# ATLAS Data Distribution

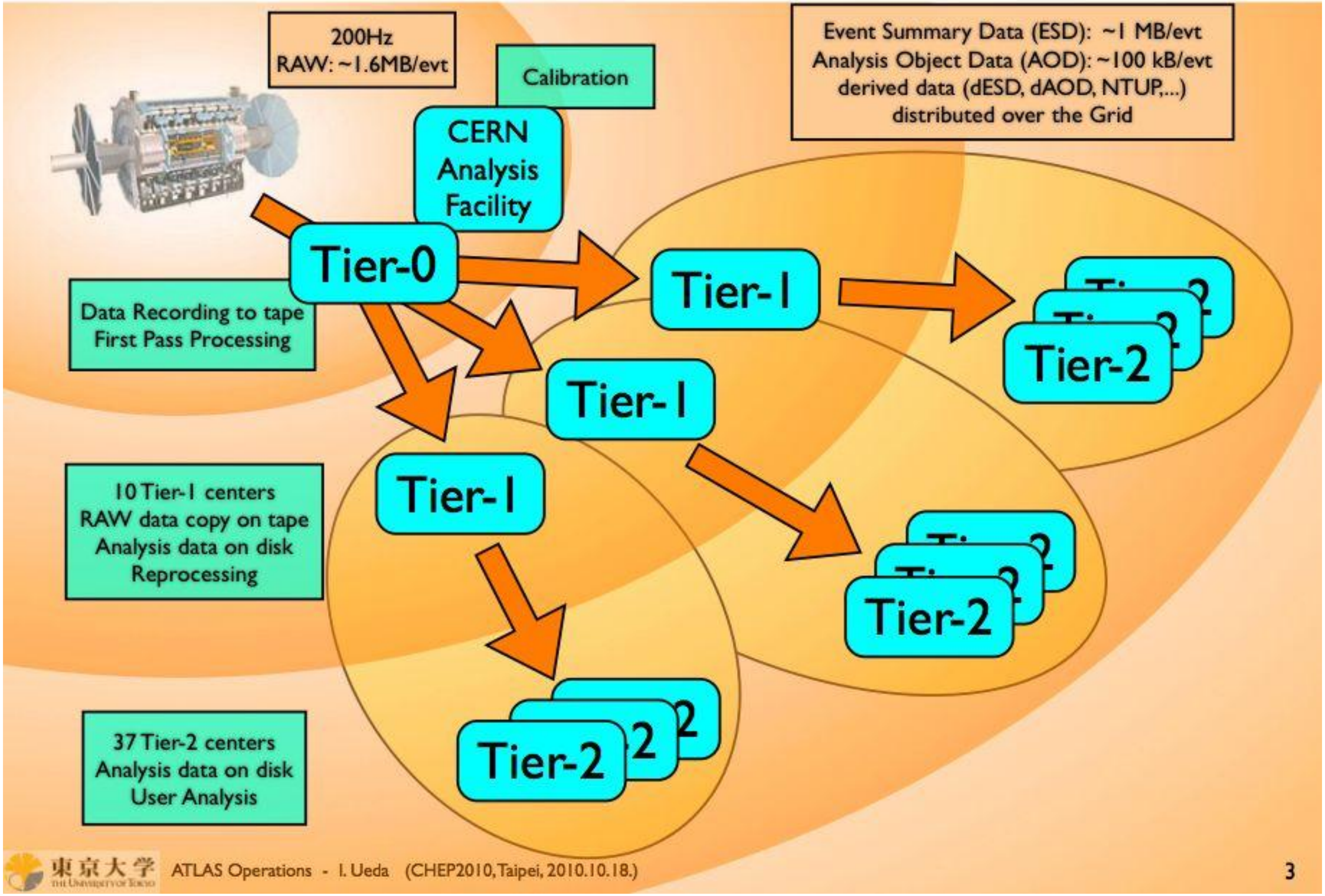
Jim Shank

BOSTON
UNIVERSITY

# Introduction

- Overview of data distribution
  - Slides from
    - I. Ueda CHEP talk last week:
    - http://117.103.105.177/MaKaC/materialDisplay.py?contribId=5&sessionId=104&materialId=slides&confId=3
- Panda Dynamic Data Placement (PD2P)
  - K. De talk last ATLAS week:
    - http://indico.cern.ch/materialDisplay.py?contribId=23&sessionId=13&materialId=slides&confId=66744

# Introduction: ATLAS Data Flow



200Hz
RAW: ~1.6MB/evt

Event Summary Data (ESD): ~1 MB/evt
Analysis Object Data (AOD): ~100 kB/evt
derived data (dESD, dAOD, NTUP,...)
distributed over the Grid

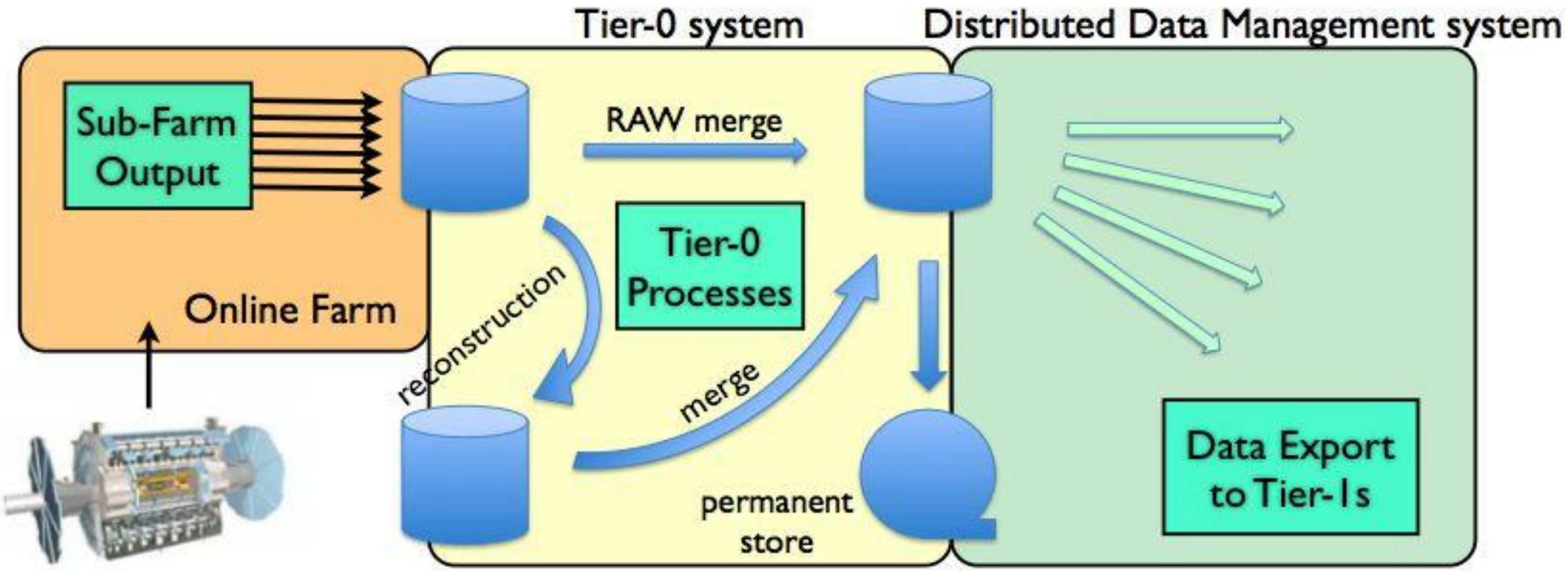
Data Recording to tape
First Pass Processing

10 Tier-1 centers
RAW data copy on tape
Analysis data on disk
Reprocessing

37 Tier-2 centers
Analysis data on disk
User Analysis

東京大学 THE UNIVERSITY OF TOKYO ATLAS Operations - I. Ueda (CHEP2010, Taipei, 2010.10.18.) 3

