User perspective of a data analysis flow with the ATLAS detector

G. Carrillo-Montoya, on behalf of the ATLAS Collaboration
University of the Witwatersrand

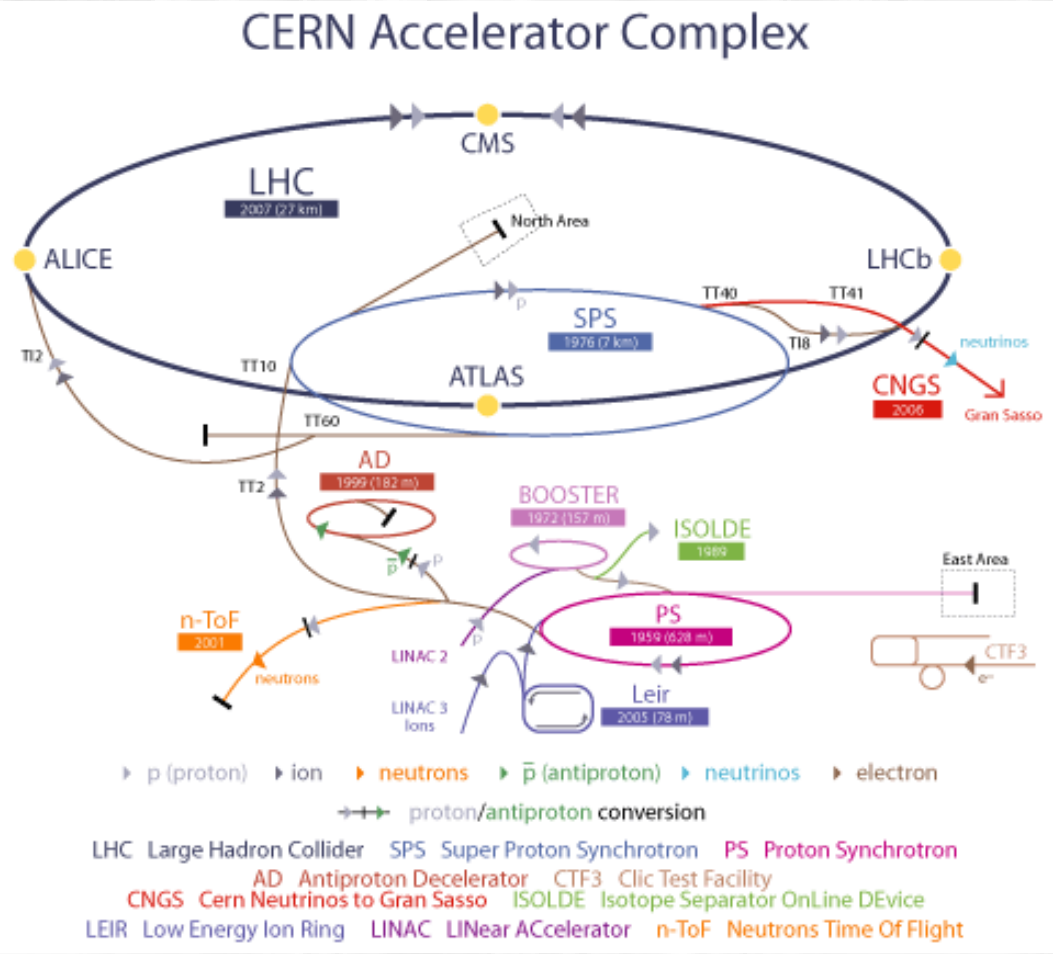
High-performance Signal and Data Processing Workshop
South Africa
January 28th - 2014

Introduction

- My goal is to drive you through the different steps a of typical (Higgs) analysis: search/measurement.
- More than the global strategy and infrastructure details → point out examples and few relevant technical aspects.
- Higgs discovery → nicely presented by Peter yesterday
- The ATLAS computing model → will be described in detail by Sahal (next talk)
- ATLAS readout/upgrade → to be presented by Alberto
- Elecronics: Tile Calorimeter → to be presented by Carlos
- ...

The **L**arge **H**adron **C**ollider

- 26.659 meters
- 9.300 magnets
- 60 tonnes of Liquid He
- 30+ years
- 6 main experiments
- 2 Kelvin



centre of mass energy: designed $\sqrt{s}$ =14 TeV (ran at 7 TeV in 2011 & 8 TeV in 2012)

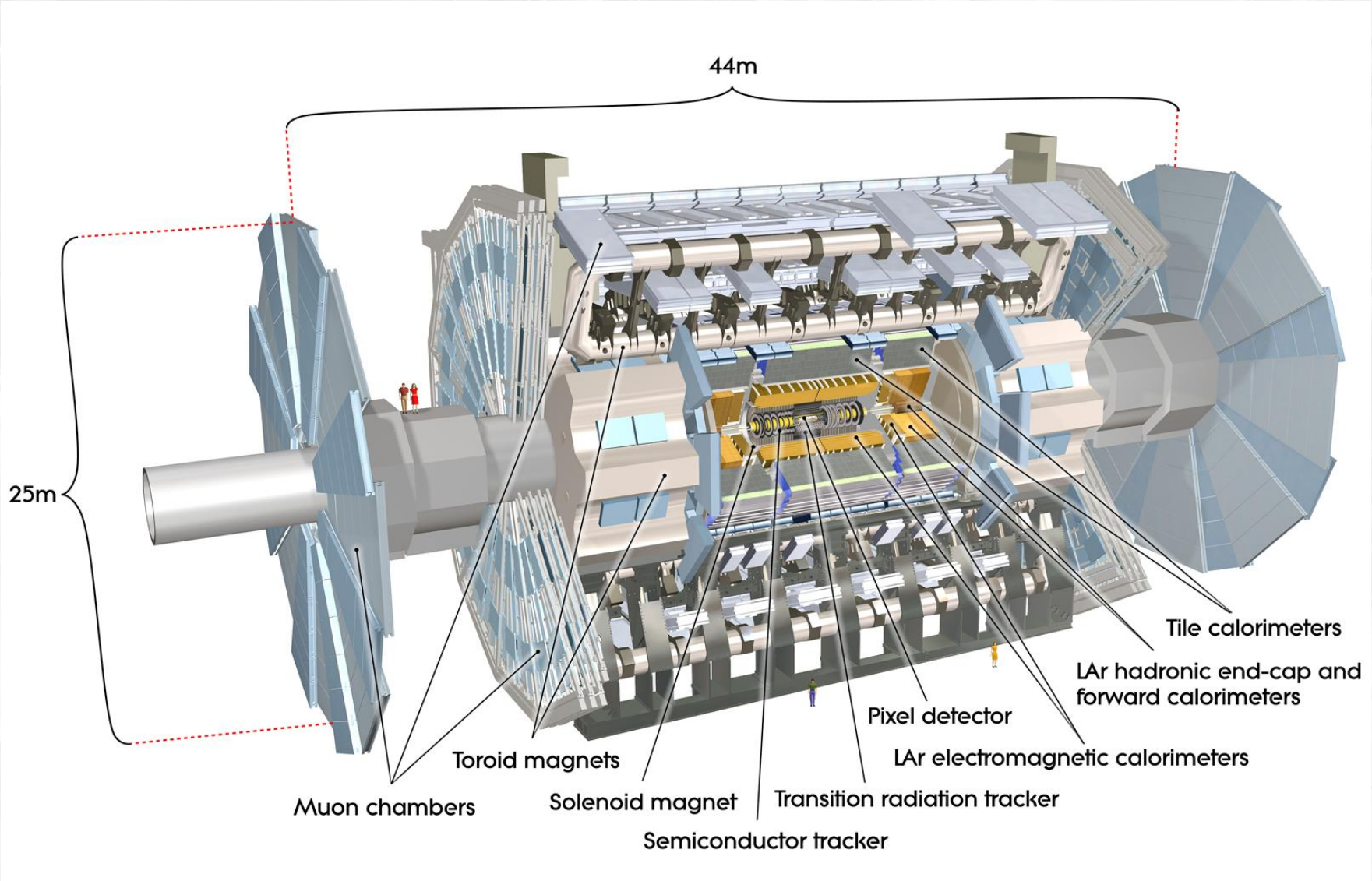
Peak instantaneous luminosity: designed $> 10^{34}$ cm^{-2}s^{-1}

