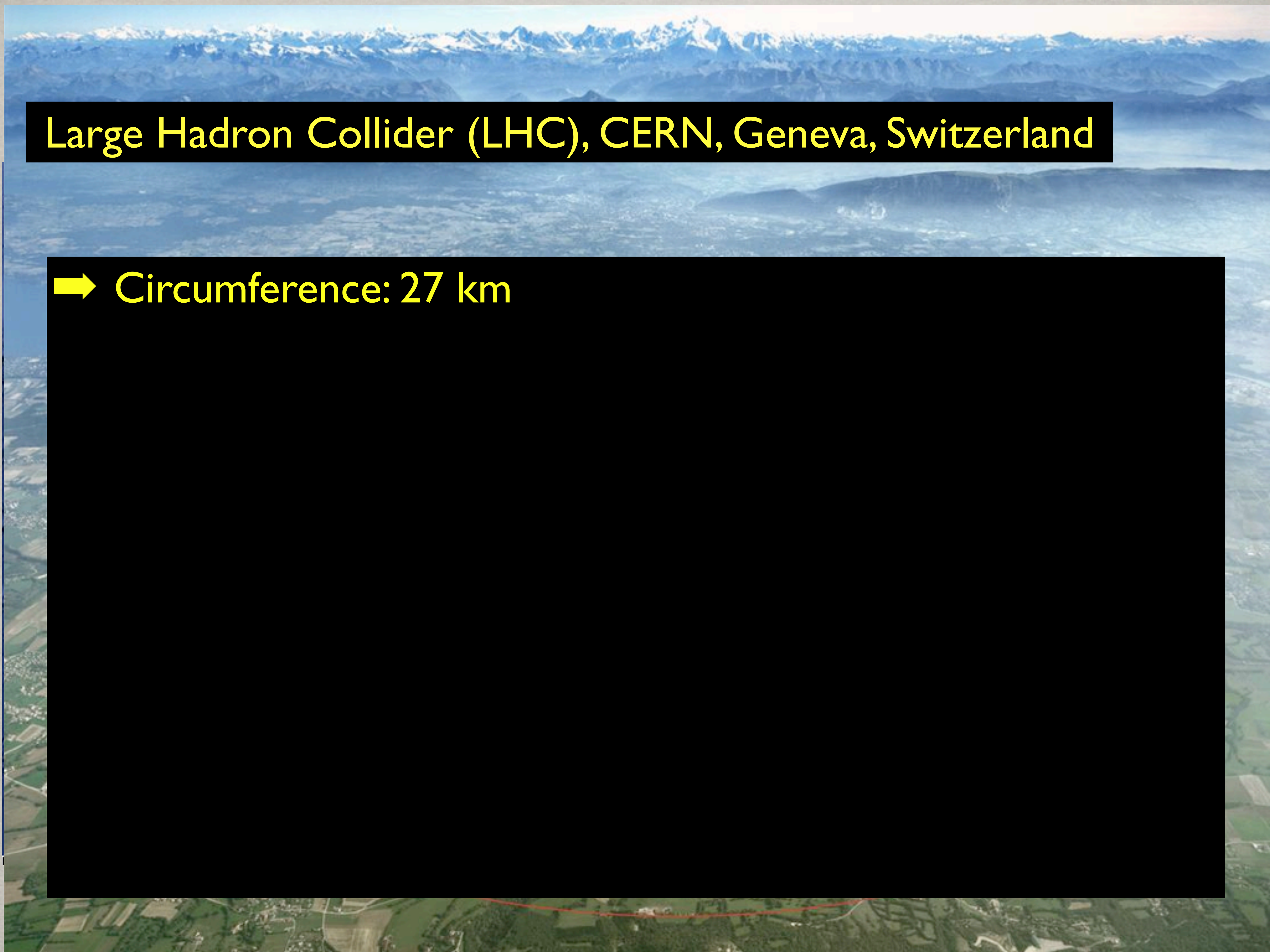
Data Management Challenges at CMS

Mike Hildreth
Kevin Lannon

Large Hadron Collider (LHC), CERN, Geneva, Switzerland

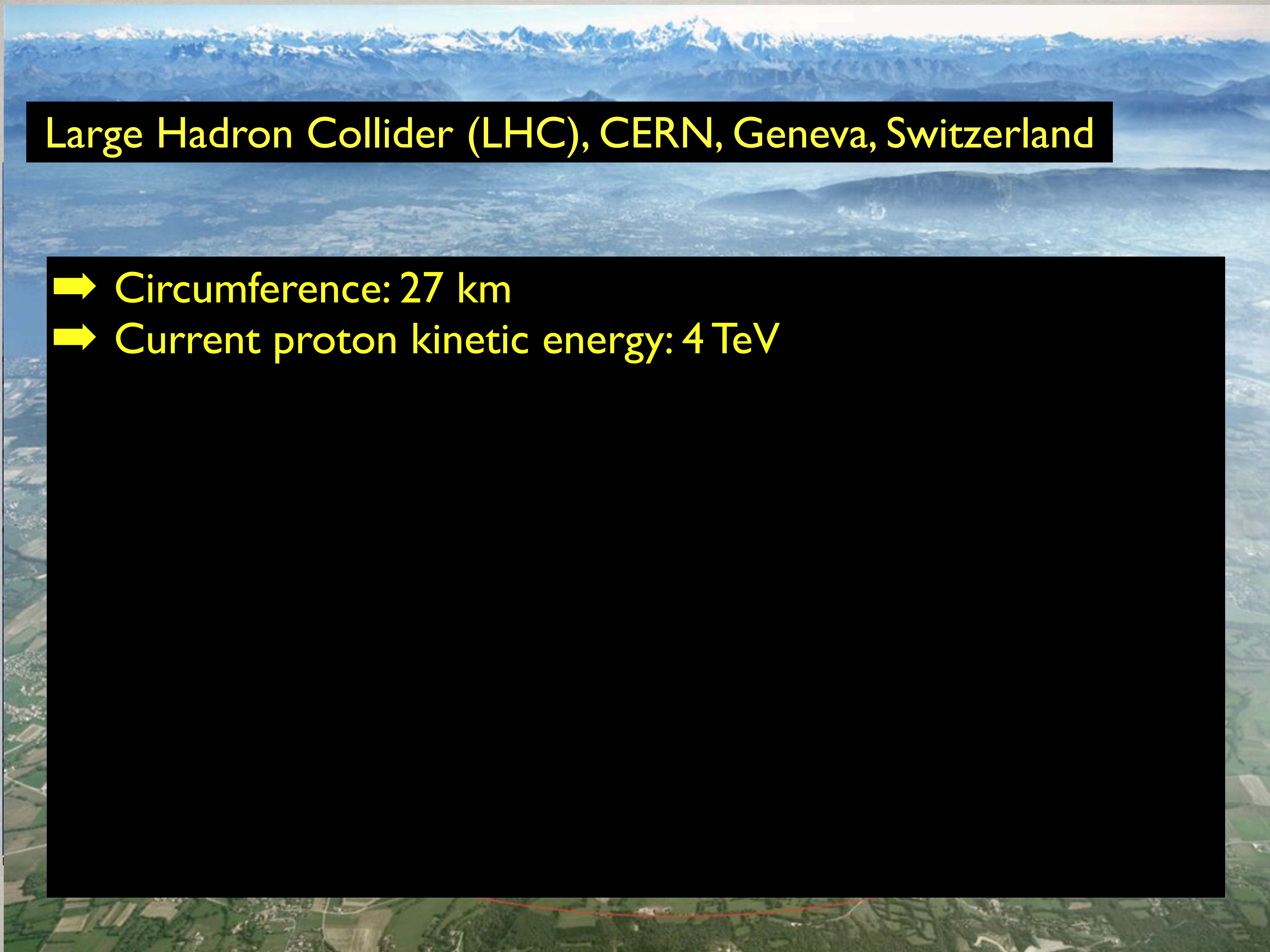
Large Hadron Collider (LHC), CERN, Geneva, Switzerland



- Circumference: 27 km

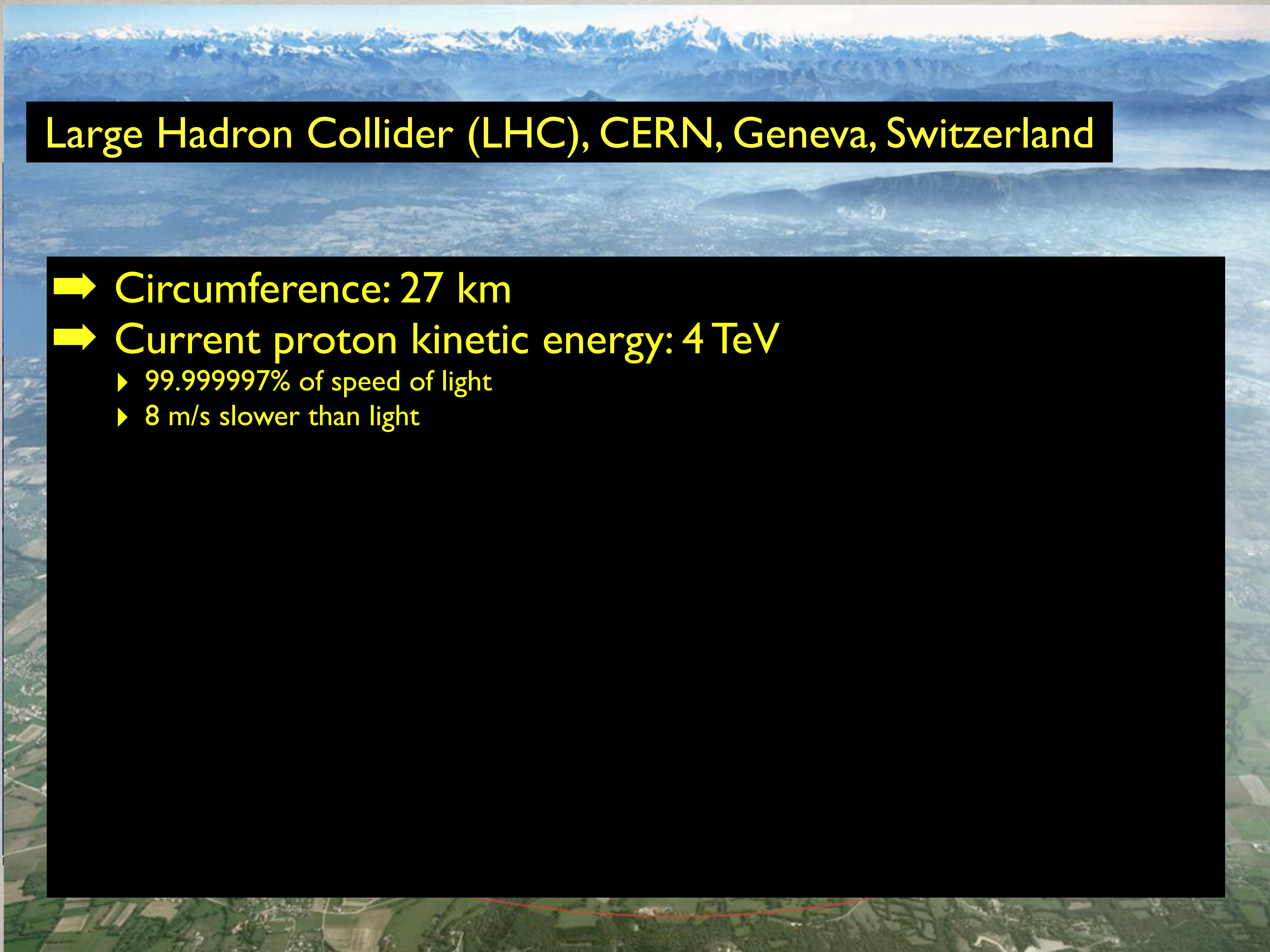
Large Hadron Collider (LHC), CERN, Geneva, Switzerland

- ➡ Circumference: 27 km
- ➡ Current proton kinetic energy: 4 TeV



Large Hadron Collider (LHC), CERN, Geneva, Switzerland

- Circumference: 27 km
- Current proton kinetic energy: 4 TeV
  - 99.999997% of speed of light



Large Hadron Collider (LHC), CERN, Geneva, Switzerland

- ➡ Circumference: 27 km
- ➡ Current proton kinetic energy: 4 TeV
  - 99.999997% of speed of light
  - 8 m/s slower than light

Large Hadron Collider (LHC), CERN, Geneva, Switzerland

- Circumference: 27 km
- Current proton kinetic energy: 4 TeV
  - 99.999997% of speed of light
  - 8 m/s slower than light
- Current total energy in beam: 135 MJ

Large Hadron Collider (LHC), CERN, Geneva, Switzerland

- ➡ Circumference: 27 km
- ➡ Current proton kinetic energy: 4 TeV
  - 99.999997% of speed of light
  - 8 m/s slower than light
- ➡ Current total energy in beam: 135 MJ
  - Equivalent to an aircraft carrier moving at 3.8 MPH

Large Hadron Collider (LHC), CERN, Geneva, Switzerland

- Circumference: 27 km
- Current proton kinetic energy: 4 TeV
  - 99.999997% of speed of light
  - 8 m/s slower than light
- Current total energy in beam: 135 MJ
  - Equivalent to an aircraft carrier moving at 3.8 MPH
  - or a Subaru Impreza moving at 1045 kilometers/hour (nearly the speed of sound)

Large Hadron Collider (LHC), CERN, Geneva, Switzerland

- Circumference: 27 km
- Current proton kinetic energy: 4 TeV
  - 99.999997% of speed of light
  - 8 m/s slower than light
- Current total energy in beam: 135 MJ
  - Equivalent to an aircraft carrier moving at 3.8 MPH
  - or a Subaru Impreza moving at 1045 kilometers/hour (nearly the speed of sound)
  - Or calories in seven 8” Cold Stone Creamery “Cheesecakes Named Desire”

Large Hadron Collider (LHC), CERN, Geneva, Switzerland

- ➡ Circumference: 27 km
- ➡ Current proton kinetic energy: 4 TeV
  - 99.999997% of speed of light
  - 8 m/s slower than light
- ➡ Current total energy in beam: 135 MJ
  - Equivalent to an aircraft carrier moving at 3.8 MPH
  - or a Subaru Impreza moving at 1045 kilometers/hour (nearly the speed of sound)
  - Or calories in seven 8" Cold Stone Creamery "Cheesecakes Named Desire"
- ➡ Superconducting magnet temperature is 2 K (colder than outer space)

