

# EVOLUTION OF EXPERIMENTS COMPUTING: ATLAS

Johannes Elmsheuser

Ludwig-Maximilians-Universität München

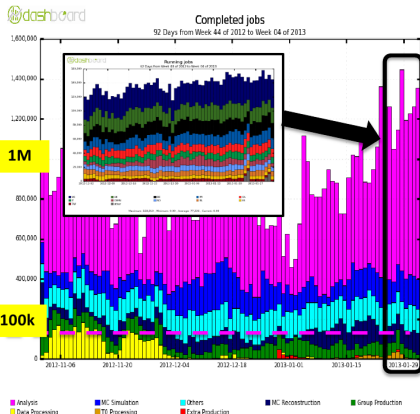
3 December 2013/7th Annual Workshop of the Helmholtz Alliance  
"Physics at the Terascale"



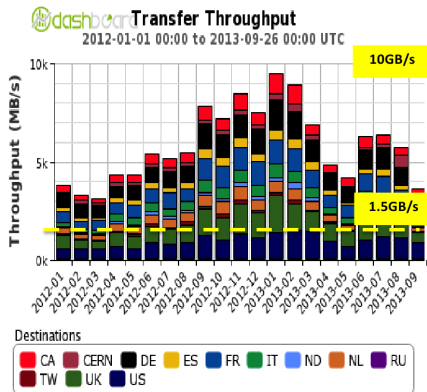
(Some content from Simone Campana's presentation at CHEP'13)

# RUN 1: WORKLOAD AND DATA MANAGEMENT

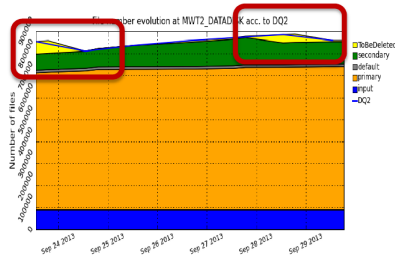
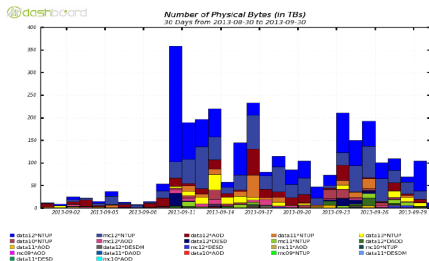
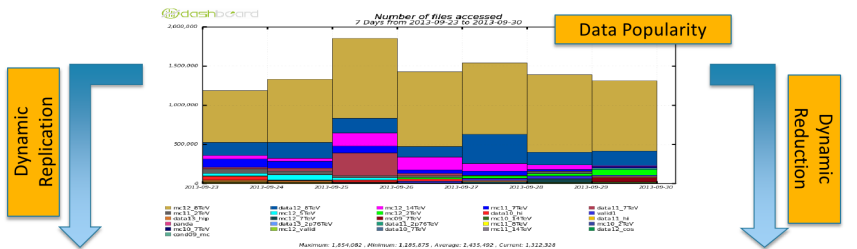
1.4M jobs/day, 150K concurrently running  
(2007 gLite WMS acceptance tests: 100K jobs/day)



Nearly 10GB/s transfer rate  
(STEP09 target: 1.5GB/s)



# RUN 1: DYNAMIC DATA REPLICATION AND REDUCTION



# CHALLENGES OF RUN 2

Trigger rate: from 550Hz to 1kHz:

- Therefore, more events to record and process

Luminosity increase: event pile-up from 25 to 40:

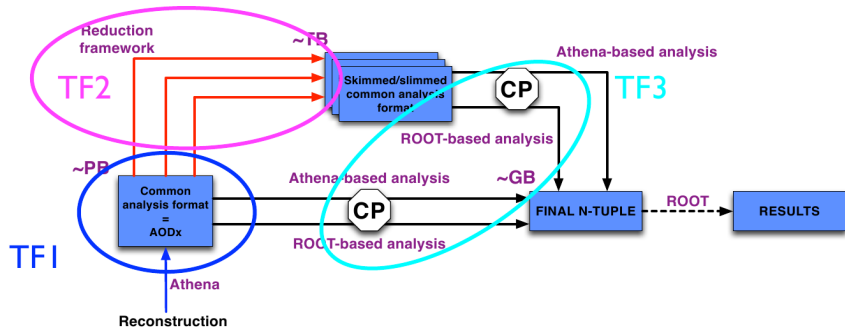
- so more complexity for processing and +20% event size

Flat resource budget:

- For storage, CPUs and network (apart for Moores law)
- For operations manpower
- The LHCC recommends that the hypothesis of flat future resources be removed from the assumptions; instead physics motivated needs should be stated.

