Exercise: Introduction to CUDA

So far we have mainly talked about the difficulties of measuring the worst case execution times of tasks (WCET). Then we have discussed the particularities of real-time operating systems, such as process scheduling and memory management, such that the **overall system can guarantee that the deadlines of all tasks are met** (if the task set is schedulable at all → schedulability test).

Another aspect of real-time sytems is that the **overall system should be as fast as possible**. Deadlines can easier be met if the computations are accelerated by special hardware, as, e.g. a GPU.

For this, we want to discuss some basics of accelerating computations on GPUs.

Read the following excellent tutorial by Mark Harris:

https://devblogs.nvidia.com/even-easier-introduction-cuda/

In this tutorial a first simple task (adding two vectors / arrays) is being parallelized with the help of CUDA.

Then try to compile the example from this tutorial with the help of the `nvcc` compiler.

Then answer the following question: How are the two vectors added in parallel?

If you have no CUDA compatible GPU: there are many GPU Cloud Computing platforms where one can rent a machine with a GPU on an hourly basis. Some platforms also offer free trials.