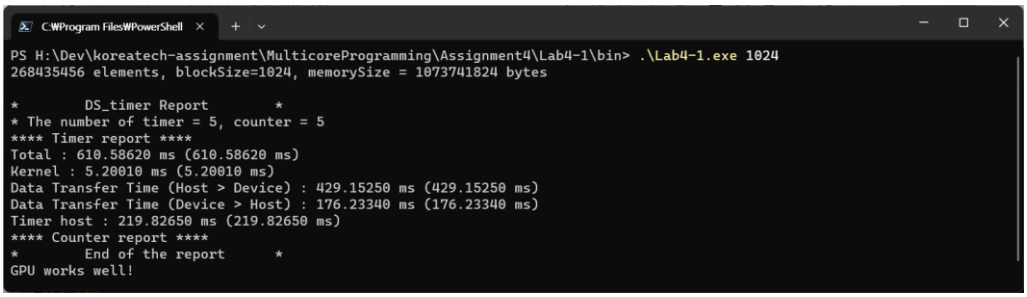
목차

1. 실행 결과

Lab 4-1. Matrix x Vector



cudaMalloc, cudaMemcpy, cudaFree등의 함수를 이용하여 디바이스 측에 CPU 메모리에 있는 데이터를 적절히 복사하고, __global__ 예약어와 함께 선언한 vectorAdd 함수를 이용하여 GPU 측에서 벡터합 연산을 구현하였고, 이 과정에서 blockIdx, blockDim를 이용하여 적절한 tID를 계산하여 사용하였다.

dataSize	Total	Kernel	Data Transfer Time (Host > Device)	Data Transfer Time (Device > Host)	Host
1	2.9192	0.3564	1.7053	0.8574	1.2497
2	5.6175	0.4589	3.5371	1.6213	2.8203
4	10.3014	0.4776	6.966	2.8574	5.1191
8	19.8954	0.5729	13.7524	5.57	10.4615
16	39.8494	0.7007	27.9965	11.1518	20.8217
32	83.4914	0.9806	55.4371	27.0272	43.8481
64	162.829	1.548	108.7351	52.545	86.2251
128	323.899	2.75	217.8379	103.3105	167.758

[데이터를 2배씩 증가시켜가며 테스트한 프로그램 실행 결과 (RTX 3080)]

dataSize: 프로그램에서 사용하는 데이터 크기 (1024 * 1024 * dataSize)

Total: Kernel + Data Transfer Time + Host 시간

Kernel: 커널 부분 (벡터 합) 부분만의 실행 시간

Data Transfer Time: 데이터 이동간 걸리는 시간

Host: 호스트의 벡터 합 소요 시간

*. 시간 단위는 밀리 초를 사용하였다.

데이터 크기가 최대 128배가 되는 동안에서도 커널 부분의 실행 시간은 7~8배 밖에 늘어나지 않았고, 데이터 크기에 따라서 Data Transfer Time이 선형적으로 증가하여 Host와 비교하였을 때 결과적으로 실행 시간이 2배 가까이 소요된 것을 확인할 수 있었다.

Lab 4-2. Matrix Addition

```
#define TO_INDEX(x, y, z, w, dimX, dimY, dimZ) (x + dimX * (y + dimY * (z + dimZ * w)))

__global__ void vectorAdd(const int *a, const int *b, int *c, int size) {
    unsigned int tID = TO_INDEX(threadIdx.x, threadIdx.y, blockIdx.x, blockIdx.y, blockDim.x, blockDim.y, gridDim.x);
    if (tID < size) {
        c[tID] = a[tID] + b[tID];
    }
}
```

8192 x 8192 행렬을 1차원 배열로 선언하고, Lab4-1과 유사한 방법으로 초기화하였다. 2차원 grid와, 2차원 block 크기를 4차원 벡터로 생각하고, 이를 1차원 색인으로 변환하는 TO_INDEX 매크로를 구현하여 벡터합 연산을 구현하였다.

```
int count = 0;
for (int gridX = 0; gridX <= LIMIT - 2; gridX++) {
    for (int gridY = 0; gridY <= LIMIT - gridX - 1; gridY++) {
        for (int blockX = 0; blockX <= LIMIT - gridX - gridY; blockX++) {
            int blockY = LIMIT - gridX - gridY - blockX;
            if (blockX + blockY <= 10 && gridY < 16) {
                // blockX + blockY <= 1024 && gridY < 65536
                dim3 dimBlock(1 << blockX, 1 << blockY, 1);
                dim3 dimGrid(1 << gridX, 1 << gridY, 1);
                printf("[#%d] ", ++count);
                runVectorAdd(dimGrid, dimBlock);
                timer.printToFile((char *) toReportFileName(gridX, gridY, blockX, blockY).c_str());
            }
        }
    }
}
```

