



# US ATLAS Computing Operations

**Kaushik De**

**University of Texas At Arlington**

**U.S. ATLAS Tier 2/Tier 3 Workshop, FNAL**

**March 8, 2010**

# Overview

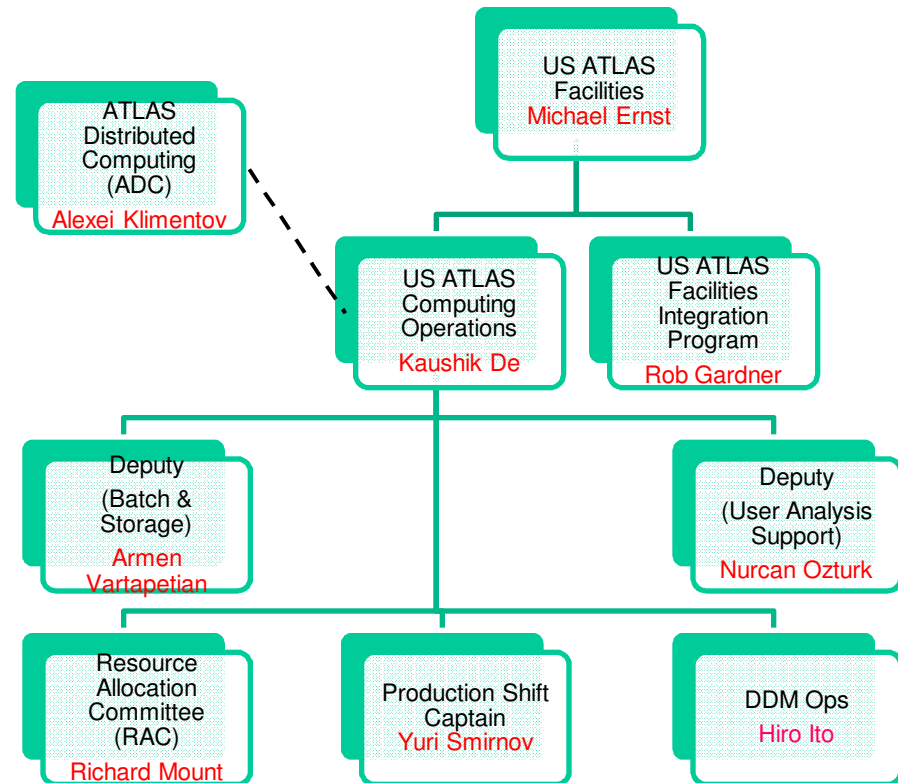


- We expect LHC collisions to resume next week
- ATLAS is ready – after 16 years of preparations
- The distributed computing infrastructure must perform
  - US facilities are required to provide about one quarter of ATLAS computing (though historically we have often provided one third)
  - US also responsible for PanDA software used ATLAS wide
  - Smooth US computing operations is critical for ATLAS success
  - We did very well with the short LHC run in 2009
  - We have experienced people and systems in place
  - But the stress on the system will be far greater in 2010-11
  - Focus of computing operations will be timely physics results
  - We have to adapt quickly to circumstances, as they arise

# US Distributed Computing Organization



- See Michael Ernst's talk for facilities overview & overall ATLAS plans
- Site integration covered in Rob Gardner's talk
- Operations group started two years ago – work will get a lot more intense now



# Scope of US Computing Operations



- Data production – MC, reprocessing
- Data management – storage allocations, data distribution
- User analysis – site testing, validation
- Distributed computing shift and support teams
- Success of US computing operations depends on site managers – we are fortunate to have excellent site teams

# Integration with ADC



- U.S. Operations Team works closely with (and parallels) the ATLAS Distributed Computing team (ADC)
  - Overall goals and plans are synchronized
  - US ATLAS is fully integrated with and well represented in ADC
  - Jim Shank – deputy Computing Coordinator for ATLAS
  - Alexei Klimentov – ADC coordinator
  - KD, Pavel Nevski – processing and shifts coordinator
  - Nurcan Ozturk – Distributed Analysis support team co-coordinator
  - Hiro Ito and Charles Waldman – DDM support and tools
  - Torre Wenaus – PanDA software coordinator
  - ADC team reorganization for 2010-2011 still ongoing

