Networks for the LHC data challenge

SCCTW'2016 - Tbilisi

7 October 2016

edoardo.martelli@cern.ch

CERN

Agenda

LHC data challenge

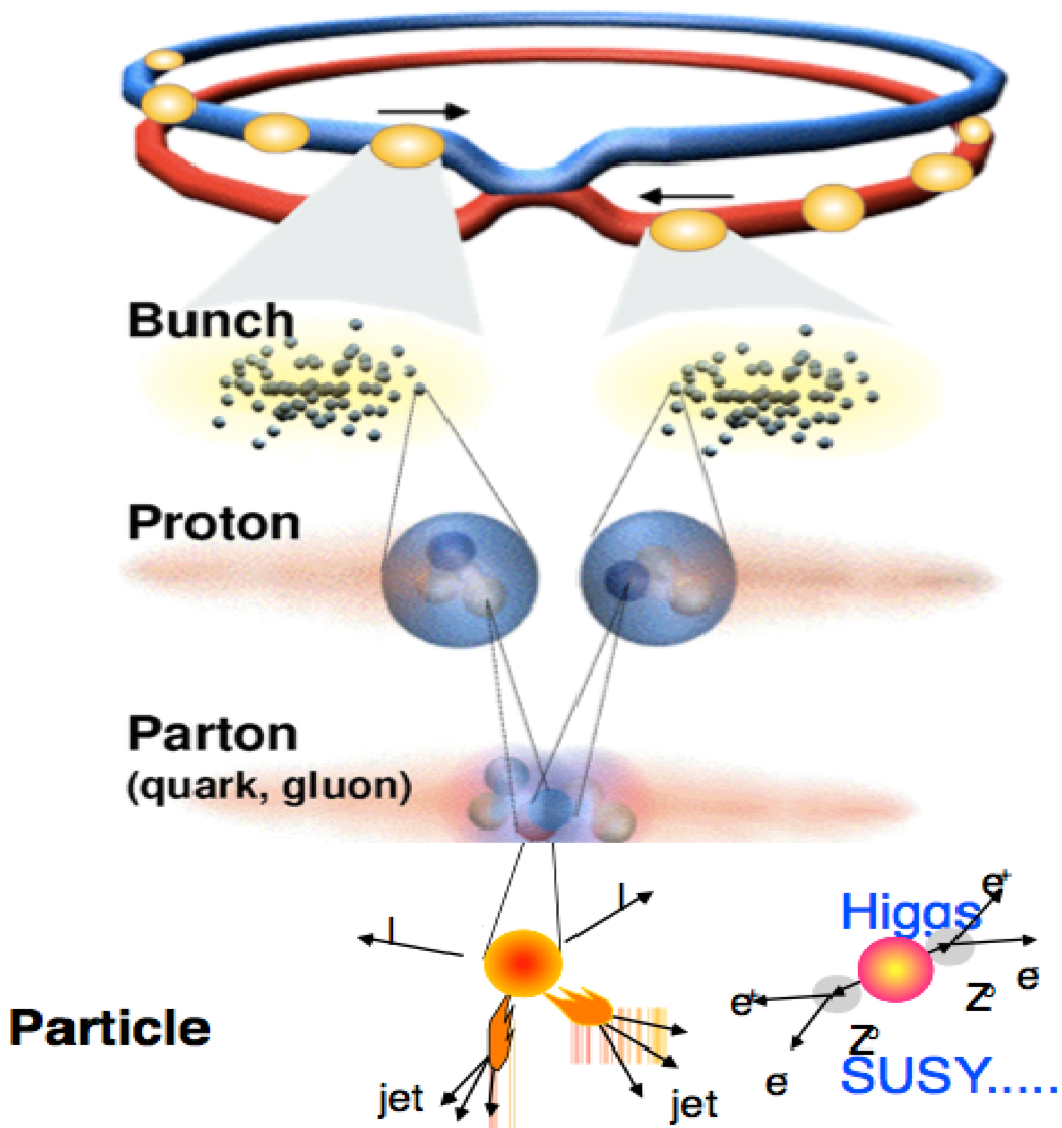
LHC data's journey

- LHC and Experiments
- CERN datacentre
- LHCOPN
- LHCONE

Commercial Cloud Services

LHC data challenge

Collisions in the LHC

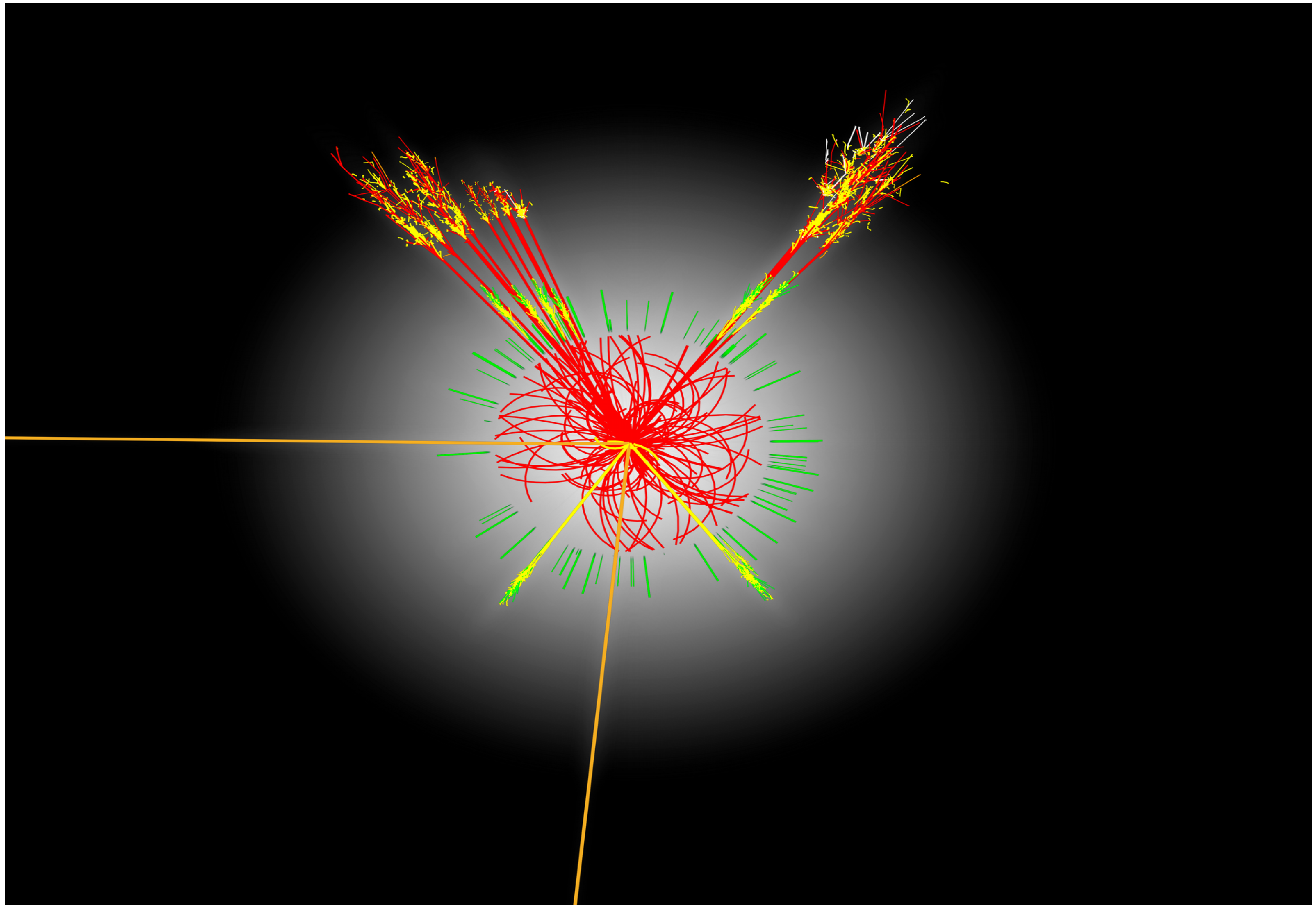


Proton - Proton	2808 bunch/beam
Protons/bunch	10^{11}
Beam energy	7 TeV (7×10^{12} eV)
Luminosity	$10^{34} cm^{-2}s^{-1}$
Crossing rate	40 MHz
Collision rate ≈	10^7-10^9

New physics rate ≈ .00001 Hz

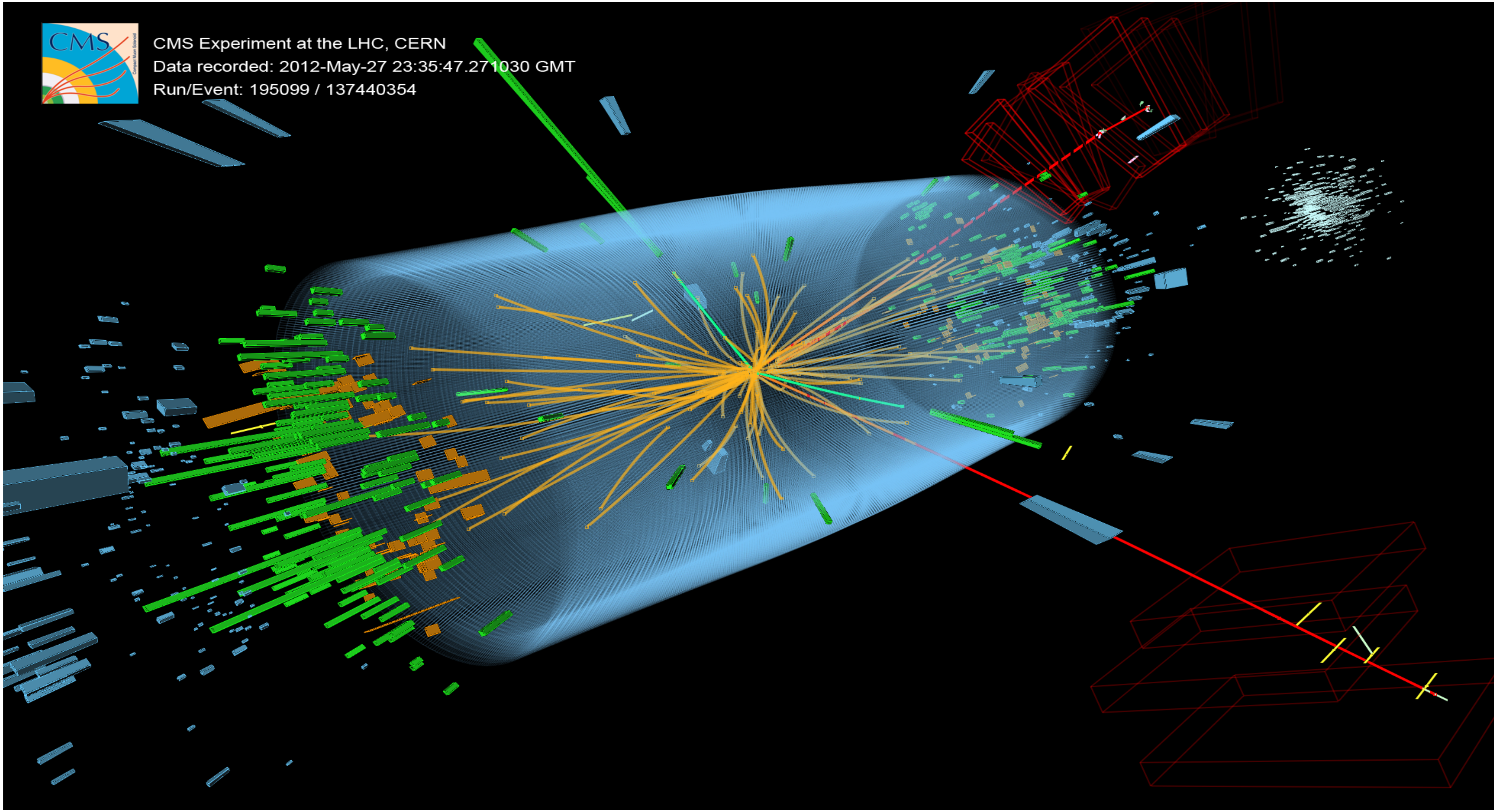
Event selection:
1 in 10,000,000,000,000

Comparing theory...



Simulated production of a Higgs event in ATLAS

.. to real events



Higgs event in CMS

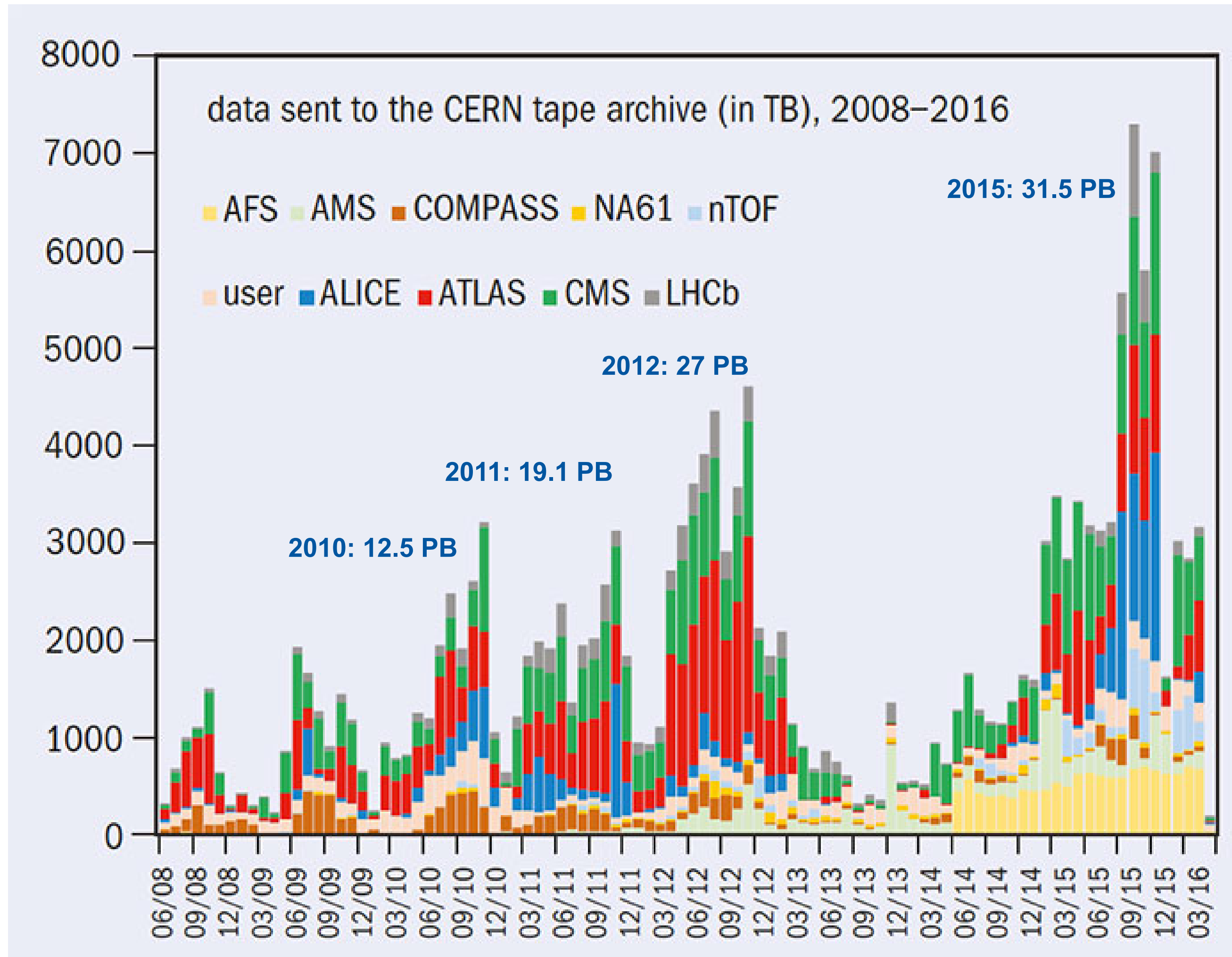
Data Challenge

- 40 million collisions per second

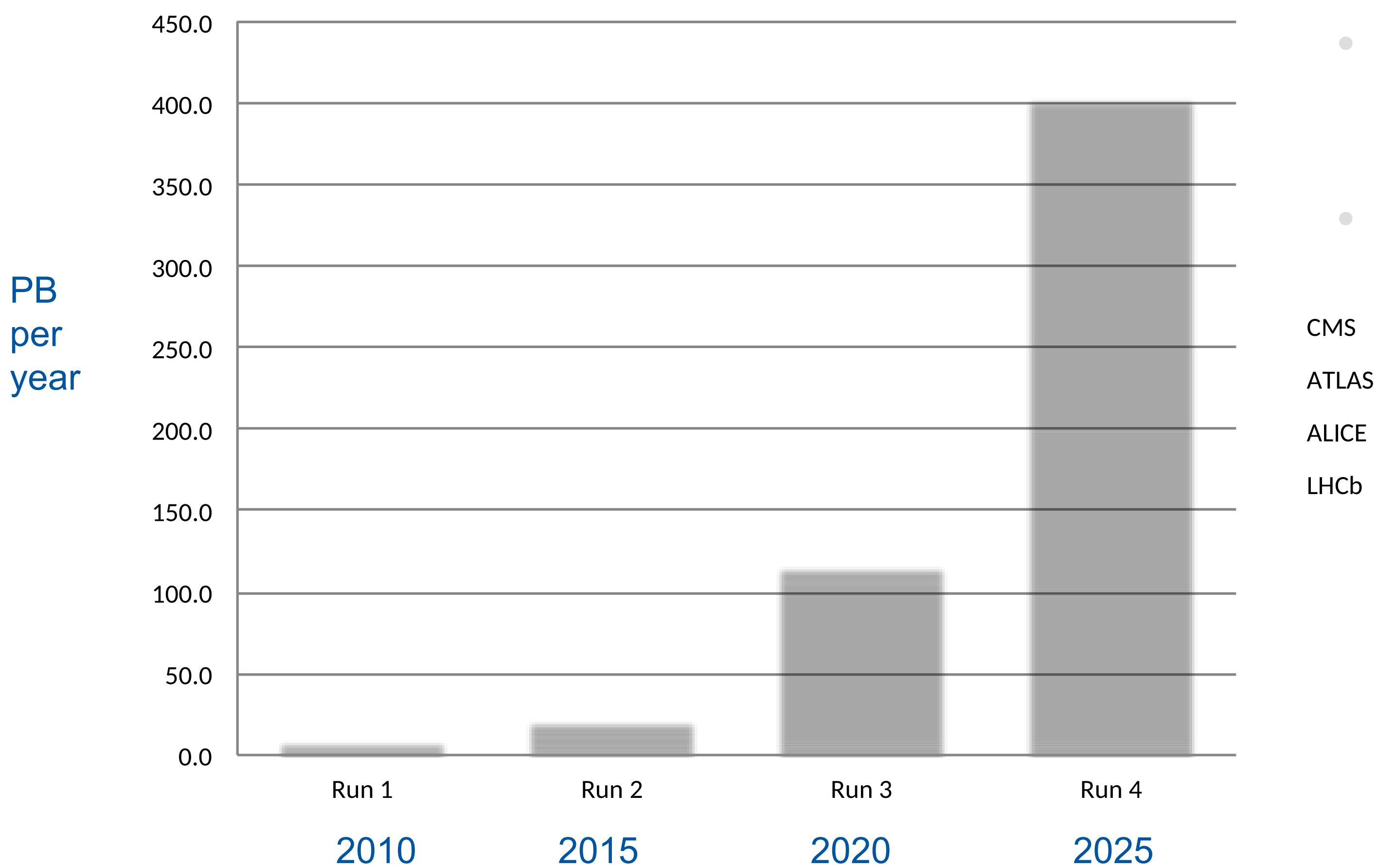
- After filtering, 1000 collisions of interest per second

