ATLAS Computing: the Run-2 experience

Fernando Barreiro Megino
on behalf of ATLAS Distributed Computing

KEK, 4 April 2017

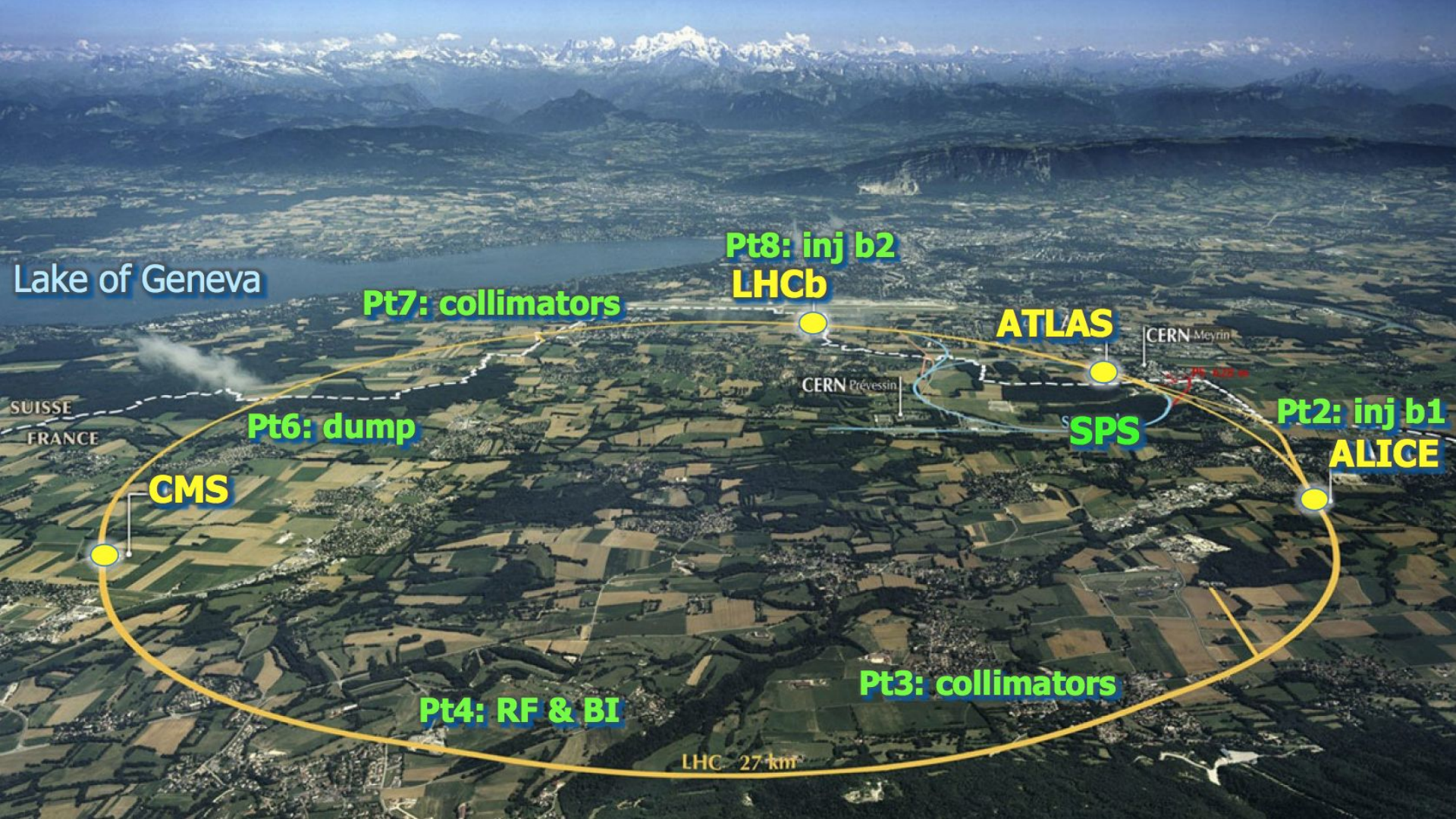
About me

- SW Engineer (2004) and Telecommunications Engineer (2007), Universidad Autónoma de Madrid (Spain)
- Working on ATLAS Distributed Computing (ADC) since 2008
  - 2008-2012: Distributed Data Management developer
  - 2012-13: ATLAS Cloud Computing co-coordinator
  - 2015-now: Workload Management developer and co-coordinator since April 2016
- 2013-2014: JP Morgan Technology Division in Geneva

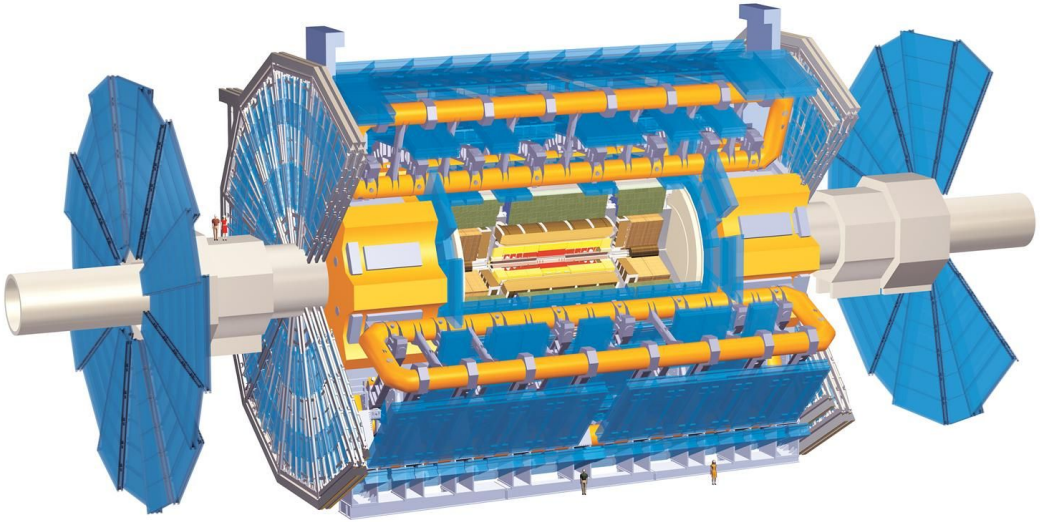
Outline

- ATLAS workflows: the data processing chain
- ATLAS Distributed Computing in Run 2
  - The Worldwide LHC Computing Grid (WLCG)
  - Distributed Data Management
  - Distributed Workload Management
- ATLAS Distributed Computing: operations and support
- Conclusions

ATLAS workflows: the data processing chain

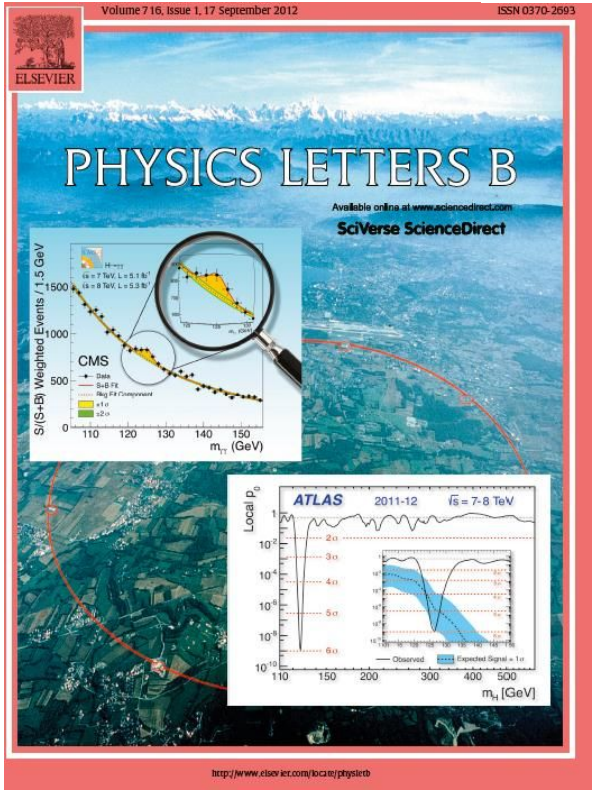

Lake of Geneva
Pt8: inj b2
LHCb
Pt7: collimators
ATLAS
CERN Meyrin
CERN Prévessin
SUISSE
FRANCE
Pt6: dump
SPS
Pt2: inj b1
ALICE
CMS
Pt3: collimators
Pt4: RF & BI
LHC 27 km

From collisions to papers in ATLAS

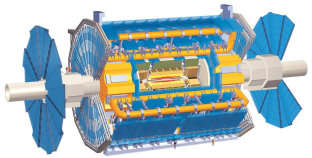


150M sensors
Collisions at 40MHz

?



The data processing chain



Trigger

Raw data (RAW)

Reconstruction

Analysis Object Data (AOD)

Derivation

Generation

Event generator output (EVNT)

Simulation

Simulated interaction with detector (HITS)

Digitization+Trigger

Simulated detector output (RDO)

Reconstruction

Analysis Object Data (AOD)

Derivation

Organized production

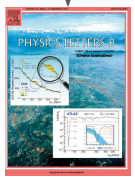
Derived AOD (DAOD)

Chaotic analysis

Detector data

Analysis

Simulated data



Online

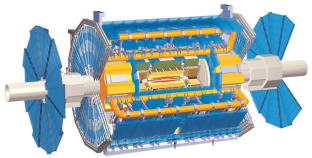
Tier-0

Grid

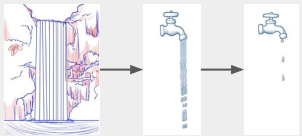
UNIVERSITY OF TEXAS ARLINGTON

PanDA

The data processing chain



2 level, online system (HW+SW)



Reduce event rate from 40 MHz (60TB/s) to 1kHz (1.6GB/s) based on signatures
Event size ~1.6MB

Trigger

Raw data (RAW)

Reconstruction

bject Data (AOD)

Derivation

Generation

Event generator output (EVNT)

Simulation

Simulated interaction with detector (HITS)

Digitization+Trigger

Simulated detector output (RDO)

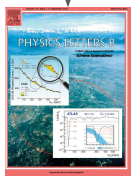
Reconstruction

Analysis Object Data (AOD)

Derivation

Derived AOD (DAOD)

Analysis

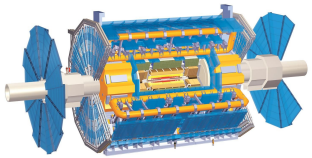


Online
Tier-0
Grid

UNIVERSITY OF TEXAS ARLINGTON

PanDA

The data processing chain



Trigger

Raw data (RAW)

Reconstruction

ject Data (AOD)

Derivation

Data quality assessment
Calibration
Mast out noisy channels
Reconstruct trajectory of particles through detector

Generation

Event generator output (EVNT)

Simulation

Simulated interaction with detector (HITS)

Digitization+Trigger

Simulated detector output (RDO)

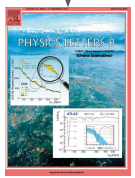
Reconstruction

Analysis Object Data (AOD)

Derivation

Derived AOD (DAOD)

Analysis

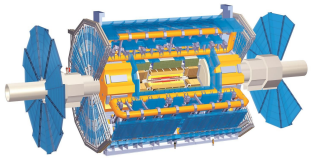


Online
Tier-0
Grid

UNIVERSITY OF TEXAS ARLINGTON

PanDA

The data processing chain



Trigger

Raw data (RAW)

Reconstruction

bject Data (AOD)

Derivation

Generation

Event generator output (EVNT)

Simulation

Simulated interaction with detector (HITS)

Digitization+Trigger

Simulated detector output (RDO)

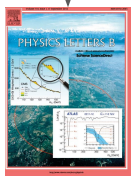
Reconstruction

Analysis Object Data (AOD)

