LCG Internal Review

Computing Fabric Overview

Goal

- The goal of the Computing Fabric Area is to prepare the T0 and T1 centre at CERN. The T0 part focuses on the mass storage of the raw data, the first processing of these and the data export (e.g. raw data copies), while the T1 centre task is primarily the analysis part.

- There is currently no physical or financial distinction/separation between the T0 installation and the T1 installation at CERN. (roughly 2/3 to 1/3)

- The plan is to have a flexible, performing and efficient installation based on the current model, to be verified until 2005 taking the computing models from the Experiments as input (Phase I of the LCG project).

Strategy

- **Continue, evolve and expand the current system**
  profit from the current experience : number of total users will not change,
  Physics Data Challenges of LHC experiments, running Experiments (CDR of COMPASS + NA48 up 150 MB/s, they run their level 3 filter on Lxbatch)

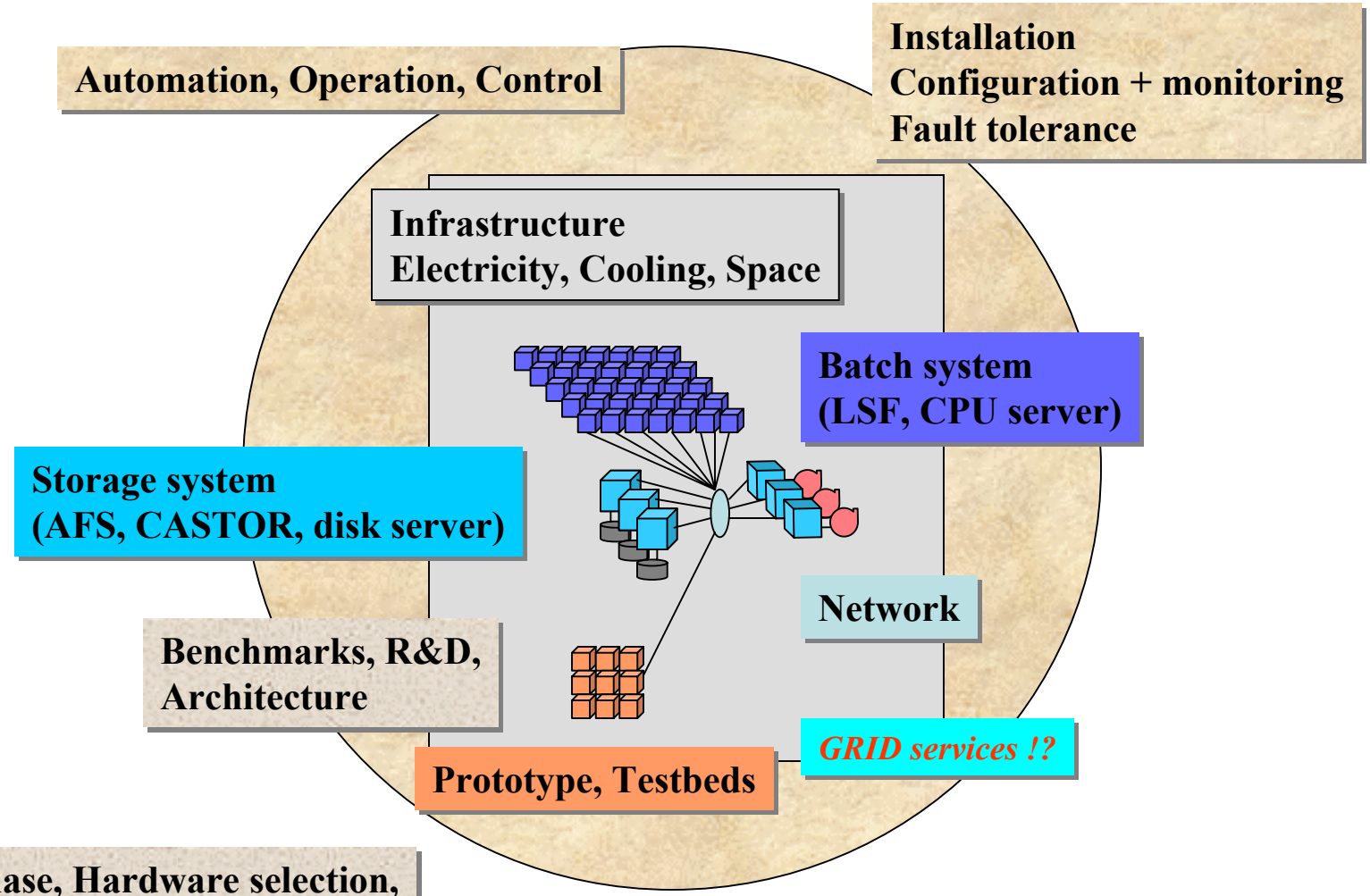
BUT do in parallel :

