



# UPPSALA UNIVERSITET

Accelerator-Based Programming - 1TD055

Final Project

Jyong-Jhih Lin

October 18, 2022

## 0 Hardware information

snowy CPU:

```
1 Architecture:          x86_64
2 CPU op-mode(s):        32-bit, 64-bit
3 Byte Order:             Little Endian
4 CPU(s):                 16
5 On-line CPU(s) list:    0-15
6 Thread(s) per core:     1
7 Core(s) per socket:     8
8 Socket(s):              2
9 NUMA node(s):           2
10 Vendor ID:              GenuineIntel
11 CPU family:             6
12 Model:                  45
13 Model name:              Intel(R) Xeon(R) CPU E5-2660 0 @ 2.20GHz
14 Stepping:                7
15 CPU MHz:                 1200.000
16 CPU max MHz:             2200.0000
17 CPU min MHz:             1200.0000
18 BogomIPS:                4388.80
19 Virtualization:          VT-x
20 L1d cache:               32K
21 L1i cache:               32K
22 L2 cache:                256K
23 L3 cache:                20480K
24 NUMA node0 CPU(s):       0-7
25 NUMA node1 CPU(s):       8-15
26 Flags:                   fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat pse36 clflush dts
                           acpi mmx fxsr sse sse2 ss ht tm pbe syscall nx pdpe1gb rdtscp lm constant_tsc arch_perfmon pebs bts
                           rep_good nopl xtopology nonstop_tsc aperfmperf eagerfpu pni pclmulqdq dtes64 monitor ds_cpl vmx smx est
                           tm2 ssse3 cx16 xtpr pdcm pcid dca sse4_1 sse4_2 x2apic popcnt tsc_deadline_timer aes xsave avx lahf_lm
                           epb ssbd ibrs ibpb stibp tpr_shadow vnmi flexpriority ept vpid xsaveopt dtherm ida arat pln pts
                           md_clear spec_ctrl intel_stibp flush_l1d
```

snowy memory:

```
1 Handle 0x1100, DMI type 17, 40 bytes
2 Memory Device
3 Array Handle: 0x1000
4 Error Information Handle: Not Provided
5 Total Width: 72 bits
6 Data Width: 64 bits
7 Size: 32 GB
8 Form Factor: DIMM
9 Set: None
10 Locator: PROC 1 DIMM 1
11 Bank Locator: Not Specified
12 Type: DDR3
13 Type Detail: Synchronous LRDIMM
14 Speed: 1333 MT/s
15 Manufacturer: HP
16 Serial Number: Not Specified
17 Asset Tag: Not Specified
18 Part Number: 647654-081
19 Rank: 4
20 Configured Memory Speed: 1333 MT/s
21 Minimum Voltage: 1.35 V
22 Maximum Voltage: 1.5 V
23 Configured Voltage: 1.35 V
24
25 #####
26 DDR3-1333 4-channel memory total bandwidth = 1333e6(T/s) * 64(bits) * 4(channels) / 8e9(GBytes/s)
27 = 42.656(GBytes/s)
```

nvidia T4, nvidia-smi:

```
1 +-----+
2 | NVIDIA-SMI 515.65.01      Driver Version: 515.65.01      CUDA Version: 11.7      |
3 |-----+-----+-----+
4 | GPU   Name               Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
5 | Fan  Temp  Perf    Pwr:Usage/Cap|      Memory-Usage | GPU-Util  Compute M. |
6 |                                           | MIG M.         |
7 |=====+=====+=====+
8 |    0   Tesla T4               On      | 00000000:08:00.0 Off |                    0 |
9 | N/A   30C    P8             14W /  70W |    2MiB / 15360MiB |          0%      Default |
10 |                                           |                      | N/A              |
11 |-----+-----+-----+
12
13 +-----+
14 | Processes: |
15 | GPU   GI    CI          PID    Type    Process name                        GPU Memory |
16 |      ID    ID                                 Usage |
17 |-----+-----+-----+
18 |
```

```

18 | No running processes found |
19 +-----+