The LHC experiments

- ATLAS, CMS $\rightarrow$ multipurpose detectors
- LHCb $\rightarrow$ forward, asymmetric, vertex (probe *CP*-violating and rare decays)
- ALICE $\rightarrow$ Heavy ion collisions, occupancy - large event multiplicity (QGP)
- Totem LHCf $\rightarrow$ far from IP (σ_{Tot}, proton structure, atmospheric shower)

- Exploring
  - Understanding the Nature of EW symmetry breaking (Hunting Higgs boson)
  - Super Symmetry
  - Matter-Anti Matter asymmetry
  - Quark-gluon plasma
  - Precision measurements $\rightarrow$ New Physics (∞)

A **T**oroidal **L**HC **A**pparatu**S** - 4π solid angle

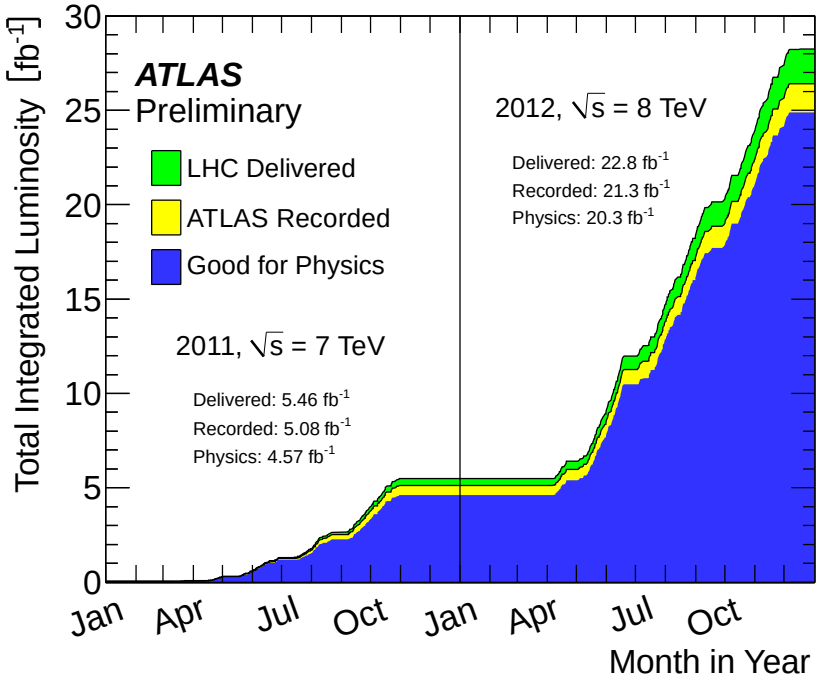


ATLAS Detector DQ Status

Subdetector	Number of Channels	Approximate Operational Fraction
Pixels	80 M	95.0%
SCT Silicon Strips	6.3 M	99.3%
TRT Transition Radiation Tracker	350 k	97.5%
LAr EM Calorimeter	170 k	99.9%
Tile calorimeter	9800	98.3%
Hadronic endcap LAr calorimeter	5600	99.6%
Forward LAr calorimeter	3500	99.8%
LVL1 Calo trigger	7160	100%
LVL1 Muon RPC trigger	370 k	100%
LVL1 Muon TGC trigger	320 k	100%
MDT Muon Drift Tubes	350 k	99.7%
CSC Cathode Strip Chambers	31 k	96.0%
RPC Barrel Muon Chambers	370 k	97.1%
TGC Endcap Muon Chambers	320 k	98.2%

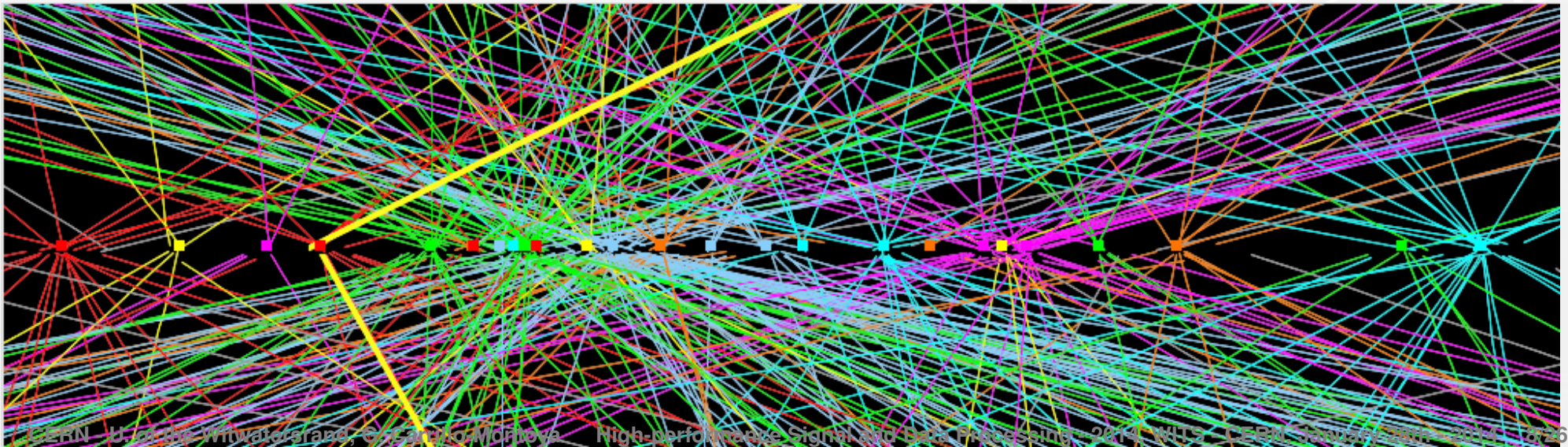
Content

Luminosity → pileup interactions



- During 2012, the LHC provided us larger integrated luminosities, still with 50 ns bunch separation:

- More interactions per bunch crossing.
- Reconstruction of objects need even more robust methods → performance of physics analysis:
  - More CPU required to process events
  - More disk needed
  - Run-II will be even more challenging!



