

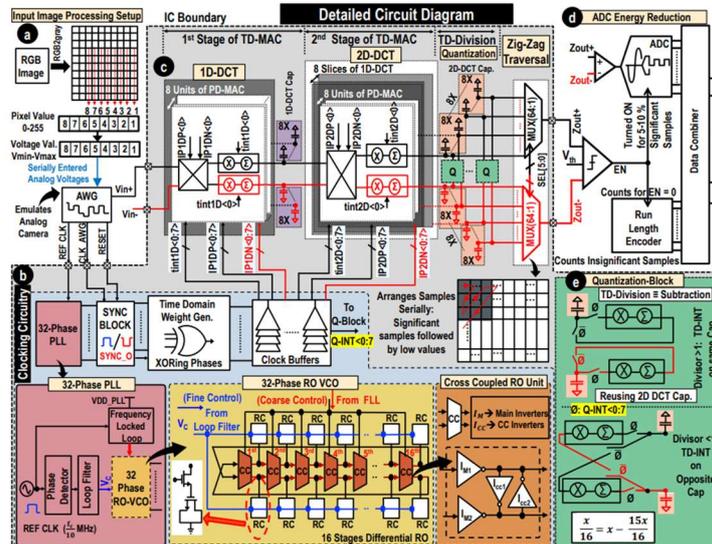
2025 IEEE CICC Review

경북대학교 전자전기공학부 박사과정 박승현

Session 17 Next-Generation Systems: From Datacenters to the Edge

CICC 2025 세션 17의 논문들은 각기 다른 시스템들의 성능을 저전력 설계 관점에서 창의적으로 향상시키는 방법을 제시한다. 첫 번째 논문은 아날로그 도메인에서 직접 JPEG 압축을 수행해 디지털화 이전의 연산으로 전력 소모를 획기적으로 줄였고, 두 번째 논문은 강화학습 기반 자율주행에 특화된 프로세서를 설계해 실시간 제어와 효율적인 학습을 동시에 가능하게 했다. 세 번째 논문은 데이터센터 고속 통신에 필요한 FEC 조건을 정의하고 이를 만족하는 BCH 디코더를 구현해 고성능과 에너지 효율을 모두 확보했다. 각각의 접근 방식이 문제의 본질을 정확히 짚고 실용적인 해법을 제시하고 있어 인상 깊었다.

#17-1 TD-dAJC: A 2pJ/pixel Time-Domain Weight and Integrating-MAC based direct-Analog-to-MJPEG Compression for Video Sensor Nodes



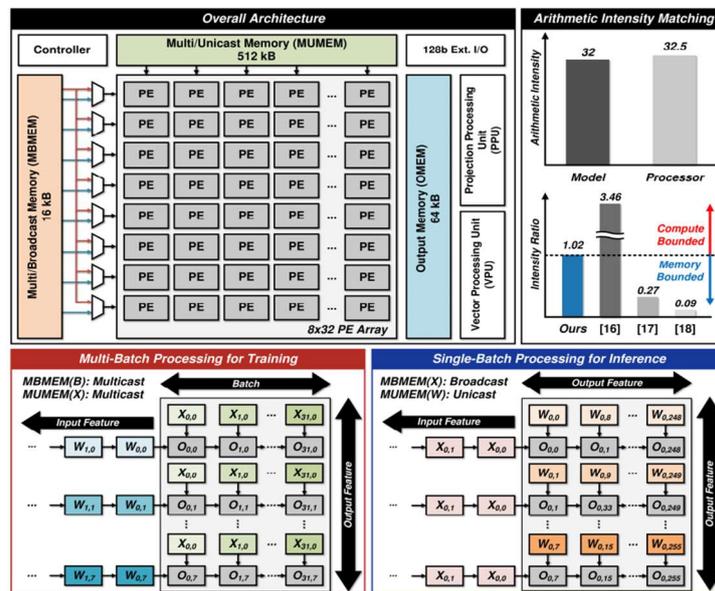
[그림 1] 제안하는 아날로그 기반 MJPEG Compression 회로 다이어그램

본 논문은 초저전력 IoT 영상 센서 노드를 위한 새로운 아날로그 기반 JPEG 압축 구조를 제안한다. 기존의 디지털 카메라 시스템은 영상 데이터를 완전히 디지털화한 후 압축을 수행하는 방식으로, 높은 ADC 전력 소모, 중간 저장소의 면적/전력 부담, 복잡한 연산 과정 등의 문제를 안고 있다. 특히 옛지 디바이스에서는 이러한 전력과 자원의 비효율성이 심각한 제약으로 작용하며, 실시간 처리가 요구되는 상황에서 MJPEG과 같은 압축 방식도 여전히 높은 연산 복잡도와 저장소 부담이 존재한다. 저자는 이러한 문제를

해결하기 위해, DCT 기반의 JPEG 압축을 디지털화 이전 단계, 즉 센서 출력 직후의 아날로그 도메인에서 직접 수행하는 TD-dAJC(Time-Domain direct Analog-to-JPEG Compression) 구조를 고안하였다.

