Problem 1 - PALEO, FLOPs, Platform Percent of Peak (PPP) 15 points

1. Why achieving peak FLOPs from hardware devices like GPUs is a difficult propostion in real systems? How does PPP help in capturing this inefficiency captured in Paleo model. (3)

Answer

- To achieve peak FLOPs, it usually requires customized libraries developed by organizations with intimate knowledge of the underlying hardware. Even these specially tuned libraries may fall short of peak execution by as much as 40%
- PPP helps in capturing the inefficiency by displaying the current FLOPs as a percentage of the theoretical maximum afforded by the hardware and software.

2. Lu et al. showed that FLOPs consumed by convolution layers in VG16 account for about 99% of the total FLOPS in the forward pass. We will do a similar analysis for VGG19. Calculate FLOPs for different layers in VGG19 and then calculate fraction of the total FLOPs attributed by convolution layers. (5)

- I used ptflops python library to compute the total FLOPs
- The total FLOPs of VGG19 are 19.67 billion
- It did not match because asynchronous computation is allowed by CUDA, however, the cores on the GPU were synchronized, which resulted in the timing differences.
- The approach adopted is run matrix multiplications asynchronously for multiple iterations to make sure that the impact of the overhead is negligible.
- Add % of Flops of all Conv2d layers
    - 0.457% + 9.421%+ 4.710%+ 9.413%+ 4.706%+ 9.409%+ 9.409%+ 9.409%+ 4.704%+ 9.407%+ 9.407%+ 9.407%+ 2.352%+ 2.352%+ 2.352%+ 2.352%
    - Total % of flops of all convolution layers equal to 89.85%
- Below includes the Raw output of the flops calculation by running “python flop_count.py”

```
(/scratch/zz3904/envs_dirs/html) [zz3904@gv015 part1]$ python flop_count.py
Warning: module Dropout is treated as a zero-op.
Warning: module VGG is treated as a zero-op.
VGG(
  143.667 M, 100.000% Params, 19.668 GMac, 100.000% MACs,
  (features): Sequential(
    20.024 M, 13.938% Params, 19.544 GMac, 99.371% MACs,
    (0): Conv2d(0.002 M, 0.001% Params, 0.09 GMac, 0.457% MACs, 3, 64, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
Warning: module Dropout is treated as a zero-op.
    (1): ReLU(0.0 M, 0.000% Params, 0.003 GMac, 0.016% MACs, inplace=True)
    (2): Conv2d(0.037 M, 0.026% Params, 1.853 GMac, 9.421% MACs, 64, 64, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (3): ReLU(0.0 M, 0.000% Params, 0.003 GMac, 0.016% MACs, inplace=True)
    (4): MaxPool2d(0.0 M, 0.000% Params, 0.003 GMac, 0.016% MACs, kernel_size=2, stride=2, padding=0, dilation=1, ceil_mode=False)
    (5): Conv2d(0.074 M, 0.051% Params, 0.926 GMac, 4.710% MACs, 64, 128, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (6): ReLU(0.0 M, 0.000% Params, 0.002 GMac, 0.008% MACs, inplace=True)
    (7): Conv2d(0.148 M, 0.103% Params, 1.851 GMac, 9.413% MACs, 128, 128, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (8): ReLU(0.0 M, 0.000% Params, 0.002 GMac, 0.008% MACs, inplace=True)
    (9): MaxPool2d(0.0 M, 0.000% Params, 0.002 GMac, 0.008% MACs, kernel_size=2, stride=2, padding=0, dilation=1, ceil_mode=False)
    (10): Conv2d(0.295 M, 0.205% Params, 0.926 GMac, 4.706% MACs, 128, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (11): ReLU(0.0 M, 0.000% Params, 0.001 GMac, 0.004% MACs, inplace=True)
    (12): Conv2d(0.59 M, 0.411% Params, 1.85 GMac, 9.409% MACs, 256, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (13): ReLU(0.0 M, 0.000% Params, 0.001 GMac, 0.004% MACs, inplace=True)
    (14): Conv2d(0.59 M, 0.411% Params, 1.85 GMac, 9.409% MACs, 256, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (15): ReLU(0.0 M, 0.000% Params, 0.001 GMac, 0.004% MACs, inplace=True)
    (16): Conv2d(0.59 M, 0.411% Params, 1.85 GMac, 9.409% MACs, 256, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (17): ReLU(0.0 M, 0.000% Params, 0.001 GMac, 0.004% MACs, inplace=True)
    (18): MaxPool2d(0.0 M, 0.000% Params, 0.001 GMac, 0.004% MACs, kernel_size=2, stride=2, padding=0, dilation=1, ceil_mode=False)
    (19): Conv2d(1.18 M, 0.821% Params, 0.925 GMac, 4.704% MACs, 256, 512, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (20): ReLU(0.0 M, 0.000% Params, 0.0 GMac, 0.002% MACs, inplace=True)
    (21): Conv2d(2.36 M, 1.643% Params, 1.85 GMac, 9.407% MACs, 512, 512, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (22): ReLU(0.0 M, 0.000% Params, 0.0 GMac, 0.002% MACs, inplace=True)
    (23): Conv2d(2.36 M, 1.643% Params, 1.85 GMac, 9.407% MACs, 512, 512, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (24): ReLU(0.0 M, 0.000% Params, 0.0 GMac, 0.002% MACs, inplace=True)
    (25): Conv2d(2.36 M, 1.643% Params, 1.85 GMac, 9.407% MACs, 512, 512, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (26): ReLU(0.0 M, 0.000% Params, 0.0 GMac, 0.002% MACs, inplace=True)
    (27): MaxPool2d(0.0 M, 0.000% Params, 0.0 GMac, 0.002% MACs, kernel_size=2, stride=2, padding=0, dilation=1, ceil_mode=False)
    (28): Conv2d(2.36 M, 1.643% Params, 0.463 GMac, 2.352% MACs, 512, 512, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (29): ReLU(0.0 M, 0.000% Params, 0.0 GMac, 0.001% MACs, inplace=True)
    (30): Conv2d(2.36 M, 1.643% Params, 0.463 GMac, 2.352% MACs, 512, 512, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (31): ReLU(0.0 M, 0.000% Params, 0.0 GMac, 0.001% MACs, inplace=True)
    (32): Conv2d(2.36 M, 1.643% Params, 0.463 GMac, 2.352% MACs, 512, 512, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (33): ReLU(0.0 M, 0.000% Params, 0.0 GMac, 0.001% MACs, inplace=True)
    (34): Conv2d(2.36 M, 1.643% Params, 0.463 GMac, 2.352% MACs, 512, 512, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (35): ReLU(0.0 M, 0.000% Params, 0.0 GMac, 0.001% MACs, inplace=True)
    (36): MaxPool2d(0.0 M, 0.000% Params, 0.0 GMac, 0.001% MACs, kernel_size=2, stride=2, padding=0, dilation=1, ceil_mode=False)
  )
  (avgpool): AdaptiveAvgPool2d(0.0 M, 0.000% Params, 0.0 GMac, 0.000% MACs, output_size=(7, 7))
  (classifier): Sequential(
    123.643 M, 86.062% Params, 0.124 GMac, 0.629% MACs,
    (0): Linear(102.765 M, 71.530% Params, 0.103 GMac, 0.522% MACs, in_features=25088, out_features=4096, bias=True)
    (1): ReLU(0.0 M, 0.000% Params, 0.0 GMac, 0.000% MACs, inplace=True)
    (2): Dropout(0.0 M, 0.000% Params, 0.0 GMac, 0.000% MACs, p=0.5, inplace=False)
    (3): Linear(16.781 M, 11.681% Params, 0.017 GMac, 0.085% MACs, in_features=4096, out_features=4096, bias=True)
('19.67 GMac', '143.67 M')
```

3. Study the tables showing timing benchmarks from Alexnet (Table 2), VGG16 (Table 3), Googlenet (Table 5), and Resnet50 (Table 6). Why the measured time and sum of layerwise timings for forward pass did not match on GPUs ? What approach was adopted in Sec. 5 of the paper to mitigate the measurement overhead in GPUs. (2+2)

Answer

- CUDA supports asynchronous programming. Before time measurement, an API (cudaDeviceSynchronize) has to be called to make sure that all cores have finished their tasks. This explicit synchronization is the overhead of measuring time on the GPUs.Therefore, the sum of layerwise timing on GPUs is longer than a full forward pass.
- The paper proposed a estimation tool called Augur to mitigate the measure overhead in GPUs
    - 1, Augur parses the descriptor of a CNN.
        - Based the type and setting of each layer, it calculates the minimal memory needed to run the CNN. The memory includes data, parameters, and workspace.

