Introduction to Physics Computing

Rudi Frühwirth
Institute of High Energy Physics
Austrian Academy of Sciences

CSC 2005, St. Malo

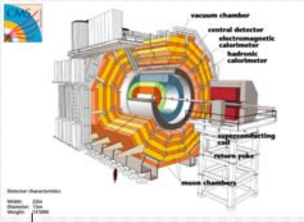
Outline of the lectures

- Introduction
- Filtering
- Calibration and alignment
- Event Reconstruction
- **Event Simulation**
- **Physics Analysis and Event Viewing**
- Data Flow and Computing Resources

What is Physics Computing?

- Input: A few Petabytes of data
- Output: A few hundred physics papers
- Data reduction factor of 10^7 to 10^8 !!
- How is it done?

It’s simple … is it?



```
Paper paper15
Data higgsdata
...
paper15=make_paper(higgsdata)
...
```

Electroweak phase transition in an extension of the standard model with a real Higgs singlet

Abstract

1. Introduction

We need…

- Millions of lines of code (C++,Java)
- Hundreds of neural networks (BNNs, not ANNs)
- Large infrastructure
  - PC farms
  - Database and storage systems
  - Distributed analysis facilities
  - The grid

What happens to the data?

- Data filtering,tagging and storage
- Calibration, alignment
- Event reconstruction
- Event simulation
- Physics analysis

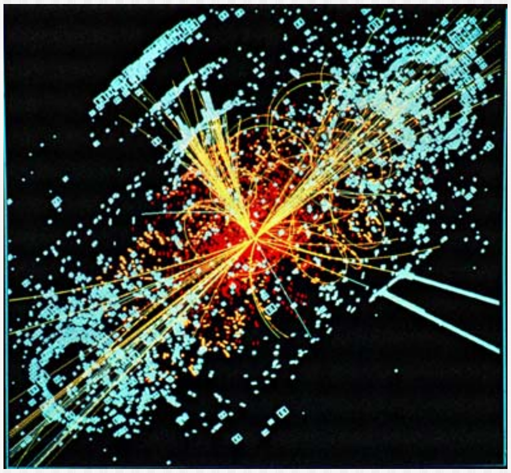
Step by step

- Each step involves some data reduction
  - data are compressed
  - data are discarded or ignored
- In each step the data get closer to be interpretable in physical terms
- Some steps are repeated many times until the output is satisfactory

The LHC challenge

- Very high event rate
- Large event size
- Large background of uninteresting events
- Large background in each event
  - many interactions in each BX
  - many low-momentum particles
- Large number of physicists doing analysis

The LHC challenge (cont)



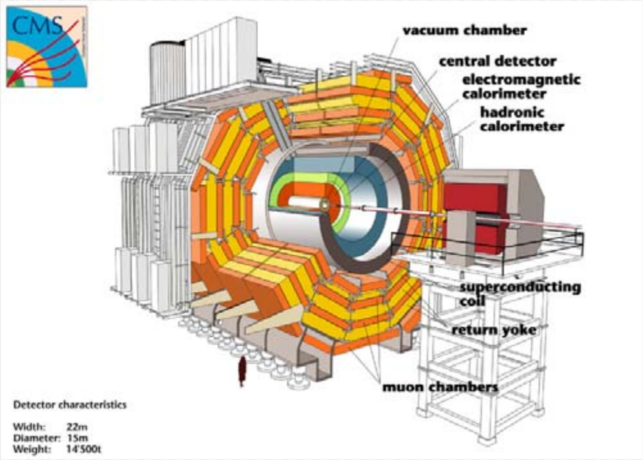
Data filtering

- Primary collision rate: 40 Megahertz
- Recording rate: 100 Hertz
- How is this achieved?
  - Multilevel trigger – chain of yes/no decisions
  - Very fast first level: (Programmable) hardware
  - Slower second level: Software on fast processors
- Reliable: Rejected data are lost forever
- Cautious: Do not lose new physics

Multilevel trigger

- Dead-time has to be minimized
- Many collisions can be discarded very quickly – Level 1
- Only the surviving ones are scrutinized more carefully – High Level Trigger(s)
- Triggers are tailored to specific physics channels (Higgs, top, WW, ZZ, …)

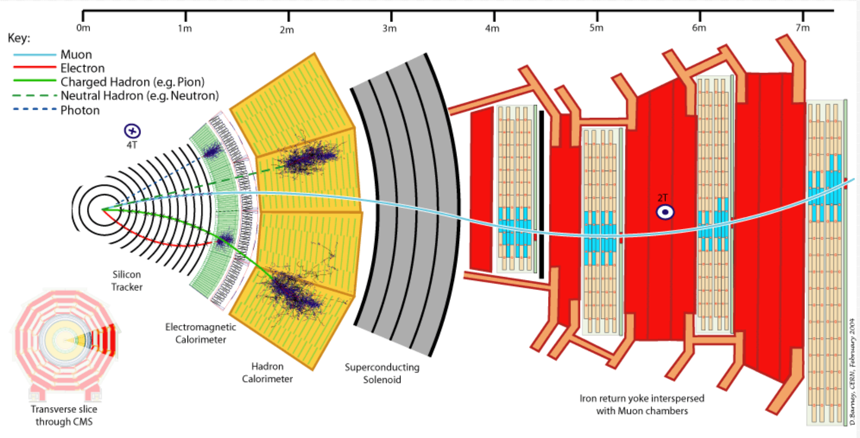
Example: CMS



What CMS subdetectors measure

- Inner tracker (pixels+strips)
  - Momentum and position of charged tracks
- Electromagnetic calorimeter
  - Energy of photons, electrons and positrons
- Hadronic calorimeter
  - Energy of charged and neutral hadrons
- Muon system
  - Momentum and position of muons

What CMS sub-detectors measure



CMS L1 trigger

- Input rate: 40 megahertz
- Output rate: 30 – 100 kilohertz
- Latency: 3.2 μs (128 BX)
- Pipelined, dead-time < 1%
- Available time for calculations: 1.25 μs
- 2 detector systems: muons/calorimeters
- 3 main steps: local/regional/global

CMS L1 calorimeter trigger

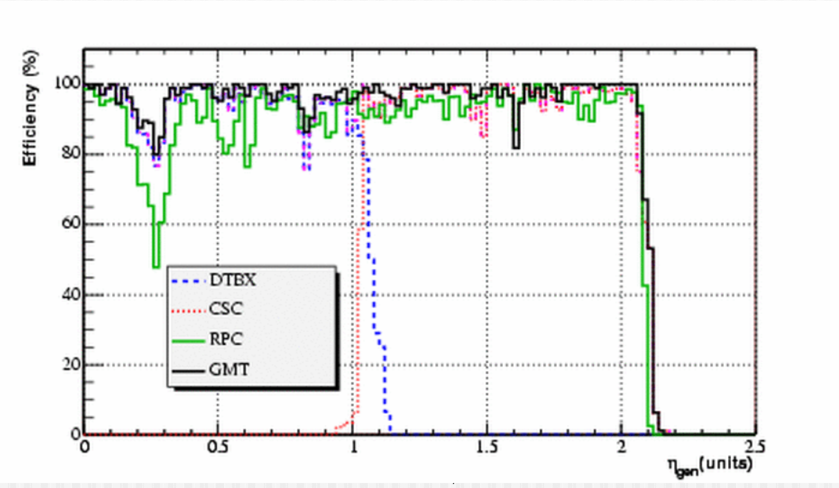
- Calorimeter trigger:
  - Two types of calorimeters
  - Local: Computes energy deposits
  - Regional: Finds candidates for electrons, photons, jets, isolated hadrons; computes transverse energy sums
  - Global: Sorts candidates in all categories, does total and missing transverse energy sums, computes jet multiplicities for different thresholds

CMS L1 muon trigger

- Muon trigger:
  - Three types of muon detectors
  - Local: Finds track segments
  - Regional: Finds tracks
  - Global: Combines information from all regional triggers, selects best four muons, provides energy and direction

---

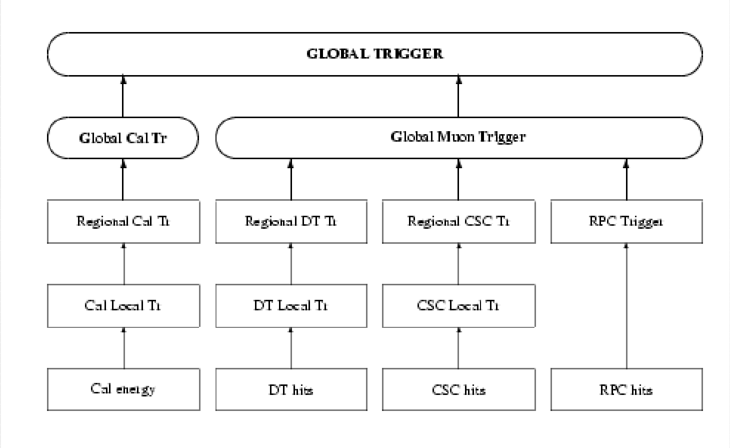
Efficiency of global muon trigger



CMS L1 global trigger

- Global trigger:
  - Final decision logic
  - 28 input channels (muons, jets, electrons, photons, total/missing E_T)
  - 128 trigger algorithms running in parallel
  - 128 decision bits
  - Apply conditions (thresholds, windows, deltas)
  - Check isolation bits
  - Apply topology criteria (close/opposite)

CMS L1 trigger

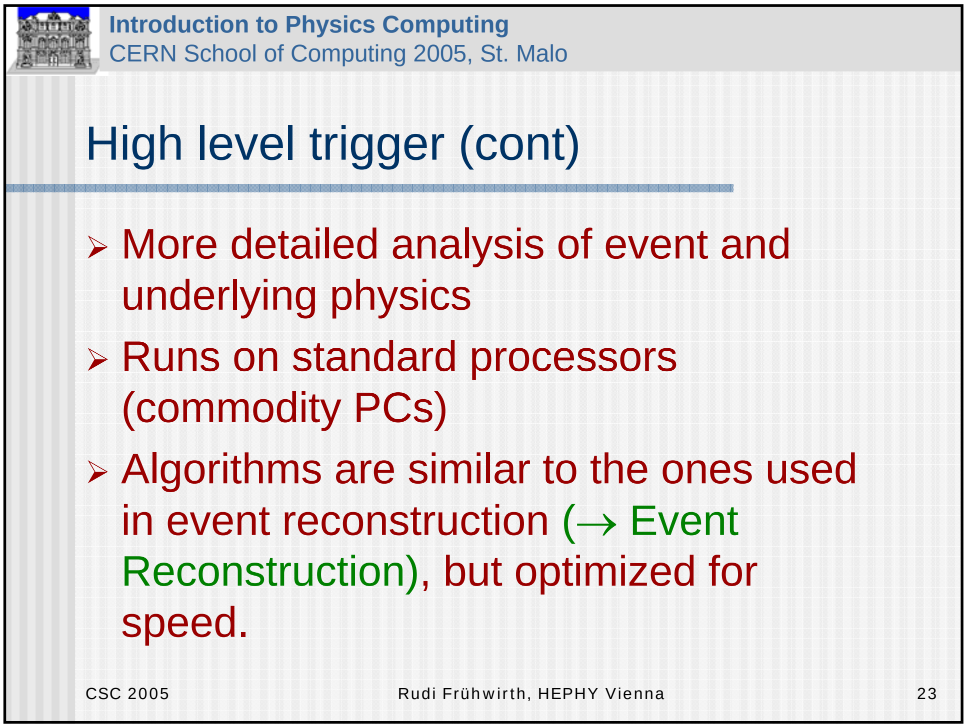


CMS L1 trigger software

- Algorithms are developed in C++
- They are tested by extensive simulation studies (→ Event Simulation)
- Manual translation into VHDL (***V***ery high speed integrated circuit ***H***ardware ***D***escription ***L***anguage)
- Comparison with C++ implementation

