LCB Workshop

Marseille, 27/9-2/10 1999

- Event Filter Farms
- Distributed Computing and Regional Centres
- Architecture
- Round Table on Software Process
- Simulation
- Persistency at LHC
- Data Analysis
- Technology Tracking

H.Hoffmann
CERN

Planned Review of Computing

- Review of the Progress and Planning of the Computing Efforts at CERN and in the Experiments for the LHC Start-Up

  – *World-wide Analysis/Computing model*
  – *Software project: Design and development*
  – *Management & Resources*

Event Filter Farms

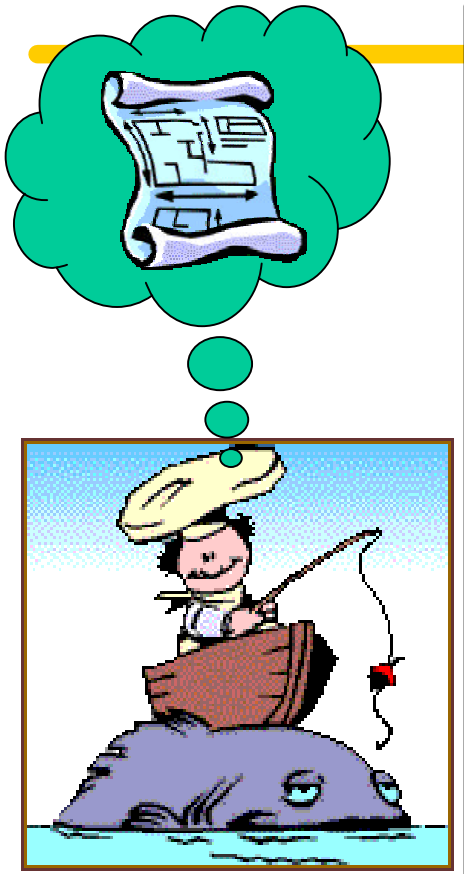
- **Session Chair : François Touchard**
- **Review of Existing Farms**
- STAR, PHENIX, HERA-B, CDF, DØ, BaBar
- ***State of the Art* Farm Computing**
- **The LHC situation**

Event Filter Farms (contd)

- Nobody really discussed event filter farms
  - Situation did not really change
  - Experiments presented their DAQ
  - PDP talked about generic farms
    - Offline oriented
    - Monitoring
    - Single system illusion
    - Process scheduling (NQS & Co.)
    - Failure management
    - Parallel computing
    - Disk/File servers

J.Jaen
IT/PDP

Key Issues

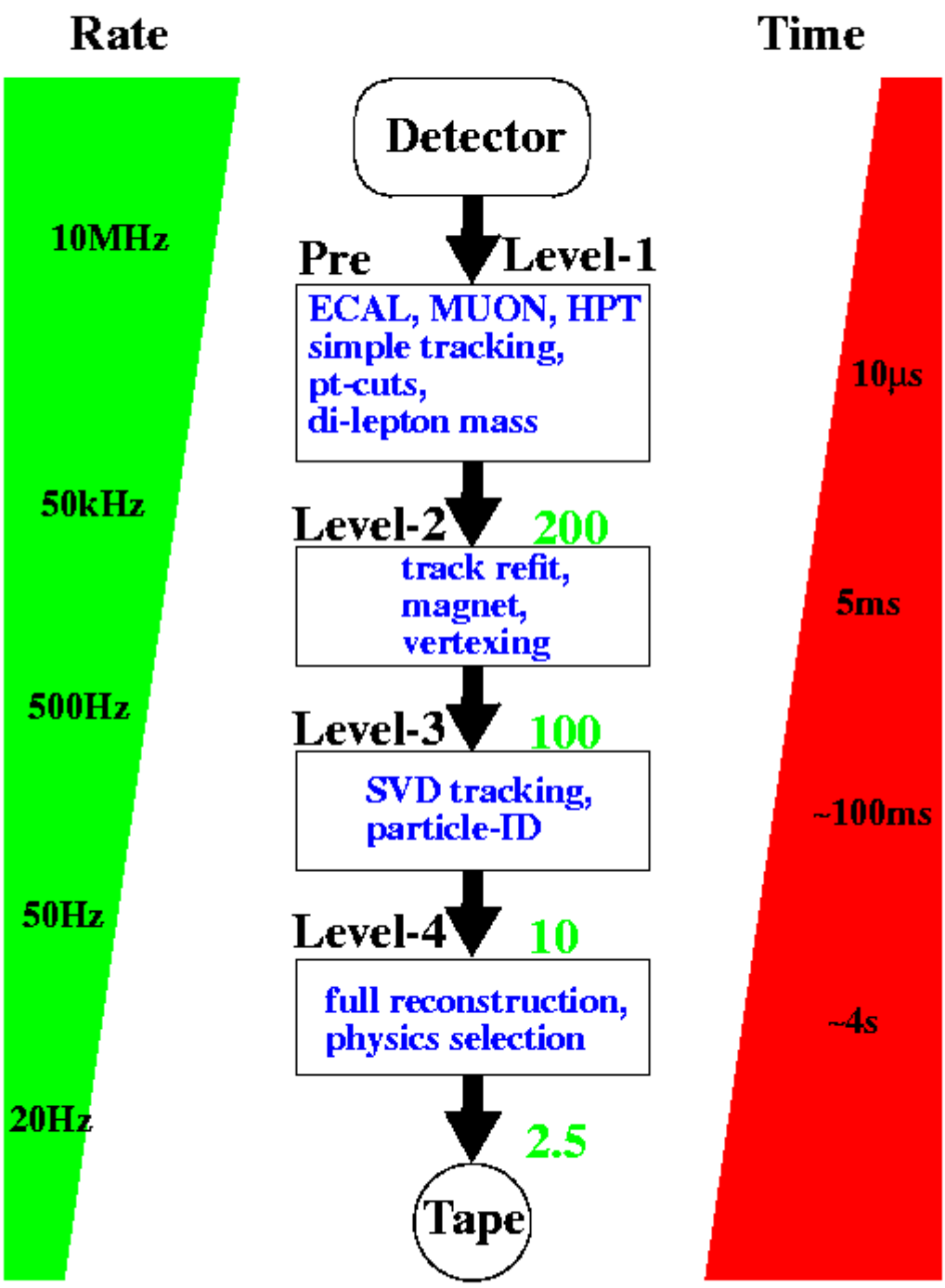


- Size Scalability (physical & application)
- Enhanced Availability (failure management)
- Single System Image (look-and-feel of one system)
- Fast Communication (networks & protocols)
- Load Balancing (CPU, Net, Memory, Disk)
- Security and Encryption (farm of farms)
- Distributed Environment (Social issues)
- Manageability (admin. and control)
- Programmability (offered API)
- Applicability (farm-aware and non-aware app.)

What I'd expected

- ☒ Hardware installation topics
  - ☒ n*1000 PCs take up some space
  - ☒ How to use the cheapest commodity items (slow control…)
- ☒ Level2/3 parameter management (SCADA,…)
- ☒ Trigger programs will run with some FSM
  - ☒ Process control
  - ☒ Filter programs must be in synch. with DAQ
- ☒ Some ideas to collect information
  - ☒ e.g. Monitoring histograms
  - ☒ ….

Rates & Times



Level-2/3 versus Level-4

- **Level-2:** → latency and bandwidth dominated
  - RoI based (event fractions $O(1\%)$)
  - reduction factor ~100
  - high input rate of 50kHz
  - very low processing time $O(1\ \text{ms})$
- **Level-3:** → still latency and bandwidth dominated
  - full event unit available (after event building)
  - reduction factor ~10
  - still high input rate of ~500 Hz
  - low processing time $O(100\ \text{ms})$
- **Level-4:** → processing time dominated
  - full event reconstruction
  - reduction factor ~1
  - moderate inpute rate of 50 Hz
  - long processing time $O(1\ \text{s})$

HERA-B reduced reconstruction time O(10) by optimizing algorithms !!!