Table 15. Technical Specifications per Compute Capability

	Compute Capability													
Technical Specifications	5.0	5.2	5.3	6.0	6.1	6.2	7.0	7.2	7.5	8.0	8.6	8.7	8.9	9.0
Maximum number of resident grids per device (Concurrent Kernel Execution)	32		16	128	32	16	128	16	128					
Maximum dimensionality of grid of thread blocks	3													
Maximum x -dimension of a grid of thread blocks	2^{31}-1													
Maximum y- or z-dimension of a grid of thread blocks	65535													
Maximum dimensionality of thread block	3													
Maximum x- or y-dimensionality of a block	1024													
Maximum z-dimension of a block	64													
Maximum number of threads per block	1024													

Compute Capability 표에서 지원하는 범위 내에서 가능한 모든 Grid와, Block 차원의 Thread Layout을 테스트했다. (2의 배수 단위로 샘플링함, RTX 3080에서 테스트됨)

측정 결과 정리 파일: https://l.abstr.net/mcp-lab4-result

realGridX: 그리드 X 차원 크기

realGridY: 그리드 Y 차원 크기

gridSize: 그리드 총 크기 (X * Y)

realBlockX: 블록 X 차원 크기

realBlockY: 블록 Y 차원 크기

blockSize: 블록 총 크기 (X * Y)

*. 기타 시간을 나타내는 열 부분은 Lab4-1의 설명과 동일하다.

realGridX	realGridY	gridSize	realBlockX	realBlockY	blockSize	Total	Kernel	Data Transfer Time (Host > Device)	Data Transfer Time (Device > Host)	Host
65536	1	65536	1024	1	1024	149.159	1.5389	105.9215	41.6982	32.8133
131072	1	131072	512	1	512	151.0839	1.5374	107.2567	42.2895	35.1862
262144	1	262144	256	1	256	149.2994	1.5634	105.9303	41.8053	32.8615
524288	1	524288	128	1	128	149.7313	1.5613	106.4681	41.7016	33.6141
1048576	1	1048576	64	1	64	149.1004	1.5622	105.8509	41.687	32.7029
2097152	1	2097152	32	1	32	150.8554	1.7197	106.8752	42.26	33.8877
4194304	1	4194304	16	1	16	153.9683	3.0034	108.6254	42.3391	33.5411
8388608	1	8388608	8	1	8	153.102	5.4986	105.913	41.6901	35.2776
16777216	1	16777216	4	1	4	159.3569	10.4042	107.2648	41.6875	32.9044

[1D Grid x 1D Block]

realGridX	realGridY	gridSize	realBlockX	realBlockY	blockSize	Total	Kernel	Data Transfer Time (Host > Device)	Data Transfer Time (Device > Host)	Host
1024	1024	1048576	64	1	64	149.0834	1.5442	105.8096	41.7294	35.2411
1024	2048	2097152	32	1	32	150.1313	1.7116	106.6946	41.7246	32.7985
1024	4096	4194304	16	1	16	150.0222	3.0516	105.2628	41.7074	32.6623
1024	8192	8388608	8	1	8	153.0677	5.4591	105.8588	41.7492	33.4715
1024	16384	16777216	4	1	4	158.662	10.3994	106.5709	41.6914	32.6286
1024	32768	33554432	2	1	2	167.3419	20.3754	105.2659	41.7	33.191
1024	64	65536	1024	1	1024	151.9221	1.482	106.9659	43.4251	33.2131
1024	128	131072	512	1	512	150.1855	1.5312	106.9414	41.7126	33.1696
1024	256	262144	256	1	256	151.3628	1.5252	108.152	41.6851	33.7085
1024	512	524288	128	1	128	150.1005	1.527	106.8677	41.7052	33.1698
2048	1024	2097152	32	1	32	151.7289	1.6898	107.6913	42.3467	34.032
2048	2048	4194304	16	1	16	151.5136	3.0665	106.102	42.3445	32.7887
2048	4096	8388608	8	1	8	154.6873	5.4344	107.5522	41.7005	33.5044
2048	8192	16777216	4	1	4	157.8145	10.3995	105.7149	41.6998	32.8088
2048	16384	33554432	2	1	2	168.0877	20.3701	105.9757	41.7415	33.5081
2048	32	65536	1024	1	1024	150.1371	1.5898	106.8624	41.6847	32.7894
2048	64	131072	512	1	512	149.9983	1.5611	106.7342	41.7027	32.996
2048	128	262144	256	1	256	148.3337	1.5647	105.0757	41.693	33.6632
2048	256	524288	128	1	128	151.3278	1.5337	107.602	42.1919	33.91
2048	512	1048576	64	1	64	150.3072	1.5484	106.996	41.7626	32.895
4096	1024	4194304	16	1	16	150.6437	3.0551	105.6377	41.9507	33.5068
4096	2048	8388608	8	1	8	155.6189	5.5053	107.3799	42.7333	33.6015
4096	4096	16777216	4	1	4	159.9548	10.8034	106.9578	42.1931	33.262
4096	8192	33554432	2	1	2	170.1283	20.4069	107.4082	42.3124	33.5081
4096	16	65536	1024	1	1024	151.2654	1.5519	107.967	41.7458	32.9447
4096	32	131072	512	1	512	150.6089	1.5783	106.6936	42.3363	35.44
4096	64	262144	256	1	256	150.5656	1.5667	107.3187	41.6799	33.1055
4096	128	524288	128	1	128	149.3765	1.5603	106.112	41.7039	32.9265
4096	256	1048576	64	1	64	149.8538	1.5869	106.5798	41.6868	32.7519
4096	512	2097152	32	1	32	150.7138	1.7147	107.2942	41.7046	32.768

