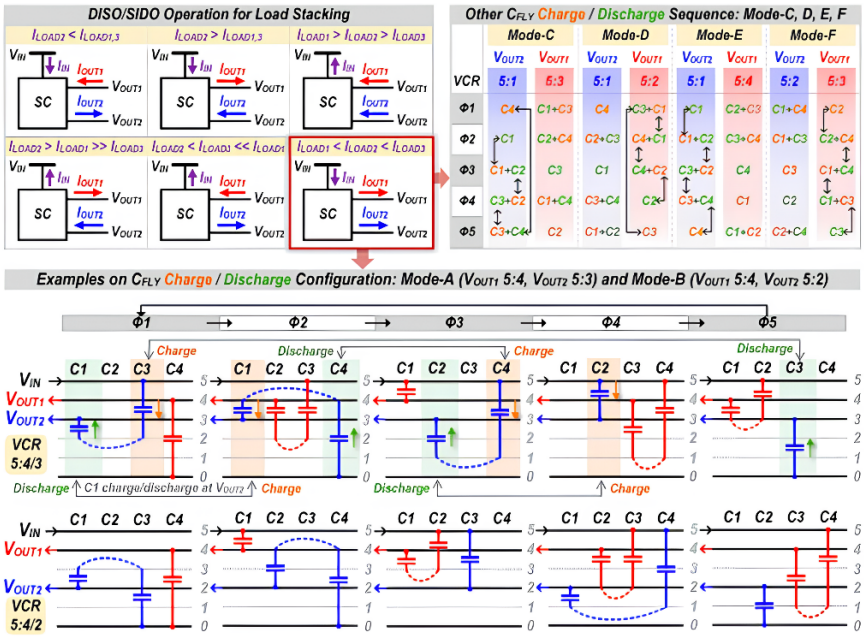
# A-SSCC 2025 Review

KAIST 전기 및 전자공학부 박사과정 박수언

## Session 2 Switching-Based Power Converters

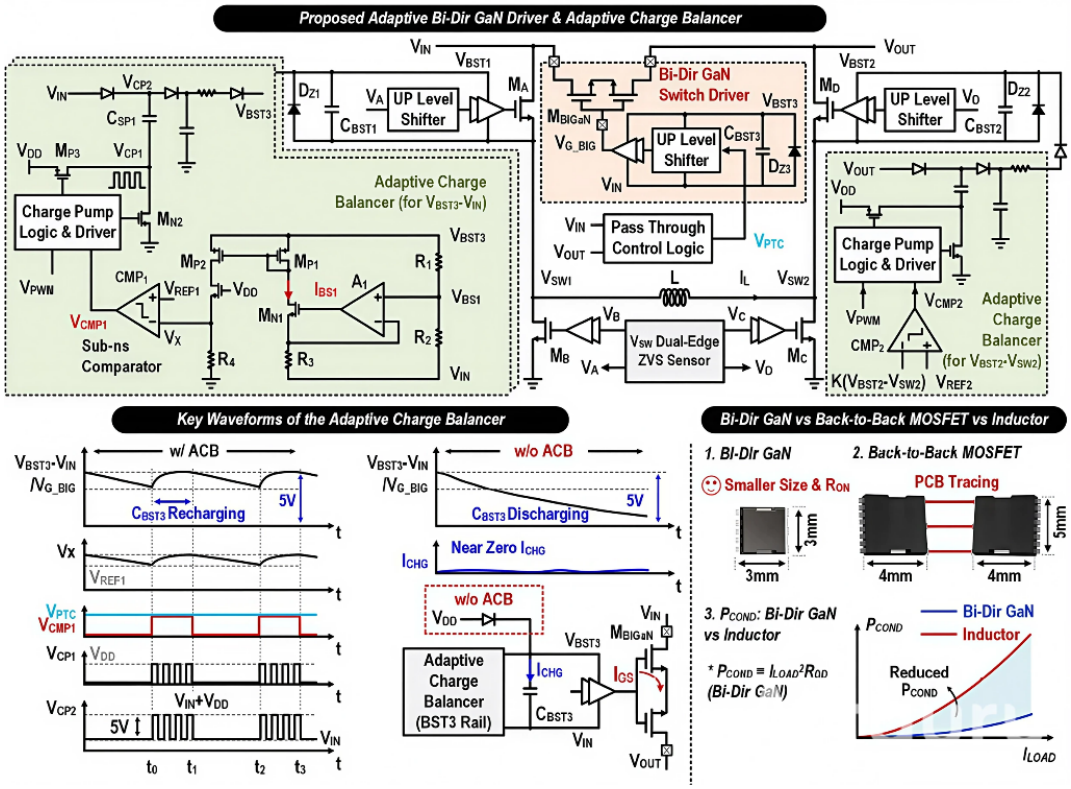
이번 2025 A-SSCC의 Session 2에서는 차세대 Switching-Based Power converter 기술을 중심으로 총 4편의 논문이 발표되었다. 고효율, 안정성을 목표로 한 다양한 기술들이 소개되었으며, 2025년에는 Load-stacking의 cross-regulation 완화 및 고속 DVS 기능을 갖춘 SIDO/DISO 및 IVR의 안정도 조정 기법 등 차세대 기술을 다룬 논문들이 채택되었다.

**#2-2** 논문은 Macau 대학에서 발표한 논문으로, Load-stacking 시스템을 위한 Cross-Regulation (상호 간섭) 억제 및 고속 DVS를 지원하는 SIDO/DISO SC 컨버터를 설계한 것이다. Load-stacking은 시스템 효율을 높이지만, 부하 전류의 Series-연결, 공유로 인해 출력 간의 상호 간섭 문제와 느린 DVS 속도가 해결해야할 사항이다. 제안한 핵심 기술은 각 출력 VOUT1,2 에 대해 플라잉 커패시터의 충/방전이 독립적인 시퀀스로 완료되도록 하여 출력 간 상호 간섭을 억제한다. 또한, 부하 불균형 발생 시 커패시터를 활용한 즉각적인 전력 재할당을 모드 변환을 통해 이루어내며 4개의 플라잉 커패시터로 6개의 VCR 모드를 지원하여 빠른 DVS 속도 및 시스템 효율을 개선하였다. 결과적으로, Cross Regulation을 기존 대비 약 6배 개선 및 46.6mV/ns의 빠른 응답 속도를 달성하였다.




[그림 1] 독립적 플라잉 커패시터 충/방전을 이용한 출력 상호간섭 억제 및 전력 재할당 동작

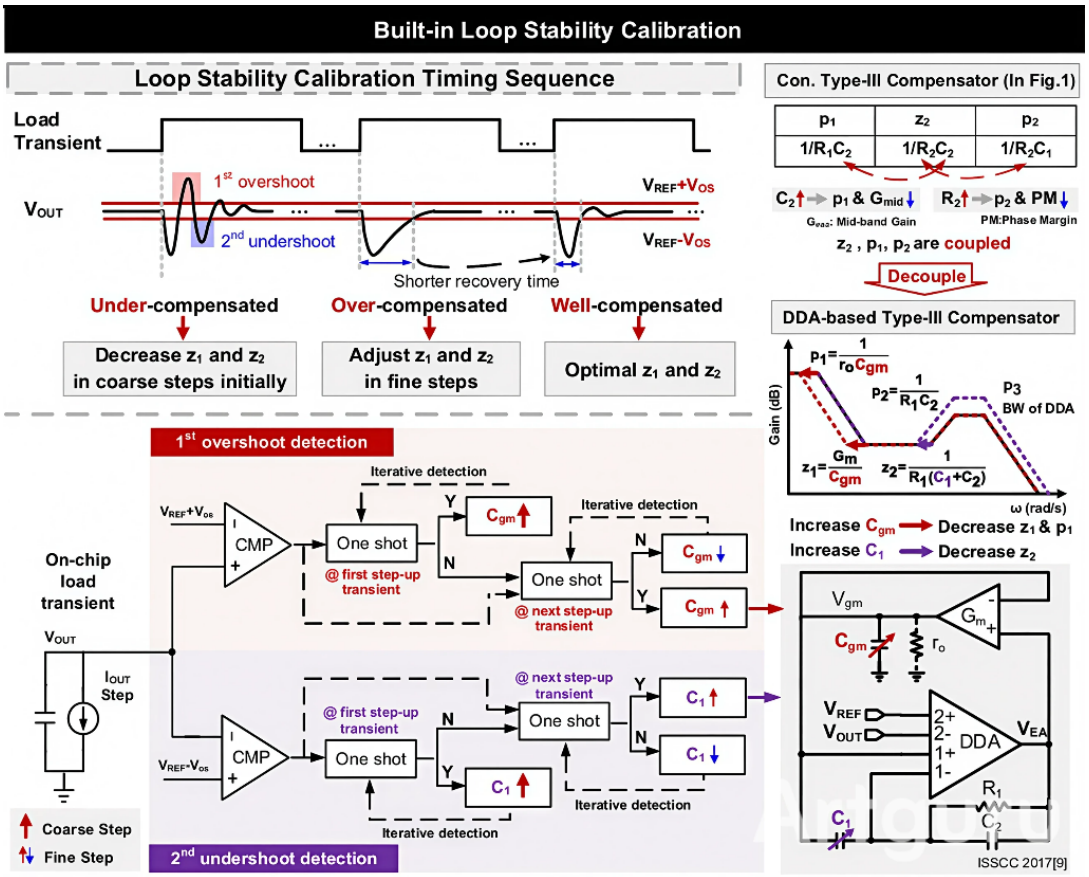
**#2-3** 논문은 이동형 태양광 패널을 위한 Soft-switching이 가능한 Bi-directional 벅-부스트 컨버터 논문이다. 벅-부스트 컨버터는 입, 출력 전압이 비슷한 상황에서 High-Side (HS) NMOS 스위치를 계속 켜 두면 부트스트랩 커패시터의 방전으로 시스템이 꺼지는 위험이 존재한다. 또한, 하드 스위칭으로 발생하는 스위칭 손실은 효율을 크게 떨어뜨린다. 제안한 핵심 기술은 입, 출력전압이 같은 상황에서 부트스트랩 전압을 모니터링하고, 부족 시 차지 펌프 동작으로 재충전하여 HS 스위치의 always-on을 유지하여 시스템-off를 방지한다. 또한, 스위칭 노드 전압의 순간적인 변화를 검출하며, 4개의 GaN 스위치를 추가적인 소자 사용 없이 ZVS 턴-온을 가능하게 한다. 그 결과 3~65V의 넓은 입력 전압 범위에서 300W의 출력을 내며 ZVS 동작으로 Peak 97.8% 효율 달성 및 입, 출력 전압이 비슷한 상황에서의 안정적인 전압 제어를 가능하게 하였다.




[그림 2] 입, 출력 전압이 비슷한 상황에서의 부트스트랩 전압 모니터링/유지 회로

**#2-4** 논문은 Loop Stability Calibration 기능 및 개선된 전류 센싱 기능을 갖춘 6-Phase IVR 논문이다. IVR은 특성 상 패시브 소자 (L, C)의 공정 편차로 인해 제어 루프가 불안정해질 수 있으며, 이를 외부 장비 없이 안정도를 보정하기 어렵고, CPU의 부하량을 확인할 수 있는 기존 전류 센서는 스위칭 노이즈와 센스-앰프의 대역폭 제한으로 큰 오차를 갖는 문제가 있다. 제안한 핵심 기술은 Load-Transient 시 발생하는 over/undershoot을 감지하여, Type-III 보상기의 zero 위치를 자동적으로 튜닝하여 안정성을 확보한다. DDA를

이용한 type-III 보상기로, zero와 pole 분리를 통해 zero 주파수만을 이동시키도록 하여 제어 안정도를 조정할 수 있는 장점이 있다. 또한, 오프-타임의 1/2 지점을 타겟팅/샘플링하여 100ps 지연에도 1% 미만의 오차를 달성하는 전류 센싱 정확도를 달성하였다. 그 결과 안정적인 출력 전압 regulation 및 기존 전류 센싱 대비 24.4%의 3-sigma 정확도를 개선하는 결과를 얻었다.




[그림 3] zero/pole 분리된 DDA-Based Type-III Compensator 및 안정도 조정 알고리즘

## 저자정보


**박수언 박사과정 대학원생**

- 소속 : KAIST
- 연구분야 : Power Management IC 설계
- 이메일 : tndjs12221@kaist.ac.kr
- 홈페이지 : https://icdesignlab.net/students

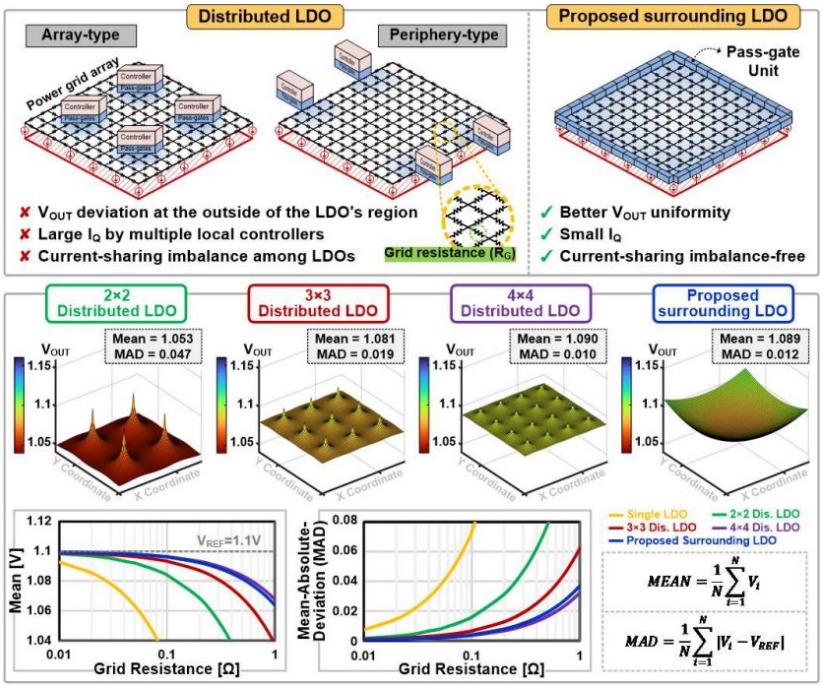
# A-SSCC 2025 Review

고려대학교 전기전자공학과 박사과정 이윤호

## Session 5 Monitoring, Regulation, and References

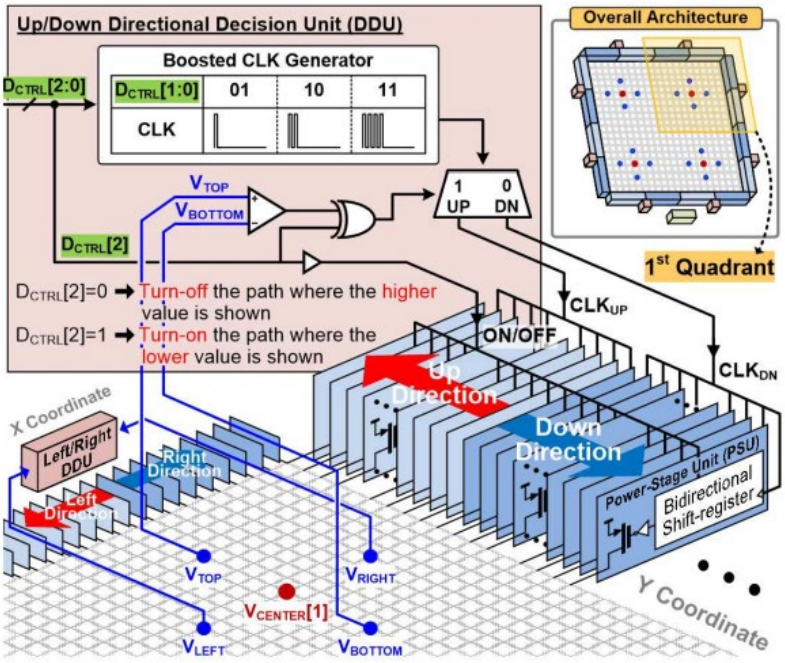
이번 A-SSCC 2025의 Session 5는 모니터링 및 레귤레이션 등과 관련된 총 5편의 논문이 발표되었다. 본 논문 리뷰에서는 이들 중 surrounding pass-gate를 이용한 디지털 분산형 LDO와 GaN HEMT의 특성을 이용한 하이브리드 LDO 논문에 대해 알아보고자 한다.

**#5-2** 본 논문은 고성능 컴퓨팅(HPC) 등의 발전으로 디지털 시스템의 코어(Core) 수가 증가함에 따라 전력 소모가 급증하는 문제에 대응하여, 전력 관리의 효율성을 극대화하고 칩 내부의 국부(Local) 전압 변동을 최소화하기 위한 Surrounding Pass-gate 기반의 디지털 분산형 LDO를 제시하였다. 아래 그림 1과 같이 기존 분산형 LDO(Distributed LDO) 구조는 LDO 간의 전류 불균형 문제와 많은 로컬 컨트롤러 사용으로 인한 높은 대기 전류 증가로 인해 전류 효율이 저하되는 문제가 있었다. 또한, 전체 출력 전압 제어가 로컬 LDO의 지정된 소수 센싱 노드에 국한되어, 센싱 되지 않은 영역에서 국부적인 전압 편차가 크게 발생하는 근본적인 한계가 있었다.



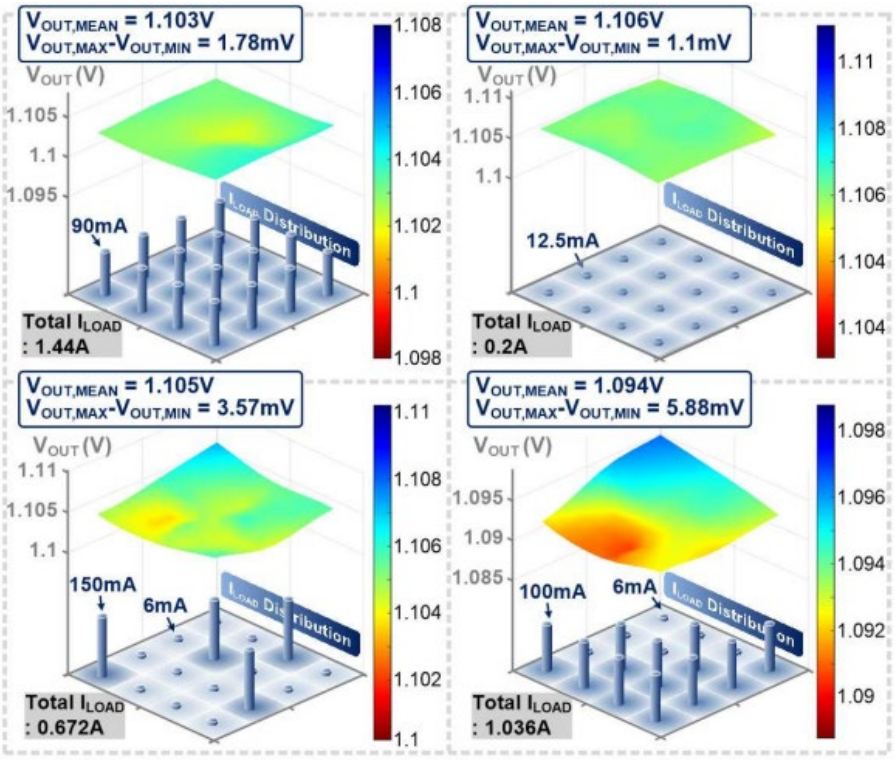

[그림 1] 기존의 분산형 LDO 구조와 제안하는 방식의 LDO 구조 간의 차이(위) 및 국부(local) 전압 불균형에 따른 출력 전압(VOUT) 편차(아래)

이러한 문제를 해결하기 위해, 본 논문은 이러한 문제를 해결하기 위해 칩 외곽을 둘러싸는(Surrounding) Pass-gate를 이용한 새로운 분산형 LDO 구조를 제안하였다. 1,600개의 파워 스테이지 유닛(PSUs)을 칩 주변에 분산 배치하고, 국부 영역에 대한 전압을 메인 컨트롤러에서 제어하는 방식을 통해 국부 전압 편차 문제를 해결하였다. 이 중 핵심은 방향 선택적 중첩 사분면 섹션 제어(Directional-Selective Nested Quadrant Section Control) 기술이다. 이 기술은 아래 그림 2와 같이 4개의 중앙 노드와 각 사분면의 방향 결정 유닛(DDU)의 센싱 전압을 활용하여, 전압 강하가 심한 방향으로만 PSU를 선택적으로 활성화함으로써 대규모 디지털 코어 영역에서도 균일하고 안정적인 출력 전압을 유지하도록 보장한다.




[그림 2] 제안하는 surrounding 방식의 분산형 LDO 구조의 동작방식을 나타내는 블록 다이어그램

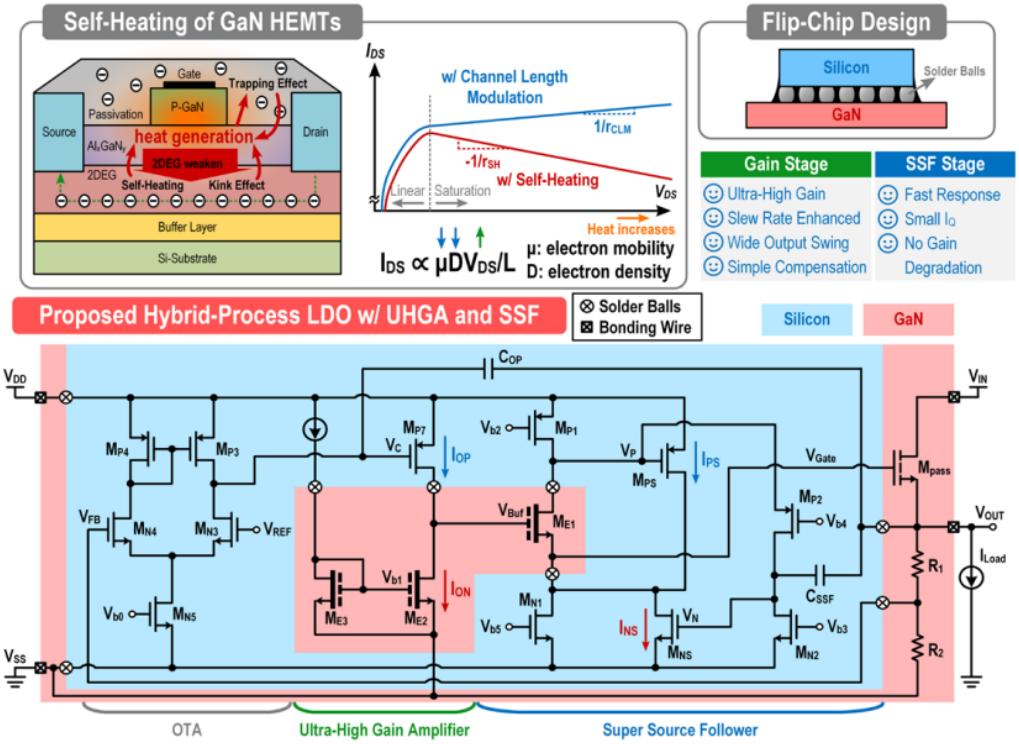
아래 그림3의 본 논문의 LDO 측정 결과와 같이, 제안하는 LDO 제어 방식을 통해 전체 출력 전류 범위에 대하여, 최대 1.78mV-to-5.88mV의 국부 VOUT 편차를 유지한다. 28nm CMOS 공정으로 제작된 제안된 LDO는 0.8V-to-1.15V의 입력에서 50mV의 낮은 dropout을 갖으며, 최대 2.7A의 출력으로 31.88A/mm2의 우수한 전류 밀도를 보여준다. 전체적인 성능지표를 나타내는 FoM (Figure-of-Merit) 또한, 출력 전류에 따라 0.77-to-2.25로 비교표 내의 최신 분산형 LDO 중 2번째로 좋은 수치를 보여준다.




[그림 3] 제안하는 분산형 LDO 구조의 부하 변화에 따른 파워 그리드 상의 VOUT 측정 결과

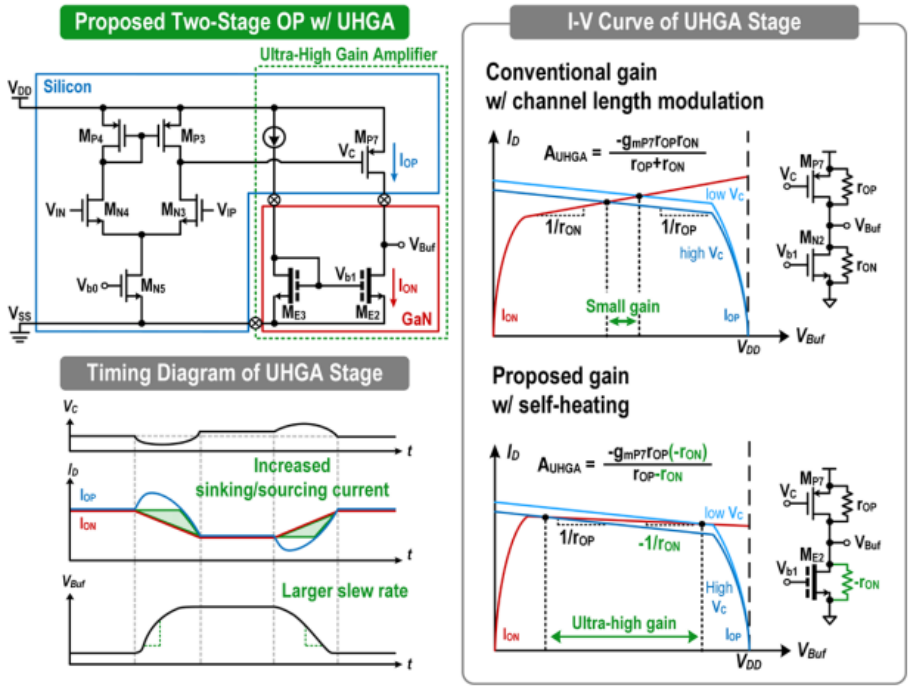
**#5-3** 본 논문은 GaN HEMT 소자의 self-heating 특성에서 기인하는 소신호 네거티브(-) 저항을 회로 이득으로 적극 활용한 하이브리드 공정의 capless LDO를 제안하였다. 본 논문은 기존 실리콘 공정 기반 LDO에서 필연적으로 요구되던 출력 커패시터 및 다단 증폭기 구조를 배제하면서도, GaN HEMT의 특성을 이용한 초고이득 루프와 빠른 과도 응답, 우수한 line 및 load regulation을 동시에 달성한 점이 특징이다. 아래 그림 4의 상단과 같이 GaN HEMT는 드레인-소스 전압이 증가함에 따라 self-heating에 의해 전자 이동도가 감소하고, 그 결과 포화 영역에서 I–V 특성이 음의 기울기를 갖는 특징이 나타난다. 본 논문에서는 이를 기존 실리콘 기반의 source-follower 구조와 조합하여 매우 큰 유효 출력 저항과 초고 DC 이득을 구현하였다.

아래 그림 4의 하단과 같이 제안된 LDO는 GaN HEMT를 이용한 ultra-high-gain amplifier(UHGA)와 super source follower(SSF)의 두 블록으로 구성된다. Pass device로는 음의 임계 전압을 갖는 depletion-mode GaN HEMT를 사용하여, 일반적인 NMOS 패스 소자에서 발생하는 VGS 드롭 문제를 제거하였다. UHGA는 eGaN 소자를 포함한 CS 구조를 기반으로 하며, 기존 실리콘 공정의 채널 길이 변조(CLM)에 의존하지 않고 높은 출력 임피던스를 형성함으로써, 54.7 dB의 단일 스테이지 DC 이득을 달성하였다.




[그림 4] GaN HEMT소자의 self-heating 특성 (위) 제안하는 하이브리드 공정의 LDO 구조 (아래)

아래 그림 5와 같이 기존의 실리콘 기반 증폭기 구조에서는 출력 전압이 변화할수록 아래쪽 NMOS의 전류도 함께 변화하여, 위쪽 PMOS와 아래쪽 NMOS의 전류가 균형을 이루는 지점에서 출력 전압이 결정된다. 이때 입력이 조금 변하더라도 균형점의 이동이 크지 않아, 출력 변화와 증폭 이득이 제한된다는 한계가 있다. 반면, 제안된 GaN 기반 구조에서는 NMOS 대신 GaN HEMT를 사용한다. GaN 소자는 전압이 증가하면 내부 발열로 인해 오히려 전류가 줄어드는 특성을 보인다. 그 결과, 작은 전압 변화에도 전류 균형점이 크게 이동하게 되고, 출력 전압이 입력 변화에 매우 민감하게 반응한다. 이러한 특징 덕분에, 기존 실리콘 구조보다 해당 노드의 slew-rate를 크게 증가시킬 수 있다.




[그림 5] 실리콘 공정 방식과 GaN HEMT 방식의 DC 특성 차이 및 UHGA의 transient 응답 특성

제안된 LDO는 GaN HEMT를 이용한 높은 이득 덕분에 0.297 mV/A의 load regulation과 0.024 mV/V의 line regulation을 달성하여 multi-stage amplifier 없이도 우수한 레귤레이션 특성을 보인다. 또한, 1mA-to-900mA (450 ns edge) 부하 변동에서, heavy-load 전환 시 95 mV 언더슛과 90 ns settling time을, light-load 전환 시 98 mV 오버슛과 54 ns의 빠른 복구 시간을 기록하였다.

## 저자정보

**이윤호 박사과정 대학원생**

- 소속 : 고려대학교
- 연구분야 : Power management ICs
- 이메일 : uknow@korea.ac.kr
- 홈페이지 : https://sites.google.com/site/kubasiclab/home

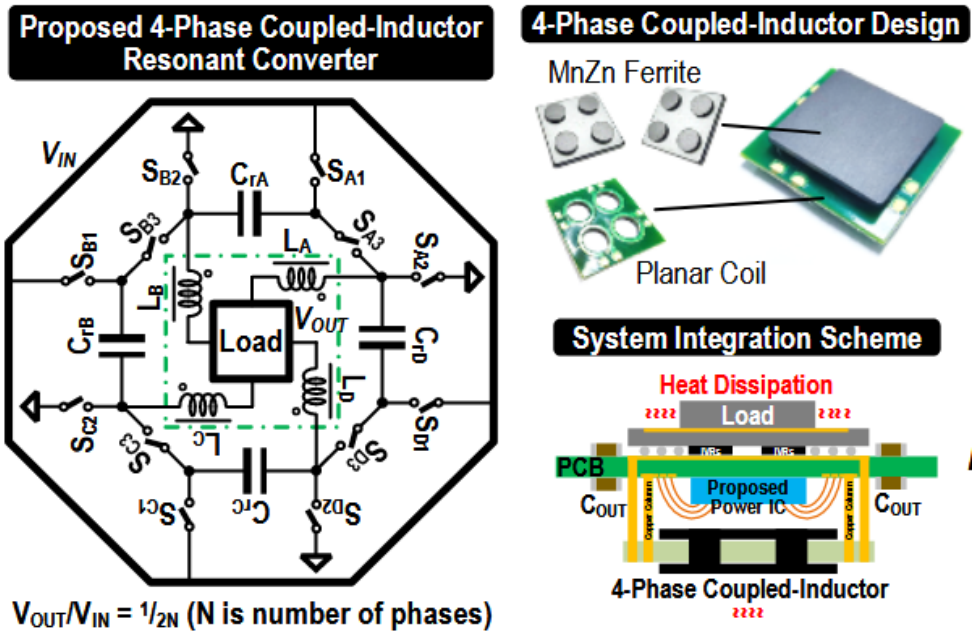
# A-SSCC 2025 Review

성균관대학교 반도체융합공학과 석사과정 이승언

## Session 14 Hybrid DC-DC Converter

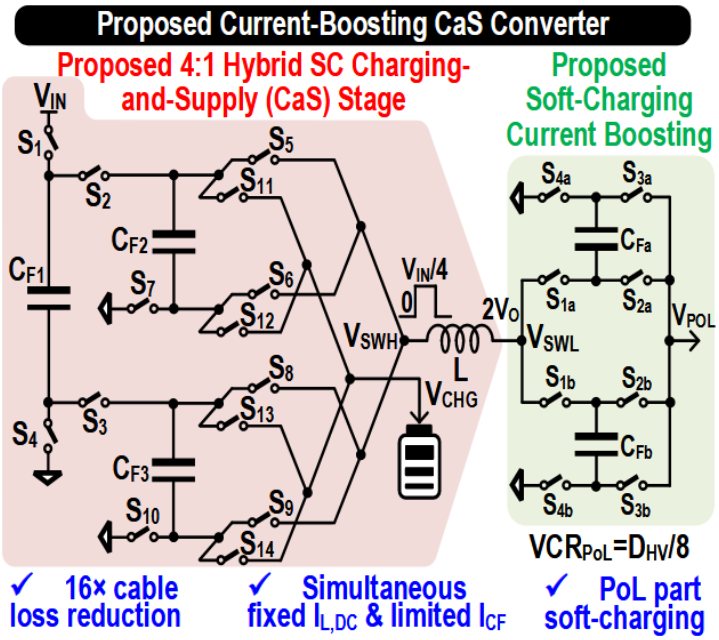
이번 ASSCC 2025의 Session 14는 Hybrid DC-DC Converter를 주제로 총 4편의 논문이 발표되었다. 이 세션에서는 다양한 PMIC 어플리케이션에서 높은 power transfer capability를 확보하는 기술뿐 아니라, 높은 VCR과 넓은 VIN 범위에서도 high efficiency를 유지하기 위한 접근들이 소개되었다. 특히 cable/inductor DCR loss, hard-charging noise, multi-output regulation까지 함께 고려한 system-level hybrid power architecture가 중요한 설계 이슈로 부각되고 있음을 잘 보여준다. 그중 14.1과 14.2는 각각 high-power 전원 시스템과 USB-PD 기반 charging 시스템을 대상으로 hybrid converter의 최신 설계 흐름을 잘 보여주는 사례라 할 수 있다.

**#14-1**은 중국 마카오 대학에서 발표한 논문으로, 8–32V 입력, 58W/15A급 SiP-integrated octuple step-down hybrid resonant converter를 제안한다. Dual-path 기반 hybrid SC 구조가 wide VCR에서 갖는 hard-charging noise와 loss 문제를 해결하기 위해, 4-phase coupled inductor를 이용한 resonant 동작으로 soft input/output conduction과 ZVS를 동시에 달성한 것이 핵심이다. 네 개의 phase-to-phase SC resonant cell을 고리 형태로 연결하고, 네 인덕터를 하나의 planar coupled-inductor 코어로 묶어 nominal VCR 8:1을 구현하며, power stage와 함께 vertical SiP integration을 적용해 보드 면적과 기생 성분을 줄였다. 측정 결과 12V 입력 3A 부하에서 92.53% peak efficiency, 32V 입력 15A 최대 부하에서 80.1% 효율을 달성해, high VIN 및 high current 영역에서도 resonant hybrid 구조가 실질적인 efficiency와 power density 이점을 제공할 수 있음을 보여준다.



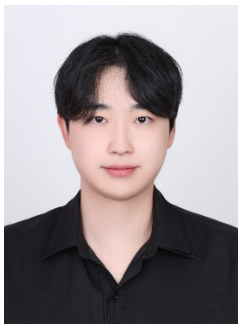

[그림 1] 14.1에서 제안한 4-Phase Coupled-Inductor 기반의 Resonant Converter

**#14-2**도 중국 마카오 대학에서 발표한 논문으로써 45 W, 12–40 V 입력에서 1S/2S battery charging과 PoL supply를 동시에 지원하는 hybrid CaS (Charging-and-Supply) converter를 제안한다. USB-PD 환경에서 케이블 저항 때문에 PCable이 PCHG를 심하게 잠식하는 문제를 짚고, 4:1 SC stage와 flying capacitor floating technique를 사용해 VCHG와 VSWH를 VIN/4로 제한함으로써 케이블 손실을 16× 줄이고 PoL 쪽은 soft-charging으로 구동하는 것이 특징이다. Duty-cycle lockout-release operation과 IL,DC current boosting 기법을 더해 wide VIN에서 fixed IL,DC와 제한된 inrush ICF를 동시에 만족하면서, 플러그 연결 시에는 케이블로부터 직접 PoL을 구동하고 동시에 배터리를 고속 충전하는 동시 CaS 동작을 구현하였다. 시스템은 94.4% peak efficiency와 90.5% PoL 변환 효율을 달성하며, USB-PD 기반 high-power charger에서 cable loss, hard-charging loss, wide-range VCR 문제를 동시에 완화하는 hybrid 아키텍처의 가능성을 보여준다.




[그림 2] 14.1에서 제안한 4-Phase Coupled-Inductor 기반의 Resonant Converter

## 저자정보



**이승언 석사과정 대학원생**

- 소속 : 성균관대학교
- 연구분야 : Power Management ICs
- 이메일 : se38lee@g.skku.edu
- 홈페이지 : https://sites.google.com/view/ecslab

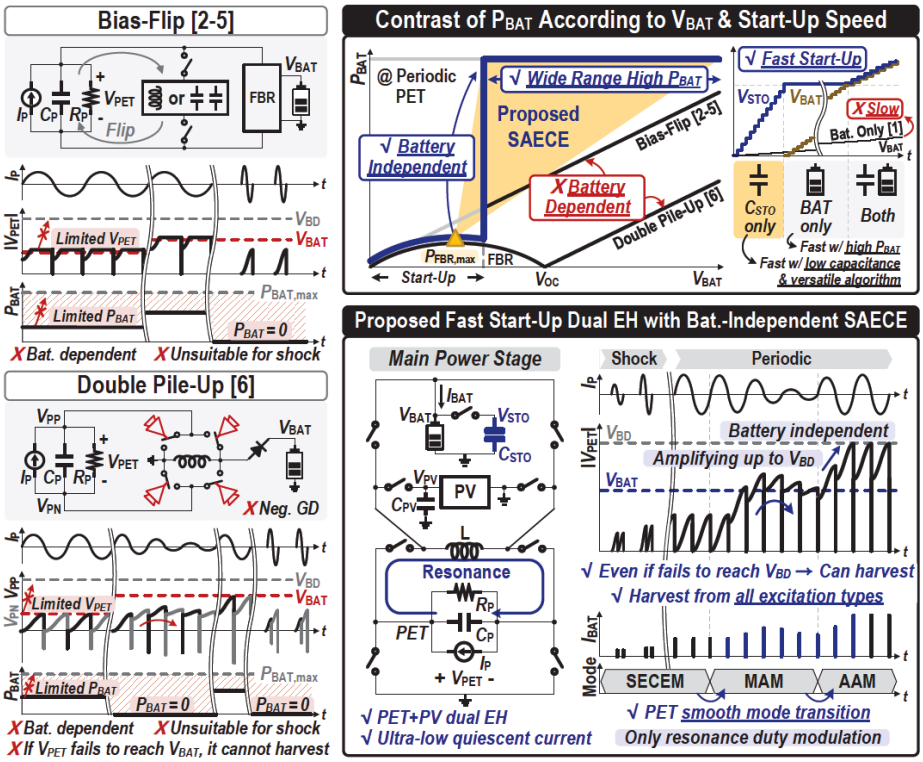
# A-SSCC 2025 Review

KAIST 전기및전자공학부 박사과정 이준기

## Session 18 Energy-Harvesting and Amplifiers

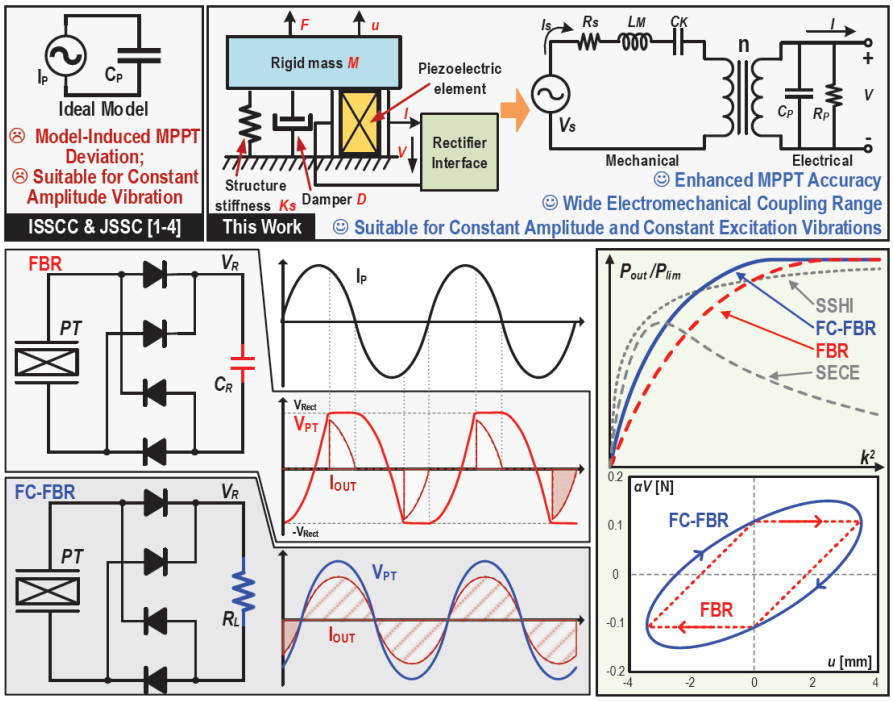
이번 A-SSCC 2025의 Session 18에서는 Energy-Harvesting and Amplifiers라는 주제로 총 5편의 논문이 발표되었다. Energy-Harvesting 분야에서는 다양한 동작 환경에서도 높은 전력 추출을 달성하기 위한 연구가 이루어졌다. 논문 18.1은 배터리 전압 상태와 무관하게 최대 전력 추출이 가능할 뿐만 아니라 빠른 스타트업을 지원하는 하베스팅 IC를 제안하였다. 논문 18.2에서는 압전 소자의 전기기계 결합에 따른 기존의 하베스팅 방식의 한계점을 지적하면서 이를 개선한 구조를 제안하였다. Amplifier 분야에서는 multi-level class-G supply modulator(논문 18.3), power-efficient dynamic amplifier(논문 18.4), linearity-enhanced wideband GaN amplifier(논문 18.5)와 같이 다양한 주제의 논문이 제출되었다. 이번 후기에서는 Energy-Harvesting 분야의 연구 2편을 자세하게 살펴보고자 한다.

**#18-1**은 POSTECH에서 발표한 논문으로, 배터리 전압에 독립적인 synchronous accumulated electric charge extraction(SAECE) 기술을 적용한 하베스팅 IC를 제안한다. 압전 소자(PET)에서 기존 에너지 하베스팅 기술은 정류 손실을 줄일 수 있다는 장점이 있지만, 추출 가능한 전력이 배터리 전압($V_{BAT}$)에 의해 제한된다는 한계가 있었다. 본 논문에서는 이러한 한계를 개선하기 위하여, 배터리 전압과 무관하게 압전 소자의 출력 전압($V_{PET}$)을 IC의 최대 허용 전압($V_{BD}$)까지 증폭시킬 수 있는 battery-independent SAECE를 제안하였다. 이 기술은 $V_{PET}$에 따라서 세 가지 모드를 지원하며, resonance duty cycle($D_R$)을 조절하여 부드러운 모드 전환을 달성하였다. 이를 통해 압전 소자의 진동 세기나 배터리 전압 크기에 상관없이 항상 최적의 에너지 추출이 가능하다. 또한, 빠른 스타트업을 위해 초기 구동 시 대용량 배터리 대신, 작은 용량의 보조 캐패시터($C_{STO}$)를 먼저 충전하는 알고리즘을 적용하였다. $C_{STO}$에 저장된 에너지를 이용해 IC를 구동하고 이후 배터리를 충전함으로써, 배터리 전압에 관계없이 빠른 스타트업이 가능하다. 추가적으로 단일 인덕터를 time-multiplexing하여 압전 소자와 태양 전지의 두 가지 에너지원으로부터 상호 간섭 없이 에너지를 추출하였다. 결과적으로 제안하는 하베스팅 IC는 기존의 full-bridge rectifier (FBR) 대비 780% 높은 최대 전력 추출 성능을 달성하였다.

