The LHC Computing Grid

The Computing Environment for LHC Data Analysis

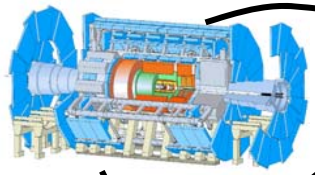
CHIPP Phoenix Cluster Inauguration

Manno, Switzerland

30 May 2008

Les Robertson
IT Department - CERN
CH-1211 Genève 23

Data Handling and Computation for Physics Analysis



detector

event filter (selection & reconstruction)

reconstruction

raw data

event summary data

processed data

event reprocessing

batch physics analysis

analysis

analysis objects
(extracted by physics topic)

event simulation

simulation

interactive physics analysis

les.robertson@cern.ch

The Computing System for LHC DATA ANALYSIS

Capacity and Evolution

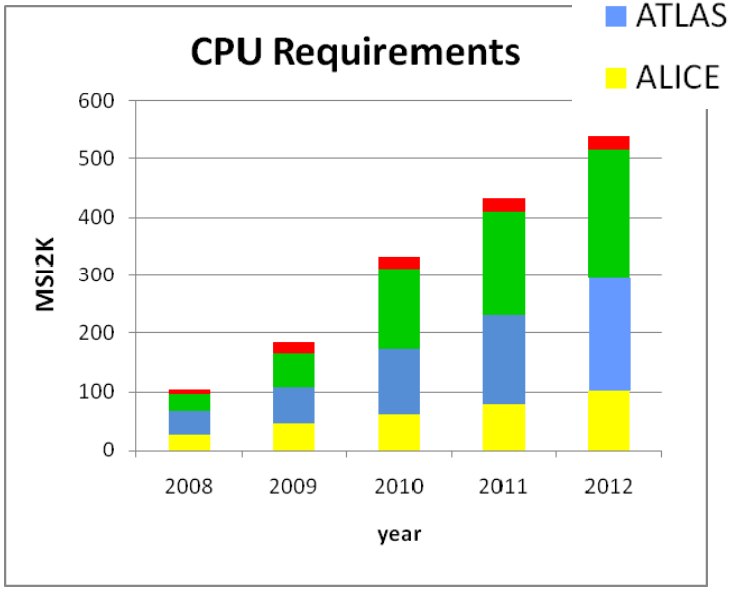
Computing requirements for all four experiments in first full year (2009)

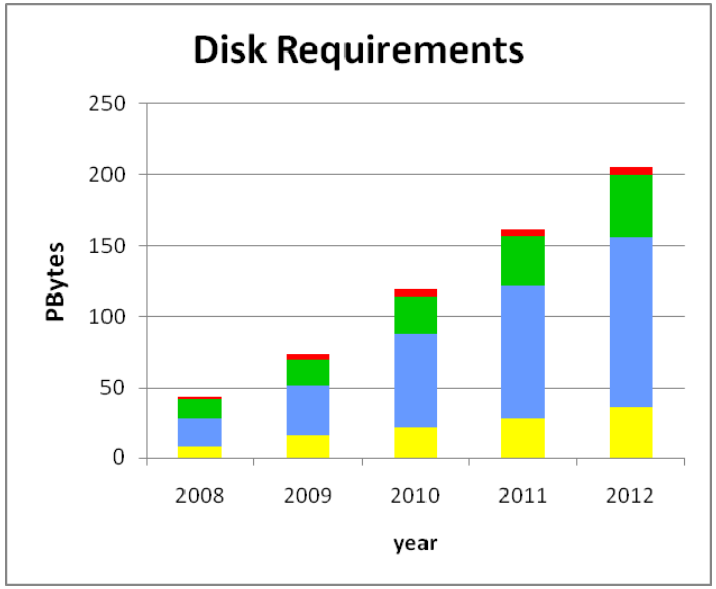
- **~70 PetaBytes disk**
- **~100K processor cores (2009)**
- **used by > 5,000 scientists & engineers**

Growth driven by new data, accelerator enhancements and improving efficiency, new analysis techniques, expanding physics scope, ...

- **disk storage growing at ~40 PB/year**

BUT – evolution of access patterns unclear





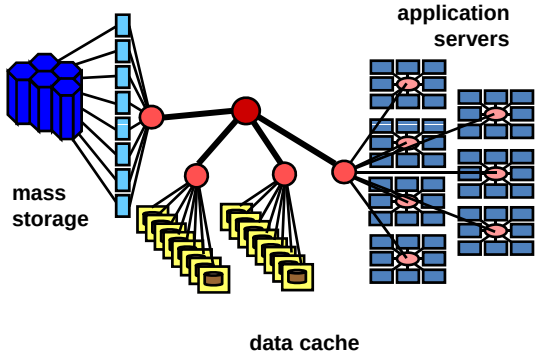
les.robertson@cern.ch

The key characteristics of experimental HEP data analysis that dictate the design of the computing system

- independent events
  → **easy parallelism**
- codes have
  -- modest memory needs (~2GB)
  -- modest floating point content
  → **perform well on PCs**

BUT --

- enormous data collections
  → **PetaBytes of new data every year**
- **shared by very large user collaborations**, many different groups, independent approaches to analysis
  → **unpredictable data access patterns**



- a simple distributed architecture developed ~1990 enabled experimental HEP to migrate from supercomputers and mainframes to clusters
- with the flexibility for easy evolution to new technologies
- and benefit from the mass market driven growth in the performance and capacity of PCs, disks, and local area networking

Why did we decide on a geographically distributed computing system for LHC?

- CERN's budget for physics computing was insufficient
- Easy parallelism, use of simple PCs, availability of high bandwidth international networking .... make it ***possible*** to extend the distributed architecture to the wide area ....

AND

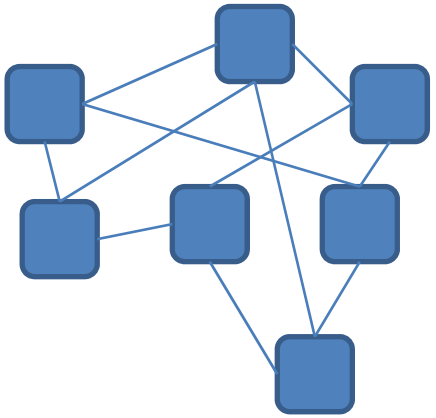
- The ~5,000 LHC collaborators are distributed across institutes all around the world with access to local computing facilities, ...
  ... and funding agencies prefer to spend at home if they can
- Mitigates the risks inherent in the computing being controlled at CERN, subject to the lab's funding priorities and with access and usage policies set by central groups within the experiments

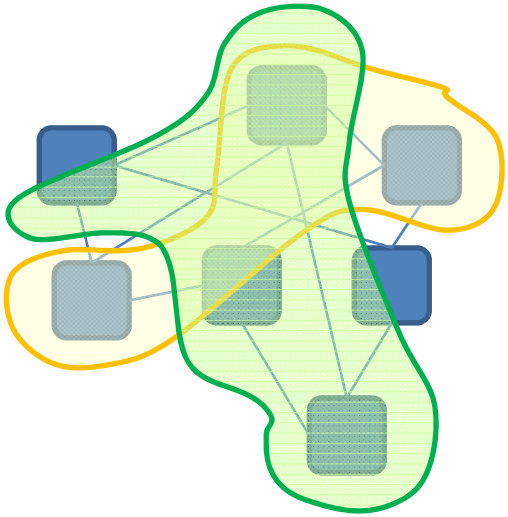
ALSO

- Active participation in the LHC computing service gives the institute (not just the physicist) a continuing and key role in the data analysis
  -- which is where the physics discovery happens
- Encourages novel approaches to analysis ....
  ... and to the provision of computing resources

What do we mean by a Computing Grid**?

