과제 #4

홍경인

M1522.006700 확장형 고성능 컴퓨팅 (001)

November 18, 2024

1 GPU 정보 확인하기

(a) 주어진 커맨드를 실행하여 얻은 결과는 아래와 같다.

```
>>> srun --partition=class1 --gres=gpu:4 nvidia-smi
srun: job 996293 queued and waiting for resources
srun: job 996293 has been allocated resources
Mon Nov 11 14:41:14 2024
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 520.61.05 Driver Version: 520.61.05 CUDA Version: 11.8 |
|-------------------------------+----------------------+----------------------+
| GPU Name Persistence-M| Bus-Id Disp.A | Volatile Uncorr. ECC |
| Fan Temp Perf Pwr:Usage/Cap| Memory-Usage | GPU-Util Compute M. |
| | | MIG M. |
|===============================+======================+======================|
| 0 NVIDIA TITAN RTX On | 00000000:18:00.0 Off | N/A |
| 41% 31C P8 30W / 280W | 0MiB / 24576MiB | 0% Default |
| | | N/A |
+-------------------------------+----------------------+----------------------+
| 1 NVIDIA TITAN RTX On | 00000000:3B:00.0 Off | N/A |
| 41% 24C P8 13W / 280W | 0MiB / 24576MiB | 0% Default |
| | | N/A |
+-------------------------------+----------------------+----------------------+
| 2 NVIDIA TITAN RTX On | 00000000:86:00.0 Off | N/A |
| 41% 24C P8 7W / 280W | 0MiB / 24576MiB | 0% Default |
| | | N/A |
+-------------------------------+----------------------+----------------------+
| 3 NVIDIA TITAN RTX On | 00000000:AF:00.0 Off | N/A |
| 41% 24C P8 2W / 280W | 0MiB / 24576MiB | 0% Default |
| | | N/A |
+-------------------------------+----------------------+----------------------+

+-----------------------------------------------------------------------------+
| Processes: |
| GPU GI CI PID Type Process name GPU Memory |
| ID ID Usage |
|=============================================================================|
| No running processes found |
+-----------------------------------------------------------------------------+

>>> srun --partition=class1 --gres=gpu:4 nvidia-smi -q
srun: job 996319 queued and waiting for resources
srun: job 996319 has been allocated resources

==============NVSMI LOG==============

Timestamp : Mon Nov 11 14:44:59 2024
Driver Version : 520.61.05
CUDA Version : 11.8

Attached GPUs : 4
GPU 00000000:18:00.0
    Product Name : NVIDIA TITAN RTX
    Product Brand : Titan
    Product Architecture : Turing
    Display Mode : Disabled
    Display Active : Disabled
    Persistence Mode : Enabled
    MIG Mode
        Current : N/A
        Pending : N/A
    Accounting Mode : Disabled
    Accounting Mode Buffer Size : 4000
    Driver Model
        Current : N/A
        Pending : N/A
    Serial Number : 1324419051655
    GPU UUID : GPU-d7d12c0c-9406-6ae6-beea-8ed0d8d58a71
    Minor Number : 0
    VBIOS Version : 90.02.2E.00.0C
    MultiGPU Board : No
    Board ID : 0x1800
    GPU Part Number : 900-1G150-2500-000
```

```
Module ID : 0
Inforom Version
    Image Version : G001.0000.02.04
    OEM Object : 1.1
    ECC Object : N/A
    Power Management Object : N/A
GPU Operation Mode
    Current : N/A
    Pending : N/A
GSP Firmware Version : N/A
GPU Virtualization Mode
    Virtualization Mode : None
    Host VGPU Mode : N/A
IBMNPU
    Relaxed Ordering Mode : N/A
PCI
    Bus : 0x18
    Device : 0x00
    Domain : 0x0000
    Device Id : 0x1E0210DE
    Bus Id : 00000000:18:00.0
    Sub System Id : 0x12A310DE
    GPU Link Info
        PCIe Generation
            Max : 3
            Current : 1
        Link Width
            Max : 16x
            Current : 16x
    Bridge Chip
        Type : N/A
        Firmware : N/A
    Replays Since Reset : 0
    Replay Number Rollovers : 0
    Tx Throughput : 0 KB/s
    Rx Throughput : 0 KB/s
Fan Speed : 41 %
Performance State : P8