# Tier-0 Processes



Accepting data from the online system and ensuring it is archived to tape

- Merging small files to adequate size for tape archiving

Processing RAW data (event reconstruction) and archiving the products to tape

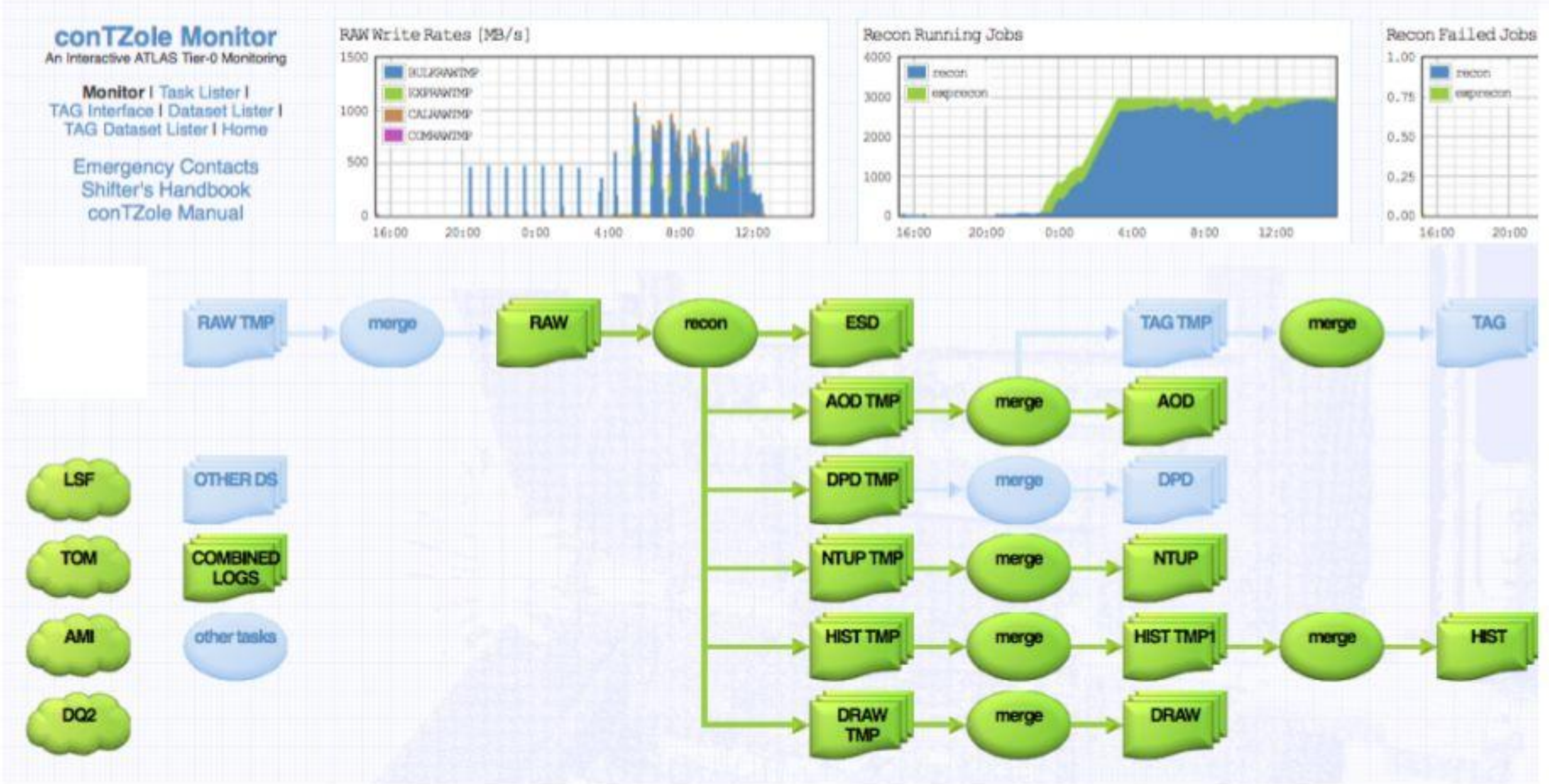
- Express stream for prompt calibration and alignment
- First-pass processing of all streams after 36h with calibration and alignment

Registering data to the ATLAS Distributed Data Management system

- Export data to Tier-1 and calibration Tier-2s, as well as CAF

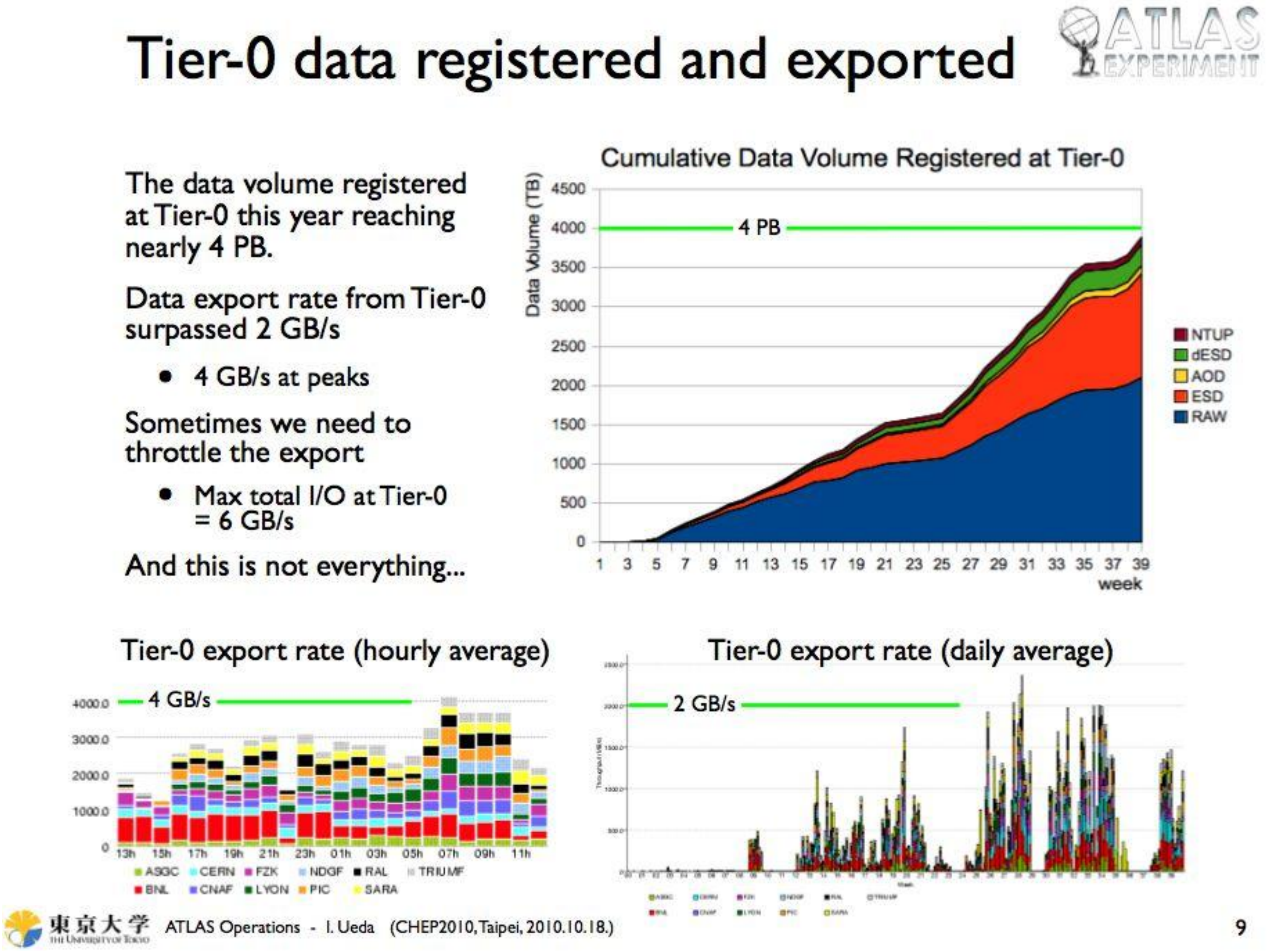
Maximum overall I/O: 6GB/s -- including internal accesses within Tier-0

# Tier-0 Workflow Monitoring



conTZole : The new monitoring system based on Web 2.0 architecture (AJAX)

- primarily aimed to be used by shifters and Tier-0 operations team
- but also useful to any ATLAS members to see how processing of a certain run goes



# Tier-0 data registered and exported

The data volume registered at Tier-0 this year reaching nearly 4 PB.

