LHC Data Analytics

Artur Barczyk
California Institute of Technology
BIG DATA Workshop
NIST, Gaithersburg, June 14, 2012

DISCLAIMER

This presentation intends to give an overview of LHC data processing, based on samples and general notions. It is as such intrinsically incomplete, as it’s impossible to cover this vast area in a short time without idiosyncratic bias.

References to detailed information were intended, and where missing can be obtained from the presenter.

OUTLINE

LHC and its data

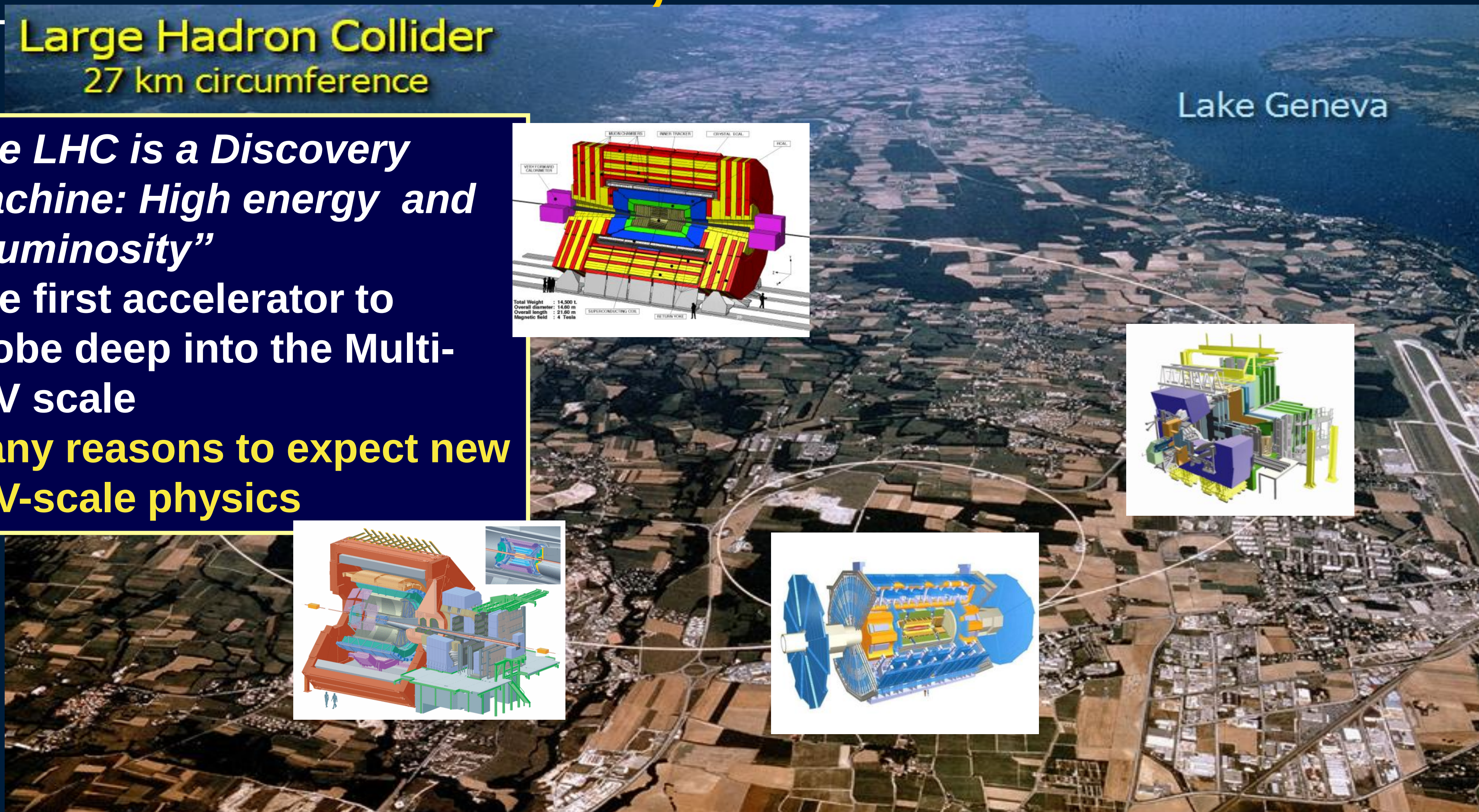
LHC data processing and analysis chain

Data sizes, Data rates

Computing Infrastructures

THE LHC AND ITS COMPUTING INFRASTRUCTURE

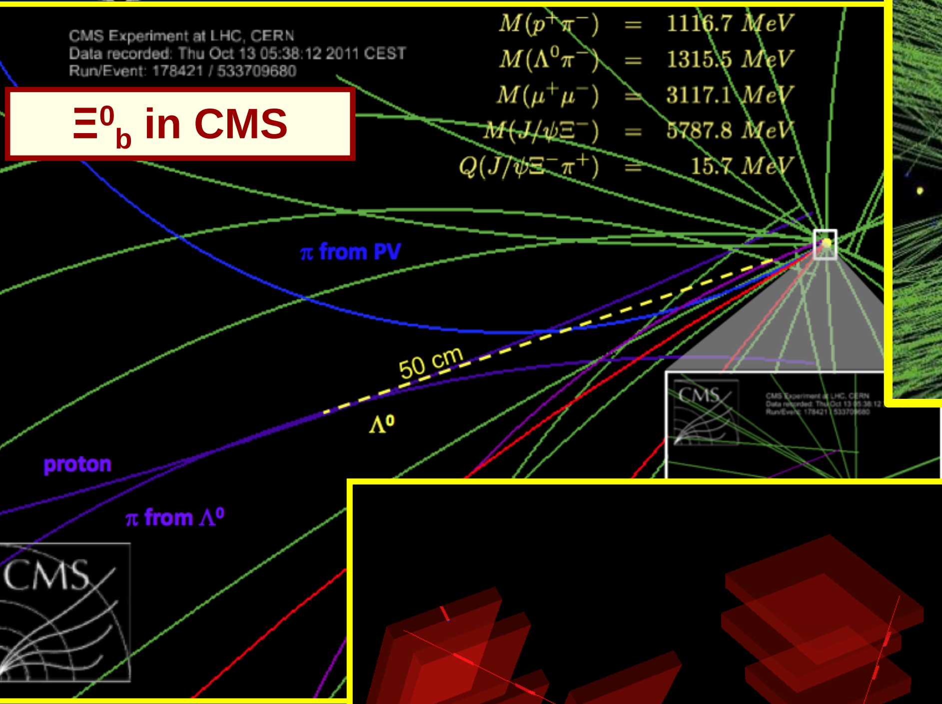
Large Hadron Collider
7 TeV in 2010-11, 8 TeV in 2012



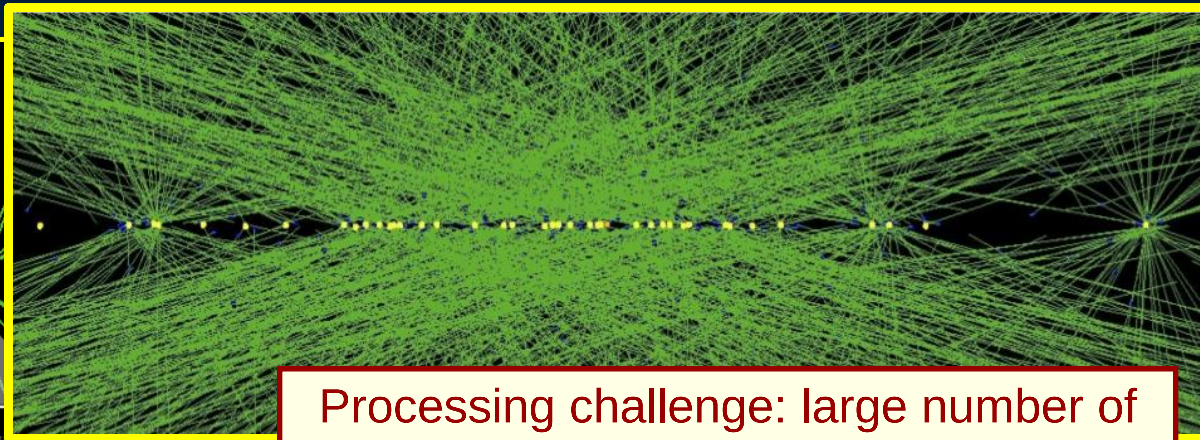
- *The LHC is a Discovery Machine: High energy and "Luminosity"*
- **The first accelerator to probe deep into the Multi-TeV scale**
- **Many reasons to expect new TeV-scale physics**

Higgs, SUSY, Substructures, CP Violation, QG Plasma, *… Gravitons, Extra Dimensions, Low Mass Strings, … the Unexpected*

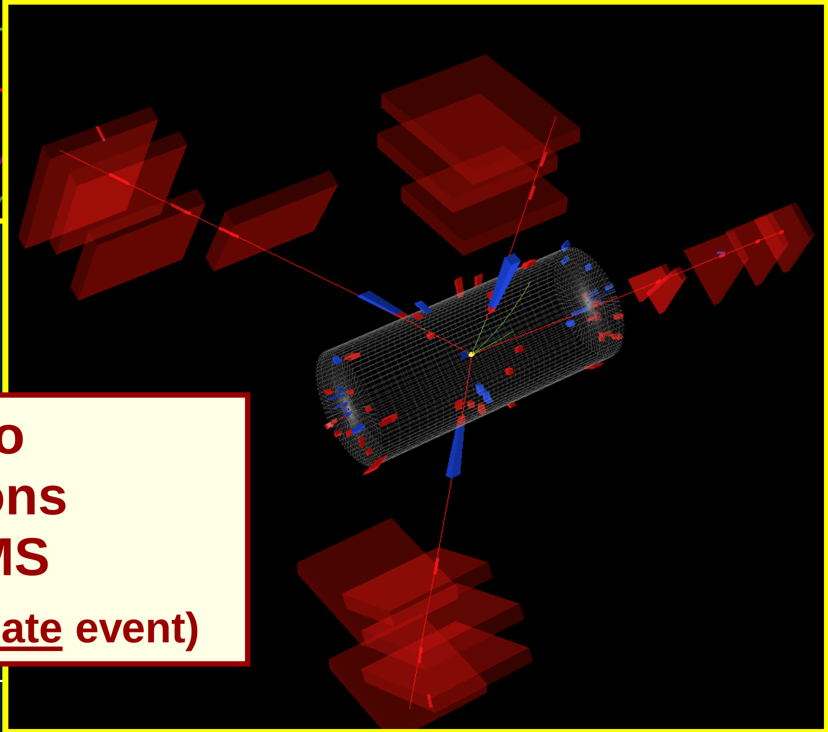
HEP Events: the basic processing unit



Ξ^0_b in CMS

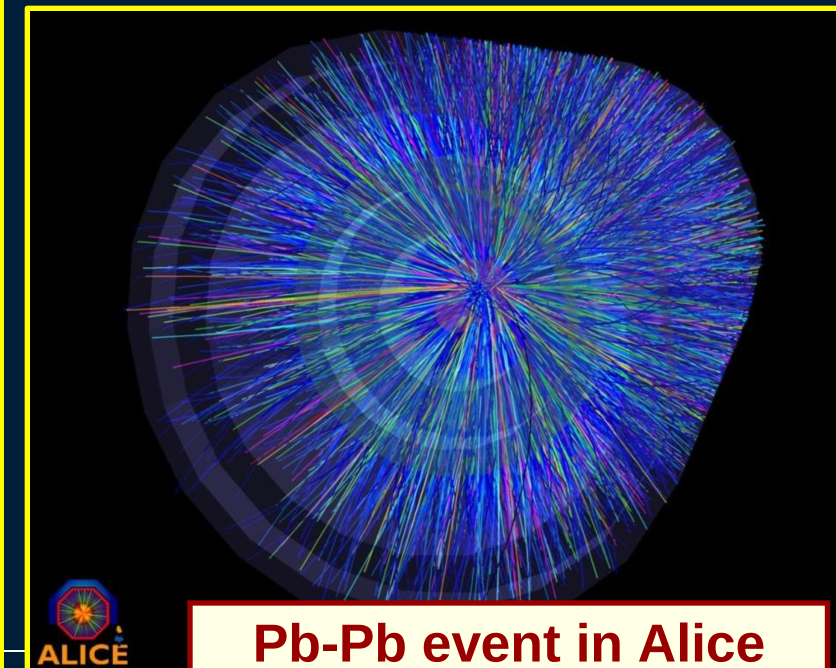


Processing challenge: large number of primary vertices in single bunch crossing



ZZ to 4 Muons in CMS

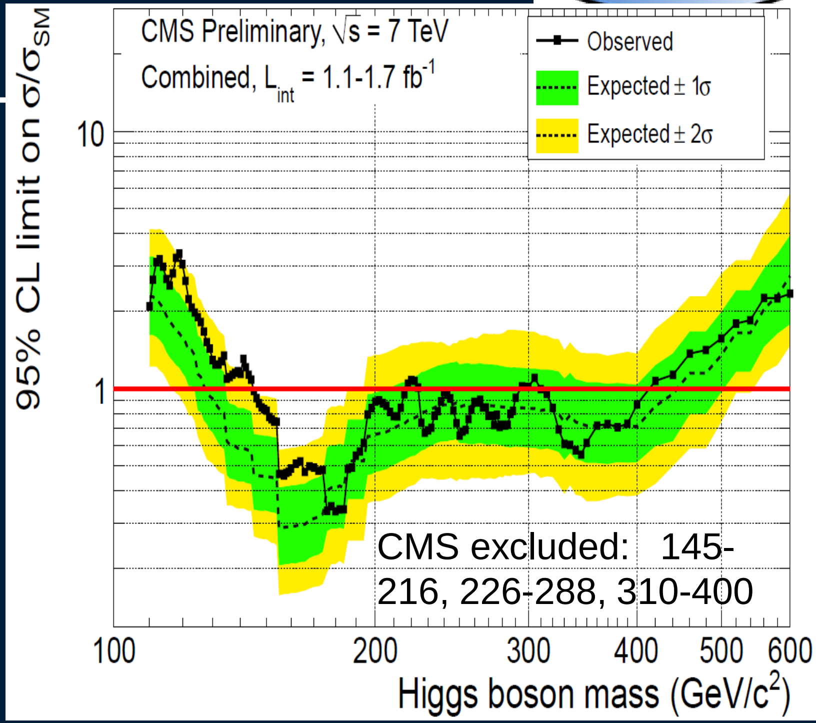
(Higgs <u>candidate</u> event)

