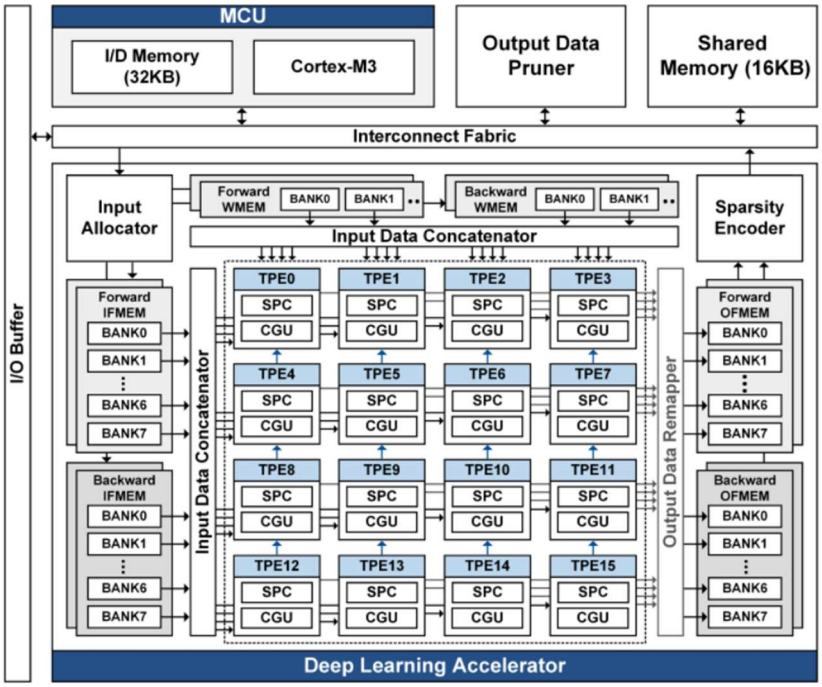
A-SSCC 2024 Review

경북대학교 전자전기공학부 박사과정 박승현

Session 11 AI Accelerators

이번 A-SSCC 2024의 11번째 세션, AI Accelerators에서는 대규모 신경망 처리 및 효율적인 학습 가속화를 위한 다양한 하드웨어 아키텍처와 최적화 기법들이 중점적으로 논의되었다. 세션의 논문들은 공통적으로 데이터 처리 효율성, 전력 및 면적 효율성, 확장성을 높이는 데 초점을 맞췄다. 또한 다양한 AI 모델과 애플리케이션을 지원하기 위한 하드웨어 최적화, 데이터 재사용 전략, 그리고 고효율 연산 구조 설계가 중요하게 다루어졌다.

