CALICE Data Processing
(From Raw to Physics Data)

Roman Pöschl
LAL Orsay

- Calice Testbeam Data Taking
- Data Management
- Event Building and Reconstruction Software
- Pros and Cons ...
- Requirements for Fermilab Data Taking
- Monter Carlo Tools
- Summary and Outlook

Prologue

This talk uses the CALICE data taking as an example and maybe model for other testbeam efforts within the ILC community

Therefore I don't consider it as a CALICE talk but rather as a general proposal for the handling of testbeam data and on how to integrate hardware and software efforts within ILC R&D

CALICE Testbeam Data Taking

CALICE collaboration is preparing/performing large scale testbeam
Data taking between 1.August and 31.October 2006

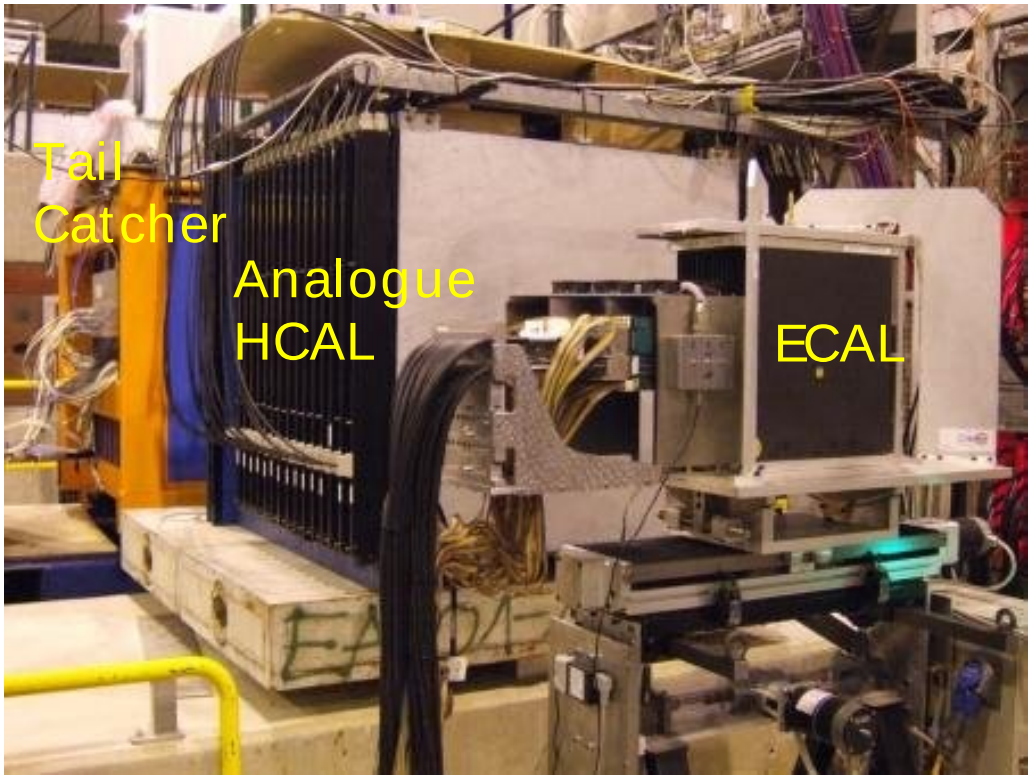
Testbeam program poses software/computing "challenges"

- Data processing from Raw Data to final Clusters in a coherent way

- Handling of Conditions Data Detector Configuration Calibration, Alignment etc.

