Experiences with Grid based data movement using the CMS data export tool PhEDEx

Tim Barrass – University of Bristol
Jens Rehn – CERN
Lassi A. Tuura – Northeastern University

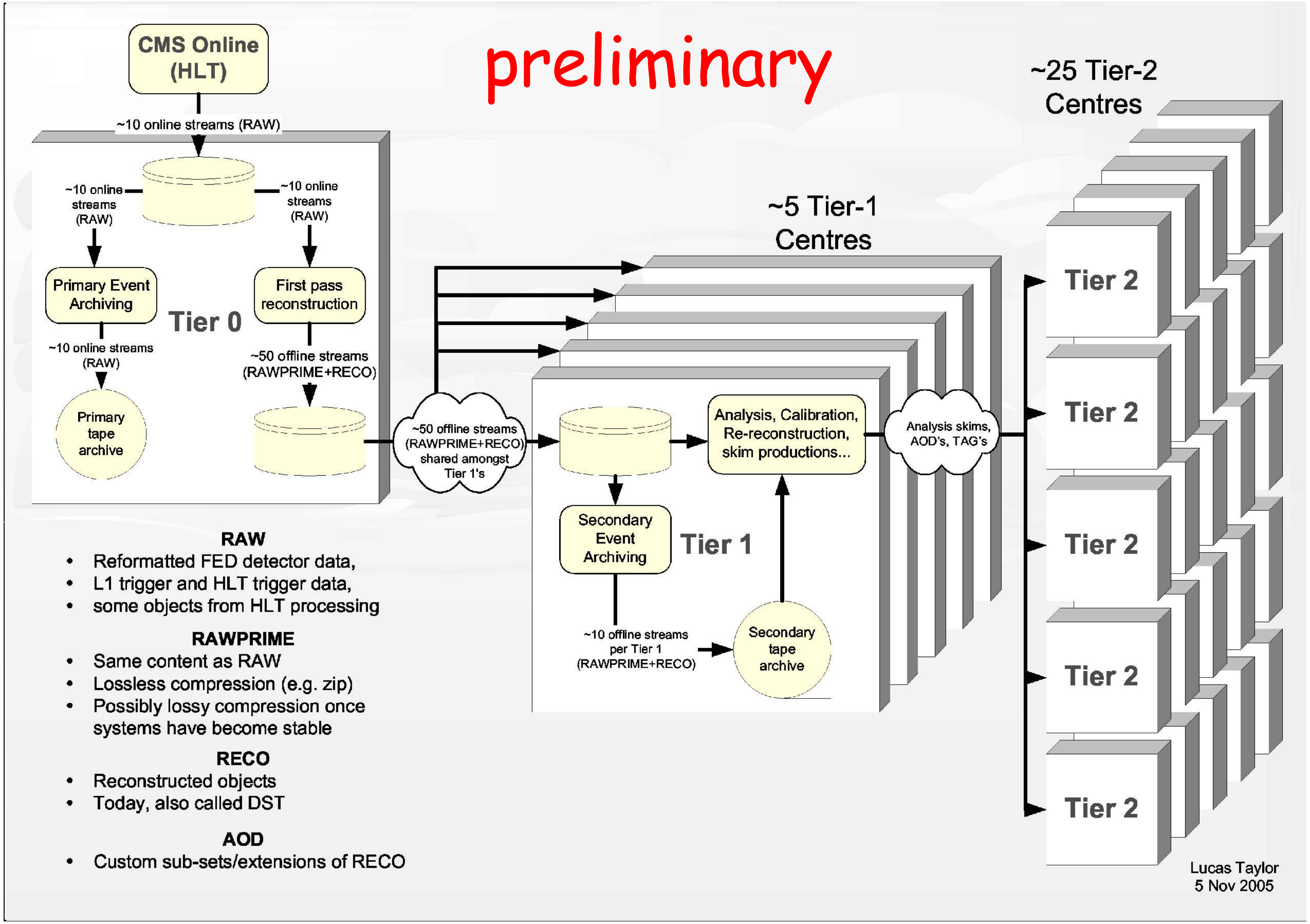
Outline

- CMS computing model
  - data organisation
  - data flow from Cern to T1 centers
- PhEDEx – a tool for mass data movement
- Data flow from Cern to GridKa
  - import / export strategies
  - some plots and numbers

The CMS computing model types of data

- RAW data
  - 0.6 +/- 0.3 MB per event (low lumi phase)
  - 1.5 +/- 0.5 MB per event (high lumi phase)
- RAW Prime data
  - lossless compressed RAW data
  - later maybe lossy compression
- Reconstructed data (= DST)
- FEVT (full events) consist of RAW Prime + DST
  - shipped to T1 centers

The CMS computing model streams of data



RAW

- Reformatted FED detector data,
- L1 trigger and HLT trigger data,
- some objects from HLT processing

RAWPRIME

- Same content as RAW
- Lossless compression (e.g. zip)
- Possibly lossy compression once systems have become stable

RECO

- Reconstructed objects
- Today, also called DST

AOD

- Custom sub-sets/extensions of RECO

PhEDEx - challenges and features (1)

- 100k files with 5 replicas each per month
- Reliable transfers
  - transfer states: system writes current state to disk
  - checking file-size and cksums
  - multi-hop transfers with fall back routes
- Fulfill transfer needs
  - offering push, **pull** and stream models
  - data subscriptions; independent from file origin
  - web interface for transfer requests & subscriptions