#11-1

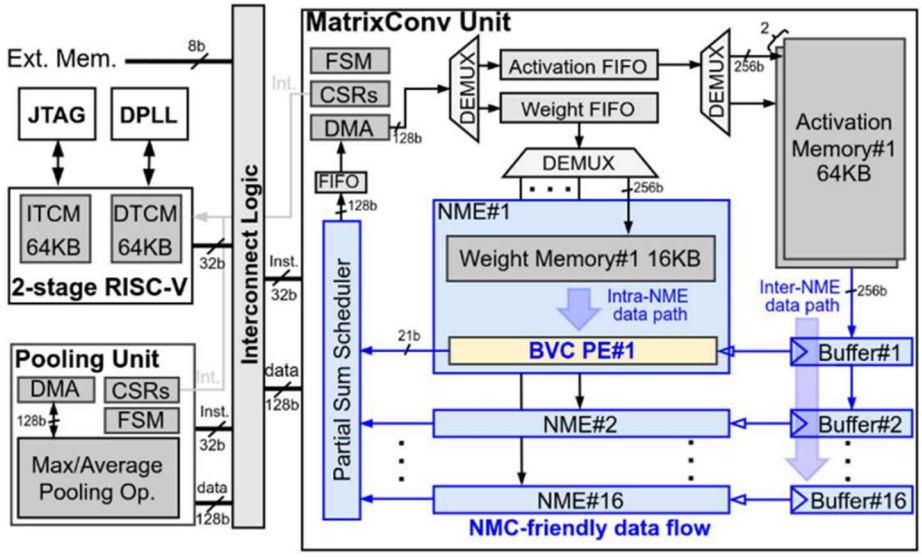


[그림 1] 제안하는 딥러닝 SoC의 시스템 아키텍처

이 논문은 대규모 신경망의 높은 계산 요구 사항을 효과적으로 처리하기 위해 확장 가능하고 재구성 가능한 딥러닝 SoC를 제안한다. 멀티칩 솔루션의 한계로 지적되는 낮은 하드웨어 활용도와 데이터 이동의 비효율성을 해결하기 위해 채널 및 평면 기반 데이터 분할, 양방향 데이터 흐름, 그리고 희소성 데이터를 활용하는 최적화된 아키텍처를 설계하였다. 이를 통해 SoC는 칩 간 데이터 이동량을 최소화하고, 계산 효율성을 극대화하며,

대규모 신경망의 다양한 구조를 지원할 수 있게 되었다. 특히, 제안된 아키텍처는 다차원 하드웨어 병렬성을 적용하여 신경망의 깊은 층에서도 높은 칩 활용도를 유지한다.
SoC의 구조는 딥러닝 가속기, 출력 데이터 필터링 장치, Coretex-M3 MCU로 구성되며, 딥러닝 가속기는 4x4 텐서 프로세싱 요소(TPE) 배열로 구성된다. TPE는 희소성 데이터를 효율적으로 처리하기 위한 희소성 처리 코어(SPC)와 채널 그룹 유닛(CGU)을 포함한다. 또한, TPE 배열은 다차원 병렬성을 고려한 설계로 CNN, GCN, Transformer 모델에서 높은 성능을 발휘한다. 예를 들어, 희소성 데이터의 비율이 75% 이상인 경우 CNN과 Transformer 모델에서 최대 75%의 지연 시간을 줄였으며, GCN에서는 93%의 지연 시간 감소를 달성하였다. 추가적으로, 데이터 분할 및 양방향 데이터 흐름은 칩 간 데이터 충돌을 방지하고, 칩 활용도를 83%까지 향상시켜 16칩 시스템에서 12.6배의 처리량 향상을 달성하였다.
이 SoC는 28nm CMOS 공정으로 제작되었으며, 칩 면적은 7.8mm², 소비 전력은 10.4-273.8mW로, 최고 성능 9.83TOPS를 기록하였다. 면적 효율은 1.26TOPS/mm², 에너지 효율은 141.4TOPS/W로, 기존 멀티칩 시스템 대비 면적에서 5.7배에서 최대 21배, 에너지 효율에서 1.7배에서 45.6배 우수한 성능을 보였다. 이 연구는 고성능 및 고효율의 딥러닝 SoC 구현 가능성을 입증한다.

