International Conference on Computing in High Energy (and Nuclear) Physics

18-month period (Europe, USA, World), history:
September 2001: Beijing / P.R. China
February 2000: Padova / Italy
October 1998: Chicago / USA
April 1997: Berlin / Germany
September 1995: Rio de Janeiro / Brazil
April 1994: San Francisco / USA
September 1992: Annecy / France
March 1991: Tsukuba / Japan
April 1990: Santa Fe / USA
April 1989: Oxford / Great Britain
Introduction

CHEP’01
3.-7. September 2001
Friendship Hotel
Beijing / P.R. China

Organized by IHEP of the Tsinghua University Beijing

Plenary Program

Monday (Future):

Status of Preparation for LHC Computing (M. Delfino / CERN)
Geant4 toolkit: status & utilization (J. Apostolakis / CERN)
Software Frameworks for HEP Data Analysis (V. Innocente / CERN)
The LHC Experiments’ Joint Controls Project (JCOP) (W. Salter / CERN)

Tuesday (Running Systems):

The BARBAR Database: Challenges, Trends and Projections (J. Becla / SLAC)
New data acquisition and data analysis system for the Belle experiment (R. Itoh / KEK)
The CDF Computing and Analysis System: First Experience (S. Lammel / FNAL)
The D0 Data Handling System (V. White / FNAL)
Plenary Program (cont.)

Wednesday (Grid):

Grid Technologies & Applications: Architecture & Achievements) (I. Foster / ANL)
Review of the EU-DataGrid project (F. Gagliardi / CERN)
US Grid Projects: PPDG nad iVDGL (R. P. Mount / SLAC)
The Asia Pacific Grid (ApGRID) Project and HEP Application (S. Sekiguchi / AIST)
Grid Computing with Grid Engine Juxta, and Jini (S. See / Sun Microsystems)

Friday (Grid/Networks/Clusters/other fields):

Grid Computing at IBM (G. Wang / IBM)
Present and Future Networks for HEPN (H. Newman / Caltec)
From HEP Computing to Bio-Medical Research and (M. G. Pia / INFN)
Large Scale Cluster Computing Workshop (Alan Silverman / CERN)
Virtual network computing environment - challenge to high performance computing GAO Wen (China academy of sciences)
Conference Summaries (Session convenors)

Parallel Tracks

1. Commodity Hardware & Software (10)
2. Control Systems (9)
3. Data Analysis & Visualization (28)
4. Data Handling & Storage (27)
5. Simulation (8)
6. Information Systems & Multimedia (5)
7. Local & Wide Area Networking (10)
8. Software Methodologies & Tools (26)
9. Triggering & Data Acquisition (28)
10. Grid Computing (26)

Total (177)

Large number of talks especially from US speakers where canceled or held by other authors
**DESY Contributions**

28 participants, 19 contributions, 14 talks, 5 posters

**DESY Hamburg:**

IT: C. Beyer (Printing), P. Fuhrmann (dCache), A. Gellrich (Linux)
K. Woller (Linux), K. Ohrenberg, B. Hellwig

IPP: J. Bürger (EDMS), L. Hagge, J. Kreuzkamp (AMS)

HI: U. Berthon (OO), G. Eckerlin (Control), R. Gerhards (OO Framework)
F. Niebergall

ZEUS: U. Behrens (DAQ), U. Fricke (Database), R. Mankel,
K. Wrona (Farms)

HERA-B: V. Amaral (Database)

TESLA: H. von der Schmitt (Tesla Computing)

**DESY Zeuthen:**

H. Leich (PITZ-Interlock), K.-H. Sulanke, H. Vogt,
P. Wegner (APEmille), T. Naumann (H1-Alignment),
J. Hernandez (Hera-B Farms), A. Spiridonov (Hera-B Reconstruction)

---

**Experiments at CERN / Simulation in many places / Analysis everywhere**

1-100 GB/sec

$\begin{array}{c}
\text{Event Filter} \\
\text{(selection \& \ reconstruction)}
\end{array}$

$\begin{array}{c}
\text{One Experiment} \\
\text{~200 MB/sec}
\end{array}$

$\begin{array}{c}
\text{Event Summary Data} \\
\text{350K Sigs}
\end{array}$

$\begin{array}{c}
\text{Processed Data} \\
\text{64 GB/sec}
\end{array}$

$\begin{array}{c}
\text{Event Reconstruction} \\
\text{~100 MB/sec}
\end{array}$

$\begin{array}{c}
\text{Raw data} \\
\text{230K Sigs}
\end{array}$

$\begin{array}{c}
\text{Event Simulation} \\
\text{1 PB/year}
\end{array}$

$\begin{array}{c}
\text{350K Sigs} \\
\text{500 TB}
\end{array}$

$\text{200 TB/year}$

$\text{Thousands of scientists worldwide}$
LHC - Data Grid Hierarchy 2005 (H. Newman)

