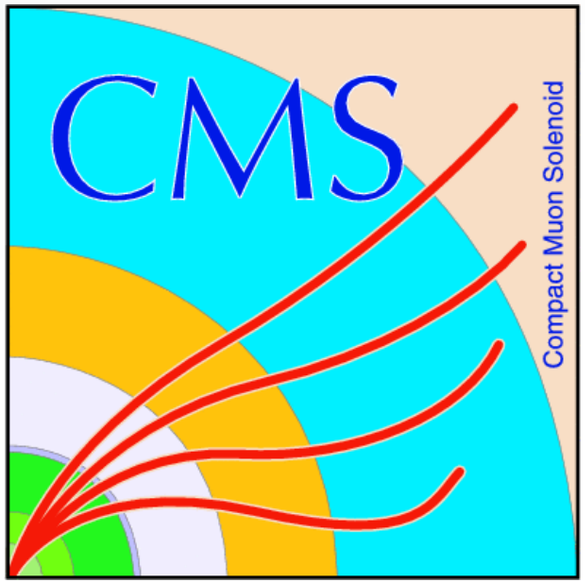


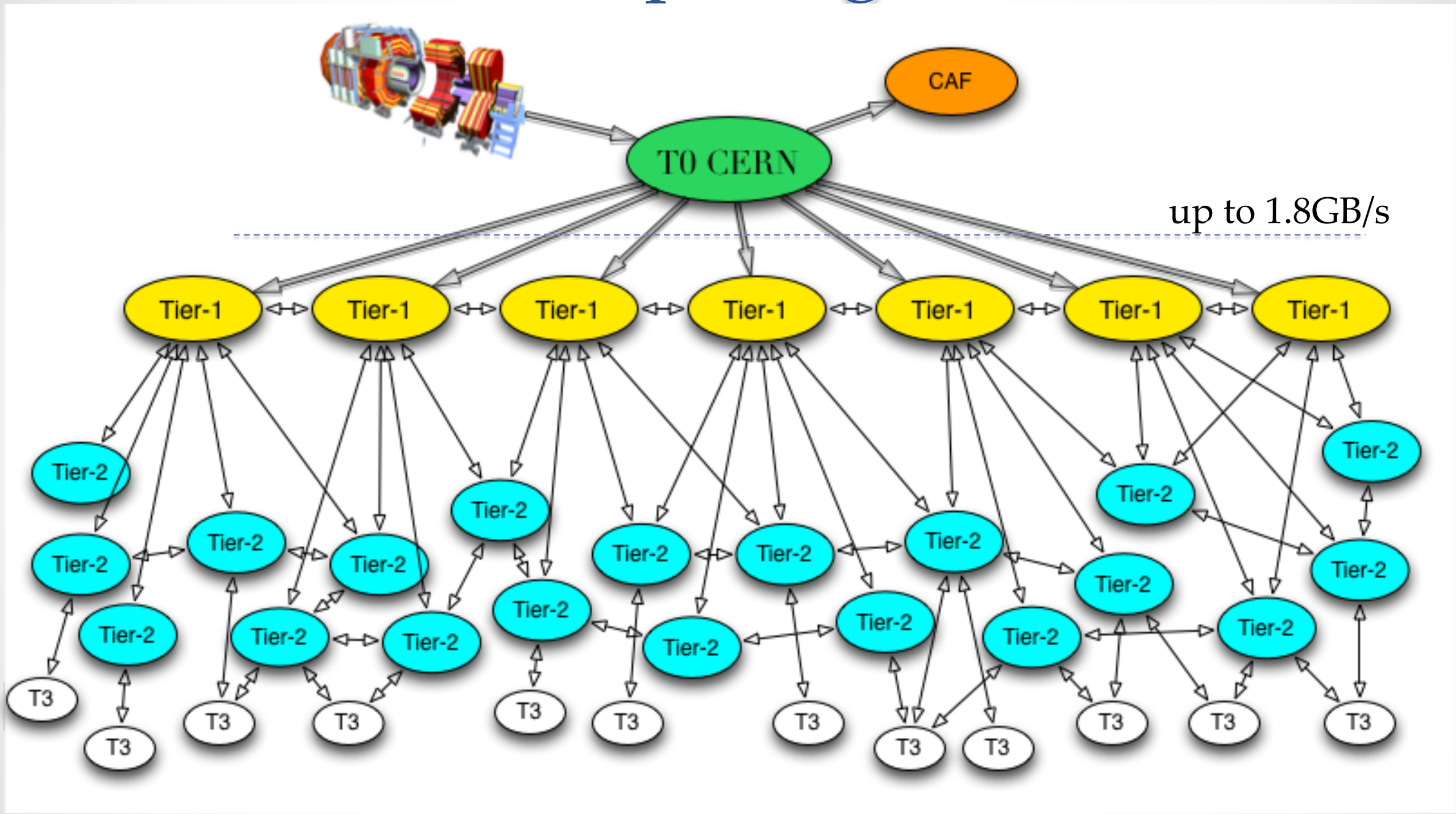
CMS Data Transfer operations after the first years of LHC collisions

D.Bonacorsi, J.Flix, O.Gutsche, R.Kaselis, P.Kreuzer, J.Letts, S.Liu, N.Magini, S.Piperov, N.Ratnikova, A.Sartirana, M.Yang

Outline

- Data Transfers Operations
  - introduction to operations;
  - monitoring and problems
  - troubleshooting
- Improving transfers quality
  - LHCONE
- Data Consistency
  - operations;
  - challenges;

