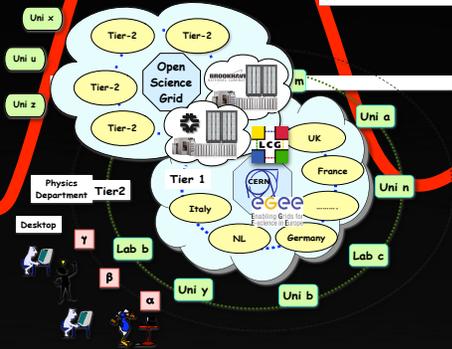
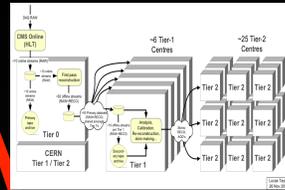
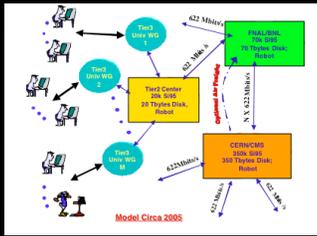


New Computing Models and LHCOne

22 May 2012

Ian Fisk

Evolution



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

Old Model Impact



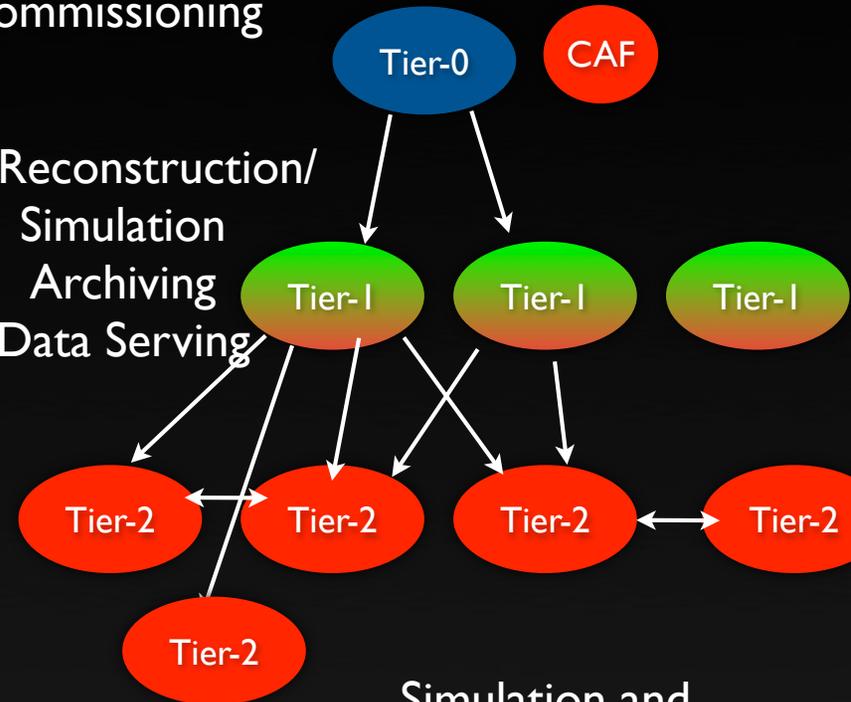
- ➔ Connections are seen as not sufficient or reliable
- Data needs to be pre-placed
 - Nothing can happen utilizing remote resources on the time of a running job
 - Data comes from specific places
 - Sites have specific functions

New Model

- Strict hierarchy of connections becomes more of a mesh
- Divisions in functionality especially for chaotic activities like analysis become more blurry
- More access over the wide area

Prompt Reconstruction
Storage
Commissioning

Re-Reconstruction/
Simulation
Archiving
Data Serving



Simulation and
User Analysis

- ▶ Model changes have been an evolution
- ▶ Not all experiments have emphasized the same things
- ▶ Each pushing farther in particular directions

Grid Services

- During the evolution the low level services are largely the same
- Most of the changes come from the actions and expectations of the experiments

Experiment Services

WMS

BDII

FTS

VOMS

Higher Level Services



Connection to batch (Globus and CREAM based)

CE

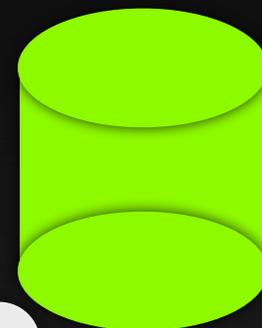
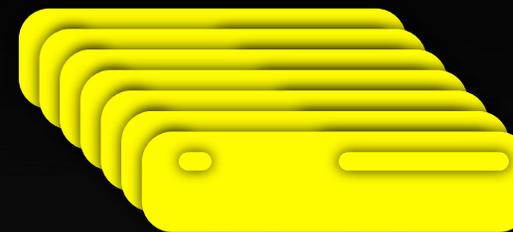
Information System

