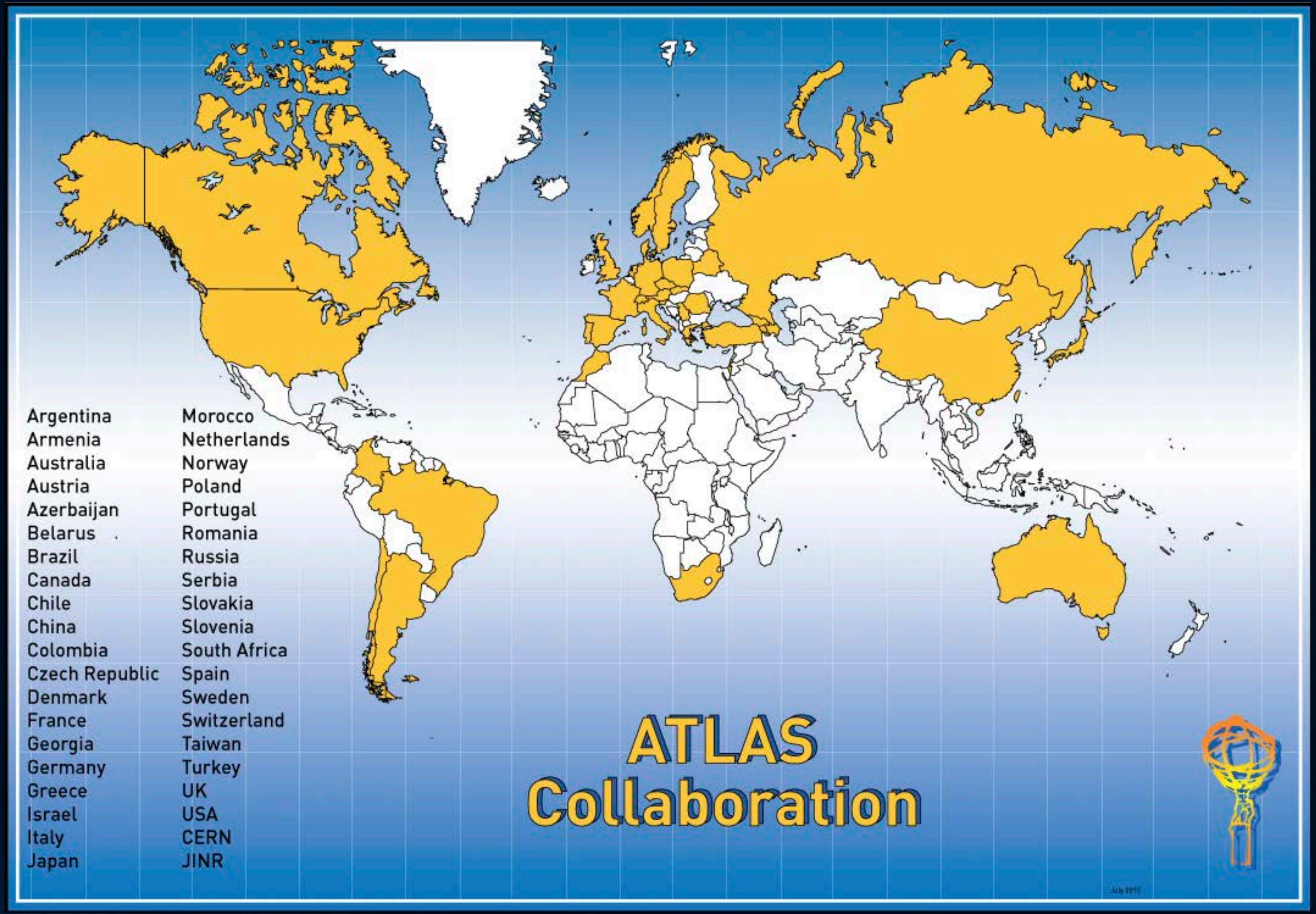
The ATLAS Experiment at CERN's LHC:

How do you analyze 40 million physics pictures per second?

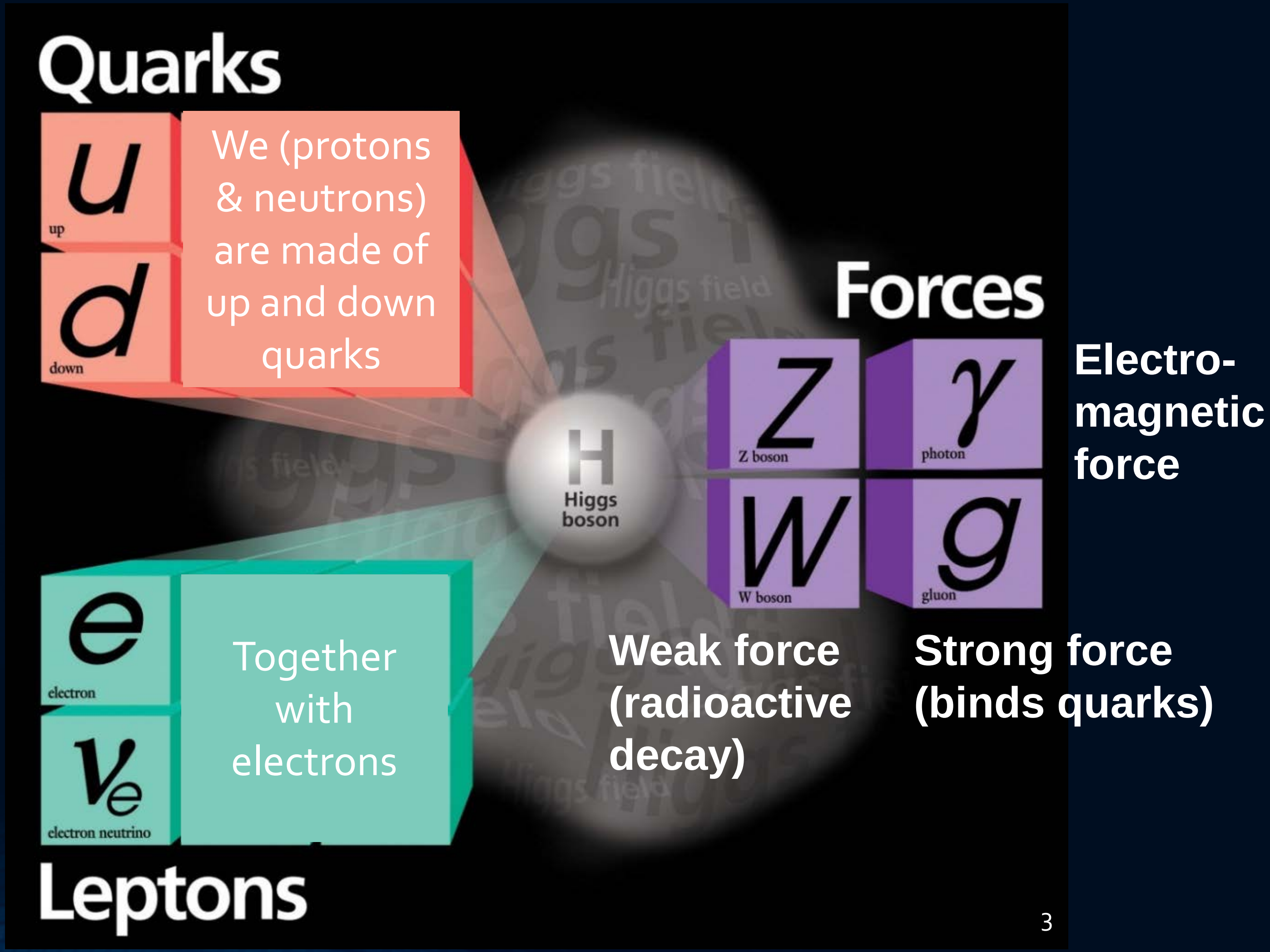
MICHAEL TUTS
COLUMBIA UNIVERSITY
NY DATA SUMMIT
JUNE 13, 2019

The ATLAS Collaboration

- Designed, built and operates the ATLAS Detector
  - ~3,000 physicists (including about 1,200 graduate students) from 38 countries and 183 institutions (44 from here in the US)
  - Over 850 publications to date
- Located at CERN in Geneva, Switzerland
- Uses the Large Hadron Collider (LHC) accelerator
  - 4 experiments there: ATLAS, CMS, ALICE and LHCb

