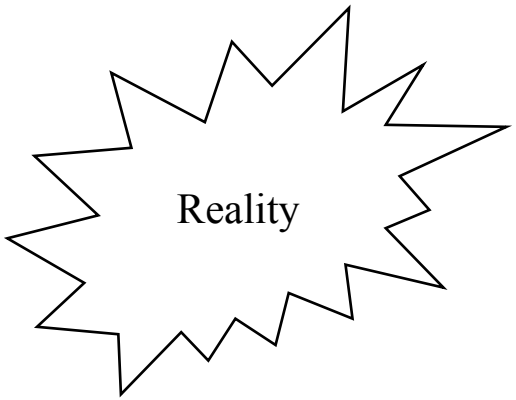
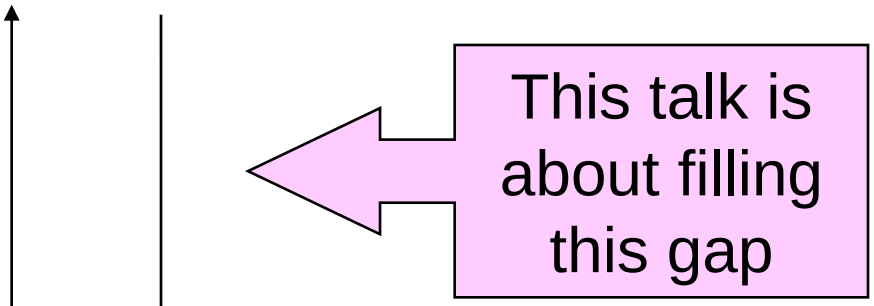
LHC data processing challenges: from the detector data to the physics results

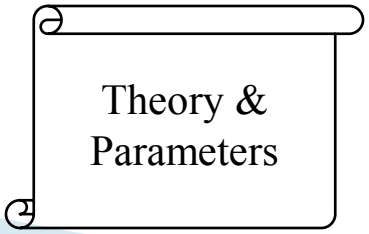
Pere Mató, PH Department, CERN

June 2012



We use experiments to inquire about what "reality" does.





The goal is to understand in the most general; that's usually also the simplest.

- A. Eddington

Theory

Particle Data Group,
Barnett et al

10. ELECTROWEAK MODEL AND CONSTRAINTS ON NEW PHYSICS

Revised August 1999 by J. Erler and P. Langacker (Univ. of Pennsylvania).

10.1. Introduction

The standard electroweak model is based on the gauge group [1] SU(2) × U(1), with gauge bosons W^i_μ, $i = 1, 2, 3$, and B_μ for the SU(2) and U(1) factors, respectively, and the corresponding gauge coupling constants g and g'. The left-handed fermion fields $\psi_i = \begin{pmatrix} \nu_i \\ \ell_i^- \end{pmatrix}$ and $\begin{pmatrix} u_i \\ d_i' \end{pmatrix}$ of the i^{th} fermion family transform as doublets under SU(2), where $d_i' \equiv \sum_j V_{ij}\, d_j$, and V is the Cabibbo-Kobayashi-Maskawa mixing matrix. (Constraints on V are discussed in the section on the Cabibbo-Kobayashi-Maskawa mixing matrix.) The right-handed fields are SU(2) singlets. In the minimal model there are three fermion families and a single complex Higgs doublet $\phi \equiv \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}$

After spontaneous symmetry breaking the Lagrangian for the fermion fields is

$$\mathscr{L}_F = \sum_i \overline{\psi}_i \left(i\, \partial\!\!\!/ - m_i - \frac{g m_i H}{2 M_W} \right) \psi_i$$
$$- \frac{g}{2\sqrt{2}} \sum_i \overline{\psi}_i\, \gamma^\mu\, (1-\gamma^5)(T^+\, W^+_\mu + T^-\, W^-_\mu)\, \psi_i$$
$$- e \sum_i q_i\, \overline{\psi}_i\, \gamma^\mu\, \psi_i\, A_\mu$$
$$- \frac{g}{2\cos\theta_W} \sum_i \overline{\psi}_i\, \gamma^\mu (g_V^i - g_A^i \gamma^5)\, \psi_i\, Z_\mu\,. \qquad (10.1)$$

"Clear statement of how the world works formulated mathematically in term of equations"

Additional term goes here

Experiment

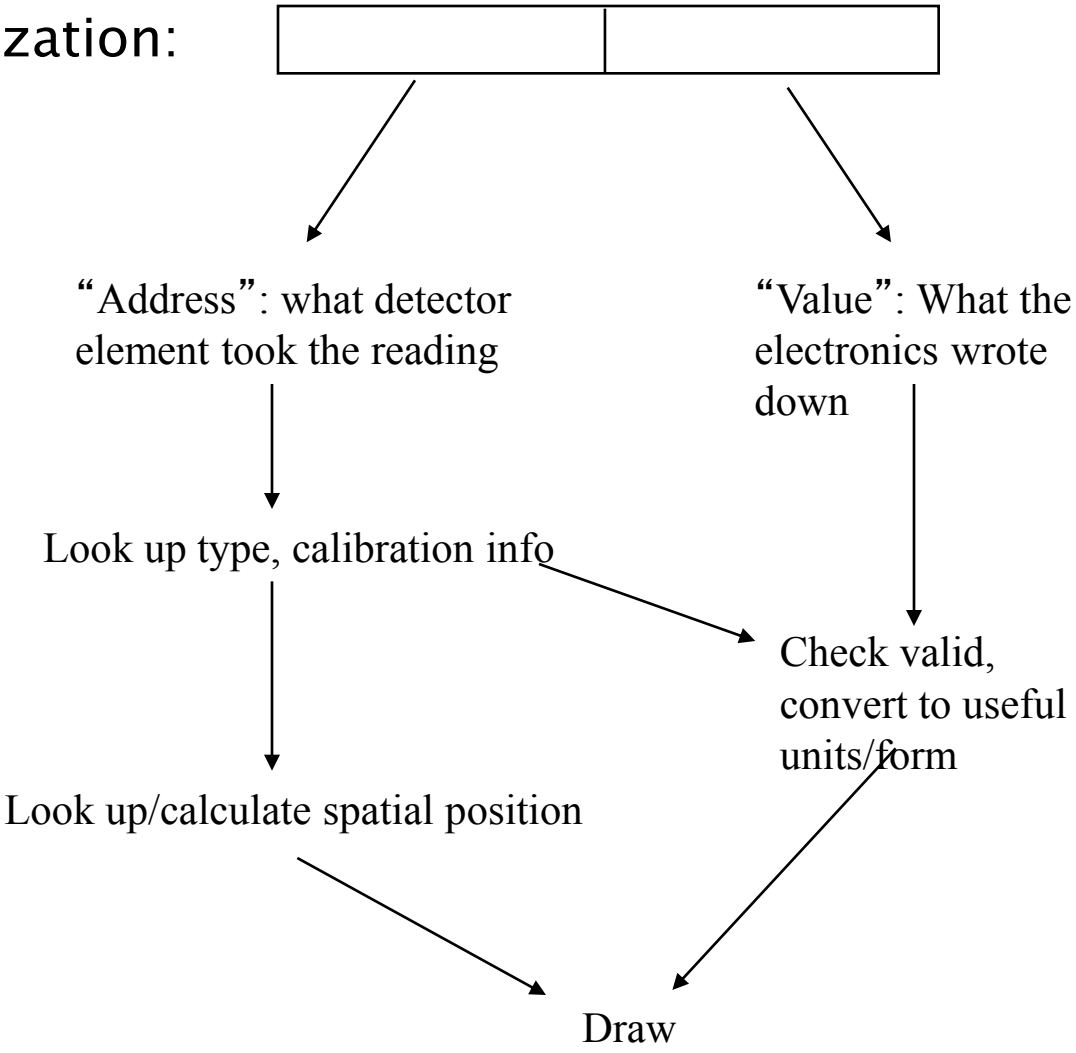
```
0x01e84c10:  0x01e8 0x8848 0x01e8 0x83d8 0x6c73 0x6f72 0x7400 0x0000
0x01e84c20:  0x0000 0x0019 0x0000 0x0000 0x01e8 0x4d08 0x01e8 0x5b7c
0x01e84c30:  0x01e8 0x87e8 0x01e8 0x8458 0x7061 0x636b 0x6167 0x6500
0x01e84c40:  0x0000 0x0019 0x0000 0x0000 0x0000 0x0000 0x01e8 0x5b7c
0x01e84c50:  0x01e8 0x8788 0x01e8 0x8498 0x7072 0x6f63 0x0000 0x0000
0x01e84c60:  0x0000 0x0019 0x0000 0x0000 0x0000 0x0000 0x01e8 0x5b7c
0x01e84c70:  0x01e8 0x8824 0x01e8 0x84d8 0x7265 0x6765 0x7870 0x0000
0x01e84c80:  0x0000 0x0019 0x0000 0x0000 0x0000 0x0000 0x01e8 0x5b7c
0x01e84c90:  0x01e8 0x8838 0x01e8 0x8518 0x7265 0x6773 0x7562 0x0000
0x01e84ca0:  0x0000 0x0019 0x0000 0x0000 0x0000 0x0000 0x01e8 0x5b7c
0x01e84cb0:  0x01e8 0x8818 0x01e8 0x8558 0x7265 0x6e61 0x6d65 0x0000
0x01e84cc0:  0x0000 0x0019 0x0000 0x0000 0x0000 0x0000 0x01e8 0x5b7c
0x01e84cd0:  0x01e8 0x8798 0x01e8 0x8598 0x7265 0x7475 0x726e 0x0000
0x01e84ce0:  0x0000 0x0019 0x0000 0x0000 0x0000 0x0000 0x01e8 0x5b7c
0x01e84cf0:  0x01e8 0x87ec 0x01e8 0x85d8 0x7363 0x616e 0x0000 0x0000
0x01e84d00:  0x0000 0x0019 0x0000 0x0000 0x0000 0x0000 0x01e8 0x5b7c
0x01e84d10:  0x01e8 0x87e8 0x01e8 0x8618 0x7365 0x7400 0x0000 0x0000
0x01e84d20:  0x0000 0x0019 0x0000 0x0000 0x0000 0x0000 0x01e8 0x5b7c
0x01e84d30:  0x01e8 0x87a8 0x01e8 0x8658 0x7370 0x6c69 0x7400 0x0000
0x01e84d40:  0x0000 0x0019 0x0000 0x0000 0x0000 0x0000 0x01e8 0x5b7c
0x01e84d50:  0x01e8 0x8854 0x01e8 0x8698 0x7374 0x7269 0x6e67 0x0000
0x01e84d60:  0x0000 0x0019 0x0000 0x0000 0x0000 0x0000 0x01e8 0x5b7c
0x01e84d70:  0x01e8 0x875c 0x01e8 0x86d8 0x7375 0x6273 0x7400 0x0000
0x01e84d80:  0x0000 0x0019 0x0000 0x0000 0x0000 0x0000 0x01e8 0x5b7c
0x01e84d90:  0x01e8 0x87c0 0x01e8 0x8718 0x7377 0x6974 0x6368 0x0000
```

