Wide Area Network Data Access Requirements in Analysis

Doug Benjamin

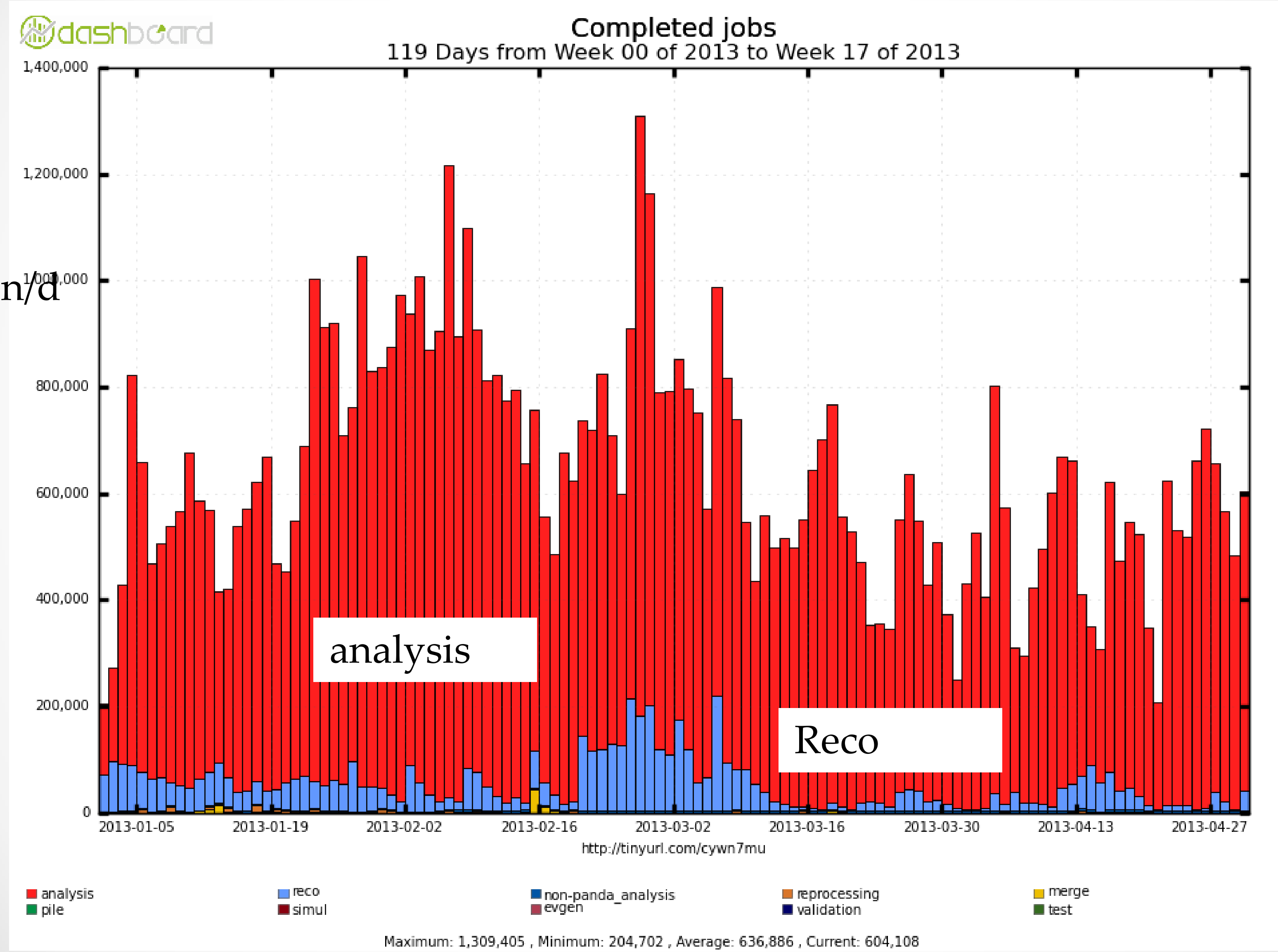
Duke University

Setting the stage

- Since the beginning of the year, how much analysis was done?
- Why Jan 1 to May 1?
  - has analysis ramp up ahead of major Winter conferences (Moriond series)
  - Post conference activity perhaps indicative of steady state activity well into LS1 period
- What data do analyzers use?

Completed Jobs (1-Jan to 1 May 2013)

