Monte Carlo simulations for future collider studies

S. Chekanov
+ contributors (www link)

HEP/ANL

ICHEP 2016. Chicago

Aug. 4, 2016
## Timeline of particle collision experiments

<table>
<thead>
<tr>
<th>Year</th>
<th>Experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>2020</td>
<td>LHC Run 3</td>
</tr>
<tr>
<td>2025</td>
<td>HL-LHC</td>
</tr>
<tr>
<td>2030</td>
<td>ILC/CLIC ?</td>
</tr>
<tr>
<td>2035</td>
<td>HE-LHL? LHeC?</td>
</tr>
<tr>
<td>2040</td>
<td>CEPC-ee (2028-)?</td>
</tr>
<tr>
<td>2045</td>
<td>CEPC -SppC ?</td>
</tr>
</tbody>
</table>

In the next decade we will deal with explorations of physics reach, detector parameters and new technology options for post-LHC era.

Requires detailed simulation of physics processes and detector responses.
Simulations for particle-collision experiments

**Physics modeling**
- Known particle properties & established Standard Model (SM)
- Event generators for Standard Model and beyond (LO, NLO, NNLO, NLO matched to NLO)

**Simulation of detector response**
- Interactions of particles with detector material
- Many parametrized cross sections (when exact theory is unknown)
- Simulation packages (Geant4, etc.)

**Detector geometry**

**Computing (HPC, grid)**

**Performance studies (analysis)**
Why do we need simulations? Higgs example

- 100 TeV collider will hunt for M~30 TeV particles decaying to Higgs/W/Z bosons
- Completely new kinematic regime → very challenging for detector designs
- The detector must be optimized to reconstruct Higgs with pT>1 TeV

**SM predictions:** \( \sim 100,000 \) Higgs / ab\(^{-1}\) for pT>1 TeV

**Just kinematics:**
- \( p_T(H)>2 \) TeV \( \rightarrow \sim 5 \) deg between \( \gamma \)'s
- \( p_T(H)>10 \) TeV \( \rightarrow \sim 1 \) deg between \( \gamma \)'s

**Instrumental challenges:**
- identify 2 photons separated by 1 degree
- reject \( \pi^0 \rightarrow \gamma\gamma \) background at the same time!
- similar problems for electron, b-jets decays

Monte Carlo simulations for future collider experiments.  S.Chekanov (ANL)
Next step after Snowmass 2013: Public Repository with Fast and Full Monte Carlo simulations

MC models → EVGEN (event-generator level) → web-optimized archives with theoretical data

- Non-proprietary software
- Open data access
- Simple deployment on personal computers (Windows / Linux / Mac)

Fast detector simulation → Full (Geant4-based) simulation with easy-to-use detector description

Simulation & event reconstruction

OPEN ACCESS http://
Long-term availability & preservation
HepSim event simulations
http://atlaswww.hep.anl.gov/hepsim/

<table>
<thead>
<tr>
<th>Generator</th>
<th>Name</th>
<th>Process</th>
<th>Topic</th>
<th>Info</th>
<th>Link</th>
<th>Created</th>
</tr>
</thead>
<tbody>
<tr>
<td>PYTHIA8</td>
<td>pp 100</td>
<td>Higgs production</td>
<td>Higgs</td>
<td>Info</td>
<td>URL</td>
<td>2016/01/07</td>
</tr>
<tr>
<td>MADGRAPH/HW6</td>
<td>pp100</td>
<td>Higgs+ttbar (NLO+PS)</td>
<td>Higgs</td>
<td>Info</td>
<td>URL</td>
<td>2015/11/13</td>
</tr>
<tr>
<td>FPQC</td>
<td>pp 8</td>
<td>Exclusive WW production</td>
<td>SM</td>
<td>Info</td>
<td>URL</td>
<td>2015/03/23</td>
</tr>
<tr>
<td>HERWIG++</td>
<td>pp 100</td>
<td>QCD djets, pT&gt;2700 GeV</td>
<td>SM</td>
<td>Info</td>
<td>URL</td>
<td>2015/04/11</td>
</tr>
<tr>
<td>HERWIG++</td>
<td>pp100</td>
<td>QCD djets, pT&gt;300 GeV</td>
<td>SM</td>
<td>Info</td>
<td>URL</td>
<td>2015/04/10</td>
</tr>
<tr>
<td>HERWIG++</td>
<td>pp100</td>
<td>QCD djets, pT&gt;900 GeV</td>
<td>SM</td>
<td>Info</td>
<td>URL</td>
<td>2015/10/03</td>
</tr>
<tr>
<td>HERWIG++</td>
<td>pp100</td>
<td>QCD djets, pT&gt;2700 GeV</td>
<td>SM</td>
<td>Info</td>
<td>URL</td>
<td>2016/01/07</td>
</tr>
<tr>
<td>PYTHIA8</td>
<td>pp100</td>
<td>QCD djets, pT&gt;8 TeV</td>
<td>SM</td>
<td>Info</td>
<td>URL</td>
<td>2015/10/21</td>
</tr>
<tr>
<td>MADGRAPH/HW6</td>
<td>pp 100</td>
<td>Top</td>
<td>Info</td>
<td>URL</td>
<td>2015/11/13</td>
<td></td>
</tr>
<tr>
<td>MADGRAPH/HW6</td>
<td>pp 100</td>
<td>Top</td>
<td>Info</td>
<td>URL</td>
<td>2015/04/10</td>
<td></td>
</tr>
</tbody>
</table>

