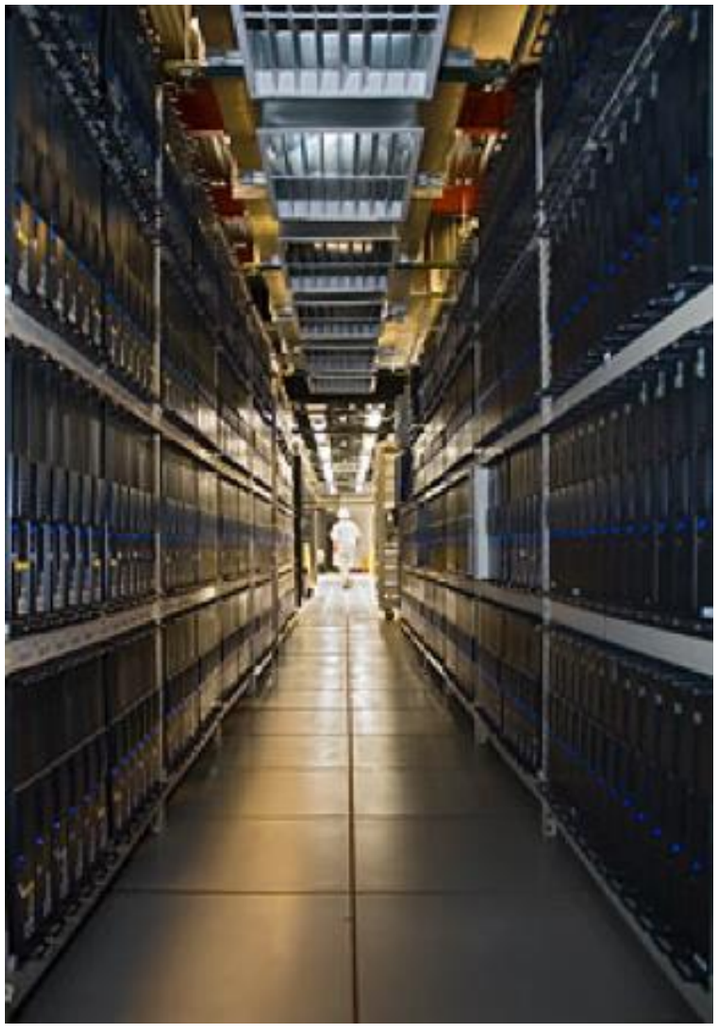
The Future of Distributed Data and Workflow Management

Ian Fisk
CHEP2016
October 10, 2016

Today was the Future at some point in the Past

The present is shaped by a couple of big changes in the past

- Move to Linux clusters
  - Suddenly systems were dramatically cheaper and faster but more complex
- Move to distributed computing
  - Support of computing