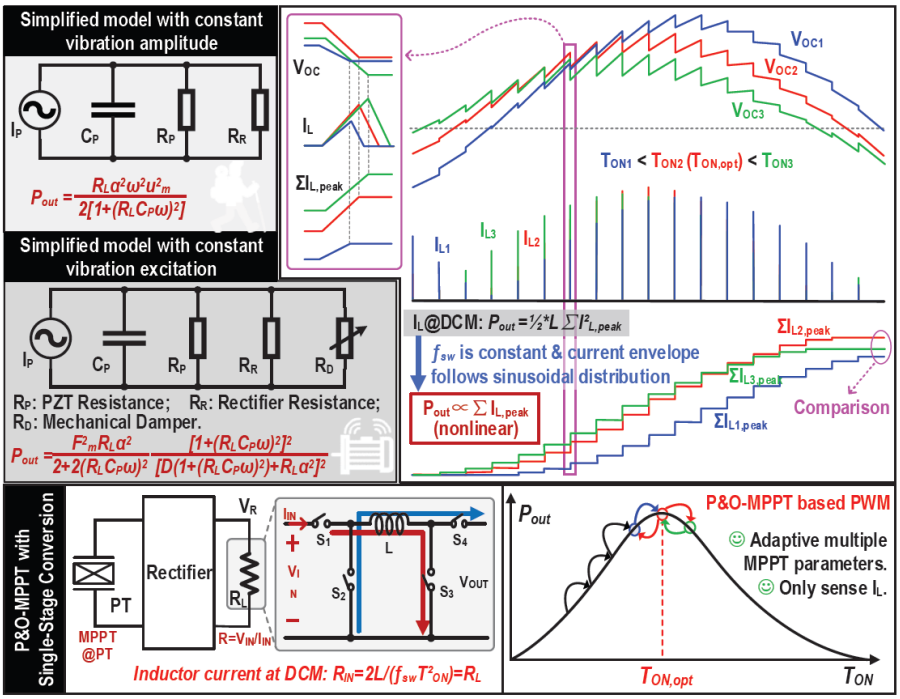



[그림 1] 제안하는 Battery-independent SAECE의 동작 원리 및 기존 기술과의 비교.

**#18-2**는 xidian university에서 발표한 논문으로, 압전 소자의 에너지 하베스팅을 위한 full-cycle full-bridge rectifier(FC-FBR)와 전류 누적 기반의 hill-climbing maximum power point tracking(MPPT) 기술을 제안한다. 기존의 하베스팅 기법은 압전 소자의 전기기계 결합이 약한 상황에서는 효과적이지만, 강한 결합 상황에서는 오히려 단순한 full-bridge rectifier 보다 효율이 떨어지는 문제가 있다. 본 논문은 이러한 문제를 해결하기 위하여 full-cycle FBR 구조를 제안하였다. 기존 FBR은 압전 소자 전압이 정류 전압보다 높을 때만 에너지를 수확하여 높은 손실이 발생하였으나, 제안하는 FC-FBR은 전체 사이클에서 전류 출력을 달성하여 더 넓은 에너지 변환 영역과 높은 출력을 확보하였다. 또한, 전류 누적 기반의 MPPT 기법을 새롭게 제안하였다. 출력 전력을 직접 측정하는 대신, 인덕터 피크 전류의 합이 출력 전력과 비례한 다는 점을 이용하여 복잡한 전력 감지 회로 없이도 P&O(Perturbation and Observation) 알고리즘을 통해 MPPT를 달성하였다. 측정 결과, 제안하는 회로는 91.3%의 최대 전력변환 효율을 달성하였으며, 기존 단일 스테이지 컨버터들이 100mV 이상의 출력전압 리플을 보이는 것과 달리, 20mV 미만의 매우 낮은 출력 리플을 달성하였다. 추가적으로 clock-multiplexing을 통하여 두 개의 압전 소자에 대하여 상호 간섭 없이 동시에 에너지 수확이 가능한 것을 실험적으로 검증하였다.




[그림 2] 제안하는 Full-cycle FBR(FC-FBR)의 동작 원리 및 기존 FBR과의 비교.




[그림 3] 제안하는 전류 누적 기반의 MPPT 방식의 동작 원리.

## 저자정보


**이준기 박사과정 대학원생**

- 소속 : KAIST
- 연구분야 : Power Management ICs
- 이메일 : leejune@kaist.ac.kr
- 홈페이지 : https://www.icdesignlab.net/

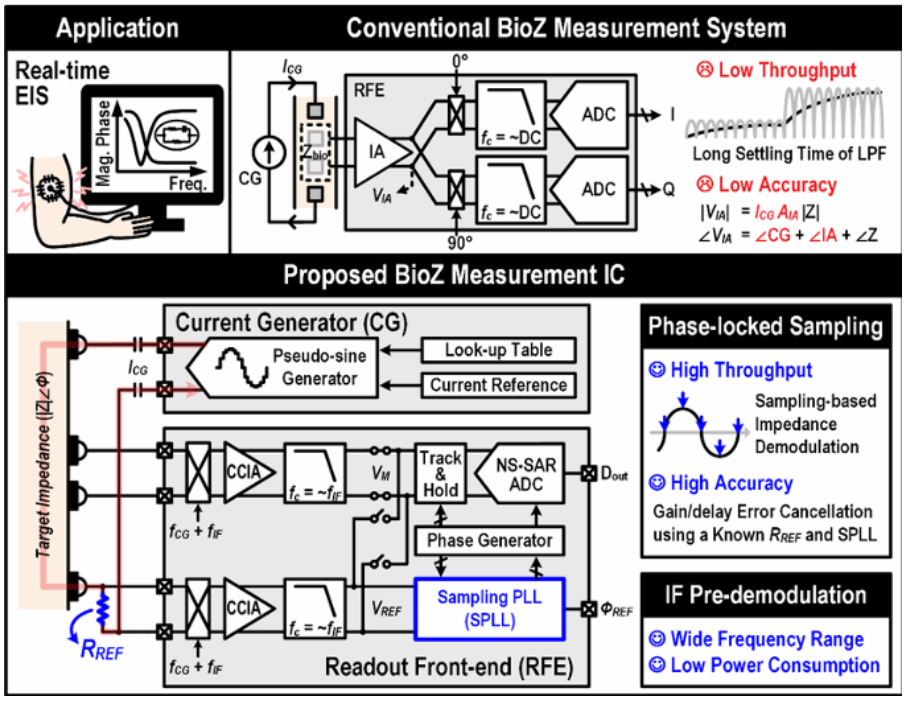
# ASSCC 2025 Review

고려대학교 전기및전자공학부 박사과정 안재웅

## Session 27 Precise and Robust Biomedical Interfaces

이번 2025 ASSCC의 Session 6에서는 Imagers라는 주제로 총 4편의 논문이 발표되었다.

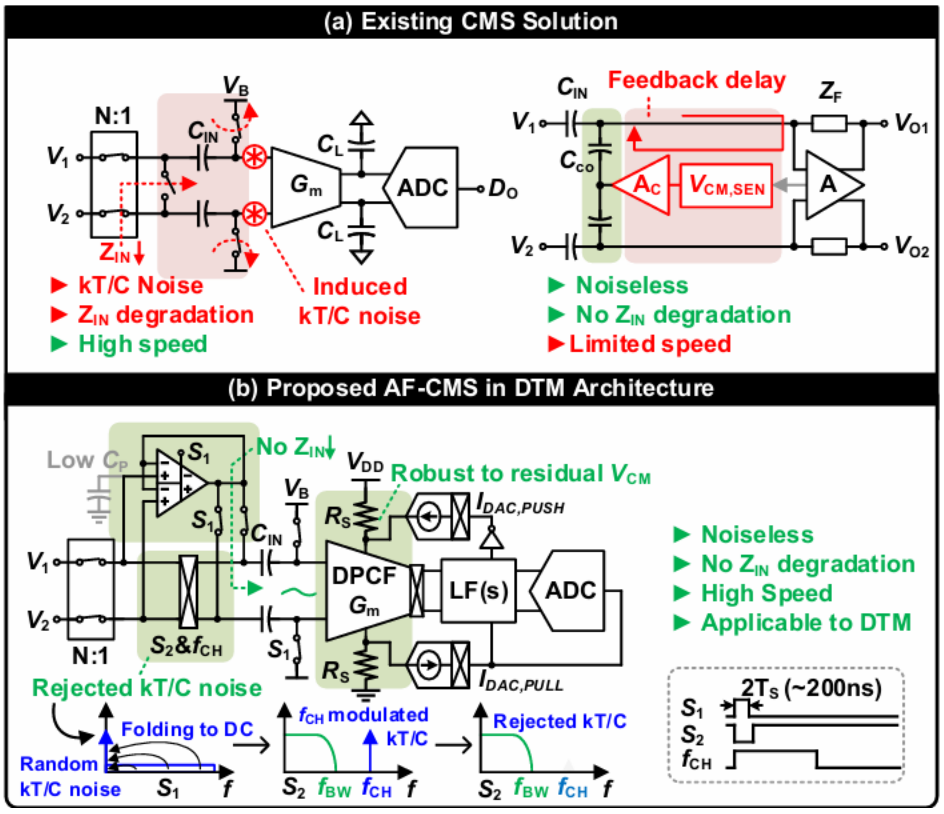
**#27-1** 이 논문은 기존 I/Q 기반 bio-impedance readout이 가지는 두 가지 한계, 즉 1) CG와 RFE 간의 위상지연·이득오차로 인해 발생하는 magnitude/phase 불일치, 2) DC 하향변환 이후 LPF settling 시간 때문에 발생하는 문제를 해결하기 위한 새로운 phase-locked sampling(PLS) 기반 EIS IC를 제안한다. 제안된 구조는 pseudo-sine current generator에 더해, target impedance와 기준 저항(RREF)을 동시에 측정하는 dual-AFE를 사용한다. 각 경로는 IF 대역에서 capacitive-coupled IA를 통해 증폭·필터링된 뒤, VREF의 zero-crossing에 위상 동기된 sampling clock을 이용해 VM과 VREF를 교차 샘플링한다. 한 주기 동안 0°/90°/180°/270°의 4-위상을 추출하여 실수·허수 성분을 한 주기안에 구할 수 있다. 결과적으로 LPF의 대역을 DC 보다 더 높게 설정할 수 있어 빠른 ODR이 가능해 진다. Sampling PLL(SPLL)도 사용이 되었는데, VREF가 VCM과 만날 때의 phase locked 된 타이밍을 이용하여 phase delay 문제를 해결하였다.SPLL은 sampling-phase detector와 frequency-locked loop(FLL)를 결합해, 큰 위상오차에서는 FLL이 빠르게 frequency를 수렴시킨다. 180 nm 공정으로 제작된 칩은 4 kHz IF에서 동작하며, 4 kS/s ODR을 실현한다. 20 Ω~4 kΩ 범위에서 저항 측정 오차는 0.4% 이하이며, 4 kHz~2 MHz 생체 임피던스 모델 측정에서 magnitude 오차 1.78%, phase 오차 1.8°를 달성한다. SPLL 잠금 안정도는 한 주기 이내에서 확보되며, 100 kΩ 환경에서 39.4 dB 이득 조건에서 40,000개의 샘플이 리코딩되었고 표준편차는 34.9 mΩ/√Hz를 보였다.




[그림 1] #27-1에서 기존의 구조와 제안한 bioZ 측정 IC

**#27-2** 고집적(high-density), 다채널(high-channel-count) 신경 기록 프런트엔드(FE)에서는 면적 최소화와 채널 간 균일성이 중요하며, 이를 위해 direct time-division multiplexing(DTM)-FE 구조가 널리 사용되어 왔다. 하지만 DTM은 빠른 스위칭에 의해 공통모드 간섭(CMI)이 상향 변조되어 커지기 때문에 왜곡과 포화가 기존 구조보다 훨씬 쉽게 발생한다. 이 논문은 이러한 문제를 해결하기 위해 active feedforward CMS(AF-CMS)를 제안한다. S1 동안 각 채널의 CMI를 입력 커패시터($C_{IN}$)에 샘플링해 저장하고, S2에서는 $C_{IN}$이 CMI를 제거한 뒤 differential-mode(DM) 신호만 통과시킨다. DM 신호는 continuous-time 2nd-order incremental ADC(IADC)의 fine loop(DPCF-Gm)와 coarse loop(DM tracker)를 통해 처리된다. 샘플링 과정에서 발생하는 kT/C noise는 chopping을 통해 DM 경로에서 제거되고, $C_{IN}$은 버퍼에 의해 구동되기 때문에 입력 임피던스 저하도 없다. 또한 AF-CMS는 샘플링 기반이므로 Gm variation에 따라 residual CMI가 남을 수 있다. 이를 해결하기 위해 논문은 Gm을 루프 안에 포함시키는 구조를 채택하였다. IDAC을 Rs를 통해 피드백 구조를 사용함으로서 폐루프 이득이 (IDAC·Rs)−1로 고정되어 Gm 변화에 영향을 받지 않는다. 하지만 Gm degeneration으로 인한 낮은 Gm과 증가된 노이즈를 보완하기 위해, 논문은 dual-path current-feedback(DPCF)-Gm 을 제안해 유효 Gm을 두 배로 높이고, 선형성을 향상시키는 구조를 적용하였다. 180nm CMOS로 구현된 8채널 DTM-DD FE는 AF-CMS 적용 시 10 mVpp, 40 Hz CMI 조건에서 119 dB의 peak CMRR

을 달성해 평균 34 dB의 개선 효과를 보였다. 또한 100 Hz, 700 mVpp의 매우 큰 CMI 상황에서도 포화되지 않고 51.4 dB SNDR을 유지해, 넓은 common-mode 범위에서도 안정적인 동작이 가능함을 입증하였다. IRN도 AF-CMS 적용 여부에 따른 차이가 거의 없어, 제안된 CMS가 노이즈를 증가시키지 않음을 확인하였다. 검증된 neural signal 재생 실험(LFP, spikes)에서도 원신호와의 높은 상관(R=0.998), 동일한 spike 검출 개수를 보였으며, 기존 DTM FEs 대비 가장 높은 CMRR과 가장 넓은 CM 범위를 제공한다는 점에서 경쟁력을 확인하였다.




[그림 2] 기존의 CMS solution과 제안한 AF-CMS in DTM 구조

## 저자정보


**안재웅 박사과정 대학원생**

- 소속 : 고려대학교
- 연구분야 : 디스플레이 드라이버 / 픽셀 보상 / 터치 센서
- 이메일 : ajw1104@korea.ac.kr
- 홈페이지 : https://sites.google.com/site/kubasiclab/home

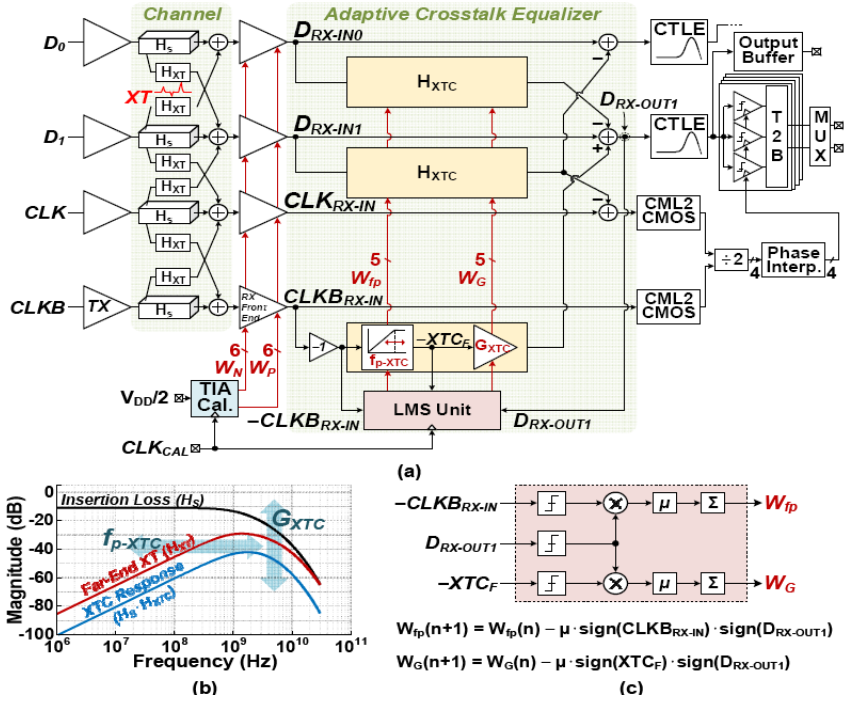
# A-SSCC 2025 Review

서강대학교 전자공학과 석박통합과정 박종민

## Session 30 Analog and Digital Interface

2025 ASSCC Session 30는 Transceiver 2편, 온도센서 1편, Hall sensor 1편으로 총 4편의 다양한 영역에서의 interface를 소개하고 있다. 실리콘 칩의 소형화 및 센서 집적화에서 중요한 요소인 환경변화에 Adaptive한 솔루션을 공통된 주제로 제시하고 있고, 그 과정에서 저전력으로 에너지효율성을 강조하는 논문들이 발표되었다. 그 중 최근 많은 주목을 받고 있는 Chiplet interface의 난제들을 다루고 있는 논문 2개를 살펴볼 예정이다.

**#30-1** 본 논문은 카이스트, 캐나다 Marvell 그리고 SK Hynix에서 공동 발표한 논문으로 2.5D chiplet packaging에서 이슈가 되는 Crosstalk과 온도변화에 대한 tolerance를 갖춘 interface solution을 제안한다. 작은 면적의 칩에서 높은 데이터속도의 transceiver를 구현하기 위해 single-ended data lane의 간격이 점점 좁아지고, 그로 인해 Capacitive coupling으로 인한 Crosstalk(XT)이 심화된다. PAM-4에서 이 현상은 더 두드러지며 SNR성능이 감소하게 되는데, 이를 해결하기 위해 본 논문은 adaptive Crosstalk Equalizer를 도입하였다. XT의 특성과 동일한 High Pass Filter를 활용하여 Data에 포함된 XT성분을 제거하는 방식이며 Least Mean Square(LMS) 알고리즘을 통해 filter의 cutoff frequency와 DC gain을 조절하게 된다. LMS를 통한 filter 특성의 최적화는 XT를 유발하는 Aggressor의 신호와 XT의 영향을 받은 Victim신호의 Equalizing 결과를 종합해 이루어지게 되는데, Victim의 XT가 다시 aggressor에 영향을 주는 mutual XT를 방지하기 위해 Inversion Clock signal을 aggressor로 활용하였다. 추가적으로 calibration이 포함된 replica TIA를 RX Front-end에 적용해 -25~115℃의 넓은 온도변화에 대한 Common-mode mismatch를 3mV수준으로 줄였다. 그 결과 PAM-4 신호 레벨 간 간격의 불일치를 나타내는 Ratio-Level Mismatch(RLM)을 83%에서 97%로 향상시켰으며 crosstalk이 포함된 clock의 rms jitter 성능을 10.49ps에서 2.17ps로 개선시켰다. 1E-12의 BER에서 0.11UI의 Bathtub 지표를 보였고, 넓은 온도변화에서 0.02UI의 적은 Eye opening variation을 보였다.




[그림 1] Adaptive crosstalk equalizer를 이용한 PAM-4 Transceiver의 구성

**#30-2** 는 광운대학교에서 발표한 논문으로, 입력 데이터의 Transition을 기반으로 동작하는 새로운 송수신기 구조를 제안한다. 본 연구는 Die-to-Die 인터페이스용 송수신기 설계에서, Capacitor 기반 수신기 구동 방식이 갖는 면적 증가와 제한적인 런길이(run length) 문제를 해결하기 위해 세 가지 접근법을 제시한다.

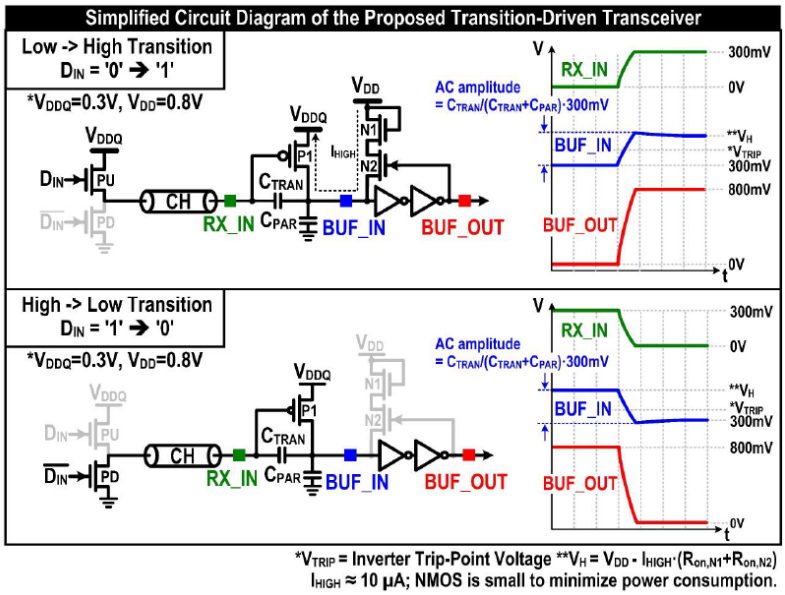
첫째, AC-coupling Capacitor를 송신기(TX)와 수신기(RX)가 공유하도록 하여 데이터 Transition을 감지하는 방식을 도입했다.

둘째, 입력 데이터의 DC 레벨에 따라 Low level에서는 PMOS를, high level에서는 피드백 기반 NMOS를 선택적으로 사용해 데이터를 안정적으로 수신하는 구조를 제안했다.

셋째, 간단한 inverter 기반 RX를 활용하여 입력 데이터의 middle voltage를 inverter의 trip-point 근처로 정렬시켜 동작 안정성을 확보했다.

특히 High→Low Transition시 피드백 기반 NMOS에서 발생할 수 있는 AC-coupling 캐패시터 induced transition feed-through로 인해 eye의 Low DC level 아래로의 undershoot 현상을 보상하기 위해 capacitor를 추가함으로써 Eye margin을 확보하였다.

제안된 송수신기는 약 0.002 $m^2$의 매우 작은 면적과 118.89 Tb/s/mm/pJ/bit의 높은 FoM(Figure of Merit, 파워 대비 Effective Bandwidth)을 기록하며 우수한 성능을 입증했다.




[그림 2] 제안된 Transition-driven transceiver의 operation

## 저자정보

**박종민 석박통합과정 대학원생**

- 소속 : 서강대학교
- 연구분야 : Reference-less CDR/High-Speed Wireline Interface
- 이메일 : park_john@naver.com
- 홈페이지 : https://sc.sogang.ac.kr/melab

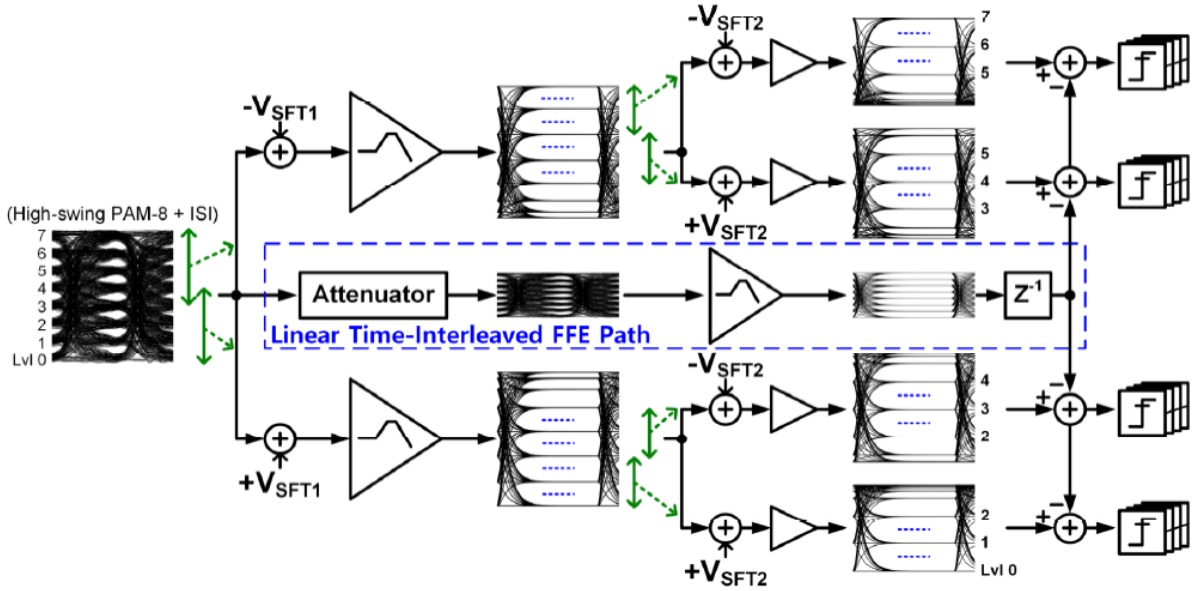
# A-SSCC 2025 Review

서강대학교 전자공학과 석박통합과정 박종민

## Session 12 Ultra high-speed transceiver

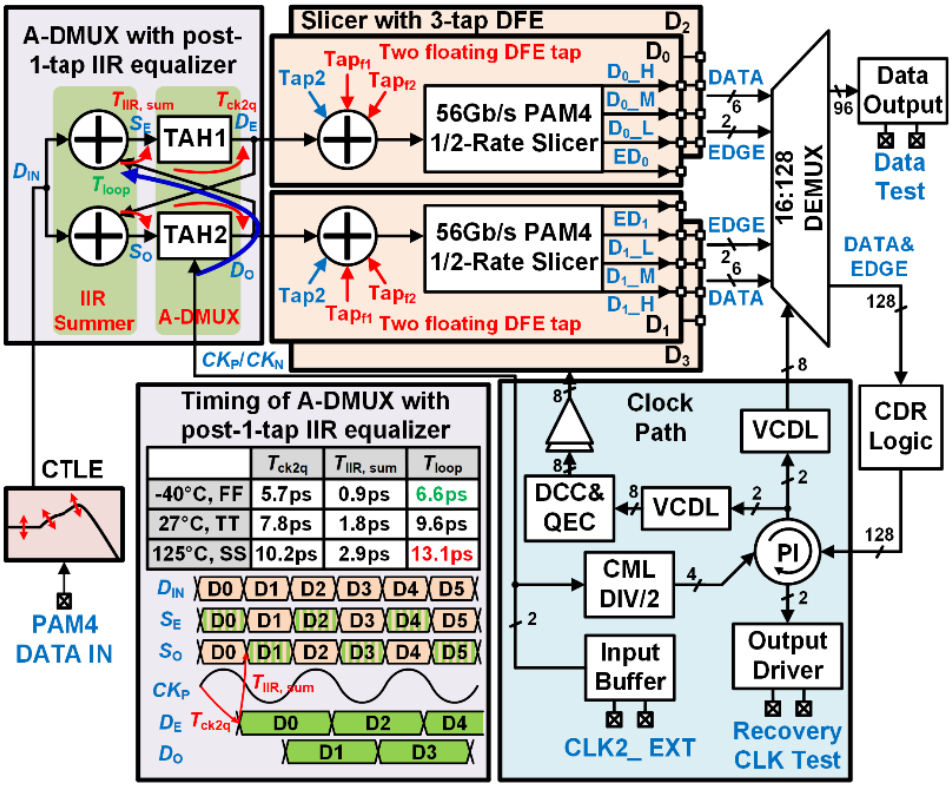
2025 ASSCC Session 12는 총 4편의 논문이 발표되었다. "Beyond 100Gb/s"의 공통된 흐름 속에 초고속 데이터 전송을 위한 여러 연구들이 수신기 2편, 광통신 1편, 송신기 1편로 나누어 소개가 되었다. 특별히 PAM-4/PAM-8/QAM와 같은 고차변조기법에 필수적인 정밀한 Equalization 기법들이 소개되었고, 높은 선형성 및 에너지 효율에 초점이 맞추어지며, 단순한 Data rate 확대가 아닌 완성도 높은 송수신 시스템을 요구하는 산업의 흐름을 볼 수 있다. 그 중 Receiver와 관련된 논문 2편을 살펴보고자 한다.

**#12-1** 본 논문은 한양대학교에서 발표한 논문으로 108Gb/s의 PAM-8 Receiver의 높은 linearity를 보장하기 위해 Multi-path CTLE+FFE구조의 Equalization system을 제안한다. High-swing으로 입력되는 PAM-8 signal에서의 linearity를 만족하기 위해 Linear time-interleaved FFE가 도입되었으며, High-swing에 대한 linearity 확보를 위한 Attenuator, ISI를 보상하기 위해 입력신호의 레벨을 Sampling하는 Track and Hold circuit(TAH) 그리고 FFE의 coefficient를 조절하기 위한 VGA로 구성이 되어있으며, TAH에서는 timing margin을 확보하기 위해 75% Duty Cycle clock signal을 활용하였다. 한편 CTLE의 구조에서는 Amplifier를 활용한 4-way sub-ranging technique을 통해 8개의 signal level을 PAM-3형태로 나누어 결과적으로 동일한 간격의 Signal level를 만족하는 high-linearity PAM-8 Equalizing System을 구현하였다. 구현된 시스템을 통해 1.4V의 차동입력 신호에 대하여 1.95pJ/bit의 입력전압 대비 높은 전력효율을 보였고 Nyquist frequency에서 10.7dB의 loss 특성을 가진 channel에 대하여 1E-7의 BER을 만족하며 기존 CTLE+2-tap FFE system 대비 개선된 BER특성을 달성했다.




[그림 1] 4-way sub-ranging CLTE + Linear Time-Interleaved FFE의 구성

**#12-2** 는 시안 자오퉁 대학교에서 발표한 논문으로 IIR Equalizer와 3-tap Direct DFE를 이용한 time-interleaved 방식의 112Gb/s PAM4 Receiver를 소개한다. 100G급 PAM4 signal에 활용되는 기존 mixed-signal receiver는 DFE의 timing margin 확보의 어려움 및 Clock signaling의 복잡도로 인해 Short-reach 기반의 높은 Data rate에 적용하는 데에 한계가 있었다. 이를 해결하기 위해 Post-1-tap IIR summer와 High-linearity Track and Hold circuit(TAH)를 활용한 1:2 Analog DEMUX를 구성하여 DFE가 가지고 있는 1-tap timing budget을 완화하고, loop-unrolled 구조에서 발생하는 추가적인 파워소모를 줄였다. 이후 long tail로 발생하는 잔여 ISI의 미세 보정을 위해 tap-2 direct DFE와 2 floating tap direct DFE를 활용하여 넓은 bandwidth에서 ISI를 보정할 수 있도록 하였다. 제안된 방식을 활용하여 Nyquist frequency에서 28.4dB의 loss특성을 보이는 channel에서 1E-12수준의 BER 및 1.32pJ/bit의 에너지 효율을 달성하며, channel loss에 대한 에너지효율 FoM인 0.046pJ/bit/dB의 에너지 효율을 달성했다.




[그림 2] 112Gb/s time-interleaved PAM4 RX의 Architecture

## 저자정보

**박종민 석박통합과정 대학원생**

- 소속 : 서강대학교
- 연구분야 : Reference-less CDR/High-Speed Wireline Interface
- 이메일 : park_john@naver.com
  홈페이지 : https://sc.sogang.ac.kr/melab

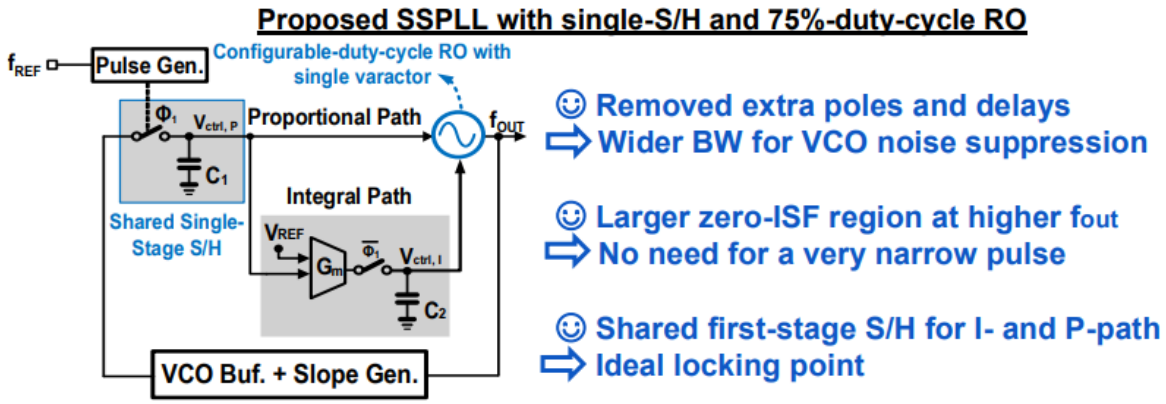
# A-SSCC 2025 Review

단국대학교 파운드리공학과 석사과정 임재영

## Session 22 Advanced Timing Recovery

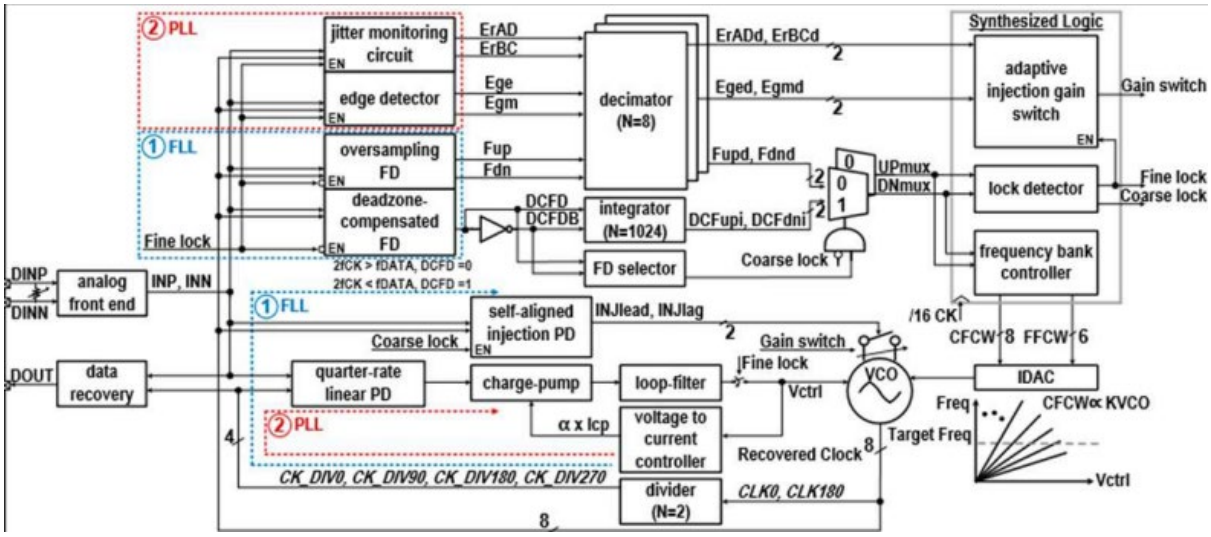
이번 A-SSCC 2025의 Session 22에서는 Advanced Timing Recovery를 주제로 고속, 저전력 시스템에서의 클록 생성 및 복구 과정에서 발생하는 timing jitter, reference spur, phase noise 누적, 그리고 lock 안정성 문제를 다루는 5편의 논문이 발표되었다. 본 세션의 논문들은 subsampling, duty-cycle 제어, injection-locked 동작, 시간 영역 처리 기법 등을 활용하여 timing recovery 성능을 제한하는 요인들을 구조적으로 해결하는 접근을 제시한다. 다음에서는 이 중 timing recovery 메커니즘을 서로 다른 관점에서 다루는 #22.1, #22.2, #22.4 논문의 구조와 동작 방식을 중심으로 정리한다.

**#22-1** 본 논문은 일본 도쿄대학교에서 발표한 논문으로, RO 기반 subsampling PLL에서 발생하는 reference spur와 timing jitter의 발생 메커니즘을 구조적으로 다룬다. 기존 subsampling PLL에서는 reference clock을 이용해 VCO 출력을 직접 샘플링함으로써 timing jitter를 줄이는 구조가 사용되어 왔다. 이때 RO 기반 PLL을 적용하면 LC 기반 구조 대비 회로 집적도 측면에서 이점이 있으나, 일반적인 50% duty-cycle RO에서는 reference spur의 크기가 duty-cycle과 pulse width에 민감하게 의존하며, timing jitter와 spur 간의 trade-off가 발생한다. 또한 기존 subsampling PLL에서는 I-path와 P-path가 서로 다른 sample-and-hold(S/H) 경로를 사용함에 따라 샘플링 시점 불일치로 인한 timing mismatch가 발생할 수 있다. 이를 완화하기 위한 방법으로 multi-stage S/H 구조나 보정 회로가 사용되어 왔으나, loop 복잡도가 증가하고 추가적인 timing 오차 요인이 생긴다. 본 논문에서는 이러한 문제를 고려하여 configurable duty-cycle ring oscillator를 사용하는 subsampling PLL 구조를 제안하였다. RO의 duty-cycle을 가변적으로 조절함으로써 reference spur의 크기를 구조적으로 제어할 수 있도록 하였으며, spur 억제와 timing jitter 간의 균형을 조절한다. 또한 기존 two-stage S/H 구조 대신 single-stage S/H 구조를 적용하고, I-path와 P-path가 shared S/H를 통해 동일한 샘플링 시점을 공유하도록 구성하였다. 이를 통해 경로 간 timing mismatch를 제거하고 loop 동작 시 발생하는 timing 오차를 줄인다. 측정 결과, 제안된 PLL은 2.6GHz 출력에서 reference spur -70dBc와 RMS jitter 186fs를 달성하였다.



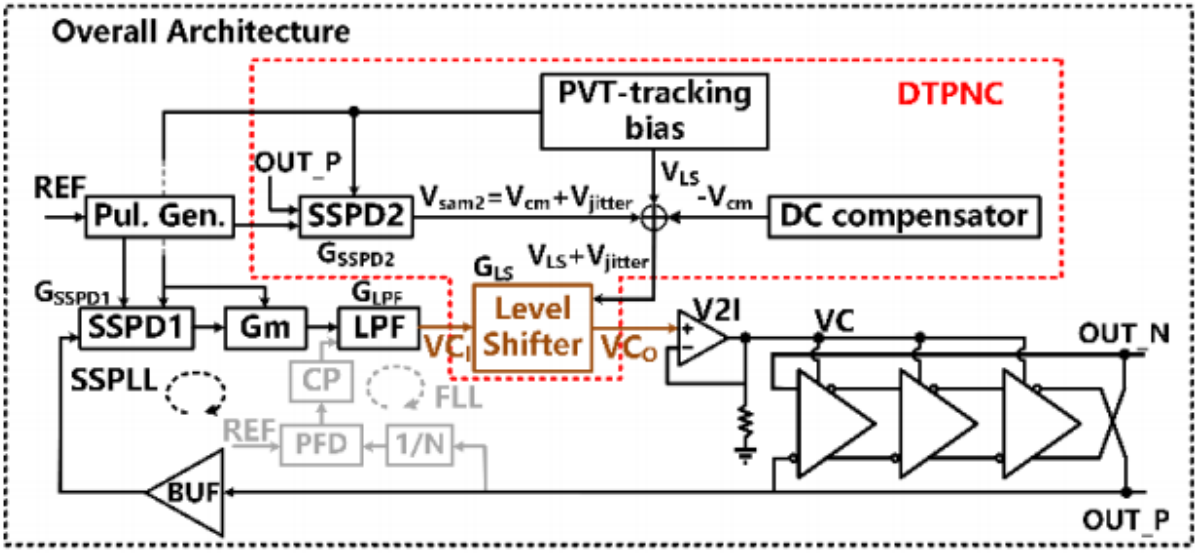

[그림 1] 제안된 single-S/H 와 75% duty-cycle RO가 적용된 SSPLL

**#22-2** 본 논문은 대한민국 성균관대학교와 삼성전자에서 발표한 논문으로, reference-less 환경에서 timing lock 안정성 문제를 해결하기 위한 CDR 구조를 제안하였다. 기존 reference-less CDR에서는 입력 데이터 스트림을 이용한 injection-locked 동작이 널리 사용되어 왔으나, 입력 데이터의 jitter 및 PVT 변화에 따라 injection 조건이 쉽게 변하면서 timing lock이 불안정해질 수 있다. 특히 고정된 injection gain을 사용하는 경우, 초기 lock을 위한 충분한 pull-in과 steady-state에서의 안정적인 timing 정렬을 동시에 만족시키기 어렵다. 이를 보완하기 위해 본 논문에서는 dual-loop 구조의 reference-less CDR를 제안하였다. 제안된 구조에서는 주파수 복구를 담당하는 frequency detection 경로와 injection 위상 정렬을 담당하는 경로를 분리하여 구성함으로써, timing recovery 과정에서 각 경로의 역할을 명확히 구분한다. 또한 데이터 에지의 기울기를 이용하는 Injection Pulse Detection(IPD) 기법을 적용하여 injection 조건을 판단한다. injection과정에서의 timing 안정성을 확보하기 위해 Adaptive Injection Gain Switching(AIGS) 기법이 도입되었으며, lock 상태에 따라 injection gain을 변경함으로써 초기 lock 구간에서는 충분한 injection을 제공하고 steady-stage에서는 timing disturbance를 줄인다. 제안된 CDR은 5Gb/s부터 12.5Gb/s까지의 데이터 속도 범위에서 안정적인 timing lock 동작을 보였으며, 12.5휴/ㄴ 기준 PLL 모드에서 약 14.9mW의 전력 소모를 나타낸다.



[그림 2] adaptive injection gain switching을 적용한 제안된 reference-less CDR 구조

**#22-4** 본 논문은 RVCO 기반 PLL에서 발생하는 multi-phase timing jitter 누적 문제를 다룬다. RVCO 기반 PLL에서는 다상 클록을 생성하는 과정에서 각 phase의 phase noise가 누적되며, PVT 변화에 따라 phase 간 mismatch가 발생하여 clock generation 성능이 제한된다. 기존의 phase noise 저감 기법은 아날로그 보정 회로나 복잡한 calibration 절차를 요구하는 경우가 많다. 본 논문에서는 이러한 한계를 고려하여 Discrete-Time Phase Noise Cancellation(DTPNC) 기법을 적용한 RVCO 기반 SSPLL 구조를 제안하였다. 제안된 구조에서는 RVCO 출력의 phase 정보를 시간 영역에서 샘플링하고, 이를 이용해 timing noise 성분을 디지털 방식으로 제거한다. 이를 통해 아날로그 영역의 추가적인 보정 회로 없이 timing jitter를 줄이는 접근을 취한다. 또한 dual mismatch adjustment를 통해 다상 경로에서 발생하는 불일치를 단계적으로 보정하고, adaptive gain mismatch adjustment 기법을 함께 적용하여, multi-phase 경로 간 gain mismatch로 인한 timing 불균형을 완화한다. 이를 통해 PVT 변화에 따른 timing jitter 변화를 완화한다. 측정 결과, 제안된 SSPLL은 2.4GHz 출력주파수에서 FoM_jitter -247.4dBc/Hz를 나타낸다.



