Worldwide LHC Computing Grid Project

Computing Systems for the LHC Era

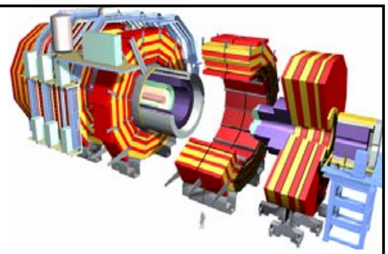
CERN School of Computing 2007

Dubrovnik
August 2007

Les Robertson
WLCG Project Leader

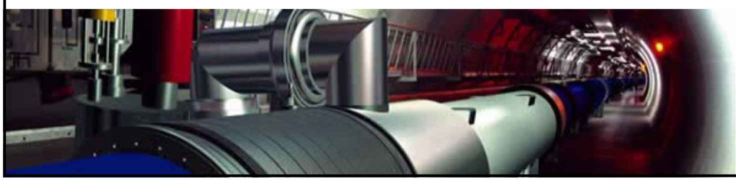
Outline

- LHC computing "problem"
- Retrospective – from 1958 to 2007
- Keeping ahead of the requirements for the early years of LHC → a Computational Grid
- The grid today – what works and what doesn't
- Challenges to continue expanding computer resources
- -- and Challenges to exploit them

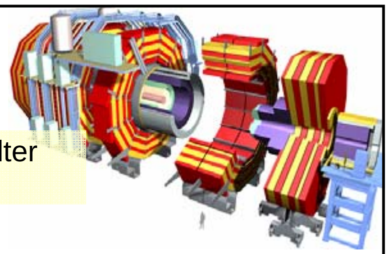


The LHC Accelerator

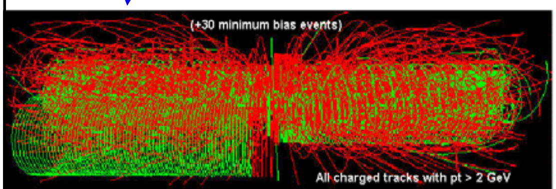
The accelerator generates 40 million particle collisions (events) every second at the centre of each of the four experiments' detectors



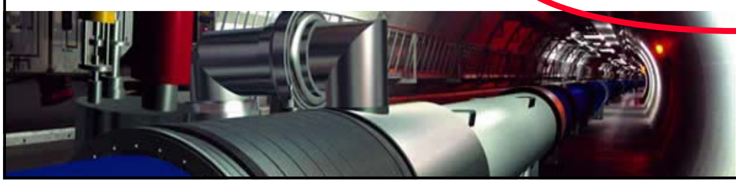
LHC DATA



This is reduced by online computers that filter out a few hundred "good" events per sec.



Which are recorded on disk and magnetic tape at 100-1,000 MegaBytes/sec → ~15 PetaBytes per year for all four experiments



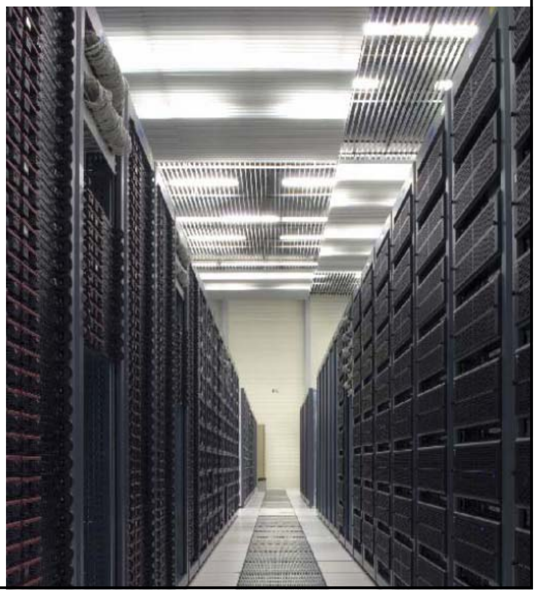
LHC DATA ANALYSIS

Experimental HEP codes
key characteristics –

- modest memory requirements
- perform well on PCs
- independent events
  → easy parallelism
- large data collections (TB → PB)
- shared by very large user collaborations

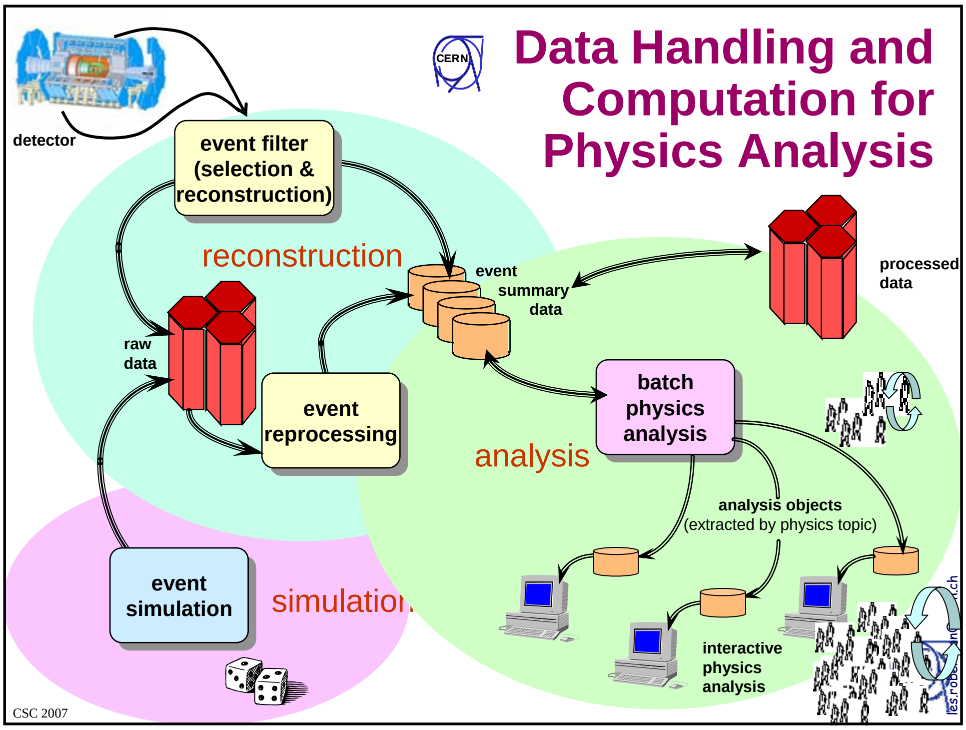
For all four experiments

- ~15 PetaBytes per year
- ~200K processor cores
- > 5,000 scientists & engineers



