Computing in HEP

Andreas Gellrich

DESY IT Group - *Physics Computing*

DESY Summer Student Program 2005

Lectures in HEP, 11.08.2005

http://www-it.desy.de/

Program for Today

- Computing in HEP
- The DESY Computer Center
- Grid Computing
- Tour through the Computer Center (12.00h)

http://www-it.desy.de/

Three Questions

You *should* be able to find somewhat conclusive answers to the following three questions at the end of the talk:

Where does computing enter into HEP experiments?

What are the main components in a HEP Computer Center?

What is Grid Computing all about?

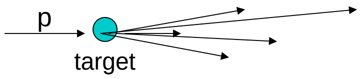
http://www-it.desy.de/

Computing in HEP

"From the physics motivation of an experiment to its computing requirements."

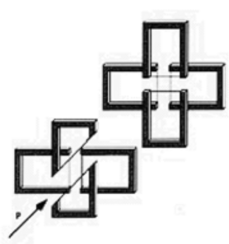
A Typical Scenario

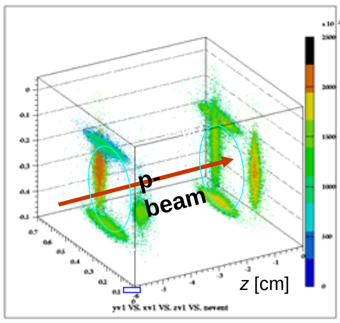
- Study the structure of matter.

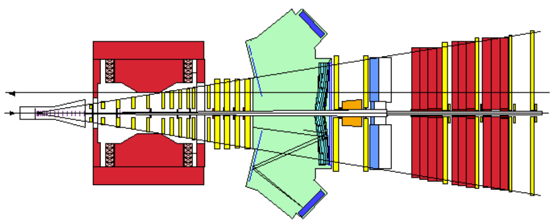


- An effect was observed or predicted.
- Interplay between experimentalists and theorists.
- Example: CP violation or the question of imbalance between matter and antimatter.
- The HERA-B experiment at DESY was planned to study this.
- In this race BaBar (SLAC) and BELLE (KEK) were the winners.
- We take HERA-B as an example because its technology prototypes the LHC experiments.