[그림 3] Discrete-Time Phase Noise Cancellation을 적용한 제안된 RVCO 기반 SSPLL의 전체구조

## 저자정보

**임재영 석사과정 대학원생**

- 소속 : 단국대학교
- 연구분야 : clock generators
- 이메일 : ljy72250338@dankook.ac.kr
- 홈페이지 : https://sites.google.com/dankook.ac.kr/acs-lab

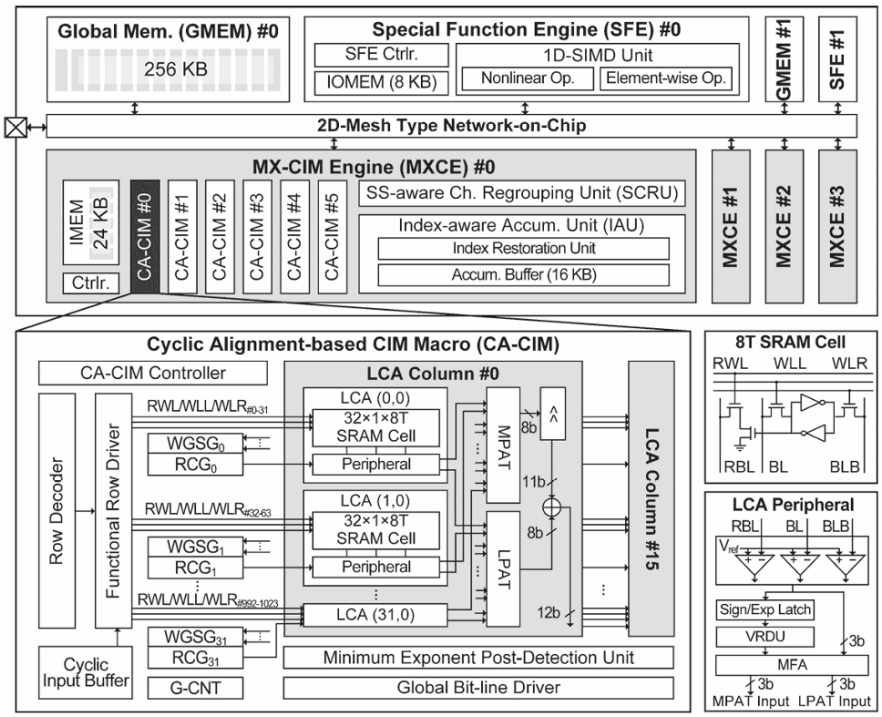
# A-SSCC 2025 Review

연세대학교 전기전자공학부 박사과정 여민준

## Session 11 Advanced Digital Compute-In-Memory For Edge AI

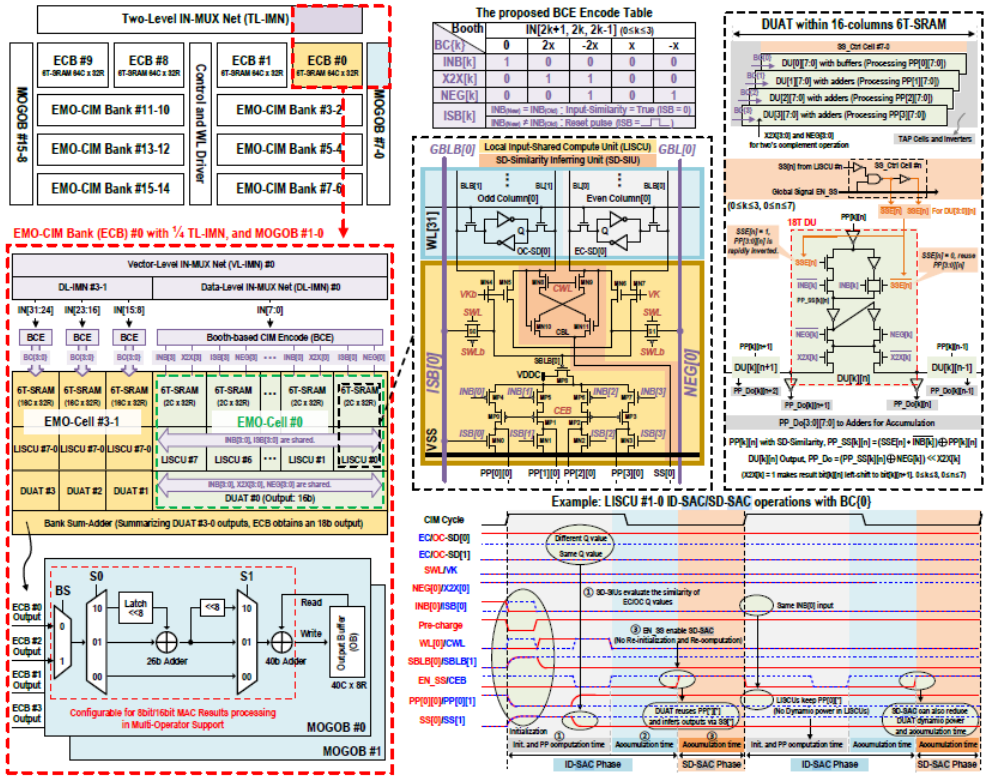
이번 A-SSCC 2025의 Session 11은 Edge AI를 위한 디지털 인메모리 컴퓨팅(Compute-In-Memory, CIM)을 주제로 총 4편의 논문이 발표되었다. 하나의 세션 안에서 생성형 AI용 sub-8bit Microscaling data format 기반 디지털 CIM 가속기(MIDAS), CNN/Attention/DW를 단일 매크로에서 지원하는 유사도 기반 EMO-CIM, RRAM–eDRAM 융합 메모리를 사용하는 부동소수점(FP) CIM 엔진(CENTAUR), 그리고 BF16 포맷을 사용하는 초고효율 디지털 CIM 매크로까지, 정수–BF16–부동소수점 전 영역을 포괄하고 있다는 점이 특징이다. 공통적으로는 (1) CNN뿐 아니라 Attention, Depthwise Conv, Diffusion 등 최신 네트워크의 다양한 연산자를 단일 하드웨어에서 지원하려는 "multi-operator" 지향, (2) 데이터 포맷(Microscaling, BF16, FP)과 Dataflow(3D-MAC, SAC, RCBS 등)을 활용해 에너지·면적 효율(EF/AF)을 극대화하려는 시도가 돋보인다.

**#11-1** KAIST에서 발표한 28nm 디지털 CIM 기반 생성형 AI 가속기 MIDAS로, Microscaling data format과 "Cyclic Alignment" data flow를 활용해 가중치 저장과 CIM Array 효율을 동시에 개선한 점이 핵심이다. Microscaling 포맷에 맞춘 Cyclic alignment 기반 CIM은 동일한 메모리 용량에서 더 많은 가중치를 수용하면서 에너지 효율 1.72배, 면적 효율 2.47배 향상을 달성하였다. 또한 최소 지수(min-exponent)를 검출하는 relative counter 기반 포스트 프로세싱을 통해 디지털 구현 대비 전력 68.8%, 면적 63.2%를 절감하였고, 센스앰프 기반 dynamic bit-significance 처리 및 shared scale-aware 채널 리그룹핑으로 추가 27.3%의 에너지 절감을 얻었다. 28nm 실리콘 측정 결과, 기존 최첨단 디지털 CIM 가속기 대비 시스템 레벨 에너지 효율 1.58배, 매크로 레벨 에너지 효율 3.32배를 보고하였다. 기존 디지털 CIM 연구들이 주로 이미지 분류와 같은 classification 중심 워크로드를 대상으로 했던 것에 비해, MIDAS는 Microscaling 포맷과 Cyclic alignment를 활용해 생성형 AI까지 적용 범위를 확장했다는 점에서 의미가 크다



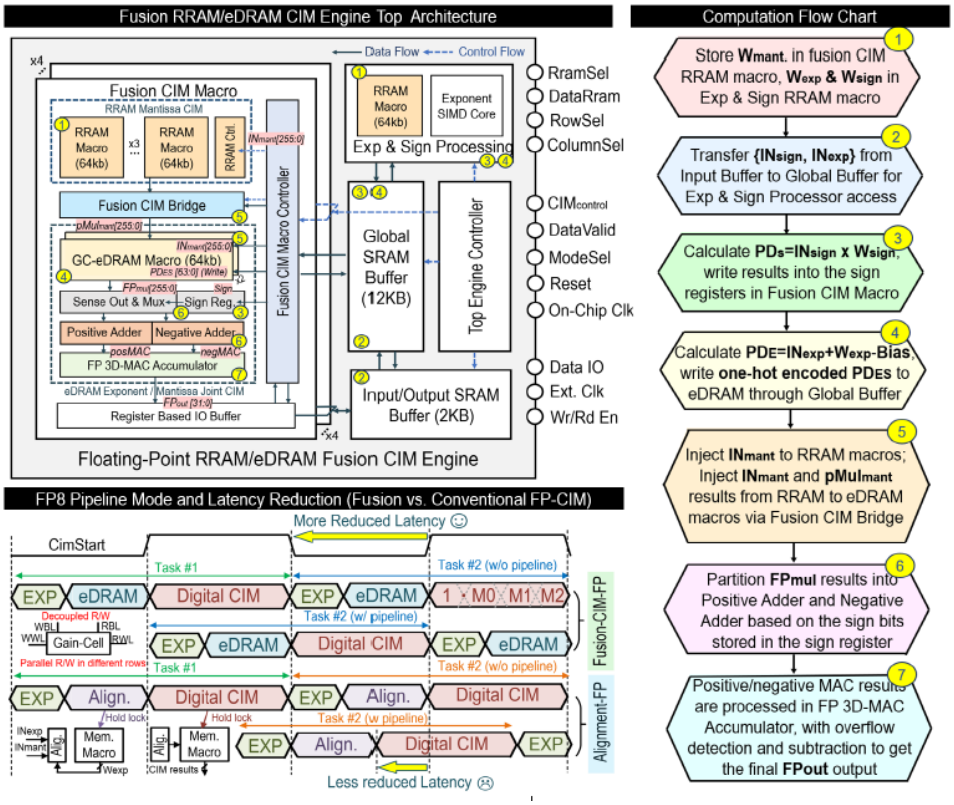

[그림 1] "MIDAS: An Energy-Efficient Microscaling Digital Compute- In-Memory-Based Accelerator with Spatio-Temporal Cyclic Alignment for Generative AI Inference" 전체 아키텍처 구조

**#11-2** Southeast University의 28nm SRAM 기반 EMO-CIM 매크로로, 하나의 CIM 매크로에서 CNN·Attention·Depthwise Conv를 모두 지원하는 "multi-operator" 구조를 제안한다. 입력과 정지 데이터의 비트 유사도를 추출하는 LISCU와 Booth 기반 인코더를 결합한 SAC 연산을 통해 불필요한 MAC을 줄이고, RVSA 기반 data flow로 연산 길이가 달라져도 에너지·면적 효율을 유지하도록 설계한 것이 특징이다. 또한 16개의 EMO-CIM bank와 TL-IMN, MOGOB로 구성된 37-kb 유닛 매크로 구조를 통해 CNN/Attention/DW 모드 간 전환을 하드웨어 차원에서 유연하게 지원한다. 입력/정지 데이터 유사도에 기반한 SAC 기법 덕분에 동일 공정의 기존 디지털/하이브리드 CIM 대비 에너지·면적 동시 지표(FoM)를 향상시킨 것으로 보고되며, 0.7 V에서 CNN/Attention 93.5 TOPS/W, DW 39.3 TOPS/W 및 최대 4.97 TOPS/mm² 수준의 효율을 달성한다. Mobile-ViT, YOLO-v10 기반 Edge-AI 벤치마크에서도 실용적인 정확도를 달성하였다.



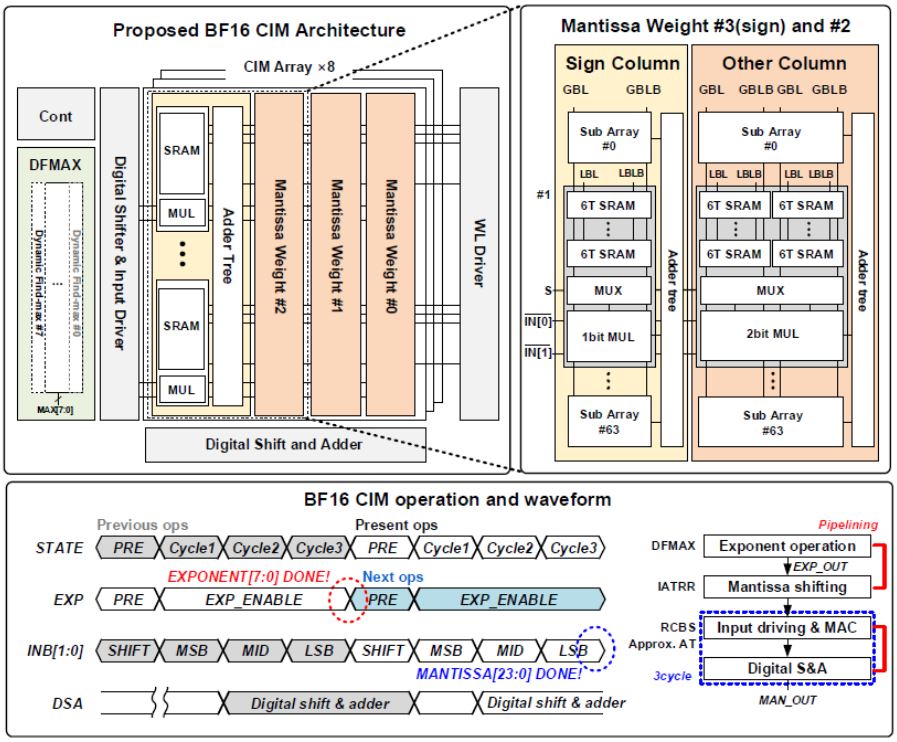

[그림 2] "EMO-CIM: An Input/Stationary-Data Similarity-Aware Computing-In-Memory Macro for Variable Vector-Wise Computation in Edge Multi-Operator AI Acceleration" 전체 아키텍처 구조

**#11-3** Purdue University 주도의 40nm RRAM–eDRAM 융합 부동소수점 CIM 엔진(CENTAUR)으로, Mantissa 연산은 RRAM, Exponent 처리는 eDRAM에서 담당하는 "FP 3D-MAC" dataflow를 통해 기존 NVM 기반 FP-CIM의 병목인 지수 정렬(pre-alignment) 오버헤드와 낮은 동작 주파수 문제를 완화한 점이 특징이다. 3T1C gain-cell 기반 shift-vector 3-operand MAC 구조를 도입해 메모리 내에서 곱셈·가산·시프트를 결합적으로 수행함으로써, 외부 디지털 로직으로의 데이터 이동을 줄이고 MAC 파이프라인을 깊게 가져갈 수 있도록 했다. 그 결과, 40nm 공정임에도 불구하고 600 MHz에서 38.5 TFLOPS/W의 에너지 효율을 달성하고, 동일 세대의 기존 SRAM-/RRAM-CIM 대비 종합 FoM이 각각 약 4.8배, 8.5배 향상시켰다. Tiny-ViT 기반 비전 워크로드를 대상으로 baseline 대비 약 1.7% 수준의 정확도 손실만을 보고하여 실용적인 정확도를 확인하였지만, 더 복잡한 dataset으로 검증 범위를 확장한다면, 제안한 엔진의 효용성이 더욱 뚜렷해 질 것으로 보인다.




[그림 3] "CENTAUR: A 38.5-TFLOPS/W 600MHz Floating-Point Digital Compute-In-Memory Engine with 40nm Fusion RRAM-eDRAM Macros Featuring 3D-MAC Operation" 전체 아키텍처 구조

**#11-4** 성균관대학교에서 발표한 28nm BF16 디지털 CIM 매크로로, BF16 포맷의 넓은 지수 범위와 상대 오차 특성을 활용해 지수·가수 연산을 근사화하면서 에너지와 레이턴시를 동시에 줄인 설계이다. Dynamic Find-Max(DFMAX) 유닛을 통해 64개 입력 쌍에서 최대 지수를 동적 로직과 마스킹으로 효율적으로 탐색하고, Reduced-Cycle Bit-Serial(RCBS) 구조를 도입해 비트-직렬 연산을 기존 4사이클에서 3사이클로 단축함으로써 연산 지연과 에너지 소모를 크게 줄인다. 여기에 Vth drop 문제를 완화한 16T/18T full-adder 기반 approximate adder tree와 input-aware toggling rate reduction(IATRR) 기법을 적용하여, 동일 면적 대비 스위칭 활동을 낮추면서 연산량을 극대화한 것이 특징이다. 그 결과, 0.0576 mm² 매크로에서 0.4 V 기준 77.2 TFLOPS/W, 0.324 TFLOPS/mm²의 높은 에너지·면적 효율을 달성하고, CIFAR-100/ResNet-18 벤치마크에서도 약 0.21% 이내의 정확도 손실만을 보여 BF16 기반 근사 CIM의 실용성을 입증하였다.




[그림 4] "A 28nm 77.2 TFLOPS/W Digital Floating-Point Compute-In-Memory Macro Employing Dynamic Find-Max and Reduced-Cycle Bit-Serial Architecture with Approximation" 전체 아키텍처 구조

## 저자정보


**여민준 박사과정 대학원생**

- 소속 : 연세대학교
- 연구분야 : 저전력 고 신뢰성 SRAM 설계, Computing-in-memory 설계
- 이메일 : ymj5887@yonsei.ac.kr
- 홈페이지 : http://vlsisys.yonsei.ac.kr/

# A-SSCC 2025 Review

한양대학교 신소재공학과 석박통합과정 송충석

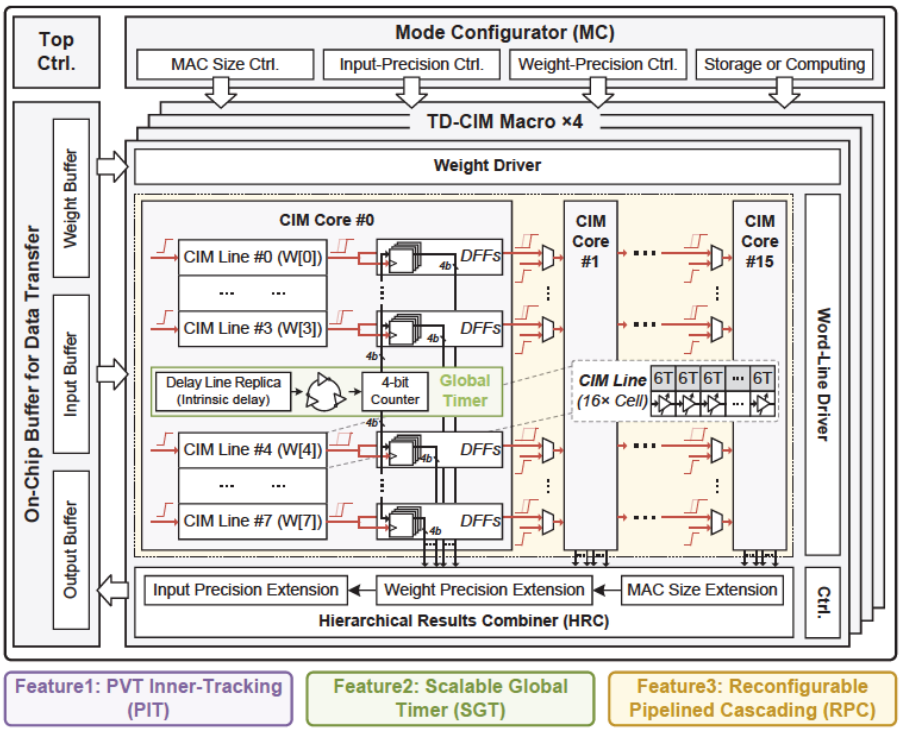
## Session 25 Energy Efficient Mixed Signal CIM Circuits

이번 2025 IEEE A-SSCC의 Session 25는 Energy Efficient Mixed Signal CIM Circuits 라는 주제로 총 5편의 논문이 발표되었다. 높은 효율성을 달성하기 위해 아날로그 기반 CIM이 주로 채택되었으며 아날로그 연산의 고질적인 문제인 낮은 정확도, PVT 변동, IR drop, Sensing margin 부족 등을 해결하기 위한 아이디어가 제시되었다. 본 Review에서는 25-1, 25-2, 25-4, 25-5 총 4편을 리뷰하고자 한다.

**#25-1** 논문은 time-domain (TD)에서 multiply-accumulate (MAC) 연산을 하는 CIM 매크로로 낮은 전압에서도 높은 에너지 효율과 확장성을 제공하는 장점이 있지만, PVT (공정, 전압, 온도) 변화에 매우 민감하고, TD 양자화로 인한 추가회로가 에너지 및 면적에서 오버헤드를 유발하며. 다양한 CNN 모델에서 CIM의 utilization이 떨어져 에너지효율성 측면에서 최대 성능과 평균 성능의 차이가 크다는 문제점이 있었다. 이를 해결하기 위해 본 논문에서는 PVT Inner-tracking 방법을 도입해 CIM cell과 TD Quantizer의 PVT 반응을 동일하게 맞춰 PVT 강건성을 확보하고, 기존 TD Quantizer들의 높은 에너지소모와 큰 면적을 감소시킬 수 있는 SGT회로를 적용해 유연한 누적계산을 가능케 하였다. 마지막으로 RPC 기법을 통해 여러 CIM 라인을 유연하게 병렬연산시켜 속도 저하 없이 큰 누적계산을 처리하였다.

28nm 공정으로 제작된 본 논문의 칩은 4개의 매크로와 매크로당 16개의 코어로 구성되었다. -10% 에서 10% 까지의 전압변화와 -50℃ 에서 100℃ 까지의 온도변화에서도 MAC 에러가 거의 없는 PVT 안정성을 보여주었으며 4b 모드에서 82.82 – 236.5 TOPS/W, 8b 모드에서 20.4 – 58.7 TOPS/W의 매우 높은 에너지 효율을 달성하였다.

CIM 매크로에서 아날로그 연산의 정확도를 높이기 위한 회로를 PVT 강건설계를 중심으로 제안하였으며 정확도 측면에서 유의미한 결과를 남겼다. 그러나 edge device에 적용할 목적으로 depthwise CNN으로 테스트하였지만 이는 다소 제한적인 환경에서만 결과를 도출했으며, 좀 더 다양한 네트워크 환경에서의 정확도 입증이 필요해 보인다.



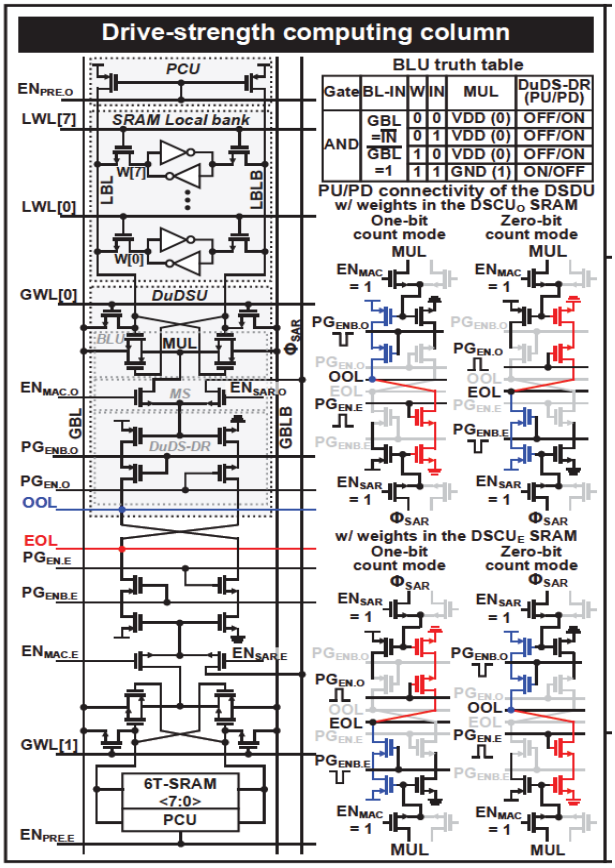

[그림 1] 논문 25.1의 제안한 TD-CIM 의 전체 구조

**#25-2** 논문은 SRAM 기반 아날로그 방식으로 MAC 연산을 하는 CIM macro이다. 기존 SRAM 기반 CIM 매크로는 vision transformer와 같이 전체 attention 연산을 수행하여 partial sum (PSUM) 오차가 쉽게 증폭되어 정확도 저하를 야기시키며, 특히 비트라인의 제한된 sensing margin으로 인해 PSUM 영역에서 오차가 계속 증가하고, 이는 shift-and-add 단계에서 이러한 에러가 누적 및 증폭되는 문제가 있었다. 이러한 문제를 해결하기 위해 본 논문에서는 Dual Drive Strength 기반 구조를 제안한다. PSUM 값이 커질수록 비트라인에 연결되는 스위치의 수가 많아져 sensing margin이 감소하게 되는데 one-bit/zero-bit count mode를 사용해 '1'의 개수에 따라 PSUM이 작을 때는 '1의 개수', PSUM이 클 때는 '0'의 개수를 읽는 방식으로 전환하여 sensing margin을 유지하게 한다.

제안하는 DuDS-CIM 매크로는 65nm LP CMOS에서 제작되었으며, 측정결과 높은 PSUM 영역에서 RMSE를 2.46에서 0.51로 낮추었으며 ResNet-20 (CIFAR-10) 에서 91.26%, ViT-S/16 (CIFAR-10) 95.57%의 높은 정확도를 달성하여 dense 및 sparse 모델 모두에서 안정적인 PSUM 연산이 가능함을 입증하였다. 에너지 효율은 942 TOPS/W, 면적 효율은 2.13 TOPS/mm^2, 동작 전압은 1.1V, 주파수 50MHz에서 동작한다.

SRAM 기반 CIM의 근본적인 취약점인 sensing margin 부족 문제를 PSUM 분포 기반의 대칭형 구조로 해결했다는 점에서 기술적 기여도가 높다. 특히, ViT 네트워크 환경에서 안정적인 동작을 실측한 것에 의의가 있다고 본다. 그러나 50MHz라는 상당적으로 느린

동작 주파수에서의 결과만 보여주고 있으며 shmoo plot을 게재하지 않았다는 점이 아쉬운 부분으로 남는다.

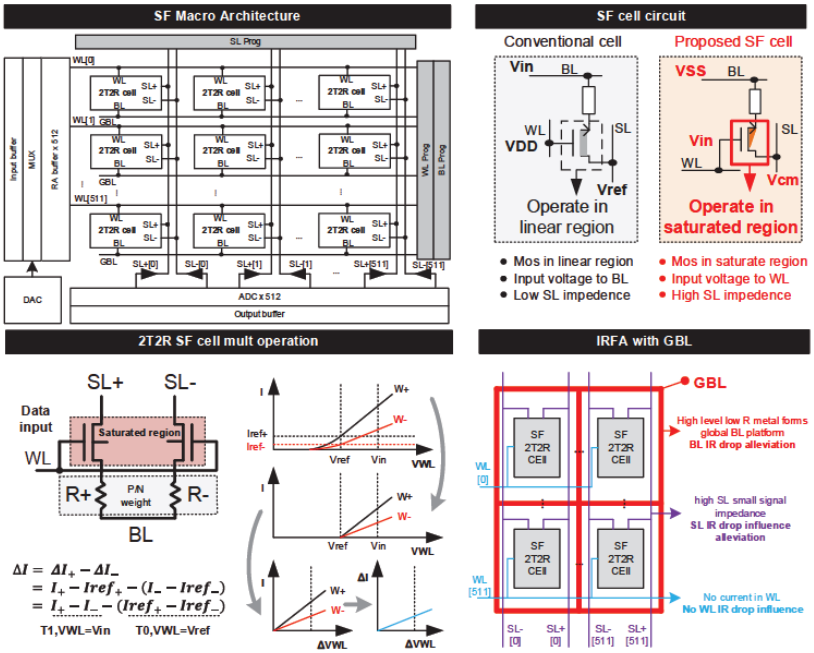



[그림 2] 논문 25.2의 제안한 Drive-strength computing column 회로

**#25-4** 논문은 RRAM 기반 CIM 매크로로, RRAM은 비휘발성 저장과 아날로그 MAC 연산을 수행할 수 있어 Edge 기반 장치에 차세대 메모리로써 각광받고 있지만 실제 하드웨어에서는 IR drop, weight cell의 선형영역 동작, 아날로그 MAC 연산의 결과전압 오차, 높은 주변부회로의 전력 소모 등의 구조적 문제가 있다. 이러한 문제를 해결하기 위해 본 논문에서는 Source-Follower (SF) 기반 weight cell과 IR-Drop-Freed Array (IRFA) 구조를 새롭게 제안한다. SF cell은 2T2R 구조의 트랜지스터를 포화영역에서 동작시켜 연산 전류를 MAC 연산의 결과전압과 분리시키고, IRFA는 1T1R의 각 라인(BL, SL, WL)에서의 전압 강하를 구조적으로 제거하여 대규모 병렬 MAC 연산에서도 안정적인 연산과 높은 처리량을 확보할 수 있도록 설계한 것이 특징이다.

제안하는 매크로는 28nm 공정으로 1Mb RRAM 어레이로 구성하여 17.0 TOPS/mm^2의 높은 연산 밀도를 달성했으며 RRAM 특유의 아날로그 변동성을 극복하여 실사용 모델에서 의미 있는 정확도를 유지하였다. 다만 실사용 모델에 대해서의 결과는 시뮬레이

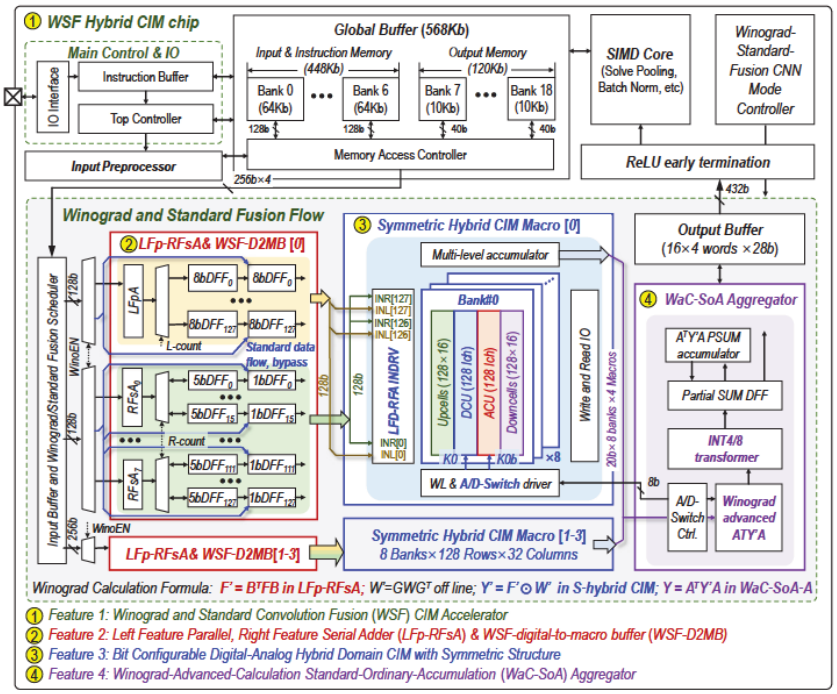
션을 기반으로 진행된 점이 아쉬운 점이다. 그럼에도 불구하고 RRAM 기반 CIM의 가장 근본적인 문제인 IR drop 과 전압 의존성 문제를 해결한 점에서 기술적 완성도가 있었으며, 추가적인 회로에 대한 면적과 파워 오버헤드에 대한 데이터, 병렬성(Parallelism)에 따른 데이터 또한 제시하고 있어 실제 칩 기반 end-to-end 연산 결과가 제시된다면 설득력이 더 강해질 것으로 판단된다.




[그림 3] 논문 25.4에서 제안한 2T2R 구조의 RRAM기반 CIM 매크로와 IRFA.

**#25-5** 논문은 CNN 연산에서 널리 사용되는 Winograd Convolution 연산 시 발생하는 하드웨어 오버헤드를 줄이는 회로를 제안한다. Winograd Convolution은 연산량을 최대 2.25배 절감할 수 있지만, 실제 하드웨어 적용에서는 추가적인 변환 연산과 메모리 요구량 증가, 정밀도 손실 등 해결해야 과제가 남아있다. 이를 해결하기 위해 본 논문은 Winograd Domain (WD) 와 Standard Domain (SD) 연산을 하나의 하드웨어 안에서 효율적으로 결합시키기 위한 Winograd-Standard Fusion 아키텍처를 제안한다. 이 아키텍처는 CIM 내에서 WD/SD 양쪽 계산을 모두 처리할 수 있도록 구성하였다. 특히 WD transform에 유리한 부분과 SD에서 정밀도를 유지해야 하는 부분을 대칭적으로 그룹화된 CIM 도메인에서 분산 수행하여 정확도 손실을 최소화하면서도 전체 CNN 연산 효율을 극대화하도록 설계하였다.

제안하는 CIM 매크로는 28nm CMOS 공정에서 제작되었으며, 기존 Winograd 전용 가속기 대비 변환 오버헤드를 줄이고, SD 연산보다 높은 연산 밀도를 확보하여 244.45 TOPS/W의 매우 높은 에너지 효율을 달성하였다. 그러나 WD와 SD를 하나의 도메인에서 처리함으로써 발생하는 추가적인 회로의 오버헤드 및 차지하는 면적 비율 등의 추가적인 데이터가 필요해 보인다.




[그림 4] 논문 25.5의 제안한 매크로의 전체 구조

## 저자정보


**송충석 석박통합과정 대학원생**

- 소속 : 한양대학교
- 연구분야 : 딥러닝 가속기 설계
- 이메일 : scs940430@hanyang.ac.kr
- 홈페이지 : https://sites.google.com/site/dsjeonglab1/home

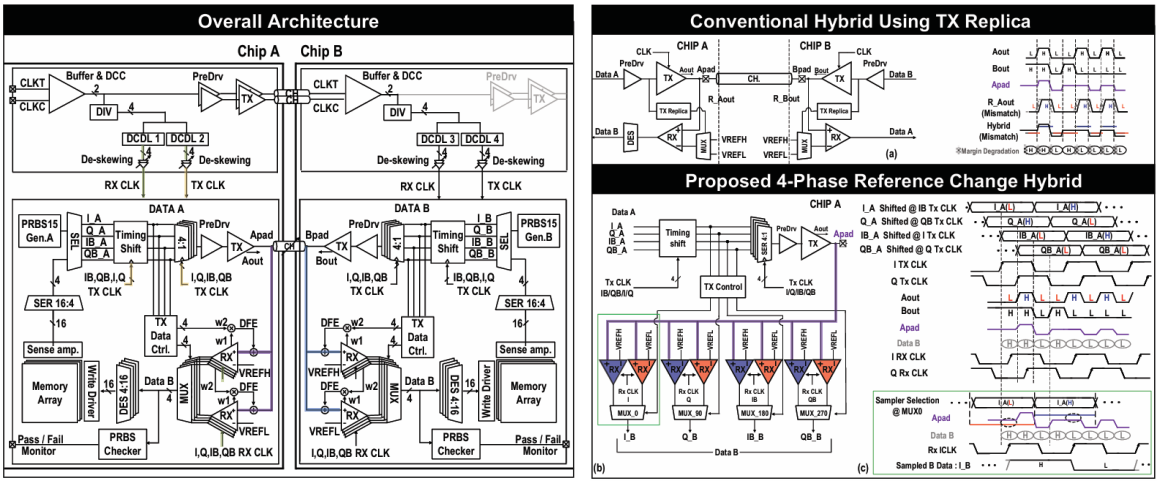
# A-SSCC 2025 Review

한국과학기술원 전기및전자공학부 박사과정 윤웅노

## Session 29 High Speed Circuit and Interface for Memory

이번 2025 IEEE A-SSCC의 Session 29에서는 High Speed Circuit and Interface for Memory 라는 주제로 총 4편의 논문이 발표되었다. 발표된 논문 중 3편(29.1, 29.2, 29.4)는 기존 채널들의 물리적인 한계를 극복하기 위하여 NRZ 방식을 탈피하거나 개선하는 연구이다. 나머지 한 편(29.3)은 high speed SRAM 동작에 필수적인 ECC 회로에 대한 논문으로 고속 동작으로 생기는 에러에도 데이터의 신뢰성을 보장하기 위한 연구를 보여주었다.

**#29-1** 해당 논문은 좁은 면적에서 고속 통신을 요구하는 차세대 HBM4 인터페이스를 위한 기술로 주목받는 SBD (Simultaneous Bidirectional) signaling에서 생기는 문제들을 해결하고자 하였다. SBD signaling으로 핀 당 통신 속도를 2배로 높였으나 channel delay로 인하여 신호 파형이 복잡하여 signal eye가 좁아지는 문제가 존재하였다. 이를 해결하기 위해 첫째, timing alignment 기법을 적용하였다. DCDL을 이용하여 channel delay를 타겟 주파수의 1/2 UI 배수가 되도록 조절함으로써, superimposed signal의 파형을 단순화시켜 신호 복원 난이도를 낮췄다.



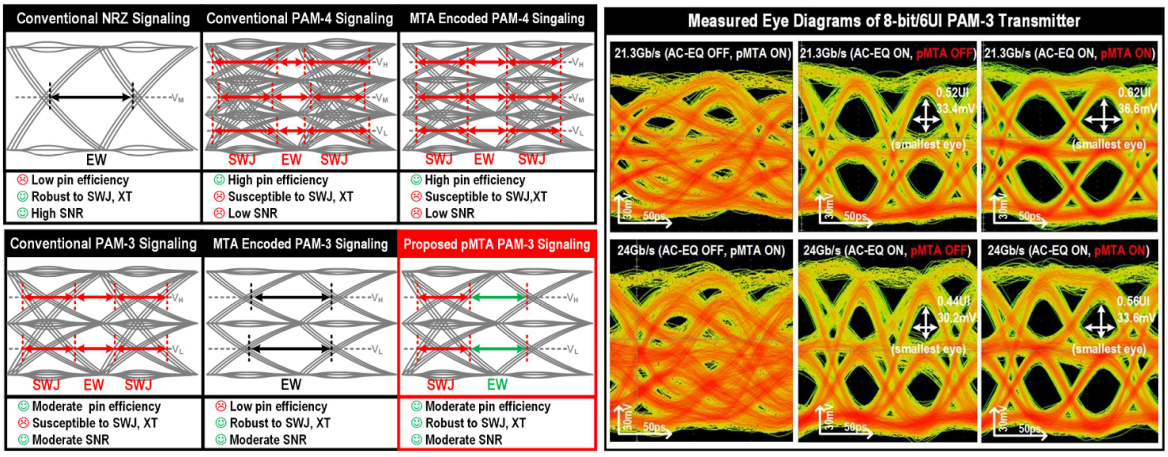

[그림 1] 제안된 SBD signaling 방식의 전체 아키텍처 (좌), 기존 TX replica 방식과 제안된 4-Phase 기준 방식의 비교 (우)

둘째, 기존 TX Replica 방식의 mismatch 한계를 극복하기 위하여 4-phase hybrid scheme을 제안하였다. 이는 replica 회로 없이, 이미 알고 있는 4-UI 단위의 송신 데이터에 따라

서로 다른 기준 전압을 가진 수신기를 선택하는 방식이다. 이를 통하여 타이밍 마진을 대폭 개선하였다. 마지막으로, dual equalization과 Ron 최적화를 적용하였다. IR drop의 영향을 줄이기 위해 TX driver 저항을 45 ohm으로 최적화하고, 이로 인해 약해진 신호 구동력과 ISI 이슈는 수신 신호용 DFE와 송신 신호용 XOR 기반 equalizer를 결합하여 보상하였다. 이러한 방식으로 18.4Gb/s/pin의 속도를 달성할 수 있었다.

**#29-4** 차세대 메모리 인터페이스는 높은 데이터 전송율에 더불어 에너지 효율성을 극대화하는 것이 핵심 과제이다. 이를 위해서 NRZ 대비 높은 대역폭 효율을 지니는 다중 레벨 signaling을 위한 PAM-4, PAM-3 구조가 도입되었으나 그 한계가 존재하였다. PAM-4는 eye height가 NRZ의 1/3로 줄어들어 SNR이 낮고, switching jitter (SWJ)와 crosstalk(XT)에 취약해 signal integrity (SI) 성능 하락 이슈가 있다. PAM-3는 PAM-4에 비해서는 좋은 SNR을 지녔으나 여전히 SI 하락 문제가 남아있었다. 이에 더불어 기존 PAM-3 방식은 이론적으로 NRZ 대비 150%의 속도 향상이 가능하지만, DDR5나 HBM3와 같은 실제 메모리 시스템의 burst length가 2의 지수 단위로 동작하기 때문에 데이터 매핑 시 실질적인 효율은 133%로 제한되는 구조적 비효율성이 존재하였다.

본 논문은 메모리 인터페이스에서 다중 레벨 송신기를 구현함에 있어서 생기는 이슈들을 저전력으로 해결하고자 partial Maximum transition avoidance (pMTA) 방식을 제안하였다. 기존의 MTA 방식과는 다르게, 신호가 2-UI 내에서의 최대로 변화하는 경우를 제거하도록 LUT를 통해 인코딩한다. 잘 최적화된 encoding mapping을 통하여 전반적인 전류 소모를 낮추면서도, 8-bits/6-UI가 가능하도록 구현하였다. 이는 곧 data rate per pin은 유지하며 switching jitter가 발생하더라도 eye width가 기존에 비하여 13.8%P 향상되는 결과로 이어졌다.




[그림 2] NRZ, PAM-4, PAM-3 방식과 TMA, pMTA에 따른 eye diagram의 비교 (좌), Signaling 방식에 따라 측정된 eye diagram (우)

또한, 해당 논문은 0.54pJ/bit의 최고 수준의 에너지 효율과 0.051pJ/bit/dB의 FoM을 달성하였는데 이는 0.3V의 무척 낮은 VDDQ를 이용하였기 때문으로 사료된다. 특히, 낮은 VDDQ에서 드라이빙을 위하여 N-over-N voltage-mode driver를 사용하여 low voltage swing terminated logic (LVSTL)을 구현한 것도 주목할만한 부분이다.

## 저자정보


**윤응노 박사과정 대학원생**

- 소속 : 한국과학기술원 전기및전자공학부
- 연구분야 : Sensor Interface ICs, Frequency Generation ICs
- 이메일 : voogi3925@kaist.ac.kr
- 홈페이지 : https://impact.kaist.ac.kr/

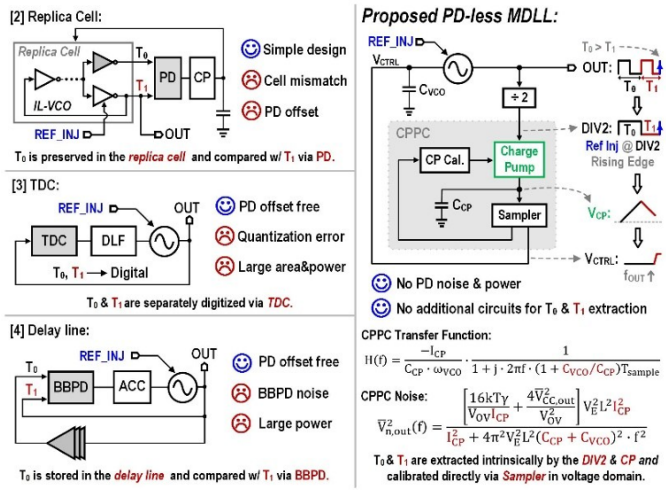
# A-SSCC 2025 Review

한국과학기술원 전기및전자공학부 박사과정 윤웅노

## Session 6 High-Performance Frequency Generation

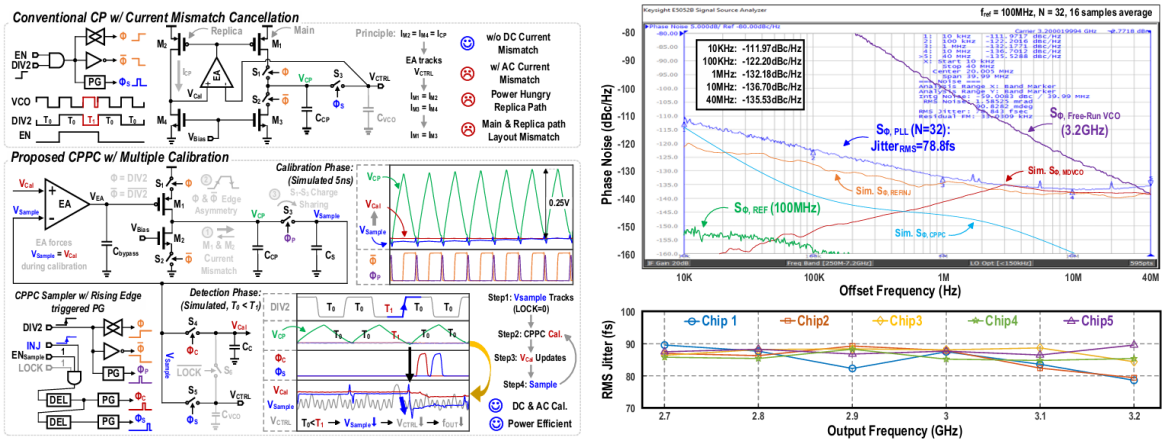
이번 2025 A-SSCC의 Session 6, High-Performance Frequency Generation은 총 4 편의 논문이 발표되었다. 6.1과 6.2에서는 spur와 noise를 줄이기 위한 보상 기법을 도입하였으며, 6.4는 sampling PLL에서 DSM으로 인한 quantization noise cancellation을 위한 방안을 제안하였다. 또한, 6.1과 6.4에서는 각각 PD-less, DTC-less 구조를 제안하여 기존에 사용하던 아날로그 블록들을 제거하고 그 역할을 내재화하려는 방향성을 보여주었다.

**#6-1** Multiplying delay-locked loop(MDLL)은 injection locked loop의 한 종류로, ring oscillator(RO)의 phase noise suppression과 동시에 넓은 BW를 챙길 수 있는 장점을 가진다. 그러나 reference frequency를 직접 injection 시켜주는 경우 reference spur가 크게 유발되는 이슈가 있다. 이를 해결하기 위해 최근 연구들은 frequency tracking loop (FTL)을 도입하여 RO free-running 주기 (T0)를 reference 주기와 일치시키려는 시도를 해왔다. 대표적인 예시로는, 1) 직접 RO에 injection 하는 것이 아닌, replica cell에 injection을 시킨 후 PD/CP를 통해 차이를 보정하는 방식, 2) TDC를 이용하여 reference와 T0 각각을 디지털 값으로 변환하여 보정하는 방식, 3) T0와 동일한 크기의 delay를 가진 delay line을 도입하여 T0를 저장하고 reference와 비교하는 방식 등이 존재하였다. 하지만 이들은 각각 cell mismatch에 취약함, TDC의 resolution-power trade-off, BBPD의 noise 기여와 delay line의 resolution-power trade-off라는 한계를 지니고 있다.