    - 2, Augur extracts matmuls from the computation of the CNN.
    - 3, Augur calculates the compute time of individual matmuls and then uses their summation as the estimate of the compute time of the CNN.

4. In Lu et al. FLOPs for different layers of a DNN are calculated. Use FLOPs numbers for VGG16 (Table 3), Googlenet (Table 5), and Resnet50 (Table 6), calculate the inference time (time to have a forward pass with one image) using published Tflops number for K80 (Refer to NVIDIA TESLA GPU Accelerators) both for single-precision and double-precision calculations. (3)

Answer

- Tesla K80: double precision peak performance of 1.87 T flops.
- VGG requires 15503M FLOPs:
    - 1 forward pass takes (15503 * 10^6) / (1.87 * 10^12) = 0.00829037433s
    - Throughput = 1/0.00829037433s = 120 images/second
- GoogLeNet requires 1606 M FLOPs:
    - 1 forward pass takes (1606 * 10^6) / (1.87 * 10^12) = 0.00085882353s
    - Throughput = 1/0.00085882353s = 1164 images/second
- Resnet requires 3922 M FLOPs:
    - 1 forward pass takes (3922 * 10^6) / (1.87 * 10^12) = 0.0020973262s
    - Throughput = 1/0.0020973262s = 477 images/second

Problem 2 - TTA metric, Stability, Generalization 15 points

Calculate the coefficient of variation of TTA for both the hardware configurations. Compare the value you obtain with that reported in Table 3(a) in the paper by Coleman et al for Resnet-50, 1xTPU.

Number	**RTX8000**	**V100**
1	62.76	45.72
2	77.38	52.47
3	76.89	48.85
4	63.17	49.22
5	63.35	48.98
coefficient of variation	11.2004938%	4.8739164%

- Deatiled result: https://drive.google.com/drive/folders/1JZUzk0xCMUPTIUt7k5Lvc-pYbtfykmmQ?usp=sharing
- Comparing the coefficient of variation of TTA reported in Table 3(a), RTX8000 is higher and V100 is lower than reported.

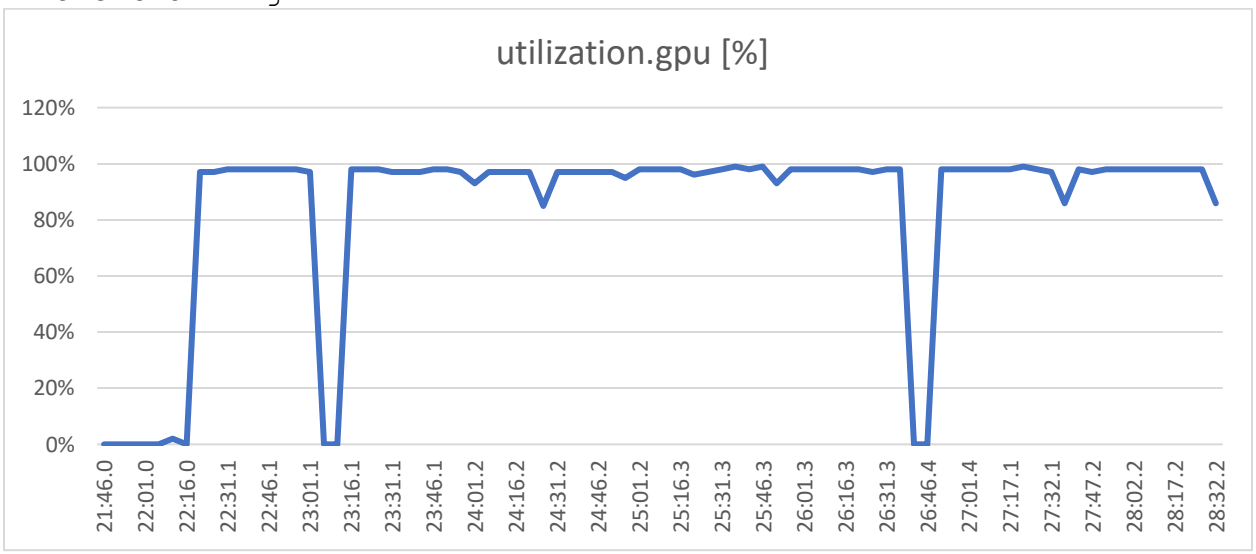
2. Collect 5 images from the wild for each of the 10 categories in CIFAR10 and manually label them.
Again in a group of 5 students, each one of you can collect 5 images for 2 categories. In this way each one of you need to collect only 10 images (5 for each of the 2 categories that you choose). These images should not be from CIFAR10 dataset. Next test the trained models for these 50 images. What is the accuracy obtained from each of the 10 trained models ? Quantify the mean and standard deviation of accuracy obtained using the 10 models for RTX8000 and the 10 models for V100. (4+4)

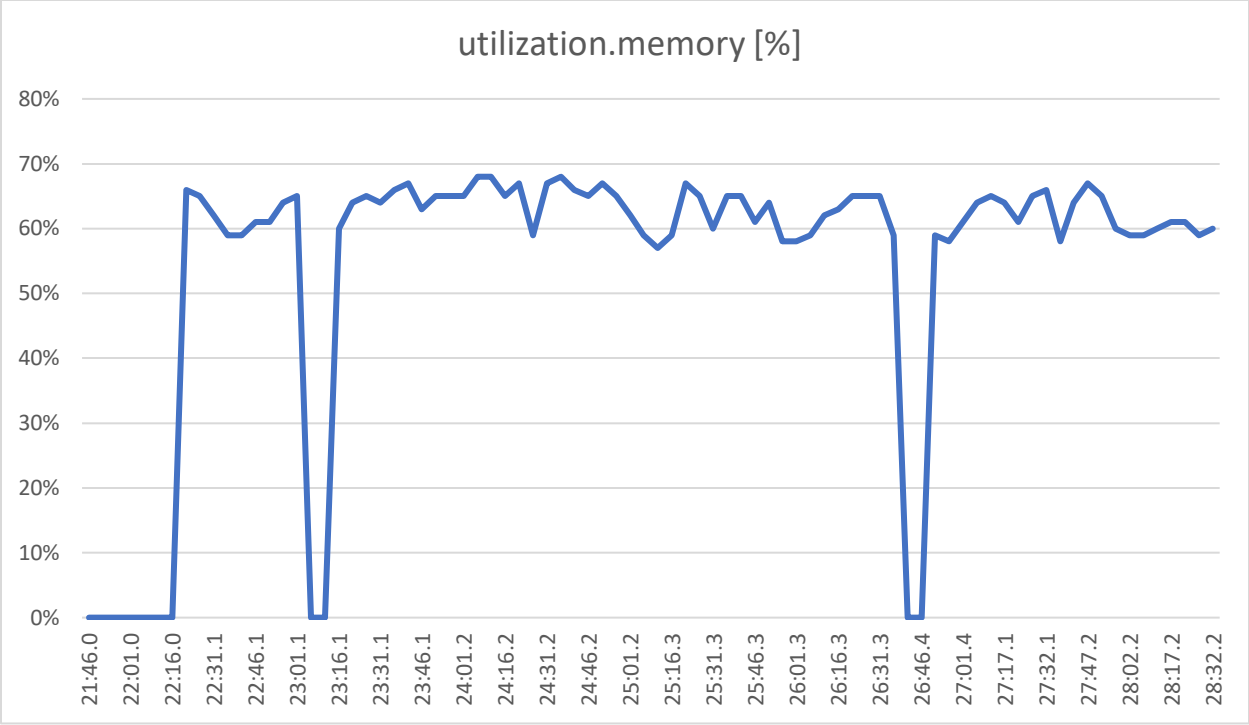
Number	**RTX8000**	**V100**
1	100%	100%
2	100%	95%
3	100%	95%
4	100%	95%
5	100%	95%
Mean	100%	96%
Standard deviation	0	0.020412415

