ATLAS Simulation using Real Data: Embedding and Overlay

Andrew Haas, on behalf of the ATLAS collaboration
NYU Physics, 4 Washington Place, New York, NY 10003
andy.haas@nyu.edu

Abstract. For some physics processes studied with the ATLAS detector, a more accurate simulation in some respects can be achieved by including real data into simulated events, with substantial potential improvements in the CPU, disk space, and memory usage of the standard simulation configuration, at the cost of significant database and networking challenges. Real proton-proton background events can be overlaid (at the detector digitization output stage) on a simulated hard-scatter process, to account for pileup background (from nearby bunch crossings), cavern background, and detector noise. A similar method is used to account for the large underlying event from heavy ion collisions, rather than directly simulating the full collision. Embedding replaces the muons found in $Z \rightarrow \mu \mu$ decays in data with simulated taus at the same 4-momenta, thus preserving the underlying event and pileup from the original data event. In all these cases, care must be taken to exactly match detector conditions (beamspot, magnetic fields, alignments, dead sensors, etc.) between the real data event and the simulation. We will discuss the status of these overlay and embedding techniques within ATLAS software and computing.

1. Introduction
In each LHC filled bunch crossing during proton-collision mode there are soft pp collisions called “pileup”, in addition to the hard-scatter pp interaction that typically causes the event to be triggered. In Run1 (2009-2012) there were up to 20 pileup interactions per bunch crossing. This has risen to about 40 during Run2 (2015-18), and is foreseen to rise further to 200-400 during high-luminosity LHC running (2025-). Along with pileup there is also “cavern background”, a gas of neutrons and photons from collision byproducts that lives many microseconds and dominates muon chamber backgrounds, and detector noise from thermal and electronic effects.

ATLAS subdetectors are also sensitive to not only the triggered bunch crossing, but the surrounding bunch crossings in time (see Figure 1). For instance, due to the drift times in the LAr calorimeters, they are sensitive to as many as 32 bunch crossings in the past and 6 in the future. These backgrounds are complicated, and thus difficult to model accurately and computationally expensive in the detector simulation.

The event overlay technique is an alternative way of modeling pileup, cavern background, and detector noise during the digitization of simulated events. In the current pileup simulation technique used by default at ATLAS, hard interactions are simulated, e.g. in Geant4 (G4), and then mixed with simulated pileup MC during digitization. Detector noise is also emulated during digitization. The overlay method, explored in this talk, instead mixes in a real data event during digitization, to account for the pileup background, cavern background, and detector noise. As will be discussed, this real event is a “zerobias” event, which is a random sampling of detector conditions during the filled bunch crossings. Advantages and disadvantages of overlay vs. pileup simulation will be discussed below, but one advantage of overlay is its inherent speed. Whereas simulated pileup takes ~1 minute to digitize
per event in Run2 conditions, and the time required rises linearly with the number of pileup interactions, overlay takes a nearly constant ~10 seconds per event, since the detector data is already digitized and just active channels must be merged.

![Figure 1: Sub-detector sensitivity to bunch crossings.](image)

2. Background data sample
The background (zerobias) data is collected using a dedicated trigger, which fires one LHC turn after a high-$p_T$ L1 EM trigger fires. This ensures that the zerobias data is triggered proportional to the luminosity in each bunch crossing. This trigger was prescaled to ~10 Hz in Run2 at ATLAS. About 3 Hz of this rate was selected using the HLT to have a jet with $p_T>40$ GeV, to have a larger sample of high-$p_T$ zerobias events, such that they do not need to be reused as frequently. No zero-suppression of the readout is used, where possible (nearly everywhere except in silicon). A total of ~100M zerobias events have been collected per year. Each zerobias event is about ~2MB after compression.

Offline, the zerobias data are sampled from lumi-blocks (1-minute time periods) in the desired time period to reproduce the luminosity profile of a high-$p_T$ trigger (accounting for dead time, prescales, the mix of HLT jet triggers in the zerobias sample, etc.). 50k randomly-chosen events are written to each zerobias data stream. The events in each stream are ordered to be monotonically increasing in time and run/event number, to later ease the conditions access.

