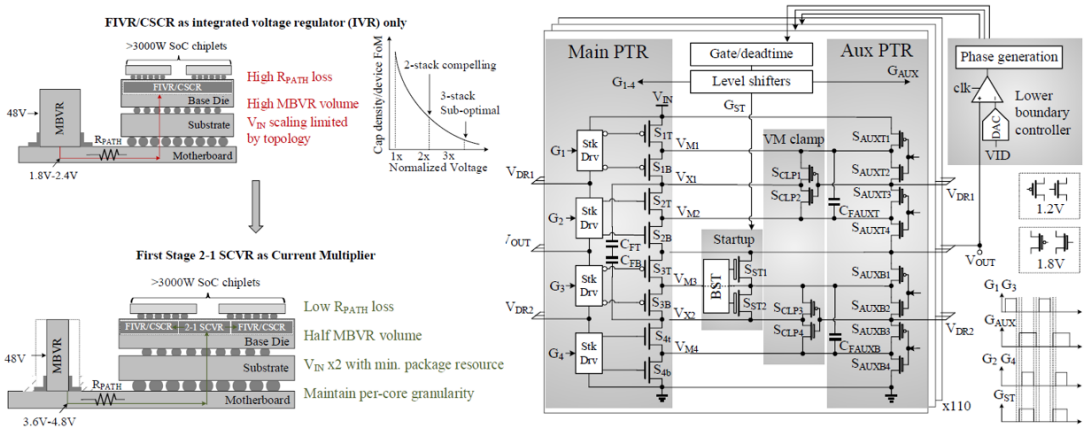
2026 VLSI Review

성균관대학교 반도체융합공학과 석사과정 이승언

Session 2 High-Voltage-Ratio and Scalable Power Delivery

이번 VLSI 2026의 Session C2는 데이터센터 및 xPU 시스템의 전원 공급을 주제로 한 고집적 DC-DC 컨버터 설계를 다루었다. AI 프로세서의 소비 전력이 급격히 증가하면서, 낮은 출력 전압에서 대전류를 공급할 때 전달 경로에서 발생하는 손실과 패키지 자원 부담을 줄이는 것이 공통된 과제로 부각되고 있다. 이를 위해 전력을 높은 전압으로 소자 근처까지 전달하는 vertical power delivery, 그리고 여러 컨버터와 IC를 확장해 대전류를 감당하는 scalability가 이번 세션을 관통하는 설계 화두로 자리잡고 있음을 잘 보여준다.

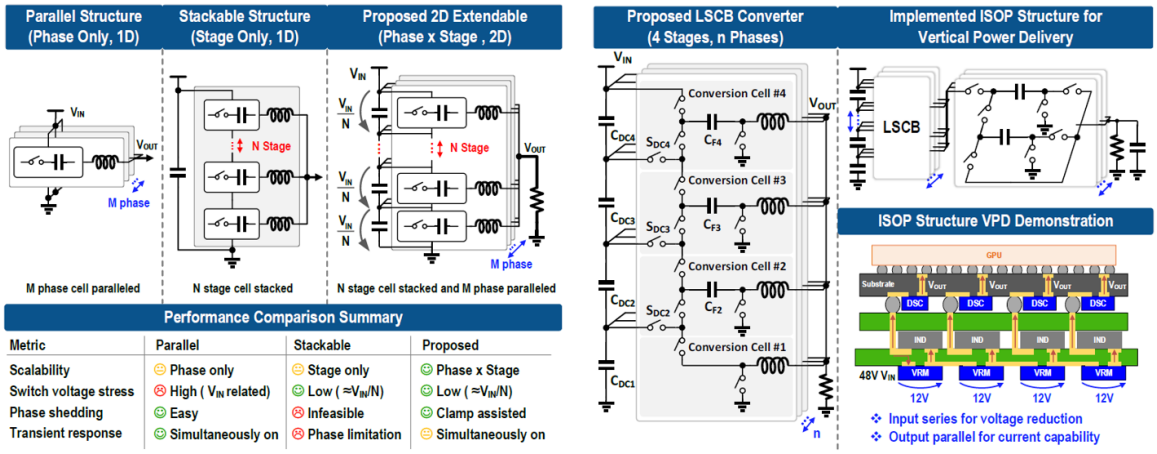
#C2-1 은 미국 Intel에서 발표한 논문으로, SoC 전원 트리에 monolithic 2-1 SCVR을 first-stage current multiplier로 삽입하는 구조를 제안한다. 기존에는 MBVR 출력(1.8~2.4V)이 곧바로 FIVR/CSCR로 전달되었으나, 그 사이에 2-1 SCVR을 두어 MBVR 출력 전압을 3.6~4.8V로 두 배 높일 수 있게 한 것이 핵심이다. 이를 통해 동일 전력에서 MBVR이 공급하는 전류가 절반으로 줄어 R_{PATH} 손실과 MBVR 부피, 패키지 자원 부담을 함께 낮춘다. 저전압 소자만 지원하는 공정 제약 하에서 고입력전압을 감당하기 위해 stacked-device 구조를 채택하며, 이때 발생하는 drive voltage 생성과 stacked-device 구동 손실, startup dead zone 문제를 auxiliary power train과 bootstrapped startup cell로 해결한다. 16nm급 공정에서 110개 모듈로 구현되어 3.6V 입력에서 94.8% peak efficiency를, 4.8V 입력 33A 부하에서 20W/mm² 전력 밀도와 84.4% 효율을 달성하며, monolithic fixed-ratio SCVR 중 가장 높은 peak efficiency와 전력 밀도를 보여준다.



[그림 1] 기존 구조와 제안하는 first-stage 2-1 SCVR current multiplier 구조 및 시스템 구성

#C2-2는 중국 Tsinghua 및 마카오 대학에서 발표한 논문으로, 48V를 1V로 변환하는 ladder-series-capacitor buck(LSCB) 컨버터를 제안한다. 기존 병렬 구조는 입력 전압이 높을수록 스위치 전압 스트레스가 커지고, 스택형 SC 구조는 conversion ratio 범위가 좁으며 idle 상태에서 일부 phase를 끌 때 flying-capacitor 전압이 유지되지 않아 phase shedding이 어렵다는 한계가 있다. 제안 구조는 phase 병렬과 stage 스택을 하나의 구조에서 동시에 지원해, 스위치 전압 스트레스를 VIN/N으로 낮추면서 출력 전류를 확장할 수 있는 것이 핵심이다. 이때 idle 상태에서는 DC 커패시터가 각 flying-capacitor 전압을 클램핑해 phase shedding과 빠른 과도 응답을 가능하게 한다.

나아가 이 구조를 단일 모듈로 쓰지 않고 3:1 Dickson SC와 cascade한 ISOP 구성으로 확장해 vertical power delivery를 구현하였다. LSCB는 180nm BCD, Dickson SC는 65nm 공정으로 구현되었으며, 단상 LSCB에서 91.2% peak efficiency를, 2상 LSCB의 20A 출력에서 87.6% 효율을, ISOP 시스템에서 최대 25A 출력과 2132W/in³ 전력 밀도를 달성한다.

