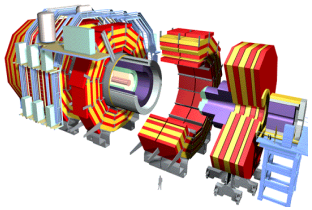
WAN area transfers and networking: a predictive model for CMS

WLCG Workshop, July 7-9, 2010

Marie-Christine Sawley, ETH Zurich

CMS Data Flow in the Computing Grid



Prompt Reconstruction

TIER-0

CAF

CASTOR

Calibration
Express-Stream
Analysis

600MB/s

50-500MB/s

TIER-1

TIER-1

TIER-1

Re-Reconstruction
Skims

50-500MB/s

~20MB/s

Simulation
Analysis

TIER-2

Purpose of the predictive model

- Need to
  - Re-examine the computing model according to conditions, and adapt whenever needed
  - Keep track of the deployment of the resources at different Tiers
- Develop a tool which would yield reliable information for the ramping up for the years to come
  - Tool has no value without dialog nor comparison with the measured rates

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= 2 MB x 300Hz x 1.4 = 840 MB/sec, or 6.75 Gb/sec.

- Each T1 receives a share according to its relative size in CPUs

- Proportional to the trigger rate, event size and Tier-1 relative size

- In 2010 we will continue to send more than 1 copy of the data, but the event size is smaller

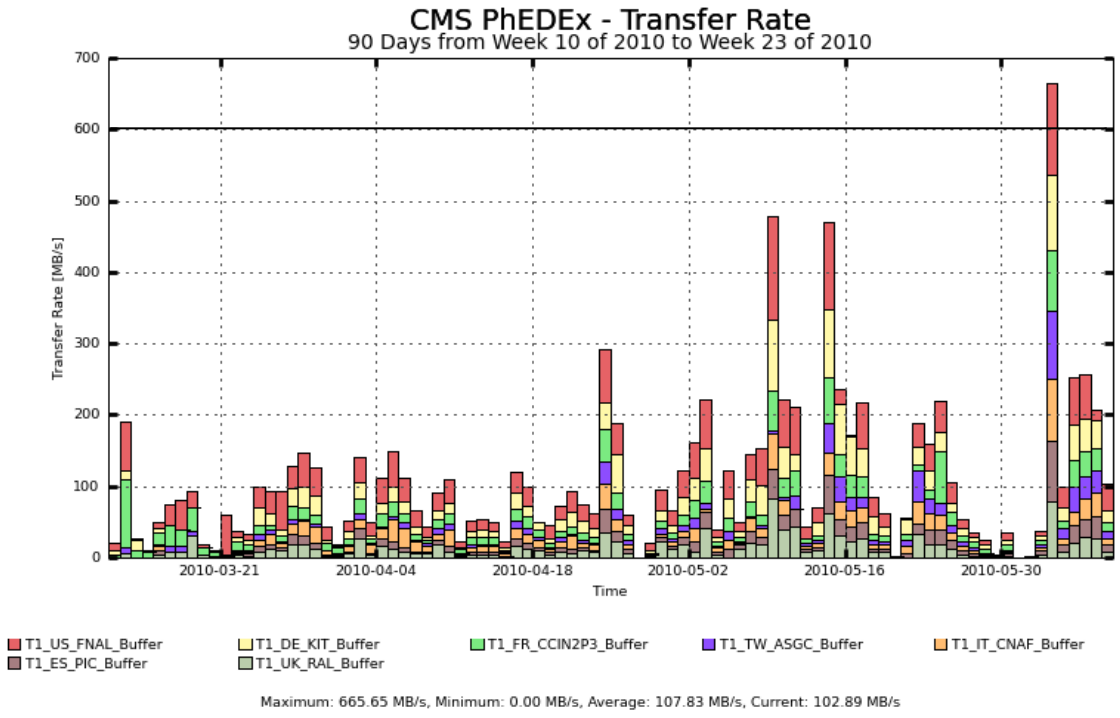
CERN to Tier-1 in 2010

- 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 would like to track the changes so that planning matches measured rates
  - Dimension according to expected bursts or peaks