1/5000th of an event in the CMS detector

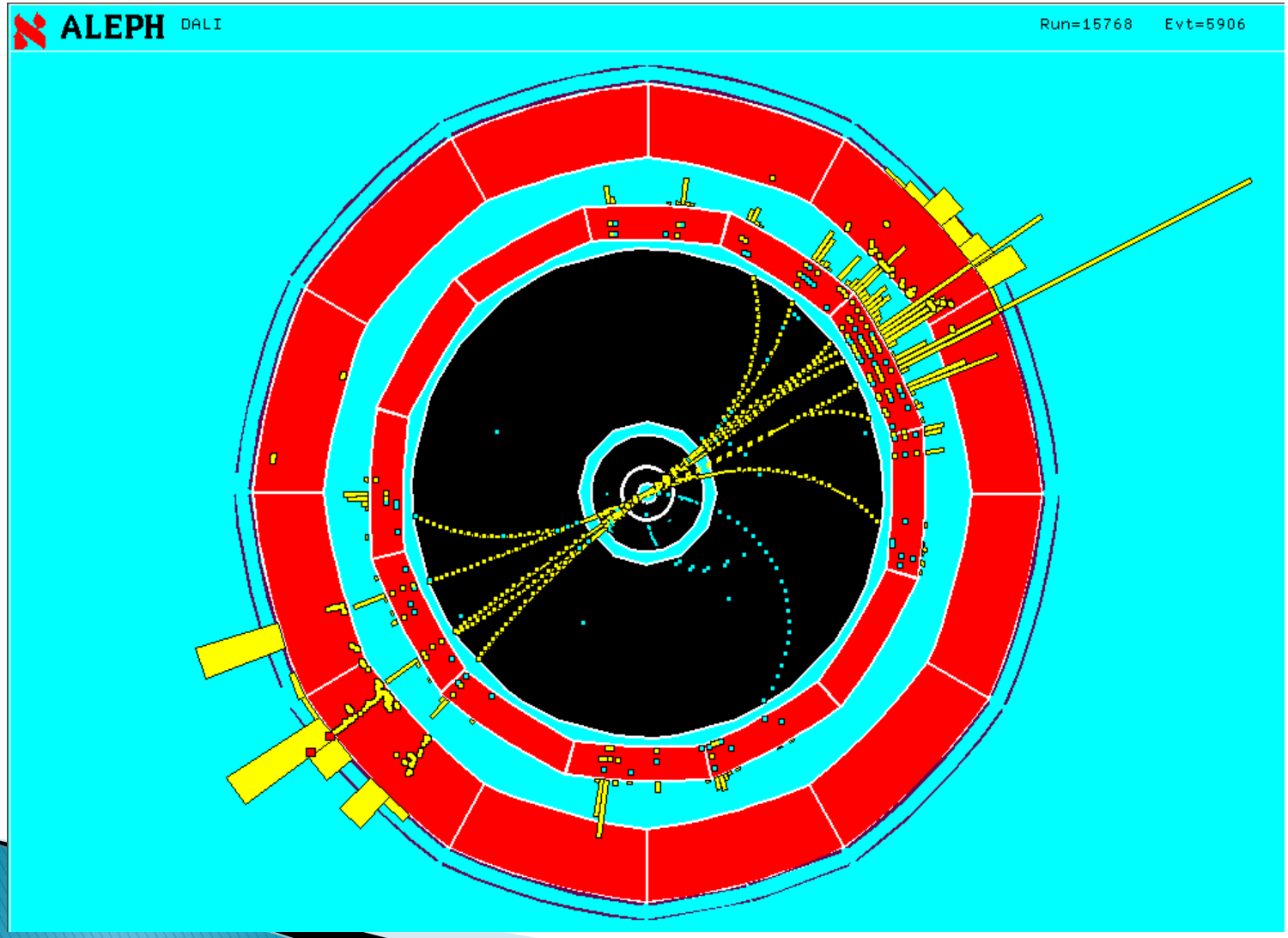
- Get about 200 events per second

What does the Data Mean?