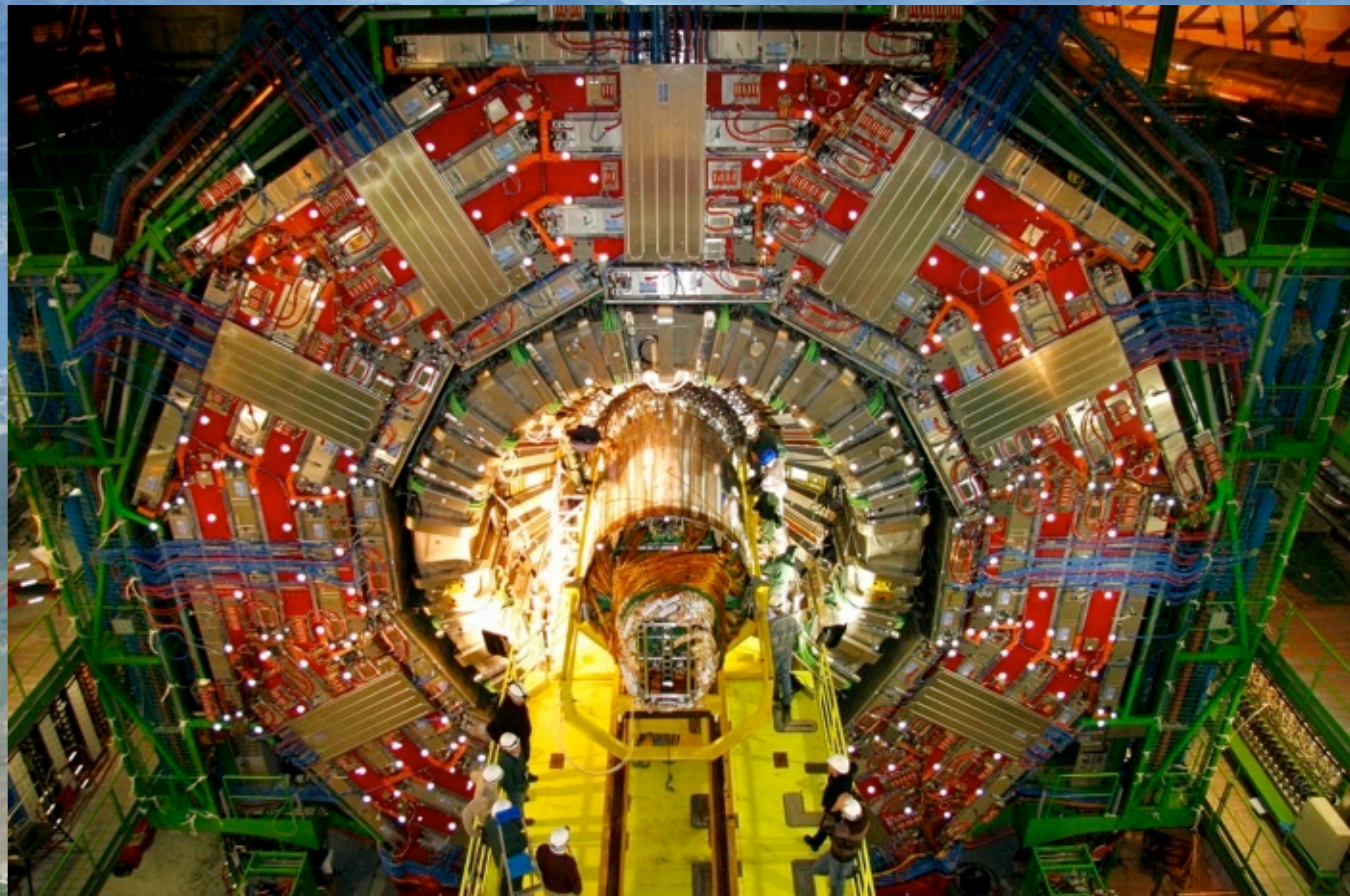
Large Hadron Collider (LHC), CERN, Geneva, Switzerland

- Circumference: 27 km
- Current proton kinetic energy: 4 TeV
  - 99.999997% of speed of light
  - 8 m/s slower than light
- Current total energy in beam: 135 MJ
  - Equivalent to an aircraft carrier moving at 3.8 MPH
  - or a Subaru Impreza moving at 1045 kilometers/hour (nearly the speed of sound)
  - Or calories in seven 8” Cold Stone Creamery “Cheesecakes Named Desire”
- Superconducting magnet temperature is 2 K (colder than outer space)
- Colliding protons like shooting two needles at each other from a distance of 6 miles and having them hit in the middle

Large Hadron Collider (LHC), CERN, Geneva, Switzerland

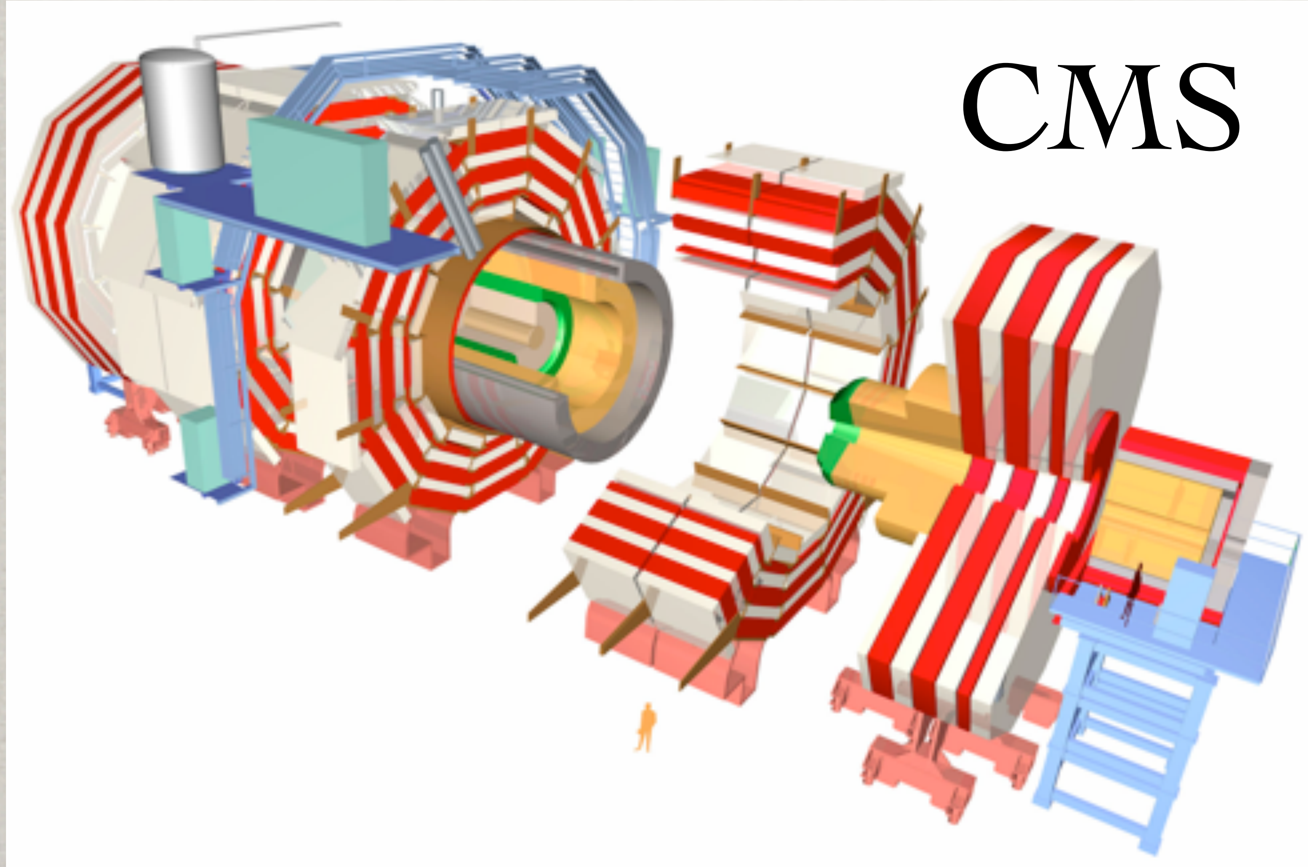
- ➡ Circumference: 27 km
- ➡ Current proton kinetic energy: 4 TeV
  - 99.999997% of speed of light
  - 8 m/s slower than light
- ➡ Current total energy in beam: 135 MJ
  - Equivalent to an aircraft carrier moving at 3.8 MPH
  - or a Subaru Impreza moving at 1045 kilometers/hour (nearly the speed of sound)
  - Or calories in seven 8" Cold Stone Creamery "Cheesecakes Named Desire"
- ➡ Superconducting magnet temperature is 2 K (colder than outer space)
- ➡ Colliding protons like shooting two needles at each other from a distance of 6 miles and having them hit in the middle
- ➡ Truly international effort: >10,000 scientists from over 100 countries


LHCb
LHCb
CMS
CMS
ATLAS
ATLAS
ALICE
ALICE



The CMS Experiment at the LHC (ND involvement)

The CMS Experiment



Total Weight: 12500 T
Diameter: 15 m (50 ft)
Length: 21.5 m (70 ft)

- Weighs the same as
  - 30 jumbo jets
  - 2500 African elephants
- Tracking detector
  - World's largest silicon detector: enough to cover a tennis court
  - 76 million readout channels
- Detector is 100 m underground
  - Constructed in pieces on surface, and lowered
  - Largest piece: ~2000 T
- Collaboration
  - Over 3000 scientists and engineers
  - 172 Universities and Labs
  - 41 countries

Physics Goals

- Study incredibly rare processes
  - Higgs boson, new types of matter?
- Need to isolate these from much more plentiful (but less interesting) processes
  - Processes of interest can occur once every ~ 1 billion collisions or more
  - collisions occur at 20 MHz
- Need to collect and analyze as many collisions as possible

How fast do we need to Go?

Basic facts:

➡ Data from detector: ~250 kB/ collision

➡ Processing time for analysis: 5 sec (basic)

		Proton Collisions in Detector	Level 1 Trigger	High Level Trigger
	Data Rate	20 MHz	60 kHz	300 Hz
For 1 year's worth of data	Data Collected	50 EB	200 PB	1-2 PB
For 1 year's worth of data	Processing time	45 Million CPU years!	170 Thousand CPU years	860 CPU years

How fast do we need to Go?

Basic facts:

- ➡ Data from detector: ~250 kB/ collision
- ➡ Processing time for analysis: 5 sec (basic)

	Proton Collisions in Detector	Level 1 Trigger	High Level Trigger
Data Rate	20 MHz	60 kHz	300 Hz
Data Collected	50 EB	200 PB	1-2 PB
Processing time	45 Million CPU years!	170 Thousand CPU years	860 CPU years

For 1 year's worth of data

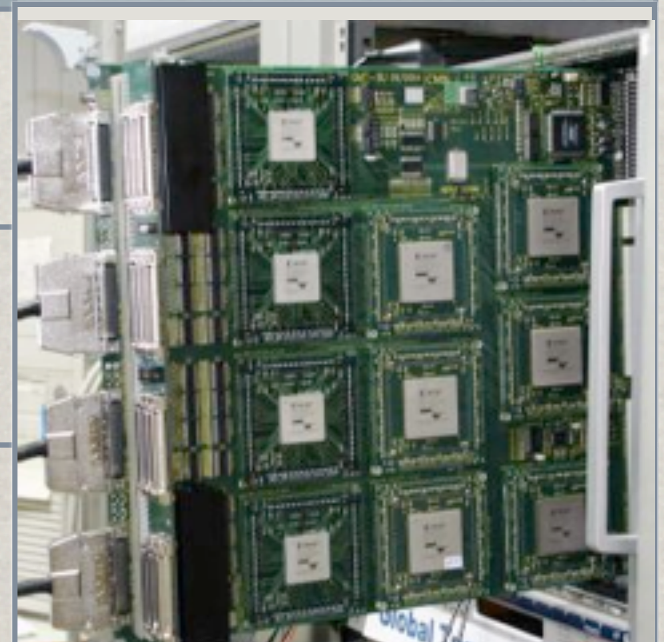
1 Million TB!

How fast do we need to Go?

Basic facts:

- ➡ Data from detector: ~250 kB/ collision
- ➡ Processing time for analysis: 5 sec (basic)

For 1 year's worth of data

	Proton Collisions in Detector	Level 1 Trigger
Data Rate	20 MHz	
Data Collected	50 EB	
Processing time	45 Million CPU years!	

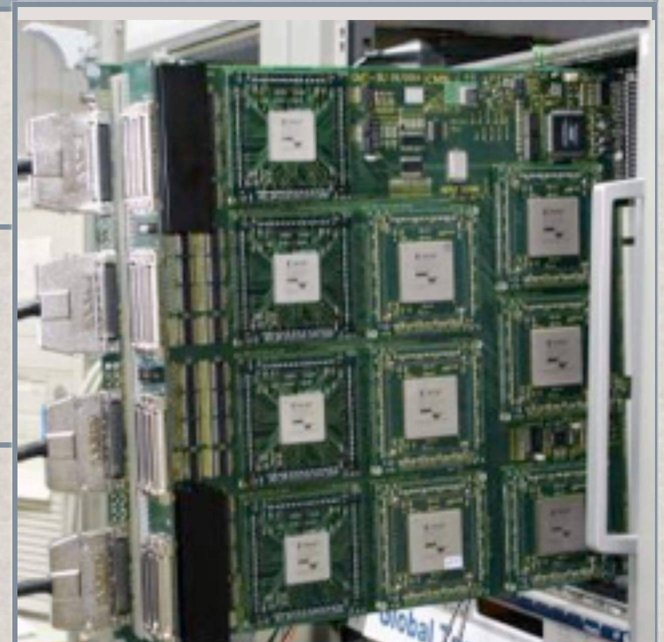
How fast do we need to Go?

Basic facts:

➡ Data from detector: ~250 kB/ collision

➡ Processing time for analysis: 5 sec (basic)

FPGA Chips do very simple analysis
~ μs to analyze data

	Proton Collisions in Detector	Level 1 Trigger
Data Rate	20 MHz	
Data Collected	50 EB	
Processing time	45 Million CPU years!	

