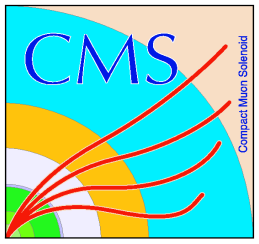
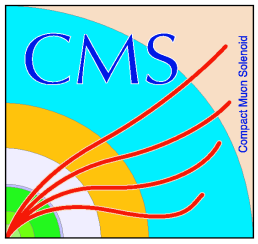


LISHEP 2006
Rio de Janeiro

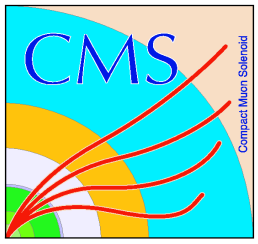
PhEDEx –
reliable and scalable data distribution on the Grid

Jens Rehn, CERN
On behalf of numerous
PhEDEx contributors



Outline

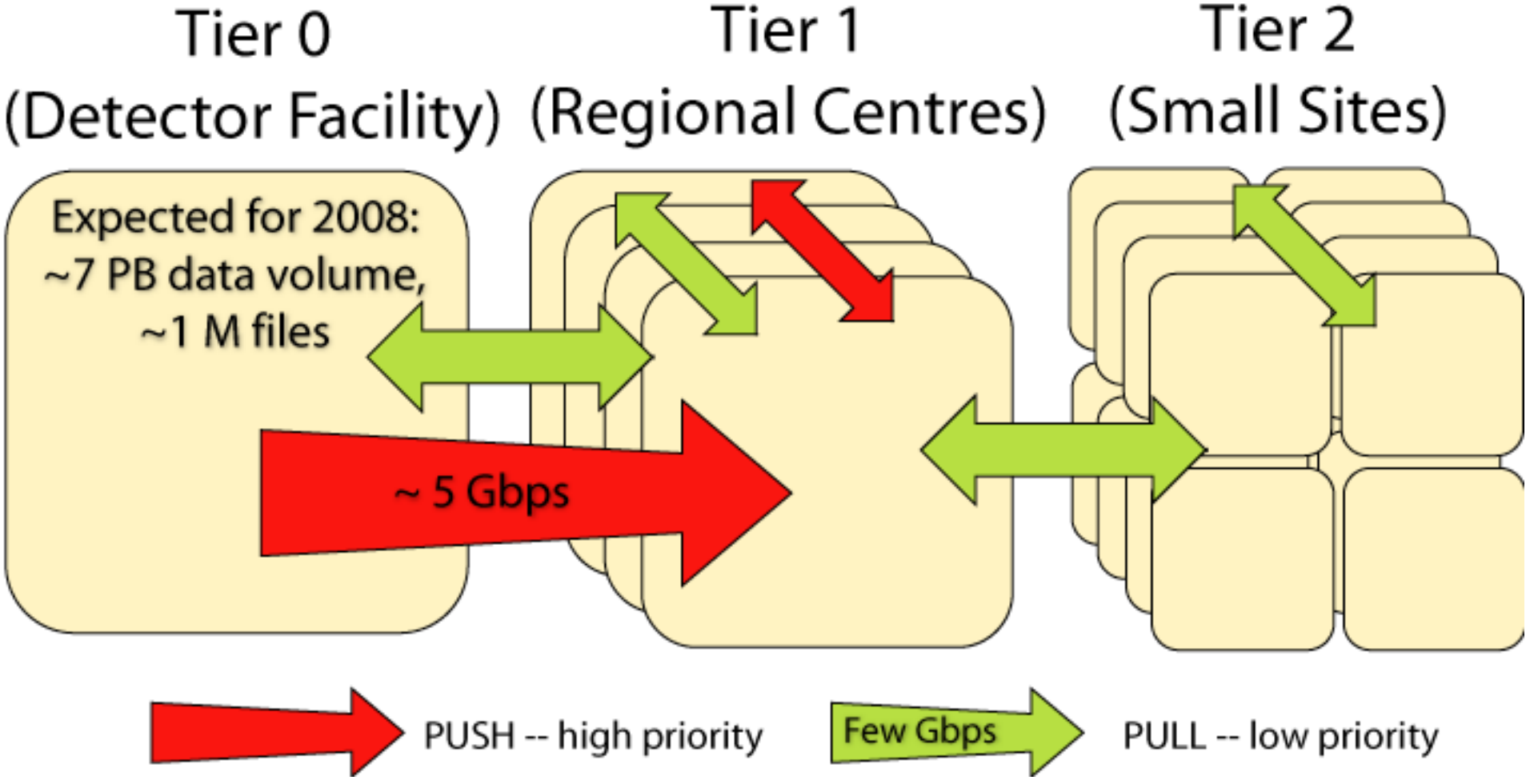
- Introduction to PhEDEx
  - Traditional and PhEDEx data transfers
  - Features and functionality
- Operating and monitoring a live PhEDEx system
- Practical examples from the last service challenge
  - Performance and scalability
  - Analysis of transfer related problems
- How to set-up and run PhEDEx
- Conclusions and outlook



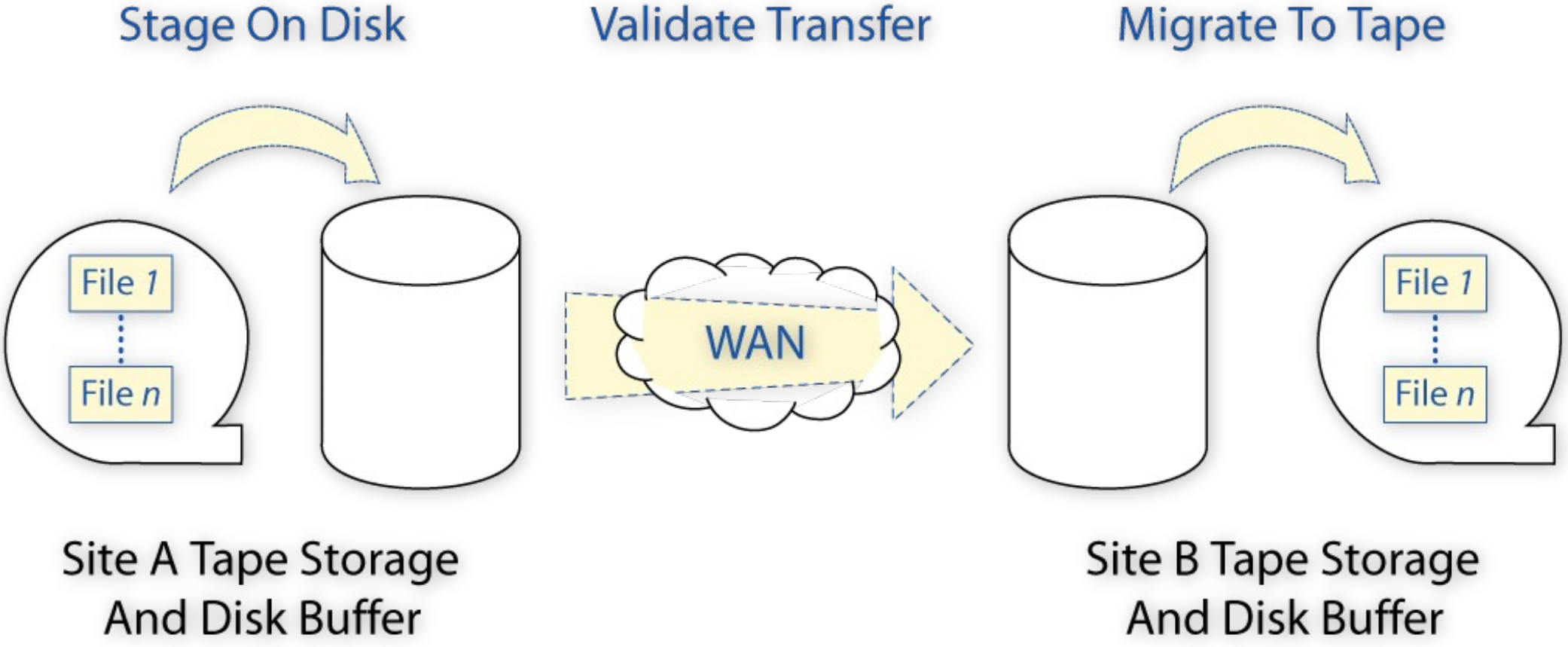
CMS data flow

- Detector data distribution @ high priority
  - One copy at Cern; one distributed copy at regional centres
  - Expected transfer volume for 2008: ~ 7 PB ≈ O(10M) files
  - Required transfer speed for 2008: ~ 5 Gb/s
- Simulated data distribution @ low priority
  - Among and between regional and local centres
  - Expected bandwidth utilisation: few Gb/s per regional centre
- Data structured in blocks of files
  - dataset, datatiers

Tiered data flow



Traditional HEP data replication



- Each step done by hand
- Manpower-intensive
- Feasible only for small amount of files

PhEDEx data replication

Web User Interface

Transfer Request Management

Block Subscription Management

A

Block Level Data Transfers

A A

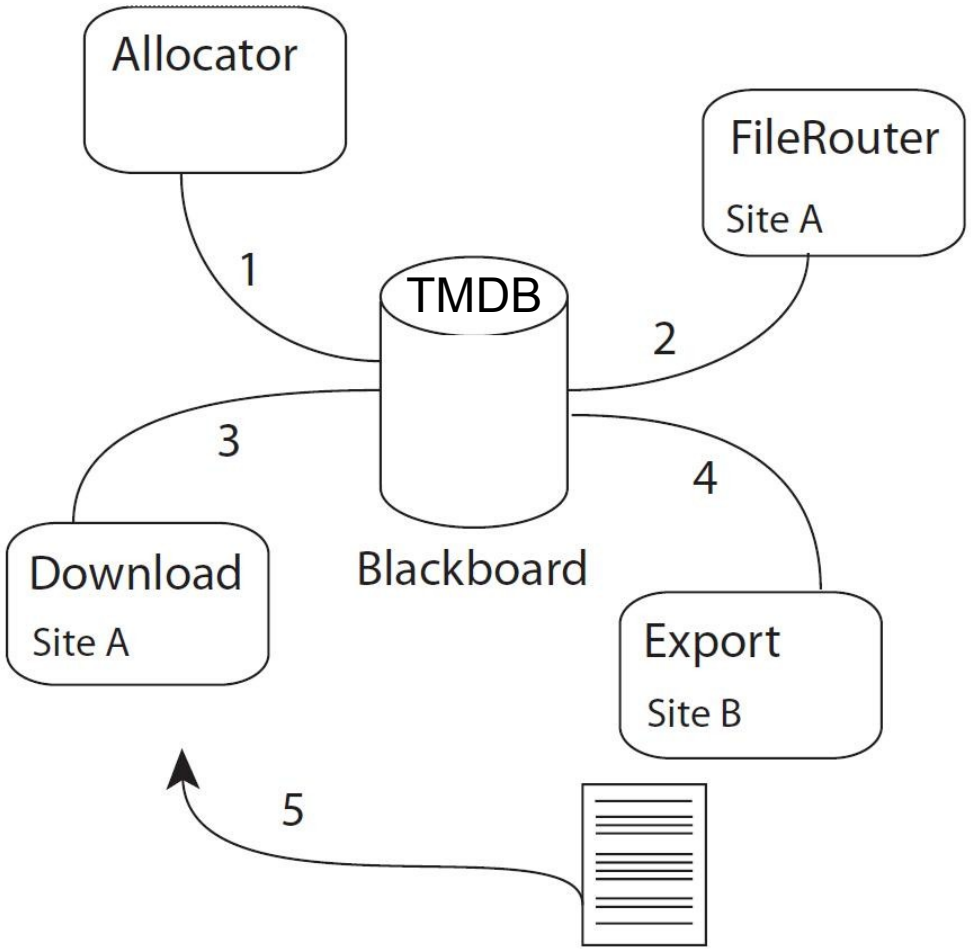
Logical Storage Nodes with PhEDEx Agents

Unreliable Point-to-Point Transfers

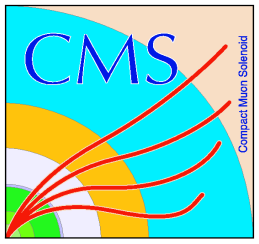
Reliable Point-to-Point Transfers (Single-Hop)

