CERN openlab II

CERN and Intel: Today and Tomorrow

Sverre Jarp
CERN openlab CTO
13 February 2008

Overview of CERN

What is CERN?

- CERN is the world's largest particle physics centre
- Particle physics is about:
  - elementary particles, the constituents all matter in the Universe is made of
  - fundamental forces which hold matter together
- Particles physics requires:
  - special tools to create and study new particles
    - Accelerators
    - Particle Detectors
    - Powerful computers

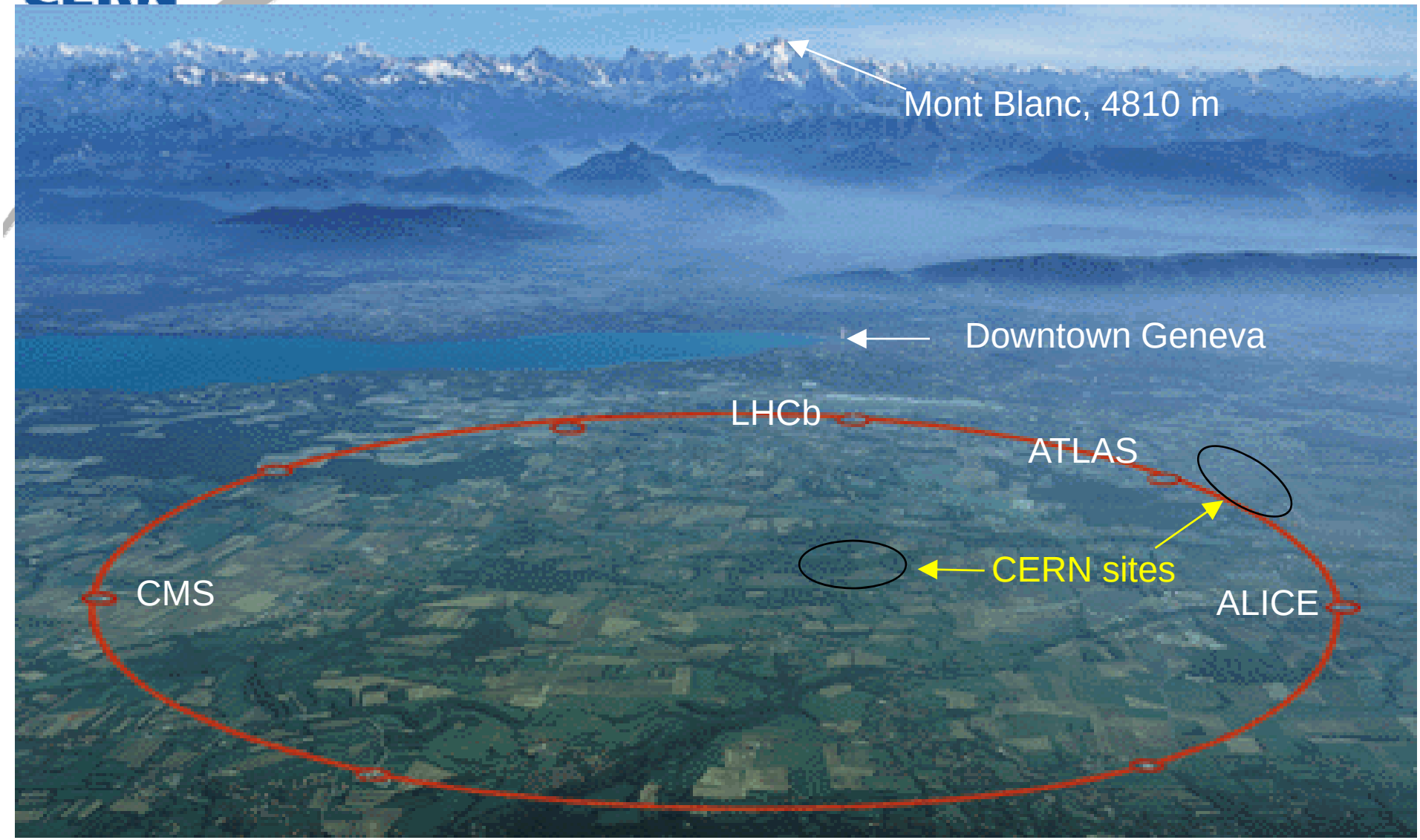
CERN is also:

-2500 staff (physicists, engineers , technicians, …)

- Some 6500 visiting scientists (half of the world's particle physicists)

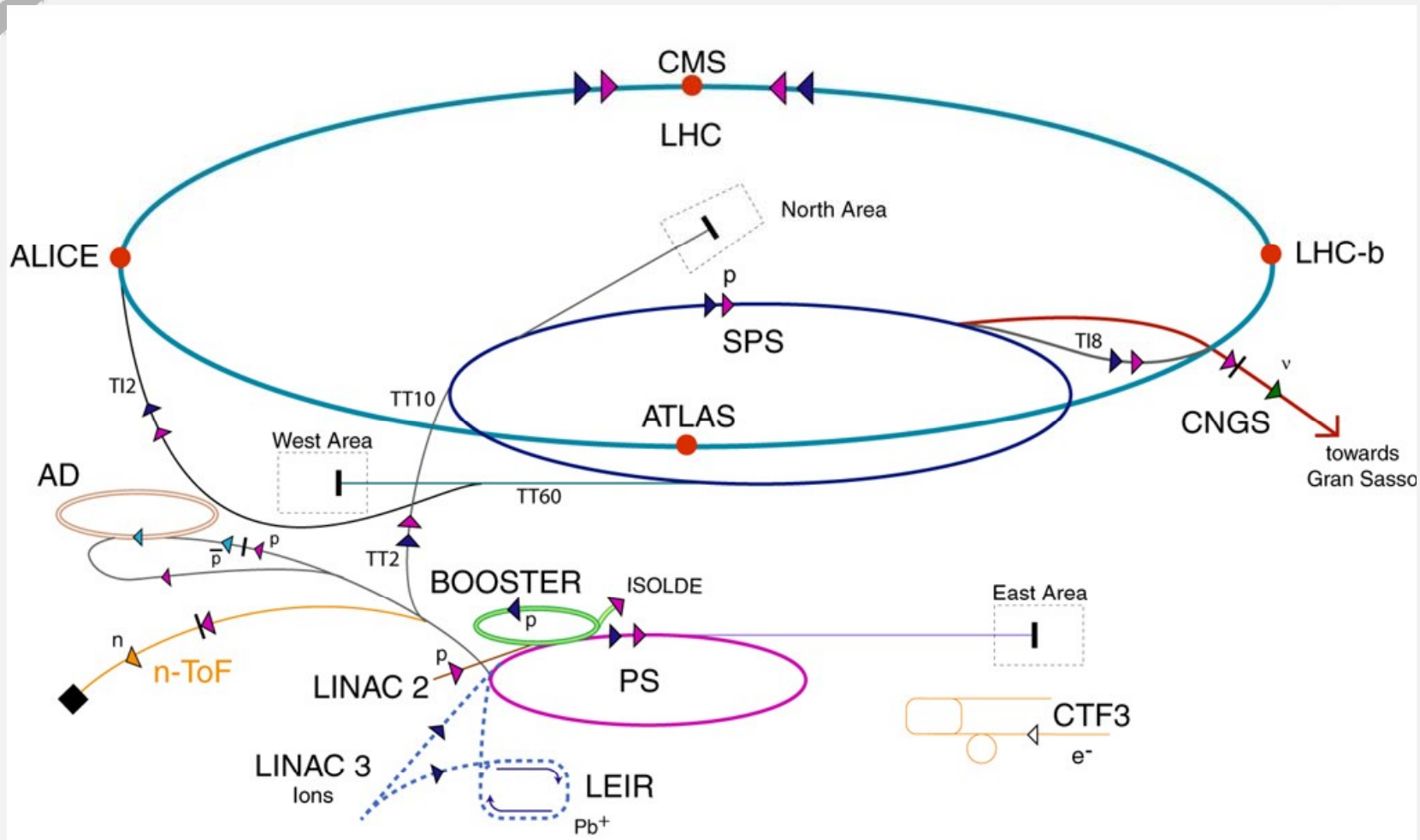
They come from 500 universities representing 80 nationalities.

The CERN Site



CERN's accelerators

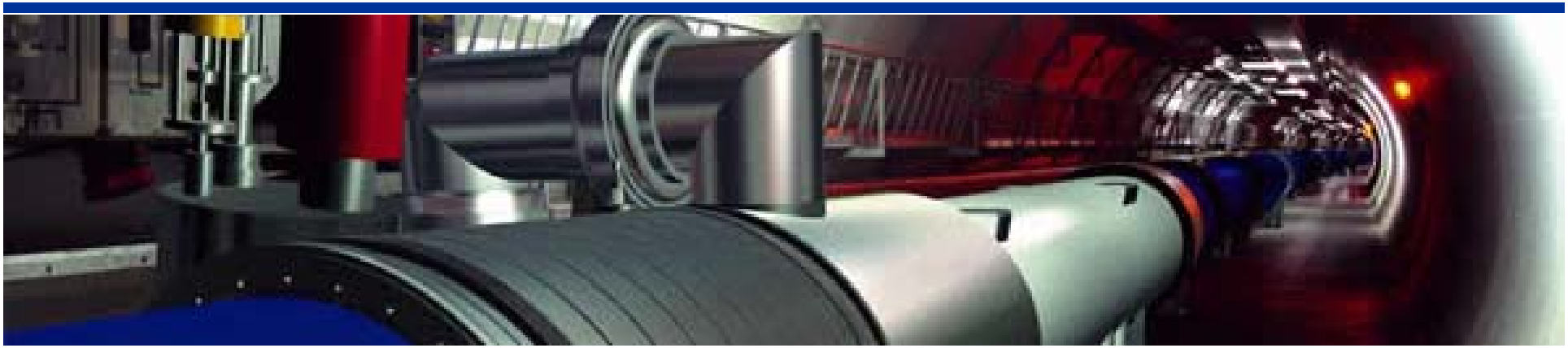
The world's most complete accelerator complex



What is LHC?

