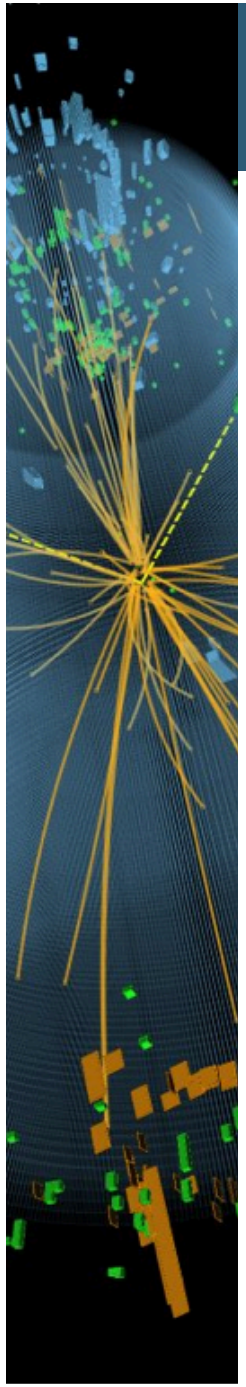
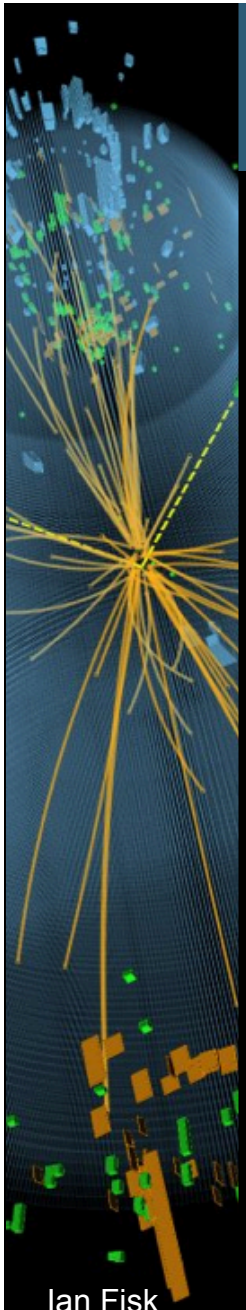


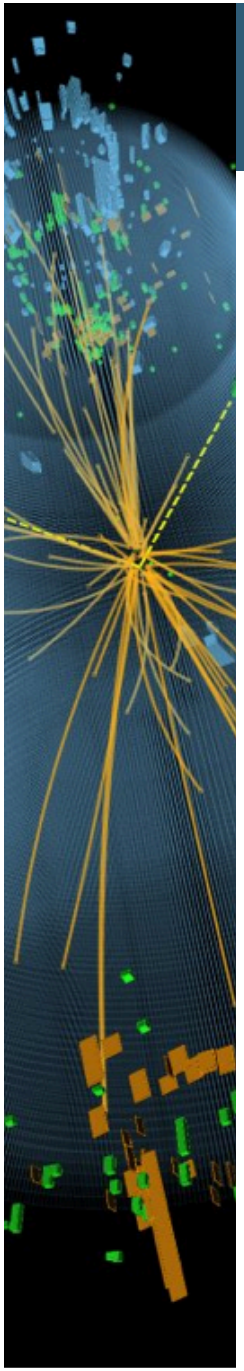
Future Directions



Future Directions

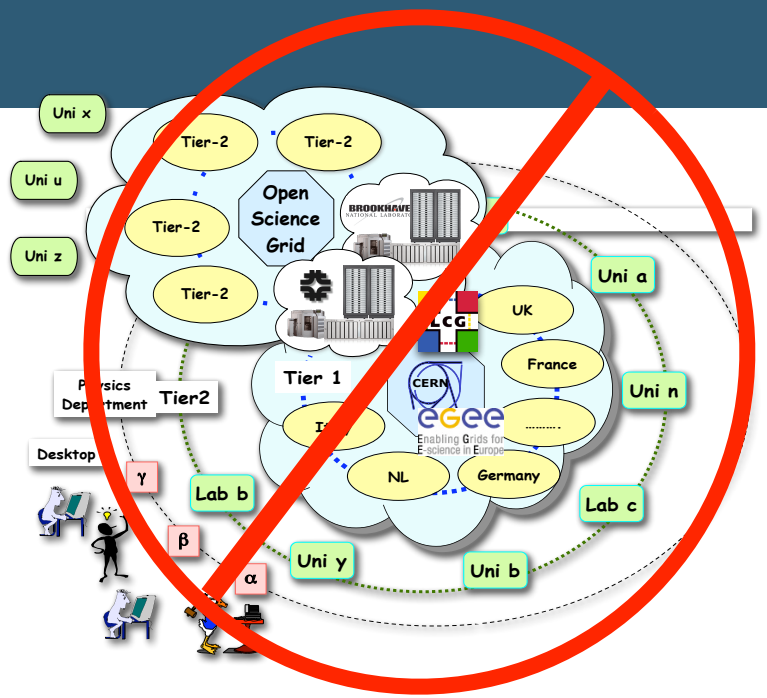
- The Future is hard to predict and is driven by technology trends, unforeseen events, special needs, previous commitments, policy, mistakes, innovation, etc.
  - What follows is some personal observations. Only when we get to the future do we get to see how accurate they were.