[2D Grid x 1D Block]

realGridX	realGridY	gridSize	realBlockX	realBlockY	blockSize	Total	Kernel	Data Transfer Time (Host > Device)	Data Transfer Time (Device > Host)	Host
1024	1024	1048576	1	64	64	150.1653	1.563	106.8645	41.7374	32.8567
1024	1024	1048576	2	32	64	149.251	1.6038	105.8909	41.7561	32.9327
1024	1024	1048576	4	16	64	150.0914	1.5475	106.8136	41.73	32.7644
1024	1024	1048576	8	8	64	149.8771	1.5848	106.5563	41.7358	32.7738
1024	1024	1048576	16	4	64	149.272	1.5453	105.9858	41.7406	34.6212
1024	1024	1048576	32	2	64	149.829	1.5546	106.5367	41.7374	32.8915
1024	1024	1048576	64	1	64	149.0834	1.5442	105.8096	41.7294	35.2411
1024	2048	2097152	1	32	32	149.9759	1.7173	106.5124	41.7459	33.7192
1024	2048	2097152	2	16	32	149.5361	1.7242	106.0771	41.7344	32.9455
1024	2048	2097152	4	8	32	149.4766	1.7152	106.0255	41.7356	33.2193
1024	2048	2097152	8	4	32	151.5839	1.719	108.1326	41.7321	33.1783
1024	2048	2097152	16	2	32	149.5076	1.7168	106.0431	41.7475	33.5158
1024	2048	2097152	32	1	32	150.1313	1.7116	106.6946	41.7246	32.7985
1024	4096	4194304	1	16	16	149.7402	2.9987	104.9789	41.7623	33.4328
1024	4096	4194304	2	8	16	150.4573	3.0247	105.7011	41.731	32.8686
1024	4096	4194304	4	4	16	150.4603	3.0197	105.6587	41.7809	36.1012
1024	4096	4194304	8	2	16	150.8299	3.0313	106.061	41.7373	33.0808
1024	4096	4194304	16	1	16	150.0222	3.0516	105.2628	41.7074	32.6623
1024	8192	8388608	1	8	8	152.4965	5.4689	105.2816	41.7457	32.8539
1024	8192	8388608	2	4	8	154.219	5.4865	106.9808	41.7514	33.0574
1024	8192	8388608	4	2	8	153.2124	5.4803	105.9951	41.7367	32.7117
1024	8192	8388608	8	1	8	153.0677	5.4591	105.8588	41.7492	33.4715
1024	16384	16777216	1	4	4	161.1374	10.5035	107.1187	43.5146	32.8504
1024	16384	16777216	2	2	4	159.0376	10.4916	106.1616	42.384	33.5009
1024	16384	16777216	4	1	4	158.662	10.3994	106.5709	41.6914	32.6286
1024	32768	33554432	1	2	2	168.4777	20.3813	106.4001	41.696	32.73
1024	32768	33554432	2	1	2	167.3419	20.3754	105.2659	41.7	33.191

