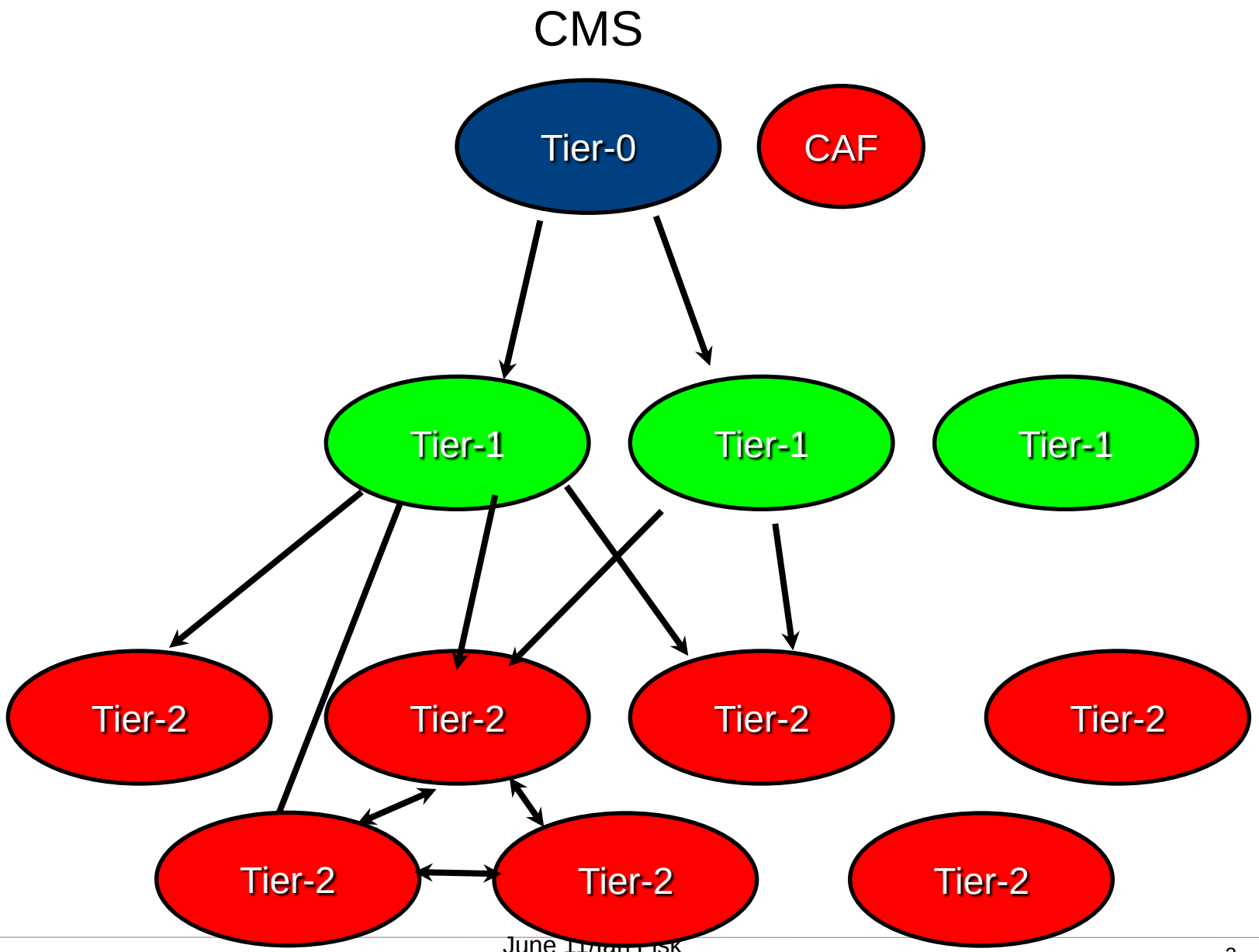
CMS Model and the Network

Ian Fisk

CMS Data Distribution Model

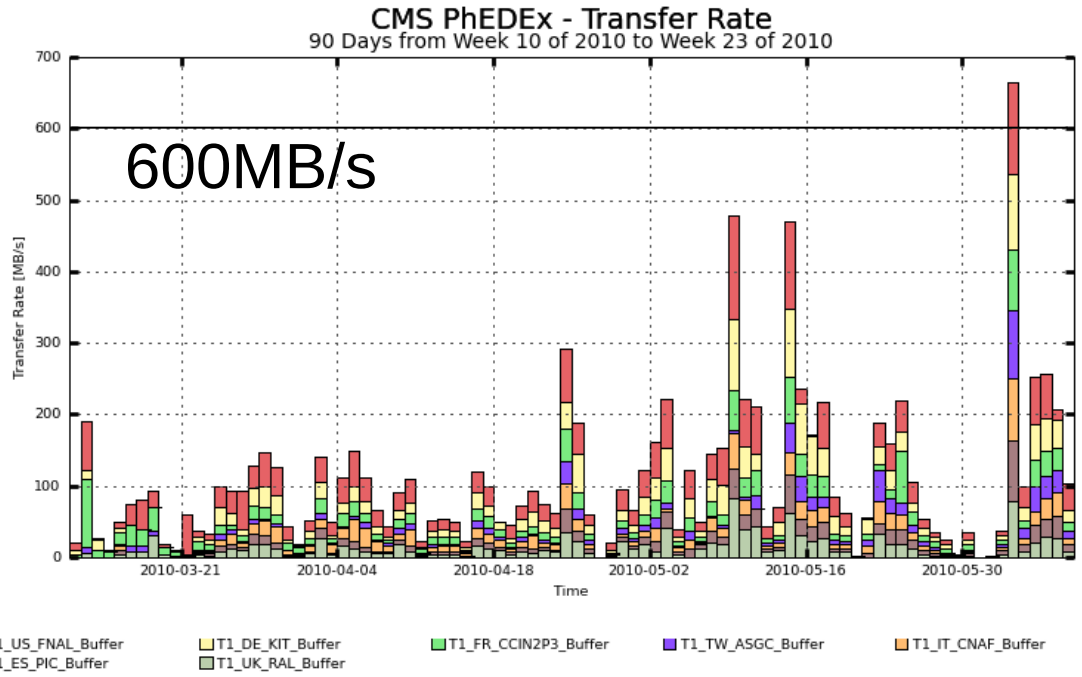


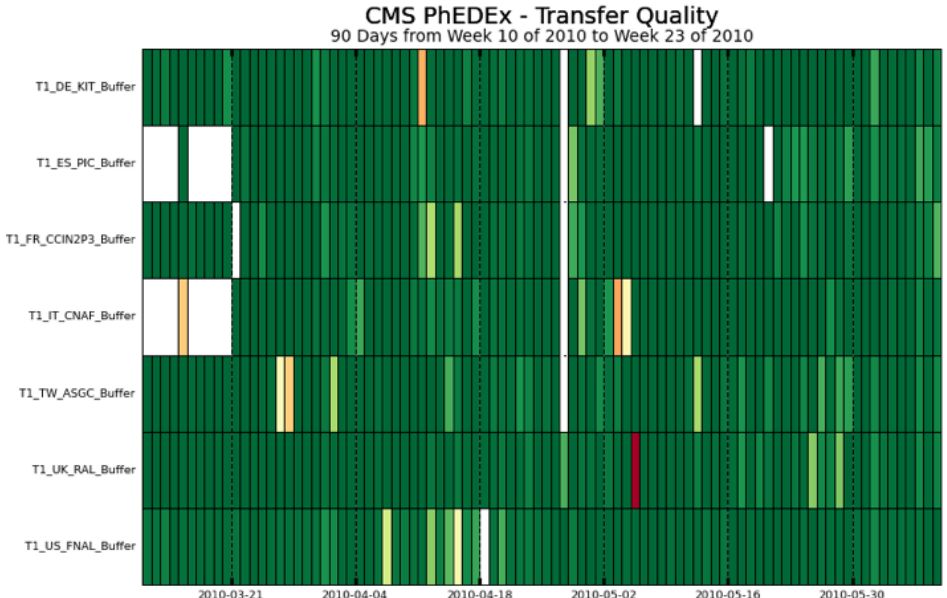
STARTING POINT

CONSTANTS	
Trigger rate	300 Hz
RAW Size	0.5MB
SimRAW	2MB
RECO size	0.5MB
AOD size	0.2MB
Total number of events	2360 MEvents
Total number of simulated events	2076MEvents
Overlap between PD	40, then 20%
Total size of RAW	3538 TB
Total size RECO	1180 TB
Total Primary AOD	472 TB

Tier-0 to Tier-1

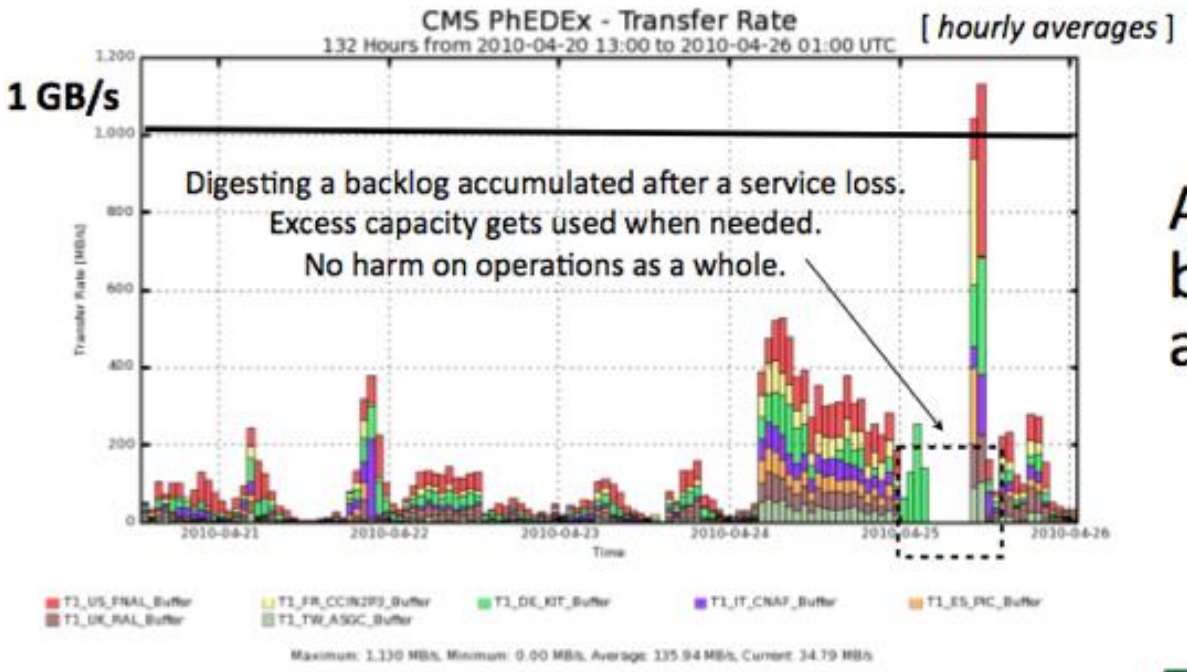
- **CERN to Tier-1 Average since beginning of 2010 run**