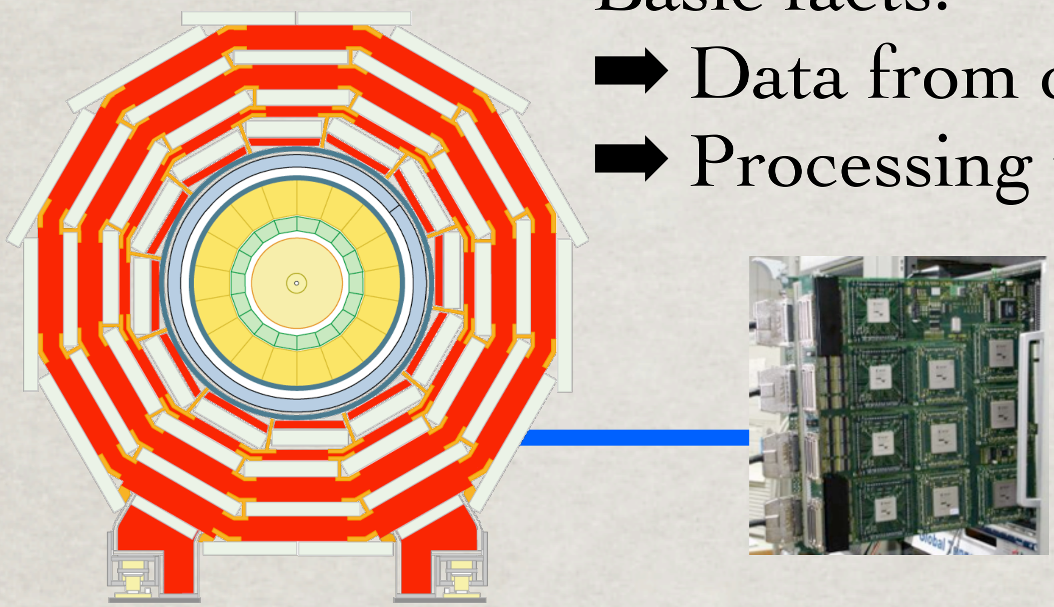
For 1 year's worth of data

UNIVERSITY OF NOTRE DAME

How fast do we need to Go?

Basic facts:

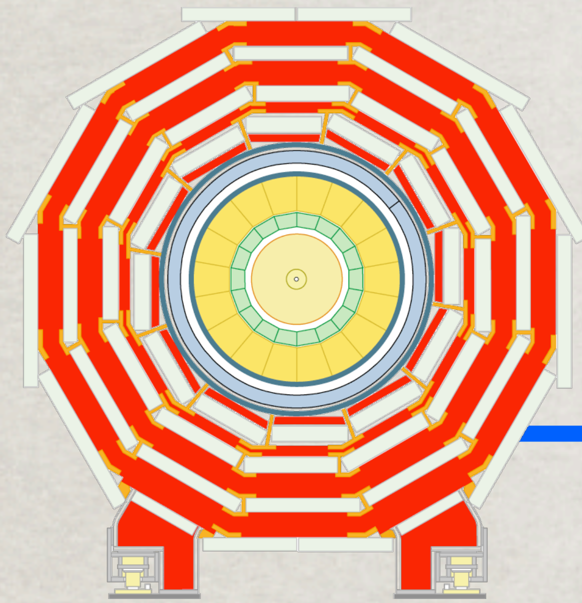
- ➡ Data from detector: ~250 kB/ collision
- ➡ Processing time for analysis: 5 sec (basic)



	Proton Collisions in Detector	Level 1 Trigger
Data Rate	20 MHz	60 kHz
Data Collected	50 EB	200 PB
Processing time	45 Million CPU years!	170 Thousand CPU years

For 1 year's worth of data

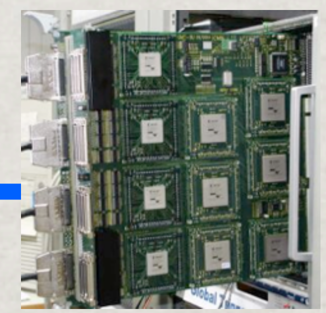
How fast do we need to Go?



Basic facts:

➡ Data from detector: ~250 kB/ collision

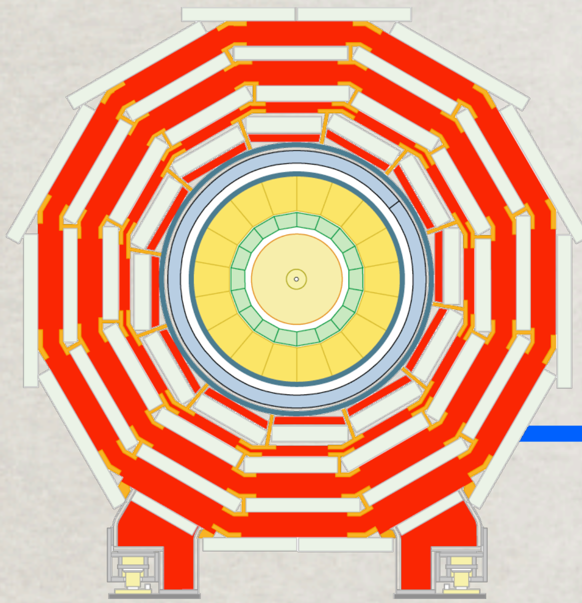
➡ Processing time for analysis: 5 sec (basic)

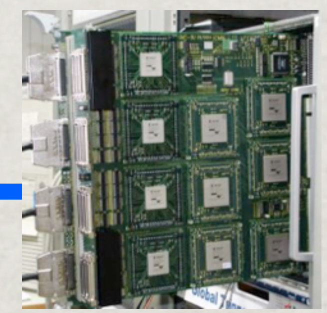


		Proton Collisions in Detector	Level 1 Trigger	High Level Trigger
	Data Rate	20 MHz	60 kHz	Computer farm with ~ 1000 CPUs High speed network/ switch
For 1 year's worth of data	Data Collected	50 EB	200 PB	
	Processing time	45 Million CPU years!	170 Thousand CPU years	

UNIVERSITY OF NOTRE DAME

How fast do we need to Go?





Basic facts:

➡ Data from detector: ~250 kB/ collision

➡ Processing time for analysis: 5 sec (basic)

Simplified analysis code
~ ms to analyze data

For 1 year's worth of data

	Proton Collisions in Detector	Level 1 Trigger	High Level Trigger
Data Rate	20 MHz	60 kHz	Computer farm with ~ 1000 CPUs High speed network/ switch 
Data Collected	50 EB	200 PB	
Processing time	45 Million CPU years!	170 Thousand CPU years	

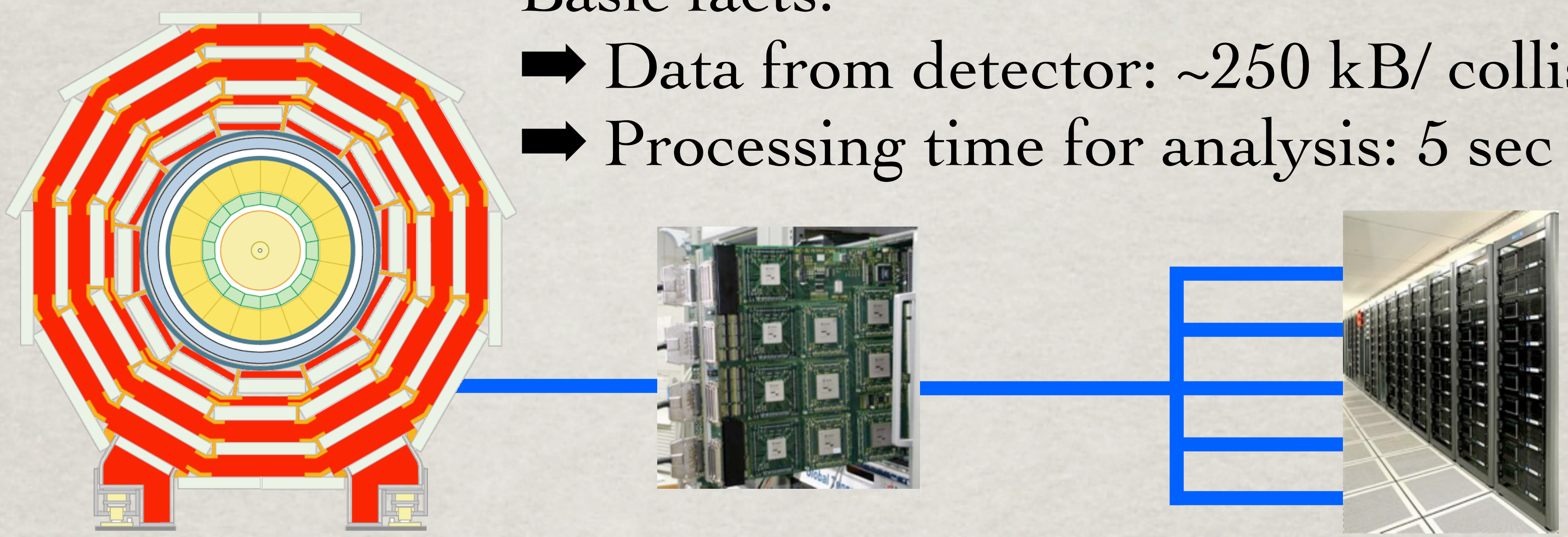
UNIVERSITY OF NOTRE DAME

How fast do we need to Go?

Basic facts:

➡ Data from detector: ~250 kB/ collision

➡ Processing time for analysis: 5 sec (basic)



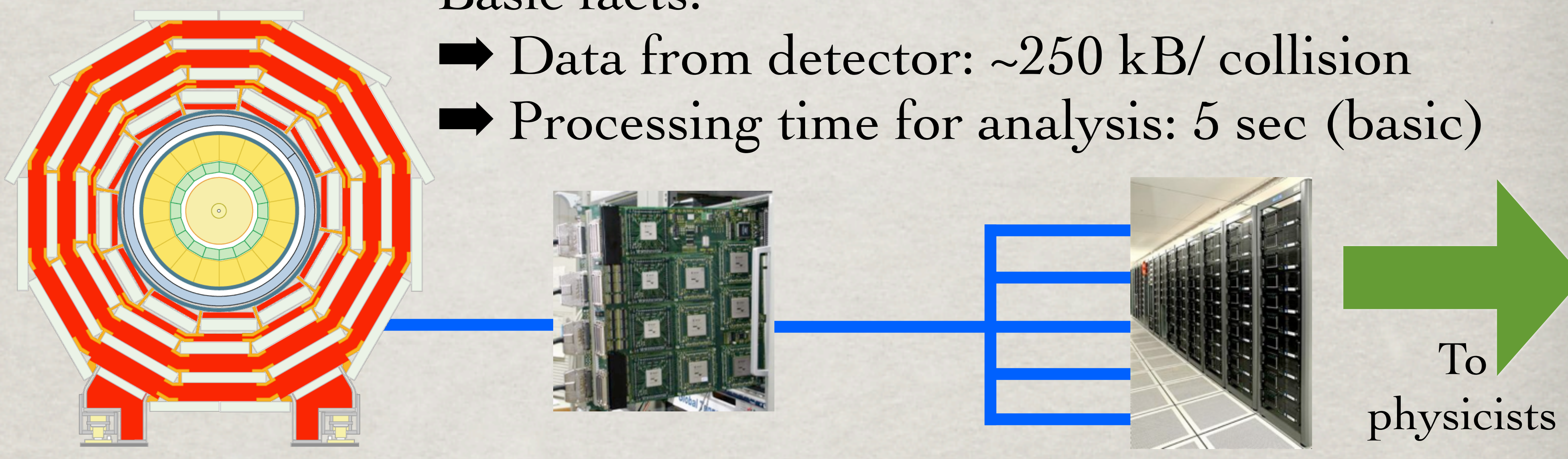
For 1 year's worth of data

	Proton Collisions in Detector	Level 1 Trigger	High Level Trigger
Data Rate	20 MHz	60 kHz	300 Hz
Data Collected	50 EB	200 PB	1-2 PB
Processing time	45 Million CPU years!	170 Thousand CPU years	860 CPU years

How fast do we need to Go?

Basic facts:

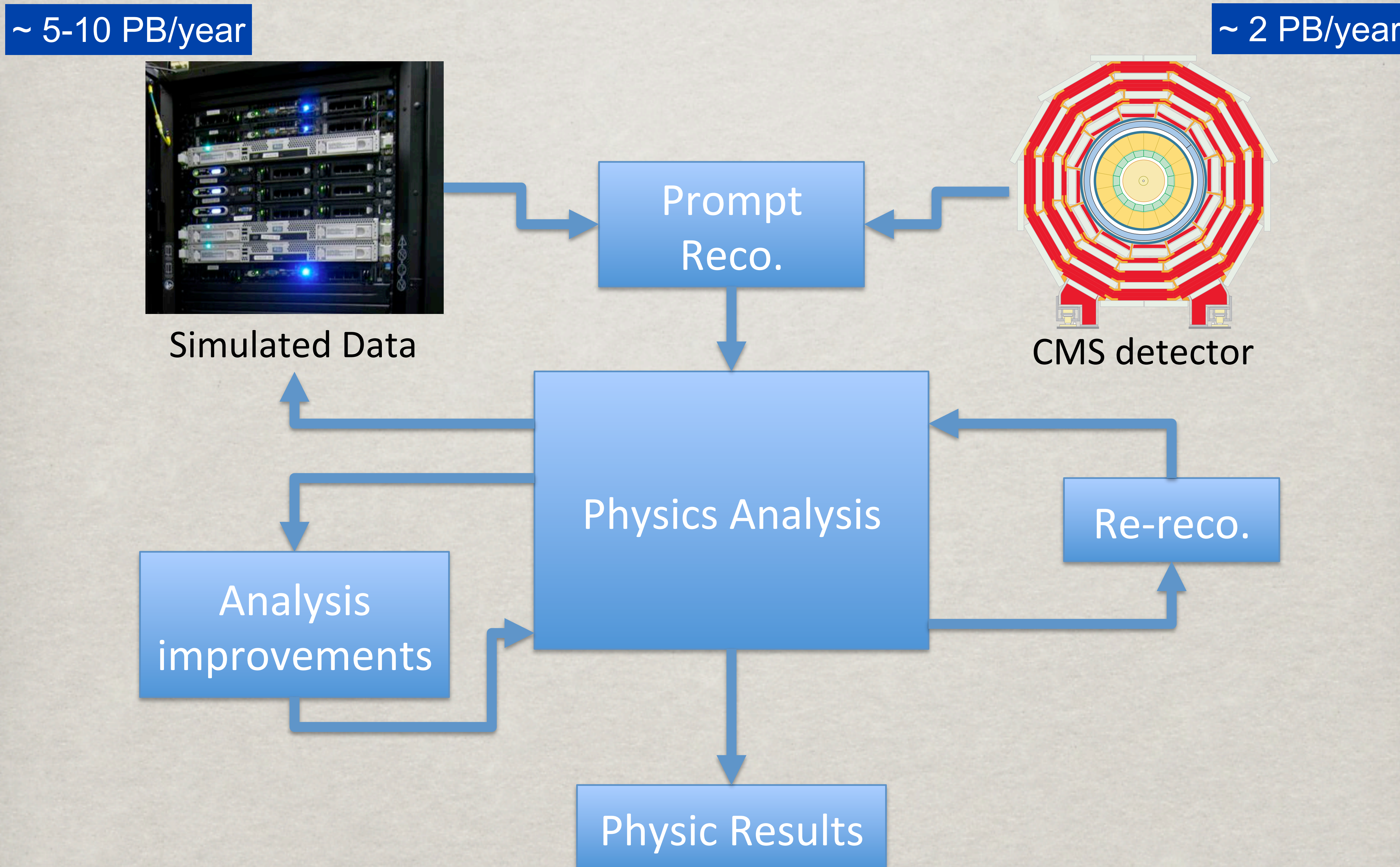
- ➡ Data from detector: ~250 kB/ collision
- ➡ Processing time for analysis: 5 sec (basic)