Data export rate from Tier-0 surpassed 2 GB/s

- 4 GB/s at peaks

Sometimes we need to throttle the export

- Max total I/O at Tier-0 = 6 GB/s

And this is not everything...

Cumulative Data Volume Registered at Tier-0

Tier-0 export rate (hourly average)

Tier-0 export rate (daily average)

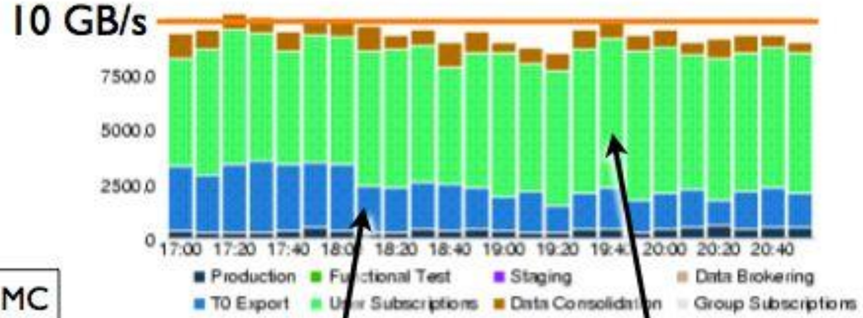
東京大学 THE UNIVERSITY OF TOKYO ATLAS Operations - I. Ueda (CHEP2010, Taipei, 2010.10.18.) 9

# Data Distribution over the Grid

In addition to the Tier-0 export, we have;

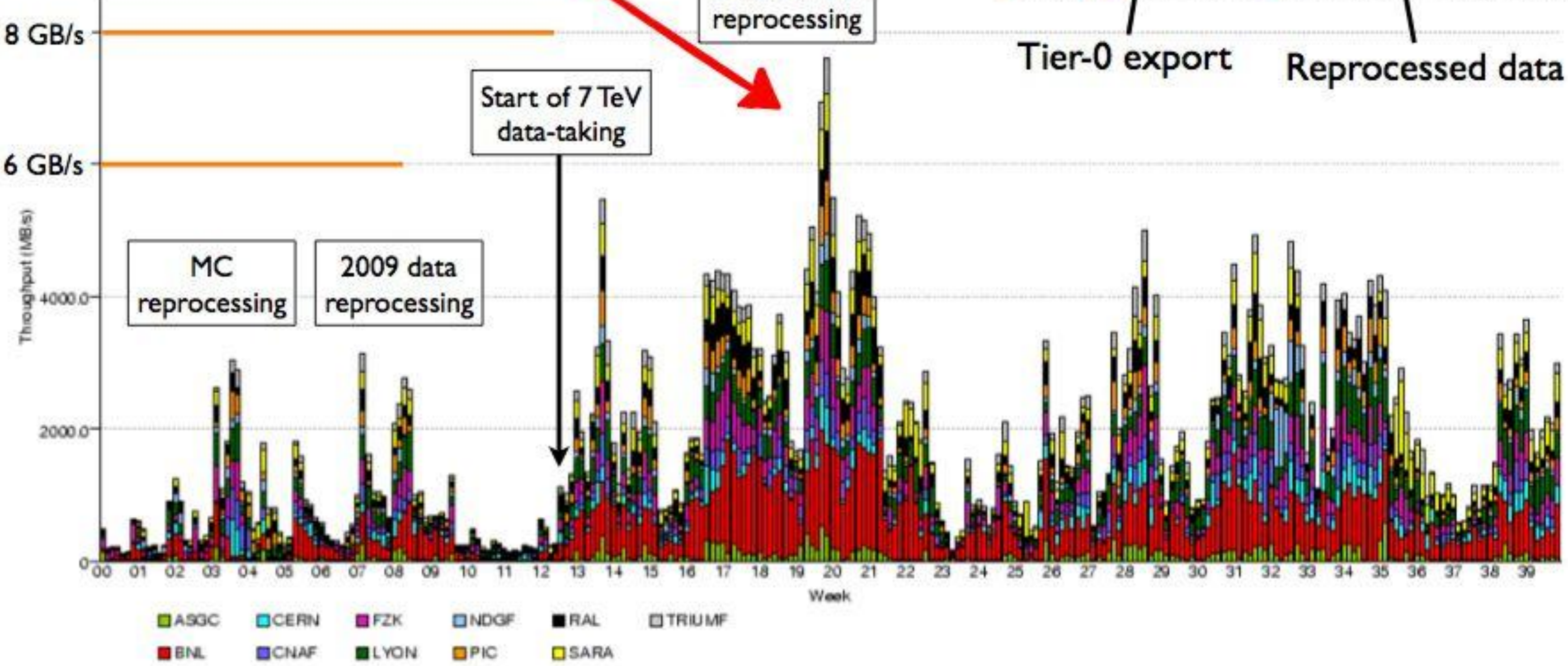
- Data movement for MC data production
- Distribution of reprocessed data

And they came all at once



Tier-0 export

Reprocessed data



Data and MC reprocessing

Start of 7 TeV data-taking

MC reprocessing

2009 data reprocessing

東京大学 THE UNIVERSITY OF TOKYO ATLAS Operations - I. Ueda (CHEP2010, Taipei, 2010.10.18.) 10

# The ATLAS Computing Model

- The ATLAS Computing Model:

  https://twiki.cern.ch/twiki/bin/viewauth/Atlas/ComputingModel

- Multiple copies of derived data (AOD, ESD, DESD, etc.) are distributed worldwide (T1, T2, T3) in order to facilitate physics analysis
- We needed to predict usage patterns in order to estimate resource needs for funding agencies.
  - Sometimes 5 years in advance
  - This is very hard to get right!

# Current Computing Model Assumptions

Number of copies distributed

| Tier 1 | Number of equivalent reprocessings of whole set | | | |
|---|---|---|---|---|
| Raw | 3628 | 2540 | | 1 |
| ESD (current) | 3690 | 1292 | 1337 | 2 |
| ESD (Previous) | 1845 | 1292 | 0 | 1 |
| AOD (Currect) | 746 | 261 | 263 | 2 |
| AOD (Previous) | 0 | 261 | 0 | 0 |
| TAG | 21 | 1 | | 10 |
| DPD | 746 | 22 | 23 | 2 |
| MC Raw | 171 | 2400 | | 0.1 |
| MC ESD (Current) | 754 | 264 | 639 | 2 |
| MC ESD(last) | 377 | 528 | 0 | 1 |
| MC AOD (current) | 463 | 162 | 266 | 2 |
| MC AOD (last) | 0 | 324 | | 0 |
| MC TAG | 9 | 1 | | 10 |
| MC DPD | 139 | 0 | 6 | 2 |
| **Tier 2** | | | | |
| Raw | 363 | | | 0.1 |
| ESD (current) | | | | |
| ESD (Previous) | | | | |
| AOD (Currect) | 3732 | | | 10 |
| AOD (Previous) | 3732 | | | 10 |
| TAG | 21 | | | 10 |
| DPD | 3732 | | | 10 |
| MC Raw | 171 | | | 0.1 |
| MC ESD (Current) | | | | |
| MC ESD(last) | | | | |
| MC AOD (current) | 2314 | | | 10 |
| MC AOD (last) | 926 | | | 4 |
| MC TAG | 9 | | | 10 |
| MC DPD | 694 | | | 10 |