제안된 시스템은 Time-Domain 기반의 Multiply-and-Accumulate(MAC) 연산 및 Division 구조를 활용하여, 영상 데이터를 디지털화 없이 압축 가능하게 한다. DCT 가중치는 시간 펄스로 변환되어 공정/전압/온도(PVT) 변화에 강인하게 동작하며, 기존 SC 방식에서 필요하던 큰 커패시턴스나 버퍼 없이 소형 저전력 구현이 가능하다. 또한, 선택적으로 유의미한 신호에만 ADC를 작동시키는 구조를 통해 ADC 전력 소모를 최소화한다. 구현된 시스템은 4K 12fps 해상도를 지원하며, 기존 디지털 방식 대비 25배, 기존 아날로그 방식 대비 13배 향상된 2pJ/pixel의 에너지 효율을 달성하였다. PSNR 30dB 수준의 영상 품질도 확보하였으며, 0.856mm²의 소형 칩 영역 내에 모든 연산을 집적함으로써 초소형 엣지 비전 시스템에 실질적인 활용 가능성을 제시하였다. 본 논문은 아날로그 컴퓨팅과 시간영역 신호 처리를 융합하여, 센서 노드의 전력 병목을 근본적으로 해결할 수 있는 새로운 영상 압축 패러다임을 제시한다는 점에서 큰 의미가 있다.

#17-2 A 28-nm Real-Time Reinforcement Learning Processor for Mapless Autonomous Navigation with Unified Actor-Critic Network and Inference-on-Request Scheduling



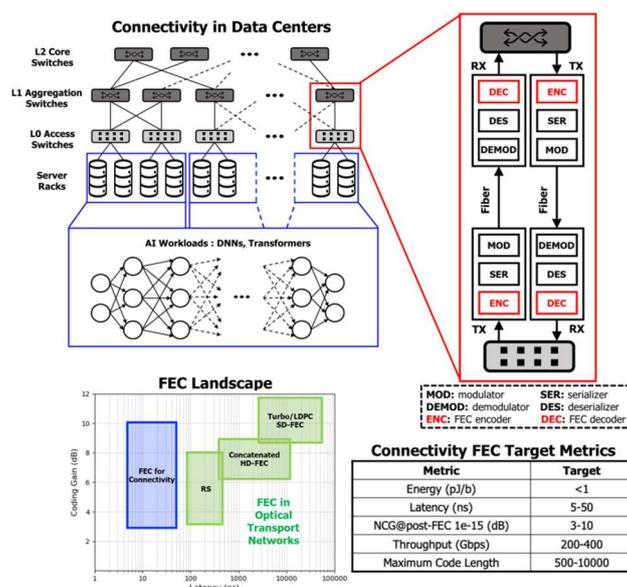
[그림 2] 제안한 강화학습 프로세서의 전체 아키텍처

본 논문은 지도 없이 자율 주행을 수행하는 로봇을 위한 실시간 강화학습 전용 프로세서를 제안한다. 지도 기반의 자율 주행이 어려운 환경에서는 로컬 센서 정보와 목표 위치만을 기반으로 하는 mapless navigation 기술이 요구되며, 이때 강화학습(Deep Reinforcement Learning, DRL)은 환경 적응력 측면에서 효과적인 접근법이다. 그러나 실제

로봇에 DRL 알고리즘을 적용하는 데에는 연산량, 외부 메모리 접근, 에너지 소모 등의 문제가 존재한다. 본 논문은 이러한 문제를 해결하기 위해, DRL 알고리즘 중 Distributional RL(D4PG)를 지원하며, 학습과 추론을 효율적으로 통합 수행할 수 있는 전용 하드웨어를 설계하였다.

제안된 프로세서는 Actor-Critic 네트워크를 통합한 구조, Inference-on-Request 스케줄링, 배치 크기 적응형 데이터 흐름, Zero-skipping 기반 범주 투사(Categorical Projection) 등을 통해 최적화되어 있다. Actor와 Critic의 구조적 유사성을 활용하여 대부분의 신경망 계층을 공유함으로써 연산량과 EMA를 약 85% 줄였으며, 실시간 제어를 위해 inference는 센서 입력 시마다 요청되고, training은 백그라운드에서 유휴 시간에 수행되는 구조로 설계되었다. 이 Inference-on-Request 스케줄링은 하드웨어 유휴 시간을 최소화하고, 필요한 동작 주파수를 86.4%까지 낮추는 데 기여한다. 8x32 PE 배열 기반의 연산 구조는 네트워크와 아키텍처 간 산술 강도를 일치시켜 병목 없이 연산을 수행하며, 실제 로봇에 적용하여 지도 없이 장애물을 회피하며 주행하는 데 성공하였다. 28nm 공정으로 제작된 칩은 2.68mW의 저전력으로 동작하며, 이는 이전 강화학습 전용 프로세서 대비 최고 수준의 PE 활용도(71%)와 EMA 절감 효과(84.1%)를 달성하였다. 본 논문은 실시간 자율 주행을 위한 에너지 효율적이며 학습까지 가능한 강화학습 프로세서의 새로운 기준을 제시한 논문이라 할 수 있다.

