2026 International Solid-State Circuits Conference

(ISSCC) Review

연세대학교 전기전자공학부 최승규 교수

Topic : Digital

Session 2 : Processors

이번 ISSCC 2026의 Session 2는 'Processors'라는 주제로 총 10편의 논문이 발표되었다. 본 세션에서는 범용 GPU 및 NPU부터 특정 응용처에 최적화된 가속기까지 폭넓은 프로세서 아키텍처가 소개되었다. 특히 고성능과 전력 효율을 동시에 달성하기 위한 3D Stacking 및 2.5D Chiplet 통합 기술이 주요하게 다루어졌다. 세부적으로는 AI 애플리케이션을 위한 3nm 기반 3D Stacking GPU, 대규모 AI 추론을 위한 Quad Chiplet AI SoC를 비롯하여 자율주행, 3D/4D Gaussian Splatting, Diffusion Model 가속기 등 최신 알고리즘에 최적화된 설계 기법들이 발표되었다.

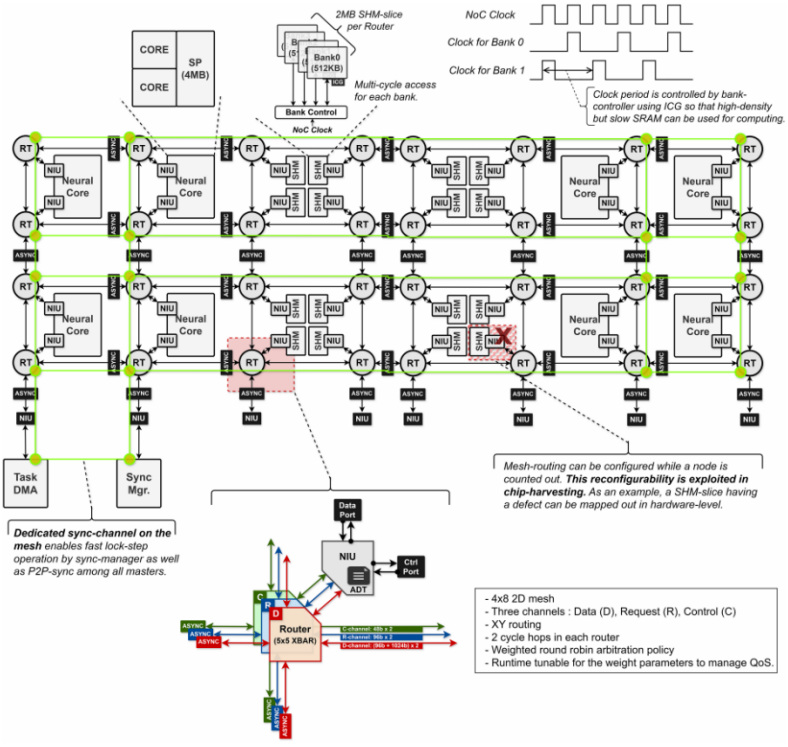
#2.1은 AMD에서 발표한 데이터센터용 AI 가속기인 Instinct MI350 Series GPU에 관한 논문이다. 기존 2D 구현의 한계를 뛰어넘기 위해 TSMC 3nm 공정으로 제작된 8개의 XCD와 6nm 공정의 IOD를 3D CoWoS-S 패키징으로 결합하였다. 이전 세대인 CDNA 3 대비 Transistor Density를 높여 하드웨어 리소스를 2배로 늘렸으며, 최근 AI 트렌드에 맞춰 FP6 및 FP4 Datatype을 새롭게 지원한다. 또한, 3nm 공정 전환에 따른 Voltage Droop 문제를 Adaptive Clocking과 개선된 Power Delivery Network를 통해 극복하였다. 해당 GPU는 이전 세대 대비 Energy Efficiency를 높이고 연산 처리량을 증가시킴으로써 광범위한 AI 추론 워크로드에서 3배 이상의 성능 향상을 달성하였다.

Computation	FLOPS/clock/CU		Peak Theoretical		MI355X Peak Speedup Over MI300X[1]
	MI300X	MI355X	MI300X	MI355X	
Matrix FP16/BF16	2048	4096	1.3 PF	2.5 PF	1.9x
Matrix FP8	2048	4096	2.6 PF	5 PF	1.9x
Matrix INT8	4096	8192	2.6 POPs	5 POPs	1.9x
Matrix MXFP6	NA	16384	NA	10 PF	New to MI350
Matrix MXFP4	NA	16384	NA	10 PF	New to MI350

Peak theoretical performance without sparsity

[그림 1] 세대별 Peak Theoretical Performance Improvement 비교

#2-2는 Rebellions에서 발표한 대규모 LLM 추론용 Quad-Chiplet SoC이다. Llama 3과 같은 초대형 모델에서 발생하는 Prefill과 Decode Phase의 병목 현상을 해결하기 위해, 4nm 기반 NPU Chiplet 4개와 4개의 HBM3E를 16Gbps UCIe Interface로 연결하여 거대한 Virtual Monolithic System을 구성하였다. 여러 코어가 동시에 활성화될 때 발생하는 High di/dt로 인한 Power Integrity 문제를 해결하기 위해 하드웨어 기반 Staggered Startup 기법과 ISC를 도입하였다. 이를 통해 Chiplet 구조의 전력 및 동기화 문제를 극복하였으며, LLaMA v3.3 70B 모델 기준 단일 칩 셋업에서 56.8 TPS의 추론 성능을 달성하였다.



[그림 1] "Neural cores, DMA, Synchronization manager를 탑재한 Full-chip scale mesh 및 3개의 Logically independent channels 연결 구조

#2.3은 UNIST와 연세대학교에서 공동으로 발표한 28nm 기반 자율주행용 EED 프로세서이다. 다중 센서 데이터를 융합하는 최신 트랜스포머 기반 EED 모델은 Sparsity가 불규칙하고 Temporal Attention 처리 시 EMA가 폭증하는 단점이 있다. 이를 극복하기 위해 과거 프레임의 정보를 바탕으로 현재 프레임의 Sparsity를 예측 및 생성하는 SRU를 고안하여 코어 활용도를 극대화하였다. 또한, 불필요한 Temporal Memory를 점진적으로 폐기하는 LSTMU 구조를 적용하였다. 결과적으로 Temporal Attention 단계의 EMA를 92.8% 감소시켰으며, 71.3mJ/Frame의 에너지 소비량으로 최첨단 자율주행 SoC 대비 218배 높은 에너지 효율과 10.3fps의 처리 속도를 달성하였다.

#2.4는 KAIST, POSTECH, University of Michigan에서 공동 발표한 B5G/6G AI-RAN용 Channel Estimation Accelerator이다. 기존 통신 시스템의 Channel Estimation을 ViT로 대체할 때 발생하는 연산량과 HRLLC 요구사항을 만족시키기 위해 설계되었다. 채널 특성을 고려한 CAC 기법을 적용해 가중치를 95% 줄여 전체 모델을 On-chip SRAM에 적재함으로써 EMA를 제거하였다. 또한, 비선형 연산을 단순화한 RSA를 도입하여 지연 시간을 크게 줄였으며, 주파수 대역에 따라 Patch Embedding Dimension을 유연하게 스위칭하는 구조를 채택하였다. 본 프로세서는 28nm 공정에서 최첨단 채널 추정 시스템 대비 16~39배 높은 14.4Gb/s의 Throughput과 7.3pJ/b의 에너지 효율을 달성하였다.

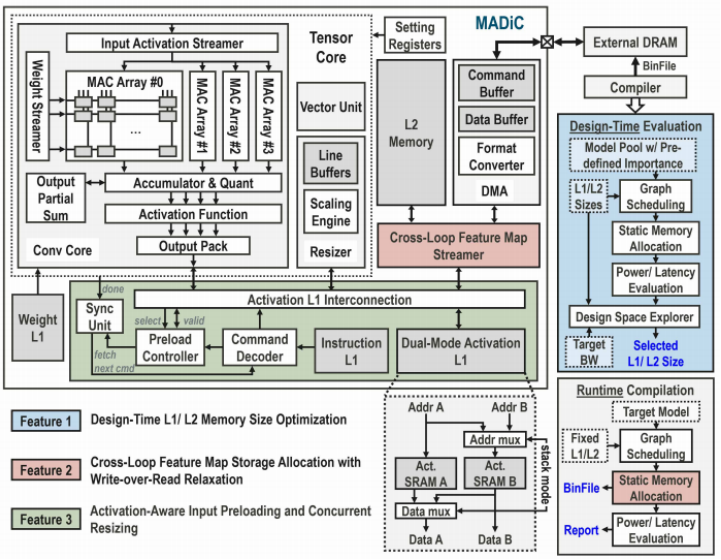
#2-5는 University of Michigan, POSTECH, KAIST에서 제안한 URLLC용 FEC Decoder이다. 짧은 코드 길이에서 높은 신뢰성을 얻기 위해 OSD가 사용되지만, 복잡한 GE 연산으로 인해 Latency가 길어지는 한계가 존재했다. 본 논문은 Soft-decision 기반 Chase Decoder와 Order-1 OSD를 결합한 Two-stage Decoder 구조를 채택하였다. 이와 함께 사전에 정렬된 비트들을 바탕으로 연산하는 PSBU 기법을 도입하여 기존 GE 연산 대비 Latency를 75% 이상, 면적을 39% 단축시켰다. 16nm 공정으로 제작된 해당 디코더는 기존 최고 수준 기술 대비 53배 개선된 14.4ns의 평균 Latency를 기록하였으며, 13.1pJ/b의 에너지 효율로 15.2Gbps의 Information Throughput을 달성하였다.

#2-6은 IBM Research에서 발표한 Enterprise Workload에 최적화된 PCIe Form Factor AI Accelerator 'Spyre'이다. 서버의 단일 PCIe 슬롯이 갖는 Tight Power Budget 내에서 성능을 극대화하기 위해 다중 전압 도메인 전략을 도입하였으며, 연산 밀집 영역에는 0.55V의 낮은 전압을 할당하였다. 특히, 긴 시간과 짧은 시간 단위의 Peak Power를 서로 다르

게 조절하는 Dual-loop Control을 적용하여 Power-limited Throttling을 최소화하고 코어 활용률을 크게 높였다. 본 프로세서는 단일 루프 제어 대비 최대 28%의 Throughput 개선을 이루었으며, Encoder-class 모델 기준 GPU 대비 2~3배 향상된 Power/Performance 수치를 달성하였다.