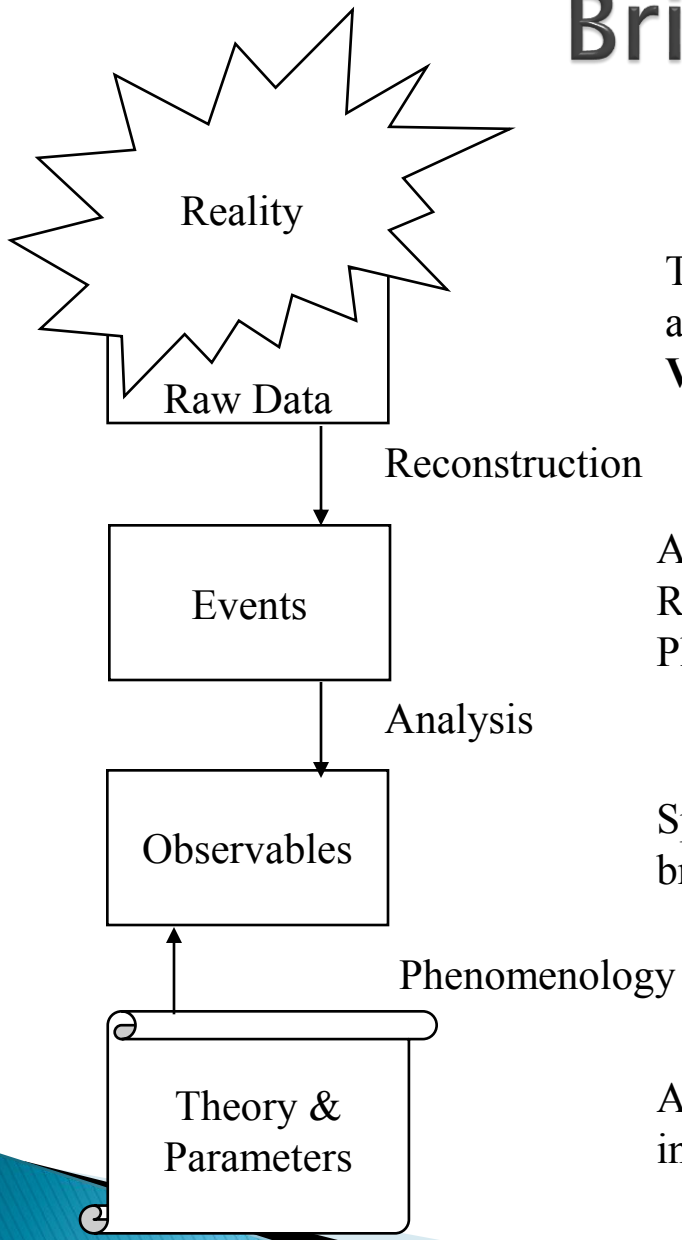
- Digitization:



Graphical Representation



Bridging the Gap



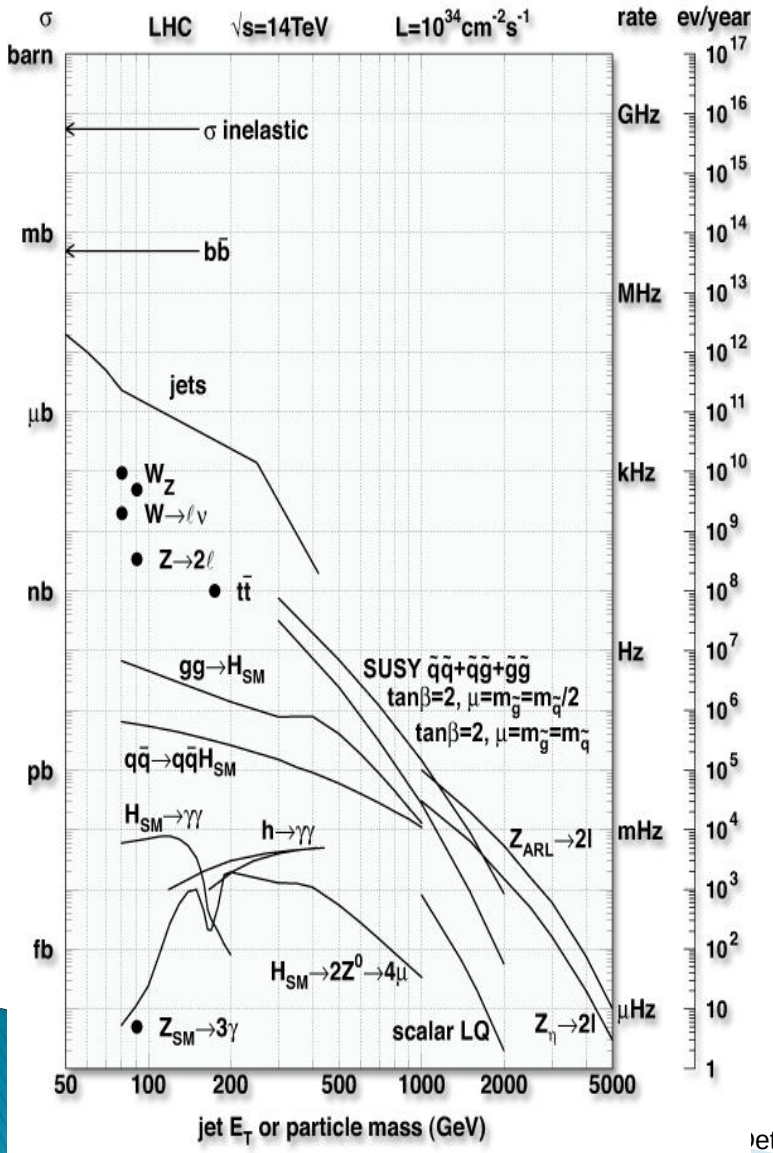
The imperfect measurement of
a (set of) interactions in the detector
Very strong selection

A unique happening:
Run 21007, event 3916 which contains a H -> xx decay
Physical quantities: positions, momentum, energy, etc.

Specific lifetimes, probabilities, masses,
branching ratios, interactions, etc

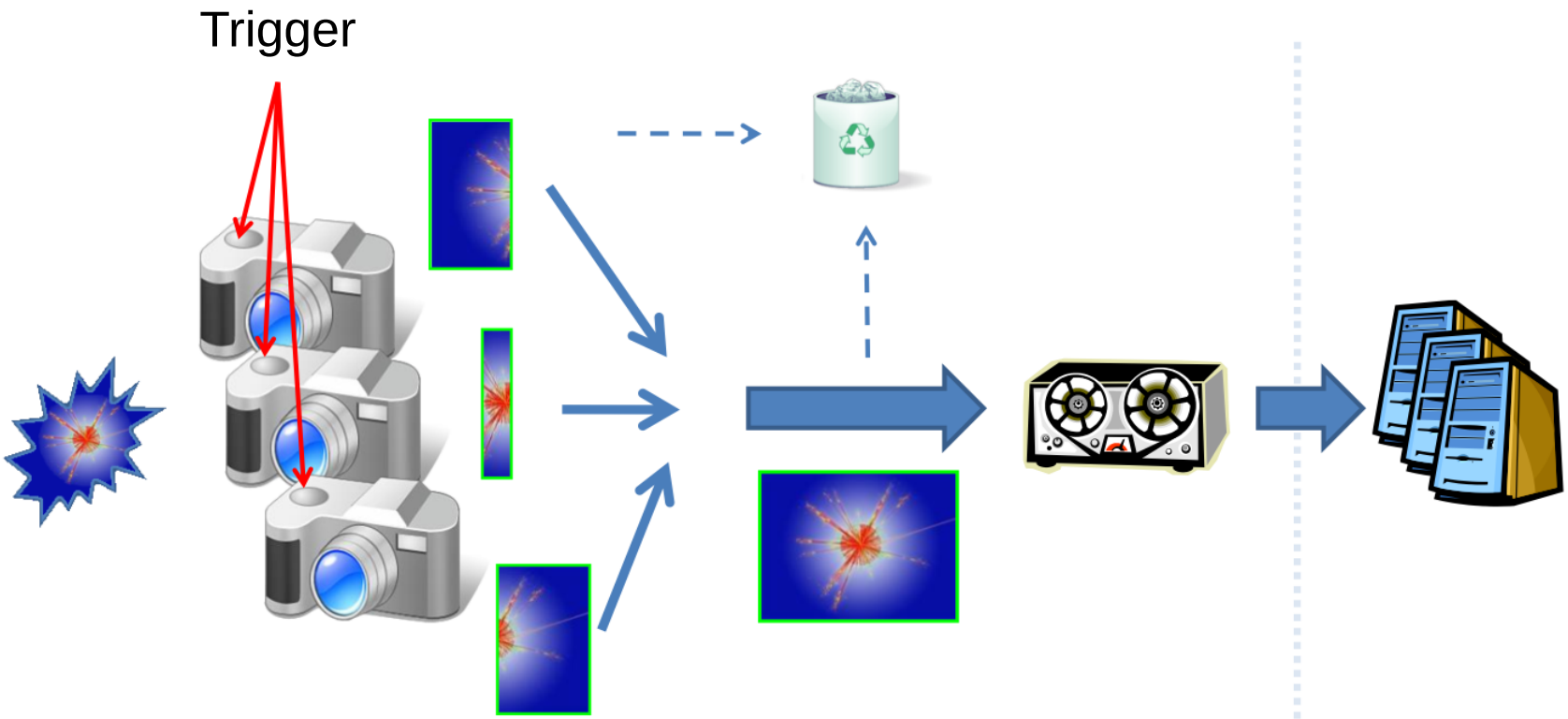
A small number of general equations, with specific
input parameters (perhaps poorly known)