Tier structure

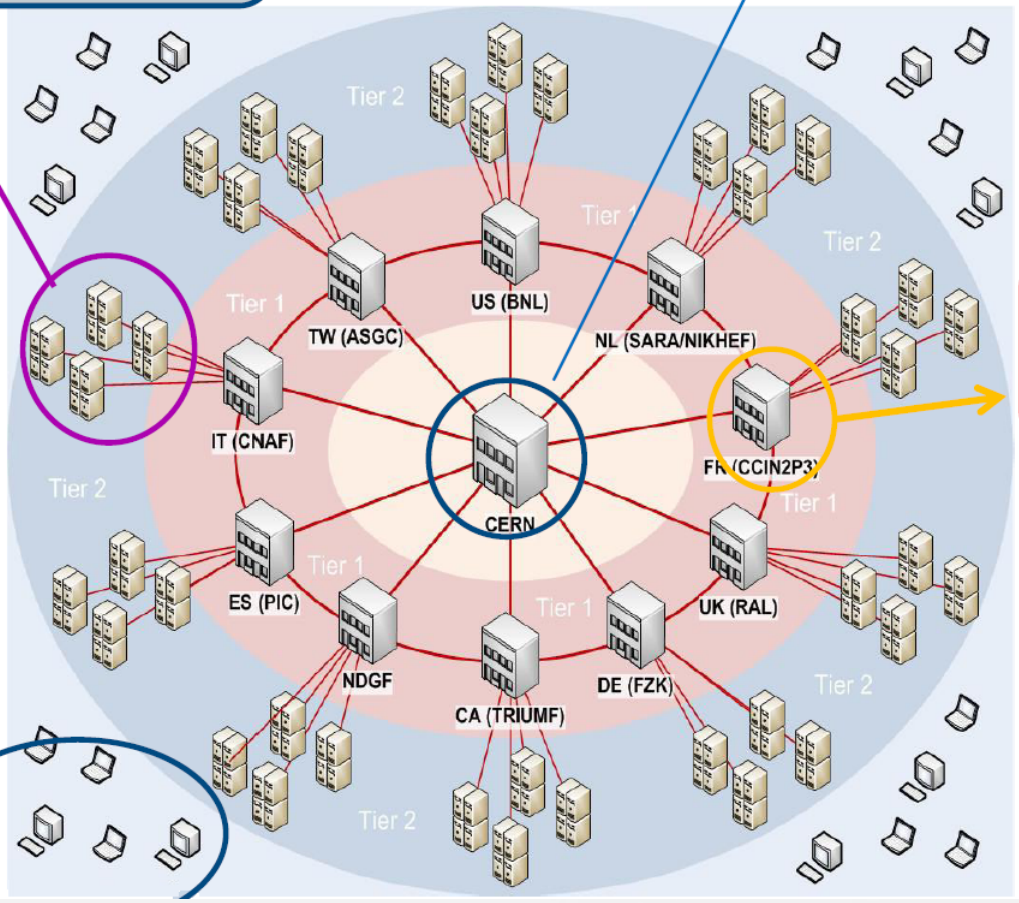
Access to all ATLAS users!

- Monte Carlo Production
- Physics Analysis
- Store Analysis Objects

Tiers2

- Prompt calibration and alignment
- Reconstruction
- Store complete set of RAW data

Tier0



Tiers1

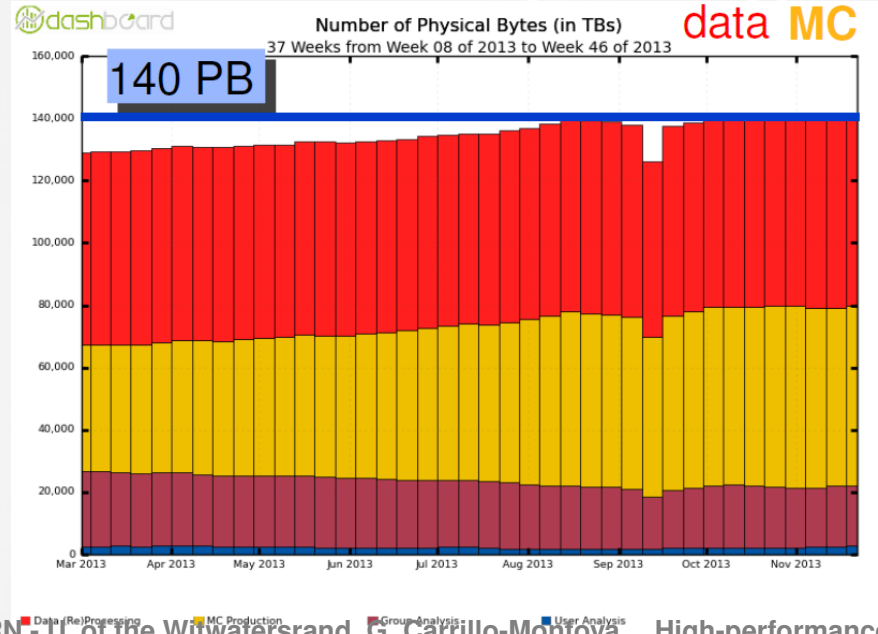
- Data Reprocessing
- Archive RAW and Reconstructed data

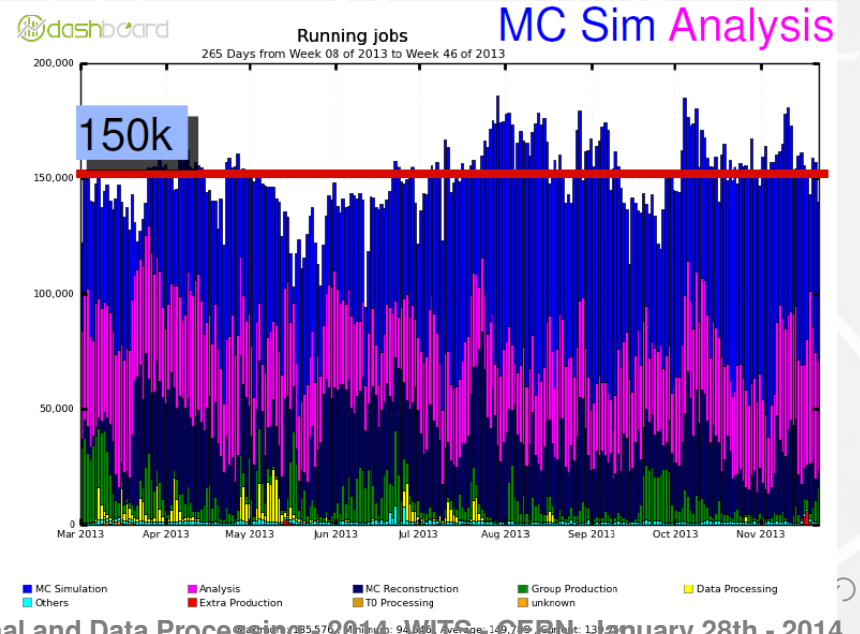
- Physics Analysis,
- Interactive Studies

