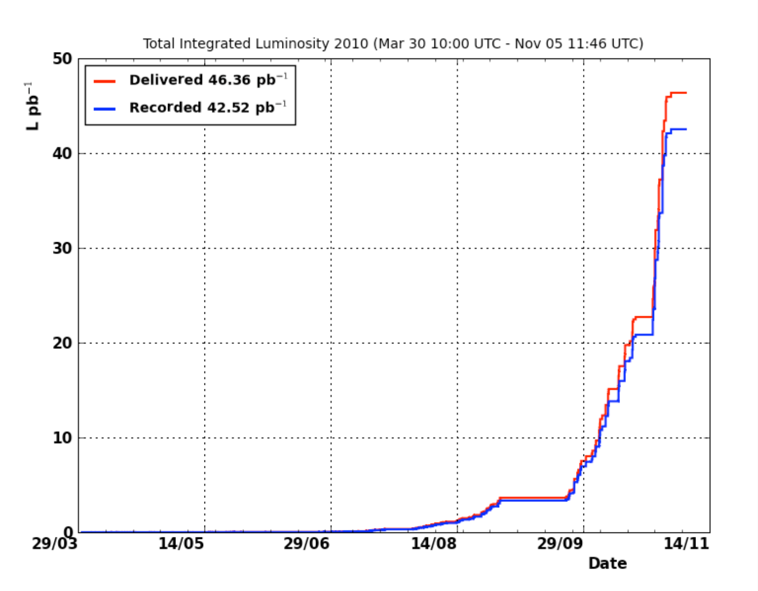
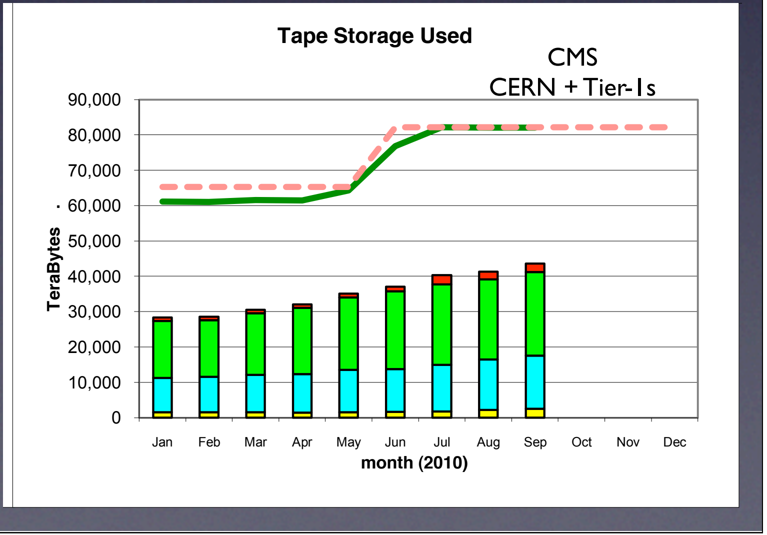
CMS Challenges and Needs

Nov 5., 2010
Ian Fisk

CMS

- Finished the first year of Proton-Proton on Wednesday
  - We have 42pb-1 recorded
    - About 75% of which was collected in Oct.
  - Estimates for next year are 20 times larger
- At the same time CMS has written about 20PB to tape
  - will write about 25PB per year