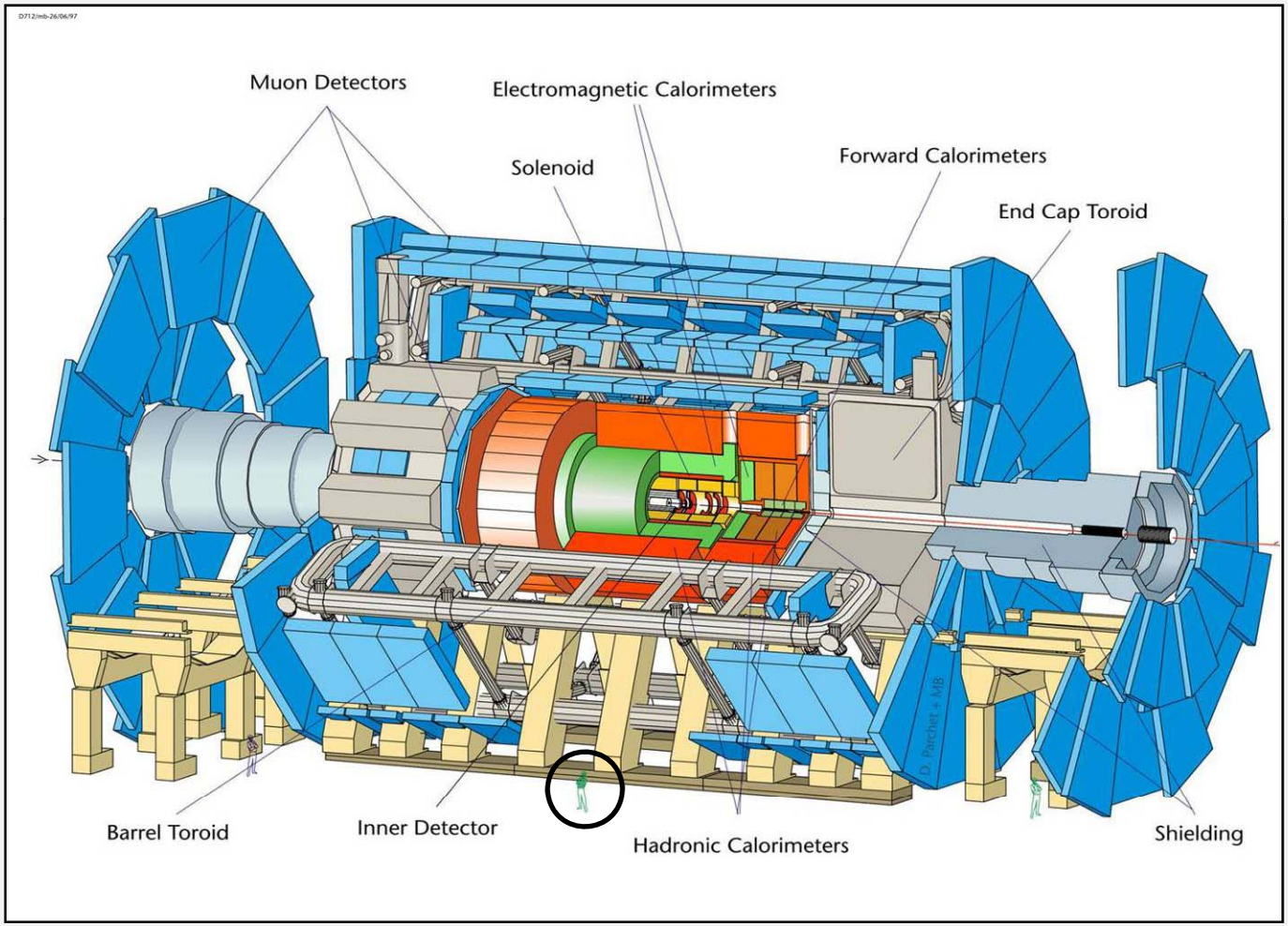
- The Large Hadron Collider will collide beams of protons at an energy of 14 TeV (in the summer of 2008)
- Using the latest super-conducting technologies, it will operate at about – 271°C, just above the temperature of absolute zero.
- With its 27 km circumference, the accelerator will be the largest superconducting installation in the world.

Four experiments, wit h detectors as 'big as cathedrals':
ALICE
ATLAS
CMS
LHCb



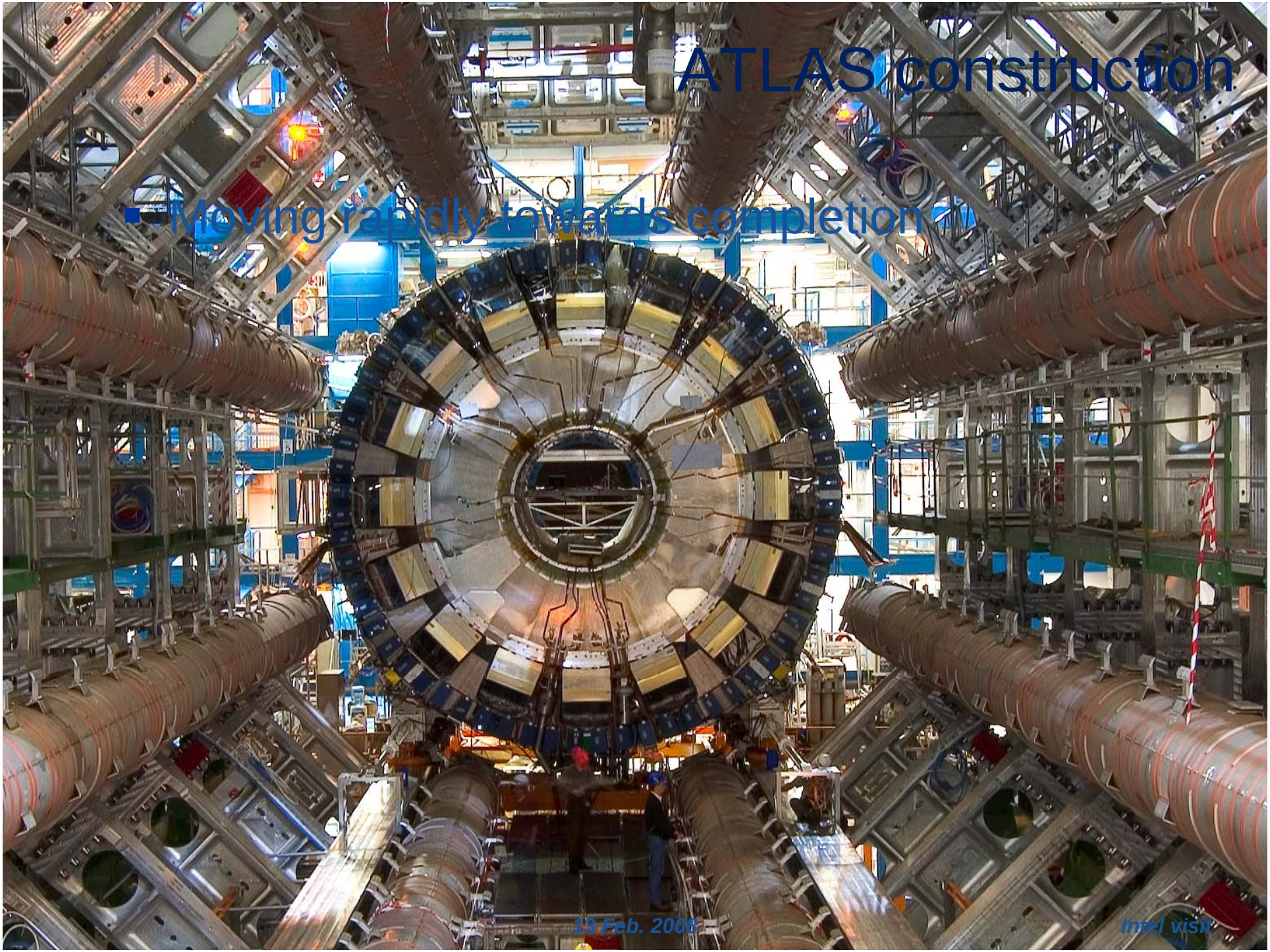
ATLAS

- General purpose LHC detector – 7000 tons

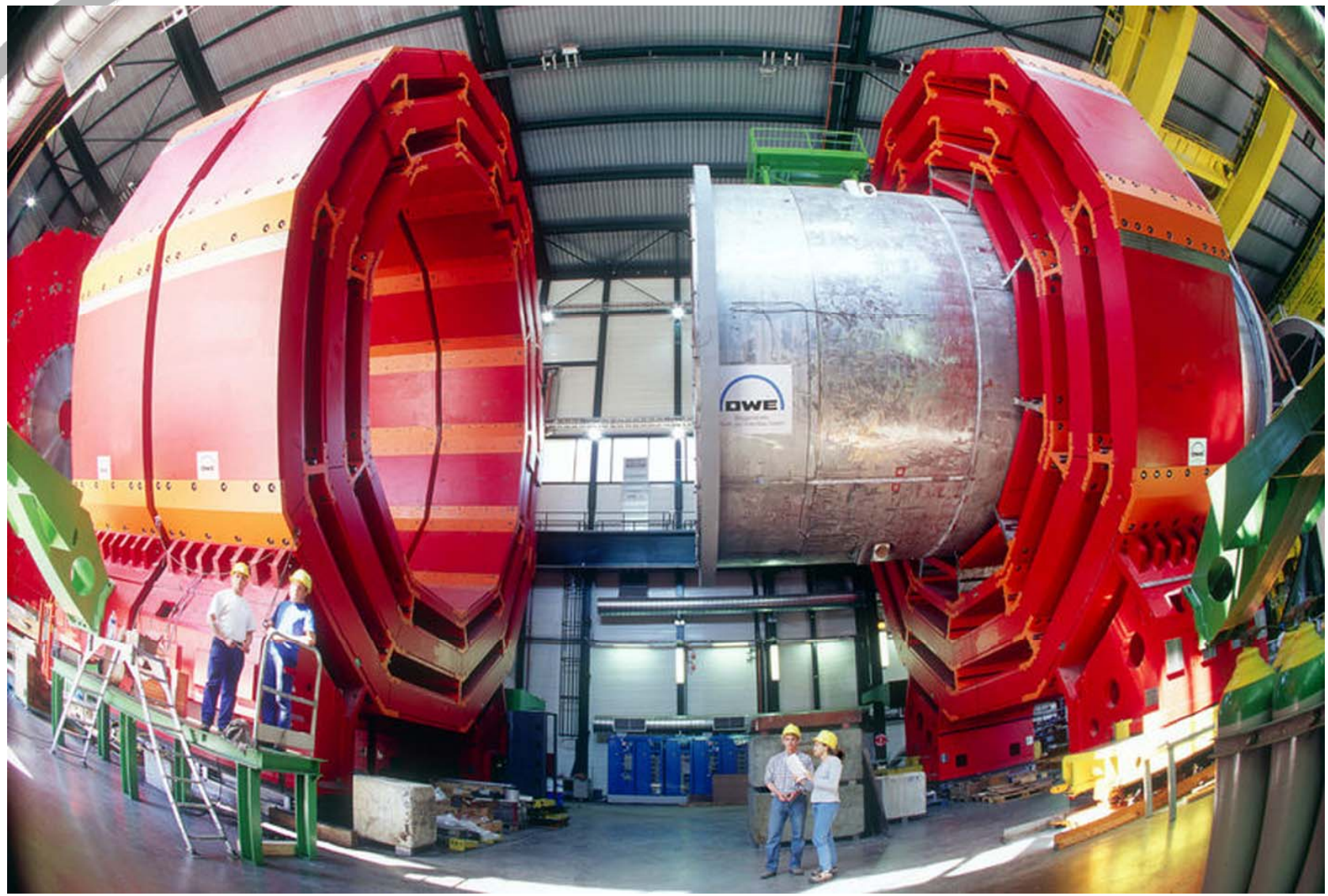


ATLAS construction

- Moving rapidly towards completion



Compact Muon Solenoid
(CMS)



Data management and computing

LHC data (simplified)

Per experiment:

