Big Science in an Era of Large Datasets – BNL Experience and Perspective in Nuclear and Particle Physics Data Analysis

Michael Ernst

Brookhaven National Laboratory

CLAS12 Software Workshop

University of Richmond

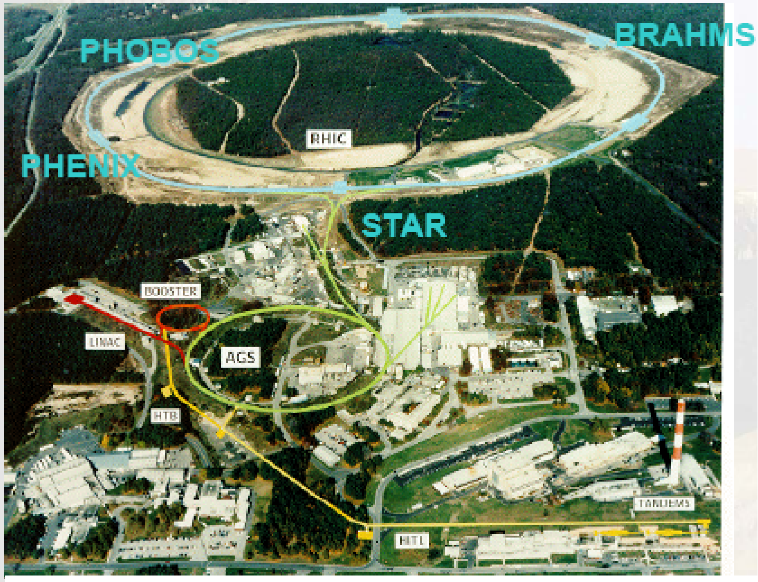
25 - 26 May 2010

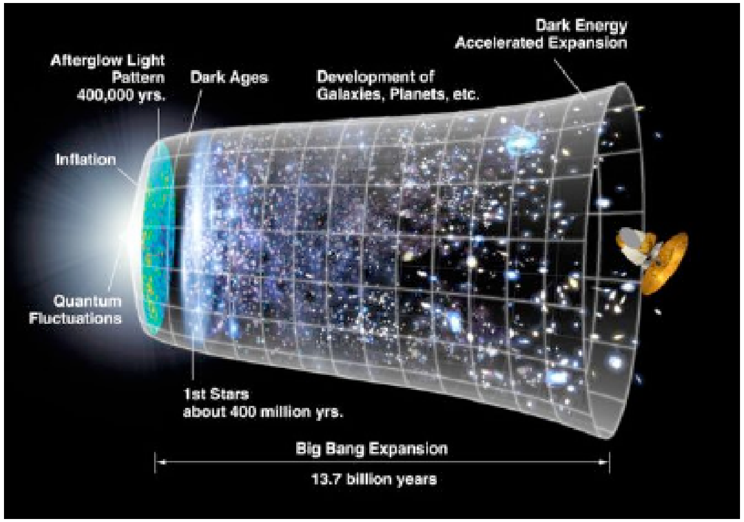
BROOKHAVEN
NATIONAL LABORATORY

The Relativistic Heavy Ion Collider (RHIC)

Scientific program in Heavy-Ion and Spin program

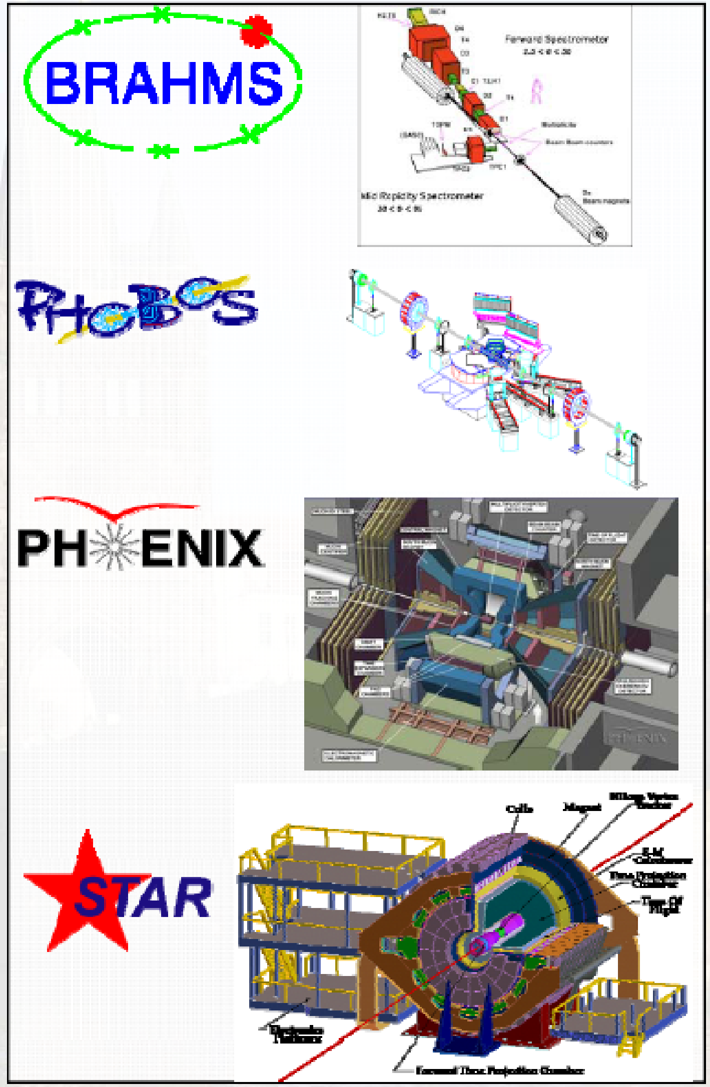
- Heavy Ion: QGP
  - provide unique insight into how quark and gluons behaved collectively at the very first moment our universe was born.
  - Critical temperature Tc ≈ 2.10^{12} K
    - The sun core is ~ 10^{7} K
    - Tc ⇔ 170 MeV
- Spin program
  - understanding how mass and spin combine into building blocks of nature
- **Versatile machine- Flexibility is key to understanding complicated problems**
  - *Polarized* protons sqrt(s_{NN}) = 50-500 GeV
  - Nuclei from d to Au (U), sqrt(s_{NN}) = 20-200 GeV



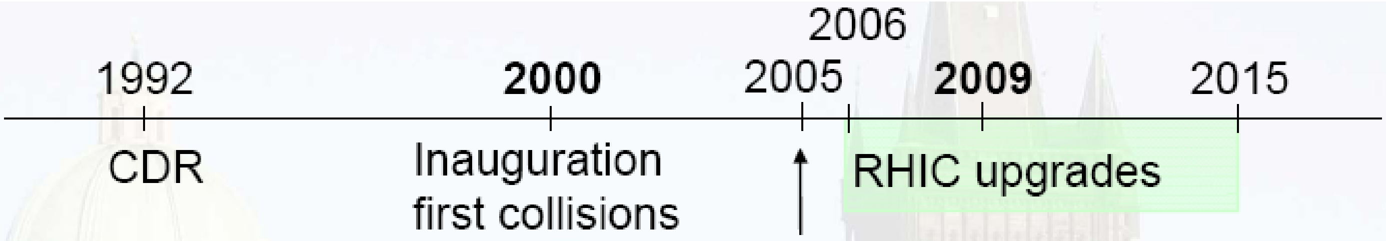


RHIC Experiments

- 4 Experiments at the start of the Program (2000)
  - BRAHMS (particle production over large rapidity range, phased out in 2006)
  - PHOBOS (4 pi multiplicity and correlations, phased out in 2006)
  - PHENIX (tracking, electromagnetic probes near – mid rapidity)
  - STAR (precision global tracking and calorimetry over large acceptance)
    - Collaboration size: 500+ Participants
    - Distribution: 50+ Institutions

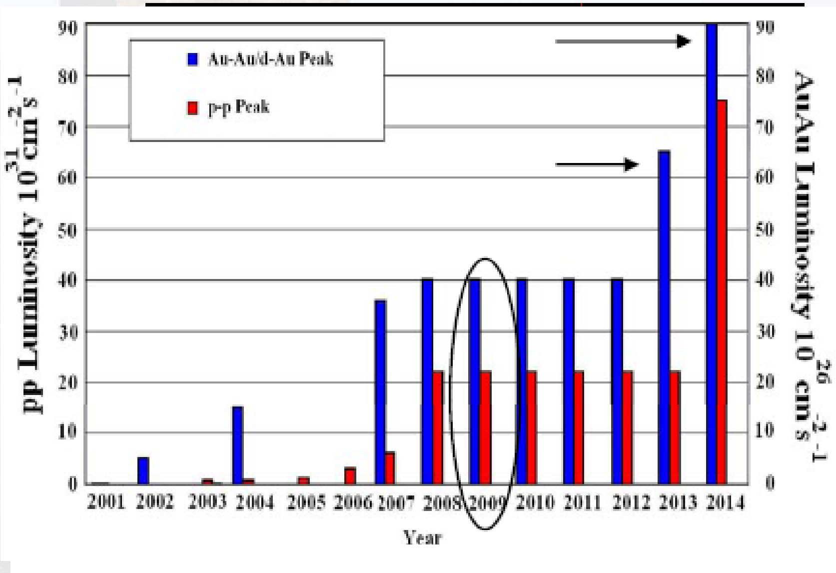


RHIC Status and Plans

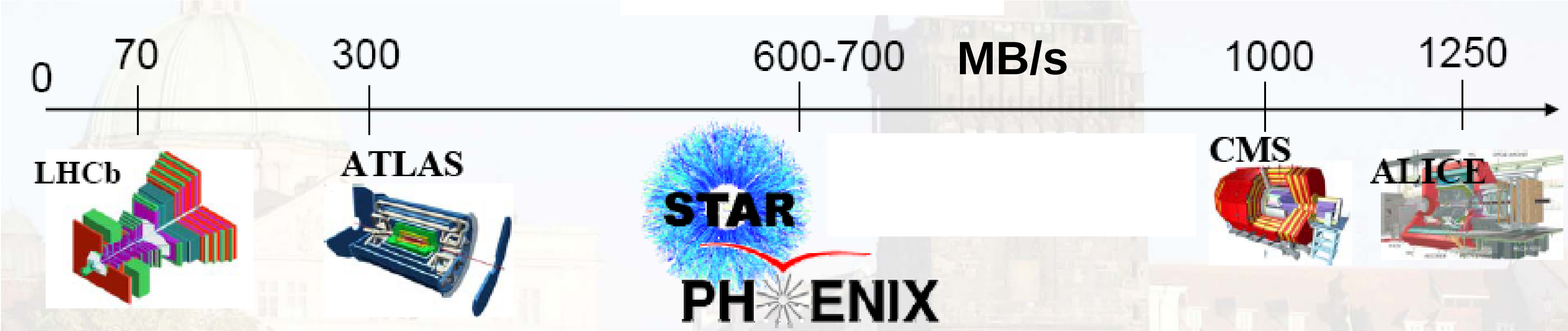


PHENIX and STAR

- Detector upgrades to address the more complex Physics
  - Heavy Flavor, Silicon Vertex, Reaction plane, forward Physics
- Machine / luminosity upgrade
  - First installment in 2007
- DAQ upgrades:
  - Early high rate for PHENIX
  - Staged DAQ upgrade for STAR (x100 in 2004, x1000 in 2008)

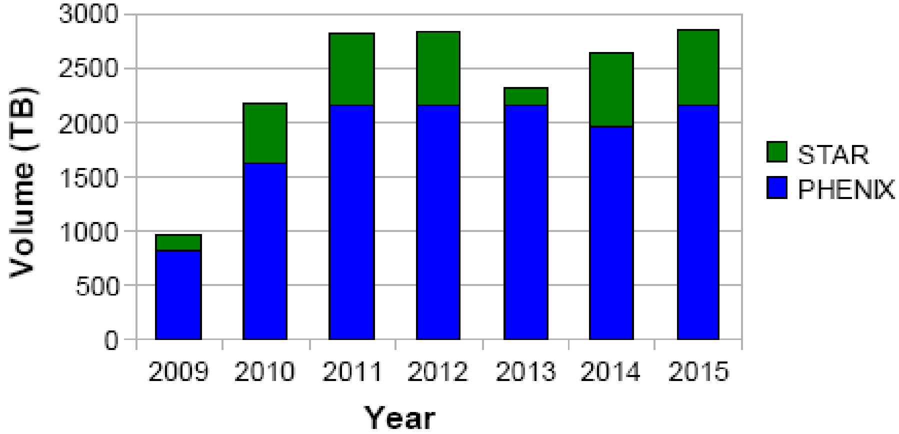


PHENIX and STAR DAQ Rates



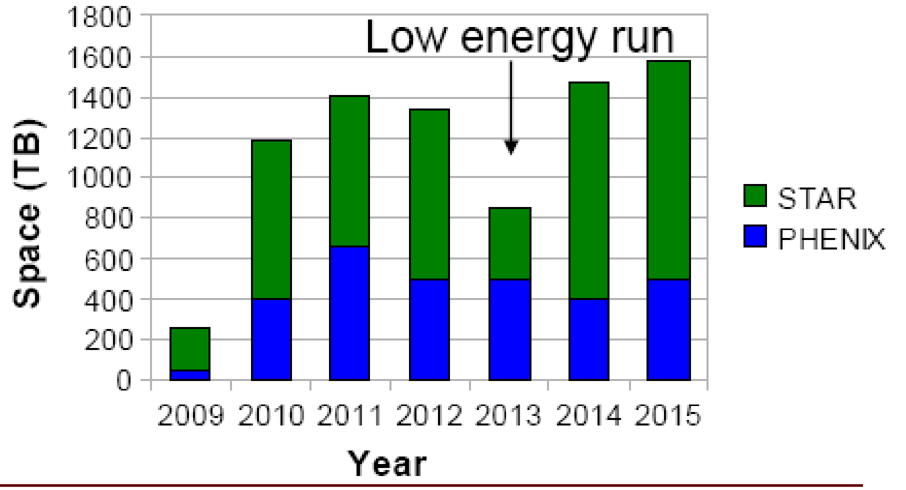
- **Dataset size**
  - **RHIC (from p+p to Au+Au) within LHC range (p+p or Pb+Pb)**
  - **PHENIX and STAR ~1 PB RAW dataset each in Run10 (2010)**
  - **RHIC will double Luminosity in ~2014**