- Comparison with simulated data
'Physics' Output

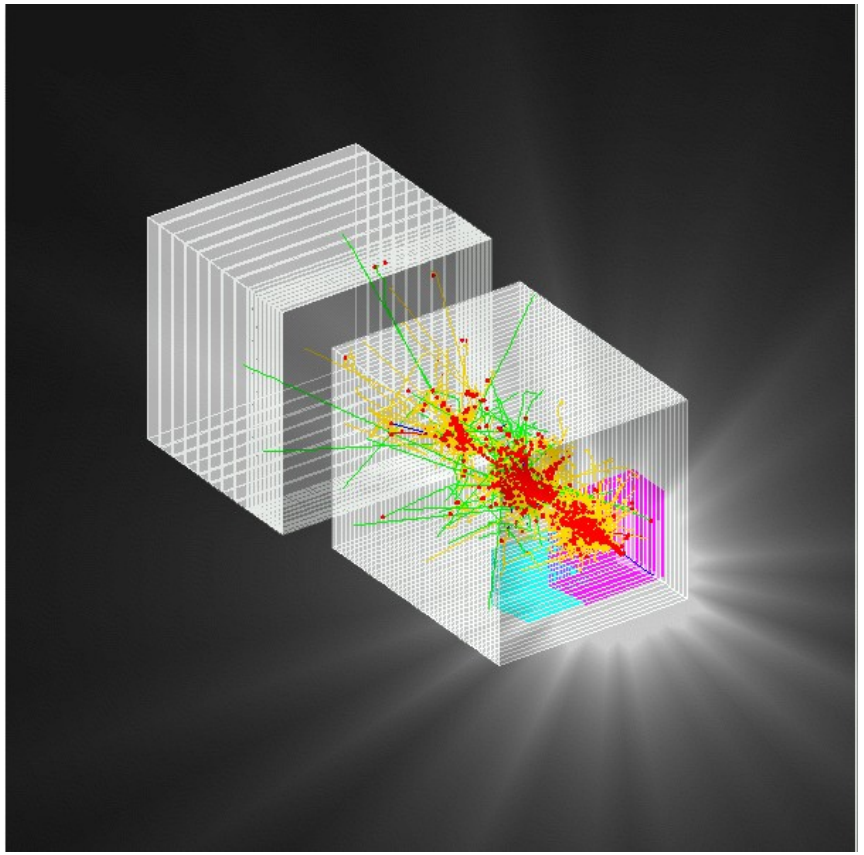
Testbeam Setup at CERN



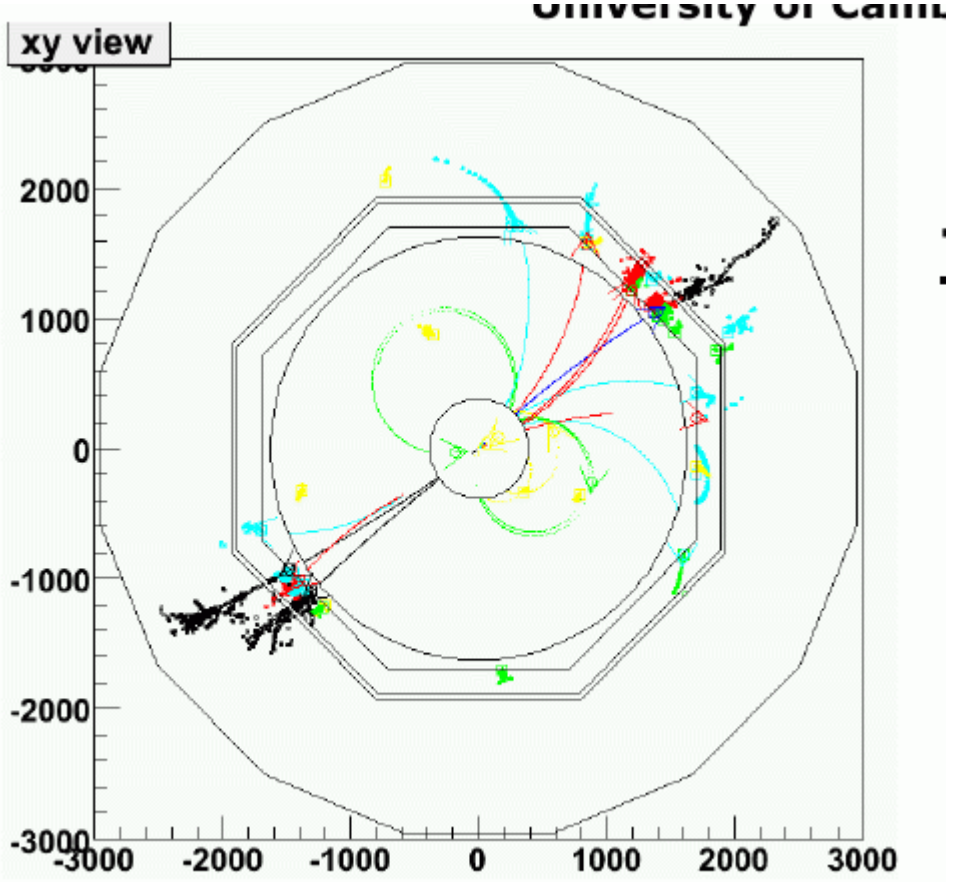
O(15000) calorimeter cells readout by Calice DAQ
No Zero Suppression
r/o speed 5 Mbyte/s continously

Development of Particle Flow Algorithms

... in the Testbeam



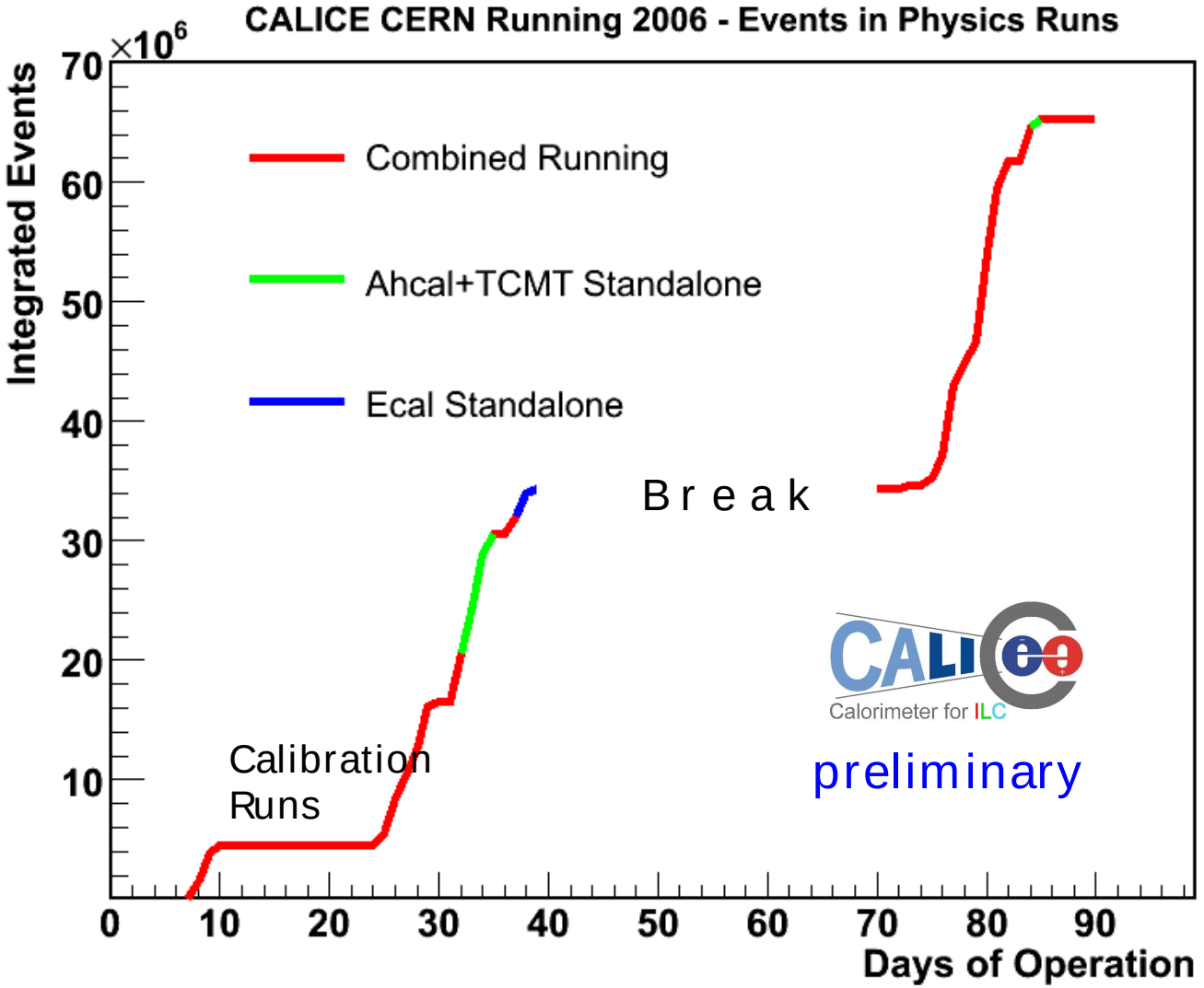
... for the real thing...



Integration of hardware and software R&D

Algorithms to be developped in detector simulation studies and applied to testbeam data and vice versa