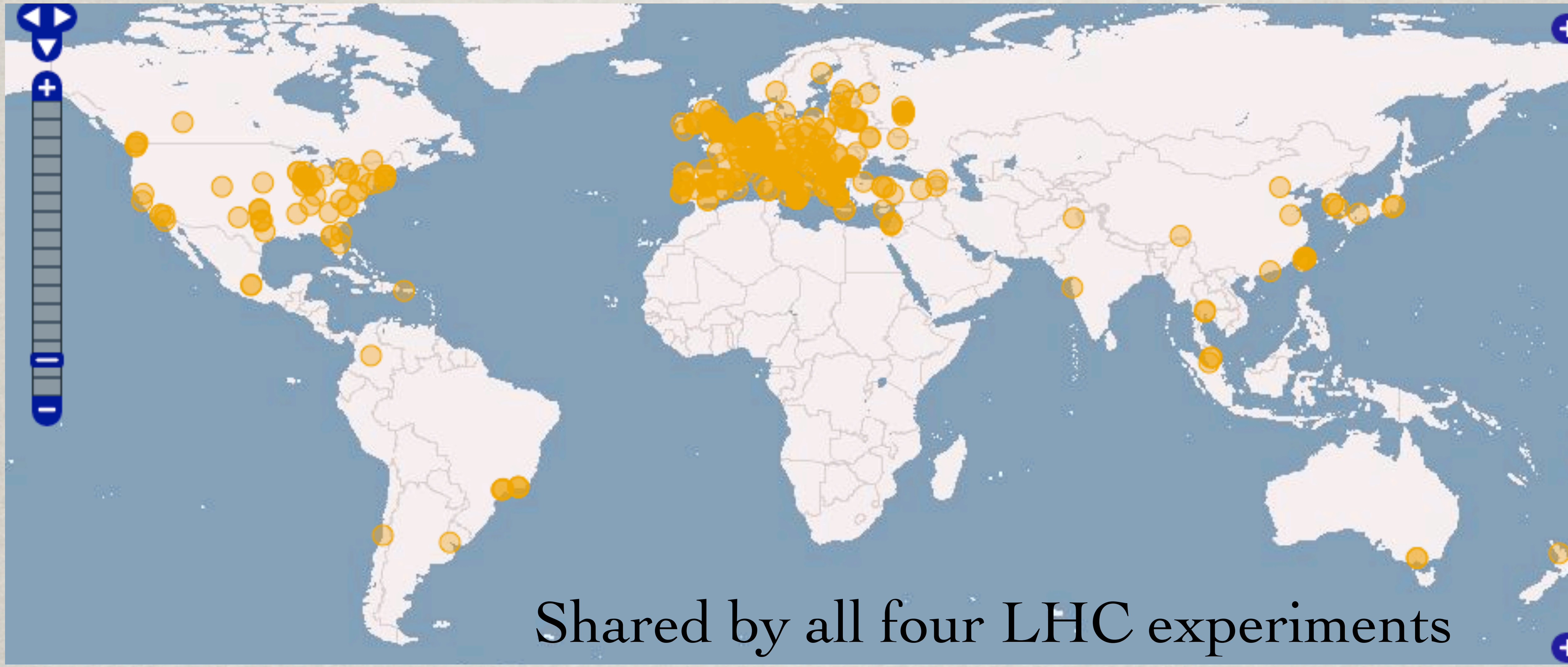
	Proton Collisions in Detector	Level 1 Trigger	High Level Trigger
Data Rate	20 MHz	60 kHz	300 Hz
Data Collected	50 EB	200 PB	1-2 PB
Processing time	45 Million CPU years!	170 Thousand CPU years	860 CPU years

For 1 year's worth of data

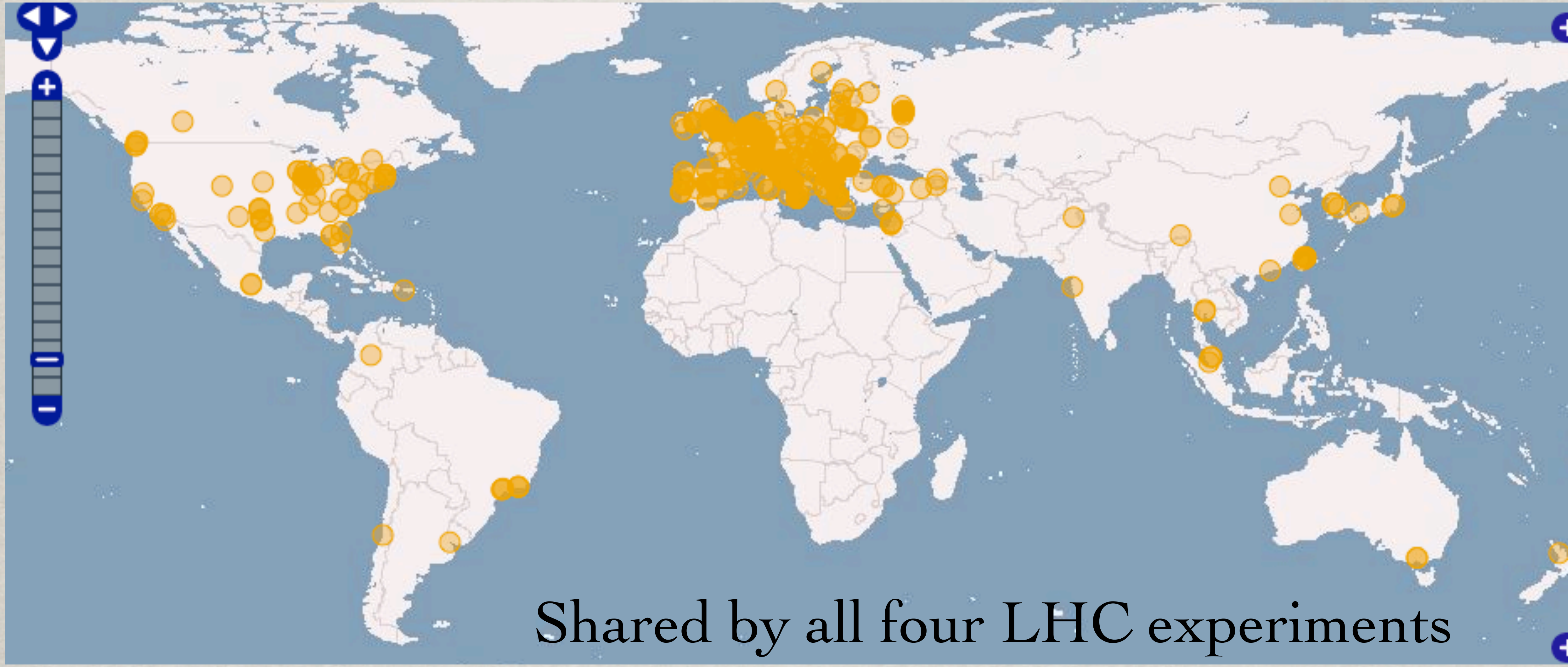
Processing CMS Data



Worldwide LHC Computing Grid

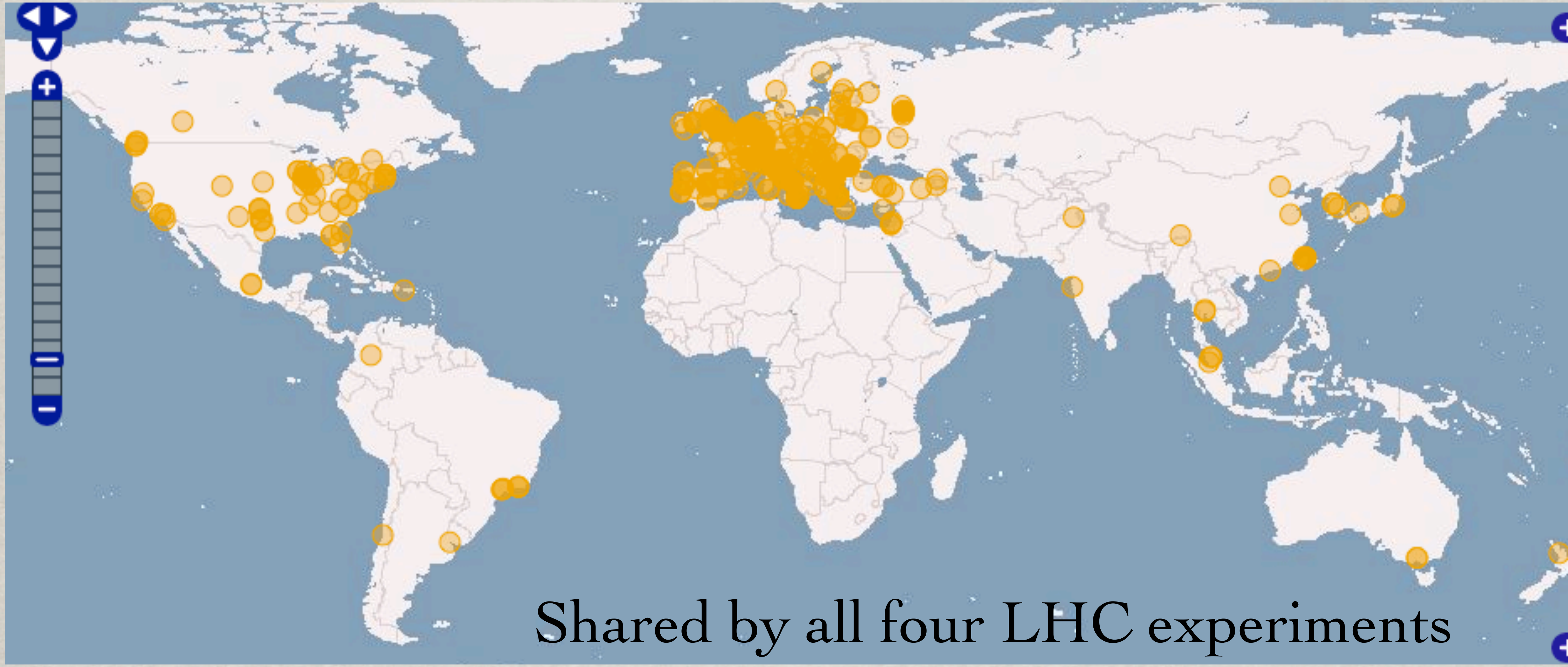


WORLDWIDE LHC COMPUTING GRID



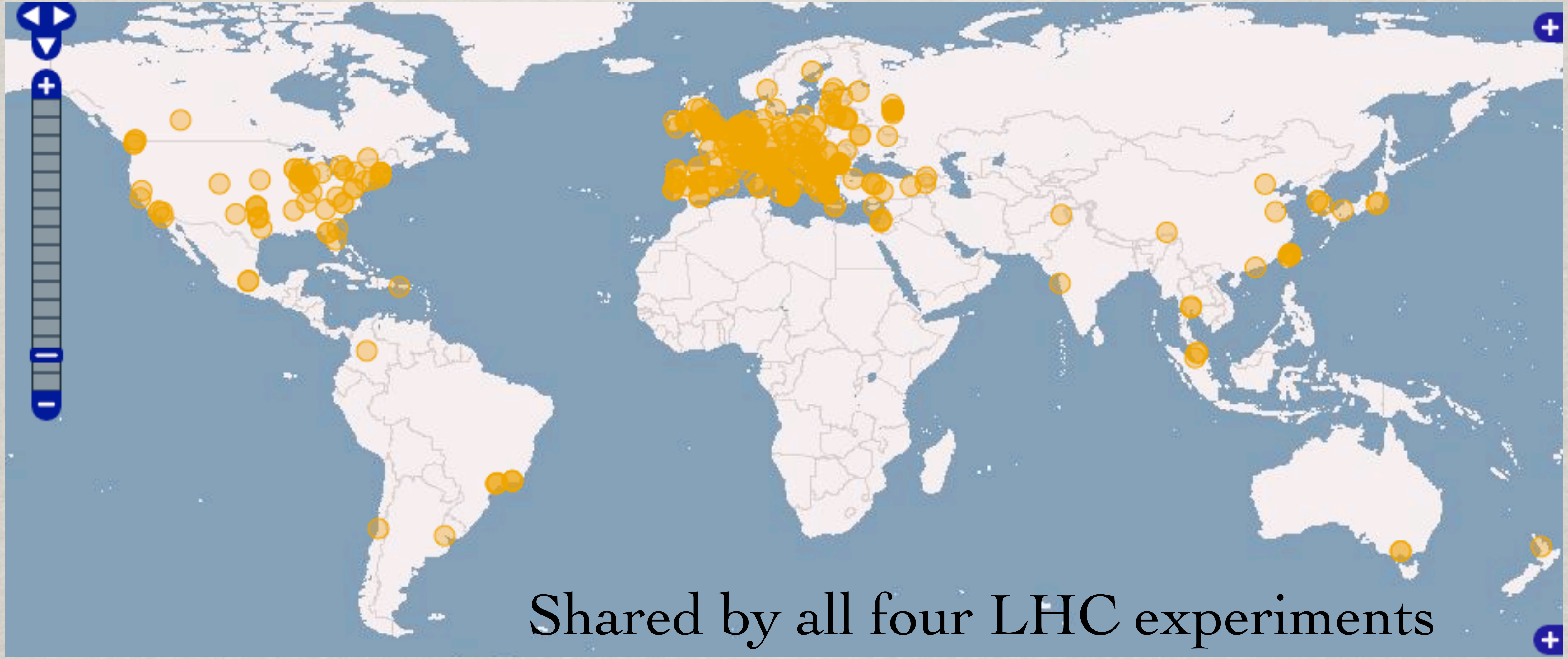
- Over 160 sites around world (including OSG sites in US)