Available: EVGEN files (LO,NLO, etc), fast simulations, full Geant4 simulations

NERSC, CERN mirrors
LHC run 1/2
HL-LHC
SPPC, FCC-hh
ILC, CEPC
samples for detector performance studies
Monte Carlo simulations for future collider experiments. S. Chekanov (ANL)

HepSim repository. How it works

Event Generators

- PYTHIA6
- PYTHIA8
- HERWIG++
- Madgraph5
- MCFM
- JetPhox
- FPMC
- NLOjet++
- LEPTO/Ariadne

large-scale computing resources

Event Generators

PYTHIA6
PYTHIA8
HERWIG++
Madgraph5
MCFM
JetPhox
FPMC
NLOjet++
LEPTO/Ariadne

Event Generators

PYTHIA6
PYTHIA8
HERWIG++
Madgraph5
MCFM
JetPhox
FPMC
NLOjet++
LEPTO/Ariadne

ProMC files on several public web servers

- Delphes fast simulation (ROOT files)
- SLIC (Geant4) full simulation and reconstruction software (LCIO files)

HepSim

- index files
- create metadata
- prepare for batch download
- validate with Jython scripts
- create search database
Software for full simulations

Simulator for the Linear Collider (SLIC) software

- Optimized for the SiD detector at SLAC (T.Johnson, N.Graf, J.McCormick, J.Strube)
- Re-purposed for future pp collider studies (S.C., A.Kotwal, J.Strube)
- Integrated with HepSim. Deployed on Open-Science Grid (OSG)
- Analysis: C++/Root or **Jas4pp** (ANL,S.C,E.May). Based on Jas3 (SLAC)

SLIC (Geant4, version 10.1) → Track reconstruction → Pandora PFA → Analysis (Jas4pp)

- Compact geometry description (XML) and LCDD Geometry
- Ongoing optimization

http://atlaswww.hep.anl.gov/hepsim/

Monte Carlo simulations for future collider experiments. S.Chekanov (ANL)
HepSim event statistics
(excluding fast and Geant4 detector simulations)

~210 Monte Carlo samples
~1.6 billion EVGEN events
  ~ 10% after fast simulations (Delphes)
  ~ 0.1% after Geant4 simulations

<table>
<thead>
<tr>
<th>Event Type</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of public file servers</td>
<td>3</td>
</tr>
<tr>
<td>Number of event samples</td>
<td>208</td>
</tr>
<tr>
<td>Number of NLO samples</td>
<td>17</td>
</tr>
<tr>
<td>Number of NLO+PS samples</td>
<td>17</td>
</tr>
<tr>
<td>Number of LO (+PS) samples</td>
<td>144</td>
</tr>
<tr>
<td>Number of events</td>
<td>1560741507</td>
</tr>
<tr>
<td>NLO events</td>
<td>5830000000</td>
</tr>
<tr>
<td>NLO+PS events</td>
<td>32860595</td>
</tr>
<tr>
<td>LO (+PS) events</td>
<td>859498212</td>
</tr>
<tr>
<td>Total size (GB)</td>
<td>6897.468</td>
</tr>
<tr>
<td>NLO size (GB)</td>
<td>238.06</td>
</tr>
<tr>
<td>NLO+PS size (GB)</td>
<td>348.693</td>
</tr>
<tr>
<td>LO (+PS) size (GB)</td>
<td>6292.482</td>
</tr>
<tr>
<td>Number of files</td>
<td>334606</td>
</tr>
</tbody>
</table>

Platforms for event generations (EVGEN)
- 10% → BlueGene/Q (ANL/Mira) (Jetphox, MCFM)
- 50% → HEP-ANL (mainly Madgraph)
- 40% → OSG-CI grid and USATLAS CI (for phase II)
Monte Carlo simulations for future collider experiments. S.Chekanov (ANL)

CPU usage for SLIC (Geant4) simulations

OSG-Connect “FutureColliders” project for HepSim jobs

HepSim simulations (SLIC/Geant4)
~ 16 GB RAM per job

2.5 million CPU*h in 2016 using OSG-grid for Geant4 simulations
(equivalent to ~10 million CPU*h on HPC BlueGene/Q Mira)
SiD detector for ILC