The Target

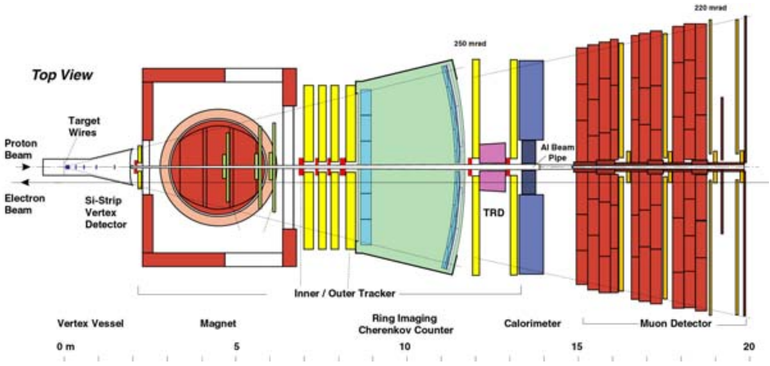






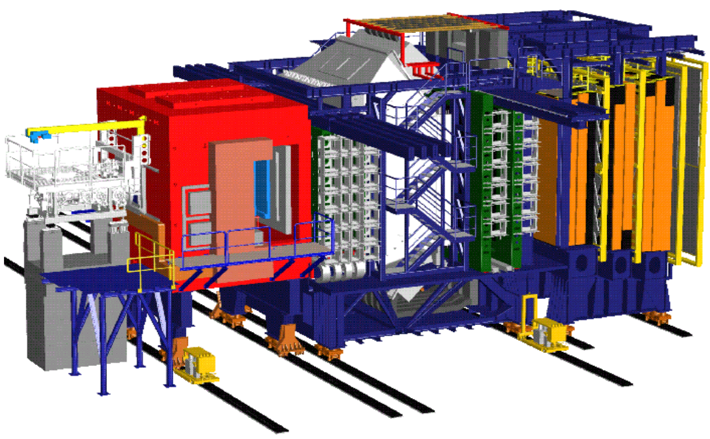
The Spectrometer

http://www-it.desy.de/



Yet Another View

http://www-it.desy.de/



http://www-it.desy.de/

The Numbers …

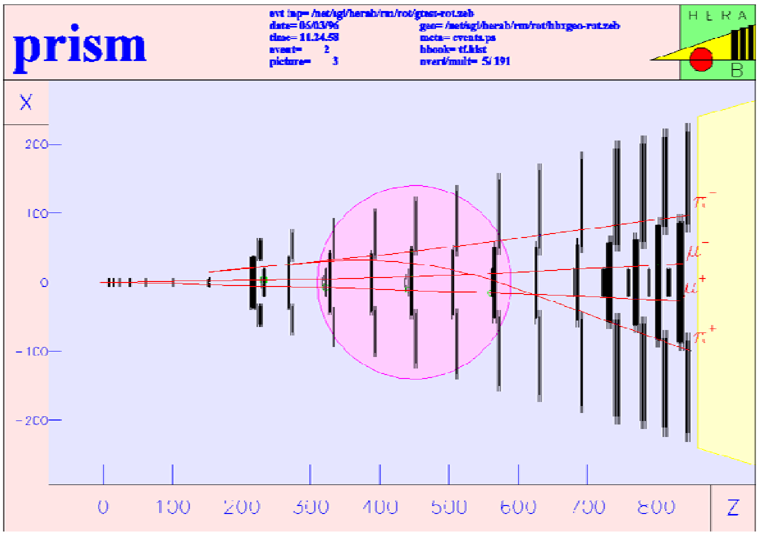
- We are interested in the CP violating decay channel: $B^0 \rightarrow J/\psi\ K^0_s$ and its antiparticles.
- We measure the asymmetry of particle and antiparticle decay: $A = (N_+ - N_-) / (N_+ + N_-)$
- Assuming an asymmetry of A = 0.1 at least 1000 reconstructed decays are needed to gain significant results.
- Only 1 of 1,000,000 pN interactions yields a bbbar quark pair ($\sigma_{bb} / \sigma_{pN} = 9.2 * 10^{-7}$)
- Only 1 of 250,000 b-quarks yields a $B^0 \rightarrow J/\psi\ K^0_s$ ($R = 4.2 * 10^{-6}$).

http://www-it.desy.de/

… The Numbers

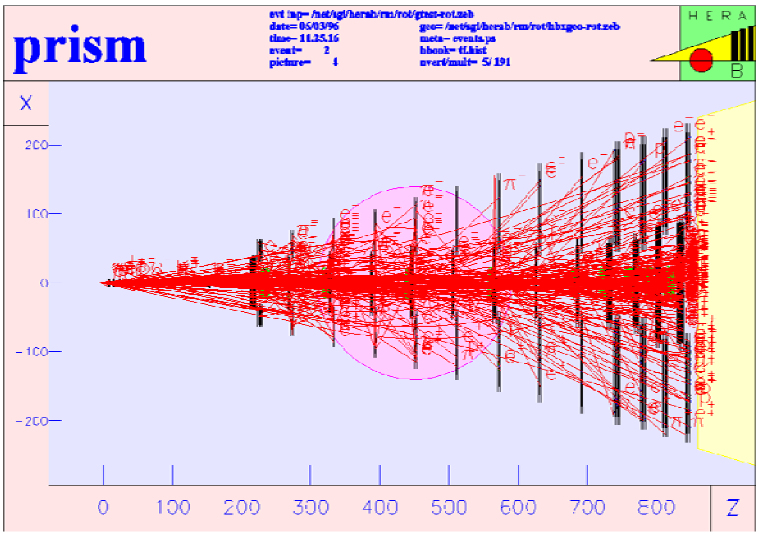
- We need to produce 40,000,000 pN *per sec* for one year (10^7 sec) to gain roughly 1000 such decays:
  $40 * 10^6$ / sec * 10^7 sec * $9.2 * 10^{-7} * 4.2 * 10^{-6} = 1545$
- This is 1 decay every 2 hours.
- Accounting for other physics topics finally leads to an interesting rate of physics of 1 Hz in a mess of 40 MHz of events.
- Of the 600,000 detector channels, on average 20% deliver data (12,000 channels).
- For each channel, 8 bytes (two words: value + index) are read out which deliver: 8 B * 12,000 = 96 kB per event.
- Full readout *would* produce: 40 MHz * 96 kB = 3.84 TB / sec

Events …



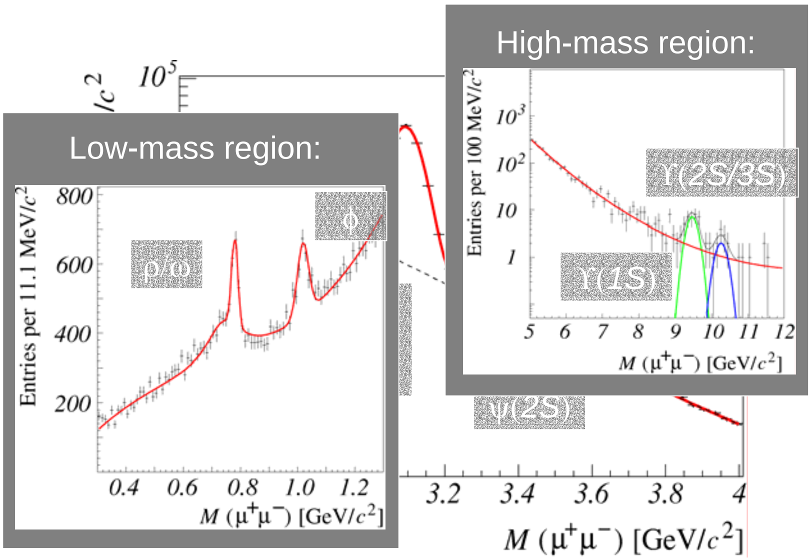
http://www-it.desy.de/

… Events



http://www-it.desy.de/

Results



The Challenge

- The detector must be designed, built, and operated.
- The data must be taken, analyzed, and understood.