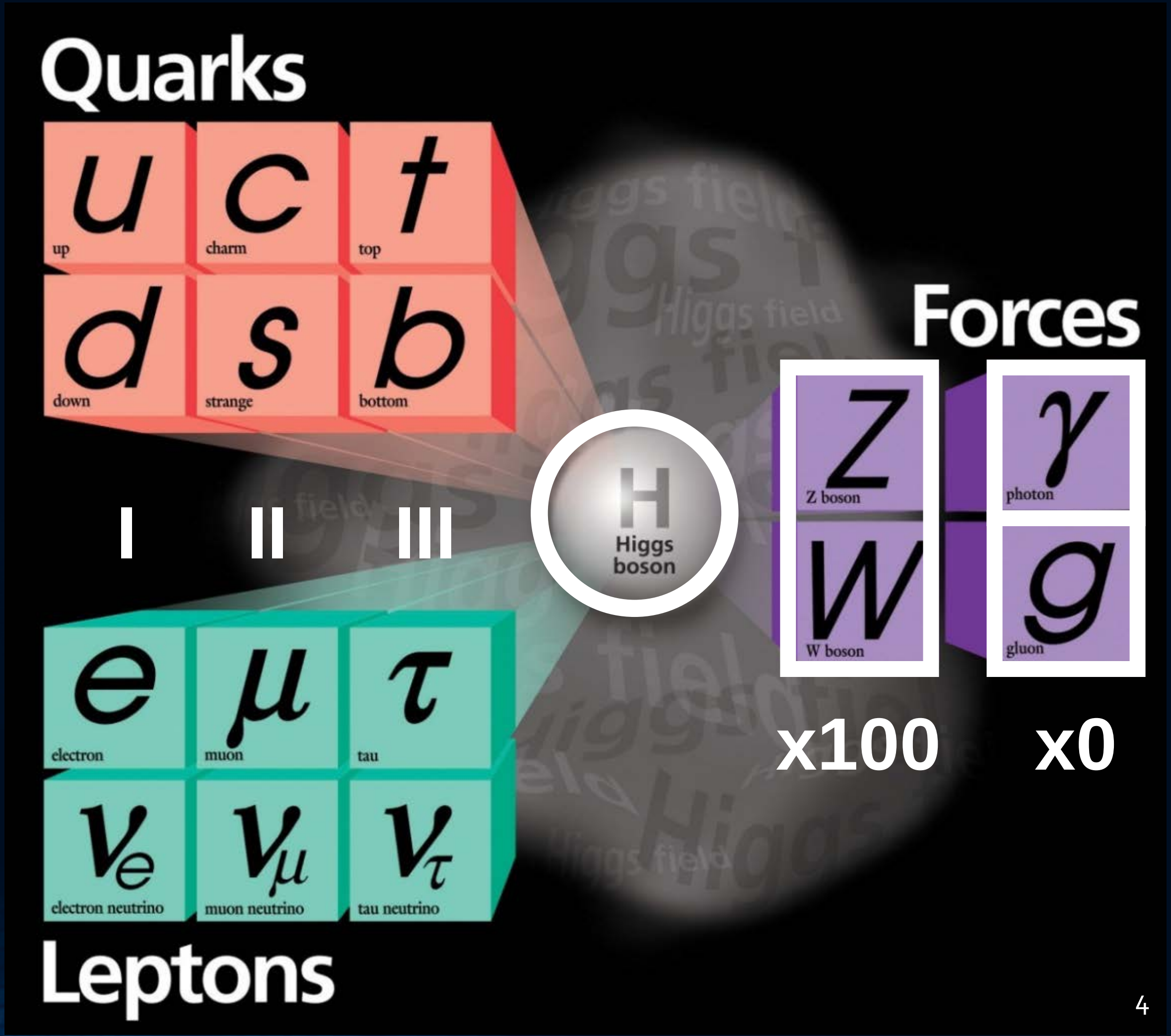


The World of Particle Physics ("The Standard Model")



But Questions Remain...

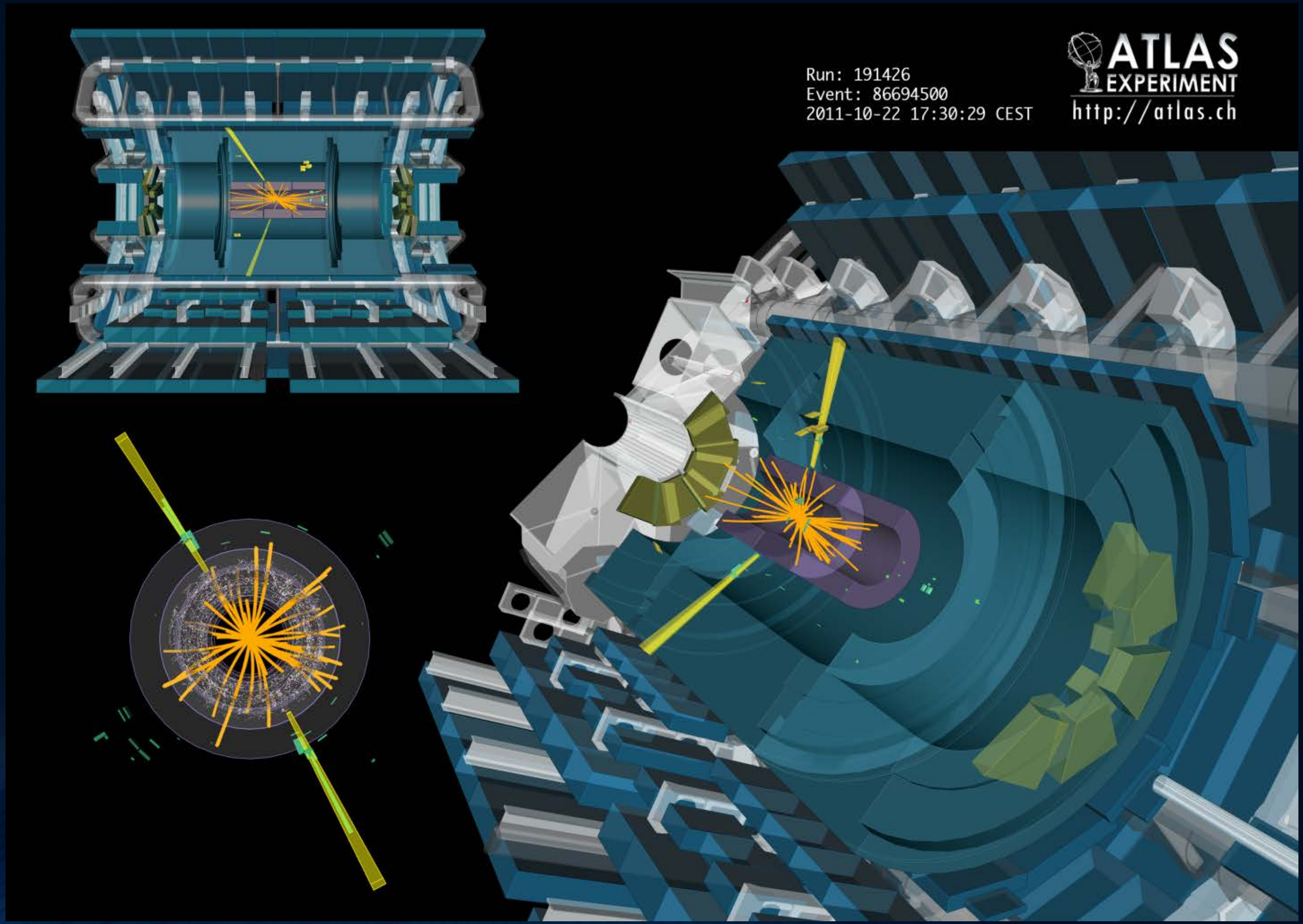
- Physicists measure mass (~weight) in units of energy. 1 proton weighs 1 GeV
- Why are the 3 families of quarks and leptons?
- Why are the masses of these elementary particles so different? How do they acquire their mass?
  - We have the beginning of the answer – the Higgs boson
- What lies beyond the Standard Model?



