컴퓨터 구조 – 과제#1

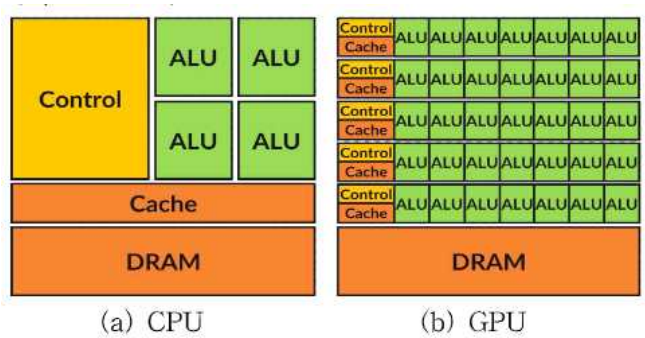
GPU 조사

ICT융합공학부 202204010 공성택

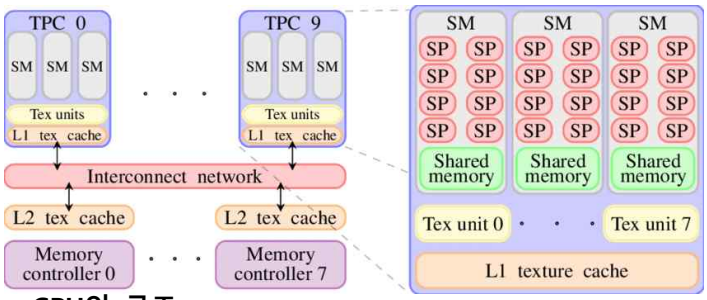
GPU의 구조는?

GPU란 컴퓨터 시스템에서 그래픽 연산을 빠르게 처리해 결과 값을 모니터에 출력하는 연산장치이다. 여러 명령어를 동시에 처리하는 병렬 처리 방식을 사용하고 있다. GPU는 크게 SP, SM, TPC 3가지의 구성요소로 볼 수 있다. SP(Steaming Processor)란 GPU의 기본 유닛 단위로, GPU 내에서 병렬 처리를 담당한다. SM(Steaming Multiprocessor)는 병렬 처리를 지원하는 프로세스 그룹이다. SM은 명령어를 해독하고 SM에 포함되어 있는 각각의 SP에 명령을 배정하여 SP들은 각각 작은 산술 및 논리 연산을 수행하며 병렬처리 작업을 한다. TPC(Texture/Processor Cluster(TPC)는 GPU 내에서 그래픽 처리를 위한 기능을 담당한다. TPC는 텍스트 유닛의 집합으로 텍스처 유닛은 그래픽 처리에서 텍스처를 읽고 처리해 그래픽 객체에 적용한다. GPU는 이런 다수의 코어를 갖고 있어, 병렬 데이터 처리에 매우 효과적이며, 이를 통해 고성능 컴퓨팅 작업을 빠르게 처리할 수 있다.

CPU와 비교하면 CPU는 범용 계산 처리를 목적으로 고성능 코어를 소수 탑재해 직렬과 병렬 처리가 가능하지만 병렬처리는 GPU에 밀린다. 반대로 GPU는 그래픽 처리를 목적으로 저성능 코어 다수를 탑재했다. 때문에 부동 소수점 연산을 통한 병렬 처리에 최적화되어있어 대량의 데이터를 효율적으로 처리할 수 있다. 또한, GPU는 메모리 대역폭을 최적화하기 위해 고대역폭 메모리(HBM)등의 특수한 메모리 구조를 사용하고, 고성능과 대량의 처리를 우선하기 때문에, 전력 소비가 높다.