Distributed computing infrastructure



7 Tier-1s; 54 Tier-2s; over 60 Tier-3s

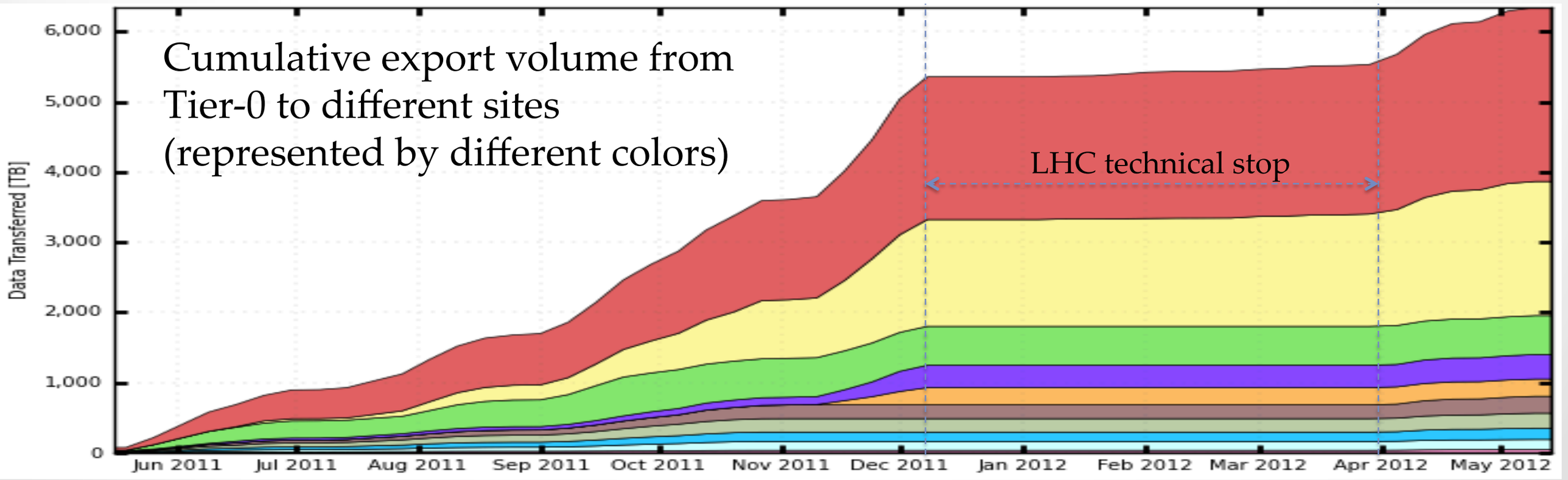
over 2550 active links

PhEDEx

- Physics Experiment Data Export is at the core of all CMS data transfers.
- Distributed database-centric architecture.
- Keeps track of data stored at sites.
- Central agents, which are calculating routes, harvest historical data, etc., are running at CERN
- Each CMS site runs a set of site software agents.
- Web based monitoring and control
  - helps to observe failing transfers, debug issues;
  - data subscriptions and approvals managed by responsible site administrators.
- Web service providing machine readable information.

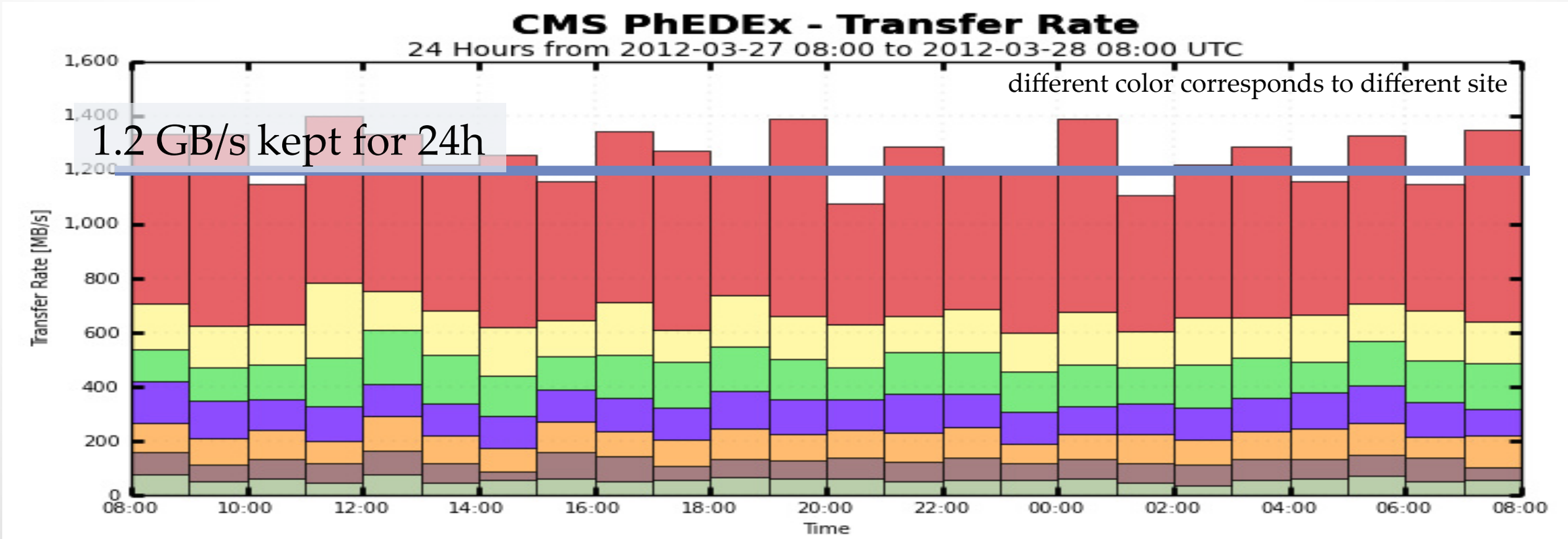
Quantities

- In data taking mode:
  - 25-30 TB are being exported from Tier-0 daily during protons' runs.
  - 60 TB on avg. exported daily during heavy-ions' runs (with record 120 TB/day, ~1.4 GB/s) from Tier-0.
  - 30-35 TB are being exported from Tier-1s to Tier-2s daily.
  - 15 TB daily Tier-2s to Tier-2s.
  - ~1400 TB transferred in one week