Tier-0 to Tier-1

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



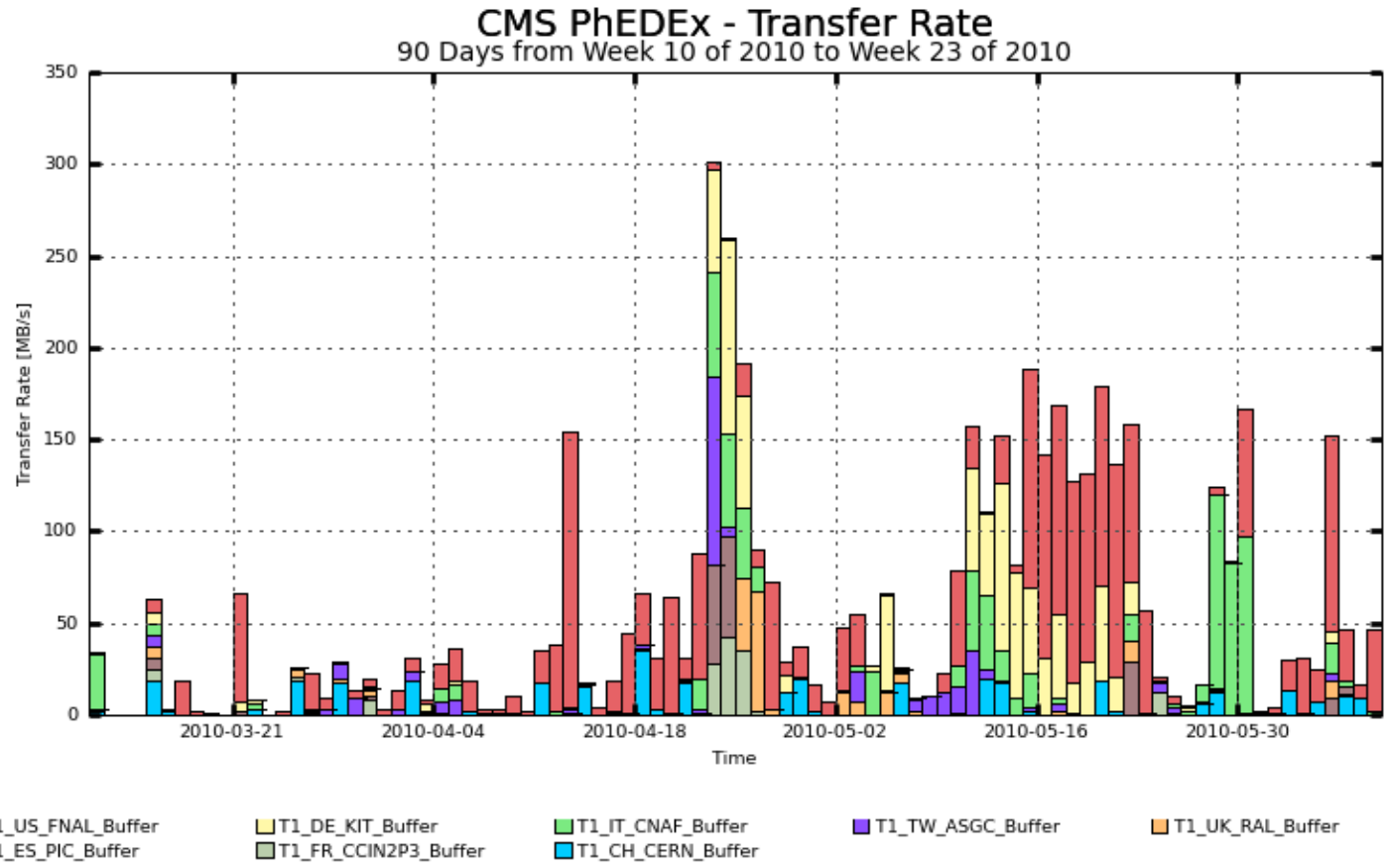
600MB/s

Tier-1 to Tier-1 in 2010

- 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 1 place only
  - Could also take longer with the normal risk of double failure

