CHEP2015
OKINAWA, japan

Pooling the Resources of the CMS Tier-1 Sites

Oliver Gutsche (FNAL), Nicolo Magini (FNAL), Christoph Wissing (DESY)
for the CMS Collaboration

April 14th, 2015

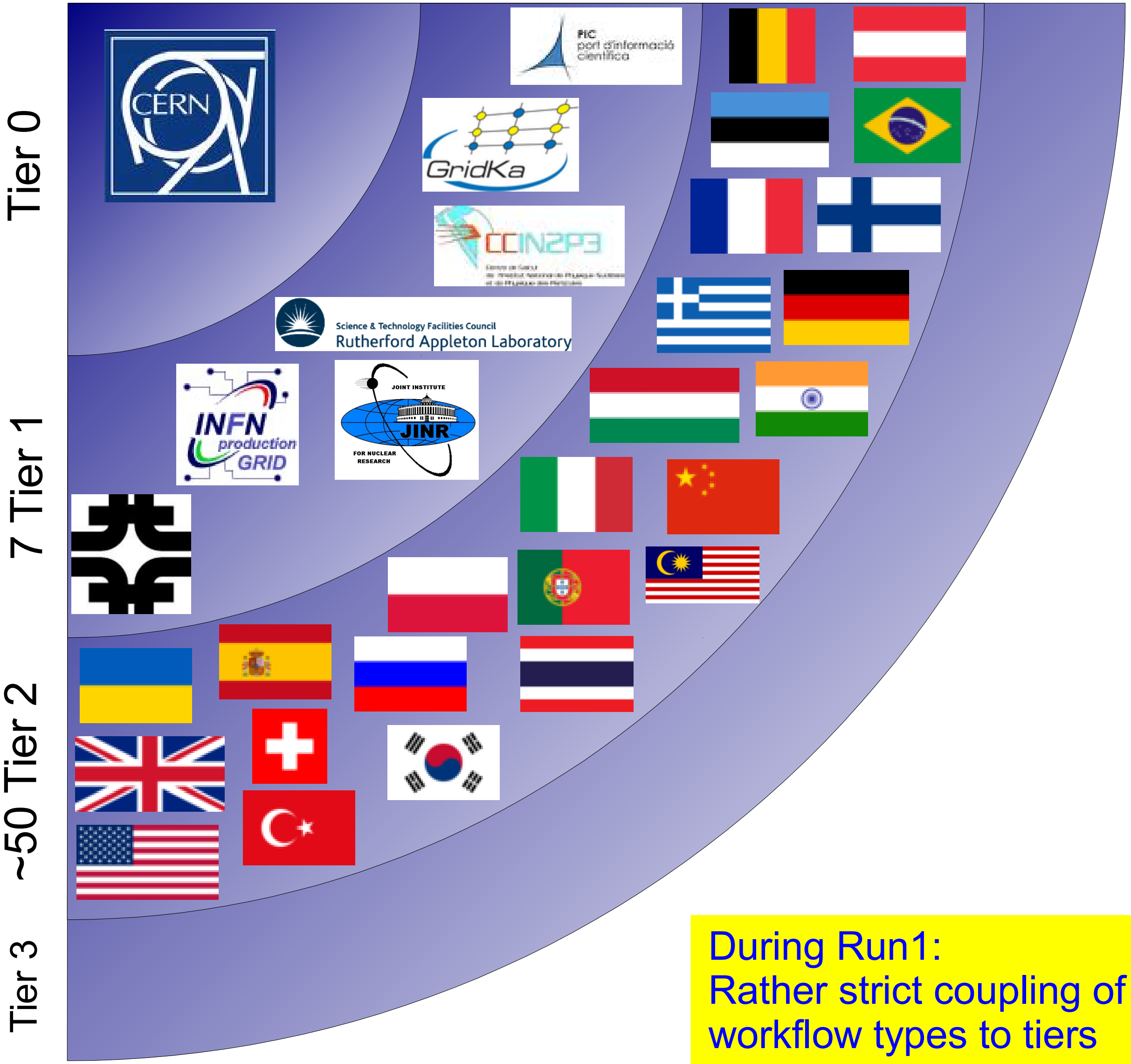
HELMHOLTZ
ASSOCIATION

Fermilab

DESY

Distributed Computing Infrastructure

More than 50 CMS centers, in more than 20 countries



Tier 0
7 Tier 1
~50 Tier 2
Tier 3
CERN
PIC port d'informació científica
GridKa
CCIN2P3
Science & Technology Facilities Council
Rutherford Appleton Laboratory
INFN production GRID
JOINT INSTITUTE FOR NUCLEAR RESEARCH
JINR

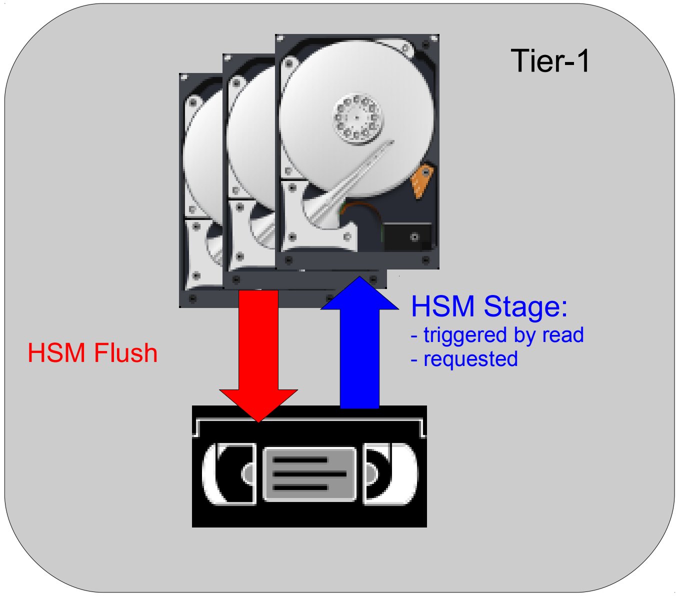
During Run1:
Rather strict coupling of workflow types to tiers

Flags taken from Wikipedia:
http://de.wikipedia.org/wiki/Liste_der_Nationalflaggen

- Tier 0
  - Main task in Run1:
    - Prompt reconstruction
    - Store RAW data and export to T1s
  - Disk and tape storage
- Tier 1
  - Main tasks in Run1:
    - Re-reconstruction & MC production
    - Long term storage of RAW and MC files
  - Disk and tape storage
- Tier 2
  - Main tasks in Run1:
    - MC production
    - User analysis
  - Only disk storage

Tape Configuration and Operations in Run1

- Disk and tape space coupled through HSM
  - Files written to tape automatically
    (immediately or as soon as possible)
  - Disk (usually) gets flushed from disk when space is needed on buffer disks
- Staging form tape: 3 cases
  - On demand: when file gets requested
  - Through SRM request
  - In practice often by ticket to site
- Pinning on disk: 2 cases
  - Through SRM commands
  - Again using tickets



