ATLAS Grid Computing - Introduction

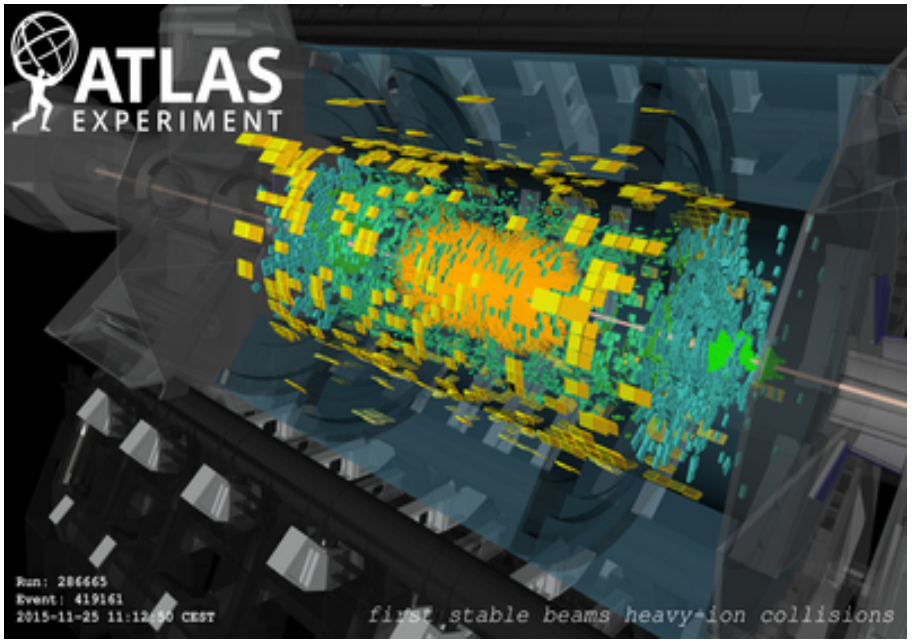
Practical Course on Parallel Computing – Sose18, 04.2018

Gen Kawamura, II. Physics Institute

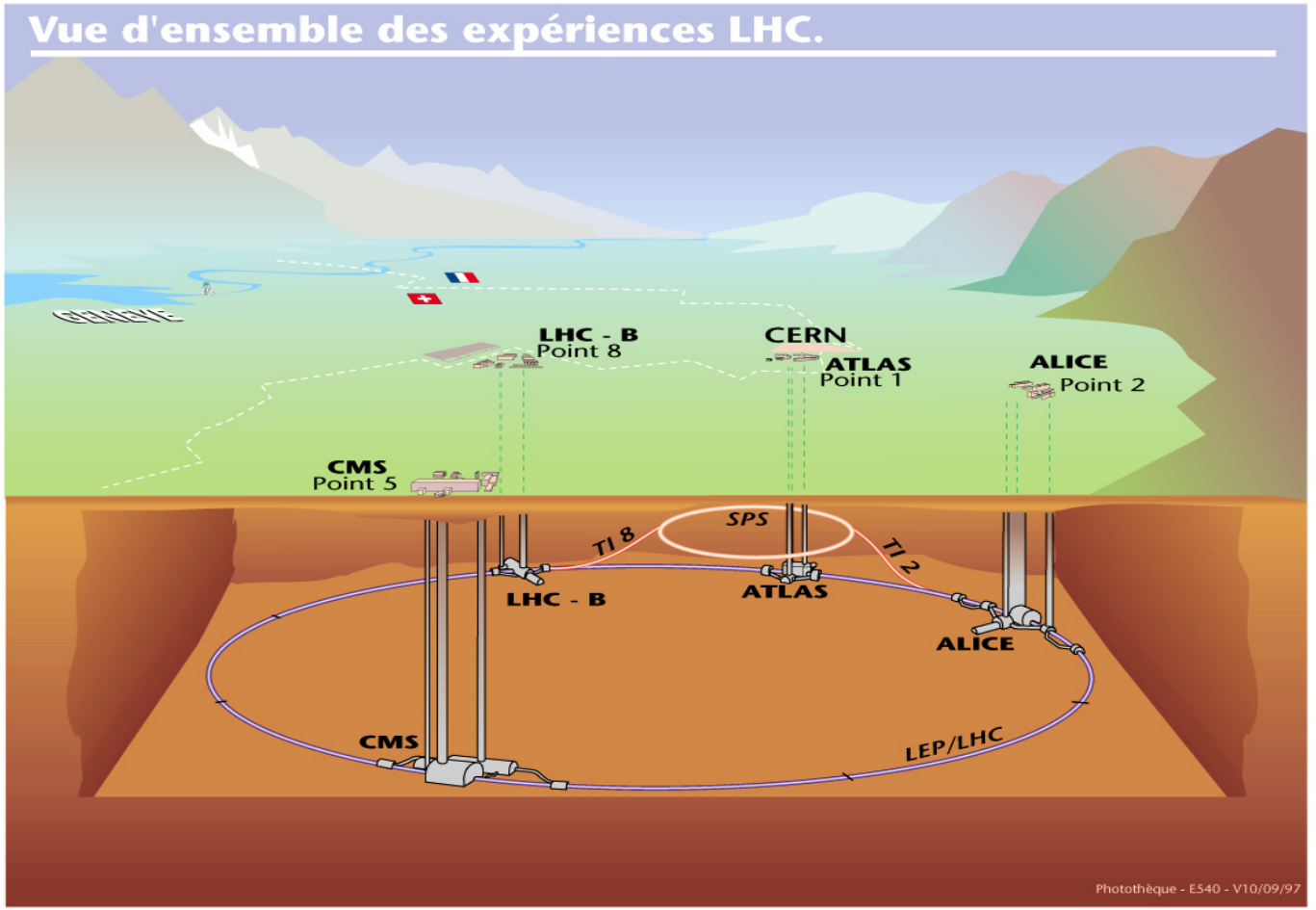
Overview of Introduction

- Introduction to ATLAS data analysis model
  - What are the LHC experiments?
  - The ATLAS detector
  - Basic Physics
  - ATLAS Grid computing and WLCG resources
  - Data flow in ATLAS
  - Basis of data analysis in high energy physics
  - Event Data Model (EDM) in ATLAS
  - Basis of data analysis
  - Basic tools: ATLAS CLI and CVMFS
- ATLAS Metadata Interface (AMI)
- Rucio (ATLAS Data Management System)
- Introduction to World LHC Computing Grid (WLCG)
- PanDA (ATLAS Job Management System)

Introduction to ATLAS data analysis model



What are the LHC experiments? - 1



The biggest hadron collier on this planet

- Hadron-Hadron collisions in TeV scale (14TeV in 2018)
- 100m below ground
- 27km circular ring
- The world largest distributed computing platform across 36 countries
- 4 main experiments (ATLAS, CMS, Alice and LHCb)
- **Already not super easy!**

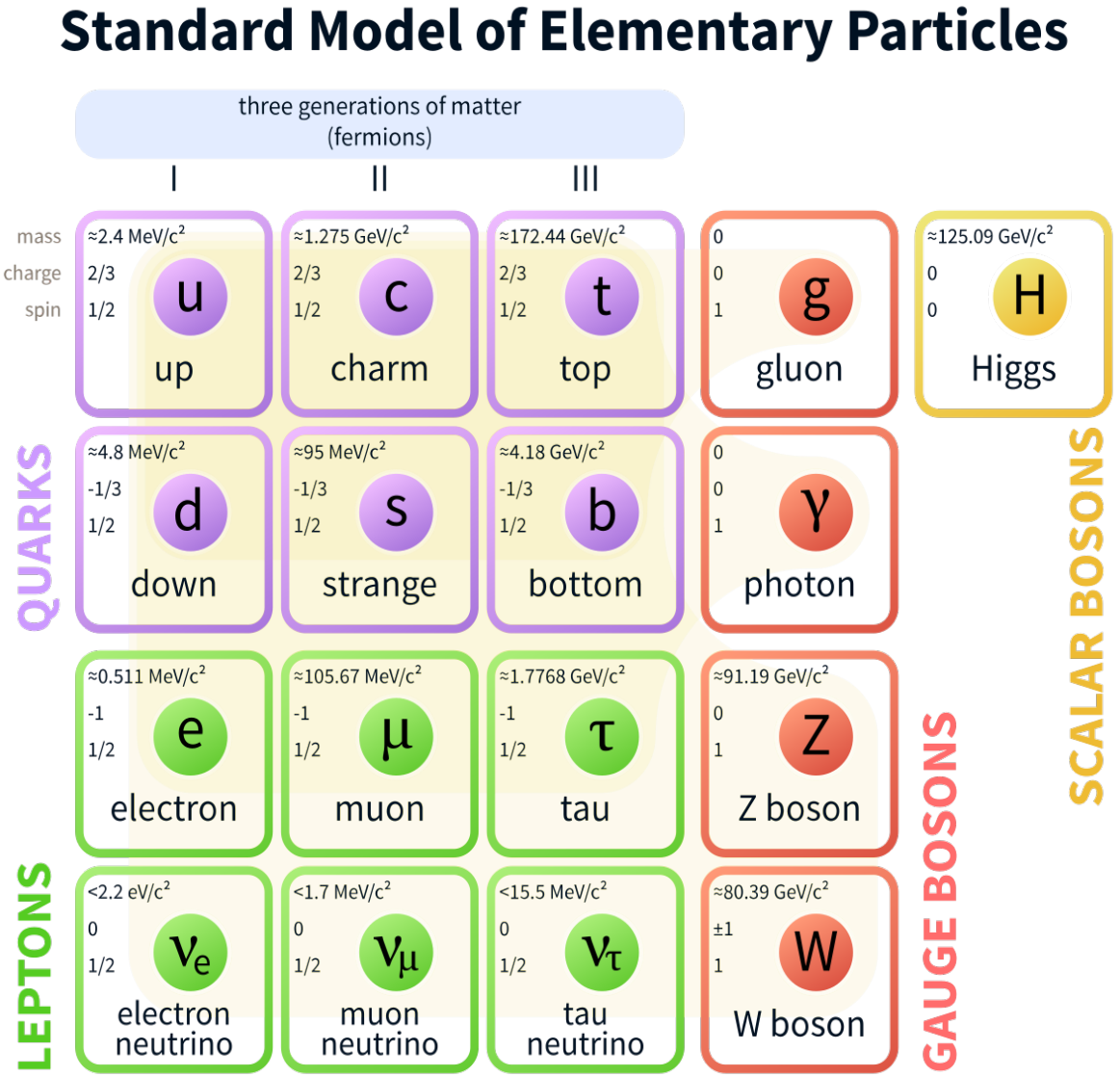
Expectations

- New physics
- Understanding of early universe and fundamental laws
  - Validation of many theoretical models

What are the LHC experiments? - 2

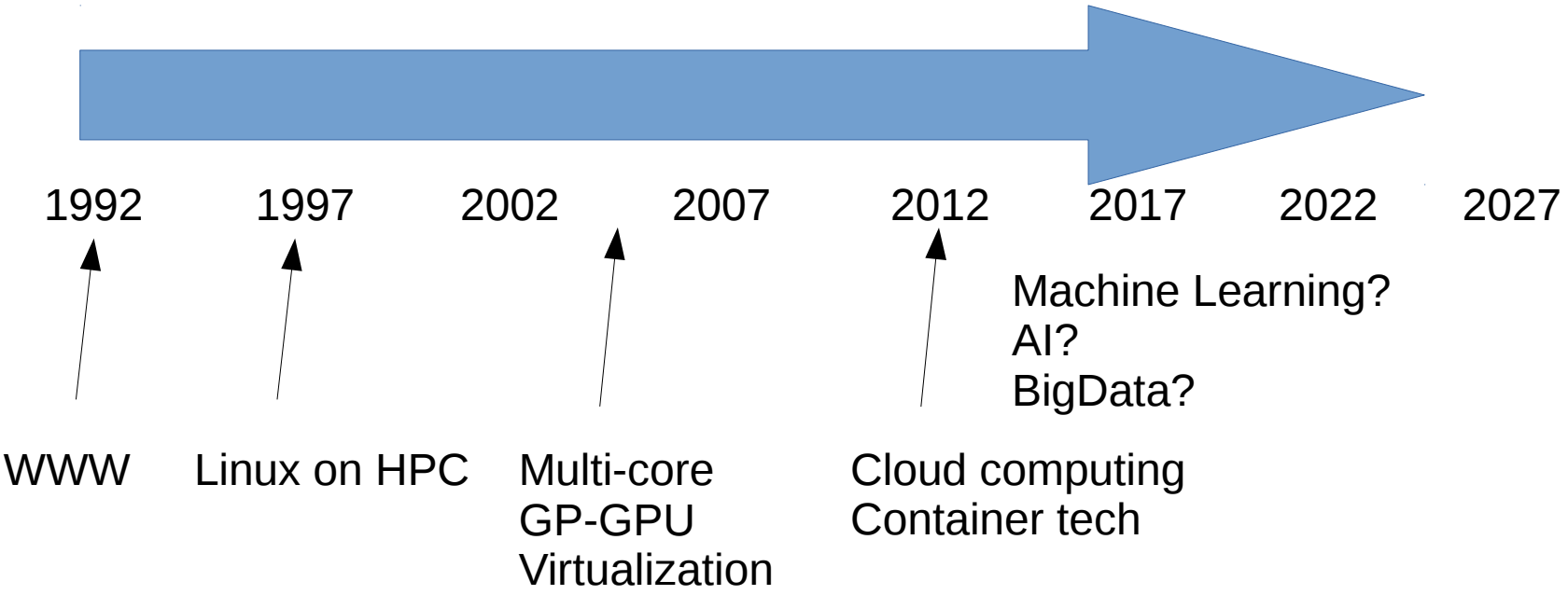
- New discovery was announced by ATLAS and CMS in 2012
  - *Higgs boson* (was expected by Standard Model)
  - Nobel prize 2013

What are the LHC experiments? - 3



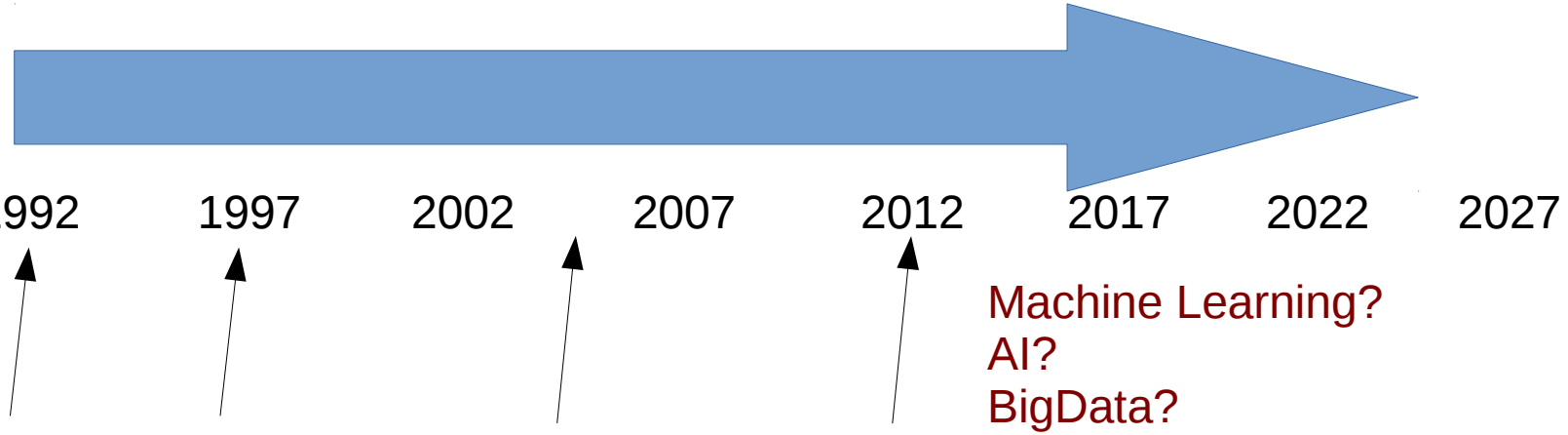
What are the LHC experiments? - 4

- Has continued since 1992



What are the LHC experiments? - 4

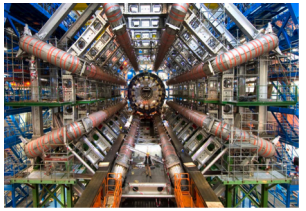
- Has continued since 1992



TDR (1994)

ATLAS

gLite (β)



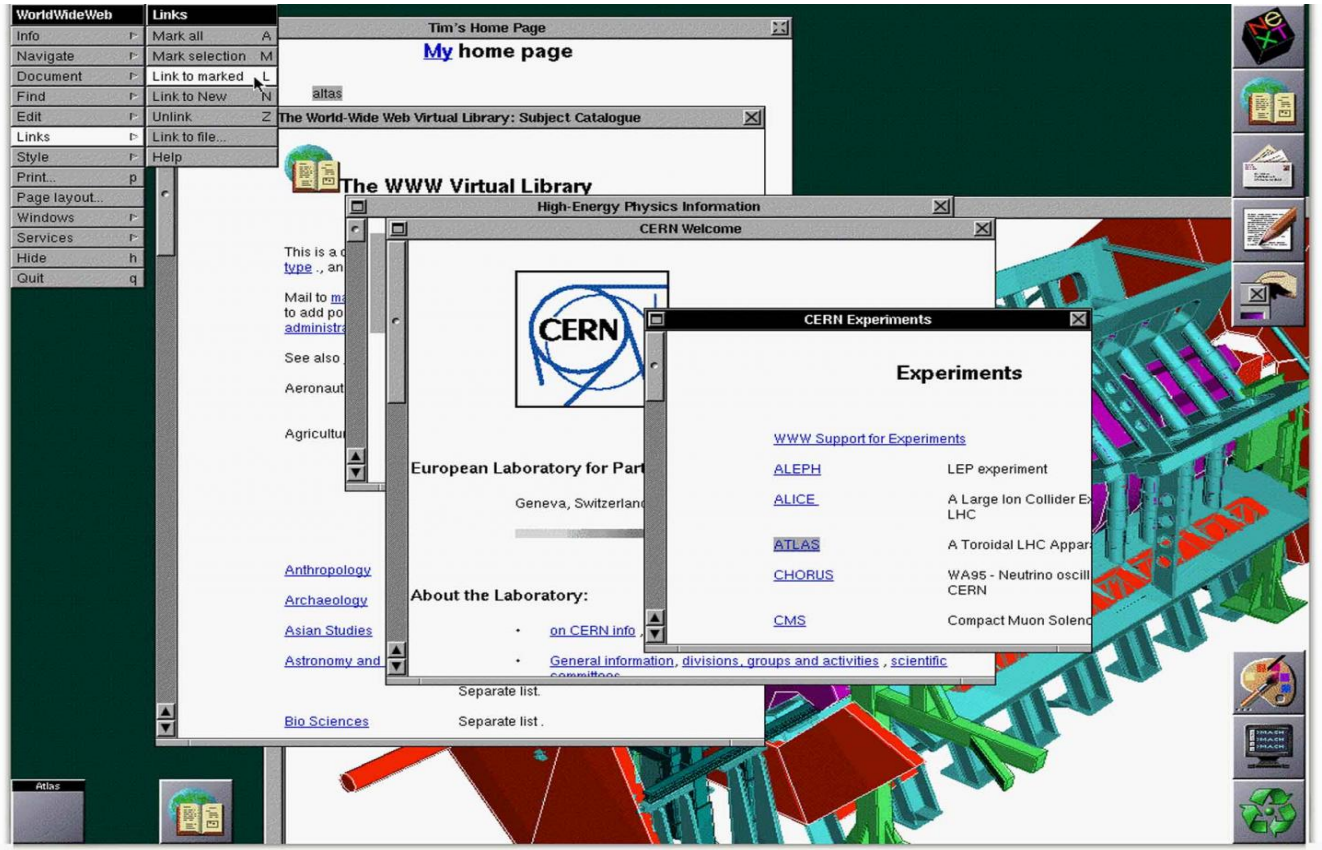
Run-1 Run-2 Run-3

HL-LHC

Discovery of Higgs Particle

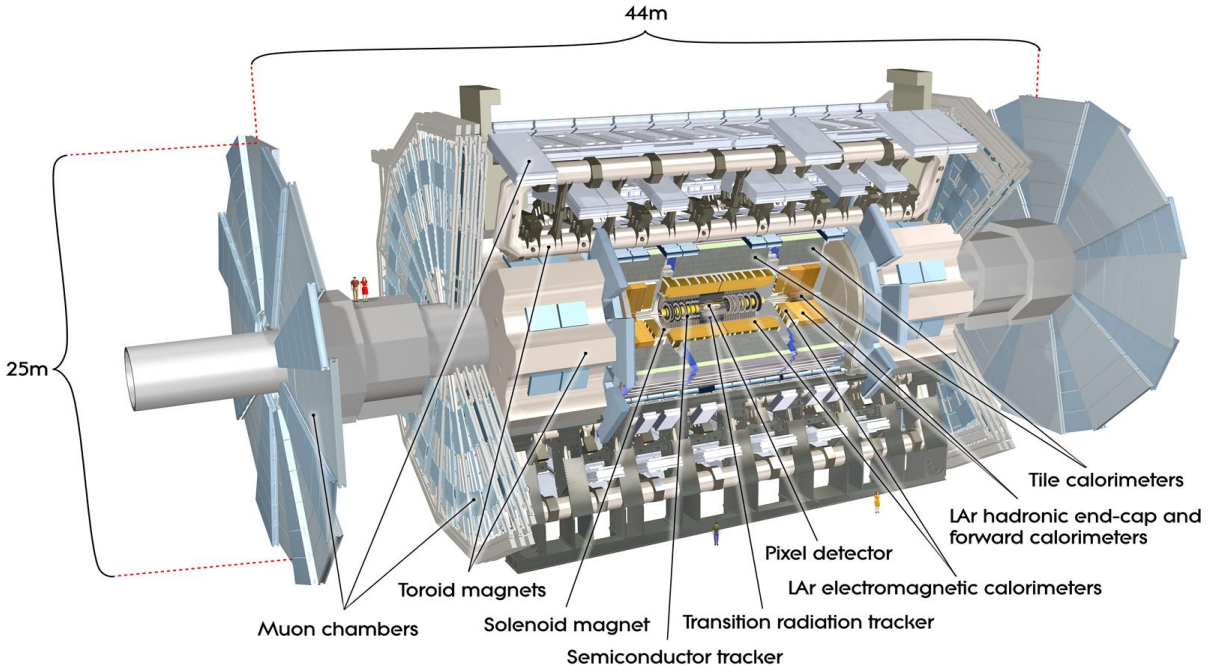
What are the LHC experiments? - 4

- Has continued since 1992



Screenshot of the original NeXT web browser in 1993 (Image: Berners-Lee/CERN)