Tier-1 to Tier-1

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

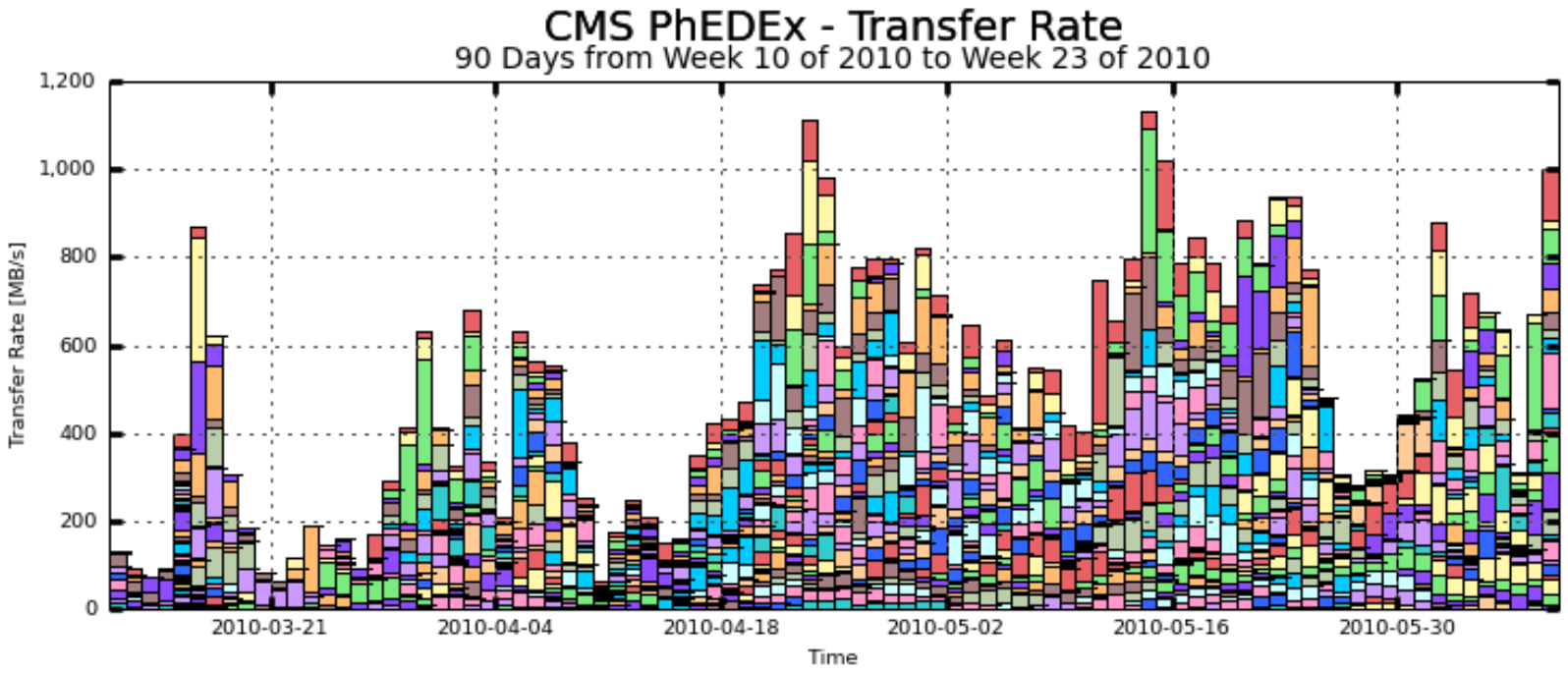


Tier-1 to Tier-2

- 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-1 to Tier-2

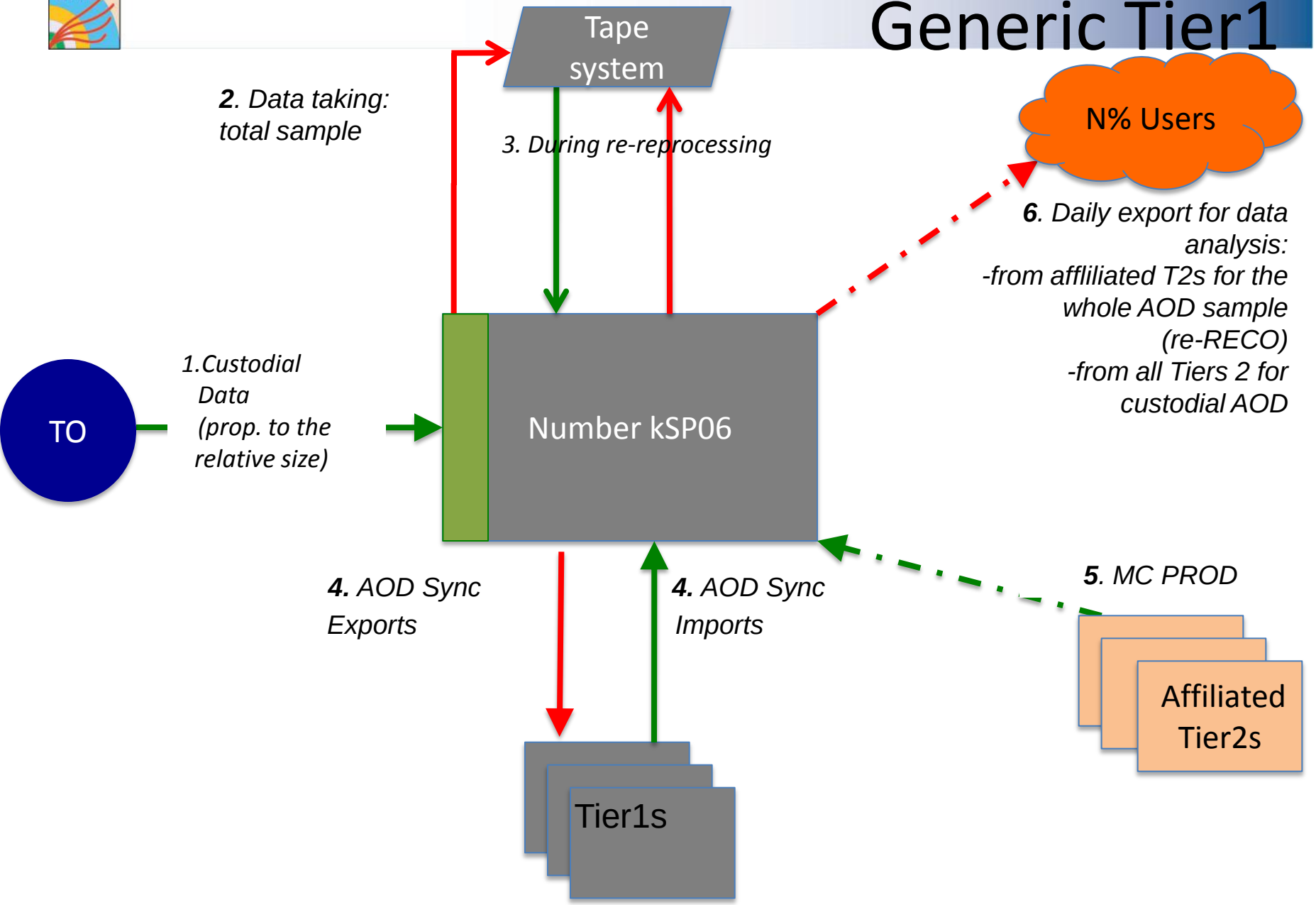
- CMS is very close to completing commissioning the full mesh of Tier-1 to Tier-2 transfers at a low rate
  - Working on demonstrating more links at 100MB/s
  - Daily average exceeding 1GB/s



2010-2011 Conditions

CONSTANTS	
Trigger rate	300 Hz
RAW Size	.500 MB
SimRAW	2.00 MB
RECO size	.500 MB
AOD size	.200 MB
Total number of events	2360 MEvents
Overlap between PD	40, then 20 %
Total number of simulated events	2076 MEvents
Total size of RAW	1474 TB
Total size RECO	1474 TB
Total Primary AOD	472 TB

