Report from HEPCCC

Tobias Haas

XV HTASC

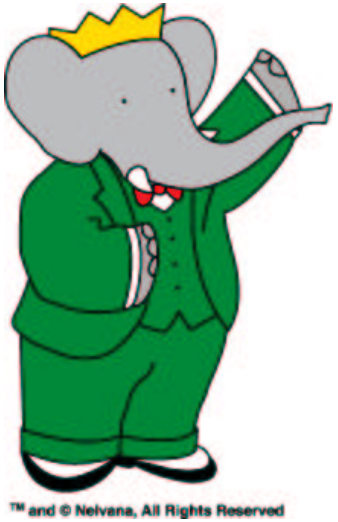
2 October, 2003

CERN

HEPCCC Agenda
(27/28 June, CC-IN2P3, Lyon)

- T. Kachelhoffer, CC-IN2P3: Life in a multi-grid multi-experiment multi-disciplinary centre
- D. Boutigny, LAPP: Tier1/A experience from a remote site (BABAR case) ☞
- W. von Rüden: EDG, LCG and EGEE Status Reports ☞
- A. Boehnlein: Update on D0 and CDF computing models and experience
- W. von Rüden: Spam Fighting at CERN ☞
- G. Wormser: Progress with I_HEPCCC ☞ *(tomorrow)*
- L. Price: What's next for ESnet? ☞
- W. von Rüden: CHEP04 and Cern School of Computing ☞
- Tobias Haas: HTASC report

Tier1/A Experience from a Remote Site the BaBar case



Dominique Boutigny

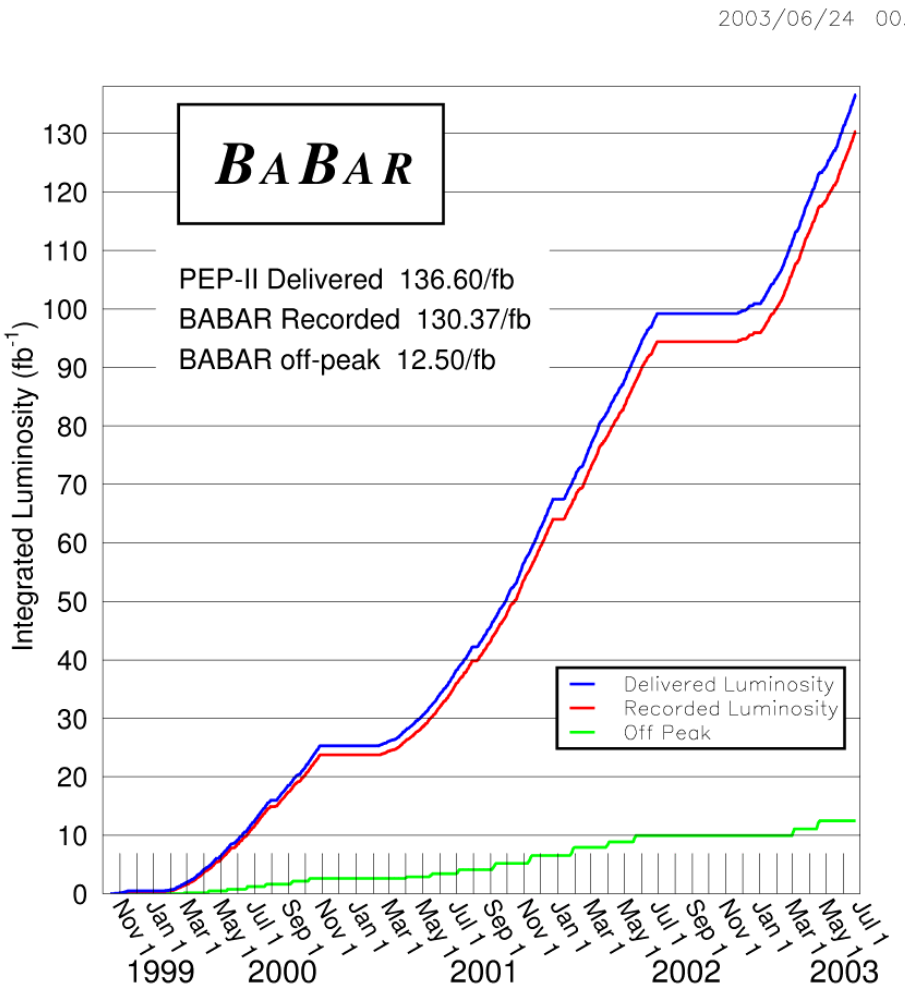
LAPP – CNRS/IN2P3

HEPCCC meeting

CCIN2P3 - June 27, 2003

Some Numbers on the BaBar Experiment (1)

- BaBar is taking physics data since 1999
  - Recorded Lumi: ~130 fb^{-1}
  - ~128 Million bb pairs
  - > 560 Million hadronic events
- 2 formats:
  - Micro: ~7 kB/evt
    - Suitable for standard analyses
  - Mini: ~10 kB/evt (on top of Micro)
    - For detector studies and detailed analyses
- MC simulation:
  - 3× # real data events



Some Numbers (2)

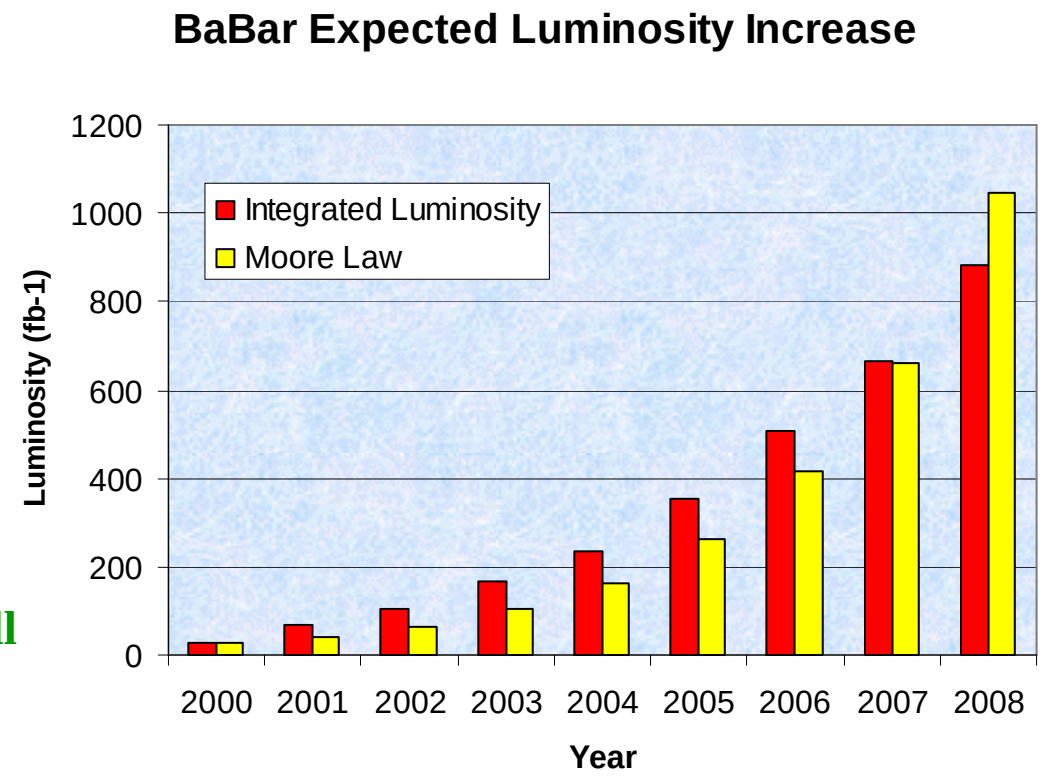
- The Integrated luminosity is increasing faster than the Moore law up to 2006

  > 500 fb^{-1} in 2006

  ⇒ BaBar computing cost is increasing with time

- <u>Interesting fact:</u> The Moore law wins after 2006

➔ This is nice as LHC computing will compete with BaBar at that time



- Necessity to setup a computing model which is making the best possible use of the resources available throughout the collaboration

The Distributed Computing Model

- Since 2001 BaBar is using a multi-Tier computing model similar to LHC ones
  - Tier A are major centers providing analysis access to the whole collaboration.
    - Tier A alone or all together have a copy of all the data (Micro + Mini)
    - Each Tier A holds a significant fraction of the whole dataset
    - Filtering on large or whole data sample is possible
  - Tier B were foreseen in the initial model
    - They have not been really implemented in BaBar
    - Similar to UK centers having significant computing resources
  - Tier C are small individual sites, they can copy a small fraction of the data from a Tier-A
    - Local analyses
    - N-tuple analyses
- The MC production is performed in ~23 sites (Tier A or C).
  - Produced events are sent back to SLAC and possibly re-distributed to other Tier-A

Formal Aspects of the Tier-A model

- Each Tier-A is formally discussed within the collaboration
  - MOU signed up by Tier-A/ country and BaBar
  - The Tier-A hardware contributions are financially evaluated yearly
    - Mechanism to deduce part of the Tier-A value from the country's common fund contribution
- The Computing Steering Group is in charge to evaluate and control the Tier-A contributions
  - Report to the International Finance Committee

The French Tier-A in Lyon

- CCIN2P3 (Lyon) has been the first Tier-A setup in BaBar
  - Providing accounts to every BaBar members ~150 accounts created for non French BaBar members
  - MOU foresees increased computing capacity up to 2005
    - CPU
    - Disk space
- Very successful up to now

- Only Objectivity data are present up to now
- Will move to Root data as soon as the new format is in place

- Manpower
  - 1 person dedicated to BaBar at CCIN2P3 ➔ Crucial !
  - +2 × 0.5 FTE from BaBar physicists to run the Tier-A
  - More people were necessary to setup the system initially

Made the choice from the beginning not to become a clone of SLAC

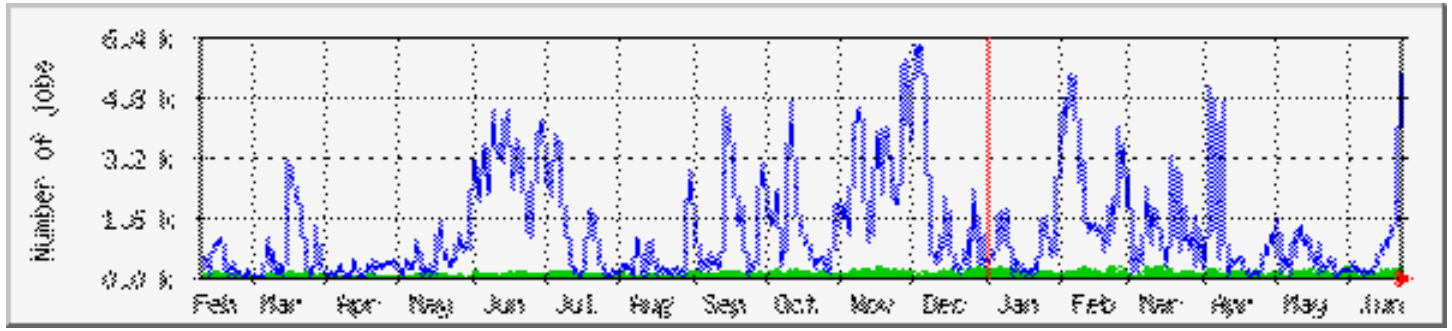
It is important to be able to implement our own solutions in agreement and in good cooperation with the rest of BaBar