Clocks Throttle Reasons
    Idle : Active
    Applications Clocks Setting : Not Active
    SW Power Cap : Not Active
    HW Slowdown : Not Active
```

```
        HW Thermal Slowdown : Not Active
        HW Power Brake Slowdown : Not Active
    Sync Boost : Not Active
    SW Thermal Slowdown : Not Active
    Display Clock Setting : Not Active
FB Memory Usage
    Total : 24576 MiB
    Reserved : 355 MiB
    Used : 0 MiB
    Free : 24220 MiB
BAR1 Memory Usage
    Total : 256 MiB
    Used : 2 MiB
    Free : 254 MiB
Compute Mode : Default
Utilization
    Gpu : 0 %
    Memory : 0 %
    Encoder : 0 %
    Decoder : 0 %
Encoder Stats
    Active Sessions : 0
    Average FPS : 0
    Average Latency : 0
FBC Stats
    Active Sessions : 0
    Average FPS : 0
    Average Latency : 0
Ecc Mode
    Current : N/A
    Pending : N/A
ECC Errors
    Volatile
        SRAM Correctable : N/A
        SRAM Uncorrectable : N/A
        DRAM Correctable : N/A
        DRAM Uncorrectable : N/A
    Aggregate
        SRAM Correctable : N/A
        SRAM Uncorrectable : N/A
        DRAM Correctable : N/A
        DRAM Uncorrectable : N/A
Retired Pages
```

```
    Single Bit ECC : N/A
    Double Bit ECC : N/A
    Pending Page Blacklist : N/A
Remapped Rows : N/A
Temperature
    GPU Current Temp : 32 C
    GPU Shutdown Temp : 94 C
    GPU Slowdown Temp : 91 C
    GPU Max Operating Temp : 89 C
    GPU Target Temperature : 84 C
    Memory Current Temp : N/A
    Memory Max Operating Temp : N/A
Power Readings
    Power Management : Supported
    Power Draw : 30.16 W
    Power Limit : 280.00 W
    Default Power Limit : 280.00 W
    Enforced Power Limit : 280.00 W
    Min Power Limit : 100.00 W
    Max Power Limit : 320.00 W
Clocks
    Graphics : 300 MHz
    SM : 300 MHz
    Memory : 405 MHz
    Video : 540 MHz
Applications Clocks
    Graphics : 1350 MHz
    Memory : 7001 MHz
Default Applications Clocks
    Graphics : 1350 MHz
    Memory : 7001 MHz
Max Clocks
    Graphics : 2100 MHz
    SM : 2100 MHz
    Memory : 7001 MHz
    Video : 1950 MHz
Max Customer Boost Clocks
    Graphics : N/A
Clock Policy
    Auto Boost : N/A
    Auto Boost Default : N/A
Voltage
    Graphics : N/A
    Processes : None
...

>>> srun --partition=class1 --gres=gpu:4 clinfo
srun: job 996321 queued and waiting for resources
srun: job 996321 has been allocated resources
Number of platforms 1
  Platform Name NVIDIA CUDA
  Platform Vendor NVIDIA Corporation
  Platform Version OpenCL 3.0 CUDA 11.8.88
  Platform Profile FULL_PROFILE
  Platform Extensions cl_khr_global_int32_base_atomics
      cl_khr_global_int32_extended_atomics cl_khr_local_int32_base_atomics
      cl_khr_local_int32_extended_atomics cl_khr_fp64 cl_khr_3d_image_writes
      cl_khr_byte_addressable_store cl_khr_icd cl_khr_gl_sharing
      cl_nv_compiler_options cl_nv_device_attribute_query cl_nv_pragma_unroll
      cl_nv_copy_opts cl_nv_create_buffer cl_khr_int64_base_atomics
      cl_khr_int64_extended_atomics cl_khr_device_uuid cl_khr_pci_bus_info
      cl_khr_external_semaphore cl_khr_external_memory
      cl_khr_external_semaphore_opaque_fd cl_khr_external_memory_opaque_fd
  Platform Host timer resolution 0ns
  Platform Extensions function suffix NV

  Platform Name NVIDIA CUDA
Number of devices 4
  Device Name NVIDIA TITAN RTX
  Device Vendor NVIDIA Corporation
  Device Vendor ID 0x10de
  Device Version OpenCL 3.0 CUDA
  Driver Version 520.61.05
  Device OpenCL C Version OpenCL C 1.2
  Device Type GPU
  Device Topology (NV) PCI-E, 18:00.0
  Device Profile FULL_PROFILE
  Device Available Yes
  Compiler Available Yes
  Linker Available Yes
  Max compute units 72
  Max clock frequency 1770MHz
