Track finding in ATLAS using GPUs

Johannes Mattmann and Christian Schmitt on behalf of the ATLAS Collaboration

GPUs at a glance / Track reconstruction principle

The following image is a schematic representation of the current Nvidia CUDA architecture. The GPU contains a set of "streaming multiprocessors" which are equipped with local storage and control units. Each one again contains a number of CUDA cores.

Track reconstruction

These cuda cores are computing elements containing each a floating point and an integer storage and control units. Each one again the GPU contains a set of "streaming multiunit and some control logic. Operations are executed in parallel on a subset of CUDA cores simultaneously. Access latencies to global memory or special functions (sin, cos, sqrt,...) can be covered by hardware thread handling allowing for quasi-lossless thread switches. A hierarchy of memory areas lets the developer adjust data locality vs. memory requirements.

Seed finder

Currently hit triples on the three innermost central barrel layers are checked if they form a viable seed for track finding. To reduce combinatorics by a priori excluding invalid seed candidates, seed search is constrained to a defined set of angular and longitudinal segments on each layer.

Propagation and Kalman filter

In general the implementation requires a CPU/ GPU interplay (see outer left figure). For each hit data (local x/y position, errors, detector identifier) and reconstructed seeds are transferred to global device memory. Propagation and Kalman filter application is performed in parallel on a per layer basisin each step. Afterwards all found tracks are reconstructed from the track parameters on each layer:

- create new track object
- starting on outermost layer: follow "origin references" of each parameter set, add all subsequent parameters
- use resulting parameters as new initial parameters for next step

Results and Outlook

Runtime comparison seed finder:
- for entire test range of single muon and simulated ttbar sample
- CPU outperforms GPU for more than 15 single muon tracks
- fit result for \( N_{\text{tracks}} \geq 100 \):
  - single core CPU: \( \text{speedup } \approx 13.7 \pm 0.5 \text{ const. speedup} \)

Propagation/Kalman filter:
- working for single muon tracks
- higher runtime fluctuations, X server state dependence observable
- for \( N_{\text{tracks}} \geq 8 \): CPU/GPU time ratio run disproporionate: 165 ± 30 max. speedup factor (50 single µ tracks)

Outlook:
- extend current model to full feature set: include endcaps and transition region, implement all special cases from official framework (in progress, to be finished by the end of 2012)
- implement optimized seed finder; thus optimize memory usage and speed of execution
- build an official ATLAS reconstruction software package from GPU implementation code
- promote equipment of data centers with GPUs during shutdown period