#17-3 Forward Error Correction Requirements for Data Center Connectivity



[그림 3] 데이터 센터 연결을 위한 FEC 시스템

이 논문은 차세대 AI 모델 학습을 위한 데이터센터 간 고속 연결에서 요구되는 FEC(Forward Error Correction) 조건을 체계적으로 분석하고, 이에 적합한 설계로 16nm FinFET 공정 기반의 BCH(255,207) 디코더를 구현한 연구이다. 최근 데이터센터는 수천 개의 GPU가 동시에 협업하여 학습을 수행하는 대규모 분산 플랫폼으로 진화하고 있으며, 이러한 시스템에서는 높은 데이터 전송률, 낮은 지연 시간, 그리고 에너지 효율적인 통신 인프라가 필수적이다. FEC는 아날로그 프런트엔드의 SNR 요구를 완화시켜 수백 Gbps급 전송 속도를 가능하게 하는 핵심 기술로 자리잡고 있지만, 데이터센터 연결을 위한 정확한 FEC 요구 조건은 충분히 규명되지 않았다. 이 논문은 이를 해결하기 위해 1) FEC 설계 사양 정의, 2) BCH 코드의 우수성에 대한 가설 제시, 3) 실제 BCH 디코더 설계를 통한 검증을 수행하였다.

제안된 BCH(255,207,6) 디코더는 0.29 pJ/b의 에너지 효율, 14.4ns의 짧은 지연 시간, 30.8 Gbps의 처리 속도, 그리고 1.04 GHz에서 동작 가능한 고성능 디코더로, 데이터센터 단거리(<100m) 광링크에 필요한 3~10dB의 코딩 게인, sub-50ns 지연, sub-pJ/b 에너지 요구 조건을 모두 만족한다. 설계는 RiBM 기반 Berlekamp-Massey 알고리즘을 활용하여 복잡도를 낮췄고, odd iteration을 생략함으로써 사이클 수를 46% 절감했다. 또한 51개의 병렬 처리 엔진을 활용해 고속 처리를 가능하게 하였으며, 에러 탐지와 성능 검증을 위한 내장 테스트 회로도 포함하였다. 결과적으로 본 디코더는 LDPC나 Polar, GRAND 기반 구조보다 훨씬 높은 영역 효율(1922 Gbps/mm²)과 지속적이고 일정한 처리 시간을 제공하며, 데이터센터 환경에서 FEC의 기준점으로서 매우 적합한 설계임을 입증하였다. 이 연구는 고속 네트워크에서의 FEC 설계가 단순히 부가 기능이 아닌 핵심 인프라의 성능을 결정짓는 요소임을 강조한다는 점에서 높은 의의가 있다.

저자정보



박승현 박사과정 대학원생

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

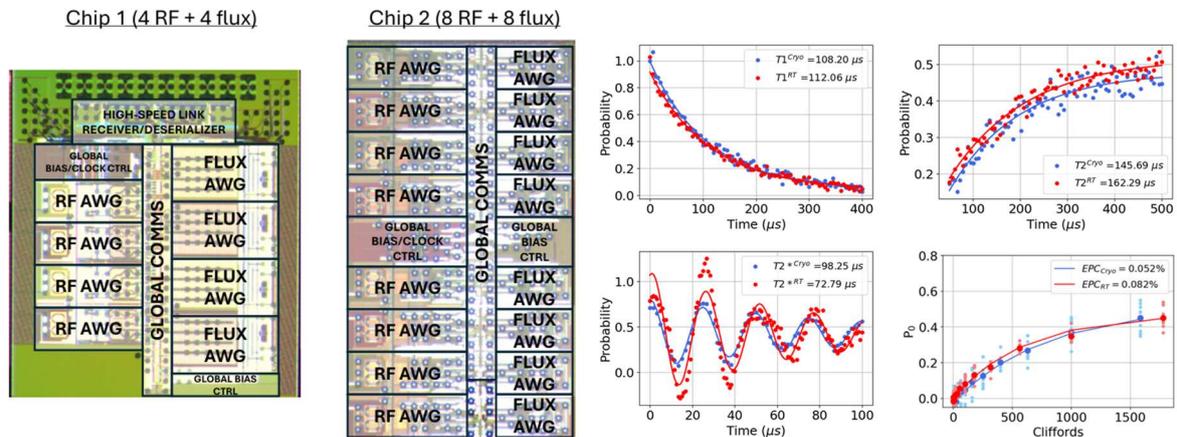
2025 IEEE CICC Review

포항공과대학교 반도체대학원 박사과정 박은빈

Session 28 Next-Generation Systems: Hardware for Quantum and Secure Computing

이번 CICC 2025의 Session 28에서는 양자 컴퓨팅과 보안 컴퓨팅을 위한 차세대 시스템 아키텍처를 주제로 총 4편의 논문이 발표되었다. 이 세션은 물리적 복제 방지(PUF), 측면 채널 공격(SCA) 방지 회로, 보안 해시 가속기, 그리고 초전도 논리 회로 등 다양한 보안 및 양자 하드웨어 기술을 다루며, 고신뢰성, 고속 연산, 회로 규모 효율을 동시에 달성하려는 최신 연구 흐름을 반영하였다

#28-1 본 논문은 UC 버클리와 Google Quantum AI 팀이 공동으로 발표한 것으로, 초전도 큐비트 제어를 위한 Cryo-CMOS 기반 다채널 제어 및 측정 회로를 제안한다. 기존 양자 컴퓨터는 수천 개의 큐비트를 제어하기 위해 다수의 고성능 AWG와 ADC 장비를 사용해야 하며, 이는 상온 장비에서 저온 냉각 시스템까지의 케이블 연결에 따른 복잡한 인터페이스, 전력 소모, 지연 문제를 초래한다.