# Data Caching for Distributed Analysis

Kaushik De

Univ. of Texas at Arlington

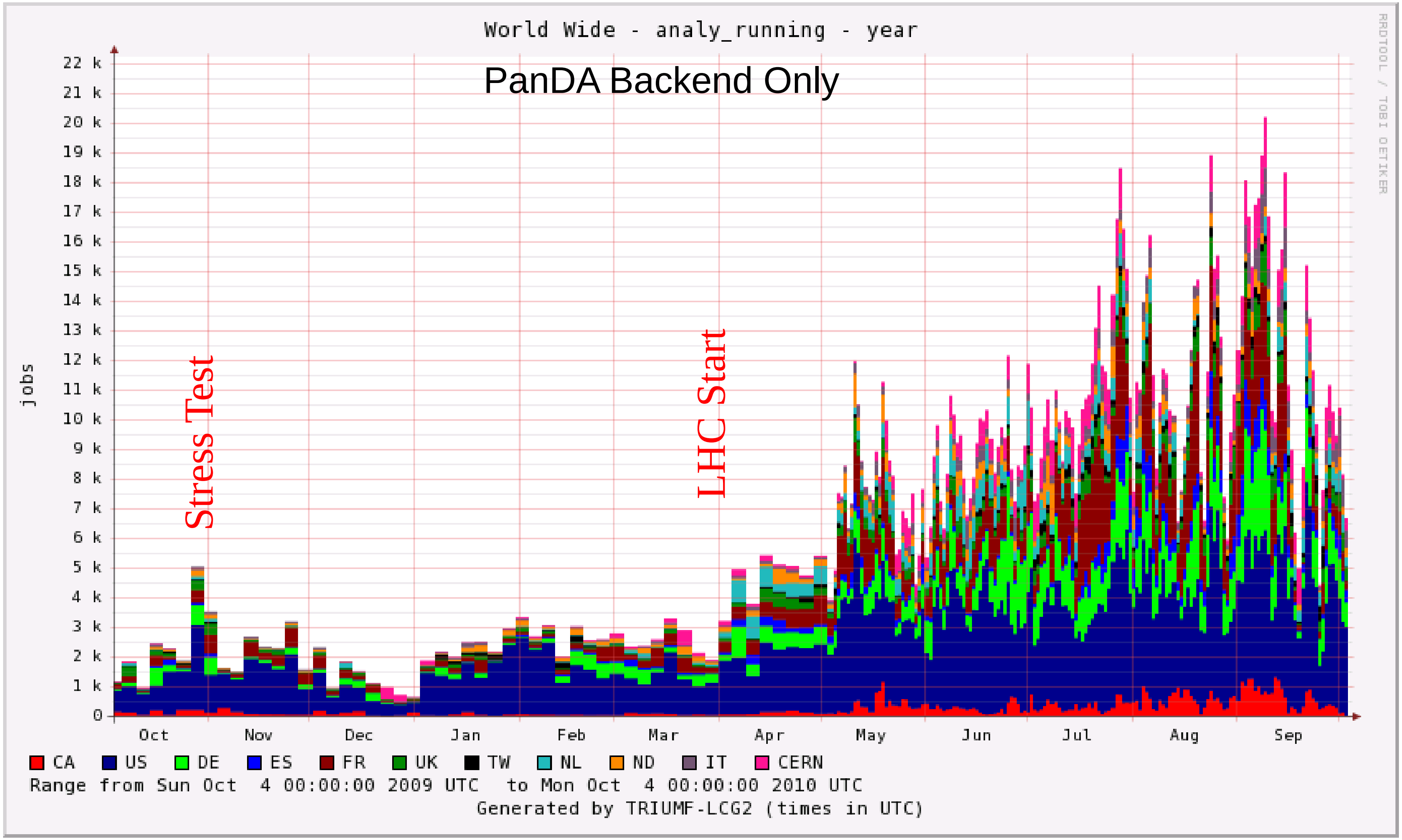
ATLAS Week, CERN

Oct 5, 2010

# Introduction

- Distributed User Analysis is often called Chaotic Analysis
  - Because it is unpredictable: number of users, number of jobs, duration of jobs, file types used, number of files… all fluctuate wildly
- We have been very successful in spite of the complexity
  - Over 1,300 different users of PanDA during past 6 months
  - Millions of jobs are run every week at hundreds of sites
  - Many people working in the background make this possible
  - But we had to be nimble and make changes since LHC started
  - In this talk, I will describe one of the newest (and biggest) changes in how we do distributed computing in ATLAS

# Huge Rise in Analysis Activity

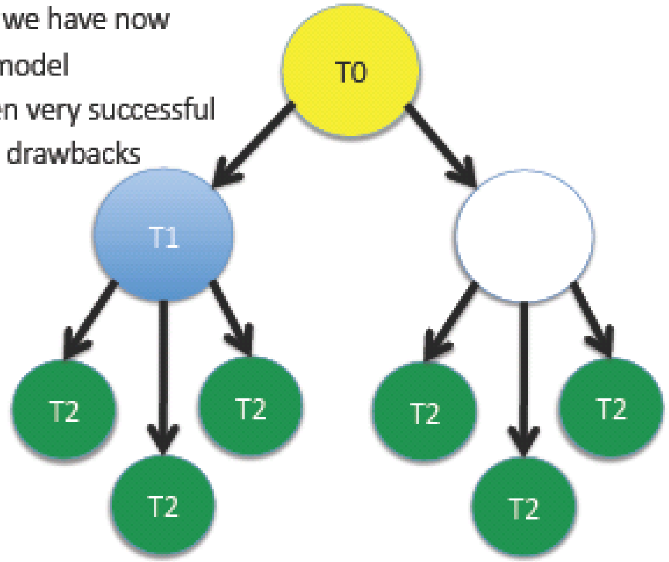


# Data Distribution is Very Important

- Most user analysis jobs run at Tier 2 sites
  - Jobs are sent to data
  - We rely on pushing data out to Tier 2 sites promptly
  - Difficult since many data formats and many sites
  - We adjusted frequently the number of copies and data types in April & May
  - But Tier 2 sites were filling up too rapidly, and user pattern was unpredictable
  - Most datasets copied to Tier 2's were never used

Data placement model
The"Monarch Model"

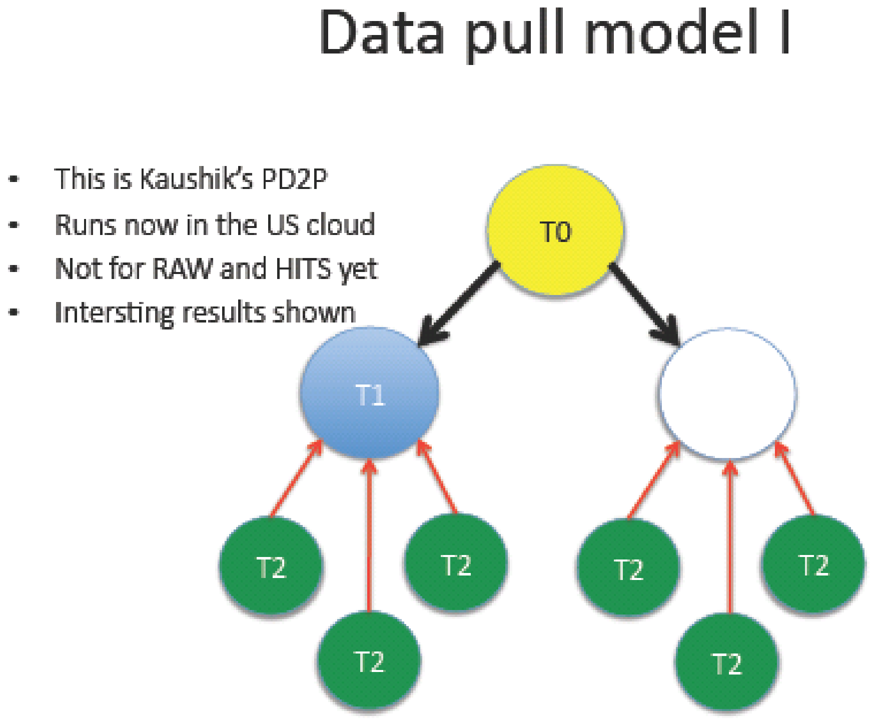
- This is what we have now
- It is a push model
- And has been very successful
- But has also drawbacks