- 40 million beam interactions per second
- After filtering, 100 collisions of interest per second
- A Megabyte of digitized information for each collision = recording rate of 0.1 Gigabytes/sec
- 1 billion collisions recorded = 1 Petabyte/year

1 Megabyte (1MB)
A digital photo

1 Gigabyte (1GB)
= 1000MB
A DVD movie

1 Terabyte (1TB)
= 1000GB
World annual book production

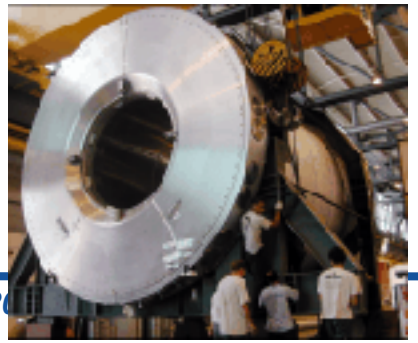
1 Petabyte (1PB)
= 1000TB
The annual production by one LHC experiment

1 Exabyte (1EB)
= 1000 PB
World annual information production

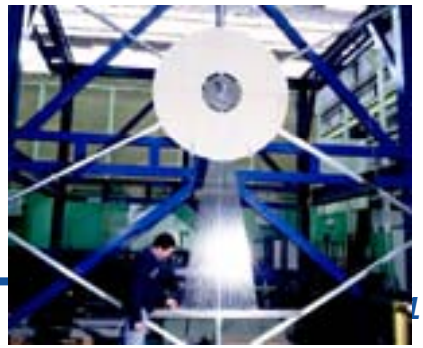
CMS

LHCb

ATLAS



ALICE



Data Handling and Computation for Physics Analysis

selection & reconstruction

Reconstruction

event summary data

10%

processed data

100%

raw data

event reprocessing

batch physics analysis

Analysis

1%

analysis objects
(extracted by physics topic)

event simulation

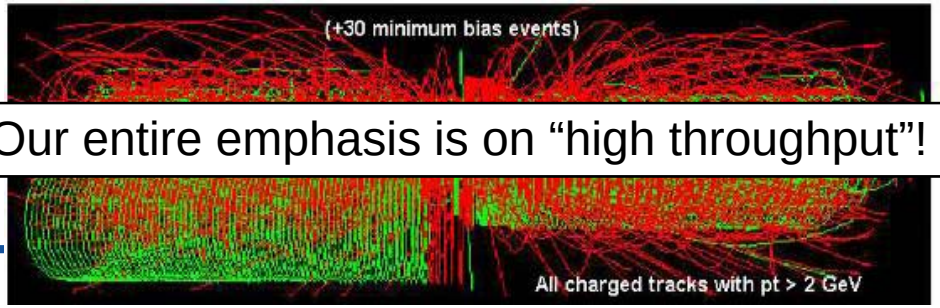
Simulation

interactive physics analysis

High Energy Physics Computing Characteristics

- Independent events (collisions of particles)
  - trivial (read: pleasant) parallel processing
- Bulk of the data is read-only
  - versions rather than updates
- Meta-data in databases linking to “flat” files
- Compute power scales with SPECint (not SPECfp)
  - But good floating-point (30% of total) is important!
- Very large aggregate requirements:
  - computation, data, input/output
- Chaotic workload –
  - research environment - physics extracted by iterative analysis, collaborating groups of physicists

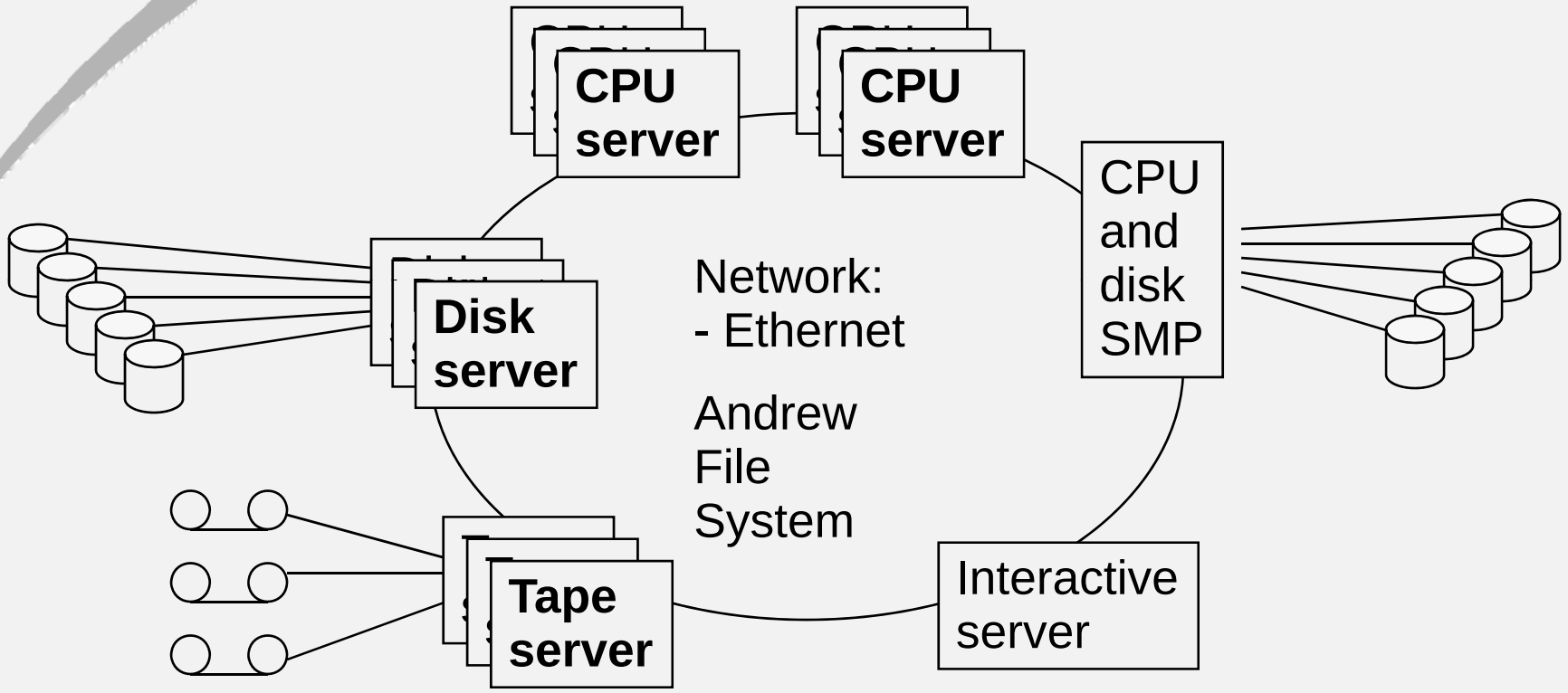
  → Unpredictable → unlimited demand



Our entire emphasis is on “high throughput”!

SHIFT architecture

(Scalable Heterogeneous Integrated Facility)



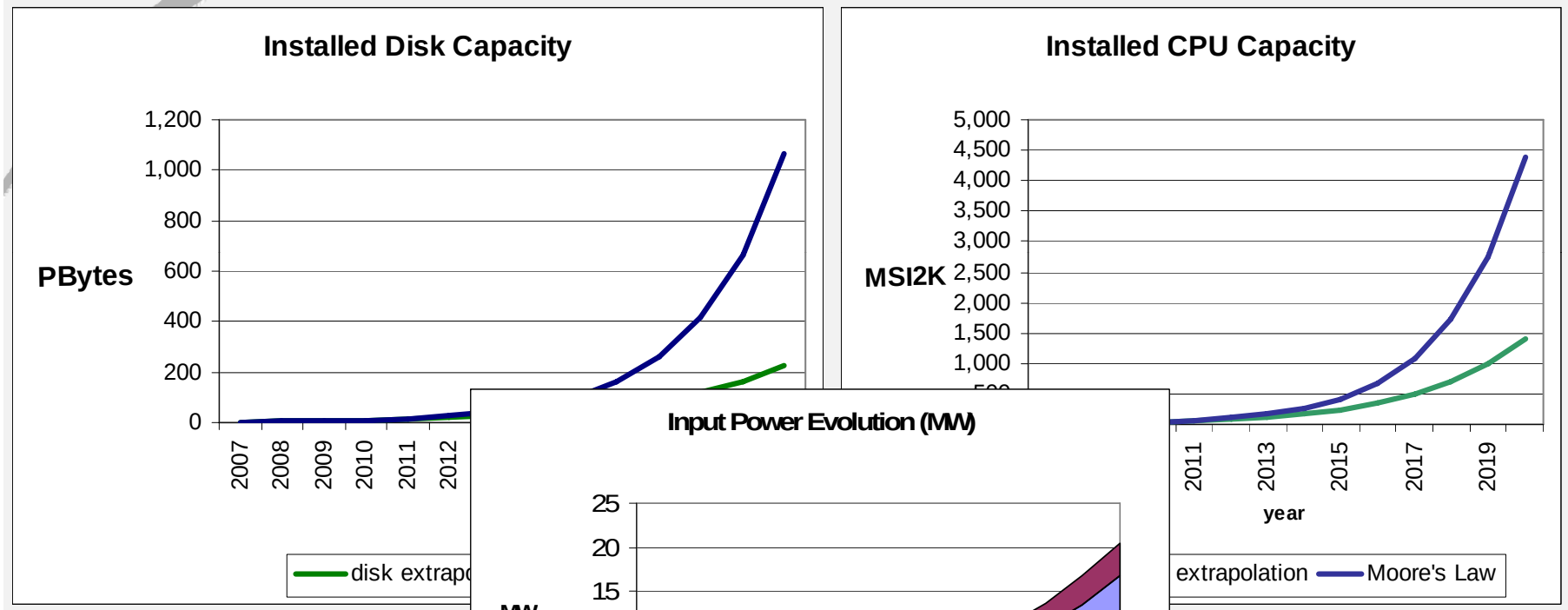
In 2001 SHIFT won the 21st Century Achievement Award issued by Computerworld

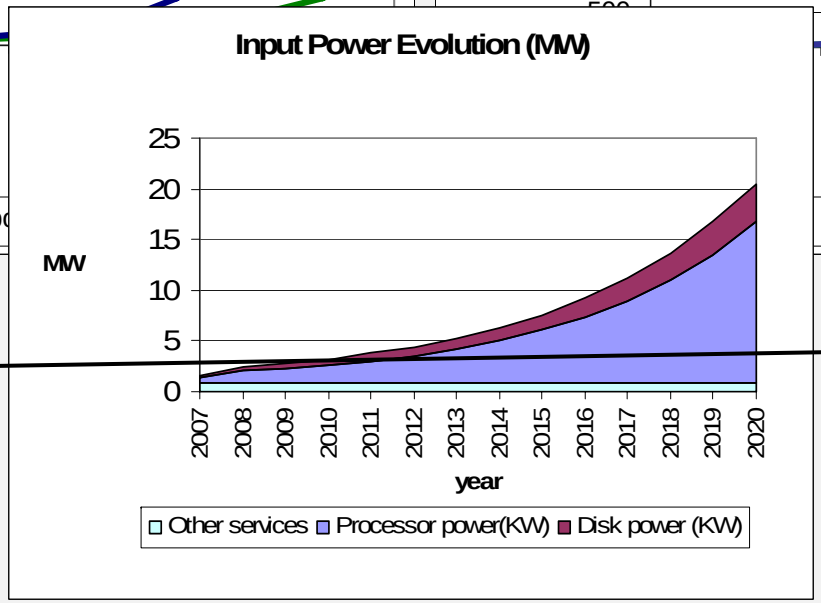
Computing at CERN today

- High-throughput computing based on reliable “commodity” technology
- About 3000 dual-socket PC servers running Linux
- More than 5 Petabytes of data on tape; 20% cached on disk

