360.252 - Computational Science on Many-Core Architectures

WS 2020 - Exercise 3

Christian Gollmann, 01435044

Last update: November 7, 2020

Contents

1 Strided Memory Access

I first wrote an implementation working on a small dataset to check if my kernel was operating correctly. Then I moved to the actual timing benchmark.

In both cases there can be seen a clear trend. I cannot really think of a reason why summing every second element is so much slower than say summing every or every sixth element. But in general it definitely makes sense that the runtime goes down when operating on less vector entries.

That the bandwidth goes down so drastically was something I didn't expect, but can also be explained. I therefor refer to an analogy when working on the CPU. There, the smallest amount of loadable memory is the cacheline. So even if only one 8 byte element is needed, the whole (mostly 64byte) cacheline hast to be loaded. That is why from a performance point of view, it is so important to keep an eye on data locality.

In the example here, we destroy the data locality concept by leaving empty spaces in between our vector usage. That's why we are no longer using what we have very efficiently.

In the runtime it can be seen that there are several plateaus, this effect becomes even clearer when turning k up to 127. This must also origin from the GPU's memory layout.

As a recommendation for more complex cases, I would say max out the use of data locality.

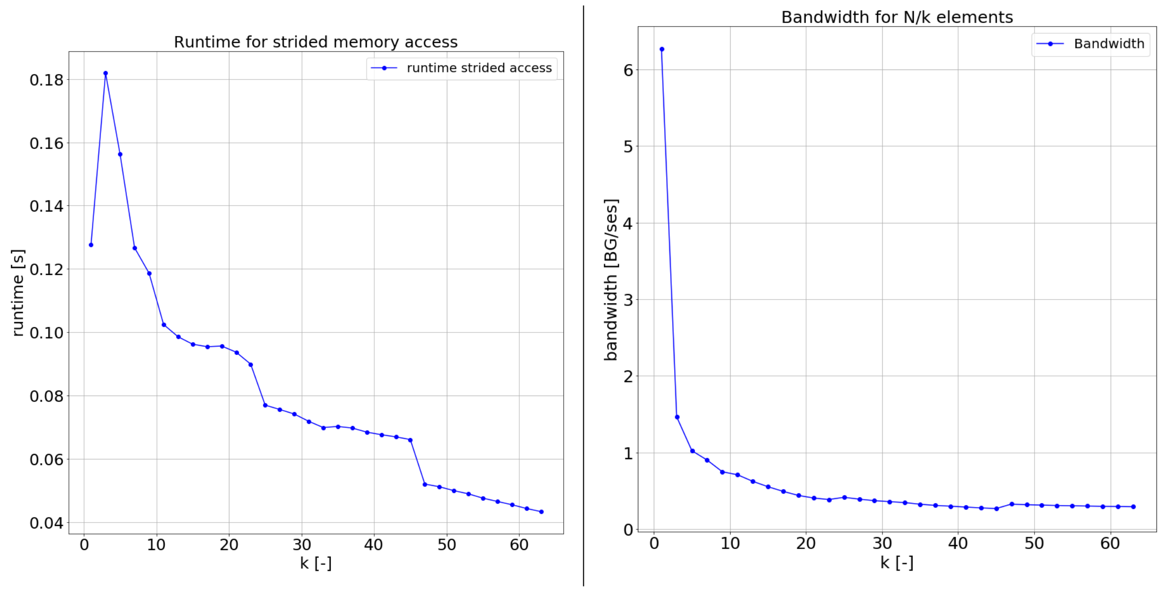

Figure 1: Runtime and bandwidth for strided memory access

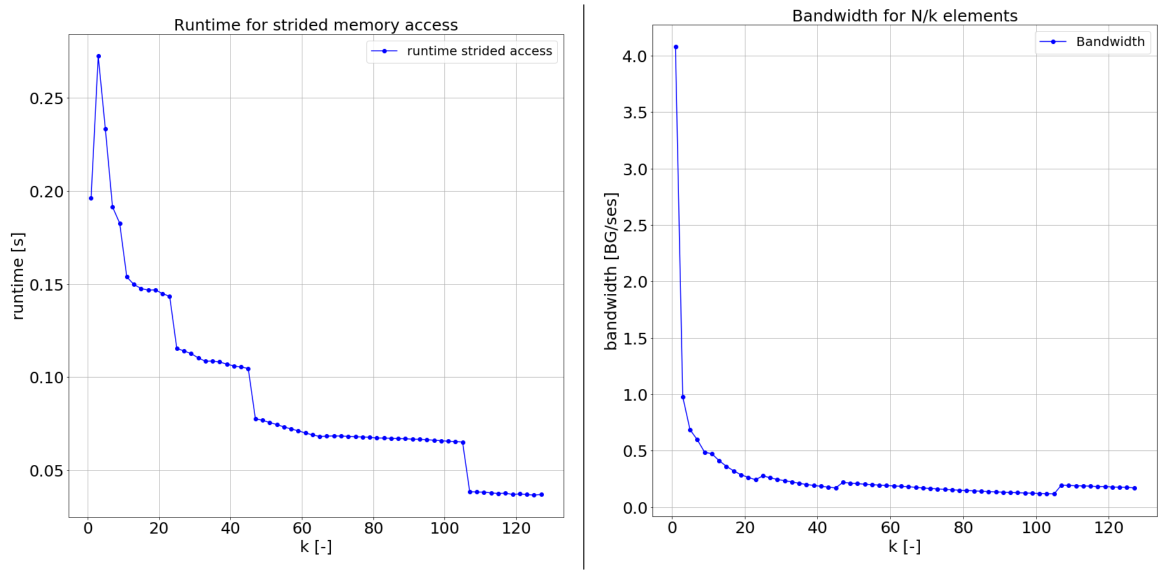

Figure 2: Runtime and bandwidth for strided memory access

Listing 1: Validation of Code for strided memory access

```
#include <stdio.h>
#include "timer.hpp"
#include <iostream>
#include <algorithm>
#include <vector>


#define MAGIC_NUMBER 10