Actual and Expected Data Volume Growth



- Initial model:
  - fraction of data from previous years on disk and/or analyzed?
  - **WRONG!**
- RHIC Experience:
  - **nearly all data from all years are being constantly analyzed, cross-compared, merged (analysis)**
- ~ ½ of the cost in storage

Derived data volume (TB)



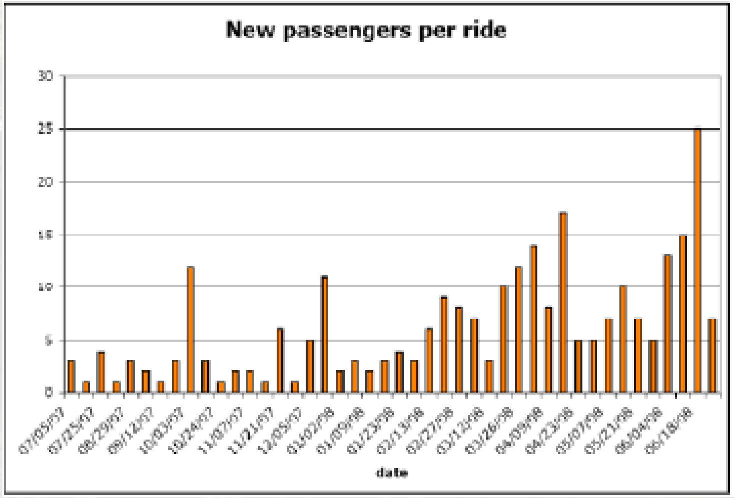
PHENIX & STAR Software Frameworks

- Decided from early on to use the ROOT Framework
  - Addresses all basic needs – histograms, NTuple, IO, versioning, framework, visualization
  - Foundation for Experiment-specific part – OO throughout
    - PhAT (Phobos Analysis Toolkit), PHool (PHenix OO Library), BRAT) BrahmsAnalysis Tool), root4star
  - Frameworks found save time and development effort
    - Analysis, calibration, data production, and eventually simulation
  - Focusing on Freeware & Open Source packages
    - The Objectivity DB lesson

Almost no change in frameworks for several years

Analysis – Chains, Trains, Taxis

- General idea: group analysis together
  - Run once over the data, multiple analysis done in one pass
  - N users, N reads => N users, 1 read
- Initial starts
  - Early start in STAR (day 1 design) – un-maintainable after 2 years
  - PHOBOS model based on a few users & Proof data access – some success
- PHENIX: Analysis trains, best success after initial tuning
  - Spin over pre-staged and pinned partial data, replace data, go-to next sample, re-launch the train. Datasets of interest covered in ~ 3 weeks
  - Larger cache added in 2006/2007 allows for a 1 week turn around

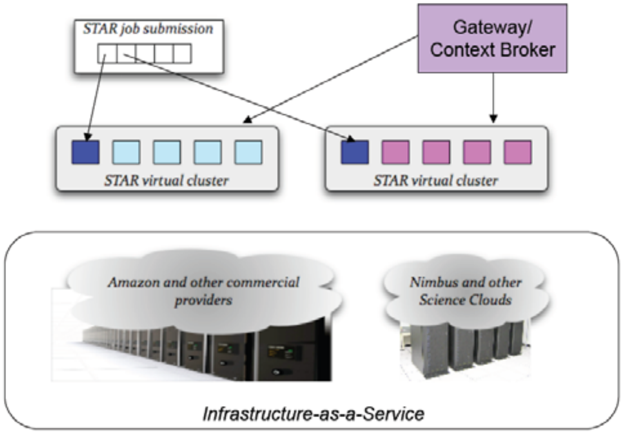


The technology of Analysis Trains has matured over time and so has the community. We expect an expansion with increasing resource demands

Extending the (centralized) Model

- Use of Grids by RHIC Experiments
  - STAR is using Open Science Grid (OSG) for MC production
    - Observe >97% efficiency
  - But, Grids are complex and too dynamic for production environment
  - Troubleshooting is difficult
  - Using dedicated sites with pre-installed software
    - Little to no opportunistic use

- What about Clouds?
  - STAR uses Amazon/EC2
  - Nimbus Infrastructure-as-a Service
  - Truly opportunistic, instantly available



RHIC Data – Evolution of derived Data Formats

DST = Data Summary Tables

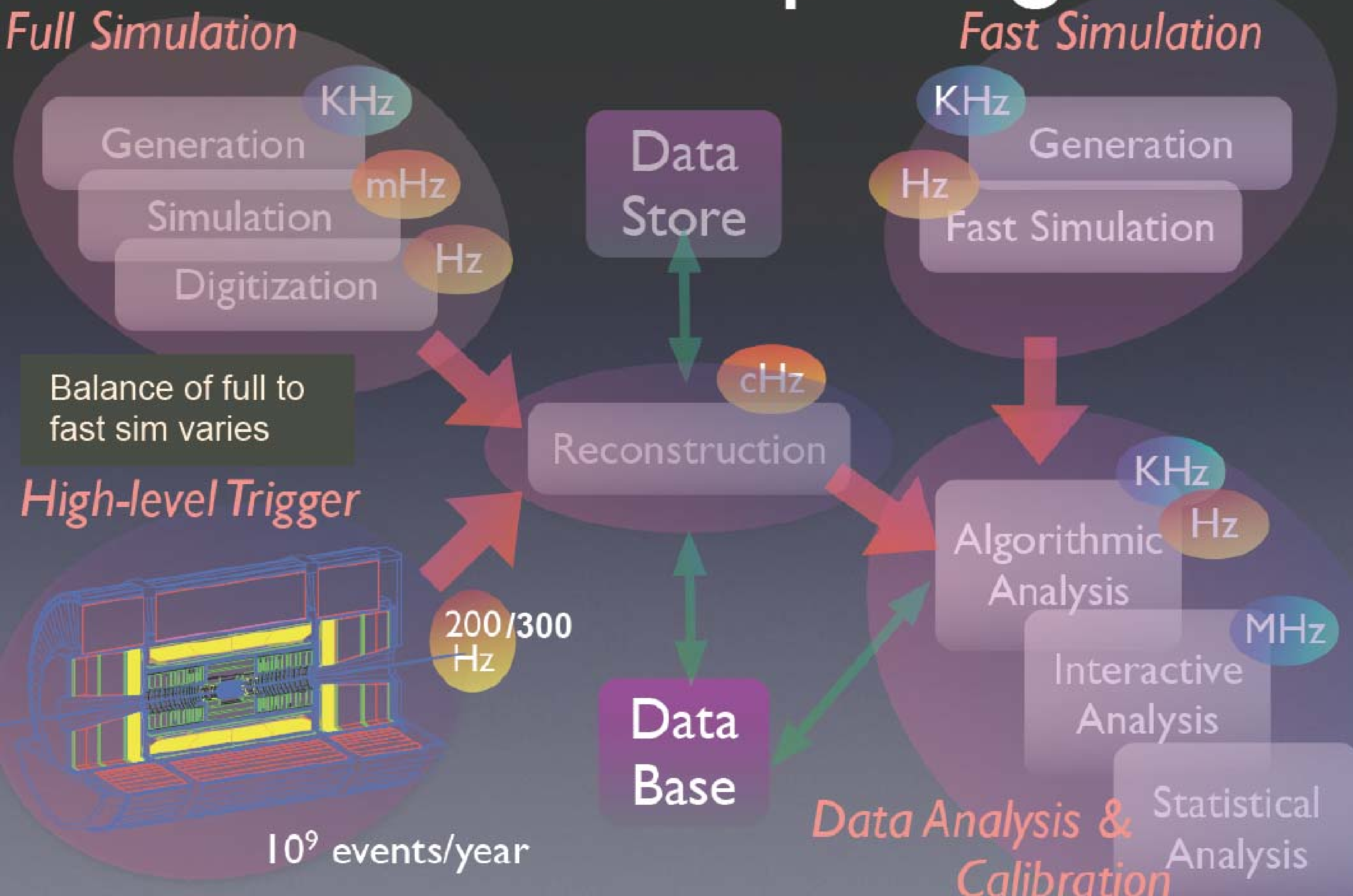
- Typical "data flow": **DAQ** $\xrightarrow[2:1]{1:1}$ **DST** $\xrightarrow[10:1]{5:1}$ **MuDST** … {pico|nano}DST,

DST, multiple formats — 2000
MuDST — 2002
MuDST on disk, picoDST — 2004
2006
Drop of DST, … — 2008

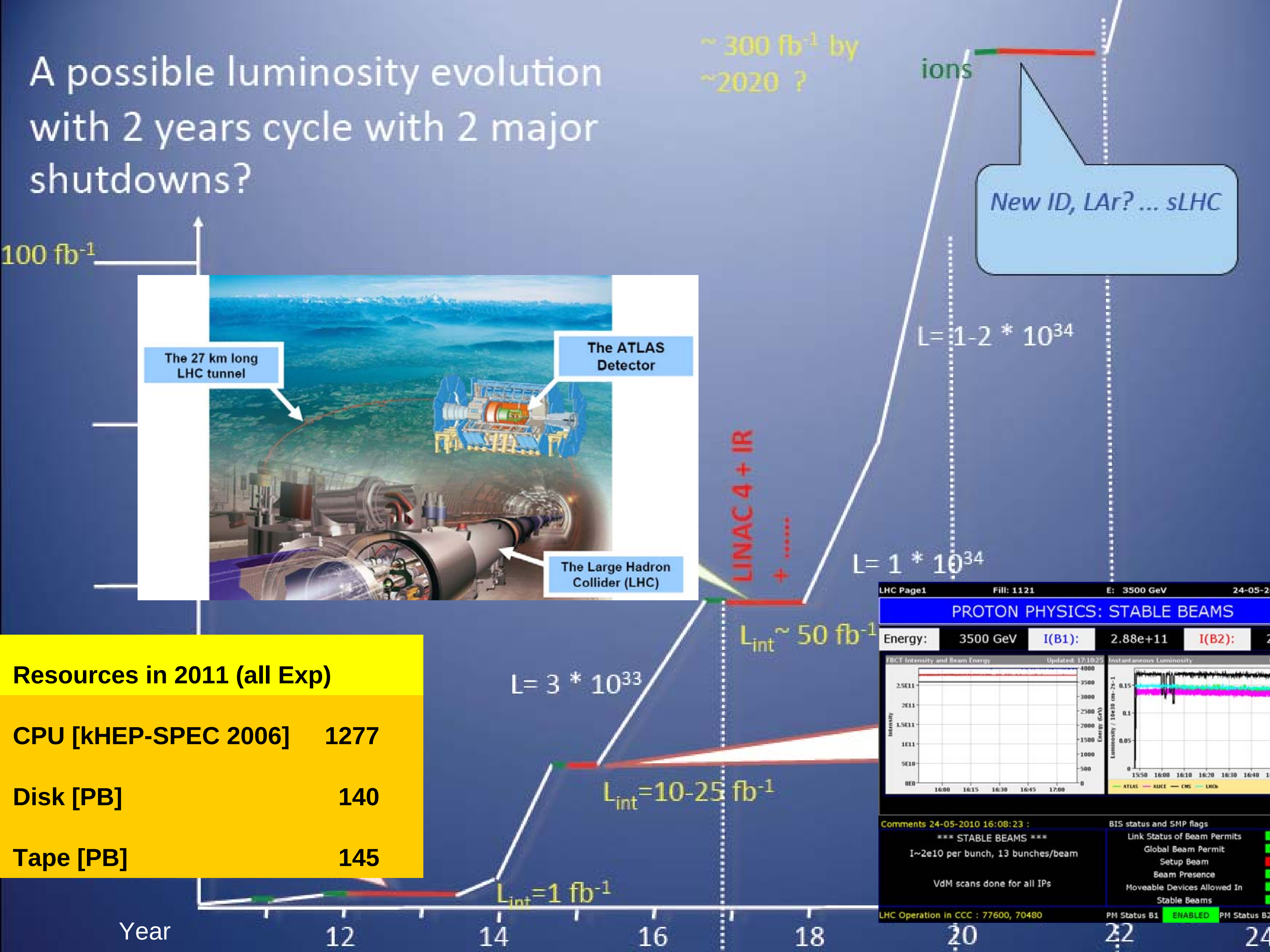
Object level access
Did not take-off – no real need? Too early?

- **New:** DST dropped after content matured (4-5 years)
  - Only a fraction kept for calibration checks purposes
  - Embedding simulation process raw signal merging

HEP Computing



A possible luminosity evolution with 2 years cycle with 2 major shutdowns?