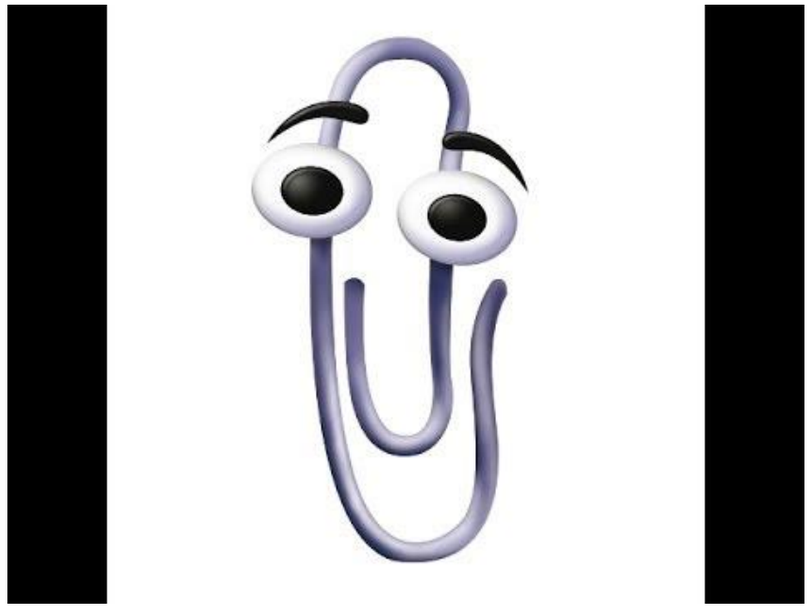
Looking forward

The path to having “visions” turns out to be well established

- I tried to prepare accordingly for this talk

Place of reflection

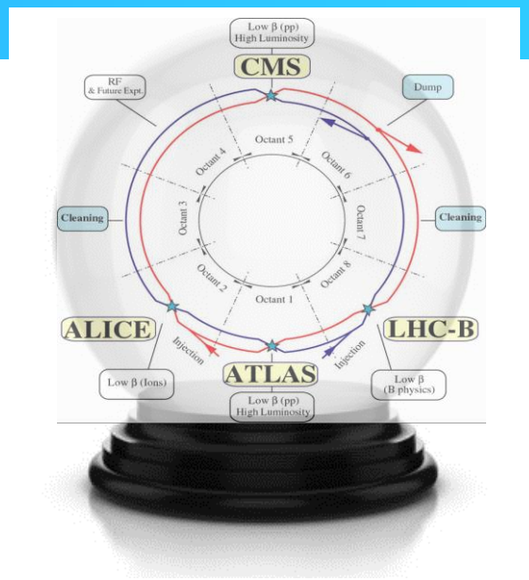
Psychedelic Agent



Spirit Guide

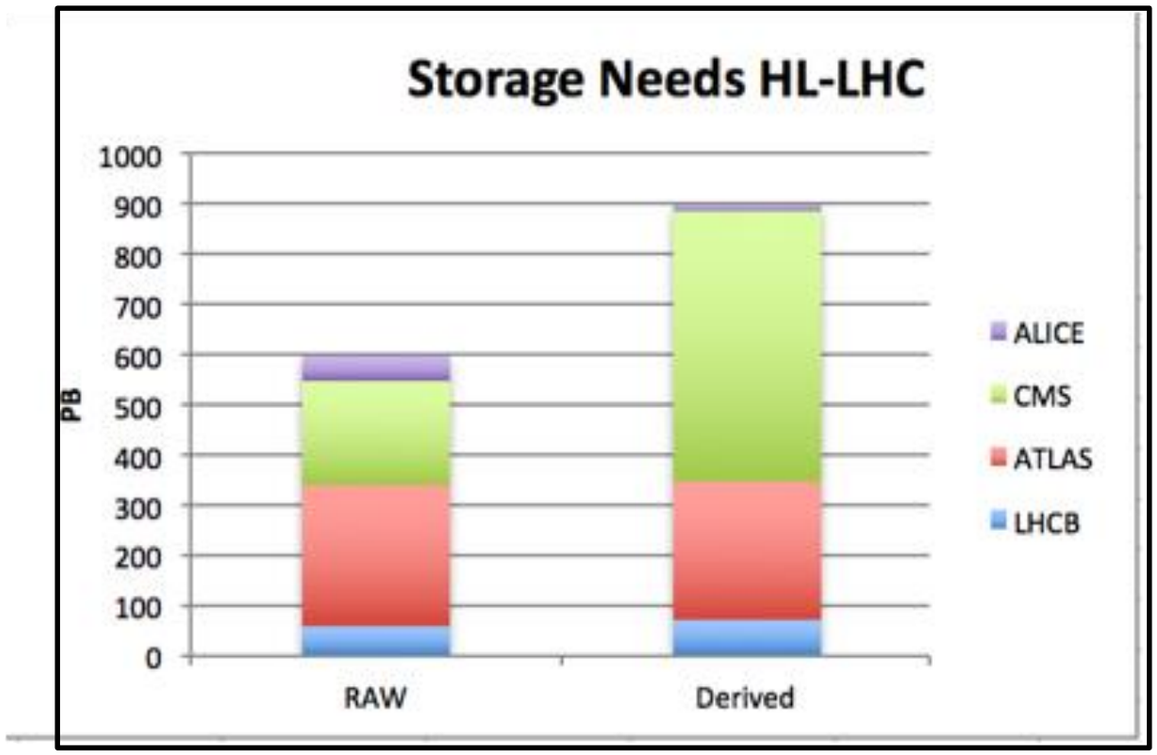
The Future

We don't know many things about the future, but there will be an HL-LHC

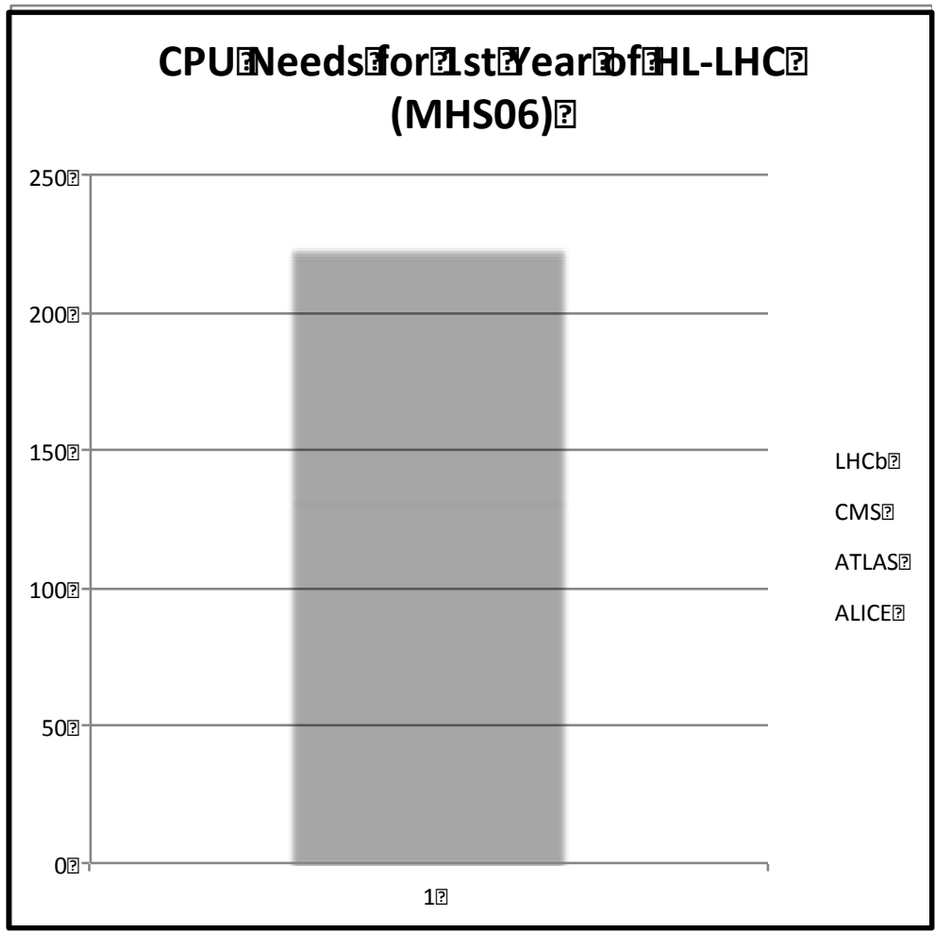


By the HL-LHC we can expect the technology improvements will give a factor of 6-10 improvement