Derivation

Derived AOD (DAOD)

Analysis



Reduce size of outputs to facilitate analysis.
Skimming: remove events
Thinning: remove objects within events
Slimming: remove variables within objects

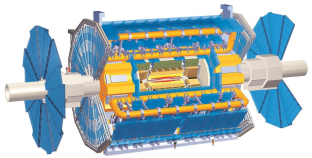
This step is done centrally to reduce resource needs by users

Online
Tier-0
Grid

UNIVERSITY OF TEXAS ARLINGTON

PanDA

The data processing chain



Trigger

Raw data (RAW)

Reconstruction

Analysis Object Data (AOD)

Derivation

Generation

Event generator

Simulation

Simulated interac

Digitization+Trigger

Simulated detector

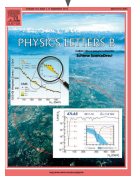
Reconstruction

Analysis Object Da

Derivation

Derived AOD (DAOD)

Analysis



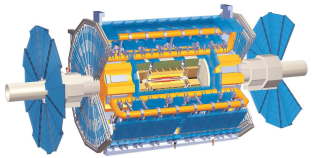
Monte Carlo production
Event generation: simulate particle interactions
Detector simulation: calculate how particles interact with detector material
Digitization: turn simulated energy into detector response that looks like the raw data from the detector

Online
Tier-0
Grid

UNIVERSITY OF TEXAS ARLINGTON

PanDA

The data processing chain



Trigger

Raw data (RAW)

Reconstruction

Analysis Object Data (AOD)

Derivation

Generation

Event generator output (EVNT)

Simulation

Simulated interaction with detector (HITS)

Digitization+Trigger

Simulated detector output (RDO)

Reconstruction

Analysis Object Data (AOD)

Derivation

Organized production

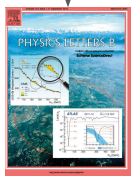
Derived AOD (DAOD)

Chaotic analysis

Detector data

Analysis

Simulated data



Online

Tier-0

Grid

UNIVERSITY OF TEXAS ARLINGTON

PanDA

ATLAS Distributed Computing (ADC) for Run 2

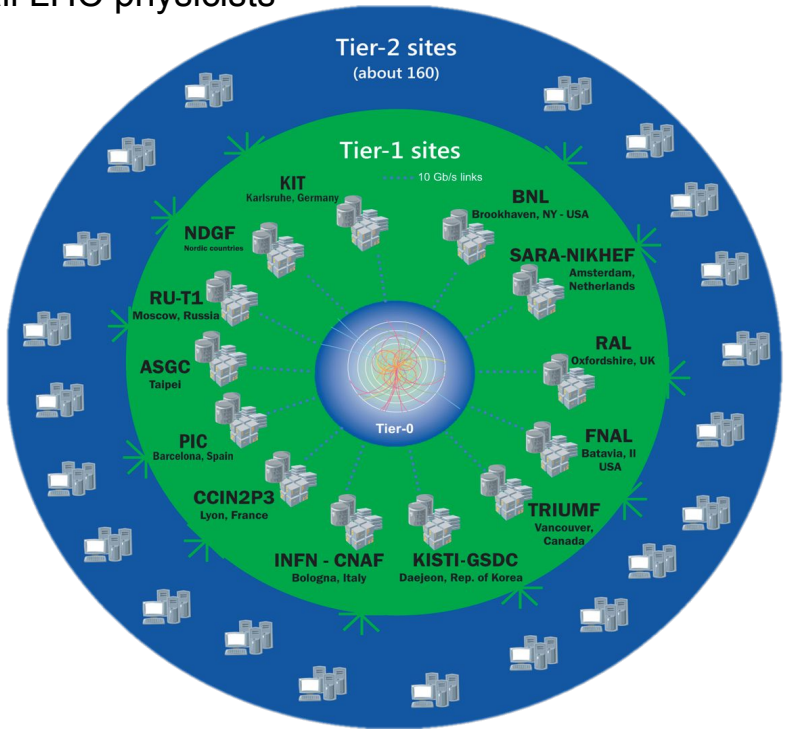
Worldwide LHC Computing Grid

- International collaboration to distribute and analyse LHC data
- Integrates computing centres worldwide that provide **computing** and **storage** resource into a single infrastructure accessible by all LHC physicists

- **Tier-0 (CERN)**: data recording and archival, prompt reconstruction, calibration and distribution

- **Tier-1s**: T0 overspilling, second tape copy of detector data, more intensive tasks

- **Tier-2s:** Processing centers, being the differences with T1s increasingly blurry - more later



For **all** experiments:

- nearly 170 sites
- ~350k cores
- 200 PB of disk
- 10 Gb links and up

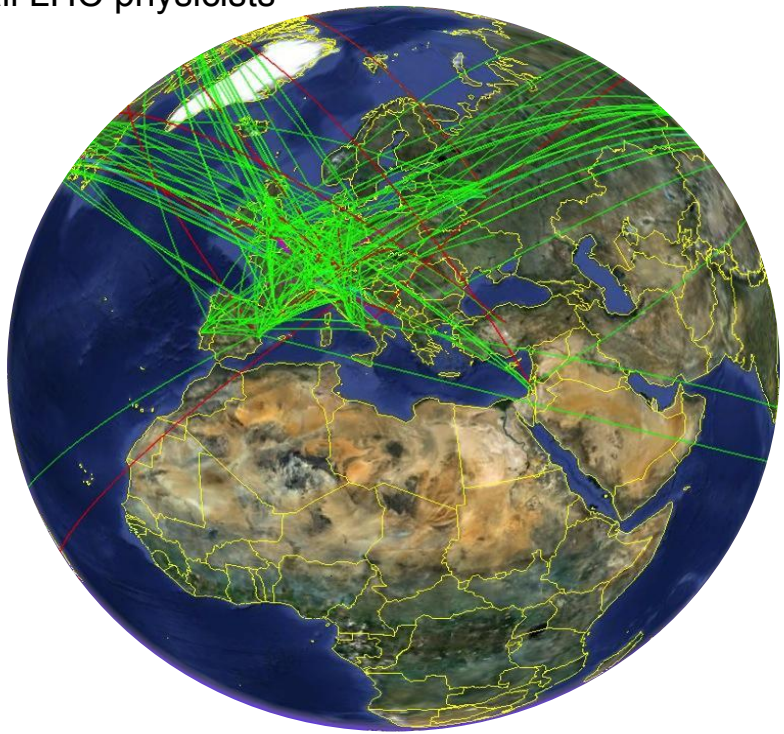
Worldwide LHC Computing Grid

- International collaboration to distribute and analyse LHC data
- Integrates computing centres worldwide that provide **computing** and **storage** resource into a single infrastructure accessible by all LHC physicists

- **Tier-0 (CERN)**: data recording and archival, prompt reconstruction and calibration and distribution

- **Tier-1s**: T0 overspilling, second tape copy of detector data, memory and CPU intensive tasks

- **Tier-2s:** Processing centers, being the differences with T1s increasingly blurry - more later



For all experiments:
- nearly 170 sites
- ~350k cores
- 200 PB of disk
- 10 Gb links and up

ATLAS Grid Information System: AGIS

VO name	Experiment Name	State	Tier	Site	PANDA Sites	Regional Center	CLOUD
atlas	BEIJING-LCG2	ACTIVE	T2D	BEIJING-LCG2	BEIJING-LCG2	CN-IHEP	FR
atlas	GRIF-IRFU	ACTIVE	T2D	GRIF	GRIF-IRFU	FR-GRIF	FR
atlas	GRIF-LAL	ACTIVE	T2D	GRIF	GRIF-LAL	FR-GRIF	FR
atlas	GRIF-LPNHE	ACTIVE	T2D	GRIF	GRIF-LPNHE	FR-GRIF	FR
atlas	IN2P3-CC	ACTIVE	T1	IN2P3-CC	IN2P3-CC, IN2P3-CC_HPC, IN2P3-CC_OPENSTACK	FR-CCIN2P3	FR
atlas	IN2P3-CC-T2	ACTIVE	T2	IN2P3-CC-T2	IN2P3-CC-T2	FR-IN2P3-CC-T2	FR
atlas	IN2P3-CC-T3	ACTIVE	T3	IN2P3-CC	IN2P3-CC-T3	FR-CCIN2P3	FR
atlas	IN2P3-CPPM	ACTIVE	T2D	IN2P3-CPPM	IN2P3-CPPM	FR-IN2P3-CPPM	FR
atlas	IN2P3-LAPP	ACTIVE	T2D	IN2P3-LAPP	IN2P3-LAPP	FR-IN2P3-LAPP	FR
atlas	IN2P3-LPC	ACTIVE	T2D	IN2P3-LPC	IN2P3-LPC	FR-IN2P3-LPC	FR
atlas	IN2P3-LPSC	ACTIVE	T2D	IN2P3-LPSC	IN2P3-LPSC	FR-IN2P3-LPSC	FR
atlas	RO-02-NIPNE	ACTIVE	T2	RO-02-NIPNE	RO-02-NIPNE	RO-LCG	FR
atlas	RO-07-NIPNE	ACTIVE	T2	RO-07-NIPNE	RO-07-NIPNE	RO-LCG	FR
atlas	RO-14-ITIM	ACTIVE	T2	RO-14-ITIM	ITIM	RO-LCG	FR
atlas	RO-16-UAIC	ACTIVE	T2	RO-16-UAIC	RO-16-UAIC	RO-LCG	FR
atlas	TOKYO-LCG2	ACTIVE	T2	TOKYO-LCG2	TOKYO-LCG2	JP-Tokyo-ATLAS-T2	FR

List of attached sites

Each site consists of multiple storage endpoints and batch queues

DDM endpoint info

Operations: Clone DDM Endpoint | Update DDM Endpoint information | Show Changes log

Name:	TOKYO-LCG2_DATADISK
Type:	DATADISK
SRM:	token:ATLASDATADISK:srm://lcg-se01.icepp.jp:8446/srm/managerv2?SFN=/dpm/icepp.jp/home/atlas/atlasdatadisk/
Token:	ATLASDATADISK
Phys Group:	
Is Rucio enabled:	No
Domain:	.*icepp.jp.*/atlasdatadisk/.*
mkdir:	No
Is cache:	No
Is Deterministic:	Yes
Is Volatile:	No
Space method:	lcg-stmd
Space Usage:	*not set*
Tape:	No
Pledged:	No
Tool Assigner::	lcg
LFC:	CERN-PROD_RUCIO_Catalog
Site:	TOKYO-LCG2
ATLAS Site:	TOKYO-LCG2
SE info:	
Resource:	New Storage Relation: NULL (NULL)
Storage element:	TOKYO-LCG2-SRM-lcg-se01.icepp.jp (srm://lcg-se01.icepp.jp:8446/srm/managerv2?SFN=)

Storage endpoint configurations

RC Site | ATLASSite | DDMEndpoint | PANDA Queue | Service | Central Services | DDM Groups | PandaQueueObject Info

PandaQueue Object details

PanDa Queue name: TOKYO-LCG2-all-ce-atlas-lcgpbs
PanDA resource name: TOKYO
PanDA resource type: GRID
PanDA Site: TOKYO-LCG2
ATLAS Site: TOKYO-LCG2
Parent object: TOKYO-LCG2_VIRTUAL
is_virtual: No

Type: production
– **is_default:** Yes
Status: online
Status control: manual
Comment: no active blacklisting rules defined

Capability: score
HC param: AutoExclusion
HC Suites: PFT
Pilot Manager: APF
CVMFS: Yes

