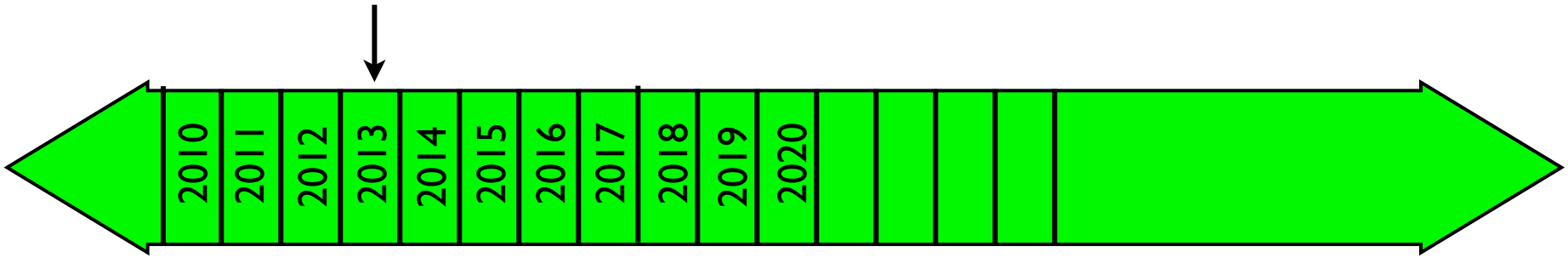
Strawman Model for Data Access and Management