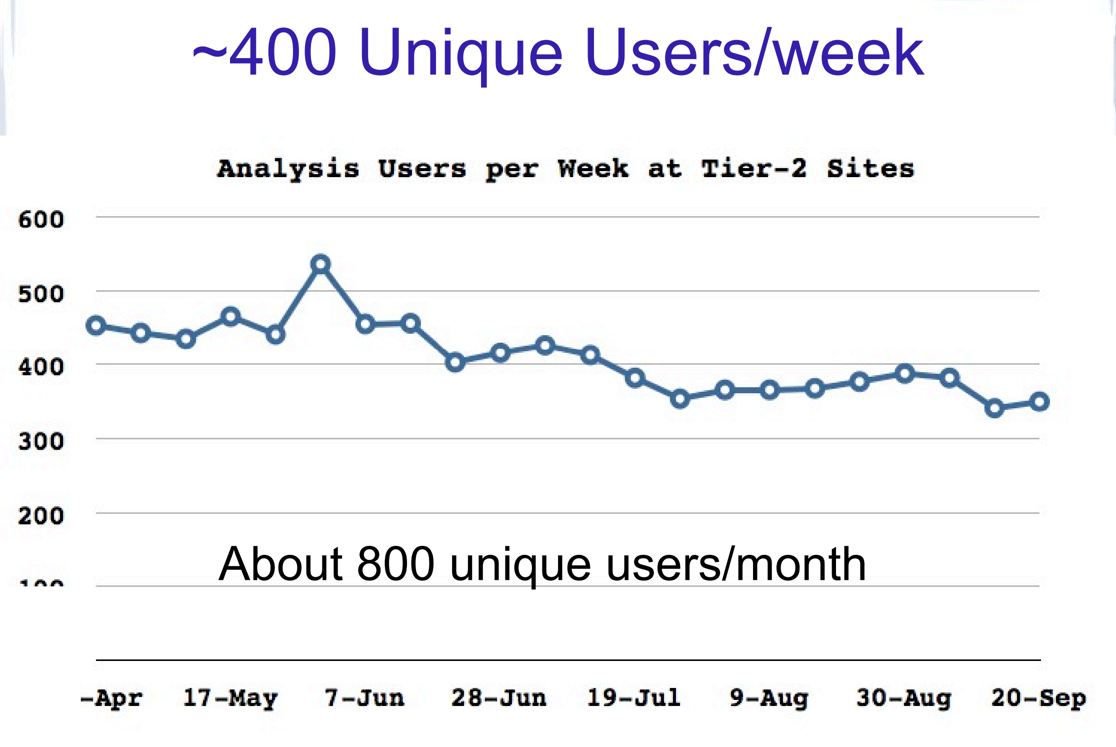
Changes of Scale

- Decreases in the cost of disk and technology to run big disk farms
  - CMS relies more heavily on staging

	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

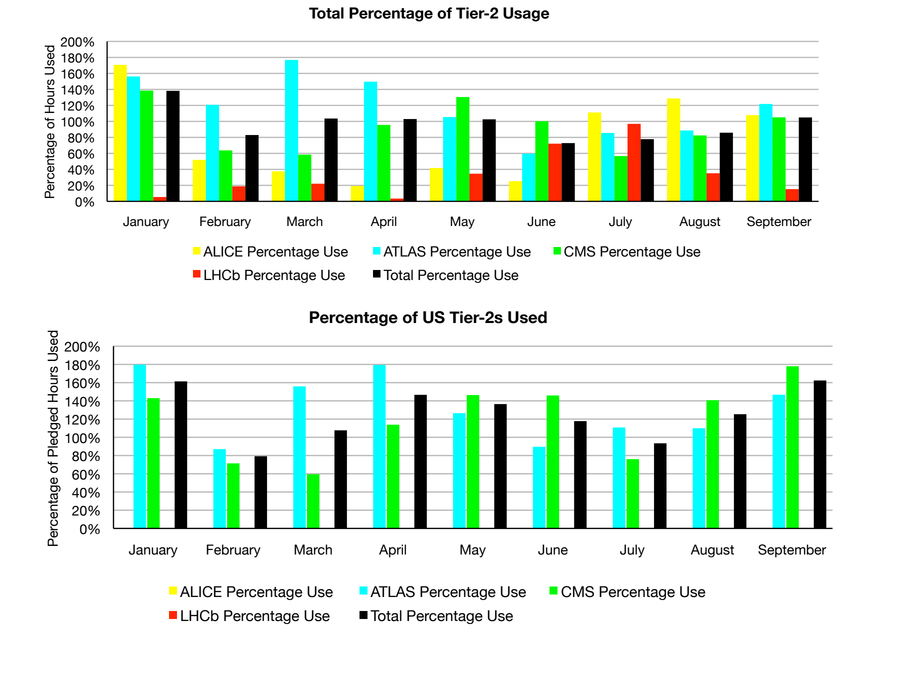
  - In 2011 majority of the currently accessed data could be disk resident