- **R&D activities and Technology evaluations**
  SAN versus NAS, iSCSI, IA64 processors, ….
  PASTA, infiniband clusters, new filesystem technologies,…..

- **Computing Data Challenges to test scalabilities on larger scales**
  “bring the system to it’s limit and beyond “
  we are very successful already with this approach, especially with the “beyond” part

- **Watch carefully the market trends**

View of different Fabric areas



Coupling of components through hardware and software

Infrastructure

There are several components which make up the Fabric Infrastructure :

- **Material Flow**
  organization of market surveys and tenders, choice of hardware, feedback from R&D, inventories, vendor maintenance, replacement of hardware
  → major point is currently the negotiation of different purchasing procedures for the procurement of equipment in 2006

- **Electricity and cooling**
  refurbishment of the computer center to upgrade the available power from 0.8 MW today to 1.6 MW (2007) and 2.5 MW in 2008
  → development of power consumption in processors problematic

- **Automation procedures**
  Installation+Configuration+Monitoring+Fault Tolerance for all nodes
  Development based on the tools from the DataGrid project
  Already deployed on 1500 nodes, good experience, still some work to be done
  several Milestones met with little delay

further details in the next talk

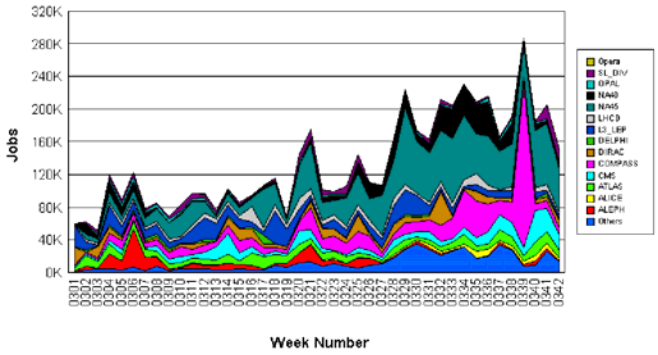
Services

The focus of the computing fabric are the services and they are integral part of the IT managerial infrastructure

- Management of the farms
- Batch scheduling system
- Networking
- Linux
- Storage management

but the service is of course currently not only for the LHC Experiments IT supports about 30 Experiments, engineers, etc. Resource usage dominated by punctual LHC physics data challenges and running experiments (NA48, COMPASS,….)

Batch Scheduler



- Using LSF from Platform Ccomputing, commercial product
- deployed on 1000 nodes,
  10000 concurrent jobs in the queue on average,
  200000 jobs per week
- very good experience, fair share for optimal usage of resources
  current reliability and scalability issues are understood
  adaptation in discussion with users
  → average throughput versus peak load and real time response
- mid 2004 to start another review of available batch systems
  → choose in 2005 the batch scheduler for Phase II

Network

- **Network infrastructure based on ethernet technology**
- **Need for 2008 a completely new (performance) backbone in the centre based on 10 Gbit technology. Today very few vendors offer this multiport, non-blocking, 10 Gbit router.**
  We have an Enterasys product already under test (openlab, prototype)
- **Timescale is tight :**

Q1 2004 market survey
Q2 2004 install 2-3 different boxes, start thorough testing
→
prepare new purchasing procedures, finance committee
vendor selection, large order
→
Q3 2005 installlation of 25% of new backbone
Q3 2006 upgrade to 50%
Q3 2007 100% new backbone