- Collaborating computing centres
- Interconnected with good networking
- Interfaces and protocols that enable the centres to advertise their resources and exchange data and work units
- Layers of software that hide all the complexity from the user
- So the end-user does not need to know where his data sits and where his jobs run
- The Grid does not itself impose a hierarchy or centralisation of services
- Application groups define ***Virtual Organisations*** that map users to subsets of the resources attached to the Grid





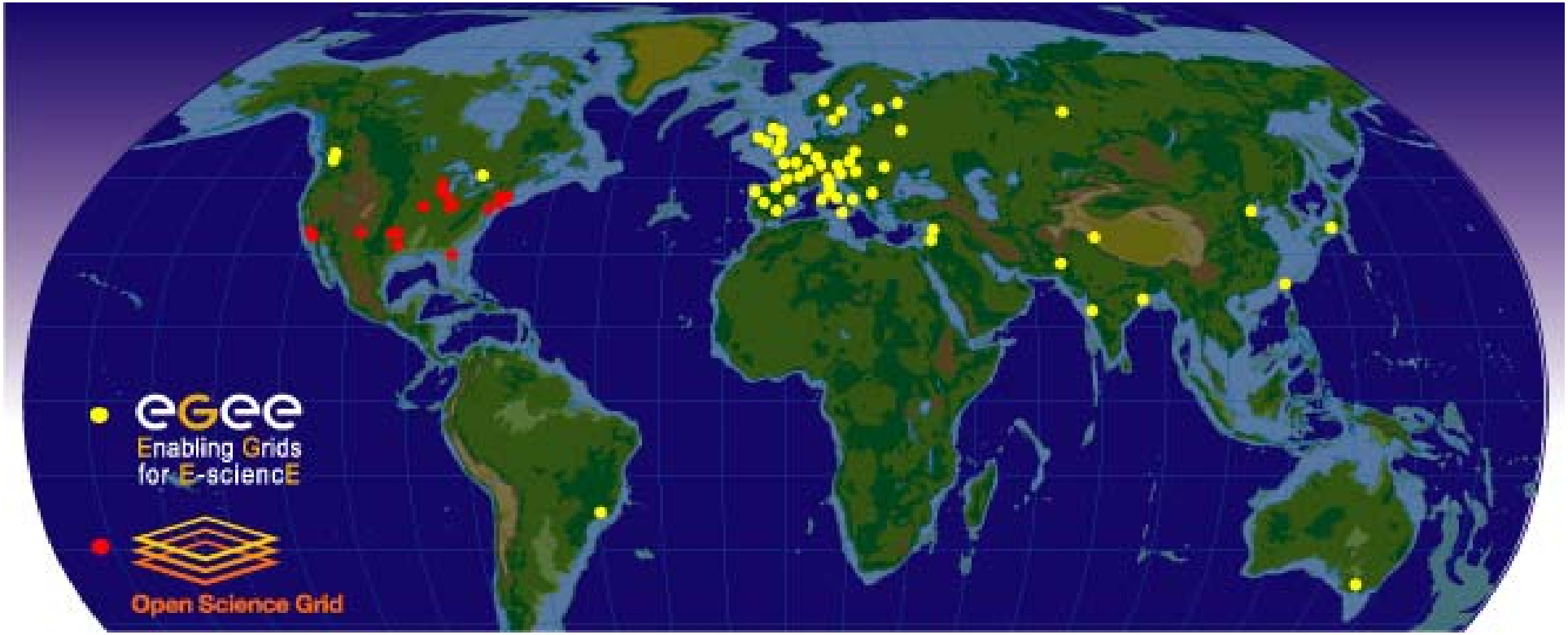
** *There are many different variations on the term Grid – this is the HEP definition*

- The advantage for the *computer centre* is that the basic services can be provided in a standard way for different application groups
  - e.g. user authentication, job submission, storage access, data transfer…
  - ATLAS, CMS, LHCb, DZERO, …., BioMed, Fusion, ….
- The advantage for the *application group* is that it can integrate resources from different centres and view them as a single service without having to support all of the software layers, negotiate the installation of special software, register users on each site, etc.
- But they have the *flexibility* to pick and choose – replace software layers with their own products, decide which services are provided at which sites, ….

LCG depends on two major science grid infrastructures

EGEE - Enabling Grids for E-Science *(with EU funding)*

OSG - US Open Science Grid *(with DoE and NSF funding)*



The Middleware for the Baseline Services needed for the LHC Experiments

- Information system
- Security framework
- Storage Element
- SRM interface to Mass Storage
  dCache, DPM, CASTOR, STORM
- Basic data transfer tools – Gridftp, srmCopy.
- Reliable file transfer service – FTS
- Catalogue services – LFC, Globus RLS
- Catalogue and data management tools – lcg-utils
- Compute element – Globus/Condor-G based CE, Cream (web services)
- Reliable messaging service
- Virtual Organisation Management Services
- Database distribution services – ORACLE streams, SQUID
- POSIX-I/O interfaces to storage
- Workload Management – EGEE Resource Broker, VO-specific schedulers
- Job monitoring tools
- Grid monitoring tools
- Application software installation
- GUIs for analysis, production GANGA, CRAB, PANDA, ..

For LCG, grid interoperability is required at the level of the baseline service
→ same software or standard interfaces or compatible functionality

The Middleware for the Baseline Services needed for the LHC Experiments

- Information system
- Security framework
- Storage Element
- SRM interface to Mass S
  dCache, DPM
  STORM
- Basic data tr
  Gridftp, s
- Reliable file t
  FTS
- Catalogue services –
  LFC, Globus RLS
- Catalogue and data management tools – lcg-utils

GRID – a simple idea – Not so simple to implement it!

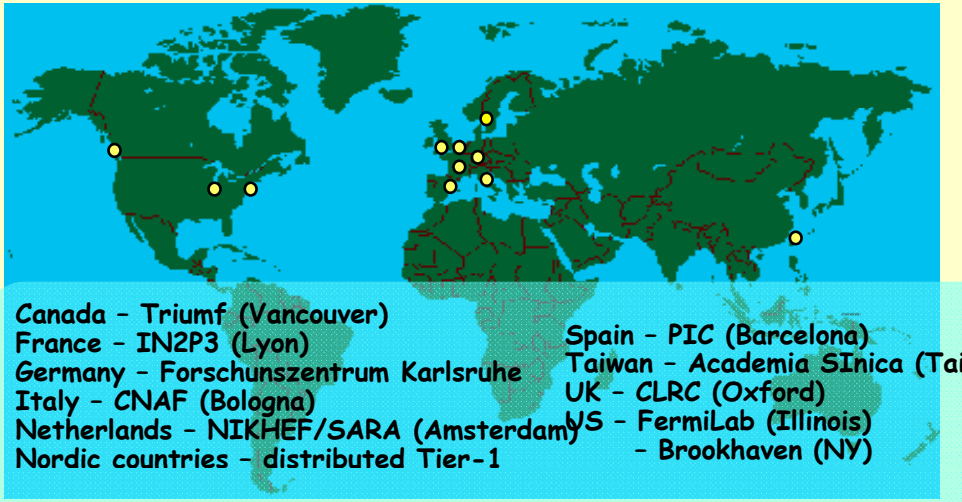
- Compute element –
  Globus/Condor-G based CE,
  Cream (web services)
- messaging service
- Organisation Management
  ces
- distribution services –
  CLE streams, SQUID
- POSIX-I/O interfaces to storage
- Workload Management –
  EGEE Resource Broker,
  VO-specific schedulers
- Job monitoring tools
- Grid monitoring tools
- Application software installation
- GUIs for analysis, production
  GANGA, CRAB, PANDA, ..