1 million/day

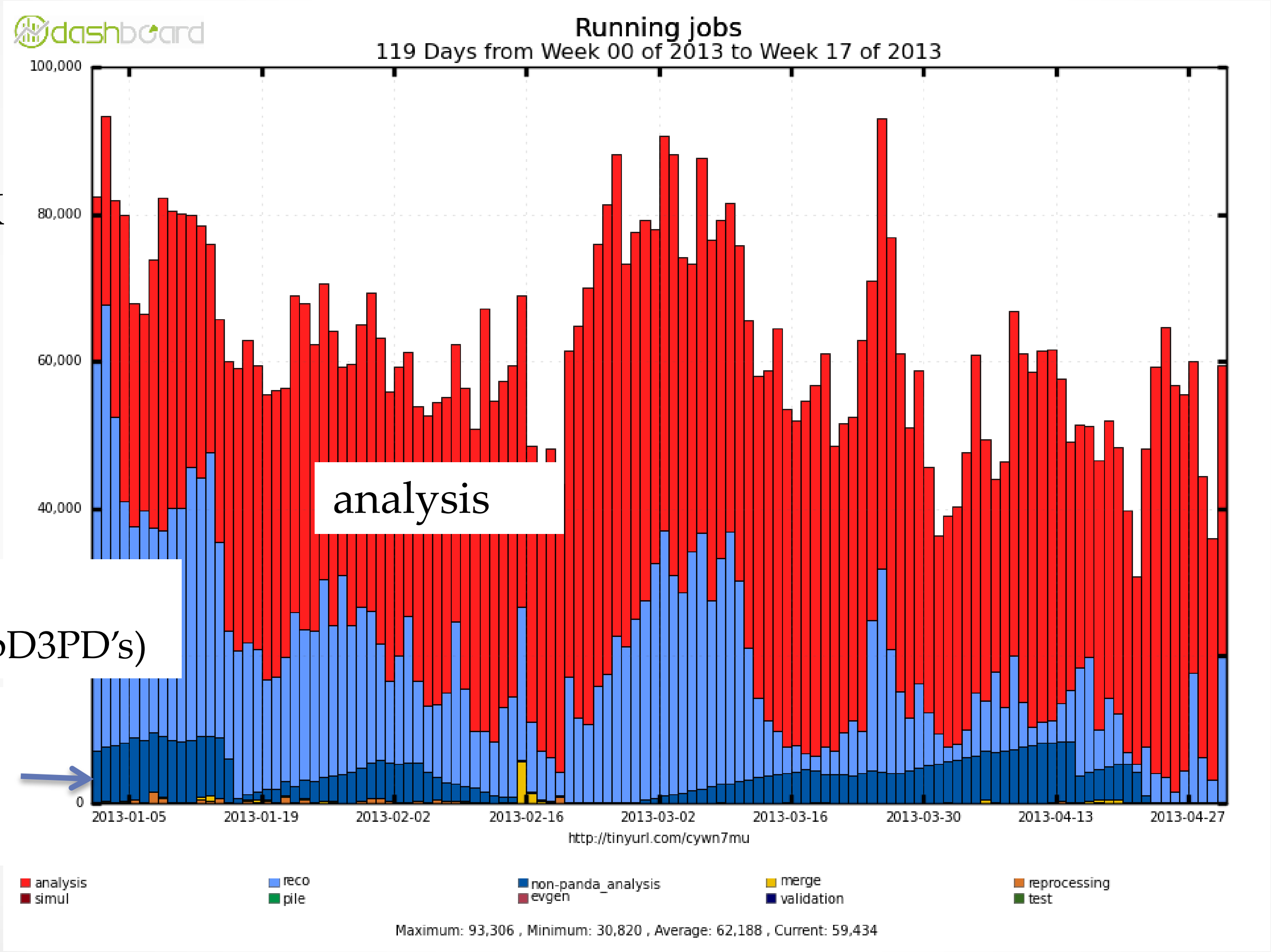


analysis

Reco

Job Type - Group Analysis, Group production, User analysis

User Analysis jobs 1-Jan – 1-May 2013



80K

analysis

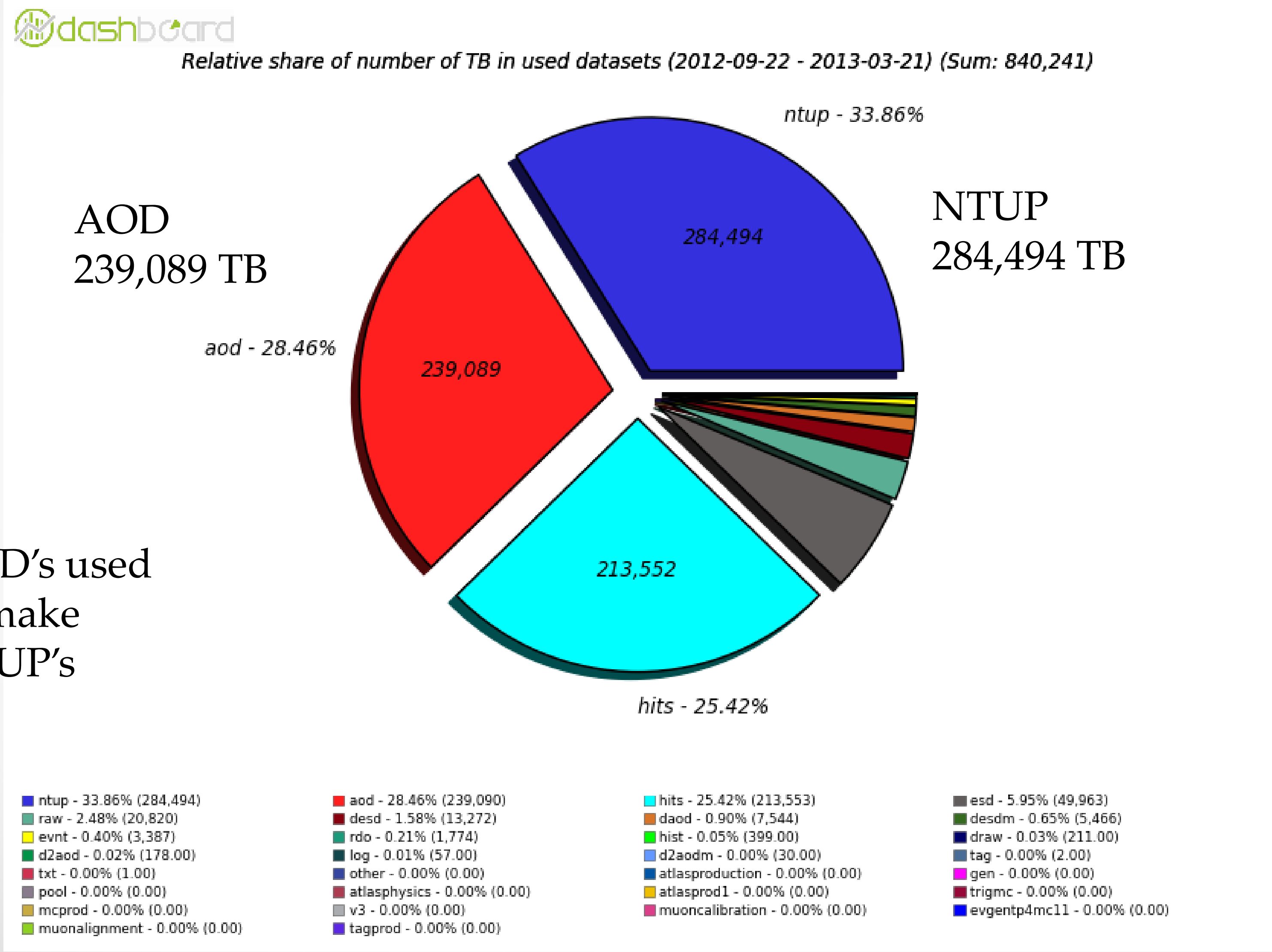
Reco (groupD3PD's)

Non-Panda analysis

Job Type - Group Analysis, Group production, User analysis

Data popularity

(used datasets 2012-09-22 to 2013-03-21)



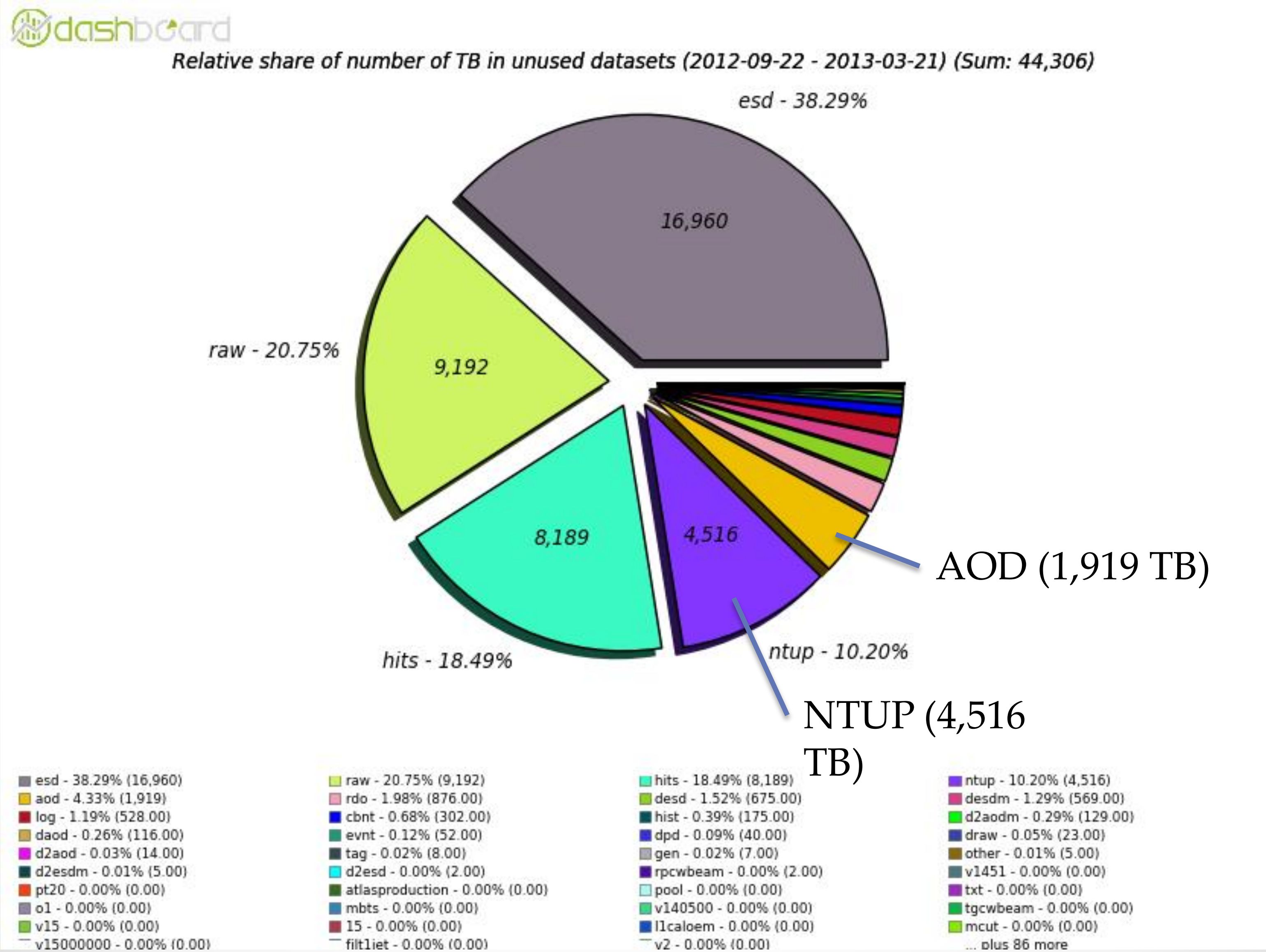
AOD
239,089 TB

NTUP
284,494 TB

AOD's used
to make
NTUP's

Data unpopularity

(Unused datasets)