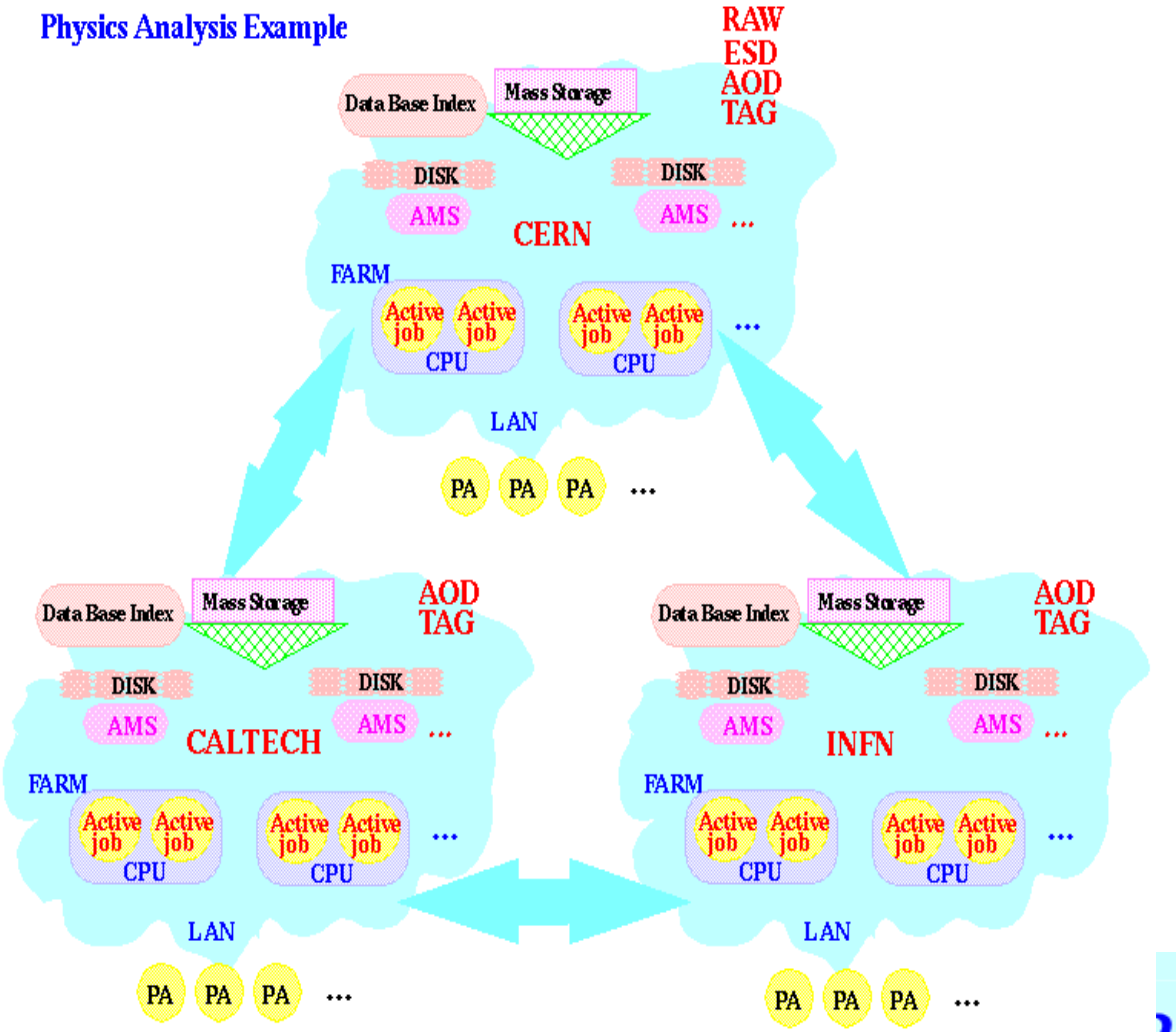
Distributed Computing and Regional Centres

- Rapporteur talk
  Harvey Newman
- MONARC Working Groups status report
  Iosif Legrand
- DPSS Distributed Parallel Storage System
  Brian Tierney

H.Newman
CIT

What is MONARC Doing ?

- Similar data processing jobs are performed in several RCs
- Each Centre has “TAG” and “AOD” databases replicated.
- Main Centre provides “ESD” and “RAW” data
- Each job processes AOD data, and also a a fraction of ESD and RAW.



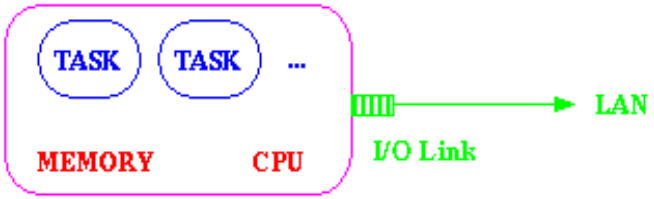
J.Legrand
CIT

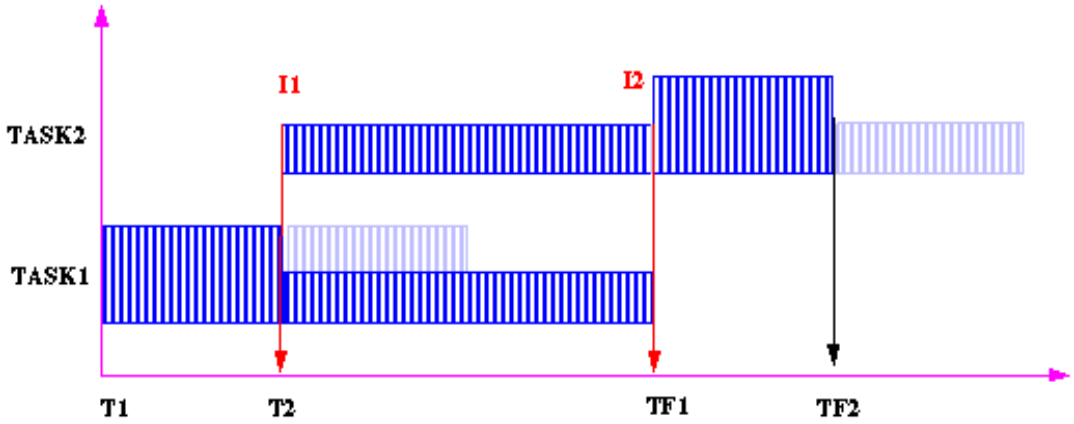
Process Modeling

Concurrent running tasks share resources (CPU, memory, I/O)

"Interrupt" driven scheme:

For each new task or when one task is finished, an interrupt is generated and all "processing times" are recomputed.





It provides:

An efficient mechanism to simulate multitask processing.

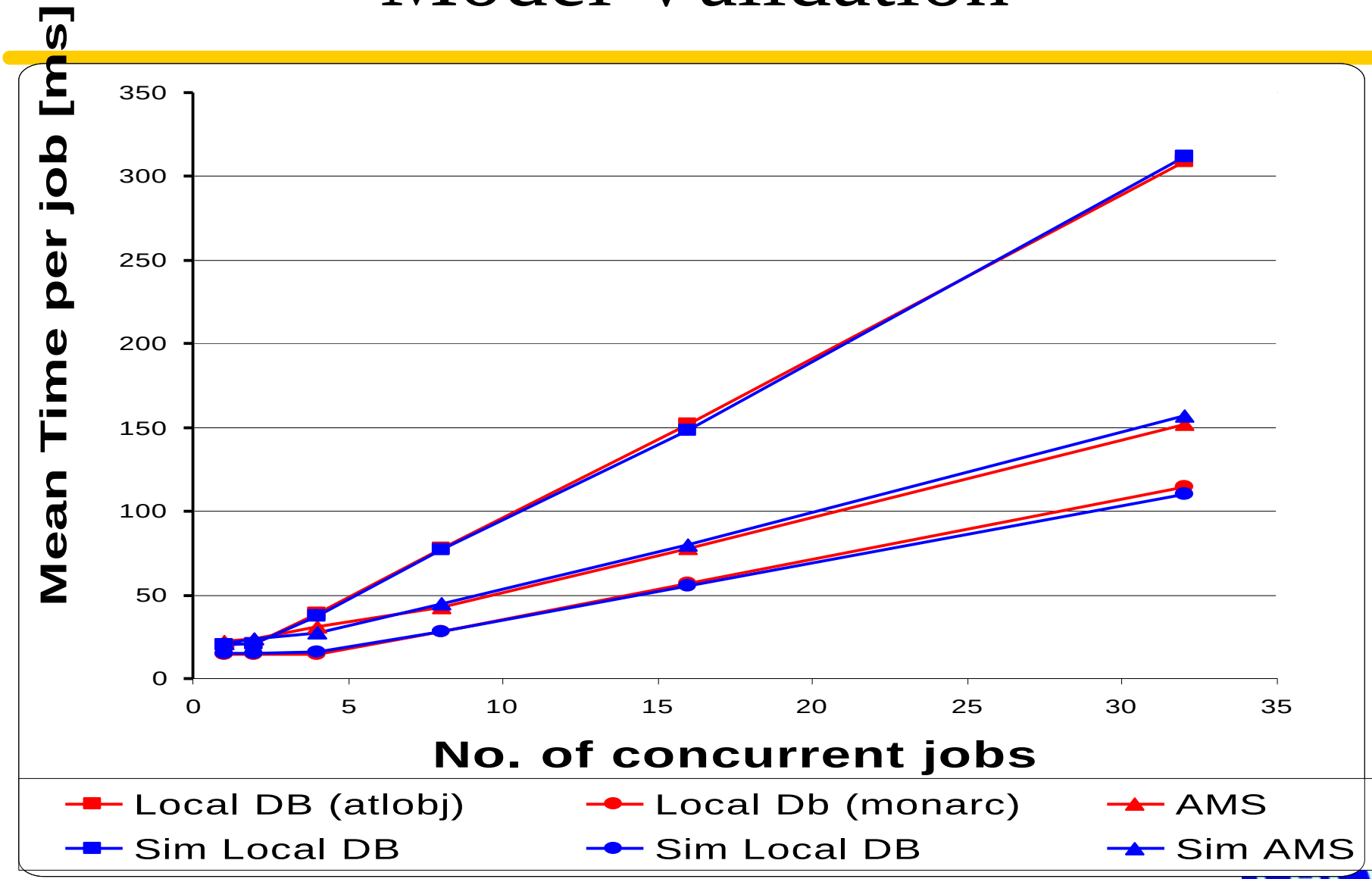
Handling of concurrent jobs with different priorities.

An easy way to apply different load balancing schemes.

