

# Computing for LHC: First Run and Beyond

Ian Fisk

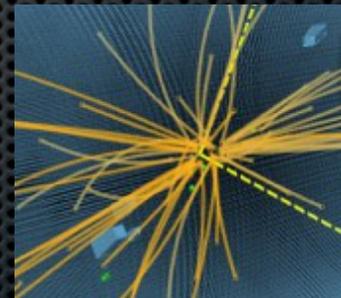
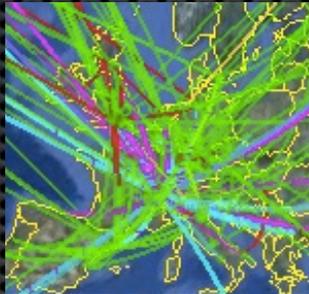
October 10, 2012

# About Me

- ✦ I work here
  - ✦ 10 year anniversary is in April
  - ✦ I've been stationed at CERN as CMS Computing Coordinator since January of 2010
- ✦ As CMS Computing Coordinator I am responsible for
  - ✦ Data Loss, Data Access Problems, Failed Processing Requests, Late Samples, Documentation, Computing Security, Accidental Data Deletion, Missing functionality in Computing Services

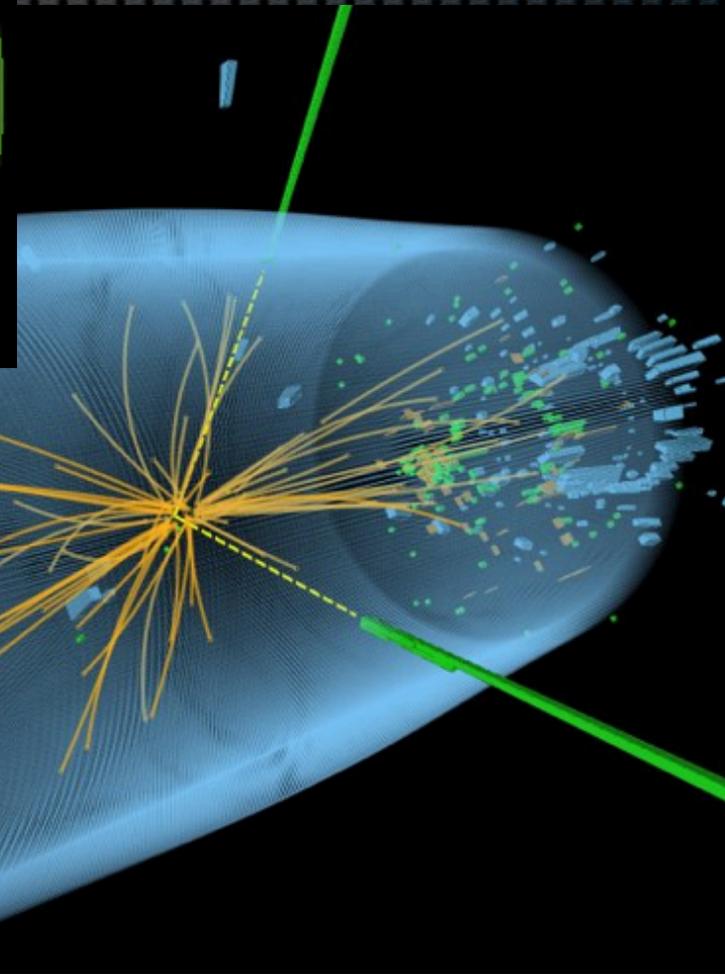
# Final Steps

- ✦ Software and Computing is the final in a long series to realize the physics potential of the experiment
  - ✦ Storage and Serve the data
  - ✦ Reconstruct the physics objects
  - ✦ Analyze the events
- ✦ As the environment has become more complex and demanding, computing and software have had to become faster and more capable

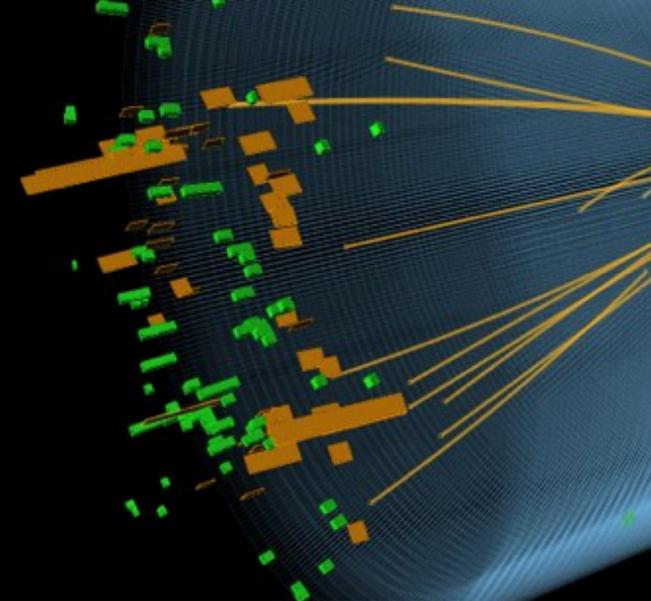




✦ You need this



✦ To get this



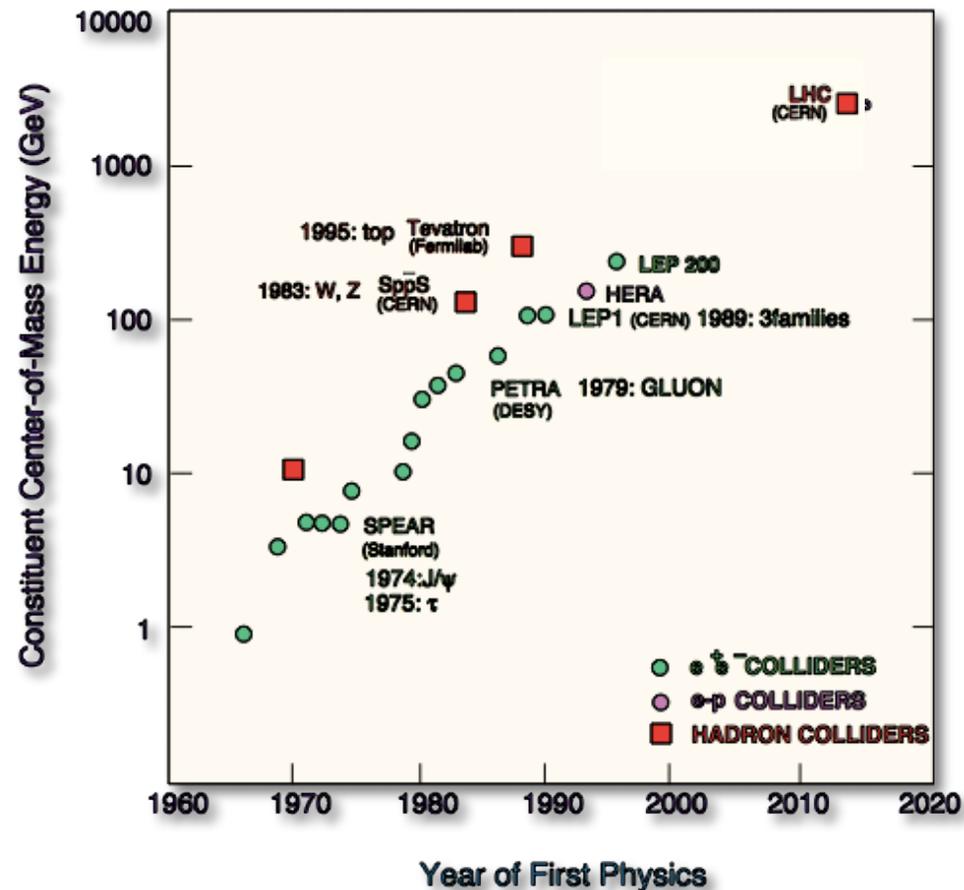
# Progress

- About 15 years ago the field switched to Linux
  - Large number of clustered inexpensive commodity machines instead of expensive single systems
- About 10 years ago the MONARC distributed Computing model was developed
  - Distributed computing around a common goal
- About 5 years ago the distributed Computing models were deployed
  - Lots of testing and work
- Recent Progress?
  - Computing Scale and Networking

# The Scale of the Problem

Looking at Tevatron a few years into Run2 and LHC on year 3

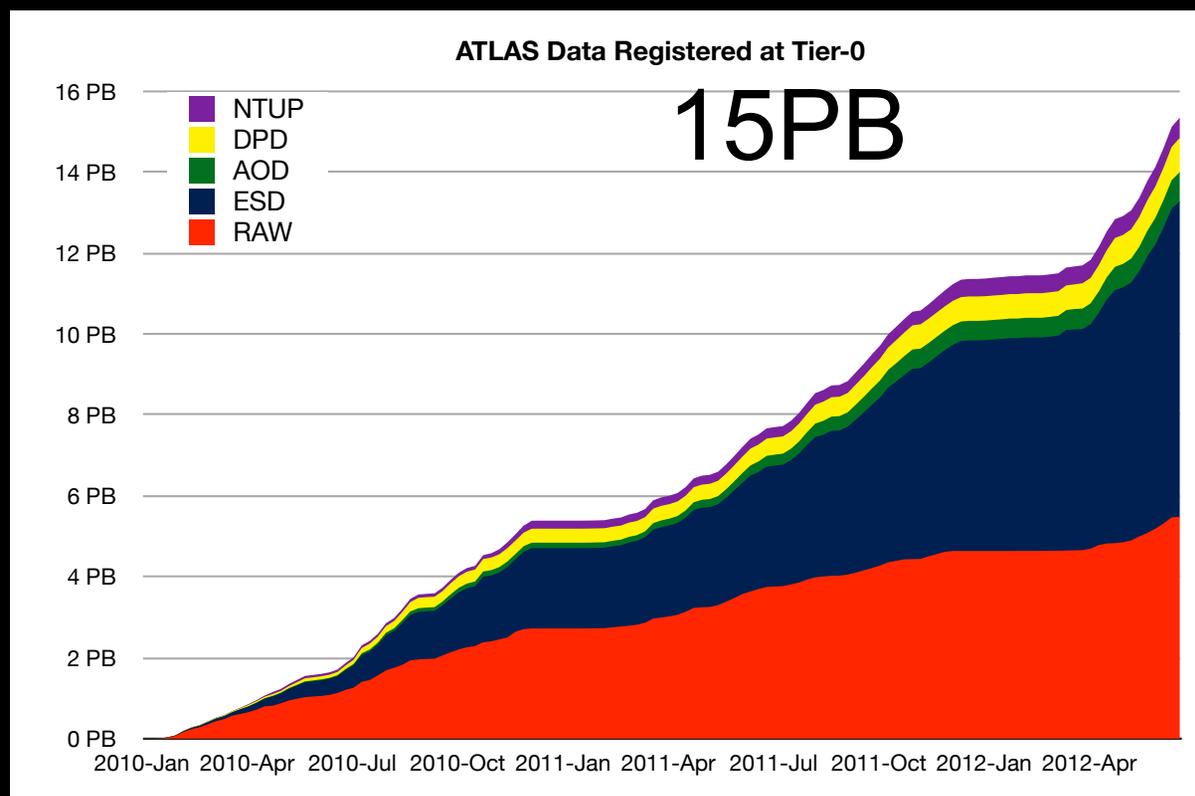
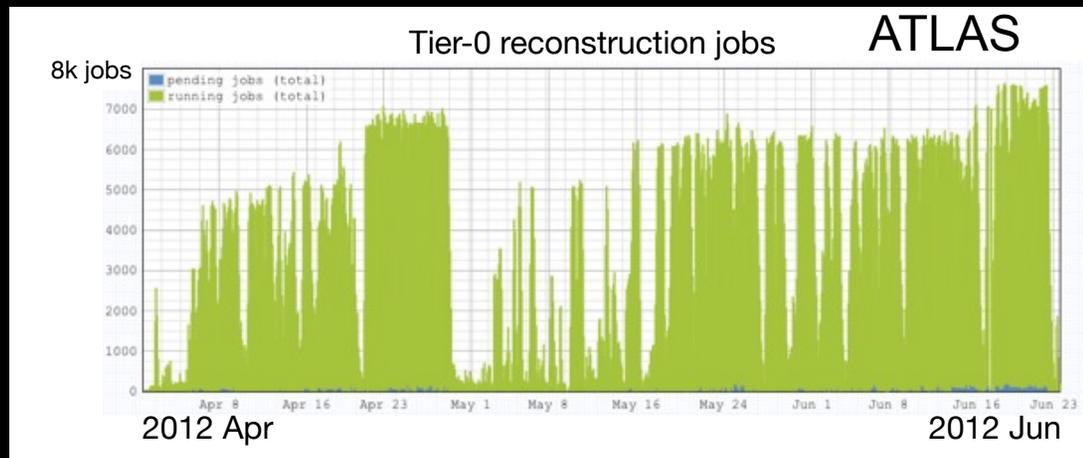
|                 | Tevatron                       | LHC                                   |
|-----------------|--------------------------------|---------------------------------------|
| Trigger         | 50Hz                           | ATLAS 500Hz<br>CMS 350Hz<br>LHCb 2kHz |
| RAW Event Size  | 150k                           | ATLAS 1.5MB<br>CMS 0.5MB              |
| RECO Event Size | 150k                           | ATLAS 2MB<br>CMS 1MB                  |
| Reco Speed      | 1-2 seconds on CPU of the time | 10s on CPU of the time                |