- 10^{10} collisions recorded each year
  > 25 Petabytes/year of data

Stored data


data sent to the CERN tape archive (in TB), 2008–2016
AFS
AMS
COMPASS
NA61
nTOF
user
ALICE
ATLAS
CMS
LHCb
2010: 12.5 PB
2011: 19.1 PB
2012: 27 PB
2015: 31.5 PB
8000
7000
6000
5000
4000
3000
2000
1000
0
06/08
09/08
12/08
03/09
06/09
09/09
12/09
03/10
06/10
09/10
12/10
03/11
06/11
09/11
12/11
03/12
06/12
09/12
12/12
03/13
06/13
09/13
12/13
03/14
06/14
09/14
12/14
03/15
06/15
09/15
12/15
03/16

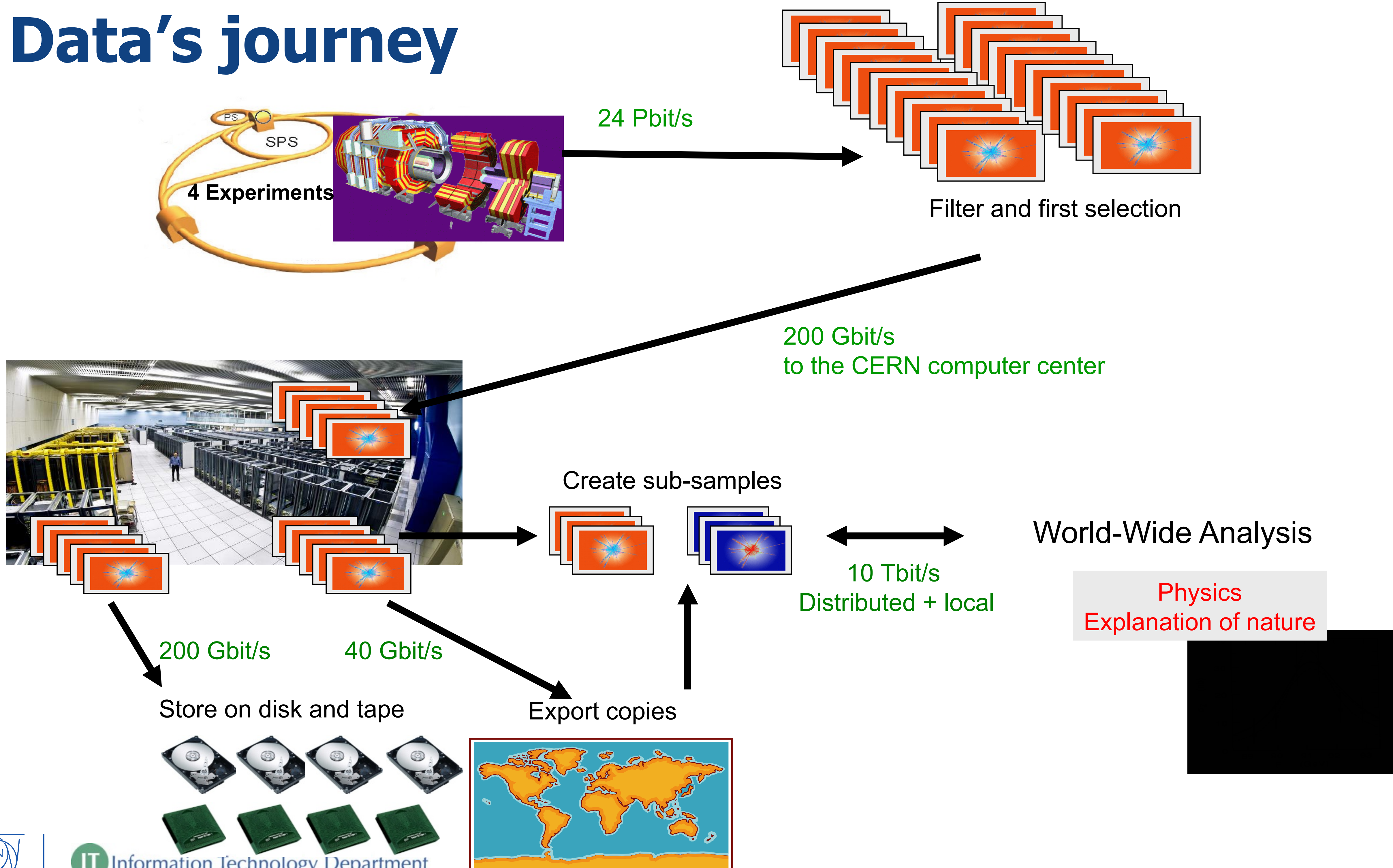
LHC data growth



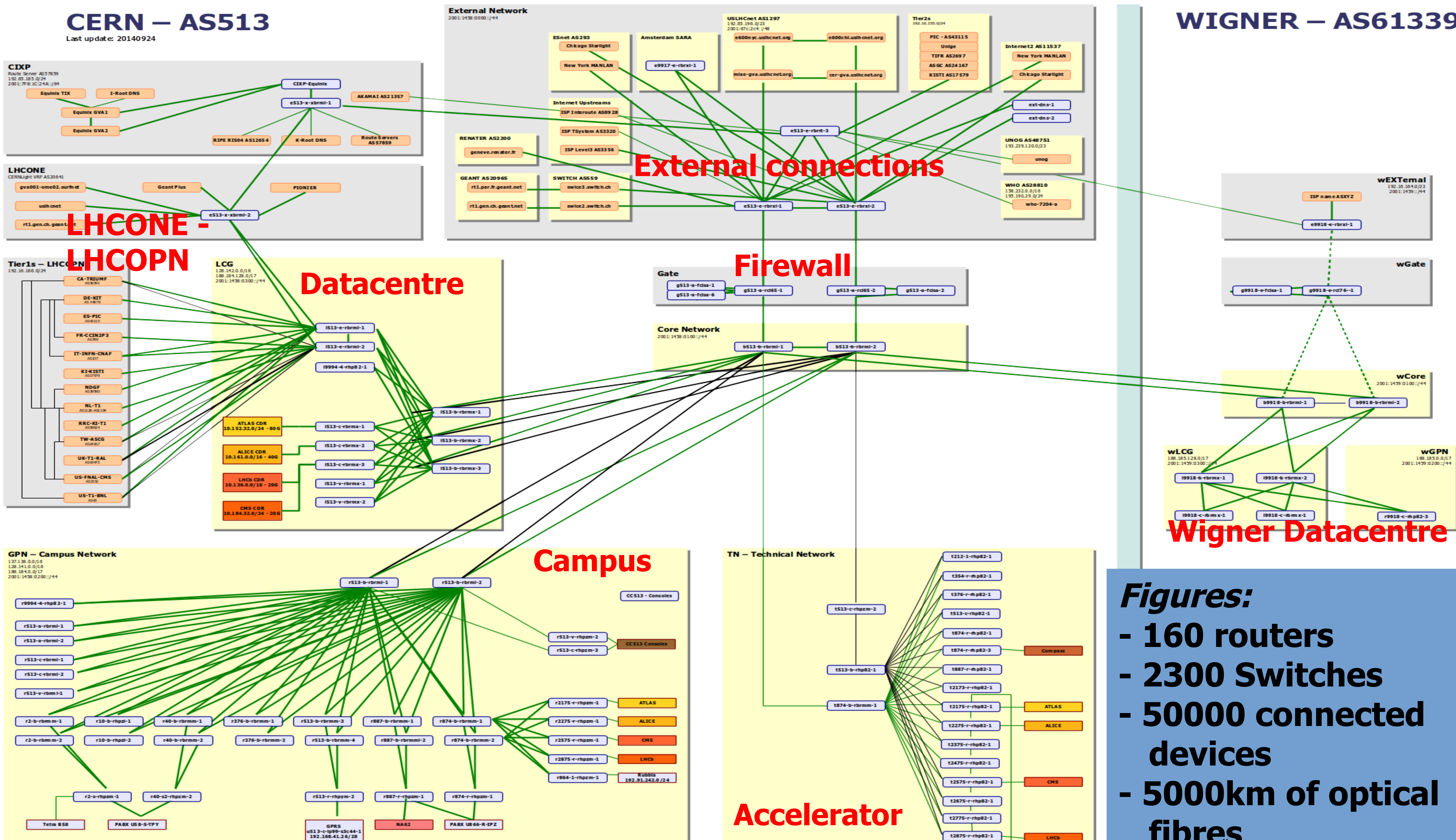
- Expecting to record 400PB/year by 2023
- Compute needs expected to be around 50x current levels, if budget available

LHC data’s journey

Data's journey



CERN Data networks in details

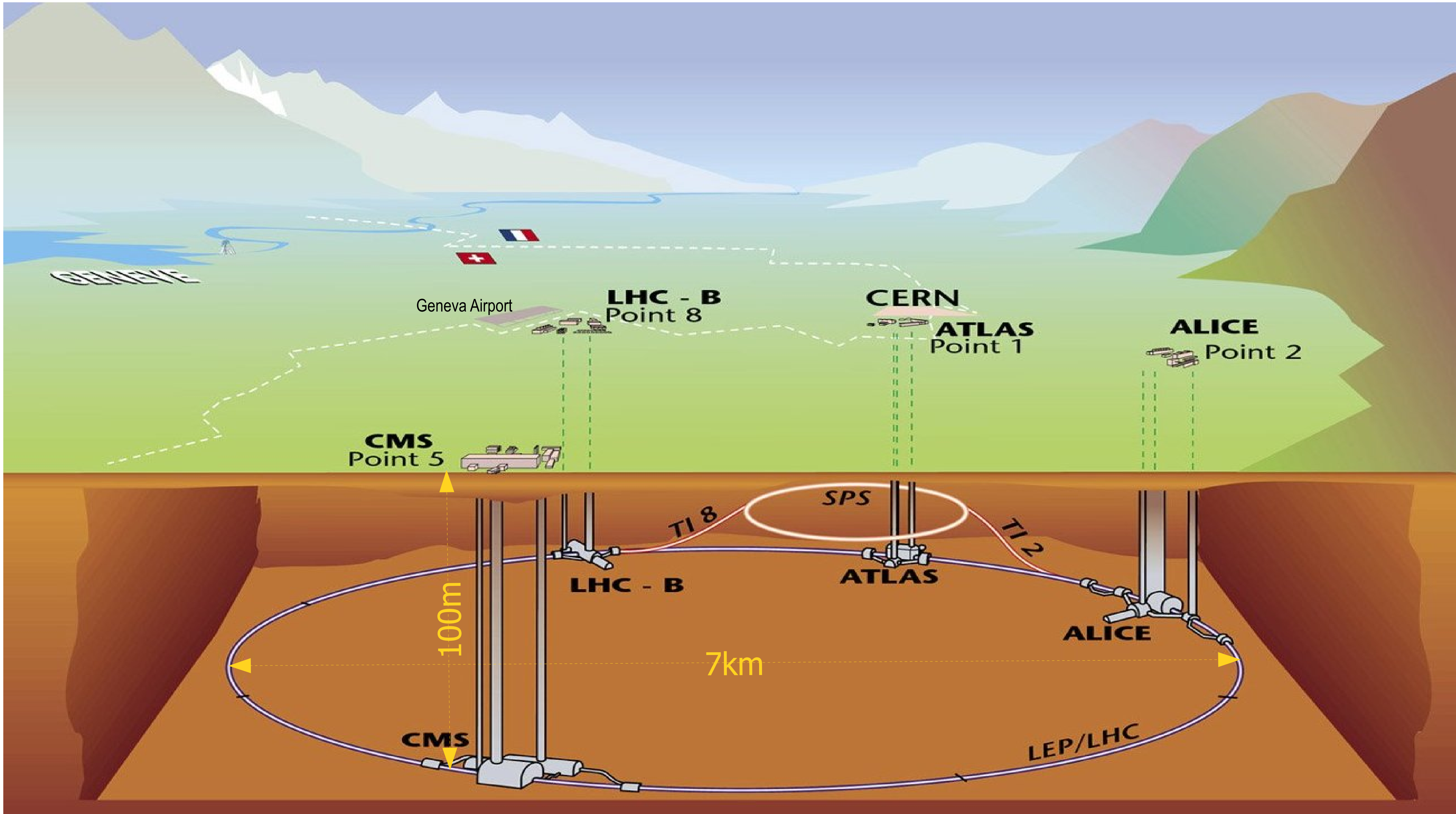


Figures:

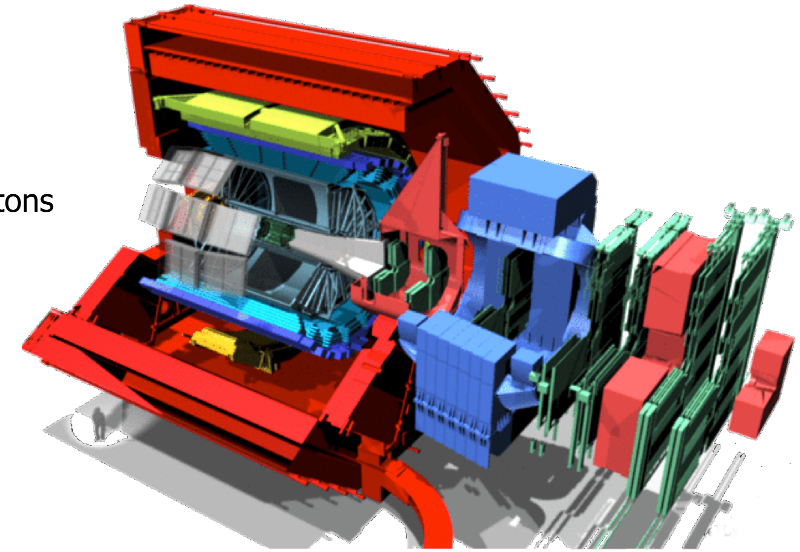
- **160 routers**
- **2300 Switches**
- **50000 connected devices**
- **5000km of optical fibres**

1^{st} stage
Data production at LHC Experiments

The LHC facility

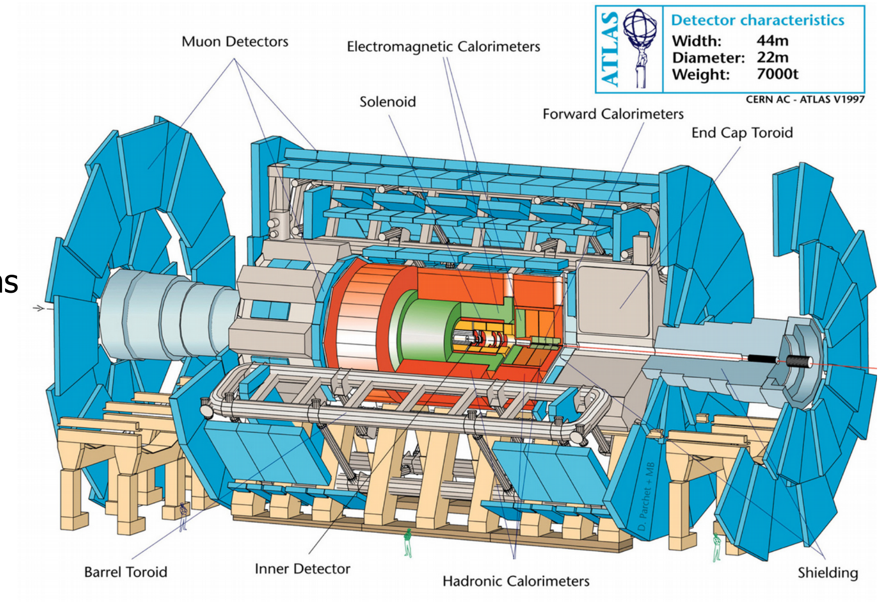


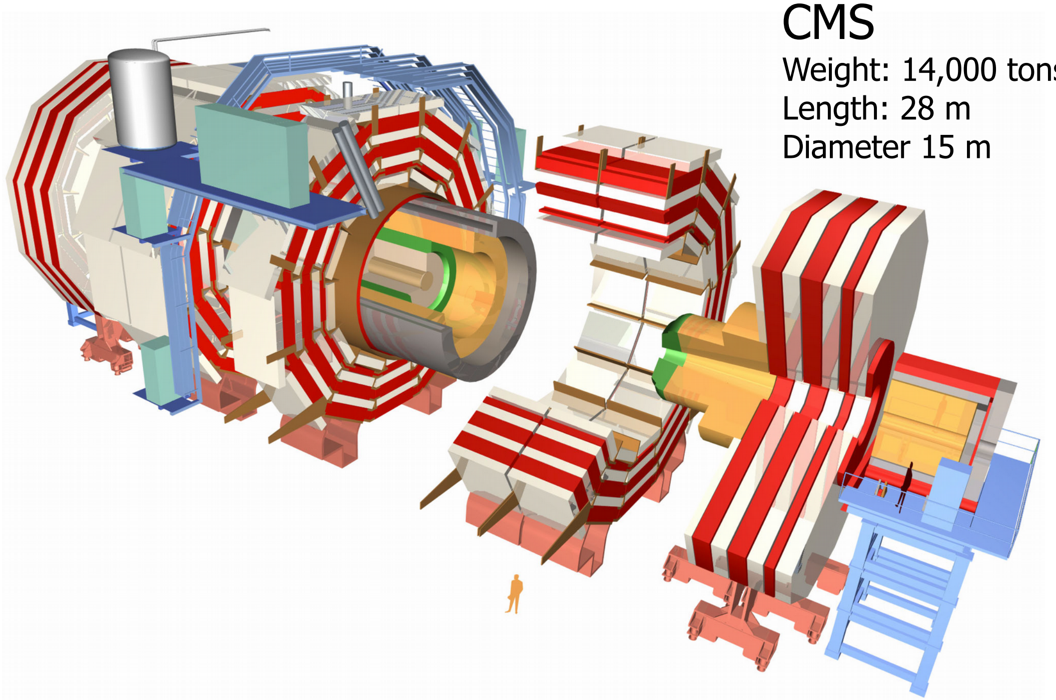
LHC major experiments



ALICE
Weight: 10,000 tons
Length: 26 m
Diameter 16 m

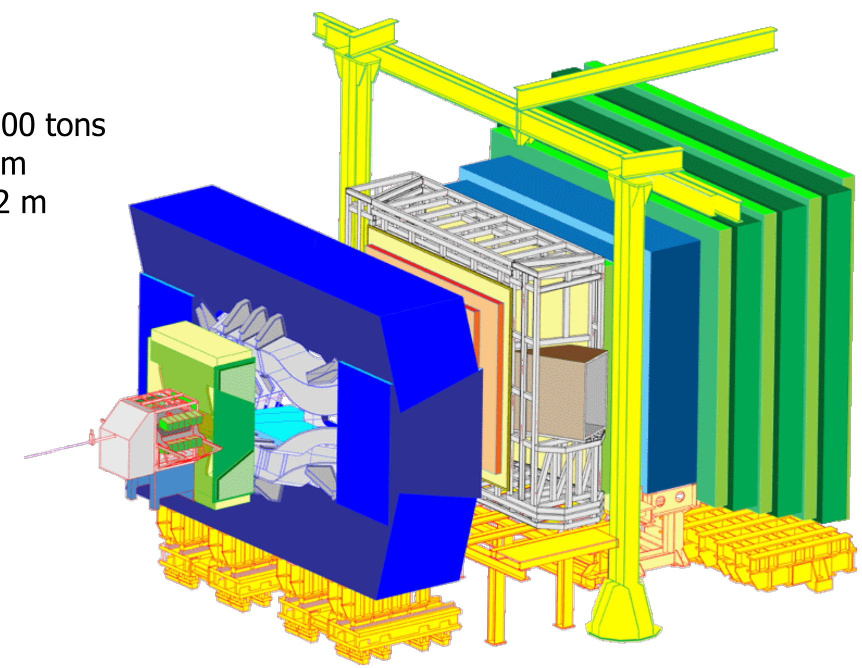
ATLAS
Weight: 7,000 tons
Length: 44 m
Diameter 22 m