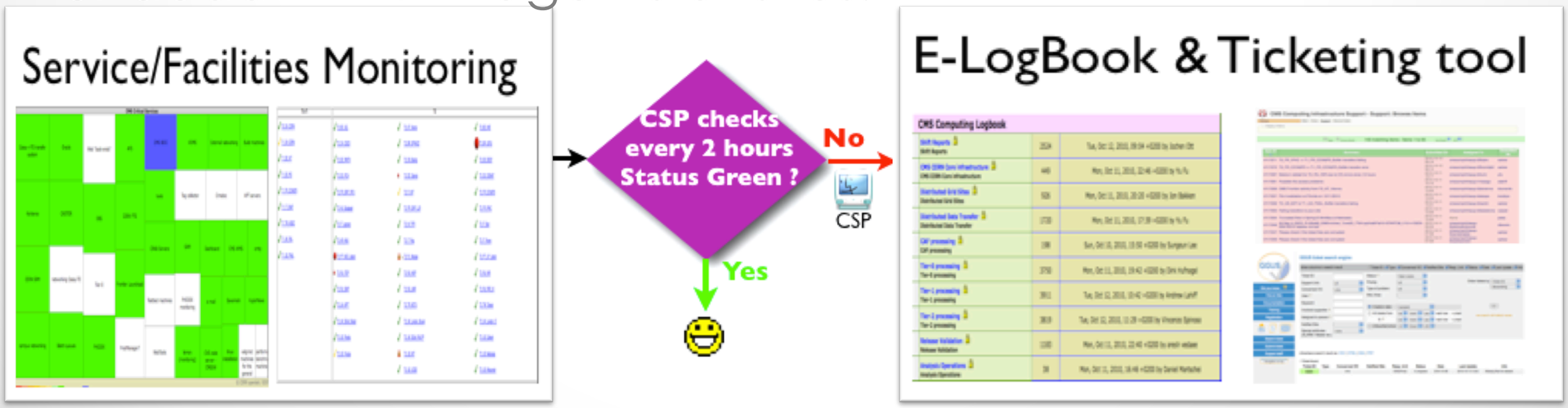
Preparation for 2012

- During the most recent test (Tier-0 ➔ Tier-1s), 1.2GB/s export rate has been achieved (export planned for 2012)
  - furthermore, available infrastructure can keep up to 1.8GB/s (retained for 12h)
- Regular transfer tests are going on all the time
- Commissioning links
  - Export from Tier-1 at 20MB/s for 24h
  - Export from Tier-2 at 5MB/s for 24h



Monitoring transfers

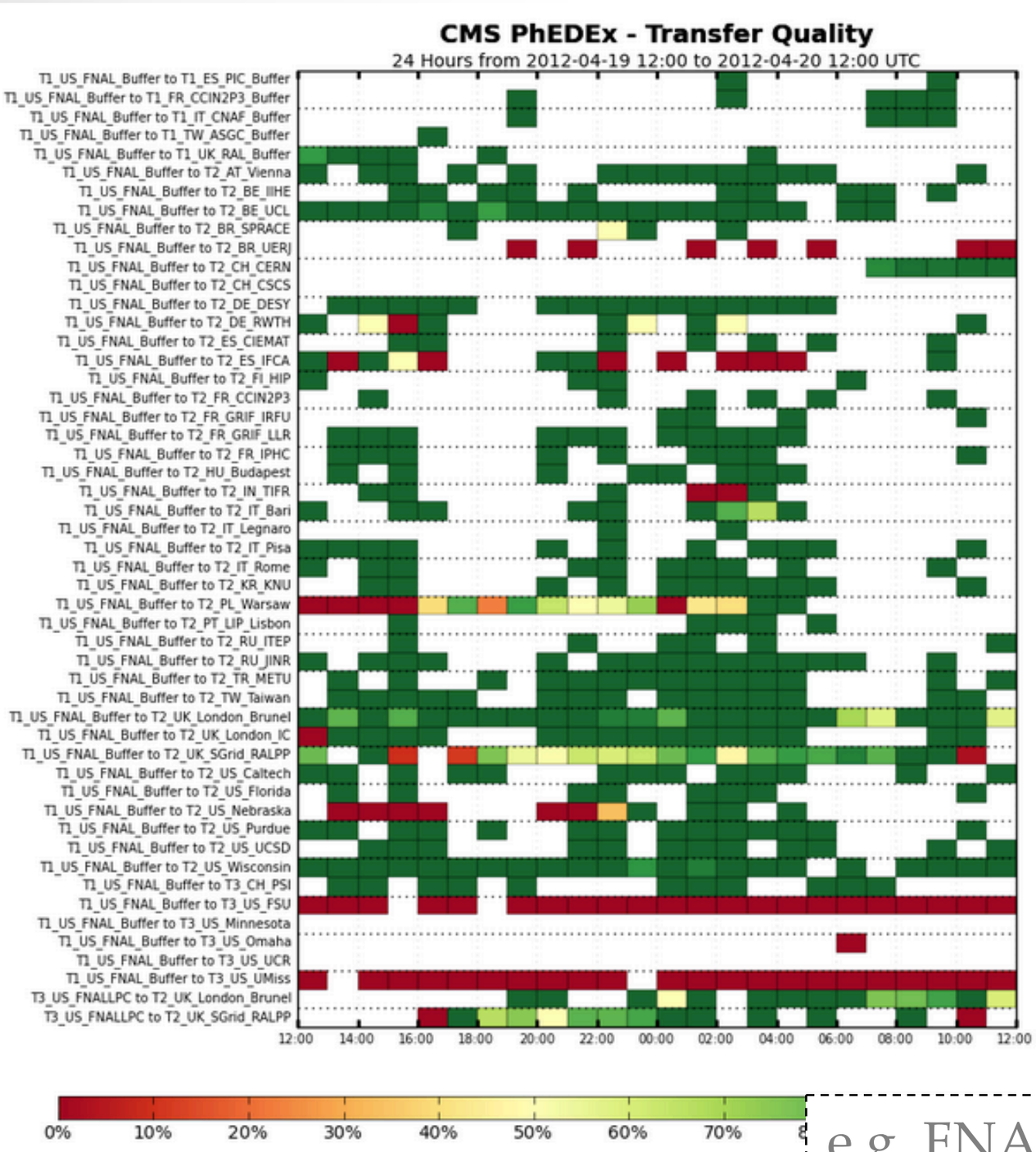
- Mostly looking at PhEDEx web page:
  - transfers quality plots.
  - transfers rate table.
- More than 2550 active links.
- 4 central operators.
- Basic checks are being done by computing shifters*, who monitor 24/7 and are very helpful detecting stuck transfers or dead PhEDEx agents at sites.