- Multi-purpose detector for the ILC
- Conceived at SLAC (USA LC Physics Group)
- The key characteristics:
  - 5 Tesla solenoid
  - Silicon tracker: 25/50 um readout pitch
  - ECAL: (0.35 cm cell size, W / silicon)
  - HCAL:
    - 1x1 cm cell size (RPC for LOI3*)
    - 40 layers for barrel (HCAL) ~4.5 $\lambda_i$
- Optimized for particle-flow algorithms (PFA)
- Fully configurable using SLIC software

→ See talk by Andrew White
‘All-silicon’ design concepts supported in HepSim

SiD
\((e^+e^- \text{ up to } 1 \text{ TeV})\)

SiCPEC
\((e^+e^- \text{ 250 GeV})\)

SiEIC
\((ep, 141 \text{ GeV})\)

SiFCC
\((\text{FCC-hh, pp 100 TeV})\)

- 1 T solenoid
  ~3 \(\lambda_i\) HCAL
  Smaller size

- 4 T solenoid
  ~4 \(\lambda_i\) digital HCAL

- 5 T solenoid
  ~12 \(\lambda_i\) Scint/Fe HCAL
  Larger size

HepSim provides single-particle and physics events after simulation & reconstruction

Monte Carlo simulations for future collider experiments. S.Chekanov (ANL)
Example I: FCC-hh detector studies

- See the ICHEP presentations by Shin-Shan Yu
- ICHEP16 posters by Nhan. V Tran and Sourav Sen

Large number of back-splash interactions in ECAL/HCAL
~4 CPU*h to simulate/reconstruct, 16 GB RAM
→ CPU intensive!

Studies of tracking at multi-TeV scale:
(J.Zuzelski, S.C., J.Proudfoot)

Collaboration to study a high-granularity calorimeters:
A.Kotwal (Fermilab/Duke), L.Gray (Fermilab), J.Strube (PNNL), N.Tran (Fermilab), S. Yu (NCU), S.Sen (Duke), J.Repond (ANL), J.McCormick (SLAC), J.Proudfoot (ANL), A.M.Henriques Correia (CERN), C.Solans (CERN), C.Helsens (CERN), S.C, etc.

Studies of jets and tracks up to ~20 TeV!
Example II: Designing a detector for CEPC (e+e- 250 GeV)

SiD detector is designed for ~500 GeV particles/jets (0.5-1 TeV CM energy)
But CEPC will measure particles/jets up to 125 GeV (250 CM energy)

Possible optimizations:

- **HCAL:**
  - barrel: 4.5 $\lambda_1$ (40 layers) → 4.0 $\lambda_1$ (35 layers)
  - endcap: 5 $\lambda_1$ (45 layers) → 4.0 $\lambda_1$ (35 layers)

- **Tracking:**
  - 5 Tesla → 4 Tesla

40 layers ~ 4.5 $\lambda_1$

**Possible optimizations:**

Design a light, cost-optimized version of the SiD detector for CEPC and use several physics processes to benchmark its performance

Example II: Simulations for CEPC (e^+e^- 250 GeV CM energy)

Available full simulations for the SiD and SiDCC (for CPC) detectors:
- Z → e^+e^-, Z → tau tau, Z → mu^+mu^-, Z → b\bar{b}
- H(125) → b\bar{b}, H(125) → γγ, H(125) → ZZ^* → 4l, H(125) → tau^+tau^-
Comparing SiD with SiDCC1

- Benchmark processes for e+e- (250 GeV)
  - $Z \to e^+e^-$, $Z \to b\bar{b}$, $Z \to \tau^+\tau^-$, $H \to \gamma\gamma$
  - $H \to 4l$, $H \to b\bar{b}$, QCD jets

- Particle flow objects to reconstruct invariant masses and jet energy resolutions (Durham jets)

Simplification of the SiD detector does not compromise physics performance

Summary

- First public Monte Carlo repository with fast and full detector simulations
- Enable physics & detector-performance studies for current & future colliders + community outreach
  - 1.6 billion events at the EVENT level for public downloads
  - Significant number of fast and fully reconstructed events for $ep$, $\mu\mu$, $ee$, $pp$ collisions (13-100 TeV) & single-particle samples for detector studies

- 14 articles, ~25 presentations since 2014 (linked to WWW):
  - Physics reach studies for HL-LHC, HE-LHC, FCC-hh etc.
  - Calorimeter studies (cell granularity)
  - Tracking optimization at multi-TeV scale
  - Software development

- 2.5 million CPU*h from OSG-grid. OSG-Connect support from UChicago
- Contributions from 17 students/scientists
- Your contribution is welcome!