- With flat budgets at 20% per year



About 10x more data



About 60x more processing

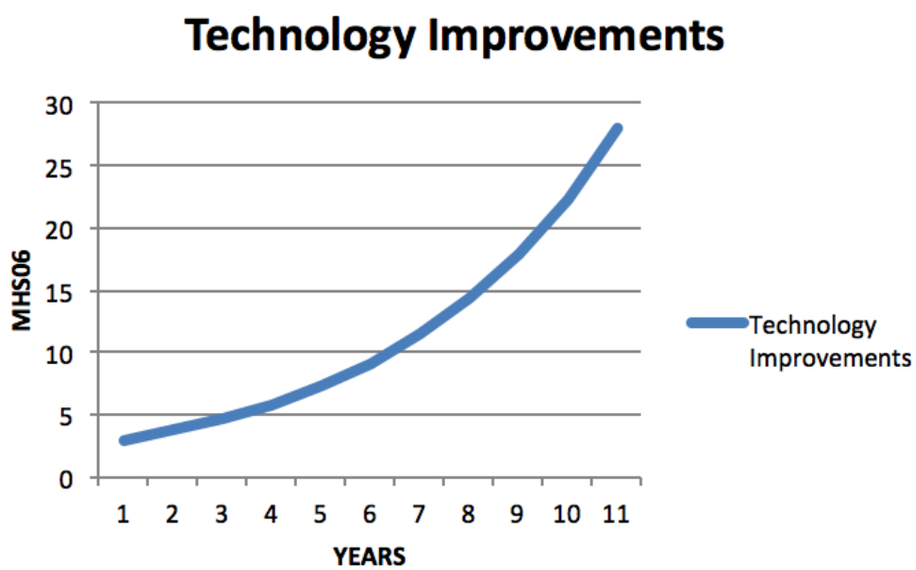
Ian Bird WLCG2016

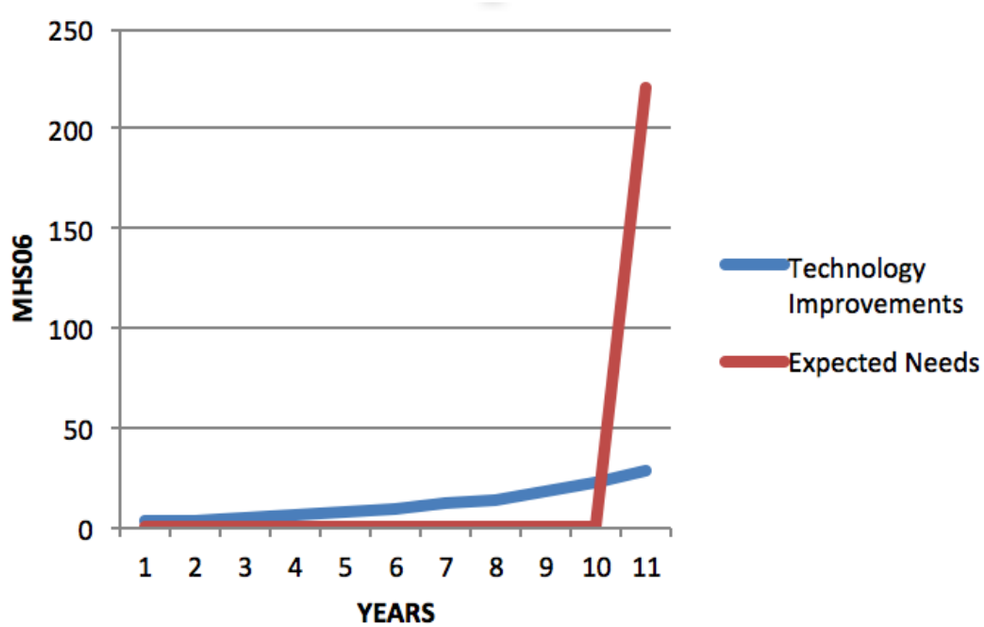
Evolution

Storage grows by a factor 10 in both raw and derived data

- Could be accommodated with technology Improvements

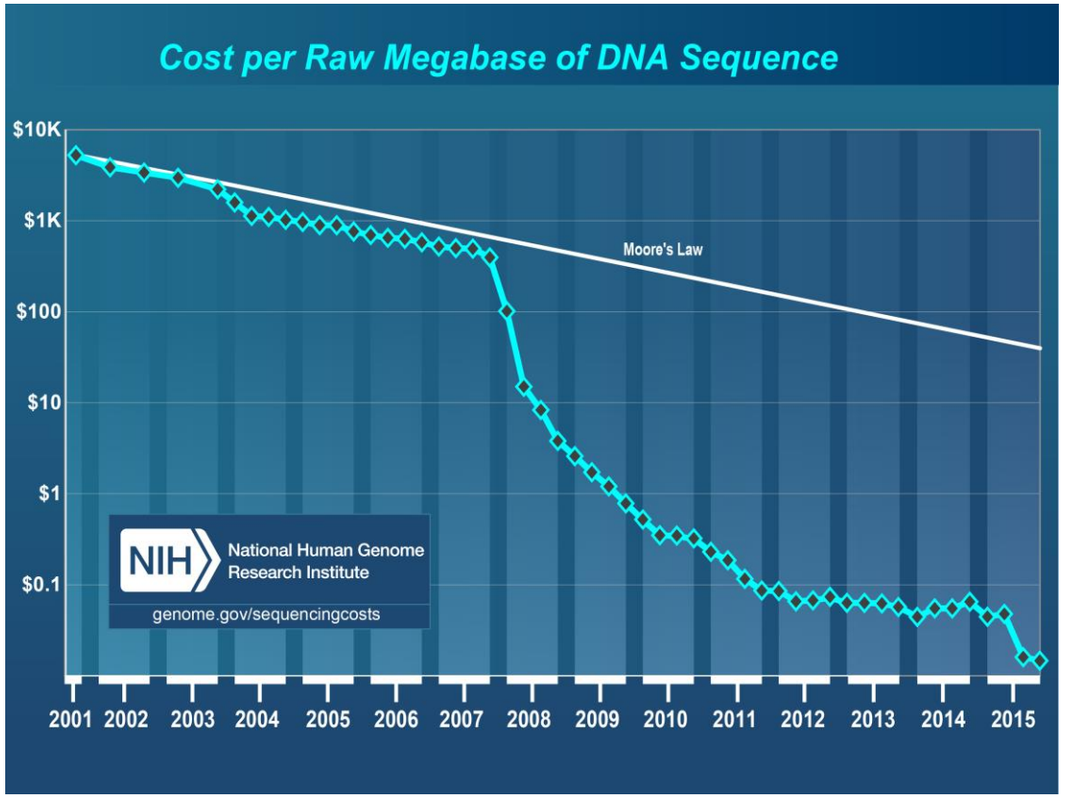
Processing requires 6-10x more than what technology evolution provides