Roughly a factor of 10 in the relevant quantities

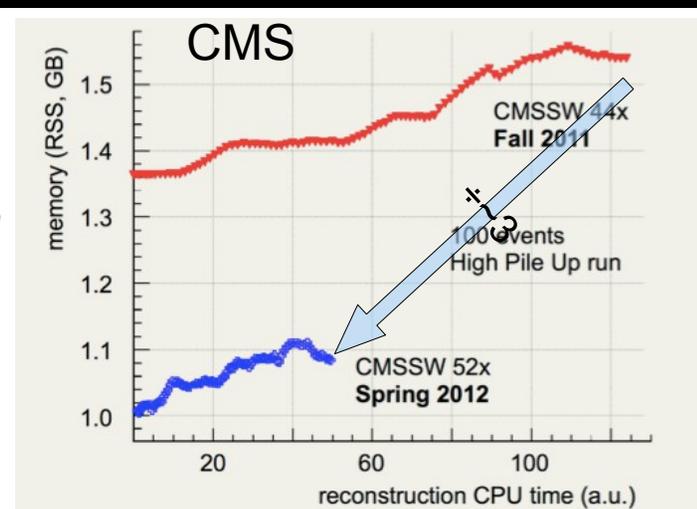
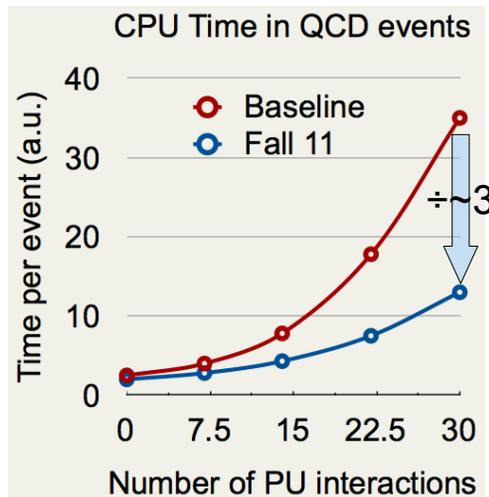
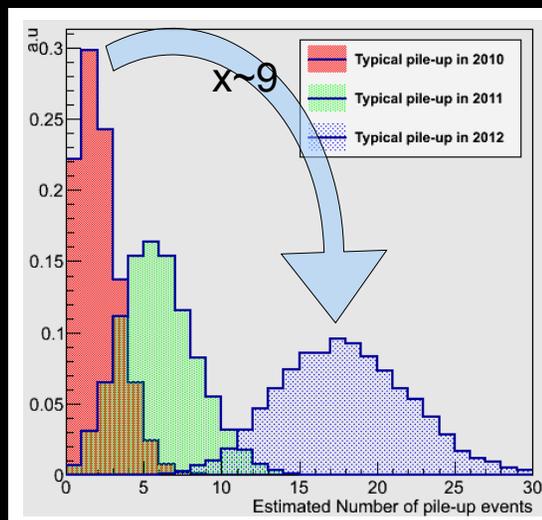
# Enormous Data Volumes

- Higher trigger rates, larger events sizes and complicated events



# Complicated Environments

- The LHC is running at a higher number of interactions per crossing than design
  - Experiments have high trigger rates due to the interesting physics
  - Has required capacity and efficiency

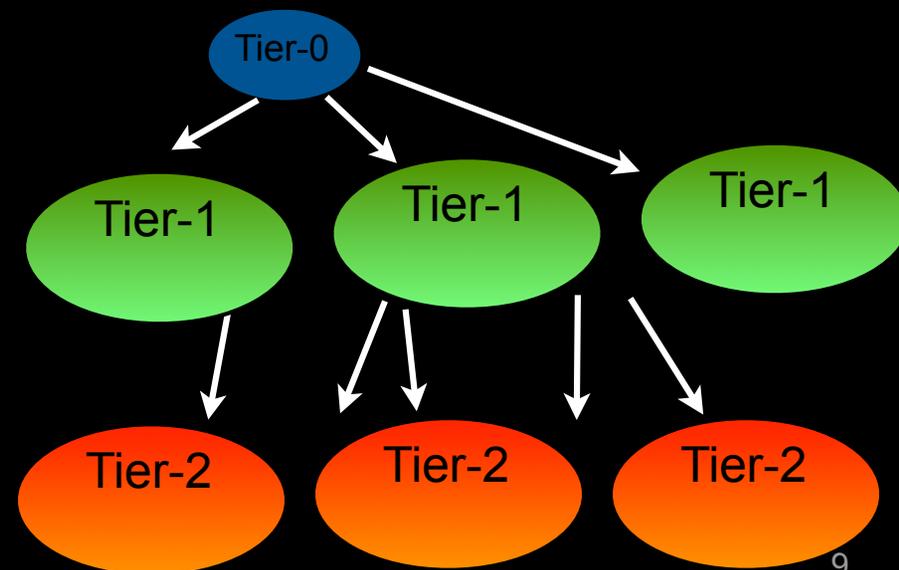
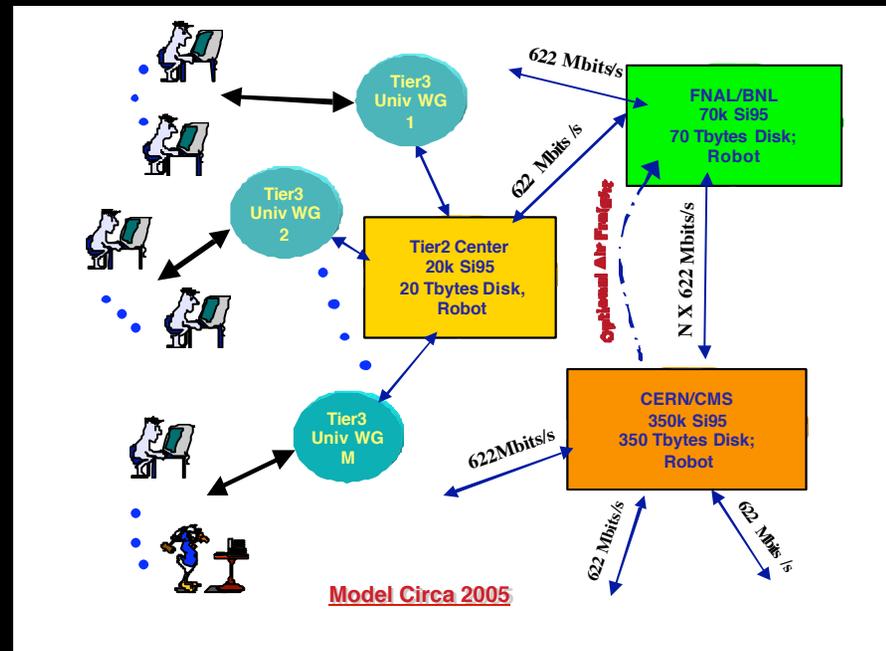


# Distributed Computing

- Computing models are based roughly on the **MONARC** model

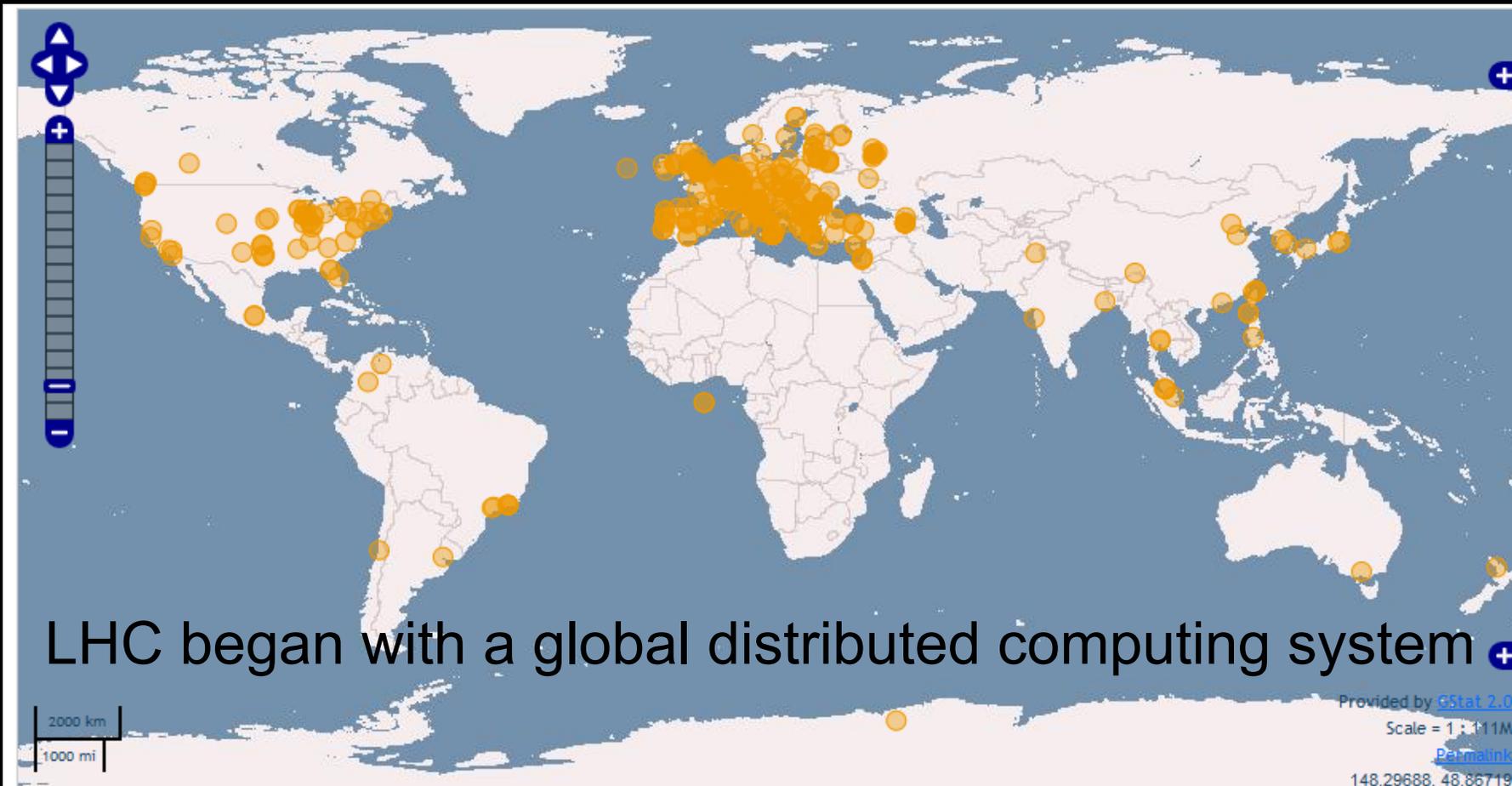
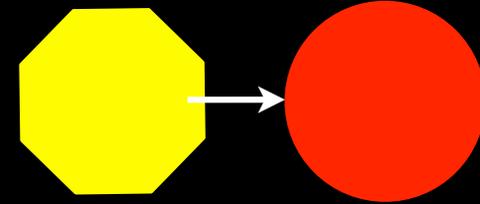
- Developed more than a decade ago
- Foresaw Tiered Computing Facilities to meet the needs of the LHC Experiments

- Assumes poor networking
- Hierarchy of functionality and capability

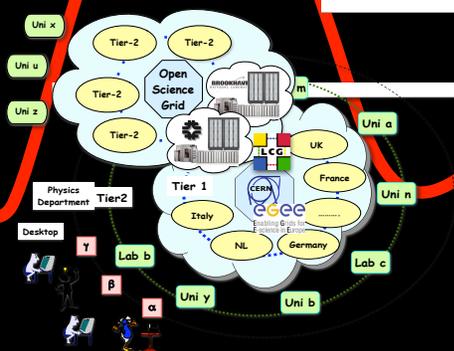
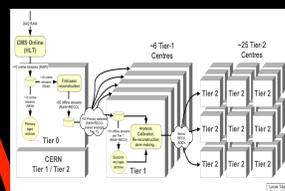
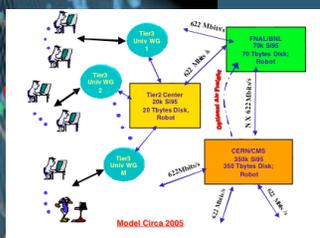


# Distributing Computing at the Beginning

- Before LHC most of the Computing Capacity was located at the experiment at the beginning
  - Most experiments evolved and added distributed computing later