* see "Towards higher reliability of CMS Computing Facilities" poster by José Flix

Monitoring transfers 2

Quality plot



Rate table

Last 2 hours

To		From	Files	Total Size	Rate	Errors	Expired	Avg. Est. Rate	Avg. Est. Latency
.*	and	FNAL							
T2_CH_CERN		T1_US_FNAL_Buffer	806	4.6 TiB	668.9 MiB/s	16	33	677.0 MiB/s	3h05
T1_US_FNAL_MSS		T1_US_FNAL_Buffer	672	1.7 TiB	250.2 MiB/s	-	-	433.6 MiB/s	0h10
T2_US_Wisconsin		T1_US_FNAL_Buffer	47	169.6 GiB	24.1 MiB/s	-	-	19.9 MiB/s	0h43
T2_UK_London_Brunel		T1_US_FNAL_Buffer	18	55.0 GiB	7.8 MiB/s	6	32	4.5 MiB/s	5d5h15
T2_UK_London_Brunel		T3_US_FNALLPC	12	39.3 GiB	5.6 MiB/s	4	-	10.8 MiB/s	1d7h54
T2_US_Caltech		T1_US_FNAL_Buffer	6	19.4 GiB	2.8 MiB/s	-	1	1.3 MiB/s	2d22h16
T2_TR_METU		T1_US_FNAL_Buffer	2	4.5 GiB	661.9 kiB/s	-	1	167.6 kiB/s	1d17h03
T2_UK_SGrid_RALPP		T1_US_FNAL_Buffer	2	3.3 GiB	479.7 kiB/s	34	-	10.0 MiB/s	1d10h16
T2_US_Purdue		T1_US_FNAL_Buffer	1	570.9 MiB	81.2 kiB/s	-	-	125.5 kiB/s	9h07
T2_TW_Taiwan		T1_US_FNAL_Buffer	2	69.6 MiB	9.9 kiB/s	-	-	11.0 kiB/s	1d8h09
T2_DE_RWTH		T1_US_FNAL_Buffer	1	66.8 MiB	9.5 kiB/s	-	2	10.6 kiB/s	3d20h06
T2_KR_KNU		T1_US_FNAL_Buffer	1	6.3 MiB	923.2 iB/s	-	-	1.8 kiB/s	3d18h13
T2_RU_JINR		T1_US_FNAL_Buffer	2	5.7 MiB	832.1 iB/s	-	2	2.2 kiB/s	5d3h16
T2_FR_CCIN2P3		T1_US_FNAL_Buffer	1	4.4 MiB	641.6 iB/s	-	-	4.4 kiB/s	0h28
T2_IT_Rome		T1_US_FNAL_Buffer	1	3.7 MiB	543.6 iB/s	-	-	5.1 kiB/s	0h09
T2_RU_ITEP		T1_US_FNAL_Buffer	1	3.7 MiB	543.6 iB/s	-	-	1.7 kiB/s	0h33
T2_AT_Vienna		T1_US_FNAL_Buffer	1	2.9 MiB	415.9 iB/s	-	3	577.8 iB/s	7d0h00
T2_IT_Pisa		T1_US_FNAL_Buffer	1	2.9 MiB	415.9 iB/s	-	-	869.5 iB/s	4d8h04
T2_US_UCSD		T1_US_FNAL_Buffer	1	2.9 MiB	415.4 iB/s	-	-	2.7 kiB/s	0h16
T2_UK_London_IC		T1_US_FNAL_Buffer	1	2.8 MiB	414.4 iB/s	-	4	1.9 kiB/s	7d0h00
T2_US_Florida		T1_US_FNAL_Buffer	1	2.8 MiB	414.4 iB/s	-	2	460.9 iB/s	7d0h00
T2_FR_GRIF_IRFU		T1_US_FNAL_Buffer	1	2.8 MiB	414.4 iB/s	-	-	1.2 kiB/s	0h33
T2_FR_IPHC		T1_US_FNAL_Buffer	1	2.8 MiB	414.4 iB/s	-	-	671.0 iB/s	1h41
T2_BE_IIHE		T1_US_FNAL_Buffer	1	1.8 MiB	263.5 iB/s	-	-	1.1 kiB/s	6d2h21
T2_BR_UERJ		T1_US_FNAL_Buffer	-	-	-/s	19	3	-/s	6d12h48
T3_US_UMiss		T1_US_FNAL_Buffer	-	-	-/s	3	5	-/s	7d0h00
T3_US_FSU		T1_US_FNAL_Buffer	-	-	-/s	2	1	-/s	5d0h00
T2_UK_SGrid_RALPP		T3_US_FNALLPC	-	-	-/s	1	-	3.9 MiB/s	0h14
T1_UK_RAL_Buffer		T1_US_FNAL_Buffer	-	-	-/s	-	23	100.1 kiB/s	1d0h41
T3_US_UCR		T1_US_FNAL_Buffer	-	-	-/s	-	13	-/s	6d12h48
T3_CH_PSI		T1_US_FNAL_Buffer	-	-	-/s	-	12	469.5 iB/s	7d0h00
T2_DE_DESY		T1_US_FNAL_Buffer	-	-	-/s	-	5	1.2 kiB/s	7d0h00
T2_HU_Budapest		T1_US_FNAL_Buffer	-	-	-/s	-	3	287.1 iB/s	7d0h00
T2_ES_IFCA		T1_US_FNAL_Buffer	-	-	-/s	-	2	410.2 iB/s	6d12h48
T2_US_Nebraska		T1_US_FNAL_Buffer	-	-	-/s	-	2	589.6 iB/s	4d2h35
T2_BR_SPRACE		T1_US_FNAL_Buffer	-	-	-/s	-	1	618.1 iB/s	6d3h00
		US_FNAL_Buffer	-	-	-/s	-	1	973.2 iB/s	4d9h00
			1583	6.6 TiB	960.7 MiB/s	85	151	-/s	0h00

e.g. FNAL → all sites

Transfer problems

- Storage issues
  - corrupted tape
  - crashed storage node
  - data loss
- Network issues
  - cable cut
  - misbehaving router on the path
  - timeouts
- Authorization issues
  - expired certificate/proxy
  - certificate/proxy doesn't have appropriate roles/extensions
- Configuration issues
  - Improper PhEDEx agents' configuration
  - FTS channels configuration

Troubleshooting

- Check if there is a link between hosting file site and receiving site?
  - additional inspections why path can not be calculated.
- Check what status file is in?
  - might be a problem staging from tape.
  - might not report properly as staged on disk.
- Check if there are errors on that link?
  - errors might immediately tell what the problem is and where is it.
- Open a ticket to a site. Give some hints what might be causing problems, ask to solve the them.
  - ~1300 tickets were opened during the last year.
  - On avg. 3.5 tickets per day (5 tickets/day excl. weekends).
  - 15-20 tickets on average are open at any given moment.