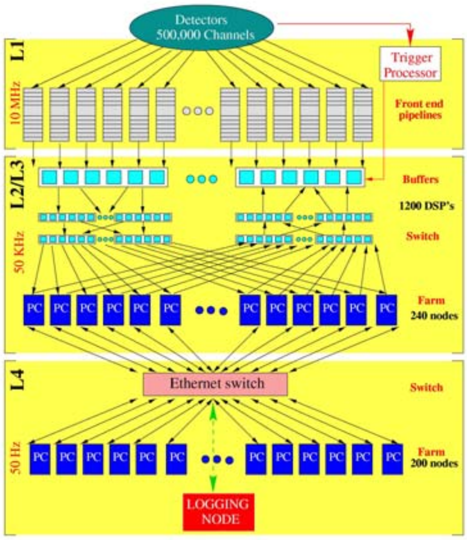
- Most of the events must be rejected as soon as possible, while keeping the few interesting ones.

- A sufficiently performant data acquisition system handles the high data rates.
- A selective trigger system filters the events.

- A clear offline computing strategy is needed which:
  - can be realized by modern computing technologies,
  - is scalable and flexible and fits the needs of hundreds of physicists, analyzing the data.

Multi-Level Trigger System

http://www-it.desy.de/



Rate	Time
10 MHz	dead-time free (μsec)
50 kHz	5 msec
50 Hz	4 sec

Level 4 Farm

http://www-it.desy.de/

- Online event reconstruction
- 50 Hz * 4 sec = 200 CPUs
- Commodity hardware
- Linux PCs
- Dual-PentiumIII / 550 MHz
- 50 Hz * 200 kB = 10 MB/sec
- 10 MB/sec = 100 TB/year

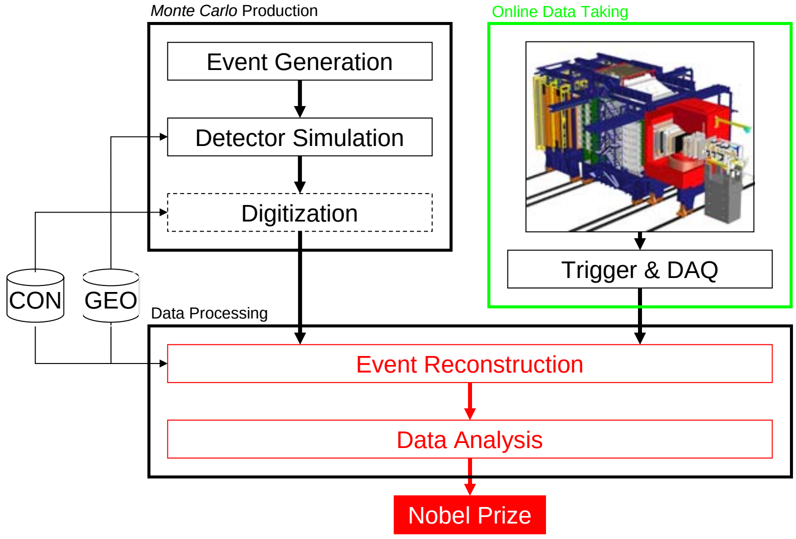
http://www-it.desy.de/

Offline Computing Strategies

- Offline computing starts after the selected events are stored.
- Offline computing includes:
  - Event generation (*Monte Carlo*)
  - Detector simulation
  - Data (re-)processing
  - Physics analysis
  - Data presentation
  - Software development
- The borderline between online and offline computing with respect to software developments is mainly gone (thanx to Linux).

http://www-it.desy.de/

The HEP Data Chain



http://www-it.desy.de/

Typical HEP Jobs

Monte Carlo:

- Event Generation: no I; small O; little CPU
- Detector Simulation: small I; large O; vast CPU
- (Digitization: usually part of simulation)

Event Reconstruction:

- (First processing: online or semi-online)
- Reprocessing: full I; full O; large CPU
- Selections: large I; large O; large CPU

Analysis:

- General: large I; small O; little CPU
- Performed by many users, many times!

http://www-it.desy.de/

Some Numbers …

DESY Experiment:

- Event Size: 200 kB
- Event Rate: 50 Hz
- => Event number: 500 M/year
- => Data Rate: 10 MB/sec
- => Data Volume: 100 TB/year (1 year = 10^7 sec)

Online Processing:

- Reconstruction Time: 4 sec/event
- Event Rate: 50 events/sec
- => Bandwidth: 10 MB/sec
- => Farm Nodes: 200 (200 = 50 * 4)

… Some Numbers

Offline Computing:

