Scientific data processing at global scale

The LHC Computing Grid

fabio hernandez

fabio@in2p3.fr

第十五届全国科学计算与信息化会议暨现代物理信息化论坛

Chengdu (China), July 5th 2011

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE DES AFFAIRES ÉTRANGÈRES ET EUROPÉENNES

中國科學院高能物理研究所
Institute of High Energy Physics
Chinese Academy of Sciences

CCIN2P3

Who I am

- Computing science background
- Working in the field of computing for high-energy physics since 1992

  software development for scientific data management (data transfer over high-throughput networks, mass storage & retrieval, cataloguing, ...) and operations of IT services for research
- Involved in grid computing projects since 2000 and in particular in planning, prototyping, deploying and operating the computing infrastructure for the LHC

  technical leadership of the French contribution to the LHC computing grid

  served in the management board and grid deployment board of the WLCG collaboration
- Served as deputy director of IN2P3 computing centre, which hosts and operates the French WLCG tier-1
- Since August 2010, visiting IHEP computing centre

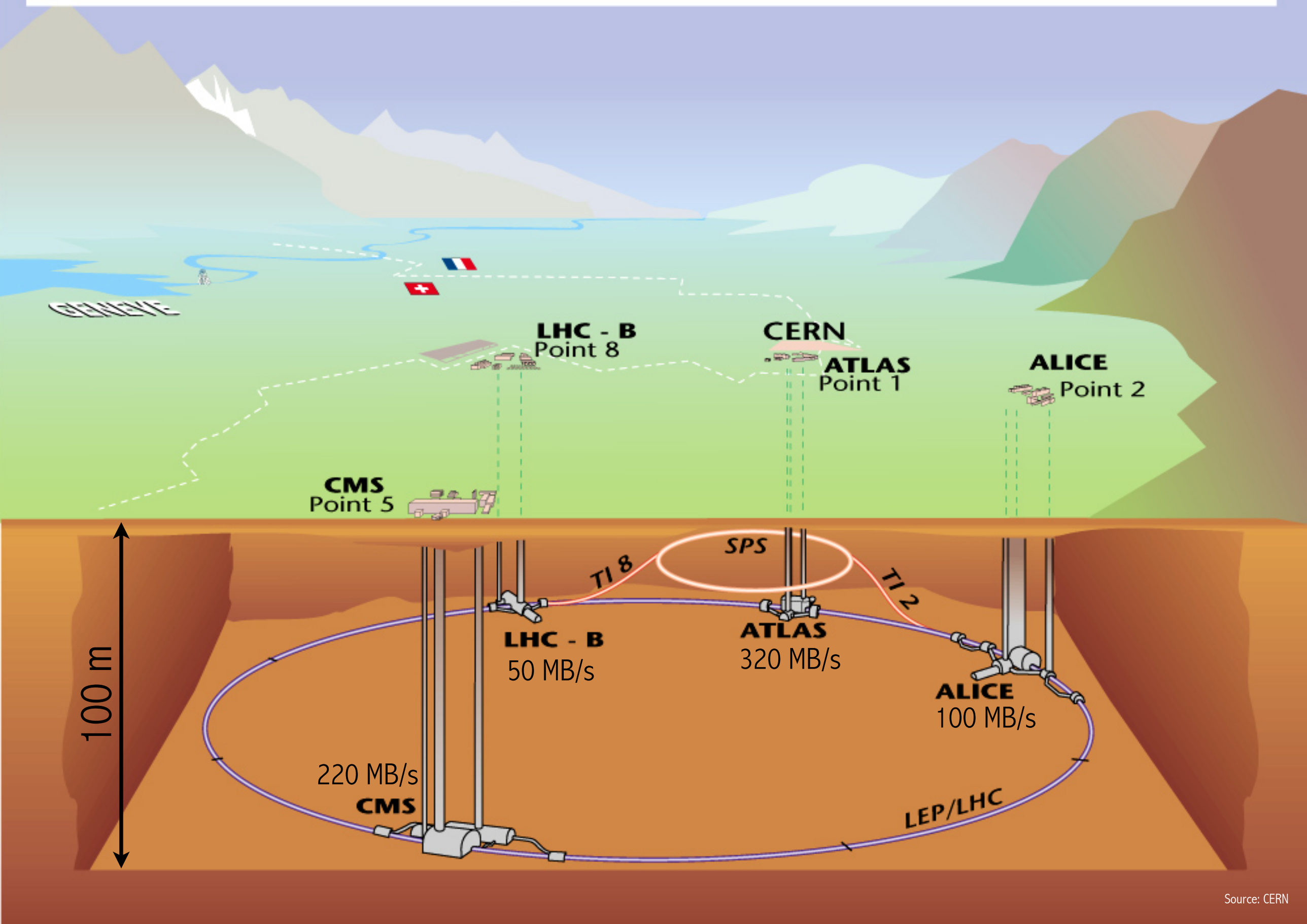
Contents

- LHC overview
- LHC Computing Grid
- Data distribution
- Data processing
- Perspectives
- Questions & Comments


LHCb
ATLAS
CERN Meyrin
CERN Prévessin
SPS 7 km
ALICE
SUISSE
FRANCE
CMS
LHC 27 km

Photo © CERN

Vue d'ensemble des expériences LHC.



Source: CERN

Scientific research at global scale

- High energy physics research

  Size, ***complexity*** *and* ***cost*** *of instruments (particle accelerators and detectors)*

  Highly skilled set of ***large teams*** *to build and operate the instruments and the research infrastructure*

- Teams, instruments and funding are multinational and **highly distributed**

  Required coordinated effort to answer specific research questions

CMS Collaboration

June 2008

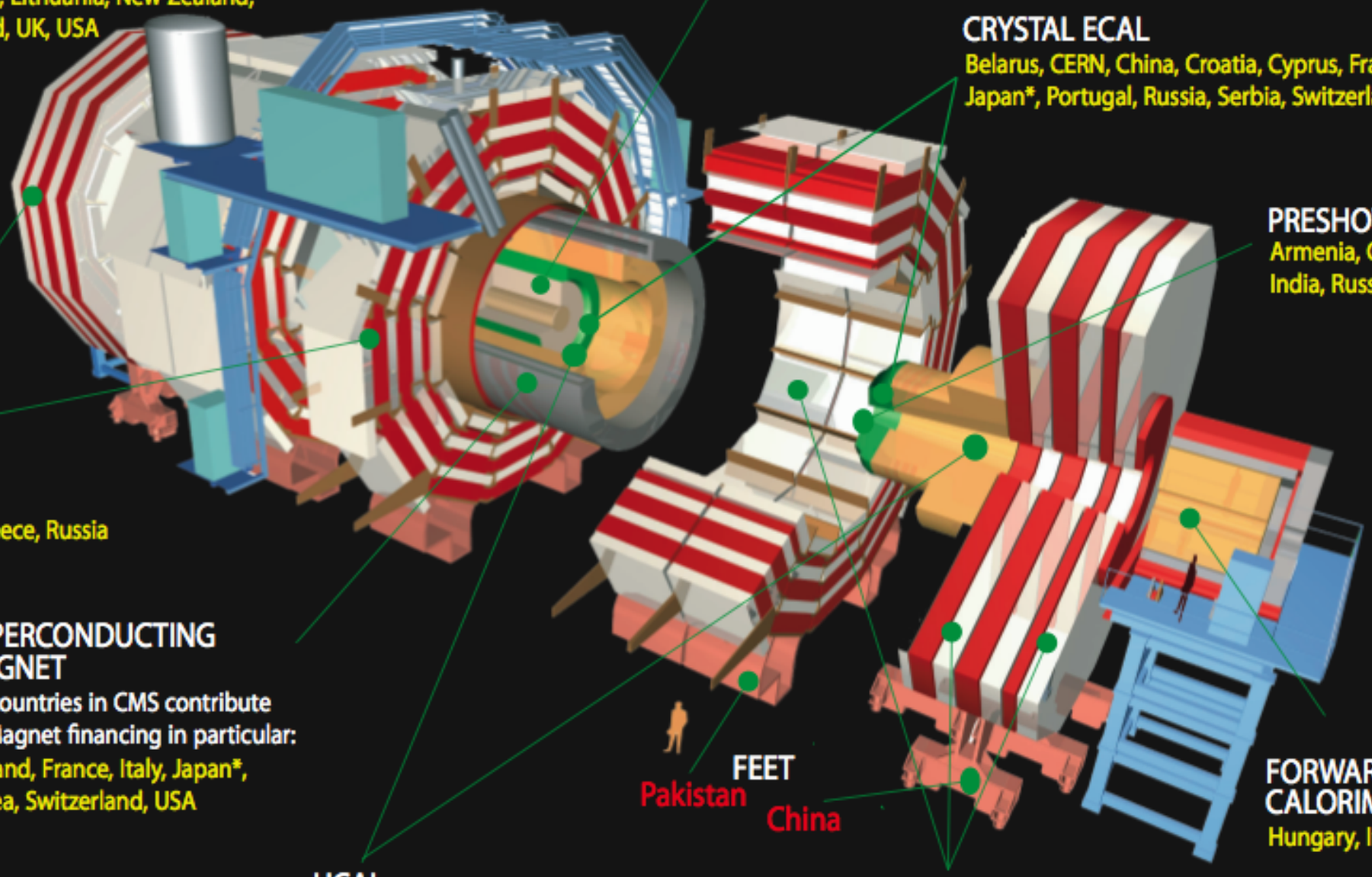
38 Countries, 183 Institutes, 3000 scientists and engineers (including 400 students)