Worldwide LHC Computing Grid



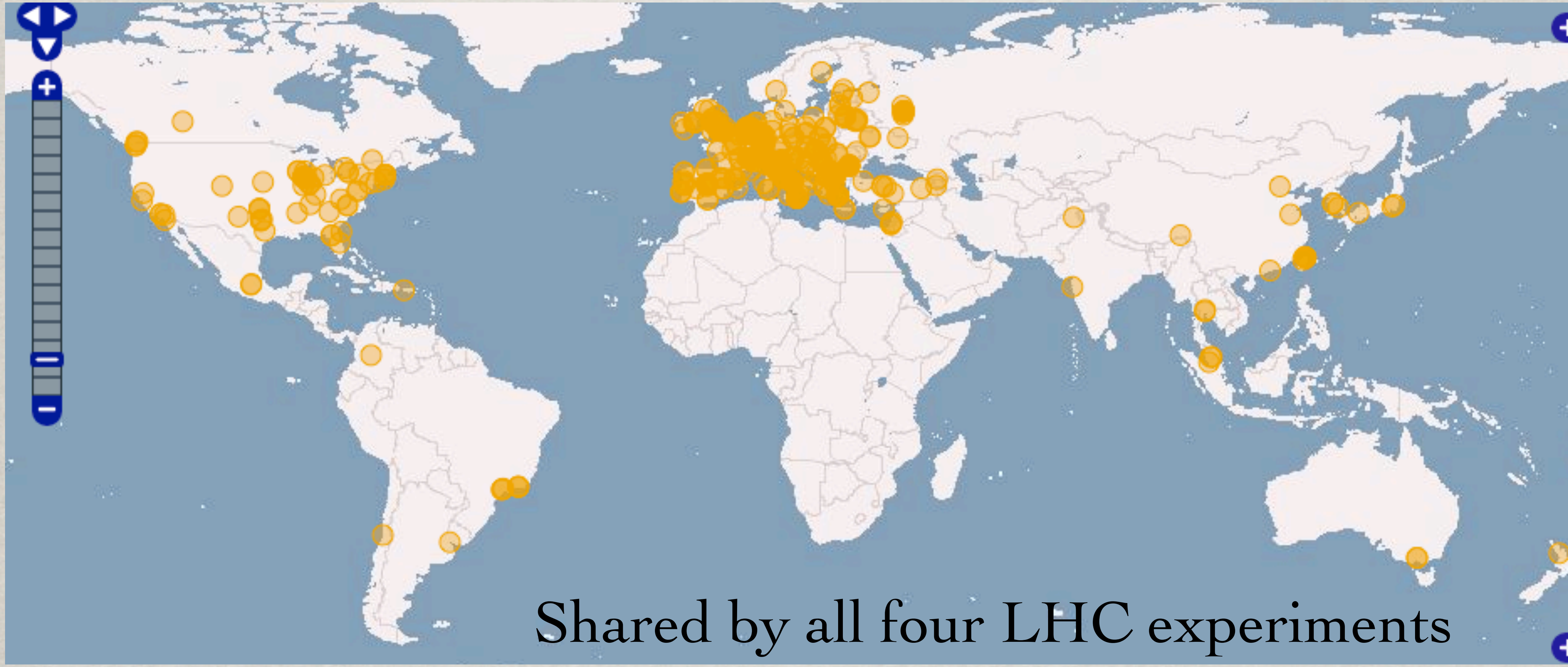
- Over 160 sites around world (including OSG sites in US)
- > 200k CPU cores available

Worldwide LHC Computing Grid



- Over 160 sites around world (including OSG sites in US)
- > 200k CPU cores available
- As many as 1 million jobs submitted in a single day

Worldwide LHC Computing Grid



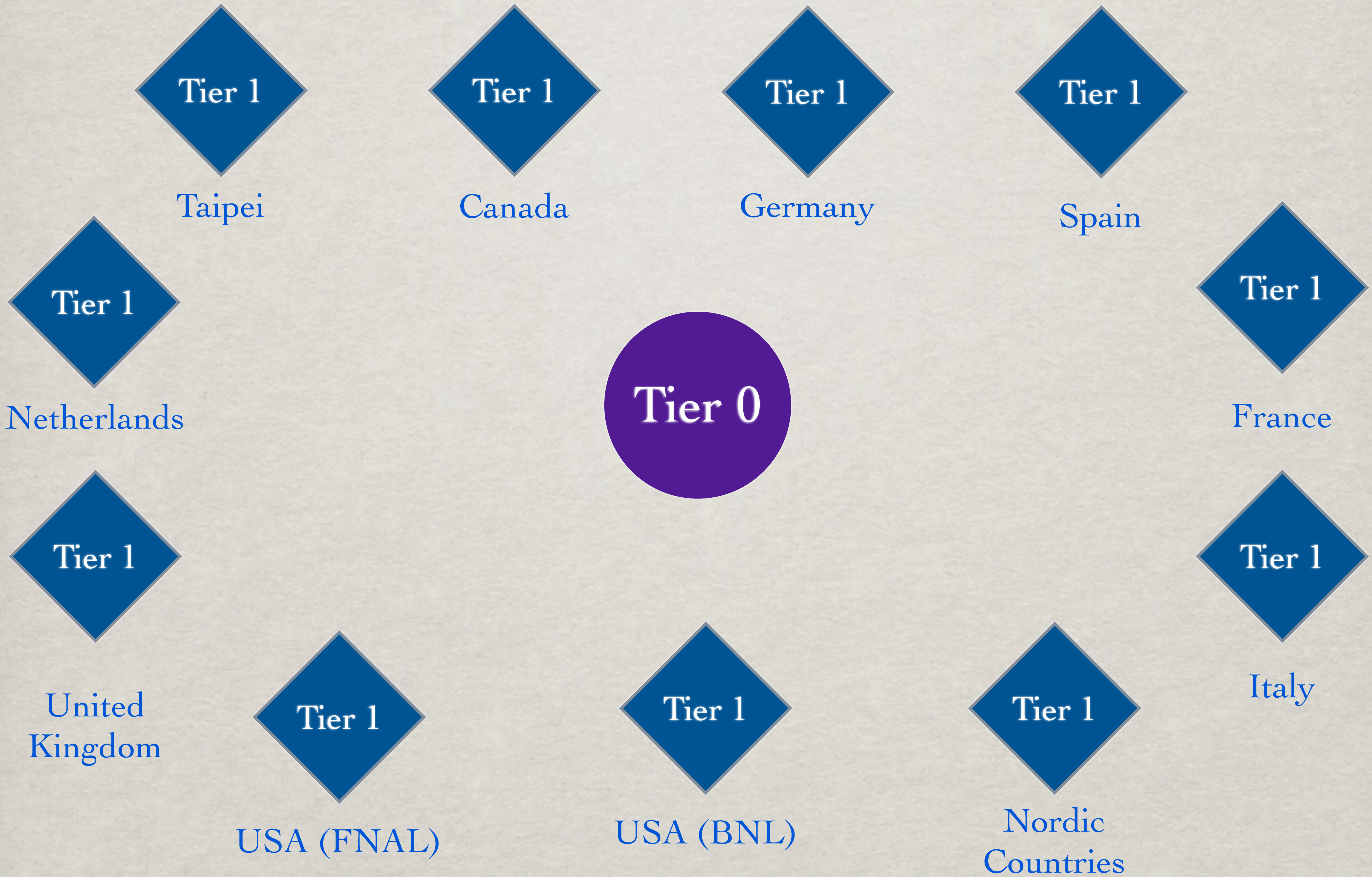
- Over 160 sites around world (including OSG sites in US)
- > 200k CPU cores available
- As many as 1 million jobs submitted in a single day
- > 300 PB of total storage available

Data Flow Organization

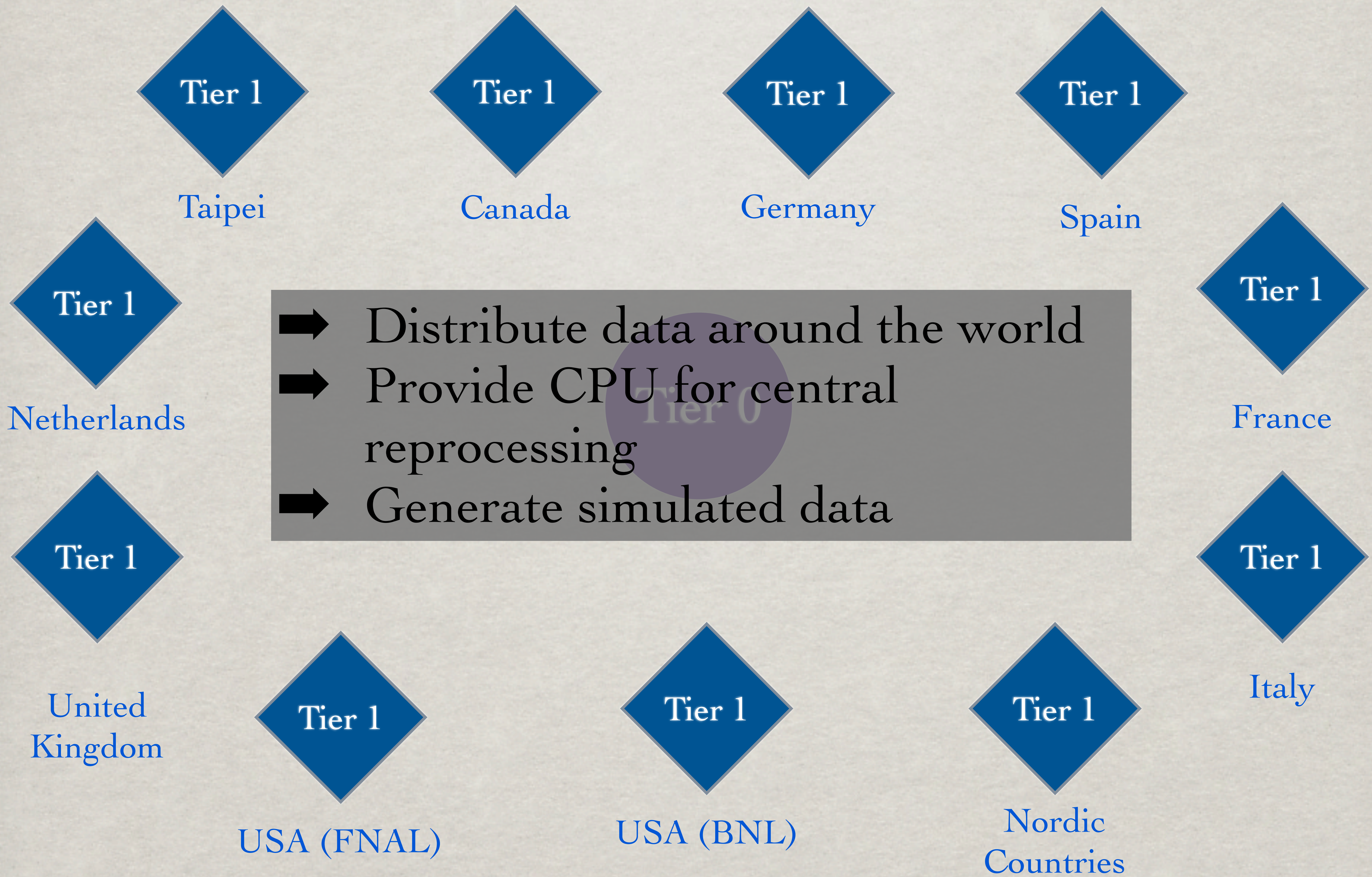
Data Flow Organization

- All LHC data passes through T0 for initial processing
- Provides less than 20% of total CPU resources for LHC experiments
- Basic data processing common to all analyses