Physics Selection at LHC



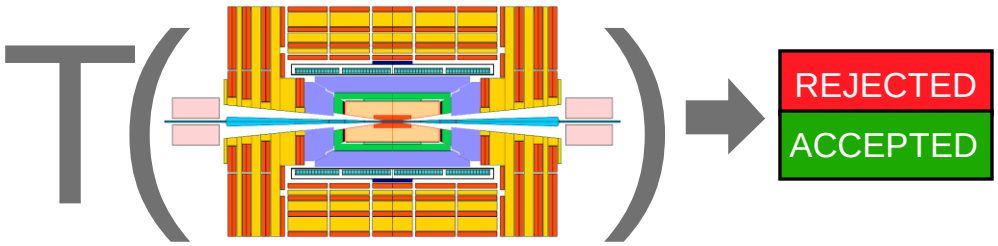
- Cross sections of physics processes vary over many orders of magnitude
  - Inelastic: 10^9 Hz
  - W$\rightarrow \ell \nu$: 10^2 Hz
  - t t production: 10 Hz
  - Higgs (100 GeV/c^2): 0.1 Hz
  - Higgs (600 GeV/c^2): 10^{-2} Hz
- QCD background
  - Jet E_T ~250 GeV: rate = 1 kHz
  - Jet fluctuations $\rightarrow$ electron bkg
  - Decays of K, π, b $\rightarrow$ muon bkg
- Selection needed: 1:10^{10-11}
  - Before branching fractions…

Trigger and Data Acquisition



Trigger

- Task: inspect detector information and provide a first decision on whether to keep the event or throw it out
- The trigger is a function of event data, detector conditions and parameters



• Detector data not (all) promptly available
• Selection function highly complex
⇒T(...) is evaluated by successive approximations

TRIGGER LEVELS

Trigger Levels

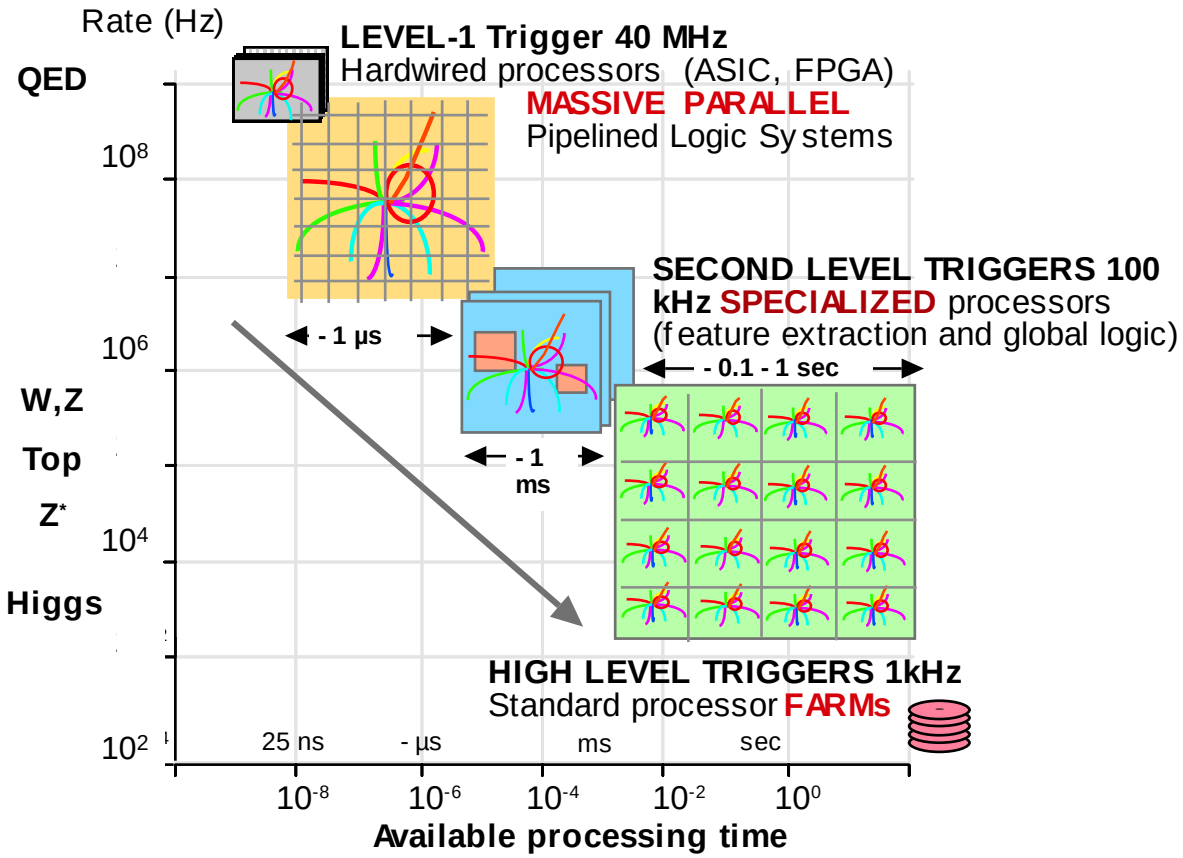
- Level-1
  - Hardwired processors (ASIC, FPGA, …)
  - Pipelined massive parallel
  - Partial information, quick and simple event characteristics (pt, total energy, etc.)
  - 3-4 μs maximum latency

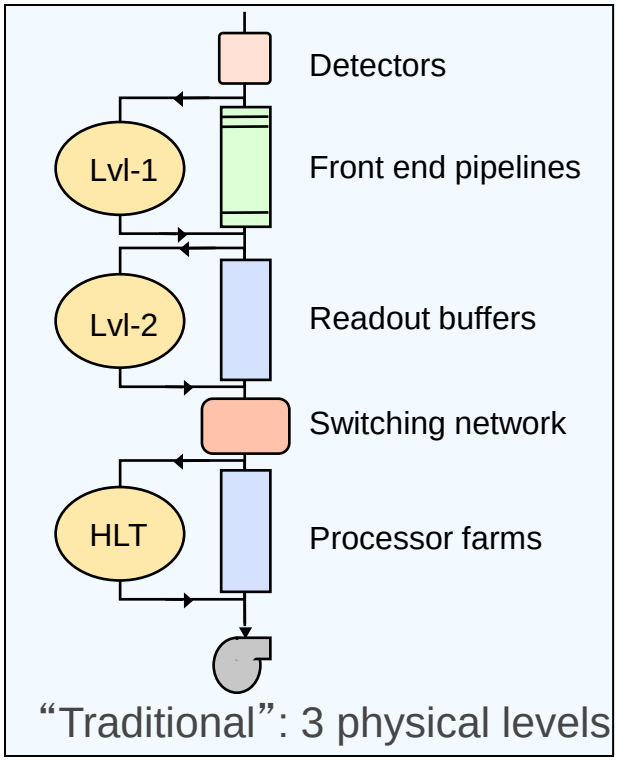
  ~ $1{:}10^4$
- Level-2 (optional)
  - Specialized processors using partial data