# US Resource Allocation Committee



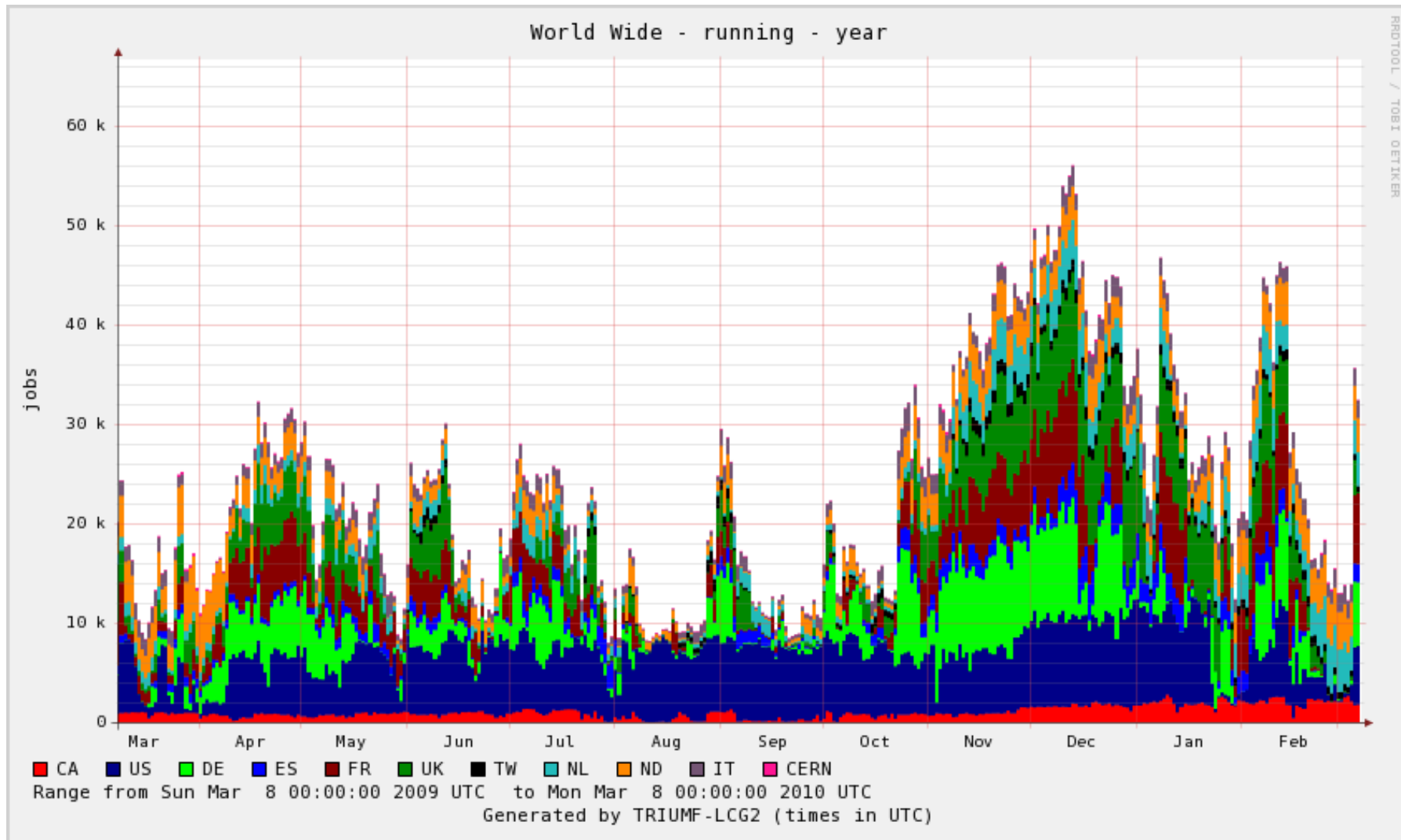
- RAC is becoming increasingly important as LHC starts
- Under new management and very active
- <https://www.usatlas.bnl.gov/twiki/bin/view/AtlasSoftware/ResourceAllocationCommittee>
- Core team meets weekly, full team meets monthly
- **Members**
  - R. Mount (RAC Coordinator), K. Black, J. Cochran, K. De, M. Ernst, R. Gardner, I. Hinchliffe, A. Klimentov, A. Vartapetian, R. Yoshida
  - Ex-officio : H. Gordon, M. Tuts, T. Wenaus, S. Willocq

# MC Production and Reprocessing



- Cornerstone of computing operations
- Experienced team of more than 6 years in the US
- Responsible for:
  - ❑ Efficient utilization of resources at Tier 1/2 sites
  - ❑ Monitor site and task status 7 days a week – site online/offline
  - ❑ Monitor data flow
  - ❑ Report software validation issues
  - ❑ Report task, and distributed software issues
- Part of ADCoS shift team:
  - ❑ **US Team:** Yuri Smirnov (Captain), Mark Sosebee, Armen Vartapetian, Wensheng Deng, Rupam Das
  - ❑ Coverage is 24/7 by using shifts in 3 different time zones
  - ❑ Added 3 new shifters from central and south America

# US Production





# MC Production Issues



- Computing resources (CPU's) were mostly idle in Jan-Feb 2010, because of lack of central production requests
- Meanwhile, US physicists need more MC simulations
- In the future, we will back-fill resources with US regional production, fast-tracking requests from US physicists
  - We have successful experience of large scale US regional production from summer 2009, and from previous years
  - Mechanism for regional production same as central production, hence data can be used for official physics analysis
  - Central production kept all queues busy in second half of 2009, hence regional production dried up – need to restart again