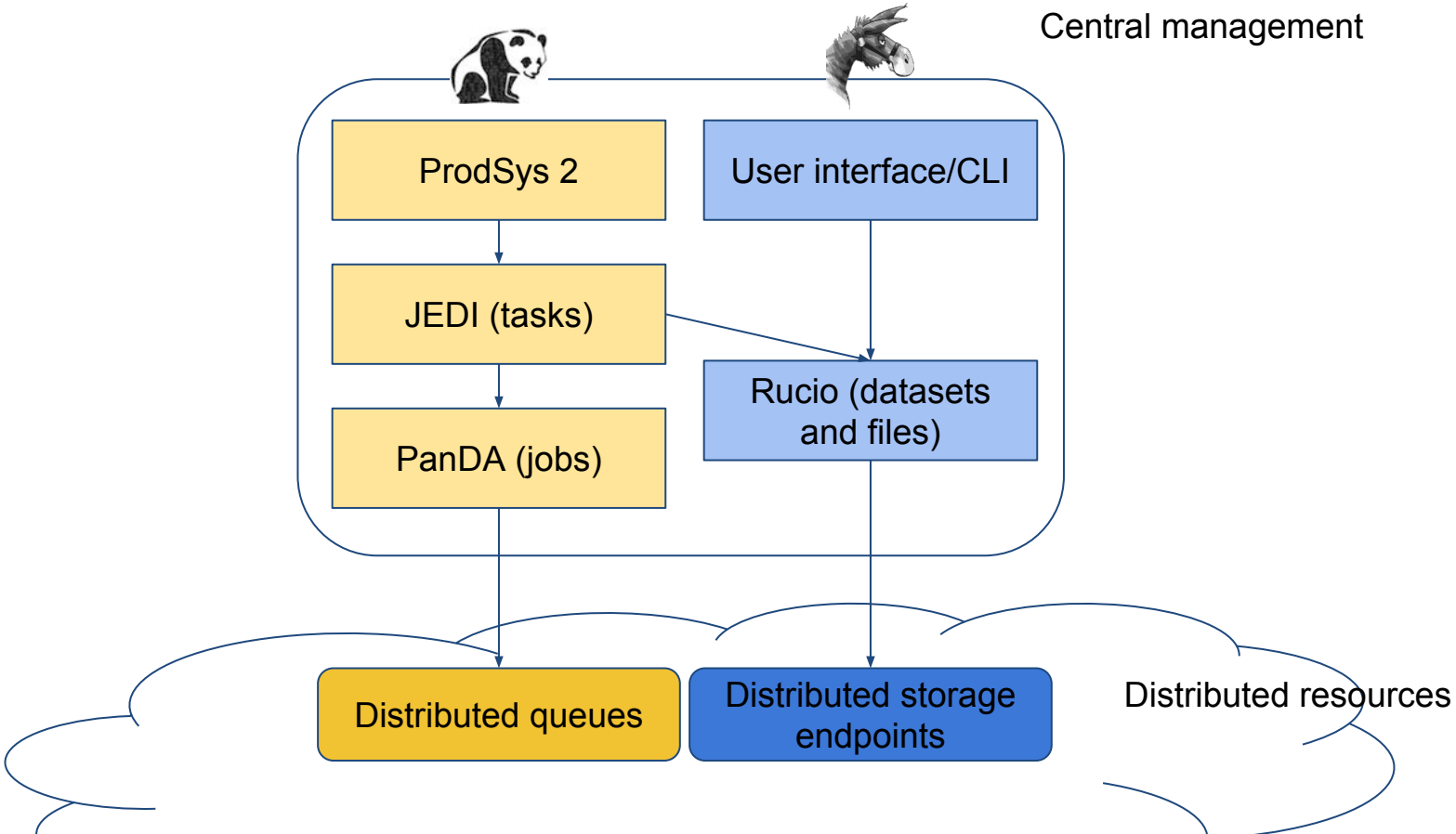
Last Modified: 2017-02-06 10:56
State: ACTIVE
State updated: 2014-09-26 12:17
State Comment: AUTO migrated from old PQ object TOKYO-LCG2-all-ce-atlas-lcgpbs

Description: *Not set*

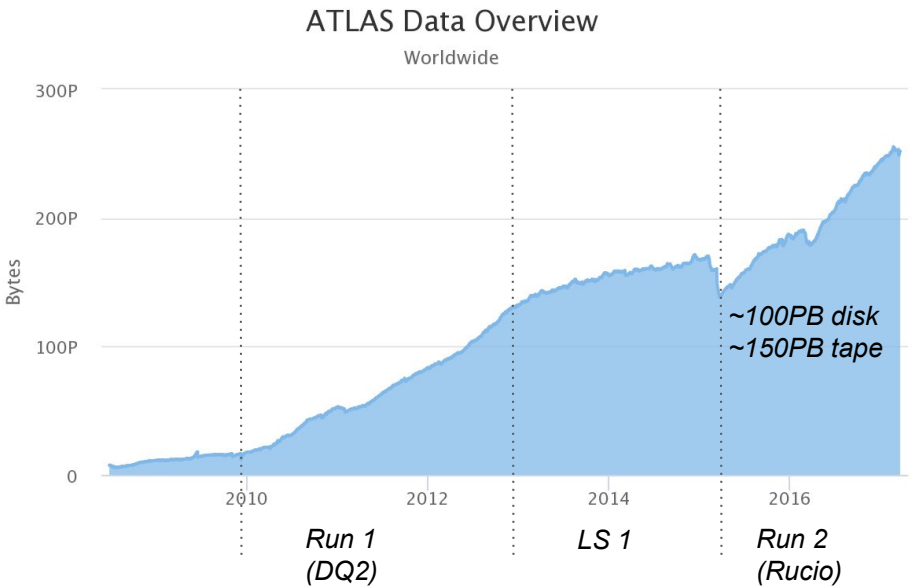
Update PandaQueueObject data | Clone PandaQueueObject | Jobs monitor

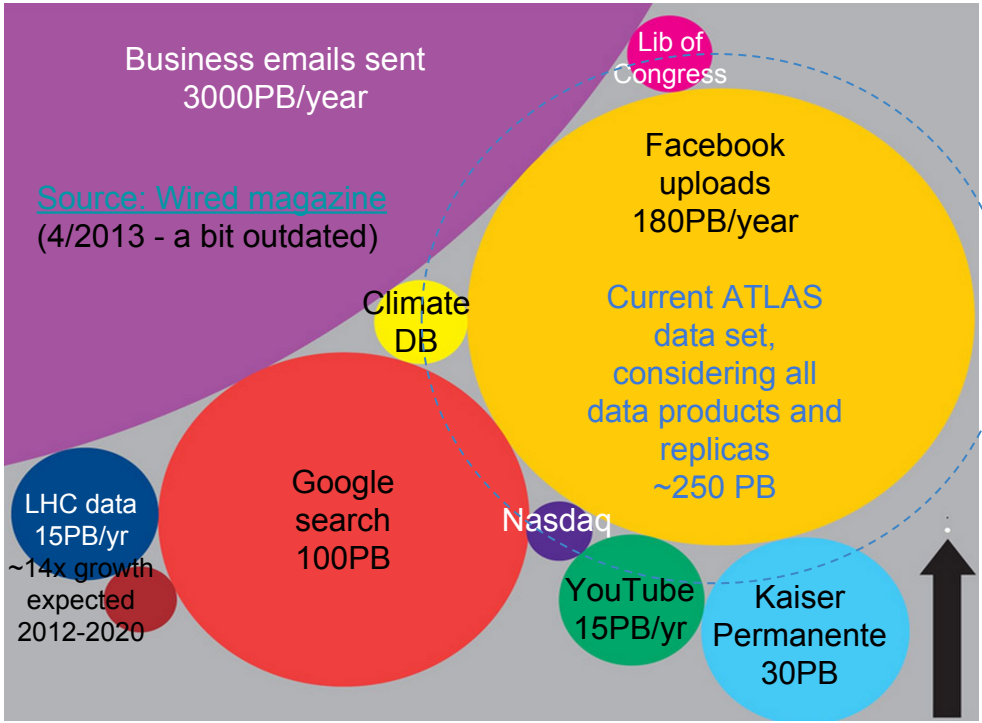
Batch/PanDA queue configurations

Workload and data management system

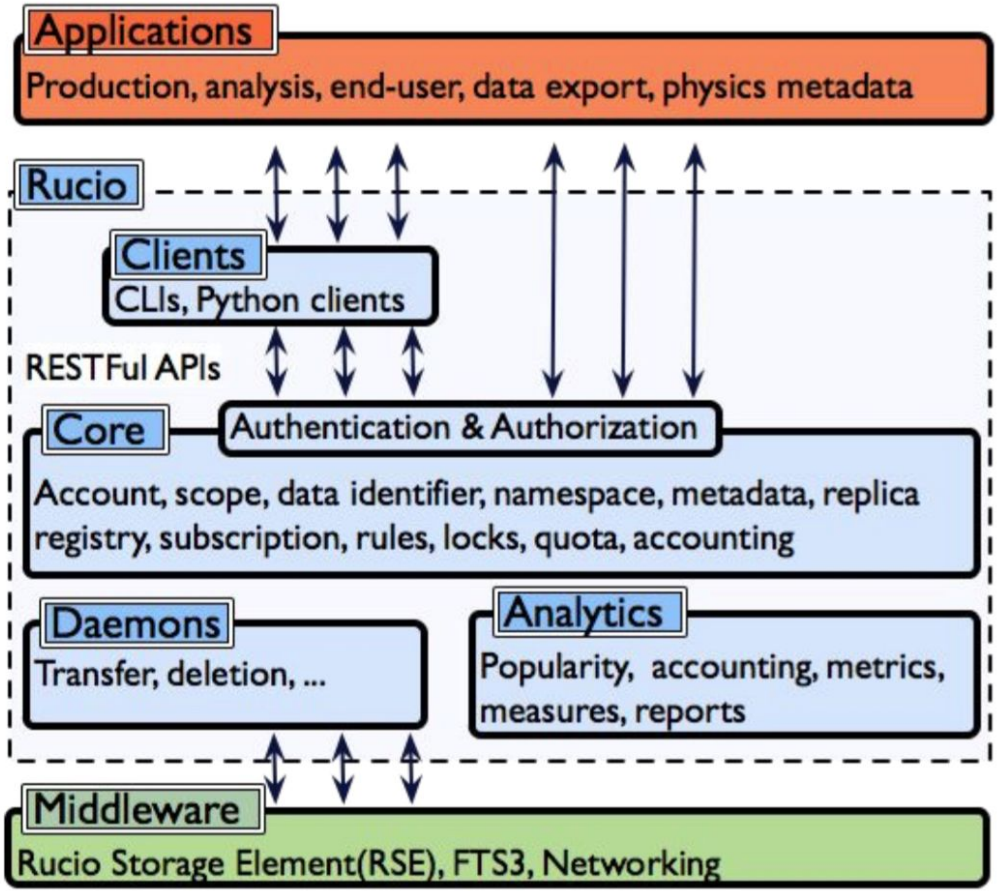


ATLAS Distributed Data Management: Rucio





Data Management: Rucio architecture

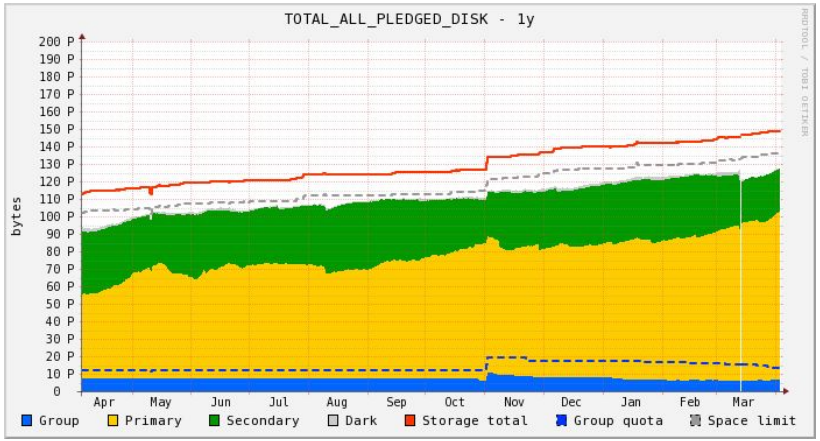


Rucio features and concepts

- Rucio accounts can be mapped to users or groups (e.g. Higgs)
- Namespace is partitioned by scopes (users, groups and other activities)
- Data ownership for users and groups: possibility to enable quota systems
- Replica management: rules define number of replicas and conditions on sites
- Granular data handling at file level - no external file catalogs
- Support of multiple protocols for file handling (access/copy/deletion)
  - SRM, HTTP/WebDAV, gridFTP
- Metadata storage: extensible key-value implementation
  - System-defined: size, checksum, creation time
  - Physics: number of events
  - Production: job/task that created the file