  ~ $1{:}10^1$
- High Level
  - Software running in processor farms
  - Complex algorithms using complete event information
  - Latency at the level of factions of second
  - Output rate adjusted to what can be afforded

  ~ $1{:}10^2$

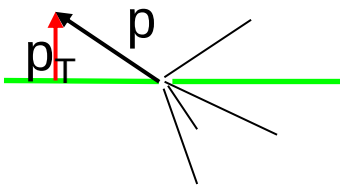
Trigger Levels and Rates





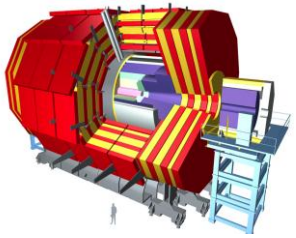
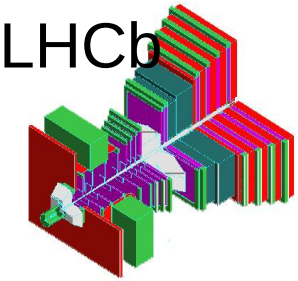
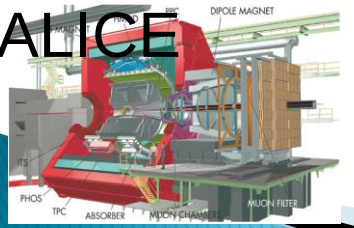
Trigger Level-1 Algorithms

- Physics facts:
  - pp collisions produce mainly hadrons with p_T~1 GeV
  - Interesting physics has particles (leptons and hadrons) with large transverse momenta:
    - $W \rightarrow e\nu$: $M(W) = 80$ GeV/c^2; $P_T(e)$ ~ 30-40 GeV
    - $H(120\ \text{GeV}) \rightarrow \gamma\gamma$: $P_T(\gamma)$ ~ 50-60 GeV



- Basic requirements:
  - Impose high thresholds on particles
    - Implies distinguishing particle types; possible for electrons, muons and "jets"; beyond that, need complex algorithms

Trigger/DAQ Summary for LHC

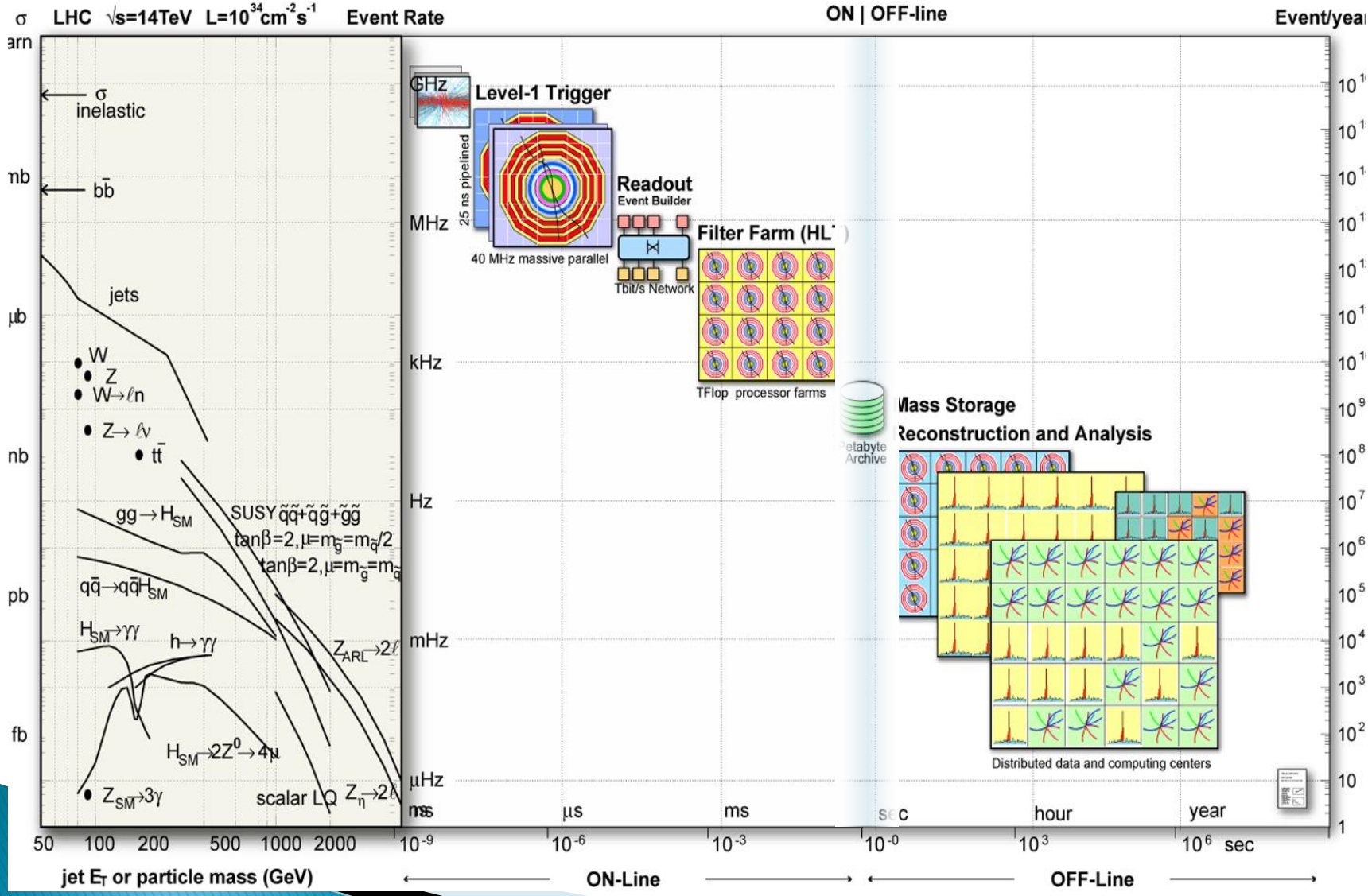
	No.Levels Trigger (HW/SW)		Level-1 Rate (kHz)	Event Size (MB)	Readout Bandw.(GB/s)	Filter Out MB/s (Event/s)
ATLAS 	1/2		75	1.5	10	300 (200)
CMS 	1/1		100	1	100	100 (100)
LHCb 	1/1		1000	0.04	40	80 (2000)
ALICE 	3/1	Pb-Pb p-p	10 200	86.5 2.5	5	1250 (14) 200 (80)

