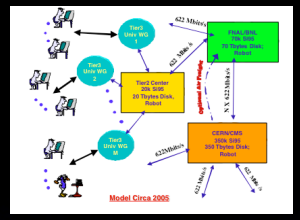
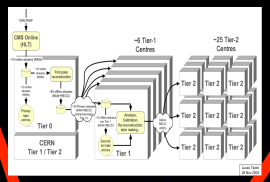
Challenges and Evolution of the LHC Production Grid

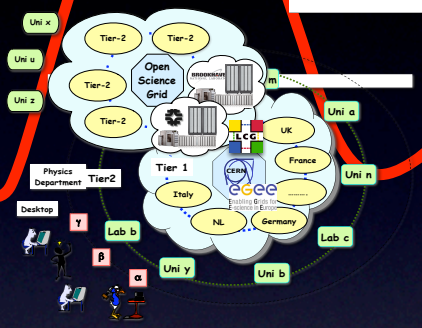
April 13, 2011

Ian Fisk

Evolution







CMS Full Mesh

ALICE Remote Access

PD2P/ Popularity

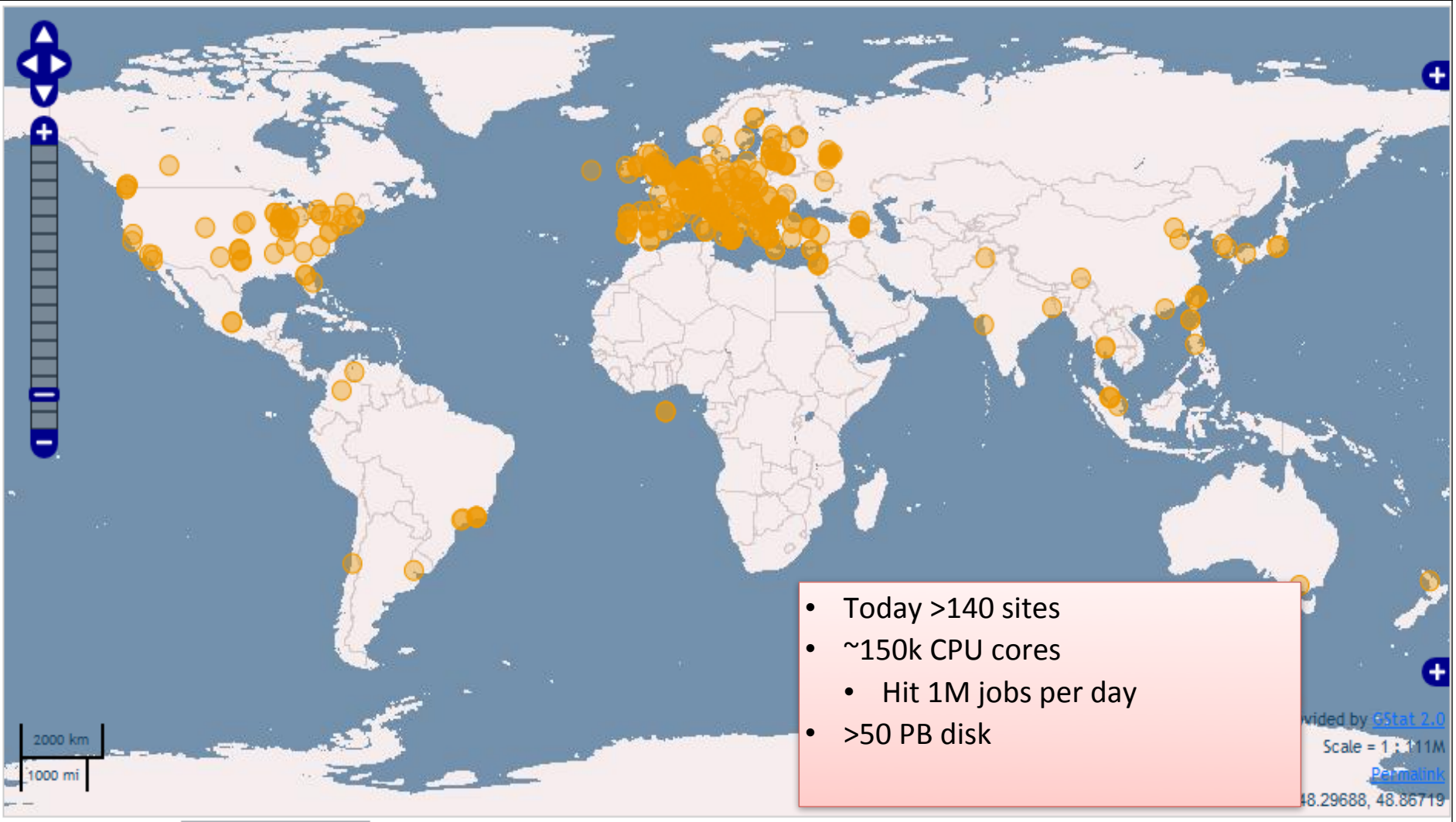
- Over the development the evolution of the WLCG Production grid has oscillated between structure and flexibility
  - Driven by capabilities of the infrastructure and the needs of the experiments

Evolution

- LHC Computing has grown up with Grid development
  - Many previous experiments have achieved distributed computing
  - LHC experiments started with a fully distributed

- LHC Computing Grid was approved by CERN Council Sept. 20 2001
  - First Grid Deployment Board was Oct. 2002
  - LCG was built on services developed in Europe and the US.
  - LCG has collaborated with a number of Grid Projects
- It evolved into the Worldwide LCG (WLCG)
  - EGEE, EGI, NorduGrid, and Open Science Grid
  - Services Support the 4 LHC Experiments


WLCG Today



- Today >140 sites
- ~150k CPU cores
  - Hit 1M jobs per day
- >50 PB disk

Architectures

- To greater and lesser extents LHC Computing model are based on the MONARC model
  - Developed more than a decade ago
  - Foresaw Tiered Computing Facilities to meet the needs of the LHC Experiments