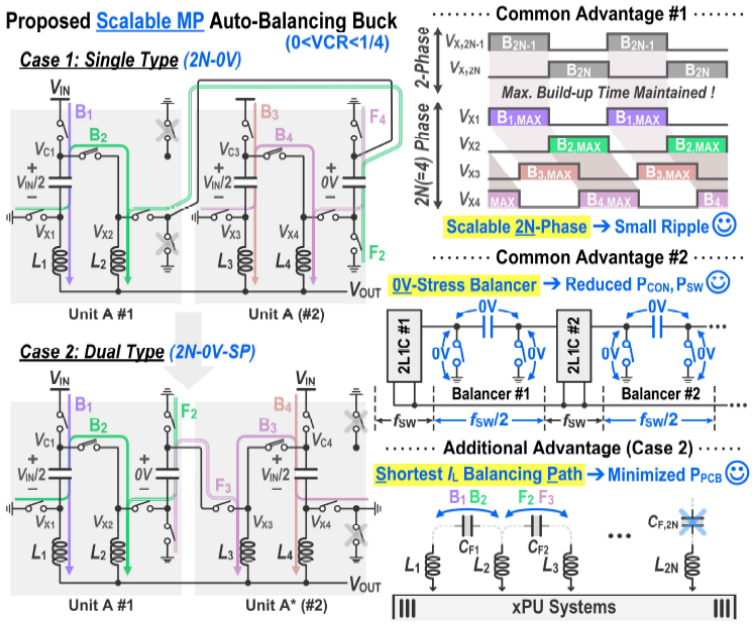


[그림 2] 제안한 2D 확장형 LSCB 컨버터와 vertical power delivery를 위한 ISOP 구조

#C2-3 은 대한민국 서강대학교에서 발표한 논문으로, xPU 시스템의 병렬 IC 확장을 위한 scalable 2N-phase buck 컨버터를 제안한다. 커패시터 기반 IL balancing을 사용하는 기존 multi-phase buck은 인덕터를 추가할수록 VCR이 $0<VCR<1/N^2$로 제약되어 phase 확장이 어렵고, 이를 개선한 이전 scalable auto-balancing buck은 VCR 제약을 피하려 2-phase로 제한되어 출력 리플이 크며 first와 last 인덕터를 잇는 end-to-end balancing 경로 탓에 PCB 트레이스 길이와 도통 손실이 증가하는 한계가 있다.

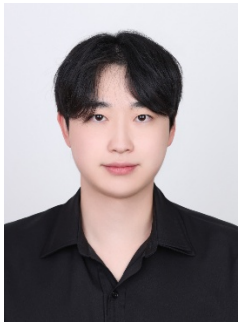
제안하는 2N-0V-SP는 인접한 2L1C 단위를 커패시터 하나와 스위치 두 개로 연결하는 0V-stress balancer를 도입하여, 기존이 build-up duty 구간에서만 IL을 balancing하던 것과 달리 전체 주기에 걸쳐 balancing함으로써 인덕터를 늘려도 VCR 범위를 $0<VCR<1/4$로 유지한 채 phase를 2N으로 확장하고 interleaving으로 출력 리플을 낮춘다. 또한 좌우 대칭 단위를 교대 배치해 인접 인덕터를 직선으로 연결하는 최단 balancing 경로를 구성함으로써 PCB 트레이스 길이와 도통 손실을 추가로 줄인다. 4-phase 프로토타입에서

1.0V 89.2%, 1.2V 90.4%의 peak efficiency를 달성하여 인접 IL balancing과 phase 확장성을 동시에 자동으로 만족하는 구조임을 보여준다.



[그림 3] 제안한 2N-0V-SP buck 컨버터의 구조와 기존 구조 대비 장점

저자정보



이승언 석사과정 대학원생

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

VLSI 2026 Review

고려대학교 전기전자공학과 박사과정 이윤호

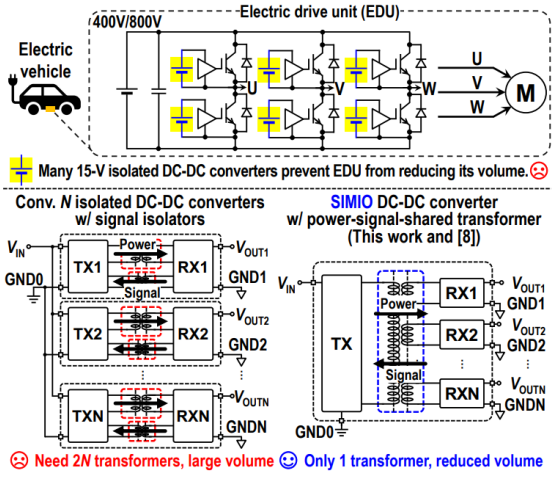
Session 5 Isolated and Multi-Domain Power Conversion

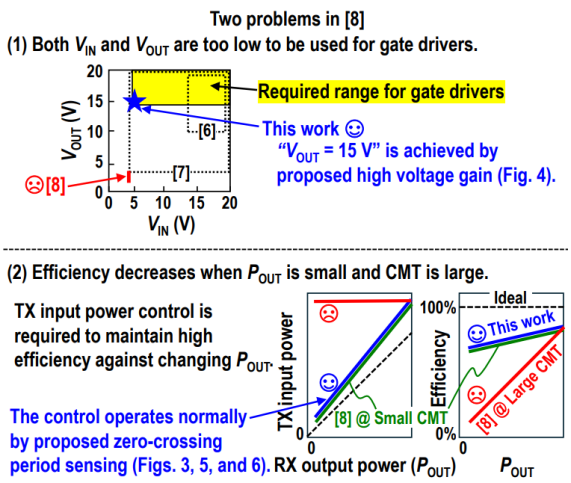
이번 VLSI 2026 Session 5에서는 전기차 구동 시스템과 고성능 컴퓨팅 칩에 요구되는 절연 전원 및 multi-domain power conversion 기술을 다룬 총 4편의 논문이 발표되었다. 본 리뷰에서는 이 가운데 하나의 PCB transformer를 통해 전력과 제어 신호 정보를 함께 전달하는 gate-driver용 SIMIO converter (5-1)와, stacked voltage domain 사이의 부하전류 불균형을 보상하는 single-inductor multiple-output (SIMO) 기반 VSIMO current-mismatch compensator (5-4)를 중심으로 살펴보고자 한다.

#5-1 본 논문은 전기차 EDU(Electric Drive Unit)의 게이트 드라이버용 절연 전원을 공급하는 과정에서 발생하는 컨버터의 부피 증가와 CMT(Common-Mode Transient)에 의한 효율 저하 문제를 해결하기 위해, 하나의 PCB transformer를 통해 전력과 제어 신호를 함께 전달하는 단일입력에 6개의 절연 출력을 갖는 DC-DC converter를 제안하였다. 아래 그림 1-(a)와 같이 전기차의 3상 인버터에는 일반적으로 6개 이상의 power switch가 사용되며, 각각의 게이트 드라이버를 구동하기 위해 15V 수준의 절연 전원이 필요하다.

기존 구조에서는 각 gate driver마다 별도의 isolated DC-DC converter가 요구하며, 전력 전달용 transformer와 출력 제어를 위한 신호를 전달하기 위해 별도의 signal isolator가 각각 필요하다. 따라서 N개의 절연 출력을 구현하기 위해 큰 부피의 transformer 2N개가 요구되므로 EDU의 전체 부피가 증가한다는 한계가 있었다. 이러한 문제를 개선하기 위해 기존 연구에서는 하나의 PCB transformer를 전력 및 신호 전달에 함께 사용하는 SIMIO(Single-Input Multiple-Isolated-Output) 구조를 제안하여, 별도의 signal isolator 없이 하나의 transformer만으로 다수의 절연 출력을 생성할 수 있었다.