- Detailed result: https://drive.google.com/file/d/1lieEEjHLF8xtyiR1ZrTNg0nKj5BE_MyQ/view?usp=sharing

3. Measure the GPU utilization using nvidia-smi while training with V100 and report both the time series of GPU core and memory utilization and their average over a period of 3 mins of training. Is the GPU utilization close to 100% ? (2)

- The GPU utilization is close to 100% during some time of the training

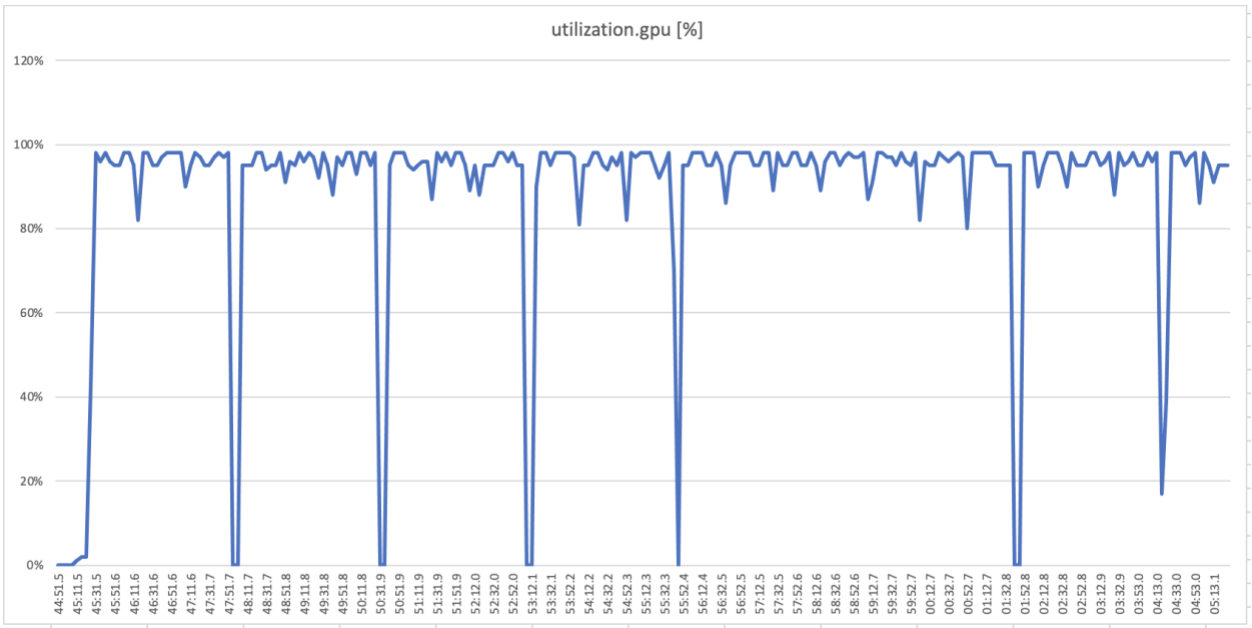


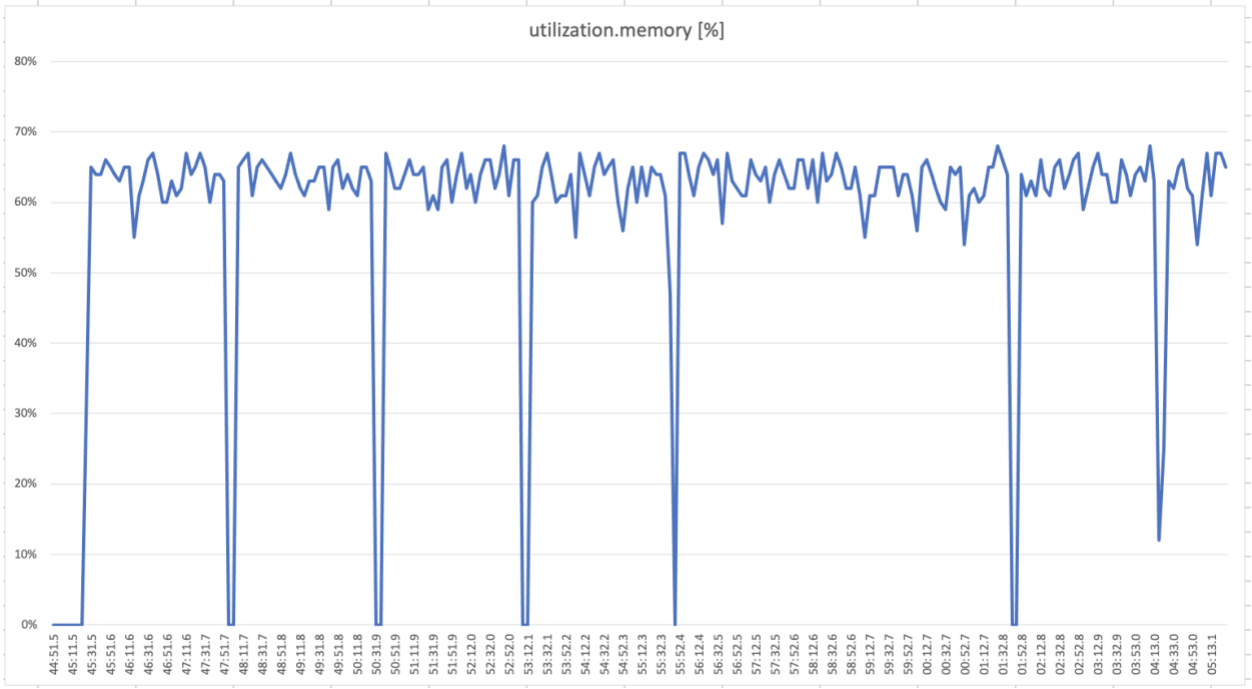


4. If the GPU utilization is low, what can you do to increase the GPU utilization? Try your trick(s) and report if you are successful or not in driving GPU utilization close to 100%. (2)

- I used DataParallel
- I set cudnn.benchmark to True to increase GPU utilization.

```
self.net = torch.nn.DataParallel(self.net)
cudnn.benchmark = True
```





Problem 3 - C++, CUDA, Unified Memory 20 points

Question 1-3:

- `README file includes all the instructions regarding how to execute my code for those questions`