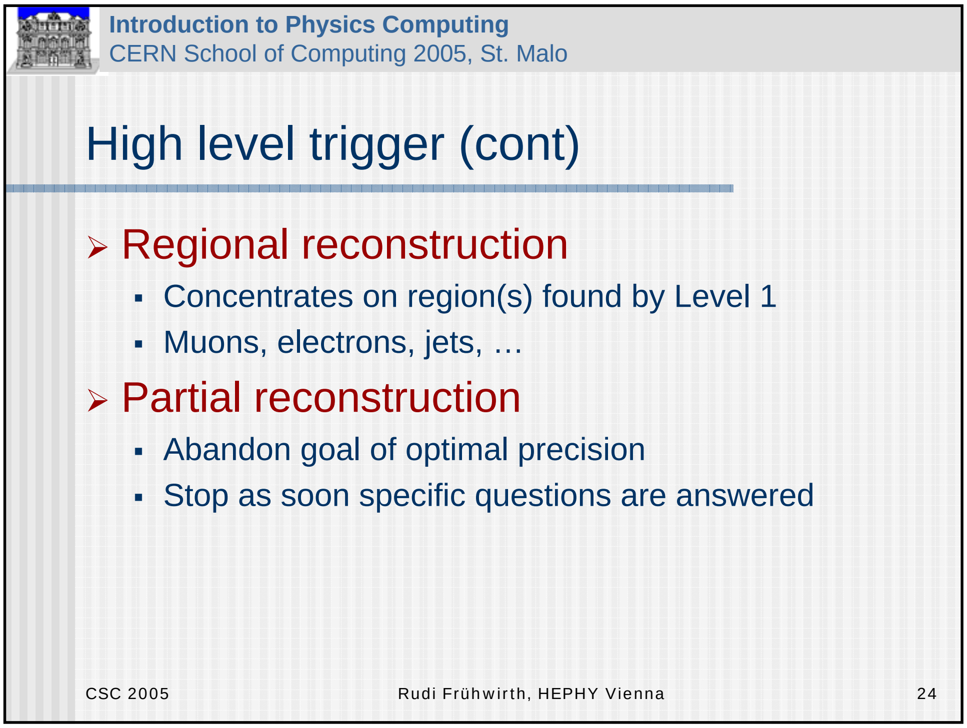
High level trigger

- Further data filtering:
  - 30 – 100 kilohertz input rate
  - 100-150 Hertz output rate
- Event tagging:
  - Reconstruct physics objects
  - Mark events having interesting features



High level trigger (cont)

- More detailed analysis of event and underlying physics
- Runs on standard processors (commodity PCs)
- Algorithms are similar to the ones used in event reconstruction (→ Event Reconstruction), but optimized for speed.



High level trigger (cont)

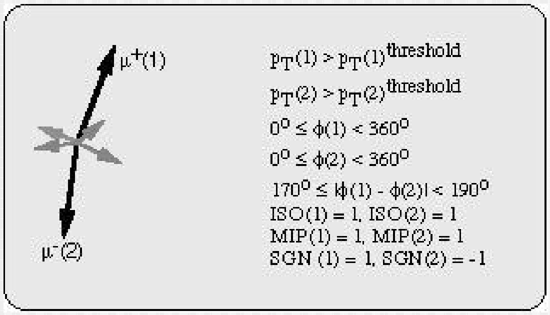
- Regional reconstruction
  - Concentrates on region(s) found by Level 1
  - Muons, electrons, jets, …
- Partial reconstruction
  - Abandon goal of optimal precision
  - Stop as soon specific questions are answered

CMS High level trigger

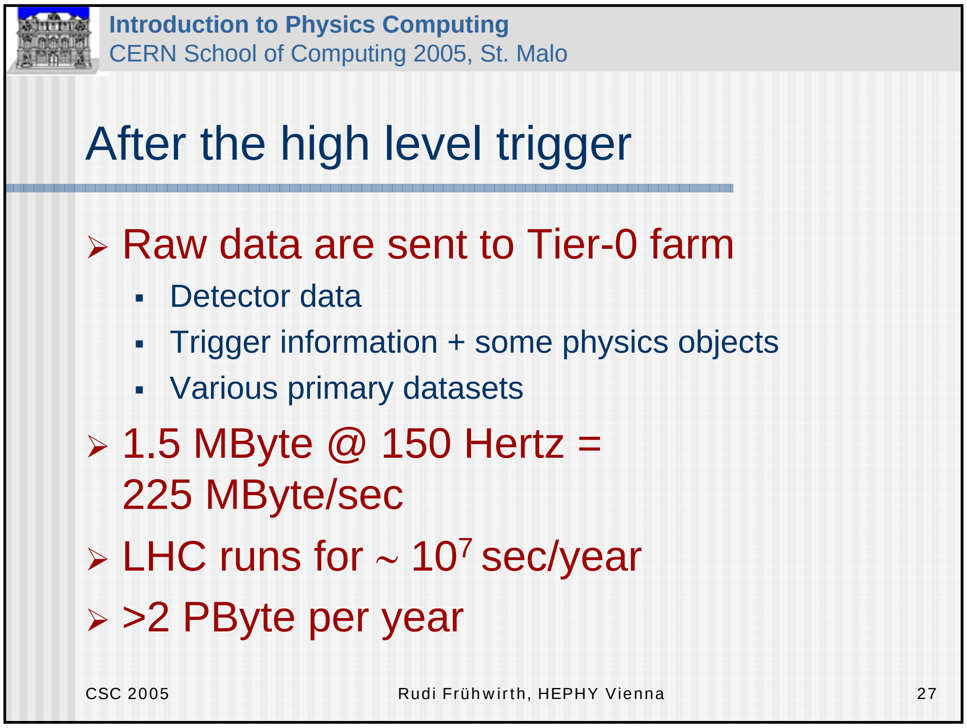
- Has to keep pace with the L1 Output
- Filter farm with about 1M SpecInt95
  - About 25000 Pentium III @ 1 GHz
  - About 2000 CPUs at startup (2007)
- Organized in subfarms
- Same software framework as in “offline” reconstruction
- Transparent exchange of algorithms

CMS HLT example

- back-to-back opposite sign isolated muons

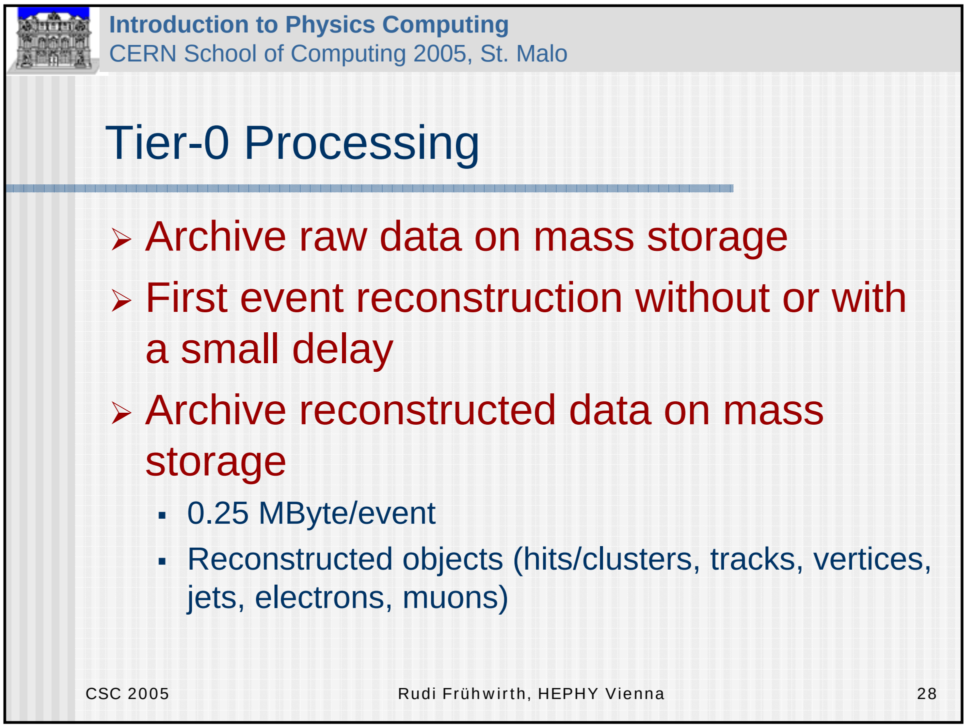


$p_T(1) > p_T(1)^{\text{threshold}}$
$p_T(2) > p_T(2)^{\text{threshold}}$
$0^0 \leq \phi(1) < 360^0$
$0^0 \leq \phi(2) < 360^0$
$170^0 \leq |\phi(1) - \phi(2)| < 190^0$
ISO(1) = 1, ISO(2) = 1
MIP(1) = 1, MIP(2) = 1
SGN (1) = 1, SGN(2) = -1



After the high level trigger

- Raw data are sent to Tier-0 farm
  - Detector data
  - Trigger information + some physics objects
  - Various primary datasets
- 1.5 MByte @ 150 Hertz = 225 MByte/sec
- LHC runs for ~ 10^7 sec/year
- >2 PByte per year



Tier-0 Processing

- Archive raw data on mass storage
- First event reconstruction without or with a small delay
- Archive reconstructed data on mass storage
  - 0.25 MByte/event
  - Reconstructed objects (hits/clusters, tracks, vertices, jets, electrons, muons)

From bits to GeV and cm

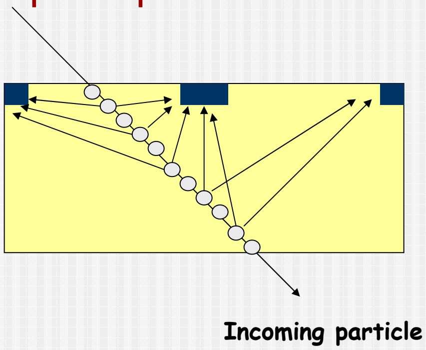
- Raw data are mostly ADC or TDC counts
- They have to be converted to physical quantities like energy or position
- Very detector dependent
- Every detector needs calibration
- Calibration constants need to be stored and updated

Calorimeter calibration

- Kinetic energy of incoming particle is converted into light or electric charge
- Destructive measurement
- Relation between deposited charge and energy needs to be known
- Long term drifts need to be monitored
- Huge amounts of data are accumulated

Silicon Tracker calibration

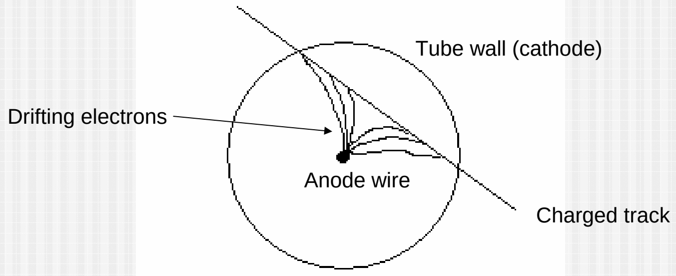
- Incoming particle creates electric charge in strips or pixels



Silicon Tracker calibration (cont)

- Charge distribution depends on location of crossing point and crossing angle
- Solve inverse problem: reconstruct crossing point from charge distribution and crossing angle
- Test beam, real data

Drift tube calibration



Drift tube calibration (cont)

- Incoming particle ionizes gas in tube
- Electrons/ions drift to anode/cathode
- Drift time is measured
- Must be converted to drift distance
- Time/distance relation must be determined (not always linear)
- Test beam, real data

Where are the detectors?

