IDEC Chip Design Contest

Graph Neural Network Accelerator

한국과학기술원 전기 및 전자공학부

최한진, 장성현, 정명수

서론

- 그래프 신경망(Graph neural network, GNN)은 SNS, 추천 시스템 등 다양한 분야에서 사용되고 있는 AI 모델임
- 특히, GNN 기반 추천 시스템 및 SNS는 사용자 입력에 대한 응답을 즉시 보내야 해, 서비스 지연 시간 단축이 중요함

GNN 가속기의 설계

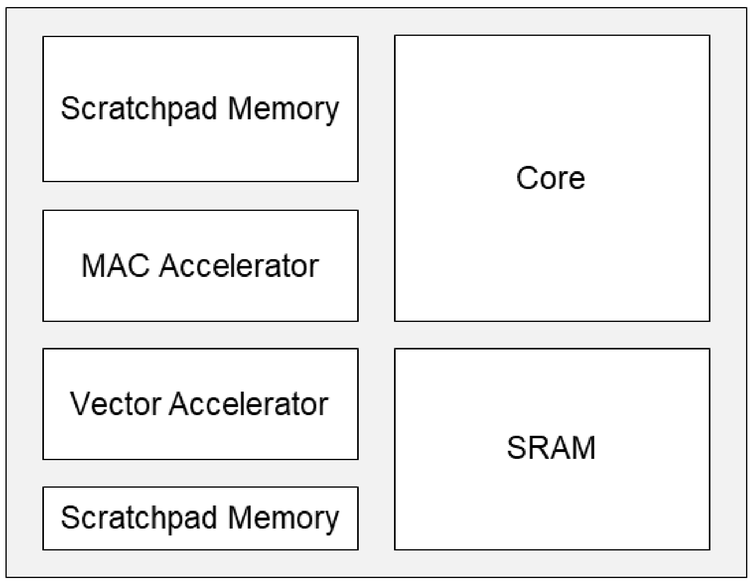

그림 1-(a) SoC 아키텍처

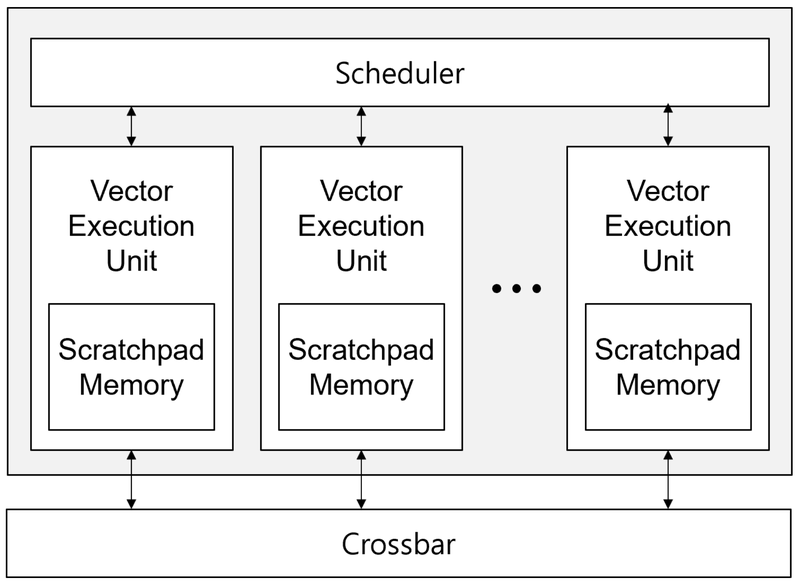

그림 1-(b) 벡터 가속기

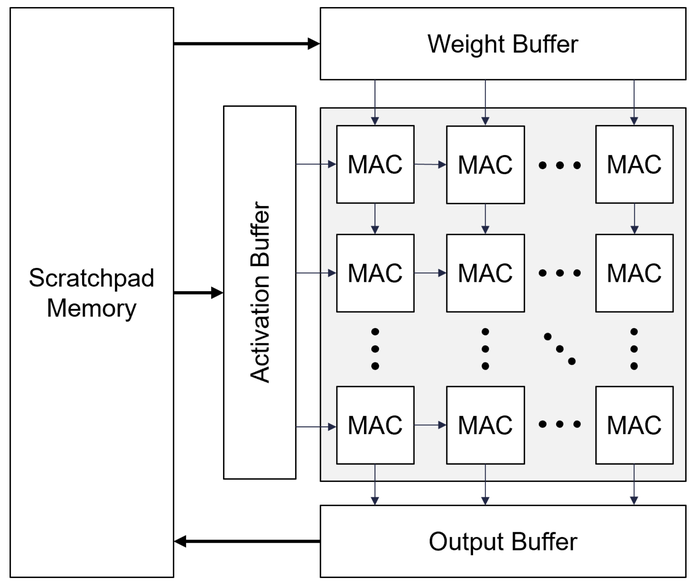

그림 1-(c) 행렬곱 가속기

SoC 아키텍처

- GNN의 핵심 연산은 행렬 곱과 벡터 연산으로 이루어져 있어, GNN 가속기에는 각 연산에 대한 가속기가 모두 필요함
- 따라서, 그림 1-(a)에서 보여지는 GNN 가속기는 행렬 곱 가속기와 벡터 연산 가속기를 포함하도록 설계하였음