[그림 1] 기존 MDLL에서의 FTL들 (좌), 제안된 PD-less MDLL (우)

본 논문은 PD-less RO-based MDLL 구조에 charge-pump phase corrector (CPPC)를 통한 FTL을 구현하여 reference spur를 억제하면서도 noise source를 최소화하는 FTL을 제안하였다. 65nm 공정에서 0.014mm$^2$의 작은 면적으로 구현되었으며, 78.8fs의 우수한 RMS jitter 성능을 달성하였다. 핵심이 되는 CPPC는 reference가 injection되는 경우 CP의 up-down pulse width가 바뀌는 특성을 이용하여 시간 차이를 전압으로 바꾸는 역할을 한다. 그렇게 바뀐 전압 값을 기준으로 RO의 주파수를 변경하여, 결과적으로는 reference 주파수와 동일하게 동작하도록 한다. 회로 구현 측면에서는 charge sharing에 의한 에러를 방지하기 위해 calibration phase를 도입하여 전압을 유지하는 기법과, 홀수 N 배수 동작 시 위상 불일치로 인한 오동작을 막기 위한 mode selection 기법을 적용하여 완성도를 높였다.

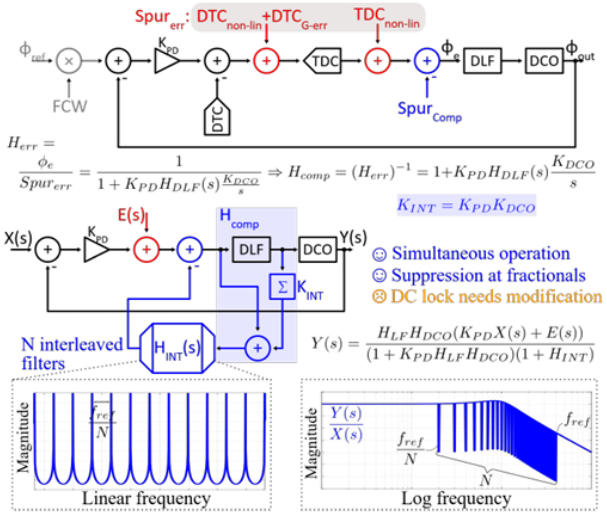



[그림 2] 기존 CP 전류 mismatch cancellation 구조와 제안된 CPPC에서 기울기/전류 mismatch의 해결 (좌) 측정/시뮬레이션 phase noise 결과와 5개 칩에 대해 측정된 RMS jitter (우)

성능적으로는 -250.2 FoM$_R$을 달성하여 power 소모 대비 매우 우수한 RMS jitter 값을 지닌다. 특히, power breakdown을 보아도 79%에 해당하는 전력을 RO가 소모하고 있기에 noise-power 최적화가 잘 이루어진 설계로 판단된다. 다만, 홀수 N 모드에서는 Loop BW가 절반으로 바뀌는 특성이 있으며, comparison table에 제시된 RMS jitter 값은 측정된 5개의 샘플 중 가장 좋은 성능을 보인 칩의 결과라는 점은 성능 해석에 감안해야 할 부분이다.

**#6-2** Fractional-N Digital PLL (DPLL)은 scalability와 작은 면적 측면에서 유리하여 무선 transceiver를 위해 많이 사용되지만, 신호 왜곡과 간섭을 유발하며 phase noise 성능을 저하시키는 fractional spur가 고질적인 이슈로 지적되어 왔다. Fractional spur는 주로 PD의 non-ideality로 인하여 발생하며, 기존의 digital pre-distortion (DPD)를 통해 해결하는

접근은 DTC의 정확도, dithering을 쓰는 방법은 높아진 noise floor가 issue로 존재하였다. 본 논문은 이러한 한계를 all-digital spur compensation (ADSC) 기법을 통하여 해결하고자 하였다. 이 기법은 에러 신호가 loop-dynamics에 포함되어 직접 조정하기 어렵다는 점에서 착안하여, 디지털 신호에 inverse transfer function을 적용해 보상 신호를 생성한다.




[그림 3] TDC 기반 DPLL 모델링과 제안된 ADSC의 원리 모식도

구체적으로는, N개의 interleaved integrators를 사용하여 PLL의 transfer function에서 fractional tone이 나타나는 지점마다 zero를 만들어 각각의 spur를 필터링하는 방식이다. 이러한 방식으로 이전의 아날로그 DTC calibration과는 다르게 그 어떤 왜곡 신호이든 주기적이면 모두 보상할 수 있기에, 복잡한 LMS 알고리즘 없이도 DTC gain error, 신호의 기울기 왜곡, ADC에서의 cap. Mismatch 등을 모두 보상할 수 있다는 점에서 장점을 지닌다. 결과적으로는 65nm 공정에서 0.26mm$^2$의 면적으로 211fs RMS jitter에 -68.1dBc fractional spur의 우수한 성능을 보여주었다.

## 저자정보

**윤응노 박사과정 대학원생**

- 소속 : 한국과학기술원 전기및전자공학부
- 연구분야 : Sensor Interface ICs, Frequency Generation ICs
- 이메일 : voogi3925@kaist.ac.kr
- 홈페이지 : https://impact.kaist.ac.kr/

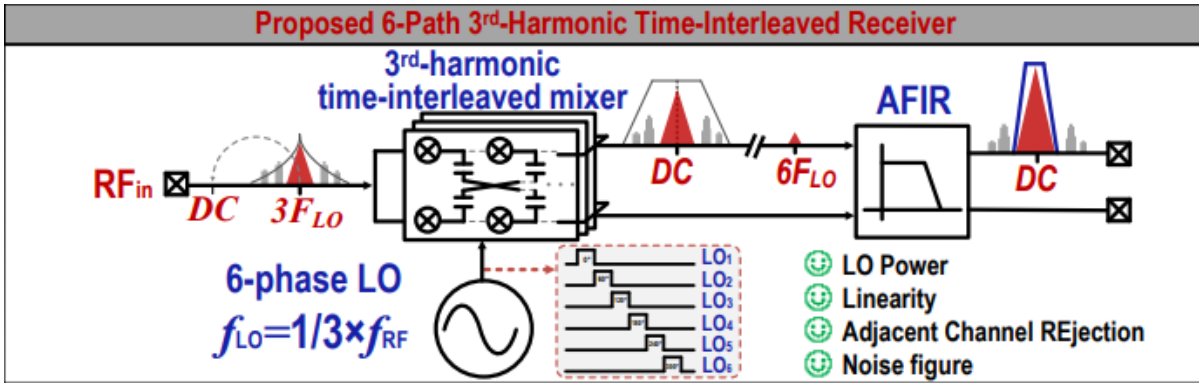
# A-SSCC 2025 Review

단국대학교 파운드리공학과 석사과정 임재영

## Session 10 Low-Power Transceivers

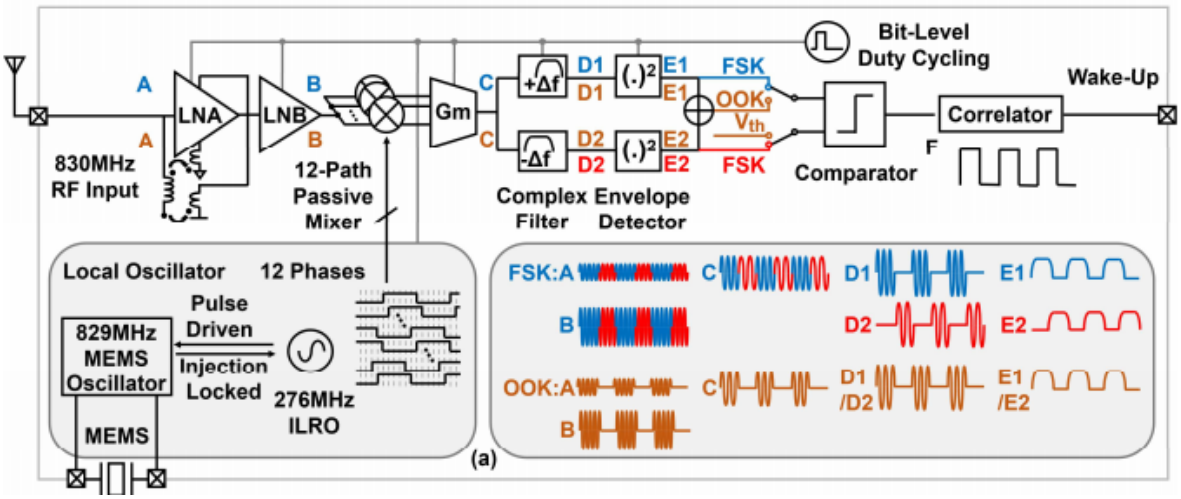
이번 A-SSCC 2025의 Session 10에서는 저전력 무선 수신기 및 송수신기를 주제로 총 4편의 논문이 발표되었다. 고선형성 RF 수신기, always-on wake-up receiver(WuRX), multi-band LPWAN 트랜시버, polar receiver 구조 등 다양한 저전력 무선 통신 아키텍처가 소개되었다. 본 리뷰에서는 #10.1과 #10.2 논문의 회로 구조와 동작 방식을 중심으로 정리한다.

**#10-1** 본 논문은 중국 톈진대학교에서 발표한 논문으로, BLE 수신기에서 선형성과 전력 소모를 동시에 고려한 passive mixer 기반 RF front-end 구조를 제안하였다. 기존 zero-IF BLE 수신기에서는 RF와 동일한 주파수의 LO를 사용하는 passive mixer 구조로 인해, LO 고조파 성분과 switching nonlinearity로 인해 강한 blocker 환경에서 선형성 저하가 발생하는 문제가 있다. 이를 완화하기 위해 본 논문에서는 3rd-harmonic time-interleaved passive mixer 구조를 도입하였다. 제안된 수신기는 6-phase non-overlapping LO를 이용한 6-path mixer를 기반으로 하며, LO 주파수를 RF 주파수의 1/3로 낮추어 LO buffer의 동작 주파수와 전력 소모를 감소시킨다. 또한 time-interleaving을 통해 mixer 스위칭에 따른 비선형 성분이 시간적으로 분산되도록 구성하였다. Mixer 이후에는 reconfigurable AFIR(Analog FIR) 필터를 적용하여 baseband 대역폭 설정과 인접 채널 신호 억제를 수행한다. AFIR 구조를 통해 baseband 필터링을 아날로그 영역에서 구현함으로써, 높은 선택도와 저전력을 동시에 달성한다. 측정 결과, 본 수신기는 OOB_IIP3 26.5dBm, SFDR 80dB, 전력 소모 415uW를 달성하였다.




[그림 1] 제안된 6-path 3rd-Harmonic Time-Interleaved Receiver

**#10-2** 본 논문은 싱가포르 난양공과대학교와 ASTAR에서 발표한 논문으로, OOK 및 FSK 신호를 대상으로 하는 초저전력 wake-up receiver(WuRX) 구조를 제안하였다. 기존 WuRX 구조에서는 RF front-end와 LO가 연속적으로 동작하는 경우가 많아, 항상 활성화된 상태에서의 전력 소모가 시스템 수명을 제한하는 요인으로 작용한다. 특히 연속 발진하는 LO와 RF front-end는 wake-up 신호가 존재하지 않는 대부분의 시간 동안에도 전력을 소모하게 된다. 이를 줄이기 위해 본 논문에서는 sub-sampling 기반 RF front-end와 duty-cycled 동작 방식을 채택하였다. RF 입력 신호는 12-path passive mixer를 통해 시간 영역에서 sub-sampling되며, 이후 저주파 성분으로 변환된다. 이 과정에서 FR front-end는 연속 동작하지 않고, 신호 검출에 필요한 구간에서만 활성화된다. Mixer 이후에는 complex envelope detector가 배치되어 RF 신호의 진폭 정보를 추출한다. OOK 신호의 경우 진폭 변화가 직접적으로 envelope에 반영되며, FSK 신호의 경우 주파수 차이에 따른 envelope 패턴의 차이가 후단에서 구분된다. 추출된 envelope 신호는 comparator를 통해 디지털 신호로 변환되며, correlator에서 미리 정의된 패턴과 비교되어 wake-up 여부가 결정된다. LO 생성부에는 pulse-driven MEMS oscillator가 사용되었다. MEMS 공진기는 injection-locked 방식으로 구동되며, 연속 발진 대신 펄스 형태의 구동 신호를 이용하여 평균 구동 에너지를 줄인다. 또한 12-phase ILRO 구조를 통해 sub-sampling mixer 구동에 필요한 non-overlapping clock이 생성된다. 제안된 WuRX는 RF front-end, LO, baseband 블록이 모두 bit-level duty cycling 방식으로 동작하도록 구성되어 있다. 측정 결과, 본 수신기는 930nW 전력 소모에서 OOK -93dBm, FSK -90dBm의 sensitivity를 달성하였다.




[그림 2] 제안된 pulse-driven MEMS oscillator 기반 WuRX 구조

## 저자정보

**임재영 석사과정 대학원생**

- 소속 : 단국대학교
- 연구분야 : clock generators
- 이메일 : ljy72250338@dankook.ac.kr
- 홈페이지 : https://sites.google.com/dankook.ac.kr/acs-lab

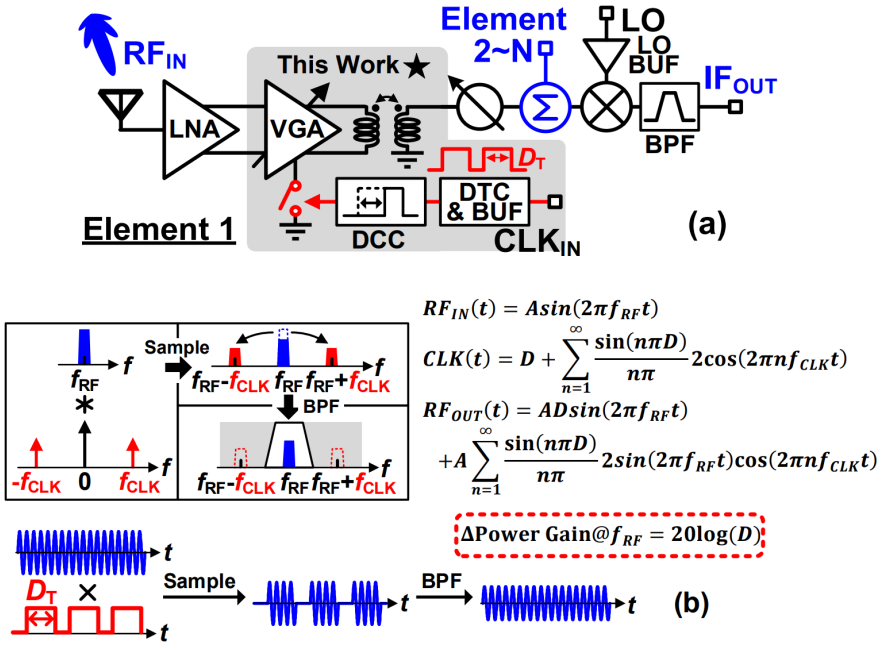
# A-SSCC 2025 Review

고려대학교 전기전자공학부 석사과정 심승우

이번 2025 IEEE ASSCC Symposium에서는 Wireless 관련 다섯 개의 세션이 열렸다. 이 중 본 리뷰에서는 Session 13에서 K,Ka-band 대역의 phase array system 두 편, 그리고 Session 17에서 Ka-band oscillator 논문까지 총 세 편을 다룬다

## Session 13 Phased-Array System and Components

**#13-1** 이번에 리뷰할 논문은 도쿄공업대학의 Kenichi Okada 교수님 그룹에서 발표한 "A Ka-Band Time-Modulated Variable Gain Amplifier with 30-dB Gain Tuning and <0.1-Degree Phase Variation via Duty Cycle Control"이다. 6G 및 위성 통신을 위한 Phased array system에서 gain control과 beamforming을 위해 variable gain amplifier(VGA)가 필수적이다. 하지만 current-steering이나 gilbert-cell 같은 기존의 기법들은 gain tuning range가 제한적이고, gain이 변할 때 phase variation이 심한 단점이 있습니다. 본 논문은 이러한 문제를 해결하기 위해 Time-Modulated Scheme를 적용했습니다. 핵심 아이디어는 별도의 clock signal의 duty cycle을 조정하여 VGA의 gain을 바꾼다는 아이디어이다. 그림 1에서 볼수있듯 rf신호가 clock 신호에의해 sampling이 되는 과정에서 원치 않는 harmonic 성분들이 발생하지만, 추가적인 band pass filter를 통해 원하는 RF 신호만 출력한다는 것이다. 이 방식의 이론적 차이는 Duty Cycle(D)에 따라 20logD로 결정된다. 또한 duty cycle의 정밀한 조정을 위해 그림 2에 나오는 dcc(duty cycle control) loop과 PMC(pulse modification cell)을 설계 및 구현하였다. CMOS 65nm 공정으로 설계되었고, 코어 면적은 1150 um x 280 um이며, gain에 따라 4.9 mW에서 103 mW의 파워를 소모한다. 측정결과 maximum gain은 29.6dB, 24 GHz에서 29.7 GHz의 3dB bandwidth와 DCC를 통해 30 dB의 gain control range를 가지고 또한 gain이 변할 때 phase는 0.1도 미만으로 변했으며 OP1dB는 28GHz에서 12.5dBm이 측정되었다. time modulated 구조를 통해 ka-band에서 30dB의 넓은 gain control range내에서도 phase variation이 0.1도 미만으로 유지되어, 우수한 phase stable한 기법을 제안하였다.




$$RF_{IN}(t) = A\sin(2\pi f_{RF}t)$$

$$CLK(t) = D + \sum_{n=1}^{\infty} \frac{\sin(n\pi D)}{n\pi} 2\cos(2\pi n f_{CLK}t)$$

$$RF_{OUT}(t) = AD\sin(2\pi f_{RF}t) + A\sum_{n=1}^{\infty} \frac{\sin(n\pi D)}{n\pi} 2\sin(2\pi f_{RF}t)\cos(2\pi n f_{CLK}t)$$

$$\Delta\text{Power Gain}@f_{RF} = 20\log(D)$$

그림 1.(a) Architecture of the time-modulated RX; (b) Illustration of time-modulated VGA operation in frequency and time domains

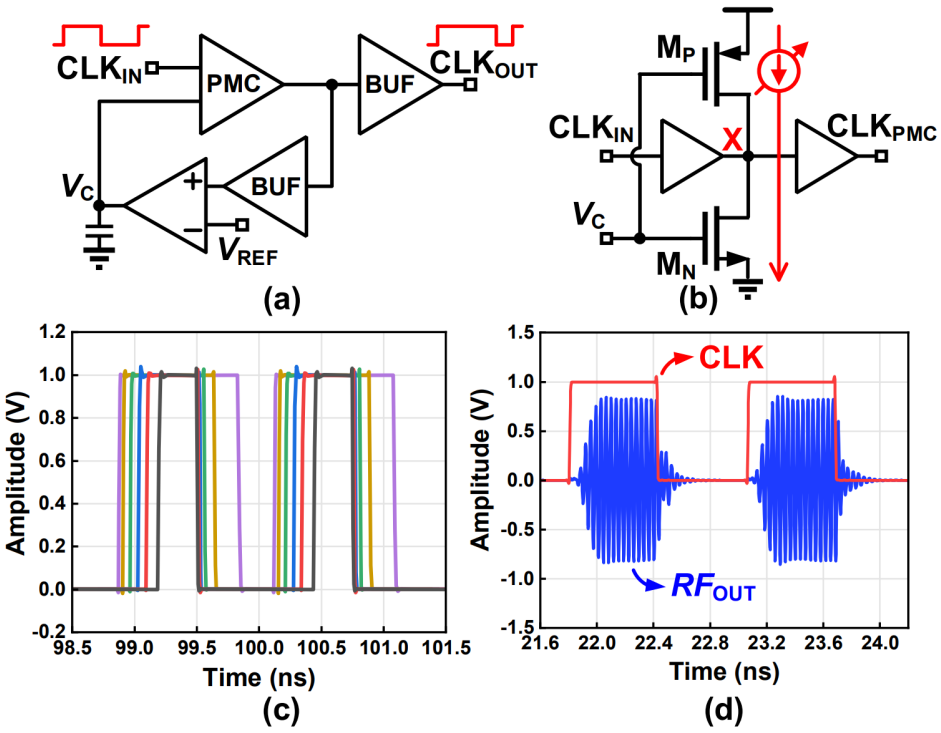



그림 2. Schematics of (a) DCC loop and (b) pulse modification cell; Simulated results of (c) duty cycle tuning by VREF and (d) transient

**#13-2** 이번에 리뷰할 논문은 칭화대학교의 Baoyong Chi 교수님 그룹에서 발표한 "A K-Band 8-Beam 4-element Phased-Array Transmitter Using GCPW-based Beam-combining Network and Compact 8-shaped Coils for Satellite Communications"이다. 저궤도(LEO) 위성통신과, 5G 비지상망(NTN)의 down-link로 사용되는 K-band에서는 높은 path loss로 인해 multi-beam phased array system이 필요하다. 하지만 on-chip multi beam 시스템을 구현시에 다수의 functional block 집적에 따른 레이아웃의 복잡성, Block 간 isolation의 어려움, 우주 극한 환경에서의 신뢰성 문제가 있다. 본 논문은 이것을 해결하기위해 GCPW(grounded coplanar waveguide) wiring 구조와, 개선된 8-shaped coil technique를 제안하였다. 그림3에 전체 블록도처럼 8개의 beam이 들어오고, 각 4개의 channel을 거쳐 32개의 output이 합쳐져 4개의 balanced power amplifer를 통해 신호가 출력되는데, 이때에 Flip chip packaging에 사용되는 bump를 path사이에 배치하여 isolation을 향상시켰고 우주 환경의 방사선에 의한 SEU(single event upset) 오류를 방지하고자 Majority Voting Logic 기반의 Fault-tolerant Resister cell을 설계하여 SPI(serial peripheral interface)의 신뢰성을 높였다. 그림 GCPW combining network에 jumper들이 line의 bend마다 대칭적으로 배치하여 gain matching을 향상시키고 dummy trace들로 signal간의 coupling을 막아 0.15 dB의 gain mismatch 감소와 17dB의 Inter-beam isolation을 향상시켰으며, 그림 3에 나와있는 개선된 8-shape coil구조를 통해 34%의 면적감소와 0.7dB의 Attenuation 정확도 향상을 보였다. CMOS 65nm 공정으로 설계되었고 전체 칩은 7.32mm x 4.52mm이며 채널당 45.3mW를 소모하였다. 측정결과 17.2 GHz에서 21 GHz의 3dB bandwidth와 32개 채널간의 0.8dB의 Gain mismatch를 보였다. 제안된 구조를 통해 동시에 8개의 beam을 지원하는 phased array transmitter를 구현하였고 multi beam beamformer 디자인에 효율적인 기법을 제안하였다.

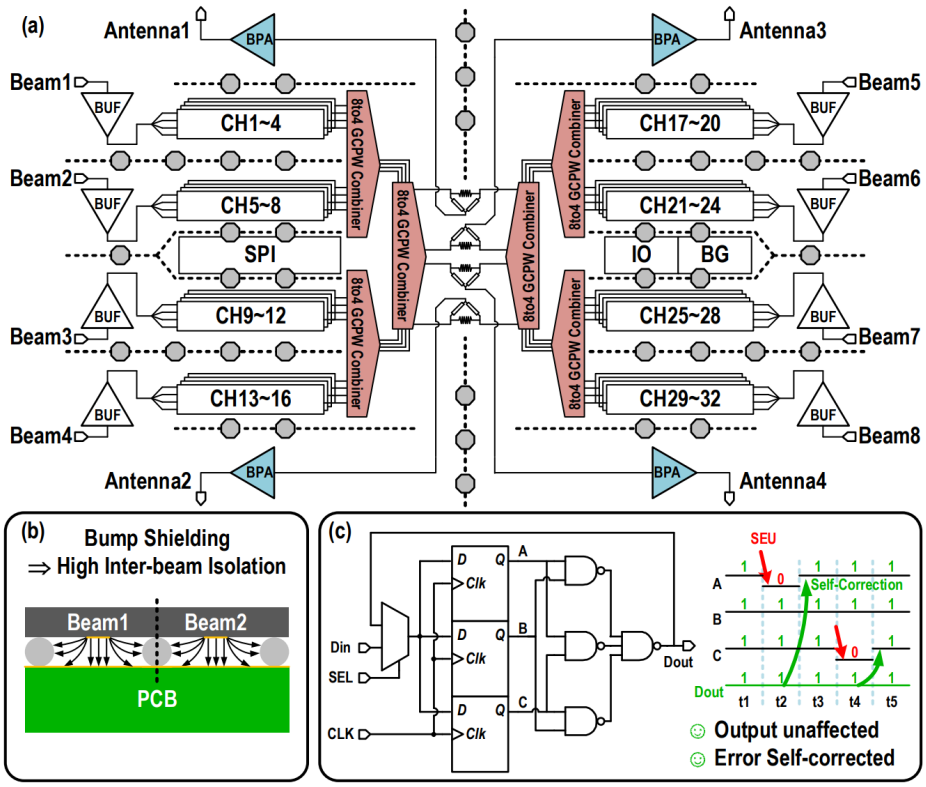
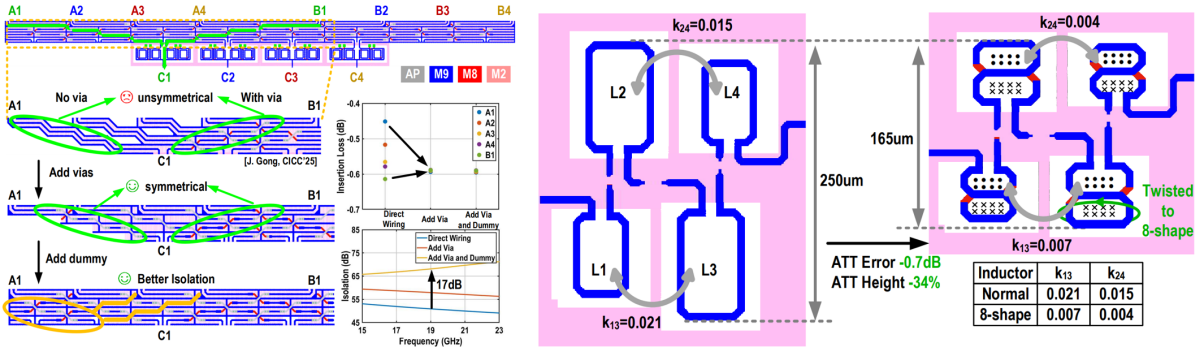



그림 3.(a) Block diagram of the proposed phased-array TX (b) bump shielding between beams (c) SEU-resilient register cell.




(a) (b)

그림 4.(a) The proposed GCPW-based 8-to-4 beam-combining network and comparison of different GCPW wiring schemes, (b) The proposed 8-shape coil technique.

**Session 17 Oscillators**

**#17-2** 이번에 리뷰할 논문은 중국 과학기술대학교(USTC)의 Yizhe Hu 교수님 그룹에서 발표한 "A 25.48-29.25 GHz Rotary Traveling-Wave Oscillator Achieving -191 dBc/Hz FoM at 10MHz Offset Using Ring-Interleaved N/P Cross-Coupled Pairs in 22-nm CMOS"이다. 6G 통신이나 AI 프로세서같은 시스템에서 High-speed multi-phase clock generation이 필수적이다. RTWO(rotary traveling wave oscillator)는 multi-phase 신호를 생성하는데 구조적으로 장점이 있지만, Switched capacitor(sw-cap)와 -Gm cell이 결합되어 있었는데 그로인해 기존 LC oscillator에 비해 Phase noise성능이 안좋다는 단점이 있습니다. 본 논문은 Ring-Interleaved Nmos/Pmos Cross-Coupled Pairs 구조를 제안했습니다. 기존의 Back to back 인버터 구조가 8개의 코어를 이용하여 8개의 differential 신호 즉 16개의 phase를 생성한 반면, 본 논문은 4개의 Nmos pair와 4개의 Pmos pair를 Interleave하여 배치하였고, 그로인해 코어들간의 matching inductor인 Mobius Ring을 Common mode inductor로 사용하여 추가면적 없이 높은 Common mode impedance를 형성하여 Flicker noise upconversion을 억제하였고, sw-cap을 Nmos와 Pmos pair의 중앙에 배치하여 3차 하모닉성분이 Capactive path대신 Inductive path로 흘러들어가 추가적으로 Flicker noise upconversion을 막아 phase noise를 향상시켰다. CMOS 22nm공정을 이용하여 설계되었고 25.48GHz에서 29.25GHz로 tuning range가 13.7%이고 1MHz에서 -105.1dBc/Hz의 phase noise가 측정되었고, 전력소모도 8.8mW로 매우 낮아 RWTO 구조중에 처음으로 FoM(figure of merit)가 -190 dBc/Hz를 넘겼다. 본 논문은 Nmos,Pmos pair를 통해 LC oscillator 급의 성능을 가진 RWTO 기법을 제안하였다.

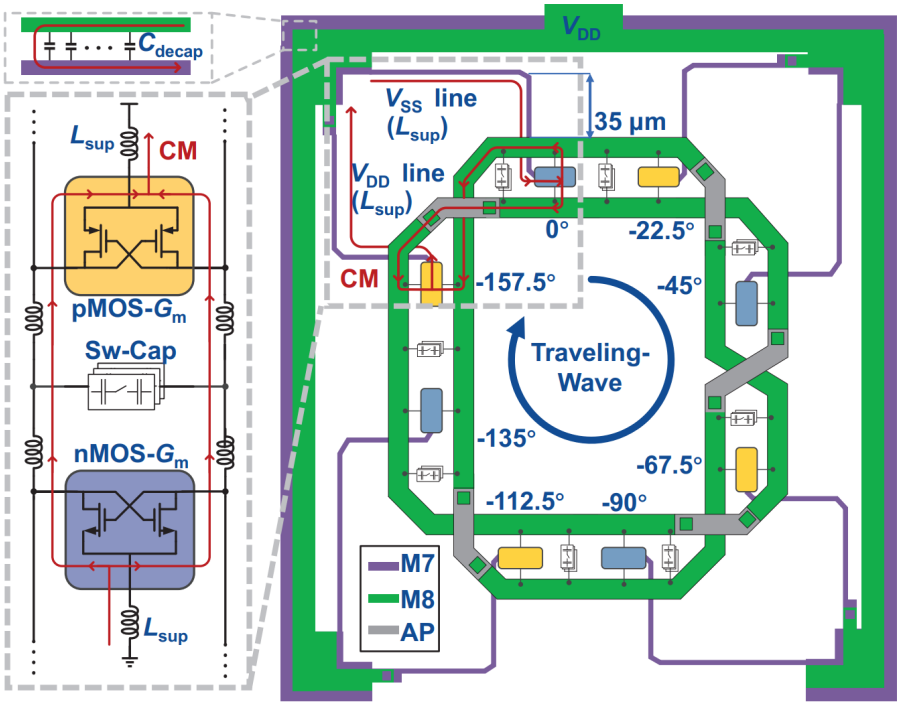



그림 5.(a) The proposed RTWO with ring-interleaved N/P cross-coupled MOS pairs and a dedicated supply scheme.

## 저자정보

**심승우 석사과정 대학원생**

- 소속 : 고려대학교 전기전자공학과
- 연구분야 : mm-Wave IC design
- 이메일 : lisang@korea.ac.kr
- 홈페이지 : https://arfsl.korea.ac.kr

# A-SSCC 2025 Review

KAIST 전기및전자공학부 석박사통합과정 이동윤

## Session 28 Advanced Transceivers

최근 무선 통신 시스템은 더 높은 데이터 전송률과 극한 환경에서의 안정성을 동시에 요구한다. Session 28에서는 이러한 요구에 부응해, 기존 수신기 아키텍처의 한계를 새로운 루프 제어 기술로 돌파한 Polar PT-RX와, 밀리미터파 대역에서 온도 변화에 따른 성능 저하를 원천 차단한 Temperature-Compensated LNA 기술이 발표되었다.

### #28-2 Coherent Polar PT-RX

28.2 "A Coherent Polar PT-RX for 32-APSK/16-QAM/GFSK Demodulation with High ACR and Relaxed I/Q Generation"은 기존의 Phase-Tracking Receiver (PT-RX)가 가진 구조적 한계를 극복하고, 고차 변조 방식(High-order Modulation)과 강력한 간섭 제거(ACR) 성능을 동시에 달성한 새로운 수신기 아키텍처를 제안한다.

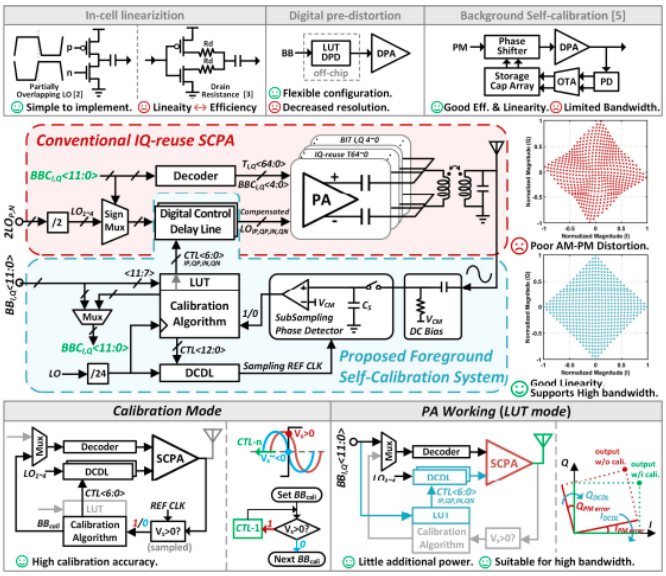
기존 PT-RX는 저전력 특성이 뛰어나지만, 위상 동기 루프(PLL) 기반 동작 특성상 Loop Bandwidth와 Stability(안정성) 간의 트레이드오프가 심했다. 대역폭을 좁히면 간섭 신호는 잘 거르지만 루프가 불안정해지고, 대역폭을 넓히면 루프는 안정되나 인접 채널 간섭(ACR)에 취약해지는 딜레마가 있었다. 또한, 위상 정보만 추적하기 때문에 진폭 정보가 포함된 QAM 같은 고차 변조 신호를 복조하기 어렵다는 치명적인 단점이 있었다.

이 논문은 두 가지 핵심 기술로 이 문제를 해결했다.

Additional Zero for Stability: 루프 내에 Delay-Locked Loop (DLL) 기반의 추가적인 Zero를 삽입하는 기법을 도입했다. 이를 통해 루프 대역폭을 좁혀 간섭 제거 성능을 극대화하면서도, 충분한 위상 마진을 확보해 시스템이 발진하지 않도록 안정성을 잡았다. 그 결과 Closed-loop RX 중 최고 수준의 ACR 성능을 달성했다.

Coherent Polar Demodulation: 진폭(Amplitude)과 위상(Phase)을 분리해 처리하되, 이를 다시 결합하여 복조하는 Coherent Polar 구조를 채택했다. 덕분에 기존 PT-RX에서는 불가능했던 32-APSK나 16-QAM 같은 진폭 변조가 포함된 신호도 정확히 복조할 수 있게 되었으며, I/Q 경로 간의 부정합(Mismatch) 문제에서도 자유로워졌다.

결과적으로 이 수신기는 2.4GHz 대역에서 동작하며, 기존 PT-RX의 저전력 장점을 유지하면서도 고성능 통신 규격을 만족시키는 차세대 수신기 솔루션으로 평가받는다.




[그림 1] 본 연구에서 제시하는 DPA 구조와 선행연구의 비교

**#28-5 Temperature-Compensated mmWave LNA (PILTOM)**

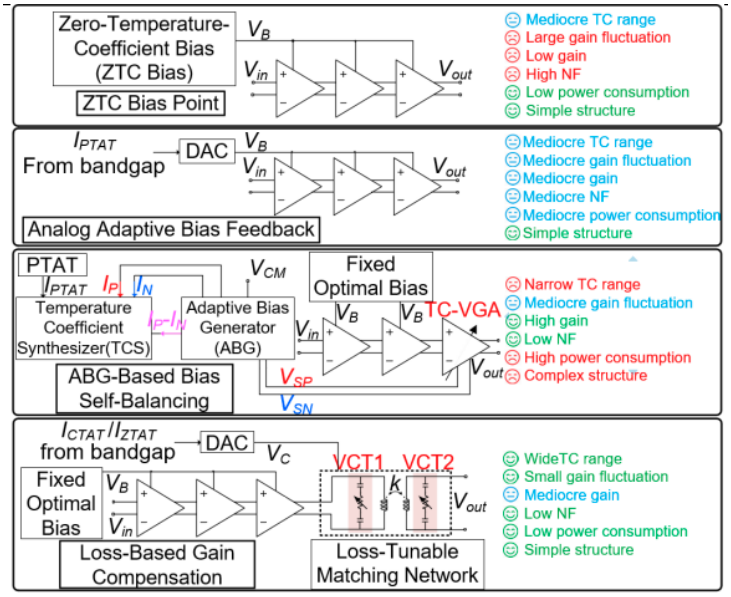
28.3 "A 26.2-41.3 GHz Temperature-Compensated LNA with Phase-Invariant Loss-Tunable Output Matching Achieving ±0.0011 dB/°C Gain Variation Across −55 °C to 125 °C"은 5G NR(28/39GHz) 및 위성 통신(Ka-band)을 위한 밀리미터파 LNA가 극한의 온도 환경(-55°C ~ 125°C)에서도 일정한 성능을 유지하도록 하는 혁신적인 보상 기술을 제안한다.

밀리미터파 시스템은 옥외 기지국이나 위성 탑재체처럼 온도 변화가 극심한 환경에 노출된다. 기존에는 온도가 변하면 트랜지스터의 바이어스 전압을 조절해 이득(Gain)을 맞추려 했으나, 이 경우 잡음지수(NF)나 선형성(Linearity)의 최적 동작점이 틀어지는 문제가 발생했다. 혹은 별도의 가변 이득 증폭기(VGA)를 추가해 이득을 보정하기도 했지만, 이는 전력 소모와 칩 면적을 증가시키고 위상(Phase)이 틀어지는 부작용을 낳았다.

핵심 원리는 다음과 같다. 트랜지스터의 바이어스는 온도가 변해도 항상 최적의 성능(최저 잡음, 최고 선형성)을 내는 지점에 고정한다. 대신, LNA의 출력 매칭 네트워크에 손실(Loss)을 미세하게 조절할 수 있는 기능을 넣어서, 온도가 낮아져 이득이 커지면 손실을 늘리고, 온도가 높아져 이득이 줄면 손실을 줄이는 방식으로 전체 이득을 일정하게 유지한다.

그리고 단순히 손실만 조절하면 위상이 변할 수 있는데, 이 회로는 가변 저항 역할을 하는 VCT와 Transformer를 정교하게 결합해, 손실을 조절해도 위상은 변하지 않도록 설계했다.

이 기술이 적용된 65nm CMOS LNA는 26.2~41.3 GHz라는 매우 넓은 대역에서 동작하며, -55°C에서 125°C까지 온도가 변하는 동안 이득 변화율이 ±0.0011 dB/°C에 불과할 정도로 완벽에 가까운 온도 보상 성능을 보여주었다. 이는 추가적인 전력 소모나 면적 증가 없이도 시스템 신뢰성을 획기적으로 높일 수 있는 기술로, 향후 6G 및 우주 통신 분야에서의 활용도가 매우 높을 것으로 기대된다.




[그림 2] 본 연구에서 제시하는 H-PCA 구조와 선행연구의 비교

## 저자정보

**이동윤 석박사통합과정 대학원생**

- 소속 : KAIST
- 연구분야 : Body-Channel-Communication Transceiver Design for Body-Area Network in Biomedical Application
- 이메일 : dongyoon.lee@kaist.ac.kr
- 홈페이지 : https://impact.kaist.ac.kr

# A-SSCC 2025 Review

KAIST 전기및전자공학부 석박사통합과정 이동윤

## Session 19 High-speed, High-resolution SAR and Pipelined ADCs

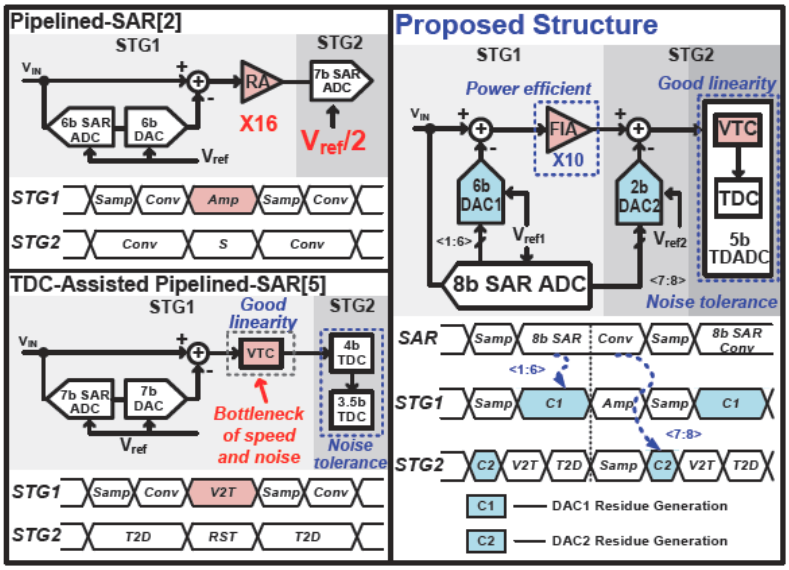
ADC 연구는 꾸준히 high speed, high resolution, low power를 동시에 달성하는 방향으로 진화해 왔다. 통신용 프론트엔드에서는 TI-SAR, pipelined-SAR, ring amplifier 기반, TDC-assisted, hybrid 구조 등이 고속·중고해상도 ADC의 주된 연구 축을 이루어 왔고, IoT·센서 인터페이스 영역에서는 NS-SAR, Incremental/Zoom ADC 등의 에너지 효율 중심 구조가 주로 연구되었다. 이와 병행해 CIS column ADC나 대규모 array 기반 구조, 그리고 cryo/edge/AI 가속기 등 특수 응용에 최적화된 비정형 ADC까지 등장하며, 응용 도메인별로 요구 사양과 설계 철학이 세분화되는 추세다. ASSCC 2025 Session 19는 이 중에서도 특히 통신·무선 SoC를 겨냥한 고속 ADC와, 멀티모드·array 구조를 통한 시스템 레벨 효율 향상에 초점을 맞춘 논문들로 구성되어 있다.

### #19-3 TDC-assisted Pipelined-SAR ADC

19.3 "A 3.4mW 64.5dB SNDR 800MS/s 12b Pipelined-SAR/TDC ADC with Parallel Amplification and Quantization"은 Pipelined-SAR에서 가장 까다로운 블록인 Residue Amplifier(RA)의 설계 부담을 줄이기 위해 등장한 TDC-assisted Pipelined-SAR 계열 구조가, 다시 VTC/TDC 쪽에서 speed·noise 병목을 맞는 상황을 정면으로 해소하는 논문이다. 기본 컨셉은 2-stage 구조에서 1단을 "SAR 기반 Coarse Quantizer + 단순 RA"로 두고, 2단을 "VTC/TDC 기반 Fine Quantizer"로 두어 고속성과 RA 요건 완화를 동시에 얻는 것이다.

이 논문에서 제안하는 구조는 1st stage에 Auxiliary 8b SAR ADC를 두고, 그 코드 중 상위 6b는 메인 MDAC(CDAC1)에 feedback하여 coarse residue를 만들고, 하위 2b는 RA 이후의 작은 CDAC2 쪽으로 넘겨 2 LSB를 동시(Parallel)로 처리하는 것이 핵심이다. 이렇게 하면 SAR 8b를 전부 MDAC에 직접 쓰지 않고 6b만 앞단 MDAC1에 쓰기 때문에, MDAC의 스위칭 횟수와 비교 횟수에 따른 지연을 줄일 수 있고, 나머지 2b는 RA 증폭 구간과 병렬로 진행되므로 전체 conversion latency를 단축할 수 있다. Residue는 단순 Open-loop Floating Inverter Amplifier(FIA)로 약 10배 정도 증폭된 뒤, 작은 CDAC2를 겸한 VTC에 의해 시간 영역으로 변환되고, 5b TDC가 fine quantization을 담당해 최종적으로 6+2+5−1(bit redundancy)=12b 출력을 만든다.

