Accelerator Service

Matti, Salvatore

Previous Problems

- From “Service Prototype” to “Accelerator Service”:
  - Move code to HeterogeneousCore/AcceleratorService package in CMSSW
  - Deploy on the new machine provided by TechLab;
  - Implement a unit test running a kernel on a real GPU
    - Still synchronously
- Evaluating features of CUDA streams, and possible usage in the AcceleratorService:
  - Using streams as “queue of device work”:
    - The host places work in the queue and continues on immediately
    - Device schedules work from streams when resources are free
  - CUDA operations are placed within a stream
    - Concurrent copy of “host pinned memory”
    - Launching kernel executions
  - Operations within the same stream are ordered (FIFO) and cannot overlap
  - Operations in different streams are unordered and can overlap
  - CPU post-processing with stream callbacks.

Next Problem(s)

- Further improvements for the “Accelerator Service”:
  - Disentangle “Resource definition” from “job scheduling”:
    - A CUDA specific Service able to deal with CUDA streams/workloads
    - Schedule job asynchronously using CUDA streams
  - Run the Raw2Digi step using the Service
    - Possible interface change/extension in order to deal with CUDA streams
  - Code restructuring/cleanup/documentation.