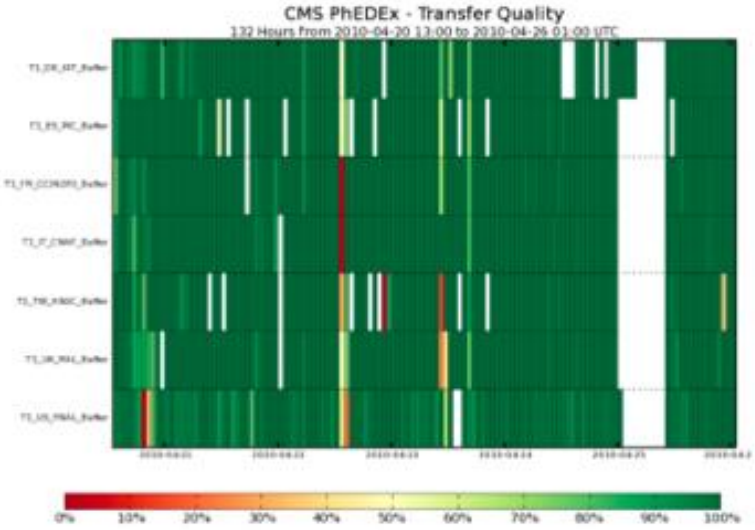
- **Transfer Quality is excellent**

Recovery Tier-0 to Tier-1



A nice example of a backlog digestion after a service incident...

... with a very high transfer quality before, and again soon after it recovered.



CERN to Tier-1

- Rate is defined by the accelerator, the detector and the data distribution policy
  - Livetime of the machine is lower than we expect for the future ↓
    - System is specified to recover between fills
  - Data is over subscribed ↑
    - Will continue as resources allow
  - RAW event size is smaller than our estimates ↓
  - Event rate is defined by the physics program →
- We expect the average rate from CERN to Tier-1s will increase, but we expect the rate is predictable and until there is a fundamental change in the input it should match the planning
  - Peaks are roughly what we would expect

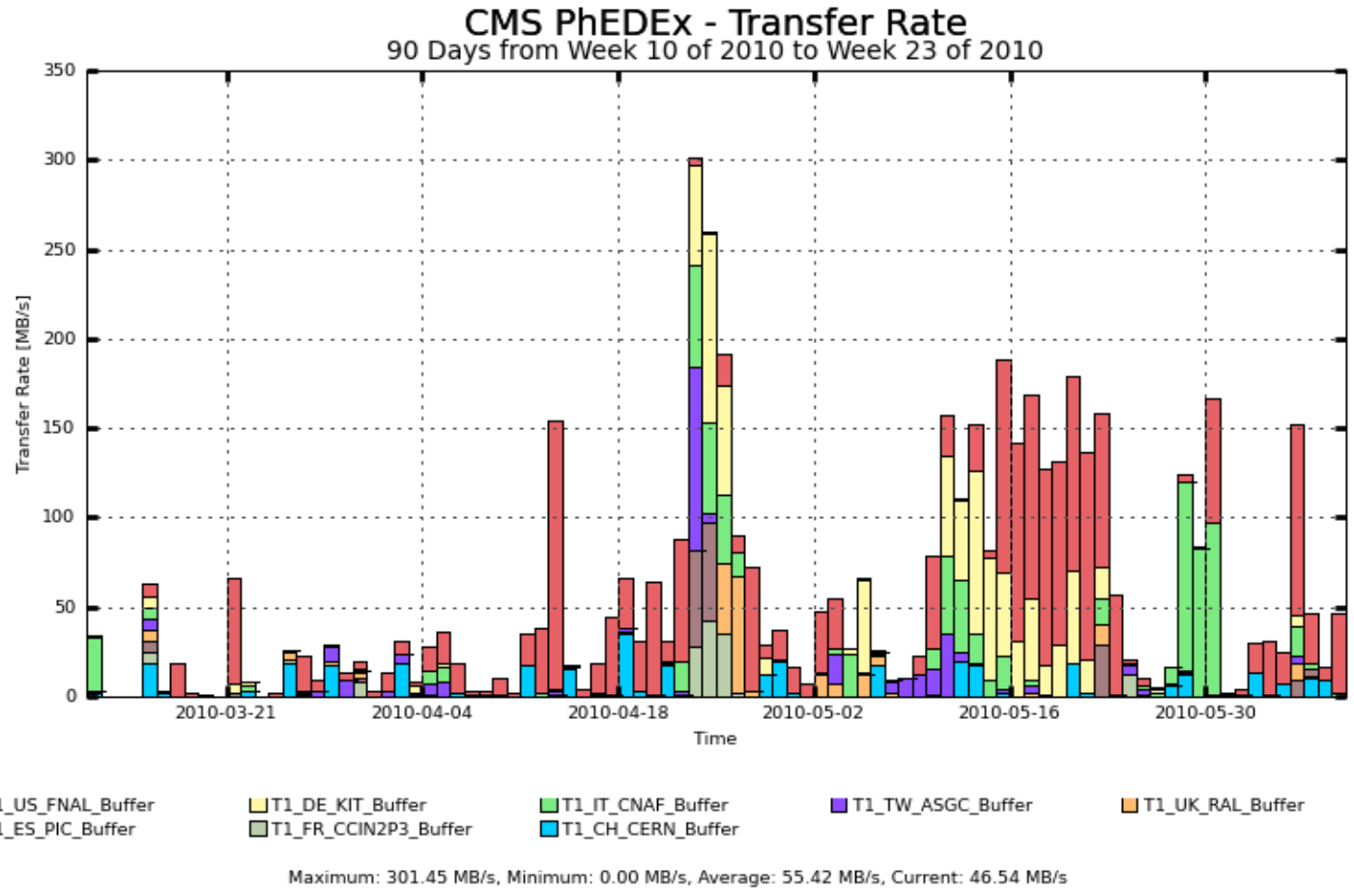
1. Exporting custodial data: methodology

- T0—> T1s : exporting FEVT
  - BW=(RAW+RECO) x Trigger frequency x (1+overlap factor). For the chosen parameters, this yields:

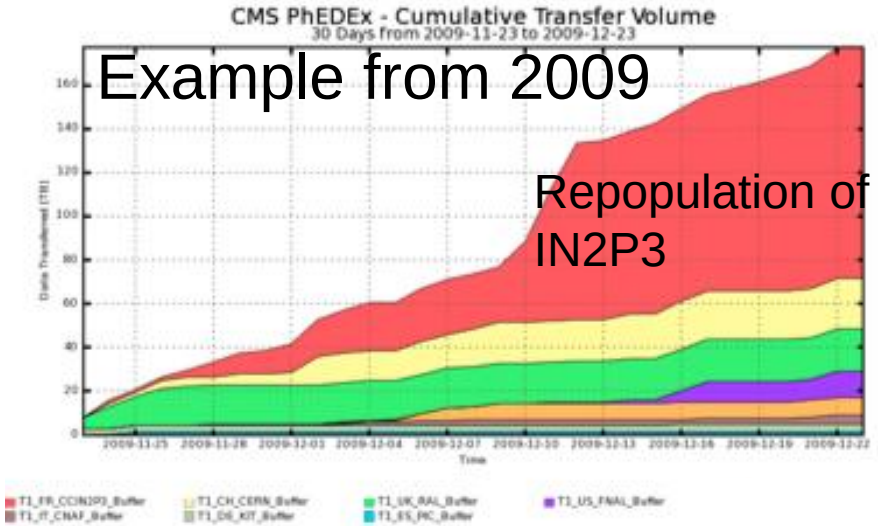
  BW= 1.0MB x 300Hz x 1.4 = 420 MB/sec, or 3.3 Gb/sec. We expect at least 2 copies of the data from CERN for 2010 so 6.6Gb/s
- Each T1 receives a share according to its relative size in CPUs
- Proportional to the trigger rate, event size and Tier-1 size
- **In 2010 We will continue to oversubscribe the data as resources allow**

Somewhat More Interesting T1 to T1

- **Tier-1 to Tier-1 transfers are used to replicate raw, reco and AOD data, recover from losses and failures at Tier-1 sites**

