CERN openlab
for DataGrid applications

1954-2004
50
CERN

openlab overview

CHEP2004

Sverre Jarp
openLab Technical Manager
IT Department
CERN

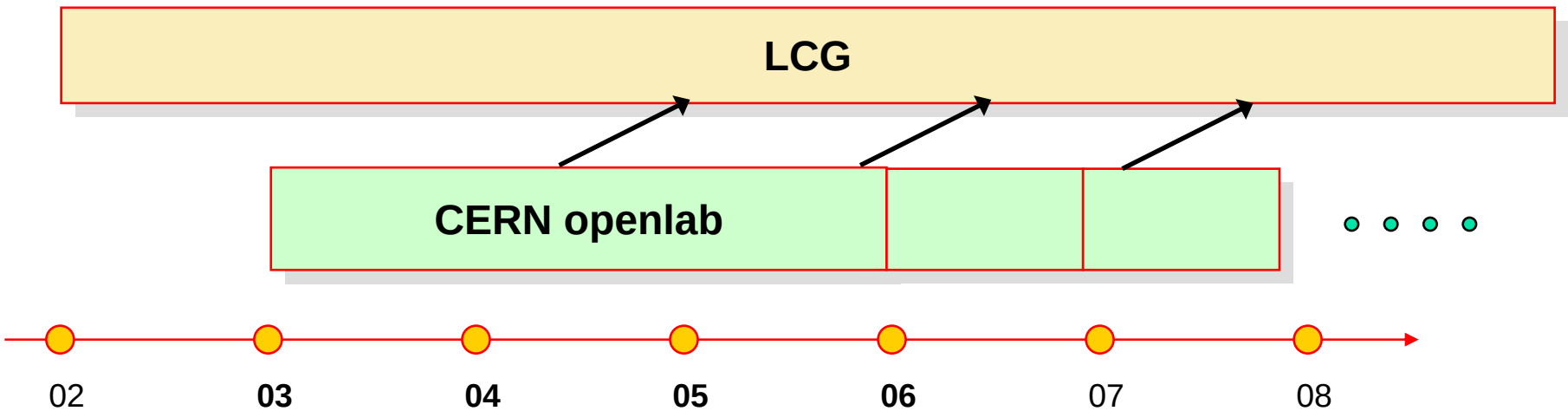
CERN

openlab for DataGrid applications

In partnership with

CERN openlab

- **Department's main R&D focus**
- **Framework for collaboration with industry**
- **Evaluation, integration, validation**
  - of cutting-edge technologies that can serve LCG
- **Initially a 3-year lifetime**
  - As of 1.1.2003
  - Later: Annual prolongations