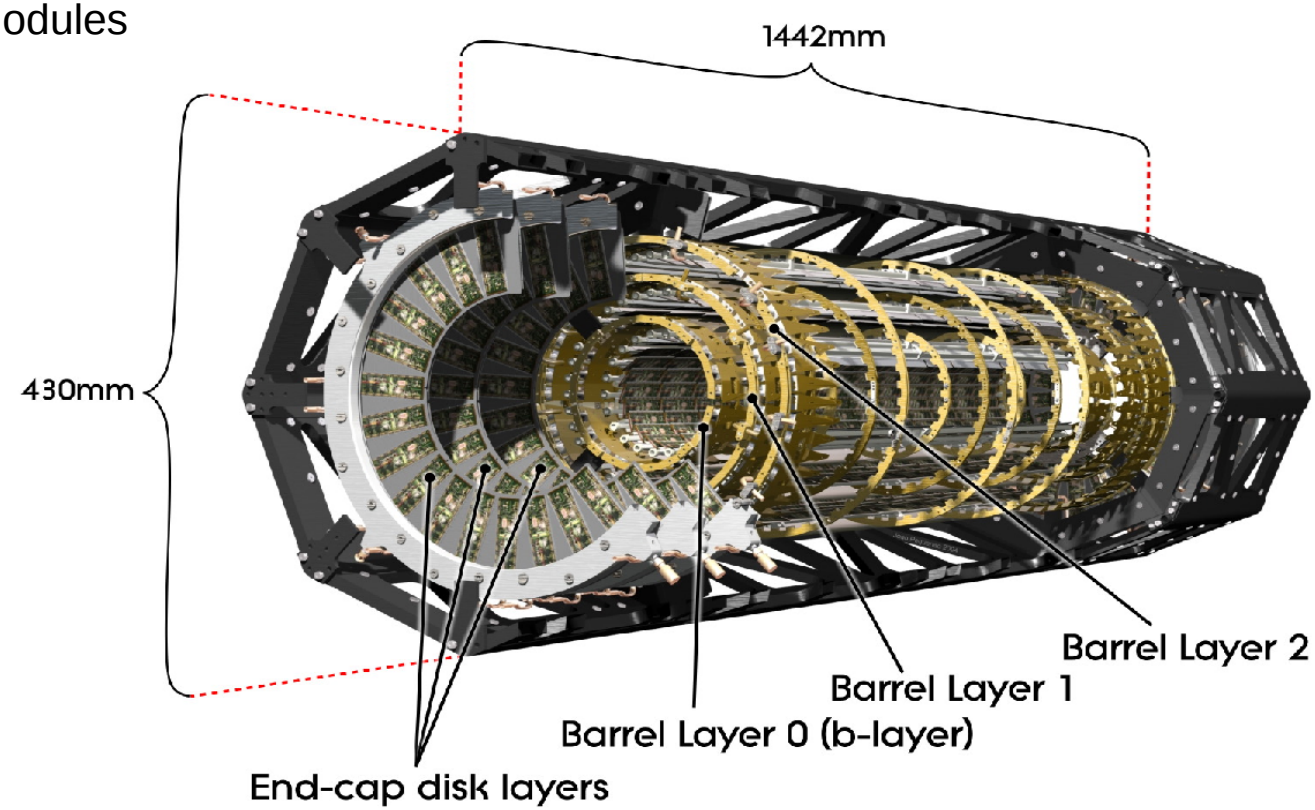
The ATLAS detector



- General-purpose HEP experiment
  - Reconstruction of outgoing particles from proton-proton collisions at 13TeV
  - pp-rate: 40MHz
  - Length: 46m, Height 25m, Weight 7kt, B-field 4T

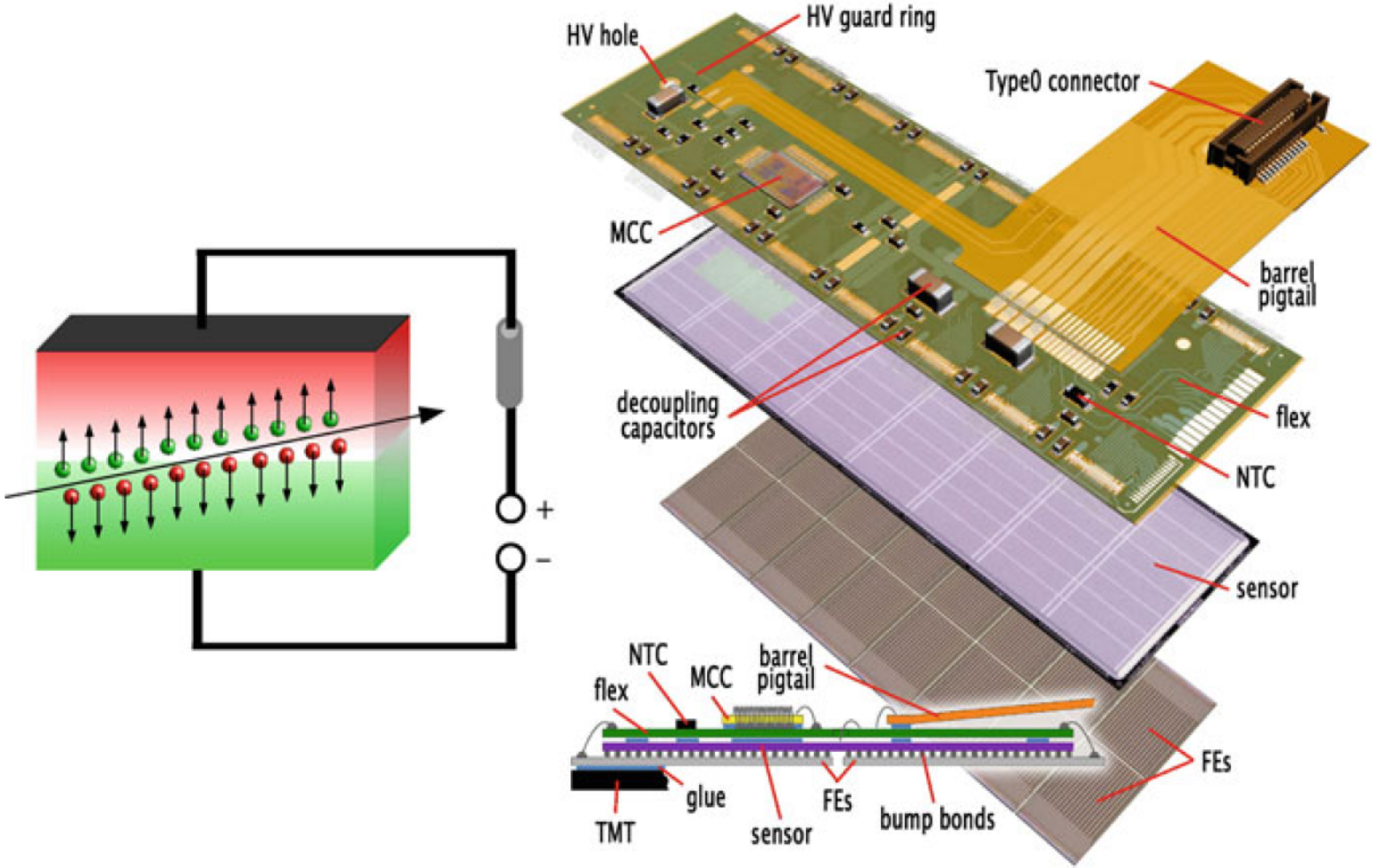
Inner Pixel Detector - 1

- 80 million pixels (80 million channels). Area 1.7m^2. 15 kW power consumption
- Barrel has 1,744 modules (10cm^2) with 46,080 readout channels per module
- Pixel size 50 x 400μm^2 . Resolution 14 x 115μm^2
- Three Pixel disks (in each endcap) have 6.6 million channels
- 3 barrel layers: 1,456 modules
- 3 disks in each end-cap: 288 modules

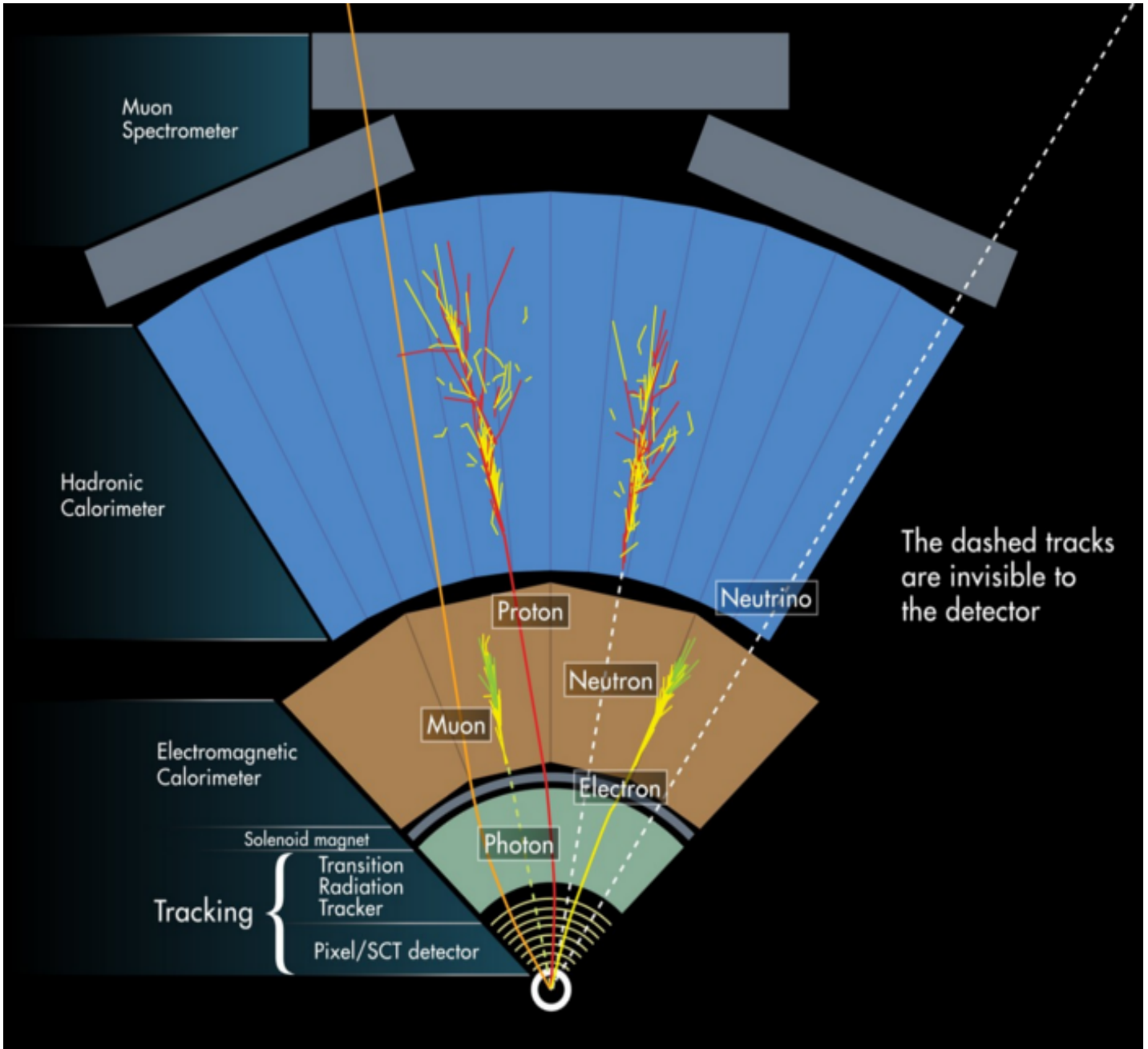


Inner pixel detector - 2

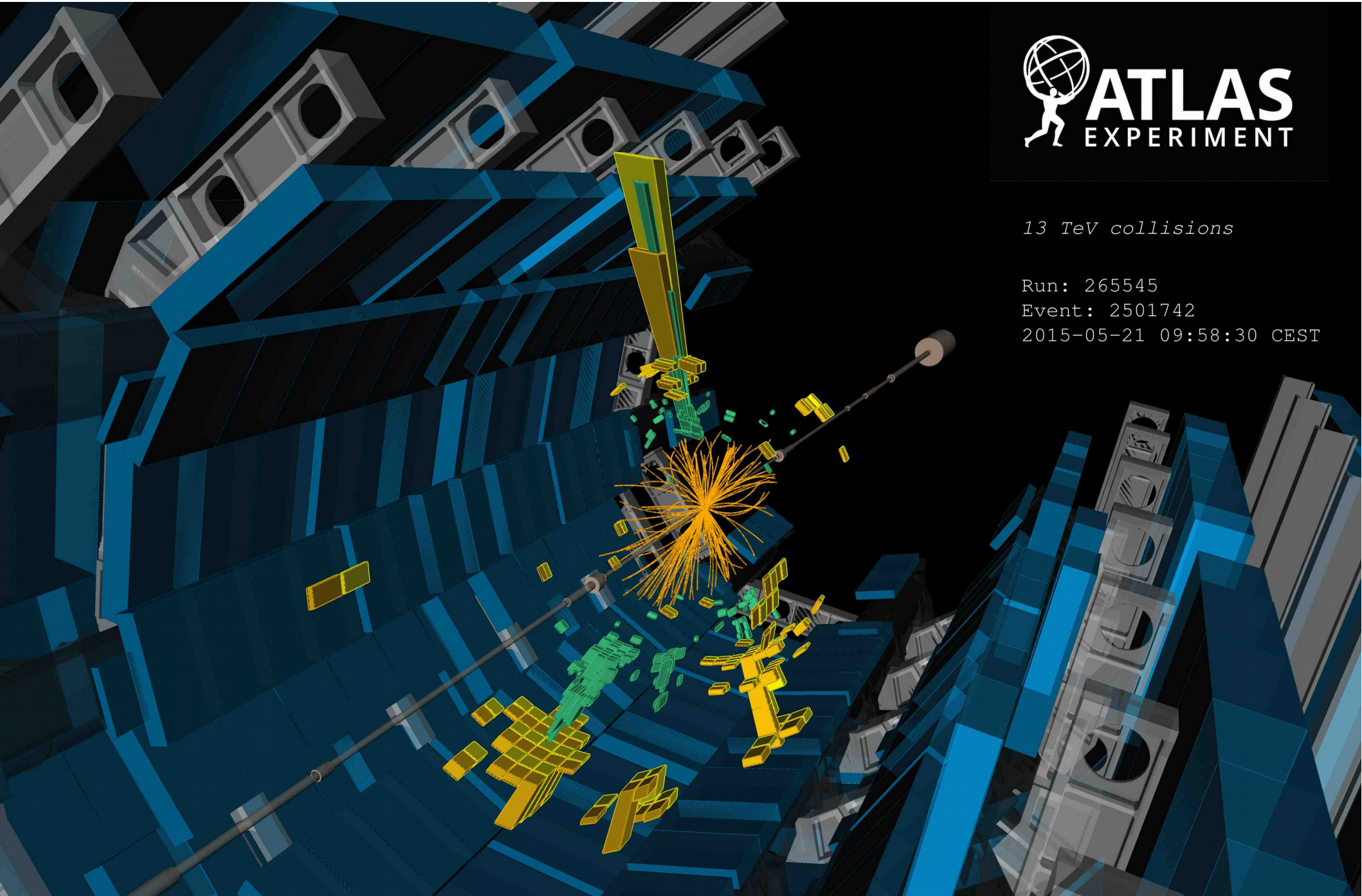
- Works like (very complicated) CCD camera



Silicon Tracker & Trigger - 2

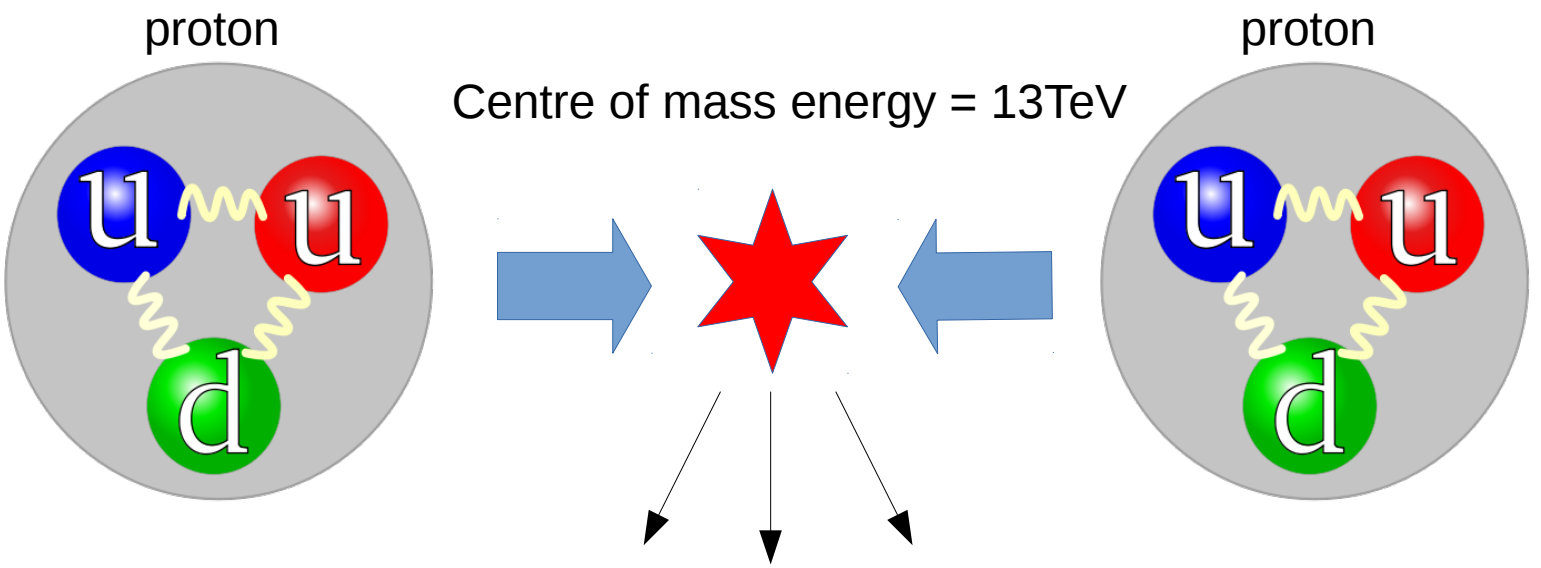


One event


ATLAS
EXPERIMENT
13 TeV collisions
Run: 265545
Event: 2501742
2015-05-21 09:58:30 CEST

Basic Physics - 1

- Proton-Proton (Hadron-Hadron) Collisions
  - Proton = 2 Up quarks, 1 Down quark bound by Gluons (strong force)



Productions of elementary particles

even unknown particles and unknown mechanisms (expectation!)