SE

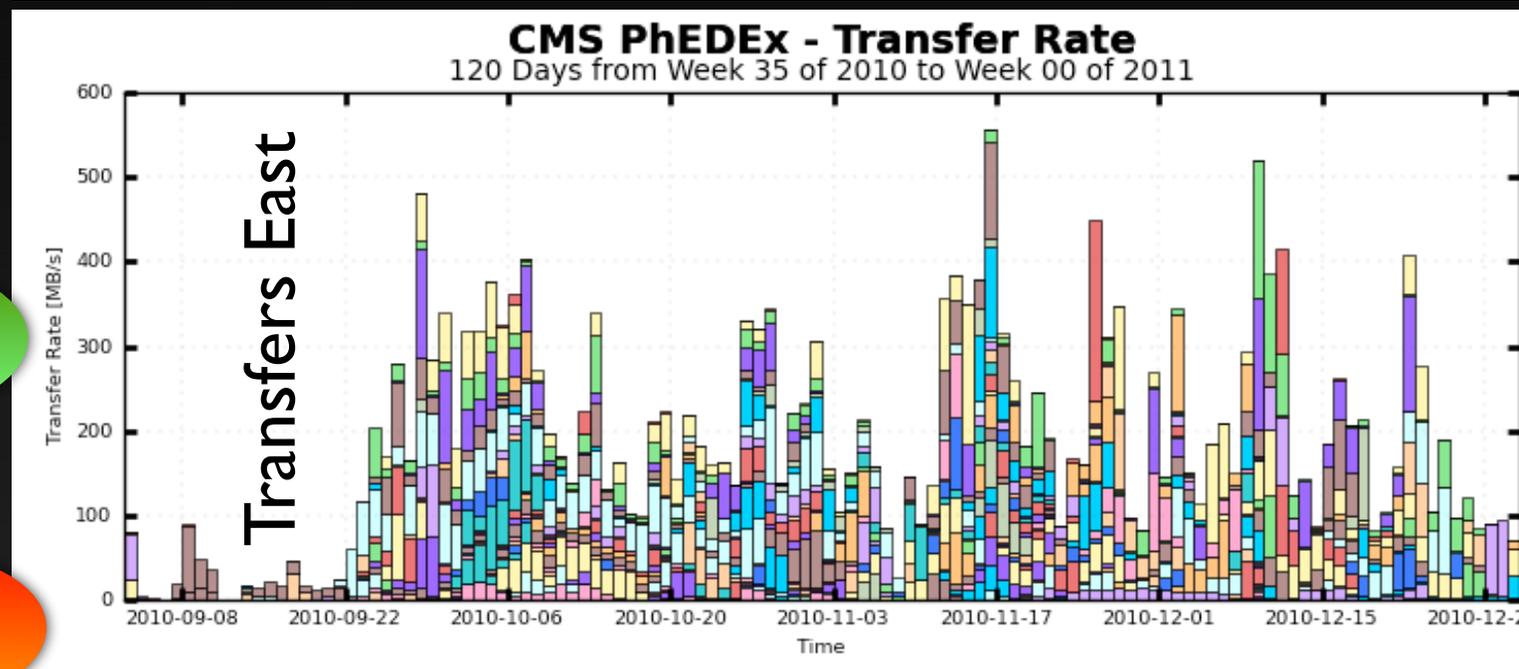
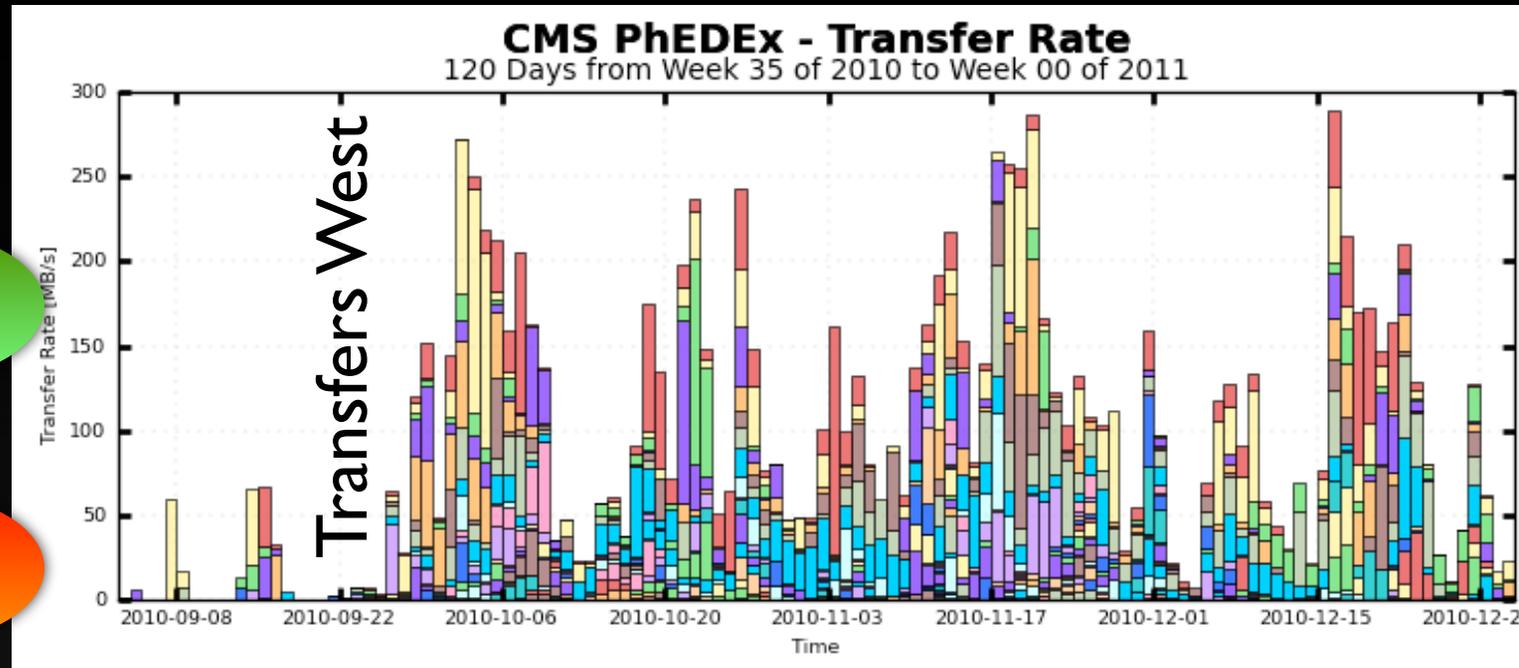
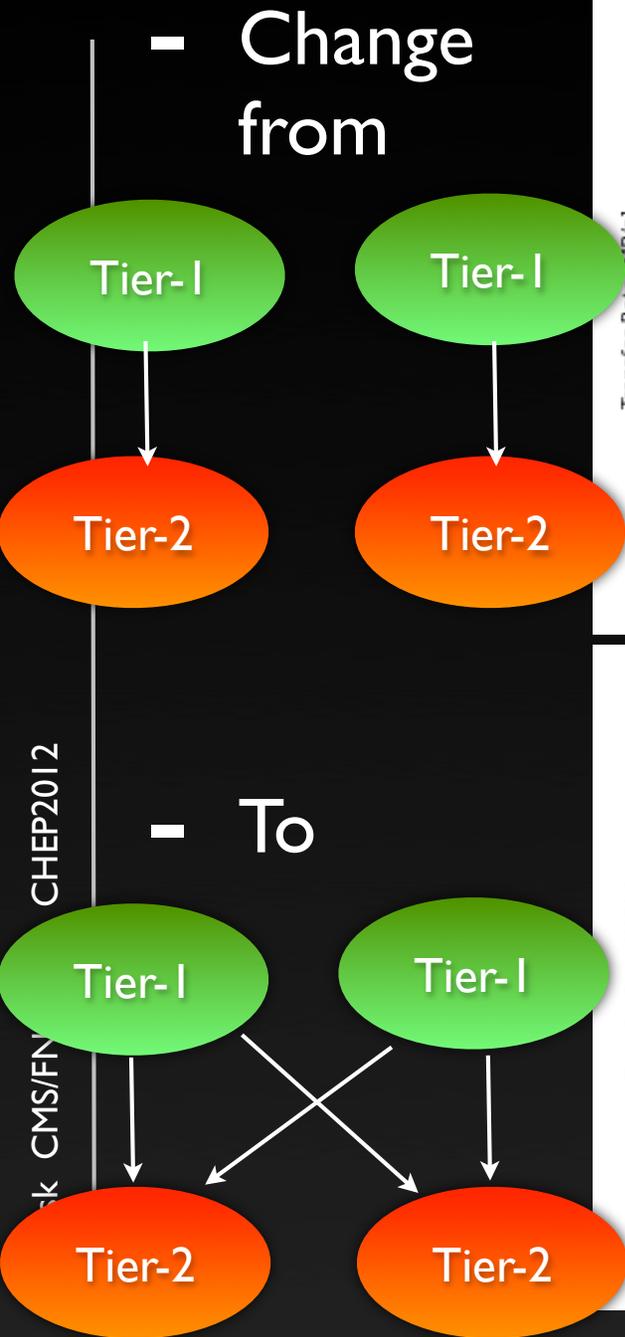
Connection to storage (SRM or xrootd)