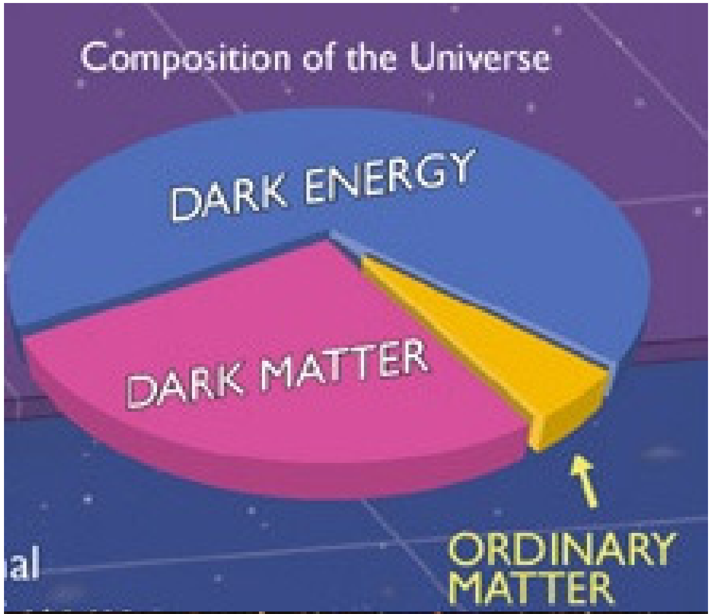
Resources in 2011 (all Exp)	
CPU [kHEP-SPEC 2006]	1277
Disk [PB]	140
Tape [PB]	145

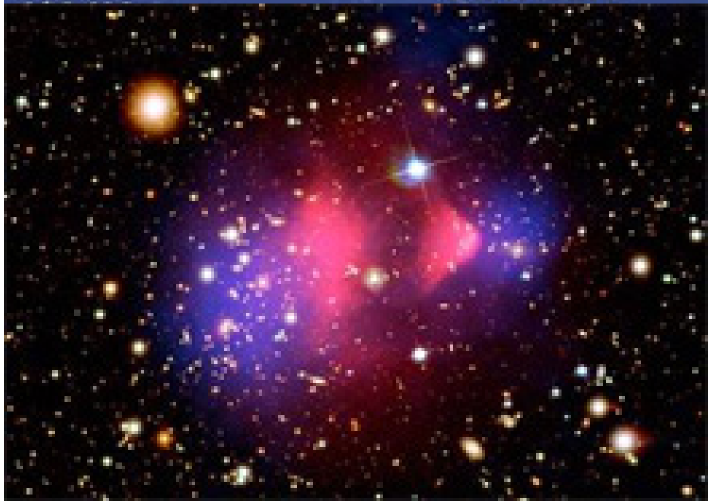
Potential Discoveries in ATLAS

- The LHC recreates, on a small scale, the conditions of the Universe just after the Big Bang in order to learn why the Universe is like it is today. ATLAS will investigate why the matter of the Universe is dominated by an unknown type of matter called dark matter.

- If the constituents of dark matter are new particles, ATLAS should discover them and elucidate the mystery of dark matter

- Evidence for dark matter can be seen in the collision of two clusters of galaxies



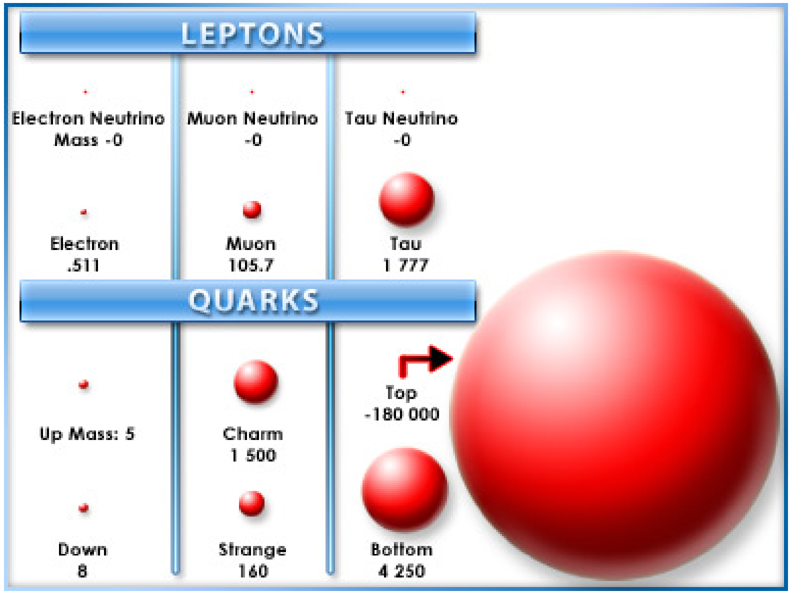


Potential Discoveries in ATLAS

➢ Why do fundamental particles have such different masses? Two of the greatest mysteries are how particles gain mass and how mass and energy are related.

➢To explain these mysteries, theories predict a new particle, the Higgs particle.

➢If this particle exists, ATLAS will discover it and provide great insight into the origin of mass.

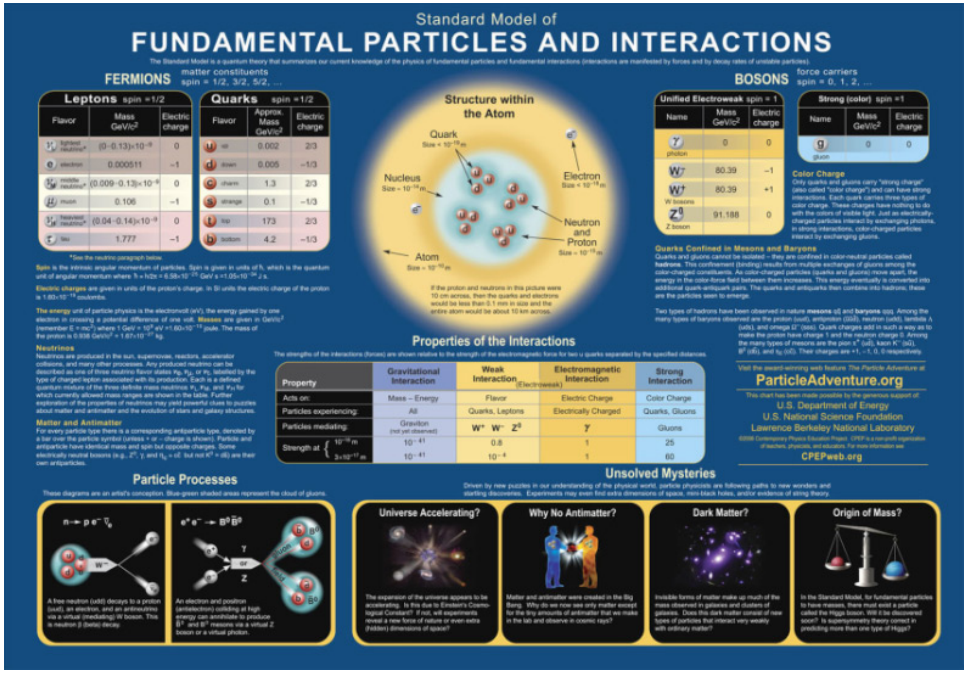


Prof. Peter Higgs Visiting the ATLAS Detector

Potential Discoveries in ATLAS

➢ Physicists have developed a theory called the **Standard Model** that explains what the world is and what holds it together. It is a theory that explains all the hundreds of particles and complex interactions with only:

- **6 Quarks**
- **6 Leptons** (such as the electron)
- **Force carrier particles** (like the photon)



➢All the known matter particles are leptons or composites made of quarks, and the interact by exchanging force carrier particles.

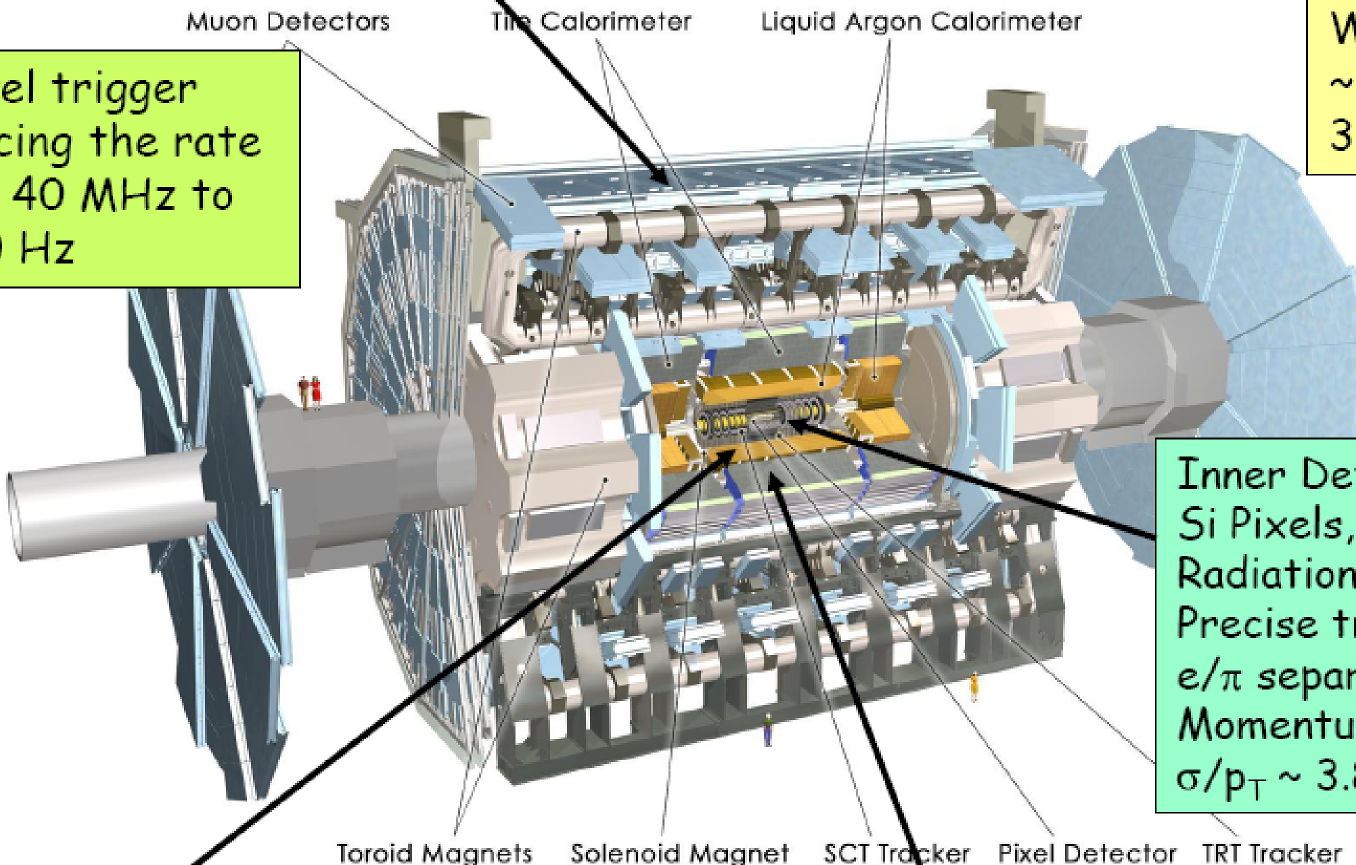
➢A primary goal of ATLAS is to look for discoveries that the Standard Model cannot explain.

Muon Spectrometer ($|\eta|<2.7$) : air-core toroids with gas-based muon chambers
Muon trigger and measurement with momentum resolution < 10% up to E_μ ~ 1 TeV

The ATLAS Detector

Length : ~ 46 m
Radius : ~ 12 m
Weight : ~ 7000 tons
~10^8 electronic channel
3000 km of cables