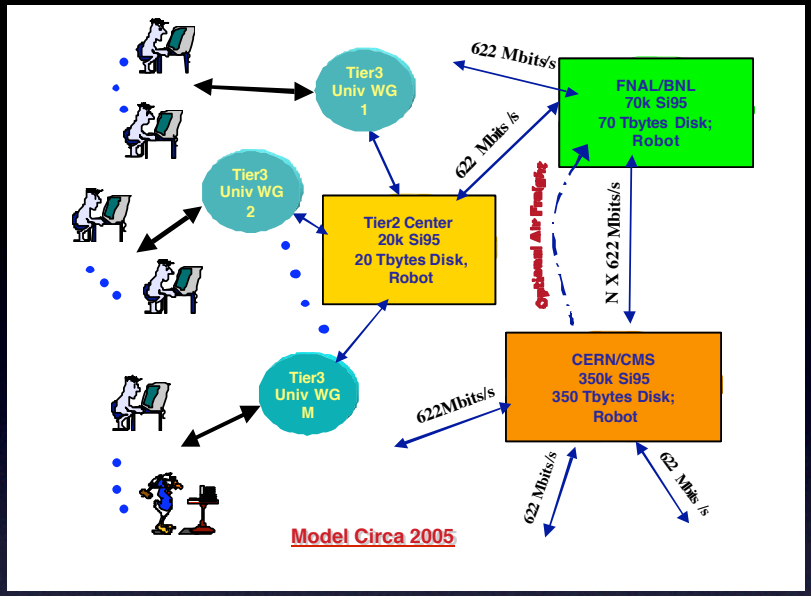
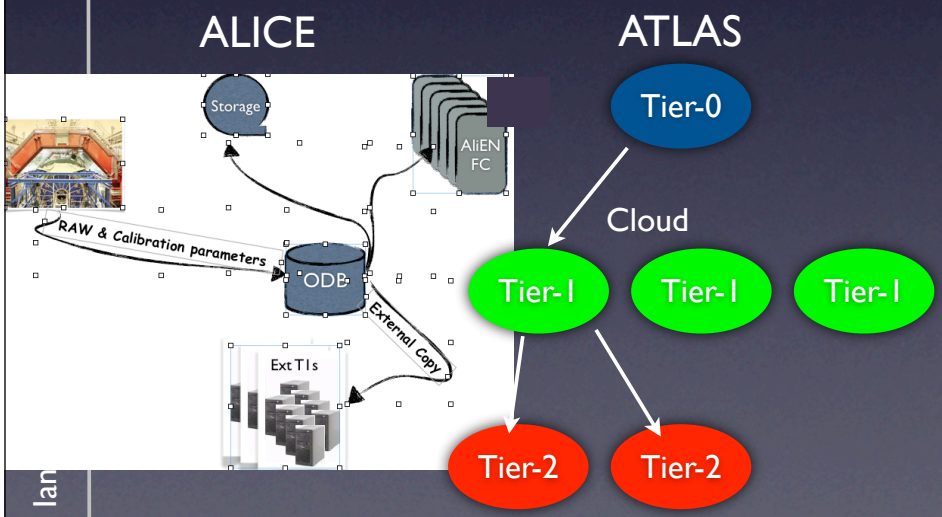
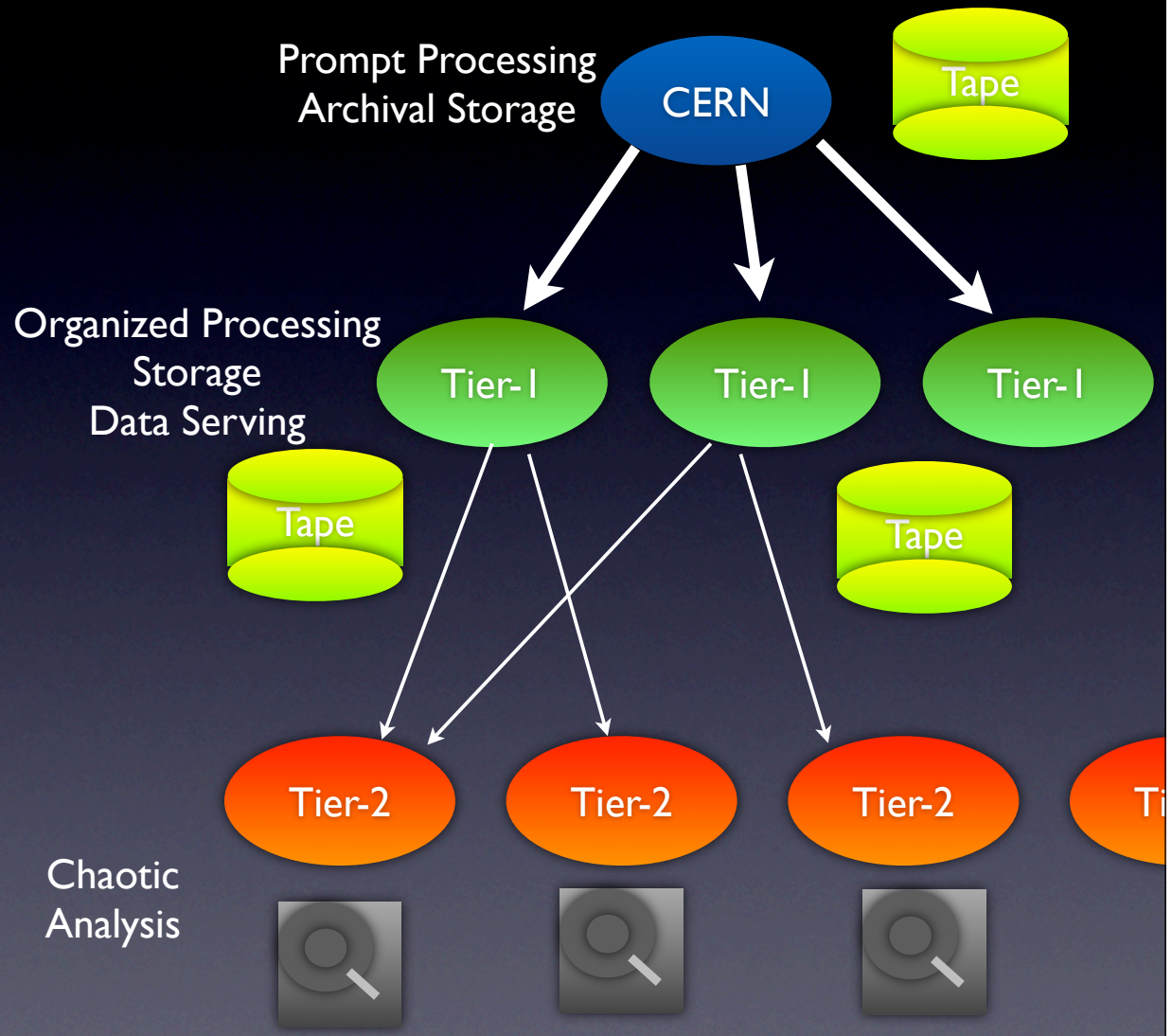


Fig. 4-1 Computing for an LHC Experiment Based on a Hierarchy of Computing Centers. Capacities for CPU and disk are representative and are provided to give an approximate scale).



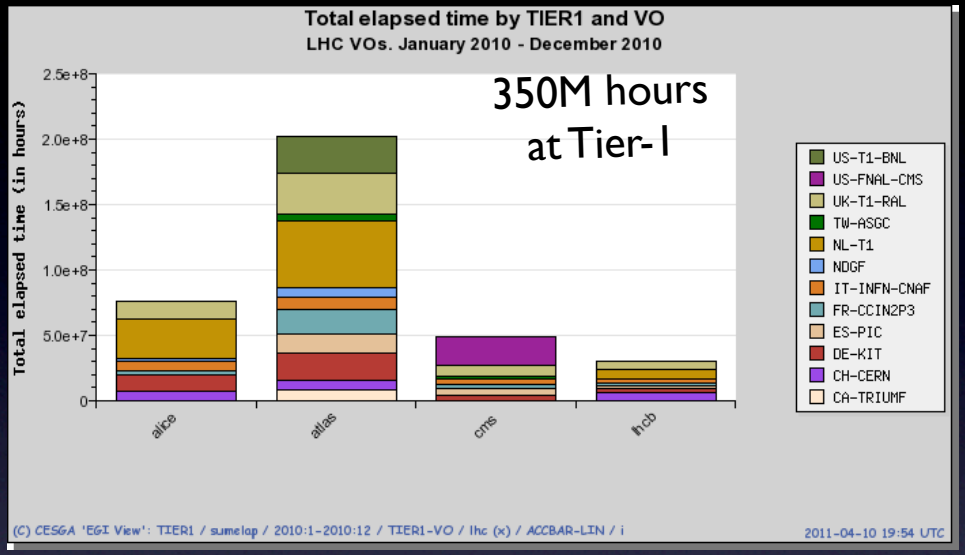
Working Today

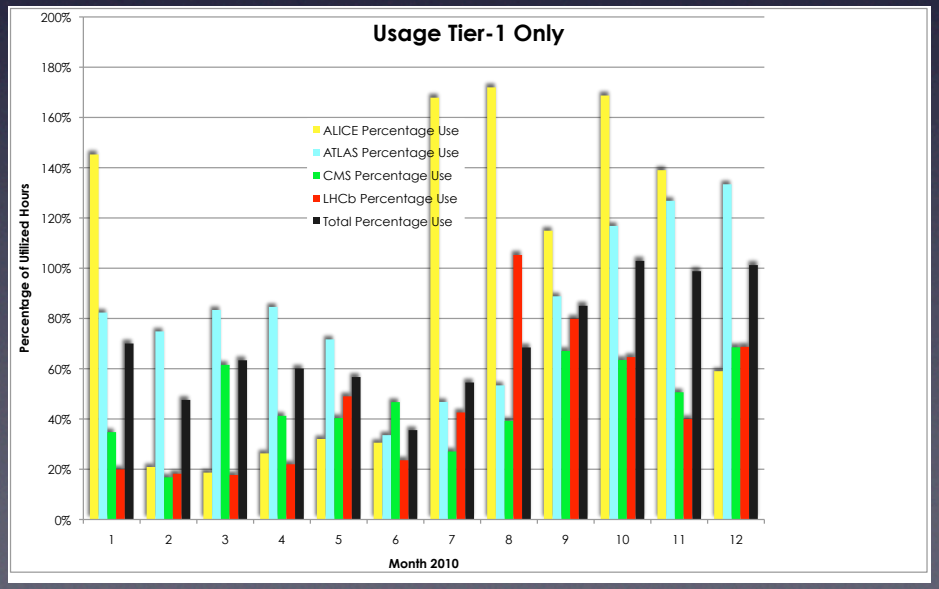
- At the LHC most analysis work is conducted far away from the data archives and storage is widely distributed