그러나 그림 1-(b)와 같이 기존 SIMIO 구조는 3.3V의 입력 전압으로부터 1.8V~3.3V의 낮은 전압만 출력가능하기에, 최소 5V의 입력과 15V의 출력이 요구되는 gate-driver 전원에는 적용할 수 없다는 한계가 있다. 또한 TX쪽 capacitor 양단의 최대 전압을 감지하여 TX의 입력 전력을 제어하기 때문에, 큰 CMT가 transformer의 기생 커패시턴스를 통해 유입되면 신호가 왜곡되어 출력 부하를 잘못 감지할 수 있고, 효율이 크게 저하되는 문제가 발생한다.

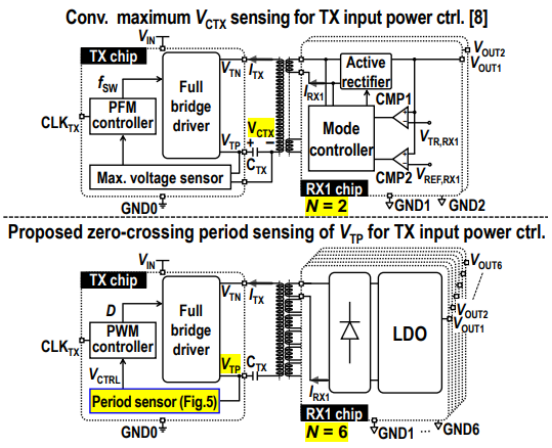


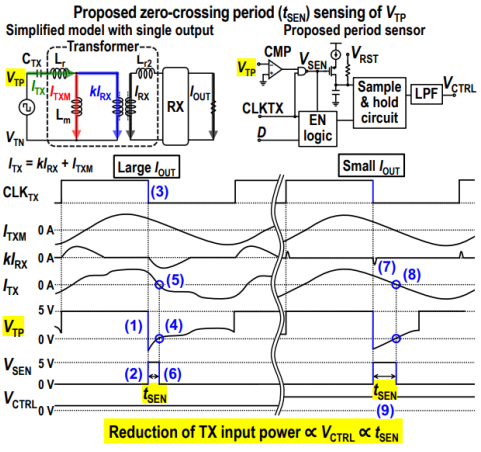


[그림 1-(a)] [그림 1-(b)]

[그림 1] 기존 N isolated DC-DC converter (a)와 그에 따른 구조적 동작 한계 사항 (b)

본 논문은 이를 해결하기 위해 아래 그림 2-(a)와 같이 높은 전압 이득을 갖는 fixed-frequency PWM 방식을 적용하고, CMT에 강인한 zero-crossing-period sensing을 이용하여 별도의 signal isolator 없이 5V 입력에서 6개의 15V 절연 출력을 생성한다. 그림 2-(b)는 출력 부하를 별도의 signal isolator 없이 TX 측에서 감지하는 원리를 보여준다. 절연형 converter에서는 RX측 전류가 transformer를 통해 TX 측에 KI_{RX}로 반영된다. 따라서 TX측 전류 I_{TX}는 자화전류 I_{TXM}과 KI_{RX}의 합으로 결정되며, 출력전류가 변하면 I_{TX}와 TX 측 센싱 노드 V_{TP}의 zero-crossing 시점도 함께 변한다. 출력전류가 클 때는 V_{TP}가 빠르게 0V에 도달하므로 sensing period인 T_{SEN}이 짧아진다. 반대로 출력전류가 작을 때는 zero-crossing이 지연되어 T_{SEN}이 길어지는데, Period sensor는 이러한 변화를 아날로그 신호인 V_{CTRL}로 변환하여, 중부하에서는 충분한 전력을 전달하고 경부하에서는 불필요한 TX 입력 전력을 감소시키도록 PWM을 제어한다.

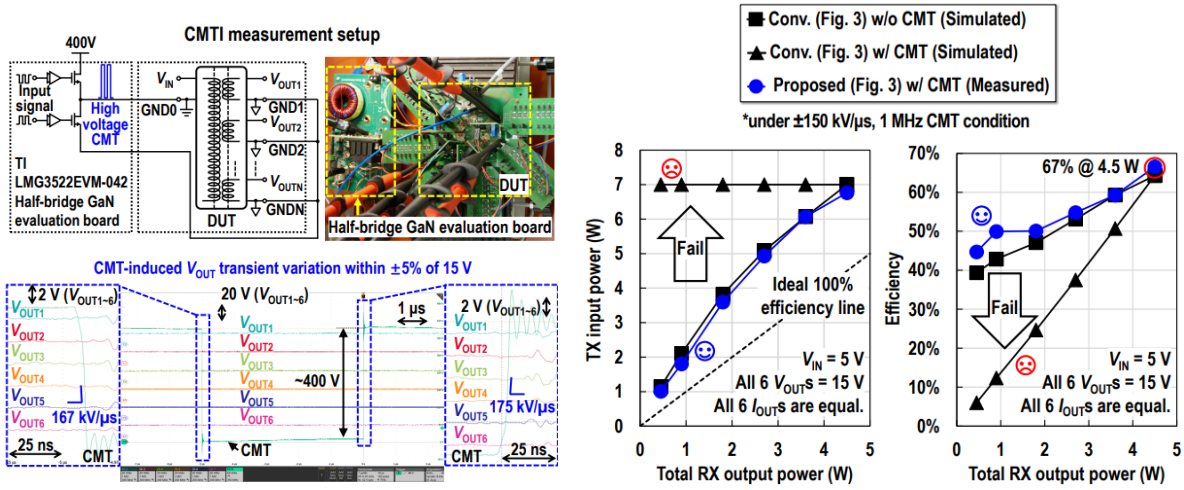




[그림 2-(a)] [그림 2-(b)]

[그림 2] 기존(위) 및 본 논문의 converter 구조(아래) (a) 및 제안하는 zero-crossing sensing 동작 (b)

측정 결과, 아래 그림 3-(a)와 같이 400V의 CMT 조건에서 6개의 출력이 모두 ±5% 이내로 유지되었으며, +175/−167kV/μs의 CMTI를 달성하였다. 또한 그림 3-(b)와 같이 ±150kV/μs, 1MHz의 반복적인 CMT 조건에서도 TX측 전력이 RX측 전력에 따라 조절되었으며, 5V 입력에서 6개의 15V 절연 출력과 4.5W의 최대 출력전력, 67%의 최대 효율을 달성하였다. 아울러 3kV 이상의 절연 성능과 출력당 314mm³의 부피를 구현하여, 별도의 signal isolator 없이 높은 CMT 내성과 다중 절연 출력을 제공하는 전기차 gate-driver용 전원 구조의 유효성을 입증하였다.