- Users: ~300
- Power Users: ~30

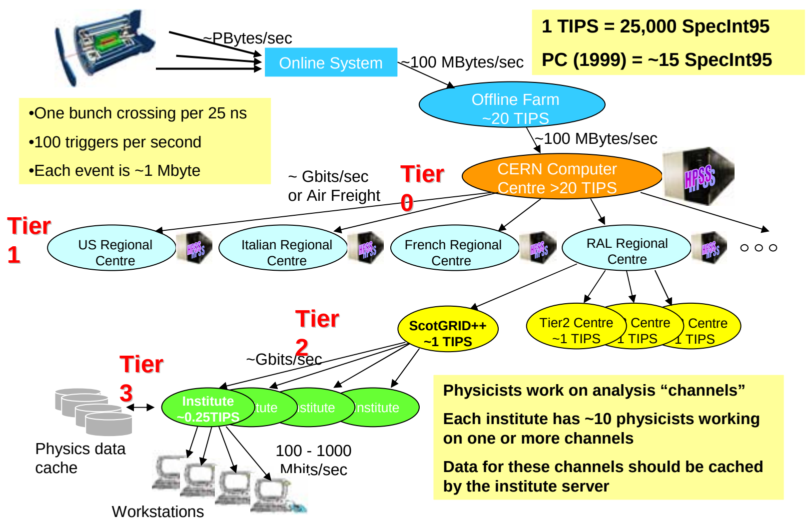
- Active Data Volume: ~10 TB
- File Server: ~10
- Offline Nodes: ~100

Typical Analysis Job:

- Event Selection: ~1 M events
- Data Selection: ~10 kB/event
- Processing Time: ~100 msec/event
- => Processing Rate: ~10 Hz
- => Bandwidth: 100 kB/sec

The LHC Computing Model

http://www-it.desy.de/



http://www-it.desy.de/

The DESY Computer Center

“Providing computing infrastructure and services to the DESY experiments and the users.”

http://www-it.desy.de/

Overview …

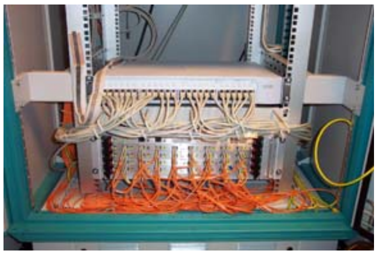
Support for the data taking of the experiments:
- Data management
- Resource provisioning (tape, disk, CPU)
- Workstation clusters (farms / clusters)

Special systems:
- High-performance cluster for lattice QCD (PC-based)
- custom cluster (apeNext, Zeuthen)

Software support for the users:
- CERN libraries
- Compilers (FORTRAN, C/C++, Java)
- ROOT
- GEANT4

http://www-it.desy.de/

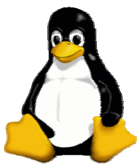
High Performance Cluster



http://www-it.desy.de/

… Overview

General infrastructure:



- Network (WAN / LAN)
- Desktops (Linux / Windows XP)
- Installation support (Linux x86, amd64 / Solaris)
- Mail (mail.desy.de / ntmail.desy.de)
- Web services (http://www.desy.de/)
- AFS home/group directories (/afs/desy.de/)
- Printing
- Large file store
- Directory services (NIS/ YP / LDAP) (ldap://ldap.desy.de/)
- Backup
- Security
- Licensing
- Telecommunications (+49 40 8998 – 0)

Numbers

http://www-it.desy.de/

• Sites:	Hamburg (Zeuthen)
• User:	~6500 (600)
• IP-addresses:	~13000
• Machines in CC:	~1000 (100)
• Unix Computer:	~700
• Linux Desktops: 	~1000
• Data on tape:	~500 TB / 3 STK robots
• Desaster recovery:	1 STK robot (building 3)
• Windows Accounts:	~2000
• Windows PCs:	~2000
• AFS:	O(1 TB)
• LAN:	10 Gbit / sec backbone
• WLAN:	11 Mbit / sec
• WAN:	1 Gbit / sec

Computing Trends

http://www-it.desy.de/

OO:
- Persistency: ROOT
- Analysis: ROOT, JAS
- Languages: C++, Java
- Scripting: Perl, Python

Commodity computing (PCs):
- No more mainframes; Intel x86 (and amd64) instead
- Linux all over the place
- PCs and Linux already in online (trigger) systems

Distributed resources:
- Few big accelerators / experiments
- O(10,000) CPUs; O(10 PB) data

http://www-it.desy.de/

Grid Computing

"Sharing resources within Virtual Organizations in a global world."

http://www-it.desy.de/

Introduction of the Grid

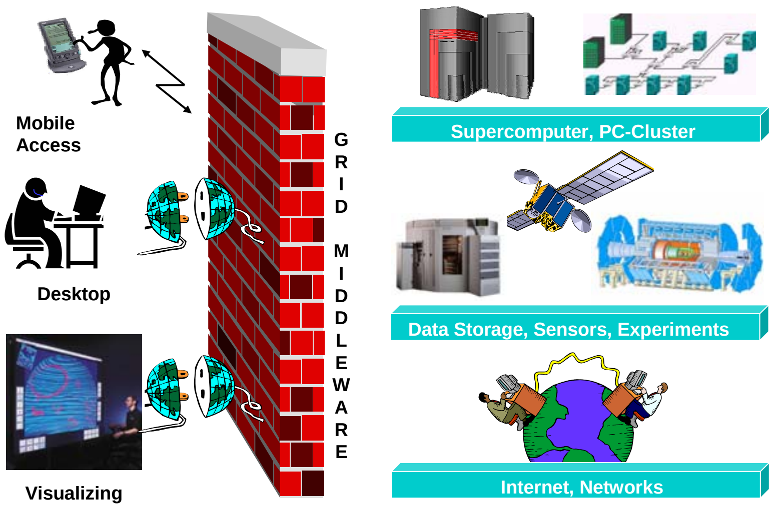
„We will probably see the spread of ´computer utilities‘, which, like present electric and telephone utilities, will service individual homes and offices across the country.“ Len Kleinrock (1969)