From Kors, SW Week

# We Changed Data Distribution Model

- Reduce pushed data copies to Tier 2's
  - Only send small fraction of AOD's automatically
  - Pull all other data types, when needed by users
  - Note: for production we have always pulled data as needed
- But users were insulated from this change
  - Did not affect the many critical ongoing analyses
  - No delays in running jobs
  - No change in user workflow

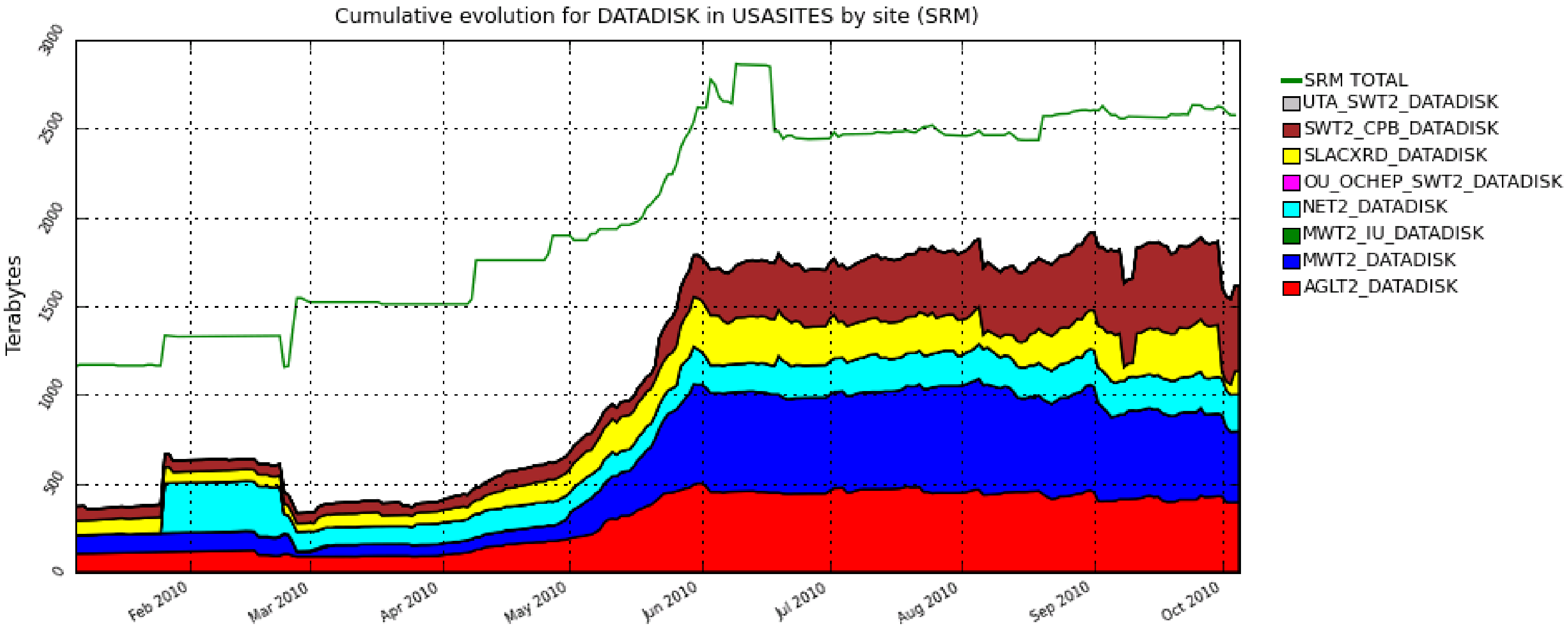


Data pull model I

- This is Kaushik's PD2P
- Runs now in the US cloud
- Not for RAW and HITS yet
- Intersting results shown

From Kors, SW Week

# Data Flow to Tier 2's



- Example above is from US Tier 2 sites
  - Exponential rise in April and May, after LHC start
  - We changed data distribution model end of June – PD2P
  - Much slower rise since July, even as luminosity grows rapidly

# What is PD2P

- Dynamic data placement at Tier 2’s
  - Continue automatic distribution to Tier 1’s – treat them as repositories
  - Reduce automatic data subscriptions to Tier 2’s – instead use PD2P
- The plan
  - Panda will subscribe a dataset to a Tier 2, if no other copies are available (except at a Tier 1), as soon as any user needs the dataset
    - User jobs will still go to Tier 1 while data is being transferred – no delay
  - Panda will subscribe replicas to additional Tier 2’s, if needed, based on backlog of jobs using the dataset (PanDA checks continuously)
  - Cleanup will be done by central DDM popularity based cleaning service (as described in previous talk by Stephane)
- Few caveats
  - Start with DATADISK and MCDISK
  - Exclude RAW, RDO and HITS datasets from PD2P
  - Restrict transfers within cloud for now
  - Do not add sites too small (storage mainly) or too slow