__global__ void sumVectors(double *x, double *y, double *z, int N, int k)
{
    int thread_id = blockIdx.x * blockDim.x + threadIdx.x;

    for(size_t i = thread_id; i < N/k; i += blockDim.x * gridDim.x)
        z[k*i] = x[k*i] + y[k*i];

}

int main(void)
{
    int N = 100;
    int k = 20;

    double*x, *y, *z, *d_x, *d_y, *d_z;
    Timer timer;

    x = new double[N];
    y = new double[N];
    z = new double[N];


    for (int i = 0; i < N; i++)
    {
        x[i] = 1;
        y[i] = 2;
```

```
            z[i] = 0;
        }


        cudaMalloc(&d_x, N*sizeof(double));
        cudaMalloc(&d_y, N*sizeof(double));
        cudaMalloc(&d_z, N*sizeof(double));
        cudaMemcpy(d_x, x, N*sizeof(double), cudaMemcpyHostToDevice);
        cudaMemcpy(d_y, y, N*sizeof(double), cudaMemcpyHostToDevice);
        cudaMemcpy(d_z, z, N*sizeof(double), cudaMemcpyHostToDevice);

        cudaDeviceSynchronize();
        timer.reset();
        std::vector<double> timings;

        for(int reps=0; reps < MAGIC_NUMBER; ++reps)
        {
            sumVectors<<<256, 256>>>(d_x, d_y, d_z, N, k);
            cudaDeviceSynchronize();
            timings.push_back(timer.get());
        }

        std::sort(timings.begin(), timings.end());
        double time_elapsed = timings[MAGIC_NUMBER/2];

        cudaMemcpy(z, d_z, N*sizeof(double), cudaMemcpyDeviceToHost);

        printf("Addition took %g seconds", time_elapsed);

        std::cout << std::endl << "z[0] = " << z[0] << std::endl;
        std::cout << "z[1] = " << z[1] << std::endl;
        std::cout << "z[k] = " << z[k] << std::endl;
        std::cout << "z[2*k] = " << z[2*k-1] << std::endl;
        std::cout << "z[2*k+1] = " << z[2*k-1+1] << std::endl;

        cudaFree(d_x);
        cudaFree(d_y);
        cudaFree(d_z);
        delete x;
        delete y;
        delete z;

        return EXIT_SUCCESS;
}
```

