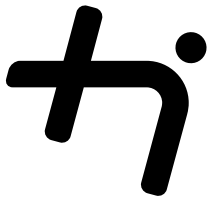
Technische Hochschule Ingolstadt
Prof. Dr.-Ing. Richard Membarth



29. März 2023

GPU Programming
Assignment 2

Submission deadline for the exercises: 03. April 2023

2.1 CUDA Vector Addition

The purpose of this exercise is to get familiar with the development environment and the CUDA API by implementing vector addition.

a) The `vec-add.cu` file contains the host and device code for the vector addition. Implement the `vec_add` CUDA kernel that adds two vectors `a` and `b` with `N` elements and stores the result to `c`.

```
1 __global__ void vec_add(const int* a, const int* b, int* c, int N) {
2     // ...
3 }
```

Add the missing CUDA API calls on the host side in order to launch the kernel:

- allocate device memory
- copy host memory to the device
- launch the kernel
- copy the device memory back to the host
- free device memory

Make sure that your code checks for CUDA API errors and works for different kernel input sizes.

b) Build and execute your program using `nvcc`, the CUDA compiler:

```
1 nvcc vec-add.cu -o vec-add
2 ./vec-add
```

For future programming tasks, we will use the cross-platform CMake tool for building our programs. CMake allows to specify the build dependencies in a `CMakeLists.txt` file and supports building outside of the source tree. Build and test your program using CMake:

```
mkdir build
cd build
cmake ..
make
./vec-add
```

c) Profile the execution time of your program using `nsys profile --stats=true`. Note that `nsys` is available at `/usr/lib/nsight-systems/bin/nsys` in the CIP pool G308. In case `nsys` does not work, you can use the CUDA event API:

```
float time;
cudaEvent_t start, stop;
cudaEventCreate(&start);
cudaEventCreate(&stop);
cudaEventRecord(start, 0);

// launch the kernel

cudaEventRecord(stop, 0);
cudaEventSynchronize(stop);
cudaEventElapsedTime(&time, start, stop);
```

How long does your vector addition take for 1000000 elements?
Report the execution time for the following block sizes:

block size	16	32	64	128	256	512	1024	2048
time (us)								

How does the block size influence the execution time?