



CMS Transfer Team

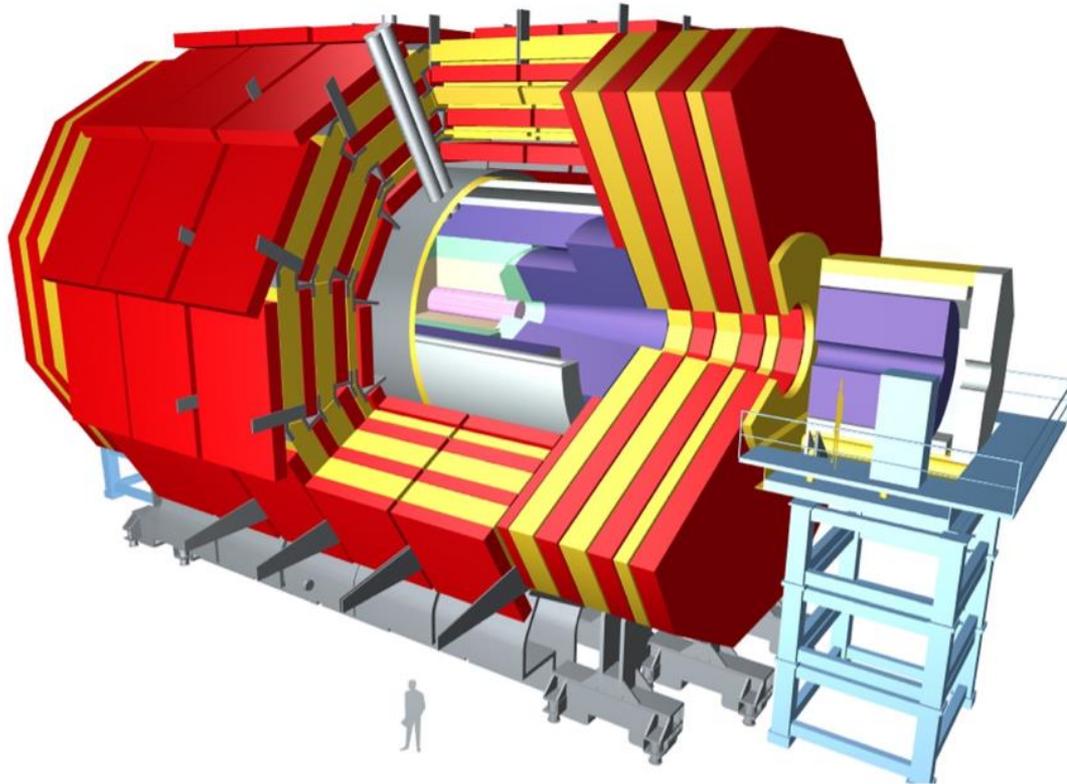
Jorge Alberto Diaz Cruz

High Energy Physics Seminar

April 27, 2016



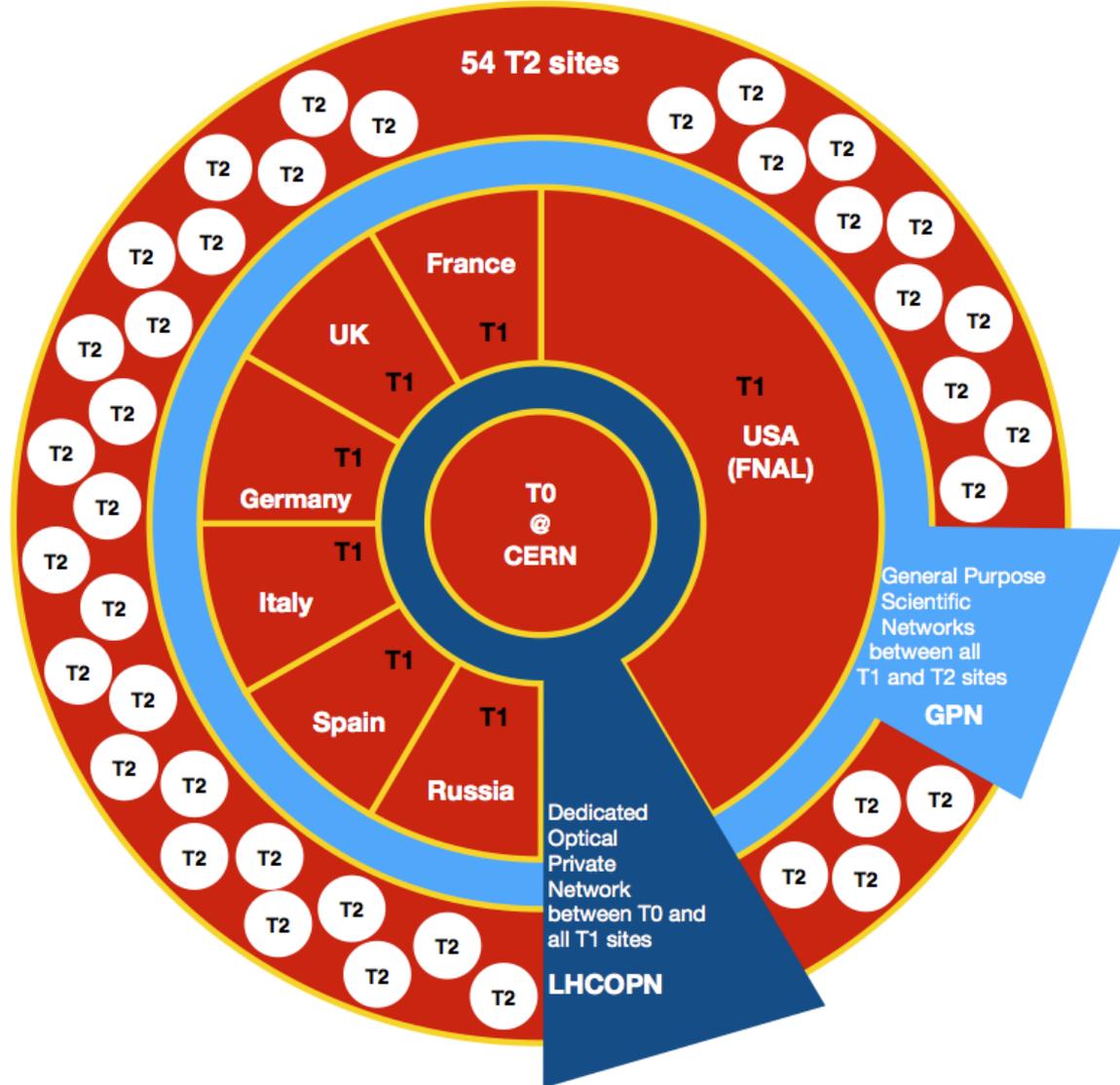
Compact Muon Solenoid - CMS



Total Weight = 14K tonnes
 Overall Diameter = 15m
 Overall Length = 28.7m
 Magnetic Field = 4T
 LHC Temperature = -271.3°C

- CERN: Conseil Européen pour la Recherche Nucléaire
- 40 Million Collisions per Second
- Data Taking Rate 500MB/s
- Run 1: Higgs Boson's spin (zero), parity (positive), and electric charge (neutral)
- Monte Carlo: Simulated 4.25 billion events in 2011 and 2.22 billion in 2012
- 2.5PB is transferred per week between all CMS sites.
- 120K cores processing events in parallel
- ~75PB disk + ~130PB tape
- Run 2: 6.5TeV per beam. [Brout-Englert-Higgs mechanism](#), [dark matter](#), [antimatter](#) and [quark-gluon plasma](#)

Worldwide LHC Computing Grid - WLCG



LHCOPN



ESnet

ENERGY SCIENCES NETWORK



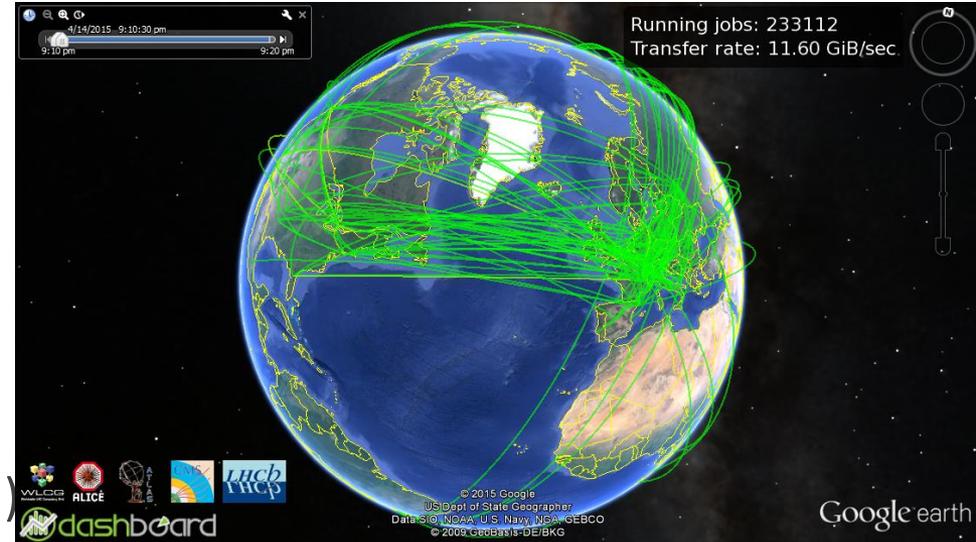
CMS Transfer Team

- Maintains large scale data movements between the different centers. ~2.5PB per week
- Work together with the admins of the over 60 computing sites to maintain functional data transfer services and resources
- Monitor the transfer system, maintain its health and debug transfer problems

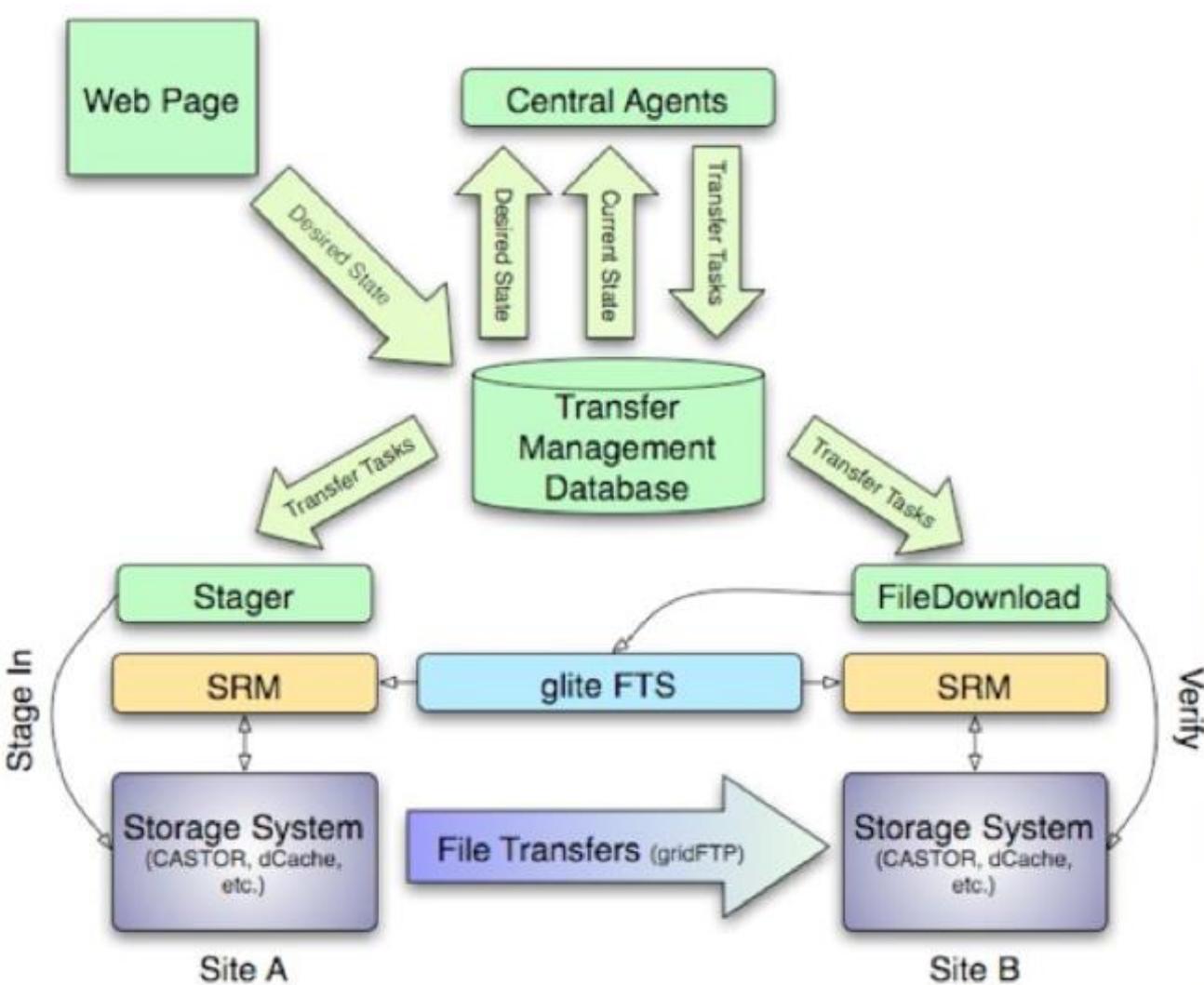


PhEDEx – Physics Experiment Data Export

- Take care of large scale data transfers across the Grid
- TMDB: Transfer Management DataBase. High availability Oracle database cluster hosted at CERN
- Agents: Software daemon processes connect directly to TMDB. Current state to desired state.
- Use the File Transfer Service (FTS)
- Ensure reliable transfers by verifying each file after transfer
- Real life monitoring through a web status display
- Daily in use since 2004
- PhEDEx data service: web service to access information (JSON, XML and perl)

