Service Challenge Tests of the LCG Grid

Andrzej Olszewski
Institute of Nuclear Physics PAN
Kraków, Poland

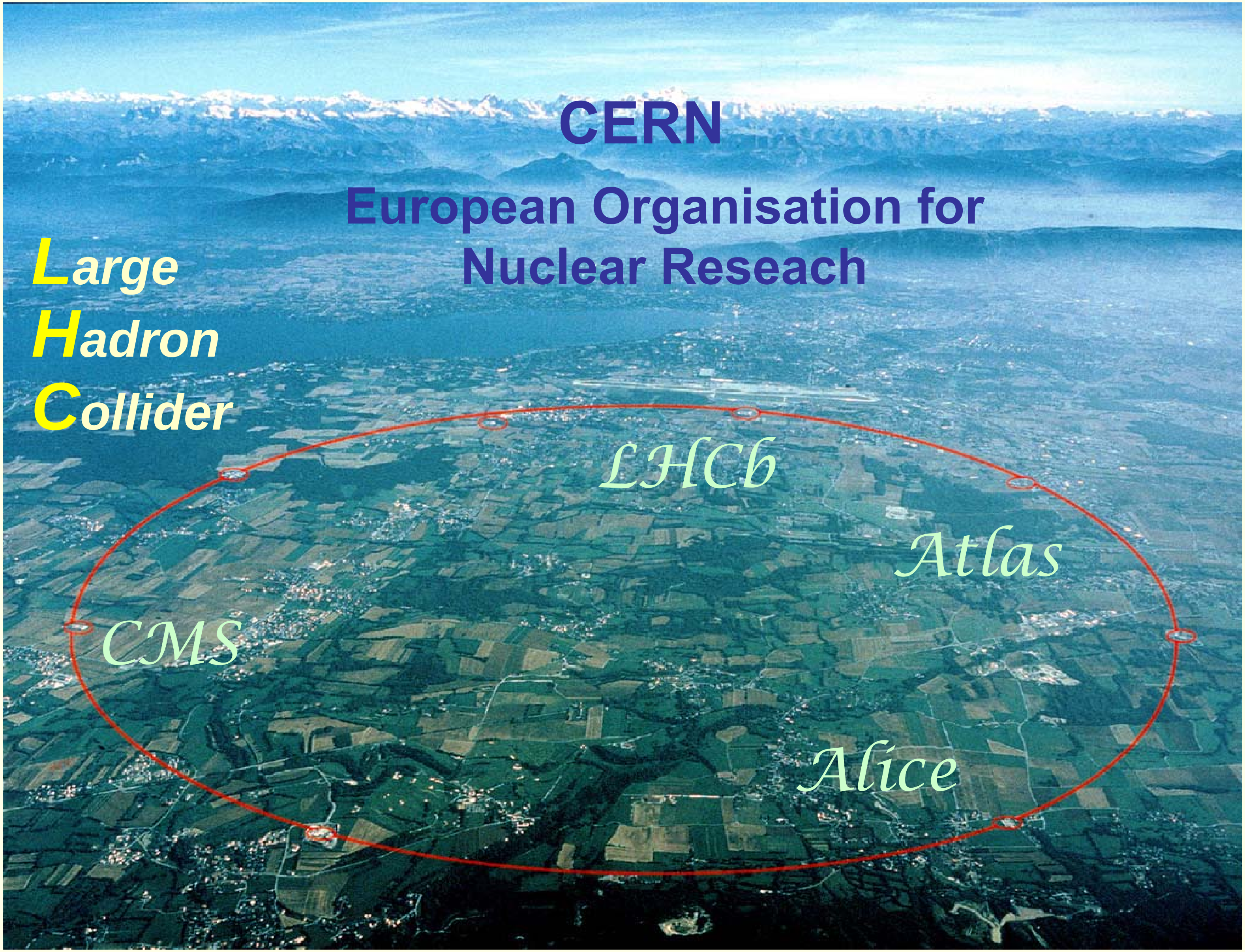
Cracow '05 Grid Workshop

22nd Nov 2005

The materials used in this presentation come from many sources
In particular I have used slides from presentations of LHC experiments and LHC Computing Grid

CERN

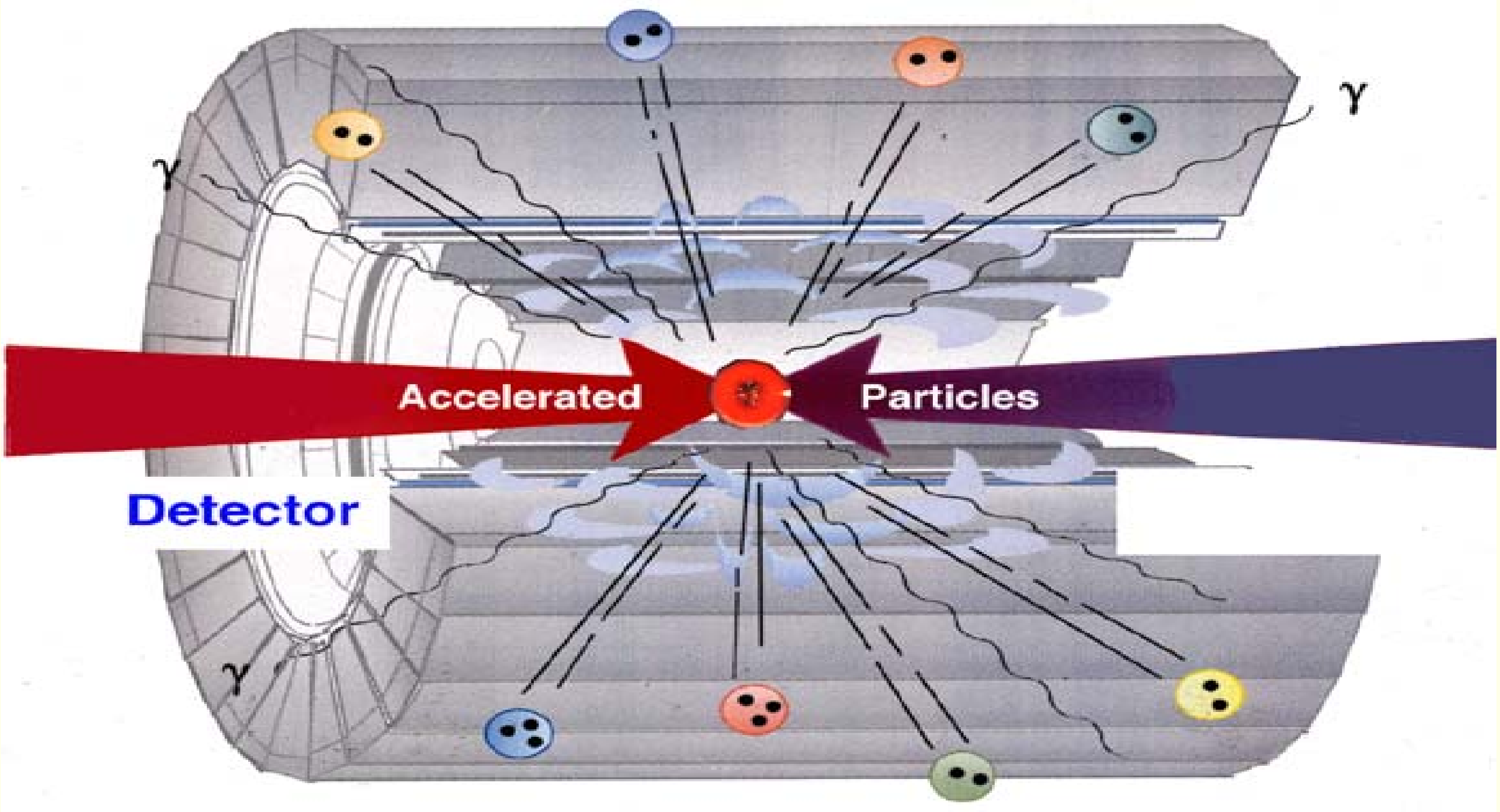
European Organisation for Nuclear Reseach