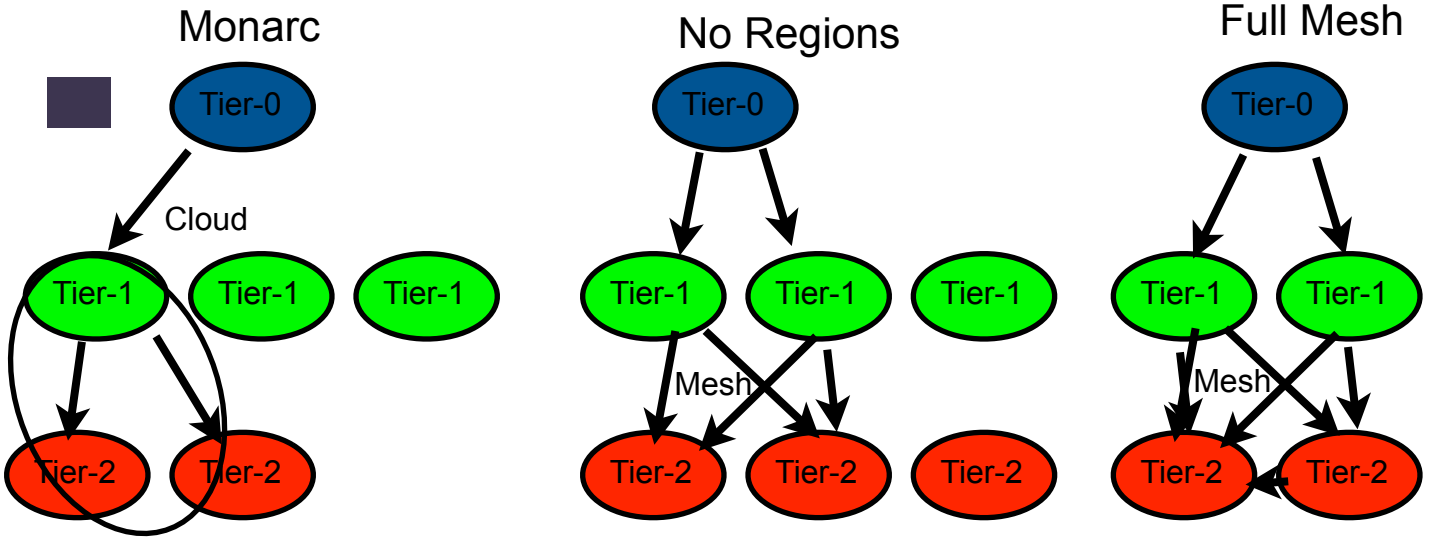
Ian Fisk
FNAL/CD

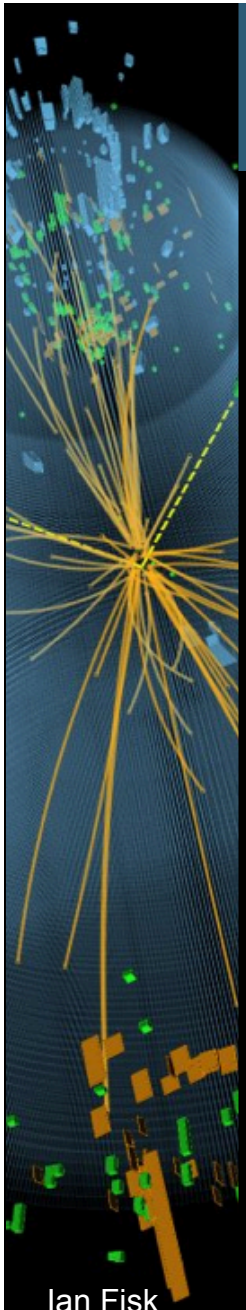


Data Driven

- ATLAS and CMS have driven their activities based on the location and flow of the data
- Very deterministic