LHC computing capacity development

Development of computing capacity with a constant budget, given the increased cost for power and cooling

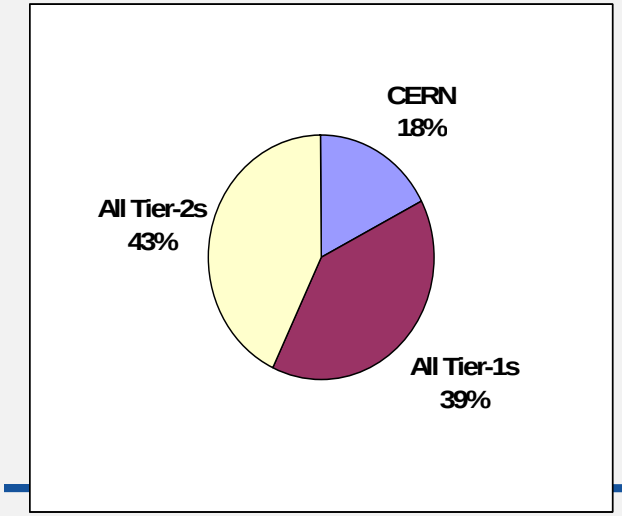


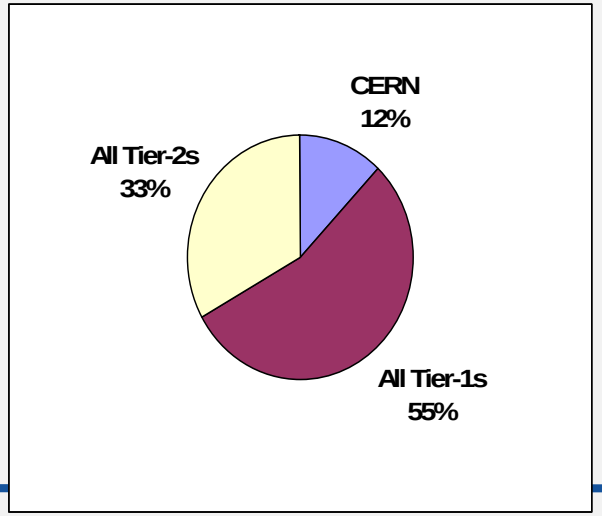


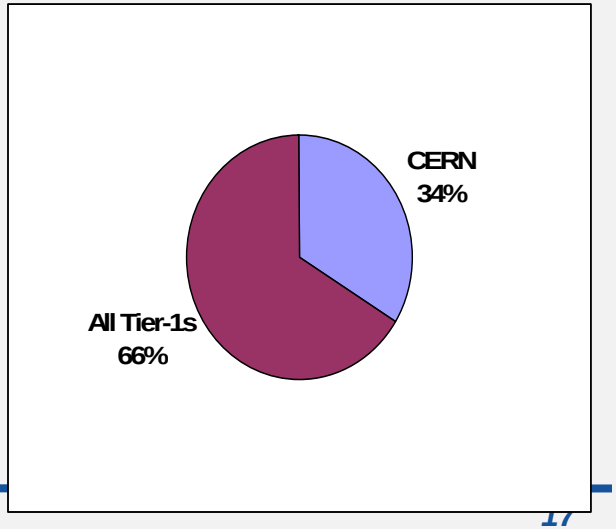
Power limit of present Computer Centre

Why do we need a Grid?

- The LHC Computing requirements are simply too huge:
  - Political resistance to putting everything at CERN
  - Impractical to build such a huge facility in one place
  - The users are in any case not necessarily at CERN
  - Modern wide-area networks have made distances shrink
    - But, latency still has to be kept in mind
- So, we are spreading the burden!



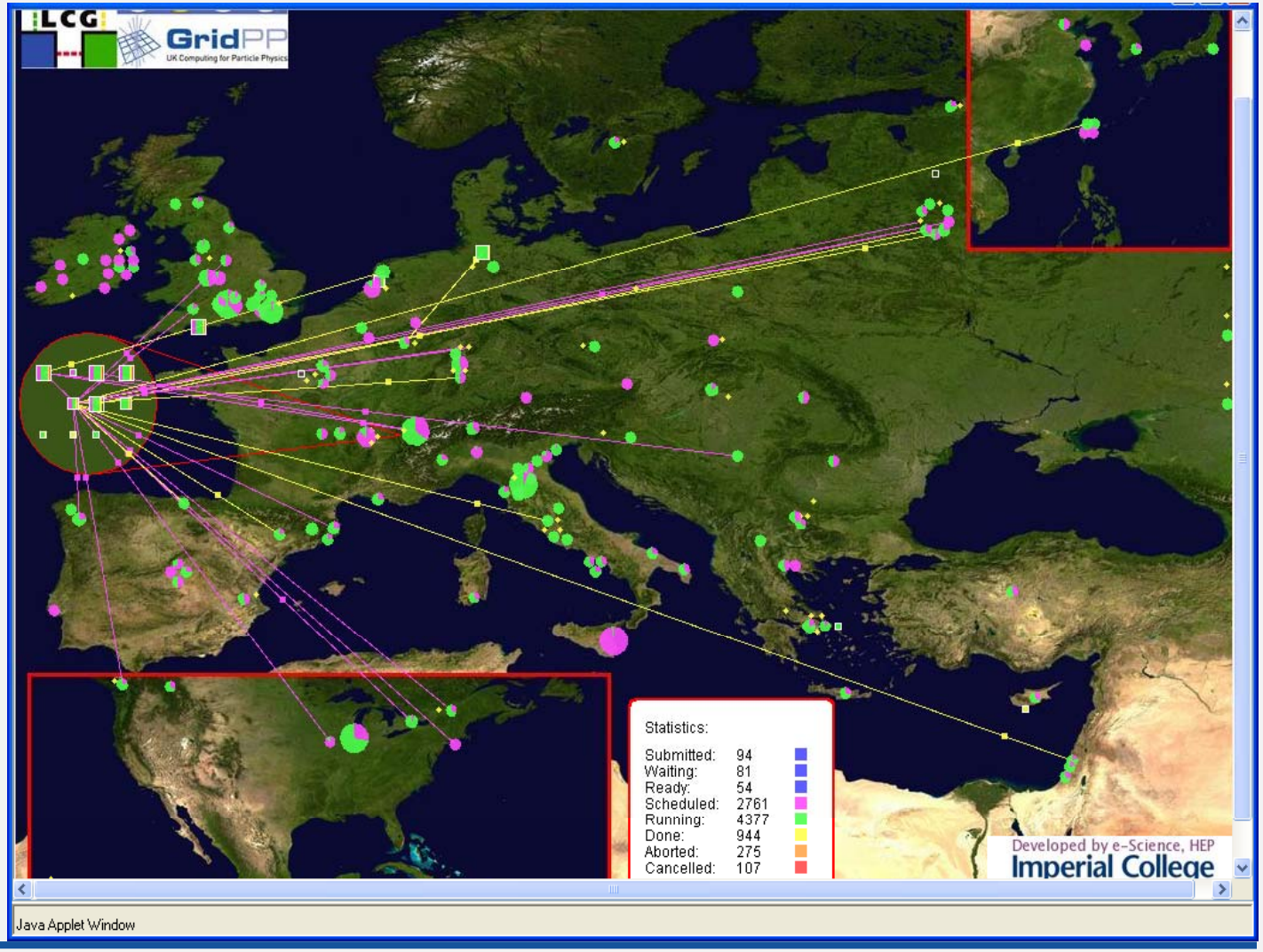




WLCG

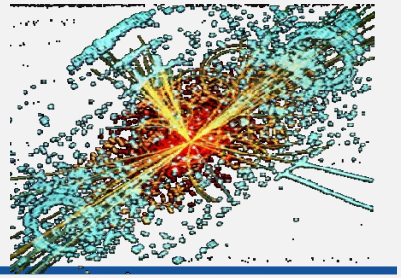
- Largest Grid service in the world !

- **Almost 200 sites in 39 countries**
- **37’000 IA-32 processors (w/Linux)**
- **Tens of petabytes of storage**



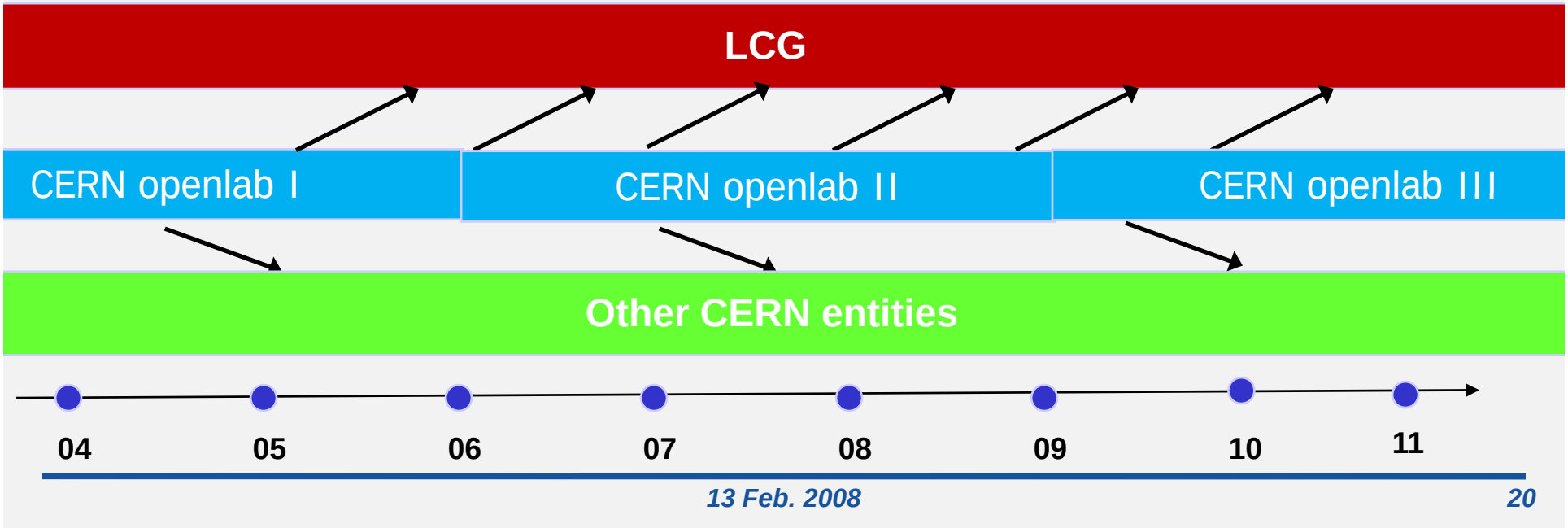
Background to the CERN openlab

- Information Technology has ALWAYS moved at an incredible pace
- During the LEP era (1989 – 2001) CERN changed its computing infrastructure twice:
  - Mainframes (1x) → RISC servers (30x) → PC servers (1000x)
- In openlab, we collaborate to harness the advantages of a continuous set of innovations for improving scientific computing, such as:
  - 10 Gigabit networks, 64-bit computing, Virtualization
  - Performance improvements (Moore's law): HW and SW
  - Many-core throughput increase, Thermal optimization
- We work with a long-term perspective:
  - LHC will operate until at least 2020!



