The European Data Grid Project

4th HNF-Europe Workshop

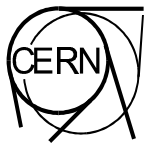
on

High Performance Networking

CERN, Geneva

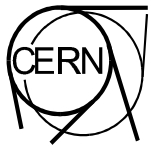
26-27 September 2001

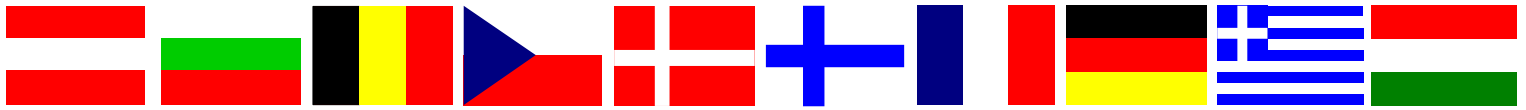
Ben Segal
CERN
Information Technology Division
B.Segal@cern.ch



Acknowledgements

The choice of material presented is my own, however a lot of the material has been taken from presentations made by others, notably Leanne Guy, Les Robertson and Manuel Delfino.

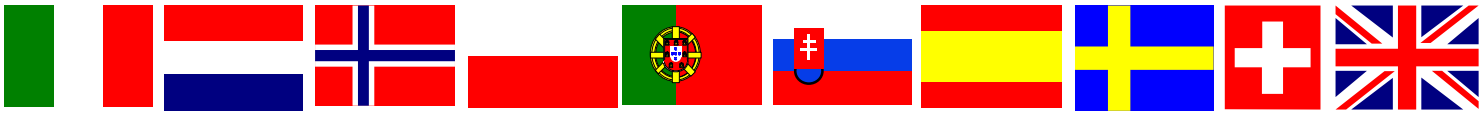




CERN - The European Organisation for Nuclear Research

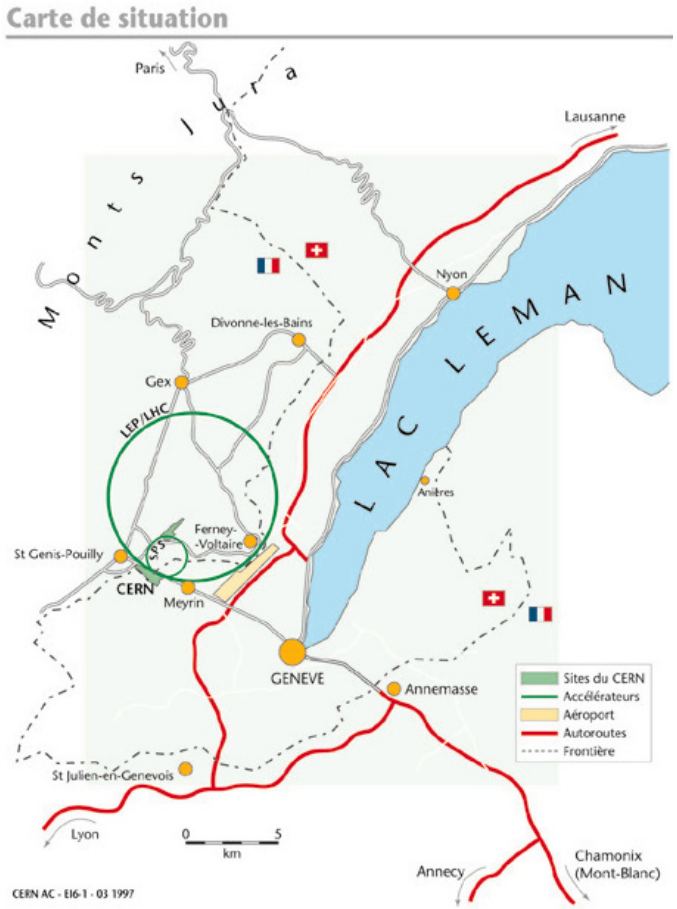
The European Laboratory for Particle Physics

- Fundamental research in particle physics
- Designs, builds & operates large accelerators
- Financed by 20 European countries
- SFR 950M budget - operation + new accelerators
- 3,000 staff
- 6,000 users (researchers) from all over the world
- Experiments conducted by a small number of large collaborations:
  LEP experiment (finished) : 500 physicists, 50 universities,
  20 countries, apparatus cost SFR 100M
  LHC experiment (future ~2005) : 2000 physicists, 150 universities, apparatus costing SFR 500M



CERN

CERN


Carte de situation
Paris
Jura
Monts
Lausanne
Nyon
LAC LEMAN
Divonne-les-Bains
Gex
LEP/LHC
Ferney-
-Voltaire
SPS
St Genis-Pouilly
CERN
Meyrin
Anières
GENEVE
Annemasse
Sites du CERN
Accélérateurs
Aéroport
Autoroutes
Frontière
St Julien-en-Genevois
Lyon
0
5
km
Annecy
Chamonix
(Mont-Blanc)
CERN AC - EI6-1 - 03 1997

The LEP accelerator

➔World's largest particle collider, ran for 11 years.