- Tracking detectors are very precise instruments
- Silicon strip detector: ~ 50 μm
- Pixel detector: ~ 10 μm
- Drift tube: ~ 100 μm
- Position needs to be known to a similar or better precision

---

Example: CMS tracker

Alignment

- Mechanical alignment
- Measurements taken before assembly
- Switching on the magnetic field
- Laser alignment
- Alignment with charged tracks from collisions, beam halo and cosmic rays
- Continuous process

Environmental data

- Calibration data
- Alignment data
- Temperatures, gas pressures, …
- Machine parameters
- Need to be made persistent

Detector related software

- Configuration
  - Load trigger files, set thresholds, set HV, set amplifier gains, …
- Slow control
  - Measure and adjust temperature, gas pressure, dark currents, …
- Monitoring
  - Check trigger rates, detector efficiency, cluster sizes, wire maps, …

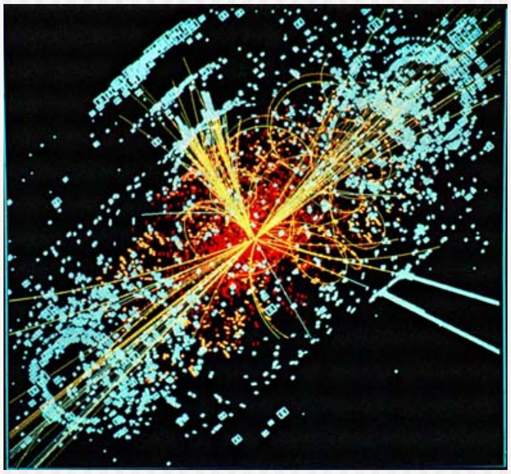
Event reconstruction

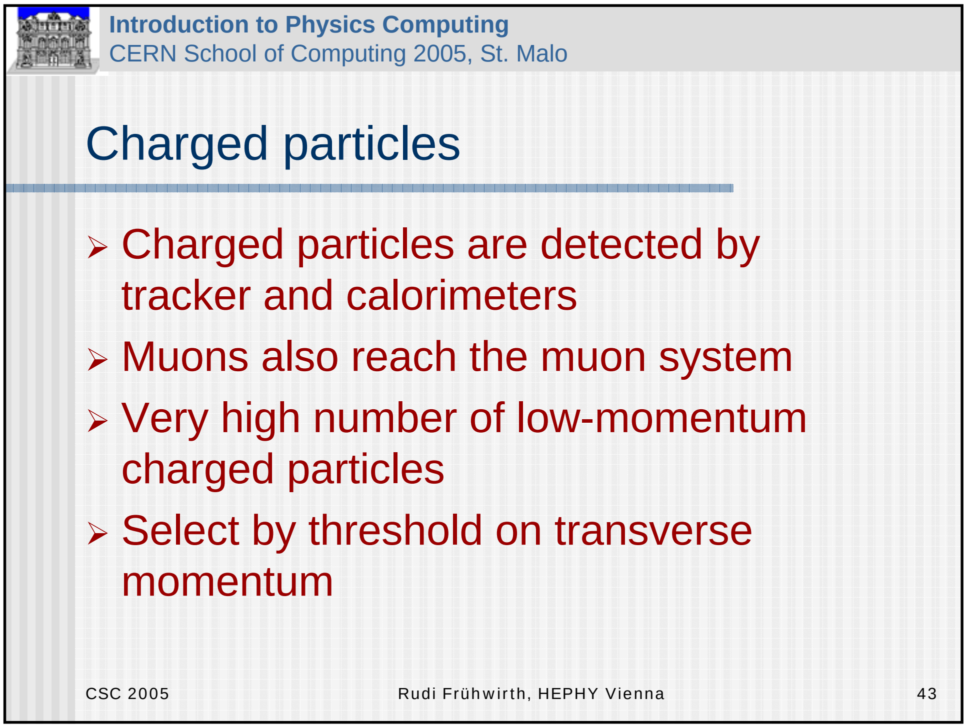
- Find out which particles have been created where and with which momentum
- Some of them are short-lived and have to be reconstructed from their decay products
- Some of them (neutrinos) escape without leaving any trace

Event reconstruction (cont)

- Reconstruct charged particles
- Reconstruct neutral particles
- Identify type of particles
- Reconstruct interaction points (vertices)
- Reconstruct kinematics of the interaction
- Not trivial, very time-consuming …

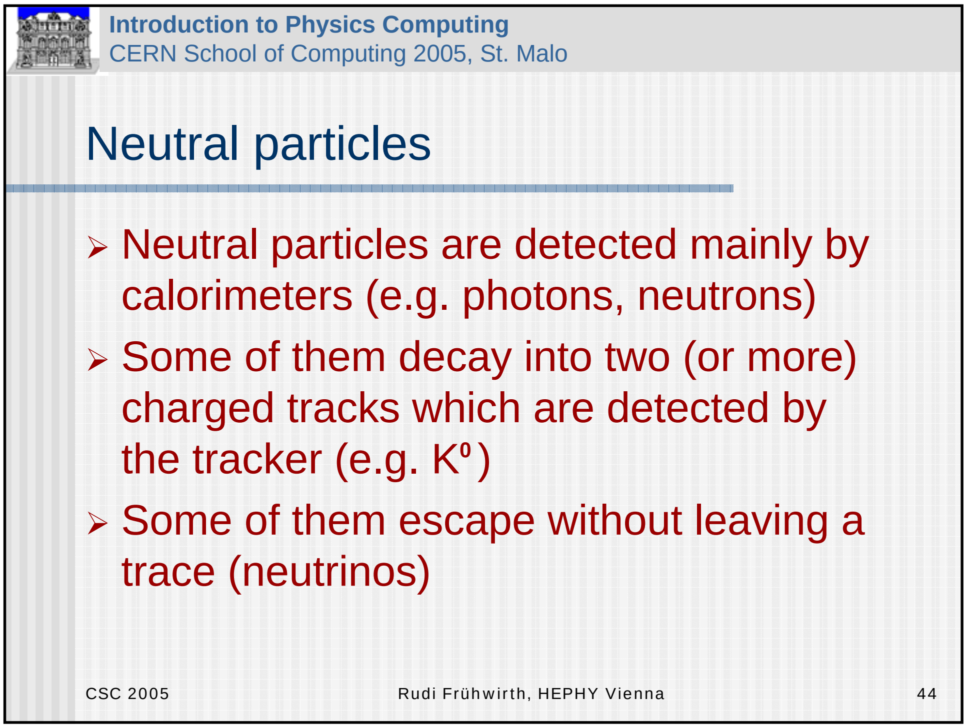
CMS: Higgs decay into two jets





Charged particles

- Charged particles are detected by tracker and calorimeters
- Muons also reach the muon system
- Very high number of low-momentum charged particles
- Select by threshold on transverse momentum



Neutral particles

- Neutral particles are detected mainly by calorimeters (e.g. photons, neutrons)
- Some of them decay into two (or more) charged tracks which are detected by the tracker (e.g. K^0)
- Some of them escape without leaving a trace (neutrinos)

Reconstruction of charged particles

- Trajectory is curved because of the magnetic field
- Position is measured in a number of places –“hits”
- Determine track parameters (location, direction, momentum) from the position measurements
- Data compression

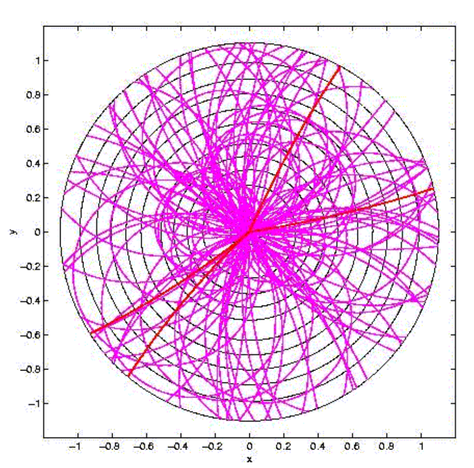
The difficulties

- Charged particles interact with all the material, not only the sensitive parts
- Multiple Coulomb scattering
  - Changes direction, but not momentum
- Energy loss by ionization
  - Changes momentum, but not energy
- Energy loss by bremsstrahlung
  - Only for electrons

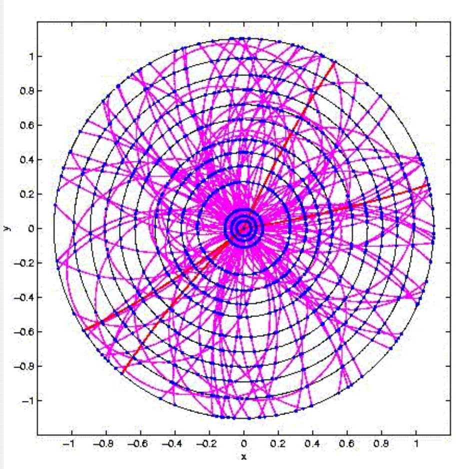
More difficulties

- Assignment of hits to particles is unknown
- Huge background from low-momentum tracks
- Additional background from other interactions in the same beam crossing and from adjacent beam crossings