Listing 2: Actual Code used in benchmark

```
#include <stdio.h>
#include "timer.hpp"
#include <iostream>
#include <algorithm>
#include <vector>


#define MAGIC_NUMBER 10

__global__ void sumVectors(double *x, double *y, double *z, int N, int k)
```

```
{
    int thread_id = blockIdx.x * blockDim.x + threadIdx.x;

    for(size_t i = thread_id; i < N/k; i += blockDim.x * gridDim.x)
        z[k*i] = x[k*i] + y[k*i];

}

int main(void)
{
    int N = 100000000;

    double*x, *y, *z, *d_x, *d_y, *d_z;
    Timer timer;

    x = new double[N];
    y = new double[N];
    z = new double[N];


    for (int i = 0; i < N; i++)
    {
        x[i] = 1;
        y[i] = 2;
        z[i] = 0;
    }


    cudaMalloc(&d_x, N*sizeof(double));
    cudaMalloc(&d_y, N*sizeof(double));
    cudaMalloc(&d_z, N*sizeof(double));
    cudaMemcpy(d_x, x, N*sizeof(double), cudaMemcpyHostToDevice);
    cudaMemcpy(d_y, y, N*sizeof(double), cudaMemcpyHostToDevice);
    cudaMemcpy(d_z, z, N*sizeof(double), cudaMemcpyHostToDevice);

    for(int k = 1; k < 64; k += 2)
    {
        cudaDeviceSynchronize();
        timer.reset();
        std::vector<double> timings;

        for(int reps=0; reps < MAGIC_NUMBER; ++reps)
        {
            sumVectors<<<256, 256>>>(d_x, d_y, d_z, N, k);
            cudaDeviceSynchronize();
            timings.push_back(timer.get());
        }

        std::sort(timings.begin(), timings.end());
        double time_elapsed = timings[MAGIC_NUMBER/2];

        std::cout << time_elapsed << std::endl;
    }

    cudaFree(d_x);
    cudaFree(d_y);
    cudaFree(d_z);
```

```
    delete x;
    delete y;
    delete z;

    return EXIT_SUCCESS;
}
```