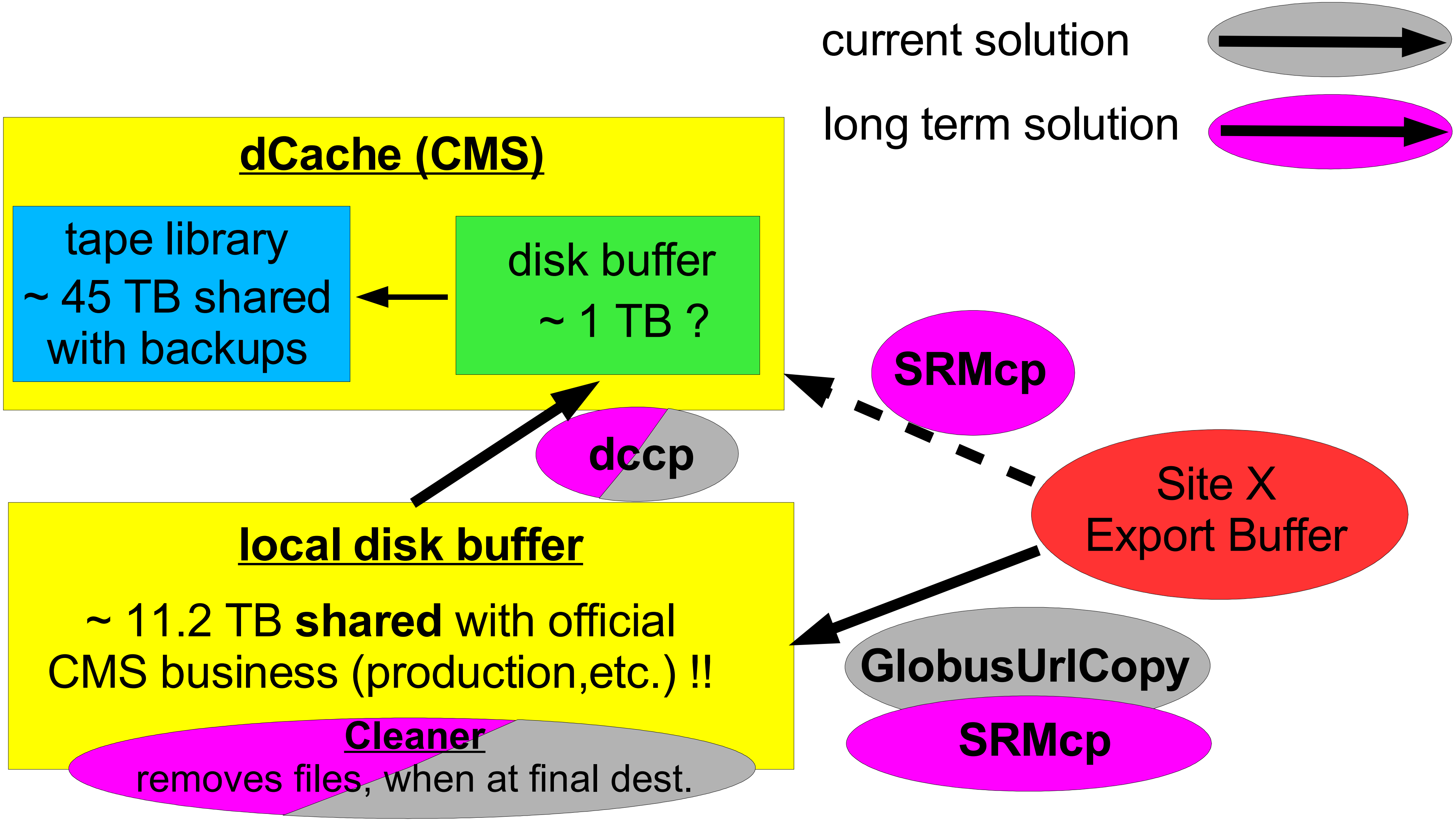
PhEDEx - challenges and features (2)

- Buffer space management
  - cleaner: removes files from disk
  - stage pool management (not yet impl.)
- Monitoring
  - status web page
  - interface to MonaLisa monitoring
- Protocol matching
  - support multiple backends: g-u-c, srmcp, dccp, lcg-rep
  - automatic protocol matching (not yet impl.)

Data movement
GridKa import

- Current import strategy
  - based on globus-url-copy to local disk
  - dccp to dCache interfaced MSS
- Future import strategy
  - using srmcp only for data replication
  - still via local disk-buffer ?
    - transfer tools unreliable (1-5%) => lots of delete operations
    - would stress dCache and tapes ?

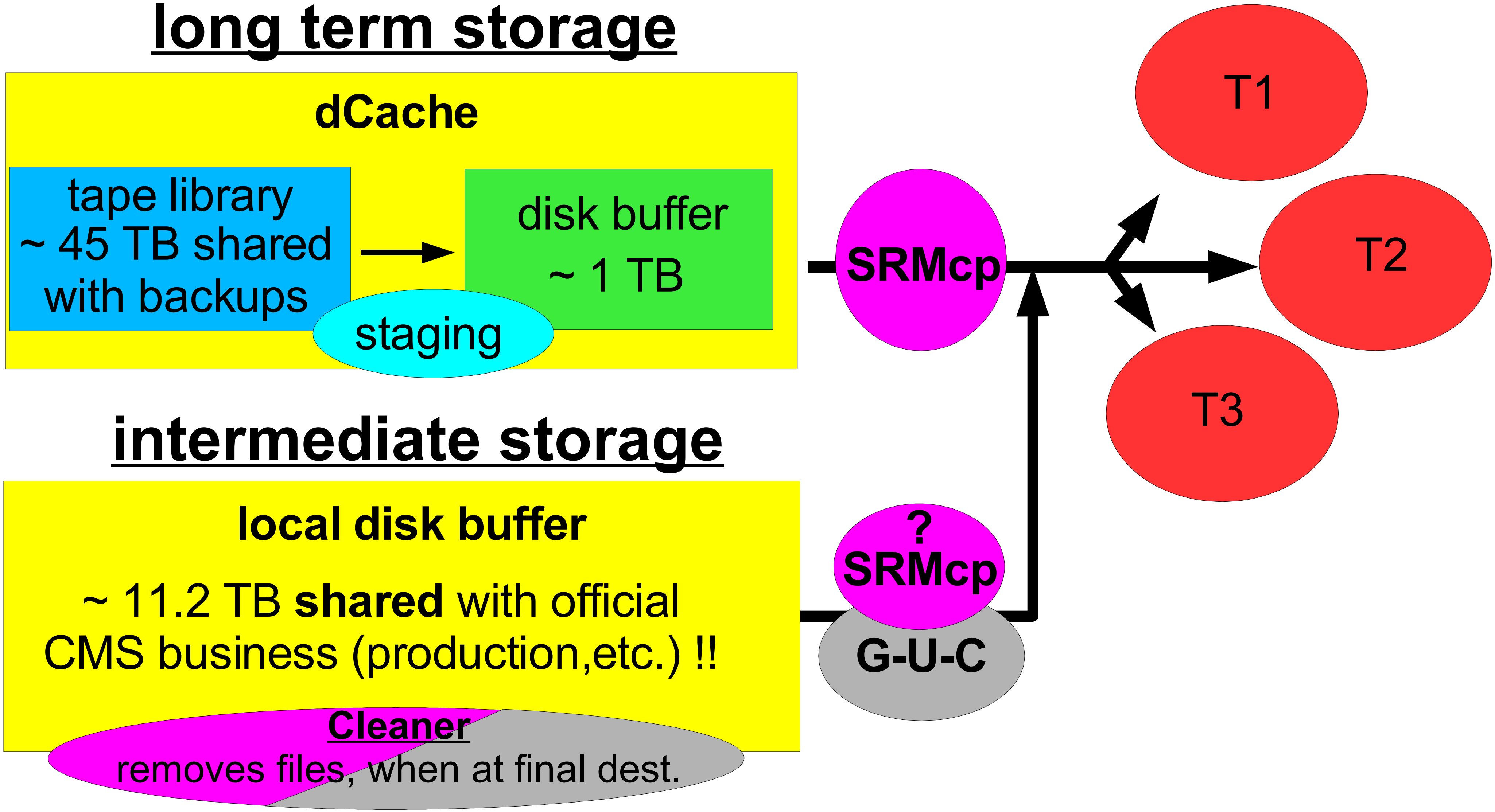
Data movement GridKa import



Data movement GridKa export

- Export from MSS via dCache
  - manages its buffer space automatically
  - no explicit cleaning of stage pool necessary
  - intelligent prestaging with auto balanced replicas
  - dCache interfaced with SRM
- Export from buffer disk
  - only for intermediate transfers (FZK not final dest.)
  - files will get deleted when at final destination