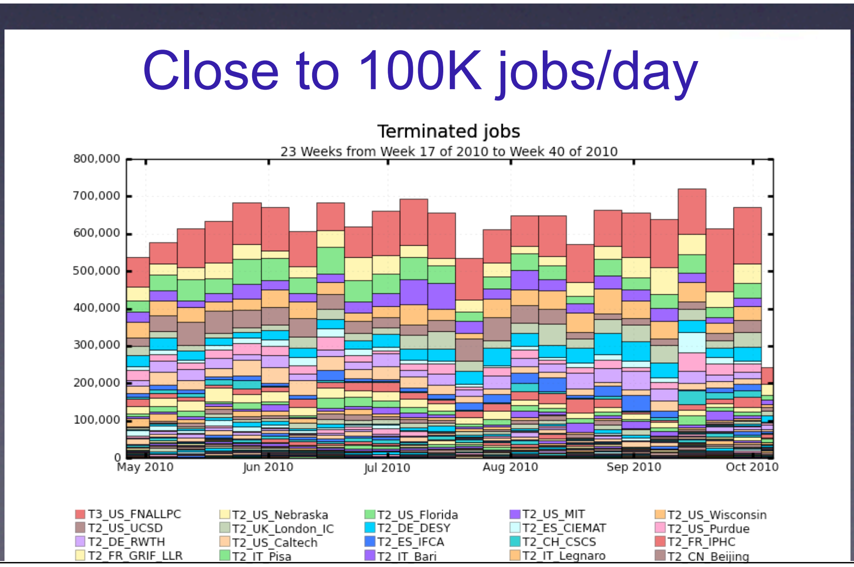
Large Analysis Activity



~400 Unique Users/week

About 800 unique users/month



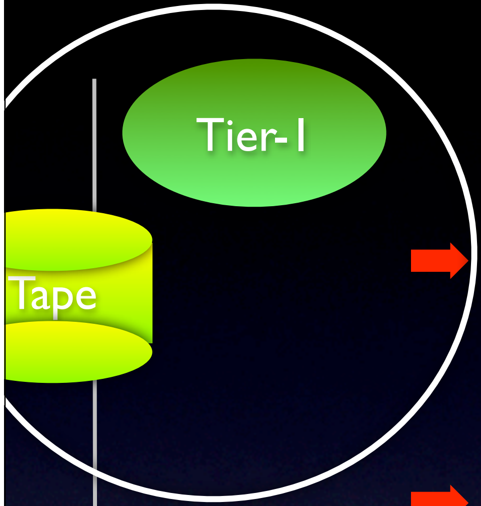


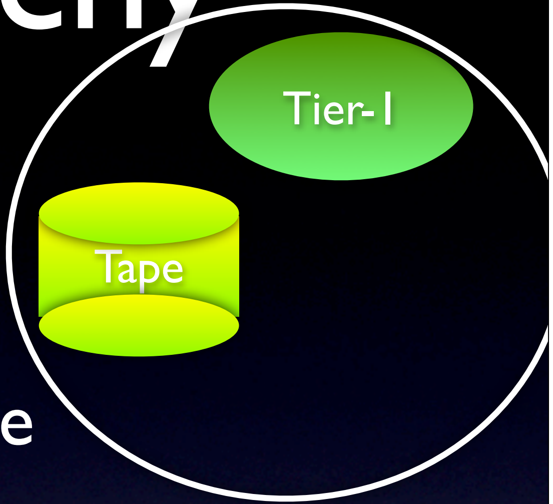
Close to 100K jobs/day

Challenges

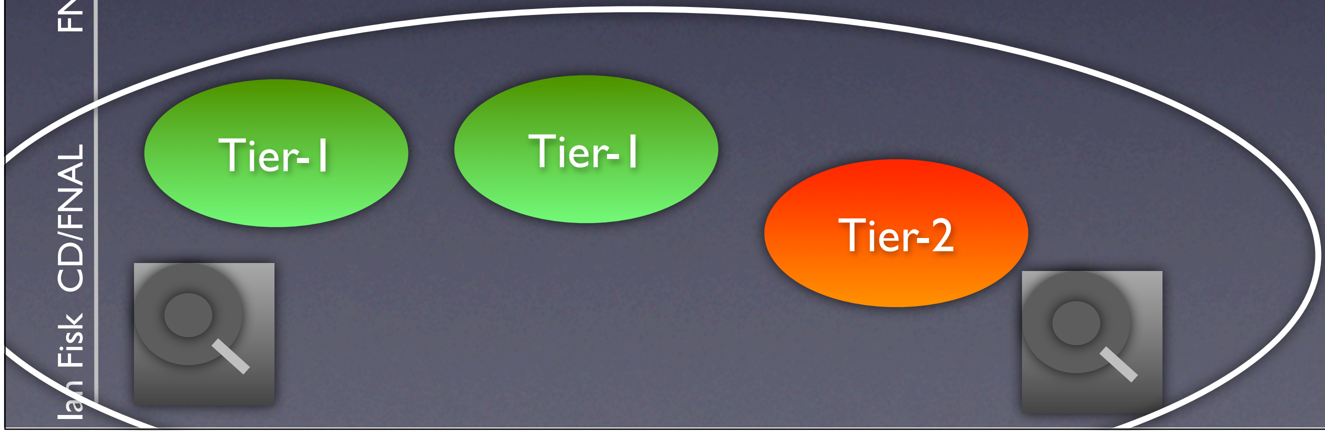
- ➡ Storage and Storage Hierarchy
- ➡ Increasing Geographic Distribution
- ➡ Data Placement and Data Access
- ➡ Resource Prioritization and Aggregation

Storage Hierarchy





- We view the Tier-1s as a global system, but a system made up of sites and not a coherent resource
- 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



What would be needed?