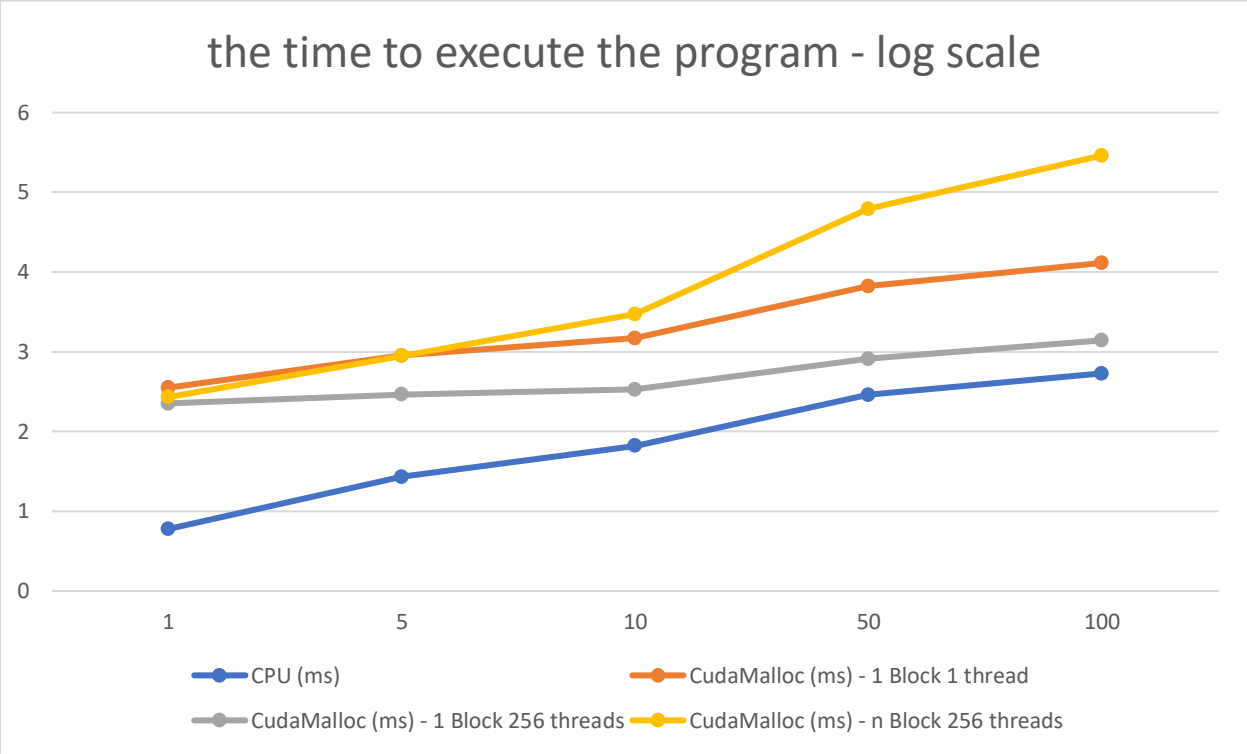
Question 4. Plot two charts one for Step 2 (without Unified Memory) and one for Step 3 (with Unified Memory). The x-axis is the value of K (in million) and y-axis is the time to execute the program (one chart for each of the three scenarios). In both the charts also plot the time to

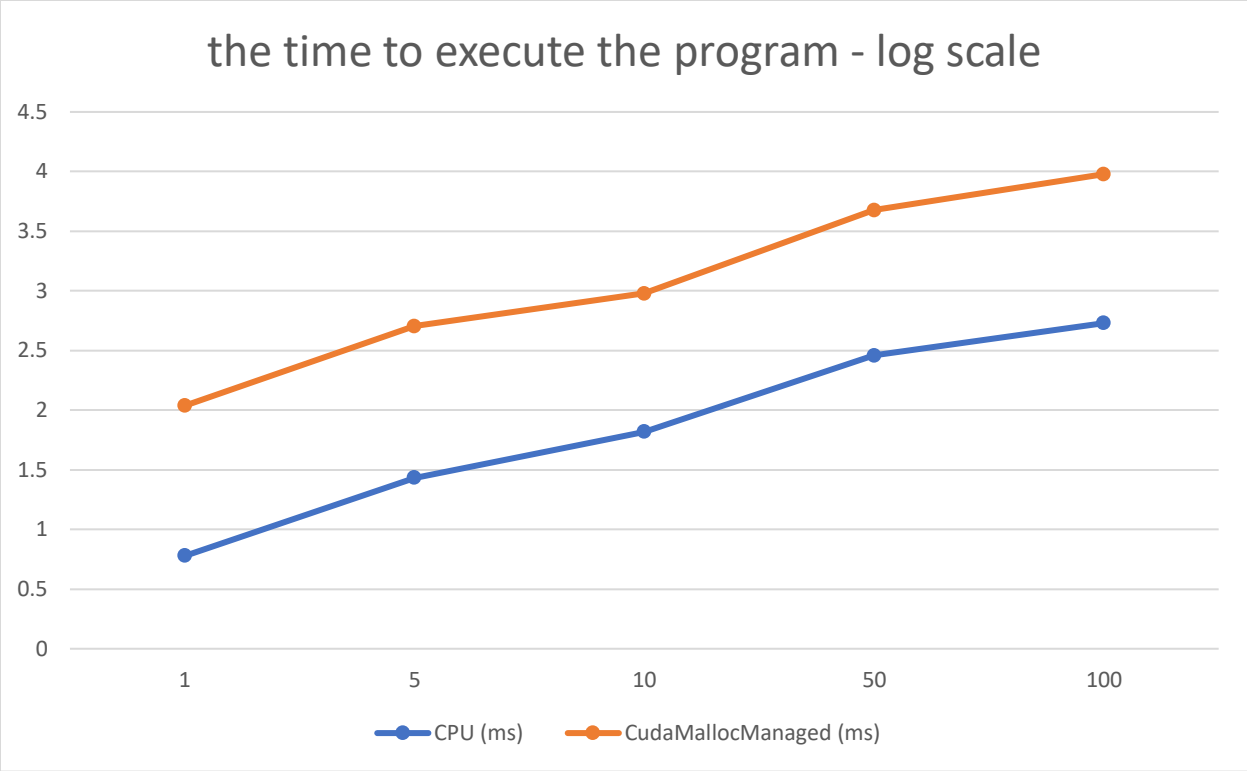
execute on CPU only. Since the execution time scale on CPU and GPU may be orders of magnitude different, you may want to use log-log scale for y-axis when plotting. (4)

K	CPU (ms)	CudaMallocManaged (ms)	CudaMalloc (ms) - 1 Block 1 thread	CudaMalloc (ms) - 1 Block 256 threads	CudaMalloc (ms) - n Block 256 threads
1	6	109.03	355.037	225.4931	270.238
5	27	504.54	894.405	290.167	888.551
10	66	950.68	1485.137	335.608	2960.307
50	288	4759.47	6631.614	819.15	62211.33
100	536	9510.2	13003.64	1392.12	289175.829

log scale

K	CPU (ms)	CudaMallocManaged (ms)	CudaMalloc (ms) - 1 Block 1 thread	CudaMalloc (ms) - 1 Block 256 threads	CudaMalloc (ms) - n Block 256 threads
1	0.77815125	2.037546012	2.550273615	2.353133257	2.431746418
5	1.43136376	2.702895603	2.951534218	2.46264802	2.94868236
10	1.81954394	2.978034358	3.171766518	2.525832305	3.471336752
50	2.45939249	3.677558594	3.82161924	2.913363436	4.793869486
100	2.72916479	3.97818965	4.114064938	3.143676673	5.461161989





Raw Data - Screenshot

CPP

1M

```
real    0m0.014s
user    0m0.006s
sys     0m0.008s
```

5M

```
real    0m0.058s
user    0m0.027s
sys     0m0.031s
```

10M

```
real    0m0.113s
user    0m0.066s
sys     0m0.047s
```

50M

```
real    0m0.555s
user    0m0.288s
sys     0m0.265s
```

100M

```
real    0m1.112s
user    0m0.536s
sys     0m0.574s
```

cudaMalloc – 1 block, 1 thread

1M

```
==3775981== NVPROF is profiling process 3775981, command: ./q3_2
==3775981== Profiling application: ./q3_2
==3775981== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   95.58%  66.515ms         1  66.515ms  66.515ms  66.515ms  vector_add(float*, float*, float*, int)
                    2.40%  1.6717ms         1  1.6717ms  1.6717ms  1.6717ms  [CUDA memcpy DtoH]
                    2.01%  1.4008ms         2  700.41us  700.19us  700.64us  [CUDA memcpy HtoD]
      API calls:   74.64%  214.14ms         3  71.379ms  138.33us  213.85ms  cudaMalloc
                   24.85%  71.310ms         3  23.770ms  875.28us  69.532ms  cudaMemcpy
                    0.17%  492.16us         1  492.16us  492.16us  492.16us  cuDeviceTotalMem
                    0.17%  491.06us         3  163.69us  131.19us  217.60us  cudaFree
                    0.13%  364.59us       101  3.6090us     120ns  162.86us  cuDeviceGetAttribute
                    0.02%  61.867us         1  61.867us  61.867us  61.867us  cuDeviceGetName
                    0.01%  39.171us         1  39.171us  39.171us  39.171us  cudaLaunchKernel
                    0.00%  6.5720us         1  6.5720us  6.5720us  6.5720us  cuDeviceGetPCIBusId
                    0.00%  1.1840us         3     394ns     161ns     862ns  cuDeviceGetCount
                    0.00%     785ns         2     392ns     125ns     660ns  cuDeviceGet
                    0.00%     498ns         1     498ns     498ns     498ns  cuDeviceGetUuid
```