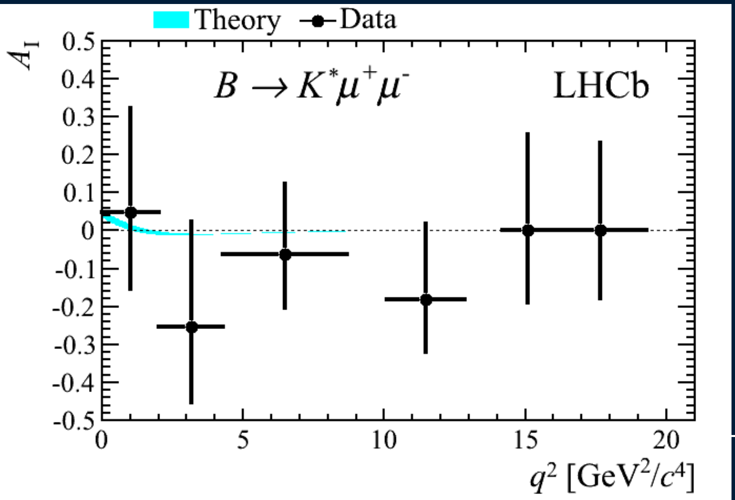


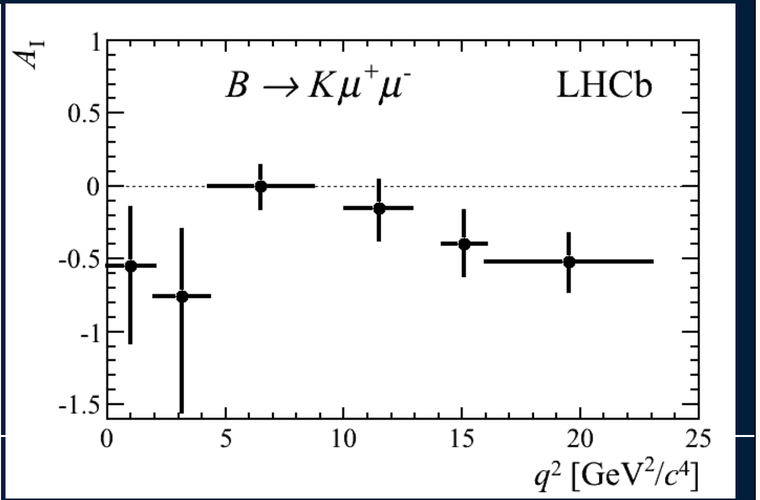
Pb-Pb event in Alice

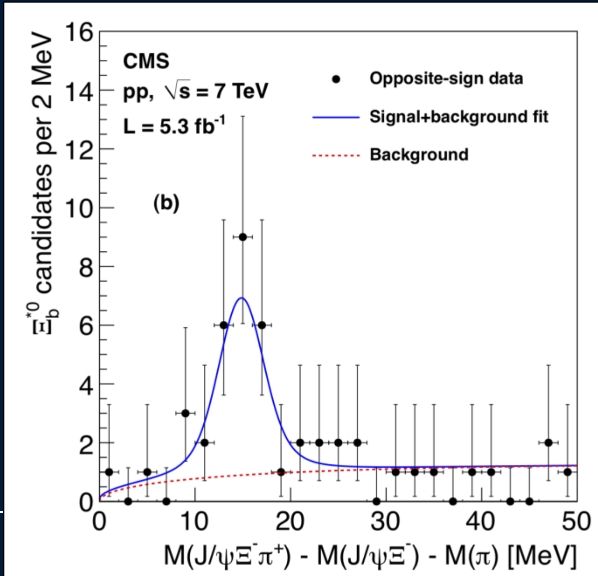
How do we search for New Physics? (Simplified)

- **Statistical Analysis**
  - Very low S/B ratio (10^{-9} – 10^{-13})
- **Looking for the predicted/expected**
  - Select events with expected signatures; Evaluate signal over background estimate (Monte Carlo simulation)
- **Looking for the unexpected**
  - Use known processes, do precision measurement, look out for deviations





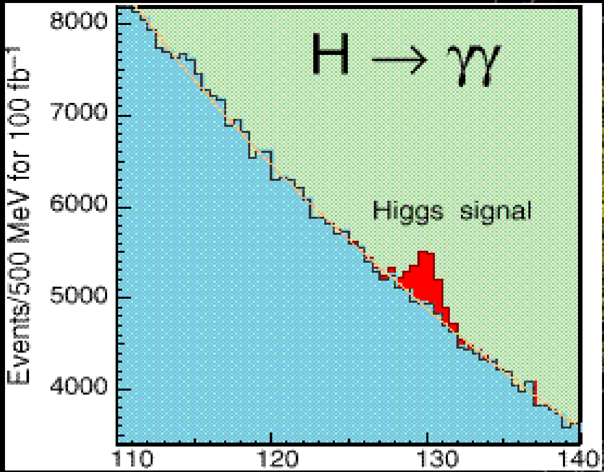


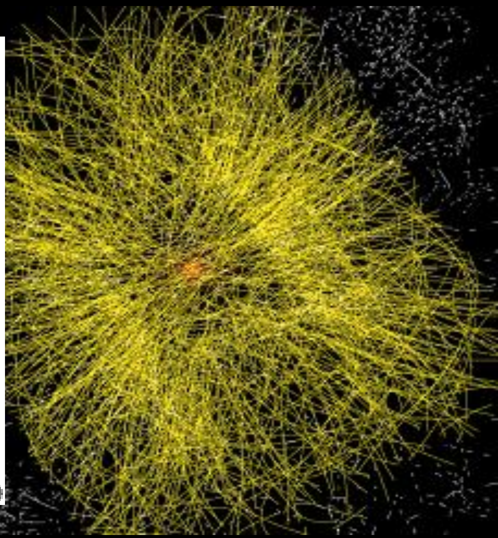


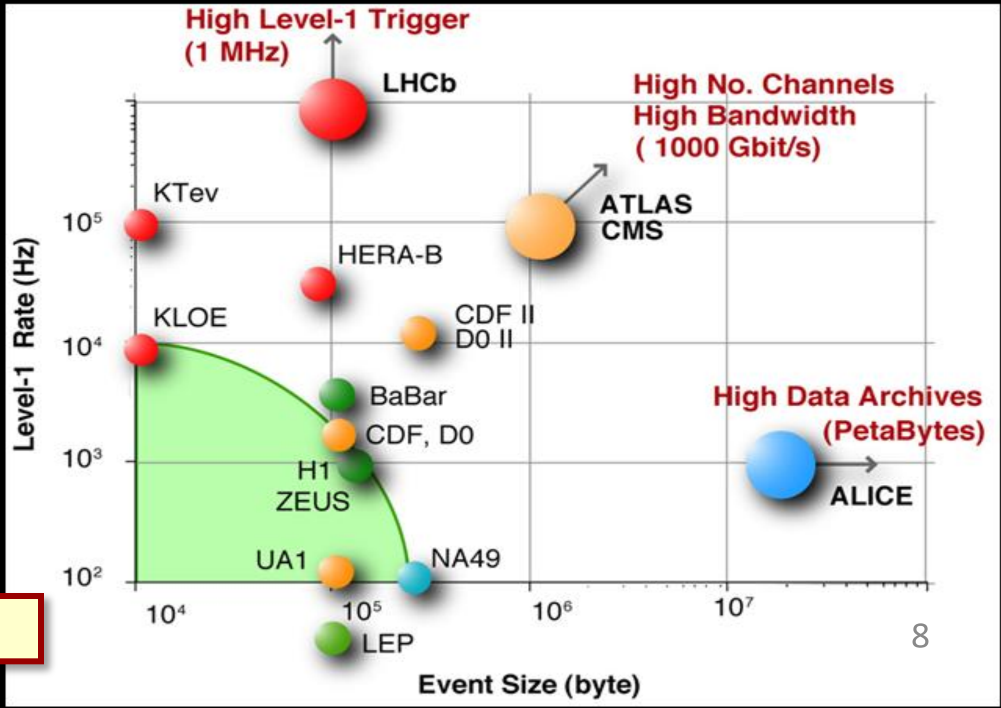
The LHC Computing Challenge

- Signal/Noise: 10^{-13} (10^{-9} offline)
- Data volume
  - High rate * large number of channels * 4 experiments
  - → 15 PetaBytes of new data each year → 23 PB in 2011
- Compute power
  - Event complexity * Nb. events * thousands users
  - → 200 k CPUs → 250 k CPU
  - → 45 PB of disk storage → 150 PB
- Worldwide analysis & funding
  - Computing funding locally in major regions & countries
  - Efficient analysis
  - → GRID technology