# Evolution



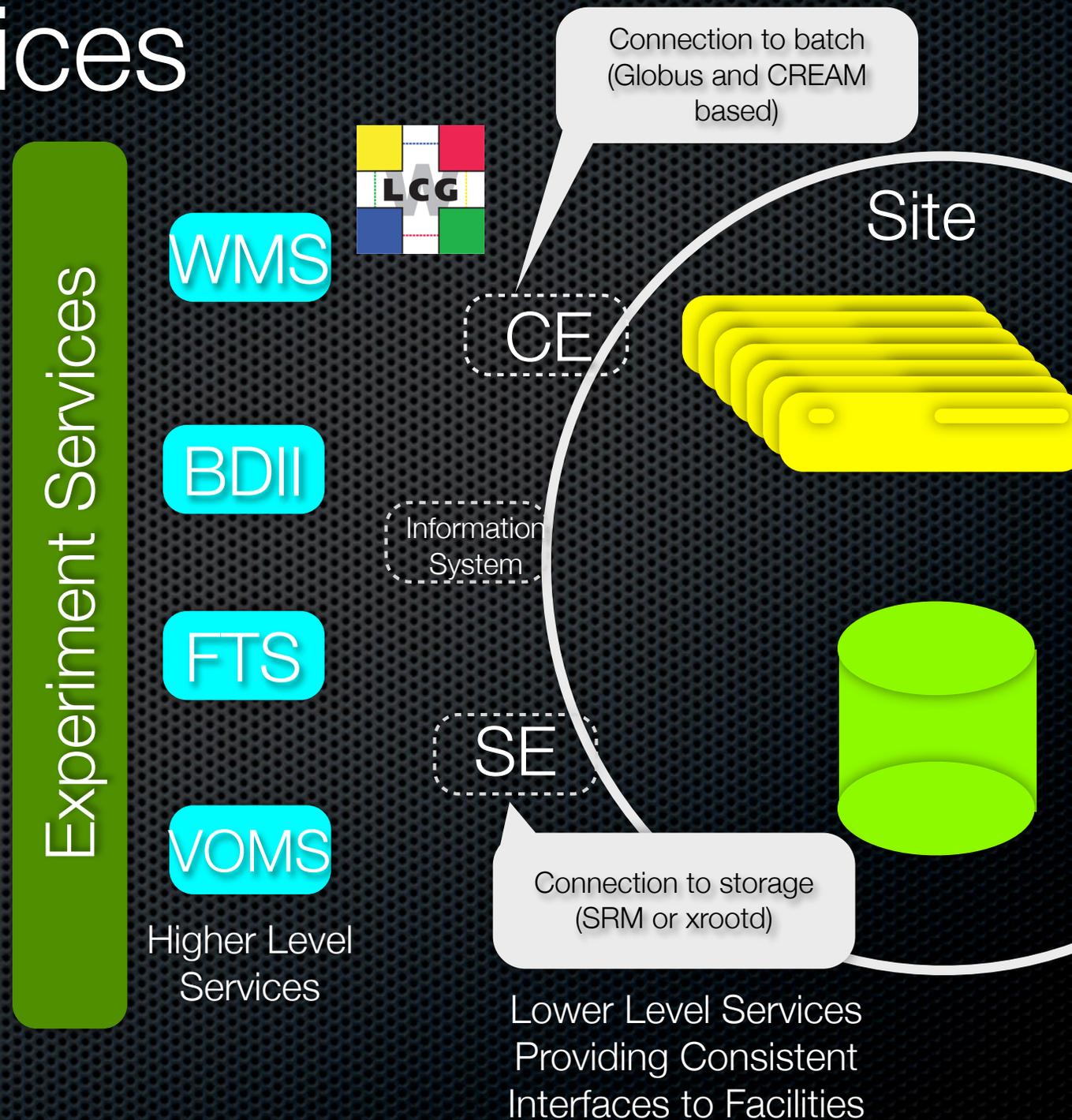
ALICE  
Remote  
Access

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

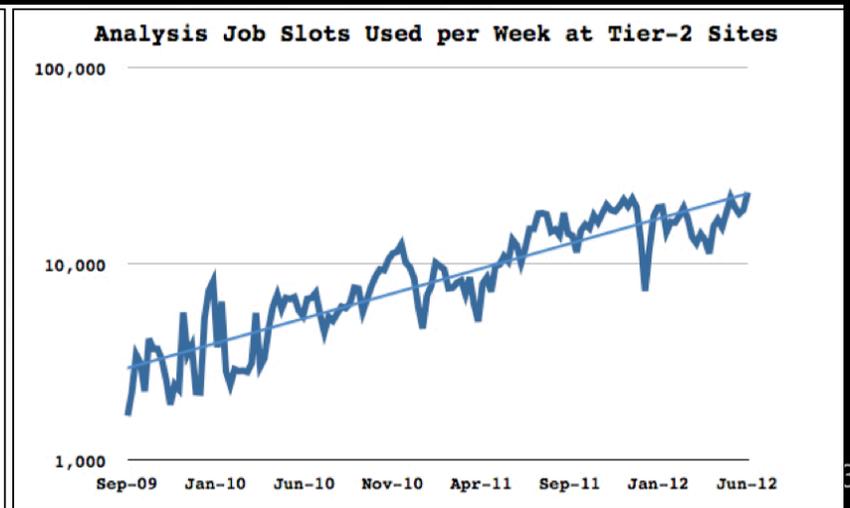
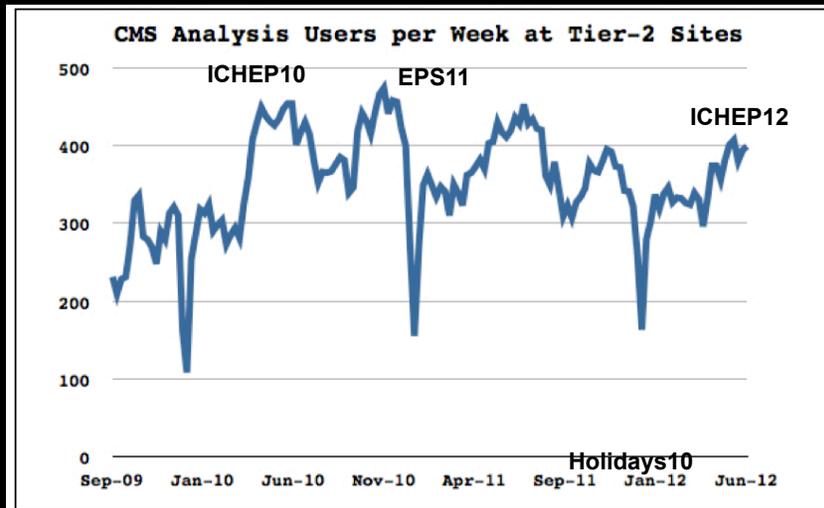
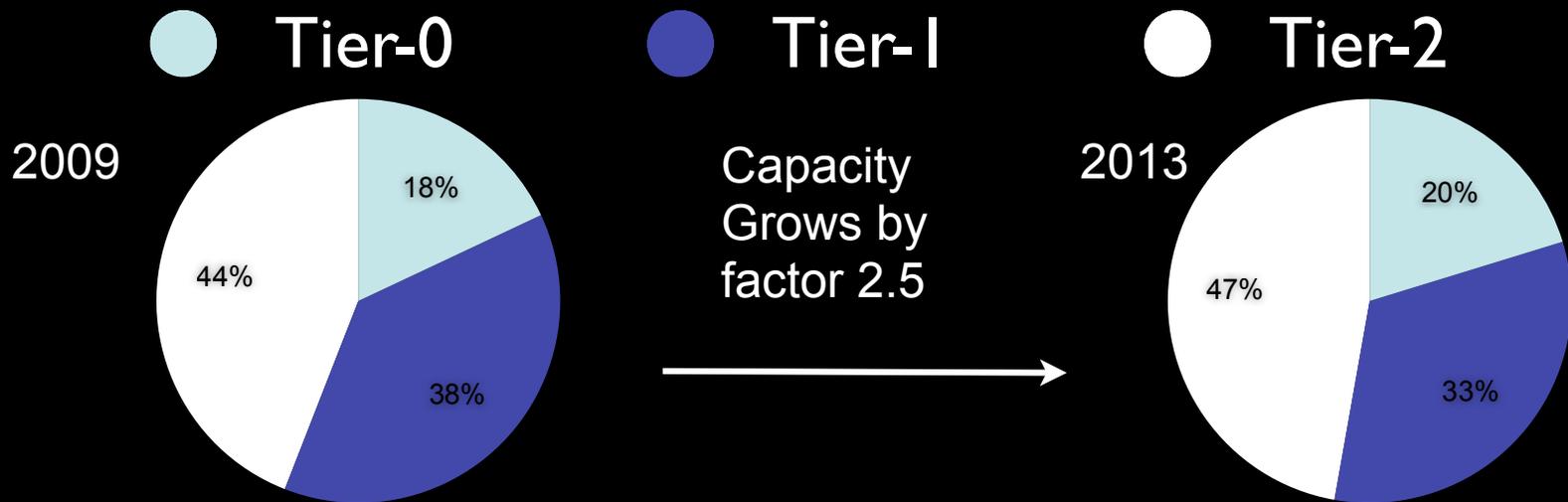
# Grid Services

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



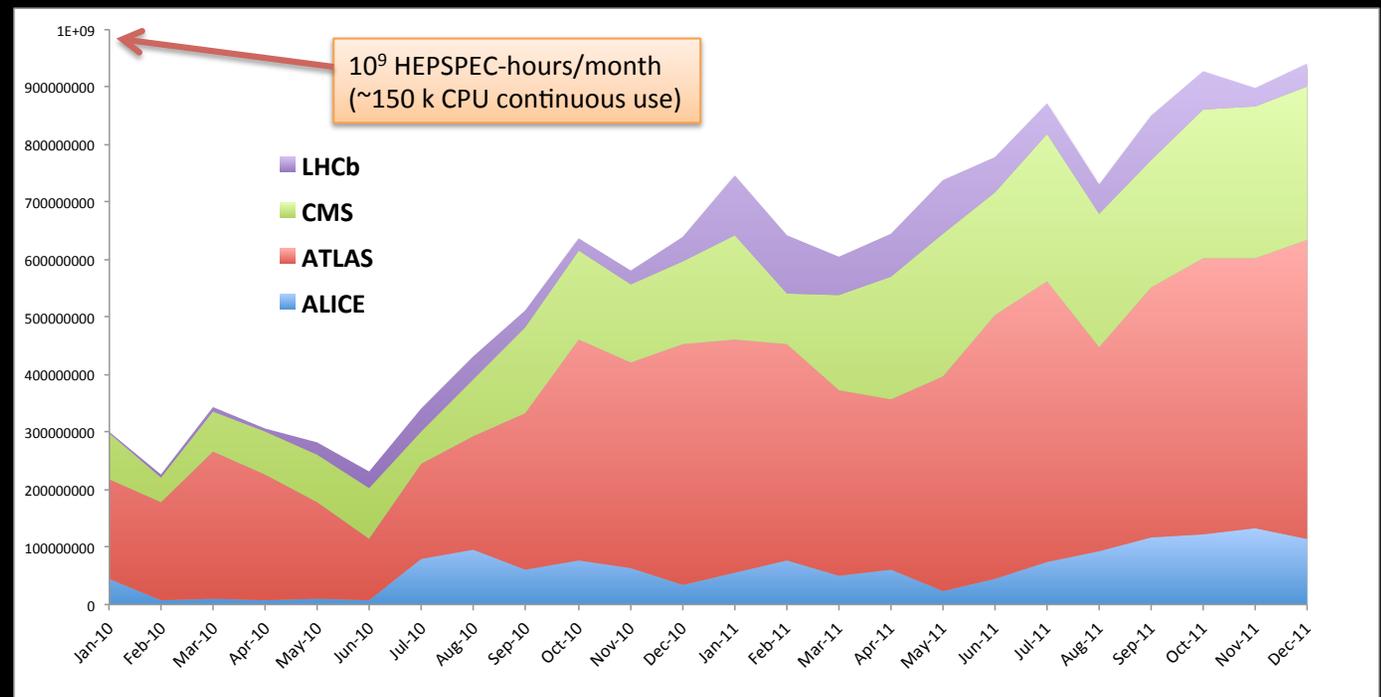
# Successes

- When the WLCG started there was a lot of concern about the viability of the Tier-2 Program
  - A university based grid of often small sites