3. Overlay method
The steps used for producing an overlaid MC event are as follows (see Figure 2).

1. Read in an input zerobias RAW data event (after offline lumi-weighting, etc.).
2. Simulate a hard-scatter G4 event, with conditions matching each selected data event (beamspot/tilt, alignments, magnetic fields, etc.). Alignments are updated at the initialization of each G4 job of 100 events, and 500 jobs complete each 50k sample. Using real-data detector alignments does lead to overlapping G4 volumes in some cases. The geometry description (of dead material) has been updated to eliminate all major ones. Minor adjustments to the alignments do not lead to new overlaps.
3. Overlay each zerobias data event with a matching G4 event at the detector channel level, then digitize combined signals. The way in which overlapping channels are handled during overlay is sub-detector dependent, but in general the signals in MC and background data are each un-
digitized to estimate the analog waveforms. These waveforms are then added, to simulate the combined signal, and this combined signal is then digitized.

4. Reconstruct the combined event as data. Some MC conditions are used, where simulation cannot model data accurately, e.g. drift radius vs time for straw tracker / muon tubes. We thus use MC constants for digits that have some energy contribution from simulated hits, and data constants elsewhere for background digits – not ideal for digits with energy from background and simulation.

![Figure 2: Overlay steps](image)

### 4. Validation and testing methods

Various outputs of the overlay simulation have been studied, to isolate modeling of the background, hard-scatter signal, and their combinations:

1. Compare directly reconstructed zerobias input RAW dataset to overlay MC with just a single neutrino, using data conditions and alignments. This isolates just the pileup/noise background effects, with no signal energy.
2. Compare standard pileup simulation to overlay simulation, using MC alignments but data conditions, with no background hits overlaid and no pileup MC collisions added. This isolates just the hard-scatter MC effects, with no background energy.
3. Compare overlay of energy at the digitization step to direct addition of energy/time G4 hits, with MC alignments and MC conditions:
   a) overlay simulation: signal hits + (MC hits → digitization) → overlay
   b) standard simulation: (signal hits + MC hits) → digitization
   This isolates the overlay digitization algorithms, for channels with background and hard-scatter energy, to see e.g. effects of tracker zero-suppression.
4. Compare overlay MC to pileup MC, for various processes.
5. Compare overlay MC to data (and pileup MC), with a clean selection, e.g. Z→ $\mu\mu$.

### 5. Drawbacks compared to pileup simulation

Overlay has several drawbacks, compared to pileup simulation:

- Less accurate when combining overlapping background and signal on the same channel for some subdetectors (e.g. silicon), since zerobias data is zero-suppressed in silicon.
- In overlay, a background hit below threshold cannot add to a G4 hit below threshold, to make a hit above threshold.
- In some cases, the background data does not record enough samples to reconstruct the background pulse precisely. For instance, the L1calo trigger uses 7 samples online, but only 5 stored.
• MC conditions are used in a few places, even when some energy is coming from background data (overlapping straw-tracker hits, e/gamma shower shapes, scale-factors for JES, etc.). (We could use truth to tell when the object is mostly background, but have not so far.)
• Potential G4 geometry overlaps when using data alignments.
• Harder to simulate future high luminosity (multiple overlay possible in most sub-detectors, but challenging in the calorimeters).
• Must wait until all data is collected for a given period before generating simulations.
• More challenging to produce. There is zerobias data preparation and large database access needs during G4/overlay and reconstruction.
• Don't have the background truth information since it’s data.

6. Advantages compared to pileup simulation
Overlay has several advantages, compared to pileup simulation:
• Real pileup events from data are used, so no generator tuning is needed.
• The N_vertex, z_vertex, and instantaneous luminosity match data precisely.
• No event weighting is needed, so there is higher “efficiency” of MC use.
• Faster (and less memory) at high luminosity than standard pileup digitization.
• Realistic mix of instantaneous luminosity variations and bunch structure and in-time/out-of-time pileup contributions (including satellite collisions).
• True detector noise and occupancy (see Figure 3), including luminosity-weighted variations and correlations between channels (like noise bursts).
• Conditions (beamspot, dead channels, etc.) are from data, including luminosity-weighted correlations. For instance, a realistic fraction of events where a section of one sub-detector is malfunctioning and a nearby section of another sub-detector is also malfunctioning.