Troubleshooting 2

Show paths on links to T2_CH_CERN and from .* with any priority Show Invalid Paths

Filter blocks .* Update

Destination	Source	Block Name	Priority	Routed Files	Routed Bytes	Transfer Attempts	Average Attempts	Earliest Request
T2_CH_CERN	T1_US_FNAL_Buffer	/ZeroBias4/Run2012A-PromptReco-v1/RECO#c862147e-8877-11e1-a9bb-003048caaace	normal	153	934.2 GiB	153	1.00	7h45 ago
T2_CH_CERN	T1_US_FNAL_Buffer	/ZeroBias2/Run2012A-PromptReco-v1/REC						
T2_CH_CERN	T1_US_FNAL_Buffer	/ZeroBias3/Run2012A-PromptReco-v1/REC						
T2_CH_CERN	T1_US_FNAL_Buff	Bias4/Run2012A-PromptReco-v1/REC						
T2_CH_CERN	T1_US_FNAL_Buffer	Bias2/Run2012A-PromptReco-v1/REC						
T2_CH_CERN	T1_US_FNAL_Buffer	Bias3/Run2012A-PromptReco-v1/REC						
T2_CH_CERN	T1_US_FNAL_Buffer	onetaMinus/Fall11-standard_443p1-v						
T2_CH_CERN	T1_US_FNAL_Buffe	taPlusX0Max/Fall11-X0Max_443p1						
T2_CH_CERN	T1_US_FNAL_Buffer	/PhotonetaPlus/Fall11-standard_443p1-v1/						
T2_CH_CERN	T1_US_FNAL_Buffer	/SingleElectronPt35X0Max/Fall11-X0Max_4						
T2_CH_CERN	T1_US_FNAL_Buffer	/SingleElectronPt35/Fall11-standard_443p1						
T2_CH_CERN	T0_CH_CERN_Export	/ZeroBias1/Run2012A-v1/RAW#c86d2a88-						
T2_CH_CERN	T0_CH_CERN_Export	/ZeroBias1/Run2012A-v1/RAW#83c56736-						
T2_CH_CERN	T0_CH_C							
T2_CH_CERN	T0_CH_C							
T2_CH_CERN	T0_CH_C							
T2_CH_CERN	T0_CH_C							
T2_CH_CERN	T0_CH_C							
T2_CH_CERN	T0_CH_C							
T2_CH_CERN	T0_CH_C							
T2_CH_CERN	T0_CH_C							
T2_CH_CERN	T0_CH_C							
T2_CH_CERN	T0_CH_C							

1

Transfer State Details

Age	To Node		From Node	State	Transfer	Priority	N Files	Size	Detail
	T2_CH_CERN	or	T2_CH_CERN						
current	T1_US_FNAL_Buffer		T2_CH_CERN	transferring	remote	high	1	742.2 kiB	(Files)
current	T2_CH_CERN		T0_CH_CERN_Export	assigned	remote	normal	10	23.3 GiB	(Files)
current	T2_CH_CERN		T0_CH_CERN_Export	exported	remote	normal	29	963.9 MiB	(Files)
				ing	remote	normal	37	14.6 GiB	(Files)
					remote	high	570	3.3 TiB	(Files)
						assigned	580	3.3 TiB	

2

Error 9

To Node: T2_CH_CERN
Time Assigned: 2012-04-15 15:45:56 UTC (0m00 since assigned) (-3d2h16 from now)
Time Exported: 2012-04-15 16:16:05 UTC (30m08 since assigned) (-3d1h46 from now)
Time Pumped: 2012-04-15 16:17:53 UTC (31m56 since assigned) (-3d1h44 from now)

From Node: T0_CH_CERN_Export
Time Transfer Start: 2012-04-15 16:18:23 UTC (32m26 since assigned) (-3d1h44 from now)
Time Transfer Done: 2012-04-15 18:18:25 UTC (2h32 since assigned) (-2d23h44 from now)
Time Transfer Expires: 2012-04-15 23:58:18 UTC (8h12 since assigned) (-2d18h04 from now)

Report Code: -258
Transfer Code:

To PFN: /eos/cms/store/data/Run2012A/TauPlusX/RECO/PromptReco-v1/000/191/086/E22B32C0-0887-E111-
From PFN: /castor/cern.ch/cms/store/data/Run2012A/TauPlusX/RECO/PromptReco-v1/000/191/086/E22B32C0-0
Space Token: (none)

Transfer Log: (3 lines) (477 chars)

```
2012-04-15 16:18:23 castor_eos_cp(19420): Executing: /data/ProdNodes/SITECONF/CH_CERN/
2012-04-15 18:18:23 castor_eos_cp(19420): d out, sending signal 1
2012-04-15 18:18:23 castor_eos_cp(1942 ted with status code signal 1 (1) af
```

3

Detail Log: (1 lines) (51 chars)

```
transfer timed out after 7200 seconds with signal 1
```

Validate Log: (7 lines) (820 chars)

```
2012-04-15 18:18:23 EOS.CAF.FileValidate(3899): Executing: /data/ProdNodes/SITECONF/CH
2012-04-15 18:18:23 EOS.CAF.FileValidate(3899): pfn is /eos/cms/store/data/Run2012A/Ta
2012-04-15 18:18:23 EOS.CAF.FileValidate(3899): Size mismatch
2012-04-15 18:18:23 EOS.CAF.FileValidate(3899): disk= db=12261017879
2012-04-15 18:18:23 EOS.CAF.FileValidate(3899): Checksum mismatch
2012-04-15 18:18:23 EOS.CAF.FileValidate(3899): disk=0x, db=0xccf8b6c1
2012-04-15 18:18:24 EOS.CAF.FileValidate(3899): Job exited with status code 3 (768) af
```

CMS Computing Infrastructure Support - Support: Browse Items

Group | Main | Docs | Support | Source Code

(+) Display Criteria

27 matching items - Items 1 to 27

Item ID	Summary	Submitted On	Assigned To	Submitted By
#128110	Transfer errors from KIPT	2012-04-24 14:53	cmscompinfrasup-t2uakipt	piperov
#128109	Transfer errors T2_PT_NCG_Lisbon -> T1_UK_RAL	2012-04-24 14:47	cmscompinfrasup-t2ptncglisbon	piperov
#128100	Transfer errors to T1_TW_ASGC	2012-04-24 10:51	cmscompinfrasup-t1twasgc	piperov
#128098	Failing transfers from ITEP	2012-04-24 10:43	cmscompinfrasup-t2ruitep	cassel
#128090	Failing transfers to METU	2012-04-24 09:11	cmscompinfrasup-t2trmetu	cassel
#128089	Expiring transfers to MIT	2012-04-24 08:43	cmscompinfrasup-t2usmit	cassel
#128070	T2_RU_PNPI->T1_UK_RAL transfer	2012-04-23 19:13	cmscompinfrasup-t2rupnpi	monicava
#128058	Transfer errors T2_US_Nebraska -> T1_UK_RAL	2012-04-23 16:21	cmscompinfrasup-t2usnebraska	piperov
#128056	T1_TW->T2_SGrid_RALPP transfer	2012-04-23 14:48	cmscompinfrasup-t1twasgc	monicava
#128052	Failing transfers from Caltech	2012-04-23 12:35	cmscompinfrasup-t2uscaltech	cassel
#128048	Timing out transfers to RALPP	2012-04-23 10:50	cmscompinfrasup-t2uksgridralpp	cassel

