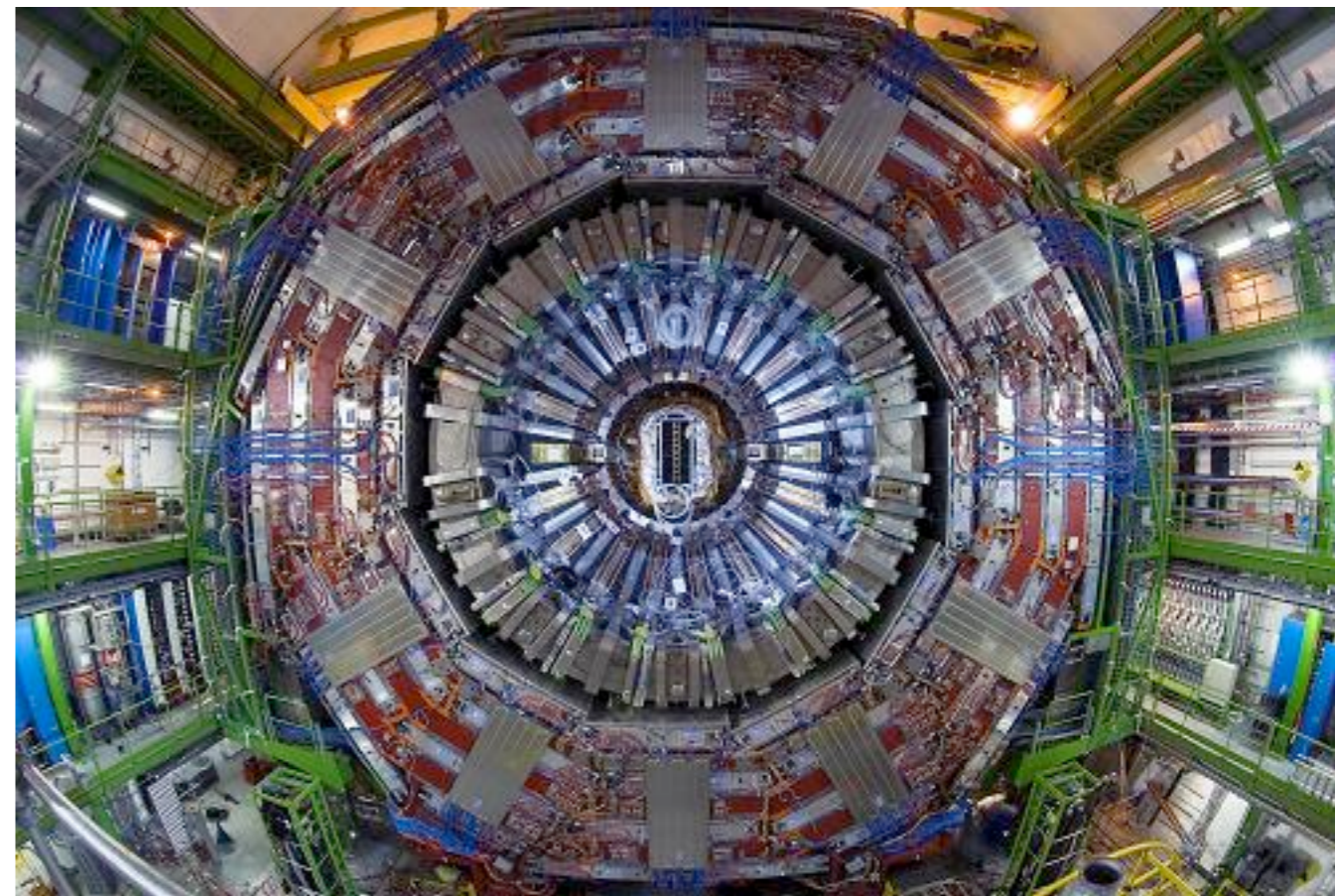