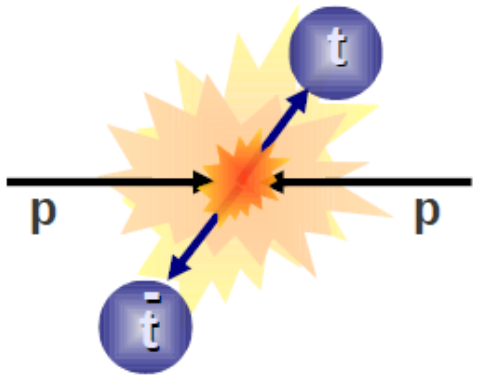
Basic Physics - 2

- Production of Teilchenzoo (particle zoo)

Basic Physics - 3

- For example, Top-Quark production

Production:

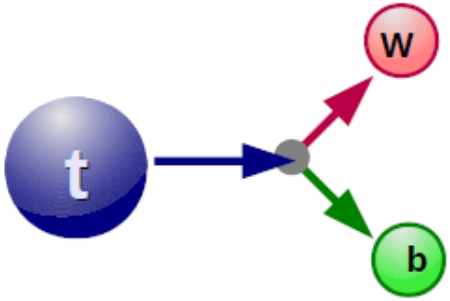


- Production rate of different processes *(strong and EW)*
- Differential distributions
- New production mechanisms

Intrinsic properties:

- Top quark mass → constraints on Higgs mass
- Charge
- Lifetime

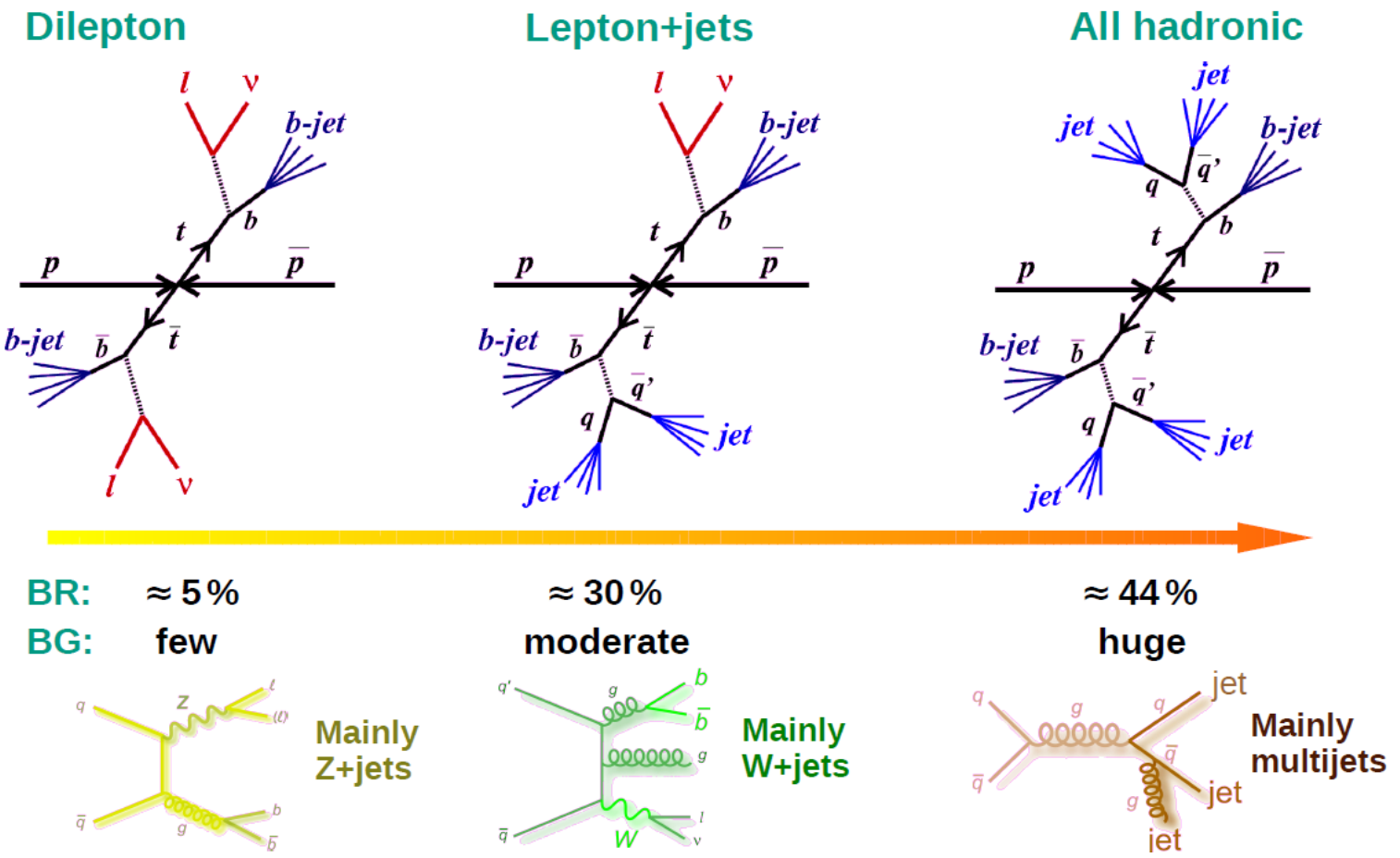
Decay:



- Decay channels *(SM and new)*
- Couplings

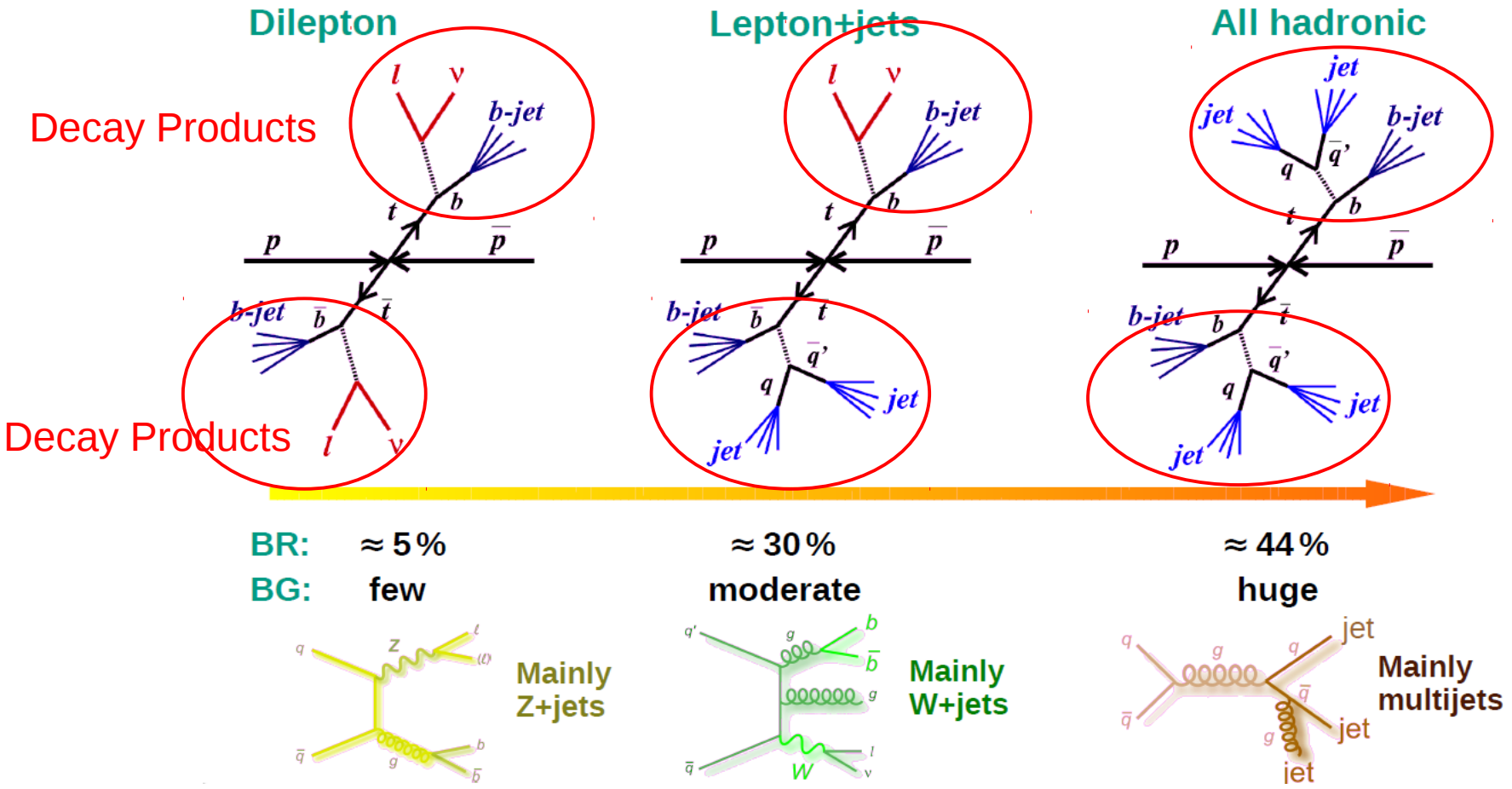
Basic Physics - 4

- For example, Top-Quark production
  - Event signatures



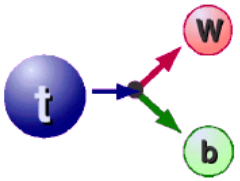
Basic Physics - 5

- For example, Top-Quark production
  - Event signatures



Basic Physics - 6

- For example, Top-Quark production
  - Decay products



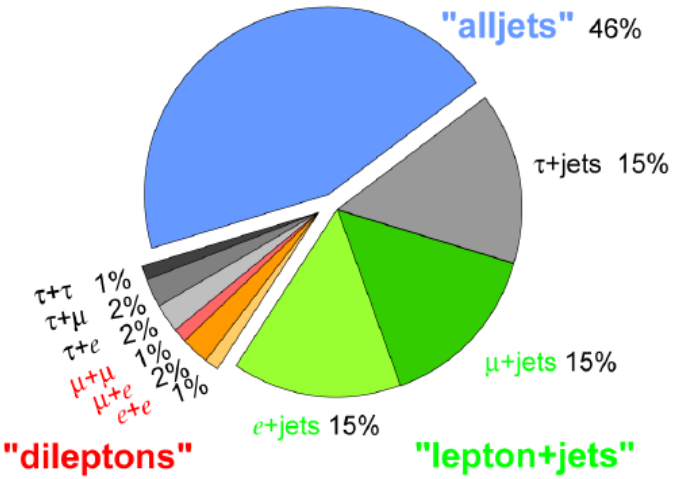
SM:

t → b W ≈ 100 %

Decay of top quark pairs

W decay	e^+	μ^+	τ^+	$u\bar{d}$	$c\bar{s}$
e^-	ee	$e\mu$	$e\tau$	electron+jets	
μ^-	$e\mu$	$\mu\mu$	$\mu\tau$	muon+jets	
τ^-	$e\tau$	$\mu\tau$	$\tau\tau$	tau+jets	
$\bar{u}d$	electron+jets	muon+jets	tau+jets	all-hadronic	
$\bar{c}s$					

dileptons

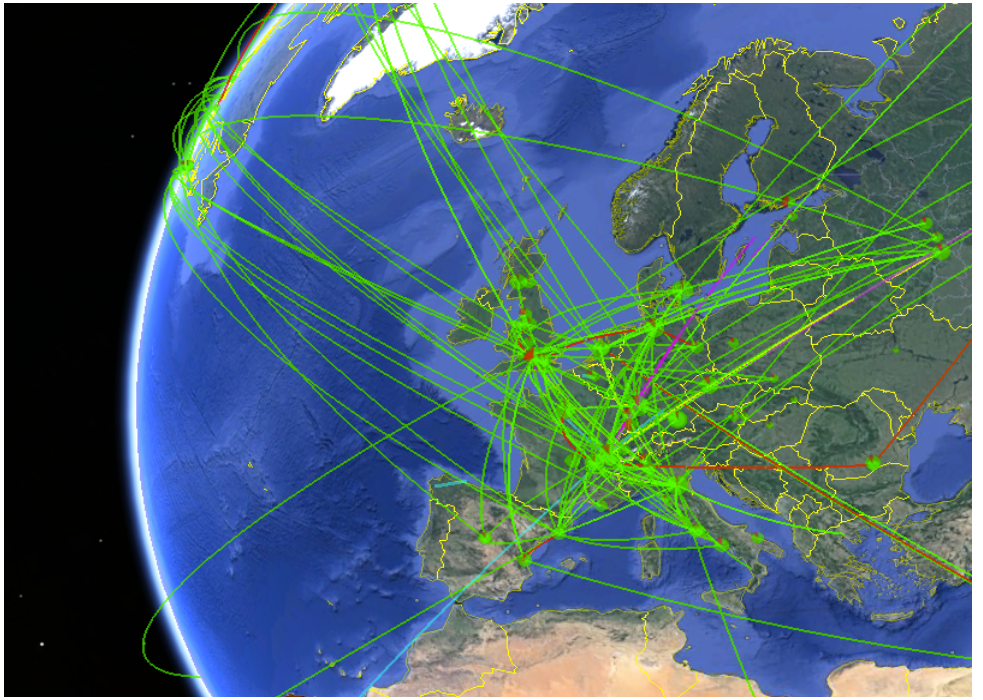


"lepton+jets"-channel:
(e+jets and μ+jets)

- Moderate background
- Relative high branching ratio (~30%)

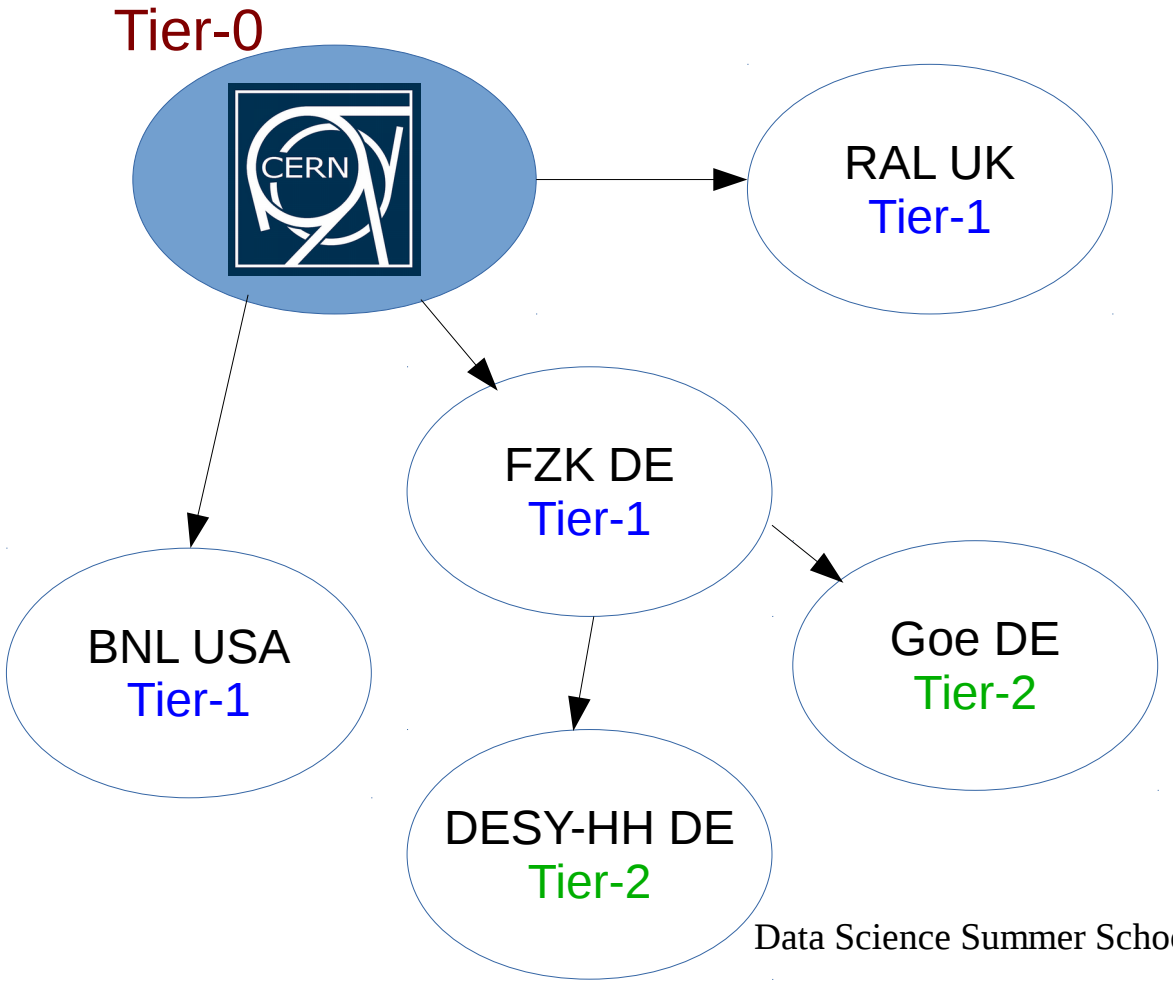
ATLAS Grid computing and WLCG resources - 1

- World LHC Computing Grid (WLCG)
  - The most sophisticated data-taking & analysis system ever built for science, providing near real-time access to LHC data
  - 2 million computing jobs per day
  - 42 countries
  - 170 computing centres



ATLAS Grid computing and WLCG resources - 2

- LHC multi-tier structure
  - WLCG = Worldwide LHC Computing Grid

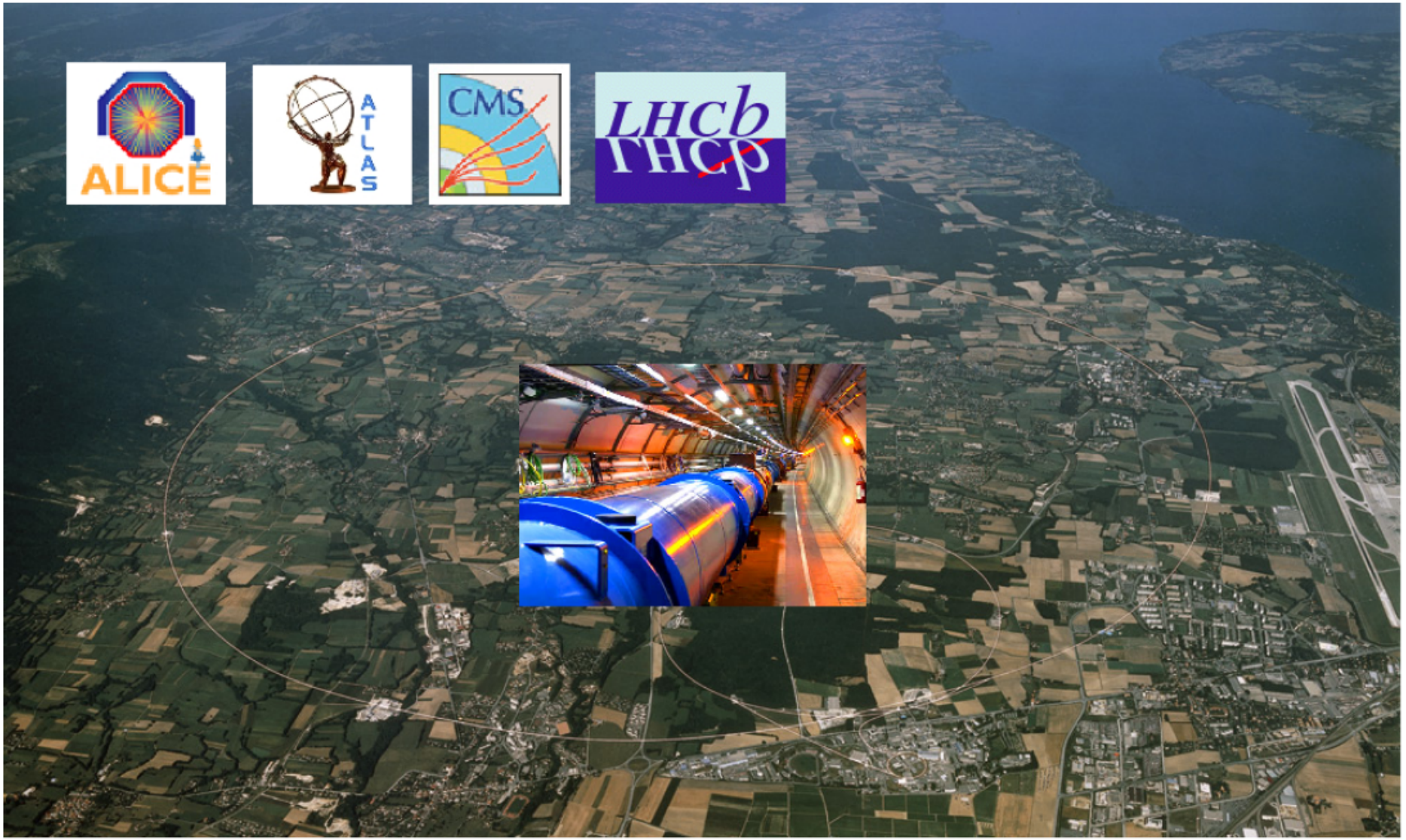


Tier-0: Raw data, Data store (in tape), Pre-processing, Reconstruction

Tier-1: National LCG-Centre, faster network connections and lager storage spaces (e.g. Tape), MC production, user analysis, etc.

Tier-2: University or Facility level computing sites. MC production, user analysis, etc.