Lower Level Services Providing Consistent Interfaces to Facilities

Site



Mesh Transfers



Tier-2

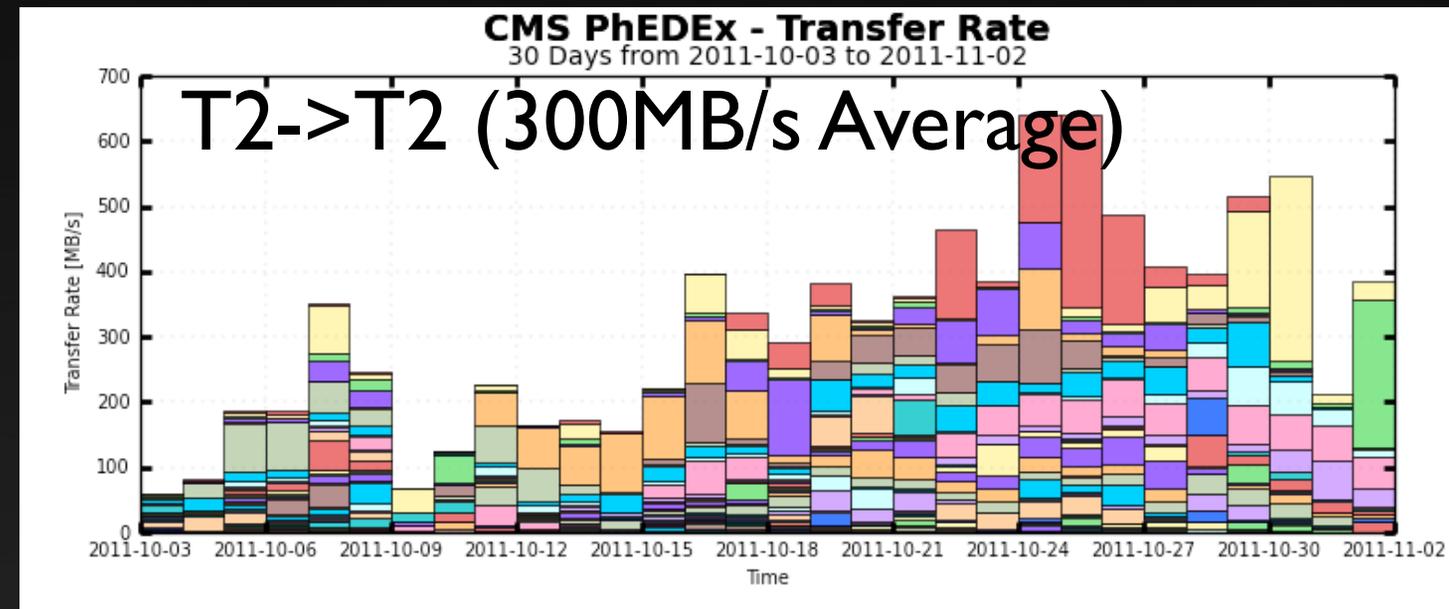
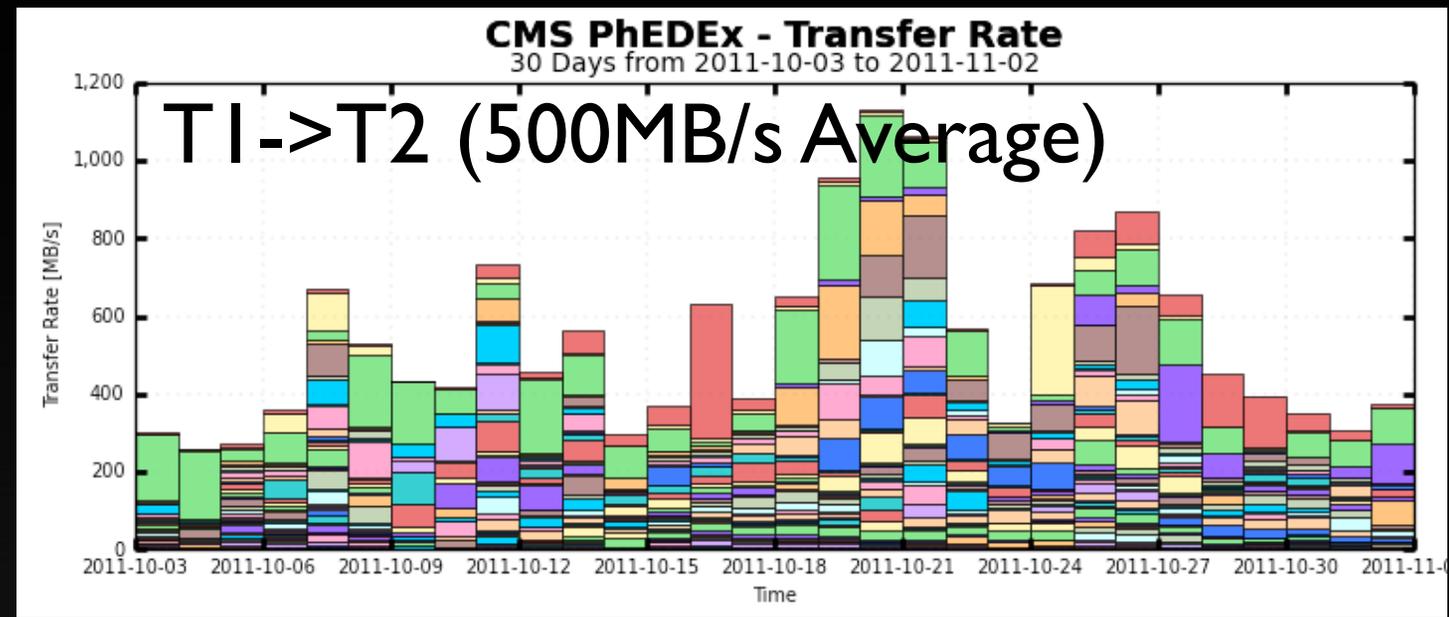
Tier-2

Completing the Mesh

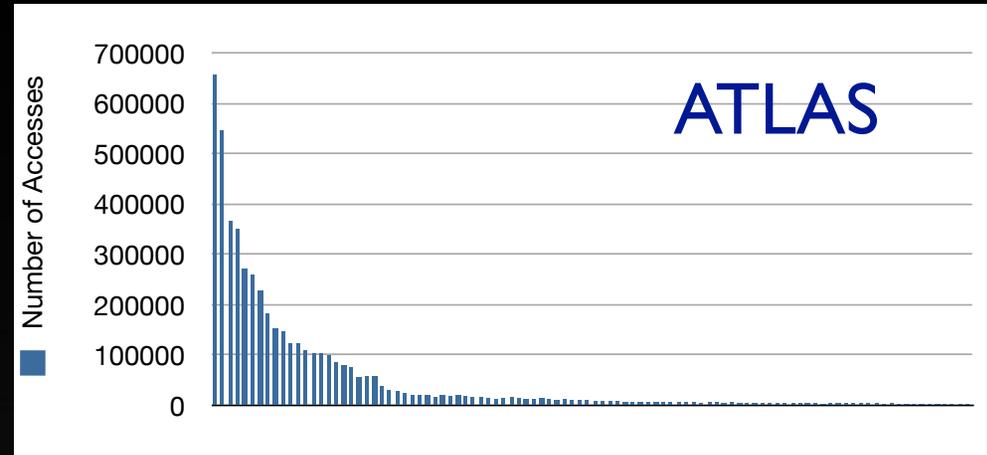
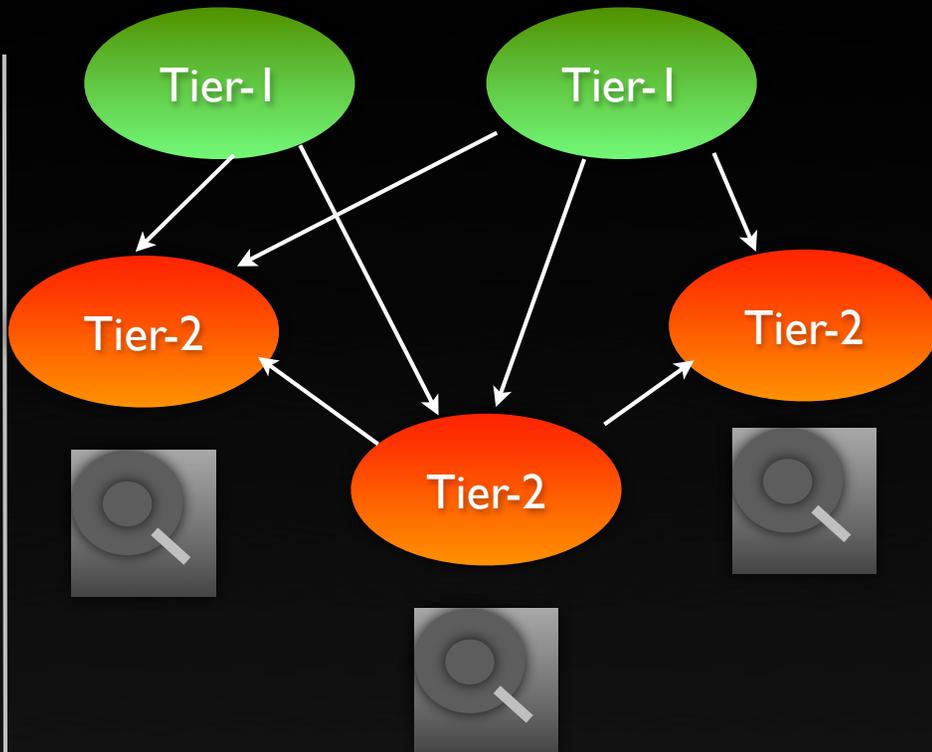
Tier-2