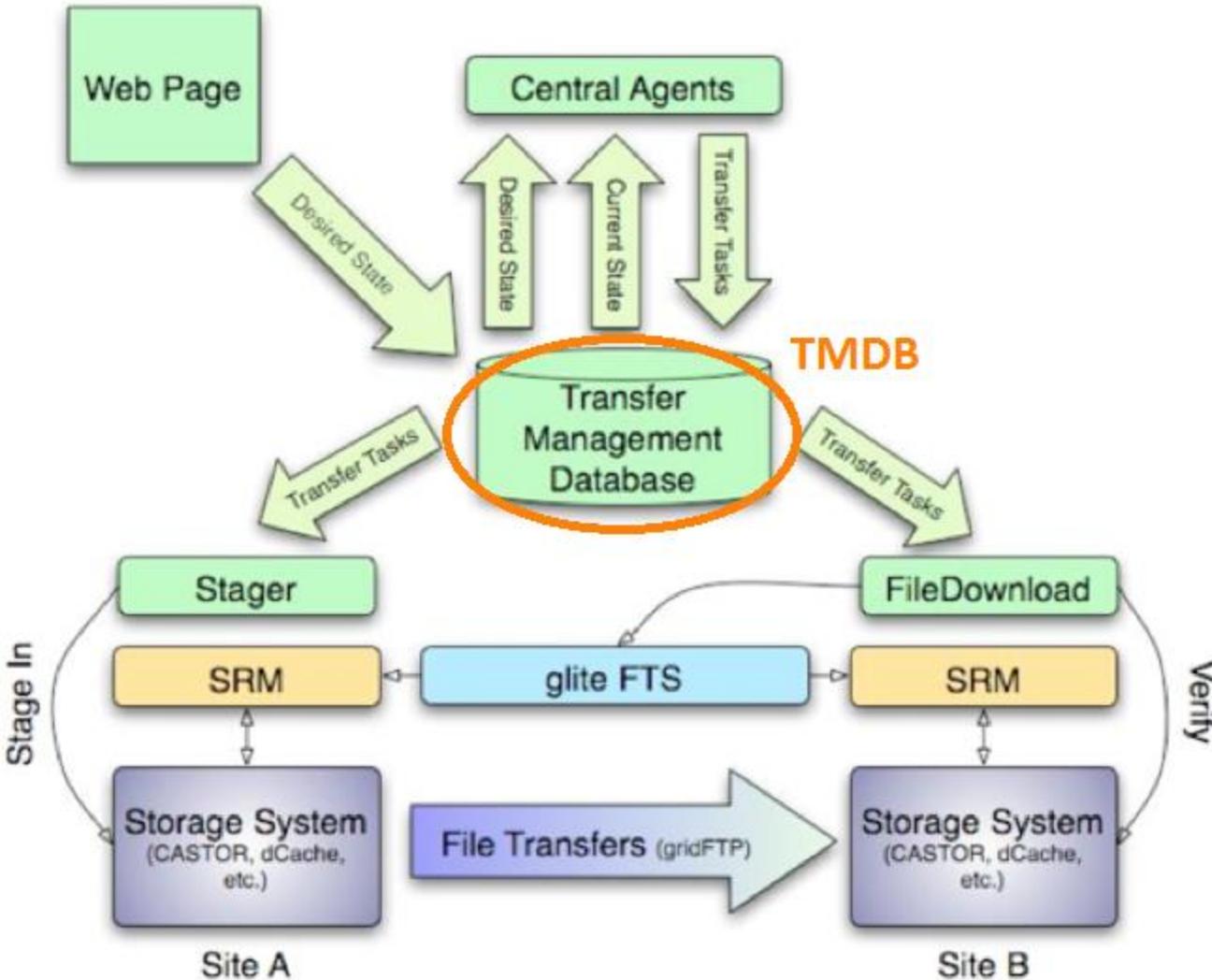


PhEDEx – Physics Experiment Data Export



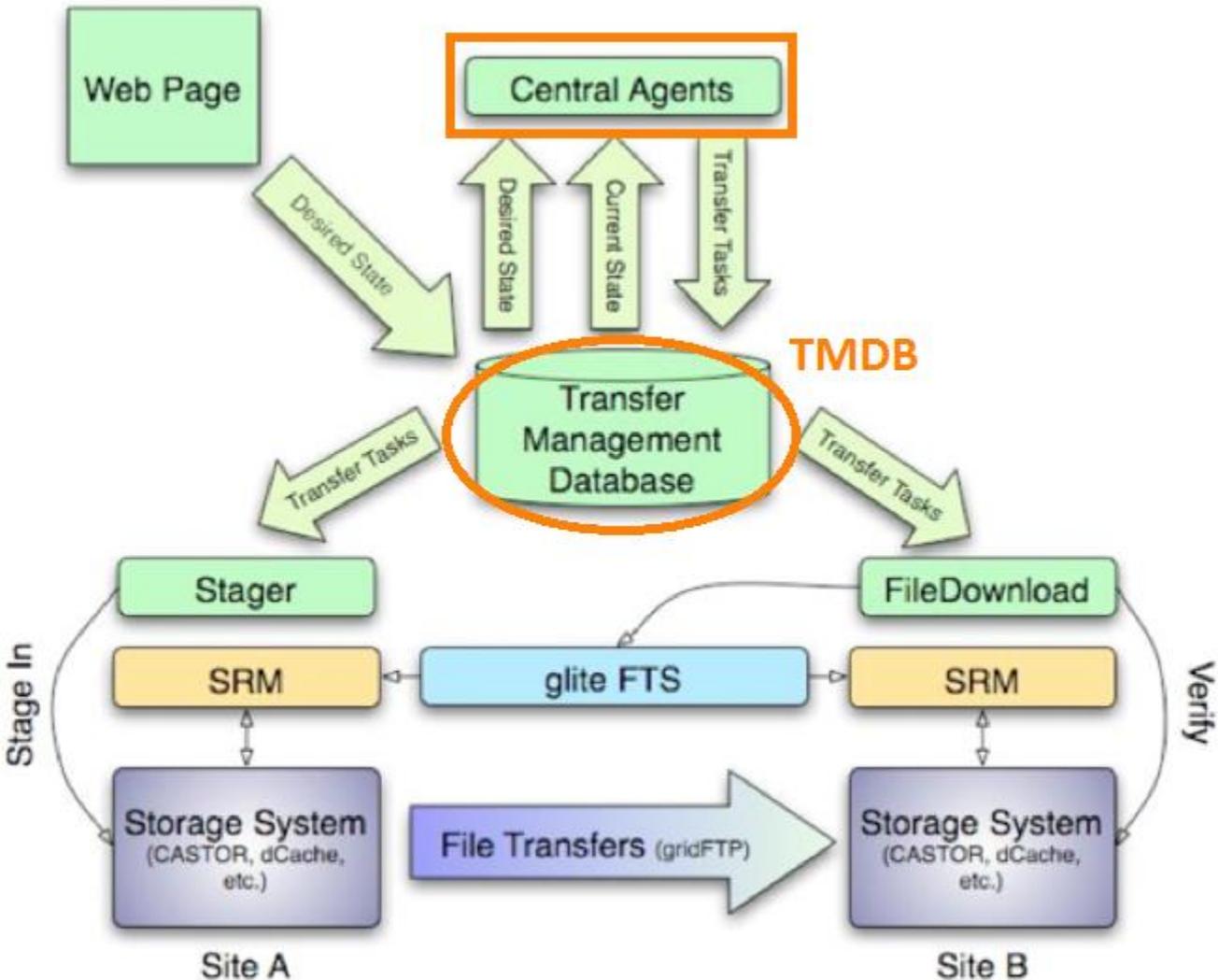
PhEDEx – Physics Experiment Data Export

- TMDB: Blackboard for the system state

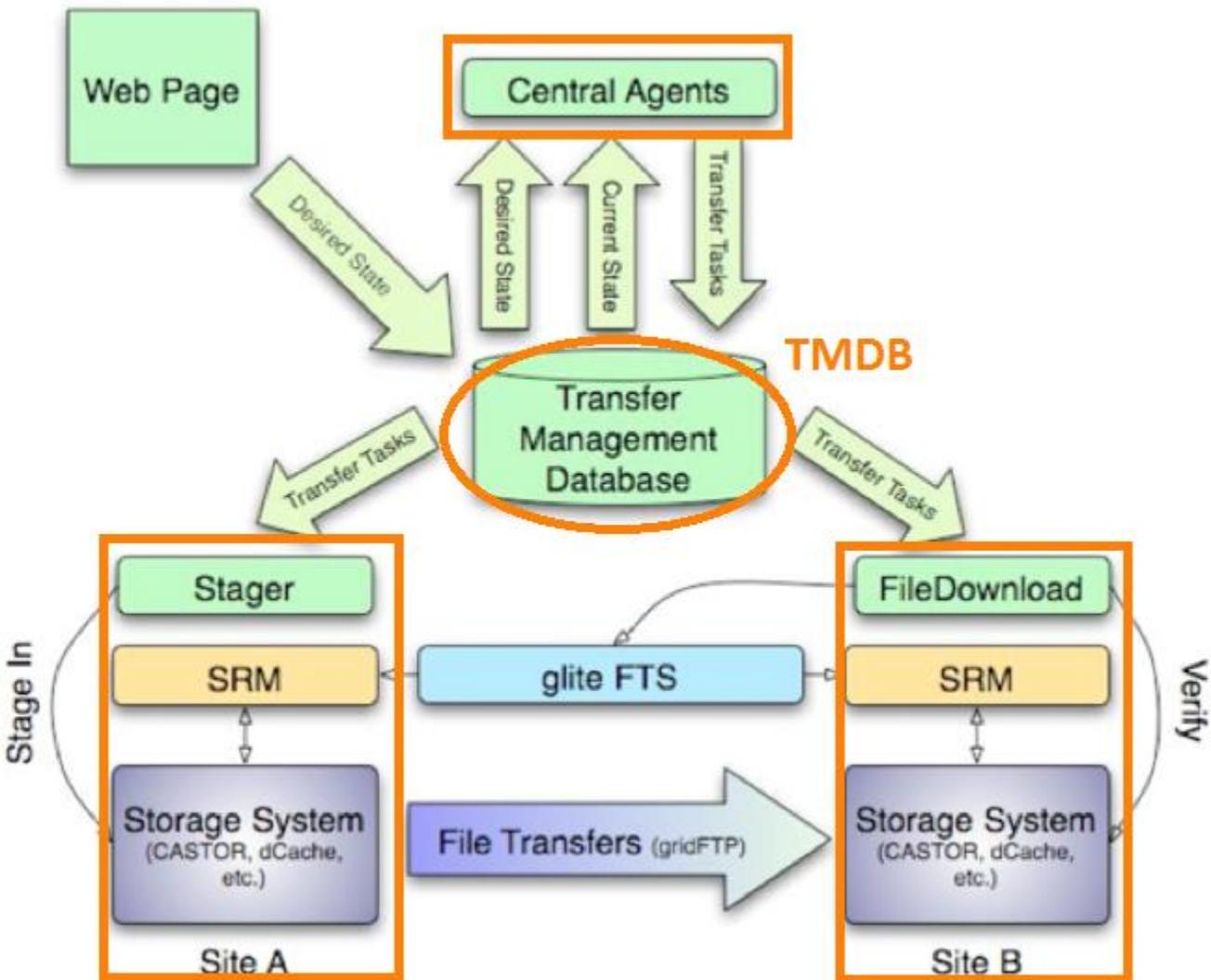


PhEDEx – Physics Experiment Data Export

- TMDB: Blackboard for the system state
- CAgents at CERN. FileRouter Agent. Allocator agent.