CPU (1)

BaBar is sharing resources with other experiments:

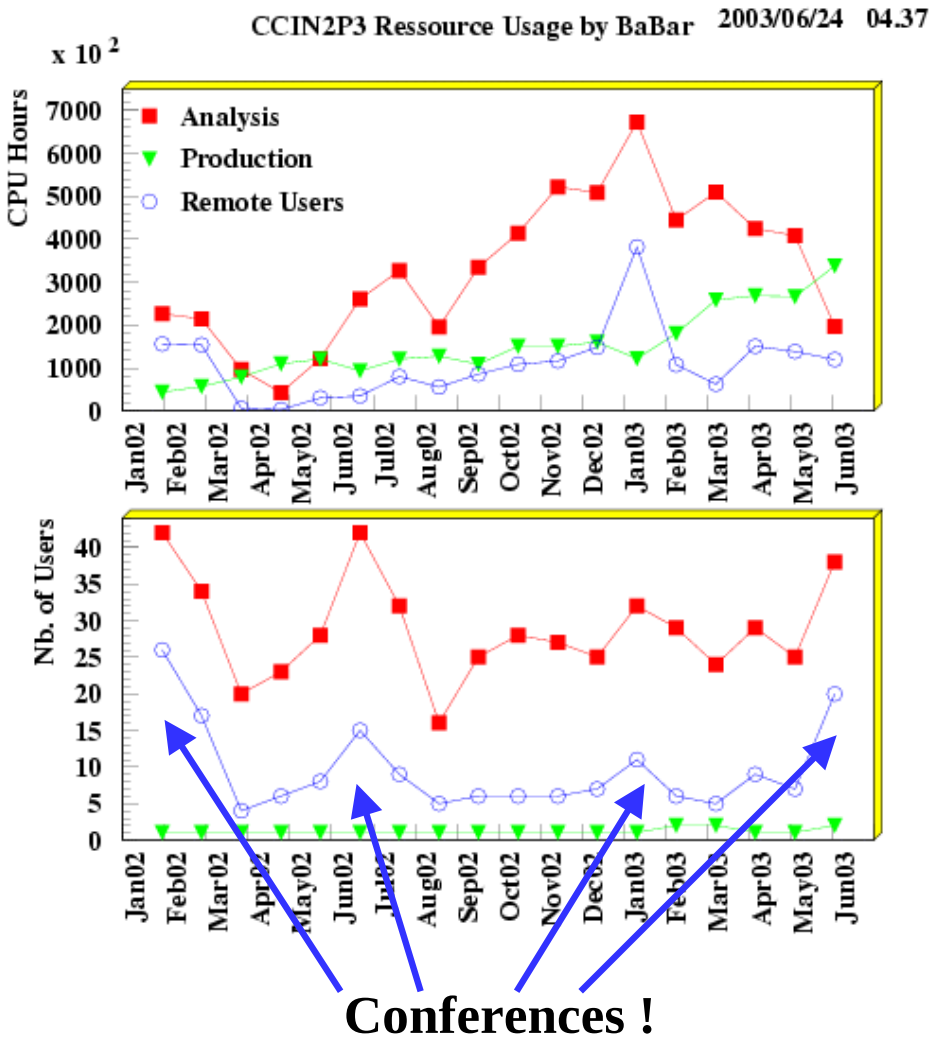
453 dual CPU machines (mostly Linux + a few SUN) ➔ ~47000 SpecInt95

Number of BaBar jobs over 16 months



- **Up to 6200 jobs in queue**
- **Up to 440 jobs running in // mostly on Linux**
- **An average of 169 jobs running over 16 months**

CPU (2)



1 hour on a 1GHz Linux box ≈ 8 normalized hours

Up to 0.7 M hours /month for Analysis

Up to 0.35 M hours / month for MC production

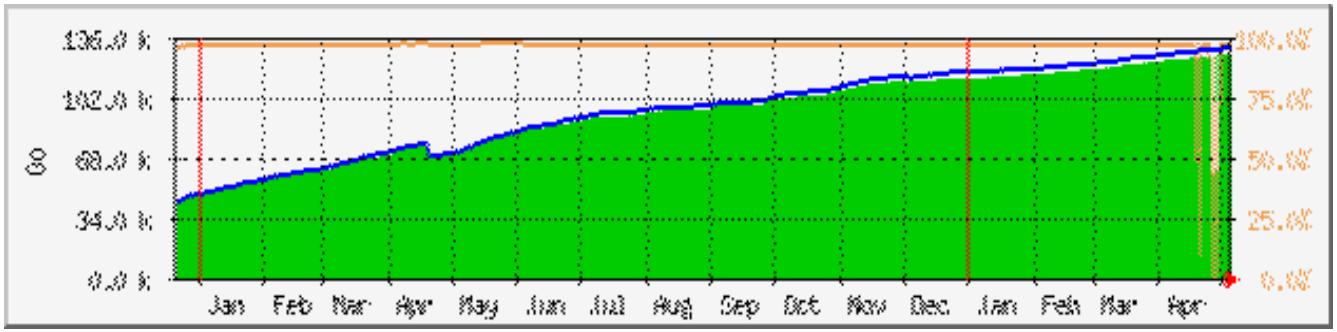
Relatively small number of very active users. Highly correlated to conference preparation

<u>Typical user's profile:</u> Run very large production ➔ skimming and /or N-tuple production for an Analysis Working Group

BaBar Data Set

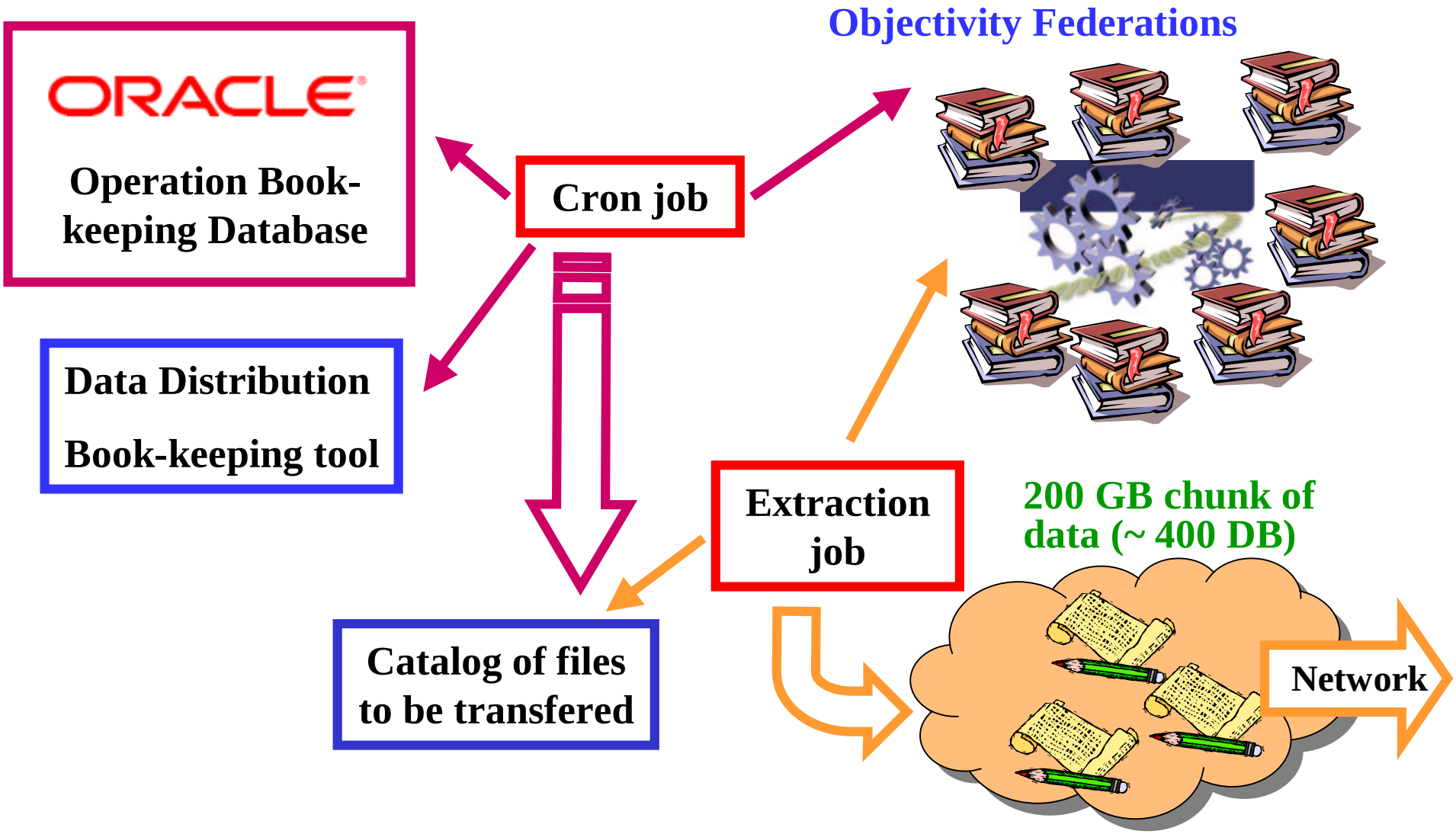
- CCIN2P3 is aiming at holding a complete copy of the data (Micro + Mini – Data + MC)
  - Run 1+2 data complete (2000-2002)
  - Large part of Run 3 data (2003) almost complete

- Fully automatic system to transfer BaBar data between SLAC and IN2P3
- 2 dedicated servers at SLAC and 1 at in2p3 for data transfer (SUN 4-processor on Gbit LAN)



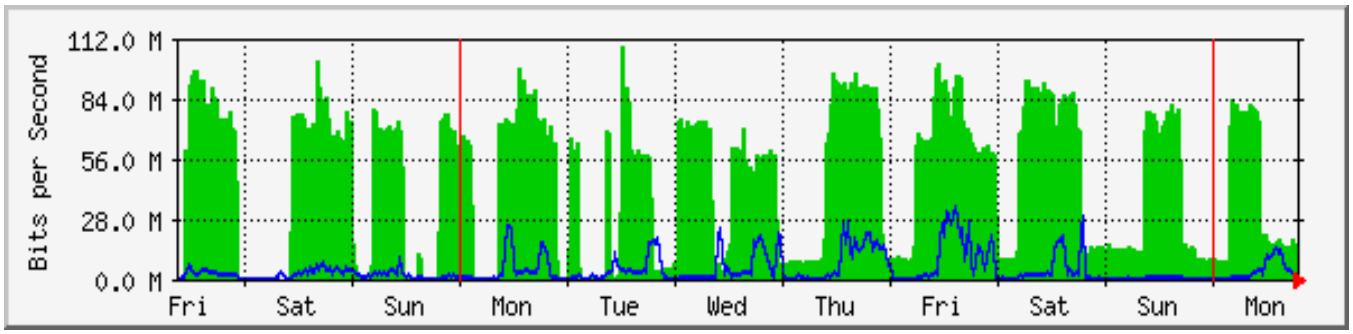
~135 TB of BaBar data in HPSS

Data Distribution



Network

- Using 2 network connections in parallel (2×622 Mbit/s)
  - IN2P3 – RENATER – ESNET or Internet 2
  - IN2P3 – CERN – ESNET
  - Allows load balancing between both links (data / MC for instance)