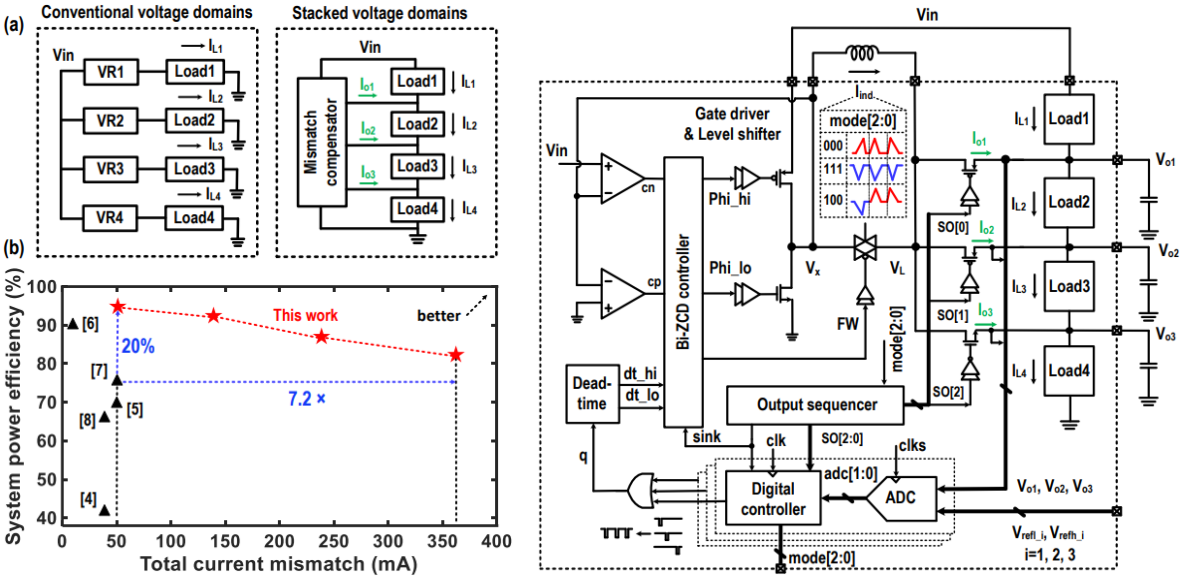
[그림 3-(a)] [그림 3-(b)]

[그림 3] CMTI 측정 셋업 및 결과 (a), 기존 및 제안하는 구조의 RX/TX 측 파워 및 효율 그래프 (b).

#5-4 본 논문은 AI accelerator 및 고성능 컴퓨팅 칩에 사용되는 다중 voltage domain의 전력 공급 과정에서 발생하는 전력변환 손실과, stacked voltage domain 간의 부하전류 불균형 문제를 해결하기 위해 하나의 inductor를 여러 출력이 공유하는 VSIMO 기반 current-mismatch compensator를 제안하였다. 아래 그림 4-(a)의 왼쪽과 같이 기존 구조에서는 각 voltage domain이 개별 voltage regulator(VR)를 통해 입력전압을 필요한 전압으로 변환하므로, 모든 부하전력이 regulator를 거치면서 추가적인 전력변환 손실이 발생한다. 이러한 손실을 줄이기 위해 기존 연구에서는 그림 4-(a)의 오른쪽과 같이 여러 voltage domain을 직렬로 연결하는 stacked voltage-domain 구조를 제안하였다. 이 구조에서는 입력전압을 여러 domain이 나누어 사용하므로 각 domain을 위한 별도의 step-down converter가 필요하지 않으며, 부하전류가 모두 동일한 이상적인 조건에서는 100%에 가까운 system efficiency를 달성할 수 있다. 또한 voltage regulator(VR)가 전체 부하전력을 처리하지 않고 domain 간의 전류 차이만 보상하므로, 기존 병렬형 전원 구조보다

전력변환 손실을 크게 줄일 수 있다.

그러나 직렬로 연결된 각 domain의 부하전류가 서로 다르면 중간 voltage rail에 전류가 남거나 부족해져 각 domain의 전압이 변하게 된다. 따라서 각 중간 rail에서 남는 전류는 흡수하고 부족한 전류는 공급하는 current-mismatch compensator가 필요하다. 기존 LDO 기반 compensator는 50%의 낮은 시스템 효율을 보이며, switched-capacitor 방식 또한 그림 4-(a)의 아래와 같이 처리 가능한 total current mismatch가 20~50mA, 시스템 효율 역시 약 65~75%로 제한되는 한계가 있었다.



[그림 4-(a)] [그림 4-(b)]

[그림 4] 기존 및 stacked voltage-domain 구조와 compensator 성능 비교 (위) 제안하는 VSIMO current-mismatch compensator의 전체 구조(b).

본 논문은 이를 해결하기 그림 4-(b)의 VSIMO와 같이 인덕터 기반의 4개의 stacked voltage domain 구조를 제안하였다. VSIMO의 핵심은 각 출력을 필요에 따라 supply mode 또는 sink mode로 동작 시킬 수 있다는 점이다. 어떤 rail에 전류가 부족한 경우에는 일반적인 인덕터 전류를 생성하여 전류를 공급하고, 반대로 전류가 남는 rail에서는 인덕터 전류를 입력 방향으로 형성하여 전류를 회수한다. 이때, 회수된 에너지를 인덕터에 저장한 후 다른 출력에 전류를 공급하는 데 이용한다. 따라서 남는 전력을 열로 소모하는 LDO 방식과 달리 domain 사이에서 에너지를 재사용할 수 있어 부하 불균형이 큰 조건에서도 높은 효율을 유지할 수 있다.

아래 그림 5-(a)는 각 출력의 전압과 보상전류를 결정하는 1.5-bit ADC 및 digital controller의 구조를 보여준다. 1.5-bit ADC를 이용하여 출력전압이 기준 범위보다 높은지, 낮은지 판단하여, encoder를 통해 에러 코드를 출력한다. 이 후 PID controller로 duty code를 생성 후, 1.024GHz로 동작하는 10-bit counter의 값과 비교하여 각 출력을 제어한

다. 동일한 voltage 에러라도 supply mode와 sink mode에서는 필요한 전류 방향이 서로 반대이므로 feedback의 부호도 함께 변경되어야 하는데, Supply mode에서는 출력전압이 낮아질수록 양의 보상전류가 증가하도록, sink mode에서는 출력전압이 높아질수록 음의 보상전류를 생성하도록 encoder가 ADC 에러의 부호를 반전한다.

그림 5-(b)는 세 출력의 supply 및 sink mode 조합에 따른 동작을 보여준다. Mode=000에서는 세 출력이 모두 supply mode로 동작하여 모든 구간에서 양의 inductor current가 흐르며, mode=111에서는 세 출력이 모두 sink mode로 동작하여 음의 inductor current가 형성된다. 또한 bidirectional zero-current detector(Bi-ZCD)는 supply mode에서 inductor current가 음의 방향으로 넘어가거나, sink mode에서 양의 방향으로 넘어가는 것을 감지하여 input power FET을 차단한다. 전류가 0A에 도달한 이후에는 freewheeling transistor를 켜서 switching node의 불필요한 진동을 억제한다.