Processing Scale

- 2010 was the first full year of running
  - Adding Tier-1 and Tier-2 computing time LHC used roughly 80 CPU millennia in 2010





Scale of Storage

- 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

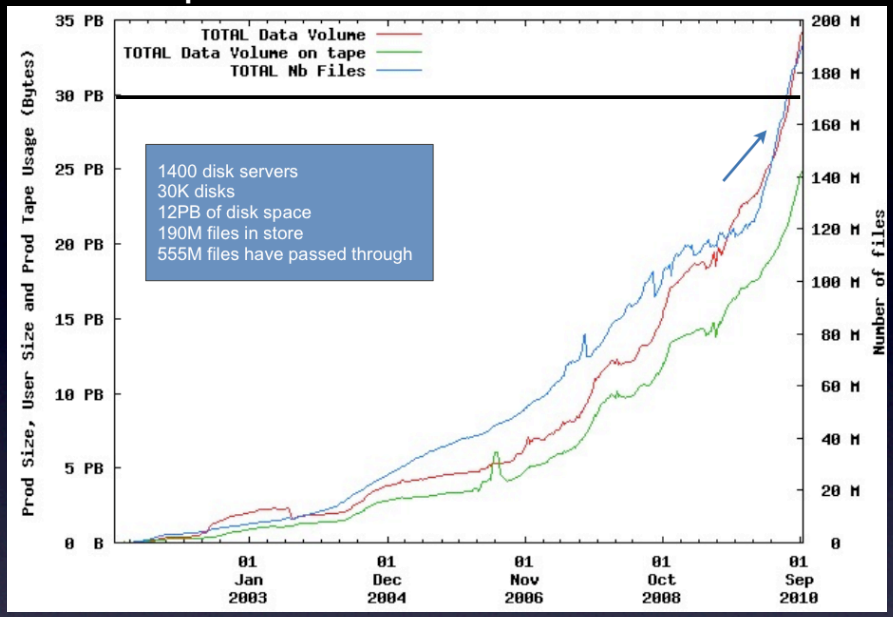
DZero	CDF
~500	~500
5900	6600

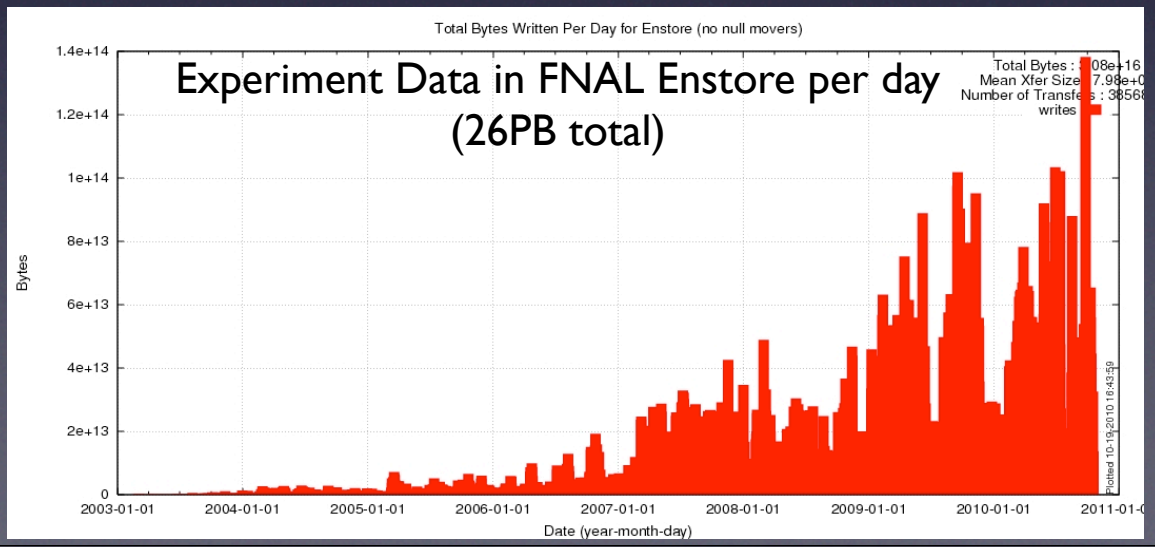
  - In 2011 majority of the currently accessed data could be disk resident

Scale of Archival Storage

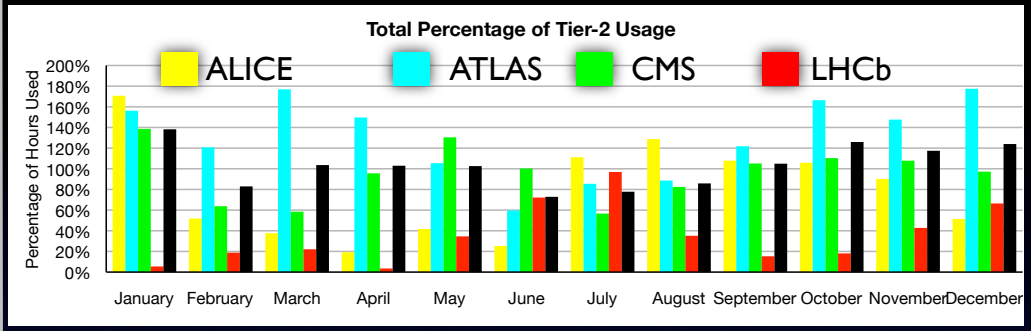
- Challenge is growing volume of data that is produced
  - With the current technology evolution CERN will have robotic capacity for half an exabyte

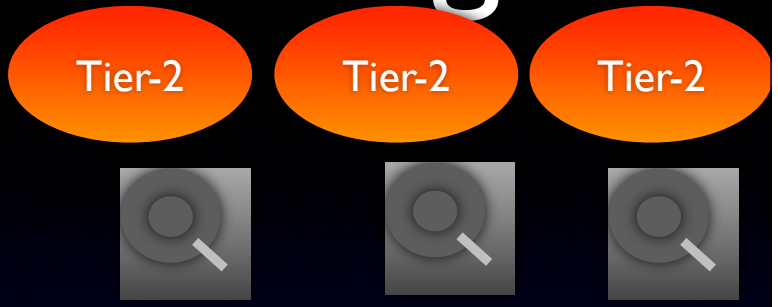
Experiment Data in CERN Castor





Analysis Disk Storage





- Example from LHC
  - Tier-2s are very heavily utilized
  - 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 are more than 60PB of T2 Disk in LHC