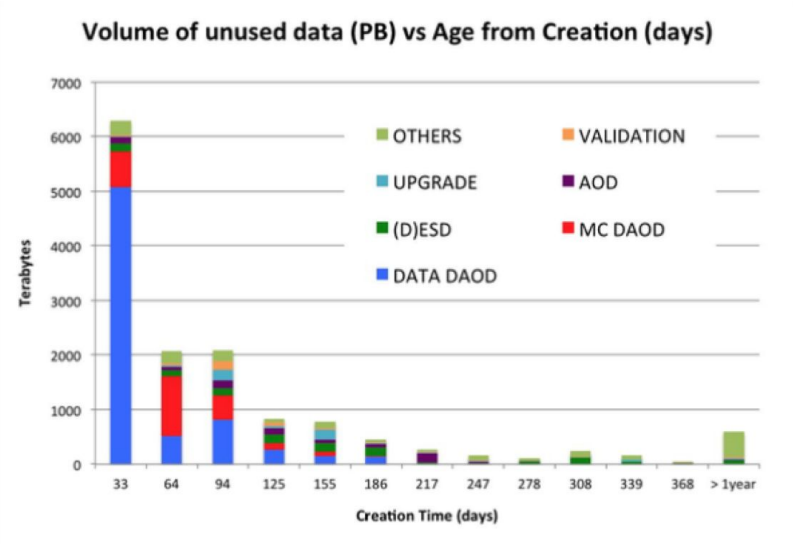
Data policies and lifecycle

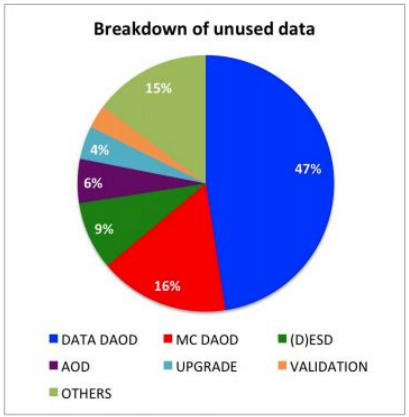
- Run 2 is very tight on space and relies on fully dynamic data replication and deletion
- Minimalistic pre-placement of only 2 replicas
- Data categories:
  - Primary (resident): base replicas guaranteed to be available on disk. Not subject to automatic clean up
  - Secondary (cache): extra replicas dynamically created and deleted according to the usage metrics
- Data rebalancing: redistribution of primary copies of popular datasets to disk resources with free space



Data policies and lifecycle

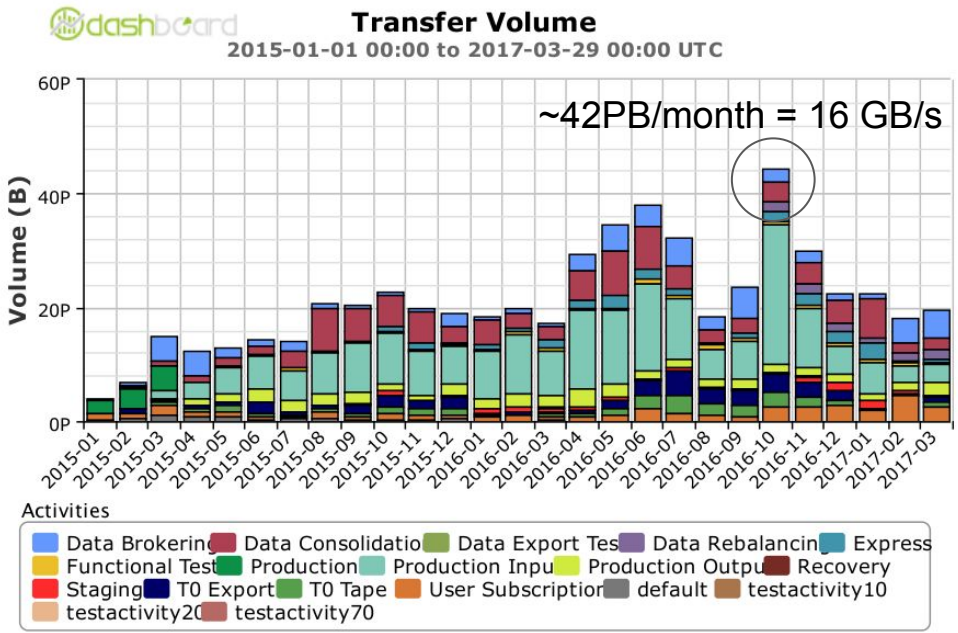
- Every dataset has a lifetime set at creation
  - 6 months for Analysis inputs (DAODs) - fast turnaround
  - 2-3 years for Monte-Carlo simulations - expensive to regenerate
  - Infinite for RAW
- Lifetime can be extended if the data is accessed
- Expired datasets can disappear any time





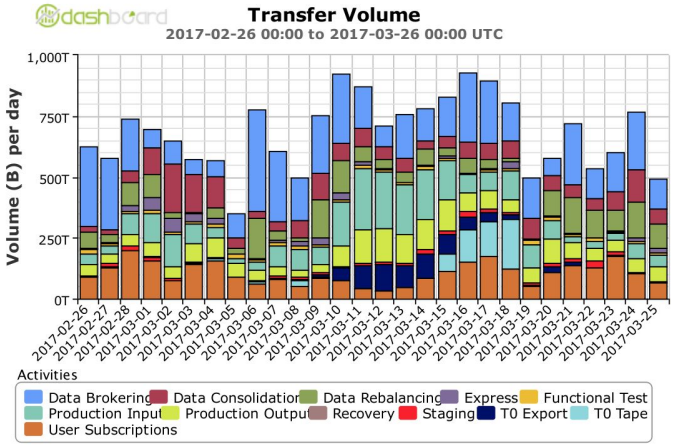
More tape and network usage

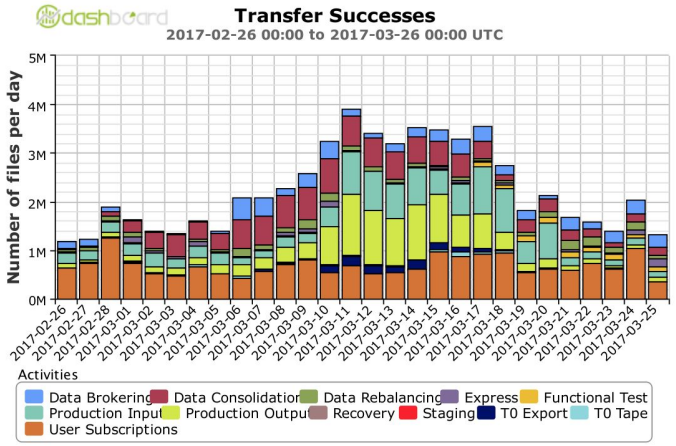
- More data and highly dynamic lifecycle: rely on tape and network
- Ongoing tests to explore the usage of tape
  - Tape pledges not reached (~70%)
  - Run derivations from tape
  - Optimization of tape access needed
- Transfer volumes keep increasing
  - LHCOPN fully utilized, including secondary network



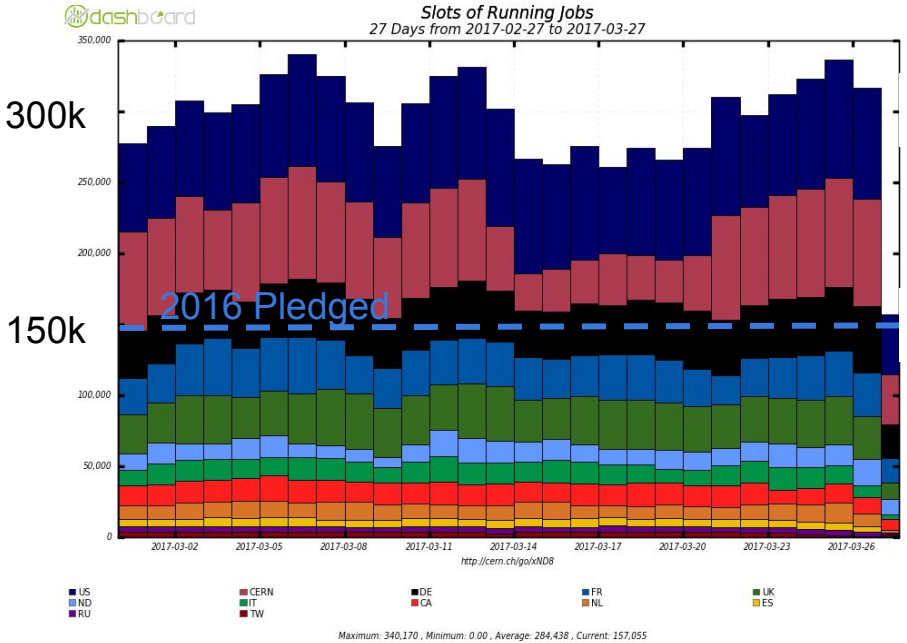
Data Management: some metrics

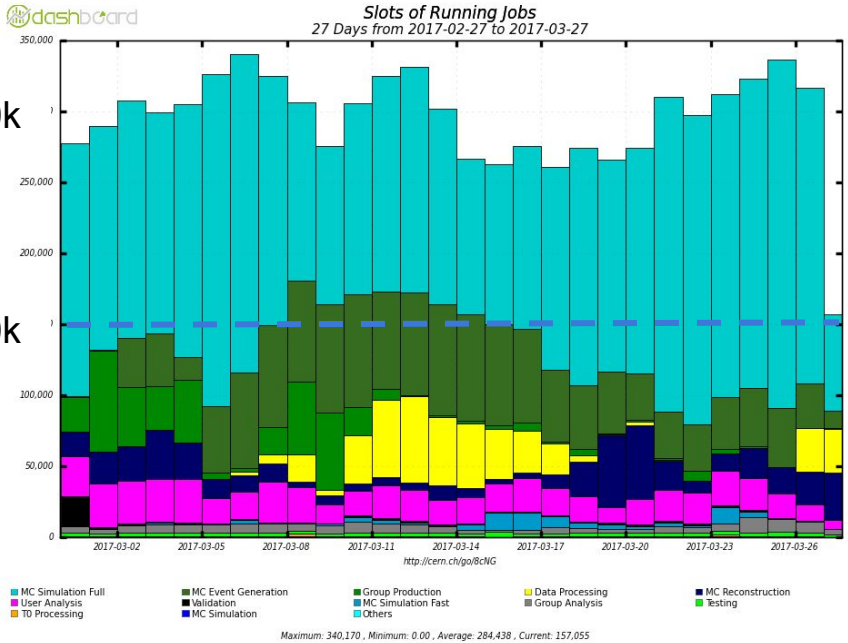
- Transfers
  - >40M files/month
  - Up to 40 PB/month
- Download
  - 150M files/month
  - 50 PB/month
- Deletion
  - 100M files/month
  - 40 PB/month