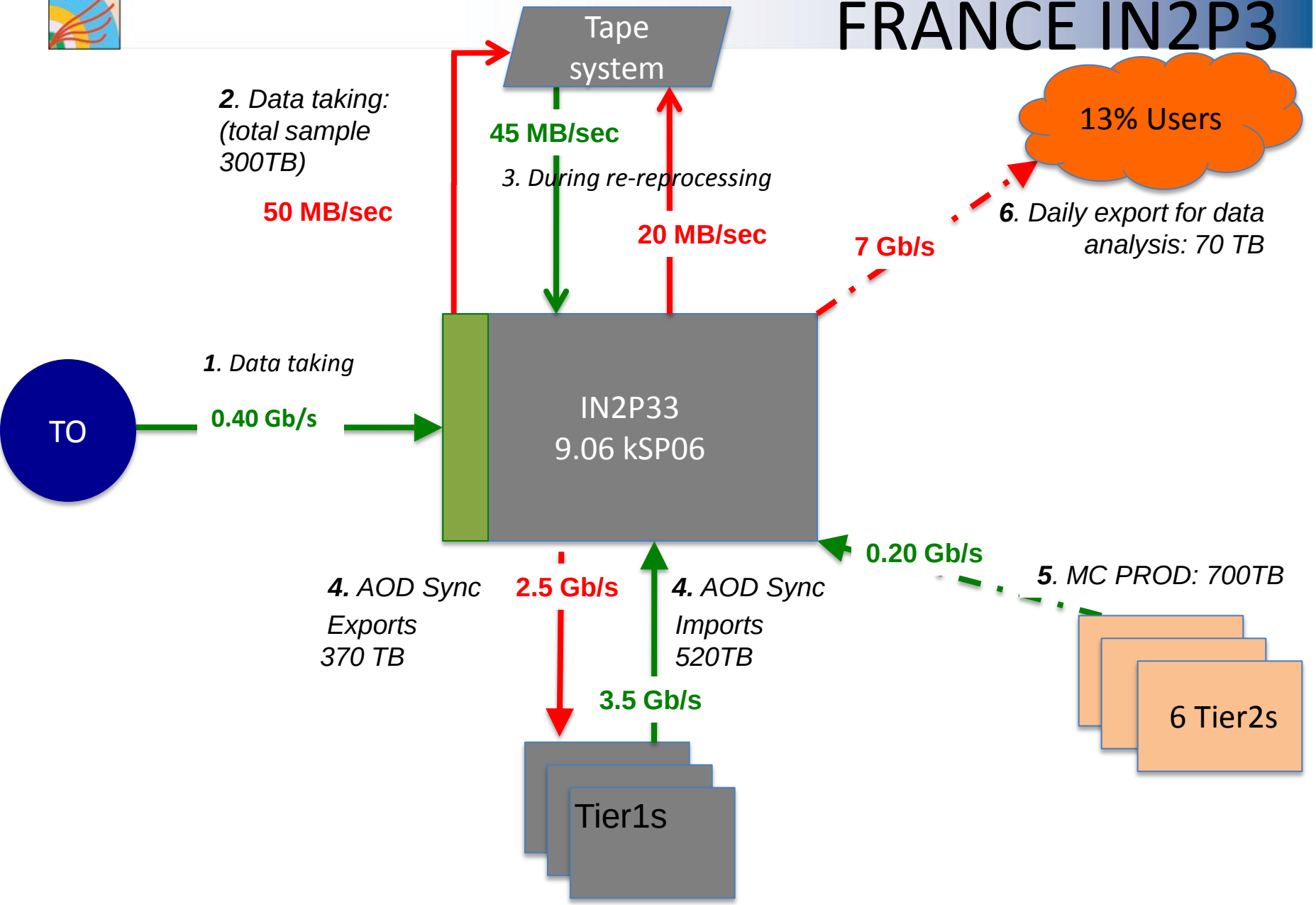
Generic Tier1



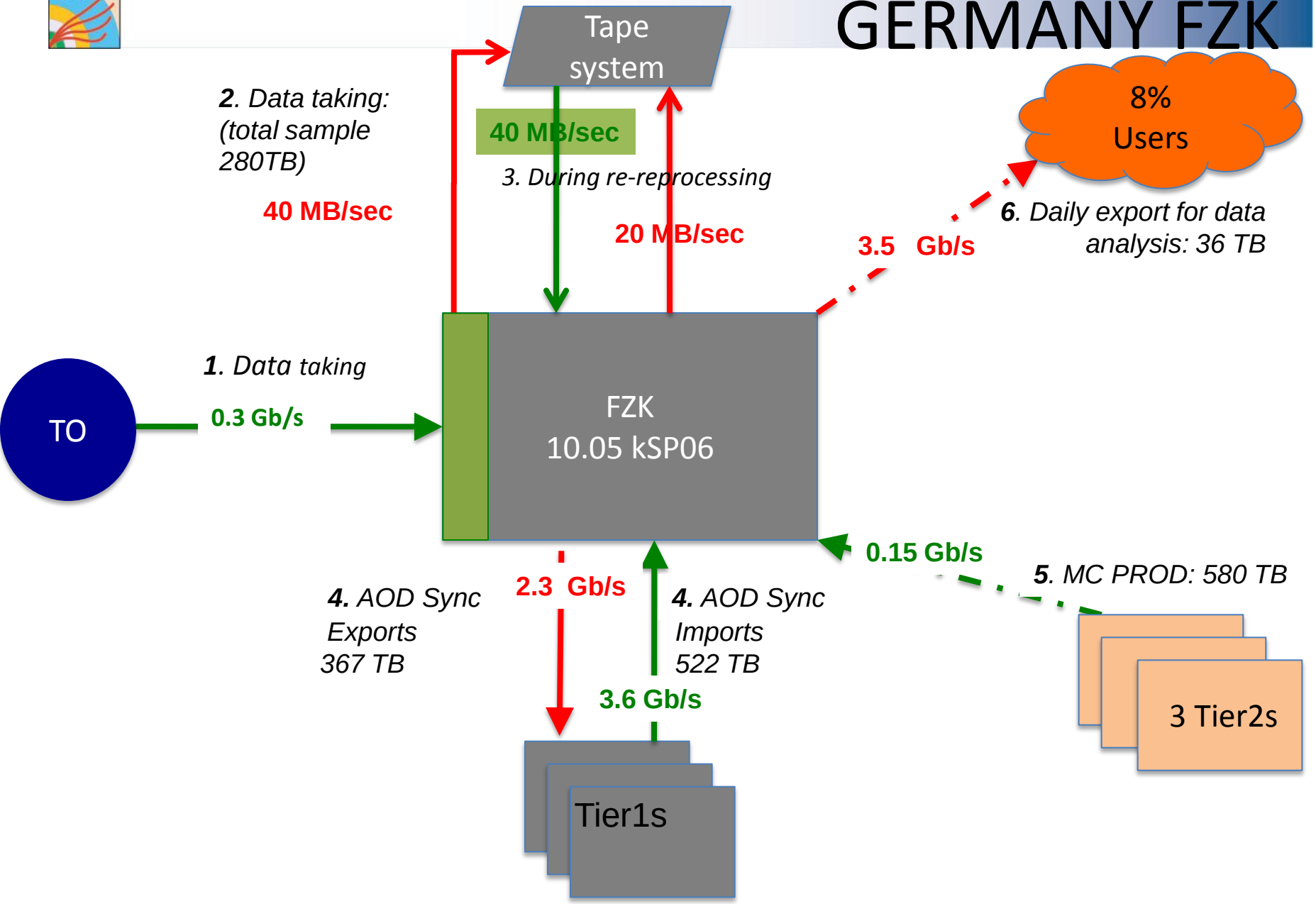
The results

- 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

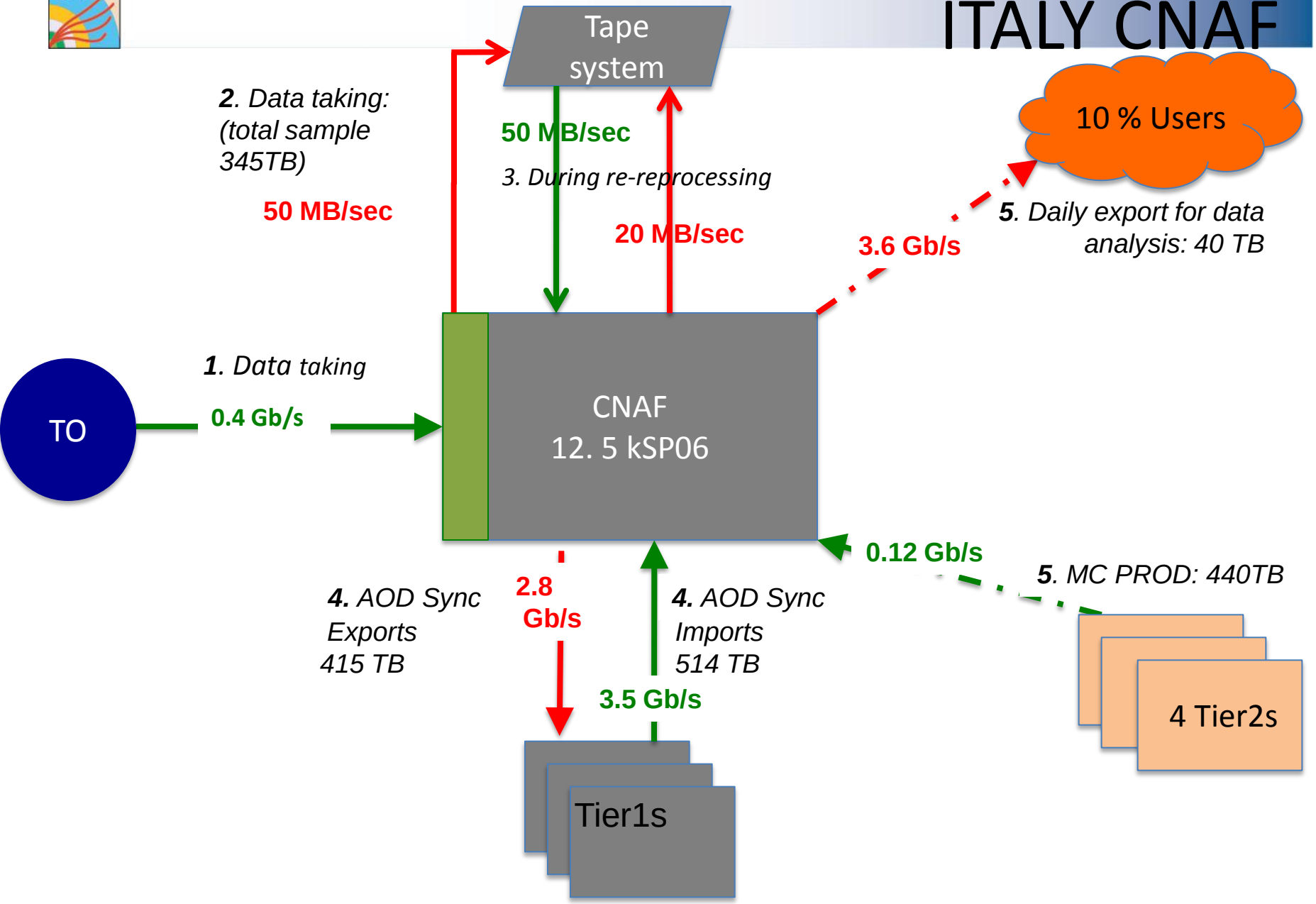
FRANCE IN2P3