Tiers3

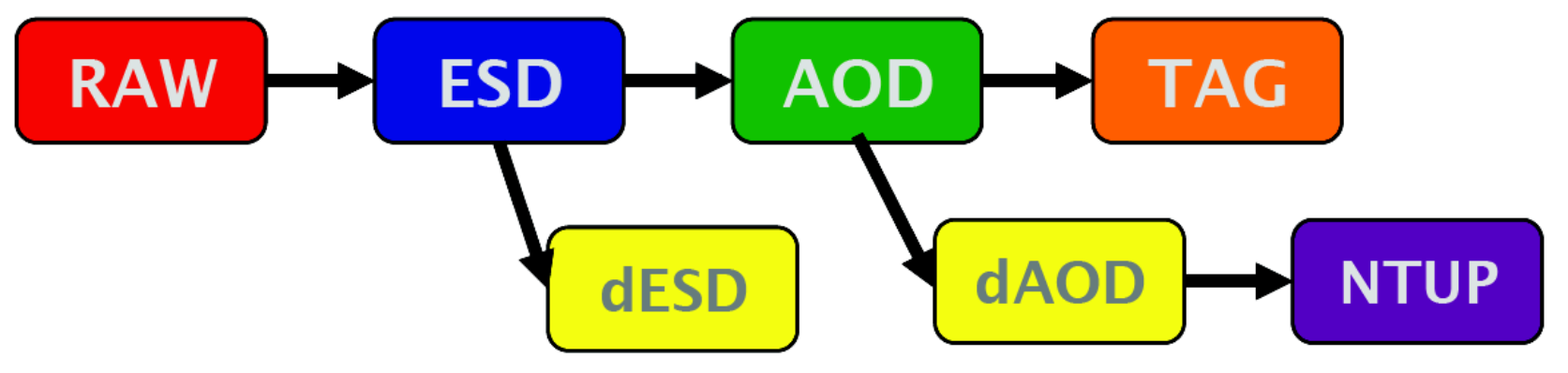
Simulated and Data events - Replicated

- Data:
  - Combining $\sqrt{s} = 7$ TeV and $\sqrt{s} = 8$ TeV (physics and delayed streams) : $\sim$7.6 billion events ($p - p$), $\sim$7.4 PB data volume
- Monte Carlo simulations:
  - $>$ 10 billion events (7 TeV, 8 TeV, various configurations) $\rightarrow$ (10-20 million simulated evnts/day)
  - Each event $\sim$ minutes to be simulated (various kinds of simulations from very fast to very detailed "Full")
  - Sizes vary depending on the format, from few kB to few MB per event





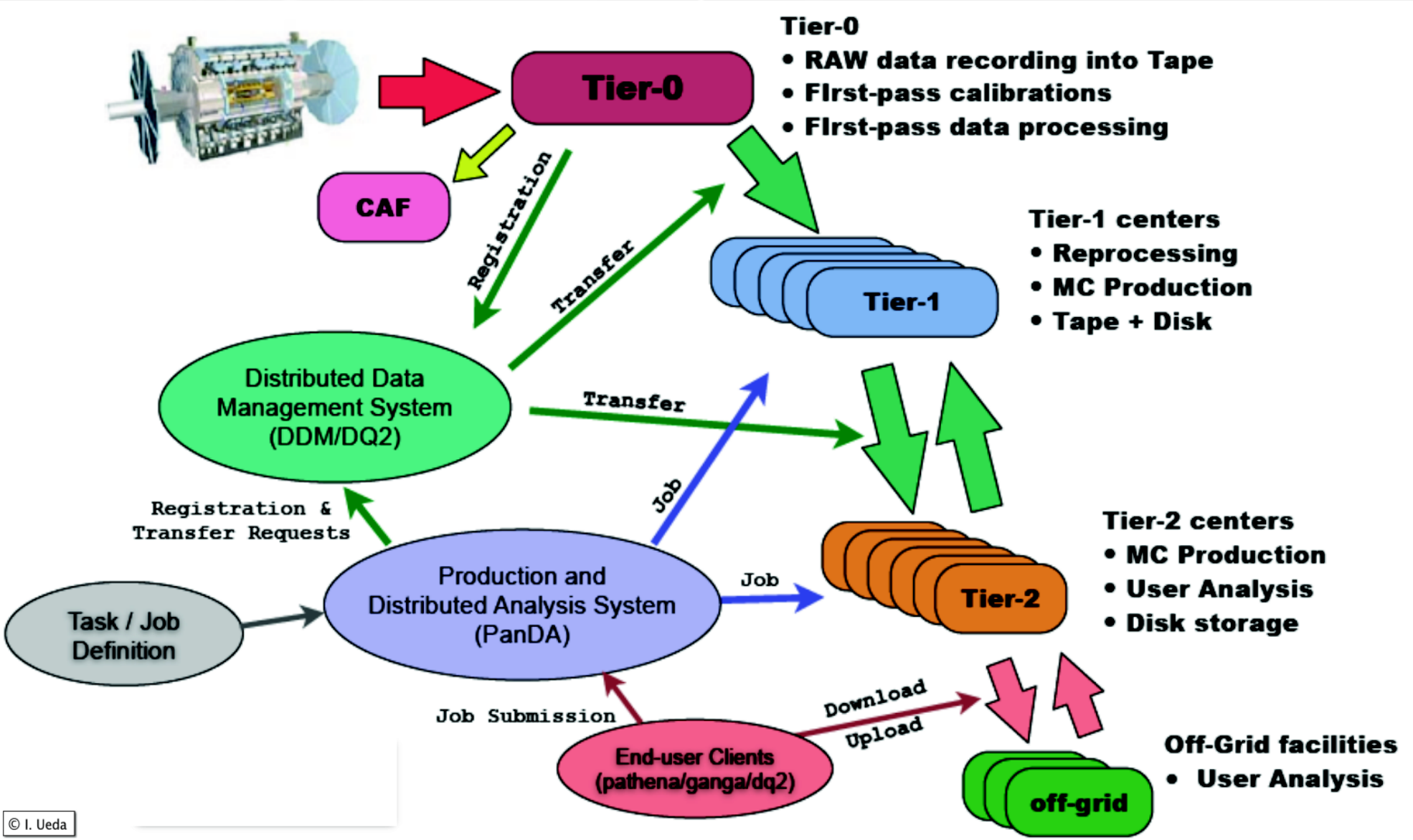
Data Flavors



- RAW: raw data from the detector
- ESD (Event Summary Data): output of reconstruction
- AOD (Analysis Object Data): event representation with reduced information for physics analysis (ATLAS-wide format)
- DPD (Derived Physics Data): representation for end-user analysis. Produced for working groups or individual end-users (group-specific format)
- dESD (performance groups), dAOD (physics groups), NTUP (physics groups and end-users)
- TAG: event-level metadata (event tags), short event summaries primarily for event selection

Content

ATLAS computing system - overview



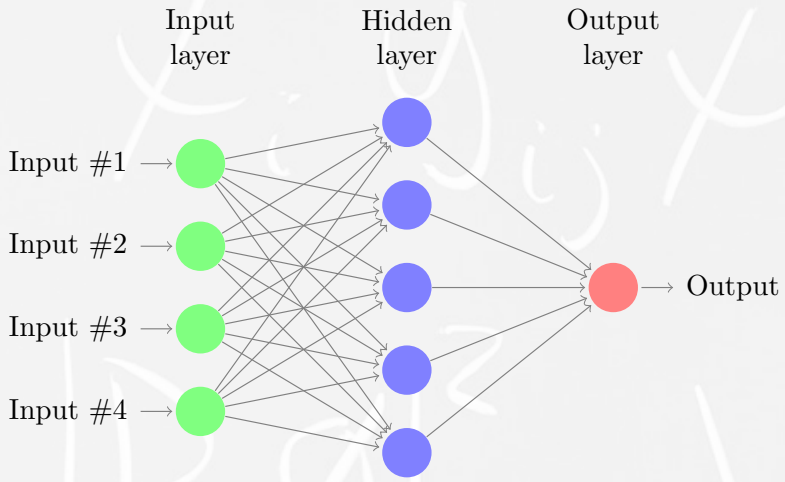
Analysing data

- Final state: like for instance $H \rightarrow ZZ \rightarrow 4\ell$
- Select events with a particular topology (know how is your signal: oppositely charged same flavour lepton pairs with certain kinematics)!
- The most expensive part of an analysis is the design of the strategy:
  - Monte Carlo simulations $\rightarrow$ as much statistics as possible (usually never enough)
  - Iterate over (hundreds of **millions or even billions** of) simulated events **over and over again**:
    - Find optimal selection (Significance like for instance)
    - Figure out possible systematic uncertainties (detector/theory)
    - Design and test methods to extract background processes from data
    - Corrections, calibrations, alignments, etc
    - ...
    - Correct bugs
- mostly **I/O bound** (but not always)