dCache

DPM

Scalla

hadoop

lustre

GPFS

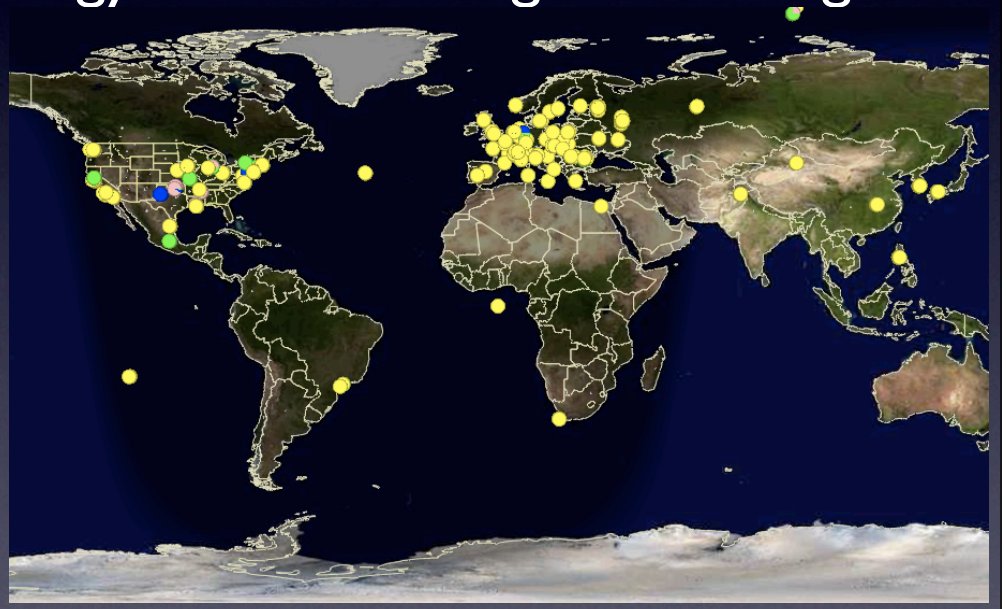
GFS

BLUEARC

panasas

Evolving Challenge - Data Management

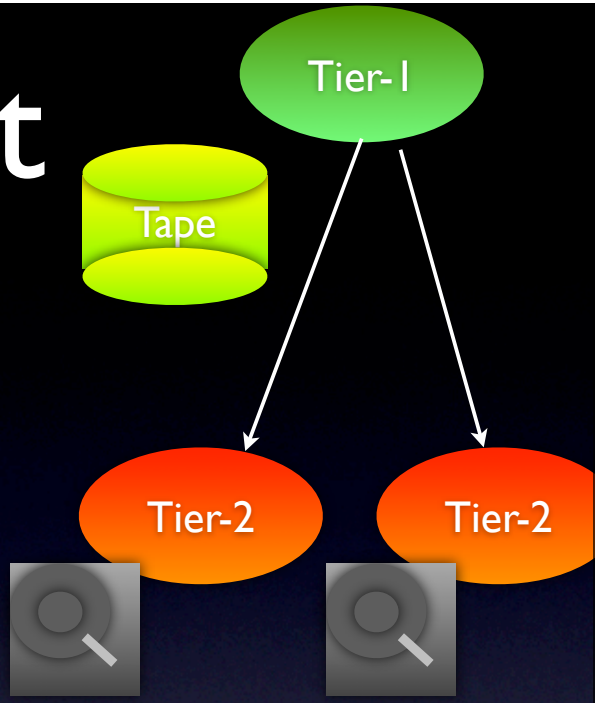
- Data is backed up on tape. Organized processing centers have substantial disk caches
- Analysis centers have large disk resources
- Good options in technology for virtualizing disk storage
- What's the problem?

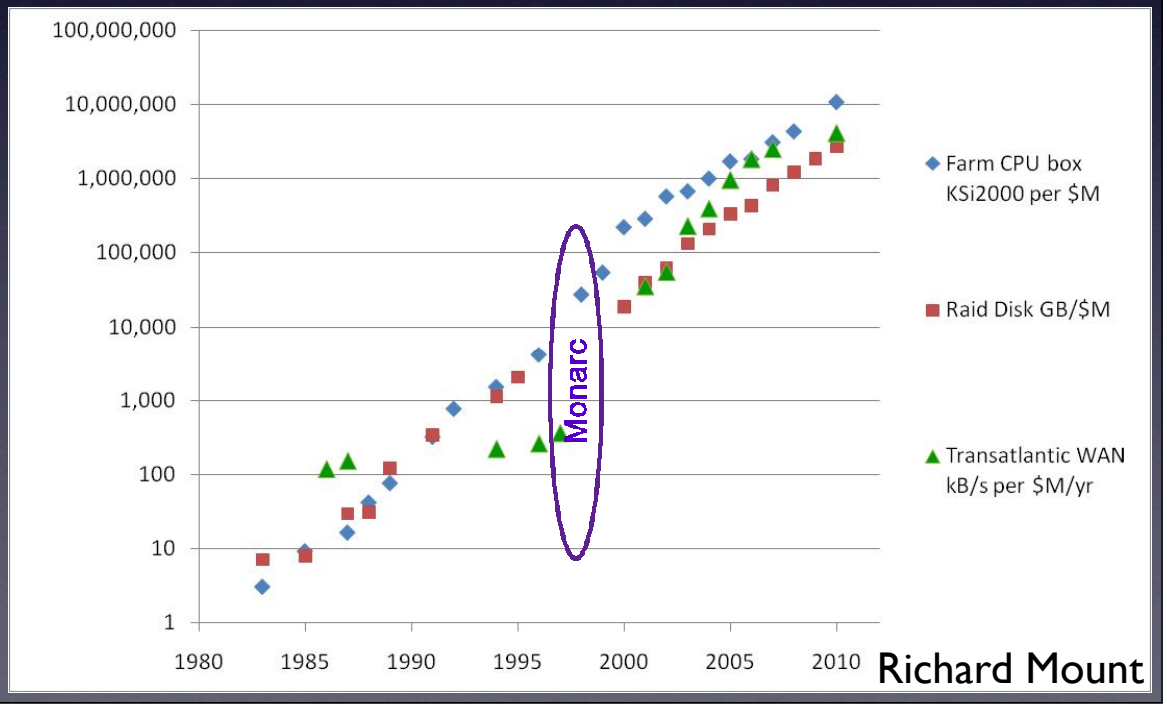


  - There are almost 100 Tier-2 sites that make up WLCG
    - Managing the space accessed by users efficiently is an interesting problem

Placement

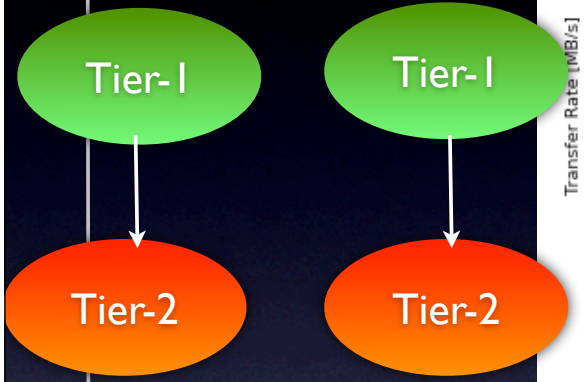
- Computing models for LHC were based on to greater and lesser extents on the MONARC computing model of 2000 and relied heavily on data placement
  - Jobs were sent to datasets already resident on sites
  - Multiple copies of the data would be hosted on the distributed infrastructure
  - General concern that the network would be insufficient or unreliable