#2-7은 National Tsing Hua University에서 발표한 Transformer 기반 Diffusion Model Processor 'Tiamat'에 관한 논문이다. 고품질 이미지 생성을 위해 CFG 기법과 MX 데이터 포맷이 도입되고 있지만, 막대한 양의 EMA와 FP Residual Connection으로 인한 품질 저하 및 버퍼 크기 증가 문제가 존재했다. 이를 해결하기 위해 CFG 배치 간 유사성을 활용하는 CBDP 흐름을 적용해 Weight EMA를 크게 줄였다. 또한, Weight 정렬 기반 양자화인 WRQ와 TSBFA를 결합하고, Hybrid MX Blocking Datapath인 HMBD 및 TSQ를 도입하여 FP 수준의 품질을 유지하면서 On-chip Buffer 요구량을 대폭 낮췄다. 해당 프로세서는 16nm FinFET 공정으로 제작되어 400MHz 기준 7.37 TOPS의 연산량과 4.99 TOPS/W의 에너지 효율을 기록하며 DiT-XL/2 및 PixArt-α 모델에서 각각 98ms/Step과 134ms/Step의 생성 속도를 달성하였다.

#2-8은 MediaTek에서 제안한 3nm 기반 Generative Diffusion Accelerator 'MADiC'이다. 최신 공정에서는 Logic 대비 SRAM의 Area Scaling 효율이 떨어져 On-chip Memory를 무작정 늘리기 어렵다는 현실적인 제약이 있다. 이를 극복하기 위해 Compiler 단에서 EMA 목표치를 만족하면서 SRAM Area Cost를 최소화하는 Design-time L1/L2 Memory 최적화를 수행하였다. 더불어 WoR Relaxation을 적용한 Cross-loop FM Allocation 전략으로 L2 Bandwidth를 극대화하고, Activation-aware Input Preloading과 Concurrent Resizing을 통해 Operator Parallelism을 구현하여 Hardware Utilization을 크게 끌어올렸다. 3nm 공정에서 불과 0.338mm²의 면적만을 차지하는 이 프로세서는 0.575V 전압에서 Diffusion ConvNet 구동 시 17.4 TOPS/W의 에너지 효율과 7.4 TOPS/mm²의 면적 효율을 기록하며 0.116s의 빠른 Inference Latency를 달성하였다.



[그림 1] " A 3nm 7.4TOPS/mm2, 17.4TOPS/W Generative Diffusion Accelerator Enabled by Hardware-Compiler Co-Optimization of Memory Hierarchy and Operator Parallelism" 전체 아키텍처 구조

#2-9는 Tsinghua University에서 발표한 4D Gaussian Splatting (4DGS) 전용 프로세서이다. 4DGS는 동적 씬 렌더링에 탁월하지만, 막대한 메모리 접근, Opacity 계산의 연산 중복성, 직렬 픽셀 렌더링으로 인한 낮은 PE Utilization 문제가 있었다. 본 논문은 메모리 집약적인 Pre-processing을 이웃 프레임의 2차 보간으로 대체하는 AQFI 유닛을 도입해 Memory Access Overhead를 대폭 줄였다. 또한, 중복 연산을 재사용하는 FROC 유닛과, 직렬 렌더링 의존성을 깨고 병렬 처리를 가능하게 하는 Tree-based TAPR 코어를 설계하여 연산 효율을 극대화하였다. 28nm 공정으로 제작된 해당 프로세서는 175MHz 및 0.65V 구동 환경에서 16.27 TFLOPS/W의 최고 에너지 효율을 달성하였으며, D-NeRF 데이터셋 기준 0.24mJ/Frame의 렌더링 에너지로 365.3 fps의 실시간 성능을 달성하였다.

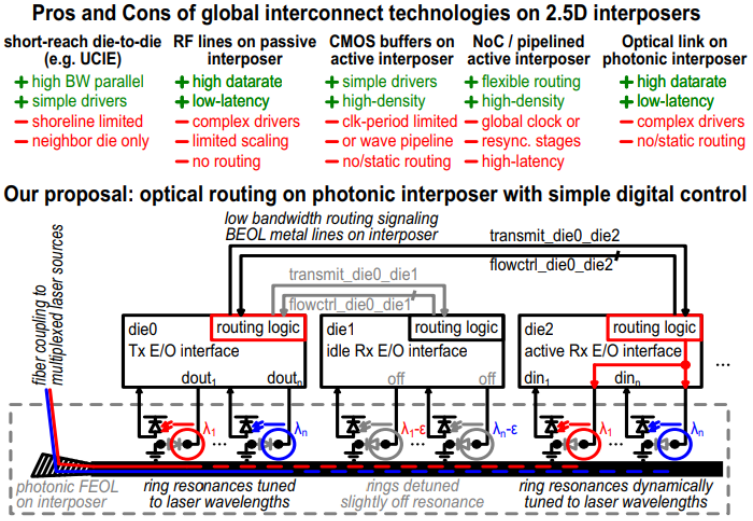
#2-10은 Tsinghua University에서 개발한 Modeling 및 Rendering 통합형 3D Gaussian Splatting (3D GS) 프로세서이다. 기존 Feedforward 3D GS 방식은 국소적인 작은 Gaussian들을 생성하여 Coarse-grained Rasterization에서 연산 낭비가 심했고, 2D Gaussian의 불규칙한 형상 탓에 막대한 EMA가 발생했다. 이를 해결하기 위해 SBTU 기반의 DFGRE를 도입하여 중복 연산을 피하고, SWB Canvas를 활용한 LOUR 워크플로우를 통해 Memory Gathering 단계를 제거하여 EMA를 낮췄다. 더 나아가, Modeling과 Rendering 모드 간 데이터플로우 유사성을 활용해 하드웨어 자원을 공유하는 URA를 채

택함으로써 Area Overhead를 최소화하였다. 28nm 공정에서 4.76mm² 면적으로 구현된 본 프로세서는 680MHz에서 1286 fps의 렌더링 Throughput을 기록하였으며, 80MHz 구동 시 0.15mJ/Frame의 압도적인 에너지 효율과 함께 0.30s의 빠른 Modeling Latency를 달성하였다.

Session 18 : Technology and Circuits for Domain-Specific Accelerators

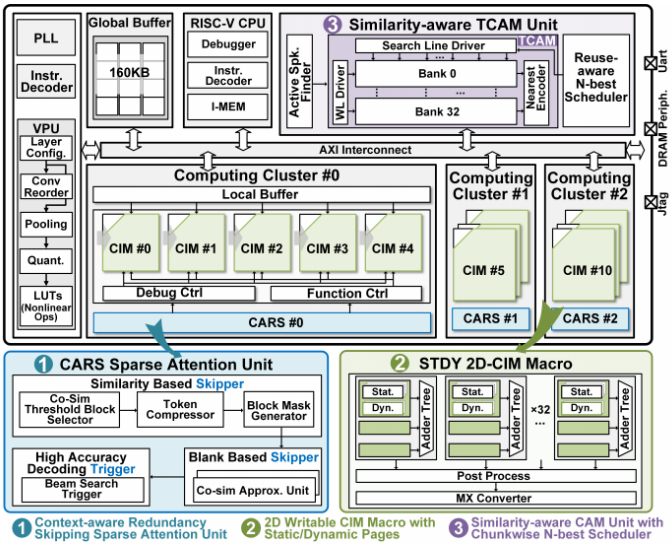
이번 ISSCC 2026의 Session 18에서는 Domain-Specific Accelerators를 주제로 총 5편의 논문이 발표되었다. 올해 해당 세션의 기술적 트렌드는 크게 세 가지로 요약된다. 첫째, Multi-die 가속 플랫폼을 위한 고효율 Chip-to-Chip Communication 기술 (논문 18.1). 둘째, LLM, Generative AI 및 Multi-speaker ASR과 같은 고도화된 머신러닝 워크로드의 특성(Sparsity, Redundancy, Data Distribution)을 정밀하게 분석하여 하드웨어 Utilization을 극대화한 구조 (논문 18.2, 18.3, 18.5). 셋째, 센서와 컴퓨팅, 메모리를 긴밀하게 결합하여 데이터 이동을 최소화하고 On-chip Learning의 한계를 극복한 Neuromorphic System이다 (논문 18.4).

#18.1은 CEA-List와 CEA-Léti에서 공동 발표한 Photonic Interposer를 위한 Electro-Optical Router에 관한 논문이다. Chiplet 스태킹 시 면적이 커짐에 따라 기존 전기적 Die-to-Die Interconnect는 도달 거리와 Latency에 한계를 보이며, 기존 Photonic Interposer 역시 복잡한 드라이버와 Static Routing에 의존해야 하는 문제가 있었다. 이를 해결하기 위해 단순한 Digital Control만으로 Frame-level Dynamic Routing을 수행하는 Lightweight 구조를 제안하였다. Tx의 Microring과 Rx의 Resonance Shift를 활용하여 1-to-6 Wavelength-flexible Link Capacity를 확보하고, Dynamic Threshold Adjustment 기능을 갖춘 Rx Front-end를 통해 신호 무결성을 높였다. 28nm FD-SOI 공정으로 제작된 해당 라우터는 18ns의 짧은 Link Setup Time과 4Gbaud의 Datarate를 달성하였으며, 0.007mm²의 매우 작은 Active Area 내에서 3.19pJ/b의 에너지 효율을 달성하였다.