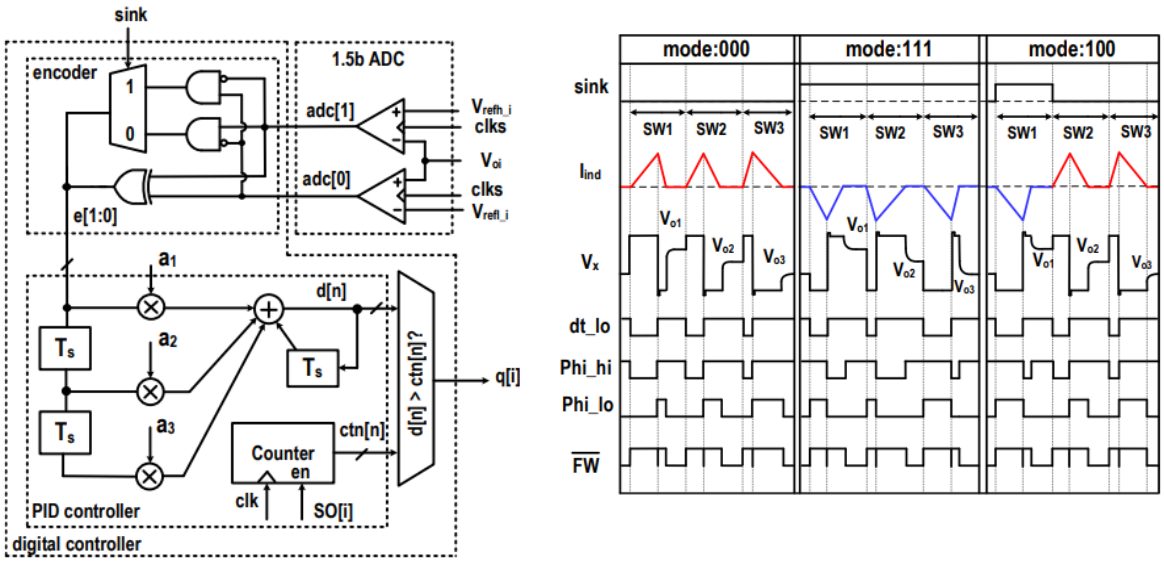


그림 5-(a)] [그림 5-(b)]

[그림 5] 본 논문의 Bi-ZCD controller 스키메틱 (a) 및 각 mode 별 VSIMO의 동작에 대한 타이밍 다이어그램(b).

측정 결과, 65nm CMOS 공정으로 제작된 VSIMO는 1.6~2.4V의 입력전압 범위에서 4개의 stacked voltage domain을 구동하였으며, 최대 432mW의 mismatch compensation power를 한다. 부하전류가 균형을 이루는 조건에서는 최대 98.5%의 시스템 효율을 달성하였고, 50~360mA의 total current mismatch 조건에서도 82~96%의 효율을 유지하였다. 특히 인접한 domain 사이에서 최대 120mA, 세 intermediate rail의 합으로는 최대 360mA의 mismatch를 보상하면서도 82% 이상의 효율을 달성하였다. 이를 통해 제안한 구조는 기존 compensator보다 최대 7.2배 큰 total current mismatch를 처리하거나, 유사한 mismatch 조건에서 약 20% 높은 시스템 효율을 달성하였다. 결과적으로 VSIMO는

하나의 inductor만으로 여러 rail의 surplus current를 회수하고 부족한 전류를 공급함으로써, stacked voltage domain의 높은 이론적 효율을 실제 불균형 부하 조건에서도 유지할 수 있음을 입증하였다.

저자정보



이윤호 박사과정 대학원생

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

2026 VLSI Review

단국대학교 파운드리공학과 석사과정 장준서

Session 11 : Wireless Power Transfer & Energy Harvesting Interfaces

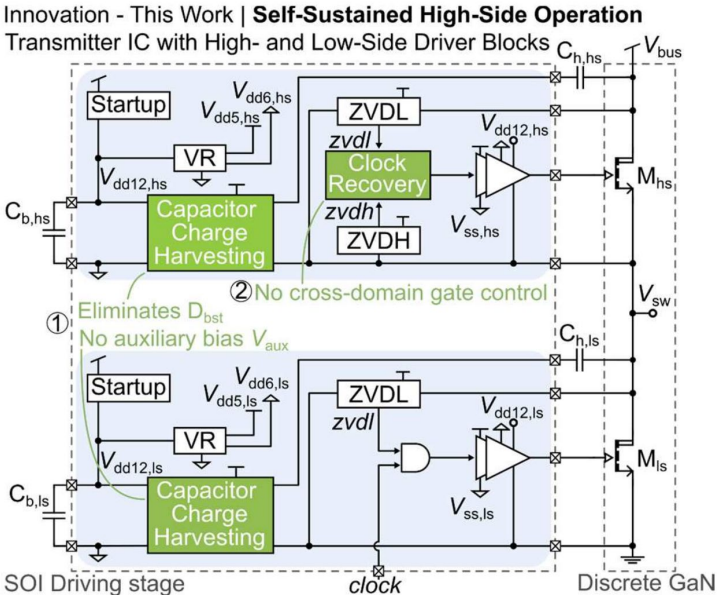
이번 2026 VLSI의 Session 11에서는 무선전력전송과 에너지 하베스팅을 주제로 5편의 논문이 발표되었다. 효율 향상과 함께 구동 회로의 단순화와 입력 조건에 따른 적응형 제어가 주요 경향으로 나타났다. 본 리뷰에서는 이러한 방향을 대표하는 Class-D 송신기(C11.1)와 머신러닝 기반 압전 하베스팅 인터페이스(C11.5)를 살펴본다.

#11-1 Leibniz University Hannover에서 발표한 논문으로, 보조 전원과 레벨 시프터 없이 동작하는 6.78 MHz Class-D 송신기를 제안하였다. 기존 구조에서는 고측 스위치의 부유 전원과 전압 영역 간 제어 신호 전달이 필요하여 고주파 구동 회로의 복잡도와 손실이 증가한다.

그림 1의 송신기는 스위칭 노드의 상승 및 하강 에지에서 전하를 회수하여 고측과 저측 드라이버의 바이어스를 공급한다. 한쪽 에지만 이용할 때보다 주기당 회수 전하가 증가하여 필요한 커패시터 크기를 줄일 수 있으며, 보조 전원과 부트스트랩 다이오드도 제거하였다.

고측 클록은 스위칭 노드의 전압을 감지하여 복원한다. 저측 동작 시간을 감지용 커패시터에 기록하고 고측 타이밍 생성에 활용해 레벨 시프터를 제거하였다. DC 버스와 하나의 외부 클록만으로 기동 및 정상 동작이 가능하다.

0.18 μm SOI 공정의 드라이버와 외부 GaN 스위치로 구현한 시스템은 48 V 입력 및 40 mm 전송 거리에서 최대 25 W 출력과 91%의 피크 DC–DC 효율을 달성하였다. 스위칭 파형을 바이어스와 타이밍 생성에 함께 활용한 점이 특징이다.



[그림 1] 자체 바이어스 및 고측 클록 복원을 이용한 Class-D 송신기 구조

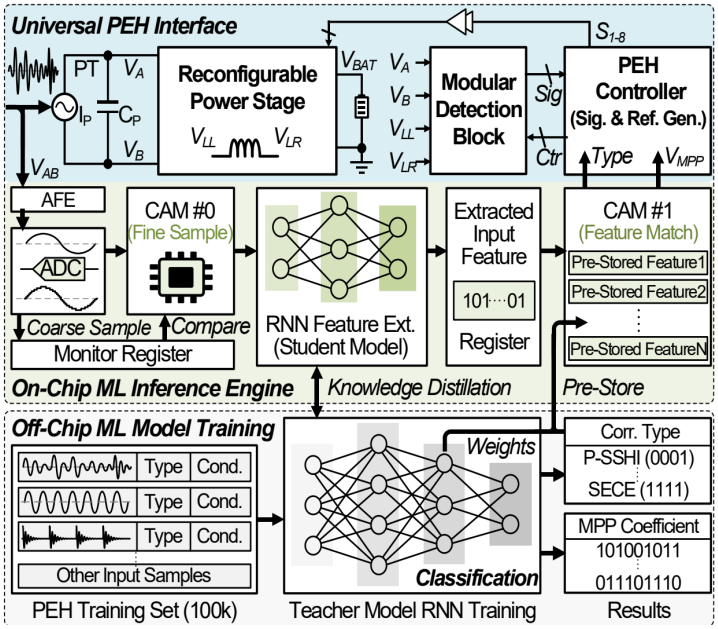
#11-5 University of Macau에서 발표한 논문으로, 머신러닝 기반 입력 식별과 토폴로지 간 최대 전력점 추적을 결합한 압전 하베스팅 인터페이스를 제안하였다. 주기 진동과 충격, 다차원 진동은 입력 파형이 달라 하나의 토폴로지로 모든 조건에서 높은 출력을 얻기 어렵다.