➔27 km circumference, 100 m underground

➔ Counter circulating beams of electron positron bunches

➔ Four experiments have confirmed Standard Model predictions to high precision

➔Maximum collision energy of 209 GeV

Many questions still remain Þ LHC

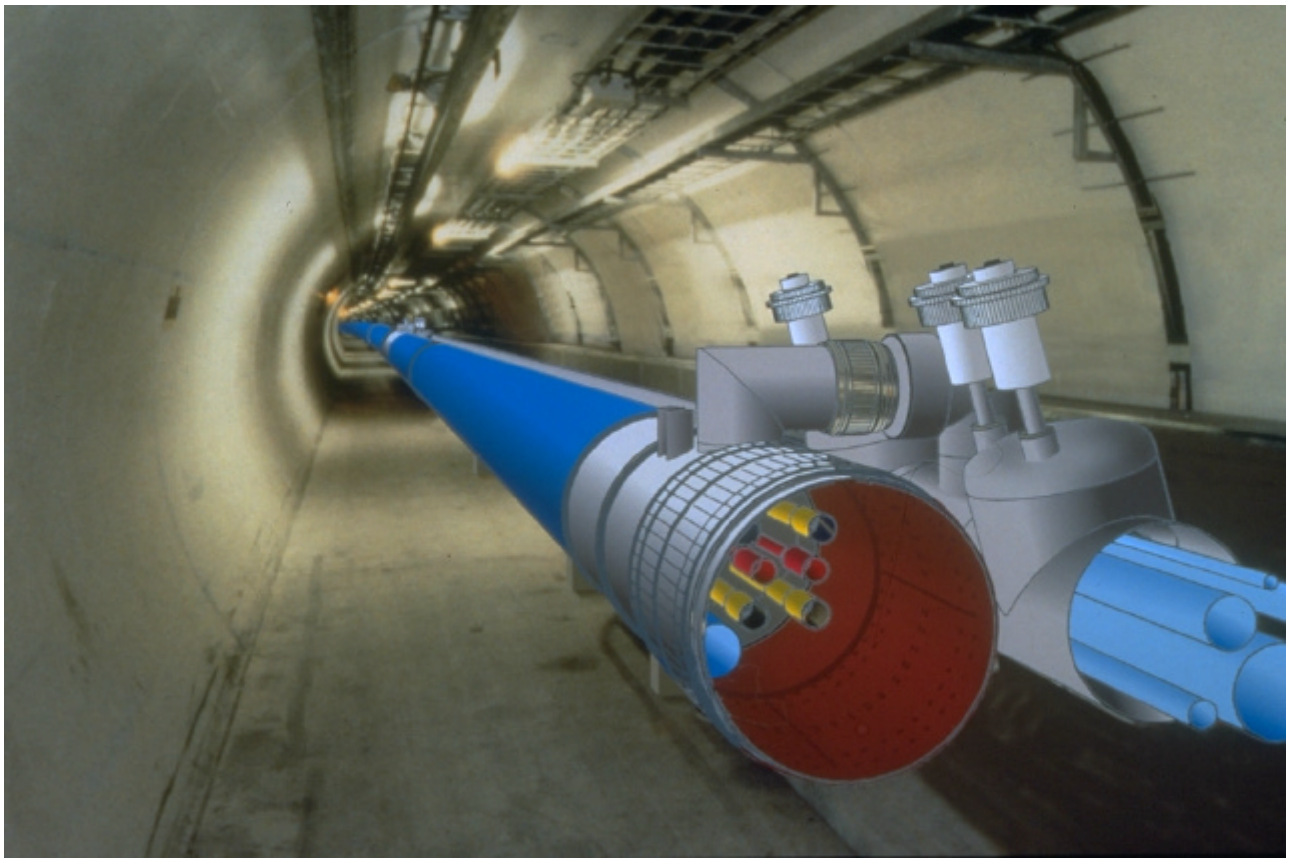
LHC in the LEP Tunnel

➔ Counter circulating beams of protons in the same beampipe.

➔ Centre of mass collision energy of 14 TeV.

➔ 1000 superconducting bending magnets, each 13 metres long, field 8.4 Tesla.