For LCG, grid interoperability is required at the level of the baseline service
→ same software or standard interfaces or compatible functionality

The LHC Computing Grid - A Collaboration of 4 Experiments + ~130 Computer Centres

Tier-0 – the accelerator centre

- Data acquisition & initial processing
- Long-term data curation
- Distribution of data → Tier-1 centres



Canada – Triumf (Vancouver)
France – IN2P3 (Lyon)
Germany – Forschunszentrum Karlsruhe
Italy – CNAF (Bologna)
Netherlands – NIKHEF/SARA (Amsterdam)
Nordic countries – distributed Tier-1
Spain – PIC (Barcelona)
Taiwan – Academia SInica (Taipei)
UK – CLRC (Oxford)
US – FermiLab (Illinois)
– Brookhaven (NY)

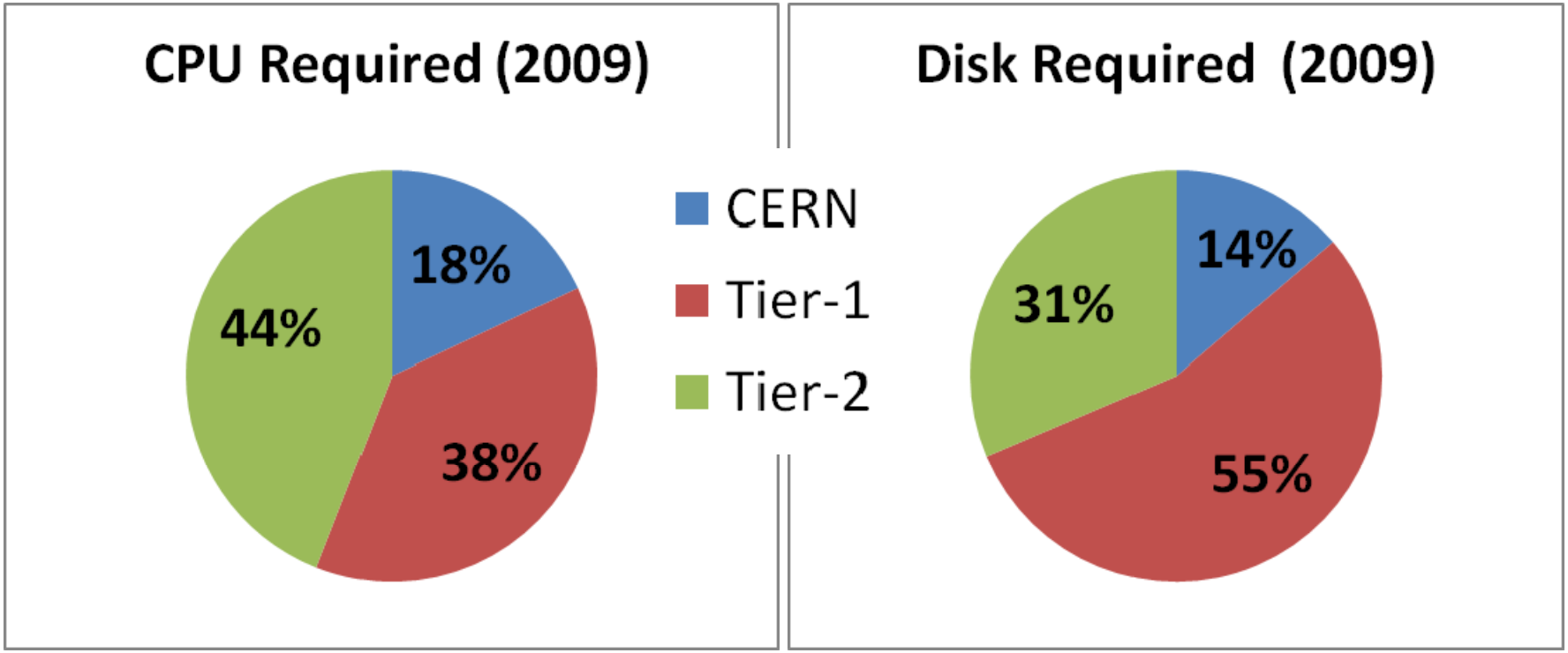
11 Tier-1 Centres – "online" to the data acquisition process → high availability

- Managed Mass Storage – → grid-enabled data service
- Data-heavy analysis
- National, regional support

Tier-2 – 120 Centres in 60 Federations in 35 countries

- **End-user (physicist, research group) analysis** – where the discoveries are made
- Simulation

Distribution of Resources across Tiers



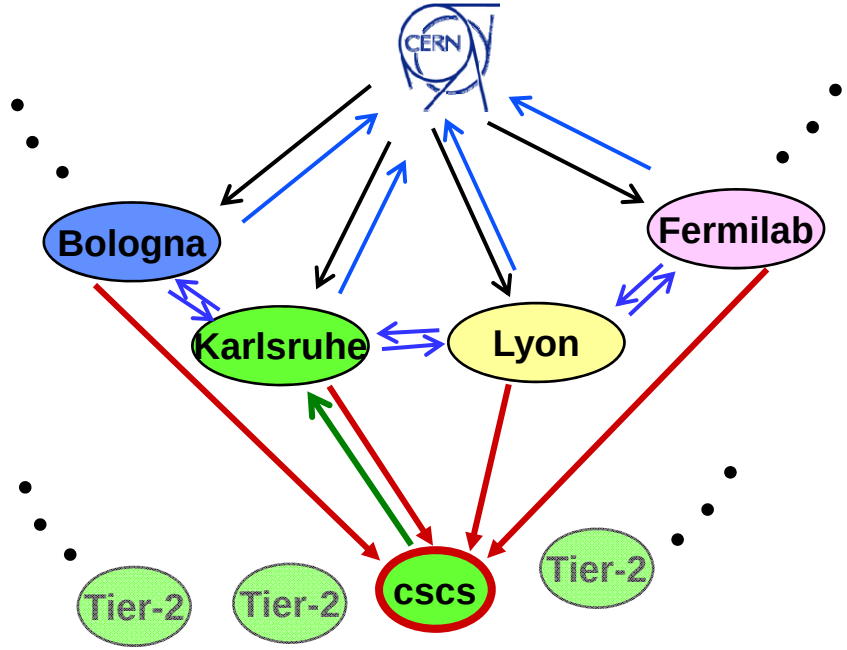
- 2009 – the first full year of data taking
- <20% at CERN – less than half of that at the Tier-2s

→ the distributed system must work from Day 1

Middleware & Software

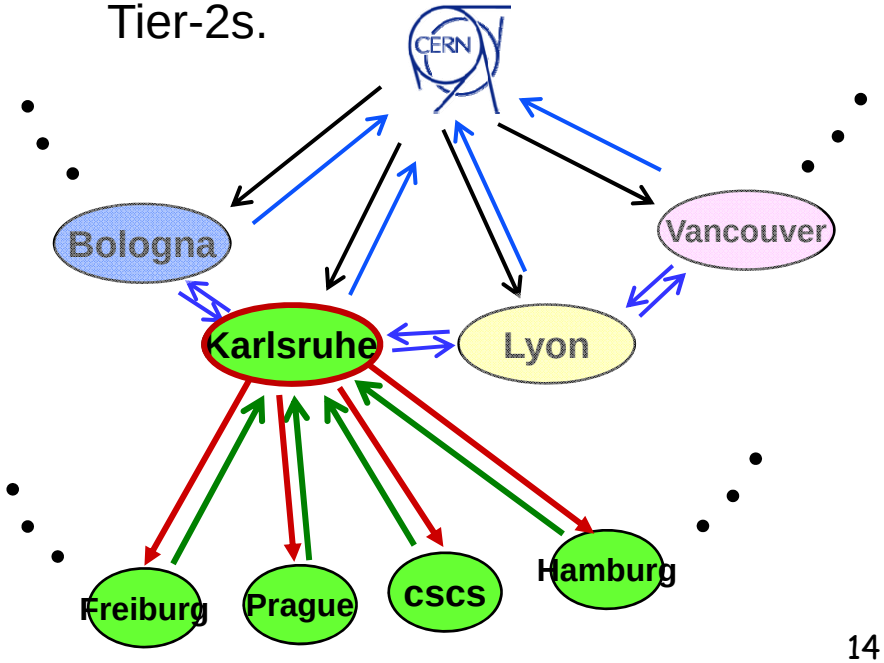
- From many sources: Globus, Condor, EGEE gLite, High Energy Physics common tools and packages, experiments, open-source projects, proprietary packages, …
- Two fundamental middleware packages are integrated, tested and distributed by the infrastructure grids
  - gLite by EGEE
  - built on the Virtual Data Toolkit by OSG
- The mass storage systems are crucial but complicated components of the LCG service – HEP developments
- And a thick layer of software is maintained and distributed by the experiments (data management, resource scheduling, analysis GUIs, ..)

