Cours/Lecture Series

1992 – 1993 ACADEMIC TRAINING PROGRAMME

LECTURE SERIES

SPEAKERS : J. KNOBLOCH, M. MAPELLI / CERN-ECP, R. MOUNT / CERN-PPE, & L. ROBERTSON / CERN-CN

TITLE : LHC Era Computing

TIME : 10, 11, 12 & 13 May, 11.00 to 12.00 hrs

PLACE : Auditorium

ABSTRACT

This course aims to show why computing for the LHC era will be difficult, and also why it will be interesting! An LHC experiment will be faced by some billion events per second. Even after the most rigorous selection, an experiment may record 1 Terabyte per day. How will this be organized, how shall we find the computing power? How shall we make the data easily accessible for analysis by thousands of physicists all over the world? And how shall we develop the software we need to do the job? We give an outline of these problems and the new approaches and technologies that will help us to solve them.
LHC Era Computing
The first step: define the problem

Livio Mapelli
CERN

- What determines computing needs
- Experimental parameters
  - Physics
  - Accelerators
  - Detectors
  - Rates
- Trigger/DAQ
  - DAQ Architecture Model
  - Distributed computing
- Software
  - Overall approaches
  - S/w development environment
  - Operating System requirements
- CERN DRDC program
- Conclusions

CERN Academic Training
May 1993
What determines computing needs

Physics -> Experimental program

Algorithms

Amount of data

Type of data

Computing

DAQ

Selection

Analysis

Online Computing

Offline Computing

Hierarchical distribution of computing

H/W: Processing Architecture

S/W: Dev. environment

OS environment

Prod. environment
Experimental parameters

Study how experimental conditions evolve in response to new physics demands.

Influence of parameter evolution on computing requirements and Trigger/DAQ architectures.

- **Physics** (rejection required)
- **Accelerator** (time structure)
- **Detectors** (size, no channels)
- **Rates** (events, triggers, data volumes)

1980 pp 1990 LEP HERA 2000 LHC/SSC
Physics

Discovery of Intermediate Vector Bosons (rej. $10^9$)

Beautiful confirmation of electroweak theory (Standard Model)
The next step

Still many unknowns in Standard Model:
• origin of masses
• symmetry breaking
• generation hierarchy

Possible solution:
existence of Higgs boson

Obvious next step:

hunt for Higgs

Unpredictable mass
need widest energy range

Small x-section
need highest luminosity

Other explorations (new energy domain):
• supersymmetry
• compositeness
• new particles

Higgs or no Higgs
at 1 TeV 'something' must happen
Physics requirement evolution

Overall rejection power required

- 1980
- 1988
- 1992
- 2000

powers of 10
Accelerators

Multi-Tev hadron collisions

are the only technically accessible way of producing today
~ 1 TeV energies at the constituent level

Two machines under design:

**LHC**  
Large Hadron Collider  
CERN - Geneva  
proton-proton  
16 TeV  
$10^{34}$ cm$^{-2}$ s$^{-1}$

**SSC**  
Superconducting Super Collider  
SSC Lab - Texas  
proton-proton  
40 TeV  
$10^{33}$ cm$^{-2}$ s$^{-1}$
The Supercolliders

The higher the luminosity the more challenging for Trigger/DAQ.

LHC in p-p mode (nominal configuration)

Revolution time: 88.9 \mu s
Revolution rate: 11.2 kHz

Injection kicker: 1 \mu s (x11)
Extraction kicker: 3.17 \mu s

No. bunches/beam: 4725
4\sigma bunch length: 31 cm

Xing rate: 66.8 MHz
Inter-bunch: 15 ns

\( \mathcal{L} = 1.7 \times 10^{34} \)
Today's accelerator timing

- **pp (UA2)**
  - 250 kHz
  - 50 kHz
  - 100 Hz
  - 10 Hz

- **Level 1**

- **Level 2**
  - 1 ms

- **Readout**
  - 10 ms

- **Level 3**
  - 100 ms

- **LEP**
  - 30 kHz
  - ≈ kHz
  - 22 μs

- **Tevatron**
  - 280 kHz
  - ≈ 100 kHz
  - 3.5 μs

- **HERA**
  - 10 MHz
  - ≈ 10 kHz
  - 96 ns

L. Mapelli - Cern Academic Training - May 93
**LHC/SSC timing**

- Level 1 trigger time exceeds bunch interval
- Detector cell memory greater than 15 ns
- Event overlap & Signal pileup
- Very high number of channels
- Level 1 trigger rate at LHC:
  - Full event rate at previous machines
Accelerator timing evolution

Frequency of bunch crossings

1980 1990 2000

pp LEP HERA LHC/SSC

L. Magelli - Cern Academic Training - May 93
### Detectors

<table>
<thead>
<tr>
<th>Particle type</th>
</tr>
</thead>
<tbody>
<tr>
<td>e</td>
</tr>
<tr>
<td><img src="image" alt="Central detector" /></td>
</tr>
</tbody>
</table>

4π complementary detector system combining the information coming at the same time.
Detector layout