4

LHCONE

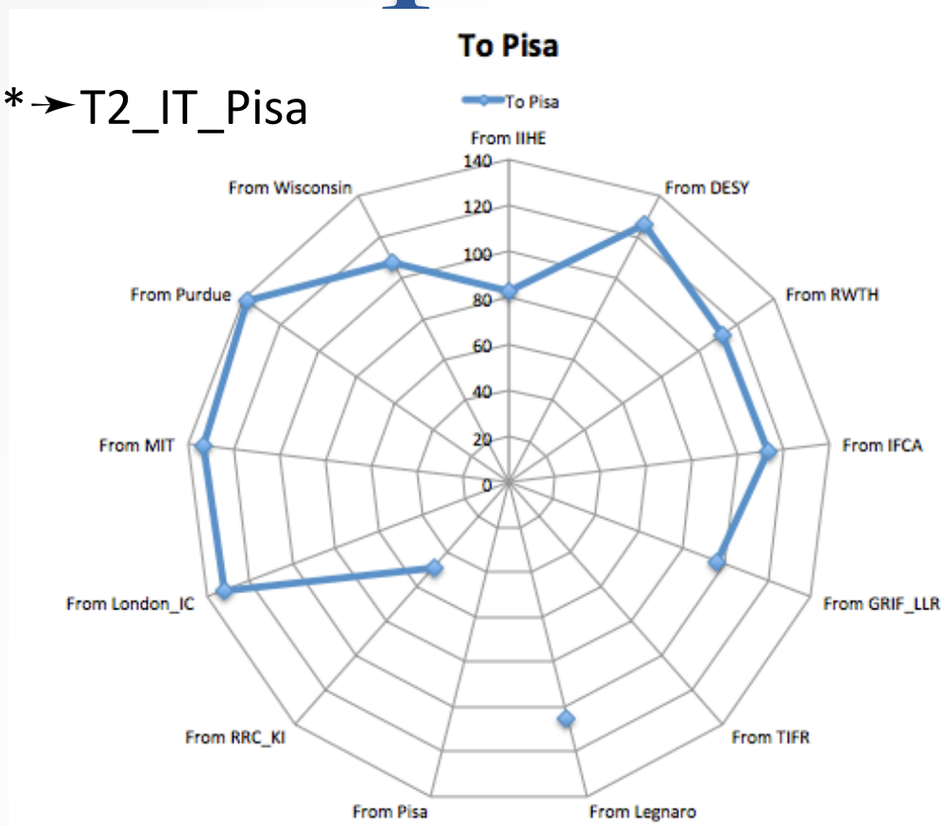
- LHC Open Network Environment.
- CMS transfers must rely on a stable and reliable network behind, but should be ready to face any changes in the underlying network infrastructure
- The objective of LHCONE is to provide a collection of access locations that are effectively entry points into a network that is private to the LHC T1/2/3 sites.
  - LHCONE is not intended to replace the LHCOPN but rather to complement it;
  - it addresses Tier-2/3 levels, on GPN infrastructures in different nations so far;
  - LHCONE is intended to grow as a robust and scalable solution for a global system serving LHC Tiers` needs and to fit in less-hierarchical computing models

Transfer tests in LHCONE

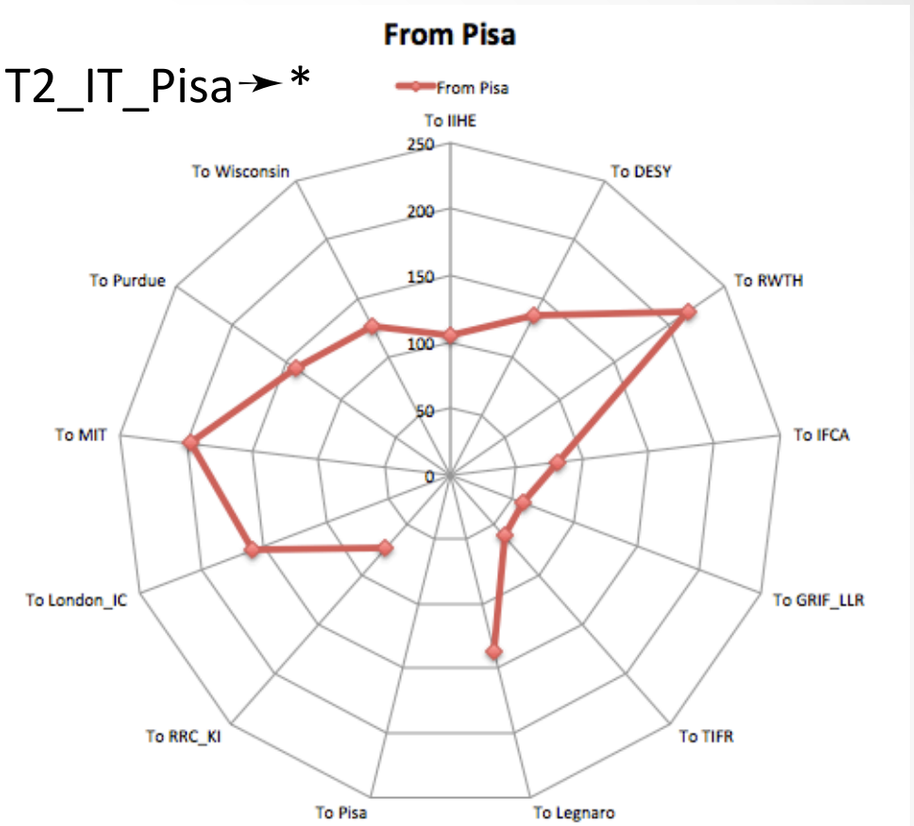
- The T2-T2 commissioning work in CMS was done in 2010.
  - connections need to be tested BEFORE and AFTER any change to the network
- CMS decided to adopt a strategy consisting of two complementary approaches
  - one based on PhEDEx LoadTest infrastructure;
  - one based on FTS/FTM;
  - the first approach is CMS specific, the second is general;
  - the first approach allows to map the performances of the T2-T2 connections and monitor them over time
    - in terms of transfer rates, transfer quality, transfer latency to transfer a few TBs sample
- The first round of such test activities is done. We are now in the process of testing connections from/to sites which enter the LHCONE prototype, one by one