5M

```
==3776069== NVPROF is profiling process 3776069, command: ./q3_2
==3776069== Profiling application: ./q3_2
==3776069== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   93.53%  317.99ms         1  317.99ms  317.99ms  317.99ms  vector_add(float*, float*, float*, int)
                    4.06%  13.791ms         1  13.791ms  13.791ms  13.791ms  [CUDA memcpy DtoH]
                    2.42%  8.2143ms         2  4.1072ms  4.0967ms  4.1176ms  [CUDA memcpy HtoD]
      API calls:   61.37%  341.25ms         3  113.75ms  4.2961ms  332.63ms  cudaMemcpy
                   38.33%  213.16ms         3  71.053ms  157.25us  212.84ms  cudaMalloc
                    0.10%  552.03us         3  184.01us  141.43us  266.01us  cudaFree
                    0.10%  529.89us         1  529.89us  529.89us  529.89us  cuDeviceTotalMem
                    0.09%  475.81us       101  4.7110us     122ns  270.09us  cuDeviceGetAttribute
                    0.01%  44.420us         1  44.420us  44.420us  44.420us  cuDeviceGetName
                    0.01%  31.691us         1  31.691us  31.691us  31.691us  cudaLaunchKernel
                    0.00%  5.8950us         1  5.8950us  5.8950us  5.8950us  cuDeviceGetPCIBusId
                    0.00%     916ns         3     305ns     145ns     613ns  cuDeviceGetCount
                    0.00%     736ns         2     368ns     154ns     582ns  cuDeviceGet
                    0.00%     327ns         1     327ns     327ns     327ns  cuDeviceGetUuid
```

10M

```
==3776127== NVPROF is profiling process 3776127, command: ./q3_2
==3776127== Profiling application: ./q3_2
==3776127== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   93.00%  593.59ms         1  593.59ms  593.59ms  593.59ms  vector_add(float*, float*, float*, int)
                    4.41%  28.141ms         1  28.141ms  28.141ms  28.141ms  [CUDA memcpy DtoH]
                    2.59%  16.546ms         2  8.2730ms  8.2655ms  8.2804ms  [CUDA memcpy HtoD]
      API calls:   75.31%  639.56ms         3  213.19ms  8.4734ms  622.61ms  cudaMemcpy
                   24.41%  207.30ms         3  69.099ms  172.67us  206.94ms  cudaMalloc
                    0.10%  843.77us         1  843.77us  843.77us  843.77us  cuDeviceGetName
                    0.08%  637.24us         3  212.41us  159.72us  312.64us  cudaFree
                    0.06%  498.20us         1  498.20us  498.20us  498.20us  cuDeviceTotalMem
                    0.04%  349.34us       101  3.4580us     116ns  152.98us  cuDeviceGetAttribute
                    0.00%  32.247us         1  32.247us  32.247us  32.247us  cudaLaunchKernel
                    0.00%  5.4970us         1  5.4970us  5.4970us  5.4970us  cuDeviceGetPCIBusId
                    0.00%  4.3660us         2  2.1830us     152ns  4.2140us  cuDeviceGet
                    0.00%  1.0830us         3     361ns     149ns     730ns  cuDeviceGetCount
                    0.00%     279ns         1     279ns     279ns     279ns  cuDeviceGetUuid
```

50M

```
(base) [zz3904@gv016 part1]$ ./q3_2.sh
==3776169== NVPROF is profiling process 3776169, command: ./q3_2
==3776169== Profiling application: ./q3_2
==3776169== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   92.96%  2.97964s         1  2.97964s  2.97964s  2.97964s  vector_add(float*, float*, float*, int)
                    4.42%  141.81ms         1  141.81ms  141.81ms  141.81ms  [CUDA memcpy DtoH]
                    2.61%  83.764ms         2  41.882ms  41.825ms  41.939ms  [CUDA memcpy HtoD]
      API calls:   93.52%  3.20687s         3  1.06896s  42.025ms  3.12271s  cudaMemcpy
                    6.40%  219.53ms         3  73.176ms  296.94us  218.92ms  cudaMalloc
                    0.05%  1.7318ms         3  577.27us  524.95us  662.35us  cudaFree
                    0.01%  380.22us       101  3.7640us     122ns  181.71us  cuDeviceGetAttribute
                    0.01%  335.67us         1  335.67us  335.67us  335.67us  cuDeviceTotalMem
                    0.00%  52.221us         1  52.221us  52.221us  52.221us  cuDeviceGetName
                    0.00%  40.375us         1  40.375us  40.375us  40.375us  cudaLaunchKernel
                    0.00%  5.8120us         1  5.8120us  5.8120us  5.8120us  cuDeviceGetPCIBusId
                    0.00%  1.1780us         3     392ns     152ns     720ns  cuDeviceGetCount
                    0.00%     986ns         2     493ns     145ns     841ns  cuDeviceGet
                    0.00%     279ns         1     279ns     279ns     279ns  cuDeviceGetUuid
```

100M

```
(base) [zz3904@gv016 part1]$ ./q3_2.sh
==3776258== NVPROF is profiling process 3776258, command: ./q3_2
==3776258== Profiling application: ./q3_2
==3776258== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   92.92%  5.94236s         1  5.94236s  5.94236s  5.94236s  vector_add(float*, float*, float*, int)
                    4.45%  284.30ms         1  284.30ms  284.30ms  284.30ms  [CUDA memcpy DtoH]
                    2.64%  168.56ms         2  84.278ms  84.066ms  84.490ms  [CUDA memcpy HtoD]
      API calls:   96.76%  6.39673s         3  2.13224s  84.312ms  6.22773s  cudaMemcpy
                    3.20%  211.69ms         3  70.562ms  446.85us  210.79ms  cudaMalloc
                    0.03%  1.6775ms         3  559.15us  462.53us  737.47us  cudaFree
                    0.01%  354.31us       101  3.5080us     121ns  156.66us  cuDeviceGetAttribute
                    0.01%  331.98us         1  331.98us  331.98us  331.98us  cuDeviceTotalMem
                    0.00%  48.009us         1  48.009us  48.009us  48.009us  cuDeviceGetName
                    0.00%  38.101us         1  38.101us  38.101us  38.101us  cudaLaunchKernel
                    0.00%  4.9310us         1  4.9310us  4.9310us  4.9310us  cuDeviceGetPCIBusId
                    0.00%     926ns         3     308ns     128ns     631ns  cuDeviceGetCount
                    0.00%     688ns         2     344ns     151ns     537ns  cuDeviceGet
                    0.00%     361ns         1     361ns     361ns     361ns  cuDeviceGetUuid
(base) [zz3904@gv016 part1]$
```

cudaMalloc – 1 block, 256 threads

1M

```
==3775411== NVPROF is profiling process 3775411, command: ./q3_2
==3775411== Profiling application: ./q3_2
==3775411== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   36.97%  1.6764ms         1  1.6764ms  1.6764ms  1.6764ms  [CUDA memcpy DtoH]
                   32.07%  1.4546ms         1  1.4546ms  1.4546ms  1.4546ms  vector_add(float*, float*, float*, int)
                   30.96%  1.4039ms         2  701.95us  701.69us  702.21us  [CUDA memcpy HtoD]
      API calls:   96.45%  214.71ms         3  71.572ms  138.37us  214.42ms  cudaMalloc
                    2.81%  6.2481ms         3  2.0827ms  878.06us  4.4572ms  cudaMemcpy
                    0.39%  878.04us         3  292.68us  263.36us  310.77us  cudaFree
                    0.16%  353.74us       101  3.5020us     116ns  159.54us  cuDeviceGetAttribute
                    0.15%  332.93us         1  332.93us  332.93us  332.93us  cuDeviceTotalMem
                    0.02%  45.791us         1  45.791us  45.791us  45.791us  cuDeviceGetName
                    0.01%  30.891us         1  30.891us  30.891us  30.891us  cudaLaunchKernel
                    0.00%  5.4150us         1  5.4150us  5.4150us  5.4150us  cuDeviceGetPCIBusId
                    0.00%  1.0430us         3     347ns     136ns     752ns  cuDeviceGetCount
                    0.00%     743ns         2     371ns     143ns     600ns  cuDeviceGet
                    0.00%     238ns         1     238ns     238ns     238ns  cuDeviceGetUuid
```