Muon detector

Central detector

EM Calorimeter

Hadron Calorimeter
ATLAS

S.C. Air Core Toroids

S.C. Solenoid

Hadron Calorimeters

Forward Calorimeters

Muon Detectors

EM Calorimeters

Inner Detector
Detector evolution

Number of electronics channels

![Graph showing the evolution of the number of electronics channels from 1980 to 2000. The graph indicates a significant increase in the number of channels between 1990 and 2000.](image)

- **1980:**
  - pp
- **1990:**
  - LEP
  - HERA
- **2000:**
  - LHC/SSC
**Cross section -> event rate**

Assume:
- $\sigma_{\text{total}} = 100 \text{ mb}$
- $\sigma_{\text{elastic}} = 30 \text{ mb}$
- $\sigma_{\text{single-diffr.}} = 10 \text{ mb}$

\[ \sigma_{\text{inelastic}} = \sigma_t - \sigma_{\text{el}} - \sigma_{\text{s.d.}} = 60 \text{ mb} \]

**Rate** (interactions/s) = $\sigma_{\text{In}}$ (mb=10^{-27} cm²) $\times$ $\mathcal{L}$ (interactions/cm²⋅s)

= 60⋅10^{-27} cm² ⋅ (10^{33}-4⋅10^{34}) interactions/cm²⋅s

= $6 \cdot 10^7 - 4 \cdot 10^9$ interactions/s

From bunch structure of beams (b-c = 15ns)

**Rate** = $6.7 \cdot 10^7$ b-c/s
(with ≈1-40 overlapping events)
Rates

Estimate of cross sections
Model of general purpose detector
Physics Monte Carlo
Trigger rate + event size

Event rate

(LHC) $\sigma_{\text{inel}} = 60 \text{ mb}$ \hspace{1cm} $L = 2 \cdot 10^{34} \text{ cm}^{-2}\text{s}^{-1}$

Rate = $2 \cdot 10^9$ interactions / s
From beam bunch structure ($b$-xing = 15 ns)
Rate = $6.7 \cdot 10^7$ $b$-xing / s
$\approx 20$ overlapping events / $b$-xing

Event size

General purpose detector
- $e$, $\mu$, jet, $P_{T\text{miss}}$
- $|\eta| \leq 3$ ($\leq 5$ coarse)
- $\mathcal{L} = 2 \cdot 10^{34} \text{ cm}^{-2}\text{s}^{-1}$

Inner tracking : $\approx 700$ tracks / 15 ns
Calorimetry : $\approx 1400$ particles / 15 ns
Muons : $< 0.1$ tracks / 15 ns

Trigger rate

Background of typical physics channels
(lepton, jet, $p_T^{\text{miss}}$ for top, Higgs, Susy)

LVL-1 $10^4 - 10^5$ Hz
LVL-2 $10^2 - 10^3$ Hz
LVL-3 100 MB/s (?)
Table 5.1: Some benchmark physics processes and possible level-1 trigger criteria which are efficient for these signatures.

<table>
<thead>
<tr>
<th>Process</th>
<th>Level-1 trigger</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Higgs \rightarrow \gamma \gamma$ ($80 &lt; m_H &lt; 130$ GeV)</td>
<td>$2 \gamma$ with $p_T &gt; 20$ GeV</td>
</tr>
<tr>
<td>$Higgs \rightarrow 4\ell$ ($120 &lt; m_H &lt; 800$ GeV)</td>
<td>$e-e, \mu-\mu$ or $e-\mu$ with $p_T &gt; 20$ GeV for both leptons</td>
</tr>
<tr>
<td>$Higgs \rightarrow 2\ell + 2\nu$ (very large $m_H$)</td>
<td>$e-e, \mu-\mu$ or $e-\mu$ with $p_T &gt; 20$ GeV for both leptons</td>
</tr>
<tr>
<td>$Top \rightarrow 3$ jets + lepton</td>
<td>$e$ or $\mu$ with $p_T &gt; 40$ GeV</td>
</tr>
<tr>
<td>$Top \rightarrow 2$ leptons</td>
<td>$e-e, \mu-\mu$ or $e-\mu$ with $p_T &gt; 20$ GeV for both leptons</td>
</tr>
<tr>
<td>W-Z pairs</td>
<td>$e-e, \mu-\mu$ or $e-\mu$ with $p_T &gt; 20$ GeV for both leptons</td>
</tr>
<tr>
<td>SUSY jets + $E_T^{miss}$</td>
<td>$\geq 3$ jets with $p_T &gt; 200$ GeV + $E_T^{miss} &gt; 200$ GeV</td>
</tr>
<tr>
<td>SUSY cascade decay to leptons</td>
<td>$e-e, \mu-\mu$ or $e-\mu$ with $p_T &gt; 20$ GeV for both leptons</td>
</tr>
<tr>
<td>$Z', W' \rightarrow$ leptons</td>
<td>$e$ or $\mu$ with $p_T &gt; 40$ GeV</td>
</tr>
<tr>
<td>$Z', W' \rightarrow$ jets</td>
<td>$2$ jets with $p_T &gt; 200$ GeV</td>
</tr>
</tbody>
</table>

Table 5.2: Level-1 trigger rates at $\mathcal{L} = 1.7 \times 10^{34}$ cm$^{-2}$ s$^{-1}$.