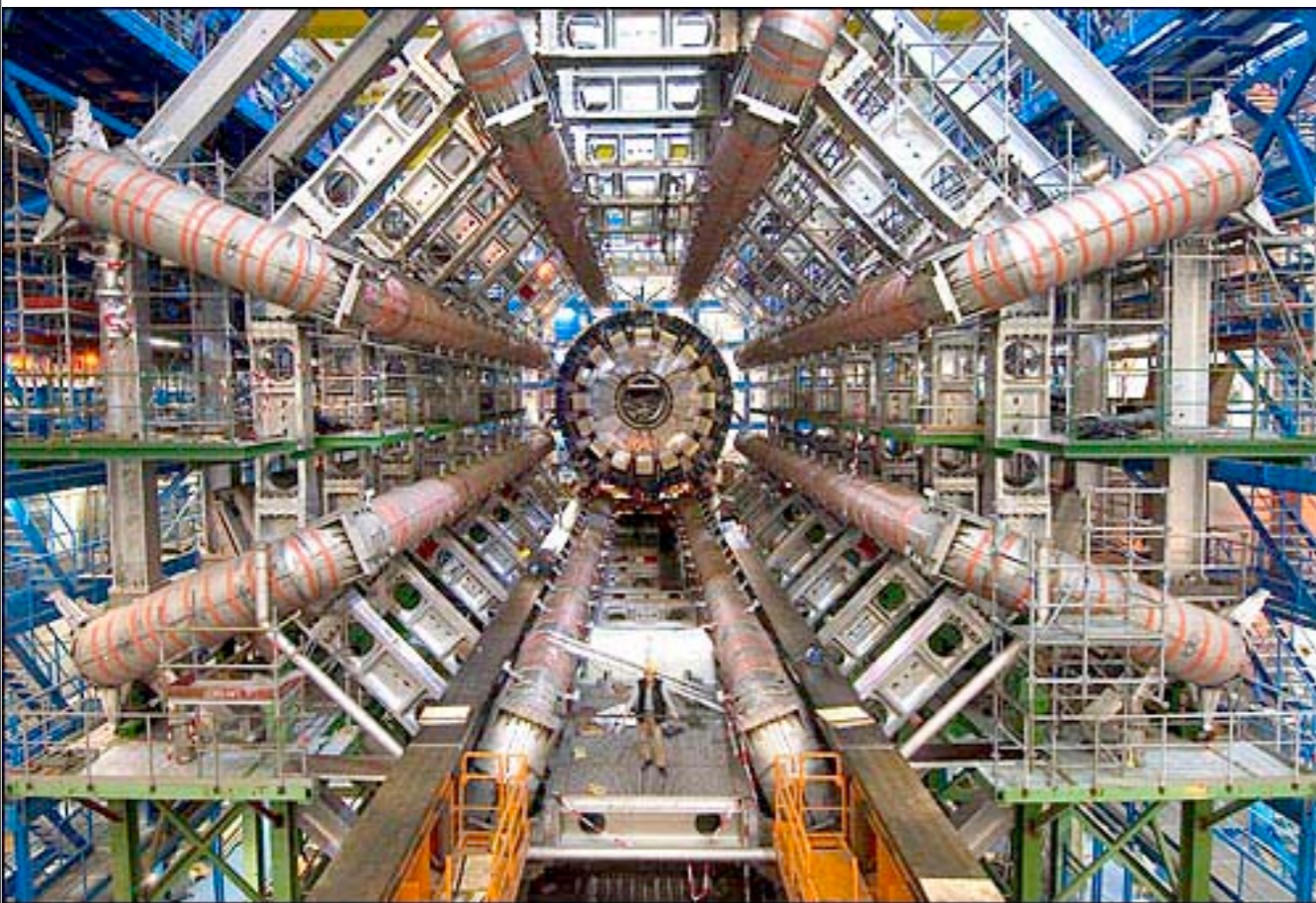