TRIGGER, DATA ACQUISITION & OFFLINE COMPUTING
Austria, Brazil, CERN, Finland, France, Greece, Hungary, Ireland, Italy, Korea, Lithuania, New Zealand, Poland, Portugal, Switzerland, UK, USA

TRACKER
Austria, Belgium, CERN, Finland, France, Germany, Italy, Japan*, Mexico, New Zealand, Switzerland, UK, USA

CRYSTAL ECAL
Belarus, CERN, China, Croatia, Cyprus, France, Italy, Japan*, Portugal, Russia, Serbia, Switzerland, UK, USA

PRESHOWER
Armenia, CERN, Greece, India, Russia, Taiwan



RETURN YOKE
Barrel: Estonia, Germany, Greece, Russia
Endcap: Japan*, USA

SUPERCONDUCTING MAGNET
All countries in CMS contribute to Magnet financing in particular:
Finland, France, Italy, Japan*, Korea, Switzerland, USA

FEET
Pakistan
China

FORWARD CALORIMETER
Hungary, Iran, Russia,Turkey, USA

HCAL
Barrel: Bulgaria, India, Spain*, USA
Endcap: Belarus, Bulgaria, Georgia, Russia, Ukraine, Uzbekistan
HO: India

MUON CHAMBERS
Barrel: Austria, Bulgaria, CERN, China, Germany, Hungary, Italy, Spain,
Endcap: Belarus, Bulgaria, China, Colombia, Korea, Pakistan, Russia, USA

Total weight	: 12500 T
Overall diameter	: 15.0 m
Overall length	: 21.5 m
Magnetic field	: 4 Tesla

* Only through Industrial contracts

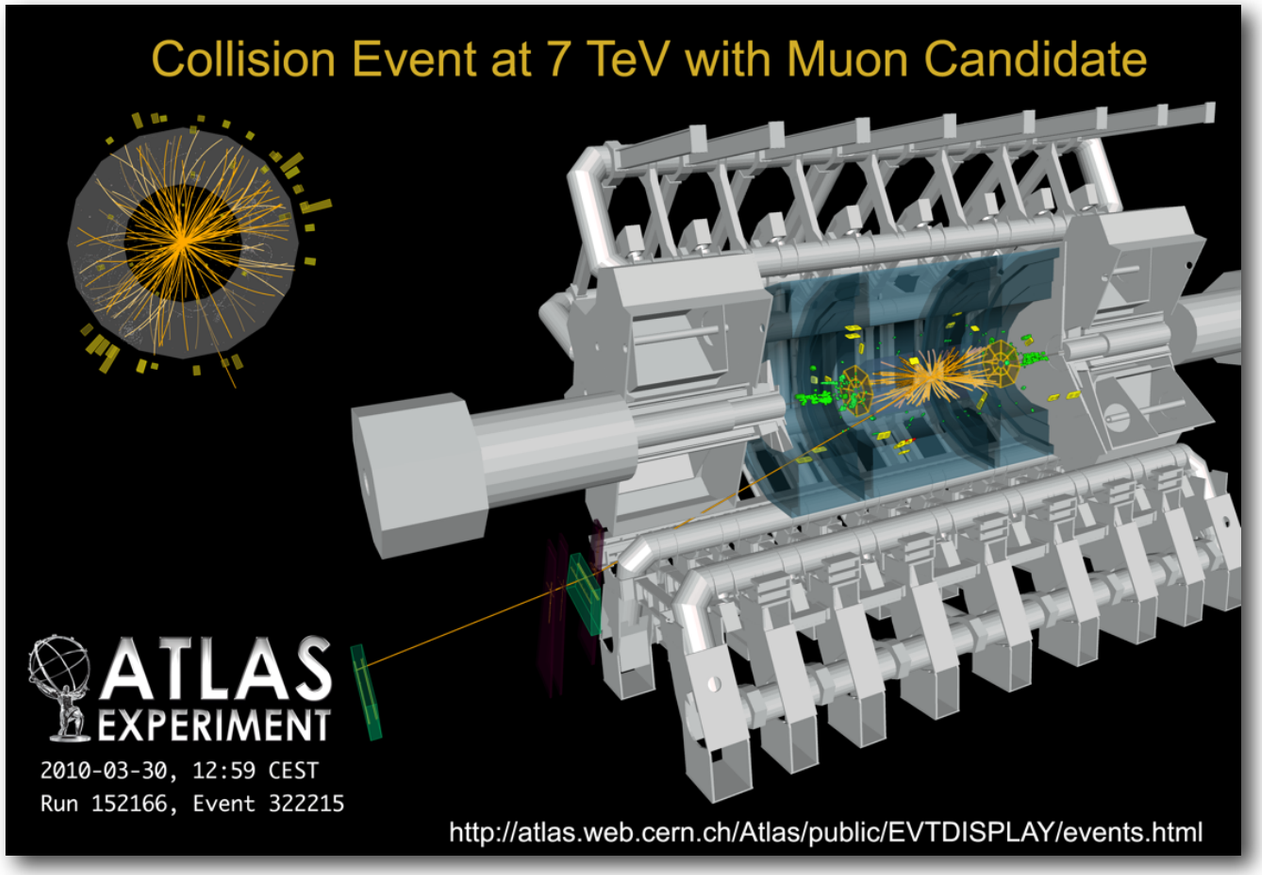
Data acquisition

- Design event rates

  40 million collisions/second

  After hardware- and software-based filtering, 1 out of 200.000 collisions stored

Experiment	Data Rate [MB/sec]
Alice	100
Atlas	320
CMS	220
LHCb	50
∑ All Experiments	690