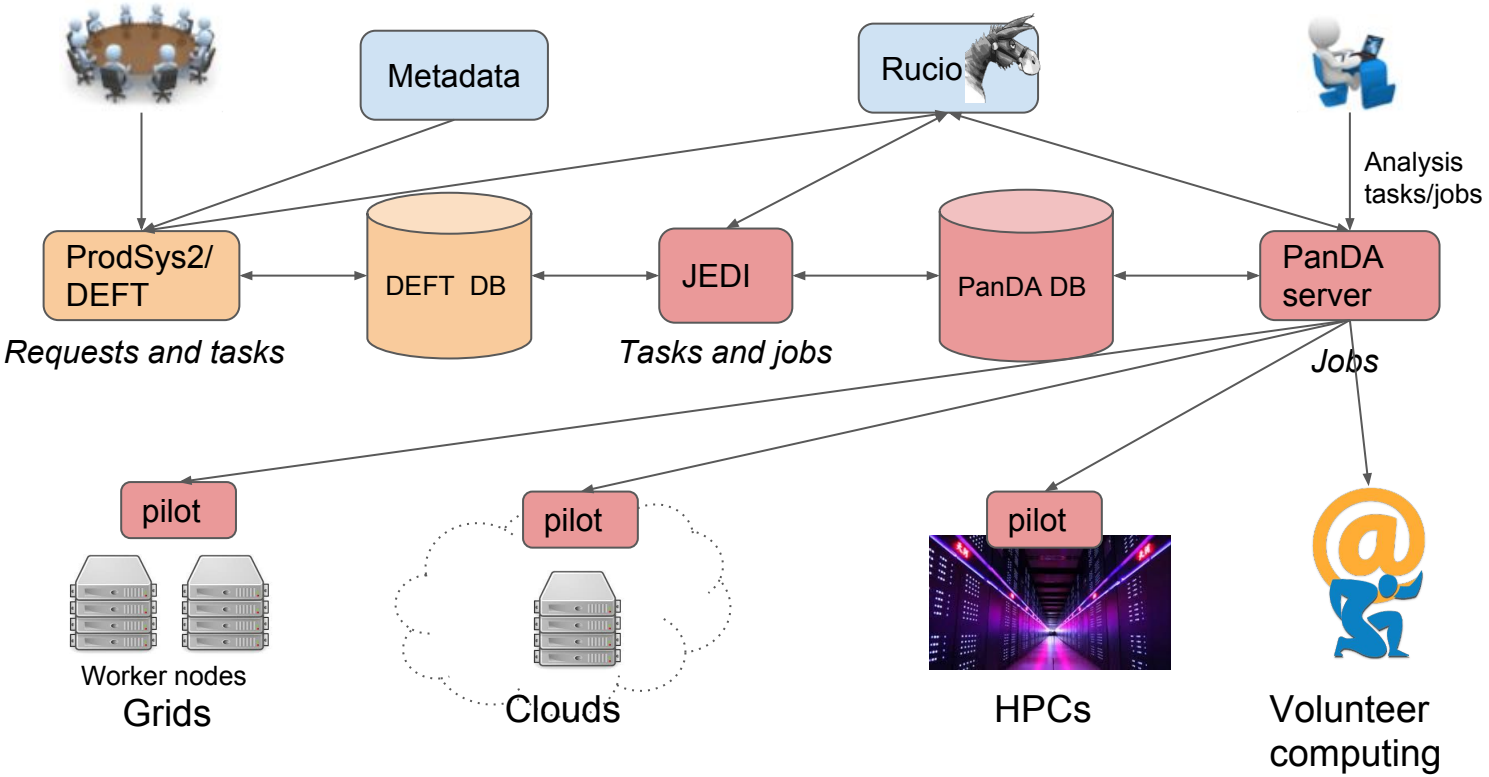
ATLAS Distributed Workload Management: PanDA





- Full grid utilization
- Resources on T1 and T2 sites are exploited beyond pledge (200% for T2s)
- Various types of resources: grid, cloud and HPCs

From requests to jobs



JEDI/PanDA workflow

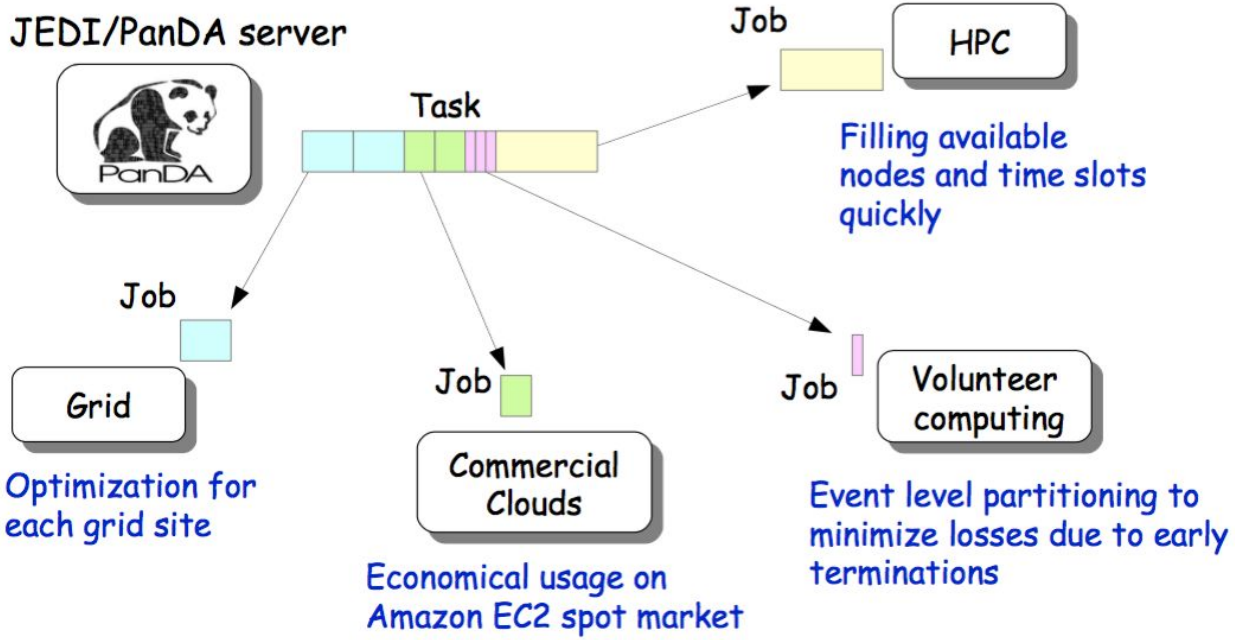
1. **Task brokerage**: tasks are assigned to the **nucleus** site that will collect the task output.
    - Assignment based on data locality, remaining work, free storage, capability to run the jobs...
2. **Job generation**
3. **Job brokerage:** jobs are assigned to processing **satellite** queues
    - Matching queue description: walltime limits, memory limits, #cores
    - And other dynamic metrics: free space, transfer backlog to nucleus, data availability, #running/#queued jobs, network connectivity to nucleus
4. **Job dispatch:** queued jobs are dispatched based on **Global Shares** targets

Task and job parameter auto-tuning

- Task and job parameters are tuned automatically
- **Scout jobs** collect real job metrics like memory and walltime
  - ~10 scout jobs are generated at the beginning of each task
  - Parameters for successive jobs in the task are optimized based on these metrics
- **Retrial module** acts on failed jobs
  - Extending memory and walltime requirements for related types of errors
  - Preventing jobs with irrecoverable errors - don’t waste CPU time retrying jobs that will never succeed
  - Rules for error codes and actions are configurable through ProdSys User Interface

Dynamic job definition

- Dynamically split workload for optimal usage of resources
- Manages workload at task, job, file and event level



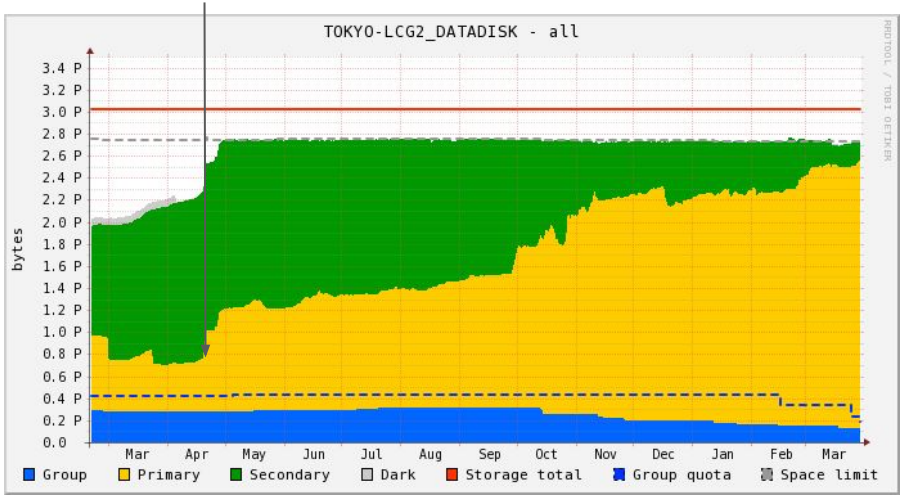
WORLD cloud

- Original ATLAS Computing Model was designed as static clouds (mostly national or geographical groupings of sites), setting data transfer perimeters
  - Tasks had to be inflexibly executed within a static cloud
  - Output of tasks had to be aggregated in the Tier 1s (O(10))
- This model had a series of shortcomings
  - WLCG networks have evolved significantly in the last two decades and limiting transfers within a cloud is no longer needed
  - Tier 2 storage was not optimally exploited and only contained secondary data
  - High priority tasks were occasionally stuck at small clouds

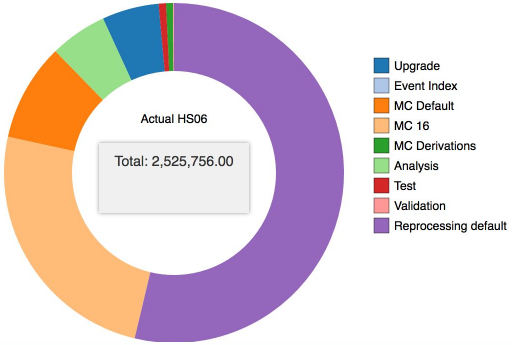
WORLD cloud

- WORLD cloud new site concepts:
  - Task nucleus: Any stable site (Tier 1 or Tier 2) can aggregate the output of a task. The capacity of being a nucleus is assigned manually based on past experience
  - Task satellites: Will process jobs and send the output to the nucleus. The satellites are defined dynamically for each task and are not confined inside the cloud
- Fully activated March 2016 and nuclei progressively added