Reliable Routed Transfers (Multi-Hop)

PhEDEx file replication

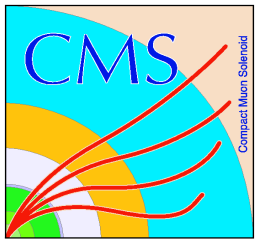


1. Allocator: allocate files to destinations
2. FileRouter: maintains & determines best routes
3. Download: marks files „wanted“ from site B
4. Export: initiate staging & provide contact information
5. Download: transfer file



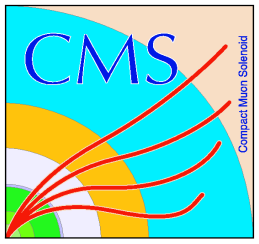
HEP requirements for a data distribution system

- Managed & structured data flow
- Reliability
  - Robustness & self-healing
    - Error recovery, automatic back-off, etc
  - Integrity of replicated data
- Flexibility
  - Different transfer models: push and pull
  - Support of common transfer protocols & storage systems
- Monitoring



PhEDEx – design
Key features (1)

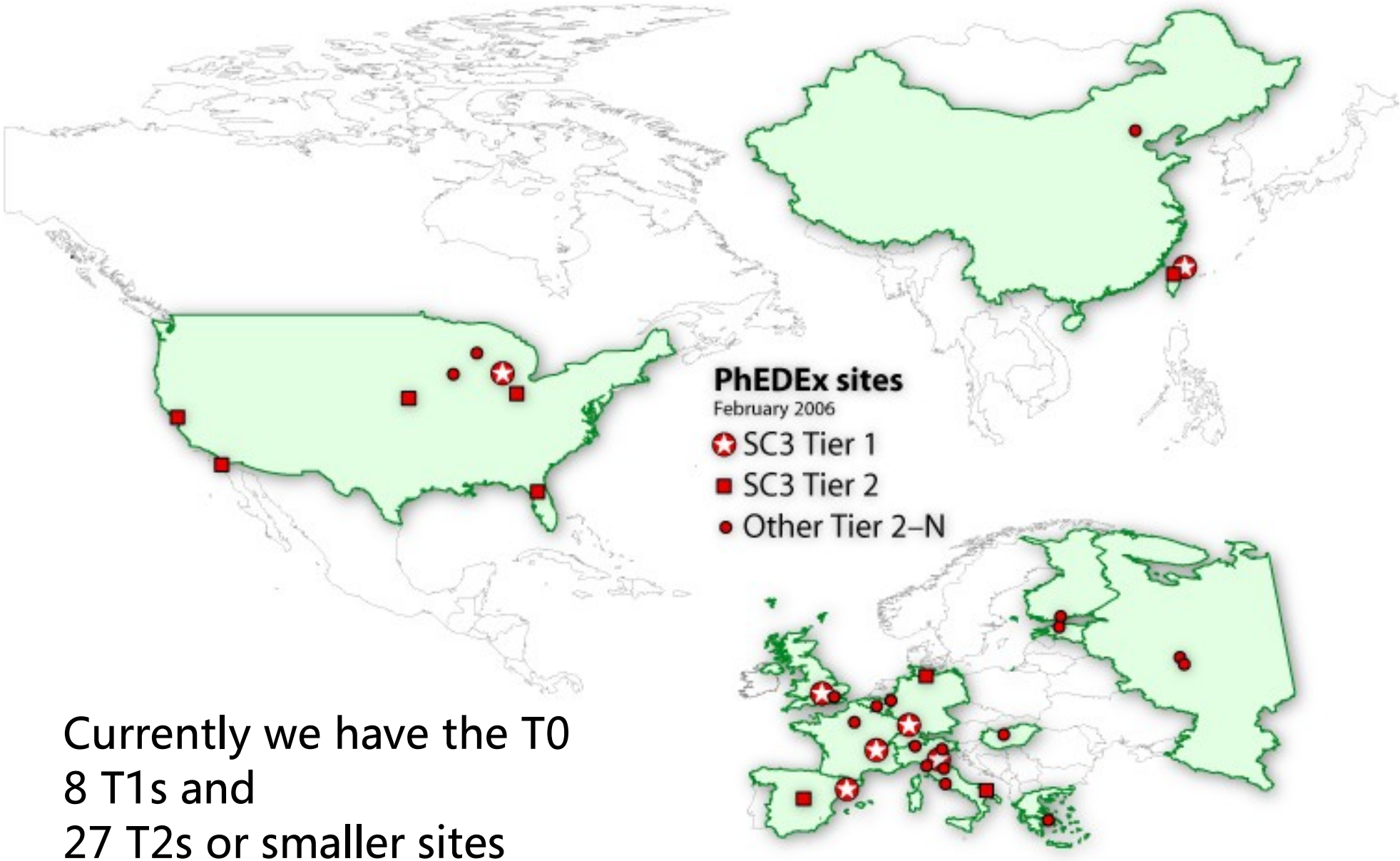
- Reliability
  - Transfer status monitored
  - File size check after each replication
  - Cksum for every file in TMDB available for further checks
  - Automatic cool off for failed transfers
  - Self-throttling: limits amount of parallel transfers
  - Designed under assumption: any operation might fail
- Monitoring
  - Status web page: http://cern.ch/cms-project-phedex



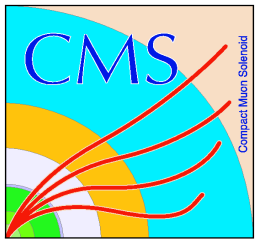
PhEDEx – design
Key features (2)

- Flexibility
  - Push and pull models supported: logical implementation
    - Push: data subscribed to destination by site hosting replica
    - Pull: destination site subscribes data to itself
  - Automatic protocol matching: G-U-C, srmcp, dccp, rfcp, FTS
  - Intelligent routing with fall-back mechanism
- Operability
  - Easy to handle deployment
  - Linux *inetd* like start/stop mechanism for agents
  - Try to provide easy to understand log messages

PhEDEx distribution network



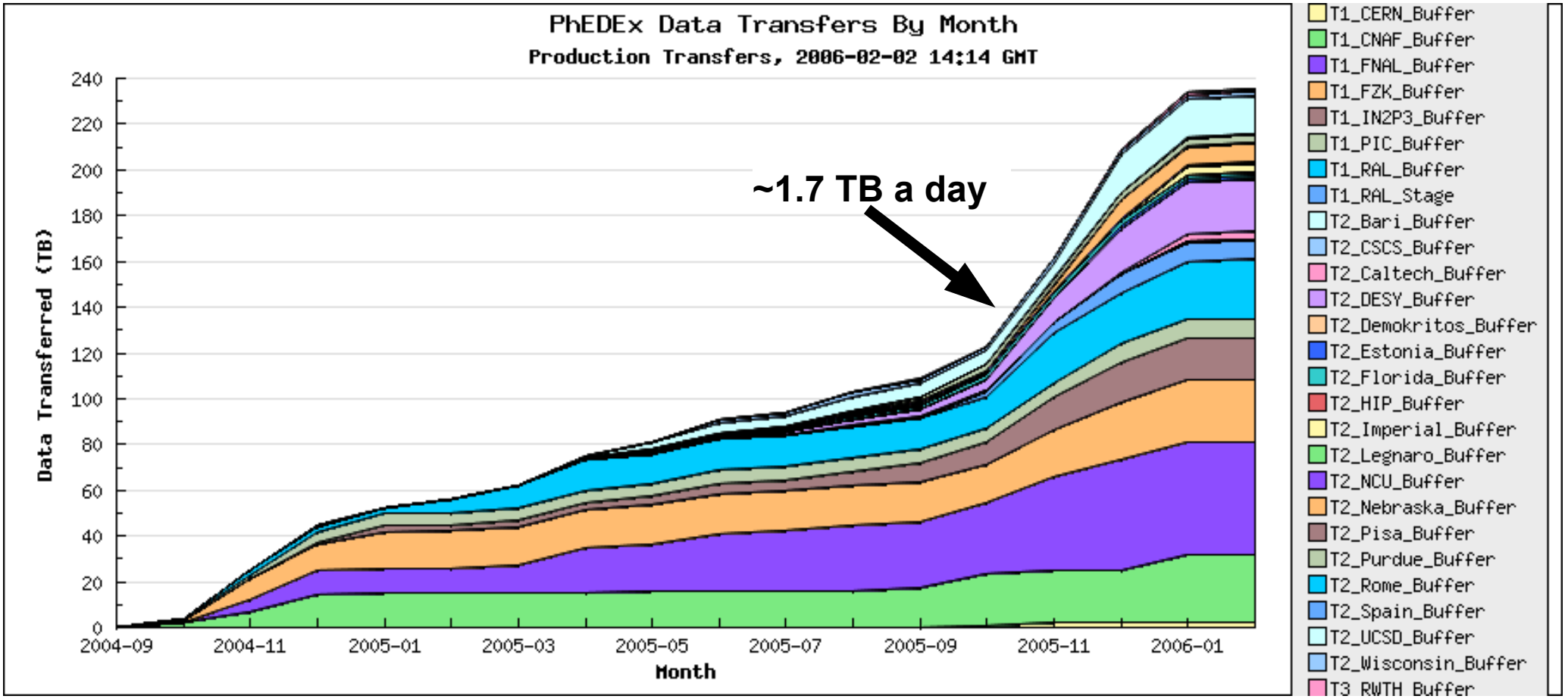
Currently we have the T0
8 T1s and
27 T2s or smaller sites



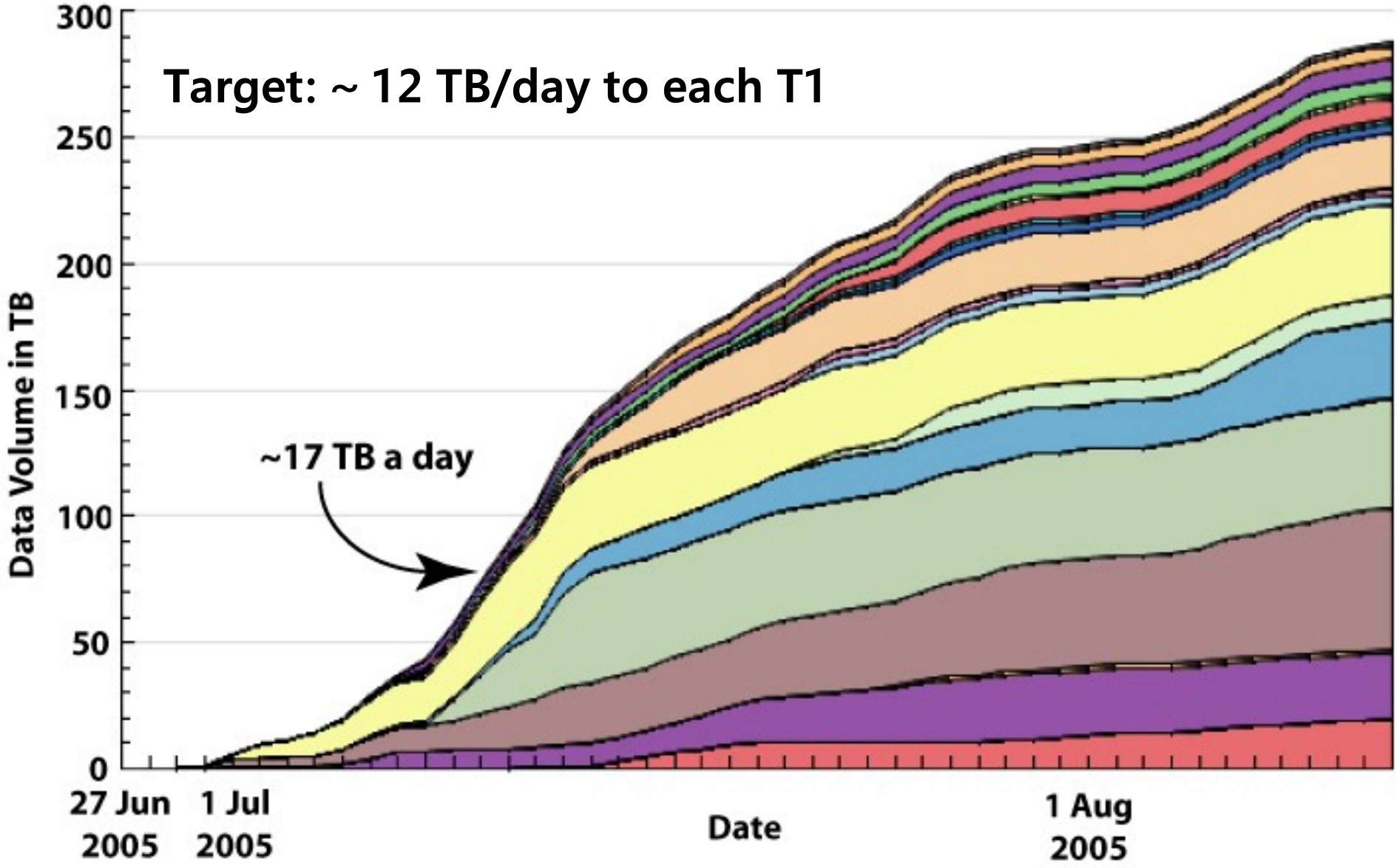
PhEDEx – design
Intelligent routing

- Routing agent determines best route: source → destination
- Routes are ranked automatically
  - According to amount of intermediate nodes: hops
  - Hops can be weighted
- IP-like routing to route files to destination
  - In case of outage, fallback routes chosen via other nodes
  - Dead nodes noticed by neighbours; no heartbeat

