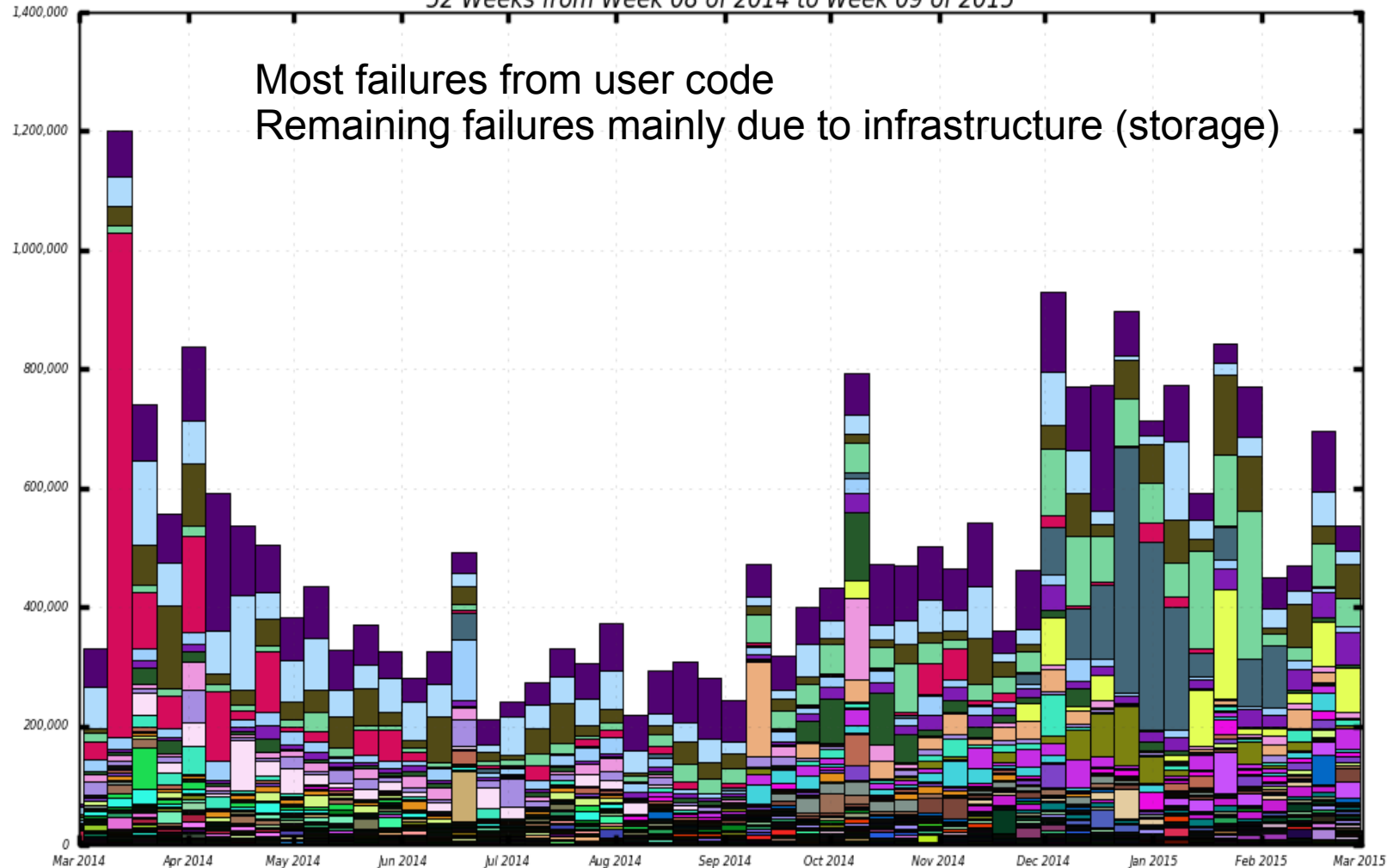


Maximum: 35,015,744,574 , Minimum: 0.00 , Average: 13,561,994,221 , Current: 2,031,532,287

Failures analysis jobs - March 2014 - 2015



Panda Failures by ExitCode
52 Weeks from Week 08 of 2014 to Week 09 of 2015



■ Unspecified error, consult log file ← Athena/ROOT crash (30%) → ■ Athena crash - consult log file Put/Get error (15%)
■ Put error: Local output file missing ■ Get error: Staging input file failed
■ Expired three days after submission ■ Get error: Replica not found
■ Put error: Error in copying the file from job workdir to k ■ lost heartbeat
■ Athena crash - consult log file__Athena ran out of memory ■ DQ2 server error
■ Get error: Failed to get LFC replica ■ New trf: Transform received signal SIGABRT
■ Transformation not installed in CE ■ Athena core dump or timeout, or conddb C ddb DB conne
■ TRF_SEGVIO - Segmentation violation ■ Payload core dump
■ Undocumented TaskBuffer Error Code : taskbuffer113 ■ New trf: Transform received signal SIGABRT; Old trf: Athena core dump or timeout, or conddb DB conne__Ather
■ Reached batch system time limit ... plus 757 more

Maximum: 1,200,036 , Minimum: 0.00 , Average: 478,705 , Current: 24,067

Conclusions

- Distributed analysis:
 - **get users results as fast as possible**
 - Ease central operations
- Many **new** components for run 2
 - Central infrastructure: [distributed data management](#) and [job submission](#)
 - Common [data format](#), [derivation framework](#) and [analysis framework](#)
- Consolidated from run 1:
 - Automatic exclusion of problematic sites from brokerage with [HammerCloud](#)
 - user support with [DAST](#)
- Ingredients are all there, system performances are stable: waiting for exciting physics from run 2!

Backup

ATLAS Grid jobs March 2014 - 2015