CSC 2007

Data Handling and Computation for Physics Analysis



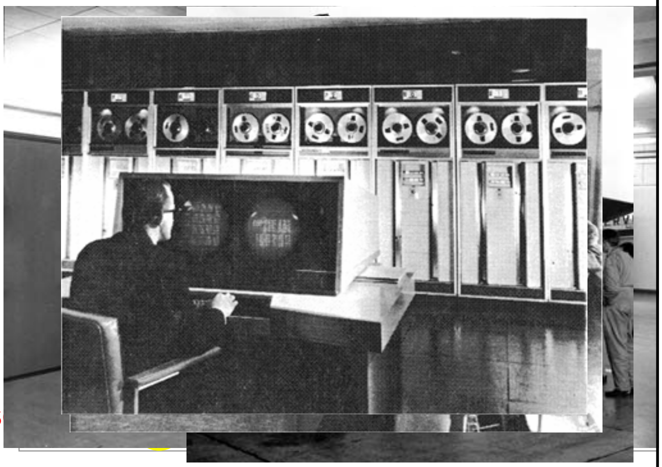
CSC 2007

Evolution of CPU Capacity at CERN

The early days
The fastest growth rate!

Technology-driven

- **Ferranti Mercury**
  1958 5 KIPS
- **IBM 709**
  1961 25 KIPS
- **IBM 7090**
  1963 100 KIPS
- **CDC 6600** - *the first supercomputer*
  1965 3 MIPS



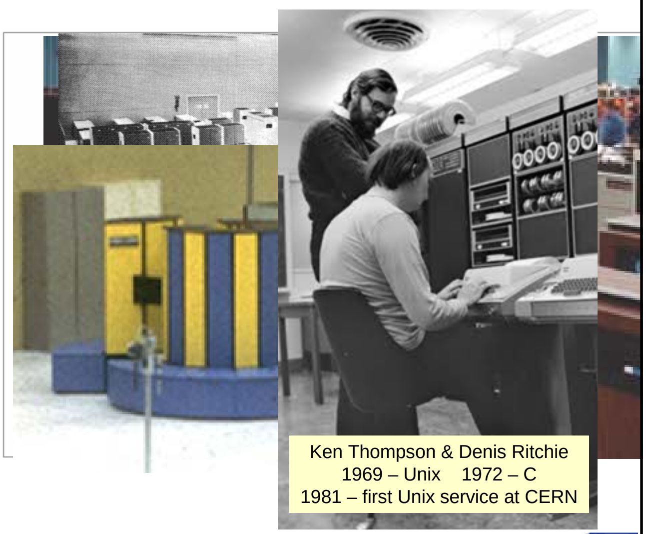
3 orders of magnitude in 7 years

The Mainframe Era

budget constrained

proprietary architectures maintain suppliers' profit margins → slow growth

- **CDC 7600**
  1972 13 MIPS
  for 9 years the fastest machine at CERN, finally replaced after 12 years!
- **IBM 168**
  1976 4 MIPS
- **IBM 3081**
  1981 15 MIPS
- **CRAY X-MP** - *the last supercomputer*
  1988 128 MIPS



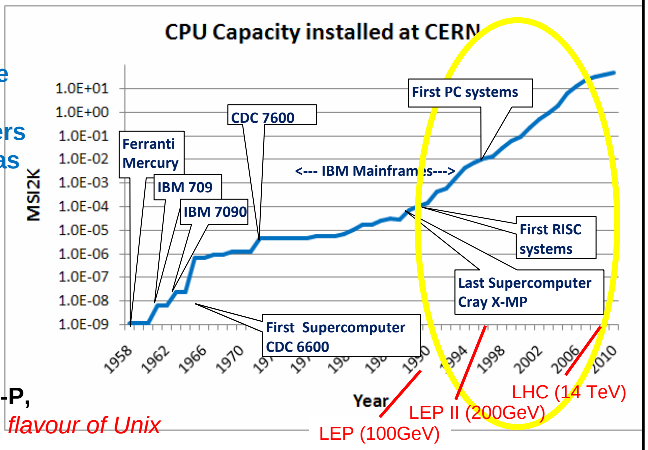
Ken Thompson & Denis Ritchie
1969 – Unix 1972 – C
1981 – first Unix service at CERN