„A computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational capabilities.“ I. Foster, C. Kesselmann (1998)

„The sharing that we are concerned with is not primarily file exchange but rather direct access to computers, software, data, and other resources, as is required by a range of collaborative problem-solving and resource brokering strategies emerging in industry, science, and engineering. The sharing is, necessarily, highly controlled, with resources providers and consumers defining clearly and carefully just what is shared, who is allowed to share, and the conditions under which sharing occurs. A set of individuals and/or institutions defined by such sharing rules what we call a virtual organization.“ I. Foster, C. Kesselmann, S. Tuecke (2000)

http://www-it.desy.de/

The Grid Dream



http://www-it.desy.de/

The *Fuzz* about Grids

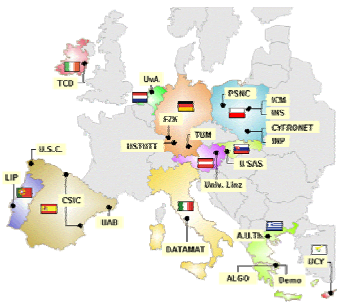
DataGRID
GGF
ILDG
crossgrid
egee Enabling Grids for E-sciencE
LCG
TERAGRID
NORDUGRID
GriPhyN
GridPP UK Particle Physics
INFN GRID
DataTAG
PPDG
iVDgL

HEP Grids Worldwide

http://www-it.desy.de/

HEP Grids in Europe

http://www-it.desy.de/



Grid Types

http://www-it.desy.de/

Data Grids:

- Provisioning of transparent access to data which can be physically distributed within *Virtual Organizations* (VO)

Computational Grids:

- allow for large-scale compute resource sharing within Virtual Organizations (VO)

Information Grids:

- Provisioning of information and data exchange, using well defined standards and web services

The Grid Definition

http://www-it.desy.de/

I. Foster: *What is the Grid? A Three Point Checklist* (2002)

"A Grid is a system that:

coordinates resources which are not subject to centralized controls …

integration and coordination of resources and users of different domains vs. local management systems (batch systems)

… using standard, open, general-purpose protocols and interfaces …

standard and open multi-purpose protocols vs. application specific system

… to deliver nontrivial qualities of services."

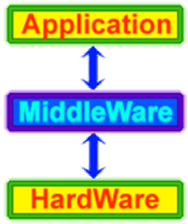
coordinated use of resources vs. uncoordinated approach (world wide web)

Grid *Middleware*

http://www-it.desy.de/

Globus:
- Toolkit
- Argonne, U Chicago

EDG (EU DataGrid):
- Project to develop Grid middleware
- Uses parts of Globus
- Funded for 3 years (01.04. 2001 - 31.03.2004)



LCG (LHC Computing Grid):
- Grid infrastructure for LHC production
- Based on stable EDG versions plus VDT etc.
- LCG-2 for Data Challenges

EGEE (Enabling Grids for E-Science in Europe)
- Started 01.04.2004 for 2 + 2 years

Grid Ingredients

http://www-it.desy.de/

Authentication:
- Use method to guarantee authenticated access only

Authorization:
- Users must be registered in a Virtual Organization (VO)

Information Service:
- Provide a system which keeps track of the available resources

Resource Management:
- Manage and exploit the available computing resources

Data Management:
- Manage and exploit the data

Certification

http://www-it.desy.de/

- Authorization and authentication are essential parts of Grids
- By means of a *certificate (X.509 standard)*
- A certificate is an encrypted electronic document, digitally signed by a *Certification Authority* (CA)
- A Certificate Revocation List (CRL) published by the CA
- Users and service hosts must be certified
- The *Globus Security Infrastructure* (GSI) is part of the *Globus Toolkit* GSI is based on the *openSSL Public Key Infrastructure* (PKI)
- In Germany FZ Karlsruhe (GridKa) is the national CA.
- Example: /O=GermanGrid/OU=DESY/CN=Andreas Gellrich

Virtual Organization

http://www-it.desy.de/

- A *Virtual Organization* (VO) is a *dynamic collection of individuals, institutions, and resources* which is defined by certain sharing rules.
- Technically, a user is represented by his/her certificate.
- The collection of authorized users is defined on every machine in /etc/grid-security/grid-mapfile .
- This file is regularly updated from a central server.
- The server holds a list of all users belonging to a collection.
- It is this collection we call a VO!
- The VO a user belongs to is *not* part of the certificate.
- A VO is defined in a central list, e.g. a LDAP tree.
- DESY maintains VOs for experiments and groups.