  Compute Capability (NV) 7.5
  Device Partition (core)
    Max number of sub-devices 1
    Supported partition types None
  Supported affinity domains (n/a)
Max work item dimensions 3
Max work item sizes 1024x1024x64
Max work group size 1024
Preferred work group size multiple 32
Warp size (NV) 32
Max sub-groups per work group 0
Preferred / native vector sizes
  char 1 / 1
  short 1 / 1
  int 1 / 1
  long 1 / 1
  half 0 / 0 (n/a)
  float 1 / 1
  double 1 / 1 (cl_khr_fp64)
Half-precision Floating-point support (n/a)
Single-precision Floating-point support (core)
  Denormals Yes
  Infinity and NANs Yes
  Round to nearest Yes
  Round to zero Yes
  Round to infinity Yes
  IEEE754-2008 fused multiply-add Yes
  Support is emulated in software No
  Correctly-rounded divide and sqrt operations Yes
Double-precision Floating-point support (cl_khr_fp64)
  Denormals Yes
  Infinity and NANs Yes
  Round to nearest Yes
  Round to zero Yes
  Round to infinity Yes
  IEEE754-2008 fused multiply-add Yes
  Support is emulated in software No
Address bits 64, Little-Endian
Global memory size 25396969472 (23.65GiB)
Error Correction support No
Max memory allocation 6349242368 (5.913GiB)
Unified memory for Host and Device No
Integrated memory (NV) No
Shared Virtual Memory (SVM) capabilities (core)
  Coarse-grained buffer sharing Yes
  Fine-grained buffer sharing No
  Fine-grained system sharing No
  Atomics No
Minimum alignment for any data type 128 bytes
Alignment of base address 4096 bits (512 bytes)
Preferred alignment for atomics
  SVM 0 bytes
  Global 0 bytes
  Local 0 bytes
Max size for global variable 0
Preferred total size of global vars 0
Global Memory cache type Read/Write
Global Memory cache size 2359296 (2.25MiB)
Global Memory cache line size 128 bytes
Image support Yes
  Max number of samplers per kernel 32
  Max size for 1D images from buffer 268435456 pixels
  Max 1D or 2D image array size 2048 images
  Max 2D image size 32768x32768 pixels
  Max 3D image size 16384x16384x16384 pixels
  Max number of read image args 256
  Max number of write image args 32
  Max number of read/write image args 0
Max number of pipe args 0
Max active pipe reservations 0
Max pipe packet size 0
Local memory type Local
Local memory size 49152 (48KiB)
Registers per block (NV) 65536
Max number of constant args 9
Max constant buffer size 65536 (64KiB)
Max size of kernel argument 4352 (4.25KiB)
Queue properties (on host)
  Out-of-order execution Yes
  Profiling Yes
Queue properties (on device)
  Out-of-order execution No
  Profiling No
  Preferred size 0
  Max size 0
Max queues on device 0
Max events on device 0
Prefer user sync for interop No
Profiling timer resolution 1000ns
Execution capabilities
```

```
    Run OpenCL kernels Yes
    Run native kernels No
    Sub-group independent forward progress No
    Kernel execution timeout (NV) No
  Concurrent copy and kernel execution (NV) Yes
    Number of async copy engines 3
    IL version (n/a)
  printf() buffer size 1048576 (1024KiB)
  Built-in kernels (n/a)
  Device Extensions cl_khr_global_int32_base_atomics
      cl_khr_global_int32_extended_atomics cl_khr_local_int32_base_atomics
      cl_khr_local_int32_extended_atomics cl_khr_fp64 cl_khr_3d_image_writes
      cl_khr_byte_addressable_store cl_khr_icd cl_khr_gl_sharing
      cl_nv_compiler_options cl_nv_device_attribute_query cl_nv_pragma_unroll
      cl_nv_copy_opts cl_nv_create_buffer cl_khr_int64_base_atomics
      cl_khr_int64_extended_atomics cl_khr_device_uuid cl_khr_pci_bus_info
      cl_khr_external_semaphore cl_khr_external_memory
      cl_khr_external_semaphore_opaque_fd cl_khr_external_memory_opaque_fd
...
```