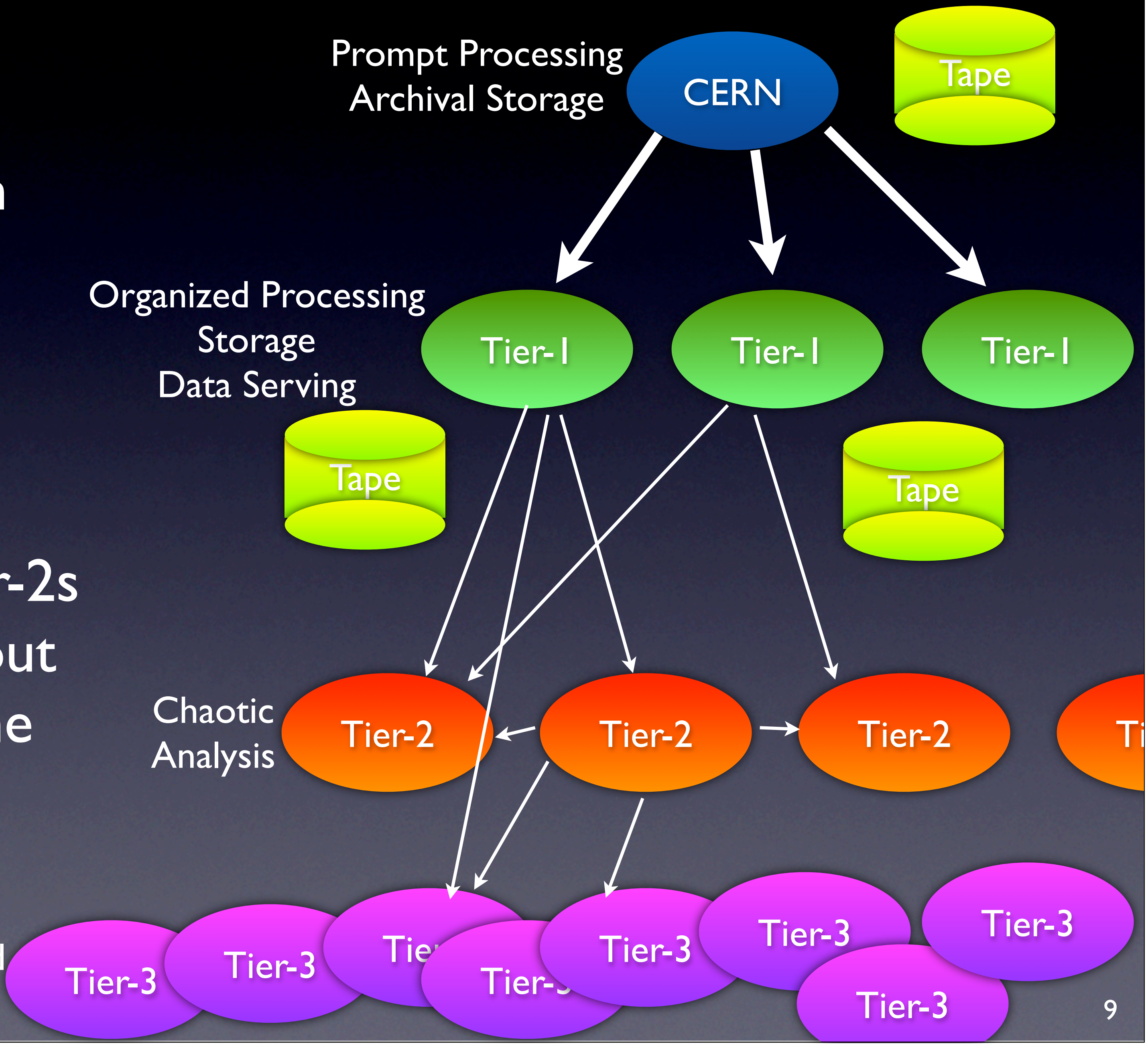
- Mostly issues of IO, Data Access, and Data Management
  - Faster file open and transfer protocols than SRM
  - Better consistency about files available at each site
  - Making sure resources are available for transfer
    - Scheduling Networks to gateways to disk systems
-

Potential Direction

- ➡ Less dynamic access to tape
  - Data migrations are scheduled events
- ➡ Softer boundaries between computing centers
  - Storage seen as a cloud between facilities

Geographic Distribution

- Tier-3s are the growth industry in CMS
  - More Tier-3s than Tier-2s (52), about half in the US