2 orders of magnitude in 24 years

Clusters of Inexpensive Processors

requirements driven

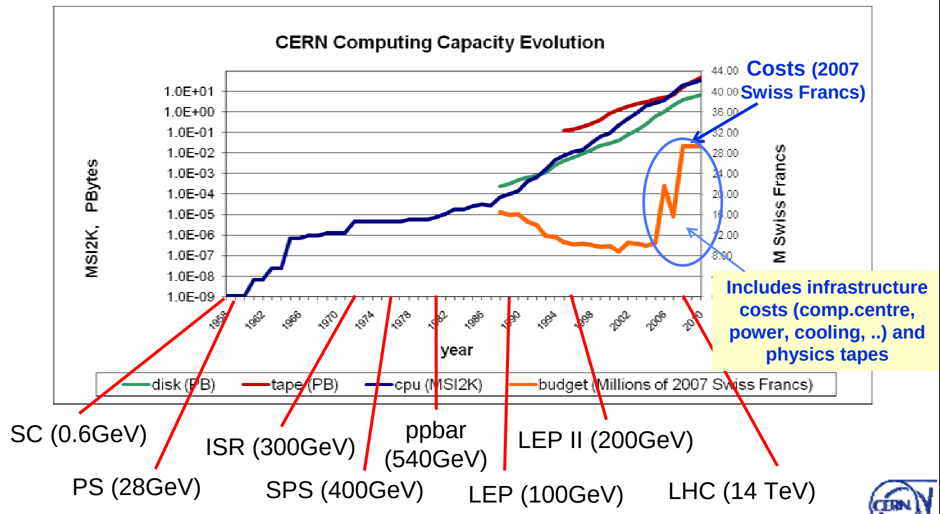
- We started this phase with a simple architecture that enables sharing of storage across cpu servers
- that proved stable and has survived from RISC thru Quad-core
- Parallel, high throughput
- Sustained price/perf improvement ~60% /yr
- Apollo DN10.000s
  1989 20 MIPS/proc



- 1990--- SUN, SGI, IBM, H-P, DEC, ... *each with its own flavour of Unix*
- 1996 – the first PC service with Linux
- 2007 – dual quad core systems
  → 50K MIPS/chip → 10**8 MIPS available == 2.3 MSI2K

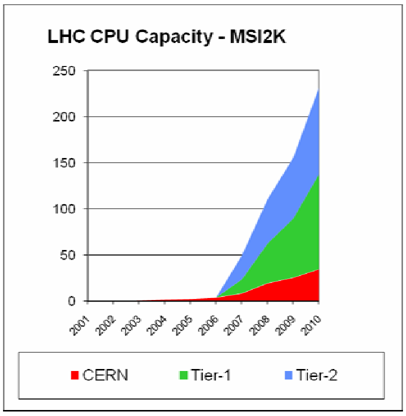
5 orders of magnitude in 18 years

Evolution of CPU Capacity at CERN



Ramping up to meet LHC requirements

- We need two orders of magnitude in 4 years – or an order or magnitude more than CERN can provide at the 220% per year growth rate we have seen in the "cluster" era, even with a significant budget increase
- But additional funding for LHC computing is possible if spent "at home"
- *A distributed environment is feasible given the easy parallelism of independent events*
- The problems are –
  - how to build this as a coherent service
  - How to make a distributed massively parallel environment usable



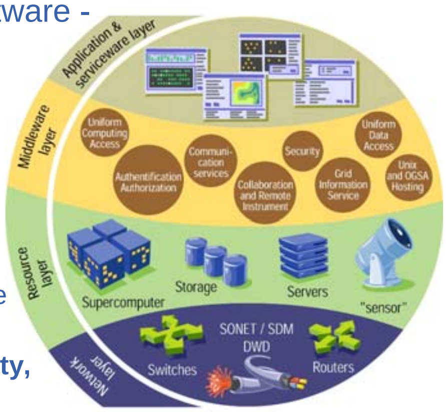
→ → **Computational Grids**

The Grid

- The **Grid** – a virtual computing service uniting the world wide computing resources of particle physics
- The **Grid** provides the end-user with seamless access to computing power, data storage, specialised services
- The **Grid** provides the computer service operation with the tools to manage the resources, move the data around

How does the Grid work?

- It relies on special system software - **middleware** – which:
  - keeps track of the location of the **data** and the **computing power**
  - balances the load on various resources across the different sites
  - provides common access methods to different data storage systems
  - handles: **authentication, security, monitoring, accounting, ...**

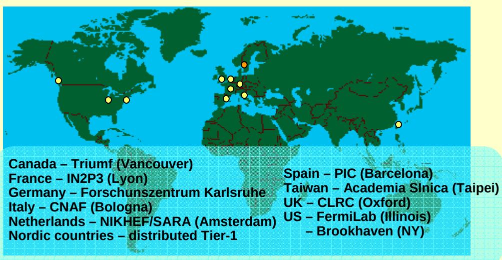


→a virtual computer centre

LCG Service Hierarchy

Tier-0 – the accelerator centre

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



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

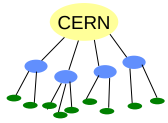
- Managed Mass Storage – → grid-enabled data service
- Data-heavy analysis
- National, regional support

Tier-2 – ~130 centres in ~35 countries

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

LHC Computing → Multi-science Grid

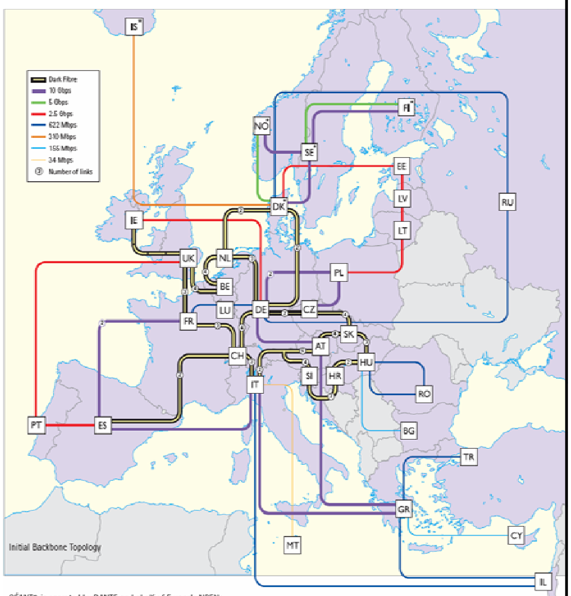
- 1999 - MONARC project
  - First LHC computing architecture – hierarchical
  distributed model
- 2000 – growing interest in grid technology
  - HEP community main driver in launching the DataGrid project
- 2001-2004 - EU DataGrid project
  - middleware & testbed for an operational grid
- 2002-2005 – LHC Computing Grid – LCG
  - deploying the results of DataGrid to provide a
  production facility for LHC experiments
- 2004-2006 – EU EGEE project phase 1
  - starts from the LCG grid
  - shared production infrastructure
  - expanding to other communities and sciences



