Evolving WLCG data access strategy and storage management strategies

May 25, 2011

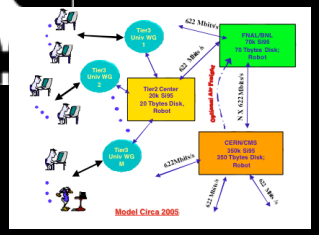
Ian Fisk

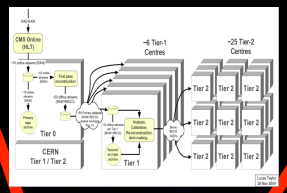
The Large Hadron Collider

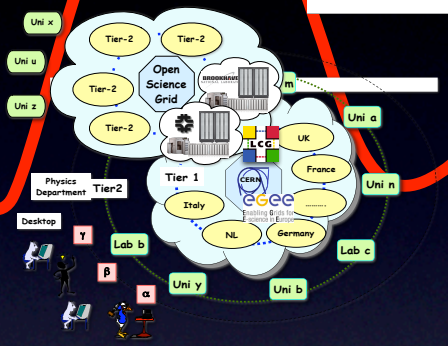
Energy frontier, high Luminosity *p*-*p*-collider at CERN, Geneva, Switzerland

WLCG is the Computing Infrastructure to support the scientific discovery

Evolution







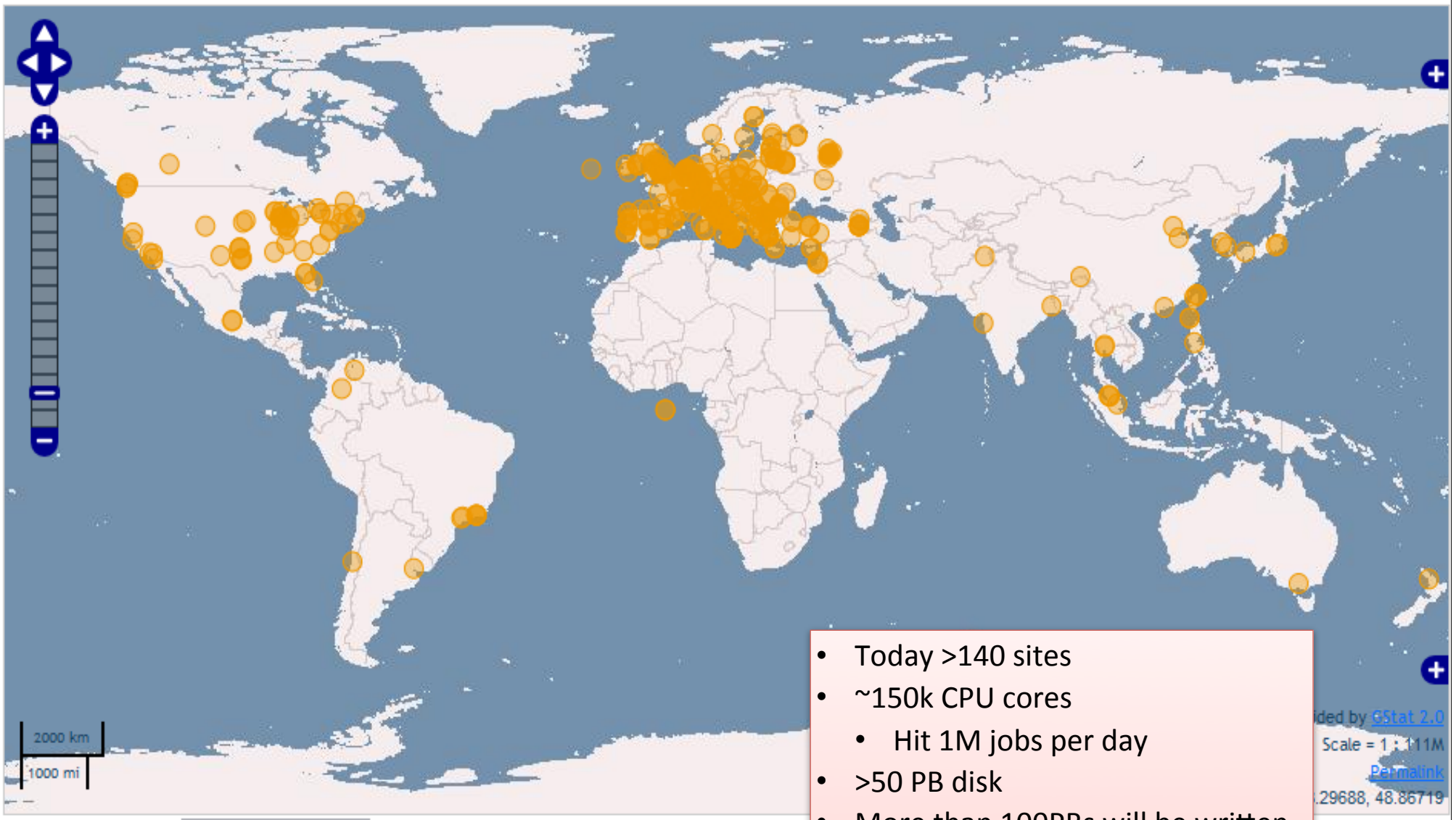
CMS Full Mesh

ALICE Remote Access

PD2P/ Popularity

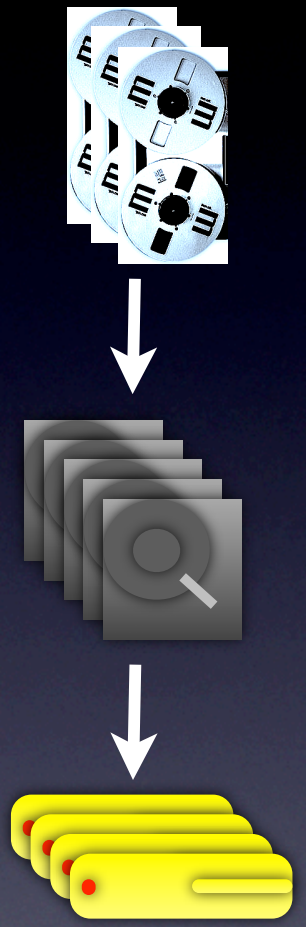
- Over the development the evolution of the World Wide LHC Computing Production grid has oscillated between structure and flexibility
  - Driven by capabilities of the infrastructure and the needs of the experiments

WLCG Today



- Today >140 sites
- ~150k CPU cores
  - Hit 1M jobs per day
- >50 PB disk
- More than 100PBs will be written into the archive by the end of 2011

Traditional Storage



- Typically High Energy Physics has relied heavily on hierarchical mass storage
  - Large datasets only a portion on disk
  - Organized Data access concentrated in centers

Changes of Scale

- Decreases in the cost of disk and technology to run big disk farms
  - LHC is no longer talking about 10% disk caches

	ALICE	ATLAS	CMS	LHCb
T0 Disk (TB)	6100	7000	4500	1500
T0 Tape (TB)	6800	12200	21600	2500
T1 Disk (TB)	7900	24800	19500	3500
T1 Tape (TB)	13100	30100	52400	3470
T2 Disk (TB)	6600	37600	19900	20
Disk Total (TB)	20600	69400	43900	5020
Tape Total (TB)	19900	42300	74000	5970