5M

```
==3775514== NVPROF is profiling process 3775514, command: ./q3_2
==3775514== Profiling application: ./q3_2
==3775514== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   46.91%  13.794ms         1  13.794ms  13.794ms  13.794ms  [CUDA memcpy DtoH]
                   27.76%  8.1636ms         2  4.0818ms  4.0757ms  4.0878ms  [CUDA memcpy HtoD]
                   25.33%  7.4478ms         1  7.4478ms  7.4478ms  7.4478ms  vector_add(float*, float*, float*, int)
      API calls:   87.49%  230.10ms         3  76.699ms  157.46us  229.78ms  cudaMalloc
                   11.66%  30.661ms         3  10.220ms  4.2711ms  22.103ms  cudaMemcpy
                    0.37%  967.16us       101  9.5750us     121ns  458.44us  cuDeviceGetAttribute
                    0.20%  530.15us         3  176.72us  139.87us  245.76us  cudaFree
                    0.18%  474.54us         1  474.54us  474.54us  474.54us  cuDeviceTotalMem
                    0.08%  216.86us         1  216.86us  216.86us  216.86us  cuDeviceGetName
                    0.01%  35.930us         1  35.930us  35.930us  35.930us  cudaLaunchKernel
                    0.00%  5.6430us         1  5.6430us  5.6430us  5.6430us  cuDeviceGetPCIBusId
                    0.00%  1.0550us         3     351ns     173ns     679ns  cuDeviceGetCount
                    0.00%     883ns         2     441ns     129ns     754ns  cuDeviceGet
                    0.00%     270ns         1     270ns     270ns     270ns  cuDeviceGetUuid
```

10

```
==3775561== NVPROF is profiling process 3775561, command: ./q3_2
==3775561== Profiling application: ./q3_2
==3775561== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   47.11%  27.966ms         1  27.966ms  27.966ms  27.966ms  [CUDA memcpy DtoH]
                   28.00%  16.620ms         2  8.3099ms  8.3043ms  8.3155ms  [CUDA memcpy HtoD]
                   24.89%  14.772ms         1  14.772ms  14.772ms  14.772ms  vector_add(float*, float*, float*, int)
      API calls:   77.66%  215.62ms         3  71.874ms  180.81us  215.26ms  cudaMalloc
                   21.84%  60.630ms         3  20.210ms  8.4970ms  43.611ms  cudaMemcpy
                    0.22%  618.72us         3  206.24us  164.32us  284.63us  cudaFree
                    0.13%  367.64us       101  3.6390us     118ns  161.53us  cuDeviceGetAttribute
                    0.12%  332.23us         1  332.23us  332.23us  332.23us  cuDeviceTotalMem
                    0.02%  46.386us         1  46.386us  46.386us  46.386us  cuDeviceGetName
                    0.01%  34.289us         1  34.289us  34.289us  34.289us  cudaLaunchKernel
                    0.00%  6.2360us         1  6.2360us  6.2360us  6.2360us  cuDeviceGetPCIBusId
                    0.00%     966ns         3     322ns     154ns     581ns  cuDeviceGetCount
                    0.00%     804ns         2     402ns     156ns     648ns  cuDeviceGet
                    0.00%     245ns         1     245ns     245ns     245ns  cuDeviceGetUuid
```

50M

```
==3775645== NVPROF is profiling process 3775645, command: ./q3_2
==3775645== Profiling application: ./q3_2
==3775645== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   47.44%  141.32ms         1  141.32ms  141.32ms  141.32ms  [CUDA memcpy DtoH]
                   28.18%  83.949ms         2  41.975ms  41.757ms  42.192ms  [CUDA memcpy HtoD]
                   24.38%  72.611ms         1  72.611ms  72.611ms  72.611ms  vector_add(float*, float*, float*, int)
      API calls:   57.26%  299.55ms         3  99.849ms  41.961ms  215.20ms  cudaMemcpy
                   42.38%  221.72ms         3  73.906ms  304.42us  221.11ms  cudaMalloc
                    0.21%  1.0740ms         3  357.98us  301.81us  466.17us  cudaFree
                    0.07%  387.69us       101  3.8380us     120ns  189.59us  cuDeviceGetAttribute
                    0.06%  333.62us         1  333.62us  333.62us  333.62us  cuDeviceTotalMem
                    0.01%  58.052us         1  58.052us  58.052us  58.052us  cuDeviceGetName
                    0.01%  37.746us         1  37.746us  37.746us  37.746us  cudaLaunchKernel
                    0.00%  5.1910us         1  5.1910us  5.1910us  5.1910us  cuDeviceGetPCIBusId
                    0.00%  4.7020us         2  2.3510us     146ns  4.5560us  cuDeviceGet
                    0.00%  1.2450us         3     415ns     146ns     877ns  cuDeviceGetCount
                    0.00%     262ns         1     262ns     262ns     262ns  cuDeviceGetUuid
```

100M

```
==3775686== NVPROF is profiling process 3775686, command: ./q3_2
==3775686== Profiling application: ./q3_2
==3775686== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   48.17%  284.09ms         1  284.09ms  284.09ms  284.09ms  [CUDA memcpy DtoH]
                   28.38%  167.41ms         2  83.705ms  83.671ms  83.740ms  [CUDA memcpy HtoD]
                   23.45%  138.31ms         1  138.31ms  138.31ms  138.31ms  vector_add(float*, float*, float*, int)
      API calls:   73.48%  591.31ms         3  197.10ms  83.870ms  423.48ms  cudaMemcpy
                   26.23%  211.08ms         3  70.362ms  564.02us  209.95ms  cudaMalloc
                    0.19%  1.5394ms         3  513.14us  473.73us  591.00us  cudaFree
                    0.04%  354.61us       101  3.5100us     120ns  157.95us  cuDeviceGetAttribute
                    0.04%  331.51us         1  331.51us  331.51us  331.51us  cuDeviceTotalMem
                    0.01%  46.430us         1  46.430us  46.430us  46.430us  cuDeviceGetName
                    0.00%  35.484us         1  35.484us  35.484us  35.484us  cudaLaunchKernel
                    0.00%  6.0020us         1  6.0020us  6.0020us  6.0020us  cuDeviceGetPCIBusId
                    0.00%  4.9720us         2  2.4860us     167ns  4.8050us  cuDeviceGet
                    0.00%     942ns         3     314ns     153ns     634ns  cuDeviceGetCount
                    0.00%     236ns         1     236ns     236ns     236ns  cuDeviceGetUuid
```

cudaMalloc – 256 – n blocks

1M – 3906 blocks

```
==3775185== NVPROF is profiling process 3775185, command: ./q3_2
==3775185== Profiling application: ./q3_2
==3775185== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   81.32%  13.520ms         1  13.520ms  13.520ms  13.520ms  vector_add(float*, float*, float*, int)
                   10.17%  1.6907ms         1  1.6907ms  1.6907ms  1.6907ms  [CUDA memcpy DtoH]
                    8.51%  1.4156ms         2  707.82us  703.93us  711.71us  [CUDA memcpy HtoD]
      API calls:   92.28%  235.21ms         3  78.403ms  142.32us  234.92ms  cudaMalloc
                    7.22%  18.401ms         3  6.1335ms  891.14us  16.582ms  cudaMemcpy
                    0.20%  516.29us         3  172.10us  125.93us  252.17us  cudaFree
                    0.14%  353.09us       101  3.4950us     122ns  156.88us  cuDeviceGetAttribute
                    0.13%  330.52us         1  330.52us  330.52us  330.52us  cuDeviceTotalMem
                    0.02%  45.281us         1  45.281us  45.281us  45.281us  cuDeviceGetName
                    0.01%  33.193us         1  33.193us  33.193us  33.193us  cudaLaunchKernel
                    0.00%  5.4600us         1  5.4600us  5.4600us  5.4600us  cuDeviceGetPCIBusId
                    0.00%  1.3970us         3     465ns     173ns  1.0150us  cuDeviceGetCount
                    0.00%     709ns         2     354ns     133ns     576ns  cuDeviceGet
                    0.00%     249ns         1     249ns     249ns     249ns  cuDeviceGetUuid
(base) [zz3904@gv016 part1]$
```