Ian Bird, CERN



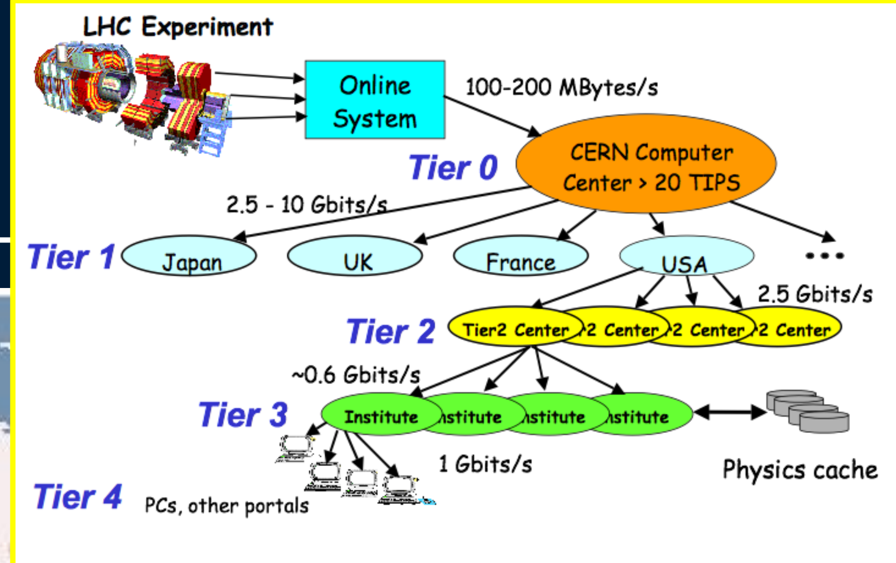


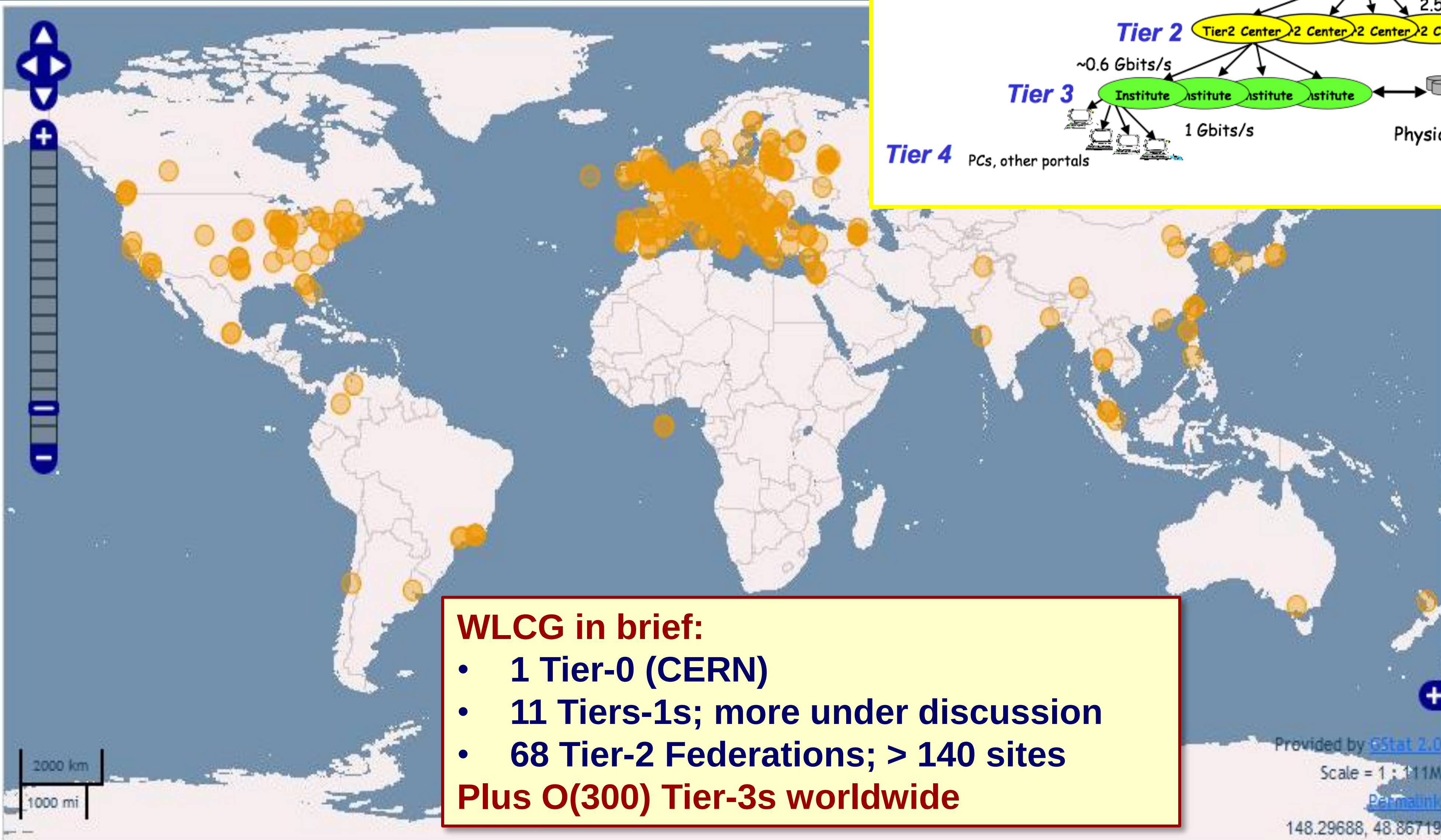


WLCG Worldwide LHC Computing Grid

LCG

LHC Computing Infrastructure





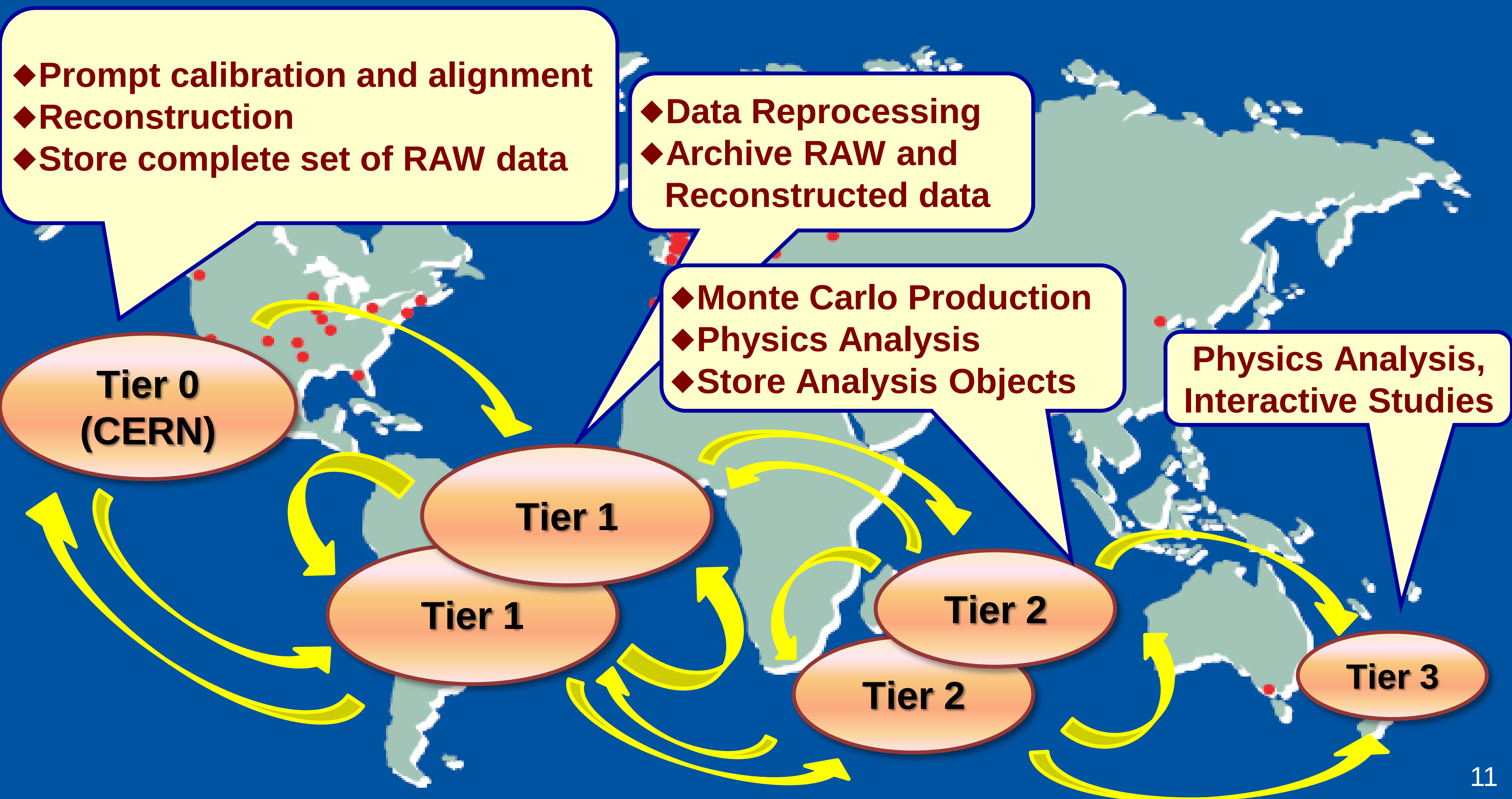
WLCG in brief:
- **1 Tier-0 (CERN)**
- **11 Tiers-1s; more under discussion**
- **68 Tier-2 Federations; > 140 sites**