Lyon – CERN network usage

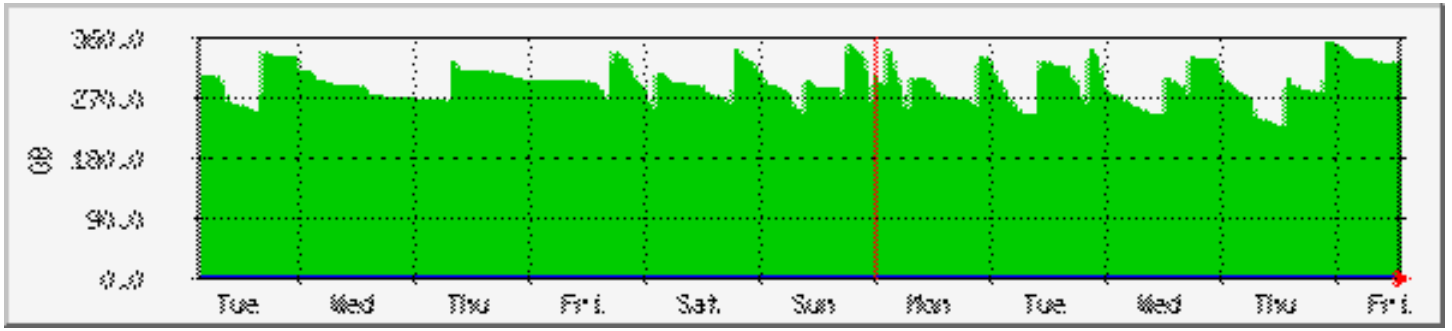
It is difficult to get more than 100 Mbit/s throughput with bbftp

But OK at the moment:

- **Limited by HPSS speed to migrate data to MSS**

HPSS Usage

- **BaBar is using HPSS in production @ Lyon ➔ Very difficult to tune**
- **Data are stored in dedicated servers (13 × Sun Solaris 4 CPU system + 2 TB disk cache)**
  - ***Will add 5 new 4-CPU SUN-480 with 4 TB disk space on each server***
  - **Running AMS (Advanced Multi-threaded Server)**
  - **If the data are not present on the server, they are automatically retrieved from HPSS**
    - **Transparent for the user**
    - **Automatic server side or client side decompression**
    - **Very positive experience ➔ Allows to get very good performances even with a relatively small cache disk space**

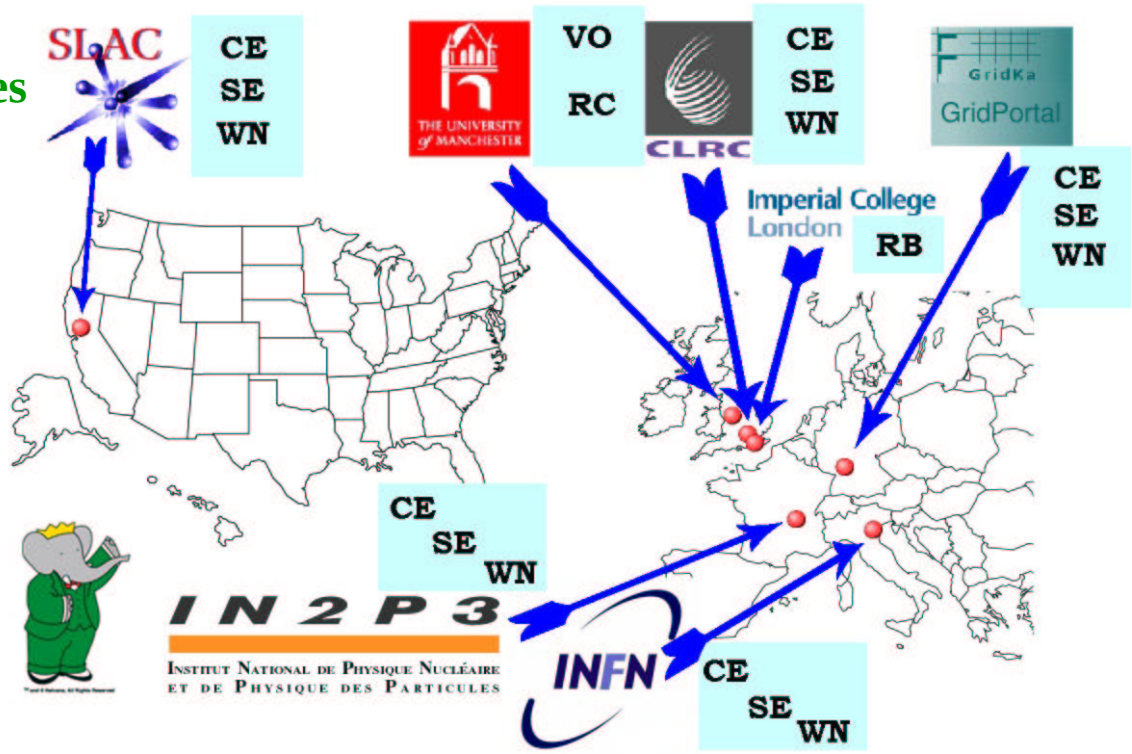


HPSS is also available for users to store large n-tuples

BaBar Grid

- The BaBar Grid is under development
- The main motivation is to develop a system where BaBar physicists don't need to take care:
  - Where are the data
  - Where are the CPU resources
- 2 main activities:
  - Porting analysis on the Grid
  - MC Production
- We already have working prototypes
- Expects to start grid MC production beginning of 2004.

- CCIN2P3 is already an active member of the EDG test-bed
- The whole batch system is accessible through the Grid



Keys for a Success

- Develop monitoring tools for all components
  - Anticipate problems
  - Scaling is always an issue
- Use reliable hardware
  - This is an increased source of concerns
  - A single network glitch (a DNS update for instance) can result in ~100 job crashes
  - Future HEP software will have to be fault tolerant
- Do not tolerate any backlog in the data transfer
  - Sometimes difficult to achieve !
- Create accounts within 24 hours
  - A new account request outside SLAC ~always correspond to a real and urgent need
    - ➔ Conferences
- Write and maintain easy startup documentation
- Maintain communication with the Tier 0 (SLAC)
- Maintain communication with the users
- Allow minimum time to answer to a user's problem

Conclusions

- CCIN2P3 is now a crucial piece in the BaBar distributed computing model
  - Analysis
    - ➔ Many top level analyses have been fully run at in2p3
  - MC Production
- Open to every BaBar user
- Very flexible
- A lot of competent people to help

- Have been pioneers in running HPSS in production for an HEP experiment

- 2 FTE are now running the Lyon Tier-A for BaBar

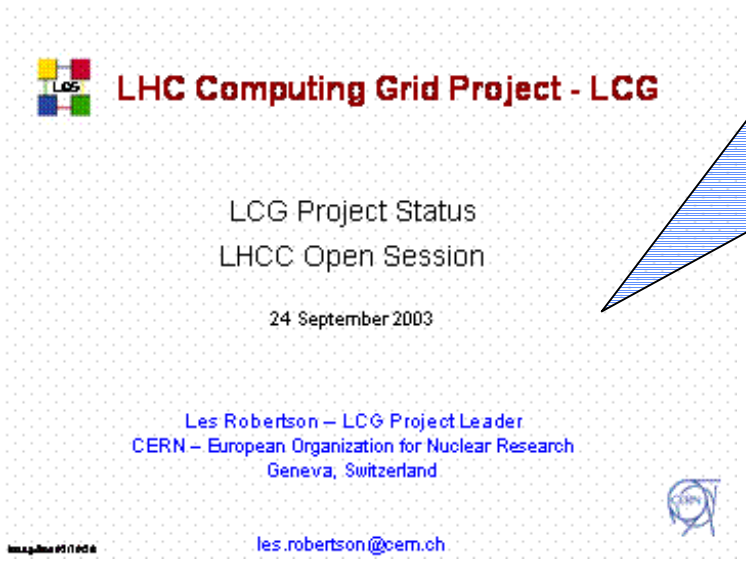
- GRID is coming into the game. It should be a tool to simplify the analysis on a distributed system, not to complicate it !

Update on LCG for HEP_CCC

Use info from LHCC open session 24 September 2003

Foils prepared by Les Robertson

Presented by Wolfgang von Rüden



last update: 02/10/2003 12:32

Applications Area Projects

- **Software Process and Infrastructure (SPI)** *(A.Aimar)*
  - Librarian, QA, testing, developer tools, documentation, training, …
- **Persistency Framework (POOL)** *(D.Duellmann)*
  - Relational persistent data store
- **Core Tools and Services (SEAL)** *(P.Mato)*
  - Foundation and utility libraries, basic framework services, object dictionary and whiteboard, math libraries
- **Physicist Interface (PI)** *(V.Innocente)*
  - Interfaces and tools by which physicists directly use the software. Interactive analysis, visualization
- **Simulation** *(T.Wenaus)*
  - Generic framework, Geant4, FLUKA integration, physics validation, generator services

---

- *Close relationship with -- **ROOT*** *(R.Brun)*
  - ROOT I/O event store; analysis package
- Group currently working on **distributed analysis** requirements – which will complete the scope of the applications area

POOL Status

- First production release of the POOL object persistency system made on time in June
  - Level 1 milestone of the LCG project
  - The base functionality requested by experiments for the data challenges in 2004
- First experiment integration milestones – met at end July
  - use of POOL in CMS pre-challenge simulation production
- Completion of first ATLAS integration milestone scheduled for this month

- POOL is now being deployed on the LCG-1 service
- Close collaboration organised between POOL team and experiment integrators

- Take-up by the experiments now beginning

SEAL and PI

- Core Libraries and Services (SEAL) –
  - libraries and tools, basic framework services, object dictionary, component infrastructure
  - implementing the new component model following the architecture blueprint
  - facilitates coherence of LCG software (POOL, PI) and integration with non-LCG software
  - uses/builds on existing software from experiments (e.g. Gaudi, Iguana elements) and C++, HEP communities (e.g. Boost)
  - first release with the essential functionality needed for it to be adopted by experiments made in July
  - working closely with experiment integrators to resolve bugs and issues exposed in integration

- Physicist Interfaces (PI)
  - Initial set of PI tools, services and policies in place
  - Incremental improvement based on feedback underway
  - Full ROOT implementation of AIDA histograms