<table>
<thead>
<tr>
<th>Trigger</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\geq 1$ isolated em cluster with $p_T &gt; 40$ GeV (isolation not required for clusters with $p_T &gt; 65$ GeV).</td>
<td>31 kHz</td>
</tr>
<tr>
<td>$\geq 2$ isolated em clusters, each with $p_T &gt; 20$ GeV (loose isolation cut)</td>
<td>16 kHz</td>
</tr>
<tr>
<td>$\geq 1\mu$ with $p_T &gt; 20$ GeV</td>
<td>8 kHz</td>
</tr>
<tr>
<td>$\geq 2\mu's$, each with $p_T &gt; 20$ GeV</td>
<td>67 Hz</td>
</tr>
<tr>
<td>$\geq 2$ jets, each with $p_T &gt; 200$ GeV</td>
<td>5 kHz</td>
</tr>
</tbody>
</table>

Figure 5.1: Inclusive trigger rates as a function of $p_T$ threshold for (a) the $e$/$\mu$ trigger (with and without an isolation requirement), (b) the $d$-$t$-jet trigger. The calculation is for $\mathcal{L} = 1.7 \times 10^{34}$ cm$^{-2}$ s$^{-1}$. 
Data volumes & bandwidths

- Estimate meaningful on a "generic detector"?
- Guess based on subsystem examples.

Expected occupancy

Muon detection
Calorimetry
Inner tracking

LVL - 1
$10^{10} - 10^{11}$ Bytes/s

Presumably the bulk of this will not be moved out of local memories, if LVL-2 algorithms based on LVL-1 triggers

LVL - 2
$10^8 - 10^9$ Bytes/s

We need Gbits to GBytes/s bandwidth to empty local memories, build event and transfer data to processor farm

LVL - 3
10 - 100 MBytes/s

Assuming a x10 reduction: "Final" physics analysis?

L Mapeili - Cern Academic Training - May 93
Rates evolution

Data volumes for permanent storage


MB/s

1980 1990 2000

pp  LEP HERA LHC/SSC

L. Mapelli - Cern Academic Training - May 93
# The LHC/SSC jump

<table>
<thead>
<tr>
<th></th>
<th>$\bar{p}p$ Colliders</th>
<th>LHC/SSC</th>
<th>HERA: 96 ns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bunch Xing</td>
<td>4 $\mu$s</td>
<td>15 ns</td>
<td></td>
</tr>
<tr>
<td>Luminosity</td>
<td>$10^{31} \text{ cm}^{-2} \text{s}^{-1}$</td>
<td>$&gt;10^{34} \text{ cm}^{-2} \text{s}^{-1}$</td>
<td></td>
</tr>
<tr>
<td>Event rate</td>
<td>$10^4$ Hz</td>
<td>$10^9$ Hz</td>
<td></td>
</tr>
<tr>
<td>Prompt trigger</td>
<td>100 Hz</td>
<td>$10^5$ Hz</td>
<td></td>
</tr>
<tr>
<td>Physics reduction</td>
<td>$10^{-9}$</td>
<td>$10^{-13}$</td>
<td></td>
</tr>
<tr>
<td>No. channels</td>
<td>&lt;$10^5$</td>
<td>&gt;$10^7$</td>
<td>LEP: &lt;$10^6$</td>
</tr>
<tr>
<td>Raw data volume</td>
<td>$10^7$ B/s</td>
<td>$10^{11}$ B/s</td>
<td></td>
</tr>
<tr>
<td>Final data rate</td>
<td>100 kB/s</td>
<td>100 MB/s</td>
<td></td>
</tr>
<tr>
<td>No. physicists</td>
<td>100</td>
<td>&gt;$10^3$</td>
<td>LEP: 500</td>
</tr>
</tbody>
</table>
DAQ Architecture Model

(our current prejudice)

<table>
<thead>
<tr>
<th>Module</th>
<th>Rate [Hz]</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRACKING</td>
<td>$10^8$ 0.001</td>
</tr>
<tr>
<td>CALO</td>
<td>$10^6$ 0.01</td>
</tr>
<tr>
<td>MUON</td>
<td>$10^5$ 0.000001</td>
</tr>
<tr>
<td>LEVEL 1</td>
<td></td>
</tr>
<tr>
<td>Pipeline</td>
<td>&lt; 2 µsec</td>
</tr>
<tr>
<td>MUX</td>
<td></td>
</tr>
<tr>
<td>LEVEL 2</td>
<td></td>
</tr>
<tr>
<td>Digital Buffer</td>
<td>10 - 100 Ev.</td>
</tr>
<tr>
<td>LEVEL 3</td>
<td></td>
</tr>
<tr>
<td>Event Trigger</td>
<td>10 - 100 (10-100 Mbyte/Ev.)</td>
</tr>
</tbody>
</table>

Mass Storage & Analysis
Experimental conditions of next generation HEP will impose severe demands on performance of trigger/DAQ systems to cope with unprecedented rates, rejections and data volumes.

The system must be designed for extreme performance

- Technology evolution will determine overall architecture
- Multi-level processing:
  - where specialised?
  - where general purpose with OS?
**Frontend Configurations**

Design of electronics is related to the engineering of the detectors.

