거대언어모델 가속화를 위한 이기종 컴퓨팅 시스템 구조 연구

김찬훈, 정기석*
한양대학교

kch1103@hanyang.ac.kr, *kchung@hanyang.ac.kr

A Study on Heterogeneous Computing System Architecture for Large Language Models Acceleration

Chan Hoon Kim, Ki-Seok Chung*
Hanyang University, Seoul, Korea

요 약

트랜스포머 기반 거대언어모델은 우수한 성능 덕분에 다양한 인공지능 어플리케이션에서 널리 활용되고 있다. 그러나 최근 거대언어모델의 발전은 GPU 의 계산 및 메모리 성능의 한계를 드러내고 있으며, 특히 어텐션 레이어에서 발생하는 메모리 대역폭 병목 현상과 데이터 이동의 과도한 비용이 주요한 문제로 대두되고 있다. 이러한 한계를 해결하기 위하여 GPU 와 고대역폭 메모리(HBM) 기반 프로세싱 인 메모리(PIM)를 결합한 이기종 시스템을 활용하고자 하는 시도가 있다. 따라서 본 논문에서는 거대언어모델 추론에서 GPU 단독 시스템과 GPU 와 HBM-PIM 의 이기종 시스템 간의 성능을 비교 분석하여, 이기종 시스템의 성능 향상을 확인한다. GPU 와 HBM-PIM 이기종 시스템인 AttAcc 을 활용하여 OPT 6.7B, 13B 모델에 적용하여 배치 크기 8, 입출력 시퀀스 길이 1024 일 때 GPU 단독 시스템 대비 최대 x1.46 지연시간 감소와 x1.46 처리량 향상을 보임을 확인하였다.

Ⅰ. 서 론

트랜스포머 기반 거대언어모델은 인공지능 챗봇, 텍스트 요약, 코드 생성 등 자연어 처리 분야뿐만 아니라 이미지, 비디오, 음성 처리와 같은 다양한 어플리케이션에도 활용되며 우수한 성능을 보이고 있다 [1]. 이러한 모델의 추론 과정은 한 번의 Prefill 단계와 반복적인 Decode 단계로 구성되며, 각 단계는 다수의 어텐션 레이어와 완전 연결 (Fully Connected, FC) 레이어로 이루어져 있다. 이때 모델의 순차적 특성으로 인하여 Decode 단계가 일반적으로 전체 실행 시간에서 대부분의 비중을 차지하게 된다. 예를 들어, 입력 시퀀스와 출력 시퀀스의 길이가 각각 1024 일 때, Decode 단계에서의 어텐션 레이어와 FC 레이어는 하이엔드 GPU 가 제공하는 메모리 대역폭 보다 훨씬 더 높은 메모리 대역폭을 요구한다 [2,3]. 이러한 메모리 대역폭 병목 현상으로 인해 기존의 GPU 단독 시스템으로는 거대언어모델의 추론을 효율적으로 가속화하는 데 한계가 존재하게 된다. 이를 해결하기 위하여 NeuPIM [2], AttAcc [3] 등 NPU 또는 GPU 와 프로세싱 인 메모리(PIM)를 결합한 이기종 시스템의 연구가 진행되었으며 Decode 단계의 어텐션 레이어에서 발생하는 데이터 이동 비용과 메모리 대역폭 병목 문제를 완화하고자 하였다. 이런 배경에서 본 논문에서는 AttAcc[3]을 활용하여 GPU 단독 시스템과의 비교 분석을 통하여 성능 향상 여부를 확인하고 결과를 분석한다.

Ⅱ. 본론

2.1 거대언어모델 추론 과정

트랜스포머 기반 거대언어모델의 추론 과정은 그림 1 과 같다. Prefill 단계는 입력 시퀀스 전체에 대한 병렬 연산을 수행하는 과정으로, 첫 출력 토큰을 생성한다. Decode 단계는 반복적으로 진행되며 새로운 출력 토큰을 순차적으로 생성한다. 두 단계에서의 FC 레이어는 배치의 크기를 증가하게 되면 행렬-행렬 연산으로 변환되며 GPU 의 병렬 연산을 최적으로 활용할 수 있다. 반면, Decode 단계에서의 어텐션 레이어는 Prefill 단계의 어텐션 레이어와는 다르게 배치의 크기를 증가하더라도 행렬-벡터 연산으로 수행하게 되어 GPU 의 연산 유닛을 효율적으로 활용할 수 없게 되며 메모리 대역폭 병목 현상이 나타나게 된다 [2,3].