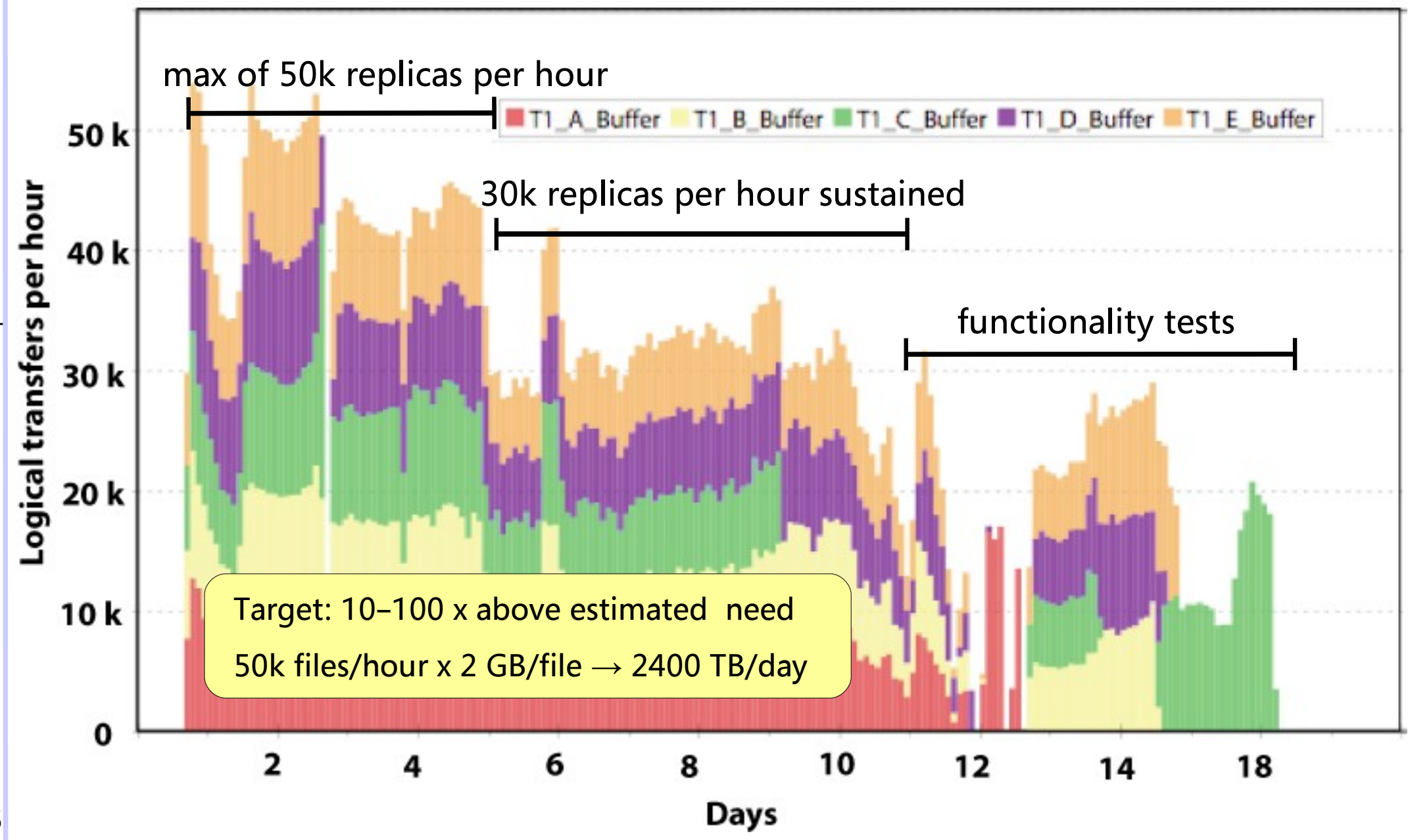
PhEDEx transfer volume
last 17 month



PhEDEx in LCG SC3 throughput phase

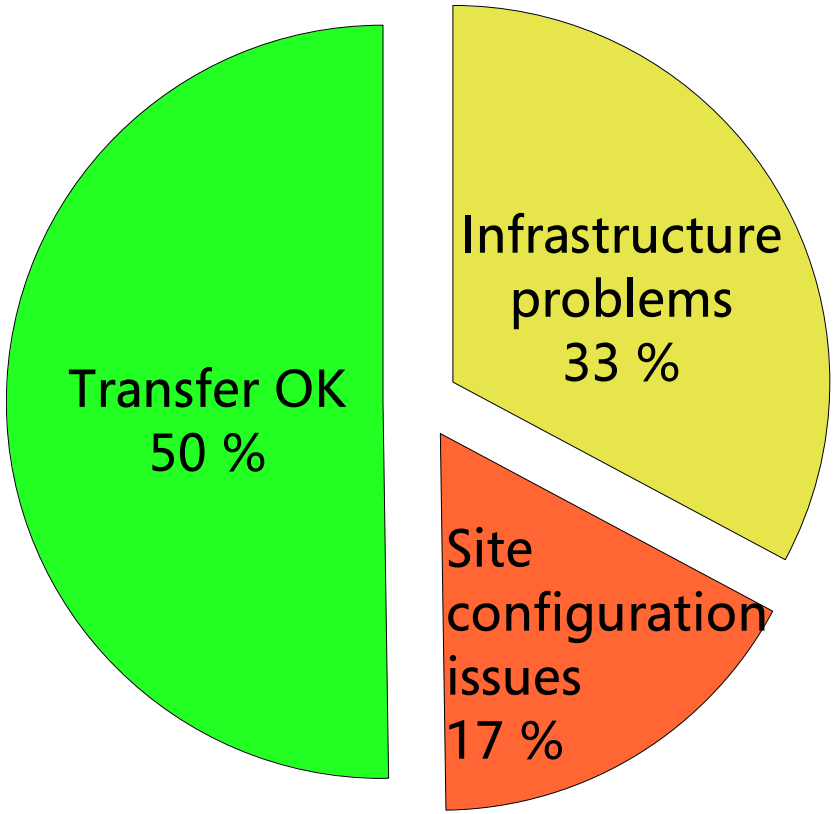


PhEDEx scalability exercise



Reliability: impossible odds?

Case study: transfer level

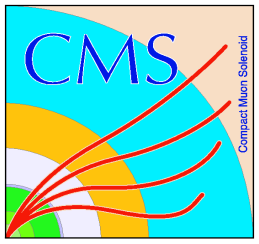


- High failure rate on new SRM/storage infrastructure
- 50% of the transfers successful on the first try
- Main problems
  - Configurations changed or wrong at sites
  - Problems related to network or storage infrastructure

Reliability: against the odds!

Case study: after PhEDEx failure recovery

- All failures recovered, eventually
- Files retransferred
- No data lost :-)
- Recovery fully automatic
  - Absolute must: in 2007 CMS will transfer ~2-10k files per day
  - Manual recovery infeasible: 1 ‰ permanent error rate ≃ 2 hrs daily maintenance

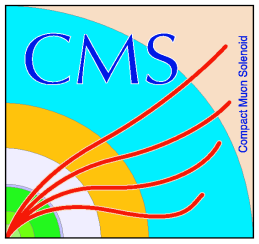


PhEDEx – in practice
Monitoring & subscription

How to monitor transfers

How to subscribe to CMS datasets

http://cern.ch/cms-project-phedex



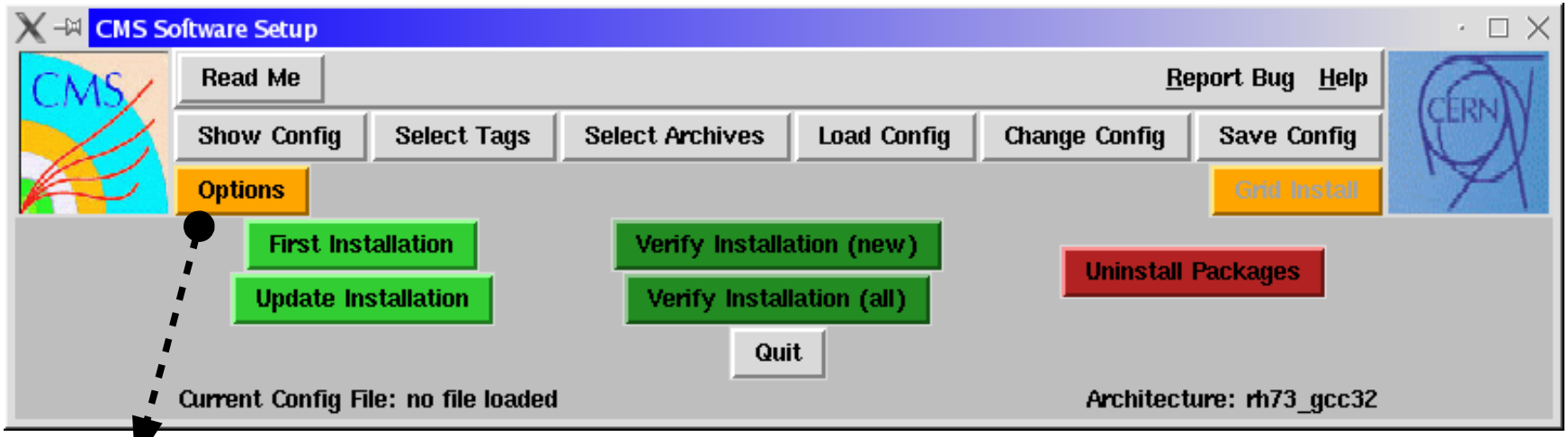
PhEDEx – deployment
Overview

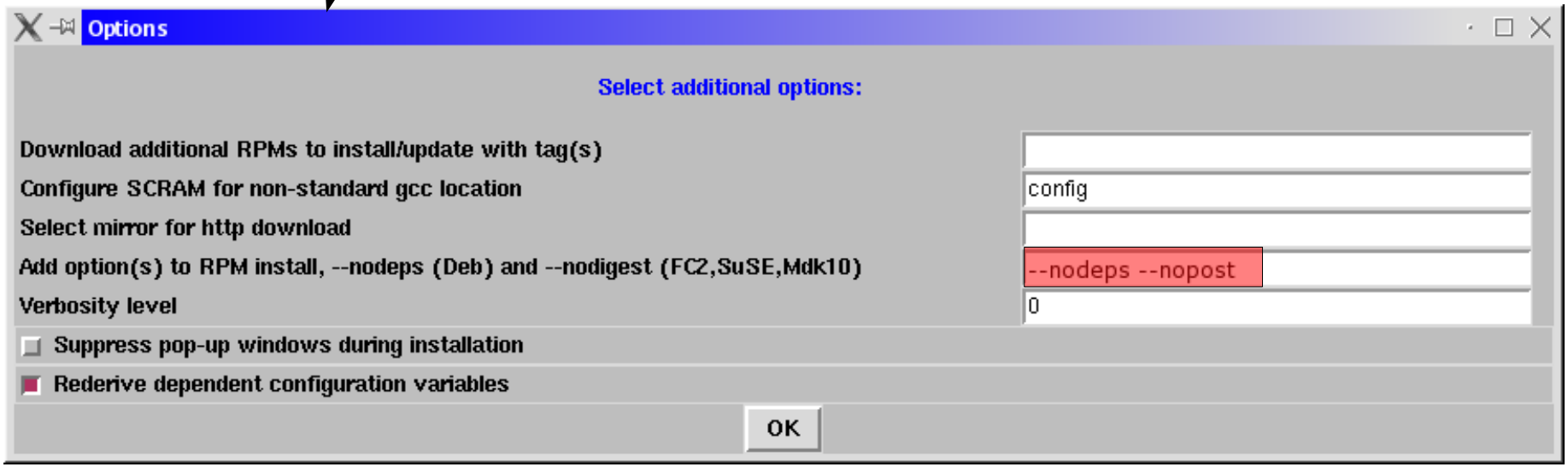
- Each site runs agents close to their storage
  - Modest resource requirements
  - Usually hosted on CMS-dedicated server
  - PhEDEx and tool installation with XCMSi
  - Underlying transfer utilities like srmcp, fts, etc.
  - Grid services: certificates and proxy renewal
  - Configuration: site registration and site specific settings
- Operated by local CMS community, in close communication with site's adminstrators

