IDEC Chip Design Contest

2026 IDEC Congress CDC
Performance Evaluation of a Bandwidth-Efficient Systolic Array with Adaptive Block-wise Data Reuse

Youngjun Hwang, Youngsik Kim
Department of Computer and Electronic Engineering,
Handong Global University, Pohang, Korea

Introduction

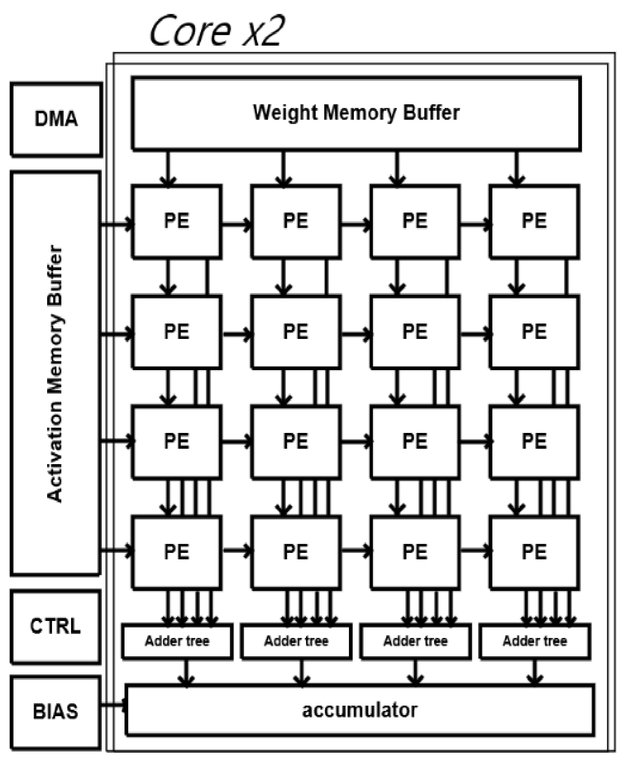
TPU는 AI 추론 연산에서 강력한 처리량을 제공하며, 엣지 및 임베디드 환경에서 유망한 가속기 솔루션으로 주목받고 있다. 그러나 TPU는 가중치 및 활성값 데이터를 외부 메모리에서 반복적으로 가져와야 하는 메모리 대역폭 병목 문제가 존재하며, 이는 전체 시스템 효율을 크게 저하시킨다.

이를 해결하기 위해 본 연구에서는 TPU 메모리 구조에서의 데이터 재사용 방안을 연구하였다. Tile DMA를 통한 블록 단위 활성값, 가중치 재사용으로 외부 메모리 접근 횟수를 줄이고, Ring Buffer 구조를 활용하여 온칩 데이터 지역성을 극대화함으로써 Alex-net의 Conv 레이어에서 메모리 대역폭 요구량을 크게 감소시켰다.

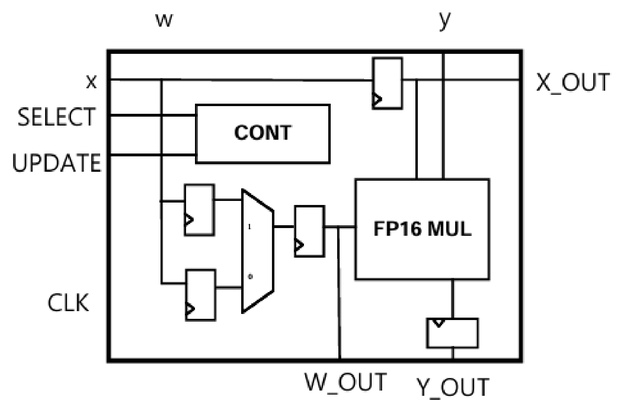
또한 Double Buffering Systolic Architecture를 사용하여 on chip 내부의 가중치 로딩 타임을 줄여 내부 latency를 최적화 하였다.

제안된 아키텍처는 FDSOI 28nm 공정으로 칩 테이프아웃을 진행하였으며, ZCU104 FPGA 플랫폼에서 성능 검증을 수행하였다.