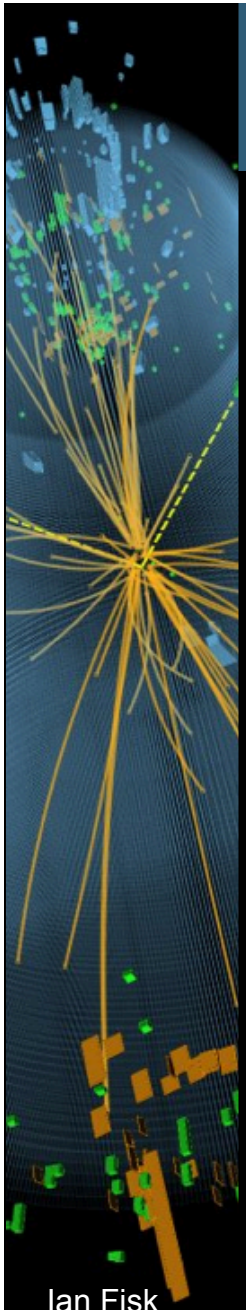




Continues Today

- Changes in the how we treat and store data are also driving the future directions
  - How we think of functionality is changing, and the concept of locality
- Generally flattening of the Tiers

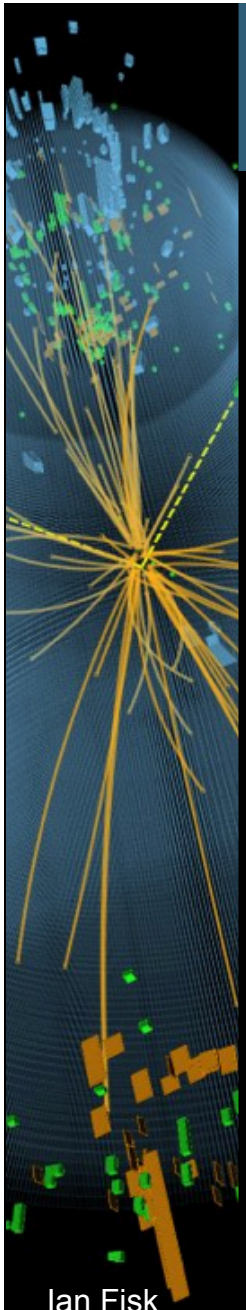
Ian Fisk
FNAL/CD



Data Management Changes

- The traditional concept of hierarchical mass storage where data can be recovered automatically from tape comes with limits
  - In the case of CMS, we restricted users from T1s to prevent accidental restores from tape
  - Data written automatically to tape is safe, but slow to recover the media if it was never intended to be there
  - A lot of work went into concepts of “storage classes” to predict when data would be on disk and tape
- CMS is in the process asking Tier-1 sites to physically separate disk and tape

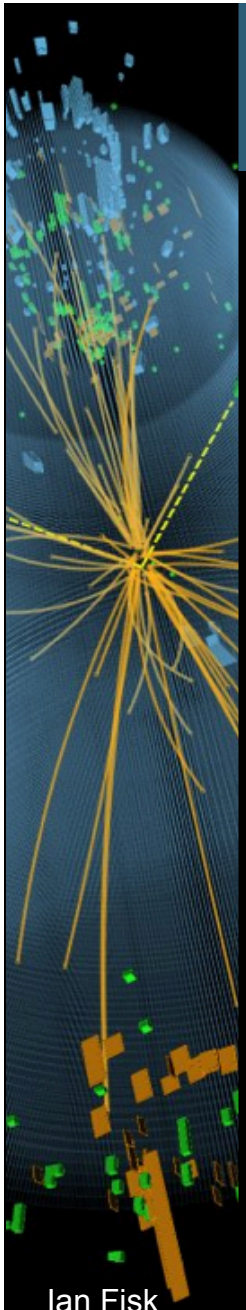
Ian Fisk
FNAL/CD



Disk/Tape Separation

- CMS will ask that all Tier-1 sites introduce two PhEDEx end points one disk and one tape
  - The disk end point should write only to disk
  - The tape end point should write to tape, but could also be hosted on disk
- Like any other endpoint data subscribed should be resident until deleted
  - Subscribing to disk will be the equivalent of the prestage and pin
  - Subscribing to tape will trigger and transfer to archive
  - Deleting from disk will release the cached space

Ian Fisk
FNAL/CD



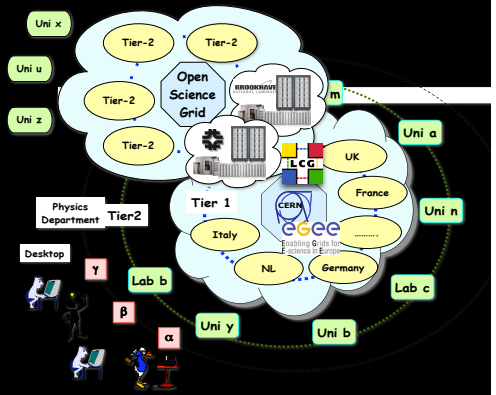
Impact

- Once you have split the archives there is no reason for a strict one to one mapping of disk and archive at Tier-1s
  - Archives could be used to stage datasets to any disk facility
    - The quantum of data we let the archive manage is dataset. (TBs rather than files GBs)
- Need to ask the question how many archival facilities do you need?
  - More than 1, but probably less than 10