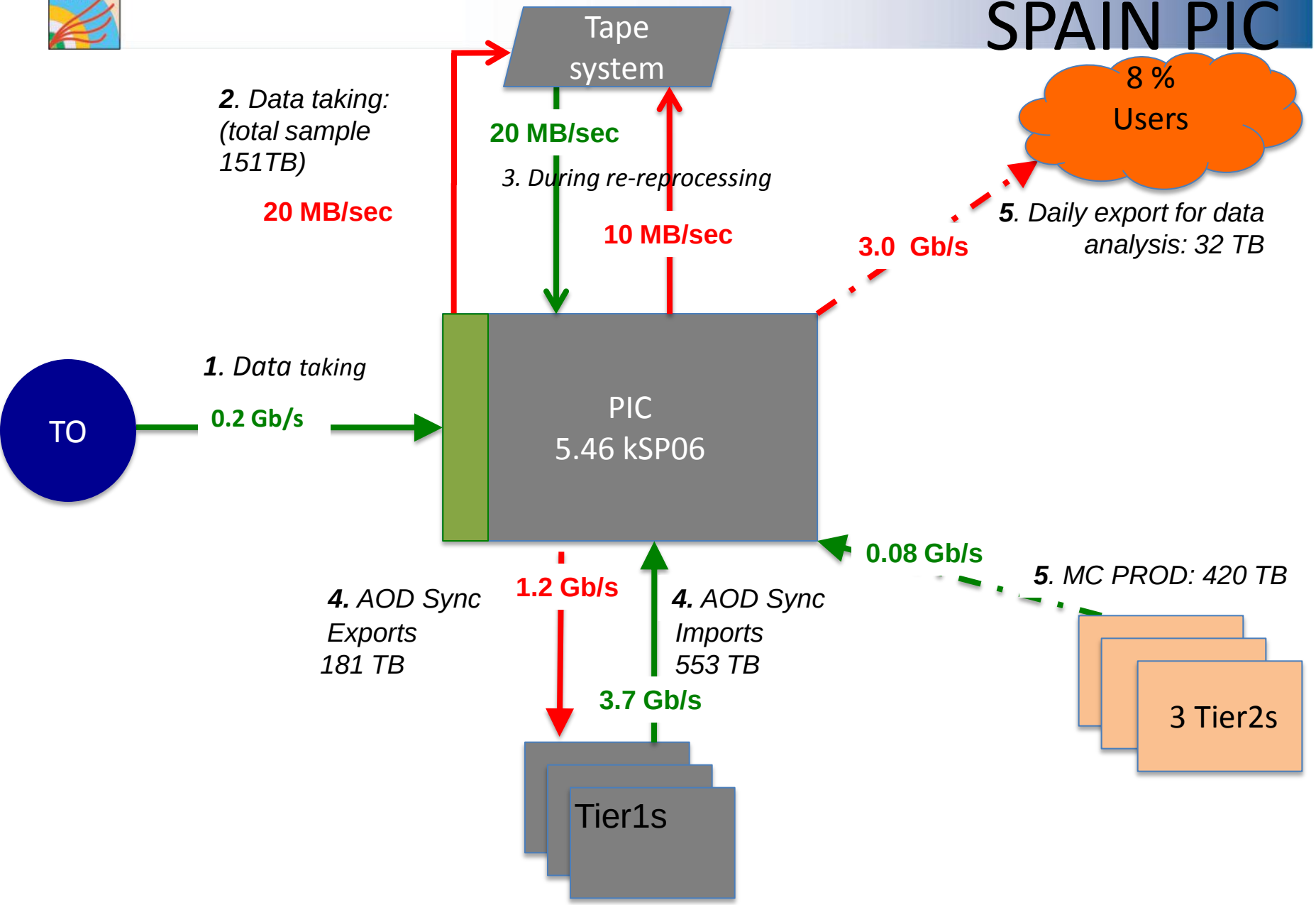
GERMANY FZK


Tape system
2. Data taking: (total sample 280TB)
40 MB/sec
40 MB/sec
3. During re-reprocessing
20 MB/sec
8% Users
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