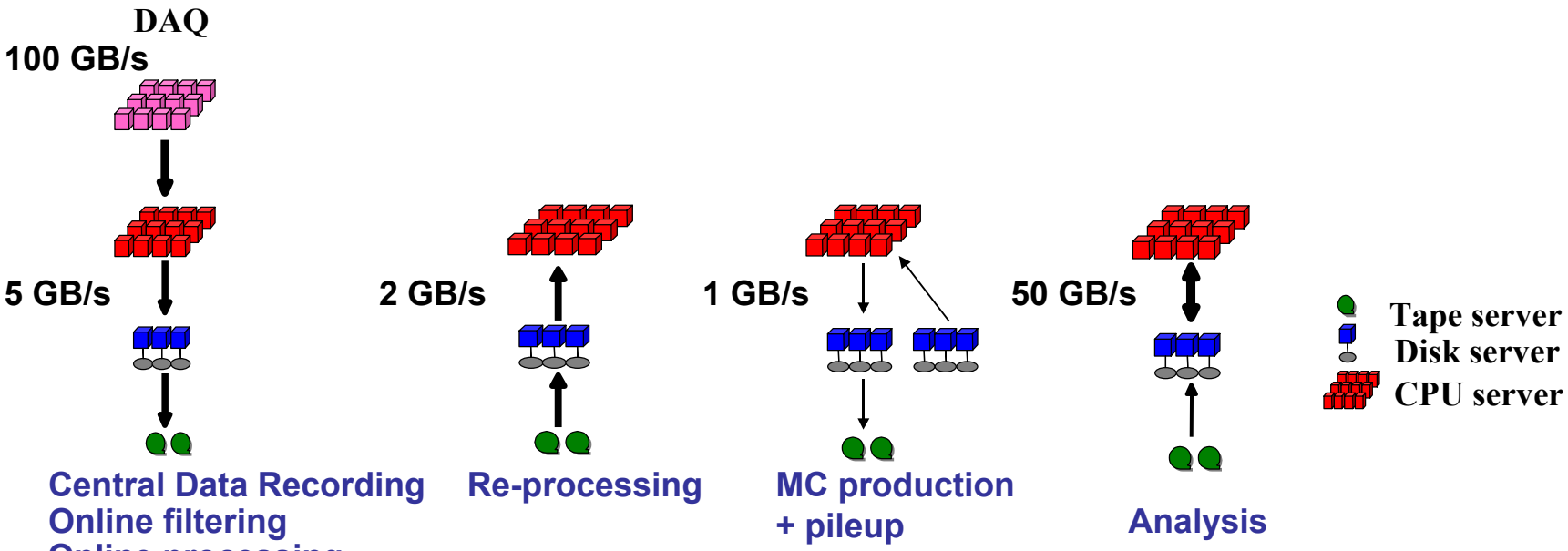
Dataflow Examples

scenario for 2008

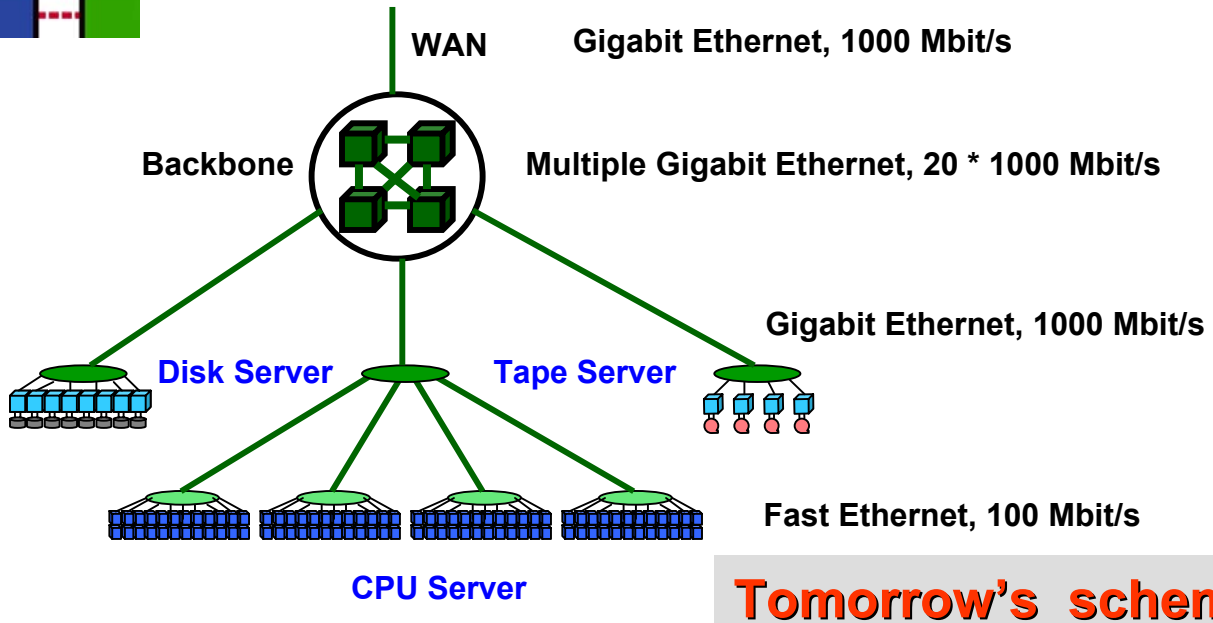
Implementation details depend on the computing models of the experiments

