Introduction to CUDA and OpenCL

Ilona Tomkowicz, Zofia Wpisz swoje nazwisko

October 17, 2019

Contents

1 Communication with GPU

In order to check details about the GPU used by VM on labolatories (GeForce GTX 1060 6GB) we used a template project deviceQuery.

1.1 How to establish connection with device

To etract information about the devices connected to host we use funcion

```
cudaGetDeviceCount(&deviceCount)
```

cudaGetDeviceCount(&deviceCount), which returnes the flag message and changed the passed argument according to the number of devices that were found. Setting onection is done by calling:

```
cudaSetDevice(dev_index);
```

Where dev_index is the index of GPU found (for only 1 GPU it will be 0).

1.2 How to fetch information about GPU

To fetch information about GPU, after completing steps in previous point, we can call:

```
cudaDeviceProp deviceProp;
cudaGetDeviceProperties(&deviceProp, dev_index).
```

to get information about device properties or

```
cudaDriverGetVersion(&driverVersion);
cudaRuntimeGetVersion(&runtimeVersion);
```

to check driver and runtime version. After that we can extract specific data like memory size, clock rate, etc using these functions parameters.

2 Experiments with data size and grid layout

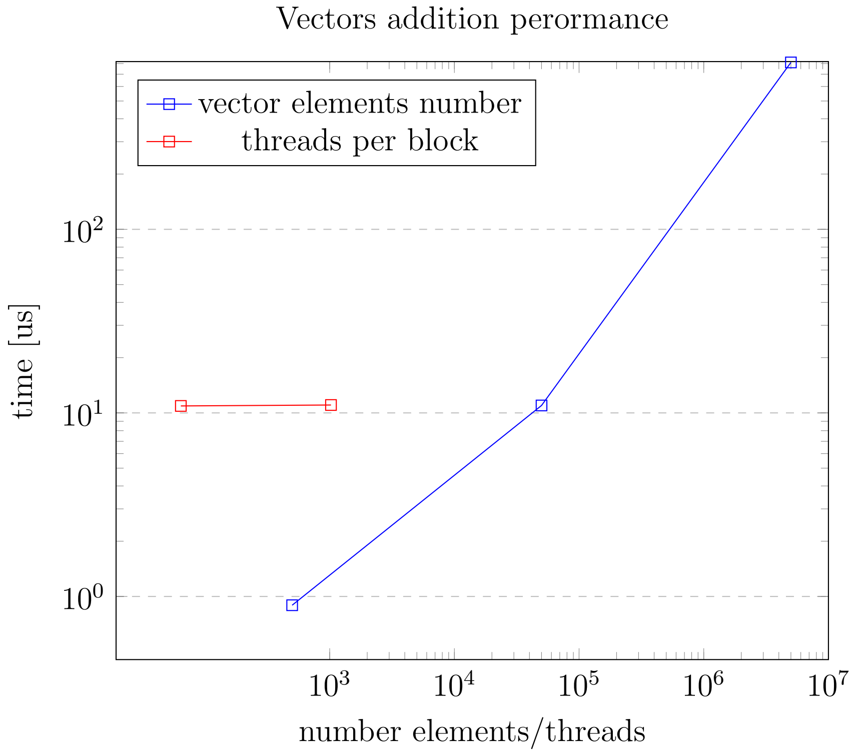
Experiments with performance were made using template vectorAdd. This template originally

- allocates memory in host and device,
- copies the host input vectors A and B into device memory,

- sets grid layout,
- uses created kernel to make caculations,
- copies the result back into host,
- veifies if the result is correct
- frees memory on both, device and host.