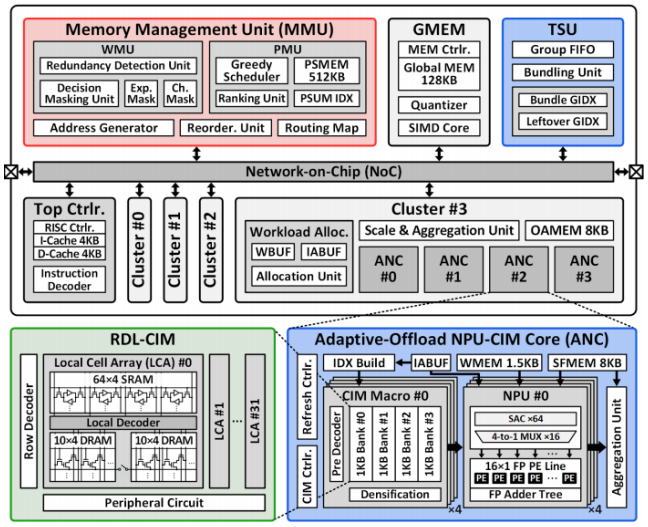
[그림 1] 3D-stacking 기술의 Scaling Perspective 및 Interposer 기반 Routed Optical Links

#18.2는 Peking University에서 제안한 Streaming MS-ASR Accelerator이다. 실제 다자간 대화 음성 인식 환경에서는 Silent Pause나 Speaker Alternation으로 인해 Block-level Computation Redundancy가 발생하고 2-Pass Decoding에 따른 메모리 병목이 심각하다. 본 논문은 Online Sparse Block Prediction을 활용해 불필요한 연산을 건너뛰는 CARS 기법을 제안하였다. 또한 MX Format 연산과 Native In-memory Matrix Transpose를 동시 지원하는 STDY-2D-DCIM 구조와, 화자 임베딩 및 N-best Rescoring을 빠르게 수행하는 SA-TCAM을 결합하여 지연 시간을 크게 단축시켰다. 22nm CMOS 공정으로 제작된 본 가속기는 582MHz 및 1.0V 구동 환경에서 1.87ms/Frame의 빠른 처리 속도와 0.158mJ/Frame의 에너지 소비량을 기록하며, MXFP8 포맷 기준 37.50 TFLOPS/W의 우수한 시스템 에너지 효율을 달성하였다.



[그림 1] Overall Architecture of Proposed ASR Accelerator 및 3가지 Main Features

#18.3은 KAIST와 MIT에서 공동으로 발표한 Sparse MoE-based Speculative Decoding LLM Processor인 'SMoLPU'이다. MoE 기반 Speculative Decoding은 EMA를 줄이는 데 효과적이지만, 예측 실패 토큰으로 인한 Expert Fetching 낭비와 런타임에 동적으로 변하는 INT-FP Ratio 때문에 Hardware Utilization이 크게 떨어지는 단점이 있다. 이를 극복하기 위해 Redundant Expert Activation을 제거하는 TaER 기법을 적용하고, 동적인 Workload 환경에서도 높은 가동률을 유지하도록 ANC 및 TSU를 도입하였다. 이에 더해 BCD를 활용한 RDL-CIM을 적용하여 Adder Tree의 Power 및 Area Overhead를 최소화하였다. 28nm FD-SOI 공정으로 구현된 해당 프로세서는 50MHz 및 0.7V 조건에서 122.1μJ/Token의 뛰어난 에너지 효율을 달성하였으며, 20.25mm² 면적에서 107.3 TOPS/W의 Peak Energy Efficiency를 달성하였다.



[그림 1] "122.1μJ/Token Sparse MoE-Based Speculative Decoding Language Processing Unit with Adaptive-Offload NPU-CIM Core" 전체 아키텍처 구조

#18.4는 HKUST, North China Research Institute of Electro-Optics, SynSense가 공동 연구한 Event-driven Spiking Processor 'SpikeRAM'이다. Edge 디바이스에서 SNN과 EVS를 결합하면 에너지와 개인정보 보호 측면에서 유리하지만, 센싱과 컴퓨팅의 분리로 인한 데이터 이동 오버헤드와 OCL 수행 시 발생하는 eNVM Endurance 및 Power 문제가 존재했다. 본 논문은 On-chip EVS와 SNN Core를 직접 결합한 Memory-centric Asynchronous Near-sensor Computing 구조를 도입하였다. 또한 Single Time-window만으로 학습을 수행하는 e-OTBP 알고리즘과 Ternary Gradient, Gray-code Weight를 적용하여 MRAM Update 횟수를 줄였다. 65nm Inference Core와 40nm On-chip Learning Core가 결합된 이 플랫폼은 1.1V 및 10MHz 환경에서 8.28mW의 시스템 전력을 소모하며, 48.1pW/Synapse/Bit의 Power Density와 Event-based Signature Verification 태스크에서 94.3%의 정확도를 달성하였다.

#18.5는 Tsinghua University에서 발표한 VAR Accelerator에 관한 논문이다. 이미지 생성에 특화된 VAR 모델은 우수한 성능을 보이나, Attention Noise 연산의 낭비, 비대칭적인 Data Distribution, 그리고 Sequential Generation 특성으로 인해 막대한 지연 시간이 발생했다. 연구진은 DVAA를 통해 Attention Noise를 억제하고 불필요한 연산을 생략하는 Speculative Execution 기법을 도입하였다. 또한 데이터 분포에 맞추어 동적으로 정밀도를 조절하는 Full-path Optimized MXINT PE를 설계하고, 공간적 상관관계를 이용해 여러 픽셀을 동시에 생성해 내는 CRPG 기법을 적용하여 병목을 해소하였다. 28nm 공정으로

5.76mm²의 면적에 구현된 해당 칩은 400MHz, 0.9V 환경에서 동작하며, DeiT/ViT VAR 가속 시 47.3 TFLOPs/W의 에너지 효율과 2.75 TFLOPs/mm²의 면적 효율을 달성하였다.

저자정보

최승규 교수

- 소 속 : 연세대학교 전기전자공학부
- 연구분야 : AI accelerator design
- 이 메 일 : seungkc@yonsei.ac.kr
- 홈페이지 : acalabys.github.io

2026 International Solid-State Circuits Conference

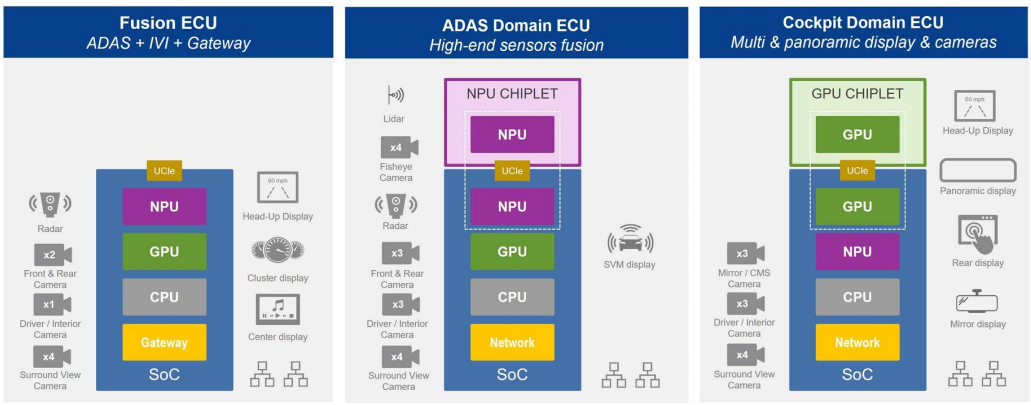
(ISSCC) Review

광주과학기술원 전기전자컴퓨터공학과 김상진 교수

Topic : Digital

Session 10: Digital Processing and Circuit Techniques

ISSCC 2026 Session 10은 특정 응용 하나에 집중된 세션이라기보다, 디지털 시스템 전반의 병목을 회로와 아키텍처 수준에서 어떻게 풀 것인가를 보여준 세션으로 정리할 수 있다. 먼저 10.5와 10.6은 전원 무결성, 데이터 이동, 다이 간 통신처럼 연산기 바깥의 문제를 정면으로 다루며, 최근 디지털 설계의 초점이 단순 연산 성능보다 system-aware optimization으로 이동하고 있음을 보여준다. 또한 10.1과 10.6은 자동차용 SoC와 hybrid-bonded 3D DNN processor를 통해, 다양한 기능 블록과 칩렛·적층 구조를 유연하게 결합하는 scalable heterogeneous integration 흐름을 드러낸다. 한편 10.7~10.10은 SAT 및 조합최적화 문제를 위한 전용 하드웨어를 제시하며, 디지털 회로 기법의 적용 범위가 전통적인 CPU·NPU를 넘어 domain-specific digital solver로 확장되고 있음을 보여준다. 즉, 이번 Session 10은 디지털 회로 연구가 미세한 블록 최적화에 머무르지 않고, 전력, 통신, 적층, 문제 특화성까지 함께 고려하는 시스템 지향적 방향으로 진화하고 있음을 잘 보여준 세션이라 할 수 있다.

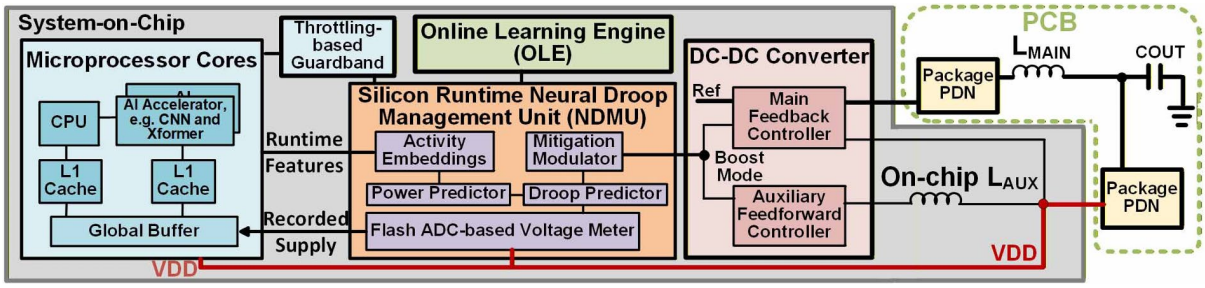


[그림 1] "10.1. SoC Chiplet Support for ASIL D Automotive Cross-Domain Applications" 개념도

#10-1 은 Renesas Electronics 에서 발표한 SDV(Software-Defined Vehicle)용 SoC 로, 이번 Session 10 에서 나타난 대규모 이종 시스템 통합 흐름을 잘 보여주는 논문이다. 기존