Plus O(300) Tier-3s worldwide

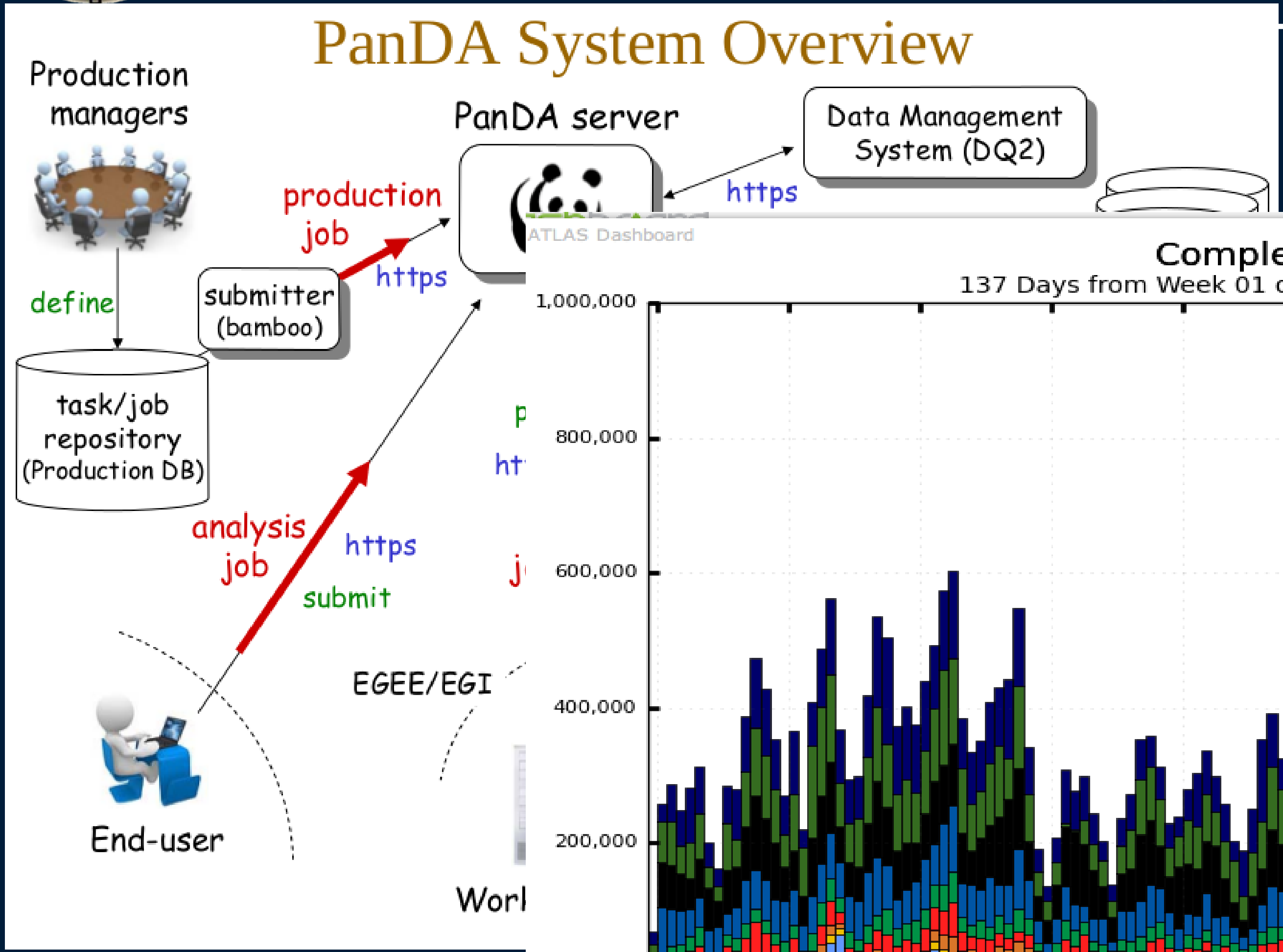
COMPUTING MODELS

The transition

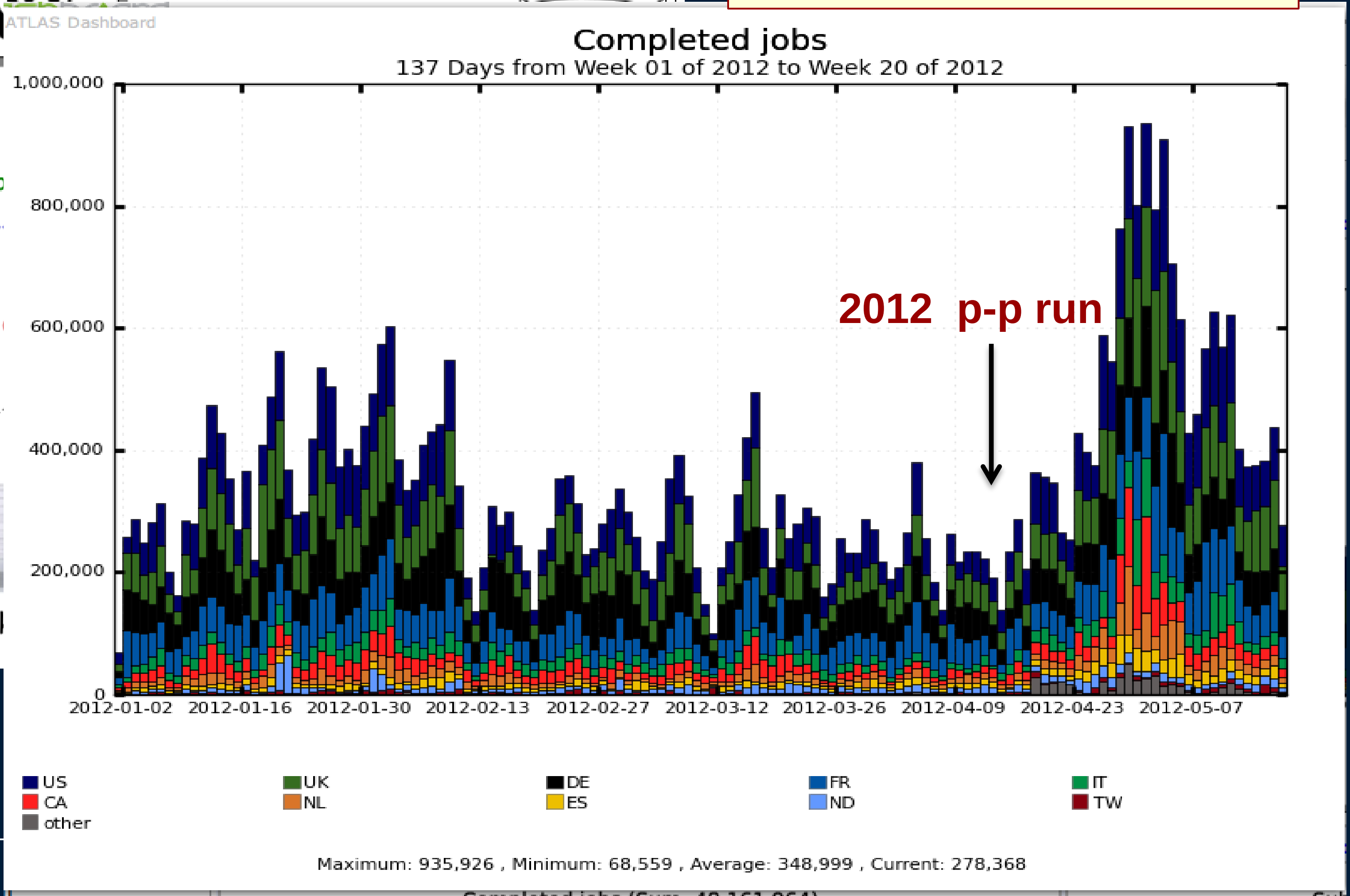
Computing Site Roles (so far)



Grid Jobs - ATLAS example: PanDA Workload Management



~35'000 Atlas jobs running concurrently at any moment

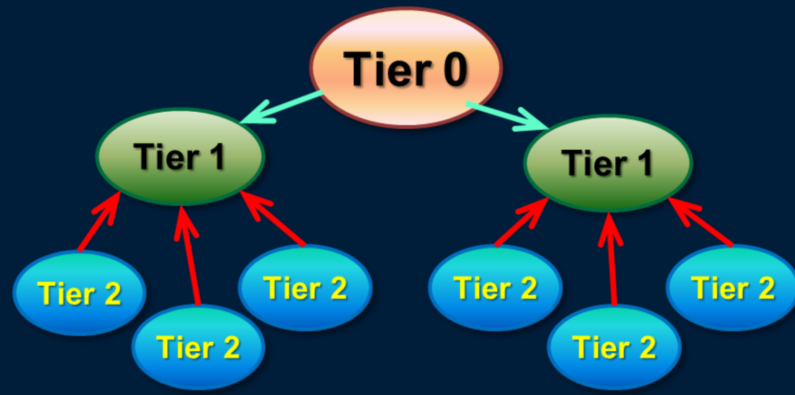


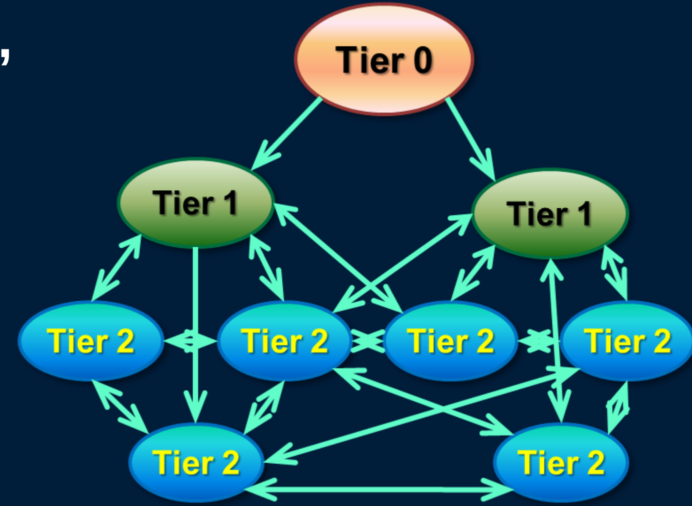
Parallel processing

- **At event granularity:**
  - LHC experiments typically use 1 core per job with sequential event processing
  - The Grid is perfectly matching this process workflow
    - Jobs dispatched to the Grid, running over event files (present at the executing site)
  - No inter-node or inter-site process synchronization (distributed, but independent computation)
- **New approaches being investigated:**
  - Make more efficient use of multi-core, multi-cpu architectures
  - Possibly make use of massively parallel hardware (GPUs)
    - E.g. in event reconstruction
  - Granularity remains at single event level
    - No clear advantage of processing single event on multiple nodes

Computing Models Evolution

- Moving away from the strict MONARC model
- Introduced gradually since 2010
- 3 recurring themes:
  - Flat(ter) hierarchy: Any site can use any other site as source of data
  - Dynamic data caching: Analysis sites will pull datasets from other sites "on demand", including from Tier2s in other regions
    - Possibly in combination with strategic pre-placement of data sets
  - Remote data access: jobs executing locally, using data cached at a remote site in quasi-real time
    - Possibly in combination with local caching
- Variations by experiment
- Increased reliance on network performance!

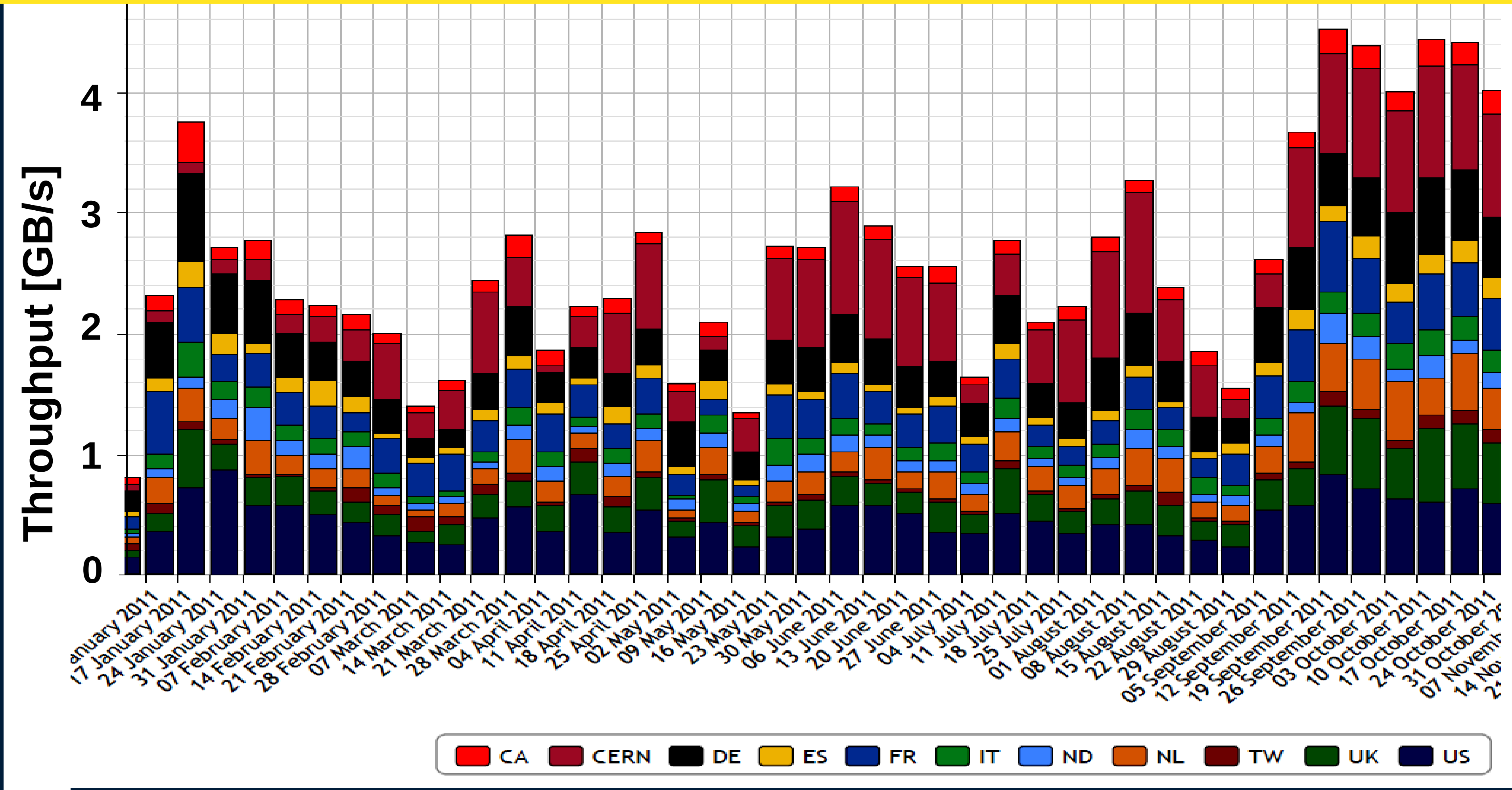




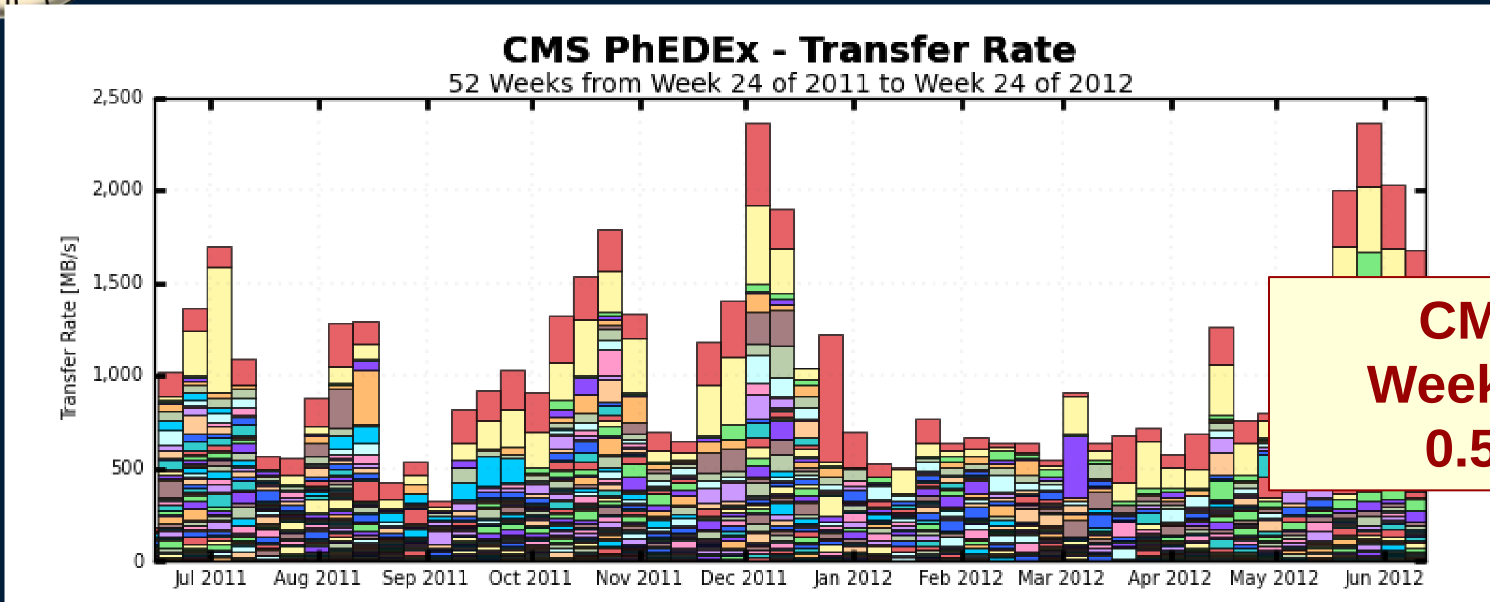
ATLAS Data Flow by Region: Jan. – Nov. 2011