Ian Fisk
FNAL/CD

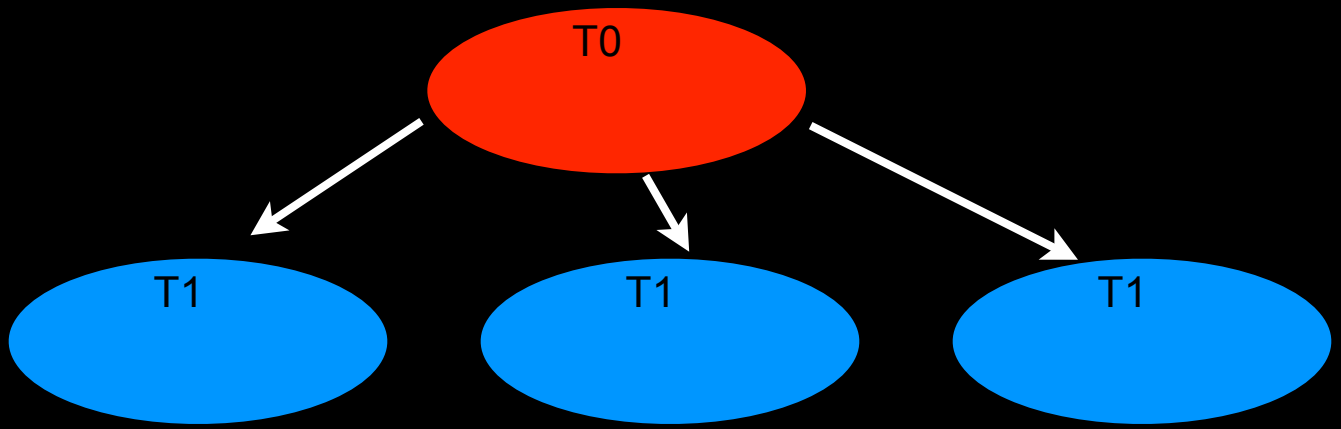
Changes how we think of tiers

- Once you introduce the concept of an archival services that is decoupled from the Tier-1
  - The functional difference between Tier-1 and TIer-2 is based more on availability and support than size of services
    - Difference between Tier-1 and Tier-2 from a functional perspective is small
  - Model begins to look less Monarc-like



Stretches into Other elements

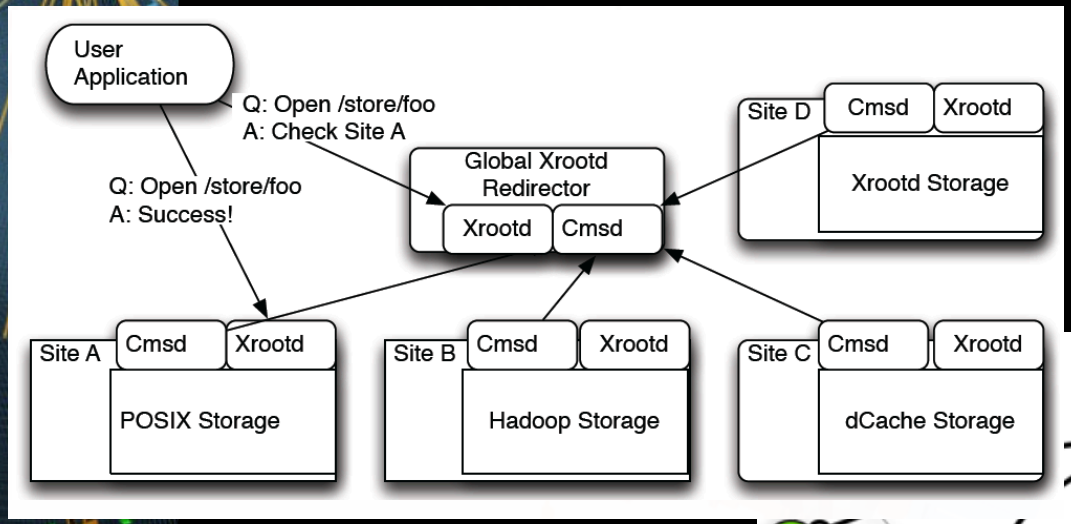
- After Long Shutdown 1, CMS will likely reconstruct about have the data the first time at Tier-1s in close to real time
  - Very little unique about the functionality of the Tier-0
    - Some prompt calibration work that uses Express data, but even that could probably be exported