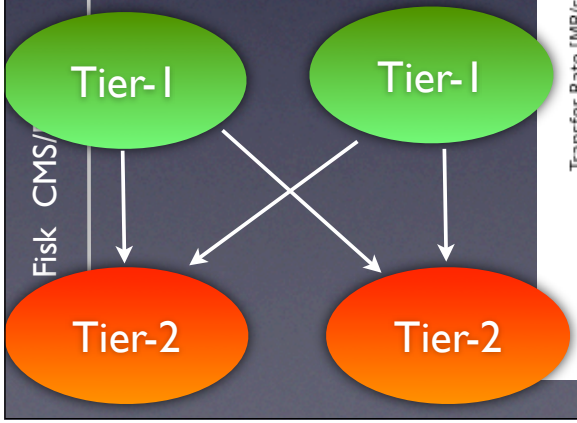


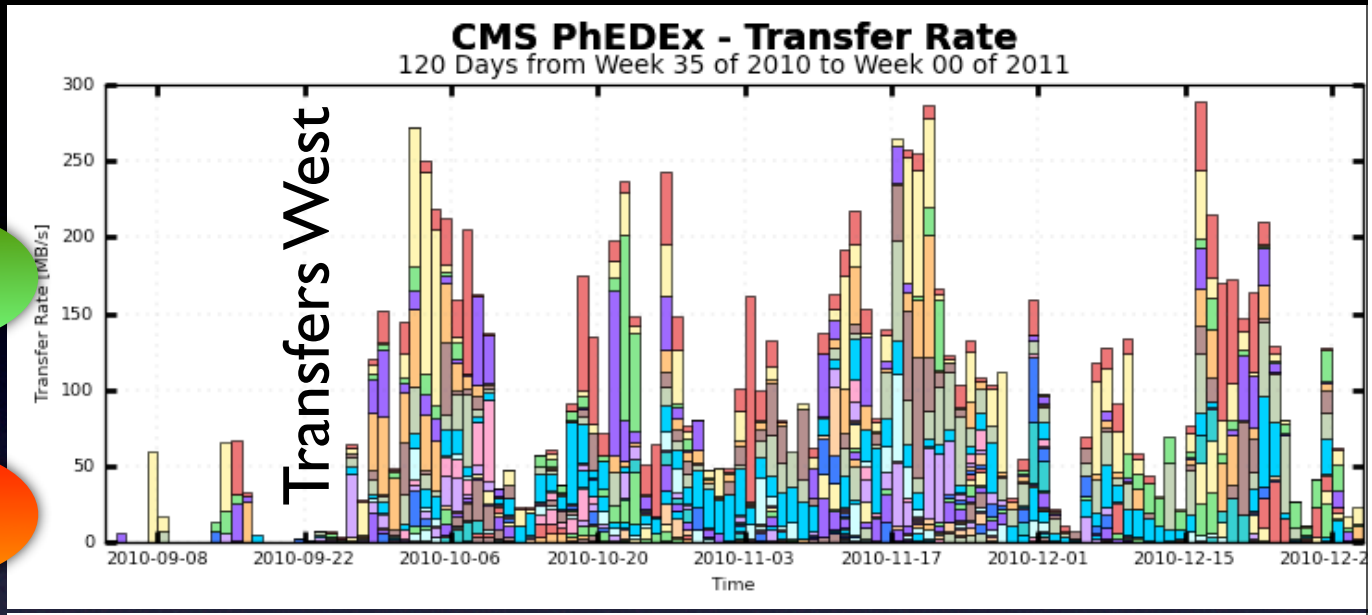
Distribution

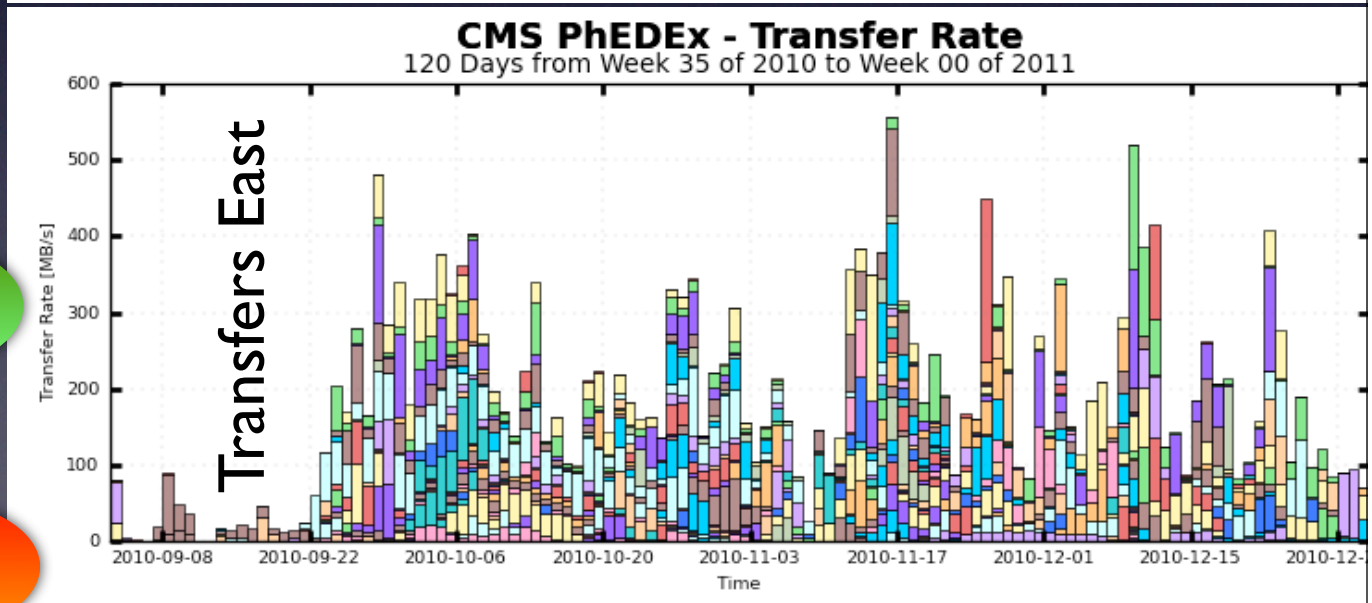
- Change from



- To

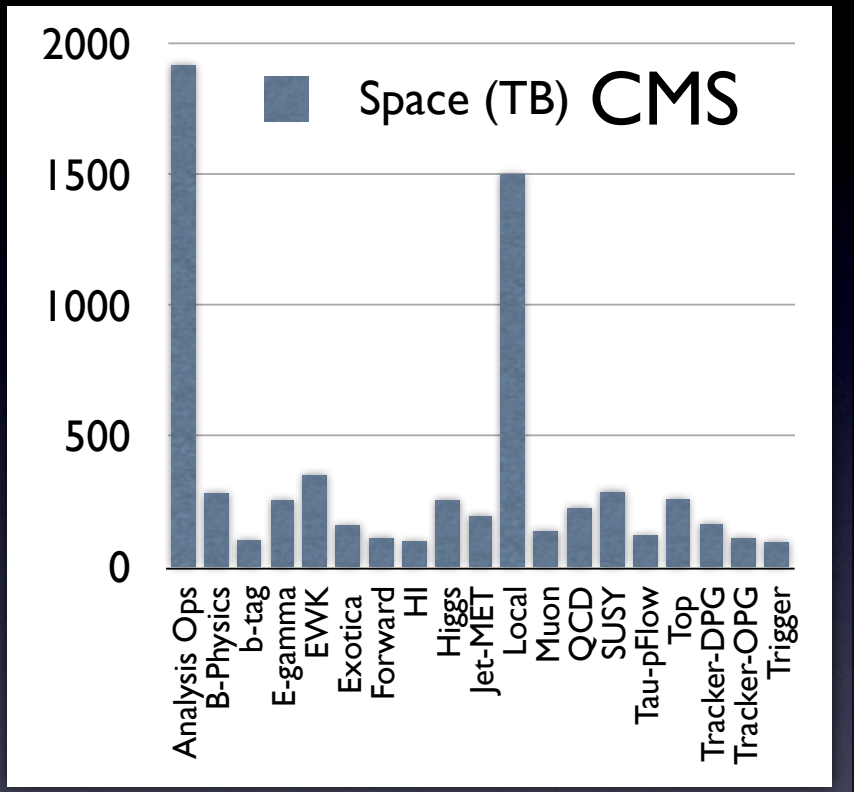






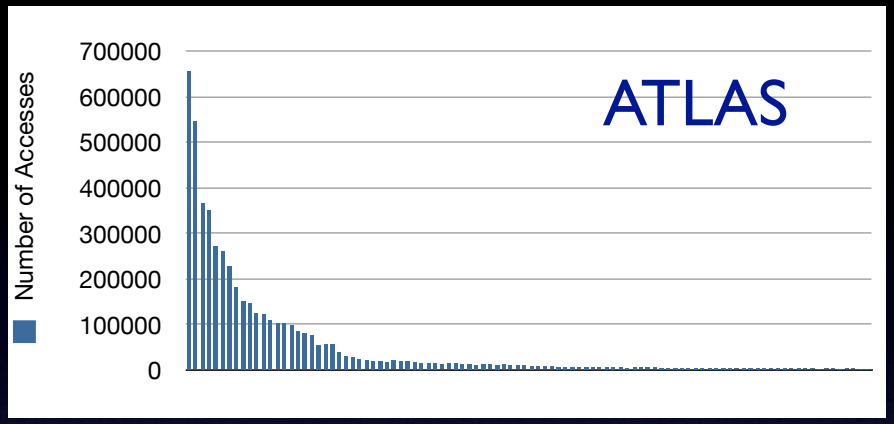
Data Management

- Experiments have chosen a variety of philosophy
  - ATLAS started with replication of nearly all data out to regions
  - CMS divided into Central background samples, physics groups, and the local community