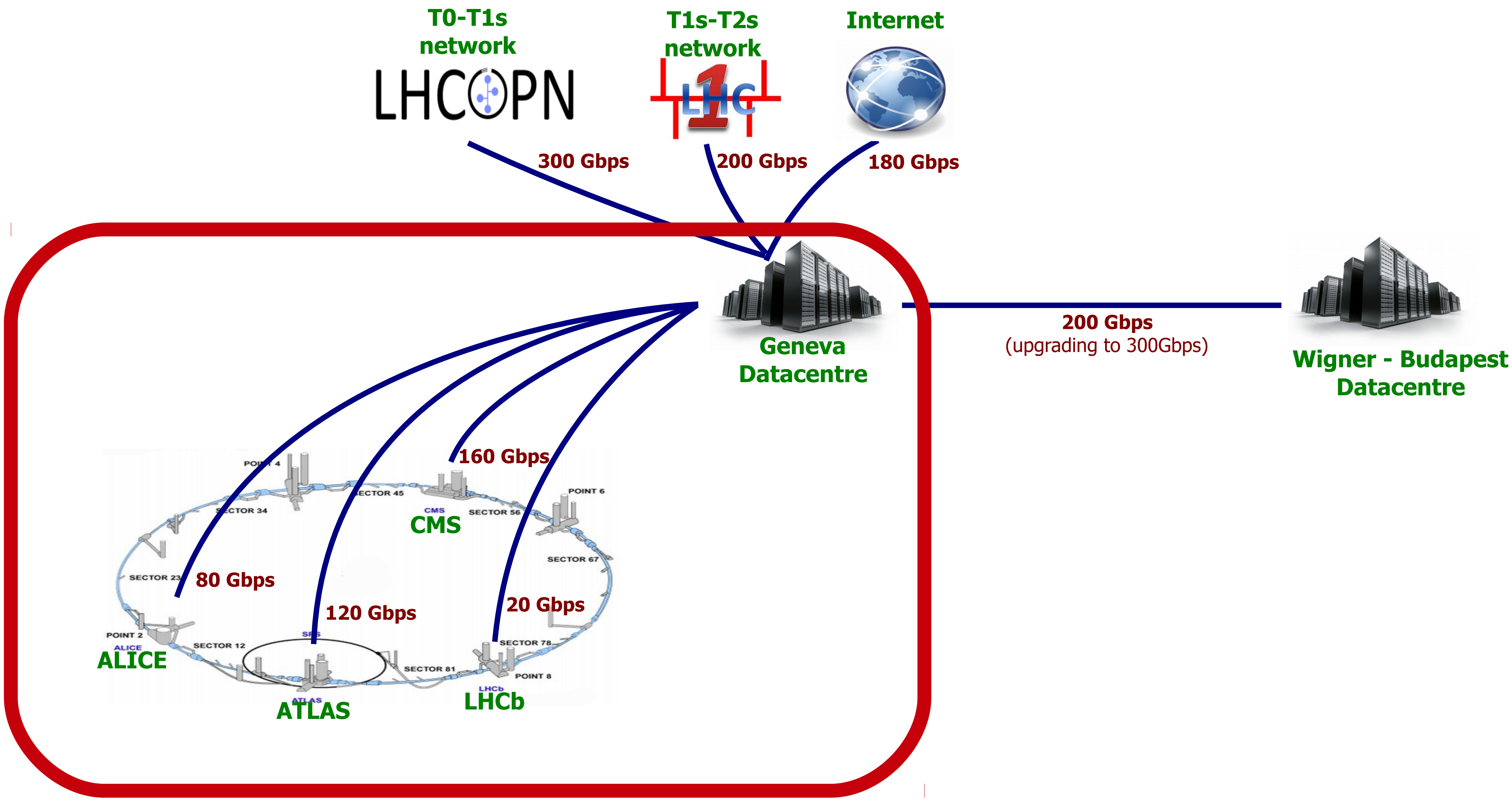


CMS
Weight: 14,000 tons
Length: 28 m
Diameter 15 m

LHCb
Weight: 5,600 tons
Length: 21 m
Diameter 12 m

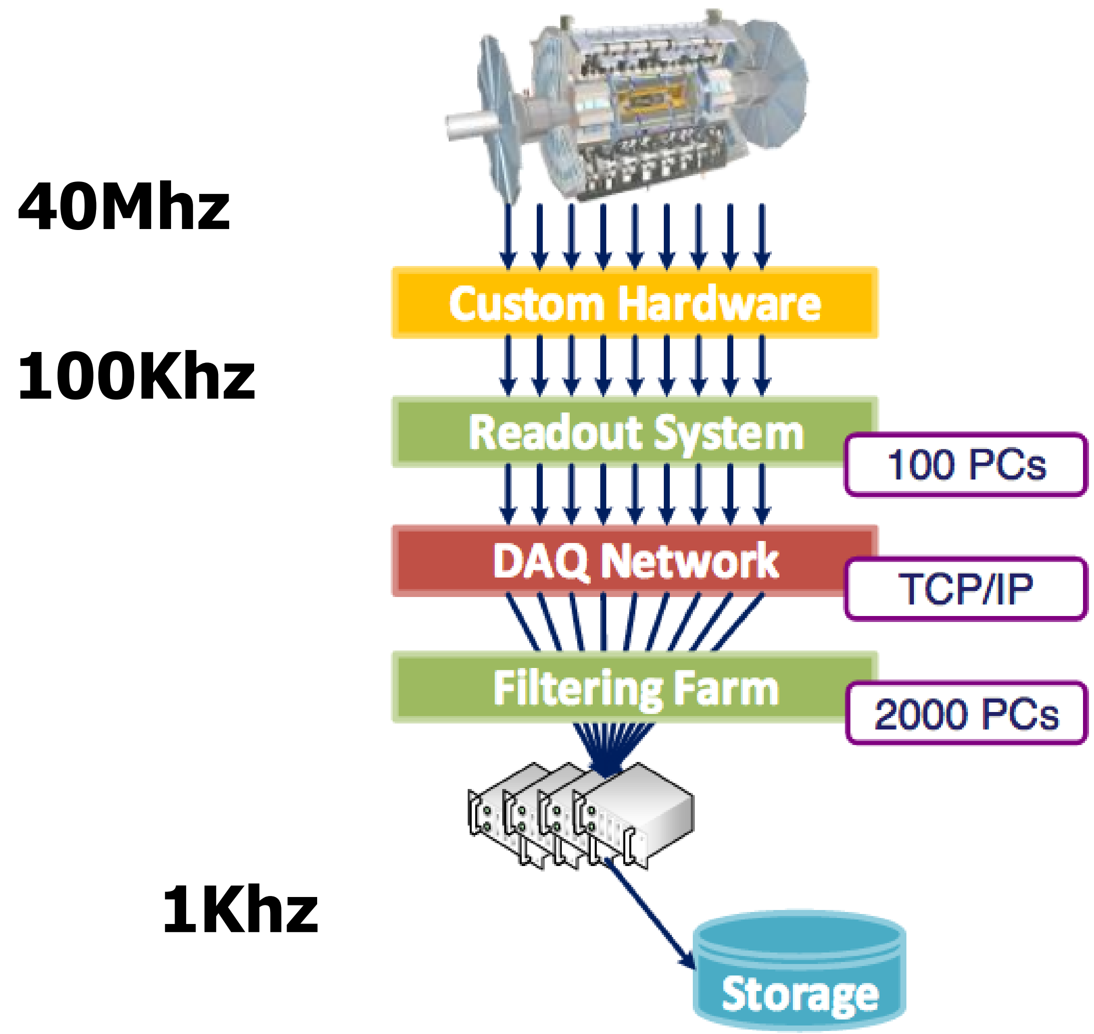


Connectivity for Experiments' data



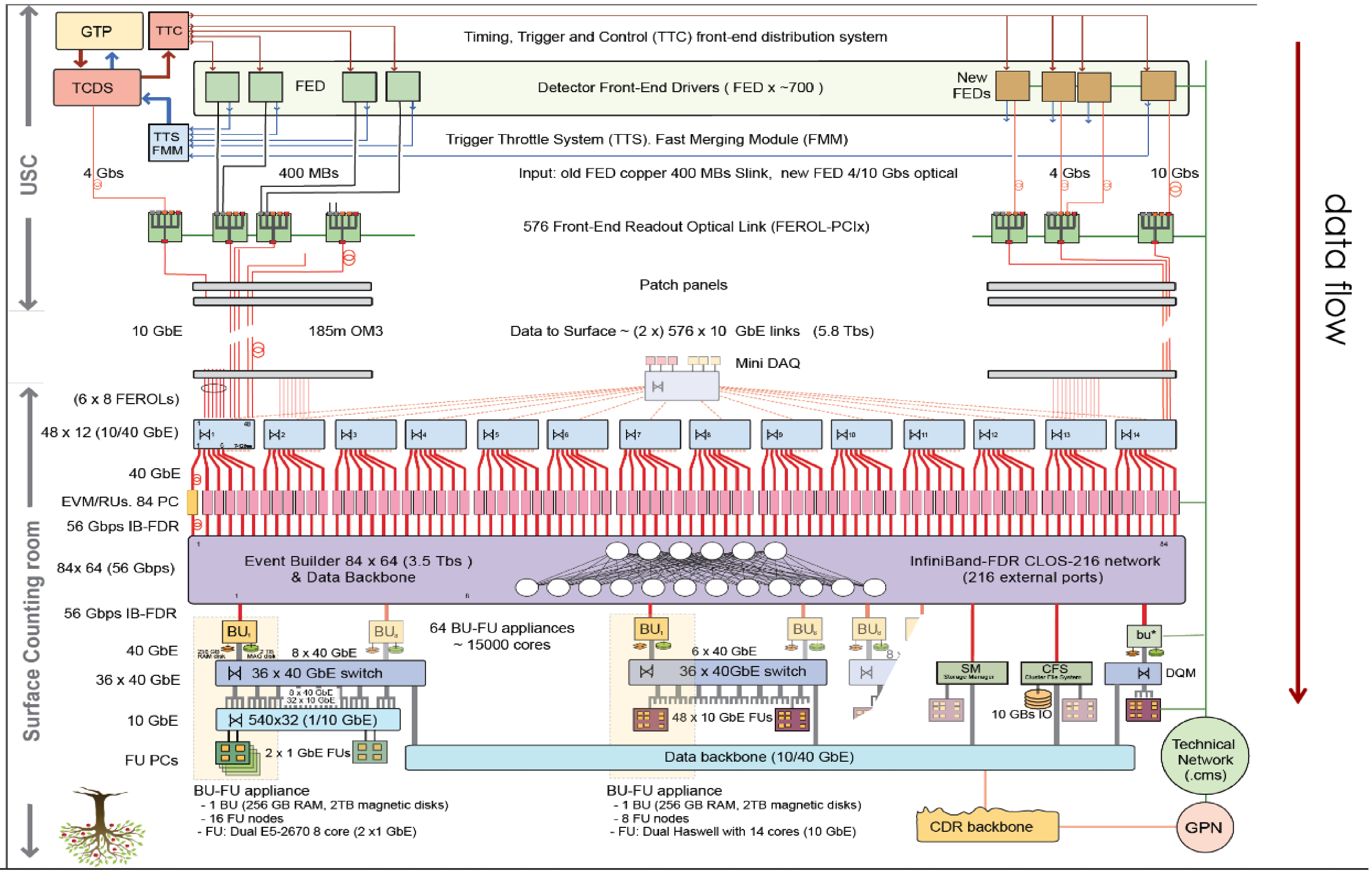
ATLAS DAQ

Data Acquisition System



CMS DAQ

Run II event builder overview



Source http://www.hpcadvisorycouncil.com/events/2015/swiss-workshop/pdf/day3/4_CERN.pdf

Alice

Data acquisition network:
- Increased capacity of the link to IT datacentre to 80Gbps (8x10Gbps)
- Increased buffers size on receiving routers to be able to fill the links

Evaluations in preparation of Run3:
- 40G Ethernet cards of various vendors (Chelsio, Intel, Mellanox) on DAC copper or QSFP fiber links
- 100G server adapters with DAC
- 100G-only switches (e.g. 32 ports 100G in a 1U top-of-rack switch)
- aggregation switches (40G and 100G ports)
- SM/LR 100G optics

ATLAS

Control network of the Data Acquisition System
- Upgraded network equipment to HP8200 chassis and Brocade ICX ToR
- Implemented full redundancy for critical nodes

Data network of the Data Acquisition System
- CDR link capacity increased to 120Gbps (3x40Gbps)
- CDR link redundancy enhanced by using ECMP to two different routers in IT Data Centre
- Upgraded equipment to Brocade MLXe32 chassis and Dell S60 ToR (ultra deep buffers)
- Architecture reviewed: Router Cluster solution implemented with Multi Chassis Trunking

Experiment network (ATLAS Technical and Control Network)
- Upgraded equipment to HP 8200 chassis and HP3500
- Increased security with access control lists

CMS

Hardware upgrades

- HP5400 chassis replaced with Brocade MLXE16
- ToR switches upgrade to Brocade ICX with 48x10Gbps ports and 40Gbps uplinks
- router interconnections will be increased to 40Gbps links
- doubled links to Campus and Technical Network to increase redundancy

Data acquisition network

- CDR connection upgraded to 160Gbps (4x 40Gbps)
- links connected to two different routers to increase redundancy

LHCb

New router to connect Control room to GPN

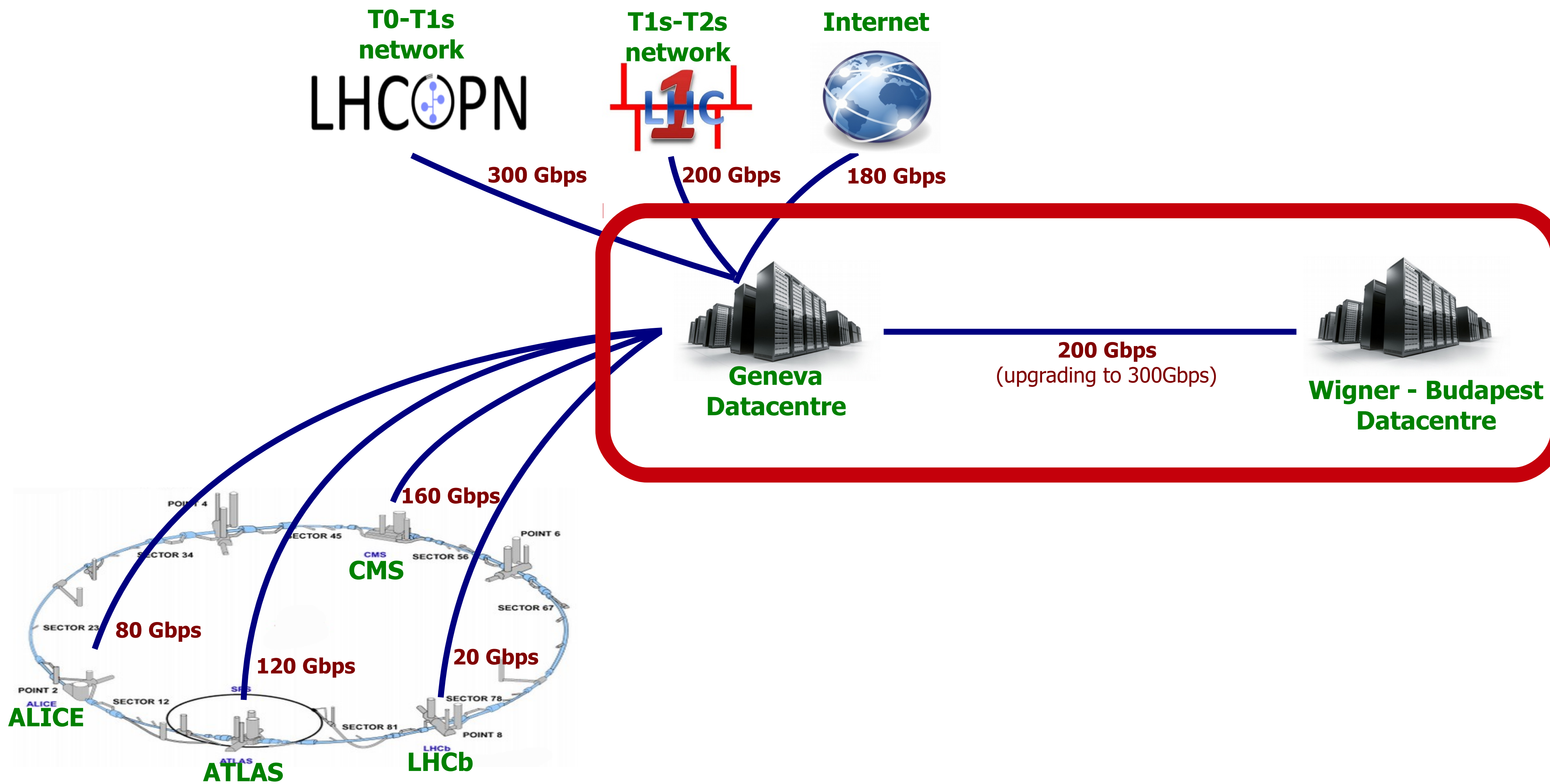
Planning to use IT Network services

Planning and testing for Run3 and Run4

Foreseeing Terabit connection to IT datacentre for Data Acquisition

2nd stage
Storage at CERN Datacentre

Distributed datacentre



2x100Gbps circuits CERN-Wigner

T-Systems	(white line)
GEANT	(red line)