Experiment computing models define specific data flows between CERN, Tier-1s and Tier-2s

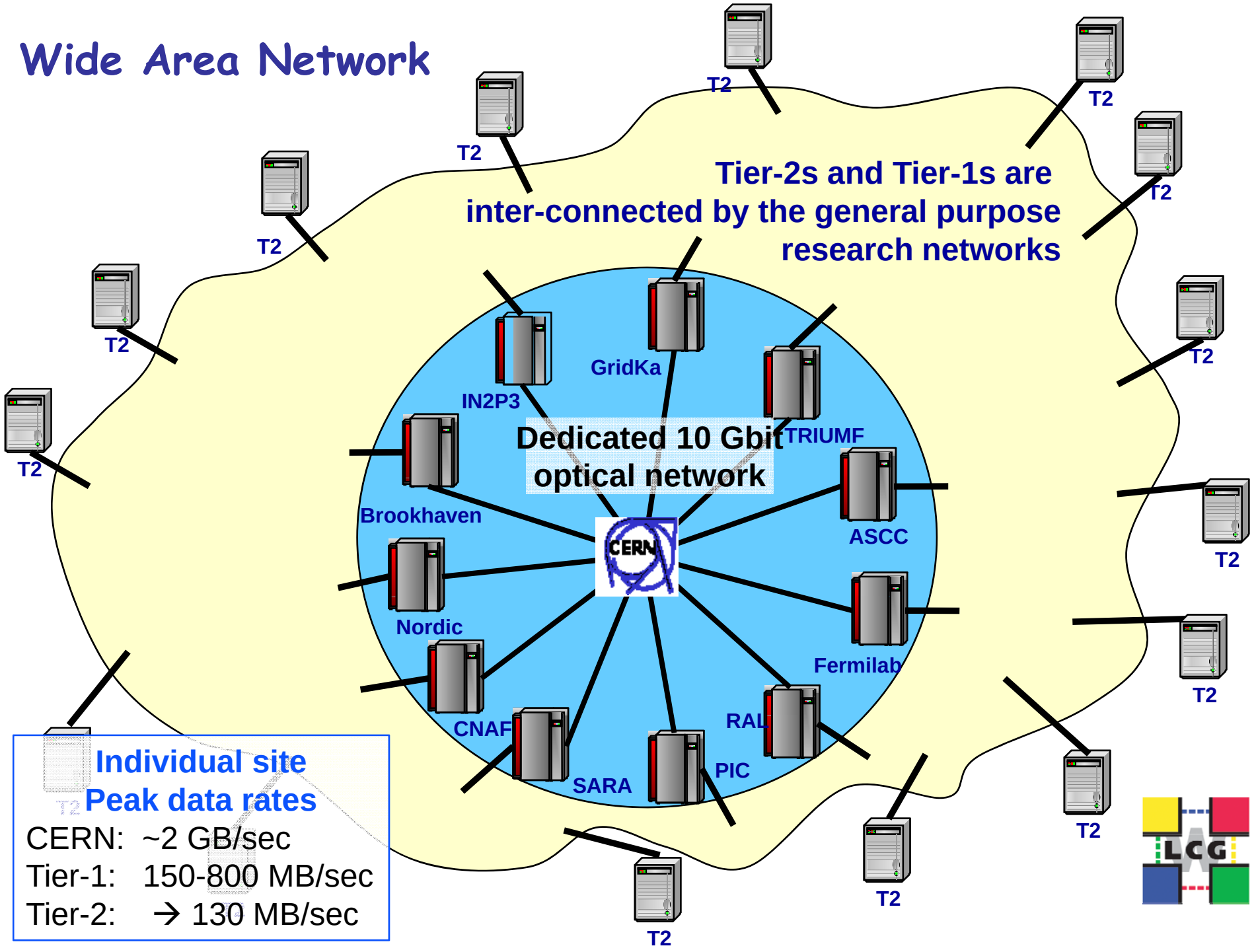


CMS – Tier-2s send simulated data to a specific Tier-1, but obtain data for analysis from any of the 7 Tier-1s: *Taipei, Bologna, Fermilab, Karlsruhe, Lyon, Barcelona, Ruther ford Lab*

ATLAS – EachTier-1 acts as the data repository for a "cloud" of associated Tier-2s.