Tier-2

→ Tier-2 to Tier-2 transfers are now similar to Tier-1 to Tier-2 in CMS

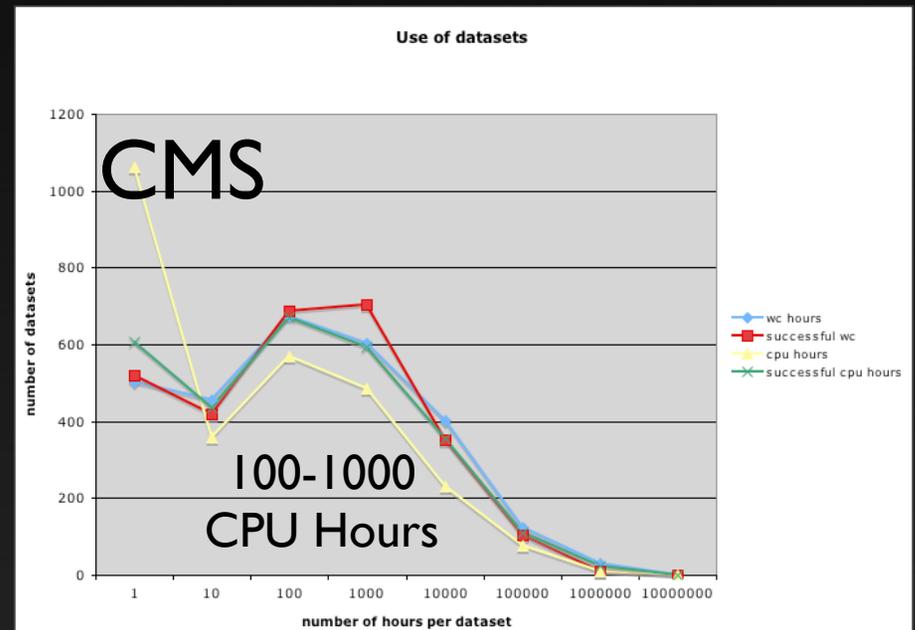


Access

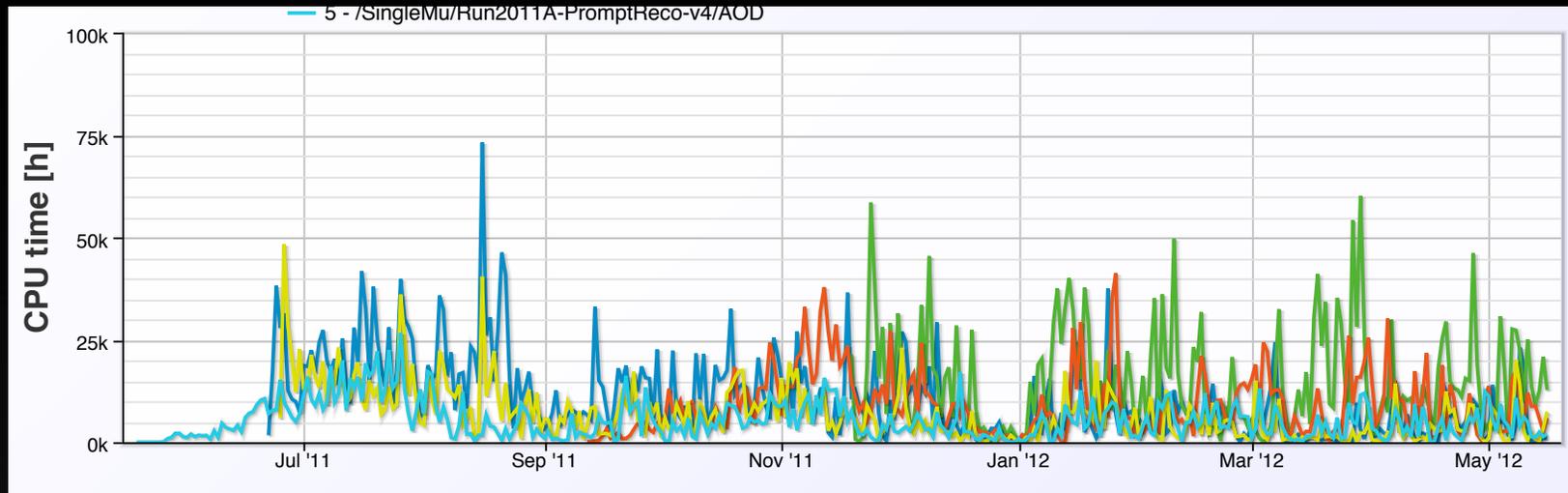


In CMS 30 % of samples subscribed by physicists not used for 3 months during 2010

- ➔ Situation is improved
 - less artificial separation in where data comes from
 - But data is still placed at sites

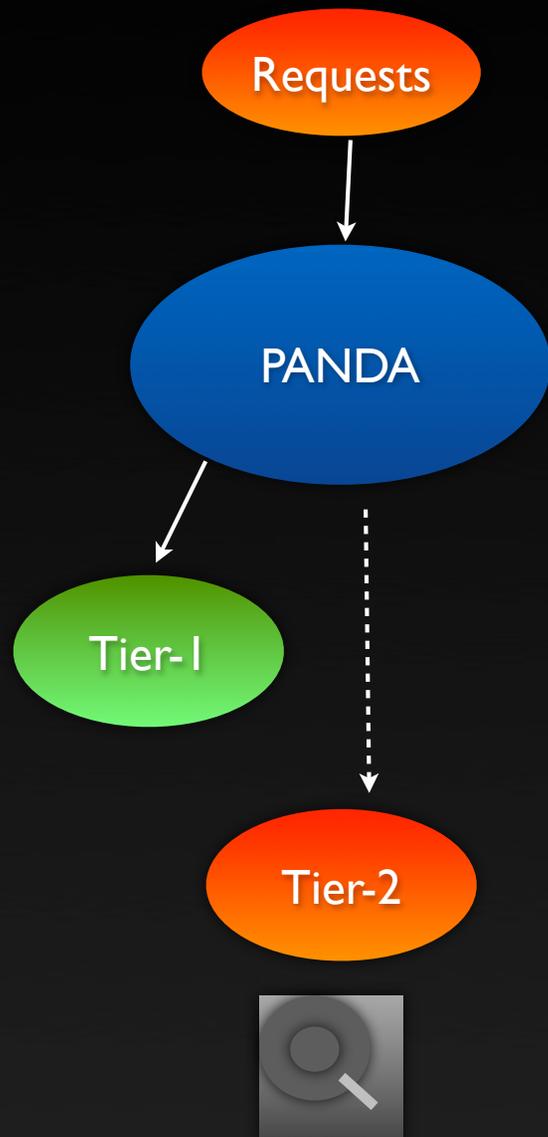


Popularity



- ➔ Services like the Data Popularity Service track all the file accesses and can show what data is accessed and for how long
 - Over a year, popular data stays that way for reasonable long periods of time