이 아키텍처에서 중요한 insight는 "TD ADC는 작은 입력 스윙에서 노이즈·선형성이 유리하다"는 점을 이용해, RA의 이득 요구와 선형성·잡음 요구를 크게 낮춘 것이다. 2nd stage가 SAR가 아니라 TDC이므로, 1st stage RA는 전통적인 고이득·저왜곡 ringamp 대신, 매우 단순한 open-loop FIA로 충분하고, 실제 설계에서도 전체 ADC 전력 3.4mW 중 약 11%만을 RA에 할당하면서 12b 급 정확도를 달성한다. 또한 VTC의 충전 커패시터를 메인 CDAC1과 분리해 작은 CDAC2(40fF)를 쓰도록 함으로써 V2T 변환 시간을 줄이고, 전체 변환 속도를 800MS/s까지 끌어올렸다. 결과적으로 이 구조는 28nm에서 단일 채널 12b, 800MS/s, 3.4mW, Walden FoM 4fJ/conv-step, Schreier FoM 173dB를 달성해 Wi-Fi 7(4K QAM, 최대 320MHz BW)의 베이스밴드 ADC 요구를 매우 효율적으로 만족시키는 구현 사례를 보여준다.




[그림 1] 본 연구에서 제시하는 TDC-assisted SAR 구조와 선행연구의 비교

**#19-4 Heterogeneous Programmable Converter Array (H-PCA)**

19.4 "A Heterogeneous Programmable A/D Converter Array Covering 2–500MHz BW and 81.4–58.7dB SNDR with over 170dB FoMs"는 최근 급증하는 multi-standard/multi-band 통신 규격을 하나의 SoC에서 동시에 지원해야 하는 요구에 대한 보다 시스템적인 해법을 제시한다. 기존 접근은 각 규격(예: GSM, LTE, NR, Wi-Fi)마다 전용 ADC를 두거나, 단일 고성능 broadband ADC를 쓰는 방식이었는데, 전자는 실리콘 면적·개발비가 크고 후자는 소비전력이 과도하다는 문제가 있다. Multi-mode / reconfigurable ADC는 이 간극을 줄이는 방향으로 연구돼 왔으나, 대부분 모드 간 spec 범위가 제한적이거나, 효율이 떨어지거나, 동시에 여러 모드를 켜기 어렵다는 한계가 있었다.

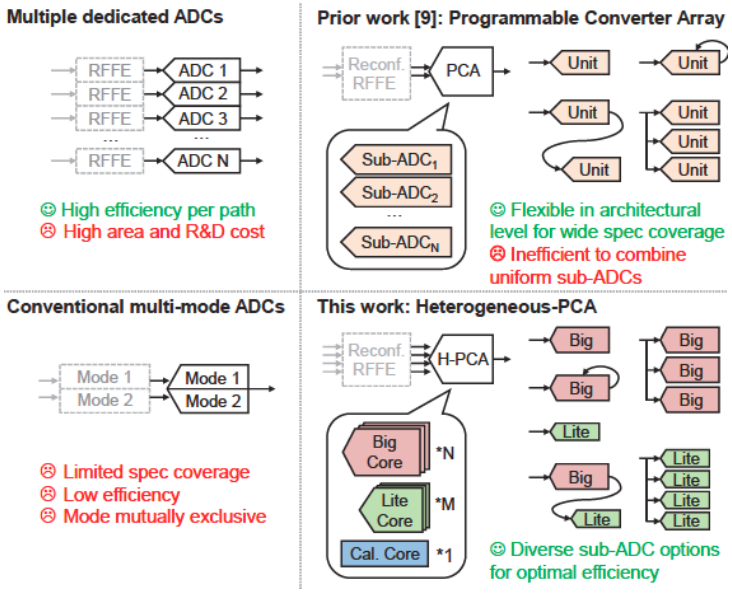
이와 대조적으로 Programmable Converter Array(PCA)는 여러 개의 sub-ADC를 array로 배치하고 필요에 따라 직렬·병렬로 재구성하여 다양한 모드를 만들 수 있는 개념인데, 종래 PCA는 (1) 모든 sub-ADC를 균일하게 설계(Uniform partition)해 아키텍처별 최적화가 어렵고, (2) residue 전달이 전압 모드에 fully-connected 입력이라 기생 RC에 매우 민감해 확장성이 떨어지는 문제가 있었다. 19.4 논문은 이를 해결하기 위해, (1) 서로 다른 특성을 가진 이기종(Heterogeneous) Core를 섞어 쓰는 H-PCA 개념과, (2) group-connected input bus + low-Zin current-mode residue bus를 도입하여 확장성을 대폭 끌어올린 것이 핵심이다.

H-PCA는 세 가지 핵심 요소로 구성된다. 첫째, Big Core는 1.5pF CDAC와 GM-C 프리앰프를 갖춘 13b SAR로, NS-SAR, 고정밀 SAR, 또는 Pipe-SAR의 1단으로 동작하며, 큰 커패시터와 전치 증폭을 통해 70dB 이상의 SNR을 요구하는 모드를 커버한다. 둘째, Lite Core는 0.25pF CDAC를 사용하는 11b SAR로, 저전력 모드 또는 Pipe-SAR 후단 스테이지에 최적화되어 약 60dB SNR 수준에서 높은 에너지 효율을 목표로 설계되었다. Big/Lite 코어 모두 시간 인터리빙이 가능해, Big Core 3개를 TI로 묶어 200MHz BW, 66dB SNDR, Lite Core 4개를 TI로 묶어 500MHz BW, 약 59dB SNDR을 달성하는 등 다양한 모드를 구성할 수 있다. 셋째, Dither Core는 공용 보정 코어로, PRNG 기반 전압/전류 dither를 Big Core와 residue bus에 주입하여 각 TIA의 이득과 전체 GM×TIA gain을 순차적으로 추출하고, Core 간 gain error를 background로 정렬하는 데 사용된다.

이 구조를 시스템적으로 가능하게 만드는 열쇠는 저입력임피던스(Current-mode) residue bus이다. 종래 PCA에서는 residue를 전압 모드로 전달하거나, 비교적 높은 Zin(예: 480Ω)의 TIA를 이용한 전류 모드를 사용했기 때문에, 버스 배선에 깔리는 기생 C와 곱해지는 시상수(예: Cp=100fF일 때 48ps)가 크고, Core 간 물리적 거리에 따라 속도·SNDR이 민감하게 변했다. 19.4에서는 gm-boosting과 Kelvin connection을 적용한 공통 게이트 기반 입력 구조로 TIA의 Zin을 약 40Ω까지 낮춰, 동일 Cp에서 시간 상수를 4ps 수준으로 줄였다. 그 결과 residue bus 상의 추가 RC가 사실상 무시 가능해져, Big/Lite Core 사이의 거리(50~420μm)가 달라져도 Pipe-SAR 모드 SNDR이 거의 일정함을 실측으로 보인다. 이는 "Core를 칩 위에 자유롭게 배치하고 여러 조합으로 연결해도 성능이 깨지지 않는" PCA의 확장성을 처음으로 설득력 있게 입증한 결과다.

프로토타입 H-PCA는 28nm CMOS에서 3개의 Big Core와 4개의 Lite Core로 구성되어, NS-SAR(2MHz BW, 81.4dB SNDR), Big-core SAR(10MHz, 70dB), Lite-core SAR(133MHz, 60dB), Pipe-SAR(100MHz, 71dB), Noise-shaped Pipe-SAR(50MHz, 76dB), 3×TI Big-SAR(200MHz, 66dB), 4×TI Lite-SAR(500MHz, 58.7dB) 등 총 7개 대표 모드를 하나의 IP로 구현한다. 모

든 모드에서 Schreier FoM 170dB 이상, 최대 176.5dB(100MHz BW)까지 달성하며, 동일한 규격 범위를 개별 single-mode ADC들로 커버하는 경우 대비 활성 면적을 약 56% 절감하는 것으로 보고된다. 또한 IQ 채널 4개를 동시에 처리하는 Wi-Fi/NR 시나리오에서 4 concurrent mode 동작을 시연해, "실제 단말 SoC에서 하나의 Reconfigurable ADC IP로 전 대역·전 표준을 커버한다"는 PCA 비전이 실현 가능한 수준에 도달했음을 보여준다.




[그림 2] 본 연구에서 제시하는 H-PCA 구조와 선행연구의 비교

## 저자정보


**이동윤 석박사통합과정 대학원생**

- 소속 : KAIST
- 연구분야 : Body-Channel-Communication Transceiver Design for Body-Area Network in Biomedical Application
- 이메일 : dongyoon.lee@kaist.ac.kr
- 홈페이지 : https://impact.kaist.ac.kr

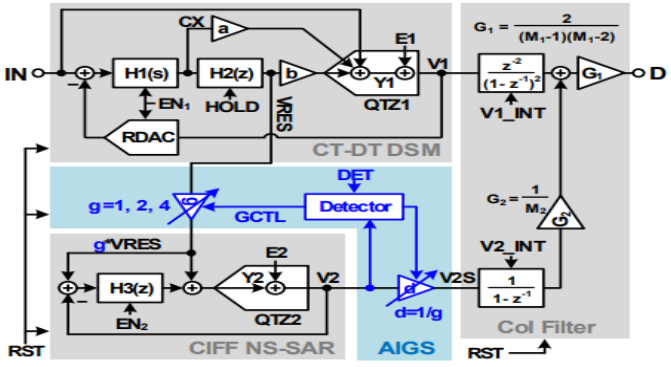
# A-SSCC 2025 Review

단국대학교 파운드리공학과 석사과정 장준서

## Session 9: High Precision ADCs

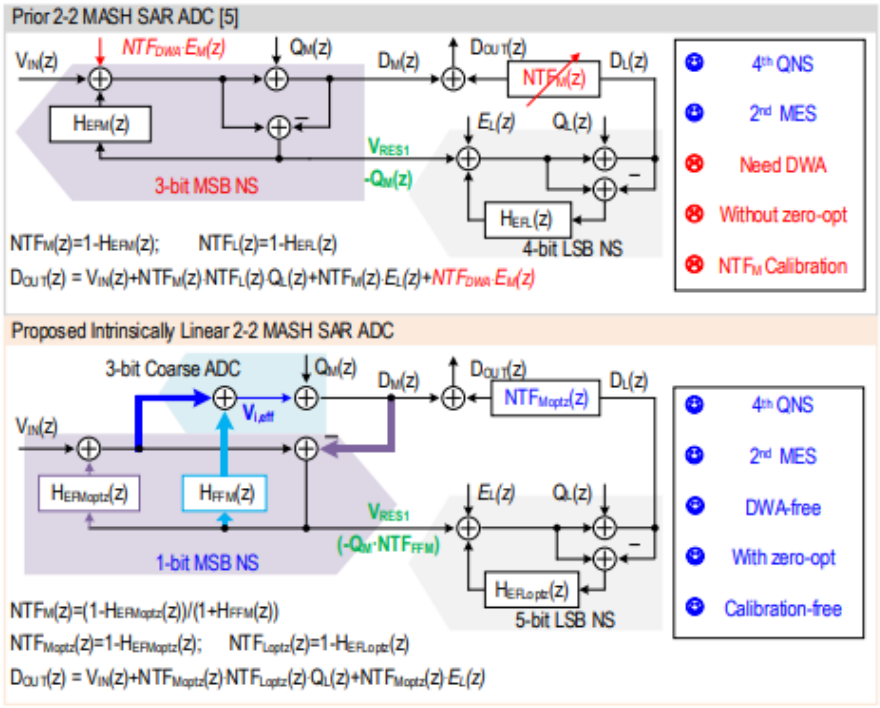
Session 9에서는 총 4편의 High Precision ADC 논문이 발표되었다. 본 세션의 논문들은 공통적으로 저주파 대역에서의 높은 정확도와 선형성, 그리고 공급 전압 및 공정 변화에 대한 안정성을 어떻게 확보할 것인가에 초점을 맞추고 있다. 특히 Incremental ΔΣ ADC, SAR ADC, 그리고 이들의 하이브리드 구조를 활용하여 정확도와 대역폭 간의 트레이드오프를 완화하려는 접근이 두드러졌다. 특히 기존 구조의 한계를 구조적으로 극복하려는 시도가 두드러졌으며, 본 리뷰에서는 이러한 흐름을 잘 대표한다고 판단되는 두 편의 논문을 선정하였다.

**#9-1** KAIST에서 발표한 논문으로, Incremental ΔΣ ADC와 MASH 구조를 결합한 Two-Step 하이브리드 ADC를 제안한다. Incremental ΔΣ ADC의 높은 DC 정확도를 유지하면서도 대역폭 확장 시 성능 저하가 발생하는 문제를 해결하기 위해, 스테이지 간 이득을 자동으로 조절하는 Automatic Inter-Stage Gain Selection(AIGS) 기법을 도입하였다. 측정 결과, 제안된 ADC는 160 kHz 대역폭에서 약 97 dB 수준의 SNDR과 100 dB 이상의 DR을 달성하였으며, AIGS 적용 시 고정 이득 설정 대비 성능이 유의미하게 향상됨을 확인하였다. 또한 변환 단계 일부를 비활성화하는 방식으로 전력 소모를 줄이면서도, 공급 전압 변화에 따른 성능 변동이 작아 안정적인 동작 특성을 보였다. 이러한 결과는 Incremental ΔΣ ADC가 하이브리드 구조와 적응형 이득 제어를 통해 정밀 계측 응용에서 정확도와 대역폭을 동시에 확장할 수 있음을 보여준다.




[그림 1] AIGS를 적용한 Incremental Two-Step 하이브리드 DSM+CIFF NS-SAR ADC 구조

**#9-3** Xidian University에서 발표한 논문으로, 구조적 기법을 통해 높은 선형성을 확보한 MASH SAR ADC를 제안한다. 기존 MASH SAR ADC는 높은 해상도를 제공할 수 있으나, 아날로그 증폭기 비선형성과 잔여 전하 누적에 의해 SFDR 확보에 한계가 존재하며, 이를 보완하기 위해 DWA나 디지털 캘리브레이션에 의존하는 경우가 많았다. 본 논문에서는 MSA(Multi-Step Amplification) Enhancement와 Incremental Correlated Level Shifting 기법을 적용하여, 별도의 DWA나 디지털 캘리브레이션 없이도 높은 선형성을 구조적으로 확보하였다. 측정 결과, 제안된 ADC는 100 kHz 대역폭에서 105 dB 이상의 SFDR을 달성하였으며, 입력 주파수 전반에 걸쳐 약 89 dB 수준의 SNDR을 유지하였다.
특히 고주파 입력 조건에서도 SFDR 열화가 제한적으로 나타나, 구조적 선형성 확보의 효과가 실측으로 확인되었다. 추가적으로, 제안된 구조는 MASH SAR ADC에서 문제가 되기 쉬운 잔여 전하(residue) 누적과 증폭 단계 비선형성을 MSA 기반 다단 증폭과 correlated level shifting을 통해 효과적으로 억제하였다. 이로 인해 입력 주파수 변화에 따른 선형성 저하가 비교적 완만하게 나타났으며, 고해상도 MASH SAR ADC에서 흔히 요구되는 복잡한 보정 회로를 배제하고도 정밀 센서 인터페이스에 적합한 성능을 확보할 수 있음을 보여준다.

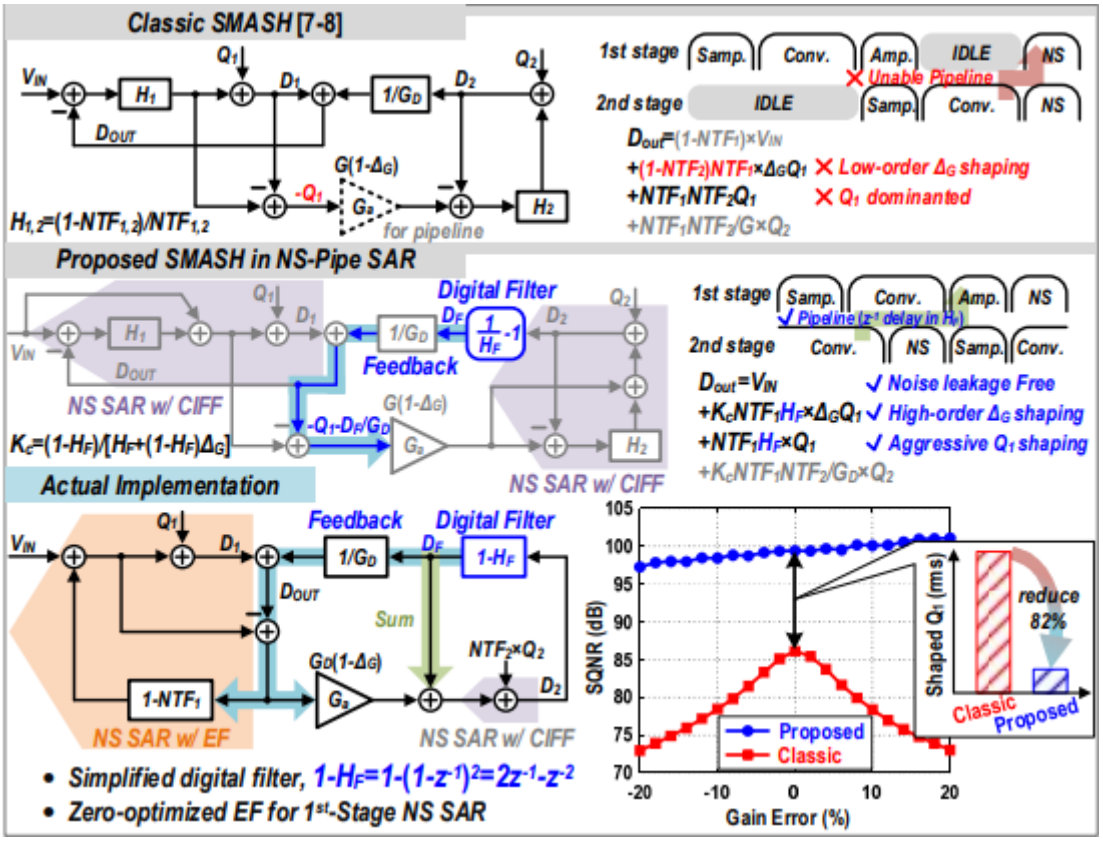



[그림 2] 기존 MASH SAR ADC와 제안된 구조적 선형 MASH SAR ADC의 비교

## Session 23: High Precision ADCs

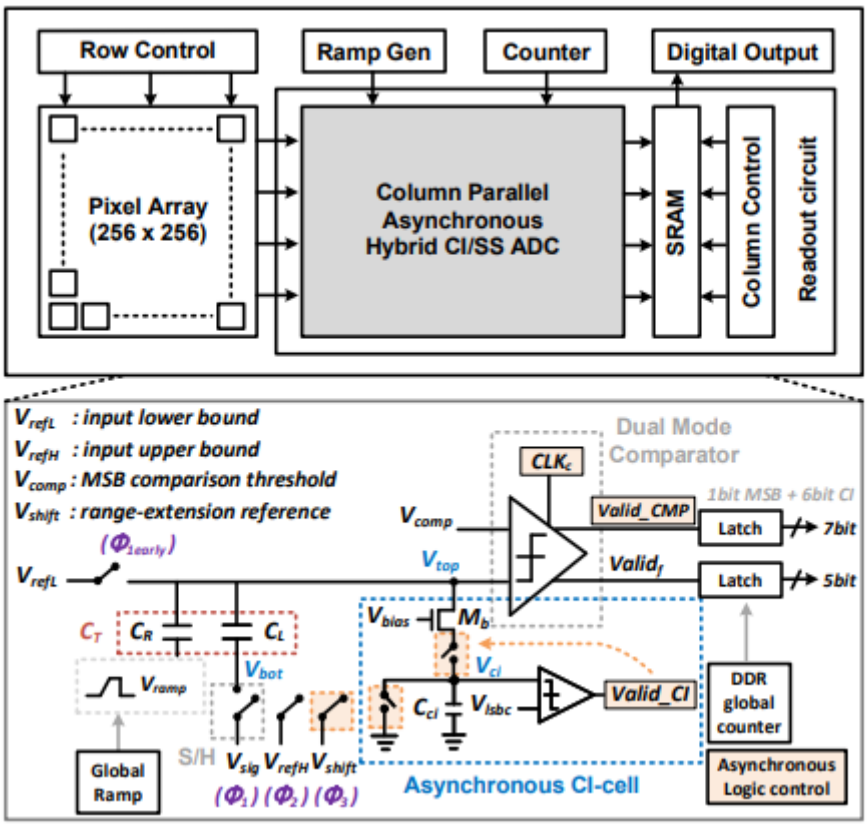
Session 23에서는 총 5편의 High Performance ADC **논문**이 발표되었으며, 고속 동작 환경에서 대역폭, 변환 속도, 전력 효율, 그리고 공정·온도·전압 변화에 대한 강건성을 동시에 확보하려는 연구들이 주를 이루었다. Pipeline SAR, Noise-Shaping SAR, 그리고 CIS용 Column-Parallel ADC 등 응용 지향적인 구조가 다수 제시되었다. 본 리뷰에서는 이 중에서도 고속 ADC에 고차 Noise-Shaping을 안정적으로 적용한 구조, 실제 시스템 레벨에서 고속, 고선형 동작을 검증한 사례라는 관점에서, 두 편의 논문을 선정하였다.

**#23-1** University of Macau에서 발표한 논문으로, Pipeline SAR ADC에 4차 Noise-Shaping을 적용한 고속 ADC 구조를 제안한다. 고속 동작 환경에서 고차 Noise-Shaping 적용 시 발생할 수 있는 루프 안정성 및 스테이지 간 이득 오차 문제를 구조적으로 해결하는 것이 본 논문의 핵심이다. 측정 결과, 제안된 ADC는 15 MHz 대역폭에서 약 80 dB 수준의 SNDR을 달성하였으며, 출력 스펙트럼을 통해 명확한 4차 Noise-Shaping 특성이 확인되었다. 또한 온도 및 공급 전압 변화에 따른 성능 변동이 제한되어, 고속 ADC에서 요구되는 PVT 강건성을 확보하였다. 이는 Pipeline SAR 구조에서도 고차 Noise-Shaping이 실용적으로 적용될 수 있음을 보여준다.



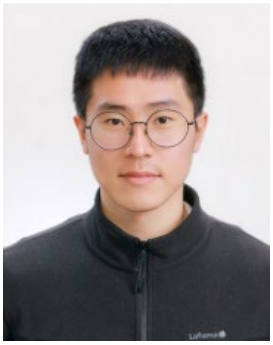

[그림 3] 기존 SMASH 구조와 제안된 NS-Pipeline SAR용 SMASH 구조의 신호 흐름 및 실제 구현

**#23-5** National Tsing Hua University에서 발표한 논문으로, Column-Parallel Hybrid CI/SS ADC를 집적한 256 × 256 CMOS Image Sensor(CIS)를 제안한다. Column-parallel ADC에서 요구되는 고속 변환 성능과 컬럼 간 균일성을 동시에 확보하는 것이 본 논문의 주요 목표이다. 제안된 구조는 비동기(Asynchronous) Charge-Injection(CI) 아키텍처와 Hybrid Quantization(CI/Single-Slope) 방식을 결합하여, 고속 동작 시에도 비교기 및 디지털 제어에 따른 지연과 비선형성을 최소화하였다. 또한 Dark Frame Calibration을 통해 컬럼 고정 패턴 잡음(FPN)을 효과적으로 제거하였다. 측정 결과, 제안된 ADC 어레이는 435kS/s 변환 속도에서 약 9.6-bit 수준의 ENOB를 달성하였으며, 컬럼 간 비균일성은 입력 범위 전반에서 1.2 LSB 이내로 제한되었다. 실제 이미지 캡처 결과에서도 컬럼 고정 패턴 잡음이 효과적으로 제거됨을 확인하였으며, 이는 대규모 병렬 ADC 어레이에서도 안정적인 동작이 가능함을 보여준다. 이러한 결과는 제안된 비동기 CI 기반 Hybrid 구조가 고속·고해상도 CIS 응용에 적합한 설계 방향임을 실측으로 입증한다.




[그림 4] CIS 프로토타입 구조와 제안된 컬럼 병렬 하이브리드 CI/SS ADC

## 저자정보




**장준서 석사과정 대학원생**

- 소속 : 단국대학교
- 연구분야 : Analog front end 설계
- 이메일 : cah7781@naver.com
- 홈페이지 : https://sites.google.com/dankook.ac.kr/acs-lab/home

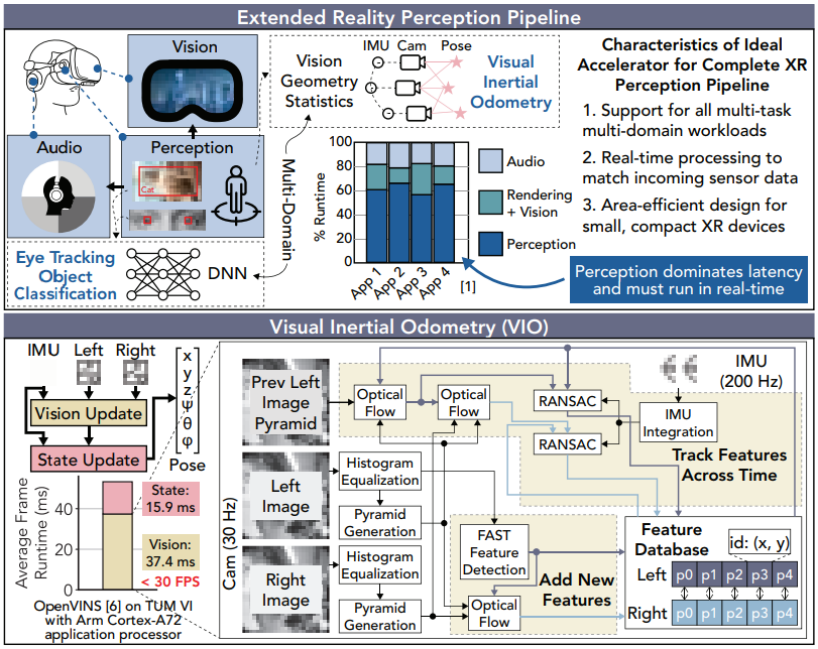
# A-SSCC 2025 Review

경북대학교 전자전기공학부 박사과정 박승현

## Session 3 Domain Specific Accelerators

이번 A-SSCC 2025 세션 3의 논문들은 서로 다른 응용 분야를 다루고 있음에도, 공통적으로 특정 도메인의 연산 패턴을 정확히 파악해 하드웨어 구조를 그에 맞게 최적화한 설계 접근이 돋보인다. 전반적으로 기존 기법을 실시간, 저전력 조건에 맞게 실제 칩으로 구현하는 데 초점을 맞춘 실용적 연구들이라는 인상을 준다.

### #3-2 Birch: A Real-Time Accelerator for Multi-Task Mixed-Domain Extended Reality Perception Workloads




[그림 1] 실시간성이 중요한 XR에서 VIO와 DNN의 가속

이 논문은 extended reality (XR)에서 지각(perception) 파이프라인을 온전히 하드웨어로 끌어안으려는 시도를 보여주는 흥미로운 연구이다. Birch는 XR 시스템에서 공통적으로 발생하는 문제, 즉, visual inertial odometry (VIO)와 DNN 기반 인식 작업이 서로 다른 성격의 연산을 요구하면서도 동시에 실시간으로 처리되어야 한다는 점에 대해 하나의 SoC에서 균형 있게 대응하는 설계 결과이다. 기존 모바일 프로세서에서는 VIO의 vision update 단계가 상당한 연산량을 차지해 30 FPS 카메라 입력조차 안정적으로 처리하기 어려웠는데, Birch는 이 부분을 하드웨어 가속으로 분리함으로써 실시간성을 확보했다.

Vision accelerator의 구조적 특징은 비교적 명확한 편이다. FAST 기반 특징점 검출, 피라미드 생성, optical flow는 모두 반복적이고 데이터 접근 패턴이 뚜렷하여 하드웨어 파이프라인에 적합한 연산인데, Birch는 라인 버퍼, 정렬 버퍼, 더블 버퍼와 같은 고전적인 영상처리 기법을 각 단계에 적용해 CPU 대비 큰 지연 감소를 얻었다. 특히 Optical flow에서 더블 버퍼를 사용해 특징점 윈도우 계산과 위치 업데이트를 겹치는 방식은 구현 난도가 높지 않으면서도 latency 절감 효과가 꽤 크게 나타났다. 이러한 최적화 덕분에 vision update가 약 7ms 수준으로 줄었는데, 이는 단순히 '빠르다'는 의미보다는 카메라 프레임레이트를 안정적으로 수용하는 데 필요한 최소 조건을 충족했다는 점에서 현실적인 의미를 갖는다.

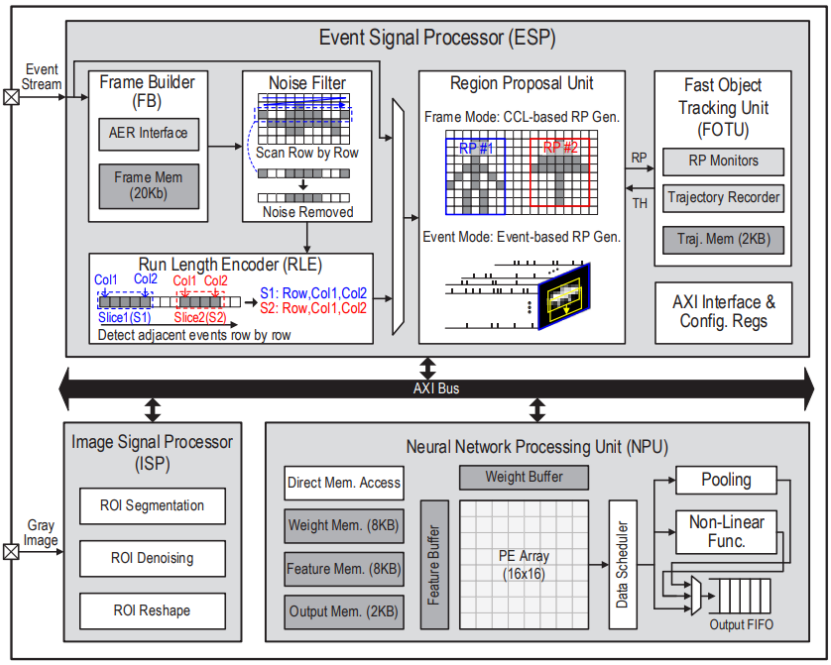
Birch가 DNN 연산 쪽에서도 실시간 성능을 확보한 점은 특이한 결과라기보다는, FP8 systolic array와 BF16 벡터 유닛을 함께 사용한 구조 덕분으로 보인다. XR에서 사용하는 DNN 모델의 규모는 대규모 모델에 비해 상대적으로 작고 고정된 패턴을 갖기 때문에, 32×32 systolic array 정도 규모의 매트릭스 유닛이라도 충분히 성능이 나온다. Eye-Gaze 모델이 1ms 미만으로 처리되는 결과나 ResNet-18이 10ms 이내에 처리되는 점은 이런 구조적 특징을 반영한 결과라고 할 수 있다. 다만 이는 Birch가 특별히 높은 연산 자원을 갖췄다기보다, XR perception에 필요한 모델들이 그리 크지 않기 때문에 현실적으로 가능한 수치로 이해할 수 있다.

흥미로운 부분은 BF16 벡터 유닛을 VIO의 RANSAC 계산에도 그대로 활용한다는 점이다. 이로 인해 별도의 전용 행렬 연산 블록 없이 CPU 대비 3배가량 빠르게 RANSAC을 수행할 수 있는데, 이는 면적 효율 측면에서 의미가 있다. XR과 같은 소형 디바이스 환경에서는 유닛을 공유하는 구조가 유리하며, Birch는 이런 선택을 적절히 적용한 사례로 보인다.

정확도 측면에서도 크게 무리 없는 수준이다. TUM VI와 EuRoC에서의 위치 추정 오차는 기존 VIO 연구 수준에 준하는데, 이는 Birch가 Vision Update만 가속하고 State Update는 CPU 기반 알고리즘을 그대로 유지하는 구조적 이유도 있다. 즉, Birch가 새로운 VIO 알고리즘을 제시한 것이 아니기 때문에 기존 소프트웨어 VIO의 정확도를 따라가는 결과가 나온 것으로 해석할 수 있다.

전체적으로 Birch는 XR perception 워크로드에서 가장 부담이 되는 Vision Update를 하드웨어로 옮기고, DNN 연산을 별도의 유닛으로 처리해 각 도메인에서 요구하는 성능을 실시간 수준으로 맞춘 일종의 균형 잡힌 설계로 평가할 수 있다. XR 기기에서 필요로 하는 조건을 무리 없이 충족한 실용적 접근을 보여주고 있다.

**#3-3 A 96pJ/Frame/Pixel and 61pJ/Event Anti-UAV System with Hybrid Object Tracking Modes**




[그림 2] 제안하는 anti-UAV 시스템의 하드웨어 아키텍처

이 논문은 이벤트 카메라 기반의 UAV 탐지 및 추적 시스템에서 흔히 나타나는 두 가지 문제—고속 이동 물체에서의 성능 저하와, 이벤트 기반 ODT (object detection & tracking)의 높은 전력—를 해결하기 위해, 프레임 기반과 이벤트 기반 방식을 상황에 따라 전환하는 하이브리드 아키텍처를 제안하고 있다. UAV처럼 작고 빠르게 움직이는 물체는 전통적인 프레임 기반 검출에서는 모션 블러가 문제이고, 반대로 이벤트 기반 방식은 항상 활성화되어 있어 전력 소모가 크다는 특성이 있기 때문에, 두 방식을 적절히 섞는 접근을 취한다.

전체 시스템은 ESP(Event Signal Processor), ISP(Image Signal Processor), NPU로 구성되며, 특히 ESP가 항상 켜진 상태에서 저전력으로 이벤트 스트림을 처리하고, 필요한 경우에만 NPU가 동작하도록 설계한 점이 눈에 띈다. 이 구조 덕분에 고정 전력 소모를 줄이면서도 빠른 Object Tracking이 가능해진다. ESP 내부에서 Run-Length Encoding을 사용해 이벤트 프레임을 slice 단위로 압축하고, Region Proposal Unit(RPU)이 이 slice들을 기반으로 객체를 찾고 추적하는데, 이 단계의 하드웨어가 논문의 핵심이라고 할 수 있다.

프레임 기반과 이벤트 기반을 전환하는 방식은 비교적 단순하지만 효과적이다. RPU는 먼저 프레임 모드에서 CCL 기반으로 객체 후보를 만든 뒤, 특정 크기 기준을 넘는 물체가 나타나면 이벤트 모드로 전환하여 주변 이벤트의 공간적 분포를 이용해 추적을 수행한다.

이 접근은 고속 이동 시 프레임 기반 검출의 불안정성을 보완하는 정도의 역할이며, 구현 복잡도 대비 성능 향상이 명확하다는 점에서 실용성이 있다. 실제로 이벤트 기반 RP 업데이트는 지정된 event threshold(TH)에 도달했을 때만 발생하게 하여 불필요한 업데이트를 줄이고, 객체 속도와 크기에 따라 TH를 조정해 RP가 과도하게 흔들리는 현상을 막는 식의 소규모 최적화가 포함되어 있다. 이런 방식은 특별히 새로운 알고리즘이라고 보긴 어렵지만, 하드웨어 구조에서는 오히려 이런 단순한 규칙 기반 방식이 안정적으로 동작할 수 있다는 장점이 있다.

고속 UAV에서 흔히 발생하는 문제는 RP가 실제 물체의 위치와 어긋나는 현상인데, 논문에서는 이를 FOTU(Fast Object Tracking Unit)로 대응한다. FOTU는 각 RP의 변화량을 모니터링하면서 TH 값을 유동적으로 조절하는데, 결과적으로 RP misalignment가 줄어들고 추적 품질이 안정된다는 실험 결과가 제시된다. 이런 방식은 기본적으로 heuristic 기반 조정이기 때문에 정교한 최적화라고 할 수는 없지만, 이벤트 기반 센서 데이터가 워낙 노이즈가 심하고 비균일한 특성을 갖기 때문에, 정교한 ML 모델보다 이러한 단순 조정이 더 안정적으로 동작할 수 있다는 점에서 타당한 접근으로 보인다.

NPU 부분은 전체 시스템에서 비중이 크지는 않지만, 전체 전력의 절반 이상이 이 유닛에서 발생한다는 점이 특징이다. 특히 NPU는 각 객체에 대해 한 번만 동작하도록 gating을 적용하여, 객체 수가 많지 않은 상황에서 상당한 전력 절감 효과를 얻는다. 전체적으로는 NPU를 매우 적극적으로 최적화하기보다는, 필요할 때만 켜는 구조를 통해 결과적으로 시스템 전력 효율을 높인 셈이다.

실험 결과로, 프레임 기반 모드에서 96pJ/(frame·pixel), 이벤트 모드에서 61pJ/event라는 수치는 저전력 Always-on 감지 시스템으로서는 합리적이며, UAV 탐지 정확도도 80% 후반대 수준으로 보고된다. UAV 속도가 높아질수록 이벤트 기반 trajectory 분류가 프레임 기반 패치 분류보다 안정적인 성능을 보인다는 결과도, 하이브리드 구조의 타당성을 지지하는 근거로 보인다.

종합적으로 보면 이 논문은 이벤트 기반 센서의 장단점을 하드웨어 레벨에서 자연스럽게 보완하는 구조적 설계 방법을 제시하였다. 프레임 기반과 이벤트 기반 중 어떤 방식이 더 우월한가를 논하기보다는, 두 방식을 상황에 따라 적절히 선택하는 것이 실제 시스템에서 더 안정적이고 전력 효율적인 접근이라는 점을 실험적으로 보여준 사례로 볼 수 있다. 이런 관점에서, UAV나 소형 고속 물체 감지 분야에서 하드웨어-알고리즘 공동 설계의 방향성을 확인할 수 있는 연구라고 정리할 수 있다.

## 저자정보

**박승현 박사과정 대학원생**

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

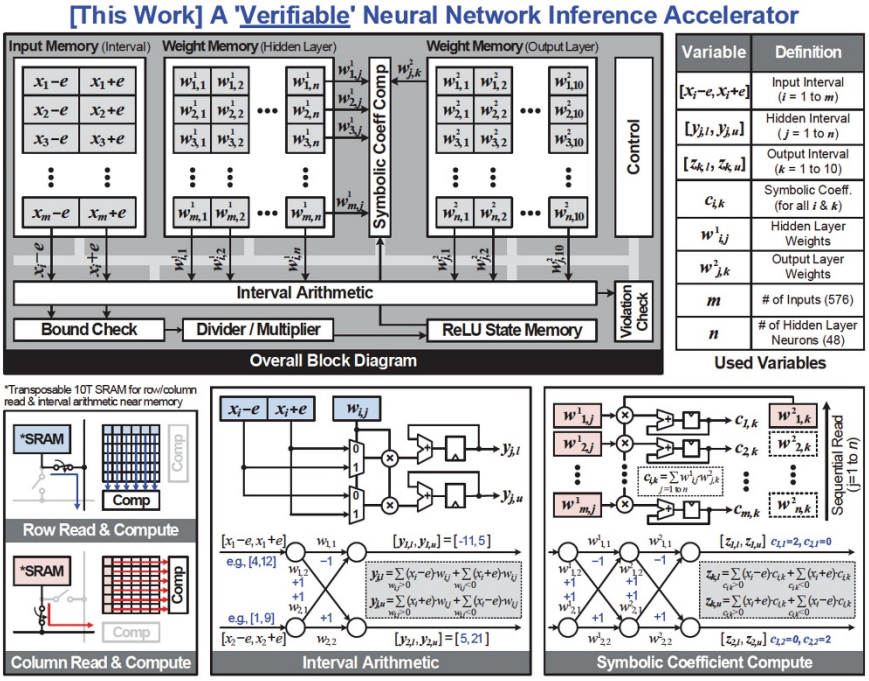
# A-SSCC 2025 Review

연세대학교 전기전자공학과 석박통합과정 김동욱

## Session 15 Techniques for Emerging Hardware Accelerators

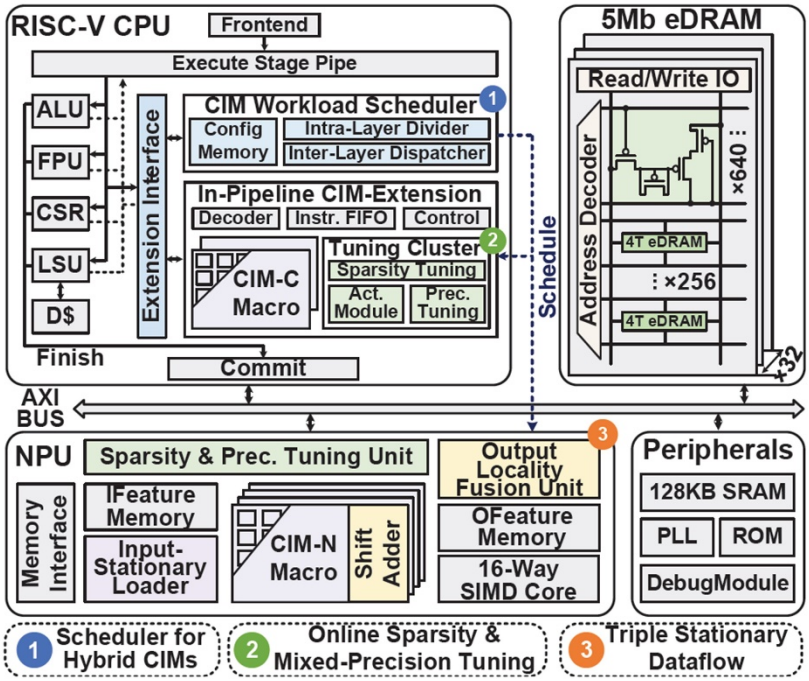
2025 A-SSCC의 Session 15는 Techniques for Emerging Hardware Accelerators라는 주제로 총 5편의 논문이 발표되었다. 이 세션은 edge AI 시스템의 궁극적인 성능 및 전력 효율 목표를 달성하기 위한 차세대 하드웨어 기술과 아키텍처에 중점을 두고 있다.

**#15-1** KAIST와 University of California, Santa Barbara의 연구진이 발표한 본 논문 [Verifica: A 65nm In-Memory Symbolic Interval Computing Accelerator for a Formal Neural Network Verification]은 edge 디바이스를 위한, ASIC 기반의 formal neural network verification 하드웨어 가속기를 제안했다. 딥러닝 모델은 뛰어난 성능을 보이지만 노이즈에 의해 오류를 일으킬 수 있는 취약성을 가지는데, 이로 인해 edge 디바이스에서는 추론 가속기를 완전히 신뢰할 수 없다는 문제가 있었다. 본 연구에서는 이를 위해 딥러닝 모델의 신뢰성을 실시간으로 검증하는 하드웨어 측면의 해결 방안을 제시했다. 기존 소프트웨어 기반 검증 방식들은 많은 연산을 필요로 하고 독립적인 입력변수 판단으로 인해 느리고 부정확하다는 한계가 있다. 이를 보완하기 위해 symbolic interval analysis를 적용하지만, 본 연구는 소프트웨어 측면의 적용을 넘어 edge 디바이스 환경에 보다 적합하도록 이를 하드웨어로 최적화했다. 함께 적용된, 비선형함수인 ReLU의 선형 완화도 검증력 극대화에 기여했다. 또한 row와 column 연산이 모두 가능한 transposable 10T SRAM cell을 사용해서, 검증 연산에 대해 near-memory computation이 가능한 아키텍처로 구성했다. 본 연구의 이러한 알고리즘-하드웨어 co-optimization 결과, 65nm 테스트 칩에서 97.8% 이상의 검증 가능성과 1.22ms 미만의 검증 시간을 확인했다. 추가적으로, 에너지 소모 또한 검증 당 2.14μJ의 낮은 소비를 달성하여 edge 디바이스에 적용가능한 방법임을 입증했다.