벡터 가속기

- 그림 1-(b)벡터 가속기는 두 벡터 간의 연산 또는 벡터 스칼라 연산을 가속하는 모듈임
- 특히, 벡터 연산 유닛 내부 스크래치패드 메모리에 데이터 버퍼링이 가능하여, 연산 시 메모리 접근 지연시간을 최소화 할 수 있음

행렬곱 가속기

- 그림 1-(c)의 행렬곱 가속기는 곱셈 누적 연산기(Multiply and accumulate, MAC)를 활용하여, 행렬 간의 곱셈을 가속하는 모듈임
- 가속기 내 파이프라이닝된 형태로 연결된 곱셈 누적 연산기는, 매 클럭마다 데이터를 주고 받으며 메모리 접근 횟수를 최소화 함

ASIC 설계

- ASIC 설계 결과 칩 스펙은 표 1에서 확인할 수 있으며, 삼성 파운드리 28nm FD-SOI 공정을 활용하여 설계를 진행하였음
- 다양한 시나리오에서의 칩 정상 동작을 위해, FF/SS/TT코너에서 사인오프를 진행함

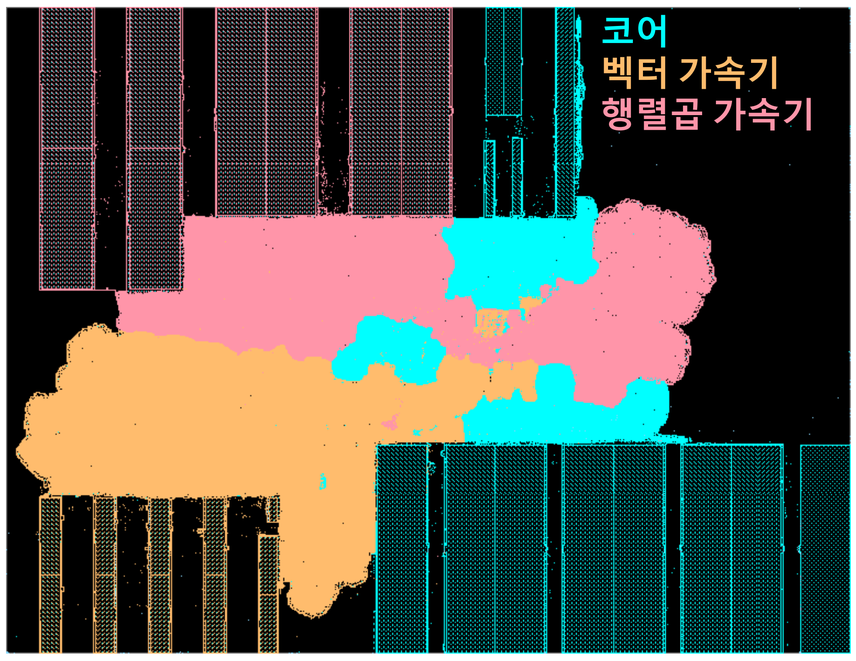

그림 2 레이아웃

결론

- 이전 연구들의 가속기는 하나의 연산만 가속 가능한 반면, 본 연구는 GNN 전 과정 가속을 위해 벡터 가속기와 행렬곱 가속기 모두를 SoC에 적용한 차이가 있음
- 그림 3의 칩 동작 검증 결과를 통해, SoC내 각 모듈의 정상 동작을 확인할 수 있음
- 가속기가 포함된 칩의 CPU 대비 성능은 그림 4와 같음

감사의 글

- 본 연구는 IDEC에서 MPW와 EDA Tool을 지원받아 수행하였습니다

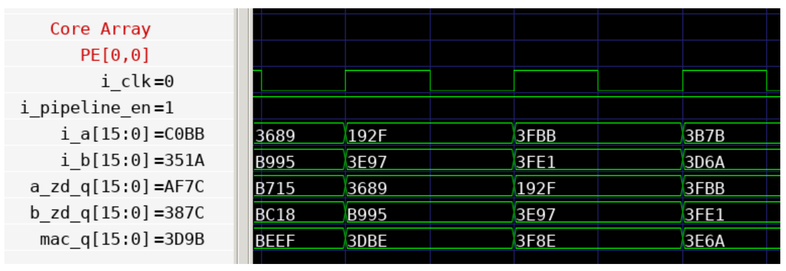

그림 3 칩 동작 검증 결과

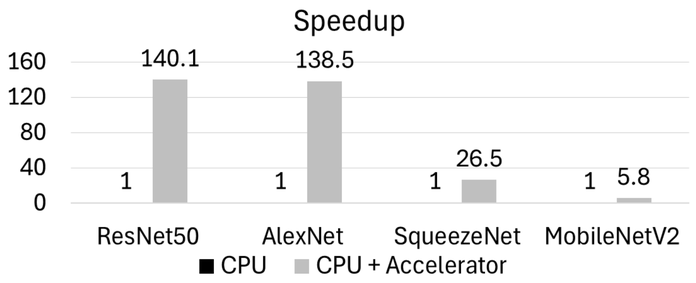

그림 4 칩 예상 성능

Item	Description
Core Area	• $5{,}002{,}092\ \mu m^2$ ($W \times H = 2{,}556\ \mu m \times 1{,}957\ \mu m$)
Supply Voltage	• $1\ V$ core VDD • $1.8\ V$ I/O DVDD
Frequency	• $50\ MHz$
Cell Instance Information	• Total Number of Cell: $1{,}084{,}771$ • Total Cell Area: $3{,}938{,}713\ \mu m^2$

표 1 설계 사양