What Data is ATLAS Recording?

- Basically the ATLAS detector is a large 100 megapixel ‘digital’ camera that takes 40 million electronic ‘pictures’ of the colliding proton bunches per second (with 100 billion protons in each bunch) – below is a picture leading to the production of two quarks (‘jets’)
- Every 25 nano-seconds the bunches collide there can be (eventually) up to 200 p-p collisions

“Picture” of a Higgs Boson Candidate

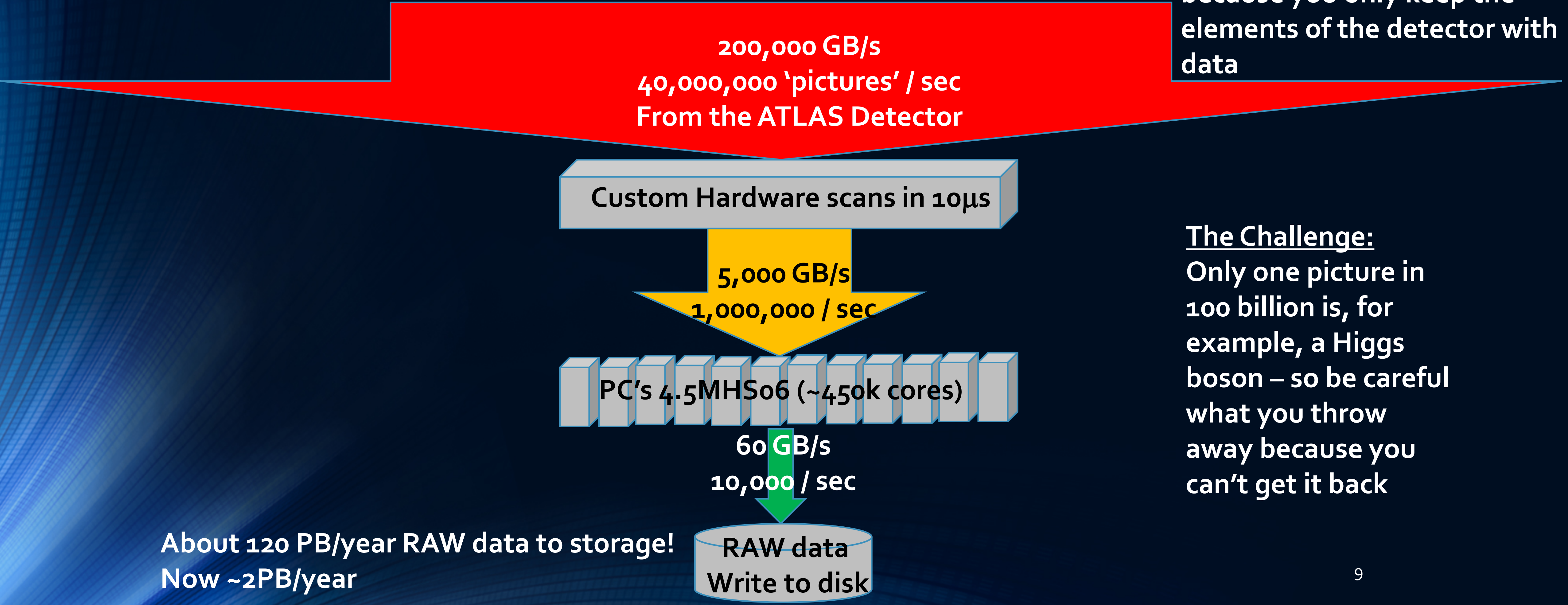


Current and Future Plans

- The ATLAS detector has been collecting data since 2010
- The LHC accelerator is producing ever more proton-proton collisions
- We are planning a major upgrade of our ATLAS detector to be ready in ~2026, when the LHC will produce ~x10 more collisions
- What I will show next is a mix of what we plan for that upgrade and how we are performing today…

Get Rid of the Junk but Keep the Interesting Pictures!
"Trigger & Data Acquisition" – HL-LHC upgrade

Each 'picture' is ~5MB in size because you only keep the elements of the detector with data