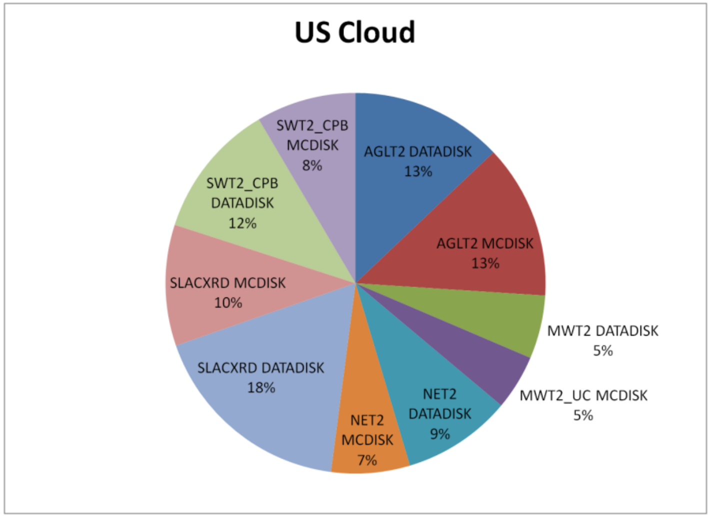
# Main Goals

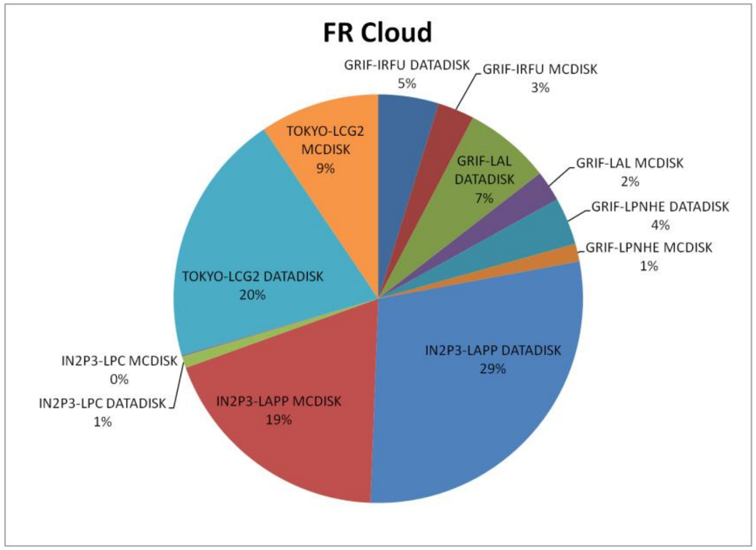
- User jobs should not experience delay due to data movement
- First dataset replication is ‘request’ based
  - Any user request to run jobs will trigger replication to a Tier 2 chosen by PanDA brokering – no matter how small or large the request
- Additional dataset replication is ‘usage’ based
  - Send replicas to more Tier 2’s if a threshold is crossed (many jobs are waiting for the dataset)
- Types of datasets replication are ‘policy’ based
  - We follow Computing Model – RAW, RDO, HITS are never replicated to Tier 2’s by PanDA (we may have more complex rules later, to allow for small fraction of these types to be replicated)
  - PanDA does replication only to DATADISK and MCDISK, for now
- Replication pattern is ‘cloud’ based
  - Even though subscription source is not specified, currently PanDA will only initiate replication if source is available within cloud (we hope to relax this in the next phase of tests)

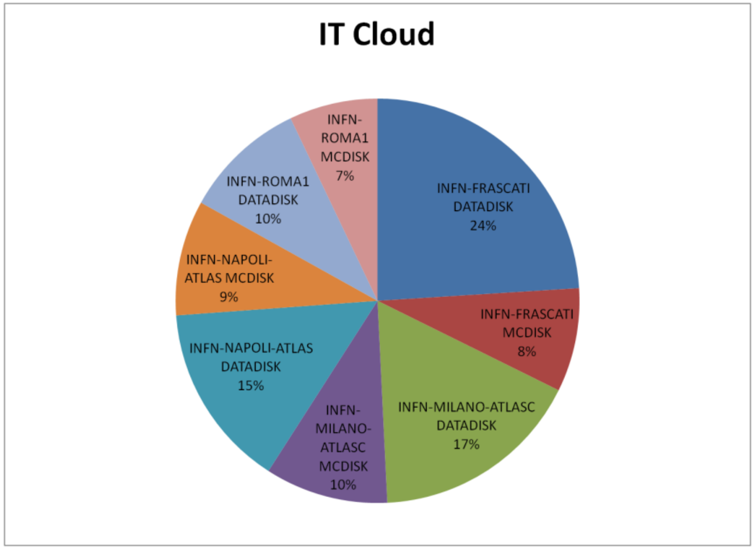
# Some Statistics

- Running for 3+ months now – since Jun 15
- Started in US cloud, and then FR cloud, now IT cloud
- 5870 datasets subscribed so far
  - Most datasets are never used and therefore never copied to Tier 2
  - Majority of datasets copied by PD2P still not reused at Tier 2
    - This will change soon because of automatic rebrokering
  - However, those which are reused, are reused often
  - 1,634,272 files were reused by other user jobs, so far in 3+ months
- Now lets look at some PD2P results/plots