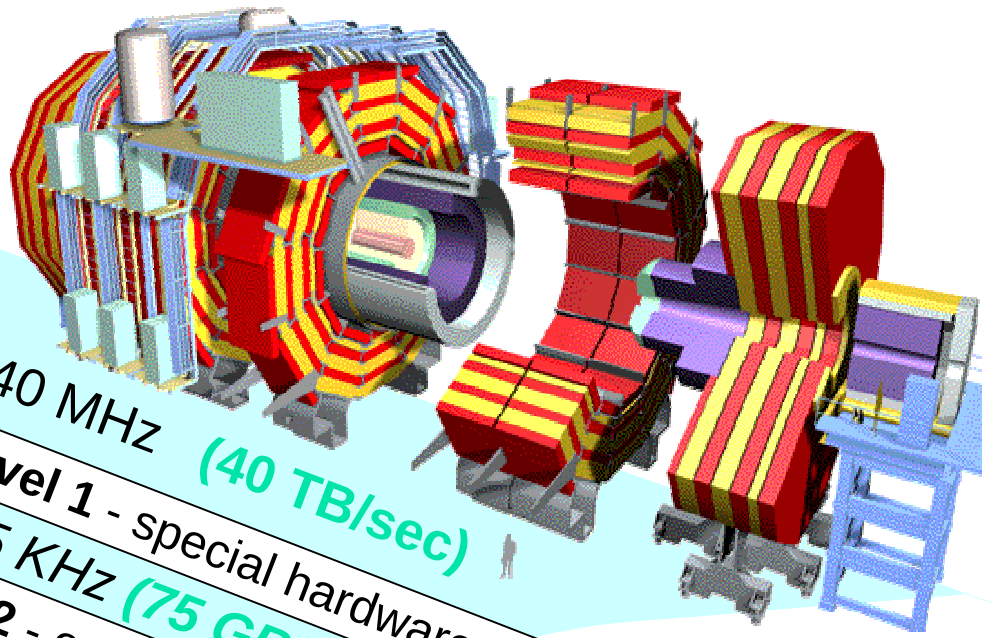
➔Super-fluid Helium cooled to 1.9^0 K



World's largest superconducting structure

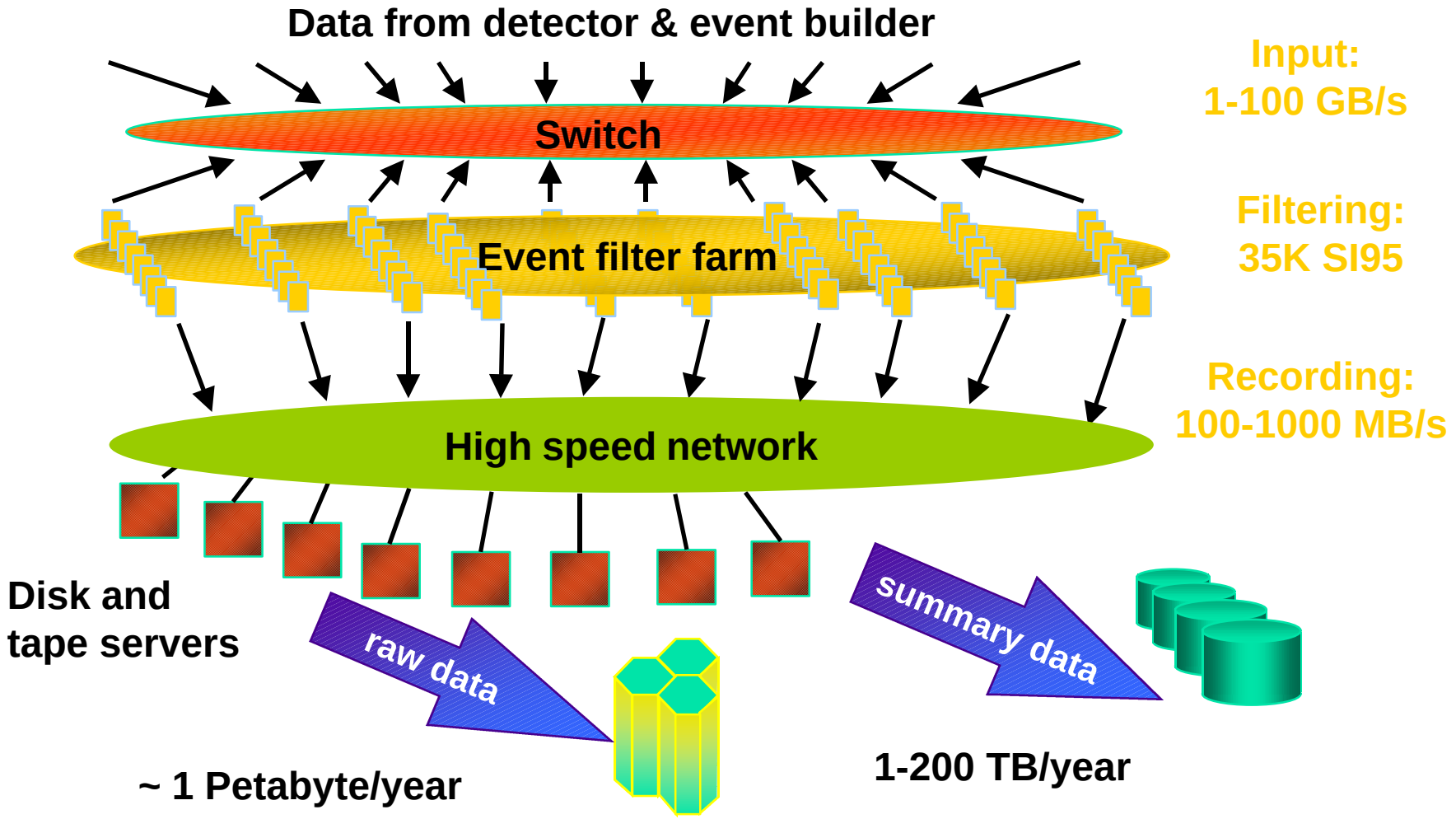
Online system

- Multi-level trigger
  - Filter out background
  - Reduce data volume
  - Online reduction 10^7
- Trigger menus
  - Select interesting events
  - Filter out less interesting



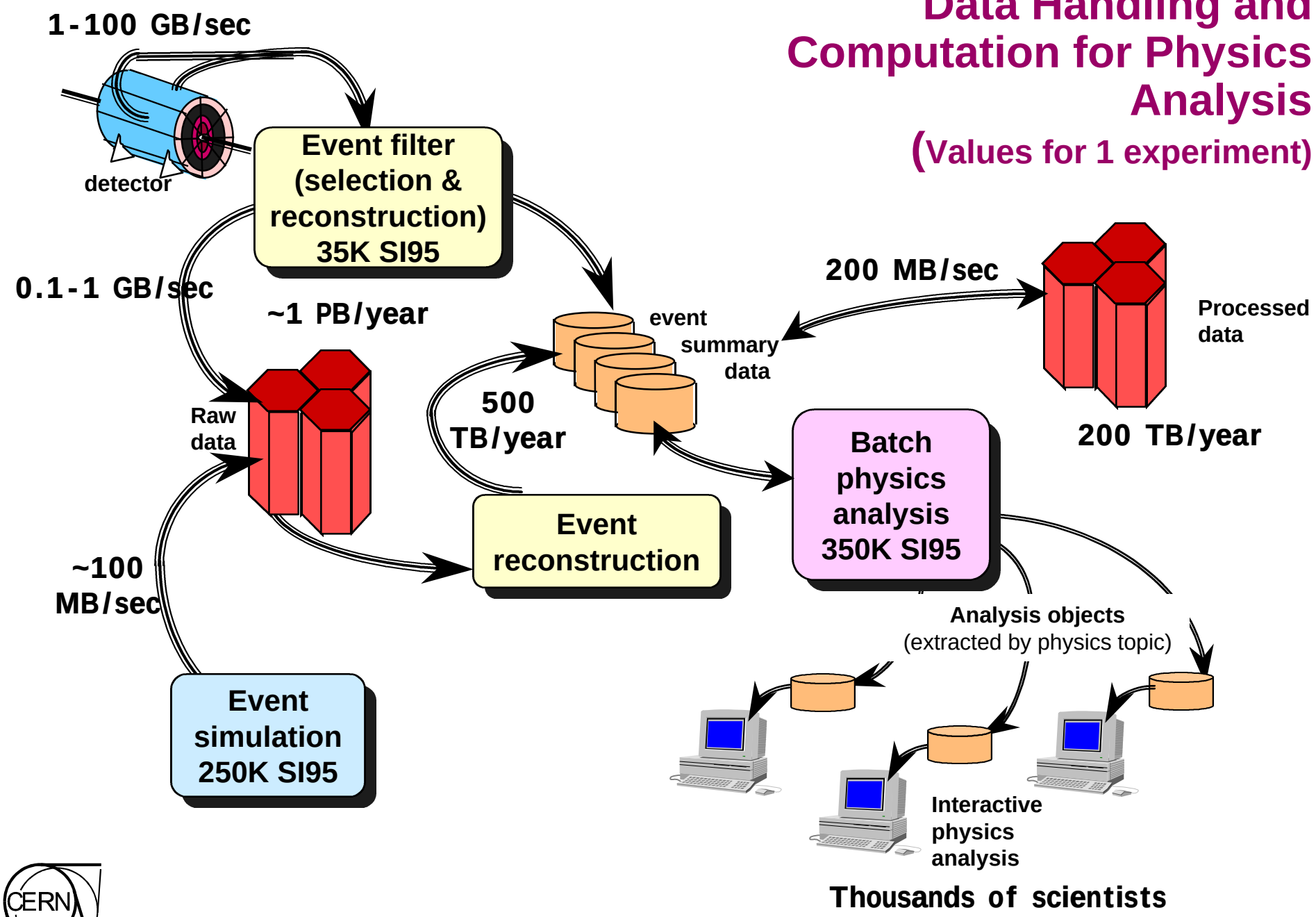
Event filter and data recording

(one experiment)



Data Handling and Computation for Physics Analysis

(Values for 1 experiment)