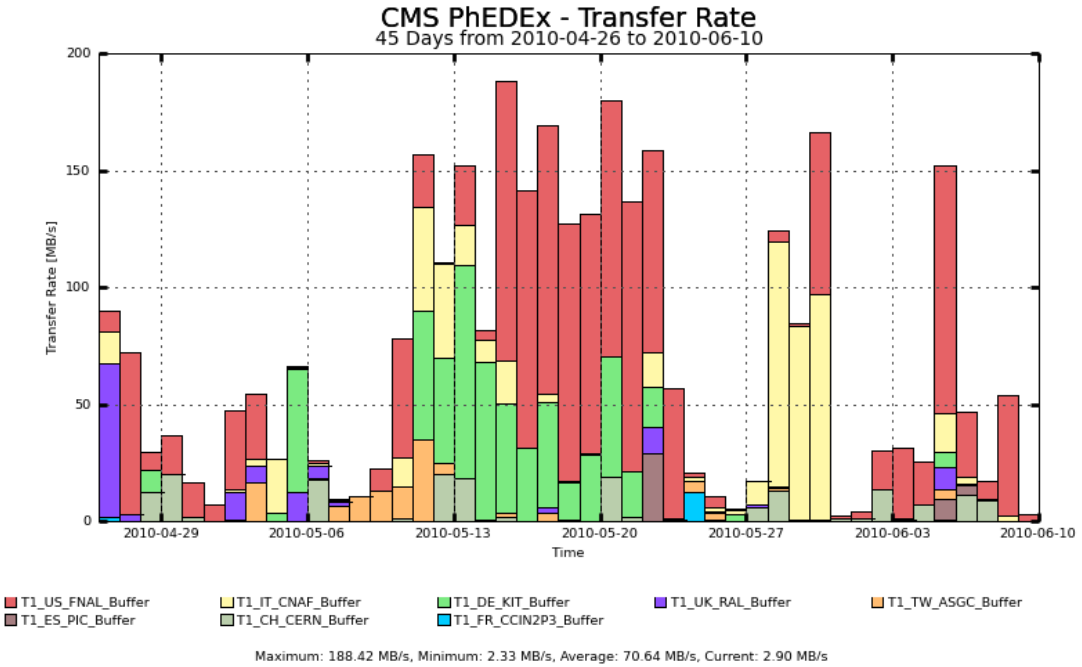


Tier-1 to Tier-1



Destination Site	Total Transfer Volume [TB]
T1_DE_KIT	0.39
T1_ES_PIC	1.51
T1_FR_CCIN2P3	105.15
T1_IT_CNAF	4.55
T1_UK_RAL	19.27
T1_US_FNAL	12.29
	143.16

24 Hour averages
since in 2010
data running T1 - T1

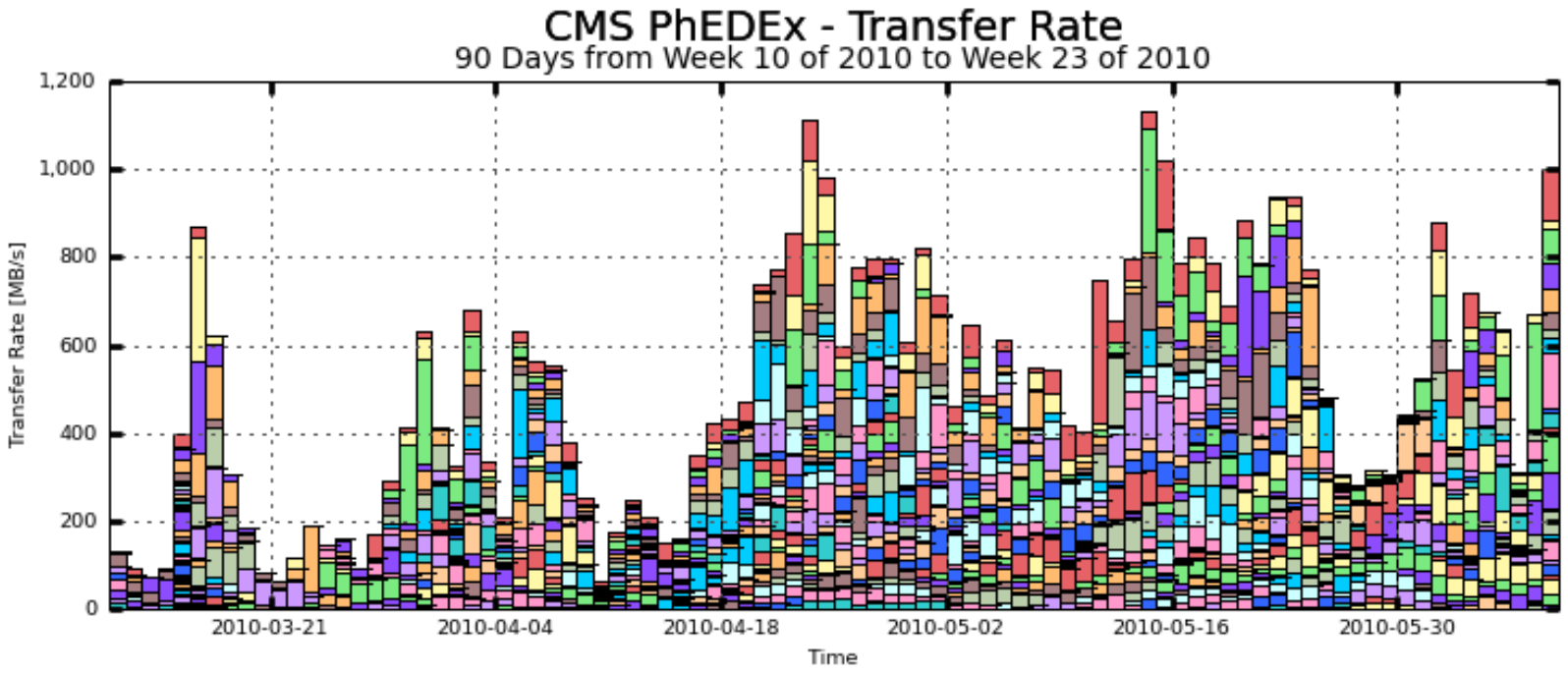


Tier-1 to Tier-1

- The CMS plan currently is ~ 3.5 copies of the AOD
  - After an refresh of the full sample of a year's running this is 1.6PB of disk to update
  - Using 10Gb/s that takes 20 days.
  - Achieving 30Gb/s is a week
    - The Computing TDR had 2 weeks
    - In 2010 we will also be replicating large samples of RECO
- Recovering from a data loss event at a Tier-1 is more challenging because the data might be coming from one place
  - Could also take longer with the normal risk of double failure

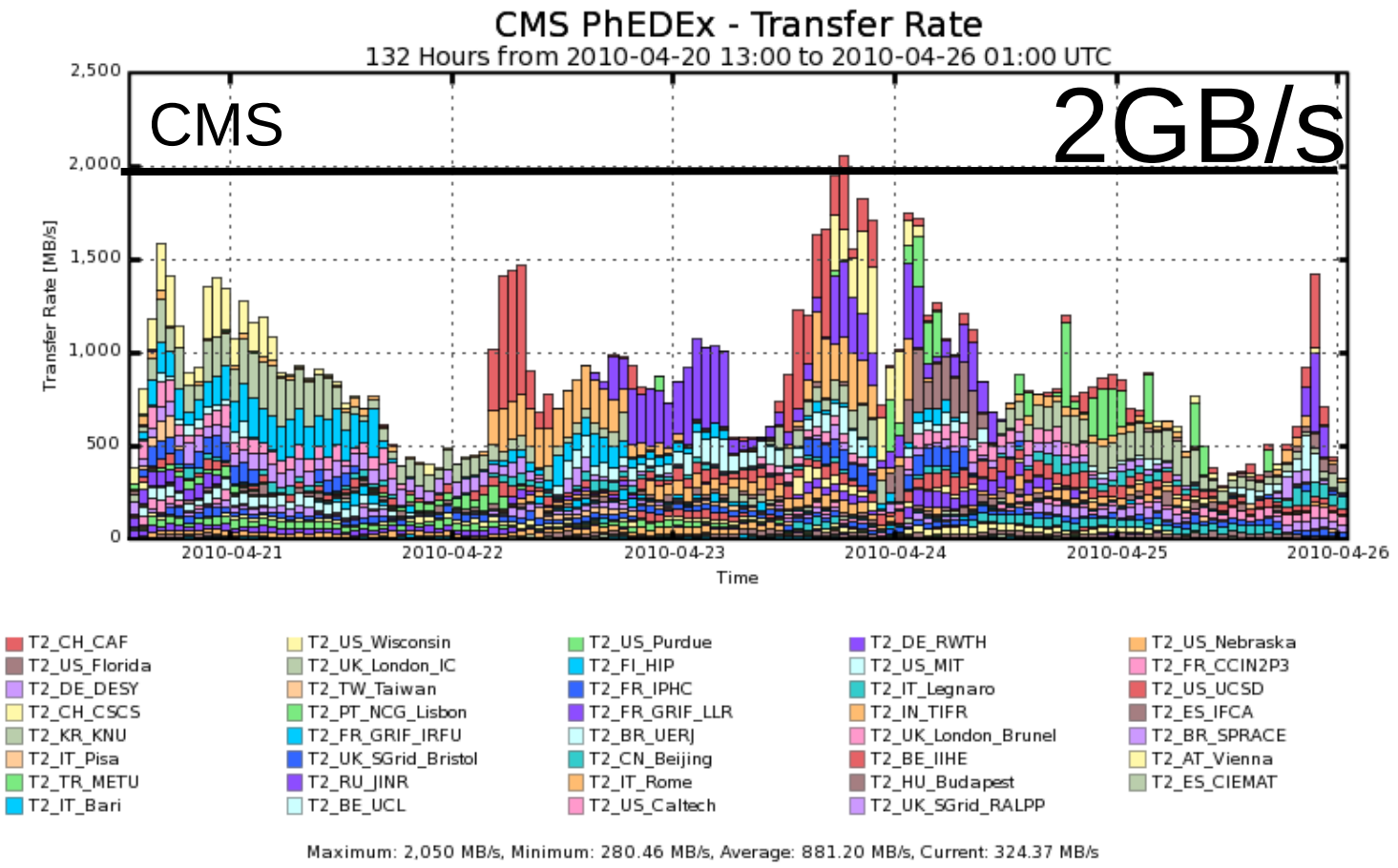
Tier-1 to Tier-2

- Getting more interesting
- CMS is very close to completing commissioning the full mesh of Tier-1 to Tier-2 transfers at a low rate
  - Working on demonstrating more links at 100MB/s
  - Daily average exceeding 1GB/s