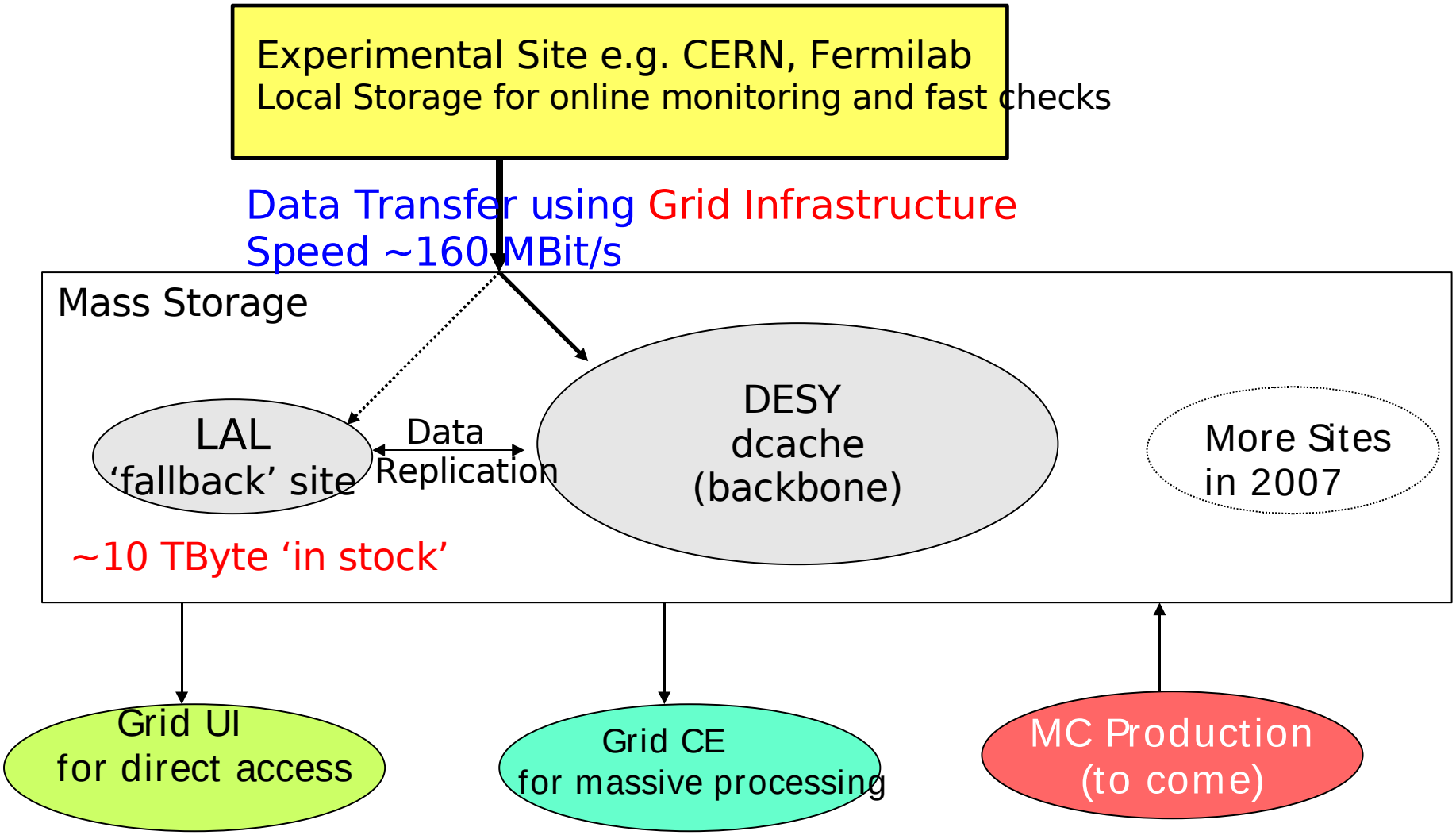
CALICE - CERN Data taking 2006



~65 Mio Events
in 'Physics' Runs
+
O(35 Mio). Muon
Calibration Events)

Efficient and fast
way of data distribution
and processing ?

Data Handling and Processing



Data available to whole collaboration ~20 Min. - 1h after run end
Data access independant of experimental site
Grid is the only 'environment' where all data are available

Why using the Grid ?

Neither ILC nor Calice nor other R&D projects
have an 'experimental center' like CERN, DESY etc.
and maybe will never have

World wide distributed R&D effort requires distributed computing

- Easy sharing of data by common data storage accessible by everyone from everywhere

- Exploiting the Grid allows for quick data processing, e.g. Several reconstruction iterations for calice testbeam data

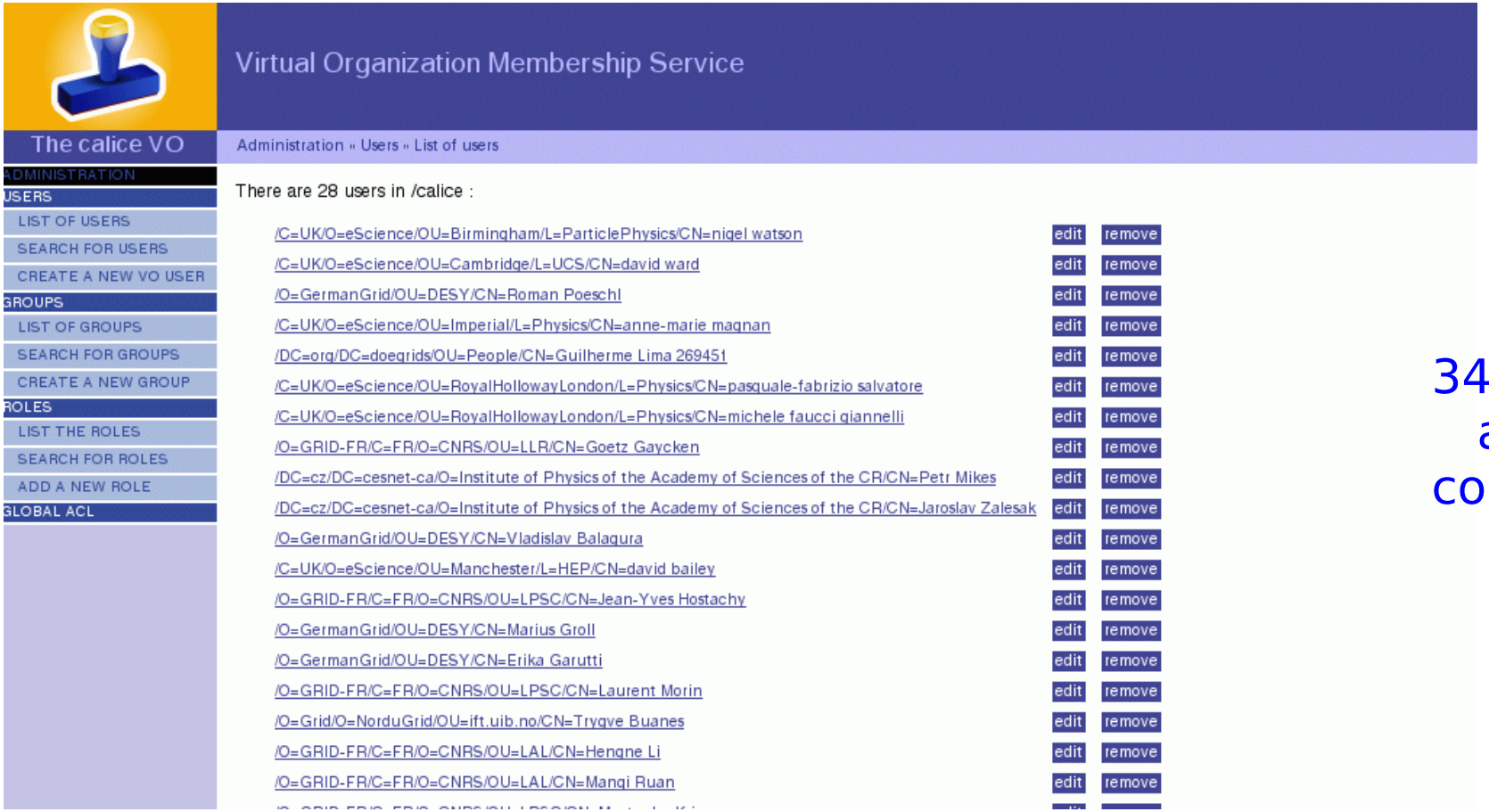
- Large simulation effort to come for the ILC requires large computing ressources
  Again no experimental center, the potential experimental centers like DESY, CERN Fermilab et al. have identified the Grid as computing platform
  General strategic decision by HEP community and science politics to invest in Grid computing

- Exploring the Grid can be regarded as an engineering/R&D effort for the ILC just as hardware development or simulation studies (which in turn demand significant computing power)

Software (and computing) infrastructure is part of the ILC Project !!!!