Simulation Project

Leader Torre Wenaus

Andrea Dell'Acqua

John Apostolakis

Alfredo Ferrari

Fabiola Gianotti

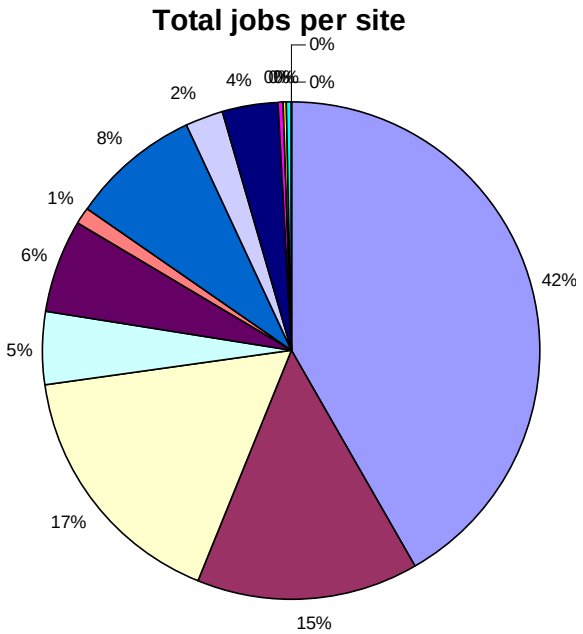
Paolo Bartalini

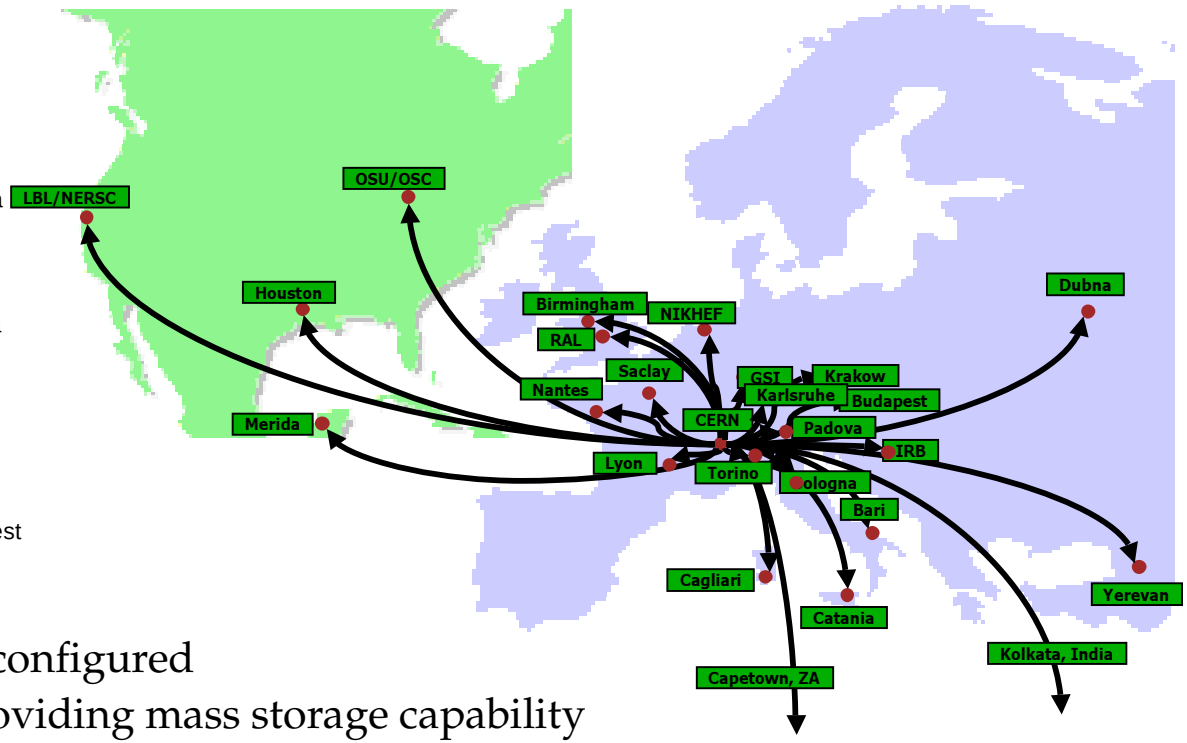
- Principal development activity: **generic simulation framework**
  - Expect to build on existing ALICE work; currently setting the priorities and approach among the experiments
  - Current status - **early prototyping beginning**
- Incorporates longstanding CERN/LHC **Geant4** work
  - aligned with and responding to needs from LHC experiments, **physics validation**, generic framework
- **FLUKA** team participating in
  - framework integration, physics validation
- Simulation **physics validation** subproject **very active**
  - Physics requirements; hadronic, em physics validation of G4, FLUKA; framework validation; monitoring non-LHC activity
- **Generator services** subproject also **very active**
  - Generator librarian; common event files; validation/test suite; development when needed (HEPMC, etc.)

Grid usage by experiments in 2003

last update: 02/10/2003 12:32

ALICE Physics Performance Report production – using AliEn





- 32 (was 28) sites configured
- 5 (was 4) sites providing mass storage capability
- 12 production rounds
- 22773 jobs validated, 2428 failed (10%)
- Up to 450 concurrent jobs
- 0.5 operators

last update 02/10/2003 12:32

Grid in ATLAS DC1
(July 2002 – April 2003)

US-ATLAS

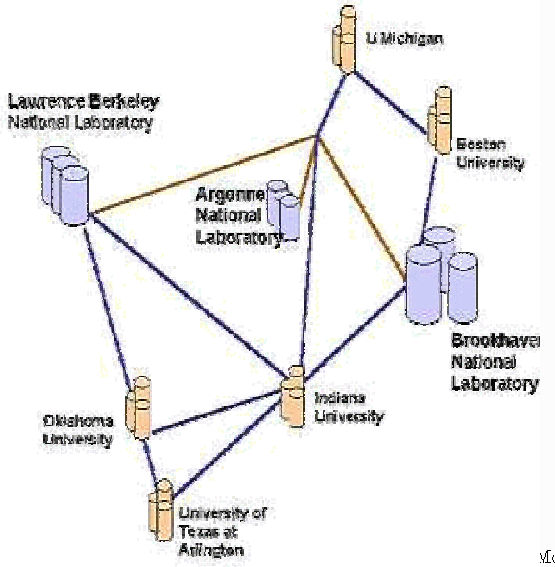
DC1:
Part of simulation;
Pile-up; reconstruction

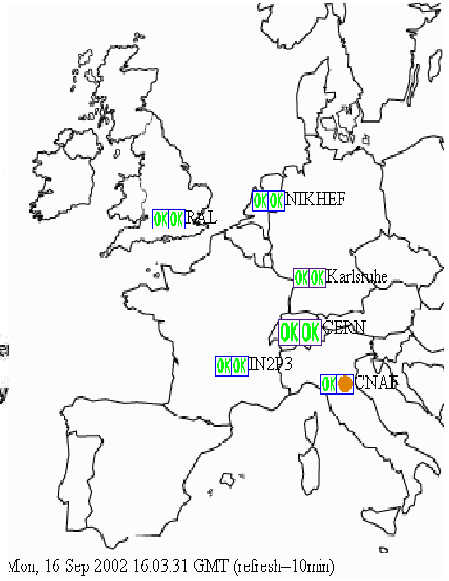
EDG

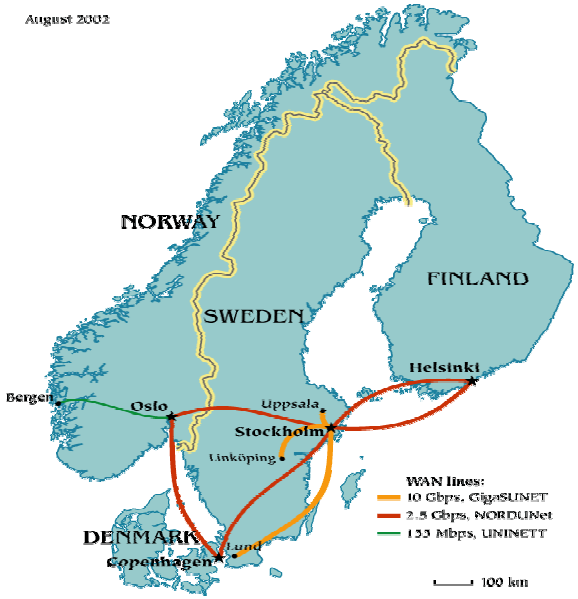
DC1:
several tests
(1st test in August02)

NorduGrid

DC1:
full production







CMS/LCG-N testbed

- CMS/LCG-0 is a CMS-wide testbed based on the LCG pilot distribution (LCG-0), owned by CMS
  - Red Hat 7.3 (7.3.2 with CERN kernel recommended)
  - Components from VDT 1.1.6 and EDG 1.4.X (LCG pilot)
  - Components from DataTAG (GLUE schemas and info providers)
  - Virtual Organization Management: VOMS
  - RLS in place of the replica catalogue (uses rlscms by CERN/IT)
  - Monitoring: GridICE by DataTAG
  - R-GMA (as BOSS transport layer for specific tests)
- Dynamic: install+test new components useful to CMS
- Support (outside CMS)
  - DataTAG WP4
  - LCG Experiment Integration and Support (EIS) team
- Currently configured as a CMS RC and producing data for DC04
- It allows us to do our software integration while awaiting LCG-1

The LHC Grid Service

last update: 02/10/2003 12:32

Goals for the Pilot Grid Service for LHC Experiments – 2003/2004

- Provide the principal **service** for Data Challenges in 2004
- Learn how Regional Centres can collaborate closely
- Develop experience, tools and process for **operating** and **maintaining** a global grid
  - Security
  - Resource planning and scheduling
  - Accounting and reporting
  - Operations, support and maintenance
- Adapt LCG so that it can be **integrated** into the sites' **mainline physics computing services**
  - Minimise level of intrusion
- For next 6 months the focus is on **reliability**
  - Robustness, fault-tolerance, predictability, and supportability take precedence; additional functionality gets prioritised

The LCG Service

Main Elements of a Grid Service

- **Middleware:**
  - Integration, testing and certification
  - Packaging, configuration, distribution and site validation
- **Operations:**
  - **Grid infrastructure** services
  - Local **Regional Centre operations**
  - **Operations Centre(s)** – trouble and performance monitoring, problem resolution, global coverage
- **Support:**
  - Integration of experiments' and Regional Centres' support structures
  - **Grid Call Centre**(s); documentation; training

Coordination and Management

Deployment Manager– Ian Bird (CERN)

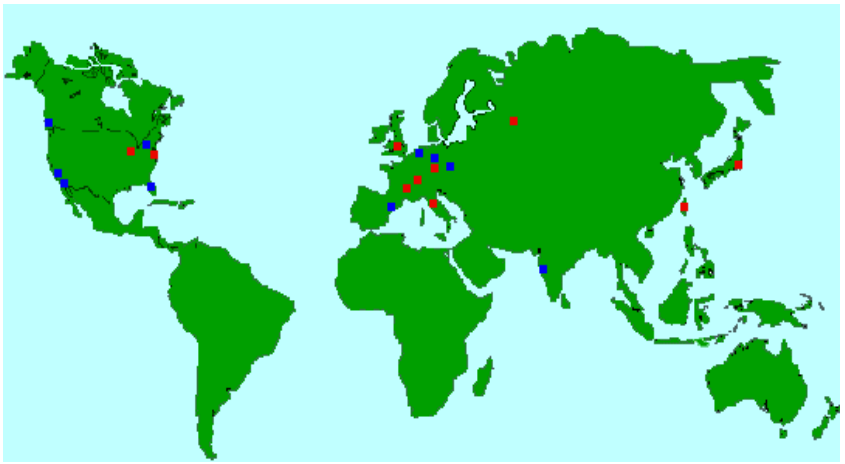
- **Grid Deployment Board**
  chair – Mirco Mazzucato (Padova)
  - National membership
  - Policies, resources, registration, usage reporting
- **Security Group**
  chair – David Kelsey (RAL)
  - Security experts
  - Close ties to site security officers
  - Security model, process, rules
- **Daily Operations**
  - Site operations contacts
  - Grid operations centre
  - Grid call centre