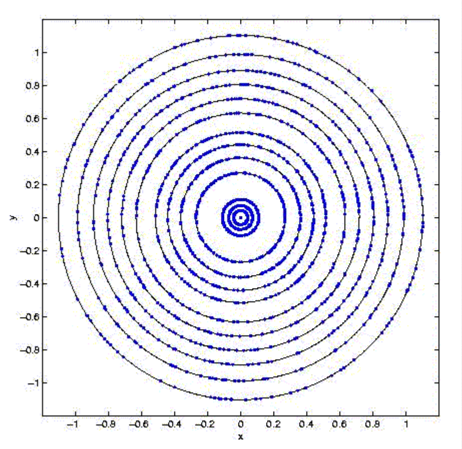
Tracks only

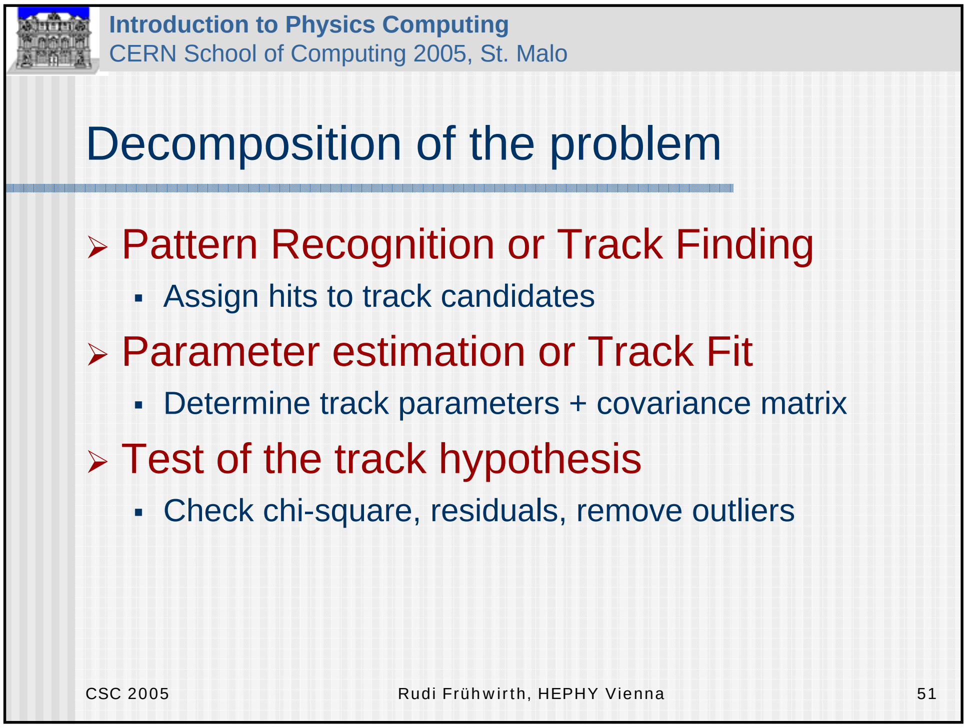


Tracks with hits



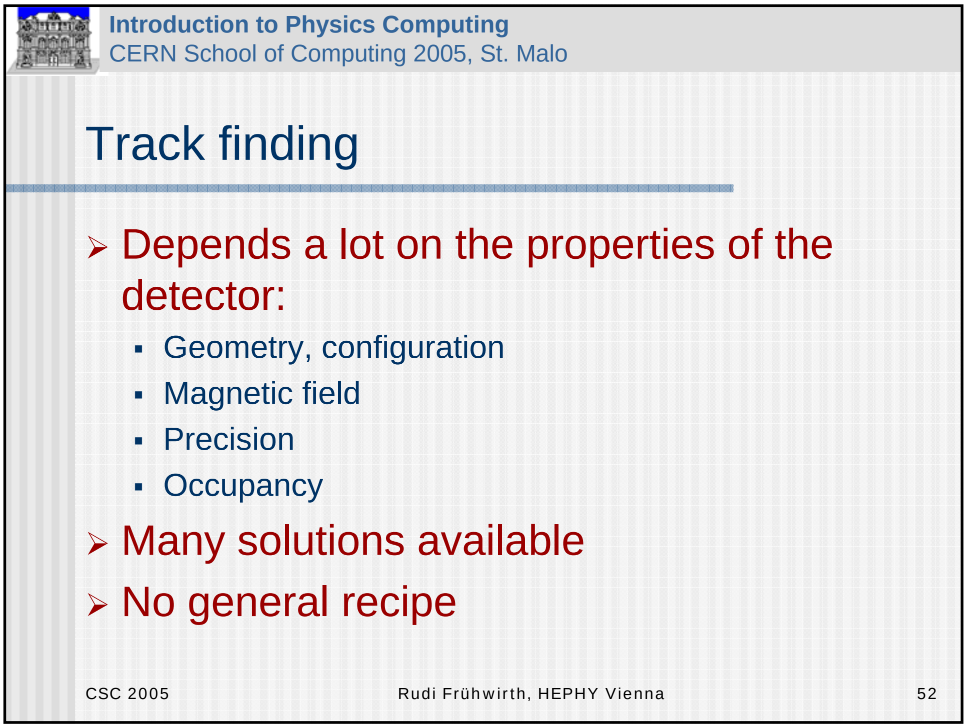
Hits only





Decomposition of the problem

- Pattern Recognition or Track Finding
  - Assign hits to track candidates
- Parameter estimation or Track Fit
  - Determine track parameters + covariance matrix
- Test of the track hypothesis
  - Check chi-square, residuals, remove outliers



Track finding

- Depends a lot on the properties of the detector:
  - Geometry, configuration
  - Magnetic field
  - Precision
  - Occupancy
- Many solutions available
- No general recipe

A few track finding algorithms

- Track following
- Simple Kalman filter
- Combinatorial Kalman filter
- Track road
- Hough transform
- Hopfield network

Track fit

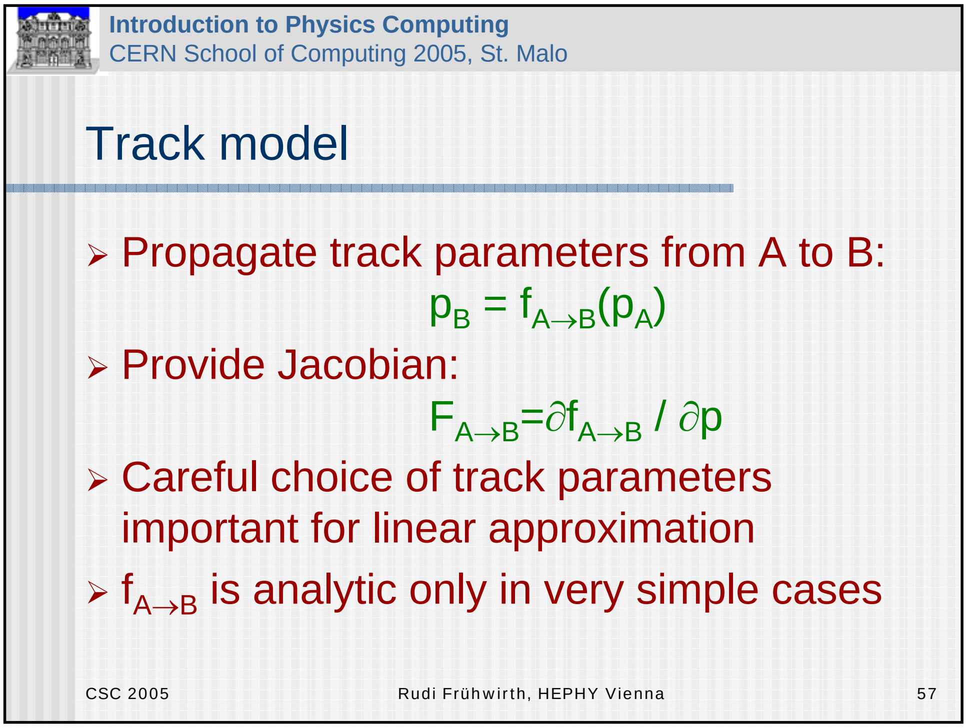
- Determine track parameters
- Determine covariance matrix
- Test track hypothesis
- Reject outliers
    - Distorted hits (cluster fusion, δ- electron)
    - Extraneous hits
    - Electronic noise

Ingredients

- Magnetic field
  - Constant or variable
- Track model
  - Solution of the equation of motion
  - Analytic (explicit) or numerical
- Error model
  - Observations errors
  - Process noise

Magnetic field

- Fast computation required
- Interpolation in a large table
- Global approximation by harmonic functions
- Local approximation by low-order polynomials
- Constant field plus correction terms



Track model

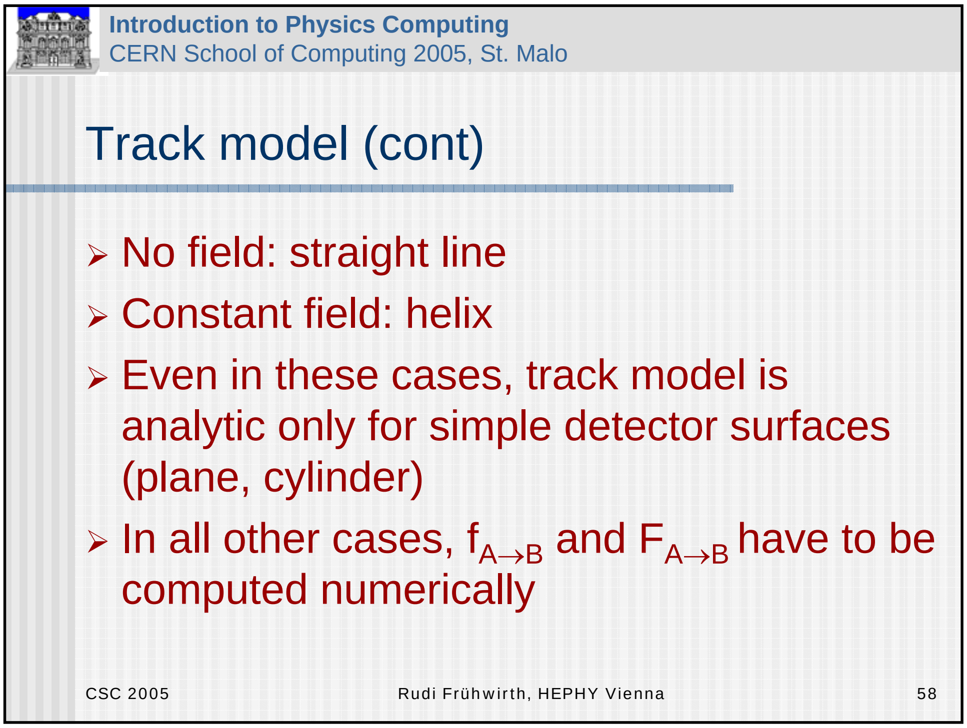
- Propagate track parameters from A to B:

$$p_B = f_{A\to B}(p_A)$$

- Provide Jacobian:

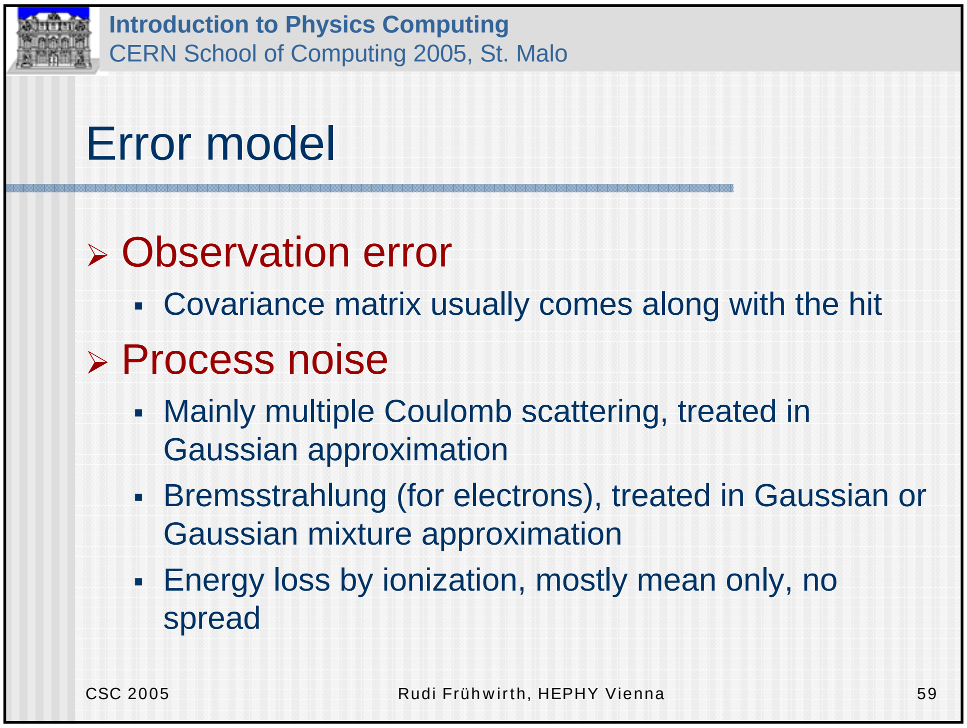
$$F_{A\to B}=\partial f_{A\to B} / \partial p$$

- Careful choice of track parameters important for linear approximation
- $f_{A\to B}$ is analytic only in very simple cases