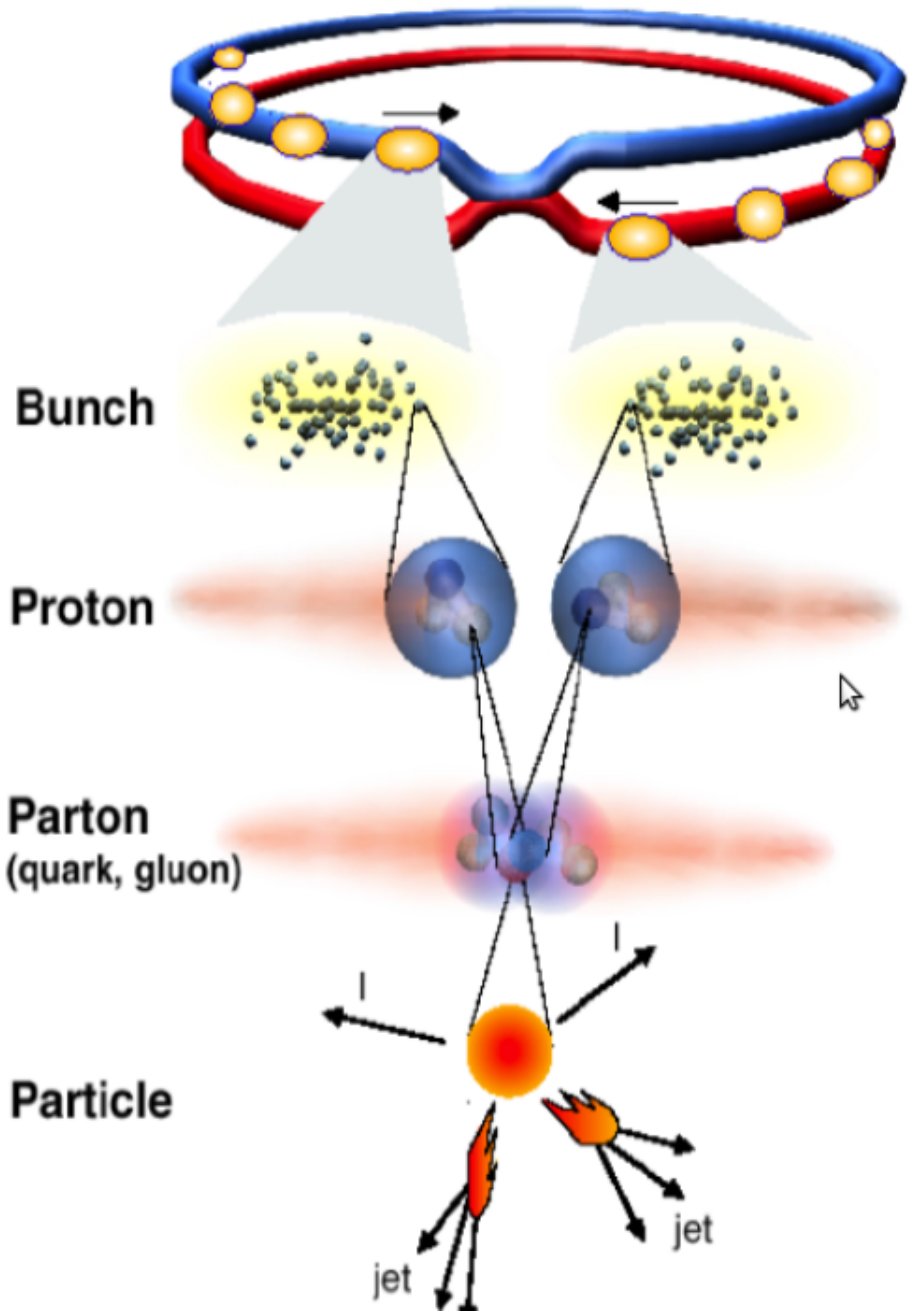
The LHC and Experiments



Detectors built and operated by a large team

Worldwide Collaboration of over 3000 physicists and engineers in ATLAS and CMS each + similar in LHCb and ALICE

Collisions at the LHC



Proton-Proton-Kollisionen
2835 Teilchenbündel (Bunch)

10^{11} Protonen / Bunch
Kollisionsrate 40 MHz (25 ns)

Schwerpunktsenergie 14 TeV
(= 7400 x Ruheenergie der
kollidierenden Teilchen)

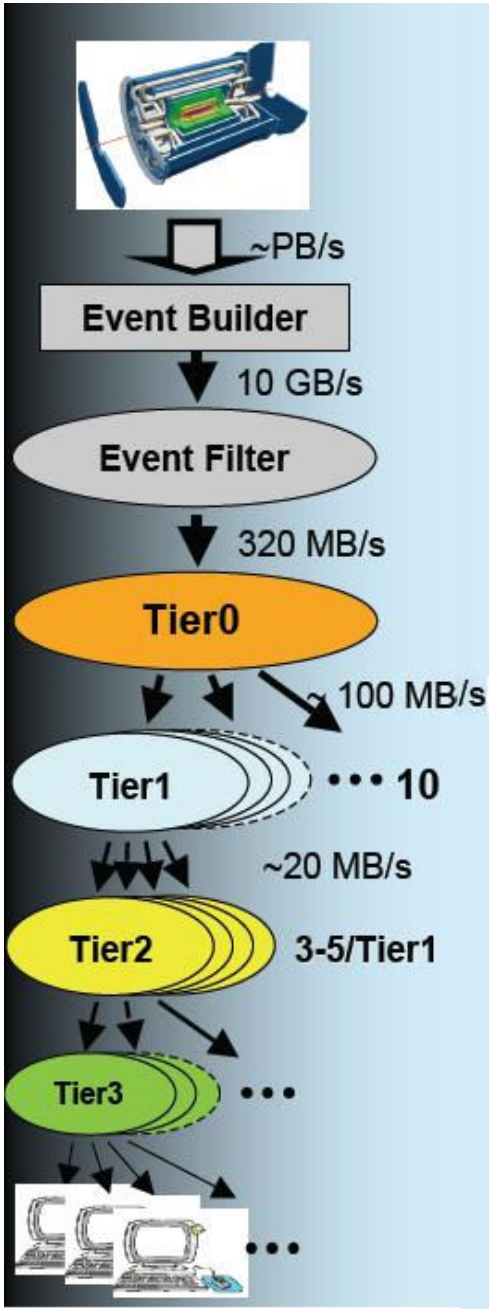
Schwerpunktsenergie der
kollidierenden Quarks und Gluonen
bis einige TeV

~25 pp-Kollisionen pro
Bunch-Kollision

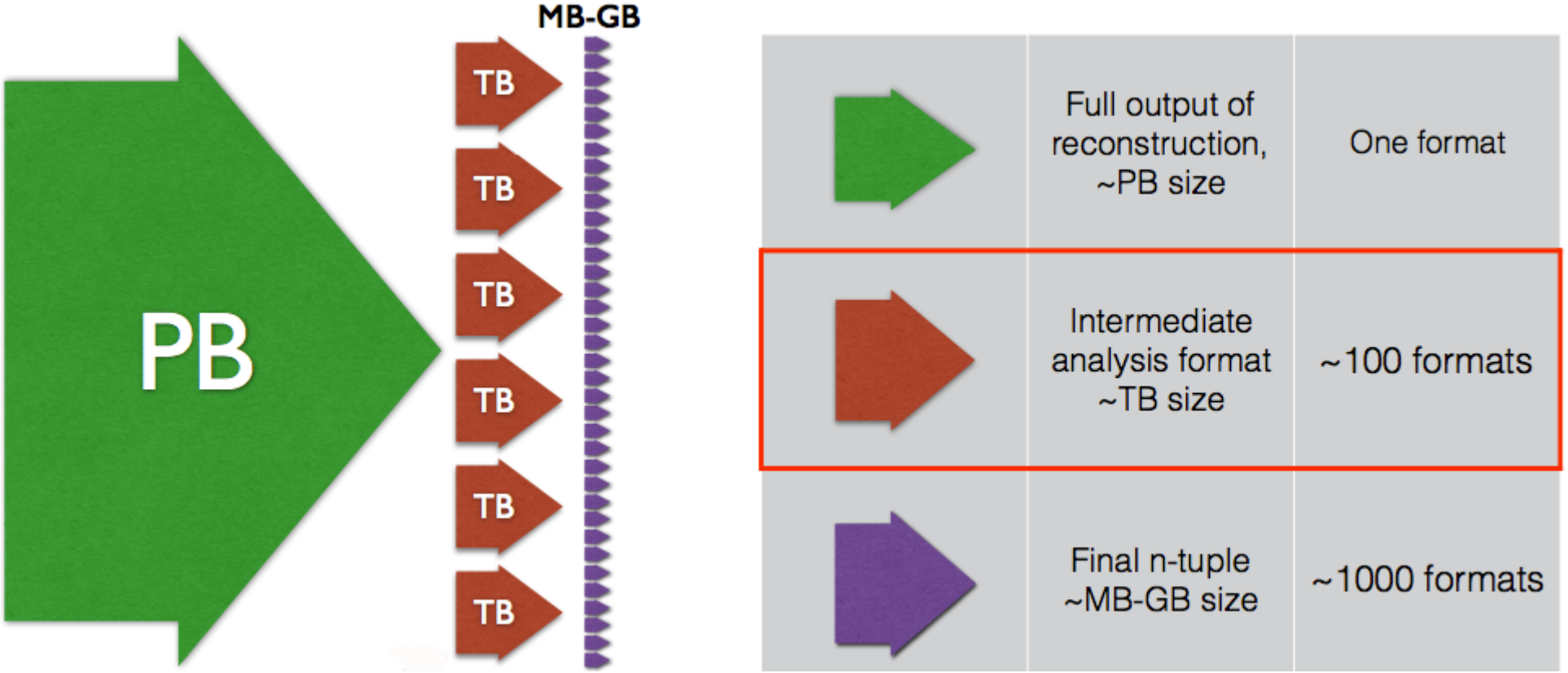
Interessante Ereignisse: 10^{-9} – 10^{-11}
unterdrückt!

Data flow in ATLAS

- RAW = about 2 MB/event
  - Original data at Tier-0 (ATLAS detector)
  - Complete replica distributed among all Tier-1s
- ESD (Event Summary Data) = 1 MB/event
  - ESDs produced by primary at Tier-1s
  - Globally 2 copies on disk
- AOD (Analysis Object Data) = 0.2 MB/event
  - Completely replicated at each Tier-1
  - Partially replicated to Tier-2s so as to have a complete set in the Tier-2s associated to each national Tier-1
  - Every Tier-2 specifies which datasets are most interesting for their reference community; the rest are distributed according to capacity
- TAG (Event Level Metadata Tags) = 1 kB/event
  - TAG files or databases are replicated to all Tier-1s (Root/Oracle)
  - Partial replicas of the TAG are distributed to Tier-2s as Root files
  - Each Tier-2 have all TAGs that correspond to the AODs storead



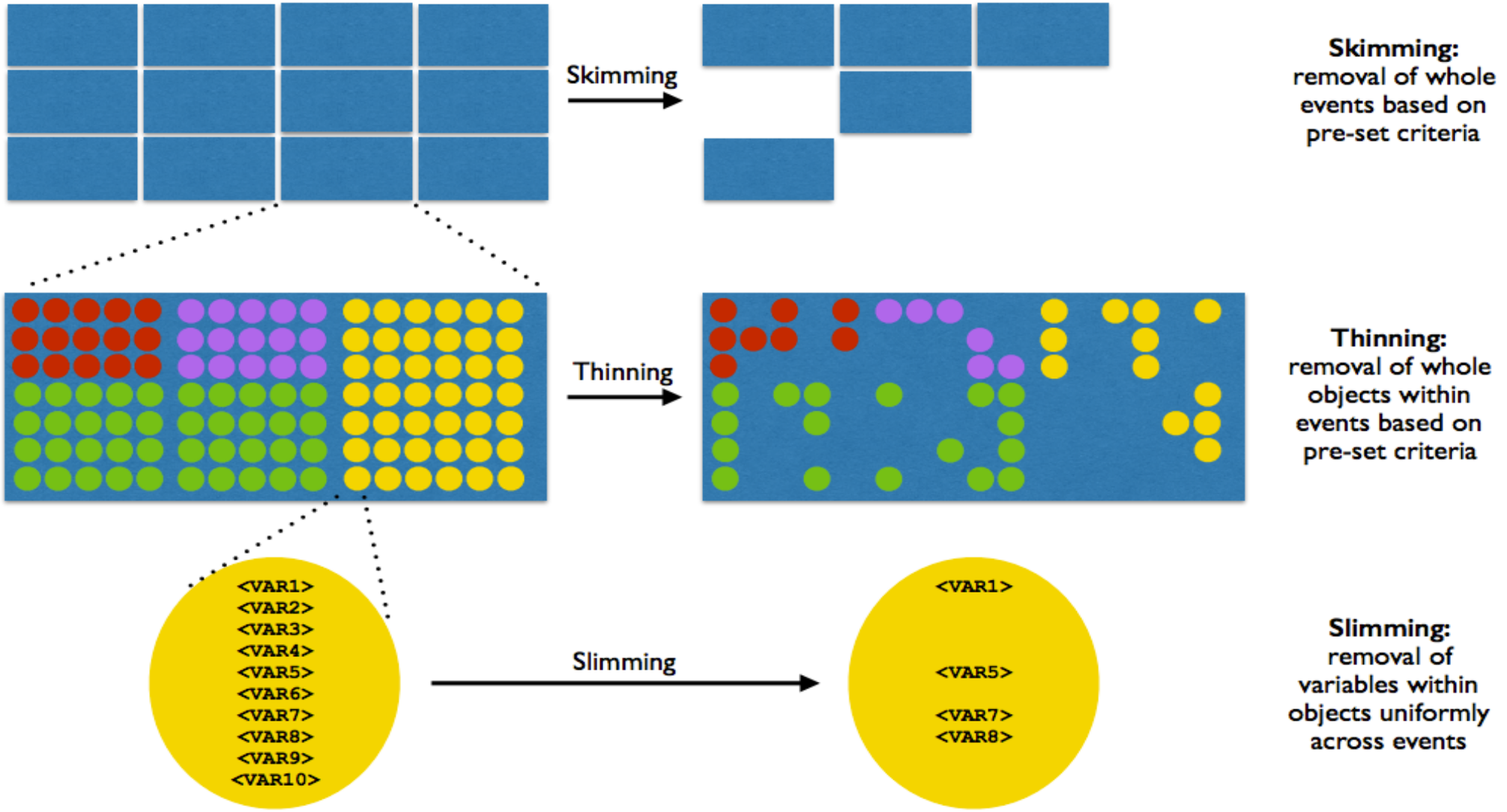
Event Data Model (EDM) in ATLAS



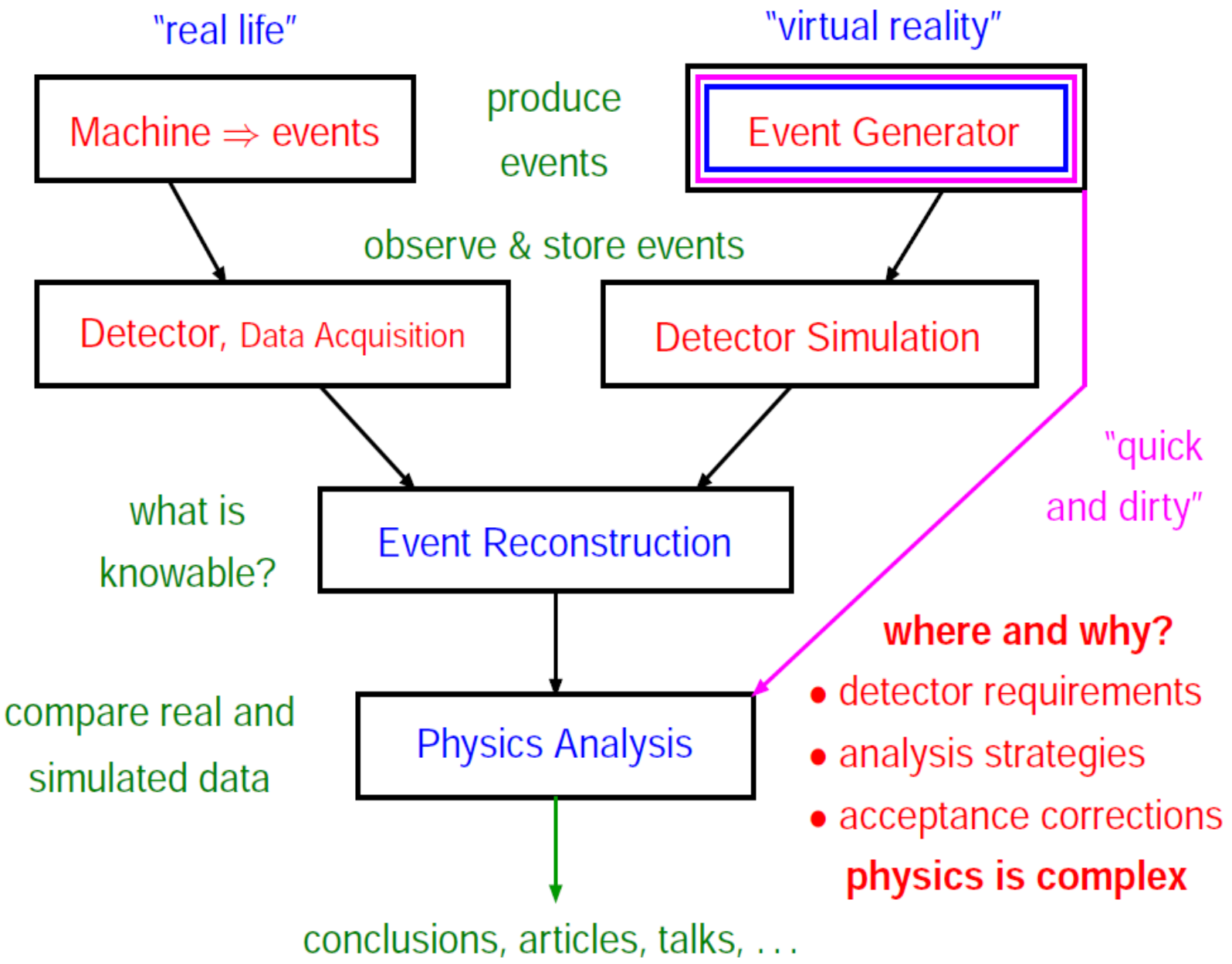
- AODs are still huge, so need to have manageable size of data → Derived AOD (DAOD)
- Derivation framework in Athena (ATLAS software framework)
- Each analysis group use own derivation framework

Event Data Model (EDM) in ATLAS

- Size reduction (= selection of events, particles and parameters)