Large
Hadron
Collider

Experiment

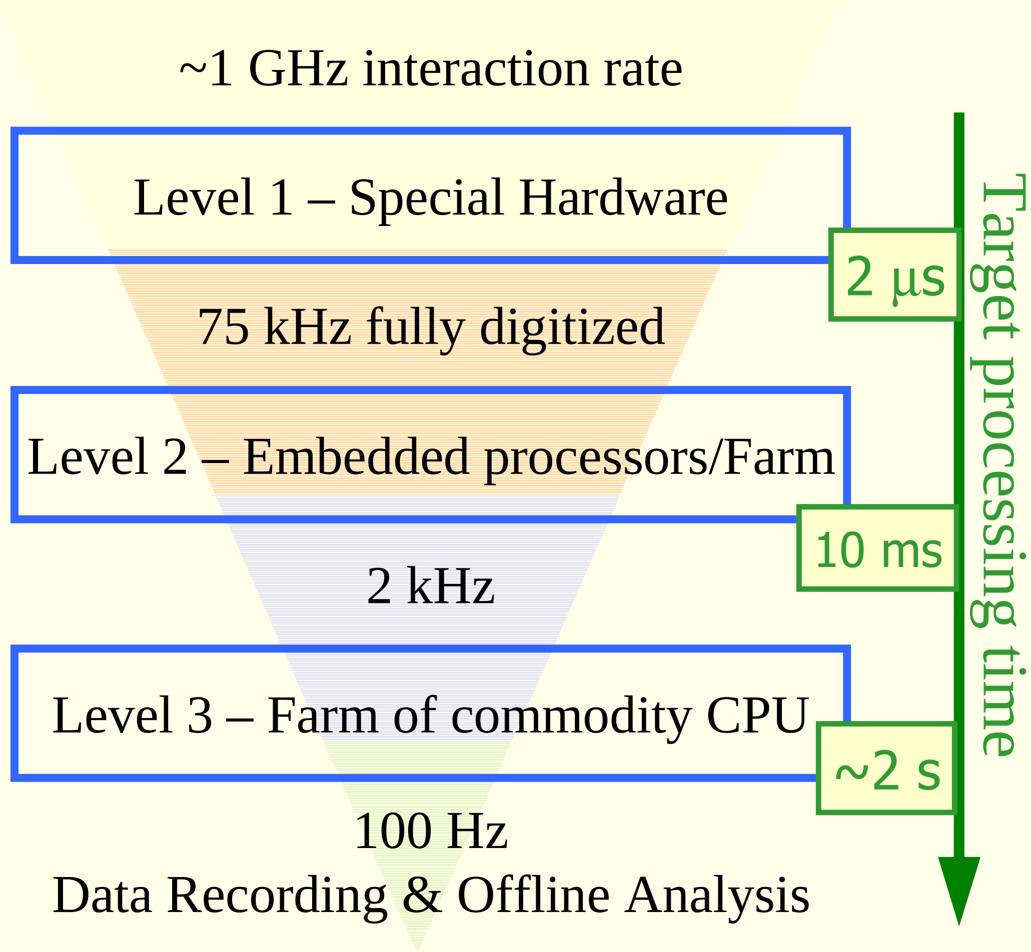
The most powerful microscope



Data Reduction

Data preselection
in real time

- many different physics processes
- several levels of filtering
- high efficiency for events of interest
- total reduction factor of about 10^7



~1 GHz interaction rate

Level 1 – Special Hardware

75 kHz fully digitized

Level 2 – Embedded processors/Farm

2 kHz

Level 3 – Farm of commodity CPU

100 Hz
Data Recording & Offline Analysis

Data Rates

	Rate	RAW	ESD	AOD	Monte	Monte
			Reco		Carlo	Carlo
	[Hz]	[MB]	[MB]	[kB]	[MB/evt]	% of real
ALICE HI	100	12.5	2.5	250	300	100
ALICE pp	100	1	0.04	4	0.4	100
ATLAS	200	1.6	0.5	100	2	20
CMS	150	1.5	0.25	50	2	100
LHCb	2000	0.025	0.025		0.5	20

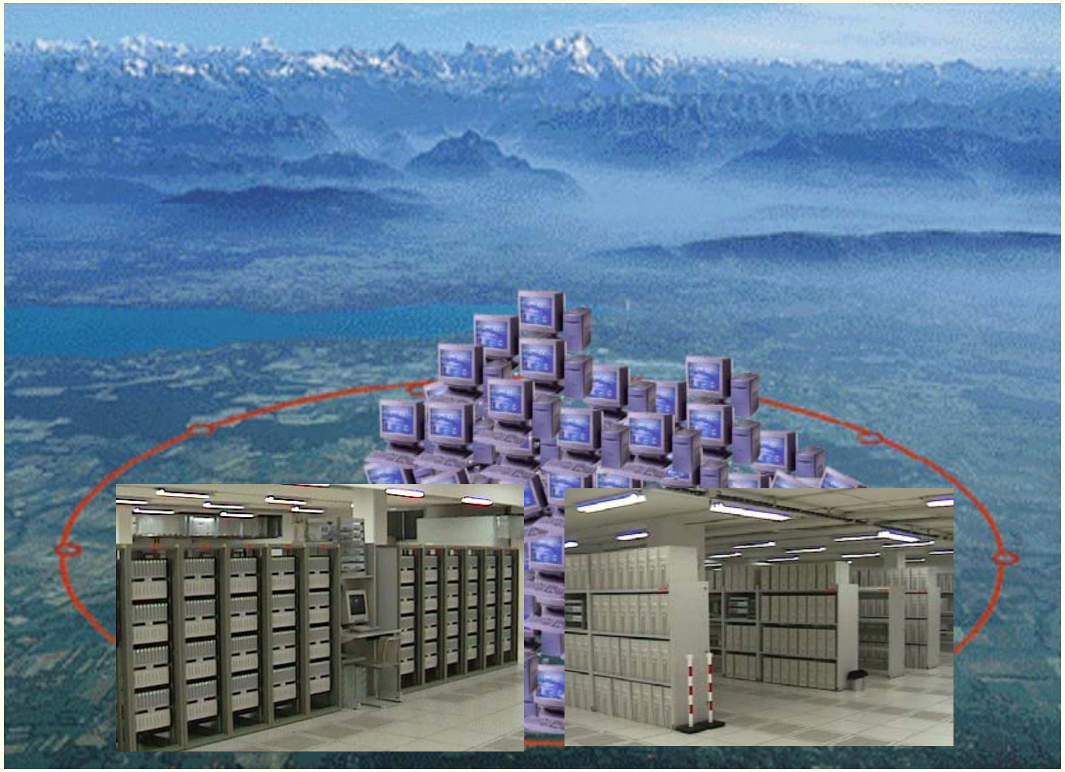
50 days running in 2007
10^7 seconds/year pp from 2008 on → ~10^9 events/experiment
10^6 seconds/year heavy ion

Mountains of CPU & Disks

For LHC computing,
100M SpecInt2000 or
100K of 3GHz Pentium 4
is needed!

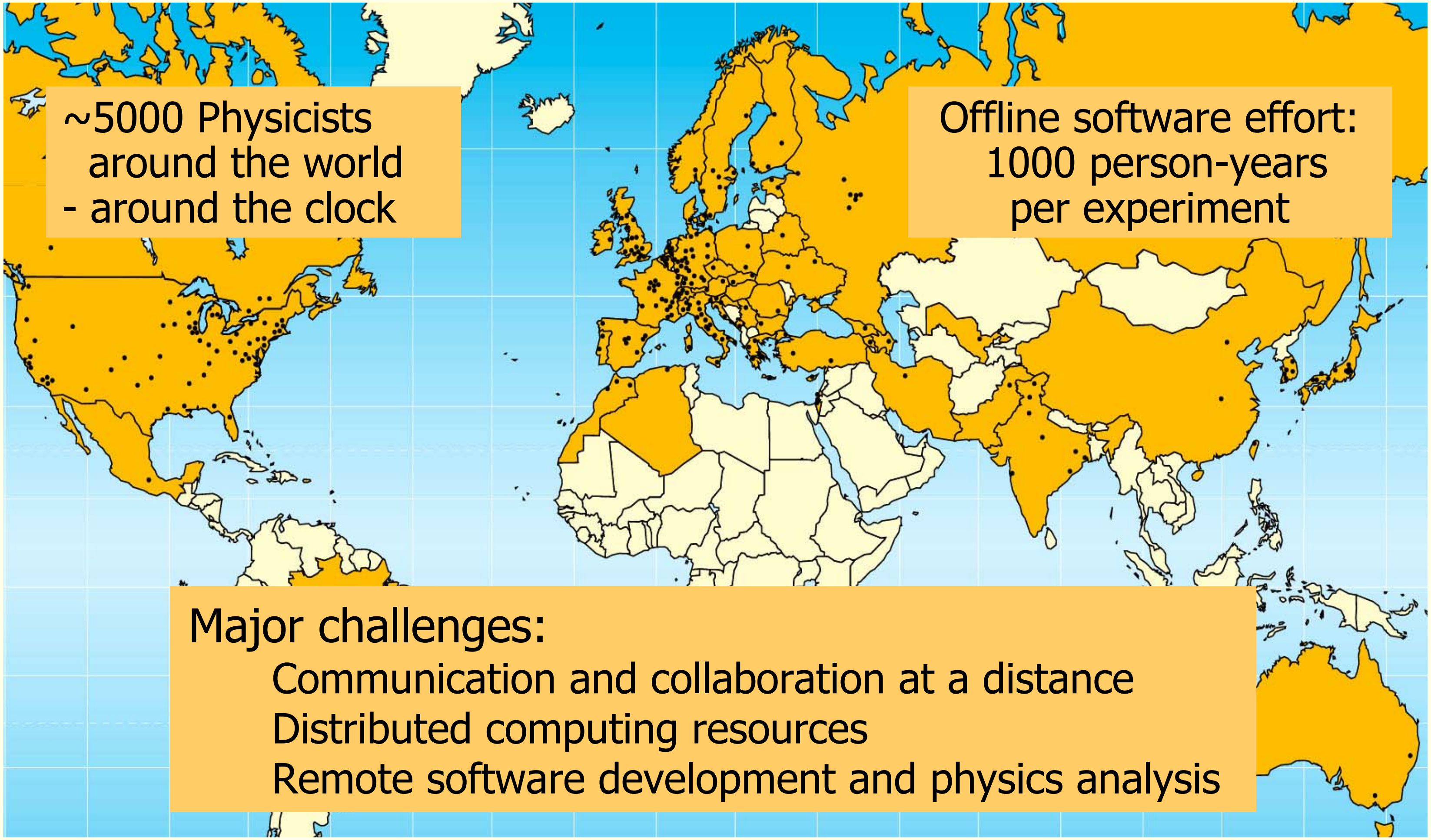
For data storage,
20 Peta Bytes or
100K of disks/tapes
per year is needed!