The Virtual Organisation - vo calice

Hosted by DESY:
Page for registration is https://grid-voms.desy.de:8443/voms/calice



34 Members
and
counting ..

VO Manager: R.P./ LAL, Deputy: A. Gellrich/ DESY

The Grid in/for Calice

Large Data Volume => Significant Computing Ressources required
Decentralized Organization <=> Decentralized Computing

Virtual Organization calice

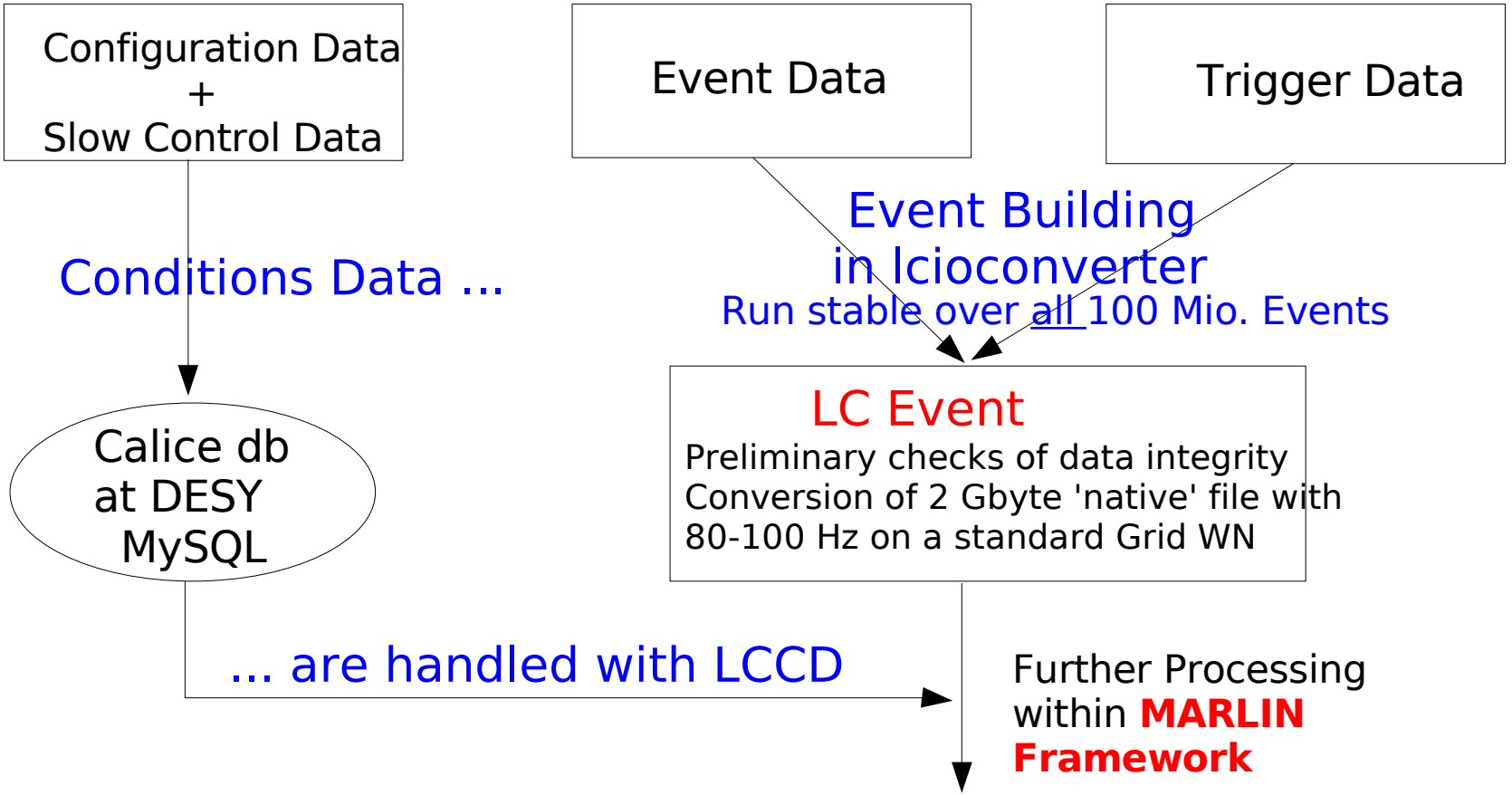
Supported by:	DESY Hamburg	Hosting, Computing and Storage
	LAL	Computing and Storage
	LLR	Computing (Storage to come)
	DESY Zeuthen	Computing and Storage
	Imperial College	Computing and Storage
	cc in2p3 Lyon	Computing (Storage to come)
	Cambridge	Computing and Storage
	Institute of Physics	Computing and Storage
	Prague	(in preparation)
	University College	Storage (Comp. To come)
	KEK	Computing and Storage (In preparation)
	Manchester	Computing and Storage (in preparation)
	CIEMAT Madrid	Offer Received
	Univ. Regina	Offer Received