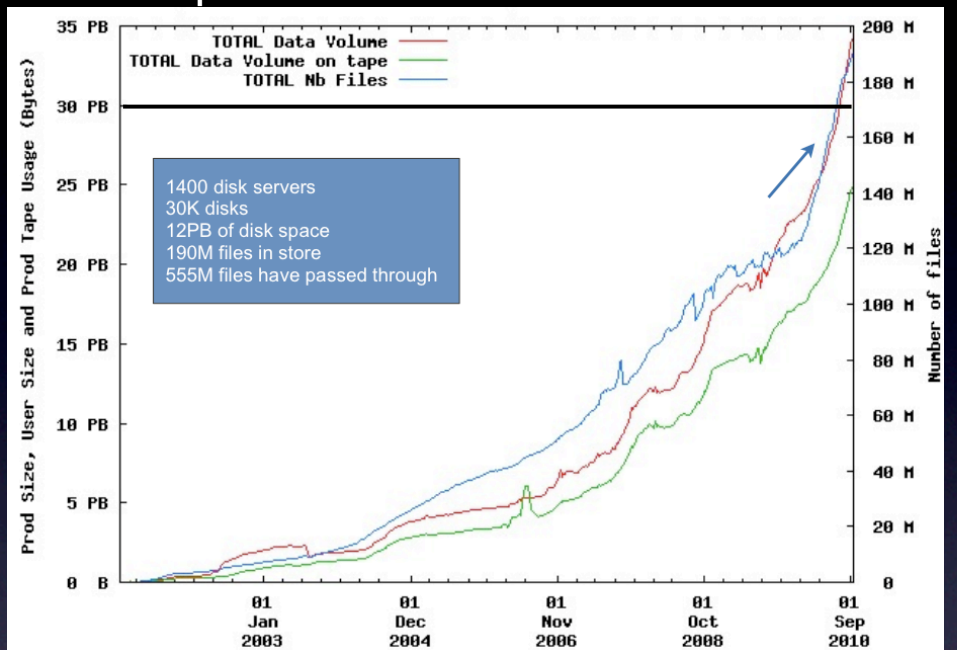
DZero	CDF
~500	~500
5900	6600

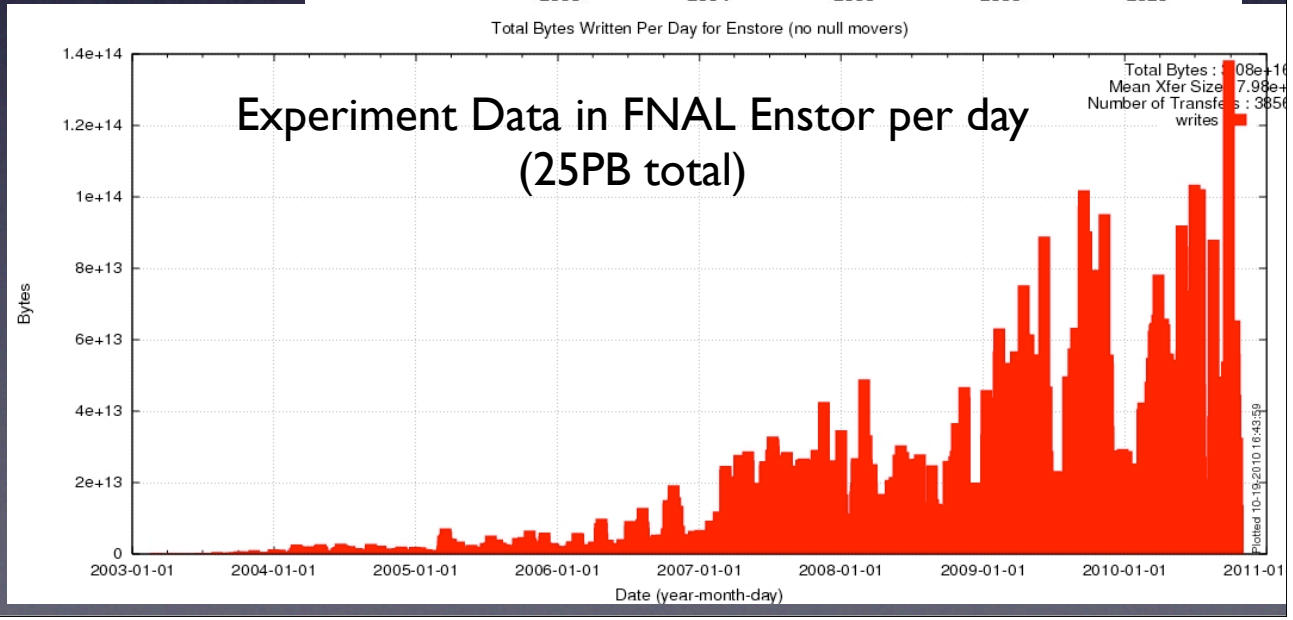
  - In 2011 majority of the currently accessed data could be disk resident

Changes of Scale

- Challenge is growing volume of data that needs to be archived
  - Continued investment by industry in higher capacity tapes
    - 5TB per tape possible
    - 50PB per robot

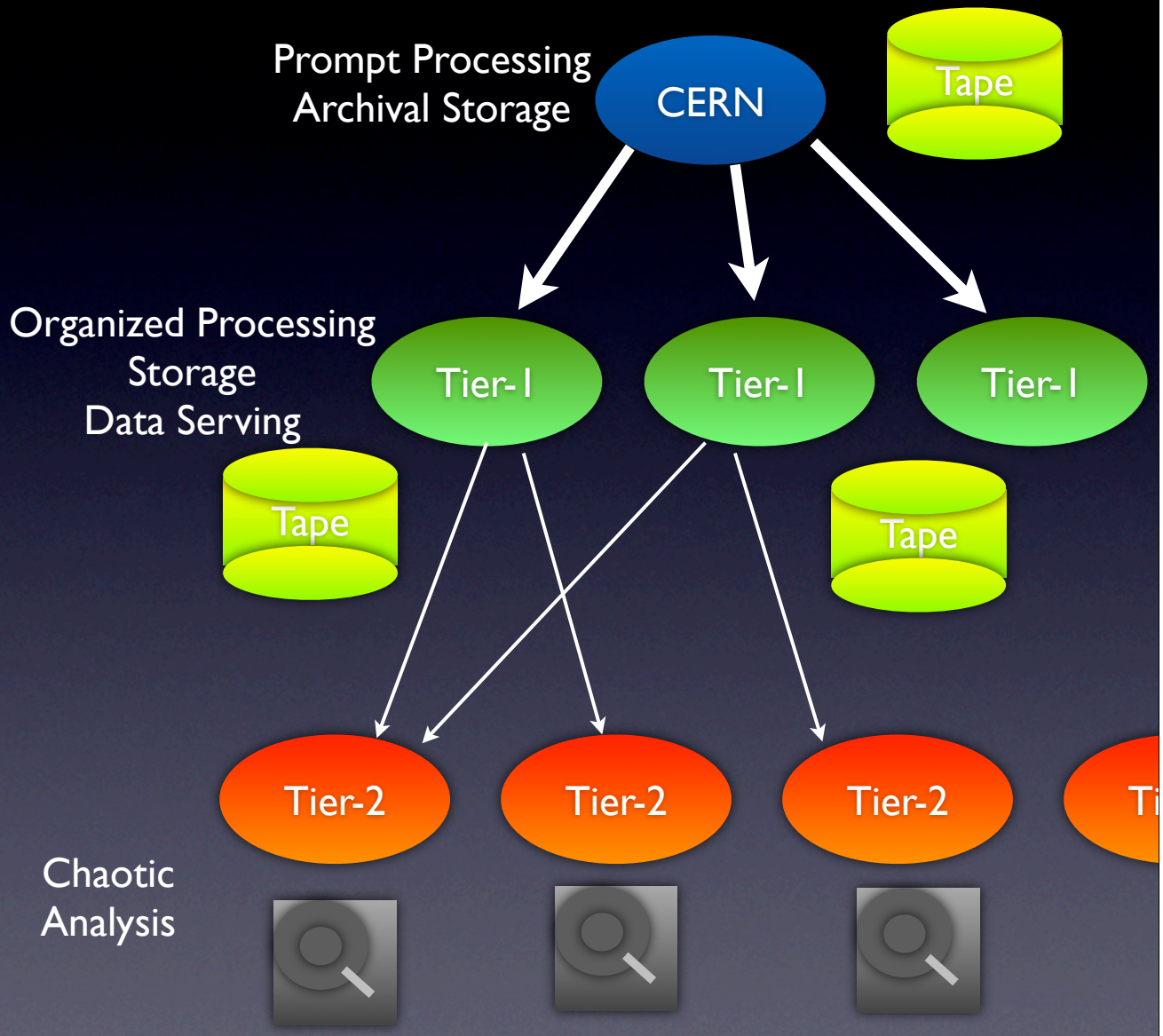
Experiment Data in CERN Castor



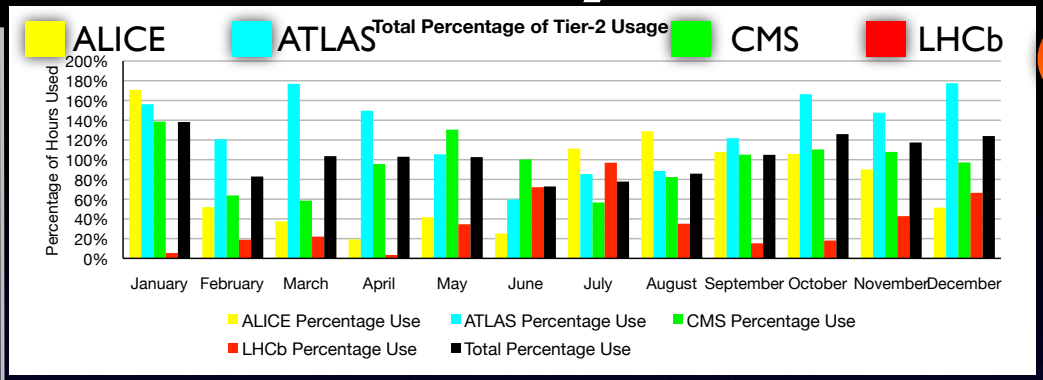


Working Today

- At the LHC most analysis work is conducted far away from the data archives



Analysis Disk Storage



- Disk based storage is heavily used
  - Many of the challenging IO Applications are conducted at centers with exclusively disk
- Tier-2s vary from 10s of TB at the smallest site to 1PB of disk at the larger sites
  - There have been many more options to manage this much space
- In 2011 there will be more than 60PB of T2 Disk in LHC