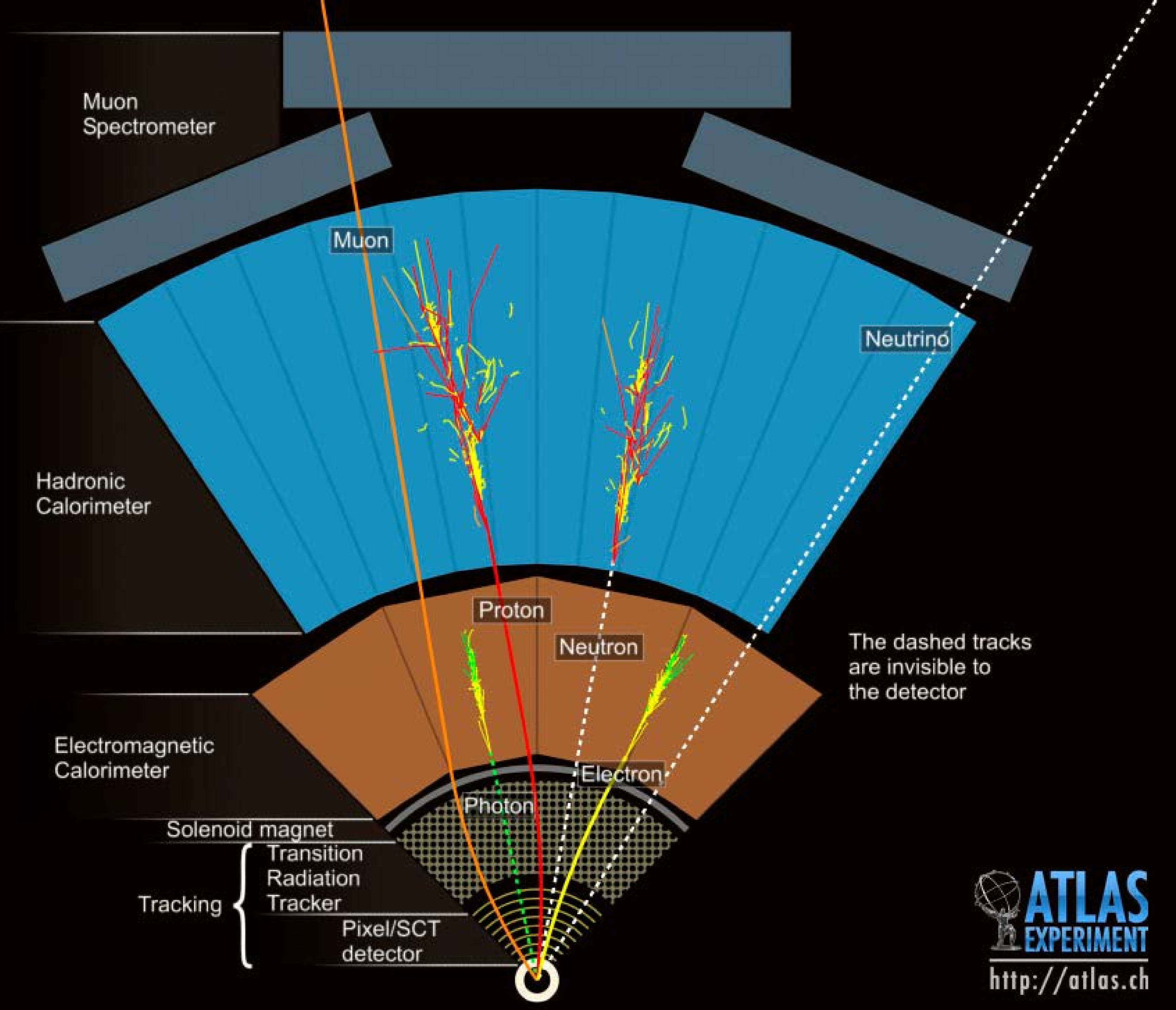


The Challenge:
Only one picture in 100 billion is, for example, a Higgs boson – so be careful what you throw away because you can't get it back

About 120 PB/year RAW data to storage!
Now ~2PB/year

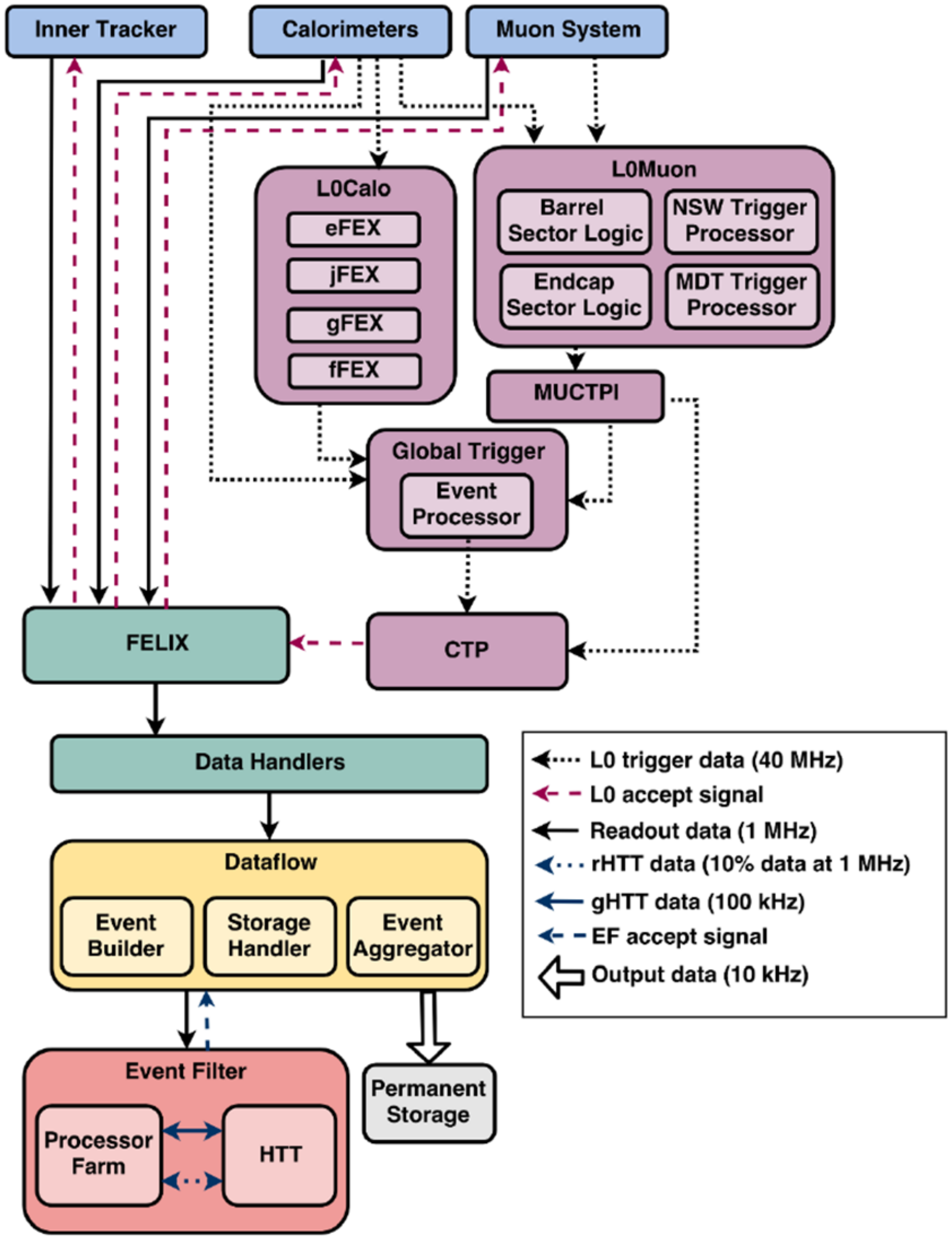
The ATLAS Subdetectors are Specialized

- Each subdetector is specialized to detect the particles produced in the collisions – recall the chart of elementary particles such as photons, electrons, muons,…
- Three main subdetectors
  - MUON SYSTEM - muons
  - CALORIMETERS – electrons & photons
  - INNER TRACKER – charged particles