PhEDEx – deployment
Software via XCMSi (1)

- Decide where to install Phedex
  - Create install dir for XCMSi ($xcmsi-base)
  - Create basedir for PhEDEx ($phedex-base)
- Download packages
  - XCMSi from http://cern.ch/cms-xcmsi
  - Untar XCMSi to $xcmsi-base
  - Get Oracle client libraries (zip): http://www.oracle.com
    - put them in sub-dir $xcmsi-base/ZIPS
- Start the installation GUI
  - cd $xcmsi-base; ./xcmsi.pl

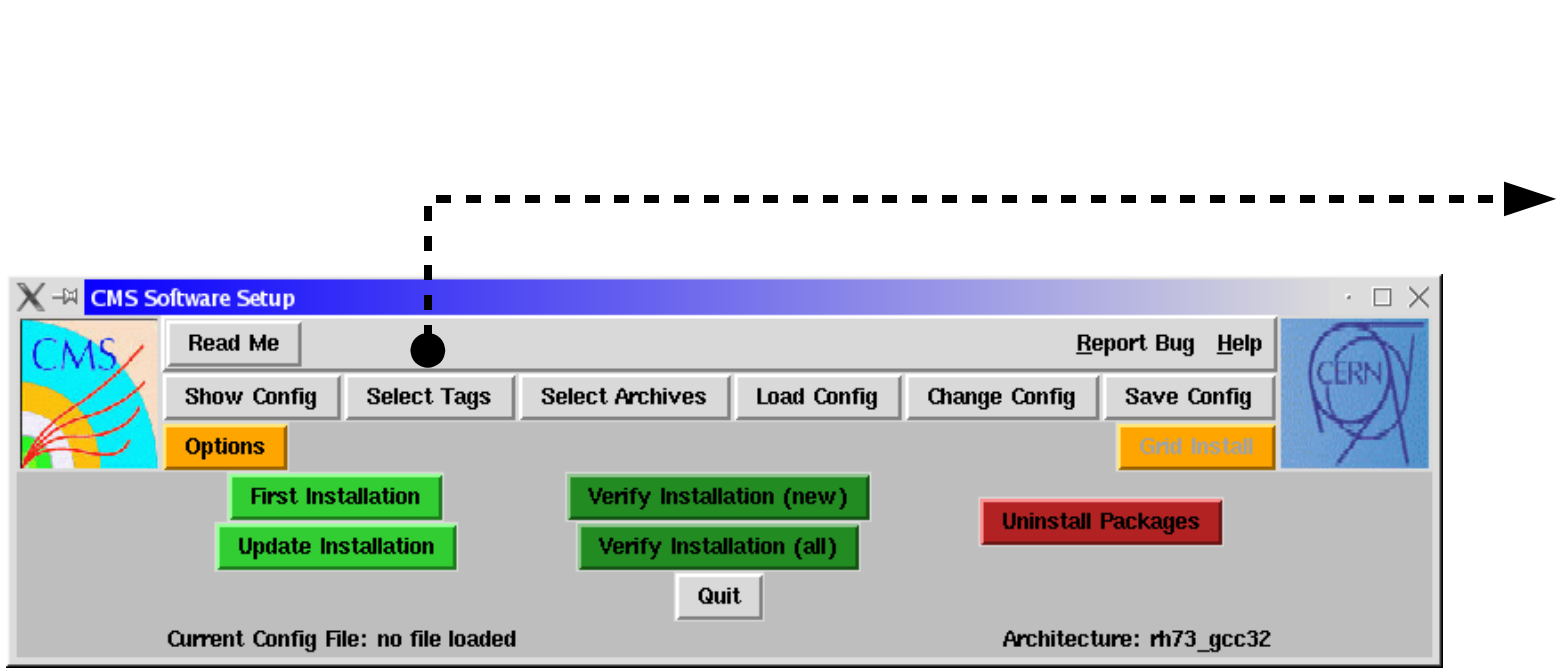
PhEDEx – deployment
Software via XCMSi (2)

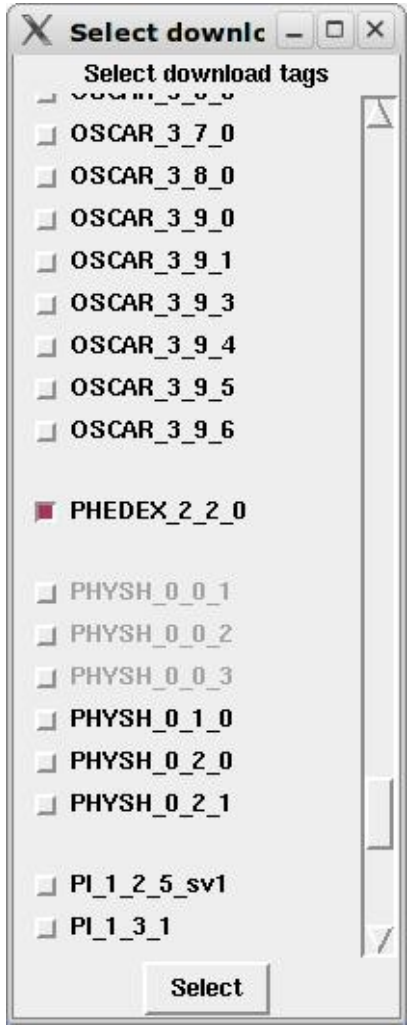




No post installation scripts and no dependency checking

PhEDEx – deployment
Software via XCMSi (3)

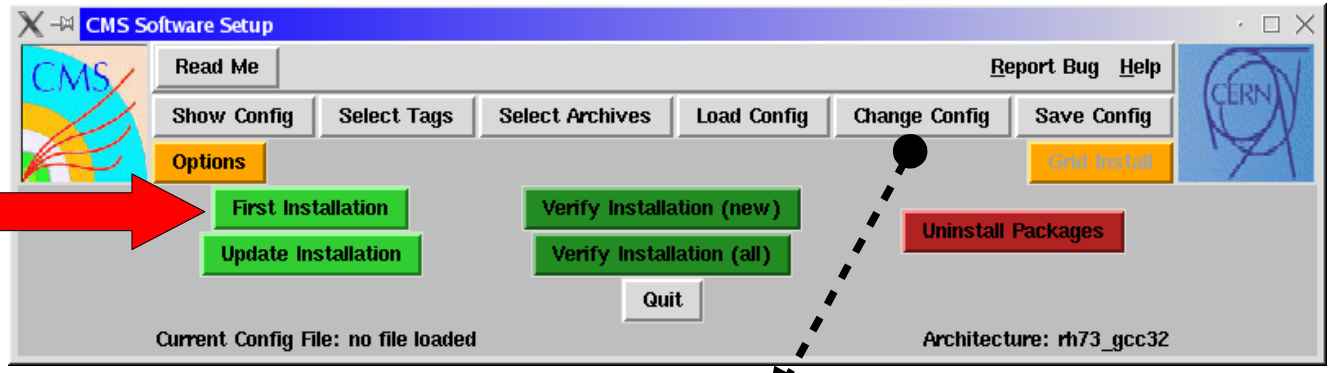


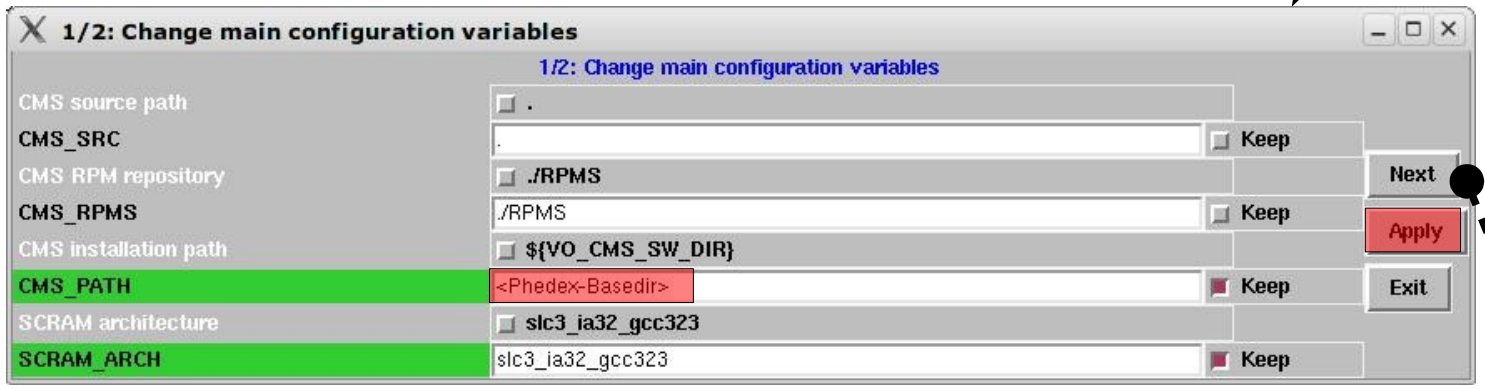


Select PhEDEx version

PhEDEx – deployment
Software via XCMSi (4)

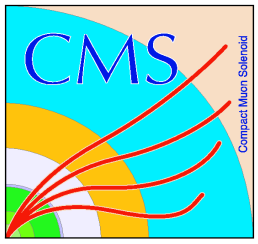
3. Finally start installation





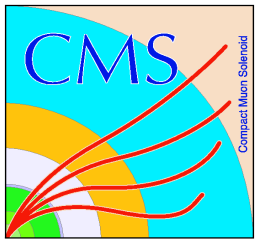
1. Select PhEDEx installation dir

2. Select PhEDEx tools dir



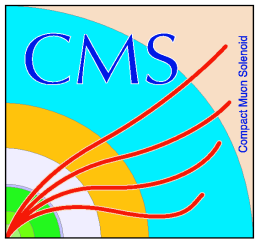
PhEDEx – deployment
Grid services

- Site local file catalogue – if you don't have one already
  - Any catalogue type is fine
  - MySQL based POOL file catalogue recommended:
    - Helper script PHEDEX/Deployment/SetupPOOLFileCatalogue
- Certificate management
  - Valid Grid certificate proxy: grid-proxy-init
  - Recommended auto-renewal via myproxy



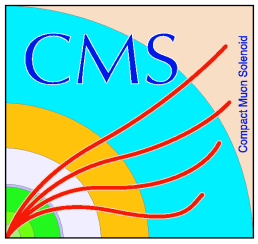
PhEDEx – deployment
Configuration

- Site registration in central DB
  - Obligatory: Documents/README/README-Deployment
  - Currently: send mail to hn-cms-phedex@cern.ch
- Site local glue scripts
  - Get a copy of templates provided in
    - Custom/Template
  - Adjust them to meet your site's requirements
    - Remove all not needed agents ! Typically only ~ 5 are needed
- Testing your installation
  - Run Deployment/TestInstallation