Data movement GridKa export



long term storage

dCache

tape library
~ 45 TB shared
with backups

disk buffer
~ 1 TB

staging

SRMcp

T1

T2

T3

intermediate storage

local disk buffer

~ 11.2 TB **shared** with official
CMS business (production,etc.) !!

Cleaner
removes files, when at final dest.

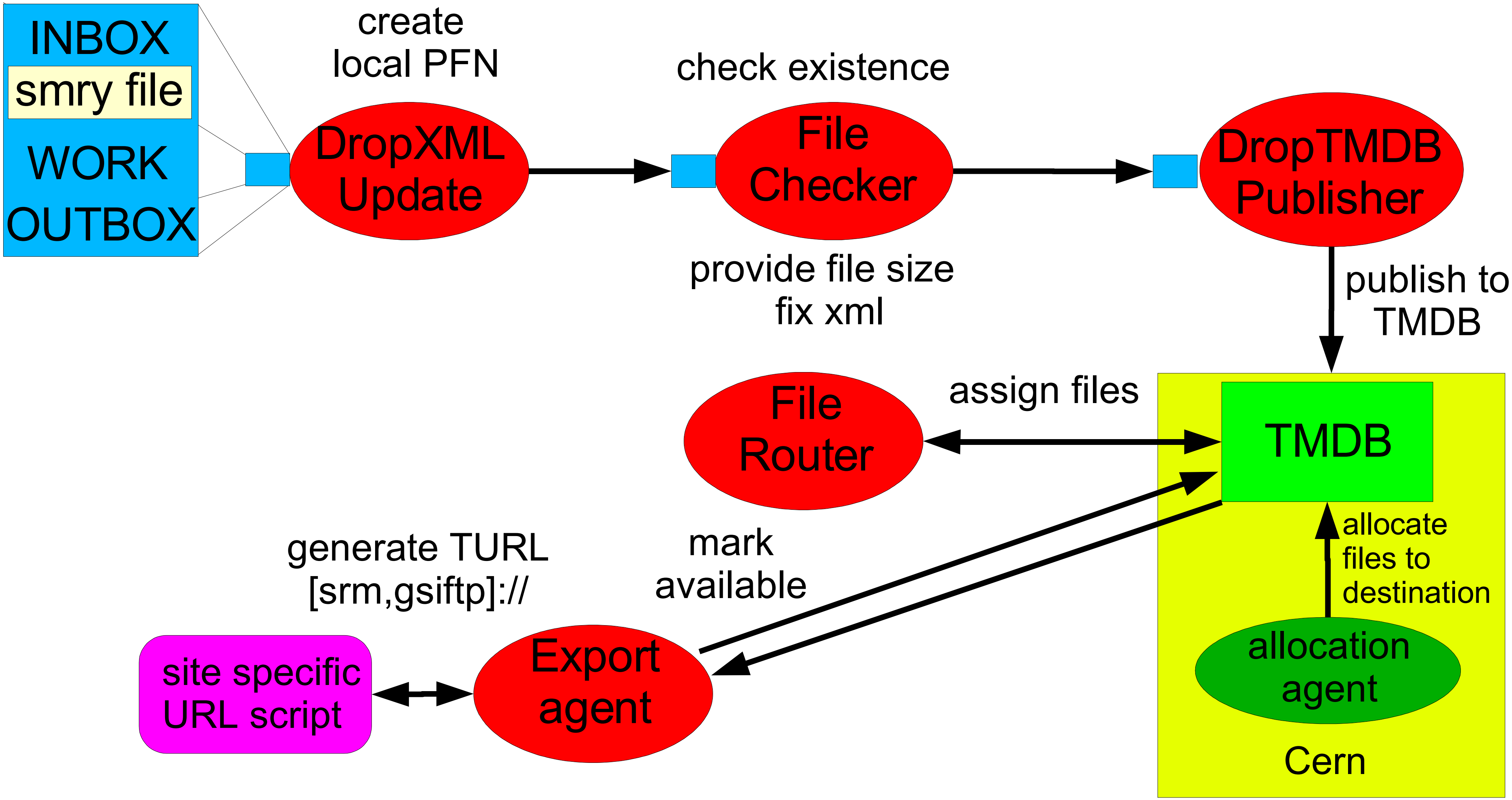
?
SRMcp

G-U-C

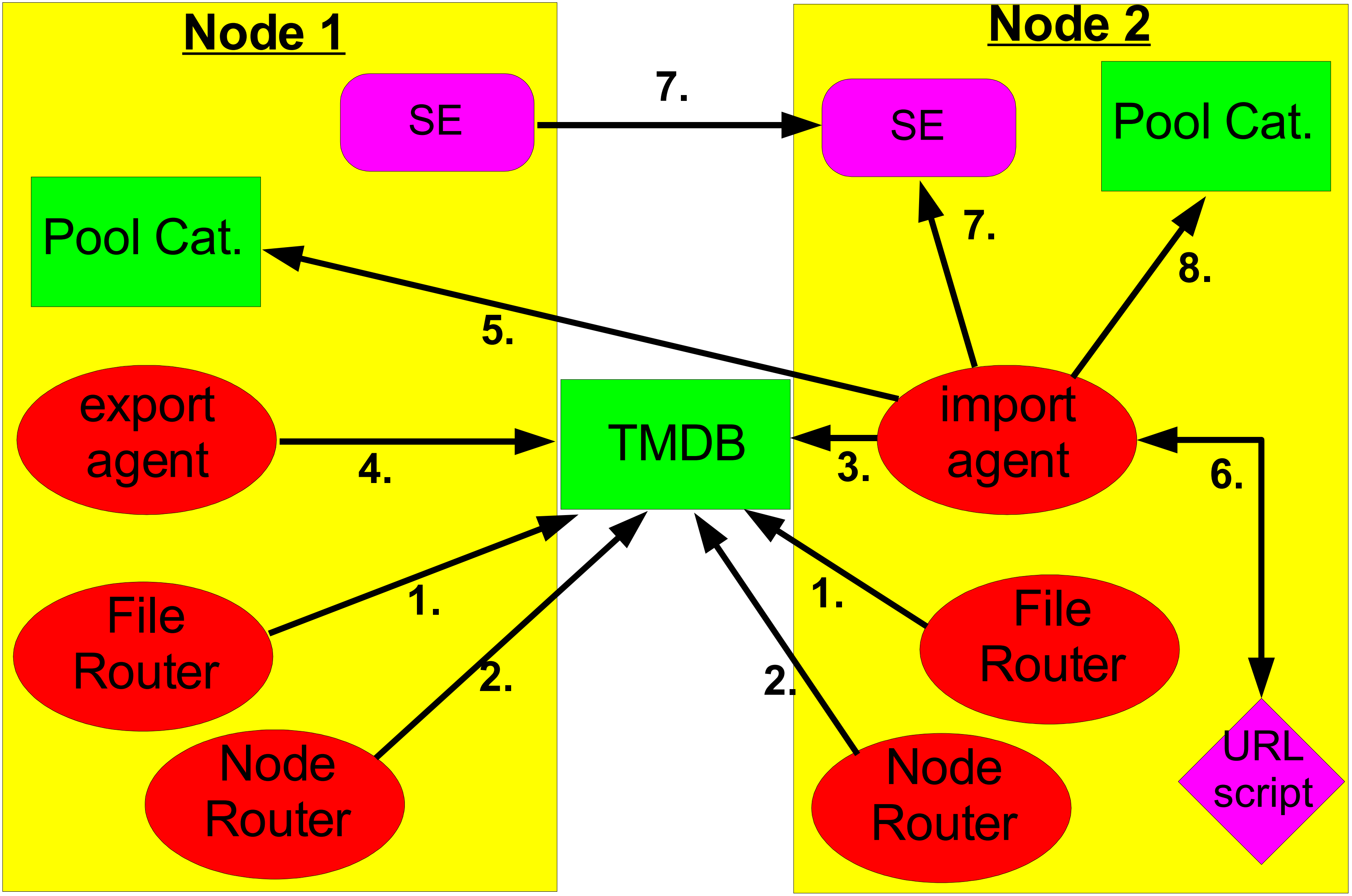
Data movement via SRM why using SRM ?

