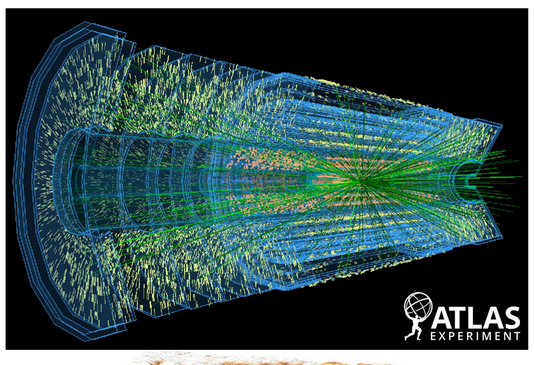
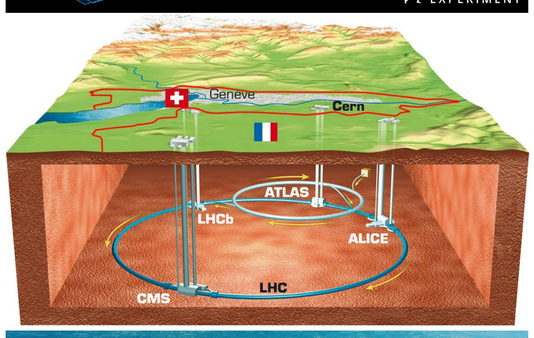
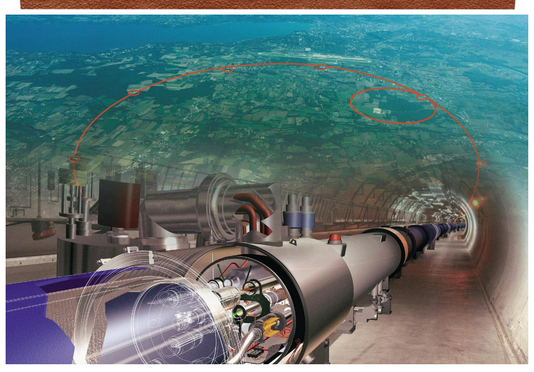
Real-Time GPU Kalman-Filter Tracking via Kernel Refactoring and INT8 Surrogates for High-Luminosity Colliders

Hao-Chun Liang*, Yuan-Tang Chou†, Bo-Cheng Lai*
National Yang Ming Chiao Tung University, Hsinchu, Taiwan*, University of Washington, USA†