**Shape/Analog Store/Digitize**
- Power dissipation
- Packaging
- Dynamic range
- Clock driven
- Calibration & Timing
- Radiation hardness

<table>
<thead>
<tr>
<th>Dynamic range</th>
<th>Cal. (15-16 bits)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resolution</td>
<td>Cal. (9-10 bits)</td>
</tr>
<tr>
<td>Power consumption</td>
<td>Inner detectors</td>
</tr>
<tr>
<td>Speed</td>
<td>All (n• 67MHz n≥1)</td>
</tr>
</tbody>
</table>

**Digitize/Digital Buffer/Process**
- Digital signal analysis
- Timing and synchronization
- Data driven
- Radiation hardness
- Power dissipation
  - Power dissipation
  - Cabling

Design of electronics is related to the engineering of the detectors.
**Systolic Trigger**

Local object identification and cut on energy sums. 
(e, μ, jets, track segments)

For a deadtimeless first level trigger, the trigger processor must itself operate as a **pipeline** at the 66 MHz bunch crossing rate (eg. **systolic processor**).

1 - 10μs Trigger latency  One event

Techniques
- Analog processors
- Systolic processor arrays
- Associative memories
- Data driven processors
Level II signatures

Full measurement and cuts on kinematics parameters.

Pattern recognition may be still needed (if fine granularity tracking detectors are used) When the selection is made on kinematical cuts then the basic operation to perform is a 3(4) dimensions scalar product.

\[ v_1 \cdot v_2 \rightarrow \sum x_i \cdot y_i \]

At this level data are buffered locally and the data flow can proceed asynchronously.

The processor architecture can be either massive parallel (with asynchronous units) or data flow or farm based.

(according to amount of data and the data collection architecture)
Refine physics selections of level–1
- Full detector granularity and precision (calorimetry / muon tracking)
- Use detectors not available at level–1 (TRD - Preshower - Tracking)
- Sharper thresholds / Improved rejection of background
- Additional requirements (e.g. muon isolation) / Physics selections

Level-1 trigger identifies regions of interest for level-2
- e.g. calorimeter cluster and muon track define road in inner tracking

Several architectures under study
- Programmable processors and high-performance data network
- General-purpose or special-purpose processors
CMS data acquisition

LV1  Level 1 Trigger  DPM  Dual Port Memory
GT   Global Trigger   DL   Data Link
TC   Timing & Control DM   Data Merger
FEP  FrontEnd Pipeline FI   Farm Interface
DR   DeRandomizer     FM   Farm Manager
ROC  ReadOut Controller EFM  Event Filter Manager
FEB  FrontEnd Bus     WS   Work Station Cluster
Level-3 trigger and Backend

bottleneck:
need a few Gbyte/s bandwidth

EAGLE event reconstruction:
a few \times 10^2 \text{ Mips} \cdot \text{s}

1-10 \text{ GB/s}
Readout / Event building

10^5-10^6 \text{ Mips}

10-100 \text{ MB/s}
Data Storage

Backend - Control

Software development not to be underestimated:
Standard Operating System
Software Engineering
Architecture modelling

Up to 100 Mbyte/s:
might need parallel writing

L. Baseli - Cern Academic Training - May 93
Software - overall approaches

Far too early to start software developments, but one could not overemphasize the importance of a well designed on-line system for the control and the monitoring of the experiment.

Given the size and the complication of the system, modern software engineering must be used to handle it.

Our community is not used to such techniques. We must learn now in order to be ready when the development begins.

Learn/develop methods for
- Specification and description language
  - Graphical and textual description
  - Tools for system design & code generation (CASE tools, meta-tools)
- Object programming
- Standard Operating System (platform independence, industry standard, UNIX-POSIX back and front end: RTUnix)
- Database. User interface.

Architecture modelling & simulation
- Behavioral simulation and animations
- Timing and statistics analysis (from gate to complex:
  - VERILOG - SYMSCRIPT - MODSIM)

Interchangeability & interoperability of software
Software requirements

A few observations leading to requirement list:

- high level rejection requires high level analysis
  multi-level distributed CPU power

- size of system requires sophisticated design and control
  modelling tools

- complexity of system similar to major s/w developments in other fields
  SE methods & tools commercial products

- project timescale requires 'longevity' of s/w development
  Interchangeability interoperability OS standardisation

Requirements at 3 levels

Software development environment  DAQ system  OS environment
**S/w development environment**

Complex development
Modern methodologies
Powerful techniques and tools

- Have to approach with SE for design, development and maintenance.

  *Individuate methods adequate for HEP online environment.*

- Modelling and simulation

- Which CASE tools?

  (... trap of being tool dependent at production level ...)

We are fighting a 'Hardware liberation movement'
... are we getting into a 'Software tool dependency'?
**DAQ system requirements**

**DAQ system**

- Multi-processor distributed
- Front-end / Back-end

**Front-end**
- for readout and RT functions

**Back-end**
- for control and management

- Platform independence
  - application portability (FE-BE)
  - programmer portability

- S/w interchangeability
  - easy replacement and truly independent development of system components

- S/w interoperability
  - recombination of components for different application

- Modularity, scalability
  - better understood in h/w (bus)
  - less clear and only recently addressed for s/w system

- Commercial products
  - take maximum advantage of industrial investment in s/w

**DAQ system**

- Data Flow, RO, analysis, recording
- Control
- Info management
- User interface

distributed in multi-processor environment and on two functional levels:
Operating system requirements

A modern O/S environment for DAQ applications must have features which span from fast real-time response to 'general purpose' data acquisition.