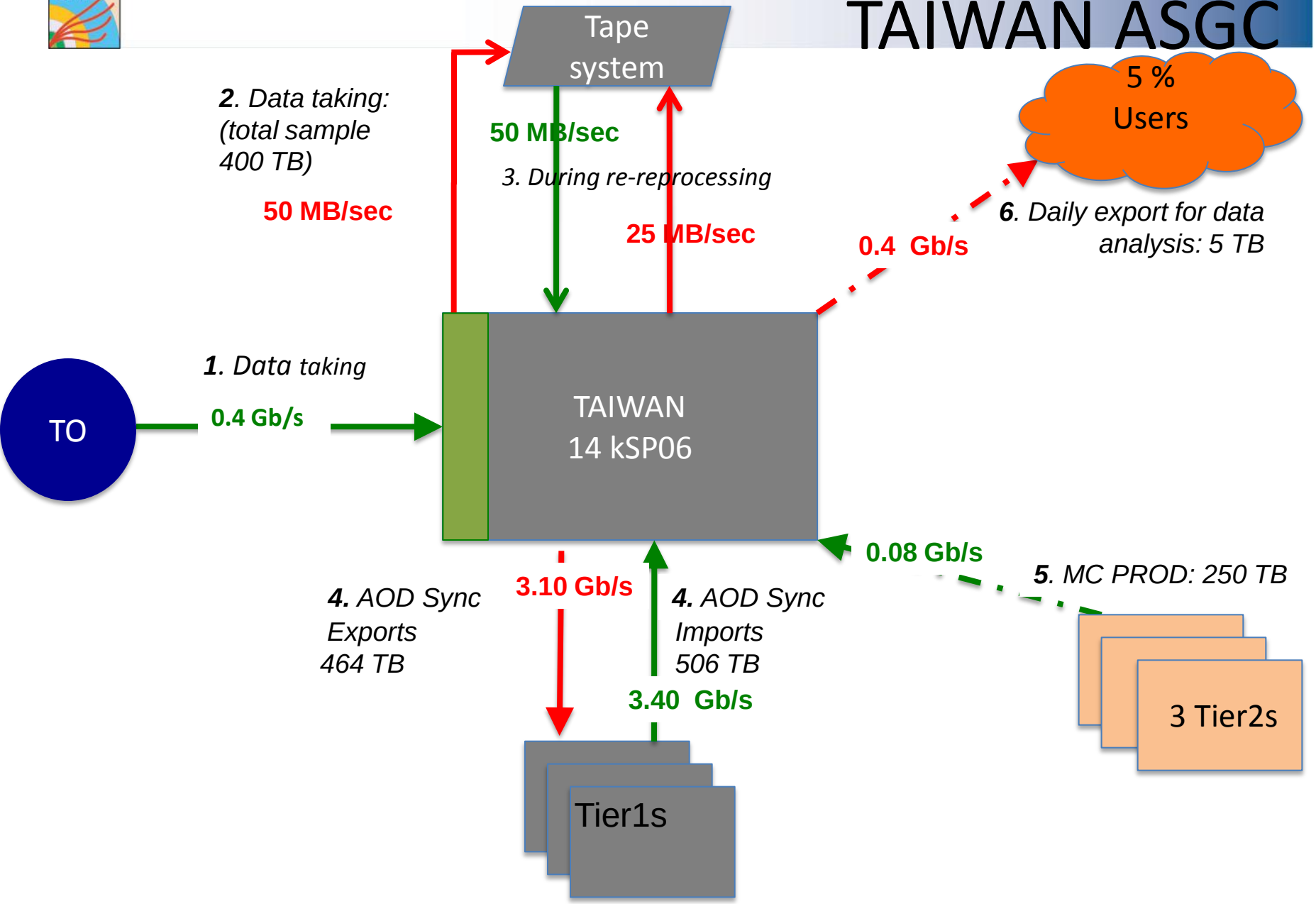
ITALY CNAF


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
1. Data taking
TO
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