Data Flow Organization

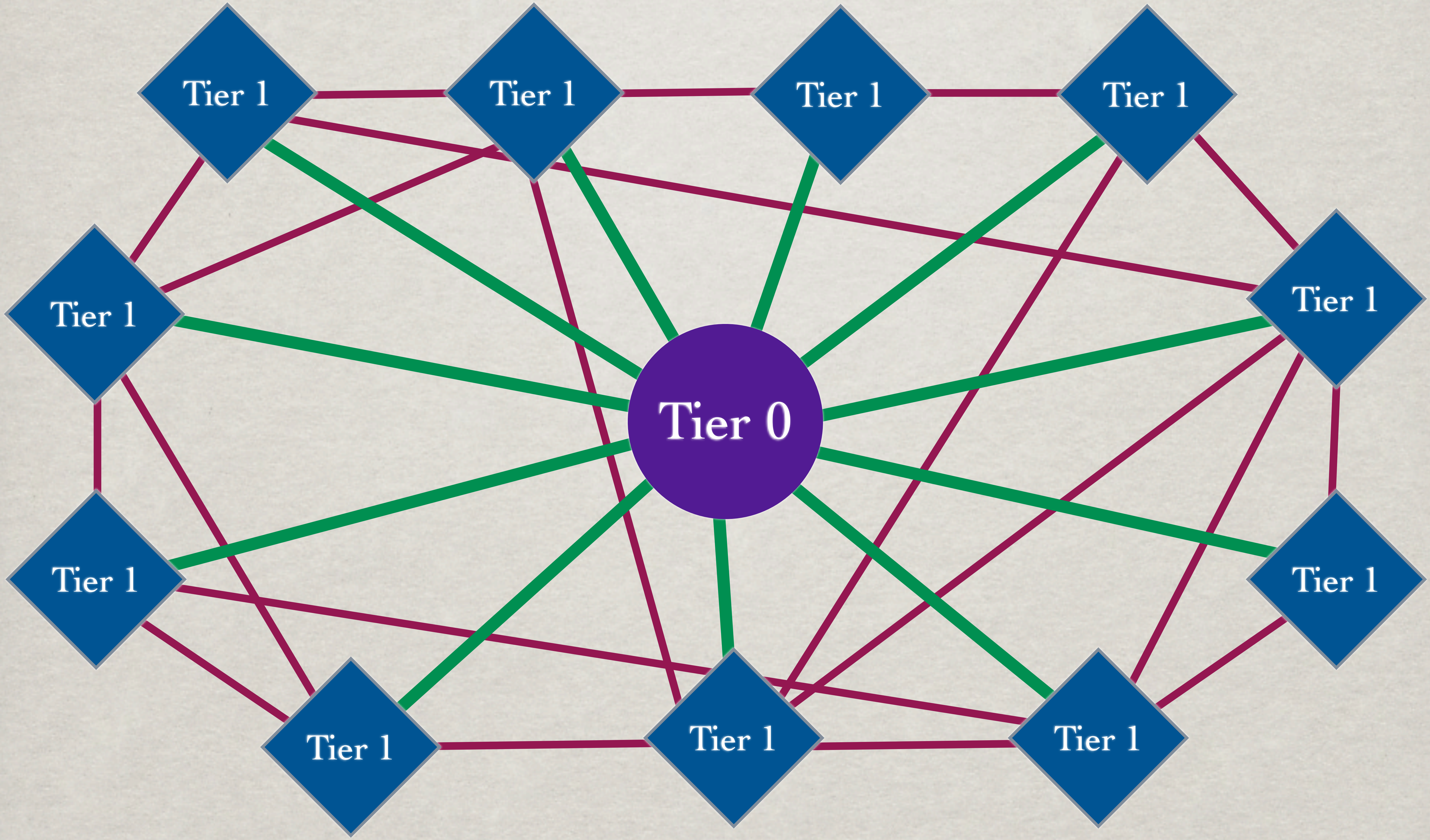


Data Flow Organization

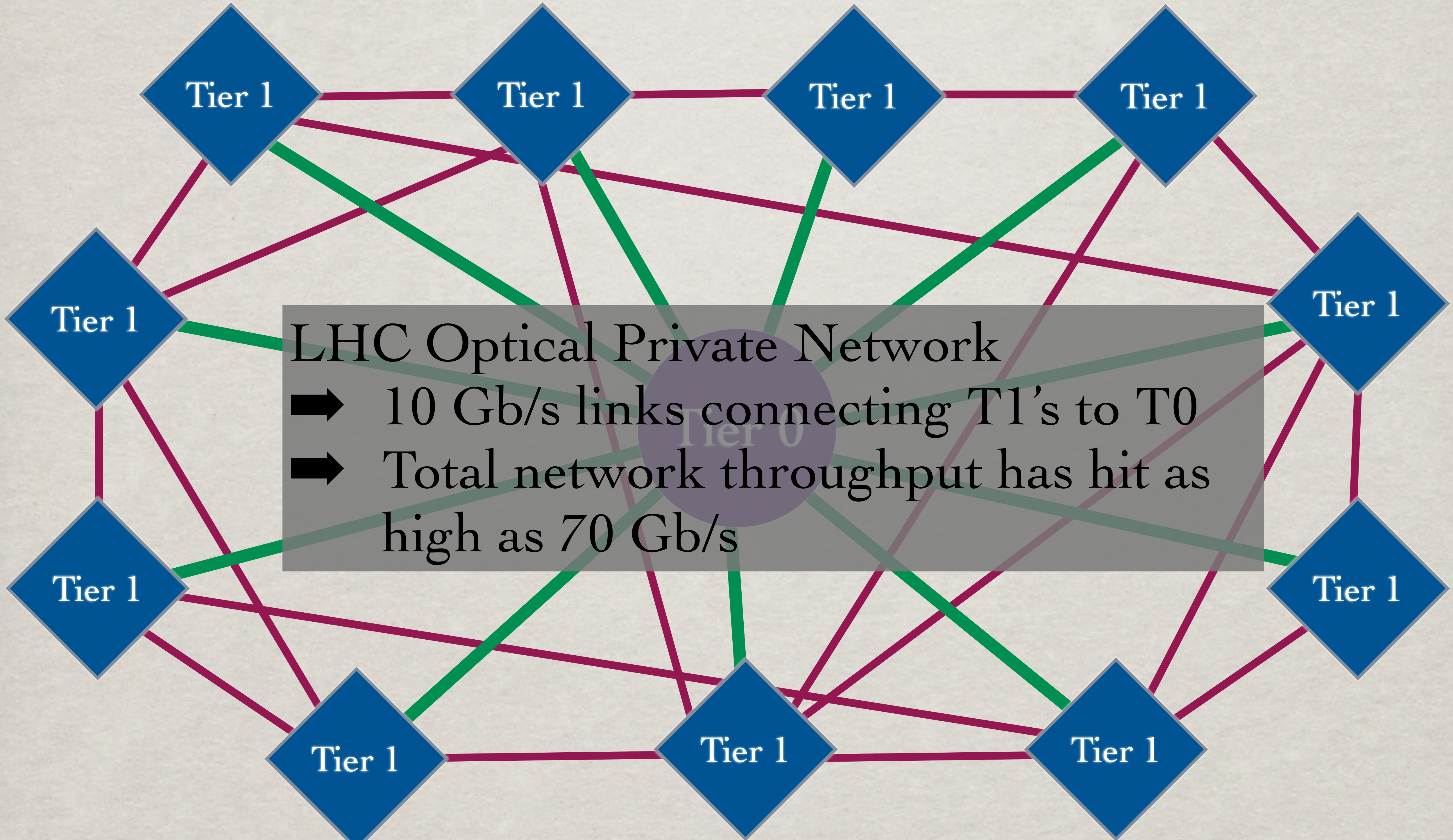


- Distribute data around the world
- Provide CPU for central reprocessing
- Generate simulated data

Data Flow Organization



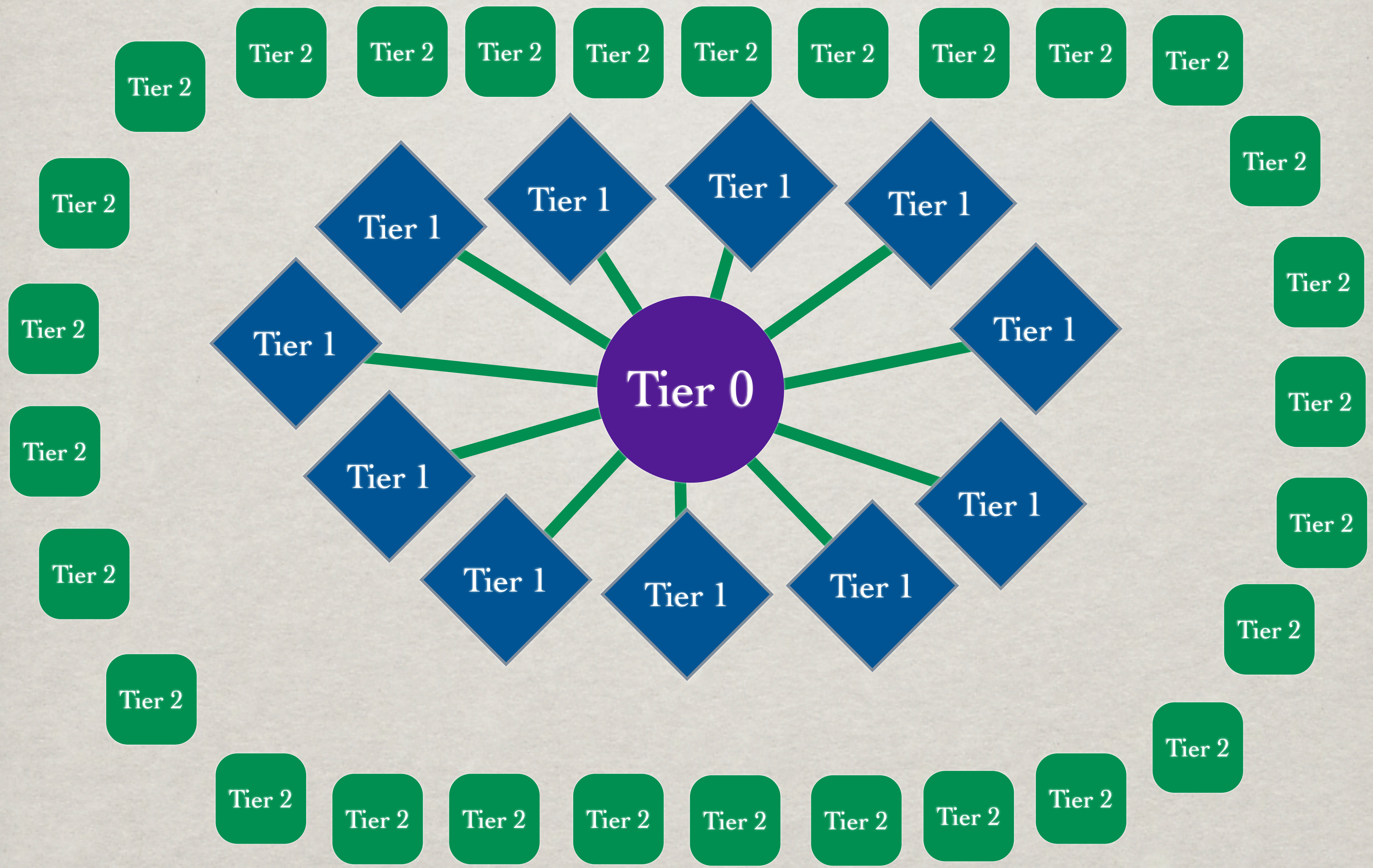
Data Flow Organization



LHC Optical Private Network

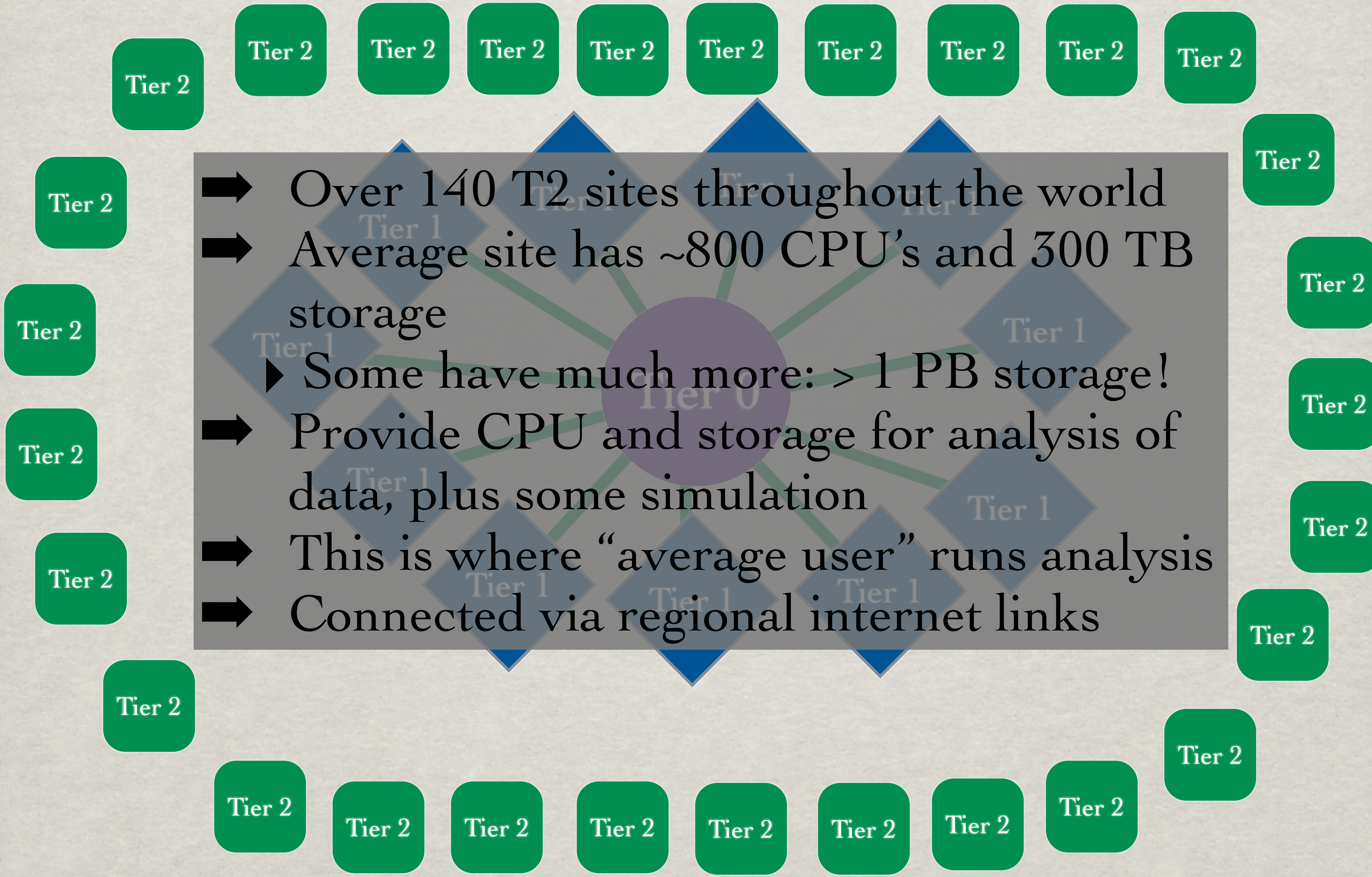
- ➡ 10 Gb/s links connecting T1's to T0
- ➡ Total network throughput has hit as high as 70 Gb/s