Grid Infrastructure …

http://www-it.desy.de/

Authentication/Authorization:
- Grid Security Infrastructure (GSI) based on PKI (openSSL)
- Globus Gatekeeper, Proxy renewal service
- Server to support VOs

Grid Information Service:
- Grid Resource Information Service (GRIS)
- Grid Information Index Service (GIIS)

Resource Management:
- Resource Broker, Job Manager, Job Submission, Batch System (PBS), Logging and Bookkeeping

Data Management: (Replica Location Services)
- Storage Elements with mass storage capabilities
- Catalogue Services (replicas, meta data)

… Grid Infrastructure

http://www-it.desy.de/

Hardware:
- => Mapping of services to logical and physical nodes.

The basic nodes are:
- User Interface (UI)
- Computing Element (CE)
- Worker Node (WN)
- Resource Broker (RB)
- Storage Element (SE)
- Catalog Service (CAT)
- Information Service (BDII, GRIS, GIIS)

http://www-it.desy.de/

Grid Set-up

Per site:

- User Interface (UI) to submit jobs
- Computing Element (CE) to run jobs
- Worker Nodes (WN) to do the work
- Storage Element (SE) to provide data
- Grid Information Index Service (GIIS)

Per VO:

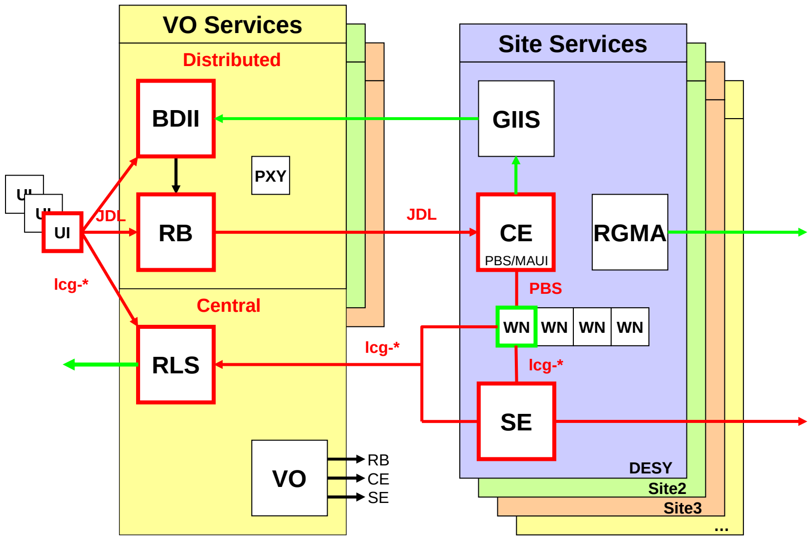
- Resource Broker (RB)
- Replica Catalog (LRC)
- Meta Data Catalog (MDC)
- VO-server (VO) / VO Membership Service (VOMS)

General:

- Certification Authority (CA)
- Network services

http://www-it.desy.de/

… DESY Grid Infrastructure …



Grid @ DESY

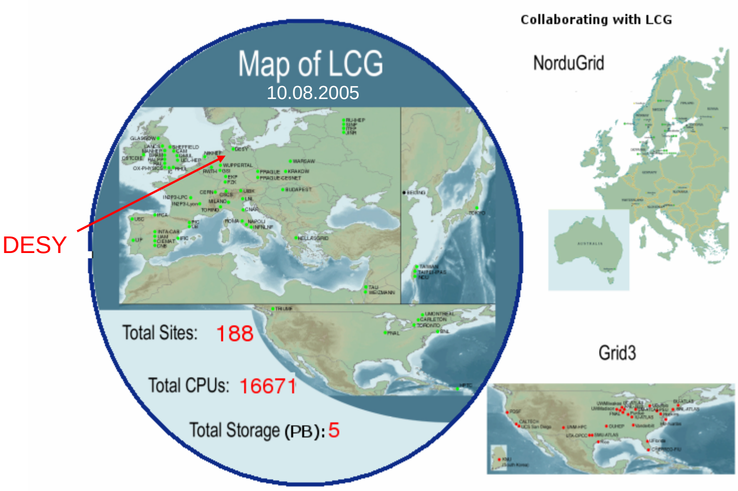
http://www-it.desy.de/

- With the HERA-II luminosity upgrade, the demand for MC production rapidly increased while the outside collaborators moved there computing resources into LCG
- H1 and ZEUS maintain VOs and *use* the Grid for MC production
- The International Linear Collider (ILC) community uses the Grid
- The LQCD group develops a Data Grid to exchange data
- DESY is about to become an LCG Tier-2 site
- EGEE and D-GRID
- dCache is a DESY / FNAL development

Enabling Grids for E-sciencE

GOC Grid Monitoring

http://www-it.desy.de/



DESY Grid Infrastructure …

http://www-it.desy.de/

- VOs *hosted* at DESY:
  - Global: '*hone*', '*ilc*', '*zeus*' (registration via LCG registrar system)
  - Regional: '*calice*', '*dcms*', '*ildg*'
  - Local: '*baikal*', '*desy*', '*herab*', '*hermes*', '*icecube*'
- VOs *supported* at DESY:
  - Global: ('*atlas*'), '*cms*', '*dteam*'
  - Regional: '*dech*'
- H1 Experiment at HERA ('*hone*')
  - DESY, U Dortmund, RAL, RAL-PP, Bham
- ILC Community ('*ilc*', 'calice')
  - DESY, RHUL, QMUL, IC, …
- ZEUS Experiment at HERA ('*zeus*')
  - DESY, U Dortmund, INFN, UKI, UAM, U Toronto, Cracow, U Wisconsin, U Weizmann