LCG Service Status

- Middleware package – components from –
  - European DataGrid (EDG)
  - US (Globus, Condor, PPDG, GriPhyN) → the Virtual Data Toolkit
- Agreement reached on principles for registration and security
- Certification and distribution process established and tested - June
- Rutherford Lab (UK) to provide the initial Grid Operations Centre
- FZK (Karlsruhe) to operate the Call Centre

- Pre-release middleware deployed to the initial 10 centres – July
- The "certified" release was made available to 13 centres on 1 September –

Academia Sinica Taiwan, BNL, CERN, CNAF, FNAL, FZK, IN2P3 Lyon, KFKI Budapest, Moscow State Univ., Prague, PIC Barcelona, RAL, Univ. Tokyo



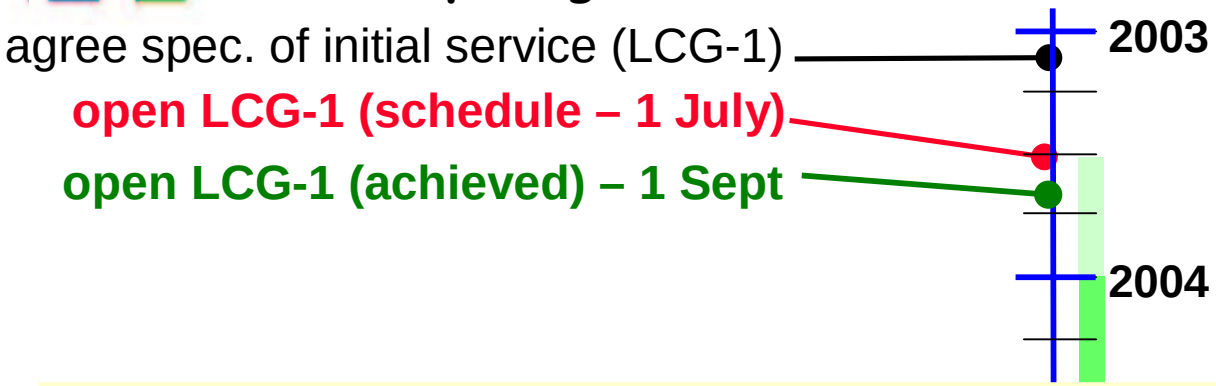
LCG Service – Next Steps

- Experiments now starting their tests on LCG-1
  - CMS target is to have 80% of their production on the grid before the end of the PCP of DC04
- Still a lot of work to be done - especially operations-related tasks
  - This will require active participation of regional centre staff
- Preparing now for adding new functionality in November to be ready for 2004
  - Implies deployment of a second multi-site testbed

- Web-site being set up at the Grid Operations Centre (Rutherford) with online monitoring information –

  see http://www.grid-support.ac.uk/GOC/

LCG Service Time-line

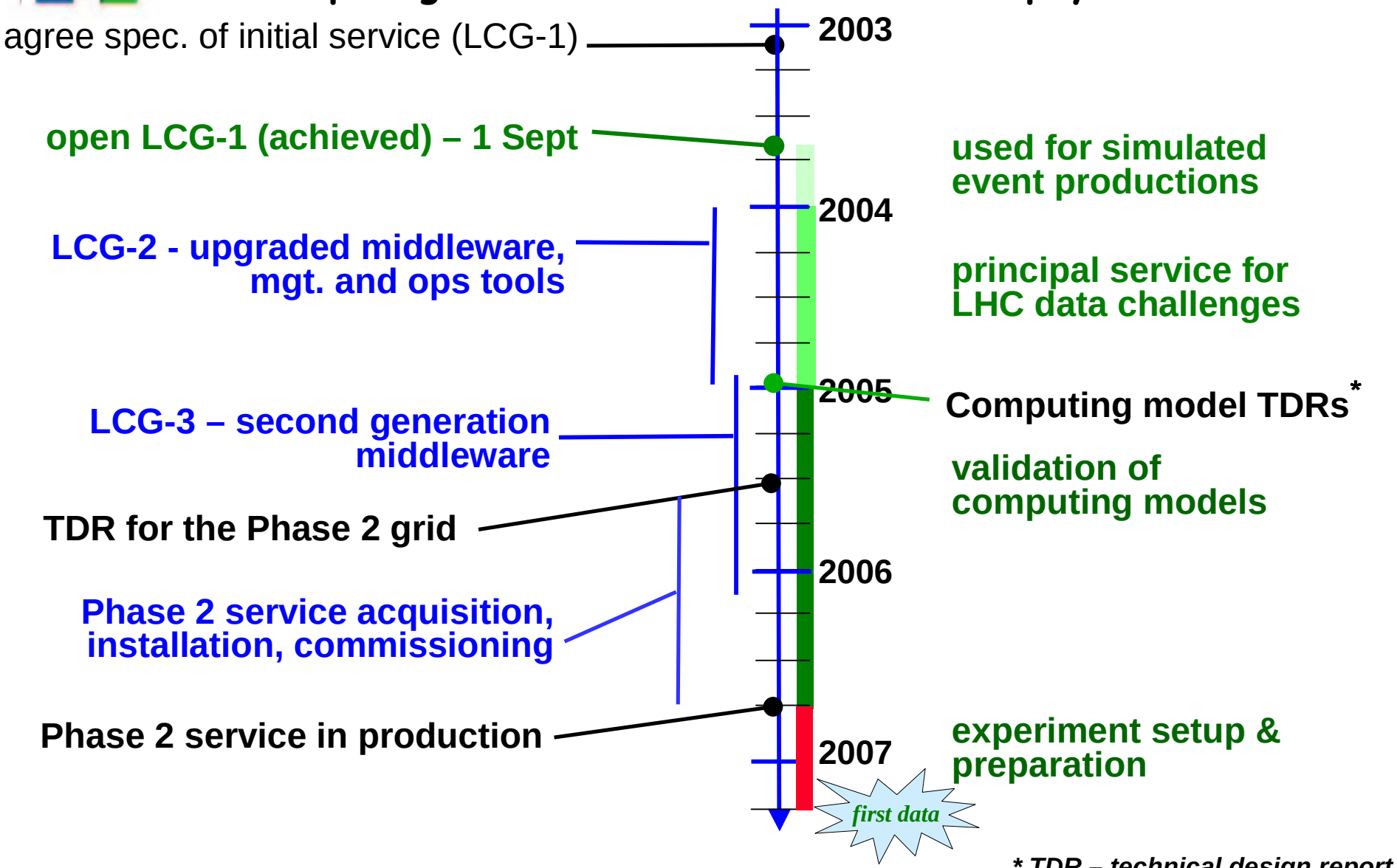


Level 1 Milestone – Opening of LCG-1 service

- 2 month delay, lower functionality than planned
- use by experiments will not start before October
- decision on final set of middleware for the 1H04 data challenges will be taken without experience of production running

→ reduced time for integrating and testing the service with experiments' systems before data challenges start next spring
→ additional functionality will have to be integrated later

LCG Service Time-line



last update 02/10/2003 12:32

** TDR – technical design report*

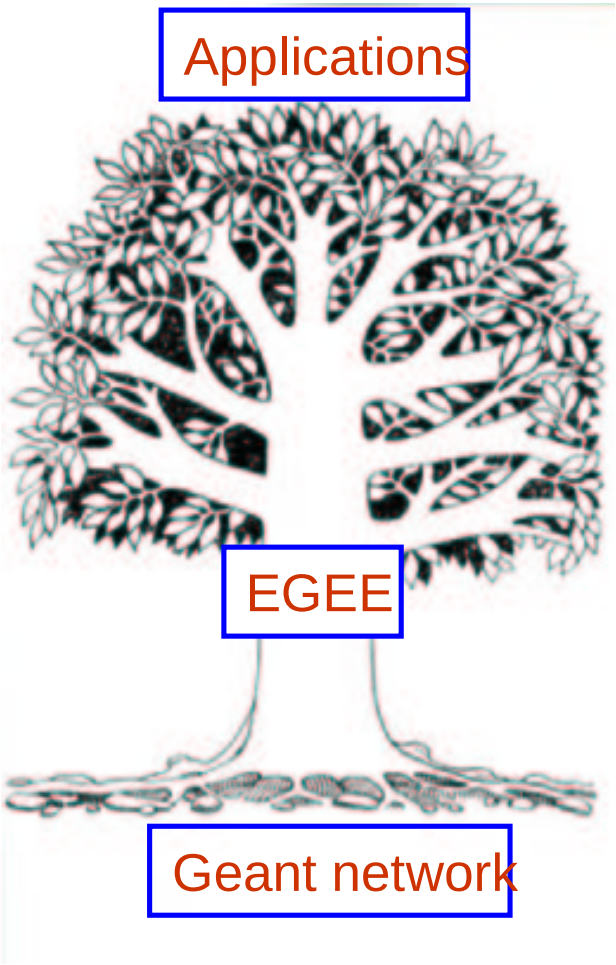
Middleware Evolution

Evolution of the Grid Middleware

- Middleware in LCG-1 ready now for use –
  - initial tests show reasonable reliability
  - scalability (performance) and stability still to be worked on
  - still low functionality.
- Early experience with the Web Services version of the Globus middleware (Globus Toolkit 3) and experience with the Open Grid Services Architecture (OGSA) and Infrastructure (OGSI) have been promising
- Good experience this year with packages linking experiment applications to grids – e.g. AliEn, Dirac, Octopus, ..
- Second round of basic Grid requirements nearing completion (HEPCAL II)
- Working group on common functionality required for distributed analysis (ARDA) nearing completion

EGEE vision:
Enabling Grids for E-science in Europe

- **Goal**
  - Create a wide European Grid production quality infrastructure on top of present and future EU RN infrastructure
- **Build on**
  - EU and EU member states major investments in Grid Technology
  - International connections (US and AP)
  - Several pioneering prototype results
  - Larg Grid development team (>60 people)
  - Requires major EU funding effort
- **Approach**
  - Leverage current and planned national and regional Grid programmes (e.g. LCG)
  - Work closely with relevant industrial Grid developers, NRENs and US-AP projects

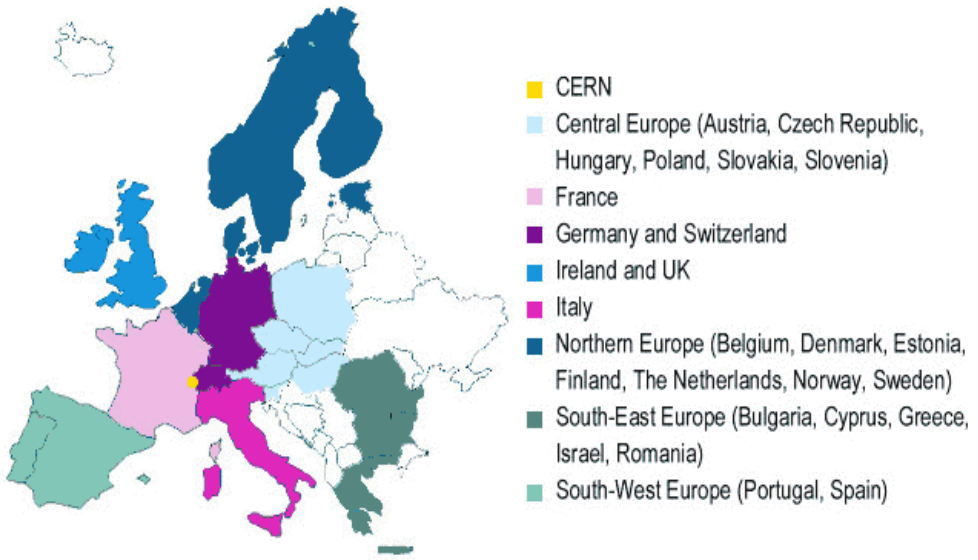


EGEE Proposal

- Proposal submitted to EU IST 6th framework call on 6th May 2003
  - Total budget request of approximately 39 M€ over 2 years
  - Executive summary (exec summary: 10 pages; full proposal: 276 pages)
    http://agenda.cern.ch/askArchive.php?base=agenda&categ=a03816&id=a03816s5%2Fdocuments%2FEGEE-executive-summary.pdf