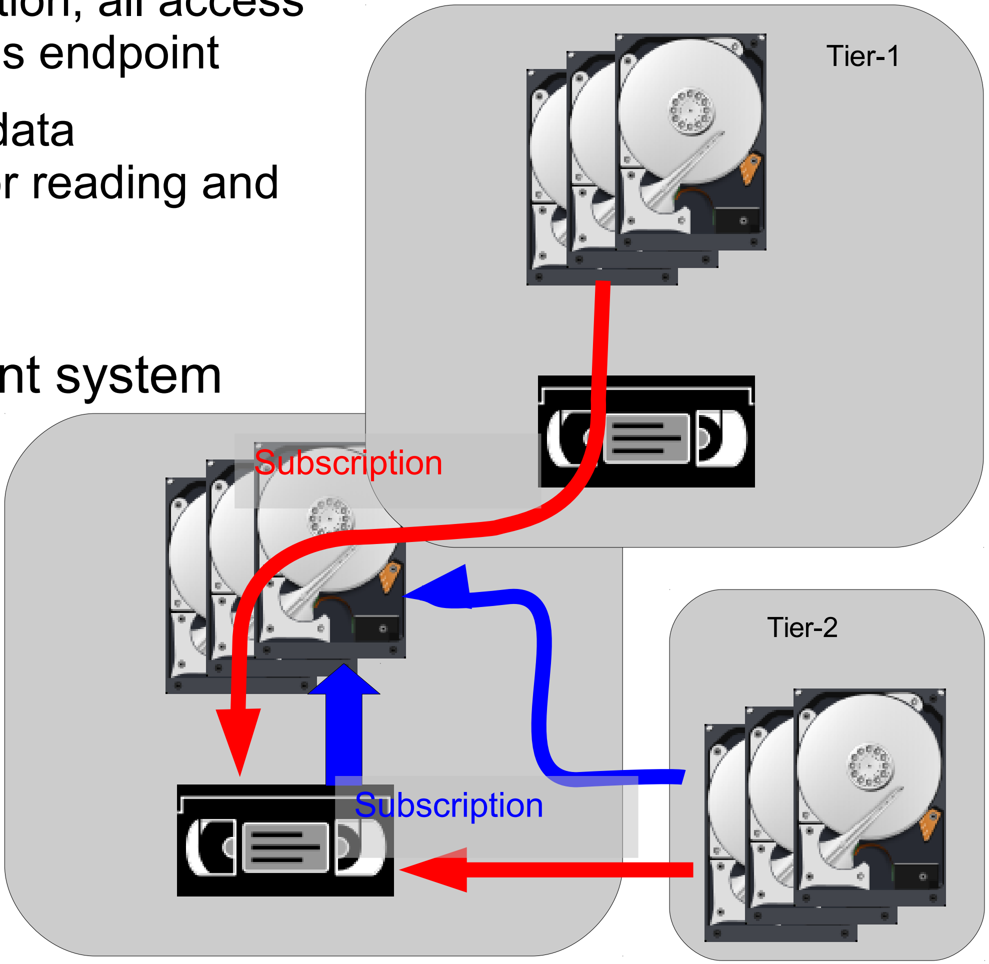
Implications of Run1 Setup

- Strict coupling of processing and tape archival of output
  - Processing always had to happen at the archiving location
  - Limiting flexibility where to run
- Limited Tier-1 access for analysis users
  - No easy way to figure out what files are on disk
  - Uncontrolled tape staging needs to be avoided
  - CMS allowed only “expert users” to run at Tier-1 using VOMS role `t1access`
- Difficult to include Tier-1 sites into AAA data federation
  - Files need to be on disk for remote access
  - Requires an easy way to determine what is on disk

Solution: Separation of disk and tape archiving at Tier-1s

Disk Tape Separation

- Basic concept
  - Separation into two logical parts
    - Disk endpoint: no automated tape migration, all access from CPU and AAA data federation to this endpoint
    - Archive: automatic tape migration, only data management system can access data for reading and writing
  - Transition from disk to tape becomes a Subscription in the data management system



- Implementation at the sites
  - Two independent storage systems
  - Split namespace

Technical Implementation

- Sites free to choose the most suitable solution for their storage systems
- Different storage instances
  - CERN: CASTOR for tape and EOS for disk
  - FNAL: Two dCache instances (+ EOS for user data)
  - JINR: Only dCache disk atm, plans another dCache instance for tape
- Two independent namespace trees on the same storage
  - RAL: CASTOR
  - KIT, CCIN2P3, PIC: dCache
  - CNAF: GPFS with StoRM
- Transfers between the two areas managed with the standard WLCG service: FTS