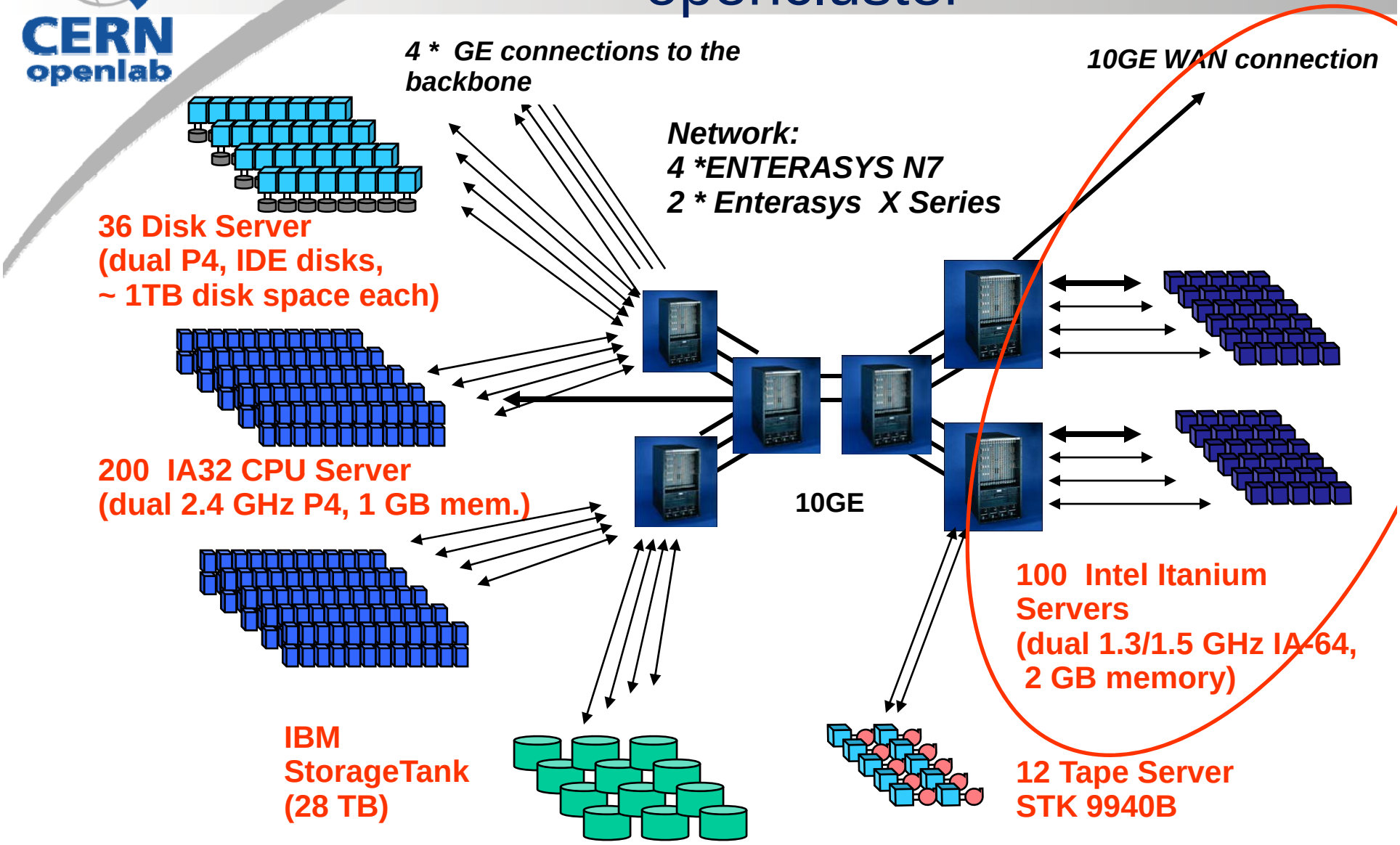
The CERN openlab

- Department's main R&D focus
- Framework for collaboration with industry
- Evaluation, integration, validation
  - **of cutting-edge technologies that can serve the LHC Computing Grid (LCG)**
- Sequence of 3-year agreements
  - **2003 – 2005: Phase I: the "opencluster" project**
  - **2006 – 2008: Phase II Platform Competence Centre**



Highlights from the CERN/Intel collaboration in openlab I (2003 – 2005)

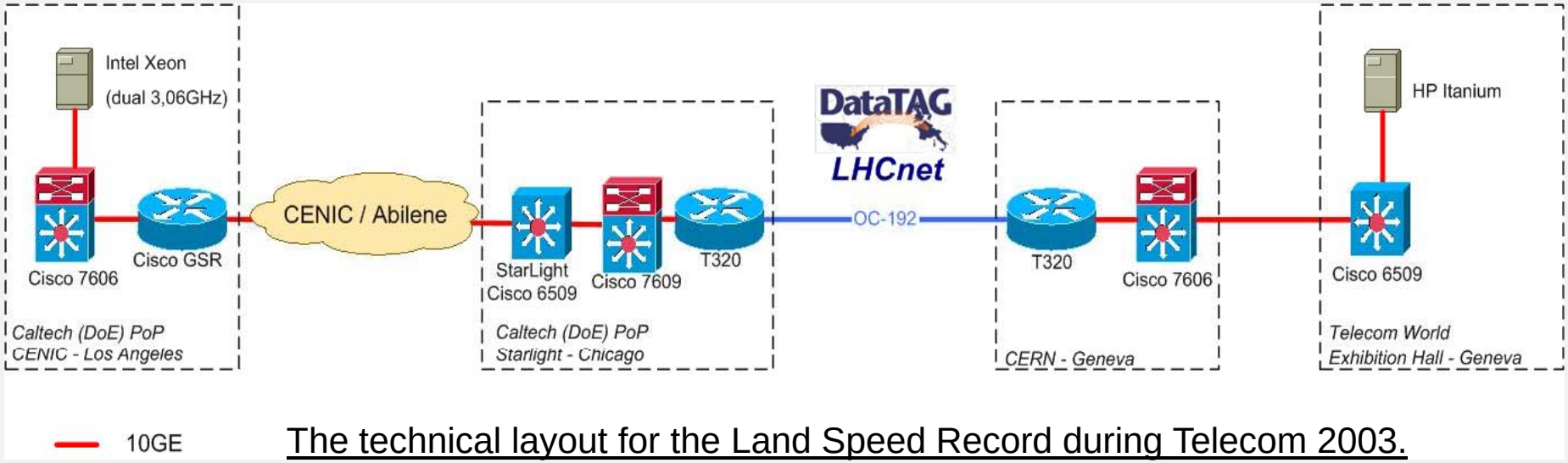
High performance test bed: the opencluster



High Througput Prototype (opencluster + LCG testbed)

Example of early success
(Land Speed records)

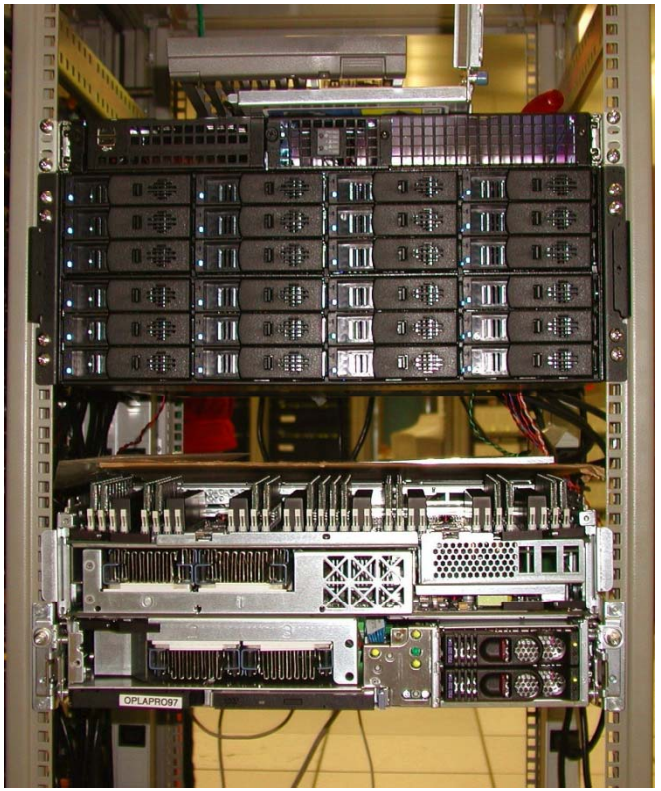
- Network speed record established during Telecom 2003
  - Aim: Transfer the equivalent of a full DVD (5 GB) with a single Itanium server and the Intel 10Gb Network Interface Card (NIC) in less than 10 seconds to California
  - Great success!
    - At Telecom: 5.44 Gbit/s over 7'067 km
    - Early in 2004: 6.57 Gbits/s over 15'766 km



The technical layout for the Land Speed Record during Telecom 2003.