We are not alone in this problem

- Other sciences grow faster than Moore's law too

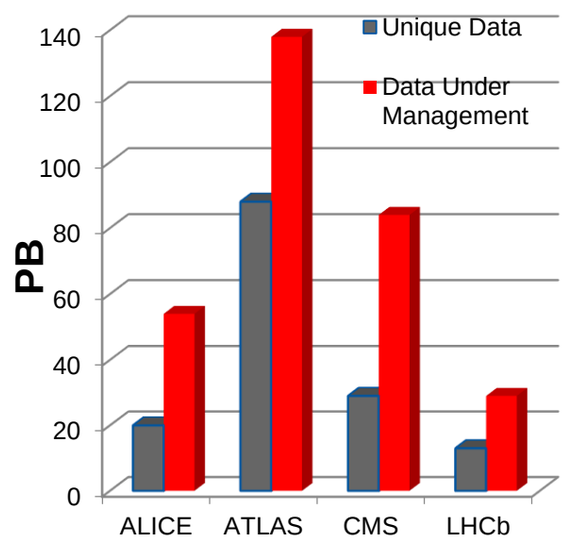


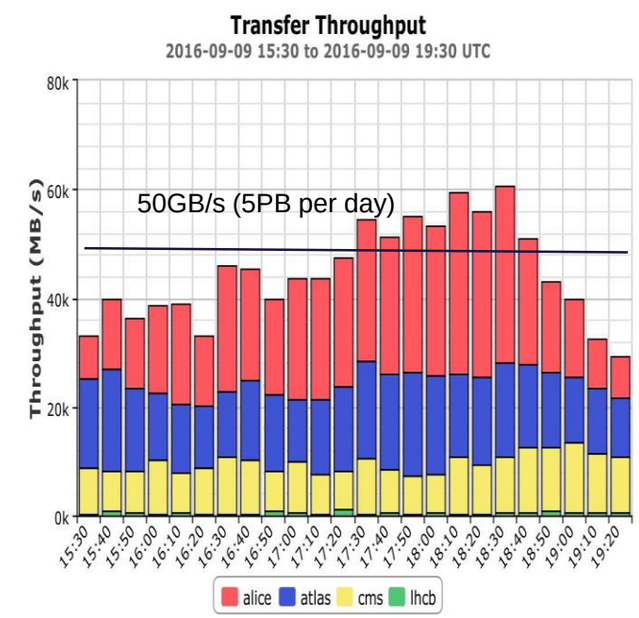
What are we actually good at

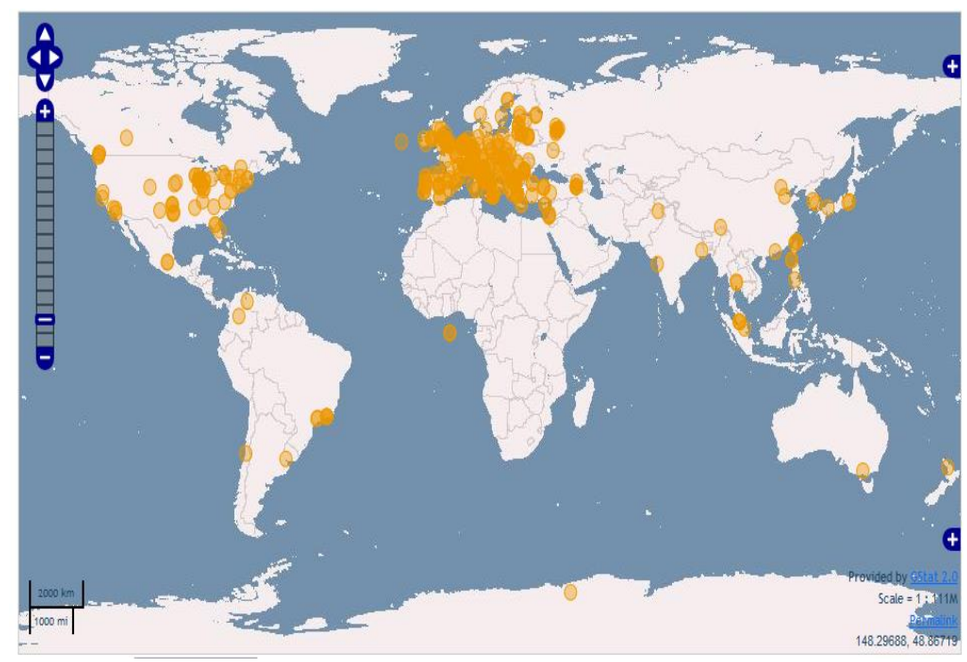
HEP as a field is good at managing and distributing data

- Datasets are large but custodially kept and protected
  - We make dynamic use of tape systems
  - We move to hundreds of sites
  - We make effective use of global network links

We remain leaders in this challenging areas







What are we not so good at?

We are not good about doing easy things cheaply

- We build up a bit system from many small sites
  - Big infrastructure providers distribute only enough to not be taken off line by the same natural disaster
- We use a very educated workforce
  - We value innovation, we hire people interested in science
- We have high overheads and infrastructure costs
  - We locate at large laboratories and we did not optimize the placement for energy utilization or cost
- We like research
  - We do not optimize for production scale computing for the lowest possible cost

Strengths of others

A number of science communities have moved to overflowing into commercial processing and storage offerings

- Generally these have been groups with lower capacity needs and less experience operating dedicated computing

Other communities have gotten farther on the adoption of big data tools

- Spark and Hadoop for data access
- Python notebooks for work, documentation and reproducibility
- Containers for interactivity and analysis preservation