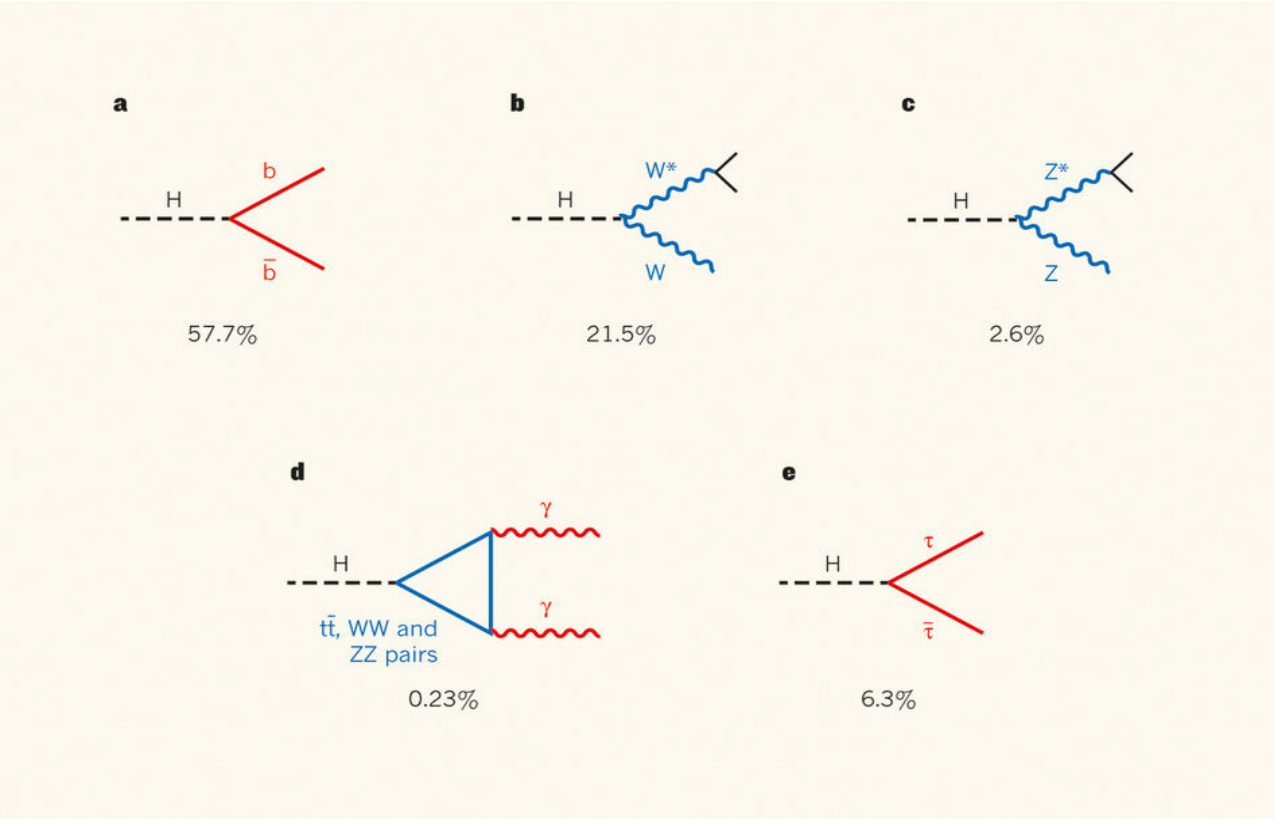


Basis of data analysis - 1



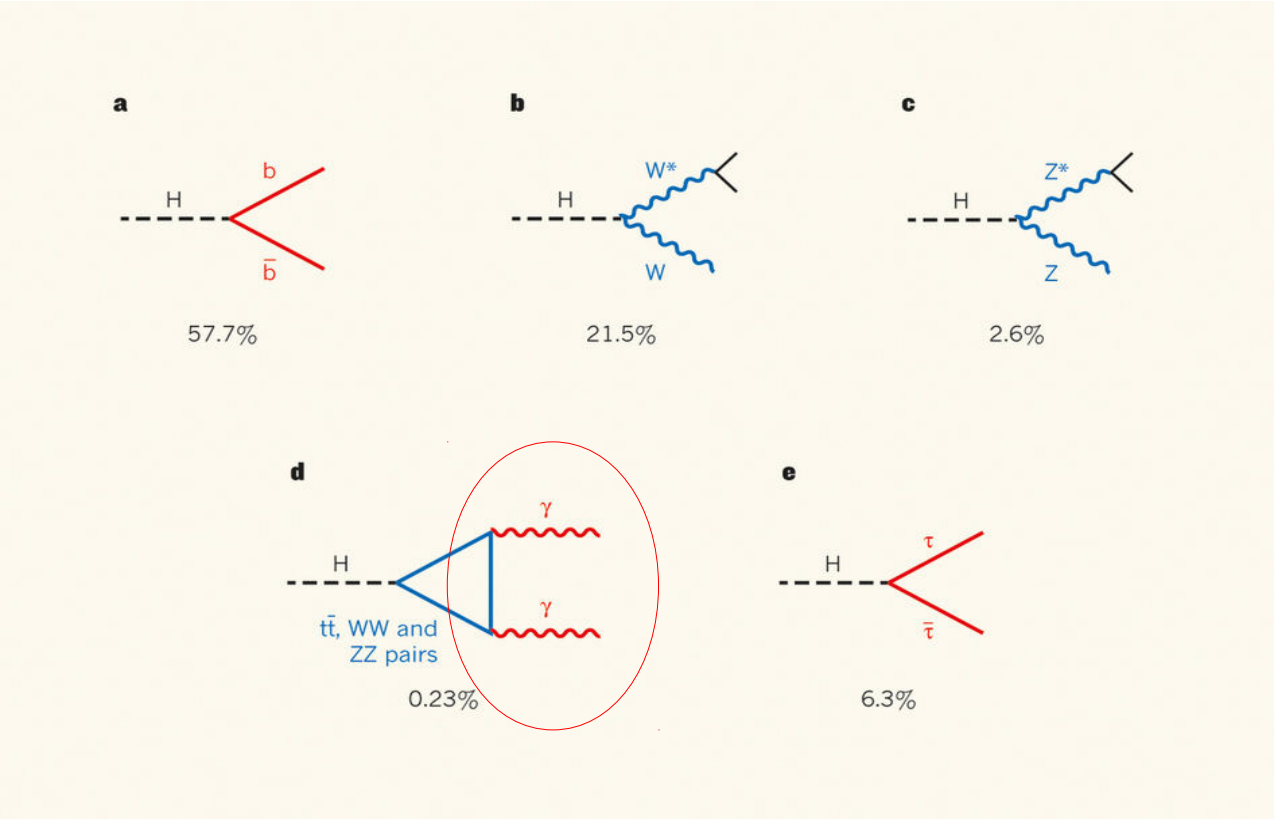
Basis of data analysis - 2

- Higgs decay channel
  - 1 EUR → 50 + 50 cent
  - 1 EUR → 50 + 20 + 20 + 10 cent
  - 1 EUR → 20 + 20 + 20 + 20 + 20 cent
  - 1 EUR → 50 + 20 + 10 + 10 + 10 cent



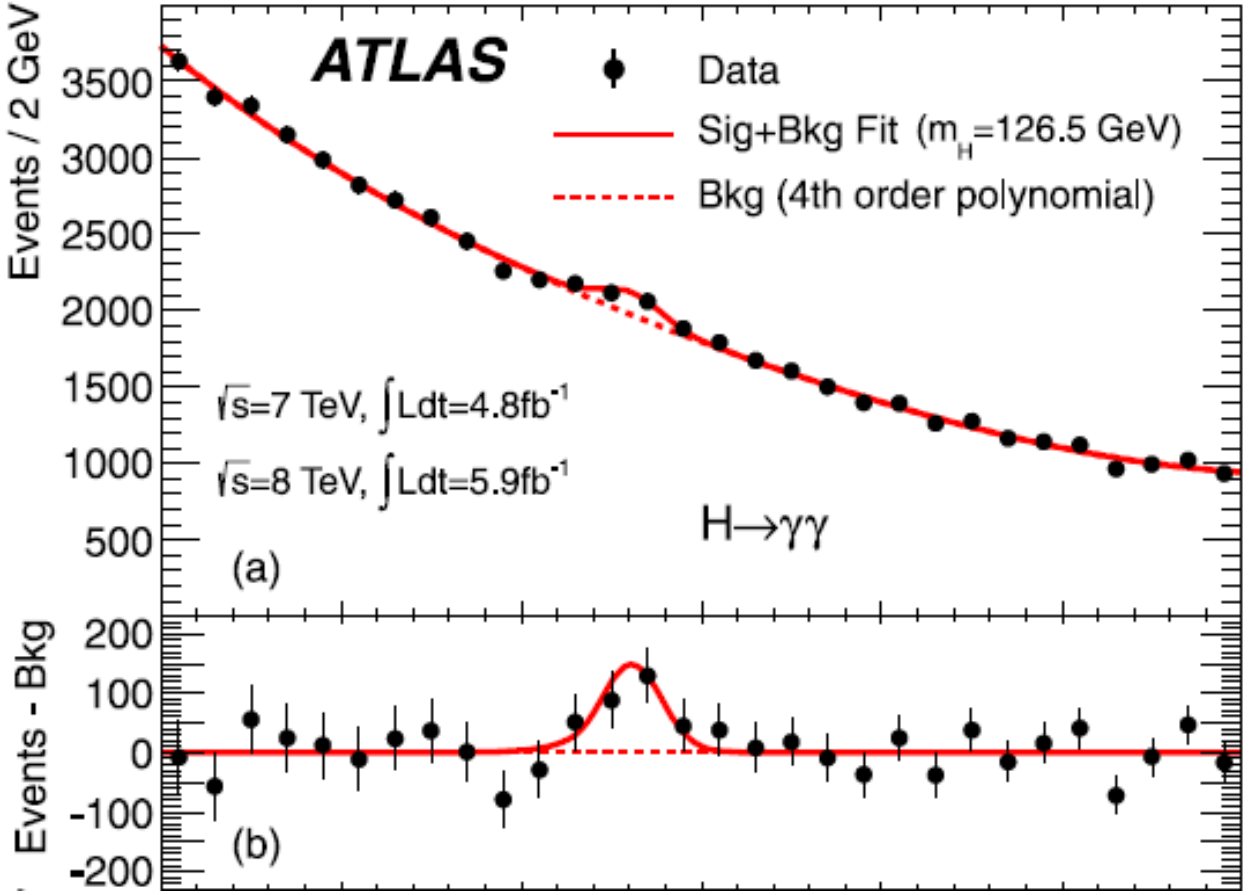
Basis of data analysis - 2

- Higgs decay channel
  - 1 EUR → 50 + 50 cent
  - 1 EUR → 50 + 20 + 20 + 10 cent
  - 1 EUR → 20 + 20 + 20 + 20 + 20 cent
  - 1 EUR → 50 + 20 + 10 + 10 + 10 cent



Basis of data analysis - 3

- Atlas Collaboration. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC." arXiv preprint arXiv:1207.7214 (2012).



Basis of data analysis - 4

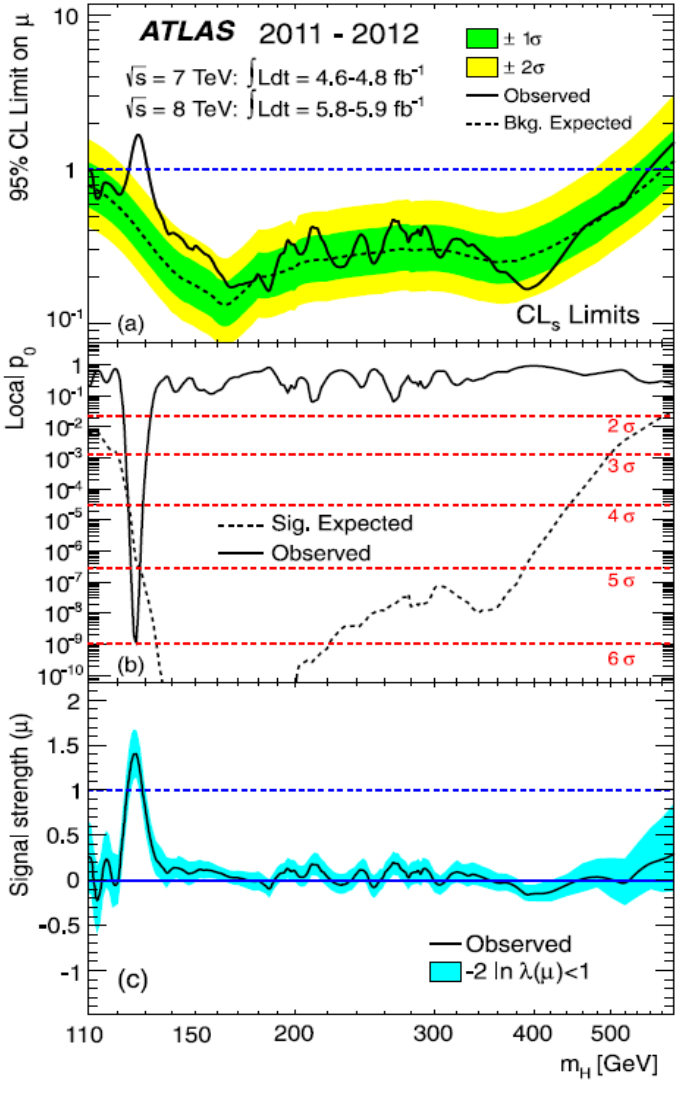


Fig. 7. Combined search results: (a) The observed (solid) 95% CL limits on the signal strength as a function of m_H and the expectation (dashed) under the background-only hypothesis. The dark and light shaded bands show the $\pm 1\sigma$ and $\pm 2\sigma$ uncertainties on the background-only expectation. (b) The observed (solid) local p_0 as a

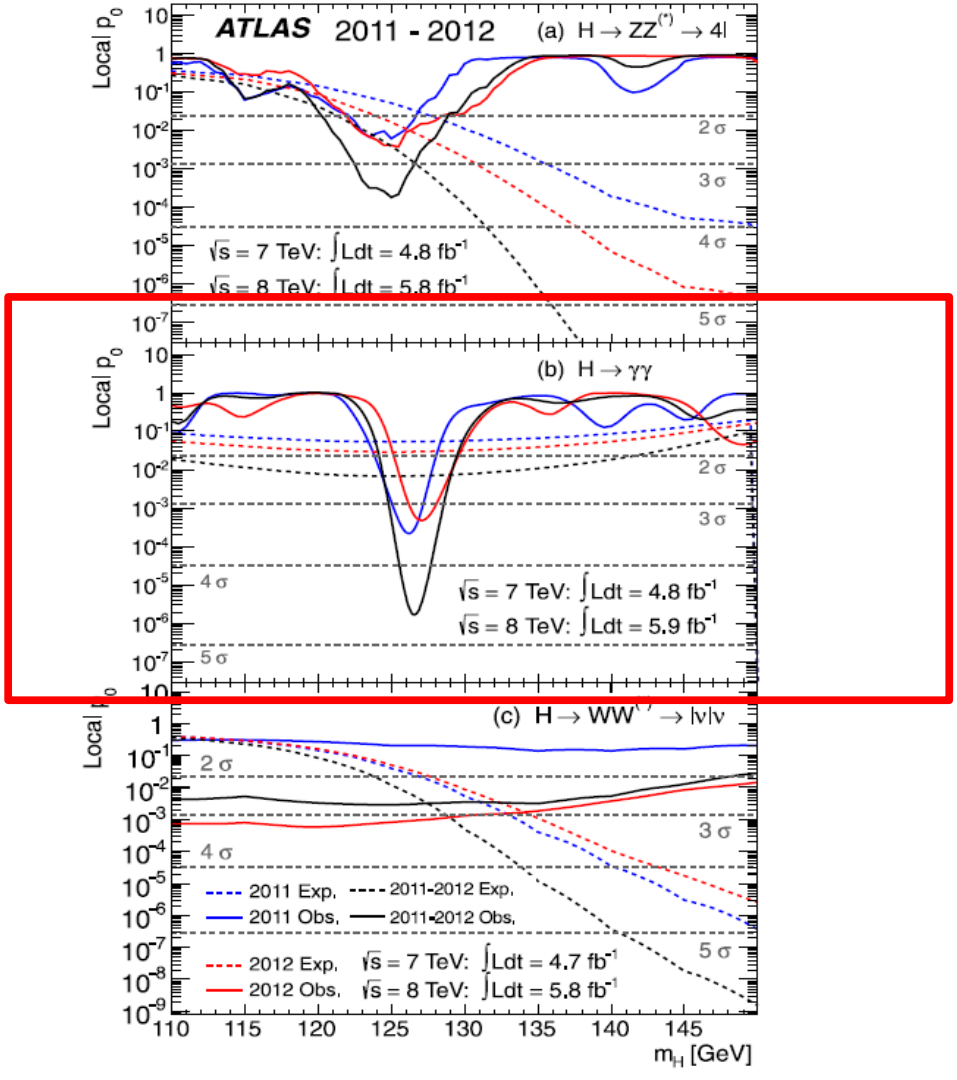


Fig. 8. The observed local p_0 as a function of the hypothesised Higgs boson mass for the (a) $H \to ZZ^{(*)} \to 4\ell$, (b) $H \to \gamma\gamma$ and (c) $H \to WW^{(*)} \to \ell\nu\ell\nu$ channels. The dashed curves show the expected local p_0 under the hypothesis of a SM Higgs boson signal at that mass. Results are shown separately for the $\sqrt{s} = 7$ TeV data (dark blue in the web version), the $\sqrt{s} = 8$ TeV data (light red in the web version)

As a result: the computing model

- In principle, no need of MPI
  - The probability of each event is independent
  - The pseudo-random number generator: affects the quality of Monte-Carlo generator, event generator
- Easily paralleled and even widely distributed among data locations
- Many end-users from different institutes
- Budgets, manpower, computing facilities, networks
  → Also distributed

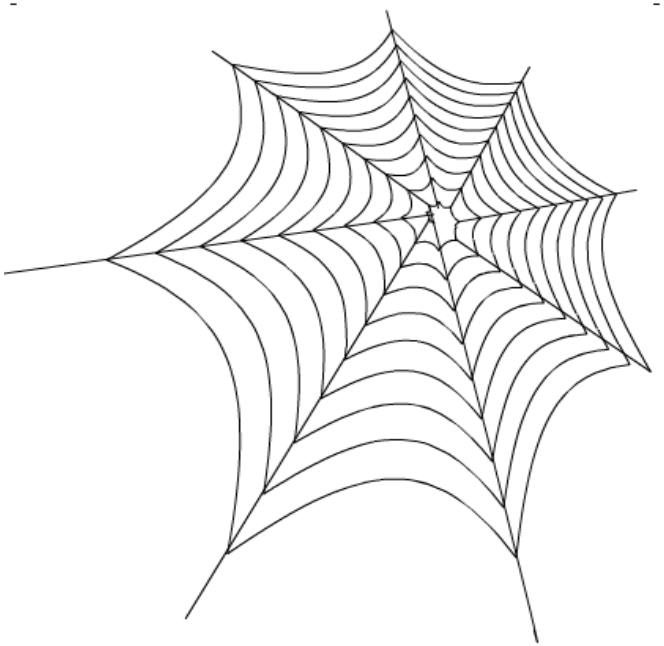
Basic Tools: ATLAS CLI and CVMFS

- Grid client software originally supported by *European Middleware Initiative* (EMI)
- CVMFS is a remote repository using FUSE file system
  - e.g. /cvmfs/atlas.cern.ch → A repository of all client software we need
  - Internally using HTTP → Need of network access
- PanDA client using ATLAS PanDA server is a tool to manage jobs
  - BigPanda monitoring system provides a Web interface
- Rucio client is a tool to interact with data
- ATLAS framework (Athena)
- ROOT (c++ HEP framework)

CVMFS (CERN VM File-System)

- Before CVMFS
  - Problems of distribution of common High Energy Physics software environments
  - Needed to build every software release in every university/facility
- *After CVMFS*
  - Everything is centrally managed/updated
  - Nightly build is also available
  - Nothing to do (e.g. build of every update) in local PCs / Clusters
  - Using general http proxy system like Squid → less network access and robust

ATLAS Metadata Interface (AMI)

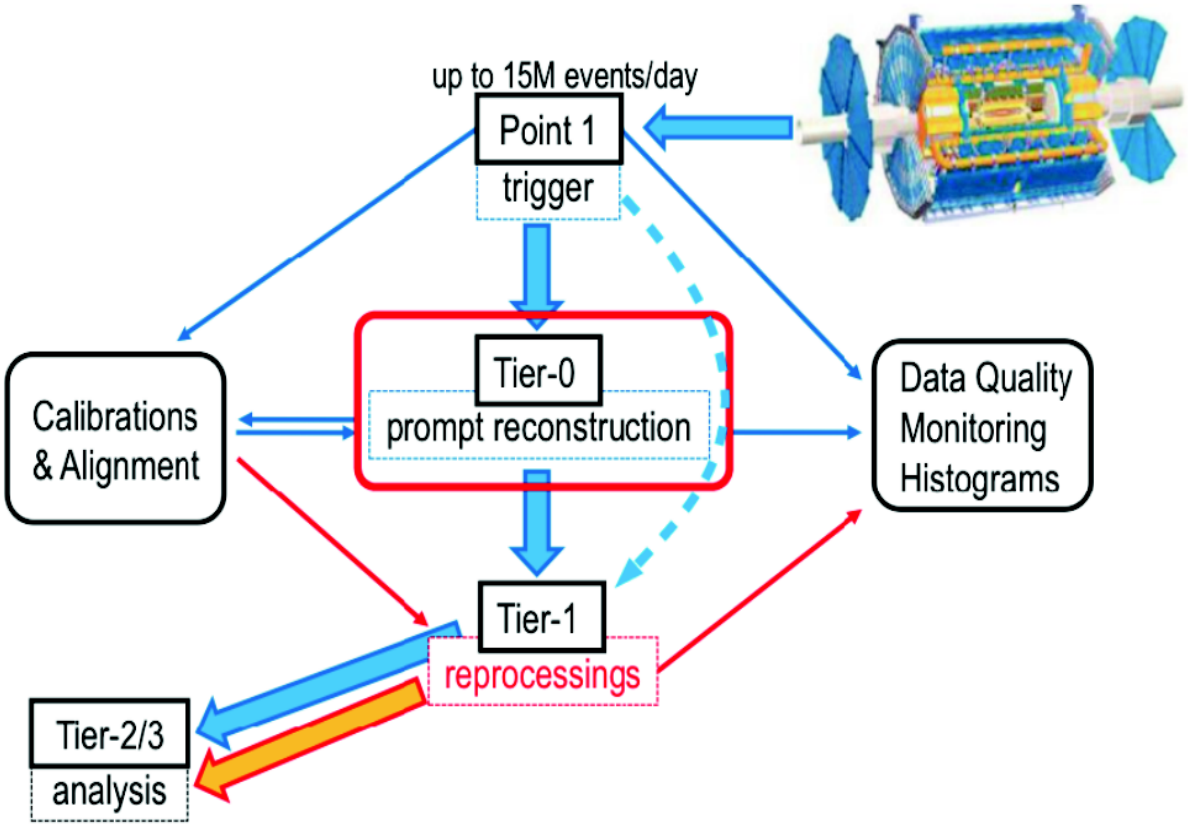


AMI (ATLAS Metadata Interface)

- Metadata = Data of Data (Description of data)
- Key of ATLAS Data Life Cycle Management
- On each step of data reconstruction in ATLAS, AMI Tags are generated
  - ATLAS data set, metadata link
    - Size and origin of dataset
    - File, number of events
    - Software parameter (AMI Tags)
    - MC parameter (PDF, generator, cross section, etc)
    - Lost files and Lumi blocks
    - Link to other applications (COMA, Rucio)
    - Data period
      - Luminosity, Trigger, Data of RAW data creation etc
  - Special Interface
    - AMI-Tags
      - Software Configuration Parameters
    - AMI-Glance
      - Data + Publication
    - Definition of data periods
    - Definition of physics containers
    - Reprocessing campaigns
    - Event count comparator

ATLAS Dataset - 2

- Tier-0
  - RAW is produced
  - Prompt data reconstruction
- ProdSys2
  - MC + Reprocessing
  - PanDA, JEDI, DEFT
- Distribution of dataset
  - Rucio