Dynamic Data Placement

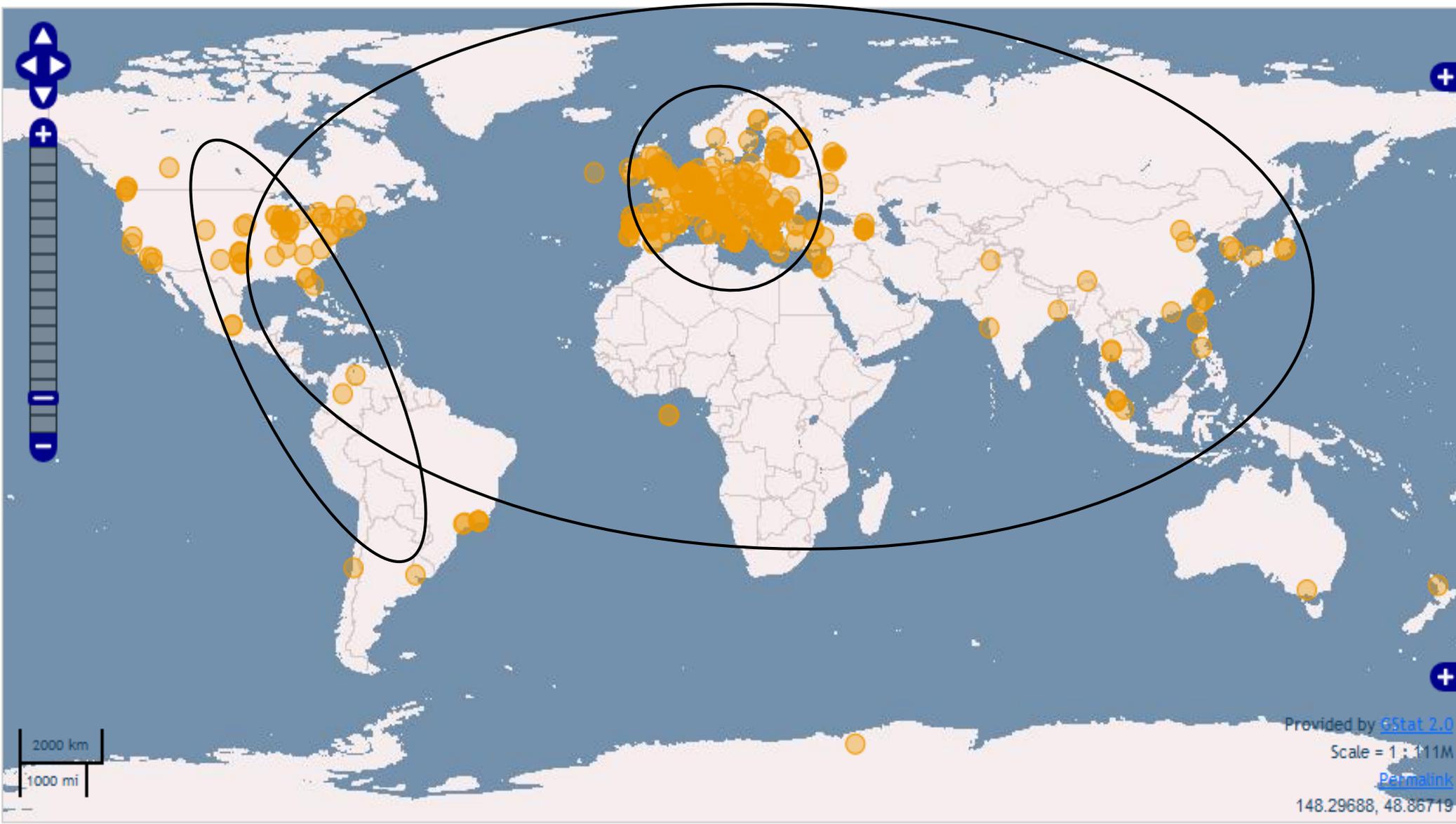


- ➔ ATLAS uses the central queue and popularity to understand how heavily used a dataset is
- Additional copies of the data can be made at sites
- Jobs re-brokered to use them
- ➔ Unused copies are cleaned
- ➔ Reduction in the amount of disk needed