Real-Time environment requires compromise between:

- **RT performance**
  - real-time features: determinism, async events, ....

- **Standardisation**
  - platform independence
  - OS standardisation

**Distributed computing ROM-ability**
**Trends in s/w world**

*From* very expensive, 'stable' h/w of low performance, therefore requiring highly optimised machine dependent s/w

*to* cheap powerful architectures supporting big s/w applications

H/w investment much less critical than need to preserve s/w investment

**Consequences**

- careful planning of s/w applications, with maintainability, interchangeability, interoperability as critical issues (modelling, SE, OOP)

- platform independence of s/w applications to take advantage of fast evolving computer h/w

- OS standardisation to support platform independence

- When CASE tool standardisation? many tools today imply use of run-time library. *Is it a necessity?*
Standardisation trends

Thanks to RISC ...

• RISC architecture boosted
  in unprecedented way price/performance ratio

• Technological limit
  possibly not yet reached

• All major manufacturers
  propose RISC based systems in highly competitive market, i.e. more scope for improvements of price/performance
  (today at 300$/MIPS, still evolving as $2^{\text{year}}$)

... strong trend to standardisation.

• Advent of RISC accelerated
  trends towards standardisation of UNIX(-like) OS, in highly dynamic market

• Large manufacturer consortiums (UI, OSF) and standardisation bodies (e.g. IEEE POSIX)
  actively working towards OS environments independent of particular (and proprietary) computer architecture
  (UI & OSF guarantee compliance to POSIX standards)
First step on OS environment

Goal: • Platform independent applications
- Application movable between host and F/E
- Easily exploit advances in computing technology (e.g. RISC)
- Freedom from one manufacturer constraint

UNIX seems the best candidate of common OS because of its generic architecture
(direction closest to industry trend uniformity and compliance to OS standards)

UNIX gives:

• O/S portability
  - UNIX has a generic (platform independent) architecture
  - Mostly written in a HLL (C)

• O/S standardization efforts point to UNIX
  - Basis for effort towards standard OS interface (IEEE POSIX)
  - Manufacturer consortiums (UI, OSF) propose common OS:
    System V.4 (unifying SystemV, BSD, Xenix)
    OSF/1 (based on Mach) (both guarantee POSIX compliance)

• Wide availability
  - All RISC computers run a UNIX brand
  - Traditional architectures propose UNIX (ULTRIX, AIX,...)

UNIX cons:

• Supposedly difficult user interface
  But commercial products exist for easy (icon based) UI

• No uniform IPC mechanism
  SVID provides SVIPC, BSD provides sockets

• Lack of real time features
  - Determinism (pre-emptable Kernel, memory locking)
  - Asynchronous I/O (SUNOS has it for ex.)
  - Asynchronous events (ASTs, ev. flags. Signals are poor substitute)
Real-time Unix

Three categories of commercial solutions for industrial demand of RT-Unix:

- **Real-Time executives**
  with some degree of **UNIX compatibility**
  e.g. X-development supported on UNIX system while target processor runs non-UNIX kernel: VxWorks, VRTX, pSOS

- **Specialised Real-Time kernels**
  to which **UNIX support** has been added
  typically SV compatibility: Chorus, DuneIX

- **UNIX systems**
  written from scratch with **Real-time capabilities**
  both modified SV or proprietary: Real-IX, RTU, LynxOS, OSF

Given trend towards OS interface standardisation one should opt for
full UNIX systems and require POSIX 1003.4 compliance.

POSIX 1003.4 (RT extension) addresses all the RT features lacking in UNIX
Today's HEP conditions

Today mostly

- **OS-9** for time critical RT (frontend)
- **VMS** on backend (control, management, ...)

One measures already on big, long experiments the **price of lack of standardisation:**

- h/w upgrades constrained by application non-portability
- h/w upgrades involve heavy s/w modifications
- stuck to a system (firm):
  cannot take advantage of market evolution
- ...

**Tendency towards Unix(-like) in new developments:**

- **VxWorks** (especially in USA)
- **real RT-Unix**, mainly **LynxOS:**
  - adopted in CERN accelerator consolidation project
  - investigated in experiment DAQ pilot projects
- **standard workstation UNIX in backend**
  uniformity throughout the system
<table>
<thead>
<tr>
<th>RD-16</th>
<th>F/E electr. for calorimetry</th>
</tr>
</thead>
<tbody>
<tr>
<td>in RD-2</td>
<td>SITP chip</td>
</tr>
<tr>
<td>in RD-6</td>
<td>TRD readout</td>
</tr>
<tr>
<td>in RD-9</td>
<td>Rad-hard SOI-CMOS</td>
</tr>
<tr>
<td>in RD-19</td>
<td>Si-pixel readout</td>
</tr>
<tr>
<td>in RD-20</td>
<td>Si-strip readout</td>
</tr>
<tr>
<td>RD-23</td>
<td>Optoelectronic data transfer</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RD-5</th>
<th>Muon trigger</th>
</tr>
</thead>
<tbody>
<tr>
<td>RD-12</td>
<td>R/O syst. test benches</td>
</tr>
<tr>
<td>RD27</td>
<td>Lvl I trigger</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RD-11</th>
<th>E.A.S.T.</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBI</td>
<td>Transputer based farm</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RD-13</th>
<th>Scalable DAQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>RD-11/13</td>
<td>HiPPI based EB</td>
</tr>
<tr>
<td>RD-24</td>
<td>SCI based DAQ</td>
</tr>
<tr>
<td>RD31</td>
<td>Nebulas</td>
</tr>
</tbody>
</table>