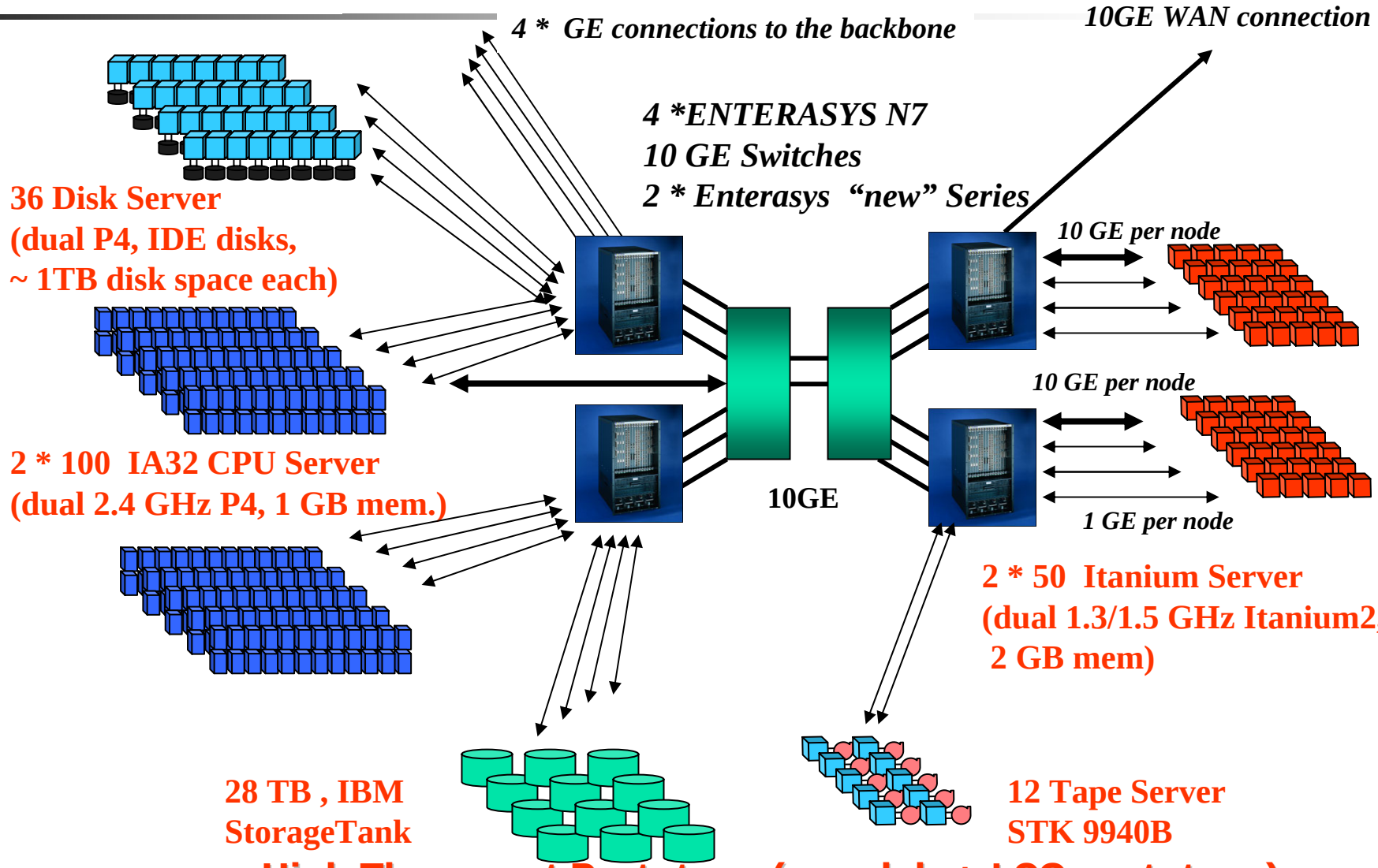


openlab participation

- **5 current partners**
  - **Enterasys:**
    - 10 GbE core routers
  - **HP:**
    - Integrity servers (103 * 2-ways, 2
    - Two post-doc positions
  - **IBM:**
    - Storage Tank file s w/metadata servers and data servers (curren
  - **Intel:**
    - Large bit Itanium processors & 10 Gbps NICs
    - 6 ystem w/PCI-Express
    - atabase software w/add-on's
    - wo post-doc positions
- **One contributor**
  - **Voltaire**
    - 96-way Infiniband switch and necessary HCAs

Because of the time limit of this talk, only certain highlights of our activities can be covered!