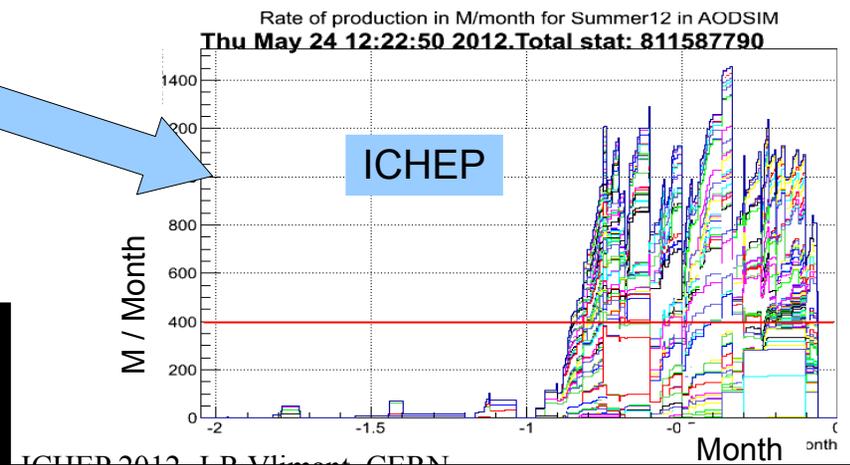
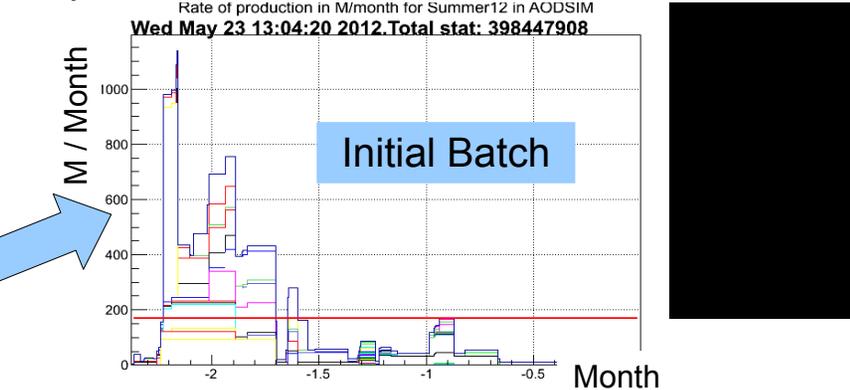
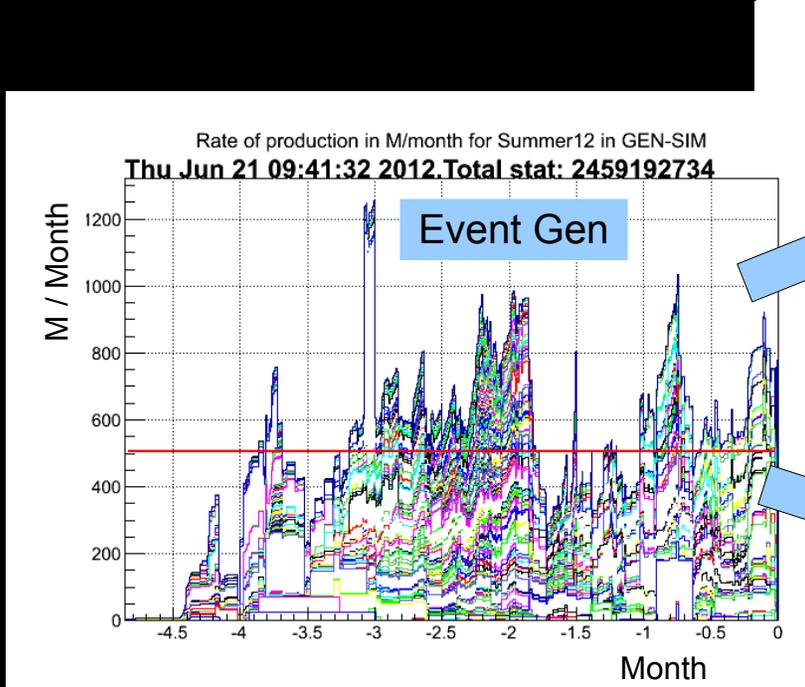
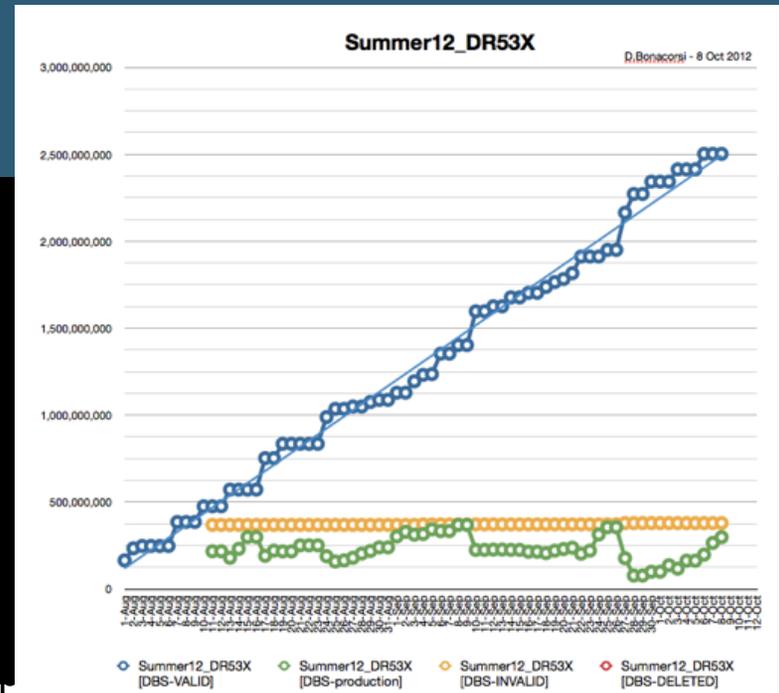
# Progress: Scale of the Solution

- Progress in distributed computing and evolution of computing capacity
  - WLCG processes 1.5M jobs on the grid per day
  - More than 150PB of both disk and tape



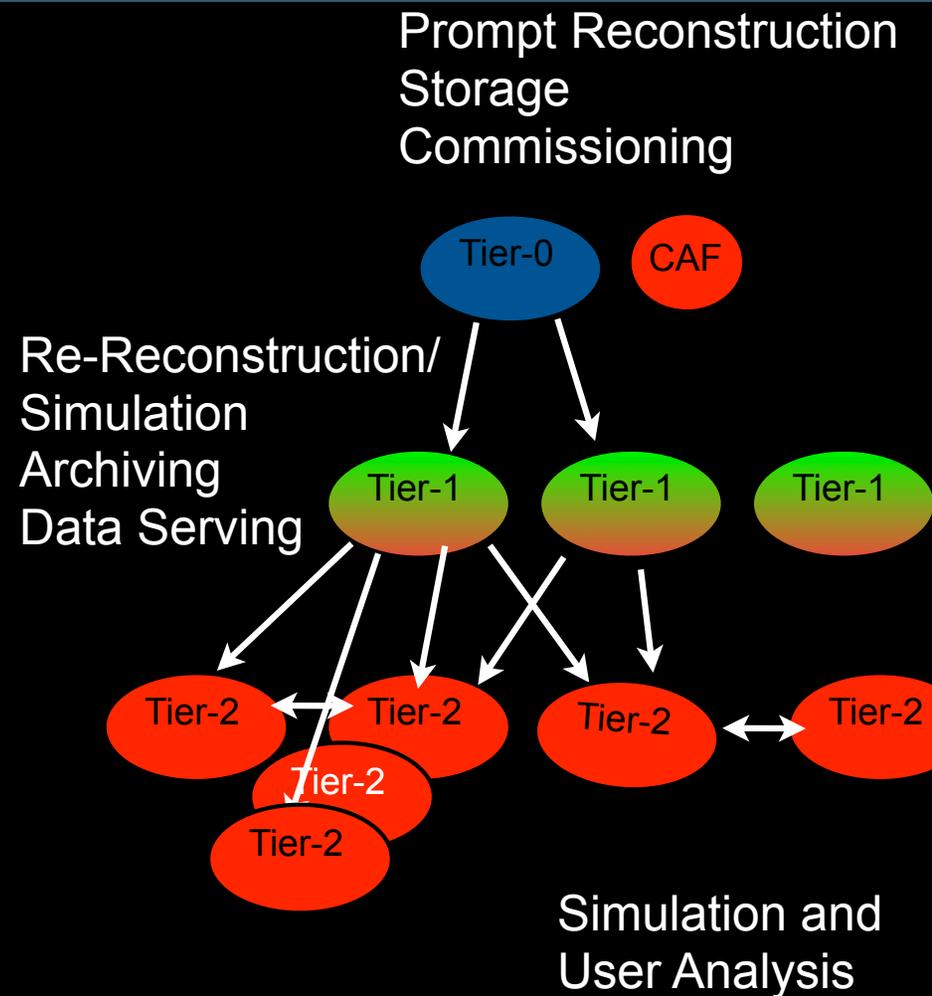
# Simulation

- Improved scale translates into
  - More simulation
  - Faster reprocessing of



# Moving Forward

- 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



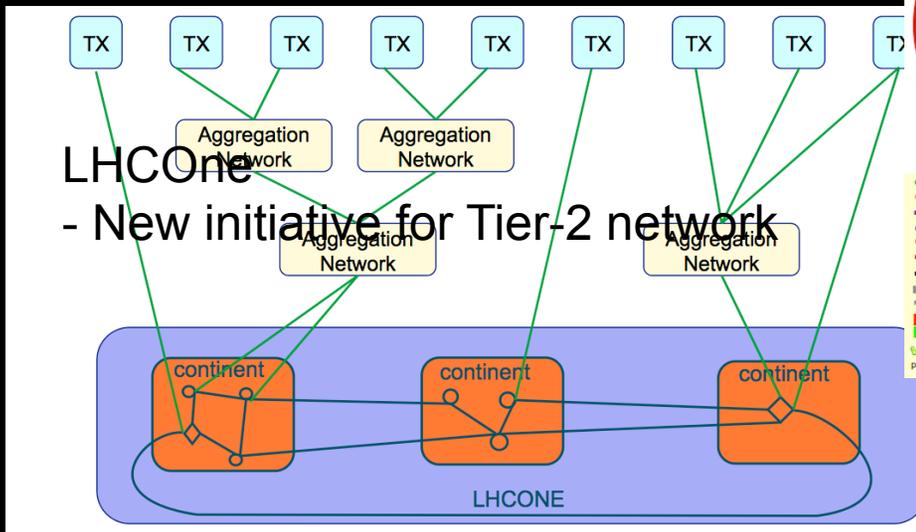
- ▶ Model changes have been an evolution
- ▶ Not all experiments have emphasized the same things
- ▶ Each pushing farther in particular directions

# Progress Networking

- One of the areas of progress has been better use of wide area networking to move data and to make efficient use of the distributed computing
  - Limited dedicated network
  - Much shared use R&E networking

## LHCOPN

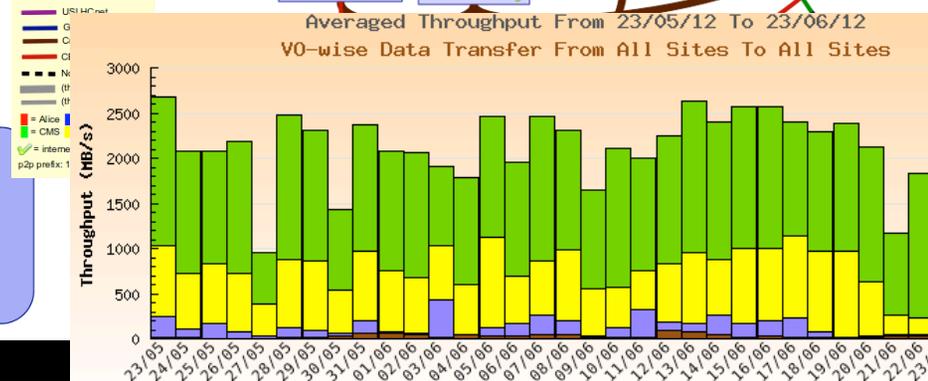
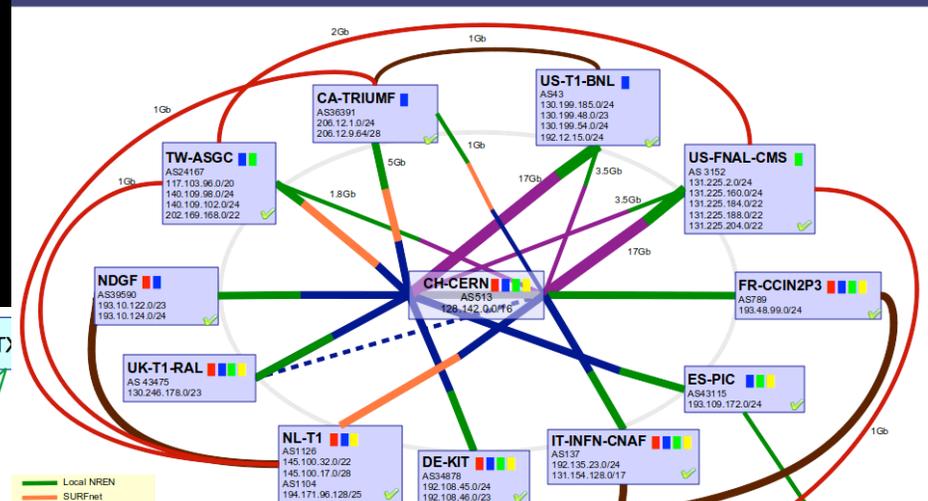
- Dedicated resource T0->T1 and T1 to T1
- LHCOne to Tier-2s



