Draft Proposal for 'Life without ESD'

(J. Boyd, B. Heinemann, RD Schaffer, D. Rousseau)

Draft v1.0

In this document we give our main proposal for reducing the storage of ATLAS data in order to allow a significant increase in the trigger rate. We also discuss a rough preliminary estimate of the resources the main proposal would take for the 2011 data, and consider some alternative scenarios. In all cases we assume a trigger rate of 400 Hz and base our resource estimate on a single run from 2010 (167607). The final resource estimate will be obtained by Computing Coordination. While we mostly discuss data from 2011 and future years we also comment on the actions that can be taken on 2010 data and on MC.

Proposal

1. No long term storage of ESDs on disk for bulk physics streams (Muon, JetTauEtmiss, Egamma, MinBias)

2. Do not store any ESD on tape at all

   A copy of ESD for all data on tape would be beneficial for a so-called "ESD reprocessing" where AODs and DESDs are remade from ESD's. However, we see no benefit in this over a full reprocessing from RAW and are prepared to abandon this option.

3. Provide at least 6 primary replicas of the AODs for the physics streams. Additional secondary copies can be created depending on user demand and disk space constraints .

4. Provide at least 4 primary replicas of the DESDs. Additional secondary copies can be created depending on user demands and disk space constraints . The DESD only needs to be kept for the most recent (N^{th}) processing. Since the DESD is primarily used for performance studies only the most recent processing needs to be provided. A possible exception is the DESD_RPVLL, which is used for physics analysis.

5. Provide 2 replicas of full ESDs from all physics streams for an approximate 10% of the data. For the Tier0 processing the data would correspond to most recent N months where we consider N=1 and N=2. We consider this *a rolling buffer* similar to the CAF. Users cannot expect that any data older than N months are available and do not need to be informed about removal of these ESD's. For reprocessing this would be one specified period, which again typically corresponds to 10% of the data (typically the latest/highest luminosity data). No automatic replication via PD2P of this sample is allowed.

Many groups have requested access to a representative fraction of the ESD (e.g. trigger, egamma) and providing the most recent data will address this. In addition the most recent data are also the most useful for detector and performance studies related e.g. to a problem that occurred during data taking. Another benefit is that in case there has been a problem in the initial DESD production this can be recovered from easily without having to again fully process the data. Physics Coordination is considering creating an organized effort to identify spectacular events (possibly due to new physics) quasi-online and this group would benefit from access to full ESD's in case they find anything. If the most recent ESD data is replicated to a Tier1 this would also allow migrating the DESD production to the Tier1's and thus reduces Tier0 resource constraints.

6. Provide 2 replicas of full ESDs for all data for some small streams (CosmicCalo, ZeroBias, StandBy, express). There is no need to keep the AOD from the express stream. The ESDs are fully provided for the Nth processing but for the (N-1)th processing only the ZeroBias and CosmicCalo is kept. Automatic creation of secondary copies of these ESD's via PD2P is allowed.

   This has been requested by detectors, e.g. CosmicCalo is needed for LAr and Tile noise studies; ZeroBias is needed for LAr and Tile pileup studies and for overlay of real data events with simulation and StandBy is needed for SCT noise studies. The express stream ESD is needed for Data Quality assessment on the CAF and many groups commented that they would also benefit from it being available on the grid. CosmicCalo is also needed for the "stopped gluino" physics analysis. The rate of these streams is small: 1Hz (ZeroBias), 10 Hz (CosmicCalo) and 10 Hz (StandBy) and should stay at this rate. The express stream rate is also budgeted at 10Hz but we also consider a 15 Hz rate in our resource estimates.

7. The ESD size will very likely be reduced by 30% compared to 2010 due to dropping of redundant objects in the calorimeter and tracking area. E.g. in run 167607 this corresponds to an average reduction among the main physics streams (Egamma, Muon, JetTauEtmiss, MinBias) from 1.4MB to 1.1MB. The size will increase in the future with increasing pileup and we need to assess this. In our resource numbers we use the current numbers, which we consider pessimistic.