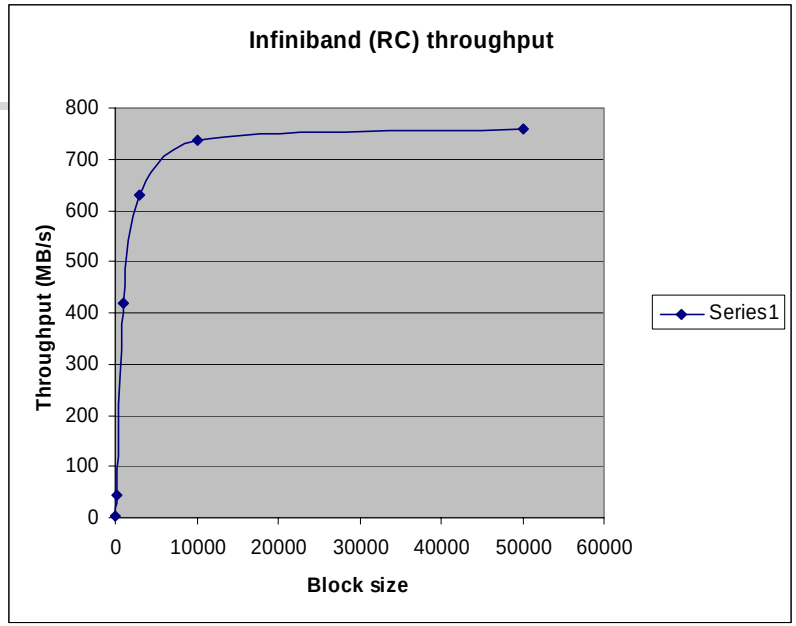
Full integration with the LCG testbed

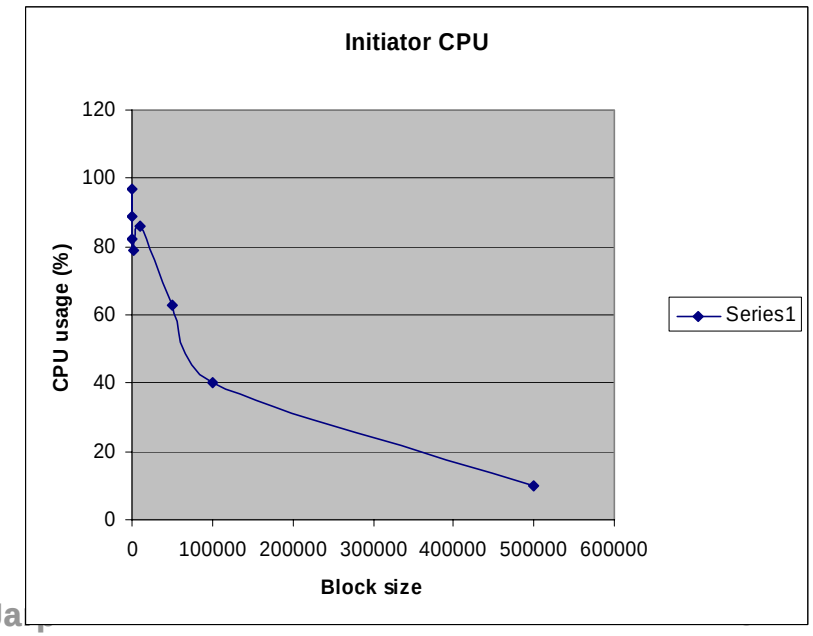


High Througput Prototype (openlab + LCG prototype)
(specific layout, October 2004)

Voltaire

- **Infiniband is a new technology with multiple strengths:**
  - **Low latency**
  - **High throughput**
  - **Low processor overhead**
  - **Connectivity to other protocols**
    - **In particular:**
      - **1 Gb Ethernet (10 GbE later)**
      - **Fibre-channel disks**