![Figure 3: Number of calorimeter clusters vs. depth in calorimeter of cluster center, from [2]. This is sensitive to calorimeter noise and pileup distributions due to seeding methods.](image)

7. Cavern background modelling
Overlay is particularly good at modelling cavern background, compared to simulated pileup. Cavern background is difficult to simulate accurately since it involves low-energy neutrons, shielding, etc. It is not even included in the default ATLAS simulations, but is automatically included in overlay simulation. Cavern background is critical to model well for some physics analyses, for instance when searching for vertices from long-lived neutral particle decays in the muon system [3], see Figure 4.
8. Heavy-ion Collisions

It is very challenging to simulate the “soft” parts of nuclear collisions. Generators such as HIJING have limitations, and simulating the resulting ~10k particles in G4 is slow. ATLAS has made extensive use of overlay simulation instead, to model the underlying event of each heavy-ion nuclear collision, see Figure 5. The method of overlay simulation is like pp production, except a minimum bias heavy-ion is collected, to ignore the large number of crossings with no collisions, and the simulated hard-scatter interaction is overlaid at the same 3D position as the reconstructed minimum bias collision.

Figure 4: Number of reconstructed muon detector hits vs. eta, for various simulations, from [3]. The remaining 10% difference is largely due to slightly different timing cuts applied during data and MC reconstruction.

Figure 5: Comparison of low-centrality (glancing) Pb+Pb heavy-ion collisions (left) to high-centrality (head on) collisions (right). The yellow is overlay simulation, which models the Pb+Pb collisions on the left well. The right then shows the effects of jet quenching, which are not in the overlay simulation used, from [4].
9. Overlay for Z →μμ embedding

Embedding takes a data event (e.g. Z→μμ) and replaces objects with another type of object (e.g. taus) to emulate a related process (e.g. Z→ττ in this case). This is critical for modeling Z→ττ background in the H→ττ analysis, and it also has other uses. So far this has been done at the reconstruction level at ATLAS, with tracks and calorimeter cells, but this has inaccuracies. Tau's are simulated and reconstructed without pileup, there are no cells below the zero-suppression threshold in data event, and one can't run the L1 trigger simulation.

We are working on using overlay techniques to perform embedding. We have recorded ~5 Hz of non-zero-suppressed Z→μμ this year, to use as overlay input (instead of zerobias). We can then simulate taus in G4 at same vertex/momenta as the reconstructed muons. Then we will overlay the MC taus and Z→μμ data event, removing digits that were used to form the reconstructed Z→μμ tracks' hits in the inner detector and muon system. Finally, we will reconstruct the overlaid event, performing subtraction of the muons' calorimeter cell energies (as in the original embedding technique).

10. Summary

Pileup, detector noise, and cavern background are difficult to model accurately at ATLAS using simulated pileup and emulated noise. Overlay simulation is a way of automatically including these backgrounds, using specially recorded data events as input. It has some drawbacks and limitations, but also many advantages in terms of speed and accuracy, compared to simulated pileup. ATLAS has used overlay for analyses where cavern background is important, and for many heavy-ion analyses. It has also found use recently for Fast-Track trigger operations, where pre-calculated banks of data track patterns must be calculated (from simulation), using alignments from the real detector (since it is used to trigger real events). We are also working to extend the overlay technique to do embedding at ATLAS.

We expect to use overlay simulation at ATLAS more regularly in the future, as the technique matures and production issues (such as conditions access) are improved. Pileup will only become more important to model accurately in the future, as instantaneous luminosity increases.

References

[1] ATLAS Collaboration: The ATLAS experiment at the CERN Large Hadron Collider, JINST 3 S08003, 2008.
[2] ATLAS Collaboration: Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1, arXiv:1603.02934 [hep-ex].
[3] ATLAS Collaboration: Search for long-lived, weakly interacting particles that decay to displaced hadronic jets in proton-proton collisions at sqrt(s)=8 TeV with the ATLAS detector, https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PAPERS/EXOT-2013-12/, Phys. Rev. D 92, 012010 (2015).
[4] ATLAS Collaboration: Measurement of the correlation of jets with high pT isolated prompt photons in lead-lead collisions at sqrt(s_NN) = 2.76 TeV with the ATLAS detector at the LHC, ATLAS-CONF-2012-121.