A new Data Format - Derived Physics Datasets

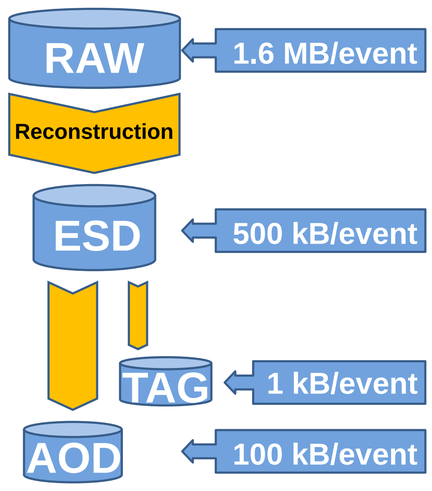
M. Barisonzi[1], S. Binet[2], U. Blumenschein[3], G. Brooijmans[4], D. Côté[1], E. Feng[5], K. Köneke[1], D. López Mateos[4,6]

[1] DESY, [2] LAL-IN2P3, [3] University Göttingen, [4] Columbia University, [5] University of Chicago, [6] California Institute of Technology

CHEP 2009 Prague, 21. – 27. 3. 2009

The ATLAS Event Data Model

- The ATLAS High-Level Trigger delivers RAW data at a rate of ~200 Hz, to be reconstructed by offline software. Several triggers separate trigger streams.
- Event Summary Data (ESD) is output of the reconstruction algorithms. Can be used for calibration and analysis.
- Analysis Object Data (AOD) is for physics analysis.
- TAG database allows event selection queries using event-level meta-data.
- A new data format for detector commissioning, performance evaluation, and physics analysis is currently defined, the Derived Physics Data (DPD), average event size of 10 kB.
- The ESD, AOD, and DPD all share the same data format.



Data Access by the Users

Users have easy access to the data that is stored on disks at the Tier 2s. In the early stages of LHC operations, understanding of the detector is paramount: calibration and commissioning take priority over data analysis. For calibration and commissioning, the ESD is the natural data format. But the ESD is not easily available to the user since only a small fraction will be on disks at Tier 2s, most of it will be on tape at Tier 1s. Something new is needed to allow easy user access to the needed information! The specialized DPDs.

The ATLAS Distributed Analysis Model

The ATLAS data is stored and distributed in a tree like structure with three layer or tiers. The ATLAS High-Level Trigger produces several data streams based on the type of the trigger decision. These RAW data streams are transferred to the Tier 0.

The ***Tier 0*** is located at CERN and provides prompt reconstruction of the data coming from the detector. This offline reconstruction produces here the ESD from the RAW data files. The AOD, DPD, and TAG datasets are produced off of the ESD and the data is archived on tape.

The data are transferred to 10 national ***Tier 1*** sites in a way that a total of two redundant copies exist. The Tier 1s are the centers where the data is reprocessed with newer software releases and/or calibration constants every 2-3 months. The TAG databases are stored here and each Tier 1 distributes the relevant data to its associated Tier 2s.

The ***Tier 2s*** are the workhorses for physics analysis where chaotic data access by the users is possible. Also, the Monte Carlo simulation is done here.

The ***Tier 3s*** are small private batch farms for data analysis.

Design specialized datasets for specific tasks:

Each specific task (commissioning a part of the detector, optimizing a specific reconstruction algorithm, searching for the Higgs boson) needs usually only a very small fraction of the total data volume. The goal is to store all needed data on disks at the Tier 2s for easy access by the users and to allow for fast data analysis without the need to process every time all data. Thus, several smaller datasets designed for specific individual needs have been designed, the ***Derived Physics Data (DPDs).***

Tools to achieve the goals of the DPD:

A framework has been utilized that allows to produce several output files running independent algorithms from a single input file. This DPD framework is fully integrated in the ATLAS data production system.