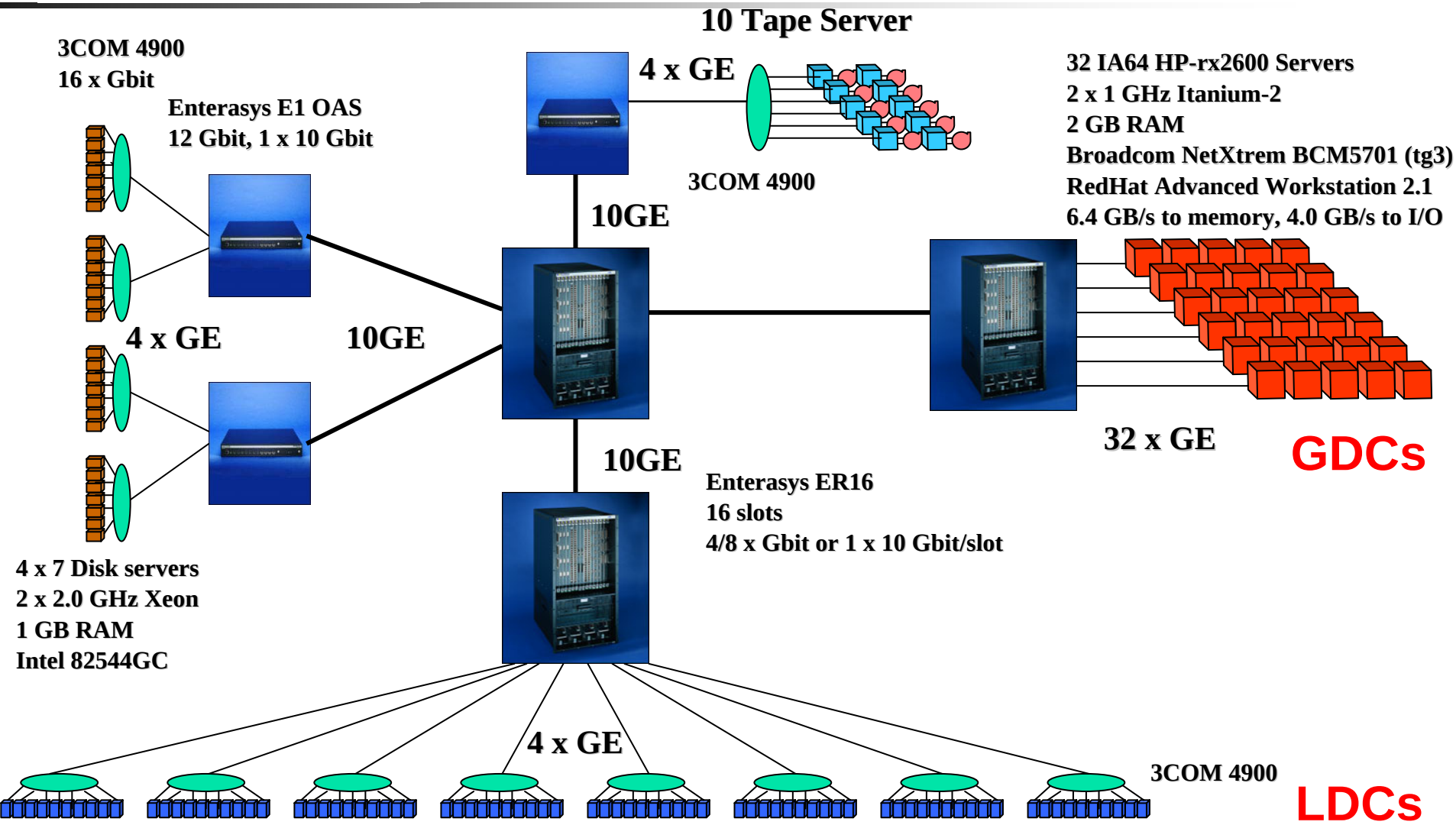




CERN openlab
for DataGrid applications

Enterasys

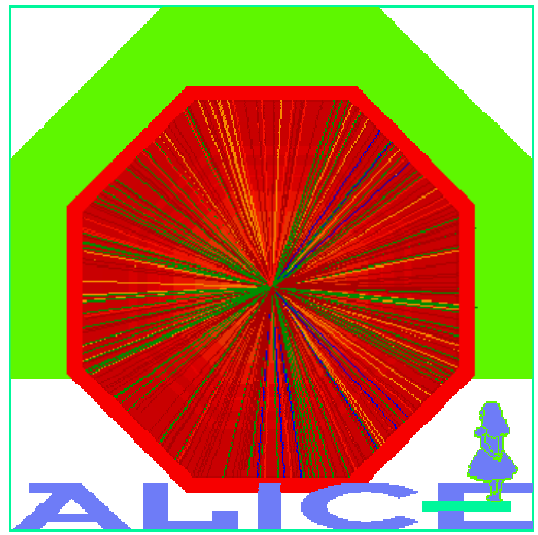
Alice Data Challenge HW Architecture



~ 80 CPU servers 2 x 2.4 GHz Xeon, 1 GB RAM, Intel 8254EM Gigabit in PCI-X 133 (Intel PRO/1000), CERN Linux 7.3.3

Achievements (as seen by Alice)

- ✘ **Sustained bandwidth to tape:**
  - **Peak 350 MB/s**
  - **Reached production-quality level only last week of testing**
  - **Sustained 280 MB/s over 1 day but with interventions [goal was 300]**