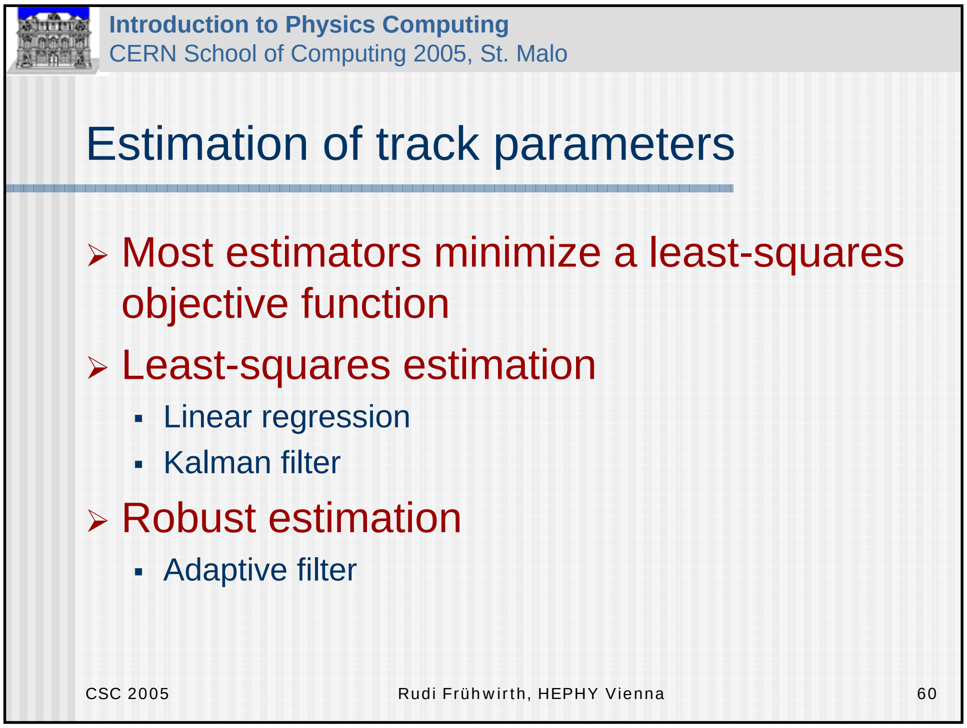
Track model (cont)

- No field: straight line
- Constant field: helix
- Even in these cases, track model is analytic only for simple detector surfaces (plane, cylinder)
- In all other cases, $f_{A\to B}$ and $F_{A\to B}$ have to be computed numerically



Error model

- Observation error
  - Covariance matrix usually comes along with the hit
- Process noise
  - Mainly multiple Coulomb scattering, treated in Gaussian approximation
  - Bremsstrahlung (for electrons), treated in Gaussian or Gaussian mixture approximation
  - Energy loss by ionization, mostly mean only, no spread



Estimation of track parameters

- Most estimators minimize a least-squares objective function
- Least-squares estimation
  - Linear regression
  - Kalman filter
- Robust estimation
  - Adaptive filter

Reconstruction of neutral particles

- Neutral particles are only seen by the calorimeters
- Photons are absorbed in the electromagnetic calorimeter
- Neutral (and charged) hadrons are absorbed in the hadron calorimeter
- Neutrinos are not detected directly

Shower finding

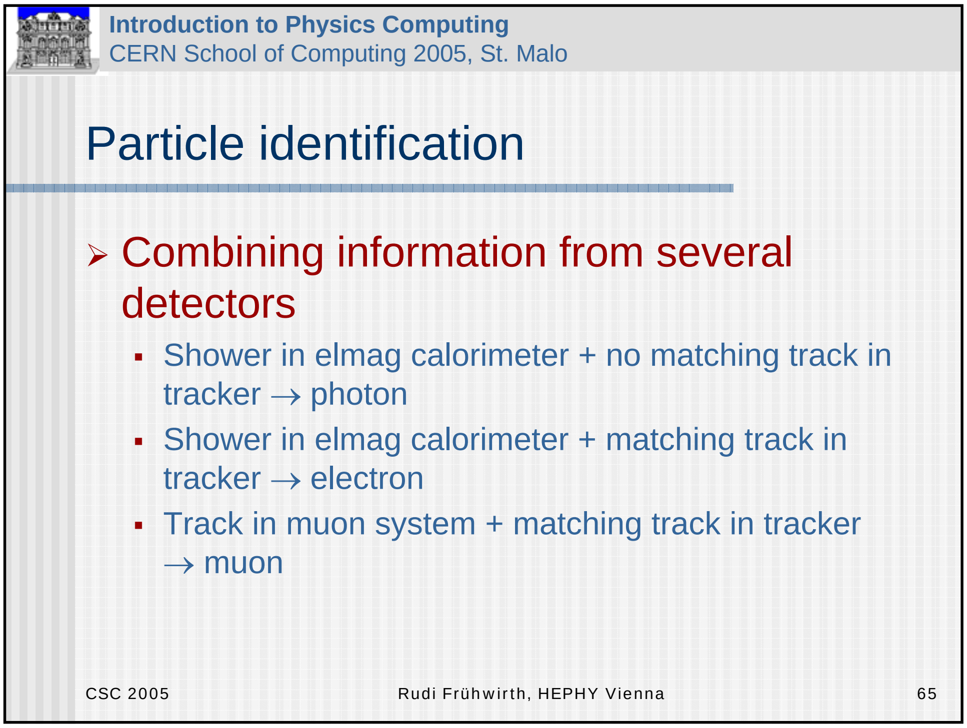
- An incident particle produces a shower in the calorimeter
- A shower is a cluster of cells with energy deposit above threshold
- Various clustering techniques are used to find showers
- Overlapping clusters must be separated

Shower finding (cont)

- The algorithms depend on various characteristics of the calorimeter
  - Type (electromagnetic or hadronic)
  - Technology (homogeneous or sampling)
  - Cell geometry
  - Granularity
  - Noise and background level

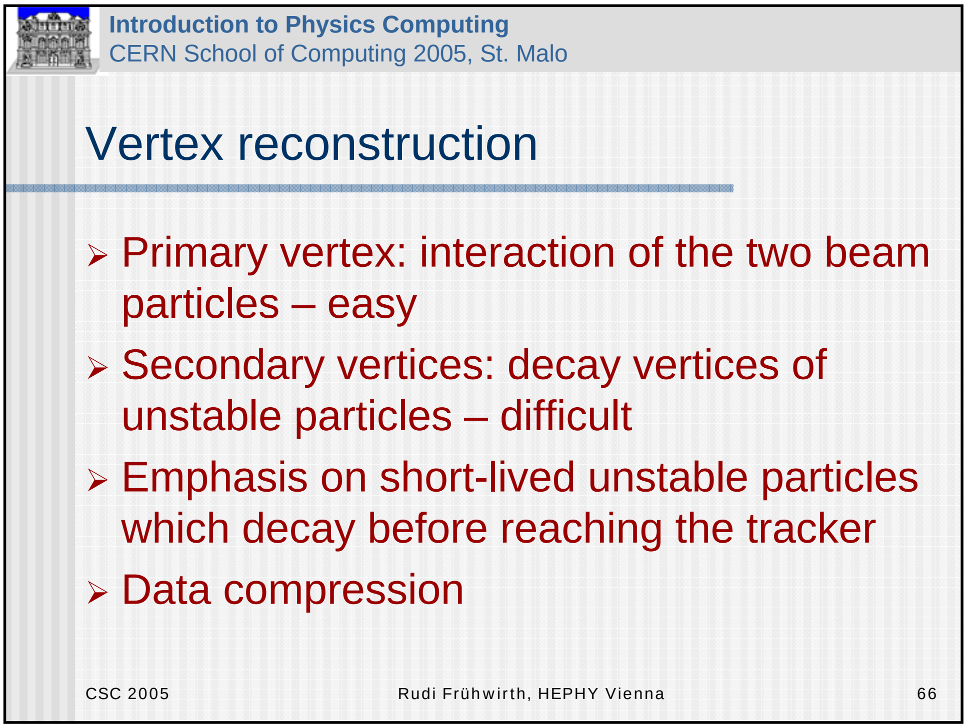
Particle identification

- Determining the type of a particle
- Dedicated detectors
  - Threshold Cherenkov
  - Ring imaging Cherenkov (RICH)
  - Transition radiation detector
  - Ionization measurements in gas or silicon



Particle identification

- Combining information from several detectors
  - Shower in elmag calorimeter + no matching track in tracker → photon
  - Shower in elmag calorimeter + matching track in tracker → electron
  - Track in muon system + matching track in tracker → muon



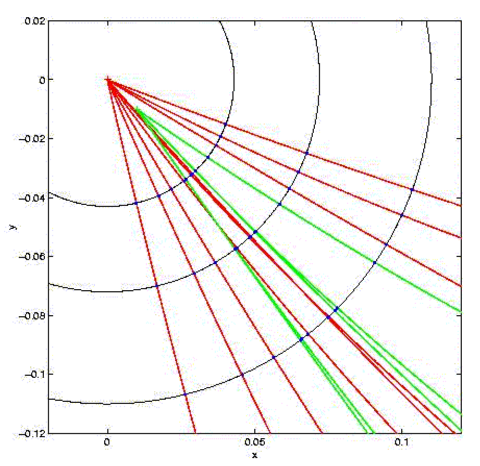
Vertex reconstruction

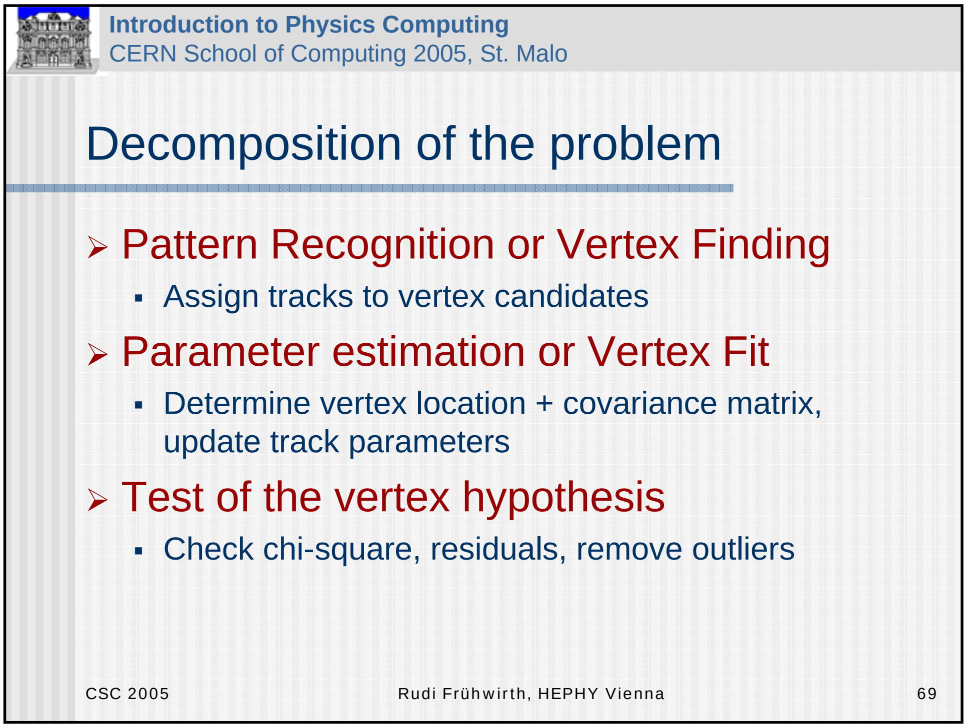
- Primary vertex: interaction of the two beam particles – easy
- Secondary vertices: decay vertices of unstable particles – difficult
- Emphasis on short-lived unstable particles which decay before reaching the tracker
- Data compression