8. Provide 1 copy of RAW data on disk. This is only required for data taken within the *last year*. No automatic replication should be allowed. The motivation for this comes from the physics requirements during an "exciting discovery". When the physics conveners and/or coordinators are informed about a credible signal of new physics there will be a big pressure to move as fast as possible. One critical aspect will be an in-depth investigation of the detailed properties of the events by the relevant detector and CP groups (e.g. the details of the LAr and Tile cell information which can only be performed on ESD or RAW). The typical number of events that people will want to look at in detail is 10-100 and these will generally be distributed randomly across runs since the analysis will typically have analysed a full year of data The requirement of having access within <24h is driven by not delaying the progress of ATLAS on understanding these events given the high pressure to publish as quickly as possible in such a scenario. A convenient user interface needs to be provided to turn a RAW event back into an

ESD etc. for user convenience but this is technically not very difficult. Creation of event displays or detailed debugging of specific events in the context of physics or performance studies also would benefit from expedient access to individual RAW data events.

9. Reduction of the RAW event size by a factor of 2. “gzip” achieves a factor of two on the RAW event size. If indeed the RAW data are kept on disk effort should be made to obtain this factor of 2 either by applying gzip or by packing the data better.

We assume in all cases that at least 2 replicas for the ESD are required to enable a fast recovery from losses of individual files due to disk failures. For the RAW we also assume that we need two replicas on disk to minimize the time required to recover a lost file. However, since the RAW data are also stored on tape a lost file can in principle also be recovered from tape with an increased latency. The latency and operational burden introduced by having just 1 RAW copy on disk should be estimated and then the number of copies required could be revised.

Preliminary rough Resource estimate

We made an estimate of the resources required in 2011 using measurements from run 167607 in 2010. This run was taken with a rate of 400 Hz and had a peak number of interactions of 3.8. The length of the run during stable beam was 11h. We consider this to be representative for 2011 although we note that the pileup will likely be larger in 2011 causing a higher event size for ID. For the default resource estimate we assume 200 days of pp running with an LHC up-time of 30%. Note, that this is only a rough preliminary estimate and the real estimate will come from putting this model into the computing model spread sheet.

Run 167607:

Stream	RAW (TB)	ESD (TB)	AOD (TB)	DESD (TB)	Nevent (10^6)
Physics_Muons	7.82	7.51	1.03	0.55	5.7
Physics_JetTauEtmiss	6.77	8.02	1.23	1.32	4.8
Physics_Egamma	5.21	5.24	0.70	0.85	3.9
Physics_MinBias	0.83	0.64	0.06	0.19	0.7
Physics_CosmicCalo	0.56	0.22	0.01	0	0.4
Physics_ZeroBias	0.09	0.03	<0.01	0	0.03
Physics_StandBy	0.23	0.18	0.02	0	0.2
Express_express	0.67	1.05	0.38	0	0.5
Sum	22.18	22.89	3.43	2.91	16.2

Thus the total data volume for the main streams (Muons, JetTauEtmiss, Egamma, MinBias) for this run is 20.6 TB for RAW, 21.4 TB for ESD's, 3.0 TB for AOD's and 2.9 TB for DESD's. We refer to these as the "bulk streams".

We now convert this single run into an annual data volume by scaling it to total of 60 full days of running (200 days with 30% LHC efficiency). Given that the above run corresponded to 11h (0.458 days) the conversion factor is 131. The resulting values are:

- RAW bulk volume: 2.7 PB
- ESD bulk volume: 2.8 PB
- AOD volume of all streams apart from express stream: 400 TB
- DESD volume: 380 TB
- ESD of express, CosmicCalo, ZeroBias+Standby: 200 TB
- The number of events is 2.1 x 10^9

These are the base numbers for our calculation. Note that these still correspond to an average events size of 1.4 MB. We use this as default but comment on the impact of a 1.1 MB size as projected. Note, that the RAW data volume can be reduces by a factor two if gzip is applied.

Overall Resource Estimate for the proposed scenario

For each data processing the following total volume for the derived formats is estimated:

- 6 AOD replicas correspond to 2.4 PB
- 4 DESD replicas correspond to 1.5 PB
- 2 ESD replicas for the bulk streams for a 1 month period (30 days with 30% LHC efficiency) correspond to a data volume of 0.4 PB. Of course 2 months would correspond to 0.8 PB etc.
- 2 replicas of the CosmicCalo, ZeroBias, StandBy and express stream ESD's correspond to 0.8 PB (this is 0.13 PB when only considering CosmicCalo and ZeroBias stream)

The total data volume of the dervied formats is thus V(N)=5.5 PB when keeping 2 months of the recent ESD's.