3-level trigger reducing the rate from 40 MHz to ~200 Hz

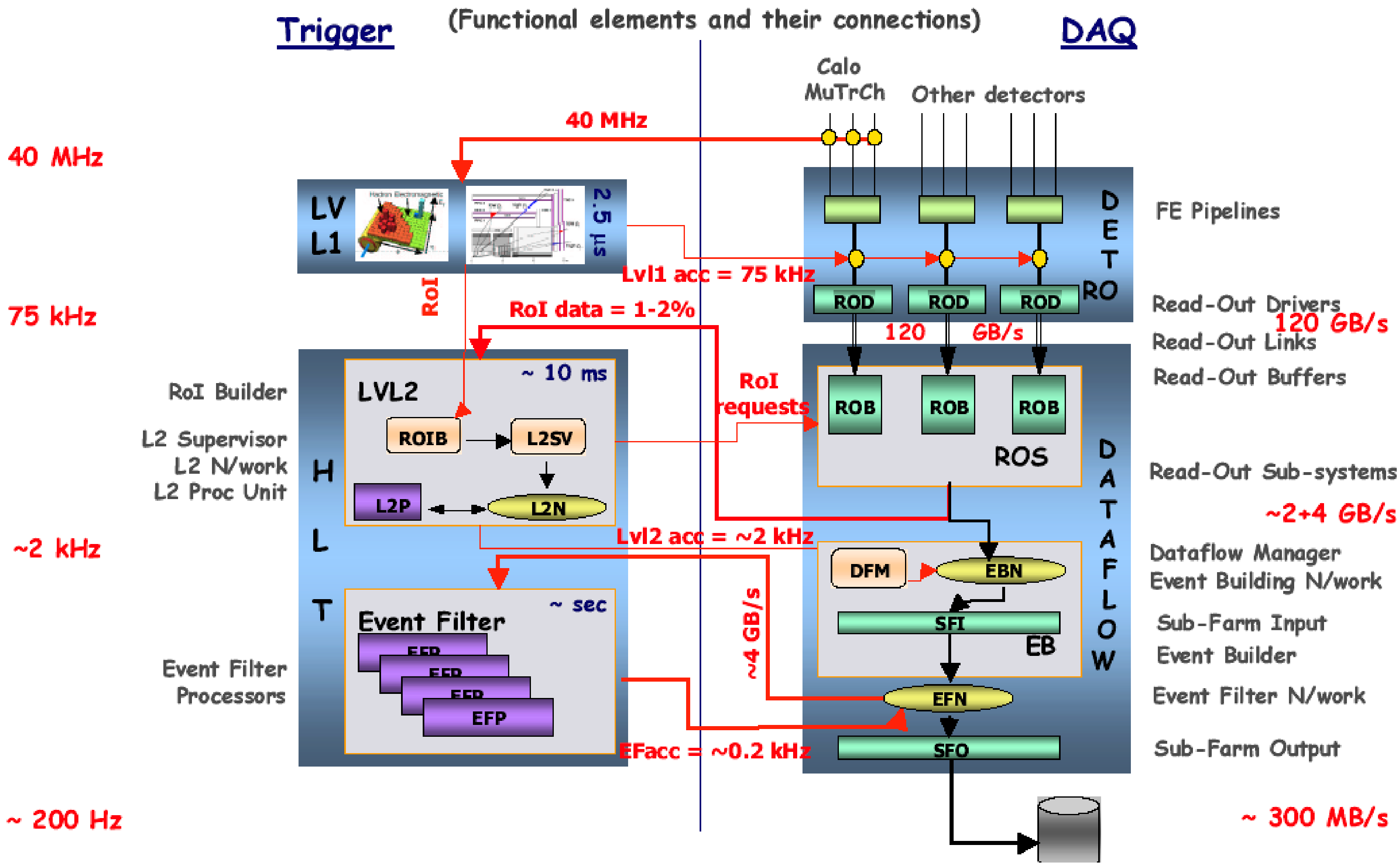


Inner Detector ($|\eta|<2.5$, B=2T):
Si Pixels, Si strips, Transition Radiation detector (straws)
Precise tracking and vertexing, e/π separation
Momentum resolution:
$\sigma/p_T \sim 3.8\times10^{-4}\, p_T(\text{GeV}) \oplus 0.015$

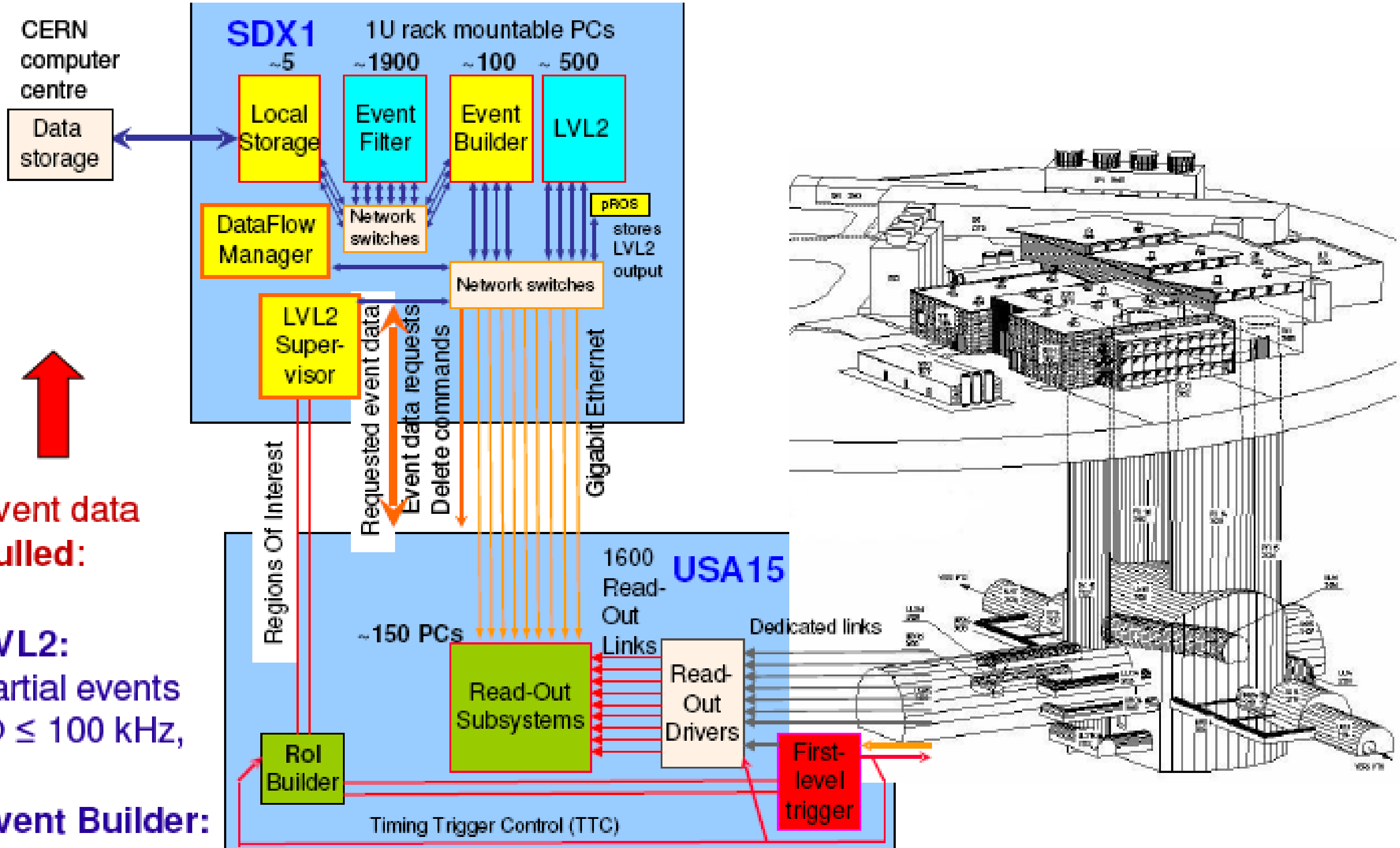
EM calorimeter: Pb-LAr Accordion
e/γ trigger, identification and measurement
E-resolution: $\sigma/E \sim 10\%/\sqrt{E}$

HAD calorimetry ($|\eta|<5$): segmentation, hermeticity
Fe/scintillator Tiles (central), Cu/W-LAr (fwd)
Trigger and measurement of jets and missing E_T
E-resolution: $\sigma/E \sim 50\%/\sqrt{E} \oplus 0.03$

ATLAS TDAQ Overview



ATLAS TDAQ Layout



Event data **pulled**:

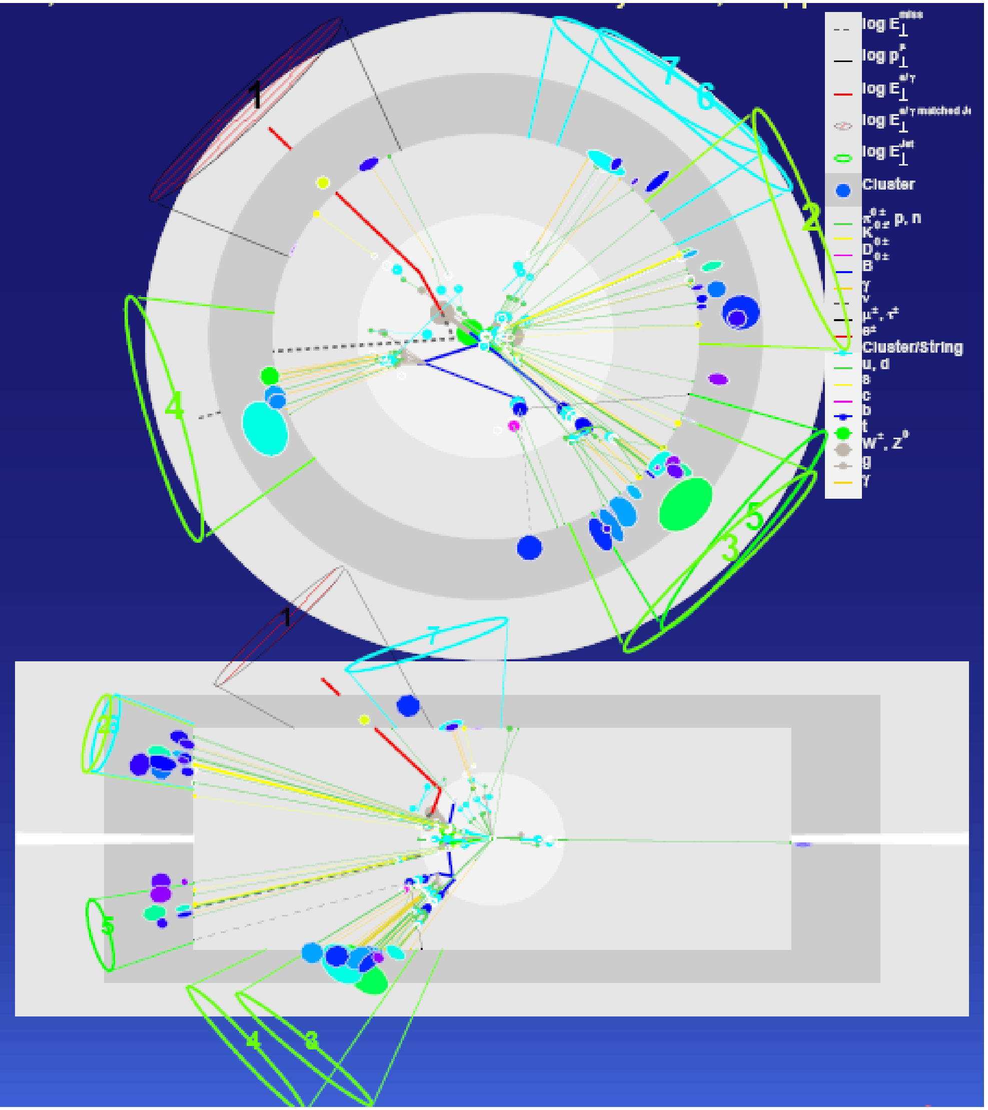
LVL2: partial events @ ≤ 100 kHz,

Event Builder: full events @ ~ 3 kHz

Event data **pushed** @ ≤ 100 kHz, 1600 fragments of ~ 1 kByte each

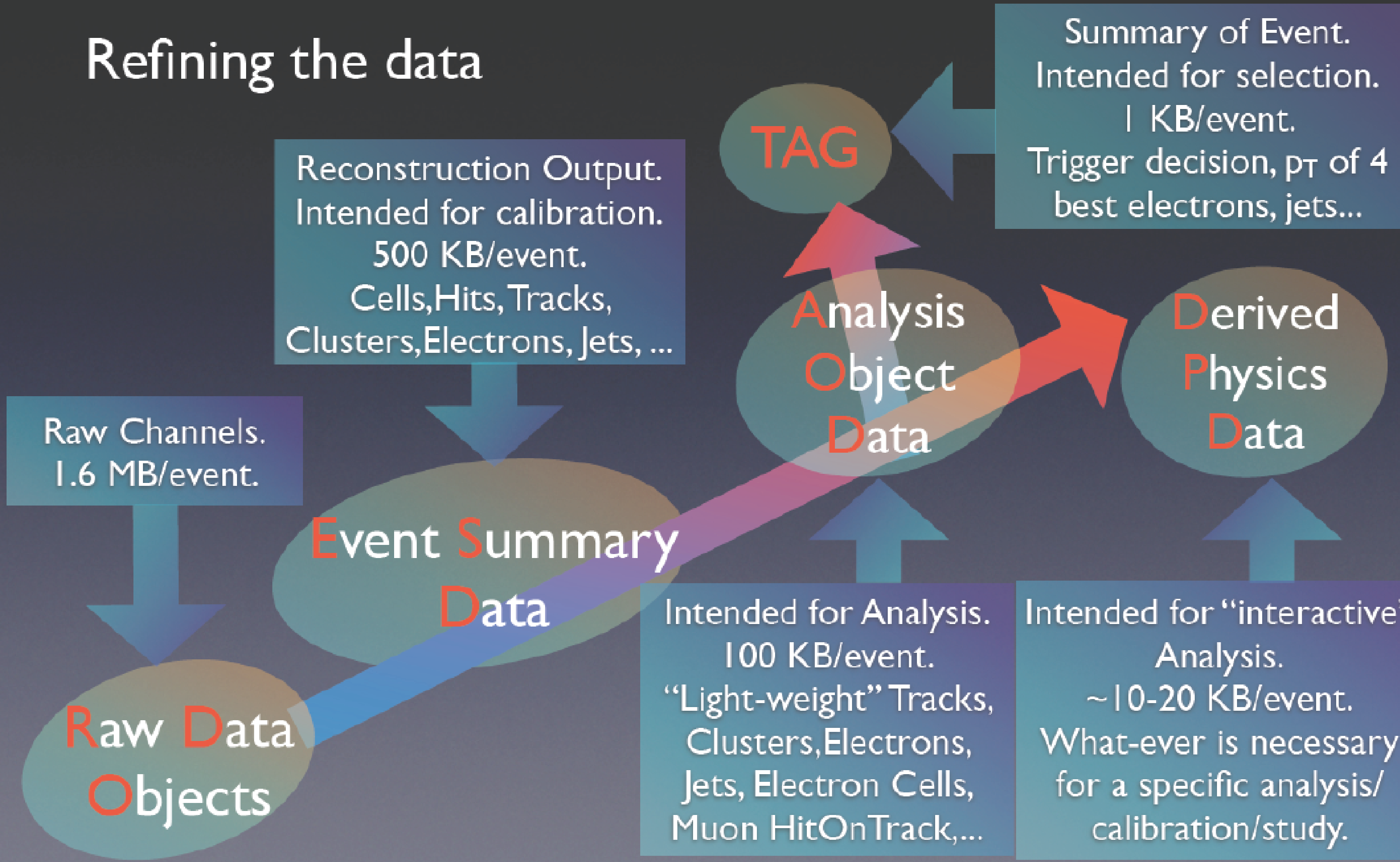
ATLAS Data and Analysis Model

- ATLAS Data formats
  - from raw to analysis data
- Derived Physics Data (DPD)
  - and tools for making them
- Metadata and TAG files
- Analysis frameworks
- `AthenaROOTAccess` and interactive `athena`
- Tools for physics analysis
- Conclusions