Acknowledged EGEE project: https://cic.in2p3.fr

Conversion to LCIO

DAQ data types are converted/wrapped into LCIO on the basis of LCGenericObjects

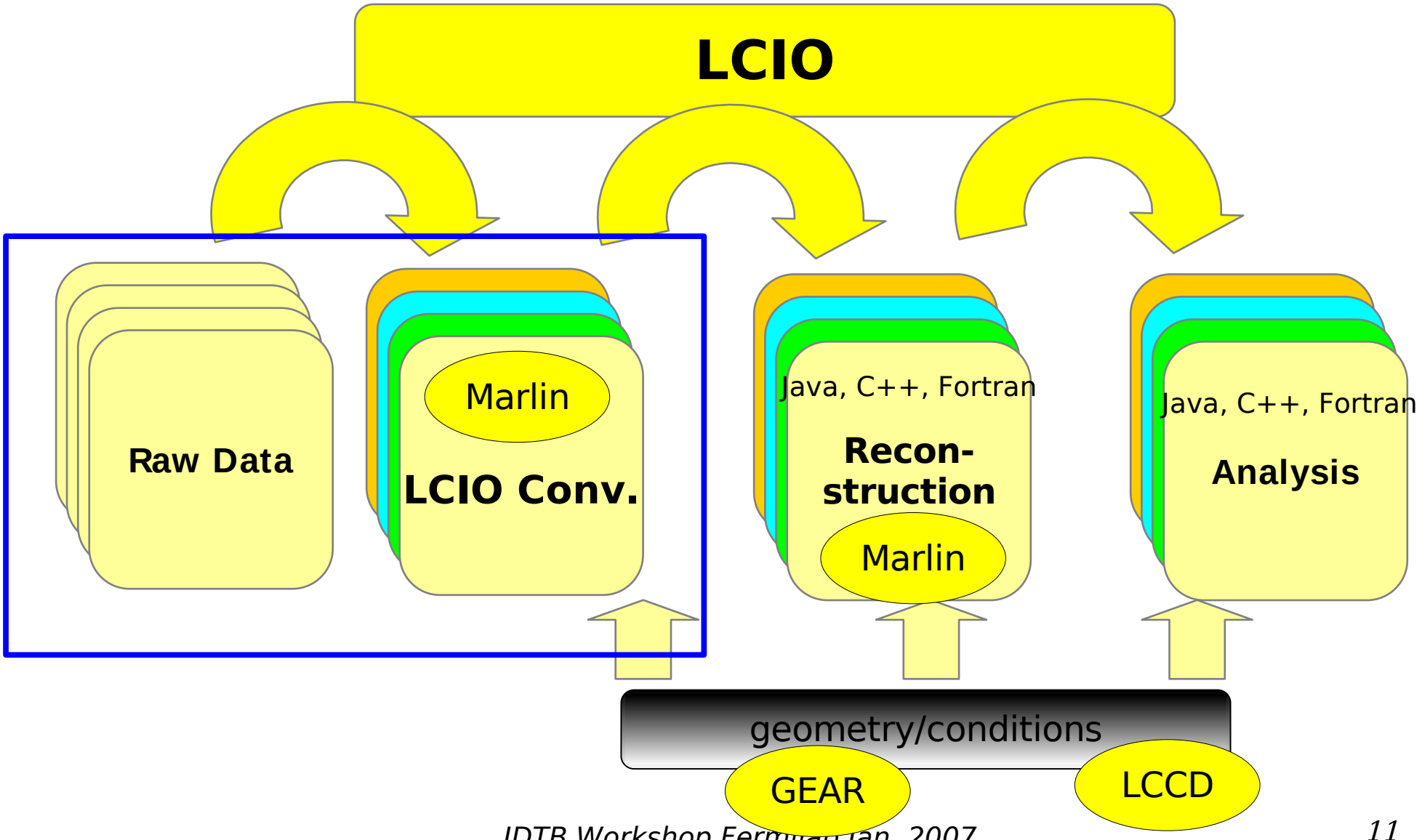
DAQ Data Files/Types



Remark: LCIO and ILC software framework is not needed to analyze calice data but using it delivers important input for future ILC s/w development
-> General ILC Concept for low level data handling

Software Tools for the LDC Study

Originally developped for simulation studies but applied in CALICE TB effort

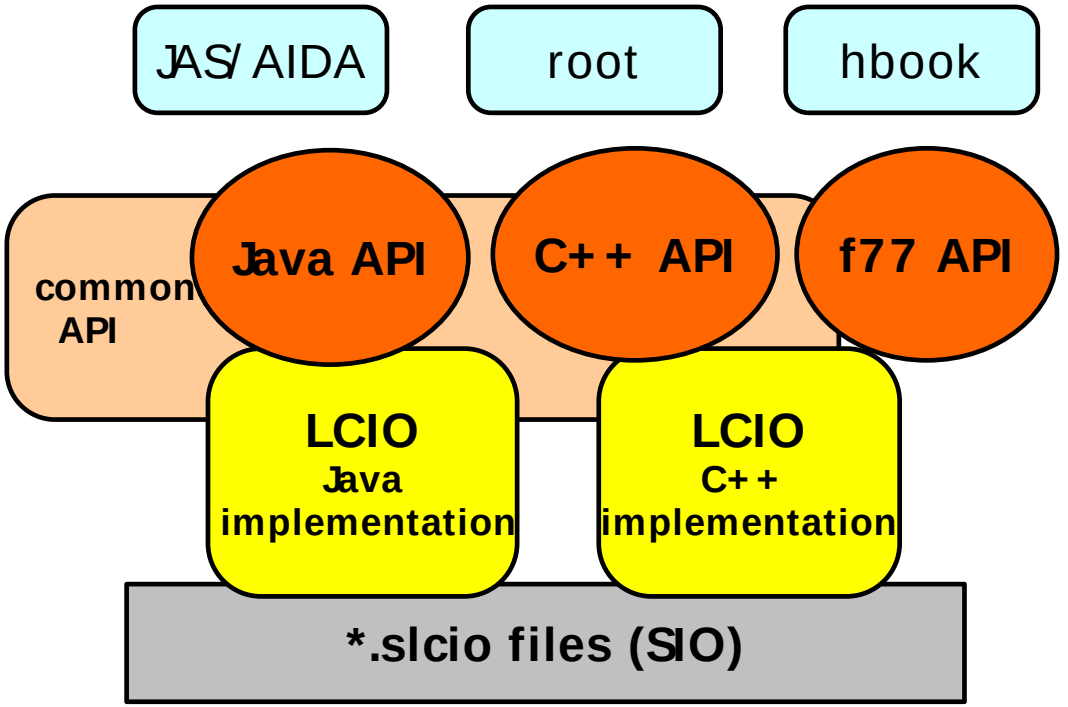


Brief overview on LCIO

- Common Project of DESY et SLAC
  - (Current) Basis of ILC software
- Principle properties
  - Java, C++ and f77 (!) API
  - Data Model for simulation studies simulation (and beam tests ?)
  - User code separated from concrete I/O format
  - No dependency on other packages