16 June 2010

Ian Fisk

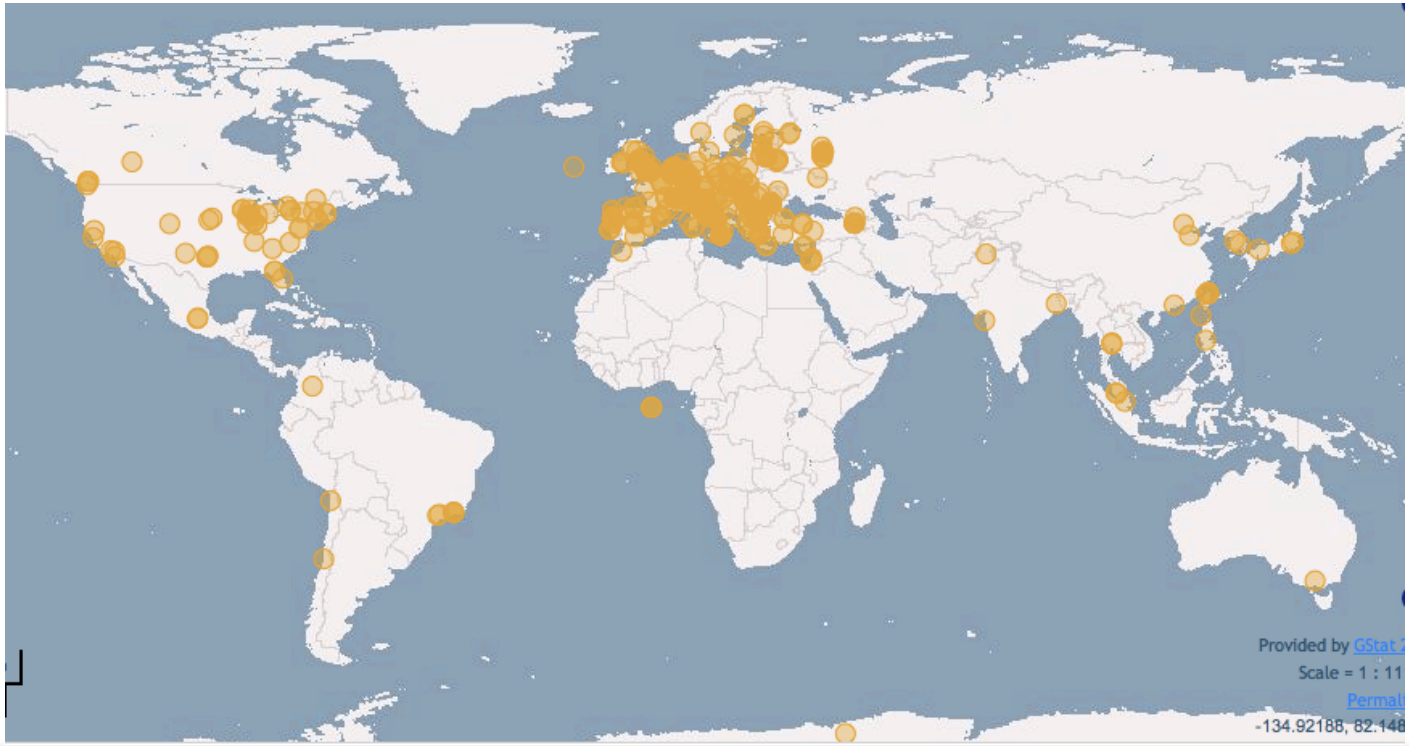
Introduction

- What follows is intended to spawn discussion
  - It is not the consensus view of any of the LHC experiments
  - Some items are opinion and some are common sense, some may be wrong or impossible to implement



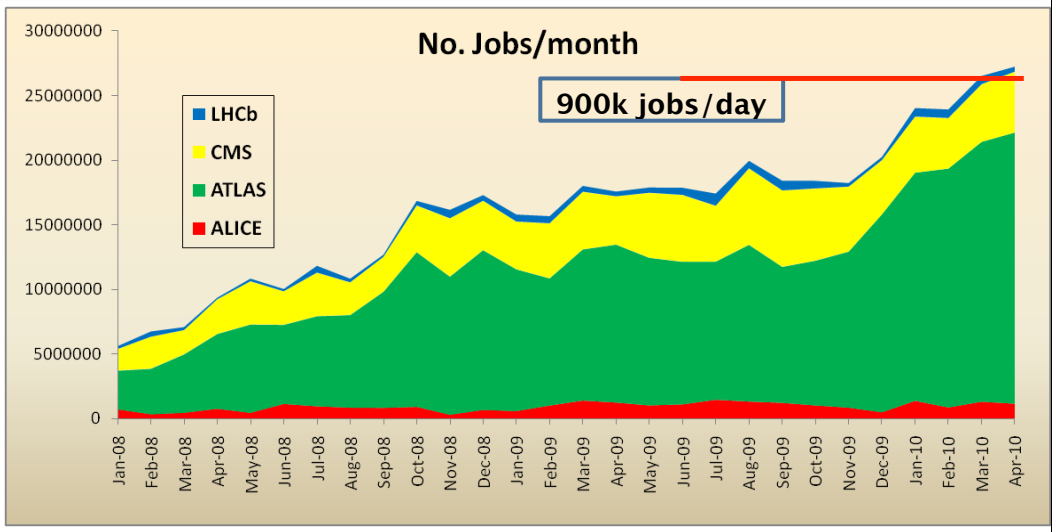
What's Good about what we have?

