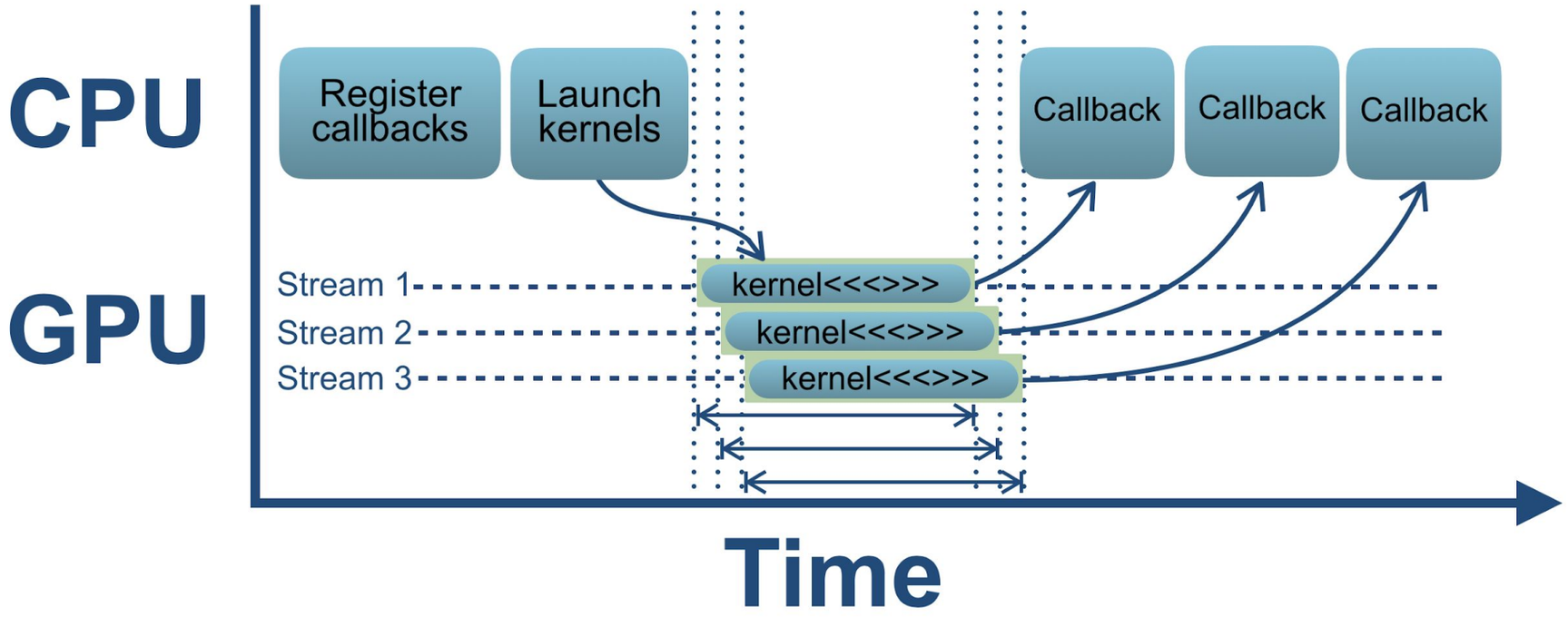
Adopted solution: CUDA Profiling Tools Interface (CUPTI)

- Register callback functions to manage buffer request/delivery of ‘activity records’
- Callbacks triggered by GPU activity
- Access returned records

holds information about GPU or operations on GPU; different kinds for kernels, memory transfers, etc.



Timing with CUPTI Callback functions consists of just a few steps:

- Initialize trace:
  - Enable collection of kernel activity records
  - Register callback functions

```
cuptiActivityEnable(CUPTI_ACTIVITY_KIND_CONCURRENT_KERNEL);
cuptiActivityRegisterCallbacks(bfrRequest, bfrCompleted);
```

- Trigger callback functions; schematically, they look like this:

```
void CUPTI API bfrRequest (uint8_t **bfr, ...)
{
      // Signal to CUPTI client that an empty buffer is needed by CUPTI
}
void CUPTI API bfrCompleted (uint8_t *bfr, ...)
{
      // Return a buffer of completed activity records to CUPTI client
}
```