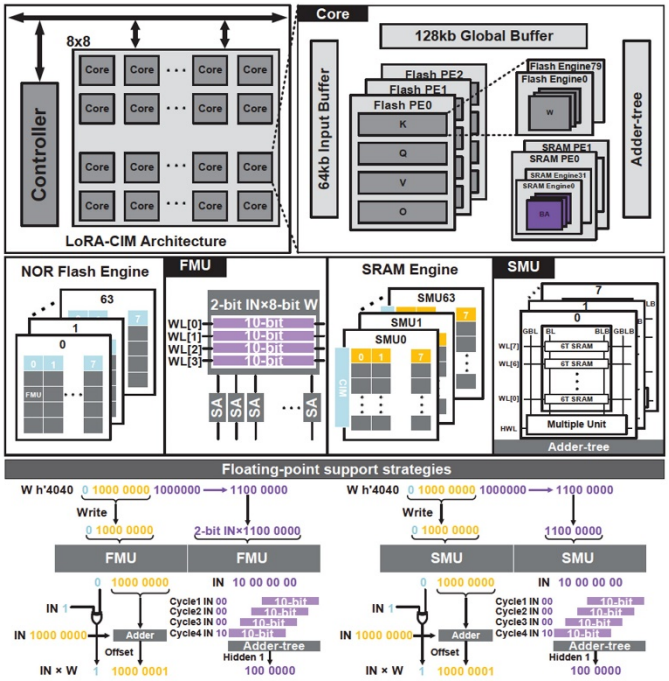

[그림 1] Verifica의 아키텍처와 회로 구성

**#15-3** Peking University, Beijing과 Southeast University, Nanjing의 연구진이 발표한 본 논문 [A 30.28-151.42TFLOPS/W@FP8 Training Accelerator with Hybrid Compute-in-Memory and Workload Scheduling for Edge AI Applications]는 edge 디바이스에서의 모델 훈련에 최적화된 가속기를 제안했다. Edge 디바이스에서의 훈련에 대한 기존 연구들은 불균형한 연산 및 낮은 하드웨어 자원 활용률, 제한된 on-chip 메모리, 그리고 높은 메모리 접근 비용으로 인한 한계를 가지고 있다. 본 연구에서는 이러한 문제를 해결하기 위해 hybrid compute-in-memory (CIM) 기반의 아키텍처와 워크로드 스케줄링 기법을 적용한 훈련용 가속기를 제시했다. 두 개의 CIM macro인 CIM-C와 CIM-N을 각각 RISC-V CPU와 NPU 내부에 두어 hybrid CIM 아키텍처를 구성했고, CIM 워크로드 스케줄러는 RISC-V CPU 내에 위치하여 명령을 hybrid CIM에게 할당할지 여부를 판단하게 했다. 스케줄러는 워크로드 스위치 및 퓨전을 통해 유휴 상태의 CIM 코어에 작업을 할당함으로써 하드웨어 활용률을 높일 수 있게 했다. 또한, 훈련의 정확성과 효율성을 위해 동적인 sparsity 및 precision 튜닝을 구현했고, 기존 weight stationary 방식의 높은 입력 메모리 접근 비용 문제를 해결하기 위해 CIM-N은 input stationary와 output stationary도 지원하는 triple-stationary dataflow를 지원하도록 구현했다. 22nm 공정으로 제작한 칩 실험 결과, 기존 CIM 훈련 가속기 대비 2.9배 높은 성능을 기록했고 하이브리드 CIM 스케줄링을 통해 CIM 활용률을 평균 2.92배 개선했으며, 면적과 에너지 효율을 모두 고려한 FoM 지표에서 기존 CIM 가속기 대비 최대 28.08배 우수한 결과를 보였다.




[그림 2] 제안하는 hybrid CIM 기반 훈련용 가속기의 아키텍처

**#15-4** Fudan University, China Flash Co.,Ltd, 그리고 Zhangjiang Laboratory의 연구진이 발표한 본 논문 [LoRA-CIM: A Fully-Digital Hybrid Flash-SRAM CIM Accelerator with LoRA Fine-Tuned LLM in Edge Devices]는 LLM을 edge 디바이스에서 효율적으로 fine-tuning 할 수 있도록 하는 hybrid CIM 가속기를 제안했다. LLM을 edge 디바이스에 배포할 때, 특정 작업에 맞게 fine-tuning하는 것은 필수적이지만 기존 CIM 기반 edge 디바이스는 작은 메모리 용량과 아날로그 CIM의 정확도 손실, 그리고 낮은 하드웨어 활용률 측면에서 문제가 있었다. 본 연구에서는 이러한 문제를 해결하기 위해 LoRA(Low-Rank Adaptation) 미세 조정 기법을 활용한 hybrid Flash-SRAM CIM 아키텍처를 제시했다. Flash CIM 엔진을 통해 용량 문제를 해결했고, 쓰기 작업이 없기 때문에 내구성을 확보할 수 있게 했다. 또한, 아날로그 CIM의 정확도 문제를 해결하기 위해 fully-digital 연산 아키텍처를 적용했다. 하드웨어 활용률 증가를 위해서는 연산 특성을 고려한 데이터플로우 최적화를 적용했는데, 트랜스포머의 레이어에 따라 CIM 코어를 파이프라인 모드로 운용하거나 병렬 모드로 운용하게 했다. 제작된 칩으로 실험을 진행한 결과, INT8 기준 1.45 TOPS/W, BF16 기준 700 GFLOPS/W의 효율을 달성했고, MobileLLM-125M 디코딩 속도를 SRAM 베이스라인 대비 7.7배 개선하였으며 에너지 감소 최대 43%를 달성했다. 본 연구를 통해 LoRA-CIM은 edge LLM 배포의 실현 가능성을 보였다.




[그림 3] Floating-point 연산 지원 방식과 함께 나타낸 LoRA-CIM 의 아키텍처

## 저자정보

**김동욱 석박통합과정 대학원생**

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

# A-SSCC 2025 Review

KAIST 인공지능반도체대학원 박사과정 하상우

## Session 21 AI Accelerators

이번 IEEE ASSCC 2025의 Session 21은 AI 프로세서를 주제로 총 5편의 논문이 발표되었다. 올해 ASSCC Session 21에서 발표된 논문들은 총 두가지의 주요 트렌드를 보였다. 첫째, 거대 언어 모델 (LLM) 가속이 핵심 주제로 부상하였다 (논문 21.1, 21.4). 특히 단순히 엣지 디바이스에서의 single batch inference를 넘어, Adelia [1]처럼 여러 유저가 사용하는 multi-batch inference까지 고려한 논문이 등장하였다 (논문 21.1). 둘째, 확산 모델 (diffusion model)을 포함한 Embodied AI 가속의 중요성이 강조되었다 (논문 21.5). 마지막으로 엣지 디바이스를 위한 BNN 및 SNN 기반 초저전력 가속기들이 제안되었다 (논문 21.2, 21.3).

**#21-1**은 칭화대학교에서 발표한 28nm 멀티 칩렛 기반 LLM 가속기로, 멀티 유저 환경에서 LLM 서빙을 효율적으로 수행하기 위해 Prefill과 Decode 단계를 물리적으로 분리하는 분산 컴퓨팅(Disaggregated Computing) 방식을 채택한 것이 가장 큰 특징이다. Prefill 단계는 입력 토큰 전체를 한 번에 처리하는 행렬-행렬 곱셈으로 연산 집약적인 특성을 가지며, decode 단계는 토큰을 하나씩 순차적으로 생성하는 벡터-행렬 곱셈으로 메모리 대역폭 집약적인 특성을 가진다.

본 논문은 다중 배치 환경을 효율적으로 서비스하기 위해 칩렛 구조를 사용한 disaggregated computing이 필요하다고 주장한다. Disaggregated computing이란 prefill과 decode를 물리적으로 분리된 클러스터에서 독립적으로 처리하는 방식으로, 최근 서버단에서 LLM 서빙을 할 때 많이 사용되는 방식이다. 본 논문은 4개의 동일한 칩렛을 연결하고, 2개는 Prefill 클러스터 (P-C), 2개는 Decode 클러스터 (D-C)로 구성하여 각 단계를 독립적으로 최적화할 수 있게 하였다. 그림 1에서 볼 수 있듯이, disaggregated computing을 엣지 칩 클러스터에 적용하면 세 가지 문제가 발생한다. 본 논문은 이를 prefill 측면, decode 측면, 클러스터 간 통신 측면 총 세 가지로 나누어 모든 부분을 해결하고자 한다.

첫 번째 문제는 prefill 단계에서의 연산 병목이다. Prefill 단계에서는 데이터 재사용률 (OP/B)이 높기 때문에 가중치를 온칩 SRAM에 올려놓고 여러 토큰이 재사용하게 된다. 이로 인해 연산기는 최대 활용률로 계속 동작하게 되고, 이러는 동안 온칩 SRAM 대역폭

은 유휴 상태로 남게 된다. 이를 해결하고자 해당 논문에선 CFHM (Copy-less Float Hybrid-LUT Multiply)을 제안하였다. CFHM은 이 유휴 SRAM을 BF16 가수부 곱셈을 위한 참조 테이블 (LUT)로 활용한다. 핵심 아이디어는 하나의 SRAM을 16개의 뱅크로 분할하고, 64개의 입력 가수부를 하위 4비트 기준으로 16개 그룹으로 분류하여 각 뱅크에서 병렬로 테이블 참조를 수행하는 것이다. 이를 통해 prefill 처리량을 31.8% 향상시켰다.

두 번째 문제는 decode 단계의 메모리 병목이다. Decode 단계에서는 데이터 재사용률 (OP/B)이 낮기 때문에 외부 메모리에서 칩으로 계속 데이터를 가져와야 하고, 이 때문에 DRAM 대역폭이 병목이 된다. 반면 P-C의 DRAM 대역폭은 prefill 특성상 상대적으로 여유가 있다. 이 문제를 해결하기 위해 해당 논문은 HCWG (Hybrid Compressed Weight Gather)을 제안했다. HCWG는 prefill을 진행하는 클러스터에서 DRAM 대역폭이 남기 때문에, P-C에서도 가중치 데이터를 불러온 다음 이를 압축해서 D-C로 넘겨주는 방식을 사용한다. 압축 방식은 두 가지로 구성된다: (1) 허프만 지수부 압축 (HEC): 지수부가 가우시안 분포를 따르는 특성을 활용하여 상위 16개 빈도 값에 대해 단순화된 허프만 부호화 적용. (2) 활성화 기반 가수부 압축 (AMC): 입력 벡터의 부호 및 가수부 정보를 기반으로 내적 결과에 기여도가 낮은 가수부를 생략. 이를 통해 41.3%의 압축률을 달성하고 decode 단계에서의 메모리 접근으로 인한 지연 시간을 줄였다.

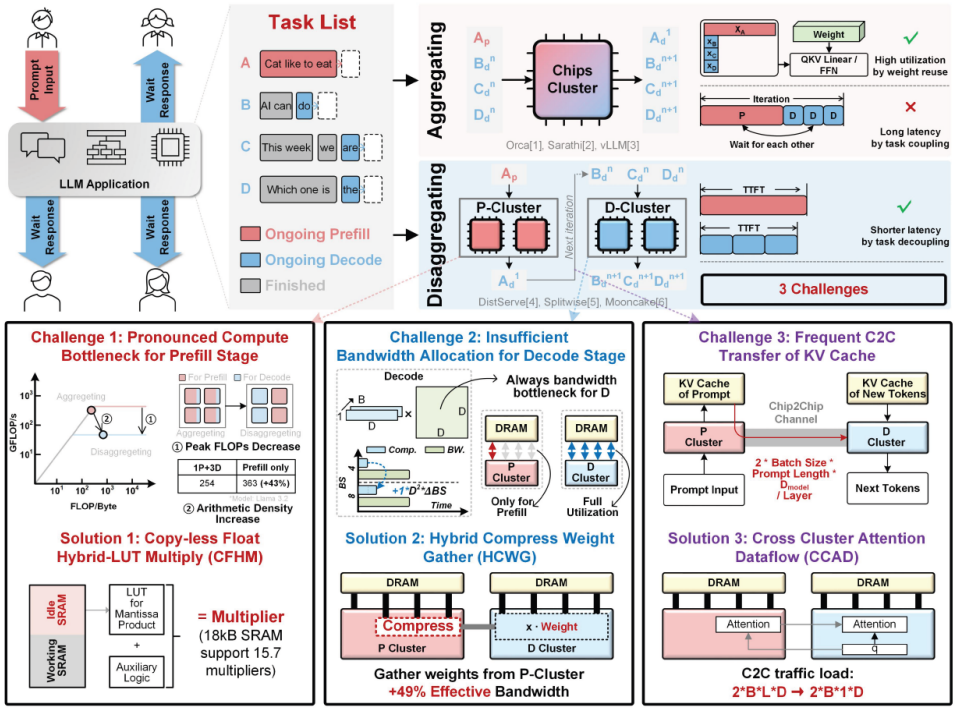
세 번째 문제는 KV 캐시의 칩 간 이동으로 인한 오버헤드이다. Disaggregated computing을 하면 P-C에서 생성된 KV 캐시를 D-C로 보내줘야 하기 때문에 에너지 및 지연 시간 오버헤드가 발생한다. 이를 해결하기 위해 본 논문은 먼저 UDB (Unified Dynamic KV Cache Balance)를 제안하였다. 어텐션 헤드 별로 KV를 P-C와 D-C에 나눠 저장하여 데이터 이동을 줄였다. 또한 CCAD (Cross Cluster Attention Dataflow)를 제안하여, P-C에 저장한 KV를 D-C로 가져와 연산하는 것이 아닌, 쿼리를 P-C로 전송하고 P-C에서 어텐션 연산을 수행하는 방식을 제시하였다. 이를 통해 칩 간 데이터 이동을 줄여 에너지 효율을 높이고 지연 시간을 46% 감소시켰다.

본 논문은 몇 가지 아쉬운 점이 있다. 첫 번째로, CFHM의 등장 배경이었던 Prefill단계의 문제점이다. 본 논문은 Prefill 단계에서 SRAM이 유휴 상태이기 때문에 비효율적이라고 주장하였다. 보통 엣지 디바이스에서 LLM을 돌리는 경우, decode 단계는 외부 메모리 대역폭이 병목이 되기 때문에 온칩 SRAM이 많이 필요하지 않다. 따라서 칩 설계 시 온칩 SRAM의 크기를 prefill 단계 요구량에 맞춰도 무방하다. 이러한 이유로 Prefill 단계에서 온칩 SRAM의 유휴가 문제가 된다는 주장은 설득력이 조금 부족하다.

두 번째 문제점은 HCWG로 인한 칩 간 데이터 이동 에너지 오버헤드가 명시되지 않았다

는 점이다. HCWG는 P-C에서 가중치를 읽어 압축 후 칩 간 링크를 통해 D-C로 전송하는 방식인데, 칩 간 데이터 이동 에너지는 일반적으로 상당히 크다. 따라서 압축으로 인한 DRAM 접근 속도 이득과 추가적인 칩 간 전송으로 인한 에너지 손실에 대한 정량적인 분석이 필요하다.

세 번째로 다양한 시나리오에 대한 검증이 부족하다. 다중 배치 환경에서는 프롬프트 길이와 생성 길이에 따라 prefill과 decode의 작업 부하 비율이 크게 달라진다. 그러나 본 연구는 다중 작업을 처리하는 경우 P-C와 D-C를 각각 2개씩 고정 할당해야 하기 때문에 이러한 작업 부하 변화에 유연하게 대응하기 어렵고 비효율적일 수 있다. 또한 CCAD와 UDB의 경우, 시나리오에 따라 사용하지 않는 것이 더 유리한 경우도 있을 것이다. 예를 들어 시퀀스 길이 L이 짧으면 KV 캐시 전송량 자체가 작아 쿼리 전송 방식의 이점이 감소하고, P-C가 prefill로 바쁜 상황에서는 어텐션 연산이 오히려 병목이 될 수 있다. 그러나 이러한 트레이드오프에 대한 분석이나 최적 지점이 어떤 경우인지에 대한 분석이 부족하다.




[그림 1] Disaggregated Computing의 문제점과 논문 21.1의 해결법

## 참고문헌

[1] J. -H. Kim et al., "Adelia: A 4nm LLM Accelerator with Streamlined Dataflow and Dual-Mode Parallelization for Efficient Generative AI Inference," 2025 Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)

## 저자정보

**하상우 박사과정 대학원생**

- 소속 : KAIST
- 연구분야 : 딥러닝 가속기 설계
- 이메일 : sangwoo_ha@kaist.ac.kr
- 홈페이지 : https://ssl.kaist.ar.k

# A-SSCC 2025 Review

서울대학교 전기정보공학부 박사과정 이재성

## Session 8 AI Accelerators and Security Circuits

이번 2025 IEEE A-SSCC의 Session 8은 AI Accelerators and Security Circuits라는 주제로 구성되었으며, 이 세션에서는 AI 연산을 위한 하드웨어 가속기와 보안 회로를 한 자리에서 조망할 수 있도록 총 6편의 논문이 발표되었다. 최근 온디바이스 AI의 적용 범위가 추론을 넘어 생성·적응·학습까지 확장되면서, 엣지 가속기의 설계 목표도 단순 연산 성능(TOPS) 경쟁에서 시스템 효율(메모리 트래픽·지연·전력) 중심으로 빠르게 이동하고 있다. Session 8의 Paper 8.1–8.3은 이러한 흐름을 압축적으로 보여준다. 확산(Diffusion) 기반 생성 모델처럼 반복 단계가 많은 워크로드는 계산량보다 반복 누적되는 데이터 이동과 비선형 연산 오버헤드가 병목이 되기 쉽고, 반대로 현장 적응을 위한 온칩 학습은 업데이트(write) 트래픽이 에너지 비용을 지배한다. 세 논문은 공통적으로 이 병목을 "연산 최적화"가 아닌 데이터플로우 재구성(재사용·스킵·퓨전·파이프라인)으로 풀어낸다.

**#8-1** 본 논문은 14-nm에서 확산(Diffusion) 기반 텍스트-투-모션(Text-to-Motion) 생성 모델을 엣지에서 실시간 수준으로 구동하기 위한 11.0 TOPS/W급 가속기를 제안한다. 생성형 AI가 이미지/텍스트를 넘어 모션·비디오·멀티모달로 확장되면서, "클라우드에서 만들어 내려받는 콘텐츠"가 아니라 "기기에서 즉시 생성해 상호작용하는 콘텐츠"에 대한 수요가 커지고 있다. 특히 XR/AR, 아바타·가상 캐릭터, 로봇의 동작 생성/보정 같은 응용에서는 지연이 곧 경험 품질을 좌우한다. 이때 diffusion 계열은 단일 pass가 아니라 수십 단계 denoising을 반복하며, 매 단계마다 Transformer 블록이 호출되므로, 단순히 연산량이 큰 수준을 넘어 반복 실행 자체가 시스템 지연과 에너지의 누적 병목으로 직결된다. 따라서 diffusion을 엣지에서 "쓸 만한 속도"로 돌리려면, 단순히 MAC을 빠르게 만드는 가속기만으로는 부족하고, 반복 구조에서 발생하는 중복과 데이터 이동을 "구조적으로" 줄이는 것이 필요하다.
확산 모델의 본질적 병목은 (i) 수십 단계 denoising에 의해 동일한 Transformer 블록이 반복 실행되며 연산/메모리 비용이 누적된다는 점, (ii) 특히 블록 내부에서 FFN이 연산량의 과반(≈60%+)을 차지해 반복 step이 많을수록 FFN 비용이 지배적으로 커진다는 점, (iii) softmax·LayerNorm 등 "비-GEMM" 연산은 FLOP보다 레지스터/버퍼 접근 에너지가 지배적일 수 있어, matmul만 최적화하면 오히려 전체 에너지 효율이 기대만큼 오르지 않는다는 점이다. 이 논문이 설득력 있는 이유는, 이 세 병목을 "각각 따로"가 아니라, diffusion의 반복 구조라는 공통 원인에서 출발해 "재사용·정밀도·비선형 연산 최적화"로

한 묶음으로 풀어냈기 때문이다.

저자들은 "반복 step 사이에는 입력/활성의 시간적 유사성(temporal similarity)이 존재한다"는 관찰을 하드웨어 최적화로 연결하며, 핵심을 세 축으로 정리한다. 첫째, Unstructured activation reuse로 FFN을 무작정 재사용하지 않고, GELU 활성 분포에서 영향이 큰(고활성) 값만 선별 재계산하고 나머지는 재사용하는 방식으로 정확도(모션 품질) 열화를 억제한다. 여기서 중요한 포인트는 "재사용" 자체가 아니라 "재사용의 기준"이다. diffusion step 간에는 어느 정도 유사성이 존재하지만, 생성 모델은 작은 오차가 누적되면 결과가 흔들릴 수 있다. 따라서 이 논문은 FFN 전체를 재사용하는 공격적 접근 대신, 품질에 민감한 부분만 다시 계산하는 선택적 정책을 취한다. 구현 관점에서는 첫 step은 dense 계산으로 기준 출력을 만들고, 이후 step은 마스크 기반으로 필요한 항만 갱신하는 "dense+selective-sparse" 실행 패턴을 구성하여 반복 비용을 구조적으로 낮춘다. 이 방식은 단순 pruning과 달리, 모델 파라미터를 희소화하는 것이 아니라 활성값의 시간적 변화량을 활용해 "이번 step에서 굳이 다시 계산할 필요가 있는가?"를 따지는 점에서, 최근 가속기 연구의 큰 흐름('정적 sparsity'보다 '동적/데이터 의존 sparsity')과 맞닿아 있다. 또한 unstructured(비정형) 희소성은 하드웨어 입장에서는 다루기 까다롭지만, diffusion처럼 반복 횟수가 큰 워크로드에서는 step마다 누적되는 이득이 커서, "까다로워도 할 만한 최적화"가 된다.

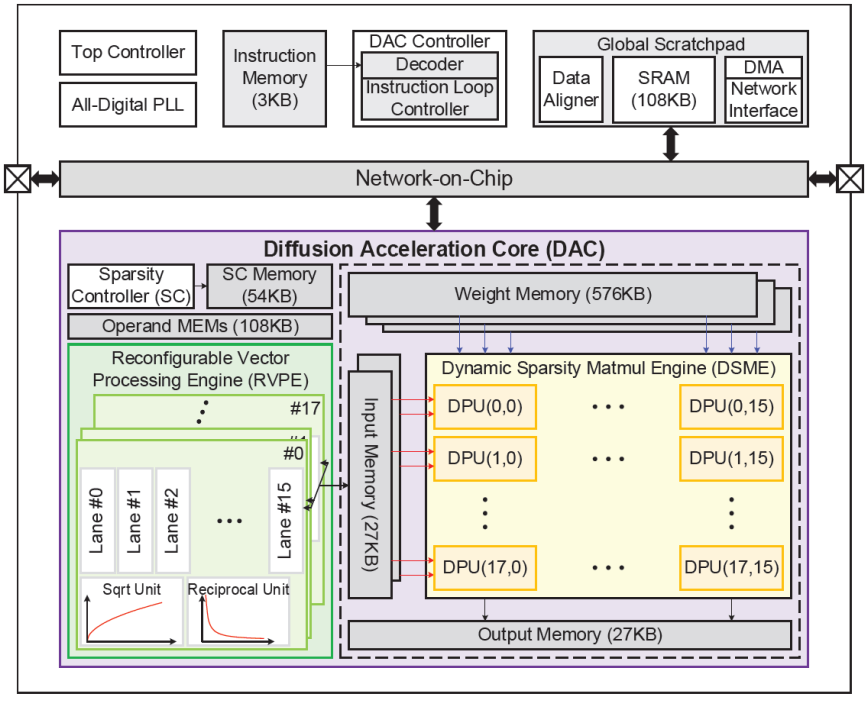
둘째, DSME(동적 희소 matmul 엔진)와 RVPE(재구성 벡터 엔진)를 분리한 DAC(Diffusion Acceleration Core) 아키텍처를 통해, matmul(특히 selective-sparse update)과 LayerNorm/activation/softmax 등 비선형·정규화 연산을 각각 최적 데이터패스로 처리한다. 이 분리는 단순한 모듈 나눔이 아니라, diffusion 가속에서 "진짜 병목이 무엇인지"를 반영한 구조적 선택이다. matmul은 처리량이 중요하지만, softmax·정규화·활성함수는 처리량보다 데이터 재배치/스케일링/정밀도 관리가 성능과 에너지에 더 큰 영향을 준다. 따라서 RVPE는 민감 연산에서 고정밀 누산/가변 정밀 모드를 제공해 중간 requantization으로 인한 품질 손상을 막는 "정확도 방어선" 역할을 한다. 이 설계는 최근 엣지 생성형 가속에서 점점 중요해지는 "혼합정밀(mixed-precision) 관리" 문제를 정면으로 다룬다. 즉, 전부를 높은 정밀도로 유지하면 전력/면적이 커지고, 전부를 낮은 정밀도로 내리면 품질이 무너지는 상황에서, 어떤 연산을 예외로 보호하고(예: LayerNorm), 어떤 연산은 공격적으로 최적화할지(예: FFN 선택적 재계산)를 하드웨어 수준에서 분업한다.

셋째, LUT 기반 softmax에서 연산 자체보다 중간값 저장/읽기(레지스터 R/W)가 더 큰 에너지 비중을 차지한다는 점을 겨냥해, max-subtraction 이후 exp 항에서 작은 값들을 threshold로 제거함으로써 "저장할 가치가 낮은 항"을 계산·저장 단계에서 동시에 줄인다. softmax는 많은 가속기에서 "대충 LUT로 때우면 된다"고 여겨지기 쉬운데, 실제로는 LUT 접근·중간값 저장·정규화 과정에서 데이터 이동이 커져 에너지가 과하게 나갈 수 있다. MoDiff의 threshold softmax는 이 지점에서 "정확도에 덜 기여하는 tail"을 잘라내어, 연산

량뿐 아니라 저장·이동량까지 동시 감소시키려는 전략이다. 이 접근은 diffusion 가속이 matmul 중심에서 벗어나 "비-GEMM을 포함한 end-to-end 효율"로 경쟁 축이 바뀌고 있음을 상징적으로 보여준다.

결과적으로 이 설계는 단순 TOPS/W 향상뿐 아니라, 확산 모델에서 중요한 end-to-end latency(반복 step 누적 지연)와 생성 품질 지표(FID)를 함께 만족시키도록 구성되어 있다. 특히 "실시간"이라는 목표는 단순 peak throughput이 아니라 반복 step 수 × step당 지연의 누적값이 관건이므로, step마다 FFN을 얼마나 덜 돌릴 수 있는지가 핵심이 된다. MoDiff는 이 포인트를 정확히 짚고 FFN 선택적 재계산으로 반복 비용을 줄이며, RVPE와 softmax 근사로 비-GEMM 오버헤드가 발목을 잡지 않게 만든다.

측정 결과, 0.63–0.94 V 및 50–600 MHz에서 동작하며, 600 MHz 기준 6.71 TOPS 수준의 처리량과 최대 11.0 TOPS/W 효율을 보고한다. HumanML3D에서 MLD/MDM 기반 텍스트-투-모션 생성 평가 시 엣지 GPU 대비 최대 17.5× 속도 향상을 달성하면서도 FID 열화는 매우 제한적이라고 제시해, "엣지 생성형 가속"이 실사용 응용(AR/VR·모션 합성 등)으로 넘어가기 위한 정량 목표를 제시한다. 다만 이 논문이 더 강해지려면, diffusion 계열이라도 텍스트-이미지/비디오 등 다른 생성 과제에서 temporal similarity 기반 재사용이 얼마나 잘 성립하는지, 또 모델 구조가 달라질 때(예: attention 비중 증가/감소, 다른 normalization) RVPE/DSME 분할이 얼마나 일반적으로 통하는지에 대한 추가 논의가 뒤따라야 한다. 그럼에도, "diffusion은 엣지에서 어렵다"는 통념에 대해, 반복 구조의 중복을 하드웨어-알고리즘 공동설계로 해체하여 실시간 가능성을 정면으로 보여준 점에서, Session 8의 방향성을 가장 강하게 드러낸 논문 중 하나로 볼 수 있다.




[그림 1] 제안한 MoDiff의 전체 칩 구조

**#8-2** 본 논문은 22-nm에서 온칩 증분학습(on-chip incremental learning)을 지원하는 비동기(asynchronous) SNN 기반 프로세서를 제안하며, 멀티센서 인식과 적응 학습을 "엣지에서 가능한 비용"으로 만들기 위한 설계 방향을 제시한다. 최근 엣지 AI의 현실적 문제는 "모델이 커서 느리다"만이 아니다. 실제 제품/현장에서는 조명 변화, 배경 변화, 센서 노화/드리프트, 사용자 개인차, 장착 위치 변화 등으로 데이터 분포가 계속 변한다. 이때 모델을 서버에서 주기적으로 재학습해 배포하는 방식은 지연·비용·프라이버시 문제를 동반한다. 결국 엣지 장치는 일정 수준의 적응 능력을 갖춰야 하며, 그 적응은 (i) 완전한 대규모 학습이 아니라, (ii) 새 환경/새 사용자에 맞춘 국소적 업데이트(증분학습/온라인 학습)의 형태로 나타나는 경우가 많다. ANP-R은 바로 이 요구를 SNN 기반으로 풀어내려는 시도이며, "온칩 학습이 가능한 뉴로모픽 프로세서"를 실제 칩 지표(pJ/SOP)로 밀어붙였다는 점에서 의미가 있다.

엣지 환경에서는 성능 유지를 위해 (i) 주기적 재학습이 아니라 기기 내부의 온라인/인크리멘털 업데이트가 필요해진다. 그러나 학습이 포함되는 순간 병목은 MAC 연산량보다 시냅스 업데이트(write) 트래픽과 그 에너지로 이동하며, 특히 SNN에서도 STDP 계열 업데이트가 잦아지면 이벤트 기반 희소성의 이점이 급격히 약화될 수 있다. 즉, "스파이크가 희소하니 저전력"이라는 장점은 추론(inference)에서 더 잘 성립하고, 학습(training/updates) 단계에서는 오히려 업데이트가 빈번해지면서 메모리 write가 지배할 수 있다. ANP-R이 중요한 이유는, 이 문제를 회피하지 않고 "학습 단계에서의 낭비를 줄이는 것"을 설계의 중심에 둔다는 점이다.

ANP-R은 "학습에서 불필요한 업데이트를 줄이고, 업데이트가 필요할 때도 저장 비용을 낮추는" 두 가지 축을 결합한다. 첫째, SSUS(Self-adaptive Synaptic Update Skipping)를 통해 학습 단계에서 의미 없는(변화가 거의 없는) 시냅스 업데이트를 스킵한다. 여기서 핵심은 "학습은 모든 시냅스가 같은 속도로 수렴하지 않는다"는 사실이다. 어떤 가중치는 초기에 빠르게 변하지만, 어떤 가중치는 곧 안정화되어 이후에는 거의 바뀌지 않는다. 그럼에도 전통적인 업데이트 구현은 매번 같은 절차로 update를 시도하기 때문에, 이미 수렴한 부분에 에너지를 계속 태운다. SSUS는 직전 업데이트 결과가 동일하거나 변화가 미미한 구간에서 skip count를 누적해 일정 임계에 도달하면 업데이트를 생략하는 방식으로, 정확도 손실을 최소화하면서도 업데이트 횟수를 큰 폭(논문에서 최대 65% 수준)으로 줄여 학습 에너지를 직접 절감한다. 중요한 건, 이 방식이 단순 "업데이트 간격을 늘리는 스케줄링"이 아니라, 시냅스별/상태별로 "필요성"을 판단하는 적응적 게이팅이라는 점이다. 따라서 동일한 네트워크라도 데이터 분포나 태스크가 바뀌면 스킵 패턴이 달라질 수 있고, 그 자체가 엣지 환경 변화에 대한 유연성을 제공한다.

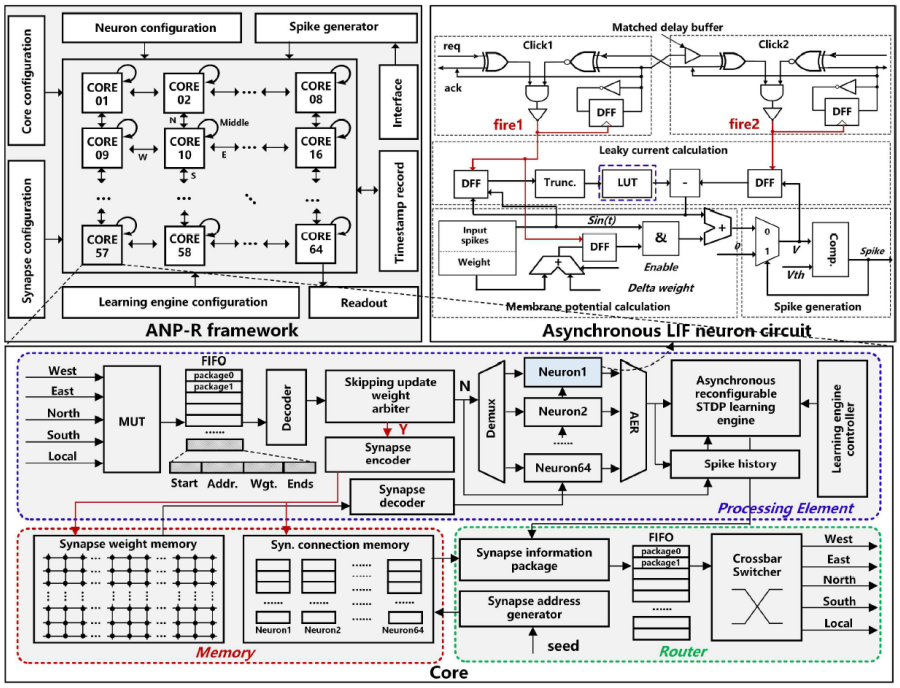
둘째, TWLW(Trained Weights Low-bit Width) coded storage를 도입해 시냅스 저장 포맷 자체를 경량화한다. 학습이 가능한 시스템에서 가중치 표현은 단순 압축이 아니라 "업데이트 통계/정확도 민감도"와 함께 설계되어야 한다. 예컨대 극단적으로 비트를 줄이면 저

장은 싸지지만 학습이 불안정해질 수 있고, 반대로 정밀도를 유지하면 write 에너지가 지배한다. TWLW는 저비트 코딩 기반 저장으로 시냅스 write 에너지(논문에서 62% 수준) 절감과 함께, 성능 열화를 제한된 범위로 관리하는 접근을 제시한다. 이 접근은 "가중치 저장 최적화"를 단순한 메모리 압축으로 보지 않고, 학습 과정의 동역학(어떤 업데이트가 얼마나 자주/얼마나 크게 발생하는가)과 묶어서 보는 흐름에 가깝다. 즉, 엣지 학습에서 중요한 건 '한 번 저장할 때의 비용'뿐 아니라 '학습 동안 몇 번 쓰는지'이므로, write 빈도/패턴을 바꾸는 SSUS와 저장 비용을 낮추는 TWLW를 함께 쓰는 조합이 자연스럽다.

아키텍처적으로는 64개의 비동기 코어와 라우터 기반 연결로 구성된 coarse-grained reconfigurable 구조를 취해, 다양한 SNN 토폴로지/학습 연산 흐름을 코어 단위로 재구성할 수 있도록 한다. 여기서 "비동기(asynchronous)"는 단순히 클록을 없앴다는 의미를 넘어, 이벤트가 없을 때는 자연스럽게 멈추고(event-driven idling), 필요한 곳만 동작할 수 있다는 점에서 always-on 센서 지능과 궁합이 좋다. 또한 coarse-grained reconfigurable이라는 선택은, 뉴로모픽이 본질적으로 다양한 연결/학습 규칙/태스크를 요구한다는 점을 반영한다. 코어당 LIF/IF 뉴런과 시냅스 블록을 갖추고, 전체적으로 4096 neurons 및 0.262M synapses 규모를 구성하여 "단일 태스크 데모" 수준을 넘어 멀티태스크/멀티센서 실험을 수행한다. 여기서 실무적으로 중요한 질문은 "확장성"이다. 코어 수가 늘어나면 라우팅/통신이 병목이 되기 쉽고, 비동기 네트워크에서 혼잡/지연 편차가 학습 안정성에 영향을 줄 수도 있다. ANP-R은 라우터 기반 연결로 확장성을 주장하지만, 향후 더 큰 규모(혹은 더 복잡한 멀티센서 융합)로 갈 때 어떤 병목이 먼저 나타날지(통신? 메모리? update?)는 후속 연구의 핵심 질문이 될 것이다.

측정 결과는 효율을 pJ/SOP 관점에서 제시하며, 0.6 V에서 0.88 pJ/SOP의 피크 효율을 보고한다. 또한 태스크별로 학습 step당 에너지와 SOP 효율을 제시해, "학습 포함 엣지 프로세서"에서 중요한 지표가 단순 처리량이 아니라 energy per step / update cost임을 강조한다. 이 점은 독자에게 매우 중요한 메시지를 준다. 엣지 학습의 평가는 종종 '정확도만' 또는 '추론 에너지'만으로 논의되지만, 실제로는 학습이 포함되면 "학습 1 step에 얼마를 쓰는가", "업데이트를 얼마나 줄였는가"가 시스템 배터리 수명과 직결된다. ANP-R은 update skipping과 coded storage로 그 비용을 직접 제어할 수 있음을 보여준다.

정리하면 ANP-R은 비동기·이벤트 기반 계산이라는 SNN의 강점을 "학습"까지 확장하면서, 업데이트 스킵과 저장 경량화로 온칩 적응의 비용을 통제해, 향후 엣지 AI가 요구할 "항상-동작(always-on) + 환경 적응"의 실질적 구현 경로를 제시한다. 특히 이 논문이 남기는 가치 있는 질문은 다음과 같다. (i) SSUS의 스킵 임계/정책은 태스크별로 어떻게 설정되는가(고정? 적응?), (ii) TWLW의 코딩 방식은 장기 학습 안정성에 어떤 영향을 주는가, (iii) 비동기 라우팅 네트워크에서 통신 지연의 분산이 학습/추론 성능에 어떤 영향을 주는가. 이 질문들은 곧 "온칩 학습 뉴로모픽"이 실사용으로 갈 때 반드시 풀어야 할 과제이며, ANP-R은 그 출발점을 꽤 설득력 있게 제시한 사례로 볼 수 있다.




[그림 2] 제안한 ANP-R의 시스템 다이어그램(좌상), 이너코어 프레임워크(하단), 그리고 이너코어에 적용된 제안 비동기 LIF 뉴런 회로(우상)

**#8-3** 본 논문은 40-nm 에서 엣지 환경에서의 DNN/SNN 학습과 추론을 모두 지원하는 Learning-in-Memory 프로세서를 제안하며, 학습 워크로드의 본질적 병목을 "연산"이 아닌 외부 메모리 접근(EMA)과 단계 분리(Forward/Backward 분리 실행)로 규정한다. 이 관점은 매우 중요한데, 실제로 학습은 inference 와 달리 (i) forward 활성 저장이 필요하고, (ii) backward/gradient 계산을 위해 중간 텐서를 다시 읽어야 하며, (iii) 업데이트를 위해 weight 및 optimizer state 까지 반복적으로 접근한다. 결과적으로 학습에서는 "연산기 처리량"이 충분해도 DRAM 왕복이 성능을 묶고, DRAM 이 에너지를 지배하는 상황이 흔하다. 특히 엣지에서는 DRAM 대역폭이 제한적이고, DRAM 왕복은 전력/발열/배터리 수명에 치명적이므로, 학습을 엣지로 가져오려면 결국 "외부 메모리 접근을 줄여야 한다"는 결론에 도달한다. Lemem 은 이 결론을 전제로, PIM 매크로만 제시하는 수준을 넘어 학습 루프 전체를 시스템적으로 재배선하려는 시도라는 점에서 주목할 만하다.

학습은 forward 뿐 아니라 error propagation, gradient generation, weight update 까지 포함하기 때문에 중간 텐서와 업데이트 트래픽이 폭증하고, 결과적으로 DRAM 왕복이 성능과 에너지의 상한을 만든다. Lemem 은 이를 해결하기 위해 (i) D-PIM(dual-mode multi-precision ping-pong PIM macro)로 데이터가 메모리 근처에서 곧바로 연산되도록

만들고, (ii) PFBC(pipelined forward-backward processing core)로 학습의 전·후방 흐름을 같은 코어 내부에서 파이프라인으로 엮어 유휴 시간을 줄이며, (iii) 레이어 퓨전(Layer-fused) 라우팅 구조(LFR/I2FR)로 레이어 경계에서의 외부 메모리 왕복을 구조적으로 차단한다. 이 세 가지는 각각 "연산 위치(메모리 근처)", "시간 구조(파이프라인)", "공간/연결 구조(퓨전 라우팅)"를 담당하며, 합쳐서 EMA 를 줄이는 방향으로 정렬된다.

먼저 D-PIM 은 '학습/추론 모두'를 염두에 둔 dual-mode, multi-precision, ping-pong 구조를 통해 데이터 이동을 줄이려 한다. 엣지 학습에서 단순히 PIM 이 유리하다는 주장만으로는 부족하고, 실제로는 (1) 정밀도 선택이 학습 안정성에 미치는 영향, (2) 버퍼링/스케줄링(핑퐁)에서 생기는 idle time, (3) 매크로 주변 로직의 오버헤드가 성능을 갉아먹지 않는지가 관건이다. Lemem 은 D-PIM 을 "학습 친화적으로" 구성하고, 나머지 파이프라인과 결합해 매크로를 최대한 놀리지 않는 방향으로 설계를 전개한다.

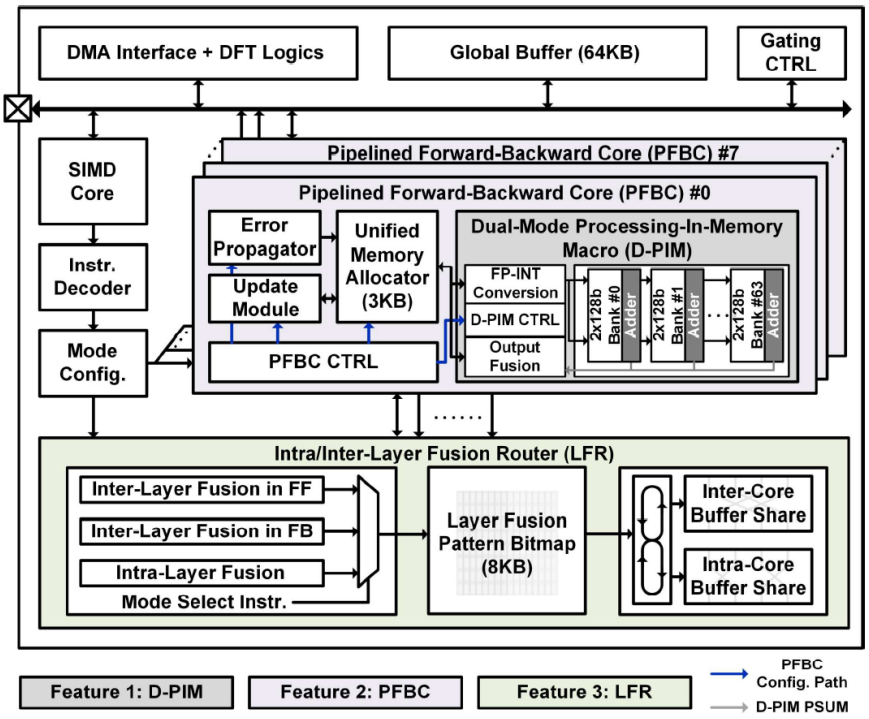
그 다음 PFBC 는 Lemem 의 핵심적인 시스템 아이디어 중 하나다. 학습의 전/후방은 전통적으로 분리되어 실행되며, 레이어 간 경계에서 activation/gradient 를 저장하고 다시 읽는다. PFBC 는 forward, error propagation, gradient generation, weight update 를 단절된 단계로 실행하지 않고, 서로 다른 모듈(D-PIM/EP/WUM 등)을 동시 구동해 "생성된 activation/feature 를 즉시 다음 단계에 소비"하게 함으로써 학습 지연과 EMA 를 동시에 줄인다. 여기서 중요한 건 "파이프라인"이라는 단어가 단순히 스테이지를 나눴다는 의미가 아니라, 학습의 종속 구조를 가능한 범위에서 겹쳐 실행(overlap) 해 유휴 시간을 줄이는 것이다. 엣지에서는 클록을 올려 해결하기 어렵기 때문에, 이런 구조적 overlap 이 처리량을 좌우한다.

마지막으로 레이어 퓨전 라우팅(LFR/I2FR)은 "왜 매번 DRAM 을 가야 하는가?"라는 질문에 대한 하드웨어적 답이다. 많은 학습 가속은 레이어 내부 최적화(예: GEMM 효율)로 끝나지만, 실제로는 레이어 사이에서 feature map 이 이동하며 트래픽이 터진다. Lemem 은 inter-layer FF fusion 과 intra-layer FB fusion 을 함께 고려해 학습 시 필요한 데이터가 온칩에서 순환하도록 만드는 데 목적이 있다. 즉, forward 에서 생성된 데이터가 가능한 한 빠르게 backward/gradient 단계에서 소비되고, 레이어 경계에서 "저장-재로딩"이 반복되는 구조를 끊겠다는 것이다. 이는 최근 학습 가속에서 자주 언급되는 "operator/layer fusion"을 라우팅/캐시 패브릭까지 확장한 형태로 볼 수 있으며, '연산기'가 아니라 '데이터 이동 경로'를 설계의 중심으로 끌어올린다는 점에서 트렌드의 정점에 있다.