LHCOne

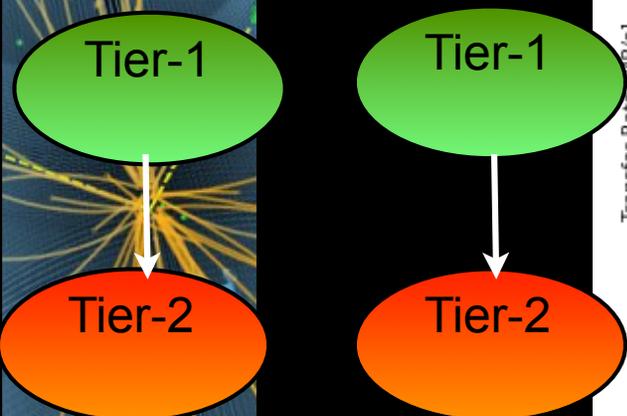
- New initiative for Tier-2 network

## LHCOPN – current status

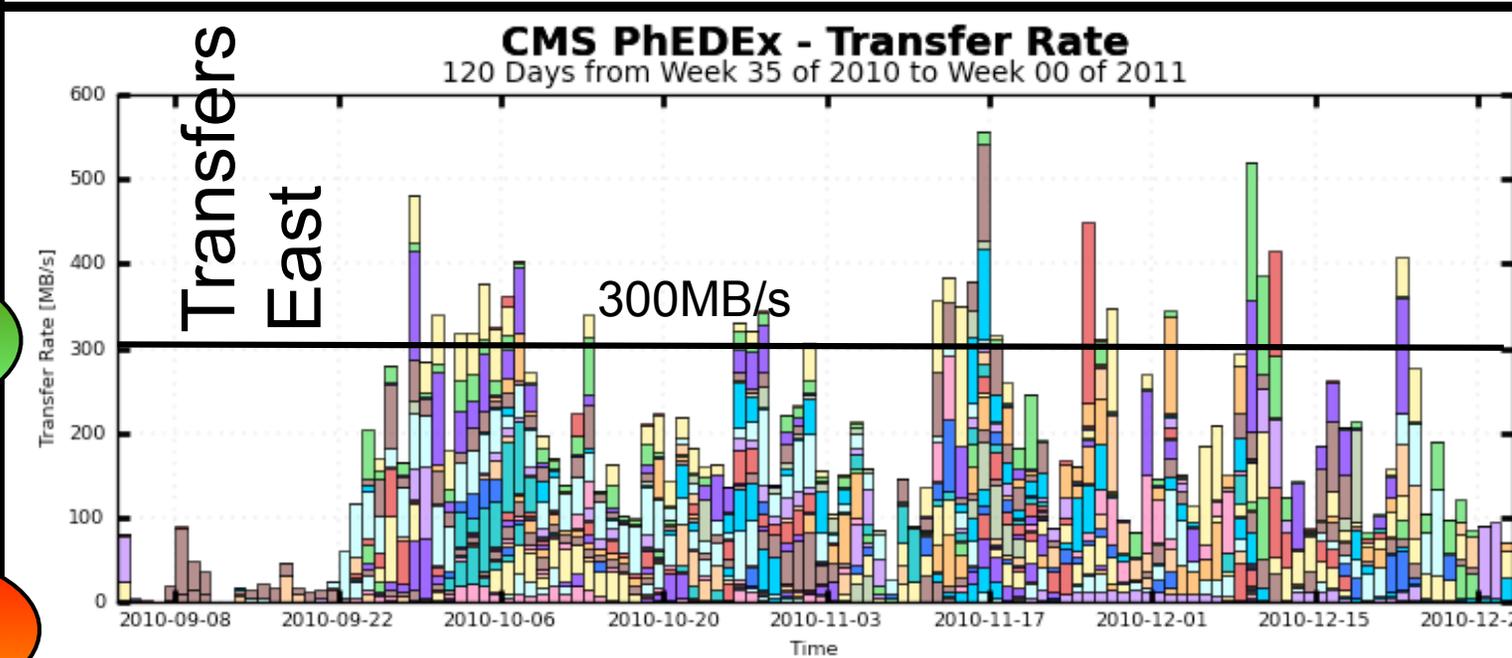
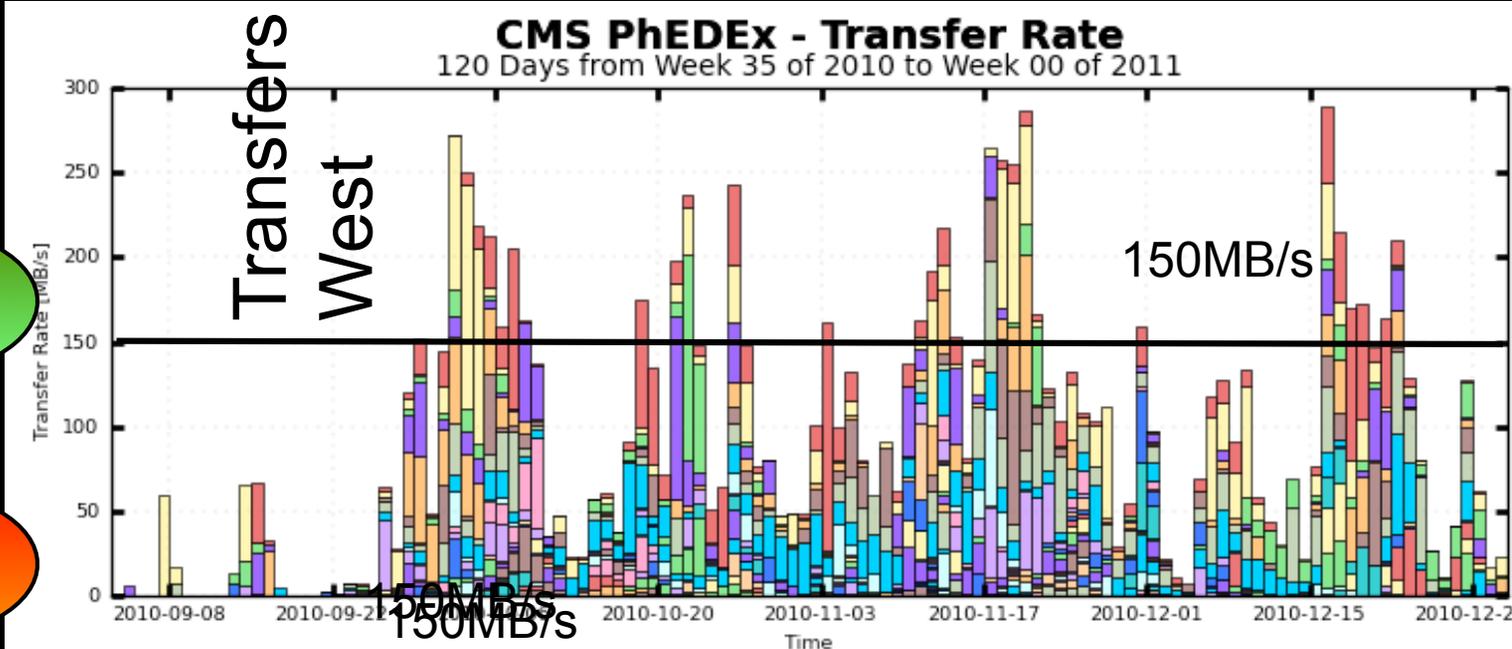
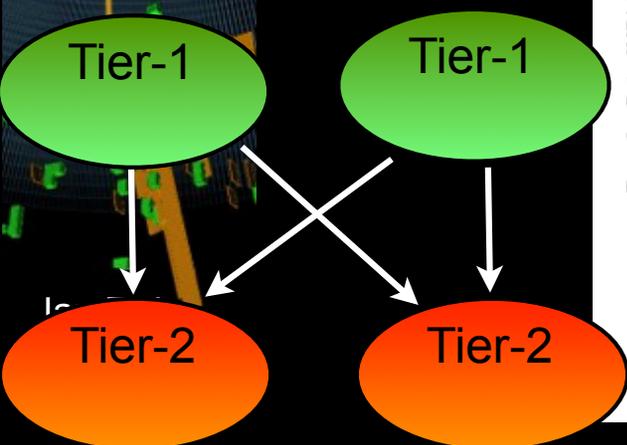


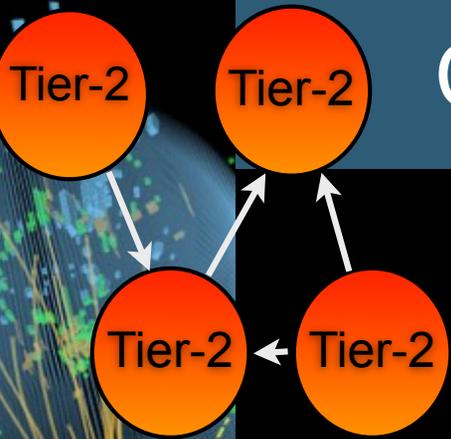
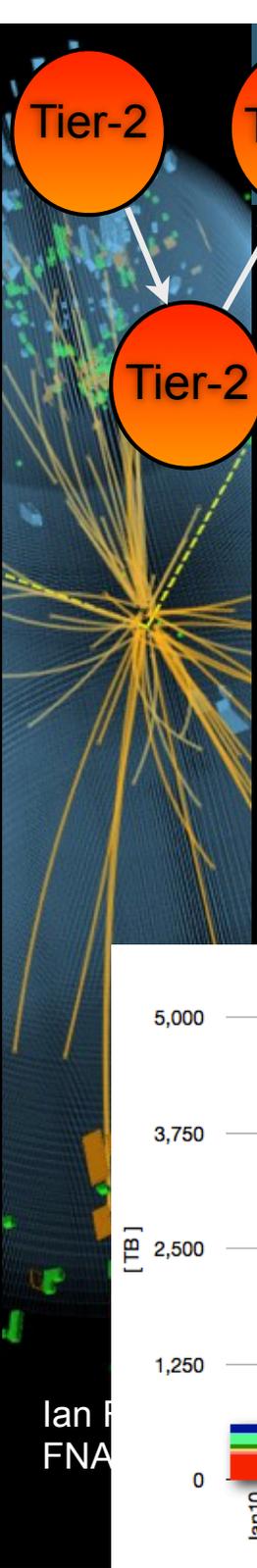
# Mesh Transfers

– Change from



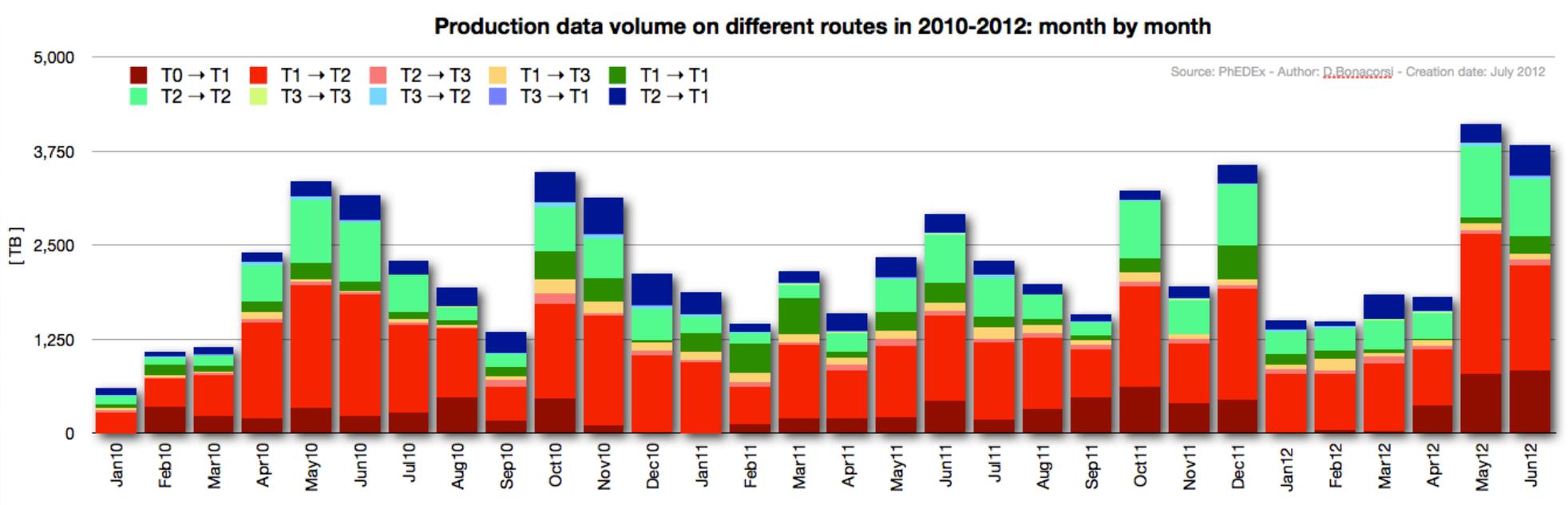
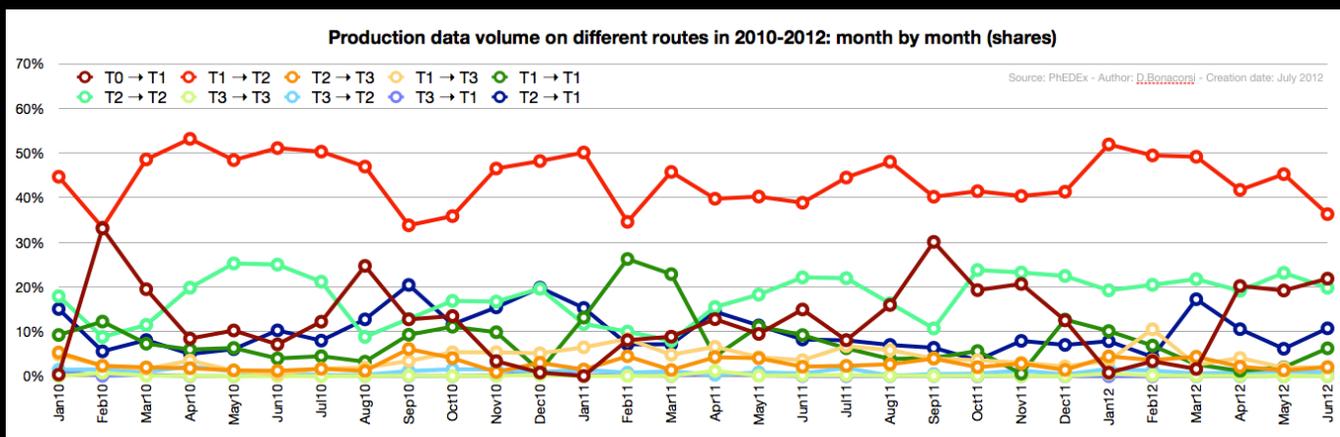
– To





