NVIDIA GPU 에서 측정 가능한 GPU 가속기 세부 성능지표 분석

신창용, 양경식, 유혁
고려대학교

{cyshin, ksyang, chuckyoo}@os.korea.ac.kr

An Analysis of Fine-grained GPU Performance Metrics on NVIDIA GPU Accelerators

Changyong Shin, Gyeongsik Yang, Chuck Yoo
Korea Univ.

요 약

최근 딥러닝 모델의 활용도가 증가함에 따라 GPU 기반 딥러닝 모델 학습 및 추론 워크로드수가 증가하고 있다. 뿐만 아니라, 개별 모델의 규모 및 이를 지원하기 위한 GPU 등 가속 하드웨어의 종류도 다양해지고 있다. 이러한 환경에서, 딥러닝 모델 학습과 추론은 높은 비용이 요구되는 온프레미스 기반의 GPU 클러스터 대신 GPU 클라우드가 적극 활용되고 있다. 하지만 GPU 클라우드는 매우 낮은 GPU 활용률을 보이는 것으로 알려져 있으며, 이를 해결하기 위해 GPU 의 세부 성능 지표 모니터링을 통해 GPU 활용 패턴을 이해하고 활용률을 개선하려는 시도가 지속되고 있다. 본 연구는 NVIDIA GPU 에서 측정 가능한 GPU 세부 성능지표를 정리 및 실측하며, 향후 GPU 활용률 개선을 위해 GPU 세부 성능지표의 활용 방안을 제시한다.

1. 서 론

최근 딥러닝 모델의 활용도가 증가함에 따라 GPU 를 사용한 딥러닝 모델의 학습 및 추론 워크로드 수가 증가하고 있다. 딥러닝 모델은 GPU 와 같은 가속기와 함께 구동된다. 이 때, 모델 학습을 위해 온프레미스(on-premise) 방식으로 GPU 클러스터를 구축하는 것은 비용적으로 부담이 크며, GPU 클러스터에서 실행중인 작업이 없을 경우 비용의 손해가 발생한다. 따라서 초기 클러스터 구축비용 절감, 온디맨드(on-demand) 요금제 등 경제적인 장점을 취할 수 있는 GPU 클라우드에서 딥러닝 워크로드를 실행한다.

그러나 최근 연구에서 GPU 클라우드에서 매우 낮은 GPU 활용률이 지적되고 있으며[1], 이에 GPU 활용률 개선을 위한 GPU 클라우드 스케줄링 연구가 진행되고 있다[2]. 이러한 연구들은 GPU 가 사용되는 양상, 즉 성능 지표의 정의 및 모니터링이 선결적으로 필요하며, 최근 연구들에서[3] 기존 GPU 모니터링 툴[4]에서 측정 가능한 지표가 coarse-grained 하며, GPU 성능 지표 모니터링을 통해 GPU 활용율이 개선될 수 있음이 드러나고 있다. 이러한 맥락에서, 본 연구는 NVIDIA 에서 제공하는 가장 대표적인 성능 프로파일링 도구인 DCGM[5]이 제공하는 수백 여 가지의 측정 지표 중, GPU 활용 분석에 유의미한 성능 지표를 선별하며 그 의미 및 측정 결과를 제시한다. 이를 기반으로, 더 나아가 GPU 성능 지표의 구체적인 활용 방안을 논한다.

2. NVIDIA GPU 하드웨어 구조

연산장치. GPU 는 본래 그래픽 처리에 최적화된 연산장치로, 동일한 명령어로 여러 개의 데이터, 또는 쓰레드를 연산 (SIMT)할 수 있도록 계층적으로 설계되었다. 구체적으로, GPU 는 다수의 streaming multiprocessor(SM) 코어로 구성되며, 각 SM 코어 내부는 FP16, 32, 64 등 특정 자료형의 연산을 처리할 수 있는 다수의 세부 코어로 구성되어 있다[6].