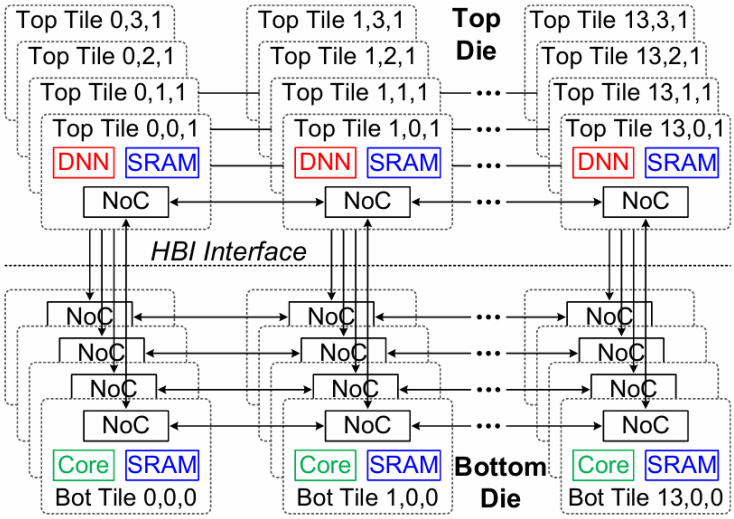
차량용 반도체가 기능별로 분리된 ECU 중심 구조에 가까웠다면, 이 논문은 autonomous driving 과 cockpit workload 를 하나의 칩 안에 함께 통합하면서도, 서로 다른 연산 자원과 안전 요구를 동시에 만족시키는 구조를 제안한 것이 특징이다. 특히 heterogeneous processing resource 와 chiplet extension 을 지원하고, automotive safety standard 를 만족시키기 위한 deterministic partitioning 과 interference control 까지 함께 고려했다는 점이 인상적이다. 또한 UCIe 기반 chiplet 연결과 RegionID 기반 접근 제어를 통해, 확장성과 기능 안전성을 동시에 확보하려는 SDV 용 시스템 아키텍처를 제시했다는 점에서 의미가 크다. 즉, 이 논문은 자동차용 반도체의 경쟁축이 더 이상 개별 연산 성능 향상에만 있지 않고, 다양한 기능 블록을 안전하고 유연하게 통합할 수 있는 시스템 수준 아키텍처로 이동하고 있음을 보여주는 대표적인 사례라고 할 수 있다.

#10-5 은 Northwestern University 에서 발표한 28nm CMOS 기반 전원 관리 SoC 로, 이번 Session 10 에서 드러난 system-aware optimization 흐름을 잘 보여주는 논문이다. 기존 droop mitigation 기법들이 주로 보수적인 guardband 설정이나 throttling 에 의존해 성능 손실을 감수했던 것과 달리, 이 논문은 workload 와 PDN variation 을 함께 고려하는 online-learning 기반 proactive power management 를 제안한 것이 특징이다. 특히 CPU, CNN, transformer accelerator 의 activity 를 바탕으로 향후 droop 을 예측하는 Neural Droop Management Unit 과, silicon 및 패키지 조건 변화에 적응하기 위한 on-device online learning engine, 그리고 이를 실제 regulation 에 반영하는 고속 DC-DC converter 를 하나의 구조 안에 통합하였다. 즉, 전원 무결성 문제를 단순한 회로 안정성 이슈가 아니라 실행 중 workload 변화에 따라 학습하고 대응해야 하는 시스템 문제로 재해석했다는 점에서 의미가 크다. 또한 다양한 패키지 조건과 workload 변화에 대해 적응적으로 동작하도록 설계되었다는 점에서, 최근 디지털 시스템 설계의 관심이 단순한 worst-case margin 확보보다 variation-tolerant, workload-aware power delivery 로 이동하고 있음을 잘 보여준다.



[그림 2] "10.5. Proactive Power Management-Based Supply Regulation with Online Learning" 구조

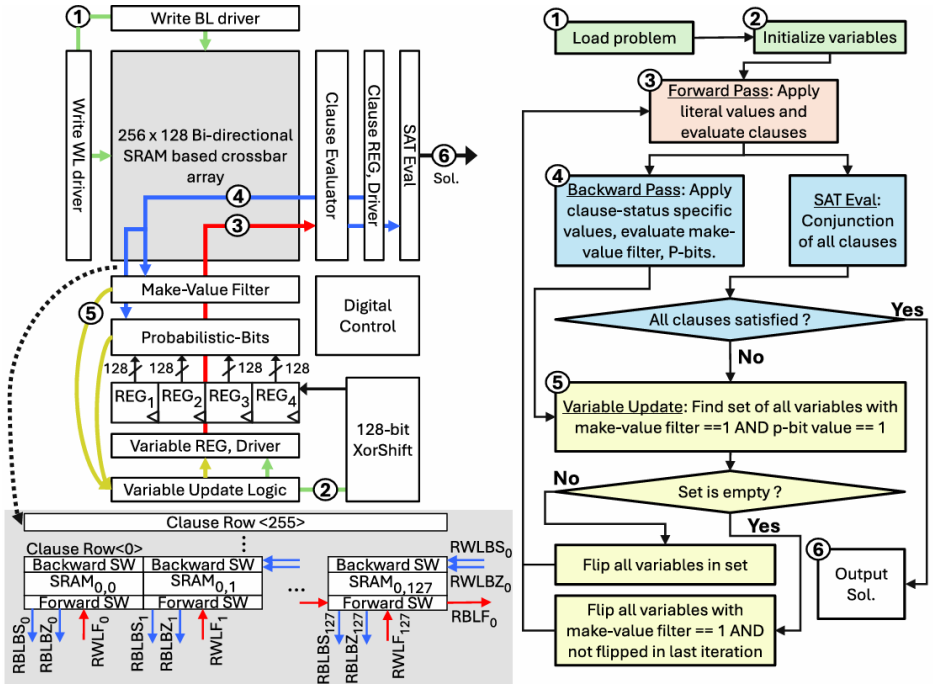
#10-6 은 Intel 에서 발표한 hybrid-bonded 3D DNN processor 로, 이번 Session 10 에서 가장 선명하게 드러난 scalable heterogeneous integration 과 data-movement-centric optimization 흐름을 대표하는 논문이다. 최근 AI 프로세서에서는 연산 밀도를 높이는 것만으로는 성능 향상이 제한되고, 코어·가속기·메모리 사이의 데이터 이동과 다이 간 통신이 오히려 더 큰 병목이 되는 경우가 많다. 이 논문은 이러한 문제를 해결하기 위해, 하단 다이에 범용 RISC-V manycore 를 두고 상단 다이에 DNN accelerator tile 을 적층한 뒤, 이를 hybrid-bonded interconnect(HBI) 와 3D NoC 로 긴밀하게 연결하는 구조를 제안한 것이 특징이다. 특히 3D 적층을 단순히 면적 절감 수단으로 사용하는 데 그치지 않고, 공유 SRAM 과 compute tile 사이의 접근을 더 촘촘하게 연결하여 메모리 대역폭과 통신 효율 자체를 높이는 방향으로 활용했다는 점이 인상적이다. 또한 bottom die 를 범용 manycore platform 으로 두고, top die 만 응용 특화형으로 교체 가능한 구조를 취함으로써, 향후 다양한 application-specific top die 로의 확장 가능성까지 고려했다는 점에서도 의미가 크다. 즉, 이 논문은 최근 디지털 시스템 설계의 관심이 더 이상 개별 연산 블록의 효율 향상에만 있지 않고, 칩 간 연결, 적층 구조, 메모리 접근 경로까지 포함한 시스템 수준 최적화로 이동하고 있음을 잘 보여주는 대표적인 사례라고 할 수 있다.



[그림 3] "10.6. A Hybrid-Bonded 12.1TOPS/mm2 56-Core DNN Processor with 2.5Tb/s/mm2 3D Network on Chip" 전체 구조

#10-10 은 University of Minnesota 에서 발표한 mixed-signal p-bit 기반 K-SAT solver 로, 이번 Session 10 에서 드러난 domain-specific digital solver 흐름을 잘 보여주는 논문이다. 최근 디지털 하드웨어 연구가 전통적인 CPU 나 AI accelerator 를 넘어, 조합최적화나

SAT 와 같은 비전통적 문제를 직접 풀기 위한 전용 구조로 확장되고 있는데, 이 논문은 그 흐름을 대표하는 사례라 할 수 있다. 특히 deterministic 한 연산 블록 대신 확률적으로 동작하는 p-bit 과 CIM 기반 병렬 변수 업데이트를 활용해, K-SAT 문제를 보다 에너지 효율적으로 탐색하려는 점이 특징이다. 즉, 기존의 정확한 순차형 탐색이나 범용 프로세서 기반 접근과 달리, 문제 자체의 확률적 특성과 병렬성을 하드웨어 구조 안에 직접 반영했다는 점에서 의미가 크다. 또한 mixed-signal CIM 을 이용해 변수 업데이트와 상호작용 계산을 밀접하게 결합함으로써, 단순 연산 성능보다 문제 구조에 특화된 계산 방식이 더 중요해지고 있음을 보여준다. 물론 이러한 구조는 적용 대상이 비교적 명확한 문제군에 한정될 수 있다는 점에서 범용성의 한계는 있지만, 반대로 보면 디지털 회로 기술의 적용 범위가 이제는 범용 컴퓨팅을 넘어 특정 문제를 빠르고 효율적으로 풀기 위한 전용 솔버로까지 확장되고 있음을 보여주는 흥미로운 사례라고 할 수 있다.



[그림 4] "10.10. Probabilistic K-SAT Solver with p-Bit-Based Parallel-Variable Update on a Mixed-Signal Compute-in-Memory Architecture" 전체 구조 및 워크플로우

Session 31: AI Accelerators

ISSCC 2026 Session 31은 생성형 AI 가속기가 단순한 연산 효율 경쟁을 넘어, 실제 추론 병목을 줄이기 위한 시스템 최적화로 확장되고 있음을 보여주었다. 특히 올해는 LLM decoding 최적화, rotation 기반 저비트화, 모델/응용의 확장, 온디바이스 개인화의 네 가지 흐름이 두드러졌다.