Data Flow Organization

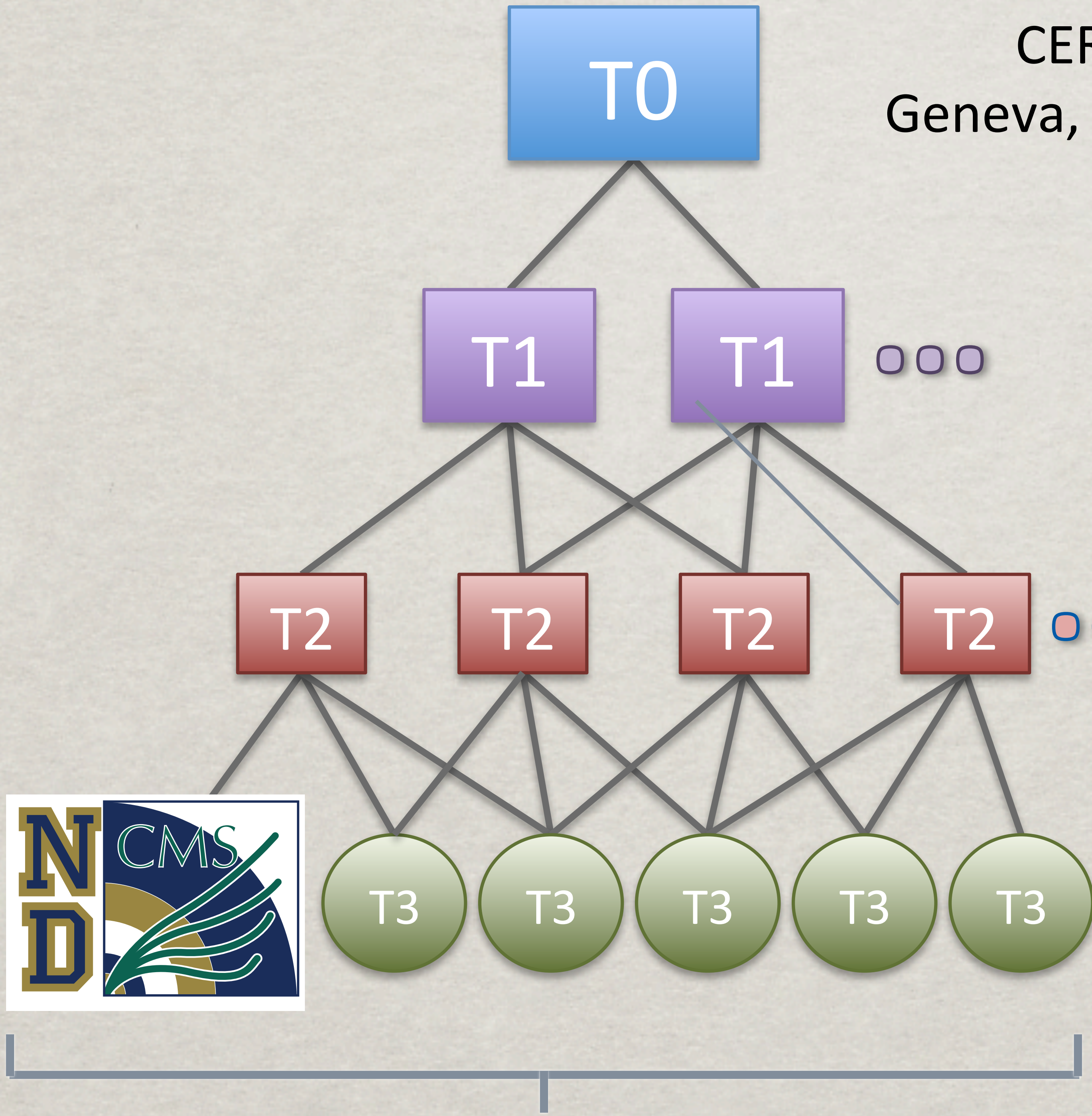


Data Flow Organization

- ➡ Over 140 T2 sites throughout the world
- ➡ Average site has ~800 CPU's and 300 TB storage
  - Some have much more: > 1 PB storage!
- ➡ Provide CPU and storage for analysis of data, plus some simulation
- ➡ This is where "average user" runs analysis
- ➡ Connected via regional internet links



Computing Tier Summary



CERN Lab
Geneva, Switzerland

National Labs
(Fermilab, etc.)

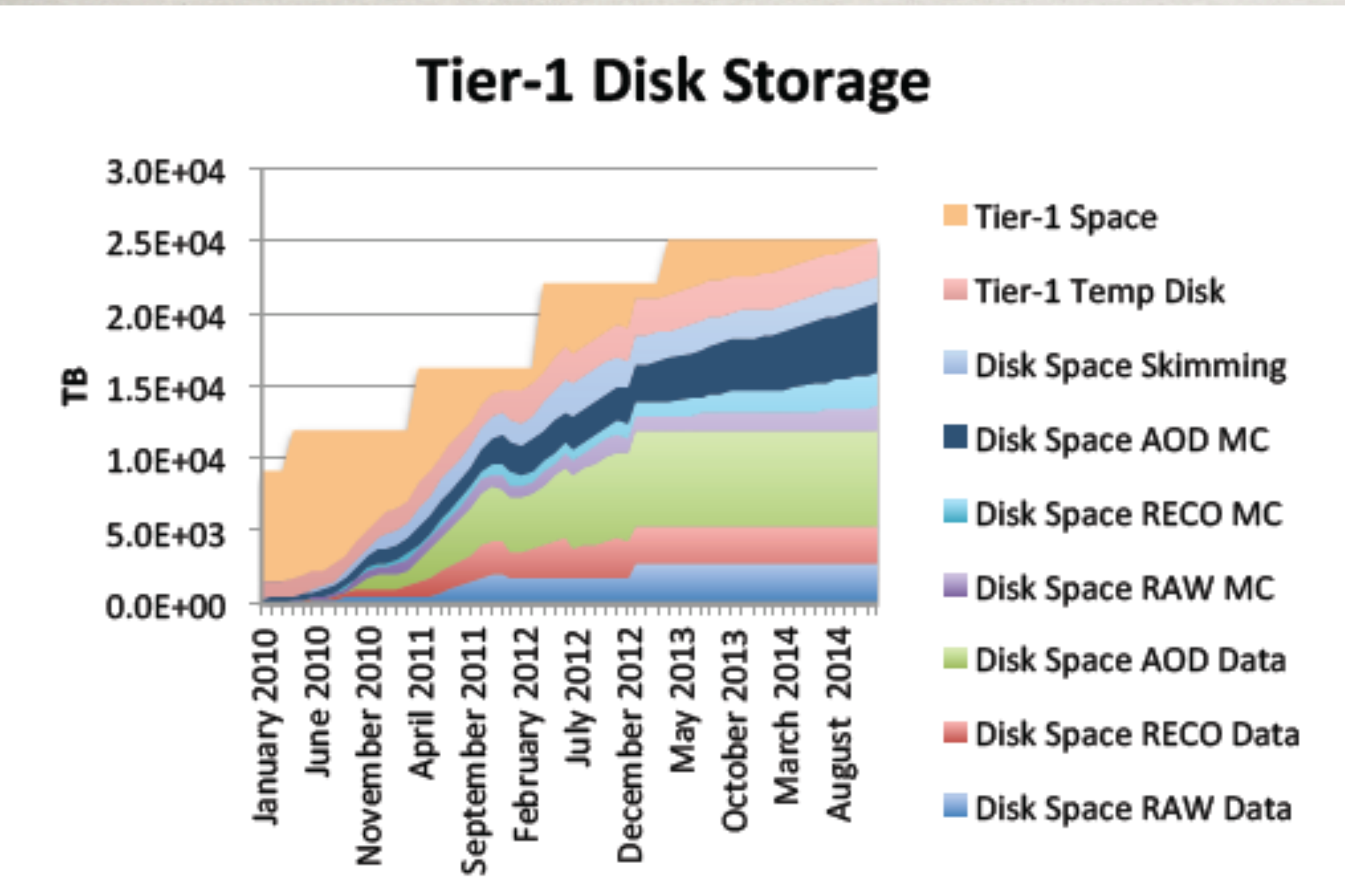
Universities
(MIT, Wisconsin,
Nebraska, Purdue, etc.)

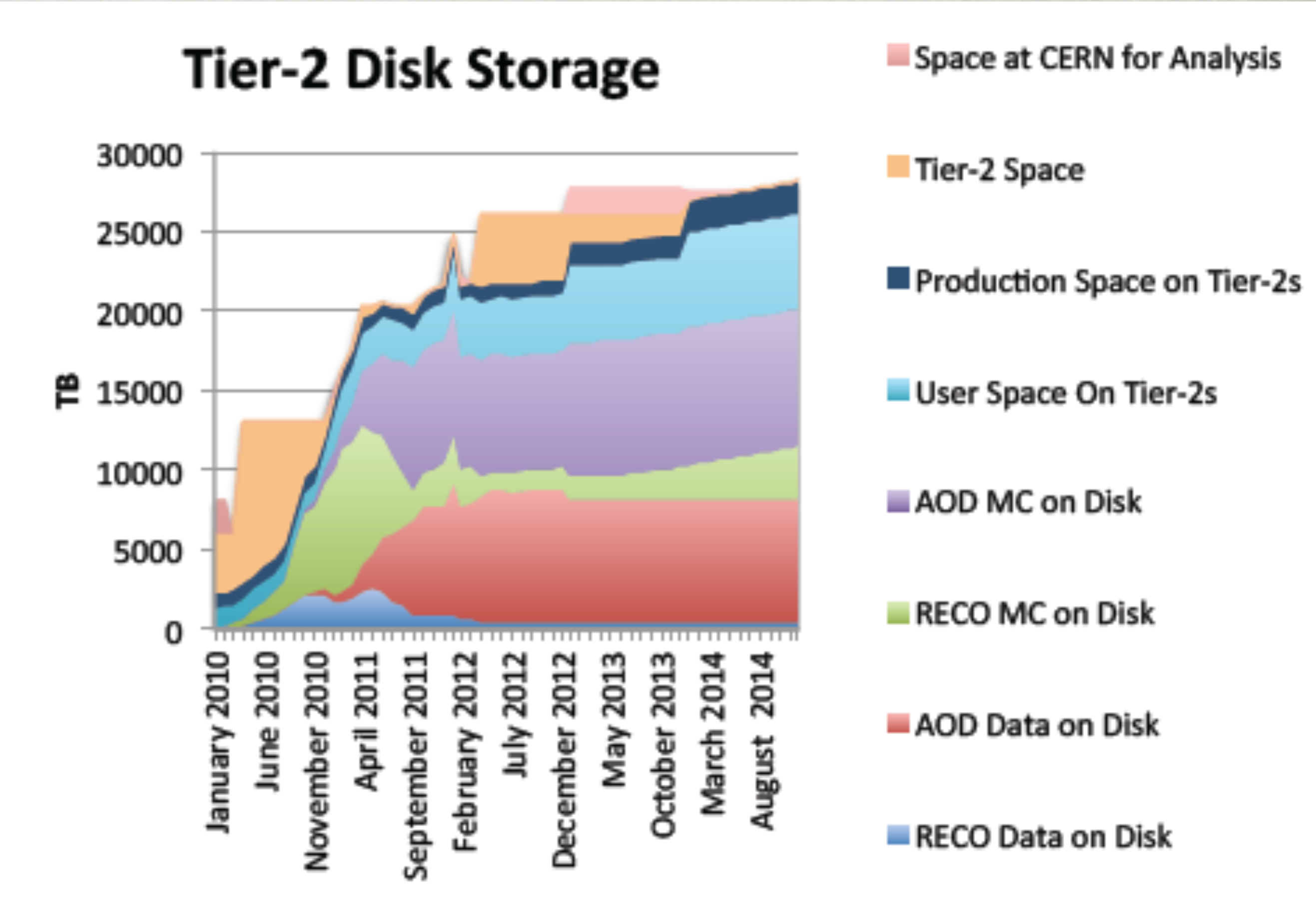
Universities
(ND, Colorado,
UMD, OSU, etc.)

Provide "Local" computing for individual groups

Detailed Planning Required

- Models of data usage needed to decide where to put data, which formats, how many copies, etc.





  - shift to higher usage of reduced data formats essential to meet current storage budget

Alternate Scenarios

- Computing and storage are currently ***the*** limiting factor on how much data CMS can collect
  - Trigger and DAQ capable of writing data at least 2x as fast as current limit
  - Forced to discard potentially interesting events
- 2012 Running: CMS is pursuing "Data Parking"
  - alternate trigger streams with a total bandwidth equal to the "high priority" triggers is being written to disk/tape
    - no prompt processing
  - will be processed later (during next year's shutdown) at the Tier 0 and Tier 1's
  - efficient use of computing resources during shutdown

Data preservation in HEP

- What to do with all of this data?
  - Irreplaceable resource
  - should be preserved, some how, for the future
- DPHEP Working Group
  - Convened by International Committee on Future Accelerators (ICFA)
  - ~ 100 members from different HEP experiments, Labs
  - Two Reports:
    - DPHEP-2009-00, http://arxiv.org/pdf/0912.0255
    - DPHEP-2012-01, May 2012, arXiv:1205.4667v1
  - Conclusions:
    - "an urgent and vigorous action is needed to ensure data preservation in HEP"
    - "A clear and internationally coherent policy should be defined and implemented"

Data Tiers

- DPHEP effort defined four data tiers:
  1. Published results, along with additional analysis-related information, leading to more complete documentation of a given analysis
  2. Processed data available in a simplified format (i.e., particle four vectors) that can be used for outreach and simplified additional analyses
  3. The full processed experimental data and simulated data and the associated software for accessing and analyzing the data
  4. The full raw data of the experiment and all of the software necessary for processing the data into a form where it can be useful for analysis
- DPHEP is planning a global coordination project
  - cooperation between national labs, stakeholders within each experiment
    - includes no-longer-running experiments like BaBar and Tevatron

Tiers and Data Preservation

Preservation Model		Use Case	
1	Provide additional documentation	Publication related info search	Documentation
2	Preserve the data in a simplified format	Outreach, simple analyses	Outreach/Science
3	Preserve the analysis level software and data format	Full scientific analysis, based on the existing reconstruction	Technical Preservation Projects/Science
4	Preserve the reconstruction and simulation software as well as the basic level data	Retain the full potential of the experimental data	Technical Preservation Projects/Science

Data Preservation

- Current efforts exist for Tiers 1 and 2:
  - supplementary INSPIRE content gives more complete information for publications (http://inspirehep.net/)
  - outreach efforts using Tier 2 data already
    - Also: **RECAST**: re-run analysis given new Monte Carlo specified by outside queries (JHEP **1104** (2011) 038 [arXiv:1010.2506])
- Serious work needed for Tiers 3 and 4
  - necessary within experiments themselves to preserve their own data for future analysis
  - outreach/public access component could be added in parallel

CMS Data Preservation

- CMS has approved a Data Preservation and Access plan
  - first LHC experiment to do so
    - other LHC experiments also considering similar policies
  - prompted by US groups needing to define "Data Management Plans" for the funding agencies
- Under Collaboration Board oversight, calls for:
  - appointment of "Data Preservation Coordinator"
    - just done: Kati Lassila-Perini will hold this position
  - "prompt" public release of Tier 1 and Tier 2 data
  - delayed release of Tier 3 data (Tier 4 will not be released)
    - hopes to release some fraction of reconstructed 2010 data in 2013
  - Creative Commons CCO waiver for released data

http://creativecommons.org/publicdomain/zero/1.0.