Worldwide Production Distributed Data Management at the LHC

Brian Bockelman
MSST 2010, 4 May 2010

At the LHC

- <http://op-webtools.web.cern.ch/op-webtools/vistar/vistars.php?usr=LHCI>
- Gratuitous detector pictures:



Forewarning...

- The LHC is a particle accelerator with 4 primary detectors.
- ATLAS, CMS, LHCb, and ALICE
- Each detector is massive, and the collaborations are massive (CMS and ATLAS have over 1,000 workers each).
- I work for CMS, so my knowledge is biased towards that experiment.

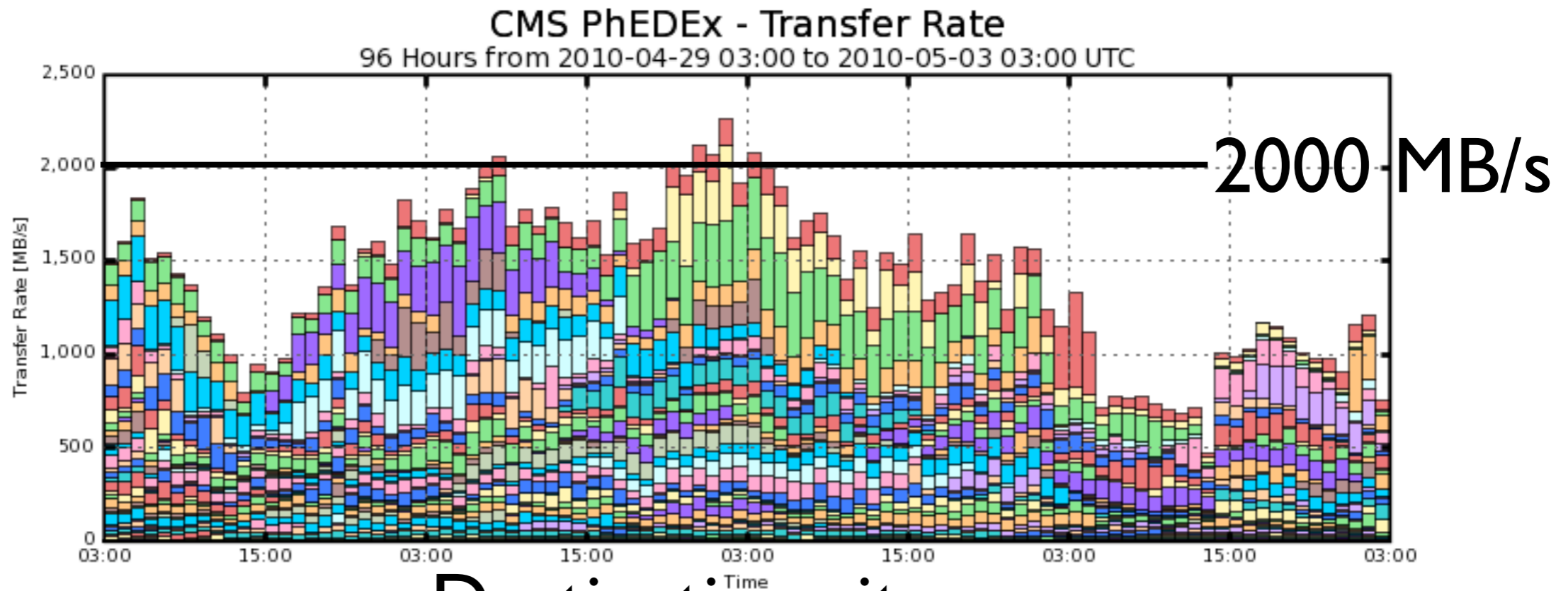
Mission for Today*

*all in about 25 minutes

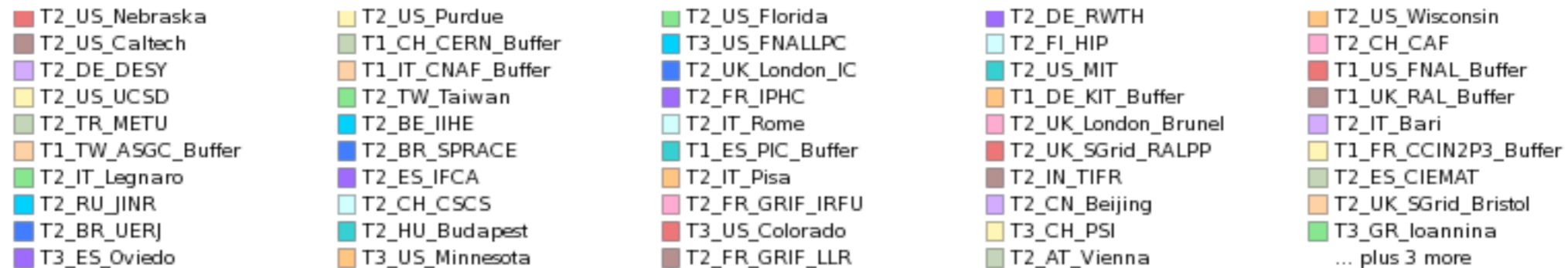
- Today, I want to convince you that:
 - LHC experiments have a massive data management challenge.
 - Data management is based on files and explicit file movement / pre-placement.
 - After many prototypes, the LHC formed around a common hierarchical model.
 - We have evolved this model in the last 5 years.

The Problem

Setting the Scale



Destination sites:



Maximum: 2,266 MB/s, Minimum: 475.62 MB/s, Average: 1,412 MB/s, Current: 749.22 MB/s

CMS on a Sunday - 52 sites (most are run on an 8x5 basis) regularly moving data at over 2200 MB/s.

Setting the Scale

- In the last 30 days:
 - 2.5 PB moved
 - 69 sites worldwide
- Overall:
 - 16.7M files in the system
 - 28 PB of replicas in the system.
- Only 1.5 months of serious data taking.
The “data management” challenge is just beginning

Data Management at a LHC experiment

- Everything fits neatly into one of two categories. Today, we discuss “offline”

Online

- Centralized
- Detectors are on this network.
- Computing work critical to operation of detectors
- Highly secured, highly walled-off from outside world.



Offline

- Very Large - thus massively decentralized
- Majority of computing resources.
 - Majority of users.
 - Processing, archival storage, analysis of data, etc.
- “the grid”

Offline Activities

- **Data-Ops:** transformation of raw data to usable data formats, grouping data according to physics content.
- **Simulation:** Simulating collisions in the detectors.
- **Analysis:** Filtering data for interesting events. Turn to simpler data format.
- **Physics:** Deep understanding of event content. Interactive and done on laptops.