Summary and outlook

- PhEDEx provides
  - Reliable and scalable data distribution on the grid
  - Flexibility to use any file replication tool, especially grid tools
  - Real life monitoring through a web status display
- Plans
  - More and improved web based data management tools
    - data subscriptions, transfer requests, agent management, deployment
  - Support transfers for physics groups and individual physicists
    - decentralisation of central database
- Hope to bring many new sites onboard :-)

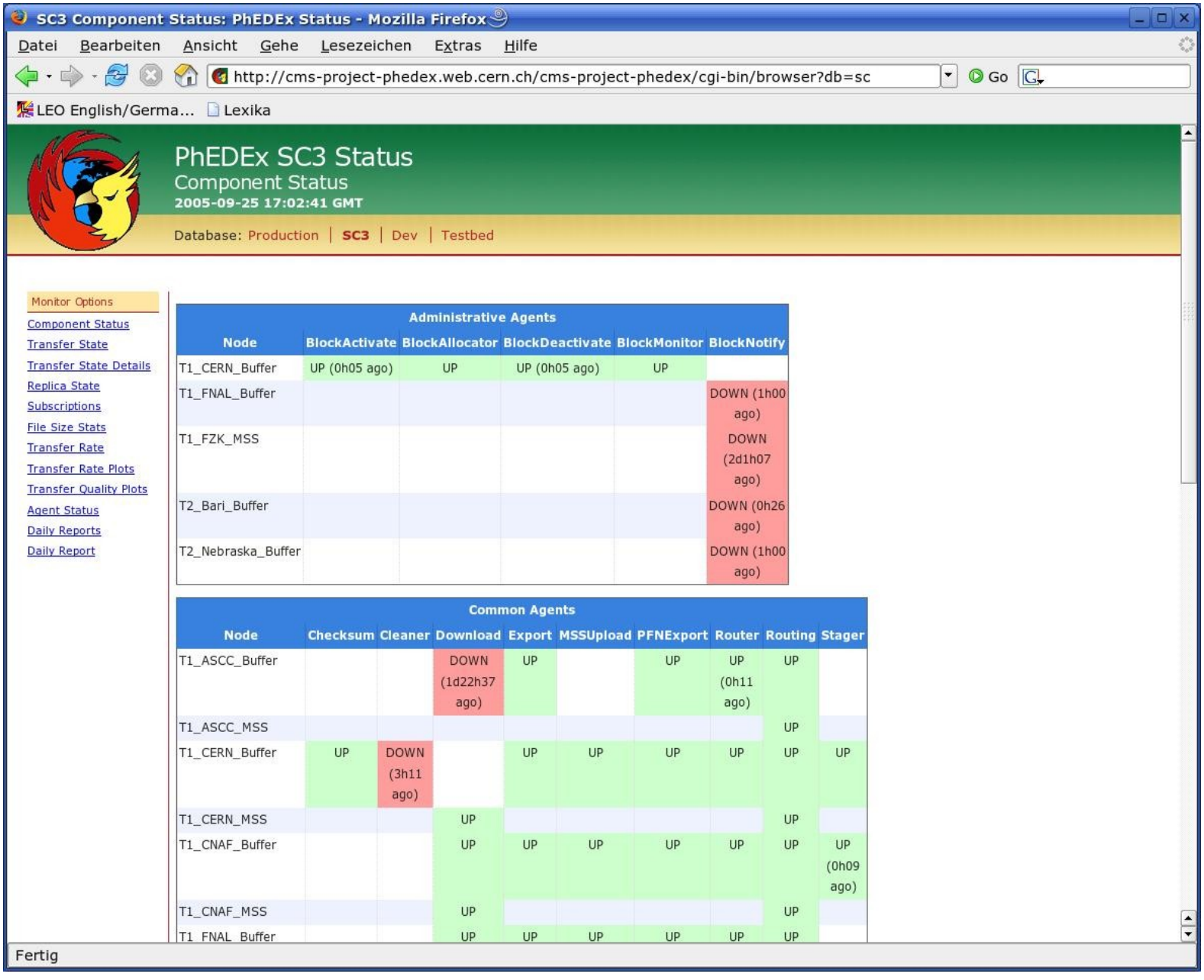


Useful links & contacts

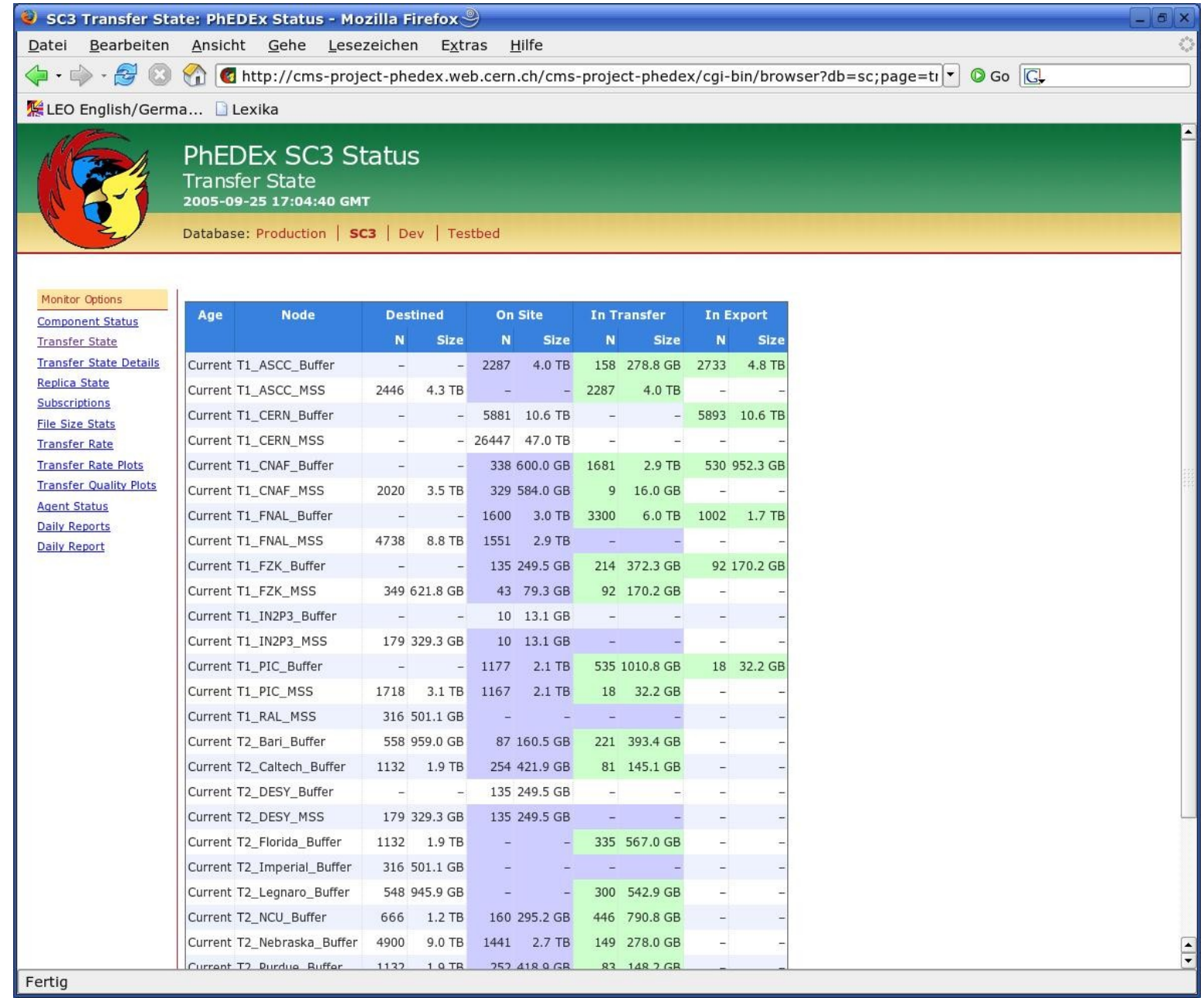
- PhEDEx project web page:
  - http://cern.ch/cms-project-phedex
  - links to documentation, monitoring & CVS repository
- PhEDEx mailing list:
  - hn-cms-phedex@cern.ch

Big Thank You to the organizers!!