~2.8 Gbytes/sec Average, 4.5 GBytes/sec Peak

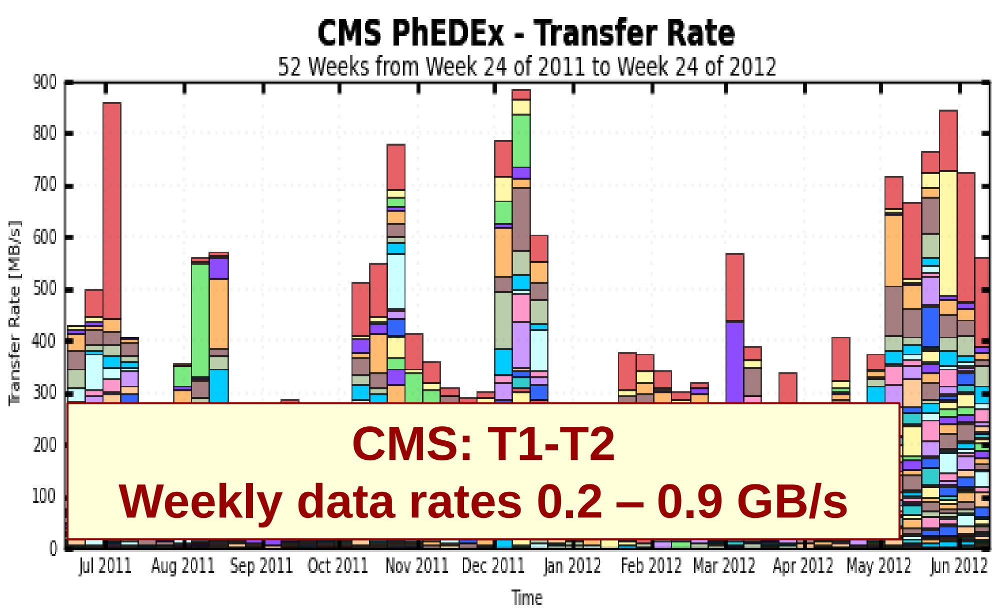
~100 Petabytes Transferred During 2011



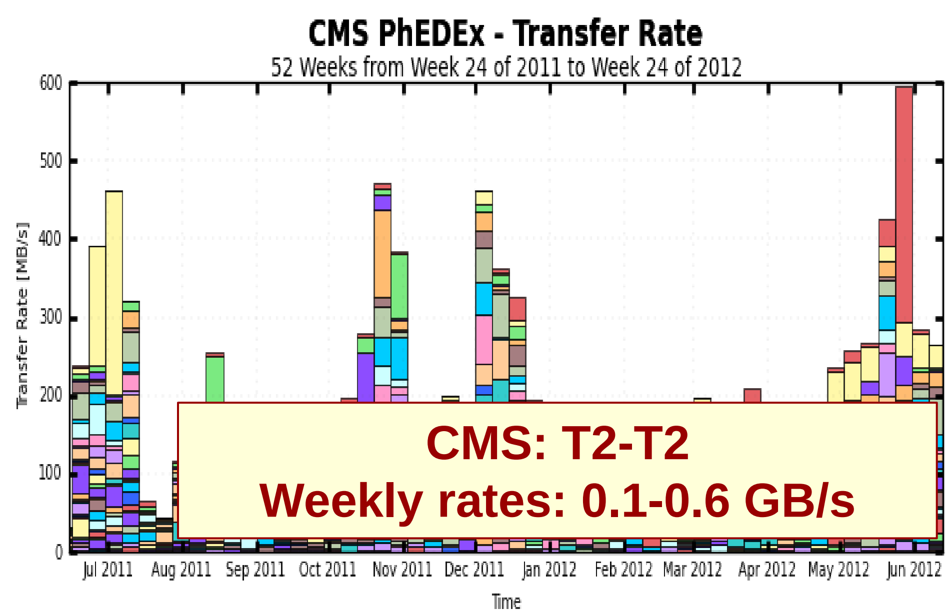
CMS data movement overview



CMS: all sites
Weekly data rates
0.5 - 2.5 GB/s



CMS: T1-T2
Weekly data rates 0.2 – 0.9 GB/s

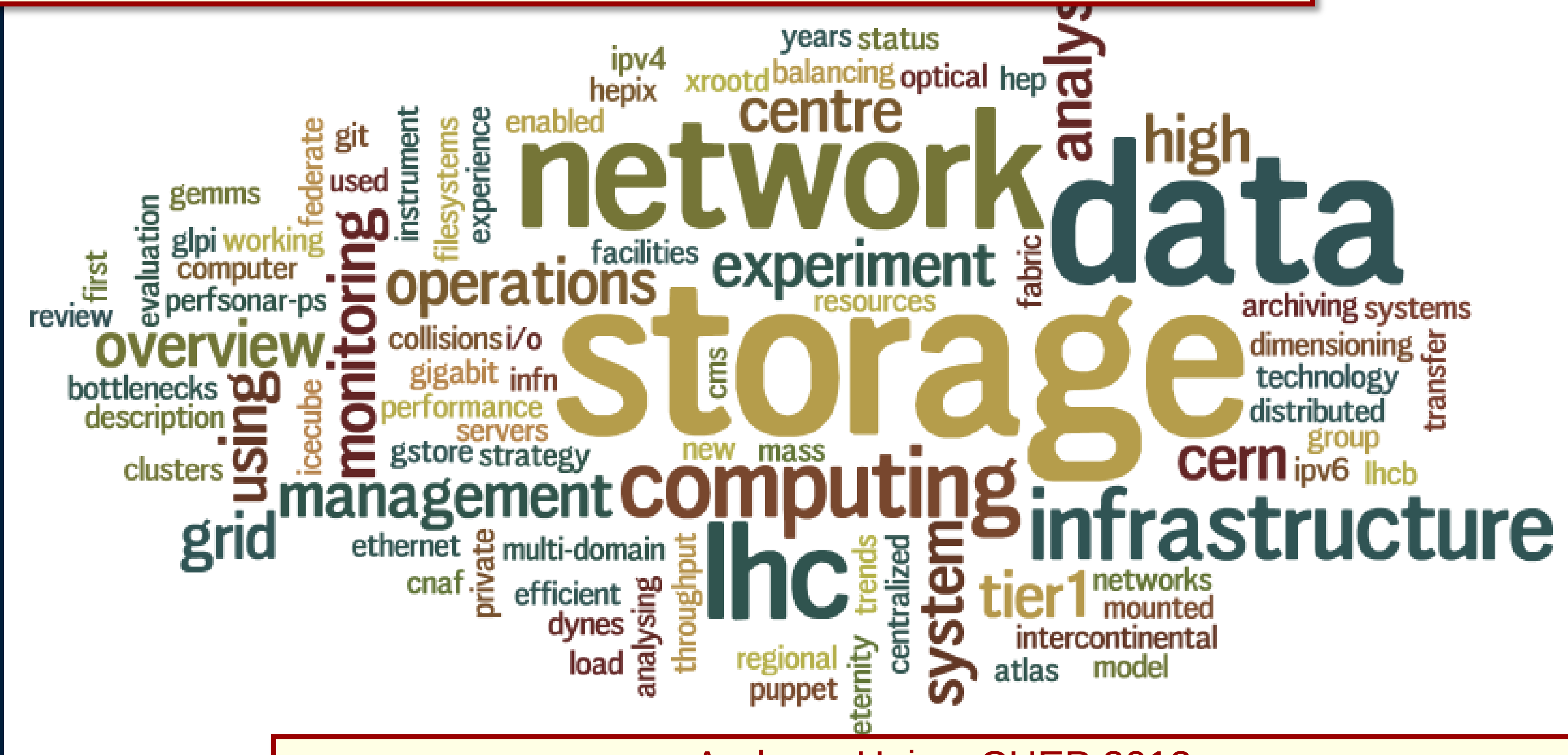


CMS: T2-T2
Weekly rates: 0.1-0.6 GB/s

NETWORK INFRASTRUCTURES FOR LHC DATA PROCESSING

HEP: Highly Networked Computing Infrastructure

CHEP 2012: Network was 2nd most discussed topic in the Computer Facilities, Production Grids and Networking track

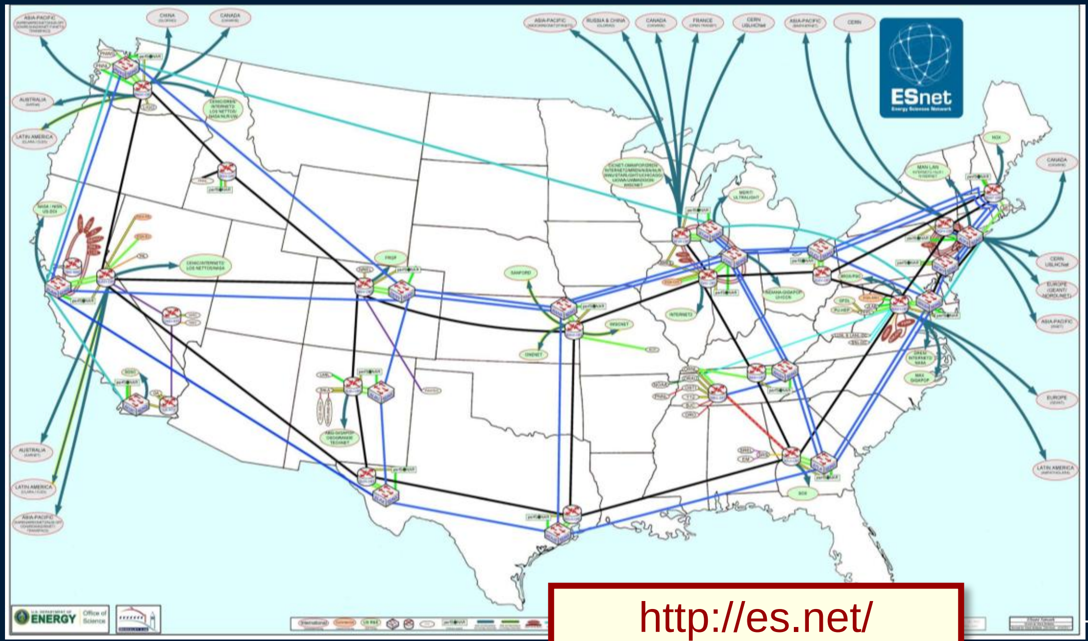


Andreas Heiss, CHEP 2012,
Track Summary: Computer Facilities, Production Grids and Networking

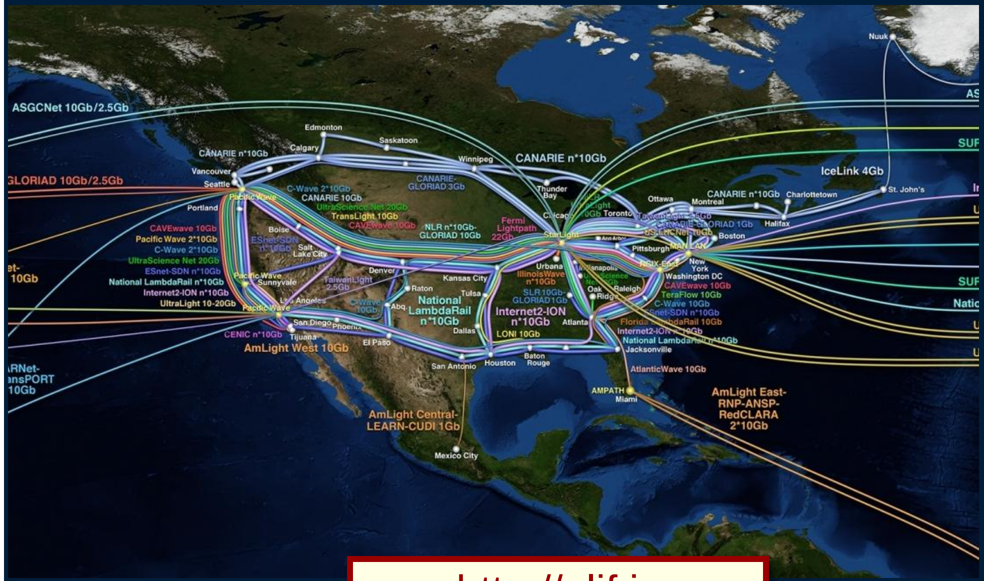
R&E Networking Landscape

- **R&E networks as substrate for LHC data movement**
  - National (e.g. ESnet, SURFNet, RNP, …)
  - Regional (e.g. MiLR, CENIC, NORDUNet, …)
  - International (e.g. NORDUnet, GEANT, ACE, …)
  - Open Exchanges (e.g. Starlight, MANLAN, NetherLight, CERNLight, …)
  - Dedicated, mission–oriented network (US LHCNet)

For more complete listing see e.g. ICFA-SCIC report at http://icfa-scic.web.cern.ch/