#11-3



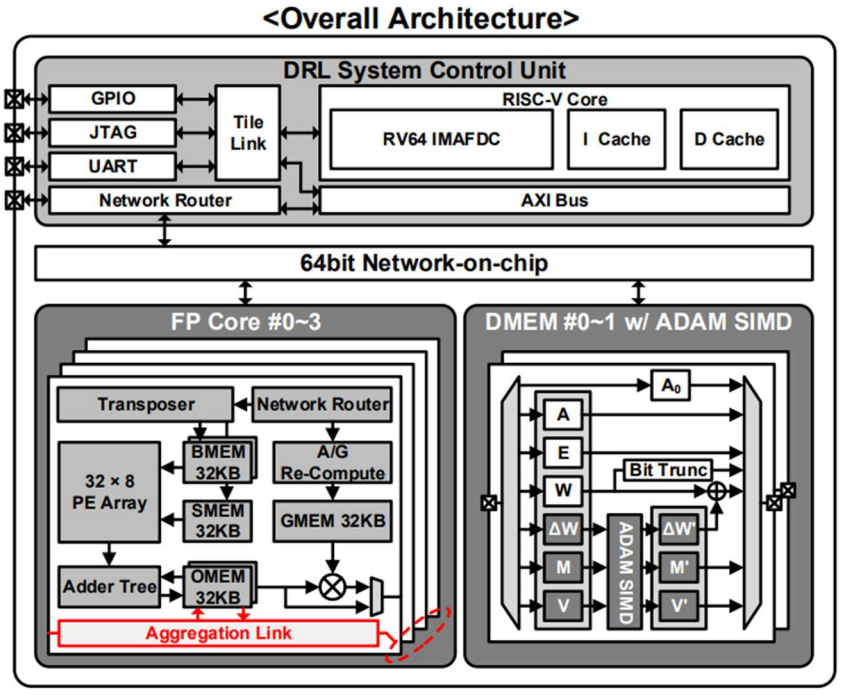
[그림 2] near-memory-computing 데이터플로우와 Booth-value-confined 다해상도 처리 아키텍처

이 논문은 엣지 AI 애플리케이션을 위해 높은 전력 효율과 처리 성능을 제공하는 Booth-value-confined(BVC) 가속기와 근접 메모리 컴퓨팅(NMC) 기반 프로세서를 제안한다. 기존의 von-Neumann 아키텍처와 비교해 NMC 기반 구조는 전력 효율에서 우위를 보였지만, 메모리 불연속성과 비효율적인 데이터 매핑 전략으로 인해 전체적인 지연과

소비 전력이 증가하는 문제가 있었다. 이를 해결하기 위해 본 연구는 BVC 근사 연산기, NMC 친화적인 데이터 흐름, 메모리 공간 연속성을 고려한 신경망 매핑 전략을 도입한다. 프로세서는 RISC-V 서브시스템, 디지털 PLL, 프로그래머블 풀링 및 매트릭스 연산 유닛으로 구성되며, BVC 근사 연산기가 포함된 16개의 NME가 주요 연산을 담당한다.

제안된 BVC 연산기는 Booth 알고리즘을 기반으로 3가지 정밀도(BVC3/6/8)를 지원하며, 전력과 면적을 기존 멀티프리시전 연산기 대비 각각 82%, 70% 절감한다. 이를 통해 각 가중치의 재사용을 극대화하고 데이터 전송량을 줄여 시스템 레벨에서 효율성을 극대화한다. 또한, 가중치와 활성화 데이터를 재사용하며 메모리 공간을 연속적으로 유지하여 메모리 액세스 지연을 최소화한다. VGG16 및 ViT-Tiny 모델에 대한 테스트 결과, 평균 전력 효율은 10.14-15.11 TOPS/W, 피크 효율은 12.92-29.11 TOPS/W를 기록하였으며, 메모리 지연을 최대 31.4% 감소시켰다. 22nm 공정으로 제작된 이 프로세서는 13.61-210.67mW의 전력에서 동작하며, Activation(INT4)과 Weight(BVC3)를 조합한 경우 최고 33.98 TOPS/W의 전력 효율을 달성하였다. 본 연구는 높은 전력 효율과 데이터 재사용성을 통해 엣지 AI 프로세서 설계에 있어 새로운 최적화 방향을 제시하였다.