# Reprocessing Issues

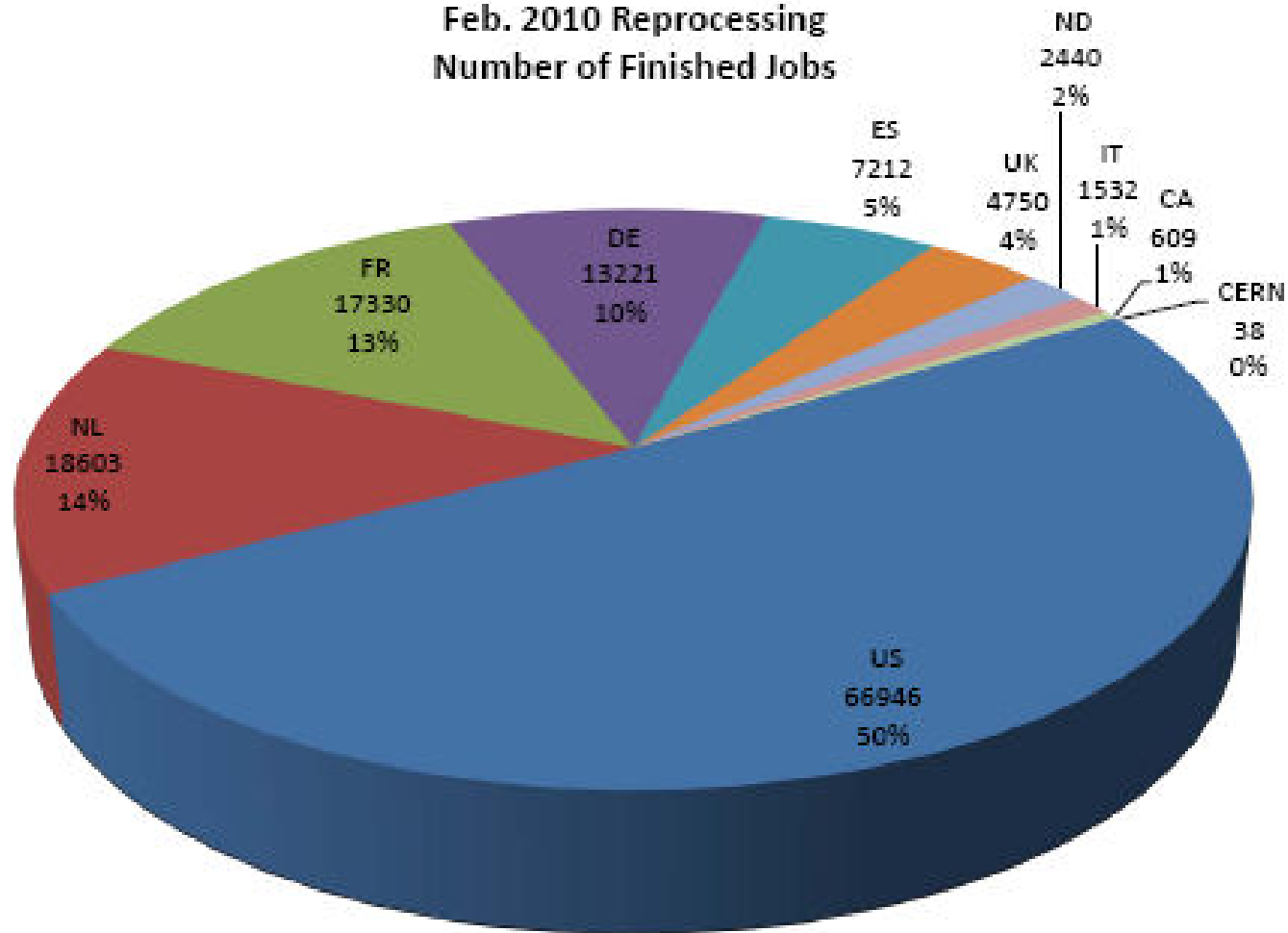


- Tape reprocessing is slow
- We try to keep as much real data on disk as possible
- Due to high I/O requirements, reprocessing is usually done at Tier 1's (ATLAS-wide)
- In the US, the option to run reprocessing at Tier 2's is also available, to speed up physics results
  - Tier 2's go through reprocessing validation
  - We have used US Tier 2's in the past (but not every time)
- Reprocessing will become increasingly challenging as we accumulate LHC data – US is doing well

# Reprocessing Experience



Feb. 2010 Reprocessing  
Number of Finished Jobs



# Storage Management



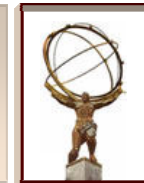
- Many tools used for data placement, data cleaning, data consistency, storage system management...
- Some useful links:
  - ❑ <http://panda.cern.ch:25980/server/pandamon/query>
  - ❑ [http://panda.cern.ch:25980/server/pandamon/query?mode=ddm\\_req](http://panda.cern.ch:25980/server/pandamon/query?mode=ddm_req)
  - ❑ [http://atlddm02.cern.ch/dq2/accounting/cloud\\_view/USASITES/30/](http://atlddm02.cern.ch/dq2/accounting/cloud_view/USASITES/30/)
  - ❑ <http://dashb-atlas-data-test.cern.ch/dashboard/request.py/site>
- Also needs many people
  - ❑ ADC Operations: led by Stephane Jezequel
  - ❑ DQ2 team: led by Simone Campana
  - ❑ ATLAS DDM operations: led by Ikuo Ueda
  - ❑ US team: KD (US Operations), Hironori Ito (US DDM), Armen Vartapetian (US Ops Deputy), Wensheng Deng (US DDM), Pedro Salgado (BNL dcache), plus Tier 2 support team – weekly meetings