M.Frank LHCb/CERN

J.Legrand
CIT

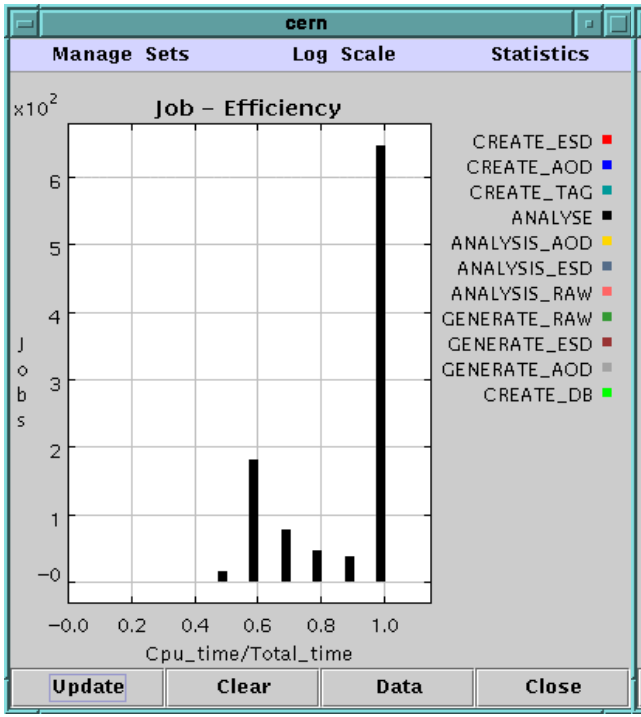
Model Validation



J.Legrand
CIT

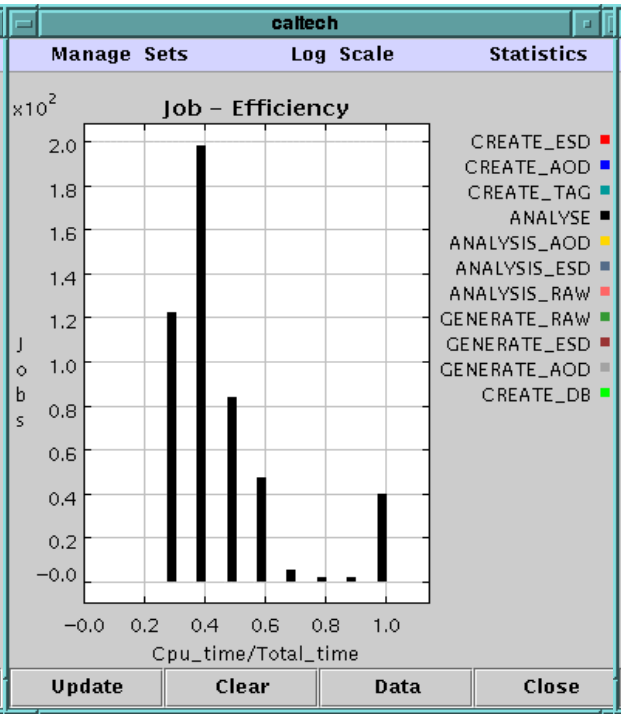
“User Happiness” Factors

“CERN”



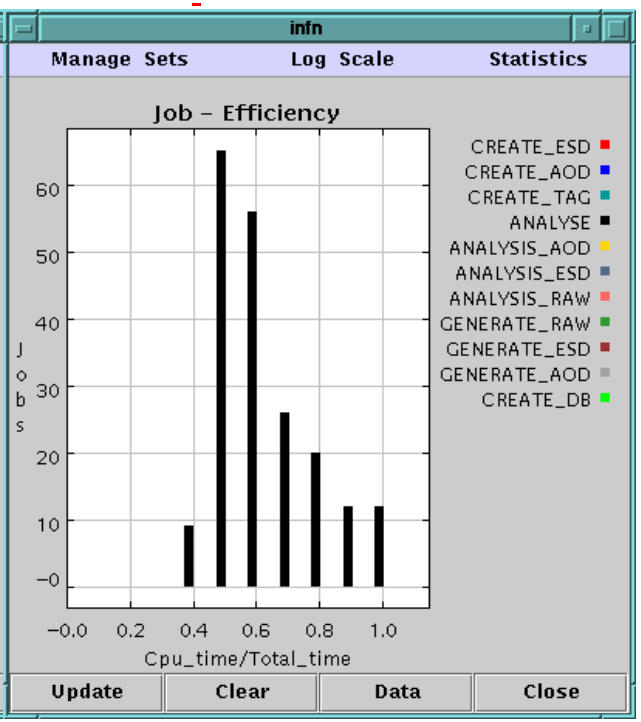
Mean 0.83

“CALTECH”



Mean 0.42

“INFN”



Mean 0.57

Objectivity Usage Statistics

Would these parameters not also fit for LHCb ?

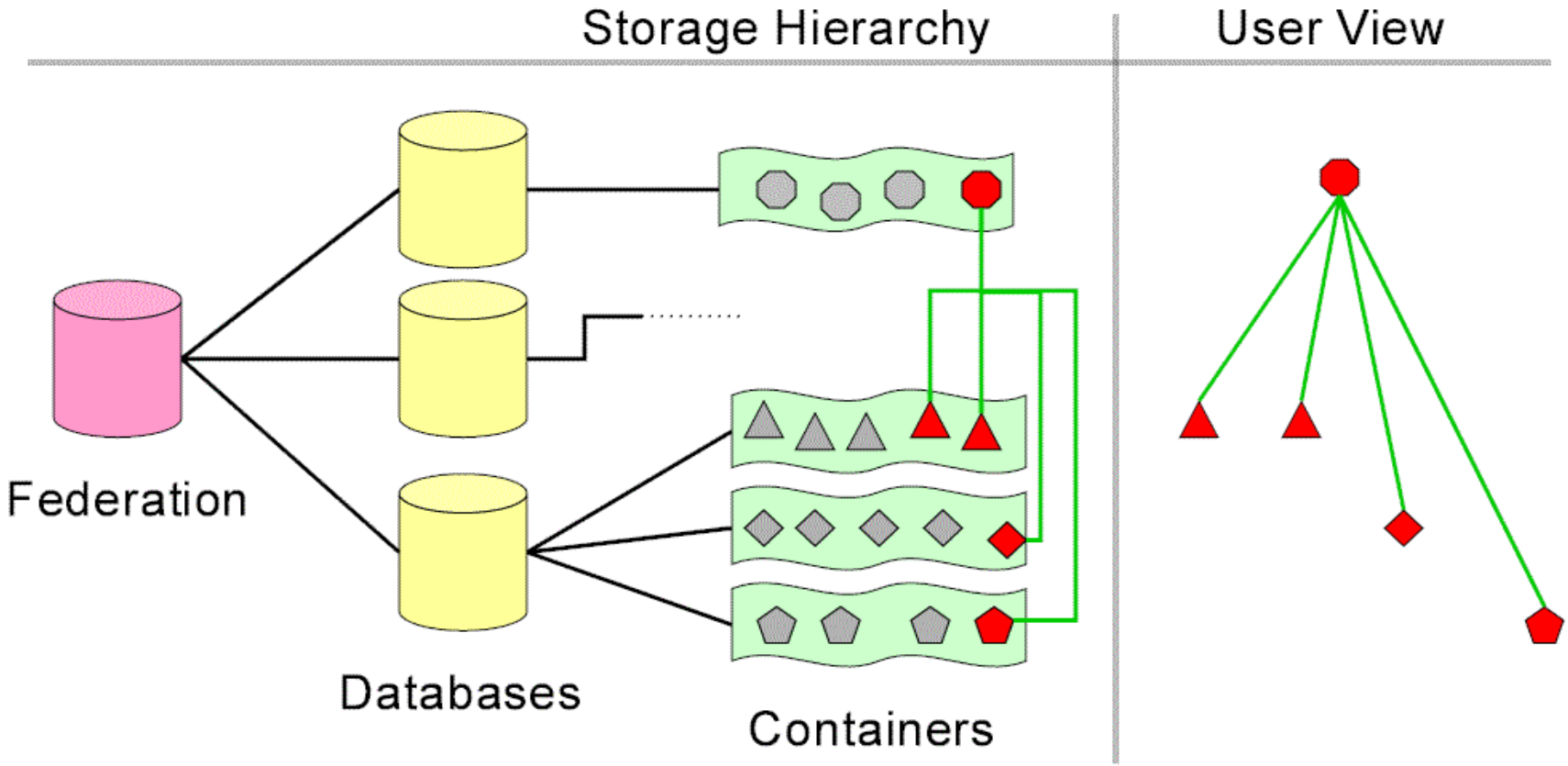
- >20 sites using Objectivity
  - USA, UK, France, Italy, Germany
- ~650 licensees
  - People who have signed the license agreement
- ~400 users
  - People who have created a test federation
- >100 simultaneous users
  - Monitoring distributed oolockmon statistics

Persistency at LHC

- Persistency at LHC
  Vincenzo Innocente
- Recent Experience at BaBar
  David Quarrie

V.Innocente
CMS

Physical Model and Logical Model



- Physical model may be changed to optimise performance
- Existing applications continue to work