Prototyped next generation disk servers (2004 – 05)

- Based on excellent equipment:
  - Two 4-way Itanium servers (RX4640)
    - Two full-speed PCI-X slots
      - 10 GbE and/or Infiniband
    - Two sets of RAID controllers
    - 24 * SATA disks with 74 GB
      - WD740 “Raptor” @ 10k rpm
      - Sustained R/W speed of 55 MB/s



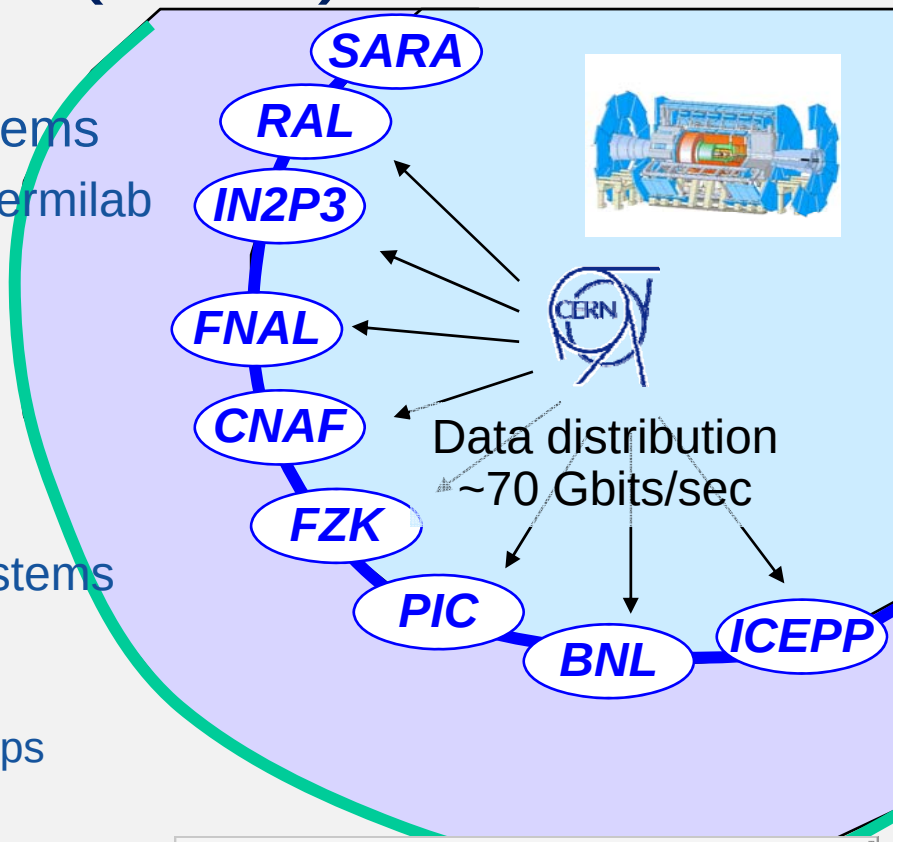
	Three 3-ware controllers	Three Areca controllers
Read speed	**~ 1100 MB/s**	**~1100 MB/s**
Write speed	**~550 MB/s**	**~900 MB/s**

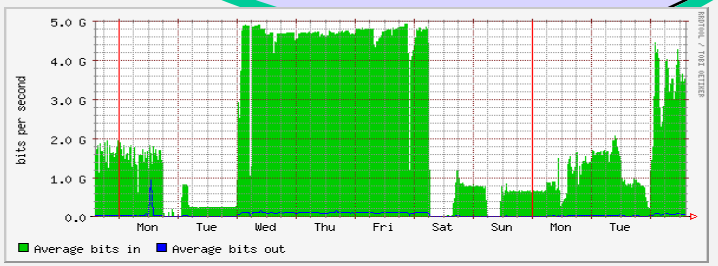
Spearheaded LCG Service Challenges (2005)

- SC1 (Data Export) used 10 Itanium systems
  - Achieved stability between CERN and Fermilab at 5 Gbps (over a 10 Gbit link)
    - Transfers via GridFTP
    - Limited by receiving side

- SC2:
  - Tests successfully with more Itanium systems and more sites
    - Fermilab, Karlsruhe @ 10 Gbps
    - CNAF, RAL, SARA, IN2P3, BNL @ 1 Gbps
    - Sustained 600 MB/s (peaks at 800 MB/s

- SC3:
  - Demonstrate “production quality” service
  - Move activities to Xeon systems
  - Incorporate CERN’s data management system

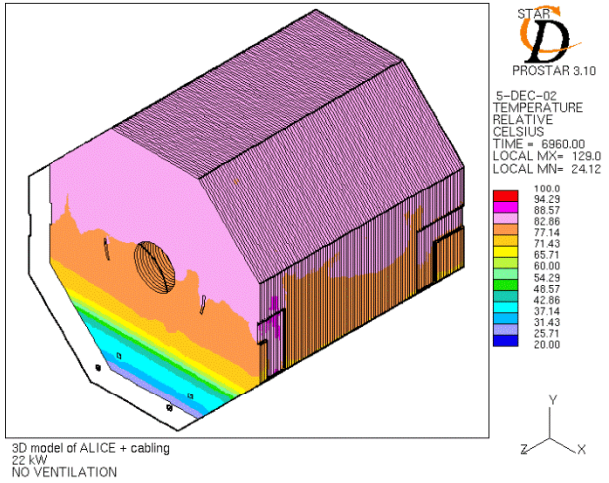


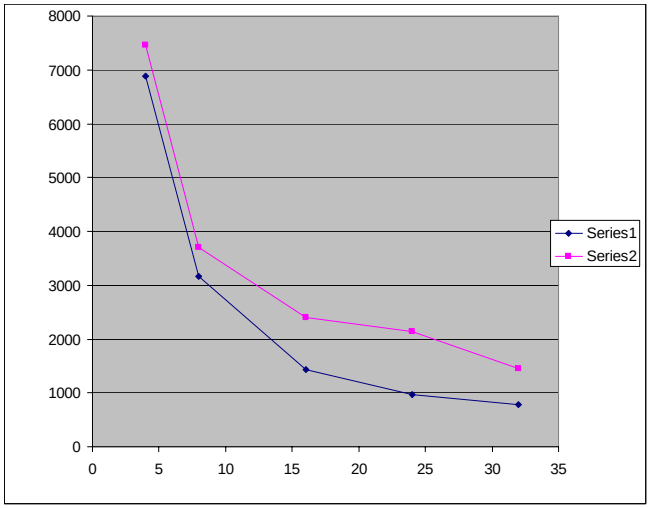


Stable data transfer at ~5Gbit/s to FNAL (US).

Computational Fluid Dynamics

- CFD on Itanium cluster
  - A numerical analysis of fluid flow, heat transfer and associated phenomena in physical systems
  - Always limited by available computing resources
  - Reduces design and engineering costs by avoiding prototype studies
  - Calculation improved by almost an order of magnitude
    - From, for instance, one month to less than four days
  - Model dimensions increased from 0.5 to 3 M cells





- Vital contribution to our LHC experiments
  - and other CERN entities
- Service is still running today

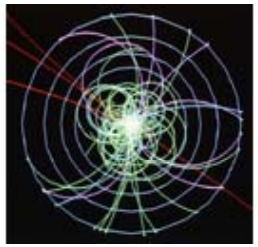
64-bit grid middleware

- In the beginning: The software chosen for LCG had been developed only with IA32 (and specific Red Hat versions) in mind
- Two openlab members worked extensively to complete the porting of LCG-2 software to Itanium
  - Result: All major components made to work on 64-bit Linux:
    - Worker Nodes, Compute Elements, Storage Elements, User Interface, etc.
  - Code, available via Web-site, transferred to HP sites (initially Puerto Rico and Bristol), as well as other interested sites
  - Changes given back to software maintenance teams
  - Porting experience summarized in white paper

All of a sudden
the Grid became heterogeneous and 64-bit capable!

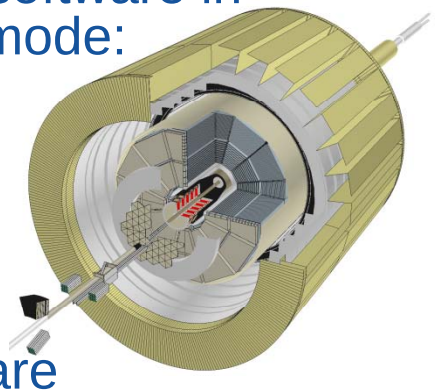
64-bit applications

- With 32 bits, a PC can address 4 Gigabytes
  - This is equivalent to the data that will come from the LHC detectors during a couple of seconds



- With 64 bits, a PC can address 18 Exabytes
  - This is equivalent to the data that would come from the LHC detectors during 100 years!

- CERN, and the HEP community, write almost all software in-house. All packages have been verified in 64-bit mode:
  - ROOT (Data analysis framework)
  - Geant4 (Physics simulation framework)
  - CLHEP (C++ Class Library)
  - CASTOR (CERN Hierarchical Storage Manager)



- LHC experiments have certified their entire software environment for this exciting capability

- All new PC servers are installed with Scientific Linux in 64-bit mode

Moving to openlab II (2006 – 2008)

Key technical contributors:
Håvard Bjerke, Andreas Hirstius, Sverre Jarp, Andrzej Nowak

Platform Competence Centre

- Intel-related activities:
  - Performance/throughput improvements
    - Compiler improvement project
    - Tuning of physics applications
    - Performance Monitoring
    - Benchmarking w/SPEC and Oracle
  - Multithreading evangelization
  - TOP500 runs
  - Virtualization
  - Thermal optimization
    - Servers and entire Computer Centre
  - 10 Gb networking

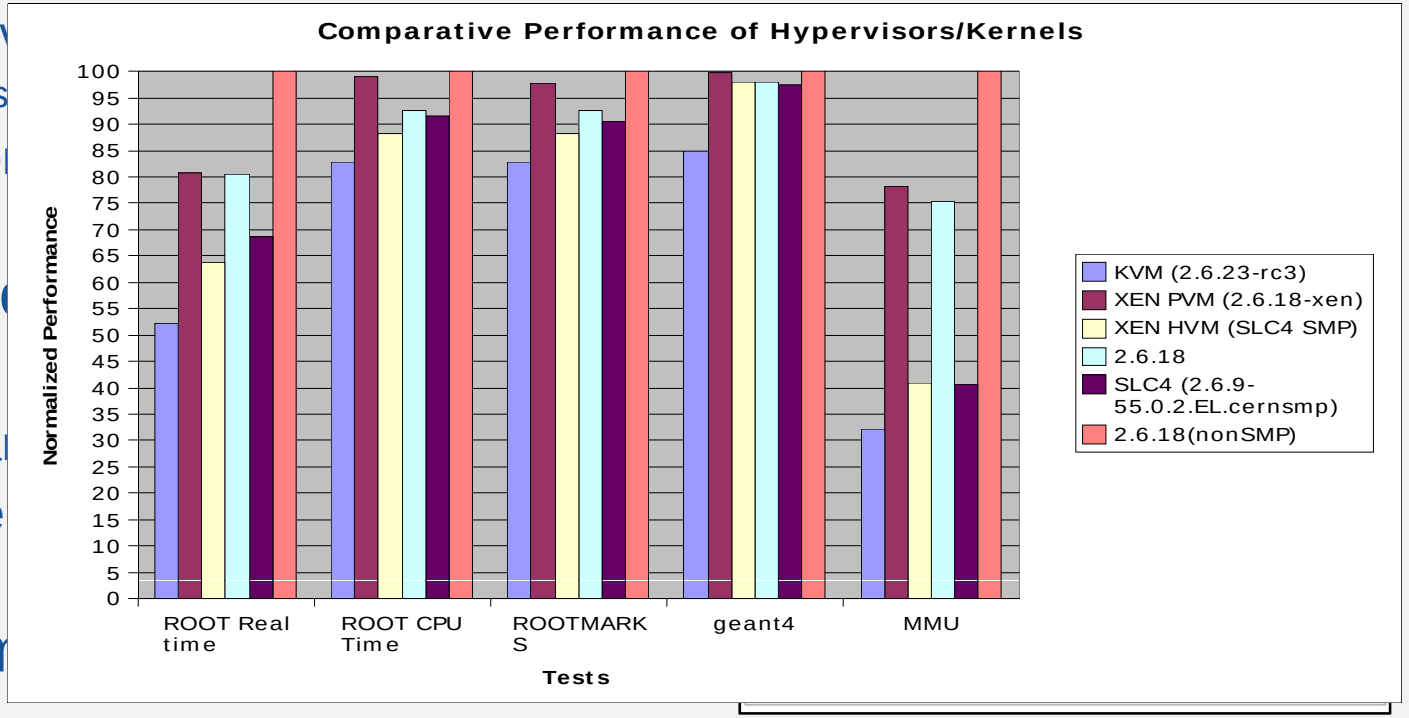
- Similar list of activities with the other partners in openlab