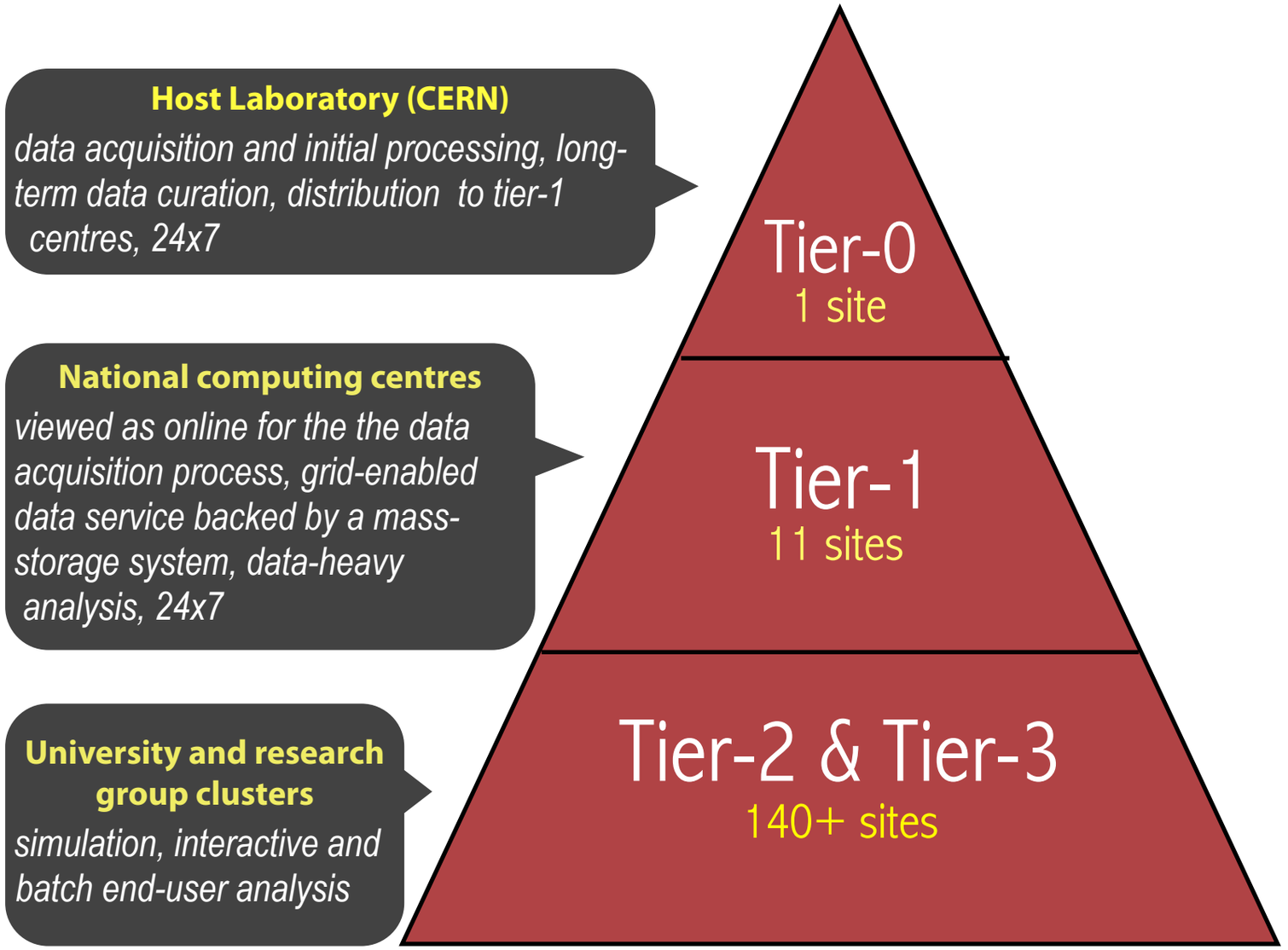
7 PB of additional raw data per nominal year*
excluding derived and simulated data

* Accelerator duty cycle assumption: 14 hours/day, 200 days/year

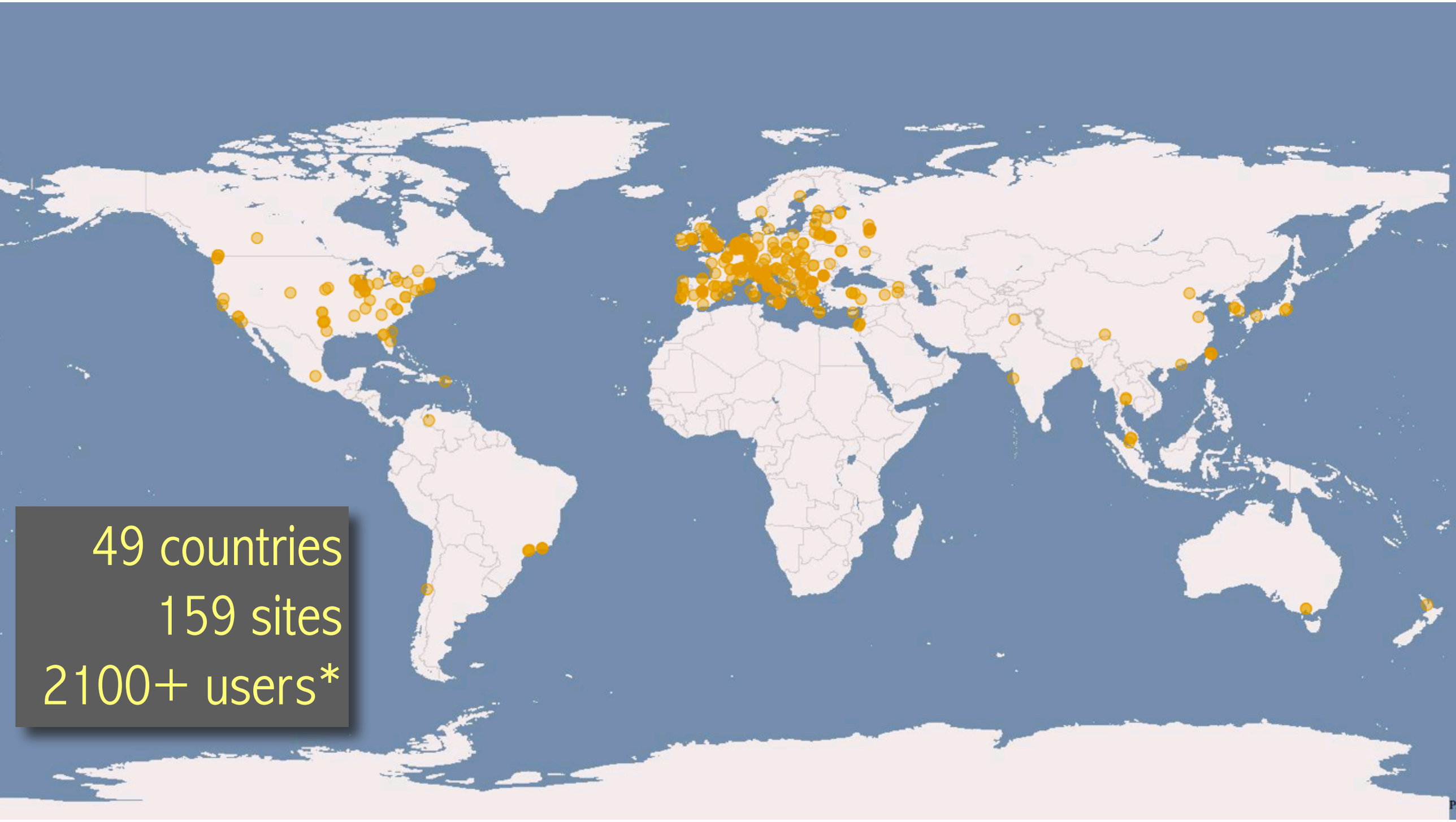
LHC Computing Grid: Architecture

- Grid-based infrastructure, composed of 150+ geographically dispersed sites, linked by high-speed networks
- Integration of resources into a coherent environment that can be used by any collaboration member

from desktop to clusters in universities to high-performance computing centres in national laboratories



Worldwide collaboration



* As of June 2011

Source: WLCG GStat

Scale: available resources

- Aggregated resources made available by the participating sites for year 2011 for all LHC experiments