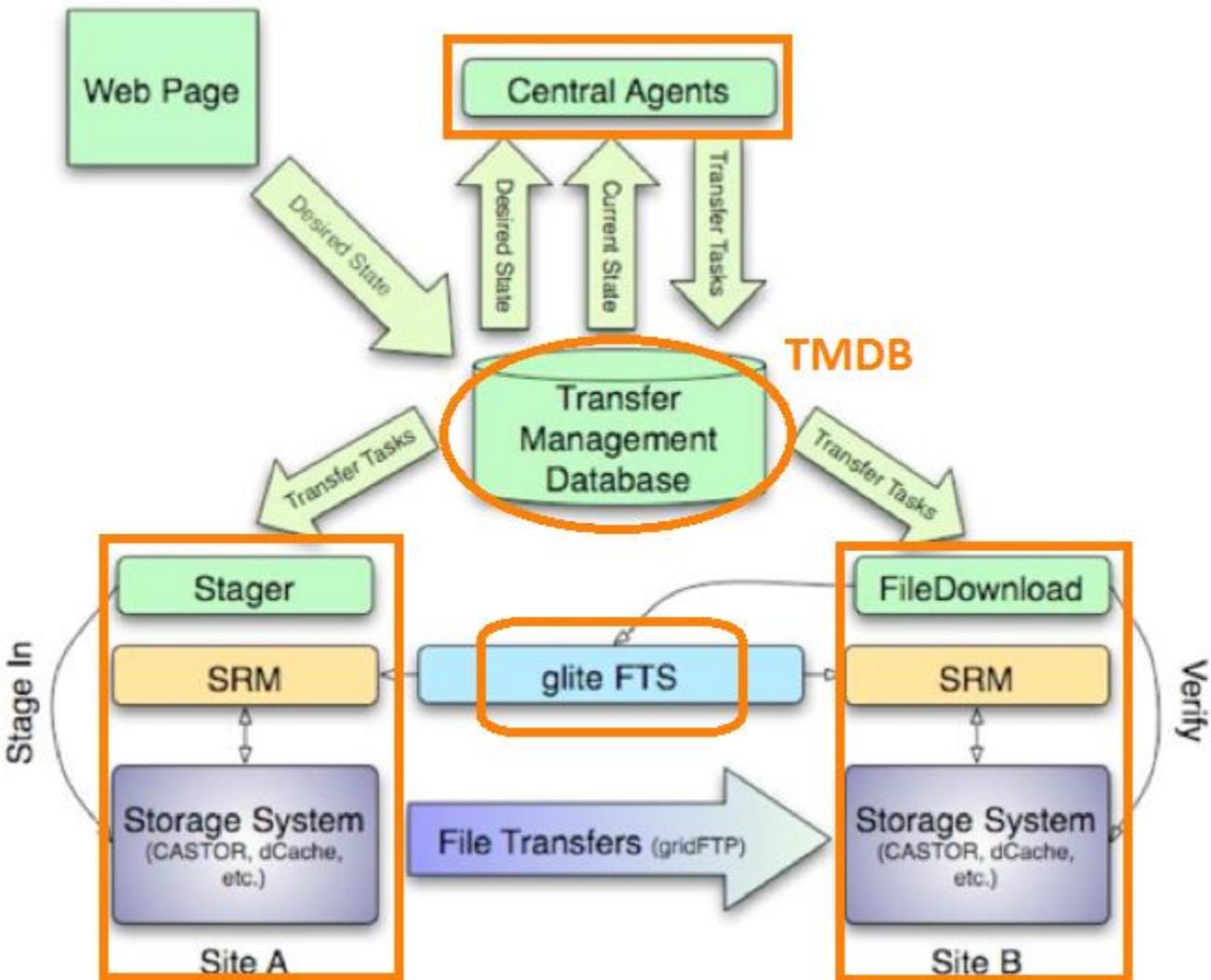


PhEDEx – Physics Experiment Data Export



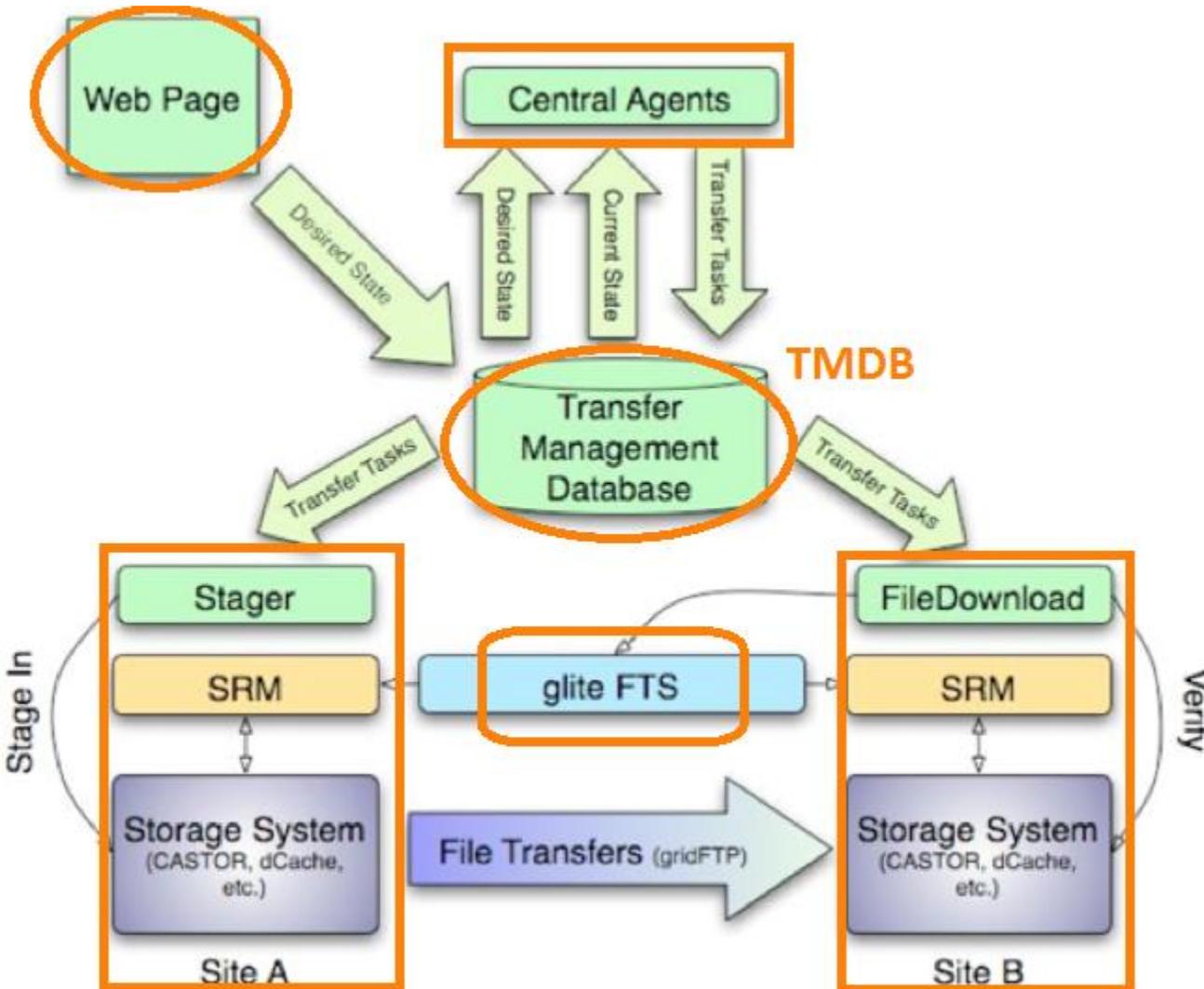
- TMDb: Blackboard for the system state
- Central Agents at CERN
 - FileRouter Agent
 - Allocator Agent
- Local Agents at Sites
 - FileExport Agent
 - FileDownload Agent
 - FileRemove Agent

PhEDEx – Physics Experiment Data Export



- TMDb: Blackboard for the system state
- Central Agents at CERN
 - FileRouter Agent
 - Allocator Agent
- Local Agents at Sites
 - FileExport Agent
 - FileDownload Agent
 - FileRemove Agent
 - BlockDownloadVerify
 - FileStger Agent
- FTS: File Transfer System

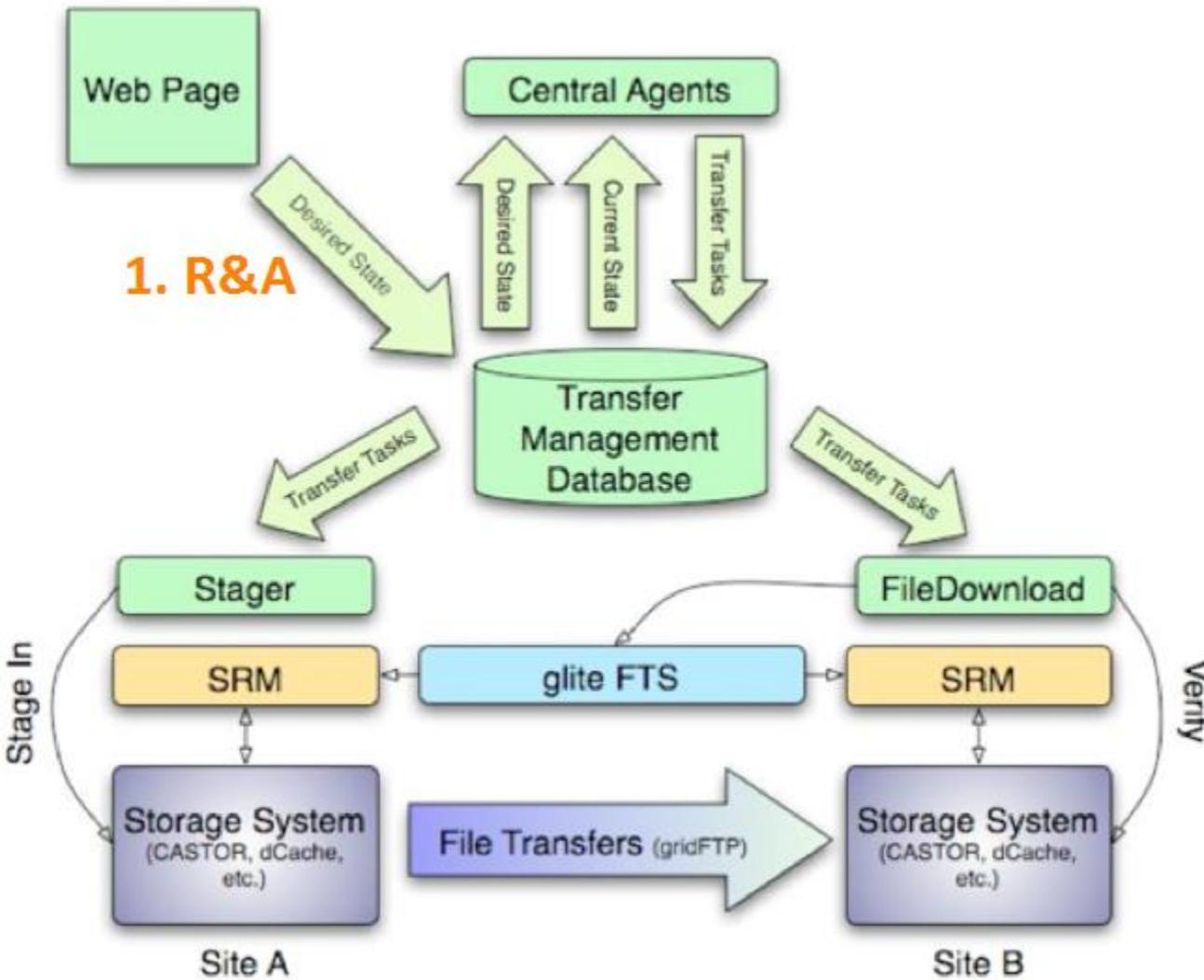
PhEDEx – Physics Experiment Data Export



- TMDb: Blackboard for the system state
- Central Agents at CERN
 - FileRouter Agent
 - Allocator Agent
- Local Agents at Sites
 - FileExport Agent
 - FileDownload Agent
 - FileRemove Agent
 - BlockDownloadVerify
 - FileStger Agent
- FTS: File Transfer System
- Web Page: User interaction. Transfer and Agent Monitoring