Standard for persistency & and data model of ILC Software

SW Architecture



Current version: **v01-08**

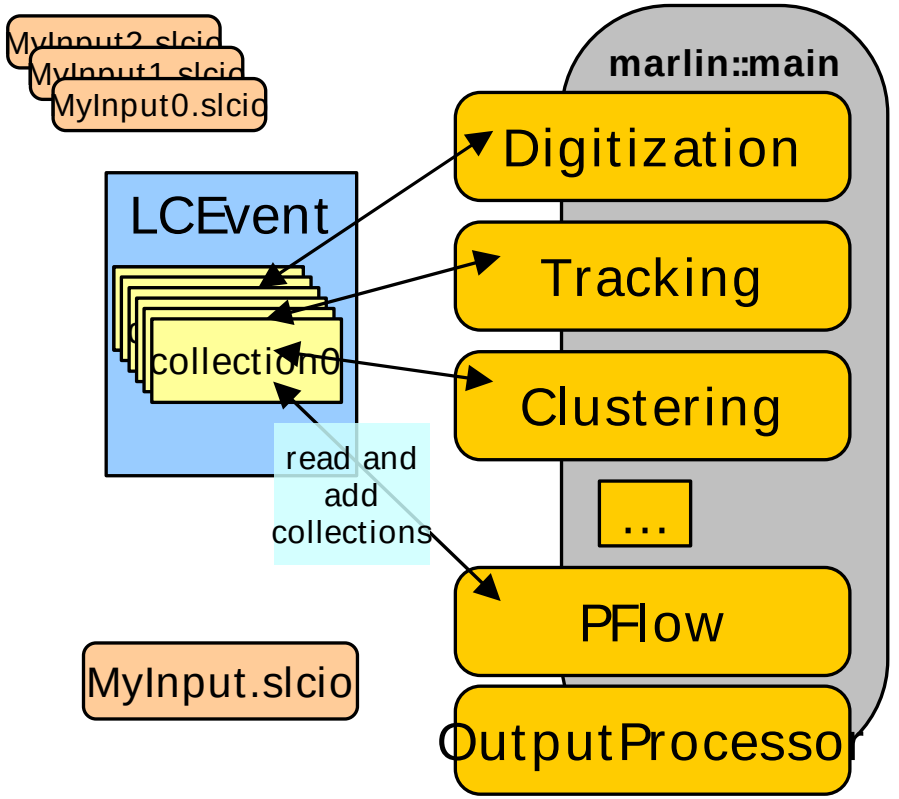
MARLIN

Modular**A**nalysis & **R**econstruction for the **L I N**ear Collider

- Modular Structure

  Language C++
- Ensemble of indepedant processors
- Provides main program
- Uses LCIO to propagate data through the sequence of processors

- Easily steerable
  - Program Flow (active processors)
  - User defined processor parameters
    - Processorwise or global
  - File I/O



Author: Frank Gaede DESY

Conditions Data Handling

- LCCD – Linear Collider Conditions Data Framework:
  - Software package providing an Interface to conditions data
    - database
    - LCIO files

  Author Frank Gaede, DESY

LCCD works and is heavily used within calice

The importance of conditions data (not only) for 'real' data renders the development of a fully functional cd data toolkit to be a fundamental !!! piece of the ILC Software
LCCD is first attempt into that direction

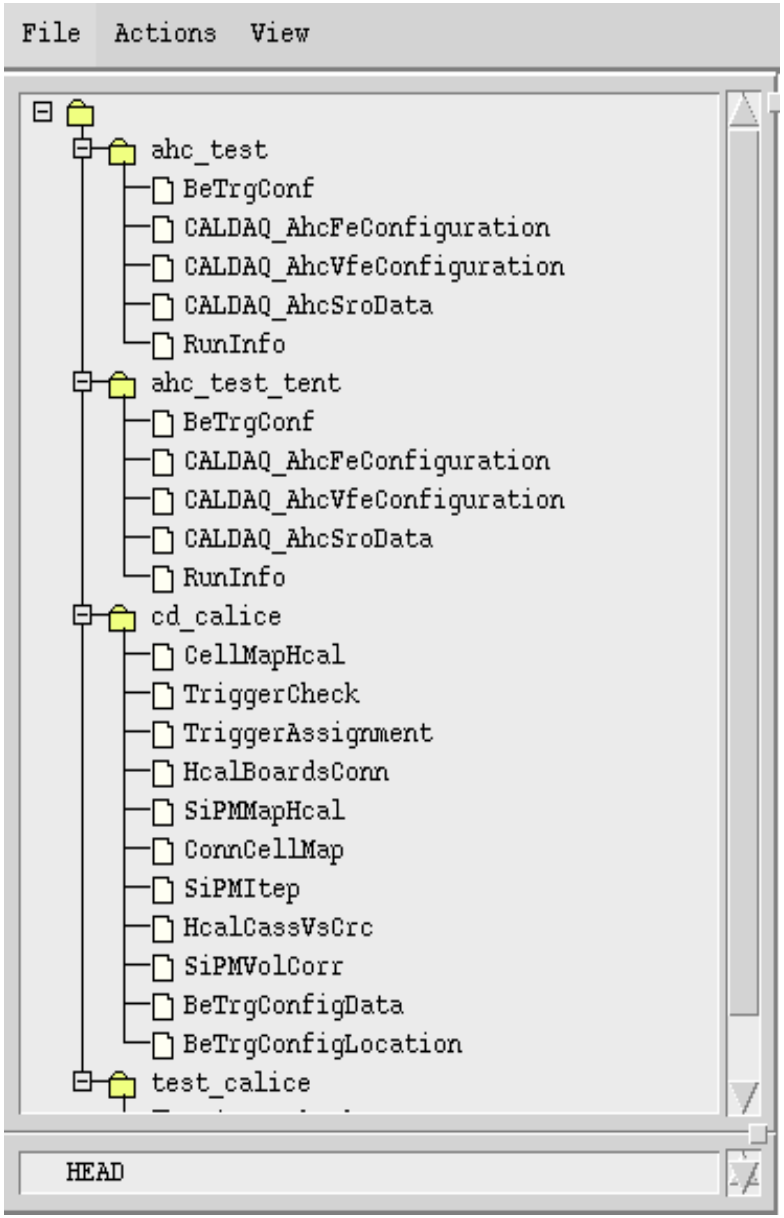
Issues to be addressed:
- Type safety
- Efficient storage and access to conditions data

  Browsing, convenient interfaces
- How to 'distribute' conditions data (e.g w.r.t to grid) ?

  BTW.: LHC does have some headache with that!

Testbeams are ideal environment to develop a working Conditions Data Handling before ILC starts

CALICE Database Hosted by DESY



Trigger Info: Assignment of triggerbits
Trigger Configuration
Info to validate Trigger information

Calibration Data

Cell Mappings: Relation electronic channel and geometrical channel
i.e. Cabling of devices

