데이터 집약적 연산의 성능을 위한 Near-Data Processing(NDP) 기반의 프로세서 아키텍처의 최근 동향

김가은[1], 오현영[2*]
[1]가천대학교 응용통계학과 학부생
[2]가천대학교 AI 소프트웨어학부 교수
rkdms2164@gachon.ac.kr, hyoh@gachon.ac.kr

Near-Data Processing (NDP) Based Processor Architectures for Data-Intensive Computing Performance

Ga-Eun Kim[1], Hyunyoung Oh[2]
[1]Dept. of Applied Statistics, Gachon University
[2]Dept. of AI · Software, Gachon University

요 약

본 논문에서는 데이터 집약적 연산의 효율성 향상을 위한 새로운 프로세서 아키텍처 접근법인 Near-Data Processing(NDP) 기술의 최근 동향을 살펴본다. AI 모델 학습과 빅데이터 분석 등에서 NDP 기술의 적용 사례를 분석하고, Processing-In-Memory(PIM)와 In-Storage Processing(ISP)를 중심으로 한 최신 연구를 소개한다. 또한, 실제 하드웨어 구현 사례로 Xilinx FPGA 개발보드의 HBM 통합 사례를 다룬다.

1. 서론

최근 AI 모델 학습, 빅데이터 분석과 같은 데이터 집약적 연산의 중요성이 증가하면서, 기존 프로세서 아키텍처의 한계를 극복하기 위한 새로운 접근법들이 제안되고 있다. 그 중 Near-Data Processing(NDP)는 데이터 이동을 최소화하고 연산 효율성을 높이는 방안으로 주목받고 있다.

본 논문에서는 NDP 기술의 최근 동향을 살펴보고, 특히 Processing-In-Memory(PIM)와 In-Storage Processing(ISP) 기술을 중심으로 연구 사례를 소개한다. 또한, 실제 하드웨어 구현 사례로 Xilinx FPGA 개발보드의 HBM(High Bandwidth Memory) 통합 사례를 다룬다.

2. 관련 기술

2.1 CPU와 GPU의 한계

CPU 와 GPU 는 현대 컴퓨팅 시스템의 중심 처리 장치로, 각각 고성능 단일 작업 처리와 대규모 병렬 처리에 특화되어 있다. 그러나 이들 모두 데이터를 메모리에서 처리 장치로 이동시켜 연산을 수행하는 구조를 가지고 있어, 데이터 전송이 빈번하게 발생한다.

특히 빅데이터 처리나 AI 모델 학습과 같은 데이터 집약적 작업에서는 이러한 구조적 한계로 인해 심각한 성능 저하가 발생한다. 대용량 데이터를 처리할 때 메모리 대역폭의 한계로 인한 병목 현상이 발생하며, 이는 전체 시스템의 성능을 크게 저하시킨다.

2.2 NDP(Near-Data Processing)

NDP 는 이러한 전통적인 컴퓨팅 구조의 한계를 극복하기 위해 제안된 새로운 패러다임이다. NDP 는 데이터가 저장된 위치 근처에서 직접 연산을 수행함으로써 데이터 이동을 최소화하고 성능을 향상시키는 것을 목표로 한다.

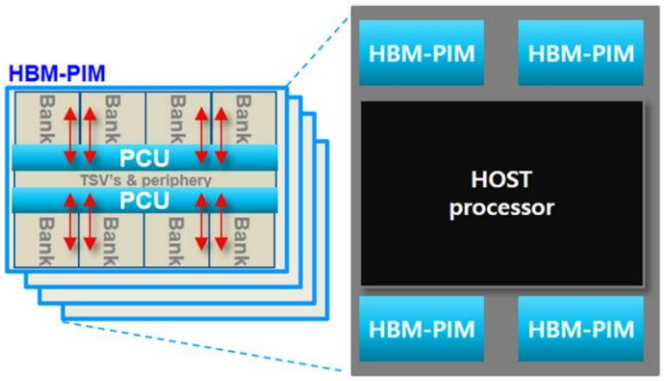
2.3 PIM(Processing In Memory)