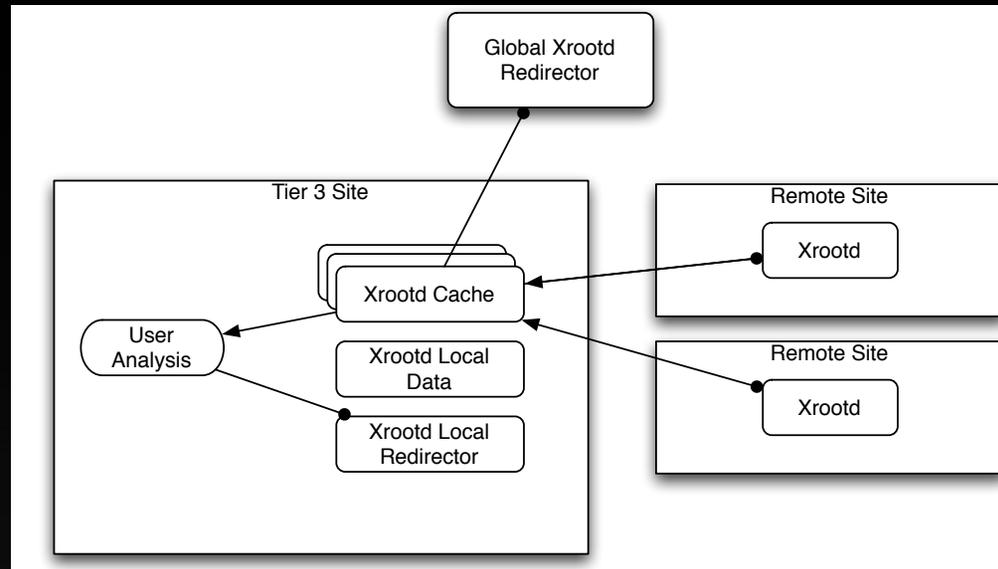
Wide Area Access

- ➔ With 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

Federations



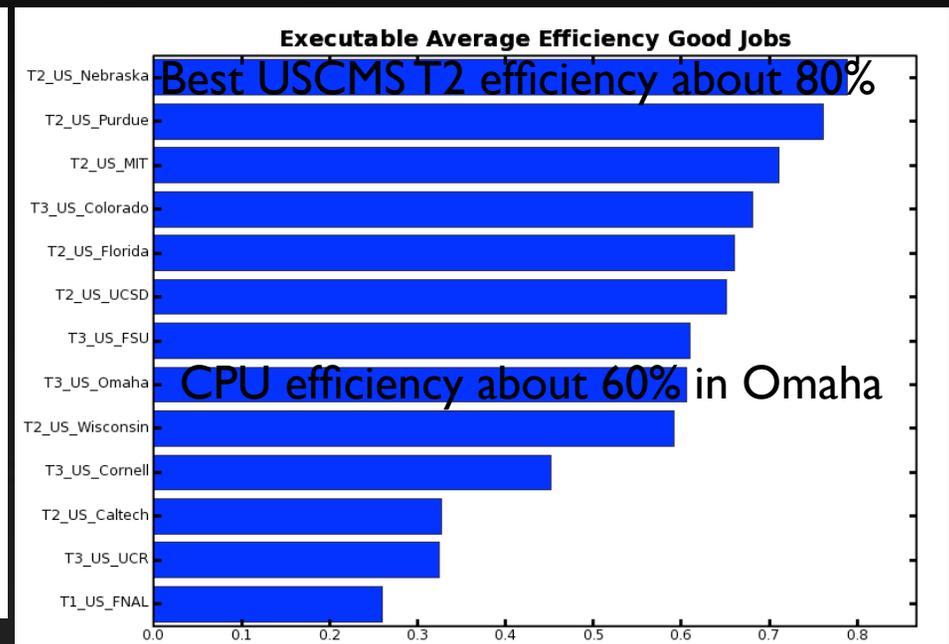
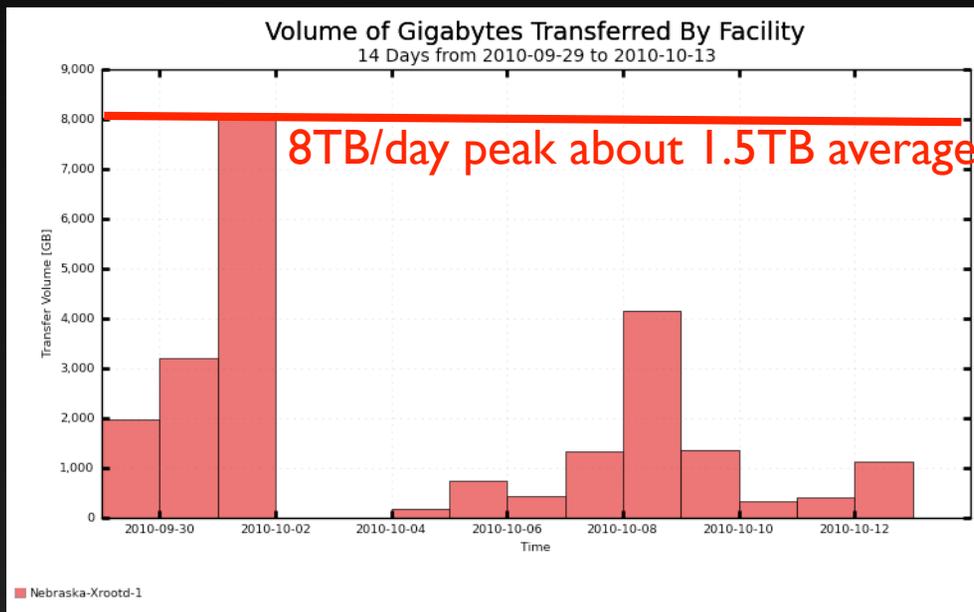
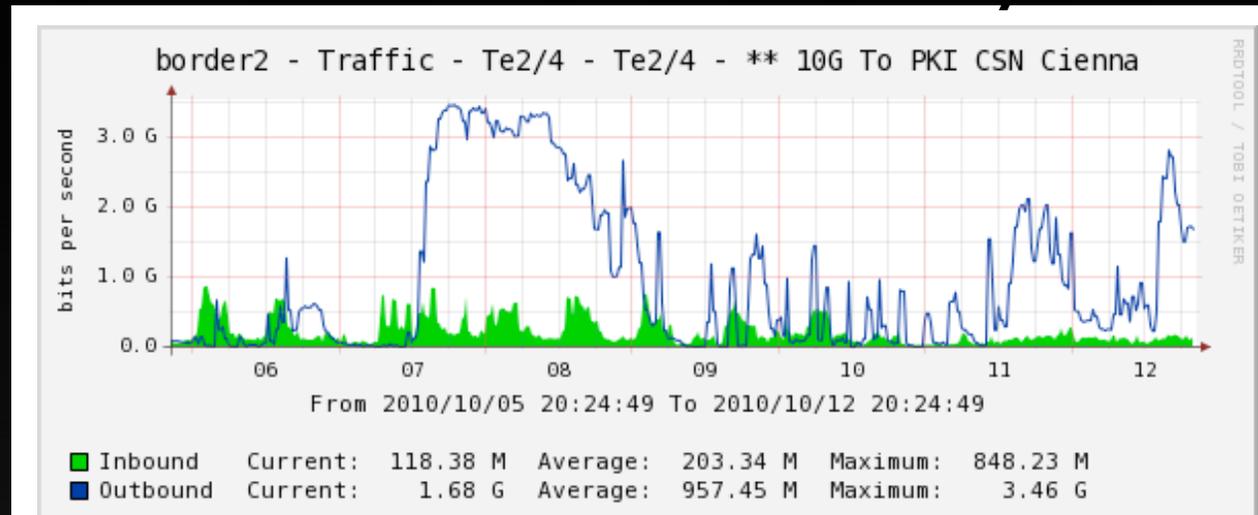
xrootd Demonstrator



- ➔ 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
 - CERN xrootd server to export large EOS pools
 - Tier-3 receiving data runs essentially diskless
- ➔ Similar installation being prepared in ATLAS

Performance

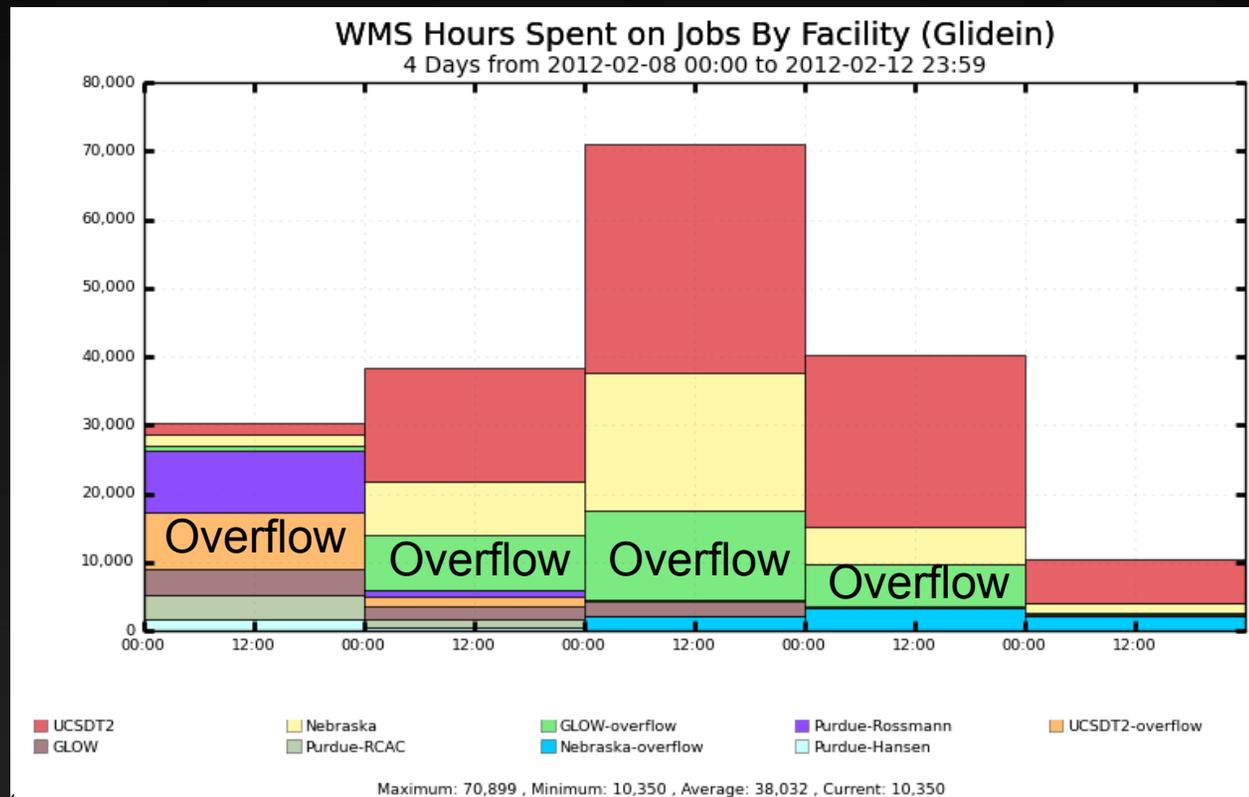
- ➔ This Tier-3 has a 10Gb/s network
- ➔ CPU Efficiency competitive