```

#### nvidia T4, *deviceQuery*:

```

1 /sw/EasyBuild/snowy/software/CUDA/10.1.243-iccifort-2019.5.281/extras/demo_suite/deviceQuery Starting...
2
3 CUDA Device Query (Runtime API) version (CUDA static linking)
4
5 Detected 1 CUDA Capable device(s)
6
7 Device 0: "Tesla T4"
8   CUDA Driver Version / Runtime Version      11.7 / 10.1
9   CUDA Capability Major/Minor version number: 7.5
10  Total amount of global memory:              14972 MBytes (15699148800 bytes)
11  (40) Multiprocessors, ( 64) CUDA Cores/MP:  2560 CUDA Cores
12  GPU Max Clock rate:                         1590 MHz (1.59 GHz)
13  Memory Clock rate:                          5001 Mhz
14  Memory Bus Width:                           256-bit
15  L2 Cache Size:                              4194304 bytes
16  Maximum Texture Dimension Size (x,y,z)      1D=(131072), 2D=(131072, 65536), 3D=(16384, 16384, 16384)
17  Maximum Layered 1D Texture Size, (num) layers 1D=(32768), 2048 layers
18  Maximum Layered 2D Texture Size, (num) layers 2D=(32768, 32768), 2048 layers
19  Total amount of constant memory:             65536 bytes
20  Total amount of shared memory per block:     49152 bytes
21  Total number of registers available per block: 65536
22  Warp size:                                   32
23  Maximum number of threads per multiprocessor: 1024
24  Maximum number of threads per block:         1024
25  Max dimension size of a thread block (x,y,z): (1024, 1024, 64)
26  Max dimension size of a grid size (x,y,z):   (2147483647, 65535, 65535)
27  Maximum memory pitch:                       2147483647 bytes
28  Texture alignment:                          512 bytes
29  Concurrent copy and kernel execution:        Yes with 3 copy engine(s)
30  Run time limit on kernels:                   No
31  Integrated GPU sharing Host Memory:          No
32  Support host page-locked memory mapping:     Yes
33  Alignment requirement for Surfaces:         Yes
34  Device has ECC support:                      Enabled
35  Device supports Unified Addressing (UVA):    Yes
36  Device supports Compute Preemption:         Yes
37  Supports Cooperative Kernel Launch:         Yes
38  Supports MultiDevice Co-op Kernel Launch:   Yes
39  Device PCI Domain ID / Bus ID / location ID: 0 / 8 / 0
40  Compute Mode:
41     < Default (multiple host threads can use ::cudaSetDevice() with device simultaneously) >
42
43 deviceQuery, CUDA Driver = CUDART, CUDA Driver Version = 11.7, CUDA Runtime Version = 10.1, NumDevs = 1,
44   Device0 = Tesla T4
45 Result = PASS

```

#### DELL Precision 7760 CPU:

```

1 Architecture:                x86_64
2 CPU op-mode(s):              32-bit, 64-bit
3 Byte Order:                  Little Endian
4 Address sizes:               39 bits physical, 48 bits virtual
5 CPU(s):                      12
6 On-line CPU(s) list:         0-11
7 Thread(s) per core:          2
8 Core(s) per socket:          6
9 Socket(s):                    1
10 NUMA node(s):                1
11 Vendor ID:                   GenuineIntel
12 CPU family:                   6
13 Model:                       141
14 Model name:                   Intel(R) Xeon(R) W-11855M CPU @ 3.20GHz
15 Stepping:                     1
16 CPU MHz:                      3200.000
17 CPU max MHz:                  4900.0000
18 CPU min MHz:                  800.0000
19 BogomIPS:                     6374.40
20 Virtualization:              VT-x
21 L1d cache:                    288 KiB
22 L1i cache:                    192 KiB
23 L2 cache:                     7.5 MiB
24 L3 cache:                     18 MiB
25 NUMA node0 CPU(s):           0-11
26 Vulnerability Itlb multihit:  Not affected
27 Vulnerability L1tf:           Not affected
28 Vulnerability Mds:            Not affected
29 Vulnerability Meltdown:       Not affected
30 Vulnerability Mmio stale data: Not affected
31 Vulnerability Spec store bypass: Mitigation; Speculative Store Bypass disabled via prctl and seccomp
32 Vulnerability Spectre v1:      Mitigation; usercopy/swapgs barriers and __user pointer sanitization
33 Vulnerability Spectre v2:      Mitigation; Enhanced IBRS, IBPB conditional, RSB filling
34 Vulnerability Srbds:           Not affected

