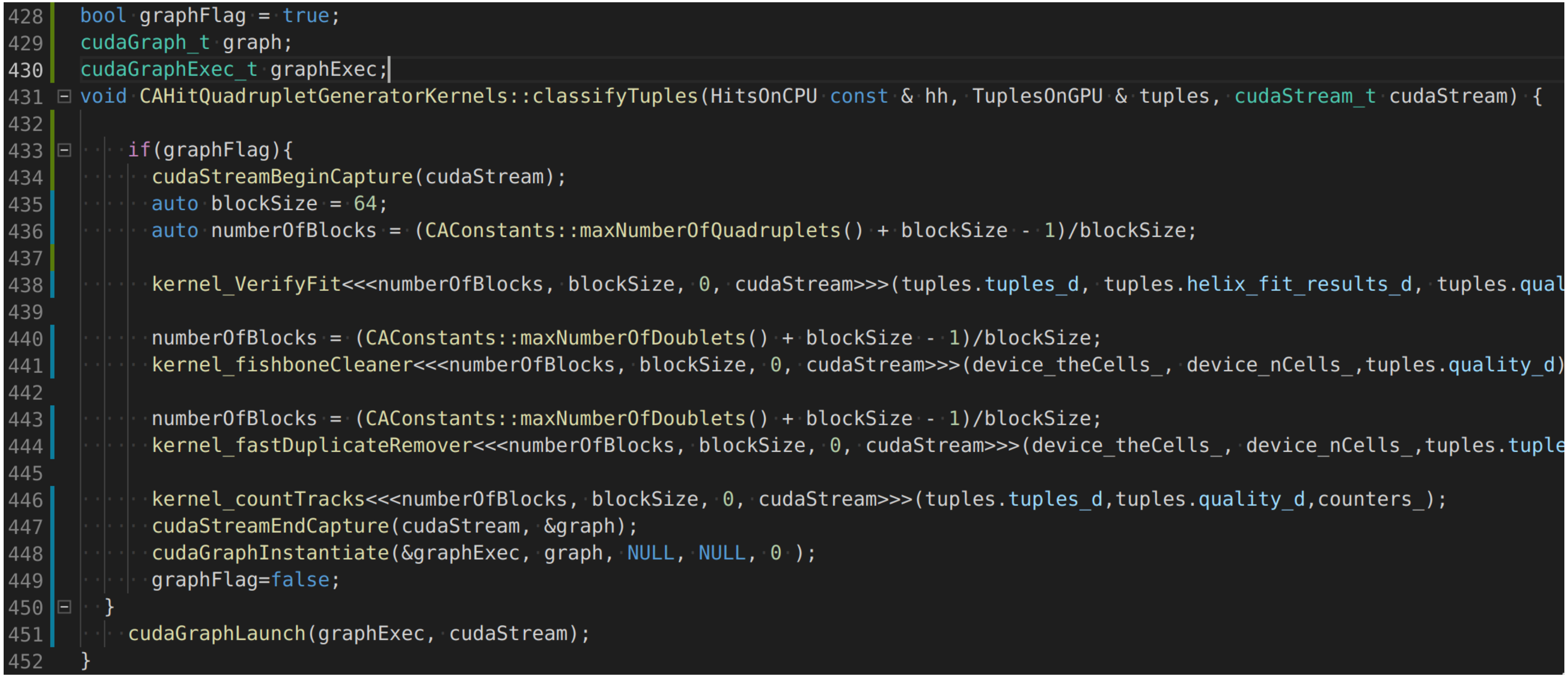
Patatrack - day 1

Andrea, Felice, Marco, Vincenzo, Vishal

CUDA Graph

- Our RAW to Tracks is running many small kernels
- Got a 4% speed up: baseline 1065Hz → 1100Hz



```
bool graphFlag = true;
cudaGraph_t graph;
cudaGraphExec_t graphExec;
void CAHitQuadrupletGeneratorKernels::classifyTuples(HitsOnCPU const & hh, TuplesOnGPU & tuples, cudaStream_t cudaStream) {

    if(graphFlag){
      cudaStreamBeginCapture(cudaStream);
      auto blockSize = 64;
      auto numberOfBlocks = (CAConstants::maxNumberOfQuadruplets() + blockSize - 1)/blockSize;

      kernel_VerifyFit<<<numberOfBlocks, blockSize, 0, cudaStream>>>(tuples.tuples_d, tuples.helix_fit_results_d, tuples.qual

      numberOfBlocks = (CAConstants::maxNumberOfDoublets() + blockSize - 1)/blockSize;
      kernel_fishboneCleaner<<<numberOfBlocks, blockSize, 0, cudaStream>>>(device_theCells_, device_nCells_,tuples.quality_d)

      numberOfBlocks = (CAConstants::maxNumberOfDoublets() + blockSize - 1)/blockSize;
      kernel_fastDuplicateRemover<<<numberOfBlocks, blockSize, 0, cudaStream>>>(device_theCells_, device_nCells_,tuples.tuple

      kernel_countTracks<<<numberOfBlocks, blockSize, 0, cudaStream>>>(tuples.tuples_d,tuples.quality_d,counters_);
      cudaStreamEndCapture(cudaStream, &graph);
      cudaGraphInstantiate(&graphExec, graph, NULL, NULL, 0 );
      graphFlag=false;
  }
    cudaGraphLaunch(graphExec, cudaStream);
}
```

Tomorrow

- Port more kernels to CUDA Graph and try to launch different streams in the same graph
- Understand how to port CUB kernel calls to Graph
- Profile more kernels