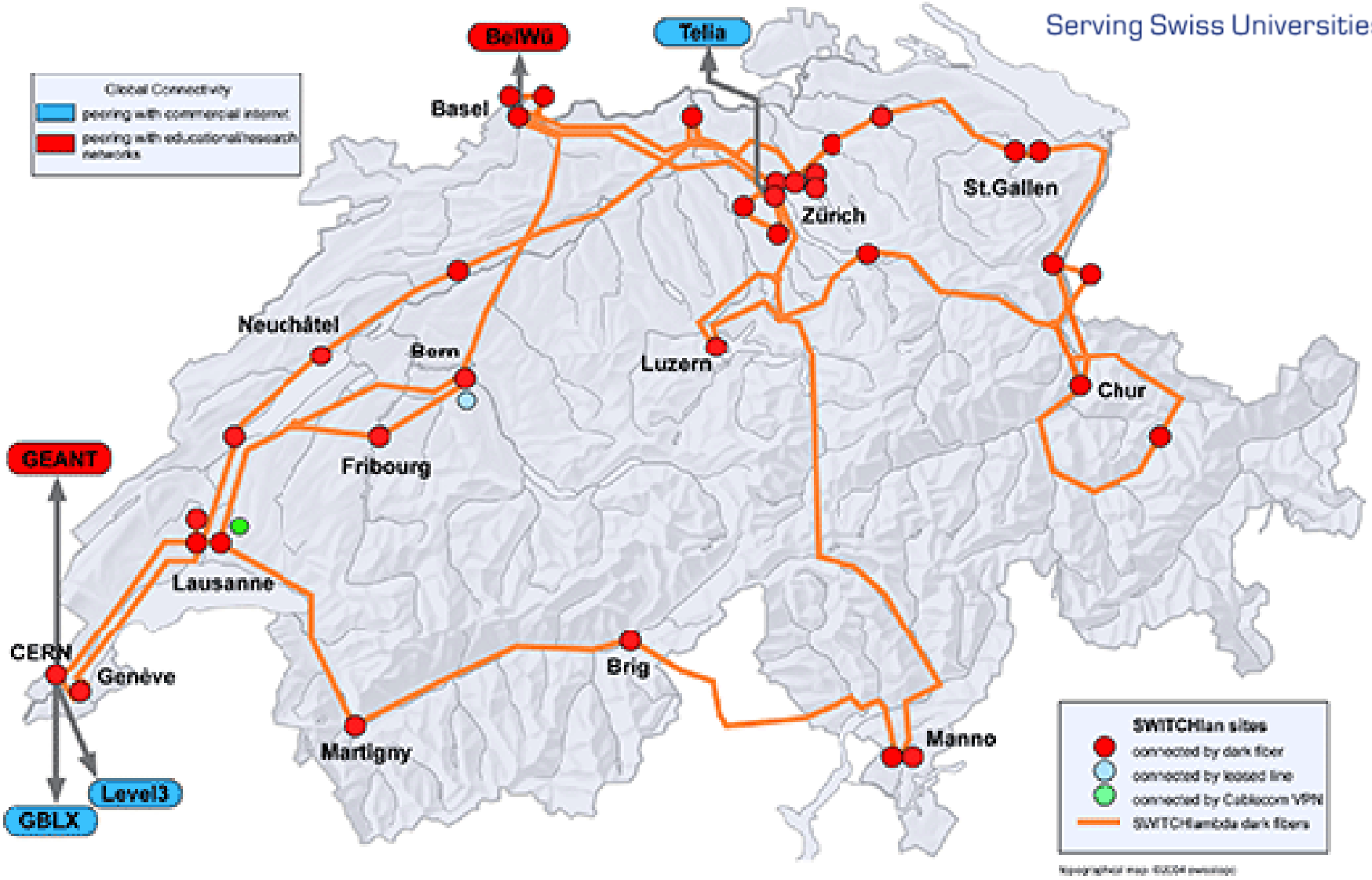
Wide Area Network

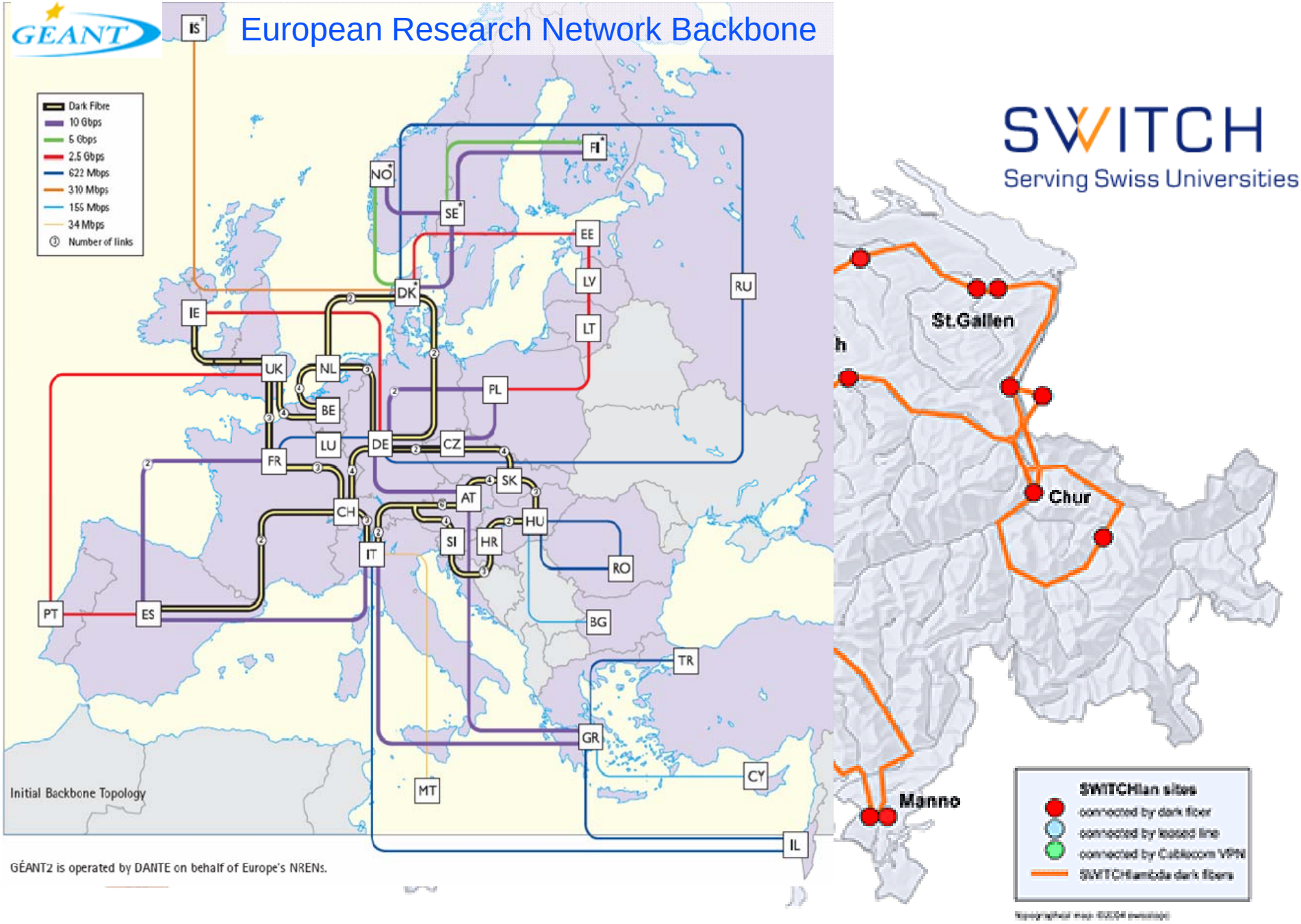


Individual site Peak data rates

CERN: ~2 GB/sec
Tier-1: 150-800 MB/sec
Tier-2: → 130 MB/sec

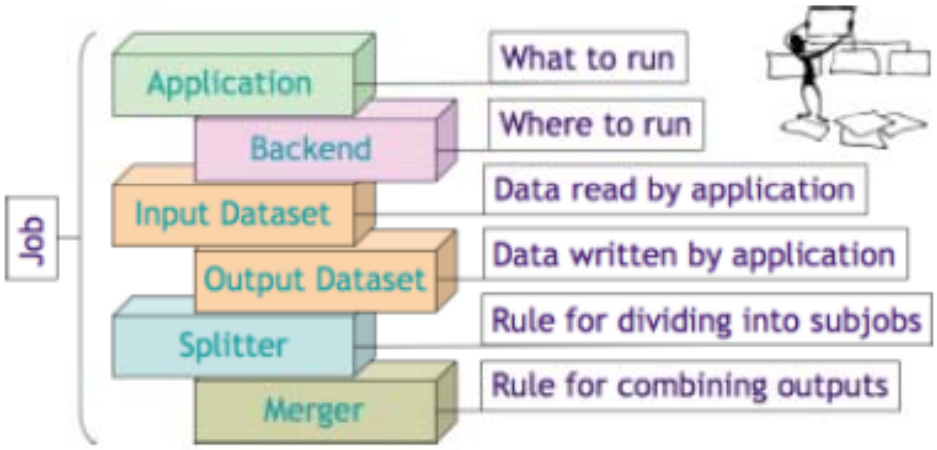
SWITCH
Serving Swiss Universities




GÉANT
European Research Network Backbone
Dark Fibre
10 Gbps
5 Gbps
2.5 Gbps
622 Mbps
310 Mbps
155 Mbps
34 Mbps
Number of links
IS
NO
SE
FI
EE
LV
LT
RU
DK
IE
UK
NL
PL
BE
LU
DE
CZ
FR
SK
AT
CH
HU
SI
HR
IT
RO
PT
ES
BG
TR
GR
CY
MT
IL
Initial Backbone Topology
GÉANT2 is operated by DANTE on behalf of Europe's NRENs.
SWITCH
Serving Swiss Universities
St.Gallen
Chur
Manno
SWITCHlan sites
connected by dark fiber
connected by leased line
connected by Cablecom VPN
SWITCHlambda dark fibers

GANGA

- Grid user interface: GANGA
  - Single tool for all Grid-based work (including analysis and small Monte Carlo productions)
  - Trivial switching between Grid running and local execution (for testing purposes)
  - Grid backends include LCG (EGEE), NorduGrid, OSG (Panda)
  - Interface via a command-line or a GUI