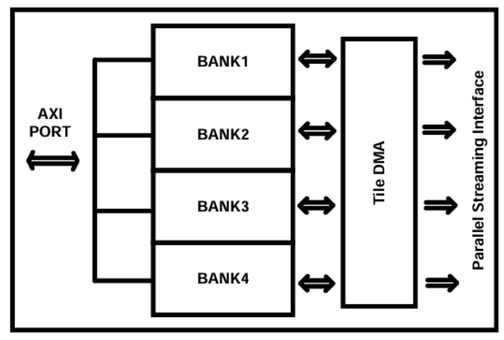
System Architecture



<Block Diagram of overall Architecture>

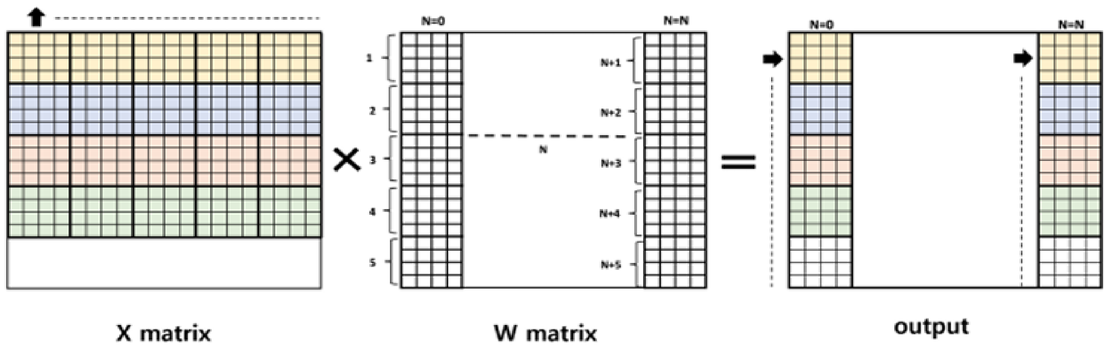


<Systolic Architecture with Double Buffering>

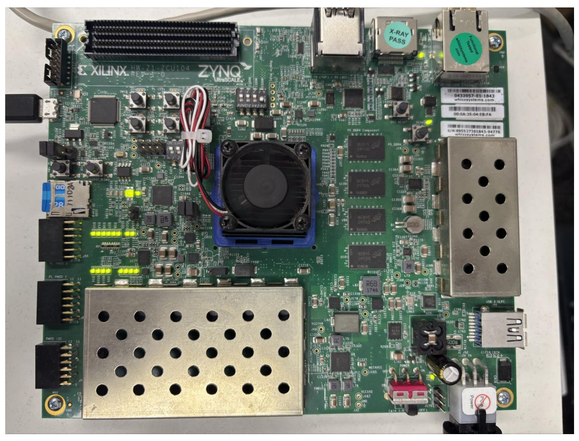


<Ring Buffer memory architecture>

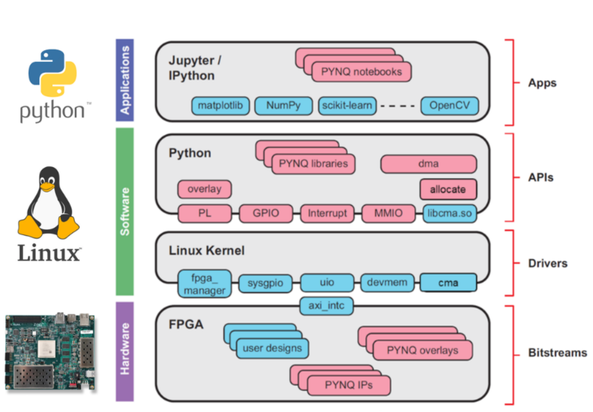
Cali



<Tiled GEMM Processing Method>

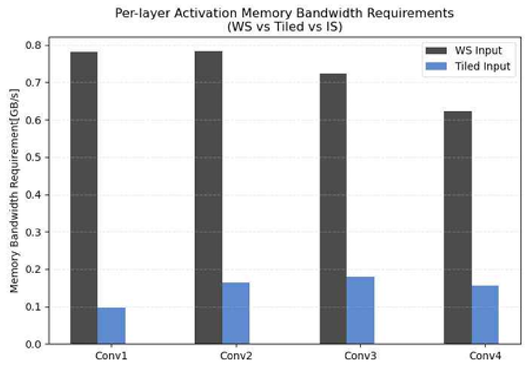


<FPGA: ZCU104>

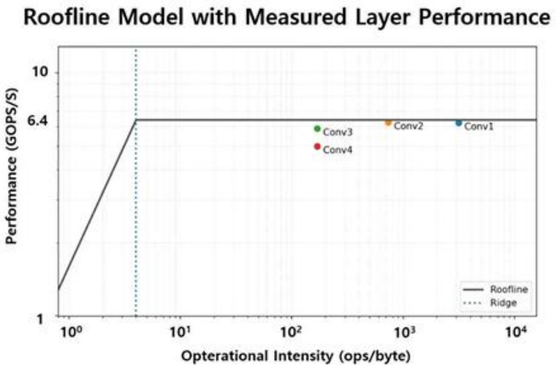


<Test Platform>

Conclusion



< Activation memory bandwidth requirements>



< Roofline for Alex-net Conv layer>

Layer	FPL2009 [8]	ASAP2009 [9]	PACT2010 [10]	ISCA2010 [11]	Our Impl.
Precision	48bits fixed	16bits fixed	Fixed point	48bits fixed	16bits float
Frequency	125 MHz	115MHz	125MHz	200MHz	100MHz
FPGA capacity	23,872 slices 126 DSP	51,840 slices 192 DSP	37,440 slices 1056 DSP	37,440 slices 1056 DSP	31,112 slices 128 DSP
Performance	2.6 GMACS	3.37 GMACS	3.5 GMACS	8 GMACS	3.2GMACS
	5.25 GOPS	6.74GOPS	7.0GOPS	16GOPS	6.4GOPS
Area-Normalized Ops per Cycle	1.76E-03 OPS/Cycle/Slice	1.13E-03 OPS/Cycle/Slice	1.52E-03 OPS/Cycle/Slice	2.15E-03 OPS/Cycle/Slice	2.20E-03 OPS/Cycle/Slice

< Comparison to previous implementation>