메모리. GPU 의 메모리 계층구조는 CPU 와 유사한 DRAM, L2 cache, L1 cache, register 등으로 구성된다. DRAM 은 SM 코어 간에 공유되며, NVIDIA V100 GPU 는 900 GB/s 의 메모리 대역폭을 지원하여 거대한 학습 데이터셋과 딥러닝 모델의 크기 증가로 인해 고대역폭이 요구되는 딥러닝 워크로드를 수행하기에 적합하다. L2 cache 는 SM 코어간, L1 cache 는 SM 코어 내 세부 코어들간 공유되며 register 는 SM 코어 내 세부 코어들간 공유되지 않는다.

3. GPU 세부 성능 지표

NVIDIA DCGM[5]은 데이터센터 내 GPU 들의 상태를 모니터링하기 위해 개발된 툴로, 실시간으로 GPU 의 연산장치 및 메모리의 활용률 조회 기능을 제공한다. 본 논문에서는 DCGM 에서 측정 가능한 수백 여 가지 GPU 성능 지표 중 연산장치 및 메모리의 활용률을 나타내는 주요 지표 5 가지를 선별하고 정리한다. 표 1 은 선별된 지표에 대한 의미와 단위 등을 요약하고 있으며, 아래 단락부터 순차적으로 논한다.

연산장치. 프로세서. SM_ACTIVE, FP32_ACTIVE 는 연산장치와 관련된 세부 성능지표이다. FP16, FP64 세부 코어에 대한 세부 성능지표의 측정 또한 가능하며, SM 코어의 clock 수 측정도 가능하다. POWER_USAGE 는 조회 시 GPU 에 공급되고 있는 전력의 크기를 의미한다.

세부 성능지표	의미	단위 및 스케일
SM_ACTIVE	단위시간 동안 각 SM 코어의 활성화된 기간의 평균	0~1
FP32_ACTIVE	단위시간 동안 각 SM 코어 내 FP32 세부 코어가 활성화된 기간의 평균	0~1
DRAM_ACTIVE	단위시간 동안 GPU 의 DRAM 으로 데이터가 송수신된 기간의 평균	0~1
PCIE_TX_BYTES	단위시간 동안 PCIe 버스를 통해 GPU 의 DRAM 에서 호스트 메모리로 전송된 데이터의 양	bytes/s
POWER_USAGE	조회 시 GPU 에 공급되는 전력의 크기	W

표 1. GPU 세부 성능지표의 종류 및 의미

메모리. DRAM_ACTIVE, PCIE_TX_BYTES 는 메모리와 관련된 세부 성능지표이다. PCIE_RX_BYTES 와 사용중인 DRAM 의 크기 또한 측정 가능하다.

4. GPU 세부 성능지표의 측정 및 활용 방안

4.1 측정 환경

본 논문은 DCGM 을 사용하여 앞서 선별한 지표를 측정한다. 각 세부 성능지표는 0.1 초 간격으로 측정한 값들의 평균이며, 총 3 회 반복 실험하여 정리했다. 또한 각 세부 성능지표의 스케일이 다르기 때문에, 각 지표의 최대값을 기준으로 정규화 한다.

측정 대상 딥러닝 모델은 대표적인 이미지분류 모델인 Inception-v4, ResNet-152, VGG-16 이다. 또한, 파라미터 서버 2 개와 V100 GPU 를 사용하는 워커 2 개로 분산 학습한다.

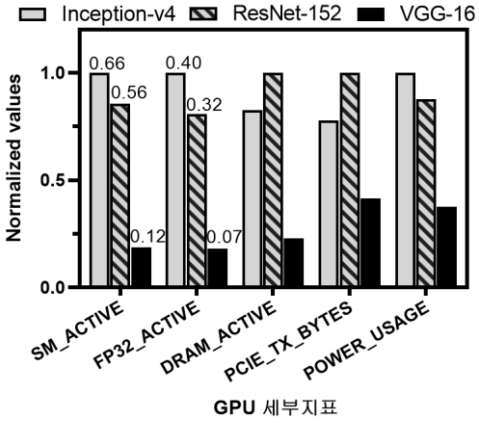

그림 1. 딥러닝 워크로드의 GPU 세부 성능지표 측정 결과

4.2 측정 결과

GPU 세부 성능지표 측정 결과, 딥러닝 모델마다 다른 GPU 세부 성능지표 활용률을 보였다(그림 2). Inception-v4 모델의 경우 연산장치 관련 지표에서 가장 높은 사용률을 보였으며, ResNet-152 모델은 메모리 관련 지표에서 가장 높은 수치를 보였다. VGG-16 모델의 경우 Inception-v4, ResNet-152 모델에 비해 비교적 낮은 연산장치, 메모리 사용량을 보였다.