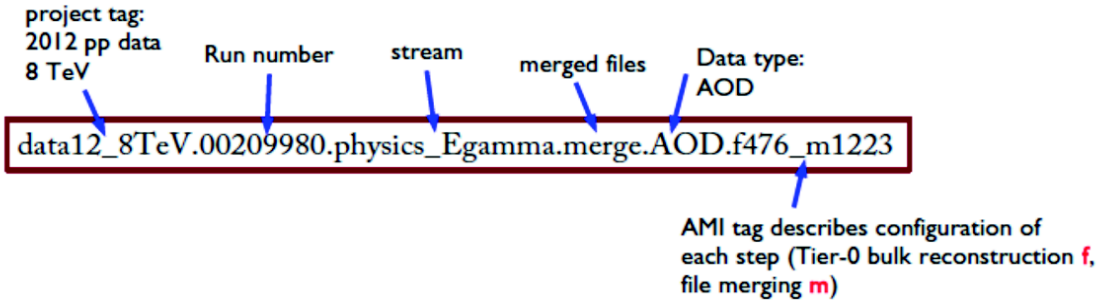
ATLAS Dataset - 3

- Data reconstruction → Example ： x353 = AMI tag
  - Reco_tf.py —AMI=x353 --inputBSFile=tier0_RawData.data
- Reprocessing Campaign
  - After improvement of ATLAS software or framework, often huge reconstruction jobs can run
    - The campaign require many computing resources
    - As a result, many different versions of AOD (Analysis Object Data)
      - Derivation framework: AOD → Derived AOD
      - About 1% size of ogirinal AOD
      - Selection of particular event, parameter etc
      - Different configurations, software and conditions

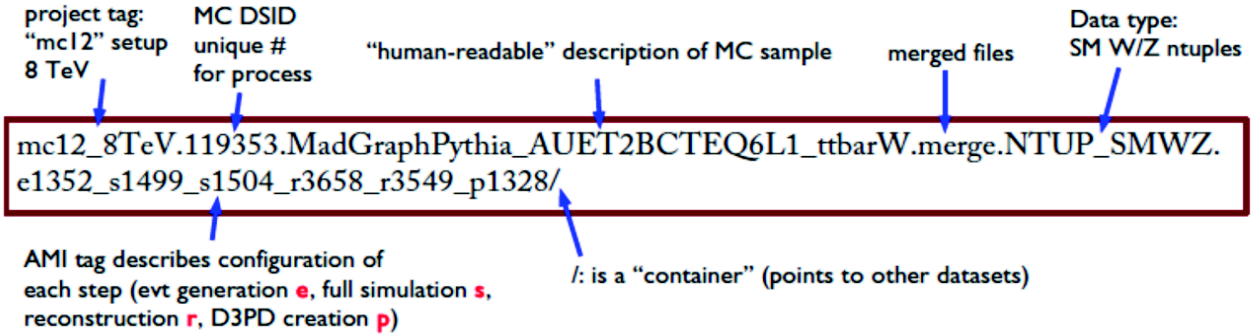
ATLAS Dataset - 4

- Dataset = Collection of files
  - Collision data (data) and Monte Carlo (mc)

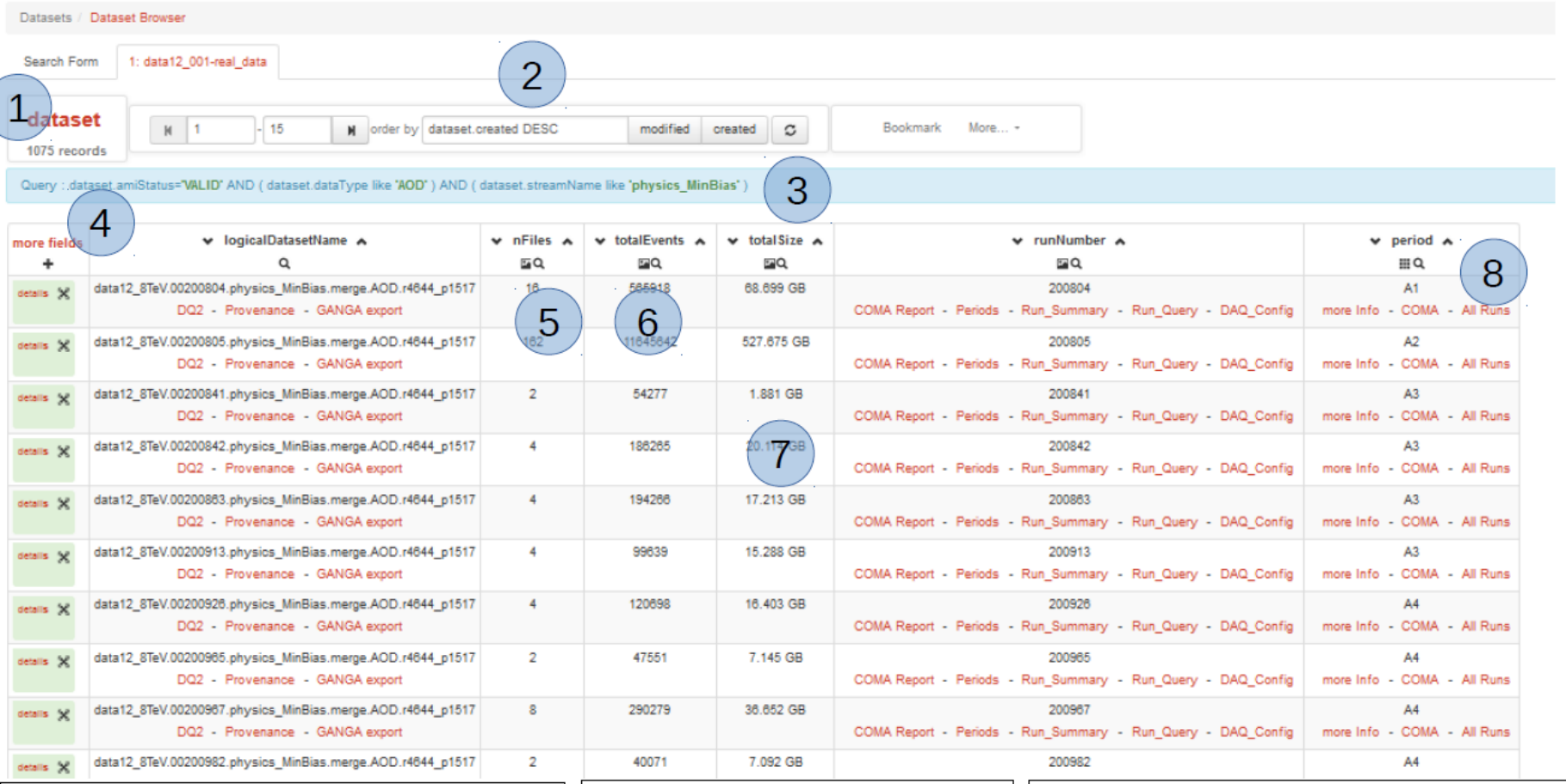
Data:



Simulation:



AMI WebUI - 1

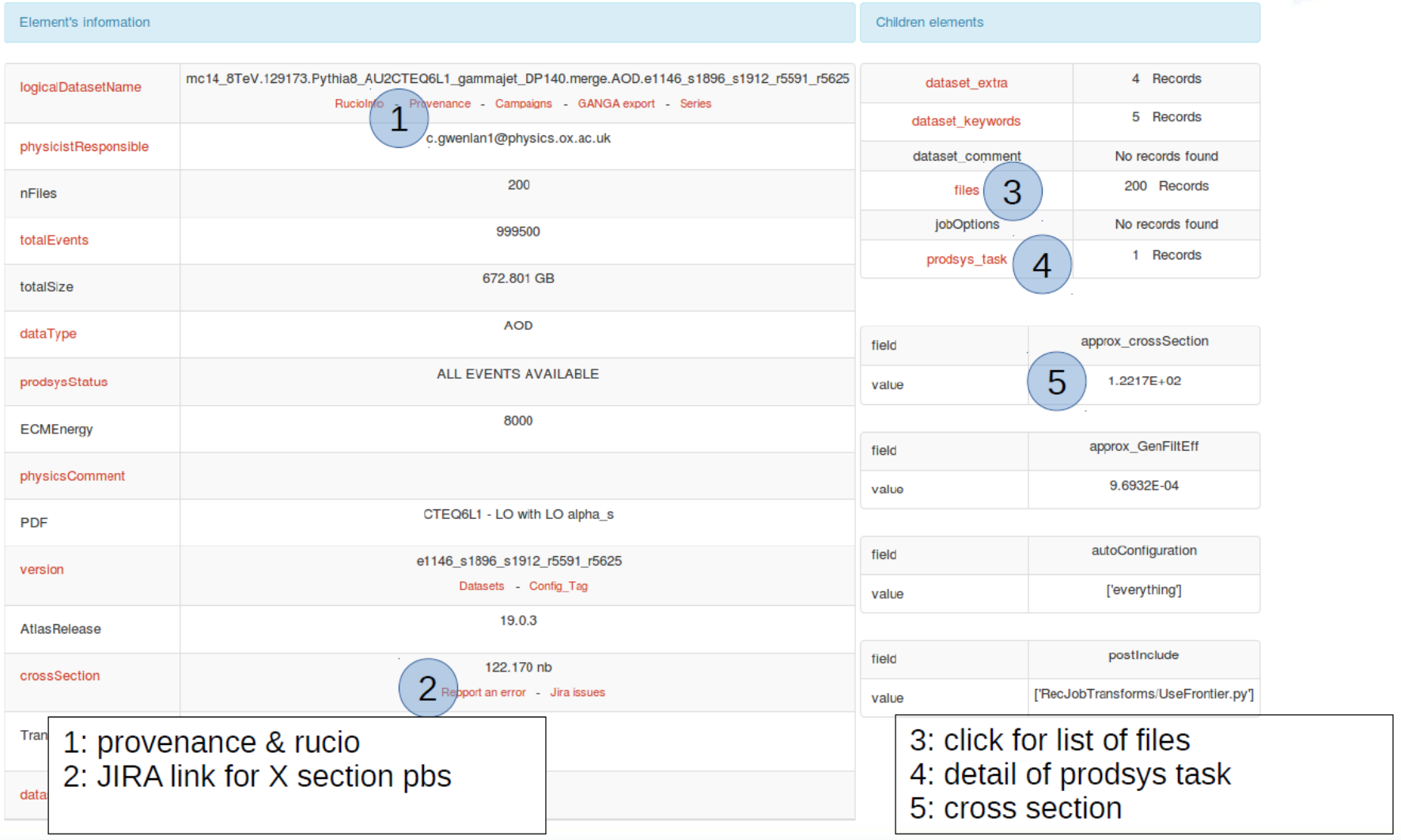


1: number of results
2: default order, more recent first

3: query clauses
4: +/- fields
5,6: filter,calculator

7: conversion of units
8: group by, order by tools

AMI WebUI - 2


Element's information
Children elements
logicalDatasetName
mc14_8TeV.129173.Pythia8_AU2CTEQ6L1_gammajet_DP140.merge.AOD.e1146_s1896_s1912_r5591_r5625
RucioInfo - Provenance - Campaigns - GANGA export - Series
1
physicistResponsible
c.gwenlan1@physics.ox.ac.uk
nFiles
200
totalEvents
999500
totalSize
672.801 GB
dataType
AOD
prodsysStatus
ALL EVENTS AVAILABLE
ECMEnergy
8000
physicsComment
PDF
CTEQ6L1 - LO with LO alpha_s
version
e1146_s1896_s1912_r5591_r5625
Datasets - Config_Tag
AtlasRelease
19.0.3
crossSection
122.170 nb
2
Report an error - Jira issues
dataset_extra
4 Records
dataset_keywords
5 Records
dataset_comment
No records found
files
3
200 Records
jobOptions
No records found
prodsys_task
4
1 Records
field
approx_crossSection
value
5
1.2217E+02
field
approx_GenFiltEff
value
9.6932E-04
field
autoConfiguration
value
['everything']
field
postInclude
value
['RecJobTransforms/UseFrontier.py']
1: provenance & rucio
2: JIRA link for X section pbs
3: click for list of files
4: detail of prodsys task
5: cross section

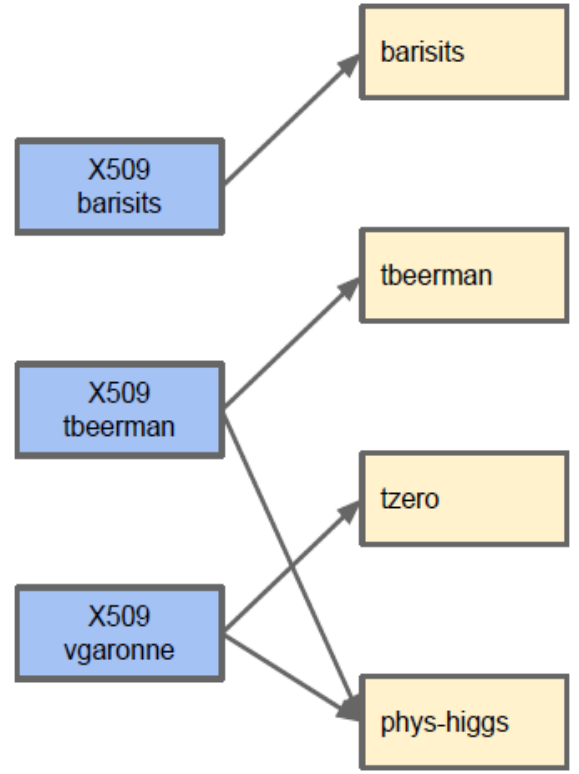
Rucio (ATLAS Data Management System)

Basic concepts - 1

- Data Management System for Run-2 in ATLAS distributed computing system
  - Used to download outputs of Grid jobs, moving data and searching for them
- Rucio CLI tools from CVMFS
  - Web interface provides similar functionality

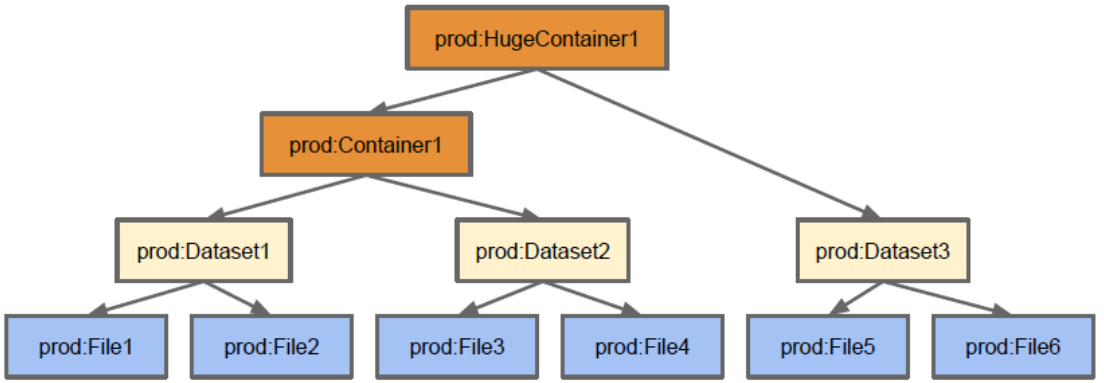
Basic concepts - 2

- Accounts
  - A Rucio Account can represent users, groups (phys-susy) or activities (panda, tzero)
  - Quota and permissions are associated to an account
  - One can connect to a Rucio account using X509 certificates/proxies, Kerberos
  - One credential can be used to map to different accounts.



Basic concepts - 3

- Rucio namespace
  - 3 types of Data Identifiers (DIDs): File, Datasets, Containers.
    - Dataset: Collection of Files
    - Containers: Collection of Datasets or Containers
  - The namespace is divided using scopes. A name is unique within a scope but can be used in other scopes. A DID is identified by a scope and a name



List DIDs

- A data Identifier is found by name and scope

To list all DIDs within a scope

rucio list-dids user.gkawamur:*

```
+-------------------------------------------------------------+--------------+
| SCOPE:NAME                                                  | [DID TYPE]   |
|-------------------------------------------------------------+--------------|
| user.gkawamur:user.gkawamur.pruntest_7168.log               | CONTAINER    |
| user.gkawamur:user.gkawamur.tut.helloworldathena.test.log   | CONTAINER    |
| user.gkawamur:user.gkawamur.20160323144306.1.log            | CONTAINER    |
| user.gkawamur:user.gkawamur.tutorial.grid.deriv.test1.log   | CONTAINER    |
| user.gkawamur:user.gkawamur.pruntest_14222.log              | CONTAINER    |
```

Using filters (search)

rucio list-dids data16_13TeV:* --filter type=DATASET,datatype=AOD

```
+-------------------------------------------------------------------------------------------------------------------+--------------+
| SCOPE:NAME                                                                                                        | [DID TYPE]   |
|-------------------------------------------------------------------------------------------------------------------+--------------|
| data16_13TeV:data16_13TeV.00293572.physics_CosmicCalo.merge.AOD.x387_m1554                                        | DATASET      |
| data16_13TeV:data16_13TeV.00297447.physics_Standby.merge.AOD.x416_m1583                                           | DATASET      |
| data16_13TeV:data16_13TeV.00297041.physics_Main.merge.AOD.f686_m1583                                              | DATASET      |
| data16_13TeV:data16_13TeV.00297041.physics_CosmicCalo.merge.AOD.f686_m1583                                        | DATASET      |
| data16_13TeV:data16_13TeV.00297041.physics_ZeroBias.merge.AOD.f686_m1583                                          | DATASET      |
```

Show metadata

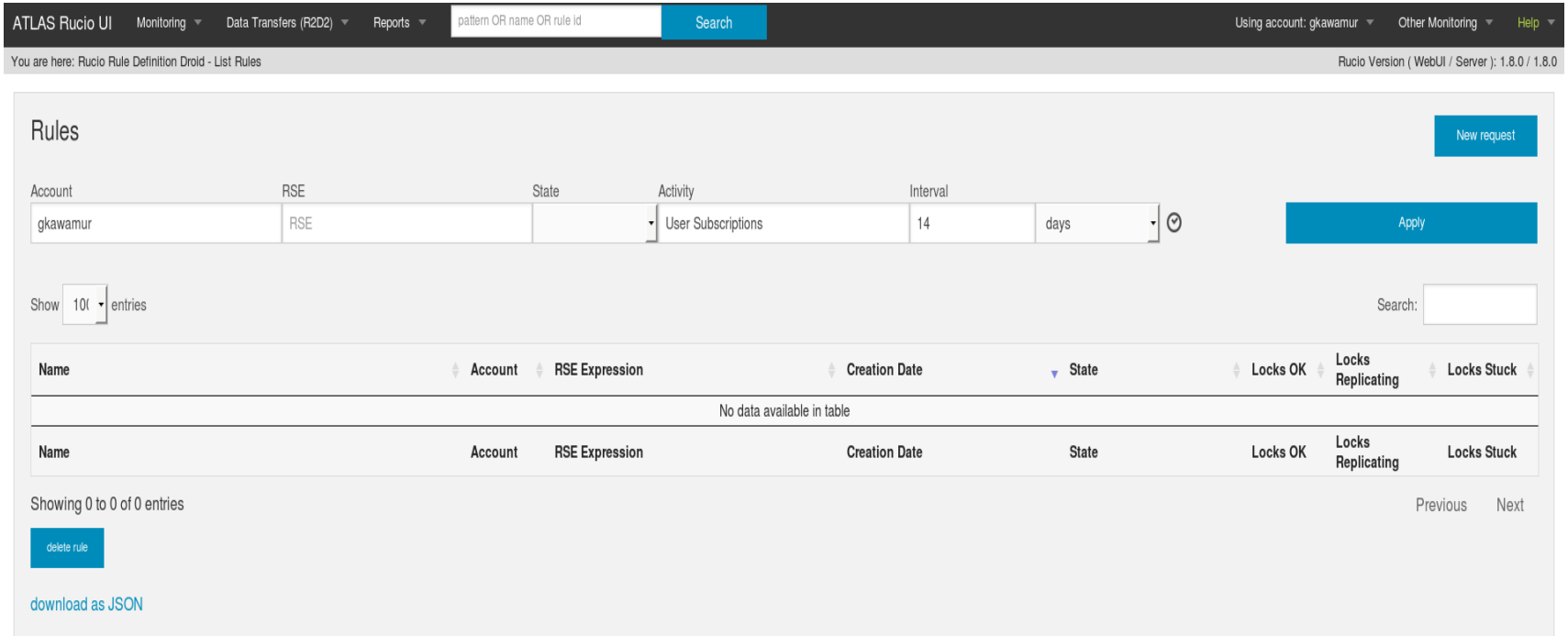
Showing metadata of a dataset

rucio get-metadata data16_13TeV:data16_13TeV.00303819.physics_Main.merge.AOD.f716_m1620

purge_replicas: None
campaign: None
is_new: None
is_open: False
closed_at: 2016-07-18 07:39:44
deleted_at: None
availability: AVAILABLE
eol_at: None
guid: None
panda_id: None
provenance: None
accessed_at: 2016-09-19 22:00:22
version: f716_m1620
scope: data16_13TeV
hidden: False
md5: None
events: 25475721
adler32: None
complete: None
lumiblocknr: None
monotonic: False
updated_at: 2016-09-19 22:00:42
obsolete: False
transient: None
did_type: DATASET
suppressed: True
expired_at: None
stream_name: physics_Main
account: tzero
run_number: 303819
name: data16_13TeV.00303819.physics_Main.merge.AOD.f716_m1620
task_id: None
datatype: AOD
created_at: 2016-07-17 04:17:53
bytes: 6462688464808
project: data16_13TeV
length: 2595
prod_step: merge
phys_group: None