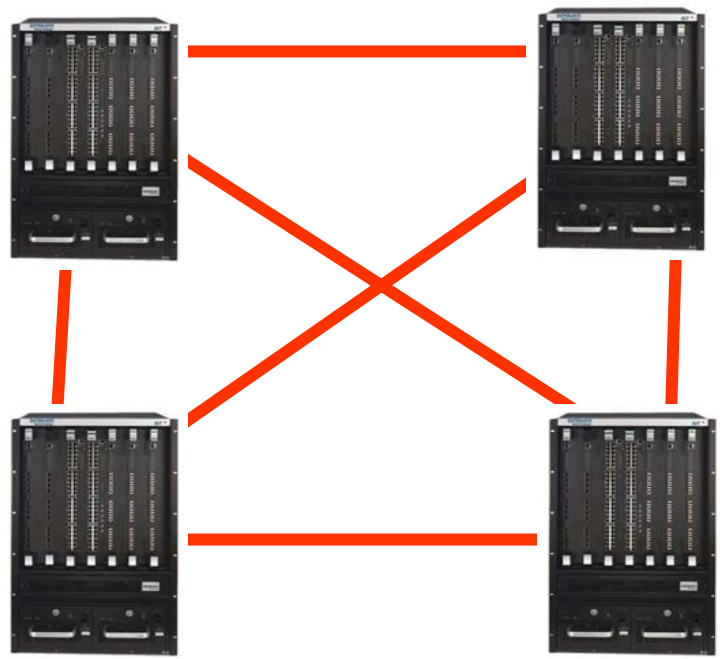


- ✓ **Itanium systems from openlab successfully integrated in the ADC**
  - ✓ **Very satisfactory stability**