The Event Data Model

Refining the data

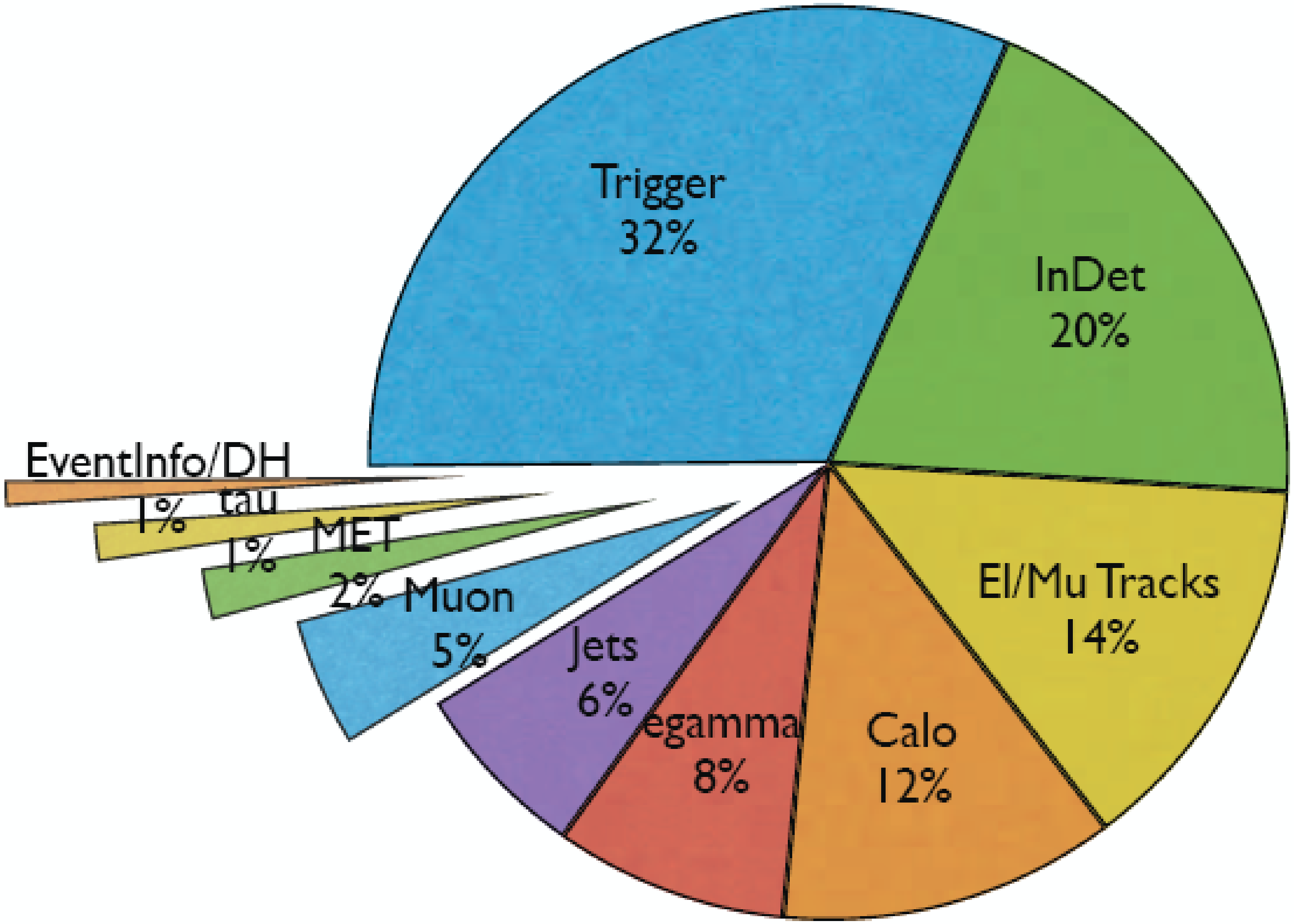


ATLAS Data Formats from RAW to Analysis

- Raw data ($\sim 200\,\mathrm{Hz}$; $1.6\,\mathrm{MB/ev}$) are written out in $4 - 7$ (RAW) physics streams based on trigger signature
- At the Tier-0 (CERN) reconstruction software runs on the raw data streams and produces
  - Event Summary Data (ESD) of about $500\,\mathrm{kB/ev}$ (target size – currently 50% more)
    - ESDs allow re-running reconstruction, particle identification, jet-finding, track re-fitting etc.
  - Analysis Object Data (AOD) of about $100\,\mathrm{kB/ev}$ (target size – currently 70% more)
    - AODs are made from ESDs and allow most common analyses
  - Derived Physics Data (DPD) of about $10\,\mathrm{kB/(AODev)}$ (assuming 10 largely disjunct primary DPD's with 10% skimming/thinning efficiency)
    - DPDs are filtered AODs with reduced information and optionally added analysis data
  - Tag Data (TAG) of about $1\,\mathrm{kB/ev}$
    - TAGs are event-level metadata allowing fast selection of events with certain signatures

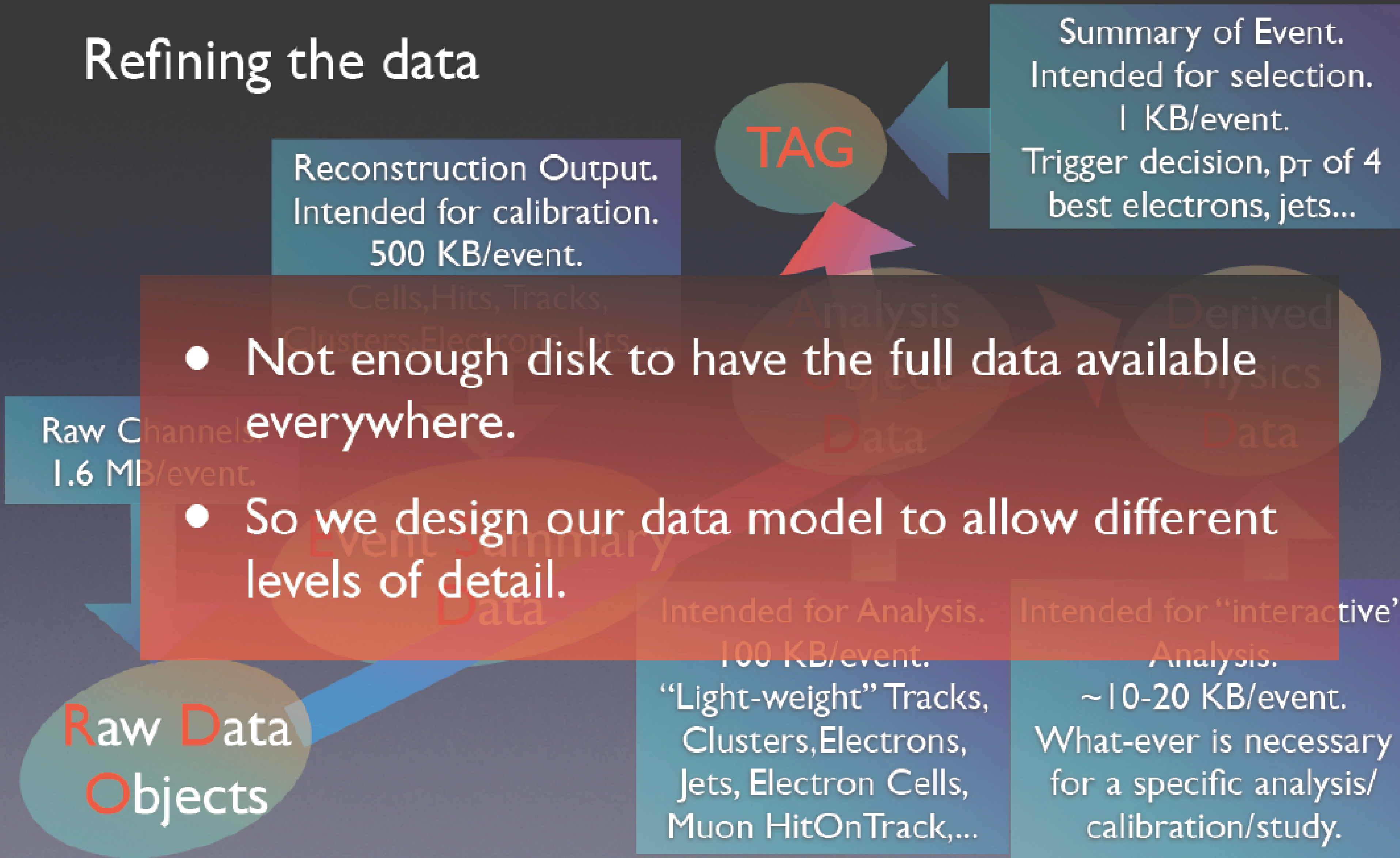
ATLAS Data Formats – Current AOD size

- Scaled Top AOD size: 167 kB/evt



The Event Data Model

Refining the data



- Not enough disk to have the full data available everywhere.
- So we design our data model to allow different levels of detail.

Analysis Activity

- *Re-reconstruction/re-calibration*- CPU intensive... often necessary.
- *Algorithmic Analysis*: Data Manipulations ESD→AOD→DPD→DPD
  - *Skimming*- Keep interesting events
  - *Thinning*- Keep interesting objects in events
  - *Slimming*- Keep interesting info in objects
  - *Reduction*- Build higher-level data which encapsulates results of algorithms
  - Basic principle: Data Optimization + CPU intensive algs → more portable input & less CPU in later stages.
- *Interactive Analysis*: Making plots/performing studies on highly reduced data.
- *Statistical Analysis*: Perform fits, produce toy Monte Carlos, calculate significance.

- *Tier 1/2 Activity*
  - Framework (ie Athena) based
  - Resource intensive
  - Large scale (lots of data)
  - Organized
  - ***Batch access only***
- *Tier 3 Activity*
  - Often exo-framework
  - ***Interactive***

Primary difference

The U.S. ATLAS Tier-1 Center in the globally Distributed ATLAS Computing Facility

- Resources Spread Around the GRID

•Interactive Analysis
•Plots, Fits, Toy MC, Studies, ...

Tier 3
BNL

DPD

•Reprocessing of full data with improved calibrations 2 months after data taking.
•Managed Tape Access: RAW, ESD
•Disk Access: AOD, fraction of ESD

30 Sites Worldwide
Tier 2

ESD
AOD
D1PD

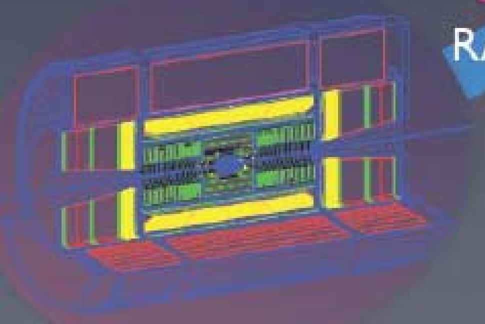
•Derive 1st pass calibrations within 24 hours.
•Reconstruct rest of the data keeping up with data taking.