# Storage Issues



- US will have ~10 PB within the next few months
  - Already have ~7.5 PB deployed
  - Fast ramp-up needed
- Storage is partitioned into space tokens by groups
  - Good – keeps data types isolated (user, MC, LHC...)
  - Bad – fragments storage, which is wasteful and difficult to manage
- Space token management
  - Each site must provide 6-10 different storage partitions (tokens)
  - This is quite labor intensive – ADC trying to automate
  - Many tools for efficient space management are still missing
  - Good management of space tokens is essential to physics analysis

# Primer on Tokens



- DATADISK – ESD (full copy at BNL, some versions at T2's), RAW (BNL only), AOD (four copies among U.S. sites)
- MCDISK – AOD's (four copies among U.S. sites), ESD's (full copy at BNL), centrally produced DPD's (all sites), some HITs/RDOs
- DATATAPE/MCTAPE – archival data at BNL
- USERDISK – pathena output, limited lifetime (variable, at least 60 days, users notified before deletion)
- SCRATCHDISK – Ganga output, temporary user datasets, limited lifetime (maximum 30 days, no notification before deletion)
- GROUPLISK – physics/performance group data
- LOCALGROUPLISK – storage for local (geographically) groups/users
- PRODDISK – only used by Panda production at Tier 2 sites
- HOTDISK – database releases

# Storage Tokens Status



Site	Used(GB)	Free(GB)	Total(GB)	USED(%)
BNL-OSG2_DATADISK	868,000	649,000	1,517,000	57
BNL-OSG2_MCDISK	1,599,000	198,000	1,797,000	89
AGLT2_DATADISK	92,555	95,295	187,850	49
AGLT2_MCDISK	182,741	67,705	250,446	72
MWT2_DATADISK	<del>24,170</del> <b>64</b>	<del>92,378</del> <b>29</b>	116,548	<del>20</del> <b>69</b>
MWT2_UC_MCDISK	138,815	5,221	144,036	96
NET2_DATADISK	106,999	21,850	128,849	83
NET2_MCDISK	140,999	11,472	152,471	92
SLACXRD_TOTAL	220,749	208,747	429,496	51
SLACXRD_DATADISK	49,004	-	-	-
SLACXRD_MCDISK	139,555	-	-	-
SWT2_TOTAL	216,838	234,133	450,971	48
SWT2_DATADISK	58,539	-	-	-
SWT2_MCDISK	123,197	-	-	-

From  
A. Vartapetian

# Data Management Issues



- DDM consolidation – see Hiro’s talk
- MCDISK is largest fraction of space usage
  - Not surprising given the delay in LHC schedule
  - Need to prune MCDISK as data arrives
  - US DDM team is studying what can be cleaned up, without affecting the work of physics users
- Some Tier 2’s are very full
  - New resources are coming soon
  - We are keeping an eye daily on these sites
- There may be space crunch again in Fall 2010