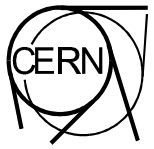


LEP to LHC

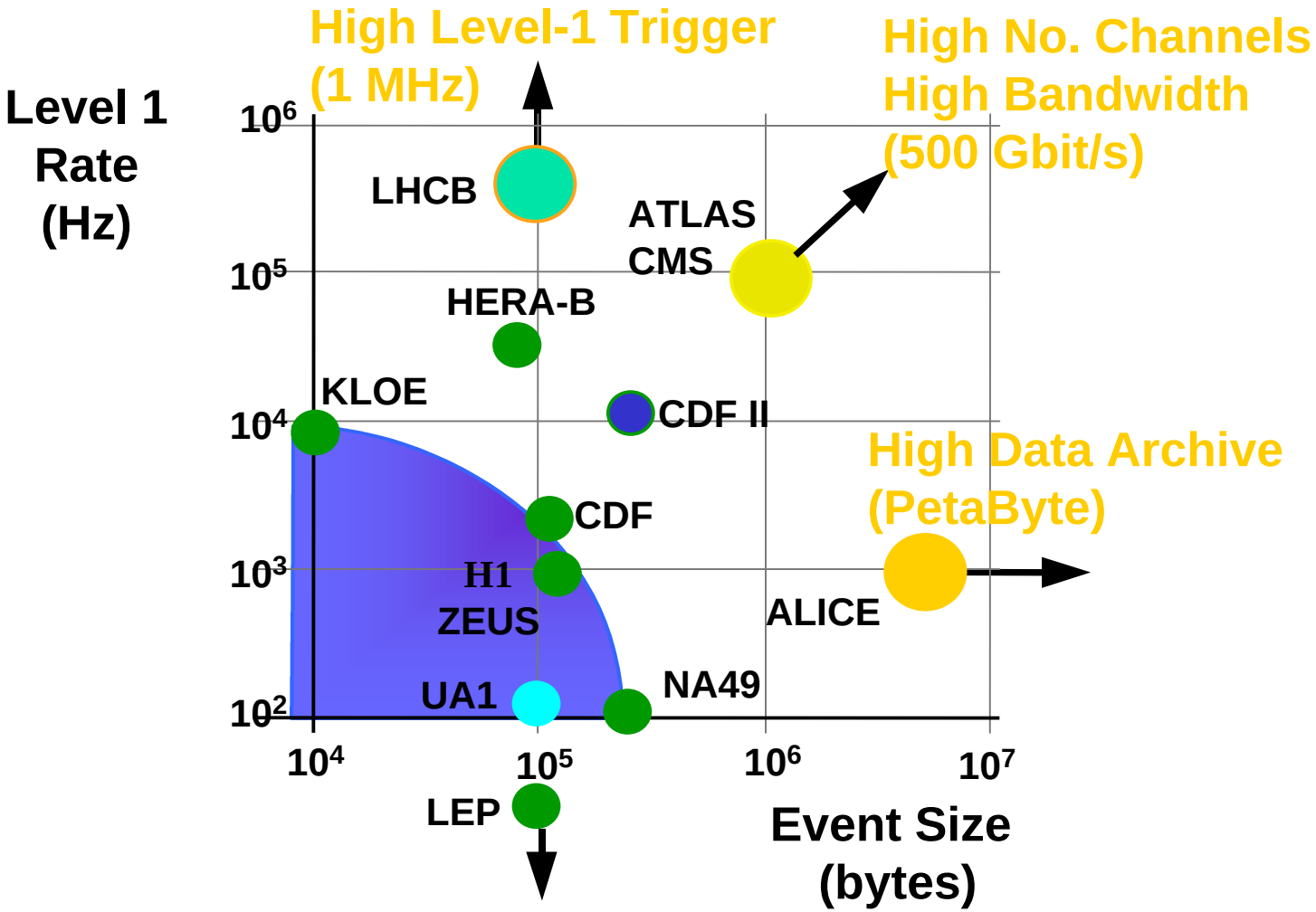
Parameter	LEP	COMPASS	LHC
Raw data rate	1MB/sec	35MB/sec	100MB/sec
Number of events/year	$<10^7$	$\sim 10^{10}$	$\sim 10^9$
Raw data volume/year	0.2-0.3 TB	300TB	1 PB
Event size	20 – 50 kB	30kB	1 MB
Event reconstruction time	2–8 Si95-secs	2 Si95-secs	500 Si95-secs
Number of users	400 - 600	~200	~2000
Number of institutes	30-50	~35	~150

Each LHC experiment requires one to two orders of magnitude greater than the TOTAL capacity installed at CERN today

All LEP: < 1TB/year Rate: 4MB/sec
All LHC: ~ 3PB/year Alice rate: 1GB/sec



How much data is involved?



Characteristics of HEP computing

Event independence

- Data from each collision is processed independently
- Mass of independent problems with no information exchange

Massive data storage

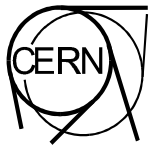
- Modest event size: 1-10 MB
- Total is very large - Petabytes for each experiment.