The Solution

CMS Circa 1996

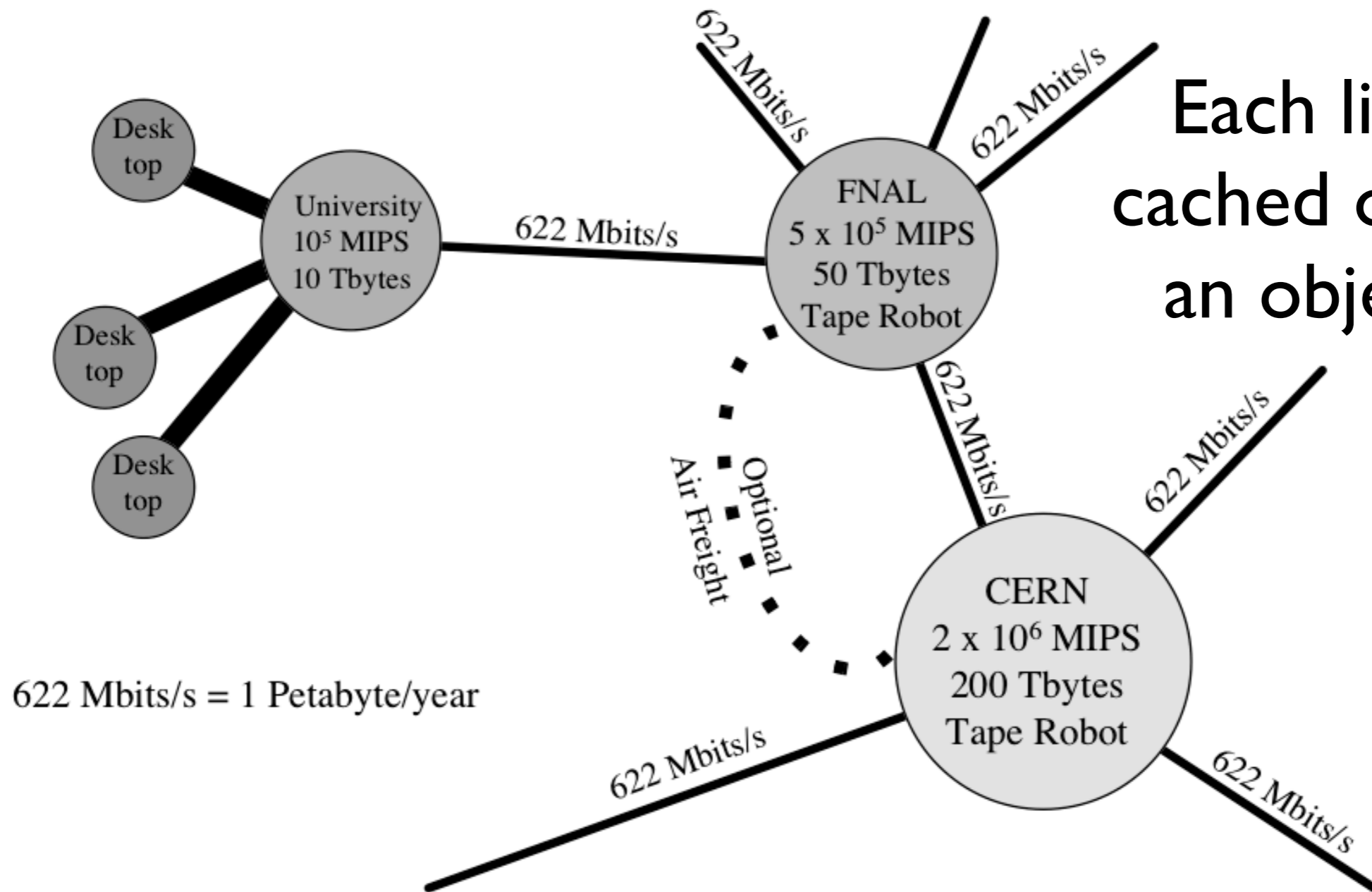


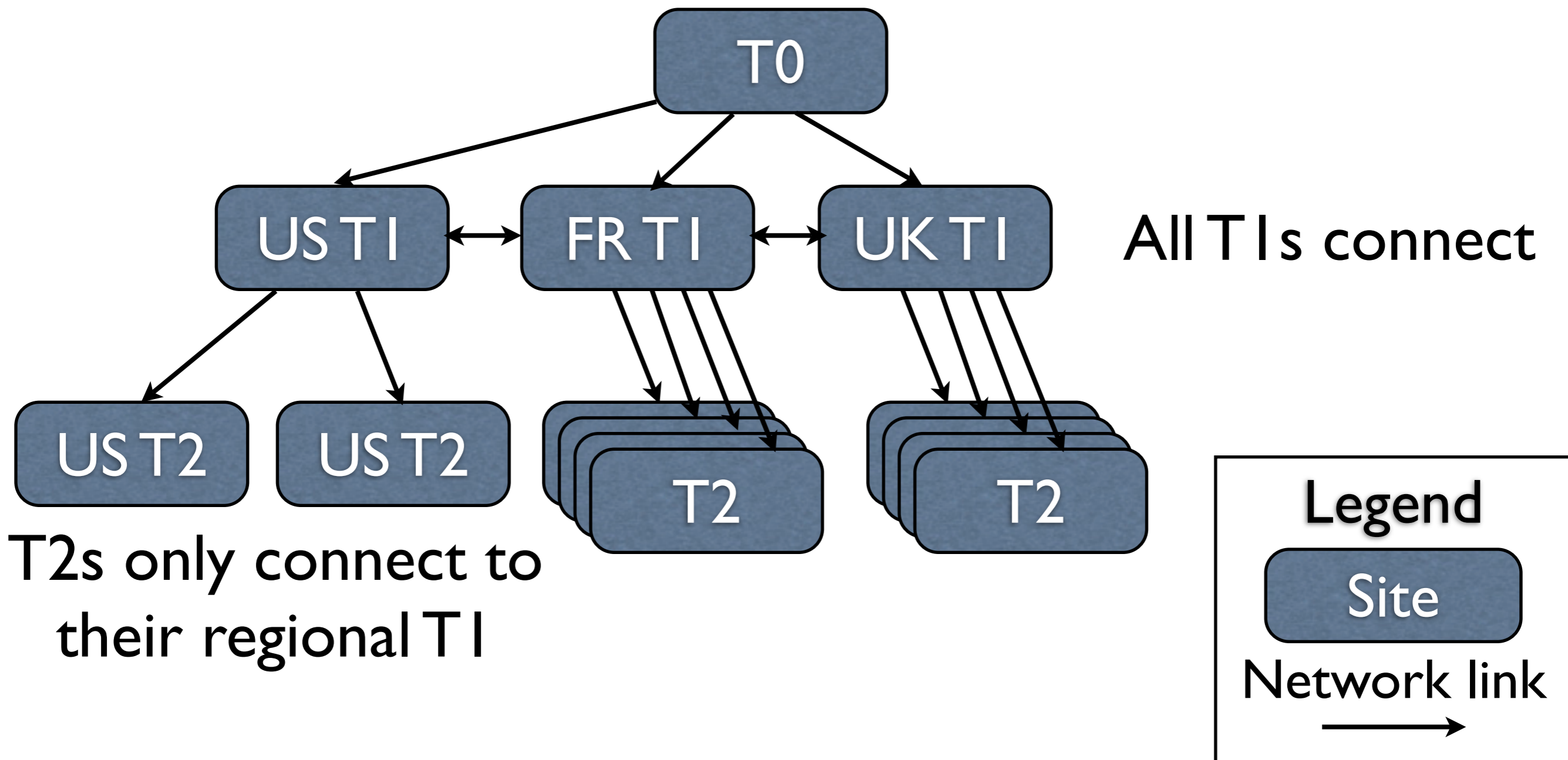
Figure 3.3: An example of how the CMS computing model might serve remote physicists in about 2005.

LHC, circa 2001

- The LHC experiments prototyped a tiered model of data distribution around 2001, based upon “grid” technologies.
 - T0: CERN (>5000 cores; many PB of disk and tape)
 - T1: Regional centers around the world (1000-5000 cores; PBs of disk and tape).
 - T2: Smaller, mostly university computational centers. (300-1000 cores; 100s of TB of disk)

LCG hierarchy

- This diagram has been re-drawn hundreds of times in thousands of presentations.



The WLCG

- In addition to the tiered data distribution, the LHC experiments adopted an idea of a common computing fabric - the “Worldwide LHC Computing Grid” (WLCG).
- Composed of smaller grids with names like EGEE (EU), OSG (US), NorduGrid.
- Took awhile to figure out what “the grid” means.

LCG Data Management

- In the tiered model, data moves downward*.
- The #1 axiom of LCG data management is that jobs move to data.
- The indivisible unit of data in particle physics is an “event” - basically, a readout of the detector when 2 particles collide.

*experiment data; simulation data is the exception, but it is order magnitude smaller