#11-4



[그림 3] EnTADRL의 아키텍처

이 논문은 연속적이고 효율적인 심층 강화 학습(DRL) 가속화를 위한 SoC(System-on-Chip)인 EnTADRL을 제안한다. DRL은 게임 AI, 자율주행 등과 같은 시퀀셜 의사결정 문제

에서 인간 수준의 성능을 달성하지만, 복잡한 연산 구조와 높은 메모리 대역폭 요구 사항으로 인해 기존 AI 가속기에 직접 매핑하기 어렵다. 이를 해결하기 위해 EnTADRL은 세 가지 주요 최적화를 제안한다: 사용자 정의 명령어 세트 아키텍처(ISA), 네트워크-온-칩(NoC) 및 스트리밍 데이터플로우를 활용한 SoC 수준 최적화, 활성화 기울기 재계산 및 Tanh 양자화를 포함하는 부동소수점(FP) 코어 최적화, 그리고 파이프라인화된 ADAM SIMD를 활용한 효율적인 가중치 업데이트 지원이다.

제안된 EnTADRL은 RISC-V 기반의 DRL 시스템 제어 유닛, 4개의 FP 코어, 그리고 2개의 데이터 메모리(DMEM)로 구성된다. FP 코어는 활성화 기울기 재계산 유닛과 32×8 2D PE 어레이를 통합하여 메모리 발자국을 14.3% 감소시키고, 계산 효율성을 높인다. 또한, ADAM SIMD는 전용 메모리 구조와 파이프라인화된 연산 흐름을 통해 높은 메모리 대역폭을 지원하며, 지연을 최소화한다. 스트리밍 데이터플로우와 이중 버퍼링 기법은 메모리 전송과 계산 사이의 병목현상을 완화하며, 전체 연산 주기를 23.2% 줄인다.

EnTADRL은 28nm CMOS 공정으로 제작되어 12.96mm²의 다이 면적과 2.8MB의 온칩 SRAM을 포함한다. 10MHz에서 200MHz의 동작 주파수와 0.65V에서 0.9V의 전압에서 작동하며, 최고 409.6 GFLOPS의 성능과 2.4 TFLOPS/W의 에너지 효율을 달성한다. 특히, DRL 가속기를 위한 기존 솔루션과 비교하여, EnTADRL은 손실 계산, 노이즈 삽입, ADAM 연산을 포함한 복잡한 DRL 연산을 온칩에서 완벽히 처리하여 실질적인 종단 간 가속화를 지원한다. 이 연구는 DRL을 위한 새로운 하드웨어 아키텍처의 가능성을 제시하며, 실제 환경에서의 DRL 성능 최적화를 크게 향상시킨다.

저자정보

박승현 박사과정 대학원생

- 소속 : 경북대학교
- 연구분야 : 딥러닝 가속기 설계
- 이메일 : ijjh0435@gmail.com
- 홈페이지 : https://ai-soc.github.io/