Some studies presented at this meeting!
Backup
Long-term preservation of theoretical calculations

- Storing Monte Carlo predictions in files makes sense if:

\[
\frac{\text{time to download & analyze on commodity computer}}{\text{CPU*h needed to create the prediction}} \equiv \epsilon \ll 1
\]

- \(\epsilon \sim 0.01-1\) - for LO MC
- \(\epsilon \ll 0.01\) - for NLO etc.

\[\epsilon \ll 1:\]
- Madgraph5 etc. (NLO+PS+hadronisation), ALPGEN
- Some fast-converging NLO calculations (MCFM, jetPHOX etc)
- MC with \(\epsilon \sim 1\) but after mixing with pile-up (CPU intensive)

\[\epsilon \sim 1:\] Less appropriate approach for:
- LO simulations (Pythia)
- Some NLO programs with slow convergence
  - requires too large data volumes to keep weighted events
New data format for EVGEN: **ProMC**

- “Archive” self-described format to keep MC events:
  - Event records, NLO, original logfiles, PDG tables etc.
- 30% smaller files than existing formats after compression

- Number of used bytes depends on values. Small values use small number of bytes

- Effective file size reduction for pile-up events
  - Particles with small momenta → less bytes used
- Installed on Mira (BlueGene/Q)
- Separate events can be streamed over the Internet:
  - similar to avi frames for web video players

http://atlaswww.hep.anl.gov/asc/promc/

*S.C., E. May, K. Strand, P. Van Gemmeren, Comp. Physics Comm. 185 (2014), 2629*
Dataset entry: \textbf{e+e- collisions (CM energy = 250 GeV). Z → e+e-}

**Repository with Monte Carlo predictions for HEP experiments**

**Information about "gev250ee_pythia6_zpole_ee" dataset**

- **Name:** gev250ee_pythia6_zpole_ee
- **Collisions:** e+e-
- **CM Energy:** 250 TeV
- **Entry ID:** 146
- **Topic:** SM
- **Generator:** PYTHIA6
- **Calculation level:** LO+PS+hadronisation
- **Process:** Z boson to e+e-
- **Total events:** 2000000
- **Number of files:** 100
- **Cross section (fb):** 1.7765 ± 0.0126 pb
- **Luminosity (fb^{-1}):** 1.1256+06 pb^{-1} (or) 1125.7949 fb^{-1} (or) 1.1258 ab^{-1}
- **Format:** ProMC
- **Submission date:** Tue Oct 13 14:28:55 CDT 2015
- **Download URL:** http://mc.hep.anl.gov/asc/hepsim/events/ee/250gev/pythia6_zpole_ee

**Validation distributions created using Python scripts on the Java platform**

**Run via JavaWeb start by streaming data over the Web**

**URL for EVGEN files (download or streaming)**

**URL with fast (DELPHES) or Geant4 (SLIC) simulations**
Benchmarks for EVGEN files

File sizes for 10,000 $t\bar{t}$ events for pp at LHC

<table>
<thead>
<tr>
<th>File format</th>
<th>File Size (MB)</th>
<th>C++ (sec)</th>
<th>CPython (sec)</th>
<th>Java (sec)</th>
<th>Jython (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ProMC</strong></td>
<td>307</td>
<td>15.8</td>
<td>980</td>
<td>11.7 (12.1 + JVM startup)</td>
<td>33.3 (35 + JVM startup)</td>
</tr>
<tr>
<td><strong>ROOT</strong></td>
<td>423</td>
<td>20.4</td>
<td>66.7 (PyROOT)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>LHEF</strong></td>
<td>2472</td>
<td>84.7</td>
<td>30.4</td>
<td>9.0 (9.6 + JVM startup)</td>
<td>-</td>
</tr>
<tr>
<td><strong>HEPMC</strong></td>
<td>2740</td>
<td>175.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>LHEF</strong> (gzip)</td>
<td>712</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>LHEF</strong> (bzip2)</td>
<td>552</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>LHEF</strong> (lzma)</td>
<td>513</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>HEPMC</strong> (gzip)</td>
<td>1021</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>HEPMC</strong> (bzip2)</td>
<td>837</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>HEPMC</strong> (lzma)</td>
<td>802</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1. Benchmark tests for reading files with 10,000 $t\bar{t}$ events stored in different file formats. For each test, the memory cache on Linux was cleared. In case of C++, the benchmark program reads complete event records using appropriate libraries. CPython code for ProMC file is implemented in pure CPython and does not use C++ binding (unlike PyROOT that uses C++ libraries). In case of LHEF files. JAVA and CPYTHON benchmarks only parse lines and tokenize the strings, without attempting to build an event record, therefore, such benchmarks may not be accurate while comparing with ProMC and ROOT.