Geneva Datacentre

Wigner Budapest Datacentre

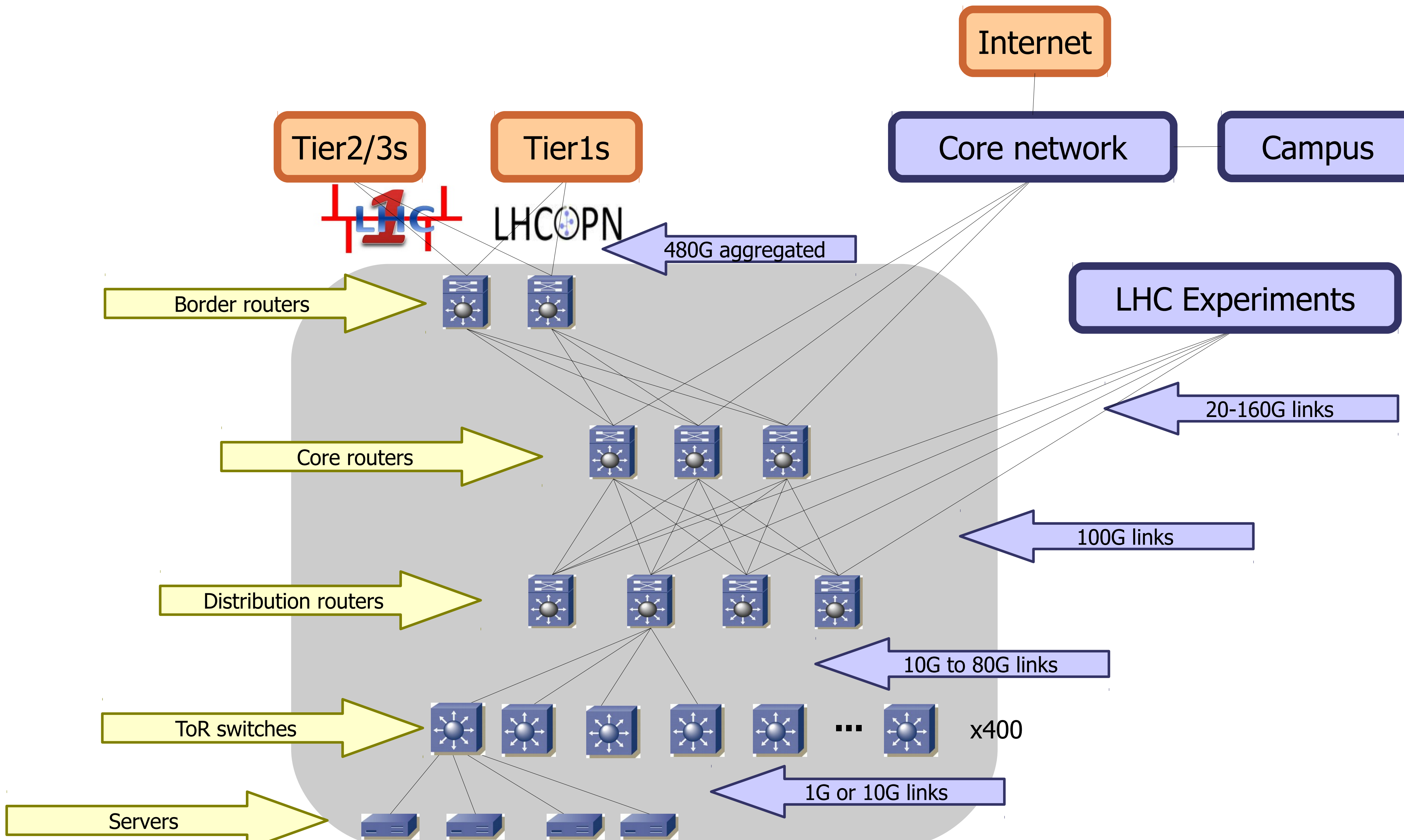
Datacentres facts

MEYRIN DATA CENTRE	last_value
Number of Cores in Meyrin	147,012
Number of Drives in Meyrin	83,660
Number of 10G NIC in Meyrin	8,849
Number of 1G NIC in Meyrin	22,371
Number of Processors in Meyrin	24,743
Number of Servers in Meyrin	13,173
Total Disk Space in Meyrin (TB)	164,184
Total Memory Capacity in Meyrin (TB)	597

WIGNER DATA CENTRE	last_value
Number of Cores in Wigner	56,000
Number of Drives in Wigner	29,698
Number of 10G NIC in Wigner	2,981
Numer of 1G NIC in Wigner	6,579
Number of Processors in Wigner	7,002
Number of Servers in Wigner	3,504
Total Disk Space in Wigner (TB)	97,333
Total Memory Capacity in Wigner (TB)	221

On 26-09-2016. Source: https://meter.cern.ch/public/_plugin/kibana/#/dashboard/elasticsearch/Overview:%20Data%20Centre

Datacentre network architecture



Servers’ connectivity

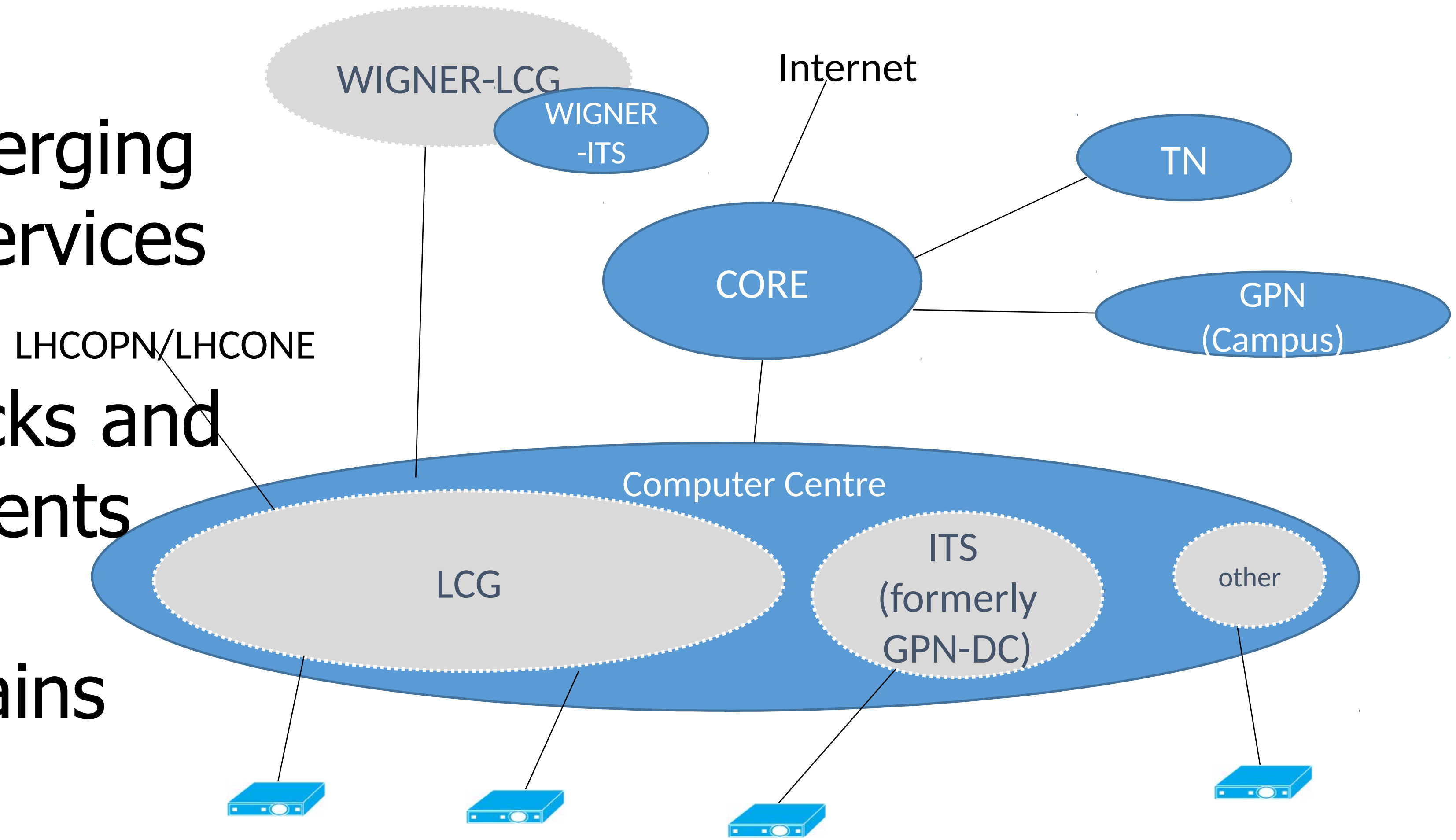
- 1G-10G connection for CPU servers
- 10G to 80G connection for disk servers
- EVPLS for VMs live migration
- Virtual routing instances (VRFs) for policy domains
- High speed access to Tier1/2/3 datacentres

Agile datacentre

- Developed solution to allow transparent VM migration based on VPLS

- Unified physical infrastructure: merging Physics and IT Services datacentres to remove bottlenecks and simplify deployments

- Multi virtual domains



Datacentre Tool Chain

puppet labs

elasticsearch

git

openstack CLOUD SOFTWARE

ceph

hadoop

RDO

kibana

FOREMAN

RUNDECK

Jenkins

Openstack Status

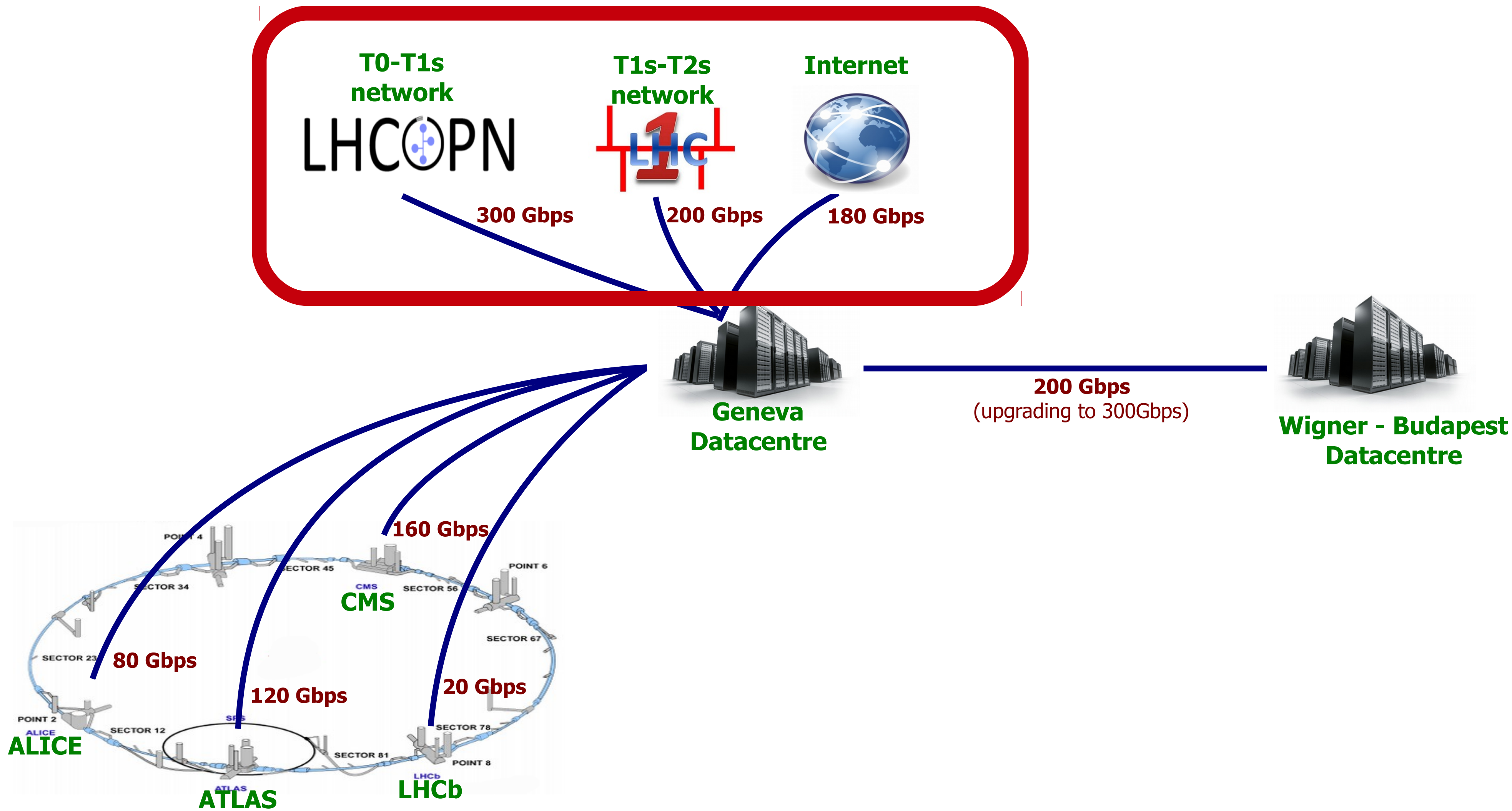
4 OpenStack clouds at CERN
- Largest is ~124,000 cores in ~5,000 servers
- 3 other instances with 45,000 cores total

Active contributor of the Openstack open-source software

Collaborating with companies at every 6 month open design summits

3rd stage
Distribution and Analyses

Data distribution



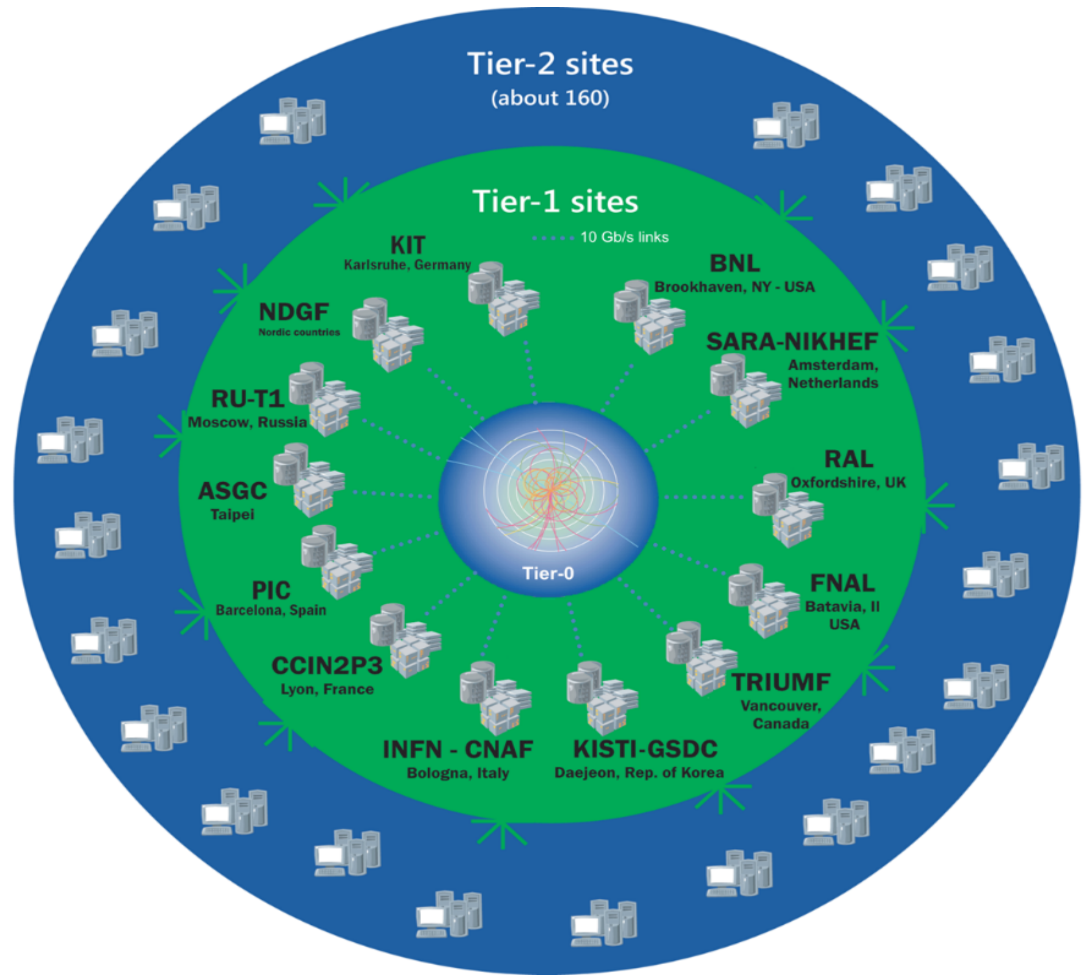
WLCG

The Worldwide LHC Computing Grid:
distributed computing infrastructure for LHC data analysis

Tier-0 (CERN): data recording, reconstruction and distribution

**Tier-1:
permanent storage,
re-processing,
analysis**

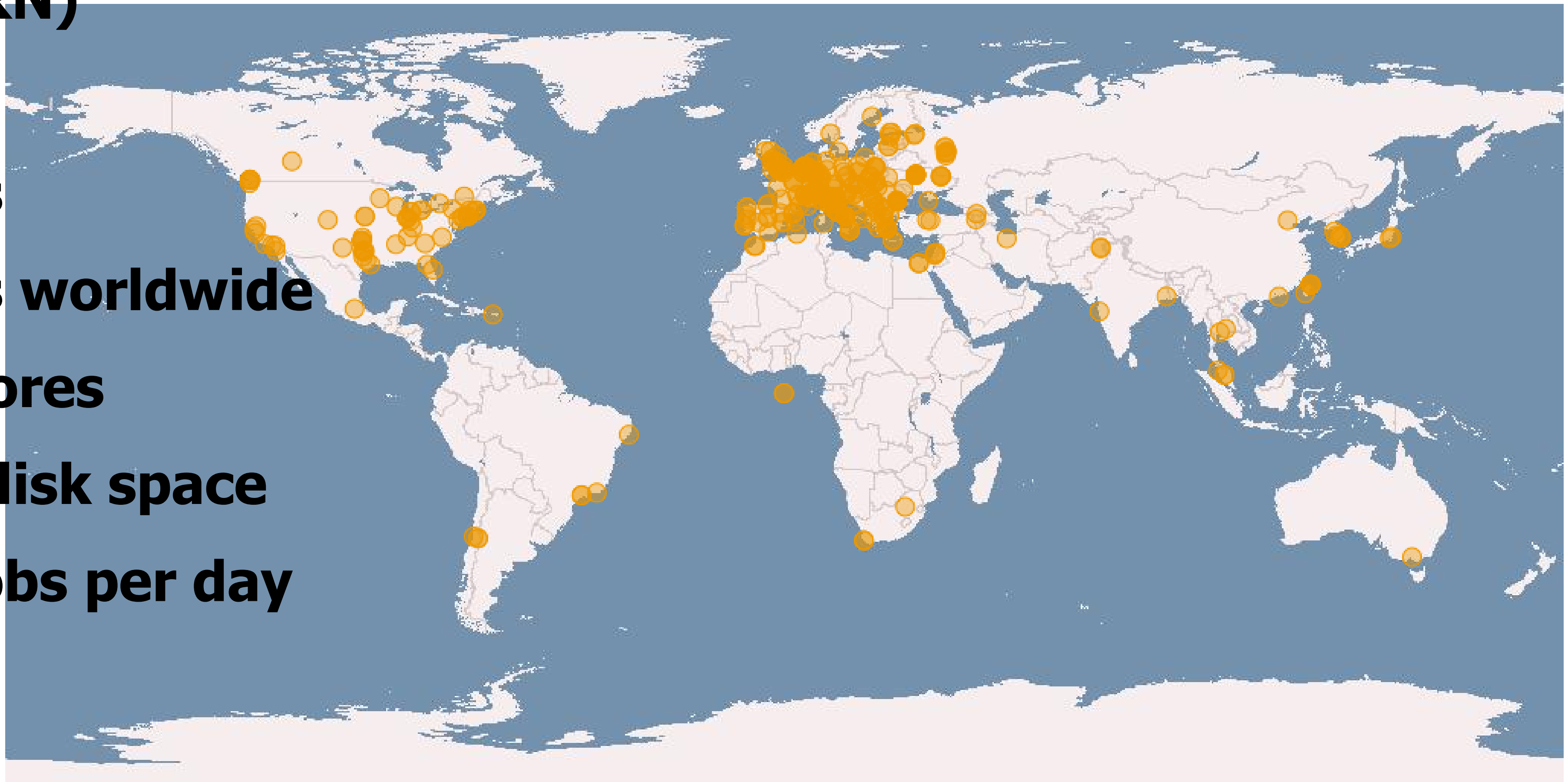
**Tier-2:
Simulation,
end-user analysis**