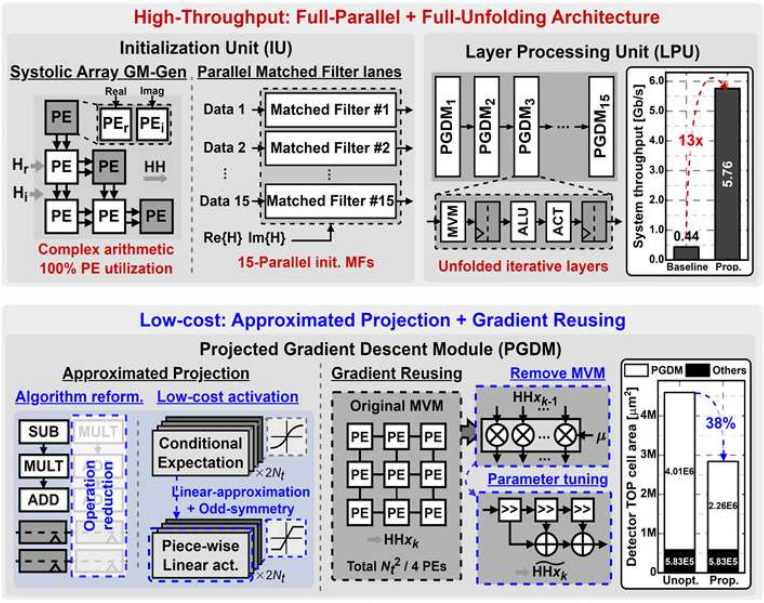
A-SSCC 2024 Review

한국과학기술원 전기및전자공학부 석사과정 권재훈

Session 27 Security/Signal Processing Systems

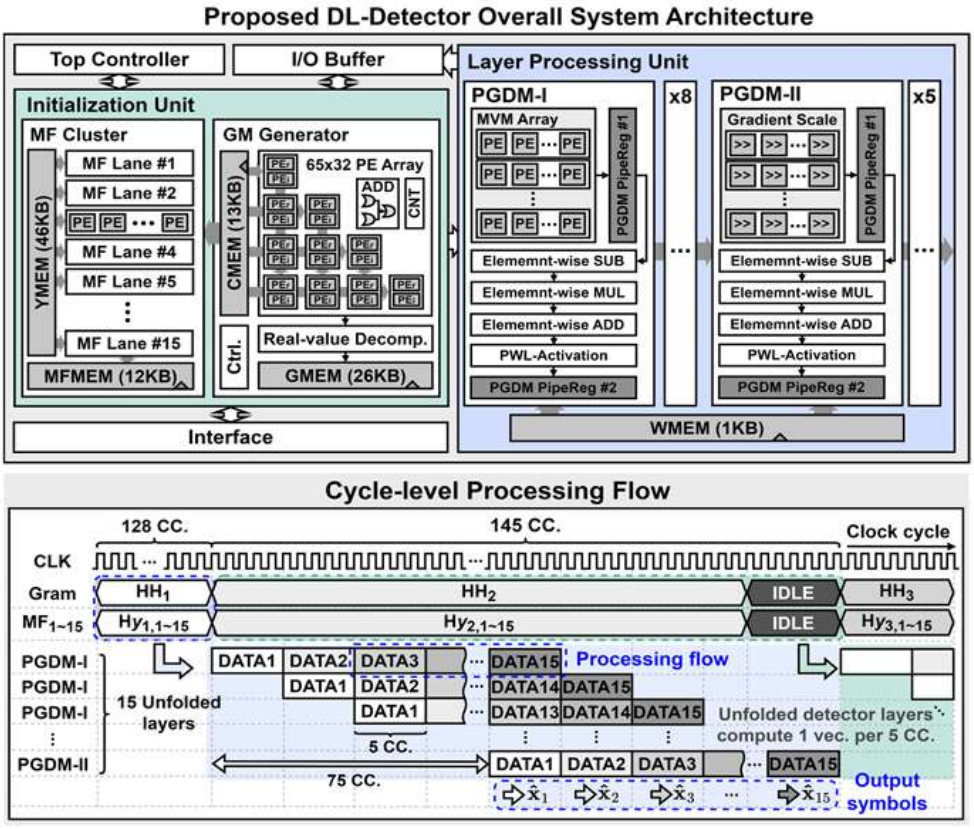
이번 2024 IEEE A-SSCC의 Session 27 Security/Signal Processing Systems에서는 SRAM 기반 Physically Unclonable Function (PUF) 설계, SOIPUF(Symmetrical Obfuscated Interconnection PUF) 기반 Hybrid Strong PUF 설계 및 Resistance 향상, DL 기반 efficient MIMO Detector 설계, HBM3와 NPU 기반의 Chiplet architecture를 위한 효율적인 Digital PHY 설계라는 주제로 총 4편의 논문이 발표되었다. 이 중 DL 기반 efficient MIMO Detector 설계, HBM3와 NPU 기반의 Chiplet architecture를 위한 효율적인 Digital PHY 설계에 대한 2개의 논문을 살펴보고자 한다.