Implementation technologies

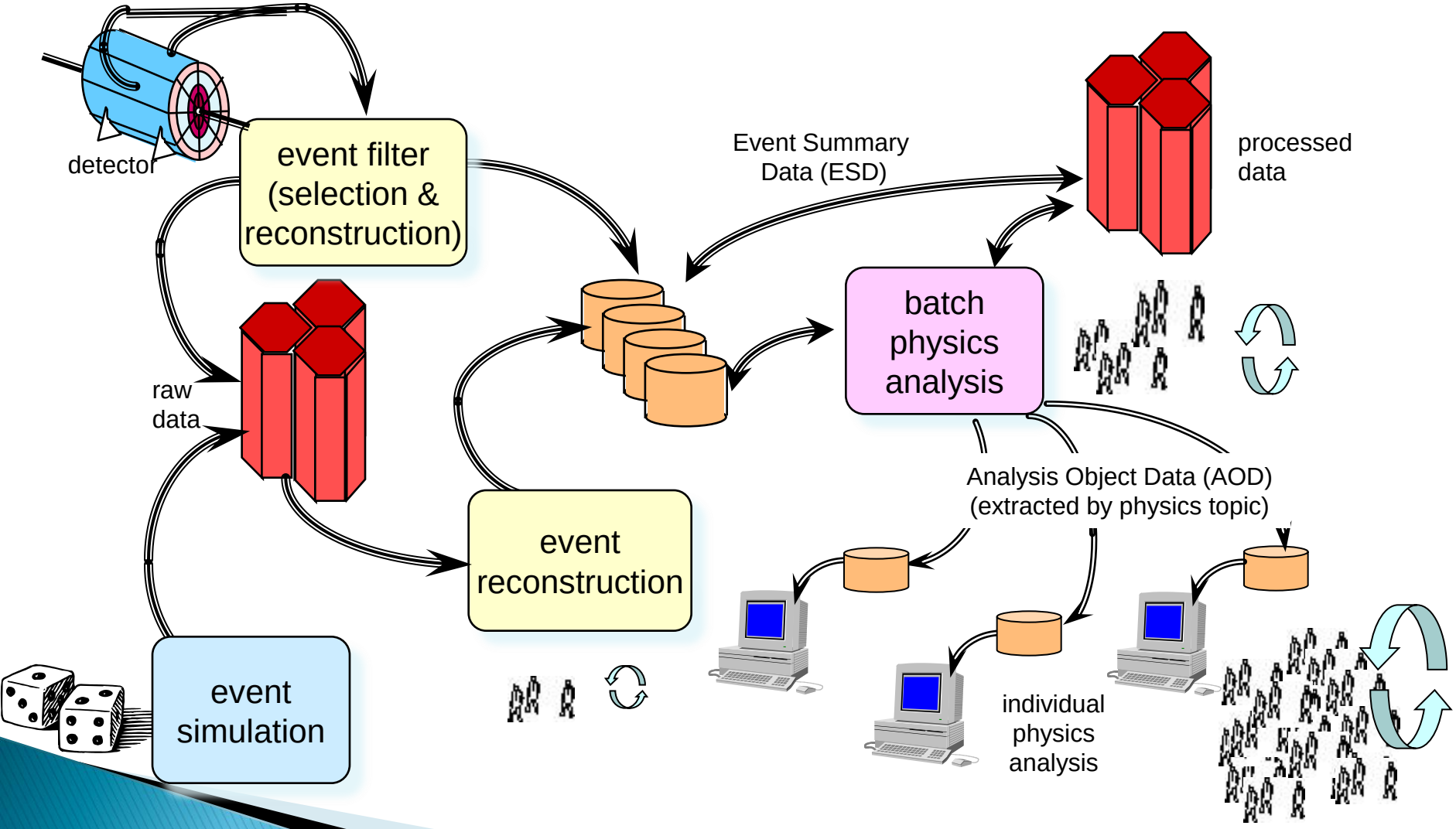
- Readout and Level-1 Trigger
  - Custom electronics (ASIC, FPGA), radiation hard/tolerant
  - Optical detector links (1-2 Gb/s)
- Event Building
  - Gigabit Ethernet links and switches
- High-Level Trigger
  - Rack mounted PCs (~2500 nodes/experiment)
- Online computing facilities
  - Large power/cooling requirements
  - Local data storage O(100 TB)

Online Computing Facilities

Selection Continues Off-line



Offline Processing Stages & Datasets

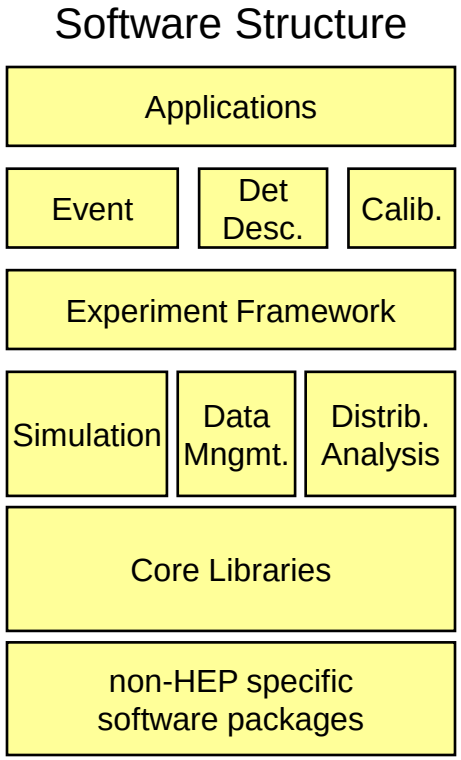


Scientific Software

- The scientific software needed to process this huge amount of data from the LHC detectors is **developed by the LHC collaborations**
  - Must cope with the unprecedented conditions and challenges (trigger rate, data volumes, etc.)
  - Each collaboration has written millions of lines of code
- Modern technologies and methods
  - Object-oriented programming languages and frameworks
  - Re-use of a number of generic and domain-specific 'open-source' packages
- The organization of this large software production activity is by itself a huge challenge
  - Large number of developers distributed worldwide
  - Integration and validation require large efforts

Common Physics Software

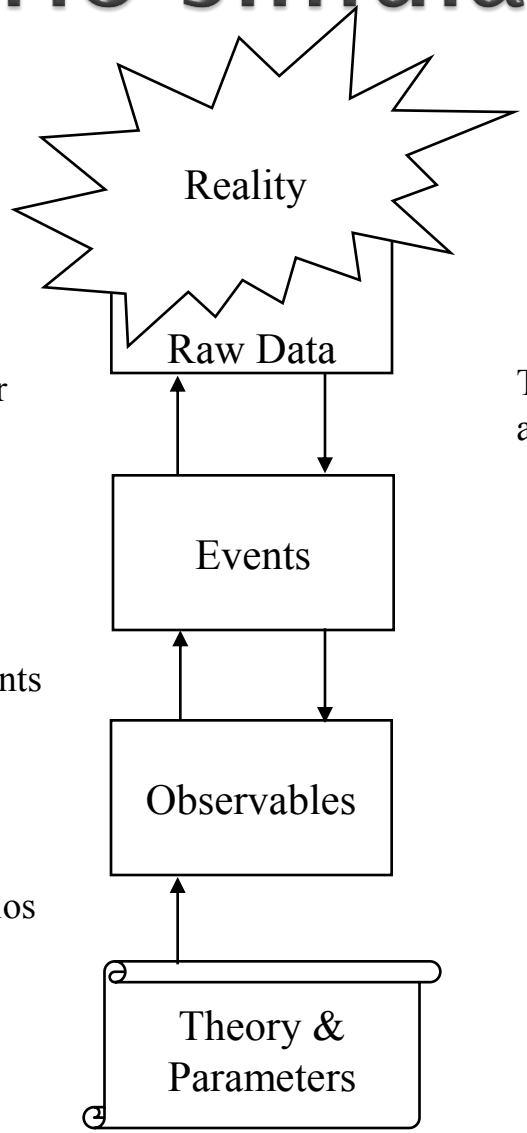
- Data processing applications are based on frameworks
  - Ensure coherency and integration
- Every experiment has a framework for basic services and various specialized frameworks:
  - Event model, detector description, visualization, persistency, interactivity, simulation, calibrations, etc.
- Core libraries and services provide basic functionality. Examples:
  - Geant4 – Simulation of particles through matter
  - ROOT – Data storage and analysis framework
- Extensive use of generic software packages
  - GUI, graphics, utilities, math, db, etc.