Some adoption of HEP tools in related fields, but not broad distribution

Relative Size of Things

Processing

Amazon has more than 40 million processor cores in EC2

Google has ~1M servers so ~20M cores

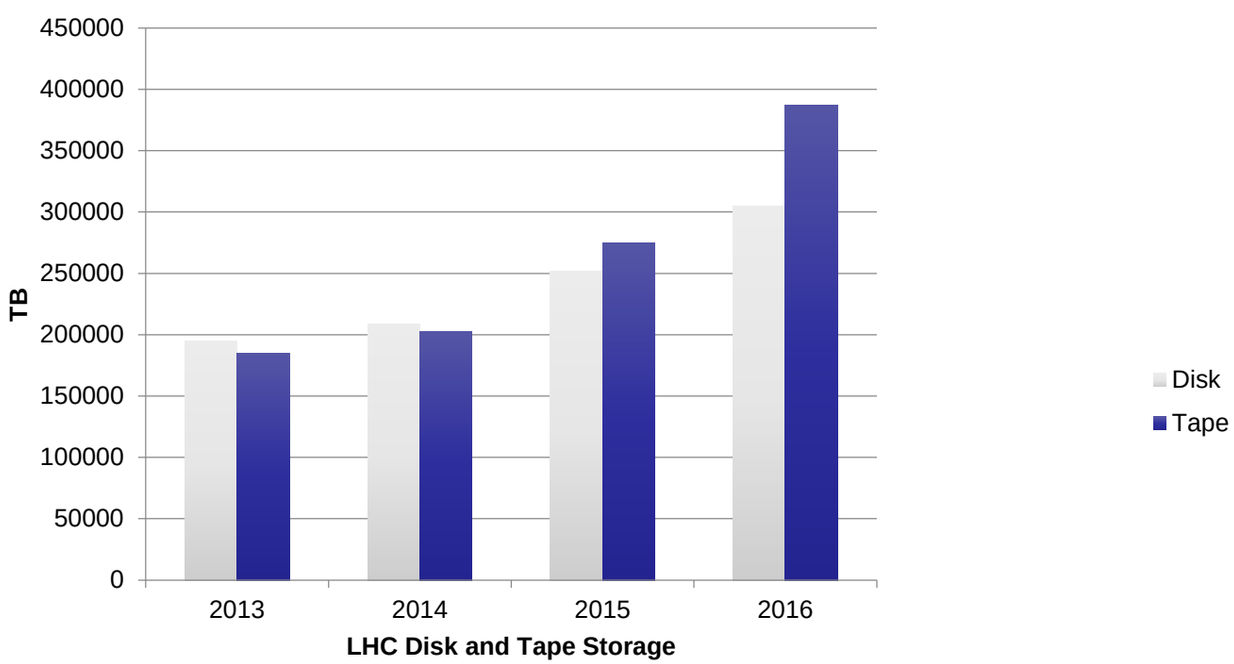
Storage

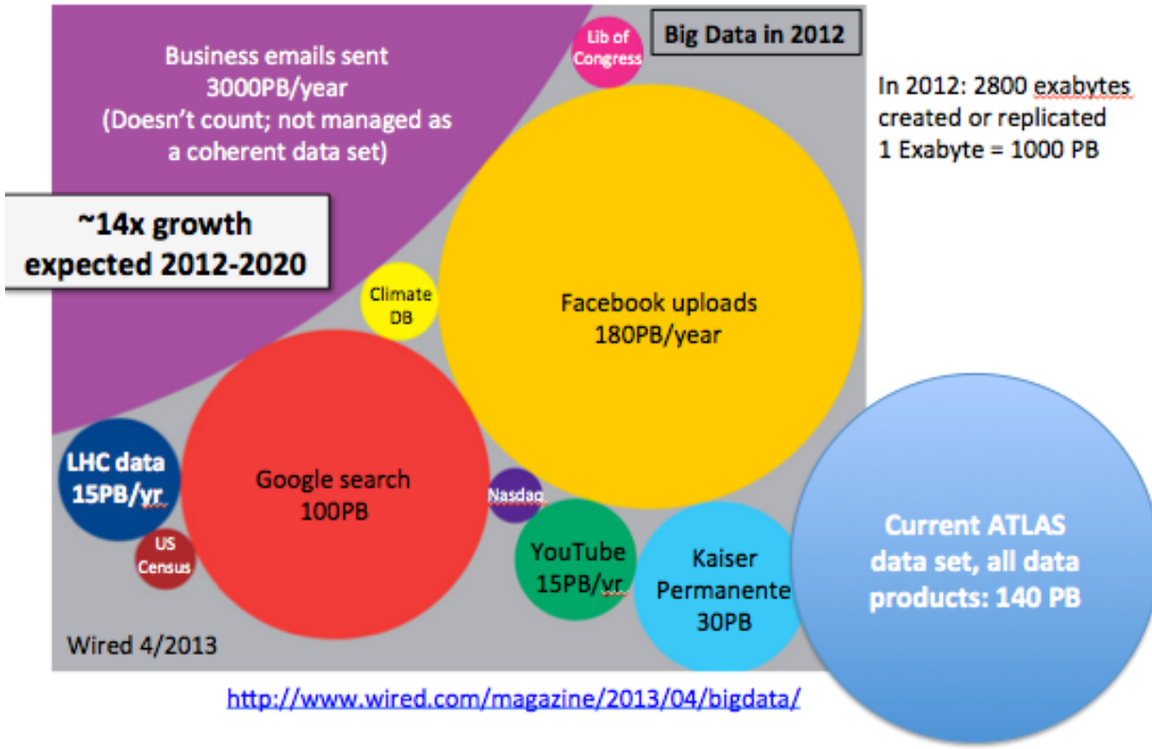
Amazon supports millions of queries per second

Google has 10-15 exabytes under management

Facebook 300PB

eBay collected and accessed the same amount of data as LHC Run1





http://www.wired.com/magazine/2013/04/bigdata/

Our data and processing problems are ~1% the size of the largest industry problems, but we still distribute more data and lead in the area of data management

Evolution of model

Commercial providers are evolving and growing incredibly fast

Starting with leasing a house model

- Reserved resources that were expensive and normally replaced with owned dedicated local resources

Moved to the rental car model

- More expensive to use than things you operate all the time, but competitive for bursts

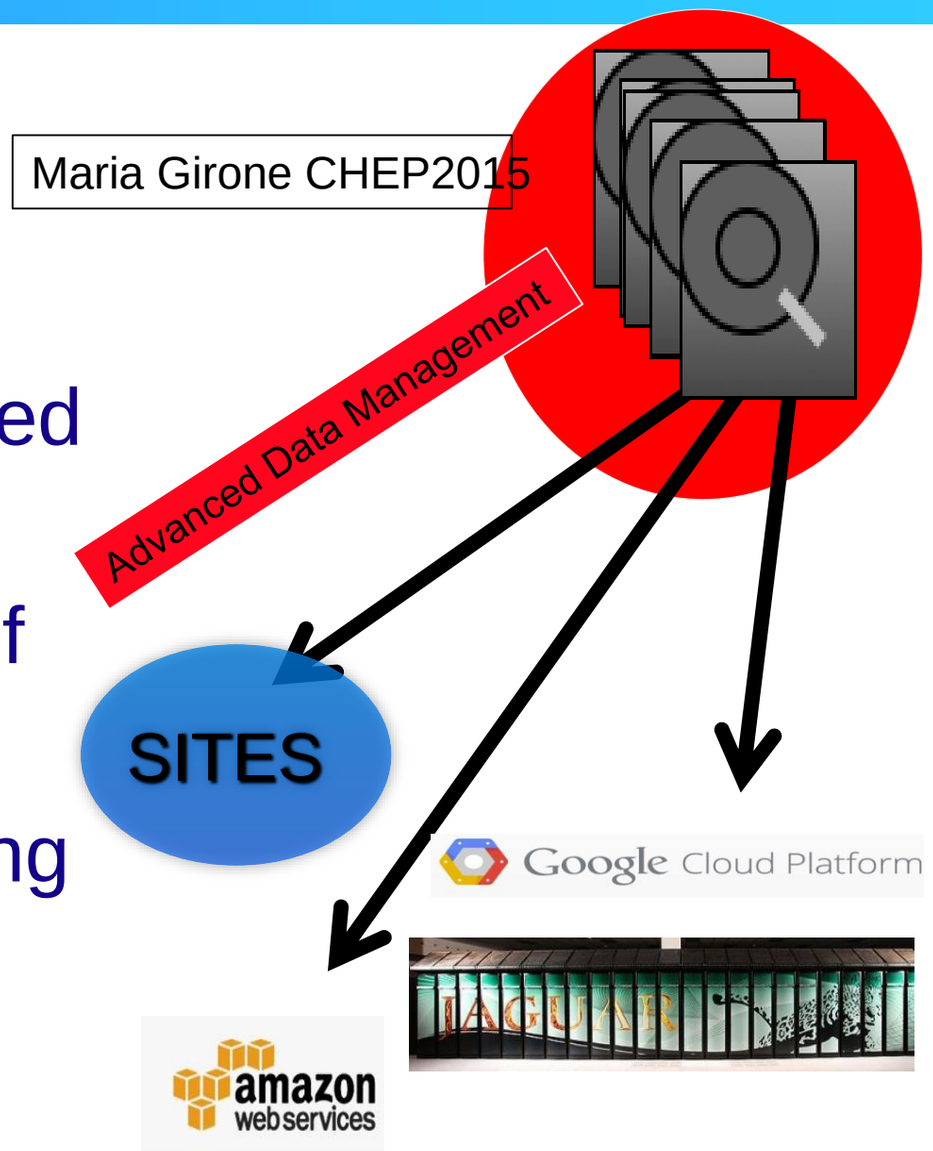
At some point in the not so distant future we will be in t Utility Provider model

- Economies of scale will mean then can just provide simplest computing more cheaply

Modifications in Data Management to Join

No external provider should be willing to accept custodial responsibility for our data

- HEP Data is valued at the cost of the accelerator detectors and operations divided by the running time.
  - Works out to hundreds to thousands of dollars a second
- Once we couple the storage and processing we lock into provider
  - Lose our leverage