Tier 1
BNL
10 Sites Worldwide

RAW/
AOD/
ESD

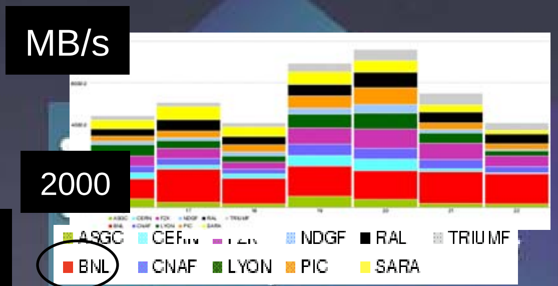
Tier 0

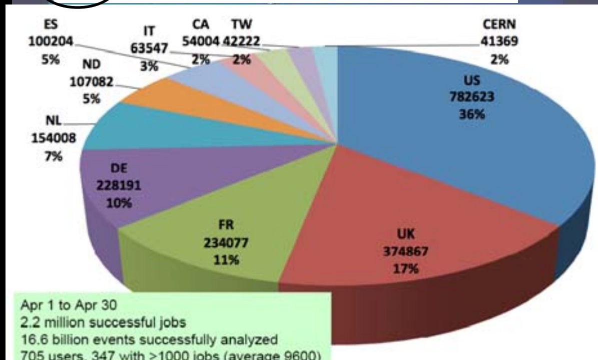
RAW



CERN Analysis Facility

The Tier-1 at BNL is the Largest LHC Analysis Center
As of May 2010
- **5500 Job Slots**
- **6.0 PB Disk**
- **9.0 PB Tape**
- **40 Gbps WAN B/W**





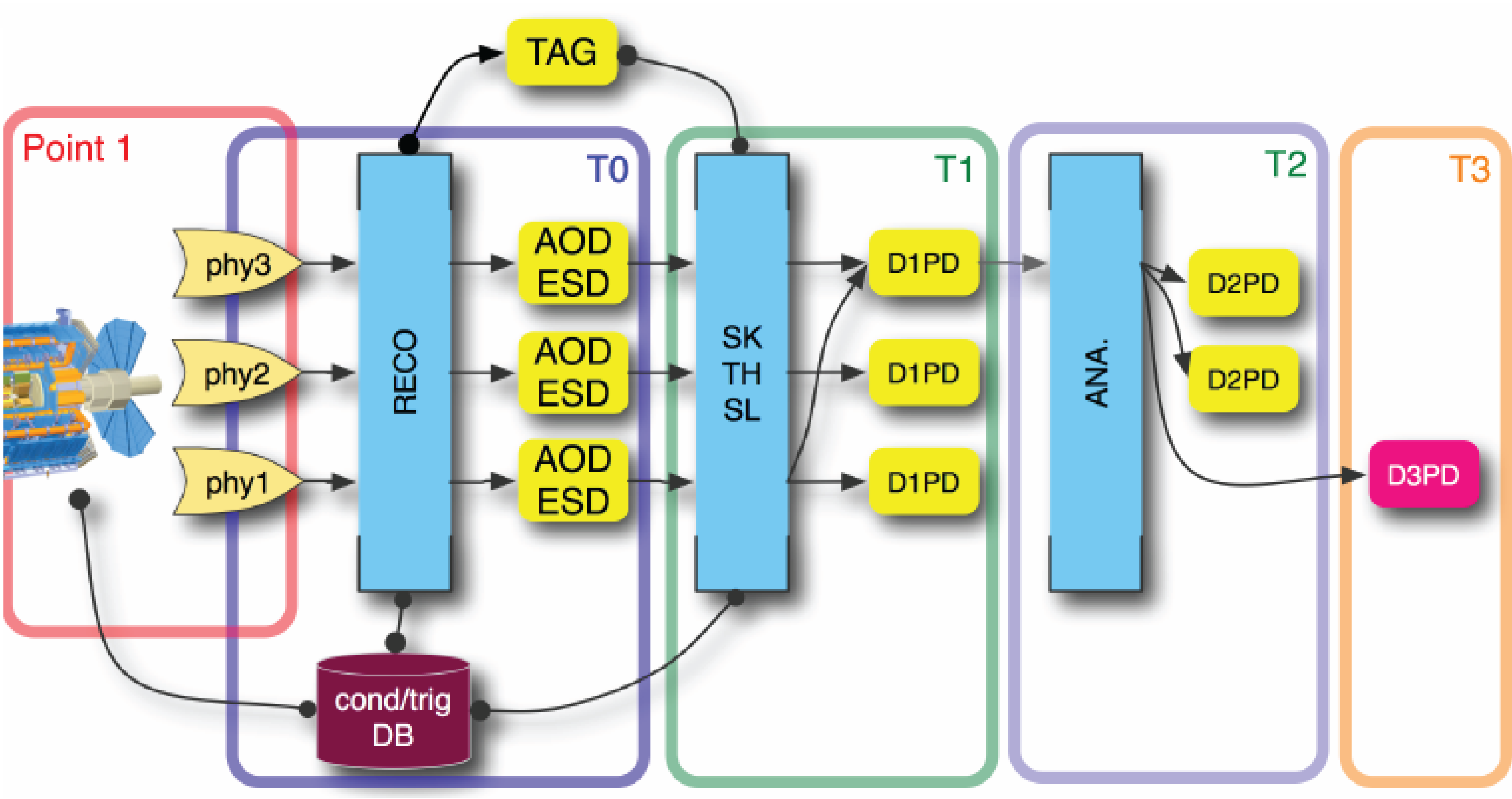
ATLAS Data Formats – from RAW to Analysis

- The Tier-0 (CERN) distributes all reconstructed data (ESD/AOD/DPD/TAG) to each of the 10 Tier-1 and on average 10% of the RAW data such that a full copy of RAW data exists
- Tier-1's distribute their data to the > 60 Tier-2's and re-reconstruct their RAW data share if needed
- Tier-2's are mainly for simulation and user-level analysis on primary DPDs to produce
  - secondary and tertiary Derived Physics Data ($D^{2/3}PD$)
    - D^2PDs are in ESD/AOD/DPD format, more refined and contain analysis data
    - D^3PDs are in flat ntuples (.root files) and are suitable for plotting final results

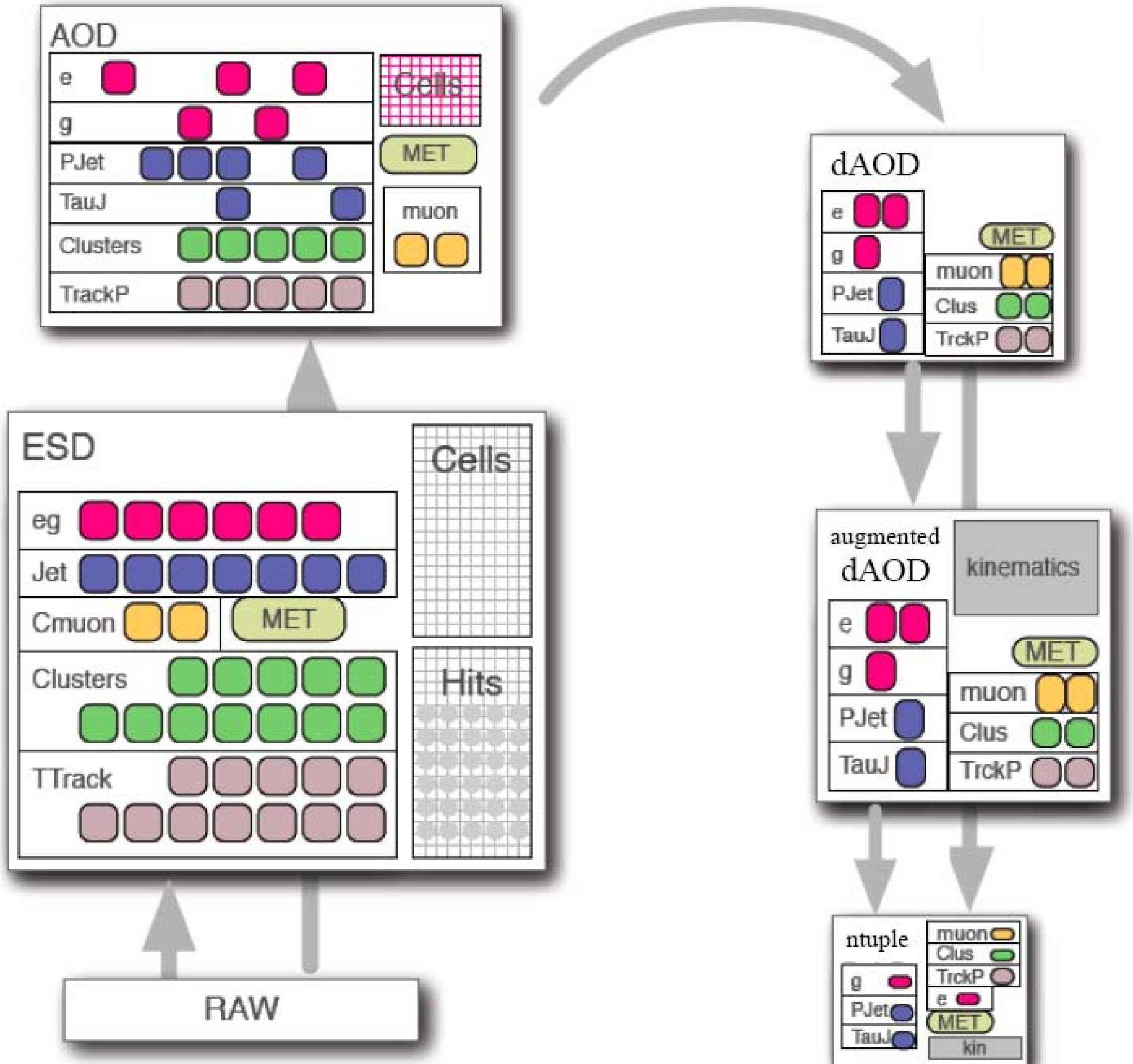
ATLAS Data – Derived Physics Data (DPD)

- `root`-access (`AthenaROOTAccess`) to AOD/ESD was made possible by separating transient and persistent representation of the data
- thus structure is provided to `root`-based analyses by the ATLAS data classes
- DPD's are consequently smaller AOD-like files where the contents is minimized by means of
  - skimming – selecting only interesting events
  - thinning – keeping only interesting objects
  - slimming – reduce the information of selected objects
- they can be re-used in the ATLAS reconstruction framework `athena`
- and be analyzed directly from `root`
- 5 – 10 primary DPDs are made centrally for each(most) physics/performance group

Analysis – From HLT to D3PD



Transforming one format to another



ATLAS Software Framework Athena

- ATLAS uses a framework called `athena` based on LHC-B's framework `Gaudi` for reconstruction of RAW data and production and re-processing of ESD/AOD/DPD
- `athena` is written in `C++` and `Python`
- The actual code is mainly provided in form of shared libraries
  - `Service`s, `Algorithm`s and `Tool`s
  - `Service`s provide access to conditions data, geometry, the internal data store (`StoreGate`), persistification, etc...
  - `Algorithm`s are invoked by `athena` if scheduled at certain points of the run/event loop
    - `initialize` before the first event
    - `execute` in every event to retrieve data from `StoreGate`, process it and write objects back in `StoreGate`
    - `finalize` after the last event
  - `Tool`s are invoked by `Algorithm`s to compute specific tasks and are scheduled by a `Service` called `ToolSvc`

ATLAS Software Framework Athena

- options to `Algorithm`s and `Tool`s are coded in so-called `properties`
  - they can be set in `C++` (defaults) and in the `Python` steering file(s)
- Most transient data classes come with a `Reflex` dictionary that binds them to `Python` or `CINT`
- Thus `athena` `Algorithm`s and `Tool`s can be written in `Python` too
- interactive analyses from `Python` or `root` have access to all transient ATLAS data
- typical size of a full `athena` installation (release kit) is 6 GB
- kits are distributed to the Tier's for grid-production and can be downloaded and `pacman`-installed on private machines (laptops)
- only linux is supported so-far
  - Mac/Windows users use virtualization

Tag Data

- TAG attributes are a very efficient way of characterizing an event
  `https://twiki.cern.ch/twiki/bin/view/AtlasProtected/TagForEventSelection14`
- SQL database and flat `ROOT` files contain TAG info
- same `athena` query syntax for both options
- Event Level Quantities
  - Run Number/Event Number
  - Event Type
  - Number of Tracks
  - Primary Vertex
  - Lumi Block
  - Missing Et, ϕ, SumEt
  - ...
  - Data Quality Flags per Subdetector
  - Trigger Information CTP decisions, Lvl1 type, Lvl2/EF masks

Tag Data

- Object Level Quantities
  - Electrons/Photons/Muons/Taus/Jets
  - falling pT ordered
  - e/γ/μ: loose pT, ϕ, η, tightness
  - tau/jet: pT, ϕ, η, likelihood (tau/B-jet)
  - ...
- Physics TAG
  - one for each phys/perf group to be defined by them
- `athena` has been extended to allow computational processing of TAG attributes without reading the event data (not even the header)
  - can check proximity of 4-vectors not only number of objects above certain pT
    - try finding the closest jet to an electron in SQL instead ...

Analysis Frameworks

- Several analysis frameworks exist on top of `athena` and/or `AthenaROOTAccess`
- `EventView` Kyle Cranmer, P.A. Delsart, Amir Farbin, Peter Sherwood, Akria Shibata et al.
  - runs in `athena`
  - inserts objects, identifies overlaps, combines objects and analyzes them
  - can be persistified in `POOL/ROOT` files
  - became more modular since the time of the AMF
- `EWPA` (Every Where Physics Analysis) Massimilano Bellomo et al.
  - a newer lightweight alternative to `EventView`
  - runs in `athena` and `AthenaROOTAccess`
  - can also be persistified on `POOL/ROOT` files
  - both `EV` and `EWPA` can use common tools and dump DPDs or flat ntuples
- `AMA` (Atlas Modular Analysis) Max Baak, Giuseppe Salamanna et al.
  - runs analysis on ESD/AOD/DPD in `athena` and `AthenaROOTAccess` and creates flat ntuples/histograms
  - is modular and uses internal EDM to access simple types for data

Analysis Frameworks

- several `python` based skeletons
  - mostly to structure output and input files
  - run standard `athena` or `AthenaROOTAccess` otherwise
  - use the `athena` data classes
  - can use common `athena` tools
  - write DPDs or flat ntuples
  - are the preferred choice in the current Analysis Model
  - look at PAT wiki for recent tutorials
    e.g. https://twiki.cern.ch/twiki/bin/view/AtlasProtected/DPDMakingTutorial140220
  - good example following this design is the `AthenaROOTAccess` based analysis by our top-mass group in CVS: `groups/MPP/ARA_Examples_Top`
- current benchmark studies indicate that analysis on a `POOL/ROOT`-type DPD is only $5\times$ slower than analysis on a flat `ROOT`-tuple
- no common validated tools exist to analyze flat `ROOT`-tuples
  ► the final flat `ROOT`-tuple should be used for plotting/trivial analysis only!