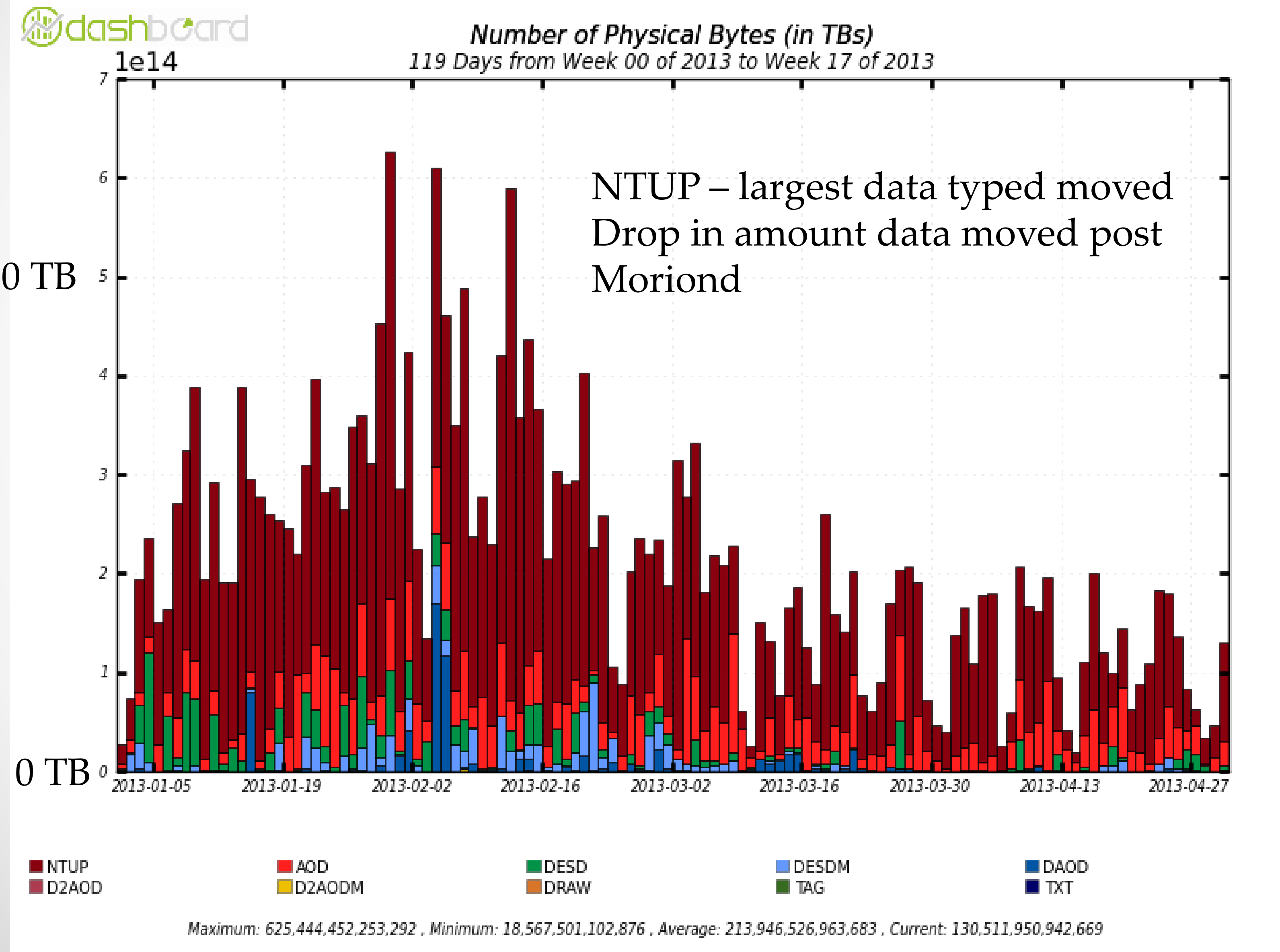
AOD (1,919 TB)

NTUP (4,516 TB)

Interim summary

- 1-Jan to 1-May 2013 – peak of over 1M analysis jobs per day
- Prior to Moriond ~800 K jobs/day
- Post Moriond ~ 400 – 600 K jobs/day
- Jobs are typically short (most < 4 hours)
- In 6 month period End – Sep 2012 to end March 2013, NTUP's most popular data type 284 PB used (~ 67 PB/month)
- In same period - 4.5 PB NTUP's unused (~ 1.5%)

Analysis triggered data motion (PD2P)