- Storing data is done all the time while you only pay for processing when it's used
  - Lose the cost advantage
- We need to be able to deliver data to resource providers dynamically

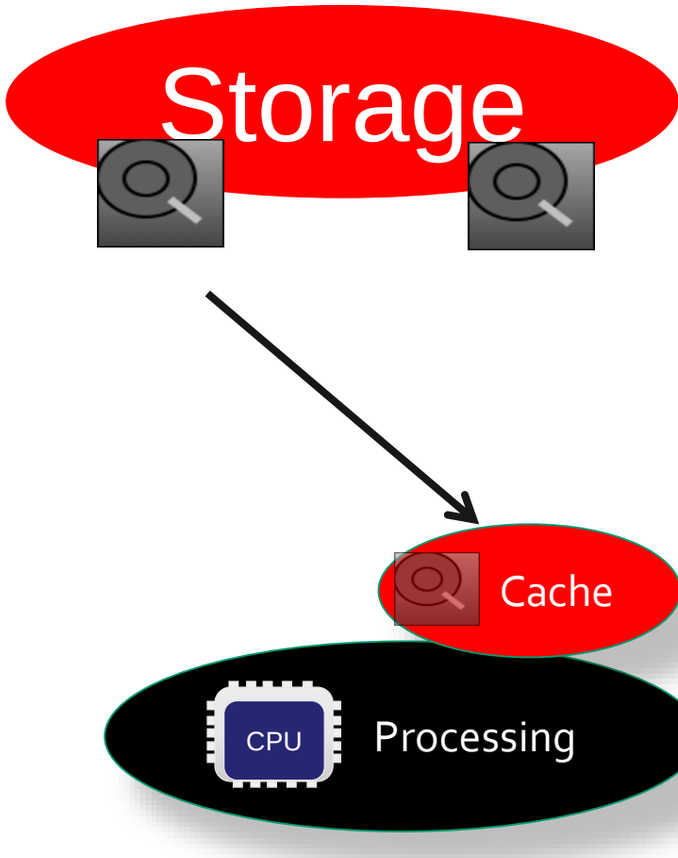
Data Management Changes

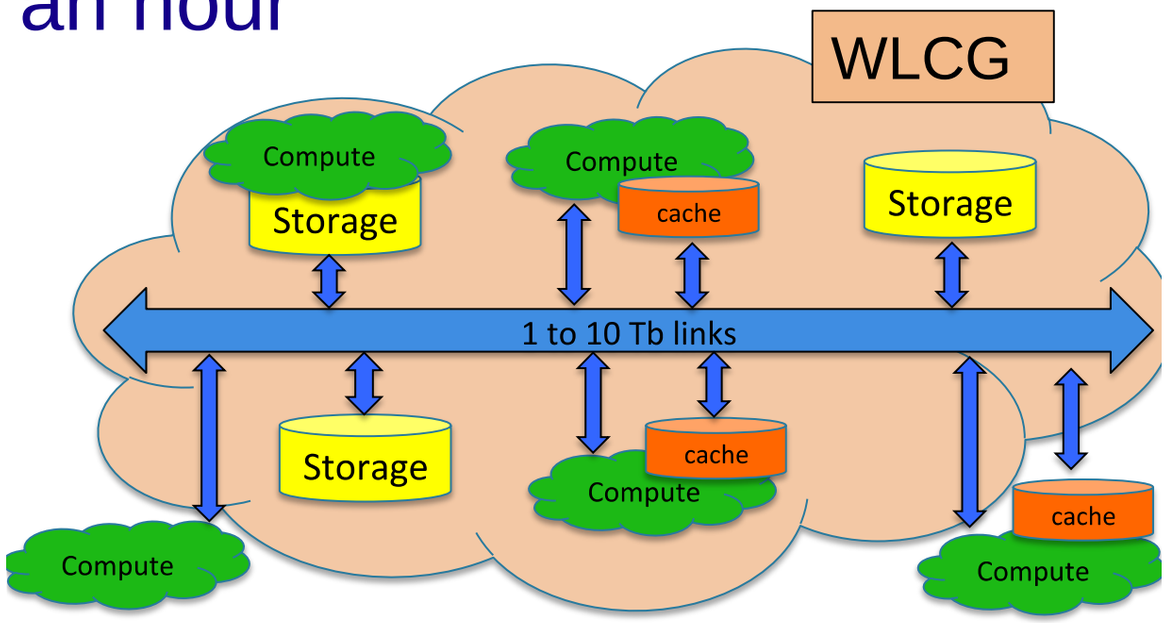
We want to keep control of the data

- Need to be able to deploy data to a diverse set of resources (Clouds, Dedicated sites, HPC Centers, etc.)
- Will need to be a combination of real time delivery and advanced data caching

In order to replicate samples of hundreds of TB in hours we will need the systems optimized end-to-end and a very high capacity network in between