Activation as nucleus



Global Shares

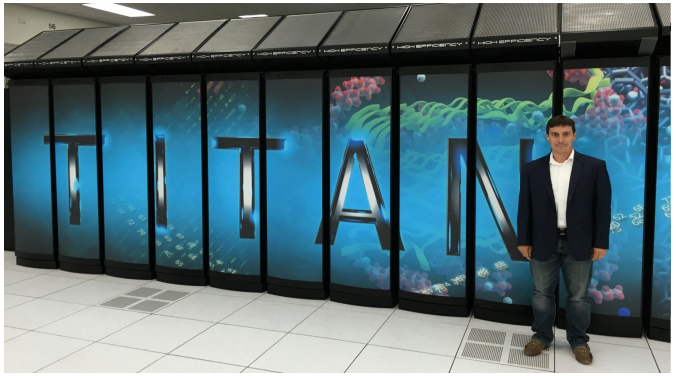


L1 Share	L2 Share	L3 Share	Actual HS06	Target HS06	Ratio	Queued
Analysis [20.0%]			137,177.95	505,151.93	27.16 %	484,131.59
Production [75.0%]			2,371,059.37	1,894,319.73	125.17 %	11,372,422.26
	MC root [17.9%]		858,642.29	451,028.51	190.37 %	6,405,418.95
		MC 16 [8.9%]	623,992.24	225,514.25	276.70 %	5,339,448.18
		MC Default [8.9%]	234,650.05	225,514.25	104.05 %	1,065,970.77
	Derivations [14.3%]		16,533.29	360,822.81	4.58 %	47,768.50
		MC Derivations [4.3%]	16,533.29	108,246.84	15.27 %	47,768.50
		Data Derivations [10.0%]	0.00	252,575.96	---	0.00
	Reprocessing [30.7%]		1,356,840.53	775,769.03	174.90 %	3,607,100.45
		Reprocessing default [24.6%]	1,356,840.53	620,615.23	218.63 %	3,607,100.45
		Heavy Ion [6.1%]	0.00	155,153.81	---	0.00
	Group production [2.9%]		0.00	72,164.56	---	13.99
	Upgrade [2.9%]		136,570.46	72,164.56	189.25 %	438,354.31
	HLT Reprocessing [2.9%]		0.00	72,164.56	---	0.00
	Validation [2.9%]		2,286.78	72,164.56	3.17 %	873,510.40
	Event Index [0.7%]		186.01	18,041.14	1.03 %	255.66
Test [5.0%]			17,522.32	126,287.98	13.87 %	55,130.64

- Distribute the currently available compute resources amongst the activities
  - Measure in currently used HS06 computing power
- Hierarchical implementation: siblings have the opportunity to inherit unused resources
- Currently only used for production shares, but in the future it will also be used for analysis vs production split

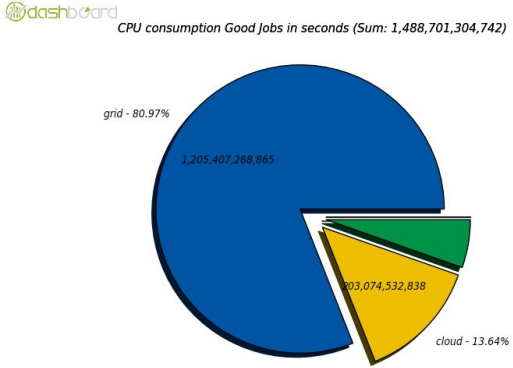
Opportunistic resources

- Centers willing to contribute to ATLAS, but not part of WLCG
  - HPC centers
  - Shared academic clusters
  - Academic and commercial Clouds
  - Volunteer computing
- Reconfiguration of ATLAS online cluster
- Some of these centers have more computing power than the WLCG altogether
  - Even a backfill of leftover cycles (no dedicated allocation) is extremely interesting for us
- Need to adapt our systems to be able to fully exploit these offers

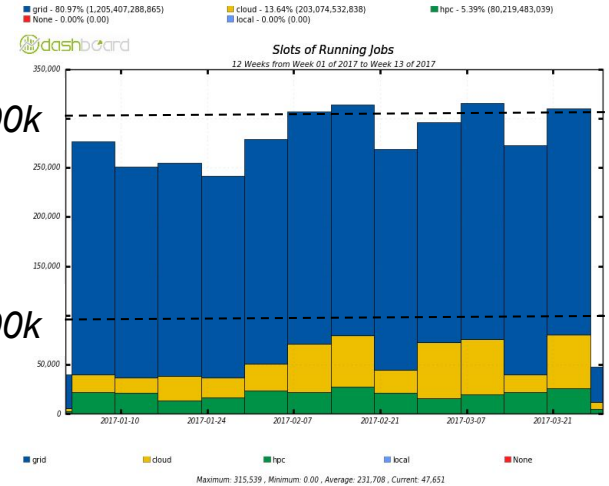


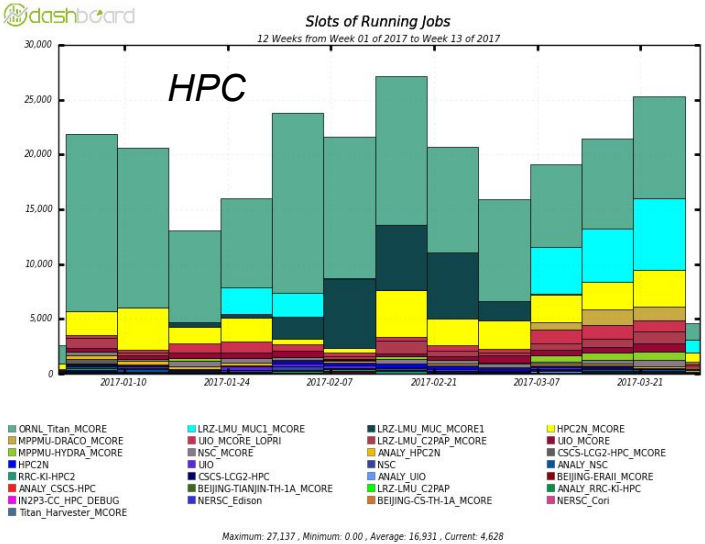
Google Compute Engine

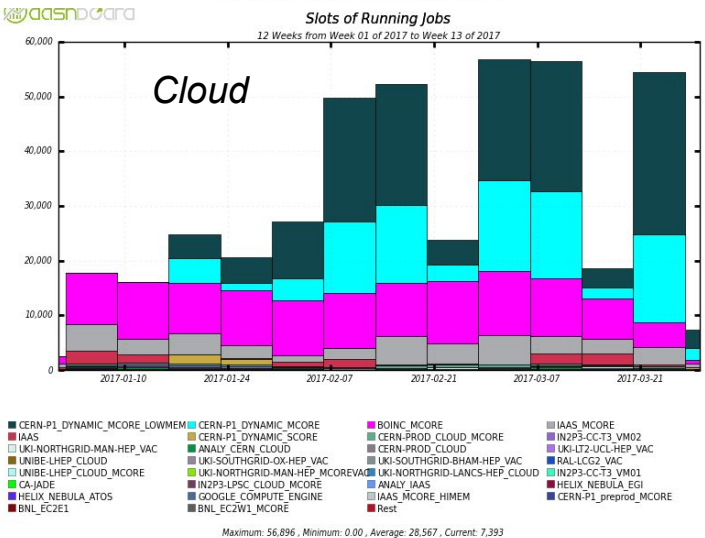
Opportunistic resources



Grid vs cloud vs HPC





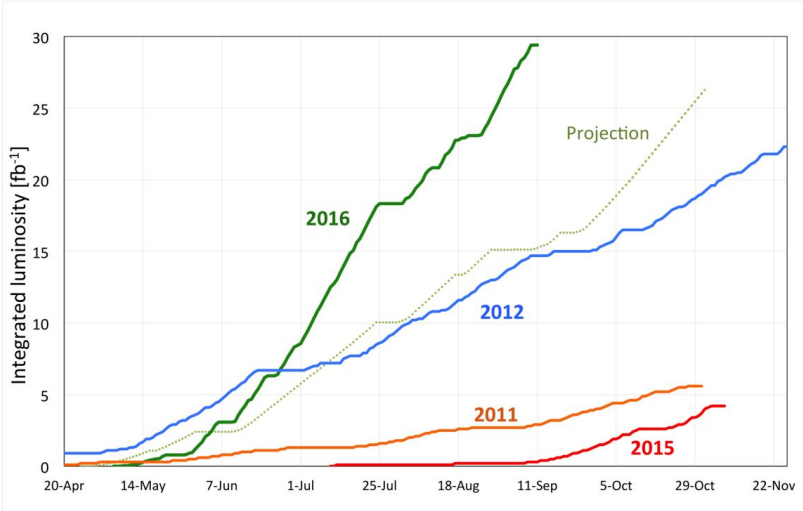


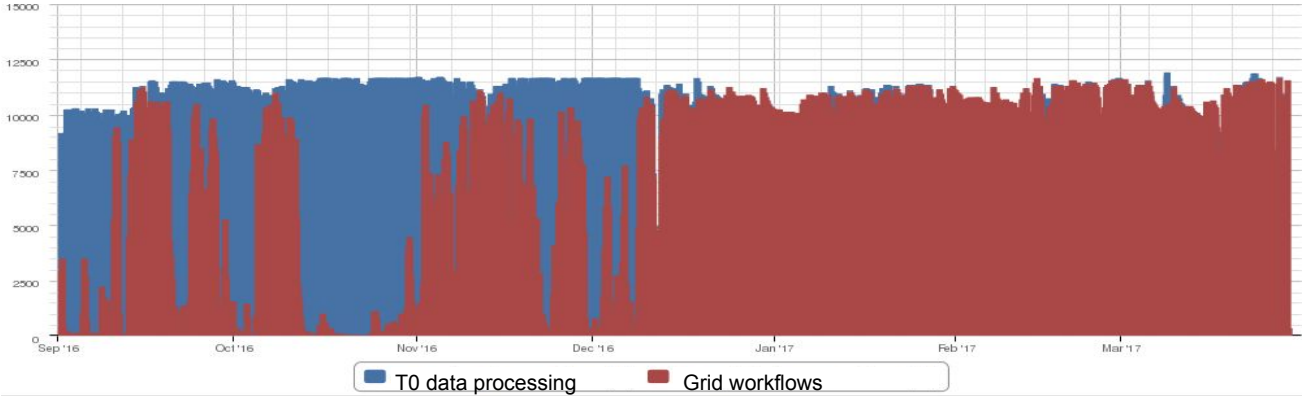
Major HPC contributor is Titan running on purely backfill mode. Constraints on tasks it can run and still a lot of backfill to exploit further

Beautiful example of how online farm is re-configured to run Grid jobs when idle. Also important, steady contribution from ATLAS@Home

Tier-0 processing

- Tier-0 facility is a powerful cluster designed to cope with the data processing needs
  - Powerful worker nodes: SSD, 4GB/core
- Switch from T0 data processing to grid workflows during periods without data