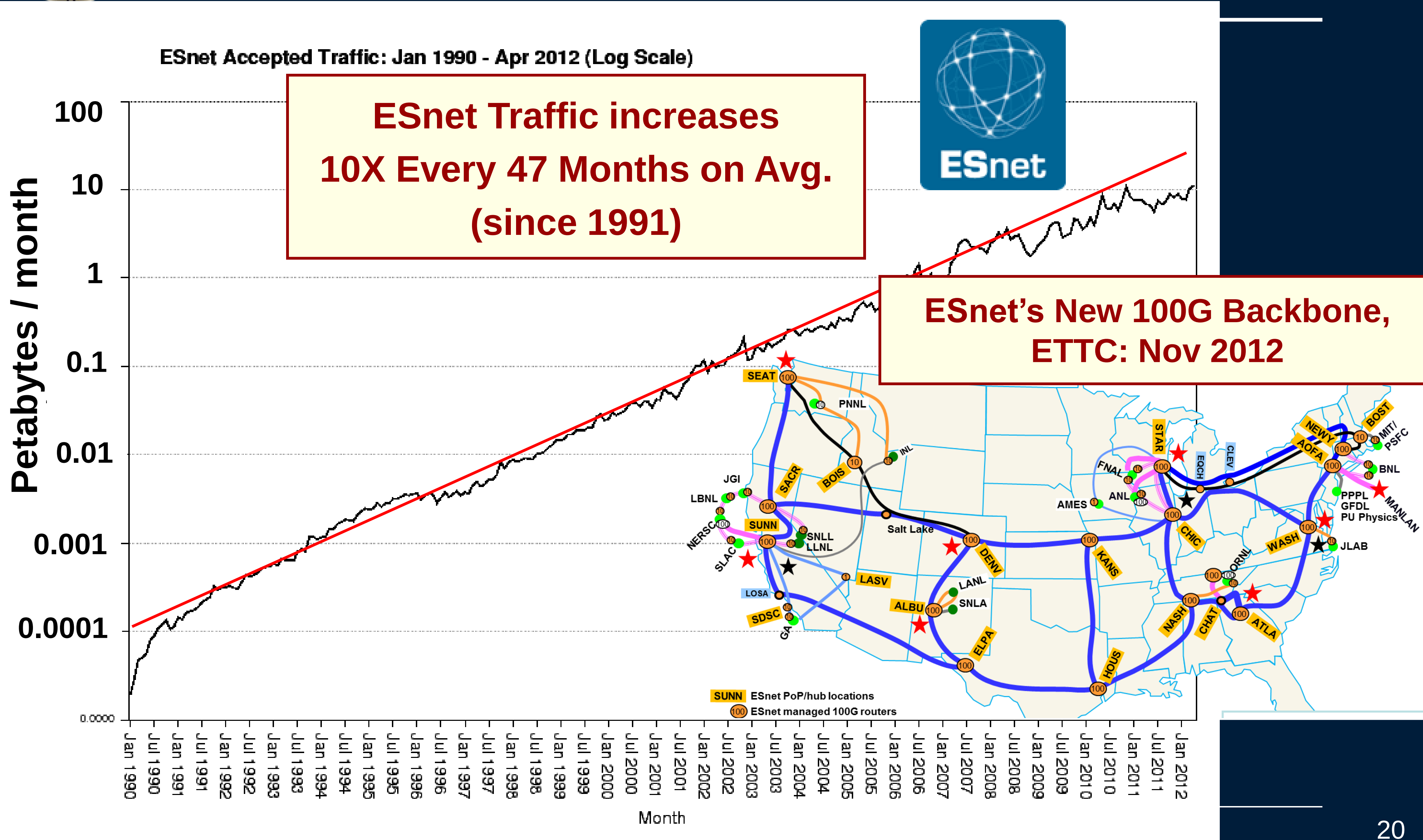


http://es.net/



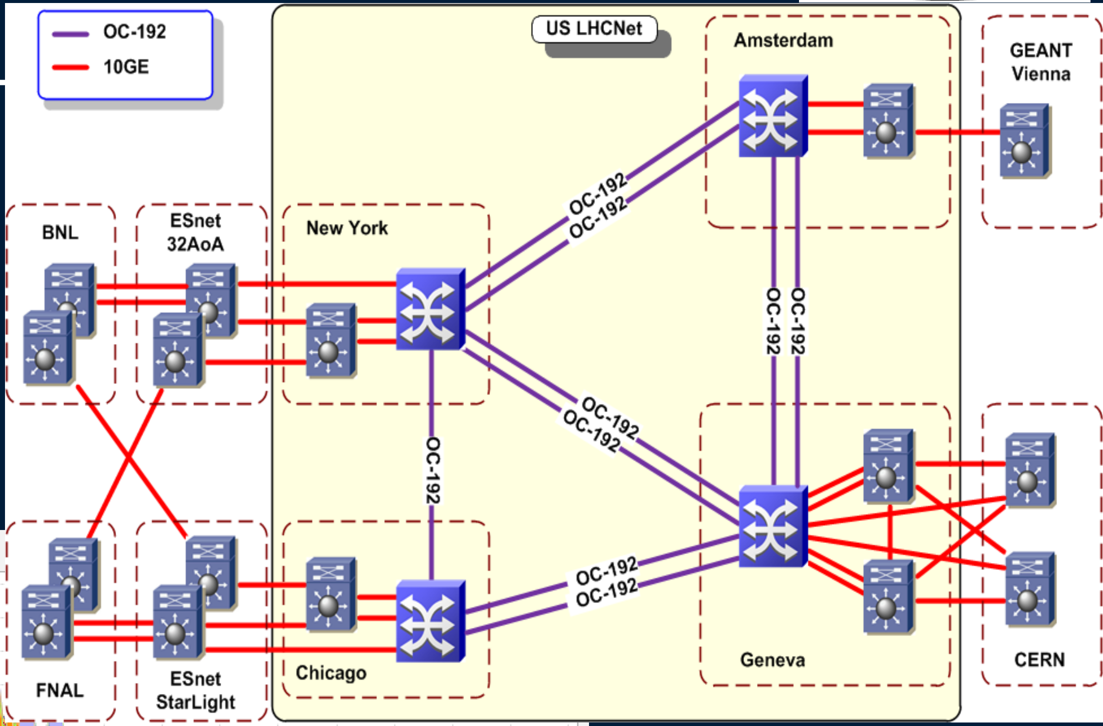
http://glif.is

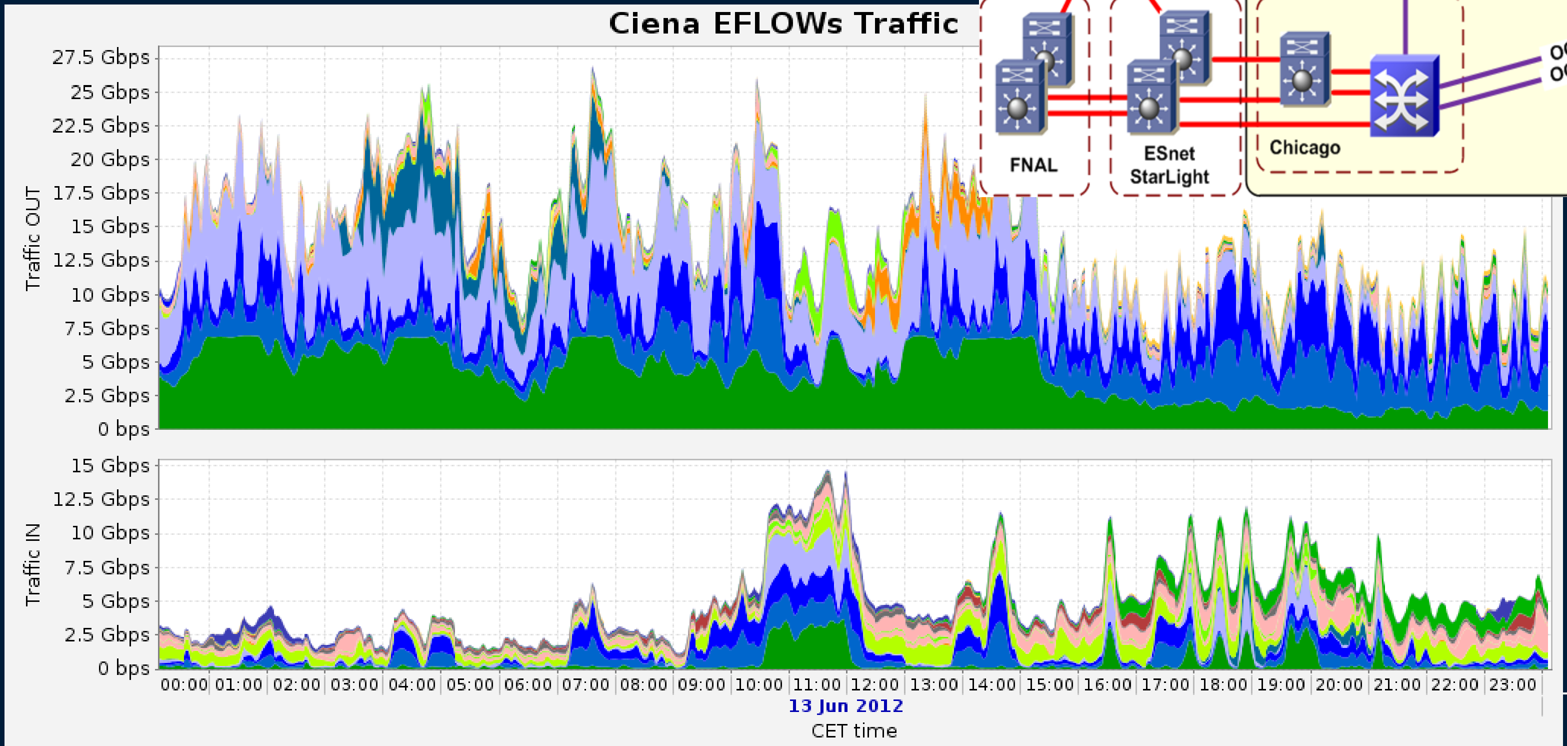
R&E Networks are vital for modern science



US LHCNet: Dedicated Network Infrastructure

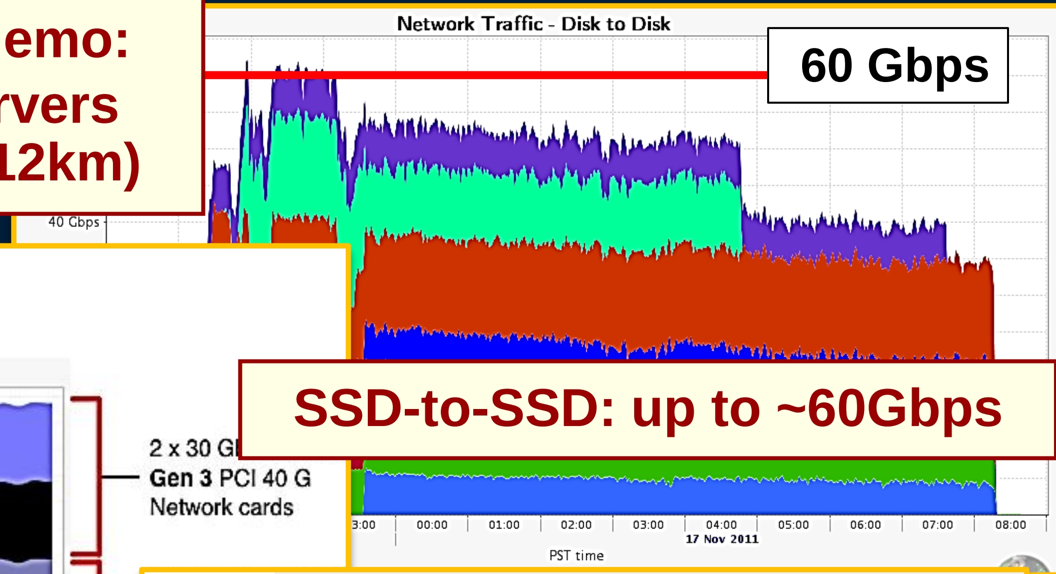
- Mission-oriented network
- Dedicated to LHC data movement between Europe and US
- High availability goal (99.95+), despite challenges of submarine environment