L Mapelli - Cern Academic Training - May 93
Conclusions

LHC experimentation

- physics: rejection required
- accelerator: very high interaction rate
- detector: huge, sophisticated, high granularity
- rate & b/width: orders of magnitude higher

certainly challenging

Recent extensive and accurate studies

- actual technology inadequacy
- complexity of the LHC project

feasible if extensive R&D

Computing does not escape tough demands

- multi-level data selection and acquisition
- distributed computing power
- software environment
- size and timescale

modern approach to computing is essential
(a lot to learn in HEP ...)

L. Mapelli - Cern Academic Training - May 93
LHC ERA COMPUTING

DATA STORAGE AND DATA ACCESS

Richard P. Mount

California Institute of Technology

CERN Academic Training Lecture
May 11, 1993
Storing, Moving and Accessing Data

- Requirements (from physics),

- Requirements (from physicists),

- Technology status and extrapolations:
  - Random access storage
  - Sequential access storage
  - Communications

- Challenges and Choices
  - File management versus database management
  - Geography

- Closing Comments.
Requirements (from Physics)
Figure 5.1: Inclusive trigger rates as a function of $p_T$ threshold for (a) the $e/\gamma$ trigger (with and without an isolation requirement), (b) the di-jet trigger. The calculation is for $L = 1.7 \cdot 10^{34}\text{cm}^{-2}\text{s}^{-1}$.

*Figure borrowed from ATLAS L0I*
Requirements (from Physics)
Data Storage

Store as much data as possible with
- tolerable inconvenience,
- tolerable cost.

Somewhere in the range:
100 Tbytes to 100,000 Tbytes.

New ‘LHC’ era unit:

‘Petabyte’ = 1000 Terabytes

ATLAS quote:
Whether or not the . . . trigger criteria would also be efficient for any unexpected process is of course impossible to estimate.
Requirements (from Physics)
Moving and Accessing Data

What do we do with the data? (1)

- Reconstruction
  - Extract as many key features of the event as possible.
  - Record relationship between features and raw data in a data structure.
  - Typical features:
    - muon candidates
    - jet candidates
    - isolated e, γ candidates
    - $E_T$
    - $E_{\text{visible}}$
    - $\vec{E}_{\text{missing}}$
  - Organize data (select, stream) according to features.

NEEDS: Sequential access to data.
Requirements (from Physics)
Moving and Accessing Data

What do we do with the data? (2)

- Analysis
  - Select events and data within events based on key features.
  - Make plots.

NEEDS: Random access to data:
How random? — example PAW.

<table>
<thead>
<tr>
<th></th>
<th>Event 1</th>
<th>Event 2</th>
<th>Event 3</th>
<th>etc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_{\text{visible}}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$E_{\text{transverse}}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$E_{\text{missing}}(x,y,z)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>muon(s) ($p_x p_y p_z$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>electron(s) ($p_x p_y p_z$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>photon(s) ($p_x p_y p_z$)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>etc.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Requirements (from Physics)
Moving and Accessing Data

We have learned to live (uncomfortably with): serial access to datasets of selected information.

We consider the height of luxury to be: random access at the event level.

We might derive revolutionary benefits from: random access to all identifiable elements of the data. (True ‘database access’)
Requirements (from Physicists)
Requirements (from Physicists)
Moving and Accessing Data

HEP will die if it ceases to be a university-driven field.

⇒ Need all (technologically/financially) possible support for full participation by physicists at universities.

Goals:
1. Workstations on desks at CERN, Cambridge, Caltech are all equally effective as interfaces to data and processing power.
2. One coherent pool of data accessible by all physicists in a collaboration.
3. Personal interaction.
Requirements (from Physicists)
Moving and Accessing Data

Implementation Options

1. Centralise the data and the I/O intensive processing.
   (May be cheapest and most efficient, but nobody likes centralisation.)

2. Distribute (over 1000s of km) the data and I/O intensive processing.
   (Requires duplicate data stores or super-fast networks. Data coherence is tricky. Appealingly de-centralised.)
Technology Status and Extrapolations
# Technology Status and Extrapolations

## Random-Access Storage

<table>
<thead>
<tr>
<th></th>
<th>On the Market</th>
<th>In the Lab.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IBM Magnetic</td>
<td>IBM Optical</td>
</tr>
<tr>
<td></td>
<td>3.5 inch</td>
<td>3.5 inch</td>
</tr>
<tr>
<td></td>
<td>1 Gbyte</td>
<td>128 Mbyte</td>
</tr>
<tr>
<td>Tracks per micron</td>
<td>0.09</td>
<td>0.29</td>
</tr>
<tr>
<td>Bits per micron</td>
<td>2.3</td>
<td>0.67</td>
</tr>
<tr>
<td></td>
<td>0.12</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>0.63</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3.16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>6.68</td>
</tr>
<tr>
<td>Number of platters</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Data rate (sustained)</td>
<td>3.0</td>
<td>0.16</td>
</tr>
<tr>
<td>Mbytes/s</td>
<td>5.2</td>
<td>0.16</td>
</tr>
<tr>
<td>Hard error rate</td>
<td>$&lt; 10^{-13}$</td>
<td>$&lt; 10^{-14}$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$&lt; 10^{-12}$</td>
</tr>
<tr>
<td>Price (SF)</td>
<td>1635</td>
<td>3130</td>
</tr>
</tbody>
</table>
Technology Status and Extrapolations
Random-Access Storage

Extrapolations

• Optical: only limited progress possible.