1) Decoding-Centric Optimization (31.1, 31.8): LLM 추론의 핵심 병목이 prefill보다 decoding 단계의 반복적 weight access와 제어 비효율에 있다는 점에 주목한 연구들이 눈에 띄었다. 31.1과 31.8은 speculative decoding을 하드웨어 구조와 결합하여 token 생성 지연과 메모리 병목을 함께 줄이려는 방향을 보여준다.

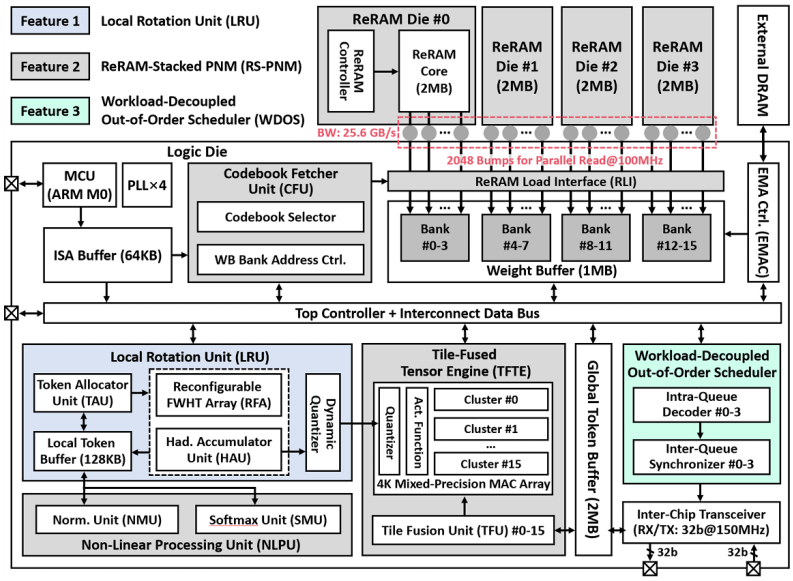
2) Rotation-Based Low-Bit Quantization (31.1, 31.2, 31.3): 저비트 정확도 저하의 원인인 activation outlier를 완화하기 위해 rotation을 적극 활용하는 흐름이 강하게 나타났다. 31.1의 local rotation, 31.2의 rotation-based group quantization, 31.3의 subspace rotation은 모두 rotation이 이제 알고리즘 기법을 넘어 하드웨어 datapath 설계의 핵심 축이 되었음을 보여준다.

3) Beyond Text LLMs: Model Expansion (31.4, 31.6, 31.7): 이번 세션은 텍스트 LLM뿐 아니라 visual autoregressive generation, vision-language model, state-space model까지 가속 대상이 넓어졌다는 점에서도 의미가 있다. 즉, 생성형 AI 하드웨어가 특정 transformer LLM에 머무르지 않고 더 다양한 모델 구조와 멀티모달 응용으로 확장되고 있음을 보여준다.

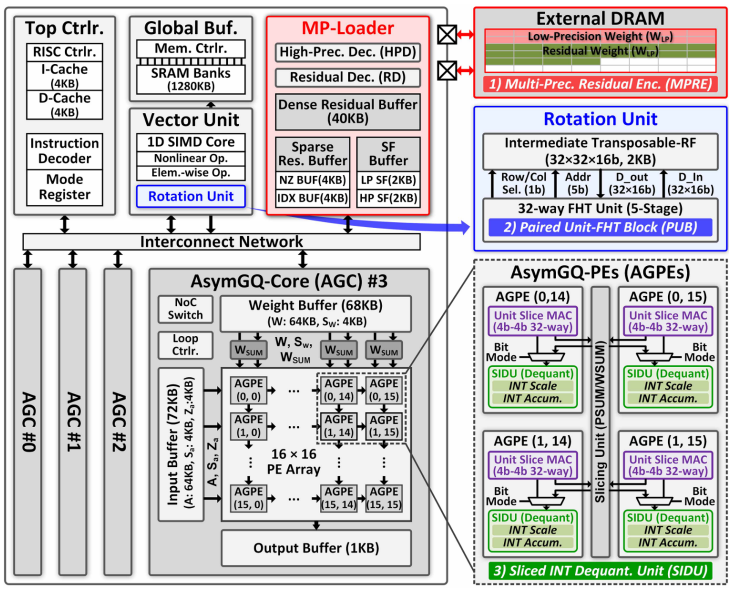
4) On-Device Personalization (31.5): 31.5는 personal LLM, RAG, fine-tuning을 모바일 단말에서 직접 수행하는 방향을 제시하며, 생성형 AI 가속기의 또 다른 축이 서버 추론을 넘어 사용자 맞춤형 온디바이스 지능으로 이동하고 있음을 보여준다. 성능뿐 아니라 privacy, responsiveness, personalization을 함께 고려했다는 점이 인상적이다.

#31-1은 홍콩과학기술대학교(HKUST)에서 발표한 LLM 가속기로, 이번 세션에서 특히 두드러졌던 speculative decoding과 memory-centric 최적화 흐름을 잘 보여주는 논문이다. 최근 LLM 추론에서는 prefill보다 decoding 단계에서 동일한 weight를 반복적으로 불러오는 과정이 더 큰 병목으로 작용하는데, 이 논문은 이를 해결하기 위해 ReRAM-on-Logic 적층 메모리, local rotation 기반 저비트 양자화, adaptive parallel speculative decoding을 하나의 시스템 안에 통합하였다. 특히 rotation을 단순한 알고리즘적 보조기법이 아니라 실제 하드웨어 datapath에 녹여냈다는 점과, speculative decoding 역시 draft model 구조와 메모리 access, scheduler까지 함께 설계해야 실질적인 효과를 얻을 수 있음을 보여준 점이 인상적이다. 즉, 이 논문은 생성형 AI 가속기의 핵심 경쟁축이 더 이상 단순한 MAC 효율 향상이 아니라, decoding 단계의 메모리 병목 완화와 시스템 수

준의 실행 최적화로 이동하고 있음을 대표적으로 보여주는 사례라고 할 수 있다. 다만 구조가 다소 복합적이어서 ReRAM 적층과 quantization, scheduling이 모두 함께 맞물려야 한다는 점에서 범용성이나 구현 복잡도 측면의 부담은 남아 있다.



[그림 1] "31.1. ReRAM-on-Logic Stacked Outlier-Free Large-Language-Model Accelerator" 전체 아키텍처 구조

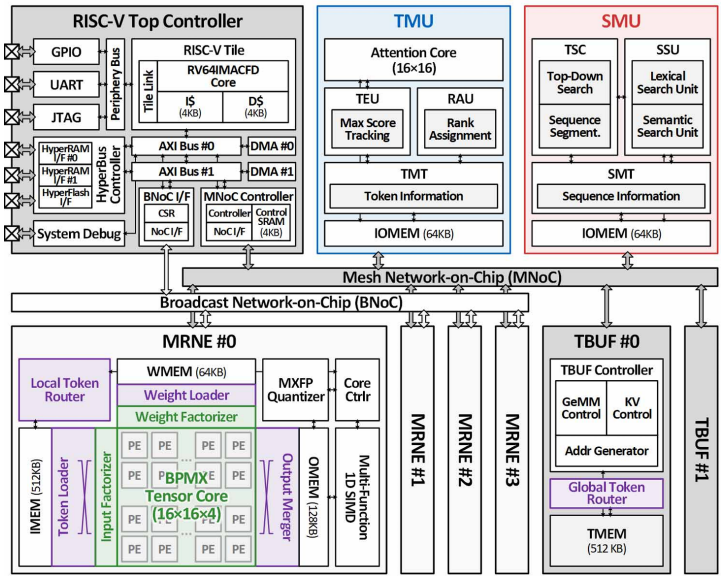


[그림 2] "31.2. Low-Bit GenAI Accelerator for Distilled-Model and CoT" 전체 아키텍처 구조

#31-2는 KAIST에서 발표한 28nm 기반 저비트 생성형 AI 가속기로, 이번 세션에서 두드러졌던 rotation 기반 양자화의 하드웨어화와 phase-aware 정밀도 제어 흐름을 가장 잘 보여주는 논문이다. 최근 LLM과 diffusion model의 저비트화에서는 단순히 bit 수를 낮추는 것만으로는 정확도 저하를 막기 어렵고, 특히 activation outlier와 group-wise scaling overhead를 함께 다뤄야 한다는 점이 중요한 과제로 떠오르고 있다. 특히 LLM에서는

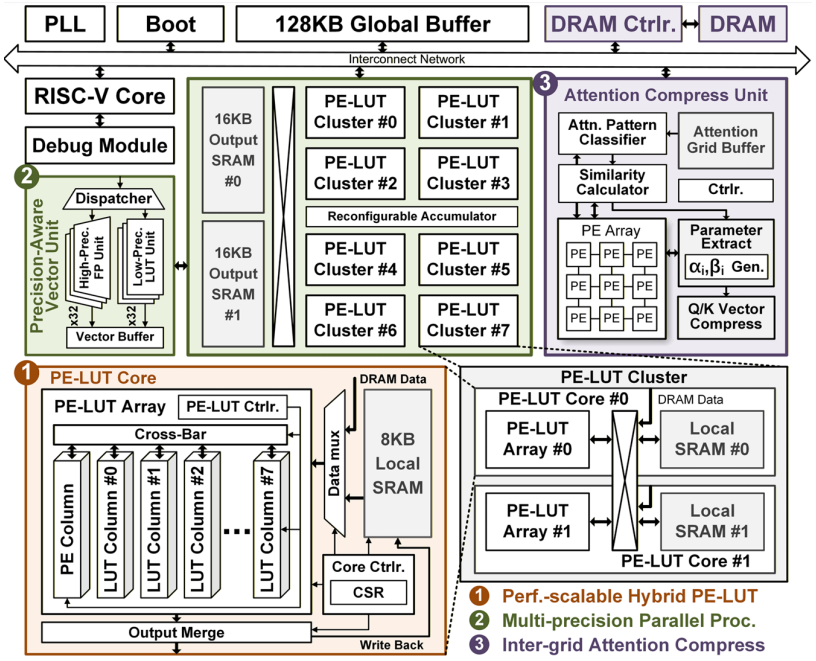
prefill과 decoding의 특성이 다르고, diffusion model에서도 early/late iteration의 민감도가 다르다는 점에 착안하여 phase별로 정밀도를 다르게 적용한 점은, 최근 GenAI 가속기가 "고정된 하나의 bit-width"에서 벗어나 실행 단계에 따라 precision을 적응적으로 조절하는 방향으로 진화하고 있음을 잘 보여준다. 또한 rotation 역시 단순 알고리즘 기법으로 남겨두지 않고, non-power-of-two 차원에서의 구현 부담을 줄이기 위한 local rotation과 permutation을 통해 실제 하드웨어 친화적인 형태로 재구성하였으며, 그 결과 scaling factor 정밀도와 dequantization 비용까지 함께 낮추려 한 점이 돋보인다. 결국 이 논문은 단순한 저비트 LLM accelerator라기보다, reasoning이 포함된 distilled model, multi-turn chat, diffusion workload까지 포괄하는 unified GenAI accelerator를 지향하면서, 앞으로의 생성형 AI 하드웨어가 단순 압축이 아니라 workload 특성, 양자화 방식, 데이터패스 구조를 함께 설계하는 공동 최적화 문제로 이동하고 있음을 보여주는 대표적인 사례라고 할 수 있다.