Examples of CPU intensive tasks

Certain tasks inputs/outputs needed are small: **CPU intensive**:

- Kinematic fits
- Multivariate analyses:
  - Matrix Element
  - Boosted Decision Trees
  - Neural Networks
  - Other machine learning algorithms



- Statistical quantification → Toy Monte Carlo (post-unblinding the data), very tight timescales

The golden channel - $H \rightarrow ZZ^{(*)} \rightarrow \ell^+\ell^-\ell^+\ell^-$

- 4-lepton (coming from Z decays: same-flavour, opposite charge) $\rightarrow$ very good resolution, high reconstruction and trigger efficiencies $\rightarrow$ mass peak can be reconstructed
- **Almost background free:** s/b between 0.9 ($4e$) and 1.6 4μ
- Very robust against systematic uncertainties
- Very small yield: signal cross section $\times$ branching ratio ($Z \rightarrow \ell\ell \sim 3\%$).
- Low P_T objects needed to maximise signal acceptance:
  - Muons:
    $P_T > 6$ GeV, $|\eta| < 2.7$
  - Electrons:
    $P_T > 7$ GeV, $|\eta| < 2.47$

Documentation:

Phys.Lett. B726 (2013) 88-119

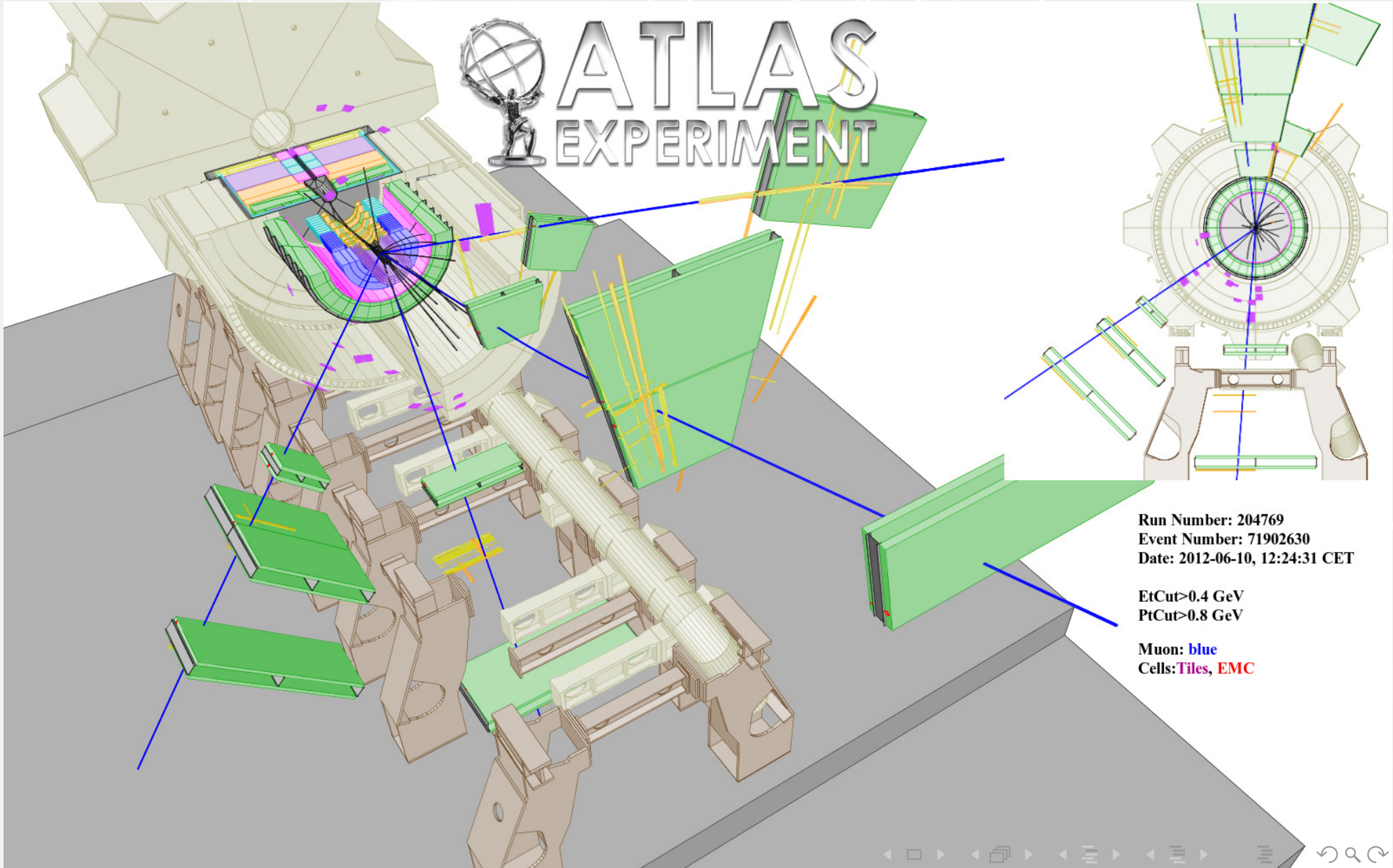
https://cdsweb.cern.ch/record/1460411/files/ATLAS-CONF-2012-092.pdf