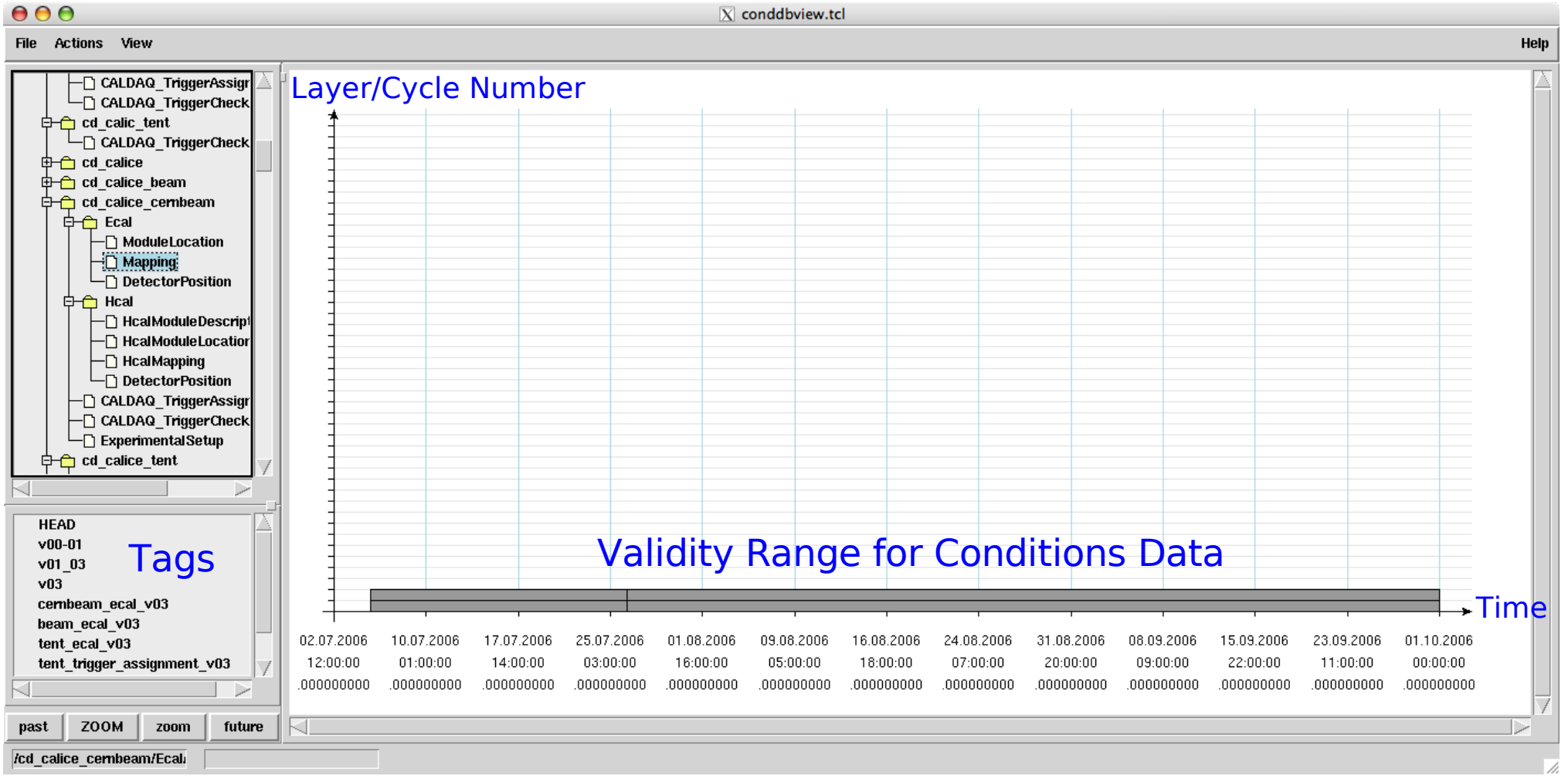
Hardware configuration during data taking.

CondDBMySQL package of Lisbon ATLAS Group
Validity Time stamp and tagging

Access via LCCD Interface

First attempt to visualize Conditions Data
(S.Schmidt, M.Schenk, R.P.)

Conditions Data in CALICE Database



CALICE Software Packages for Data Processing

- LCIO Conversion

  All data of 2006 (DESY/CERN) have been converted using the version v04-02-xx of the converter

- Reconstruction

  Many runs have been reconstructed for the Ecal using Version v04-02-01 of the reco package
  see lfn:/grid/calice/tb-xxxx/reco/rec_v0402

  'Unexperienced users' are encouraged to use these as an entry point to the data analysis

  Hcal Reconstruction v00-01-17 (S. Schmidt)
  Implementation into reco package nearly finished

- userlib (Common to all packages)

  Currently version v04-04-05

Pros and Cons using ILC Software for (Calice) Testbeam Data

Pros

Benefit from existing tools/features for/of ILC Software
e.g. LCEvent allows to gather information on event

Newcomers can work in one software framework for testbeam and physics studies

Define at an early stage of the ILC R&D the needs for a complete data processing
Coherent s/w concept at time of ILC Detector TDR
Not just guesswork!!!

It's in the spirit of the LDC CDR!!!!

BTW: The converted LCIO files can be analyzed on any OS (endianess) and on future 64bit architectures!!!

Cons

Need to wait for converted files
No quick turnaround in particular during development of DAQ and tests
Needs tight communication between DAQ and s/w developpers

Overhead generated by usage of ILC Software
- Slower program execution?
- Profiling of ILC Software needed

Source of (potential) errors unclear

Requirements to Fermilab

For data taking:

- CALICE Model of Data Processing requires excellent Network Connnectivity between 'Counting House' at Fermilab and DESY

  Equipment at CERN:

  (Calice) disk array for local buffering (cooling facility desired)
  Linux PC as grid-ui
  to manage data transfer – Will be replaced by dedicated (calice) server
  Gigabit switch (HP Pro Series)
  Optical Fibre to network star point in experimental area
  Connected to a Gigabit uplink
  Grid transfers benefit from 10Gbit Ethernet connection into DESY

- Support by Fermilab experts when setting up the chain needed

  Installing necessary devices and consultance on suited material
  Setup of (dedicated) datapath (like HTAR at CERN) ?
  Search for bottlenecks
  Tests well in advance of start of data taking (2 month before!!!!)

For data management and processing:

- Interface between different grid software (OpenScience and LCG)

- (In principle) no direct access to Fermilab mass storage needed

  ... but storing replicas on Fermilab's mass storage highly desirable

Outlook to 2007 (and beyond)

- CALICE will continue data taking with fully equipped detector in 2007
  at CERN and Fermilab (and DESY)
  20000 cells in r/o again w/o zero suppression

- Mass Production of MC will start

Expect to have 25 – 30 Tbyte of data in stock by the end of 2007
raw, converted and reconstructed data
MC files
Will grow beyond 2007!!!!