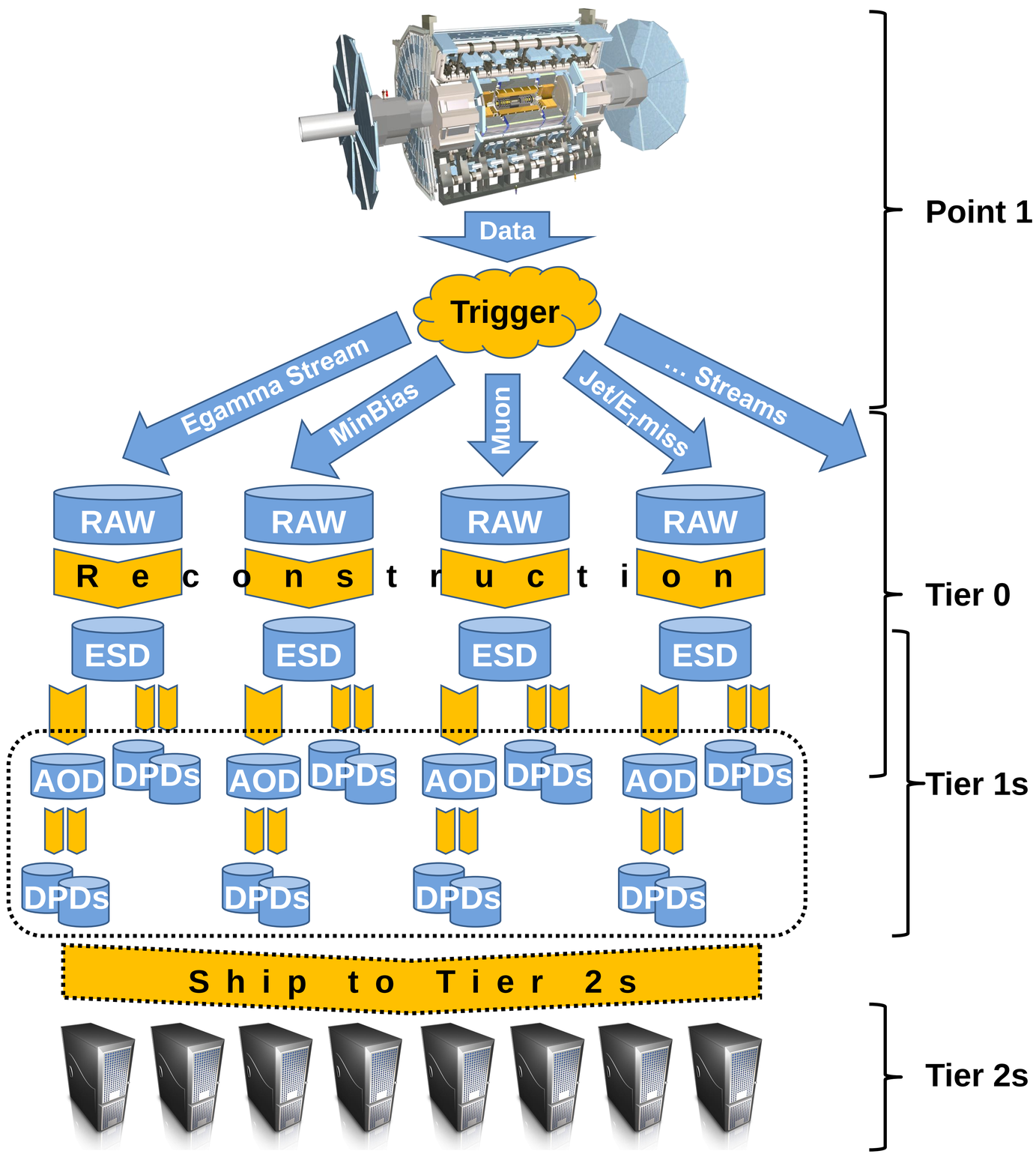
Each DPD schedules its own list of event selection algorithms to produce a pre-filtered dataset for the task at hand, *e.g.*, only selecting events with a well reconstructed high-energy photon to study jet reconstruction algorithms in the recoil.

Each DPD can also only select the object categories of interest, *e.g.*, only store the output of one standard jet algorithm or remove all calorimeter cells to save space.

Each DPD can remove individual objects of a given category, *e.g.*, only keep calorimeter cells near electrons and photons, but remove all others to study the electron and photon reconstruction and identification algorithms.

Each DPD can choose to remove parts of the individual objects, *e.g.*, remove redundant error matrices from tracks.

Every DPD automatically performs all necessary bookkeeping operations. All the luminosity information is correctly transferred to the output DPD even if no events from the input file are selected. All decisions of the event filtering algorithms are stored in meta data in the produced DPD, including the information which filter selected the event and what the filter configurations were. The DPDs are also integrated into the scheme of the TAG databases.



Commissioning DPDs:

Detector commissioning with cosmics and single beam data is already ongoing. For this, the ESD is the best format, but is too large in data volume. About 216 million cosmic events (~600 TByte) have been recorded so far. For individual tasks, only a small subset of the about 500000 events are needed, *e.g.*, events where a cluster was found in the electromagnetic calorimeter. The ESD format is kept, but only specific events are selected. Eleven distinct commissioning DPDs with a data volume of only about 1 TB or less each were produced in December and a new reprocessing will start soon.

Performance DPDs:

During the early phase of colliding beam LHC running, the highest priority in data analysis will be understanding, evaluating, and improving the reconstruction and identification algorithms of the ATLAS software and the detector performance.

For this, quantities from the ESD are needed. Nine distinct performance DPDs are produced off of the ESDs. But only interesting events for the task at hand are kept, *e.g.*, keeping events with at least one reconstructed muon. Also, only detector information for the task at hand is kept, *e.g.*, only calorimeter cells around an electron or photon candidate.

Physics DPDs:

In order to improve the efficiency of physics analysis, several DPDs are produced off of the AODs. They contain all information needed for the analysis at hand. But only events of a certain category are selected, *e.g.*, only event that contain two reconstructed electrons or events that passed a certain class of triggers.

About 15-20 distinct physics DPDs are defined, each containing only a few percent of the size of the AOD. These physics DPDs significantly increases the speed of the analysis due to the drastically reduced number of events a given analysis has to process.