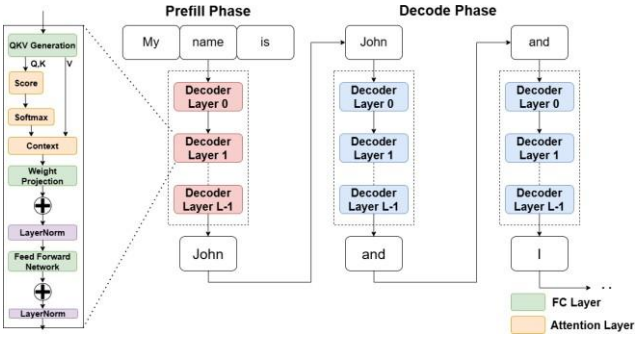

그림 1. 거대언어모델 추론 과정

2.2 GPU 와 HBM-PIM 이기종 시스템 구조

본 논문에서 활용하는 AttAcc 기반 이기종 시스템 구조는 그림 2 와 같다. Host CPU 는 GPU 와 HBM-PIM 의 동작 스케줄링을 담당하며 Prefill 단계와 Decode 단계의 FC 레이어를 담당하는 GPU 와 Decode 단계의 어텐션 레이어를 수행하는 HBM-PIM 장치로 구

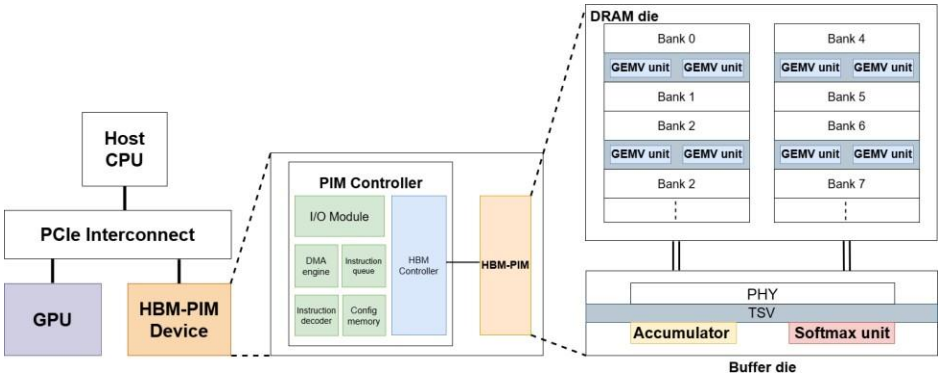

그림 2. GPU 와 HBM-PIM 이기종 시스템 및 HBM-PIM 구조

성된다. GPU 와 HBM-PIM 은 PCIe 4.0 으로 연결된다. 호스트는 HBM-PIM 장치의 제어 레지스터들과 config 메모리를 memory-mapped I/O 로 통해서 접근한다. Instruction queue 와 instruction decoder 는 각각 PIM 명령어가 저장되고 디코딩 되는 부분이며 DMA 를 통해 다른 장치에 접근하여 어텐션 연산을 위한 값들을 저장한다. HBM-PIM 장치는 8 단 HBM3 를 기반으로 하며 8 단 HBM3 는 8 개의 DRAM 다이와 이를 연결하는 버퍼 다이가 TSV 를 사용하여 3D 적층된 구조로 구성되어 있다. 4 개의 DRAM 다이는 하나의 랭크를 구성하고 각 DRAM 다이는 독립적으로 동작하는 8 개의 pseudo 채널로 이루어져 있으며, 각 pseudo 채널은 16 개의 뱅크를 포함하고, 이 뱅크들은 4 개의 뱅크 그룹으로 묶여 있다. HBM-PIM 장치는 그림 2 와 같이 두 가지 연산 유닛을 가지고 있다. 16 비트 부동소수점 기반의 행렬-벡터 연산과 소프트맥스 연산을 수행하며 각각의 연산을 위하여 행렬-벡터 연산기는 DRAM 다이의 뱅크 단위로 배치하고 소프트맥스 연산기는 버퍼 다이에 배치함으로써 HBM 의 내부 대역폭을 최대로 활용하여 연산을 수행할 수 있도록 한다.