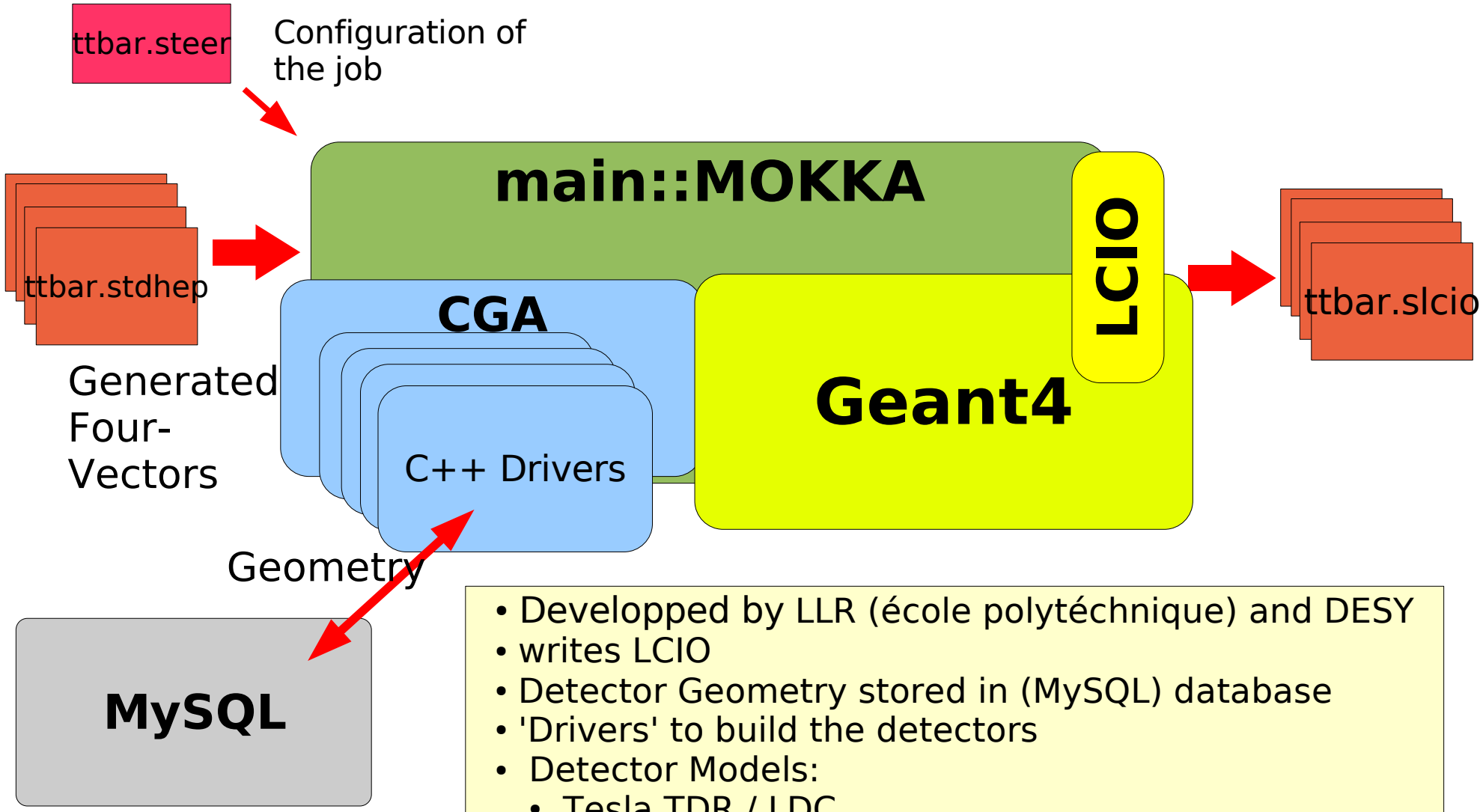
(My) Vision of CALICE Data Management

- Every country participating in CALICE should identify one site
  to store replicas of CALICE data
  Avoids network traffic and provides faster access to data

- Use local ressources (i.e. Storage) at smaller sites
  to store e.g. analysis output
  If files are visible by the grid your bright ideas can
  be made available easily to your colleagues

Fermilab might be the US site where replicas of all calice data are stored

Monte Carlo Tools - Mokka

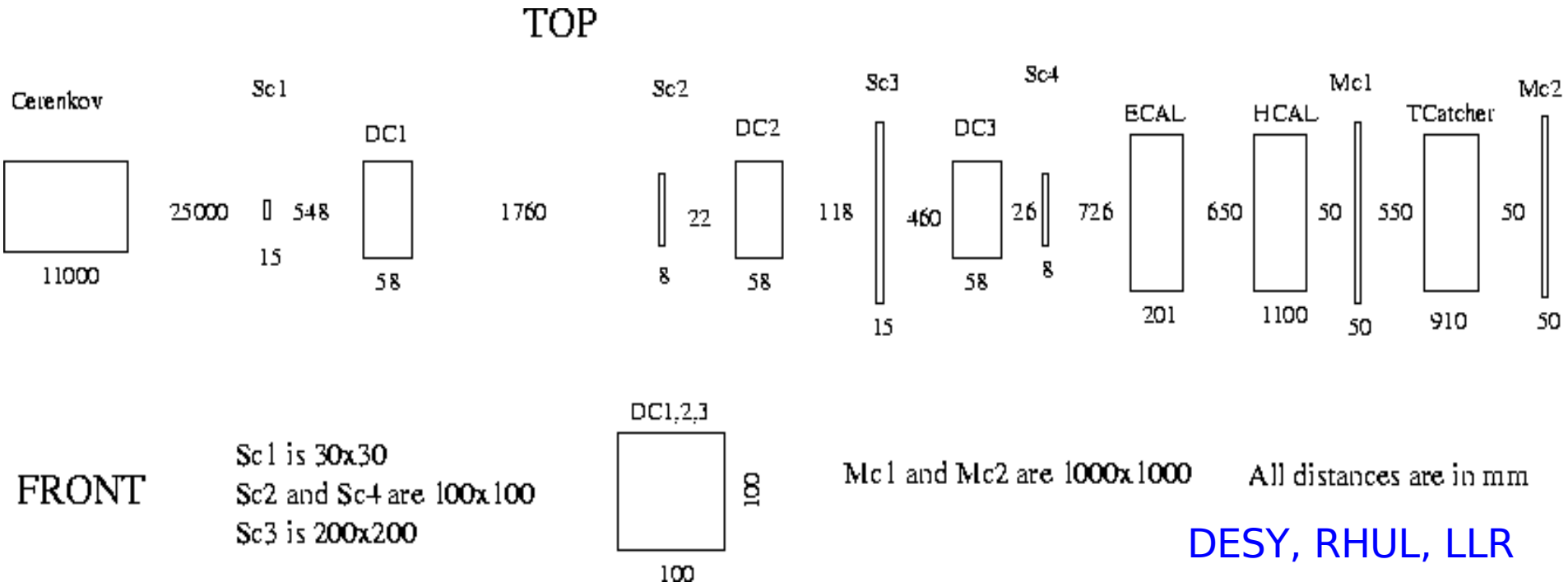


- Developped by LLR (école polytéchnique) and DESY
- writes LCIO
- Detector Geometry stored in (MySQL) database
- 'Drivers' to build the detectors
- Detector Models:
  - Tesla TDR / LDC
  - SiD
  - Calice Prototyps
- Used in LDC study (Europe)

Same package for Detector and Testbeam Studies – Integrated effort

A view to the Monte Carlo Branch

- Model for the simulation of the CERN test beam is available (in release 06-02 of Mokka)



DESY, RHUL, LLR

Will use grid for MC production

Estimation ~ 5000 kSI2kd for simulation of CERN data

Simulation will be followed by a digitisation step

Realized as Marlin Processors within Digisim Package

A.M Magnan, G. Lima

Summary and Outlook

- Calice uses ILC Software for processing of Testbeam Data

  ILC Datataking in a (big) nutshell
  Very important input for current and future developments of ILC Software
  Allows for stringent tests of the ILC Software concepts on a 'living' beast

- Calice uses systematically Grid tools

  First (and only?) R&D project within ILC effort
  24h/24h 7h/7h during CERN testbeam
  So far mostly for data management
  CPU consumption still tiny but will grow fast when starting e.g. MC production
  Calice has benefitted a lot from close collaboration with and support by IT Divisions of DESY, CERN and LAL

- Experience with testbeam data clearly reveals the needs for a coherent concept to handle 'low level' data within ILC Software

  (Latest) Next generation R&D projects should be used to develop a complete data processing/handling strategy for the ILC.
  Avoid 'island' solutions and work on an integrated effort
  Fermilab as tentative R&D Center for ILC ideal place to support that idea

CALICE does not only hardware-prototyping but also 'computing prototyping' for the ILC

Looking forward to our term at Fermilab -> Autumn 2007!!!!!