Failover to sites

- We've introduced the concept of that when we see a long queue at a site we can divert additional requests to a site that doesn't host the data
 - Triggered by operators and only triggered to sites with reasonable network connectivity to the hosting site

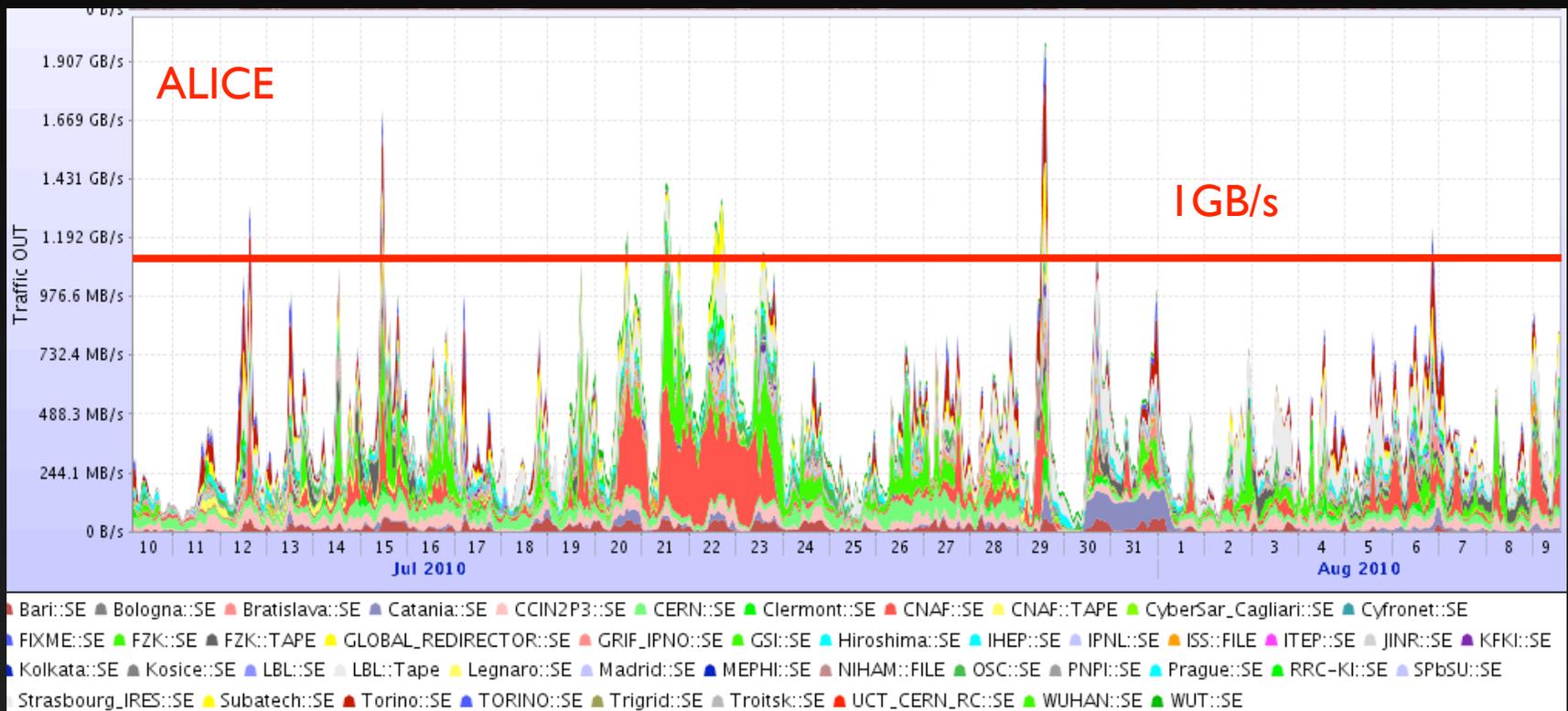
- Job failover to xrootd
- Efficiency loss is measured in 10s of percent and not factors



Networking

➔ ALICE Distributes Data in this way

- Rate from the ALICE Xrootd servers is comparable in peaks to other LHC experiments

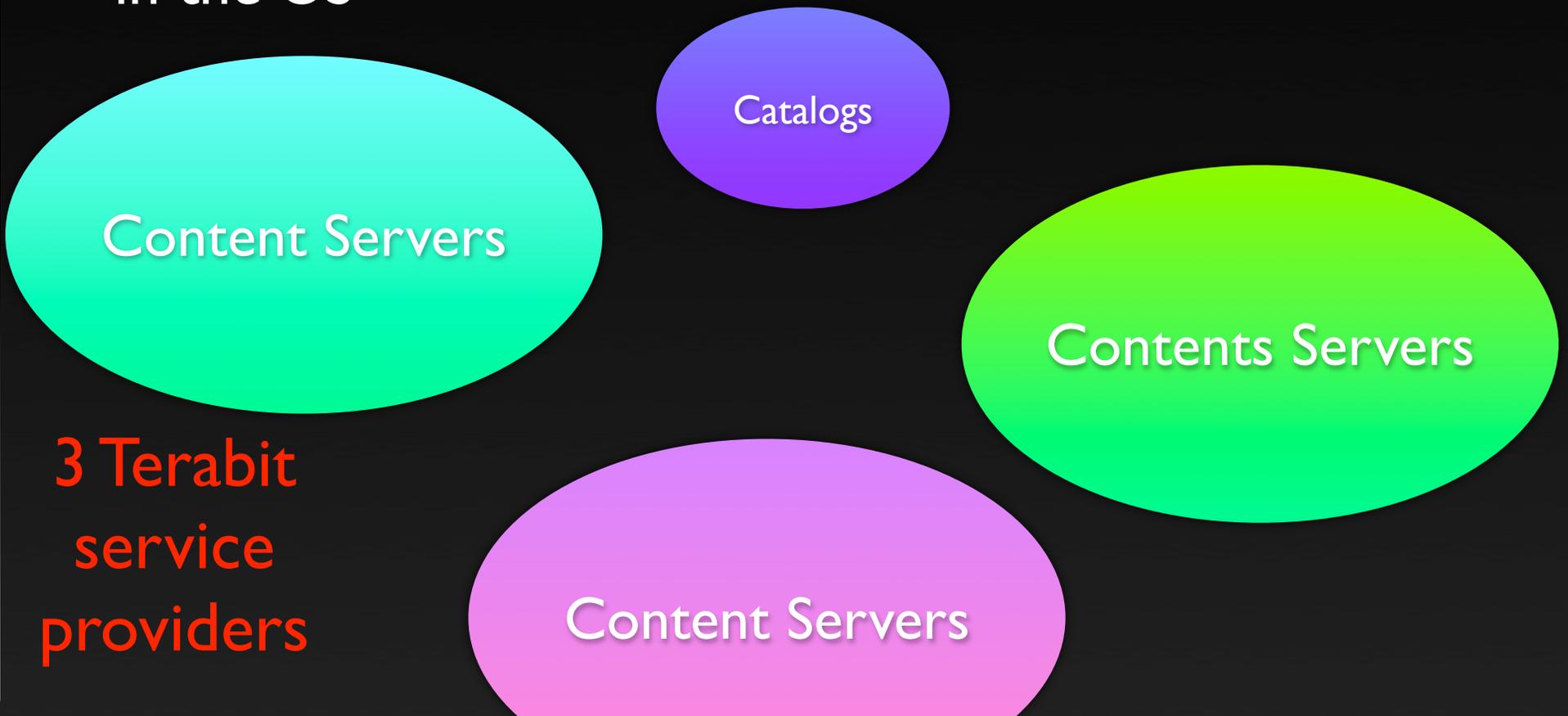


Content Delivery

NETFLIX

→ Why is our problem harder than Netflix?