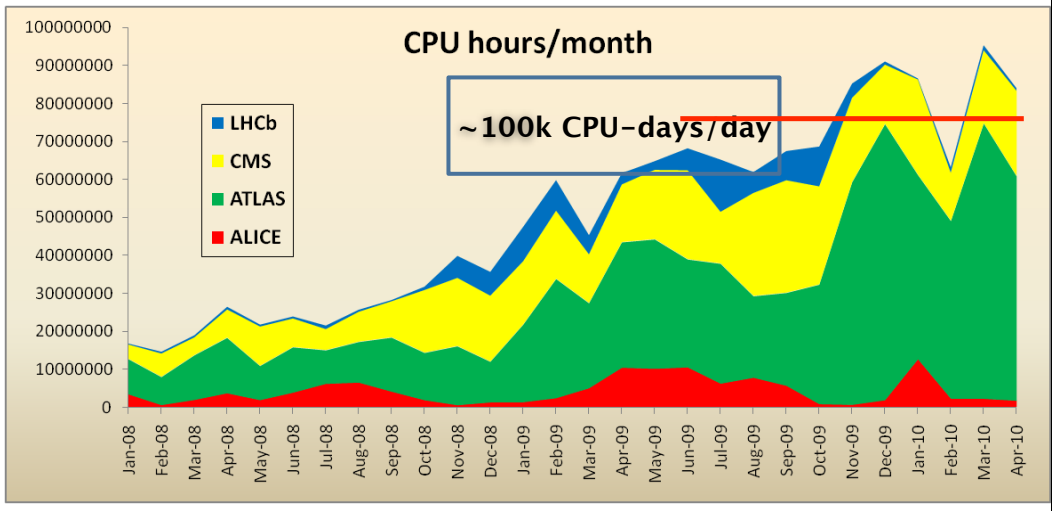
- After 9 years of implementation we have a system that can be widely deployed and scales reasonably well
  - We don't have extensive experience with analysis users
- More than 170 computing facilities in 34 countries
  - More than 100k Processing Cores
  - More than 50PB of disk



Scale

- Running increasingly high workloads:
  - Jobs in excess of 900k / day; Anticipate millions / day soon
  - CPU equiv. ~100k cores
- Workloads are
  - Real data processing !
  - Simulations
  - Analysis – more and more (new) users: several hundreds now
- Data transfers at hundreds of TB per day