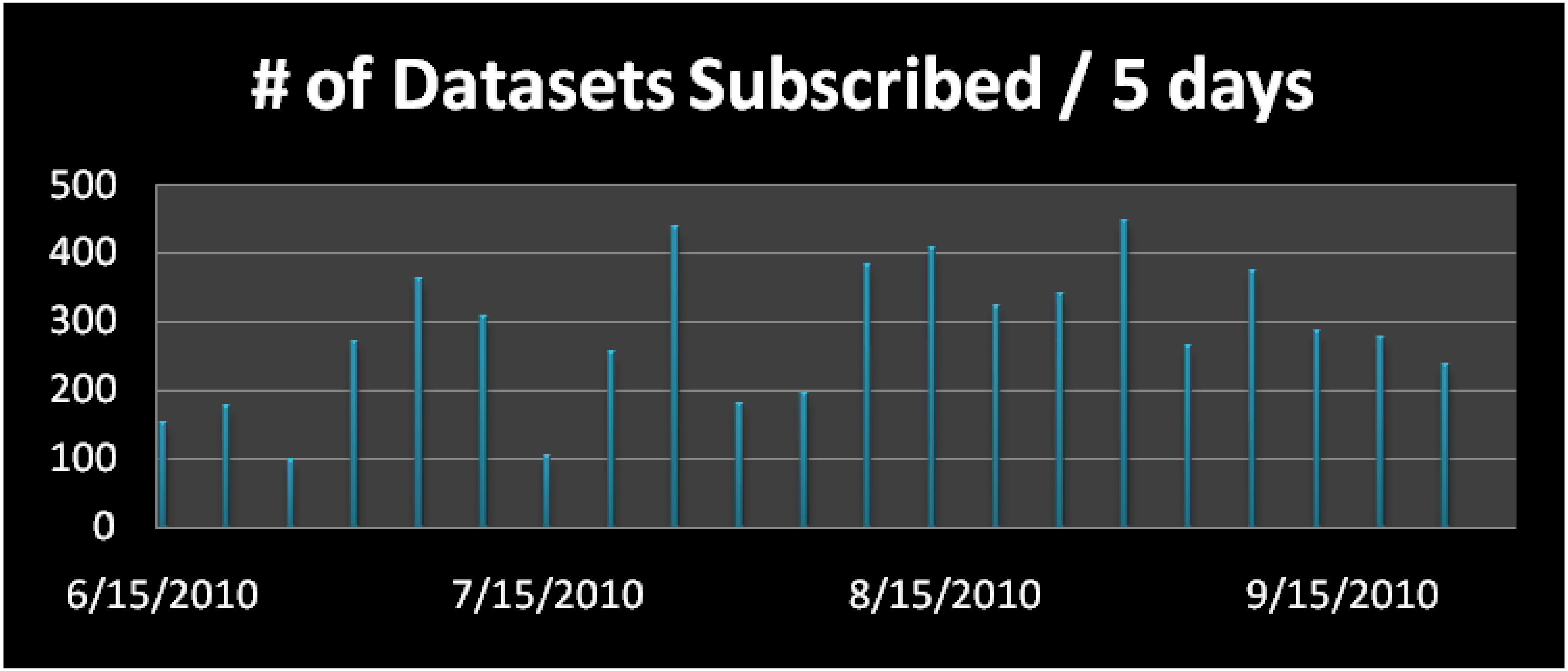
# Distribution Among Sites is Even





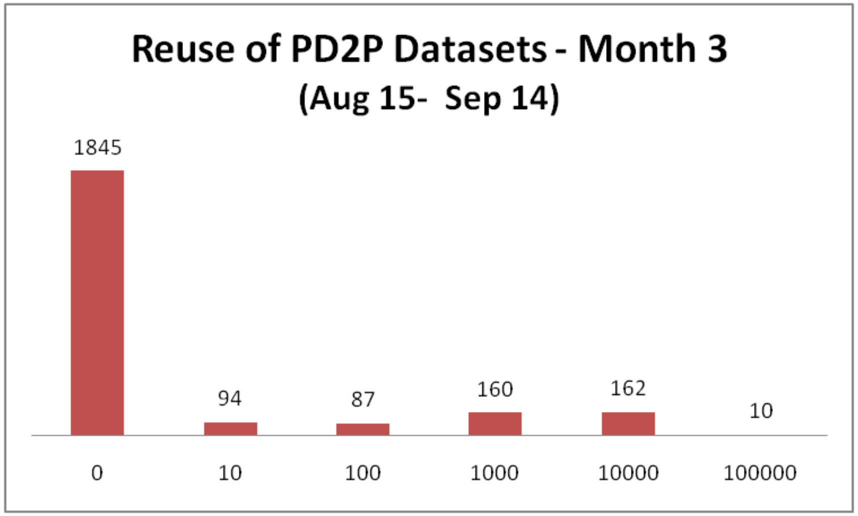


# Rate is also Even

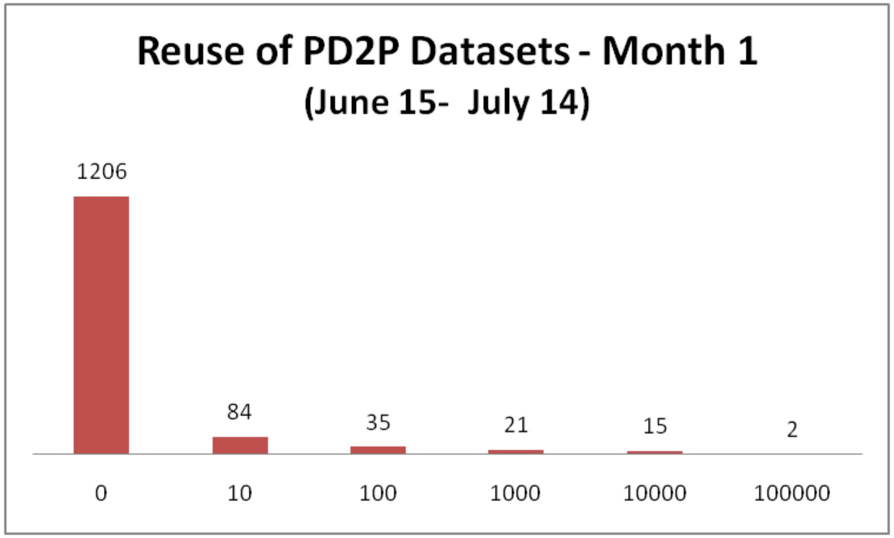


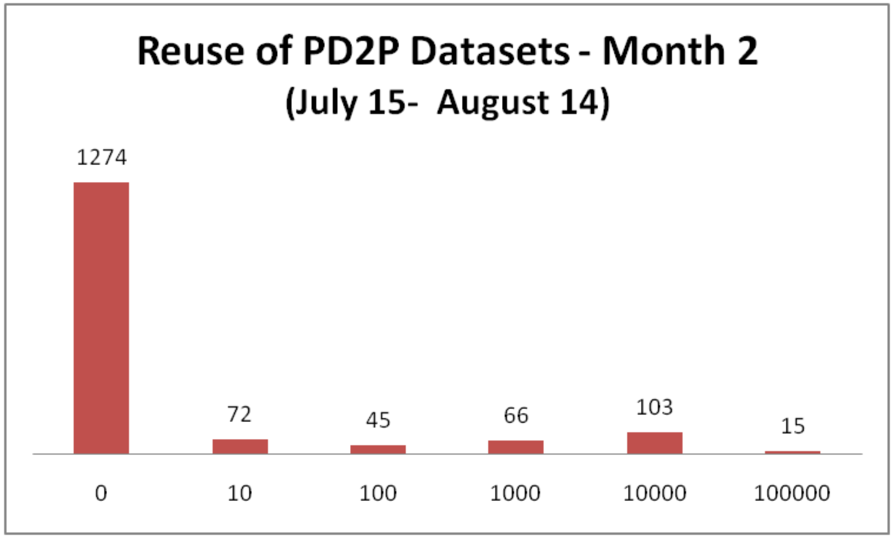
Summed over all three clouds

# Reuse of PD2P Files



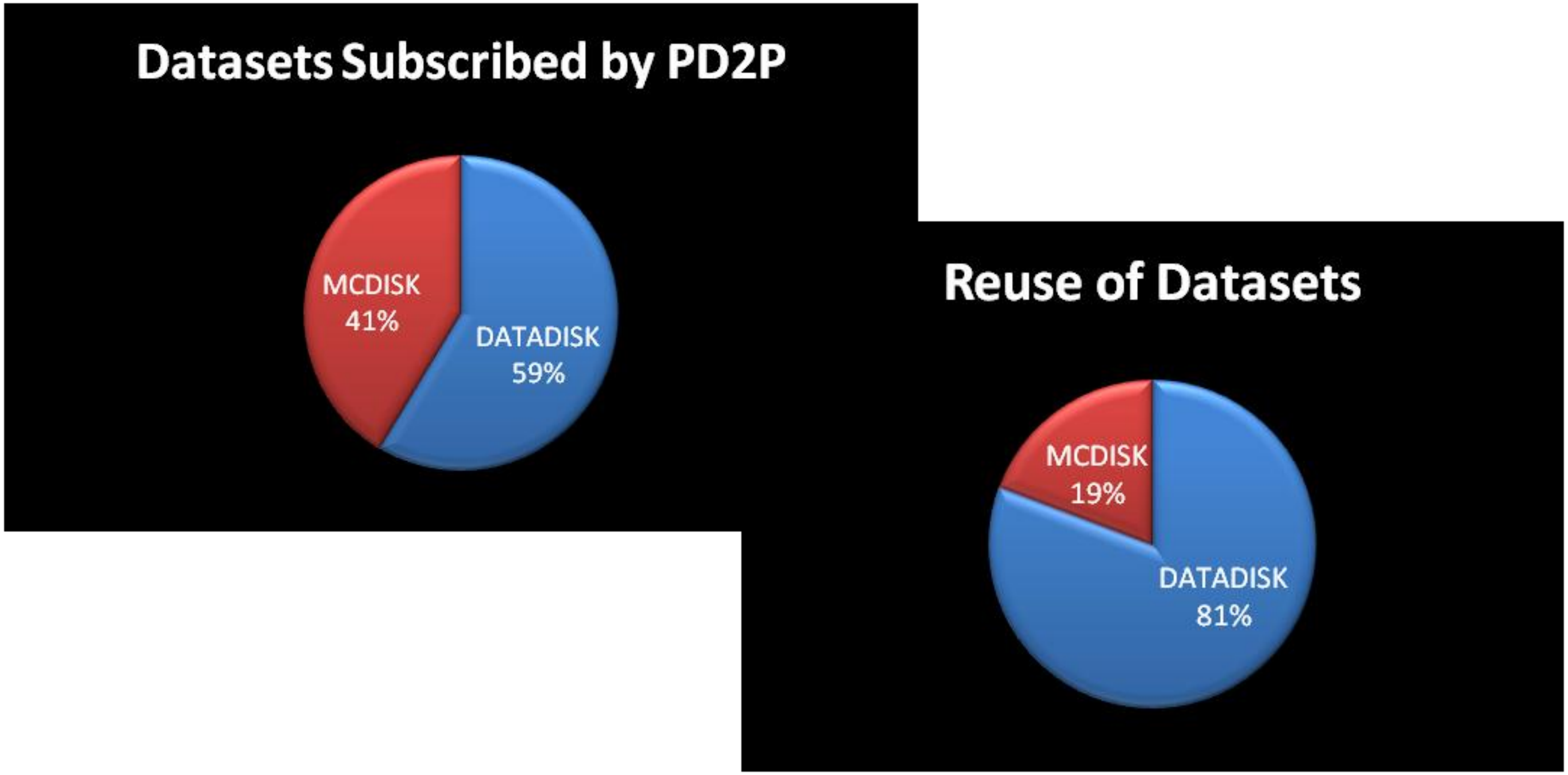
We plot here the number of datasets subscribed by PD2P which were accessed later by other users (x-axis shows number of files accessed)