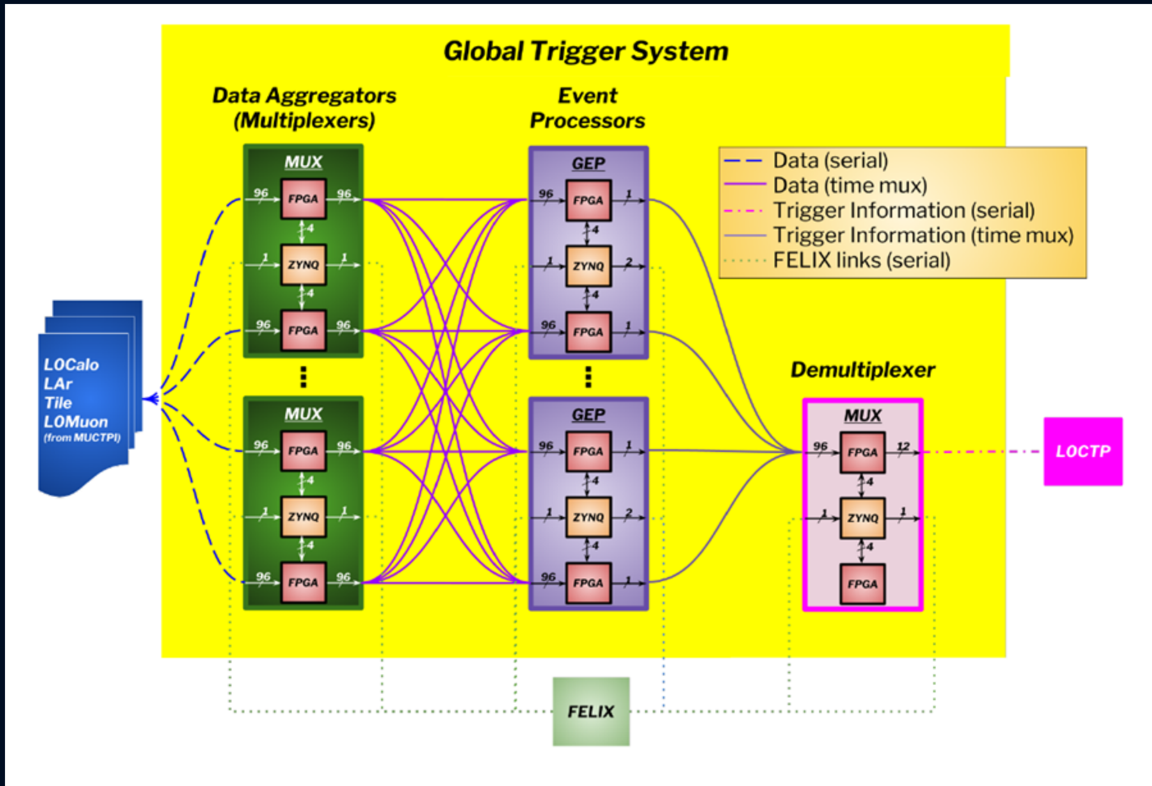
Selecting and Handling the Data

- We know when the collisions occur, so we can look at the 'picture' (we call them 'events') that occurs at that time, each is composed of data recorded by specialized sub-detectors, as it passes through 3 stages:
- 'LO TRIGGER SYSTEM': We cannot afford to record all pictures, so we must select those of interest, and we must do it quickly using custom hardware (FPGAs, ASICs,...) to select pictures of interest up to 1M/sec
- 'DATA ACQUISITION SYSTEM' (DAQ): Once we know what pictures we want we use custom hardware to format the sub-detector data for use by commodity PC servers and standard networking to fully read out (READOUT) and pass that data on (DATAFLOW) for further fast analysis
- 'EVENT FILTER' (EF): Takes that formatted data into a processor farm and together with a hardware 'track trigger' based on more sophisticated algorithms makes a decision on which pictures to record to permanent storage at a rate of 10k/sec



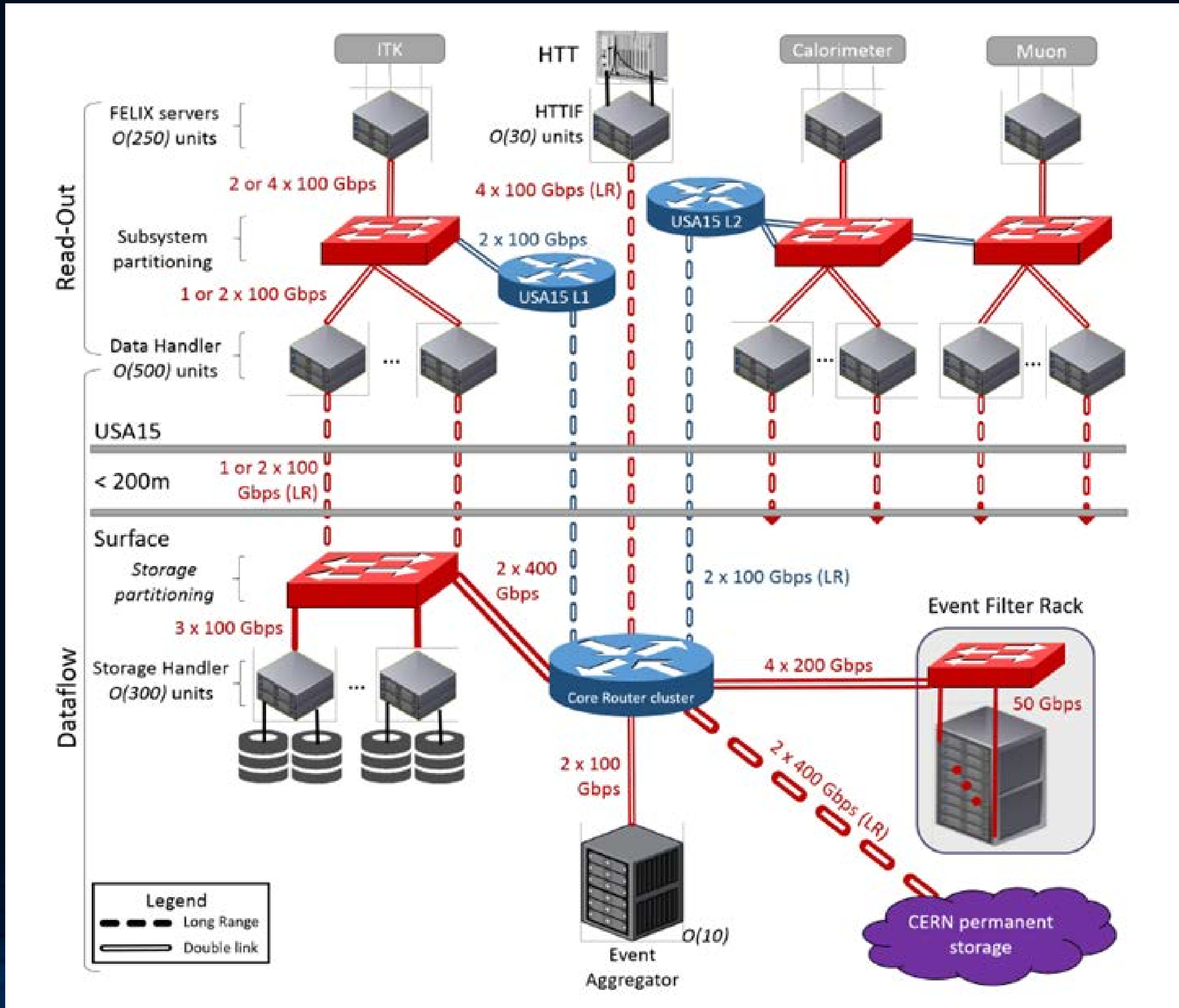
A Detail of the Custom Hardware

- This is a detail of the Global Trigger
- Takes information from the L0 trigger from the variuops subsystems, multiplexes that information to a system of Global Event Processors that receive complete data for the event to identify ‘physics objects’ (electrons, taus, jets, muons, missing transverse energy) – the work is in the firmware!
- This is a hardware system that is largely built around FPGAs (those are the cost drivers)



How do you Move the Data Around?