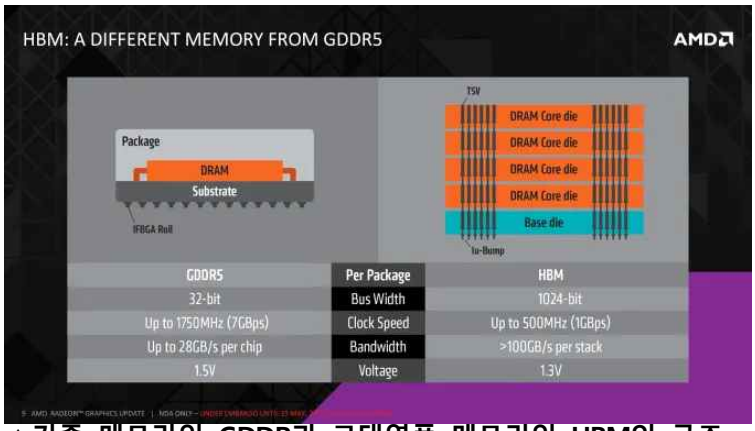
▲CPU와 GPU의 구조.



▲GPU의 구조.

GPU에서 HBM이란?

HBM(High Bandwidth Memory)이란 고대역폭 메모리로, GPU와 같은 고성능의 컴퓨터 장치에서 사용되는 고성능 메모리이다. 여러 개의 실리콘 메모리칩을 수직으로 쌓아 메모리 용량을 크게 늘리고, 정보 전달 거리를 줄여서 데이터 전송 속도를 대폭 상향했다. 이는 데이터의 효율적인 이동을 가능케 해서 메모리와 처리 장치 간의 병목 현상을 최소화할 수 있다. 기존 DDR 메모리와 비교했을 때, HBM의 대역폭이 훨씬 높기 때문에 다양한 데이터 집약 작업에서 유용하고, 그래픽 처리, 빅데이터 분석, 시뮬레이터 등의 분야에서 뛰어난 성능을 발휘하고 있다. 현재 AI가 무서운 속도로 발전하고 있는 상황에서 HBM은 인공지능에 사용되는 GPU의 필수재로 주목받고 있다. AI가 대규모 데이터를 학습하기 위해서 수많은 GPU를 활용하고, 또 방대한 데이터를 학습하려면 데이터 처리와 저장 기능이 가장 중요하다. 이런 점에서 HBM은 인공지능 시대의 열쇠로 주목받고 있으며, 삼성을 비롯한 다양한 기업들은 HBM 기술을 개발하고 상용화하여 적용하려고 하고 있다. 정교한 AI 모델이 등장하면서 계산 요구가 계속 증가함에 따라, 고대역폭이자 에너지 효율적인 HBM은 핵심 기술로써 차세대 컴퓨팅 기능을 가능하게 할 것이다.



▲기존 메모리인 GDDR과 고대역폭 메모리인 HBM의 구조.

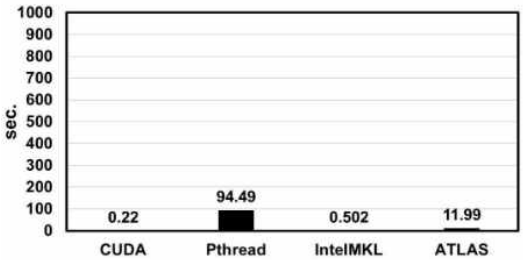
GPU가 왜 CPU보다 딥러닝을 할 때, 더 좋은 성능을 낼 수 있는가?