→ more input from the 2004 Data Challenges

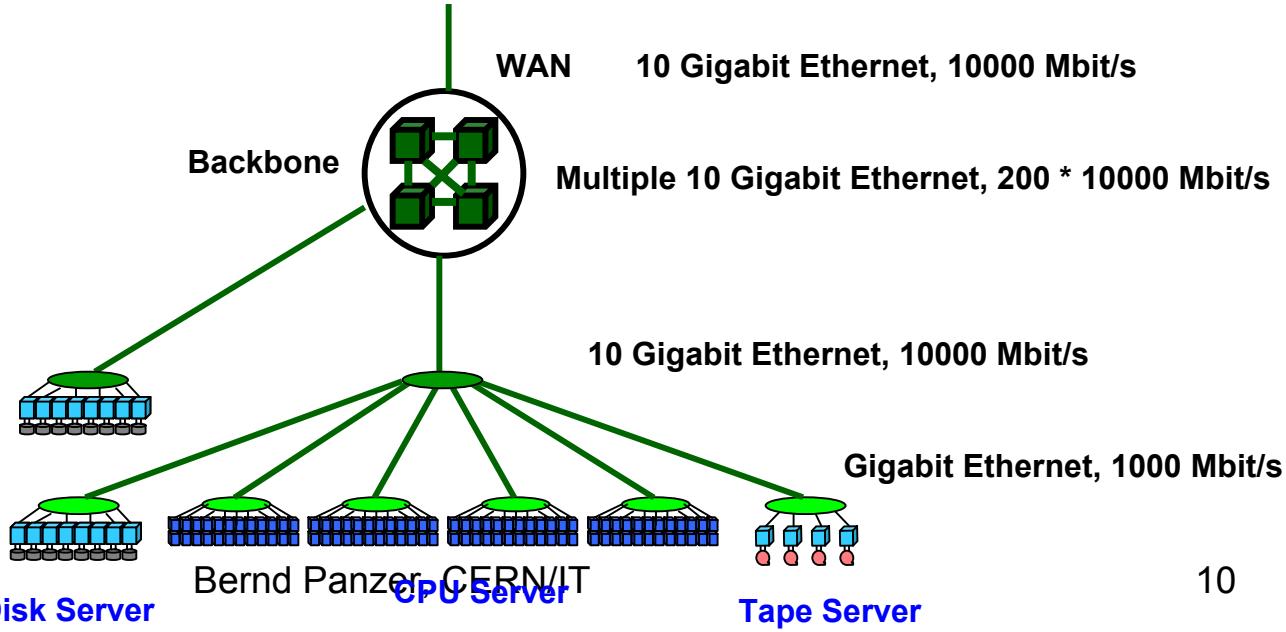
→ modularity and flexibility in the architecture are important



Today's schematic network topology



Tomorrow's schematic network topology



Wide Area Network

- Currently 4 lines 21 Mbit/s, 622 Mbits/s , 2.5 Gbits/s (GEANT),
  dedicated 10 Gbit/s line (starlight chicago, DATATAG),
  next year full 10 Gbit/s production line

- Needed for import and export of data, Data Challenges,
  todays data rate is 10 – 15 MB/s

- Tests of mass storage coupling starting (Fermilab and CERN)

- Next year more production like tests with the LHC experiments
  CMS-IT data streaming project inside the LCG framework
  tests on several layers : bookkeeping/production scheme, mass storage coupling, transfer protocols (gridftp, etc.), TCP/IP optimization