- Netflix delivers streaming video content to about 20M subscribers
- Routinely quoted as the single largest user of bandwidth in the US



By the numbers

→ We have a smaller number of clients, less distribution, and higher bandwidth per client

	NETFLIX	HEP
Bandwidth per client	1.5Mbit	1MB
Clients	1M*	100k cores
Serving	1.5Tbits	0.8Tbits
Total Data Distributed	12TB	20PB

→ They have much less data



Similar Problems
Not all files
are equally accessed

Forward
Physics

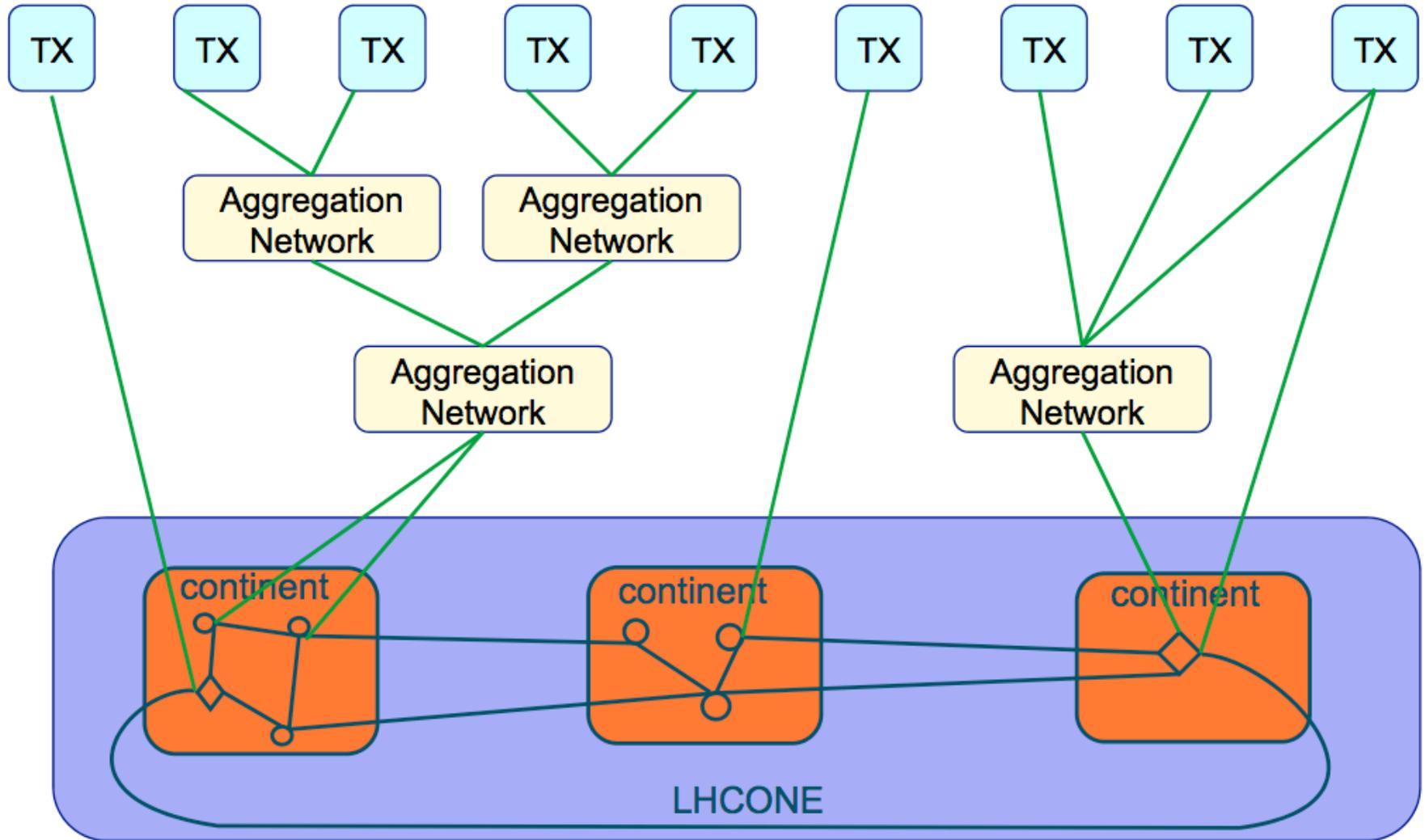
Challenge of HEP

- ➔ High Energy Physics has a lot of data in a highly distributed environment
 - Hard to make many multiple static copies
 - Need to be able to make dynamic replicas and clean up
 - Need to access data over long distances
- ➔ Trying to make networking more predictable
 - Enter LHCOne

LHCONE in a Nutshell

- ➔ LHCONE was born (out the 2010 transatlantic workshop at CERN) to address two main issues:
 - To ensure that the services to the science community maintain their quality and reliability
 - To protect existing R&E infrastructures against overuse by our traffic
- ➔ LHCONE is expected to
 - Provide some guarantees of performance
 - ◆ Large data flows across managed bandwidth that would provide better determinism than shared IP networks
 - ◆ Segregation from competing traffic flows
 - ◆ Use all available resources, especially transatlantic
 - ◆ Provide Traffic Engineering and flow management capability
 - Leverage investments being made in advanced networking

LHCONE Initial



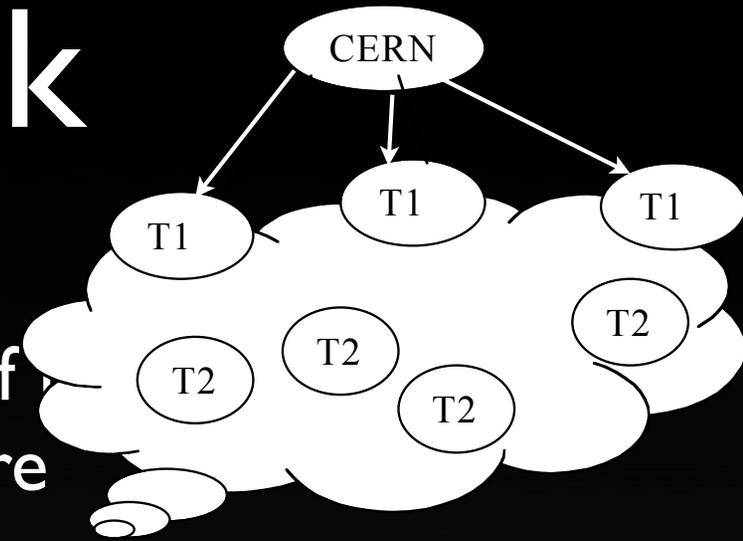
LHCOne Activities

- ➔ Start with virtual routing and forwarding (VRF)
 - To bring up a system with more predictable networking
 - Improve the monitoring
- ➔ Migrate toward more advanced technologies
 - Dynamic circuits
 - Scheduled bandwidth

Changes

- ➔ New model has less structure
 - More options for where data comes from
 - More flexibility in where activities happen
- ➔ This lack of structure places more expectations on the support services like networking and data management
 - Developing more advanced service for management and transport
 - Network has been very reliable. Programs like LHCOne try to maintain this

Outlook



- ➔ More flexible and dynamic use of resources available will make more efficient use of the resources
- ➔ All the actions trying to ensure that we don't make artificial separation
- ➔ Should put us in a better situation to make use of Computing Services we don't control
 - Clouds and Opportunistic Computing