CASTOR

CERN Advanced STORage manager

EOS

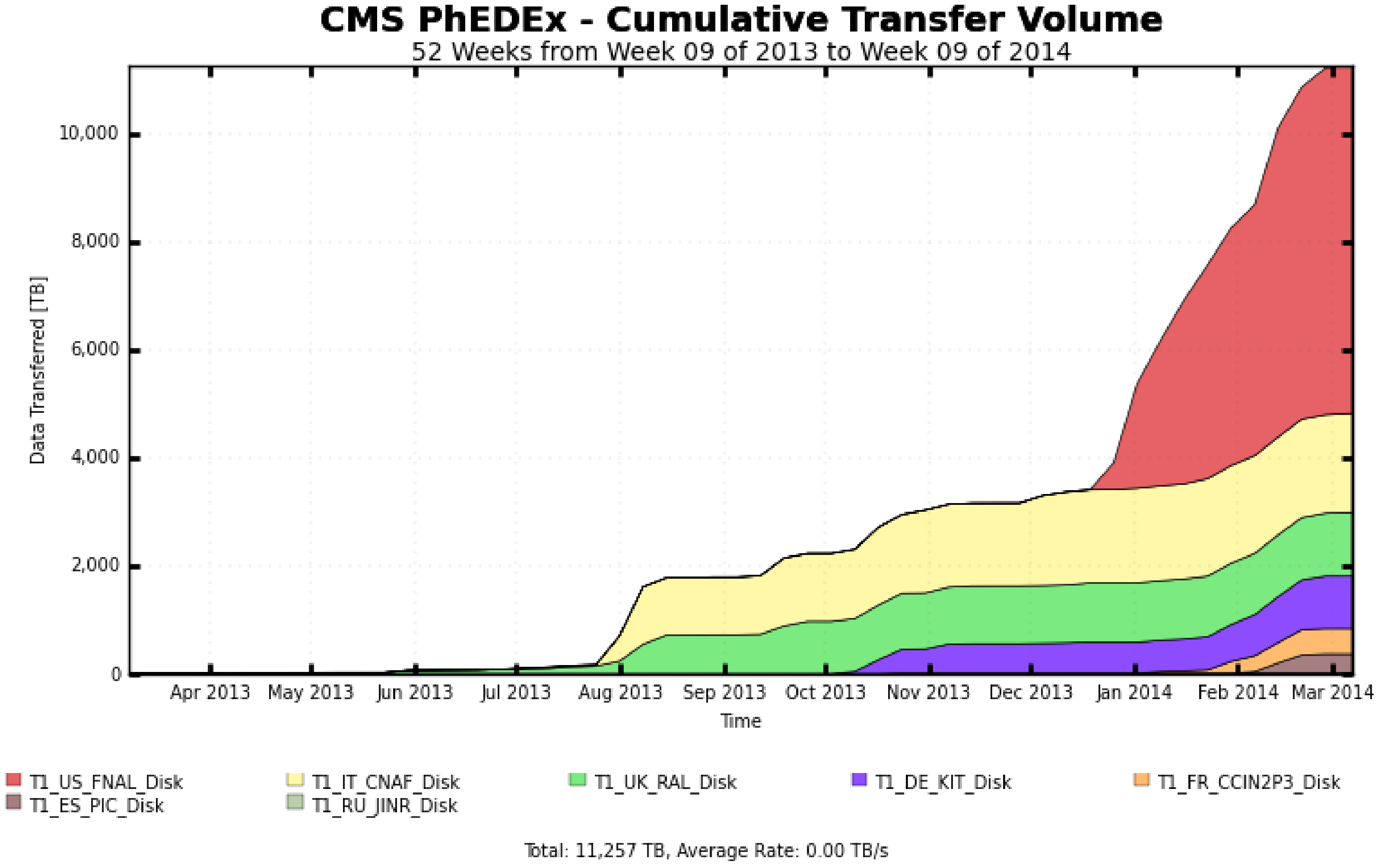
dCache

StoRM

Storage Resource Manager

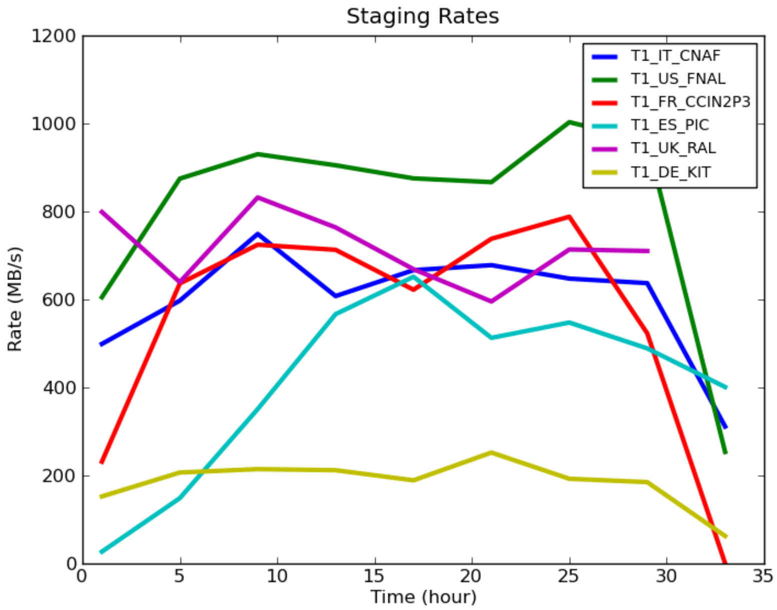
Population of new Disk Endpoints

- Pioneered by RAL in April 2013, completed at FNAL in March 2014
- New disk endpoints populated with over 10 PB of data during the migration



Commissioning of Sites and Transfer System

- Change site configuration to interact with Disk endpoint only
  - Mapping of Logical File Name (LFN) to URL via Trivial File Catalog (TFC)
    - Jobs read from/write to disk endpoint only
- Introduce additional transfer links in the transfer system
  - Connect new Tier-1 disk endpoints to other disk endpoints and tape endpoints
- Verification of functionality by test workflows
- Some recent tape staging tests:



Site	Expected Rate (MB/s)	Achieved Rate (MB/s)
FNAL	650	~900
CNAF	210	~630
JINR*	150	*
KIT	150	~200
RAL	135	~700
IN2P3	135	~650
PIC	75	~500