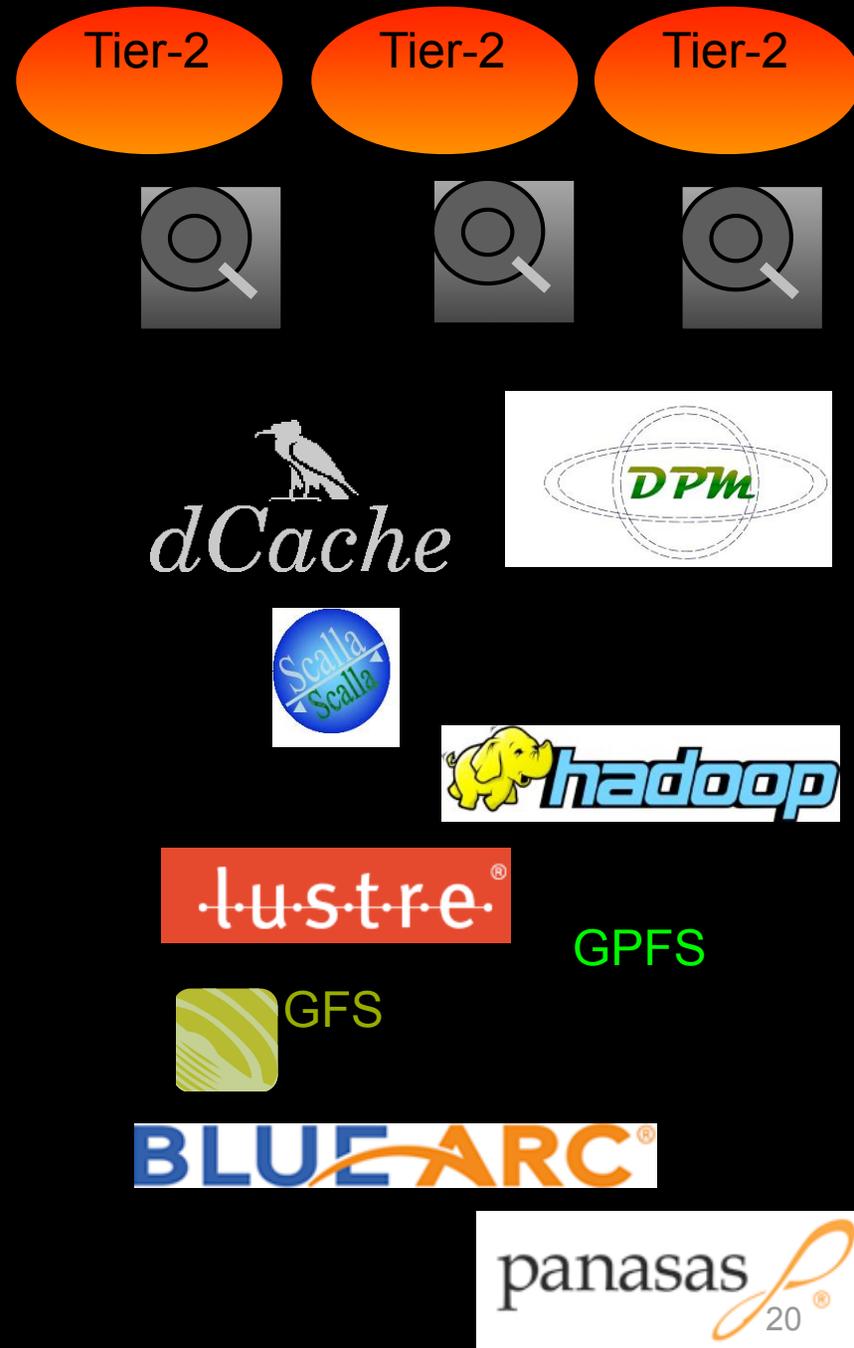
# Completing the Mesh

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

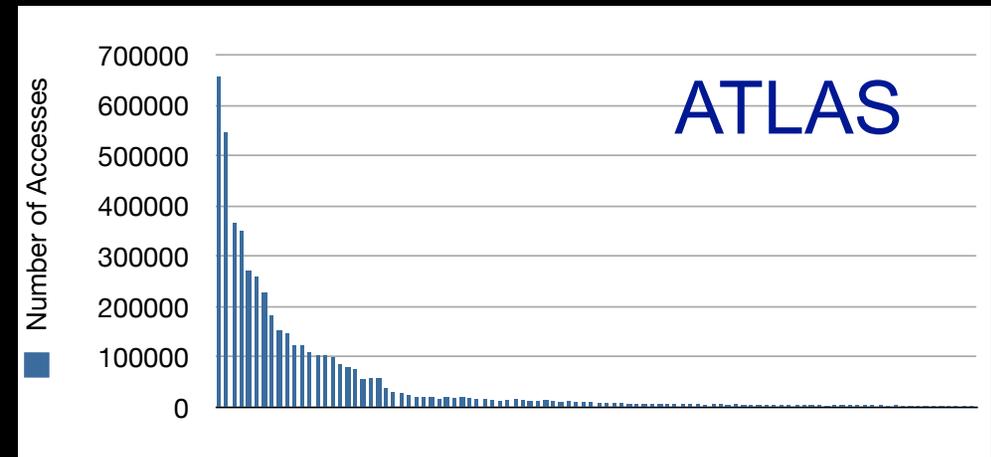
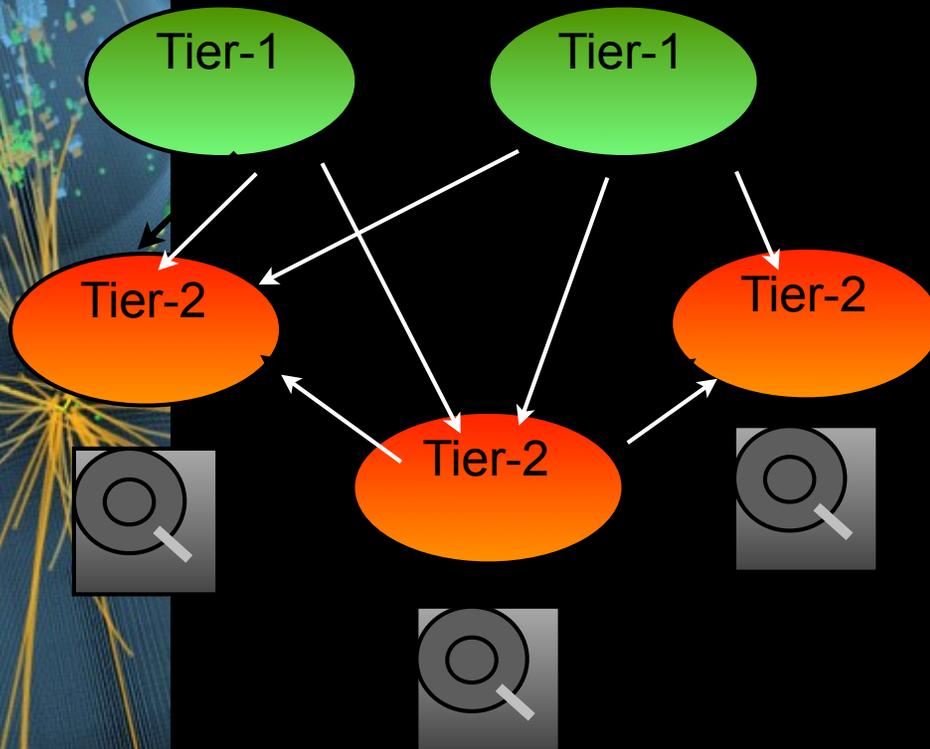


# Analysis Disk Storage

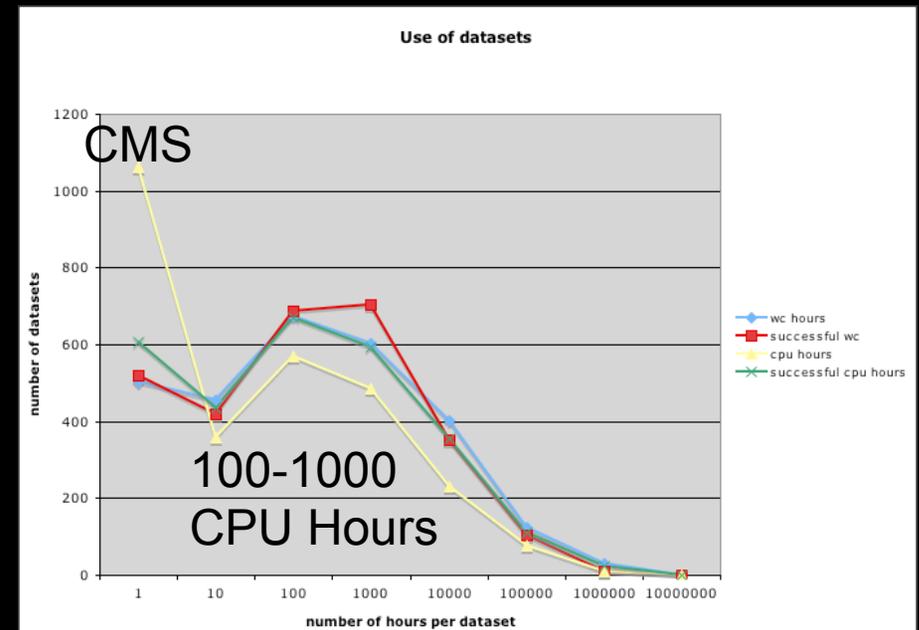
- The ability to purchase and operate large storage resources has been critical to the success of HEP computing
  - Tier-2s are very heavily utilized
  - Many of the challenging IO Applications are conducted at centers with exclusively disk
- Tier-2s vary from 10s of TB at the smallest site to 1PB of disk at the larger sites
- In 2012 there are more than 80PB of T2 Disk in LHC



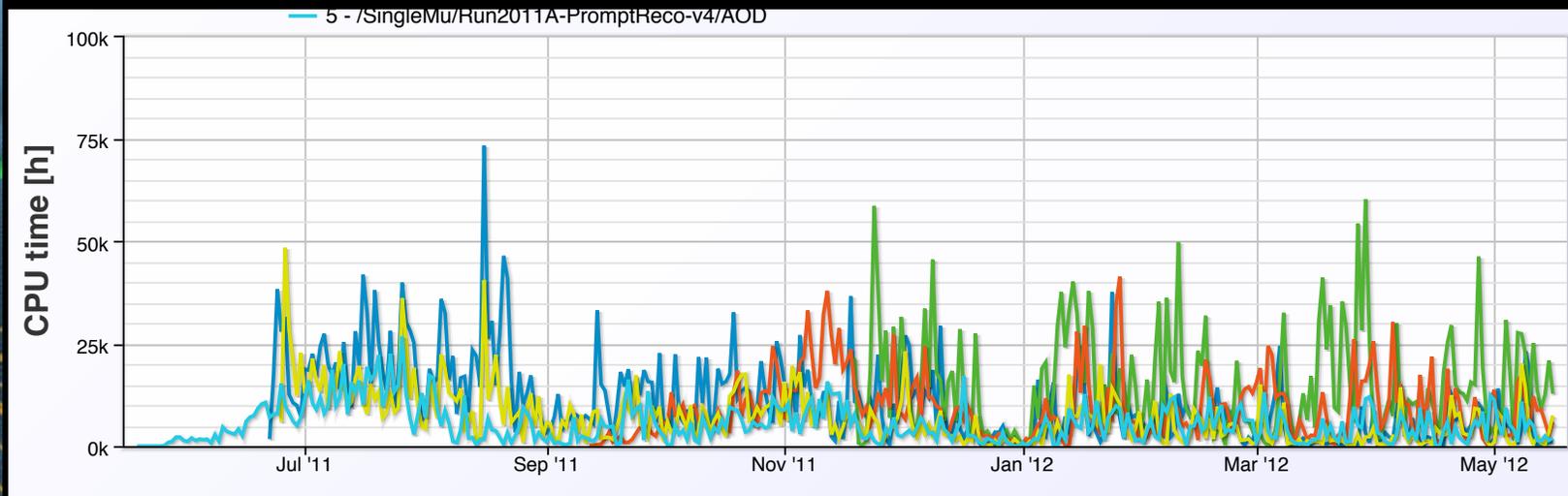
# Access



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



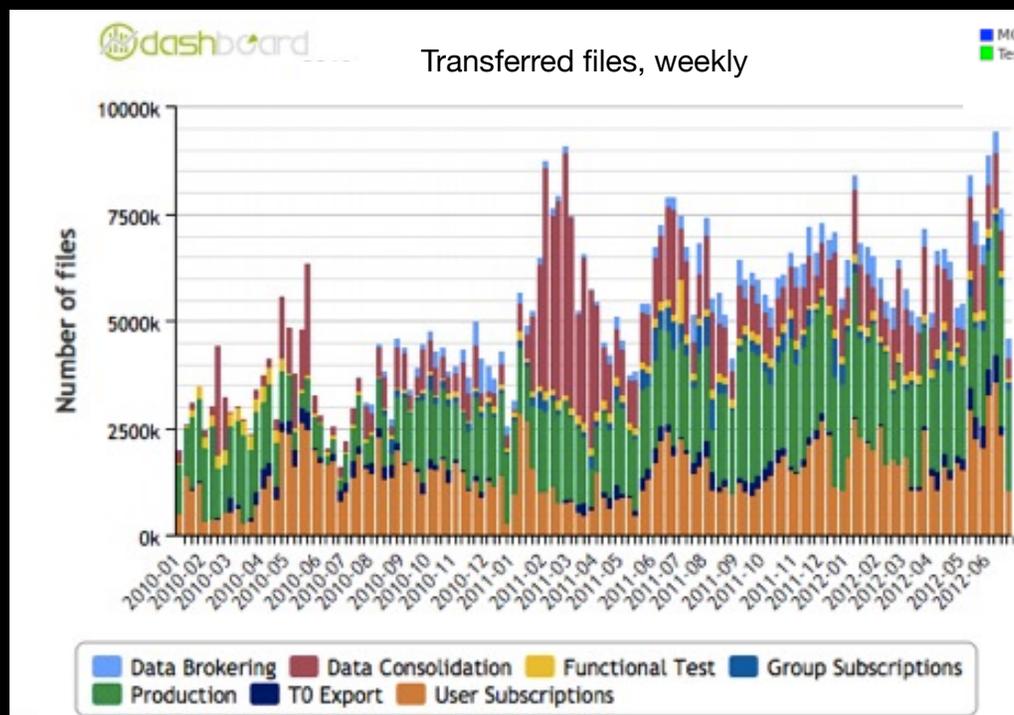
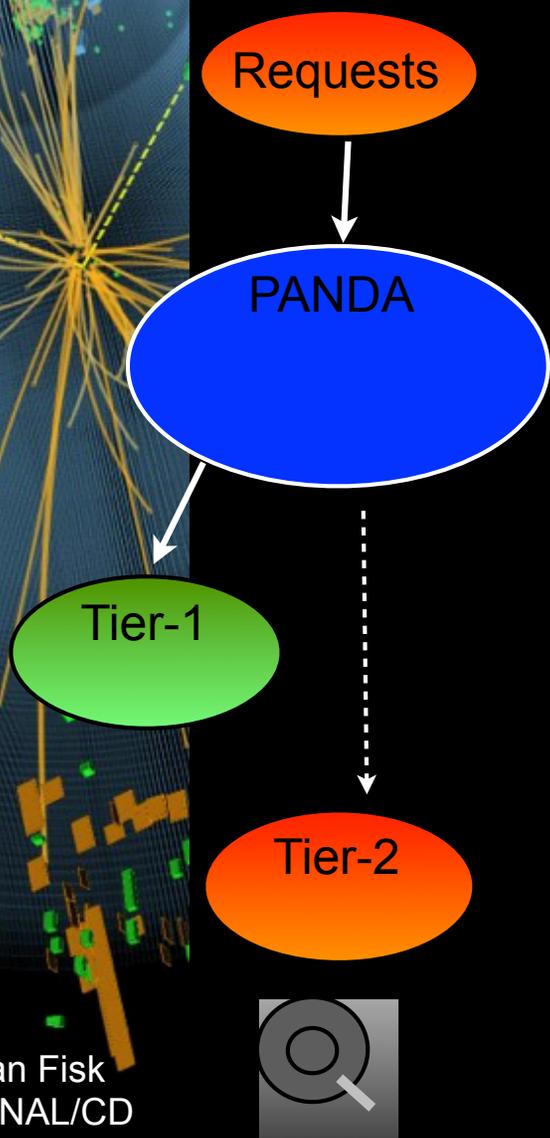
# Popularity