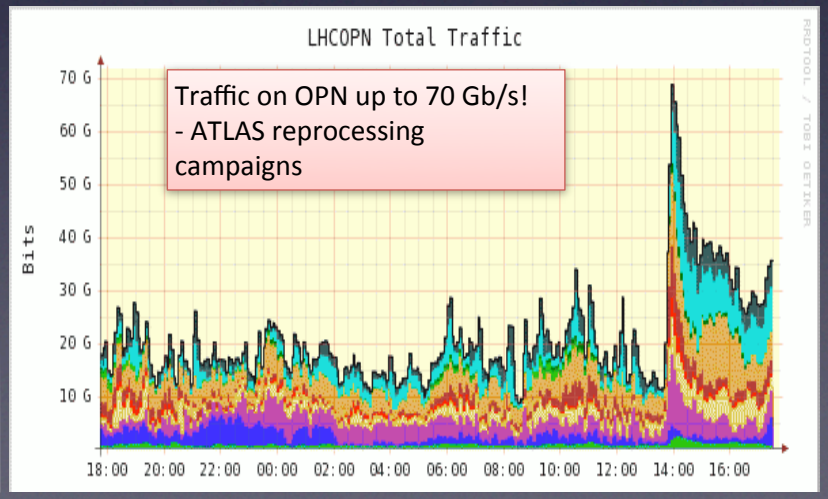


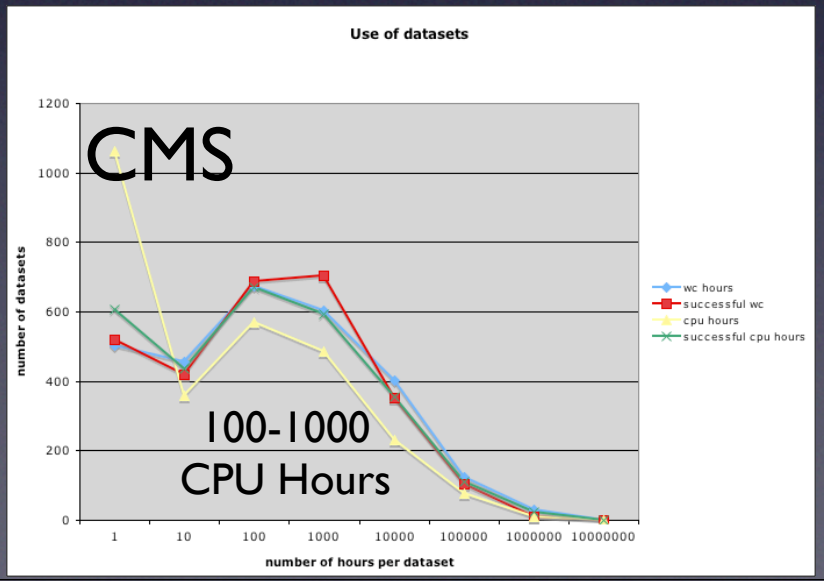
Access

- For a system intended to protect against weak networking, we're using a lot of network
  - LHC Experiments reprocessed a lot of data in 2010
  - Refreshing large disk caches requires a lot of networking

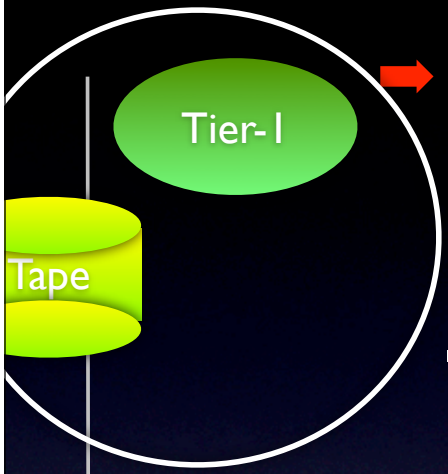


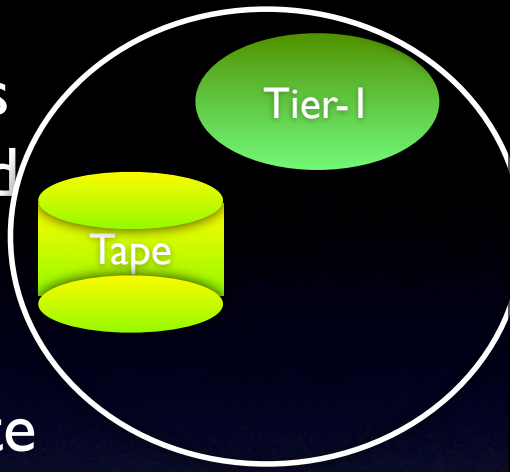
In CMS 30 % of samples subscribed by physicists not used for 3 months during 2010





Placement

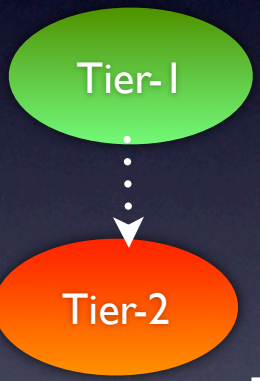




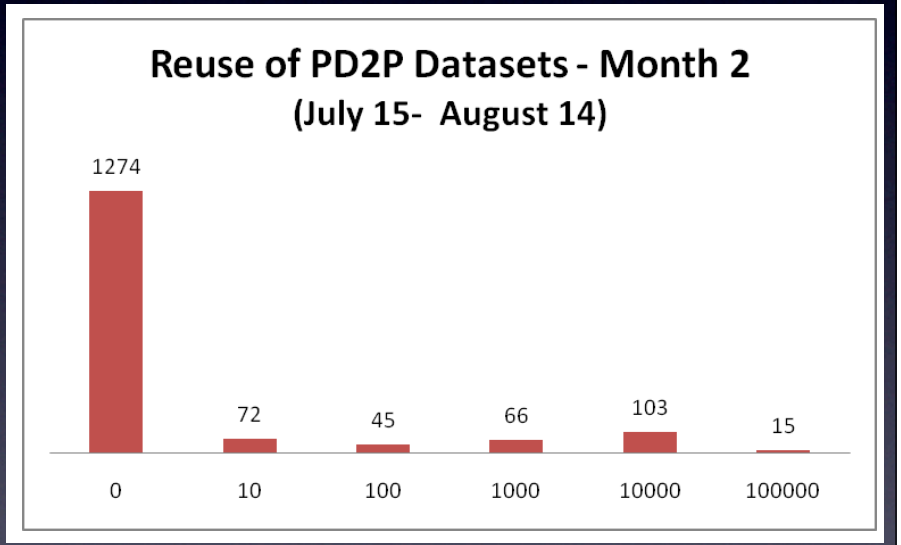
- In an environment that discounts the network the sites are treated independently
  - On the time scale of a job submitted and running on a site it is assumed the local environment cannot be changed
- From a data access perspective in 2011 data available over the network from a disk at a remote site may be closer than data on the local tape installation

Dynamic Replication

- ATLAS introduced Panda Dynamic Data Placement (PD2P)