이 접근의 효과는 정량적으로도 강하게 제시된다. 논문은 EMA를 최대 4.54× 감소시키고, DDR5 전송 대역폭을 2.0× 증가, 그 결과 에너지를 74.7% 절감하고 GPU 대비 3.48× 속도 향상을 보고한다. 또한 성능/효율 지표로 24.21 TFLOPS 처리량과 179.8 TFLOPS/W 에너지 효율을 제시해 "학습 포함 프로세서"의 경쟁력을 강조한다. 여기서 주목해야 할 것은, Lemem 이 단순히 "PIM 이니까 효율이 좋다"라고 말하는 것이 아니라, EMA 감소/대역폭 증가/에너지 절감/속도 향상이라는 시스템 레벨 연쇄 효과를 근거로 제시한다는 점이다. 엣지에서 학습을 하려면 결국 시스템 전체(메모리·네트워크·스케줄링)가 바뀌어야 하는데, 이 논문은 그 변화를 하나의 통합 설계로 보여준다.

더 나아가 Fashion-MNIST, CIFAR-10, N-MNIST, DVS-Gesture 등 DNN/SNN 을 아우르는 벤치마크 구성을 통해 범용성을 주장하며, 엣지에서 학습이 가능해지기 위한 조건(샘플당 에너지/iteration 지연 등)을 실측 수치로 제시한다. DNN 과 SNN 을 함께 다루는 점은 특히 의미가 있는데, 실제 엣지 시스템은 프레임 기반 인식(DNN)과 이벤트 기반 센싱(SNN)이 공존하는 방향으로 가고 있고, "학습 포함" 요구가 생기면 플랫폼이 분리될수록 통합 비용이 커진다. Lemem 은 이 공존을 하드웨어 구조 차원에서 수용하려는 방향성을 보여준다.

다만 Lemem 이 강한 주장을 하는 만큼, 독자가 자연스럽게 던지게 되는 질문도 있다. (i) layer-fused pipeline 을 어떤 범위까지 자동화할 수 있는가(워크로드가 달라지면 스케줄이 얼마나 바뀌는가), (ii) 다양한 네트워크 구조(복잡한 skip connection, attention-heavy 모델, 멀티모달 결합)에서도 EMA 감소가 같은 정도로 유지되는가, (iii) 학습의 수치 안정성/정확도 보존을 위해 어떤 정밀도/스케일링 정책이 필요한가. 다시 말해, 이 논문은 "엣지 학습이 가능하다"를 강하게 보여주지만, 그 가능성이 실사용에서 보편화되려면 툴체인/컴파일/모델 다양성까지 포함한 확장 논의가 뒤따라야 한다. 그럼에도 불구하고, Lemem 은 "학습은 결국 시스템"이라는 관점을 가장 정면으로 구현한 논문으로, PIM 매크로 단품 성능이 아니라 학습 루프 전체를 관통하는 데이터플로우(파이프라인·퓨전·캐시/라우팅)를 설계의 중심에 두어 엣지 학습의 실현 가능성을 한 단계 끌어올린다.




[그림 3] 제안한 Lemem 프로세서의 전체 구조

## 저자정보

**이재성 박사과정 대학원생**

- 소속 : 서울대학교
- 연구분야 : SNN 및 PIM
- 이메일 : jslee0122@snu.ac.kr
- 홈페이지 : https://sites.google.com/view/ic3-snu

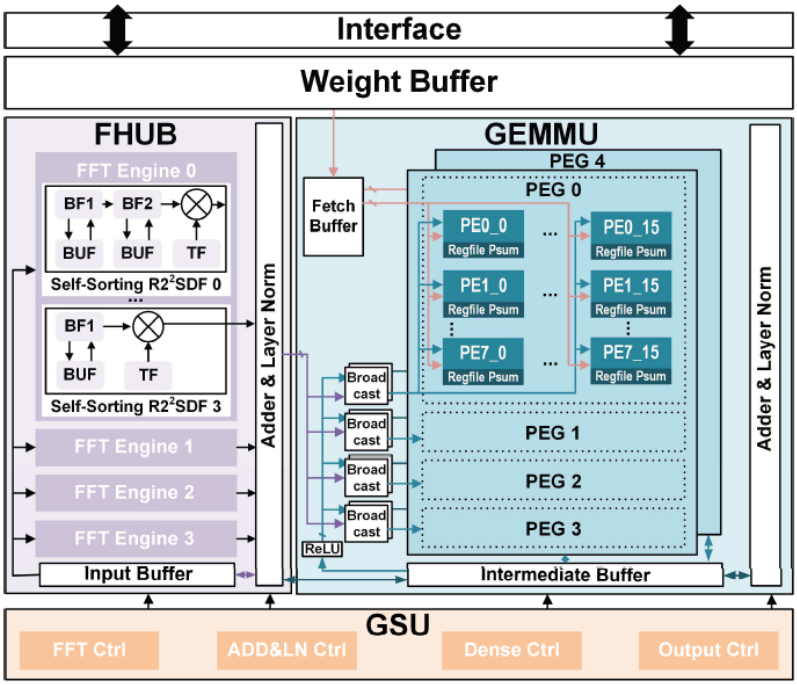
# A-SSCC 2025 Review

한국과학기술원 전기및전자공학부 석사과정 박성용

## Session 8 AI Accelerators and Security Circuits

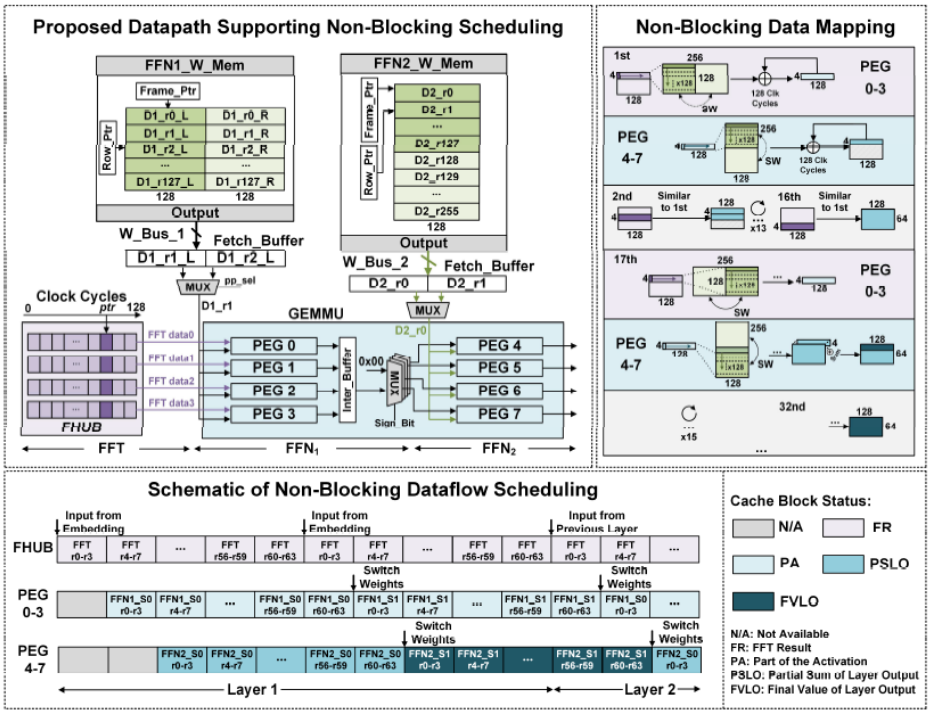
이번 A-SSCC 2025의 Session 8은 AI Accelerators and Security Circuits 라는 주제로 총 6편의 논문이 발표되었다. 이 세션에서는 AI 연산을 가속시키는 architecture와, AI를 활용한 TRNG, PUF와 같은 HW 보안 회로에 중점을 두었다.

**#8-4** 본 논문은 EdgeFTX 구조를 제안하여, Transformer 모델의 self-attention을 FFT(Fast Fourier Transform) 기반 Token Mixing으로 대체함으로써, 배터리 제약이 극심한 IoT 및 웨어러블 디바이스와 같은 엣지 컴퓨팅 환경에서 최신 Transformer 모델을 구동하고자 한다. 본 연구는 Attention 메커니즘 자체를 경량화하거나 희소성(Sparsity)을 활용하는 데에 집중하는 대신, FFT 기반의 Attention-Free 구조를 채택함으로써, 알고리즘의 시간 복잡도를 $O(N^2)$에서 $O(N \log N)$으로, 공간 복잡도를 $O(N)$으로 낮춰 모델을 경량화하였다. 본 연구는 NAS(Neural Architecture Search)를 활용하여 모델의 정확도와 하드웨어 효율성(FFT 연산 수, SRAM 접근 횟수) 간의 파레토 최적 지점을 탐색하였다. 그 결과, 베이스라인 대비 FFT 연산은 75%, SRAM 접근은 50% 줄이면서도 정확도 저하는 0.76%로 최소화하였다.




[그림 1] 제안된 Transformer 가속기 전체 구조

그림 1은 EdgeFTX의 전체 architecture에 대한 block diagram이다. EdgeFTX는 FFT 연산과 기존 FFN(Feed Forward Network)의 행렬 연산을 효율적으로 처리하기 위해, 4개의 직렬 FFT 엔진(FHUB)과 8개의 병렬 PE 배열(GEMMU)로 구성된 FFT-FFN heterogeneous 아키텍처를 설계하였다. FHUB는 Attention 메커니즘을 대체하는 Token Mixing을 수행하는 전용 모듈이다. 이 모듈은 4개의 직렬 FFT 엔진으로 구성되며, 각 엔진은 128-point FFT를 처리한다. 내부적으로는 Radix-$2^2$ SDF(Single-path Delay Feedback) 구조를 기반으로 설계되었는데, 가장 큰 특징은 Self-Sorting In-Place 회로를 사용했다는 점이다. 일반적인 FFT 하드웨어는 연산 결과인 주파수 도메인 데이터의 순서가 뒤섞이기 때문에 이를 바로잡기 위한 대용량의 재정렬 버퍼(Reordering Buffer)가 필수적이다. 그러나 FHUB 내의 버터플라이 모듈은 연산과 동시에 데이터를 제자리에 저장(In-place)하며 자동으로 정렬되도록 설계되어, 추가적인 버퍼나 정렬 로직을 제거하여 전력을 7% 절감하고 면적을 11% 줄이는 결과를 보였다. GEMMU는 Transformer의 FFN 연산, 즉 행렬 곱셈을 가속하기 위한 모듈이다. 이는 8개의 병렬 PE array Group(PEG)로 구성되며, 각 PEG는 128개의 PE를 포함한다. 각 PE는 가중치를 로드하고 MAC 연산을 수행하여 FFN의 벡터 스케일링 연산을 병렬로 처리한다.




[그림 2] Non-Blocking Dataflow Scheduling

또한, FFT-FFN 유닛 간의 데이터 의존성으로 인한 파이프라인 Stall를 막기 위해 Non-Blocking Scheduling(NBS) 기법을 사용하였으며, 자세한 dataflow 스케줄링은 그림 2에서 확인할 수 있다. 이는 FFN을 세부 태스크로 나누어, 데이터가 완전히 생성될 때까지 기다리지 않고, FFT 결과가 나오는 즉시 미세 단위(Fine-grained)로 다음 연산에 공급하여 연산 유닛의 가동률을 100%에 가깝게 유지하였다. 이를 통해 layer execution에 걸리는

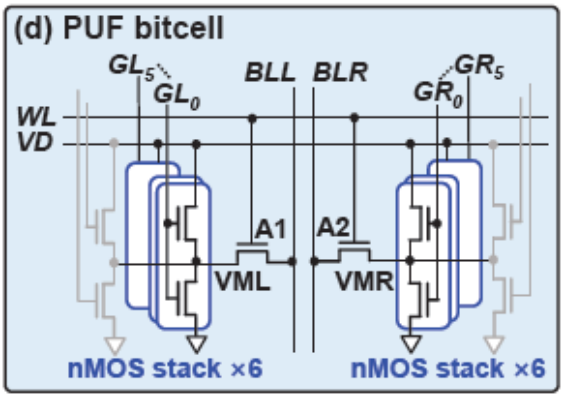
latency를 기존 직렬 스케줄링 대비 1.49배 단축하였다.

22nm CMOS 공정으로 제작된 EdgeFTX 칩은 0.51V, 100MHz 동작 조건에서 3.6mW의 전력을 소모하며, 1.39 TOPS/W의 에너지 효율을 달성하였다. SST-2 벤치마크에서 동일 규모의 BERT 모델 대비 2.3% 높은 정확도를 보였으며, 최신 General-purpose edge 가속기 대비 전력 소모를 최대 17.6배 낮춤으로써, FFT 기반 접근법이 유효함을 보여주었다.

**#8-6** 본 논문은 ML(Machine Learning)을 활용하여, PUF의 전압 및 온도 변동(VT variation)에 의한 Bit Flipping 오류를 해결하고자 하였다. 먼저 PUF(Physically Unclonable Function)는, 동일한 웨이퍼와 마스크를 사용하더라도 반도체 제조 공정에서 발생하는 Vth mismatch와 같은 Process Variation을 이용한다. 특정 입력을 인가했을 때, 미세한 mismatch를 증폭하여 0 또는 1의 값으로 출력함으로써 Chip에 고유한 지문(fingerprint)을 생성하는 기법이다. PUF는 반도체 지문으로서 고유해야 하지만, 동시에 언제 어디서나 동일한 값을 출력해야 하는 안정성이 필수적이다. 그러나, 동일한 Chip이여도 측정 시마다 온도, 전압 등의 환경 변화에 의해 bit-flipping되어 error가 발생하게 된다.

기존의 Dark-bit masking 기법은 불안정한 셀을 찾아 masking하므로 유효 셀이 줄어드는 문제가 있다. 이러한 문제를 해결하기 위한 Reconfigurable 기법 또한 선택지 제한으로 인해 BER(Bit Error Rate)가 존재하게 된다. 본 연구는 다음과 같은 방법을 통해, -40℃ ~ 120℃에서 zero-BER을 달성하는 reconfigurable PUF 구조를 제안하였다.

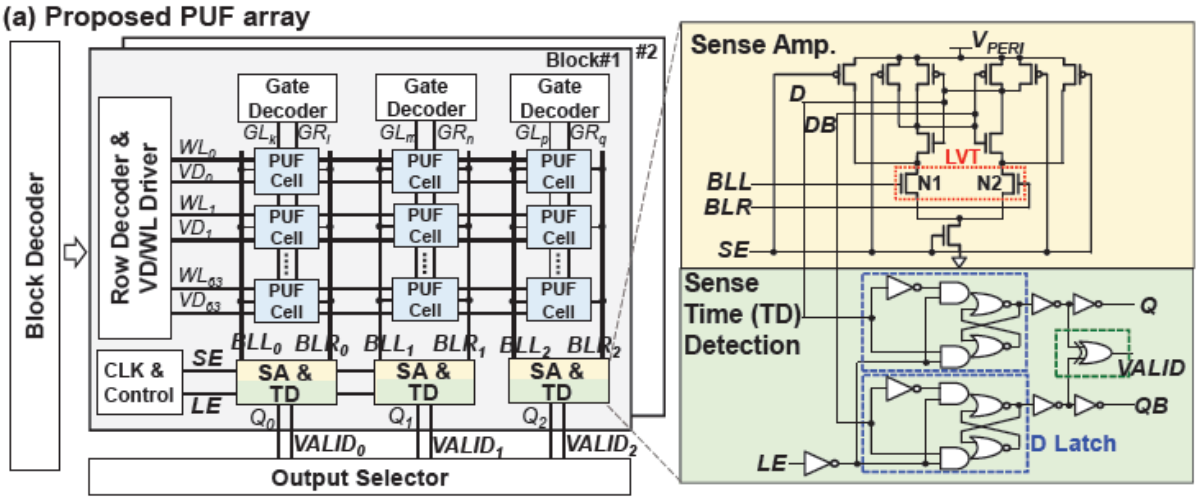
제안된 구조는 PUF 셀 하나당 36가지의 경로(Way) 구성을 지원하도록 설계되었다. 통계적 모델 분석에 따르면, 구성 가능한 경로의 수(N)가 증가할수록 해당 셀 내에서 공정 mismatch가 크고 안정적인 구성을 찾을 확률이 높아진다. N=2 또는 N=4인 기존 연구와 달리 N=36으로 확장함으로써, 불안정한 영역을 벗어난 안정적인 응답을 생성할 수 있는 확률을 높였다. 이를 통해 비트 셀을 버리지 않고도 셀당 4비트의 stable한 출력을 얻을 수 있었다.




[그림 3] PUF 비트 셀의 회로도

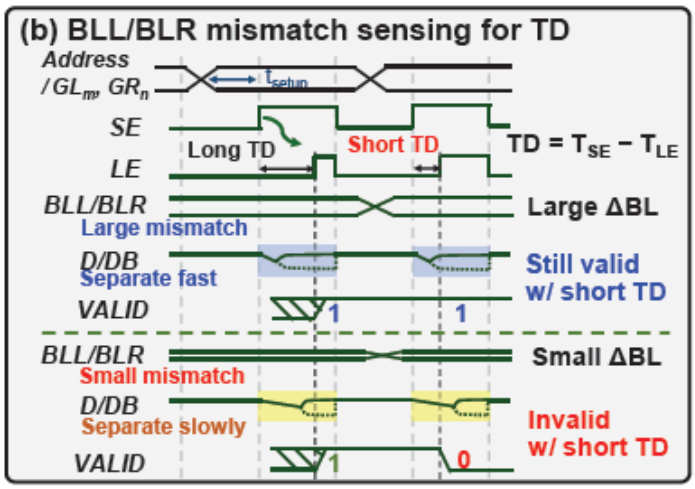
비트 셀의 구조는 그림 3과 같다. 셀은 대칭 구조로 이루어져 있으며, 좌측과 우측에 각각 6개의 직렬 NMOS 스택이 배치되어 있다. 외부에서 인가되는 게이트 선택 신호를

통해 좌우 각각 하나의 NMOS 경로를 활성화할 수 있으며, 이 조합을 통해 단일 셀 내에서 총 36가지의 서로 다른 전류 경로를 형성할 수 있다. 반도체 공정의 mismatch로 인해 선택된 좌우 NMOS 쌍 사이에는 구동 전류 차이가 발생하고, 이는 좌우 Bit line 간의 전압 차이로 나타난다.




[그림 4] PUF array의 블록도와 회로도

제안된 PUF array는 그림 4와 같다. Cell Array에는 Bit line 간의 미세한 전압 차이를 감지하여 증폭시키는 StrongARM latch로 구성된 SA(Sense Amplifier)와, 전압 차이가 증폭되어 0과 1의 안정 상태로 진입하기까지의 지연 시간을 측정하는 TD(Time Detection)가 연결되어 있다. Mismatch가 클수록 Bit line간 전압 차이가 크므로, 0과 1로 더 빠르게 분리되어 TD 값이 작아지게 된다. 따라서 측정된 TD 값을 mismatch를 측정하는 지표로 사용하게 된다. Mismatch에 따른 TD 값의 차이는 그림 5에서 확인할 수 있다. TD sweep을 수행하여 TD 값의 평균과 분산을 측정하여 mismatch의 정도를 나타내고, 동시에 전압 sweep을 수행하여 전압 민감도(VDD sensitivity)를 측정한다.




[그림 5] mismatch sensing을 통해 TD를 구하는 timing diagram

한편, 온도 변화에 따른 안정성을 확보하기 위해서는 일반적으로 높은 비용의 온도 chamber 테스트(Temperature Sweep)가 필요하다. 그러나 본 연구에서는 전압 민감도(VDD Sensitivity)가 온도 민감도(Temperature Sensitivity)와 높은 상관관계를 가짐을 확인하였다. 이를 통해 ML model은 mismatch를 나타내는 TD 값의 분포와 전압 민감도를 입력으로 받아 선형 회귀(linear regression)를 수행하여 -40℃ ~ 120℃ 온도 범위에서 BER을 예측하고, Hungarian 알고리즘을 이용하여 BER이 최소가 되는 4개의 경로 조합을 찾아내게 된다. 이러한 ML-based selection 기법을 사용하여, 해당 온도 범위에서 BER이 0에 가깝게 되도록 설계되었다.

65nm 공정에서 제작된 칩을 대상으로 광범위한 온도 테스트를 진행하여 다음과 같은 결과를 얻었다. 첫째로, 기존 휴리스틱 방식이나 단순 전압 Sweep 방식은 일부 오류를 남긴 반면, ML-based selection은 측정된 모든 온도 범위에서 BER이 6.51E−8 미만으로, 사실상 에러가 없는(zero-BER) 안정성을 보여주었다. 둘째로, 36-Way 구성을 위해 셀 내부 트랜지스터 수가 30개로 늘어났음에도 불구하고 dark-bit masking으로 버려지는 셀이 없어, 유효 비트당 면적은 674 $F^2$/bit로 기존 고신뢰성 PUF 대비 가장 작은 수준을 달성하였다. 셋째, 생성된 키의 고유성(Uniqueness)을 나타내는 해밍 거리(Hamming Distance)는 이상적인 값인 50%에 근접한 49.1%였으며, NIST의 randomness test를 모두 통과하여 보안 키로서의 적합성을 확인하였다.

## 저자정보


**박성용 석사과정 대학원생**

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

# 2025 ASSCC Review

한국과학기술원 바이오및뇌공학과 박사과정 석동열

## Session 4 Wireless Powering & Stimulation Systems for Implants

이번 2025 ASSCC의 Session 4에는 Wireless Powering & Stimulation Systems for Implants 라는 주제로 총 4편의 논문이 발표되었다. 본 세션은 체내 이식형 전자기기 구동을 위한 무선 전력전송기술과 신경계조절용 전기자극기를 다루고 있다. 무선 전력전송 분야에서 2편(4-1, 4-2), 전기자극기 분야에서 2편(4-3, 4-4)이 게재되었다. 무선전력전송 분야에서는 초음파(ultrasonics), 자기전기효과(magnetoelectronic effect)와 같이 전기장 코일 방식의 대안으로 제시되는 무선전력기술의 기존 한계를 지적하고 극복하기 위한 다양한 방법이 제안되었으며, 전기자극기 분야에서는 이식형 신경 자극기의 활용분야를 넓히고 장기적 안정성을 확보하기 위한 자극기 설계 전략, 비침습적 뇌-심부 자극을 위한 시간-간섭 자극 방식의 자유도와 정밀도를 높이기 위한 집적회로 설계를 소개하였다.

**#4-1** 논문은 중국 동부과학기술대학교와 북경대학교의 공동연구로 초음파를 통한 체내 무선전력 전송 및 역-산란(backscattering)을 활용한 업-링크 데이터 전송기술의 한계를 극복하기 위한 방법을 제안하고 있다. 이 주제에 관련하여 기존에는 하나의 초음파 트랜스듀서에서 전력전송과 데이터 송신 기능을 동시에 수행할 수 있도록 하는 방법이 없었으므로 시분할(time-division) 방식의 트랜스듀서 활용 또는 전력전송 이외의 데이터 송신용 트랜스듀서를 추가하는 방법이 사용되었다. 저자는 이러한 기존 방식에서는 전력전달이 연속적이지 않으며, 데이터 전송속도가 높지 않고, 또한 추가적인 장치를 위한 공간, 전력이 필요하다는 제약이 있음을 지적하며, 초음파를 활용하는 무선 전력전송 기술에서도 연속적인 전력전달과 역-산란 데이터 전송이 동시에 가능하게 하는 구조를 제시한다.

가장 핵심적인 원리는 전력 수신부의 트랜스듀서와 연결된 코일로 가는 경로의 스위칭 주기를 조절하여, 하나의 트랜스듀서로 전력을 연속적으로 수신하면서 동시에 수신부의 입력 임피던스를 제어하는 것으로, 논문에서는 이를 통하여 역-산란 방식의 업-링크 데이터 전송을 가능하다는 것을 보여주었다. 이는 인덕터 충전 시간에 의해 트랜스듀서가 바라보는 평균 입력 임피던스가 결정되며, 이 값이 출력 부하의 변화와 상관이 없기 때문으로 이를 통하여 전력 전달과 데이터 전송이 연속적이고 동시에 가능해진다.

또한 저자들은 채널 상태 변화에 따른 통신 신뢰도를 확보하기 위해 업-링크 전송에 다중 변조 방식을 도입하고, 수신된 전력 레벨에 따라 변조 방식과 데이터 전송률을 적응적으로 선택하는 구조를 제안한다. 제안된 시스템은 BPSK, APSK, 4ASK 변조를 지원하며, 채널 상태가 양호한 경우 최대 300 kbps의 업-링크 데이터 전송률을 달성하면서도 비트 오류율(BER) $10^{-6}$ 수준의 신뢰도를 유지함을 실험적으로 보였다. 이러한 접근은 초음파 전력전송 환경에서 발생할 수 있는 채널 간 다양성에 대응하면서도, 기존 시분할 방식 대비 높은 유효 데이터율과 낮은 지연 시간을 가능하게 한다. 실험 결과, 약 5cm 깊이의 체내 환경을 모사한 조건에서 연속적인 전력 회수와 동시에 최대 약 192 μW 수준의 전력을 수신하면서 안정적인 업-링크 데이터 전송을 달성하였으며, 4ASK 변조 기준으로는 비트당 에너지 소모 6.3pJ/bit의 높은 에너지 효율을 기록하였다.

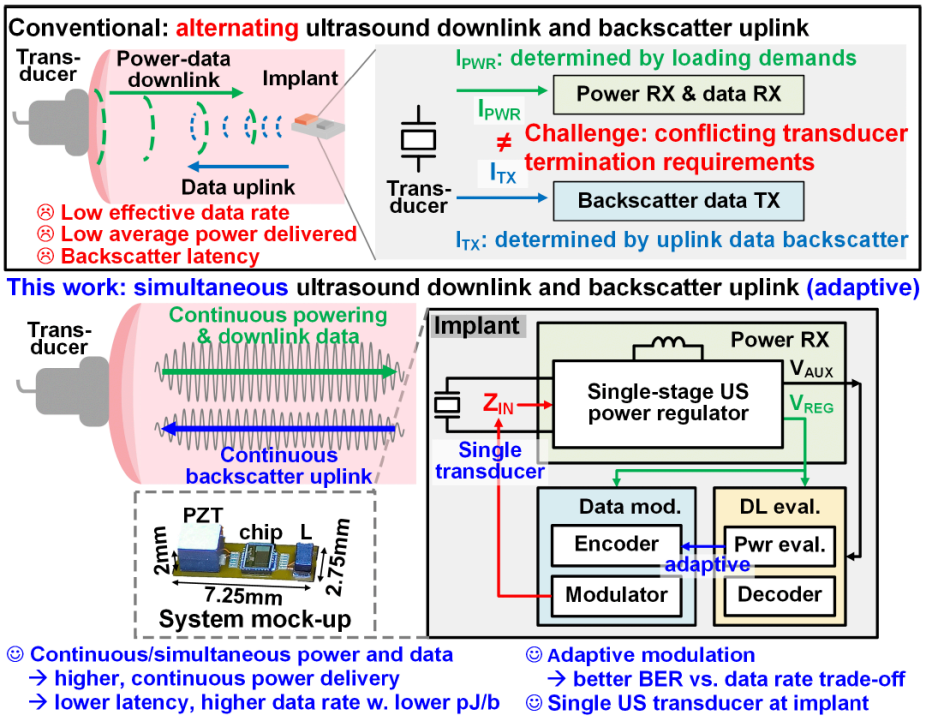



Fig. 1. Motivation and proposed simultaneous ultrasound downlink (power and data) and backscatter communication system.

[그림 1] 동시적 초음파 무선전력전달 및 데이터 통신 기술의 개념도

**#4-4** 논문은 한양대학교, KAIST 및 서울대학교가 참여한 공동연구로, 시간 간섭 자극(Temporal Interference Stimulation, TIS)을 다채널로 확장한 개념인 multi-TIS 의 구동 IC 를 제안한다. TIS 는 서로 근접한 주파수를 갖는 고주파 교류 전류를 중첩하여 심부 조직에서 저주파 포락선(envelope)을 형성함으로써, 표면 자극을 최소화하면서 심부 뇌 자극을 가능하게 하는 비침습적 신경조절 기법이다. 그러나, 기존의 TIS 시스템은 제한된 채널 수와 낮은 선형성, 채널 간 불일치로 인해 자극 공간 선택성이 충분하지 않으며, 보드 레벨 또는 단순 전류 미러 기반 회로에 의존함으로써 확장성과 정밀도에 한계가 있었다.

저자들은 이러한 문제를 해결하기 위해 고전압 증폭기 기반 전류 구동(source-sink) 구조를 채택한 12채널 multi-TIS 드라이버 IC를 제안한다. 제안된 구조는 기존 전류 미러 방식 대비 공정·전압·온도(PVT) 변화에 강인하며, 양방향 전류 구동을 통해 사인파 형태의 자극 신호를 보다 높은 선형도로 생성할 수 있다. 각 채널은 저전압 DAC 에서 생성된 정밀한 사인파 신호를 고전압 영역으로 증폭하여 전극에 인가함으로써, 다양한 부하 조건에서도 일정한 전류 자극을 유지할 수 있도록 설계되었다. 또한, 본 논문에서는 채널 간 이득 및 오프셋 불일치로 인해 발생하는 자극 왜곡을 보정하기 위한 3 단계 보정(calibration) 기법을 제안한다. 측정된 출력 파형을 기준으로 진폭 및 공통모드 오차를 추출하고, 이에 대응하는 보정 코드를 디지털적으로 적용함으로써 채널 간 불일치를 효과적으로 제거한다. 이를 통해 다채널 TIS 구동 시에도 포락선 신호의 왜곡을 최소화하고, 공간적으로 보다 정밀한 간섭 패턴 형성이 가능함을 보였다.

실험 결과, 제안된 IC 는 180 nm BCD 공정으로 제작되었으며, 최대 ±20 V 의 출력 전압 범위에서 채널당 최대 ±2 mA 의 자극 전류를 안정적으로 구동하였다. 15 kΩ 부하 조건에서 1 kHz 자극 시 SFDR 62 dB, SNDR 60 dB, THD 0.119%를 달성하였으며, 보정 이후 진폭 오차는 1% 미만으로 감소하였다. 또한 8 채널 multi-TIS 구성에서 단일 TIS 대비 약 25% 향상된 공간 선택성을 확인함으로써, 제안된 구조가 심부 뇌 영역을 보다 국소적으로 자극할 수 있음을 실험적으로 입증하였다.

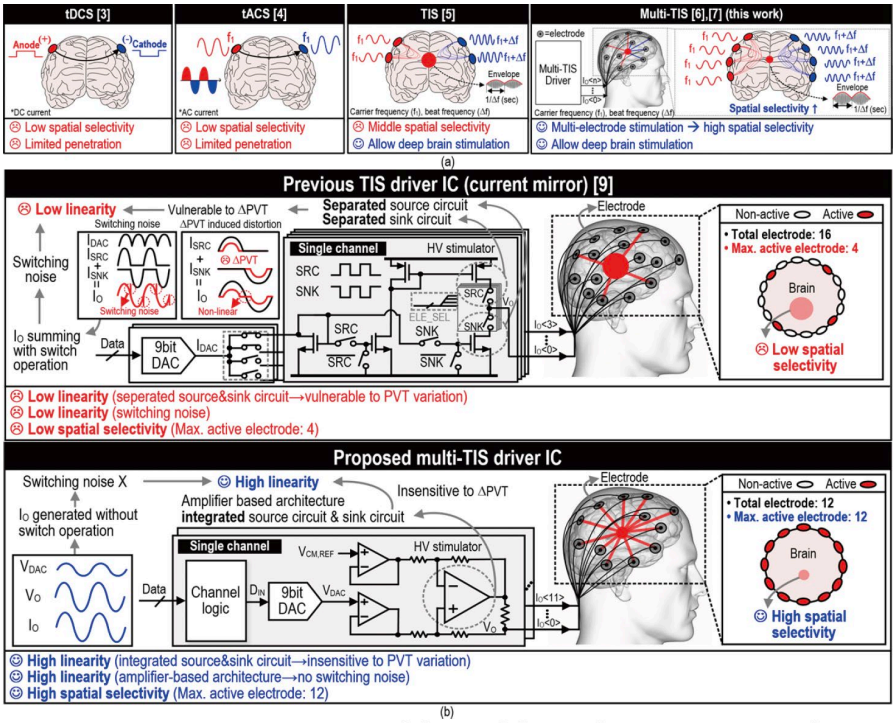



Fig. 1. (a) principles of tDCS, tACS, TIS, and multi-TIS, (b) comparison between previous and proposed TIS driver IC.

[그림 2] 제안된 다중 시간-간섭 자극기(multi-TIS) 구동회로의 개념도

## 저자정보

**석동열 박사과정 대학원생**

- 소속 : 한국과학기술원
- 연구분야 : 바이오메디컬 응용 회로설계(센서 및 신호처리)
- 이메일 : sukd10@kaist.ac.kr

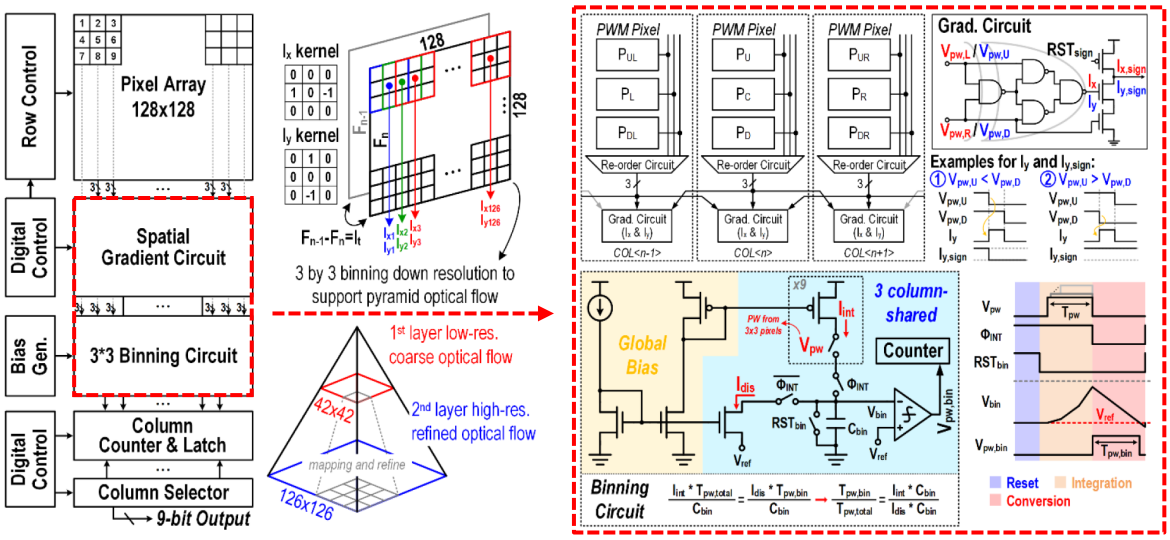
# A-SSCC 2025 Review

울산과학기술원 전기및전자공학부 석박통합과정 홍기업

## Session 16 Visual Interactive System

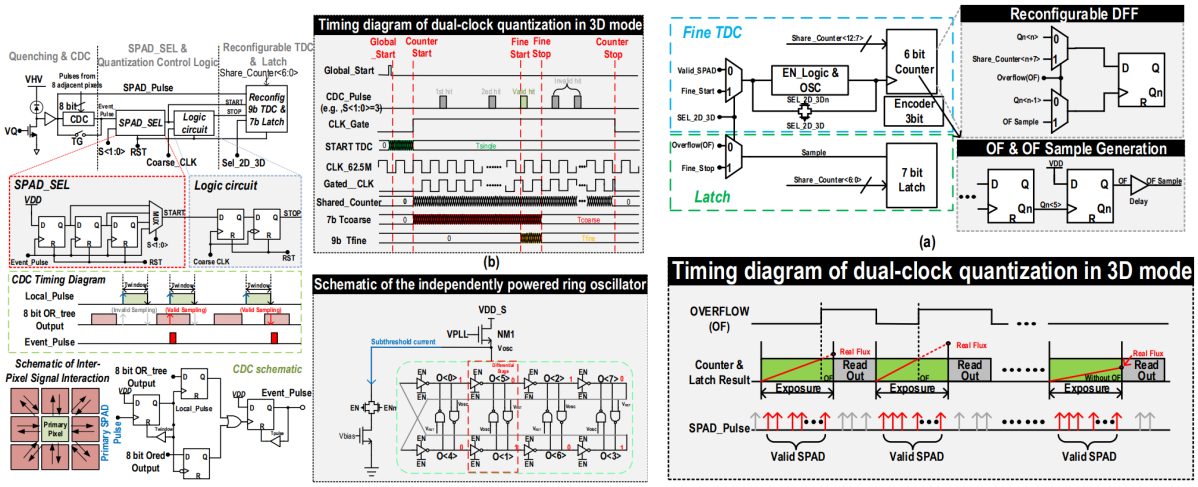
올해 ASSCC session 16에는 3개 image sensor 논문과 1개의 capacitive 터치센서 논문이 소개되었다. 이미지센서 분야에서는 구체적으로 PIS(processing-in-sensor), LiDAR, SPAD-PPD hybrid image sensor가 소개되었으며 현재 가장 활발히 연구되고있는 분야의 논문들이다. 각 논문은 연산 에너지 효율 향상(16.1), background immunity 향상(16.2), 픽셀 구조 소형화 및 dynamic range 향상(16.3)을 목표로한 센서를 제안한다.

**#16-1** National Tsing Hua University에서 발표한 논문으로 128×128 dual-resolution processing-in-sensor (PIS) 구조를 제안하며, optical-flow 기반 motion processing이 핵심이다. 픽셀은 기존에 같은 그룹에서 발표되었던 6T1C PWM 구조로 구성되어 있으며, 단일 프레임 내에서 raw image, temporal gradient, spatial gradient를 동시에 출력할 수 있도록 한다. Spatial gradient 연산은 컬럼 단에 구현된 low-power, area-efficient combinational logic을 통해 수행하며 기존의 OPAMP-based 혹은 adder-based subtractor 대비 높은 에너지 효율을 달성한다. 또한 3×3 binning 회로를 추가하여 저해상도 optical-flow pyramid 입력을 직접 센서에서 생성할 수 있어, 외부 프로세서의 연산 부담을 줄일 수 있다. 따라서 제안된 센서는 9-bit raw image, temporal/spatial gradient, 3×3 binning을 모두 지원하면서도 30fps에서 198.44 μW의 매우 낮은 전력을 달성한다.




[그림 1] 전체 chip architecture(좌)와 제안하는 spatial gradient 및 binning 회로

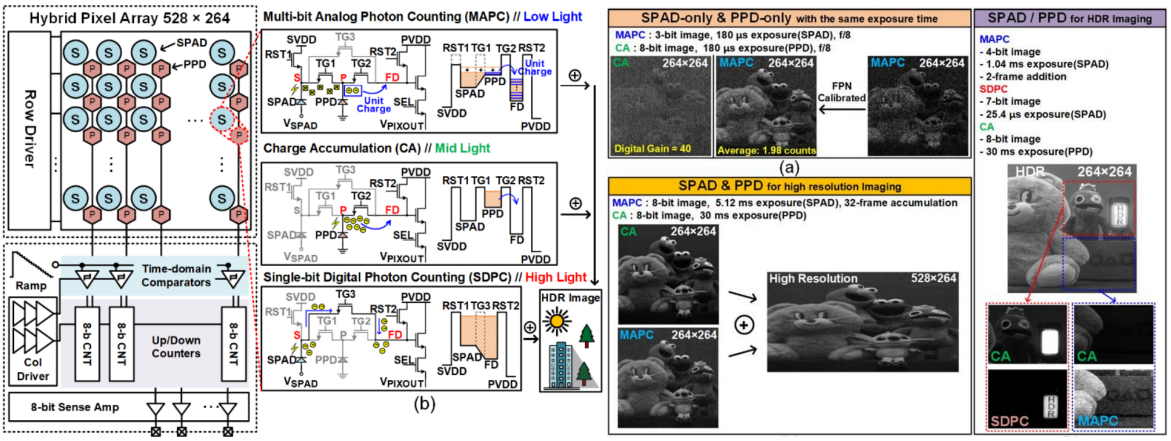
**#16-2**는 Xidian University에서 제안한 flash LiDAR sensor이다. 본 논문은 픽셀 구조를 재구성하여 high dynamic range 2D intensity와 dToF기반 3D depth를 sensing할 수 있는 센서를 디자인하였다. 본 센서는 전체 픽셀 어레이가 동시에 픽셀별 photon의 time of flight를 digitize 하는 flash type dToF sensor이다. 기존 dToF센서의 challenge였던 background에 의한 TDC 및 SPAD pile up issue를 CDC(coincidence detection circuit)와 first-last hit signal selection으로 해결하였다. 또한 3D mode에서 TDC로 활용하던 counter를 2D mode에서는 photon counting에 사용하여 2D sensing이 가능하게 하였다. 더 나아가 2D intensity dynamic range 확장을 위해 TTS(time to saturation) 방식을 이용하여 counter가 정해진 counter depth를 초과하였을 때의 시간을 픽셀에 저장하여 119dB에 달하는 dynamic range를 달성하였다.




[그림 1] 3D mode 동작 timing diagram 및 CDC, first-last hit detection circuit(좌),

HDR 2D mode 동작의 operational concept과 timing diagram(우)

**#16-3**는 Sungkyunkwan University에서 제안한 HDR 2D sensor이다. 본 논문은 인간 눈에 있는 간상세포와 원추세포를 모방한 하이브리드 픽셀 어레이를 디자인하였다. 각 세포는 각각 SPAD 와 PPD로 구현하였으며 광자 감지 민감도가 높은 SPAD와 낮은 PPD소자를 결합함으로써 넓은 조도 상황에서 능동적으로 빛을 감지할 수 있게 하였다. SPAD소자는 높은 광자 감지 민감도로 인해 일반적으로 저조도 센싱에 매우 탁월하나 photon counting counter 구조가 PPD기반 픽셀보다 비교적 복잡하다는 단점이 있다. 본 논문에서는 해당 픽셀 구조를 4개 미만의 transistor로 간소화하였으며 photon counting 필요한 charge transfer를 PPD를 이용하여 효율적으로 구현하였다. 이를 통해 110nm BSI process에서 18.295μm × 14.5 μm라는 비교적 작은 픽셀 크기를 달성할 수 있었으며, SPAD 및 PPD response를 결합함으로써 109dB라는 큰 dynamic range를 달성하였다.



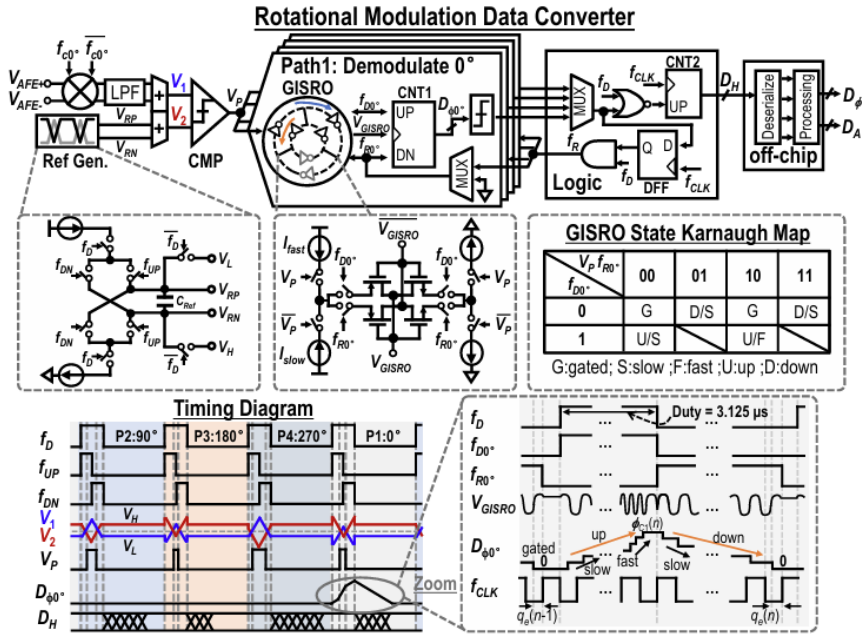

[그림 2] 센서 아키텍처 및 하이브리드 픽셀 operational concept (좌),

본 센서로 획득된 HDR 및 고해상도 이미지 (우)

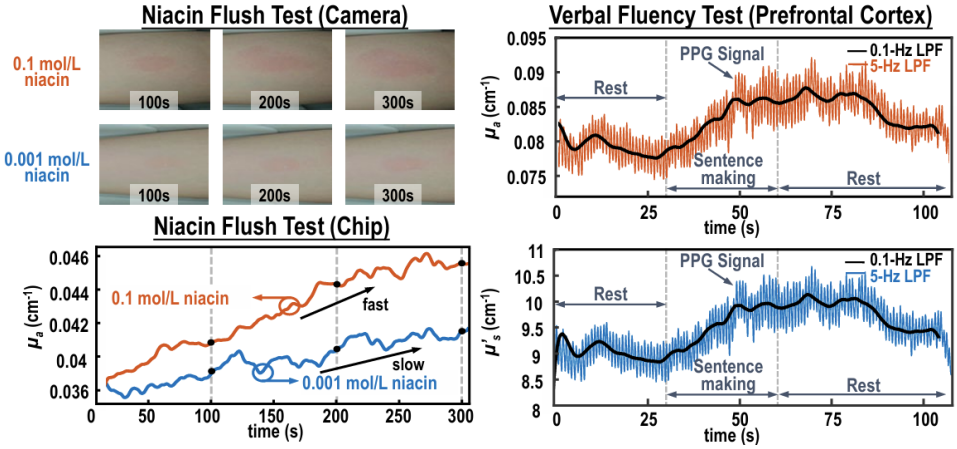
**Session 20 Circuits for Cognitive and Physiological Interfaces**

ASSCC 2025의 Session 20은 'Circuits for Cognitive and Physiological Interfaces' 주제로 총 5편의 논문이 발표되었다. 그 중 Processor, FD-NIRS, Neural recording System-on-Probe(SoP)에 대한 3편의 논문을 살펴보고자 한다.