2 Offset Memory Access

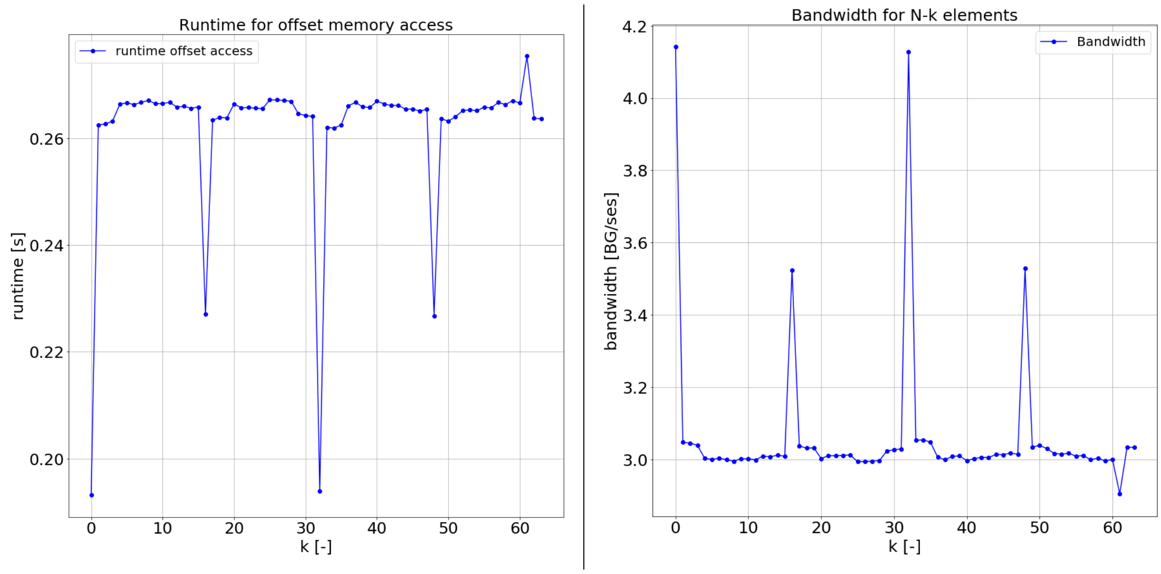

Figure 3: Runtime and bandwidth for offset memory access

Listing 3: Code for offset access

```
#include <stdio.h>
#include "timer.hpp"
#include <iostream>
#include <algorithm>
#include <vector>


#define MAGIC_NUMBER 10

__global__ void sumVectors(double *x, double *y, double *z, int N, int k)
{
    int thread_id = blockIdx.x * blockDim.x + threadIdx.x;

    for(size_t i = thread_id; i < N-k; i += blockDim.x * gridDim.x)
        z[k+i] = x[k+i] + y[k+i];

}

int main(void)
{
    int N = 100000000;

    double *x, *y, *z, *d_x, *d_y, *d_z;
    Timer timer;

    x = new double[N];
    y = new double[N];
    z = new double[N];


    for (int i = 0; i < N; i++)
```

```
    {
        x[i] = 1;
        y[i] = 2;
        z[i] = 0;
    }


    cudaMalloc(&d_x, N*sizeof(double));
    cudaMalloc(&d_y, N*sizeof(double));
    cudaMalloc(&d_z, N*sizeof(double));
    cudaMemcpy(d_x, x, N*sizeof(double), cudaMemcpyHostToDevice);
    cudaMemcpy(d_y, y, N*sizeof(double), cudaMemcpyHostToDevice);
    cudaMemcpy(d_z, z, N*sizeof(double), cudaMemcpyHostToDevice);

    for(int k = 0; k < 64; k += 1)
    {
        cudaDeviceSynchronize();
        timer.reset();
        std::vector<double> timings;

        for(int reps=0; reps < MAGIC_NUMBER; ++reps)
        {
            sumVectors<<<256, 256>>>(d_x, d_y, d_z, N, k);
            cudaDeviceSynchronize();
            timings.push_back(timer.get());
        }

        std::sort(timings.begin(), timings.end());
        double time_elapsed = timings[MAGIC_NUMBER/2];

        std::cout << time_elapsed << std::endl;
    }

    cudaFree(d_x);
    cudaFree(d_y);
    cudaFree(d_z);
    delete x;
    delete y;
    delete z;

    return EXIT_SUCCESS;
}
```

3 Conjugate Gradient - matrix vector product

Here I went with the example from the lecture. But I also took much inspiration from https://medium.com/analytics-vidhya/sparse-matrix-vector-multiplication-with-cuda-42d191878e8f which turned out to be a good reference also for later and more sophisticated implimentations.

Listing 4: CUDA kernel for calculating sparse matrix vector multiplication

```
__global__ void csr_matvec(int N, int *rowoffsets, int *colindices, double *values,
double const *x, double *y)
{
    for (int row = blockDim.x * blockIdx.x + threadIdx.x;
        row < N;
        row += gridDim.x * blockDim.x)
    {
      double val = 0;
      for (int jj = rowoffsets[i]; jj < rowoffsets[i+1]; ++jj)
      {
          val += values[jj] * x[colindices[jj]];
      }
      y[row] = val;
    }
}
```

4 Conjugate Gradient - further kernels

Listing 5: kernel for the vector operations in lines 5 and 9

```
__global__ void dot_product(double *x, double *y, double *dot, unsigned int n)
{
    unsigned int index = threadIdx.x + blockDim.x*blockIdx.x;
    unsigned int stride = blockDim.x*gridDim.x;

    __shared__ double cache[256];

    double temp = 0.0;
    while(index < n){
        temp += x[index]*y[index];

        index += stride;
    }

    cache[threadIdx.x] = temp;

    __syncthreads();

    for(int i = blockDim.x/2; i>0; i/=2)
    {
        __syncthreads();
        if(threadIdx.x < i)
            cache[threadIdx.x] += cache[threadIdx.x + i];
    }

    if(threadIdx.x == 0){
        atomicAdd(dot, cache[0]);
    }
}
```

Listing 6: kernel for the vector operations in lines 7, 8 and 12

```
__global__ void vector_plus_alpha_vector(double *x, double *y,
double *z, double alpha, int N)
{
    int thread_id = blockIdx.x * blockDim.x + threadIdx.x;

    for(size_t i = thread_id; i < N; i += blockDim.x * gridDim.x)
        z[i] = x[i] + alpha * y[i];

}
```

5 Basic CUDA e)

Now I called the vector addition for a vector with e7 elements with varying grid and block sizes. It seems that small values (16, 32, 64) lead to a not so good performance.