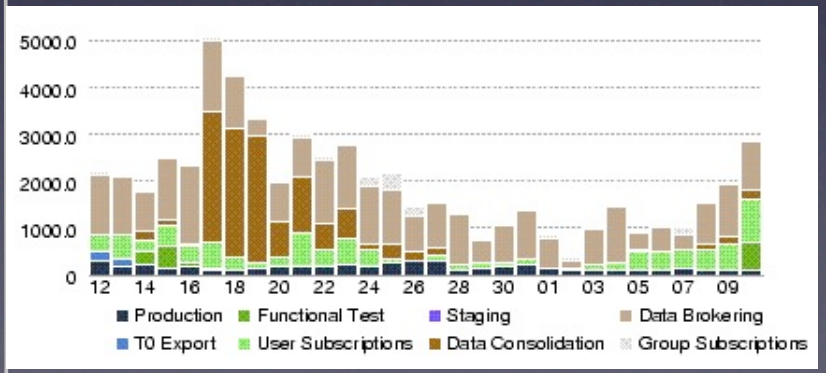


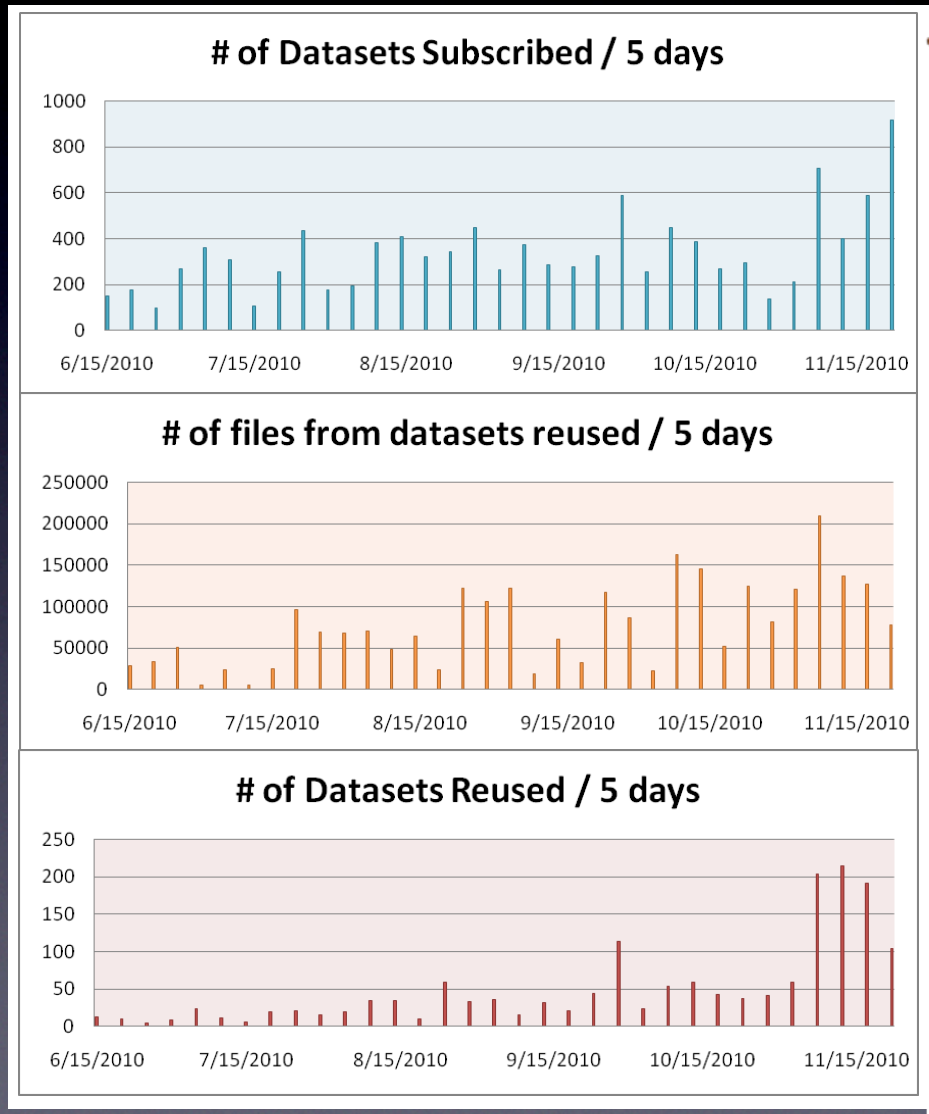
- Jobs are sent to Tier-1 and data replicated to a Tier-2 at submission time



Data Placement and ReUse

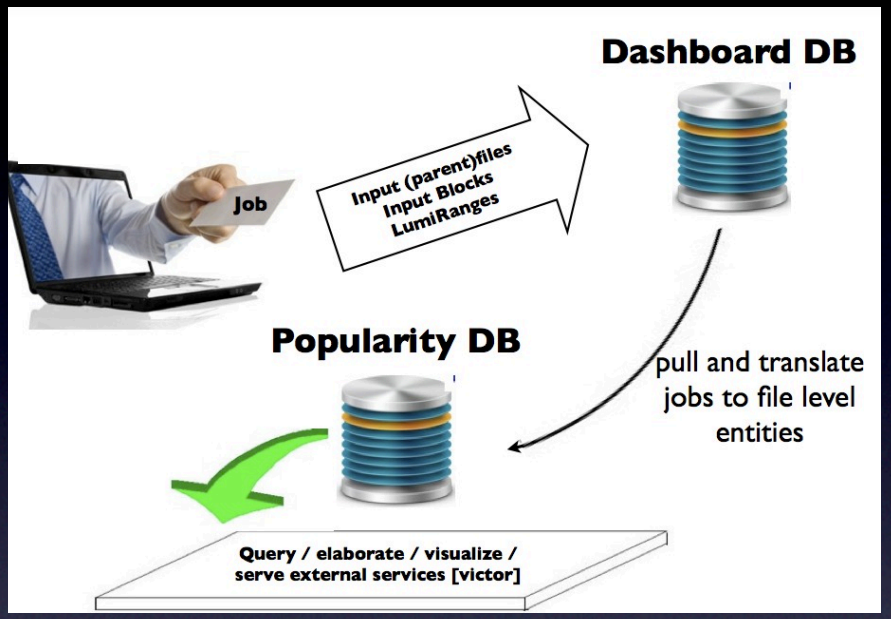
- Dynamic placement now accounts for a lot of the networking
- Re-brokering jobs is increasing the reuse of samples and the efficiency



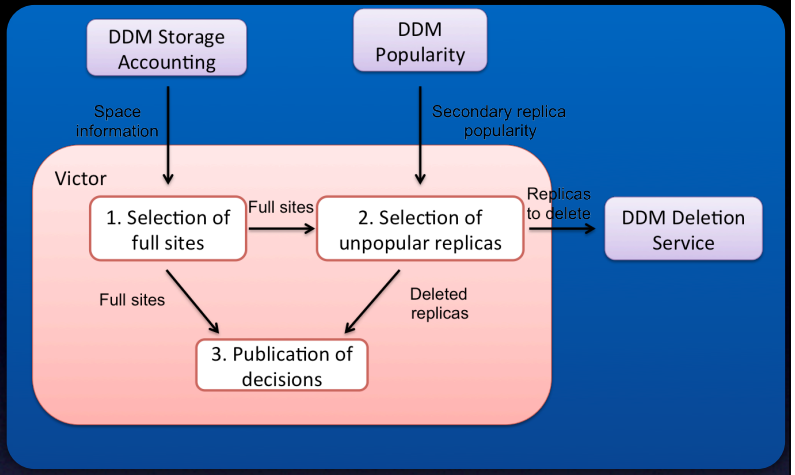


Popularity

- If you want to understand how better to manage storage space, important to know how it's used
- Interesting challenge to track the utilization of 30PB worth of files spread over more than 50 sites
  - Equally important to know what's not accessed



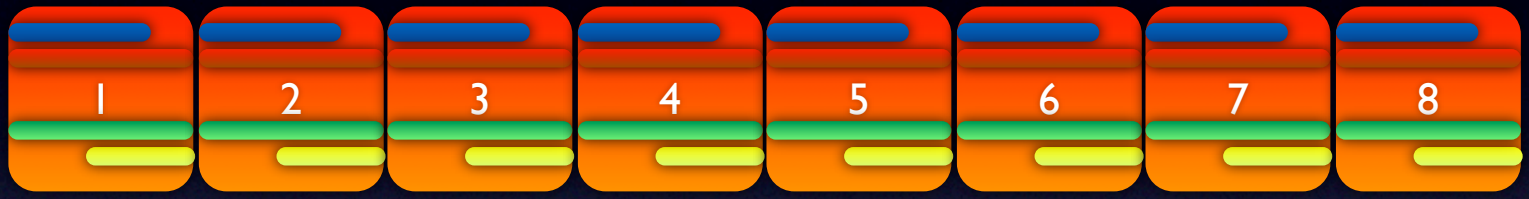
Clean-Up and Replication



- Once popularity is understood
  - Popular data can be replicated multiple times
  - Unused data replicas can be cleaned up
- Data Popularity will be tracked at the file level
  - Improves granularity and should improve reuse of the service

Analysis Data

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



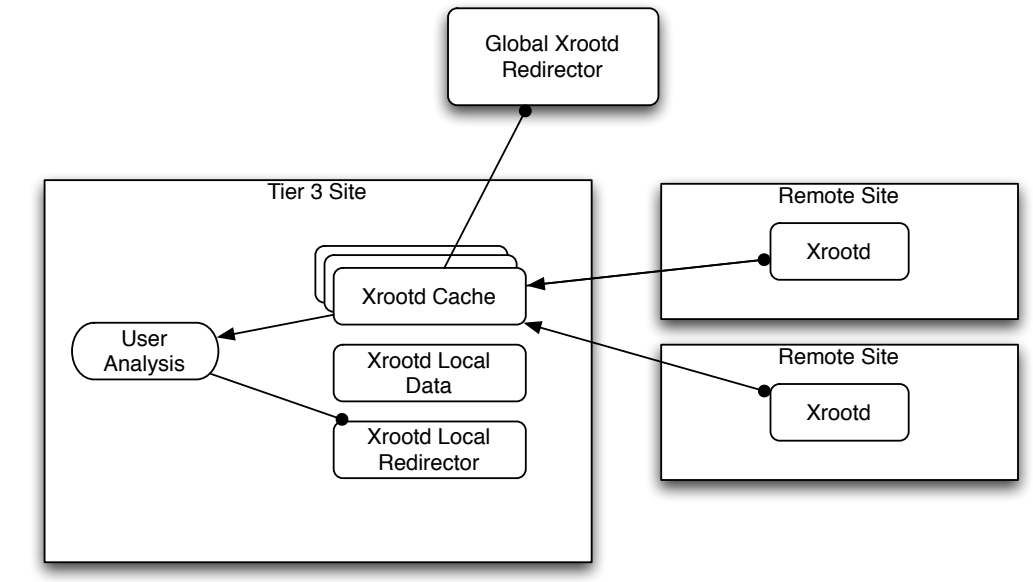
- In reality it's not how we either write or read the files
  - More like