LCG Data Management

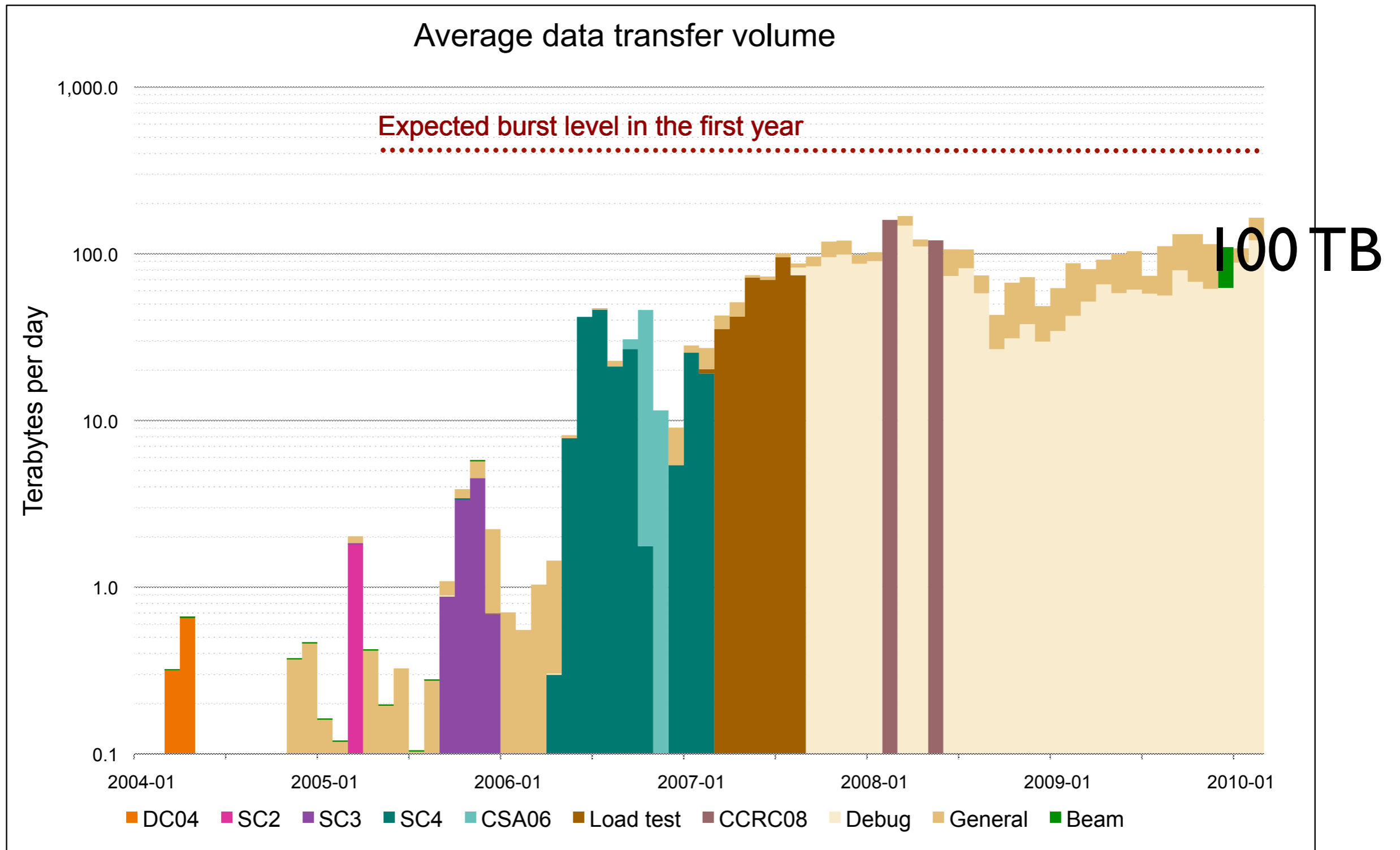
- Events are grouped into files.
- This is the smallest unit computing deals with.
- Data is explicitly managed in terms of files moving from point A to point B.
- The files/explicit movement metaphor was in response to the field's 5-10 years of object database/streaming experience:
- Wonderful reference: <http://www-db.cs.wisc.edu/cidr/cidr2005/papers/P06.pdf>

LCG TDR, 2005

- From 2001 to 2005, LCG experiments built lots of research/prototype around this model.
- In 2005, the LCG Technical Design Report set out the “baseline services” for the grid.
- Idea: during 2005-2007, these services into production through a series of increasingly-large “services challenges”

Service Challenge

Logarithmic scale; TB / day



LCG Baseline Services

- The LCG has the following data management services:
 - *File Catalogues*: Registers file replicas centrally.
 - *File Transfer Service*: Given two URLs, transfer from point A to point B.
 - *Storage Element/SRM*: Ability to interact with large storage systems through uniform grid protocol.
 - Databases for “detector conditions” - configuration of detector subsystems at a given point in time.