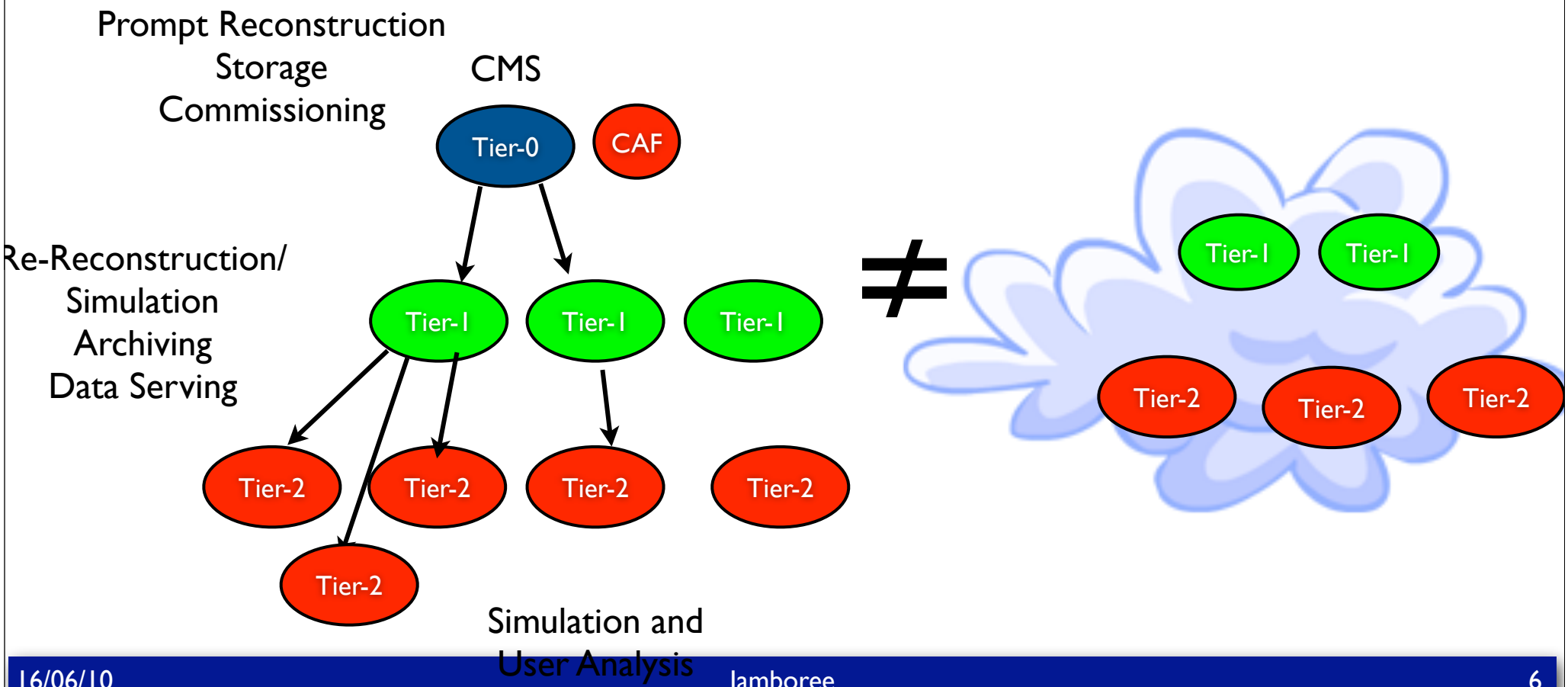


How to Improve?

- How to improve analysis for users?
  - Currently a lot of knowledge in the implementation details.
- Improve the transparency of access.
- Introduce less deterministic features to the system to improve flexibility and response

Legacy of MONARC

- A lot of the structure and hierarchy of the MONARC computing models remains for several of the LHC experiments
  - Transparency of Data Placement and Access has been replaced with a reasonably structured environment
- The MONARC Report just turned 10. There have been improvements in Computing Services and it may be time to revisit



Compromises

- A number of the choices the LHC experiments made in formulating computing models were based on limitations or assumptions of limitations
  - The networks will be the bottleneck
  - We will need hierarchical mass storage because we cannot afford sufficient local disk
  - The file catalogs will not scale
  - We will overload the source sites if we ask for transfers to a large number of sites
  - We need to send the jobs to the data to achieve efficient CPU utilization
  - We need structure and predictable utilization