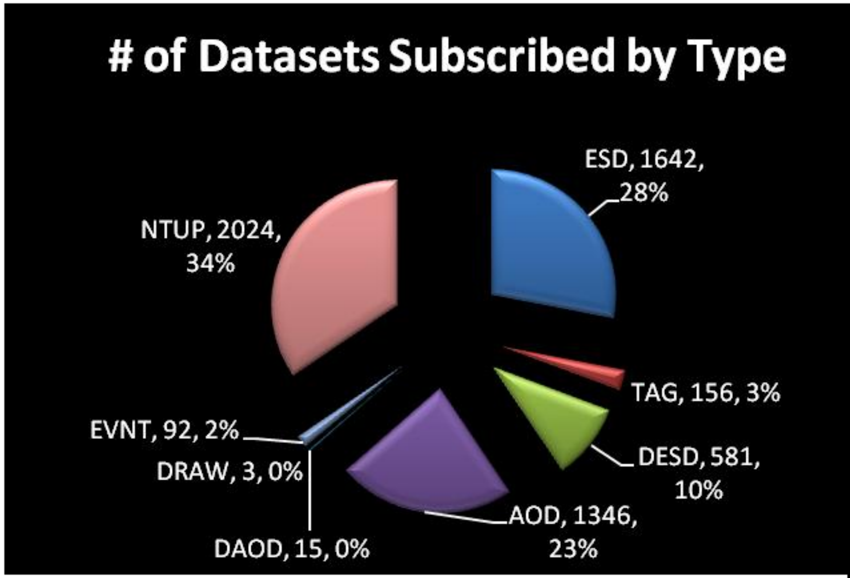
# Patterns of Data Usage – Part I

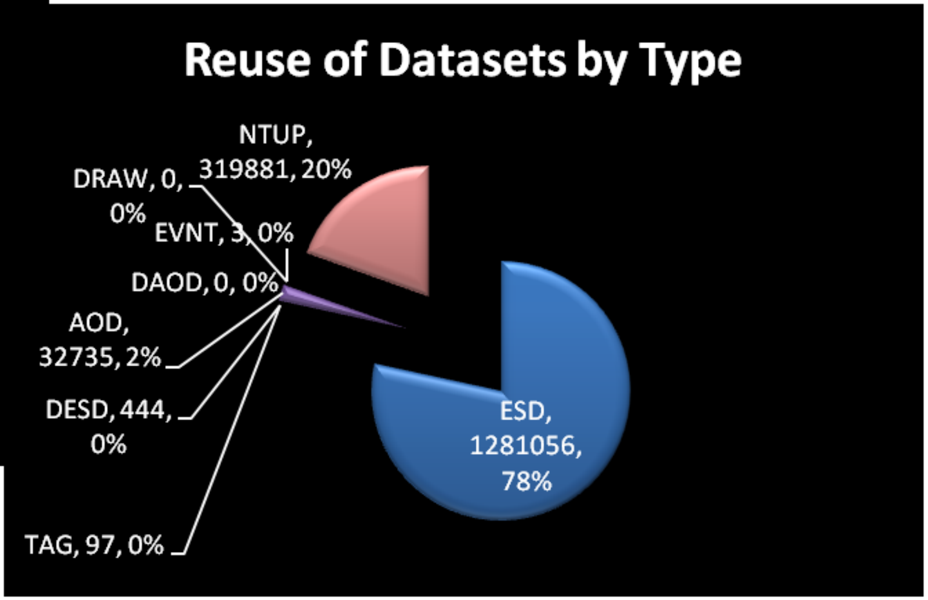
- Interesting patterns are emerging by type of data
    - LHC data reused more often than MC data – not unexpected



# Patterns of Data Usage – Part 2

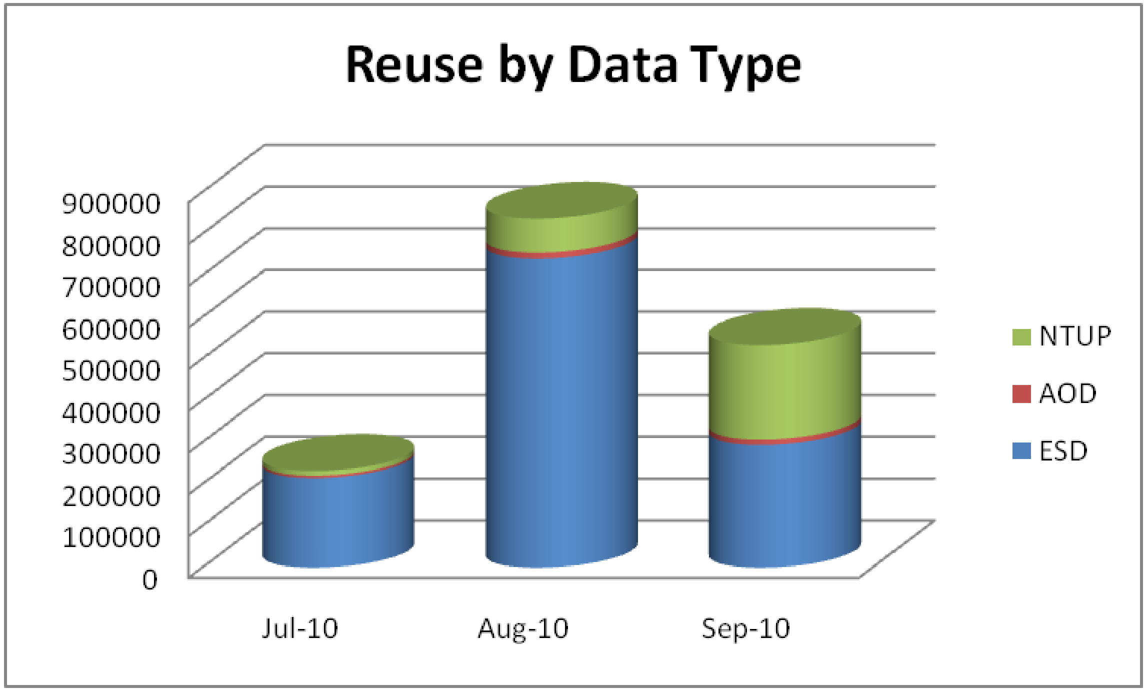
- Interesting patterns also by format of data
- During past 3+ months:
  - All types of data showing up: ESD, NTUP, AOD, DED most popular
  - But highest reuse (counting files): ESD, NTUP





# Trends in Data Reuse

- PD2P pull model does not need a priori assumption about popular data types for user analysis
- It automatically moves data based on user workflow
- We observe now a shift towards using DPD's (NTUP)



# Recent Improvements to PD2P

- Re-brokering was implemented last week
  - PanDA will now re-broker jobs to a different site, if they remain in queue too long (site problems, too many users, long jobs…)
  - Side effect – users can now use dataset containers for output
  - If dataset containers are used, sub-jobs may now be brokered to multiple sites for faster execution (in the past all sub-jobs went to a single site chosen by PanDA)
  - Results of these changes do not show up in plots yet, but will speed up user job completions, and balance the load better among sites

# What Next?

- Is it time to tune PD2P algorithm?
  - Not yet – rate of subscriptions is still low (much lower than subscribing all datasets available, as before PD2P)
  - Low threshold for first subscription helps additional users, even if the subscribed datasets are seldom reused
  - High threshold for multiple subscriptions - only copy hot datasets
  - We will monitor and optimize PD2P as data volume grows
- Can we improve and expand to other caching models?
  - Many ideas on the table
  - For example: using ROOT TreeCache
  - For example: using XRootD based caching
  - These require longer term development
  - Large Scale Demonstrator LST2010 – CERN IT and ATLAS project

# Useful Links

- The ATLAS Computing Model:

  https://twiki.cern.ch/twiki/bin/viewauth/Atlas/ComputingModel

- Shift information:
  - ADCOS:
    - https://twiki.cern.ch/twiki/bin/viewauth/Atlas/ADCoS
  - Comp@P1:
    - https://twiki.cern.ch/twiki/bin/viewauth/Atlas/CompAtP1Shift