- Negotiates transfer protocol (e.g. gridftp)
- Checks available space for file
- Assures correct file transfer (check sums)
  - but failed transfers produce exit code 0 also here ?!
  - PhEDEx does filesize and cksum checks on its own
- Initiates file staging (e.g. on dCache)

PhEDEx - drop-box data injection & export

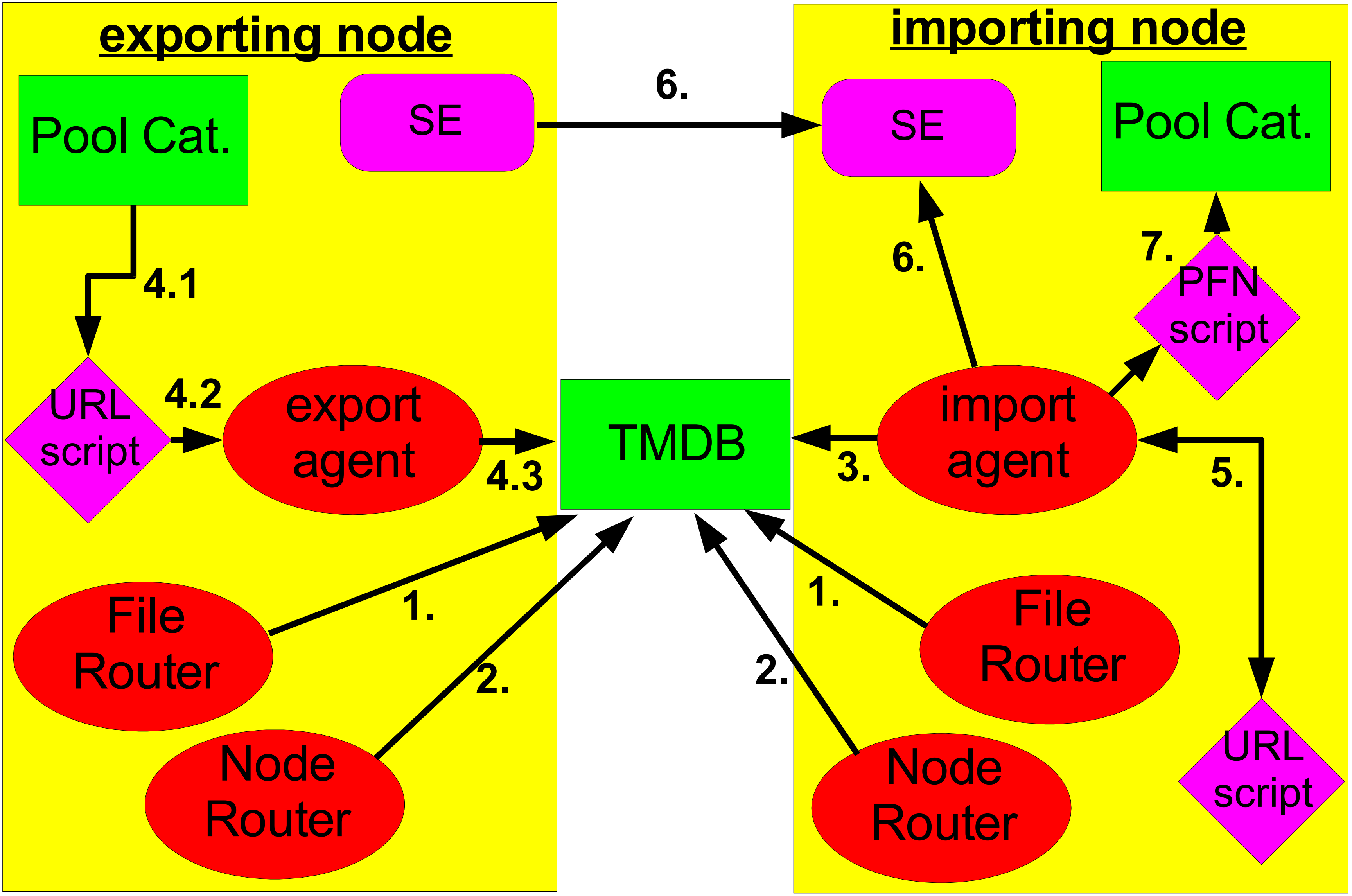


PhEDEx - drop-box data import (now)