기존의 최대 전력점 추적(MPPT)이 정해진 토폴로지 안에서 동작점을 조절하는 데 비해, 본 논문에서는 입력을 식별하여 토폴로지와 최대 전력점 계수를 함께 선택한다. 입력 조건에 따라 에너지 추출 방식 자체를 바꿀 수 있는 구조이다.

그림 2와 같이 모델의 학습은 칩 외부에서 수행하며, 칩 내부에서는 순환 신경망(RNN)을 이용해 입력 파형의 특징을 추출한다. 추출한 특징을 내용 주소화 메모리(CAM)에 저장된 정보와 비교하여 토폴로지 및 최대 전력점 계수를 결정한다. 재구성 가능한 전력단은 8개의 스위치와 하나의 인덕터로 구성되며, 전하 추출과 전압 반전 등의 동작을 조합하여 선택된 토폴로지를 구현한다.

입력 식별에 필요한 에너지를 줄이기 위해 기동 시에는 파형을 정밀하게 분석하고, 정상 상태에서는 그 길이의 1/5에 해당하는 구간만 관찰한다. 입력이 유지되면 기존 토폴로지와 동작점을 그대로 사용하여 특징 추출과 개방 회로 전압 탐색을 생략한다. 이를 통해 한 번의 입력 식별에 소모되는 에너지를 정밀 식별 대비 최대 59% 절감하였다.

180 nm 공정으로 제작된 칩은 주기 진동, 충격 및 다차원 진동에 따라 서로 다른 에너지 추출 방식으로 동작함을 보였다. 해당 연구는 머신러닝의 판단 결과를 실제 전력단 재구성과 연결하고, 정상 상태의 판단 비용까지 줄여 다양한 진동 환경에 대응하도록 구성한 점이 특징이다.



[그림 2] 입력 식별 및 토폴로지 선택을 위한 범용 압전 하베스팅 인터페이스

두 연구는 각각 스위칭 파형을 재활용한 구동 회로 단순화와 입력 분석에 기반한 전력단 재구성을 제안하였다. 전력 변환을 보조하는 회로와 제어 방식까지 함께 개선한다는 점에서 이번 세션의 설계 방향을 보여준다.

저자정보



장준서 석사과정 대학원생

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

2026 VLSI Review

KAIST 전기 및 전자공학부 박사과정 박수언

Session 15 Integrated Voltage Regulators and Power Converters

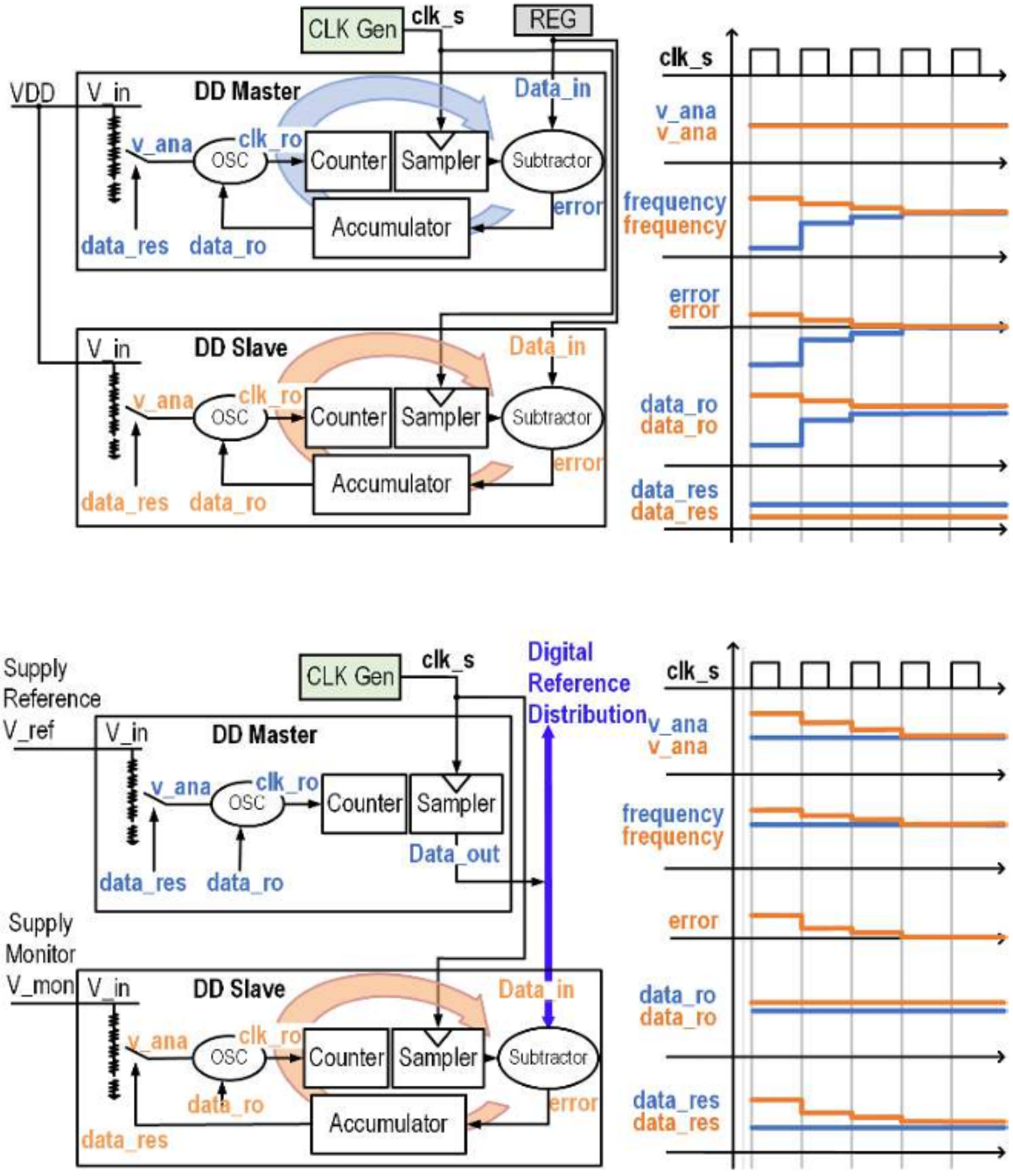
이번 15 Session은 전력 변환기 및 모니터링 기술을 주 내용으로 다루고 있다. 이 중 Infineon Technologies에서 발표한 차량용 SoC 전원 모니터링 논문(C15.1)을 살펴보며, 디지털 친화적이고 분산 가능한 Differential Supply Monitor 아키텍처의 주요 특징과 성능을 고찰하고자 한다.