- High throughput network needed
- Still will need to make choices of the network infrastructure (Ethernet, Infiniband, OmniPath)
- Shown in the diagram is based on an Ethernet choice



But Not Done Yet! Preparing the RAW Data

- RAW data has to be turned into data that can be used for physics analysis – move from online to offline
  - Reconstruct the 100M locations and values in the detector to positions and energies
    - Requires calibration
  - Convert that to ‘objects’ (electrons, muons, jets,…) that can be studied by physicists

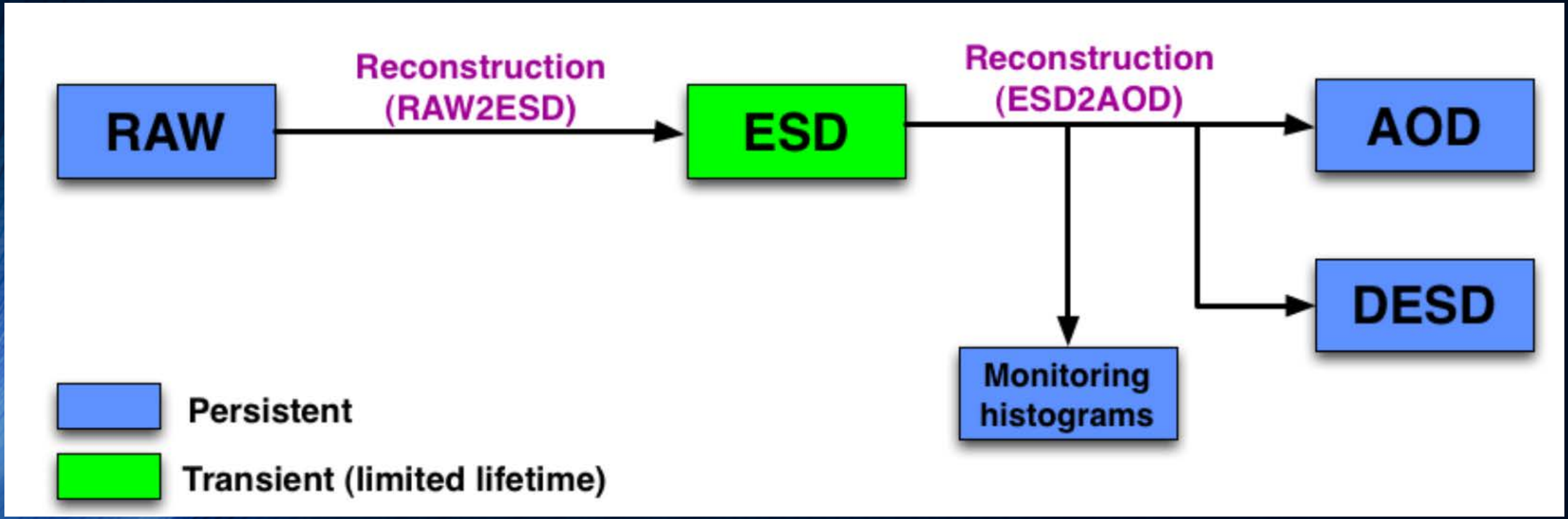
Many Different Types Data Sets to be Saved

- RAW – the raw data coming from the detector
- ESD (Event Summary Data) – create from RAW a detailed output of the detector reconstruction with sufficient information to allow particle identification, track fitting, jet calibration etc
- AOD (Analysis Object Data) – create from ESD a summary of the reconstructed event which contains sufficient information for common physics analyses
- DAOD (Derived Analysis Object Data) – create from AOD by selecting a subset for specific analyses
- And others for the simulation data: EVNT (from the physics event generators); HITS (simulate the hits in the detector); RDO (simulated detector output); AOD…

So you Save a lot More to Disk than the RAW data

File Type	'Picture' Size (kB)	Disk/Tape Versions saved
RAW	1,000	1 / 2
ESD	2,700	0.2 / 0
AOD	350	2 / 1

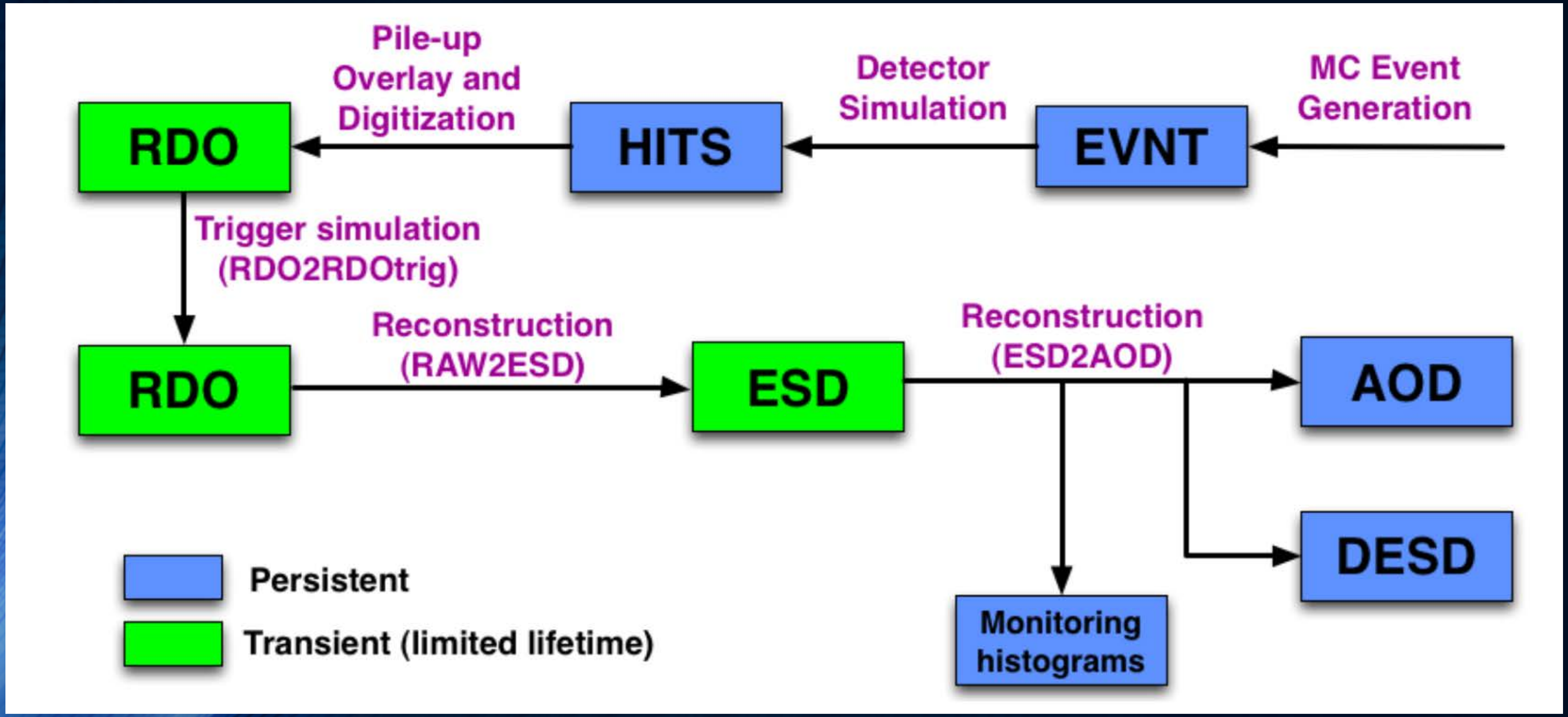
- The processed data follows this path



Physicists access the data at this level (AOD, DPD), ideally user transparent in terms of the location of resources (computing and data)