WLCG resources

WLCG resources:

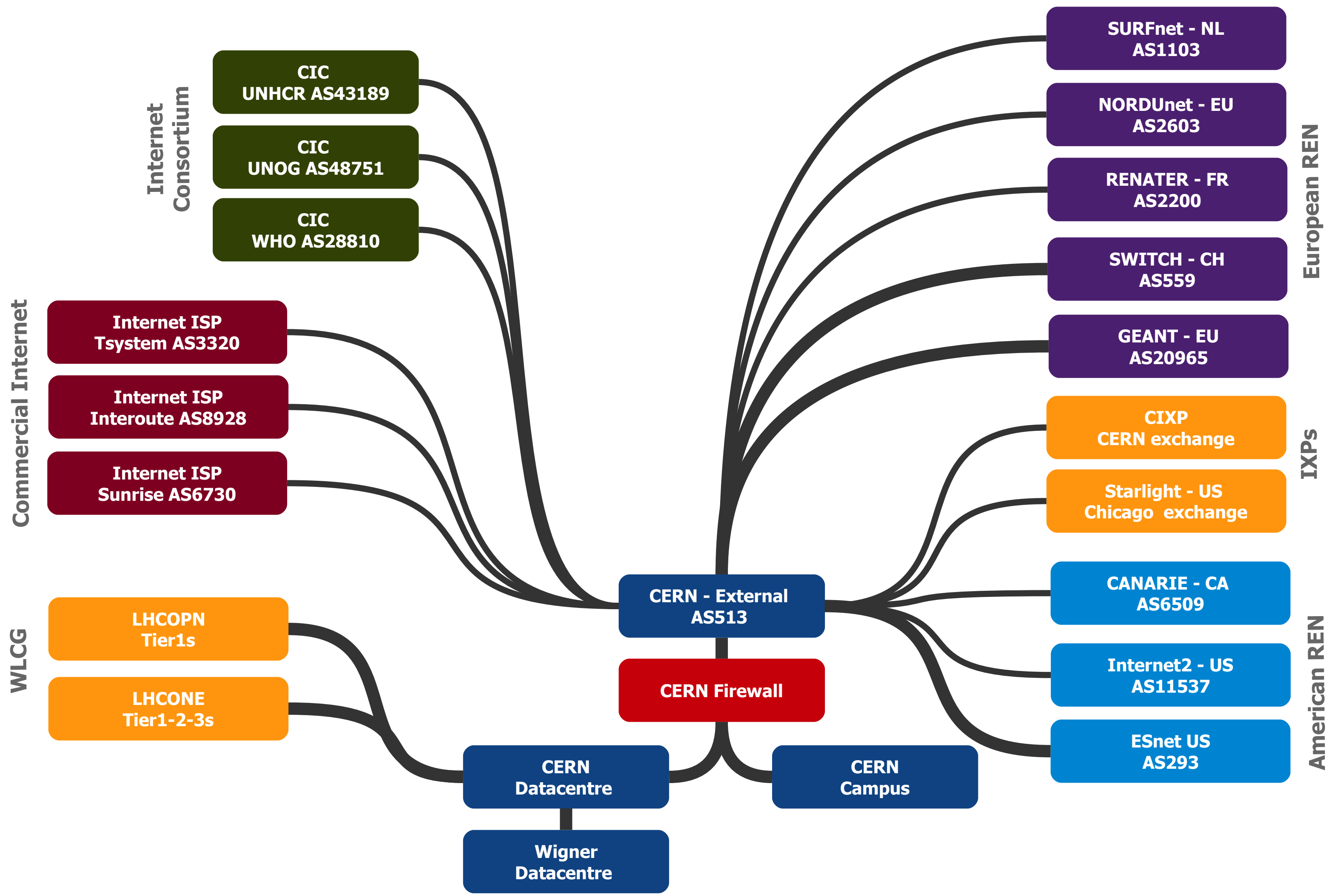
- 1 Tier0 (CERN)
- 13 Tier1s
- ~170 Tier2s
- >300 Tier3s worldwide
- ~350,000 cores
- ~500PB of disk space
- 2 millions jobs per day



Connectivity to remote sites

External Network

CERN External Network

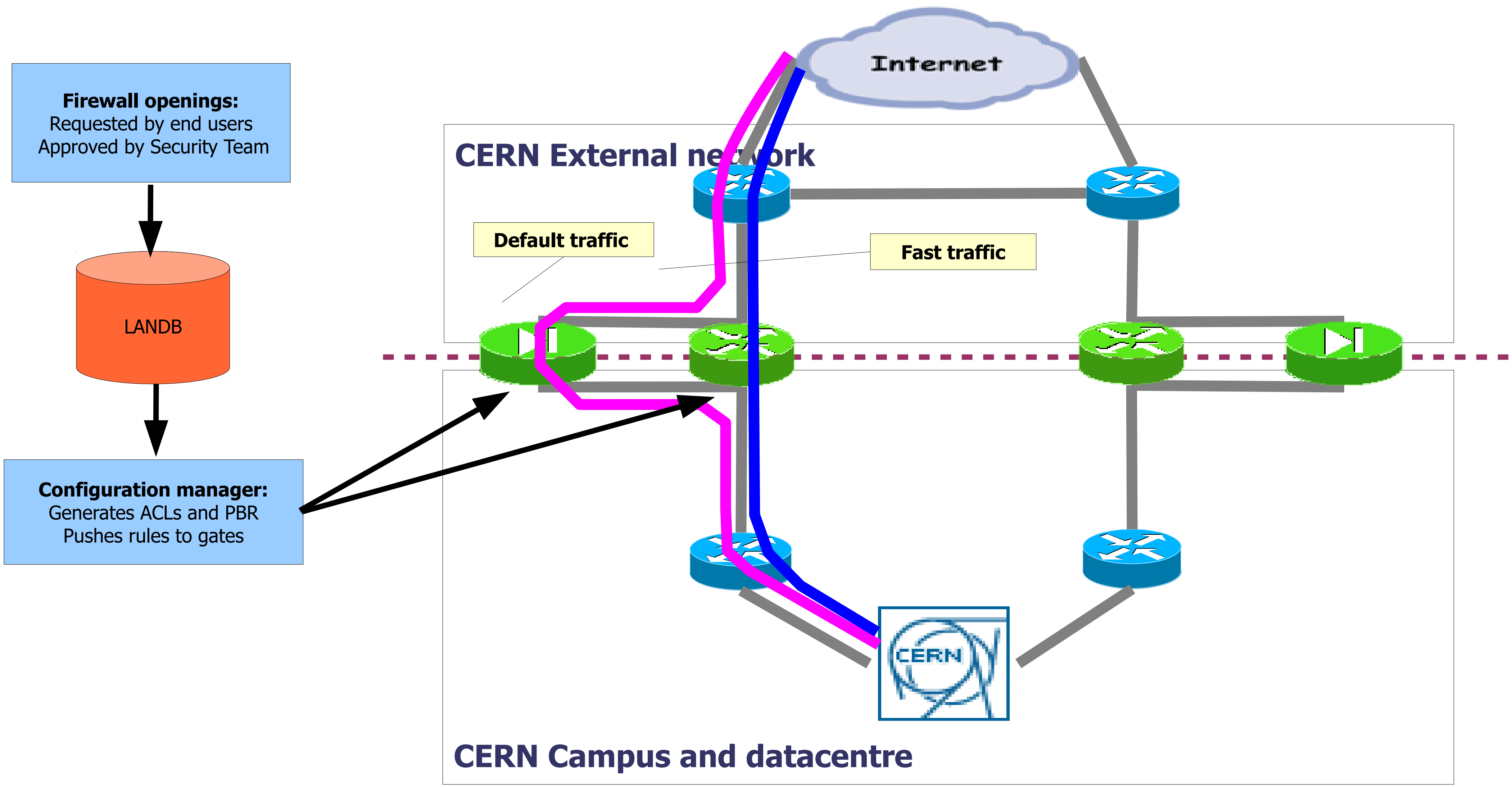


SDN Firewall

Firewall rules stored in network database

Routers and Firewall configuration updated very 15 minutes

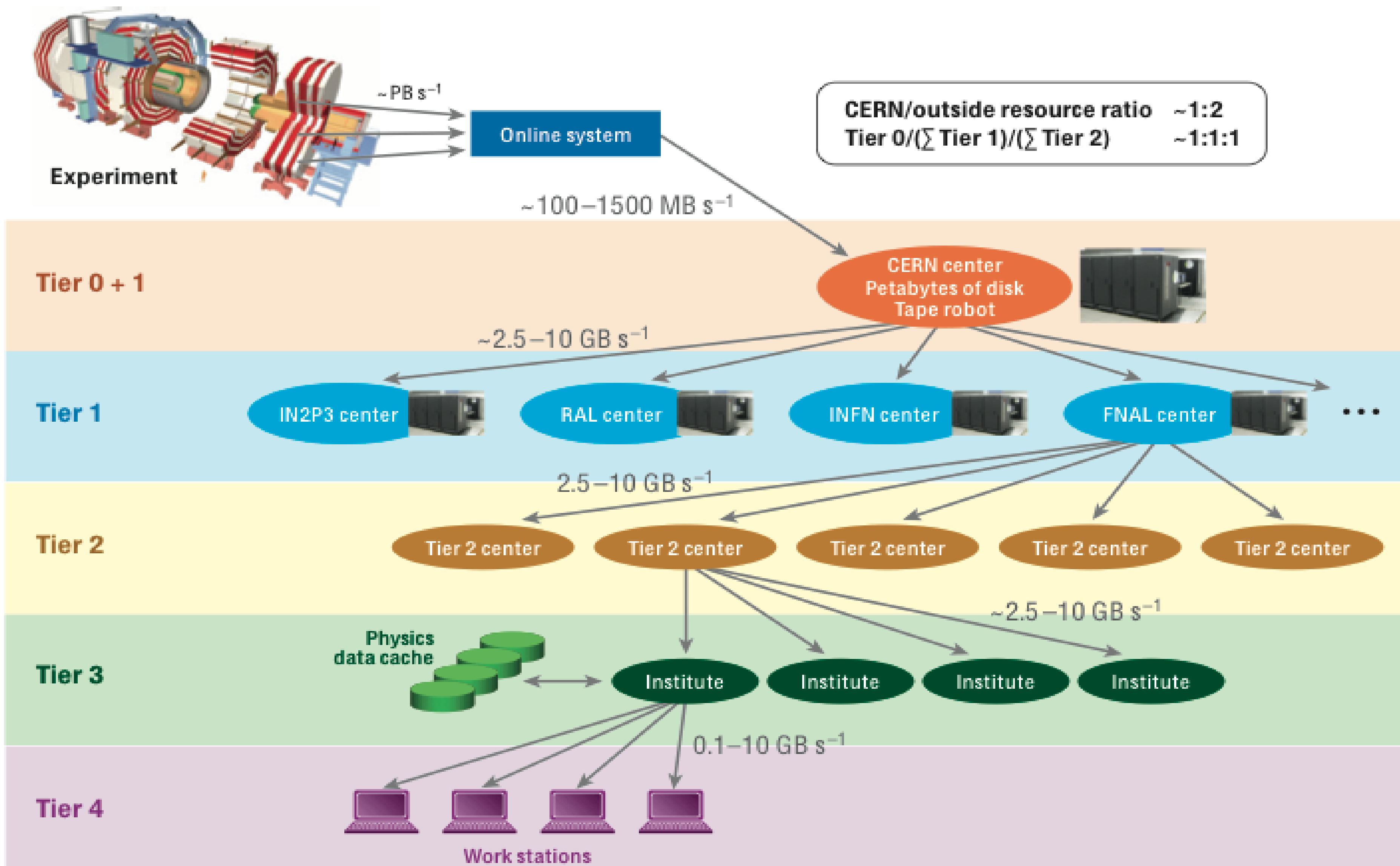
Statefull firewall bypass for large, well known flows



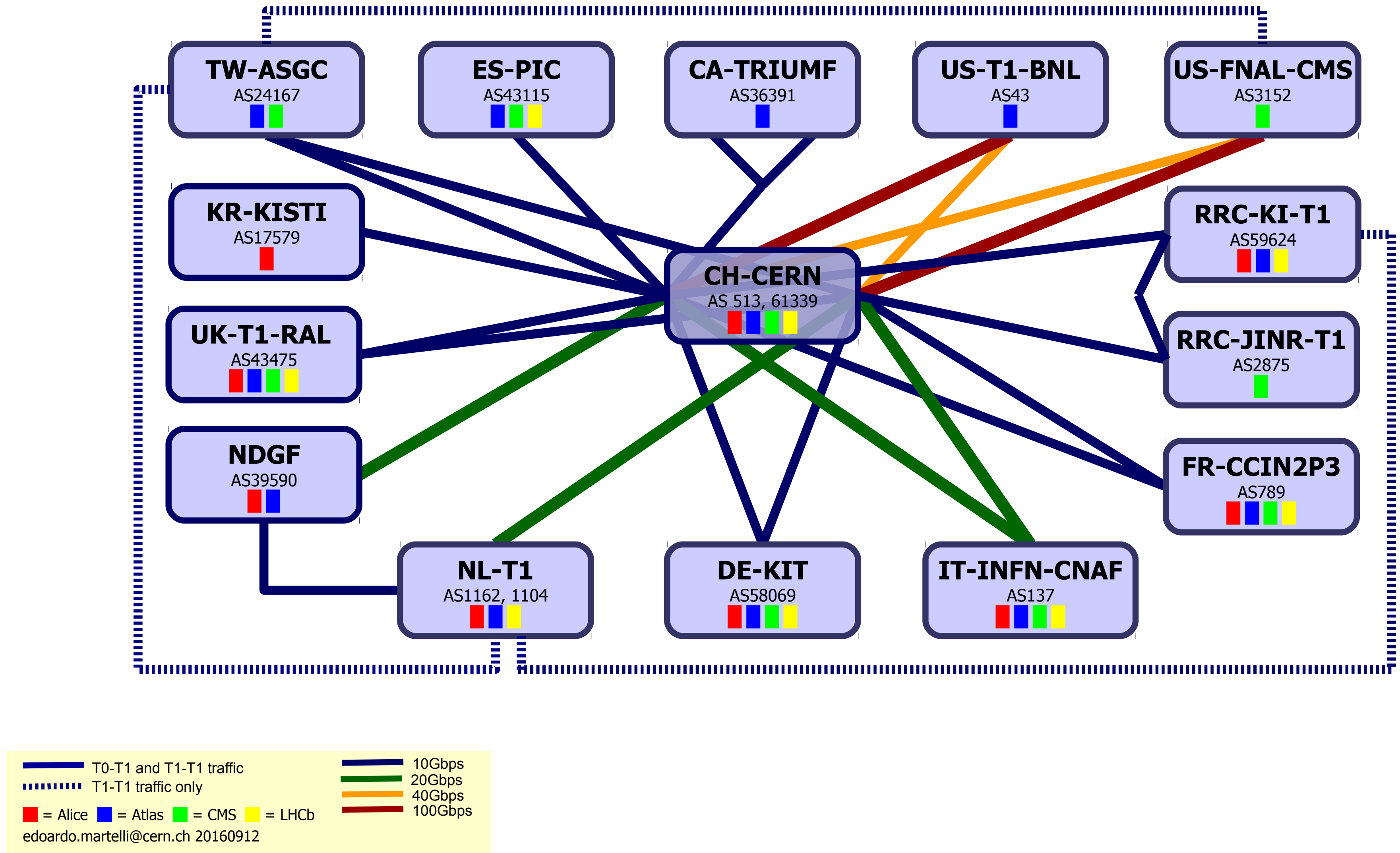
Storage and Analyses at Tier1s

LHCOPN: Tier0-Tier1s network

Original Computing Model



LHCOPN topology



LHCOPN

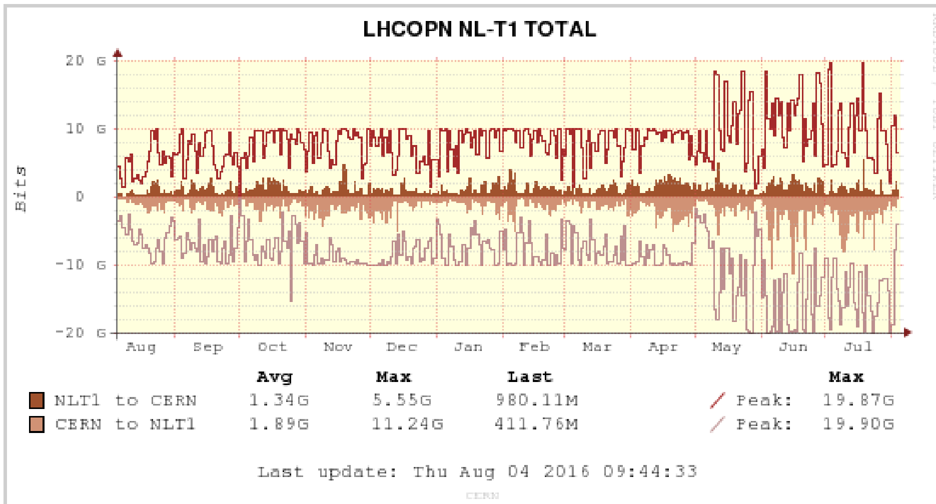
Private network connecting Tier0 and Tier1s

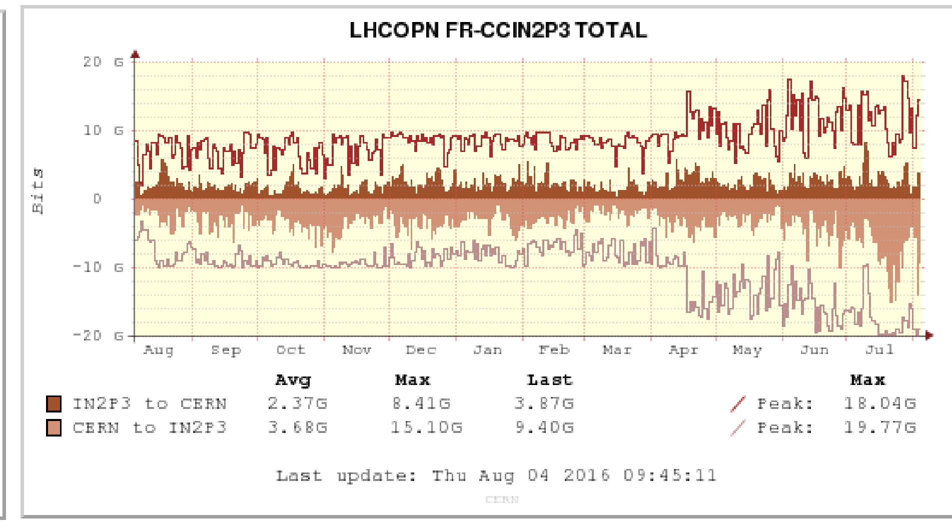
- Reserved to LHC data transfers and analysis
- Single and bundled long distance 10G and 100G Ethernet links
- Star topology
- BGP routing: communities for traffic engineering, load balancing.
- Security: only declared IP prefixes can exchange traffic.