The ATLAS Distributed Computing infrastructure needs to evolve in order to face those challenges

# RECOMMENDATIONS OF ANALYSIS MODEL STUDY GROUP



Analysis Model Study Group recommended change of ATLAS (offline) analysis model - 3 (4) task forces setup

- TF1: design new analysis merged ROOT/Athena EDM ntuple: xAOD
- TF2: data reduction framework
- TF3: analysis framework and tools

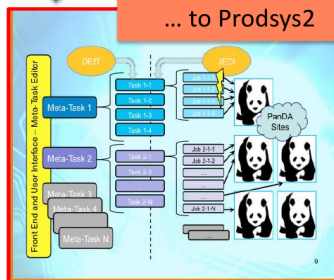
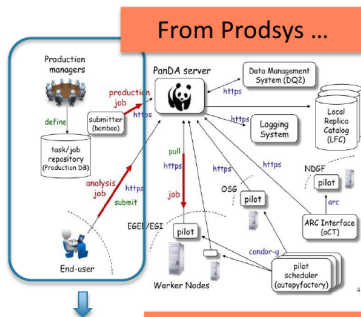
# WORKLOAD MANAGEMENT IN RUN 2: PRODSYS2

## 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



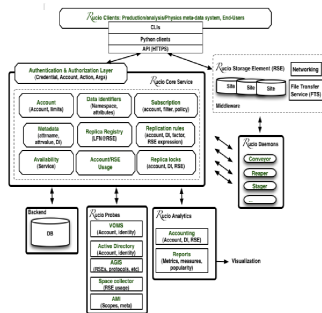
# DATA MANAGEMENT IN RUN 2: RUCIO

## Implements a highly evolved Data Management model

- File (rather than dataset) level granularity
- Multiple file ownership per user/group/activity

### Features:

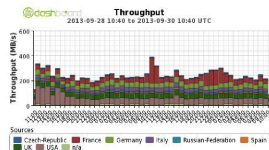
- Unified dataset/file catalogue with support for metadata
- Built-in policy based data replication for space and network optimization
- Redesign leveraging new middleware capabilities (FTS/GFAL-2)
- Plug-in based architecture supporting multiple protocols (SRM/gridFTP/xrootd/HTTP...)



# 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



# OPPORTUNISTIC RESOURCES: CLOUDS

A “Cloud” infrastructure allows to demand resources through an established interface

- (If it can) it gives you back a (virtual) machine for you to use
- You become the administrator of your cluster

## Free opportunistic cloud resources

- The ATLAS HLT farm is accessible through cloud interface during the Long Shutdown
- Academic facilities offering access to their infrastructure through a cloud interface

## Cheap opportunistic cloud resources

- Commercial Infrastructures (Amazon EC2, Google, ...) offering good deals under restrictive conditions

## Work done in ATLAS Distributed Computing

- Define a model for accessing and utilizing cloud resources effectively in ATLAS
- Develop necessary components for integration with cloud resources and automation of the workflows



# OPPORTUNISTIC RESOURCES: HPCs

HPC offers important and necessary opportunities for HEP

- Possibility to parasitically utilize empty cycles

Bad news: very wide spectrum of site policies

- No External connectivity
- Small Disk size
- No pre-installed Grid clients
- One solution unlikely to fit all

Good news: from code perspective, anything seriously tried so far did work

- Geant4, ROOT, generators
- Short jobs preferable for backfilling

HPC exploitation is now a coordinated ATLAS activity



Oak Ridge Titan System	
Architecture:	Cray XK7
Cabinets:	200
Total cores:	299,008 Opteron Cores
Memory/core:	2GB
Speed:	20+ PF
Square Footage	4,352 sq feet