[그림 1] 제안된 RF-AWG two multi-channel 칩 및 qubit 측정 결과

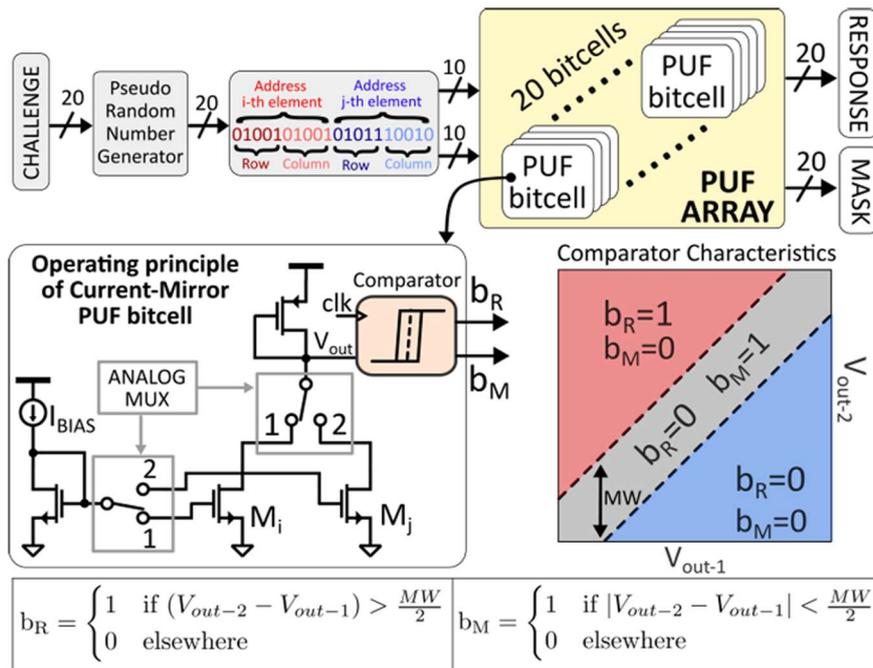
이를 해결하기 위해 본 논문은 1.6K에서 동작 가능한 CMOS 기반 양자 제어 칩을 설계하고, cryogenic 환경에서 동작하는 다채널 송수신 회로를 집적화하였다. 주요 기술로는 다음이 있다. 첫째, 4채널, 1GS/s waveform generator는 9비트 분해능의 비동기 델타-시그마 DAC 구조를 채택하여 고주파 신호 생성 시 low-power 및 low-area를 동시에 만족시킨다. 둘째, 3채널 측정 프론트엔드 회로는 1.2GS/s의 샘플링 속도를 가지는 low-noise ADC와 연동되며, in-phase/quadrature 신호를 안정적으로 분리·측정하는 데 최적화되어

있다. 셋째, 스위치 매트릭스와 on-chip biasing 회로를 통해 여러 큐비트를 선택적으로 제어·측정할 수 있는 확장성을 갖춘 구조를 제공한다.

이 회로는 GlobalFoundries 22nm FDSOI 공정으로 제작되었으며, 총 9.2mm² 면적 내에 전체 회로를 집적하였다. 냉각 환경에서의 실험 결과, 전체 시스템은 15.2mW의 전력으로 안정적으로 동작하였으며, 실제 초전도 큐비트를 사용한 실험에서도 정확한 파형 출력 및 측정 결과 재현이 가능함을 입증하였다.

본 연구는 cryogenic 환경에서 동작하는 제어 회로의 소형화, 저전력화, 고집적화라는 핵심 요구를 만족시키면서, 향후 수천~수만 큐비트 스케일의 양자 시스템 구현을 위한 핵심 하드웨어 플랫폼을 제시한 점에서 중요한 의미를 지닌다.

#28-2 본 논문은 KAIST와 삼성전자에서 공동으로 발표한 것으로, 초소형 면적과 높은 엔트로피를 동시에 달성하는 아날로그 기반 PUF(Physical Unclonable Function) 구조를 제안한다. 기존 디지털 PUF는 회로 복잡도 및 낮은 랜덤성으로 인해 보안성과 재현성에서 한계가 있으며, 아날로그 PUF는 민감한 전류/전압 편차를 기반으로 높은 엔트로피를 구현할 수 있으나, 공정/온도/전압(PVT) 변화에 취약하다는 문제가 있었다.



[그림 2] 제안된 Physical Unclonable Function 구조 및 동작원리

이를 해결하기 위해 본 논문은 sub-threshold 전류 미러 기반 아날로그 PUF 셀을 설계하고, 여기에 Bit Masking 및 Digital Post-Processing 기법을 결합하여 환경 변화에 강건한 비트 출력을 생성하였다. 특히, 인접한 셀 간 상대적 전류 크기를 비교하여 비트를 생성하는 구조를 채택해 절대 전류값의 PVT 민감도를 효과적으로 제거하였다. 이 구조는 standard 65nm CMOS 공정 기반에서 구현되었으며, 단일 셀 기준 166 F²/bit의 면적 효

동 분석(DPA)에 강인한 구조를 형성한다. 셋째, 각 단계마다 난수 마스크를 실시간으로 재생성하여 고정 마스크로 인한 공격 가능성을 제거하였다.