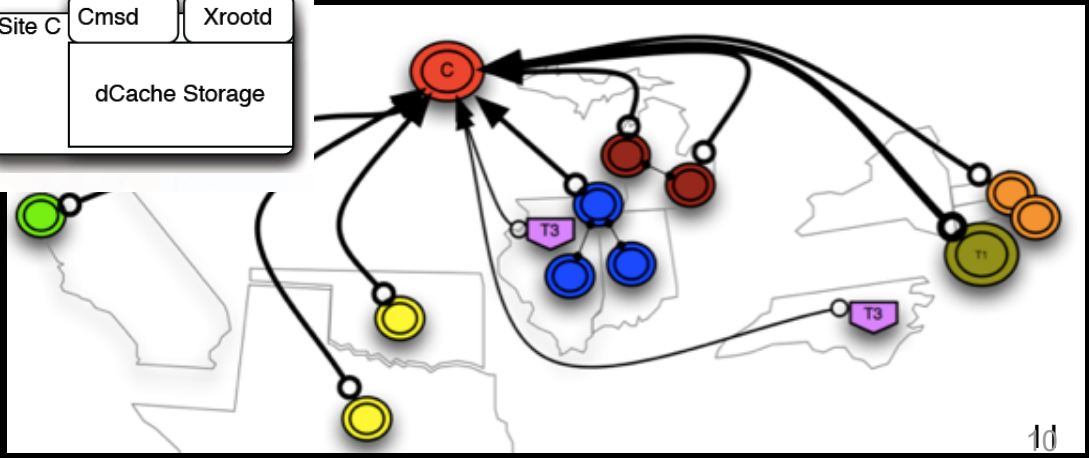


Ian Fisk
FNAL/CD

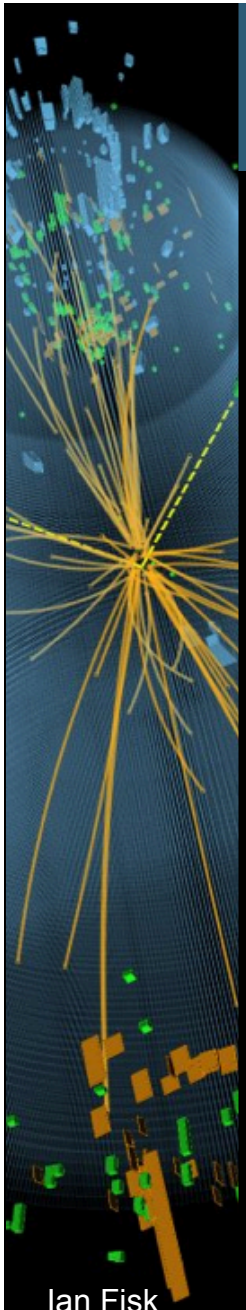
Wide Area Access

- Data Federation begins to break down the boundaries between sites
  - Sending data directly to applications over the WAN
- Not immediately obvious that this increases the wide area network transfers





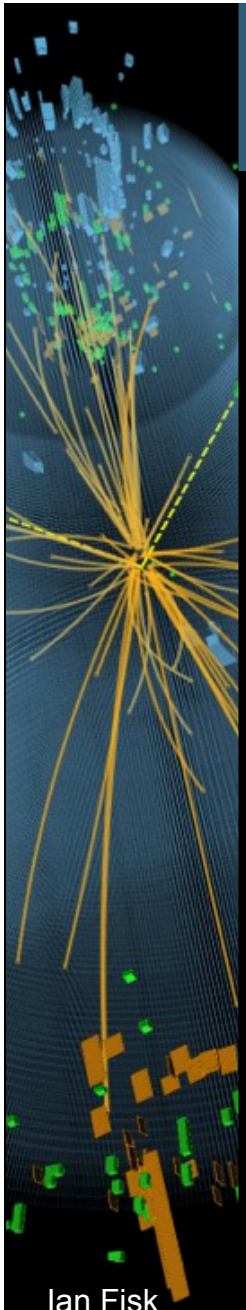
Ian Fisk
FNAL/CD



The Tier-1 Storage "Cloud"

- CMS is proposing to work on transparent access to data at Tier-1s using the OPN
  - Negotiations with sites for moving worker resources inside the OPN domain
  - Calculate and test the amount of access that could be sustained with our share
- Short term would eliminate individual workflow problems. In the long term could evolve CMS to much more dynamic use of the resources
- Instead of failing back to archive, we would fall over to xrootd if the data was accessible on another Tier-1
  - All items discussed at the Amsterdam workshop

