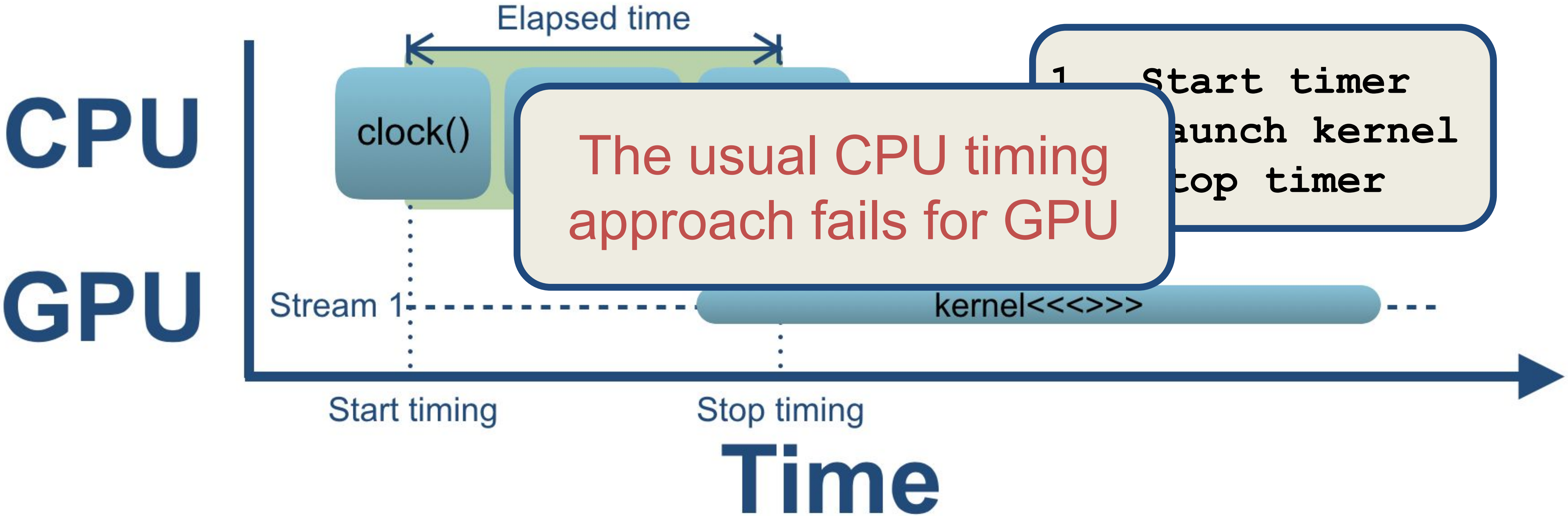
Cannot measure execution time on GPU the same way as on CPU!

- What happens if we try to measure kernel execution time naively?



```
1    Start timer
    aunch kernel
    top timer
```

The usual CPU timing approach fails for GPU

GPU work is in general asynchronous

- GPU operations are asynchronous with respect to:
  - Streams (series of operations which execute in issue order)
    - Operations across streams may be interleaved
    - While operations within a stream execute in-order, there is no relationship between issue order execution order for operations in different streams
  - Host
    - Kernel execution, e.g., is by default asynchronous with host

```
kernel<<<...>>>(...)
cpuWork(...)
```

} May overlap, as kernel launch is non-blocking