The new European Network Backbone

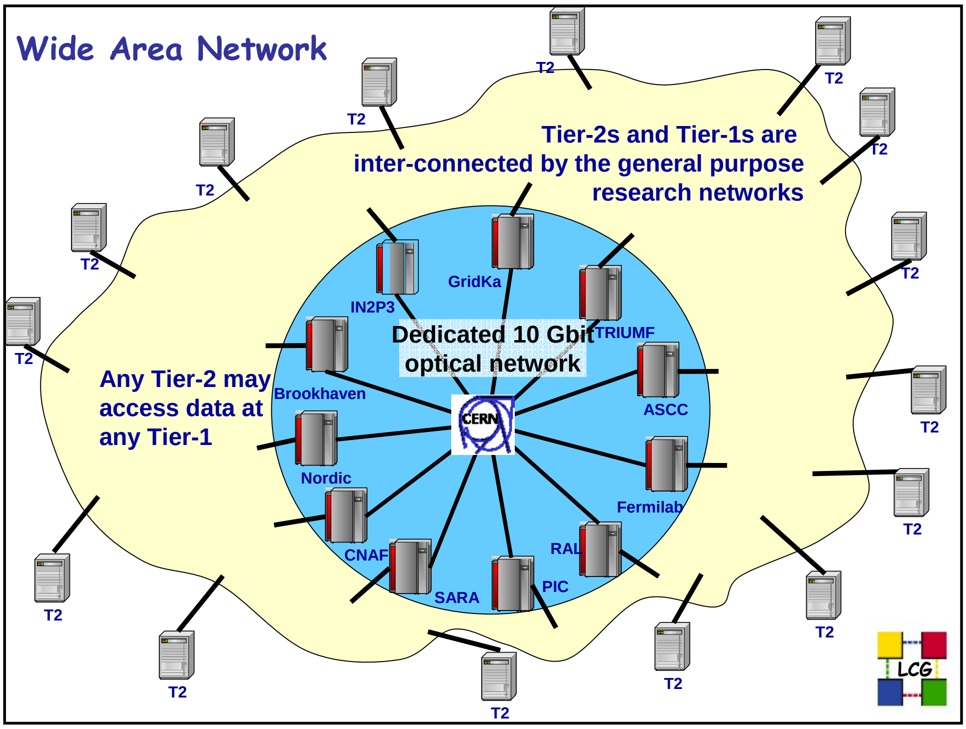
- LCG working group with Tier-1s and national/ regional research network organisations
- New GÉANT 2 – research network backbone

  → Strong correlation with major European LHC centres (Swiss PoP at CERN)
  → Core links are fibre



GÉANT2 is operated by DANTE on behalf of Europe's NRENs.

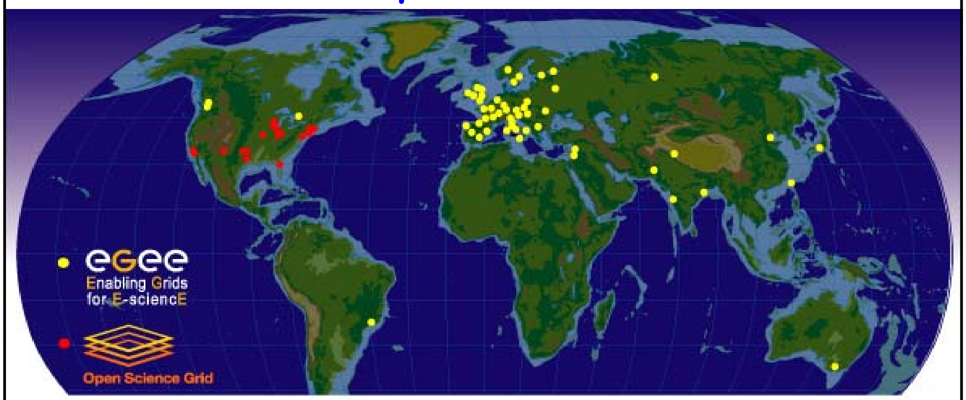
Wide Area Network



WLCG depends on two major science grid infrastructures ….

EGEE - Enabling Grids for E-Science

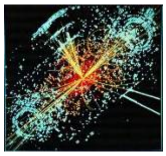
OSG - US Open Science Grid



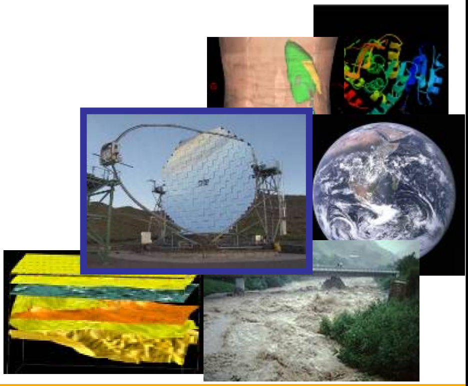
A map of the worldwide LCG infrastructure operated by EGEE and OSG.

Towards a General Science Infrastructure?