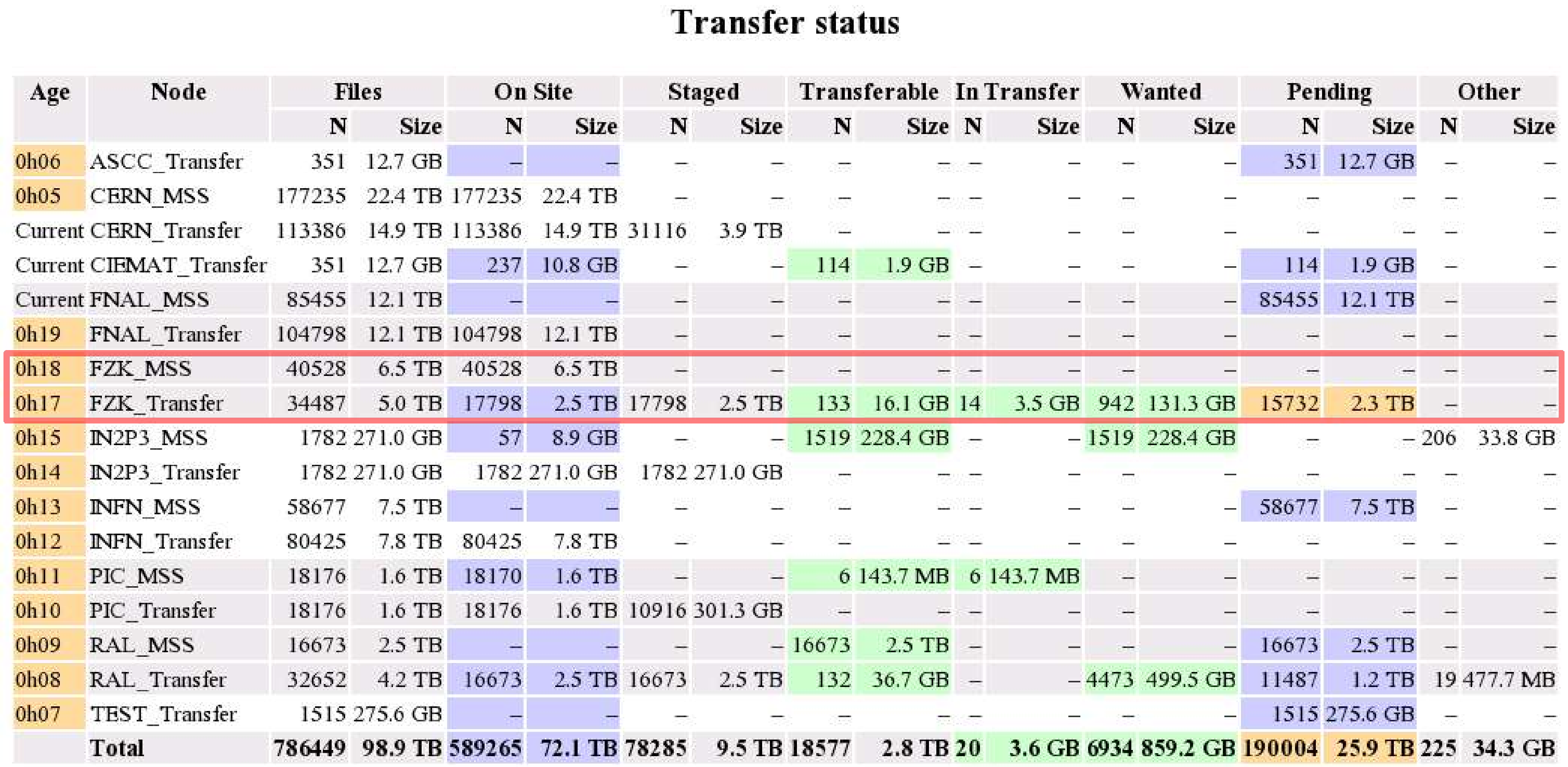
1. assign files
2. maintain routes
3. mark wanted
4. mark available
5. get S-PFN,...
6. gen. D-PFN
7. init transfer
8. store PFN

PhEDEx - drop-box data import (future)



1. assign files
2. maintain routes
3. mark wanted
4.1 get PFN
4.2 derive TURL
4.3 advert. TURL mark avail.
5. gen. loc. TURL
6. init transfer
7. derive PFN store in POOL

Overview of data moved recent PRS request



Transfer status

Age	Node	Files		On Site		Staged		Transferable		In Transfer		Wanted		Pending		Other	
		N	Size	N	Size	N	Size	N	Size	N	Size	N	Size	N	Size	N	Size
0h06	ASCC_Transfer	351	12.7 GB	–	–	–	–	–	–	–	–	–	–	351	12.7 GB	–	–
0h05	CERN_MSS	177235	22.4 TB	177235	22.4 TB	–	–	–	–	–	–	–	–	–	–	–	–
Current	CERN_Transfer	113386	14.9 TB	113386	14.9 TB	31116	3.9 TB	–	–	–	–	–	–	–	–	–	–
Current	CIEMAT_Transfer	351	12.7 GB	237	10.8 GB	–	–	114	1.9 GB	–	–	–	–	114	1.9 GB	–	–
Current	FNAL_MSS	85455	12.1 TB	–	–	–	–	–	–	–	–	–	–	85455	12.1 TB	–	–
0h19	FNAL_Transfer	104798	12.1 TB	104798	12.1 TB	–	–	–	–	–	–	–	–	–	–	–	–
0h18	FZK_MSS	40528	6.5 TB	40528	6.5 TB	–	–	–	–	–	–	–	–	–	–	–	–
0h17	FZK_Transfer	34487	5.0 TB	17798	2.5 TB	17798	2.5 TB	133	16.1 GB	14	3.5 GB	942	131.3 GB	15732	2.3 TB	–	–
0h15	IN2P3_MSS	1782	271.0 GB	57	8.9 GB	–	–	1519	228.4 GB	–	–	1519	228.4 GB	–	–	206	33.8 GB
0h14	IN2P3_Transfer	1782	271.0 GB	1782	271.0 GB	1782	271.0 GB	–	–	–	–	–	–	–	–	–	–
0h13	INFN_MSS	58677	7.5 TB	–	–	–	–	–	–	–	–	–	–	58677	7.5 TB	–	–
0h12	INFN_Transfer	80425	7.8 TB	80425	7.8 TB	–	–	–	–	–	–	–	–	–	–	–	–
0h11	PIC_MSS	18176	1.6 TB	18170	1.6 TB	–	–	6	143.7 MB	6	143.7 MB	–	–	–	–	–	–
0h10	PIC_Transfer	18176	1.6 TB	18176	1.6 TB	10916	301.3 GB	–	–	–	–	–	–	–	–	–	–
0h09	RAL_MSS	16673	2.5 TB	–	–	–	–	16673	2.5 TB	–	–	–	–	16673	2.5 TB	–	–
0h08	RAL_Transfer	32652	4.2 TB	16673	2.5 TB	16673	2.5 TB	132	36.7 GB	–	–	4473	499.5 GB	11487	1.2 TB	19	477.7 MB
0h07	TEST_Transfer	1515	275.6 GB	–	–	–	–	–	–	–	–	–	–	1515	275.6 GB	–	–
	Total	**786449**	**98.9 TB**	**589265**	**72.1 TB**	**78285**	**9.5 TB**	**18577**	**2.8 TB**	**20**	**3.6 GB**	**6934**	**859.2 GB**	**190004**	**25.9 TB**	**225**	**34.3 GB**