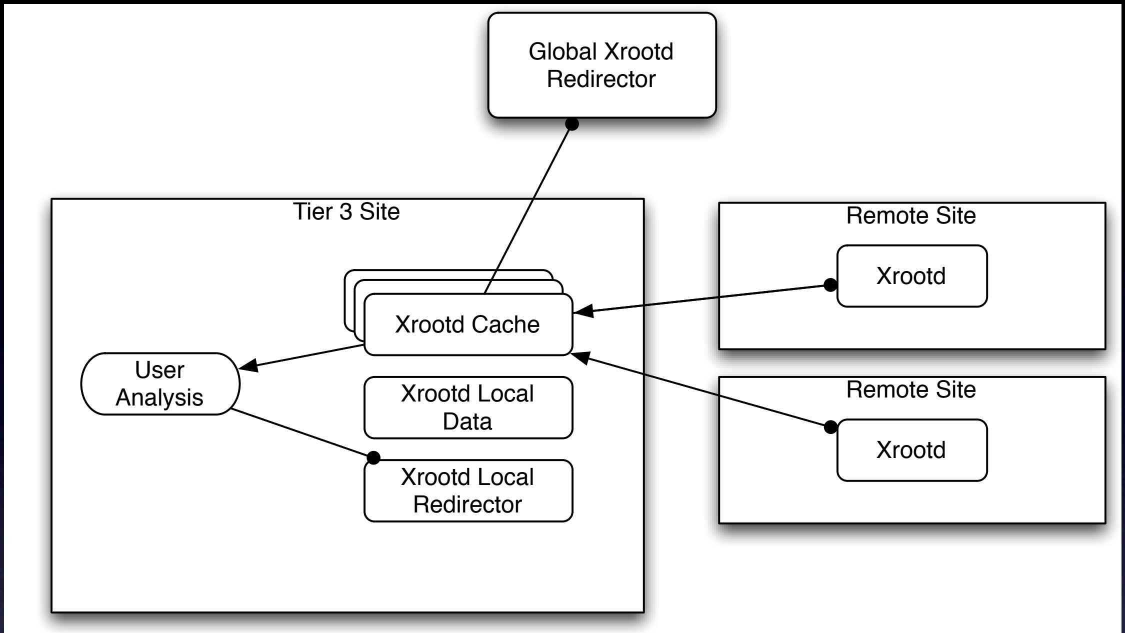
Tier-3s

- ➡ Good opportunity for additional analysis resources
- ➡ Generally smaller installations, but limited effort
- ➡ Up to now we have treated these like smaller Tier-2s
  - Services required are similar, but effort and resources deployed are smaller
  - Not clear that this is the most efficient model

Capitalizing on Tier-3s

- We have an interesting resource for Analysis
  - How to make them more efficient
    - Solve the data Management Problem
      - Single largest complaint is the need to run the experiment data management system
    - Reduce the effort to operate the grid interfaces