Resource	
CPU	1.5M HS06
Disk	132 PB
Tape	130 PB

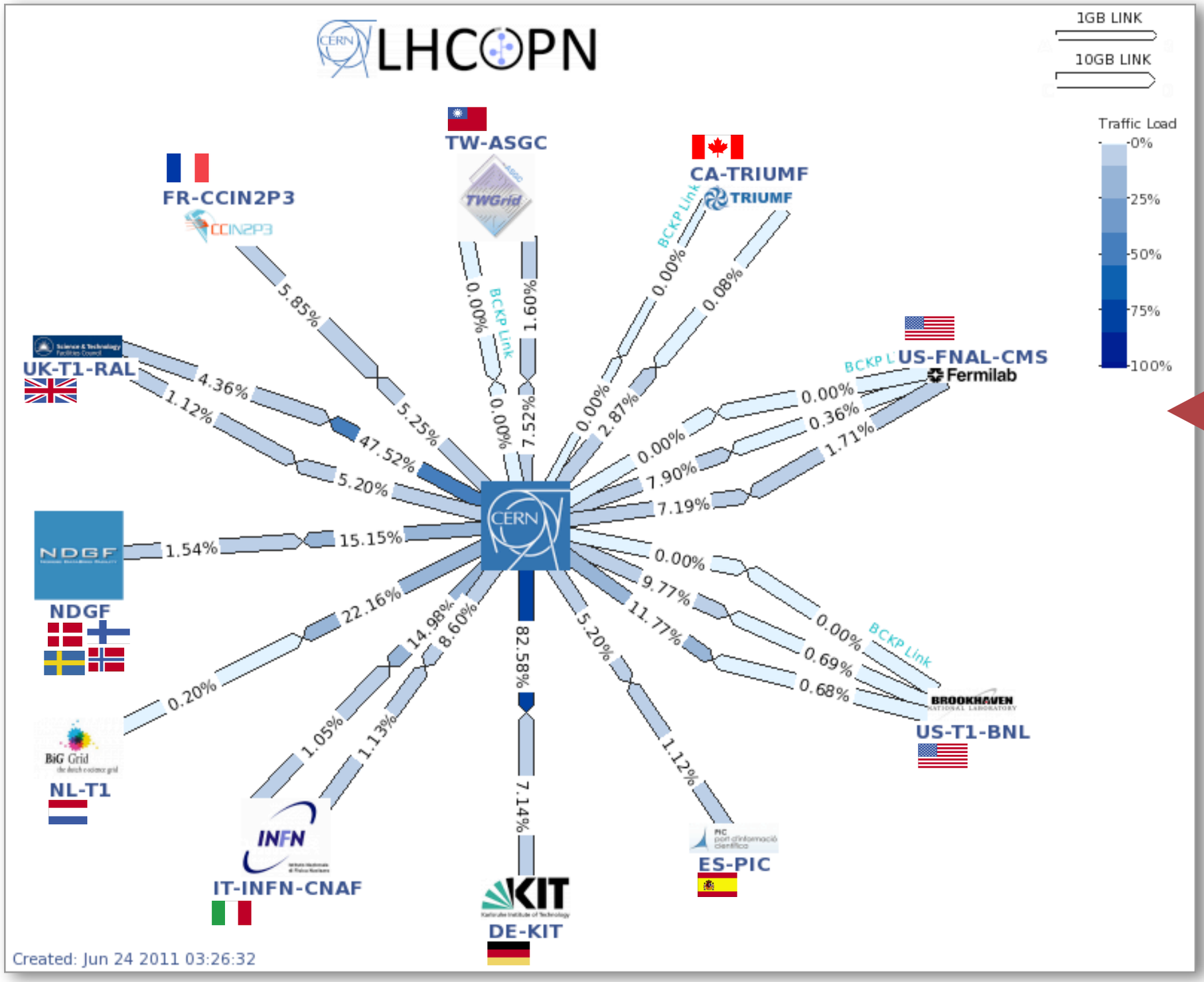
Equivalent capacity of 175.000 recent CPU cores*

Equivalent to 66.000 disk spins, 2TB each

* Intel Xeon 5650 @ 2.6 GHz

Source: WLCG Resource, Balance and Usage

Data distribution: infrastructure



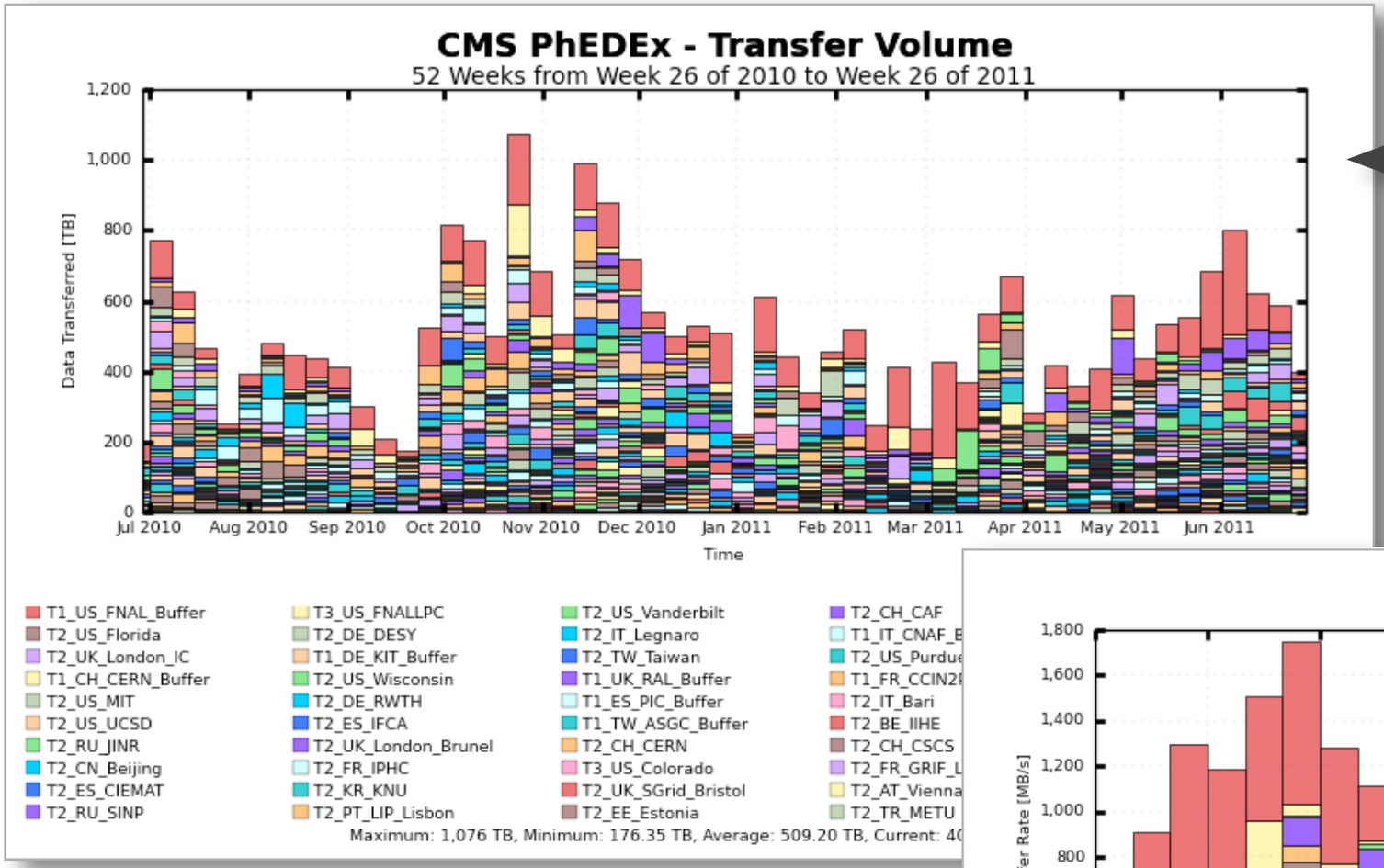
CERN is connected to tier-1 sites through dedicated and redundant 10 Gbps links, provided by national academic and research networks

Usage restricted to LHC data exchange

General purpose research networks used for data exchange with tier-2s and tier-3s

