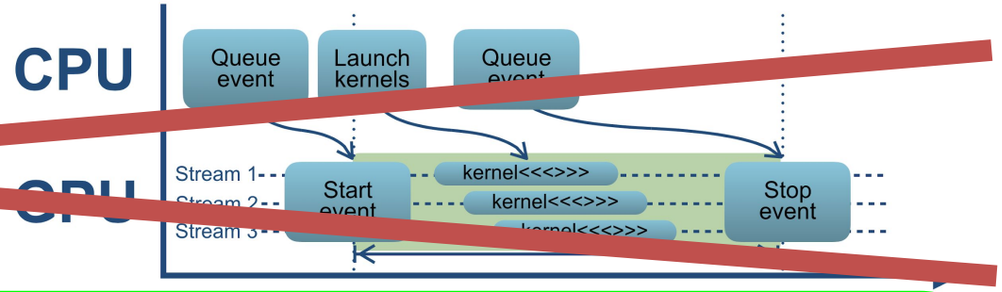
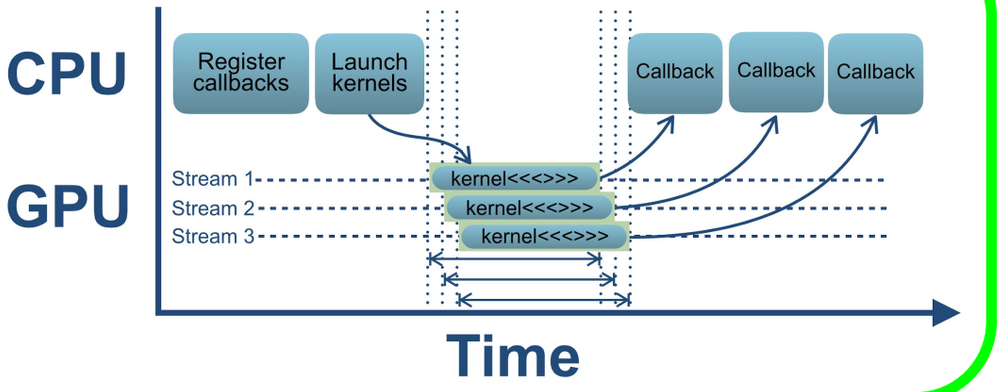
Three possible on-the-fly timing strategies

1. Count GPU clock cycles
   - Requires additional device-to-host transfers
   - Implementation may be invasive
2. CUDA Events
   - Can give ambiguous results

```
1.   Start GPU timer
2.   DO GPU kernel work
3.   Stop GPU timer
4.   Read elapsed time to host
```



3. CUDA Profiling Tools Interface (CUPTI)
   - Buffer requests and delivery of timing information handled by CUPTI
   - Gives unambiguous kernel timings



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.