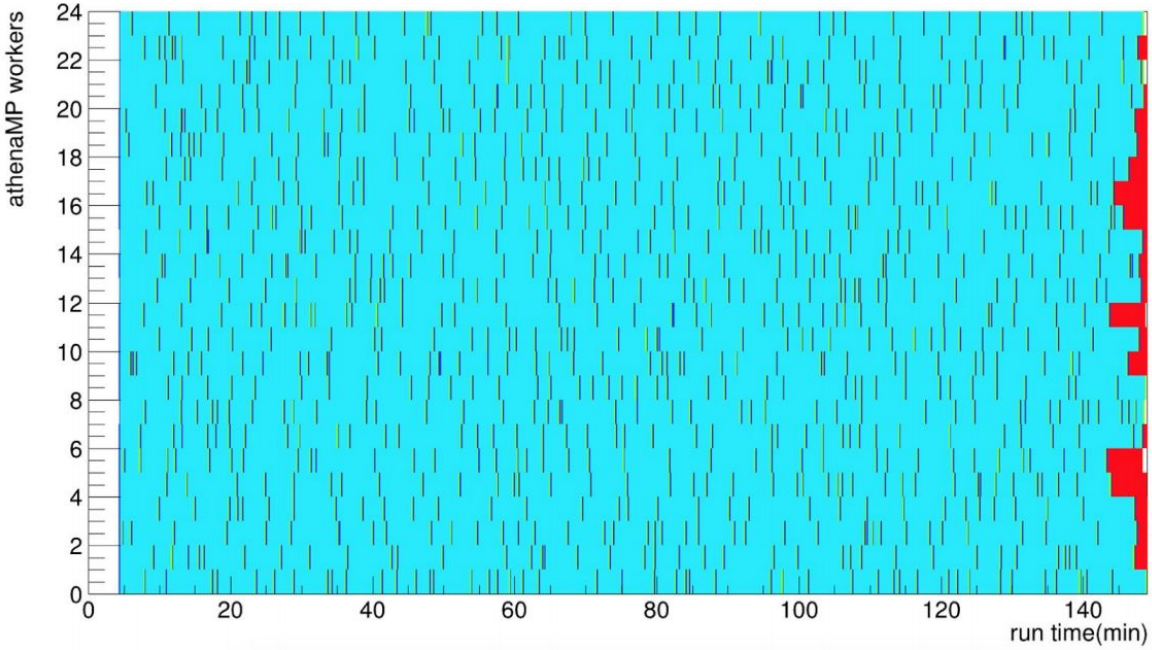




Event Service

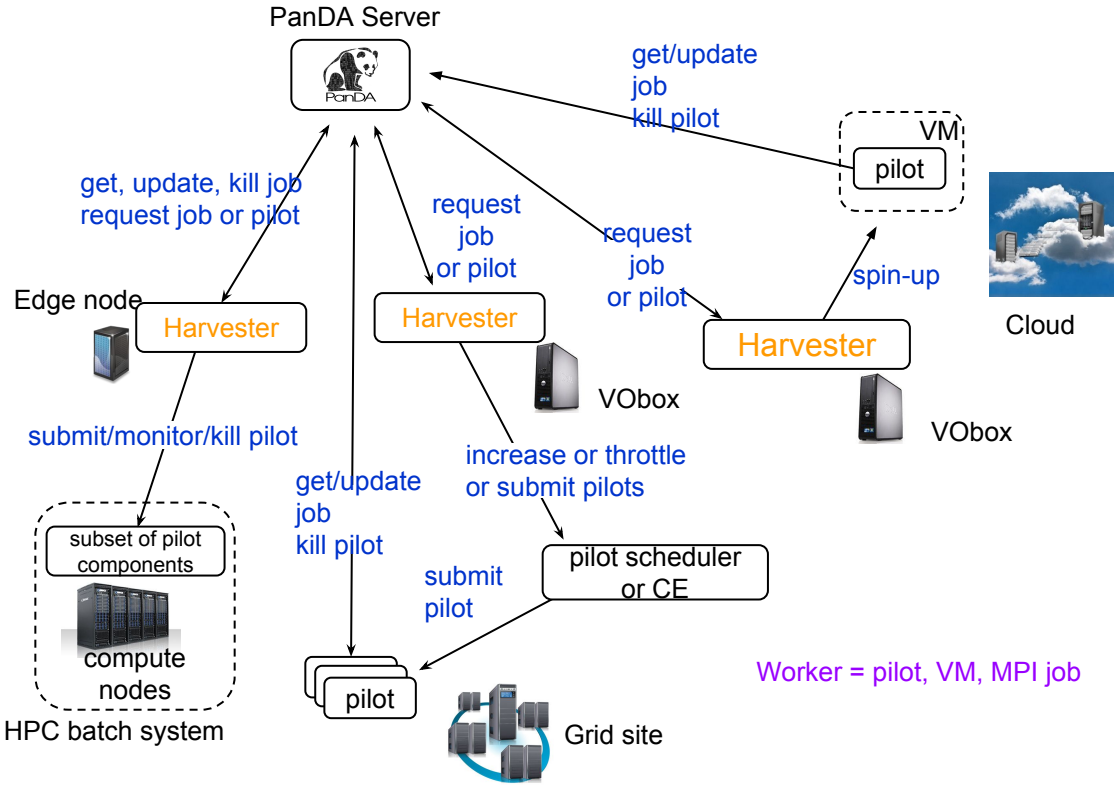
Example: optimized NERSC Edison utilization with Event Service/Yoda

- One AthenaMP job drives 24 workers, one per core
- Optimized initialization time down to ~3 minutes (white)
- Yoda feeds events to workers until the batch slot is exhausted
- Blue = productive event processing time
- Only the last incomplete event is discarded (red) when the slot terminates



Ongoing effort: Harvester

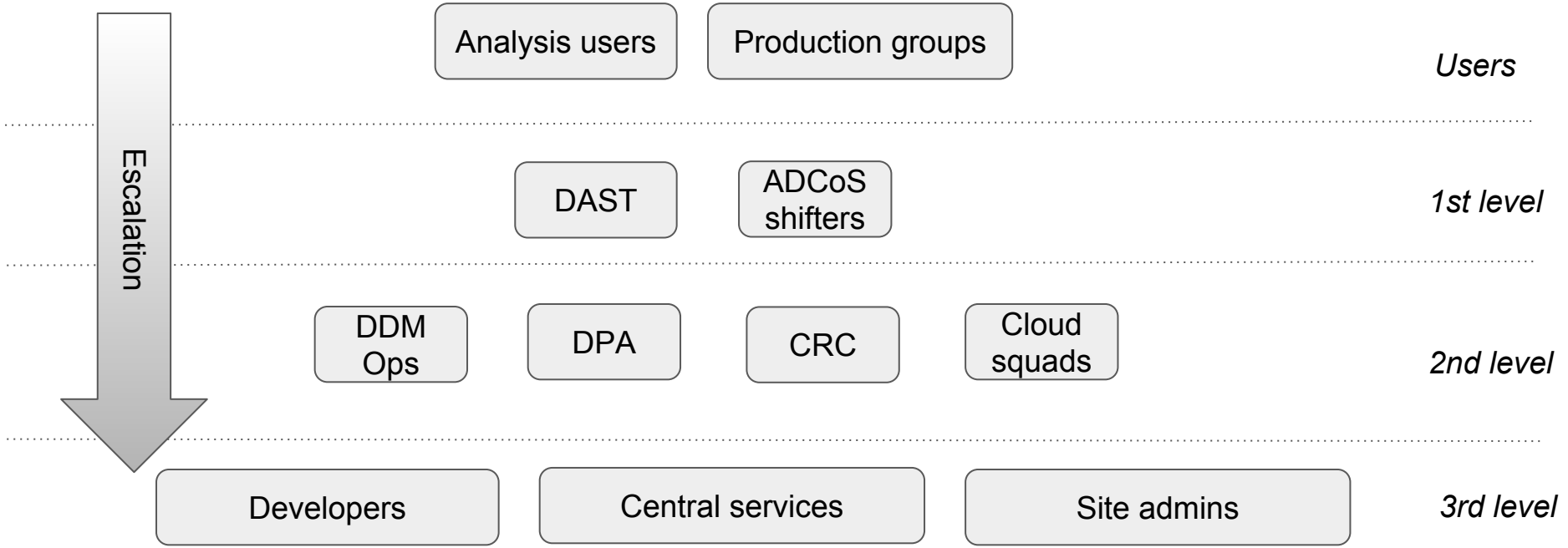
- Workload management components were adapted “ad-hoc” to the different types of resources, as we were getting familiar with them
- Harvester targets to have a common machinery for all computing resources and provide a commonality layer in bringing coherence to HPC implementations



ATLAS Distributed Computing: operations and support

ADC support & ops model

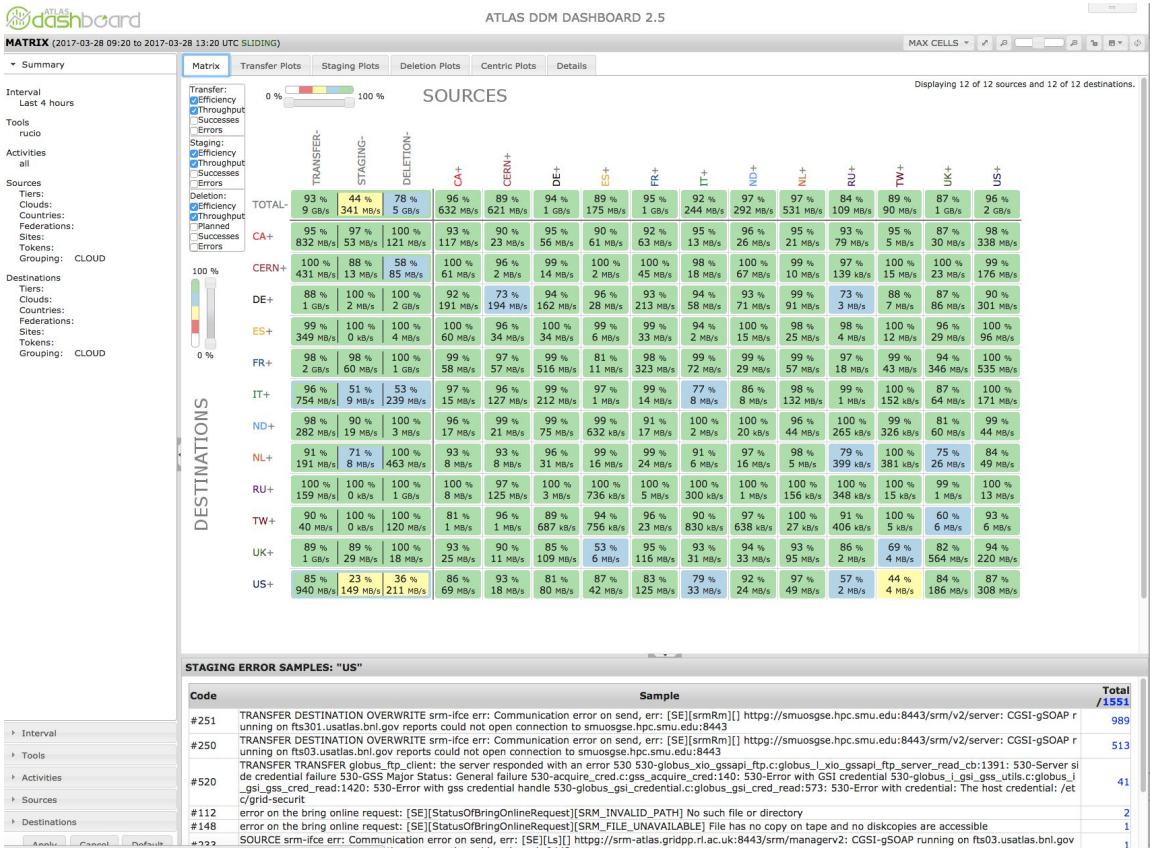
DAST: Distributed Analysis Support Team
ADCoS: ATLAS Distributed Computing operations Shift
DPA: Distributed Production and Analysis (~WM Ops)
CRC: Computing Run Coordinator