# Cleaning Up



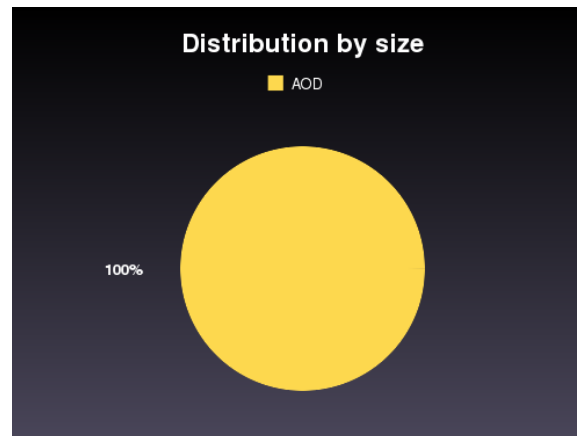
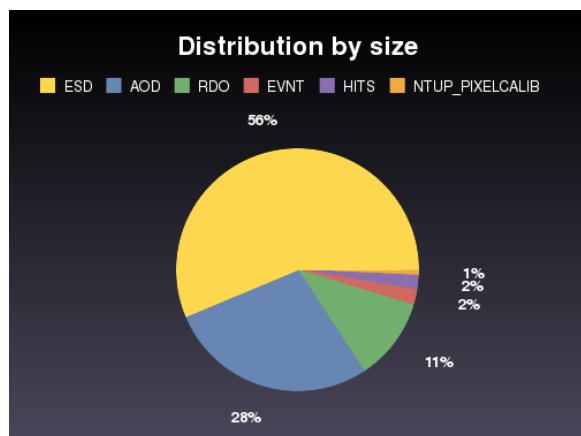
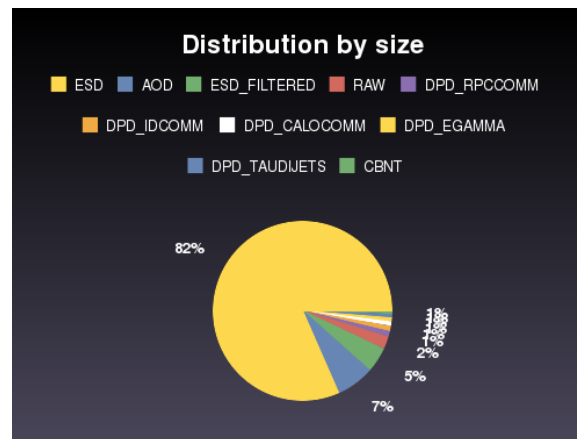
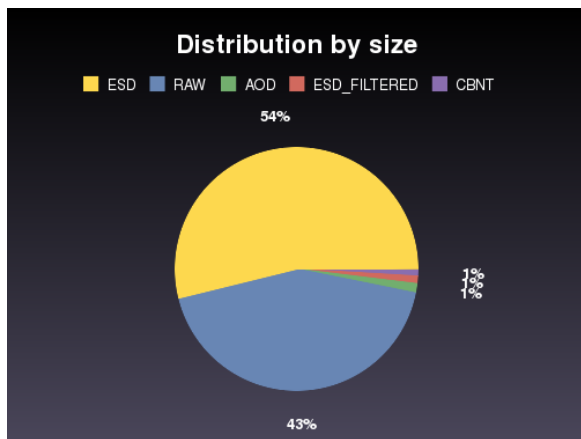
- MCDISK/DATADISK - cleaned by Central Deletion or US DDM
- MCTAPE/DATATAPE - cleaned by Central Deletion, with notification to BNL to clean/reuse tapes
- SCRATCHDISK - cleaned by Central Deletion
- GROUPODISK - cleaned by Central Deletion
- HOTDISK - never cleaned!
- PRODDISK - cleaned by site
- USERDISK - cleaned by US DDM
- LOCALGROUPODISK - cleaned by US DDM

# Some Examples of Data Distribution



## BNL DATADISK/MCDISK

## MWT2 DATADISK, SWT2 MCDISK



# GROUPDISK Endpoints in US



- All Tier 1/Tier 2 sites in the US provide physics/performance endpoints on GROUPDISK
- For example: PERF-MUONS at AGLT2 & NET2, PERF-EGAMMA at SWT2, PERF-JETS at MWT2 & WT2
- Each phys/perf group has designated data manager
- GROUPDISK endpoints host data of common interest
- Data placement can be requested through group manager or DaTRI: [http://panda.cern.ch:25980/server/pandamon/query?mode=ddm\\_req](http://panda.cern.ch:25980/server/pandamon/query?mode=ddm_req)
- For endpoint usage summary see: [http://atlddm02.cern.ch/dq2/accounting/group\\_reports/](http://atlddm02.cern.ch/dq2/accounting/group_reports/)