AthenaROOTAccess (ARA)

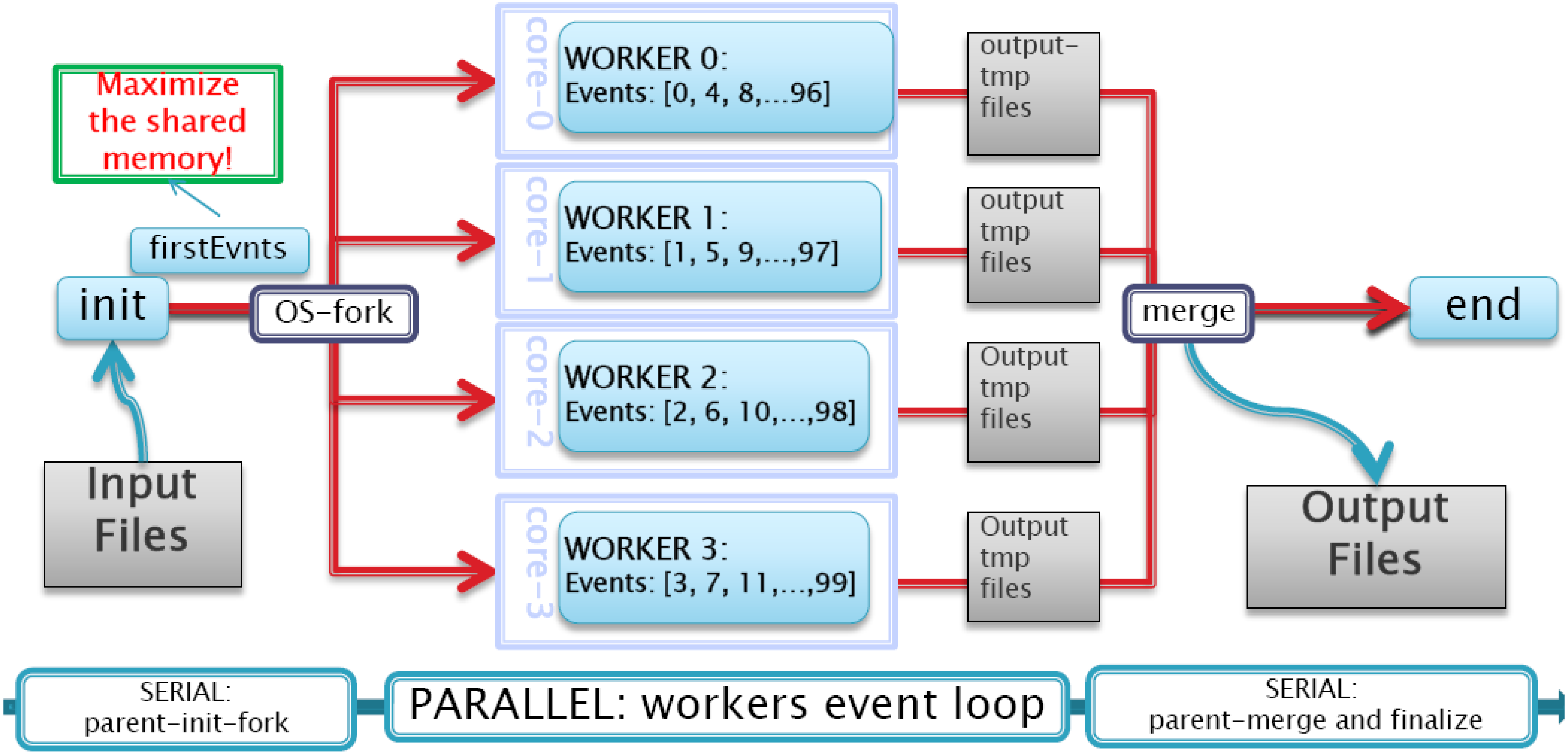
- What is `AthenaROOTAccess`?
  `https://twiki.cern.ch/twiki/bin/view/AtlasProtected/AthenaROOTAccess` by Scott Snyder et al.
  - `AthenaROOTAccess` allows you to access the objects in ESD/AOD/DPD directly from `ROOT` without the `athena` framework
  - Many athena classes (most notably the classes describing the transient objects) are available from `ROOT` and `PyROOT` via their dictionaries
  - The `athena` software has to be installed and setup but instead of `athena.py` you run `python -i` or `root`
- How does it work?
  - Minimal `test.py` start script for one AOD:

```
import user
import ROOT
import PyCintex
import AthenaROOTAccess.transientTree
f = ROOT.TFile.Open ('AOD.pool.root')
tt = AthenaROOTAccess.transientTree.makeTree(f)
```

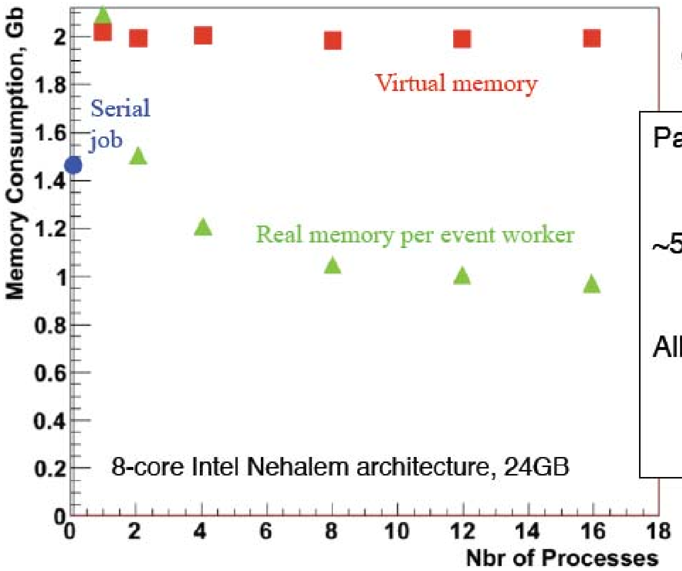
  - sets up a virtual transient tree with branches corresponding to the transient objects identified by their `StoreGate` keys
  - The transient/persistent converters (`TPCnv`) are automatically invoked when a specific entry is requested to convert from the persistent data on the ESD/AOD/DPD to the transient representation
  - for example the branch `CaloClusterContainer_p4_CaloCalTopoCluster` in the persistent `CollectionTree` on the AOD will trigger the creation of the branch `CaloCalTopoCluster` in the transient tree which points to the transient `CaloClusterContainer`

AthenaMP – Event level parallelism

$> Athena.py --nprocs=4 -c EvtMax=100 Jobo.py



Athena goes Parallel



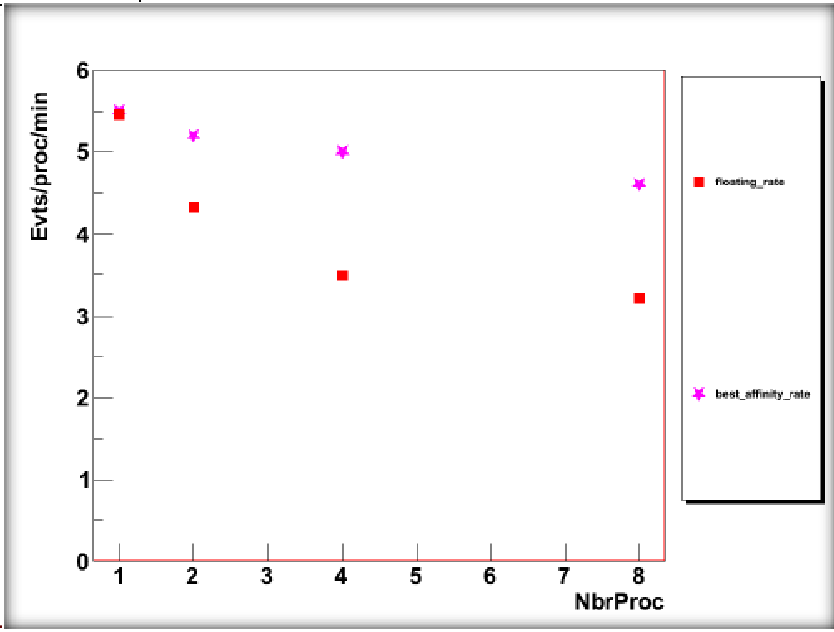
athenaMP Memory Sharing

Parallel athenaMP workers share memory
Unlike current job-parallel approach

~500MB saving per event worker running ATLAS reconstruction

Allows to either reduce hardware memory requirement, …or… (more likely?) accommodate greater memory usage by software

Typical AthenaMP performance study: the event through-put in reconstruction jobs on a eight core machine, using the default process scheduler in SLC5 (floating process sharing) and a scheduler assigning one dedicated core per process (highest core affinity). Scheduling a process on a dedicated core stays closer to the highest total event rate (flat through-put).

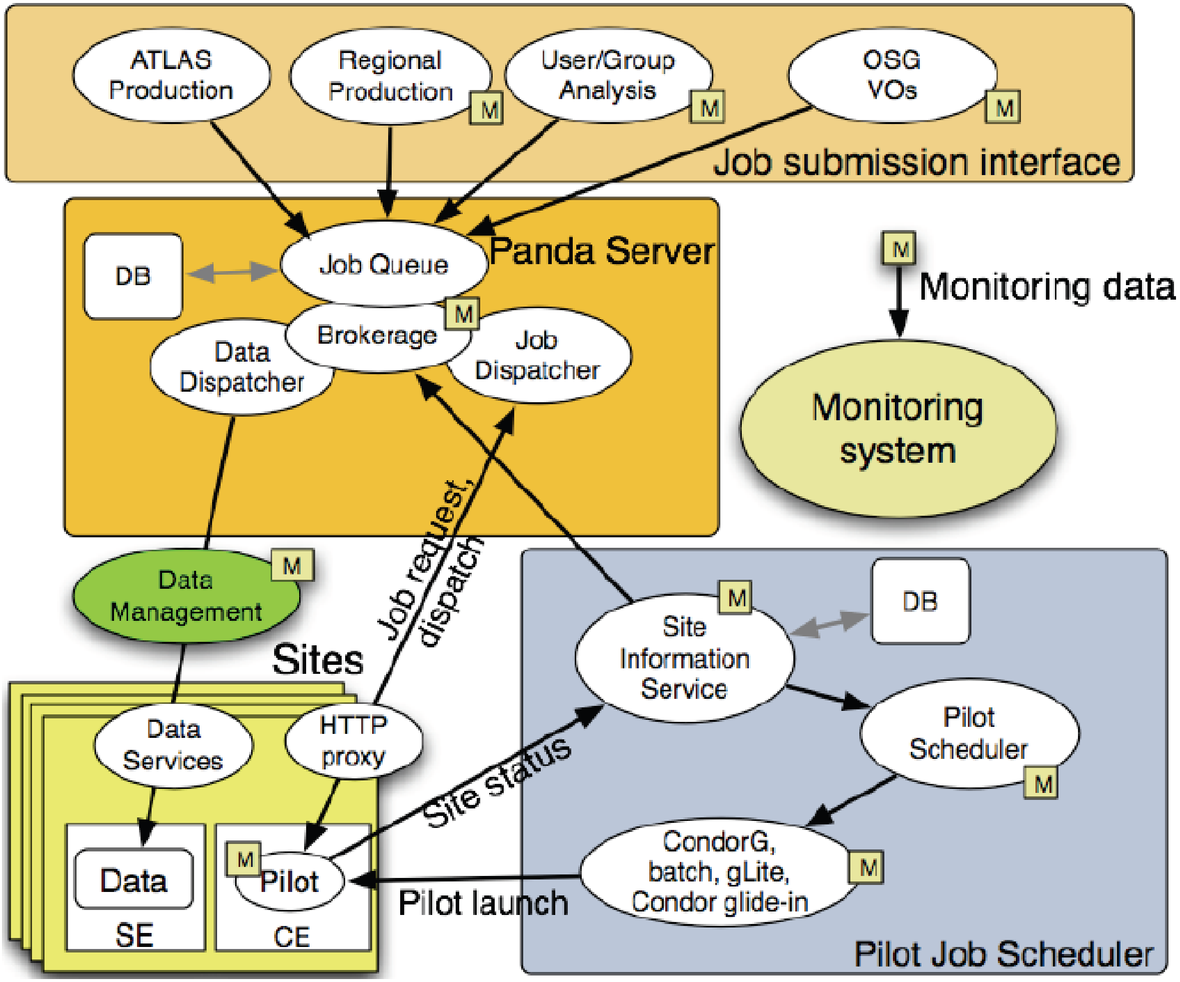


Conclusions on Software

- Data Model
  - staged model: RAW → ESD → AOD → DPD/TAG
  - initial reconstruction done at Tier-0
  - re-reconstruction at Tier-1
  - user analysis at Tier-2
- Analysis Model
  - DPD as `POOL/ROOT` file well established
  - analysis based on `C++` and `python` supported
  - first `athena` to create DPD then `AthenaROOTAccess` to analyse DPD
  - TAGs made more useful by adding computing step
- Frameworks
  - follow the simpler is better approach
  - frameworks need to be modular, use common tools and be validated
  - analysis should be done on AOD/DPD (not flat ntuples)
- Common Tools
  - rich structure of Atlas data objects allows for sophisticated common tools
  - try to cover most aspects of common analysis problems