[2D Grid x 2D Block]

realGridX	realGridY	gridSize	realBlockX	realBlockY	blockSize	Total	Kernel	Data Transfer Time (Host > Device)	Data Transfer Time (Device > Host)	Host
1024	32768	33554432	1	2	2	168.4777	20.3813	106.4001	41.696	32.73
1024	32768	33554432	2	1	2	167.3419	20.3754	105.2659	41.7	33.191
2048	16384	33554432	1	2	2	169.8434	20.3716	106.7065	42.7649	33.2341
2048	16384	33554432	2	1	2	168.0877	20.3701	105.9757	41.7415	33.5081
4096	8192	33554432	1	2	2	169.1823	20.3839	106.5525	42.2455	32.9024
4096	8192	33554432	2	1	2	170.1283	20.4069	107.4082	42.3124	33.5081
8192	4096	33554432	1	2	2	168.0052	20.3704	105.9315	41.7027	33.688
8192	4096	33554432	2	1	2	169.0774	20.3646	107.0107	41.7018	33.5293
16384	2048	33554432	1	2	2	168.026	20.3764	105.955	41.6943	32.8536
16384	2048	33554432	2	1	2	168.7329	20.3745	106.6555	41.7025	32.9328
32768	1024	33554432	1	2	2	168.066	20.3738	106.0072	41.6847	33.4067
32768	1024	33554432	2	1	2	168.0802	20.3598	106.0179	41.702	33.2877
65536	512	33554432	1	2	2	169.9348	20.3712	107.873	41.6903	33.0391
65536	512	33554432	2	1	2	167.4003	20.3768	105.2742	41.749	32.9048
131072	256	33554432	1	2	2	168.7426	20.3677	106.6815	41.6929	32.7063
131072	256	33554432	2	1	2	168.0512	20.3721	105.9852	41.6935	32.6374
262144	128	33554432	1	2	2	171.7371	21.7712	107.6648	42.3008	33.3048
262144	128	33554432	2	1	2	170.1778	20.3905	107.5487	42.2376	33.0039
524288	64	33554432	1	2	2	171.0775	20.3483	108.4748	42.254	33.167
524288	64	33554432	2	1	2	168.4384	20.3901	105.8843	42.1637	32.7211
1048576	32	33554432	1	2	2	168.5014	20.3676	106.408	41.7255	33.1525
1048576	32	33554432	2	1	2	167.6408	20.3636	105.5777	41.6989	34.3922
2097152	16	33554432	1	2	2	170.5806	20.3912	107.8518	42.3371	32.7449
2097152	16	33554432	2	1	2	170.5396	20.4049	106.3362	43.798	33.7877

[2D Grid x 2D Block, 그리드 사이즈 내림차순 정렬]

realGridX	realGridY	gridSize	realBlockX	realBlockY	blockSize	Total	Kernel	Data Transfer Time (Host > Device)	Data Transfer Time (Device > Host)	Host
1024	64	65536	1	1024	1024	149.3301	1.5463	106.0499	41.7335	33.0813
1024	64	65536	1024	1	1024	151.9221	1.482	106.9659	43.4251	33.2131
1024	64	65536	2	512	1024	150.688	1.5378	107.4073	41.7427	33.1187
1024	64	65536	4	256	1024	148.9187	1.5617	105.6066	41.7501	33.558
1024	64	65536	8	128	1024	149.4749	1.6908	105.9015	41.8824	33.2655
1024	64	65536	16	64	1024	150.0685	1.5482	106.7473	41.7726	32.8139
1024	64	65536	32	32	1024	148.8928	1.5346	105.6185	41.7395	35.6277
1024	64	65536	64	16	1024	149.8556	1.5454	106.5702	41.7399	33.1936
1024	64	65536	128	8	1024	149.0548	1.5414	105.7745	41.7387	32.8011
1024	64	65536	256	4	1024	151.3118	1.5158	107.3247	42.4711	33.9577
1024	64	65536	512	2	1024	151.0264	1.4978	107.1716	42.3565	33.2725
2048	32	65536	1	1024	1024	150.6364	1.5272	106.8744	42.2344	34.5762
2048	32	65536	1024	1	1024	150.1371	1.5898	106.8624	41.6847	32.7894
2048	32	65536	2	512	1024	150.8815	1.5092	107.0314	42.3406	32.912
2048	32	65536	4	256	1024	150.7852	1.5728	106.8831	42.329	32.9972
2048	32	65536	8	128	1024	152.1355	1.5185	108.3801	42.2367	32.9455
2048	32	65536	16	64	1024	150.762	1.5039	106.9575	42.3002	33.0058
2048	32	65536	32	32	1024	152.7535	1.514	108.9063	42.3328	33.0396
2048	32	65536	64	16	1024	150.5319	1.5151	106.7565	42.26	32.6053
2048	32	65536	128	8	1024	151.3348	1.5032	107.4514	42.3799	35.7344
2048	32	65536	256	4	1024	150.9085	1.516	106.9484	42.4437	33.6489
2048	32	65536	512	2	1024	150.0881	1.5009	106.3086	42.2784	33.1575
4096	16	65536	1	1024	1024	153.366	1.516	109.5537	42.2961	33.1492
4096	16	65536	1024	1	1024	151.2654	1.5519	107.967	41.7458	32.9447