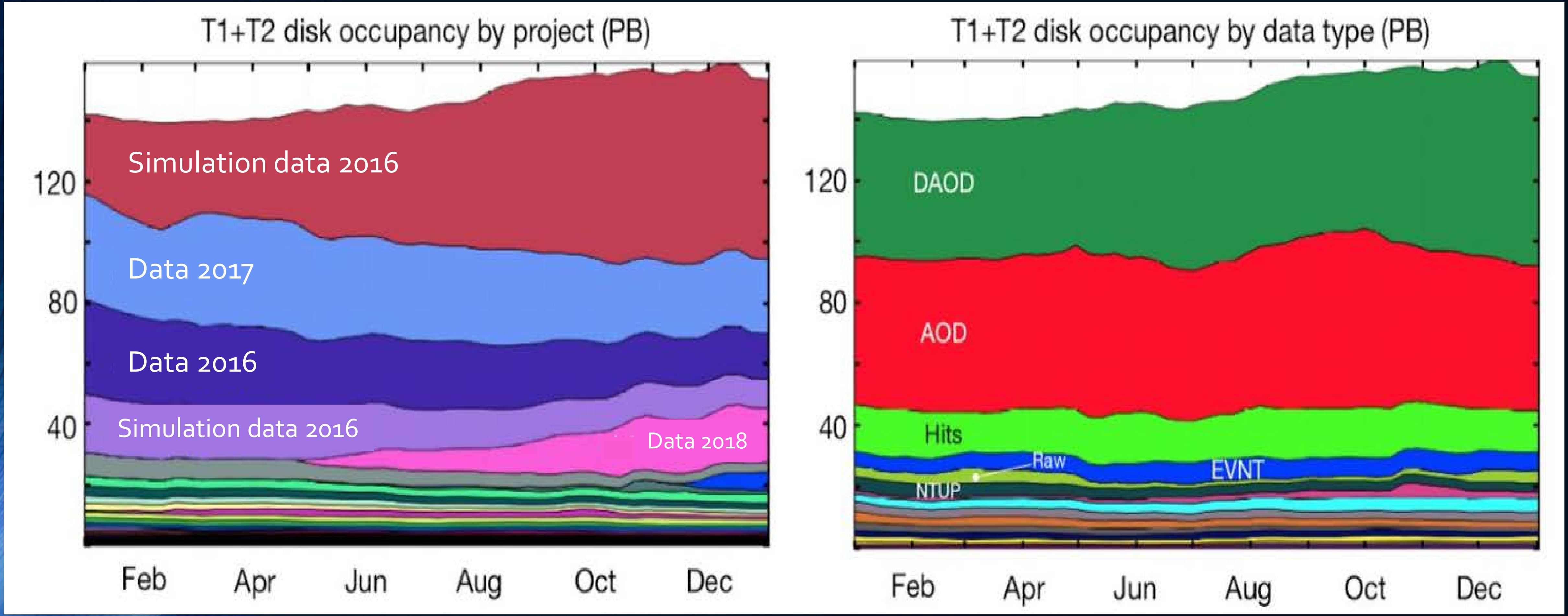
And do Something Similar for the Simulated Data ("Monte Carlo")

- The simulated data follows this path



File Type	'Picture' Size (kB)	Disk/tape Versions saved
EVNT	100	2 / 1
HITS	1,000	0 / 1
RDO	3,700	0.05 / 0
ESD	3,700	0.05 / 0
AOD	550	2 / 1

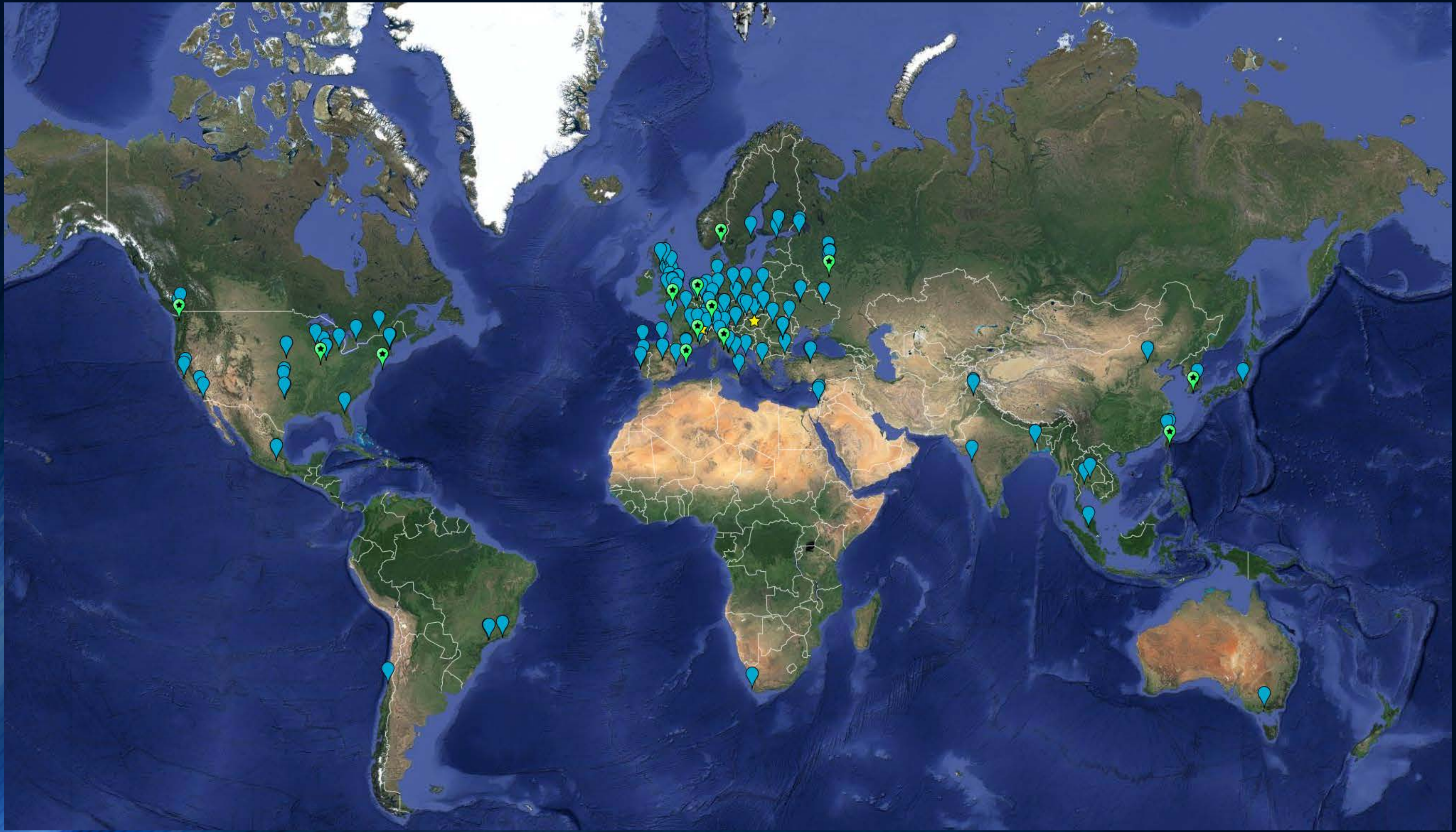
Leads to Large Amounts of ATLAS Data on Disk (~ 160 PB in 2018)



It is Big Data

- Let's look at 2016 ATLAS disk space
  - 200PB on disk + tape
  - 100,000 data sets
  - 1 billion files
  - Traffic between sites averages 20GB/s

Where do you put all the data and analyze it? Worldwide LHC Computing Grid (WLCG)