Software Components

- Foundation Libraries
  - Basic types
  - Utility libraries
  - System isolation libraries
- Mathematical Libraries
  - Special functions
  - Minimization, Random Numbers
- Data Organization
  - Event Data
  - Event Metadata (Event collections)
  - Detector Conditions Data
- Data Management Tools
  - Object Persistency
  - Data Distribution and Replication
- Simulation Toolkits
  - Event generators
  - Detector simulation
- Statistical Analysis Tools
  - Histograms, N-tuples
  - Fitting
- Interactivity and User Interfaces
  - GUI
  - Scripting
  - Interactive analysis
- Data Visualization and Graphics
  - Event and Geometry displays
- Distributed Applications
- Parallel processing
- Grid computing

Monte Carlo simulation's role



Calculate what imperfect detector would have seen for those events

Treat that as real data and reconstruct it

Randomly pick decay paths, lifetimes, etc for a number of events

Compare to original to understand efficiency

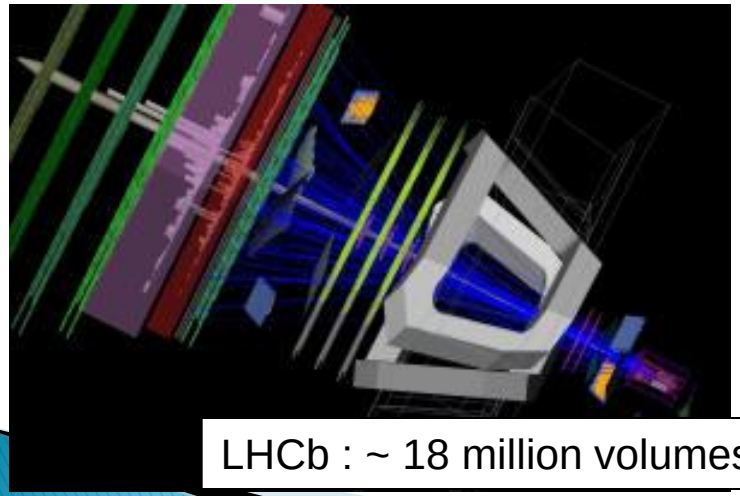
Calculate expected branching ratios

MC Generators

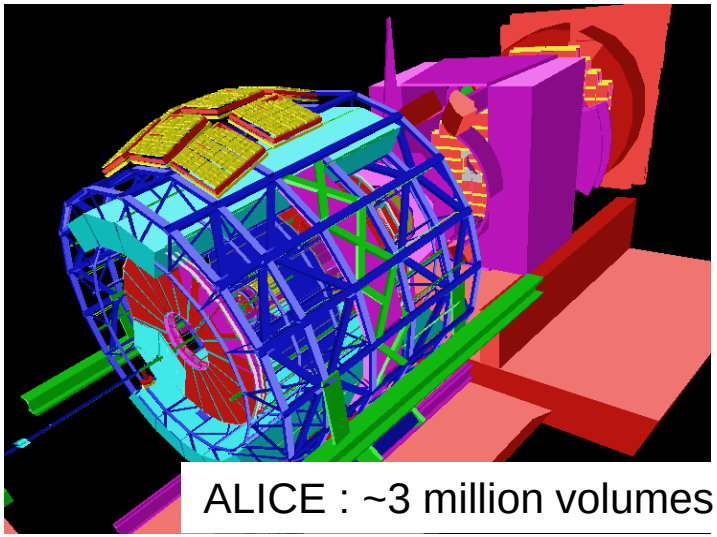
- Many MC generators and tools are available to the experiments provided by a solid community
  - Each experiment chooses the tools more adequate for their physics
- Example: ATLAS alone uses currently
  - Generators
    - AcerMC: Zbb~, tt~, single top, tt~bb~, Wbb~
    - Alpgen (+ MLM matching): W+jets, Z+jets, QCD multijets
    - Charbydis: black holes
    - HERWIG: QCD multijets, Drell-Yan, SUSY...
    - Hijing: Heavy Ions, Beam-gas..
    - MC@NLO: tt~, Drell-Yan, boson pair production
    - Pythia: QCD multijets, B-physics, Higgs production...
  - Decay packages
    - TAUOLA: Interfaced to work with Pythia, Herwig and Sherpa,
    - PHOTOS: Interfaced to work with Pythia, Herwig and Sherpa,
    - EvtGen: Used in B-physics channels.

Detector Simulation – Geant 4

- Geant4 has become the standard tool, in production for the majority of LHC experiments during the past 5 years, and in use in many other HEP experiments and for applications in medical, space and other fields
- On going work in the physics validation
- Good example of common software