#27-3 본 논문은 POSTECH에서 발표한 논문으로, Deel Learning (DL) 기반의 Massive multiple-input multiple-output (MIMO) uplink detector를 최초로 silicon으로 구현하였으며, 5.76 Gb/s의 throughput과 79.7 pJ/b의 energy efficiency을 달성했다. 128×32 MIMO 환경에서 4~256QAM modulation을 지원하며, 기존의 minimum mean square error (MMSE) 방식에 비해 4.3 dB의 성능 향상을 이뤄냈다. [그림 1]은 제안된 DL-detector의 high-throughput, low-cost를 위한 최적화 방법을 나타낸 그림이다.



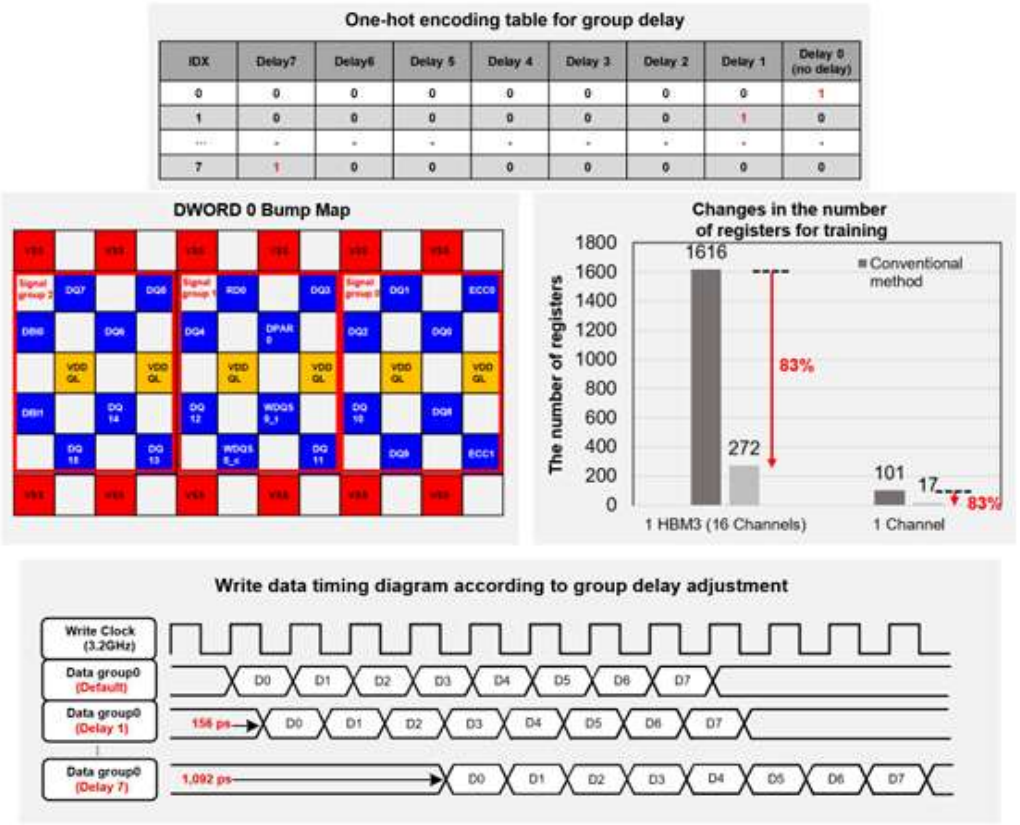
[그림 1] Overview of the proposed high-throughput and low-cost optimization strategies of DL-detector.