# EVENT SERVICE, MONITORING, GRID INFORMATION SYSTEM, DATABASES

## Event service

- Under development: Store small metadata info of every data and MC event in a database and make it accessible for special services

## Monitoring

- <http://adc-monitoring.cern.ch/>
- Converged on an “ADC monitoring architecture”
- Rationalization of monitoring system and Porting monitoring to the newly developed components (Prodsys2, Rucio)

## AGIS - Grid Information System

- Source repository of information for PanDA and DDM
- More a configuration service than an information system

## Databases

- Relational databases (mostly Oracle) are currently working well
- Many use cases might be more suitable for NoSQL solution (Hadoop: DDM accounting, Event service possible candidate)
- Frontier/Squid fully functional for all remote database access at all sites

# SUMMARY AND CONCLUSIONS

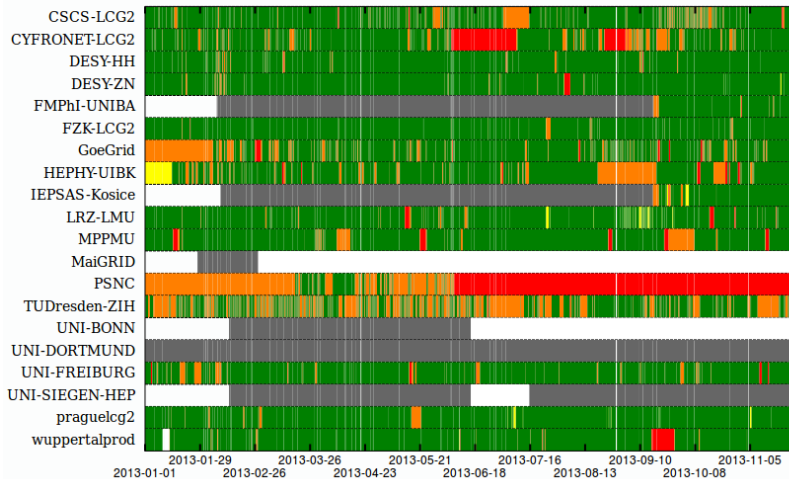
- ATLAS distributed computing development is driven by operations
- Many R&D projects:
  - quickly converge on possible usability in production
  - All R&Ds made it to production (NoSQL, FAX, Cloud Computing)
- Core components (Prodsys2 and Rucio) on schedule
- Model of incremental development steps and commissioning has been a key component for the success of Run1
- New analysis model next big challenge for 2014 before 2015 data taking