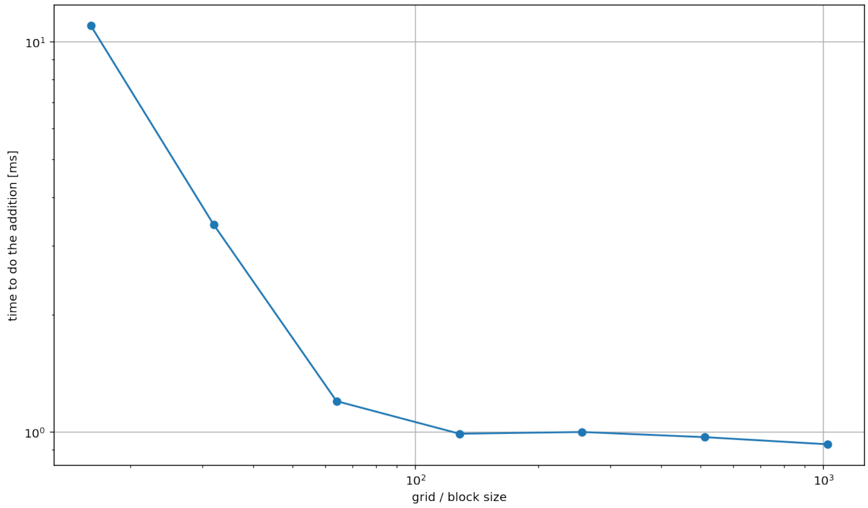

Figure 4: addition time for vector with e7 elements for different grid / block values

Listing 7: kernel call with different values x

```
sumVectors<<<x, x>>>(d_x, d_y, d_z, N);
```

6 Dot Product a)

For the Dot Product Exercises I got some inspiration from https://bitbucket.org/jsandham/algorithms_in_cuda/src/master/dot_product/ and added / changed code as needed. I am aware that my implementations are not the prettiest and might also be error prone. But they get the job done for those specific examples.

Listing 8: Dot Product with two GPU stages

```
#include <stdio.h>
#include <iostream>
#include "timer.hpp"
#include <random>


__global__ void dot_product_first_part(double *x, double *y,
double *temporary, unsigned int n)
{
    unsigned int index = threadIdx.x + blockDim.x*blockIdx.x;
    unsigned int stride = blockDim.x*gridDim.x;

    __shared__ double cache[256];

    double temp = 0.0;
    while(index < n){
        temp += x[index]*y[index];

        index += stride;
    }

    cache[threadIdx.x] = temp;

    __syncthreads();

    for(int i = blockDim.x/2; i>0; i/=2)
    {
        __syncthreads();
        if(threadIdx.x < i)
            cache[threadIdx.x] += cache[threadIdx.x + i];
    }

    if(threadIdx.x == 0){
        temporary[blockIdx.x] = cache[0];
    }
}

__global__ void dot_product_second_part(double* temporary, double* dot)
{
    for(int i = blockDim.x/2; i>0; i/=2)
    {
        __syncthreads();
        if(threadIdx.x < i)
            temporary[threadIdx.x] += temporary[threadIdx.x + i];
    }

    __syncthreads();
```

```
    if(threadIdx.x == 0){
        *dot = temporary[0];
    }

}

int main()
{
    unsigned int n = 10000;
    unsigned int x = 256;
    double *h_prod;
    double *d_prod;
    double *h_x, *h_y;
    double *d_x, *d_y;
    double *d_temporary;
    Timer timer;

    h_prod = new double[n];
    h_x = new double[n];
    h_y = new double[n];


    // fill host array with data
    for(unsigned int i=0;i<n;i++){
        h_x[i] = 1;
        h_y[i] = 2;
    }

    // start timer
    timer.reset();

    // allocate memory
    cudaMalloc(&d_prod, sizeof(double));
    cudaMalloc(&d_x, n*sizeof(double));
    cudaMalloc(&d_y, n*sizeof(double));
    cudaMalloc(&d_temporary, x*sizeof(double));


    // copy data to device
    cudaMemcpy(d_x, h_x, n*sizeof(double), cudaMemcpyHostToDevice);
    cudaMemcpy(d_y, h_y, n*sizeof(double), cudaMemcpyHostToDevice);


    dot_product_first_part<<<x, x>>>(d_x, d_y, d_temporary, n);
    dot_product_second_part<<<1, x>>>(d_temporary, d_prod);


    // copy data back to host
    cudaMemcpy(h_prod, d_prod, sizeof(double), cudaMemcpyDeviceToHost);

    // get runtime
    double time_elapsed = timer.get();


    // report results
    std::cout<<"dot product computed on GPU is: "<<*h_prod<<" and took "
```

```
    << time_elapsed << " s" <<std::endl;


    // free memory
    free(h_prod);
    free(h_x);
    free(h_y);
    cudaFree(d_prod);
    cudaFree(d_x);
    cudaFree(d_y);

}
```