```

```

35 Vulnerability Tsx async abort:    Not affected
36 Flags:                             fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat pse36
    clflush dts acpi mmx fxsr sse sse2 ss ht tm pbe syscall nx pdpe1gb rdtscp lm constant_tsc art
    arch_perfmon pebs bts rep_good nopl xtopology nonstop_tsc cpuid aperfperf tsc_known_freq pni pclmulqdq
    dtes64 monitor ds_cpl vmx smx est tm2 ssse3 sdbg fma cx16 xtpr pdcm pcid sse4_1 sse4_2 x2apic movbe
    popcnt tsc_deadline_timer aes xsave avx f16c rdrand lahf_lm abm 3dnowprefetch cpuid_fault epb cat_l2
    invpcid_single cdp_l2 ssbd ibrs ibpb stibp ibrs_enhanced tpr_shadow vnmi flexpriority ept vpid ept_ad
    fsgsbase tsc_adjust bmi1 avx2 smep bmi2 erms invpcid rdt_a avx512f avx512dq rdseed adx smap avx512ifma
    clflushopt clwb intel_pt avx512cd sha_ni avx512bw avx512vl xsaveopt xsavec xgetbv1 xsaves
    split_lock_detect dtherm ida arat pln pts hwp hwp_notify hwp_act_window hwp_epp hwp_pkg_req avx512vbmi
    umip pku ospke avx512_vbmi2 gfni vaes vpclmulqdq avx512_vnni avx512_bitalg tme avx512_vpopcntdq rdpid
    movdiri movdir64b fsrm avx512_vp2intersect md_clear flush_l1d arch_capabilities

```

#### DELL Precision 7760 memory:

```

1 Handle 0x1100, DMI type 17, 92 bytes
2 Memory Device
3   Array Handle: 0x1000
4   Error Information Handle: Not Provided
5   Total Width: 72 bits
6   Data Width: 64 bits
7   Size: 32 GB
8   Form Factor: SODIMM
9   Set: None
10  Locator: DIMM C
11  Bank Locator: BANK 0
12  Type: DDR4
13  Type Detail: Synchronous
14  Speed: 2933 MT/s
15  Manufacturer: 01980000802C
16  Serial Number: 97B0B609
17  Asset Tag: 04212100
18  Part Number: 9965657-029.A00G
19  Rank: 2
20  Configured Memory Speed: 2933 MT/s
21  Minimum Voltage: Unknown
22  Maximum Voltage: Unknown
23  Configured Voltage: 1.2 V
24  Memory Technology: DRAM
25  Memory Operating Mode Capability: Volatile memory
26  Firmware Version: Not Specified
27  Module Manufacturer ID: Bank 2, Hex 0x98
28  Module Product ID: Unknown
29  Memory Subsystem Controller Manufacturer ID: Unknown
30  Memory Subsystem Controller Product ID: Unknown
31  Non-Volatile Size: None
32  Volatile Size: 32 GB
33  Cache Size: None
34  Logical Size: None
35
36 #####
37 DDR4-2933 2-channel memory total bandwidth = 2933e6(T/s) * 64(bits) * 2(channels) / 8e9(GBytes/s)
38 = 46.928(GBytes/s)