#15-1 논문은 차량용(Automotive) SoC의 신뢰성 및 기능 안전성(ASIL) 확보를 위해, 추가적인 아날로그 센서나 레퍼런스 없이 Core Supply 도메인 내에서 작동하는 분산형 디지털 전원 모니터(Supply Monitor) 아키텍처를 제안한다.

종래의 전원 모니터링 방식(ADC 또는 링 오실레이터 기반)은 노이즈가 적은 조용한(Quiet) 아날로그 전원 도메인을 요구하거나, 온도의 영향을 보상하기 위해 밴드갭 회로, ULVT 소자, 다중 점 교정 및 룩업 테이블(LUT)이 필수적이어서 구현 및 검사 비용이 큰 단점이 있었다. 제안한 논문에서는 Master-Slave 구조의 차동 Digital Detector (DD) 네트워크를 도입하여, 아날로그 레퍼런스 배선 없이 디지털 레퍼런스 신호만을 분산 배치하는 구조를 완성하였다. 이를 통해 온도, 공정, 경화(Aging) 등 글로벌 변동 요인을 차동 구조 특성으로 자연스럽게 상쇄시켰다.

각 DD는 가변 Resistive Ladder와 Current-starved Oscillator, 카운터 및 디지털 로직으로 구성된다. 웨이퍼 테스트 시 단 1회의 1점 교정(One-time, One-point Calibration)을 통해 오실레이터 제어 코드를 설정한 후, 모니터링 단계에서는 Slave DD가 Master DD와의 오차를 줄이도록 가변 저항 비율을 디지털 피드백으로 조정한다. 이에 따라 모니터링 전압은 오실레이터 특성이 아닌 Resistive Ladder ratio에 의해 선형적으로 계산되어, PVT 변동 및 물리적 거리에 우수한 강인성을 보인다.

28-nm CMOS 공정으로 제작된 제안 모니터 IP는 $0.01\ \mathrm{mm}^2$ 의 매우 작은 면적과 540μW의 저전력을 달성하였다. 1A의 전류를 소모하는 Supply Aggressor 동작 및 -40°C ~ 160 °C의 극심한 온도 범위 조건에서도 10 mV 미만의 우수한 DC 추정 오차를 나타내었으며, $1\ \mu s$ 내에 217mV의 전압 변동을 추적하여 2.45pJ/conv-step의 FoM을 달성하였다. 이는 면적과 검사 비용을 최소화하면서 차세대 차량용 SoC의 안전성을 높이는 우수한 디지털 친화적 솔루션으로 평가된다.



[그림 1] 제안하는 Master-Slave 기반, Supply Monitor의 동작 : Calibration Phase (top), Monitor Phase (bottom).

저자정보



박수언 박사과정 대학원생

- 소속 : KAIST
- 연구분야 : Power Management IC 설계
- 이메일 : tndjs0905@gmail.com
- 홈페이지 : https://icdesignlab.net/

2026 VLSI Review

KAIST 전기및전자공학부 박사과정 이준기

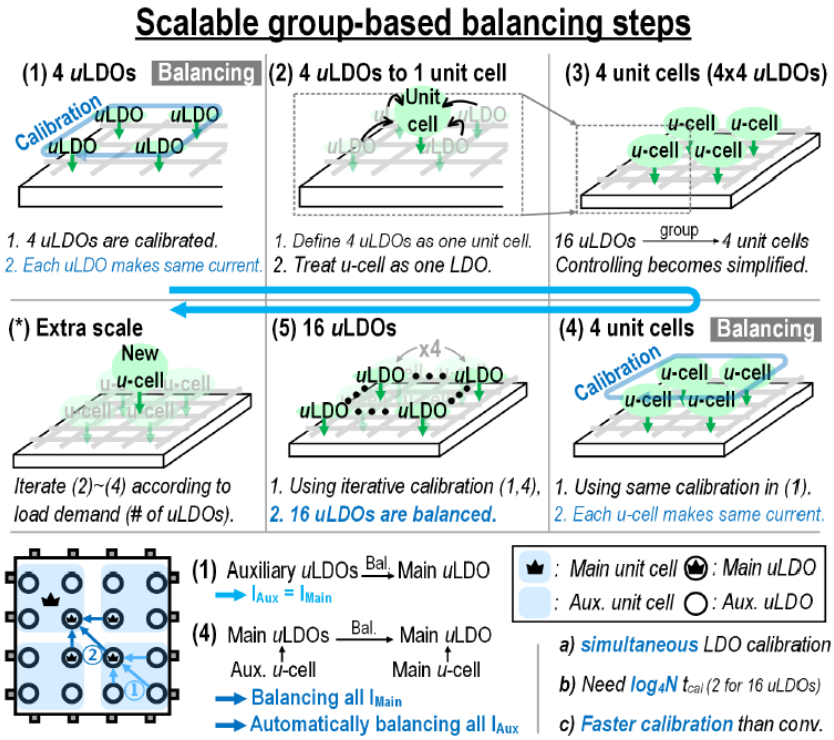
Session 22 Fast and High-PSR On-Chip LDO Regulation

이번 2026 VLSI의 Session 22에서는 Fast and High-PSR On-Chip LDO Regulation이라는 주제로 총 4편의 논문이 발표되었다. 이번 session에서는 고성능 SoC의 안정적인 전력 공급을 위한 on-chip LDO 기술들이 다루어졌으며, 빠른 과도응답을 위한 digital LDO 제어 기법부터 분산형 LDO의 전류 균일화, 넓은 주파수 범위의 전원 잡음을 억제하기 위한 high-PSR analog LDO까지 다양한 회로 기술들이 소개되었다. 본 review에서는 대규모 distributed LDO에서 빠르고 정확한 전류 균일화를 구현한 C22-2와 최대 1GHz까지 높은 PSR 성능을 확보하면서 빠른 과도응답을 달성한 C22-3을 다룬다.

#22-2는 서강대학교에서 발표한 논문으로, 대규모 distributed digital LDO에서 빠르고 정확한 전류 균일화를 달성하기 위한 scalable distributed digital LDO(SD-DLDO)를 제안하였다. Distributed LDO에서는 여러 unit LDO(uLDO)의 출력전류가 균일하지 않을 경우 특정 영역에 전류가 집중되어 local hotspot을 야기할 수 있다. 또한, 기존의 serial 또는 hierarchical calibration 방식은 N개의 uLDO 전류를 균일화하기 위해 N-1번의 순차적인 calibration이 필요하므로 LDO 개수가 증가할수록 convergence time이 크게 증가하는 한계점이 존재한다.

해당 논문에서는 이러한 문제를 해결하기 위해 analog current sensor와 group-based simultaneous calibration을 적용하였다. 각 uLDO 내부에는 하나의 power transistor와 함께 compact analog current sensor를 집적하여 실제 출력전류를 직접 감지한다. 또한, conventional serial calibration과 달리 4개의 uLDO를 하나의 group으로 구성하고, group 내부의 모든 uLDO를 동시에 calibration한다. Calibration이 완료된 4개의 uLDO를 하나의 unit cell로 취급한 뒤, 다시 4개의 unit cell을 상위 단계에서 동일한 방식으로 calibration 함으로써 구조를 계층적으로 확장할 수 있다. 이를 통해 N개의 uLDO를 균일화하는데 필요한 calibration step을 획기적으로 감소시켰다.

