Fast simulation

HSF Parallel Session
2019 Joint HSF/OSG/WLCG Workshop
Jefferson Laboratory, Newport News, VA

Vincent R. Pascuzzi
University of Toronto

On behalf of the LHC experiments

20 March 2019
Fast agenda

- Motivation
- Current fast simulation
- Future fast simulation
- Summary
Motivation

The problem

Large MC statistics to model recorded data
- Increasing luminosity poses greater challenges

Highly-descriptive detector description not always necessary
- e.g. signal samples, systematics, upgrade studies
Motivation

Energy deposits converted to digital signals

Build physics objects

Interactions of particles with matter

4-vectors
Motivation

Energy deposits converted to digital signals

Build physics objects

Data for analysis

Joint HSF/OSG/WLCG Workshop (2019-03-20)
Current fast simulation

*Sacrifice accuracy and precision for speed*

- Maintain certain level of agreement with highly-detailed full simulation
- Compatibility with offline processing chain
- Reusability and event/process recycling

Simulation
- Majority of CPU spent in calorimeters

Digitization
- Large number of inner tracker readout channels; complex modelling of readout emulation

Reconstruction
- Pattern recognition: combinatorial problem (function of $<\mu>$)
Current fast simulation

Simplified geometries

- Approximations of full Geant4-based geometry for tracking and calorimetry

- Physical volumes -- e.g. cuboids, cylinders -- built from primitive surfaces rather than complex G4VSolid s

- Detailed volume materials projected onto 2D surfaces

- Particle-material interactions on surfaces determined by surface 'thickness' measured in radiation/interaction lengths

- Database-lookup for detector properties, e.g. alignment, element extents
Current fast simulation

Calorimetry

Geant4 EM and hadronic templates
- Simulate $e, \gamma, \pi^\pm$ shower shapes/energy profiles
- $\pi$ used for all (charged+neutral hadrons)

Longitudinal and lateral profiles
- Determine $E/\eta$-dependent parameterization of shower properties: start and shape
- Fit parameterization to template and extract best parameters
- Sample energy deposits to create ‘hits’

Shower libraries
- Homogeneous assumption does not hold
- Thresholds/cutoffs for low-$E$ particles
- Geant4-based ‘libraries’ of showers
  - Compression achieved through, e.g. clustering and truncation cutoffs => rescaling
Current fast simulation

Tracking

Simulation
- Parameterizations models for particle-material interactions
- Material effects for, e.g. bremsstrahlung and photon conversions, multiple-scattering; dependent on amount of material traversed

Reconstruction
- Exploit MC generator, i.e. truth, information (ideal detector)
- Skip pattern recognition, seeding and ambiguity treatment
- Manipulate true hits to remove bias in track fitting
- Hit content and efficiencies
- Merging of near-by hits, track-sharing, ‘fake’ tracks
Philosophy

- Should spend majority of time simulating interesting processes, not rest of event (ROE)

Generate ROE once, signal $N_{\text{ReDecay}}$ times

- New $D^0$ decay vertex along fixed momentum vector
- New decay product kinematics
- Same efficiencies and resolution (by construction)

All shapes in agreement but correlations exist

- Effect of statistical uncertainties depend on studied observable
- Sample original events and take all ReDecay’d replicas
- Most observables are unaffected

*CMS simulates all pile-up using FastSim that is mixed with hard-scatter events in SimHits or RecHits. ATLAS is working on MC+MC overlay (see talk from T. Novak).
Mini-summary I

- Parameterized calorimeter simulations used for large fraction of Run-2 MC production
- Combine configurable Fast (simulation + digitization + reconstruction) Chain

- Fast tracking, calorimeter and fast emulation of track reconstruction for significant fraction of Run-2 MC production; 20~100x speed-ups
- Ongoing accuracy/precision improvements, support for Phase-2
Current fast simulation

Mini-summary II

- ReDecay used for O(100M) events already, up to 50x speed-ups; additional 30-50x from frozen showers
- Other ideas: disable subdetectors, fully parameterized detector with Delphes

- Detector simulation based on ROOT VMC
- Intending to have ability to mix multiple full simulation engines with fast simulation kernels
  - Extension to VMC
Future fast simulation

Harmonizing full + fast simulation techniques

Philosophy
- Fully simulate only signal objects or specific subdetector, fast everything else
- Streamline full+fast simulation techniques

Proof-of-concept
- Mixing of two simulation "engines" with toy model
- Check between single and concurrent Geant4 instances

*ATLAS developed the Integrated Simulation Framework (ISF) but a number of difficulties -- e.g. defining scale factors, many types of MC -- are encountered.
Existing methods effective, however...
- Average lateral shower properties, uncorrelated energy fluctuations
- Sub-clusters not well modelled
- Pre-simulated showers => finite library size

Unsupervised learning
- Learn relationships between dataset properties and classify without being told the ‘answer’
  - Principle Component Analysis (PCA)
  - Generative Adversarial Networks (GANs)
  - Variational Auto-Encoders (VAEs)