AI 모델에서는 프로세서 병렬 연산 속도가 얼마나 빠르냐를 기준으로 성능을 구분한다. 이점에서 병렬처리에 특화된 GPU는 이미지 인식, 객체 감지 및 비디오 처리 등과 같이, 반복적이고 비슷하고, 대량의 연산을 수행하는 작업에서 직렬, 병렬 범용적인 용도로 사용되는 CPU보다 월등히 작업 속도가 빠르다. 구조적인 측면에서 자세히 살펴보면, GPU는 CPU보다 산술논리연산장치(ALU)가 월등히 많다. 때문에 산술 논리연산에서 GPU가 CPU보다 성능이 좋을 수밖에 없다. CPU는 다양한 작업을 처리하기 위해서 CU, cache, ALU 등의 다양한 프로세서들로 이루어져 처리 구조가 복잡하다. 하지만 GPU는 특화된 연산만을 위해 복잡한 구조 대신 다수의 ALU로 구성되었다. 특히 12코어의 고성능 코어를 가진 CPU와 SP가 4864개 정도인 그래픽 카드가 비슷한 값이다. 딥러닝은 고수준의 연산보다는, 엄청나게 많은 저수준의 연산들로 이루어져있기 때문에, 성능이 좋은 CPU보다 더 많은 GPU를 사용하는 것이 더 좋은 성능을 낼 수 있다. 따라서 다량의 연산을 할 때에는 CPU가 GPU에게 명령을 해서 GPU를 통해 수행하는 식으로 진행된다.

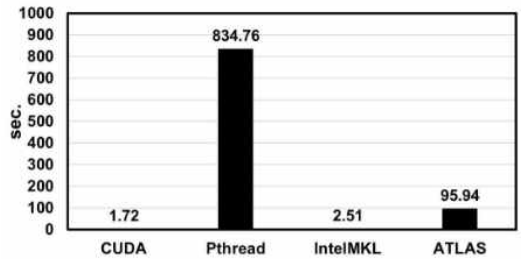
실제로 「딥 러닝 연산을 위한 GPU/CPU 성능 분석」(최진서, 강동현, 2022)이라는 논문에서 행렬 곱 연산 수행 시간을 비교하는 실험 결과에 따르면, 10코어 CPU 프로세서 2개와,

2560개의 코어로 구성된 GPU의 행렬 곱셈 연산 처리 속도의 차이는 GPU가 CPU 대비 최대 485배 높은 것으로 나타났다.

위에서 정리한 내용처럼 나는 딥러닝이라는 수행에서는 CPU의 범용 처리보다 GPU의 병렬처리 특화가 더 좋은 성능을 보여줄 것이라는 의견이고 실제로도 그렇다. 딥러닝은 복잡한 고수준의 연산보다는 이미지와 비디오를 읽고 처리하고, 객체를 인식하는 등의 단순한 반복 연산이 주를 이루기 때문에 수많은 연산 프로세서를 가진 GPU가 해당 작업에서는 월등히 뛰어난 모습을 보여줄 것이다. 그러나 이는 GPU가 CPU보다 무조건적으로 좋다는 의견이 아니고, 추후에 인공지능이 발전하고, 딥러닝이나 인공지능의 연산에 복잡한 구조가 주를 이룬다면 CPU가 GPU보다 효율적인 상황이 생길 것이다. 그러면 또 현대 과학자, 기술자들은 인공지능이라는 분야를 연구하면서 HBM과 같은 더 성능 좋은, 더 효율적인 컴퓨터 프로세서들을 개발하거나 CPU와 GPU를 모두 사용하는 GPGPU와 같이 새로운 사용방법을 찾게 될 것이고, 결국 높은 수준의 인공지능이 상용화되어 모든 사람이 인공지능을 주로 활용하는 시대가 올 것이라고 생각한다. 나 또한 이번 과제를 통해 조사하며 HBM이나 GPU의 장점, 구조 등 새로운 정보들을 공부했는데, IT 업계에서 일하고 공부하기 위해서 더 새로운 기술이나 정보들에 관심을 갖고 찾아볼 수 있도록 해야겠다고 반성했다. 특히 인공지능 분야에 대해 기술 동향 등을 관심 갖고 정보를 찾도록 할 것이다.



▲ 4096 행렬 곱 수행시간 비교.(GPU를 사용한 CUDA의 연산이 가장 빠르다.)



▲ 8192 행렬 곱 수행시간 비교.(GPU를 사용한 CUDA의 연산이 가장 빠르다.)

참고문헌

최진서, 강동현. "딥 러닝 연산을 위한 GPU/CPU 성능 분석." 한국정보과학회 학술발표논문집 2022.12 (2022): 1145-1147.