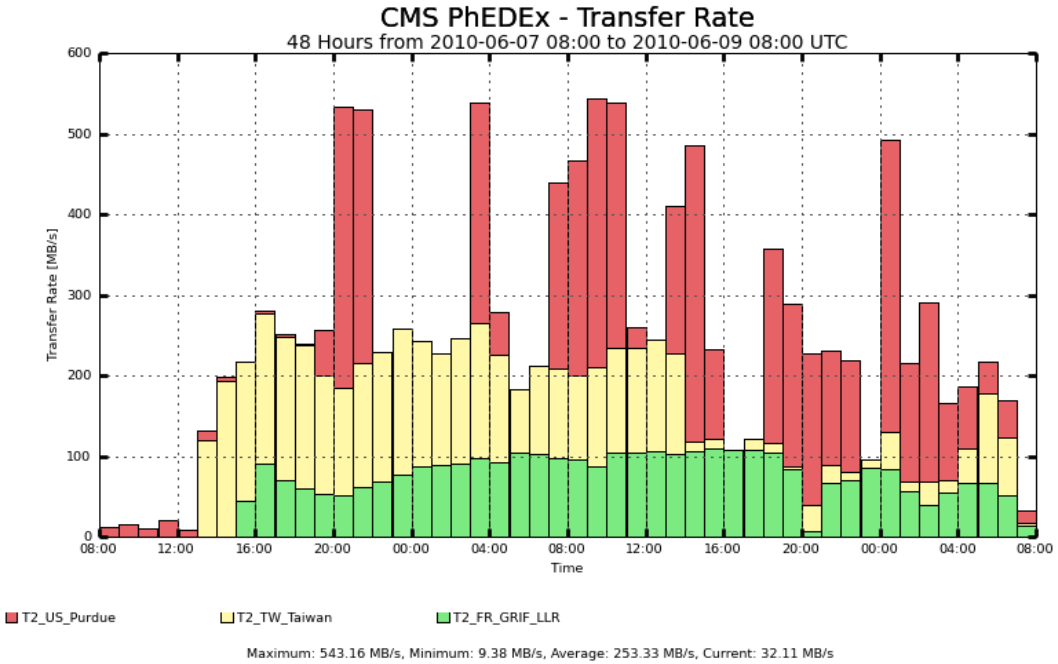
Tier-1 to Tier-2

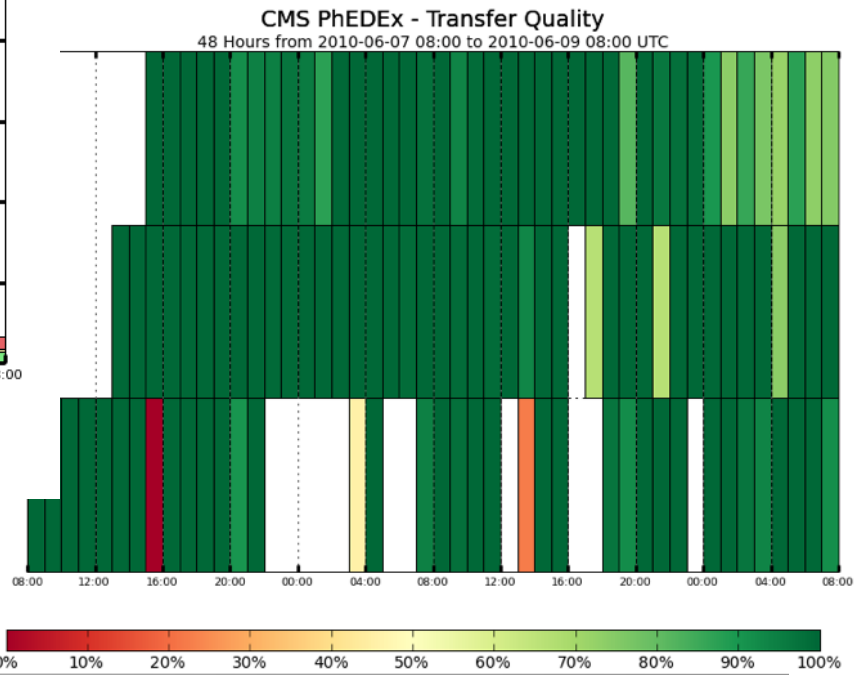
- **Looking at hourly averages is more interesting**
  - **See already several examples of 500MB/s for bursts**



This week: a case study

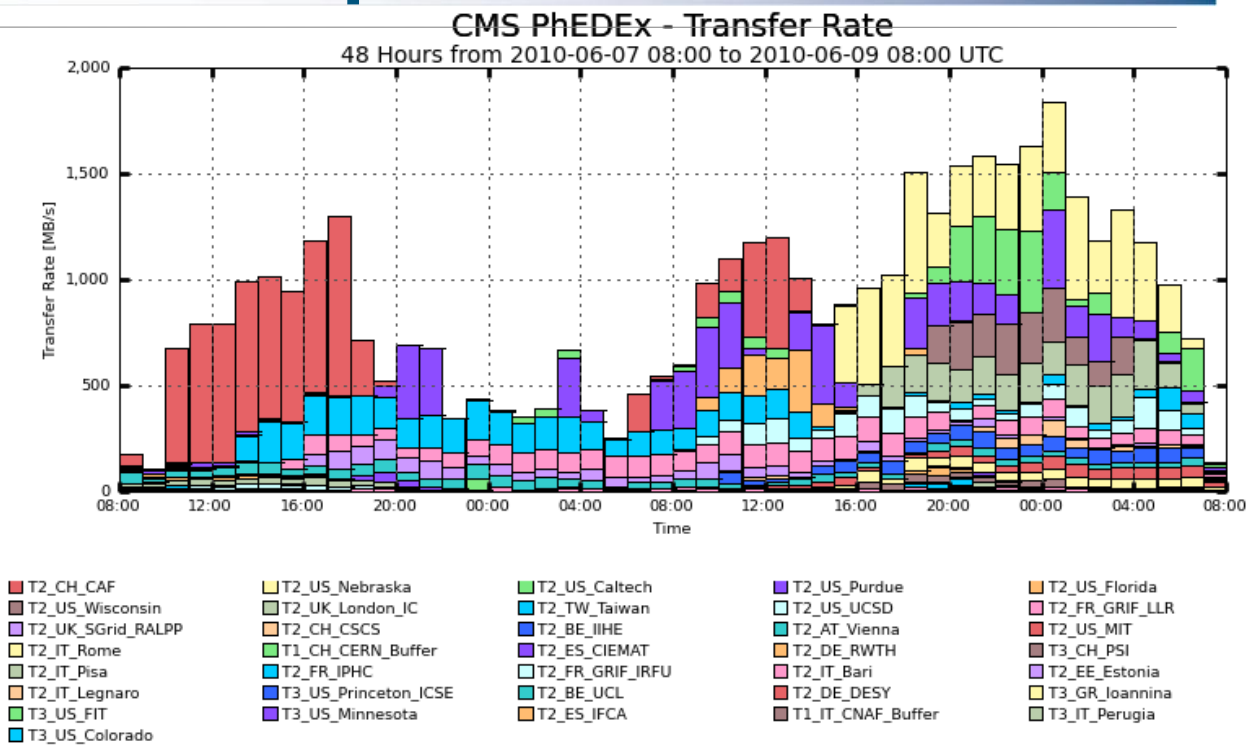
- CMS produced a 35TB skim of the data sample after a reprocessing pass
  - Skim took about 36 hours to produce
  - Data is then subscribed to analysis users

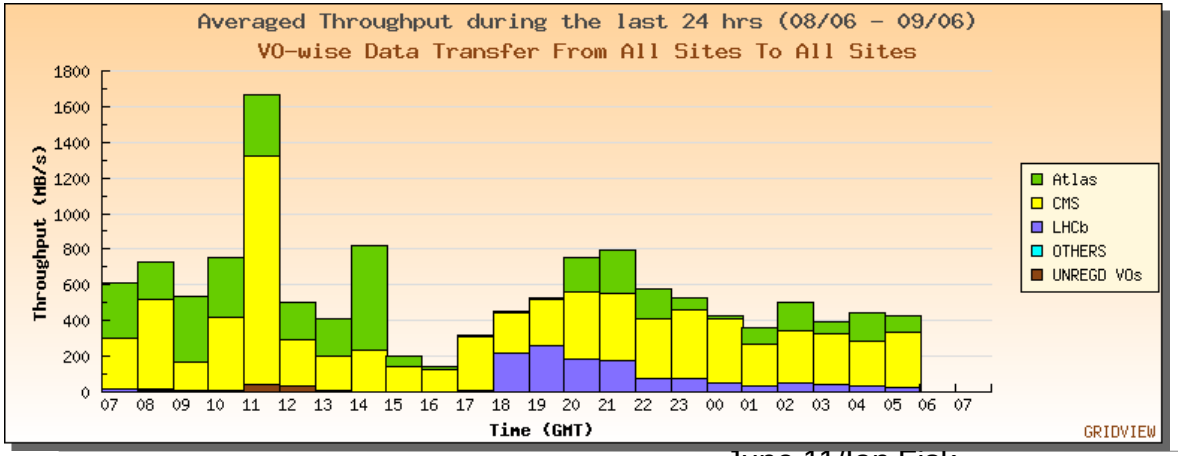




Total Export Rate

- **Source site is exporting data at more than 1.5GB/s (12Gb/s)**
- **Higher than CERN for all 4 VOs**