- Activities
  - Deployment of Grid Infrastructure (based on LCG)
  - Re-Engineering of grid middleware (OGSA environment)
  - Dissemination, Training an

9 regional federations covering 70 partners in 26 countries



EGEE Status

- EGEE proposal passed thresholds at first EU review (June 2003)
- Hearing organised at Brussels for **1st July 2003** to answer a sequence of 10 questions prepared by the EU reviewers on details of the project
  - Responses to be given as a presentation (max. 15 slides)
  - Team of 6 EGEE representatives will attend

    Fabrizio Gagliardi (CERN) Project Director Guy Wormser (IN2P3/CNRS) Applications
    Wolfgang von Rueden (CERN) CERN/LCG Anders Ynnerman (ITN) Northern Europe
    Bob Jones (CERN) Technical Director Giorgio Maggi (INFN) Administration
- Next Steps (assuming hearing is successful)
  - Informed by EU of allocated budget envelope within 2 weeks
  - Negotiate budget details during summer and produce Technical Annex (details of negotiated tasks and budgets)
  - Foreseen project start date: **1st April 2004**

    Existing EU DataGrid and DataTAG projects will be extended until this date

    Main partners are asked to open posts during summer 2003

    Main partners are requested to assign resources during summer 2003 to start OGSA engineering investigations and architecture design work so that project can start on time

LCG and EGEE

- EU project approved to provide partial funding for operation of a general e-Science grid in Europe, including the supply of suitable middleware – **Enabling Grids for e-Science in Europe – EGEE**
  EGEE provides funding for 70 partners, large majority of which have strong HEP ties
- Similar funding being sought in the US

- LCG and EGEE work closely together, sharing the management and responsibility for -
  - **Middleware** – share out the work to implement the recommendations of HEPCAL II and ARDA
  - **Infrastructure operation** – LCG will be the core from which the EGEE grid develops – ensures compatibility; provides useful funding at many Tier 1, Tier2 and Tier 3 centres
  - **Deployment of HEP applications** - small amount of funding provided for testing and integration with LHC experiments

Middleware - Next 15 months

- Work closely with experiments on developing experience with early distributed analysis models using the grid
  - Multi-tier model
  - Data management, localisation, migration
  - Resource matching & scheduling
  - Performance, scalability
- **Evolutionary** introduction of **new software** – rapid testing and integration into mainline services –
  – while maintaining a **stable service** for **data challenges**!

- Establish a **realistic** assessment of the grid functionality that we will be able to **depend on at LHC startup** –
  a fundamental input for the Computing Model TDRs due at end 2004

Grids - Maturity is some way off

- Research still needs to be done in all key areas of importance to LHC
  - e.g. data management, resource matching/provisioning, security, etc.
- Our life would be easier if standards were agreed and solid implementations were available – but they are not
- We are just entering now in the second phase of development
  - Everyone agrees on the overall direction, based on Web services
  - But these are not simple developments
  - And we still are learning how to best approach many of the problems of a grid
  - There will be multiple and competing implementations – some for sound technical reasons
- We must try to follow these developments and influence the standardisation activities of the Global Grid Forum (GGF)
- It has become clear that LCG will have to live in a world of multiple grids – but there is no agreement on how grids should inter-operate
  - Common protocols?
  - Federations of grids inter-connected by gateways?
  - Regional Centres connecting to multiple grids?

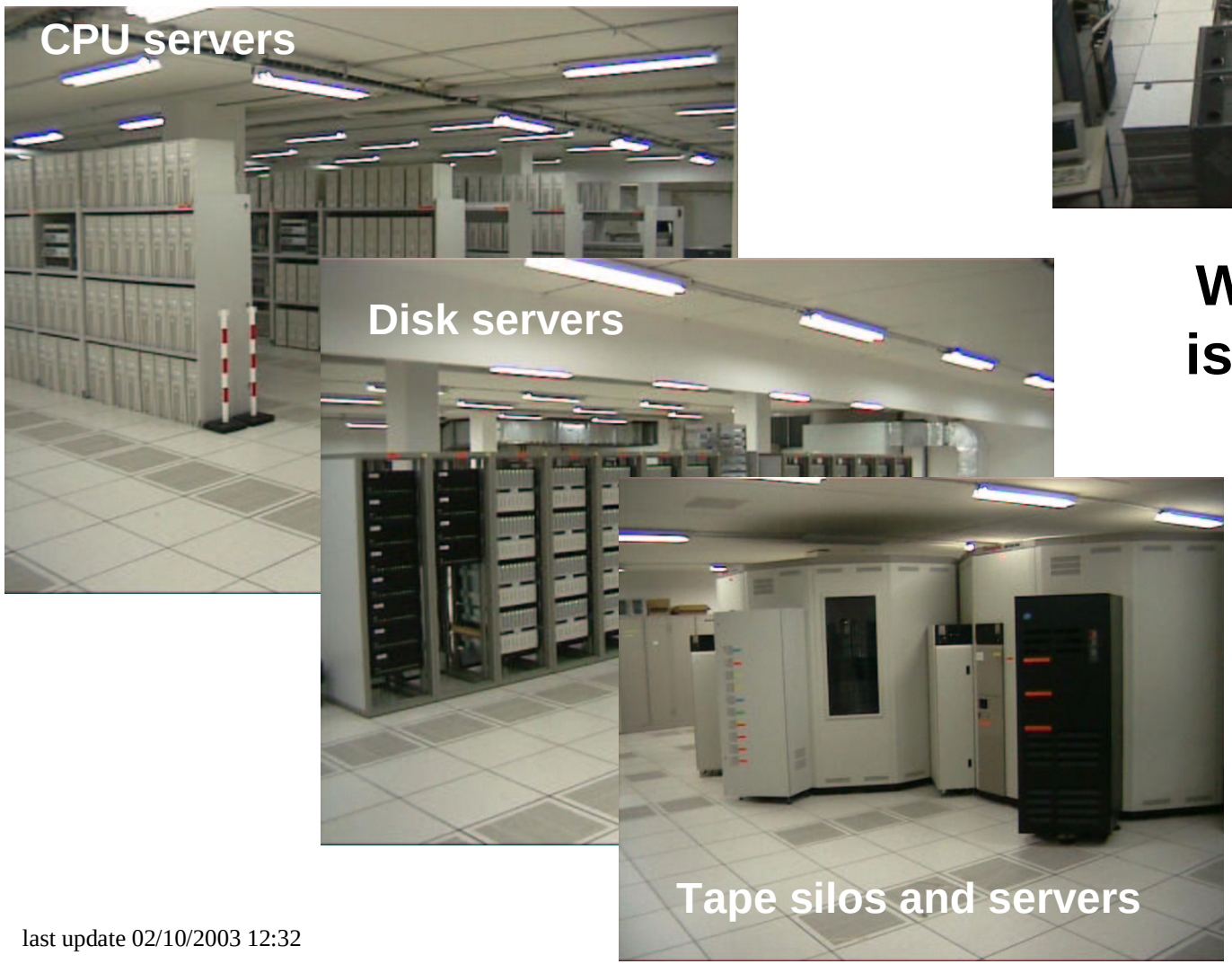
Running a service in this environment will be challenge!

CERN Fabric

LCG Fabric Area

- Fabric = Computing Centre based on big PC cluster
- Operation of the CERN Regional Centre
  - GigaByte/sec data recording demonstration in April
  - 350 MB/sec DAQ-Mass Storage milestone for ALICE
- Preparation of the CERN computing infrastructure for LHC
  - See next foil
- Technology tracking
  - 3rd round of technology tracking completed this year –
    see http://www.cern.ch/lcg → *technology tracking*
- Communication between operations staff at regional centres – uses the HEPIX organisation – 2 meetings per year

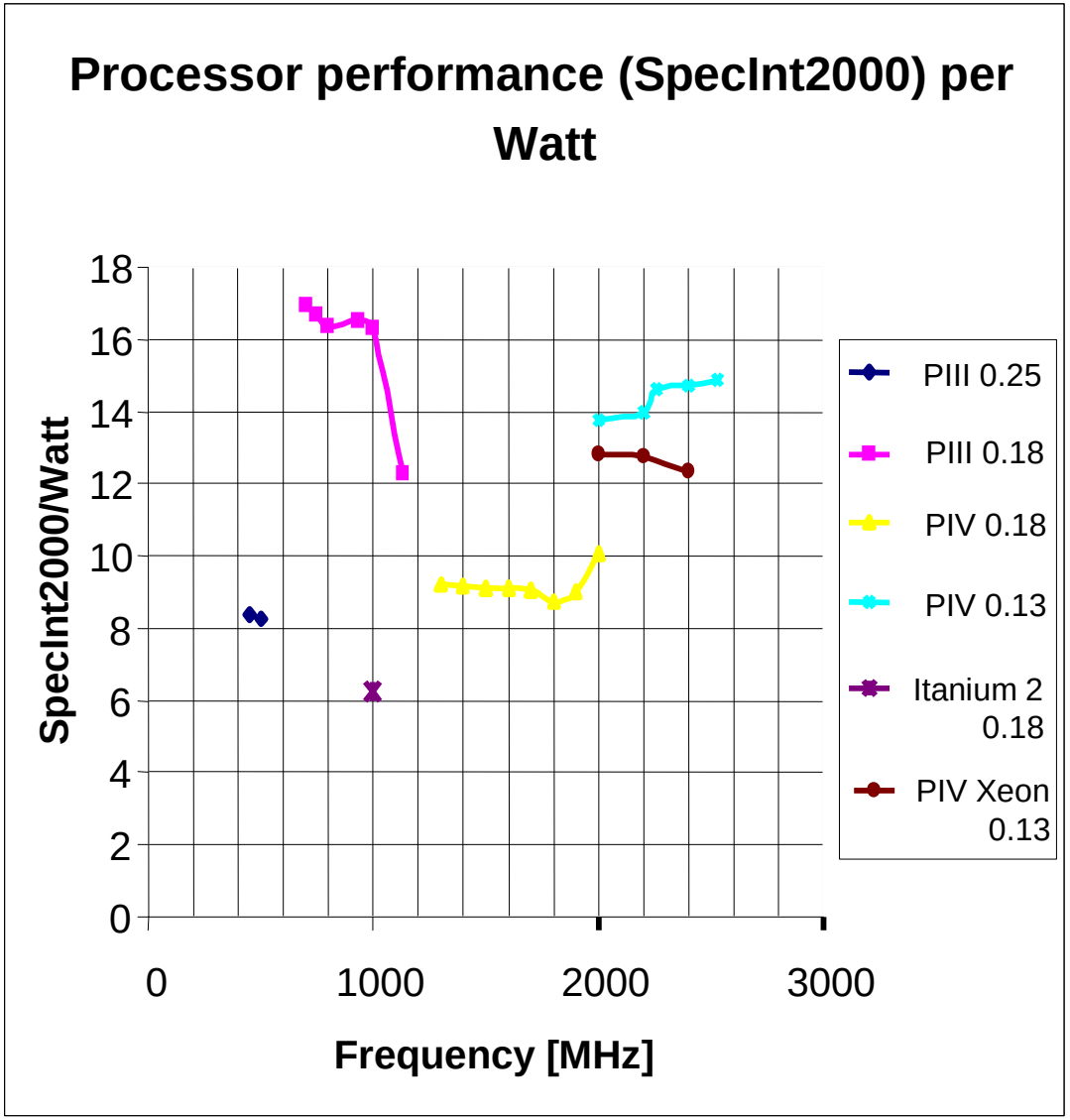
The new computer room in the vault of building 513 is now being populated