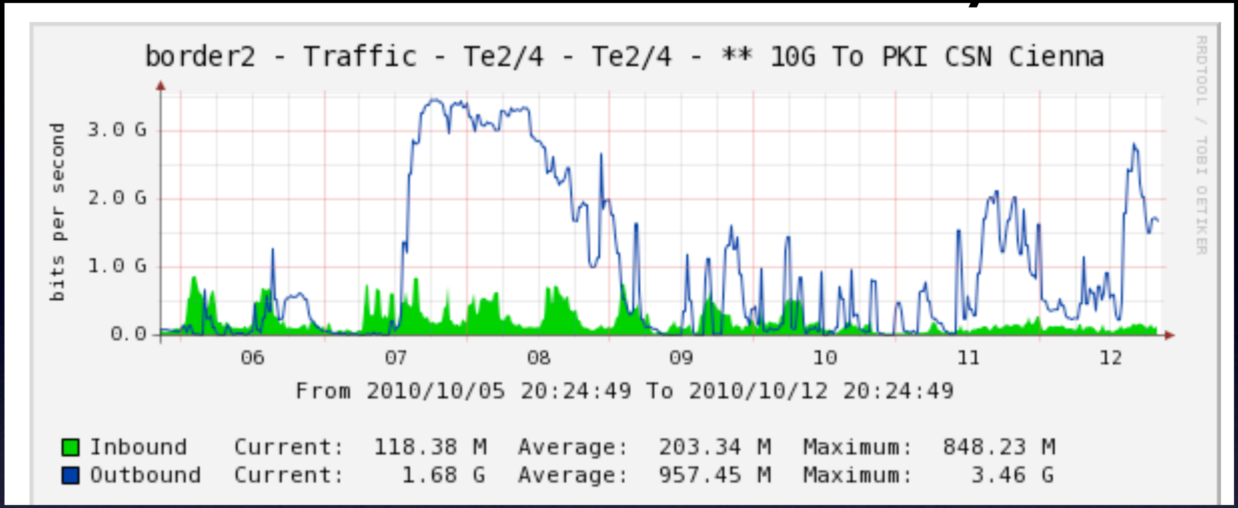
xrootd Demonstrator

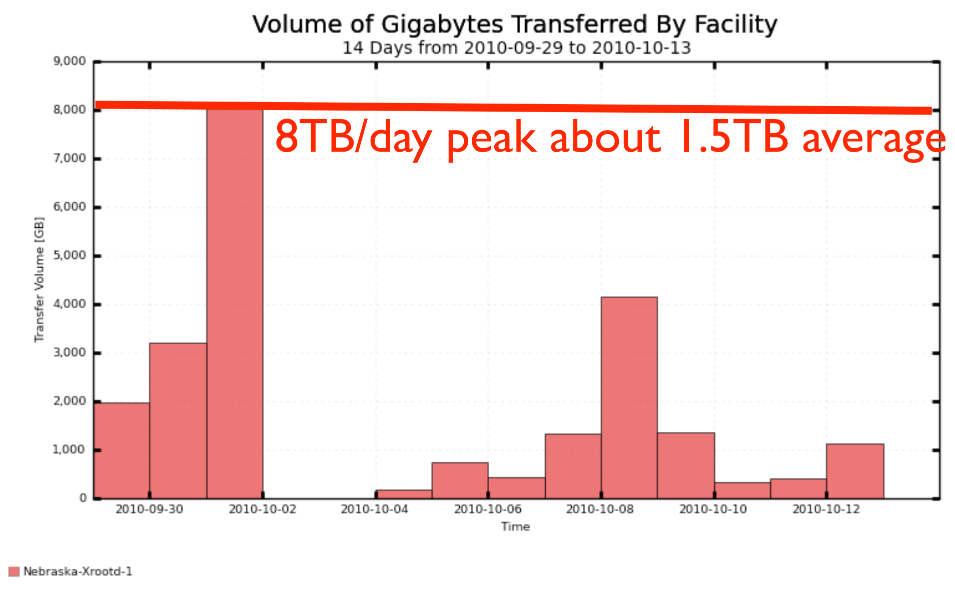


- Current Xrootd demonstrator in CMS is intended to support the Tier-3s (Lead by Brian Bockelman)
  - Facility in Nebraska with data served from a variety of locations
  - Tier-3 receiving data runs essentially diskless
- Similar installation being prepared in ATLAS