Motivation & Requirement



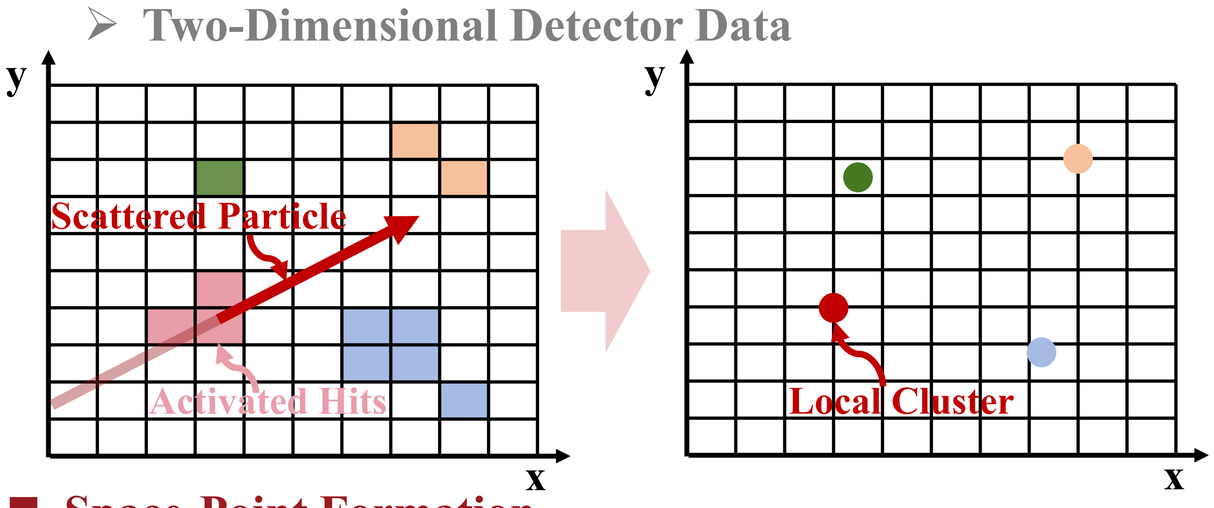




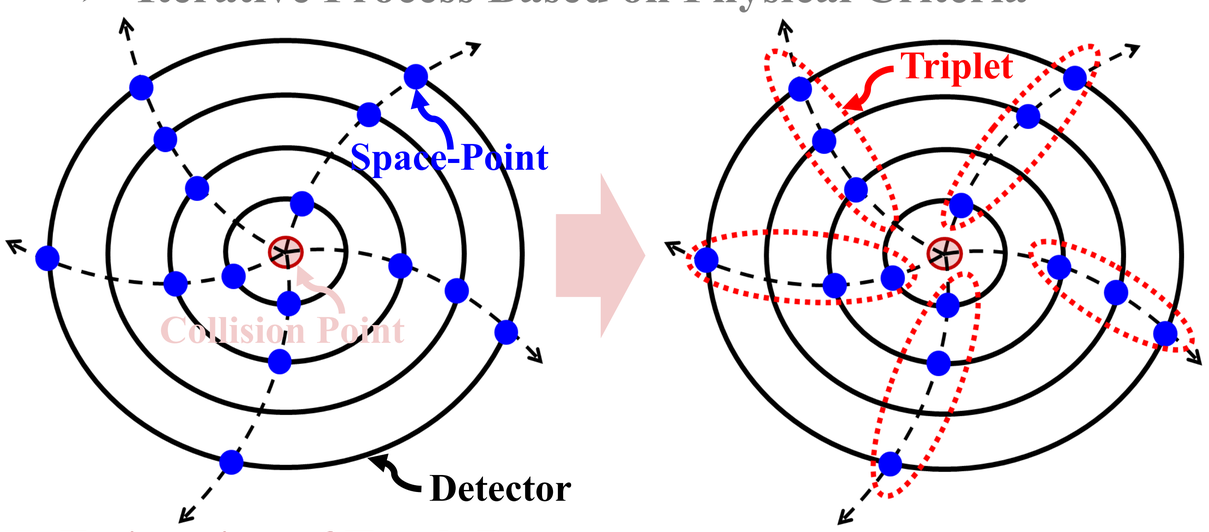
- **ATLAS Experiment**
  - ***Proton-Beam Collisions***
    - Produce Hits on Detectors
    - Extreme High Collision Rate
    - Require Selective Filtering
- **Overall Selective Filtering System**
  - Trajectory Reconstruction
  - Essential for Managing Data
  - ***(Step.1) Readout System***
    - Input: Continuous 40MHz
    - Buffer Depth: 10μs
  - ***(Step.2) Level-0 Trigger System***
    - Latency ≤ 10μs
    - Output Rate: 1MHz
  - ***(Step.3) Data Acquisition System***
    - Pipelined and Buffered
    - Throughput: 4.6TB/s
    - Event-Size: ~4-5 MB
  - ****(Step.4) Event Filtering System***
    - No Latency Constraint
    - Throughput: 50-60 GB/s
    - Event-Size: ~5-6 MB

Event Filtering CUDA Pipeline

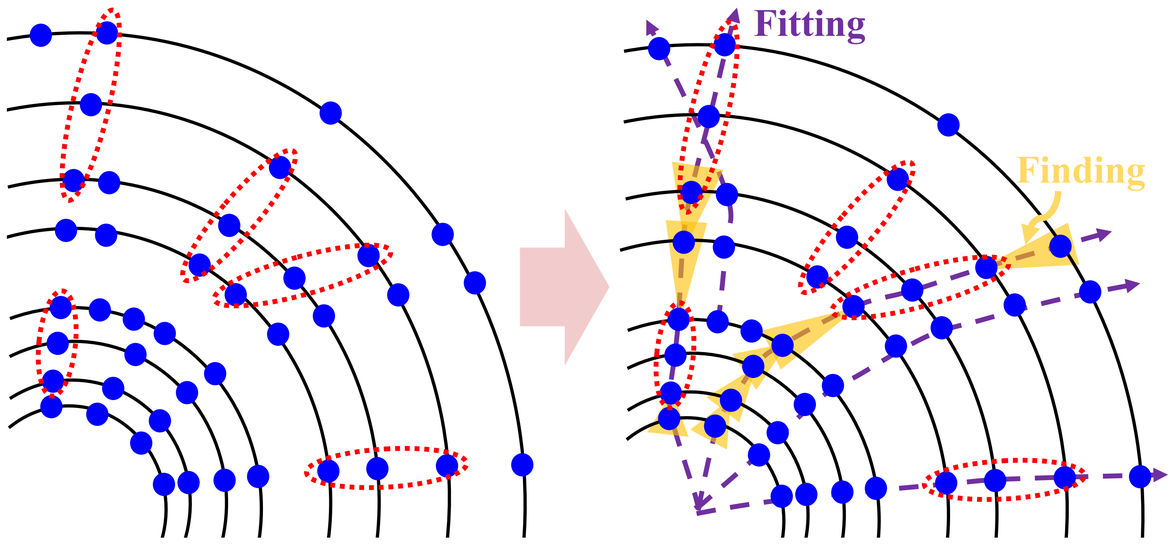
- **Sparse Connected Component Labeling (CCL)**
  - Two-Dimensional Detector Data