실험 결과, 해당 가속기는 1.7Gbps의 처리 속도와 6.1pJ/bit의 에너지 효율을 달성하며, 성능 저하 없이 마스킹을 적용한 구조임에도 불구하고 비마스킹 가속기 대비 약 73배 이상의 SCA 저항성을 실측 기반으로 입증하였다. 이는 TVLA(Test Vector Leakage Assessment) 기준을 모두 통과하였으며, 실제 프로빙 장비를 이용한 고전압 EM 분석에서도 정보 누출이 관찰되지 않았다.

본 연구는 HMAC-SHA256이라는 널리 사용되는 보안 알고리즘에 대해 ASIC 수준에서의 정량적 SCA 저항성 확보를 실현했다는 점에서, IoT 및 엣지 컴퓨팅 기기에서 요구되는 경량, 고신뢰 보안 하드웨어 설계의 실용적 이정표를 제시한다.

저자정보



박은빈 박사과정 대학원생

- 소속 : 포항공과대학교
- 연구분야 : 딥러닝 가속기 설계
- 이메일 : eunbin@postech.ac.kr
- 홈페이지 : <https://sites.google.com/view/epiclab>

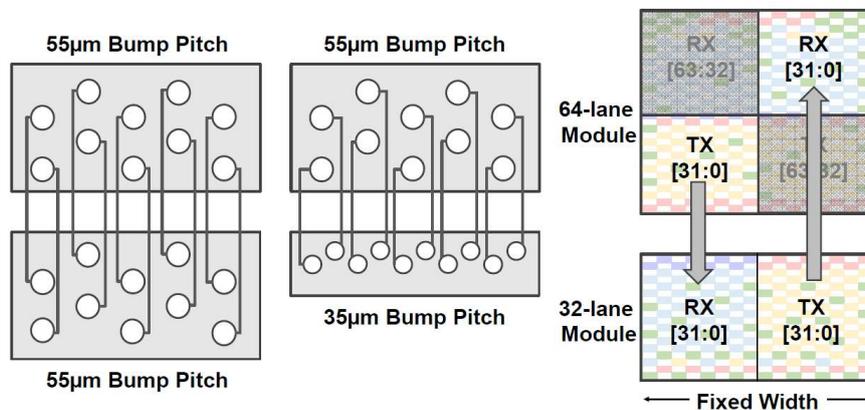
2025 IEEE CICC Review

KAIST 전기및전자공학부 박사과정 엄소연

Session 33 Advancing System Designs with Chiplet Technology

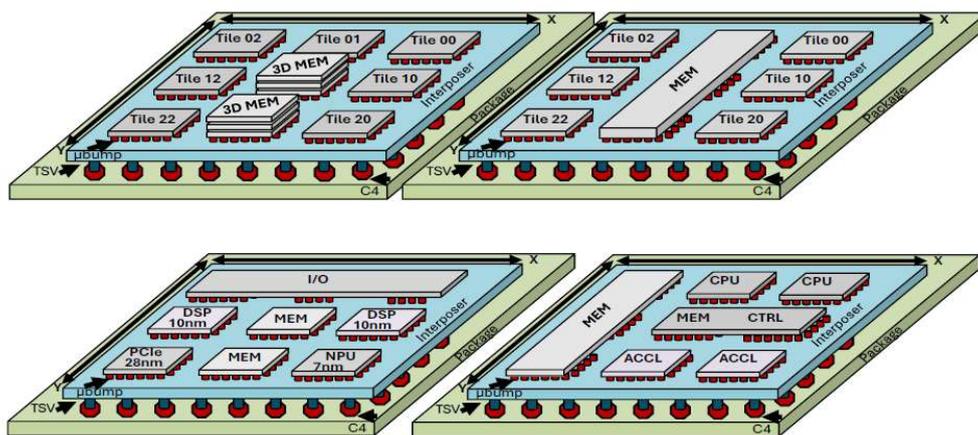
이번 2025 CICC의 Session 33은 "Advancing System Designs with Chiplet Technology"라는 주제로 총 5편의 논문이 발표되었다. 본 세션에서는 칩렛 기반 시스템의 효율성과 확장성을 개선하기 위한 다양한 접근들이 소개되었으며, 각 논문은 테스트 비용 절감, 비정형 데이터 흐름 최적화, 동적 디스패칭 구조, 고속 시리얼링크 설계, AI 추론을 위한 레이턴시-에너지 트레이드오프 등 시스템 수준의 문제 해결을 지향한다. 이 후기를 통해 각 논문을 간단히 정리하고자 한다.