… DESY Grid Infrastructure …

http://www-it.desy.de/

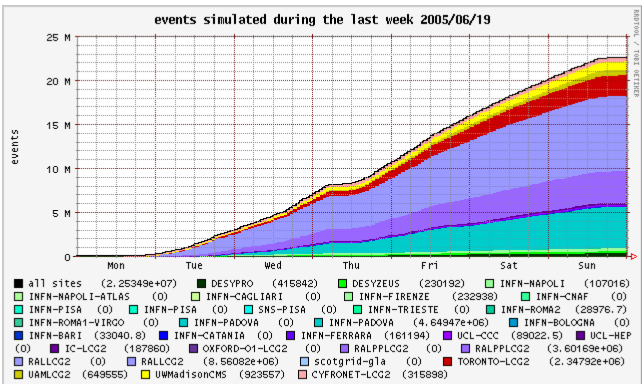
- SL 3.04
- Quattor (OS for all nodes; complete installation for WNs)
- Yaim (for all service nodes)
- LCG-2_4_0
- Central VO Services:
  - VO server (LDAP) [`grid-vo.desy.de`]
  - Replica Location Services (RLS) [`grid-cat.desy.de`]
- Distributed VO Services:
  - Resource Broker (RB) [`grid-rb.desy.de`]
  - Information Index (BDII) [`grid-bdii.desy.de`]
  - Proxy (PXY) [`grid-pxy.desy.de`]
- Site Services: [`DESY-HH`]
  - GIIS: `ldap://grid-giis.desy.de:2170/mds-vo-name=DESY-HH,o=grid`
  - CE: 24 WNs (48 CPUs, XEON 3.06 GHZ) (IT) [`grid-ce.desy.de`]
  - CE: 17 WNs (34 CPUs, XEON 1GHz) (ZEUS) [`zeus-ce.desy.de`]
  - SE: dCache-based with access to the entire DESY data space
  - Storage: disk 5 TB (+ 15 TB this summer), tape 0.5 PB media (2 PB capacity)

… DESY Grid Infrastructure

- SuperMicro Superserver
- rack-mounted 1U servers
- dual Intel P4 XEON 2.8 / 3.06 GHz
- 2 GB ECC DDRAM
- GigaBit Ethernet
- 80 GB (E)IDE system disk
- 200 GB (E)IDE data disk
- 10 Gbit/s DESY back-bone
- 1 Gbit/s WAN (G-WIN)
- 30 Worker Nodes (WN)
- 10 core service nodes
- 3 WNs ZEUS
- 2 WNS U Hamburg
- 17 WNs ZEUS

ZEUS @ Grid

- > 300 M events have been produced on the Grid since Nov 2004
- mainly produced outside DESY
- 32 sites (incl. Wisconsin and Toronto)



http://www-it.desy.de/

Grid Job Example …

Job Submission:

- Job delivers hostname and date of worker node

- Grid Environment
- Job script/binary which will be executed
- Job description by JDL
- Job submission
- Job status request
- Job output retrieval

http://www-it.desy.de/

… Authentication …

Start.

grid-ui> grid-proxy-init
Your identity: /O=GermanGrid/OU=DESY/CN=Andreas Gellrich
Enter GRID pass phrase for this identity:
Creating proxy ....
Done
Your proxy is valid until Thu Oct 23 01:52:00 2003

grid-ui> grid-proxy-info -all
subject : /O=GermanGrid/OU=DESY/CN=Andreas Gellrich/CN=proxy
issuer : /O=GermanGrid/OU=DESY/CN=Andreas Gellrich
type : full
strength : 512 bits
timeleft : 11:59:48

… Job Description …

http://www-it.desy.de/

grid-ui> less script.sh

```
#! /usr/bin/zsh
host=`/bin/hostname`
date=`/bin/date`
echo "$host | $date“
```

grid-ui> less hostname.jdl

```
Executable      = "script.sh";
StdOutput       = “stdout";
StdError        = “stderr";
InputSandbox    = {"script.sh"};
OutputSandbox = {“stdout","stderr"};
```


… Job Matching …

http://www-it.desy.de/

grid-ui> edg-job-list-match hostname.jdl

```
Connecting to host grid-rb.desy.de, port 7772

*****************************************************************************
COMPUTING ELEMENT IDs LIST
The following CE(s) matching your job requirements have been found:
- grid-ce.desy.de:2119/jobmanager-pbs-short
- grid-ce.desy.de:2119/jobmanager-pbs-long
- grid-ce.desy.de:2119/jobmanager-pbs-medium
- grid-ce.desy.de:2119/jobmanager-pbs-infinite

*****************************************************************************
```

… Job Submission …

http://www-it.desy.de/

grid-ui> edg-job-submit hostname.jdl
Connecting to host grid-rb.desy.de, port 7772
Logging to host grid-rb.desy.de, port 9002

****** edg-job-submit Success ***
The job has been successfully submitted to the Resource Broker.
Use edg-job-status command to check job current status. Your job identifier (edg_jobId) is:

https://grid-rb.desy.de:7846/131.169.223.35/134721208077529?grid-rb.desyde:7771
****

… Job Status …

http://www-it.desy.de/

grid-ui> edg-job-status
'https://grid-rb.desy.de:7846/131.169.223.35/134721208077529?grid-rb.desy.de:7771'

Retrieving Information from LB server https://grid-rb.desy.de:7846
Please wait: this operation could take some seconds.