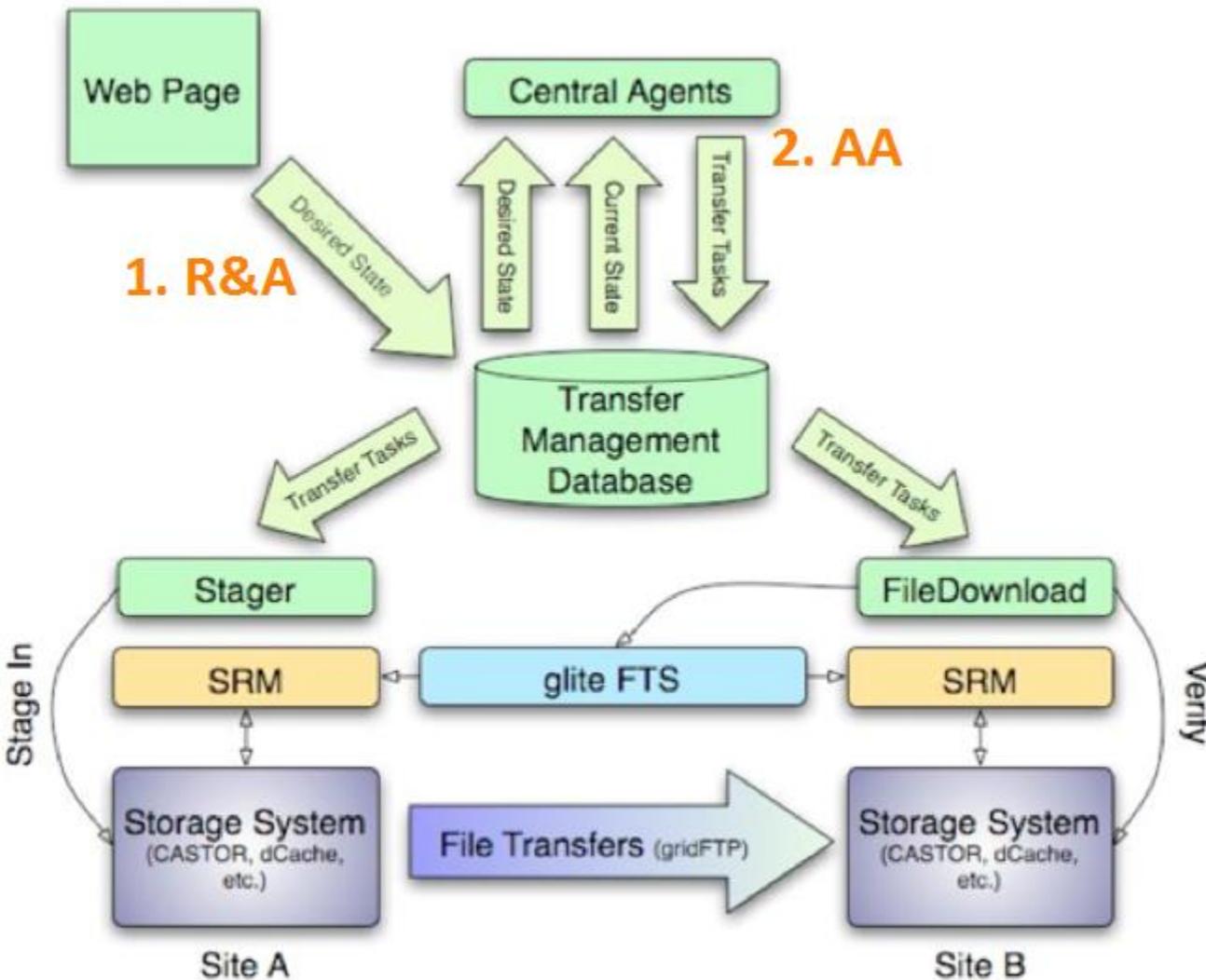
PhEDEx – Physics Experiment Data Export

1. Request & Approval



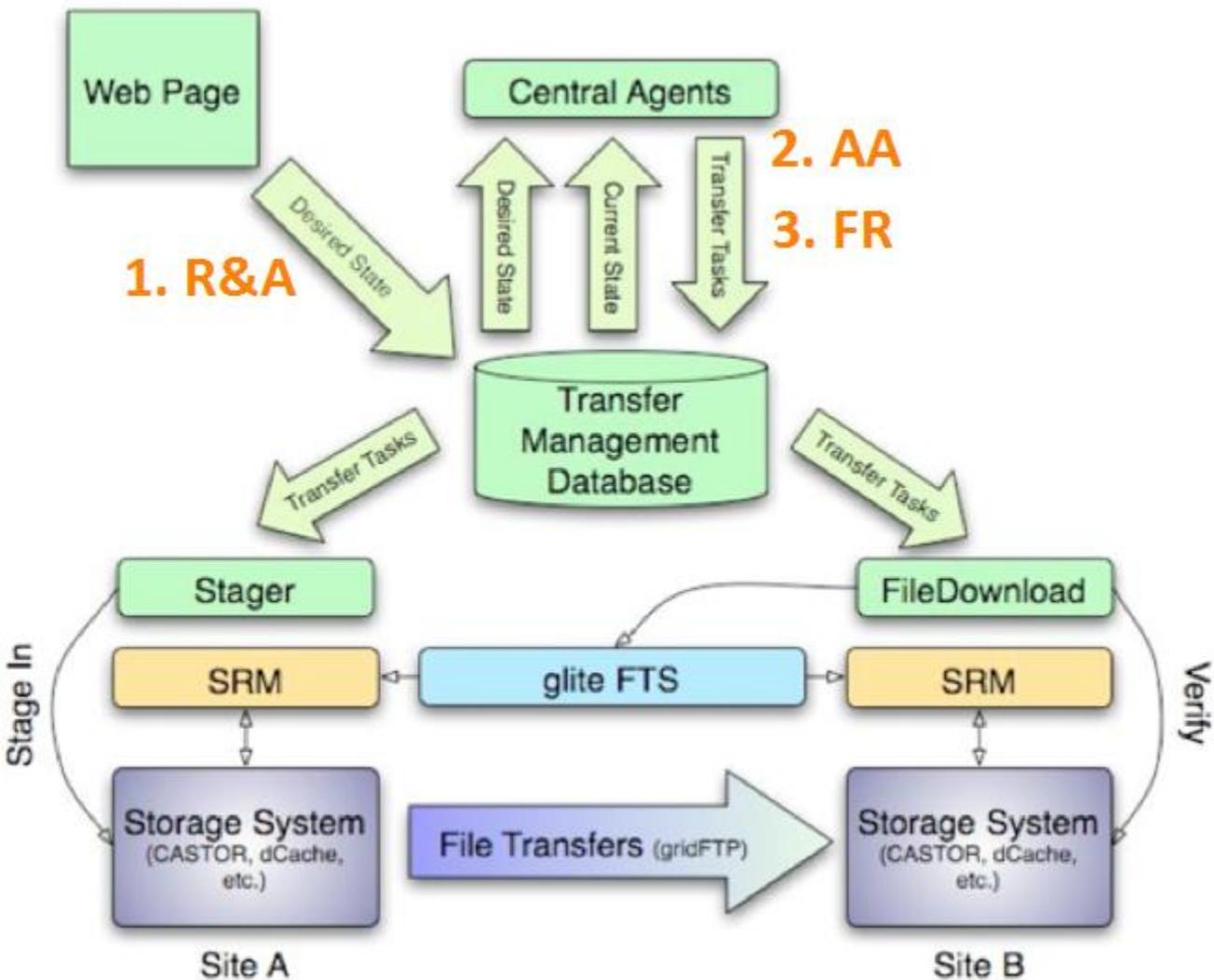
PhEDEx – Physics Experiment Data Export

1. Request & Approval
2. Allocator agent allocates files to destinations

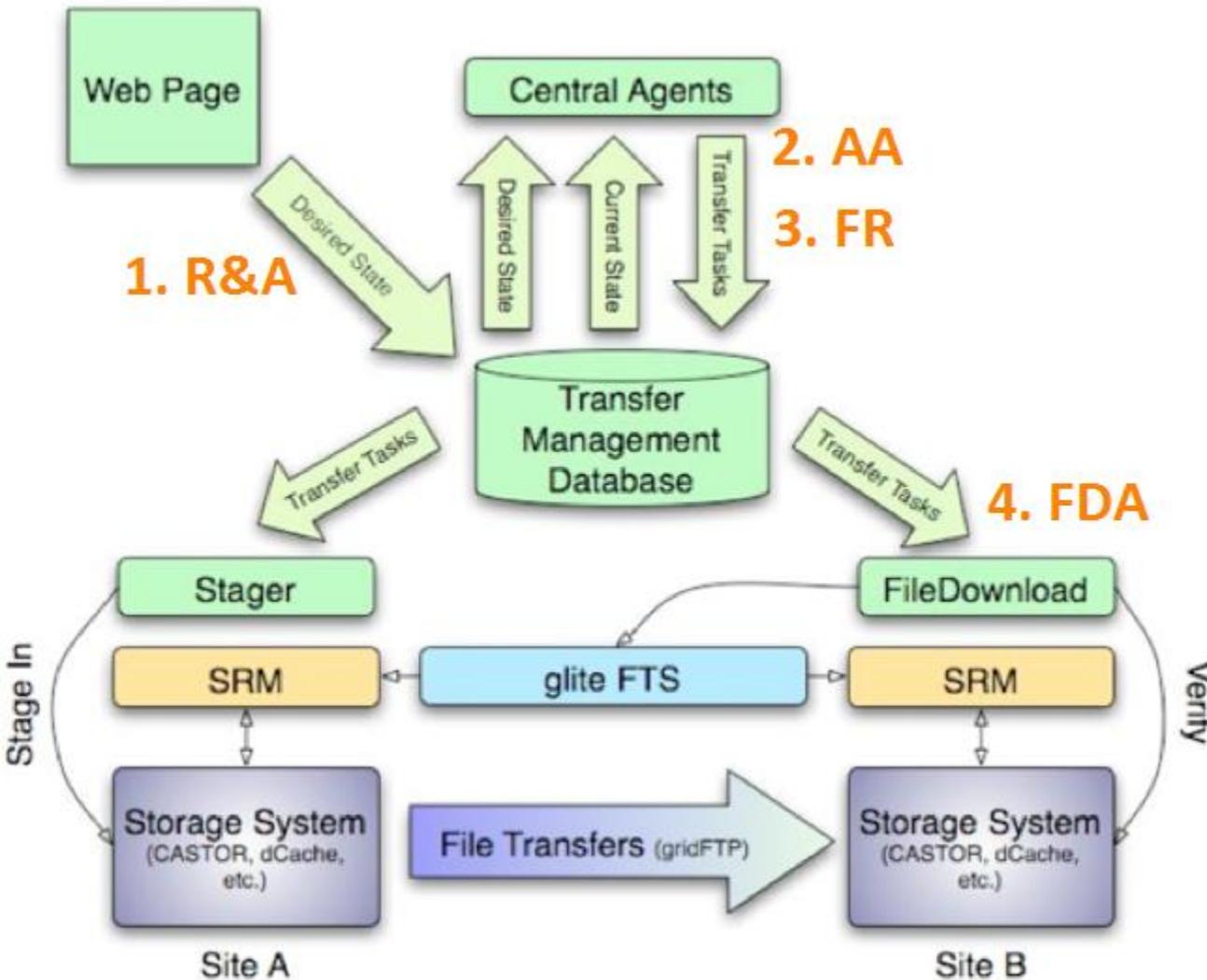


PhEDEx – Physics Experiment Data Export

1. Request & Approval
2. Allocator agent allocates files to destinations
3. FileRouter Agent determines replica

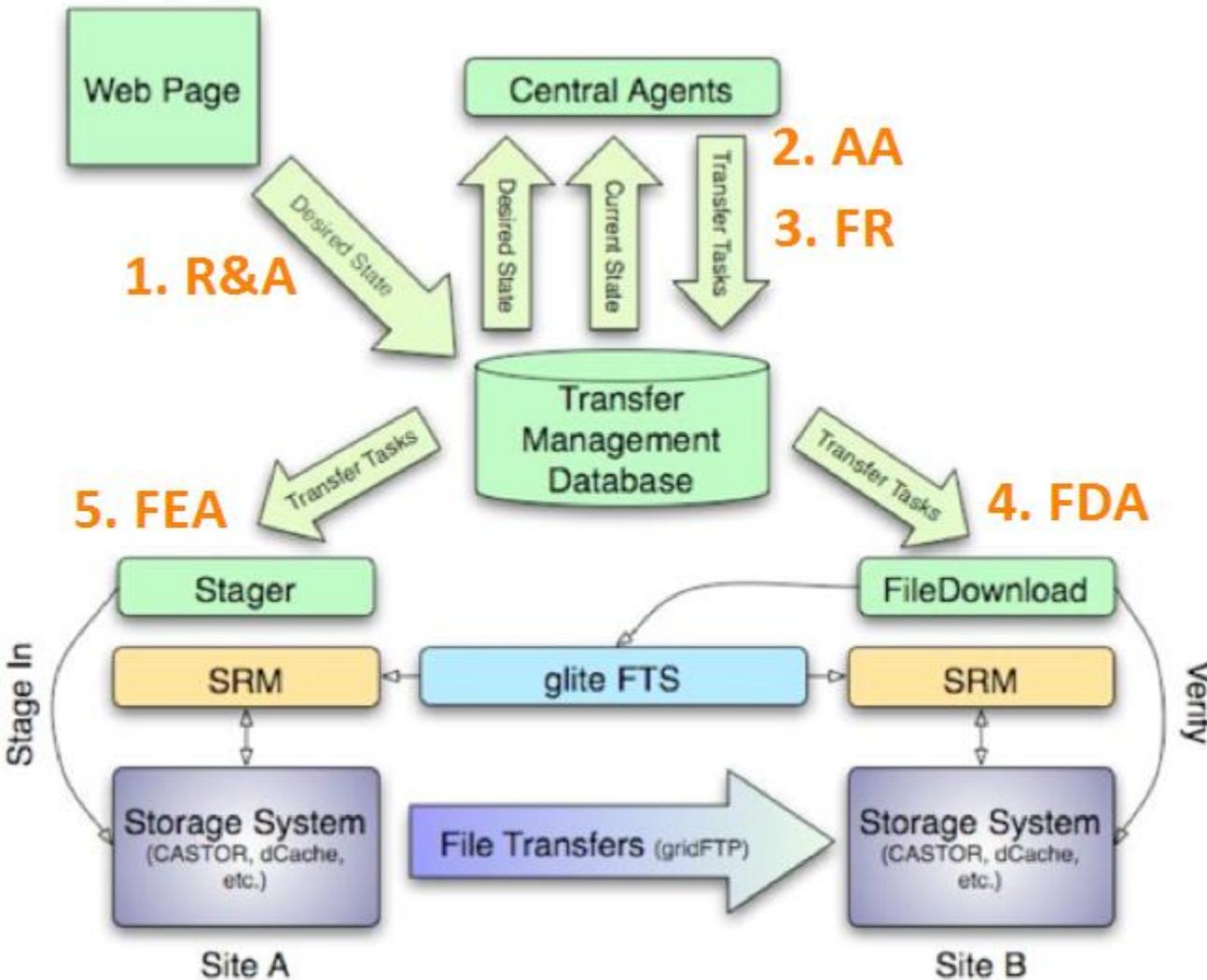


PhEDEx – Physics Experiment Data Export