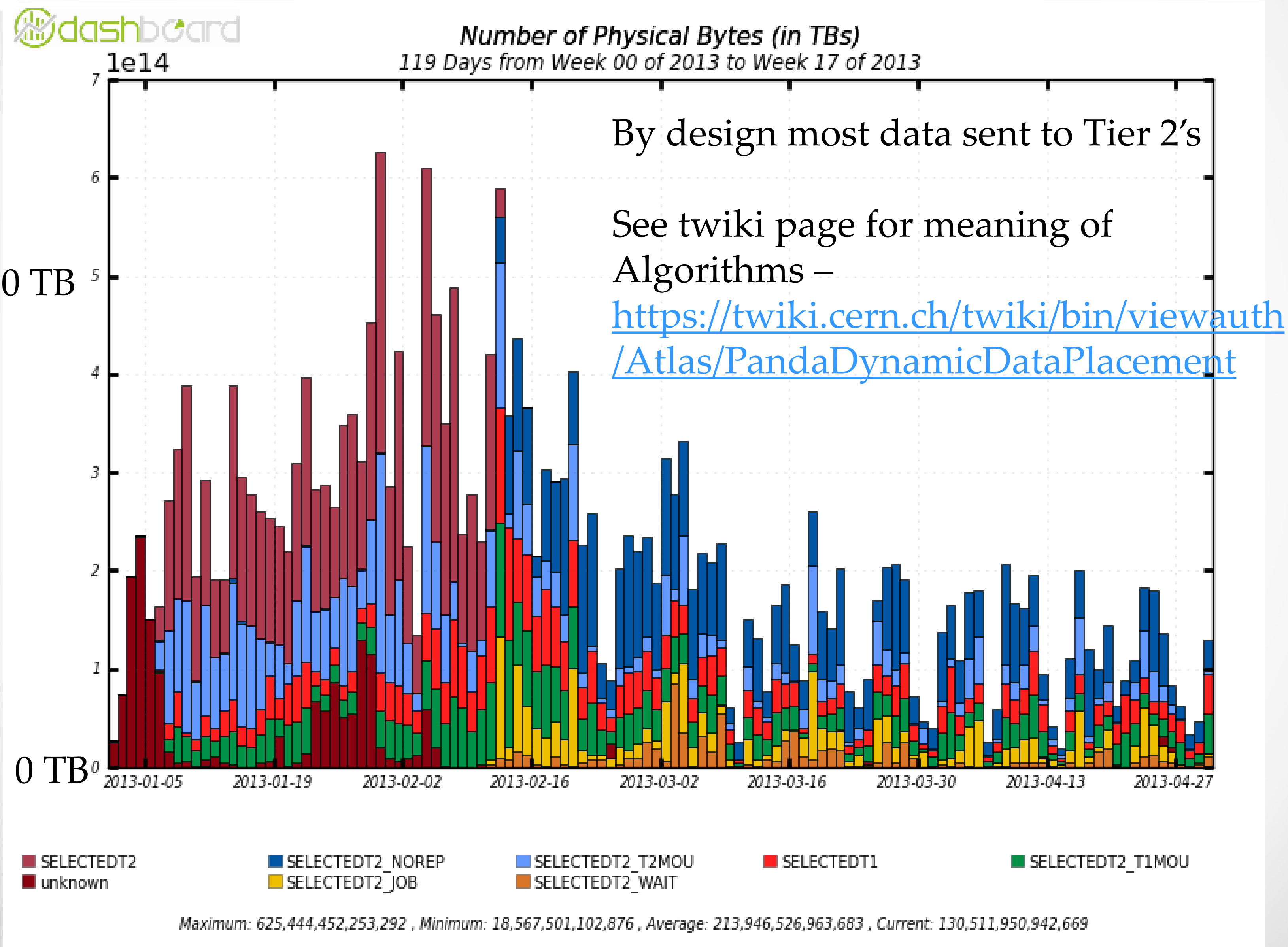


NTUP – largest data typed moved
Drop in amount data moved post Moriond

500 TB

0 TB

PD2P by Algorithm



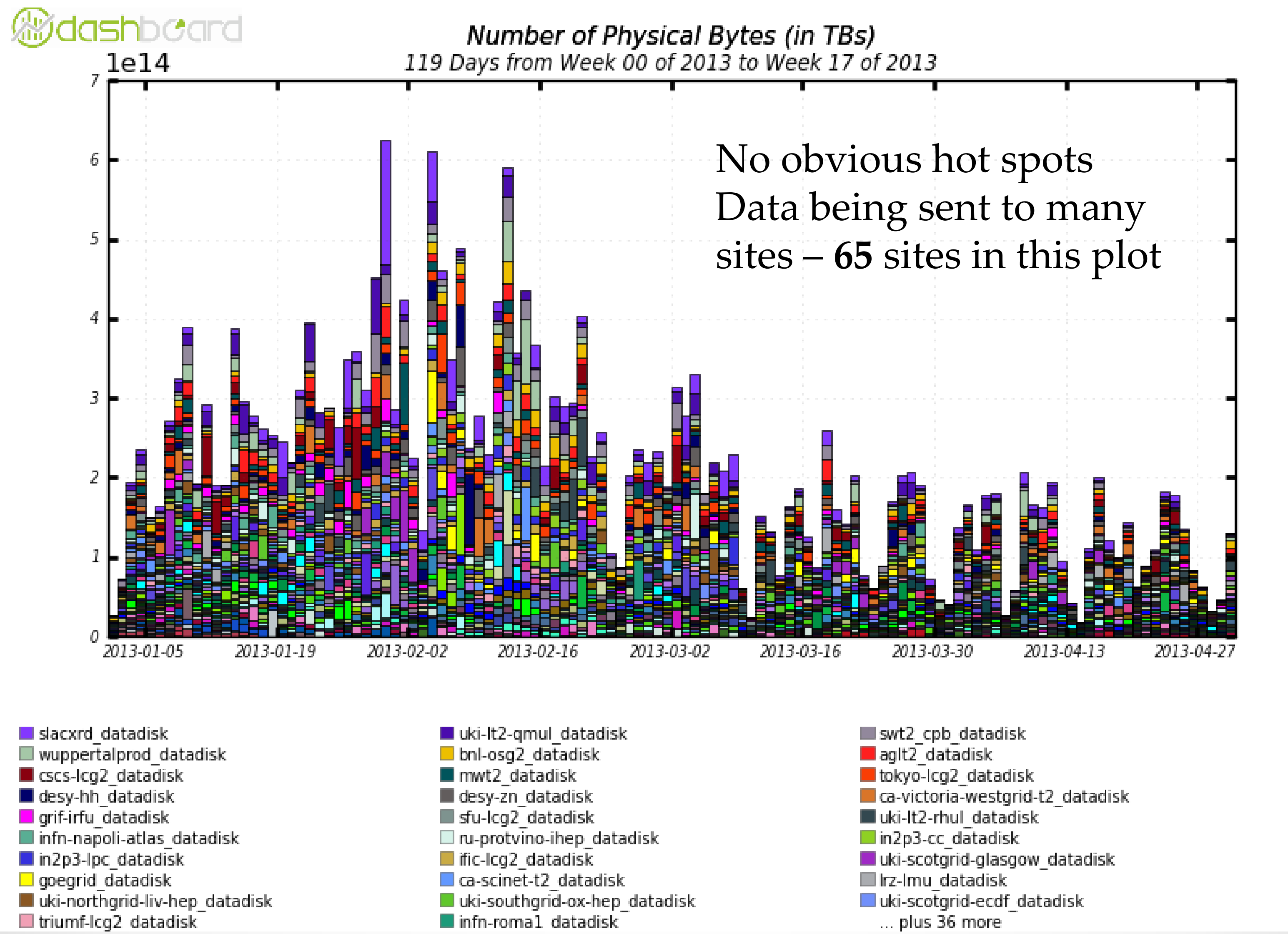
500 TB

0 TB

By design most data sent to Tier 2's

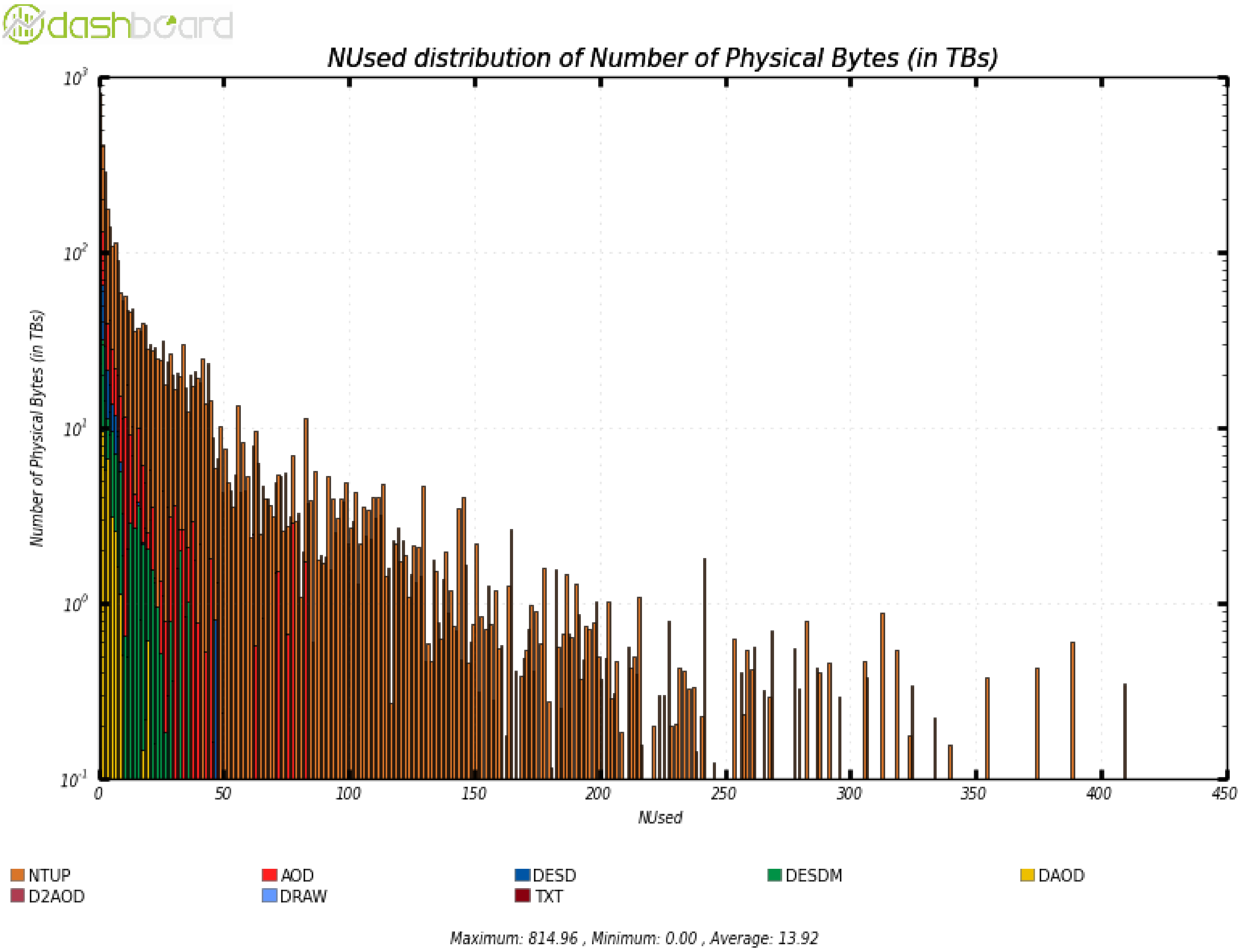
See twiki page for meaning of Algorithms – https://twiki.cern.ch/twiki/bin/viewauth/Atlas/PandaDynamicDataPlacement