Goal for ADC VI: 450 MB/s

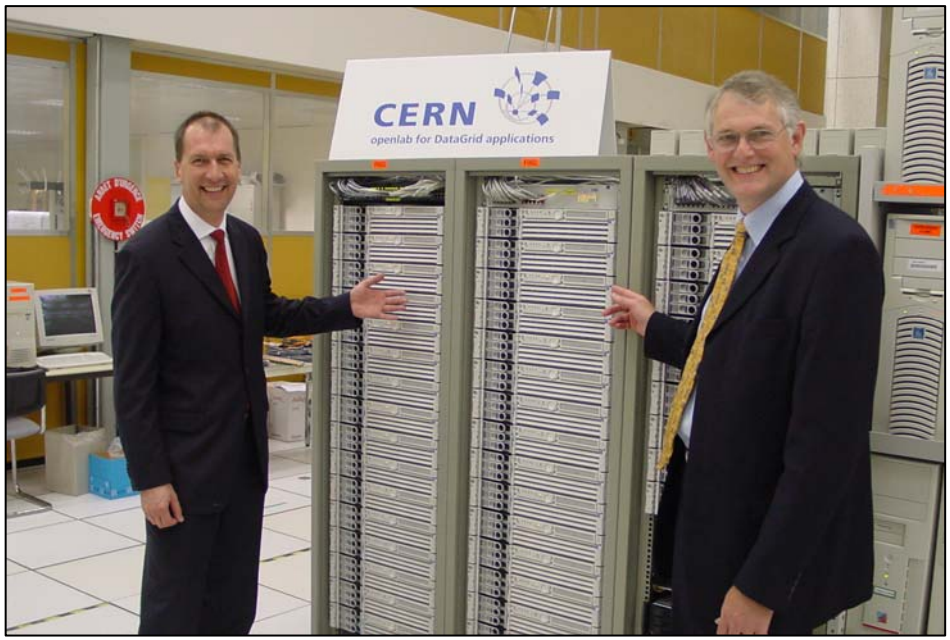
New generations

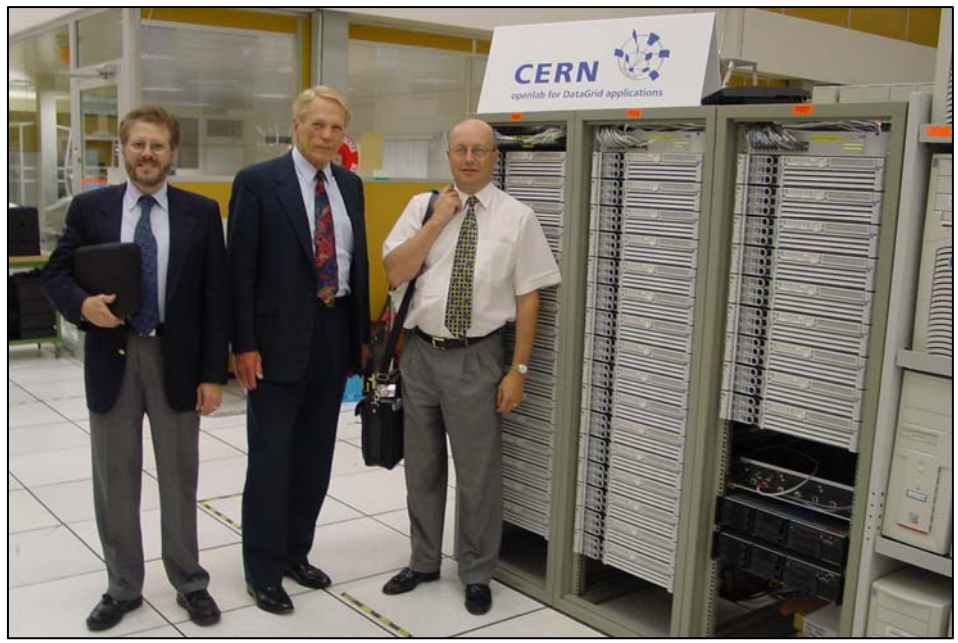
- In openlab we are currently using N7 “mid-range” routers
- Used successfully in tests with IBM StorageTank
- Recently, Enterasys has provided us with their most recent systems
  - See Enterasys’s presentations and documentation here at CHEP2004 for details



CERN openlab
for DataGrid applications

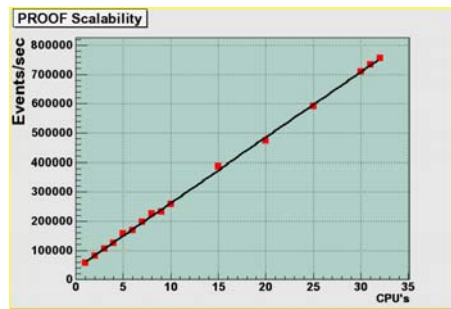
HP / Intel





Use of Itanium systems

- **Why Itanium?**
  - **Choice made in already in 2002**
    - **Neither Intel EM64T or AMD64 available**
  - **Pure 64-bit approach forces "complete conversion" to new mode**
    - **Ported: ROOT, CLHEP, GEANT4, ALIROOT, LCG2, etc.**
  - **HP Itanium servers**
    - **Have excellent stability and I/O capabilities**
  - **We use standard "Scientific Linux CERN 3"**
    - **Intel and GNU compilers**
  - **Very good performance monitoring tools**
    - **For both application and system performance**
  - **SPECint performance is adequate**
    - **~1300 SPECint**
    - **~1100 ROOTmarks with "stress"**
  - **Eagerly awaiting**
    - **Dual-core "Montecito" processors next year**
    - **Intel is also very "bullish" on evolution towards 2007**



CERN openlab
for DataGrid applications

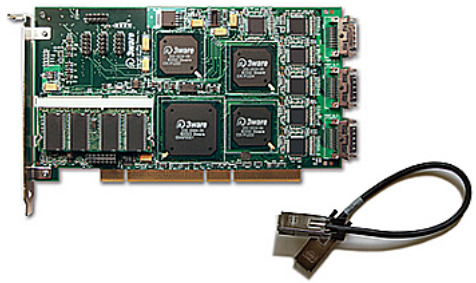
Gridification

- **A good success story:**
  - **Starting point: The software chosen for LCG (VDT + EDG) had been developed only with IA32 (and specific Red Hat versions) in mind**
    - **Consequence: Configure-files and make-files not prepared for multiple architectures. Source files not available in distributions (often not even locatable)**
  - **Stephen Eccles, Andreas Unterkircher worked for many months to complete the porting of LCG-2**
    - **Result: All major components now work on Itanium/Linux:**
      - **Worker Nodes, Compute Elements, Storage Elements, User Interface, etc.**
    - **Well tested inside the Test Grid**
    - **Code, available via Web-site, transferred to HP sites (Initially Puerto Rico and Bristol)**
    - **Changes given back to developers**
      - **VDT now built also for Itanium systems**
  - **Porting experience summarized in white paper (on the Web)**

From now on the Grid is heterogeneous!