Tier-1 to Tier-2

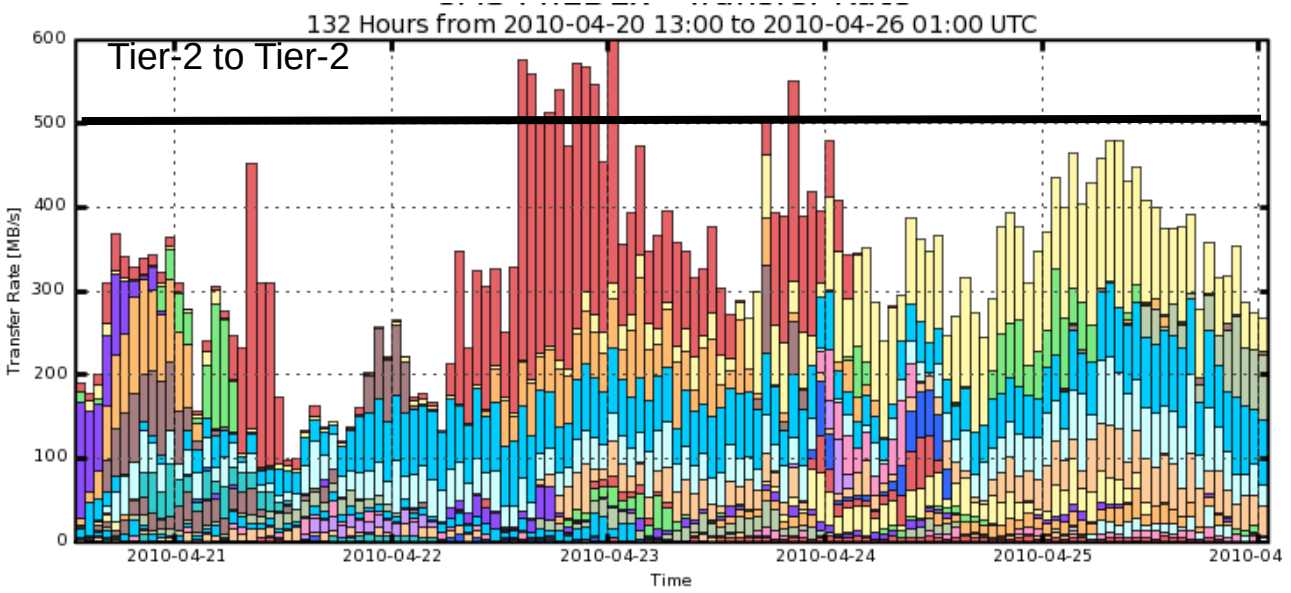
- In CMS Tier-1 to Tier-2
  - Driven by group and user requests
  - Already we have a 35TB sample users are trying to replicate and access
    - Somewhat unwieldy generally
  - Full mesh topology is challenging because there are oceans and heterogeneous environments in the way
- Data is refreshed frequently and even large samples may need to refreshed
  - 500MB/s is already demonstrated
    - Hardest use case is going to be refreshing data after reprocessing
      - Typically comes from one place and needs to go to many

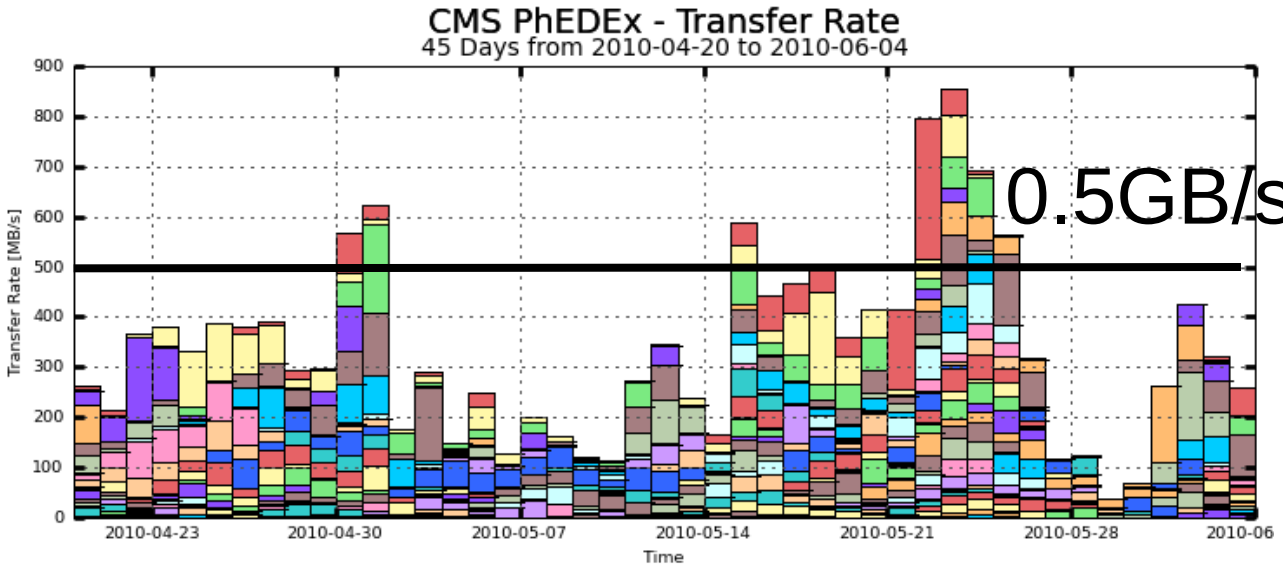
Tier-1 to Tier-2

- Data from Tier-1 to Tier-2 is driven by event selection efficiency, frequency of reprocessing, level of activity
  - All of these are harder to predict, but translate into physics potential
- The connections between data production sites and analysis tiers needs to allow prompt replication
  - CMS is currently replicating 35TB of data that took 36 hours to produce to 3 sites (~100TB)
  - These bursts are not atypical

Tier-2 to Tier-2

- **Tier-2 to Tier-2 transfers are relatively new, but a growing issue in CMS**
  - **Started with trying to replicate group produced data between supporting Tier-2 sites**





Tier-2 to Tier-2

- Making sure transfer links work has been a lot of effort
  - Many permutations between Tier-2 sites
- Tier-2 data is always on disk
- Many Tier-2s have good WAN connections
- Sometimes geographically close

Looking Forward

- Transfers between Tier-2s is already a big step toward flattening the hierarchy
  - Should decrease the latency and decrease load on the Tier-1s, but it makes more unstructured use of the networks
- CMS has been working to optimize the number of objects we read from the files and the order they are requested
  - Initially this was intended to make better use of local storage
  - Initial indications are that one can get reasonable CPU efficiency for applications reading data directly over long distances

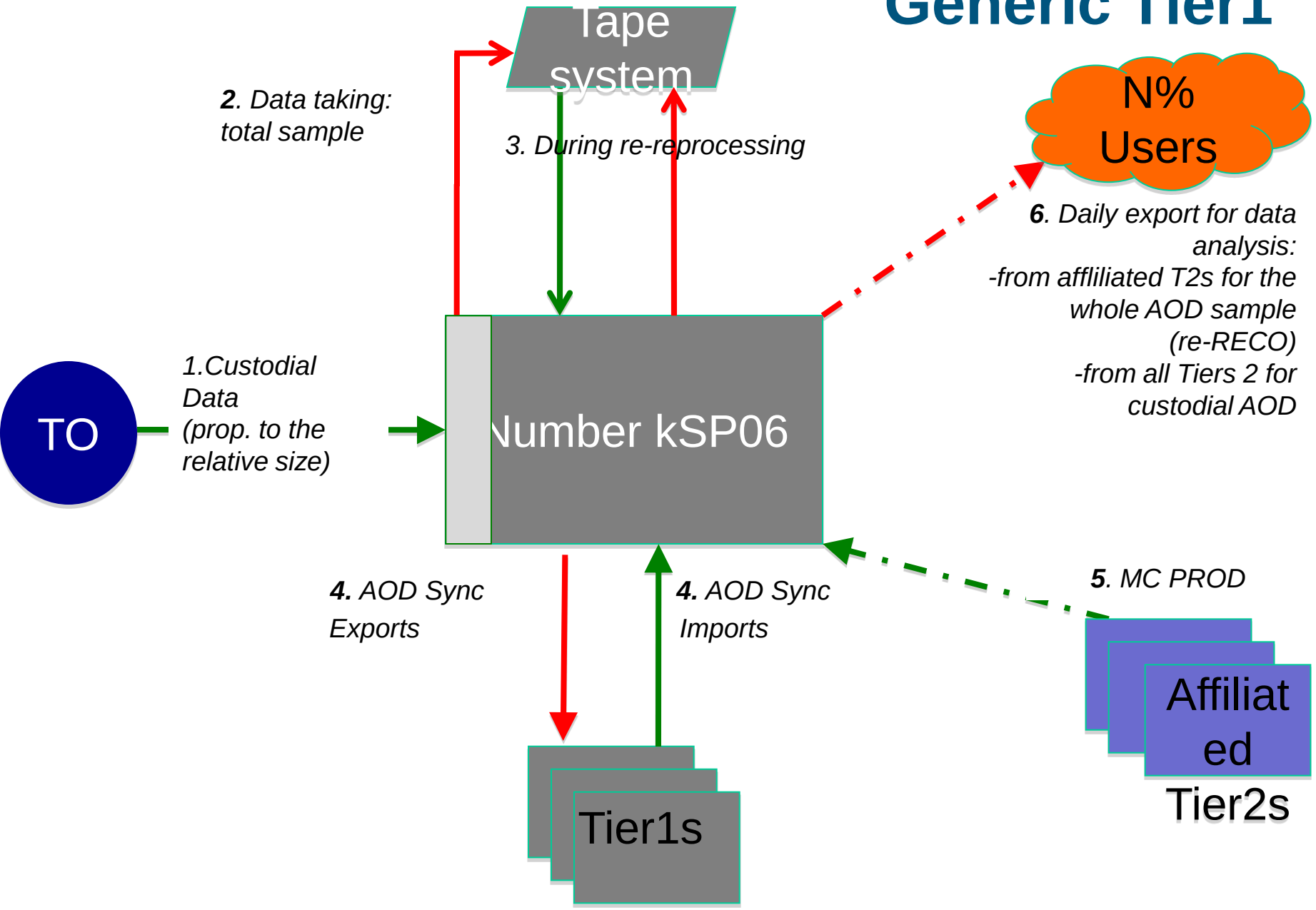
Looking Forward

- How data access over the WAN will evolve is under discussion
  - ALICE has similar functionality now, it's not clear if a location unaware solution would be efficient or desirable in CMS
  - Might be a reasonable backup channel for data
  - Might be more
- It's also not completely clear that all applications of reading over the WAN necessarily cause the networking to go up
  - Moving whole files, if you only need a fraction of the objects might not be efficient utilization.

Outlook

- CERN to Tier-1s is driven by the detector and the accelerator
  - Somewhat predictable as networking scales with instantaneous luminosity
- Tier-1 to Tier-1 is driven by need to replicate samples and to recover from problems
  - See reasonable bursts that will grow with the datasets. Bursts scale as integrated luminosity
- Tier-1 to Tier-2 is driven by activity and physics choices
  - Large bursts already. Scale as activity level and integrated lumi
- Tier-2 to Tier-2 is ramping up.