5M – 19531 blocks

```
==3775054== NVPROF is profiling process 3775054, command: ./q3_2
==3775054== Profiling application: ./q3_2
==3775054== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   93.43%  312.63ms         1  312.63ms  312.63ms  312.63ms  vector_add(float*, float*, float*, int)
                    4.12%  13.785ms         1  13.785ms  13.785ms  13.785ms  [CUDA memcpy DtoH]
                    2.45%  8.1963ms         2  4.0982ms  4.0857ms  4.1106ms  [CUDA memcpy HtoD]
      API calls:   60.48%  335.86ms         3  111.95ms  4.2855ms  327.27ms  cudaMemcpy
                   39.27%  218.08ms         3  72.692ms  158.31us  217.75ms  cudaMalloc
                    0.10%  569.36us         3  189.79us  141.02us  282.00us  cudaFree
                    0.07%  376.44us       101  3.7270us     133ns  166.87us  cuDeviceGetAttribute
                    0.07%  364.06us         1  364.06us  364.06us  364.06us  cuDeviceTotalMem
                    0.01%  51.118us         1  51.118us  51.118us  51.118us  cuDeviceGetName
                    0.01%  37.153us         1  37.153us  37.153us  37.153us  cudaLaunchKernel
                    0.00%  4.9790us         1  4.9790us  4.9790us  4.9790us  cuDeviceGetPCIBusId
                    0.00%  1.0170us         2     508ns     166ns     851ns  cuDeviceGet
                    0.00%  1.0130us         3     337ns     182ns     623ns  cuDeviceGetCount
                    0.00%     257ns         1     257ns     257ns     257ns  cuDeviceGetUuid
```

10M – 39062 blocks

```
==3774958== NVPROF is profiling process 3774958, command: ./q3_2
==3774958== Profiling application: ./q3_2
==3774958== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   96.73%  1.32499s         1  1.32499s  1.32499s  1.32499s  vector_add(float*, float*, float*, int)
                    2.05%  28.138ms         1  28.138ms  28.138ms  28.138ms  [CUDA memcpy DtoH]
                    1.22%  16.679ms         2  8.3395ms  8.3390ms  8.3401ms  [CUDA memcpy HtoD]
      API calls:   86.12%  1.37111s         3  457.04ms  8.5342ms  1.35402s  cudaMemcpy
                   13.78%  219.39ms         3  73.129ms  174.05us  219.03ms  cudaMalloc
                    0.05%  777.49us         3  259.16us  164.76us  438.10us  cudaFree
                    0.02%  354.20us       101  3.5060us     122ns  156.01us  cuDeviceGetAttribute
                    0.02%  348.74us         1  348.74us  348.74us  348.74us  cuDeviceTotalMem
                    0.00%  48.089us         1  48.089us  48.089us  48.089us  cuDeviceGetName
                    0.00%  30.563us         1  30.563us  30.563us  30.563us  cudaLaunchKernel
                    0.00%  5.0920us         1  5.0920us  5.0920us  5.0920us  cuDeviceGetPCIBusId
                    0.00%     792ns         3     264ns     138ns     493ns  cuDeviceGetCount
                    0.00%     730ns         2     365ns     164ns     566ns  cuDeviceGet
                    0.00%     259ns         1     259ns     259ns     259ns  cuDeviceGetUuid
```

50M – 195312 blocks

```
==3774768== NVPROF is profiling process 3774768, command: ./q3_2
==3774768== Profiling application: ./q3_2
==3774768== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   99.27%  30.7723s         1  30.7723s  30.7723s  30.7723s  vector_add(float*, float*, float*, int)
                    0.46%  141.36ms         1  141.36ms  141.36ms  141.36ms  [CUDA memcpy DtoH]
                    0.27%  83.920ms         2  41.960ms  41.821ms  42.099ms  [CUDA memcpy HtoD]
      API calls:   99.31%  30.9993s         3  10.3331s  42.027ms  30.9150s  cudaMemcpy
                    0.69%  214.75ms         3  71.583ms  300.39us  214.14ms  cudaMalloc
                    0.00%  1.2063ms         3  402.11us  304.52us  594.42us  cudaFree
                    0.00%  547.72us       101  5.4220us     122ns  338.15us  cuDeviceGetAttribute
                    0.00%  334.83us         1  334.83us  334.83us  334.83us  cuDeviceTotalMem
                    0.00%  54.822us         1  54.822us  54.822us  54.822us  cuDeviceGetName
                    0.00%  35.308us         1  35.308us  35.308us  35.308us  cudaLaunchKernel
                    0.00%  5.1240us         1  5.1240us  5.1240us  5.1240us  cuDeviceGetPCIBusId
                    0.00%  1.0020us         3     334ns     145ns     682ns  cuDeviceGetCount
                    0.00%     714ns         2     357ns     146ns     568ns  cuDeviceGet
                    0.00%     236ns         1     236ns     236ns     236ns  cuDeviceGetUuid
```

100M – 390625 blocks

```
==3083491== NVPROF is profiling process 3083491, command: ./q3_2
==3083491== Profiling application: ./q3_2
==3083491== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:   99.73%  143.927s         1  143.927s  143.927s  143.927s  vector_add(float*, float*, float*, int)
                    0.20%  287.22ms         1  287.22ms  287.22ms  287.22ms  [CUDA memcpy DtoH]
                    0.07%  96.113ms         2  48.056ms  47.570ms  48.543ms  [CUDA memcpy HtoD]
      API calls:   99.57%  144.312s         3  48.1039s  47.657ms  144.215s  cudaMemcpy
                    0.39%  566.75ms         3  188.92ms  635.32us  565.47ms  cudaMalloc
                    0.04%  52.849ms         3  17.616ms  717.64us  49.892ms  cudaFree
                    0.00%  1.5335ms       101  15.183us     133ns  669.76us  cuDeviceGetAttribute
                    0.00%  1.4410ms         1  1.4410ms  1.4410ms  1.4410ms  cuDeviceTotalMem
                    0.00%  206.54us         1  206.54us  206.54us  206.54us  cuDeviceGetName
                    0.00%  56.635us         1  56.635us  56.635us  56.635us  cudaLaunchKernel
                    0.00%  6.6270us         1  6.6270us  6.6270us  6.6270us  cuDeviceGetPCIBusId
                    0.00%  5.0550us         2  2.5270us     165ns  4.8900us  cuDeviceGet
                    0.00%  1.3190us         3     439ns     197ns     922ns  cuDeviceGetCount
                    0.00%     295ns         1     295ns     295ns     295ns  cuDeviceGetUuid
```

cudaMallocManaged

1M

```
(base) [223904@gv016 part1]$ ./q3_3.sh
==3777109== NVPROF is profiling process 3777109, command: ./q3_3
==3777109== Profiling application: ./q3_3
==3777109== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  109.03ms         1  109.03ms  109.03ms  109.03ms  add(int, float*, float*)
      API calls:   69.61%  255.84ms         2  127.92ms  20.036us  255.82ms  cudaMallocManaged
                   29.67%  109.03ms         1  109.03ms  109.03ms  109.03ms  cudaDeviceSynchronize
                    0.28%  1.0362ms       101  10.259us     117ns  488.18us  cuDeviceGetAttribute
                    0.24%  889.90us         2  444.95us  422.89us  467.01us  cudaFree
                    0.12%  454.01us         1  454.01us  454.01us  454.01us  cuDeviceTotalMem
                    0.07%  239.50us         1  239.50us  239.50us  239.50us  cuDeviceGetName
                    0.01%  36.979us         1  36.979us  36.979us  36.979us  cudaLaunchKernel
                    0.00%  5.0920us         1  5.0920us  5.0920us  5.0920us  cuDeviceGetPCIBusId
                    0.00%     955ns         3     318ns     152ns     610ns  cuDeviceGetCount
                    0.00%     740ns         2     370ns     137ns     603ns  cuDeviceGet
                    0.00%     252ns         1     252ns     252ns     252ns  cuDeviceGetUuid

==3777109== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
      48  162.83KB  4.0000KB  0.9961MB  7.632813MB  805.4400us  Host To Device
      12         -         -         -           -  3.531120ms  Gpu page fault groups
Total CPU Page faults: 24
```