While the old room is being cleared for renovation

Processor Energy Consumption

- Energy consumption is increasing ~linearly with achieved processor performance
- Power managed chips are a solution for the home/office market - but will probably not help significantly with round the clock, high cpu-utilisation applications
- Intel TeraHertz and TriGate R&D projects aim at significant reductions in power consumption – but we may not see products before 2007-08
- Electric power and cooling are major cost and logistic problems for computer centres – CERN is planning 2.5 MW for LHC (up from ~800 KW



Resources in Regional Centres

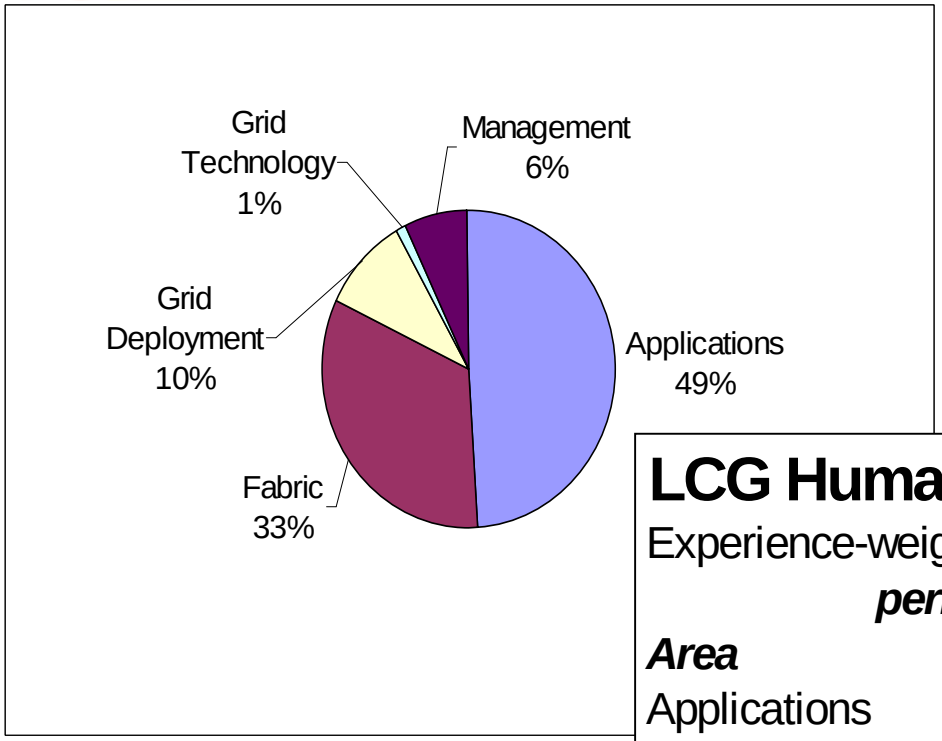
- Resources planned for the period of the **data challenges in 2004**
- CERN **~12% of the total** capacity
- Numbers have to be refined – different standards used by different countries
- **Efficiency of use** is still a major question mark – *reliability, efficient scheduling, sharing between Virtual Organisations (user groups)*
- These resources will in future be integrated into the LCG quarterly reports

	CPU	Disk TB	Support FTE	Tape TB
CERN	700	160	10.0	1000
Czech Repub	60	5	2.5	5
France	420	81	10.2	540
Germany	207	40	9.0	62
Holland	124	3	4.0	12
Italy	507	60	16.0	100
Japan	220	45	5.0	100
Poland	86	9	5.0	28
Russia	120	30	10.0	40
Taiwan	220	30	4.0	120
Spain	150	30	4.0	100
Sweden	179	40	2.0	40
Switzerland	26	5	2.0	40
UK	1656	226	17.3	295
USA	801	176	15.5	1741
Total	5600	1169	120.0	4223

Resources committed for 1Q04

Human Resources Consumed

without Regional Centres



LCG Human Resource Usage to 2Q03

Experience-weighted FTE-years - all funding sources

period	*2002*	*1Q03*	*2Q03*	*Total*
Area				
Applications	39.2	13.6	14.4	67.2
Fabric	25	10.3	10	45.3
Grid Deployment	5.3	3.8	4.9	14.0
Grid Technology	0.5	0.2	0.6	1.3
LCG Management	6.2	1.8	1.6	9.6
Total LCG	**76.2**	**29.7**	**31.5**	**137.4**

last update 02/10/2003 12:32

Summary

- POOL object persistency project is now entering real use by experiments
- Simulation project provides an LHC framework for agreeing requirements and priorities for GEANT 4 and FLUKA
- 2003 has seen increased use of grids in Europe and the US for simulation
- The first LCG service is now available for use – 2 months later and with lower functionality than planned, but we are optimistic that this can provide a stable global service for the 2004 data challenges
- The requirements for grid functionality for distributed analysis are expected to be agreed next month – in time to take advantage of the EGEE EU funding for re-engineered grid middleware for science
- The intense activity world wide on grid development promises longer term solutions and short term challenges
- The major focus for all parts of the project in the next year is demonstrating that distributed analysis can be done efficiently using the grid model

Spam Fighting at CERN

Emmanuel Ormancey

Topics

- Statistics…
- Email stealing
- Products overview
- Solution for CERN
- How it works
- User configuration
- Efficiency
- Cost

Statistics…

- At CERN:
  - 100 000 Incoming mails per day.
  - Spam filter detects from 25% to 35% as spam.
- Measurements in Europe for 2001: (NetValue users panel) :
  - Spam increased by 80% in 2001.
  - 36.8% of received mails are Spam.
- According to US AntiSpam company Brightmail:
  - Spam increased by 450% during last year.
  - 74% of received mails are Spam.
- Estimated cost for companies:
  - 8.9 billion $ for US companies, 2.5 billion $ for European companies (half productivity loss, half technical cost).
  - 500 Million investment in spam fighting.
  - 1.2 billion $ spent in User Support handling user complaints.
  - 1 spam = 1$ cost per company.
- Cost for spammers:
  - 39$ for 1 million French email addresses.

Email stealing

- Test at CERN: an “invisible” email address was published on the Mail Service Website, 37 days after the first Spam was received.
- 6 Weeks study: 275 email addresses published on 175 different supports.
  (source Federal Trade Commission, November 2002)
- In 6 weeks: 3349 Spams were received by the 275 addresses.
- Speed record: First Spam was received 9 minutes after publishing an email in a Chat room.

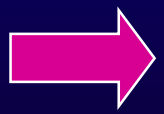
Support	Spammed emails
Chat room	100%
Newsgroup	86%
Standard Web site	86%
Personal Web Site	50%
Forum	27%
WebMail	9%

Existing products

- **Existing products are too simple**
  - **Basic tests usually based on word matching in the mail.**
  - **Action taken requires huge amount of work:**
    - Delete: helpdesk will receive user complaints if false positive.
    - Quarantine (i.e. Norton antivirus): require manual lookup to validate real spam and good mails.
- **Spammers can easily get the product and improve their technique to bypass them.**

Solution for CERN

- SpamAssassin product looked promising
- A new product was developed based on it
  - Improved version: CERN SpamKiller.
  - Adds existing rules and custom tests.
  - Easy to modify and to create new checks.
  - Multi-platform: Exchange 2000 and Sendmail.



CERN SpamKiller

- Is now running at CERN for nearly 1 year.

Technical Overview

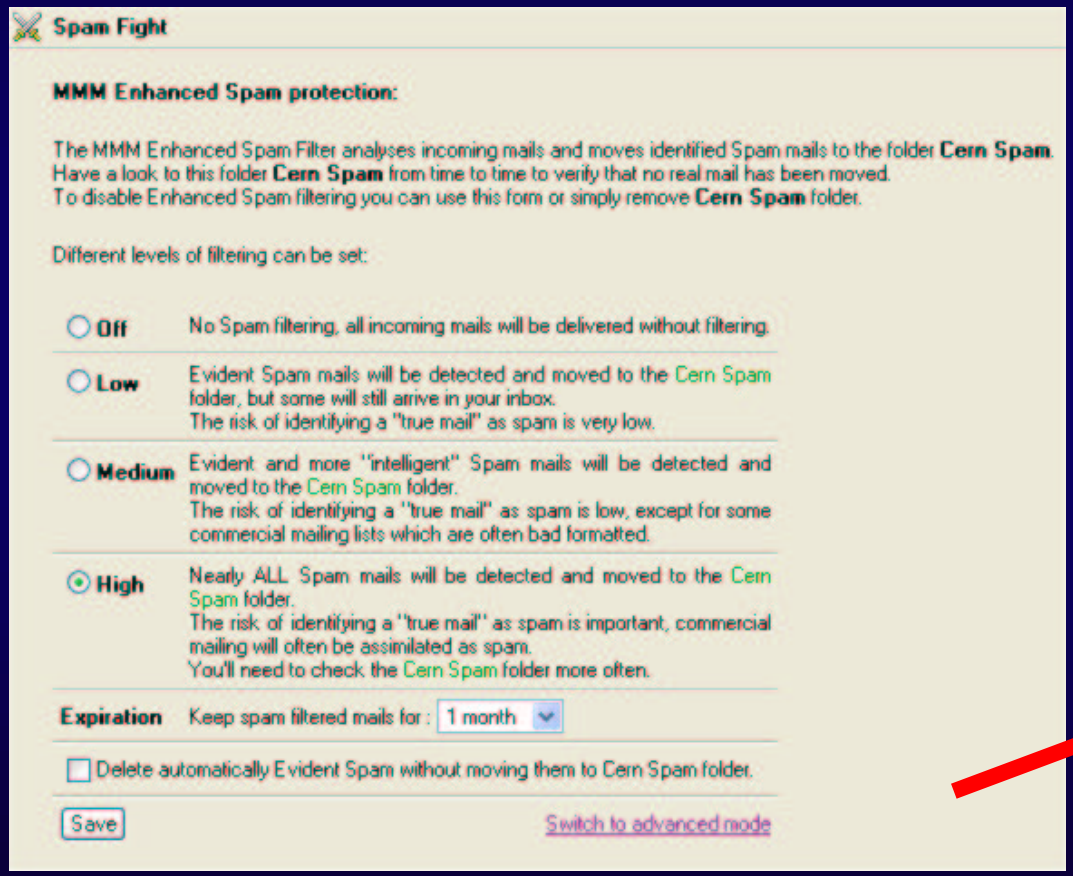
- Client / Server solution:
  - Server is written in C#, running on Windows.
  - Portable code.
  - Clients exists on Windows, Unix and can be created on any platform.
  - Uses simple SpamAssassin protocol.
- Configuration and Logs in XML files, easy to reuse.