Exploit accelerators
- Ideal use-case for GPUs and FPGAs
  - e.g. hyperparameter scans
Future fast simulation

**FastCaloSimV2**

**ATL-SOFT-PUB-2018-002**

**Improved energy and shape parameterizations**
- Decorrelate using PCA
- Speed comparable to legacy FastCaloSim
- Add forward calorimeters, improved lateral shower profiles

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>$E_{\text{tot}}$ and $E_{\text{frac}}$ from Geant4 samples</td>
</tr>
<tr>
<td>Transform</td>
<td>Construct cumulative distributions (integrating over all events) from inputs, transformed into Gaussians</td>
</tr>
<tr>
<td>PCA</td>
<td>Construct PCA matrices; first principal component (PC) is rotated in direction of maximum variance</td>
</tr>
<tr>
<td>Bin</td>
<td>Divide input shower samples into 'PCA bins' consisting of equal number of events (i.e probabilities)</td>
</tr>
<tr>
<td>Reiterate</td>
<td>Assign each sample a PCA bin and perform PCA chain again</td>
</tr>
</tbody>
</table>

Before rotation

After rotation
**Future fast simulation**

**Deep generative models**

Generate new samples following probabilistic distribution as training set
- Try to learn true data distribution -> reproduce it
- Learn 'approximation function' of true distribution

**VAEs**
- Maximize lower bound of data log-likelihood
  - Generate samples with variations
  - Depend on learning probability density
  - Low-dimensional latent space => blurry images output

**GANs**
- Equilibrium between generator and discriminator
- High-quality images output but complicated to work with

---

Joint HSF/OSG/WLCG Workshop (2019-03-20)  Vincent R. Pascuzzi | vpascuzz@cern.ch
Future fast simulation

DL efforts at the LHC

HSF Simulation Working Group

CERN openlab

ALICE

condDCGAN

3D convolutional GANs

- Inference testing on GPUs, FPGAs
- TensorFlow+MKL-DNN
  - Speedups ~ 10000
  - 1ms/shower

- Performance measured i7-6850K (single core)
- Anticipate additional order of magnitude with Titan XP

<table>
<thead>
<tr>
<th>Method</th>
<th>Machine</th>
<th>Time/Shower (msec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Simulation (geant4)</td>
<td>Intel Xeon Platinum 8180</td>
<td>17000</td>
</tr>
<tr>
<td>3d GAN (batch size 128)</td>
<td>Intel Xeon Platinum 8180 (TF 1.4)</td>
<td>7</td>
</tr>
</tbody>
</table>
Future fast simulation

**DL efforts at the LHC**

**VAEs and GANs**

- Single $\gamma, E = 65$ GeV, $0.2 < |\eta| < 0.25$
- Challenges posed by layers with small (and sparse) energy deposits

"For image generation we are usually happy if the result looks like it is desired...In science we need the result to reasonably well match the given set of requirements."
Summary

With an increasing (HL-)LHC luminosity, we need to look towards fast simulations

- Some techniques employed for nearly a decade, worked well
  - Parameterizations, reduced geometries, other simplifications
  - We already know these won’t suffice in the future
  - Some cases accuracy is also a problem

Next generation of fast simulation requires new strategies

- Integration of full+fast techniques into harmonized framework
- Much effort ongoing to exploit advanced statistical techniques and fast hardware
  - Machine learning
  - Accelerators and HPC

Deep learning shows promise

- LHC experiments devoting human-power into generative algorithms
  - Detector-agnostic implementation, use of GPUs, FPGAs, HPC
  - Still some R&D before hitting mainstream/production
  - Scientific requirements and considerations are crucial
Backup

(The joke: Calvin Pickard is/was a backup goalie.)

cbc.ca/sports/hockey
Projected LHC programme

LHC has been increasing luminosity since its inception
Current fast simulation

Ongoing work

Recycling: MC+MC overlay

Streamlining full+fast simulation techniques
Future fast simulation
Future fast simulation

Fast Digitisation for Inner Detector  

- Most time-consuming (~50%) due to high density of hits
- Implemented for Pixel, SCT, and TRT detectors

**Pixel and SCT**

- Estimate charge by projecting simulated track length onto readout surface (for each readout channel)
  - Corrections made for Lorentz angle drift
- Charge and track lengths smeared to account for multiple scattering
- Form clusters directly from track information (no cluster-finding algorithms)

**TRT**

- Emulate response using distance of nearest approach
- Smear for uncertainty estimate
- Parameterisation allows for particle ID
Future fast simulation

Fast(/truth-assisted(/seeded)) Reconstruction (/Pseudo-tracking)

- Build particle trajectories from digitised hits
  - Combinatorial in nature, exponential increase with increasing pile-up
    - Dominated by Inner Detector

Methodology

- Emulate effect of standard algorithms using MC generator (i.e. truth) information
- Skip time-consuming pattern recognition
- Manipulate hit content and efficiency
- Apply similar selection on criteria used in standard tracking