V.Innocente
CMS

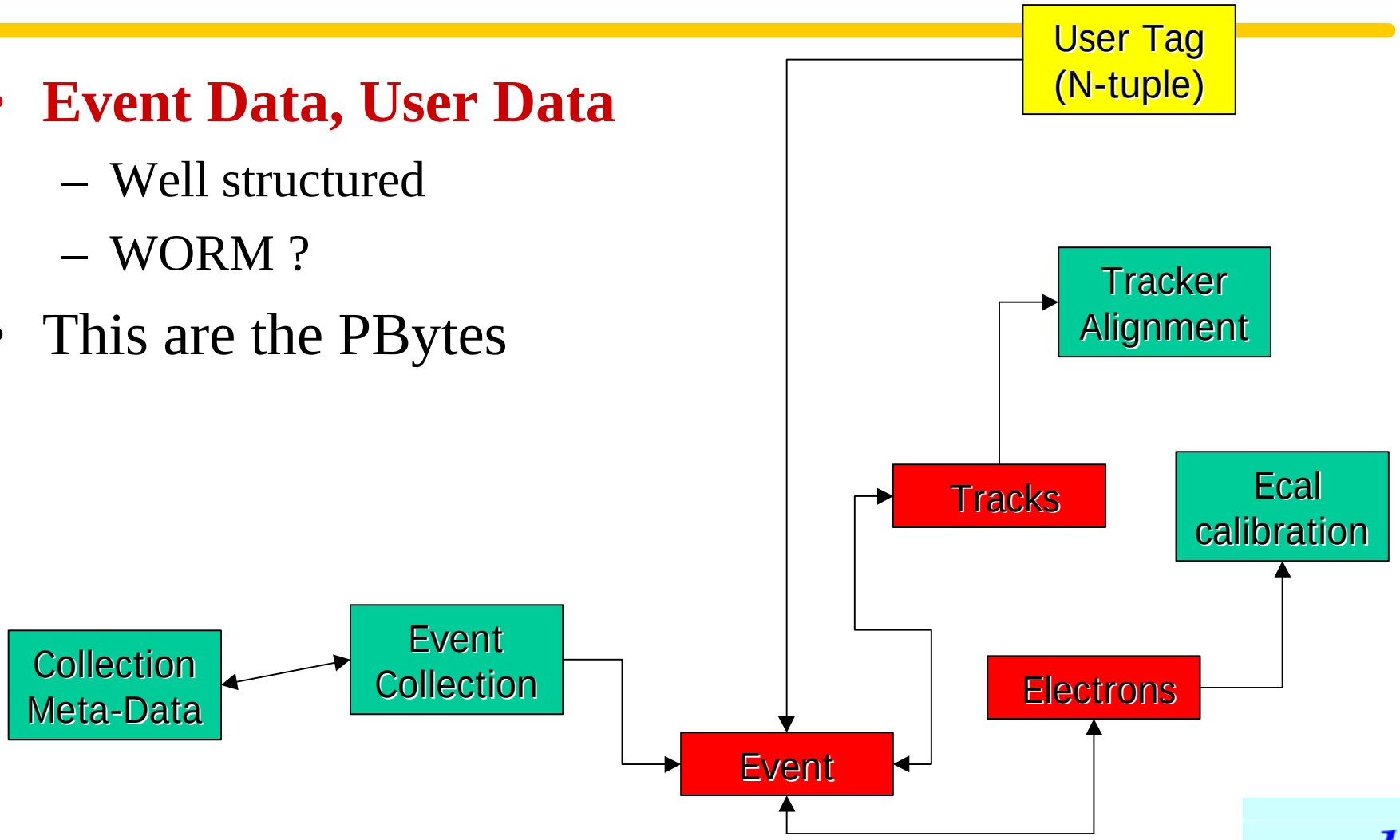
Solution Space

- ODBMS
  - Objectivity/DB
  - In-house build
- ROOT
  - “Quasi” random access to event data
  - Root alone is not sufficient
- Wrapped relational databases
- Hybrid solutions (CDF, D0, Star, Phenix)
  - sequential files/ ROOT files
  - File management using relational databases

V.Innocente
CMS

Bulk HEP Data

- **Event Data, User Data**
  - Well structured
  - WORM ?
- This are the PBytes

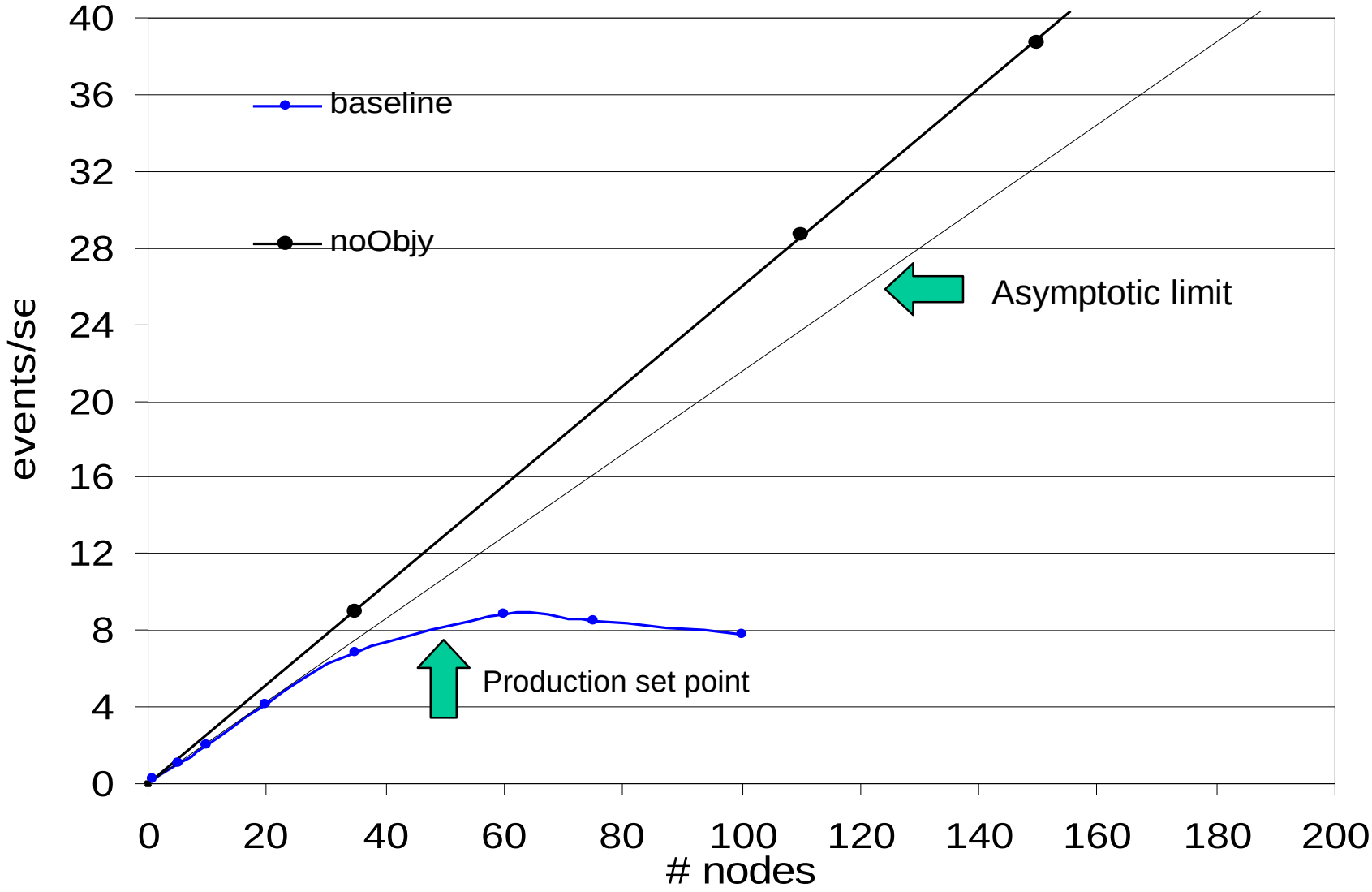


Exp. Characteristics

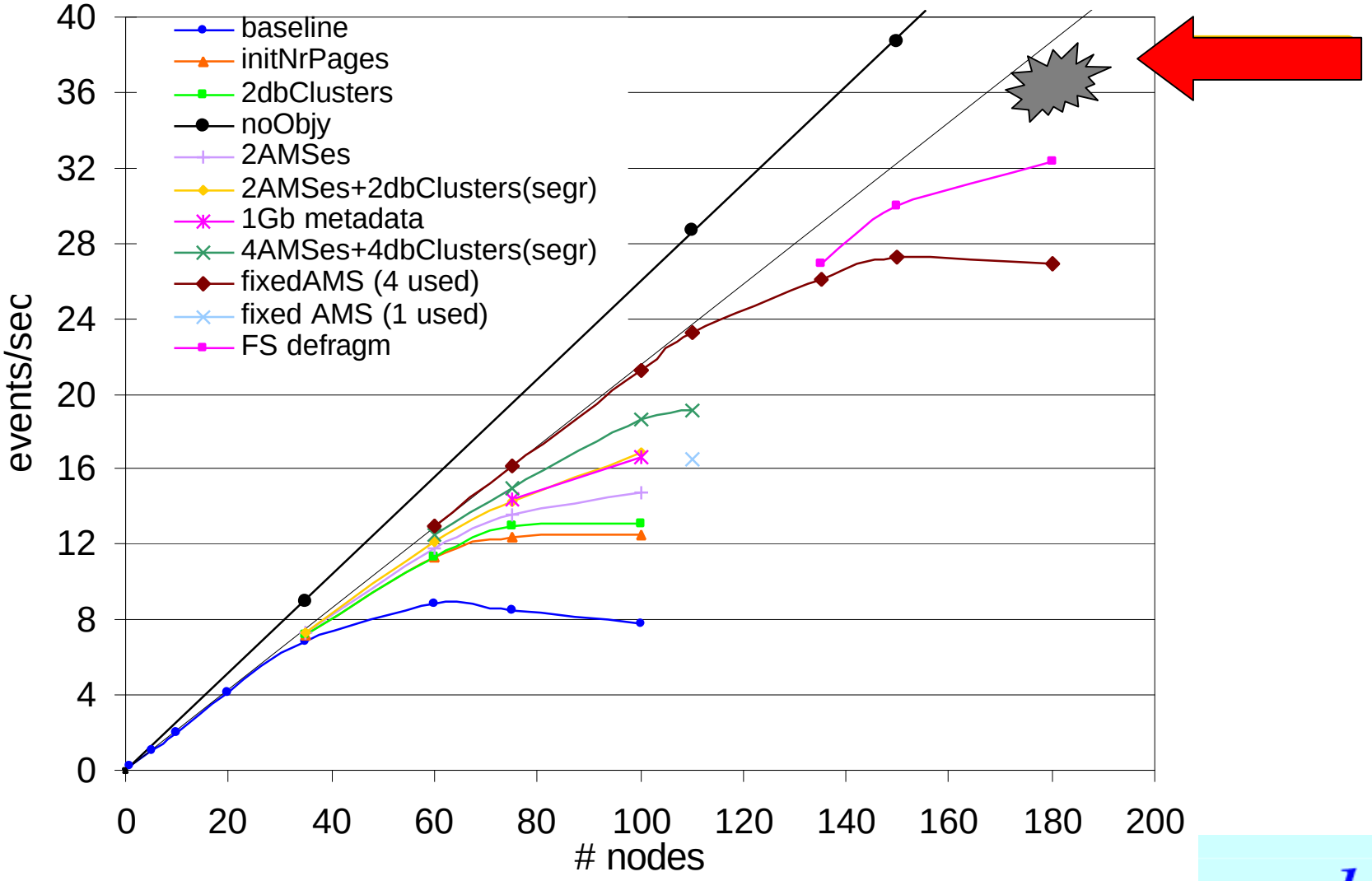
(D.Quarrie-BaBar)