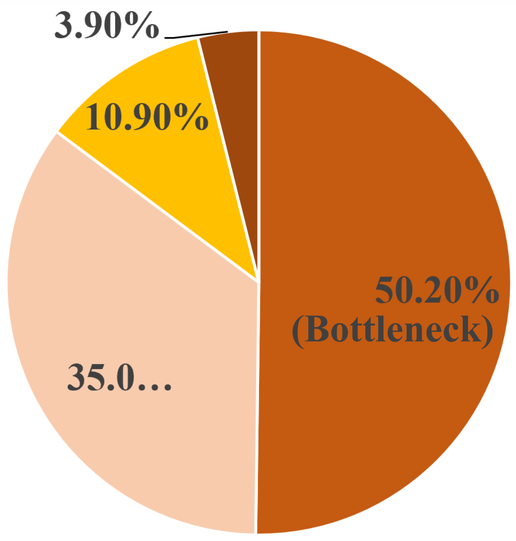
- **Space-Point Formation**
  - Local Two-Dimensional Clustering in Detectors
  - Transformation from Two to Three Dimensions
- **Seed Identification**
  - Space-Point Pair and then Triplet Formation
  - Iterative Process Based on Physical Criteria



- **Estimation of Track Parameters**
- **Track Finding and then Fitting**
  - Combinational Kalman Filter / Kalman Filter



Profiling of Event Intervals

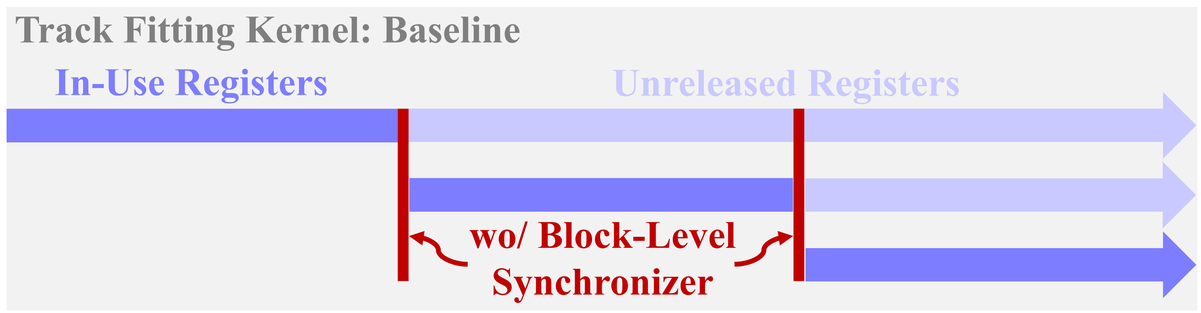


- **Track Fitting Stage: 50.20%**
  - Average Interval: 11.08 ms/event
- **Track Finding Stage: 35.00%**
  - Average Interval: 7.67 ms/event
- **CCL Stage: 10.90%**
  - Average Interval: 2.27 ms/event
- **Summation of Other Stages: 3.90%**
  - Average Interval: 0.86 ms/event