PD2P sends data everywhere

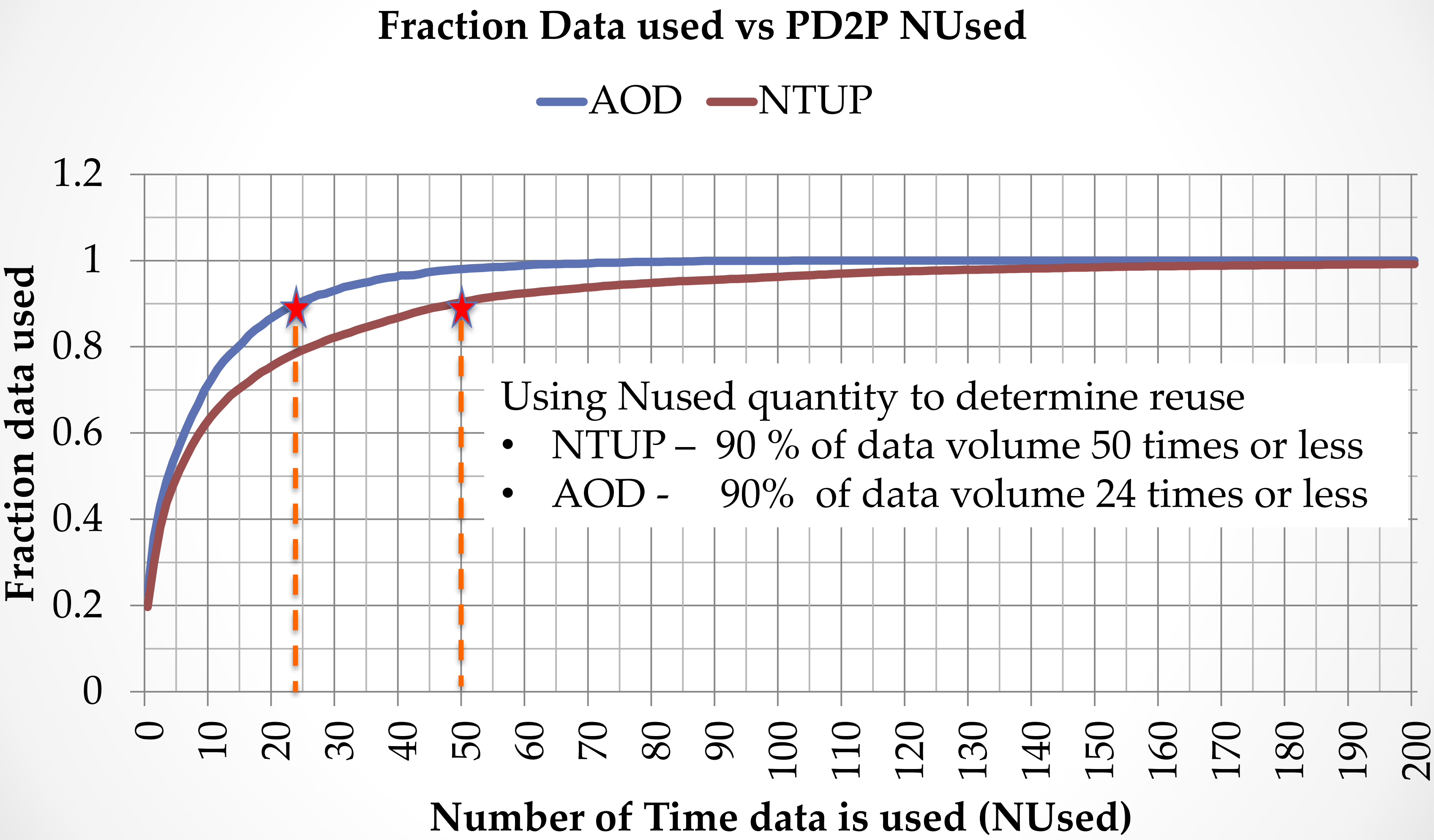


No obvious hot spots
Data being sent to many sites – **65** sites in this plot

Dataset reuse



Fraction of Data volume reused



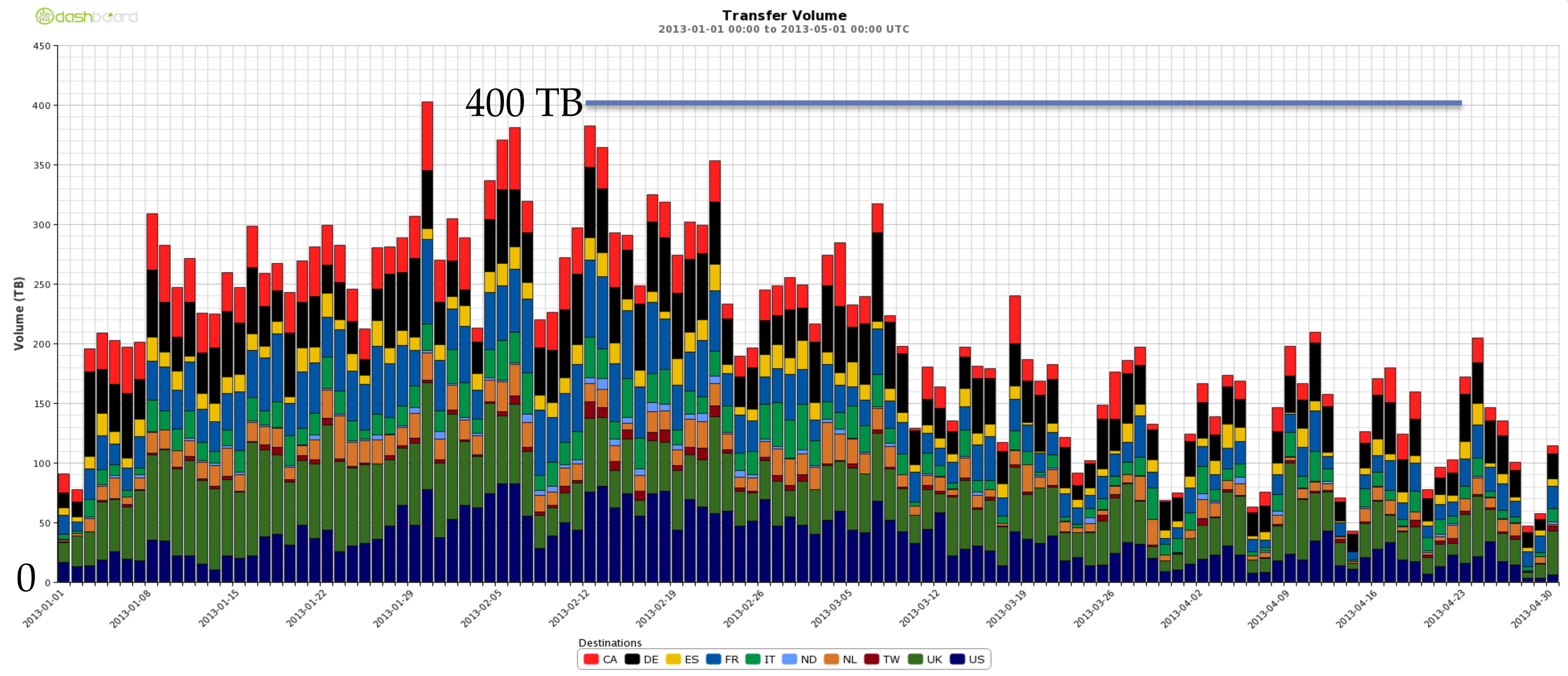
Using Nused quantity to determine reuse

- NTUP – 90 % of data volume 50 times or less
- AOD - 90% of data volume 24 times or less

PD2P interim summary

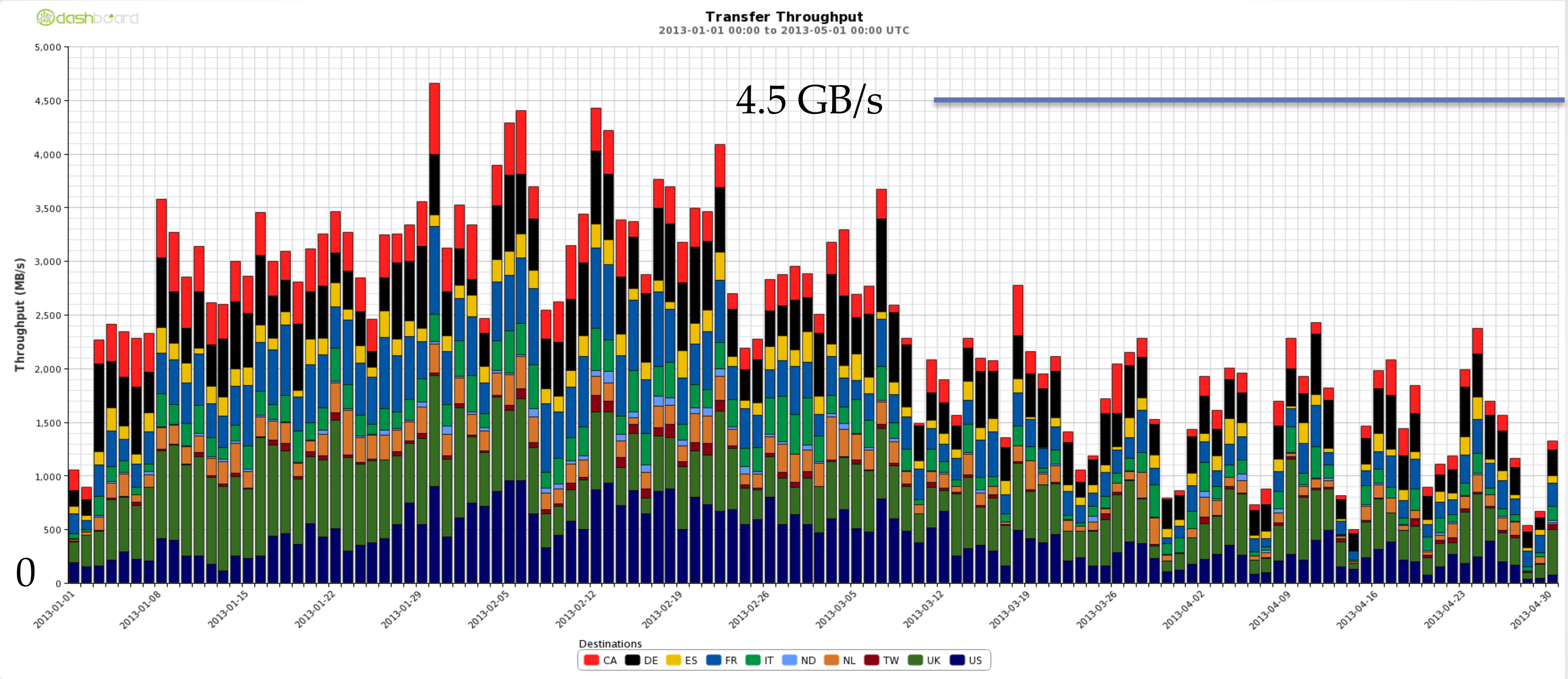
- Jan-Apr 2013 – PD2P moved at peak ~ 600 TB/day
- NTUP most popular data type to move
    - Not a surprise given the popularity plots
- Most data moved to Tier 2 sites
- Data moved to a wide variety of sites (> 67 sites)
- … something about dataset reuse….

Planned data replication



DDM – Daily Data Brokering data Volume (1-Jan to 1-May 2013)