* Tape at JINR to be commissioned

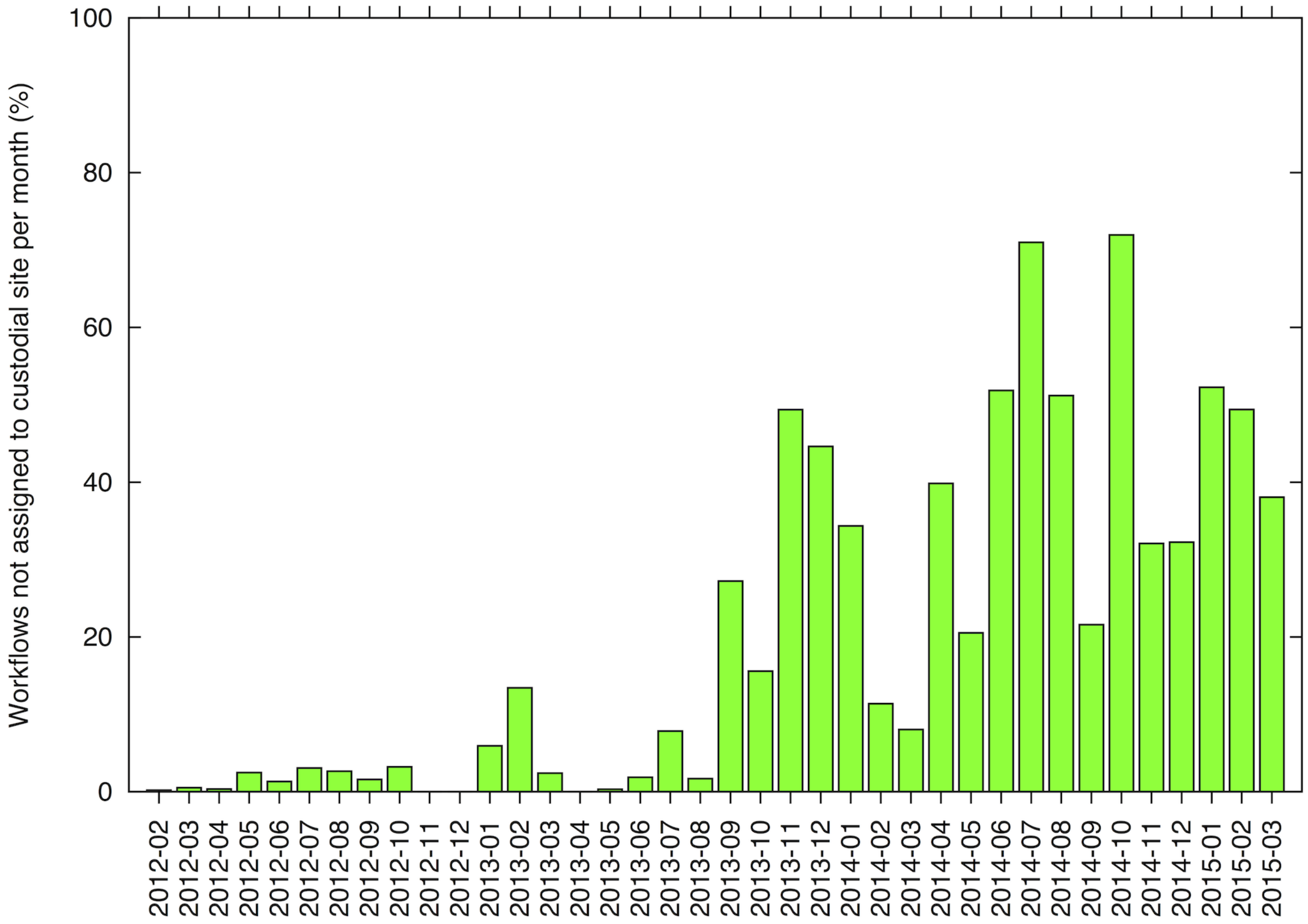
All tape rates well above needs

Big Gain in Flexibility

- Processing can start immediately
  - No need to wait for creation of tape families at archival site
- Workload can run at any Tier-1 site
  - No restriction to run at archiving Tier-1 location
- Subscription to tape can be delayed
  - Allows for check of results
  - Cleaning garbage from disk much easier than from tape
- All files on disk endpoint get published through AAA data federation
  - Allows for remote access
  - Fraction of data processing can run without local subscription
- Tier-1 sites can be opened for analysis jobs
  - Jobs can only access files on disk endpoint

Example: Flexibility in DIGI-RECO Workflow Assignment

- 50% or more get assigned to other site than archiving (=custodial) site after separation of disk and tape resources at Tier-1 sites



Summary

- In Run1 tape resources strictly coupled to local Tier-1 disk resources
  - Restricted assignment of Tier-1 workflows to archiving site
  - Prevented analysis jobs from being run at Tier-1 sites
  - Enforced tape family creation before start of actual processing
- Effort to separate disk and tape resources
  - Run separate storage instances for disk and tape
  - Separation through different trees in the namespace
  - Tape reading/writing becomes a subscription in the data management system
- Big gain in flexibility
  - Restriction from Run1 resolved