Characteristic	*Size*	
No. of Detector Subsystems		7
No. of Electronic Channels		~250,000
Raw Event Size		~32kBytes
DAQ to Level 3 Trigger	2000Hz	50MByte/sec
Level 3 to Reconstruction	100Hz	2.5MByte/sec
Reconstruction	100Hz	7.5MByte/sec
Event Rate		10^9 events/year
Storage Requirements (real & simulated data)		~300TByte/year

First Results Results



Results so far



Knobs to twiddle (tested)

- Minimize catalog operations
- Separate conditions DB server
- Separate catalog server
- Tune AMS server
- Client file descriptors
- Client cache sizes
- Initial container sizes
- Transaction lengths
- TCP configuration
- Multiple AMS processes
- Database clustering
- Autonomous partitions
- Disable filters
- Singleton Federations
- Veritas Filesystem optimization
- Decrease payload per event
- LM starvation?
- Loadbalance across datamovers
- More datamovers
- Database pre-creation
- Gigabit lockserver
- Caching handles
- Local bootfile
- Unlock instead of mini-transaction
- Run OPR with no output
- Run on shire to bypass AMS

My Impressions

- BaBar x 2-3 ≈ LHCb
- BaBar x n ≈ ATLAS or CMS (n < 10)
- ALICE is different

- Objectivity/DB works, but
  - BaBar has 6-12 people dealing with persistency

It's a nice tale that commercial SW does not need maintenance

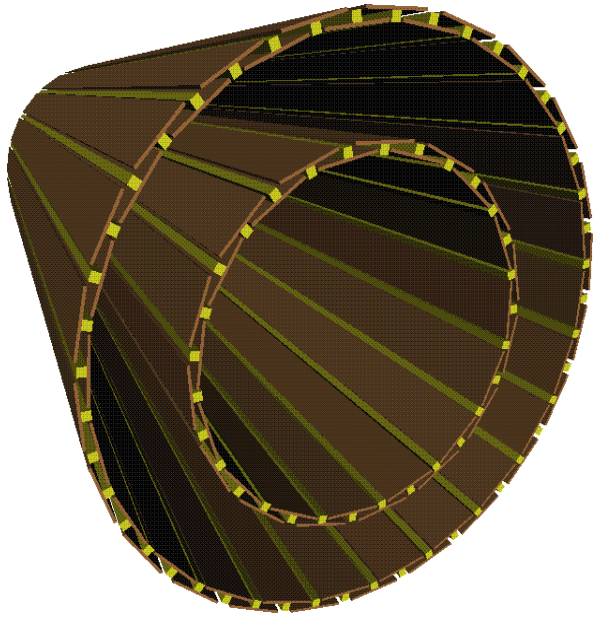
Simulation

- Rapporteur talk
  Matthias Schroeder
- Report from GEANT4 workshop
  John Apostolakis
- Base Classes for Simulation in ALICE: Generators and Segmentation
  Andreas Morsch

J.Apostolakis
Geant4/IT

…another Geant 4 Overview

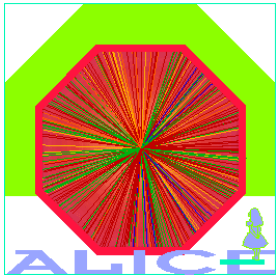
- Geant4: brief history
- Overview of Geant4
  - kernel's power
  - additional abilities
- Developments at Geant4 workshop (20-24 Sept. 1999)
- Status and plans



We have to put hands on ourselves….

A.Morsch
ALICE

Simulation Components



Feasibility Studies

Physics Signals

Fast Simulation

Detector Simulation

Physics Performance

Example: Fast Simulation

Segmentation

Detector Classes

Event-Generators

Particle Stack

MC

Example: Digitisation
Cluster Finder

Detector Performance

Tracking Codes

Hits

Digits

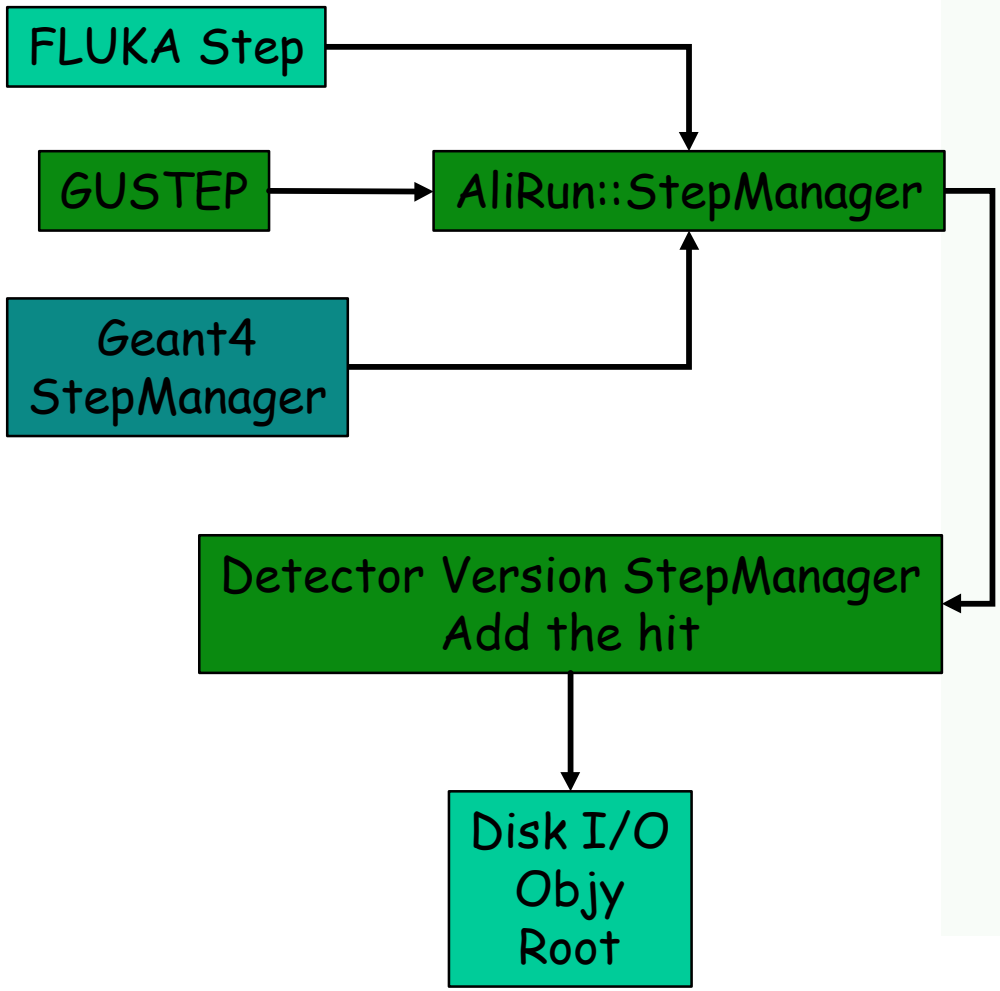
Reconstruction

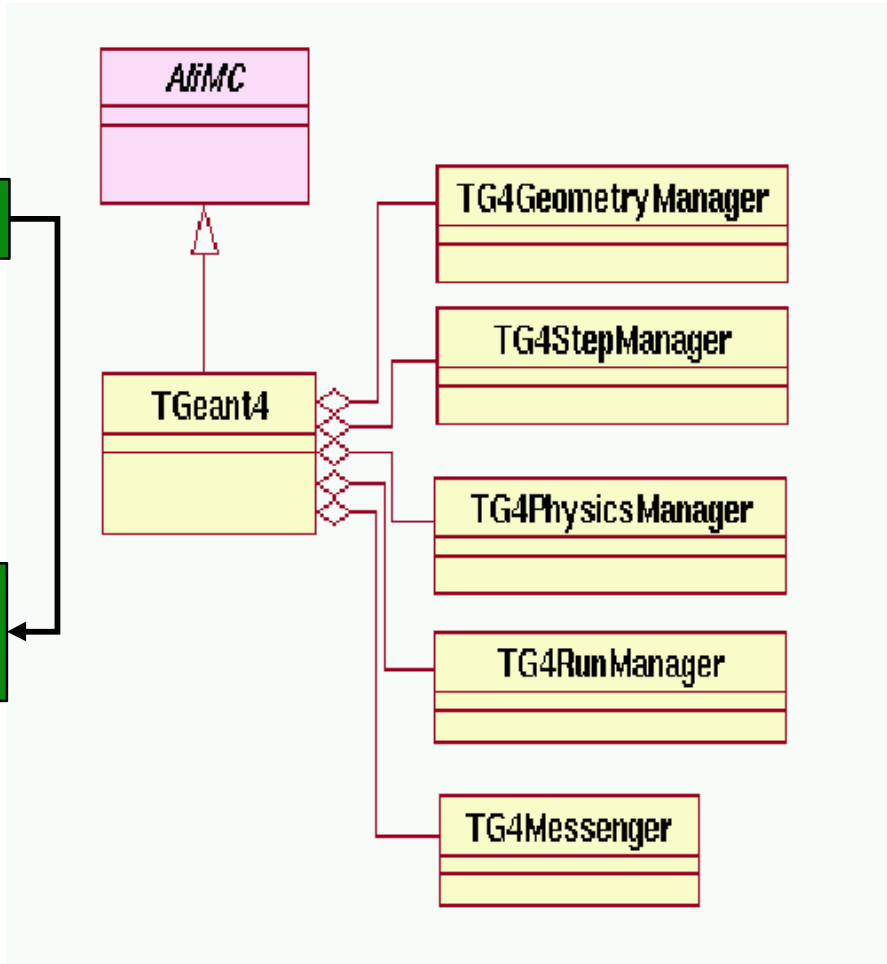
Reconstruction Optimisation

LHCb

A.Morsch
ALICE

New Tracking Schema





Round Table on Software Process

- Overview of Processes Currently Established
  Hans-Peter Wellisch
- Software Processes in BaBar
  Gabriele Cosmo
- Software Processes in Geant4
  John Apostolakis
- Panel Discussion: ALICE, ATLAS, CMS, LHCb, BaBar, Geant4