Ian Fisk
FNAL/CD

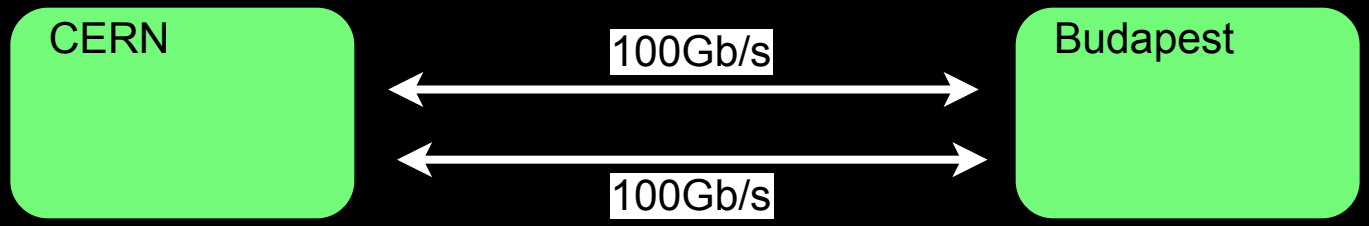


Flexibility

- Computing intensive tasks like reprocessing can be sustained reading data from remote storage
  - Input size is small compared to the size of the application
  - 50kB/s is enough to sustain the CMS application per slot
  - Even thousands of cores can be reasonably fed with Gb/s
- Works for analysis as well as long as the data format allows only the objects needed to be read

Ian Fisk
FNAL/CD

Networking

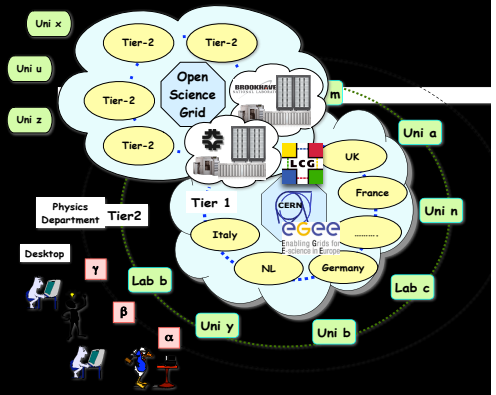


- CERN is deploying a remote computing facility in Budapest
  - 200Gb/s of networking between the centers at 35ms ping time
  - As experiments we cannot really tell the difference where resources are installed