PhEDEx – monitoring
Component status

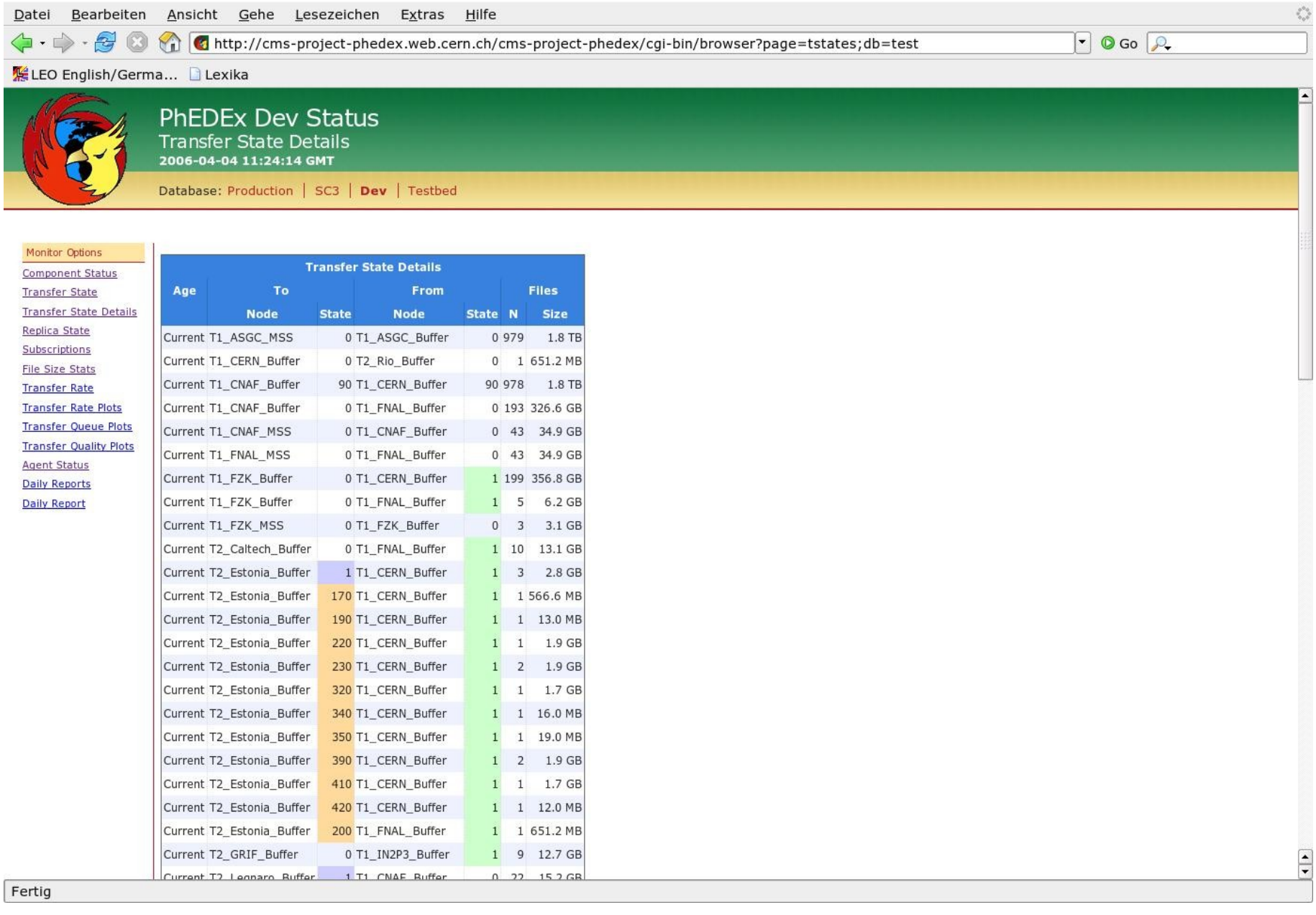


PhEDEx – monitoring
Transfer state



Age	Node	Destined		On Site		In Transfer		In Export	
		N	Size	N	Size	N	Size	N	Size
Current	T1_ASCC_Buffer	–	–	2287	4.0 TB	158	278.8 GB	2733	4.8 TB
Current	T1_ASCC_MSS	2446	4.3 TB	–	–	2287	4.0 TB	–	–
Current	T1_CERN_Buffer	–	–	5881	10.6 TB	–	–	5893	10.6 TB
Current	T1_CERN_MSS	–	–	26447	47.0 TB	–	–	–	–
Current	T1_CNAF_Buffer	–	–	338	600.0 GB	1681	2.9 TB	530	952.3 GB
Current	T1_CNAF_MSS	2020	3.5 TB	329	584.0 GB	9	16.0 GB	–	–
Current	T1_FNAL_Buffer	–	–	1600	3.0 TB	3300	6.0 TB	1002	1.7 TB
Current	T1_FNAL_MSS	4738	8.8 TB	1551	2.9 TB	–	–	–	–
Current	T1_FZK_Buffer	–	–	135	249.5 GB	214	372.3 GB	92	170.2 GB
Current	T1_FZK_MSS	349	621.8 GB	43	79.3 GB	92	170.2 GB	–	–
Current	T1_IN2P3_Buffer	–	–	10	13.1 GB	–	–	–	–
Current	T1_IN2P3_MSS	179	329.3 GB	10	13.1 GB	–	–	–	–
Current	T1_PIC_Buffer	–	–	1177	2.1 TB	535	1010.8 GB	18	32.2 GB
Current	T1_PIC_MSS	1718	3.1 TB	1167	2.1 TB	18	32.2 GB	–	–
Current	T1_RAL_MSS	316	501.1 GB	–	–	–	–	–	–
Current	T2_Bari_Buffer	558	959.0 GB	87	160.5 GB	221	393.4 GB	–	–
Current	T2_Caltech_Buffer	1132	1.9 TB	254	421.9 GB	81	145.1 GB	–	–
Current	T2_DESY_Buffer	–	–	135	249.5 GB	–	–	–	–
Current	T2_DESY_MSS	179	329.3 GB	135	249.5 GB	–	–	–	–
Current	T2_Florida_Buffer	1132	1.9 TB	–	–	335	567.0 GB	–	–
Current	T2_Imperial_Buffer	316	501.1 GB	–	–	–	–	–	–
Current	T2_Legnaro_Buffer	548	945.9 GB	–	–	300	542.9 GB	–	–
Current	T2_NCU_Buffer	666	1.2 TB	160	295.2 GB	446	790.8 GB	–	–
Current	T2_Nebraska_Buffer	4900	9.0 TB	1441	2.7 TB	149	278.0 GB	–	–
Current	T2_Purdue_Buffer	1132	1.9 TB	252	418.9 GB	83	148.2 GB	–	–

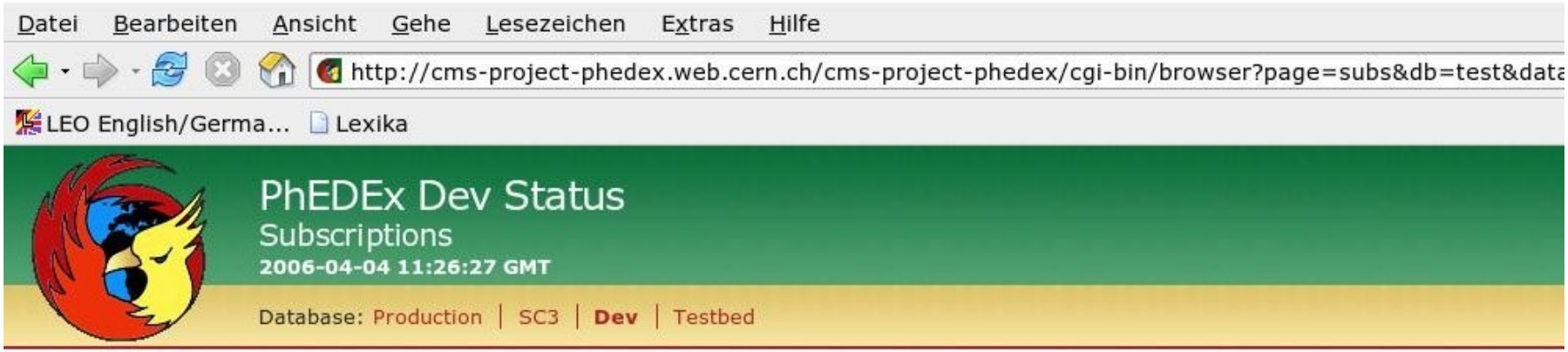
PhEDEx – monitoring
Transfer state details



Transfer State Details

Age	To Node	To State	From Node	From State	Files N	Files Size
Current	T1_ASGC_MSS	0	T1_ASGC_Buffer	0	979	1.8 TB
Current	T1_CERN_Buffer	0	T2_Rio_Buffer	0	1	651.2 MB
Current	T1_CNAF_Buffer	90	T1_CERN_Buffer	90	978	1.8 TB
Current	T1_CNAF_Buffer	0	T1_FNAL_Buffer	0	193	326.6 GB
Current	T1_CNAF_MSS	0	T1_CNAF_Buffer	0	43	34.9 GB
Current	T1_FNAL_MSS	0	T1_FNAL_Buffer	0	43	34.9 GB
Current	T1_FZK_Buffer	0	T1_CERN_Buffer	1	199	356.8 GB
Current	T1_FZK_Buffer	0	T1_FNAL_Buffer	1	5	6.2 GB
Current	T1_FZK_MSS	0	T1_FZK_Buffer	0	3	3.1 GB
Current	T2_Caltech_Buffer	0	T1_FNAL_Buffer	1	10	13.1 GB
Current	T2_Estonia_Buffer	1	T1_CERN_Buffer	1	3	2.8 GB
Current	T2_Estonia_Buffer	170	T1_CERN_Buffer	1	1	566.6 MB
Current	T2_Estonia_Buffer	190	T1_CERN_Buffer	1	1	13.0 MB
Current	T2_Estonia_Buffer	220	T1_CERN_Buffer	1	1	1.9 GB
Current	T2_Estonia_Buffer	230	T1_CERN_Buffer	1	2	1.9 GB
Current	T2_Estonia_Buffer	320	T1_CERN_Buffer	1	1	1.7 GB
Current	T2_Estonia_Buffer	340	T1_CERN_Buffer	1	1	16.0 MB
Current	T2_Estonia_Buffer	350	T1_CERN_Buffer	1	1	19.0 MB
Current	T2_Estonia_Buffer	390	T1_CERN_Buffer	1	2	1.9 GB
Current	T2_Estonia_Buffer	410	T1_CERN_Buffer	1	1	1.7 GB
Current	T2_Estonia_Buffer	420	T1_CERN_Buffer	1	1	12.0 MB
Current	T2_Estonia_Buffer	200	T1_FNAL_Buffer	1	1	651.2 MB
Current	T2_GRIF_Buffer	0	T1_IN2P3_Buffer	1	9	12.7 GB
Current	T2_Legnaro_Buffer	1	T1_CNAF_Buffer	0	22	15.2 GB

PhEDEx – monitoring
Subscriptions



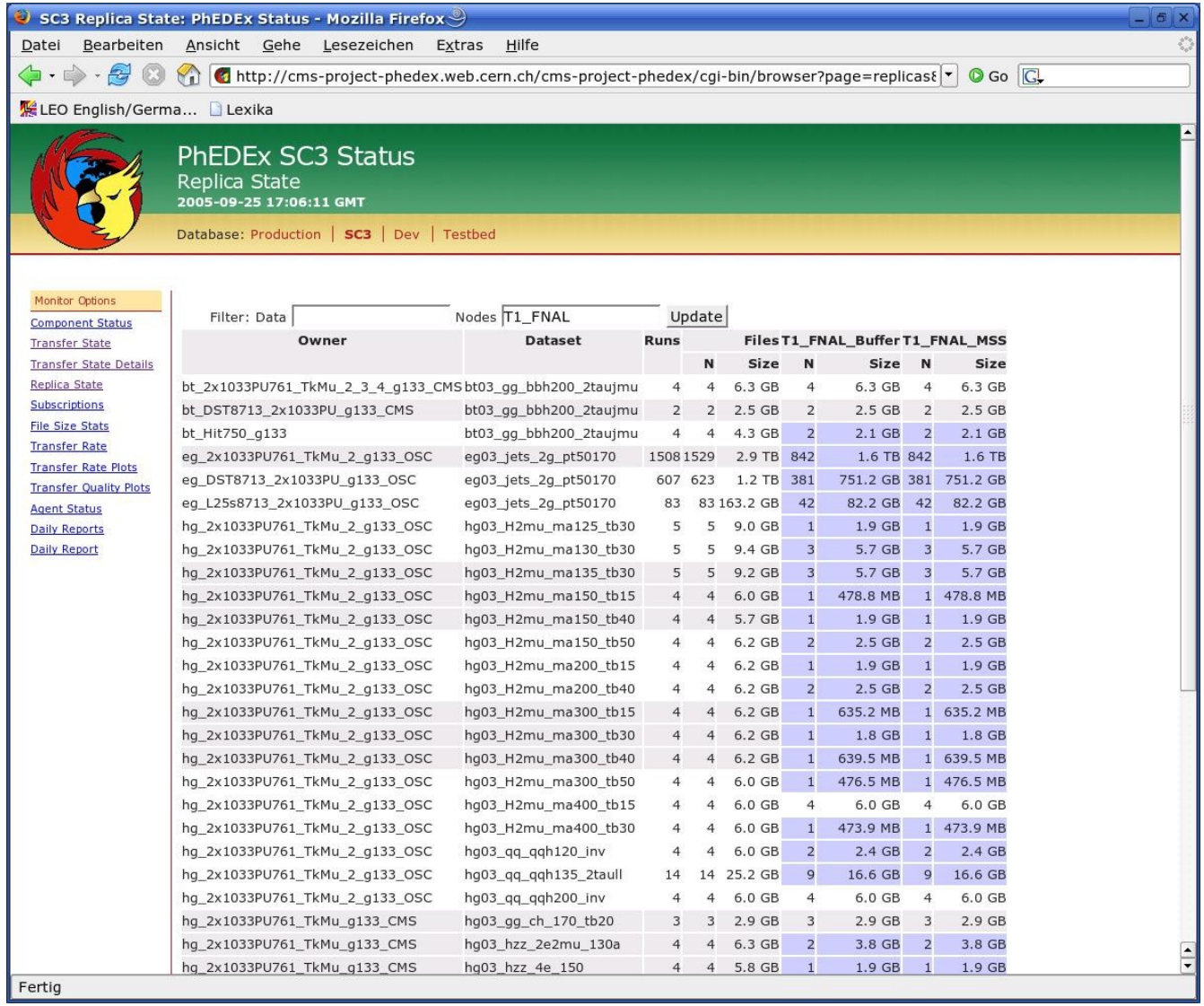
Monitor Options

- Component Status
- Transfer State
- Transfer State Details
- Replica State
- Subscriptions
- File Size Stats
- Transfer Rate
- Transfer Rate Plots
- Transfer Queue Plots
- Transfer Quality Plots
- Agent Status
- Daily Reports
- Daily Report