dCache

DPM

Scalla

hadoop

lustre

GPFS

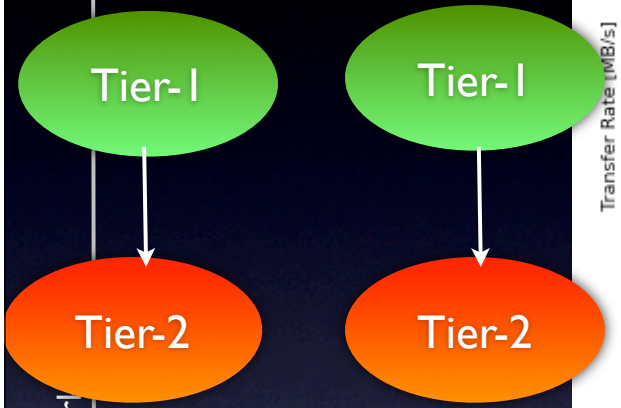
GFS

BLUEARC

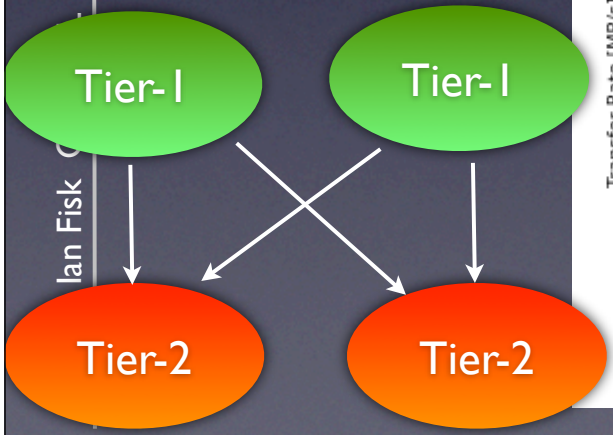
panasas

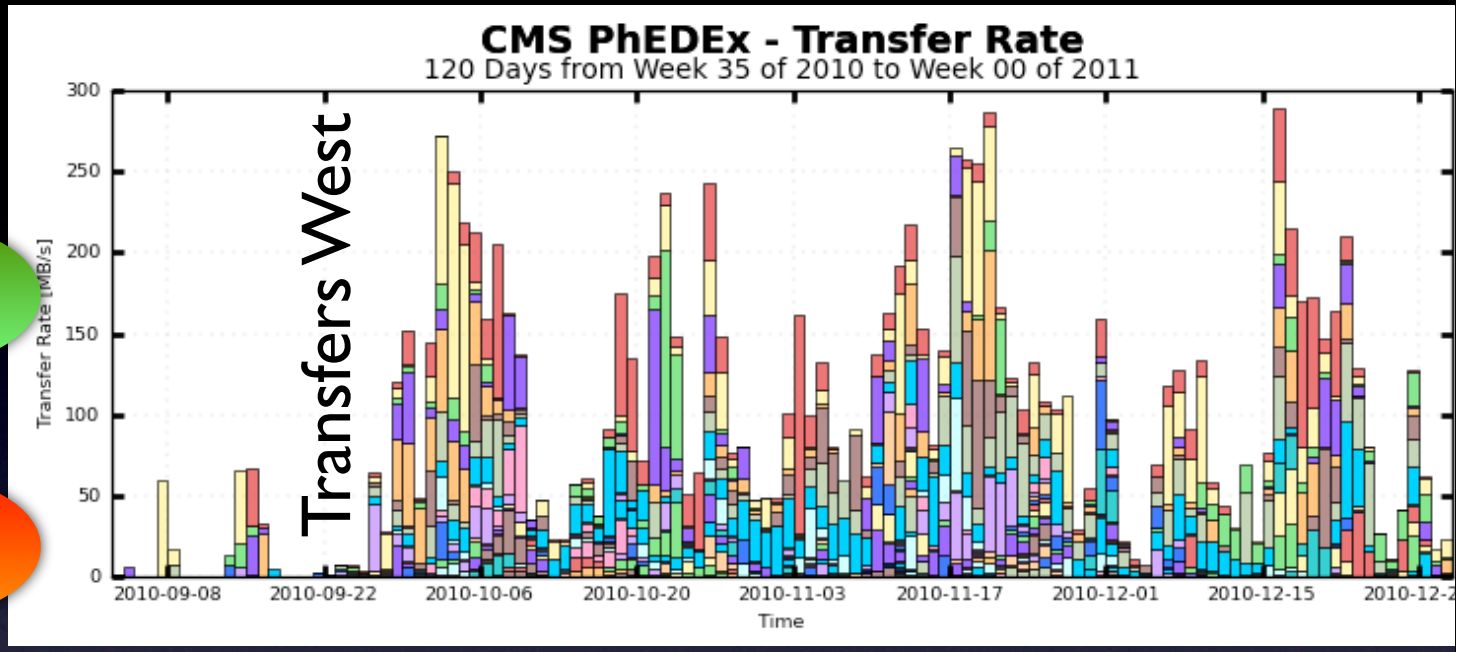
Distribution

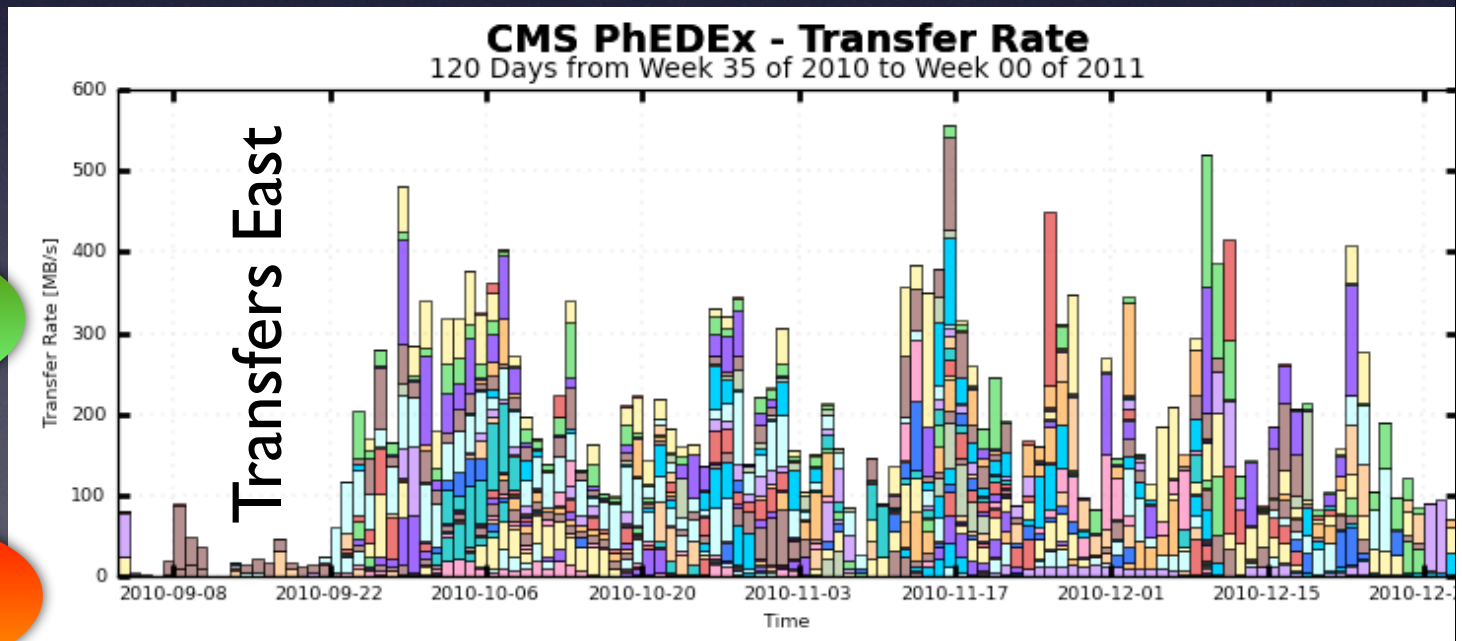
- Change from



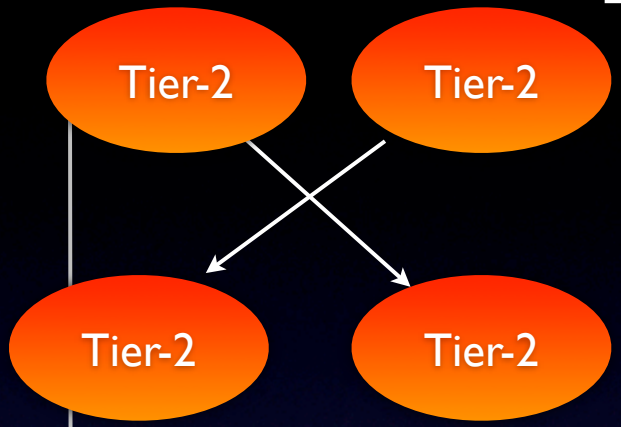
- To

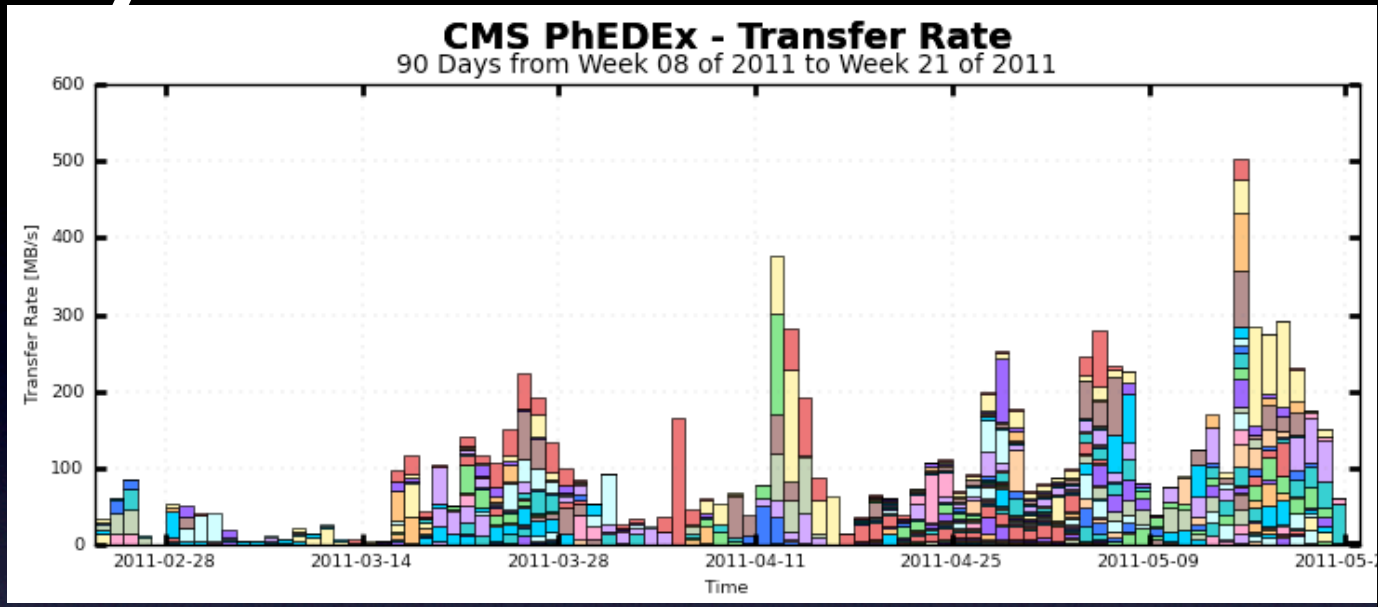




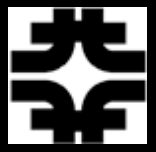


Dynamic Data





- Migration from very structured migration of data to more dynamic data placement
  - Evolved both with user expectations and the capabilities of the network