- 100Gb/s is 36TB an hour





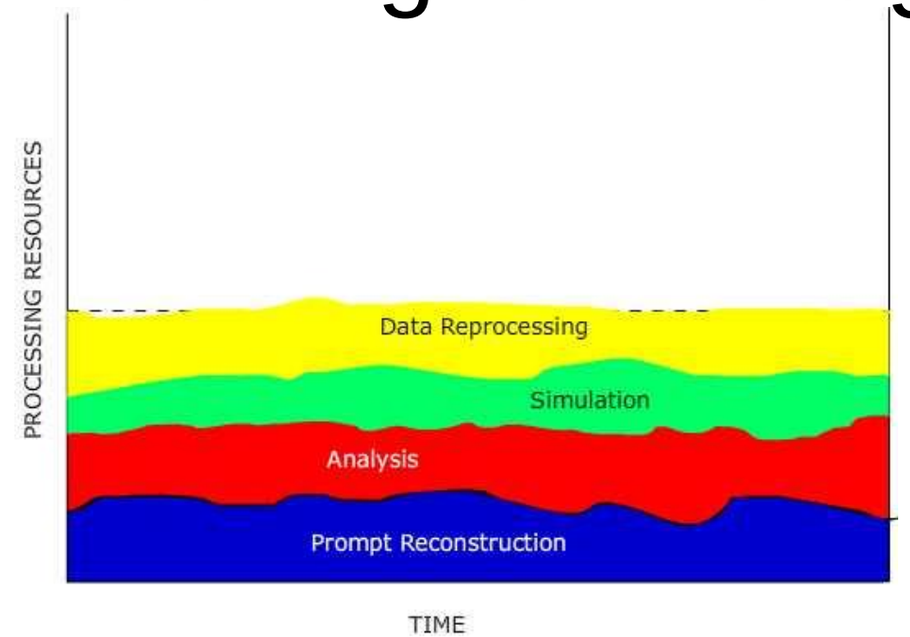
Simone Campana ECFA2016

Benefits

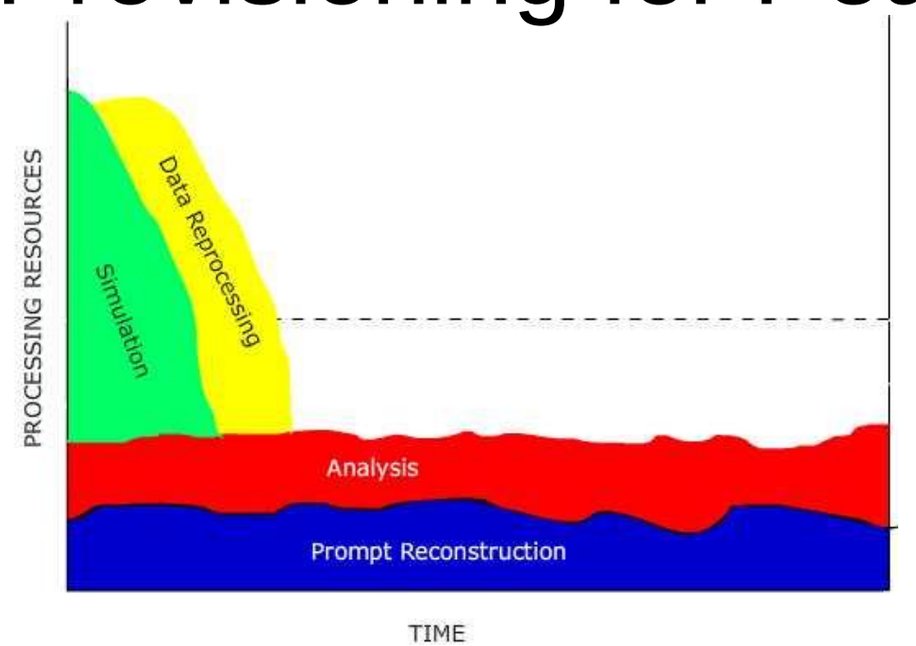
As expensive as computers are, people's time is more

- How we currently schedule processing makes little sense from the perspective of maximize the efficiency of people

Provisioning for Averag



Provisioning for Pea



One of the biggest improvements in joining to a much larger pool of resources is breaking the idea we need to lay out our resources for average load

- Workflows could be completed as they are defined and not over months

Impact on Workflow Management

In these processing models the workflow system needs to be able to scale to 5-10 times the average load

- We want to be able to burst to high values
- The least expense time to be delivered resources might be all at the same

If one is using commercially provided computing faults turn into real money

- Need to focus on potentially wasteful things
  - Infinite loops
  - Giant log output that trigger data export charges
  - CPU efficiency loss

- All things we probably should have been worrying about with our dedicated systems, but somehow when you are directly paying for

But we are still going to fail

Computing needs are growing faster than technology or budgets

- We are short by about a factor of 6 (Factor of 60 with 10 gained with technology)

Efficiency gains in scale will be measured in percentages not factors

- Getting access to alternative computing facilities, working more cheaply, buying some capacity will all help but a factor of 2 would be a miracle

Something has to give

- If we are factors from our resource needs, data will become very static
  - We will not have enough capacity to reprocess it
- We will be factors short on simulation and capacity for analysis users

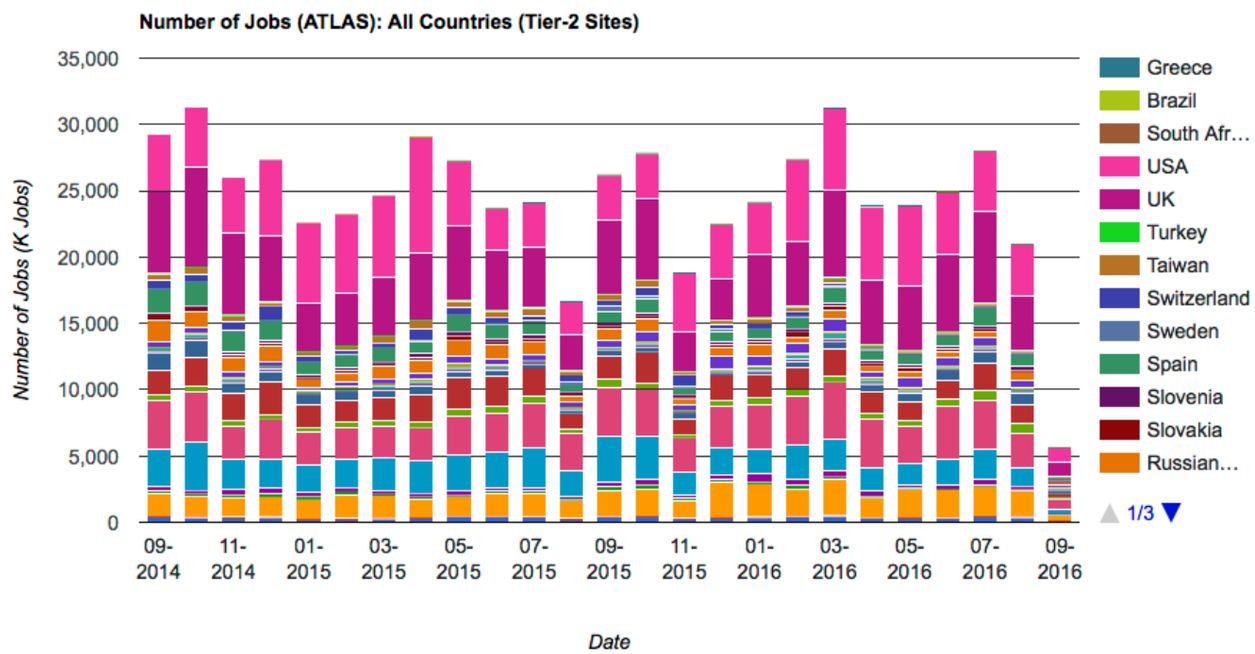
How to reduce

To significantly cut our computing needs, we need to rethink how we process and look at data

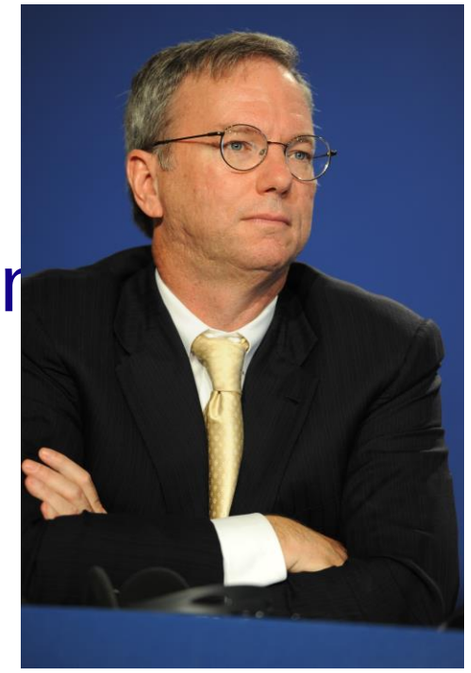
A significant fraction of our resources are spent selecting and reformatting data

- The task is done repeatedly by many users and groups
  - Big serial access to the data and creation of user defined data formats

10s of Millions of processing tasks a month



What Would Google Do (WWGD)?