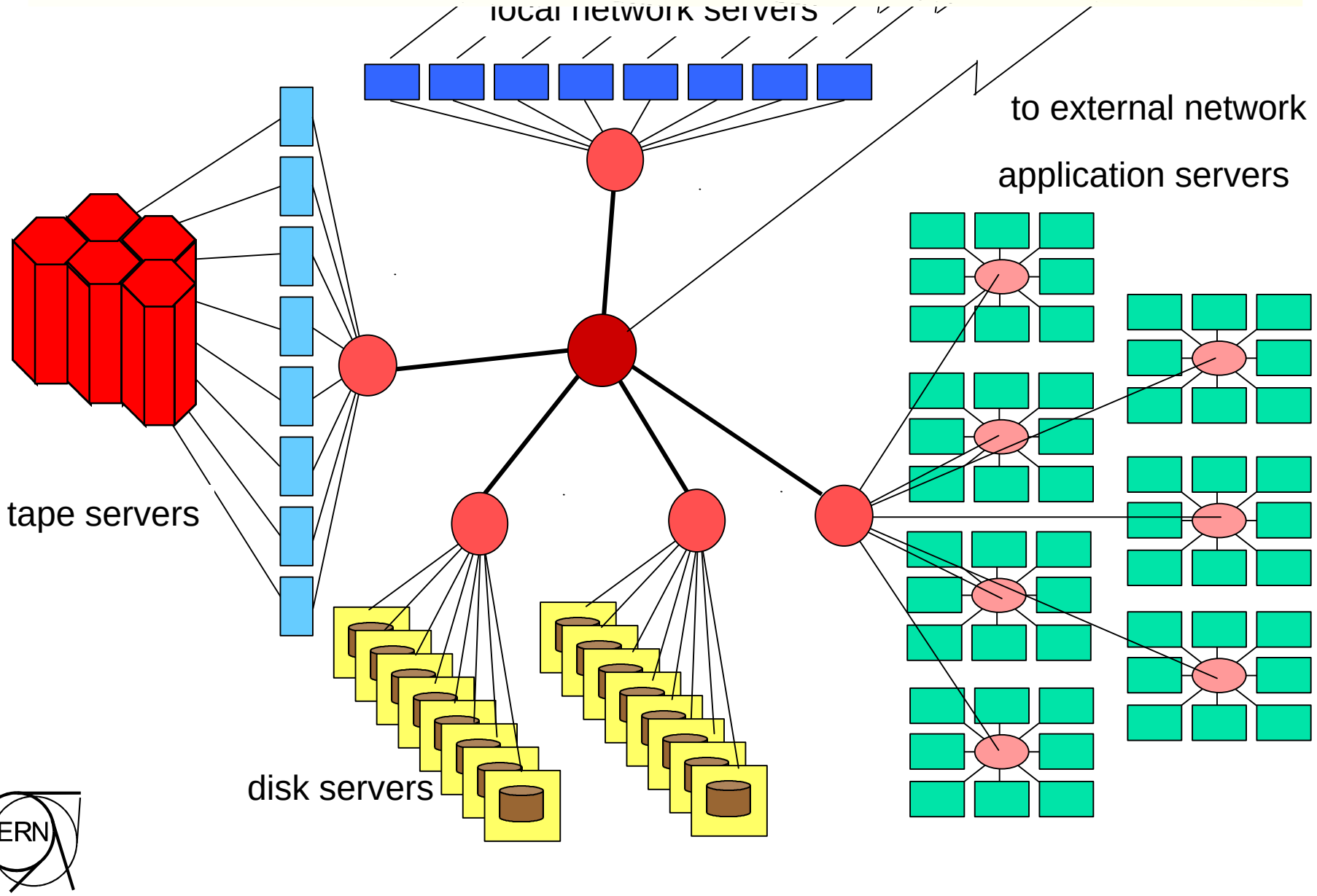
Mostly read only

- Data never changed after recording to tertiary storage
- But is read often ! cf.. magnetic tape as an archive medium

Modest floating point needs

- HEP computations involve decision making rather than calculation
- Computational requirements in SPECint95 secs




local network servers
to external network
application servers
tape servers
disk servers
CERN

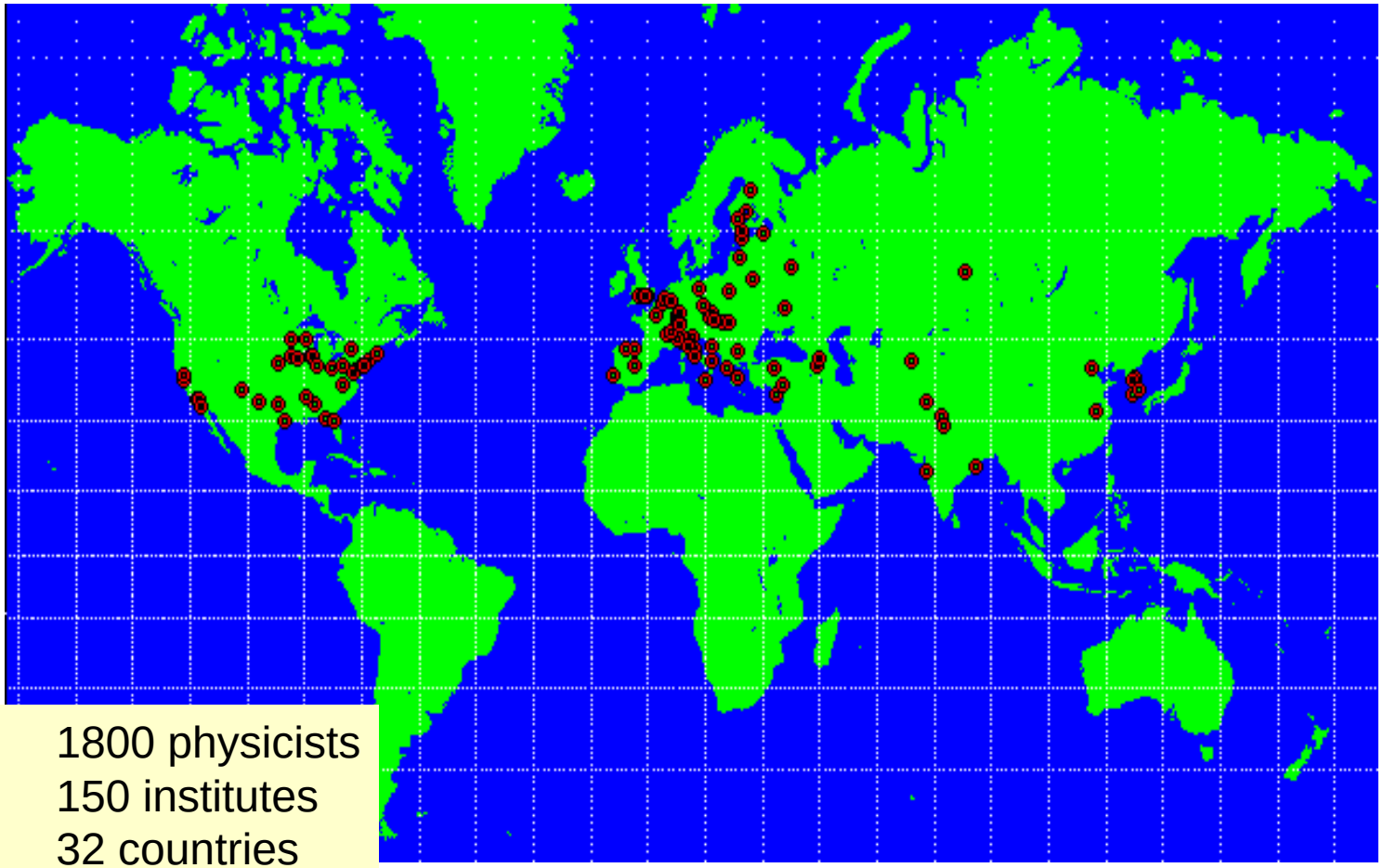
LHC Computing fabric at CERN

Estimated computing resources required at CERN for LHC experiments in 2007*							
collaboration		***ALICE***	***ATLAS***	***CMS***	***LHCB***	***Total***	***Today***
cpu capacity (KSI95)	*total 2007*	824	690	820	225	**2'559**	**10**
	annual inc. after 2007	272	228	271	74	**845**	
disk capacity (TB)	*total 2007*	530	410	1'140	330	**2'410**	**30**
	annual inc. after 2007	270	210	570	170	**1'220**	
active mag. tape capacity (PB)	*total 2007*	3.2	9.0	1.5	0.9	**14.6**	**1**
	annual inc. after 2007	3.2	9.0	1.5	0.9	**14.6**	
shelved mag. tape capacity (PB)	*total 2007*	0.0	0.0	2.6	0.3	**2.9**	
	annual inc. after 2007	0.0	0.0	2.6	0.3	**2.9**	
aggregate tape I/O rates (GB/sec)	*total 2007*	1.2	0.8	0.8	0.4	**3.2**	