GPU 세부 성능지표 간의 관계를 살펴보면, FP32_ACTIVE 지표의 수치가 SM_ACTIVE 지표의 수치의 60% 정도로 나타난다. 이는 FP32 세부 코어가 SM 코어에 내부에 존재하는 fine-grained 한 코어이며, SM 코어가 활성화된 구간 중 60% 정도는 FP32 세부 코어를 사용했음을 의미한다.

또한 각 세부 성능지표를 정규화 할 경우, 연산장치 관련 세부 성능지표인 SM_ACTIVE 와 FP32_ACTIVE, POWER_USAGE 의 Pearson R 상관계수가 각각 0.97, 0.98 로 서로 높은 상관성을 보였으며, 메모리관련 세부 성능지표들 역시 0.85 로 서로 유사한 상관성을 보였다.

4.3 활용 방안 제언

GPU 활용률을 높이기 위해 단일 GPU 에 복수개의 딥러닝 학습 워크로드를 실행하는 GPU 공유 기법이 제안되었다[2, 7]. 해당 연구에서, 동시 실행되는 딥러닝 학습 워크로드에 따라 워크로드 간 간섭의 정도가 달라짐이 보고되었다. 워크로드 간 간섭은 GPU 자원(연산장치 및 메모리)에 대한 경합으로 인해 발생하며, GPU 세부 성능지표 측정으로 워크로드 간 간섭의 정도를 가늠해볼 수 있으리라 기대된다. 예를 들어, 본 논문의 측정 결과를 기반으로 보면, Inception-v4 와 ResNet-152 를 동시 실행하는 경우보다, Inception-v4 와 VGG-16 또는 ResNet-152 와 VGG-16 모델을 동시 학습하는 경우가 학습 시간 및 간섭 측면에서 더 효과적일 것으로 판단해볼 수 있다.

5. 결 론

본 연구에서는 NVIDIA GPU 의 연산장치 및 메모리에 대해 다양한 GPU 성능 지표가 존재하며, 성능 이해에 주요한 프로세서 및 메모리의 주요 지표를 선별하였다. 대표지표에 대한 성능 지표 측정 결과 딥러닝 모델에 따라, 성능이 다양하게 나타남을 확인했다. 향후 GPU 세부 성능지표 측정 결과를 기반으로 딥러닝 워크로드 간 간섭을 예측하여 GPU 활용률 연구에 활용할 수 있다.

ACKNOWLEDGMENT

이 논문은 2022 년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업(No. NRF-2021R1A6A1A13044830)과 2022 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(No. 2015-0-00280, (SW 스타랩) 성능 및 보안 SLA 보장이 가능한 차세대 클라우드 인프라 SW 개발)을 받아 수행된 연구임.

참 고 문 헌

[1] Jeon, Myeongjae, et al. "Analysis of Large-Scale Multi-Tenant GPU Clusters for DNN Training Workloads." 2019 USENIX Annual Technical Conference (USENIX ATC 19).
[2] Xiao, Wencong, et al. "AntMan: Dynamic Scaling on GPU Clusters for Deep Learning." 14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20).
[3] Jiang, Yuting, et al. "Moneo: Non-intrusive Fine-grained Monitor for AI Infrastructure." ICC 2022-IEEE International Conference on Communications. IEEE, 2022.
[4] https://developer.nvidia.com/nvidia-system-management-interface.
[5] https://docs.nvidia.com/datacenter/dcgm/dcgm-api/.
[6] https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf
[7] Bao, Yixin, Yanghua Peng, and Chuan Wu. "Deep learning-based job placement in distributed machine learning clusters." IEEE INFOCOM 2019-IEEE conference on computer communications.