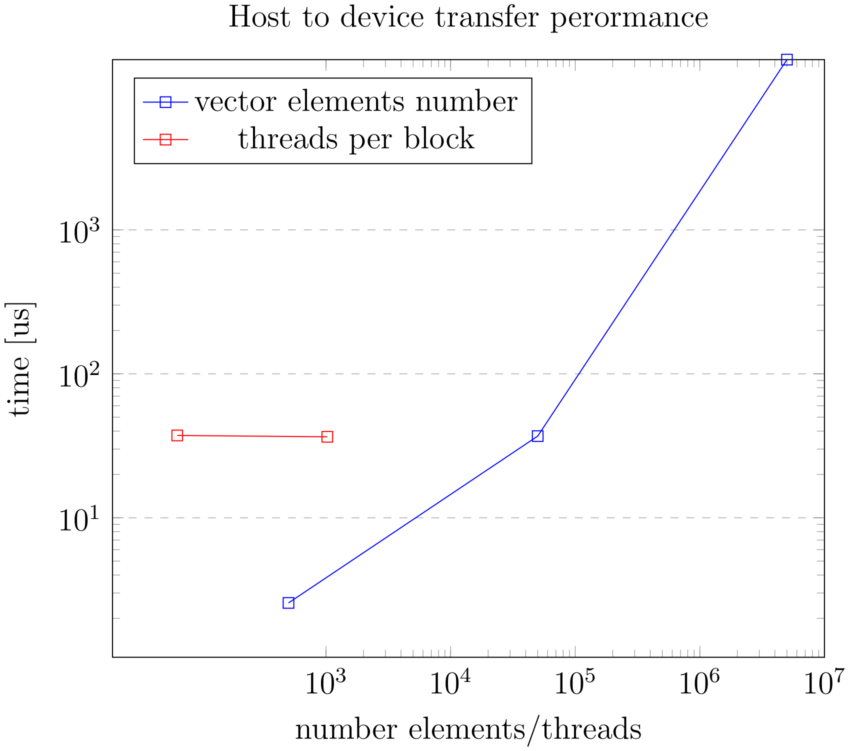
Number of elements is changed for fixed gird layout and grid layout is changed for fixed number of elements.

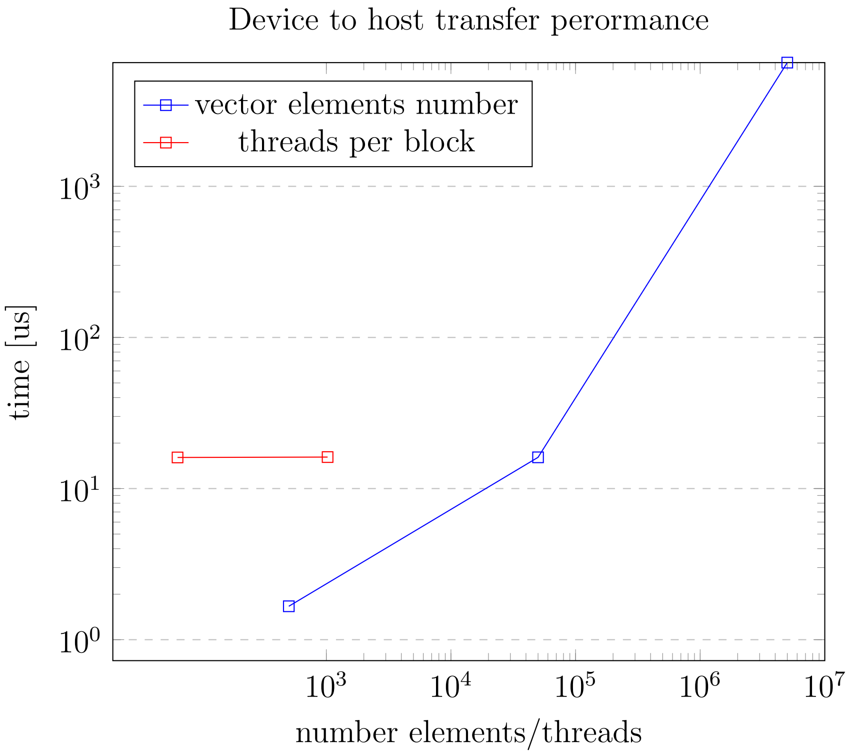
2.1 Computation time comparison



A huge difference can be seen when changing data size by two orders of magnitude, whereas changing number of threads from very small (64) to very big (1024) does not result in any major change in terms of time performance.

2.2 Data transfer time comparison





Although transfer times differ for copying from and to the device, plots look like more or less the same for both operations.

Generally, for all actions on which experiments were conucted the time change is smaller for low figures (for change from 500 to 50 000 - two orders of magnitude - time increased by c. one order of magnitude) and bigger for high figures (for change from 50 000 to 5 000 000 - also two orders - we gained c. two orders of magnitude). It indicates, that using GPU with large number of data transfered at the same time can be less beneficial than granulating this data into subsets and processing them in small portions in GPU. That is why we should try to form our computations in group of smaller calculations, but not too small to aviod too fequent and unnecessary data copying between host and device.