Next generation disk servers

- **Based on state-of-the-art equipment:**
  - **4-way Itanium server (RX4640)**
    - **Two full-speed PCI-X slots**
      - **10 GbE and/or Infiniband**
    - **24 * S-ATA disks with 74 GB**
      - **WD740 “Raptor” @ 10k rpm**
      - **Burst speed of 100 MB/s**
    - **Two 3ware 9500 RAID controllers**
  - **In excess of 770 MB/s RAID-5 read speed**
    - **Single stream, sequential**
  - **Only 340 MB/s for write w/RAID 5**



Read speed with RFIO is 650 MB/s (multiple streams)

CERN openlab
for DataGrid applications

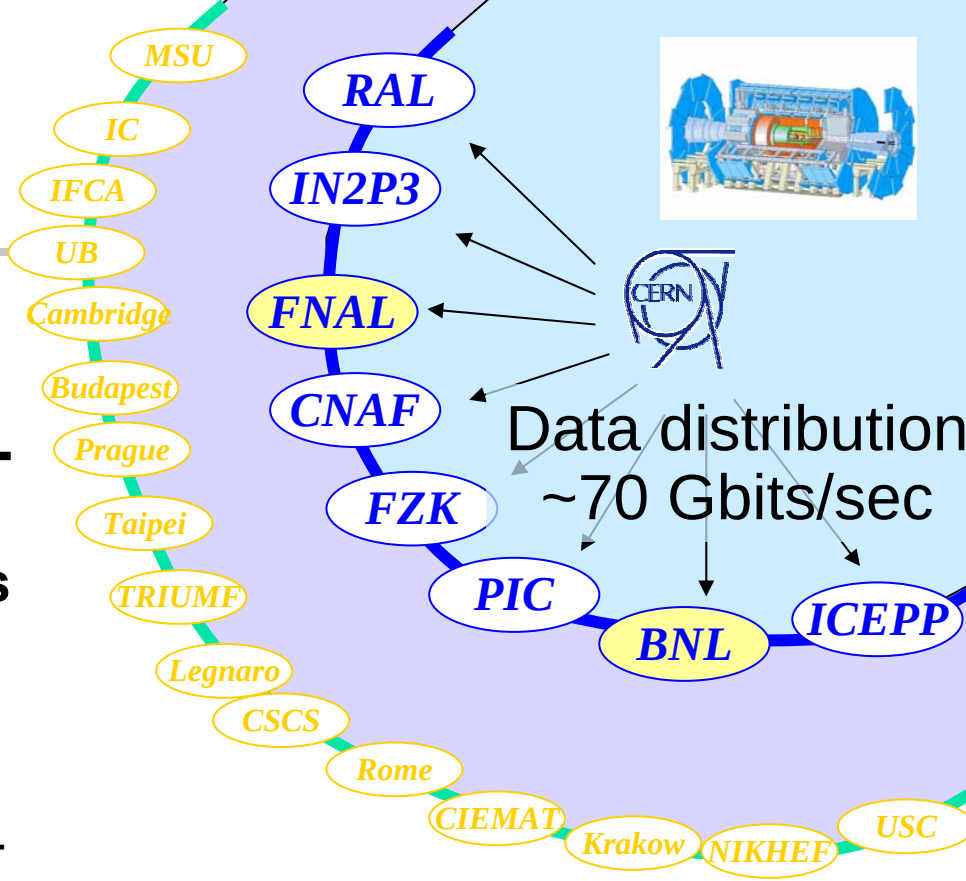
10 Gbps WAN tests
(between CERN and CalTech)

- **Initial breakthrough during Telecom-2003**
  - **with IPv4 (single/multiple) streams: 5.44 Gbps**
    - **Linux, Itanium-2 (RX 2600), Intel 10Gbps NIC**
  - **Also IPv6 (single/multiple) streams**
- **In June**
  - **Again IPv4, and single stream (Datatag/Openlab):**
  - **6.55 Gbps with Linux, Itanium-2 (RX4640), S2IO NIC**
- **In September:**
  - **Same conditions as before:**
  - **7.29 Gbps**

But SuNET with a much longer lightpath has just grabbed the record, even if they only reach 4.3 Gbps.
We will be back!