RucioUI

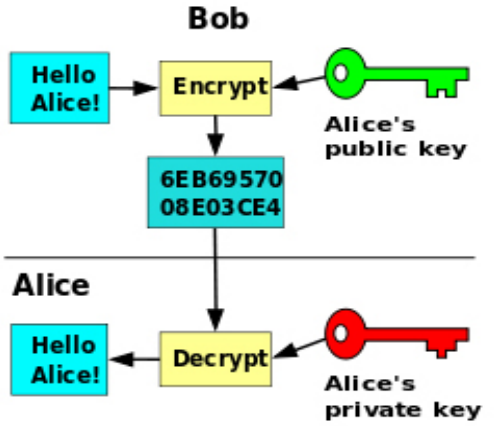
- Move data between RSEs, use Rucio replication rules
- Such request can be generated by CLI and UI
- The UI tool is *the Rucio Rule Definition Droid (R2D2)* *https://rucio-ui.cern.ch/r2d2*
- Basic need: X509 certificate in your browse



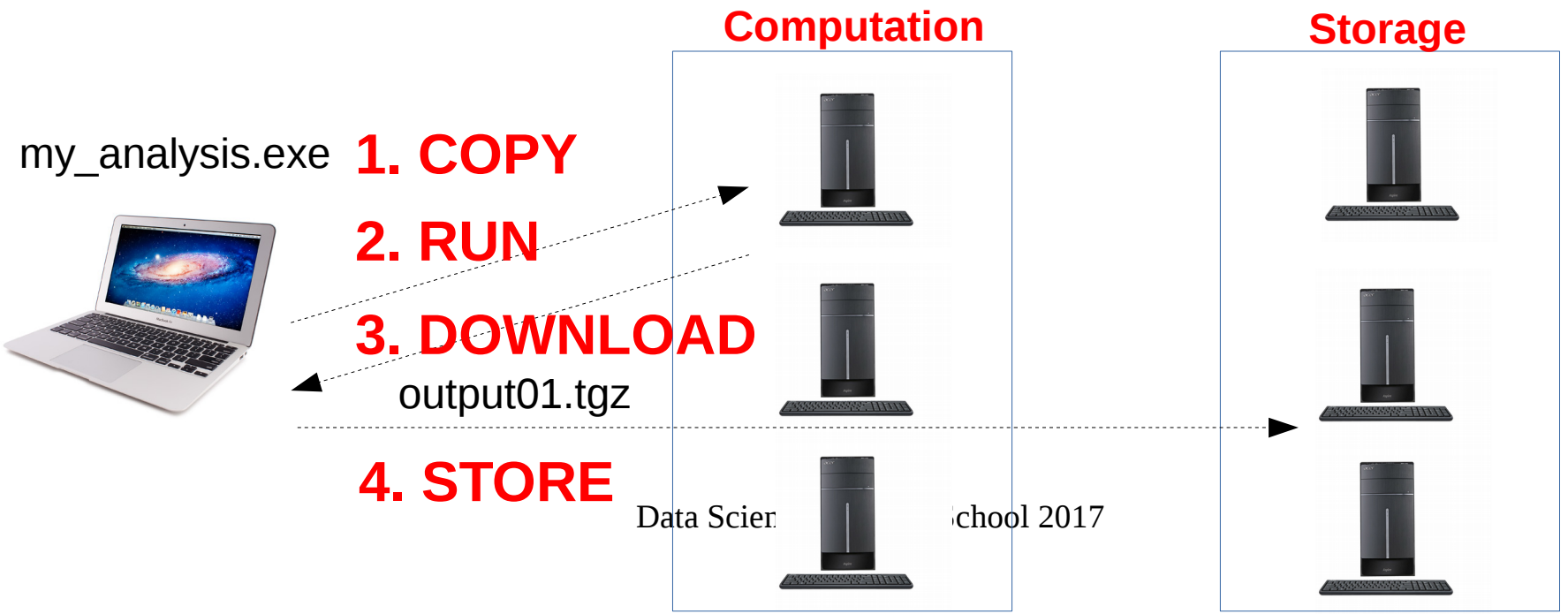
Introduction to World LHC Computing Grid

Concepts - 1

- How it works (*without Grid*)
  - Only PKI by openssl -> a simplest way




```
$ scp my_analysis.exe computer01:~/
$ ssh computer01 "~/myprogram.exe run 01"
$ scp computer01:output01.tgz ./output01.tgz
$ scp output01.tgz storage02:output01.tgz
```

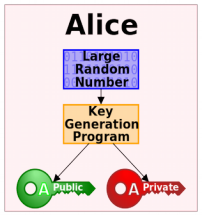


Concepts - 2

- How it works (*with Grid*)
  - Need of basic stuffs
    - Security policy based on PKI
    - Working with many users (physicists)
    - Recognition of users in affiliation, organization and experiment
    - Control of Job allocation
    - Control of storage space allocation and data replication
    - Control among geographically distributed computers
    - Interface of local computing infrastructures
    - Tree-like information search for computing clusters
      - E.g. like Domain Name Server
    - End user tools to make life easy

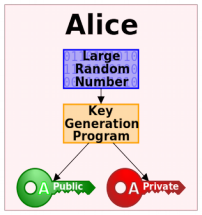
Concepts - 3

- Then, hierarchical signatures



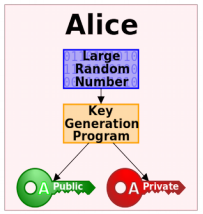
CA1

Sign



CA2

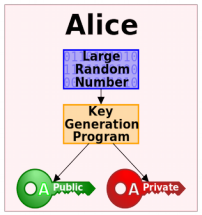
Sign



My X509 Certificate
= Grid user certificate

Concepts - 4

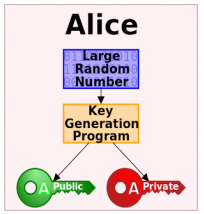
- Generating a new certificate = proxy certificate



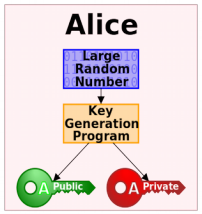
CA1

Verifiable

Proxy certificate
(with limited validity)



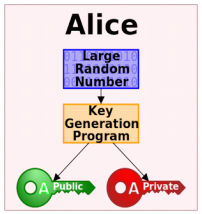
Sign



CA2

Sign

Generating primary numbers +
Signed by you + PK



My X509 Certificate
= Grid user certificate

Concepts - 5

- Virtual Organization (VO), VOMS
  - VOMS extends a certificate with a VO

Certificate Authority

Sign

User certificate

Sign

Proxy certificate
= a limited copy of your certificate

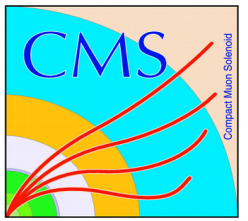
Requesting VO
voms-proy-init

VOMS server

ATLAS VOMS
CMS VOMS
LHCb VOMS

Obtaining VO extension

VOs



Concepts - 6

- A similar analogy
  - Authenticated and authorized for your tasks

Certificate Authority

Sign

User certificate

Sign

Proxy certificate
= a limited copy of your certificate

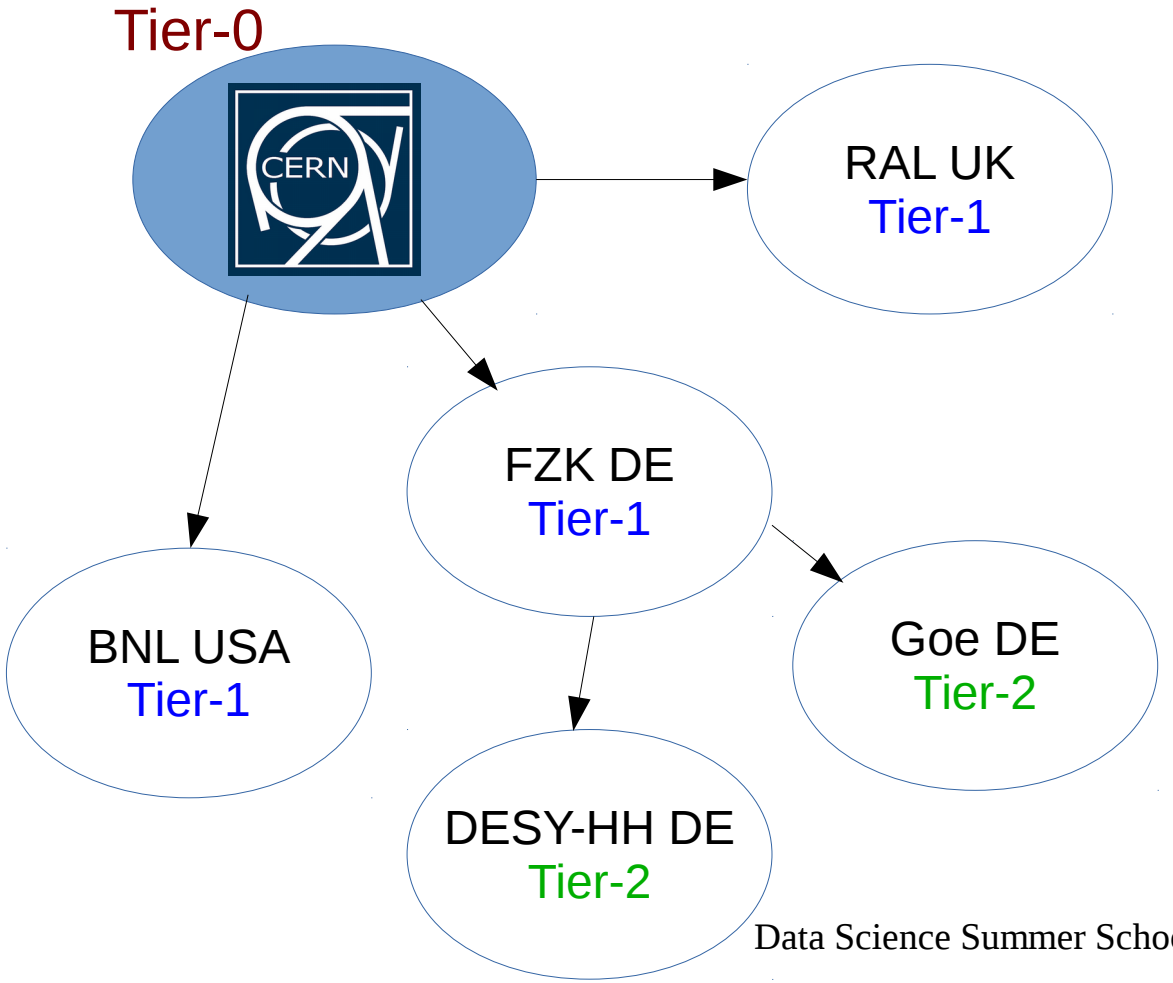
Controlled

Authentication
= who you are?

Authorization
= can enter a new land (Grid)

ATLAS Grid job - 1

- LHC multi-tier structure
  - WLCG = Worldwide LHC Computing Grid



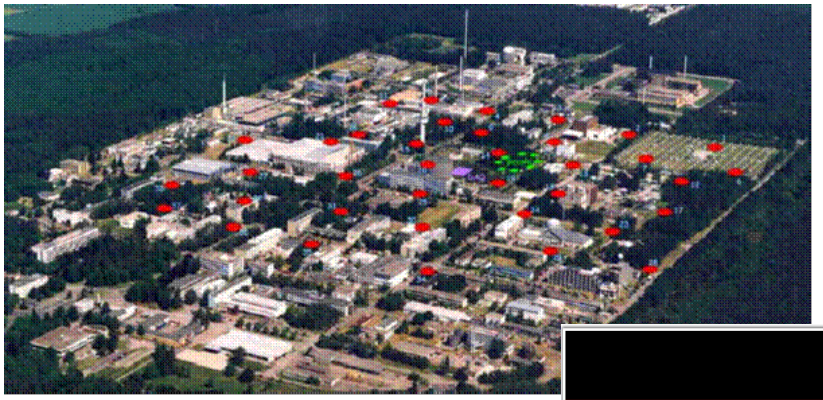
Tier-0: Raw data, Data store (in tape), Pre-processing, Reconstruction

Tier-1: National LCG-Centre, faster network connections and lager storage spaces (e.g. Tape), MC production, user analysis, etc.

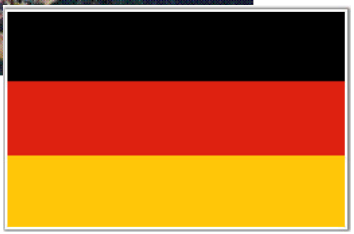
Tier-2: University or Facility level computing sites. MC production, user analysis, etc.

ATLAS Grid job - 1

- FZK Tier-1



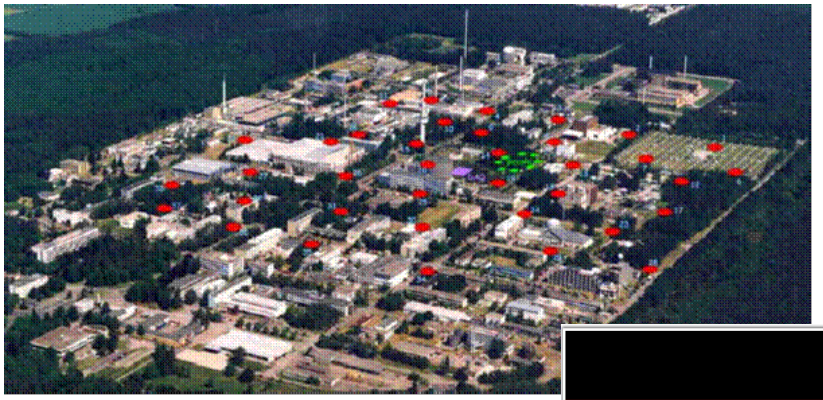
Forschungszentrum Karlsruhe
in der Helmholtz-Gemeinschaft



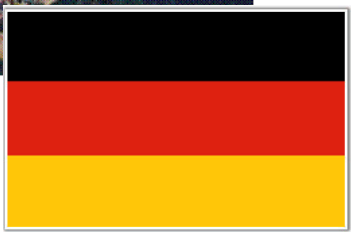
- BNL Tier-1

ATLAS Grid job - 1

- FZK Tier-1



Forschungszentrum Karlsruhe
in der Helmholtz-Gemeinschaft



Job queue endpoint

Computing resources

Storage spaces

DDM endpoint

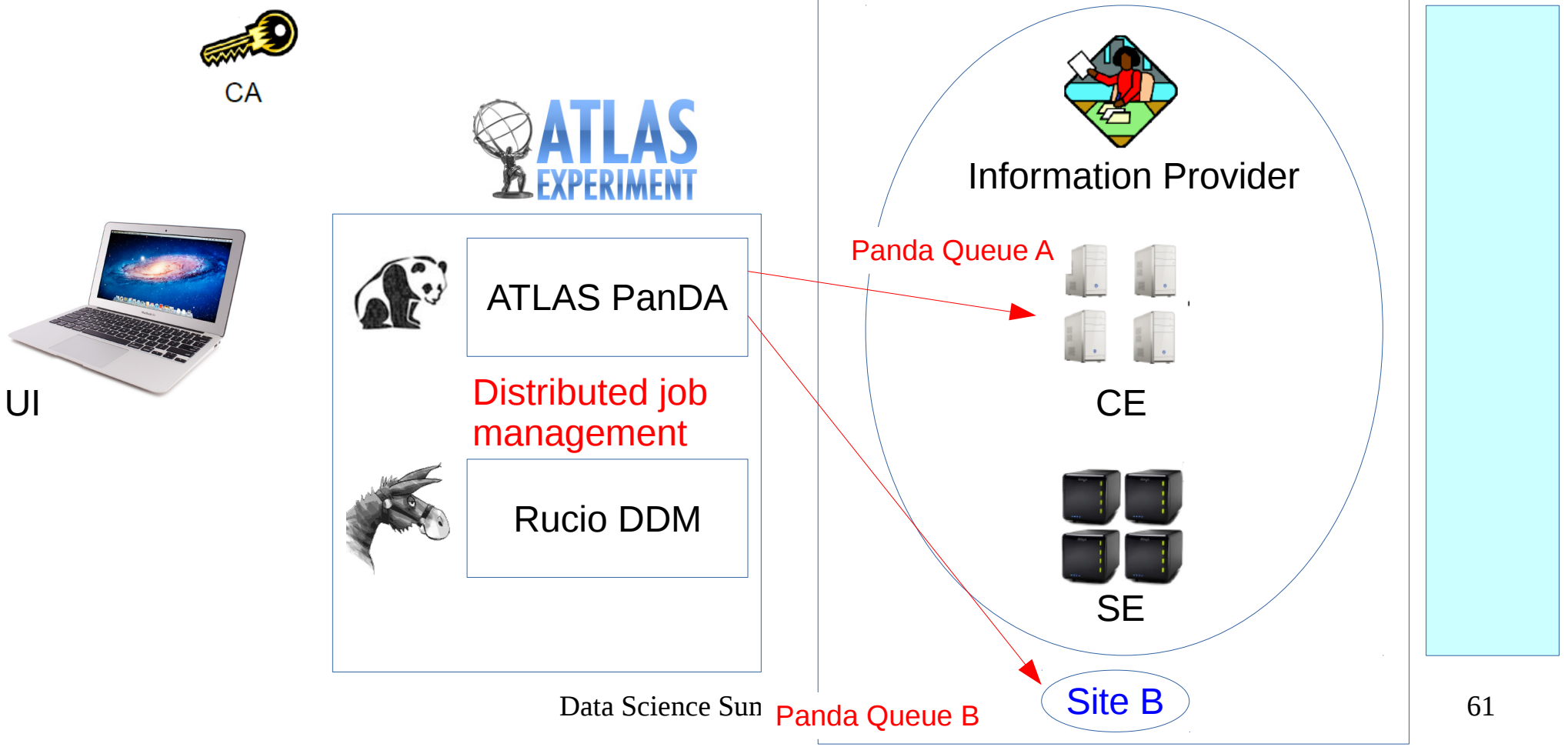
Protocols: srm, gridftp, https, xroot

FZK Computing centre

DDM = Distributed Data Management

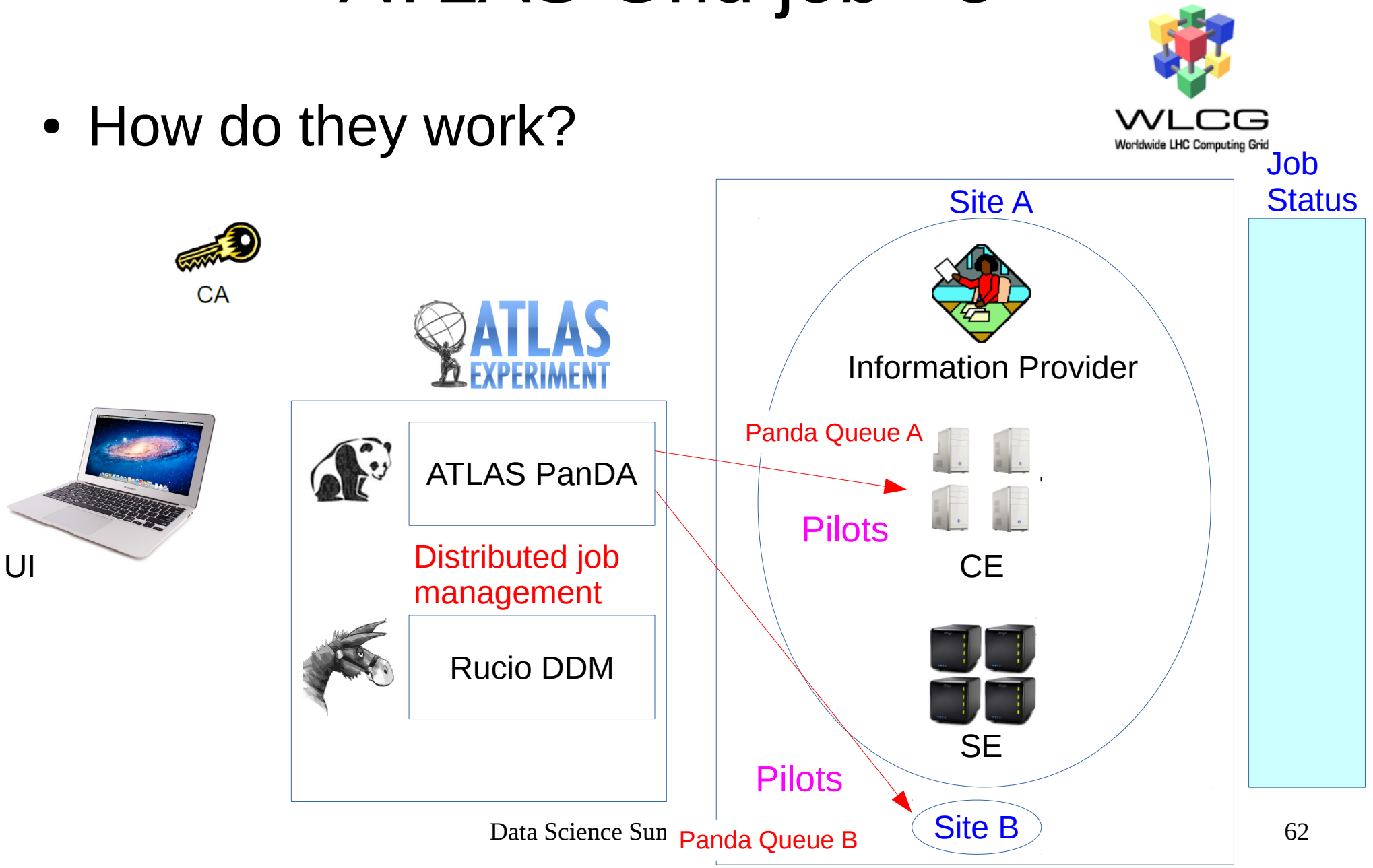
ATLAS Grid job - 2

- How do they work?