The difficulties

- Association of tracks to vertices is unknown
- Secondary tracks may pass very close to the primary vertex
    - Especially if decay length is small
- Track reconstruction may be less than perfect

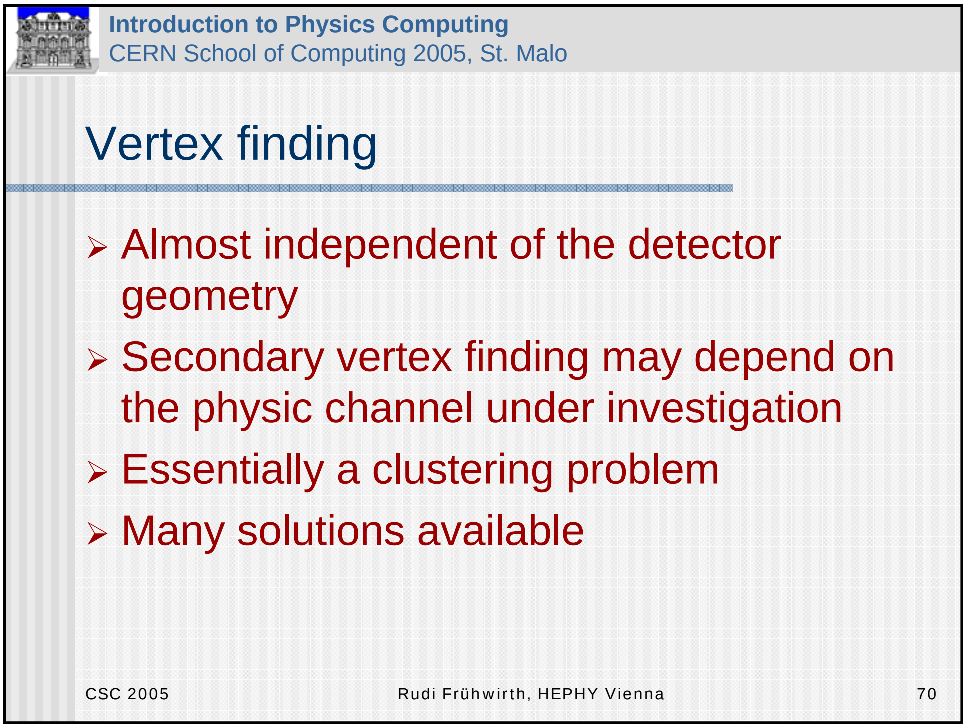
Primary and secondary tracks





Decomposition of the problem

- Pattern Recognition or Vertex Finding
  - Assign tracks to vertex candidates
- Parameter estimation or Vertex Fit
  - Determine vertex location + covariance matrix, update track parameters
- Test of the vertex hypothesis
  - Check chi-square, residuals, remove outliers



Vertex finding

- Almost independent of the detector geometry
- Secondary vertex finding may depend on the physic channel under investigation
- Essentially a clustering problem
- Many solutions available

A few vertex finding algorithms

- Hierarchical clustering
  - Single linkage, complete linkage,…
- Non-hierarchical clustering
  - k-means, robust location (mode) estimation, iterated vertex fit
- Neural network/physics inspired
  - Competitive learning, deterministic annealing, superparamagnetic clustering, quantum clustering,…

Estimation of vertex parameters

- Most estimators minimize a least-squares objective function
- Least-squares estimation
  - Linear regression
  - Kalman filter
- Robust estimation
  - Adaptive filter

Kinematic fit

- Impose constraints on a reconstructed vertex
  - Momentum conservation
  - Energy conservation (if masses are known)
  - Invariant mass of mother particle
- Put constraints into a Lagrange multiplier
- Construct least-squares objective function

Kinematic fit (cont)

- Taylor-expand objective function with respect to all momentum vectors
- Minimize objective function
- Neutral particles have to be included (calorimeter information)

Persistency

- Event reconstruction produces physics objects
  - Tracks
  - Vertices
  - Identified particles
  - Jets
  - Tags
- Need to be made persistent

Persistency (cont)

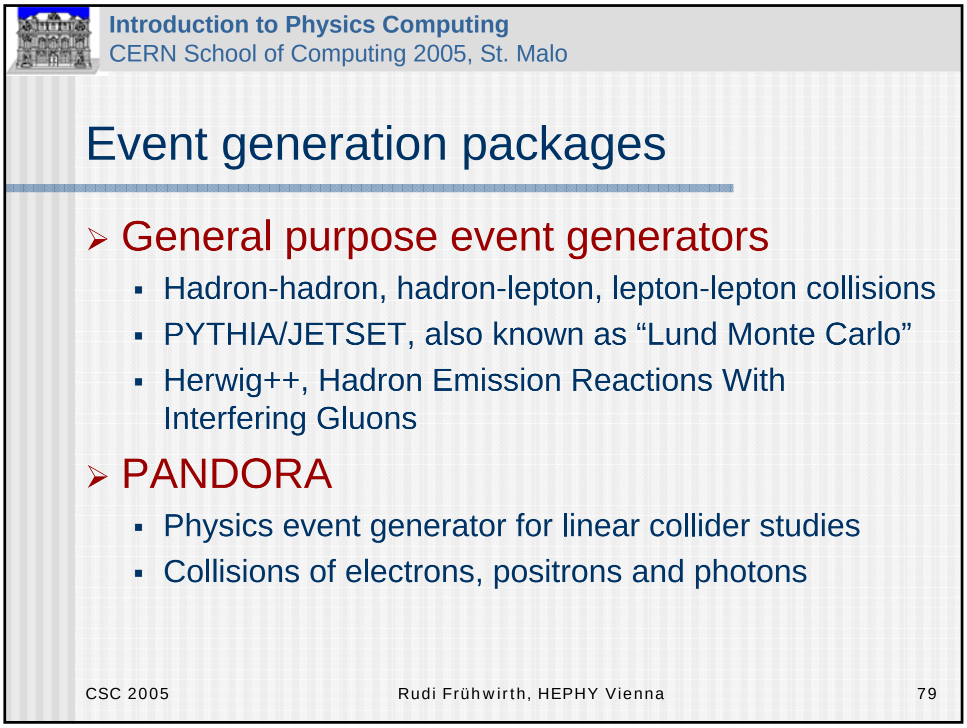
- Physics objects depend on
  - Alignment
  - Calibration
  - Version of the reconstruction program
  - Algorithm parameters
- Must be made persistent as well
- Tools: ROOT, POOL

Simulation

- Why do we need simulation?
- Optimization of detector in design phase
- Testing, validation and optimization of reconstruction algorithms
- Computation of reconstruction efficiency
- Computation of acceptance corrections
- Background studies

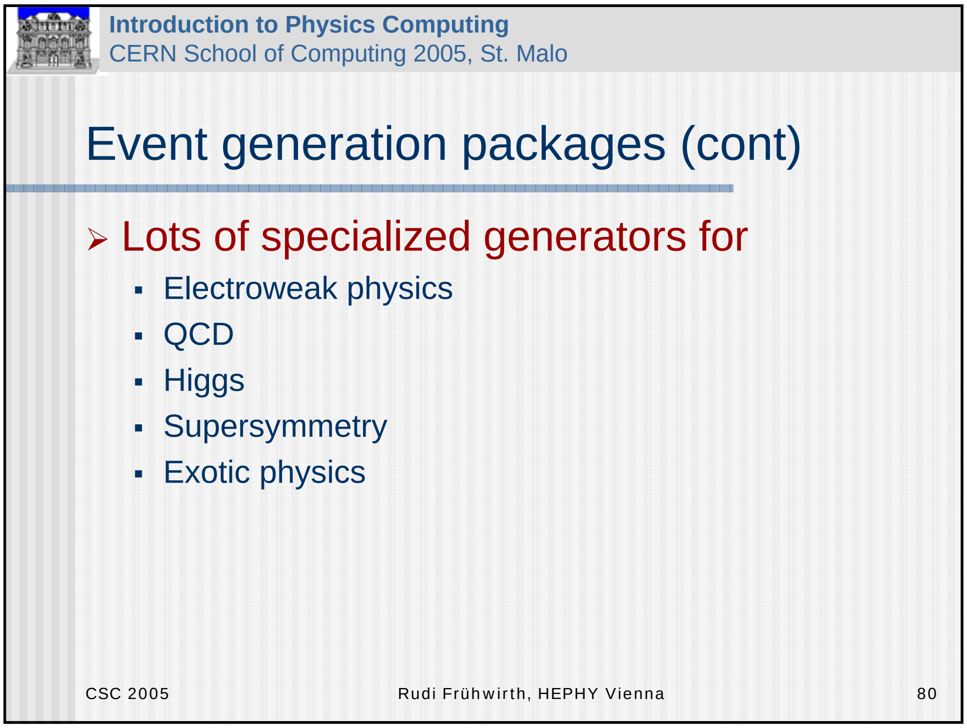
Simulation steps

- Event generation
- Tracking through the detector, using detector geometry and magnetic field
- Interaction of particles with matter
- Signal generation in sensitive volumes
- Digitization (simulate ADC or TDC)
- Digitized data and truth information are made persistent



Event generation packages

- General purpose event generators
  - Hadron-hadron, hadron-lepton, lepton-lepton collisions
  - PYTHIA/JETSET, also known as “Lund Monte Carlo”
  - Herwig++, Hadron Emission Reactions With Interfering Gluons
- PANDORA
  - Physics event generator for linear collider studies
  - Collisions of electrons, positrons and photons



Event generation packages (cont)

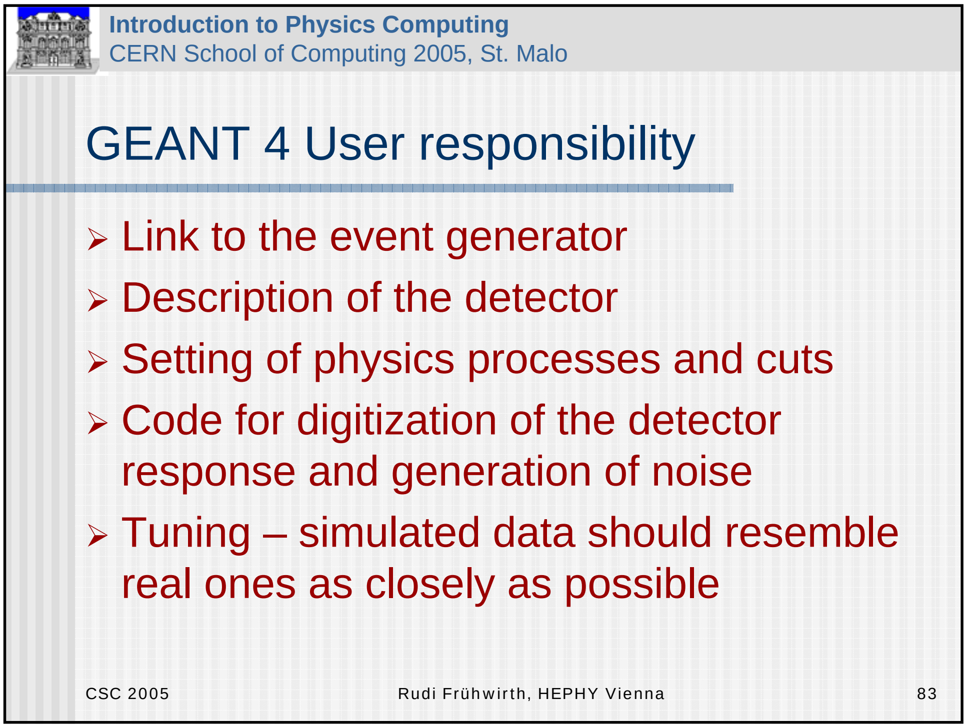
- Lots of specialized generators for
  - Electroweak physics
  - QCD
  - Higgs
  - Supersymmetry
  - Exotic physics

Detector simulation

- Was frequently (and still sometimes is) experiment-specific
- Nowadays there is a widely used standard: GEANT
  - GEANT3: FORTRAN
  - GEANT4: C++
- You will hear a LOT about GEANT4 later!