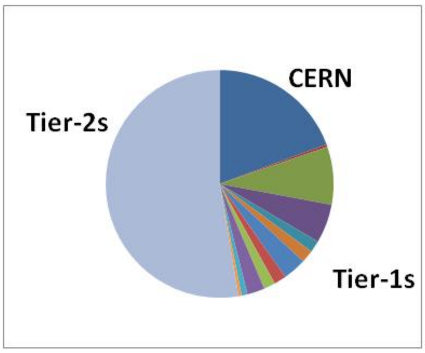
Enabling Grids for E-sciencE

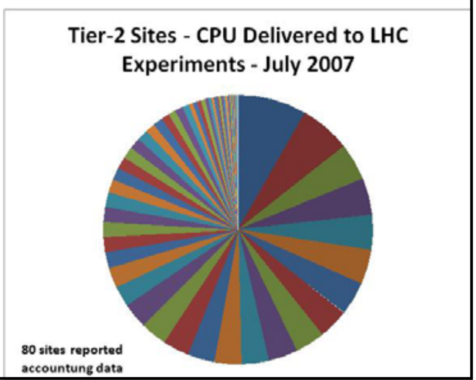


- More than 20 applications from 7 domains
  - High Energy Physics (Pilot domain)
    - 4 LHC experiments
    - Other HEP (DESY, Fermilab, etc.)
  - Biomedicine (Pilot domain)
    - Bioinformatics
    - Medical imaging
  - Earth Sciences
    - Earth Observation
    - Solid Earth Physics
    - Hydrology
    - Climate
  - Computational Chemistry
  - Fusion
  - Astronomy
    - Cosmic microwave background
    - Gamma ray astronomy
  - Geophysics
    - Industrial applications

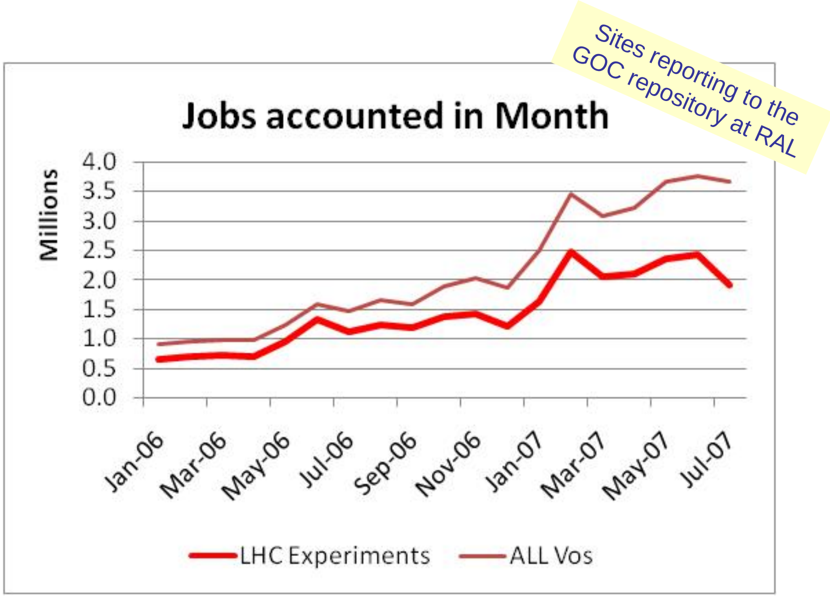


CPU Usage accounted to LHC Experiments July 2007

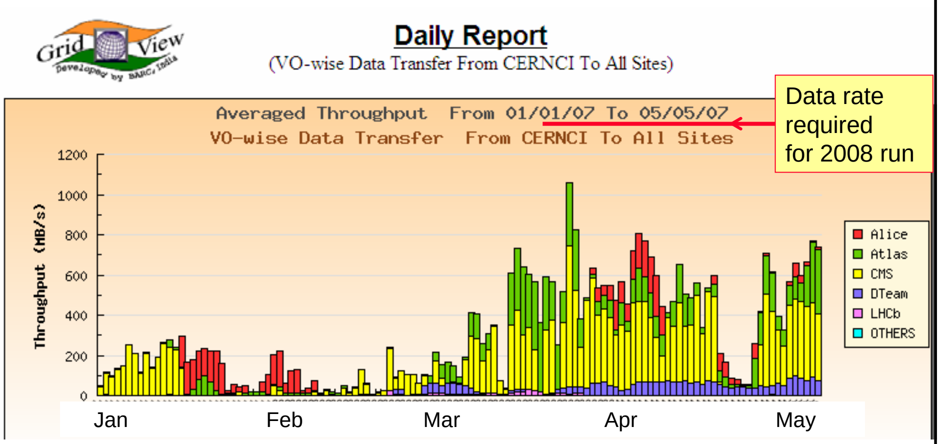




CERN	20%
11 Tier-1s	30%
80 Tier-2s	50%

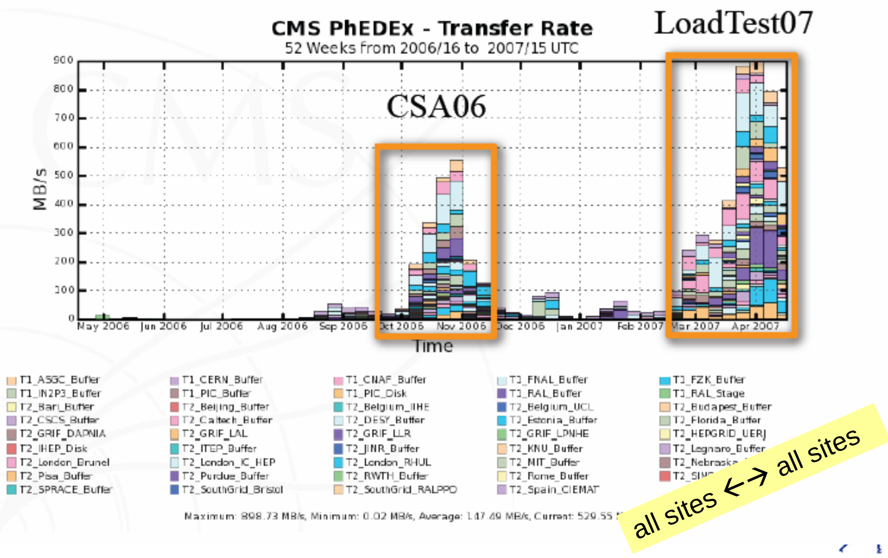


2007 – CERN →Tier-1 Data Distribution



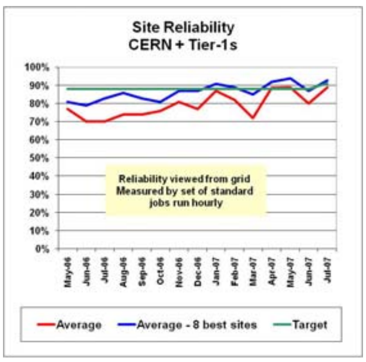
Average data rate per day by experiment (Mbytes/sec)