Ian Fisk
FNAL/CD

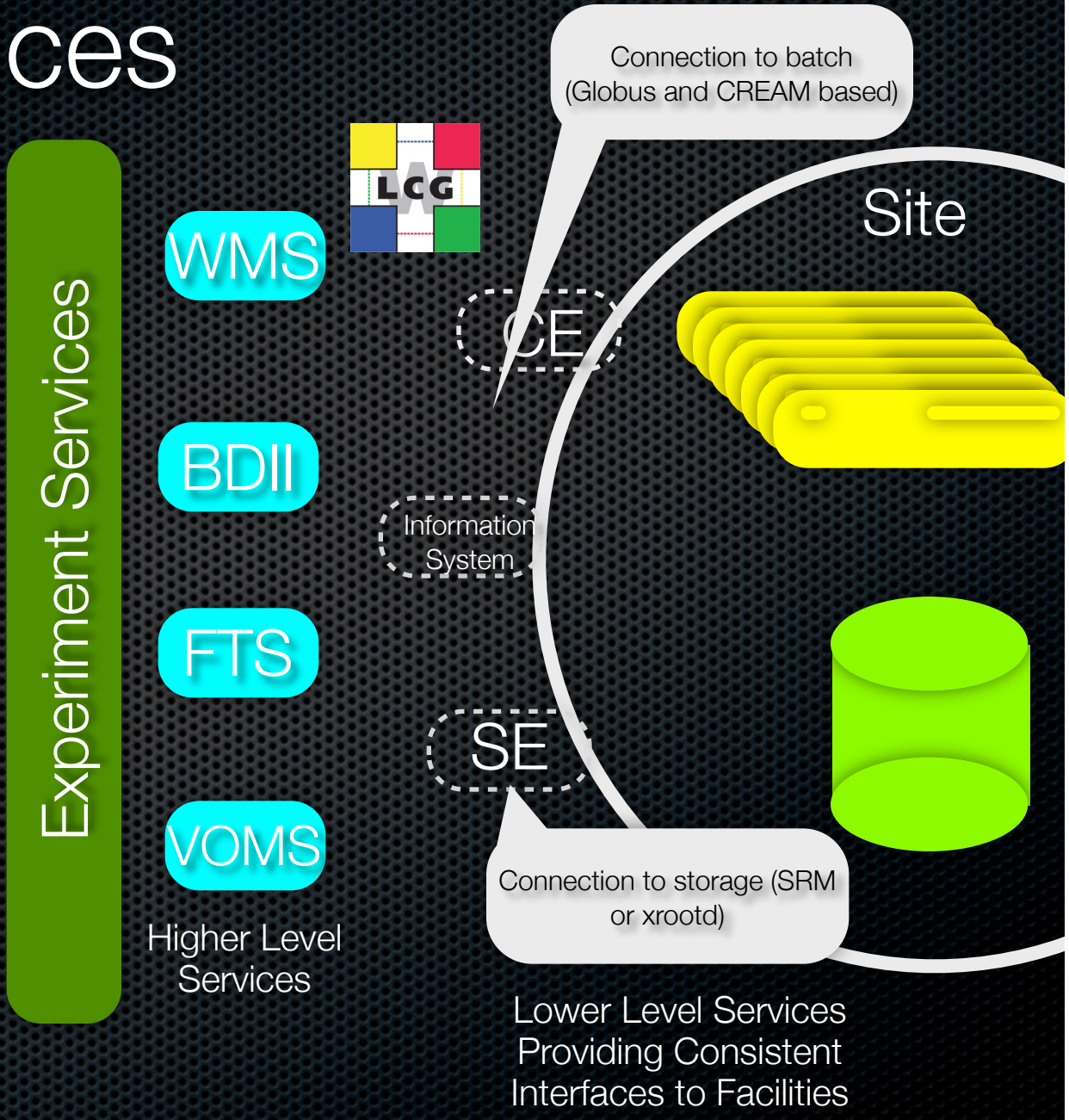
Networks

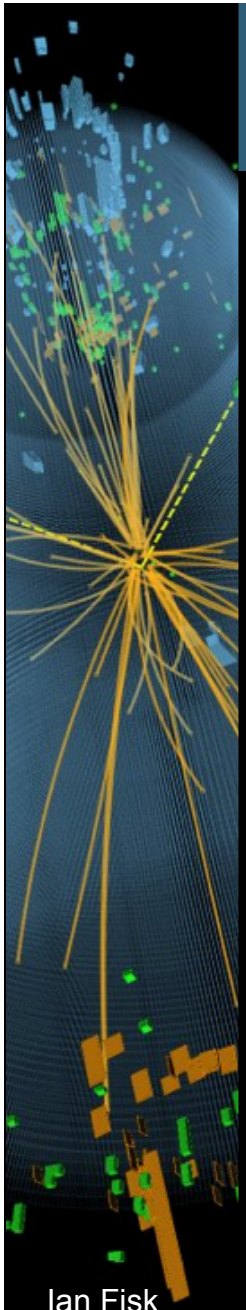
- These 100Gb/s links are the first in production for WLCG
  - Will be the first of many
- We have reduced the differences in site functionality. Then reduced the difference in data accessible. Then reduced the difference in even the perception that two sites are separate
- We can begin to think of the facility as a big center and not a cluster of centers