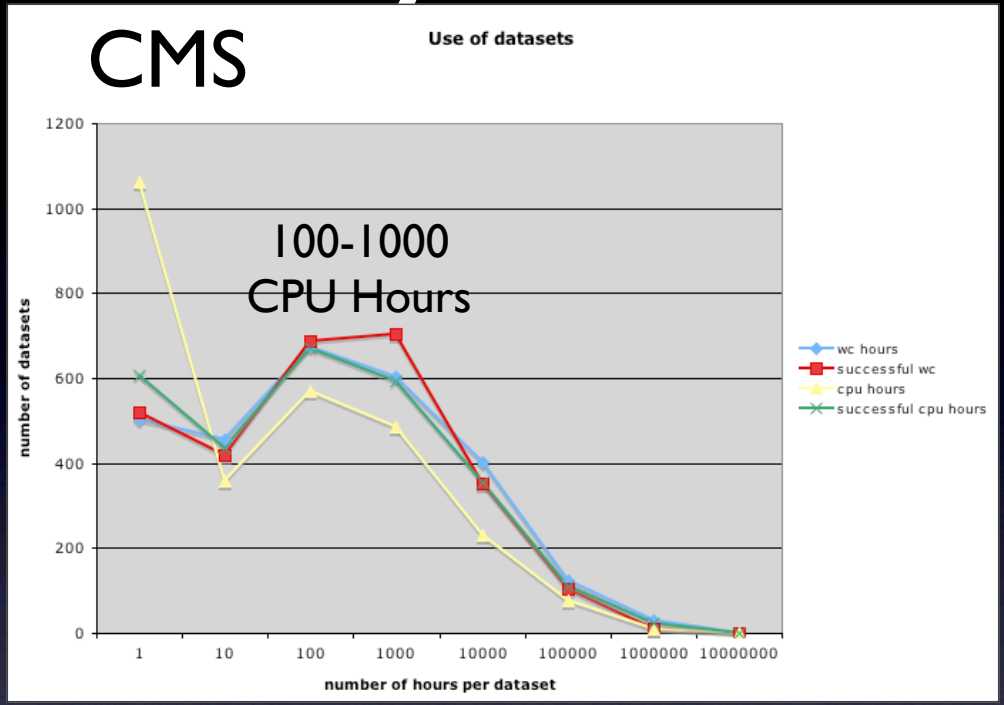
Access

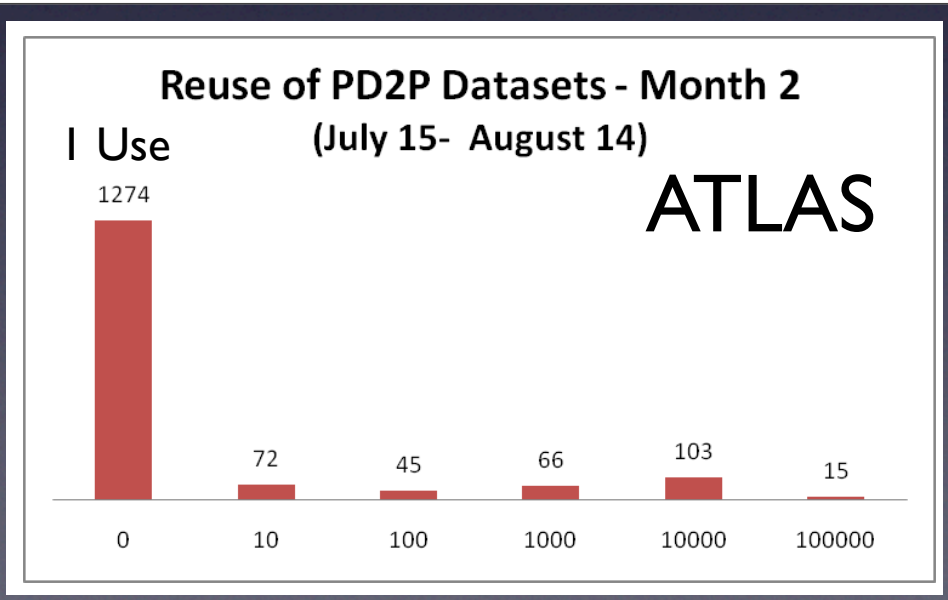
- We have more than 100 analysis sites and more than 2000 individuals a month submitting analysis jobs
  - In CMS running 20k cores for analysis with user jobs 15-30 minutes on average
  - Around 1M file accesses per day
- Trying to track the usage and make the most efficient use of the storage is a technical challenge
- Data Popularity is a joint project of CERN IT Experiment support

Popularity

- Huge variation in the access level to data
  - We will get better at predicting popularity
- We may need different strategies for data used a lot and data used once

CMS



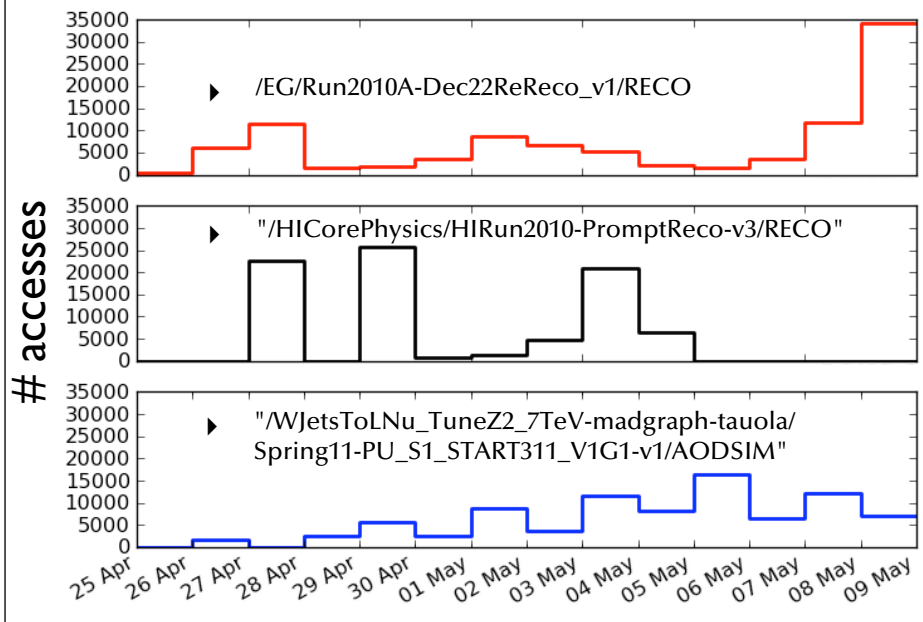


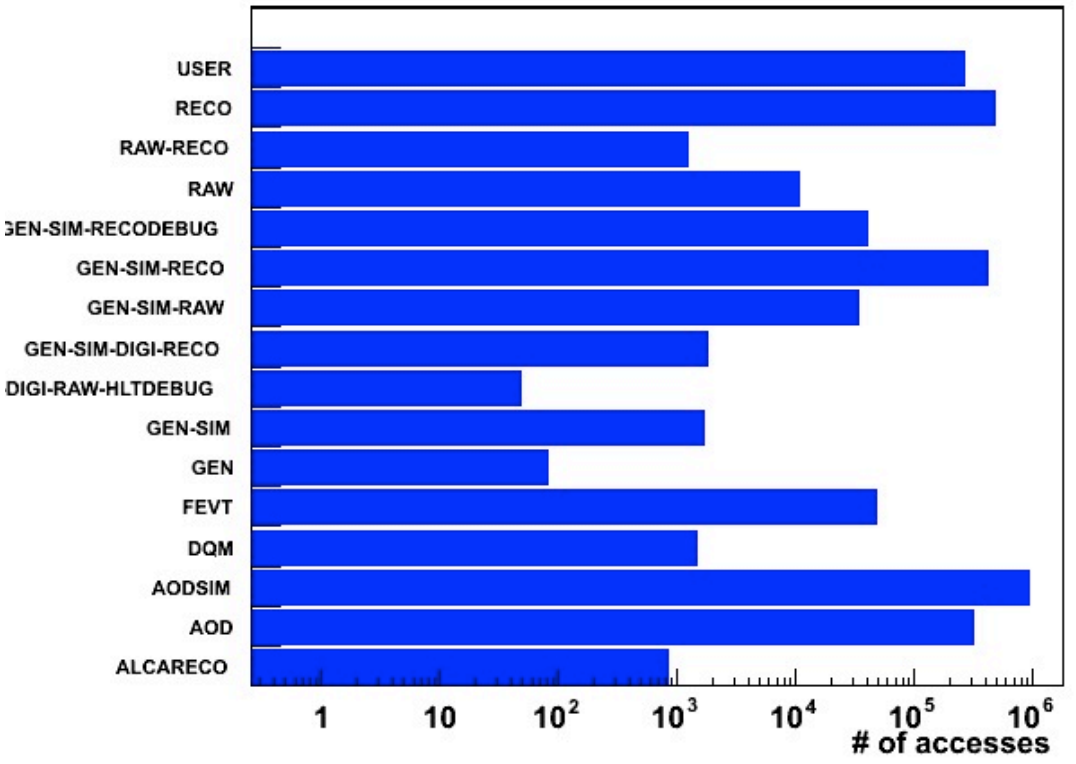
Popularity

- Can now see what data is being accessed most and how they are accessed, what tiers and what users are looking at things