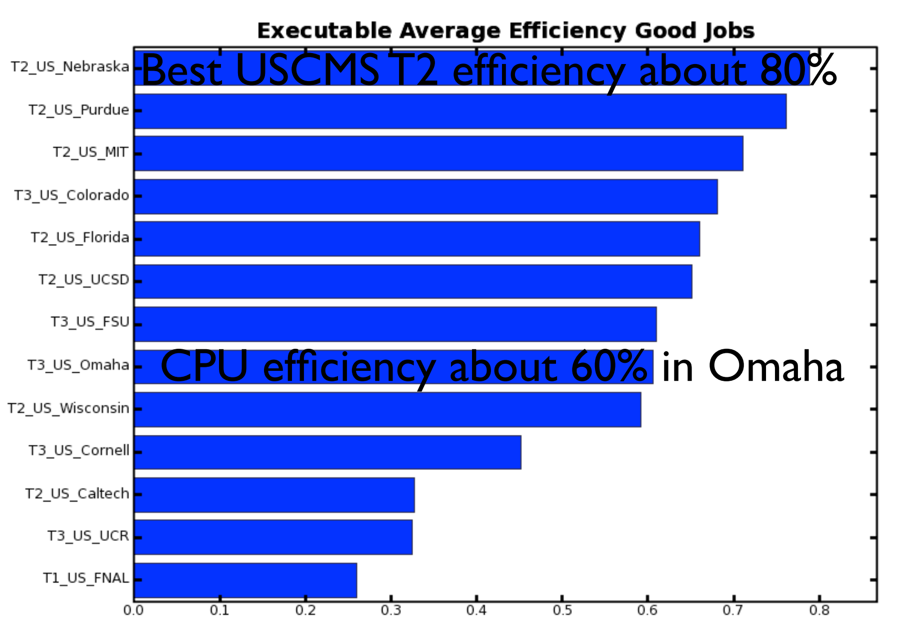
Performance

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





8TB/day peak about 1.5TB average

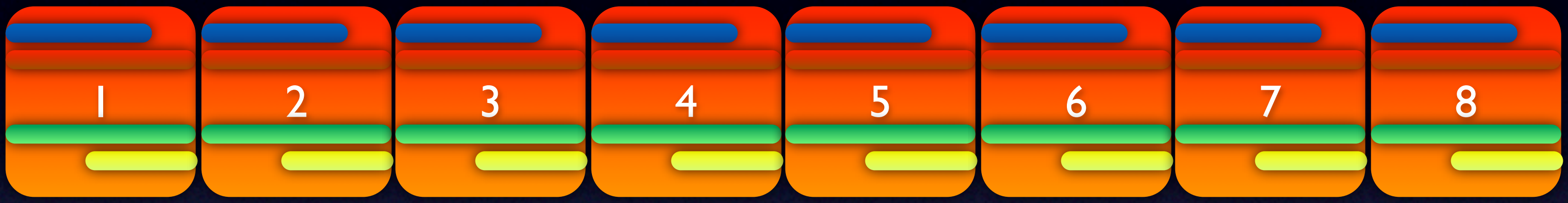


Best USCMS T2 efficiency about 80%

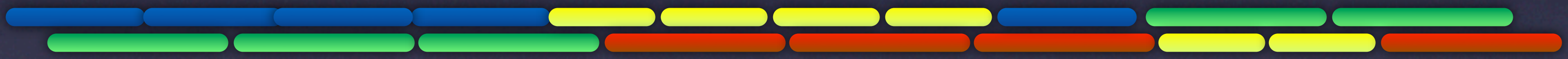
CPU efficiency about 60% in Omaha

Analysis Data

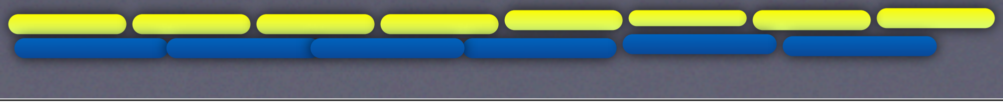
- We like to think of high energy data as series of embarrassing parallel events



- 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

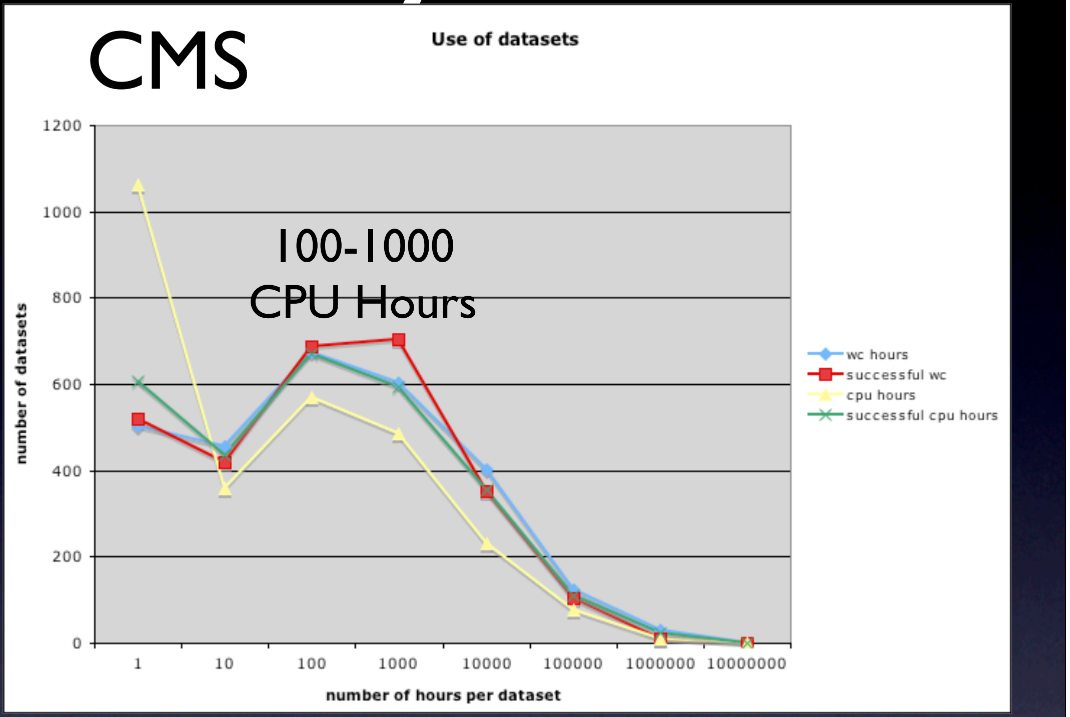
- With properly optimized IO other methods of managing the data and the storage are available
  - Sending data directly to applications over the WAN
- Not immediately obvious that this increases the wide area network transfers
  - If a sample is only accessed once, then transferring it before hand or 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