Tier 1
- FNAL Center
- IN2P3 Center
- INFN Center
- RAL Center

Tier 2
- CERN Computer Ctr
- ~25 TIPS

Tier 3
- Physics data cache
- Workstations

Tier 4
- ~0.6-2.5 Gbps
- Tier 0 + 1

Tier 0 + 1
- Tier 2 Center
- Center
- Center

Physicists work on analysis “channels”
Each institute has ~10 physicists working on one or more channels

CERN/Outside Resource Ratio ~1:2
Tier0/(Σ Tier1)/(Σ Tier2) ~1:1:1

LHC / Grid Computing

Layered Grid Architecture
(By Analogy to Internet Architecture)

"Coordinating multiple resources": ubiquitous infrastructure services, app-specific distributed services

"Sharing single resources": negotiating access, controlling use

"Talking to things": communication (Internet protocols) & security

"Controlling things locally": Access to, & control of, resources

Application

Collective

Resource

Connectivity

Fabric

Transport Internet

Link
GRID from a services view

Applications
- Chemistry
- Cosmology
- Environment
- Biology
- High Energy Physics

Application Toolkits
- Distributed computing toolkit
- Data intensive applications toolkit
- Collaborative applications toolkit
- Remote visualisation applications toolkit
- Problem solving applications toolkit
- Remote instrumentation applications toolkit

Grid Services (Middleware)
- Resource-independent and application-independent services
  - authentication, authorisation, resource location, resource allocation, events,
  - accounting, remote data access, information, policy, fault detection

Grid Fabric (Resources)
- Resource-specific implementations of basic services
  - E.g., transport protocols, name servers, differentiated services, CPU schedulers,
  - public key infrastructure, site accounting, directory service, OS bypass

LHC / Grid Computing (B. Segal at CERN HNF meeting)

10-005 Querying Large Physics Data sets over an information Grid (R. McClatchey)

Applications
- HEP Apps (WP8)
- EO Apps (WP9)
- Bio Apps (WP10)

Data Grid Services
- Worldcloud Management (WP1)
- Data Management (WP2)
- Monitoring Services (WP3)

Core Middleware
- Globus Middleware

Physical Fabric
- Fabric Management (WP4)
- Networking (WP7)
- Mass Storage Management (WP5)

EU-DataGrid
LHC / Grid Computing

Goal:
... Grid is (also) the attempt to develop a new world-wide “standard engine” to provide transparent access to resources (computing, storage, network....) for coordinated problem solving by dynamical Virtual Organizations.

Implementation:
Globus, Avaki (commercialized former Legion project)

Projects:
EU Data Grid (CERN + others, 9.8 M Euros EU funding over 3 years)
Grid Physics Network (GriPhyN)
Particle Physics Data Grid (PPDG, $ 11 M, US DOE, over 4 years)
International Virtual Data Grid Laboratory (iVDGL, $ 15 M, US NSF, over 5 years)
... and many many others

Cluster Computing

HEP wide: More and more (Linux) PC clusters are replacing large SMP and MPP ‘mainframes’ for offline simulation and analysis

Challenge: automatic management and support, no common tools, side dependent solutions
CERN: Batch farm, about 800 nodes, dual PII/PIII up to 800 MHz, dedicated and shared subfarms controlled managed via LSF (Load Sharing Facility),
Goal: One big shared batch farm with optimal resource scheduling

Only DESY is officially supporting Linux on desktops.
Red Hat – basic distribution (except DESY – SuSE)

Many online farms in HEP experiments:
CDF, D0, RHIC, H1, ZEUS, HERA-B, HERMES, BaBar
LHC: CMS, ATLAS, ALICE – Testfarms O(100), O(1000) planned from 2005/2006
Cluster Computing

Report on FNAL Cluster Workshop (by invitation, May 2000):
Review of tools with regard to PC clusters
“Administration is far from simple and poses increasing problems as cluster sizes scale” (A. Silverman, CERN)
Quote of the week - “a cluster is a great error amplifier” (Chuck Boeheim, SLAC)
DESY underrepresented, usage of cfengine, grid engine ... not mentioned