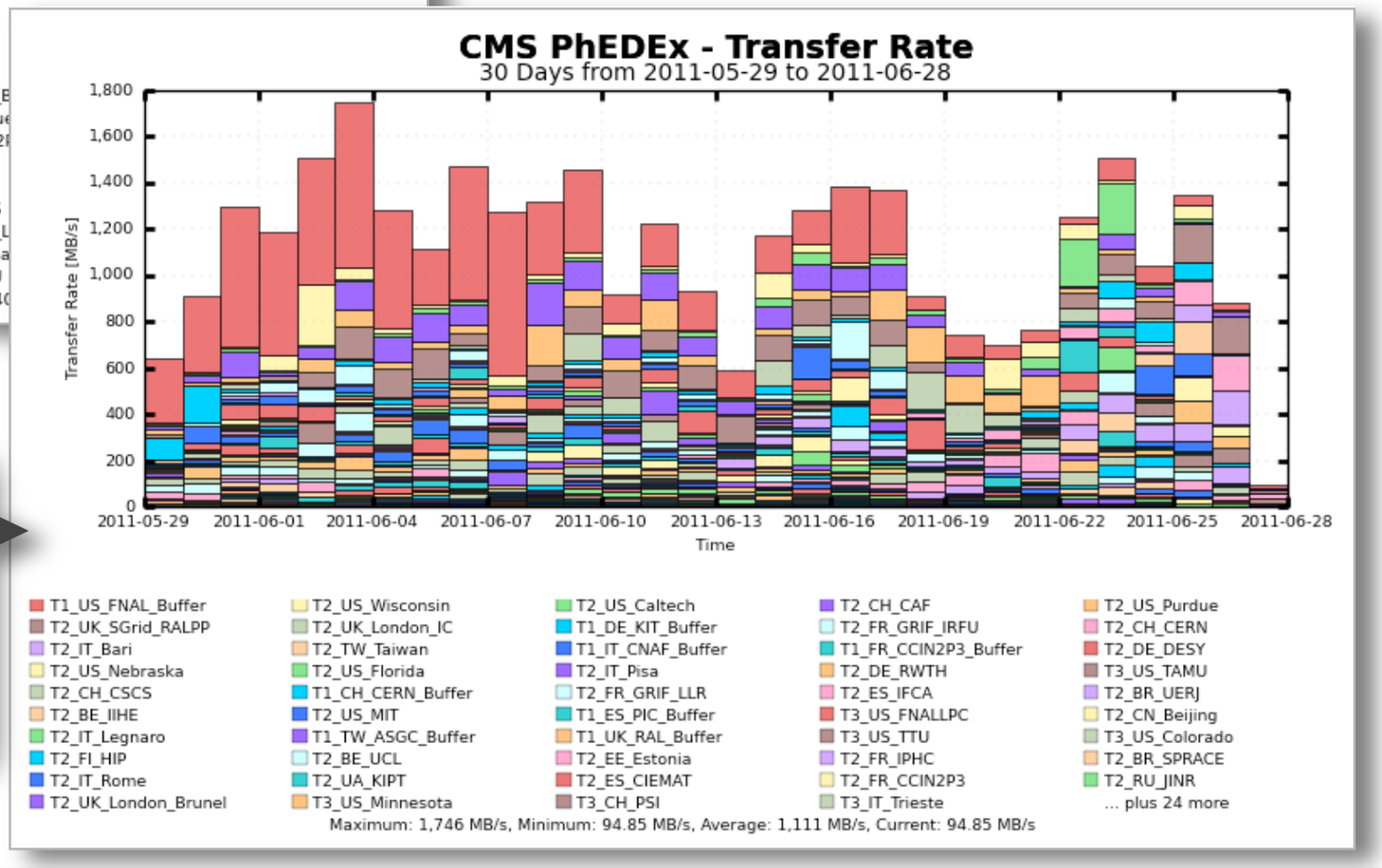
Source: LHCOPN

Data exchange: example of CMS



Volume

average aggregate daily data exchange over 60+ sites: ***500 TB***

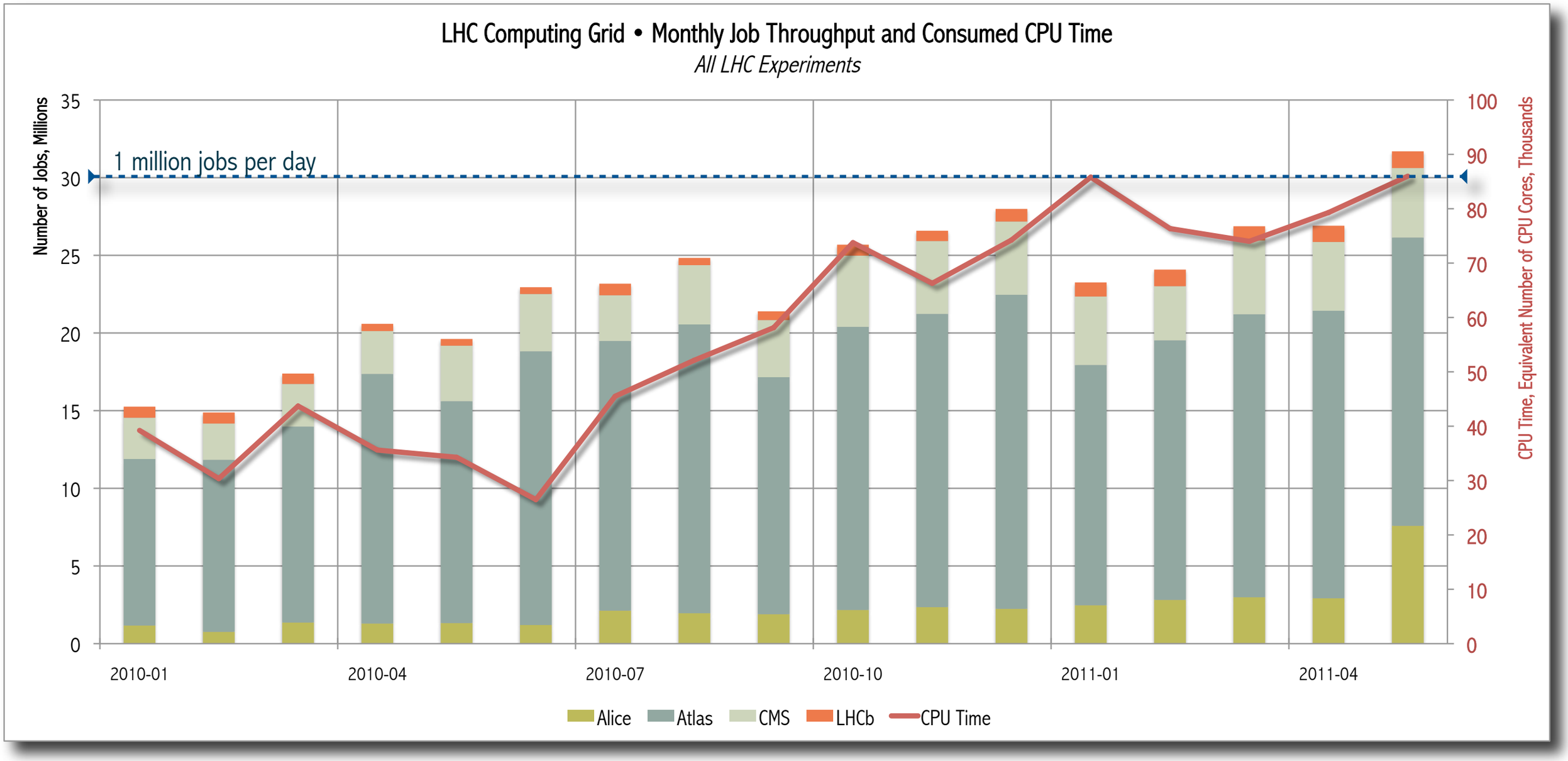


Rate

average aggregate exchange data rate: ***1.1 GB/sec***

Source: CMS PhEDEx

Grid usage: simulation + data processing



Source: EGEE Accounting Portal

Experiment-specific higher-level layers

- Experiments have developed and deployed higher-level layers on top of the infrastructure-level grid services