Architecture

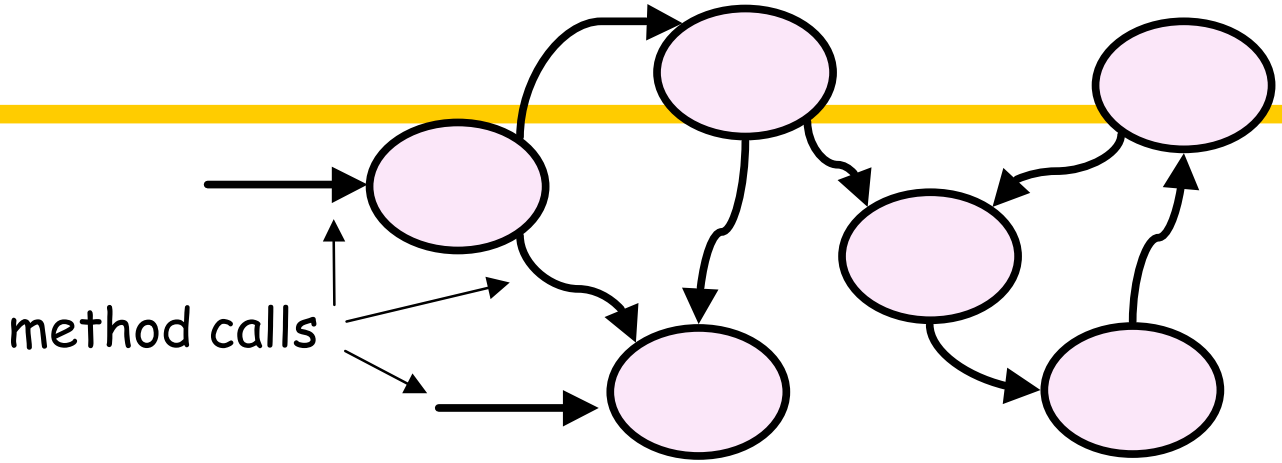
- Rapporteur talk RD Schaffer

RD.Schaffer
ATLAS

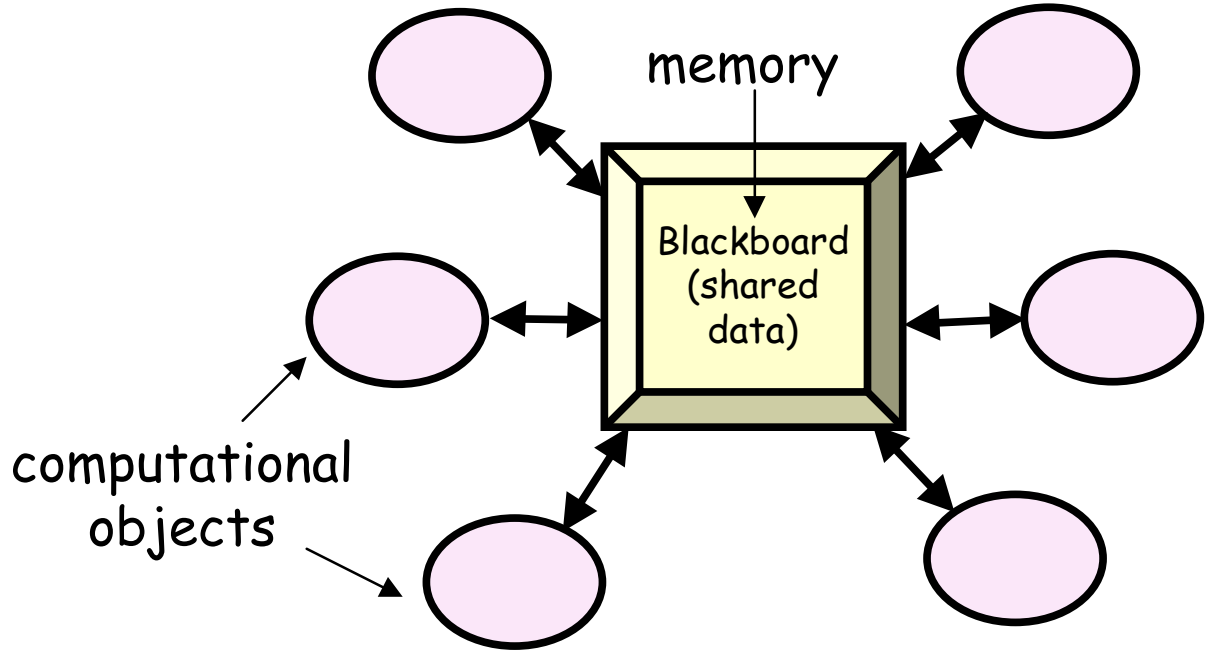
Architecture Styles

- Control/data interaction issues
  - Are control-flow and data-flow topologies isomorphic?
  - If isomorphic, is the direction the same or opposite?
- Useful examples of architectural patterns
  - Call-and-return styles — **data abstraction** (object-oriented)
  - Data flow styles — **pipe-and-filter systems**
  - Data-centered repository styles — **blackboard**
  - Interacting process styles — **implicit invocation**
- Note: few systems are purely any one of these!

