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

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
```