LHCb : ~ 18 million volumes

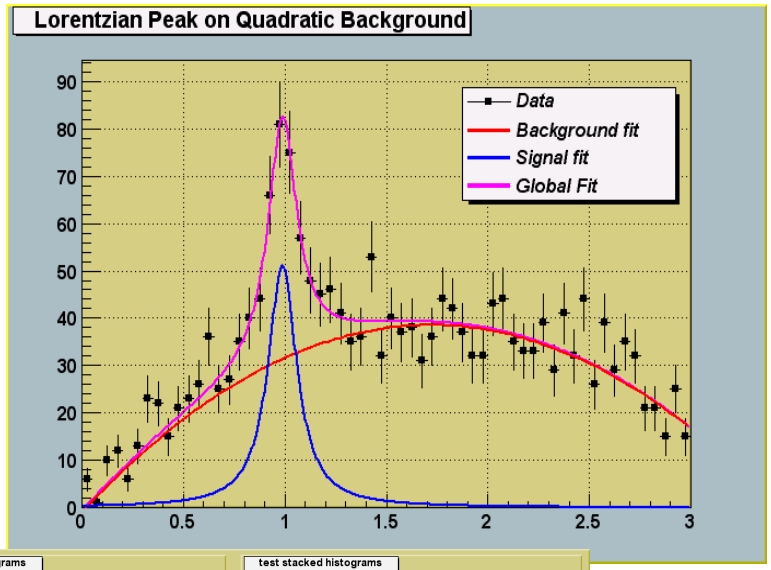


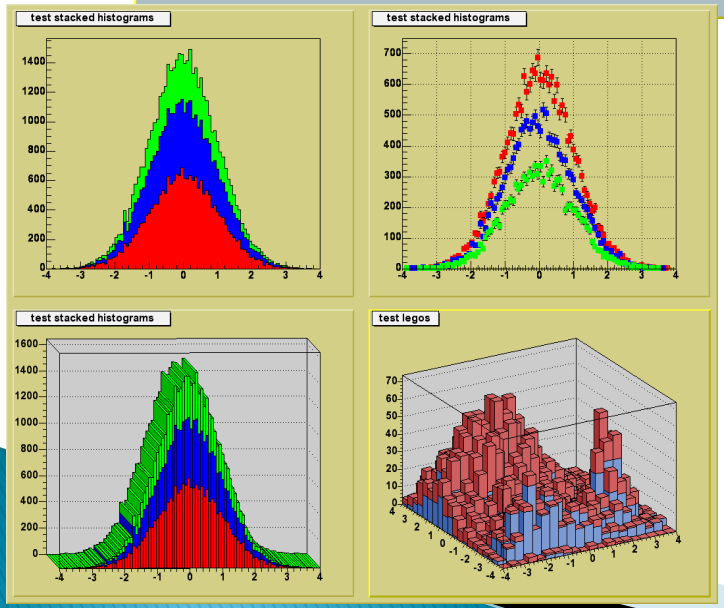
ALICE : ~3 million volumes

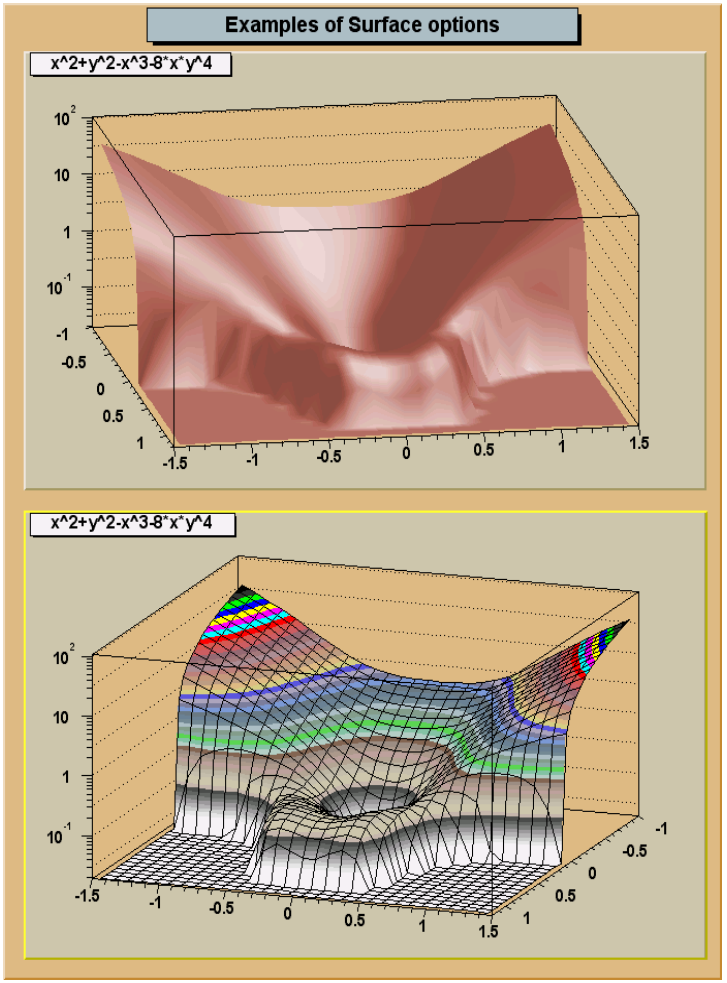
Analysis – Brief History

- 1980s: mainframes, batch jobs, histograms back. Painful.
- Late 1980s, early 1990s: PAW arrives.
  - NTUPLEs bring physics to the masses
  - Workstations with "large" disks (holding data locally) arrive; looping over data, remaking plots becomes easy
- Firmly in the 1990s: laptops arrive;
  - Physics-in-flight; interactive physics in fact.
- Late 1990s: ROOT arrives
  - All you could do before and more. In C++ this time.
  - FORTRAN is still around. The "ROOT-TUPLE" is born
  - Side promise: reconstruction and analysis form a continuum
- 2000s: two categories of analysis physicists: those who can only work off the ROOT-tuple and those who can create/modify it
- Mid-2000s: WiFi arrives; Physics-in-meeting

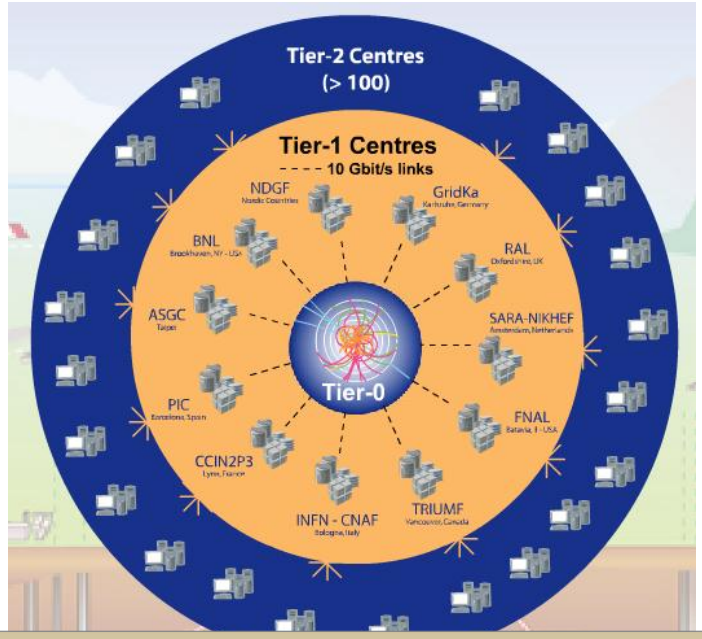
ROOT: Graphics

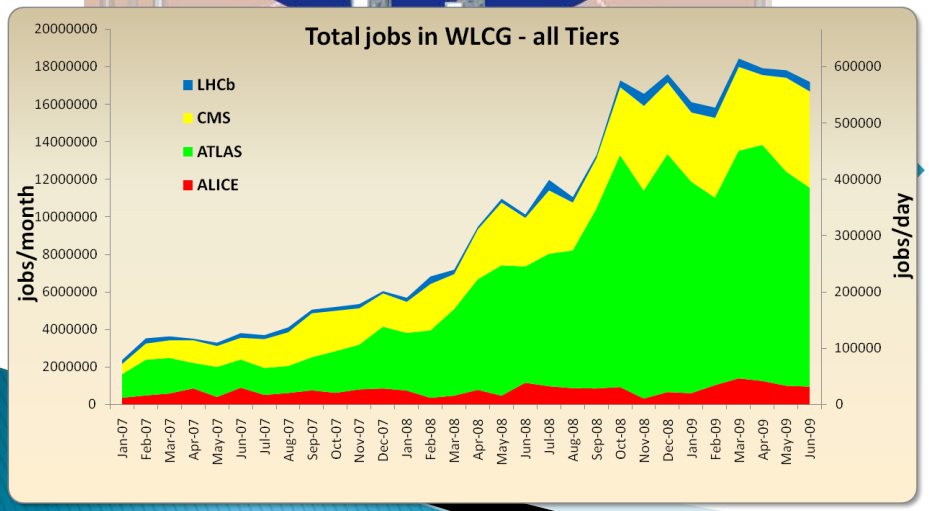






LHC Computing Grid

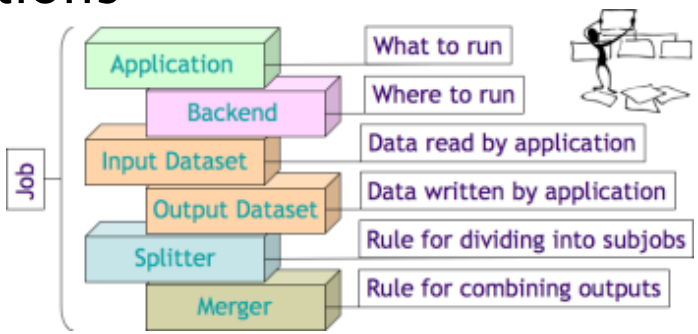




- Tier-0 (CERN): data recording, initial data reconstruction, data distribution
- Tier-1 (11 centres): permanent storage, re-processing, analysis
- Tier-2 (~130 centres): simulation, end-user analysis
- Installed Capacity (Sep-2009)
  - CPU: 78.600 kSI2K (~40.000 cores)
  - Disk: 31 PB
  - Tape: 49 PB
- 600.000 job/day
  - + 45% in a year

Application Software on the Grid

- Experiments have developed tools to facilitate the usage of the Grid
- Example: GANGA
  - Help configuring and submitting analysis jobs (Job Wizard)
  - Help users to keep track of what they have done
  - Hide completely all technicalities
  - Provide a palette of possible choices and specialized plug-ins:
    - pre-defined application configurations
    - batch/grid systems, etc.
  - Single desktop for a variety of tasks
  - Friendly user interface is essential



Summary

- The online multi-level trigger is essential to select interesting collisions (1 in 10^6–10^7)
  - Level-1: custom hardware, huge fanin/out problem, fast algorithms on coarse-grained, low-resolution data
  - HTL: software/algorithms on large processor farm of PCs
  - Large DAQ system built with commercial components
- The experiments produce about 7-10 PB/year raw data
- Reconstruction and analysis to get from raw data to physics results
  - Huge programs (10^7 lines of code) developed by 100's of physicists
  - Unprecedented need of computing resources