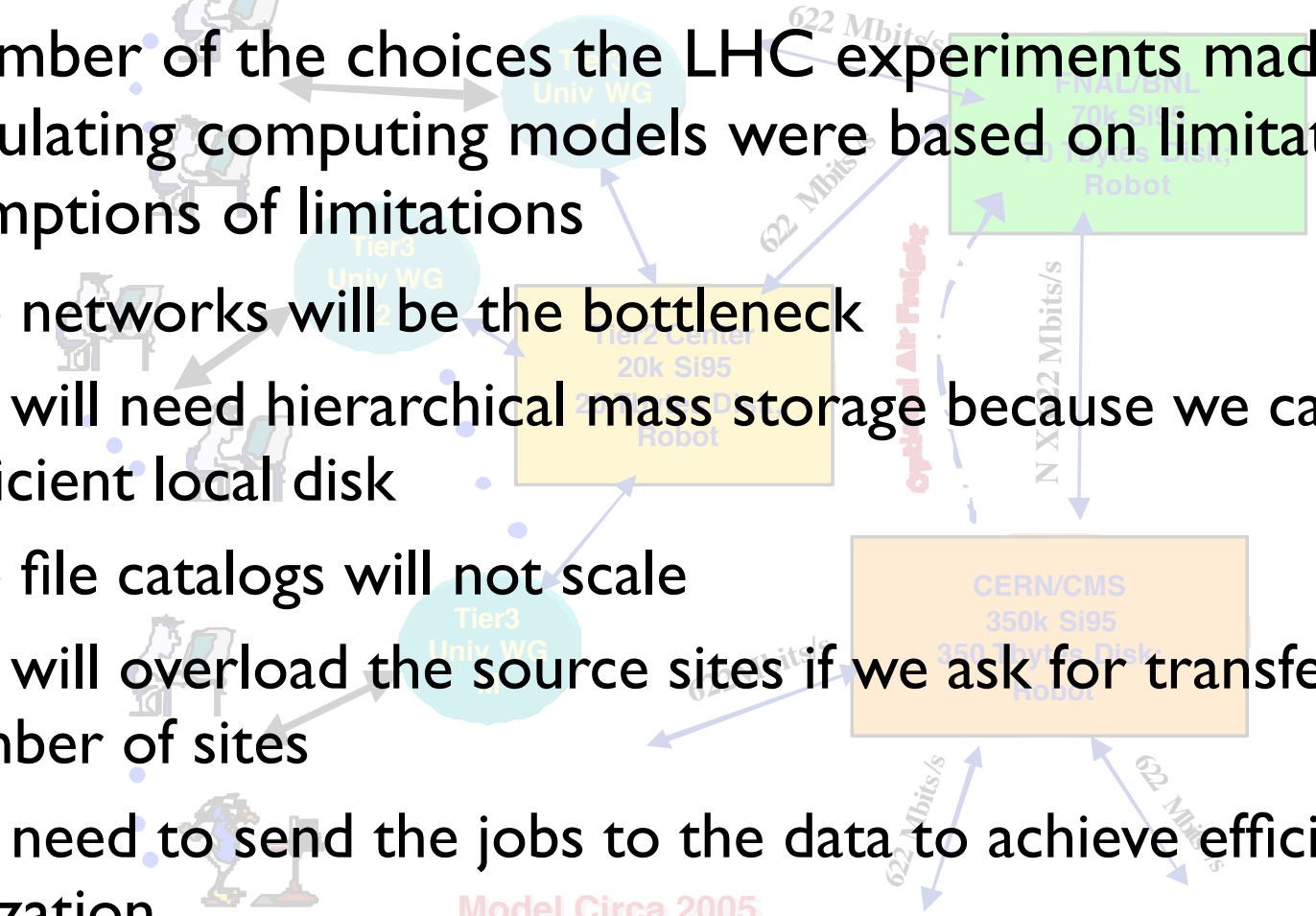
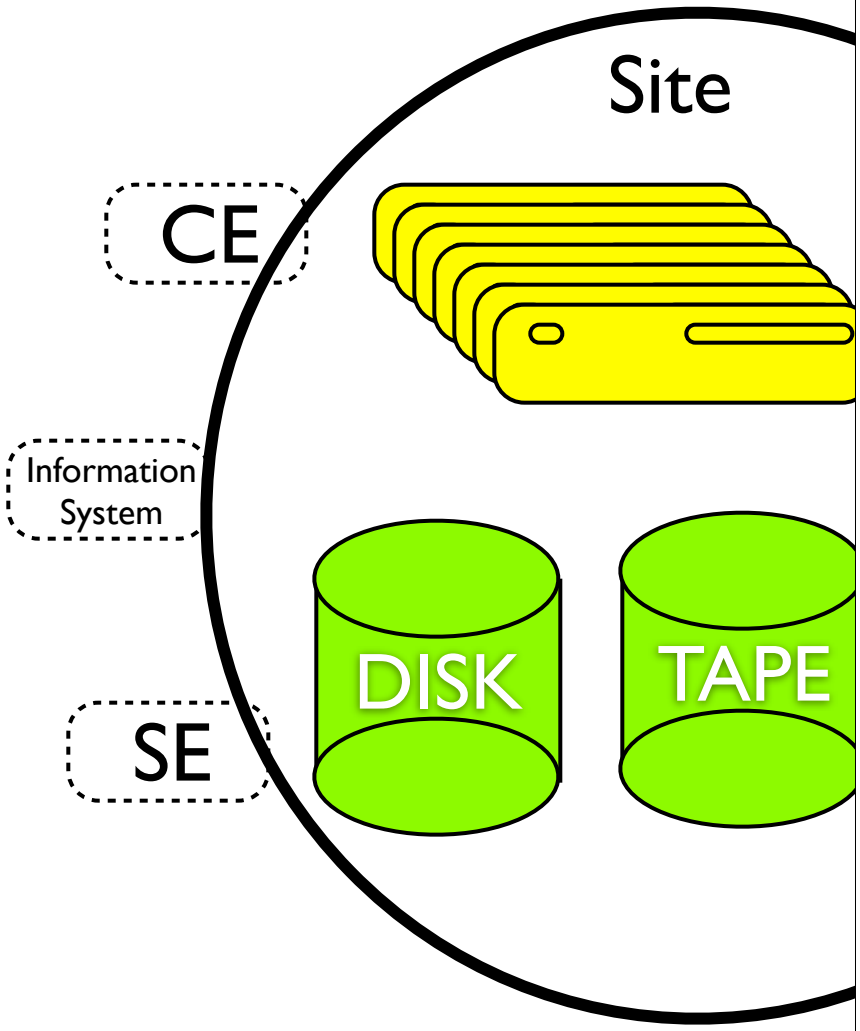