Filter: Data | Nodes CERN|FNAL|Rio | Update

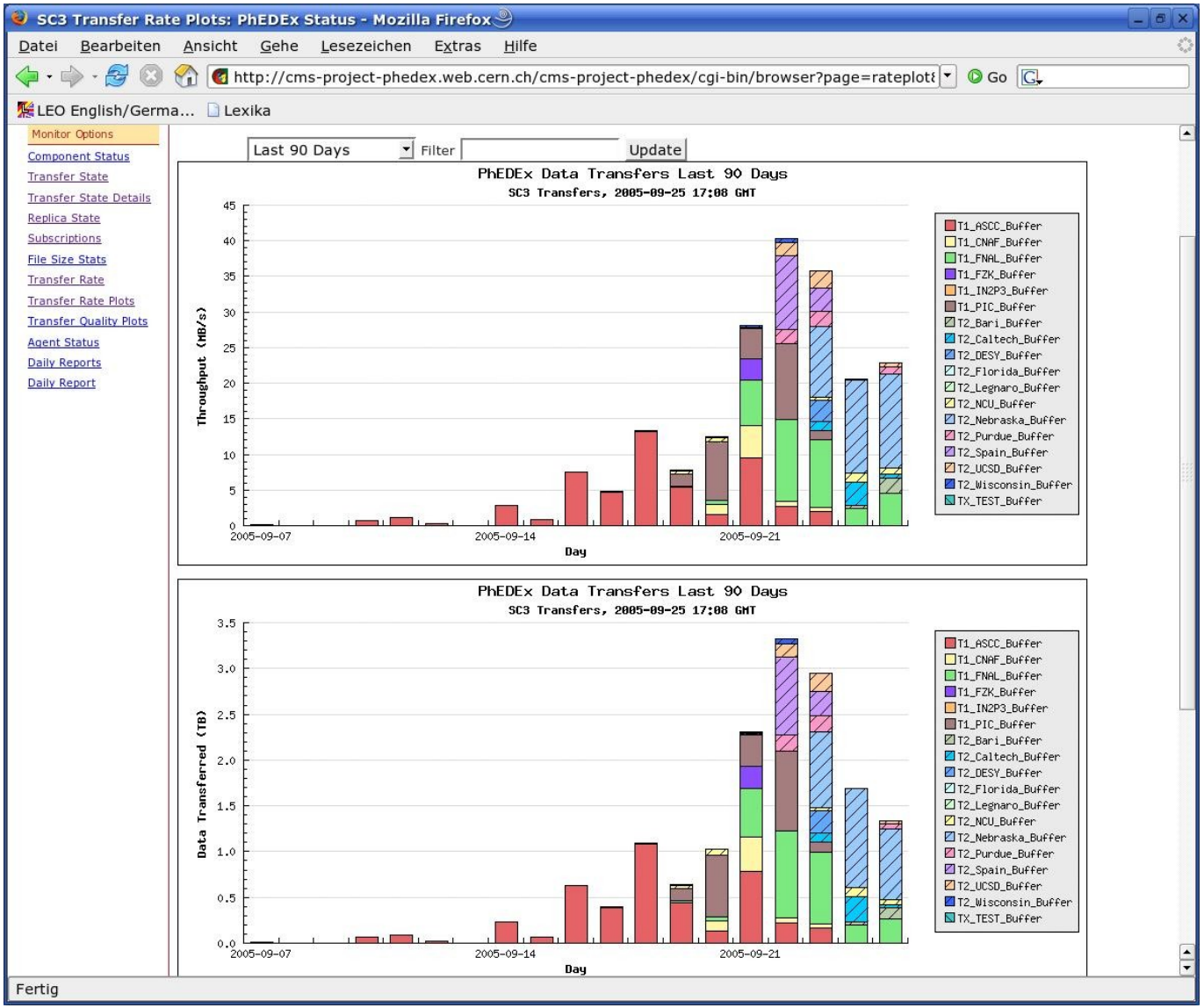
Owner	Dataset	Files N	Files Size	T1_FNAL_MSS	T2_Rio_Buffer
bt_2x1033PU761_TkMu_2_3_4_g133_CMS	bt03_qq_bbh200_2taujmu	4	6.3 GB	1.00 / 1.00	0.25 / 0.10
bt_DST8713_2x1033PU_g133_CMS	bt03_qq_bbh200_2taujmu	2	2.5 GB	1.00 / 1.00	0.00 / 0.00
bt_Hit750_g133	bt03_qq_bbh200_2taujmu	4	4.3 GB	1.00 / 1.00	0.00 / 0.00
_mtcc_test01	express	7	92.0 MB	0.00 / 0.00	
_mtcc_test02	express	15	21.8 GB	0.00 / 0.00	
mtcc_test01	express	4	70.0 MB	0.00 / 0.00	
test01	express	3	40.0 MB	0.00 / 0.00	
_mtcc_test01	normal	6	93.0 MB	0.00 / 0.00	
_mtcc_test02	normal	9	12.8 GB	0.00 / 0.00	
mtcc_test01	normal	8	110.0 MB	0.00 / 0.00	
test01	normal	2	30.0 MB	0.00 / 0.00	

PhEDEx – monitoring
Replica state

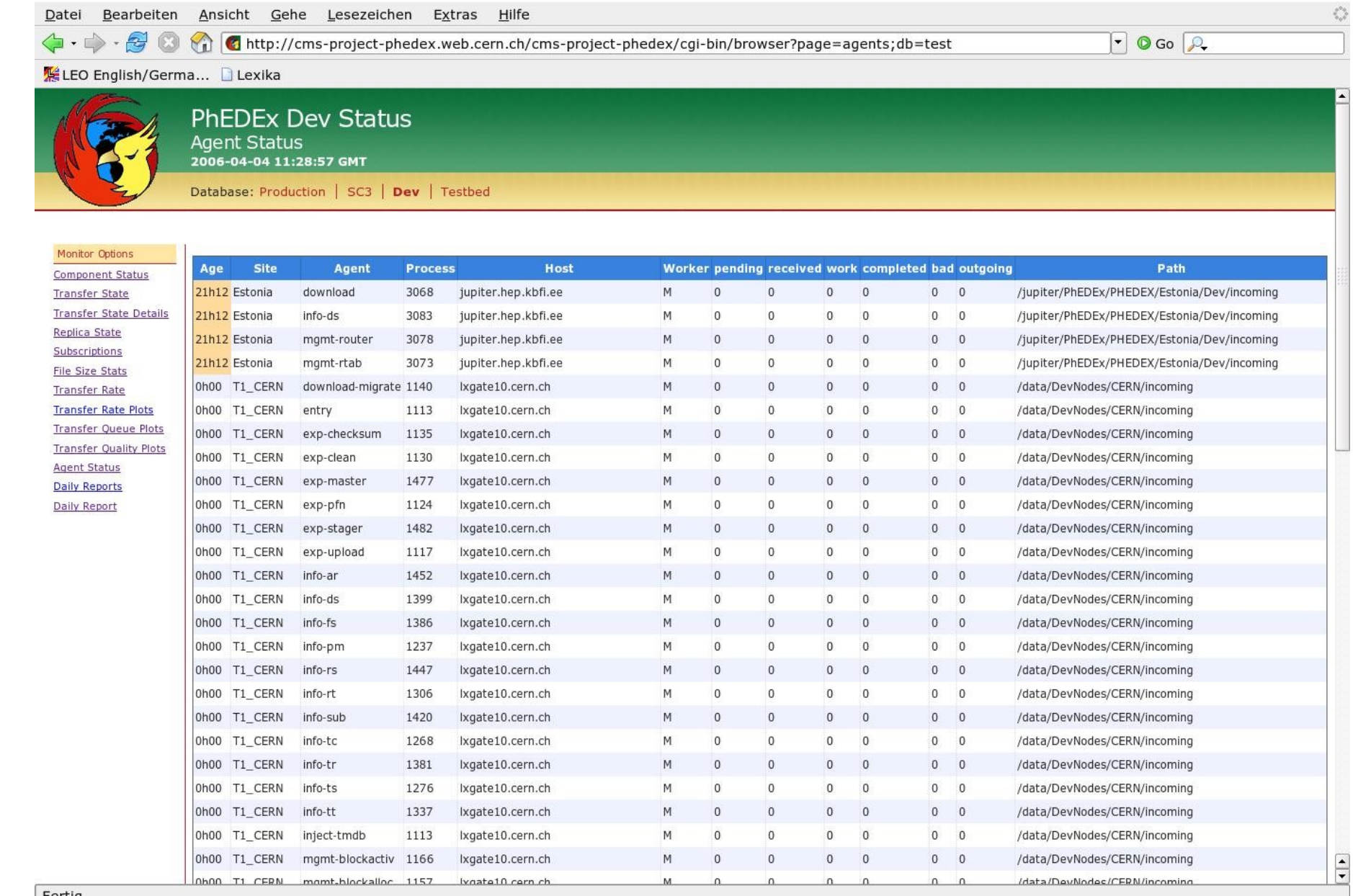


Owner	Dataset	Runs	Files		T1_FNAL_Buffer		T1_FNAL_MSS	
			N	Size	N	Size	N	Size
bt_2x1033PU761_TkMu_2_3_4_g133_CMS	bt03_gg_bbh200_2taujmu	4	4	6.3 GB	4	6.3 GB	4	6.3 GB
bt_DST8713_2x1033PU_g133_CMS	bt03_gg_bbh200_2taujmu	2	2	2.5 GB	2	2.5 GB	2	2.5 GB
bt_Hit750_g133	bt03_gg_bbh200_2taujmu	4	4	4.3 GB	2	2.1 GB	2	2.1 GB
eg_2x1033PU761_TkMu_2_g133_OSC	eg03_jets_2g_pt50170	1508	1529	2.9 TB	842	1.6 TB	842	1.6 TB
eg_DST8713_2x1033PU_g133_OSC	eg03_jets_2g_pt50170	607	623	1.2 TB	381	751.2 GB	381	751.2 GB
eg_L25s8713_2x1033PU_g133_OSC	eg03_jets_2g_pt50170	83	83	163.2 GB	42	82.2 GB	42	82.2 GB
hg_2x1033PU761_TkMu_2_g133_OSC	hg03_H2mu_ma125_tb30	5	5	9.0 GB	1	1.9 GB	1	1.9 GB
hg_2x1033PU761_TkMu_2_g133_OSC	hg03_H2mu_ma130_tb30	5	5	9.4 GB	3	5.7 GB	3	5.7 GB
hg_2x1033PU761_TkMu_2_g133_OSC	hg03_H2mu_ma135_tb30	5	5	9.2 GB	3	5.7 GB	3	5.7 GB
hg_2x1033PU761_TkMu_2_g133_OSC	hg03_H2mu_ma150_tb15	4	4	6.0 GB	1	478.8 MB	1	478.8 MB
hg_2x1033PU761_TkMu_2_g133_OSC	hg03_H2mu_ma150_tb40	4	4	5.7 GB	1	1.9 GB	1	1.9 GB
hg_2x1033PU761_TkMu_2_g133_OSC	hg03_H2mu_ma150_tb50	4	4	6.2 GB	2	2.5 GB	2	2.5 GB
hg_2x1033PU761_TkMu_2_g133_OSC	hg03_H2mu_ma200_tb15	4	4	6.2 GB	1	1.9 GB	1	1.9 GB
hg_2x1033PU761_TkMu_2_g133_OSC	hg03_H2mu_ma200_tb40	4	4	6.2 GB	2	2.5 GB	2	2.5 GB
hg_2x1033PU761_TkMu_2_g133_OSC	hg03_H2mu_ma300_tb15	4	4	6.2 GB	1	635.2 MB	1	635.2 MB
hg_2x1033PU761_TkMu_2_g133_OSC	hg03_H2mu_ma300_tb30	4	4	6.2 GB	1	1.8 GB	1	1.8 GB
hg_2x1033PU761_TkMu_2_g133_OSC	hg03_H2mu_ma300_tb40	4	4	6.2 GB	1	639.5 MB	1	639.5 MB
hg_2x1033PU761_TkMu_2_g133_OSC	hg03_H2mu_ma300_tb50	4	4	6.0 GB	1	476.5 MB	1	476.5 MB
hg_2x1033PU761_TkMu_2_g133_OSC	hg03_H2mu_ma400_tb15	4	4	6.0 GB	4	6.0 GB	4	6.0 GB
hg_2x1033PU761_TkMu_2_g133_OSC	hg03_H2mu_ma400_tb30	4	4	6.0 GB	1	473.9 MB	1	473.9 MB
hg_2x1033PU761_TkMu_2_g133_OSC	hg03_qq_qqh120_inv	4	4	6.0 GB	2	2.4 GB	2	2.4 GB
hg_2x1033PU761_TkMu_2_g133_OSC	hg03_qq_qqh135_2taull	14	14	25.2 GB	9	16.6 GB	9	16.6 GB
hg_2x1033PU761_TkMu_2_g133_OSC	hg03_qq_qqh200_inv	4	4	6.0 GB	4	6.0 GB	4	6.0 GB
hg_2x1033PU761_TkMu_g133_CMS	hg03_gg_ch_170_tb20	3	3	2.9 GB	3	2.9 GB	3	2.9 GB
hg_2x1033PU761_TkMu_g133_CMS	hg03_hzz_2e2mu_130a	4	4	6.3 GB	2	3.8 GB	2	3.8 GB
hg_2x1033PU761_TkMu_g133_CMS	hg03_hzz_4e_150	4	4	5.8 GB	1	1.9 GB	1	1.9 GB