#31-5는 KAIST에서 발표한 28nm 기반 모바일 intelligence SoC로, 이번 세션에서 두드러졌던 on-device personalization 흐름을 가장 직접적으로 보여주는 논문이다. 많은 생성형 AI 가속기들이 여전히 서버 환경에서의 추론 효율 향상에 집중하는 반면, 이 논문은 개인화된 LLM 서비스가 실제 모바일 기기 안에서 동작하려면 무엇이 필요한지를 정면으로 다루고 있다는 점이 인상적이다. 특히 단순 inference만이 아니라, 대화 이력을 활용한 RAG 기반 user interaction과 사용자 피드백을 반영한 fine-tuning 기반 user adaptation을 모두 하나의 칩에서 지원함으로써, 생성형 AI 하드웨어의 응용 축이 클라우드 추론에서 개인화된 온디바이스 지능으로 확장되고 있음을 잘 보여준다. 이를 위해 mixed-rank token processing, similarity-aware sequence processing, 그리고 저전력 MX tensor core와 같은 구조를 함께 제안하였는데, 이는 personalization이 단순히 모델을 올리는 문제가 아니라, 검색·추론·적응을 모두 아우르는 시스템 수준 최적화 문제임을 보여준다. 물론 1B급 compact LLM을 전제로 했다는 점에서 아직 대형 모델 자체를 온전히 대체하는 방향이라고 보기는 어렵지만, 사용자 맞춤형 서비스 관점에서는 오히려 이러한 경량화와 시스템 통합이 더 현실적인 방향일 수 있다. 즉, 이 논문은 생성형 AI 가속기의 경쟁이 더 이상 "얼마나 큰 모델을 얼마나 빠르게 돌리느냐"에만 있지 않고, privacy, responsiveness, personalization을 실제 단말 환경에서 어떻게 구현할 것인가로 이동하고 있음을 대표적으로 보여주는 사례라고 할 수 있다.



[그림 3] "31.5. A 9.8mW Mobile Intelligence System-on-Chip" 전체 아키텍처 구조

#31-4는 북경대학교에서 발표한 22nm 기반 visual autoregressive 생성 가속기로, 이번 세션에서 드러난 생성형 AI 하드웨어의 적용 범위 확장을 잘 보여주는 논문이다. 최근 생성형 AI 가속기 연구는 주로 텍스트 LLM이나 diffusion model에 집중되어 왔는데, 이 논문은 이미지 생성을 위한 visual autoregressive model을 직접 겨냥했다는 점에서 차별성이 크다. 특히 fixed low-bit quantization만으로는 품질 저하가 커질 수 있다는 문제를 인식하고, performance-scalable hybrid PE-LUT, runtime multi-precision processing, grid-similarity 기반 attention compression을 함께 제안함으로써, 생성 모델에서도 단순 저비트화보다 모델 특성에 맞춘 정밀도 제어와 attention 최적화가 중요하다는 흐름을 보여주었다. 또한 인접 grid 간의 유사성을 활용해 attention map을 압축한 접근은, 시각 생성 모델에서 연산량 자체뿐 아니라 attention이 유발하는 메모리와 데이터 이동 비용이 중요한 병목이 되고 있음을 잘 드러낸다. 결국 이 논문은 생성형 AI 가속기의 초점이 텍스트 중심 LLM을 넘어 시각 생성, 나아가 다양한 멀티모달 모델로 확장되고 있으며, 그 과정에서 hardware design도 단일 MAC 효율보다 precision scalability와 structure-aware compression을 함께 고려하는 방향으로 진화하고 있음을 보여주는 사례라고 할 수 있다.



[그림 4] "31.4. A Visual Autoregressive Generation Accelerator" 전체 아키텍처 구조

저자정보

김상진 교수

- 소 속 : 광주과학기술원 전기전자컴퓨터공학과
- 연구분야 : AI Semiconductor & Digital Architecture
- 이 메 일 : sangjinkim@gist.ac.kr
- 홈페이지 : https://sites.google.com/view/symple-lab

2026 International Solid-State Circuits Conference

(ISSCC) Review

포항공과대학교 반도체대학원 박사과정 박은빈

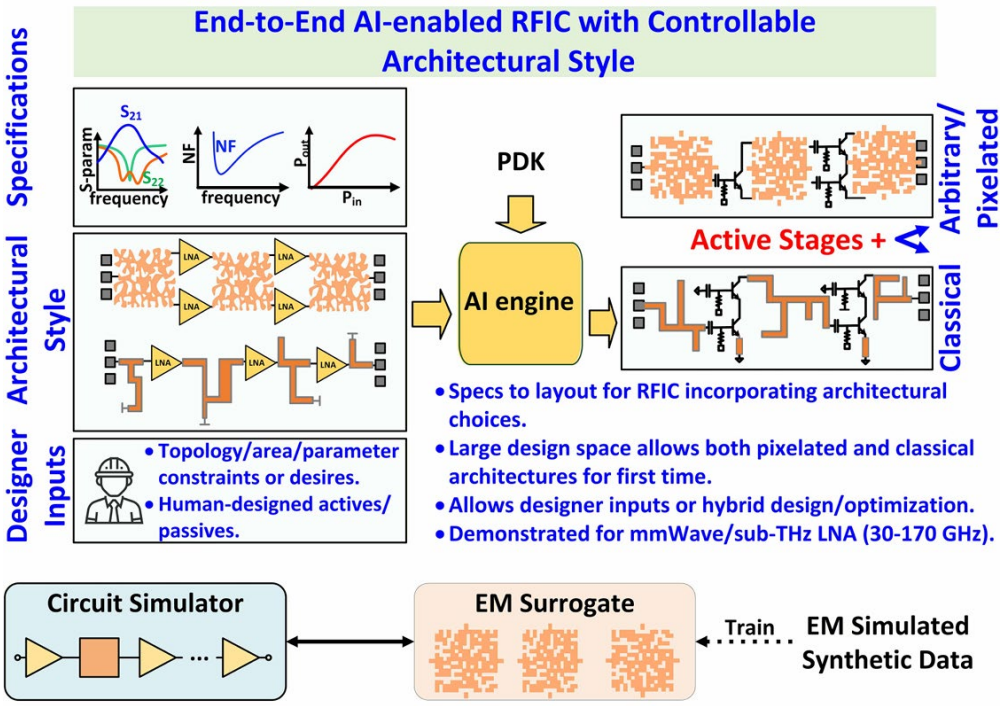
Session 13 Circuits for AI and AI for Circuits

ISSCC 2026의 Session 13에서는 AI를 위한 회로와 회로를 위한 AI라는 두 방향의 흐름을 중심으로 총 다섯 편의 논문이 발표되었다. 이 세션은 한편으로는 대규모 추천 시스템과 조합 최적화 문제를 가속하기 위한 AI 특화 하드웨어를 다루고, 다른 한편으로는 RF/mm-wave 회로 설계를 자동화하고 기존 사람이 떠올리기 어려운 구조를 찾아내는 AI 기반 회로 설계 방법론을 조명한다. 구체적으로는 hybrid compute-in-RRAM 기반 추천 시스템 가속기, quantum-inspired analog k-SAT solver, AI-enabled end-to-end RFIC 설계 플로우, inverse-designed N-path filter, 그리고 topology-optimized wideband mm-wave LNA가 포함되어 있다.

#13-2 논문에서는 AI를 이용해 LNA를 specification 단계부터 최종 layout 단계까지 자동으로 합성하는 end-to-end RFIC 설계 플로우를 제안하였다. 기존 RF/mm-wave LNA 설계는 gain, matching, noise figure, stability, linearity, power를 동시에 만족해야 하고, topology 선택과 EM 구조 설계까지 함께 맞물려 있어 설계 기간이 수개월에 이를 만큼 복잡하다. 특히 기존 방식은 설계자가 미리 알고 있는 회로 템플릿과 직관에 크게 의존하기 때문에, 사람이 떠올리기 어려운 non-intuitive 구조를 탐색하기 어렵고, 설계 공간도 제한적이었다. 본 논문은 이러한 한계를 해결하기 위해 topology, architecture, circuit parameter, EM structure를 하나의 AI 기반 프레임워크 안에서 함께 최적화하는 RFIC synthesis flow를 제시하였으며, classical transmission-line 기반 구조와 arbitrary pixelated 구조 사이를 설계자가 직접 제어할 수 있도록 하여 해석 가능성, 디버깅 용이성, 실사용성을 함께 확보하였다. 측정 결과, 제안된 플로우로 합성한 LNA는 24–90GHz에서 16dB 이득과 3.8–6.8dB NF를 달성하였고, 또 다른 AI 합성 설계는 85GHz와 160GHz 부근에서 각각 30dB와 18dB의 peak gain을 보여 주어, AI가 실제 mm-wave/sub-THz RFIC 설계에 적용 가능함을 입증하였다.