ATLAS Jobs on the Grid - PanDA

➢ Workload Management System for Production and Distributed Analysis

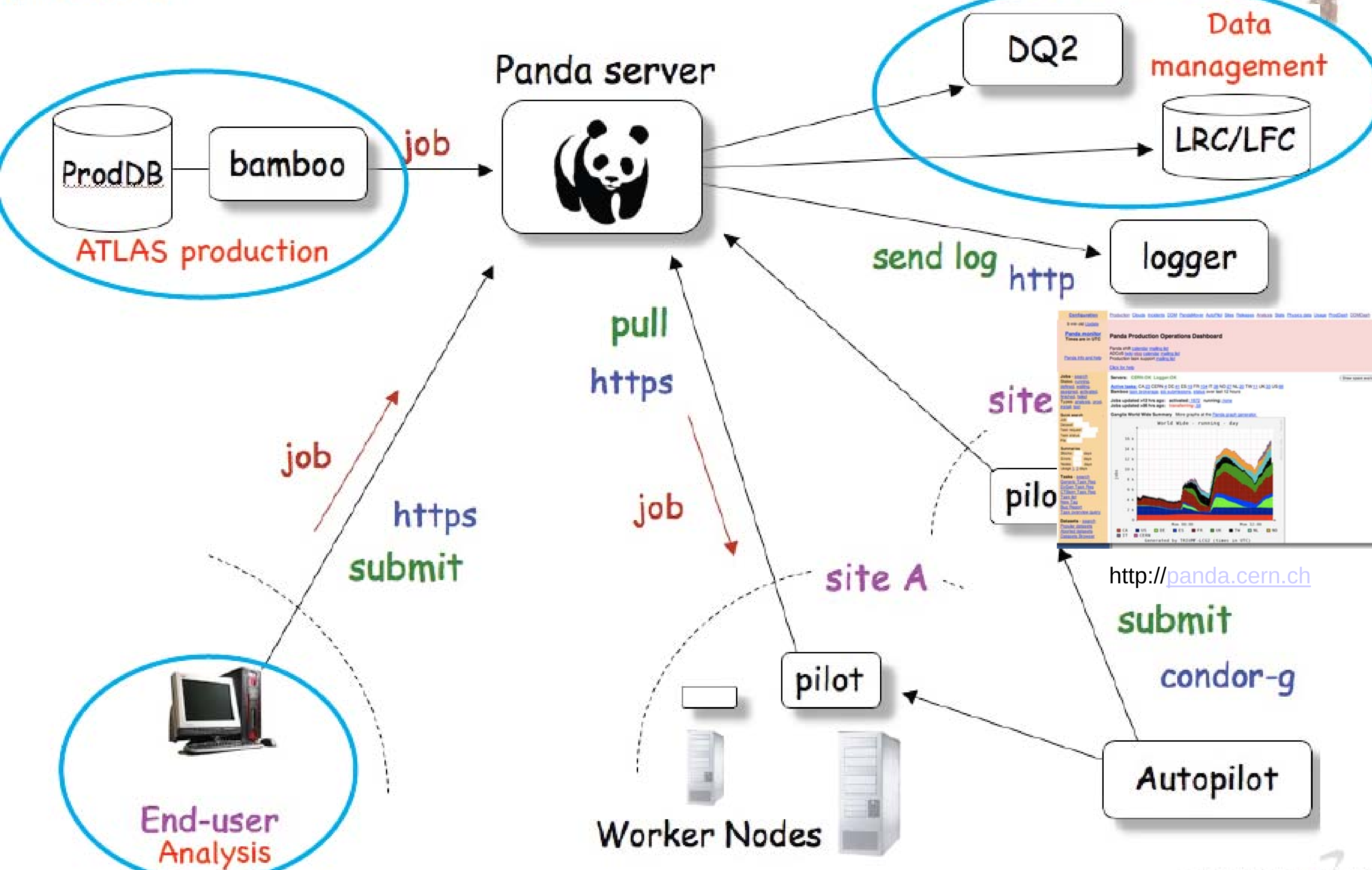


- Launched 8/05 by US ATLAS to achieve scalable data-driven WMS
- Designed for analysis as well as production
- Insulates users from distributed computing complexity
  - Low entry threshold
- US ATLAS production since late ‘05
- US analysis since Spring ’06
- **ATLAS-wide production since early ‘08**
- ATLAS-wide analysis still rolling out
- OSG WMS program since 9/06

PanDA Attributes

- Pilots for **'just in time'** workload management
  - Efficient 'CPU harvesting' prior to workload job release
  - Insulation from grid latencies, failure modes, inhomogeneity
- Tight **integration with data management** and data flow
  - Designed/developed in concert with the ATLAS DDM system
- Highly **automated**, extensive monitoring, low ops manpower
- Based on **well proven**, highly scalable, robust web technologies
- Can use any job submission service (**CondorG**, local batch, EGEE, Condor glide-ins, ...) to deliver pilots
- Global **central job queue** and management
- Fast, **fully controllable brokerage** from the job queue
  - Based on data locality, resource availability, priority, quotas, ...
- Supports **multiple system instances** for regional partitioning, scaling

PanDA Operation

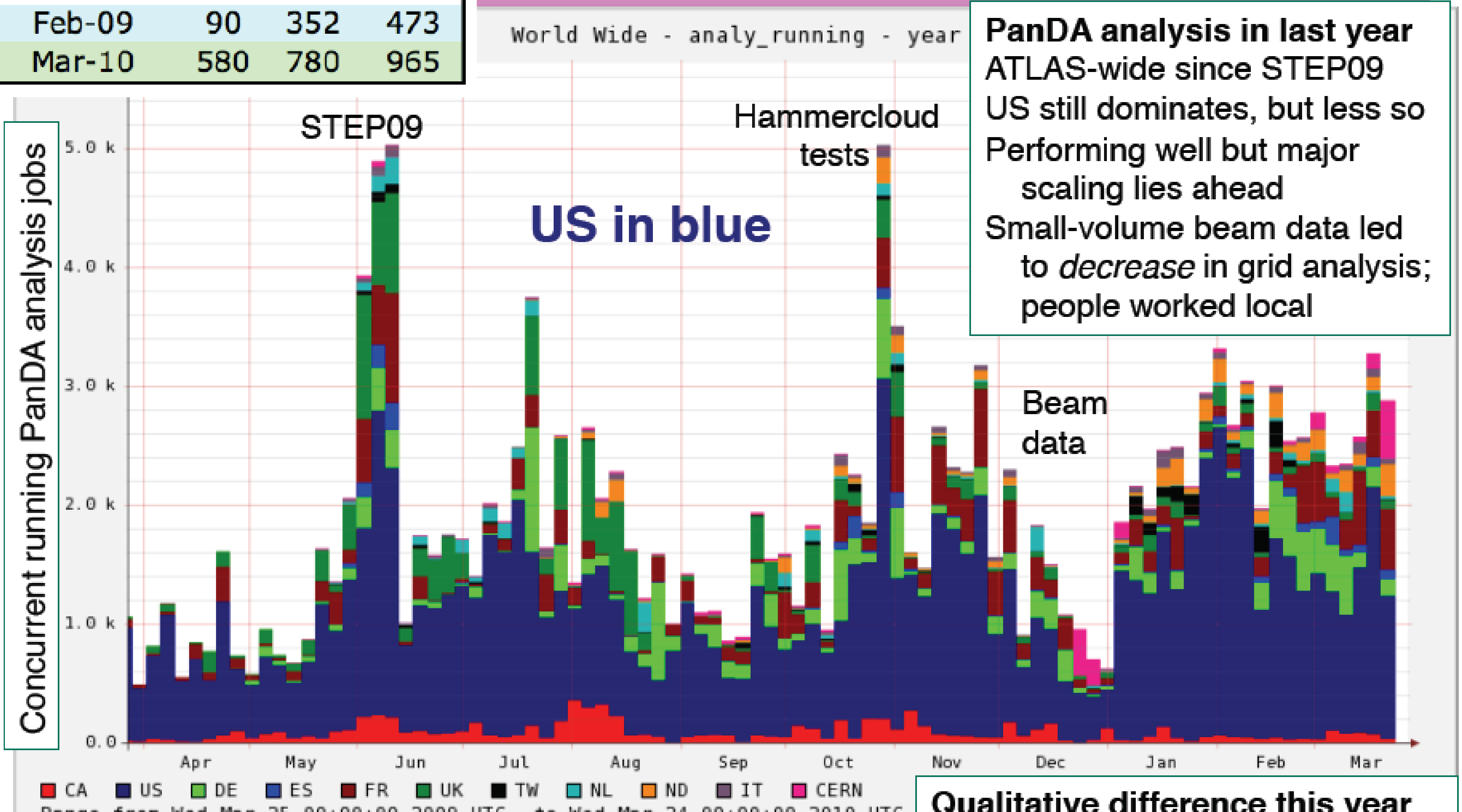


U.S. ATLAS

PanDA Distributed Analysis

PanDA: US-developed system now the ATLAS-wide standard

User counts	1mo	3mo	6mo
Feb-09	90	352	473
Mar-10	580	780	965



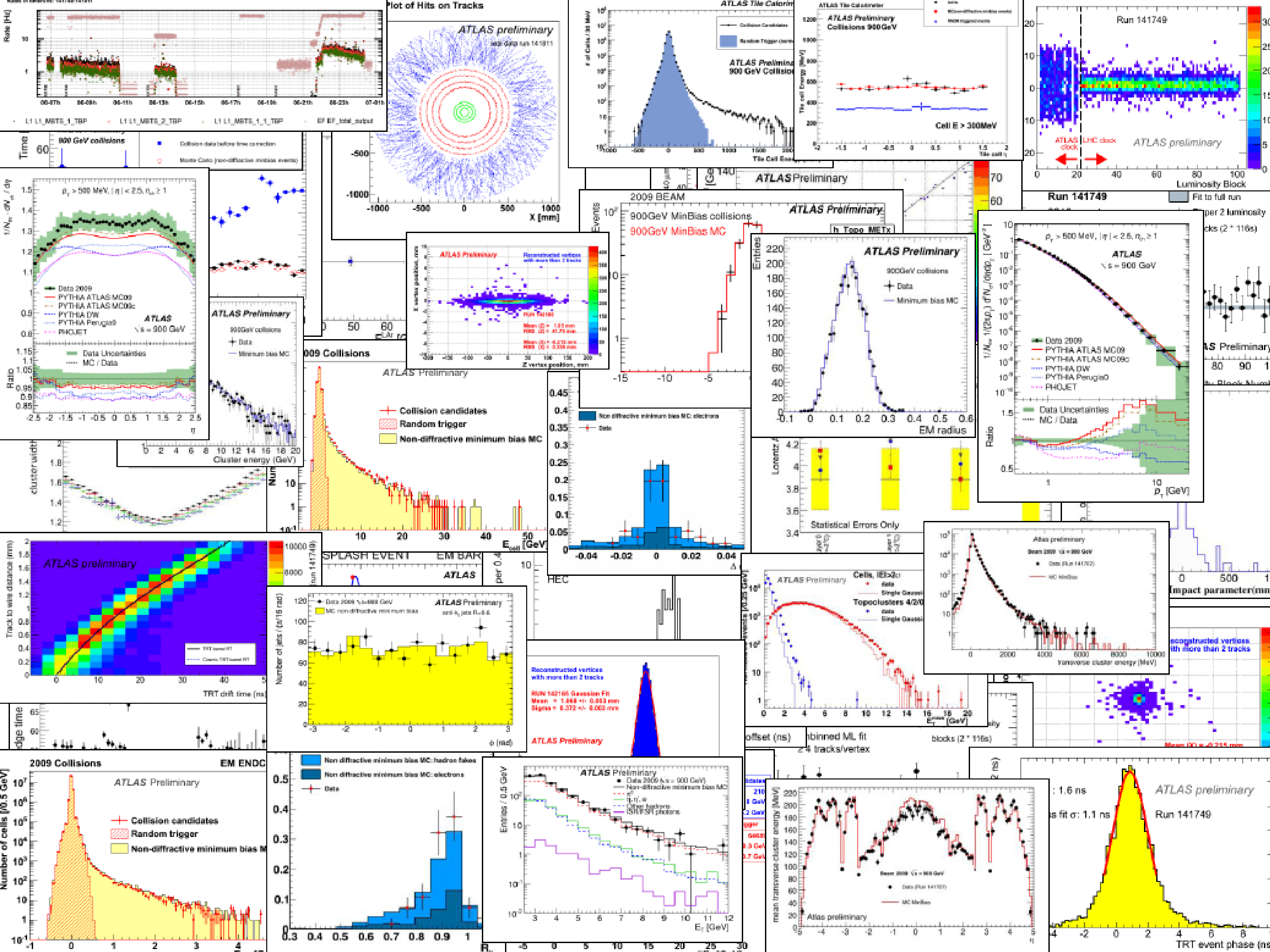
PanDA analysis in last year

- ATLAS-wide since STEP09
- US still dominates, but less so
- Performing well but major scaling lies ahead
- Small-volume beam data led to *decrease* in grid analysis; people worked local

Support by ATLAS-wide user support shift

Qualitative difference this year

Sustained step up in usage, more non-US. Much more to come



Plot of Hits on Tracks
ATLAS preliminary
ATLAS Preliminary
900 GeV collisions
Run 141749
ATLAS clock
LHC clock
Luminosity Block
2009 BEAM
900GeV MinBias collisions
900GeV MinBias MC
Data 2009
PYTHIA ATLAS MC09
PYTHIA ATLAS MC09c
PYTHIA DW
PYTHIA Perugia0
PHOJET
Data Uncertainties
MC / Data
2009 Collisions
Collision candidates
Random trigger
Non-diffractive minimum bias MC
EM radius
Minimum bias MC
Statistical Errors Only
TRT drift time (ns)
Track to wire distance (mm)
Reconstructed vertices with more than 2 tracks
Non diffractive minimum bias MC: hadron fakes
Non diffractive minimum bias MC: electrons
Data
transverse cluster energy (MeV)
Impact parameter(mm)
TRT event phase (ns)
Cluster energy (GeV)
EM ENDC
SPLASH EVENT
HEC