At CERN currently:
~2,400 processors
~2 Peta Bytes of disk
~12 PB of magnetic tape



Even with technology- driven improvements in performance and costs – CERN can provide nowhere near enough capacity for LHC!

Large, distributed community



~5000 Physicists
around the world
- around the clock

Offline software effort:
1000 person-years
per experiment

Major challenges:

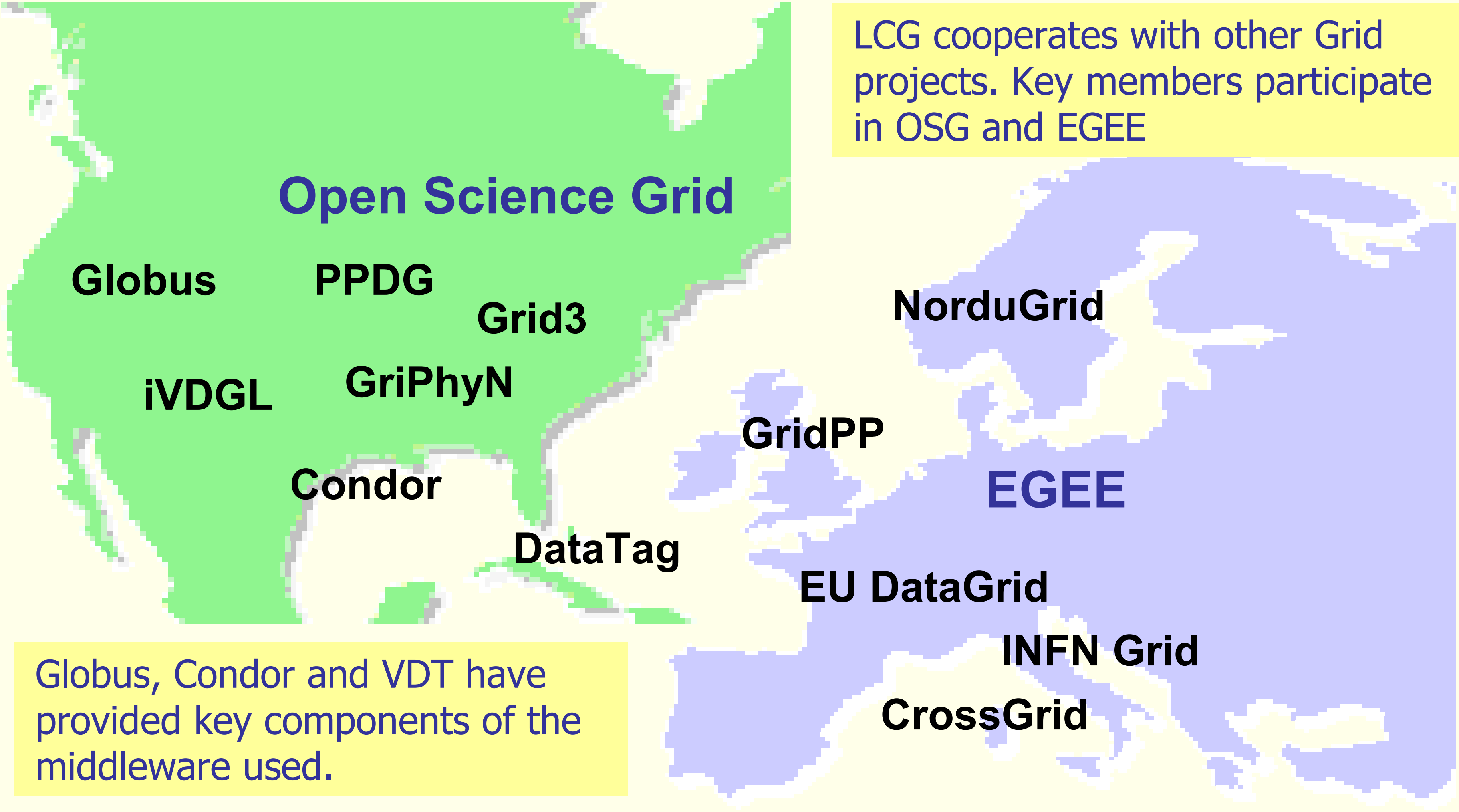
- Communication and collaboration at a distance
- Distributed computing resources
- Remote software development and physics analysis

The LCG Project

- Objectives
  - Design, prototyping and implementation of a computing environment for LHC experiments:
    - Infrastructure (for HEP it is effective to use PC farms)
    - Middleware (based on EDG, VDT, gLite….)
    - operations (experiment VOs, operation and support centres)
- Approved by the CERN Council in September 2001
  - Phase 1 (2001-2004):
    Development of a distributed production prototype that will be operated as a platform for the data challenges
  - Phase 2 (2005-2007):
    Installation and operation of the full world-wide initial production Grid system, requiring continued manpower efforts and substantial material resources.

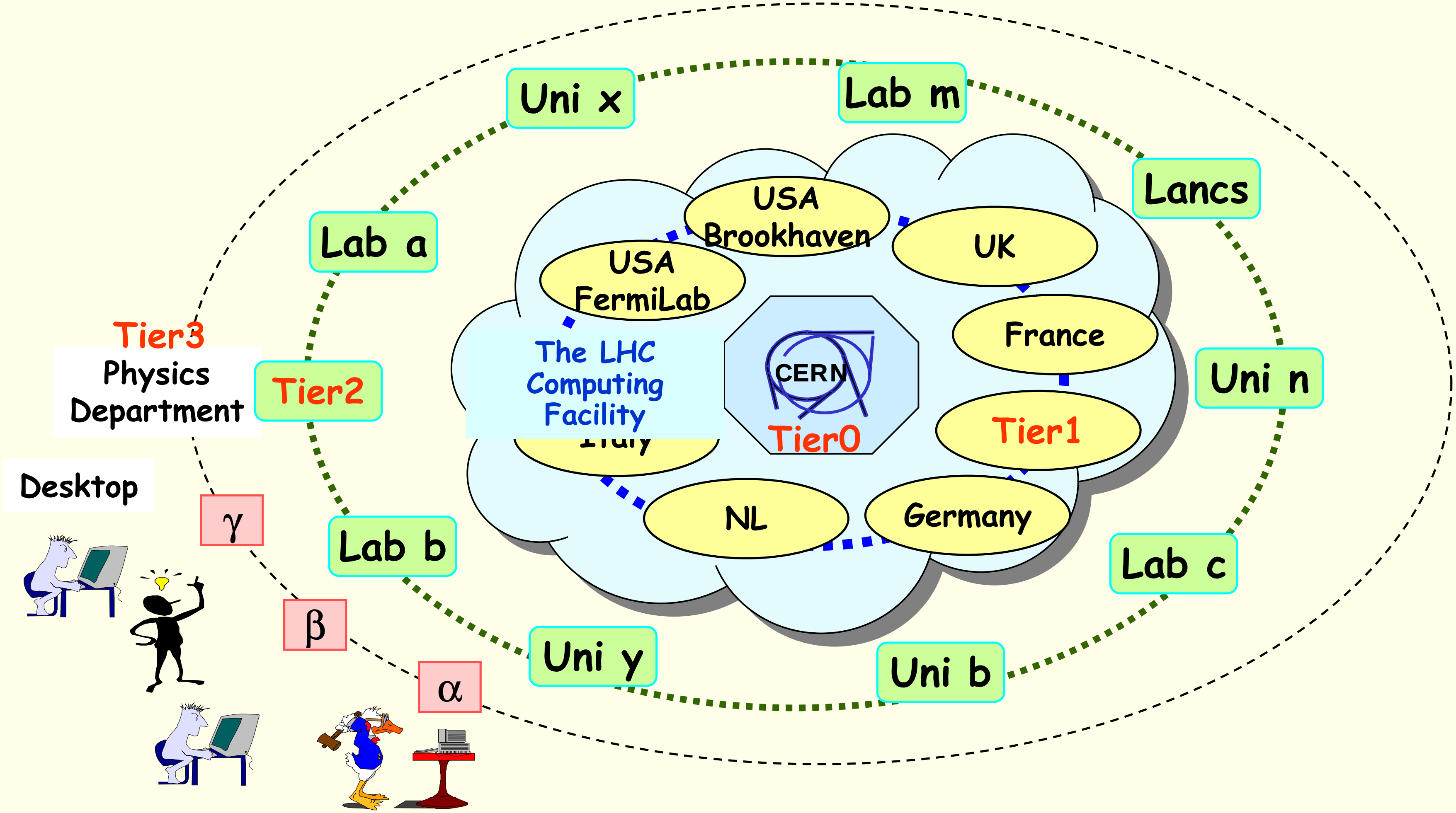
Grid Foundation Projects

LCG cooperates with other Grid projects. Key members participate in OSG and EGEE



Globus, Condor and VDT have provided key components of the middleware used.