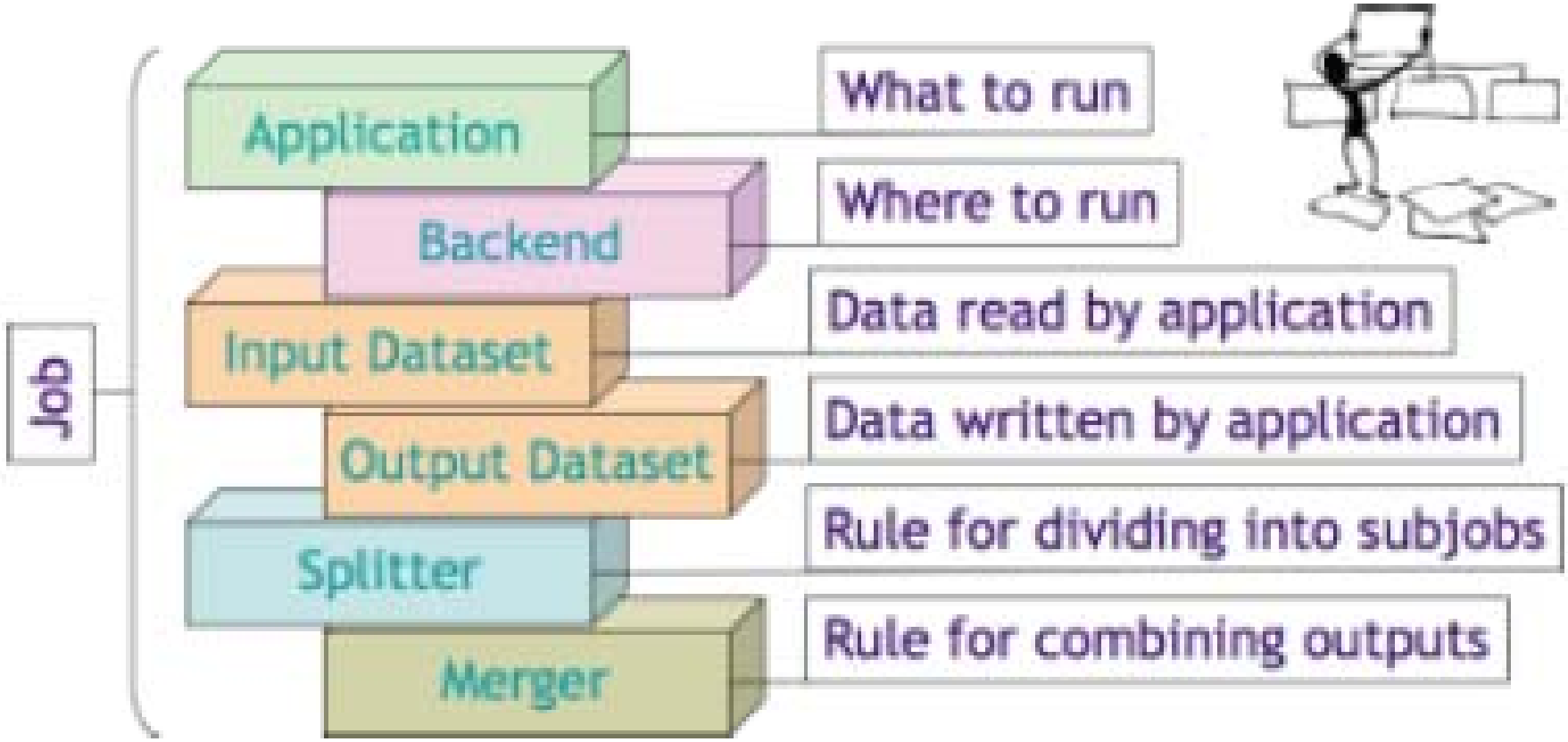


www.cern.ch/ganga

LANCASTER UNIVERSITY

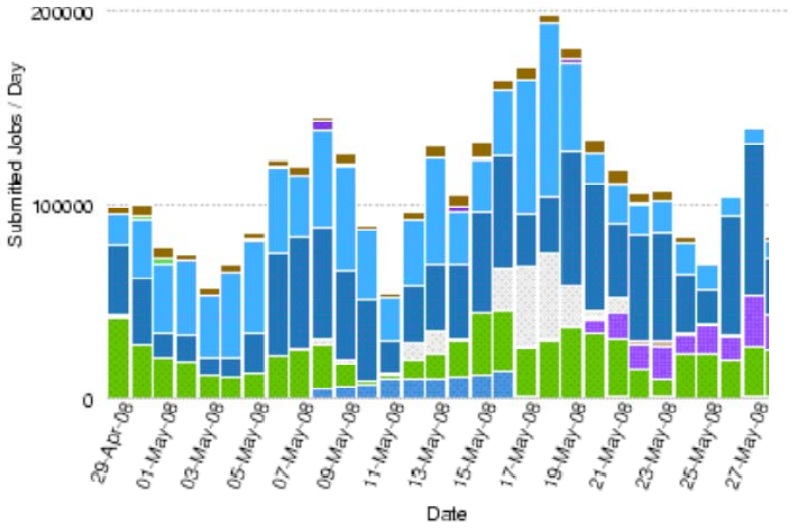
GANGA

- Grid user interface: GANGA
  - Single tool for all Grid-based work (including analysis and small Monte Carlo

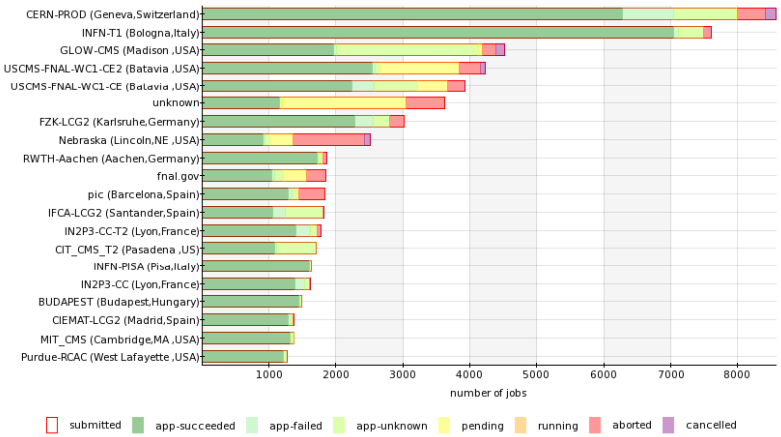


Experiment Dashboards provide tools for monitoring and debugging

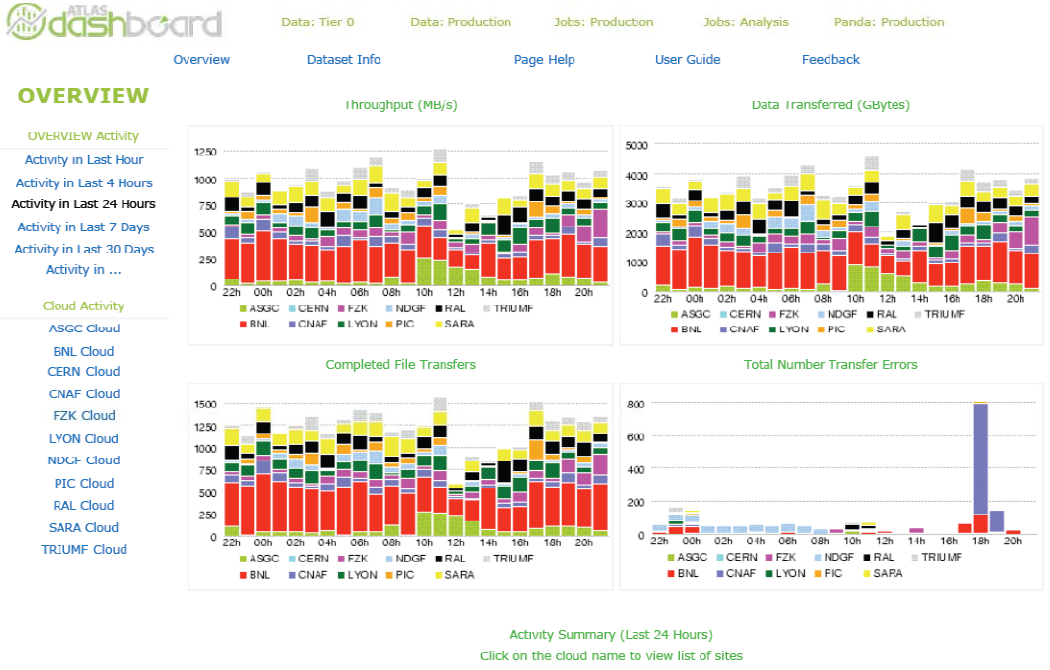
CMS jobs May 2008 sorted by activities
Up to 200K jobs per day - 35% end-user analysis



