GPU Computing

Exercise Sheet 02

Maximilian Richter, Jan Kränzke, Markus Everling,
Nicolas Schledorn, Olga Sergeyeva, Xel Pratscher

November 4, 2024

1 Raw Kernel Startup Time

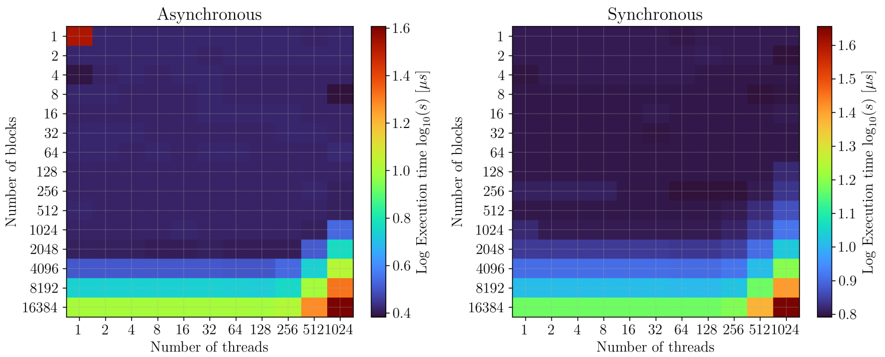

Figure 1: Asynchronous and synchronous kernel start-up time for different number of blocks and number of threads.

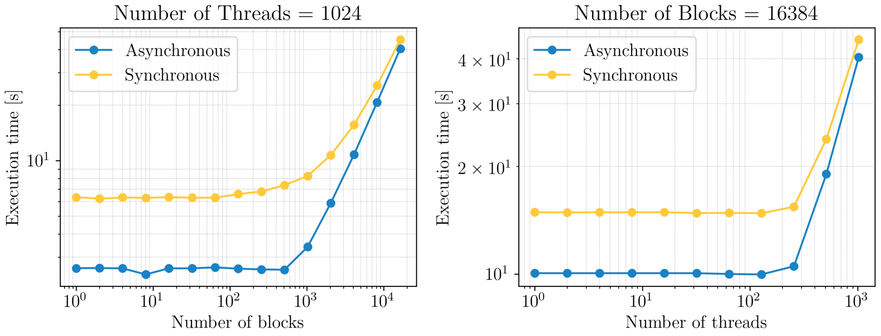

Figure 2: Asynchronous and synchronous kernel start-up time for (left) fixed number of number of threads and (right) fixed number of blocks.

As observed in Figure 1, the kernel start-up time generally increases with the number of threads and blocks. Additionally, Figure 2 reveals that synchronous start-up consistently takes longer than asynchronous start-up. This difference is expected, as synchronous execution requires waiting for each kernel to complete before moving to the next, whereas asynchronous execution can overlap processes, reducing overall latency.

2 Break-even Kernel Startup Time

Using the `long long int clock64();` function, we can count the number of clock cycles we need to wait to allow for another kernel launch call. To measure this, the kernel busy-waits until we observe that the asynchronous kernel startup time doubles. This measurement was rather inconsistent with parameters but we came out at around: **8200 cycles** (at a time of 5.38e-02 from `chTimerElapsedTime`)

3 PCIe Data Movements

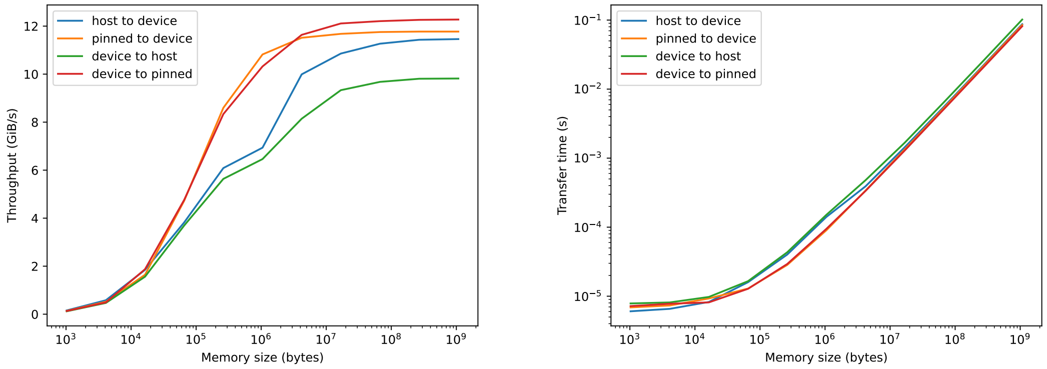

Figure 3: Comparison of the (left) throughput and (right) time for different memory allocation methods.

The plots in Figure 3 indicate that the throughput of the memory transfer between host and device is limited in all four cases. Contrary to that, the transfer time keeps increasing similarly for each allocation. This can be understood as a consequence of the limited theoretical maximum throughput of the PCIe 3.0 x16 connection. For small copies however, the throughput is not limited by the PCIe bandwidth, but rather by the overhead of dispatching the copy commands to the GPU. The most throughput achieved the *device to pinned* allocation. The worst is the *device to host* throughput. This is expected, as the GPU can copy to pinned memory directly, but cannot copy to pageable host memory without indirections.

4 Willingness to Present

- Raw Kernel Startup Time: **True**
- Break-even Kernel Startup Time: **True**
- PCIe Data Movements: **True**