  Goal: to hide the "grid plumbing" to the end-user and add experiment-specific logic

- Examples

  data placement

  data and meta-data cataloguing

  dataset bookkeeping systems

  job management & workflow engines

  monitoring & alerting

  site status monitoring

  long-lived agents

  network link commissioning

  accounting

  …

EGI & WLCG Operations Tools

Site registry: GOCDB
RAL (UK)

European Grid Infrastructure Headquarters
EGI (NL)

Ticketing system:GGUS
KIT (DE)

Dashboard, Monitoring, GridMap, Security
CERN (CH)

EGI Operations portal
CC-IN2P3 (FR)

Resource Usage: EGEE Accounting Portal
CESGA (ES)

Central tools complemented by experiment-specific tools and procedures and by the operations tools deployed by each one of the national grid infrastructures in Europe and by OpenScienceGrid in USA

Achievements & lessons learned

- Global grid infrastructure is a reality and is delivering

  Used routinely by thousands of users for processing LHC data, including analysis

  Allows experiments to deliver physics results short time after data taking

  Same platform used also by other sciences, albeit at a smaller scale

- Network traffic higher than initially expected

  Network is very reliable: redundancy is key

- A priori data placement does not scale well

  Jobs sent to site hosting the data suppose multiple copies of the same data around the world

  Lot of data never used: refreshing all those caches generates a lot of network traffic and load on storage services

- Balance between centralization and distribution

  Network reliability makes centralization (with adequate redundancy) an easier to manage option than full distribution

- Complexity and sustainability

  Ongoing effort for support is high for experiments, sites and grid infrastructure operations

Adapted from: Ian Bird, LCG-France Meeting, May 30/2011

Perspectives

- Evolution of the computing models

  *from a priori to **dynamic data placement** based on popularity: popular data is replicated when jobs are sent to a site — unused data is removed from cache*

- Use remote WAN I/O

  ***read** a file remotely over the long distance network*

  ***download** a missing file from a dataset when needed*

- Site interconnection evolution

  *use of open **network exchange points** to allow tier-2s and tier-3s to exchange data among them and with tier-1s*

  do not overload the general research and education network with LHC data

Perspectives (cont.)

- Efficient usage of available resources

  Available resources starting to constraint experiments

- Evolution of hardware building blocks

  Many-cores machines, GPUs

- Cloud technology & virtualization

谢谢您

fabio hernandez
fabio@in2p3.fr
CAS/IHEP Computing Centre
CNRS/IN2P3 Computing Centre

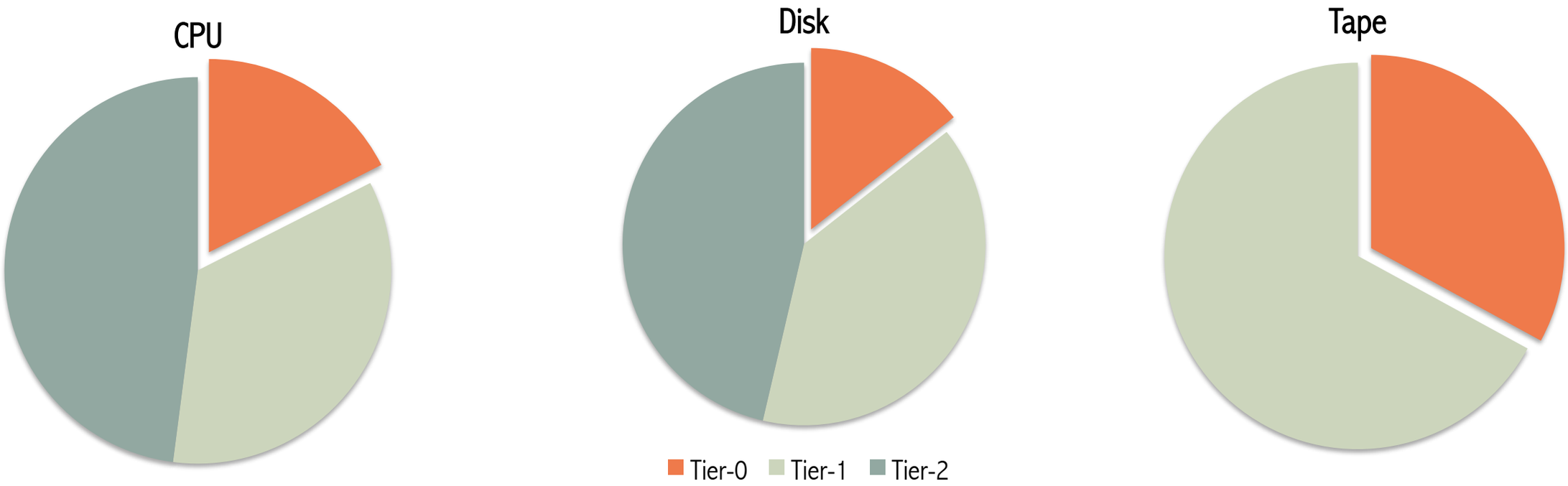
Questions & Comments

fabio hernandez
fabio@in2p3.fr
CAS/IHEP Computing Centre
CNRS/IN2P3 Computing Centre

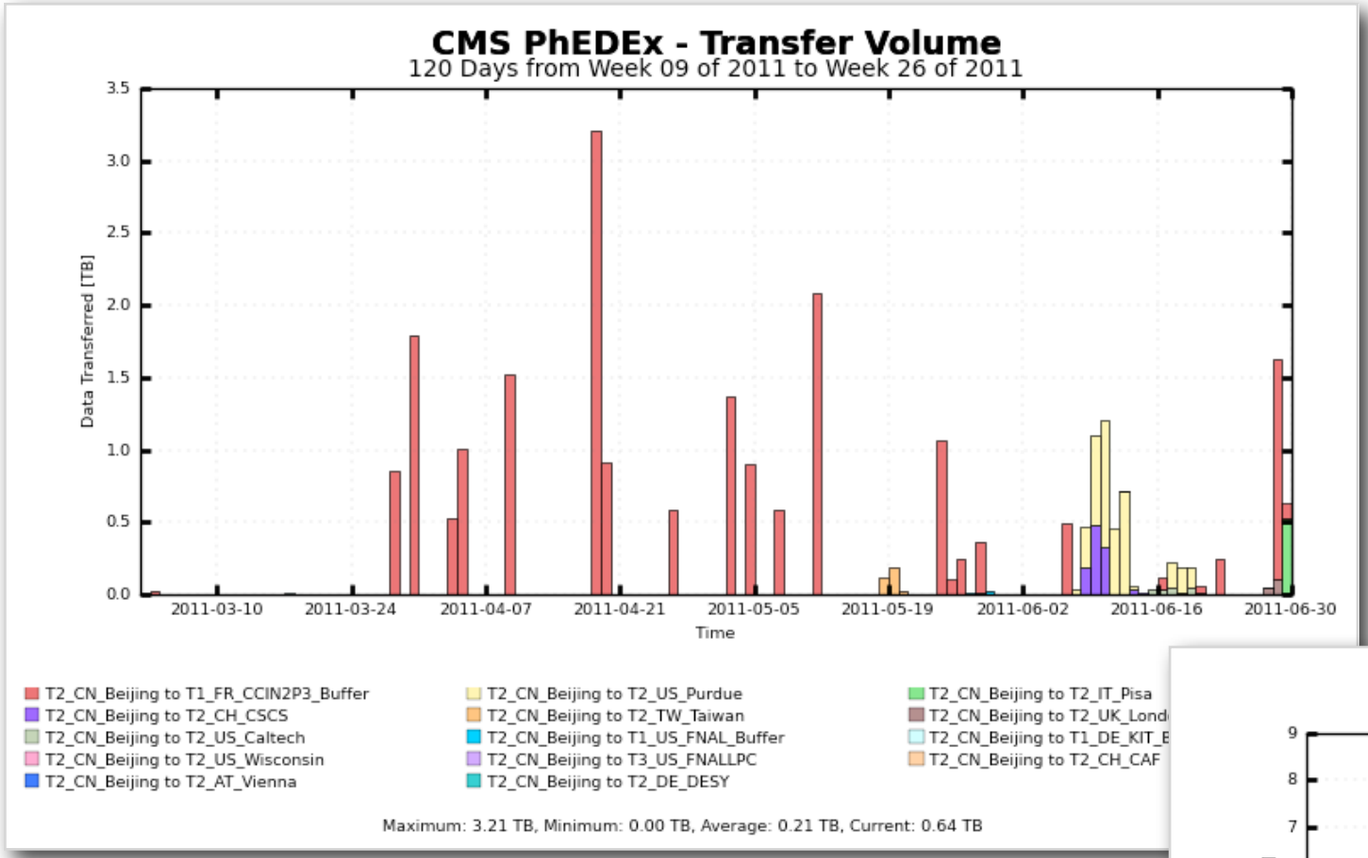
Backup Slides

fabio hernandez
fabio@in2p3.fr
CAS/IHEP Computing Centre
CNRS/IN2P3 Computing Centre