Preservation Coordination

- Next: implementation of technical infrastructure, policy, etc. to make data available
  - guidance, but no FTEs (yet) from DPHEP
    - suggestions of overall structure, but no concrete implementation plans
  - CMS will rely on internal expertise, coordinate with external agencies
  - would be most efficient to build infrastructure that is re-useable by other experiments, or even other disciplines
- Several efforts in this area exist or are in the pipeline

DASPOS

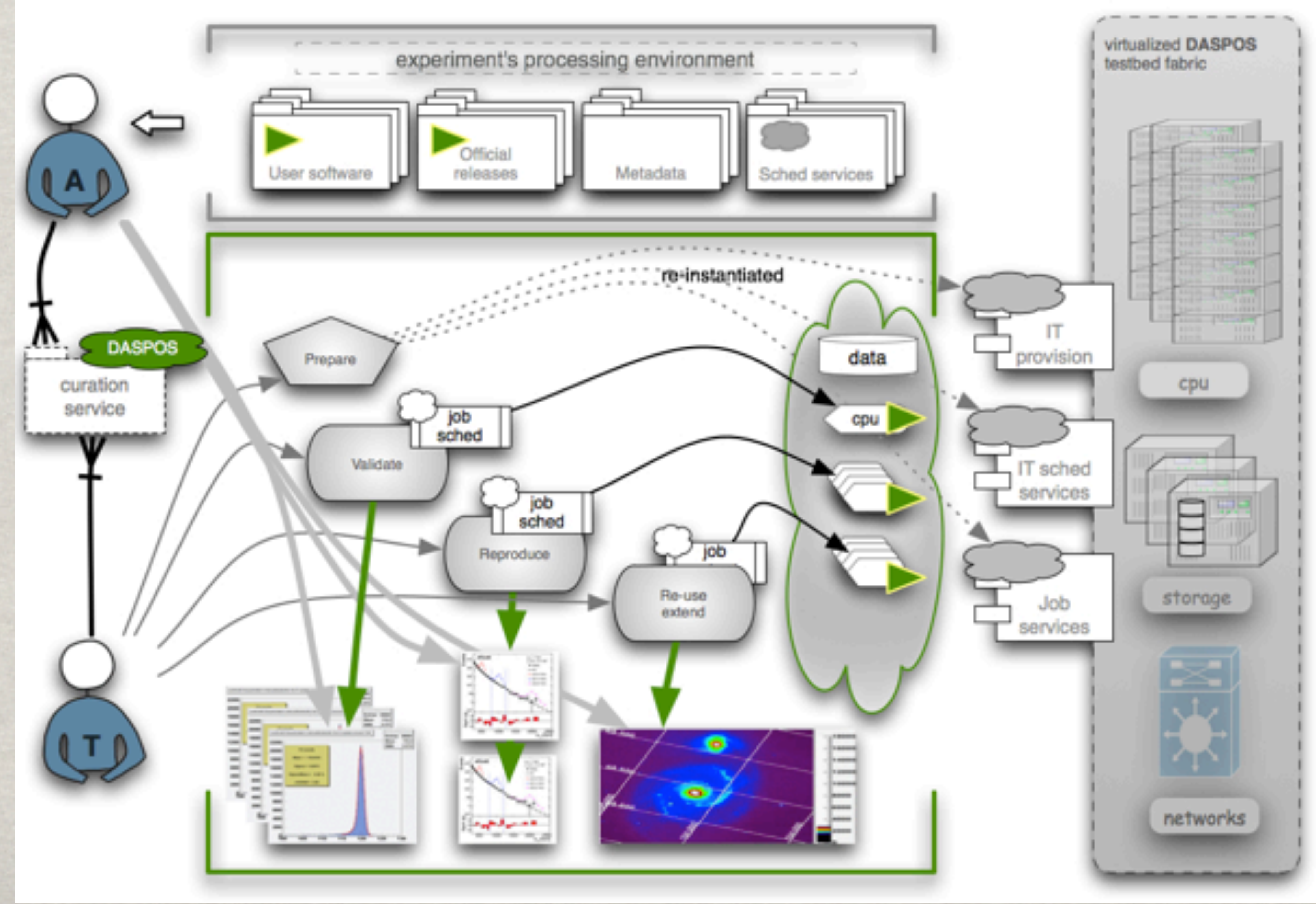
- Data And Software Preservation for Open Science
- multi-disciplinary proposal just submitted to NSF
- Links HEP effort (DPHEP+experiments) to Biology, Astrophysics, Digital Curation
  - aim to achieve some commonality across disciplines in
    - meta-data descriptions of archived data
      - What's in the data, how can it be used?
    - computational description
      - how was the data processed?
      - i.e.: follow Tier 3 reconstructed data to final physics result
    - impact of access policies on preservation infrastructure

DASPOS

- In parallel, will build test technical infrastructure to implement a data preservation system
  - “scouting party” to figure out where the most pressing problems lie, and some solutions
    - incorporate input from multi-disciplinary dialogue, use-case definitions
  - Will translate needs of analysts into a technical implementation of meta-data specification
  - Will implement “physics query” infrastructure across small-scale distributed network
  - end result: “template architecture” for data preservation systems

DASPOS

- Final Milestone: "Curation Challenge"
  - an analyst will reproduce some physics result using only curated information
  - success defined by external auditing team

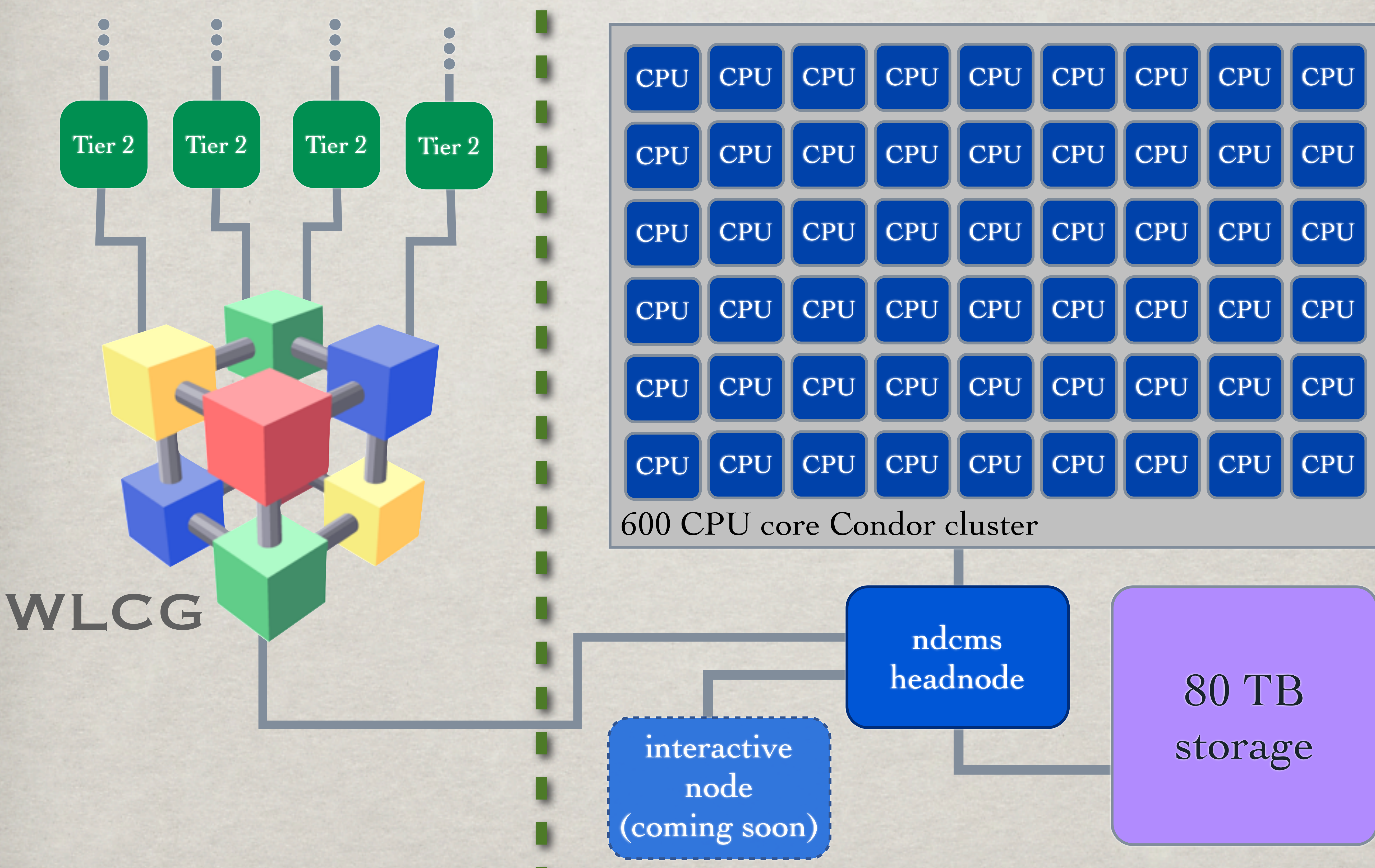


Conclusions

- CMS: Lots of data
- Global data flow, storage issues are under control
- Efficient use of resources is the main limitation
  - more $$$ would help, obviously, but...
  - creative solutions ("Data Parking") can allow more physics output with little additional cost
- Data Preservation and Access will be major issues
  - merely preserving data for re-use within the experiments will be a major challenge
  - No technical infrastructure in place to handle public release, access to data
  - DASPOS project could help

Backup Slides

CMS Tier 3 @ ND



Tier 2
Tier 2
Tier 2
Tier 2
WLCG
CPU CPU CPU CPU CPU CPU CPU CPU CPU
CPU CPU CPU CPU CPU CPU CPU CPU CPU
CPU CPU CPU CPU CPU CPU CPU CPU CPU
CPU CPU CPU CPU CPU CPU CPU CPU CPU
CPU CPU CPU CPU CPU CPU CPU CPU CPU
CPU CPU CPU CPU CPU CPU CPU CPU CPU
600 CPU core Condor cluster
ndcms headnode
80 TB storage
interactive node (coming soon)

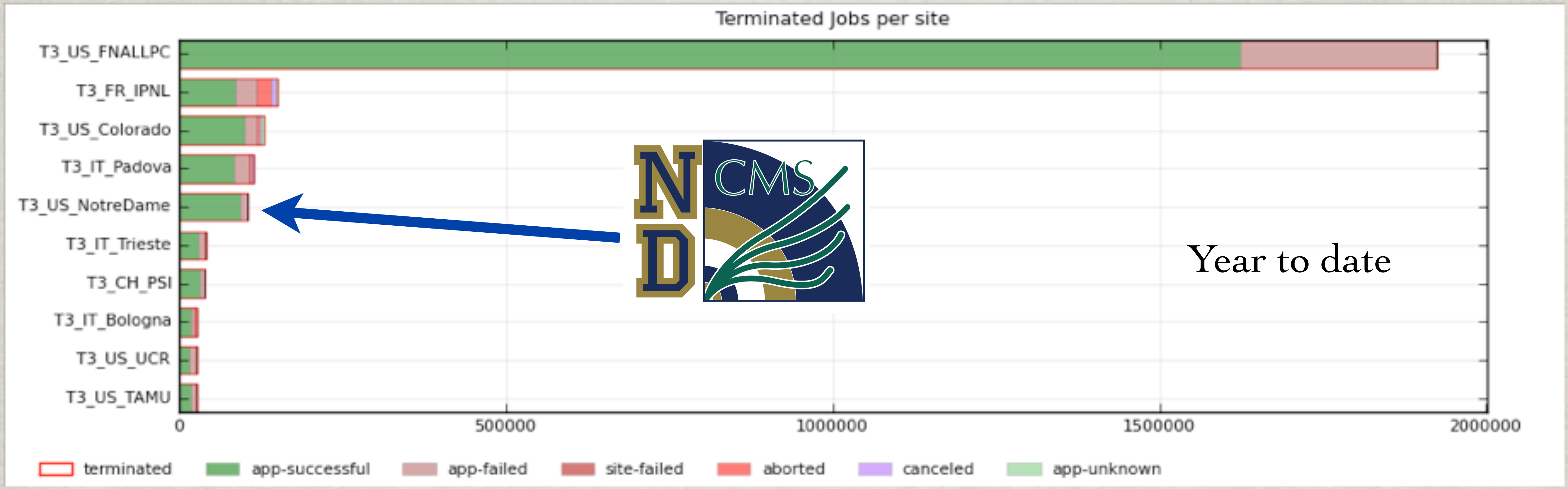
ND CMS Typical Usage

- 2-3 teams (1-2 faculty, 1 PD, 1-3 students + outside collaborators)
- Analysis workflow
  - Process data using GRID at T2 sites; transfer 15-25 TB output to ND T3
  - Further processing: generates another < 1 TB additional data
  - CPU-intensive computations: negligible additional data generated
  - Make discoveries! Publish papers
- Replace dataset with updated/larger dataset every 6-12 months

ND T3 Success Stories

- Have kept 80 TB storage full for ~ 1 year
- Primary processing and storage for several students about to graduate (Sean and Jamie)
- Shared resources with collaborators from other institutions (UVa, OSU, Milano)
  - Shared both storage and processing resources
  - Using standard CMS/GRID interfaces
- Undergrad participation in CMS research grows from 0 to 5 students in 2 years

ND Success Stories



Jobs from last month