**#20-2** fNIRS는 웨어러블 구성과 비침습 방식으로 실시간 뇌 활성도를 관찰할 수 있어, 정신의학 연구에서 주목받고 있다. Shanghai Jiao Tong 대학에서 발표한 본 논문은 dynamic TIA 기반 저전력 아키텍처와 APD DC servo loop, 4-phase 샘플링 기법을 통해 1ps 미만의 고해상도 ToF FD-NIRS 시스템을 제안하였다. 제안된 Rotational Modulation Data Converter(RMDC) 구조는 하나의 신호 주기 내 4개 지점 샘플링을 통해 과잉결정 행렬을 형성하여 별도의 보정없이 오프셋에 강인한 신호 복원을 수행하며, Differential difference PWM Frontend를 통해 비교기의 zero-crossing 동작으로 선형성을 확보하였다. 또한, GISRO 기반 위상 도메인 듀얼 슬로프 data converter는 gating시의 freeze 특성을 활용한 1차 noise shaping으로 높은 분해능을 달성한다. 이를 통해 14.6mW의 저전력으로 0.93ps ToF 분해능을 달성하였고, In-vitro 실험에서 흡수·감소 산란계수의 최대 오차 3% 미만으로 기존 기술 대비 우수한 검출 정확도를 입증했다.

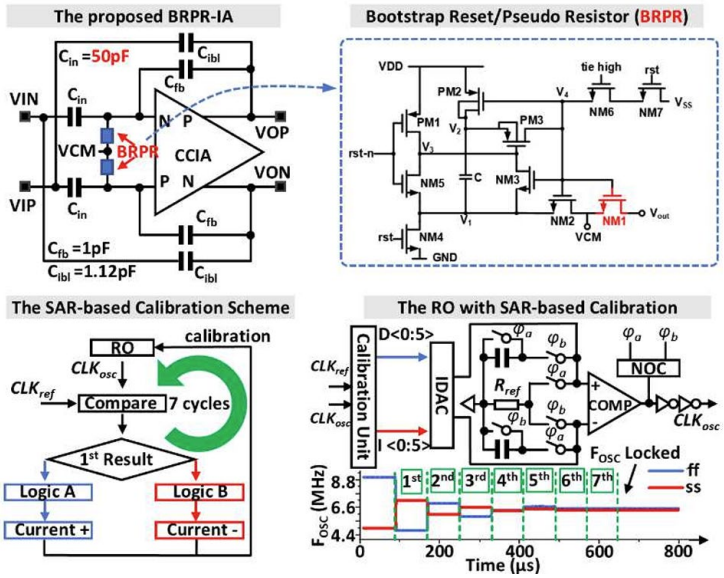



[그림 1] 고해상도 FD-NIRS를 위해 제안된 Data Converter 회로도 및 동작 원리

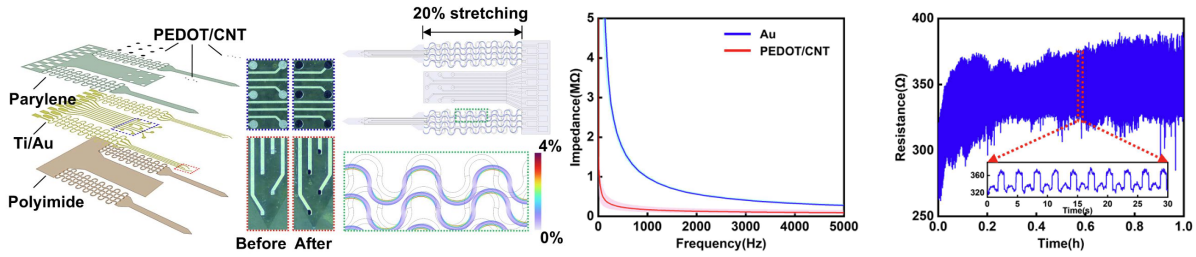



[그림 2] Niacin 홍조 반응(왼쪽)과 언어유창성 검사(오른쪽) 시연 결과

**#20-4** Guangdong Institute of Intelligence Science and Technology에서 제안하는 Neural Miner는 System-on-Probe(SoP) 구조의 16채널 신경 기록 시스템으로, 기록칩과 신축성 프로브을 단일 소형 패키지로 통합하여 다중 뇌 영역을 최소 침습으로 동시에 기록한다. 제안하는 Bootstrap Reset/Pseudo Resistors(BRPR-IA) 구조는 대면적 입력 트랜지스터와 50pF 캡을 통해 1/f 노이즈를 억제하고, 리셋 스위치와 pseudo-저항을 결합한 BRPR 구조를 도입하여 기존의 dead zone 문제를 없애고, TΩ급 고임피던스를 안정적으로 유지하도록 한다. 또한 on-chip SAR 기반 보정이 적용된 closed-loop RO를 통해 IRN 1.18μVrms, CMRR>110dB의 신호 무결성을 확보하였다. PEDOT/CNT 전극(전극 사이즈 : 35x35μm$^2$)과 서펜타인 구조 프로브를 적용하여 낮은 전극 임피던스와 기계적 유연성을 동시에 구현하였다. 측정 결과 0.12mm$^3$/ch의 소형 폼팩터에서 ECoG, LFP, AP를 In-vivo로 기록하여 고집적·저침습 신경 인터페이스로서의 가능성을 입증하였다.




[그림 3] 제안된 BRPR-IA 회로(상단)와 SAR 기반 보정의 RO 회로(하단)




[그림 4] 신축성 프로브 구조(왼쪽)와 프로브의 임피던스 측정결과(오른쪽)

## 저자정보

**홍기업 석박사통합과정 대학원생**

- 소속 : 울산과학기술원
- 연구분야 : 센서디자인, 혼성회로설계
- 이메일 : slsnsep357@unist.ac.kr
- 홈페이지 : https://sites.google.com/view/bias-sogang/home

# A-SSCC 2025 Review

고려대학교 전기전자공학과 박사과정 한창우

## Session 24: Advanced Circuits for Memory and Sensing

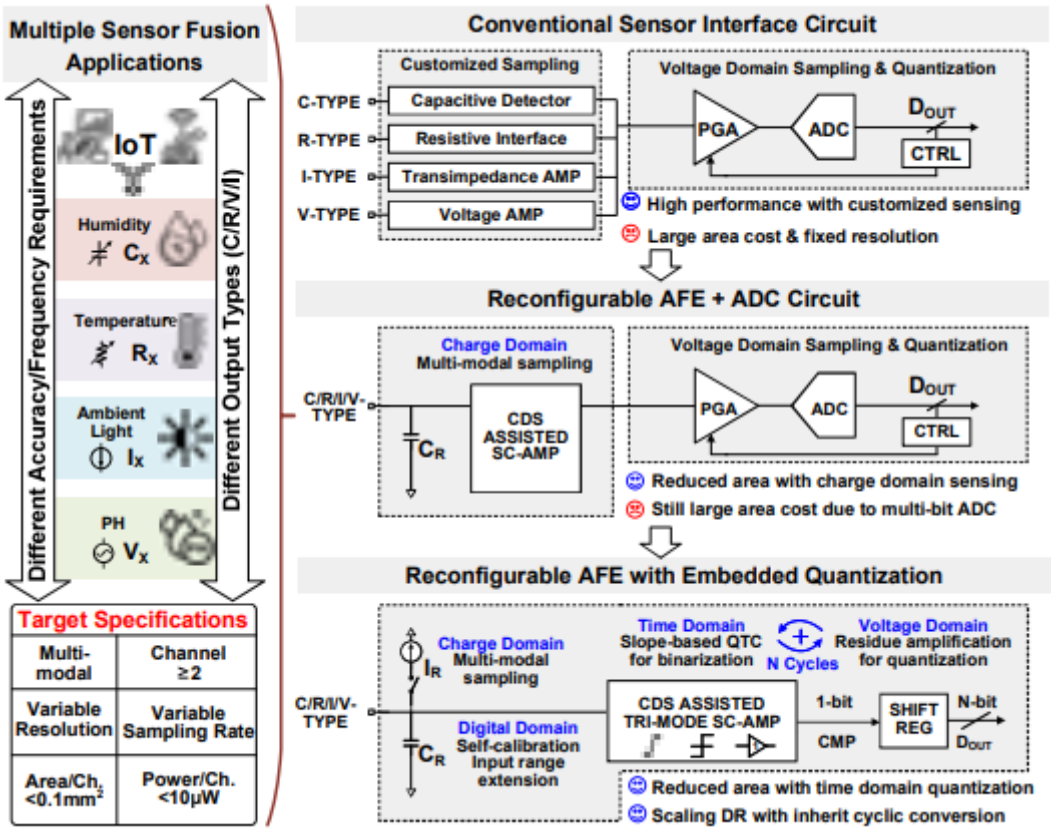
이번 A-SSCC 2025 Session 24에서는 차세대 메모리 및 센싱 회로 기술을 주제로 총 4편의 논문이 발표되었다. 본 세션은 3D DRAM 및 멀티모달 센서 인터페이스를 중심으로, 고집적·저전력 시스템 구현을 위한 아날로그 회로 설계를 다룬다. 특히 공정·전압·온도(PVT) 변화에 강인한 회로 구조를 통해 면적 및 전력 효율을 개선하는 설계 전략이 공통적으로 제시되었다. 본 리뷰에서는 Session 24 중 3D DRAM과 멀티모달 센서 인터페이스 기술을 다룬 두 편의 논문을 중심으로 살펴보고자 한다.

**#24-2** 본 논문은 Shanghai Jiao Tong University에서 발표한 연구로, 전압·전류·저항·커패시턴스(V/I/R/C) 신호를 하나의 회로에서 처리할 수 있는 멀티모달 스위치드-캐패시터(SC) 센서 인터페이스를 제안한다. IoT 및 센서 응용 환경에서는 다양한 물리량을 동시에 측정해야 하나, 기존 센서 인터페이스는 신호 종류별로 개별 아날로그 프론트엔드(AFE)와 ADC를 필요로 하여 면적과 전력 소모가 증가하는 한계가 있었다.

이를 해결하기 위해 본 논문에서는 모든 입력 신호를 전하(charge) 도메인으로 통합하고, 이를 시간 영역 기반 Charge-to-Time Conversion(QTC) 방식으로 양자화하는 새로운 구조를 제안한다. 제안된 회로에서는 V/I/R/C 입력 신호를 단일 커패시터(CR)에 샘플링한 뒤, CDS(Correlated Double Sampling)가 적용된 SC 증폭기를 통해 신호를 증폭하며, 이후 동일한 증폭기를 QTC 동작에 재사용하여 별도의 ADC 없이 디지털 출력을 생성한다.

특히 변환 과정에서 발생하는 residue 전하를 증폭하여 다음 변환 사이클의 입력으로 사용하는 cyclic residue amplification 구조를 도입함으로써, 고해상도 동작 시 변환 사이클 수가 급격히 증가하는 문제를 효과적으로 완화하였다. 또한 시간 영역 양자화의 특성을 활용한 digital self-calibration 기법을 적용하여 공정·전압·온도(PVT) 변화 및 비교기 지연에 따른 오차를 최소화하였다.

측정 결과, 제안된 회로는 55 nm CMOS 공정에서 구현되었으며, 채널당 약 0.015 $mm^2$의 면적, 6.4 μW의 저전력 소모, 그리고 V-mode에서 SNDR 38.3 dB의 성능을 달성하였다. 또한 C-mode에서는 self-calibration을 통해 최대 1 nF까지 입력 범위 확장이 가능함을 실험적으로 검증하였다. 본 논문은 하나의 SC 증폭기를 센싱과 양자화에 동시에 활용함으로써 멀티모달 센서 인터페이스의 집적도와 전력 효율을 크게 향상시킨 설계로, 저전력 IoT 센서 시스템에 적합한 실용적인 솔루션을 제시한다.




[그림 1] 기존 센서 인터페이스 회로와 제안된 재구성형 스위치드-캐패시터 기반 멀티모달 센서 인터페이스 구조

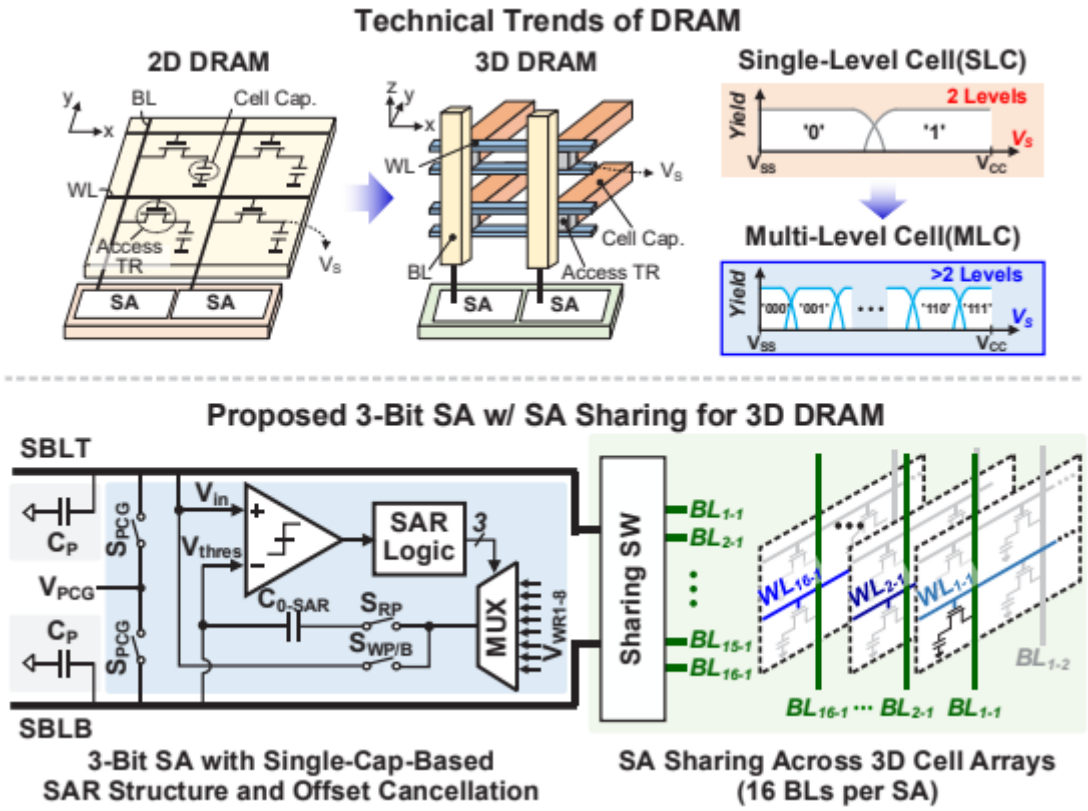
**#24-3** 본 논문은 KAIST에서 발표한 연구로, 3D DRAM 환경에서 멀티레벨 셀(MLC) 구현을 위한 3-bit sense amplifier(SA) 구조를 제안한다. 3D DRAM은 수직 적층 구조를 통해 고집적 메모리 구현이 가능하나, 셀 캐패시턴스 감소와 비트라인 기생 성분 증가로 인해 기존 단일 비트 SA 구조만으로는 안정적인 MLC 동작에 한계가 있다.

이를 해결하기 위해 본 논문에서는 단일 커패시터 기반 SAR ADC 구조를 적용한 3-bit SA를 제안하며, SA 내부에 offset cancellation 기법을 결합하여 공정 및 소자 불균일성에 따른 오프셋 문제를 완화하였다. 제안된 SA는 셀 전압을 직접 비교하는 대신, charge-domain 기반의 비교 및 디지털 변환 방식을 사용함으로써 고정밀 다중 비트 판별이 가능하도록 설계되었다.

또한 3D DRAM의 구조적 특성을 고려하여 SA sharing 기법을 도입함으로써, 하나의 SA가 여러 비트라인을 공유하면서도 안정적인 판독이 가능함을 보였다. 이는 기존 SA sharing 방식에서 문제로 지적되던 순차적 복원에 따른 타이밍 오버헤드를 제거하고, 3D DRAM 환경에 최적화된 구조임을 실험적으로 검증하였다.

측정 결과, 제안된 3-bit SA는 28 nm CMOS 공정에서 구현되었으며, 기존 구조 대비

SA 오프셋 분산을 크게 감소시키고, 낮은 공급 전압 조건에서도 안정적인 MLC 동작이 가능함을 확인하였다. 본 논문은 3D DRAM에서 요구되는 고집적·고정밀 판독을 회로 수준에서 실현한 연구로, 향후 고용량 3D DRAM 시스템에 적용 가능한 실질적인 SA 설계 방향을 제시한다.




[그림 2] 3D DRAM 기술 동향 및 SA 공유 구조를 적용한 3-bit SAR 기반 센스 앰프 구조

## 저자정보

**한창우 박사과정 대학원생**

- 소속 : 고려대학교 전기전자공학과
- 연구분야 : 차세대 반도체 소자 및 회로
- 이메일 : cwoo0105@naver.com
- 홈페이지 : https://sites.google.com/view/kudclab

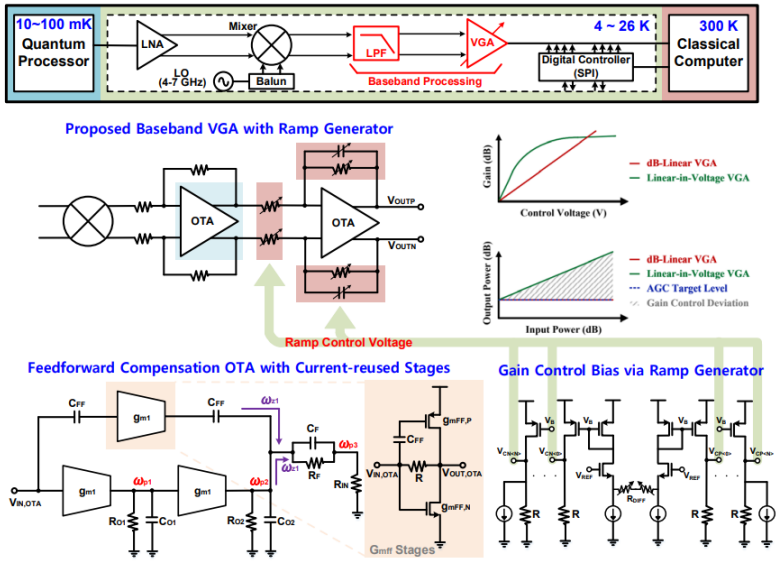
# A-SSCC 2025 Review

포항공과대학교 반도체대학원 박사과정 박은빈

## Session 26 Circuits and Systems for Quantum and Security

2025 IEEE A-SSCC의 Session 26에서는 Cryogenic CMOS 기반 양자 컴퓨팅 인터페이스, 센서 보안 아키텍처, 그리고 고성능 암호 가속기라는 세 가지 핵심 연구 축을 중심으로 총 네 편의 논문이 발표되었다. 최근 반도체와 컴퓨팅 분야에서는 양자 프로세서와의 근거리 통합을 가능하게 하는 Cryo-CMOS 기술, 그리고 양자 이후 시대를 대비한 Post-Quantum Cryptography(PQC) 연산 가속기와 같은 연구가 빠르게 확대되고 있다. Session 26의 논문들은 이러한 기술적 변화 흐름을 반영하며, 미래의 고신뢰·고효율 컴퓨팅 시스템을 구성하기 위한 회로 및 아키텍처 수준의 혁신적인 접근법들을 제시하고 있다.

**#26-1** 논문에서는 4–7 GHz 대역을 지원하며 낮은 노이즈와 높은 선형성을 제공하는 cryogenic CMOS 기반 다중 큐비트 읽기(read-out) 회로를 제안하였다. 양자컴퓨팅 시스템이 대규모로 확장되기 위해서는 수백에서 수천 개의 큐비트를 수십 마이크로초 내에 빠르게 측정해야 하는데, 기존 실온(RT) 기반 구조는 케이블 증가로 인한 열부하, 노이즈 유입, 복잡한 배선 문제 때문에 확장성이 제한적이었다. 본 논문은 이러한 문제를 해결하기 위해 극저온 환경에서 동작하는 단일칩 cryo-CMOS ROIC를 설계하였으며, 이를 통해 다중 큐비트 read-out을 위한 저잡음·고구성 효율의 RF 경로를 제공한다. 제안된 회로는 48 K noise temperature, 52.3 dB SFDR, 4–7 GHz 대역 지원을 달성하여 기존 cryogenic 인터페이스 대비 성능과 실용성을 크게 향상시켰다.
본 논문은 먼저 극저온에서 발생하는 트랜지스터 gm 변화 및 flicker noise 증가 문제를 해결하기 위해 Transformer 기반 Dual Noise-Canceling LNA 구조를 적용하였다. CG/CS 결합 구조를 변압기 기반 피드백 경로와 함께 사용함으로써 cryogenic 환경에서 크게 변동하는 잡음 성분을 효과적으로 상쇄하고, 기존 대비 44% 개선된 NF를 달성하였다. 또한, mixer 단계에서는 cryogenic 환경에서 flicker noise가 더욱 심각해지는 문제를 해결하기 위해 Transformer-coupled Current-Bleeding Mixer를 제안하였다. Current Bleeding(CB) 구조는 flicker noise를 줄이는 데 유리하지만 thermal noise를 증가시키는 단점이 있는데, 본 논문은 transformer를 이용해 이러한 thermal noise 증가를 억제하여 넓은 대역폭과 높은 변환이득을 동시에 확보하였다. 여기에 dB-linear VGA와 253 MHz IF 대역폭을 지원하는 이득 제어 블록을 추가하여 multi-qubit read-out에서 필요한 정밀 이득 조절과 넓은 IF 처리 범위를 실현하였다.



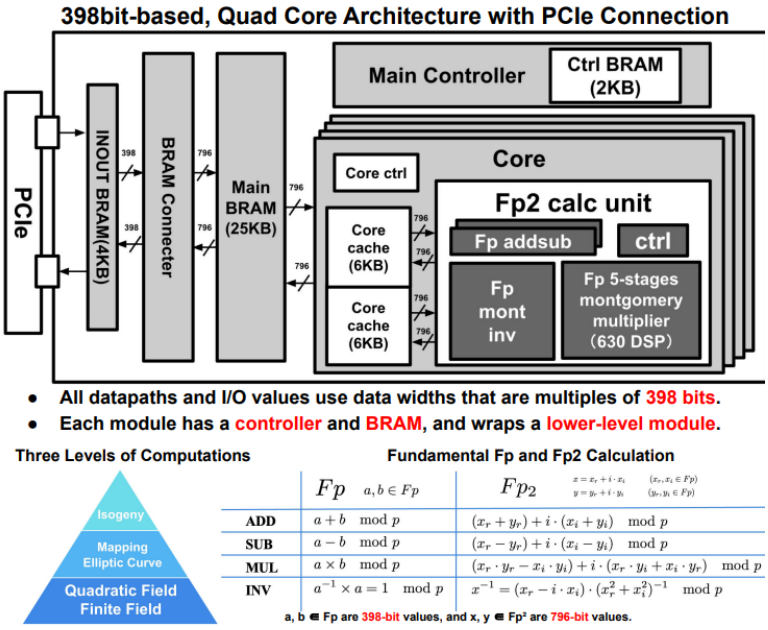

[그림 1] 제안된 cryo-CMOS 기반 multi-qubit read-out RF 수신기 구조

측정 결과, 제안된 cryo-CMOS ROIC는 극저온 환경에서도 안정적으로 동작하며, multi-qubit read-out에 필요한 고선형성·저잡음 성능을 충족함을 입증하였다. 또한 단일칩 구조를 통해 기존 실온 기반 측정 시스템이 요구하던 복잡한 배선을 줄이고, 열부하를 감소시켜 대규모 큐비트 확장성 측면에서도 중요한 장점을 확보하였다. 본 논문은 cryogenic RF 회로의 안정성, 노이즈 특성 개선, 다중 큐비트 지원 능력을 종합적으로 향상시킴으로써 대규모 양자컴퓨팅 시스템을 위한 실용적이고 확장 가능한 read-out 인터페이스 설계 방향을 제시한 연구로 평가될 수 있다.

**#26-3** 논문에서는 398bit (2,2)-isogeny 기반 Post-Quantum Cryptography(PQC)를 고속 처리하기 위한 FPGA 가속기 아키텍처를 제안하며, 기존 소프트웨어 중심 isogeny 연산의 병목을 해결하는 새로운 하드웨어 접근 방식을 제시하였다. Isogeny 기반 암호는 짧은 키 길이라는 장점에도 불구하고 대규모 정수 연산과 복잡한 곡선 변환 과정으로 인해 계산 비용이 매우 크다는 문제가 있었다. 본 논문은 이러한 한계를 극복하기 위해 398bit 데이터 경로 기반의 4-코어 병렬 FPGA 구조를 도입하여 연산 병렬성과 메모리 활용 효율을 극대화하였다. 특히, 기존 구현들이 직면했던 메모리 접근량 폭증과 낮은 스루풋 문제를 해결하고자, 각 연산 단계의 데이터 생애(lifetime)를 분석하여 메모리 공간을 재할당하는 정적 캐시 최적화 기법을 적용하였으며, 이를 통해 기존 대비 약 75%의 BRAM 사용량을 절감하였다.

본 논문은 또한 하드웨어에서 가장 큰 연산 비용을 차지하는 모듈러 곱셈을 가속하기 위해 5-stage 파이프라인 Montgomery multiplier를 설계하였다. 398bit 곱셈을 64bit DSP 블록으로 분해해 매핑함으로써 LUT 기반 구현 대비 최대 8.5배 높은 동작 속도를 확보하였다. Isogeny 체인에서 자주 등장하는 DBL·MAP·FJTJ 연산의 구조적 병렬성을 분석하여 핵심 함수들을 효과적으로 스케줄링한 점도 중요한 기여다. 특히, DBL과 MAP은 상호 독

립적으로 병렬 수행이 가능하며, 이를 최대 4코어 수준으로 동시에 처리해 전체 isogeny 체인의 실행 횟수를 최소화하였다. 이러한 최적화된 경로는 SageMath 기반 소프트웨어 구현 대비 총 연산 비용을 약 9% 추가 절감하는 성능 향상을 가져왔다.



[그림 1] 제안된 4-코어 FPGA 기반 isogeny 연산 가속기 구조

측정 결과, 제안된 FPGA 가속기는 호스트 CPU 대비 최대 87%의 지연 시간 단축(59.6 ms)을 달성했으며, 기존 C/GMP 기반 단일 스레드 구현과 비교해 현저히 빠른 처리 속도를 보여주었다. 특히, FESTA·QFESTA 등 최신 isogeny 기반 암호 스킴에서 반복적으로 등장하는 (2,2)-isogeny 연산을 효율적으로 가속함으로써, 향후 PQC 표준화 과정에서 isogeny 계열 암호의 실용화를 뒷받침할 수 있는 중요한 하드웨어적 가능성을 제시한다. 본 연구는 고비용 정수 연산, 메모리 병목, 연산 스케줄링 문제를 종합적으로 해결함으로써 차세대 보안 시스템을 위한 고성능 PQC 가속기 설계에서 의미 있는 진전을 보여준다.

## 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

# A-SSCC 2025 Review

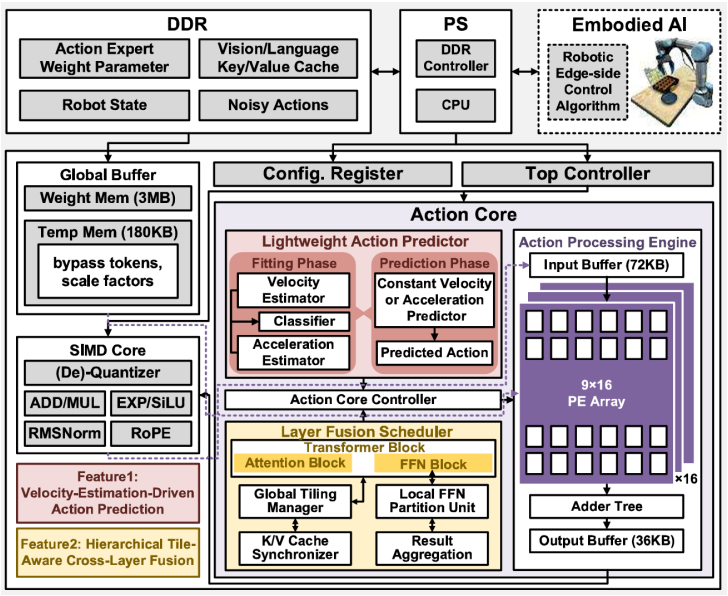
KAIST 전기 및 전자공학부 석사과정 이가은

## Session 7 Application-Driven FPGA Circuits and Systems

이번 A-SSCC 2025의 Session 7에서는 엣지 환경에서의 실시간 인공지능 가속 및 고성능 비디오·로보틱스 처리를 위한 하드웨어 가속기를 주제로 총 4편의 논문이 발표되었다. 본 세션에서는 로봇 행동 계획, 비디오 트랜스포머, 3D 렌더링 등 계산 복잡도와 메모리 요구량이 큰 AI 워크로드를 대상으로, FPGA 기반의 저지연·고효율 가속 구조를 제안한 연구들이 소개되었다. 특히 토큰 수 감소, 예측 기반 연산 생략, 캐시 및 프리패칭을 활용한 메모리 병목 완화 등 알고리즘–하드웨어 협업(Co-design) 기법이 공통적으로 강조되었으며, 엣지 디바이스에서 실시간 성능을 달성하기 위한 다양한 설계 전략이 집중적으로 논의되었다. 본 리뷰에서는 Session 7에 포함된 논문 중에서도, 2편의 논문을 살펴보고자 한다.

**#7-1** 본 논문은 Fudan University 와 ZTE Corporation 에서 발표한 연구로, Embodied Artificial Intelligence(EAI) 환경에서 로봇의 실시간 행동 계획을 수행하기 위한 FPGA 기반 하드웨어 가속기를 제안한다. 최근 로보틱스 분야에서는 시각 정보와 언어 지시를 입력으로 받아 연속적인 행동 시퀀스를 생성하는 Vision-Language-Action(VLA) 모델이 주목받고 있으나, diffusion 또는 transformer 기반 구조로 인해 계산 복잡도와 지연 시간이 커 엣지 환경에서 실시간 제어 주기를 만족시키기 어렵다. 기존 연구들은 병렬화나 모델 경량화를 통해 성능 개선을 시도해왔지만, 행동 계획과 같이 시간적으로 연속된 토큰을 생성하는 문제에서는 예측 가능한 구간에서도 동일한 신경망 연산이 수행되어 불필요한 연산과 전력 소모가 발생한다.

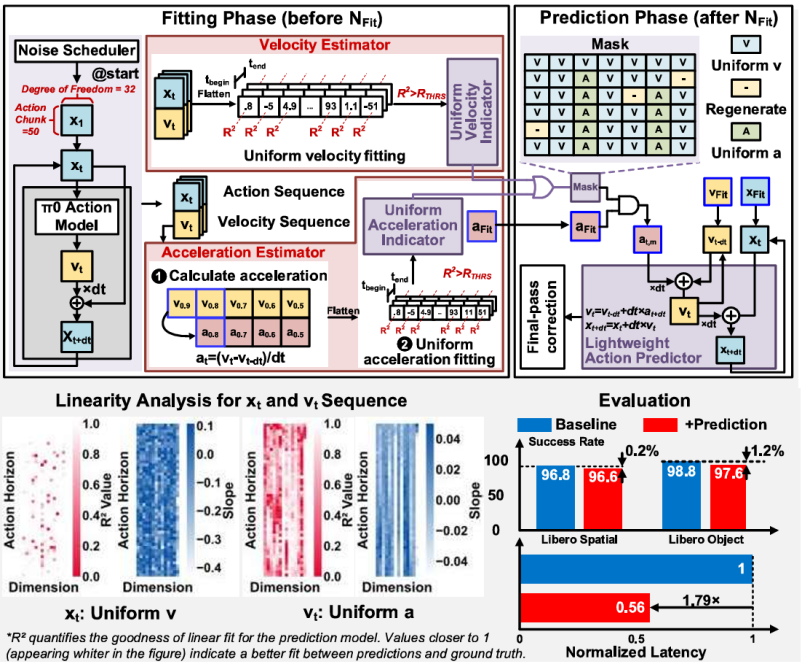
본 논문에서는 이러한 문제를 해결하기 위해 속도 추정 기반 행동 예측(Velocity-Estimation-Driven Behavior Prediction)이라는 새로운 접근 방식을 제안한다. 제안된 방법은 행동 생성 과정을 시간 축에서 분석하여, 모든 토큰을 동일하게 처리하는 대신 토큰의 동역학적 특성에 따라 처리 방식을 달리하는 것을 핵심으로 한다. 구체적으로, 로봇의 최근 행동 이력을 기반으로 각 자유도에 대한 속도 및 가속도를 추정하고, 해당 구간이 선형 운동으로 근사 가능한지를 판단한다. 이 과정에서 결정 계수($R^2$)를 사용하여 행동 변화의 선형성을 정량적으로 평가하며, 일정 임계값 이상일 경우 해당 구간을 예측 가능 구간으로 분류한다.




[그림 1] Overall architecture

예측 가능 구간에서는 신경망 기반 행동 생성 대신, 이전 상태에서 추정된 속도 또는 가속도를 이용한 외삽 연산을 통해 다음 행동을 생성한다. 반면, 장애물 회피나 방향 전환과 같이 비선형성이 큰 구간에서는 기존 diffusion 기반 행동 생성 모델을 그대로 적용한다. 이를 통해 전체 행동 시퀀스 중 상당 부분을 차지하는 선형 구간에서 고비용 신경망 연산을 제거할 수 있으며, 계산 복잡도와 전력 소모를 동시에 줄일 수 있다.

하드웨어 아키텍처 측면에서, 논문에서는 이러한 알고리즘적 아이디어를 반영한 계층적 파이프라인 구조를 FPGA 상에 구현하였다. 전체 시스템은 크게 행동 입력 처리 모듈, 속도 추정 및 선형성 평가 모듈, 예측 기반 행동 생성 모듈, 그리고 신경망 추론 모듈로 구성된다. 속도 추정 모듈은 최근 여러 시점의 행동 데이터를 저장하고, 선형 회귀 연산을 수행하여 속도 벡터를 계산한다. 이후 선형성 평가 결과에 따라 Prediction Phase 또는 Correction Phase 로 동작 경로가 분기된다. Prediction Phase 에서는 간단한 산술 연산을 통해 행동을 외삽하며, 이 과정은 매우 낮은 지연 시간과 전력으로 수행된다. Correction Phase 에서는 일정 주기마다 신경망 추론 결과를 사용하여 예측 결과를 보정함으로써, 장시간 동작 시 발생할 수 있는 누적 오차를 제한한다. 이러한 구조는 실시간성을 유지하면서도 행동 정확도를 안정적으로 확보할 수 있도록 설계되었다. 또한 FPGA 자원 활용 측면에서도, 고비용 연산 유닛의 활성 빈도를 줄여 전체 시스템 효율을 향상시킨다.

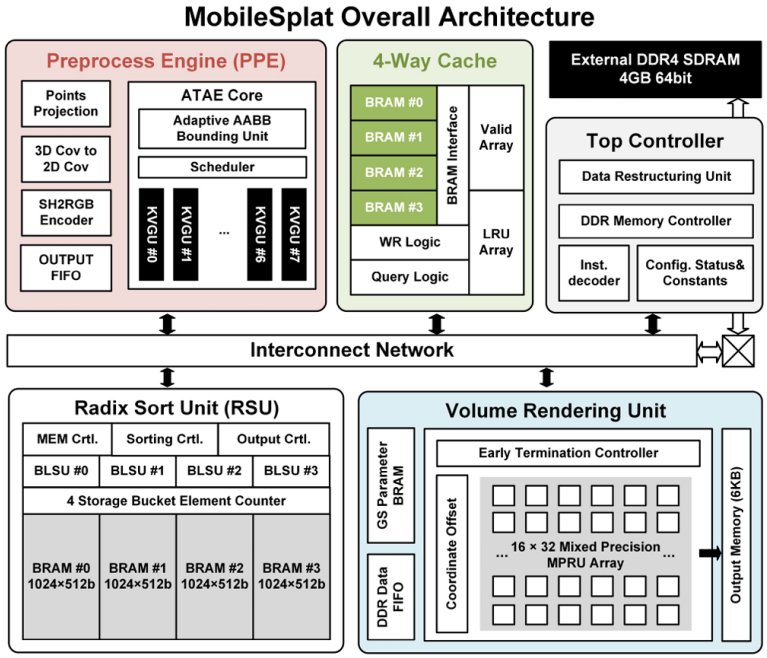



[그림 2] Velocity-estimation-driven behavior prediction mechanism,

제안된 가속기는 실제 로봇 행동 계획 시나리오에서 평가되었으며, 최대 102.5 Hz의 행동 계획 주기를 달성하였다. 또한 예측 기반 연산 생략을 통해 평균 처리 속도가 최대 1.79배 향상되었고, 행동 성공률 역시 기존 방식과 유사한 수준을 유지하였다. 종합적으로 본 논문은 Embodied AI 행동 생성 문제를 단순한 신경망 가속이 아닌, 로봇 동작의 시간적·물리적 특성을 활용한 시스템 수준 최적화 문제로 재정의하였다.

**#7-2** 본 논문은 Tsinghua University Shenzhen 캠퍼스 연구진과 AMD 의 공동 연구로, 3D Gaussian Splatting 을 모바일 및 엣지 환경에서 실시간으로 구현하기 위한 FPGA 기반 렌더링 프로세서 MobileSplat 을 제안한다. 3D Gaussian Splatting 은 NeRF 대비 빠른 렌더링 속도와 높은 시각적 품질로 다양한 3D 비전 응용에서 주목받고 있으나, 다수의 Gaussian primitive 를 타일 단위로 처리해야 하므로 연산량과 메모리 접근 비용이 커 엣지 환경에서의 실시간 구현이 어렵다.

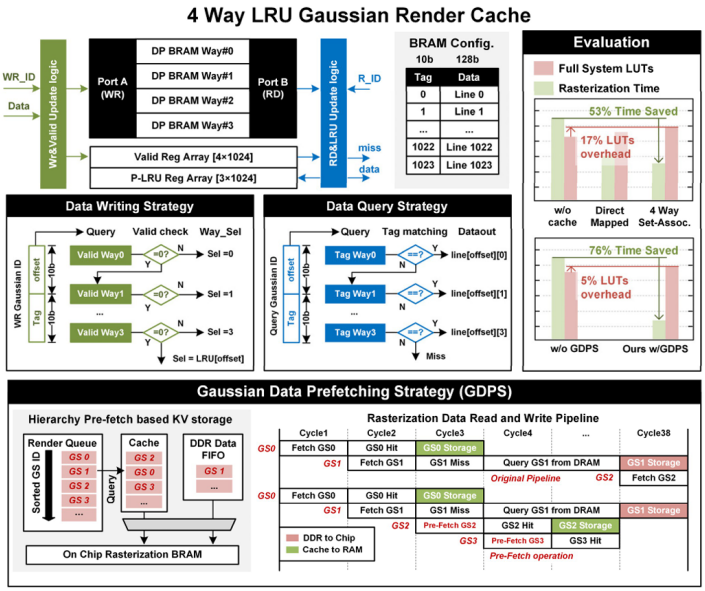
기존 연구들은 GPU 기반 병렬 처리로 성능을 확보해왔으나, 전력 소모와 시스템 비용 측면에서 모바일 환경에 적합하지 않으며, Gaussian 데이터의 반복적 접근으로 인한 메모리 병목을 근본적으로 해결하기 어렵다. 이러한 한계를 해결하기 위해 본 논문은 Gaussian Splatting 의 연산 및 데이터 접근 특성을 분석하고, 불필요한 연산과 메모리 접근을 구조적으로 제거하는 하드웨어 아키텍처를 제안한다.




[그림 3] Overall Architecture of MobileSplat

본 논문에서 제안한 MobileSplat 의 핵심 설계 철학은 연산 유닛의 처리 속도를 높이는 것보다, 렌더링 파이프라인 전반에서 처리해야 할 데이터의 양 자체를 줄이고 데이터 이동을 최소화하는 데 있다. 이를 위해 첫 번째로 도입된 기법은 Pre-culling Architecture 이다. 제안된 Pre-culling 구조는 Gaussian 의 투영 범위를 사전에 분석하여, 특정 타일에 영향을 미치지 않는 Gaussian 을 렌더링 파이프라인 초기에 제거함으로써 불필요한 연산을 방지하며, 평균적으로 처리해야 하는 타일 수를 줄인다. 두 번째 핵심 요소는 Gaussian 데이터의 반복적인 접근 특성을 활용한 온칩 캐시 구조이다. MobileSplat 은 4-way set-associative 구조의 Gaussian Render Cache 를 FPGA 내부 BRAM 으로 구현하여 최근에 사용된 Gaussian 데이터를 저장한다. 세 번째로, 본 논문에서는 Gaussian Splatting 의 시각적 특성을 고려한 Mixed-Precision Computation Path 를 제안한다. 모든 연산을 동일한 고정밀도로 수행하는 대신, 최종 화질에 미치는 영향이 상대적으로 작은 연산은 저정밀도로 처리하고 민감한 연산만 고정밀 경로를 통해 수행함으로써 연산량과 전력 소모를 줄인다. 이러한 접근은 최근 AI 및 그래픽스 하드웨어 설계에서 널리 사용되고 있는 정확도–효율 트레이드오프를 잘 반영한 사례라 할 수 있다. 메모리 병목 완화를 위해 제안된 Gaussian Data Prefetching Strategy(GDPS) 또한 본 논문의 중요한 기여 중 하나이다. GDPS 는 Gaussian 데이터 접근 패턴을 분석하여 향후 필요할 것으로 예상되는 데이터를 미리 DRAM 에서 읽어와 내부

파이프라인에 공급함으로써, 외부 메모리 접근 지연을 내부 연산과 겹쳐 숨기고 전체 렌더링 지연을 효과적으로 완화한다.




[그림 4] 4 Way LRU Gaussian Render Cache

실험 결과, 제안된 MobileSplat 은 다양한 3D 장면에 대해 최대 105 FPS 의 실시간 렌더링 성능을 달성하였다. Pre-culling 기법을 통해 평균 타일 수를 최대 35%까지 감소시켰으며, Gaussian Render Cache 와 GDPS 를 통해 rasterization 단계의 수행 시간을 크게 단축하였다. 또한 제한된 FPGA 자원 내에서 구현이 가능함을 보였고, 전력 효율 측면에서도 모바일 환경에 적합한 수준을 달성하였다. 종합적으로 본 논문은 3D Gaussian Splatting 의 병목을 연산량 증가의 문제로만 보지 않고, 데이터 이동과 메모리 접근 관점에서 재정의하였다는 점에서 의의가 크다.

## 저자정보

**이가은 석사과정 대학원생**

- 소속 : 한국과학기술원 전기 및 전자공학부
- 연구분야 : 디지털 회로 설계
- 이메일 : gelee@ics.kaist.ac.kr
- 홈페이지 : https://idec.or.k