Phys.Lett. B716 (2012) 1-29

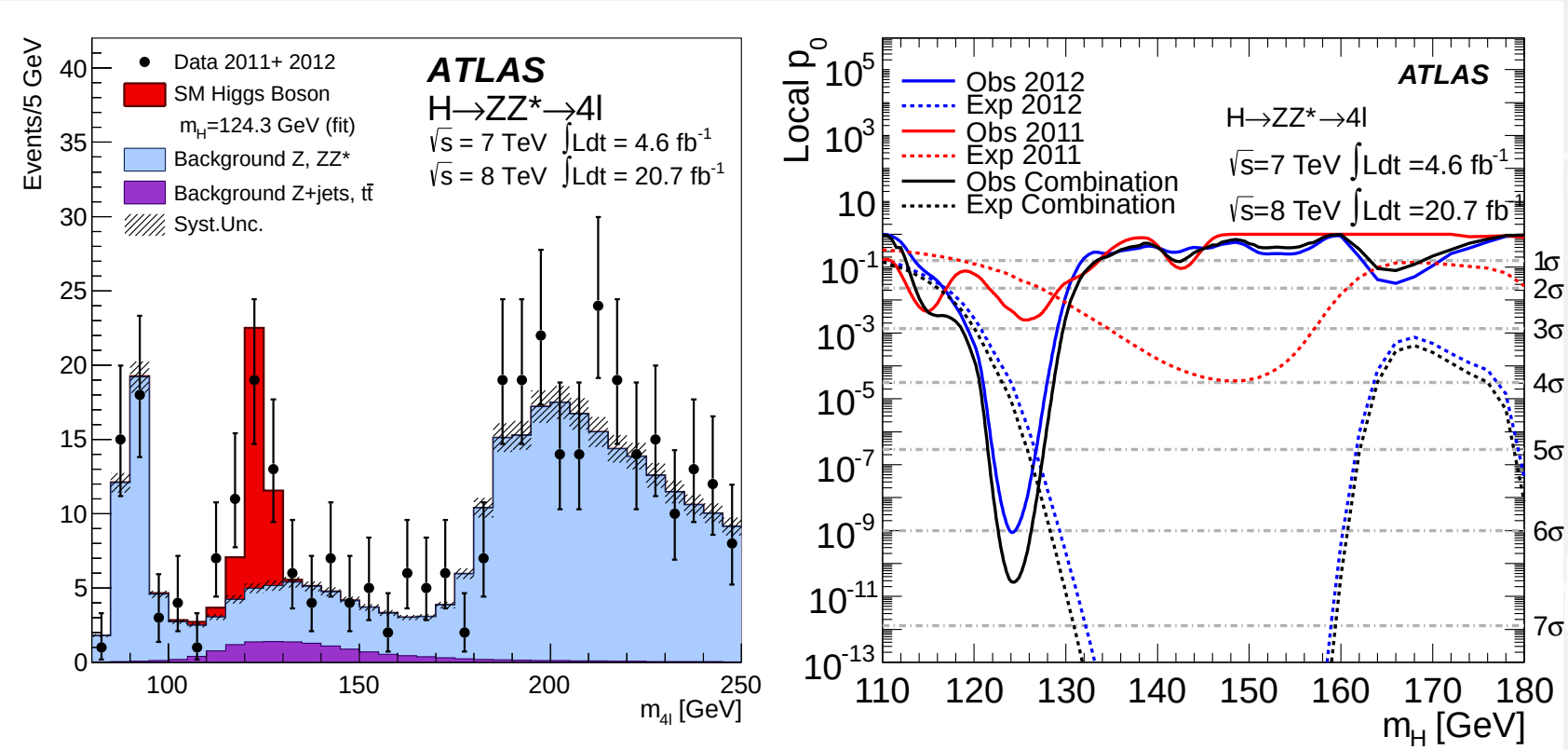
4μ candidate. $m_{4\ell} = 125.1$ GeV, $m_{12} = 86.3$ GeV, $m_{34} = 31.6$ GeV.

μ_1: $P_T = 36.1$ GeV, $\eta = 1.29$, $\phi = 1.33$ μ_2: $P_T = 47.5$ GeV, $\eta = 0.69$, $\phi = -1.65$

μ_3: $P_T = 26.4$ GeV, $\eta = 0.47$, $\phi = -2.51$ μ_4: $P_T = 71.7$ GeV, $\eta = 1.85$, $\phi = 1.65$



Events - $H \rightarrow ZZ \rightarrow 4\ell$ ... needle in a haystack

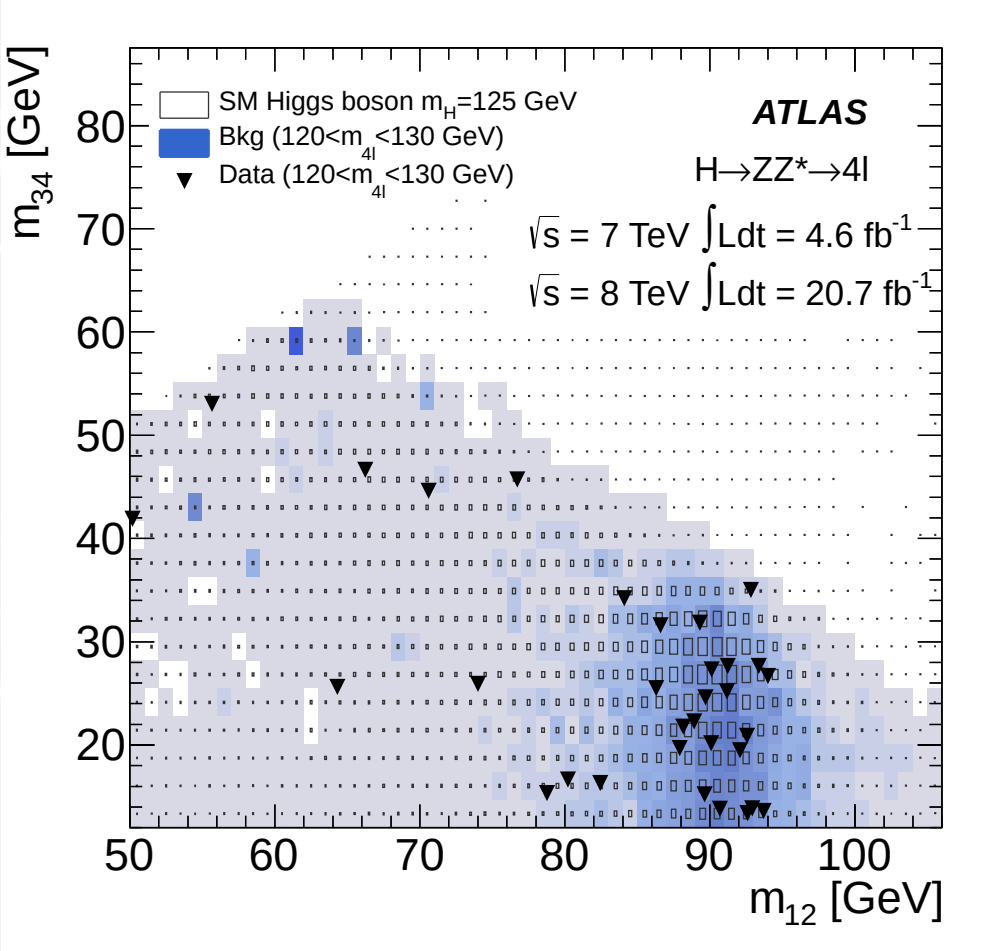


In a $m_{4\ell}$ window around 120-130 GeV:

	Signal	ZZ^*	$Z+\text{jets}, t\bar{t}$	Observed
4μ	6.3±0.8	2.8±0.1	0.55±0.15	13
$2e2\mu/2\mu2e$	7.0±0.6	3.5±0.1	2.11±0.37	13
$4e$	2.6±0.4	1.2±0.1	1.11±0.28	6