RD.Schaffer
ATLAS



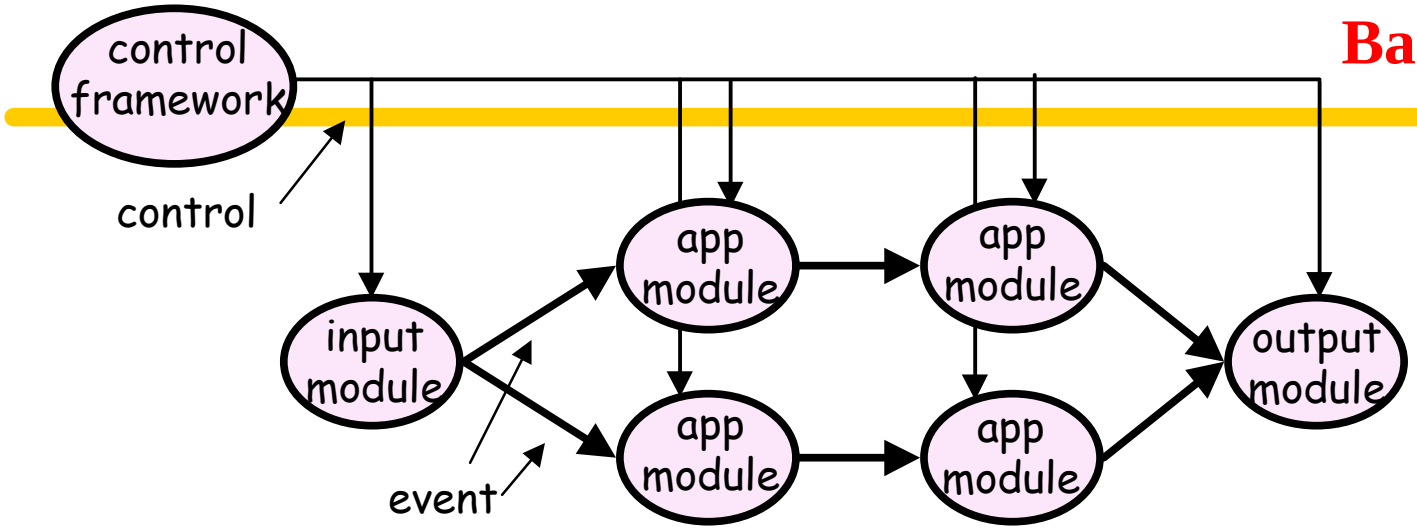
Call-and-
return style



Data-centered
repository style

RD.Schaffer
ATLAS

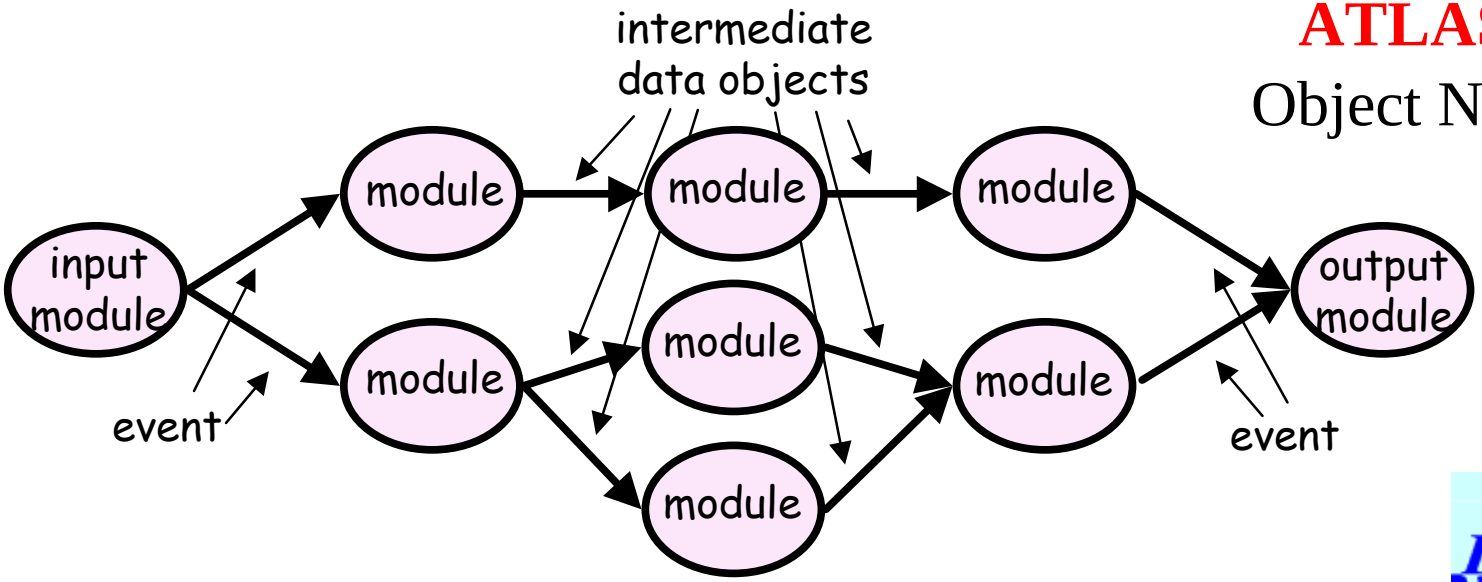
BaBar and **CDF**



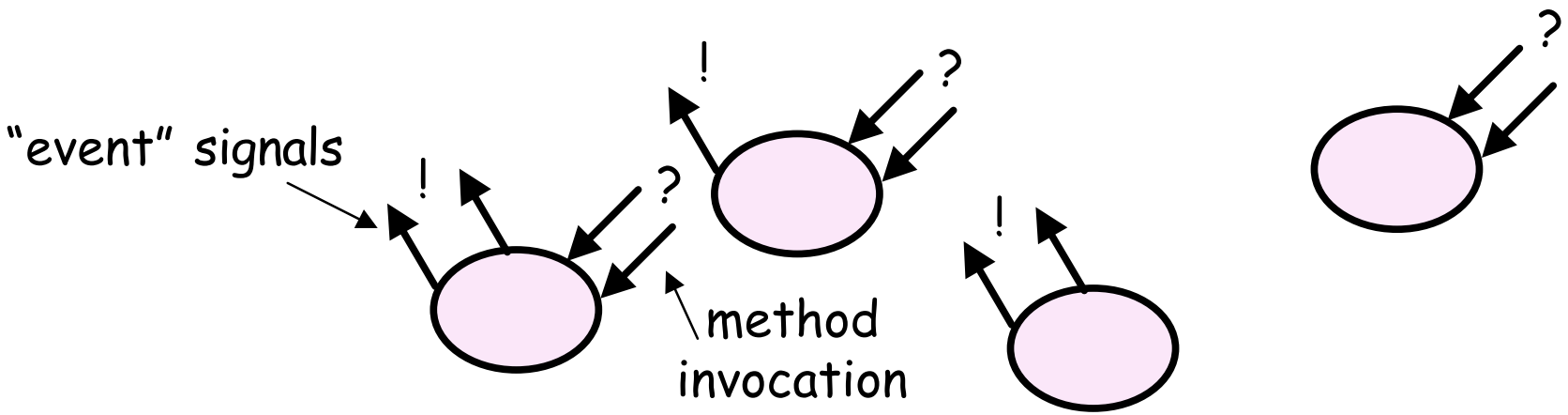
pipe-and-filter style

ATLAS

Object Network



RD.Schaffer
ATLAS

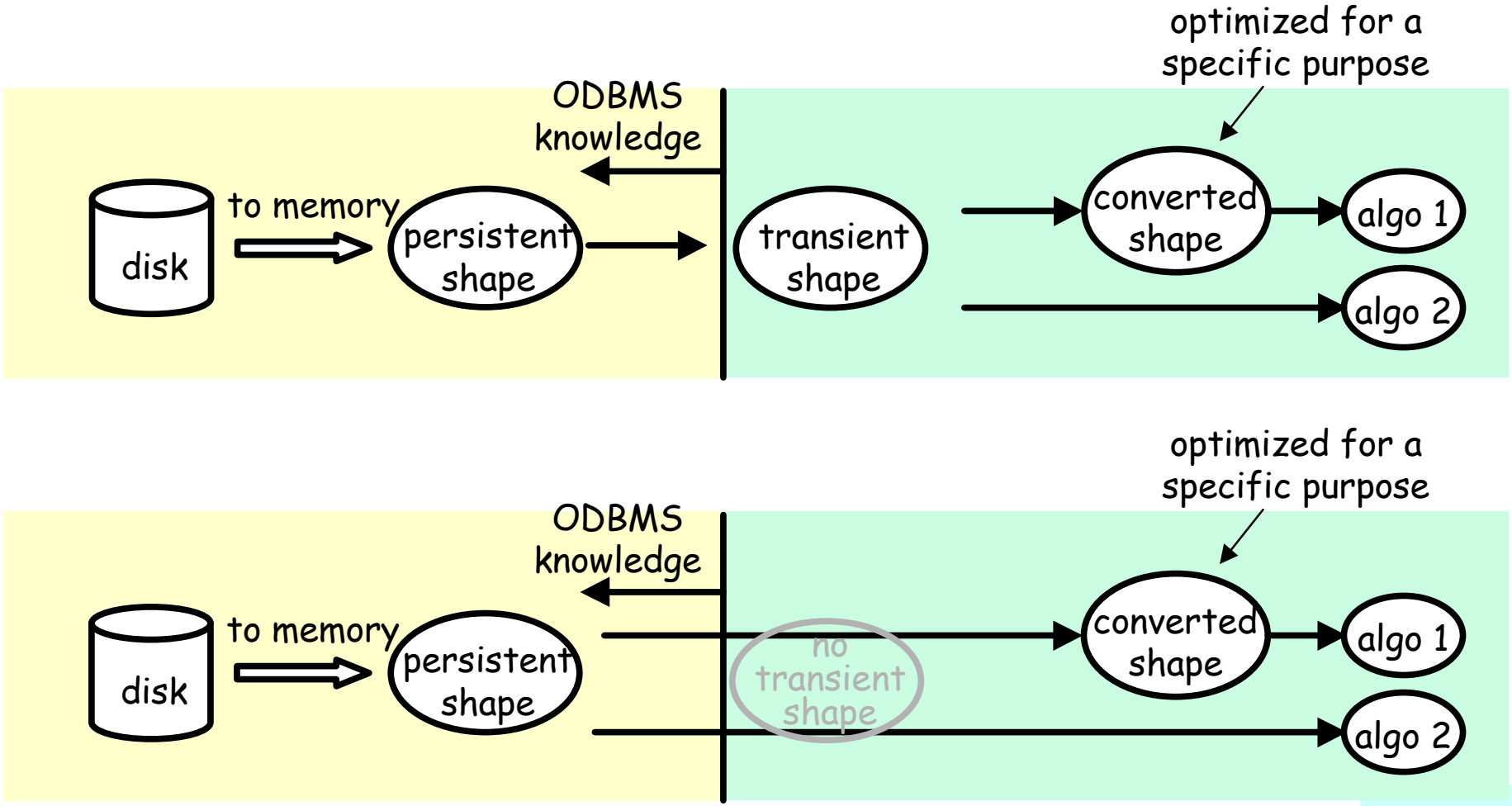


Independent
reactive objects

CMS

RD.Schaffer
ATLAS

Architectural Issues of Persistency



Data Analysis

- Rapporteur talk
  Guy Wormser
- Migration from Fortran to C++ and OO, as seen by the physicist
  Marek Kowalski
- The WIRED experience
  Julius Hrivnac
- The JAS experience
  Michael Ronan

G. Wormser
LAL Orsay

Historical perspective : PAW

- Very large ‘ productivity boost ’ in the physicists community with the introduction of a universal analysis tool program **PAW**
  - very easy to use , available everywhere
  - Ntuples, MINUIT, presentation package
  - fortran interpreter
  - macros/script (KUIP, .kumac)
- No integration within experiments framework
  - No overhead!
  - But not possible to benefit from infrastructure