```

#### nvidia A3000, nvidia-smi:

```

1 +-----+
2 | NVIDIA-SMI 510.47.03      Driver Version: 510.47.03      CUDA Version: 11.6      |
3 |-----+-----+-----+-----+-----+-----+
4 | GPU   Name               Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
5 | Fan   Temp   Perf         Pwr:Usage/Cap|               Memory-Usage | GPU-Util  Compute M. |
6 |-----+-----+-----+-----+-----+-----+-----+
7 | 0     0    NVIDIA RTX A300... Off      | 00000000:01:00:0 On    |                     N/A |
8 | N/A    58C     P0      36W /  N/A      | 1606MiB /  6144MiB |    100%      Default  |
9 |-----+-----+-----+-----+-----+-----+
10 |
11 +-----+
12
13 +-----+
14 | Processes:
15 | GPU   GI    CI          PID    Type   Process name                      GPU Memory
16 |   ID   ID                                 Usage                     |
17 +-----+-----+-----+-----+-----+-----+
18 | 0     N/A  N/A       3964      G   /usr/lib/xorg/Xorg                 109MiB |
19 | 0     N/A  N/A      12824      G   /usr/lib/xorg/Xorg                 603MiB |
20 | 0     N/A  N/A      12940      G   /usr/bin/gnome-shell               271MiB |
21 | 0     N/A  N/A      30513      G   ...308337019390783085,131072      481MiB |
22 | 0     N/A  N/A      215156     G   ...R2021a/bin/glnxa64/MATLAB        3MiB |
23 | 0     N/A  N/A      216972     G   ...GL_KHR_blend_equation_adv        5MiB |
24 | 0     N/A  N/A      284196     C   ./stream_triad_cuda                113MiB |
25 +-----+

```

#### nvidia A3000, deviceQuery:

```

1 ./deviceQuery Starting...

```

```

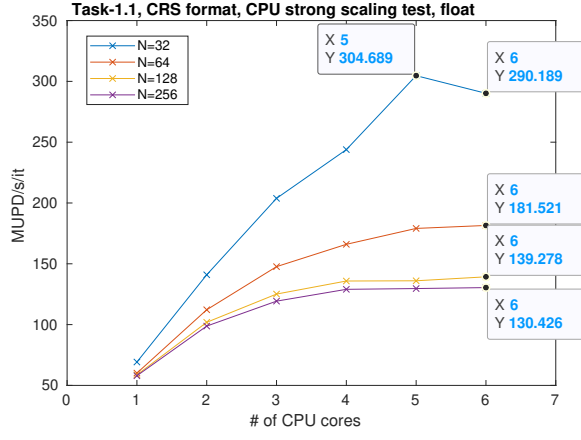
2
3   CUDA Device Query (Runtime API) version (CUDART static linking)
4
5   Detected 1 CUDA Capable device(s)
6
7   Device 0: "NVIDIA RTX A3000 Laptop GPU"
8     CUDA Driver Version / Runtime Version      11.6 / 11.6
9     CUDA Capability Major/Minor version number: 8.6
10    Total amount of global memory:              5913 MBytes (6200098816 bytes)
11    (32) Multiprocessors, (128) CUDA Cores/MP:  4096 CUDA Cores
12    GPU Max Clock rate:                         1560 MHz (1.56 GHz)
13    Memory Clock rate:                         5501 Mhz
14    Memory Bus Width:                          192-bit
15    L2 Cache Size:                             3145728 bytes
16    Maximum Texture Dimension Size (x,y,z)      1D=(131072), 2D=(131072, 65536), 3D=(16384, 16384, 16384)
17    Maximum Layered 1D Texture Size, (num) layers 1D=(32768), 2048 layers
18    Maximum Layered 2D Texture Size, (num) layers 2D=(32768, 32768), 2048 layers
19    Total amount of constant memory:             65536 bytes
20    Total amount of shared memory per block:     49152 bytes
21    Total number of registers available per block: 65536
22    Warp size:                                  32
23    Maximum number of threads per multiprocessor: 1536
24    Maximum number of threads per block:         1024
25    Max dimension size of a thread block (x,y,z): (1024, 1024, 64)
26    Max dimension size of a grid size    (x,y,z): (2147483647, 65535, 65535)
27    Maximum memory pitch:                       2147483647 bytes
28    Texture alignment:                          512 bytes
29    Concurrent copy and kernel execution:        Yes with 2 copy engine(s)
30    Run time limit on kernels:                   Yes
31    Integrated GPU sharing Host Memory:          No
32    Support host page-locked memory mapping:     Yes
33    Alignment requirement for Surfaces:          Yes
34    Device has ECC support:                      Disabled
35    Device supports Unified Addressing (UVA):    Yes
36    Device supports Compute Preemption:         Yes
37    Supports Cooperative Kernel Launch:         Yes
38    Supports MultiDevice Co-op Kernel Launch:   Yes
39    Device PCI Domain ID / Bus ID / location ID: 0 / 1 / 0
40    Compute Mode:
41      < Default (multiple host threads can use ::cudaSetDevice() with device simultaneously) >
42
43   deviceQuery, CUDA Driver = CUDART, CUDA Driver Version = 11.6, CUDA Runtime Version = 11.6, NumDevs = 1,
44   Device0 = NVIDIA RTX A3000 Laptop GPU
45   Result = PASS