A snapshot of test results

E.g. rates *➔T2_IT_Pisa



E.g. rates T2_IT_Pisa➔*



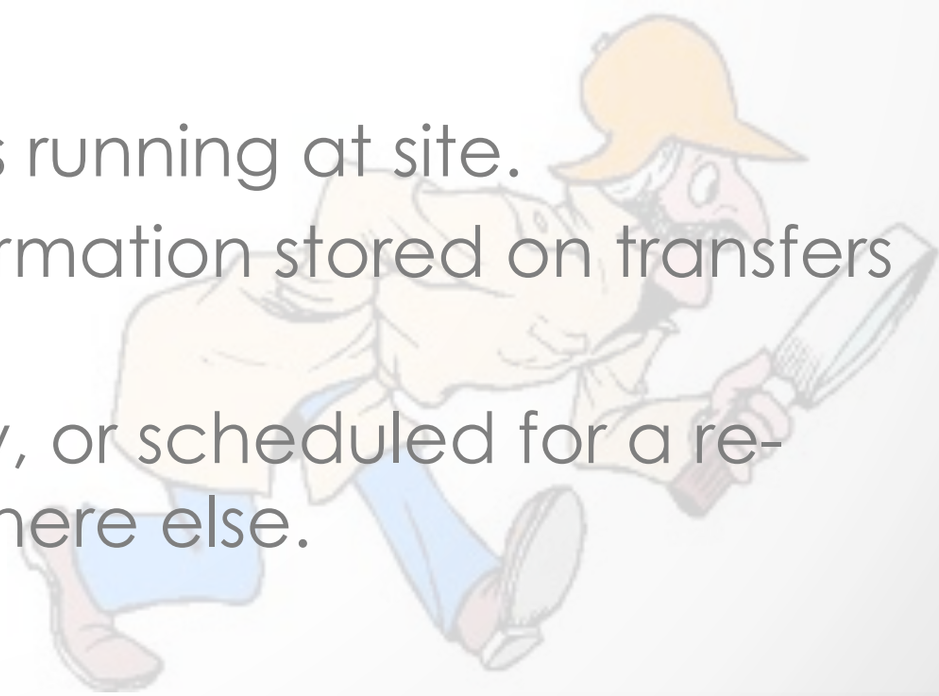
	Max rate in 1 hr [MiB/s] to / from	BE IIHE	DE DESY	DE RWTH	ES IFCA	FR GRIF LLR	IN TIFR	IT Legnaro	IT Pisa	RU RRC KI	UK London IC	US MIT	US Purdue	US Wisconsin
BE	IIHE		85	105	49	50	60	96	83	38	79	75	80	76
DE	DESY	105		518	85	62	61	148	126	65	182	260	109	256
DE	RWTH	97	144		86	108	88	164	112	74	229	255	103	183
ES	IFCA	77	85	97		61	76	87	113	66	102	122	72	136
FR	GRIF_LLR	107	132	449	76		77	145	96	57	320	279	133	368
IN	TIFR													
IT	Legnaro	87	109	196	64	57	47		105	62	79	171	114	180
IT	Pisa	105	135	217	81	59	61	137		74	160	197	141	126
RU	RRC_KI	42	68	119	28	n/a	42	51	49		117	99	77	97
UK	London_IC	64	110	414	93	88	87	139	132	63		305	116	287
US	MIT	108	89	422	84	68	65	133	133	59	39		72	428
US	Purdue	101	55	314	55	48	75	75	138	48	427	320		408
US	Wisconsin	102	105	365	81	43	86	139	108	62	100	330	85	

Data consistency

- Meta-data about all CMS production data is being stored separately (in databases) from the real data (on disks, tapes) and should stay consistent.
- Inconsistency can appear due to hardware failure.
- Several different points to check:
  - all files stored in the database should actually exist on the storage elements
  - files, stored on the storage elements, are complete
  - should be no dark data – files, that we don't know about
- Monthly consistency check are being done at all Tier-1s.
- Being extended to Tier-2s.

How checks are being done?