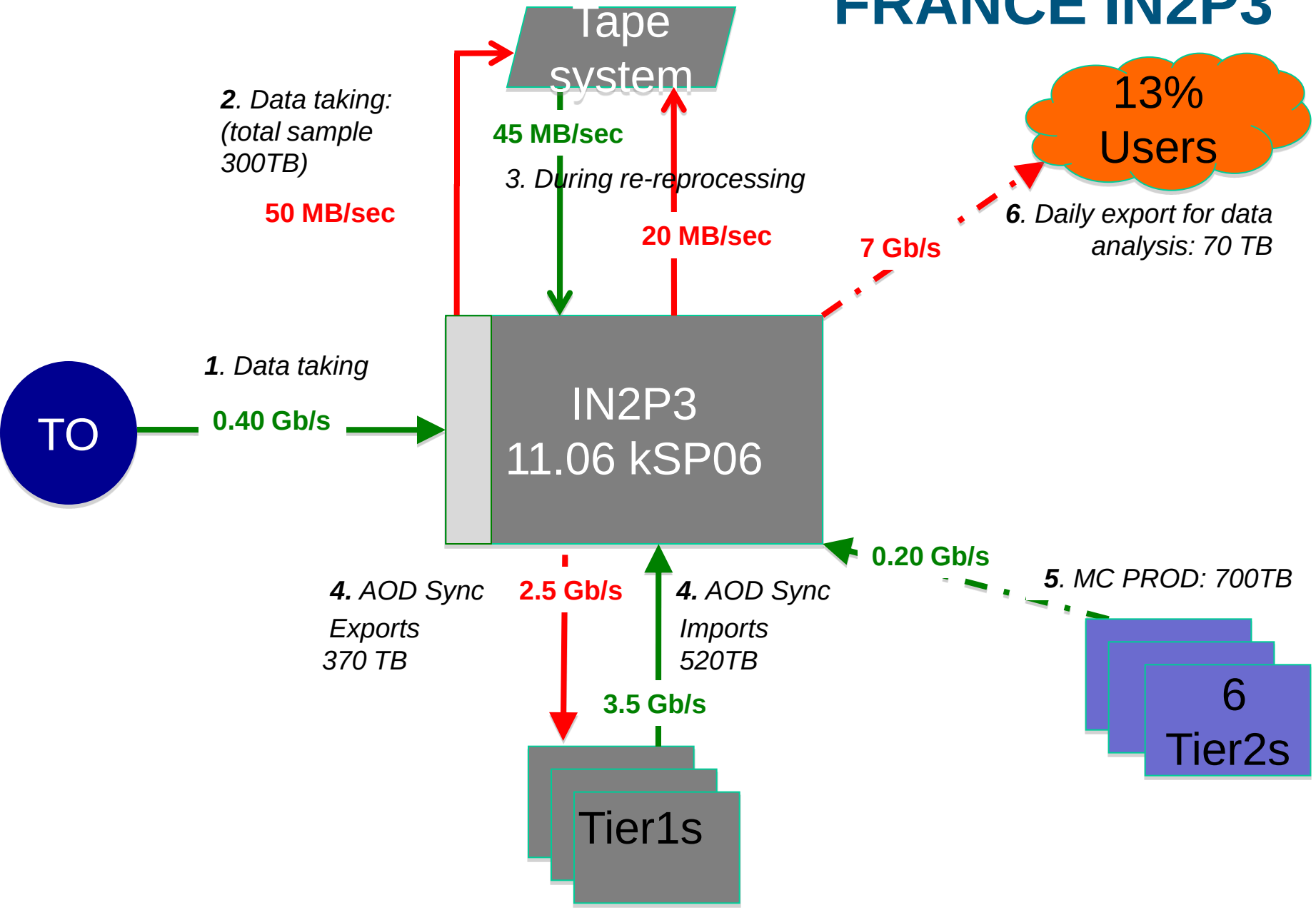
Generic Tier1



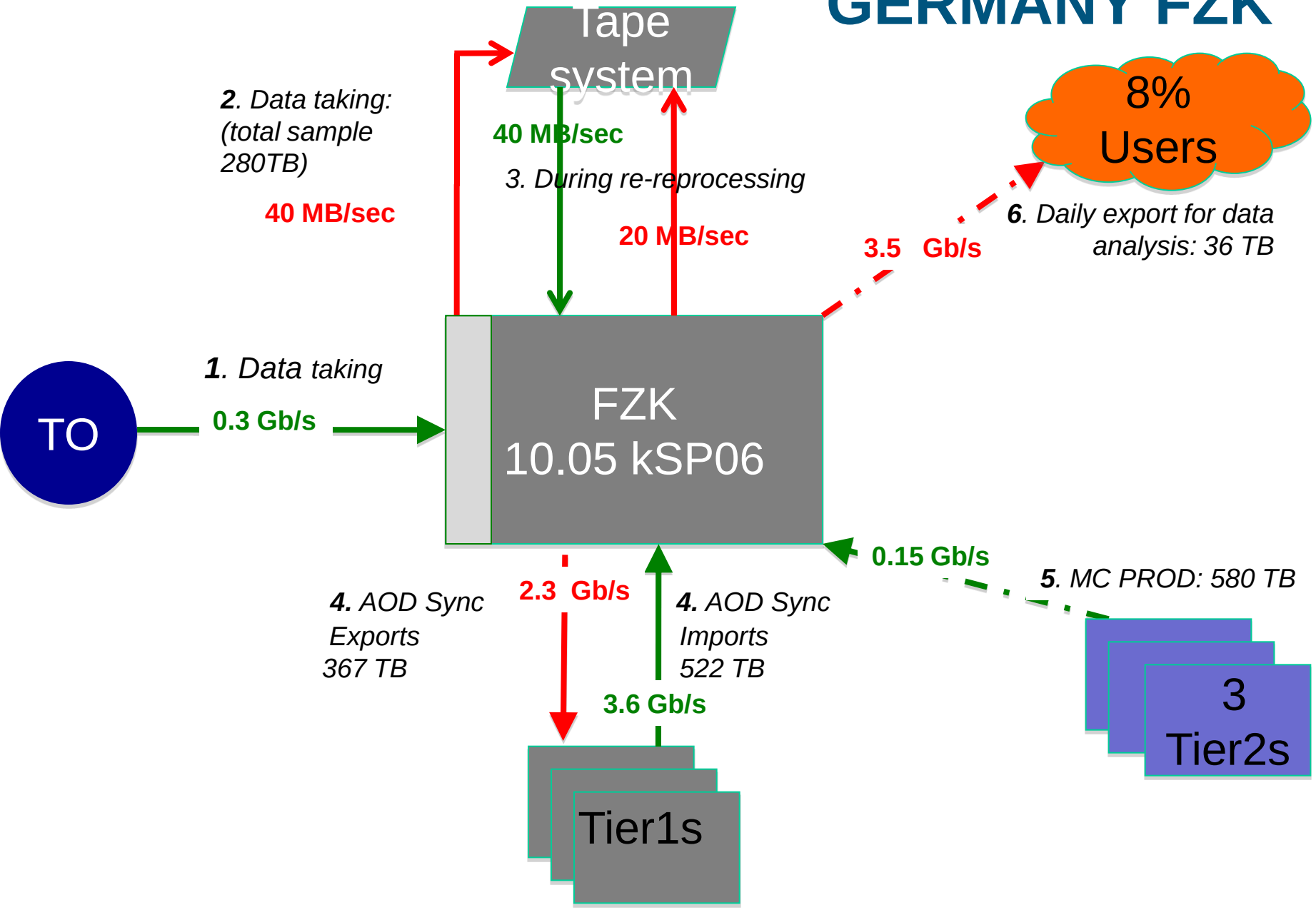
The results

- 1 slide per regional Tier1
- Pledged Cores are for 2010
- Remember: this are really raw values
- Links:
  - Solid line: sustained bandwidth (data taking and re-processing periods ONLY)
  - Broken line: peak bandwidth (may happen at any time: numbers shown is the total if it all happens at the same time)
- For each Tier 1, the fraction of served users for analysis is a combination based on
  - Relative size T2s for analyzing the share of 1srt AOD at considered Tier1, number of users based on the number of supported physics groups
  - Relative size of T1 for analyzing the full AOD