*Taken from the LHC computing review

World Wide Collaboration

Þ distributed computing & storage capacity



CMS: 1800 physicists
150 institutes
32 countries

World-wide computing

Two problems:

- Funding
  - will funding bodies place all their investment at CERN?

No

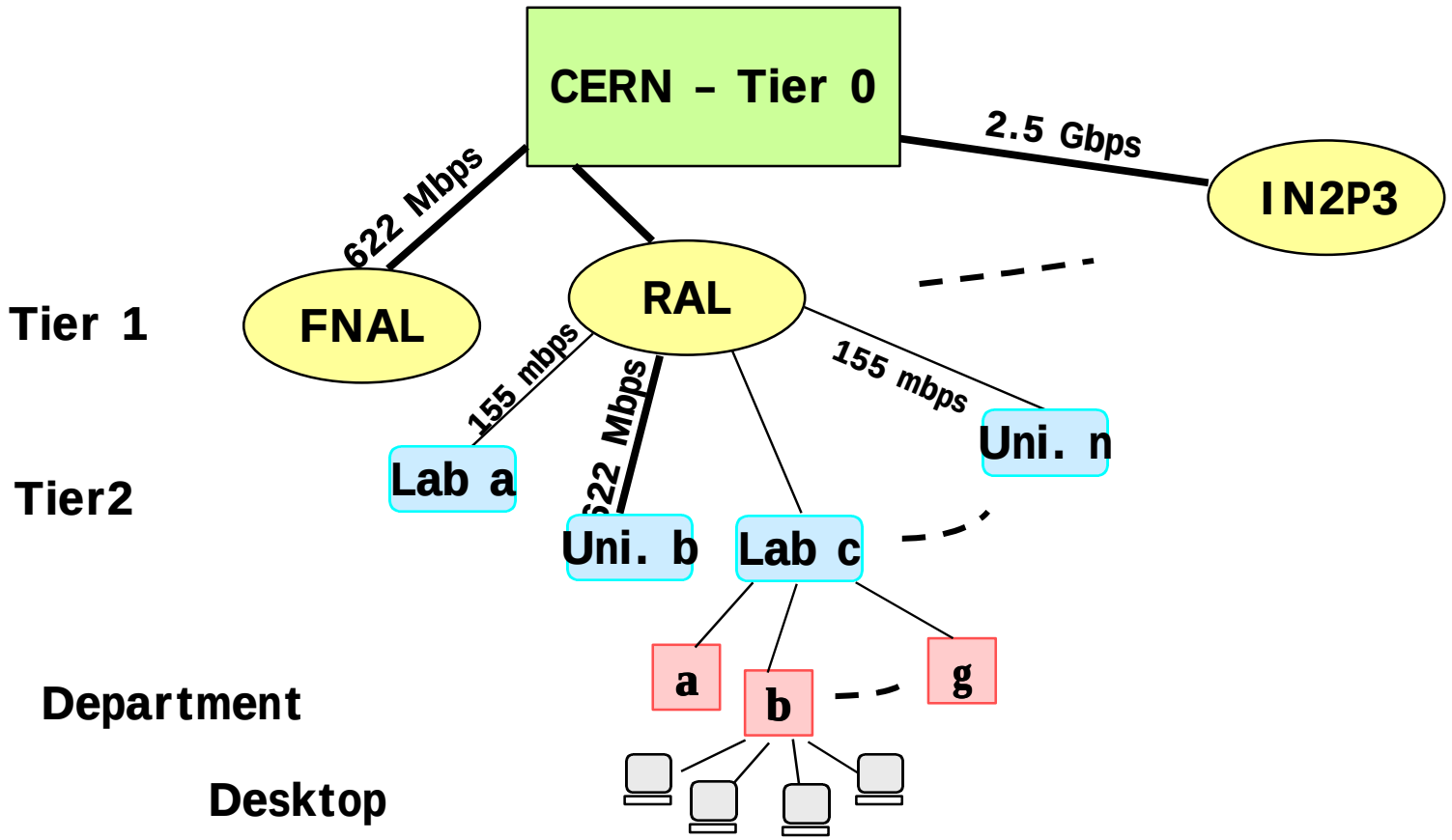
- Geography
  - does a geographically distributed model better serve the needs of the world-wide distributed community?

Maybe –

if it is **reliable** and **easy to use**

Need to provide physicists with the best possible access to LHC data irrespective of location