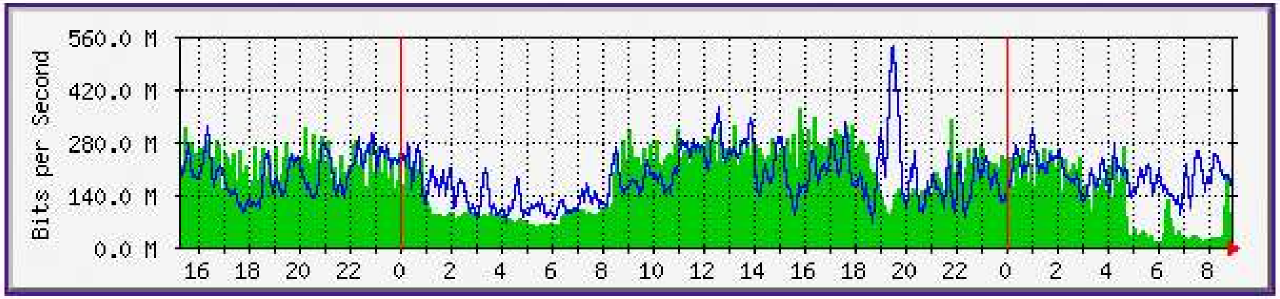
http://cms-project-phedex.web.cern.ch/cms-project-phedex