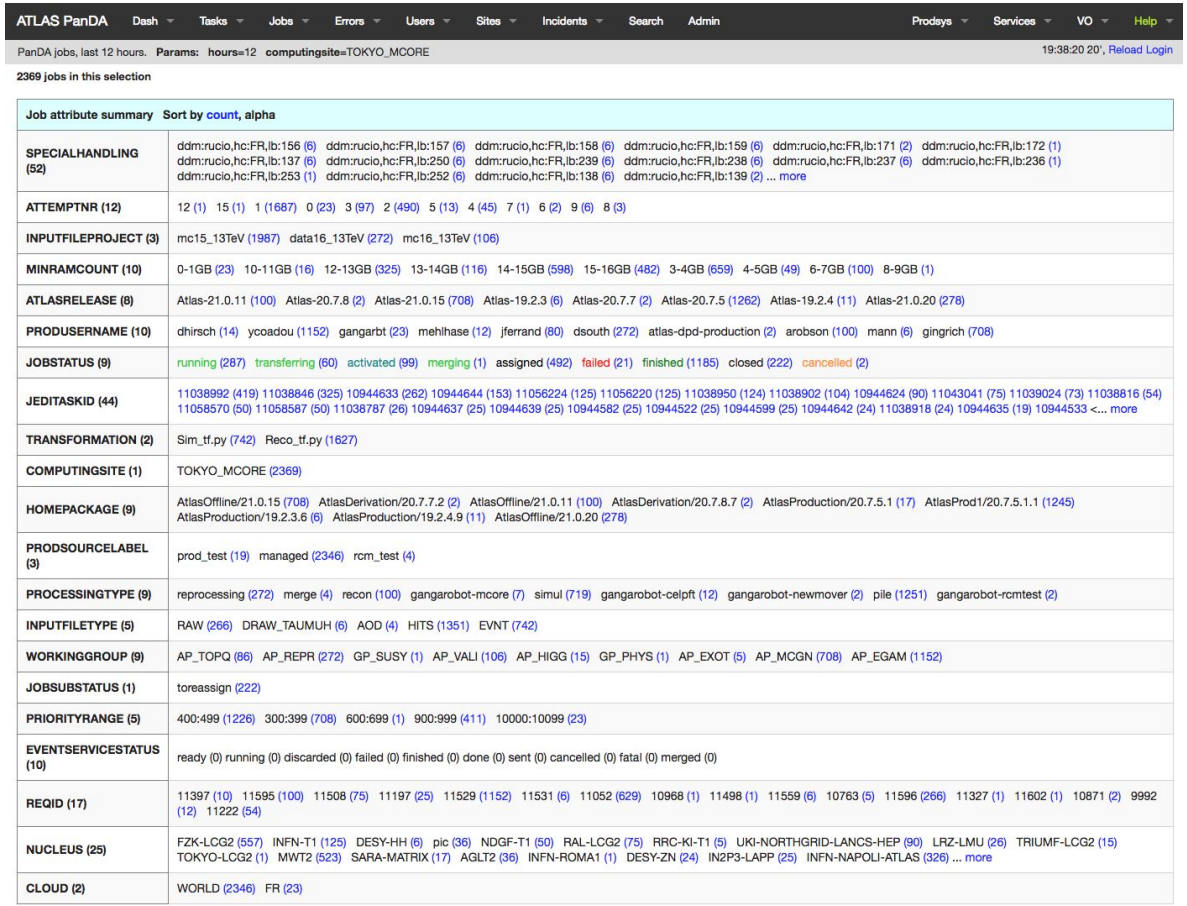
ADC shifts

- Distributed Analysis Shift Team (DAST)
  - Shifts cover EU and US time zones
  - First point of contact to address analysis questions
  - Escalate questions/issues to experts
- ATLAS Distributed Computing Operations Shifts (ADCoS)
  - 24/7 follow the sun, not presential shifts
  - Follow failing jobs and transfers, service degradations, etc.
  - Report to sites/clouds or escalate to CRC/experts
- Computing Run Coordinator (CRC)
  - Shifts are 1 week long and presential at CERN. “Stand by”
  - Coordinates daily ADC operations
  - Main link within ADC communities and representation in WLCG ops meetings
  - Facilitates communication between ADC shifters (in particular ADCoS) and the ADC experts
  - Requires a certain expertise level

Monitoring: DDM Dashboard



Monitoring: BigPanDA



ATLAS PanDA | Dash | Tasks | Jobs | Errors | Users | Sites | Incidents | Search | Admin | Prodsys | Services | VO | Help

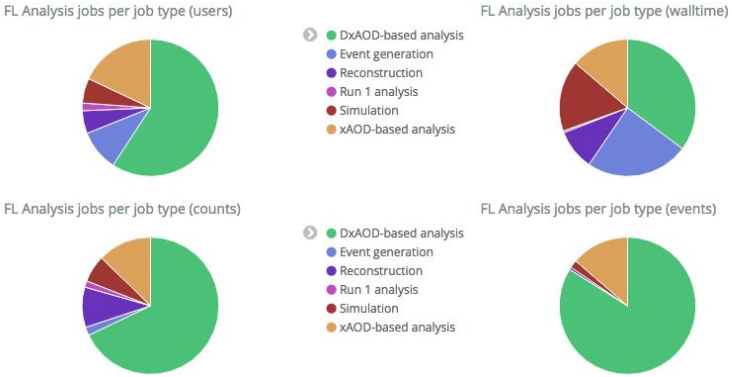
PanDA jobs, last 12 hours. **Params:** **hours**=12 **computingsite**=TOKYO_MCORE — 19:38:20 20', Reload Login

2369 jobs in this selection

Job attribute summary	Sort by count, alpha
SPECIALHANDLING (52)	ddm:rucio,hc:FR,lb:156 (6) ddm:rucio,hc:FR,lb:157 (6) ddm:rucio,hc:FR,lb:158 (6) ddm:rucio,hc:FR,lb:159 (6) ddm:rucio,hc:FR,lb:171 (2) ddm:rucio,hc:FR,lb:172 (1) ddm:rucio,hc:FR,lb:137 (6) ddm:rucio,hc:FR,lb:250 (6) ddm:rucio,hc:FR,lb:239 (6) ddm:rucio,hc:FR,lb:238 (6) ddm:rucio,hc:FR,lb:237 (6) ddm:rucio,hc:FR,lb:236 (1) ddm:rucio,hc:FR,lb:253 (1) ddm:rucio,hc:FR,lb:252 (6) ddm:rucio,hc:FR,lb:138 (6) ddm:rucio,hc:FR,lb:139 (2) ... more
ATTEMPTNR (12)	12 (1) 15 (1) 1 (1687) 0 (23) 3 (97) 2 (490) 5 (13) 4 (45) 7 (1) 6 (2) 9 (6) 8 (3)
INPUTFILEPROJECT (3)	mc15_13TeV (1987) data16_13TeV (272) mc16_13TeV (106)
MINRAMCOUNT (10)	0-1GB (23) 10-11GB (16) 12-13GB (325) 13-14GB (116) 14-15GB (598) 15-16GB (482) 3-4GB (659) 4-5GB (49) 6-7GB (100) 8-9GB (1)
ATLASRELEASE (8)	Atlas-21.0.11 (100) Atlas-20.7.8 (2) Atlas-21.0.15 (708) Atlas-19.2.3 (6) Atlas-20.7.7 (2) Atlas-20.7.5 (1262) Atlas-19.2.4 (11) Atlas-21.0.20 (278)
PRODUSERNAME (10)	dhirsch (14) ycoadou (1152) gangarbt (23) mehlhase (12) jferrand (80) dsouth (272) atlas-dpd-production (2) arobson (100) mann (6) gingrich (708)
JOBSTATUS (9)	running (287) transferring (60) activated (99) merging (1) assigned (492) failed (21) finished (1185) closed (222) cancelled (2)
JEDITASKID (44)	11038992 (419) 11038846 (325) 10944633 (262) 10944644 (153) 11056224 (125) 11056220 (125) 11038950 (124) 11038902 (104) 10944624 (90) 11043041 (75) 11039024 (73) 11038816 (54) 11058570 (50) 11058587 (50) 11038787 (26) 10944637 (25) 10944639 (25) 10944582 (25) 10944522 (25) 10944599 (25) 10944642 (24) 11038918 (24) 10944635 (19) 10944533 <... more
TRANSFORMATION (2)	Sim_tf.py (742) Reco_tf.py (1627)
COMPUTINGSITE (1)	TOKYO_MCORE (2369)
HOMEPACKAGE (9)	AtlasOffline/21.0.15 (708) AtlasDerivation/20.7.7.2 (2) AtlasOffline/21.0.11 (100) AtlasDerivation/20.7.8.7 (2) AtlasProduction/20.7.5.1 (17) AtlasProd1/20.7.5.1.1 (1245) AtlasProduction/19.2.3.6 (6) AtlasProduction/19.2.4.9 (11) AtlasOffline/21.0.20 (278)
PRODSOURCELABEL (3)	prod_test (19) managed (2346) rcm_test (4)
PROCESSINGTYPE (9)	reprocessing (272) merge (4) recon (100) gangarobot-mcore (7) simul (719) gangarobot-celpft (12) gangarobot-newmover (2) pile (1251) gangarobot-rcmtest (2)
INPUTFILETYPE (5)	RAW (266) DRAW_TAUMUH (6) AOD (4) HITS (1351) EVNT (742)
WORKINGGROUP (9)	AP_TOPQ (86) AP_REPR (272) GP_SUSY (1) AP_VALI (106) AP_HIGG (15) GP_PHYS (1) AP_EXOT (5) AP_MCGN (708) AP_EGAM (1152)
JOBSUBSTATUS (1)	toreassign (222)
PRIORITYRANGE (5)	400:499 (1226) 300:399 (708) 600:699 (1) 900:999 (411) 10000:10099 (23)
EVENTSERVICESTATUS (10)	ready (0) running (0) discarded (0) failed (0) finished (0) done (0) sent (0) cancelled (0) fatal (0) merged (0)
REQID (17)	11397 (10) 11595 (100) 11508 (75) 11197 (25) 11529 (1152) 11531 (6) 11052 (629) 10968 (1) 11498 (1) 11559 (6) 10763 (5) 11596 (266) 11327 (1) 11602 (1) 10871 (2) 9992 (12) 11222 (54)
NUCLEUS (25)	FZK-LCG2 (557) INFN-T1 (125) DESY-HH (6) pic (36) NDGF-T1 (50) RAL-LCG2 (75) RRC-KI-T1 (5) UKI-NORTHGRID-LANCS-HEP (90) LRZ-LMU (26) TRIUMF-LCG2 (15) TOKYO-LCG2 (1) MWT2 (523) SARA-MATRIX (17) AGLT2 (36) INFN-ROMA1 (1) DESY-ZN (24) IN2P3-LAPP (25) INFN-NAPOLI-ATLAS (326) ... more
CLOUD (2)	WORLD (2346) FR (23)

Analytics

- Traces and Job data is streamed to ElasticSearch
  - Facilitates analytics and easy aggregation and filters
- Example: Identify incoherent user behaviour, such as individual users running own MC production or occupying non-negligible amounts of resources, can be easily identified



produsername: Descending	Sum of Walltime times core	Count	Sum of inputfilebytes	Sum of outputfilebytes
[redacted]	109 years	89,555	4.6TB	5.7TB
[redacted]	81 years	3,109,997	6.6PB	14.6TB
[redacted]	44 years	1,258,790	6TB	518.6GB
[redacted]	31 years	156,144	458.9TB	5TB
[redacted]	25 years	9,907	24.1TB	242.6GB
[redacted]	23 years	47,278	20.8TB	4TB
[redacted]	19 years	603,890	1.7PB	2TB
[redacted]	19 years	160,732	431TB	3.7TB
[redacted]	17 years	148,551	84.6TB	611.9GB
[redacted]	16 years	452,578	701.4TB	6.7TB

Wrapping up

Conclusions

- 2016 was a very successful year for ATLAS: more data recorded than anticipated
  - ATLAS Distributed Computing was challenged, but proved successful
  - Components are heavily automated and resilient
  - Components present no scaling issues
- ATLAS Computing Model adapted to increasing resource constraints
  - Minimalistic data pre-placement, relying on dynamic transfers and deletions according to usage patterns
  - Dynamic job generation and optimization of resource usage
  - Dependence on optimal exploitation of opportunistic compute resources
  - Software moving in coherent direction, optimizing CPU and memory consumption
- HL-LHC era will be far more intense and we need to start preparing now! See Simone's presentation for details

Reference material

- ATLAS
  - J. Catmore: From collisions to papers
  - ATLAS Resource Request for 2014 and 2015
- ATLAS Distributed Computing (ADC)
  - T. Wenaus: Computing Overview
  - A. Filipcic: ATLAS Distributed Computing Experience and Performance During the LHC Run-2
  - C. Serfon: ATLAS Distributed Computing
- ADC Data Management
  - V. Garonne: Experiences with the new ATLAS Distributed Data Management System
- ADC Workload Management
  - T. Maeno: The Future of PanDA in ATLAS Distributed Computing
  - T. Maeno: Harvester
- ADC Operations and Support
  - C. Adams: Computing shifts to monitor ATLAS Distributed Computing infrastru operations