Contribution은 1번째로 Fully Unfolded DL-based MIMO Detector를 설계하여 병렬성을 높인 것이다. Unfolding technique은 loop를 제거하고 병렬화하는 설계 기술이고 이를 통해 throughput을 증가시킬 수 있다. 이에 대한 자세한 내용은 textbook [1]에서 참고할 수 있다. 기존 직렬 방식의 DL 연산과 달리, 모든 DL layer의 연산을 병렬로 실행하여 13배의 throughput 향상을 이끌어냈다. Deep Learning에 대한 자세한 설명은 textbook [2]에서 참고할 수 있다. 2번째로 Initialization을 빠르게 할 수 있도록 한 것이다. 본 논문의 초기화 단계에서는 fully parallel MF와 gram-matrix generator를 사용해 128 cycles 내에 초기화를 완료한다. 이때 gram-matrix는 Symmetric Property를 활용하여 half-sized systolic array와 메모리만으로 계산할 수 있다. 또한 15 parallel MF lane을 통해 연속적인 입력 데이터를 병렬로 처리한다. 마지막으로 Approximate Projection과 Gradient Reusing이다. Approximate Projection은 기존의 projection 연산을 최적화하여, linear connection 연산의 50% 감소를 달성했다. 또한 ReLU 기반 nonlinear unit을 단순화하여 divisor와 exponent 연산자를 제거했다. Gradient Reusing은 마지막 6개의 DL layer에서는 Gradient Reusing을 적용하여 새로운 gradient를 생성하는 것이다. 이를 통해 MVM 연산을 완전히 제거함으로써, 전체 연산 cost의 38% reduction을 달성했다. [그림 2]는 전체적인 chip architecture와 cycle-level로 processing flow를 나타낸 그림이다.



[그림 2] Overall chip architecture and cycle-level processing flow.

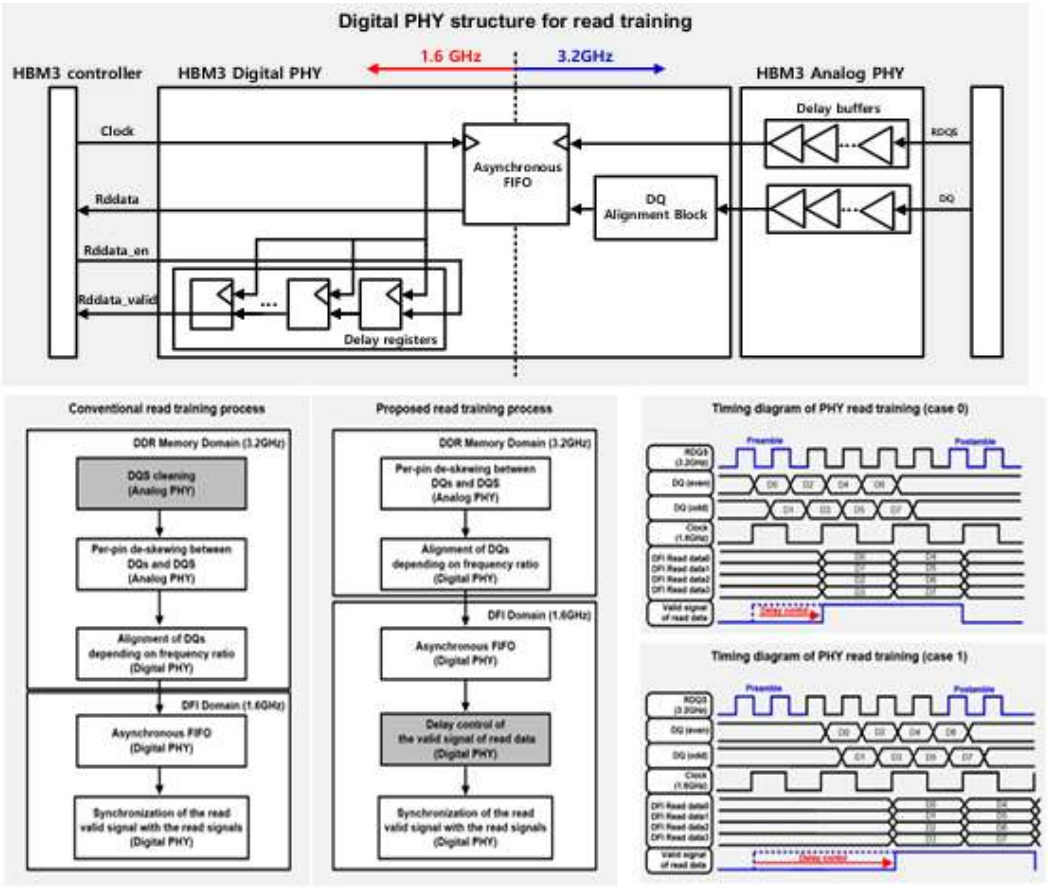
#27-4 본 논문은 AI SoC Research Division, Electronics and Telecommunication Research Institute (ETRI)에서 발표한 논문으로, HBM3 Digital PHY를 chiplet 기반 AI processor에 최적화한 설계에 대한 연구이다. 12nm CMOS 공정을 사용하였고 6.4Gb/s/pin의 데이터 전송 속도를 달성했다. 기존의 HBM2E와 비교하면, HBM3는 clock frequency가 1.2GHz에서 3.2GHz로 3배 향상되었는데, 이러한 고속 동작을 지원하기 위해 고속, 저전력의 Digital PHY 설계가 필요하여 본 논문에서 최적화를 제시한 것이다. chiplet architecture에 대한 내용은 textbook [3]에서 참고할 수 있다. 또한 textbook [4]에서 PHY 설계의 개념과, timing 문제 및 설계 기법 등을 참고할 수 있다.