- 2008 :
  multiple 10 Gbit/s lines will be availble with the move to 40 Gbit/s connections
  CMS and LHCb will export the second copy of the raw data to the T1 center ,
  ALICE and ATLAS want to keep the second copy at CERN (still ongoing discussion)

Storage (I)

AFS (Andrew File System)

- A team of 2.2 FTE takes care of the shared distributed file system to provide access to the home directories (small files, programs, calibration, etc.) of about 14000 users.
- Very popular, growth rate for 2004 : 60 % (4.6 TB → 7.6 TB)
  expensive compared to bulk data storage (factor 5-8), automatic backup, high availability (99 %), user perception different
- GRID job software environment distribution 'preferred' through shared file system solution per site
  → file system demands (performance, reliability, redundancy,etc.)
- Evaluation of different products have started
  expect a recommendation by mid 2004, collaboration with other sites (e.g. CASPUR)

Storage (II)

CASTOR

- CERN development of a Hierarchical Storage Management system (HSM)
- Two teams are working in this area : Developer (3.8 FTE) and Support (3 FTE)
  Support to other institutes currently under negotiation (LCG, HEPCCC)
- Usage : 1.7 PB of data with 13 million files,
  250 TB disk layer and 10 PB tape storage
  Central Data Recording and data processing
  NA48 0.5 PB COMPASS 0.6 PB LHC Exp. 0.4 PB
- Current CASTOR implementation needs improvements → New CASTOR stager
  - A pluggable framework for intelligent and policy controlled file access scheduling
  - Evolvable storage resource sharing facility framework rather than a total solution
- Carefully watching the tape technology developments (not really commodity)
  in depth knowledge and understanding is key

Linux

- 3.5 FTE team for Farms and Desktop
- Certification of new releases, bugfixes, security fixes,
  kernel expertise → improve performance and stability
- Certification group with all stakeholders : experiments, IT, accelerator, etc.
- Current distribution based on RedHat Linux
  major problem now : change in company strategy
  drop the free distributions and concentrate on the business with
  licenses and support for 'enterprise' distributions
- We are together with HEP community negotiating with RH
- Several alternative solutions were investigated : all need more money and/or more manpower
- Strategy is still to continue with Linux (2008 →)

Grid – Fabric coupling

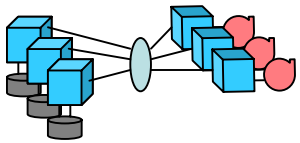
- Ideally clean interface and Grid middleware and services are one layer above the Fabric
  → reality is more complicated (intrusive)
- New research concept meets conservative production system
  → inertia and friction
- Authentication, security, storage access, repository access, job scheduler usage, etc. different implementations and concepts
  → adaptation and compromises necessary
- Regular and good collaboration between the teams established, still quite some work to be done
- Some milestones are late by several months (Lxbatch Grid integration)
  → late LCG-1 release and problem resolving in the GRID-Fabric API's more difficult than expected

further details in the next talk

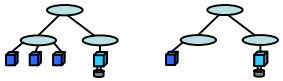
General Fabric Layout

New software , new hardware (purchase)

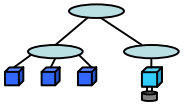
R&D cluster
(new architecture
and hardware)



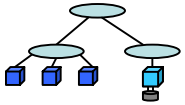
Development cluster
GRID testbeds



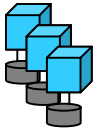
Certification cluster
Main cluster 'en miniature'



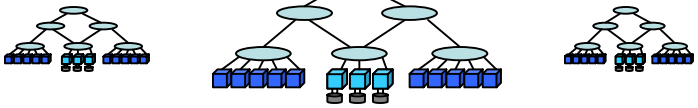
Benchmark and performance cluster
(current architecture and hardware)



Service control and management
(e.g. stager, HSM, LSF master,
repositories, GRID services,
CA, etc



