# CMS Requirements for Disk/Tape Separation

The functionality of the traditional hierarchical mass storage system was developed when the ratio of disk to tape was much smaller than it is for the LHC experiments. Instead of 10% disk caches deployed for the Tevatron experiments, LHC experiments have between 50% and 80% of the tape capacity deployed as disk at the Tier-1 centers. The large volume of disk and the large size of the active dataset are making a model of allowing the mass storage system to stage as needed on demand inefficient. At all facilities with tape, CMS is requested to perform large-scale external pre-staging operations. There is no way of monitoring, on a large scale, that a file is still on disk. The experiment is asked to manage the use of the site disk space to improve the efficiency, but has few tools to do this.

Currently when files are written to storage at Tier-1s they are also written to tape. This introduces a variety of limitations. For CMS to replicate RAW data between Tier-1s to improve the completion time of processing a dataset involves writing a second copy to tape, which then fills the archive and takes months to recover the media when the replica is no longer needed. Additionally, files written to Tier-1 sites before the dataset is completely validated are archived and if there is a problem with the sample and it needs to be deleted the media can again take months to be recovered.

To solve these two problems CMS proposes to use the data management system that has been successfully used for wide area data replication to also manage the replication of data between disk and tape. In this model the sites for CMS will deploy two PhEDEx endpoints: one disk and one archive. The disk end point should write only to disk and the tape end point would write to tape through a much smaller disk buffer. The primary use-cases are outlined below.

1.) A PhEDEx transfer from T1_XX_Archive to T1_XX_Disk is the equivalent of a prestage and indefinite pin. This will be done on the dataset and datablock level, not the file level.

2.) A PhEDEx transfer from T1_XX_Disk to T1_XX_Archive is a transfer to tape archive. A transfer from T2_XX to T1_XX_Archive is an indication the dataset is validated and ready for archive. This is used to transfer data from a production site to the Tier-1 archive directly when it’s not expected to be accessed at a Tier-1 by a processing workflow. This will also be done on the dataset and datablock level, not the file level, and may serve as helpful information in layout planning for the mass storage system.

3.) A deletion of data from T1_XX_Disk is the equivalent of a pin release. Disk space will also be released in dataset and data block chunks.

Workflows from either production or analysis users will only be submitted to datasets resident on the T1_XX_Disk endpoint, so no CMS workflow should trigger a tape recall at run time. Replicas that are archived elsewhere will be written to the disk endpoint, accessed with a workflow, and then deleted from the disk endpoint. Datasets that are in the process of validation will be written to T1_XX_Disk and then subscribed to T1_XX_Archive to indicate they are ready for archival storage.

In CMS we have referred to this solution as Disk/Tape separation because from a functionality perspective the disk and archive are separated. It is strictly speaking not necessary that the archive and disk be physically close or even provided by the same technology, though in many places they will be. Workflows may be performed and written to T1_YY_Disk and then archived to T1_XX_Archive to balance the use of tape across the collaboration resources.

CMS is asking for this functionality to achieve more flexible use of the resources and better management of the disks at Tier-1s. We also believe that for datasets that are available in multiple locations we can reduce the tape staging rate. A PhEDEx subscription of a dataset to T1_XX_Disk would allow the dataset to be pulled from multiple locations and an alternative disk location could be given preferential weight over a tape location, so pre-staging a sample could actually mean transferring over the WAN from a remote disk.

The technical implementation is largely the choice of the site because the situation varies at each. CERN accomplished the functionality by introducing a separate technology for the disk storage and the archive. Transfers between the two are actually file transfers managed by PhEDEx subscriptions. RAL accomplished this with different pools all using Castor. It is probably possible to perform this even within the existing mass storage technology through the creative use of storage classes and a custom PhEDEx agent. CMS is happy to help in any way that facilitates the transition. CMS would like to have all 7 Tier-1 sites able to support the use-cases outlined above by the end of the summer 2013.