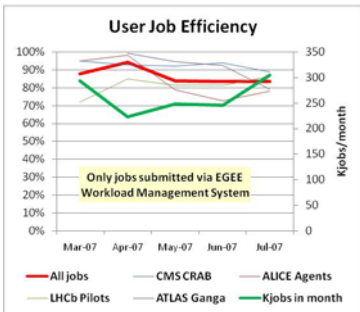
Data Transfers
Comparison with CSA06 — weekly



Reliability?

- Operational complexity is now the weakest link
  - Sites, services
  - Heterogeneous management
  - Major effort now on monitoring
  - Grid infrastructure, & how does the site look from the grid
  - User job failures
  - Integrating with site operations
- .. and on problem determination
  - Inconsistent, arbitrary error reporting
  - Software log analysis (good logs essential)





Early days for Grids

Middleware:

- Initial goals for middleware over-ambitious – but now a reasonable set of basic functionality, tools is available
- Standardisation slow –
  - Multiple implementations of many essential functions (file catalogues, job scheduling, ..), some at application level
- But in any case - useful standards must **follow** practical experience

Operations:

- Providing now a real service, with reliability (slowly) improving
- Data migration, job scheduling maturing
- Adequate for building experience – site and experiment operations

Experiments can now work on improving usability:

- a good distributed analysis application integrated with the experiment framework, data model
- a service to maintain/install the environment at grid sites
- problem determination tools – job log analysis, error interpreters, ..

So we can look forward to continued exponential expansion of computing capacity to meet growing LHC requirements, & improved analysis techniques?

A Few of the Challenges

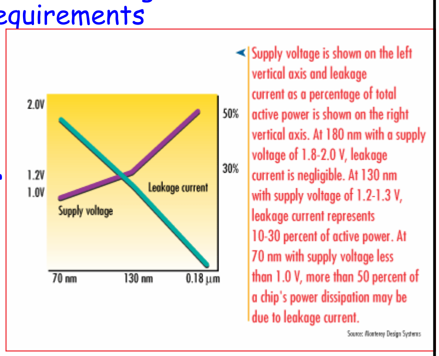
Energy

Costs

Usability

Energy and Computing Power

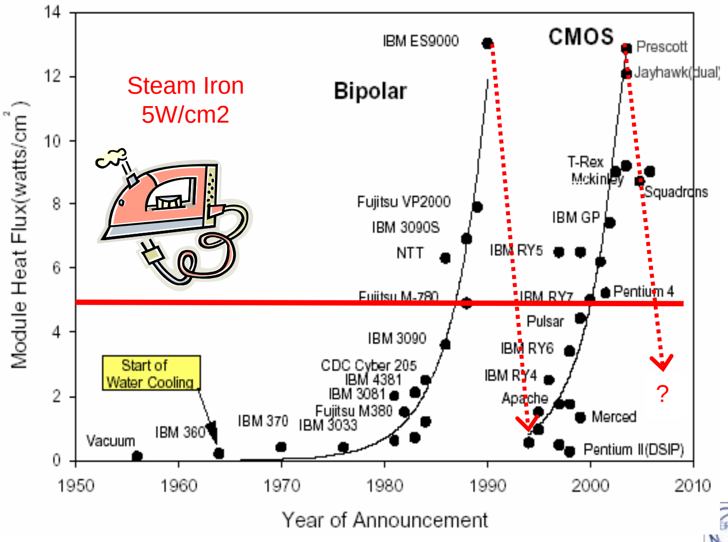
- As we moved from mainframes through RISC workstations to PCs the improved level of integration reduced dramatically the energy requirements
- Above ~180nm feature size the only significant power dissipation comes from transistor switching
- While architectural improvements could take advantage of the higher transistor counts the computing capacity improvement could keep ahead of the power consumption
- But from ~130nm two things have started to cause problems –
  - Leakage currents start to be a significant source of power dissipation
  - We are running out of architectural ideas to use the additional transistors that are (potentially) available