5M

```
==3777201== NVPROF is profiling process 3777201, command: ./q3_3
==3777201== Profiling application: ./q3_3
==3777201== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  504.54ms         1  504.54ms  504.54ms  504.54ms  add(int, float*, float*)
      API calls:   67.10%  504.55ms         1  504.55ms  504.55ms  504.55ms  cudaDeviceSynchronize
                   32.38%  243.47ms         2  121.74ms  30.989us  243.44ms  cudaMallocManaged
                    0.42%  3.1678ms         2  1.5839ms  1.5813ms  1.5865ms  cudaFree
                    0.05%  358.86us       101  3.5530us     120ns  157.50us  cuDeviceGetAttribute
                    0.04%  325.32us         1  325.32us  325.32us  325.32us  cuDeviceTotalMem
                    0.01%  50.837us         1  50.837us  50.837us  50.837us  cudaLaunchKernel
                    0.01%  45.116us         1  45.116us  45.116us  45.116us  cuDeviceGetName
                    0.00%  5.6840us         1  5.6840us  5.6840us  5.6840us  cuDeviceGetPCIBusId
                    0.00%     871ns         2     435ns     140ns     731ns  cuDeviceGet
                    0.00%     866ns         3     288ns     157ns     526ns  cuDeviceGetCount
                    0.00%     249ns         1     249ns     249ns     249ns  cuDeviceGetUuid

==3777201== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
     236  165.53KB  4.0000KB  0.9961MB  38.14844MB  4.001807ms  Host To Device
      59         -         -         -           -  15.99709ms  Gpu page fault groups
Total CPU Page faults: 118
```

10M

```
==3777278== NVPROF is profiling process 3777278, command: ./q3_3
==3777278== Profiling application: ./q3_3
==3777278== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  950.68ms         1  950.68ms  950.68ms  950.68ms  add(int, float*, float*)
      API calls:   79.50%  950.72ms         1  950.72ms  950.72ms  950.72ms  cudaDeviceSynchronize
                   19.80%  236.80ms         2  118.40ms  23.212us  236.78ms  cudaMallocManaged
                    0.59%  7.0000ms         2  3.5000ms  3.4060ms  3.5939ms  cudaFree
                    0.06%  726.11us       101  7.1890us     123ns  329.33us  cuDeviceGetAttribute
                    0.04%  440.09us         1  440.09us  440.09us  440.09us  cuDeviceTotalMem
                    0.01%  96.767us         1  96.767us  96.767us  96.767us  cuDeviceGetName
                    0.00%  44.176us         1  44.176us  44.176us  44.176us  cudaLaunchKernel
                    0.00%  4.9720us         1  4.9720us  4.9720us  4.9720us  cuDeviceGetPCIBusId
                    0.00%     779ns         3     259ns     145ns     487ns  cuDeviceGetCount
                    0.00%     693ns         2     346ns     142ns     551ns  cuDeviceGet
                    0.00%     256ns         1     256ns     256ns     256ns  cuDeviceGetUuid

==3777278== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
     464  168.38KB  4.0000KB  0.9961MB  76.29688MB  7.952891ms  Host To Device
     116         -         -         -           -  31.55559ms  Gpu page fault groups
Total CPU Page faults: 232
```

50M

```
(base) [zz3904@gv016 part1]$ ./q3_3.sh
==3777333== NVPROF is profiling process 3777333, command: ./q3_3
==3777333== Profiling application: ./q3_3
==3777333== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  4.75947s         1  4.75947s  4.75947s  4.75947s  add(int, float*, float*)
      API calls:   94.14%  4.75950s         1  4.75950s  4.75950s  4.75950s  cudaDeviceSynchronize
                    4.62%  233.76ms         2  116.88ms  33.615us  233.73ms  cudaMallocManaged
                    1.22%  61.713ms         2  30.856ms  15.255ms  46.458ms  cudaFree
                    0.01%  352.08us       101  3.4850us     121ns  156.73us  cuDeviceGetAttribute
                    0.01%  329.84us         1  329.84us  329.84us  329.84us  cuDeviceTotalMem
                    0.00%  53.485us         1  53.485us  53.485us  53.485us  cudaLaunchKernel
                    0.00%  44.696us         1  44.696us  44.696us  44.696us  cuDeviceGetName
                    0.00%  5.2610us         1  5.2610us  5.2610us  5.2610us  cuDeviceGetPCIBusId
                    0.00%  1.1900us         3     396ns     205ns     688ns  cuDeviceGetCount
                    0.00%     757ns         2     378ns     175ns     582ns  cuDeviceGet
                    0.00%     232ns         1     232ns     232ns     232ns  cuDeviceGetUuid

==3777333== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
    2300  169.84KB  4.0000KB  0.9961MB  381.4766MB  39.81117ms  Host To Device
     575         -         -         -           -  138.9393ms  Gpu page fault groups
Total CPU Page faults: 1150
```

100M

```
(base) [zz3904@gv016 part1]$ ./q3_3.sh
==3777442== NVPROF is profiling process 3777442, command: ./q3_3
==3777442== Profiling application: ./q3_3
==3777442== Profiling result:
            Type  Time(%)      Time     Calls       Avg       Min       Max  Name
 GPU activities:  100.00%  9.51020s         1  9.51020s  9.51020s  9.51020s  add(int, float*, float*)
      API calls:   97.03%  9.51023s         1  9.51023s  9.51023s  9.51023s  cudaDeviceSynchronize
                    2.39%  234.41ms         2  117.20ms  38.927us  234.37ms  cudaMallocManaged
                    0.57%  56.037ms         2  28.019ms  27.566ms  28.471ms  cudaFree
                    0.01%  563.59us       101  5.5800us     117ns  370.27us  cuDeviceGetAttribute
                    0.00%  321.60us         1  321.60us  321.60us  321.60us  cuDeviceTotalMem
                    0.00%  45.264us         1  45.264us  45.264us  45.264us  cuDeviceGetName
                    0.00%  42.030us         1  42.030us  42.030us  42.030us  cudaLaunchKernel
                    0.00%  6.2030us         1  6.2030us  6.2030us  6.2030us  cuDeviceGetPCIBusId
                    0.00%  1.0210us         3     340ns     153ns     678ns  cuDeviceGetCount
                    0.00%     788ns         2     394ns     162ns     626ns  cuDeviceGet
                    0.00%     289ns         1     289ns     289ns     289ns  cuDeviceGetUuid

==3777442== Unified Memory profiling result:
Device "Tesla V100-SXM2-16GB (0)"
   Count  Avg Size  Min Size  Max Size  Total Size  Total Time  Name
    4584  170.43KB  4.0000KB  0.9961MB  762.9453MB  79.59911ms  Host To Device
    1146         -         -         -           -  298.5592ms  Gpu page fault groups
Total CPU Page faults: 2292
(base) [zz3904@gv016 part1]$
```