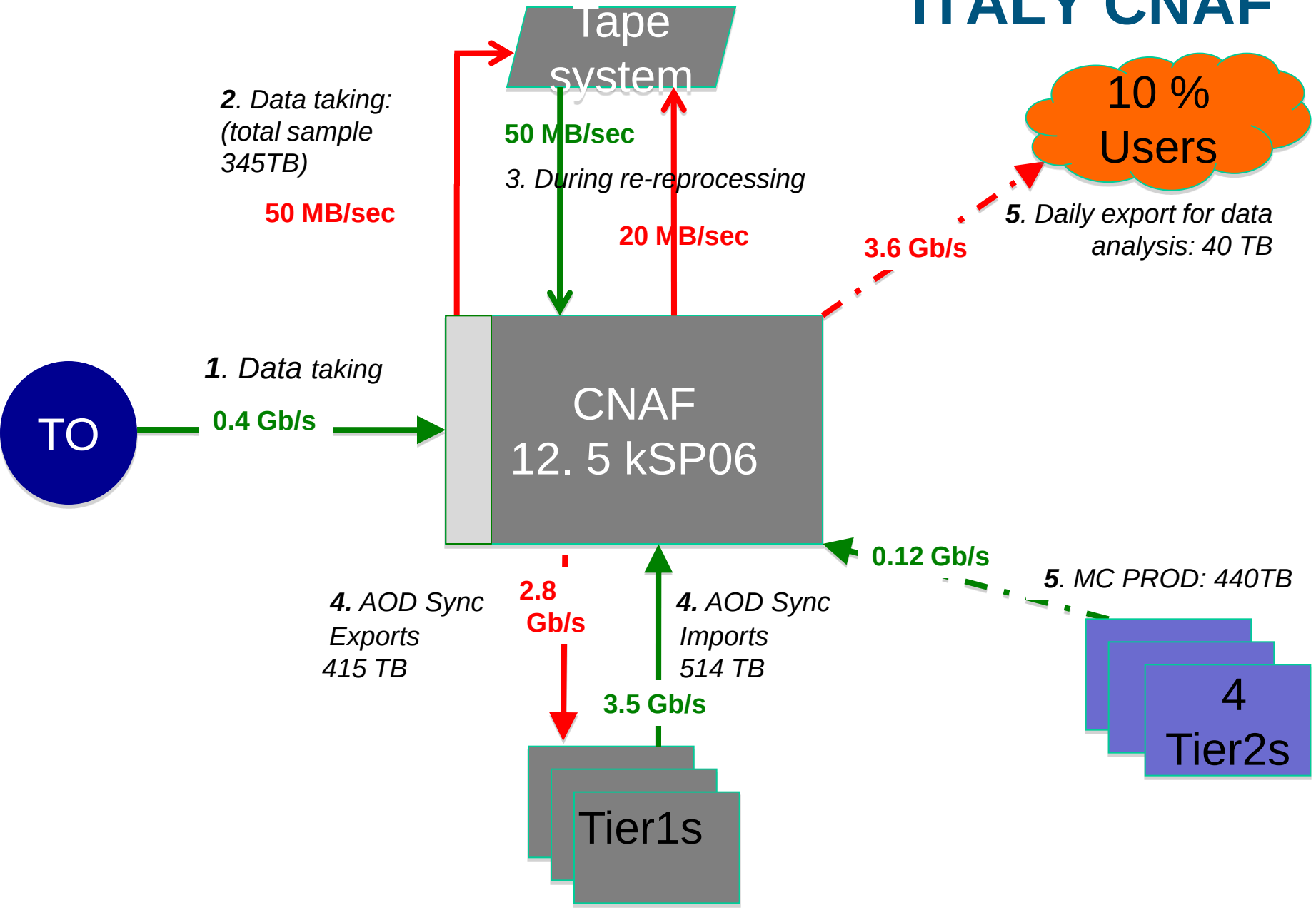
FRANCE IN2P3



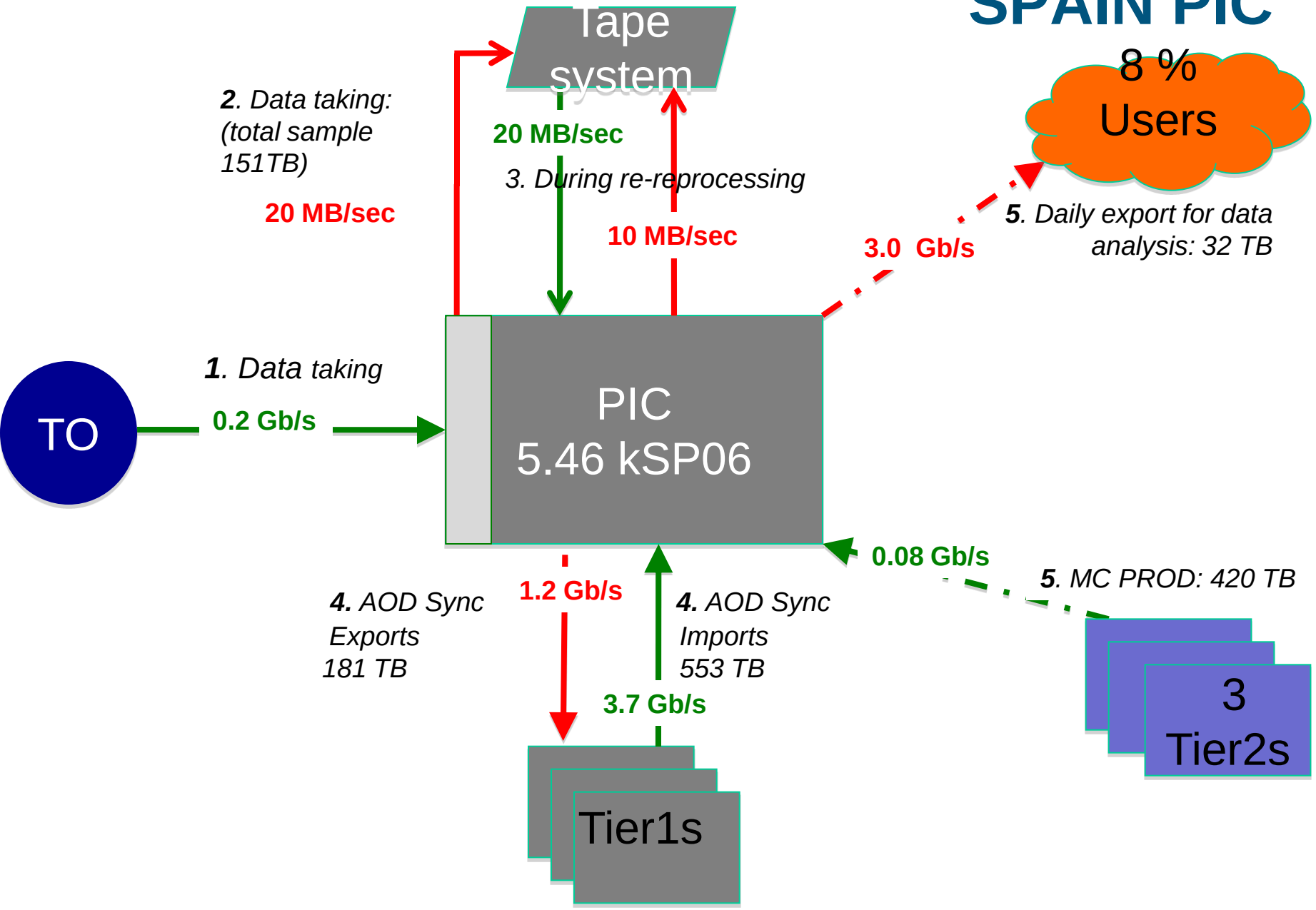
GERMANY FZK


Tape system
8% Users
2. Data taking: (total sample 280TB)
40 MB/sec
40 MB/sec
3. During re-reprocessing
20 MB/sec
3.5 Gb/s
6. Daily export for data analysis: 36 TB
TO
1. Data taking
0.3 Gb/s
FZK
10.05 kSP06
0.15 Gb/s
5. MC PROD: 580 TB
4. AOD Sync Exports 367 TB
2.3 Gb/s
4. AOD Sync Imports 522 TB
3.6 Gb/s
3 Tier2s
Tier1s

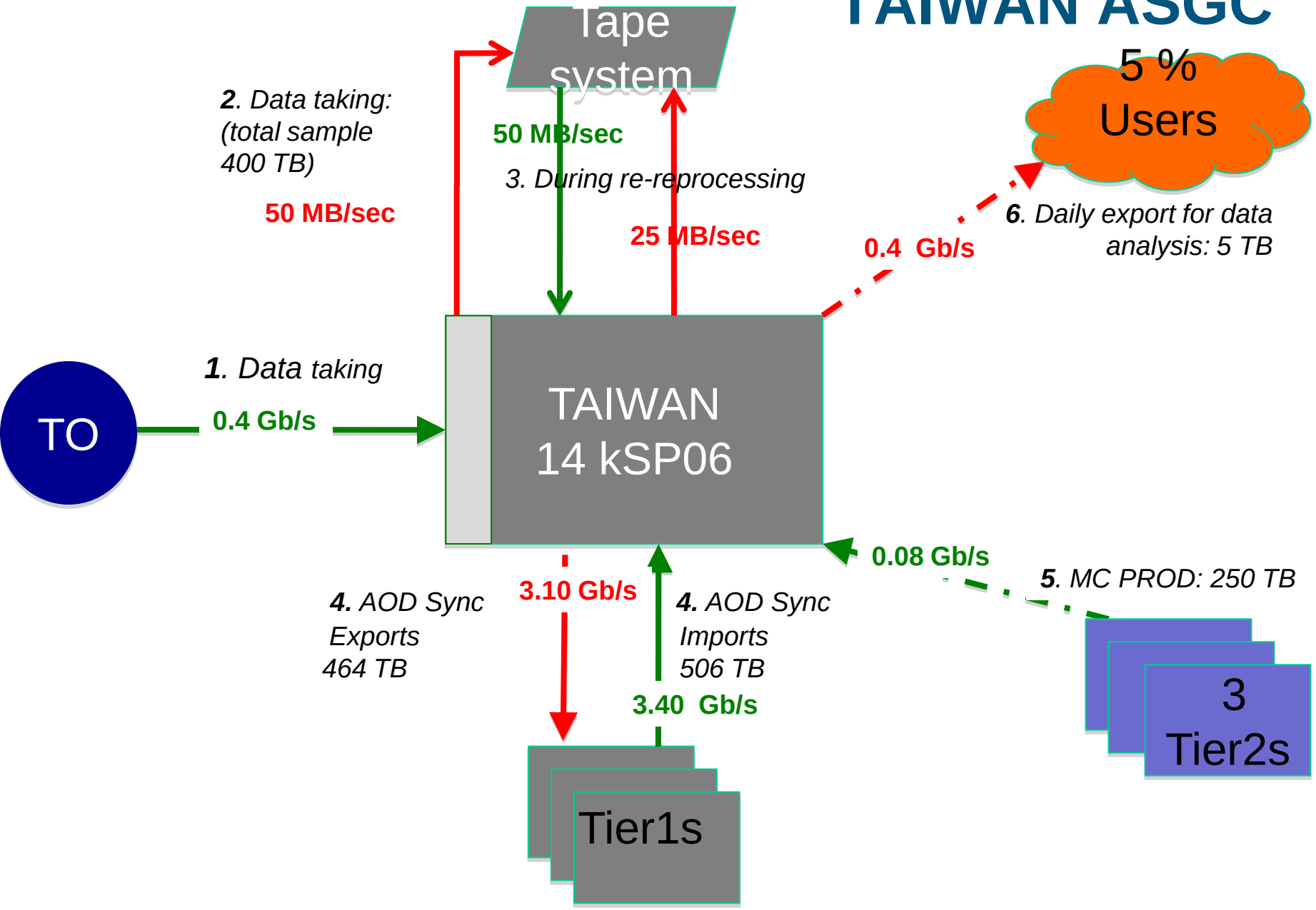
ITALY CNAF


Tape system
2. Data taking: (total sample 345TB)
50 MB/sec
50 MB/sec
3. During re-reprocessing
20 MB/sec
10 % Users
3.6 Gb/s
5. Daily export for data analysis: 40 TB
TO
1. Data taking
0.4 Gb/s
CNAF
12. 5 kSP06
0.12 Gb/s
5. MC PROD: 440TB
4. AOD Sync Exports 415 TB
2.8 Gb/s
4. AOD Sync Imports 514 TB
3.5 Gb/s
4 Tier2s
Tier1s