PhEDEx – monitoring
Transfer rate

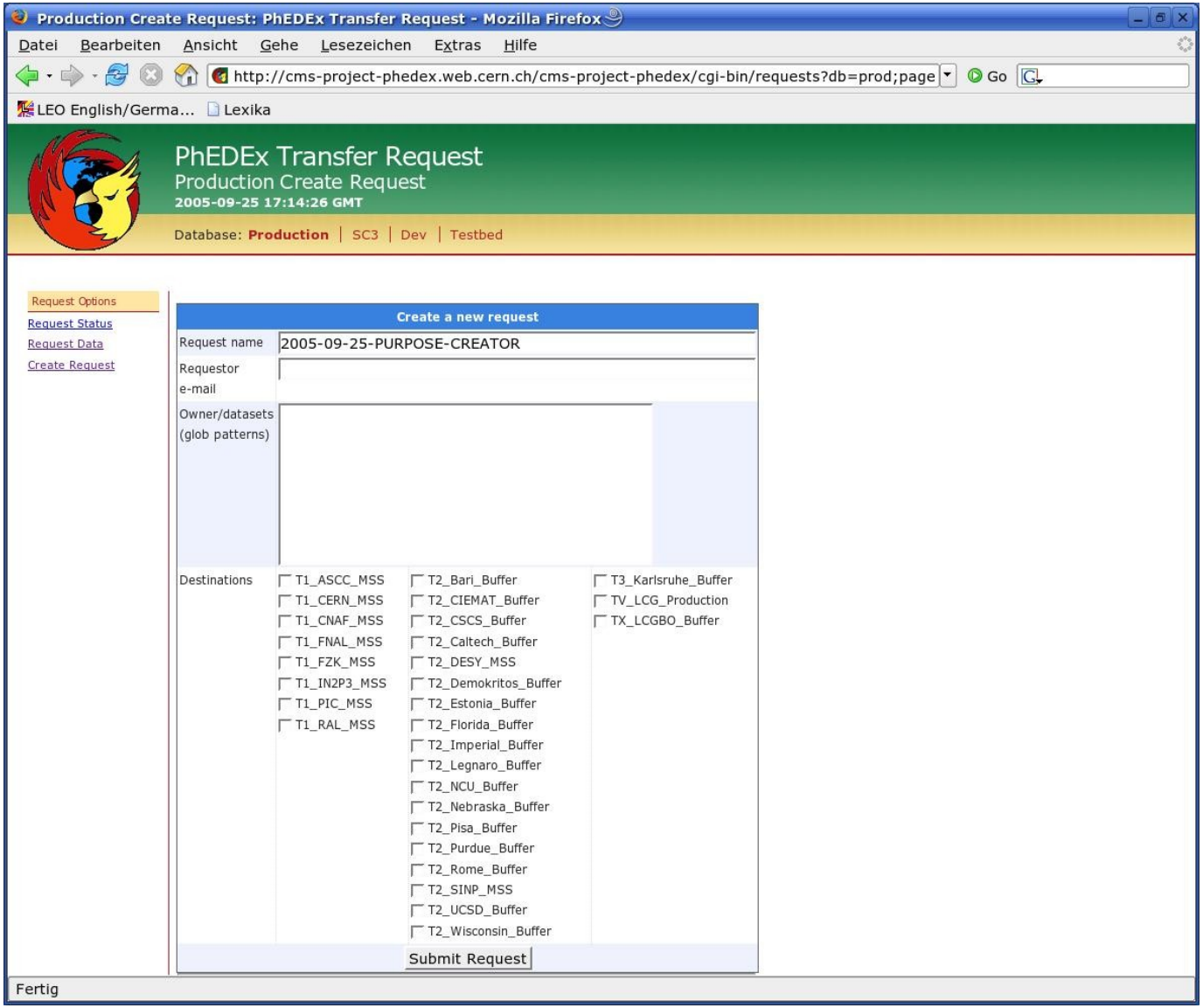
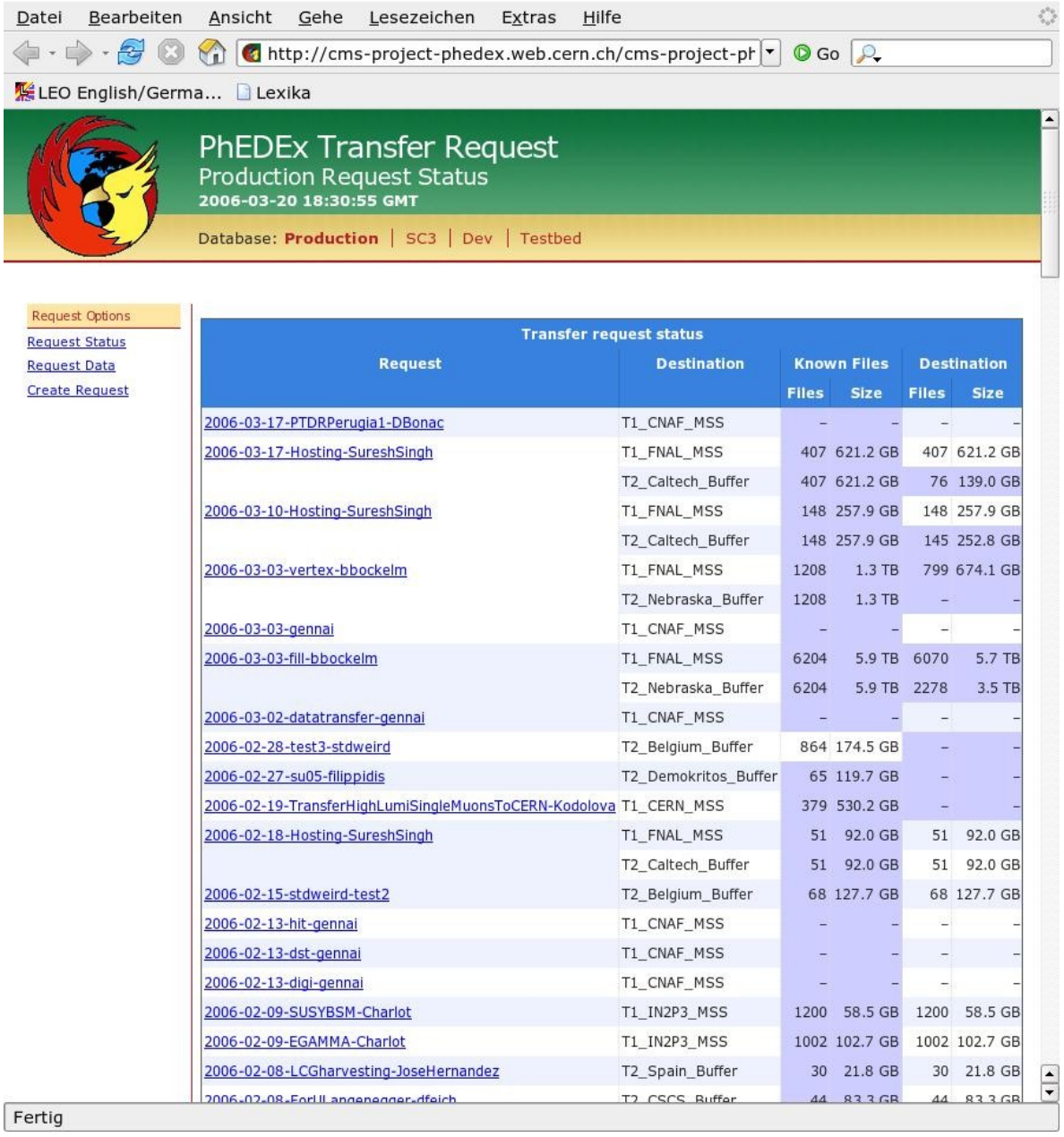


PhEDEx – monitoring
Agent status



Age	Site	Agent	Process	Host	Worker	pending	received	work	completed	bad	outgoing	Path
21h12	Estonia	download	3068	jupiter.hep.kbfi.ee	M	0	0	0	0	0	0	/jupiter/PhEDEx/PHEDEX/Estonia/Dev/incoming
21h12	Estonia	info-ds	3083	jupiter.hep.kbfi.ee	M	0	0	0	0	0	0	/jupiter/PhEDEx/PHEDEX/Estonia/Dev/incoming
21h12	Estonia	mgmt-router	3078	jupiter.hep.kbfi.ee	M	0	0	0	0	0	0	/jupiter/PhEDEx/PHEDEX/Estonia/Dev/incoming
21h12	Estonia	mgmt-rtab	3073	jupiter.hep.kbfi.ee	M	0	0	0	0	0	0	/jupiter/PhEDEx/PHEDEX/Estonia/Dev/incoming
0h00	T1_CERN	download-migrate	1140	lxgate10.cern.ch	M	0	0	0	0	0	0	/data/DevNodes/CERN/incoming
0h00	T1_CERN	entry	1113	lxgate10.cern.ch	M	0	0	0	0	0	0	/data/DevNodes/CERN/incoming
0h00	T1_CERN	exp-checksum	1135	lxgate10.cern.ch	M	0	0	0	0	0	0	/data/DevNodes/CERN/incoming
0h00	T1_CERN	exp-clean	1130	lxgate10.cern.ch	M	0	0	0	0	0	0	/data/DevNodes/CERN/incoming
0h00	T1_CERN	exp-master	1477	lxgate10.cern.ch	M	0	0	0	0	0	0	/data/DevNodes/CERN/incoming
0h00	T1_CERN	exp-pfn	1124	lxgate10.cern.ch	M	0	0	0	0	0	0	/data/DevNodes/CERN/incoming
0h00	T1_CERN	exp-stager	1482	lxgate10.cern.ch	M	0	0	0	0	0	0	/data/DevNodes/CERN/incoming
0h00	T1_CERN	exp-upload	1117	lxgate10.cern.ch	M	0	0	0	0	0	0	/data/DevNodes/CERN/incoming
0h00	T1_CERN	info-ar	1452	lxgate10.cern.ch	M	0	0	0	0	0	0	/data/DevNodes/CERN/incoming
0h00	T1_CERN	info-ds	1399	lxgate10.cern.ch	M	0	0	0	0	0	0	/data/DevNodes/CERN/incoming
0h00	T1_CERN	info-fs	1386	lxgate10.cern.ch	M	0	0	0	0	0	0	/data/DevNodes/CERN/incoming
0h00	T1_CERN	info-pm	1237	lxgate10.cern.ch	M	0	0	0	0	0	0	/data/DevNodes/CERN/incoming
0h00	T1_CERN	info-rs	1447	lxgate10.cern.ch	M	0	0	0	0	0	0	/data/DevNodes/CERN/incoming
0h00	T1_CERN	info-rt	1306	lxgate10.cern.ch	M	0	0	0	0	0	0	/data/DevNodes/CERN/incoming
0h00	T1_CERN	info-sub	1420	lxgate10.cern.ch	M	0	0	0	0	0	0	/data/DevNodes/CERN/incoming
0h00	T1_CERN	info-tc	1268	lxgate10.cern.ch	M	0	0	0	0	0	0	/data/DevNodes/CERN/incoming
0h00	T1_CERN	info-tr	1381	lxgate10.cern.ch	M	0	0	0	0	0	0	/data/DevNodes/CERN/incoming
0h00	T1_CERN	info-ts	1276	lxgate10.cern.ch	M	0	0	0	0	0	0	/data/DevNodes/CERN/incoming
0h00	T1_CERN	info-tt	1337	lxgate10.cern.ch	M	0	0	0	0	0	0	/data/DevNodes/CERN/incoming
0h00	T1_CERN	inject-tmdb	1113	lxgate10.cern.ch	M	0	0	0	0	0	0	/data/DevNodes/CERN/incoming
0h00	T1_CERN	mgmt-blockactiv	1166	lxgate10.cern.ch	M	0	0	0	0	0	0	/data/DevNodes/CERN/incoming
0h00	T1_CERN	mgmt-blockalloc	1157	lxgate10.cern.ch	M	0	0	0	0	0	0	/data/DevNodes/CERN/incoming

PhEDEx – subscription
Create request



PhEDEx – subscription
Monitor request