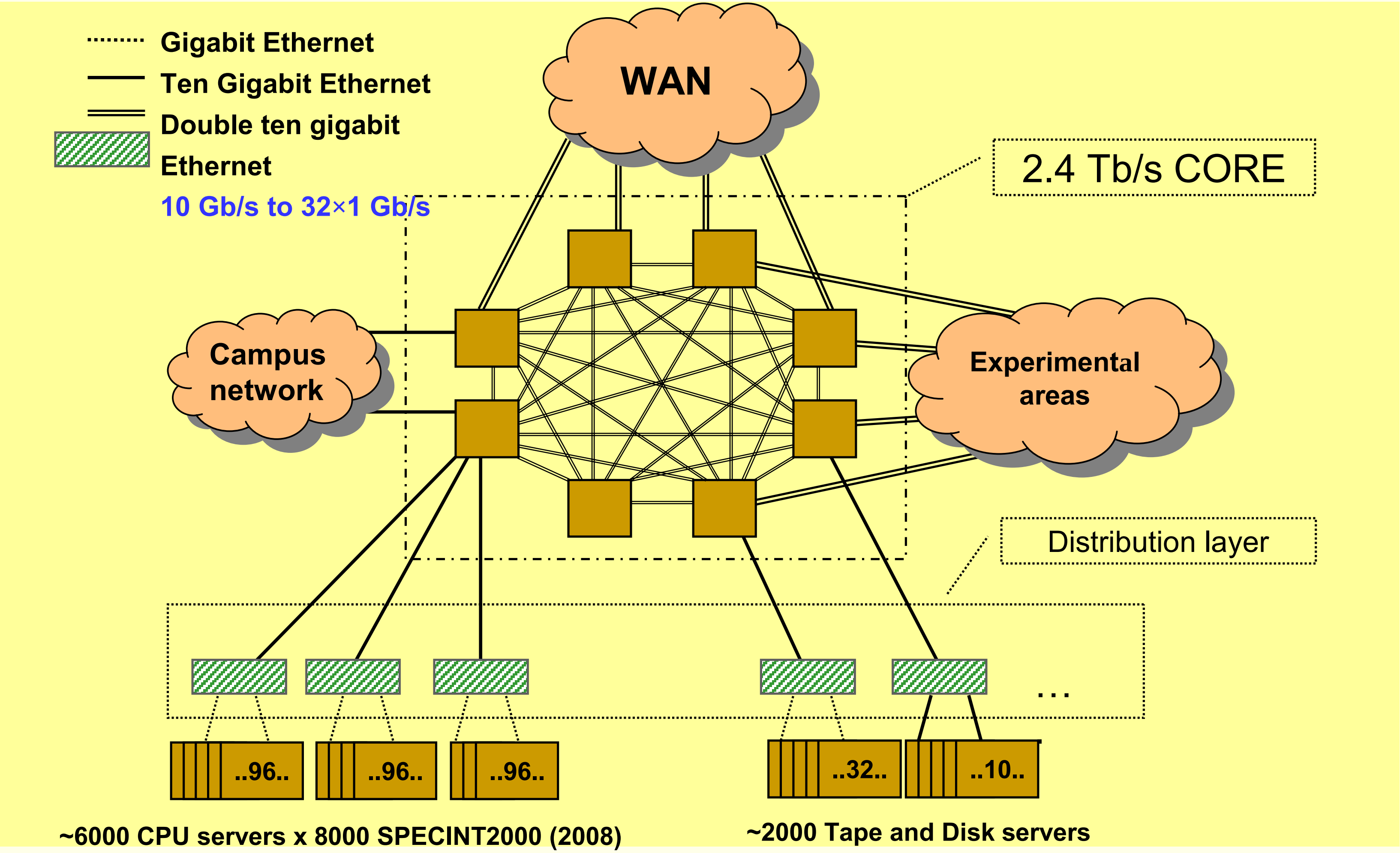
LCG Hierarchical Model



LCG Hierarchical Model

- Tier-0 at CERN
  - Record RAW data (1.25 GB/s ALICE)
  - Distribute second copy to Tier-1s
  - Calibrate and do first-pass reconstruction
- Tier-1 centers (11 defined)
  - Manage permanent storage – RAW, simulated, processed
  - Capacity for reprocessing, bulk analysis
- Tier-2 centers (> 100 identified)
  - Monte Carlo event simulation
  - End-user analysis
- Tier-3
  - Facilities at universities and laboratories
  - Access to data and processing in Tier-2s, Tier-1s

Architecture – Tier0



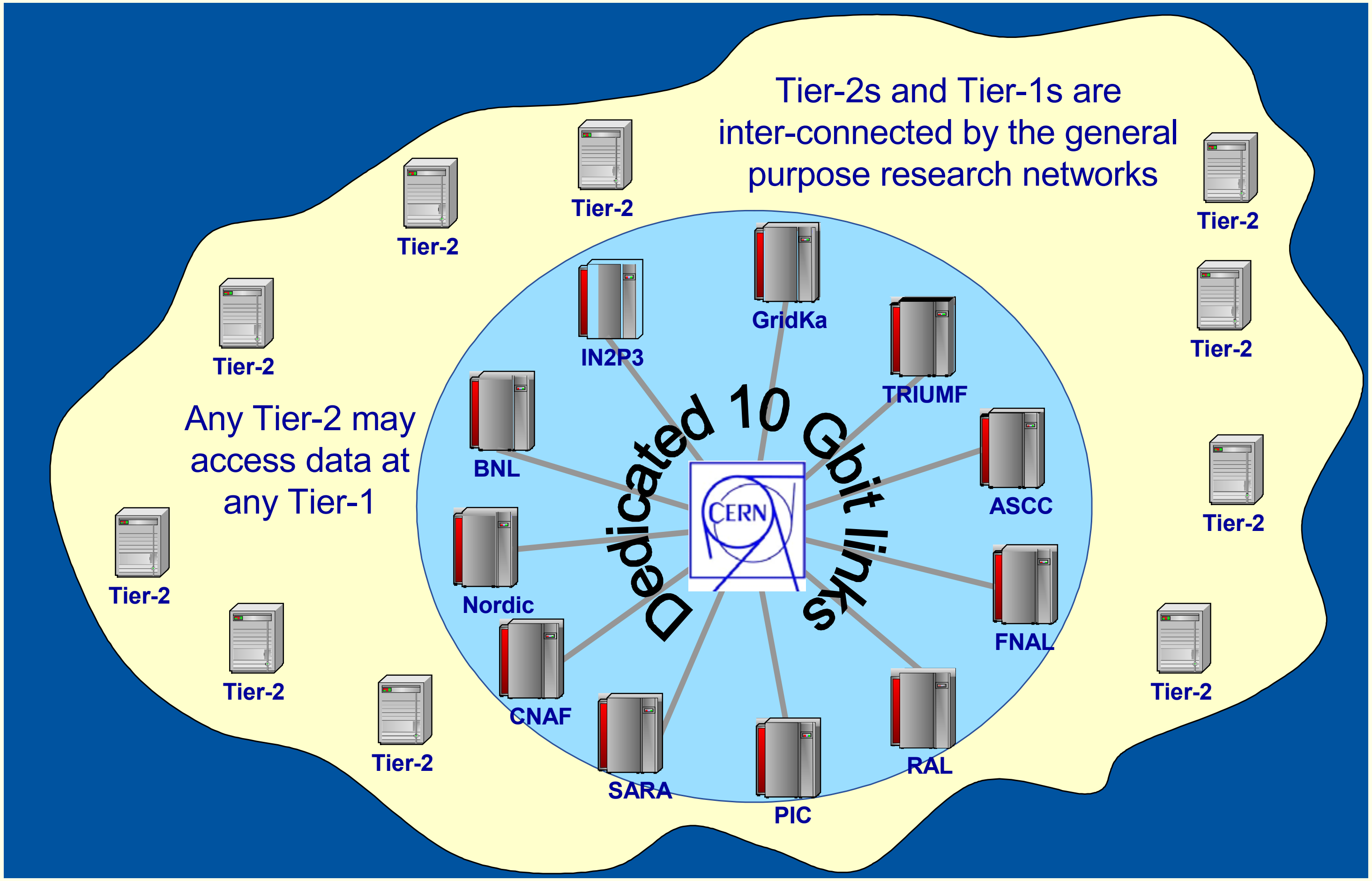
~6000 CPU servers x 8000 SPECINT2000 (2008)

~2000 Tape and Disk servers

Network Connectivity

National Reasearch Networks (NRENs) at Tier-1s:

ASnet
LHCnet/ESnet
GARR
LHCnet/ESnet
RENATER
DFN
SURFnet6
NORDUnet
RedIRIS
UKERNA
CANARIE



Service Challanges

LCG Service Challenges are about preparing, hardening and delivering the production LHC Computing Environment.

- Data recording

  CERN must be capable of accepting data from the experiments and recording it at a long term sustained average rate of 1.6 – 1.8 GBytes/sec

- Service Challenge 1 - 2

  Demonstrate reliable file transfer, disk to disk, between CERN and Tier-1 centres, sustaining for one week an aggregate throughput of 500 MBytes/sec at CERN.