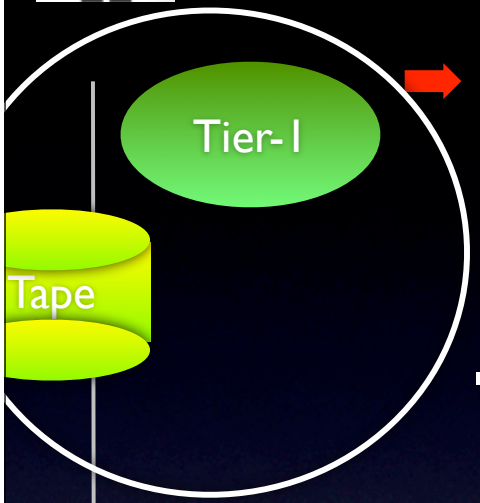
Datasets

- /EG/Run2010A-Dec22ReReco_v1/RECO **(acc 96176 , 8 blocks, 9 users)**
- /HICorePhysics/HIRun2010-PromptReco-v3/RECO **(acc 81180, 70 blocks, 5 users)**
- /WJetsToLNu_TuneZ2_7TeV-madgraph-tauola/Spring11-PU_S1_START311_V1G1-v1/AODSIM **(acc 71051, 4 blocks, 40 users)**

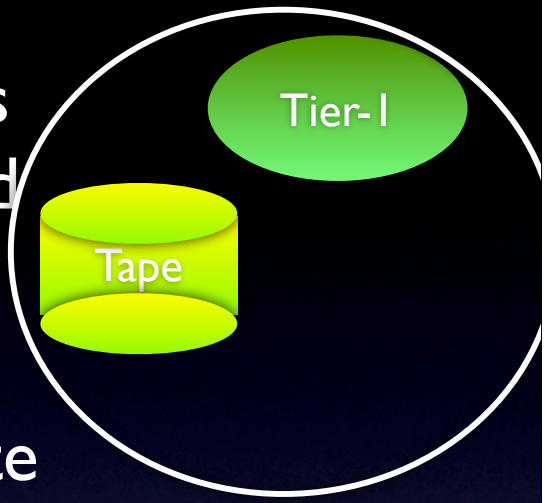




Placement



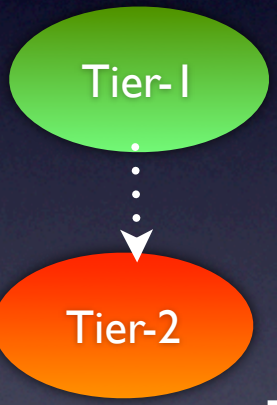
- In an environment that discounts the network the sites are treated independently
  - On the time scale of a job submitted and running on a site it is assumed the local environment cannot be changed
- From a data access perspective in 2010 data available over the network from a disk at a remote site may be closer than data on the local tape installation



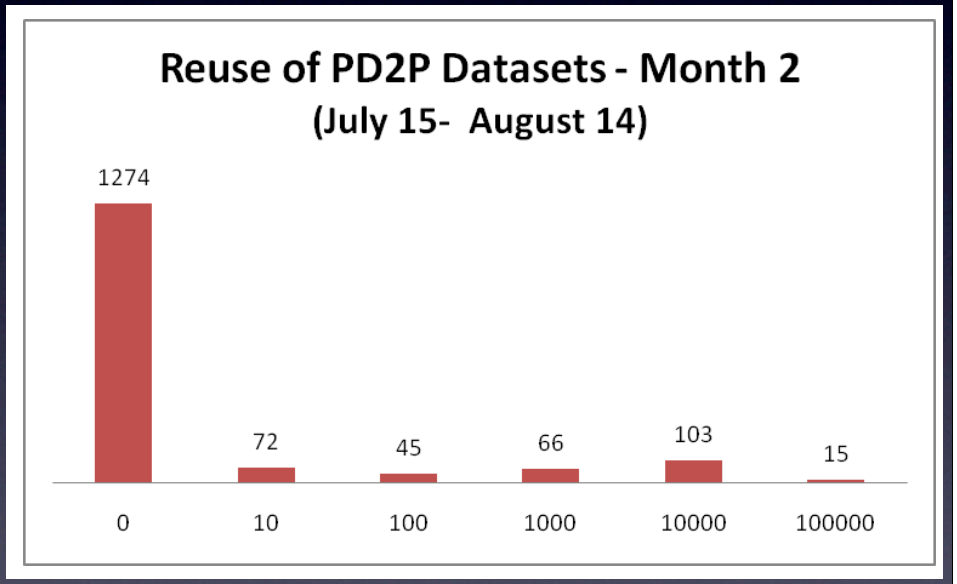
Tier-1 Tier-1 Tier-2

Dynamic Replication

- ATLAS introduced Panda Dynamic Data Placement (PD2P)

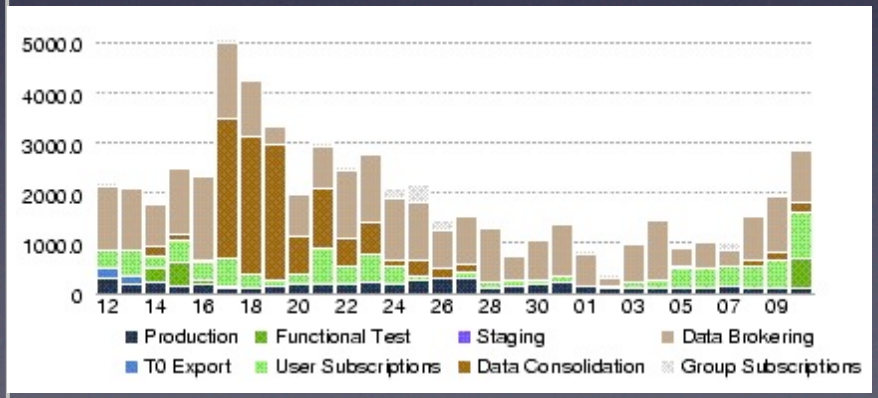


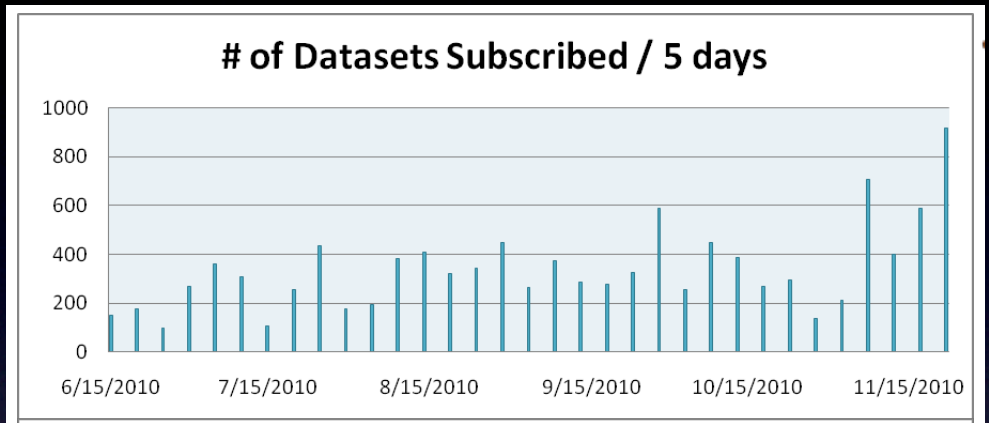
- Jobs are sent to Tier-1 and data replicated to a Tier-2 at submission time