Resource distribution

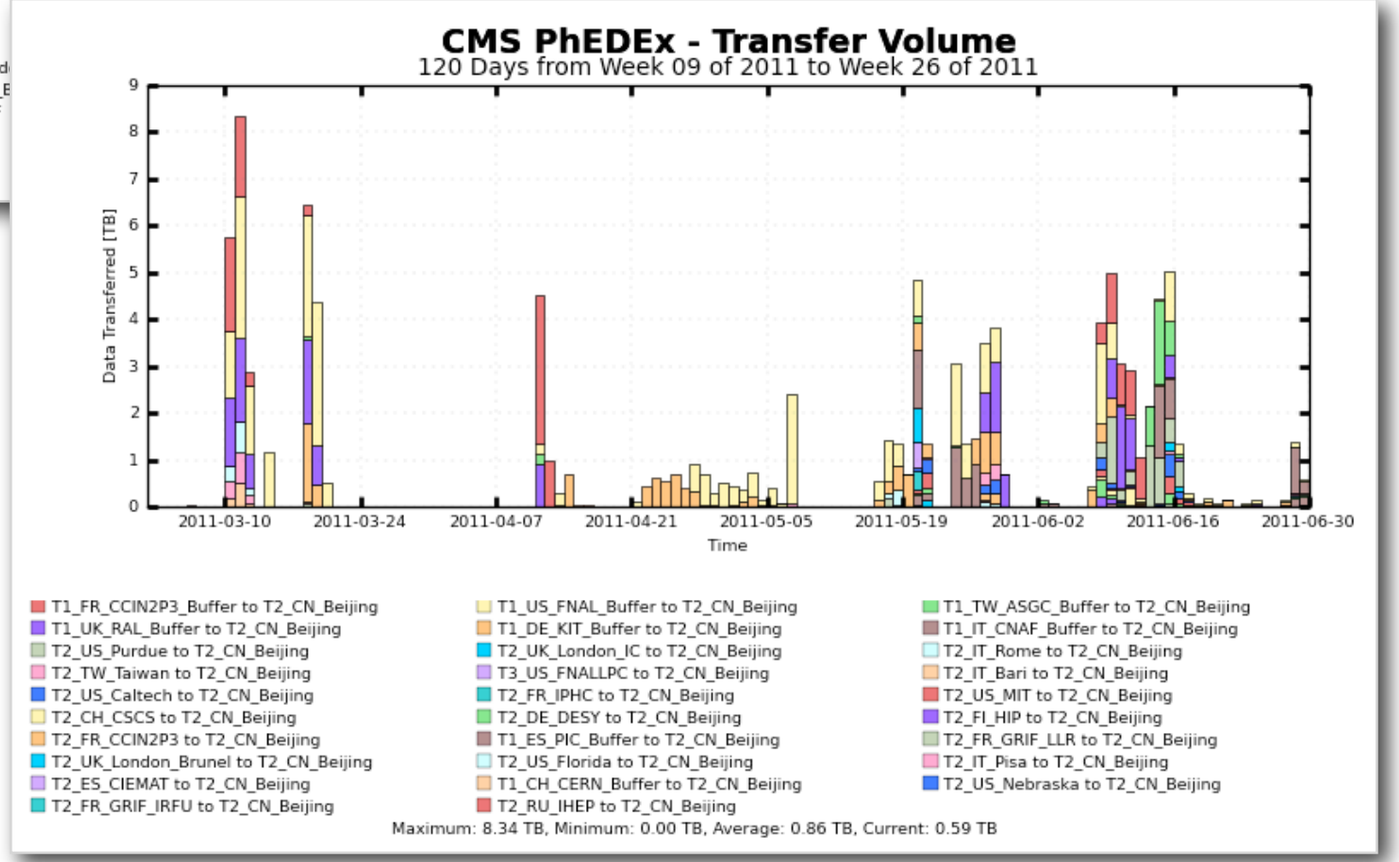


CMS data distribution: IHEP

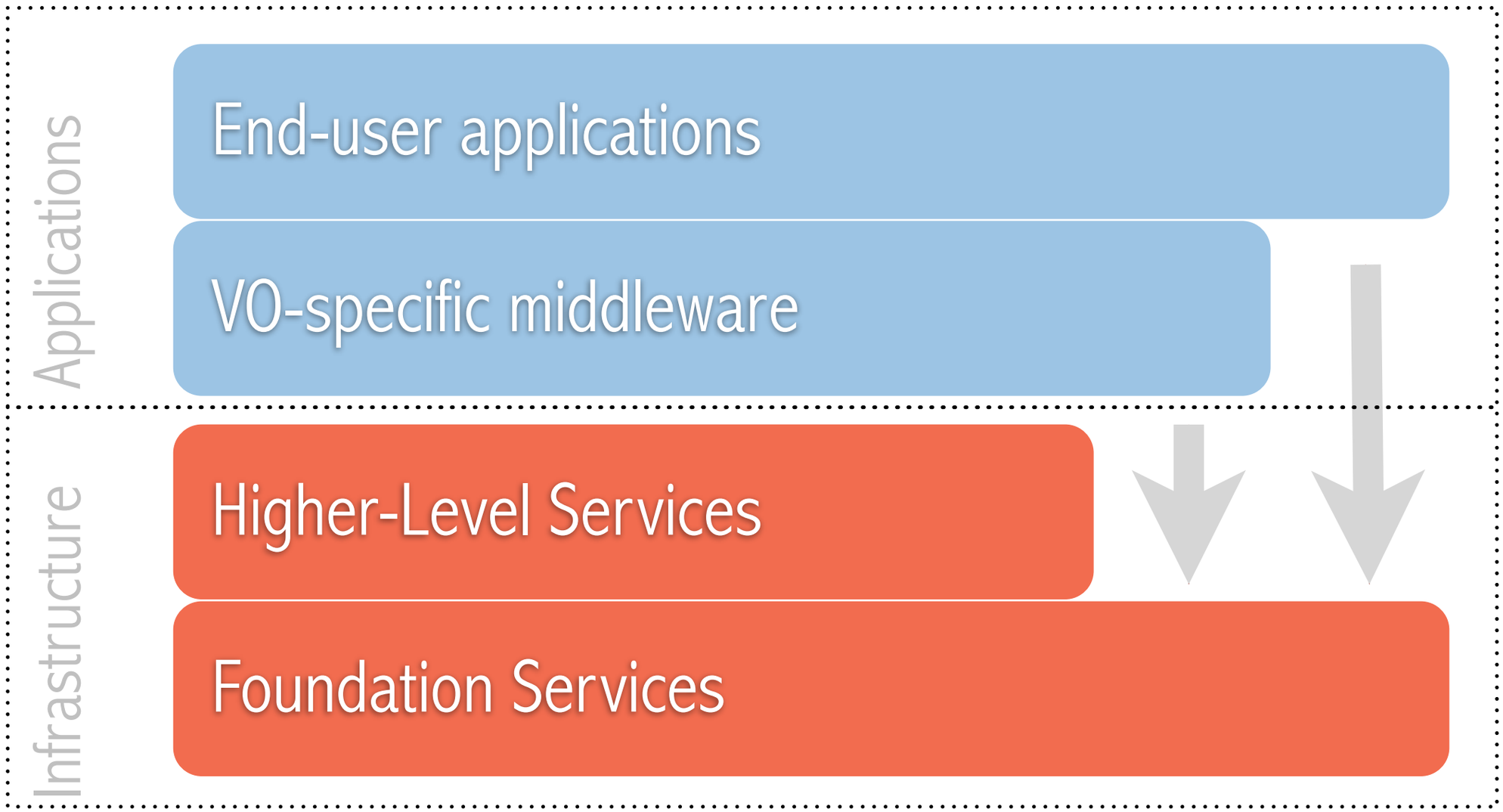


IHEP⇒other sites

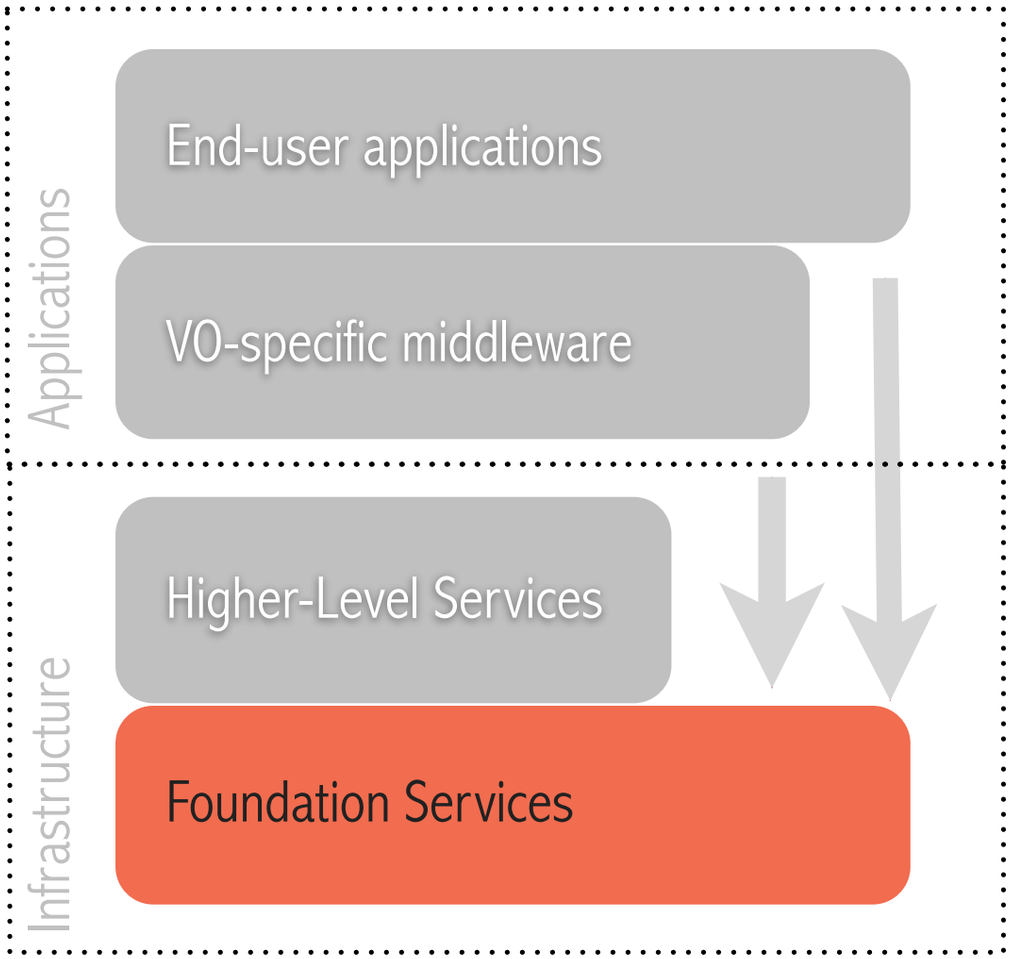
other sites⇒IHEP



Grid middleware



Grid middleware (cont.)



- Authentication

  individuals, hosts and services use X.509 certificates issued by accredited certification authorities

- Authorization

- Virtual organization membership

  provides information on the user's relationship with his/her virtual organization