# US Endpoint Summary

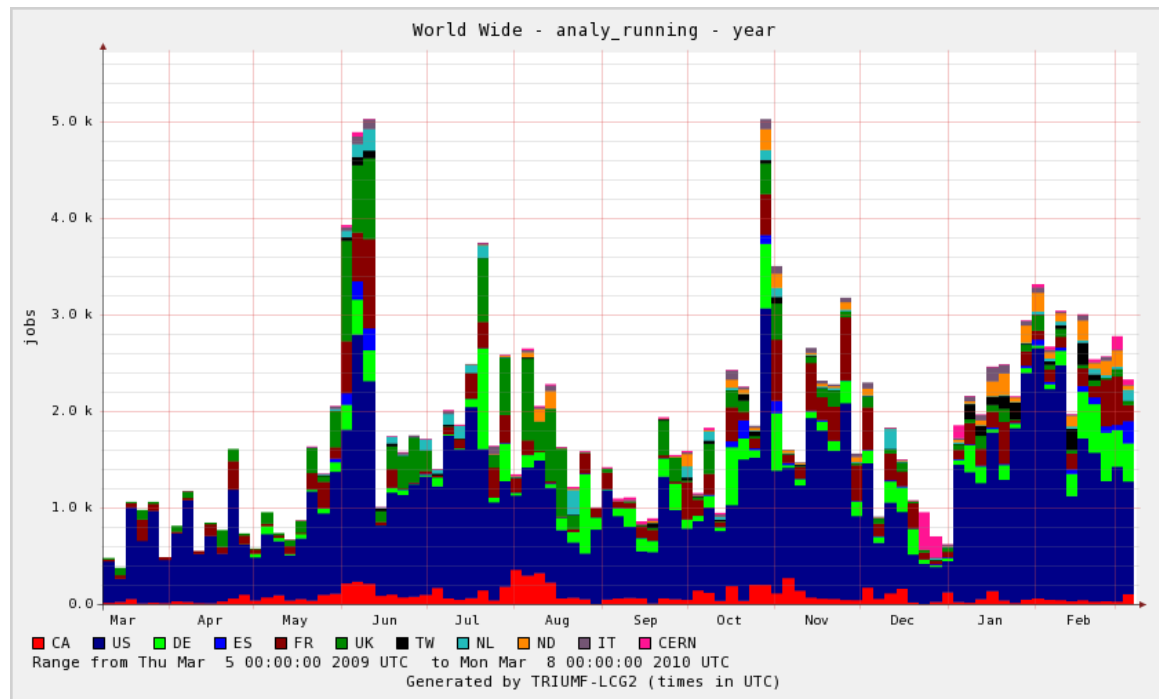


Site	Group	Role	Used by role(TB)	Booked by role(TB)	ΣUsed(TB)	ΣBooked(TB)	DQ2 Used(TB)	SRM Used(TB)
AGLT2	PERF-MUONS	/atlas/perf-muons/role=production	0.0	27.49	0.0	27.49	1.16	1.56
	PHYS-SM	/atlas/phys-sm/role=production	0.0	27.49	0.0	27.49		
	TRIG-DAQ	/atlas/trig-daq/role=production	0.0	27.49	0.0	27.49		
BNL-OSG2	PERF-EGAMMA	/atlas/perf-egamma/role=production	0.0	27.49	0.0	27.49	14.81	16.57
	PERF-FLAVTAG	/atlas/perf-flavtag/role=production	0.35	27.49	0.35	27.49		
	PERF-JETS	/atlas/perf-jets/role=production	0.0	27.49	0.0	27.49		
	PERF-MUONS	/atlas/perf-muons/role=production	0.0	27.49	0.0	27.49		
	PHYS-HI	/atlas/phys-hi/role=production	2.58	1.1	2.58	1.1		
	PHYS-HIGGS	/atlas/phys-higgs/role=production	0.15	5.5	0.15	5.5		
	PHYS-SM	/atlas/phys-sm/role=production	0.05	27.49	0.05	27.49		
	TRIG-DAQ	/atlas/trig-daq/role=production	0.0	27.49	0.0	27.49		
MWT2_UC	PERF-JETS	/atlas/perf-jets/role=production	0.0	27.49	0.0	27.49	2.01	2.2
	PERF-TAU	/atlas/perf-tau/role=production	0.0	27.49	0.0	27.49		
	PHYS-HIGGS	/atlas/phys-higgs/role=production	0.0	27.49	0.0	27.49		
NET2	PERF-MUONS	/atlas/perf-muons/role=production	0.0	27.49	0.0	27.49	3.13	132.05
	PHYS-EXOTICS	/atlas/phys-exotics/role=production	0.0	27.49	0.0	27.49		
	PHYS-TOP	/atlas/phys-top/role=production	0.49	27.49	0.49	27.49		
SLACXRD	PERF-FLAVTAG	/atlas/perf-flavtag/role=production	0.35	27.49	0.35	27.49	11.54	12.88
	PERF-IDTRACKING	/atlas/perf-idtracking/role=production	0.0	27.49	0.0	27.49		
	PERF-JETS	/atlas/perf-jets/role=production	0.0	27.49	0.0	27.49		
	PHYS-BEAUTY	/atlas/phys-beauty/role=production	0.0	27.49	0.0	27.49		
	PHYS-SM	/atlas/phys-sm/role=production	0.0	27.49	0.0	27.49		
	SOFT-SIMUL	/atlas/soft-simul/role=production	0.52	2.2	0.52	2.2		
SWT2_CPB	PERF-EGAMMA	/atlas/perf-egamma/role=production	0.0	27.49	0.0	27.49	1.46	1.61
	PHYS-SUSY	/atlas/phys-susy/role=production	0.0	27.49	0.0	27.49		
	PHYS-TOP	/atlas/phys-top/role=production	0.49	27.49	0.49	27.49		