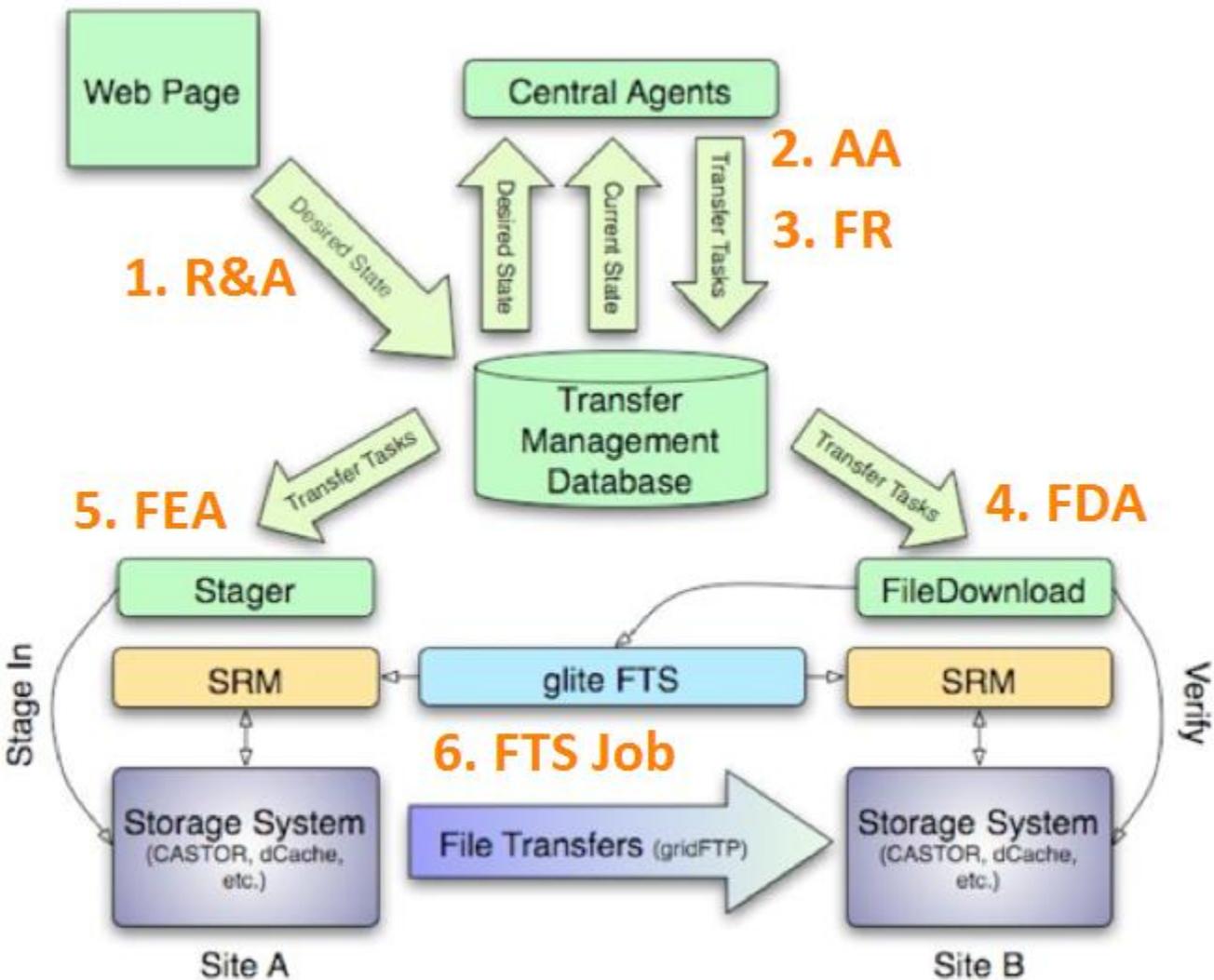
1. Request & Approval
2. Allocator agent allocates files to destinations
3. FileRouter Agent determines replica
4. FileDownload Agent marks file as “wanted”

PhEDEx – Physics Experiment Data Export



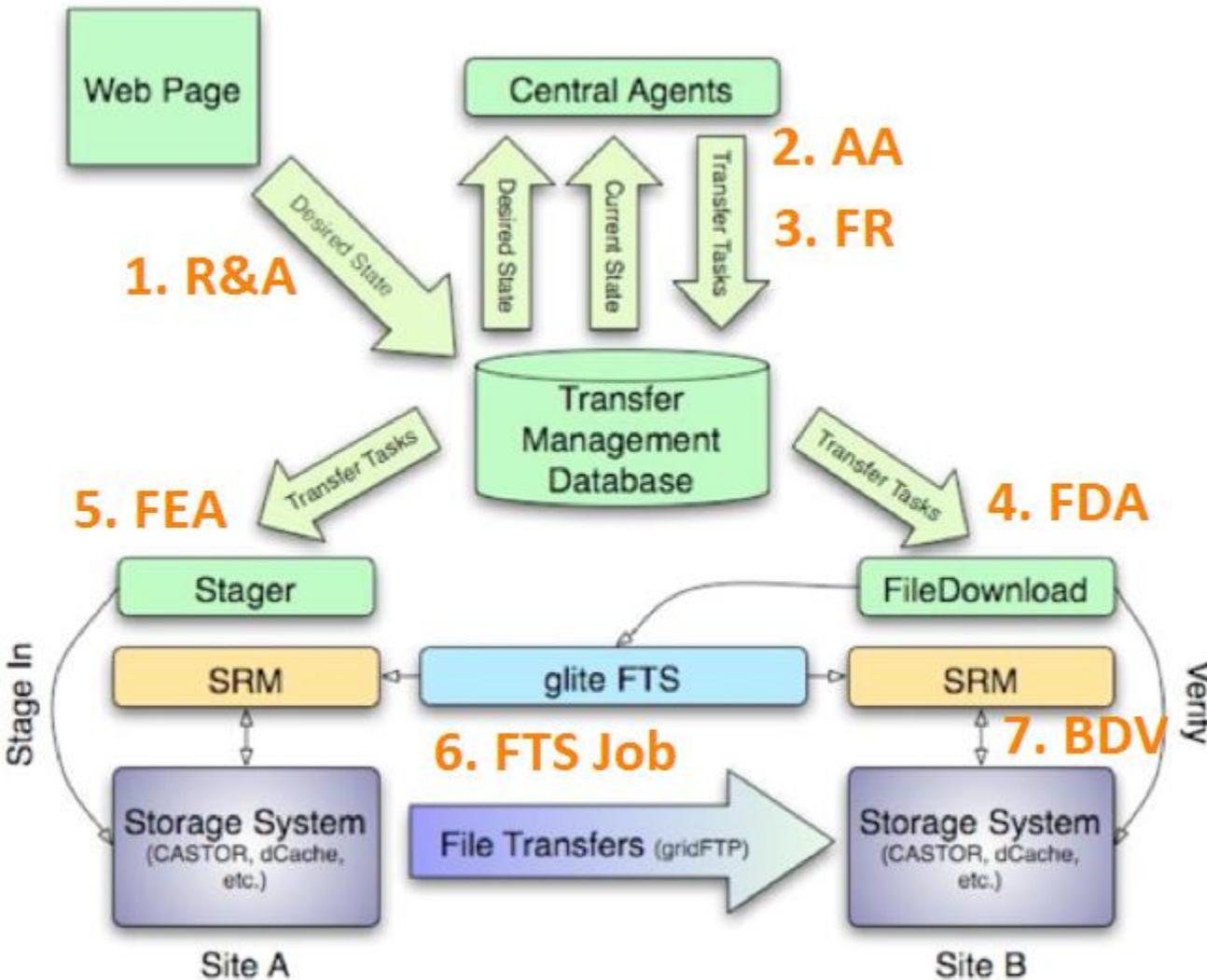
1. Request & Approval
2. Allocator agent allocates files to destinations
3. FileRouter Agent determines replica
4. FileDownload Agent marks file as “wanted”
5. FileExport Agent Initiate staging and marks file as available