Lattice QCD farm at FNAL:
2000: 80 PCs, Dual PIII 600 MHz, Myrinet
(Replacement for ACPMAPS cluster)
2002: 256 nodes P4(XEON) 2 GHz(?), Myrinet2000
2004: 1000 nodes, 1 Tflop expected, extension to 10 Tflops planned
(deployment of about 200 nodes/year)

Mass Storage

Who’s Using What

<table>
<thead>
<tr>
<th>“Experiment”</th>
<th>Event DB</th>
<th>Metadata DB</th>
<th>Mass Storage System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice</td>
<td>Root</td>
<td>MySQL</td>
<td>Castor</td>
</tr>
<tr>
<td>AMS</td>
<td>Root</td>
<td>Oracle</td>
<td></td>
</tr>
<tr>
<td>Atlas</td>
<td>Objectivity</td>
<td>MySQL</td>
<td>Castor</td>
</tr>
<tr>
<td>BaBar</td>
<td>Objectivity</td>
<td>Objectivity</td>
<td>HPSS</td>
</tr>
<tr>
<td>BES</td>
<td>Root</td>
<td>Oracle</td>
<td></td>
</tr>
<tr>
<td>CDF</td>
<td>Root</td>
<td>Oracle</td>
<td>DIM</td>
</tr>
<tr>
<td>CMS</td>
<td>Objectivity</td>
<td>Objectivity</td>
<td>Enstore, HPSS, Castor</td>
</tr>
<tr>
<td>COMPASS</td>
<td>CDR</td>
<td>Objectivity</td>
<td>Castor</td>
</tr>
<tr>
<td>D0</td>
<td>Flat Files</td>
<td>Oracle</td>
<td>SAM+Enstore, HPSS, etc</td>
</tr>
<tr>
<td>JLAB-located</td>
<td>various</td>
<td>various</td>
<td>JASMin</td>
</tr>
<tr>
<td>KEK-located</td>
<td>various</td>
<td>various</td>
<td>HPSS</td>
</tr>
<tr>
<td>KLOE</td>
<td>YBOS</td>
<td>YBOS+DB2</td>
<td>ADSM+Local</td>
</tr>
<tr>
<td>LHCb</td>
<td>Root</td>
<td>In Progress</td>
<td>Castor</td>
</tr>
<tr>
<td>Star</td>
<td>Root</td>
<td>MySQL</td>
<td>HPSS</td>
</tr>
<tr>
<td>ZEUS</td>
<td>Objectivity</td>
<td>Objectivity</td>
<td>OSM, EuroStore</td>
</tr>
</tbody>
</table>
Mass Storage

RAIT (Reliable Array of Inexpensive Tapes)
- Parity tapes to data stripes
- US DOE/ASCI project, talk by Storage Technology Corporation (STK)
- Test setup: 80 MB/sec sustained,
  1 FiberChannel Interface striped out to 8 drives
- Goal: 8 GB/sec by striping 12 RAIT systems together

dCache
- Joint project: DESY-IT, FERMI CD_INTEGRATED SYSTEMS
- “Generic tool for caching, storing and easily accessing huge amounts of data, distributed among a large set of heterogeneous caching nodes”
- Single name space (pnfs)
- Hot Spot Spreading
- OSM interface

Disk storage is the biggest cost risk for LHC computing

DAQ

More than 31 contributions ...

3x ATLAS, LHC, CERN
3x CMS, LHC, CERN
2x ALICE, LHC, CERN
1x nTOF, PS, CERN
3x DZERO, Tevatron, Fermilab
2x CDF, Tevatron, Fermilab
1x BeTeV, Tevatron, Fermilab
1x HI, HERA, DESY
1x HERA-B, HERA, DESY
1x Zeus, HERA, DESY
1x AMANDA, (DESY)
1x ANTARES, (CEA)
1x ARGO, YBJ-HACRL
1x BaBar, PEP-II, SLAC
1x BES III, BEPC-II, IHEP Beijing
1x CLEO, CESR
1x FLNP, IBR-2, JINR
1x ISTRA+, IHEP Protvino
1x PHENIX, RHIC, BNL
1x SND2000, VEPP-2000, IHEP Novosibirsk
...
**Main topic: High Rate Systems**

Typical solution: pipelined readout + hardware filter + software filter (PC farm)

**Examples:**

**Existing:**
- HSOA-B
  - 500000 Channels
  - 10 MHz, 90 GB/s
  - Hardware Filter: 10 MHz, 10 GB/s

**Future:**
- ATLAS
  - 10000000 Channels
  - 40 MHz x 25 events
  - Hardware Filter: 75 MHz, 0(200) GB/s

- L2 Software Filter (RoI):
  - 50 kHz, 250 MB/s
- L3 Software Filter (Full Data):
  - 500 kHz, 250 MB/s
- L4 Software Online Reco
  - 50 Hz, 6MB/s