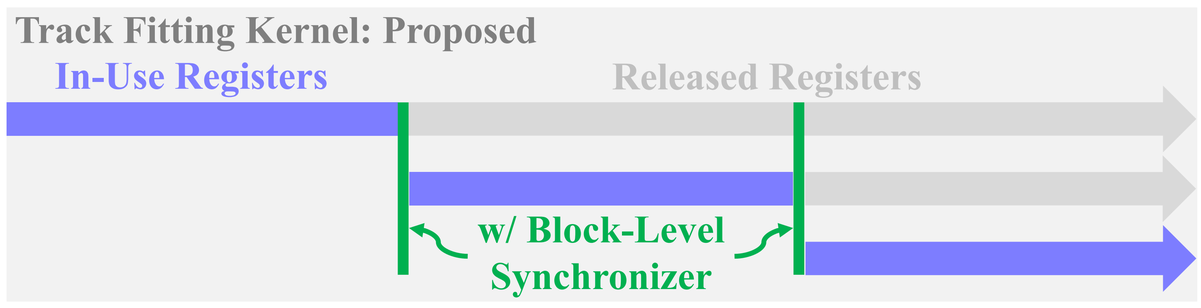
: Track Fitting : CCL : Track Finding : Others

Optimization – Kernel Refactoring

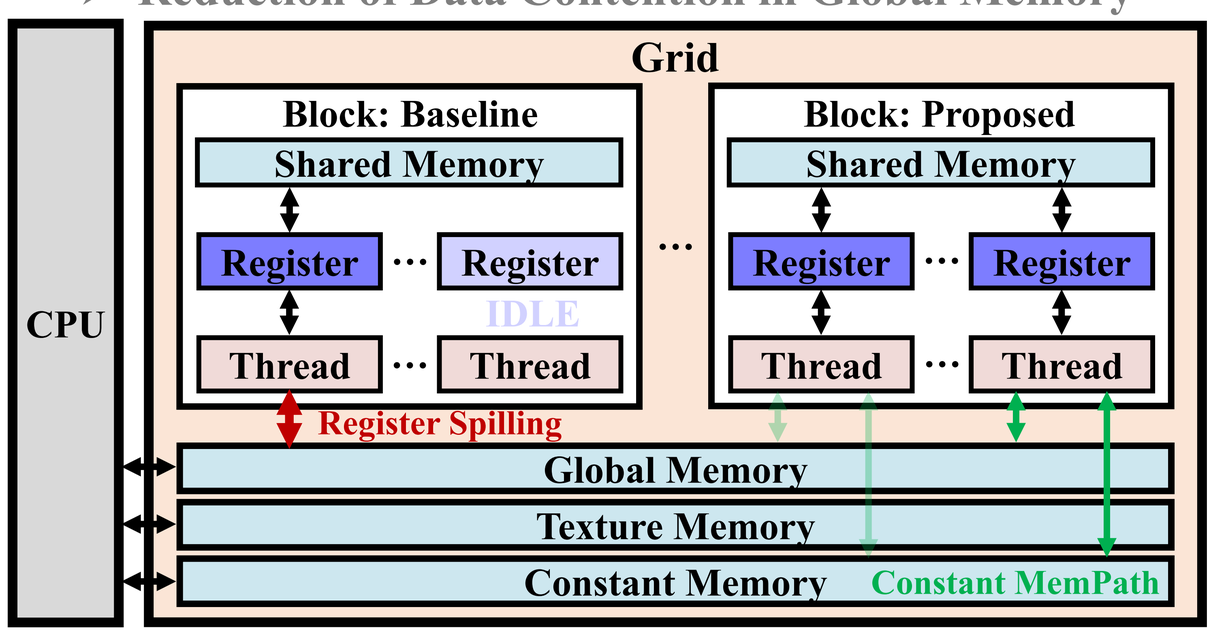
- **Spill-Inducing Register Pressure in Track Fitting Kernel**
  - Long-Lived Variables Occupying Registers
  - Round-Trip Access Latency from Global Memory



- **Phase Separation Using Block-Level Synchronizer**
  - CFG Partition and Compiler Liveness Analysis



- **Utilization of Constant Memory**
  - Reduction of Data Contention in Global Memory



Optimization – INT8 Surrogate

- **Challenges of Matrix Inversion in Kalman Filter**
  - **Serialized Warp-Level Execution**
    - Hardware-Limited Division (#Core : #SFU = 4:1)
    - Long-Latency Division (FP32: 32-48 Cycles)
  - **Divergent Warp-Level Branching**
    - Pivot Swapping for Floating-Point Precision
  - **Replacement with Multi-Layer Perceptron**
    - Teacher-Student Knowledge Distillation
    - Symmetric INT8 Quantization

Conclusion

- **Phase Separation Using Block-Level Synchronizer**
  - **Event Interval Reduction of 15%**
- **Matrix Inversion Replacement with TSKD-MLP**
  - **Event Interval Reduction of 186% with MSE = 8×10^{-5}**