PhEDEx – Physics Experiment Data Export



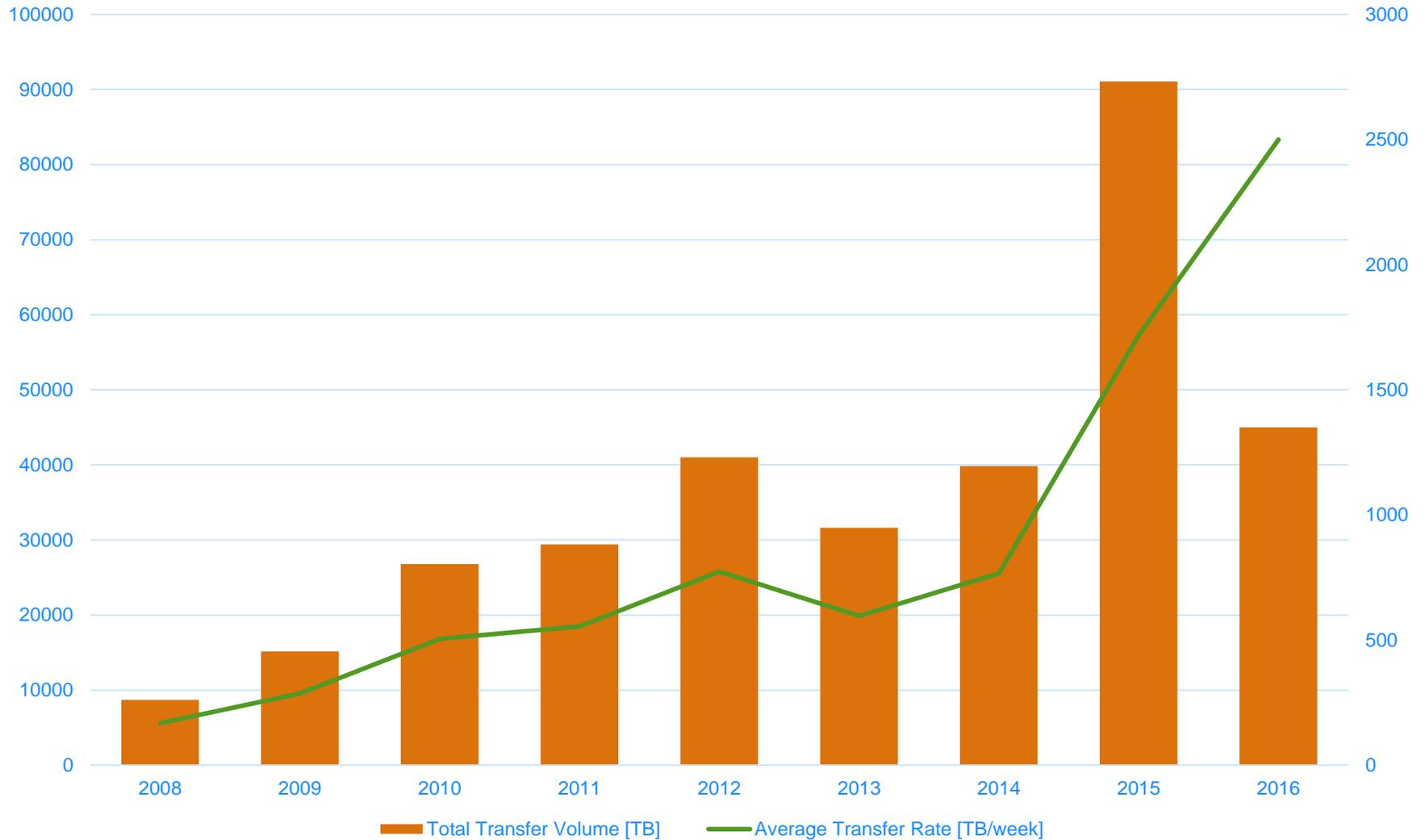
1. Request & Approval
2. Allocator agent allocates files to destinations
3. FileRouter Agent determines replica
4. FileDownload Agent marks file as “wanted”
5. FileExport Agent Initiate staging and marks file as available
6. FileDownload Agent generates a FTS job to transfer the file

PhEDEx – Physics Experiment Data Export



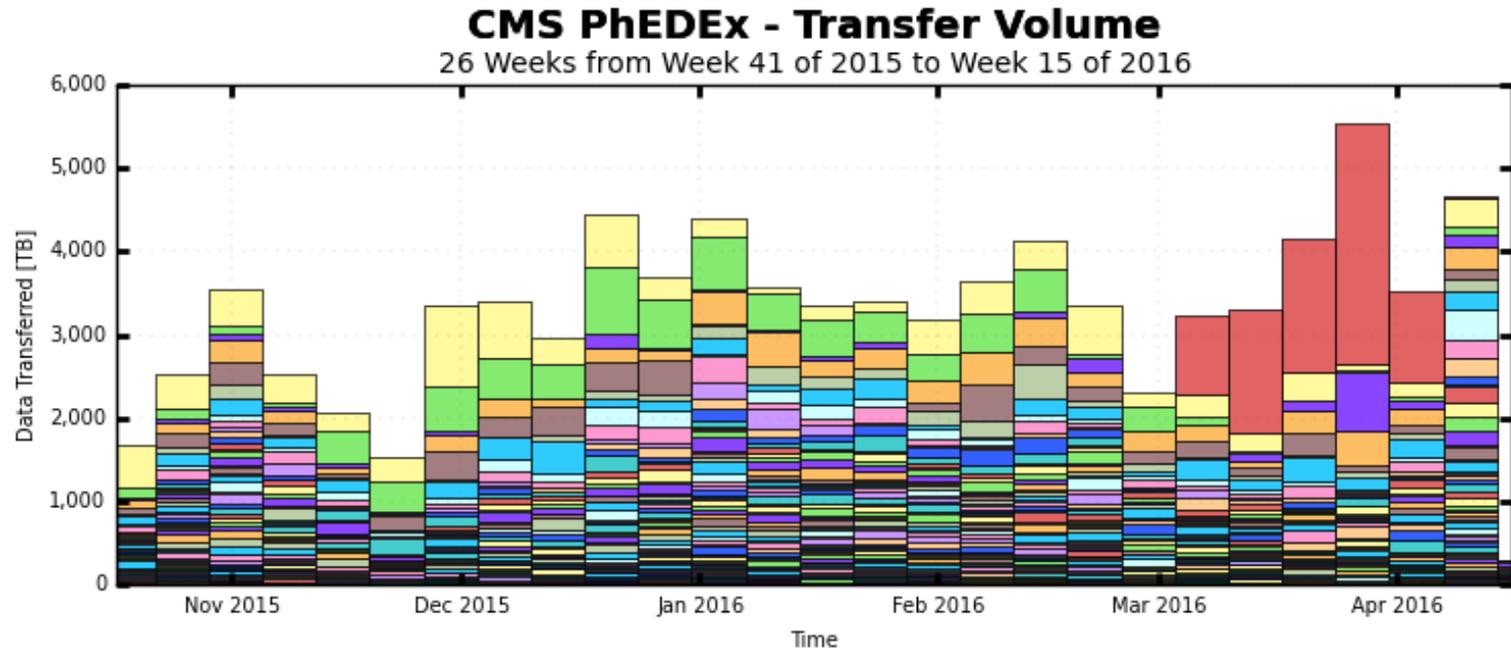
1. Request & Approval
2. Allocator agent allocates files to destinations
3. FileRouter Agent determines replica
4. FileDownload Agent marks file as “wanted”
5. FileExport Agent Initiate staging and marks file as available
6. FileDownload Agent generates a FTS job to transfer the file
7. BlockDownloadVerify verifies size&cksum

CMS Data Transfers



PhEDEx Monitoring

- PhEDEx Web Page interface

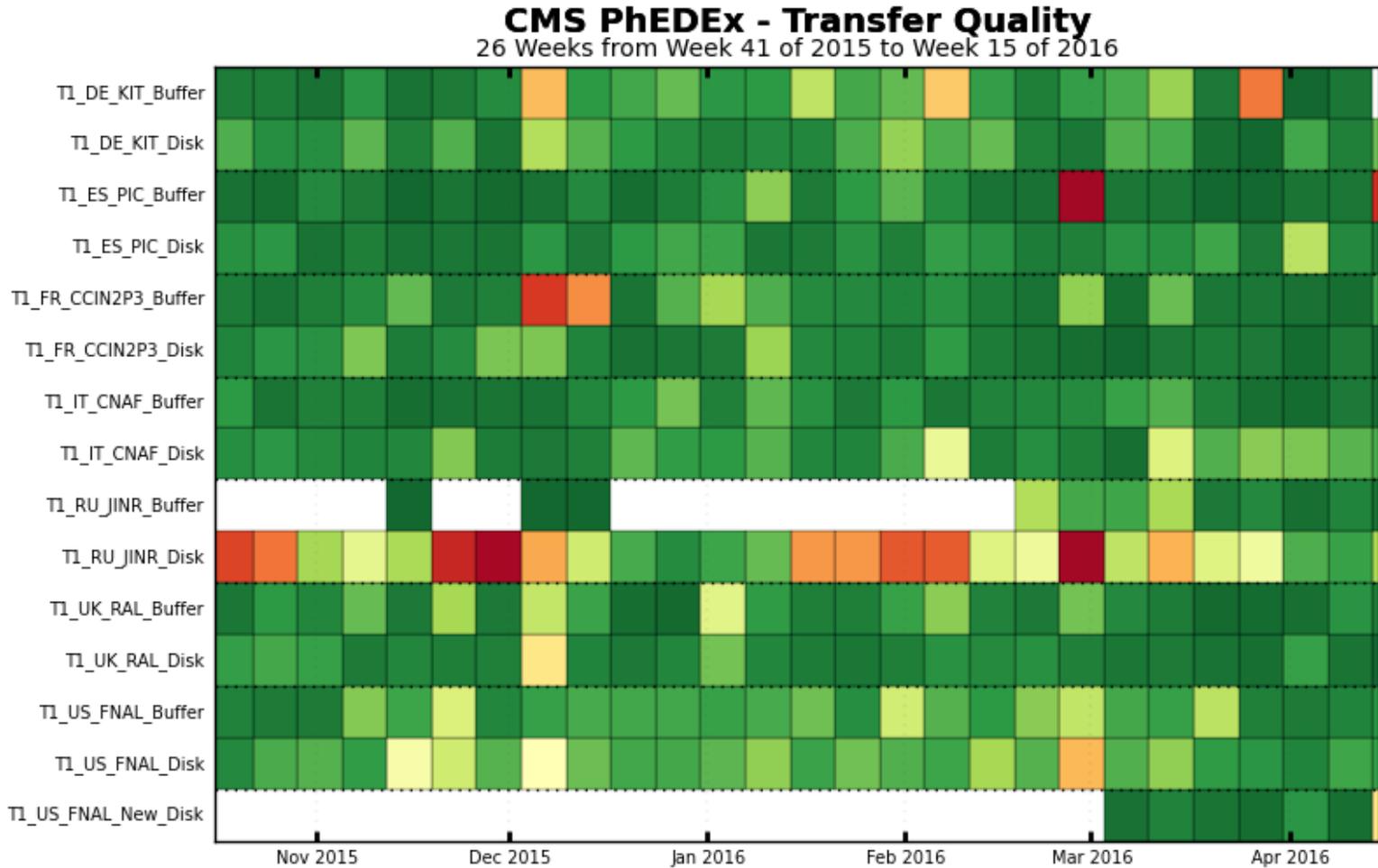


- | | | | | |
|---|---|---|---|--|
| ■ T1_US_FNAL_New_Disk | ■ T1_US_FNAL_Disk | ■ T1_US_FNAL_Buffer | ■ T2_US_Purdue | ■ T2_CH_CERN |
| ■ T0_CH_CERN_Export | ■ T2_US_Nebraska | ■ T1_IT_CNAF_Disk | ■ T1_UK_RAL_Disk | ■ T1_DE_KIT_Disk |
| ■ T2_US_Vanderbilt | ■ T1_RU_JINR_Disk | ■ T2_US_MIT | ■ T1_IT_CNAF_Buffer | ■ T2_US_Florida |
| ■ T2_IT_Legnaro | ■ T2_US_Caltech | ■ T1_FR_CCIN2P3_Disk | ■ T1_UK_RAL_Buffer | ■ T2_UK_London_IC |
| ■ T2_EE_Estonia | ■ T2_US_Wisconsin | ■ T1_FR_CCIN2P3_Buffer | ■ T2_BE_UCL | ■ T1_DE_KIT_Buffer |
| ■ T1_RU_JINR_Buffer | ■ T2_DE_DESY | ■ T1_ES_PIC_Disk | ■ T2_ES_CIEMAT | ■ T2_IT_Pisa |
| ■ T2_BR_SPRACE | ■ T2_IT_Bari | ■ T2_BE_IHÉ | ■ T2_KR_KNU | ■ T2_UK_SGrid_RALPP |
| ■ T2_US_UCSD | ■ T2_FR_IPHC | ■ T2_IN_TIFR | ■ T2_UK_London_Brunel | ■ T2_BR_UERJ |
| ■ T2_CH_CSCS | ■ T2_FR_GRIF_IRFU | ■ T2_FI_HIP | ■ T2_DE_RWTH | ■ T2_ES_IFCA |
| ■ T2_FR_GRIF_LL | ■ T2_RU_JINR | ■ T0_CH_CERN_Disk | ■ T2_HU_Budapest | ... plus 38 more |