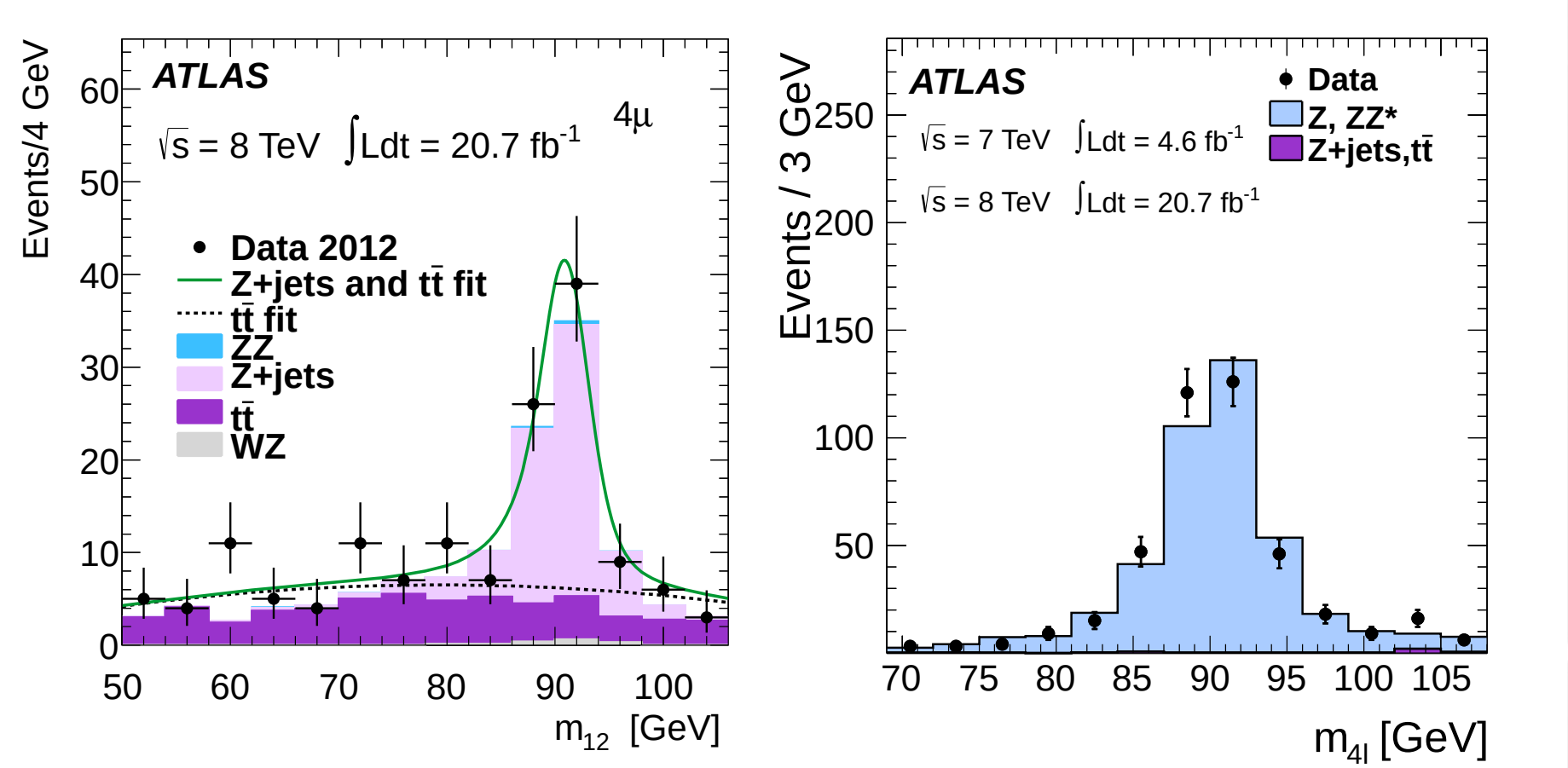
Examples: Cut optimisation

Example: Use a genetic algorithm to optimise the optimal requirement on the invariant mass of the lepton pair m_{12} and m_{34} that maximises the significance $Z = S/sqrtS + B$, in reality it is a bit more complicated than this (Poisson regime, systematic uncertainties, etc).



Optimisations aim a certain luminosity target. (with non negligible backgrounds, pretty much always the case in Higgs analysis in hadron colliders, optimal values depend on the statistics)

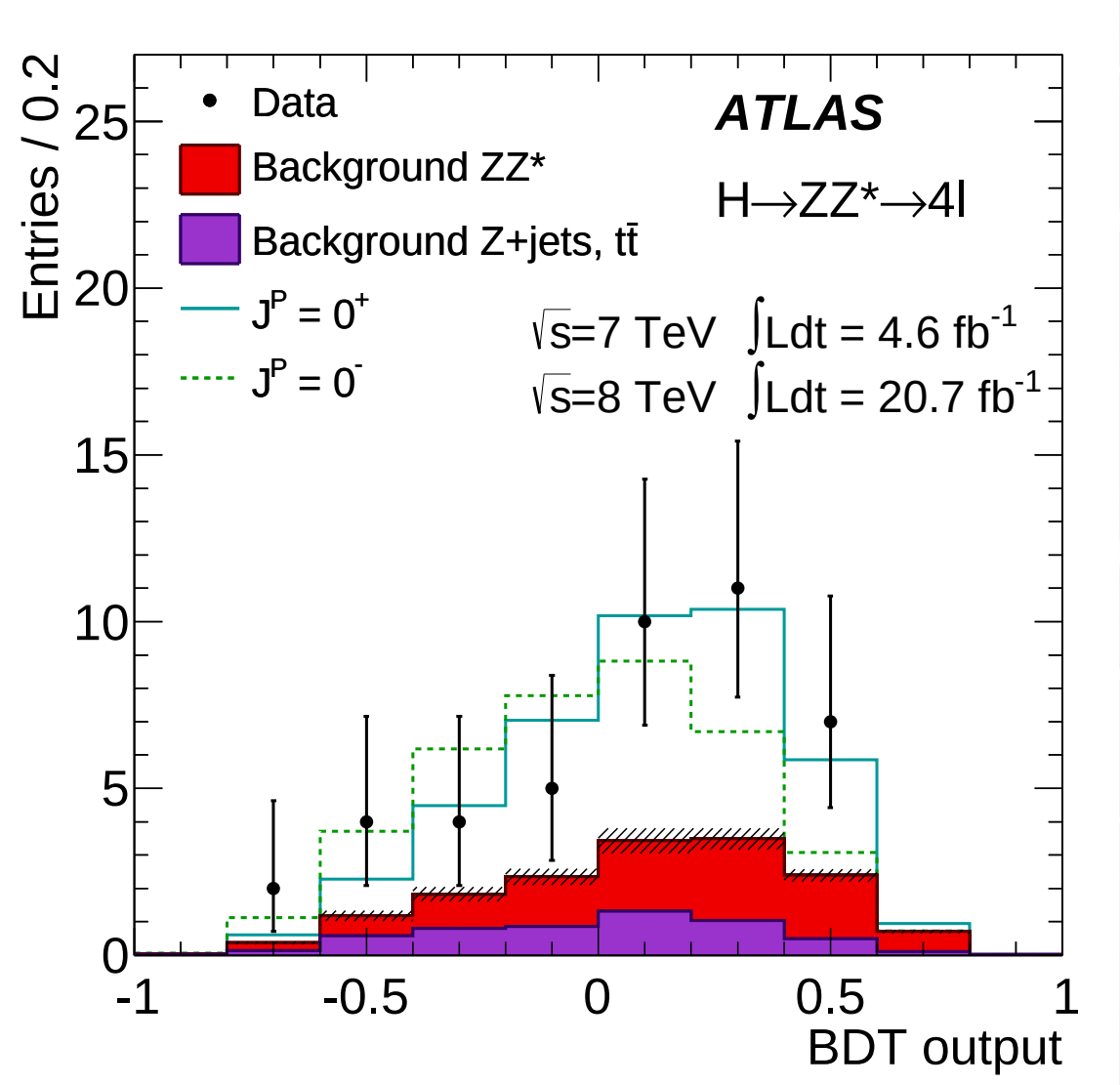
- Various control samples are used to measured contributions of reducible backgrounds (Z**+jets and** $t\bar{t}$), depending on the flavour of the sub-leading pair.
- Irreducible background (**ZZ**), constraint by fit on the full $m_{4\ell}$ range.
  Cross checked by the single-resonant production peak.