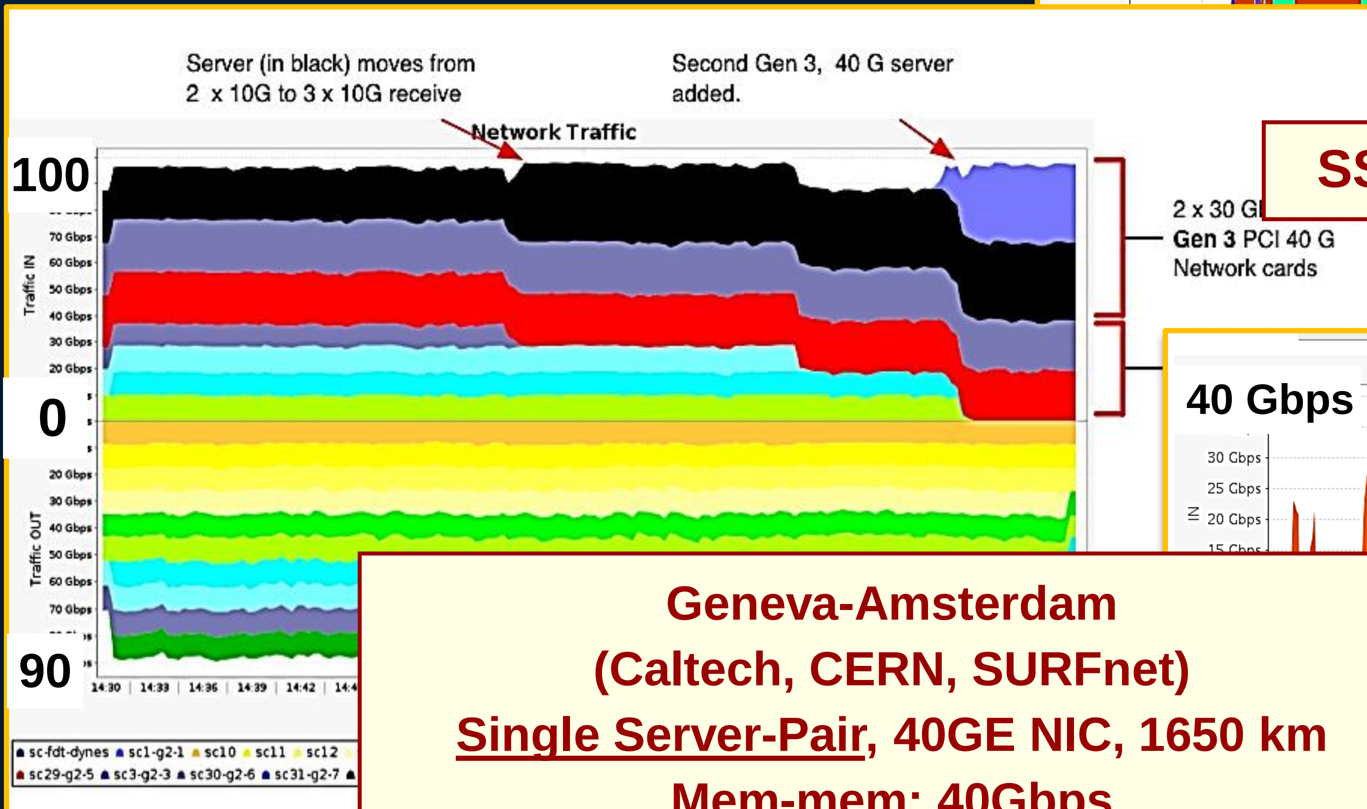


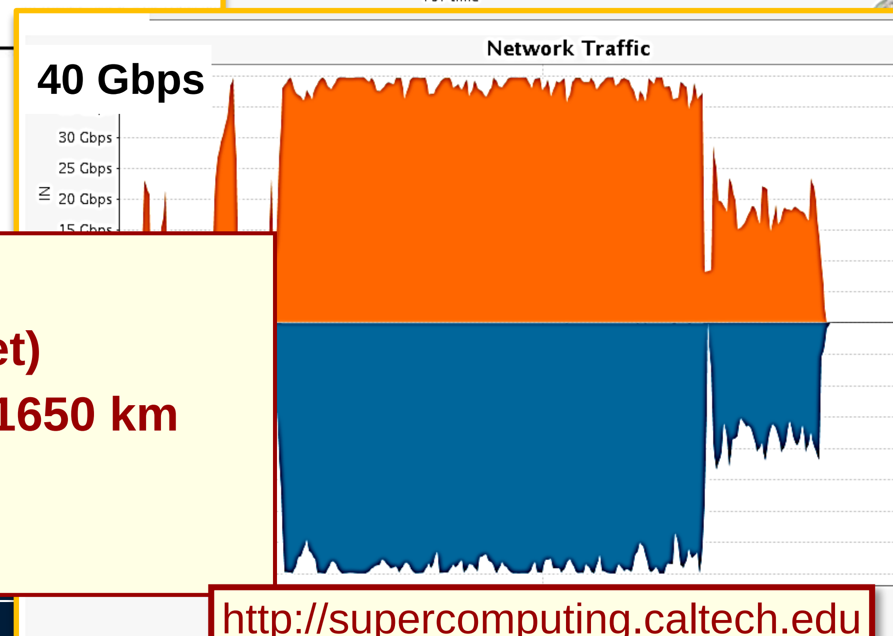
100 & 40 Gbps System Integration

SuperComputing 2012 Caltech demo:
100Gbps single wave; 40 GE servers
Seattle - University of Victoria (212km)



SSD-to-SSD: up to ~60Gbps





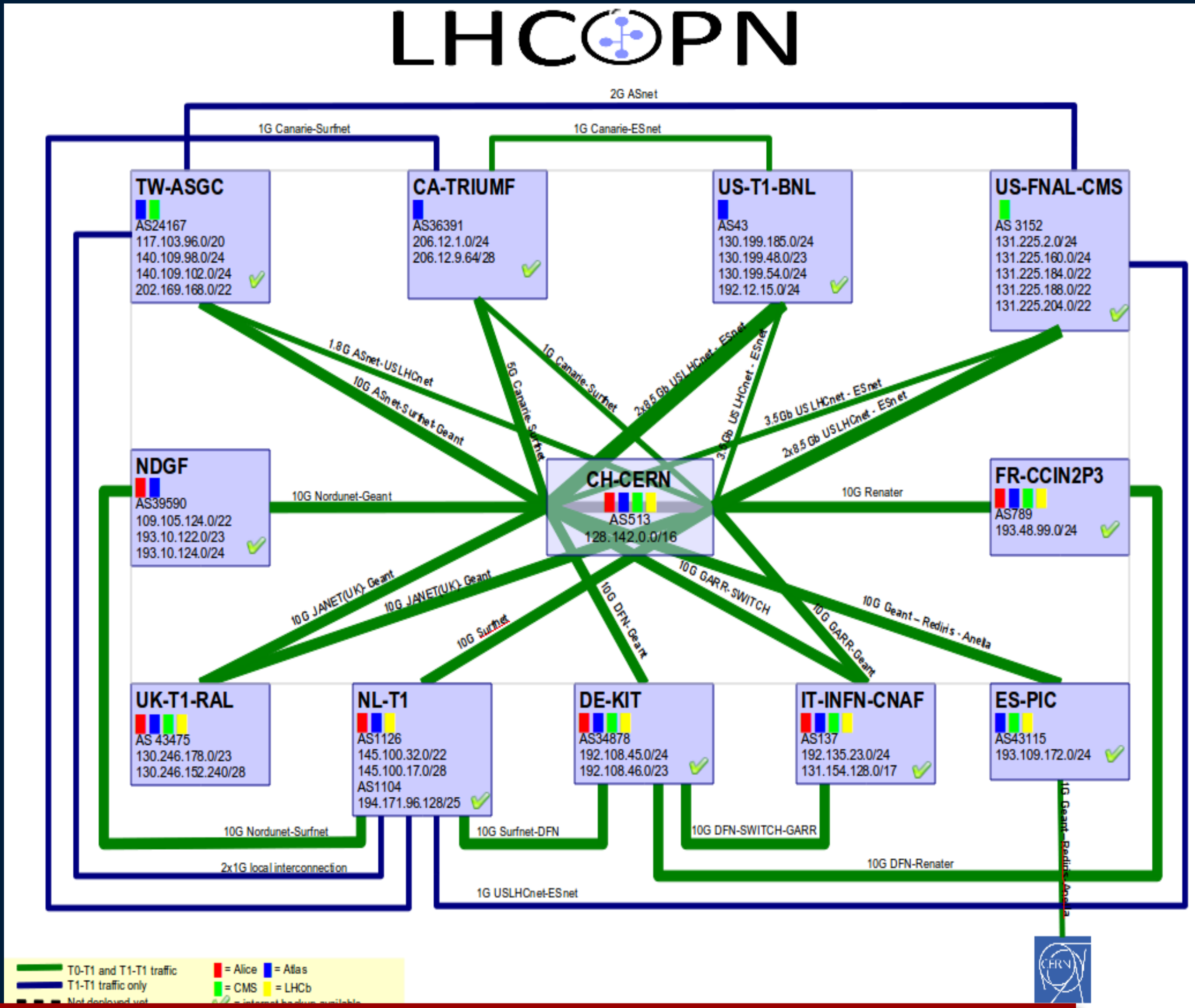
Geneva-Amsterdam
(Caltech, CERN, SURFnet)
<u>Single Server-Pair</u>, 40GE NIC, 1650 km
Mem-mem: 40Gbps
SSD-SSD: ~18Gbps

FDT: http://monalisa.cern.ch/FDT/

http://supercomputing.caltech.edu

The LHC Optical Private Network
Serving LHC Tier0 and Tier1 sites

- **Dedicated network resources for Tier0 and Tier1 data movement**
- **Layer 2 overlay on R&E infrastructure**
- **130 Gbps total Tier0-Tier1 capacity (today)**
- **Simple architecture**
  - Point-to-point Layer 2 circuits
  - Flexible and scalable topology
- **Grew organically**
  - From star to partial mesh
- **Open to technology choices**
  - have to satisfy requirements
  - OC-192/SDH-64, EoMPLS, …
- **Federated governance model**
  - Coordination between stakeholders



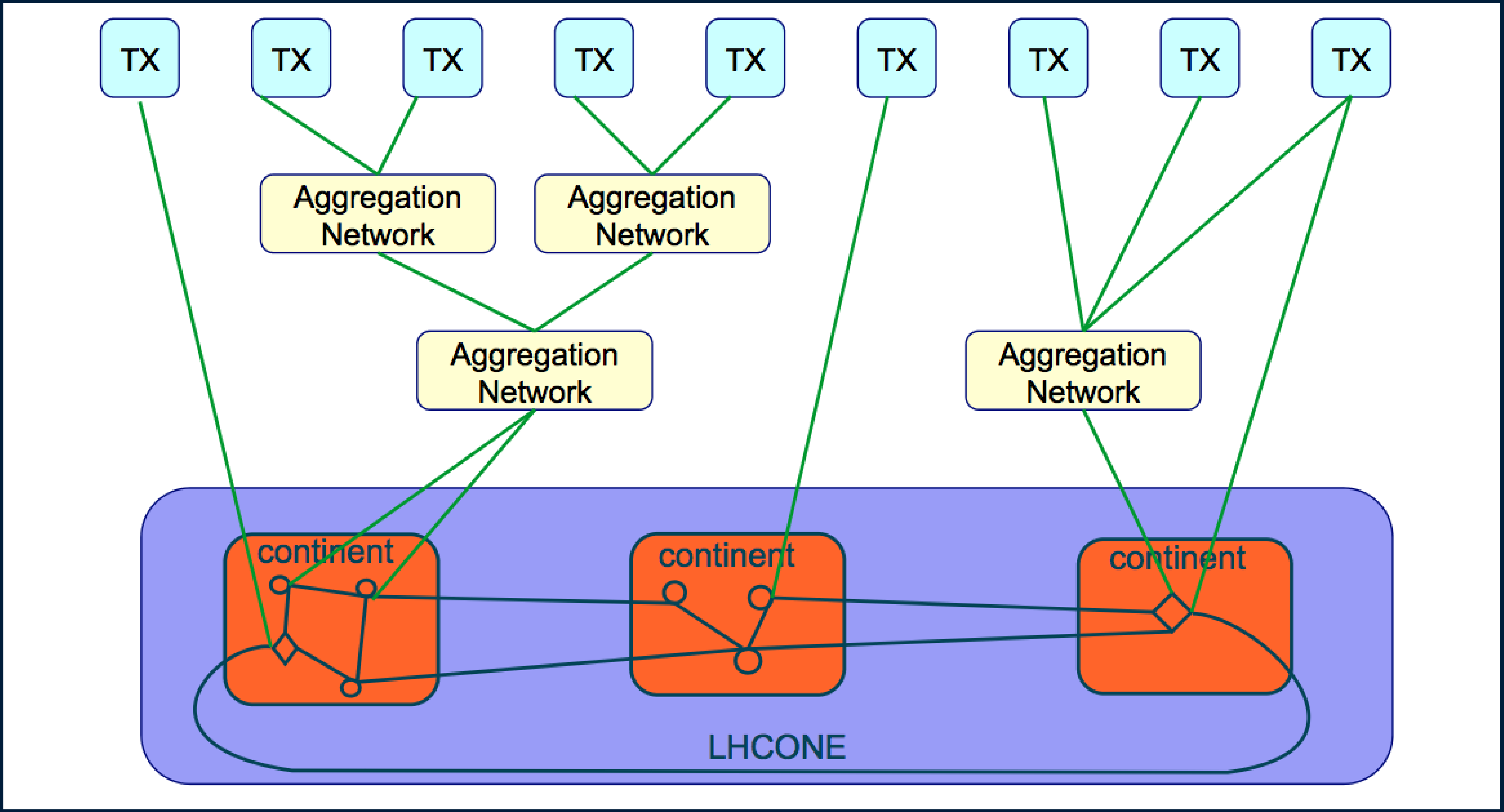
https://twiki.cern.ch/twiki/bin/view/LHCOPN/WebHome