Maximum: 5,543 TB, Minimum: 284.96 TB, Average: 3,252 TB, Current: 284.96 TB

PhEDEx Monitoring

- PhEDEx Web Page interface



PhEDEx Monitoring

- PhEDEx Web Page interface

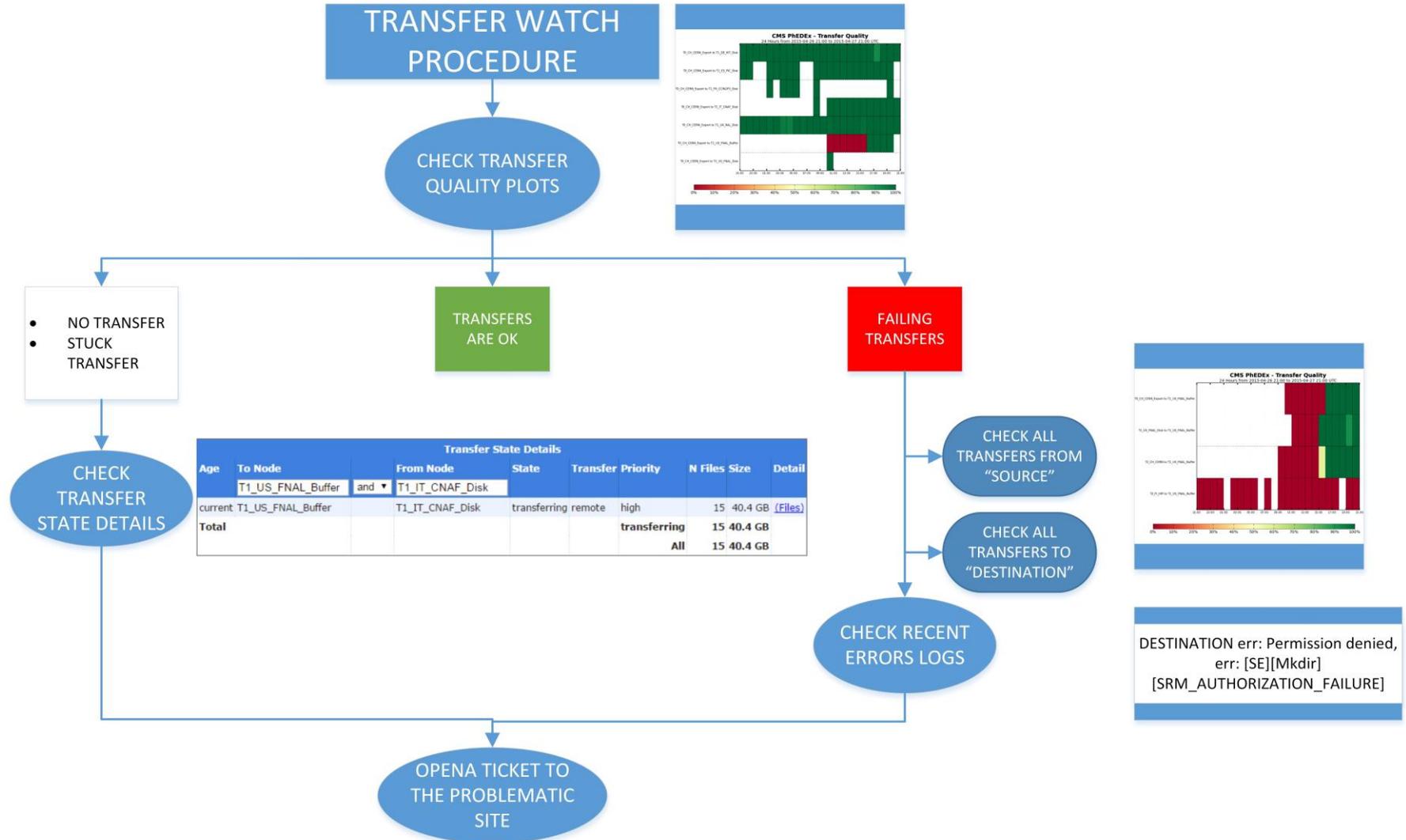
Infrastructure Agents			
Node	FileRouter	FileIssue	FilePump
PhEDEx Central	UP	UP	UP

Workflow Agents						
Node	RequestAllocator	BlockAllocator	BlockMonitor	BlockDelete	BlockActivate	BlockDeactivate
PhEDEx Central	UP	UP	UP	UP	UP	UP

Support Agents					
Node	BlockDownloadVerifyInjector	InfoFileSize	InfoStatesClean	InvariantMonitor	PerfMonitor
PhEDEx Central	UP	UP	UP	UP	UP

Site Agents						
Node	FileDownload	FileExport	FileStager	FileRemove	BlockDownloadVerify	Watchdog
T0_CH_CERN_Disk	UP (2/2 agents)	UP (3/3 agents)		UP	UP	UP
T0_CH_CERN_Export	UP (2/2 agents)	UP (2/2 agents)	UP	UP	UP	UP (2/2 agents)

PhEDEx Monitoring Example



Storage: Tape vs Disk



Tape Robots:

- Very cheap to store large amounts of data
- Slow access
- Robot Arm inserts a cartridge into a tape drive for Reading or writing

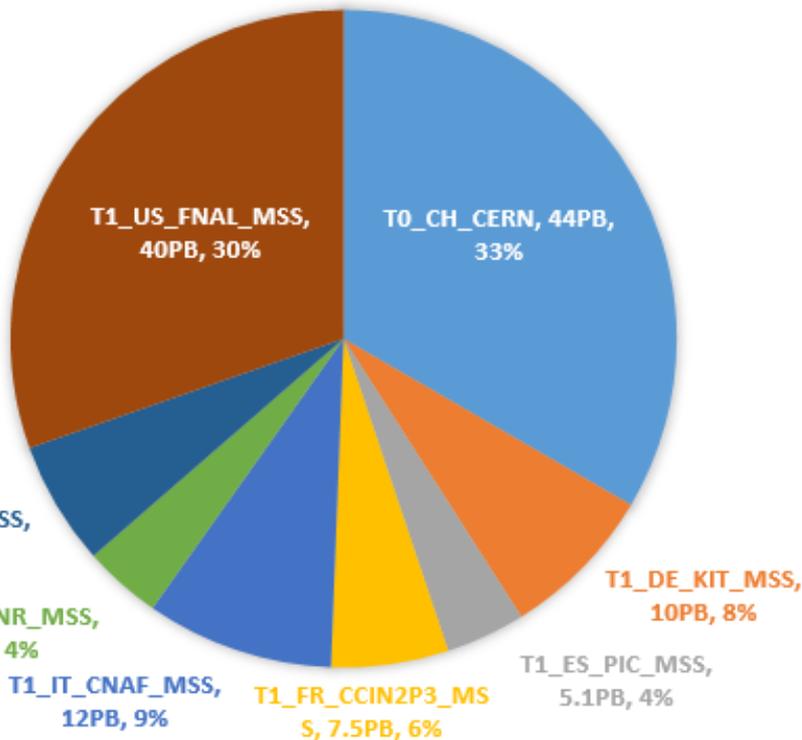
Large Disk Arrays:

- Normal Hard Drives
- Fast Access
- Store Data Temporarily
- Software Systems -> Big Hard Drive

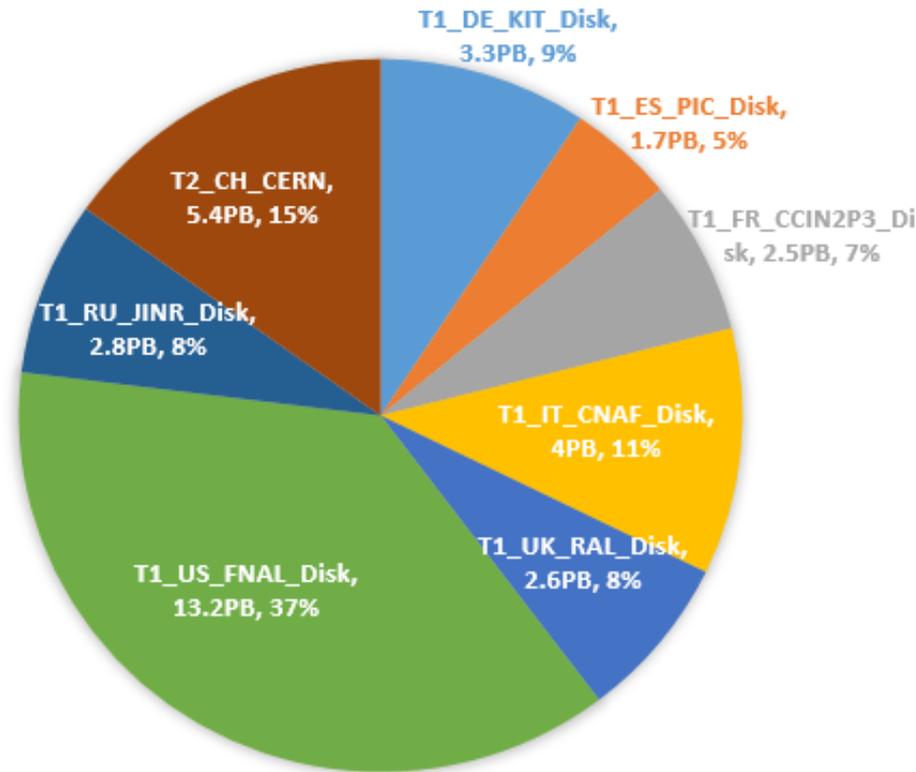


T0 – T1 Storage Overview Pledges

TAPE ENDPOINTS = 131.6PB



DISK ENDPOINTS = 35PB



Disk - Tape Separation

Old Setup:

- One MSS system controlling both disk and tape.
- Files written to tape automatically
- User analysis not allowed at Tier1s. (Only for Expert Users)
- Stage as needed on demand inefficient
- Processing always had to happen at the archiving location (Limiting flexibility where to run)



Procedure:

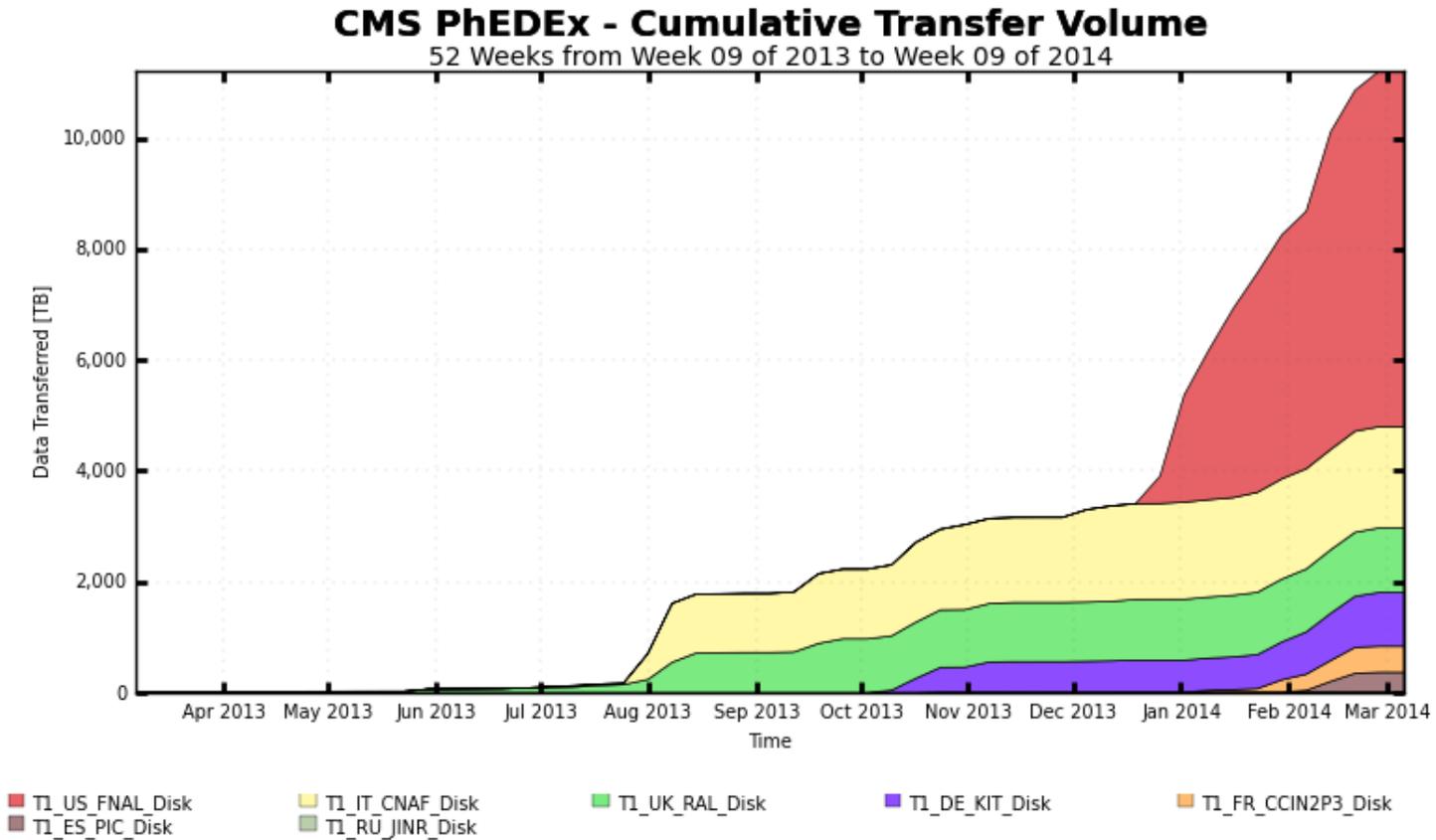
- Deploy two independent endpoints: Tape and DiskConnect new Tier1 Disk endpoints to the other sites.
- Tape reading/writing becomes a subscription