```

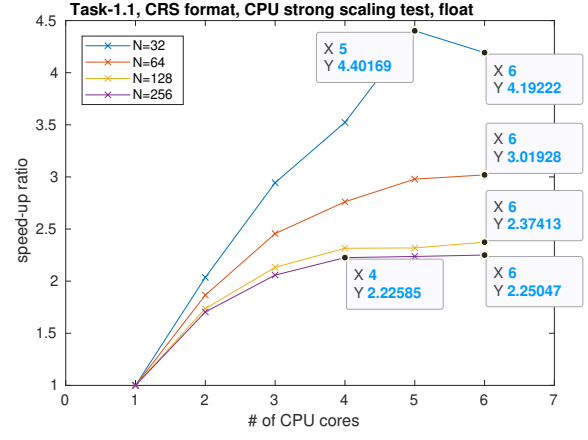
- I did this project and this report alone.
- According to the grading criteria, this project might be in grade 3. I did not finish the CELL-C-Sigma format because I did not have enough time to work on it, not because I have no motivation to achieve a higher grade. I am very willing to discuss more questions in the oral exam session to show I have learned more than a grade 3.

# 1 Task 1

## 1.1 Strong Scaling Test

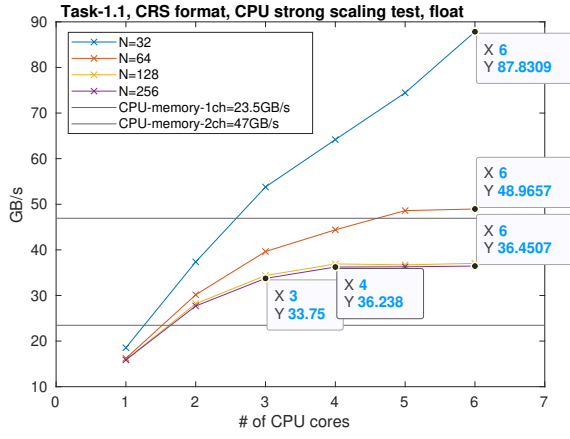


(a) MUPD

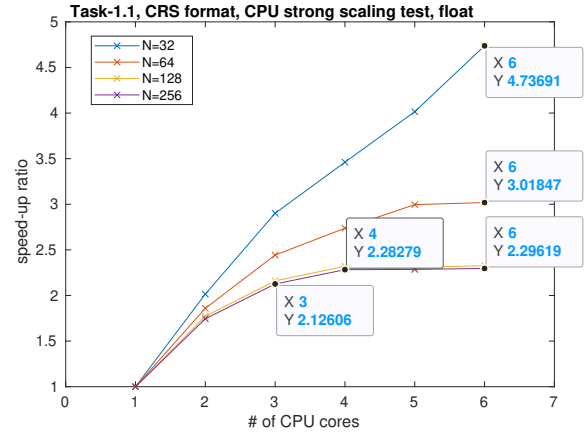


(b) speed-up

Figure 1: CPU, float, MUPD

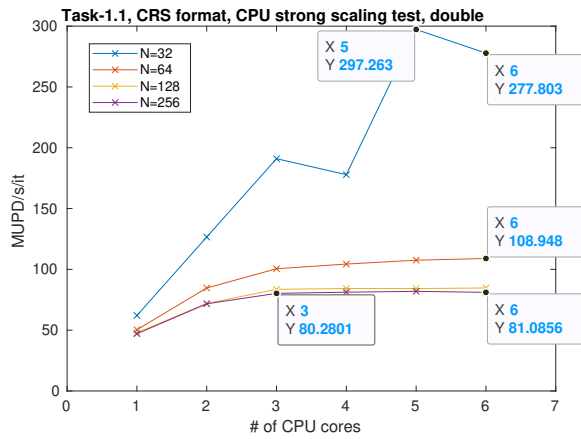


(a) GB/s

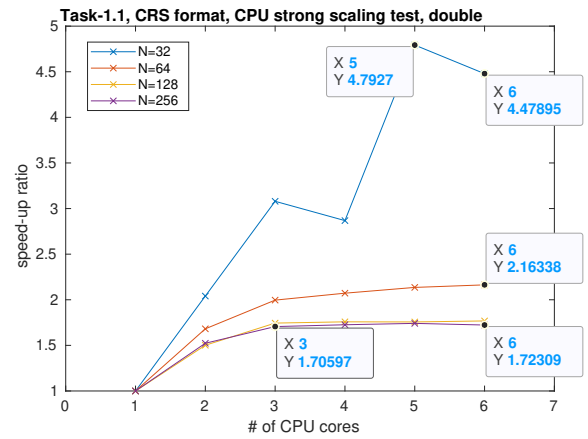


(b) speed-up

Figure 2: CPU, float, GB/s



(a) MUPD



(b) speed-up

Figure 3: CPU, double, MUPD

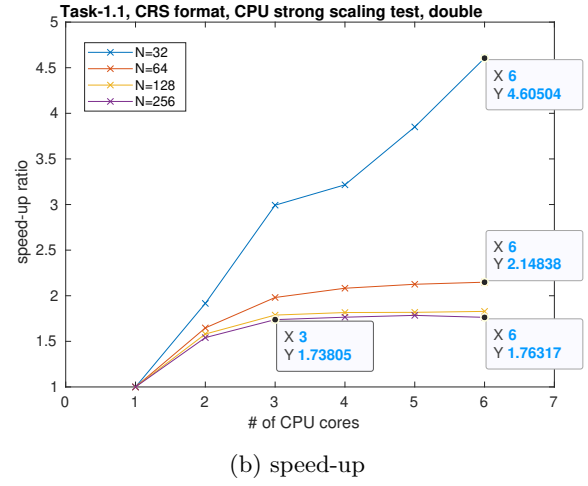
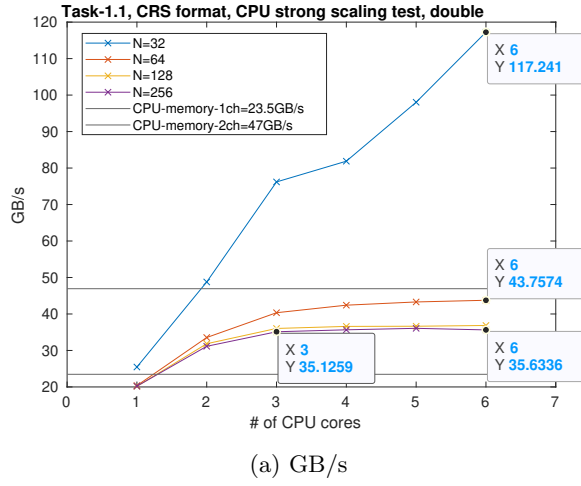