PIM 은 NDP 의 대표적인 구현 방식 중 하나로, 메모리 칩 내부에 연산 능력을 추가하는 기술이다. 이를 통해 메모리에서 직접 일부 연산을 수행하여 데이터 전송량을 줄이고 성능을 향상시킬 수 있다. 그림 1 은 삼성전자의 Aquabolt-XL(HBM2-PIM)로서 PIM 기술의 실제 적용 사례이다[1]. 이 구조에서는 DRAM 뱅크 어레이의 I/O 경계에 프로그래머블 컴퓨팅 유닛(PCU)이 통합되어 있다. 이러한 설계로 메모리 내부에서 직접 연산을 수행할 수 있어, 데이터 이동을 최소화하고 병렬 처리 능력을 향상시킬 수 있다. Kim et al.[2]의 연구에서는 PIM 환경에서 딥러닝 및 머신 러닝에 자주 사용되는 ADD 및 GEMV(Matrix-Vector Multiplication) 연산을 효율적으로 수행하기 위한 프로그래밍 기법을 제안했다. 이 연구는 PIM 을 사용한 행렬 연산이 CPU 에 비해 약 2.4 배의 성능 향상을 보여주었다.

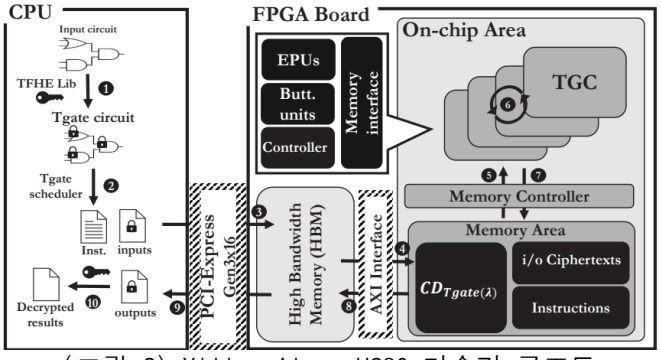
2.4 ISP(In-Storage Processing)

ISP 는 NDP 의 또 다른 구현 방식으로, 저장 장치 내부에 프로세싱 능력을 추가하여 데이터 처리를 수행하는 기술이다. Choe et al.[3]의 연구에서는 ISP-ML 이라는 새로운 SSD 플랫폼을 제안하여 머신러닝 워크로드의 인-스토리지 처리(ISP) 성능을 평가했다.

* 교신저자



(그림 1) Samsung Aquabolt-XL 구조도



(그림 2) Xilinx Alveo U280 가속기 구조도

실험 결과, EASGD 가 동기식 SGD 와 Downpour SGD 에 비해 각각 평균 5.24 배와 1.96 배 더 나은 성능을 보였으며, 채널 수를 8 개에서 16 개로 늘렸을 때 동기식 SGD 의 경우 1.48 배의 속도 향상을 달성했다. 또한 ISP 기반 접근법이 호스트 메모리가 부족한 상황에서 기존 방식보다 더 효율적임을 입증하여, NDP 의 잠재력을 확인하였다.

2.5 HBM(High Bandwidth Memory)과 FPGA 통합

HBM(High Bandwidth Memory)과 FPGA 의 통합은 NDP 개념을 실제 하드웨어에 구현한 사례로 볼 수 있다. Xilinx 의 Alveo U280 데이터센터 가속기 카드는 8GB 의 HBM2 를 다양한 워크로드 처리 하드웨를 프로그래밍할 수 있는 FPGA 칩과 동일한 패키지에 통합하여 최대 460GB/s 의 메모리 대역폭을 제공한다[4]. 이러한 HBM-FPGA 통합은 AI 연산뿐만 아니라 다양한 데이터 집약적 워크로드에 대해 최적화된 하드웨어 가속기를 구현할 수 있게 한다. Nam et al.[5] 연구에서는 TFHE(Fully Homomorphic Encryption over Torus) 연산을 가속화하기 위해 그림 2 와 같이 Xilinx Alveo U280 FPGA 에 XHEC 라는 가속기를 구현했다. XHEC 는 Near-Data Processing 의 개념을 활용하여 다수의 Tgate 연산 코어(TGC)를 병렬로 실행하고, 데이터 이동을 최소화하는 최적화된 메모리 레이아웃을 사용한다. 이러한 접근 방식으로 XHEC 는 기존 CPU, GPU, FPGA 기반 구현에 비해 처리량을 2.43 배에서 44.66 배까지 향상시켰으며, 와트당 처리량은 12.19 배에서 61.65 배까지 개선했다.