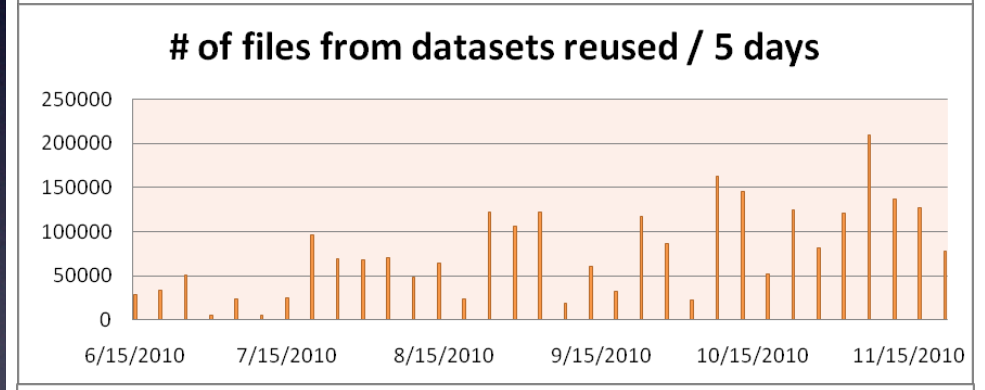


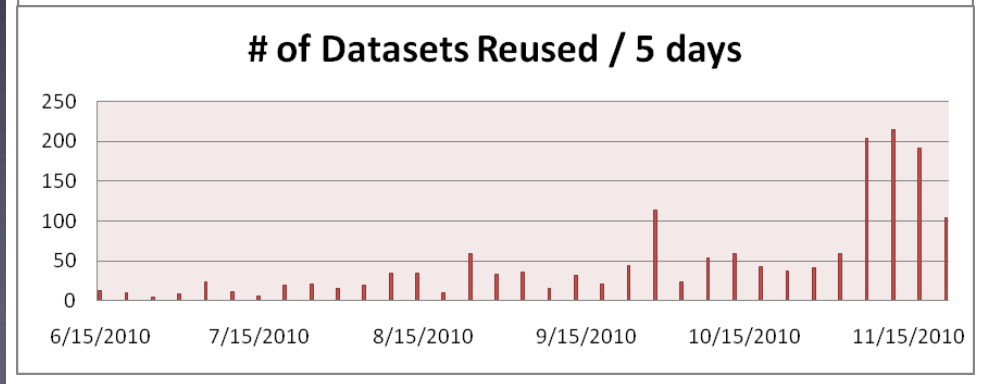
Data Placement and ReUse

- Dynamic placement now accounts for a lot of the networking
- Re-brokering jobs is increasing the reuse of samples and the efficiency