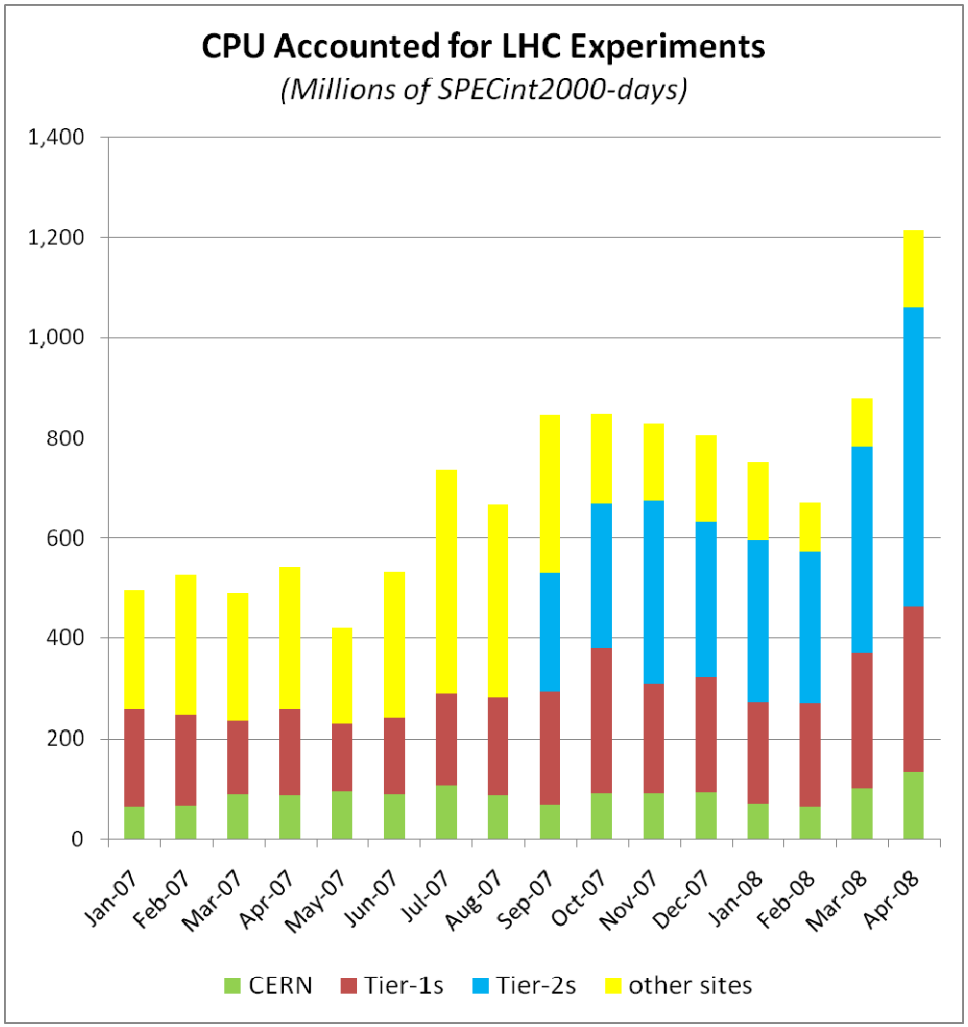
job status by site – user can drill down to find details of the errors