Main fabric cluster

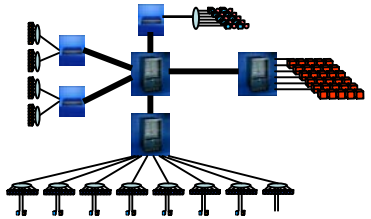


old current new

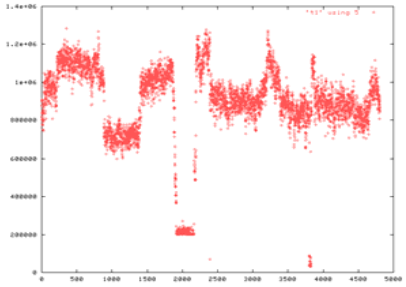
2-3 hardware generations
2-3 OS/software versions
4 Experiment environments

Data Challenges

- Physics Data Challenges (MC event production, production schemes, middleware)



- ALICE IT Mass Storage Data Challenges
  2003 → 300 MB/s, 2004 → 450 MB/s, 2005 → 700 MB/s
  preparations for the ALICE CDR in 2008 → 1.2 GB/s
- Online DCs (ALICE event building, ATLAS DAQ)



- IT scalability and performance DCs (network, filesystems, tape storage → 1 GB/s)
- WAN coupling of HSM systems, data export and import

➢ Architecture testing and verification, computing models, scalability
→ needs large dedicated resources, avoid interference with production system

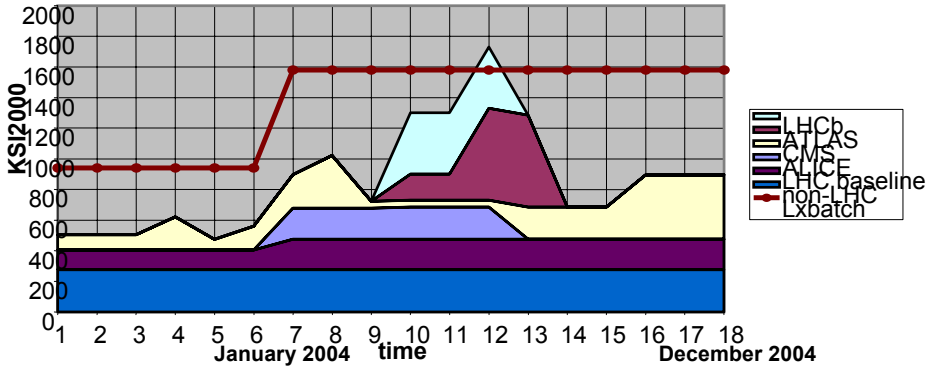
Computer center today

- **Main fabric cluster (Lxbatch/Lxplus resources)**
  - → physics production for all experiments
    Requests are made in units of Si2000
  - → 1200 CPU server, 250 disk server, ~ 1100000 Si2000, ~ 200 TB
  - → 50 tape drives (30MB/s, 200 GB cart.)
    10 silos with 6000 slots each == 12 PB capacity

- **Benchmark,performance and testbed clusters (LCG prototype resources)**
  - → computing data challenges, technology challenges, online tests, EDG testbeds, preparations for the LCG-1 production system, complexity tests
  - → 600 CPU server, 60 disk server, ~500000 Si2000, ~ 60 TB
    current distribution :
    220 CPU nodes for LCG testbeds and EDG
    30 nodes for application tests (Oracle, POOL, etc.)
    200 nodes for the high performance prototype(network,ALICE DC, openlab)
    150 nodes in Lxbatch for physics DCs

Resource Planning

- Dynamic sharing of resources between the LCG prototype installation and the Lxbatch production system.
- Primarily Physics data challenges on Lxbatch and computing data challenges on the prototype
- IT Budget for the growth of the production system will be 1.7 million in 2004 and the same in 2005.
- Resource discussion and planning in the PEB



LCG Materials Expenditure at CERN

	2001	2002	2003	2004	2005	TOTAL
PROTOTYPE						
Processors -		422	220	300	500	
PC R&D + admin tests		100		50		
Disk storage -		919	120	930	780	
Mass storage -		1323	100	200	200	
Systems admin contract -		380	200		400	
Physics WAN/LAN -			225	500	500	
LCG Associate s		290	310	360	360	
TOTAL PROTOTYPE	**2454**	**3434**	**1175**	**2340**	**2740**	**12143**
PREPARING CC FOR PHASE 2						
Vault		1665				
Substation			1590	805	550	
Main computer room			850	1880	500	
Air conditioning				300	450	
TOTAL PREPARING CC for PHASE 2	**150**	**1665**	**2440**	**2985**	**1500**	**8740**
TOTAL PROTOTYPE OPERATION	**2604**	**5099**	**3615**	**5325**	**4240**	**20883**

