**cudaDeviceSynchronize() and cudaEventSynchronize()**

The functions cudaDeviceSynchronize() and cudaEventSynchronize() are both used for **synchronization in CUDA**, but they are used for **different scopes and purposes**.

Here’s a detailed comparison:

**1. cudaDeviceSynchronize()**

**🔹 Purpose:**

* **Waits for all tasks on the entire device (GPU) to finish** — across **all streams** and kernels.

**🔹 Behavior:**

* It blocks the **host (CPU)** until all previously issued commands on **all CUDA streams** have completed.

**🔹 Use Case:**

* Used to:
  + Ensure **all GPU work** has completed before moving on.
  + Benchmark entire kernel execution (start → kernel launch → cudaDeviceSynchronize() → stop).

**🔹 Example:**

kernel<<<blocks, threads>>>();

cudaDeviceSynchronize(); // Wait until all kernels finish

**2. cudaEventSynchronize(cudaEvent\_t event)**

**🔹 Purpose:**

* **Waits for a specific CUDA event** to complete — not the entire device.

**🔹 Behavior:**

* Only blocks the CPU until the **specific event** has occurred (e.g., kernel or memcpy completion).
* More fine-grained than cudaDeviceSynchronize().

**🔹 Use Case:**

* Used for:
  + Measuring time between two events.
  + Synchronizing on **a specific stage** of a GPU workflow.

**🔹 Example:**

cudaEvent\_t start, stop;

cudaEventCreate(&start);

cudaEventCreate(&stop);

cudaEventRecord(start);

kernel<<<blocks, threads>>>();

cudaEventRecord(stop);

cudaEventSynchronize(stop); // Wait until 'stop' event finishes

**Key Differences Summary**

| **Feature** | **cudaDeviceSynchronize()** | **cudaEventSynchronize(event)** |
| --- | --- | --- |
| Scope | All GPU activities (global sync) | Single event (fine-grained sync) |
| Blocks Host Until | All kernels & memcopies complete | Specified event is complete |
| Performance Impact | Higher (can block more work) | Lower (syncs only what you need) |
| Common Use Case | Final sync before host access or timing | Timing GPU kernel execution or syncing milestones |
| Granularity | Coarse (device-wide) | Fine-grained (event-level) |

**Best Practice**

* Use cudaEventSynchronize() when measuring **timing** or waiting for **specific events**.
* Use cudaDeviceSynchronize() when you want to ensure **all GPU operations are done** before continuing (e.g., before reading output or for debugging).