Fig. 4-1 Computing for an LHC Experiment Based on a Hierarchy of Computing Centers. Capacities for CPU and disk are representative and are provided to give an approximate scale).

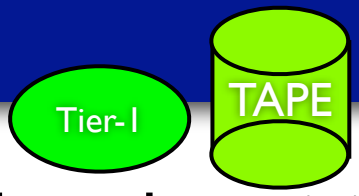
Site View

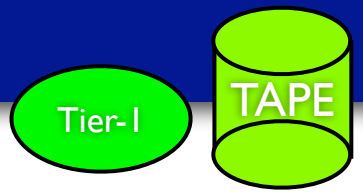
- Current site view for several of the LHC experiments begins to look like a walled city with a couple of gates
  - CE accepts jobs. We run thousands of similar requests and we authorize all of them
  - SE transfers data in and out
    - Data is preloaded into the sites and jobs come and finds it



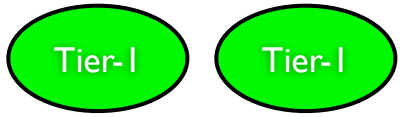
Lower Level Services
Providing Consistent
Interfaces to Facilities

Networking





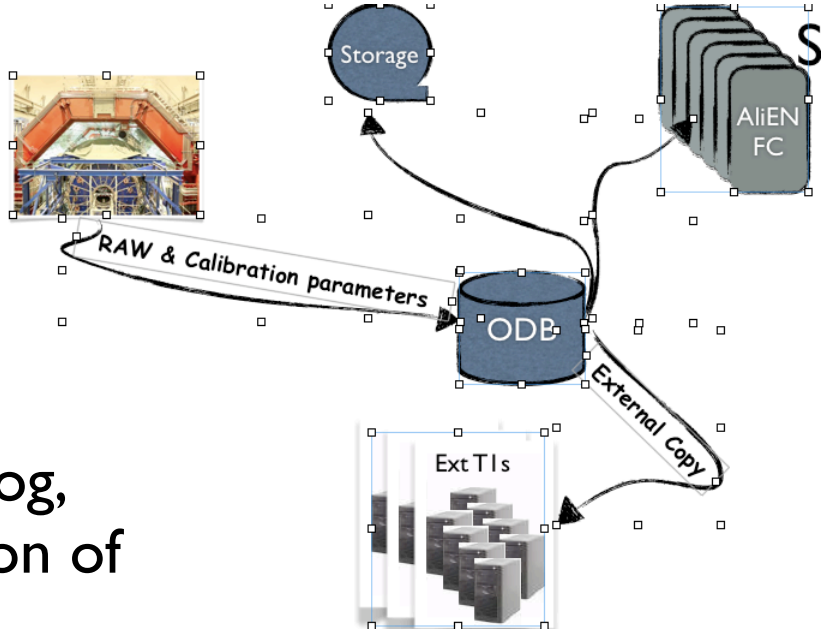
- When the original models were written the networking was perceived as unreliable, slow, or insufficient
  - Sites needed to be treated as independent because they were far apart and on reasonable time scales activities happening at the far site would not impact local activities
- The current reality is sites are much closer.
  - At many places the WAN connection and the local backbone is comparable



- For several experiments the current model assumes networking, one of the most stable resources, is the least reliable
  - How do we take better advantage of networking?