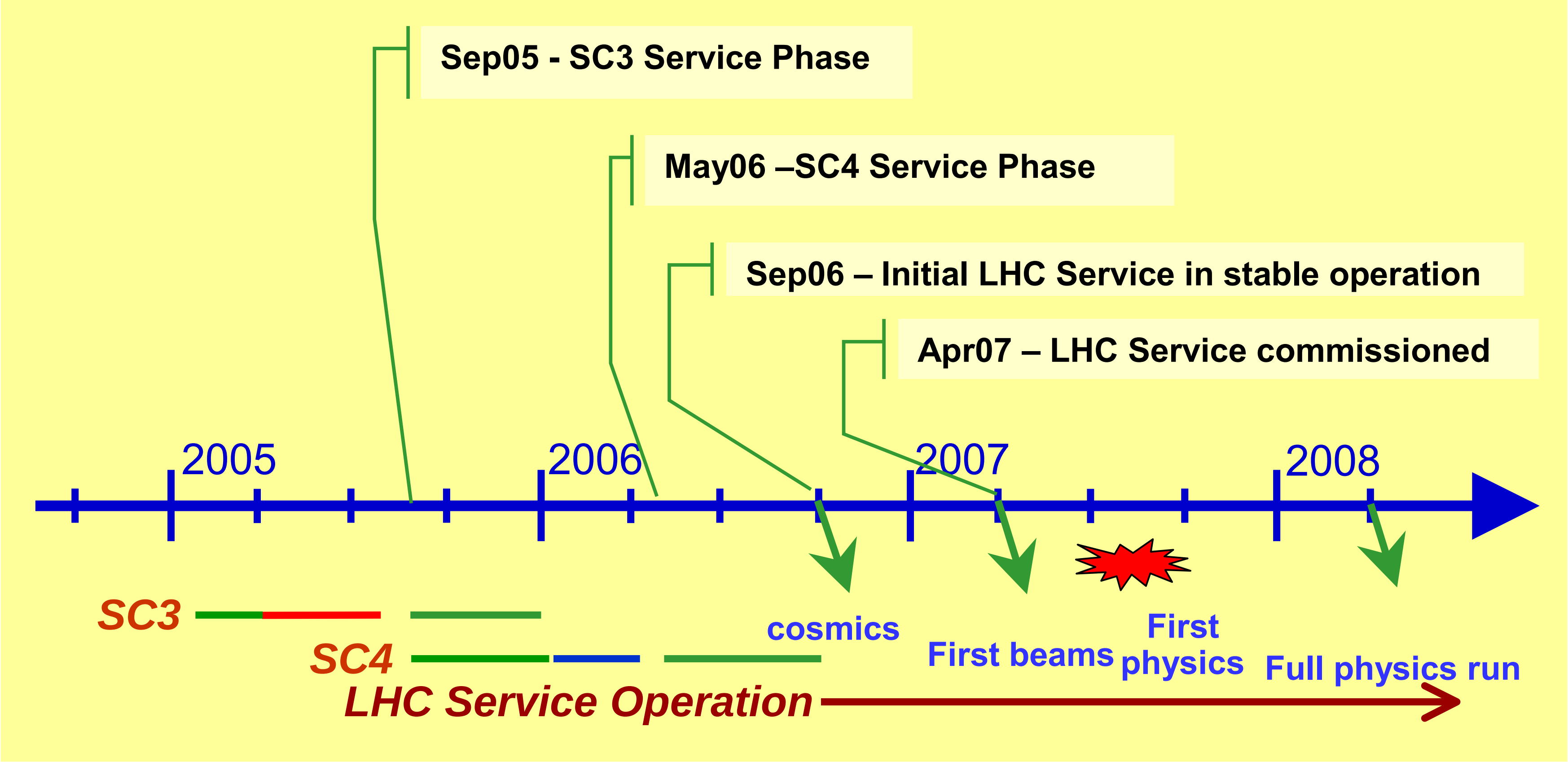
- Service Challenge 3

  Operate a reliable base service including most of the Tier-1 centres and some Tier-2s. Grid data throughput 1GB/sec, including mass storage 500 MB/sec (150 MB/sec & 60 MB/sec at Tier-1s).

- Service Challenge 4

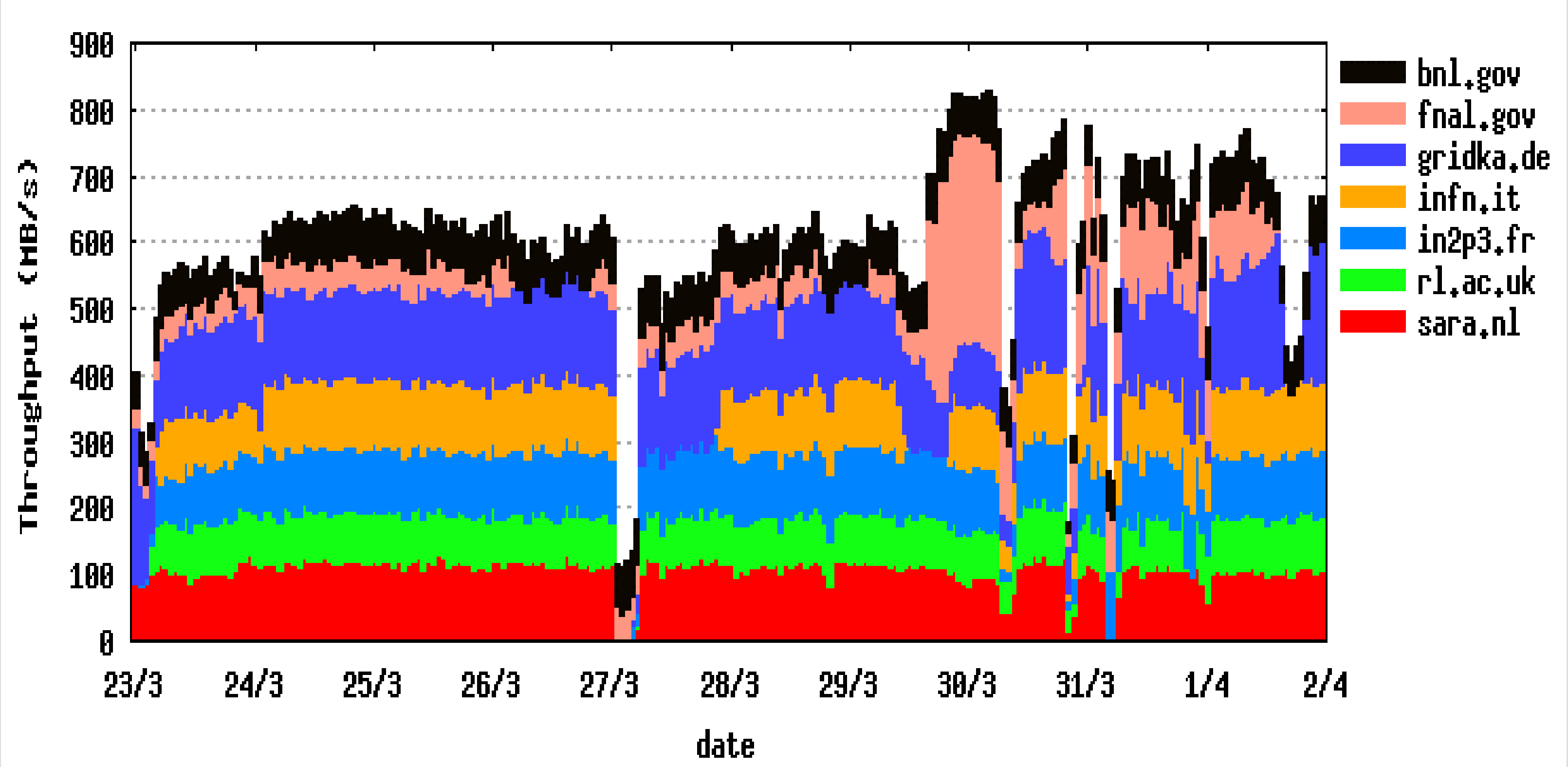
  Demonstrate that all of the offline data processing requirements expressed in the experiments' Computing Models, from raw data taking through to analysis, can be handled by the Grid at the full nominal data rate of the LHC

Schedules



- **SC3 – Currently finishing throughput phase, testing basic experiment software chains**
- **SC4 – All Tier-1s, major Tier-2s – sustain nominal final grid data throughput (~ 1.5 GB/sec)**
- **LHC Service in Operation – September 2006 – ramp up to full operational capacity by April 2007**