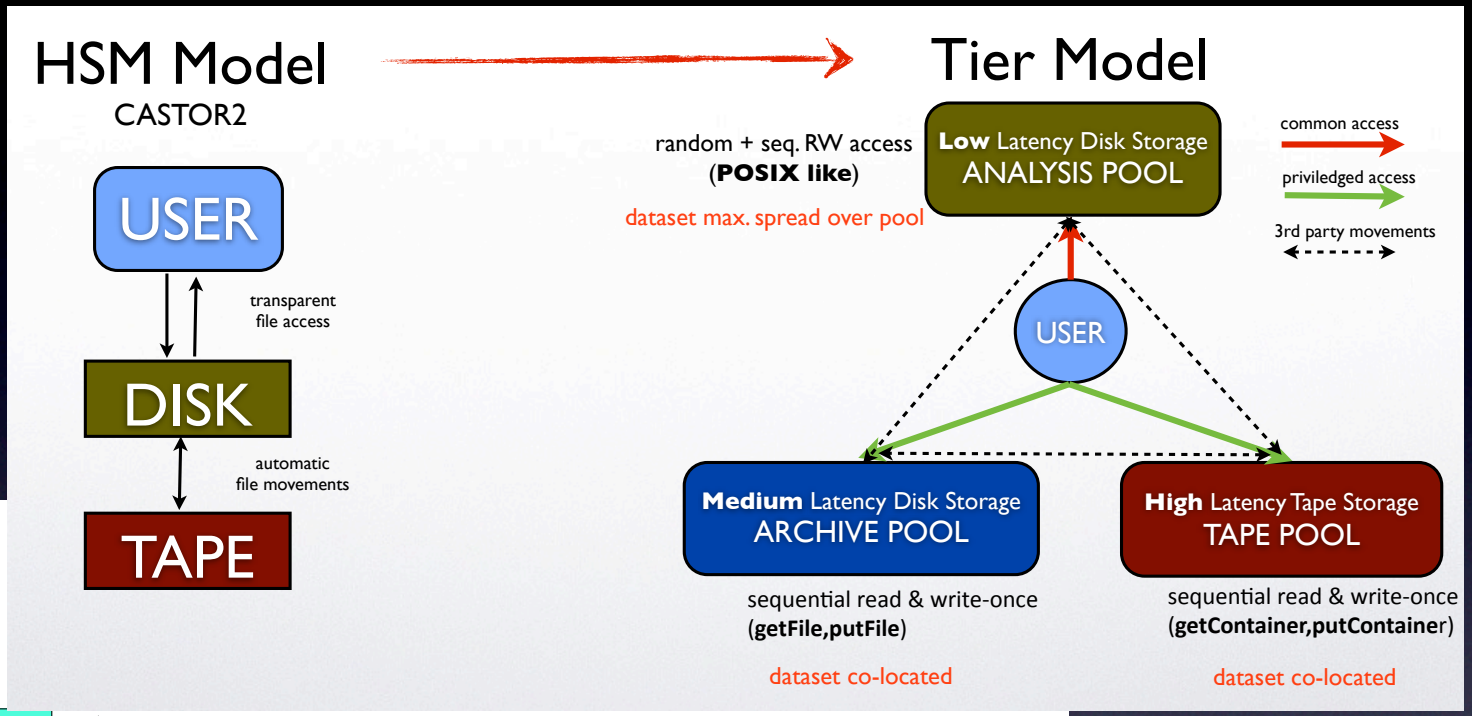


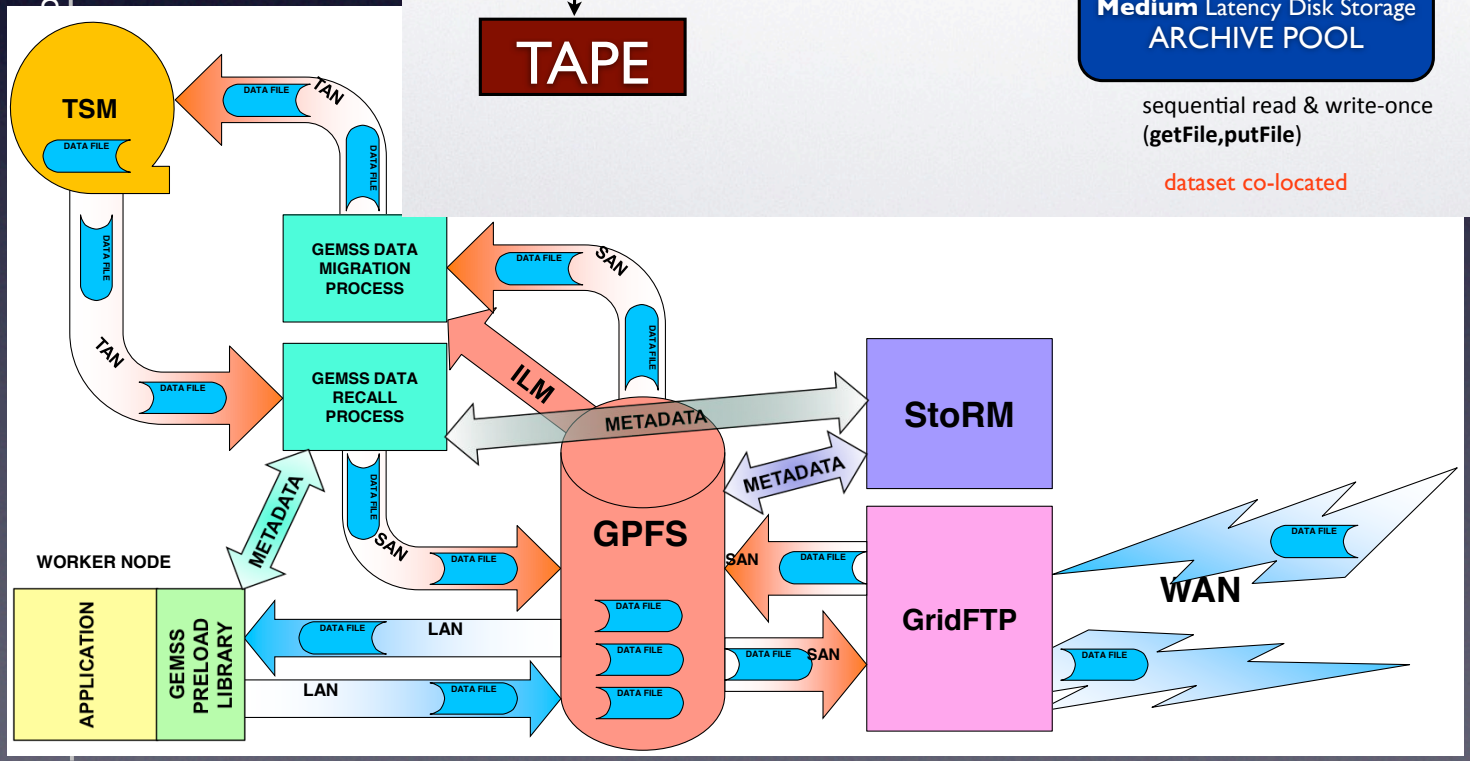


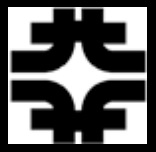




Model Transition

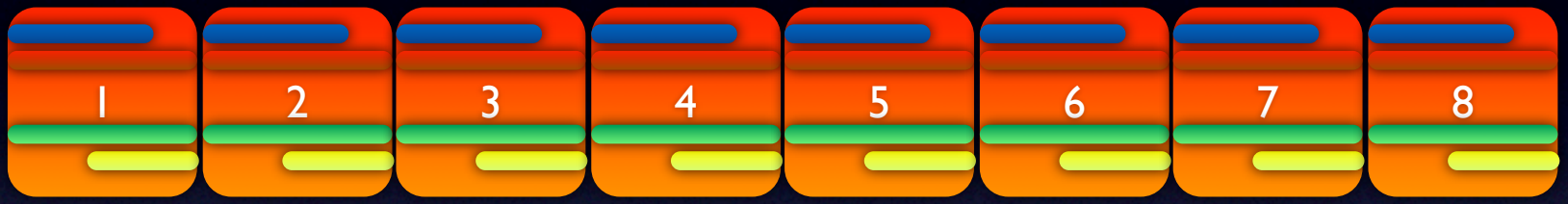






Analysis Data

- We like to think of high energy data as series of embarrassing parallel events

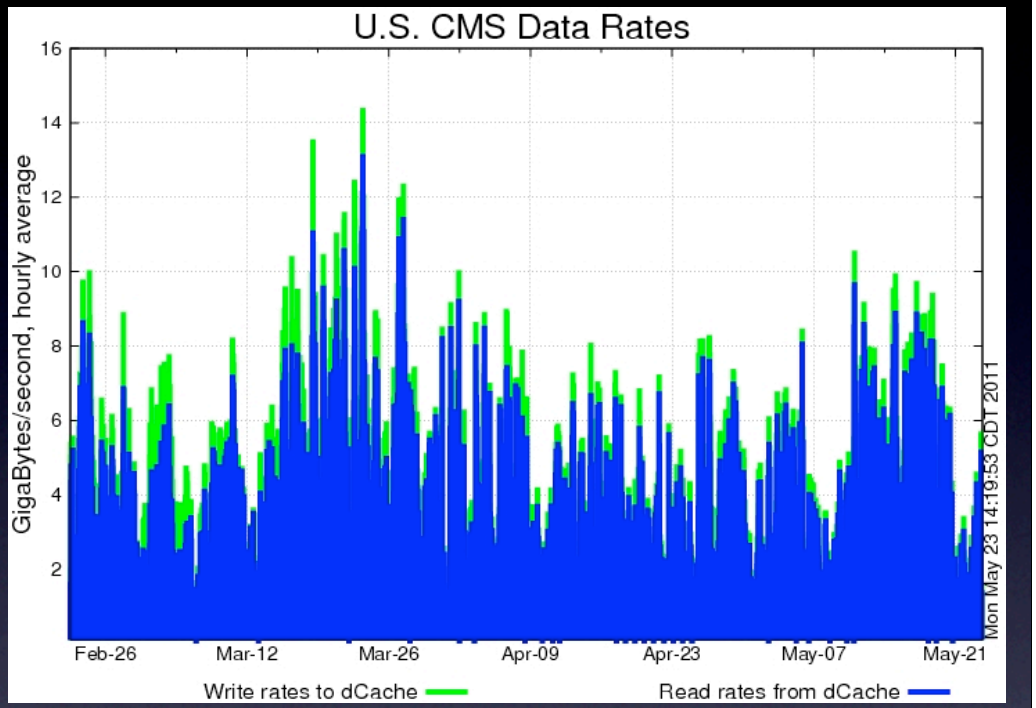


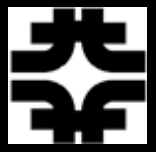
- In reality it's not how we either write or read the files
  - More like

- Big gains in how storage is used by optimizing how events are read and streamed to an application
  - Big improvements from the Root team and application teams in this area

Improvements

- Rates from dCache were 13-14GB/s at FNAL for periods last year
  - Even with more processor cores the rate is lower due to IO improvements
  - Manifests itself with better CPU efficiency and faster applications

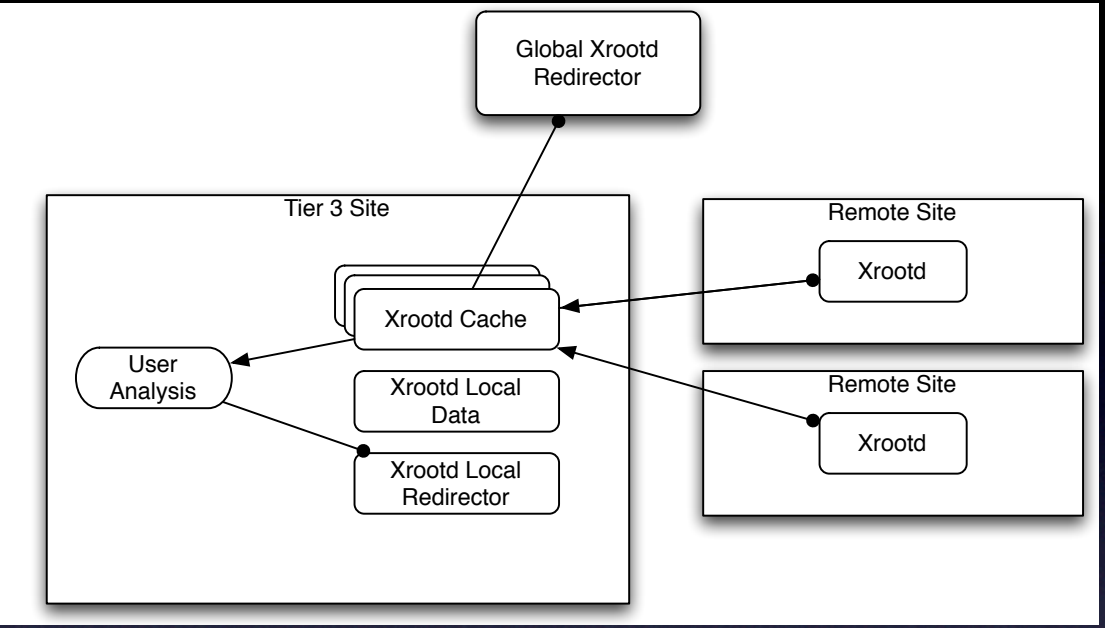




Wide Area Access

- With properly optimized IO other methods of managing the data and the storage are available
  - Sending data directly to applications over the WAN
- Not immediately obvious that this increases the wide area network transfers
  - If a sample is only accessed once, then transferring it before hand or in real time are the same number of bytes sent
  - If we only read a portion of the file, then it might be fewer bytes