Staffing

- 25.5 FTE from IT division are allocated in the different services to LHC activities. These are fractions of people, LHC experiments not yet the dominating users of services and resources
- 12 FTE from LCG and 3 FTE from the DataGrid (EDG) project are working in the area of service developments (e.g. security, automation) and evaluation (benchmarks, data challenges, etc.)
  This number (15) will decrease to 6 by mid 2004 (EDG ends in February, end of LCG contracts (UPAS, students, etc.)
  Fellows and Staff continue until 2005

Re-costing

- Re-costing exercise during April and May
- Representatives from IT and the 4 LHC experiments
- Review the equipment cost for phase 2 of LCG and 2009-2010 take into account slight changes in the model and the adjusted requirements from the experiments
  → Excel table → explanation note
- LCG seminar in July and paper in August

Re-costing results

Resource	Old 2006-08	New 2006-08	New- Old 2006-08	Old 2009-10	New 2009-10	New - Old 2009-10
CPU+LAN	17.7	19.5	1.8	6.3	6.8	0.5
Disk	6.3	11.9	5.6	2.2	2.9	0.7
Tape	22.5*	27.8	5.3	19.2	17.6	-1.6
WAN	11.4	6.0	- 4.4	6.8	4.0	-2.8
Sysadmin	7.9	3.5	- 4.4	6.6	3.0	-3.6
SUM	65.8	68.7	2.9	41.1	34.3	-6.8
Budget		60.0			34.0	

* **A bug in the original paper is here corrected**

All units in [million CHF]

Comparison

	2008 prediction	2003 status
➢ Hierarchical Ethernet network	280 GB/s	2 GB/s
➢ ~ 8000 mirrored disks	(4 PB)	2000 mirrored disk 0.25 PB
➢ ~ 3000 dual CPU nodes	(20 MSI2000)	2000 nodes 1.2 MSI2000
➢ ~ 170 tape drives	(4 GB/s)	50 drives 0.8 GB/s
➢ ~ 25 PB tape storage		10 PB

→The CMS HLT will consist of about 1000 nodes with 10 million SI2000 !!

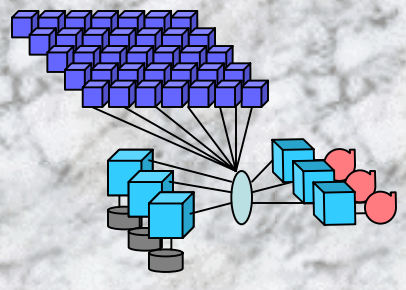
External Fabric relations

Collaboration with Industry
openlab
HP, INTEL, IBM, Enterasys, Oracle
-- 10 Gbit networking
-- new CPU technology
-- possibly , new storage technology

LCG
-- Hardware resources
-- Manpower resources

Collaboration with India
-- filesystems
-- Quality of Service

CERN IT
- Main Fabric provider



GDB working groups
-- Site coordination
-- Common fabric issues

Collaboration with CASPUR
--harware and software benchmarks and tests, storage and network

External network
-- DataTag, Grande
-- Data Streaming project with Fermilab

LINUX
-- Via HEPIX RedHat license coordination
inside HEP (SLAC, Fermilab)
certification and security