LHC Open Network Environment - The Background

- So far, T1-T2, T2-T2, and T3 data movements have been using General Purpose R&E Network infrastructure
  - Shared resources (with other science fields)
  - Mostly best effort service
- Increased reliance on network performance → need more than best effort
  - Separate large LHC data flows from routed R&E GPN
- Collaboration on global scale, diverse environment, many parties
  - Solution has to be Open, Neutral and Diverse
  - Agility and Expandability
    - Scalable in bandwidth, extent and scope
- Organic activity, growing over time according to needs
- LHCONE Services being constructed:
  - Multipoint, virtual network (logical traffic separation and TE possibility)
  - Static/dynamic point-to-point Layer 2 circuits (guaranteed bandwidth, for high-throughput data movement)
  - Monitoring/diagnostic

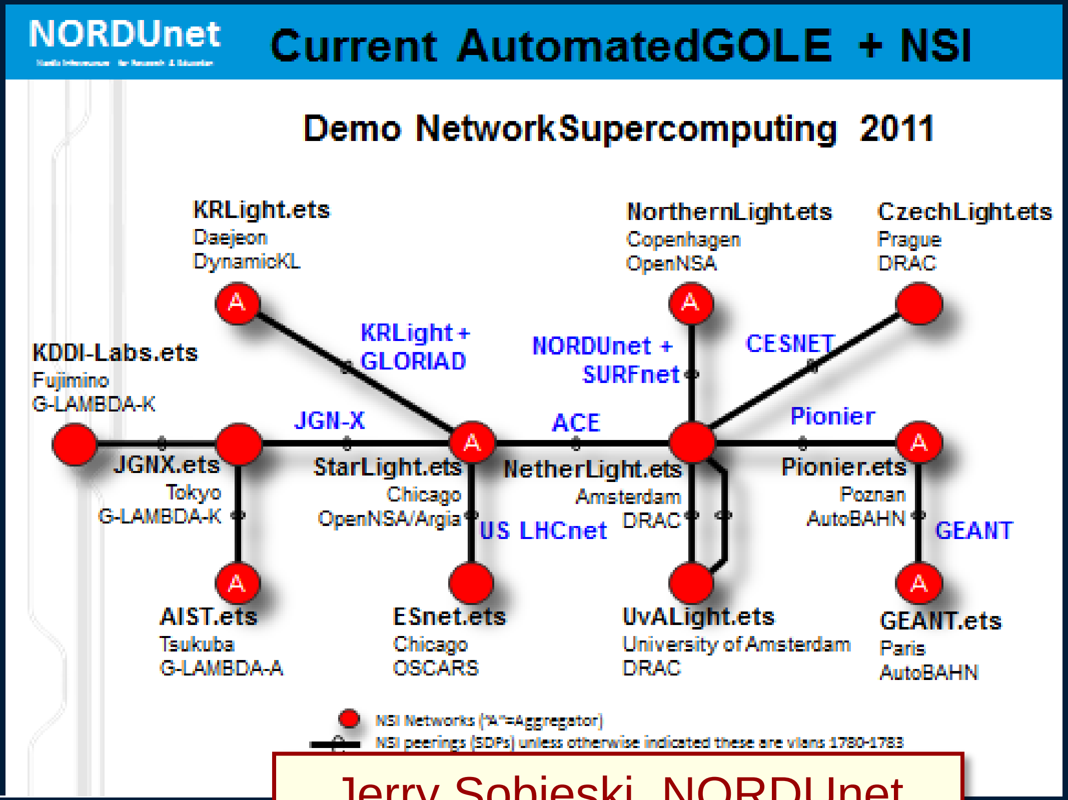
http://lhcone.net

LHC Open Network Environment
The 30’000 ft view



Following Important New Initiatives (Networking sample)

- **LHCONE Goal: Manage LHC's large data flows**
  - Workflow efficiency
  - As site capabilities progress (Nx10G, soon 100G)
  - Avoid triggering DOS alarms (and counter actions)
- **Lightpath technologies**
  - ESnet OSCARS, Internet2 ION, SURFnet/Ciena DRAC…
  - DYNES (reaching end-sites)
  - OGF NSI standards
- **Network virtualization**
  - Data center and WAN
  - Multipoint, multipath topologies
- **Software Defined Networking**
  - OpenFlow, …



Jerry Sobieski, NORDUnet

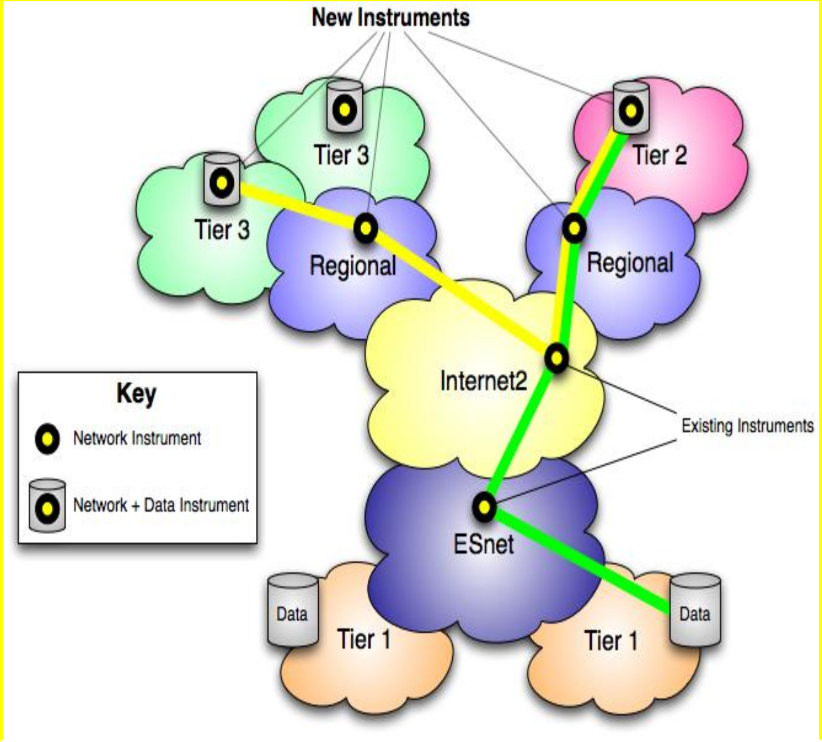
US: DYNES Project, supporting LHC data movement

- NSF funded; Internet2, Caltech, UoMichigan, Vanderbilt
- Nation-wide Cyber-instrument extending hybrid & dynamic capabilities to campus & regional networks
- Provides 2 basic capabilities at campuses and regional networks:

1. Network resource allocation such as bandwidth to ensure transfer performance
2. Monitoring of the network and data transfer performance

- Extending capability existing in backbone networks like ESnet and Internet2
- Tier2 and Tier3 sites need in addition

3. Hardware at the end sites capable of making optimal use of the available network resources



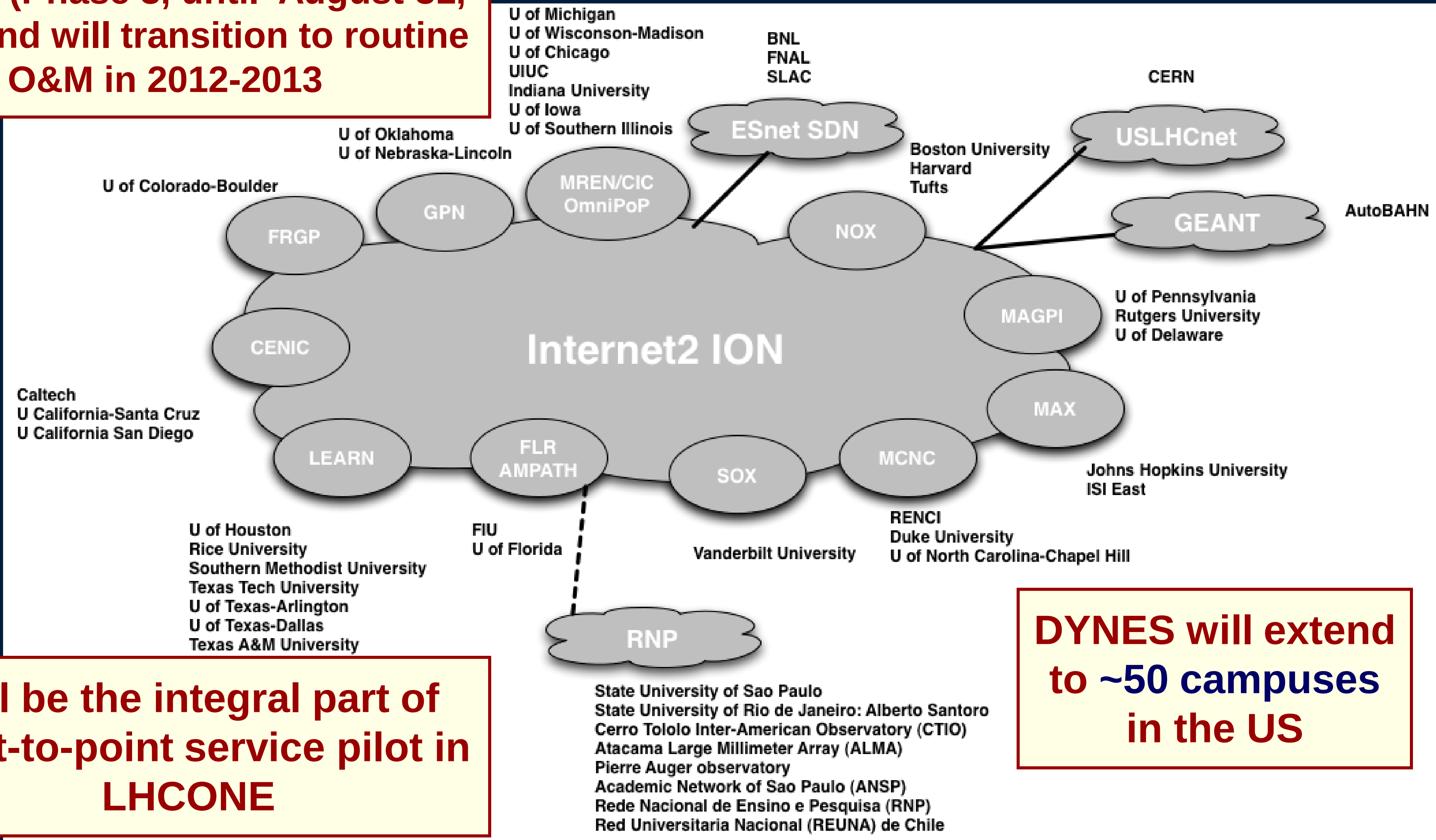
Two typical transfers that DYNES supports: one Tier2 - Tier3 and another Tier1-Tier2.

The clouds represent the network domains involved in such a transfer.

http://internet2.edu/dynes

DYNES Current Logical Topology

DYNES is currently scaling up to full size (Phase 3, until August 31, 2012), and will transition to routine O&M in 2012-2013



Will be the integral part of point-to-point service pilot in LHCONE

DYNES will extend to ~50 campuses in the US

A WORD ON ARCHIVING

Long-term Data Preservation

- **ICFA* Study Group formed in 2009**
  - DPHEP: Study Group for Data Preservation and Long Term Analysis in High Energy Physics
- **Recent end of several experiments**
  - @ LEP, HERA, PEP-II, KEKB, Tevatron
- **There is need to preserve information**

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

3 Levels in 3 areas

- **Recently published report: http://arxiv.org/pdf/1205.4667**

* International Committee on Future Accelerators, http://www.fnal.gov/directorate/icfa/

Summary

- **Excellent performance of the LHC and its Experiments**
  - Producing 10s of Petabytes of new data each year
  - Large statistics are necessary for discovery of new physics
  - Datasets distributed between >140 sites world-wide
  - WLCG is the underlying global, distributed computing infrastructure
- **Data and Computing Models are evolving**
  - More dynamic, more optimized
  - More reliant on network performance
- **Requires new approaches to networking**
  - Intelligent, holistic, systems approach is emerging
    - End-systems, dynamic optical network architectures, monitoring
  - DYNES will extend dynamic bandwidth allocation capability to 50 US campuses, and connect to partner networks abroad
  - LHCONE: virtual multi-domain network for traffic engineering LHC flows

QUESTIONS…

Artur.Barczyk@cern.ch