Figure 4: CPU, double, GB/s

N=32 is in the cache region, that is why its behavior is different from the others. In this region, the cache memory bandwidth is not an issue and the strong scaling is good.

It is in the DRAM region when N=64, 128, 256. The strong scaling test shows that it would be inefficient to use more than 4 CPU cores for float type and 3 CPU cores for double type. The reason is the DRAM memory bandwidth bound which is indicated in the figure 2 and 4.

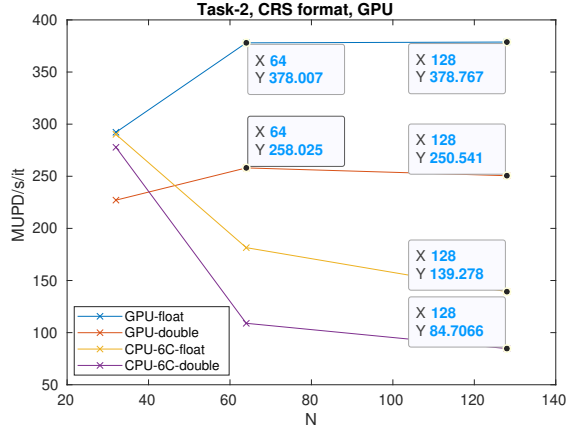
For the double type computation, it requires the bandwidth as twice as the float type. That is why it would reach the maximum speed-up earlier than the float type.

This experiment indicate an interesting thing, which is it needs to use more than 1 CPU core to reach the DDR-2ch maximum bandwidth. In the previous labs, even with an optimized AVX-512 SIMD parallelization, 1 CPU core could only reach the DDR-1ch bandwidth.

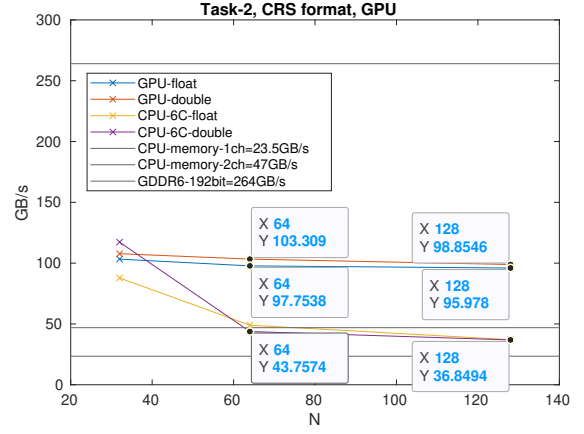
## 1.2 Weak Scaling Test

Intel Xeon W-11855 M CPU only has 6 physical cores which is not enough to perform the weak scaling test, so it is skipped.

## 2 Task 2



(a) MUPD



(b) GB/s

Figure 5: GPU&CPU, float&double, MUPD&GB/s

Nvidia A3000 GPU does not have enough memory to perform N=256 test. It is skipped.

The GPU does not reach the maximum compute performance and the maximum memory bandwidth. Also, the float and double type have basically the same bandwidth. These two results indicate that it must be another reason other than the compute bound and bandwidth bound. I think it is caused by the memory access latency and overhead during the discontinuous and non-coalesced data access. This kind of discontinuous and non-coalesced data access would tremendously reduce the number of the simultaneously working CUDA cores which lower the performance and bandwidth.

With the same memory bandwidth performance, it is reasonable that the double type performance is lower than the float type. Not mention that the number of FP64 ALU is 1/64 as the FP32 ALU for the A3000 model.