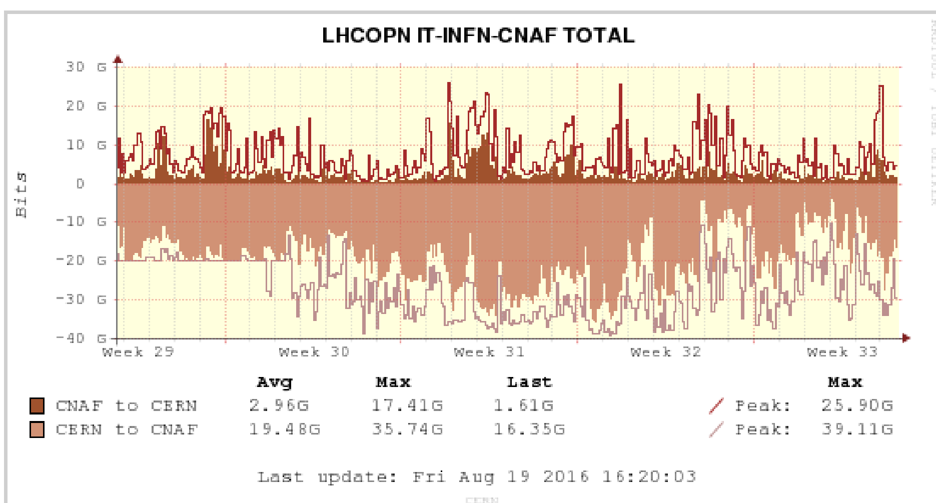
Latest developments

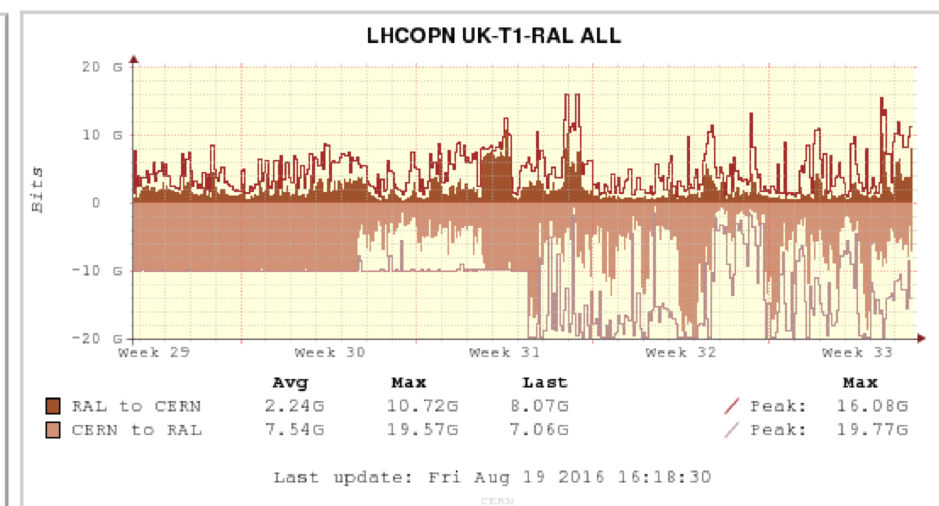
- Traffic volume has already grown 100% from the beginning of Run2
- 5 Tier1s have doubled their link capacity in the last months
- 11 sites (over 13) now connected also with IPv6

Upgraded links' utilization

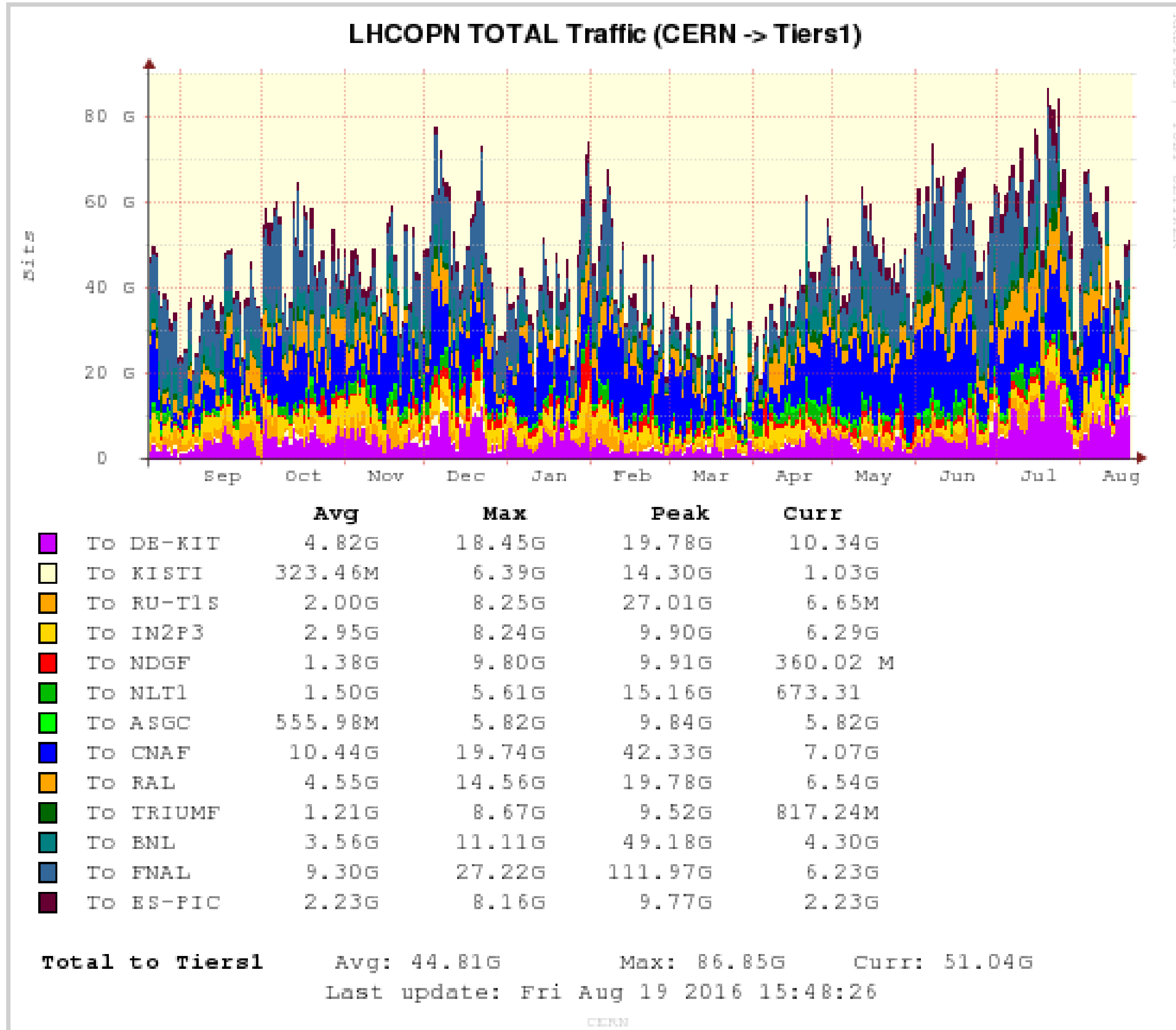








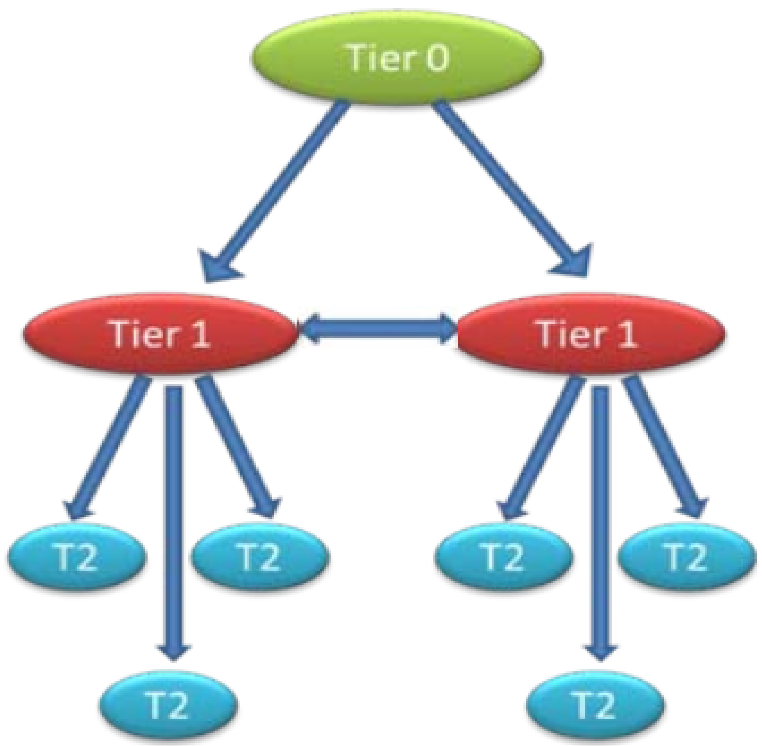
LHCOPN traffic – last 12 months



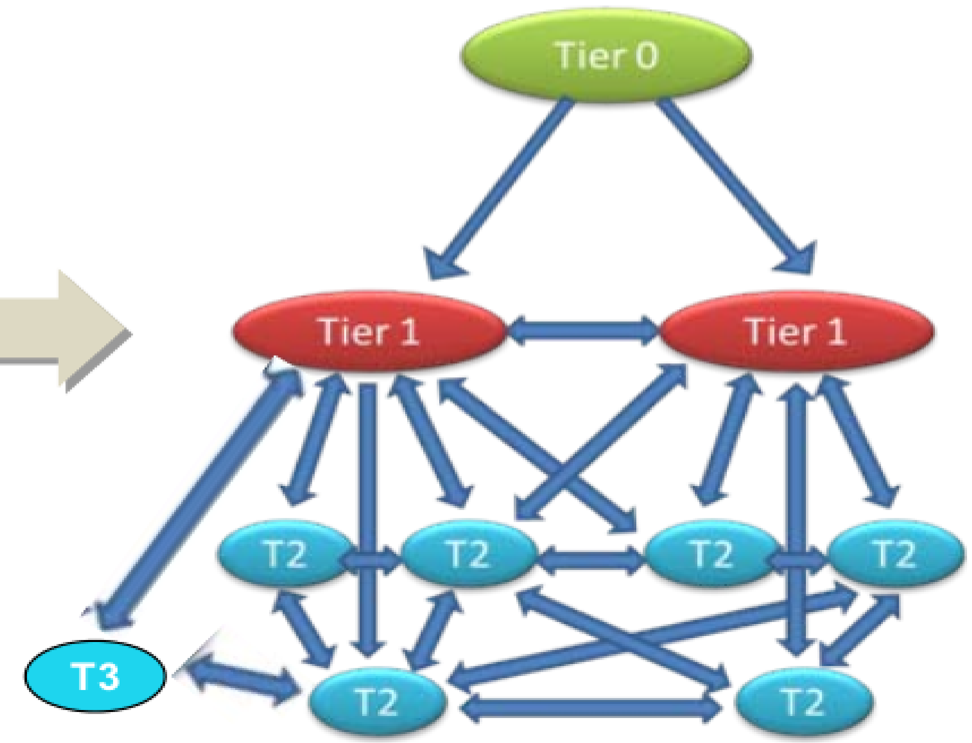
Analyses at Tier2s and Tier3s

LHCONE: Tier1/2/3 network

Computing Model evolution



Original MONARCH model



Model evolution

LHCONE concept

- Serving any LHC sites according to their needs and allowing them to grow

- Sharing the cost and use of expensive resources

- A collaborative effort among Research & Education Network Providers

- Traffic separation: no clash with other data transfer, resource allocated for and funded by the HEP community

- Trusted traffic that can bypass slow perimeter firewalls

LHCONE services

L3VPN (VRF): routed Virtual Private Network - *operational*

P2P: dedicated, bandwidth guaranteed, point-to-point links – *in development*

perfSONAR: monitoring infrastructure - *operational*

L3VPN service

Layer3 (routed) Virtual Private Network

Dedicated worldwide backbone connecting **Tier1s, T2s and T3s** at high bandwidth

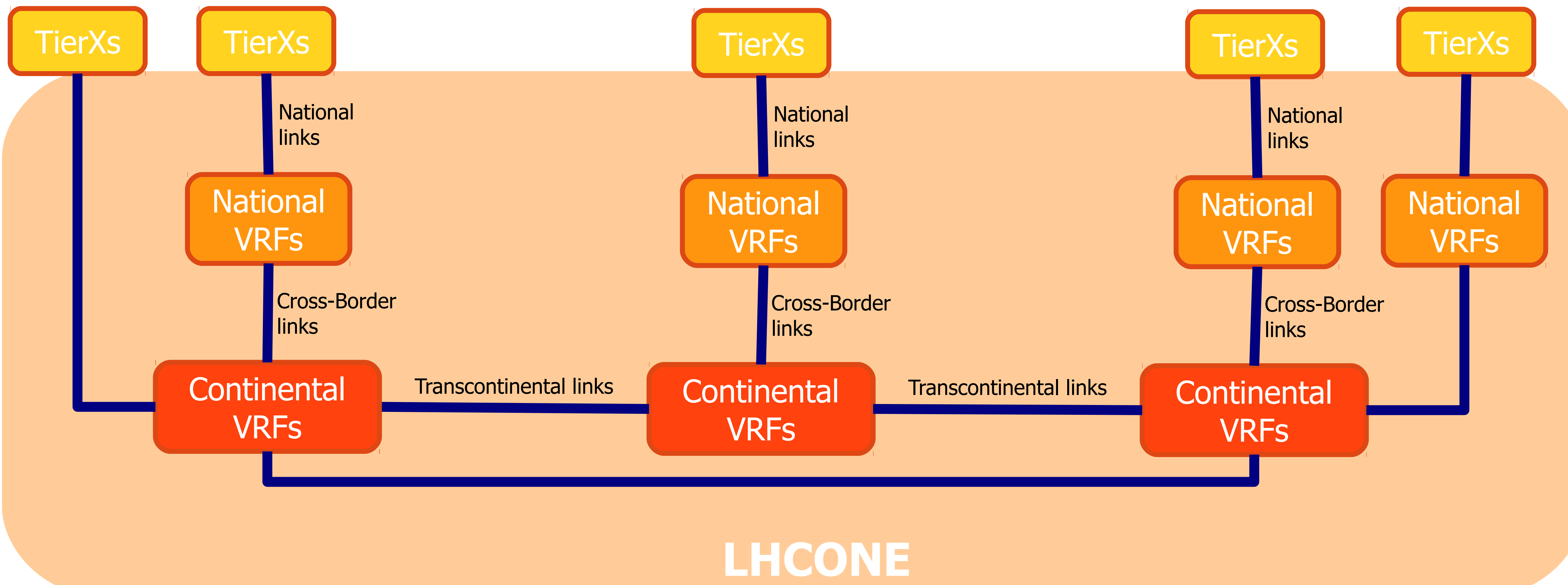
Bandwidth dedicated to LHC data transfers, no contention with other research projects

Trusted traffic that can bypass firewalls

LHCONE L3VPN architecture

- TierX sites connected to National-VRFs or Continental-VRFs
- National-VRFs interconnected via Continental-VRFs
- Continental-VRFs interconnected by trans-continental/trans-oceanic links

Acronyms: **VRF** = Virtual Routing Forwarding (virtual routing instance)



LHCONE Status

Over 17 national and international Research Networks

Several Open Exchange Points including NetherLight, StarLight, MANLAN, WIX, CERNlight and others

Trans-Atlantic connectivity provided by ESnet, GEANT, Internet2, NORDUnet and SURFnet

Trans-pacific connectivity provided by ASGCnet, KREOnet, SINET

~57 end sites connected to LHCONE:
- 10 Tier1s
- 47 Tier2s

Recent developments

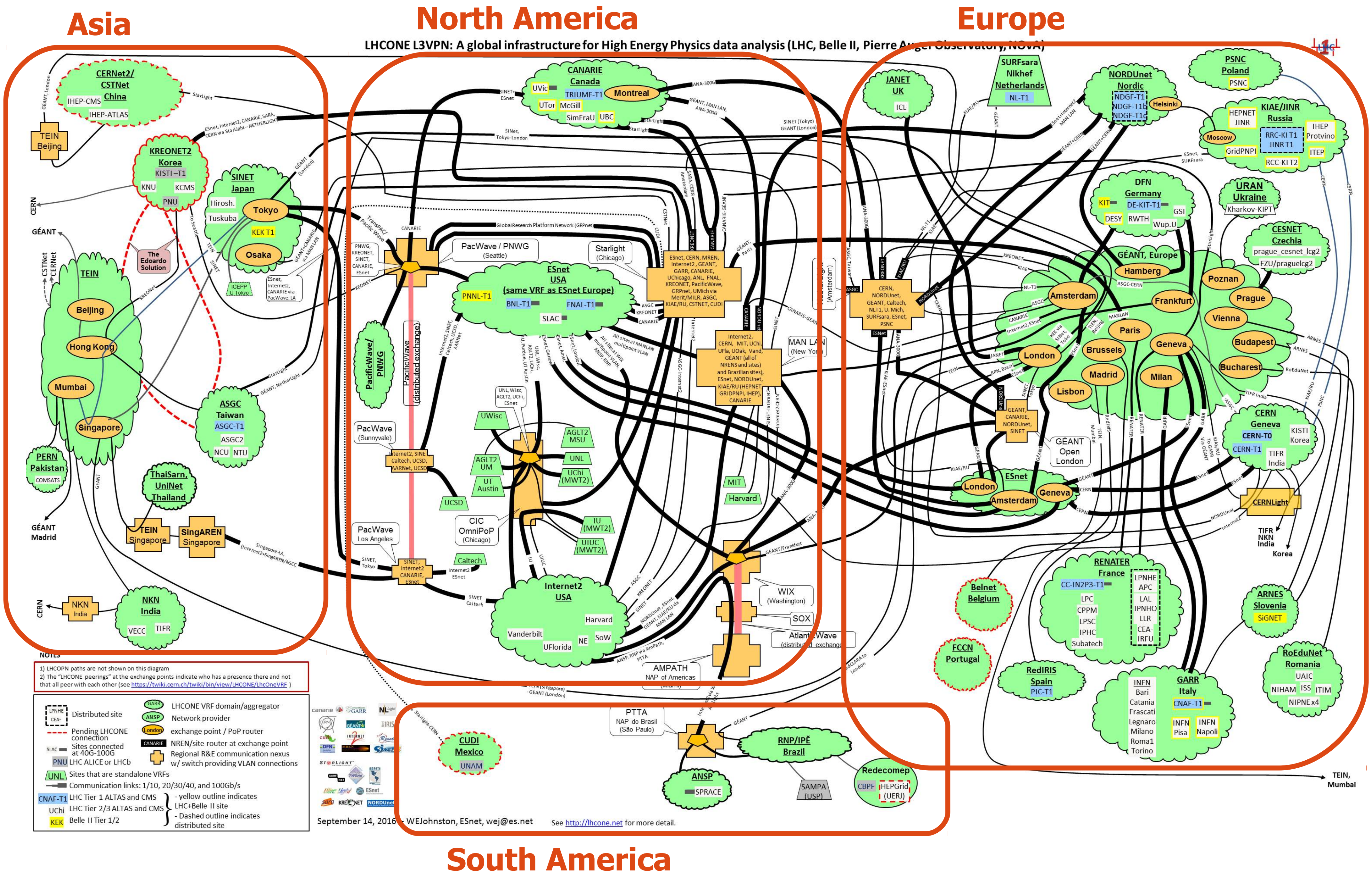
The LHCONE network is expanding
- Ukrain and Belgium are the latest European countries connected
- North American sites moving to 100G connections
- South America is now a stable partner, Chile is interested to join
- Traffic now being exchange with TEIN (Asia)
- Japan has connected with 2x10G links

Traffic within LHCONE is steadily growing
- Growth of over 65% from Q2 2015 to Q2 2016
- GÉANT has seen peaks of over 100Gbps
- ESnet moved around ~110PB of data in the last year, which count for ¼ of their traffic.