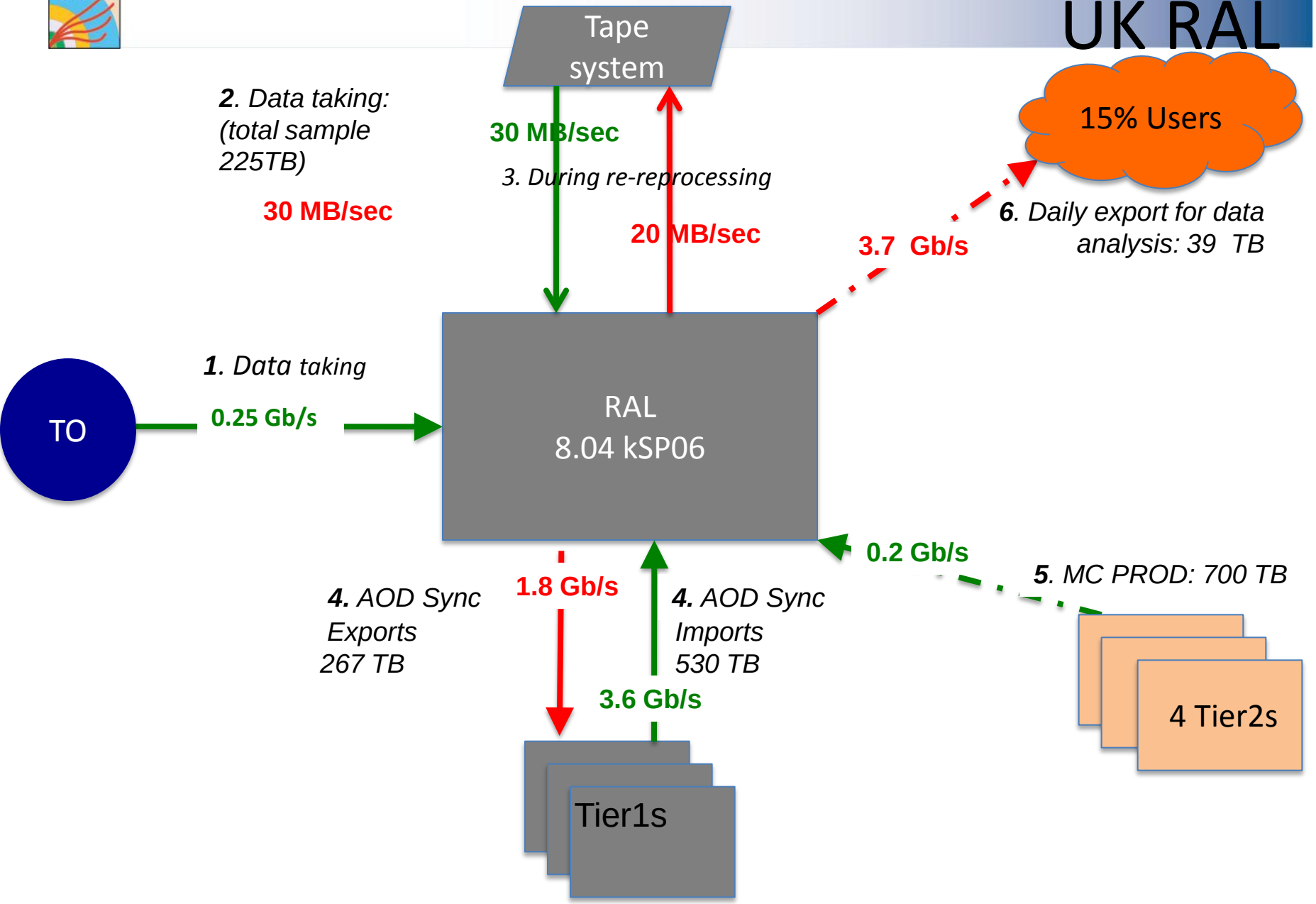
SPAIN PIC


Tape system
2. Data taking: (total sample 151TB)
20 MB/sec
20 MB/sec
3. During re-reprocessing
10 MB/sec
8 % Users
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

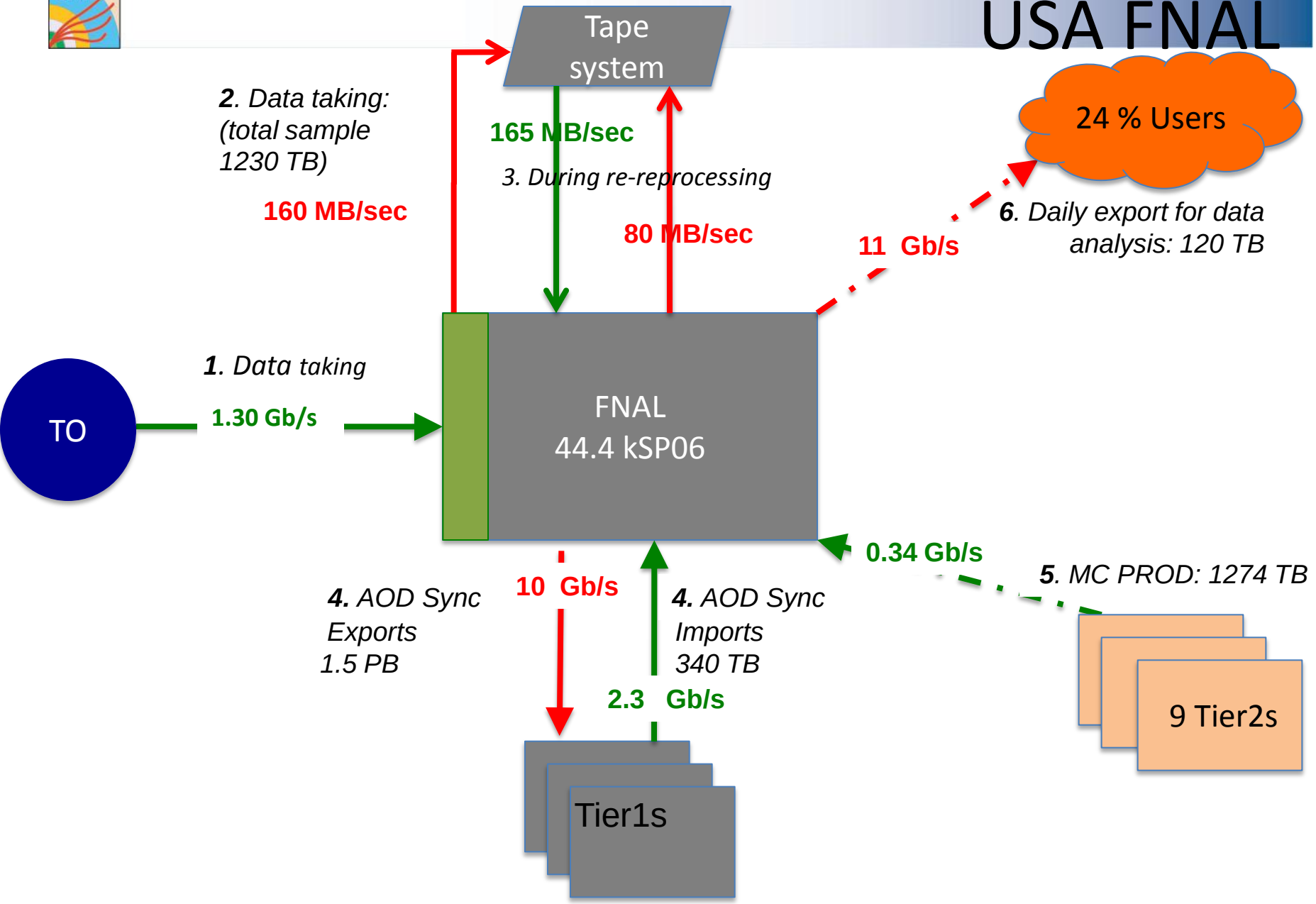
TAIWAN ASGC


Tape system
2. Data taking: (total sample 400 TB)
50 MB/sec
50 MB/sec
3. During re-reprocessing
25 MB/sec
5 % Users
6. Daily export for data analysis: 5 TB
0.4 Gb/s
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
2. Data taking: (total sample 1230 TB)
160 MB/sec
165 MB/sec
3. During re-reprocessing
80 MB/sec
24 % Users
11 Gb/s
6. Daily export for data analysis: 120 TB
1. Data taking
TO
1.30 Gb/s
FNAL
44.4 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