[2D Grid x 2D Block, 블록 사이즈 내림차순 정렬]

realGridX	realGridY	gridSize	realBlockX	realBlockY	blockSize	Total	Kernel	Data Transfer Time (Host > Device)	Data Transfer Time (Device > Host)	Host
256	256	65536	1	1024	1024	152.6984	1.457	107.5827	43.6584	33.1731
32	4096	131072	16	32	512	149.9573	1.4657	106.769	41.7224	32.7314
2048	128	262144	16	16	256	149.4238	1.466	106.2294	41.7282	32.9686
1024	64	65536	1024	1	1024	151.9221	1.482	106.9659	43.4251	33.2131
256	256	65536	1024	1	1024	151.1878	1.4836	107.3108	42.3929	33.2907
32	8192	262144	256	1	256	150.8553	1.4869	107.0875	42.2804	33.4261
32	4096	131072	512	1	512	150.2933	1.4944	106.5086	42.2899	33.3534
512	2048	1048576	64	1	64	151.0916	1.4945	107.3101	42.2865	33.596
512	512	262144	256	1	256	152.6607	1.4964	108.8768	42.287	33.4045
1024	64	65536	512	2	1024	151.0264	1.4978	107.1716	42.3565	33.2725
16	8192	131072	512	1	512	150.623	1.4989	106.8235	42.3001	33.3906
64	1024	65536	512	2	1024	150.983	1.499	107.1407	42.3429	32.8107
256	4096	1048576	8	8	64	150.7396	1.499	106.7651	42.4755	36.4712
8	16384	131072	1	512	512	150.3843	1.5007	106.5786	42.3046	33.0339
2048	32	65536	512	2	1024	150.0881	1.5009	106.3086	42.2784	33.1575
2048	32	65536	128	8	1024	151.3348	1.5032	107.4514	42.3799	35.7344
512	2048	1048576	16	4	64	149.0591	1.5037	105.8319	41.7231	32.6111
2048	32	65536	16	64	1024	150.762	1.5039	106.9575	42.3002	33.0058
16	32768	524288	32	4	128	150.8542	1.5046	107.0407	42.3087	32.9362
64	1024	65536	1024	1	1024	149.3714	1.5056	106.135	41.7303	33.5911
65536	1	65536	16	64	1024	153.2542	1.5065	109.0079	42.7393	33.1557
524288	1	524288	4	32	128	150.3732	1.5074	106.4499	42.4154	33.0605
64	16384	1048576	64	1	64	150.4922	1.5079	106.4202	42.5638	33.5895
2048	32	65536	2	512	1024	150.8815	1.5092	107.0314	42.3406	32.912

[2D Grid x 2D Block, 커널 시간 순 오름차순 정렬]

측정 결과 차원 형태에 따른 성능 변화는 거의 나타나지 않았다. 너무 작은 블록 사이즈를 사용하는 경우 성능 저하가 나타났고 관측 결과 기준 32개보다 적은 블록 크기를 사용하기 시작하면서 큰 성능 하락이 관측되었다. 커널 최고 성능이 관측된 그리드와 블록 크기는 **(Grid: 256x256), (Block: 1 x 1024)**였다. 전반적으로 **Grid Size**: 65536~1048576,

Block Size: 512~1024 구간에 있는 케이스들의 성능이 좋게 나타났다.

2. 소감

이번 과제를 수행하면서 GPU를 이용한 프로그래밍이 어떤 방식으로 이루어지는지에 대해서 배울 수 있었다. 시간을 측정하면서 실제 코어 로직 부분이 되는 커널 부분의 소요 시간이 아주 짧은 것을 계속 확인할 수 있었는데, 복잡한 문제를 GPU에서 계산하면 데이터 전송 오버헤드보다 훨씬 큰 이득을 볼 수 있을 것임을 느낄 수 있었다. 향후 과제에서 더 복잡한 문제를 GPU를 이용하여 CPU 시간보다 훨씬 빠른 시간에 문제를 해결하는 것을 직접 눈으로 볼 수 있을 것이라 생각하니 기대가 된다.