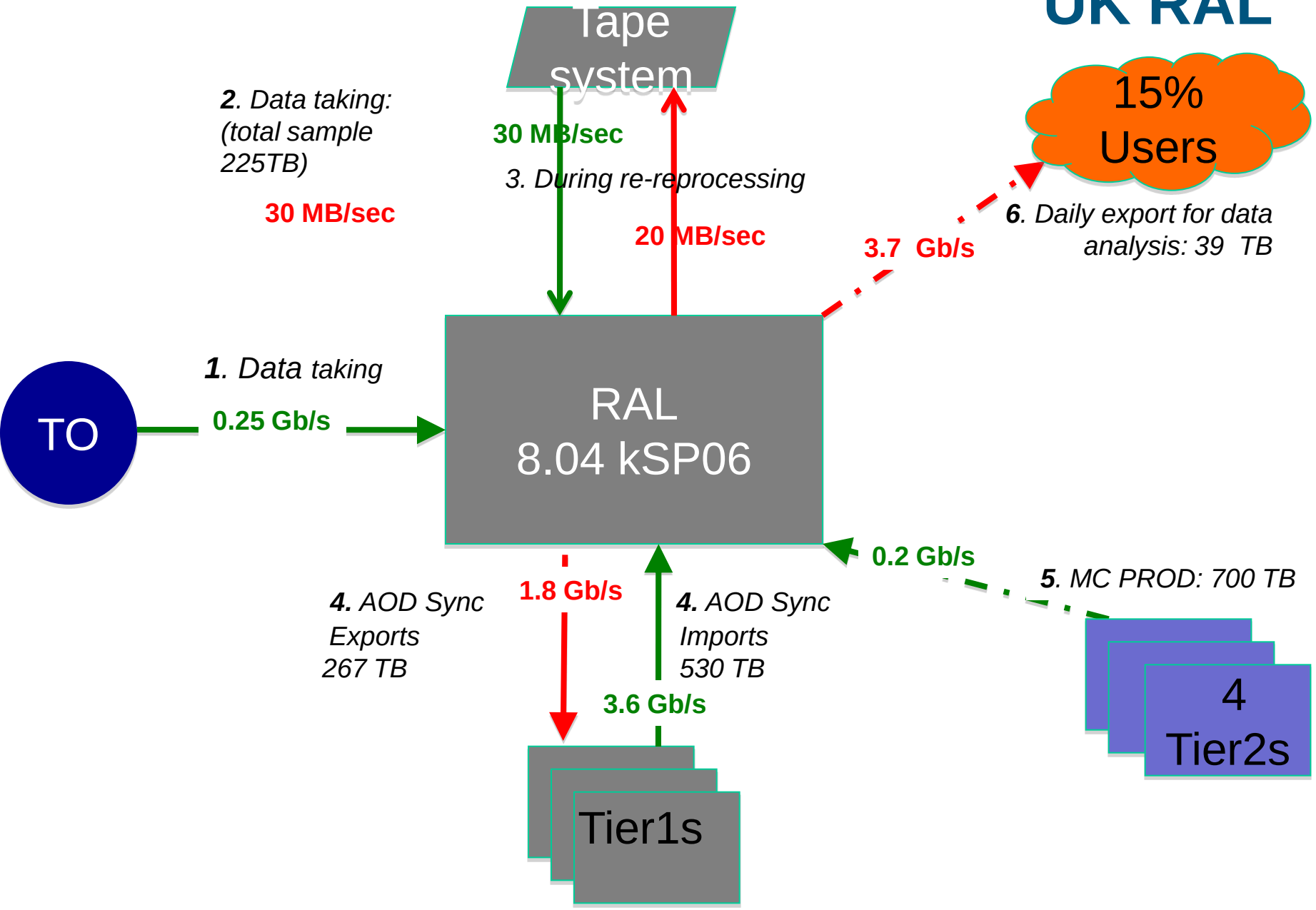
SPAIN PIC

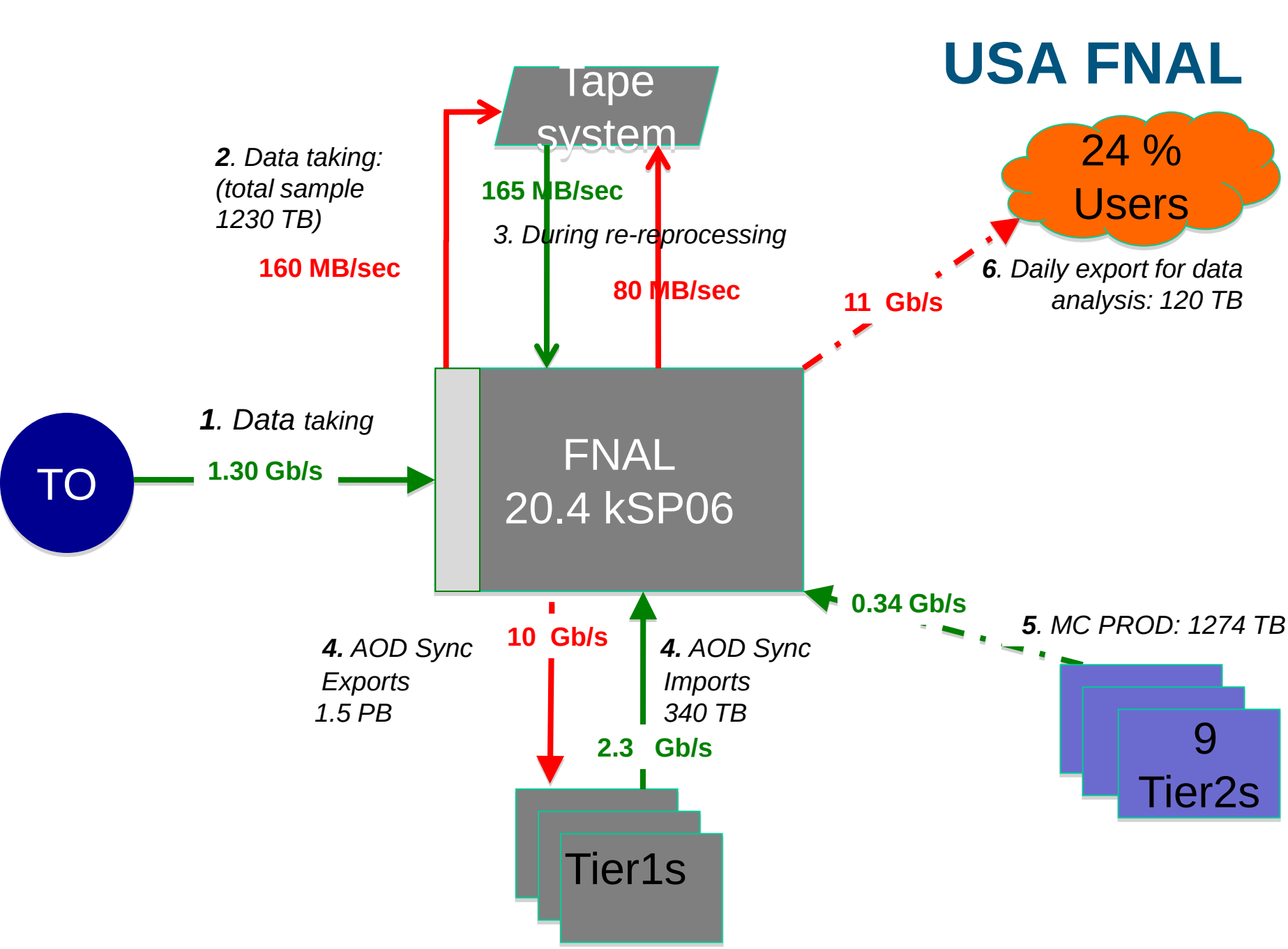

Tape system
8 % Users
2. Data taking: (total sample 151TB)
20 MB/sec
20 MB/sec
3. During re-reprocessing
10 MB/sec
3.0 Gb/s
5. Daily export for data analysis: 32 TB
1. Data taking
TO
0.2 Gb/s
PIC
5.46 kSP06
0.08 Gb/s
5. MC PROD: 420 TB
4. AOD Sync Exports 181 TB
1.2 Gb/s
4. AOD Sync Imports 553 TB
3.7 Gb/s
3 Tier2s
Tier1s

TAIWAN ASGC


Tape system
5 % Users
2. Data taking: (total sample 400 TB)
50 MB/sec
50 MB/sec
3. During re-reprocessing
25 MB/sec
0.4 Gb/s
6. Daily export for data analysis: 5 TB
1. Data taking
TO
0.4 Gb/s
TAIWAN
14 kSP06
0.08 Gb/s
5. MC PROD: 250 TB
4. AOD Sync Exports 464 TB
3.10 Gb/s
4. AOD Sync Imports 506 TB
3.40 Gb/s
3 Tier2s
Tier1s

UK RAL


Tape system
2. Data taking: (total sample 225TB)
30 MB/sec
30 MB/sec
3. During re-reprocessing
20 MB/sec
15% Users
3.7 Gb/s
6. Daily export for data analysis: 39 TB
1. Data taking
TO
0.25 Gb/s
RAL
8.04 kSP06
0.2 Gb/s
5. MC PROD: 700 TB
4. AOD Sync Exports 267 TB
1.8 Gb/s
4. AOD Sync Imports 530 TB
3.6 Gb/s
4 Tier2s
Tier1s


USA FNAL
Tape system
24 % Users
2. Data taking: (total sample 1230 TB)
165 MB/sec
3. During re-reprocessing
160 MB/sec
80 MB/sec
11 Gb/s
6. Daily export for data analysis: 120 TB
1. Data taking
TO
1.30 Gb/s
FNAL
20.4 kSP06
0.34 Gb/s
5. MC PROD: 1274 TB
4. AOD Sync Exports 1.5 PB
10 Gb/s
4. AOD Sync Imports 340 TB
2.3 Gb/s
9 Tier2s
Tier1s