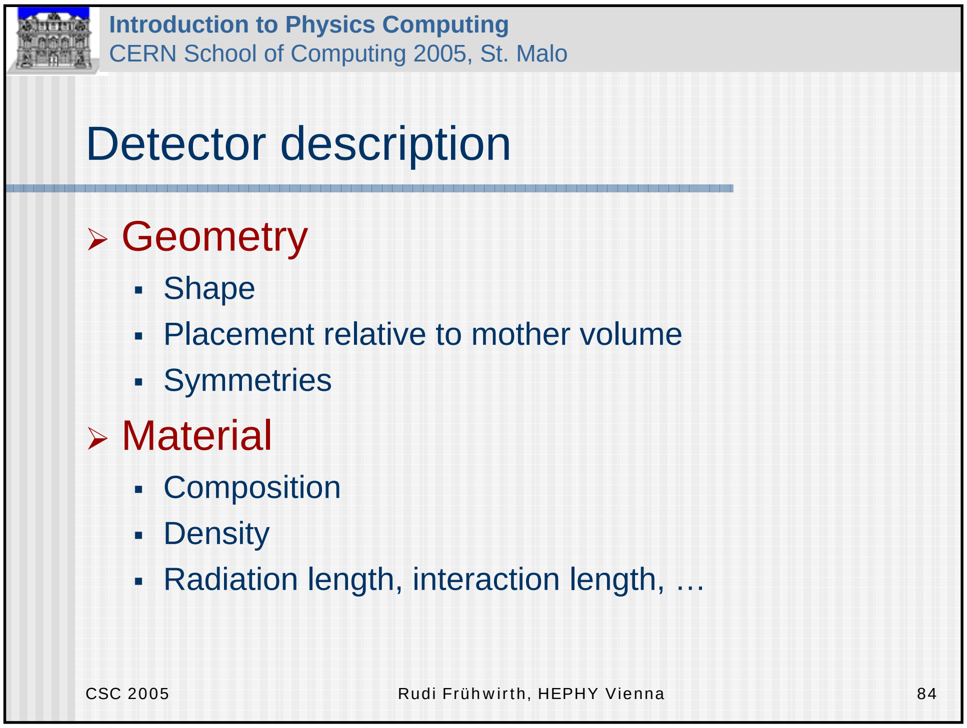
GEANT4 functionality

- Description of geometry and materials
- Particle tracking and interactions with matter
- Generation of the detector response
- Bookkeeping, metadata management
- Visualization of geometry, tracks and hits



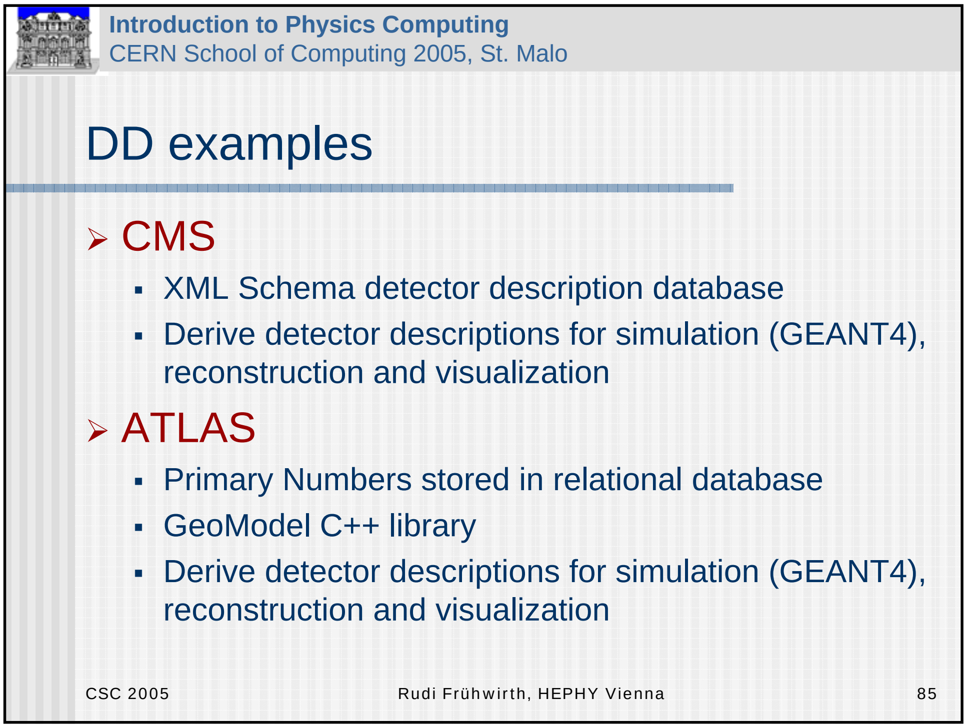
GEANT 4 User responsibility

- Link to the event generator
- Description of the detector
- Setting of physics processes and cuts
- Code for digitization of the detector response and generation of noise
- Tuning – simulated data should resemble real ones as closely as possible



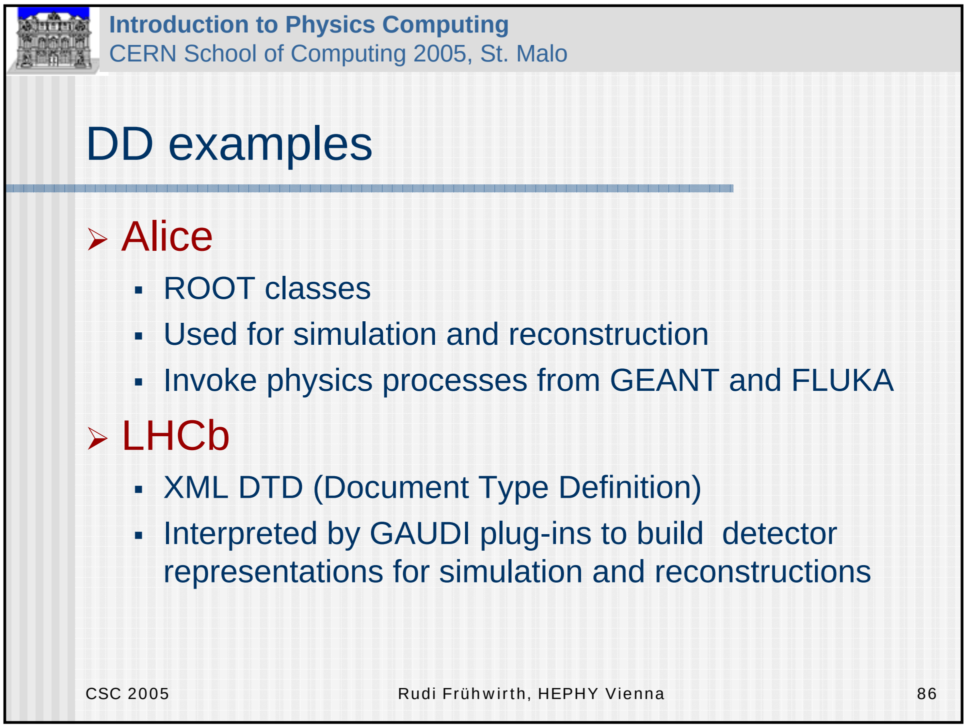
Detector description

- Geometry
  - Shape
  - Placement relative to mother volume
  - Symmetries
- Material
  - Composition
  - Density
  - Radiation length, interaction length, …



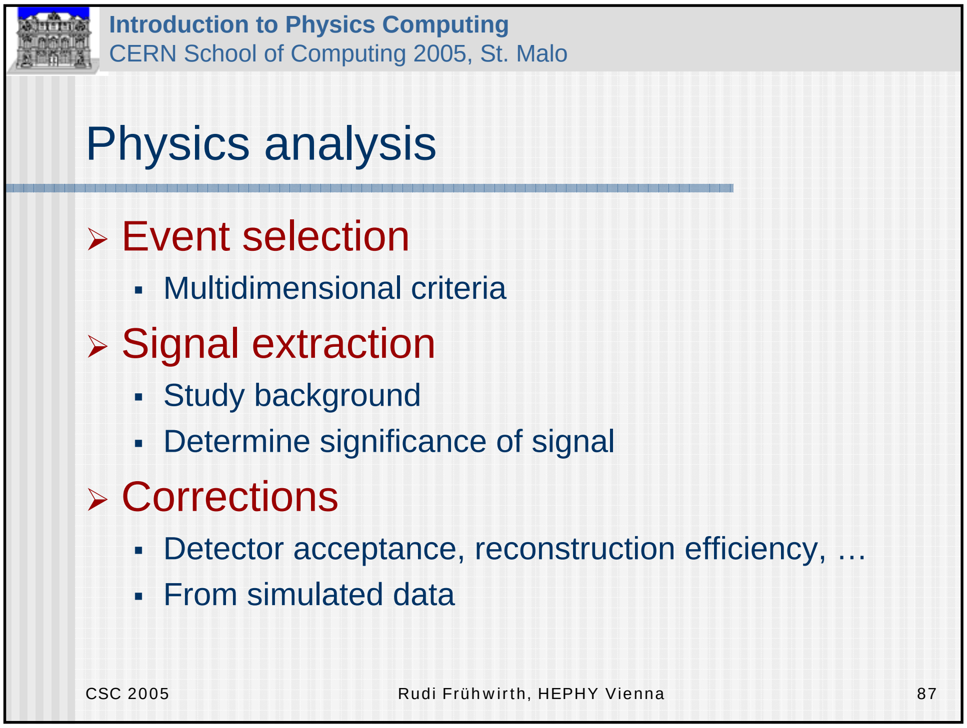
DD examples

- CMS
  - XML Schema detector description database
  - Derive detector descriptions for simulation (GEANT4), reconstruction and visualization
- ATLAS
  - Primary Numbers stored in relational database
  - GeoModel C++ library
  - Derive detector descriptions for simulation (GEANT4), reconstruction and visualization