2.6 NDP 아키텍처 및 마이크로아키텍처

표 1 은 지금까지 다룬 데이터 이동을 최소화하고 성능을 향상시키기 위한 각 아키텍처의 특성과 장점을 제시한다. 이를 통해 NDP 기술의 다양한 구현 방식이 어떻게 데이터 집약적 작업에서의 성능 향상을 도모하는지를 알 수 있으며, 각각의 아키텍처가 제공하는 주요 장점을 비교하여 특정 응용에 적용할 수 있다.

(표 1) NDP 기술의 아키텍처와 마이크로아키텍처

유형	마이크로 아키텍처 구현	기능 및 장점	응용 분야
NDP	UPMEM DRAM	메모리 내 연산 유닛 추가, 데이터 이동 최소화	빅데이터 분석, 머신러닝 가속화
PIM	Samsung HBM2(Aquabolt-XL)	DRAM 내 프로그래머블 컴퓨팅 유닛(PCU)추가, 높은 병렬 처리 성능	딥러닝 모델 학습, 이미지 처리
ISP	Xilinx Alveo U280	FPGA 와 HBM 의 통합으로 맞춤형 가속기 구현 가능, 높은 메모리 대역폭	AI 모델 학습, 암호화 연산 가속화

3. 결론

AI 모델 학습과 빅데이터 분석 같은 데이터 집약적 작업의 중요성이 증가함에 따라, NDP 기술이 주목받고 있다. PIM 과 ISP 를 중심으로 한 최근 연구들은 이 기술의 실용화 가능성을 보여주고 있다. 삼성의 Aquabolt-XL, Xilinx 의 Alveo U280, XHEC 등의 사례에서 볼 수 있듯이, NDP 기술은 데이터 이동을 최소화하고 연산을 데이터 저장 위치 근처에서 수행함으로써 성능과 에너지 효율성을 크게 향상시킬 수 있다.

그러나 NDP 의 광범위한 적용을 위해서는 하드웨어 구조의 혁신과 이를 활용할 수 있는 소프트웨어 생태계의 발전이 필요하다. 향후 연구에서는 NDP 기술의 실제 시스템 적용 사례를 확대하고, 다양한 워크로드에서의 성능 평가가 필요할 것이다. 또한, 기존 프로세서 아키텍처와 NDP 를 효과적으로 결합한 하이브리드 시스템에 대한 연구도 중요한 방향이 될 것이다.

사사문구

이 논문은 2024 년도 정부(산업통상자원부)의 재원으로 한국산업기술기획평가원의 지원(No. RS-2024-00406121, 자동차보안취약점기반 위협분석시스템개발(R&D))과 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(No. RS-2022-00166529)을 받고 과기정통부 정보통신기획평가원의 정보보호핵심원천기술개발사업(No. RS-2024-00337414)으로 수행한 결과임.

참고문헌

[1] J. H. Kim et al., "Aquabolt-XL: Samsung HBM2-PIM with in-memory processing for ML accelerators and beyond," 2021 IEEE Hot Chips 33 Symposium (HCS), 2021

[2] 김경모, 이하윤, 신동군, “Processing-in-Memory 를 위한 효율적인 행렬 연산 기법”, 한국소프트웨어종합학술대회 논문집, 2021

[3] Hyeokjun Choe, et al., “Near-Data Processing for Machine Learning”, ICLR, 2017

[4] AMD, “AMD ALVEO™ U280 Adaptable Accelerator Cards for Data Center Workloads”, [online] Available: https://www.amd.com/en/products/accelerators/alveo.html

[5] Kevin Nam et al., “Accelerating N-Bit Operations over TFHE on Commodity CPU-FPGA.” ICCAD, 2022