본 논문은 먼저 designer-controllable architecture synthesis라는 점에서 기존 AI 기반 회로 설계와 차별화된다. 단순히 목표 spec을 입력해 블랙박스 형태의 결과를 얻는 것이

아니라, 설계자가 classical 구조를 유지할지, non-intuitive pixelated 구조를 허용할지 선택할 수 있고, 그 제약 안에서 reinforcement learning 기반 탐색이 수행된다. 구체적으로 RL은 stage 수, common-emitter/common-base/cascode topology, transistor sizing, biasing, interface impedance 등을 포함한 회로 변수를 탐색하며, 이후 inverse EM engine이 목표 scattering parameter를 만족하는 실제 EM 구조를 생성한다. 이 과정에는 PPO 기반 학습이 적용되었고, 약 35만 개의 circuit-EM example을 바탕으로 학습하여 output-stage layout은 5–10분, complete LNA design은 수 시간 내에 도출할 수 있도록 하였다. 또한 제안된 플로우는 gain, NF, DC power 사이의 Pareto front를 빠르게 시각화함으로써 특정 PDK에서의 설계 가능 범위를 정량적으로 탐색할 수 있다는 장점도 가진다.

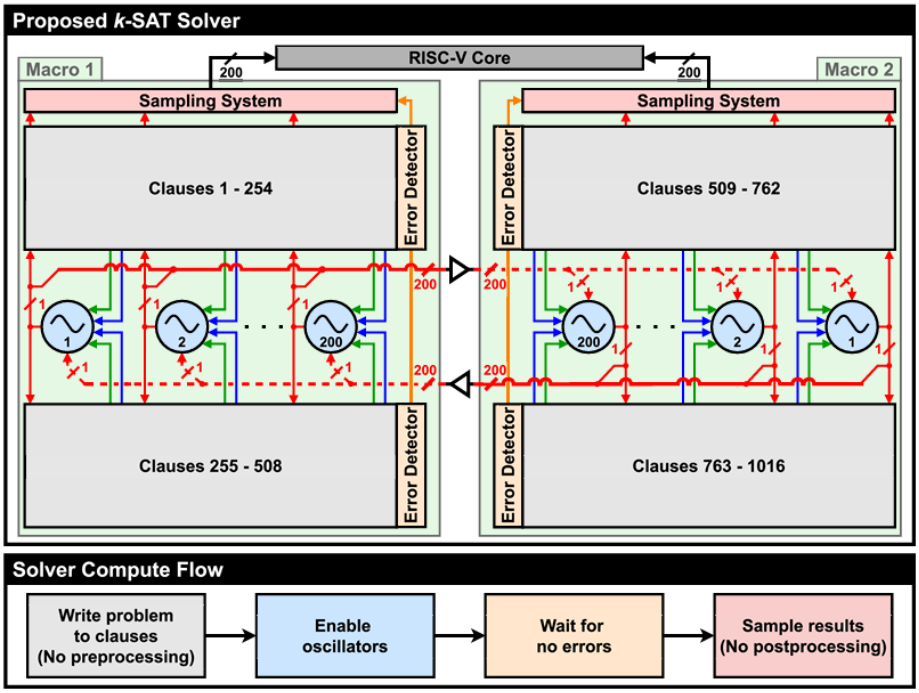


[그림 1] End-to-end AI-enabled RFIC system

논문에서는 이를 검증하기 위해 두 개의 대표적인 LNA를 제시하였다. 첫 번째는 24–90GHz frequency-diplexing broadband LNA로, 24–55GHz와 55–90GHz를 분리하는 3-port broadband matching network를 이용해 넓은 대역에서 power/noise matching을 동시에 만족하도록 설계되었다. 이 구조는 사람이 전통적 방식으로 설계하기 어려운 arbitrary multi-port pixelated EM structure를 활용하며, 측정에서 24–90GHz bandwidth, 16dB gain, 3.8–6.8dB NF를 달성하였다. 두 번째는 85–150/160GHz dual-peaking LNA로, 보다 classical한 transmission-line 기반 구조를 사용하면서도 RL이 dual resonance와 noise-optimal matching을 동시에 맞추도록 설계하였다. 이 회로는 85GHz에서 30dB peak gain과 5.8dB minimum NF를 보였고, 160GHz 부근에서도 18dB peak gain을 달성하였다. 즉

본 논문은 AI가 unconventional 구조뿐 아니라 classical 구조까지 포함하는 범용 RFIC synthesis framework로 동작할 수 있음을 보였으며, practical한 mm-wave/sub-THz LNA 설계 방법론을 제시했다는 점에서 의미가 크다.

#13-3 논문에서는 조합 최적화 문제의 대표 예인 k-SAT를 빠르고 에너지 효율적으로 해결하기 위한 quantum-inspired analog solver를 제안하였다. SAT 및 k-SAT 문제는 인공지능, 스케줄링, 자동 설계, 신약 탐색 등 다양한 분야에서 핵심적인 역할을 하지만, NP-Complete 문제이기 때문에 기존 von Neumann 기반 디지털 시스템으로는 대규모 문제를 빠르게 해결하기 어렵다. 기존 실리콘 기반 SAT solver들은 대부분 3-SAT에만 제한되거나, extensive preprocessing이 필요하거나, 문제 크기가 50 variables 수준에 머무르는 한계가 있었다. 본 논문은 이러한 한계를 해결하기 위해 mixed k-SAT를 직접 지원하면서도 최대 200 variables와 1016 clauses까지 확장 가능한 mixed-signal analog solver 'Medusa'를 제안하였다. 28nm CMOS로 구현된 프로토타입은 50-variable 3-SAT benchmark에서 평균 4.92μs solution time과 19.1nJ energy를 달성하였고, 100% solvability와 accuracy를 보이면서 기존 state-of-the-art 대비 에너지 효율은 3.5배, 시간은 3배 개선된 성능을 입증하였다. 또한 200-variable uf200-860 benchmark에서도 30.9ms의 solution time을 보여, 더 큰 문제 크기에서도 확장 가능성을 확인하였다.



[그림 1] 제안된 k-SAT Solver 구조

본 논문의 핵심은 먼저 large and mixed k-SAT를 직접 처리할 수 있는 회로 구조에 있다. 제안된 solver는 각 binary variable을 하나의 GRXO-based spin에 일대일로 대응시키고,

clause들은 mixed-signal feedback network로 구현하여 spin 상태를 직접 교란하는 방식으로 해를 탐색한다. 특히 기존 방식과 달리 추가적인 ADC나 다수의 DAC를 거치지 않고, analog feedback current를 GRXO capacitor에 직접 주입함으로써 feedback path의 지연과 에너지 소모를 줄였다. 또한 본 논문은 software SAT solver의 개념을 하드웨어적으로 가져온 Make/Break feedback을 함께 사용한다. clause가 만족되지 않았을 때는 Make current를 인가해 관련 spin이 상태를 바꾸도록 유도하고, 반대로 clause가 하나의 variable에 의해서만 겨우 만족되는 경우에는 Break current를 인가해 해당 spin이 쉽게 바뀌지 않도록 억제한다. 이 Break feedback은 시스템의 불안정한 oscillation을 줄이고 최종 만족 상태로 수렴하도록 돕는 역할을 하며, 50-variable 3-SAT 문제에서 평균 solution time을 30배 개선하는 효과를 보였다.

또 다른 중요한 기여는 확장성을 위한 distributed clause logic과 inter-macro coupling 구조이다. 각 clause는 최대 200개의 spin과 연결될 수 있도록 programmable clause cell들로 구성되며, variable selection, negation, feedback 기능이 fully decentralized하게 분산 구현된다. 이를 통해 임의의 clause size를 갖는 mixed k-SAT 문제를 지원하면서도 all-to-all connectivity를 유지할 수 있다. 또한 두 개의 compute macro를 디지털 방식으로 강하게 결합하여 대응되는 spin들이 동일한 상태를 유지하도록 만들었고, 이를 통해 단일 macro의 한계를 넘어 더 큰 문제 크기로 확장할 수 있도록 하였다. 여기에 current summation line의 parasitic delay를 줄이기 위해 active-cascode termination, clause bank splitting, selective load disconnection 같은 회로 및 레이아웃 최적화 기법을 함께 적용하여, 대규모 mixed-signal feedback 구조에서 속도 저하를 억제하였다. 결과적으로 제안된 2-macro 시스템은 2.59mm² active core area 안에서 200-variable, 1016-clause 문제를 처리할 수 있었으며, 별도의 pre-processing이나 post-processing 없이 실제 benchmark에서 높은 정확도와 빠른 수렴 속도를 달성하였다. 즉 본 논문은 quantum-inspired computing의 아이디어를 CMOS mixed-signal 회로로 현실화하여, 기존 3-SAT 중심 solver보다 더 큰 mixed k-SAT 문제를 직접 다룰 수 있는 practical combinatorial optimization accelerator를 제시했다는 점에서 의미가 크다.