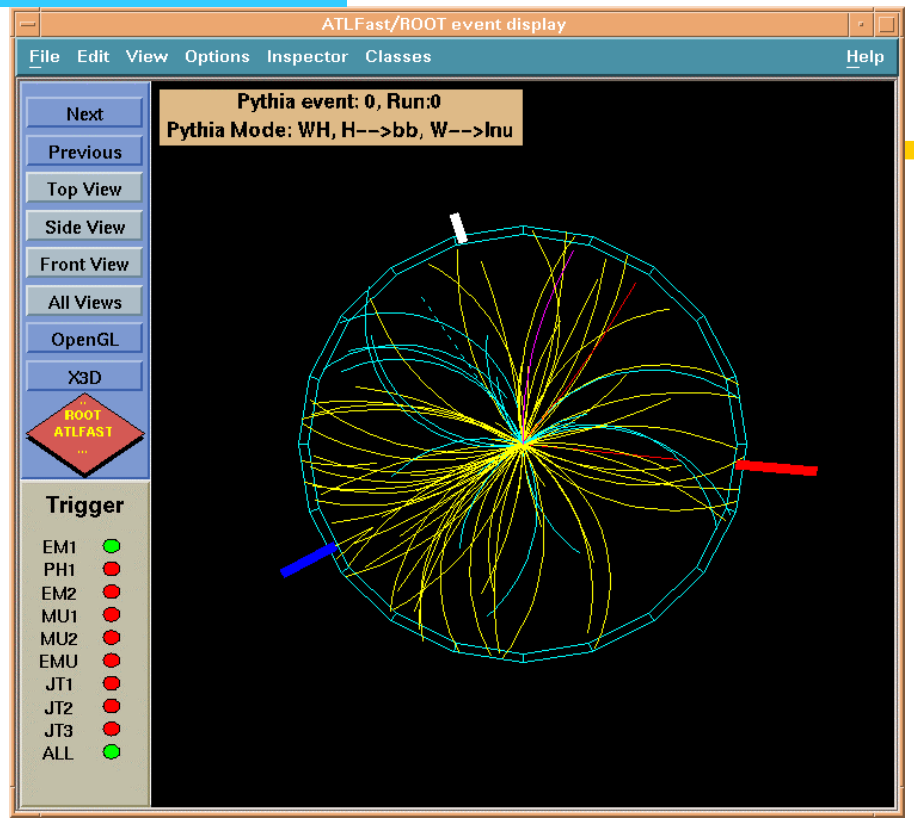
G. Wormser
LAL Orsay

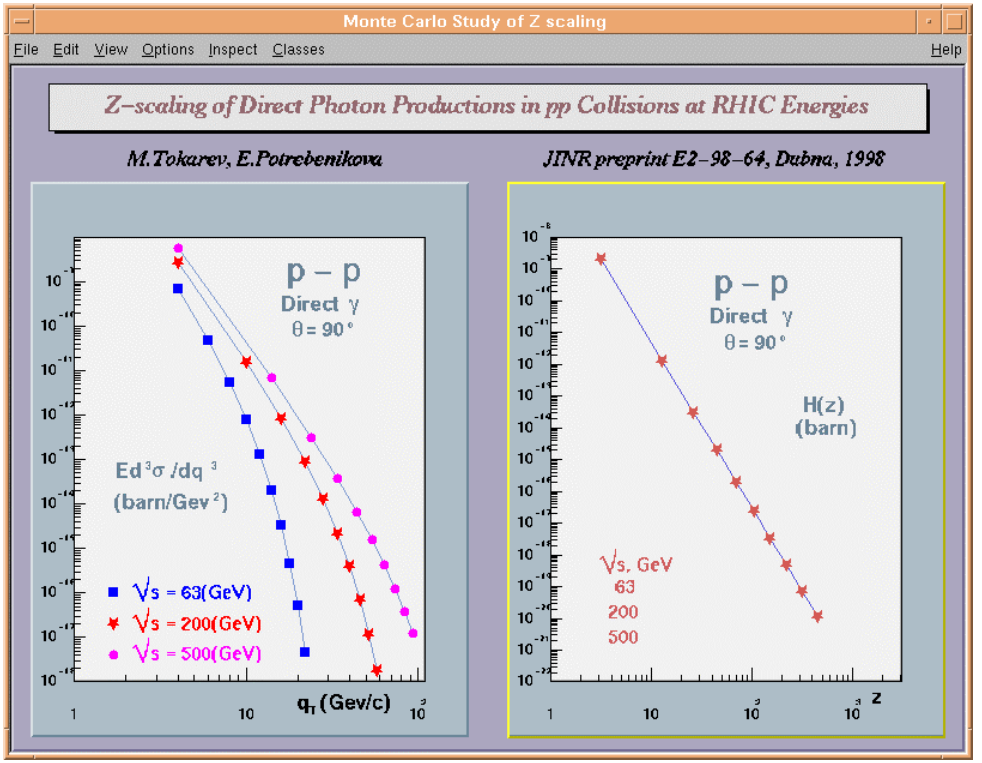
The new environment

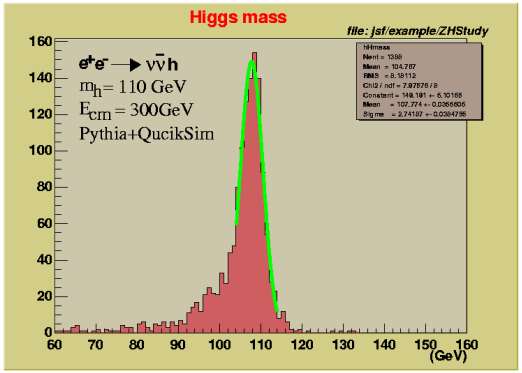
- OO Data structures (ROOT,Objectivity,etc)
- Analysis codes and tools in OO language
  - ➪ We want ‘ PAW_OO ’!
- Very large datasets
  - ➪ want Better integration within the framework
- Very powerful CPUs
  - ➪ Better interactivity

G. Wormser
LAL Orsay

ROOT examples

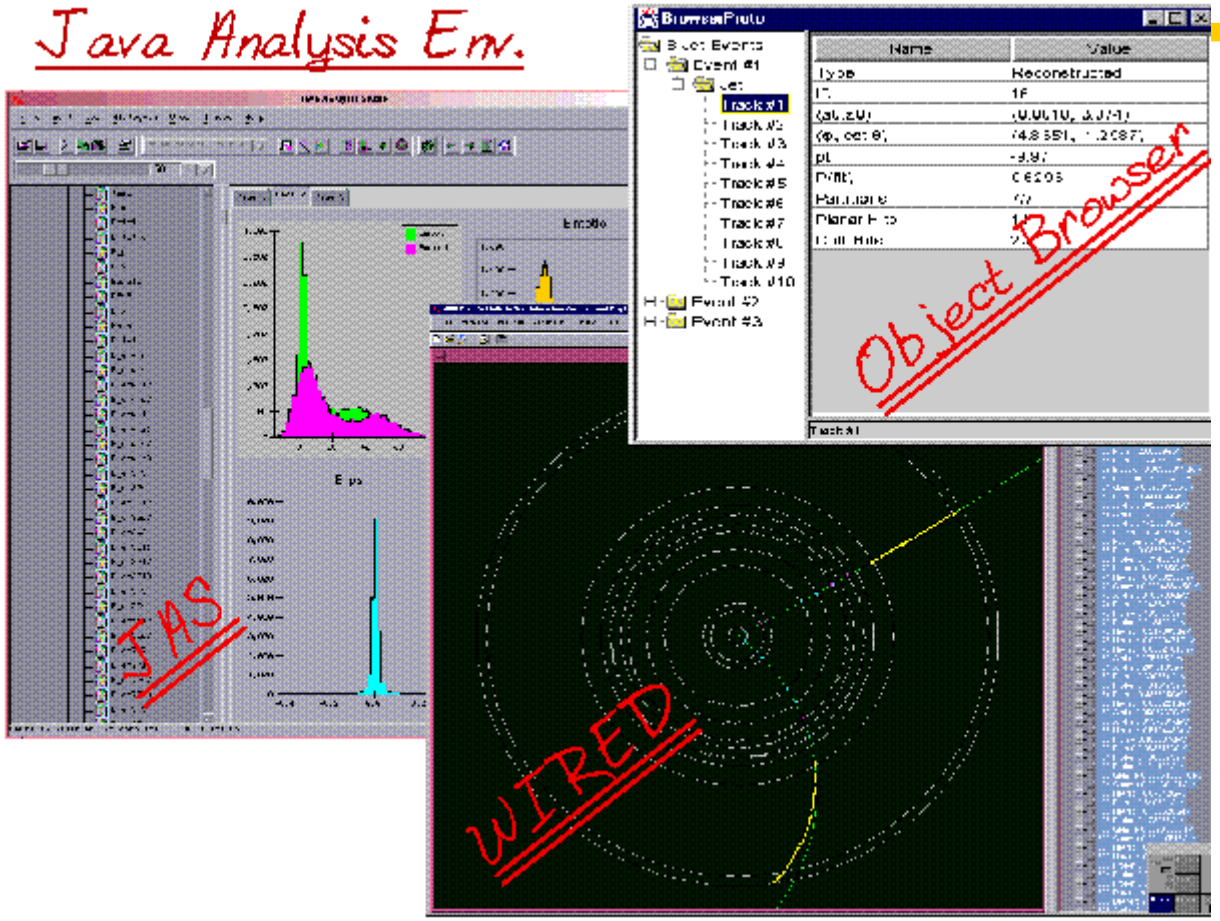






J.Hrivnac
ATLAS

Java is coming...



The JAVA Alliance

- Analysis helpers in JAVA seem to become usable
  - Easily extendable
  - JAVA language is equivalent to scripting and better than kumac
- Tony Johnson/SLAC
  - JAS: HBOOK & Co. are just plugins to access data
- Mark Donzelmann/CERN
  - WIRED: Client - Server approach to event displays

M.Ronan
LBL

JAVA@NLC

- Started with JAVA, no legacy
- Full reconstruction for detector studies
  - "Detector is a bigger ALEPH"

Technology Tracking

- Pasta - Processors, storage, architectures
  Les Robertson
- Local Area Networking
  Jean-Michel Jouanigot
- Wide Area Networking
  Olivier Martin
- Data Grid projects
  Stewart Loken

Tech.Tracking
CERN/IT

Storage, CPU & Networks

Estimated cost in 2005

- **Processors: $0.75-1.60 per CERN-Unit**
- **Disks: $2-4/GB in 2005**
  - **data rate increases only with the linear density**
- **Tapes: $0.50 per GB for raw tape**
  - **reliable drives unlikely to go below $10-20K**
  - **robotics - >=$20 per slot (no improvement)**
- **LAN: 1000BT. NIC: $200, Switch port: $500 - $1000**
- **WAN: unforeseeable**
  - **expected growth at constant cost: 15-50 % / year**

LHCb