Chip Power Dissipation

FIO

CERN IT Department

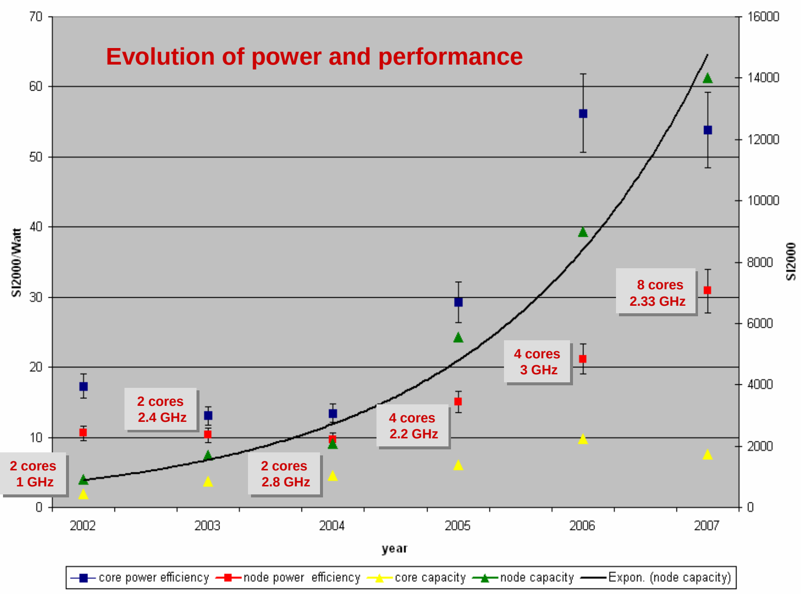


Tony Cass
CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Power Growth

FIO

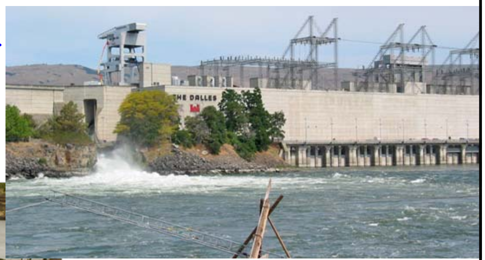
CERN IT Department



Tony Cass
CERN - IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Energy Consumption – Today's major constraint to continued computing capacity growth

- Energy is increasingly expensive
- Power and cooling infrastructure costs vary linearly with the energy content – no Moore's law effect here
- Energy dissipation becomes increasingly problematic as we move towards 30KVA/m^2 and more with a standard 19" rack layout
- Ecologically anti-social
- Google, Yahoo, MSN have all set up facilities on the Columbia River in Oregon - renewable low-cost hydro power



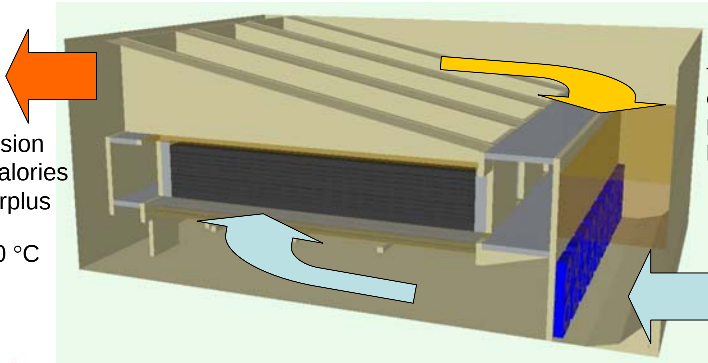
Chipping away at energy losses

- Techniques to reduce current leakage:
  - Silicon on Insulator
  - Strained silicon - more uniform → faster electron transfer
  - Stress memorisation - lower density N-channels
  - P-channel isolation using silicon-germanium
- Techniques that work fine for office and home PCs – but do not help over-loaded HEP farms
  - Power management – shut down the core (or part of it) when idle
  - Many-core processors with special-purpose cores – audio, graphics, network, .. – that are powered only when needed
- Good for HEP
  - Many-core processors – sharing power losses in off-chip components – as long as the cores are general-purpose
  - Single-voltage boards
  - More efficient power supplies

IBM

La réalisation de centres informatiques haute densité et écologiques

Un bâtiment permettant d'héberger une informatique très haute densité (30 kW/m²) et refroidi naturellement pendant 70% à 80% de l'année.



Paris 2007
DATACENTER FACILITIES & ENGINEERING CONFERENCE / EXPO

LCG

How might this affect LHC?

ON THE OTHER HAND –

- The grid environment and high speed networking allow us to place our major capacity essentially anywhere
- Will CERN install its computer centre in the cool, hydro-power-rich north of Norway?

Prices and Costs

Price = f(cost, market volume, supply/demand, ..)

For ten years the market has been ideal for HEP

- the fastest (SPECint) processors have been developed for the mass market – consumer and office PCs

Will we continue to ride the mass market wave?

- the standard (1Gbps) network interface is sufficient for HEP clusters – maybe need a couple
- Windows domination has **imposed hardware standards**
- and so there is reasonable competition between hardware manufacturers for processors storage, networking
- while Linux has freed us from proprietary software

Prices and Costs

- PC sales growth expected in 2007 (from IDC report via PC World)
  - 250M units (+12%)
  - More than half Notebook (sales up 28%)
  - But desktop and office systems down
  - And revenues grow only 7% (to ~$245B)

- With notebooks as the market driver -
  - Will energy (battery life, heat dissipation) become more important than continued processor performance?
- Applications take time to catch up with the computing power of multi-core systems
  - There are a few ideas for using 2-cores at home
  - Are there any ideas for 4-cores, 8-cores??
- Reaching saturation in the traditional home + office markets?

Prices and Costs

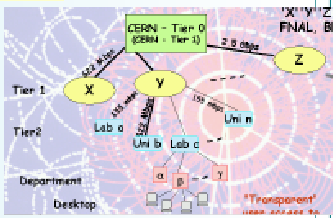
- And what about handheld devices ?
  -- will they handle the mass market needs
  -- connecting wirelessly to everything
  -- including large screens, keyboards whenever there is a desk at hand?
- But handhelds have very special chip needs –
  -- low energy, gsm, gps, flash memory or tiny disks, ...
- Games continue to demand new graphics technology
  - On specialised devices?
  - or will PCs provide the capabilities?
  - and will that come at the expense of general purpose performance growth?

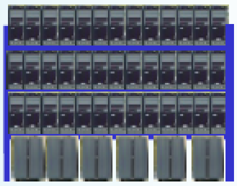
Will scientific computing slip back into being a niche market with higher costs, higher profit margins → higher prices?