10 Gb Networking

- With the first generation cards, we successfully prototyped high-throughput disk servers, but
  - Very high cost
  - Reasonable throughput required jumbo-frames
    - MTU 9KB, rather than 1.5KB (Ethernet standard)
- Production disk servers (w/1Gb NICs) have now reached their limit in terms of throughput/capacity
- Today, optimistic that 2nd generation cards will be better
  - Reasonable cost, especially CX4
  - Native speed reachable (10Gb) with standard MTU
  - Driver support native in Linux kernel

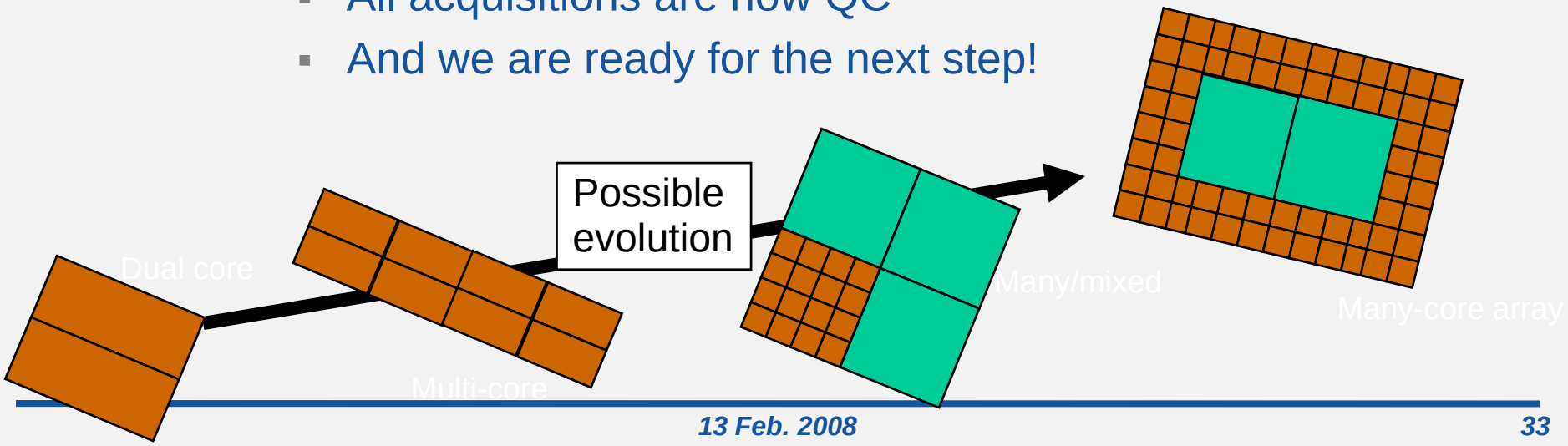
Virtualization

- Aim: Evangelize and demonstrate advantages of virtualization technology (mainly Xen)
  - System testing (actively used in LCG)
  - Server consolidation
  - Improv
    - als
  - Perso
- Also: D tools:
  - OSFa
  - Conte
- Benchm

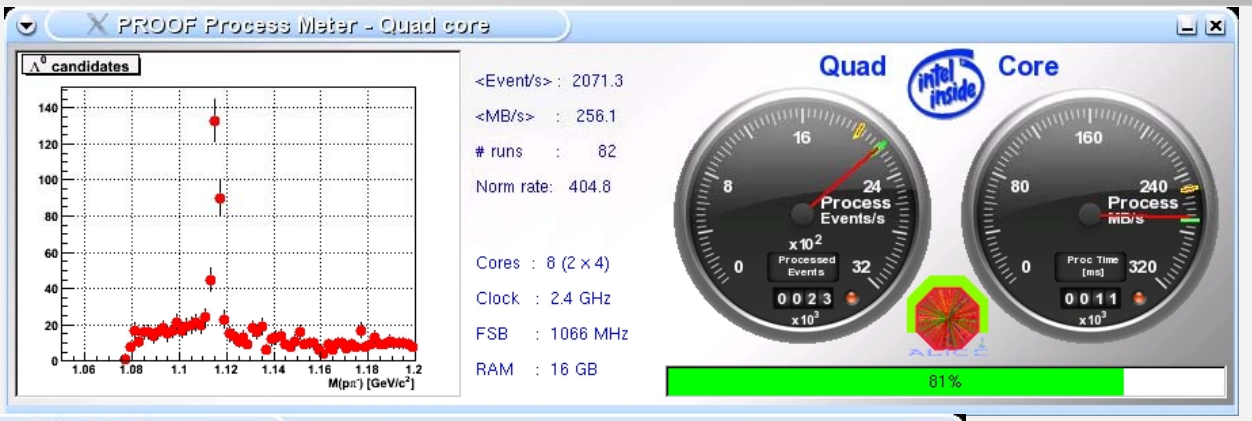


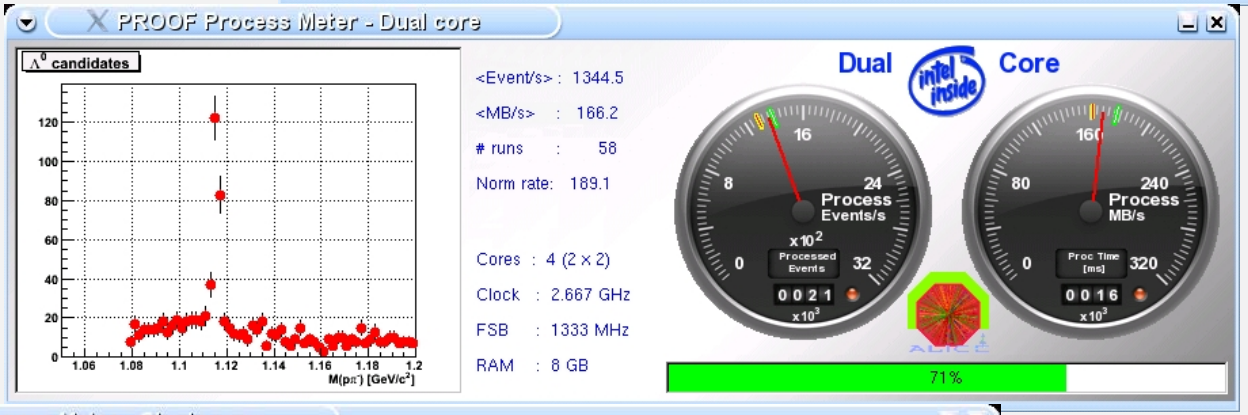
From Multi to Many

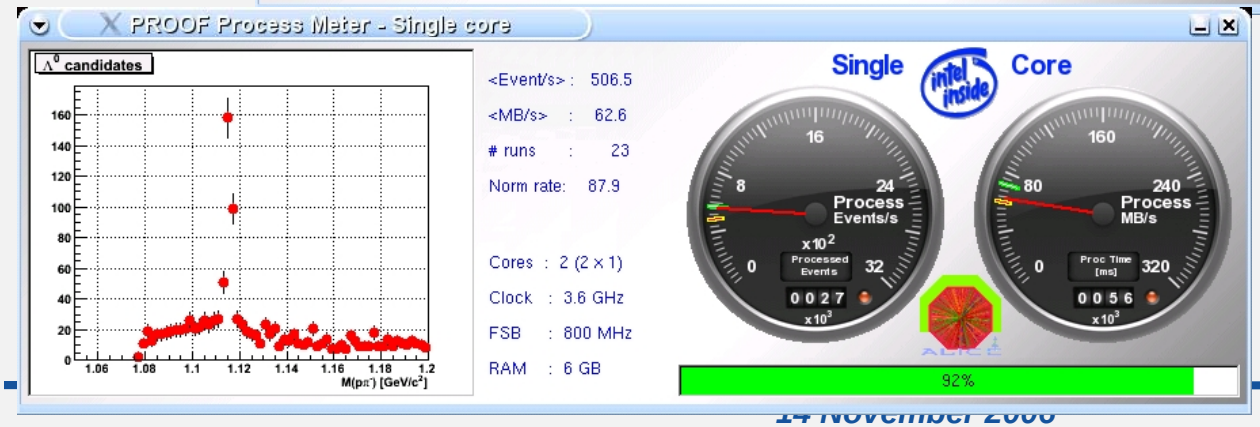
- Our “high throughput” computing model is ideally suited:
  - Independent processes can run on each core, provided that:
    - Main memory is added
    - Bandwidth to main memory remains reasonable
  - Testing, so far, has been very convincing
    - Woodcrest, Clovertown, Harpertown; Montecito
- In November 2006, we were proud to be part of Intel’s movement to Quad core
  - All acquisitions are now QC
  - And we are ready for the next step!



Multicore comparisons



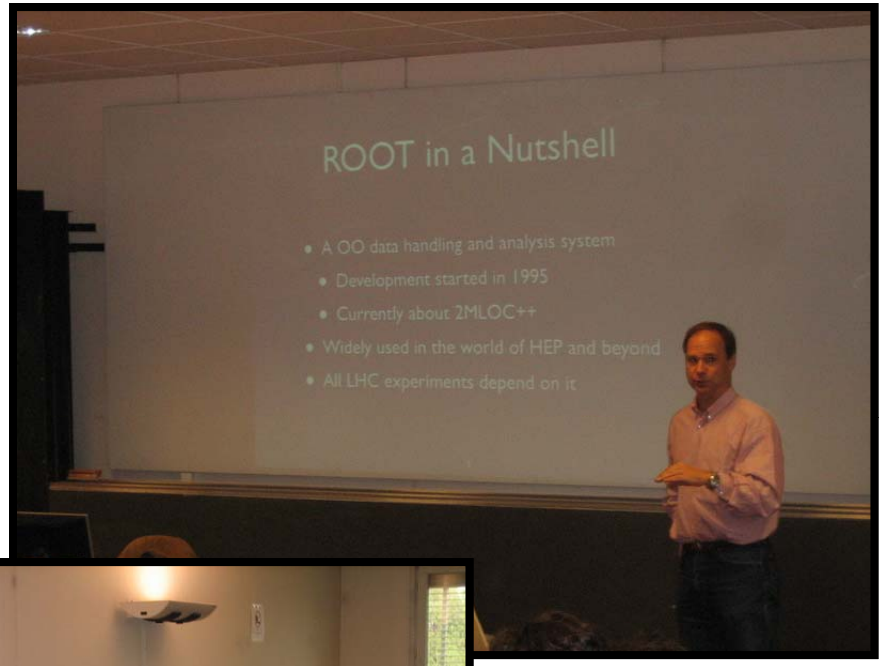




Multi-threading activities

- Aim: Evangelize/teach parallel programming
- Two workshops arranged w/Intel teachers in 2007
  - 1 day lectures, 1 day exercises
  - 5 lecturers (2 Intel + 3 CERN), 45 participants, 20 people oversubscribed
  - Survey: 100% said expectations met
  - Next workshop: Late Spring 2008
- Licenses for the Intel Threading Tools (and other SW products) available
  - to all CERN users
- Advances in Geant4 parallelization experiment