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

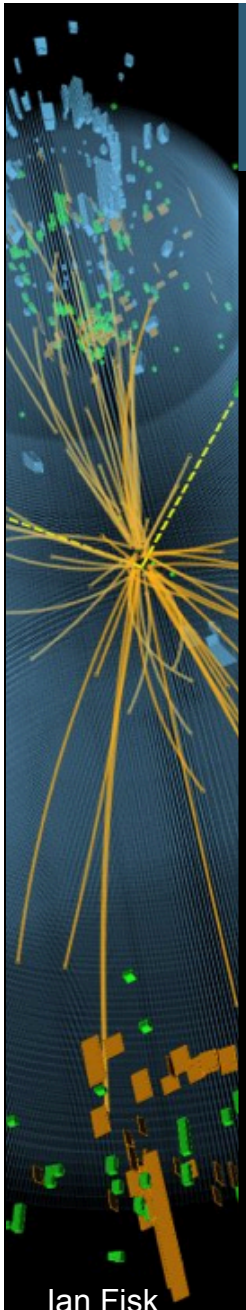




Changing the Services

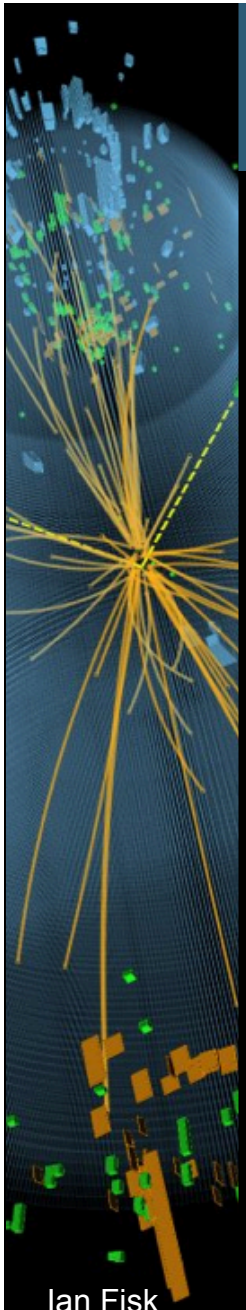
- The WLCG service architecture has been reasonably stable for over a decade
  - This is beginning to change with new Middleware for resource provisioning
- A variety of places are opening their resources to “Cloud” type of provisioning
  - From a site perspective this is often chosen for cluster management and flexibility reasons
  - Everything is virtualized and services are put on top
- There is nothing that prevents a site from bringing up exactly the same environment currently deployed for the WLCG, but maybe it’s not needed

Ian Fisk
FNAL/CD



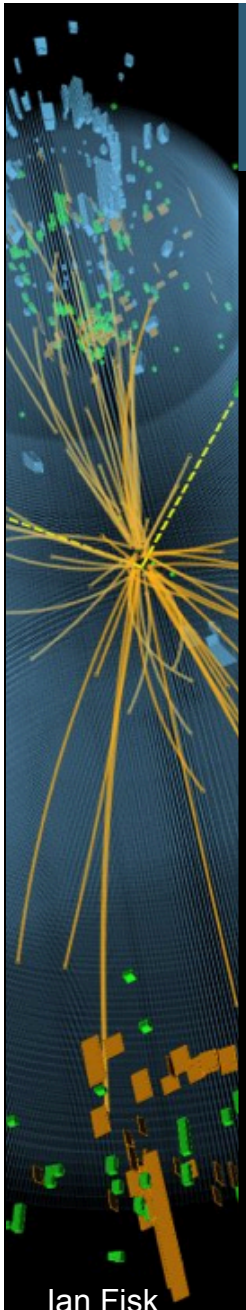
Evolution

- In the new resource provisioning model the pilot infrastructure communicates with the resource provisioning tools directly
  - Requesting groups of machines for periods of time
- A couple of improvements
  - Larger community of people is working on things like Open Stack rather than the CEs used by WLCG
  - The current architecture of authenticating every pilot individually makes little sense
  - Already we try to balance resource provisioning on longer time scales than a single job with MUPJs, but for most of the facilities that support WLCG the provisioning of resources could be even more coarse



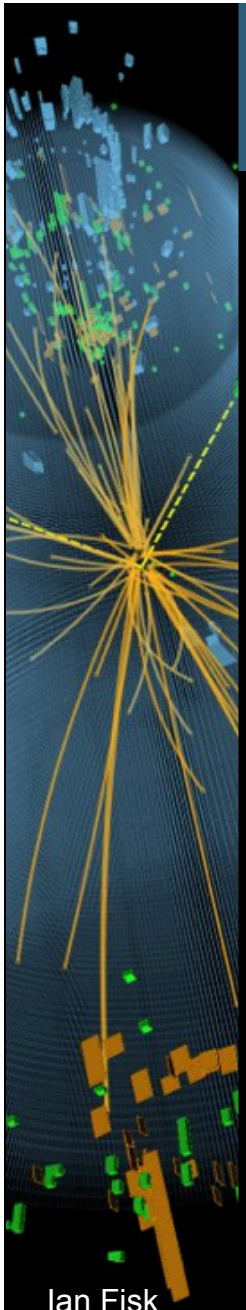
Trying this out

- CMS and ATLAS are trying to provision resources like this with the High Level Trigger farms
  - Open Stack interfaced to the Pilot systems
- In CMS this was work from Cloud Experts in IT-ES/VOS and work from the Glide-In WMS Team and workflow submission
  - We got to 3500 running cores and the facility looks like another destination, though no grid CE exists.



End Result

- I think you are going to see the resource provisioning tools change
  - You will see some sites buy their resources from a provider, and not want to bring up a Tier-2 facade
  - There will be broader contribution to tools like Open Stack, and the cluster management and the resource provisioning will begin to merge
  - The work done on breaking down the barriers between storage systems will be critical to make the system work, because in these resource provisioning models it can be very hard to tell where the hardware actually is



Outlook

- I think the general outlook for is
  - a breaking of the boundaries between sites
  - less separation of the functionality
  - the system will be more capable of being treated like a single large facility, rather than a cluster of nodes
- This will be more flexible and efficient, and able to incorporate other types of resources

Ian Fisk
FNAL/CD