Some NRENs and sites need to upgrade their connection
- GÉANT is already working with the NRENs for the upgrades

Expected to see further increases after the upgrades

Current topology



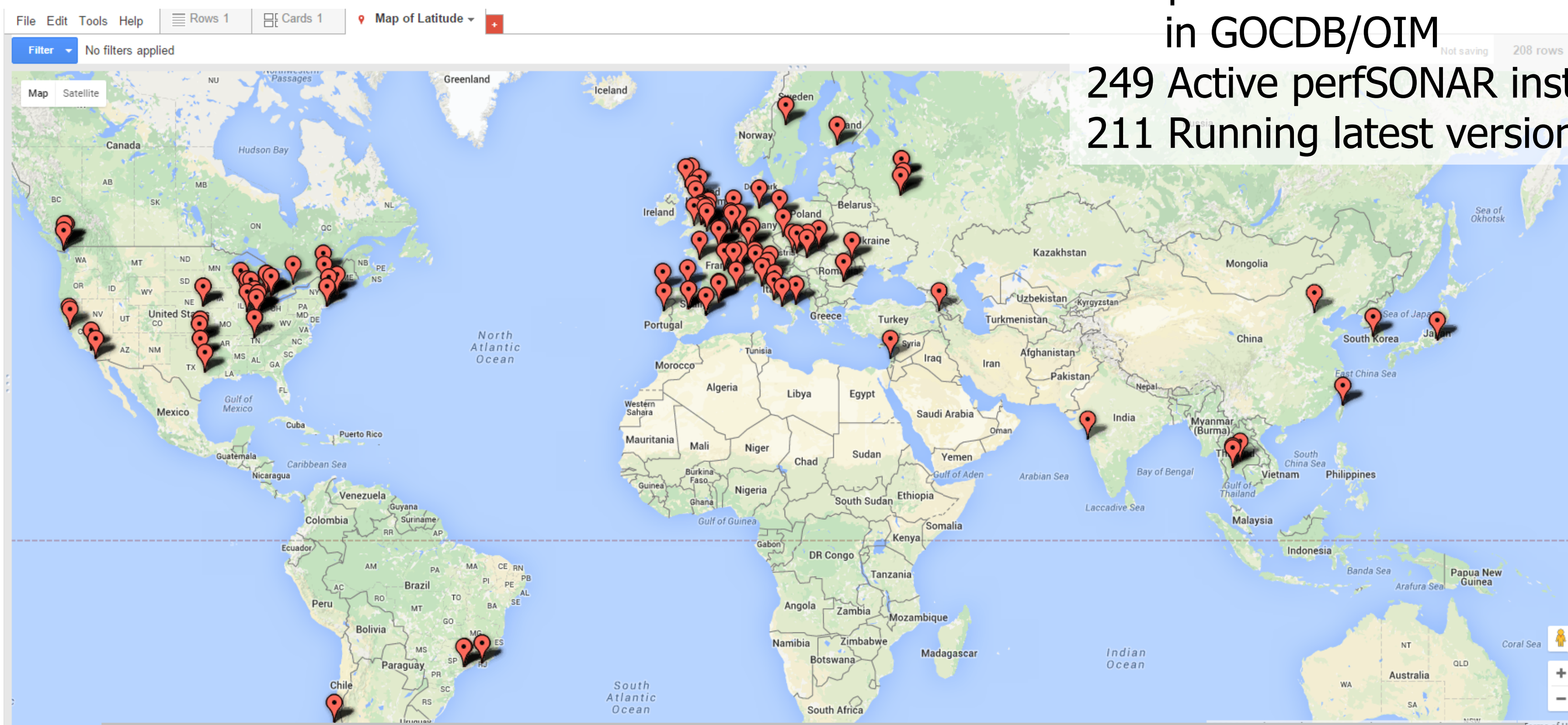
GEANT connects to Georgia and Armenia

The next countries that will connect to LHCONE?!

LHCONE perfSONAR

http://grid-monitoring.cern.ch/perfsonar_report.txt for stats

278 perfSONAR instances registered in GOCDB/OIM
249 Active perfSONAR instances
211 Running latest version (3.5+)



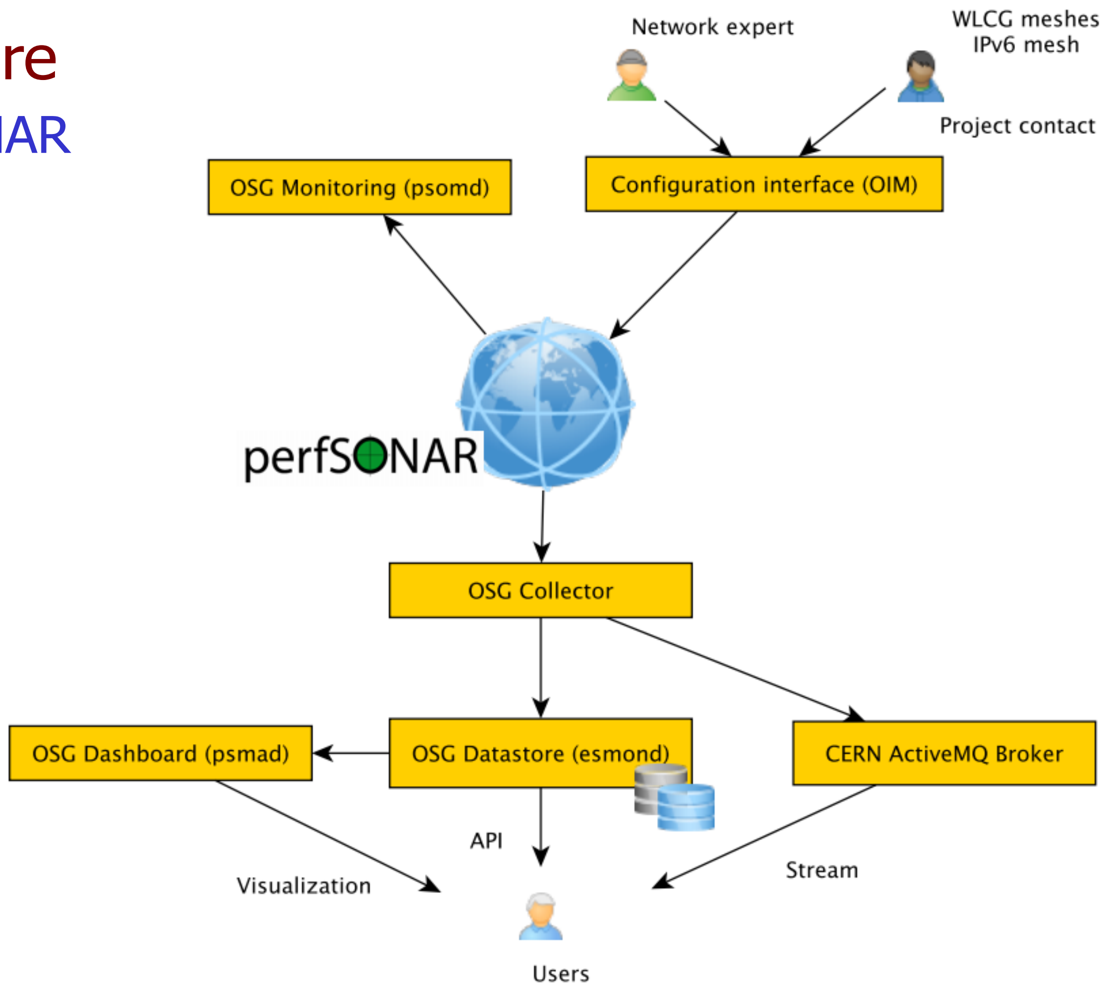
https://www.google.com/fusiontables/DataSource?docid=1QT4r17HEufkvnqhJu24nIptZ66XauYEIBWWh5Kpa#map:id=3

- Initial deployment coordinated by WLCG perfSONAR TF
- Commissioning of the network followed by WLCG Network and Transfer Metrics WG

Slide credit: Shawn McKee, University of Michigan

PerfSONAR: gathering and storing metrics

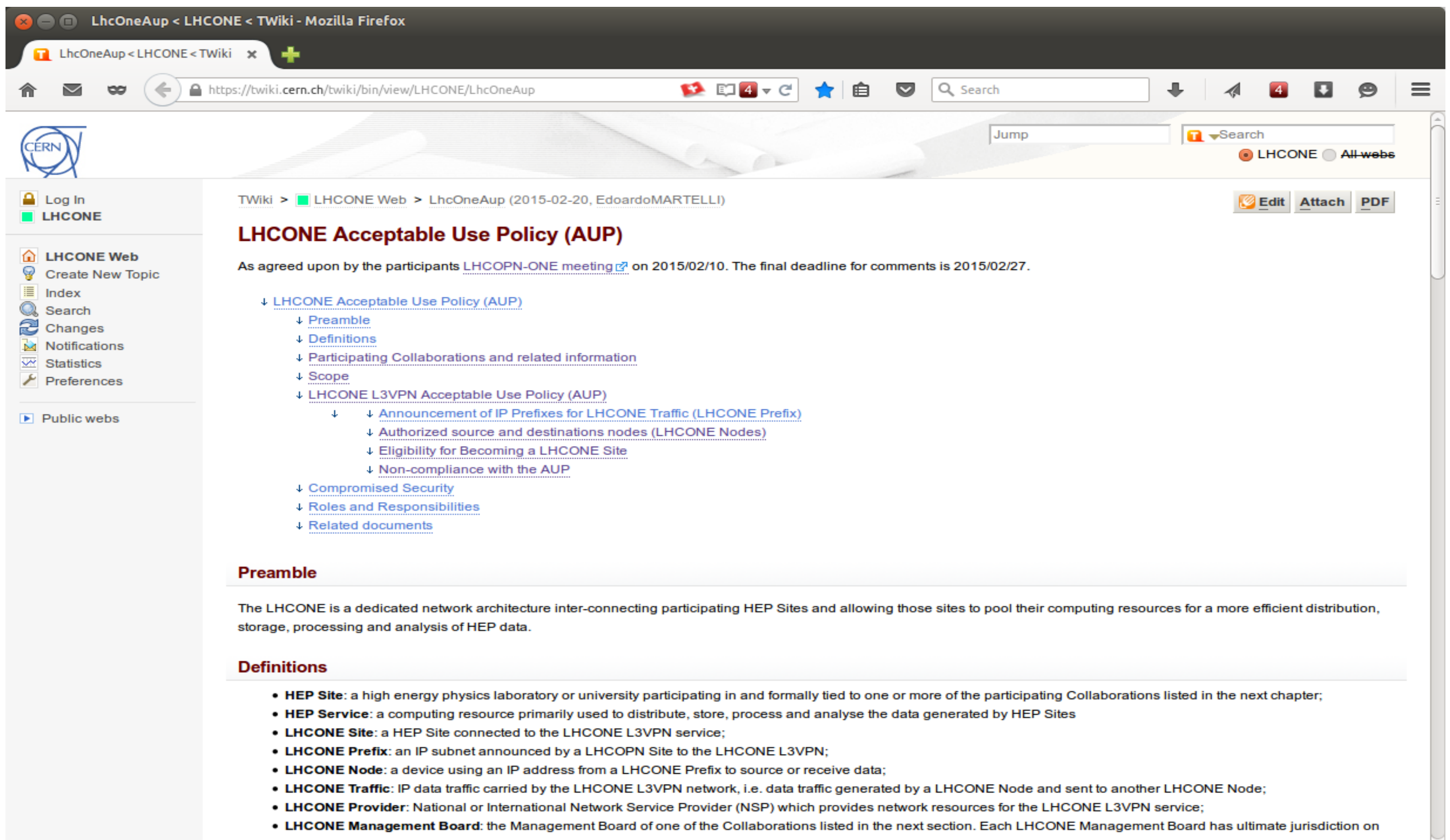
- OSG is providing network metric data for its members and WLCG via the Network Datastore
  - The data is gathered from all WLCG/OSG perfSONAR instances
  - Stored indefinitely on OSG hardware
  - Data available via Esmond API
  - In production since September 14th 2015
- The primary use-cases
  - Network problem identification and localization
  - Network-related decision support
  - Network baseline: set expectations and identify weak points for upgrading



Slide credit: Shawn McKee, University of Michigan

LHCONE Acceptable Use Policy

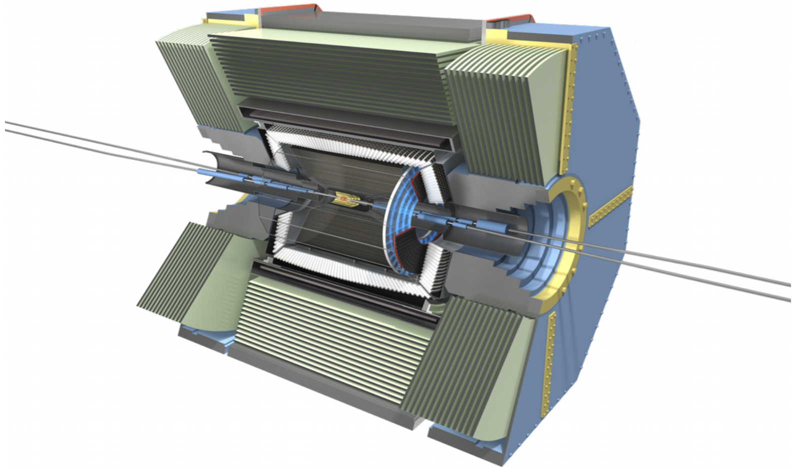
The LHCONE AUP has been defined to regulate the utilization of the L3VPN service (https://twiki.cern.ch/twiki/bin/view/LHCONE/LhcOneAup)



Open to other HEP collaborations

The L3VPN is now used also by:

- Belle II experiment



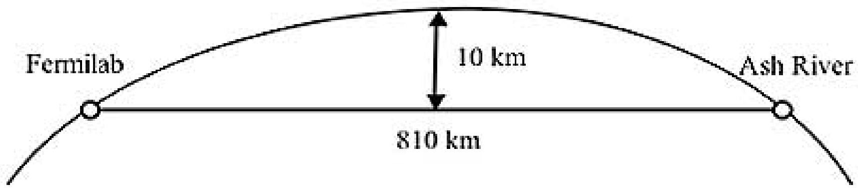
- Pierre Auger Observatory

NOvA experiment just joined

NOvA experiment

- Neutrino oscillation accelerator experiment with FD at Ash River
  - Oscillation parameters
  - CP violation

FZU (CZ) works with FNAL (US) on the NOvA experiment.



Next: XENON dark matter project

The XENON dark matter project has asked permission to use LHCONE. They will produce more information to support the request

http://xenon.astro.columbia.edu/index3.html

More information on LHCOPN and LHCONE

Latest LHCOPN/ONE meetings:
Taipei March2016: https://indico.cern.ch/event/461511/
Helsinki September 2016: https://indico.cern.ch/event/527372/

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

Latest development

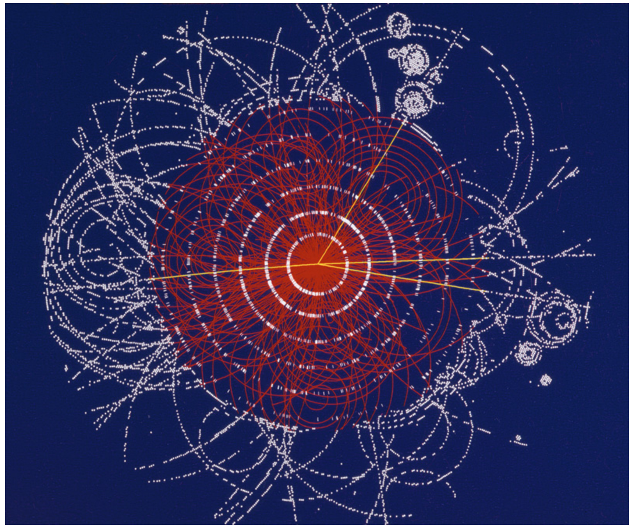
Opportunistic resources from Commercial Clouds

Commercial Cloud Services

CERN and other academic institutes have being evaluating the use of Commercial Cloud Services