• Magnetic: progress will continue.

• Price:
  – Price now: \(~$1000\) per Gbyte.
  – Rate of decrease: \(~\text{factor 1.3--1.4 per year}\).
  – Price in 2003: \($35--73\) per Gbyte.

• Possible Storage System:
  200 Gbyte now \(\implies\) 3–6 Tbyte in 2003.
NOTE: areal density of (delivered) IBM disks has risen by 1.7 per year in 1988–1993:

⇒ Superoptimistic guess: 40 Tbyte system in 2003. (10 times current laboratory performance.)
Technology Status and Extrapolations
Sequential-Access Storage

Recording on Tapes

Linear Recording

Many heads; slow tape-head speed.

Helical Recording

Few heads; high tape-head speed.
# Technology Status and Extrapolations

Sequential-Access Storage

<table>
<thead>
<tr>
<th></th>
<th>'IBM' '3480' 0.2 Gbyte</th>
<th>IBM '3490' 1 Gbyte</th>
<th>Exabyte 5 Gbyte</th>
<th>Ampex DD2 25 Gbyte</th>
<th>Creo/ICI 1000 Gbyte</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tracks per micron (bits per micron^2)</td>
<td>0.0015</td>
<td>0.003</td>
<td>0.031</td>
<td>0.028</td>
<td>(~0.7)</td>
</tr>
<tr>
<td>Hard error rate (sustained) Mbytes/s</td>
<td>(\sim 10^{-13})</td>
<td>(\sim 10^{-13})</td>
<td>(&lt; 10^{-13})</td>
<td>(&lt; 10^{-14})</td>
<td>(&gt; 10^{-18})</td>
</tr>
<tr>
<td>Prices (SF): Drive: Media per Gbyte</td>
<td>10,000</td>
<td>30,000</td>
<td>5,000</td>
<td>300,000</td>
<td>400,000</td>
</tr>
</tbody>
</table>

- Linear
- Helical
- Transverse
Technology Status and Extrapolations
Sequential-Access Storage

- Optical:
  - Speed limitation (burning holes).
  - 1 Terabyte tape NOW!
  - Will there be a market for optical mass storage?

- Magnetic:
  - Speed could be adequate (market?).
  - \(\sim 25\) Gbyte per tape NOW (helical).
  - \(\sim 50\) Gbyte per tape 1995+ (linear).
  - Increasing diversity of formats and packaging
    \[\Rightarrow\] Poor data interchange medium!
Technology Status and Extrapolations
Sequential-Access Storage

• Prediction:
  - 1 Tbyte (at least) cartridges for closed shop systems in 2003.
  - 10 Pbyte robotic storage systems in 2003.
  - Media cost 0.1–1.0 SF per Gbyte in 2003.

The days of the 2400 foot tape are gone forever.
Technology Status and Extrapolations
Sequential-Access Storage

Limiting density for surface recording techniques?

DNA stores 2 bits in 10–15 atoms
2 bits in 0.7 (nm)²
3,000,000 bits per (micron)²
Technology Status and Extrapolations
Communications Links

In use now

<table>
<thead>
<tr>
<th>Service</th>
<th>Bandwidth</th>
<th>Medium</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>LAN — Ethernet</td>
<td>10 Mbit/s</td>
<td>copper</td>
<td>SF 300</td>
</tr>
<tr>
<td>LAN — FDDI</td>
<td>100 Mbit/s</td>
<td>fibre/copper</td>
<td>SF 6000</td>
</tr>
<tr>
<td>Leased Lines</td>
<td>2 Mbit/s</td>
<td></td>
<td>SF 200K</td>
</tr>
<tr>
<td>PTT Trunks</td>
<td>2,400 Mbit/s</td>
<td>fibre</td>
<td>SF 40M</td>
</tr>
</tbody>
</table>

(approximate line prices per 1000 km per year)

In development labs

<table>
<thead>
<tr>
<th>Service</th>
<th>Bandwidth</th>
<th>Medium</th>
</tr>
</thead>
<tbody>
<tr>
<td>‘LAN’ — FibreChannel</td>
<td>1,000 Mbit/s</td>
<td>fibre</td>
</tr>
<tr>
<td>PTT or ‘user’ links</td>
<td>10,000 Mbit/s</td>
<td>fibre</td>
</tr>
</tbody>
</table>