Data Placement

- Each experiment builds their own services on top of the baseline, which might represent different approaches:
 - CMS: About 1/2 of data is subscribed centrally; about 1/2 is subscribed by physics groups. Fairly decentralized!
 - Full mesh of transfer links (inc. T2 to T2).
 - ATLAS: All data centrally subscribed. Transfer links same as original model.
- Both experiments use the *File Transfer Service* (FTS) baseline service to automatically move the data once the experiment decides the source and dest.

Crisis in Storage

Problems in Data Management

- Implementations reveal many issues:
 - Explicit movement of files is hard. Everything can fail! This confuses and frustrates users to no end.
 - Explicit placement implies keeping location data (file catalog). Global file catalogs are almost always out of sync with our local systems.
 - We are not a walled garden - we must (safely) export data off the grid to laptops and institutes.

Catalogs, oh my!

- We have up to 3 layers of catalog info:
 1. Each experiment maintains a catalog of file/dataset locations.
 2. Sites maintain a mapping from experiment name to filesystem name.
 3. Filesystem maintains mapping of names to location on one of many disk servers.

Catalogs

- Each transition between catalog introduces inconsistencies.
- Every day, humans must deal with these by hand.
- Software produced by three different layers - experiment, grid, and site - and they don't auto-sync with each other.

Solutions? None yet...
Glad we didn't design DNS!

Get the data *out*

- What to do with the output of user jobs?
- Getting the data off the grid and closer to users is essential. Science happens when “playing with data” on laptops.
- Downloading data must be brain-dead easy, so Joe Q Physicist can download data. HTTP(s)?
- Some users have institutional computing resources larger than our data centers.
- We must be able to protect ourselves from (inadvertent) DOS attacks.

Our storage elements are a bottleneck for users - good (protects sites from overload) or bad (slows science)?

Conditions distribution

- Conditions are metadata about the detector when the event data was recorded.
- Relatively small, can change quickly; being constantly improved.
- Experiments started by streaming database contents to sites.
 - This turned into copying SQL-lite (or other format) database files around.
- A huge pain to track lots of small files!

Conditions distribution

- Conditions data is now transported through hierarchy of HTTP caches.
 - Top-level is a Java web servlet at CERN.
 - Each site has a local squid cache.
 - Data transfer and access is 100% transparent to user, fast, and repeated accesses are low-latency.
- Thought: Can such a system scale from metadata to the real event data? Users understand how to use HTTP(s) more than grid protocols!

Can we think of the LHC as a content-distribution network?

Futures

- In the 1990's, we had a distributed object-oriented database model of DM.
- In the 2000's, we had a hierarchical(ish) explicit file management and movement model.
 - Solutions to “central workflows” are in hand.
- In the 2010's, it is about the physicist!
 - Implicit instead of explicit placement?
 - Reduce complexity to increase reliability for the end-user?

Futures

- We have a hard rule about data being in place before jobs are sent.
- Not very cloud-like. Done because in 1996 we thought the networks were fragile.
- But today's networks are pretty impressive!
- How can we take advantage of opportunistic CPU cycles if we can't stream data?
- Can we start jobs if 95% of the data is in place? 75?
- Is it quicker to retransfer files from another site instead of pulling it from tape? Should one copy always be on disk to avoid tape latency?

Conclusions

- The LHC has been working on managing data at the petascale for 15+ years.
- This led to a common model that solidified in 2005, and we have been working to make it “production ready” since.
- The CS research is 1/5 of the effort necessary to make something reliable in production.
- There has been continuous evolution of our data models to this point - and all signs point to continued innovation.

Departing Thoughts

The most important physics is done on the laptop, not the cluster. Data management on the LHC is about reliably and quickly transforming 1 PB of data to the 1 GB the user cares about.