Data Brokering transfer Rate

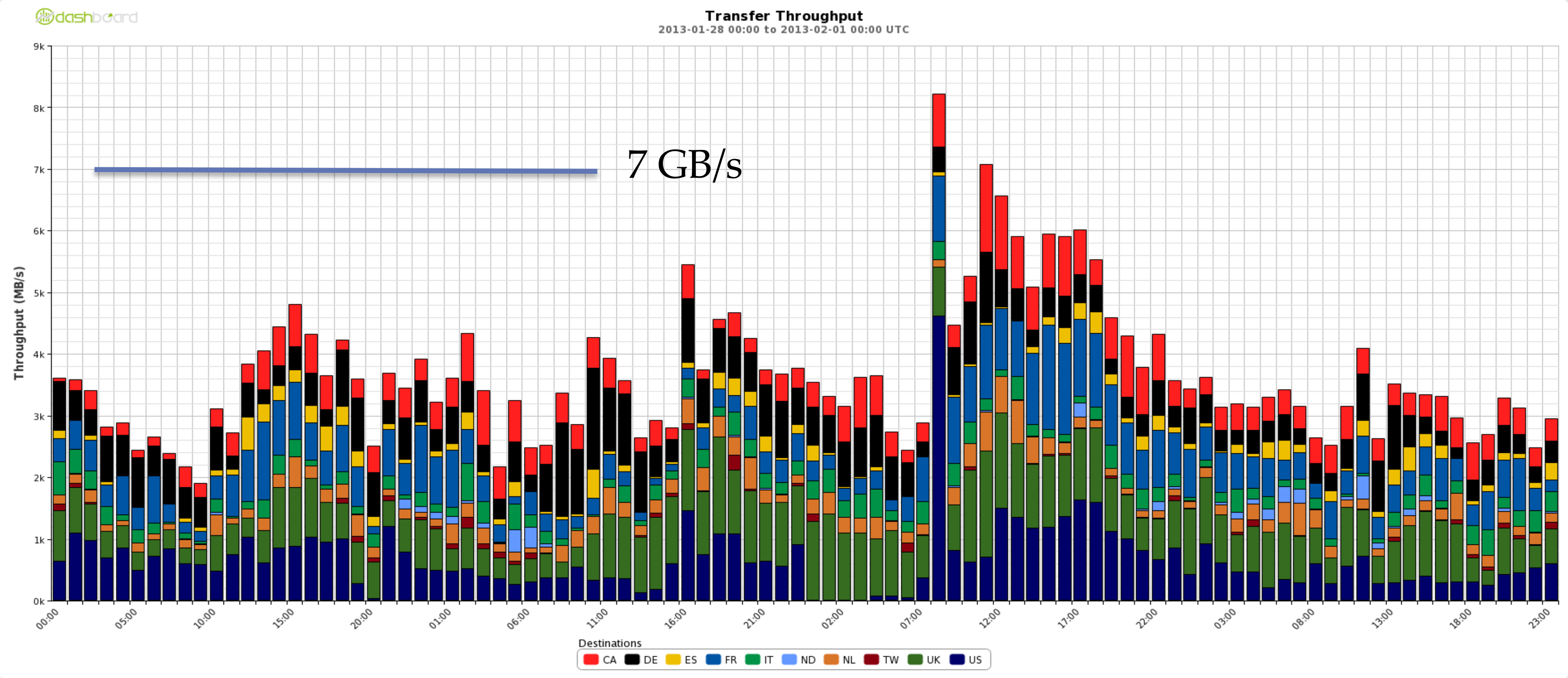


4.5 GB/s

0

Transfer Throughput (averaged over a day) 1-Jan to 1-May 2013

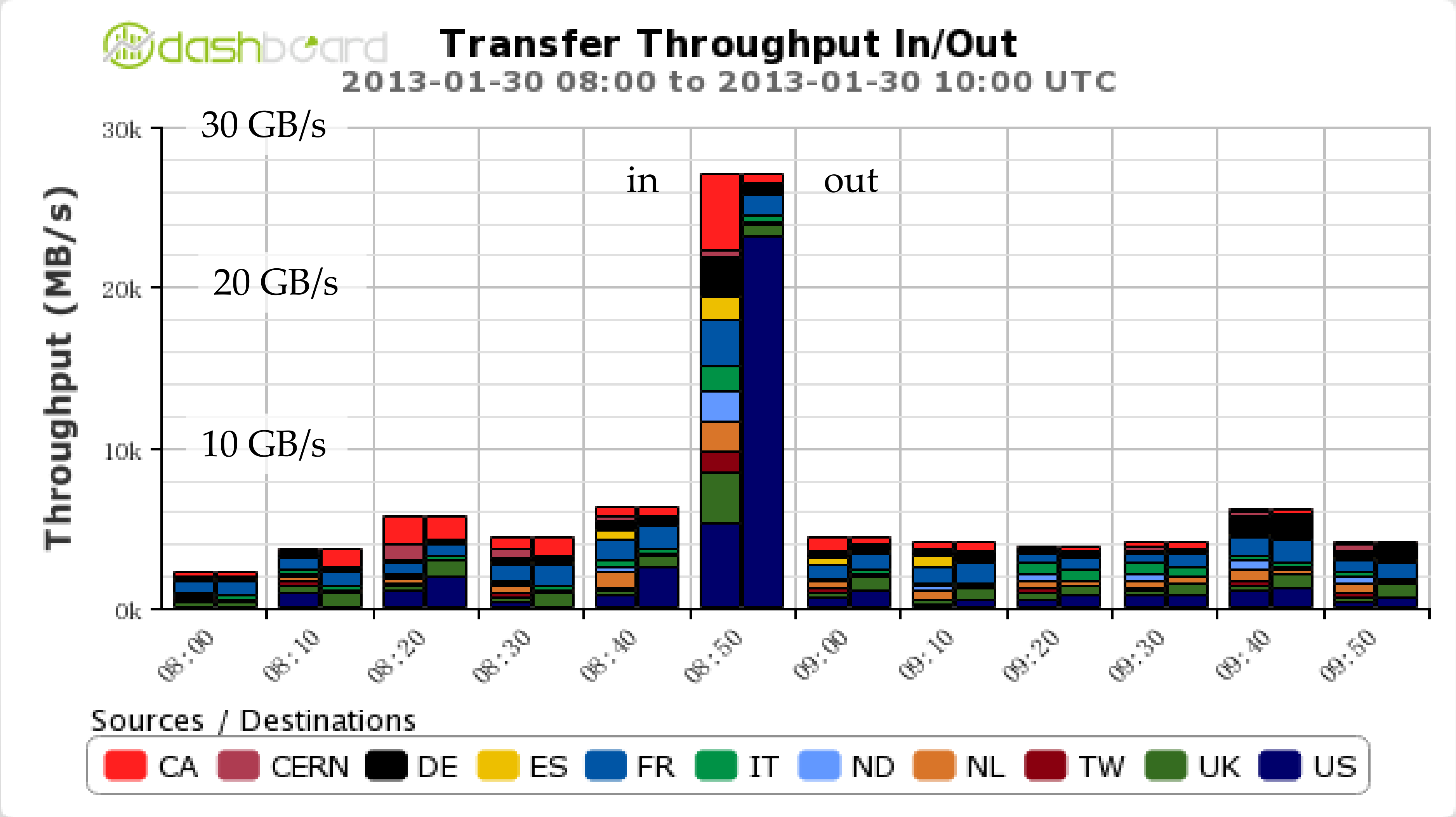
Data Brokerage transfer rate 28-Jan – 31-Jan (48 hrs)



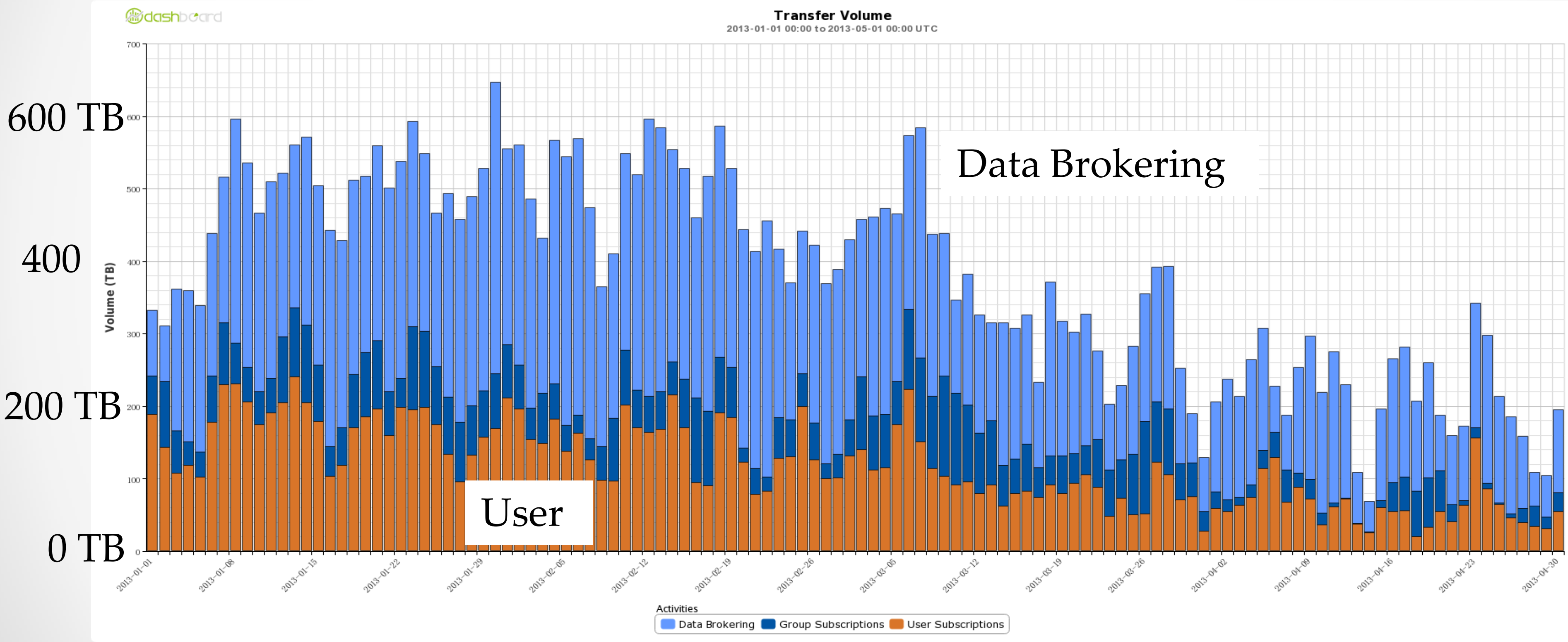
7 GB/s

Transfer Throughput (one hour bins) 28-Jan to 1-Feb 2013

Transfer rate zoom in further



Planned transfers Data Brokering, Group Subscriptions, User subscriptions



Planned Data Brokerage summary

- Peak 400 TB/day and ~ 200 TB/day post Moriond
- Transfer rates
  - o Peak 4.5 GB/s (day resolution)
  - o During busiest few days 28-Jan through 31-Jan-13 (48 hrs)