#33-1은 TSMC에서 발표한 논문으로, 차세대 칩렛 간 상호 운용성과 대역폭 밀도를 제공하는 UCle (Universal Chiplet Interconnect Express) 표준에 기반한 인터페이스 설계를 다룬다. 본 논문은 다양한 패키징 기술 및 동작 속도 조건 하에서 UCle 규격을 만족하는 인터페이스를 효율적으로 구현하기 위한 네 가지 핵심 설계 관점을 중심으로 구성된다. 첫째, 송수신 인터페이스 회로 아키텍처 설계에서는 링크의 전력과 성능 간의 균형을 고려한 구성 방식이 제시된다. 둘째, 실리콘 다이 내부에서 PHY 영역 배치를 최적화하고 패키지 경로를 단축하기 위한 floorplanning 전략이 설명된다. 셋째, 패키지 레벨 채널 최적화를 위해 redistribution layer (RDL), interposer 기반 설계, 다중 리피터 삽입 등 다양한 기술의 효과를 비교 분석한다. 넷째, power delivery network(PDN) 분석을 통해 고속 링크 동작 중 발생할 수 있는 전력 노이즈 문제를 해결할 수 있는 설계 방법이 제안된다. 본 논문은 다양한 패키지 기술(CoWoS, InFO, EMIB 등)에 대응하는 최적화된 UCle 설계 전략을 제시하며, 패키지와 실리콘을 아우르는 통합적 설계 접근의 중요성을 강조한다.



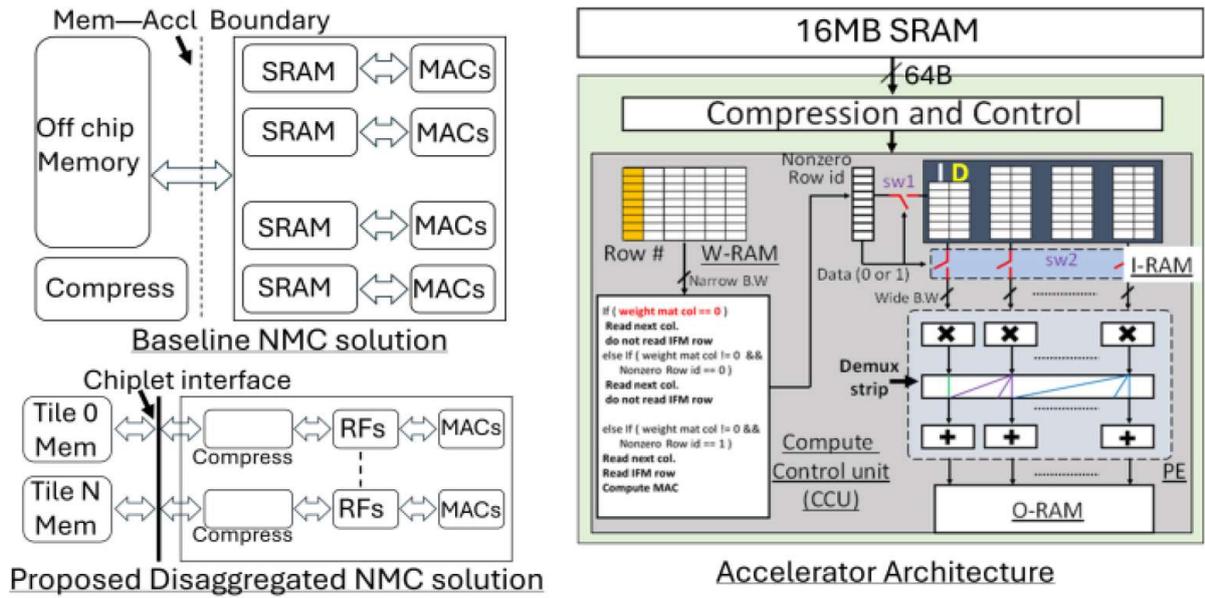
[그림 1] #33-1에서 제안한 상호운용성을 위한 UCle 범프 배치 계획

#33-2는 Intel에서 발표한 논문으로, 다이-투-다이(D2D) 통신을 기반으로 한 이기종 칩렛 기반 AI/미디어 가속 시스템을 위한 고성능 패시브 베이스 다이 설계를 다룬다. 본 논문은 다양한 프로세스 노드(TSMC, Intel 등)에서 제조된 칩렛들을 통합하는 고성능 시스템을 구현하기 위해, 재구성 가능한 다중 칩렛 구조와 커스터마이징된 D2D I/O 구조를 제안한다. 특히, 인공지능 및 비디오 인코딩 워크로드에 적합하도록 설계된 이 시스템은, UCle (Universal Chiplet Interconnect Express)와 호환되며 2Gbps/wire 전송 속도, 0.85pJ/bit의 에너지 효율을 달성하였다. 본 논문은 베이스 다이 설계부터 패키징 및 어셈블리 과정까지의 전반적인 플로우를 다루며, 상이한 파운드리 기반의 칩렛 간 통신을 위한 전기적 및 물리적 설계 최적화 전략을 소개한다. 이 구조는 향후 새로운 애플리케이션을 위한 top-die 교체에도 유연하게 대응 가능하도록 설계되었으며, 분산형 AI/미디어 가속 시스템에 적합한 확장성과 재사용성을 갖춘 베이스 플랫폼을 제공한다.