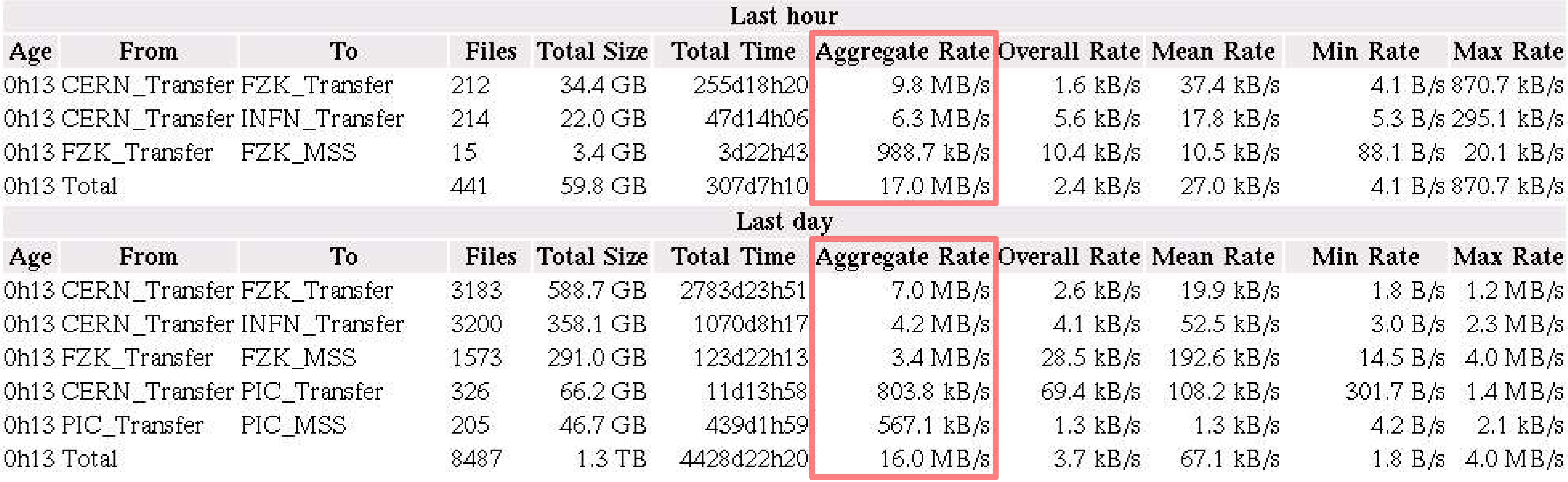
Transfer performance

up to ~2 TB a day outbound
limited by network

Transfer Rate Statistics

CERN Daily External Internet Traffic





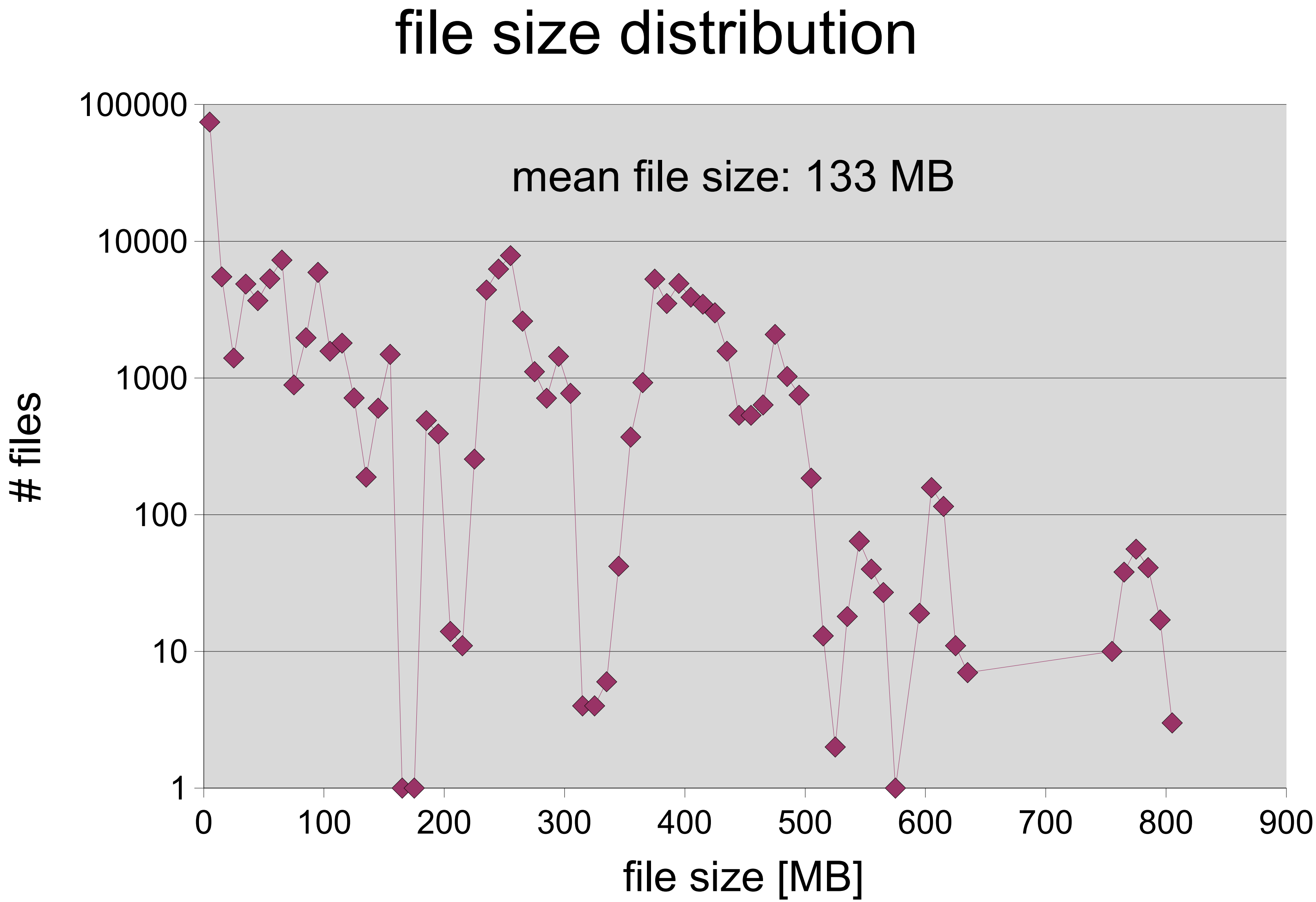
Last hour

Age	From	To	Files	Total Size	Total Time	Aggregate Rate	Overall Rate	Mean Rate	Min Rate	Max Rate
0h13	CERN_Transfer	FZK_Transfer	212	34.4 GB	255d18h20	9.8 MB/s	1.6 kB/s	37.4 kB/s	4.1 B/s	870.7 kB/s
0h13	CERN_Transfer	INFN_Transfer	214	22.0 GB	47d14h06	6.3 MB/s	5.6 kB/s	17.8 kB/s	5.3 B/s	295.1 kB/s
0h13	FZK_Transfer	FZK_MSS	15	3.4 GB	3d22h43	988.7 kB/s	10.4 kB/s	10.5 kB/s	88.1 B/s	20.1 kB/s
0h13	Total		441	59.8 GB	307d7h10	17.0 MB/s	2.4 kB/s	27.0 kB/s	4.1 B/s	870.7 kB/s

Last day

Age	From	To	Files	Total Size	Total Time	Aggregate Rate	Overall Rate	Mean Rate	Min Rate	Max Rate
0h13	CERN_Transfer	FZK_Transfer	3183	588.7 GB	2783d23h51	7.0 MB/s	2.6 kB/s	19.9 kB/s	1.8 B/s	1.2 MB/s
0h13	CERN_Transfer	INFN_Transfer	3200	358.1 GB	1070d8h17	4.2 MB/s	4.1 kB/s	52.5 kB/s	3.0 B/s	2.3 MB/s
0h13	FZK_Transfer	FZK_MSS	1573	291.0 GB	123d22h13	3.4 MB/s	28.5 kB/s	192.6 kB/s	14.5 B/s	4.0 MB/s
0h13	CERN_Transfer	PIC_Transfer	326	66.2 GB	11d13h58	803.8 kB/s	69.4 kB/s	108.2 kB/s	301.7 B/s	1.4 MB/s
0h13	PIC_Transfer	PIC_MSS	205	46.7 GB	439d1h59	567.1 kB/s	1.3 kB/s	1.3 kB/s	4.2 B/s	2.1 kB/s
0h13	Total		8487	1.3 TB	4428d22h20	16.0 MB/s	3.7 kB/s	67.1 kB/s	1.8 B/s	4.0 MB/s