Change most of the common analysis queries to map/reduce type indexing

- Allows reuse of the initial query access across users and groups

Concentrate access at a limited number of big sites

- Economies of scale
- Simplification of data management and reduce replicas

We should create data reduction facilities

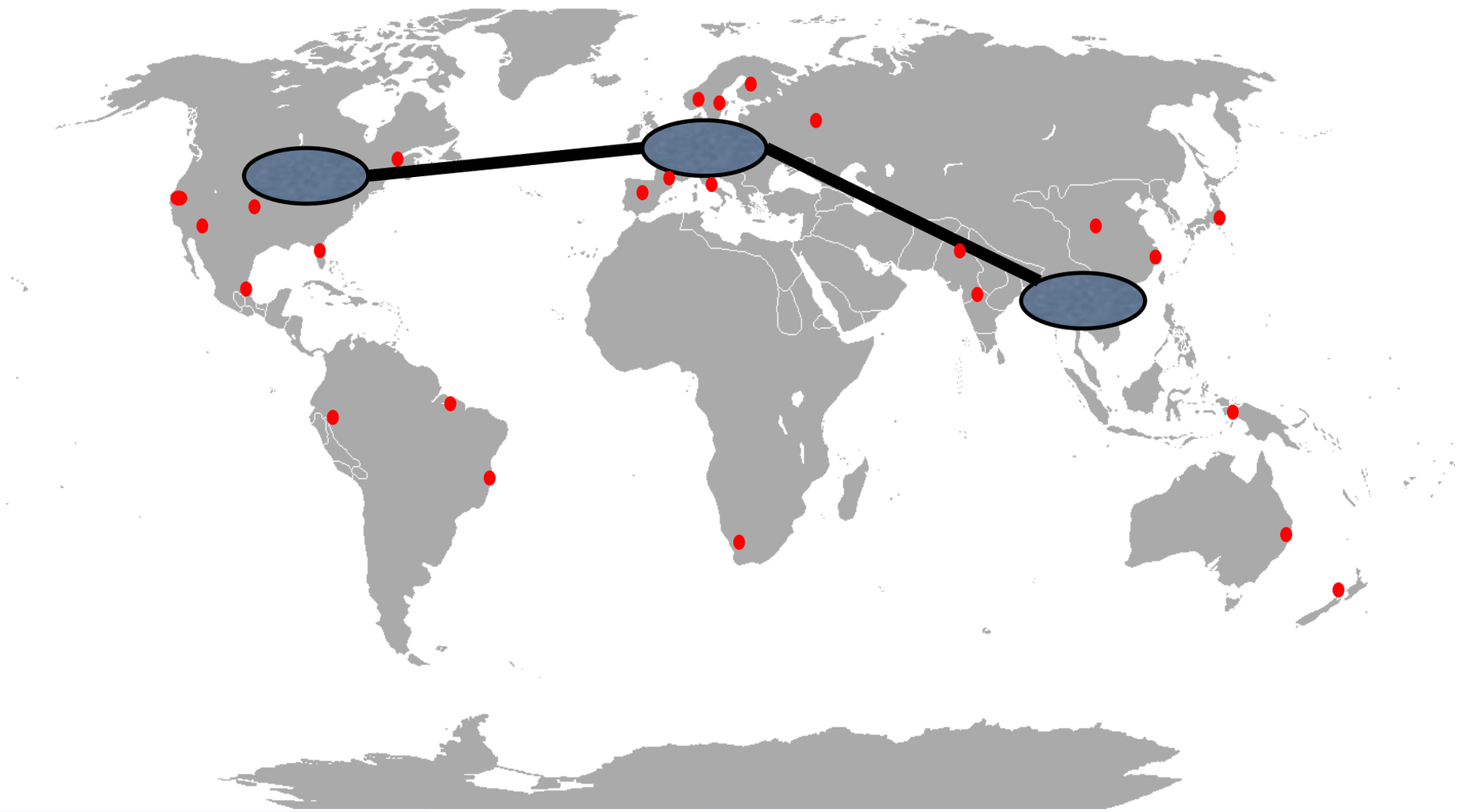
- Goal should be to support reasonably large numbers of users in reducing data by factors of 1000 in a matter of hours
  - PB goes to TB gets send to external resources

Potential for enormous gains by reusing and indexing

Workflow and Data Management

Big centers for data reduction impacts workflow and data management

- Data selection workflow sits on top of “big data” tools
  - Focusing effort on reproducibility and shared selection criteria
- Data Management involves moving small samples to end sites



And now for some blasphemy

It’s time we talk about how we process data

- Maybe the event processing loop model is not the only model

What if we think of processing as a set of data transformations?

What if we don’t need to process or reprocess every event?

Can we tie data reduction and data refresh together?

Can we store data as objects and not files?

Do we need to process all data in the same way?

Data Transformations

Data processing as a set of discreet data transformations

- Opens the possibility of new architectures
  - Not all transformations need to run on the same hardware
- If we store the data and the transformations, then processing can be done only when requested or when one of the dependencies change
  - Would naturally only consume resources for needed transformations on a selection of events
    - Reduces the need to process all the data all the time
    - Keeps data dynamic and refreshed
- Opens possibility for new languages and new external expertise

Changes to workflow and data management

Tying processing and data selection through discreet data transformations directly impacts data and workflow management

- Activity is triggered automatically
  - Needs throttling mechanisms
- The bulk of the data is placed at big sites
  - Reduced samples are moved and replicated
- Still a push to enable the processing on a variety of resources
  - Ability to burst to high capacity becomes even more important when access can trigger processing

Looking Forward

Medium Term

- We need to change to have looser coupling between processing and storage, so we open up new lower cost options for processing while keeping control of the data
- We need to move to the adoption of external resources (commercial, academic, HPC, etc.) to open more options and increase our ability to burst to larger peaks

Longer term

- Move to deploy data reduction centers
  - Will increase the efficiency of analysis access and could simplify some aspects of data management
- Investigate data transformations to replace some parts of traditional reprocessing
  - Dynamically trigger processing subsets of events

Outlook

The amount of computing we will need within 10 years will be challenging to meet

- We need to expand the pool of possible resources and to economize how we provide computing in general
- We need to continue our efforts decouple processing and storage
- Increase how quickly we can make data accessible

We have a big gap that will not be closed by technology alone

- We have to rethink how we execute resource intensive workflows like analysis and data processing
  - Efficiency gains from organized data reduction
  - Rethinking how we think about our processing operations on data