- Computing element

  remote job submission to the site's batch system

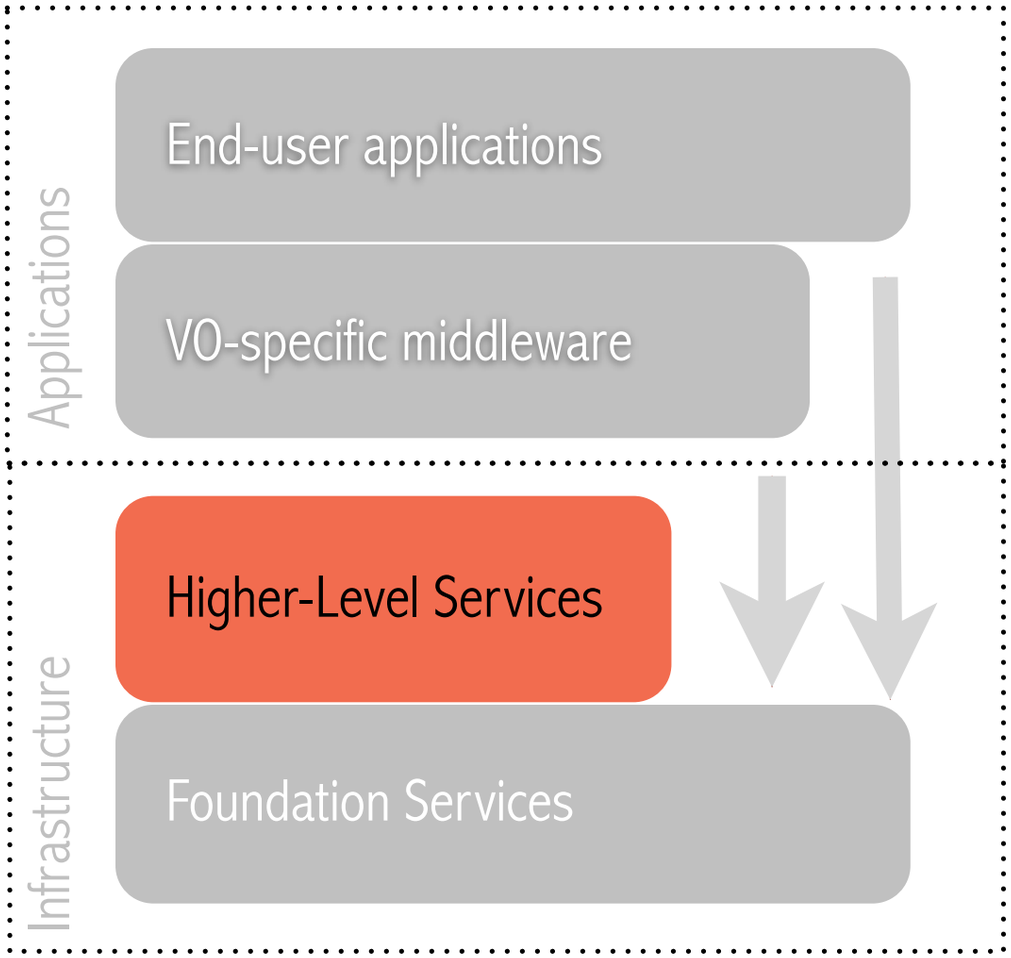
- Storage element

  inter-site file transfer to and from disk- and tape-based storage

- Information system

- Accounting

Grid middleware (cont.)



- Workload management

  job scheduling

- Data management

  scheduled file transfers

  file replica management

  meta-data management

- Virtual organization software installation