ALICE

- ALICE is different from the other LHC Experiments in that they break dependence on local files
  - It's interesting model to look at
    - Relies heavily on xrootd to implement access to files over the WAN
    - Authorization is done at a central catalog, and does not rely on site implementation of the experiment policy
  - Current implementation requires a VO box inside the firewall at each site



Concrete Steps: Tape

- Can we move to a system where tape is a true archival system?
  - Our network bandwidth off the site is typically higher than the total IO to tape on a site
    - LTO-4 tapes take a minute on average just to position
  - The latency to pull a file from a disk copy somewhere should be much better than pulling from tape
    - Requires a rethinking of how we implement the interface to storage and how we track what's on disk and tape
  - Can we get lower latency, improved CPU efficiency and lower operations load?
  - Currently at least 2 of the experiments have several times the disk space of the total expected dataset but yet we stage from tape

Concrete Steps: Access

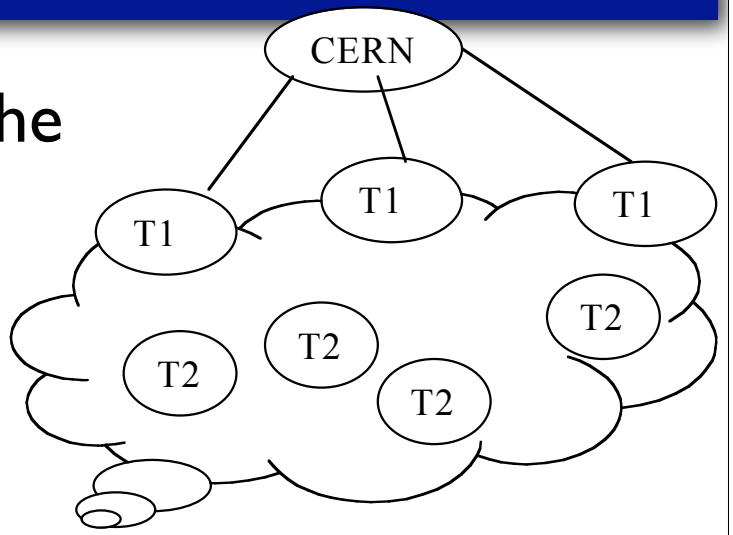
- Can we provide alternatives for the other LHC experiments to simply taking the local file?
  - Can we provide an access layer that requires less optimization in the application for all the kinds of storage?
  - Can we provide an access layer that keeps good CPU efficiency but provides the possibility to access files from another site?
    - What intelligence would be needed in this layer?

Concrete Steps: Data Management

- Why can't we have a data management system as good as that used for video piracy?
  - File size is similar. Data volume is smaller, but number of users and sites is bigger
- We should be investigating peer-to-peer technologies to handle replication of data between analysis sites

Migration

- Tier-0 to Tier-1 is a natural extension of the experiment DAQ systems
  - We probably want predictable and controlled flows of data into the remote tape archive.



- Tier-1 to Tier-2 for processing and analysis would we benefit from less structure ?
  - Some data pre-placed or transferred by the last application and some accessed at a distance
  - Talk this week about what problems are made harder and what are made easier.

Timescale

- The restart in 2013 is a reasonable target for an evolution in the system
  - Enough time to demonstrate new functionality and scale. Integrate new components and test
  - Not enough time to start from a green field
  - We shouldn't allow the work to disrupt the current operations
- Need to evolve from the functionality we have
  - Reasonable Goals:
    - More transparent access to files and better utilization of the network
    - Achieving a less deterministic system where there is the possibility for alternative data access

Why?

- We need more efficient data management and data access infrastructure for analysis for all the LHC experiments
- We need a system that the total amount of disk will not scale linearly with the total volume of data
  - The out year disk needs would likely be unachievable

Some Speculation

- The world around us will be moving more to storage whose location is unknown and considered unimportant
  - We need to figure out how to efficiently use the networks and optimize the data access to provide less reliance on local storage.
  - People have been predicting the death of tape for years. Tape is valuable to protect against failure, but less regular access would provide a more efficient system
  - Do we expect to care about data location as much in 3 years?