SC2 - Throughput to Tier1 from CERN



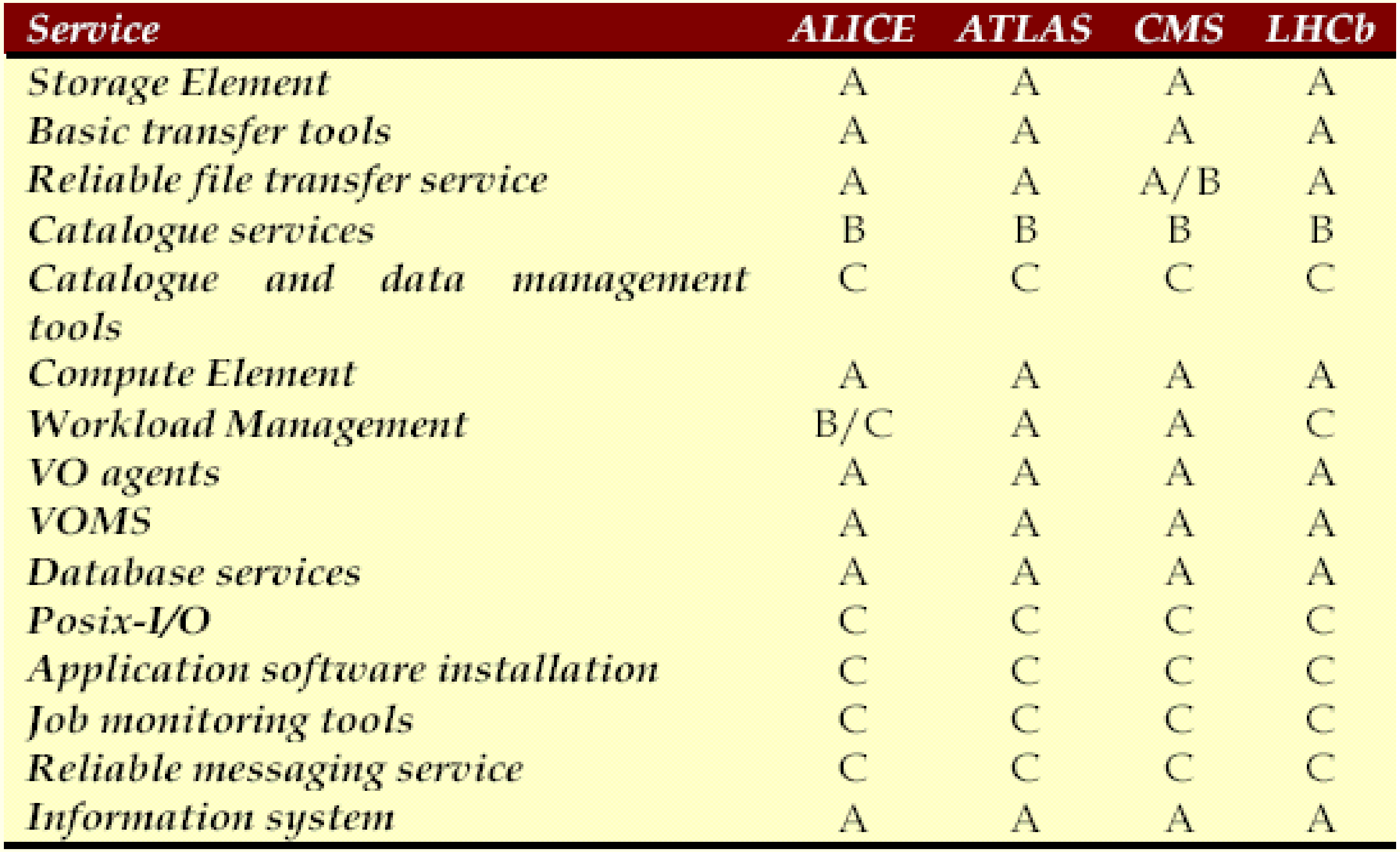
Has to use multiple TCP streams and multiple file transfers to fill up network pipe

SC3 Throughput Tests

Site	MoU Target (Tape)	Aver. MB/s (Disk)
ASGC	100	10
BNL	200	107
FNAL	200	185
GridKa	200	42
CC-IN2P3	200	40
CNAF	200	50
NDGF	50	129
PIC	100	54
RAL	150	52
NIKHEF	150	111
TRIUMF	50	34

- All Tier0 participating
- Using SRM interface
- July - low transfer rates and poor reliability of transfers between T0-T1
  - running at ~half the target of 1GB/s with poor stability
  - T1-T2 transfers – at much lower rates – on target
- Since then a better performance has been seen after resolving FTS and Castor problems at CERN

Baseline Services



Service	ALICE	ATLAS	CMS	LHCb
Storage Element	A	A	A	A
Basic transfer tools	A	A	A	A
Reliable file transfer service	A	A	A/B	A
Catalogue services	B	B	B	B
Catalogue and data management tools	C	C	C	C
Compute Element	A	A	A	A
Workload Management	B/C	A	A	C
VO agents	A	A	A	A
VOMS	A	A	A	A
Database services	A	A	A	A
Posix-I/O	C	C	C	C
Application software installation	C	C	C	C
Job monitoring tools	C	C	C	C
Reliable messaging service	C	C	C	C
Information system	A	A	A	A

Priority A: High priority and mandatory
Priority B: Standard solutions are required, but experiments could select different implementations
Priority C: Desirable to have a common solution, but not essential

LCG/EGEE Services in SC3

- Basic services (CE, SE, ...)
- New LCG/gLite service components tested in SC3
  - SRM Storage element at T0, T1 and T2
    SRM 1.1 interface to provide managed storage
  - FTS server at T0 and T1
    T0 and T1 to provide (reliable) File Transfer Service
  - LFC catalog at T0, T1 and T2
    Local catalogs to provide information about location of experiments' files and datasets
  - VOBOX at T0, T1 and T2
    For running experiment specific agents at a site

SRM Service

- SRM v1.1 insufficient
- Volatile, Permanent space
- Global space reservation: reserve, release, update (mandatory LHCb, useful ATLAS,ALICE).
- Permissions on directories mandatory
  - Prefer based on roles and not DN (SRM integrated with VOMS desirable)
- Directory functions (except mv)
- Pin/unpin capability
- Relative paths in SURL important for ATLAS, LHCb, not for ALICE

FTS Service

- First require base storage and transfer infrastructure (gridftp, SRM) to become available at high priority and to demonstrate sufficient quality of service
- Reliable transfer layer is valuable
- The gLite FTS seems to satisfy current requirements
- Experiments plan on integrating with FTS as an underlying service to their own file transfer and data placement services
- Interaction with fts (e.g catalog access) – can be implemented either in the experiment layer or integrating into FTS workflow
- Regardless of transfer system deployed – need for experiment-specific components to run at both Tier1 and Tier2
- Without a general service, inter-VO scheduling, bandwidth allocation, prioritisation, rapid address of security issues etc. would be difficult

Local File Catalog

- File Catalogues provide the mapping of Logical file names to GUID and Storage locations (SURL).
- Experiments need hierarchical name space (directories)
- Need some form of a collection notion (datasets, fileblocks, …)
- Need to have role-based security (admin, production, etc.)
- Support bulk-operations: Dump entire catalog
- Interfaces are required to:
  - POOL, Posix-like I/O service, WMS
    (e.g. Data Location Interface/Storage Index interfaces)
- LCG File Catalog fixes performance and scalability problems seen in EDG Catalogs and provides most of the required functionality
- Experiments rely on grid catalogs for locating files and datasets
- Experiment dependent information is in experiment catalogues

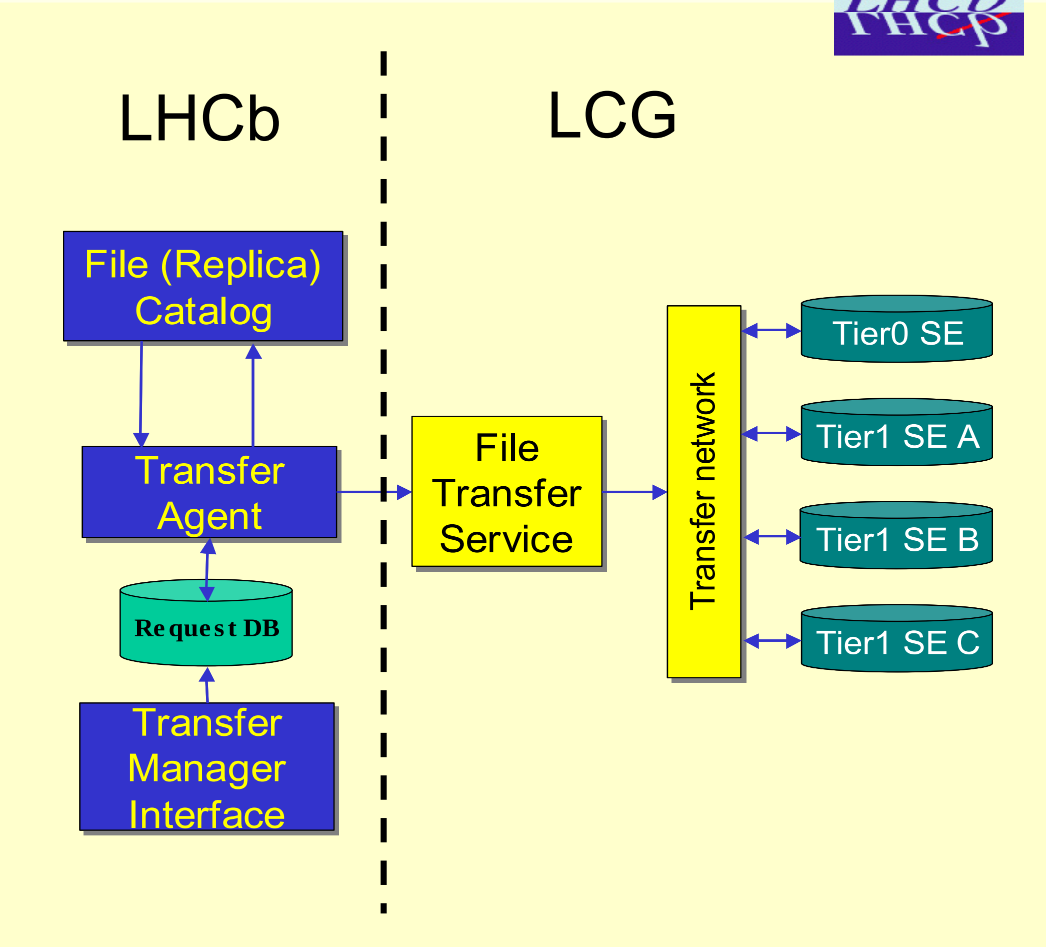
VOBox

- The VOBox is a machine where permanent-running processes (agents or services) can be deployed and where the required security, logging & monitoring can be incorporated.
- The VOBox provides a way to deploy VO specific upper layer middleware on the Site with the aim of filling the gap between existing LCG middleware and the VO needs.
- The VOBox is not to by-pass current middleware deployment but to strengthen & enhance it to meet the experiment specific requirements.