In research labs

<table>
<thead>
<tr>
<th>Service</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wavelength Division</td>
<td>32,000 Mbit/s</td>
</tr>
<tr>
<td>Multiplexing</td>
<td></td>
</tr>
<tr>
<td>Limit of WDM</td>
<td>~25 Tbit/s</td>
</tr>
<tr>
<td>Teilnehmer-Nr. / N° d'adhérent / N° d'abonneé</td>
<td>Rechnungsperiode / Période comptable / Periodo contabile</td>
</tr>
<tr>
<td>---------------------------------------------</td>
<td>----------------------------------------------------------</td>
</tr>
<tr>
<td>680159</td>
<td>01.01.93 - 28.02.93</td>
</tr>
</tbody>
</table>

Direction des télécommunications
1211 GENEVE 2

Laboratoire européen de physique des particules (CERN)
Att. Dr. Richard Mount/PPE
1211 GENEVE 23

<table>
<thead>
<tr>
<th>Taxes d'abonnement</th>
<th>MENSUEL</th>
<th>PRORATI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relevé détaillé</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Cambridge - Genève NP 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taxes pour circuits internationaux</td>
<td>33 150.00</td>
<td>66 300.00</td>
</tr>
<tr>
<td>Total</td>
<td>Relevé détaillé</td>
<td>0.00</td>
</tr>
<tr>
<td>Total</td>
<td>Taxes pour circuits internationaux</td>
<td>66 300.00</td>
</tr>
<tr>
<td>Total général</td>
<td>66 300.00</td>
<td></td>
</tr>
</tbody>
</table>
Technology Status and Extrapolations
Communications Links

Only one leading-edge technology (monomode fibre).

Lasting Problems:

- **LAN**: interface costs (volume-driven).
- **WAN**: cable/fibre installation costs (volume-driven).
  (NOTE: European costs are $5 \times \text{‘free market’ costs}.)

Prediction:

- We will be able to do anything for which there is a large market.
Technology Status and Extrapolations
Communications Switches

Available today
Network Systems Corp. PS-32 HiPPI Switch
32 × 32 crossbar
800 Mbits/s per port
~300 KSF.
(HiPPI workstation interfaces cost ~ 20 KSF)

In development
FibreChannel and ATM Switches
e.g. 1024 × 1024 crossbar
1000 Mbits/s per port

Prediction: Switches will be available with adequate performance (for data movement) at a cost per port similar to workstation interfaces.
Technology Status and Extrapolations
Communications Links

Possible Market Evolution:

Voice/text $\Rightarrow$ Images

Still Frames $\Rightarrow$ Motion

text: 1–10 kbytes per page

image: 100–10,000 kbytes
Challenges and Choices
Challenges and Choices

We have recorded 3,000 Tbytes on tape. We have 30 Tbytes of disk space. What do we do?

What do we do now?

- Use our intelligence to guess which 1% of the data we should make directly accessible.
- If (when) we guess wrongly,
  - either abandon the investigation or analysis which looks outside the 1%,
  - or mount tapes for a few weeks,
  - and improve the choice of the 1% for next time.

Minimal use of commercial solutions to the 'I want access to a slowly varying 1% of my data' problem. (Because there are almost no solutions).
Challenges and Choices

What should we do in 2003?

Two possibilities:

- File Management (Hierarchical Storage Management)
- Database Management
Challenges and Choices

File Management

- YOU divide up your data into files, such that at any one time \( \sim 1\% \) of them will be needed. (Good luck!).
- Industry standard software (and interworking hardware) gives you access to your files more-or-less transparently.
  - Products will exist.
  - Success of standardisation effort is not clear.

i.e. There is some hope to improve on VM-STAGE by 2003.
Data Handling Supporting L3 Physics Analysis
Reference Model for
Open Storage Systems Interconnection

Mass Storage System Reference Model
Version 5

Abstract: The IEEE Reference Model for Open Storage Systems Interconnection (OSSI), also
known as the Mass Storage System Reference Model Version 5, is the first in a series of standards
for application and user interfaces to open storage systems. The Reference Model identifies the high-
level abstractions that underlie modern storage systems. It defines common terminology and
concepts that allow the architectures of existing and future systems to be described and compared.
The Reference Model provides a conceptual and functional framework within which independent
teams of experts may proceed with detailed OSSI interface definitions.

Keywords: data processing, information interchange, open systems, mass storage, hierarchical
storage management, reference model, storage system architecture, computer interface
Challenges and Choices
Database Management

Excellent Conceptual match to HEP data analysis problem:

"Search 10,000 Tbytes for the bytes which satisfy my selection criteria"

BUT:

- Typical DBMSs handle a few Gbytes. (for a few $M you can buy a 1 Tbyte system)
- Tapes? What are you talking about? (only one level of mass storage (disks) handled by current systems.)
- DBMS industry has little contact with IEEE Mass Storage System standardisation effort. ("We cannot afford the performance penalty of using standards.")

- DBMS-based physics analysis requires a complete re-design of our software.
Challenges and Choices

Database Management

Reasons for (some) optimism:

- Image handling in commercial computing will drive up database sizes.
- Massively parallel systems work well as dedicated database engines.

Prediction:

- 1 Tbyte database systems with good performance and tolerable cost are very likely for 2003.
- 1000 Tbyte database systems supporting disk and tape storage are uncertain.
Challenges and Choices

File Management and Database Management

Summary

No major problem to analyse data as we do now (but on 1000 times larger data samples).

Revolutionary changes may be possible using database technology.