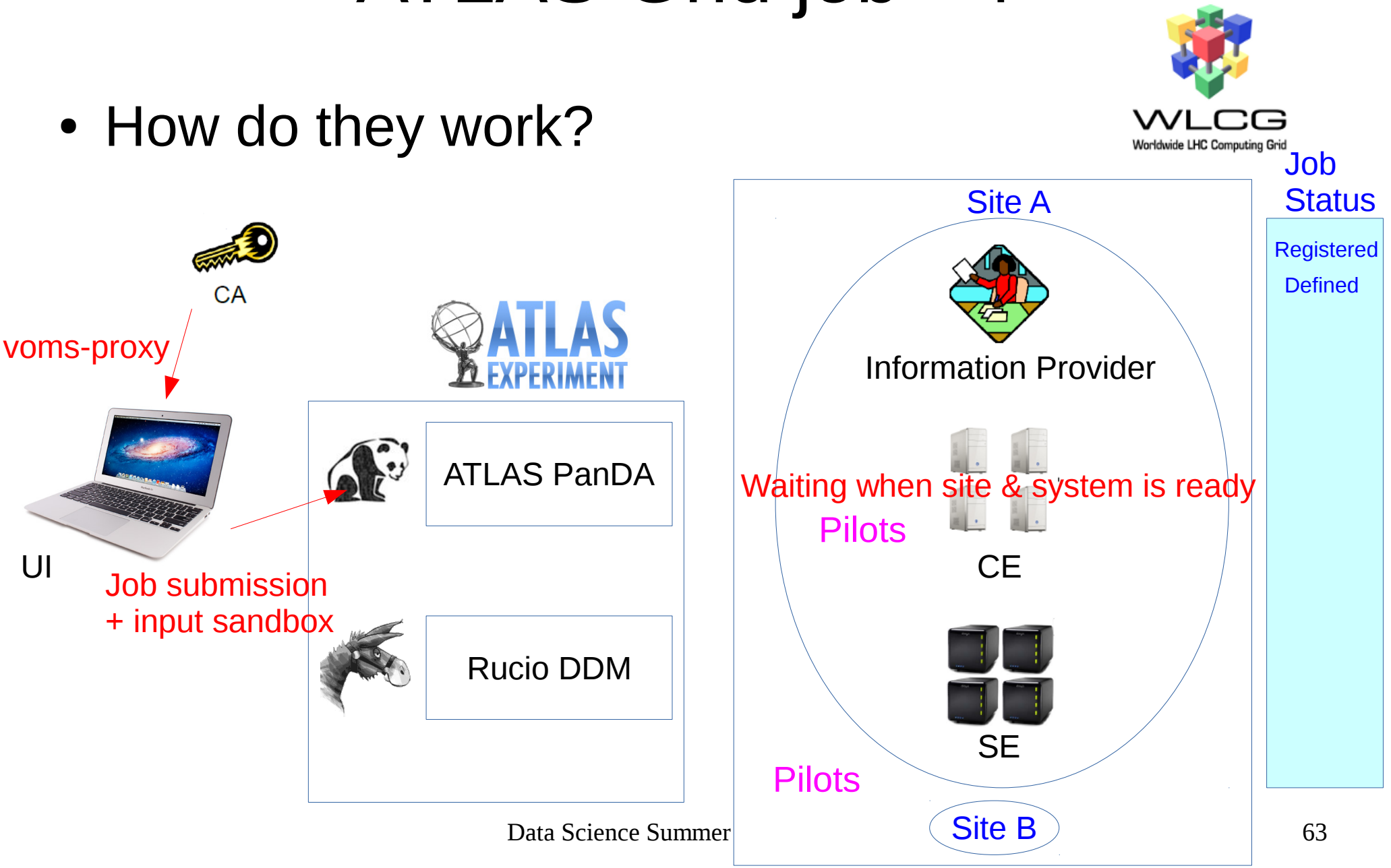
ATLAS Grid job - 3

- How do they work?



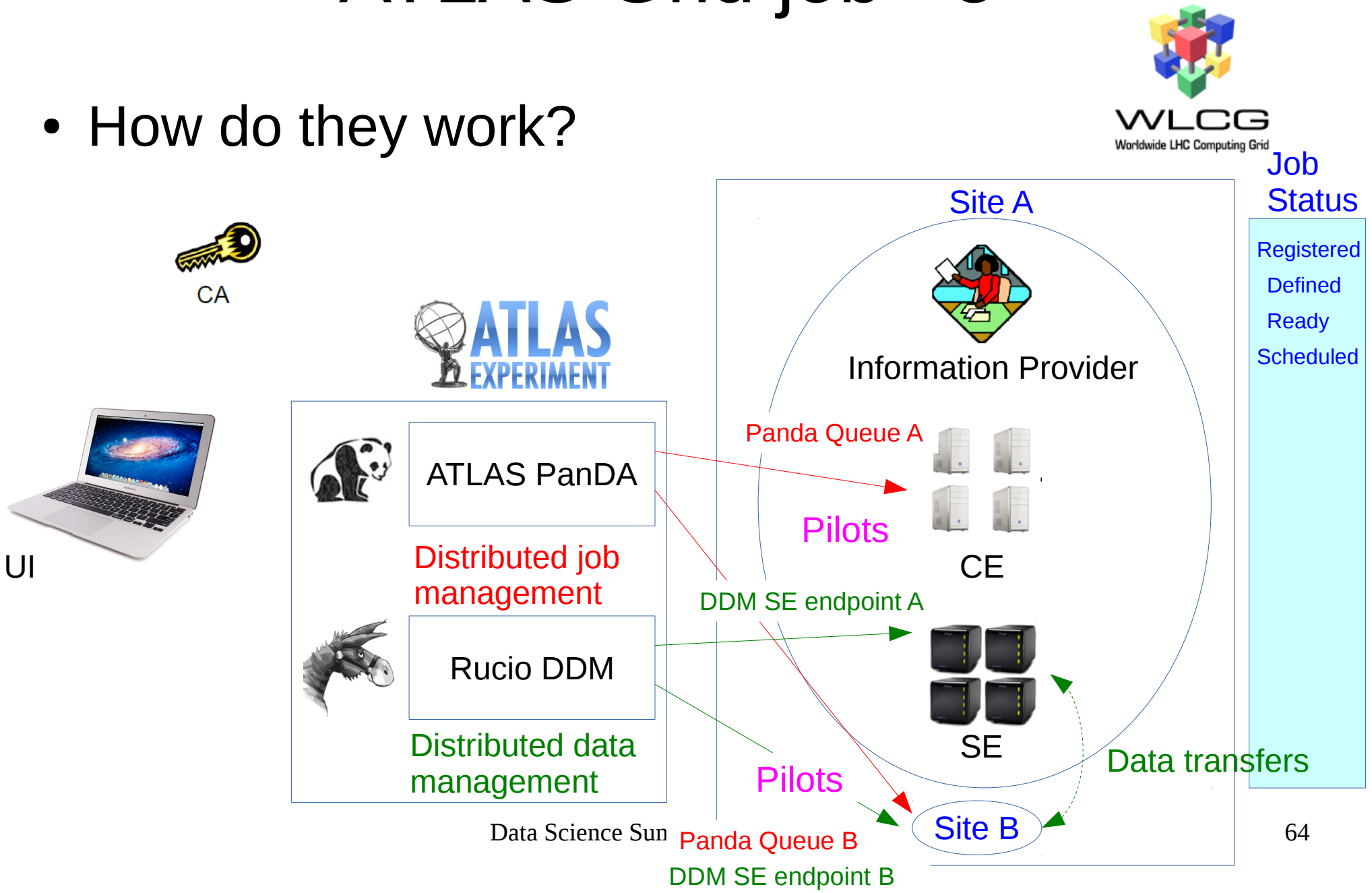
ATLAS Grid job - 4

- How do they work?



ATLAS Grid job - 5

- How do they work?



ATLAS Grid job - 6

- How do they work?

CA

UI

ATLAS PanDA

Distributed job management

Rucio DDM

Distributed data management

Site A

Information Provider

Panda Queue A

Running!

CE

DDM SE endpoint A

Analysis data

SE

Site B

Panda Queue B

DDM SE endpoint B

Job Status

Registered

Defined

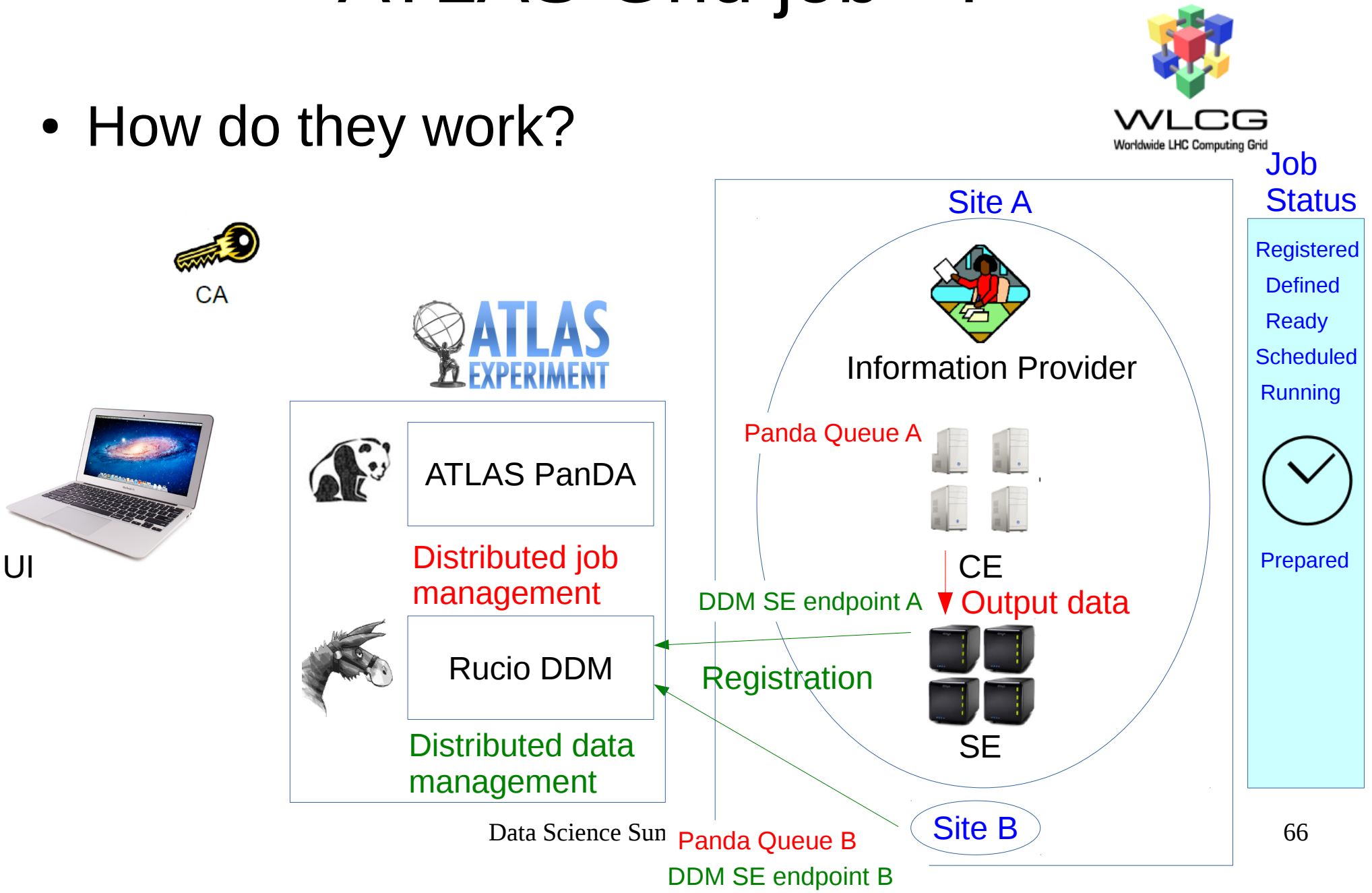
Ready

Scheduled

Running

ATLAS Grid job - 7

- How do they work?



ATLAS Grid job - 8

- How do they work?

CA

UI

ATLAS PanDA

Distributed job management

Rucio DDM

Distributed data management

Site A

Information Provider

Panda Queue A

Done!

CE

DDM SE endpoint A

SE

Site B

Panda Queue B

DDM SE endpoint B

Job Status

Registered

Defined

Ready

Scheduled

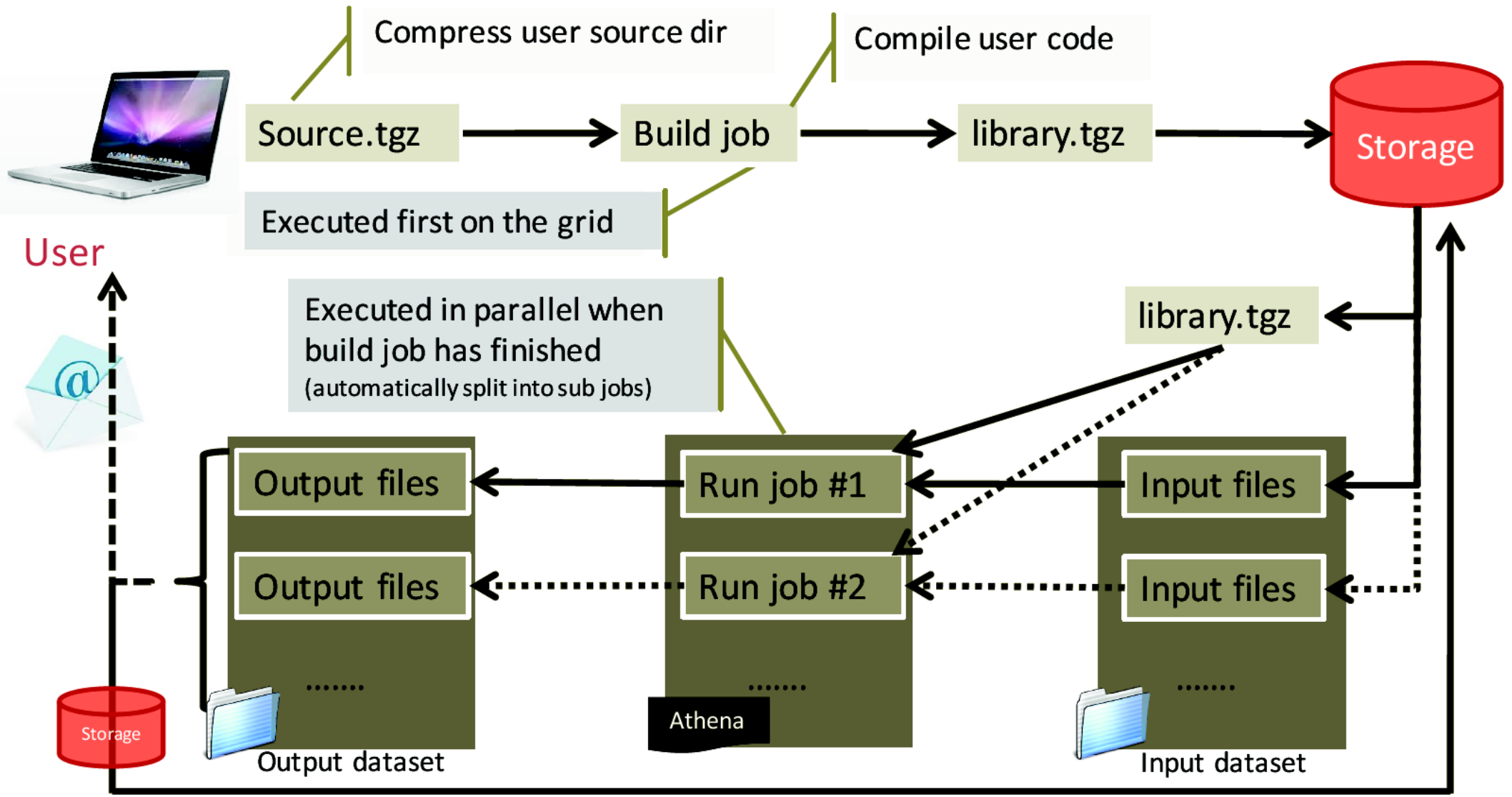
Running

Prepared

Done

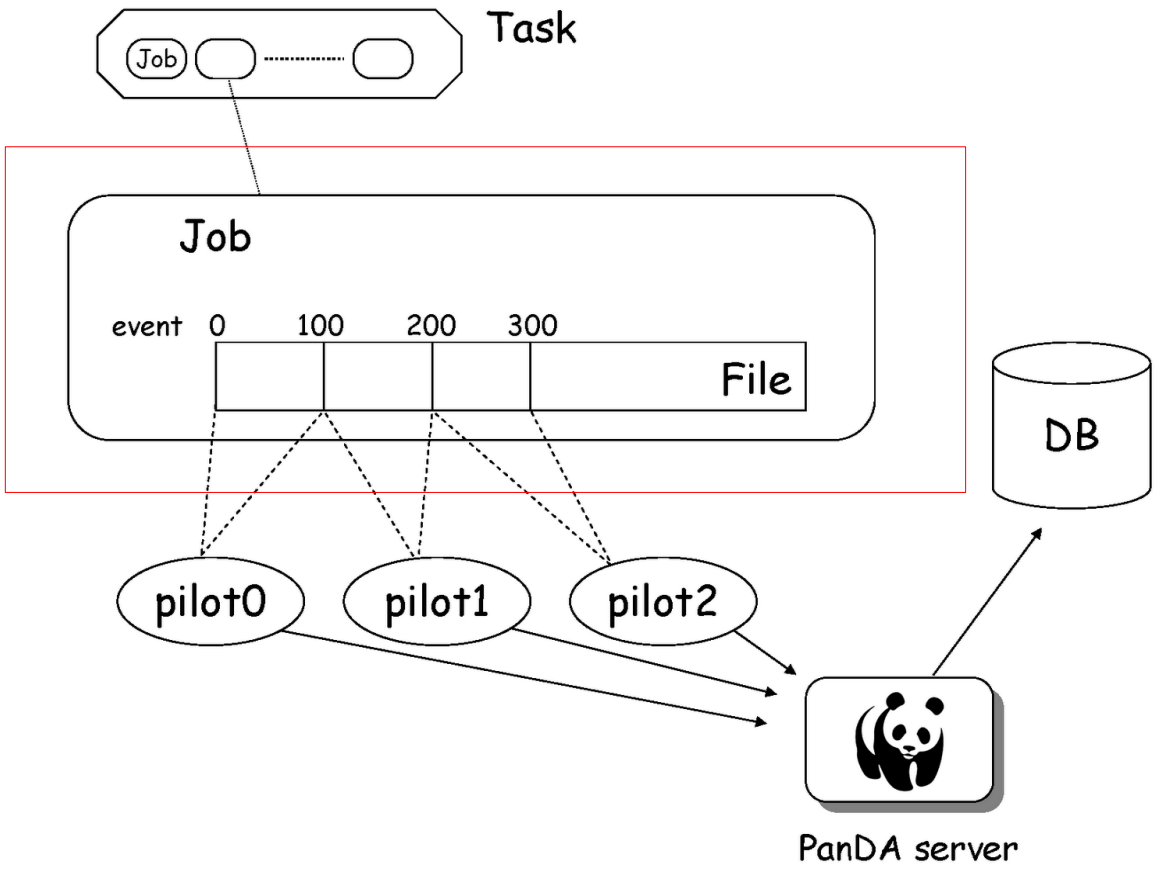
PanDA (ATLAS Job Management System)

Build job & Run job



Event level processing

- Task (calculation) can be split by evenl level handler
  - Event Level Handler can generate basic pilot jobs according to events in given file(s)



Conclusion

- Successful deployments for huge collaborations and experiments in the real world
  - Infrastructure demonstrated to be able to support LHC data taking and analysis
- Distributed parallel computing system works among different institutes, batch systems and countries
- Significant operational infrastructure behind it
- Service Oriented Architecture: A reliable and robust service of many components need

Hands-on

Open the hands-on slide.
Use a grid proxy (Note: the certificate is valid during the time slots of the hands-on)