How can we use all of this stuff effectively and efficiently

Les collaborations LHC (étapes d'analyse)

Raw Data (PetaBytes)	DAQ -> T0 -> T1
Après reconstruction (100 TeraBytes)	T1 -> T2
Pour l'analyse (10 TeraBytes)	T2 -> T3
Analyse par physicien (1 TeraByte)	





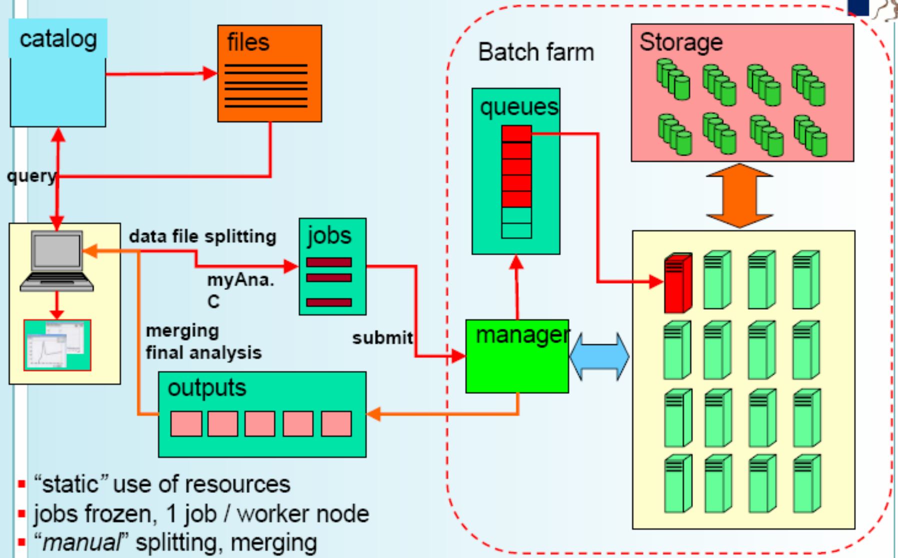
How do we use the Grid

- We are looking at ~100 computer centres
  - With an average of 100 PCs
  - Providing 2,000 cores
- So a total of ~200K cores (+ notebooks, PDAs, etc...)
- And ~100 millions files for each experiment
- Keeping track of all this, and keeping it busy is a significant challenge

We must use Parallelism at all levels

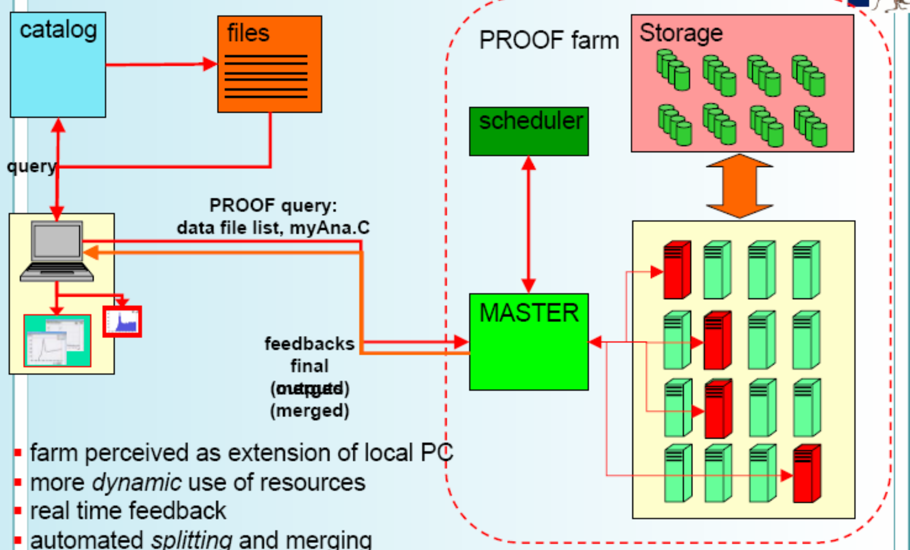
- There will be 200K cores each needing a process to keep it busy –
- Need analysis tools that
  - keep track of 100M files in widely distributed data storage centres
  - can use large numbers of cores and files in parallel
  - and do all this transparently to the user
- The technology to this by generating batch jobs is available
- But the user –
  - Wants to see the same tools, interfaces, functionality on the desktop and on the grid
  - Expects to run algorithms across large datasets with "interactive" response times

Batch: l'approche classique



- "static" use of resources
- jobs frozen, 1 job / worker node
- "*manual*" splitting, merging
- limited monitoring (end of single job)

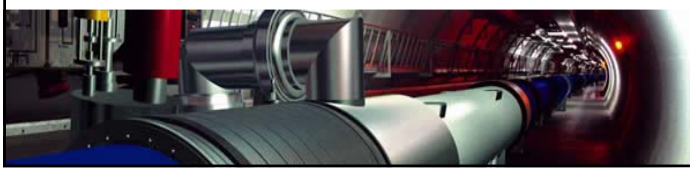
Interactif et Parallèle: ROOT/PROOF



- farm perceived as extension of local PC
- more *dynamic* use of resources
- real time feedback
- automated *splitting* and merging

Summary

- We have seen periods of rapid growth in computing capacity .. and periods of stagnation
- The grid is the latest attempt to enable continued growth – by tapping alternative funding sources
- Energy is looming as a potential roadblock – both for cost and environmental reasons
- Market forces, that have sustained HEP well for the past 18 years, may move away and be hard to follow
- But the grid is creating a competitive environment for services that opens up opportunities for alternative cost models, novel solutions, eco-friendly installations
- While enabling access to vast numbers of components that dictate a new interest in parallel processing
- This will require new approaches at the application level



Final Words

- Architecture is essential -- but KEEP IT SIMPLE
  - Flexibility will be more powerful than complexity
- Learn from history
  - So that you do not repeat it
- Develop through experience
  - First satisfy the basic needs
  - Do not over-engineer before the system has been exposed to users
  - Adapt and add functionality in response to **real needs, real problems**
  - Re-writing or replacing shows strength not weakness
- Standardisation can only follow practice
  - Standards are there to create competition, not to stifle novel ideas
- Keep focus on the science
  - Computing is the tool, not the target