[그림 3] Grouping signal delays based on HBM3 bump map.

Contribution은 1번째로 group delay 방식을 개선한 것이다. 기존에는 data, command, address 신호에 대해 개별적으로 delay를 조정했지만, 본 논문에서는 HBM3 bump map을 기반으로 grouped signal 단위로 delay를 조절하는 방식을 제안했다. 이를 통해 delay register 수를 83% 줄였고, PHY training time도 단축시킬 수 있었다. 이에 대한 figure는 [그림 3]을 통해 제시되었다. 2번째로 Async FIFO 구조를 단순화한 것이다. 기존의 async FIFO는 full flag로 상태를 모니터링해야 했지만 본 논문에서는 full flag 없이도 동작 가능한 async FIFO 구조를 제안했다. FIFO의 data depth를 8로 설정하여, full state activation을

방지했다. 또한 루프백 카운터를 latency가 적은 lookup table(LUT)로 대체해 고속 신호 동작에 적합한 구조를 설계했다. 3번째로 새로운 Read Training 알고리즘을 제안한 것이다. 기존의 RDQS cleaning을 없애고, FIFO의 clock 신호로 RDQS를 직접 연결했다. 또한 high-frequency domain(3.2GHz) 대신 low-frequency domain (1.6GHz)에서 delay 제어를 수행해 전력 소모를 줄이고 알고리즘 복잡도를 낮췄다. 이에 대한 figure는 [그림 4]에서 볼 수 있다.



[그림 4] Proposed HBM3 PHY structure for read training.

참고문헌

[1] Parhi, Keshab K. VLSI digital signal processing systems: design and implementation. John Wiley & Sons, 2007.

[2] Goodfellow, Ian. "Deep learning-ian goodfellow, yoshua bengio, aaron courville." Adapt. Comput. Mach. Learn (2016).

[3] Manna, Kanchan, and Jimson Mathew. Design and Test Strategies for 2D/3D Integration for NoC-

based Multicore Architectures. Springer Nature, 2019.

[4] Khan, Shoab Ahmed. Digital design of signal processing systems: a practical approach. John Wiley & Sons, 2011.

저자정보

권재훈 석사과정 대학원생

- 소속 : 한국과학기술원 전기및전자공학부
- 연구분야 : Digital Circuit Design, ECC Hardware Design
- 이메일 : jhkwon@ics.kaist.ac.kr
- 홈페이지 : https://ics.kaist.ac.kr/