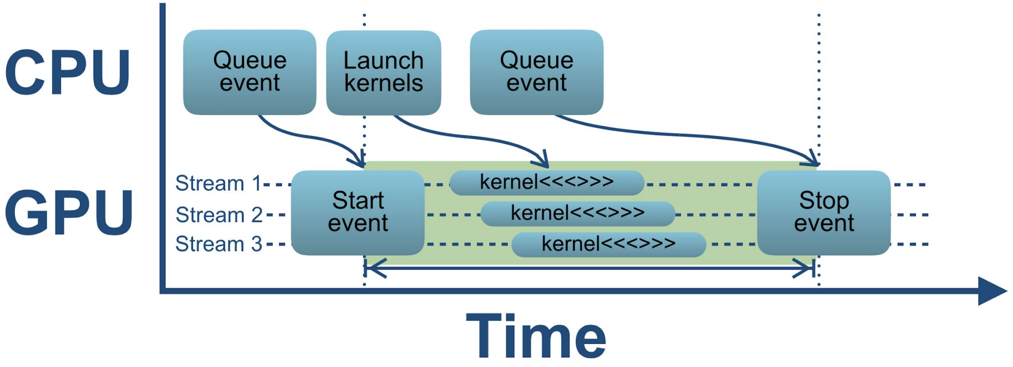
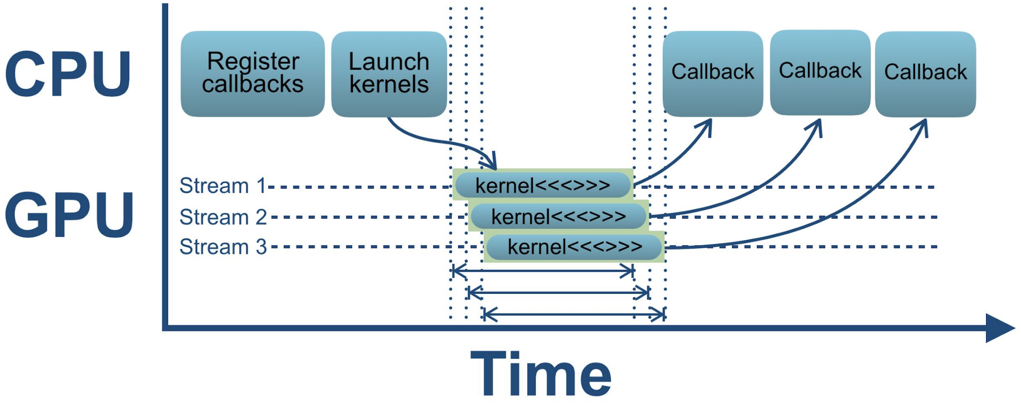
Three possible on-the-fly timing strategies

1. ~~Count GPU clock cycles~~
   - ~~Requires additional device-to-host transfers~~
   - ~~Implementation may be invasive~~
2. CUDA Events
   - Can give ambiguous results
3. CUDA Profiling Tools Interface (CUPTI)
   - Buffer requests and delivery of timing information handled by CUPTI
   - Gives unambiguous kernel timings

1. ~~Start GPU timer~~
2. ~~Do GPU kernel work~~
3. ~~Stop GPU timer~~
4. ~~Send elapsed time to host~~

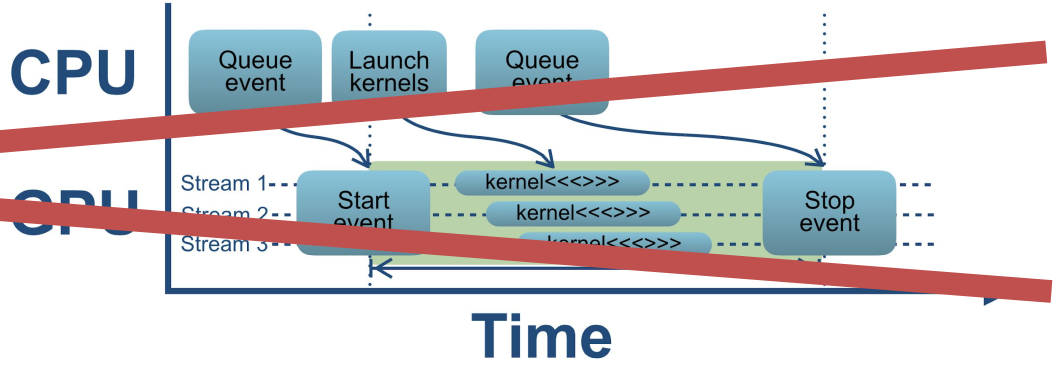


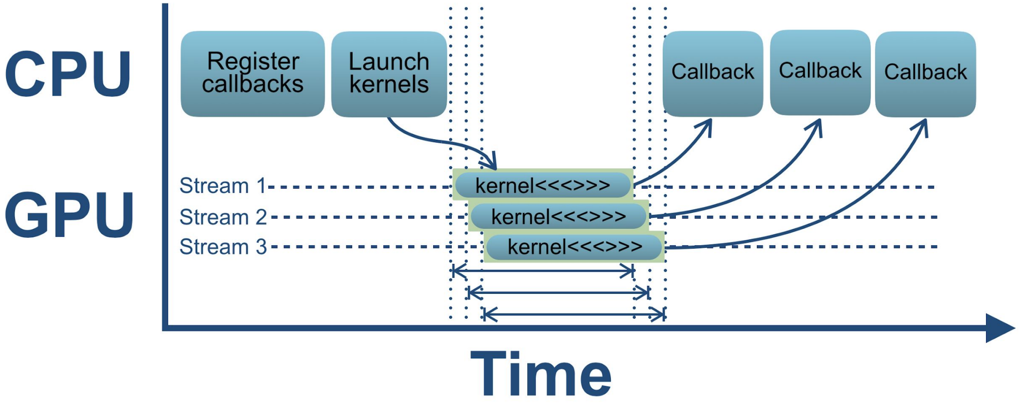


Three possible on-the-fly timing strategies

1. ~~Count GPU clock cycles~~
   - ○ ~~Requires additional device-to-host transfers~~
   - ○ ~~Implementation may be invasive~~
2. ~~CUDA Events~~
   - ○ ~~Can give ambiguous results~~
3. CUDA Profiling Tools Interface (CUPTI)
   - ○ Buffer requests and delivery of timing information handled by CUPTI
   - ○ Gives unambiguous kernel timings

~~1. Start GPU timer~~
~~2. Do GPU kernel work~~
~~3. Stop GPU timer~~
~~4. Send elapsed time to host~~


CPU
GPU
Queue event
Launch kernels
Queue event
Stream 1
Stream 2
Stream 3
Start event
kernel<<<>>>
kernel<<<>>>
kernel<<<>>>
Stop event
Time


CPU
GPU
Register callbacks
Launch kernels
Callback
Callback
Callback
Stream 1
Stream 2
Stream 3
kernel<<<>>>
kernel<<<>>>
kernel<<<>>>
Time