BACKUP

# DE CLOUD SITE AVAILABILITY JAN - NOV 13

## Status of panda analy status NEW

7896 Hours from 2013-01-01 00:00 to 2013-11-26 00:00

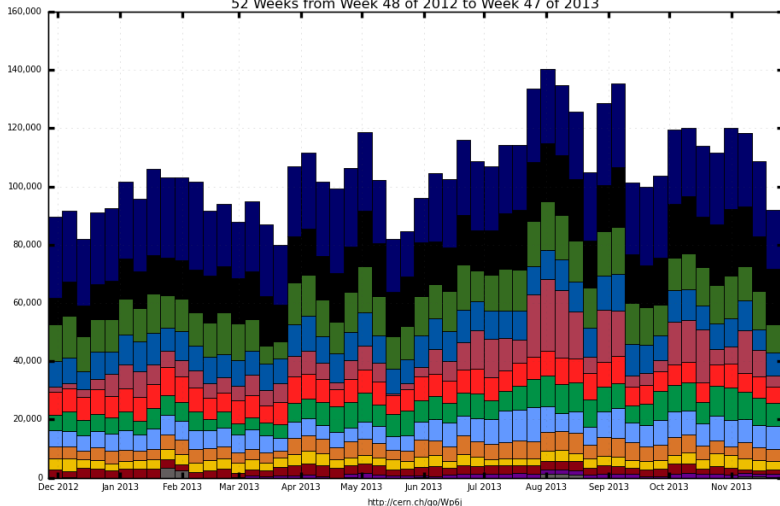


# RUNNING PRODUCTION JOBS - LAST 12 MONTH



## Running jobs

52 Weeks from Week 48 of 2012 to Week 47 of 2013



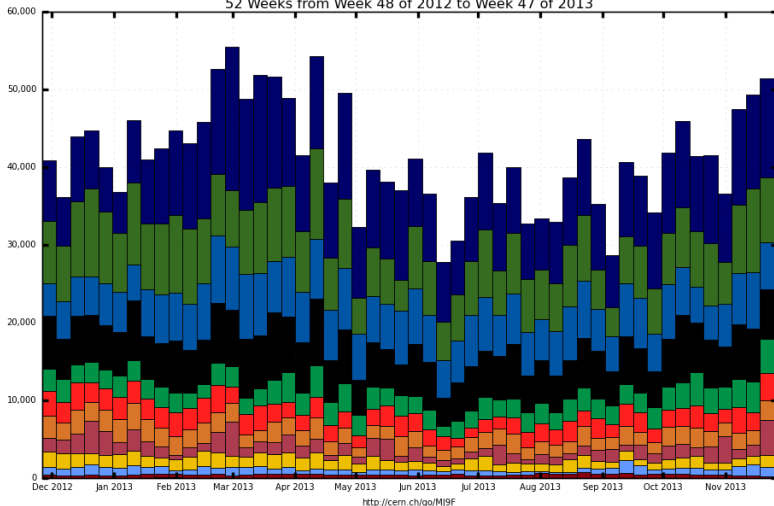
Maximum: 140,351 , Minimum: 0.00 , Average: 101,791 , Current: 13,589

# RUNNING ANALYSIS JOBS - LAST 12 MONTH



## Running jobs

52 Weeks from Week 48 of 2012 to Week 47 of 2013



Maximum: 55,458 , Minimum: 0.00 , Average: 40,067 , Current: 14,874



# DE-CLOUD TOTAL FOR ATLAS (OCT12-SEP13)

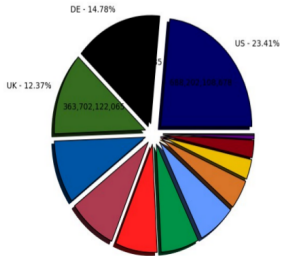
- DE total: 2<sup>nd</sup> production (14.8%); 4<sup>th</sup> for Analysis (14.2%)

(2012 2<sup>nd</sup>, 14.3%)

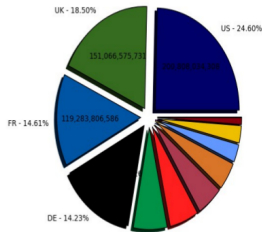
(2012: 3<sup>rd</sup>, 13.9%)



Wall Clock consumption All Jobs in seconds (Sum: 2,939,571,393,007)



Wall Clock consumption All Jobs in seconds (Sum: 816,382,943,503)



US - 23.41% (688,202,108,678)    DE - 14.78% (434,445,862,485)    UK - 12.37% (363,702,122,065)    FR - 9.45% (277,685,651,117)  
 CN - 0.42% (127,565,505,353)    CA - 7.41% (217,861,065,083)    IT - 0.75% (218,428,672,883)    NO - 0.61% (178,418,601,802)  
 NL - 4.43% (130,082,324,145)    ES - 2.99% (88,028,252,358)    JP - 0.56% (16,493,971,218)    other - 0.19% (5,678,293,478)

US - 24.60% (200,808,038,308)    UK - 18.50% (151,066,575,731)    FR - 14.61% (119,283,806,587)    DE - 14.23% (116,166,119,826)  
 IT - 0.40% (33,077,422,430)    CA - 5.57% (45,632,098,648)    CN - 0.30% (160,577,422,430)    NL - 4.39% (130,082,324,145)  
 NO - 2.81% (22,967,896,358)    ES - 2.77% (22,312,411,561)    JP - 0.07% (18,727,867,527)    other - 0.01% (86,705,802)  
 JP - 0.00% (0.00)

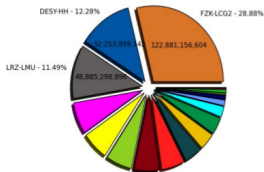
# ATLASDE CLOUD BY SITE OCT12-SEP13

- GridKa: production 29%, analysis 20%
- Desy/MPP : production 22%, analysis 27%
- DE University sites: production 30%, analysis 29%
- non-De sites: production 19%, analysis 24%