(b) GPU 모델명은 NVIDIA TITAN RTX이고 노드 당 GPU가 4개씩 부착되어 있다.

(c) GPU의 메모리 크기는 24576 MiB이다.

(d) GPU의 maximum power limit은 280W, maximum SM clock speed는 2100MHz이다.

(e) GPU의 Max work item dimension은 3, Max work item size는 1024 × 1024 × 64, Max work group size는 1024이다.

2 Matrix Multiplication with OpenCL

Matrix A와 B를 작은 타일(32 × 32) 단위로 분할하여 각 work group이 독립적으로 계산을 수행하도록 설계하였으며, 각 work group 내에서는 work item이 matrix의 특정 부분을 담당한다. 한 work item이 여러 element를 동시에 메모리에서 fetch 해오도록 하여 메모리 접근 효율성을 높였다. 또한 local memory를 활용하여 타일 간 데이터 재사용을 높였다.

matmul_initialize 함수는 행렬 곱셈 연산을 수행하기 위한 OpenCL 환경을 초기화한다.

- `clGetPlatformIDs`: OpenCL 플랫폼의 ID를 리턴함.
- `clGetDeviceIDs`: 플랫폼에서 사용할 수 있는 device를 조회함.
- `clCreateContext`: OpenCL context를 생성함.

- `clCreateCommandQueue`: 생성된 context와 device에 대한 command queue를 생성함.
- `clCreateProgramWithSource`: 소스 파일로부터 OpenCL 프로그램을 생성함.
- `clBuildProgram`: 생성된 프로그램을 선택된 device에 맞게 컴파일함.
- `clCreateKernel`: 빌드된 프로그램에서 kernel을 하나 골라 추출하여 생성함.
- `clCreateBuffer`: matrix A, B, C를 저장할 디바이스 메모리 버퍼를 생성.

matmul 함수는 실제 행렬 곱셈 연산을 수행하는 커널을 실행하고 결과를 읽어온다.

- `clEnqueueWriteBuffer`: host memory의 matrix A와 B를 device memory buffer에 전송.
- `clSetKernelArg`: kernel의 인자들을 설정.
- `clEnqueueNDRangeKernel`: 설정된 kernel을 지정된 work group 크기로 실행.
- `clEnqueueReadBuffer`: 연산이 완료된 후 device memory의 데이터를 host memory로 전송.

matmul_finalize 함수는 사용된 OpenCL 리소스를 해제하여 memory leakage를 방지한다.

- `clReleaseMemObject`: device memory buffer를 해제.
- `clReleaseKernel`: 생성된 kernel object를 해제.
- `clReleaseProgram`: 빌드된 프로그램을 해제.
- `clReleaseCommandQueue`: 생성된 command queue를 해제.
- `clReleaseContext`: 생성된 OpenCL context를 해제.

Kernel의 성능 개선은 아래 순서로 이루어졌다:

Kernel 1: Baseline

행렬 곱셈의 기본적인 구현으로, 각 work item이 행렬 C의 한 원소를 계산한다. 별도의 최적화는 적용하지 않았다.

- 각 work item이 $C[i][j]$ 한 원소를 계산
- global memory 접근만 사용
- 데이터 재사용 없음

Kernel 2: Local memory tiling

Local memory를 활용하여 데이터의 재사용을 극대화하는 타일링(Tiling) 기법을 도입하였다. 이를 통해 global memory 접근 횟수를 줄이고, 메모리 대역폭을 효율적으로 활용하였다.

- local memory에 타일 데이터를 로드하여 재사용
- `__local`을 활용항여 메모리 지정
- work group 간의 동기화를 위해 `barrier` 사용

Kernel 3: Tiling with WPT

Work Per Thread(WPT) 개념을 도입하여 각 work item이 여러 원소를 동시에 계산하도록 최적화했다. 타일 크기와 WPT의 비율은 실험을 통해 적절한 값을 사용하였다.

- WPT를 적용하여 한 work item 당 여러 C 원소 계산
- TS와 WPT 비율 조정
- local memory 접근 패턴 최적화
- 병렬 계산의 효율성 증대

Kernel 4: 2-dimensional tiling

Kernel 3의 구조를 기반으로 tile을 두 차원으로 확장했으며, 각 work item이 2차원 타일 블럭을 하나씩 채우게 하였다. work group 구성 및 메모리 접근 패턴을 더 정교화했다.

- Tiling을 2-dimension으로 구현
- work group 및 work item 배치를 최적화

실험 결과

Kernel number	Method	Throughput (GFLOPS)
Kernel 1	Baseline	229.23
Kernel 2	Local memory tiling	407.55
Kernel 3	Tiling with WPT	512.19
Kernel 4	2-dimensional tiling	1436.75