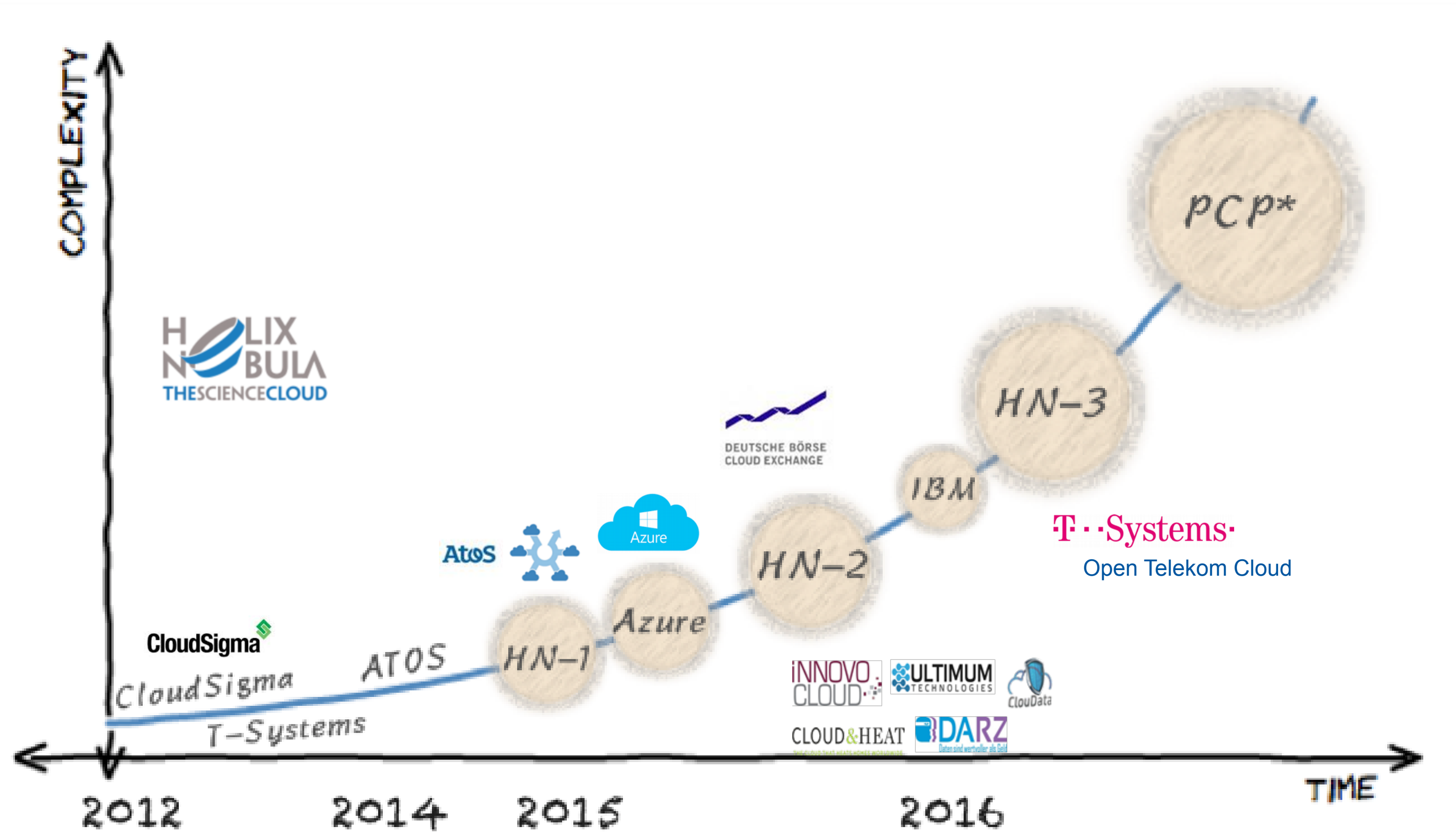
Research and Education Networks (REN) are evaluating how to connect Cloud Service Providers (CSP) to their customers

Main issues:
- deliver traffic from cloud datacentres to different continents
- avoid or not cloud-to-cloud traffic
- not all the RENs allow commercial traffic



CERN approach to Cloud procurement

Series of short procurement projects of increasing size and complexity



Slide credit: Bob Jones (CERN)

Cloud procurement so far

- HN-1: ATOS
  - Detector simulation for ATLAS
- Microsoft Azure evaluation
- HN-2: DBCE
  - Detector simulation for ATLAS, CMS, ALICE, LHCb
- IBM SoftLayer evaluation
- HN-3: T-Systems OTC)
  - Detector simulation, reconstruction and analysis for ATLAS, CMS, ALICE, LHCb

Credit: Domenico Giordano (CERN)

Requirements for VMs and Storage

VMs

- 1000VMs
- Each VM shall consist of at least 4 vCPUs
- Each VM shall have at least 8 GB of memory
- Each VM shall have 100 GB of primary local storage
- **Each VM shall be accessible remotely by all WLCG sites over the Internet via a public IPv4 address**.

Storage

- The usable storage capacity shall be at least 500TB.
- **The clustered storage shall be accessible remotely by all WLCG sites over the Internet via public IPv4 addresses.**

Requirements for WAN

- Use of NAT is only permitted for a 1:1 address translation, where each external address is assigned to exactly one internal address. Use of Port Address Translation (PAT) is not permitted.

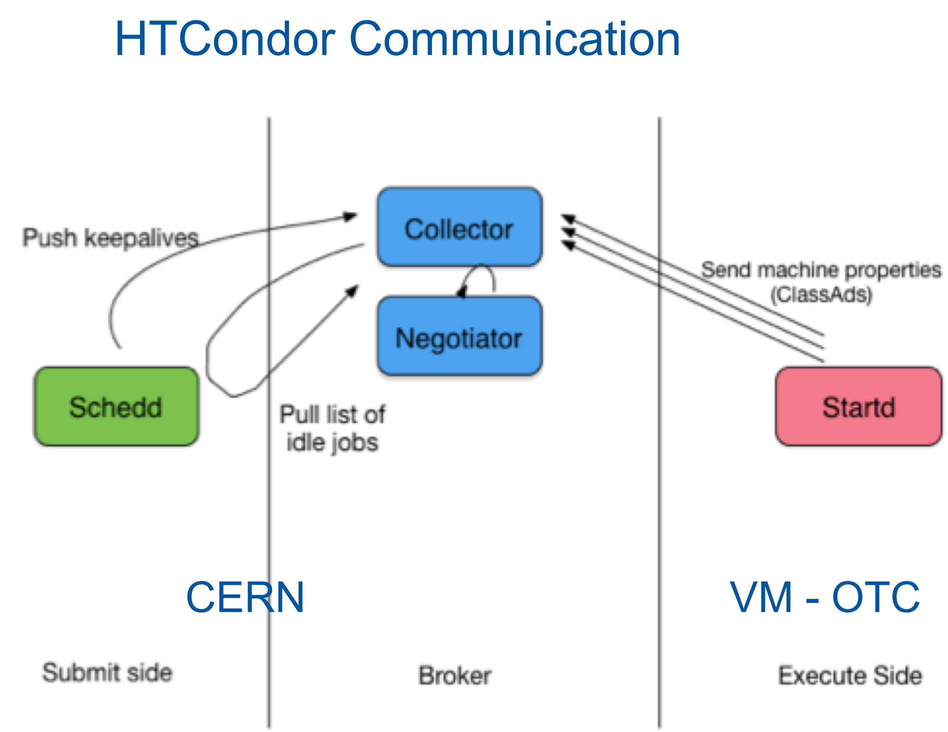
- To provide IP connectivity, the contractor's site from where the IaaS cloud resources will be made available shall be connected to or peering with at least one of the following:
  - a. an NREN, which can provide the contractor with transit to CERN;
  - b. GEANT or NORDUnet;
  - c. an IXP which CERN is also connected to (currently CIXP).

- The total reserved peak bandwidth available between CERN and VMs through the contractor’s connection shall be at least **40Gbps**.

Use: transparent extension of CERN resources

- Deployed HTCondor worker nodes in OTC

- Same HTCondor Batch instance exposing CERN on-premise resources
  - Transparent to the user, same entry point for job submission
  - Grid jobs from users are submitted to the Schedd



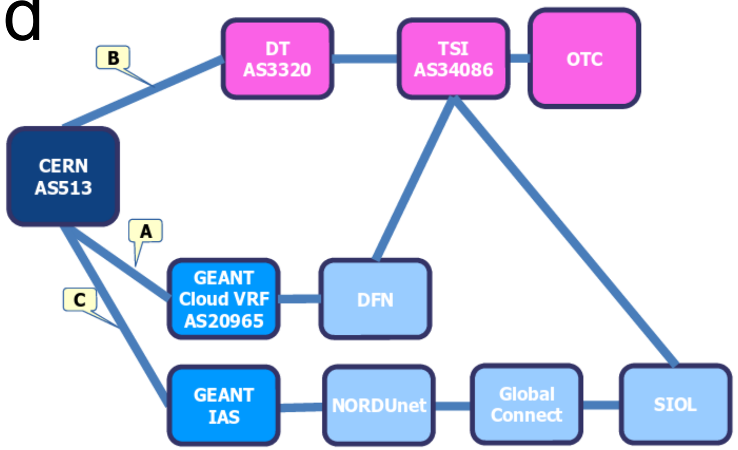
Slide credit: Domenico Giordano (CERN)

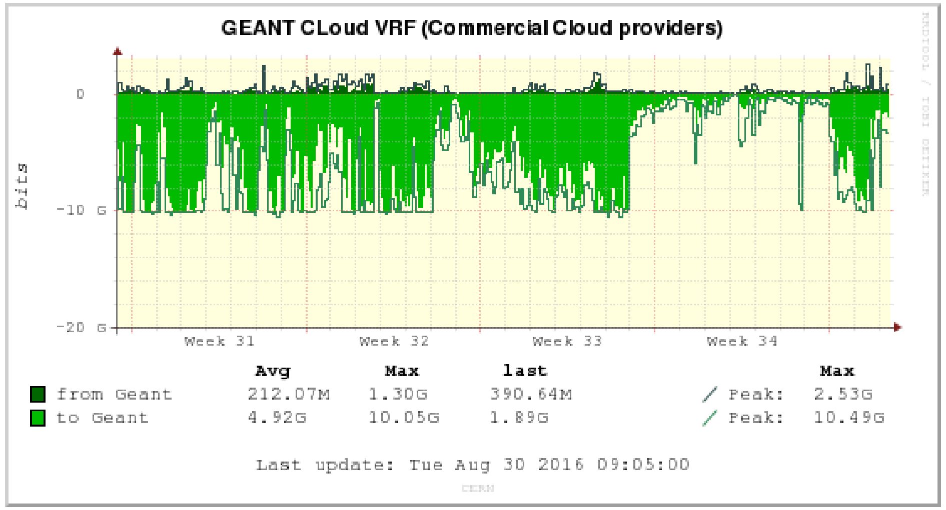
Experiences with Cloud Services

CERN has recently evaluated IBM Softlayer and T-Systems cloud services

Softlayer was already peering with GEANT in Frankfurt and provided good network performance

T-Systems was asked to connect to the GEANT Cloud VRF and that has guaranteed 10Gbps of bandwidth





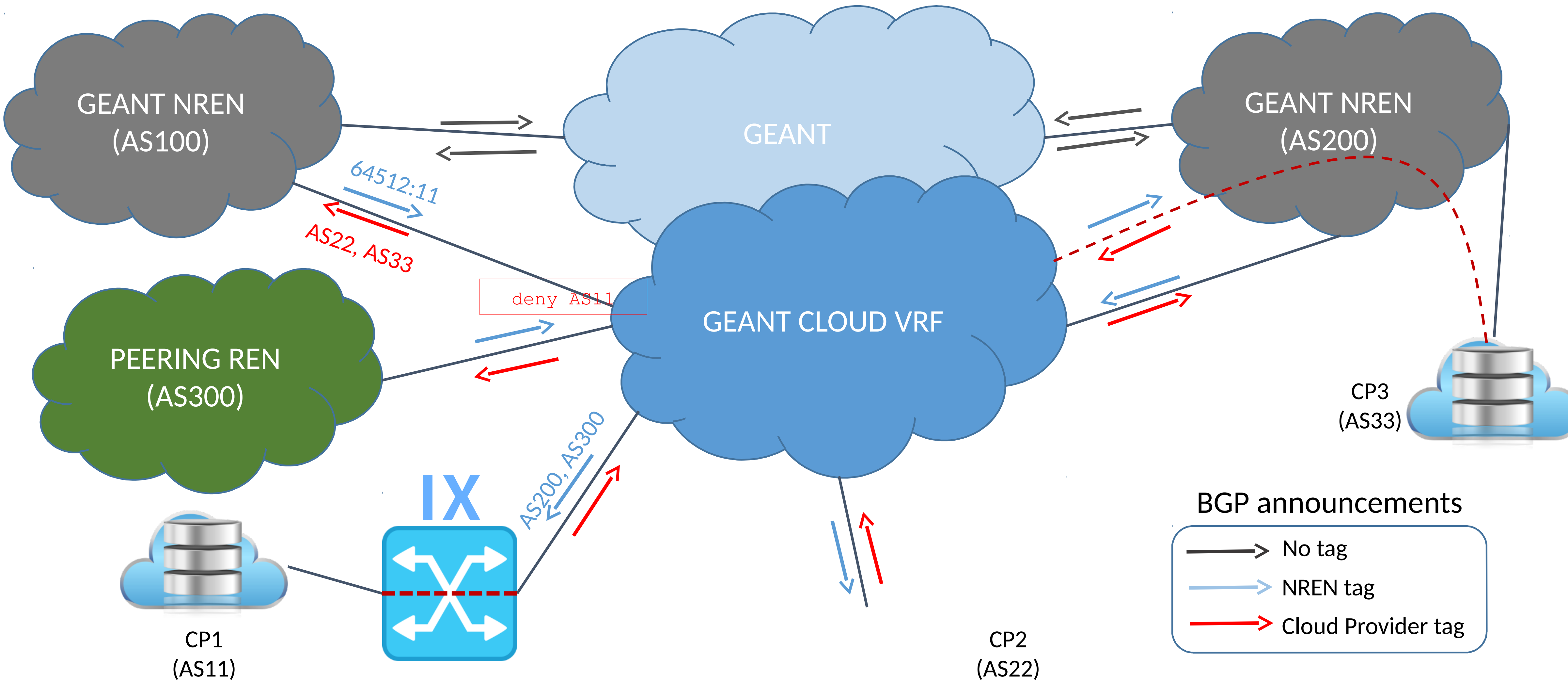
Connectivity for Commercial clouds

Research and Education Networks (REN) are evaluating how to connect commercial cloud providers to their users

Main challenges:
- deliver traffic from cloud datacentres to different continents
- avoid or not cloud-to-cloud traffic
- not all the RENs allow commercial traffic

Three solution presented:

GEANT: dedicated VRF

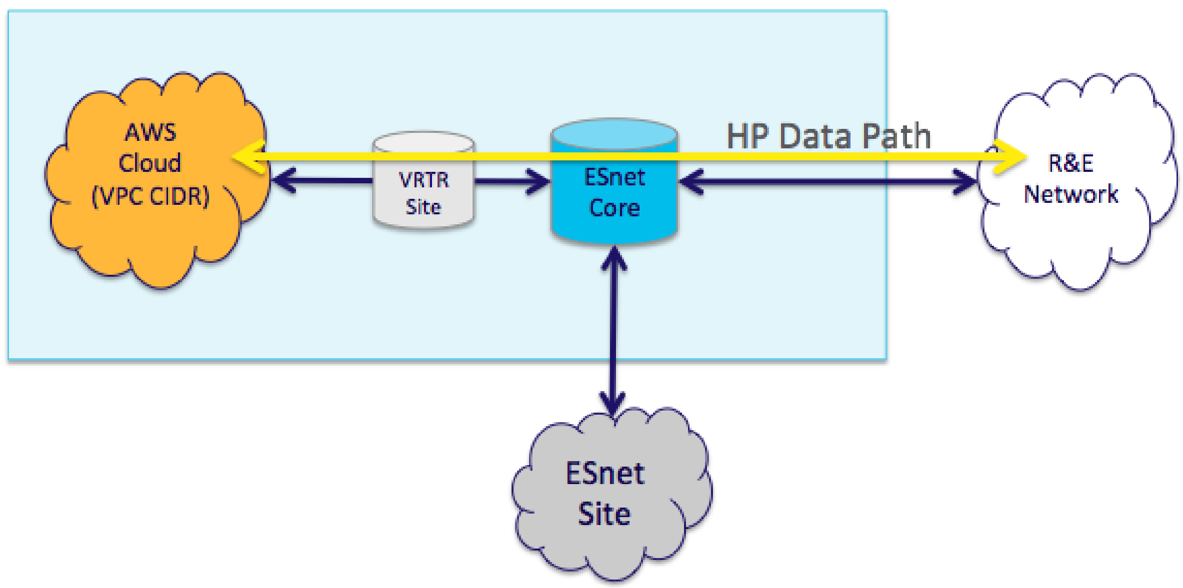


ESnet: on-net VPN termination

Virtual "Site Router" (VRTR) Service

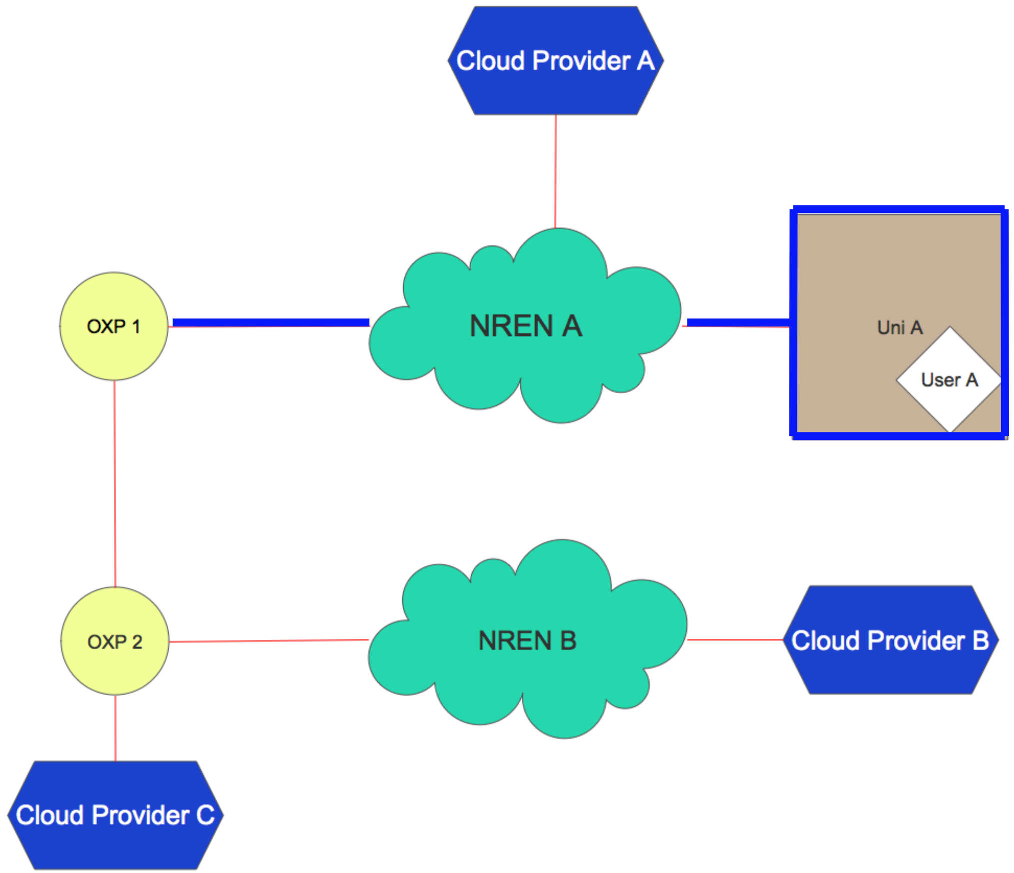
At the edge of the cloud

Virtual Site Router at AWS Exchange Point



Virtual "Site Router" improves path efficiency and takes pressure off of the site local-loop.

NORDUnet: transport to eXchange Points



Conclusion

Summary

Computer networks are a fundamental part of the LHC instrument

The excellent performance of the LHC are putting more load on all the networks

LHC Experiments: stable operations during Run2, planning for major network upgrades in Run3

LHCOPN: growing traffic, links being upgraded

LHCONE: expanding, especially in Asia

Commercial clouds: connectivity challenges

Upcoming meetings

2nd Asia Tier Center Forum

Date: 30 Nov 2016 → 2 Dec 2016
Location: Nakhon Ratchasima, Thailand
https://indico.cern.ch/event/558754/

WLCG pre-GDB meeting on networking

Date: 10th January 2017
Location: CERN, Geneva
https://indico.cern.ch/category/6890/

Next LHCONE LHCOPN meeting

Date to be defined (end of March/beginning of April 2017)
Location: BNL, New York

Questions?

edoardo.martelli@cern.ch