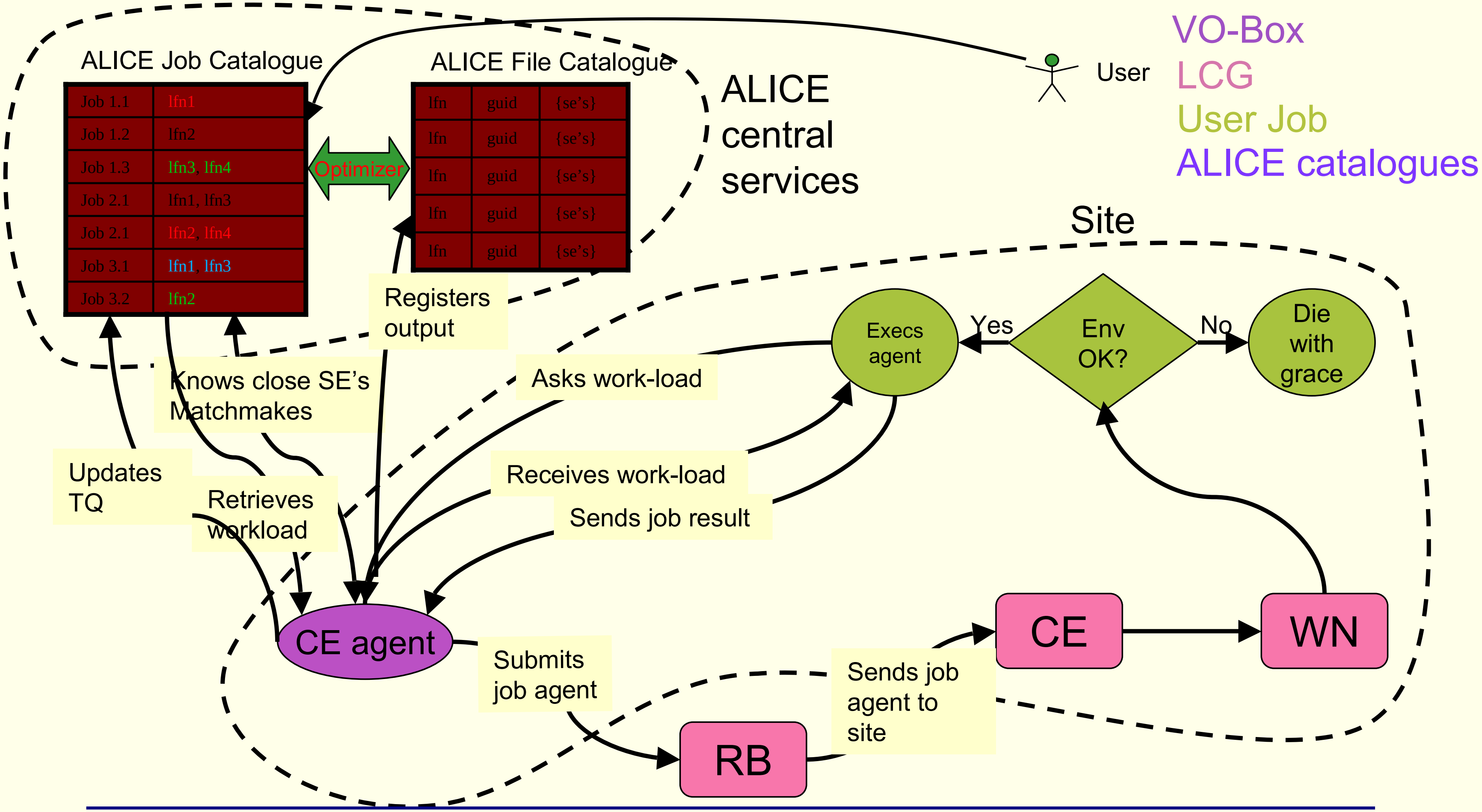
Experiment Integration

LHCb Architecture for using FTS

- Central Data Movement model based at CERN.
  - FTS+TransferAgent+ RequestDB
- TransferAgent+ReqDB developed for this purpose.
- Transfer Agent run on LHCb managed lxgate class machine



ALICE Workload Management



ALICE Job Catalogue

Job 1.1	lfn1
Job 1.2	lfn2
Job 1.3	lfn3, lfn4
Job 2.1	lfn1, lfn3
Job 2.1	lfn2, lfn4
Job 3.1	lfn1, lfn3
Job 3.2	lfn2

ALICE File Catalogue

lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}
lfn	guid	{se's}

Service Level Definition

Class	Description	Downtime	Reduced	Degraded	Availability
C	Critical	1 hour	1 hour	4 hours	99%
H	High	4 hours	6 hours	6 hours	99%
M	Medium	6 hours	6 hours	12 hours	99%
L	Low	12 hours	24 hours	48 hours	98%
U	Unmanaged	None	None	None	None

- **Downtime defines the time between the start of the problem and restoration of service at minimal capacity (i.e. basic function but capacity < 50%)**
- **Reduced defines the time between the start of the problem and the restoration of a reduced capacity service (i.e. >50%)**
- **Degraded defines the time between the start of the problem and the restoration of a degraded capacity service (i.e. >80%)**
- **Availability defines the sum of the time that the service is down compared with the total time during the calendar period for the service. Site wide failures are not considered as part of the availability calculations. 99% means a service can be down up to 3.6 days a year in total. 98% means up to a week in total.**
- **None means the service is running unattended**

Example Services & Service Levels

Service	Service Level	Runs Where
Resource Broker	Critical	Main sites
Compute Element	High	All sites
MyProxy	Critical	
BDII	Critical	Global
R-GMA	High	
LFC	High	All sites (ATLAS, ALICE) CERN (LHCb)
FTS	High	T0, T1s (except FNAL)
SRM	Critical	All sites

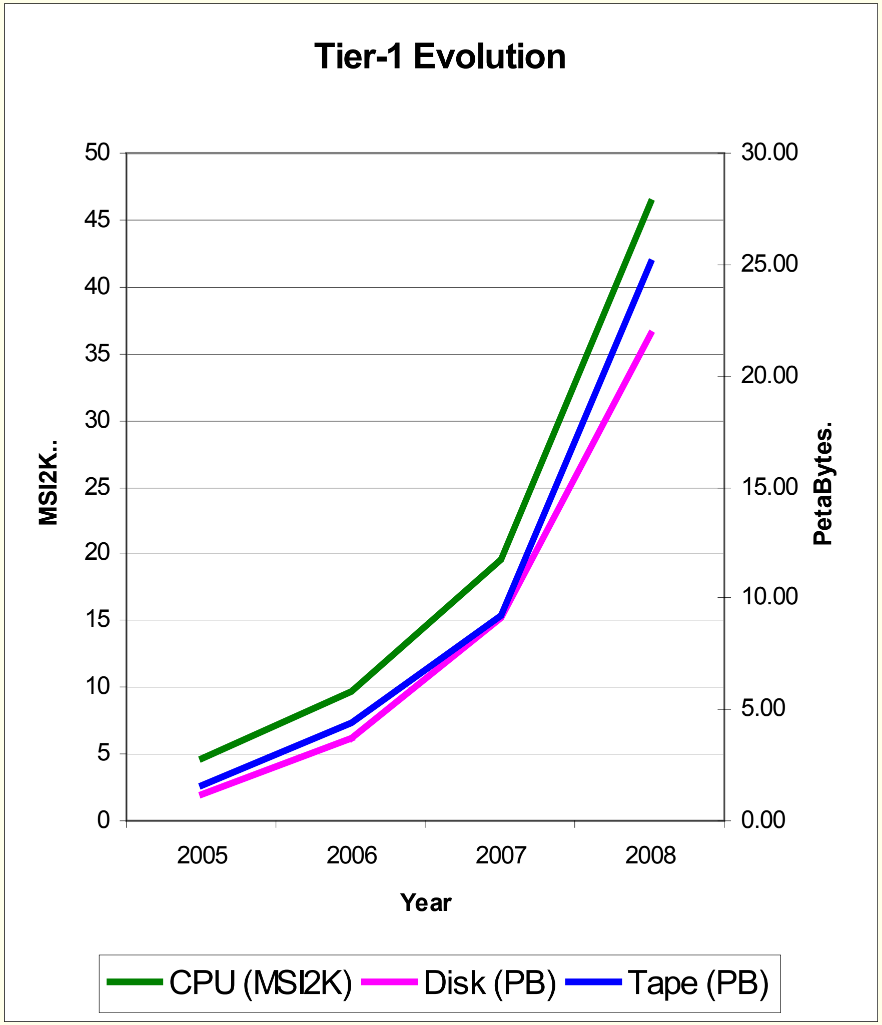
This list needs to be completed and verified