7 Dot Product b)

Listing 9: Dot Product with GPU and CPU combined

```
#include <stdio.h>
#include <iostream>
#include "timer.hpp"
#include <random>


__global__ void dot_product(double *x, double *y, double *temporary, unsigned int n)
{
    unsigned int index = threadIdx.x + blockDim.x*blockIdx.x;
    unsigned int stride = blockDim.x*gridDim.x;

    __shared__ double cache[256];

    double temp = 0.0;
    while(index < n){
        temp += x[index]*y[index];

        index += stride;
    }

    cache[threadIdx.x] = temp;

    __syncthreads();

    for(int i = blockDim.x/2; i>0; i/=2)
    {
        __syncthreads();
        if(threadIdx.x < i)
            cache[threadIdx.x] += cache[threadIdx.x + i];
    }

    if(threadIdx.x == 0){
        temporary[blockIdx.x] = cache[0];
    }
}



int main()
{
    unsigned int n = 10000;
    unsigned int x = 256;
    double *h_prod;
    double *d_prod;
    double *h_x, *h_y;
    double *d_x, *d_y;
    double *d_temporary;
    double *h_temporary;
    Timer timer;

    h_prod = new double[n];
    h_x = new double[n];
    h_y = new double[n];
```

```
    h_temporary = new double[x];


    // fill host array with data
    for(unsigned int i=0;i<n;i++){
        h_x[i] = 1;
        h_y[i] = 2;
    }

    // start timer
    timer.reset();

    // allocate memory
    cudaMalloc(&d_prod, sizeof(double));
    cudaMalloc(&d_x, n*sizeof(double));
    cudaMalloc(&d_y, n*sizeof(double));
    cudaMalloc(&d_temporary, x*sizeof(double));
    cudaMemset(d_prod, 0.0, sizeof(double));


    // copy data to device
    cudaMemcpy(d_x, h_x, n*sizeof(double), cudaMemcpyHostToDevice);
    cudaMemcpy(d_y, h_y, n*sizeof(double), cudaMemcpyHostToDevice);


    dot_product<<<x, x>>>(d_x, d_y, d_temporary, n);


    // copy data back to host
    cudaMemcpy(h_temporary, d_temporary, x*sizeof(double), cudaMemcpyDeviceToHost);

    // sum up elements
    double dot = 0;
    for(int i = 0; i < x; i++)
    {
        dot += h_temporary[i];
    }

    // get runtime
    double time_elapsed = timer.get();


    // report results
    std::cout<<"dot product computed on GPU and CPU is: "<<dot<<" and took "
    << time_elapsed << " s" <<std::endl;


    // free memory
    free(h_prod);
    free(h_x);
    free(h_y);
    cudaFree(d_prod);
    cudaFree(d_x);
    cudaFree(d_y);

}
```