- 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

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

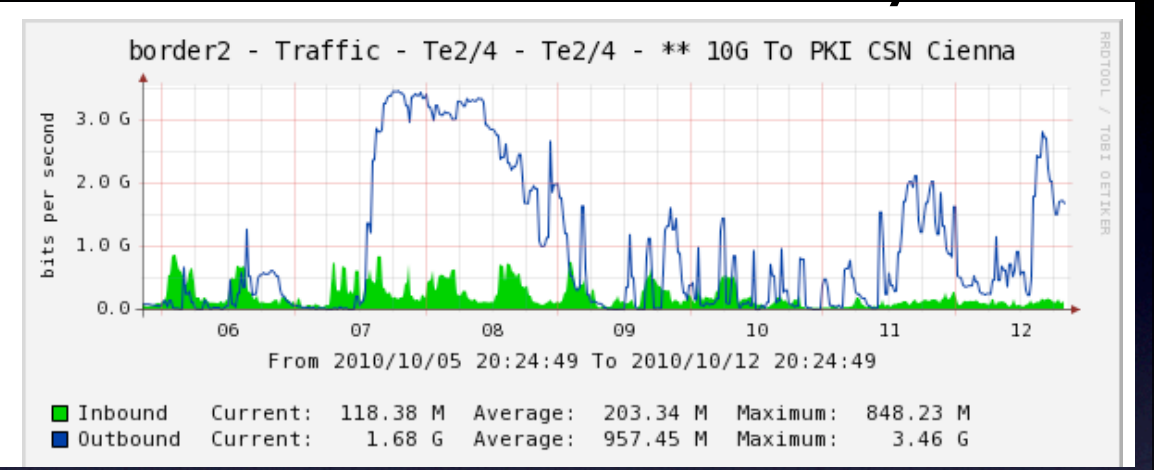
xrootd Demonstrator

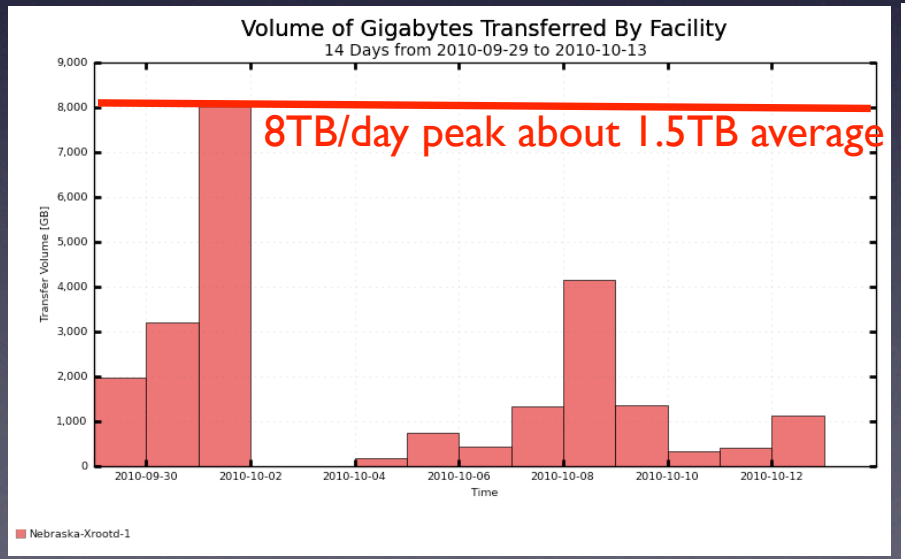


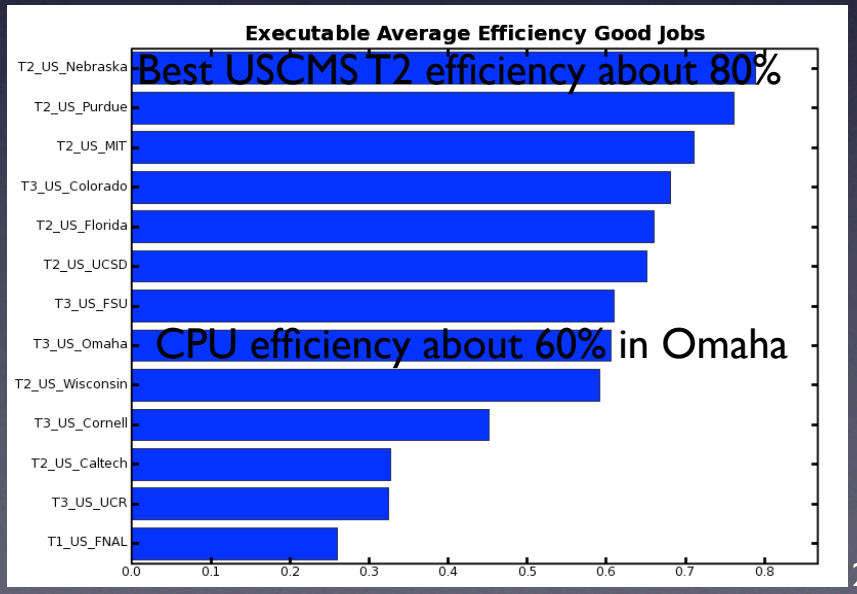
- Current Xrootd demonstrator in CMS is intended to support the university computing
  - Facility in Nebraska and Bari with data served from a variety of locations
  - Tier-3 receiving data runs essentially diskless
- Similar installation being prepared in ATLAS

Performance

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

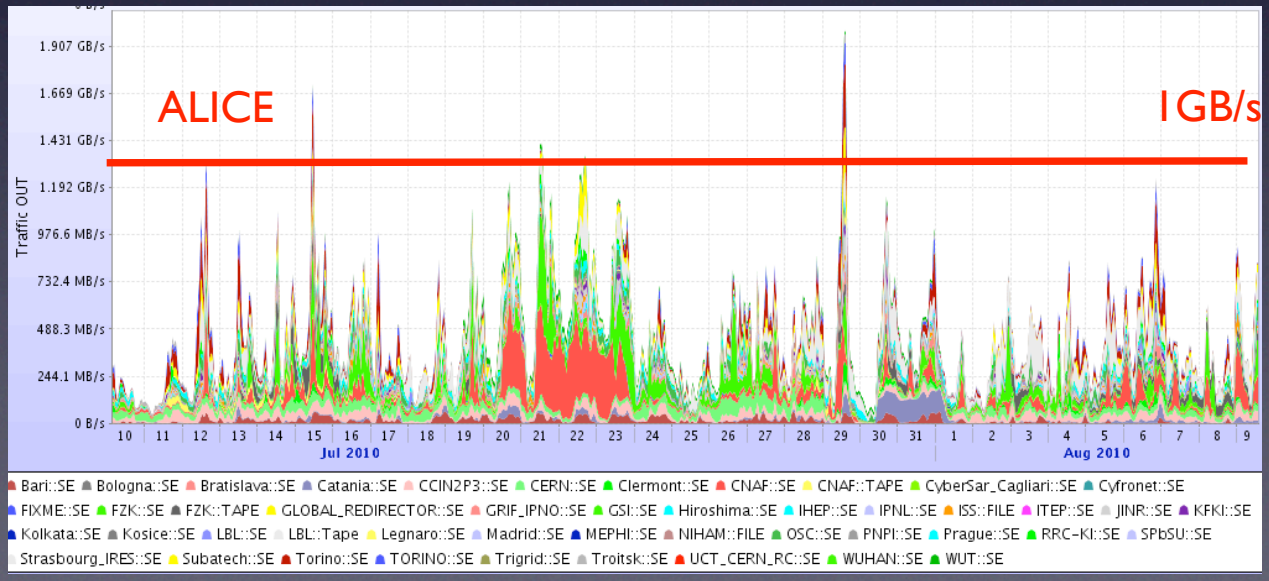






Networking

- ALICE Distributes Data in this way
  - Rate from the ALICE Xrootd servers is comparable in peaks to other LHC experiments



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

- First year of running on LHC went well
  - We are learning rapidly how to operate the infrastructure more efficiently
  - Making more dynamic use of the storage and making better use of the networking
  - 2011 and 2012 are long runs at the LHC
    - Data volumes and user activities are both increasing