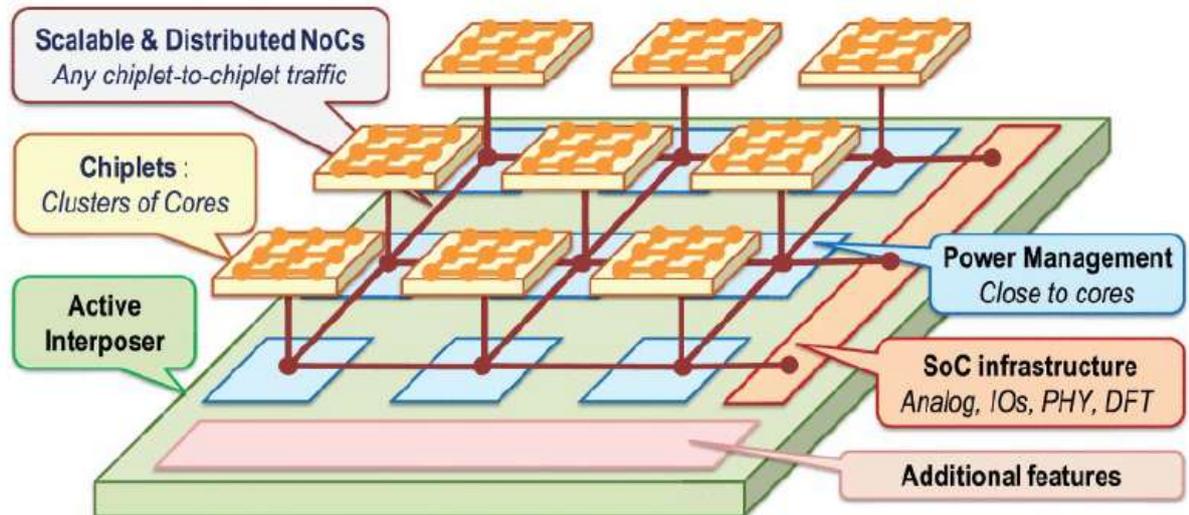
[그림 2] #33-2에서 제안한 2.5D 통합을 사용하여 패시브 기반 칩렛에 대한 방식

#33-3은 Intel에서 발표한 논문으로, 대형 AI 모델의 Sparse GEMM 연산을 고효율로 가속하기 위해, 16개의 Near Memory Compute(NMC) 칩렛을 기반으로 한 2.5D 이기종 시스템을 제안한다. 각 NMC 칩렛은 16MB SRAM과 INT8 연산을 위한 Sparse GEMM 가속기를 내장하고, 전체 시스템은 총 256MB의 온칩 SRAM과 68 TOPS/W의 효율을 달성한다. 데이터 이동을 최소화하기 위해, 입력 행렬은 압축 포맷(COO/CSC)으로 변환되어 저장되며, MAC과 RF 간 2.048Tbps의 대역폭을 통해 고속 연산이 가능하다. 또한, 각 칩렛 간 인터커넥트는 0.8pJ/bit의 에너지로 168Gbps 속도를 지원하며, 메모리-연산의 비율을 1Byte:4096 MACs로 유지한다. 실제 ResNet50 및 Llama3 8B 모델을 기반으로 한 실험에서는 50W~75%의 sparsity를 활용해 최대 3.82배의 속도 향상과 평균 8배의 throughput 개선을 보였고, 전체 시스템은 Sparse 연산 기준 68 TOPS/W를 실현하였다. 본 구조는 확장 가능성이 높으며, 다양한 sequence length 및 sparsity 조건에서도 성능 저하 없이 유연하게 대응할 수 있는 구조로 설계되었다.