- 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 made
  - Jobs re-brokered to use them
- Unused copies are cleaned



# 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



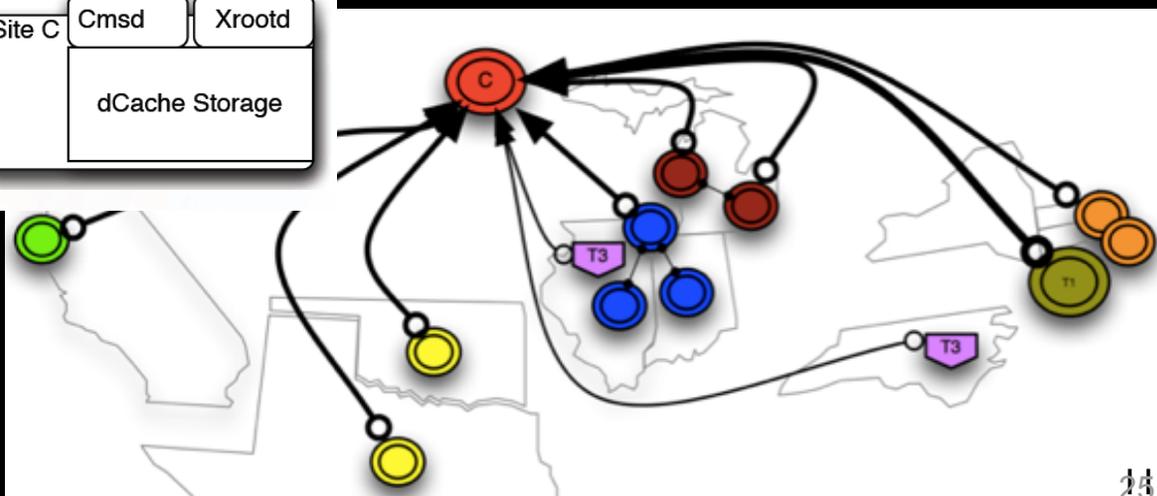
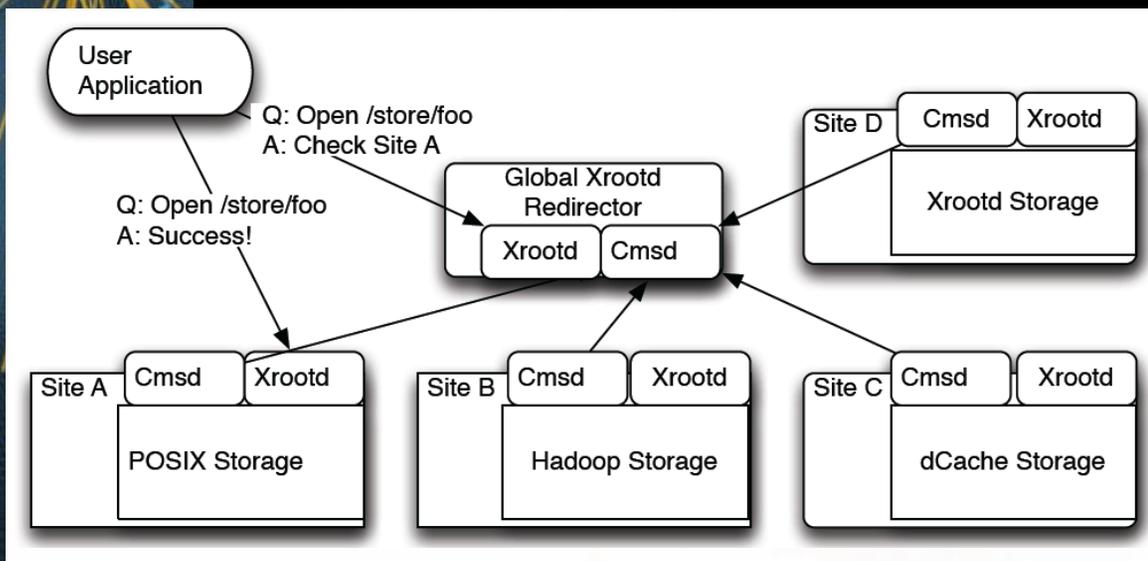
- Big gains in how storage is used by optimizing how events are read and streamed to an application

–Big improvements from the Root team and application teams in this area



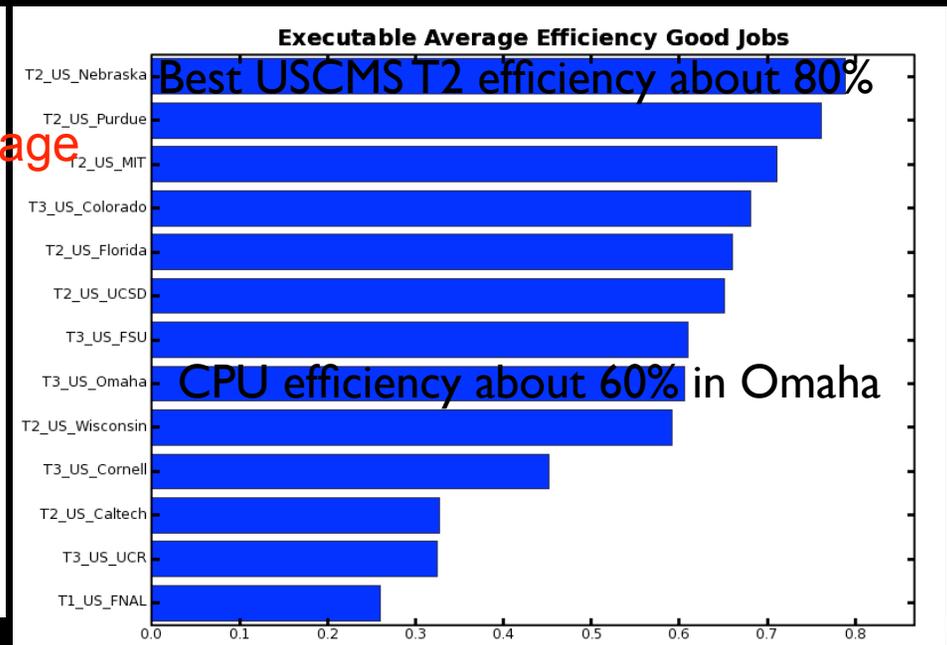
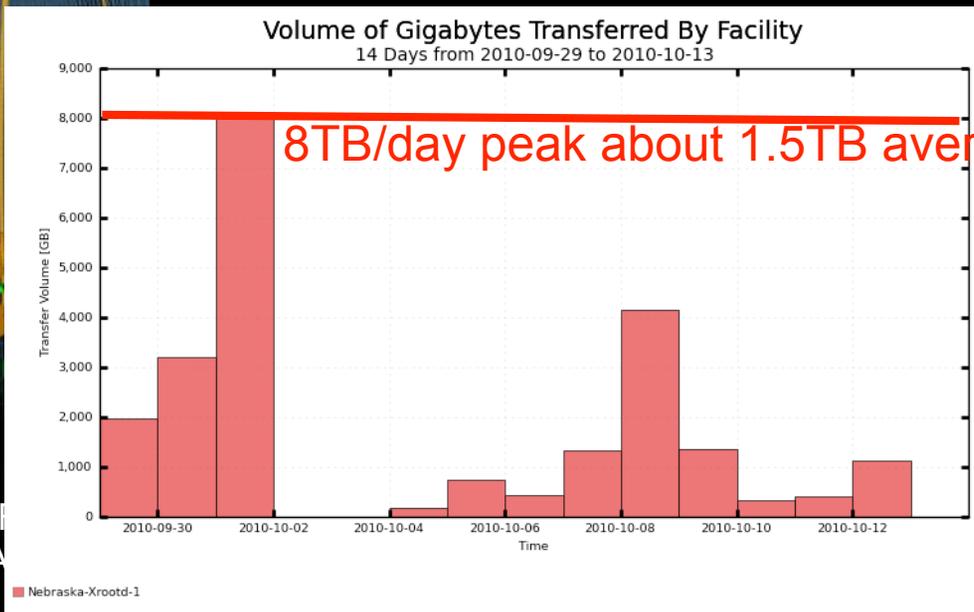
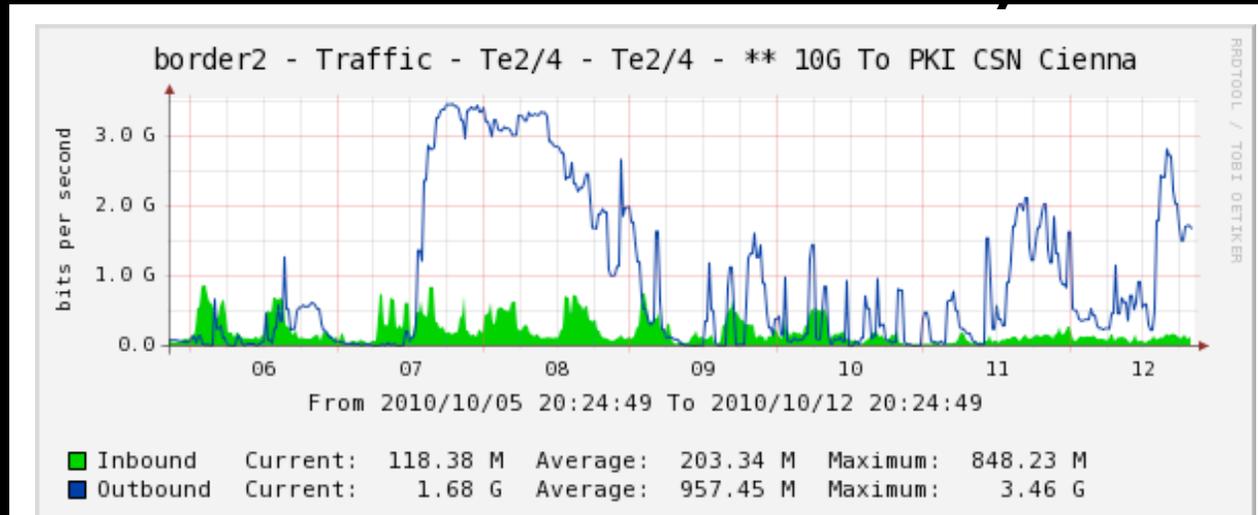
# Wide Area Access