## Production



Wall Clock consumption All Jobs in seconds (Sum: 425,521,601,431)



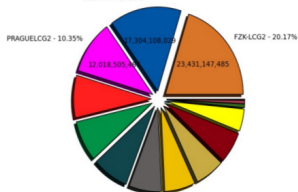
FZK-LC02 - 28.88% (122,881,156,604)  
 LRZ-LMU - 11.49% (48,885,298,896)  
 MPPMJI - 6.13% (25,583,204,144)  
 KOFORNO - 5.82% (24,148,028,807)  
 UNI-FRANBURG - 4.89% (20,387,356,588)  
 DESY-ZN - 3.82% (15,754,122,238)  
 PRIMA-LINBA - 3.23% (13,536,085,898)  
 DESY-LINBA - 0.54% (2,286,521,727)  
 TUOSPOREN-ZRH - 0.32% (1,316,422,746)  
 IAN-SEPINA-APP - 0.00% (0.000,000)

DESY-HH - 12.28% (52,253,958,141)  
 PRAGUE-LC02 - 6.77% (28,189,841,831)  
 MPPERTNA-PROD - 6.05% (25,149,347,728)  
 CSCS-LC02 - 4.49% (18,621,700,843)  
 CYFRONET-LC02 - 3.85% (15,931,134,838)  
 UN-DOORHARD - 2.27% (9,405,289,280)  
 IRIAS-AGSC - 0.89% (3,625,724,288)  
 PISC - 0.52% (2,161,216,719)  
 NAGARA - 0.33% (1,368,868,888)

## Analysis



Wall Clock consumption All Jobs in seconds (Sum: 116,164,119,926)  
DESY-HH - 14.90%

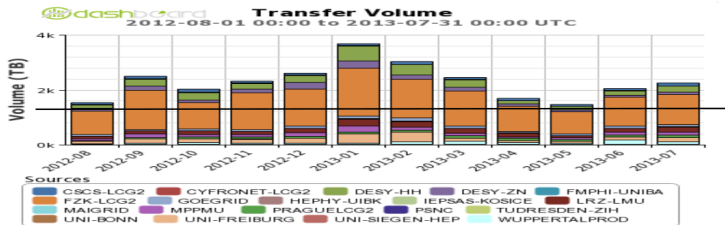
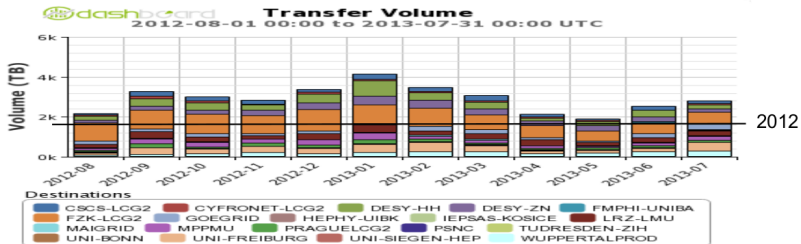


FZK-LC02 - 20.17% (23,431,147,485)  
 CSCS-LC02 - 8.17% (9,482,875,888)  
 LRZ-LMU - 6.75% (7,842,398,888)  
 KOFORNO - 4.02% (4,708,206,488)  
 TUOSPOREN-ZRH - 0.84% (984,884,888)  
 IRIAS-AGSC - 0.02% (234,199)

DESY-HH - 14.90% (17,304,108,020)  
 DESY-ZN - 7.96% (9,272,503,639)  
 CYFRONET-LC02 - 6.39% (7,422,798,828)  
 MPPMJI - 4.28% (4,973,122,428)  
 PISC - 0.31% (351,843,988)  
 NARA - 0.00% (0.000,000)

PRAGUE-LC02 - 10.35% (12,028,505,481)  
 UNI-FRANBURG - 7.53% (8,749,114,442)  
 MPPERTNA-PROD - 6.05% (7,024,451,831)  
 DESY-LINBA - 0.70% (808,178,831)  
 PRIMA-LINBA - 0.61% (718,132,727)  
 UN-DOORHARD - 0.00% (0.000,000)

# ATLAS TRANSFER TO/FROM DE BY SITE



# ATLAS TRANSFER TO/FROM DE BY CLOUD