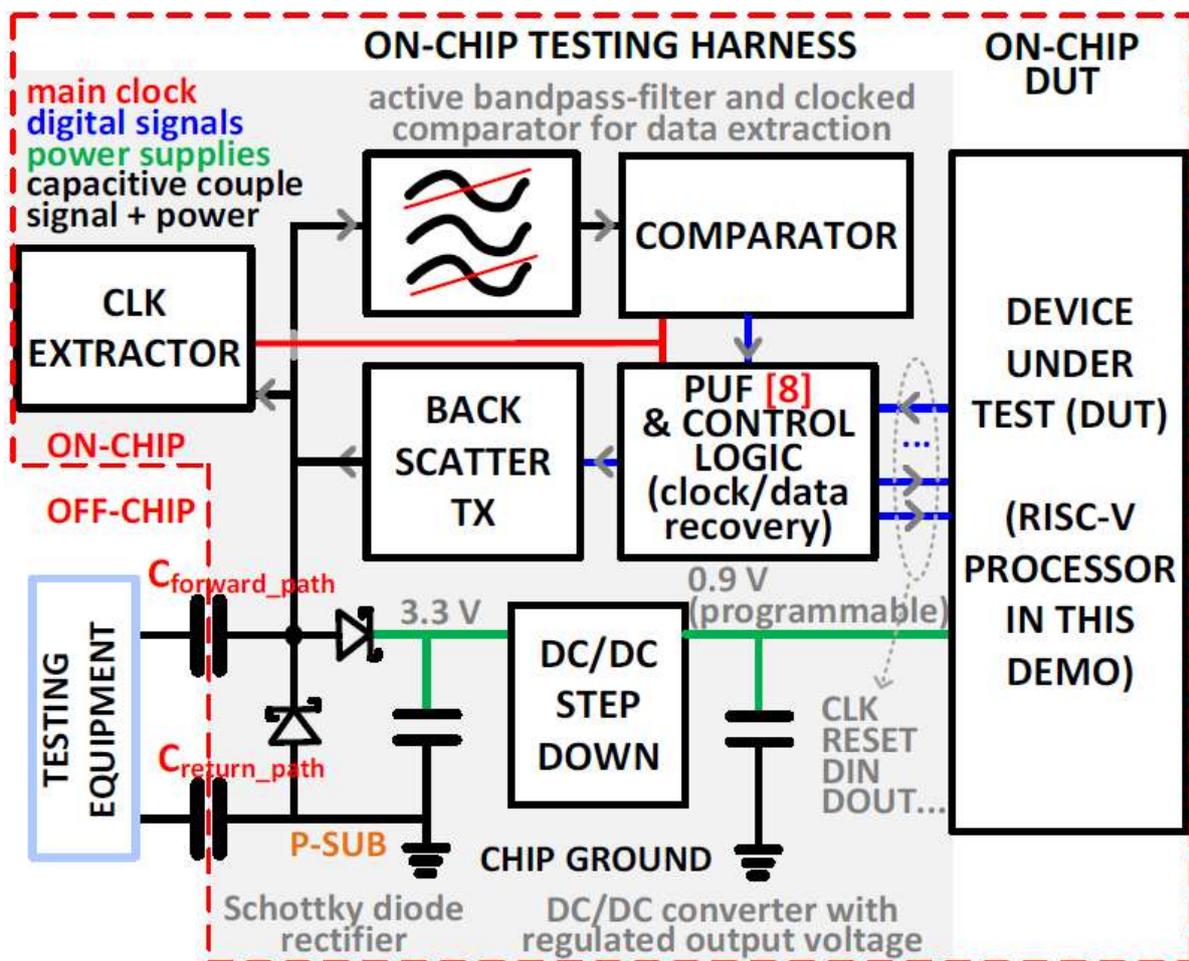
[그림 3] #33-3에서 제안한 데이터 이동 및 연산 감소 기법들을 활용하는 NMC SPGEMM 가속기

#33-4는 Tsinghua University에서 발표한 논문으로, 3D-IC 및 칩렛 기반 SoC에서 전력 공급의 주요 문제를 해결하기 위해 다양한 전력 변환 회로(LDO, SCVR, Buck DC-DC)를 통합한 전원 아키텍처를 제안한다. 고집적 이기종 통합 환경에서 패키지, 재배선층(RDL), 칩스택 레이어를 통한 전력 분배는 전압 강하, 열 문제, 공급 효율 등 복합적인 제약이 따른다. 본 논문은 이러한 전력 벽(power wall)을 극복하기 위해 세 가지 주요 전원 변환기를 집적하고, 각 방식에 적합한 변환 토폴로지, 제어 방식, 최적화 전략을 함께 분석하였다. 특히 칩렛 기반의 계층적 전력 분배를 위한 scalable integration 기법과 함께, 각 회로의 특성에 따라 효율성과 응답성을 비교하여 실질적인 설계 인사이트를 제공하였다. 제안된 프레임워크는 3D-IC 전력 설계에서 회로-시스템 수준의 통합적 접근을 통해 차세대 고성능 SoC의 에너지 효율 극대화를 위한 방향성을 제시한다.



[그림 4] #33-4에서 제안한 다중 전압 도메인 및 VPD 기능을 갖춘 액티브 인터포저

#33-5은 싱가포르국립대학교(NUS)에서 발표한 연구로, 초저가 IoT 칩셋을 위한 완전 무접촉, 위치 불변성 테스트 방법을 제안한다. 기존 칩셋 테스트는 고정밀 프로브 기반이거나 정렬 요구가 있는 비접촉 방식에 의존하여, 단가 대비 테스트 비용이 과도하다는 문제가 있었다. 본 논문에서는 칩셋 전면과 후면 모두에 정전용량(capacitive) 인터페이스를 형성하고, 별도 정렬 없이도 동작 가능한 테스트 구조를 구현하였다. 전면 전극은 최상단 금속층이나 3D 프린팅으로 형성되며, 후면은 실리콘 기판 자체를 이용한다. ASK 변조 기반 전력 및 데이터 전송, Manchester 인코딩 기반 클럭 동기화를 모두 동일 커패시터 경로에서 수행함으로써 인터페이스를 간소화하였다. 또한 각 칩셋은 PUF 기반 ID를 통해 충돌 없이 자신을 식별하고, varactor를 활용한 backscatter 통신을 통해 응답 신호를 송신한다. 제안된 구조는 65nm 공정으로 구현되었으며, 실제 테스트 결과 10mm 이내에서 위치에 관계없이 reliable한 통신 및 테스트가 가능함을 보였다.



[그림 5] #33-에서 제안한 여러 칩셋 간 동시 동작을 지원하는 블록 다이어그램

저자정보



엄소연 박사정 대학원생

- 소속 : KAIST 전기및전자공학부
- 연구분야 : Computing-In-Memory Processor
- 이메일 : soyeon.um@kaist.ac.kr
- 홈페이지 : <https://ssl.kaist.ac.kr/>