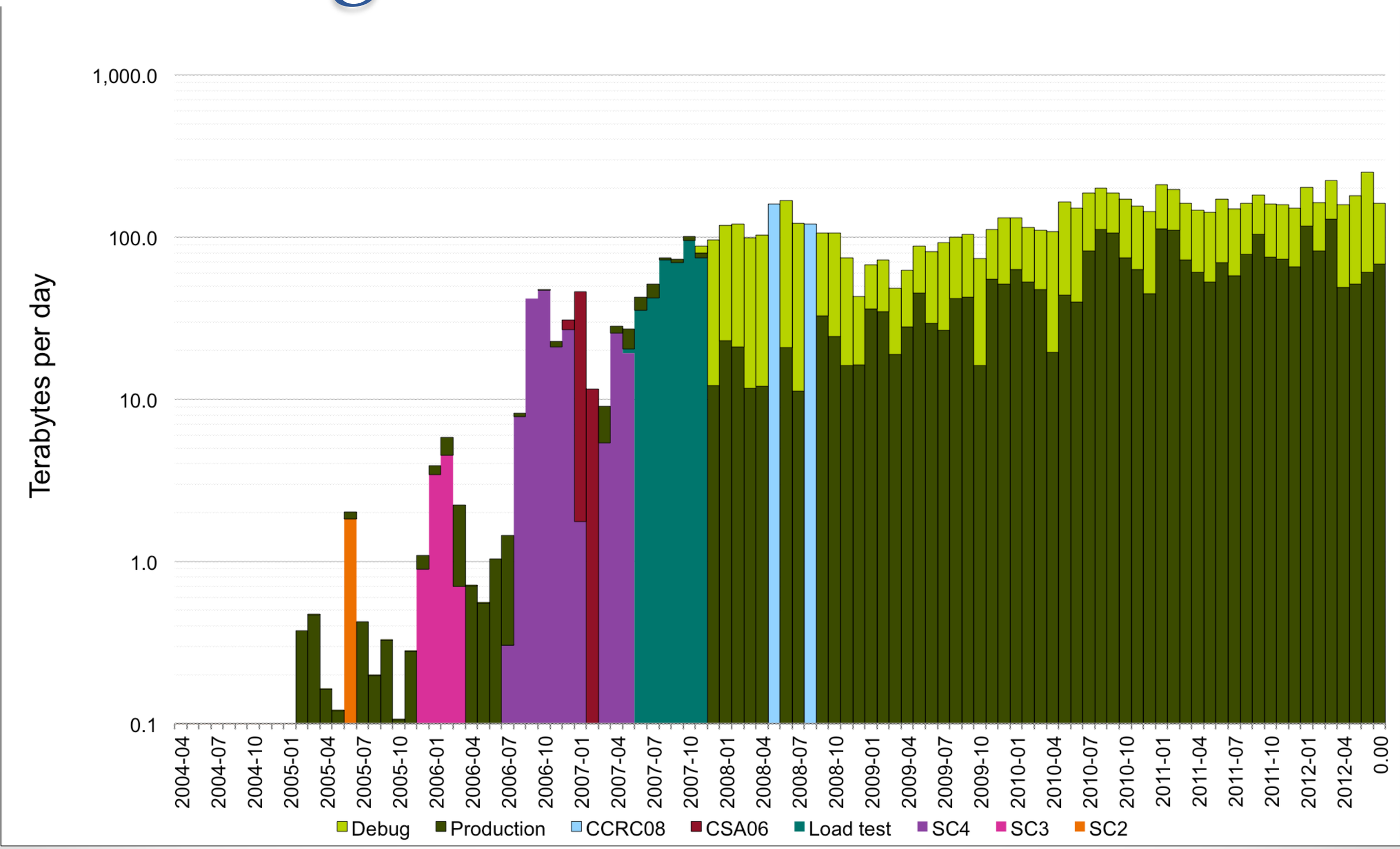
- Storage consistency check (SCC):
  - List of all files older than one month are being gathered from the site.
  - List is checked against internal databases.
  - Present files not found in databases are called orphans.
  - Orphans have to be double checked:
    - If they are 'real' orphans, they get deleted.
    - If they are 'fake' – they get registered in databases properly.
- Block download verify (BDV):
  - Done centrally, if PhEDEx BDV agent is running at site.
  - Agent checks file and its size with information stored on transfers database.
  - Failures either get deleted completely, or scheduled for a re-transfer to site, if it is available somewhere else.



Consistency challenges

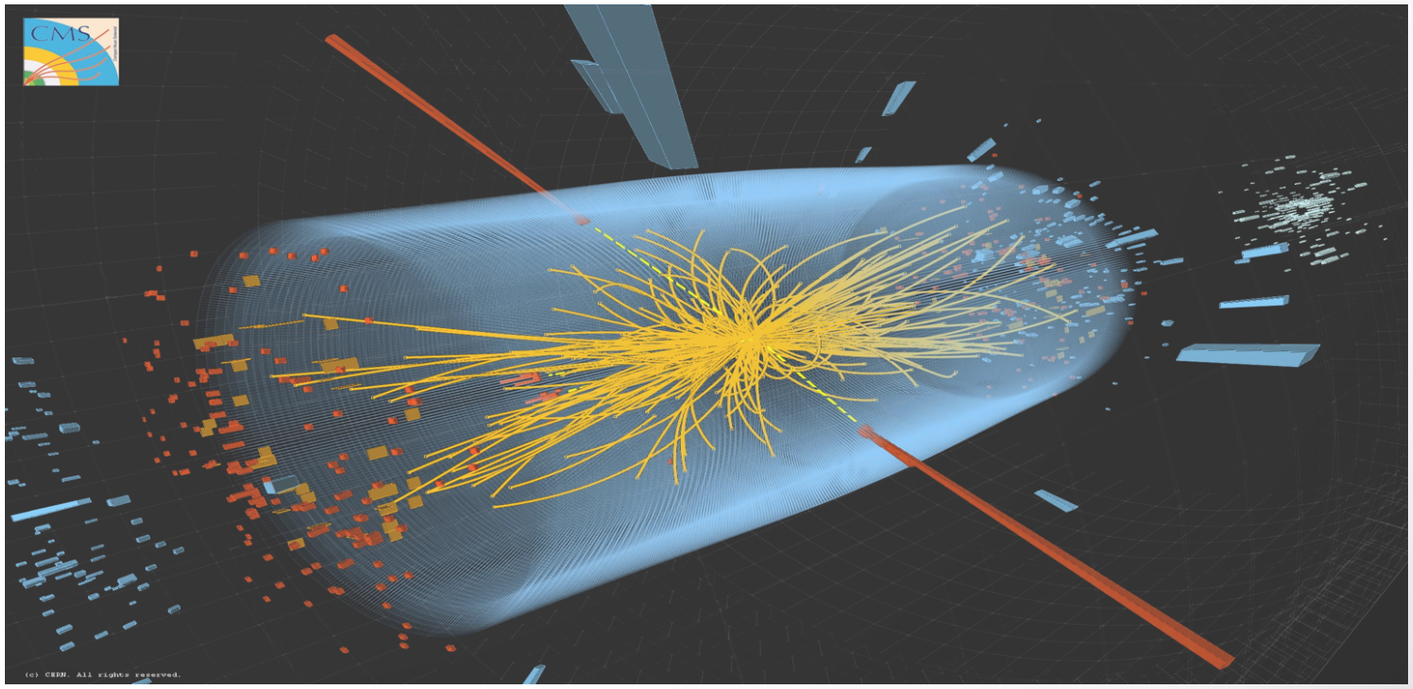
- Big number of sites (there are >50 Tier-2s).
- ~20 millions (~50PB) of files stored across T0, T1s and T2s.
- Big variation of storage technologies (CASTOR, dCache, DPM, Lustre, Hadoop…).
- Aiming for a full consistency is operationally challenging:
  - Not all sites are running PhEDEx BDV agent or running some old version not compatible with central DB or BDV agent is not properly configured, therefore either reporting many failures or not reporting at all.
  - Can not be fully automated (failures in some steps might cause data loss), but some things can be done to ease the process.

Average data transfer volume



Summary

- Transfers are running, and stand up for CMS needs.
- Monitoring takes a lot of effort in both – time and man-power to ensure smooth data transfer.
- Transfer operations can and will be partially automated
  - main issue to fully automatize - error messages from various components of the system are not unified on the same problem
    - e.g. different storage systems report different error messages
- Operators are fully prepared to meet this year transfers' challenges.



Questions?

Thank you ☺