- With 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



# Performance

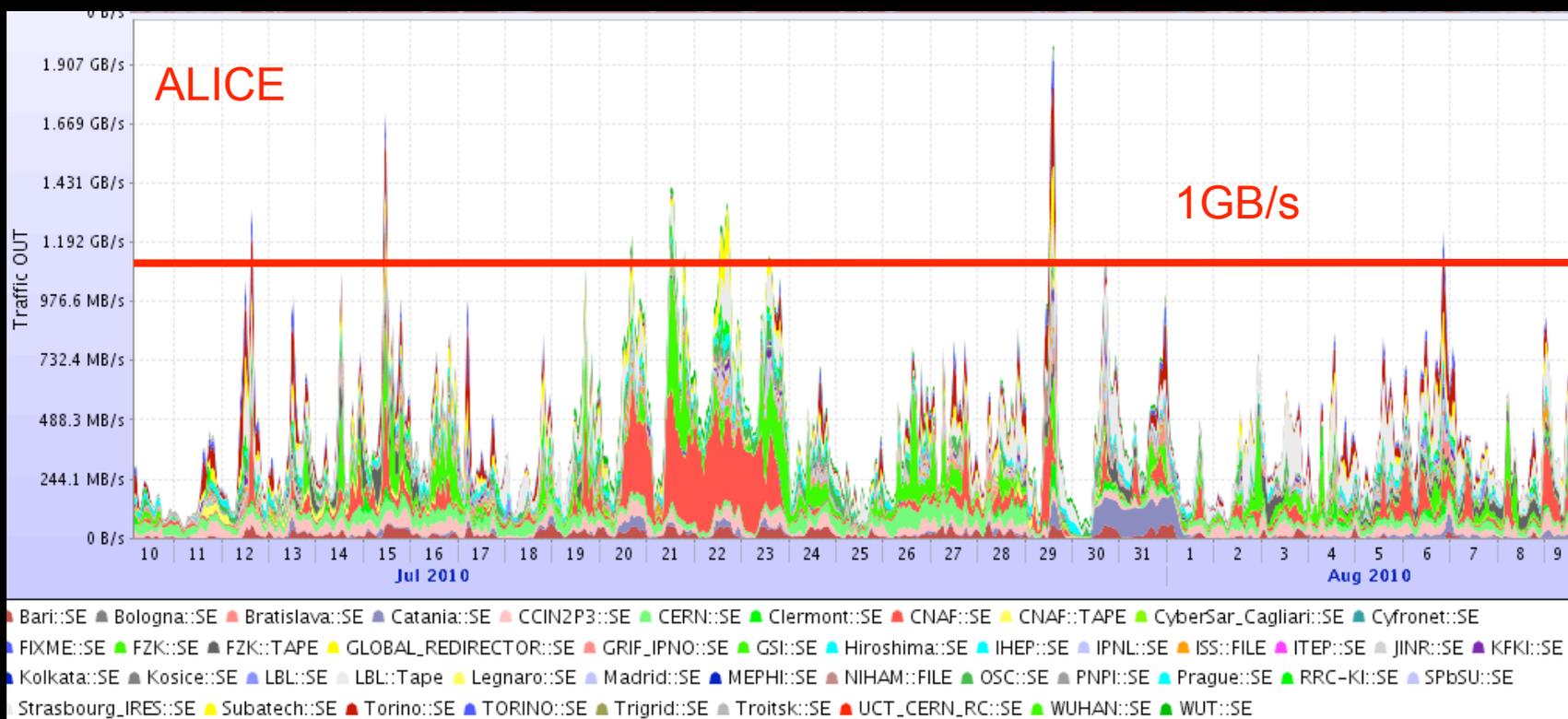
- This Tier-3 has a 10Gb/s network
- CPU Efficiency competitive



# Networking

- **ALICE Distributes Data in this way**

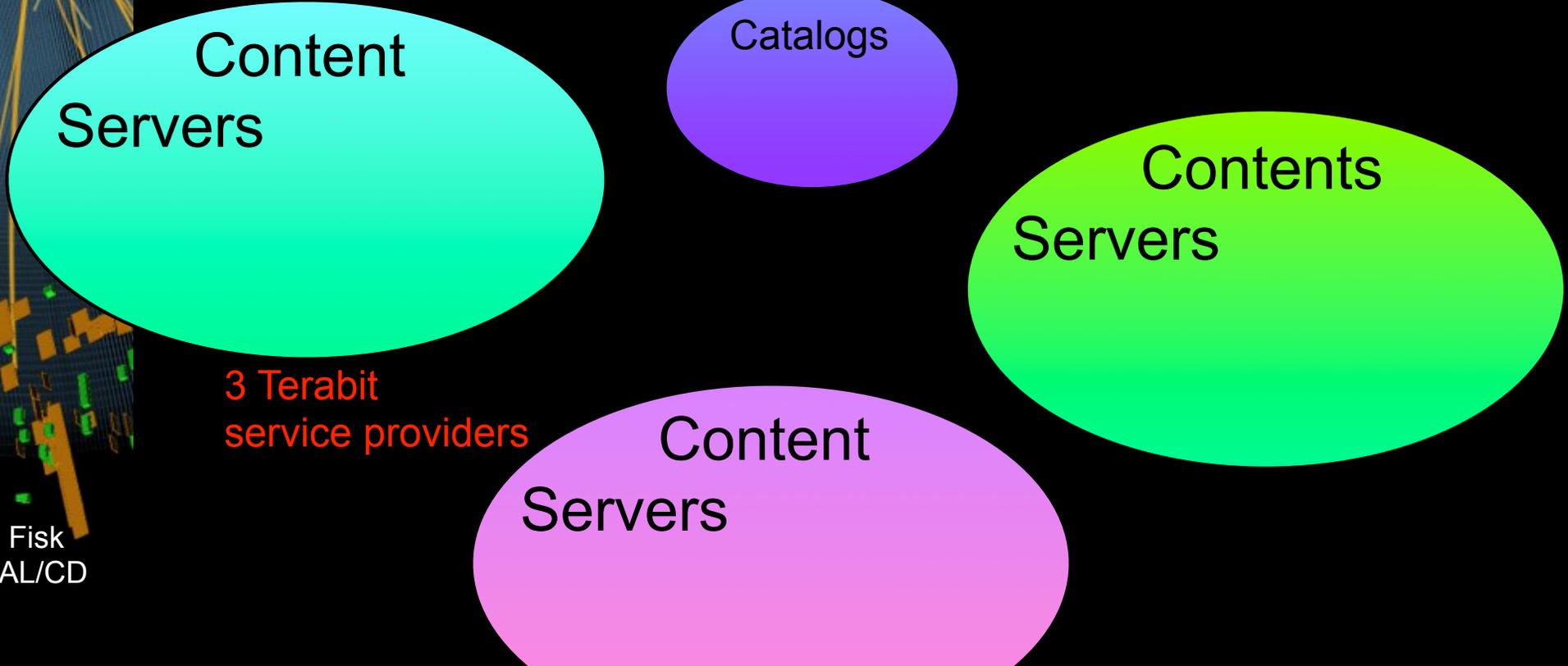
- Rate from the ALICE Xrootd servers is comparable in peaks to other LHC experiments
- Has the potential for providing access to “Any Data, Any Where, Any Time”



# Content Delivery Networks



- **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
    - More than 30% of the traffic



Content Servers

Catalogs

Contents Servers

3 Terabit  
service providers

Content Servers

# By the numbers

- We have a smaller number of clients, less distribution, and higher bandwidth per client
- They have much less data

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

Similar Problems  
Not all files  
are equally accessed

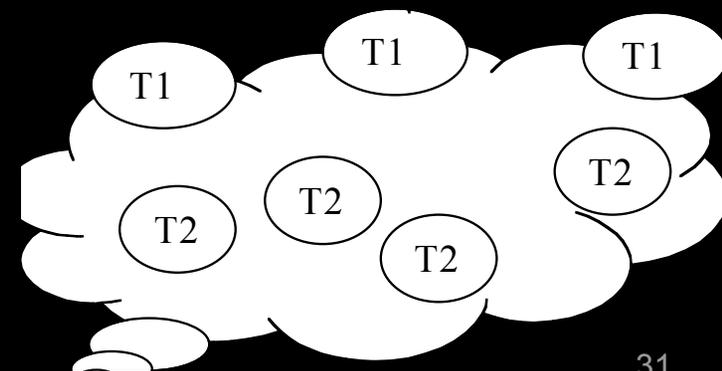
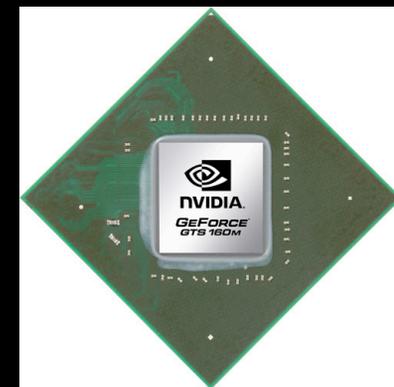


# Challenge of HEP

- 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

# Looking Forward

- Computing is at something of a cross roads
  - In one direction are clouds
    - Generic computing services that are bought, shared, or contributed
    - Computing as a service
  - In the other direction are very specialized systems
    - High performance, low power
      - Massively multi-core
      - GPUs
- Most likely we will use both depending on the needs



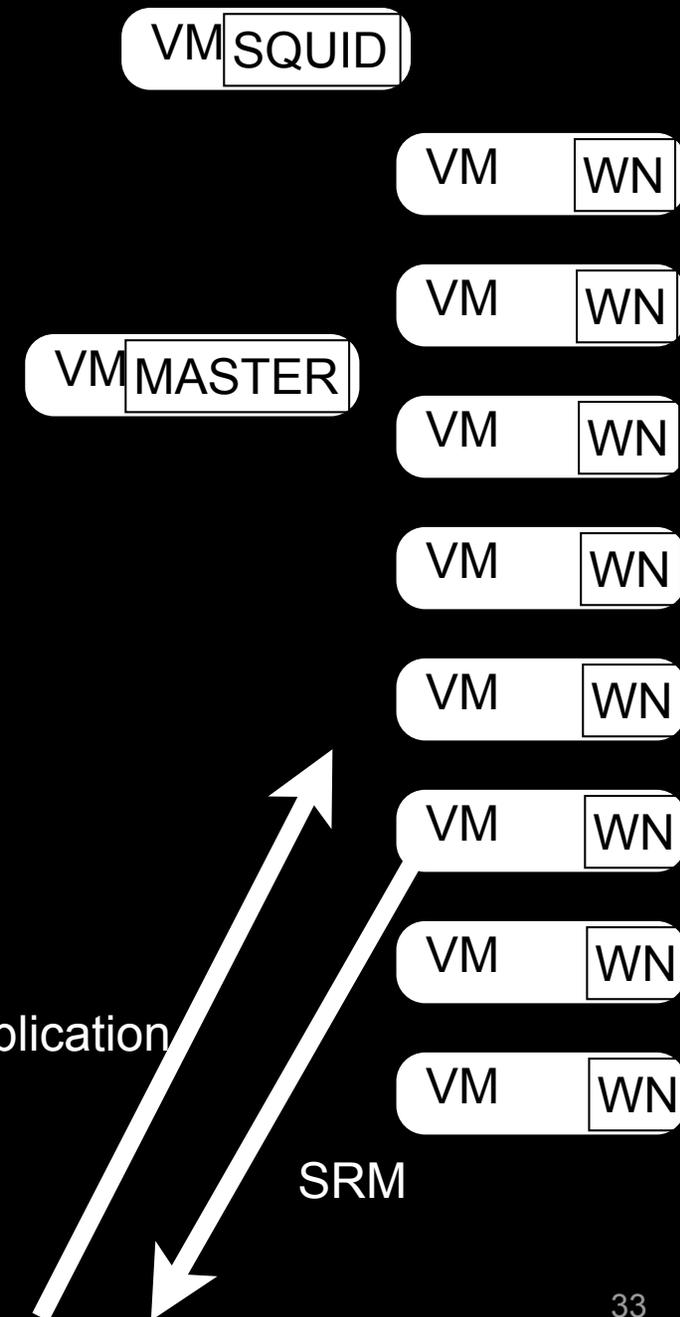
# Clouds vs Grids

- **Grids offer primarily standard services with agreed protocols**
  - Designed to be as generic as possible, but execute a particular task
- **Clouds offer the ability to build custom services and functions**
  - More flexible, but also more work



# Progress

- Large and small companies now offering Cloud services
  - ATLAS and CMS have both demonstrated standard workflows
- MC Simulation and Analysis workflows have both be demonstrated
  - Input data in over wide area transfer
  - Output over SRM



# Next Steps

- Goal is to be able to expand into addition computing resources to increase capacity dynamically
  - Possible to float into a cloud all the infrastructure associated with a site
    - Somewhat impractical and the services were not designed for dynamic deployment
  - Better to start processing services that connect and register with the workflow systems. Eliminating the need for the grid gateways
- Goal is to be able to double the processing capacity on demand

# Looking Forward

- Cloud provided computing tends to be factors more expensive than providing the resources in house for resources that are heavily used
  - The company needs to make money and you have to assume they have a huge efficiency gain over you associated with scale to make the service competitive
  - There are a variety of examples of non-commercial clouds which are interesting
  - Costs for commercial facilities is coming down
    - Interesting to cover peak periods
- Need to prepare for a time when this could be the norm
  - Unclear if Clouds follows a utility model or a rental car model

# Dedicated Hardware

- In the opposite direction of clouds is specialized hardware optimized for specialized tasks
  - Graphical processing units handle some kinds of floating point calculations extreme fast
    - Interesting to consider for things like trigger
  - Low power, high core count systems could be used for specialized massively parallel computing
    - Interesting to consider for data scouting applications
- Lose flexibility to go to specialized hardware, so the gains need to be large

# Data Preservation and Access

- High Energy Physics has not always had a strong record in data preservation
  - Data is always archived
  - 1 of the 4 LEP experiments has actually demonstrated the ability to go back and reanalyze a long time after collection
    - Interestingly the entire data for a LEP experiment fits happy on a desktop hard drive
- Tevatron is in the middle of this
- LHC has begun thinking of the preservation problem, but also releasing publicly a portion of the data
  - Open Access release

# Open Access

- CMS is planning to release a portion of the 2010 dataset
  - 2010 was a total of  $45\text{pb}^{-1}$ . We now take several times that per fill, but is an interesting sample with much simpler events
  - Data is divided into RAW, RECO, AOD, and 4 Vectors
    - Proposal is to release a portion of the AOD at this time. RECO is technically more challenging as it's larger. Proposal is not to release the RAW data, because it's not used for analysis even by CMS members.

# Providing Access to Data

- Even in the first limited release CMS is looking at distributing a few hundred TB of data
  - We don't really have a good indication of the expected level of interest
    - Students, interested citizen scientists, theorists?
  - Or the scale of the computing resources available to the interested parties
- Currently considering
  - xrootd through a data federation
    - Same technology we use distributed data in CMS, but with not authentication for these datasets
      - Assumes open access doesn't negatively impact other access
        - » Technology more common in HEP
  - Something like Torrent for peered distribution
    - A parallel infrastructure would be needed. Technology more common in a larger community

# Outlook

- Computing is the end of a long chain to realize the physics potential of the experiments
- The number of resources we have access to has grown along with the scale of the problem
  - More flexible and dynamic use of the resources available will make more efficient use of the system