Data export to LCG Tier-1/-2

- **Tests (initially) between CERN and Fermilab + BNL +NIKHEF (and others)**
  - **Multiple HP Itanium servers with dual 1 Gb NICs**
    - **Disk to disk transfers via GridFTP**
    - **Each server:**
      - **100 MB/s IN + 100 MB/s OUT**
    - **Aggregation of multiple streams across 10 GbE link**
    - **Similar tuning as Internet2 tests**
  - **Goal: Achieve 5 Gbps sustained by Dec. 04**



"Service Data Challenge"
Stability is paramount – no longer just "raw" speed

CERN openlab
for DataGrid applications

IBM

CERN openlab
for DataGrid applications

StorageTank

- Collaboration directly between CERN and IBM Almaden
  - CERN expressed the desire to use Linux clients and iSCSI protocol right from the start
    - Not available in initial SAN FS product
- Random Access test
  - CMS pileup “look-alike” developed by Rainer Többicke
- Scenario:
  - 100 GB dataset, randomly accessed in ~50kB blocks
  - 1 – 100 2 GHz P4-class clients, running 3 – 10000 “jobs”
- Hardware
  - 4 IBM x335 metadata servers
  - 8 IBM 200i controllers, 336 SCSI disks
  - Added 2 IBM x345 servers as disk controllers after the test
- Results (after one week’s running)
  - Peak data rate: 484 MB/s (with 9855 simultaneous “jobs”)
  - After the test, special tuning, 10 servers, smaller number of clients:
    - 705 MB/s

CERN openlab
for DataGrid applications

StorageTank Plans

- **Plans for Phase 2**
  - **Alice Computing Data Challenge VI**
  - **Scenario**
    - **~50 input streams, ~50 output streams**
    - **Data staged from input through disk subsystem to tape**
    - **Goal is 450 MB/s "end-to-end" for 1 week (or more)**
  - **Sizing the system**
    - **Reasonable data block size, e.g. 256 KB**
    - **Note that 450 MB/s end-to-end means 450 in + 450 out!**
    - **Sufficient margin for dead time during tape mounts, tape file switching, etc.**
    - **1.5 GB/s disk bandwidth may be required**

CERN openlab
for DataGrid applications

Oracle

CERN openlab
for DataGrid applications

Oracle

- Activities to date
  - Two Oracle-funded research fellows since early 2004
  - Work focuses on prototyping new Oracle technologies, based on the needs of critical services, such as RLS (Replica Location Service)
  - First results obtained on
    - DataGuard
      - Maintains a hot backup of the catalogue which can be used while the main instance is not available for operating system scheduled interventions
    - Streams
      - Allows asynchronous database replication
  - 10g installed on Itanium cluster node

CERN openlab
for DataGrid applications

Future Activities

- **Significant activity expected in the area of Oracle Clusters**
  - **Application Server Clusters + DB Clusters (RAC)**
- **Further exploitation of RAC, which is a technology that services a number of purposes:**
  - **High availability - proven technology used in production at CERN to avoid single points of failure with transparent application failover**
  - **Consolidation – allows a smallish number (1 – 8?, 1 – 64?) of database services to be consolidated into a single management entity**
  - **Scalability – SAN-based infrastructure means CPU nodes and storage can be added to meet demand**
- **Many opportunities**
  - **Further development of Streams**
    - **Organize a "challenge" within DB group**
  - **Move to Oracle-managed storage (ASM) ?**
  - **Move to "the Oracle Grid" for IAS/DB infrastructure?**
  - **Exploit many other 10g features?**
    - **"Big file table-space" for Ultra-Large DBs (ULDBs)**

CERN openlab
for DataGrid applications

Conclusions

- **CERN openlab:**
  - **Solid collaboration with our industrial partners**
  - **Encouraging results in multiple domains**
  - **We believe partners are getting good "ROI"**
    - **But only they can really confirm it**
  - **No risk of running short of R&D**
    - **IT Technology is still moving at an incredible pace**
  - **Vital for LCG that the "right" pieces of technology are available for deployment**
    - **Performance, cost, resilience, etc.**
  - **Likely ingredients identified for LCG (so far): 64-bit programming, iSCSI, next generation I/O (10 Gb Ethernet, Infiniband, etc.)**

During summer: addition of 6 summer students