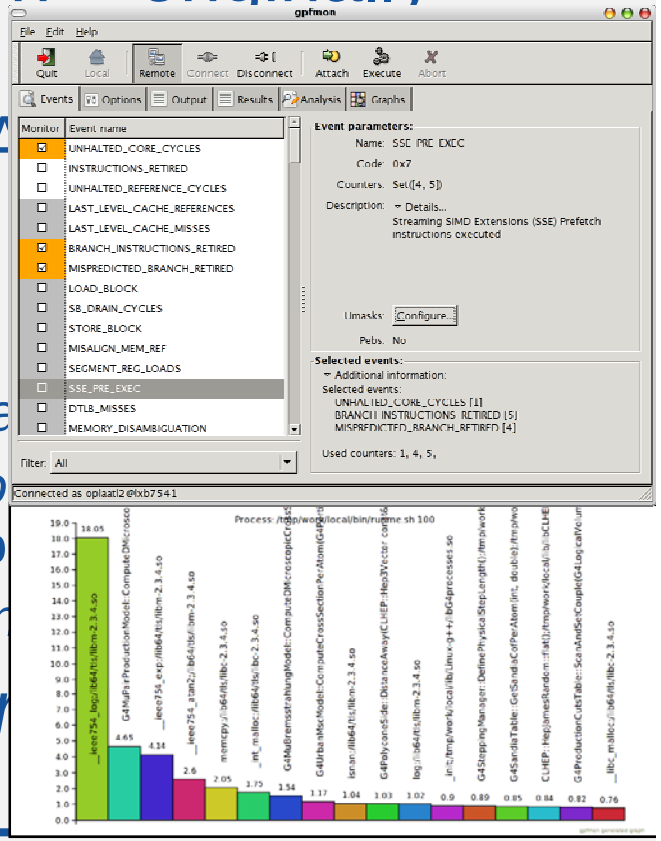
MT Workshop pictures



Herbert Cornelius and Hans-Joachim Plum from Intel

Performance Monitoring

- Started as a joint project with S.Eranian/HP Labs
- Aim: Ensure that his performance monitoring interface (*perfmon* – originally developed for Itanium) gets the Linux kernel for use on *A* platforms
- Our contributions:
  - Intense testing on Core 2 and Ita
  - Increased sophistication in *pfmo* comprehensive symbol resolutio
  - Graphical user interface: *gpfmor*
- *Also: Courses on architectur performance*

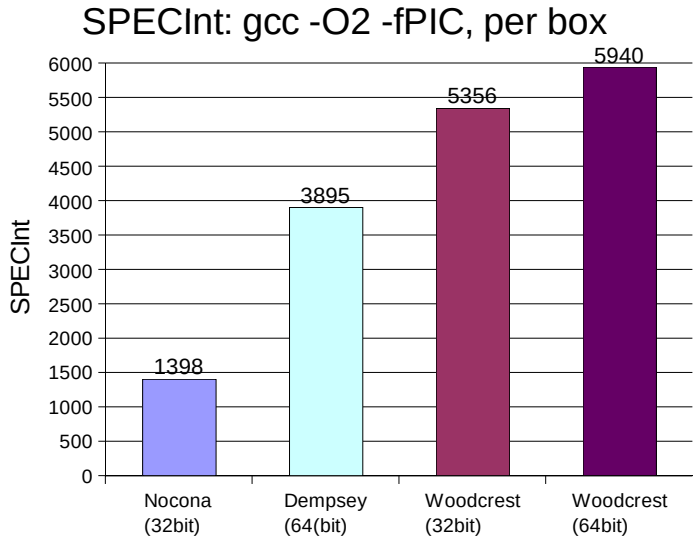


Compiler project

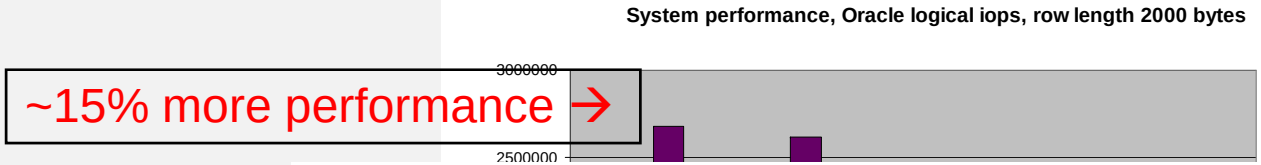
- Aim: Improved performance of jobs by influencing the back-end code generator
  - Based on our millions of lines of C++ source code
  - Also: Build test suites for performance and regression testing
- 2008:
  - Target further improvements in execution time
  - Special emphasis on additional options on top of O2
  - Expand to more complex benchmarks
    - Multithreading/TBB + SSE
  - Compiler expert from Intel visiting (Sept./Oct.)
- Project is active since the start of openlab I
  - With particular strength in in-order execution

Benchmarking

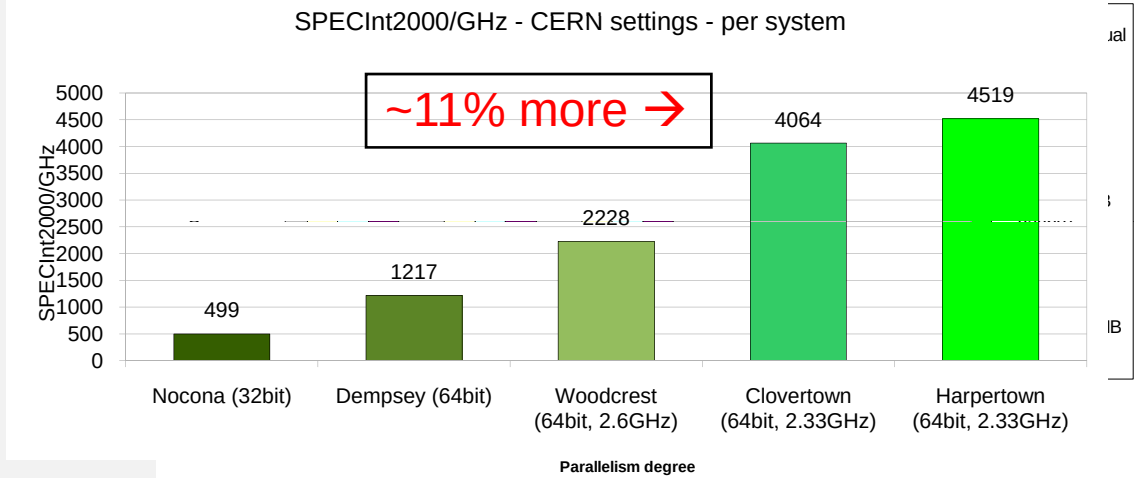
- Aim: Identify most relevant (and convenient) benchmark for acquisitions
  - Currently: Parallel SPEC2000Int (based on gcc –O2 –fpic –threads)
- Status: Works well, but more modern benchmark suite needed
- Candidates:
  - All of SPECInt2006
  - C++ part of SPEC2006
  - CERN-specific codes



Oracle performance benchmarks



- Test and validate performance of newer platform



- Oracle RDBMS performance comparison between (all dual-socket platforms):
  - E5140 (2.33Ghz, 4MB cache, “Woodcrest” - DC), current deployment platform for CERN’s Linux RACs
  - E5345 (2.33Ghz, 8MB cache, “Clovertown” - QC)
  - E5410 (2.33Ghz, 12MB cache, “Harpertown” - QC)

TOP500 runs

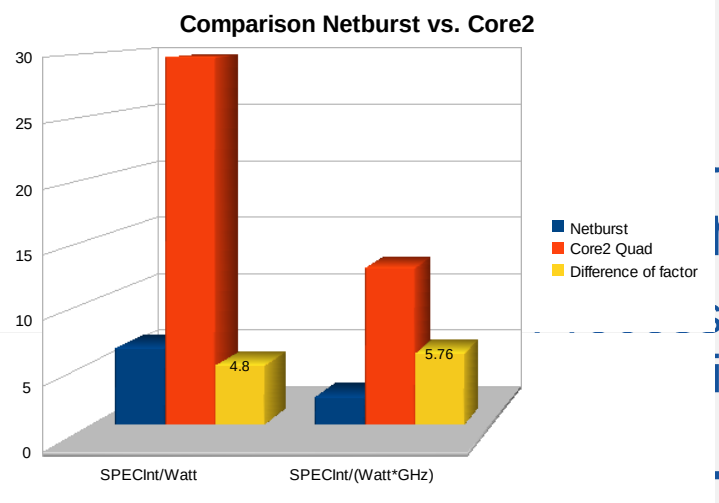
- Aim: Profit from the large acquisitions done for LHC to report the best possible number for TOP500
  - Also: Act as “burn-in” test for new systems
- Last Spring: 8.329 Tflops with 340 dual-core dual-socket servers
  - #115 in June 2007, #233 five months later (!)
- Now trying with more than 10’000 cores
  - 1300 quad-core DS servers
    - Ethernet interconnect

40-50 Tflops ?

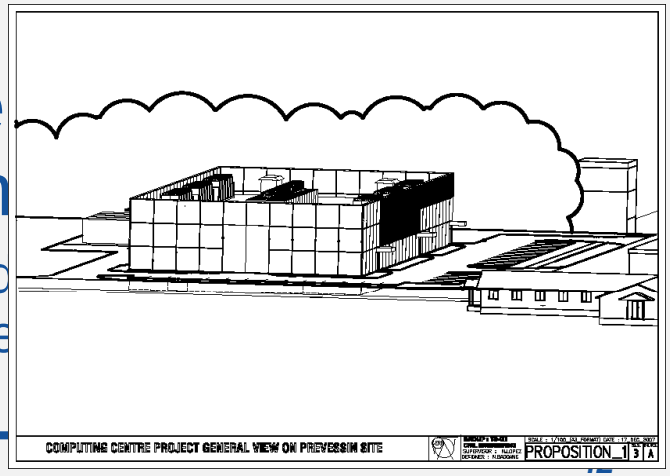
- Working closely with Sergey Shalnov (Intel)
  - Using his “hybrid” version of High Performance Linpack

Thermal control

- Working group to maximize power/thermal efficiency of current facility
  - Enclosing cold aisles for better separation of cold/hot air
  - Add “thermal penalties” in all acquisitions
- Project for new Computer Centre
  - Understand all relevant issues (before starting)
  - Aim at 5 + 5 MW
- Paper on power efficiency just completed



understand the
rver componen
sors (frequencies and
isks; I/O cards; Powe



New activities

- New processors
  - We are keenly interested in the move from multi-core to many-core!
  - Contacts with Intel developers
    - Benchmarks submitted; Scalability tests run on simulators
    - Encouraging scalability results with Geant4-derived benchmarks (CMS) and Trigger benchmark (ALICE)
  - On the software side, we participate in the review of the Ct language specifications
    - Initial specifications just received

CERN and Intel in openlab