New Setup:

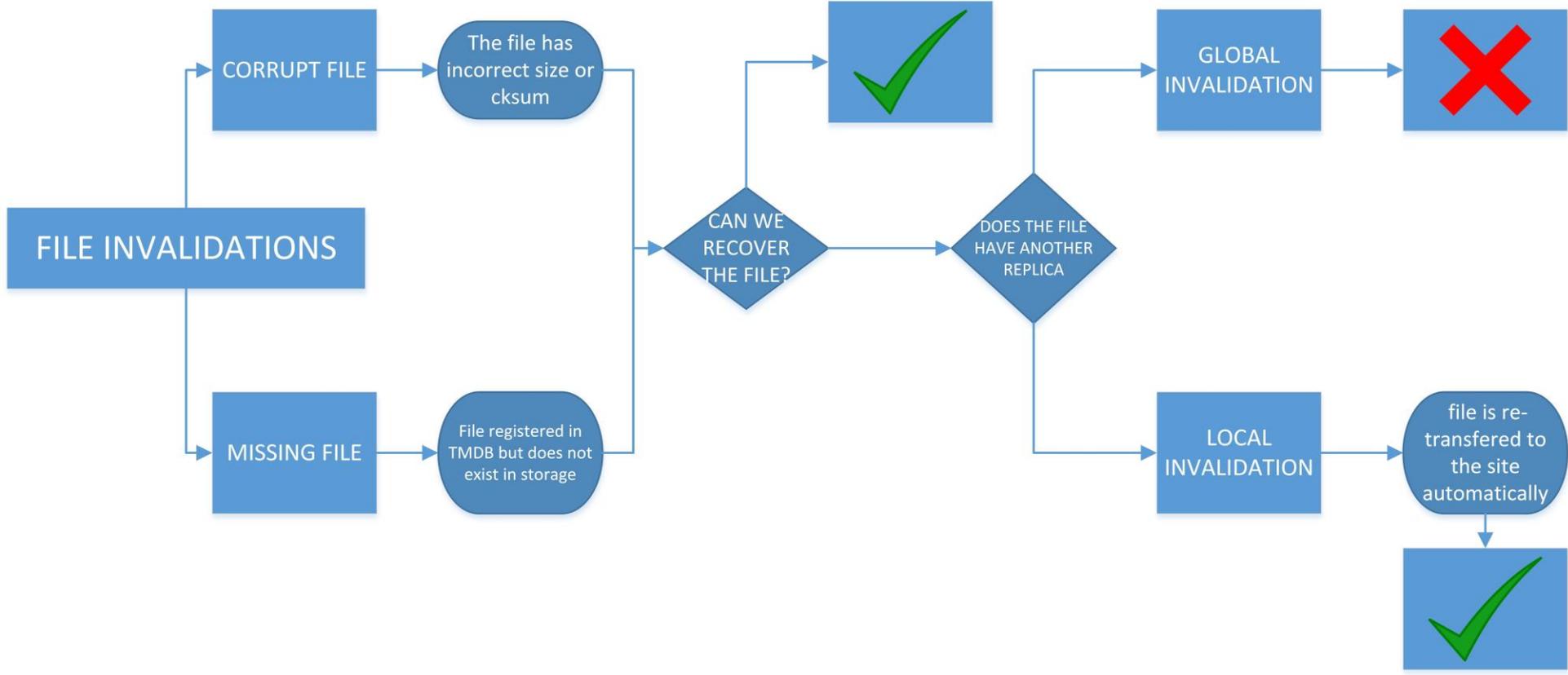
- Increase flexibility for Tier-1 processing . Enable user analysis at Tier-1s. Enable remote access of Tier-1 Disk data
- Jobs can only Access files on disk endpoints
- No automated tape migration
- Tape: Data can be accessed only for reading and writing through PhEDEx

Disk - Tape Separation

- Started by RAL in April 2013. Completed by FNAL in March 2014
- Disk endpoint Storage = 11.215TB



File Invalidation



Consistency Check

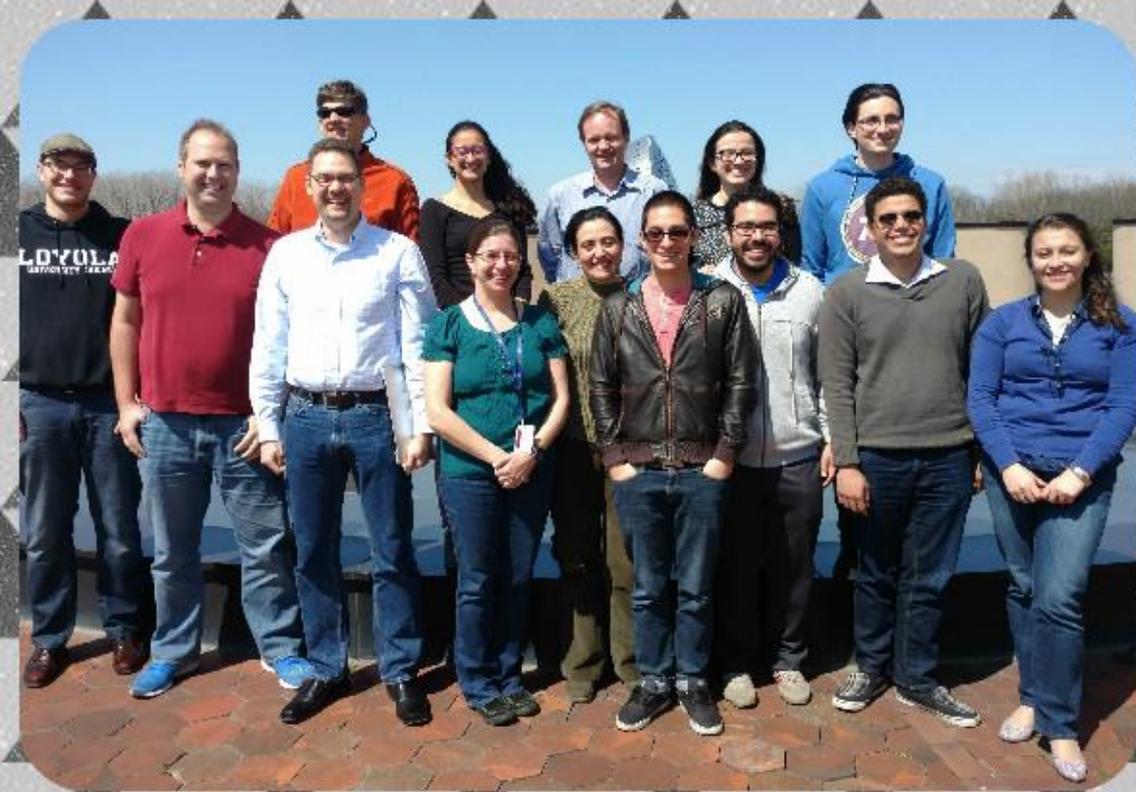
- Periodically, Transfer Team runs CCs, to synchronize TMDB and Storage (Disk and Tape)
- SCC: Storage Consistency Check. Check if all files in storage are registered in the database. 'Orphans' are removed from storage.
- BDV: Block Download Verify. Check if all files registered in the database can be accessed (Size & checksum). Files are locally or globally invalidated.
- Round April 2016.
T1_FR_CCIN2P3_MSS. 13 orphan files deleted. 1 file invalidated locally



Link Commissioning

- To be used in the Production instance of [PhEDEX](#), a data transfer link must first go through a commissioning procedure in the Debug instance using the LoadTest infrastructure.
- LoadTest Files 2.5GB
- T1 → T2: 20MB/s (680 files in 24h)
- T2 → T1: 5MB/s (169 files in 24h)

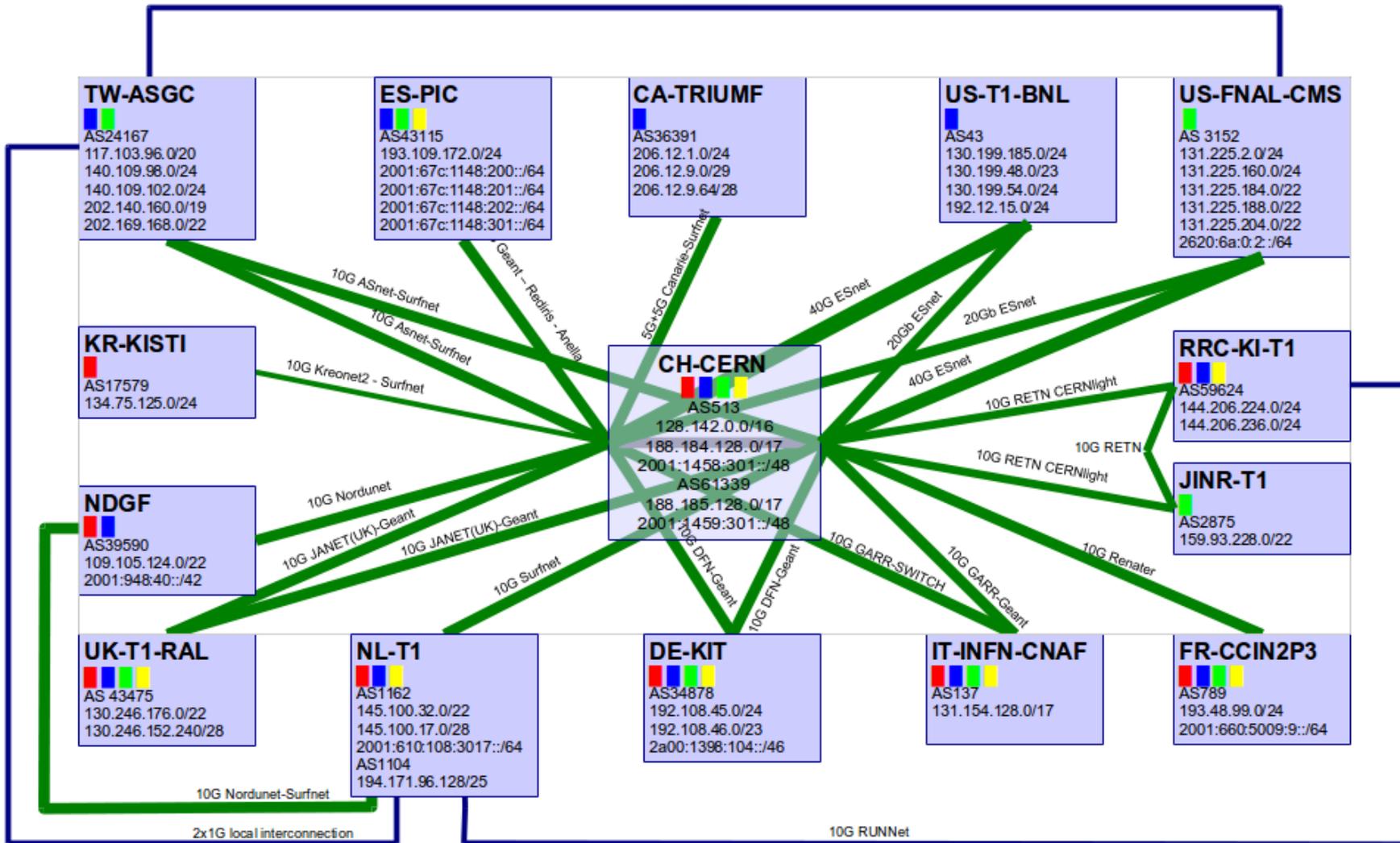




FERMILAB ROC

LHCOPN

2G ASnet



— T0-T1 and T1-T1 traffic
— T1-T1 traffic only
- - - Not deployed yet
(thick) >= 10Gbps
(thin) <10Gbps

■ = Alice ■ = Atlas
■ = CMS ■ = LHCb

p2p prefix: 192.16.166.0/24 - 2001:1458:302::/48
 edoardo.martelli@cern.ch 20160322