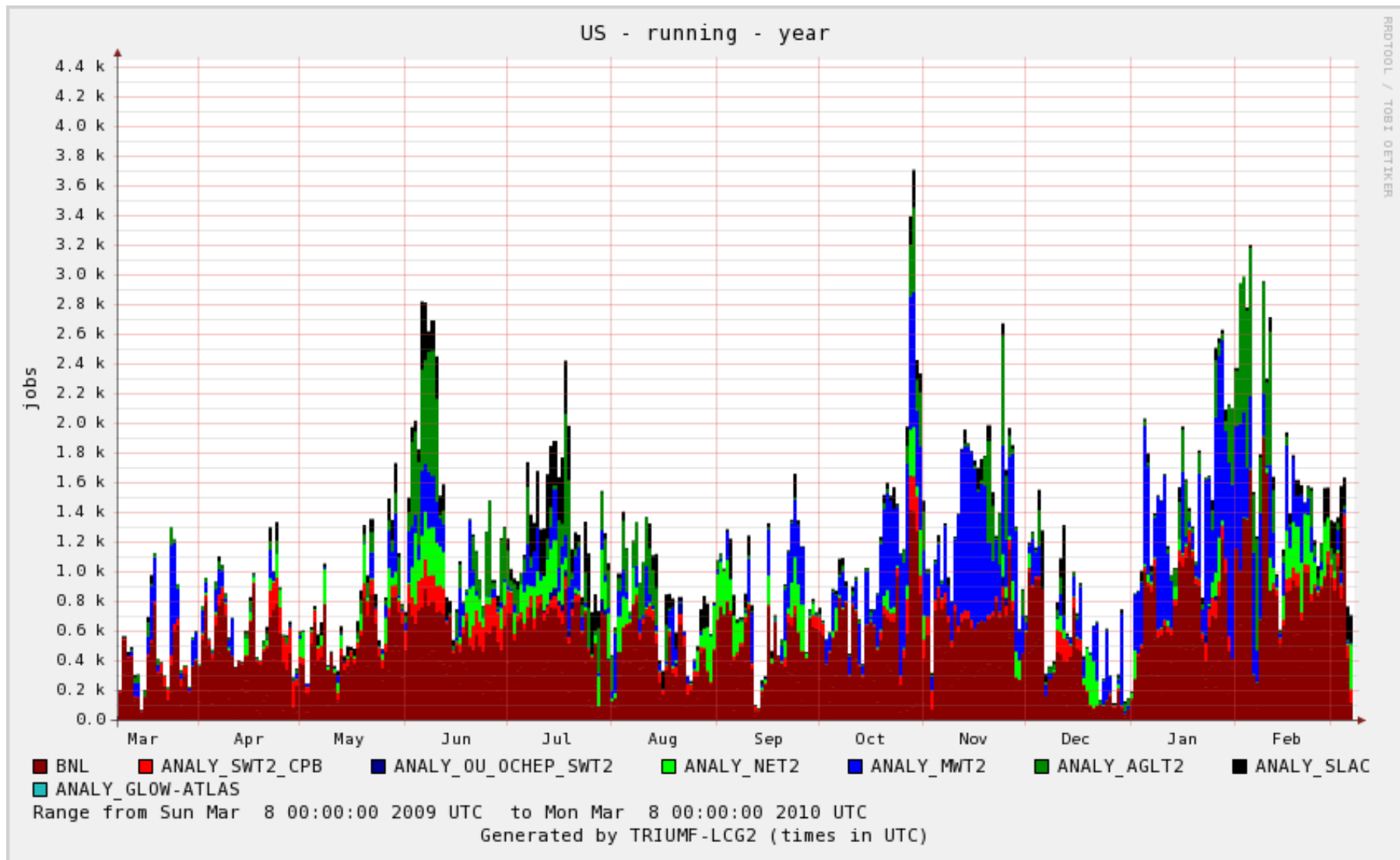
# User Analysis



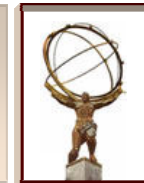
- U.S. ATLAS has excellent track record of supporting users
  - US is the most active cloud in ATLAS for user analysis
  - Analysis sites are in continuous and heavy use for >2 years
  - We have regularly scaled up resources to match user needs
  - Tier 3 issues will be discussed tomorrow