저자정보

박은빈 박사과정 대학원생

- 소속 : 포항공과대학교
- 연구분야 : HW-SW co-optimization 및 양자오류정정부호
- 이메일 : eunbin@postech.ac.kr, eunbin.epiclab@gmail.com
- 홈페이지 : https://sites.google.com/view/epiclab/member/ebpark

2026 International Solid-State Circuits Conference

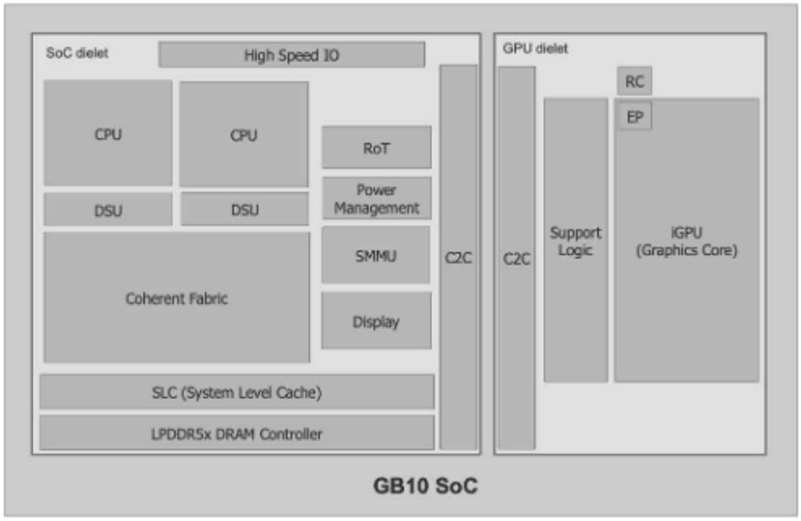
(ISSCC) Review

연세대학교 전기전자공학과 석박통합과정 김동욱

Session 17 Highlighted Chip Releases for AI

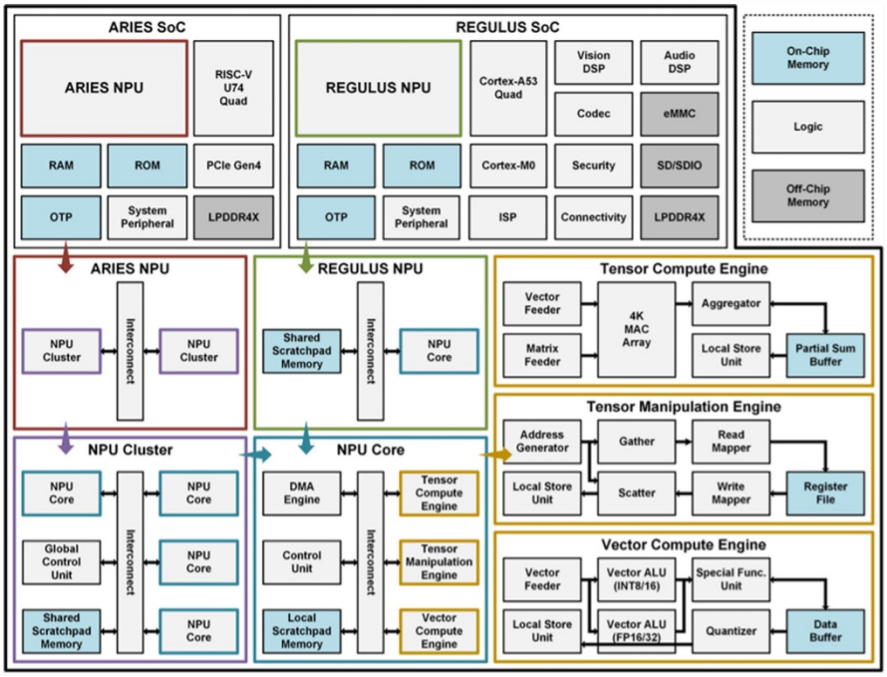
2026 ISSCC의 Session 17은 Highlighted Chip Releases for AI라는 주제로 총 4편의 연구가 발표되었다. Session은 발전하는 AI 기술을 뒷받침할 산업계의 하드웨어 솔루션 연구에 중점을 두고 있으며, 전통적인 데이터센터 향 AI chip은 물론이고 desktop AI super computer를 위한 chip과 edge 디바이스를 위한 기술도 선보였다.

#17-1 NVIDIA 연구진이 발표한 본 논문 [NVIDIA GB10: SoC Built for AI Acceleration]은 데스크탑 환경에서 데이터센터급 성능을 구현하기 위한 새로운 AI 가속기 SoC인 GB10을 제안했다. 기존 AI 개발은 고가의 데이터센터 하드웨어 또는 클라우드 환경 인프라에 의존해야 했기에 비용과 접근성 측면에서 문제가 있었다. 본 연구에서는 이를 해결하기 위해 표준 콘센트 전력만으로도 구동 가능한 DGX Spark 워크스테이션을 위한 하드웨어 솔루션을 제시했다. GB10 SoC는 NVIDIA의 고성능 GPU 설계와 MediaTek의 저전력 CPU 및 메모리 시스템 기술을 결합했고, TSMC 3nm 공정 기반의 dual-dielet 구조를 사용했다. 구체적으로는 CPU 다이에 ARM v9.2 코어 20개가 두 개의 클러스터로 나뉘어 존재하고 16MB L3 캐시를 클러스터가 공유한다. GPU 다이는 Blackwell 아키텍처 기반 iGPU가 있으며 6144개의 CUDA 코어와 5세대 Tensor 코어, 4세대 RT 코어 및 24MB의 전용 L2 캐시를 포함한다. 두 다이는 NVLink-C2C 인터페이스를 통해 600GB/s의 대역폭을 가지며 연결된다. 또한, Address Translation Services (ATS)를 통해 하드웨어 수준에서 coherency를 제공하여 통합된 메모리 공간을 효율적으로 사용할 수 있도록 지원한다. 메모리로는 고가의 HBM 대신 128GB의 LPDDR5x를 채택하였고, 301GB/s의 대역폭과 전력 및 비용 효율성을 확보했다. 결과적으로, 최대 200B 파라미터 규모 모델의 inference 및 70B 모델의 fine-tuning이 가능하게 했고 5세대 Tensor Core를 통해 FP4 기준 1PFLOPS, FP32 기준 31TFLOPS의 연산 성능을 달성했다.



[그림 1] Dual-dielet architecture의 GB10 SoC

#17-3 Mobilint 연구진이 발표한 본 논문 [ARIES and REGULUS: A Unified and Scalable Hardware-Software Co-Designed NPU SoC Family for On-Device and On-Premises Multimodal Inference]는 다양한 AI 워크로드를 효율적으로 실행하기 위한, 확장 가능한 NPU 아키텍처와 이를 적용한 두 종류의 SoC(ARIES, REGULUS)를 제안했다. 최근 on-device AI의 확산으로 연산 중심 모델뿐 아니라 메모리 집약적인 LLM을 모바일의 제한된 전력 및 대역폭 하에서 효율적으로 처리해야 하는 과제가 대두되었다. 본 연구에서는 이를 해결하기 위해 HW-SW co-design 측면의 솔루션 세 가지를 제시했다. 첫째, 정확도 손실을 최소화하면서 메모리 효율을 극대화하고자 mixed-precision 양자화 기법을 도입했다. 데이터의 분포에 최적화된 맞춤형 양자화를 수행한 것이고, 정확도를 떨어뜨리는 outlier는 고정밀도로, 비 균일한 데이터는 전용 Look-Up-Table을 통해 효율적으로 압축했다. 둘째, 복잡한 activation function을 곡률에 따라 근사 정도를 다르게 하는 방식을 적용해 연산 비용을 낮췄다. 셋째로, 단일 코어에서 멀티 칩 시스템까지 확장 가능한 통합 프로그래밍 모델을 구축했다. 하드웨어 구현 측면에서 NPU 코어는 행렬 연산을 담당하는 TCE(4K MACs 탑재), 데이터 레이아웃 변환을 위한 TME, 벡터 연산을 수행하는 VCE 등으로 구성했다. 8개 코어의 ARIES(Samsung 14nm)와 싱글 코어의 REGULUS(TSMC 12nm)를 통해 유연한 확장이 가능함을 보였고, INT8 기준 각각 25W TDP에서 80 TOPS와 5W TDP에서 10TOPS의 성능을 달성했다. 실측 결과에서 ARIES는 GPU보다 높은 전력 및 대역폭 효율성을 보였고, REGULUS는 edge 및 모바일 환경에서 중요한 성능 지표인 빠른 inference 속도를 달성해 보였다.



[그림 2] ARIES, REGULUS의 SoC 아키텍처와 NPU microarchitecture

#17-4 Microsoft 연구진이 발표한 본 논문 [Maia: A Reticle-Scale AI Accelerator]는 LLM 워크로드에 최적화된 차세대(2세대) AI 가속기 Maia 200의 아키텍처와 구현 결과를 선보였다. 본 연구에서는 AI 가속기 설계에 있어 시스템, 네트워킹, 그리고 소프트웨어를 수직적으로 통합하여 구현한 제품을 제시한다. 기존 Maia 100(1세대)이 메모리 대역폭과 연산 성능 확장 측면에서 최신 LLM 모델을 감당하기 어려워졌고, 이에 따라 2세대 제품에서 추론 및 합성 데이터 생성(synthetic data generation) 워크로드에서 비용 대비 성능을 업계 최고 수준으로 끌어올렸다고 소개한다. Maia 200은 레티클(reticle) 크기의 다이이고, TSMC 3nm 공정 및 CoWoS-S 패키징을 통해 제작되었다. 칩 내부 아키텍처에서 가장 기본은 타일이고, 각 타일은 행렬 연산을 위한 Tile Tensor Unit (TTU)과 BF16을 위한 256-way SIMD unit인 Tile Vector Processor (TVP)로 구성되며 FP8, FP6, FP4 등 다양한 narrow data type을 지원하여 연산 효율을 높였다. 또한, 19개 metal layer 기반 PDN과 전략적으로 배치한 bump pads가 다이 전역에 안정적으로 전류 공급과 신호 무결성 확보에 매우 큰 역할을 했다고 언급한다. 메모리 측면에서는 6개의 HBM3e stack을 통해 최대 7TB/s의 대역폭을 확보함과 동시에 대용량 on-die SRAM을 활용했다. 결과적으로 Maia 200은 이전 세대 Maia 100 대비 4배의 메모리(HBM) 대역폭을 달성했고, Azure의 생성형 AI 추론 워크로드에서 3~5배의 비용 대비 성능 향상을 달성하여 하이퍼스케일 데이터센터 환경에 적합한 칩임을 보였다.

	Maia 100	Maia 200
Peak Dense Tensor TOPS	6bit: 3000 9bit: 1500 BF16: 800	FP4: 10145 FP8: 5072 BF16: 1268
Peak HBM BW (TB/s)	1.8TB/s	Up to 7TB/s
Num of HBM Stack	4	6
HBM Capacity (GB)	64	216
Host PCIe BW (GB/s)	32	64
PCIe Configuration	Gen4 x16/Gen5 x8	Gen6 x8
Backend Network Bandwidth (GB/s) (bi-directional)	1200	2800
Backend Network Configuration	12x400	28x400

[그림 3] Maia 100(1 세대)과 Maia 200(2 세대)의 성능 비교

저자정보

김동욱 석박통합과정 대학원생

- 소속 : 연세대학교
- 연구분야 : 메모리 시스템
- 이메일 : dwkim3852@yonsei.ac.kr
- 홈페이지 : https://dtl.yonsei.ac.kr