    7 GB/s (one hour resolution)
  - o Jan 30, 2013 (08:00 – 10:00) peak time, in Peak 10 minute period

    30 GB/s - reading mostly from US Cloud to rest of ATLAS

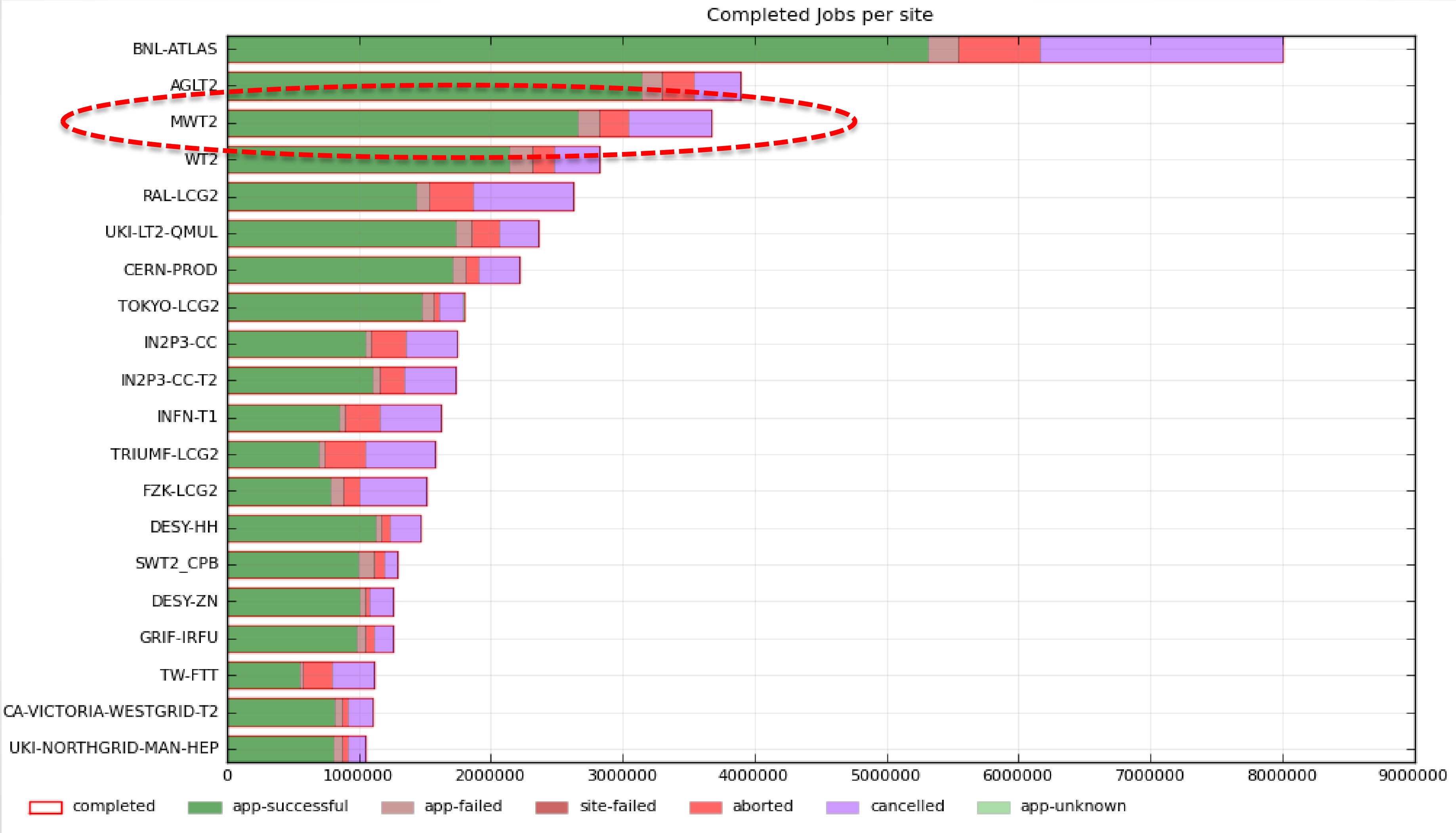
    ~ 5 GB/s other 10 minute periods
- User data subscriptions

150-200 TB/day pre Moriond

50-100 TB/day post Moriond

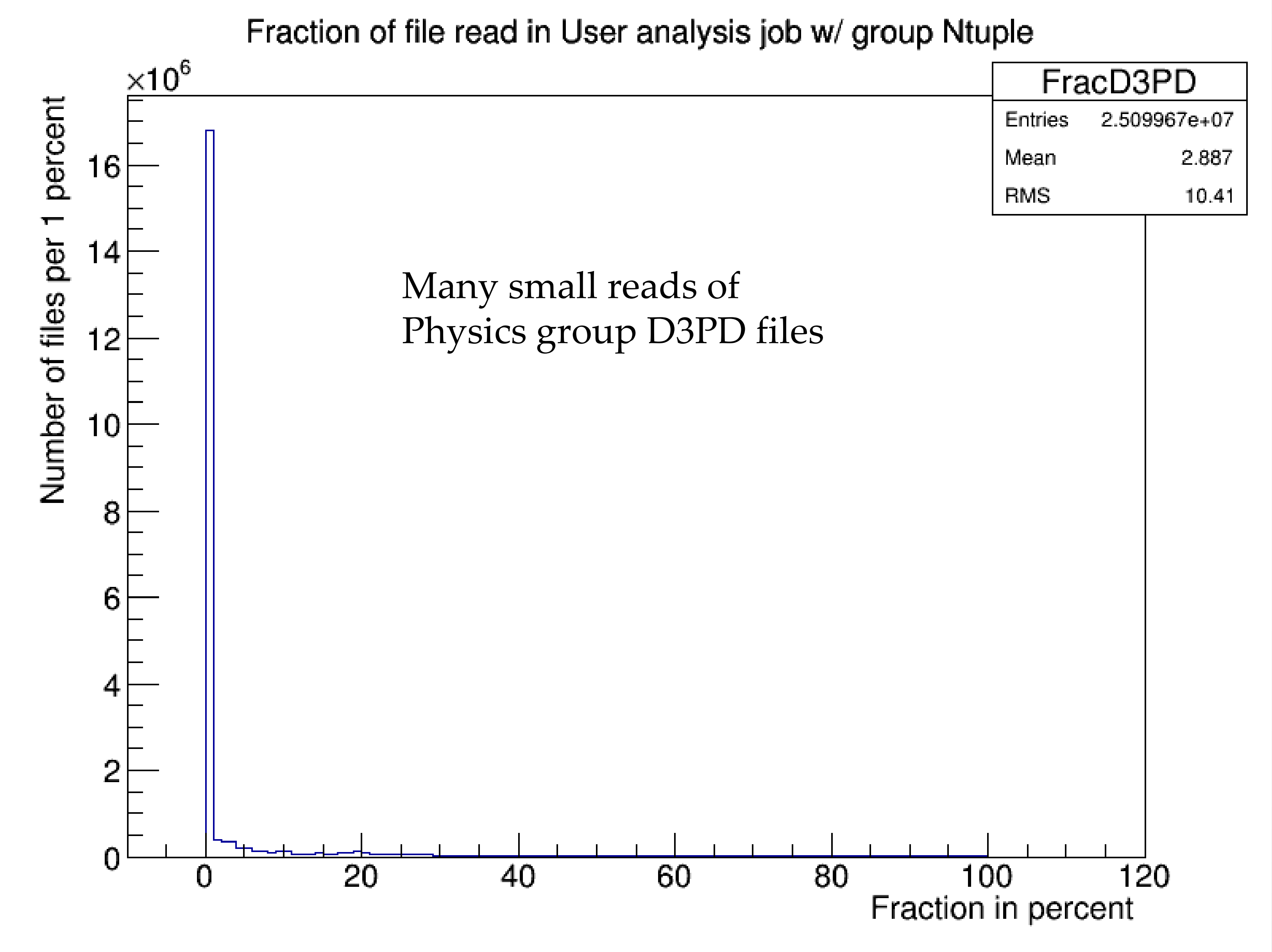
- All networking plans should include accounting for User Output