8 Dot Product c)

Listing 10: Dot Product with atomicAdd on GPU

```
#include <stdio.h>
#include <iostream>
#include "timer.hpp"
#include <random>


__global__ void dot_product(double *x, double *y, double *dot, unsigned int n)
{
    unsigned int index = threadIdx.x + blockDim.x*blockIdx.x;
    unsigned int stride = blockDim.x*gridDim.x;

    __shared__ double cache[256];

    double temp = 0.0;
    while(index < n){
        temp += x[index]*y[index];

        index += stride;
    }

    cache[threadIdx.x] = temp;

    __syncthreads();

    for(int i = blockDim.x/2; i>0; i/=2)
    {
        __syncthreads();
        if(threadIdx.x < i)
            cache[threadIdx.x] += cache[threadIdx.x + i];
    }

    if(threadIdx.x == 0){
        atomicAdd(dot, cache[0]);
    }
}



int main()
{
    unsigned int n = 10000;
    double *h_prod;
    double *d_prod;
    double *h_x, *h_y;
    double *d_x, *d_y;
    Timer timer;

    h_prod = new double[n];
    h_x = new double[n];
    h_y = new double[n];


    // fill host array with data
```

```
    for(unsigned int i=0;i<n;i++){
        h_x[i] = 1;
        h_y[i] = 2;
    }

    // start timer
    timer.reset();

    // allocate memory
    cudaMalloc(&d_prod, sizeof(double));
    cudaMalloc(&d_x, n*sizeof(double));
    cudaMalloc(&d_y, n*sizeof(double));
    cudaMemset(d_prod, 0.0, sizeof(double));


    // copy data to device
    cudaMemcpy(d_x, h_x, n*sizeof(double), cudaMemcpyHostToDevice);
    cudaMemcpy(d_y, h_y, n*sizeof(double), cudaMemcpyHostToDevice);


    dot_product<<<256, 256>>>(d_x, d_y, d_prod, n);

    // copy data back to host
    cudaMemcpy(h_prod, d_prod, sizeof(double), cudaMemcpyDeviceToHost);

    // get runtime
    double time_elapsed = timer.get();


    // report results
    std::cout<<"dot product computed on GPU is: "<<*h_prod<<" and took "
    << time_elapsed << " s" <<std::endl;


    // free memory
    free(h_prod);
    free(h_x);
    free(h_y);
    cudaFree(d_prod);
    cudaFree(d_x);
    cudaFree(d_y);

}
```

9 Dot Product d)

It can be seen that at least in my case, there is no difference in runtime for the different methods a to c. This surprises me but maybe is due to how I time the executions. Though, I wanted to take all the steps necessary (also the cudaMalloc and cudaMemcpy) into account. Again there are some similarities in respect to the number of vector entries with Figure 1 and Figure 2. So this behaviour could be memory related.

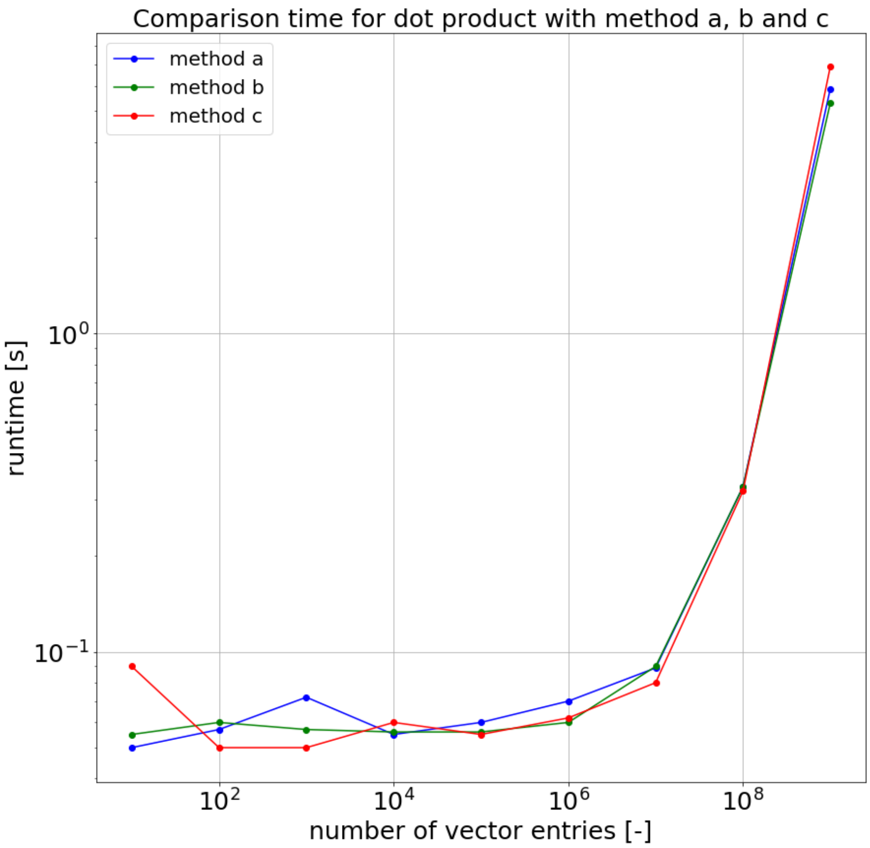

Figure 5: Runtimes for calculating dot product with different methods.