Regional centres - a multi tier model



MONARC report: http://home.cern.ch/~barone/monarc/RCArchitecture.html

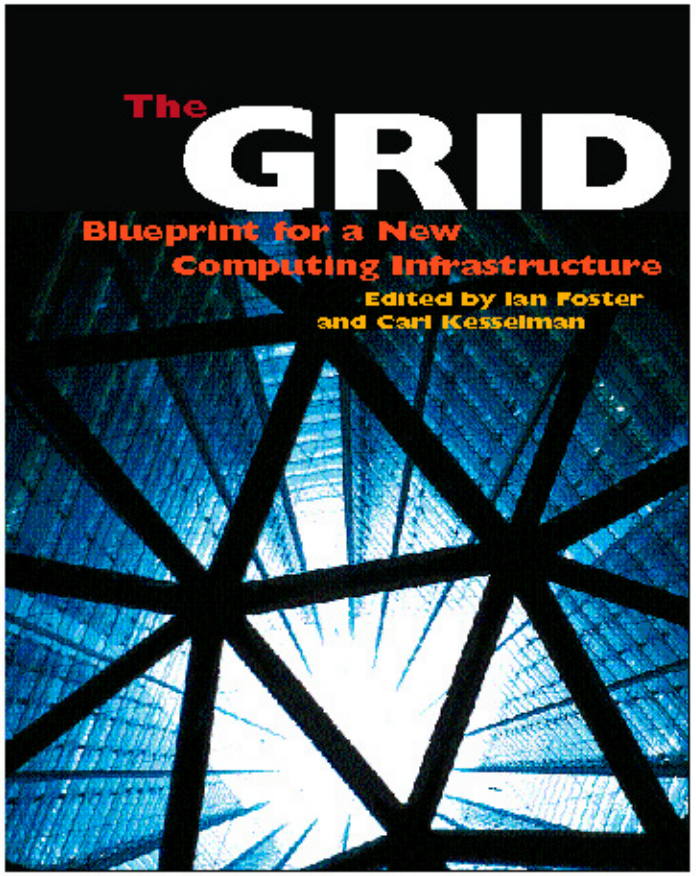
The Basic Problem - Summary

- Scalability → cost → complexity → management
  - Thousands of processors, thousands of disks, Petabytes of data, Terabits/second of I/O bandwidth, ….
- Wide-area distribution → complexity → management → bandwidth
  - WANs are only and will only be ~1-10% of LANs
  - Distribute, replicate, cache, synchronise the data
  - Multiple ownership, policies, ….
  - Integration of this amorphous collection of Regional Centres ..
  - .. with some attempt at optimisation
- Adaptability → flexibility → simplicity
  - We shall only know how analysis will be done once the data arrives

Can Grid technology be applied to LHC computing?

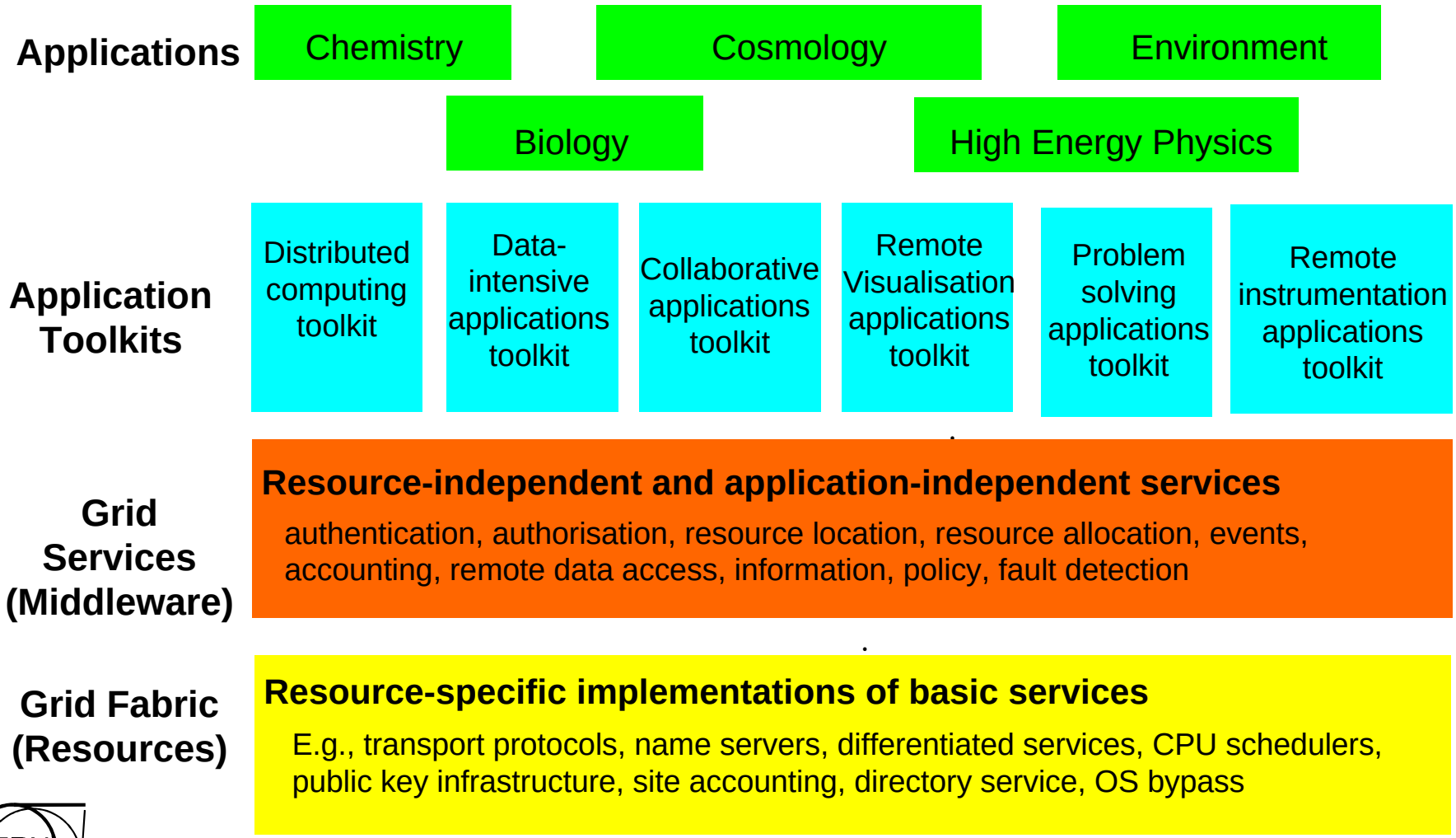
The GRID metaphor

- Analogous with the electrical power grid
- Unlimited ubiquitous distributed computing
- Transparent access to multi peta byte distributed data bases
- Easy to plug in
- Hidden complexity of the infrastructure



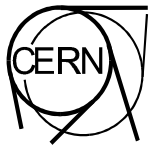
Ian Foster and Carl Kesselman, editors, "The Grid: Blueprint for a New Computing Infrastructure," Morgan Kaufmann, 1999, http://www.mkp.com/grids

GRID from a services view



What should the Grid do for you?