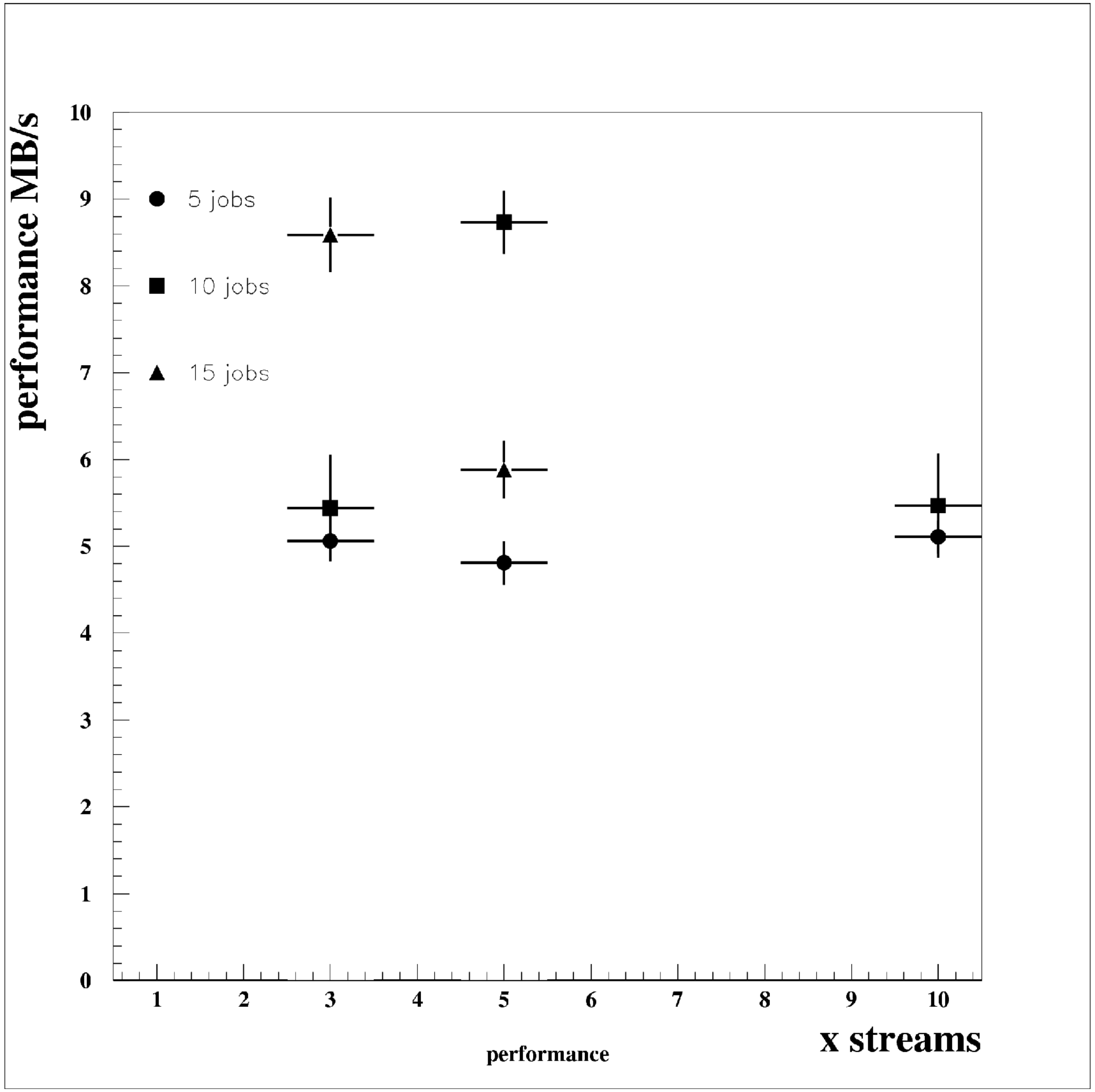
Typical size of files



Summary & outlook

- 40 TB of data successfully moved in total
  - ~10 TB during DC04
  - ~30 TB so far during recent PRS request
- System proved capable to handle requirements
  - provides reliable transfers
  - redundant distribution chain
- Central TMDB potential single point of failure
  - distribution of DB planned
  - looking into P2P solutions

performance tuning of globus-url-copy



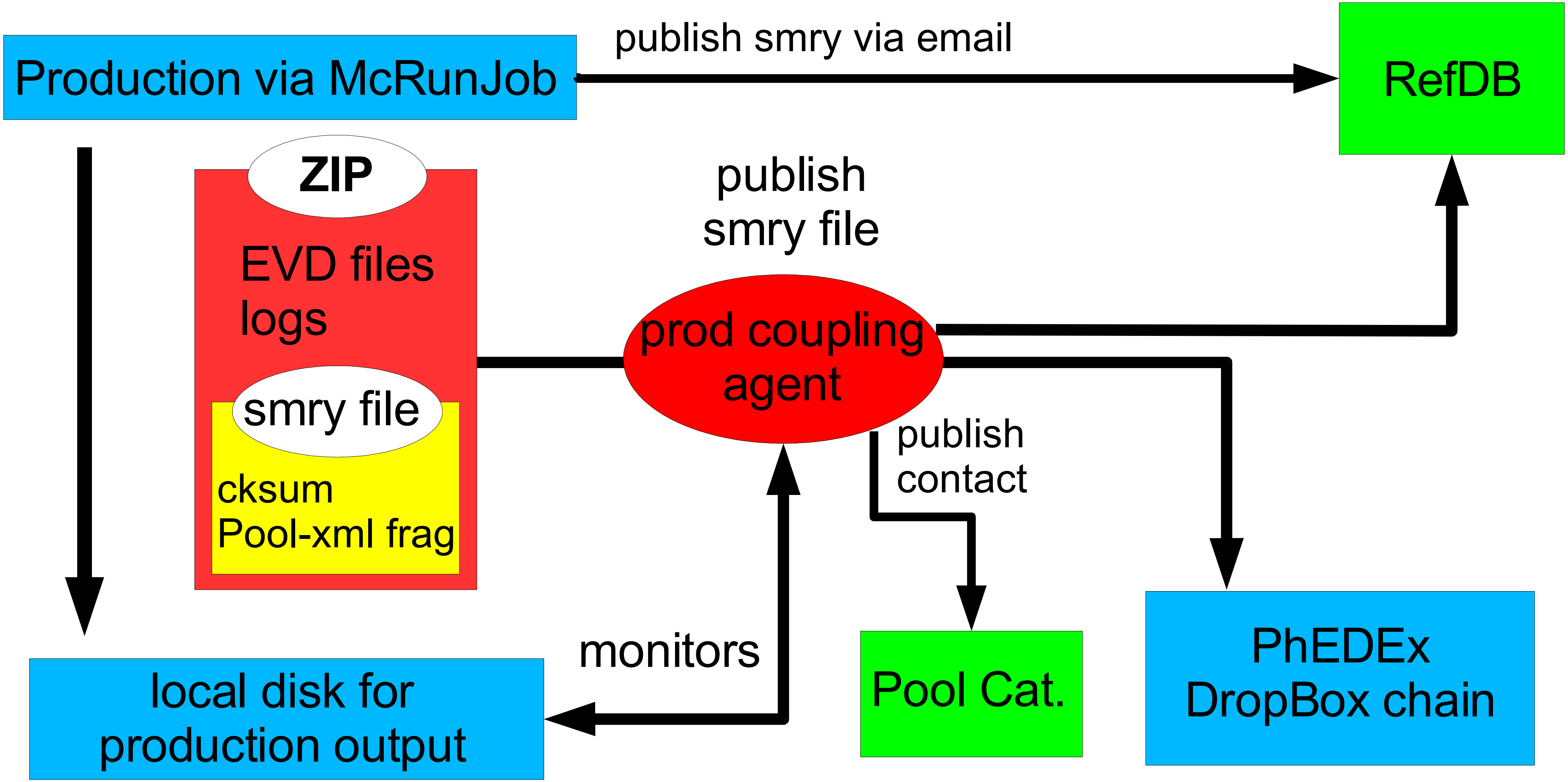
<u>tuneable parameters:</u>

- amount of g-u-c jobs
- amount of streams

<u>rule of thumb:</u>

# jobs * # streams ≈ 50

Production coupling to PhEDEx



Production via McRunJob
publish smry via email
RefDB
ZIP
EVD files
logs
smry file
cksum
Pool-xml frag
publish
smry file
prod coupling agent
publish
contact
monitors
local disk for production output
Pool Cat.
PhEDEx DropBox chain