For the four LHC experiments combined

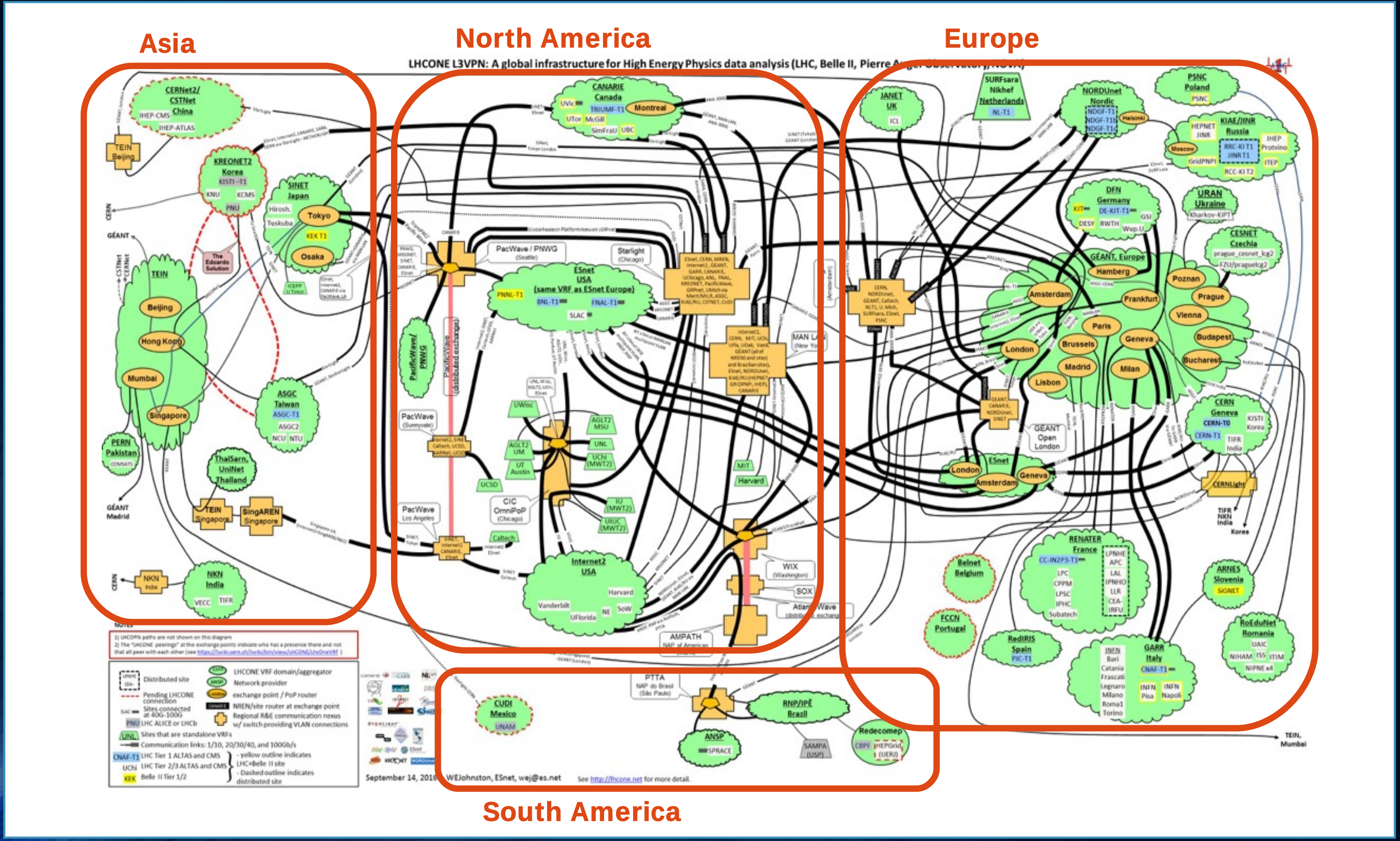
Tiered computing system – T0 (CERN), T1 (National), T2 (regional), T3 (local)

167 Computing sites
42 countries

1,300 Petabytes of storage (pledged 530PB disk + 770PB tape)

674,000 CPU cores (9.3M HS06)

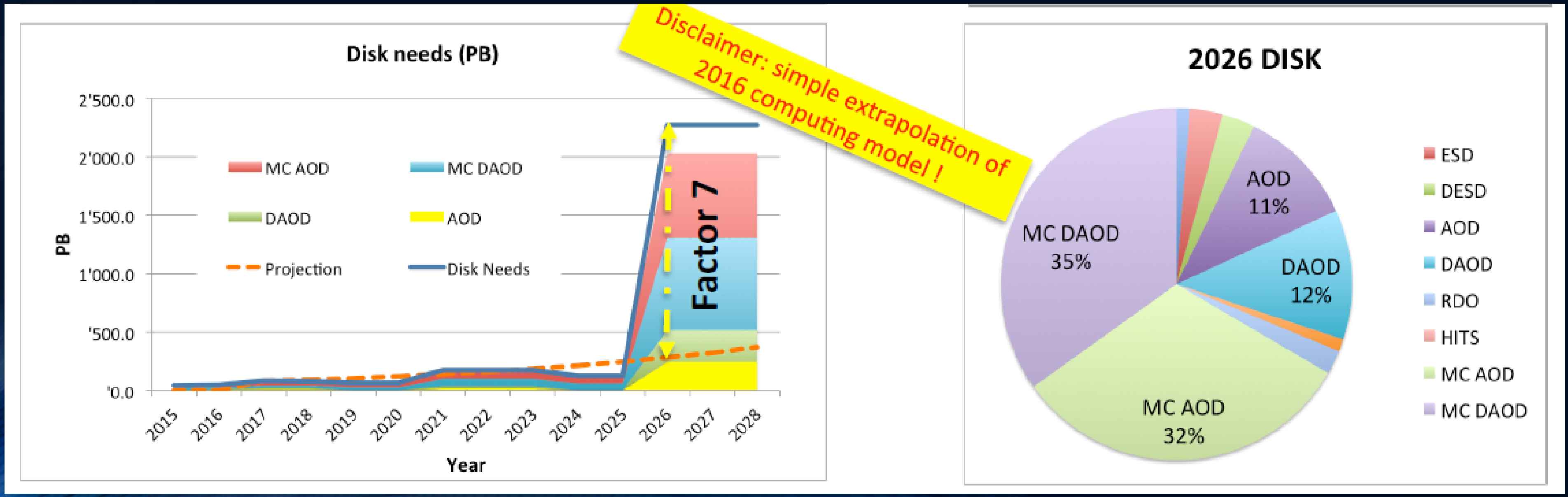
But it is Complicated! (Worldwide network)



The US piece is called the Open Science Grid (OSG) that not only services the LHC but other sciences as well

The Next Challenges

- Data Mining & Analysis – we need to get smarter
  - Use advanced techniques – deep learning etc
  - Use the CPU's we have – multithreading etc
  - Opportunistic resources – HPC, cloud,…
- We are building to the future, to upgrade the detector by 2025, to work with an improved LHC, producing x10 the number of collisions



Summary & Outlook

- The general purpose LHC Experiments (ATLAS, CMS) must select and analyze data produced in the collisions of protons on protons with high efficiency for ‘interesting’ physics events
- Need fast on-line decisions from the trigger (10μs latency) to reduce 40M/sec to 1M events/sec
- Need to further reduce that to 10,000 events/sec using sophisticated algorithms running on farm of commodity PC’s
- Must high efficiency for interesting events for the 10,000 events/sec recorded to permanent storage (disk & tape)
- Need massive off-line analysis computing grid with massive storage (thousands of petabytes)
- ATLAS has been hugely successful (many hundreds of publications, Higgs discovery), and will need to remain so when the LHC accelerator is upgraded and the detector is upgraded to keep up
- It is a challenge! We plan to meet that challenge as we upgrade the ATLAS detector for operation in 2026