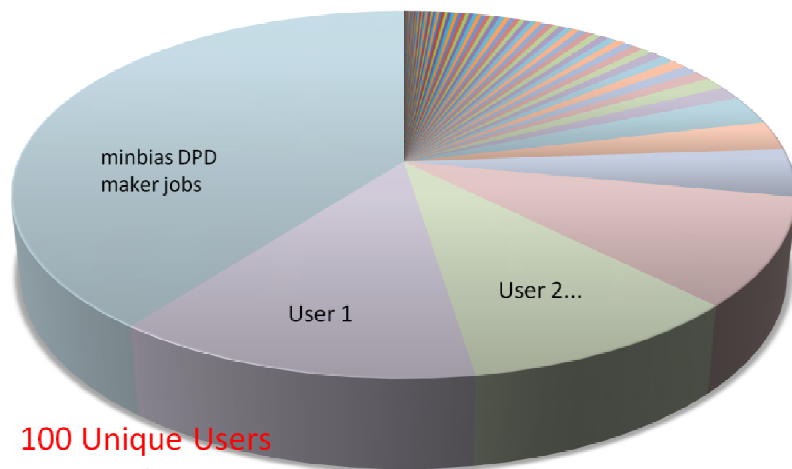
# Analysis Usage Growing



# 900GeV User Analysis Jobs

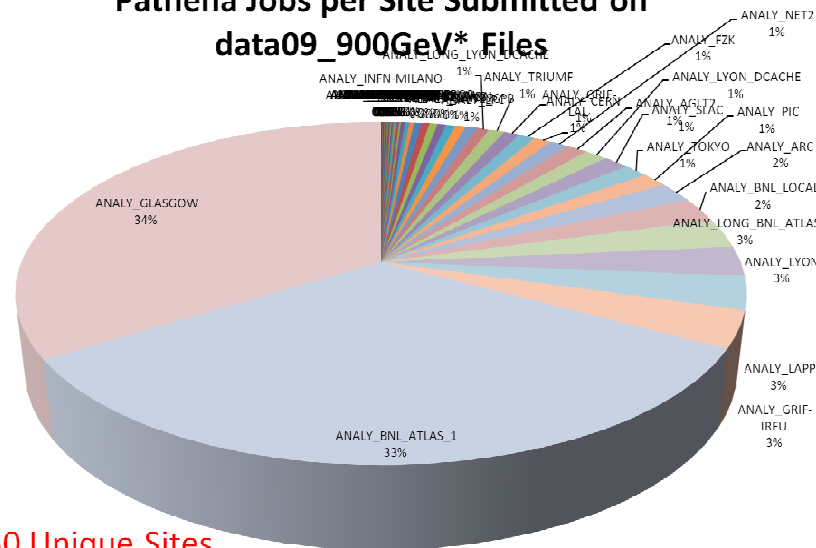


Pathena Jobs per User Submitted on data09\_900GeV\* Files



100 Unique Users  
11768 jobs

Pathena Jobs per Site Submitted on data09\_900GeV\* Files



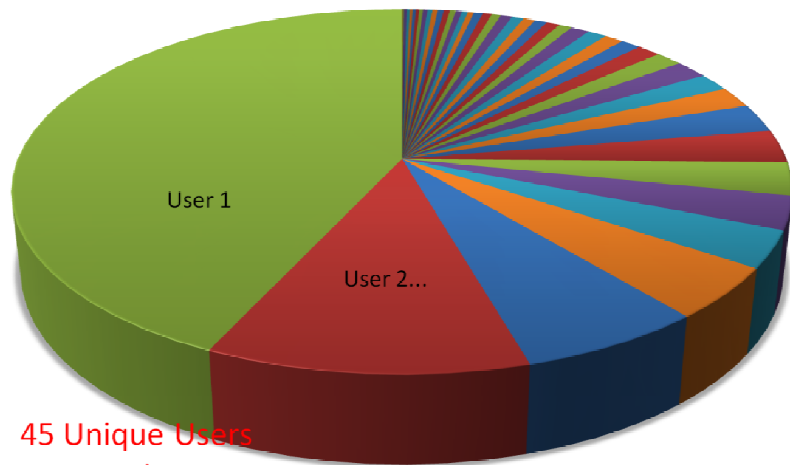
50 Unique Sites

Statistics from November 2009

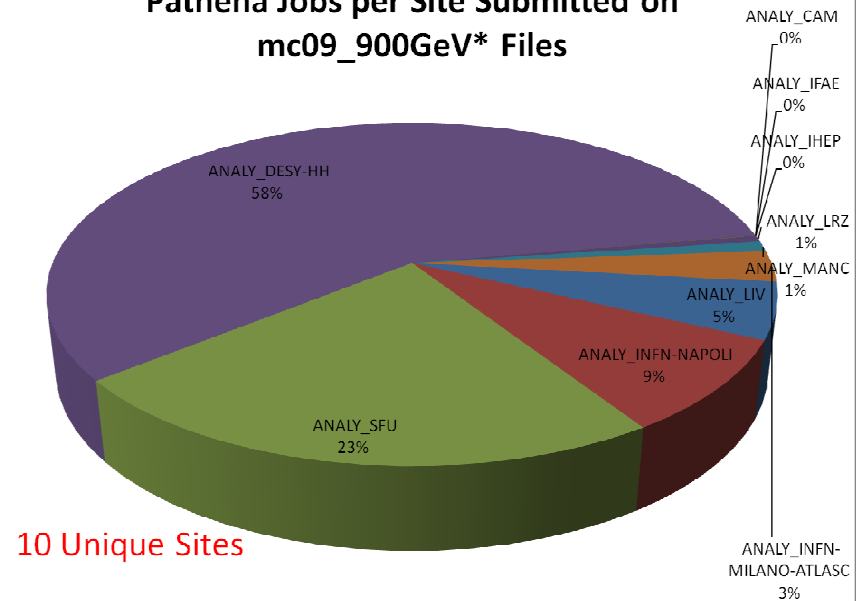
# 900GeV MC User Analysis Jobs



Pathena Jobs per User Submitted on mc09\_900GeV\* Files



Pathena Jobs per Site Submitted on mc09\_900GeV\* Files



Statistics from November 2009



# ADCoS (ADC Operations Shifts)



- ADCoS combined shifts started January 28th, 2008
- ADCoS Goals
  - World-wide (distributed/remote) shifts
  - To monitor all ATLAS distributed computing resources
  - To provide Quality of Service (QoS) for all data processing
- Organization
  - Senior/Trainee: 2 day shifts, Expert: 7 day shifts
  - Three shift times (in CERN time zone):
    - ASIA/Pacific: 0h - 8h
    - EU-ME: 8h - 16h
    - Americas: 16h - 24h
- U.S. shift team
  - In operation long before ADCoS was started

# Distributed Analysis Shift Team – DAST



- User analysis support is provided by the AtlasDAST (Atlas Distributed Analysis Shift Team) since September 29, 2008. Previously, user support was on a best effort basis provided by the Panda and Ganga software developers.
- Nurcan Ozturk (UTA) and Daniel van der Ster (CERN) are coordinating this effort.
- DAST organizes shifts currently in two time zones – US and CERN. One person from each zone is on shift for 7 hours a day covering between 9am-11pm CERN time, and 5 days a week.
- Please contact Nurcan to join this effort

# Conclusion



- Waiting for collisions!