- You submit your work …
- … and the Grid:
  - Finds convenient places for it to be run
  - Organises efficient access to your data
    - Caching, migration, replication
  - Deals with authentication to the different sites that you will be using
  - Interfaces to local site resource allocation mechanisms, policies
  - Runs your jobs
  - Monitors progress
  - Recovers from problems
  - Tells you when your work is complete
- If there is scope for parallelism, it can also decompose your work into convenient execution units based on the available resources, data distribution



European Data Grid -- R&D requirements

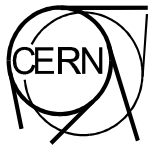
Local fabric

- Management of giant computing fabrics
  - auto-installation, configuration management, resilience, self-healing
- Mass storage management
  - multi-PetaByte data storage, “real-time” data recording requirement,
  - active tape layer – 1,000s of users, uniform mass storage interface,
  - exchange of data and metadata between mass storage systems

Wide-area

- Workload management
  - no central status, local access policies
- Data management
  - caching, replication, synchronisation, object database model
- Application monitoring

Note: Build on existing components such as Globus middleware
Foster (Argonne) and Kesselman (University of Southern California)



European Data Grid partners

Managing partners

UK: PPARC Italy: INFN France: CNRS Netherlands: NIKHEF

ESA/ESRIN CERN

Industry

IBM (UK), Compagnie des Signaux (F), Datamat (I)

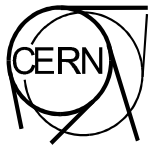
Associate partners

Istituto Trentino di Cultura (I), Helsinki Institute of Physics / CSC Ltd (FI), Swedish Science Research Council (S), Zuse Institut Berlin (DE), University of Heidelberg (DE), CEA/DAPNIA (F), IFAE Barcelona, CNR (I), CESNET (CZ), KNMI (NL), SARA (NL), SZTAKI (HU)

Other sciences

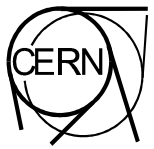
KNMI(NL), Biology, Medicine

Formal collaboration with USA being established



Preliminary programme of work

Middleware	
WP 1 Grid Workload Management	F. Prelz/INFN
WP 2 Grid Data Management	B. Segal/CERN
WP 3 Grid Monitoring services	R. Middleton/PPARC
WP 4 Fabric Management	O. Barring/CERN
WP 5 Mass Storage Management	J. Gordon/PPARC
Grid Fabric -- testbed	
WP 6 Integration Testbed	F. Etienne/CNRS
WP 7 Network Services	P. Primet/CNRS
Scientific applications	
WP 8 HEP Applications	F. Carminati/CERN
WP 9 EO Science Applications	L. Fusco/ESA
WP 10 Biology Applications	V. Breton/CNRS
Management	
WP 11 Dissemination	M. Draoli/CNR
WP 12 Project Management	F. Gagliardi/CERN



Middleware : WP 1 - WP 3: wide area

Workload Management WP 1

- Define and implement a suitable architecture for distributed scheduling and compute resource management in a GRID environment.
- Maximise the global system throughput.

Data management WP 2

- manage and share PetaByte-scale information volumes in high-throughput production-quality grid environments.
- Replication/caching; Metadata mgmt.; Authentication; Query optimisation;
- High speed WAN data access; interface to Mass Storage Mgmt. systems.

Application monitoring WP 3

- Tens of thousands of components, thousands of jobs and individual users
- End-user - tracking of the progress of jobs and aggregates of jobs
- Understanding application and grid level performance

Middleware WP 4 - WP 5 : local fabric

Fabric management WP 4

- Automated installation, configuration management, system maintenance
- Automated monitoring and error recovery - resilience, self-healing
- Performance monitoring
- Characterisation, mapping, management of local Grid resources

Mass storage management WP 5

- Multi-PetaByte data storage HSM devices
- Uniform mass storage interface
- Exchange of data and metadata between mass storage systems

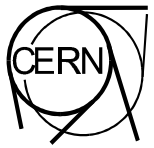
Grid fabric WP 6 - WP 7

Integration test bed WP 6

- Operate prototype test beds for applications / experiments.
- Integrate & build successive releases of the project middleware.
- Demonstrate by the end of the project, test beds operating as production facilities for real end-to-end applications over large trans-European and potentially global high performance networks.

Networking services WP 7

- Definition and management of the network infrastructure.
- Monitor network traffic and performance, develop models and provide tools and data for the planning of future networks, especially concentrating on the requirements of Grids handling significant volumes of data.
- Deal with the distributed security aspects of Data Grid.



Scientific applications WP 8 - WP 10

HEP WP 8

- Develop and/or adapt High Energy Physics applications (Simulation, Data Analysis, etc.) for the geographically distributed community using the functionality provided by the Data Grid, i.e. transparent access to distributed data and high performance computing facilities.

Four LHC experiments involved -- requirements are similar

Earth Observation WP 9

- Develop Grid-aware Earth Sciences applications
- Facilitate access to large computational power and large distributed data files for Earth Sciences applications.

Biology WP 10

- High throughput for the determination of three-dimensional macromolecular structures, analysis of genome sequences.
- Production storage and comparison of genetic information.
- Retrieval and analysis of biological literature and development of a search engine for relations between biological entities.

Management WP 11 - WP 12

Information dissemination and exploitation WP 11

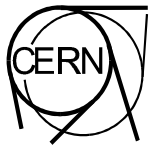
- Generation of required interest necessary for the deployment of the Datagrid Project's results
- Promotion of the middleware in industry projects
- Co-ordination of the dissemination activities undertaken by the project partners in the various European countries
- Industry & Research Grid Forum initiated as the main exchange place of information dissemination and potential exploitation of the Data Grid results

Project management WP 12

- Overall management and administration of the project
- Co-ordination of technical activity within the project
- Conflict and resource allocation resolution and external relations

Status

- Prototype work began at CERN (and in some of the collaborating institutes) before the official project start date . Globus initial installation and tests done early: several problems found and corrected.
- Proposal to the EU submitted on May 8th 2000; second draft submitted in September; accepted and signed December 29 (2000). Project started officially January 1st 2001.
- The first Project milestone is the Month 9 integration of early middleware and Globus on to the first testbed configurations. This is taking place as we speak.



EU Data Grid Main Issues

- Project is by EU standards very large in funding and participants
- Management and coordination is a major challenge
- Coordination between national (European) and EU Data Grid programmes
- Coordination with US Grid activity (GriPhyN, PPDG, Globus)
- Coordination of the HEP and other sciences' objectives
- Very high expectations already raised, could bring disappointments

Conclusions

The scale of the computing needs of the LHC experiments is **very large** compared with current experiments

- each LHC experiment requires one to two orders of magnitude greater capacity than the total installed at CERN today

We believe that the hardware technology will be there to evolve the current architecture of "commodity clusters" into large scale computing fabrics.

- But there are many management problems - workload, computing fabric, data, storage in a wide area distributed environment
- Disappointingly, solutions for local site management on this scale are not emerging from industry

The scale and cost of LHC computing imposes a geographically distributed model.

The Grid metaphor describes an appropriate computing model for LHC and future HEP computing.