Inverted d_0 requirement for one of the two subleading leptons

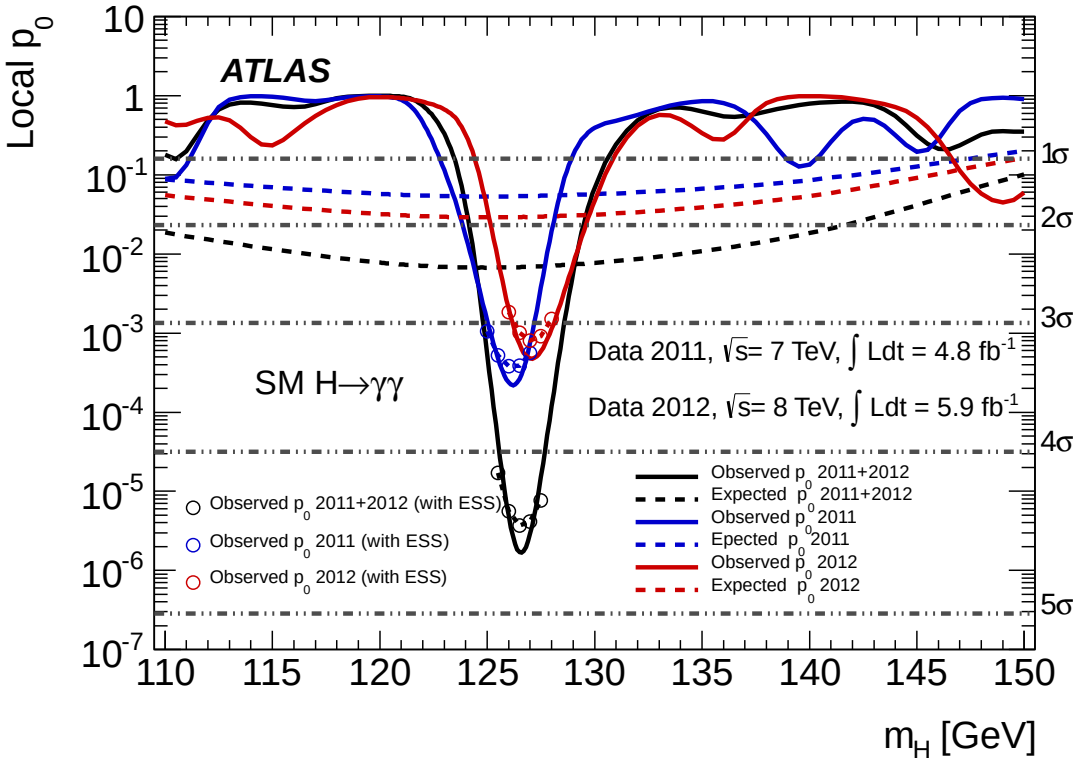
Relaxed kinematic requirements

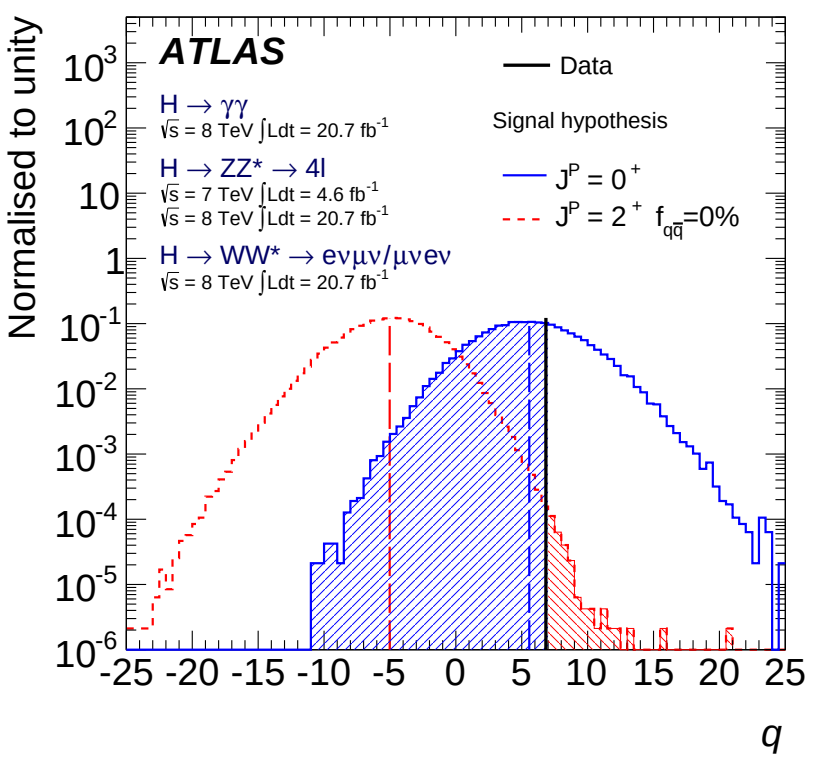
Examples: Multivariate discriminant



Example: Use a Boosted Decision Tree fed with kinematic variables (production and decay angles) to distinguish Spin and Parity of the observed events. (Train machine learning)

Statistics and Toy Monte Carlo





- In some cases we could use an asymptotic formula that save us from having to toss Toy Monte Carlo pseudo-experiments
- "Asymptotics" doesn't always work → many millions of fits needed(independent of each other)
- Usually no need of large external libraries.
- Used GPUs in some cases (or even HPC Blue-Gene @ ANL)

Summary

- The excellent performance of ATLAS has allowed to develop a very large variety of analyses, pushing the frontier of understanding
- Our computing model efficiently allows users to analyse several PB of data using hundreds of thousands of CPUs.
- A glimpse (example) of the data analysis details behind a Higgs boson search/measurement was introduced. HEP data analyses $\rightarrow$ wide spectra of tasks
  Even inside a particular analysis, many different techniques are used each with its own set of challenges and infrastructure requirements.


THANK YOU ...

BACKUP SLIDES ...

ATLAS Data-taking efficiency

ATLAS 2011 p–p run												
Inner Tracking			Calorimeters				Muon Detectors				Magnets	
Pixel	SCT	TRT	LAr EM	LAr HAD	LAr FWD	Tile	MDT	RPC	CSC	TGC	Solenoid	Toroid
99.8	99.6	99.2	97.5	99.2	99.5	99.2	99.4	98.8	99.4	99.1	99.8	99.3

Luminosity weighted relative detector uptime and good quality data delivery during 2011 stable beams in pp collisions at $\sqrt{s}$=7 TeV between March 13th and October 30th (in %), after the summer 2011 reprocessing campaign

ATLAS p-p run: April-December 2012										
Inner Tracker			Calorimeters		Muon Spectrometer				Magnets	
Pixel	SCT	TRT	LAr	Tile	MDT	RPC	CSC	TGC	Solenoid	Toroid
99.9	99.1	99.8	99.1	99.6	99.6	99.8	100.	99.6	99.8	99.5
All good for physics: 95.5%										

Luminosity weighted relative detector uptime and good quality data delivery during 2012 stable beams in pp collisions at $\sqrt{s}$=8 TeV between April 4th and December 6th (in %) – corresponding to 21.3 fb^{-1} of recorded data.