TrackML: a LHC Tracking Machine Learning Challenge

*Paolo Calafiura* (LBNL),

David Rousseau (LAL), Cecile Germain (Paris Sud), Vincenzo Innocente (CERN), Riccardo Cenci (Pisa), Michael Kagan (SLAC), Isabelle Guyon (ChaLearn), David Clark (UCB), Steve Farrell (LBNL), Rebecca Carney (Stockolm), Andreas Salzburger (CERN), Davide Costanzo (Sheffield), Markus Elsing (CERN), Tobias Golling (Geneve), Tony Tong (Harvard), Jean-Roch Vlimant (Caltech)
Tracking @ HL-LHC

Expect $O(10x)$ increase in:

- Intensity $\rightarrow$ Tracks/event
- Trigger rate $\rightarrow$ Number of events

$O(100X)$ more tracks/sec wrto LHC run 2

Hardware evolution:

$O(10x)$ more transistors $\rightarrow$ cores

Net result:

$O(10x)$ CPU deficit

**IF** can increase tracking parallelism $O(10x)$
Why ML? Why Now?

Computationally regular, adaptive approximations of non-linear phenomena

Source: Turner et al NIPS 2014

Natural to vectorize and parallelize
The TrackML Challenge

An idea born across the Bay 18 months ago, at Connecting the Dots 2015

Goal: speedup 10x HL-LHC track formation

Wider Benefits:

Engage ML community

Foster cross-experiment collaboration:
  • Generate public domain, shared HL-LHC tracking datasets
  • Develop shared methodology to evaluate tracking performance

Source: A. Salzburger
Components of a Machine Learning Challenge

1. **Starting Kit:**
   - A compelling description of the problem to solve.
   - Software needed to ingest datasets.
   - May include simple reference solution to guide competitors

2. **Datasets**
   - Training, Validation, Testing

3. **Figure of Merit:**
   - Quantitative assessment of a solution, used to grade competitors

4. **Organizer and Host Platform**
   - Manage challenge, provide/award prize, follow-up
The Problem to Solve

- **10^5 Hits**
  - Triplet Generation
- **10^9 Triplets**
  - Seed Selection
- **10^5 Seeds**
  - Track Finding
- **10^4 Tracks**

**Source:** A. Salzburger
In practice…

Challenge: given a list of 3D hits, return a list of tracks, each track being a list of 3D hits.
Datasets

Use aCTS (Track 2, Wed 11.30) to produce samples. We are adding modules to:

• Read in realistic HL-LHC events (pile-up!)

• Write out 3D hits in “findable” tracks

Estimate $O(10^B)$ Tracks $\rightarrow$ $O(1M)$ events $\rightarrow$ $O(1TB)$

Format:

$\text{[Event ID,}$
$\text{[Track ID,}$
$\text{[x, y, z}]])$

training only
Benchmarking the Solutions

Solution: \([\text{Event: [Track: [(x,y,z)]]]}\)

Three ingredients for the figure of merit:

1. Efficiency (fraction of tracks found)
   - must be uniformly high for all “findable” tracks, not just on average.

2. Fake rate (fraction of tracks “invented” from random combination of points)

3. Processing time (wall clock/track)
Measuring Processing Time

Not a concern for most ML challenges. For example Kaggle does not support it.

Very much a concern for us, but definition less straightforward than one may think:

- **Latency** (time per PU, trigger) vs **Throughput** (aggregated time, offline)

- Which **target platform**? Xeon, KNL, GPU, FPGA…

Provide a reference platform, help participants to test their solutions on it.
Software environment on reference platform must not bias for/against a certain paradigm (e.g. CNN) or tookit (e.g. theano).
Organization, Next Steps

Informal, open collaboration
Join by subscribing to trackml-challenge@googlegroups.com

Share broad goals and several collaborators with
- aCTS
- HEP.TrkX new DOE pilot project, provide framework to develop and evaluate LHC tracking algorithms, particularly ML ones.

Moving from discussions to prototyping stage to first decisions on dataset format, generic detector configuration

Lots remains to be done, e.g. in the area of performance measurement