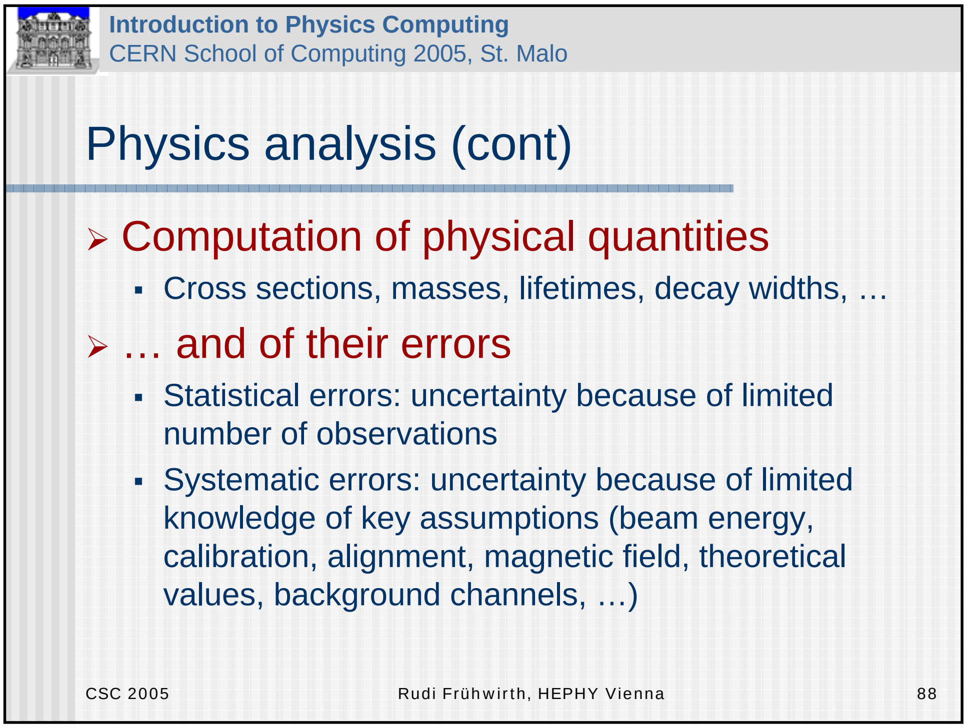
DD examples

- Alice
  - ROOT classes
  - Used for simulation and reconstruction
  - Invoke physics processes from GEANT and FLUKA
- LHCb
  - XML DTD (Document Type Definition)
  - Interpreted by GAUDI plug-ins to build detector representations for simulation and reconstructions



Physics analysis

- Event selection
  - Multidimensional criteria
- Signal extraction
  - Study background
  - Determine significance of signal
- Corrections
  - Detector acceptance, reconstruction efficiency, …
  - From simulated data



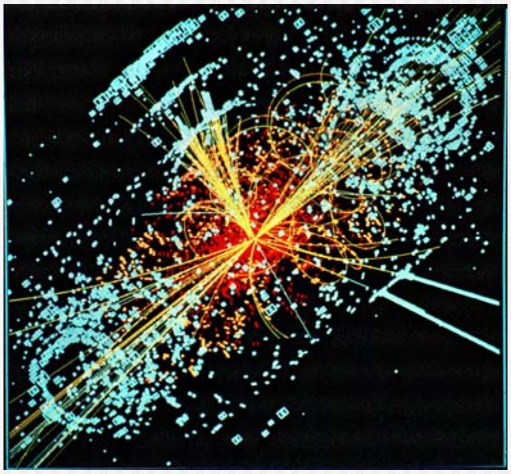
Physics analysis (cont)

- Computation of physical quantities
  - Cross sections, masses, lifetimes, decay widths, …
- … and of their errors
  - Statistical errors: uncertainty because of limited number of observations
  - Systematic errors: uncertainty because of limited knowledge of key assumptions (beam energy, calibration, alignment, magnetic field, theoretical values, background channels, …)

Analysis tools

- Need versatile tools for
  - Multidimensional selection
  - Event display and interactive reprocessing
  - Histogramming
  - Plotting
  - Fitting of curves and models
  - Point estimation, confidence intervals, limits
  - …

Event display



Analysis tools (cont)

- ROOT
  - Builds on HBOOK and PAW (CERN)
- OpenScientist
  - “Open, modular, free, portable, efficient and collaborative” (LAL Orsay)
- JAS
  - Java Analysis Studio (SLAC)
- WIRED
  - Platform independent event display (Java, SLAC)

Distributed analysis

- Physics analysis will take place in many labs all over the world
- Physicists need access to event data and corresponding calibration, alignment and bookkeeping data … and to simulated data
- We need the grid!
- You will hear a LOT about the grid next week!

Metadata challenge

- Metadata are data describing other data
- Distributed analysis needs lots of metadata to track the location and validity of alignment constants, calibration constants, reconstructed objects, …
- Frequent updates
- Frequent access

Datasets in CMS

- RAW: Raw data (1-1.5 MByte)
  - Detector data, L1 trigger results, HLT results, reconstructed HLT objects
- RECO: Reconstructed data (1.5 MByte)
  - Reconstructed objects (hits, clusters, tracks, vertices, muons, electrons, jets)
- AOD: Analysis object data (50 kByte)
  - High-level reconstructed objects (tracks, vertices, muons, electrons, jets)

Datasets in CMS (cont)

- TAG: Tagging data (10 kByte)
  - Run/event number, some high-level physics objects
- Non-event data
  - Construction data (information on sub-detectors)
  - Equipment management data (detector geometry, electronics)
  - Configuration data (front-end electronics)
  - Conditions data (run conditions, calibration, alignment)

Data flow in CMS

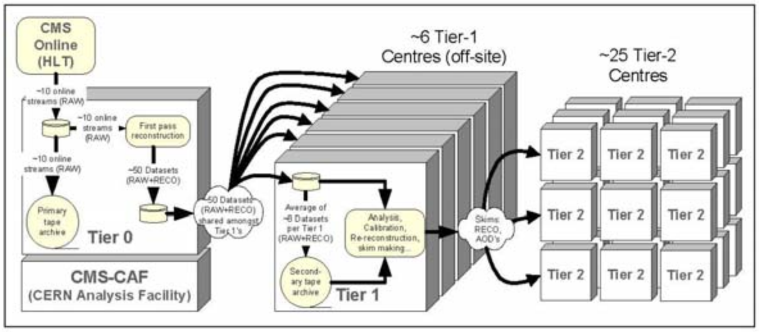


Figure 2.1: Schematic flow of bulk (real) event data in the CMS Computing Model. Not all connections are shown - for example flow of MC data from Tier-2's to Tier-1's or peer-to-peer connections between Tier-1's.

Tiered resources

- Tier-0 (CERN)
  - First pass of reconstruction
  - Primary archive on mass storage
- Tier-1 (<10 centres)
  - Global and local services
  - (Only) copies of certain samples
  - Reconstruction
  - Monte Carlo production

Tiered resources (cont)

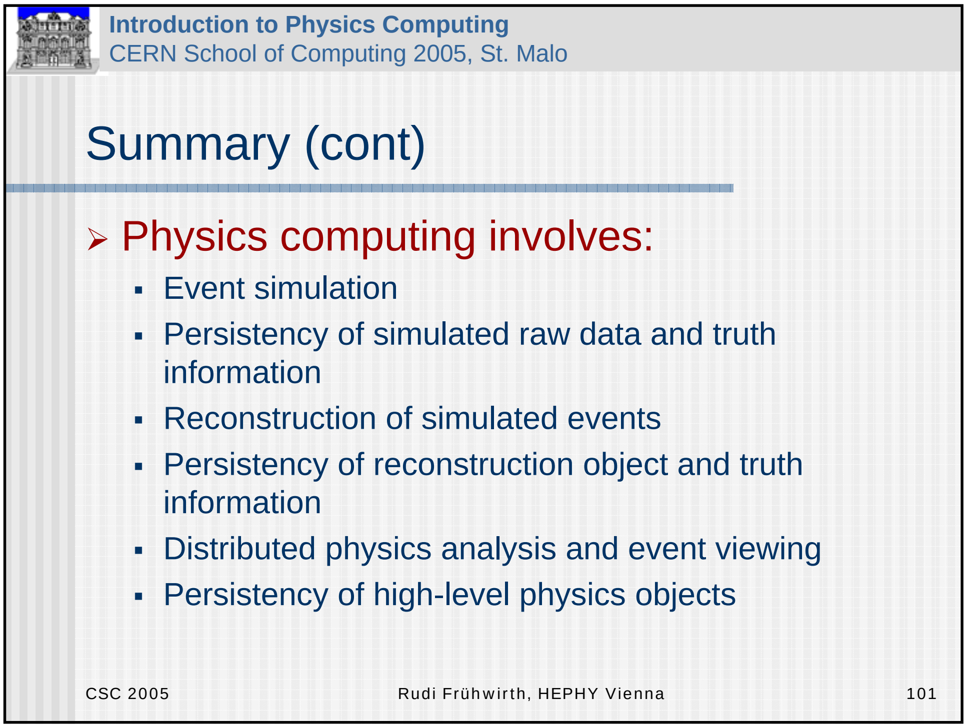
- Tier-2 (~25 centres)
  - Associated to a particular Tier-1 centre
  - Local use
  - CMS controlled use
  - Opportunistic use
- Tier-3 (~100 centres)
  - Coordination with a specific Tier-2 centre
  - Local use
  - No guaranteed support, no guaranteed availability

Additional resources

- CMS-CAF (CERN Analysis Facility)
  - Ready access to RAW and RECO data
  - Short turnaround
  - Operation critical tasks

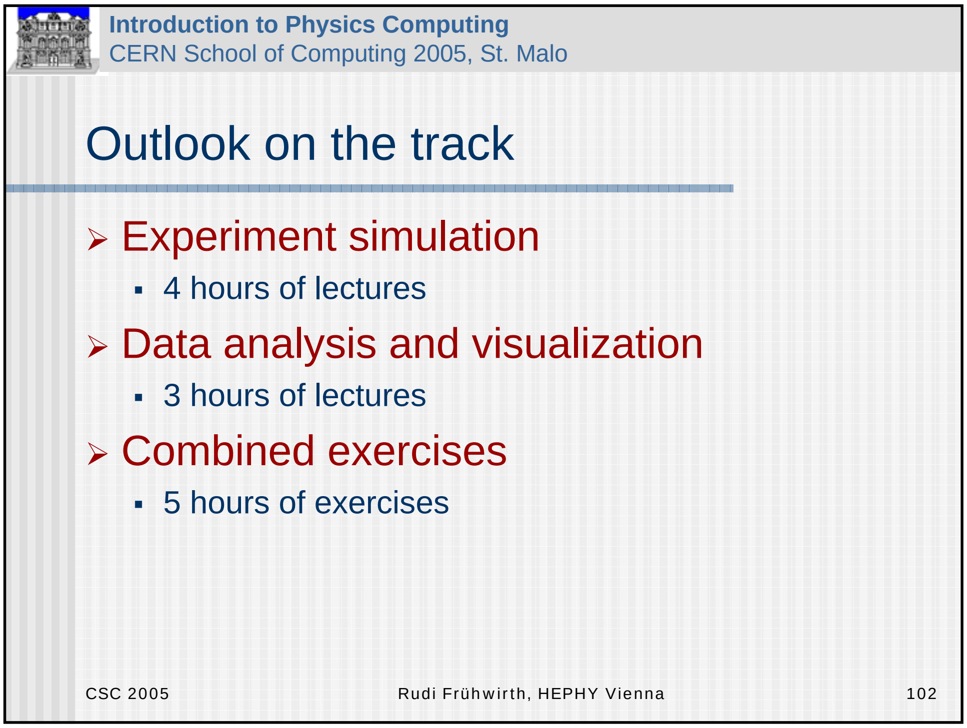
Summary

- Physics computing involves:
  - Event filtering with multilevel trigger
  - Persistency of raw data
  - Calibration and alignment
  - Persistency of calibration, alignment and environmental data
  - Event reconstruction
  - Persistency of reconstruction objects and metadata



Summary (cont)

- Physics computing involves:
  - Event simulation
  - Persistency of simulated raw data and truth information
  - Reconstruction of simulated events
  - Persistency of reconstruction object and truth information
  - Distributed physics analysis and event viewing
  - Persistency of high-level physics objects



Outlook on the track

- Experiment simulation
  - 4 hours of lectures
- Data analysis and visualization
  - 3 hours of lectures
- Combined exercises
  - 5 hours of exercises