How it works

- **SpamKiller calculates a score for a mail, based on different tests:**
  - **Text in header, body and attachments.**
  - **“Smart tests” more complex (word associations).**
  - **Open relays blacklist check.**
  - **Catalog check: compares mail with spam catalog signatures.**
  - **Bayesian Statistics calculation: probability for a mail of being spam.**
- **Each test returns a score, sum of all scores gives final note:**

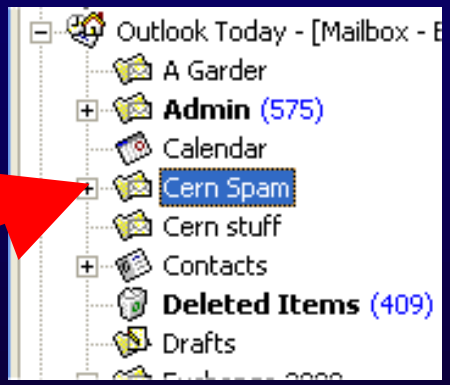
Content analysis details: (5.559 hits, 5 required)
2 points: HTML-only mail, with no text version
0.814 points: Subject has an exclamation mark
0.5 points: Spam phrases score is 00 to 01 (low)
2.035 points: 'remove' URL contains an email address

User configuration



- Configure Spam Level.
- Set expiration time.

CERN Spam folder automatically created.



Filtering

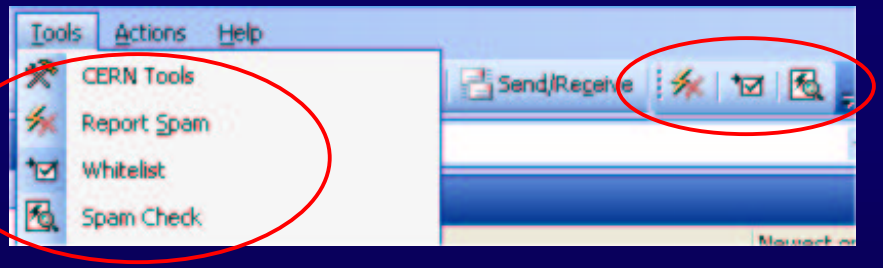
- When spam is detected:
  - Allow user to choose a spam detection level.
  - Depending on the score, mail is moved to CERN Spam folder if Spam.
- User have to check from time to time the CERN Spam folder for incorrectly classified mails.

User configuration

- Many users = many different views on Spam fighting.
- Solution is to propose customizable Spam Fighting tools:
  - Configuration level: User chooses the spam threshold
  - Option to automatically delete evident Spam
  - White list feature: patterns to match in From, Subject, To
  - Allow maximum level: everything is Spam except people I know: white lists, contacts, CERN people
- Still a lot to do:
  - Reject mail if user don’t understand the language (Japanese, Chinese or Russian written mail).
  - Propose a Bayesian statistics dictionary at user level.

Reporting Spam

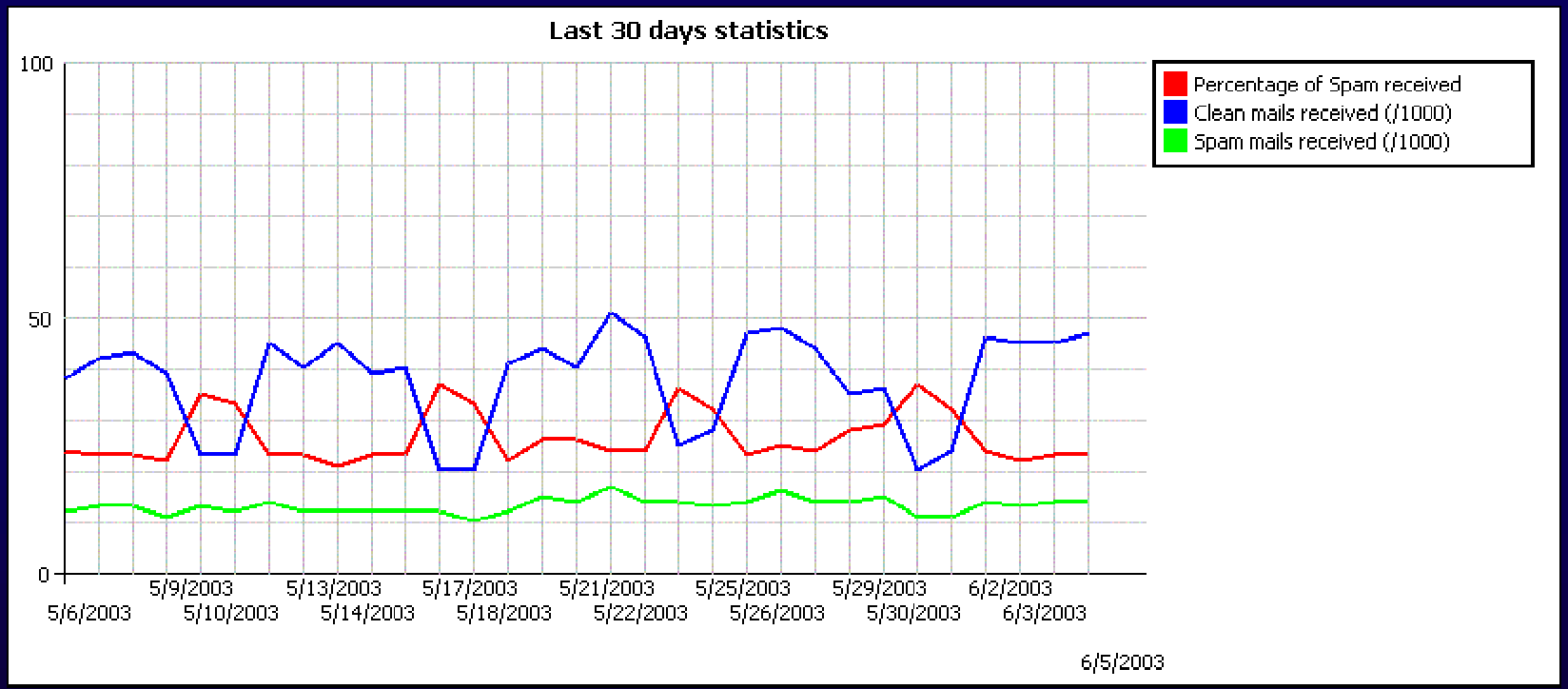
- Outlook XP special buttons:
  - delete Spams and send them to CERN Antispam team.
  - Build Whitelist from received mails.
  - Test a mail in SpamKiller and view detailed results.



- Spams reported:
  - Goes to a Public Folder in Exchange.
  - Used to improve detection rules.

Statistics

Online statistics available on SpamKiller website:



Efficiency

- False positives are very low, except for commercial lists.
  - White lists at user level can be configured to prevent this.
- Very good spam detection
  - Statistics are hard to build.
  - “My” mailbox filtering is optimized:
    - 20 to 40 spams filtered per day.
    - 1 or 2 spams still getting in Inbox per month.
- Signature catalog checking is not useful at all, random code is now omnipresent in spam mails. Even the nilsimsa algorithm cannot work fine anymore.
- 85% of Spams could have been detected only with Bayesian statistical filter.

Problems

- Viruses can be detected by SpamKiller, but are not deleted, just moved to CERN Spam folder.
- Out Of Office Assistant of Outlook replies to Spam senders even if filtered.
- CERN Spam folder checking might show to users “horrible” mails, so don’t use automatic preview for this folder (Outlook)

Costs

- Previous basic filters:
  - 1-2 hours/day to analyze logs, abuse reports and improve black lists.
  - Helpdesk cost to handle false rejection.
- SpamKiller:
  - Initial development and test phase: 1 FTE for 3 months.
  - Less than 1 hour / week for tuning and improvements.

Next…

- Long term goal: use a commercial product when available
- Current situation:
  - Young techniques, constantly evolving
  - No 100% reliable solution
  - Proprietary interfaces requiring dedicated manpower
- Still checking commercial and free products.
- Meanwhile: SpamKiller evolution
  - Includes all existing Spam detection techniques
  - Think, test and add new techniques
  - Propose a fully customizable solution at user level
  - Automatic whitelist generation by sending a generated mail to “first time” mail senders, asking them to do a specific action (See www.mail-block.com)

Conclusion

- Keep some basic low level checks on SMTP gateways for flood prevention, virus rejection.
- SpamKiller available to MMM users for enhanced Spam detection since more than 6 months:
  - General good feedback
  - Bad comments always due to configuration problems, or user customization feature missing.
- Also used to quarantine at server level Spam sent to Helpdesk.
- Future is to join forces against Spam:
  - Share rules, regular expressions patterns and Bayesian statistics dictionary with other organizations.
  - Central Antispam configuration with Live Update like antivirus definitions could be the solution.

CHEP04

- CERN is organizing next CHEP
  - "Somewhere in Switzerland not @ CERN"
  - 3rd/4th week of September 2004
  - Program chair: J. Harvey
  - Local Organizer: A. Silverman
- General frustration about the way the IAC has worked in recent years
  - Membership
  - Site choice
  - Weight of Europe
  - Last CHEP organization was very late!!
- ☞ HEPCCC asks W. v. Rüden to disolve current IAC and reform.

CERN School of Computing

- Countries where CSC has taken place in the past ☞
- Logistic model ☞
- Role of local organizers ☞
- Countries should volunteer to host the school (particularly those where it has not been before)

Visited Countries

1970	Varenna	**Italy**	1991	Ystad	**Sweden**
1972	Pertisau	**Austria**	1992	L'Aquila	**Italy**
1974	Godöysund	**Norway**	1993	L'Aquila	**Italy**
1976	La Gr. Motte	**France**	1994	Sopron	**Hungary**
1978	Jadwisin	**Poland**	1995	Arles	**France**
1980	Vraona	**Greece**	1996	E.aan Zee	**The Nether**.
1982	Zinal	**Switzerland**	1997	Pruhonice	**Czech Rep**.
1984	Aiguablava	**Spain**	1998	Funchal	**Portugal**
1986	Renesse	**The Nether**.	1999	St. Jablonki	**Poland**
1987	Troia	**Portugal**	2000	Marathon	**Greece**
1988	Oxford	**Great Britain**	2001	Santander	**Spain**
1989	Bad Herre.	**Germany**	2002	V. Equense	**Italy**
1990	Ysemonde	**Belgium**	2003	Krems	**Austria**

School Logistics Models

Model	Accommo-dation	Classes	Computer / Network Equipment	Example
Integrated	Hotel	Same Hotel	Provided / rented by Local Organising Cmtt	• Arles • Marathon • Vico Enquense
Separate	Hotel	Other location Usually University	Provided by host University	• Santander • Krems

Local organizer role (summary)

Site issues	■ Propose options for the school venue ■ Liaise with hotel / university
Technical infrastructure	■ Arrange for computers for exercises ■ Arrange for external Internet access
On-site support	■ Provide bilingual secretarial help during the School
Financing	■ Help find local sponsors (institutions, industry) (for social events, equipments, …)
Programme	■ Contribute ideas, content to the program (e.g. after dinner speakers) ■ Propose social events, excursions