ATLAS data transfer status - 28 May 2008
Throughput ~1100 MB/s



Evolution of LHC grid cpu usage



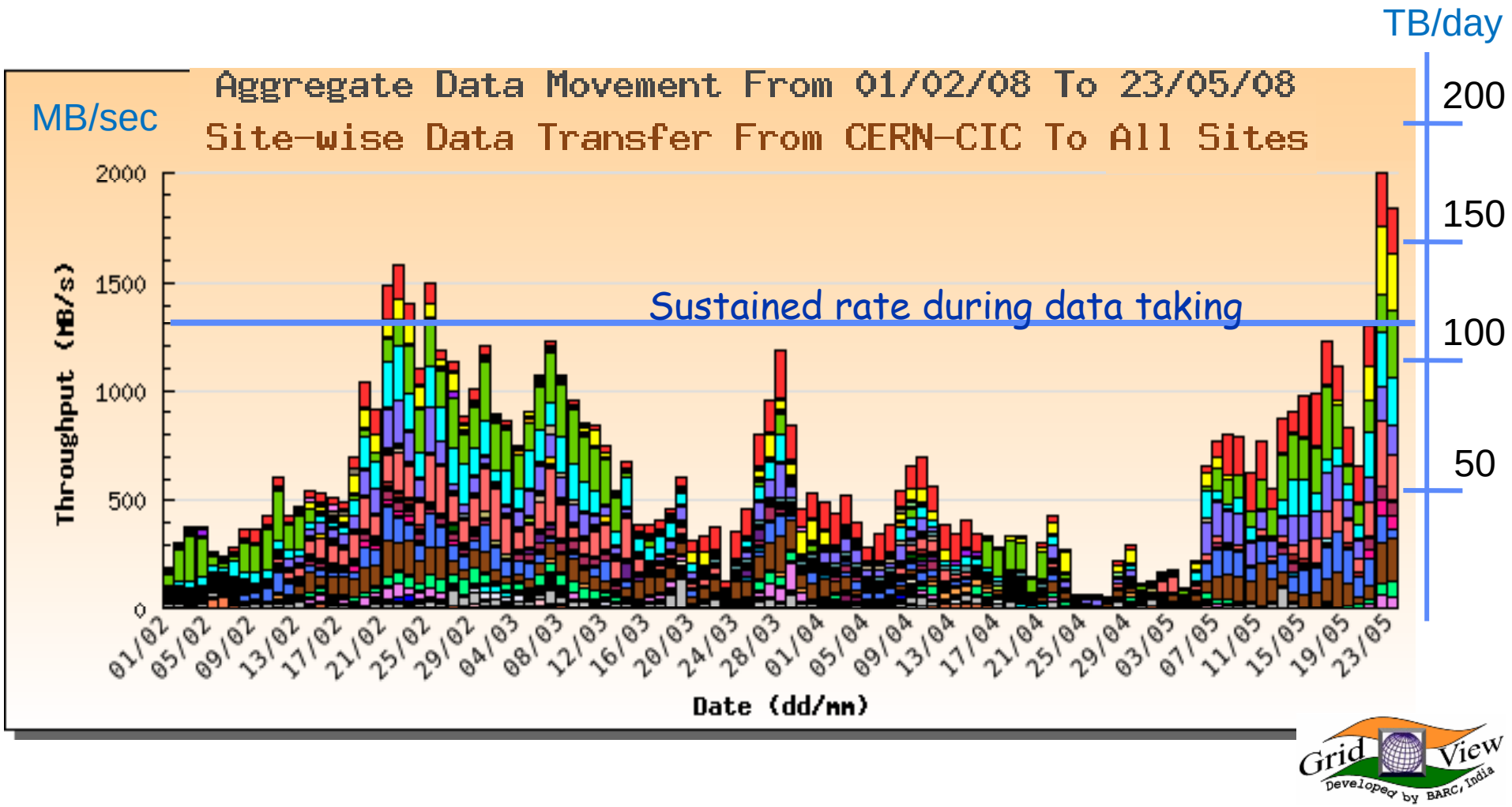
150% growth since Jan 07

More than half from Tier-2s

~800K core-days
~25K cores at 100% utilisation
~55% of committed capacity

Experiment services are still in test mode – awaiting the real data

Data Transfer CERN→Tier-1s



CMS Data Volume
TBytes/day - all sites to all sites

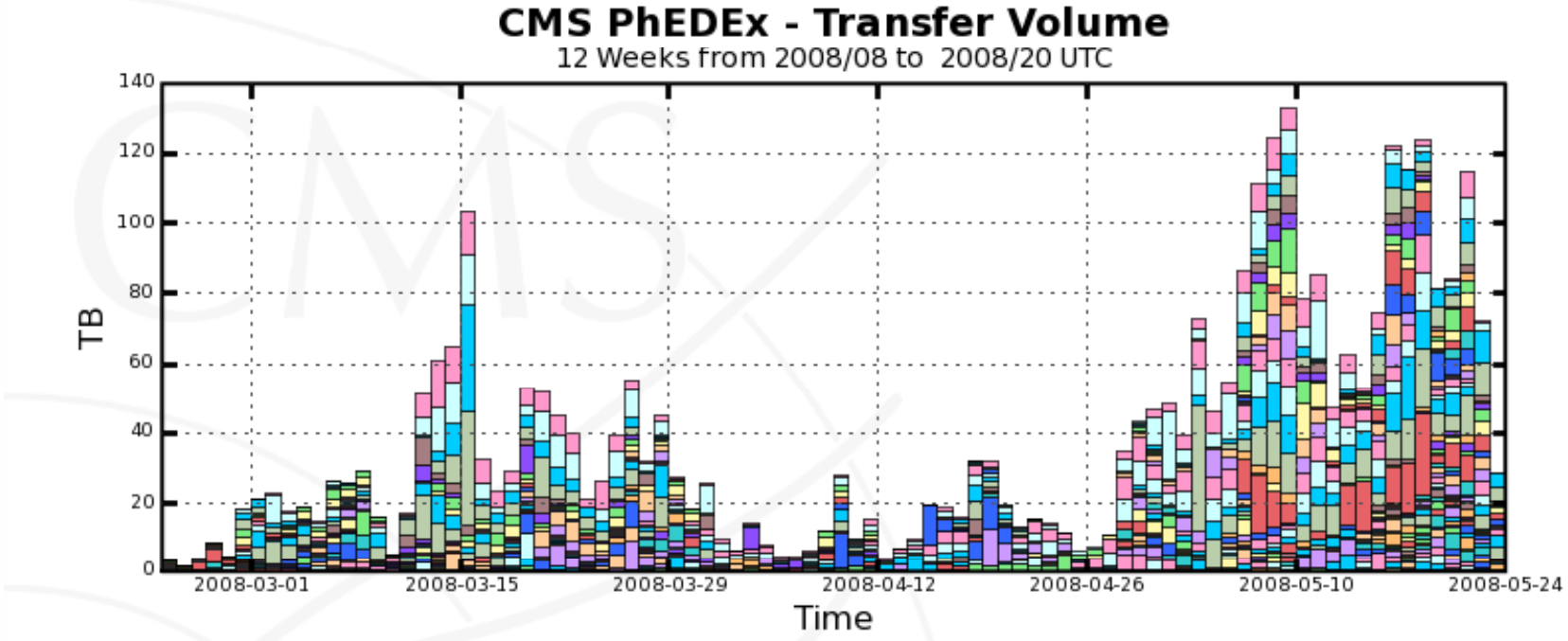
PhEDEx – CMS Data Transfers

Info | **Activity** | Data | Requests | Components | Reports

Rate | **Rate Plots** | Queue Plots | Quality Plots | Routing | Transfer Details | Deletions | Recent Errors

Graph Transfer Volume **by** Destination **filter source** **destination** show MSS nodes

Period Last 90 Days **up to** Update



LHC Computing Grid Status Summary

- The "final" pre-startup testing is now going on
  → all experiments exercise simultaneously their full computing chains - from the data acquisition system to end-user analysis at the Tier-2s – at the full 2008 scale
- No show-stoppers since before the tests began in February – day-to-day issues being rapidly resolved
- Most performance and scaling metrics have been achieved

BUT this is research ….

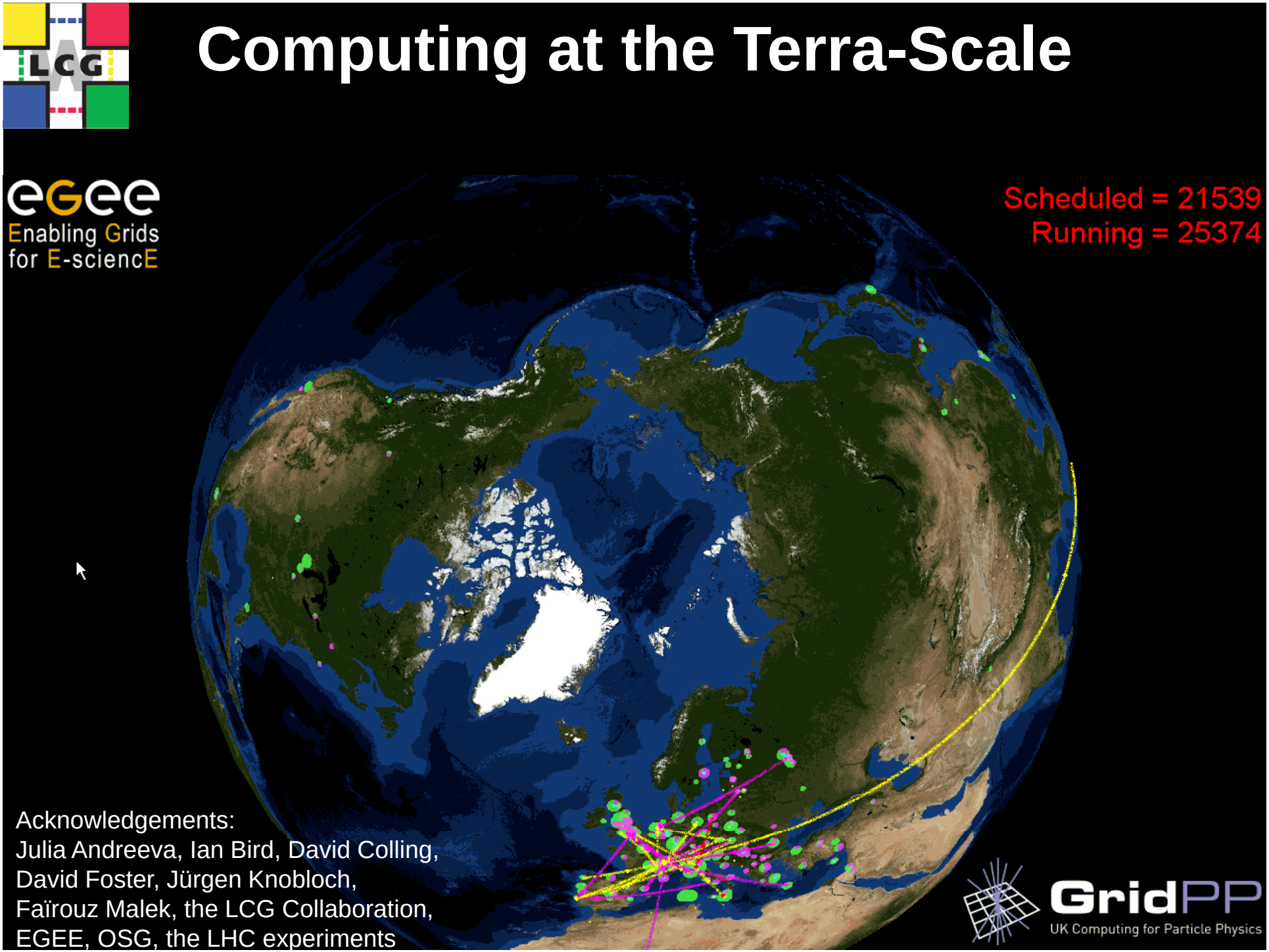
- The actual load, access patterns, user behaviour is unpredictable – depends on how physicists react to what they find in the real data
- We can look forward to an exciting time when the beam starts!!

Summary

- Grids are all about sharing
  - they are a means whereby groups distributed around the world can pool their computing resources
  - large centres and small centres can all contribute
  - users everywhere can get equal access to data and computation
  - without having to spend all of their time seeking out the resources
- Grids also allow the flexibility to place the computing facilities in the most effective and efficient places –
  - exploiting funding wherever it is provide,
  - piggy-backing on existing computing centres,
  - or exploiting cheap and renewable energy sources
- The LHC provides a pilot application –
  - with massive computing requirements, world-wide collaborations
  - that is already demonstrating that grids can deliver in production

and the scientific success of LHC will depend on the grid from day 1

Computing at the Terra-Scale



Acknowledgements:
Julia Andreeva, Ian Bird, David Colling,
David Foster, Jürgen Knobloch,
Faïrouz Malek, the LCG Collaboration,
EGEE, OSG, the LHC experiments