- 240 PCs
- 400 PCs

**Critical component: Farm communication using Gigabit Ethernet, Myrinet**

Example:
Quality of Service on Linux for the Atlas TDAQ Event Building Network:

Test setup: 4 Linux PCs with Gigabit Ethernet plus Gigabit Switch

- QoS can eliminate packet loss on UDP/IP multicast transfer (up to 60% without QoS)
- CPU usage of QoS is small on the transfer

In small size of message, QoS was 10% worse. In large size of message, the percentage increases.
**DAQ**

### Readout/Run Control/Online Software...

**Frequently used technologies:**

<table>
<thead>
<tr>
<th>OS</th>
<th>Low level software</th>
<th>High level software</th>
<th>Client-server/remote actions</th>
<th>GUI, histogramming, etc.</th>
<th>configuration</th>
<th>Scripting (run control)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VxWorks (realtime)</td>
<td>C++</td>
<td>C++</td>
<td>CORBA</td>
<td>ROOT (C++)</td>
<td>databases (Objectivity, mySQL) via JDBC</td>
<td>Python</td>
</tr>
<tr>
<td>LINUX (multithreading, since realtime typically not needed)</td>
<td>JAVA</td>
<td>JAVA</td>
<td>Java</td>
<td>JAVA</td>
<td>XML</td>
<td></td>
</tr>
<tr>
<td>Solaris</td>
<td>JAVA</td>
<td>JAVA</td>
<td>Java</td>
<td>JAVA</td>
<td>XML</td>
<td></td>
</tr>
</tbody>
</table>

**Interesting: Tendency to use many different technologies in parallel:**

**Examples:**

a) A Dataflow Meta-Computing Framework for Event Processing in H1
   - Linux PCs + C++ + Java + omniORB + jdk1.3 CORBA + ROOT + JAS + Python
b) On the way to Maturity - The CLEO III Data Acquisition and Control System
   - VxWorks/Solaris/Windows NT CPUs + C++ + Java + VisiBroker CORBA +
     Objectivity

---

**DAQ**

### Pleasant change to relax: Small systems and simple tools like...

- **TASS:** Trigger and Acquisition system simulator
  - TASS reproduces in a realistic way commercial NRIM, CAMAC and VME modules
  - Access to any hardware and software characteristics like with real modules

- **KTLinux and ROOT for data acquisition in small experiments**
  - **Hardware:**
    - PC with ISA-to-CAMAC controller
    - plus CAMAC crate etc.
    - Interrupt requests are sent to PC using I2C
  - **Software:**
    - KTLinux for hard real-time DAQ
      (Interrupt latency < 15 µs)
    - ROOT for GUI control

**Offers:**

- High DAQ speed
- A reliable and stable DAQ system
- Easy implementation of software design
- Low cost

---

- to help physicists developing trigger systems and students learning the fundamentals

TASS aims to be the 'bridge' joining the existing software packages in HEP
- Detector/GEANT - DAQ/TASS - Analysis/PAW
<table>
<thead>
<tr>
<th>URL</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://www.ihep.ac.cn/~chep01/">http://www.ihep.ac.cn/~chep01/</a></td>
<td>CHEP01 home page</td>
</tr>
<tr>
<td><a href="http://www.EU-DataGrid.org">http://www.EU-DataGrid.org</a></td>
<td>EU Data Grid</td>
</tr>
<tr>
<td><a href="http://www.globus.org/research/papers/anatomy.pdf">http://www.globus.org/research/papers/anatomy.pdf</a></td>
<td>Grid anatomy</td>
</tr>
<tr>
<td><a href="http://www.ieeeftcc.org">http://www.ieeeftcc.org</a></td>
<td>IEEE task force on cluster computing</td>
</tr>
<tr>
<td><a href="http://conferences.fnal.gov/lccws/">http://conferences.fnal.gov/lccws/</a></td>
<td>Fermilab Large Cluster Workshop</td>
</tr>
<tr>
<td><a href="http://iguana.cern.ch">http://iguana.cern.ch</a></td>
<td>IGUANA Toolkit</td>
</tr>
<tr>
<td><a href="http://www-sldnt.slac.stanford.edu/nld">http://www-sldnt.slac.stanford.edu/nld</a></td>
<td>Java Analysis Studio</td>
</tr>
<tr>
<td><a href="http://tass.roma1.infn.it/">http://tass.roma1.infn.it/</a></td>
<td>Tass DAQ simulator</td>
</tr>
</tbody>
</table>