****
BOOKKEEPING INFORMATION:
Printing status info for the Job : https://grid-rb.desy.de:7846/131.169.223.35/134721208077529?grid-rb.desy.de:7771

To be continued …

… Job Status …

… continued:

- Status = Initial
- Status = Scheduled
- Status = Done
- Status = OutputReady

… Job History …

grid-ui> edg-job-get-logging-info 'https://grid-rb.desy.de:7846/131.169.223.35/093011136900851?grid-rb.desy.de:7771'

Retrieving Information from LB server https://grid-rb.desy.de:7846
Please wait: this operation could take some seconds.

****
LOGGING INFORMATION:

Printing info for the Job :
https://grid-rb.desy.de:7846/131.169.223.35/093011136900851?grid-rb.desy.de:7771

To be continued …

http://www-it.desy.de/

… Job History …

http://www-it.desy.de/

… continued:

- Event Type = JobAccept
- Event Type = Job Transfer
- Event Type = JobMatch
- Event Type = JobScheduled
- Event Type = JobRun
- Event Type = JobDone

… Job Output …

http://www-it.desy.de/

grid-ui> edg-job-get-output
'https://grid-rb.desy.de:7846/131.169.223.35/134721208077529?grid-rb.desy.de:7771'

****
JOB GET OUTPUT OUTCOME
Output sandbox files for the job:
-https://grid-rb.desy.de:7846/131.169.223.35/134721208077529?grid-rb.desy.de:7771
have been successfully retrieved and stored in the directory:
/tmp/134721208077529
****

… Job Output

grid-ui> ls -l /tmp/134721208077529
total 4
-rw-r--r-- 1 gellrich it 0 Oct 23 15:49 hostname.err
-rw-r--r-- 1 gellrich it 39 Oct 23 15:49 hostname.out

grid-ui> less /tmp/134721208077529/hostname.out
grid101: Thu Jul 23 15:47:41 MEST 2005

Done!

Grid Data Management

- So far we have simply computed something …
- The RB has picked a CE with computing resources

- A typical job reads/writes data
- Data files shall not be transferred with the job

- Scenario:
  - The data files are always accessed from a (nearby) SE.
  - Data files are registered in a *Replica Catalogue* and can be found via a *Meta Data* information (*Logical File Name*).
  - Data management services are used to replicate the needed data files from elsewhere to the appropriate SE.

http://www-it.desy.de/

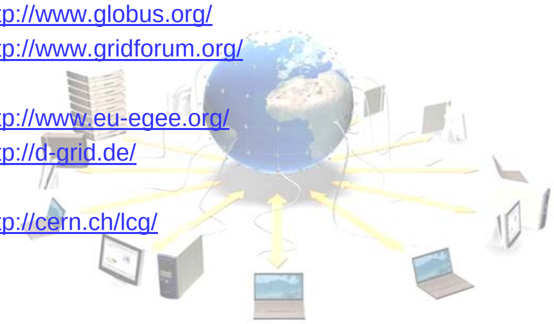
Grid Web Links

Grid:

- http://www.gridcafe.org/
- http://www.globus.org/
- http://www.gridforum.org/

- http://www.eu-egee.org/
- http://d-grid.de/

- http://cern.ch/lcg/



DESY:

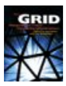
- ✓ http://grid.desy.de/
- ✓ http://www.dcache.org/
- ✓ http://www-zeus.desy.de/grid/

http://www-it.desy.de/

Grid Literature

Books:

- Foster, C. Kesselmann: *The Grid: Blueprint for a New Computing Infrastructure*, Morgan Kaufmann Publisher Inc. (1999)
- F. Berman, G. Fox, T. Hey: *Grid Computing: Making The Global Infrastructure a Reality*, John Wiley & Sons (2003)



Articles:

- I. Foster, C. Kesselmann, S. Tuecke: *The Anatomy of the Grid* (2000)
- I. Foster, C. Kesselmann, J.M. Nick, S. Tuecke: *The Physiology of the Grid* (2002)
- I. Foster: *What is the Grid? A Three Point Checklist* (2002)

Grid Summary

http://www-it.desy.de/

- http://grid.desy.de/
- The Grid has developed from a smart idea to reality.
- Grid infrastructure will be standard in (e)-science in the future.
- LHC can *not* live w/o Grids.
- DESY: LCG, ILDG; EGEE, D-GRID
- Experiments exploit the Grid for Monte Carlo production.

The Three Questions

http://www-it.desy.de/

What are your questions to the three questions I denoted at the beginning of my talk:

Where does computing enter into HEP experiments?

What are the main components in a HEP Computer Center?

What is Grid Computing all about?

The summary is up to you.