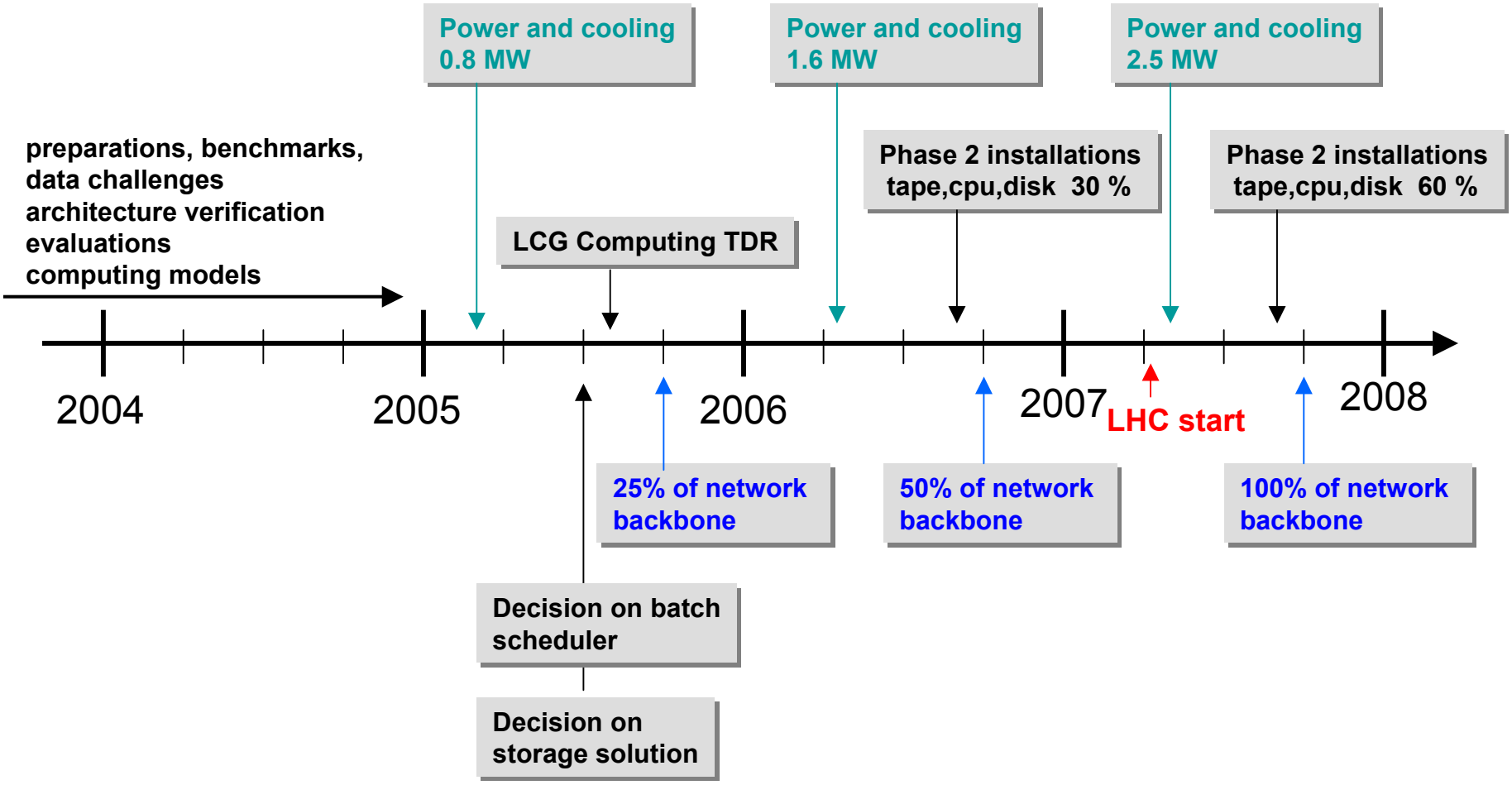
CASTOR
-- SRM definition and implementation (Berkeley, Fermi, etc.)
-- mass storage coupling tests (Fermi)
-- scheduler integration (Maui, LSF)
-- support issues (LCG, HEPCCC)

EDG, WP4
-- Installation
-- Configuration
-- Monitoring
-- Fault tolerance

GRID Technology and deployment
-- Common fabric infrastructure
-- Fabric ←→ GRID interdependencies

Online-Offline boundaries
-- workshop and discussion with Experiments
-- Data Challenges

Timeline


Power and cooling 0.8 MW
Power and cooling 1.6 MW
Power and cooling 2.5 MW
preparations, benchmarks, data challenges architecture verification evaluations computing models
Phase 2 installations tape,cpu,disk 30 %
Phase 2 installations tape,cpu,disk 60 %
LCG Computing TDR
2004
2005
2006
2007
LHC start
2008
25% of network backbone
50% of network backbone
100% of network backbone
Decision on batch scheduler
Decision on storage solution

Summary

- The computing models need to be defined in more detail.
- Positive collaboration with outside Institutes and Industry.
- Timescale is tight, but not problematic.
- Successful Data Challenges and most milestones on time.
- The pure technology is difficult (network backbone, storage), but the real worry is the market development.