Data rate intake by Tier2 estimate

- Data import simulated taking the parameters of the 2010-11 run
- The association with Physics groups is taken into consideration
- The global processing capacity of Tier1s and the relative disk space of each Tier2 are taken into consideration
- Expected rate if all PG work at the same time, on a single day
  - **Sustained** rate on a **peak** day (any better definition?)
- Purpose
  - Inform sites about usage/load, planning
  - Helping sites which may run into imbalance, such as
    - WAN likely to be a limitation, especially if site is serving >1 VO
    - Imbalance between number of PG and the amount of local resources

Data rate intake by Tier-2
Preliminary comparison with measured rate

- Data from Data Ops
  - T1 to T2, best rate from sample of measures over a few hours, between November and March (200 files -2GB each- sent between T1 and T2)
- For 27 sites, the simulation gives a number below the measured data rate → satisfactory
- For 9 sites, there are no valuable data yet to be compared
- **For 7 sites**, 1 (or more) link is above simulated data rate, however the average is below → monitor and try to understand
- **For 4 sites**, all measured links were below simulated data rate→ go deeper

Possible reasons for significant deviation between results

- Simulation may be inaccurate for that particular site
  - Model keeps being refined
- New measurements keep coming
- Real limitations may come from
  - WAN
  - A few parameters to tune at the Tier-2
  - …..

- **Still very much Work in Progress, do not jump on conclusions before further analysis**

Data rate intake for T2s

Country	Tier 2s	data rate Incoming <-- T1 (MB/sec)	Number PG	Installed WAN (Gb/sec)	Measured rate (average)	remark
Austria	T2-AT-Vienna	70	2	1	**54**	
Belgium	T2-BE-IIHE	70	2	2	80	
Belgium	T2-BE-UCL	70	2	2	**47**	
Brazil	T2-BR-UERJ	70	2	10	**40**	
Brazil	T2-BR-SPRACE	30	1	10	44	
China	T2-CN-Beijing	40	1	1	32	
Estonia	T2-EE-Estonia	47	1	2	73	
Finland	T2-FI-HIP	70	2	?	104	
France	T2-FR-IPHC	70	2	1	89	
France	T2-FR-GRIF	100	3	1	**56**	
France	T2-FR-IN2P3	130	4	10	[redacted]	
Germany	T2-DE-Desy	140	4	10	111	
Germany	T2-DE-RWTH	70	2	10	111	
Hungary	T2-HU-Budapest	40	1	1	58	

Data rate intake for T2s

Country	Tier 2s	data rate Incoming <-- T1 (MB/sec)	Number PG	Installed WAN (Gb/sec)		remark
India	T2-IN-TIFR	40	1	1	90	
Italy	T2-IT-Bari	40	1	1	**30**	
Italy	T2-IT-Legnaro	70	2	2	120	
Italy	T2-IT-PISA	70	2	2	**54**	
Italy	T2-IT-Roma	70	2	1	72	
Korea	T2-KR-KNU	40	1	20	66	
Pakistan	T2-PK-NCP	2		0.07	[redacted]	*Model yields insignificant rate as disk space is very modest*
Poland	*T2-PL-Warsaw*	*5*		*2*	*97*	*Model yields insignificant rate as disk space is very modest*
Portugal	T2-PR-LIP	1		1.	60	
Portugal	T2-PR-NGC	30	1	?	70	
Russia	*T2-RU-IHEP*	*1*		*0.10*	*45*	*Model yields insignificant rate as disk space is very modest*
Russia	*T2_RU_INR*	*1*		*?*	*29*	*Model yields insignificant rate as disk space is very modest*
Russia	T2-RU-ITEP	30	1	2.5	104	
Russia	T2-RU-JINR	70	2	1	**42**	
Russia	T2-RU-PNPI	1		0.15	41	*Model yields insignificant rate as disk space is very modest*

Data rate import for T2s

Country	Tier 2s	data rate Incoming <-- T1 (MB/sec)	Number PG	Installed WAN (Gb/sec)		remark
Russia	T2-RU-RRC_KI	1		2.5		
Russia	T2-RU-SINP	30	1	2.5	38	
Spain	T2-ES-CIEMAT	100	3	2.5	**80**	
Spain	T2-ES-IFCA	80	2	2.0	74	
Switzerland	T2-CH-CSCS	40	1	10	106	
Taiwan	*T2-TW-TAIWAN*	*3*		*10*	*77*	*Model yields insignificant rate as disk space is very modest*
Turkey	*T2-TR-METU*	*4*		*1*	*55*	*Model yields insignificant rate as disk space is very modest*
UK	T2-UK-IC	100	2.5	1	**74**	
UK	T2-UK-Brunel	50	1.5	1	76	
UK	T2-UK-Southgrid RAL	40	1	1	102	
UKRAINE	T2-UA-KIPT	2	1	0.06	27	*Model yields insignificant rate as disk space is very modest*

Data rate import for T2s

Country	Tier 2s	data rate Incoming <-- T1 (MB/sec)	Number PG	Installed WAN (Gb/sec)	
USA	T2_US_Caltech	100	3	10	107
USA	T2_US_Florida	110	3	10	**56**
USA	T2_US_MIT	100	3	10	97
USA	T2_US_Nebraska	100	3	10	**82**
USA	T2_US_Purdue	100	3	10	111
USA	T2_US_UCSD	100	3	10	137
USA	T2_US_Wisconsin	100	3	10	168

Outlook

- Development of the model is on going
- Taking regularly into account real activities
  - CERN to Tier-1s is driven by the detector and the accelerator
  - 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.
  - 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.

- Keeping the dialog with the sites and the specialists to enrich the model

Thank you

Acknowledgements
Ian Fisk, Daniele Bonacorsi, Josep Flix,
Markus Klute, Oli Gutsche, Matthias Kasemann

Question or feedback?
sawley@cern.ch