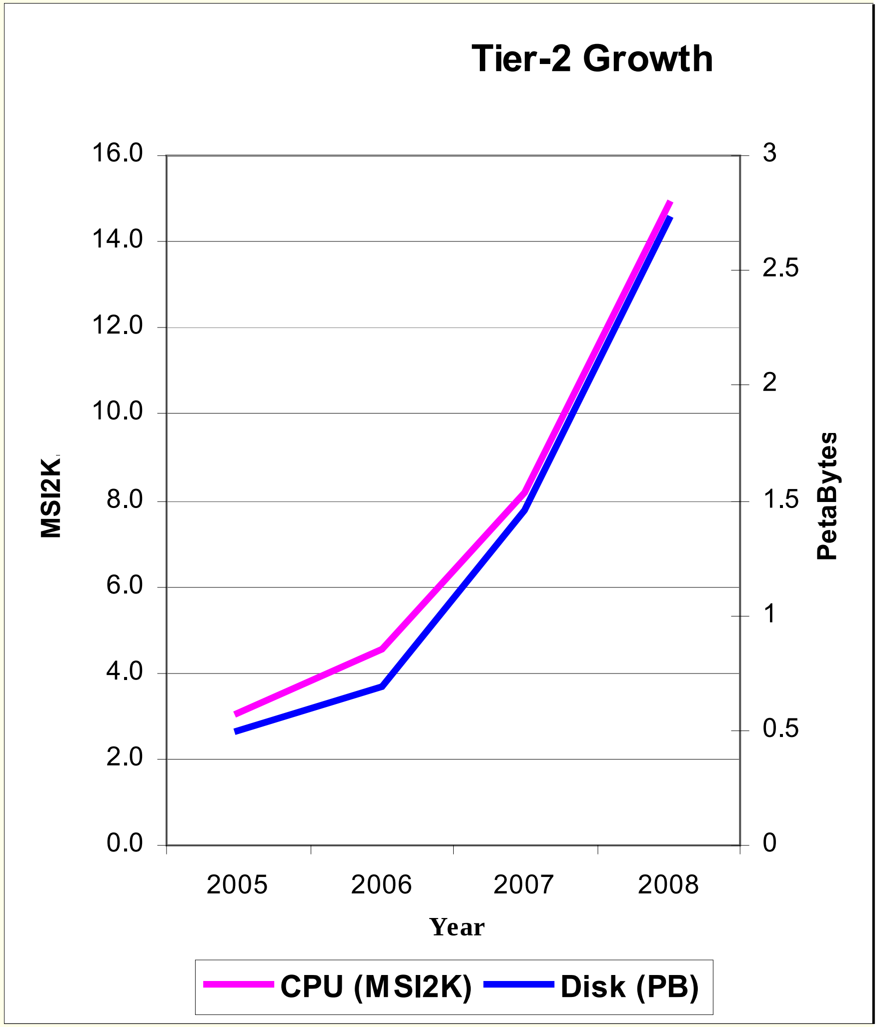
Then timescales for achieving the necessary service levels need to be agreed

Building WLCG Service

- All services required to handle production data flows now deployed at all Tier1s and participating Tier2s
- Bring the remaining Tier2 centres into the process
- Getting the (stable) data rates up to the target
- Identify the additional Use Cases and functionality
- Bring core services up to robust 24 x 7 production standard required
  - Need to use existing experience and technology...
  - Monitoring, alarms, operators, SMS to 2nd / 3rd level support…
- (Re-)implement Required Services at Sites so that they can meet MoU Targets
  - Measured through Site Functional Tests
  - Delivered Availability, maximum intervention time etc.
- Goal is to build a cohesive service out of a large distributed community

Building WLCG Service





Extra

VOBox Services: ALICE

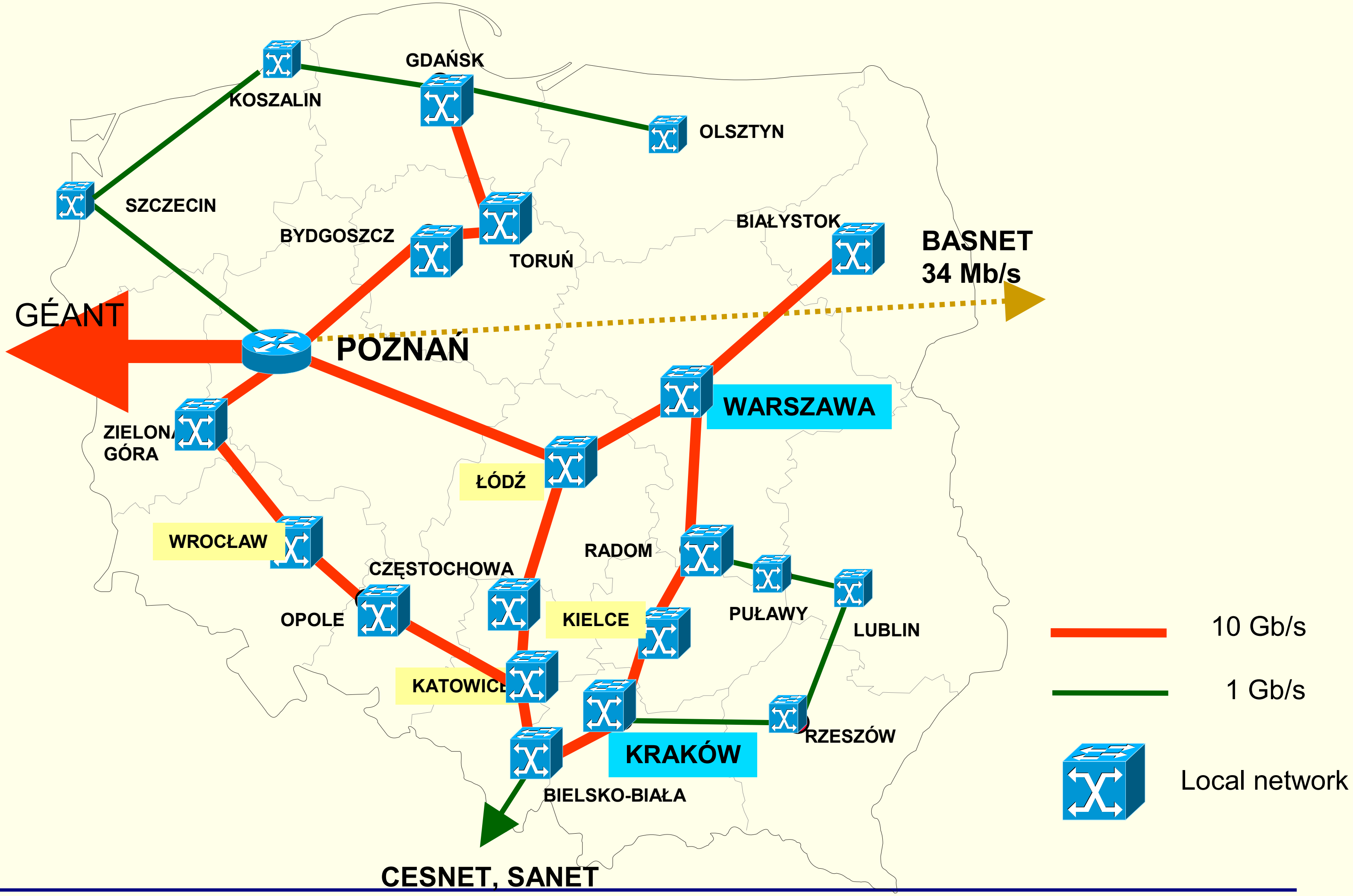
AliEn and monitoring agents and services running on the VO node:

- AliEn Computing Element (CE) (Interface to LCG RB)
- Storage Element Service (SES)
  - interface to local storage (via SRM or directly)
- File Transfer Daemon (FTD)
  - scheduled file transfers agent
    (possibly using FTS implementation)
- Cluster Monitor (CM) – local queue monitoring
- MonALISA – general monitoring agent
- PackMan (PM) – software distribution and management
- xrootd – application file access
- Agent Monitoring (AmOn)

Polish LCG/EGEE Centres

- Cracow:
  - CYFRONET – Academic Computer Centre
  - *http://www.cyfronet.pl/*
- Warsaw:
  - ICM – Interdisciplinary Centre for Mathematical and Computational Modelling
  - *http://www.icm.edu.pl/*
- Poznań
  - PCSS – Poznań Supercomputing and Networking Centre
  - *http://www.man.poznan.pl/*

Polish Network



Polish Tier2

- Poland is a federated Tier-2
- HEP LHC community: *~60 people*
- Each of the computing centres naturally will support mainly 1 experiment
  - Cracow – ATLAS
  - Warsaw – CMS
  - Poznań – ALICE
- Currently setting up for participation in SC3, SC4
- In the future each centre will probably be about 1/3 of the average/small Tier2 for a given experiment