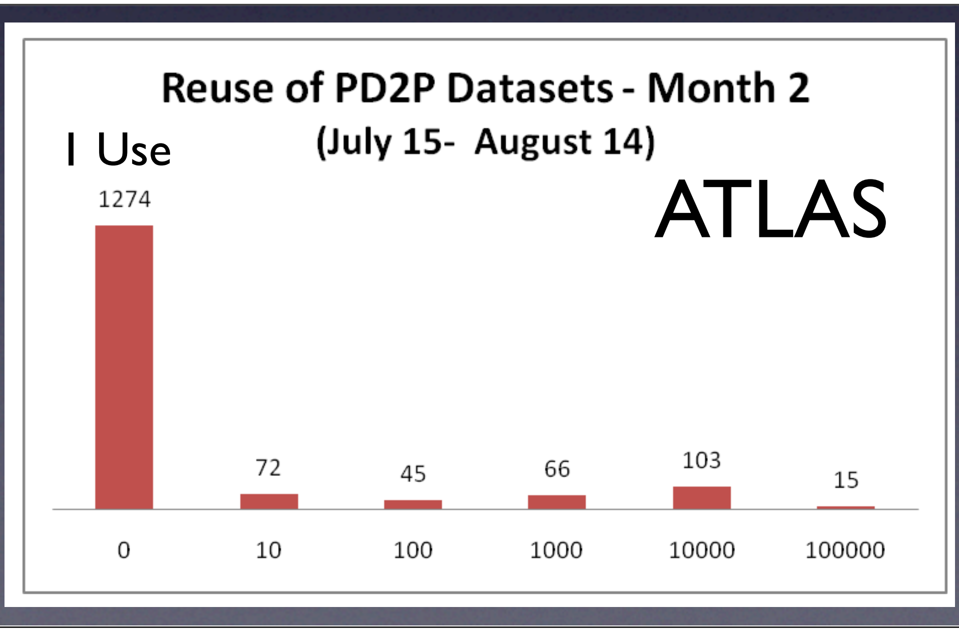
Data Placement and Access

- ➡ So is wide area access a solution for other types of data placement and access problems
  - Tier-1 access to data on this at another Tier-1?
  - Tier-2 access to non-local data?
- ➡ Maybe!

Popularity

- Huge variation in the access level to data
  - Usefulness of data in the LHC is short
- We may need different strategies for data used a lot and data used once



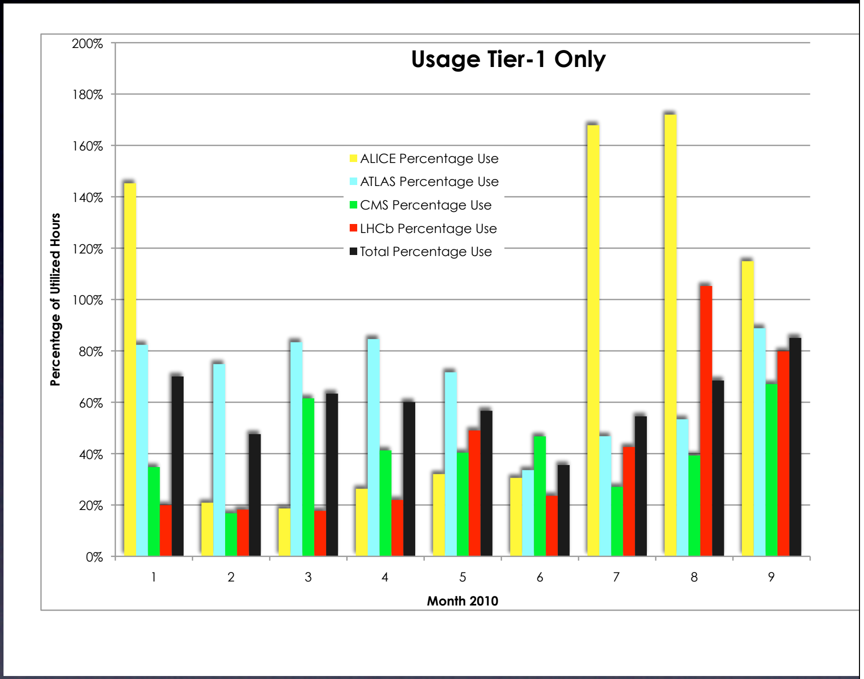


What’s Needed?

- Predicability
  - You shouldn’t tie CPUs directly to the wide area network without knowing the network is going to deliver
- Throttles
  - Protect the facility against being knocked over by remote access
- Data Management
  - Smarter systems to predict when data needs to be replicated and when it’s past it’s useful life
    - More complete monitoring of access

Resources

- CMS is not yet fully resources constrained but will get there soon
  - Challenge on how to steer the use of the resources across a globally distributed set
    - Normally this is done with central task queues
      - Scaling issues
    - CMS has done this coarsely with interactions with the sites



Many Processes and Many Cores

- Currently we have 1 process per core and track both of these
- Looking at ways of taking the whole node
  - Reducing the number of processes we need to track and increasing the efficiency of the node
    - Better memory and IO management
- Challenging aspect is the transition
  - While this is a multi-core challenge, we think most of the work is in workflow

Outlook

- CMS has a quickly growing dataset and interesting challenges in how to evolve the storage and access
  - IO and data management work
  - Wide area access with limits
- A geographically separated computing facility that continues to grow
  - Improve the utilization and efficiency
- Prepare for resource constrained priority decisions