Analysis Sites



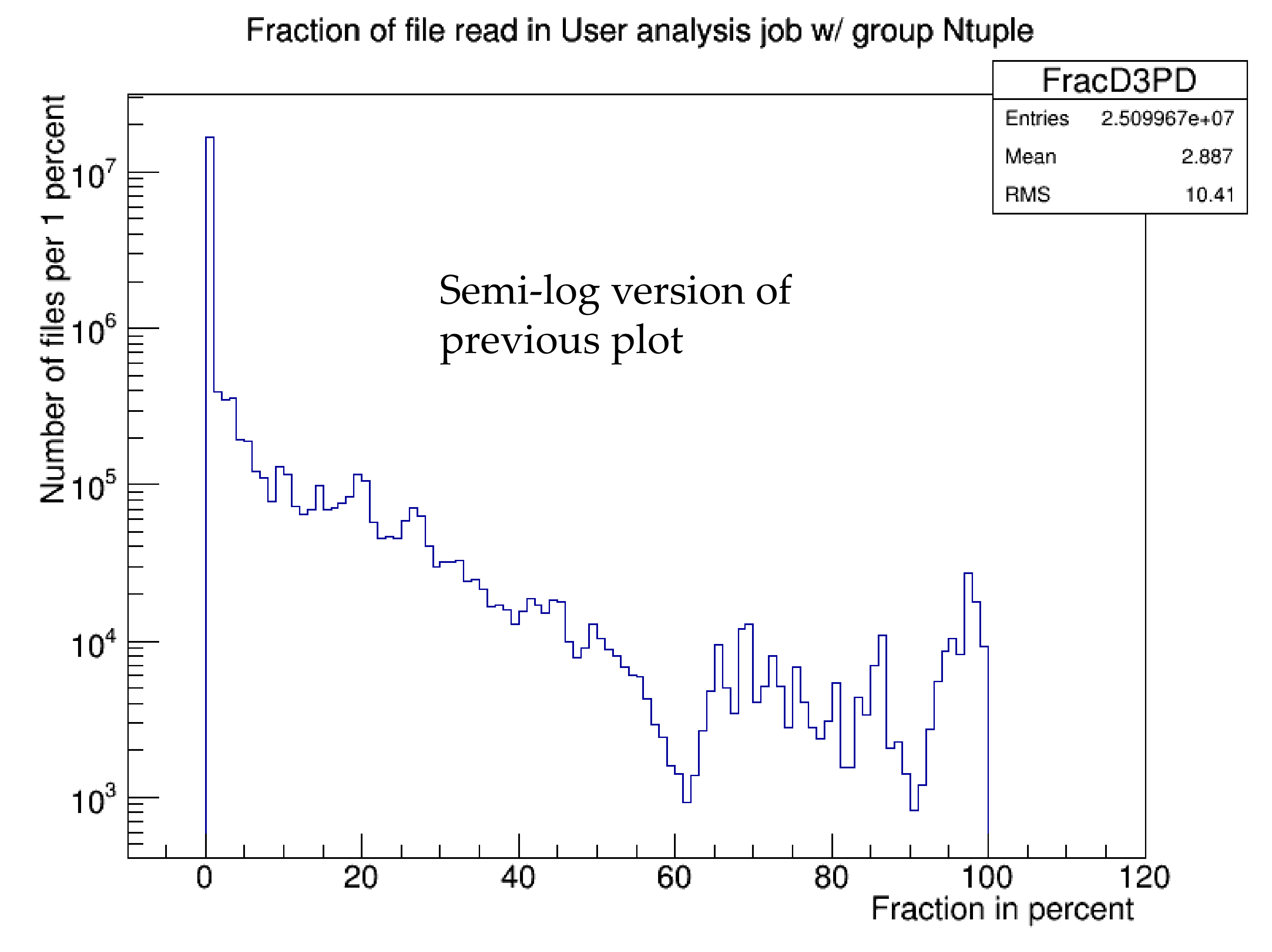
Jan 1 to April 30 2013 - MWT2 – direct read site

How much of D3PD do users read?



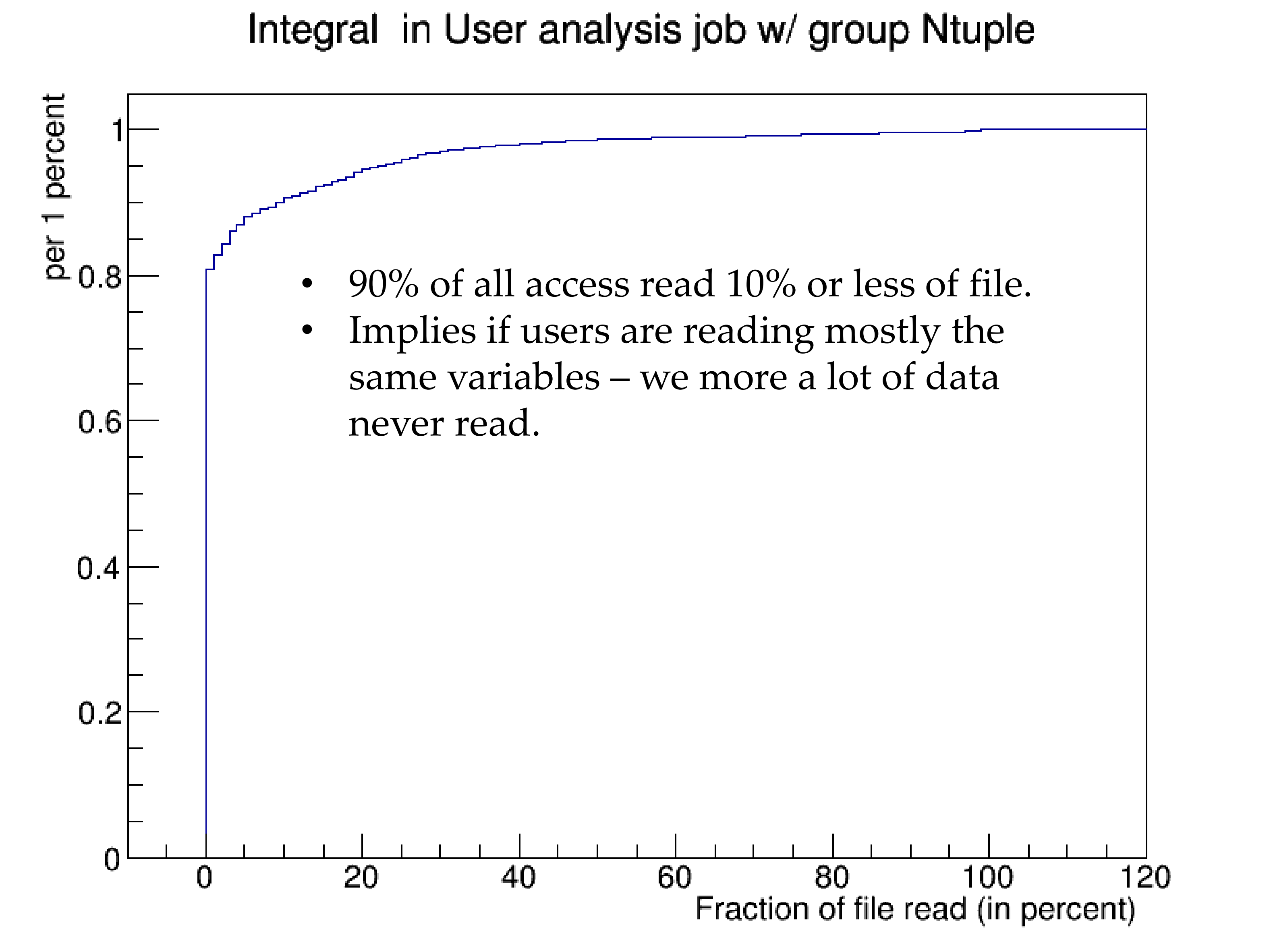
Many small reads of
Physics group D3PD files

How much of D3PD do users read?



Semi-log version of previous plot

How much of D3PD do users read?



- 90% of all access read 10% or less of file.
- Implies if users are reading mostly the same variables – we more a lot of data never read.

Interim summary user access of D3PD's

- Current D3PD's from the Physics groups typically very large
- They contain more Branches than the users typically use
- New analysis model is working to merge AOD and D3PD's
    - Implication – input files will contain much more information that the users actually read
    - Efficient Skim/Slim centralized service will be crucial
- Need some mechanism for capturing what variables the users are really using and then provide them files with mostly those variables.

Open questions needing further study

- How much PD2D data is read only once or twice after it has been replicated via PD2P?
- How long does a file stay popular (ie read frequently)?
  - We want to replicate the popular files and not the other ones?
- Can we reduce the amount of data in the Analysis files that is rarely read?
- Should we have caches for data files at the Tier 1 and Tier 2 sites ?
  - Do for analysis data what was done for database data and software releases (frontier/squid and cvmfs/squid).
  - What will it take to have partial event caches?
- Can we estimate the network growth needs when jobs and input data are in separate locations (WAN access from jobs)

Conclusions

- During busy times – 1 Million user analysis and D3PD production jobs per day, user analysis jobs are short duration.
- PD2P moves at peak ~ 600 TB/day and planned replication of similar files ~ 400 TB/day, User subscriptions ~ 200 TB/day
- 90% of NTUP datasets (by volume) are read 50 times or less. (according to PD2P values)
- Users still read a small fraction of the centrally produced D3PD's (NTUP's). Implies much of the data moved is not read
- What network issues will be seen w/ factor of 3 increase in Trigger Rate in 2015-2016?