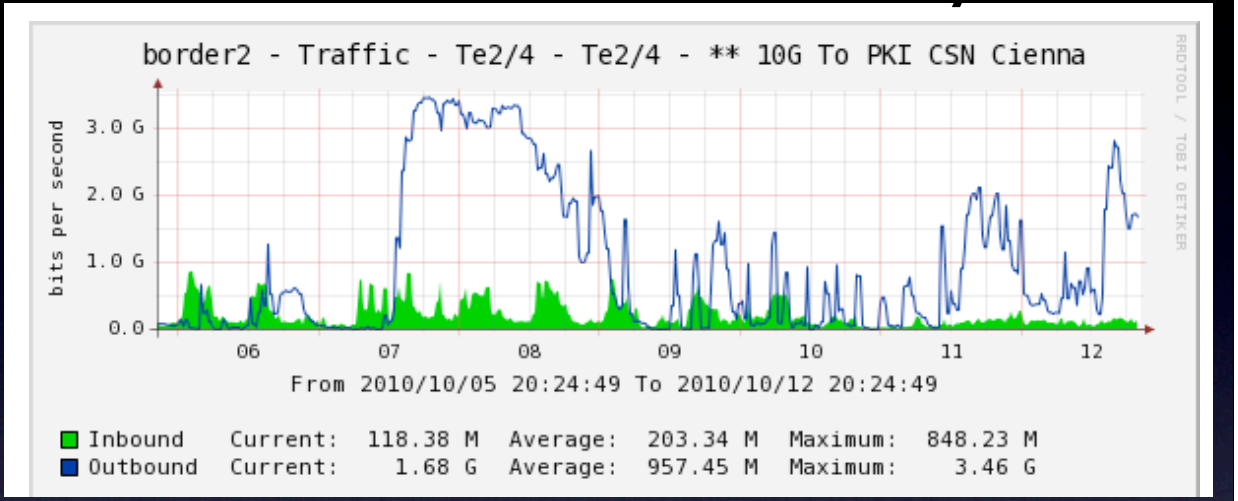
xrootd Demonstrator

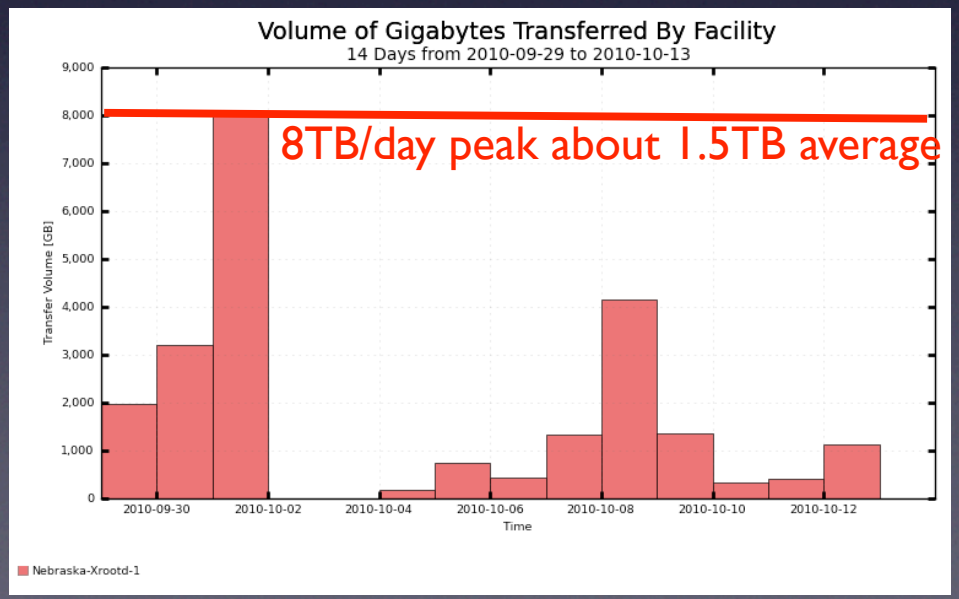


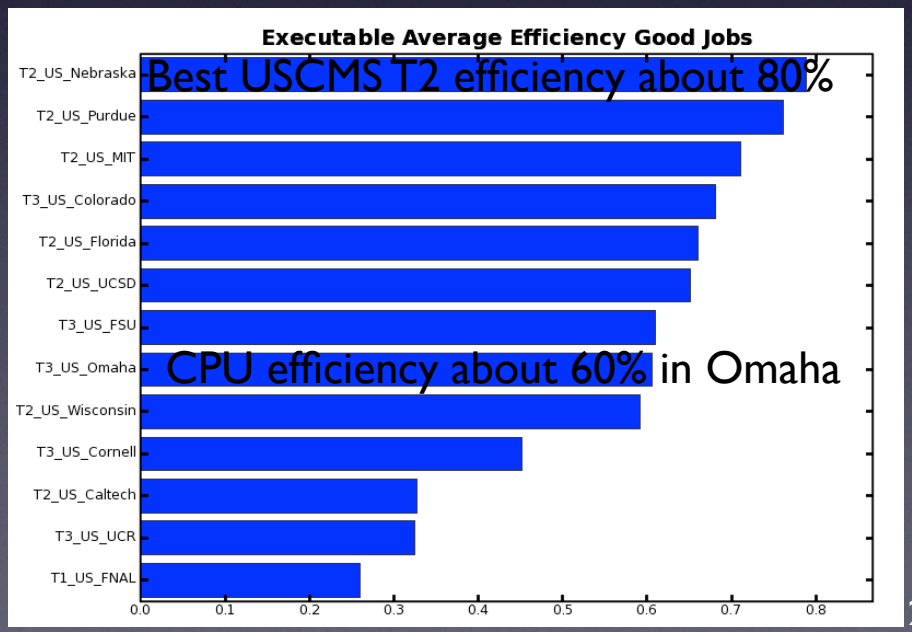
- Current Xrootd demonstrator in CMS is intended to support the university computing
  - Facility in Nebraska and Bari with data served from a variety of locations
  - Tier-3 receiving data runs essentially diskless
- Similar installation being prepared in ATLAS

Performance

- This Tier-3 has a 10Gb/s network
- CPU Efficiency competitive

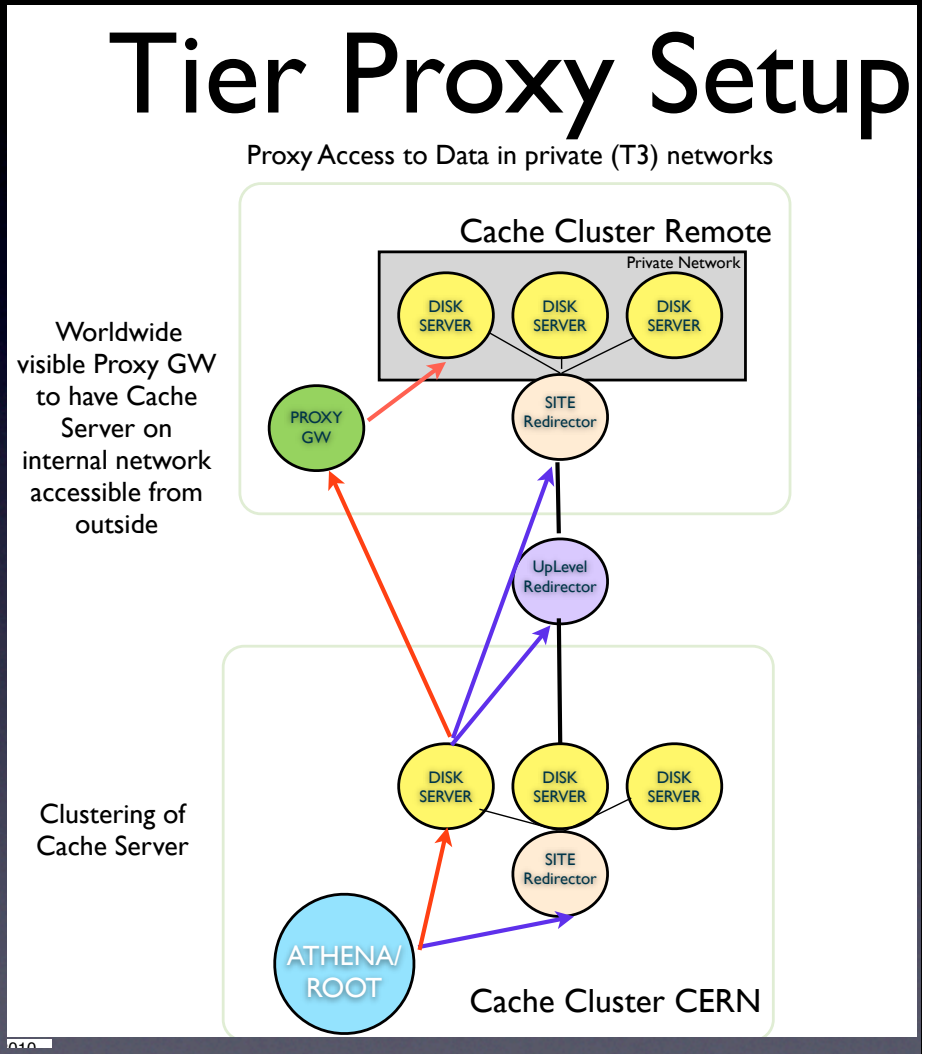






Web Caching

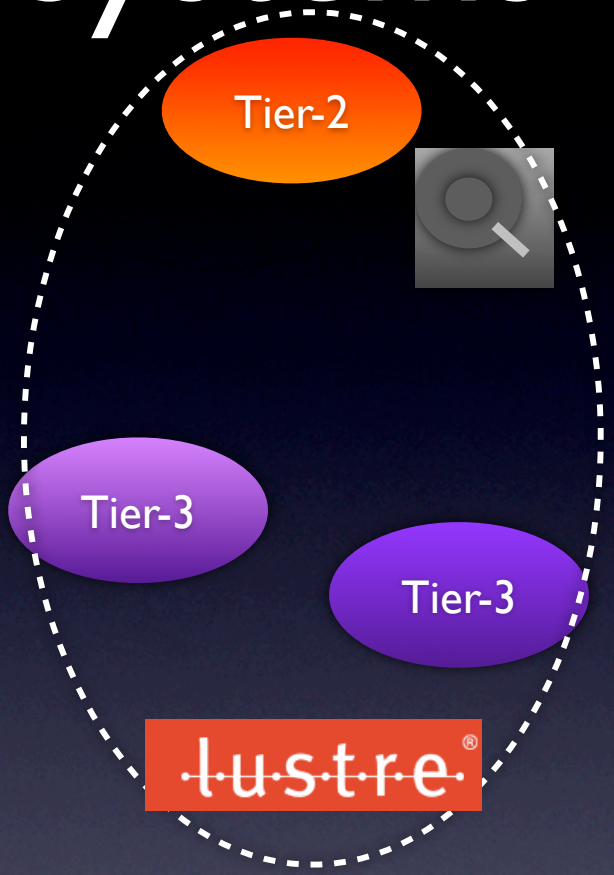
- Interesting proposal to use true data caches to serve the data
  - Can cache entire files, which looks like an automated pull model
  - Can cache portions of a file and queries, which could be helpful for analysis



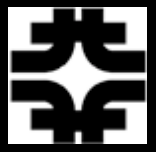
R.Brun, D. Duellmann, G. Ganis,A. Hanushevski, L.Janyst,AJ. Peters, M. Ernst, J. Hover

Wide Area File Systems

- Florida, FNAL, OSG, and TeraGrid have been working with wide area Lustre
  - Successfully demonstrated to serve data to Tier-3s in a geographical area
  - Wide area deployment for TeraGrid
  - Not yet at extremely large disk capacity
- Interesting technique that would simplify data management

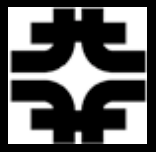


- Very interesting work also being done on NFS4.1
  - Web based redirection system



Future?

- Once you have streams of objects and optimized IO, the analysis application an application like skimming does not look so different from video streaming
  - Read in incoming stream of objects. Once in a while read the entire event
- Web delivery of content in a distributed system is an interesting problem, but one with lots of existing tools
  - Early interest in Content Delivery Networks and other technologies capable of delivering a stream of data to lots of applications



Outlook

- It's impossible to discuss data storage and management in a distributed environment without also talking about networking and access
  - Sites cannot be treated independently. A view of the system and the access is needed.
- Experiments are able to store the data and access it for organized processing and analysis
  - Modifications in the access and management might have big gains in efficiency.