When a reprocessing is done we would remove the older (N-1) versions of the DESD's and the StandBy and express stream ESD's. Thus the 2nd copy would then be reduced from 5.5 TB to V(N-1)=3.3 PB, and the total volume is V(total)=8.8 PB.

In addition one copy of RAW on disk corresponds to 2.7 PB for the above rate and LHC efficiency assumptions. It is not clear to us if for RAW we also need 2 replicas as safety net against disk failures or if 1 may be sufficient.

Alternative scenarios:

Any variations of the model can easily be calculated from the above numbers, e.g.

- assuming a total number of AOD and DESD primary replicas of 10 for the most recent processing would increase V(N) to 9.4 PB, V(N-1) would be unchanged and V(total) would thus increase to 12.7 PB.

- asuming a 30% reduction in ESD size and no change in the AOD size V(N)=4.8 PB, V(N-1)=3.1 PB and V(total)=7.9 PB

- assuming a 20% increase in the DESD and Express stream we obtain V(N)=6.1 TB, V(N-1)=3.5 PB and V(total)=9.6 PB

- Keeping the 2 replica's of the DESD also on disk for the (N-1)the processing will increase V(N-1) to 4.0 PB and thus V(total)=9.5 PB

- If the LHC efficiency is 50% instead of 30% in V(total) increases to 14.7 PB bringing it above the available resources. The 30% reduction of the ESD would decrease this to 13.1 PB. The RAW data volume would increase to 9.0 PB

- Assuming an increase of 20% of the AOD results in V(N)=6.0 PB, V(N-1)=3.8 PB and V(total)=9.8 PB

- Assuming only 3 replicas are kept for the AOD's of the (N-1)the processing results in V(N-1)=2.1 PB and V(total)=7.6 PB

- The current (2010) scenario of having 2 ESD replicas and 10 AOD and DESD replicas for both procesings, all streams apart from the express stream, and no RAW data on disk corresponds to a total volume of V(total)=27 PB

2010 data

Proposals on what could be done immediately with 2010 data:

- Remove the DESD's for rel15 (Tier0 processing) apart from the DESDM_RPVLL on March 15th. This will save about 0.5 PB.

- Reduce primary AOD and DESD replicas to 6 for both processing versions on March 15th. This will save about 3.5 PB

- Remove all ESD's other than those of periodI with a 3-month notice (i.e. by May 1st) for the rel15 processing. This will save about 3 PB.

We think that on a time scale of 6 months about 7 PB of 2010 data can be removed without compromising any analyses. After that the 2010 data will be reprocessed again and will then need to obey the same constraints as the 2011 data and the rel16 derived data can further be reduced. It appears to not be needed to keep the 2010 RAW data on disk as long as the ESD's from the rel16 processing are kept on disk.

We can also consider deleting further data from periods A-C (only 300/nb) most likely in consultation with the detector systems, physics and CP groups.

In addition older data (2009 and cosmics) can also largely be deleted again tbc with systems, CP and physics groups.

Heavy Ion data

We have not yet accounted for the Heavy Ion data. Their volume of course also will need to be considered in view of the overall space constraints.

Monte Carlo

Many groups have commented that they need some access to MC ESD's. We are not yet ready to make a concrete proposal as we need to survey the current numbers and the request in more detail.

Summary

In summary we have proposed a scenario for dropping the bulk stream ESD's that we think fulfills the requirements expressed by the systems, the CP and the Physics groups and tried to respect the constraints as we understand them from Computing Coordination. Assuming an LHC efficiency of 30% and a trigger rate of 400 Hz and using a benchmark run from 2010 we have estimated the resulting volume for derived data of 2011 varies between 7.9 and 12.7 PB depending on the assumptions on the events size and number of replicas for the individual datasets. In addition we estimate a 2.7 PB (1.4 PB if gzipped) volume of RAW data. Our default scenario results in a total data volume on disk of about 10 PB in 2011 and would be similar in future years. This represents a reduction of about a factor of 2.7 compared to the current model. If the LHC efficiency is siginifcantly higher than 30% further reductions are likely necessary which can largely be achieved by reducing the number of primary replicas further. In addition our proposal yields an reduction of the volume taken by the 2010 data by at least 7 PB on a time scale of 3 months.

Acknowledgements

We thank Richard Hawkings, Aleandro Nisati, Srini Rajagopolan and Jim Shank for their feedback on earlier versions of this document