2.3 실험 환경 및 실험 결과

실험 환경: 본 논문에서는 AttAcc 시뮬레이터 [3]를 수정하여 실험을 진행하였다. GPU 와 HBM-PIM 이기종 시스템 구성은 NVIDIA A100 80GB GPU 1 대와 16GB 메모리 용량을 가지는 HBM-PIM 1 대로 구성하였으며 시스템의 전체 메모리 용량은 96GB 이다. 모델의 모든 파라미터를 GPU 메모리에 수용하기 위하여 OPT 6.7B, 13B 모델에 대해서 이기종 시스템과 GPU 단독 시스템 간의 비교 분석을 진행하였다. 입력 시퀀스와 출력 시퀀스 길이는 모두 1024 로 총 2048 길이로, 배치 크기는 모두 8 로 설정하였다.

실험 결과: 실험 결과는 표 1 과 같다. 성능 비교는 지연시간 (latency)과 처리량 (throughput)의 비교로 진행되었으며 지연시간은 Prefill 과 Decode 연산 시간의 합으로, 처리량은 배치 크기를 지연시간으로 나눈 값으로 계산한다. OPT-6.7B 모델에 대하여 GPU 단독 시스템은 지연시간 8.49ms, 처리량 942.53 tokens/s 인 반면 GPU 와 HBM-PIM 의 이기종 시스템은 지연시간 5.81ms, 처리량 1376.19 tokens/로 GPU 단독 시스템 대비 1.46 배 지연시간 감소, 1.46 배의 처리량 향상을 이루었다. OPT-13B 모델에 대해서는 GPU 단독 시스템은 지연시간 13.49ms, 처리속도 593.08 token/s 인 반면 이기종 시스템은 지연시간 10.17ms, 처리량 786.9 tokens/s 로써 GPU 단독 시스템 대비 1.33 배의 지연시간 감소와 1.32 배의 처리량 향상을 이루었다.

Table 1: Latency and Throughput Comparison

Model	Configuration	Latency (ms)	Throughput (tokens/s)
OPT-6.7B	GPU	8.49	942.53
	GPU & HBM-PIM	5.81	1376.19
OPT-13B	GPU	13.49	593.08
	GPU & HBM-PIM	10.17	786.9

Ⅲ. 결론

본 논문에서는 기존의 거대언어모델 추론 가속을 위하여 활용되던 GPU 단독 시스템에서 벗어나 GPU 와 HBM 기반 PIM 의 이기종 시스템을 활용하여 비교 분석 및 성능 향상을 확인하였다. OPT 6.7B, 13B 모델에 대하여 GPU 와 HBM-PIM 이기종 시스템은 기존 GPU 단독 시스템 대비 최대 x1.46 의 지연시간 감소 및 x1.46 의 처리량의 성능 향상을 보임을 확인하였다.

ACKNOWLEDGMENT

이 논문은 2020 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. 2020-0-01304, 모바일 자가 학습 가능 재귀 뉴럴 네트워크 프로세스 기술 개발).

EDA Tool지원받은 경우: 본 연구는 IDEC에서 EDA Tool를 지원받아 수행하였습니다.

참 고 문 헌

[1] Ashish Vaswanmi. Et. Al., “Attention is All you Need.", Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, NeurIPS 2017.

[2] Heo, Guseul, et al. "Neupims: Npu-pim heterogeneous acceleration for batched llm inferencing." Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3. 2024.

[3] Park, Jaehyun, et al. "AttAcc! Unleashing the Power of PIM for Batched Transformer-based Generative Model Inference." Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2. 2024.