제안하는 구조는 28nm CMOS 공정에서 16개의 uLDO로 구현되었다. 측정 결과 부하전류가 1.6A인 상황에서 134.6us의 convergence time과 5.2%의 current balancing error를 달성하였으며, balancing 과정에서도 출력전압에 glitch나 ringing이 발생하지 않았다.



[그림 1] 그룹-기반 동시 교정을 이용한 scalable distributed digital LDO

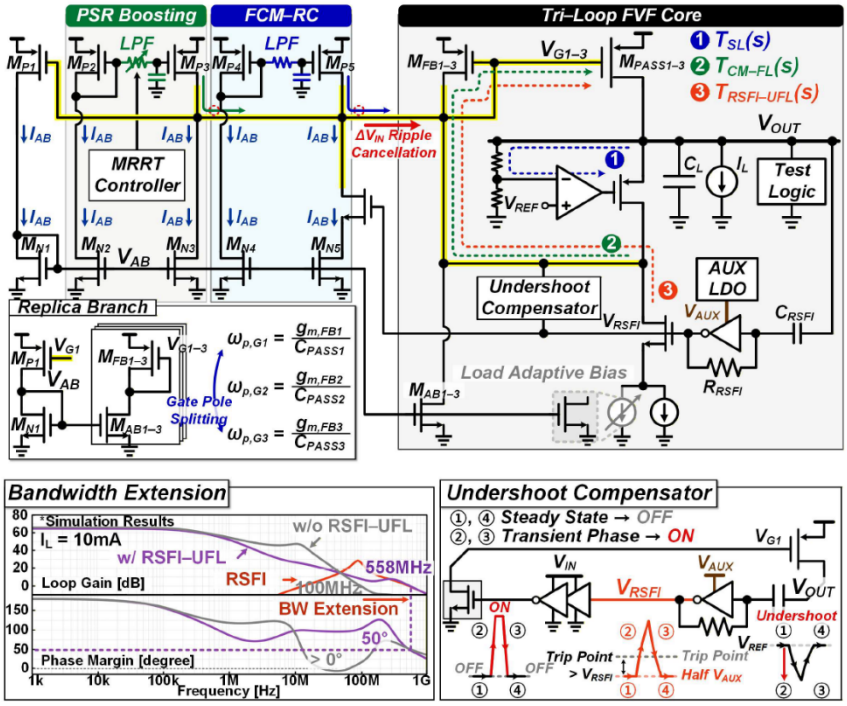
#22-3은 성균관대학교에서 발표한 논문으로, 넓은 주파수 범위의 PSR과 빠른 과도응답을 동시에 달성하기 위한 gain-bandwidth(GBW)-enhanced tri-loop analog LDO를 제안하였다. 최근 SoC에서는 switching regulator와 LDO를 함께 사용하는 전원 구조와 빈번한 active-idle mode transition으로 인해 넓은 주파수 범위의 전원 ripple과 급격한 부하 변동이 동시에 발생한다. 기존의 fully integrated LDO는 high-frequency PSR과 빠른 transient response 사이에 trade-off가 존재하기 때문에, 이를 해결하기 위해 세 종류의 regulation loop을 이용하여 두 성능을 동시에 향상시켰다.

제안하는 LDO는 gain-enhanced slow loop(SL), current-mirror-based fast loop(CM-FL), regulated shunt-feedback-inverter-based ultrafast loop(RSFI-UFL)의 세 개 loop으로 구성된다. SL은 높은 DC regulation gain과 저주파 PSR을 향상시키며, CM-FL에는 filtered current-mirror ripple cancellation을 적용하여 입력 전압에서 발생한 ripple 성분을 feedback path에 주입함으로써 pass transistor의 ΔV_{SG}를 최소화한다. RSFI-UFL은 LDO 출력의 고주파 성분을 빠르게 감지하여 regulation bandwidth를 500MHz 이상으로 확장하고, load transient 시 pass transistor의 gate를 빠르게 구동하여 transient response를 개선한다.

특히 해당 논문에서는 switching regulator의 dominant ripple을 선택적으로 억제하기 위해 notch-like PSR boosting과 maximum ripple rejection tracking(MRRT) 기법을

제안하였다. 추가적인 filtered current-mirror path의 attenuation coefficient를 조절하여 특정 주파수에서의 supply-ripple을 상쇄하였다. MRRT controller는 time-domain ripple sensor와 Vernier TDC를 이용하여 출력에 존재하는 잔여 ripple 크기를 감지한다. 이후 control word를 sweep하면서 잔여 ripple이 최소가 되는 조건을 자동으로 탐색하여 PSR boosting frequency를 입력 전압의 dominant ripple frequency 맞춘다.

제안하는 LDO는 40nm CMOS 공정에서 제작되었으며, 40pF의 on-chip output capacitor를 사용하였다. 측정 결과 10Hz부터 1GHz까지 -25.5dB보다 우수한 PSR을 유지하고, 10MHz에서는 MRRT를 통해 -52.8dB의 PSR을 달성하였다. 또한, 0.15mA에서 10mA로 변화하는 load transient에서 50mV의 undershoot과 13ns settling time을 달성하였다. 실제 pre-stage DC-DC converter와 연결한 측정 결과, MRRT를 통해 출력 ripple을 약 10배 감소시켰다.



[그림 2] GBW-enhanced tri-loop LDO의 전체 회로 구조

저자정보



이준기 박사과정 대학원생

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

2026 VLSI Review

한국과학기술원 전기및전자공학부 석사과정 이가은

Session 9 Neural Interface and Brain-Computer Interface Circuits

이번 2026 IEEE Symposium on VLSI Technology and Circuits의 Session 9는 Neural Interface and Brain-Computer Interface Circuits라는 주제로 총 4편의 논문이 발표되었다. 이 세션은 ECoG, EEG 등 뇌신경 신호로부터 유의미한 정보를 실시간으로 추출·해석하는 온칩 지능형 신경 신호 처리 회로를 중심으로, 운동 의도 디코딩, 발작 검출, 수면 조절 등 다양한 뉴럴 인터페이스 응용을 위한 저전력·고집적 SoC 아키텍처를 탐구한다. 네 논문 모두 딥러닝 기반 디코딩 모델을 온칩에 완전히 통합하여 외부 컴퓨팅 자원에 대한 의존을 제거하고, computing-in-memory(CIM) 등 메모리 중심 연산 구조를 통해 극저전력·저지연 동작을 실리콘 검증으로 입증하고 있다는 점에서 공통된 방향성을 보인다.

#9-1 본 논문은 EPFL 연구 팀에서 발표한, 최초로 완전 온칩 통합된 트랜스포머 기반 BCI SoC 연구이다. 척수손상 및 뇌졸중으로 인한 사지마비 환자의 운동 기능 회복을 위해서는 외부 컴퓨터에 의존하지 않는 실시간 온칩 디코딩이 필수적이나, 기존 신경 디코딩 SoC들은 오프칩 메모리에 의존하거나 온칩 저장·연산 자원의 한계로 인해 트랜스포머와 같은 고성능 모델의 탑재가 어려웠다는 문제의식에서 출발한다. 특히 기존의 backpack-mounted 외부 디코딩 시스템은 100ms를 초과하는 큰 레이턴시를 유발해 실사용성에 제약이 있었다는 점이 강조된다.

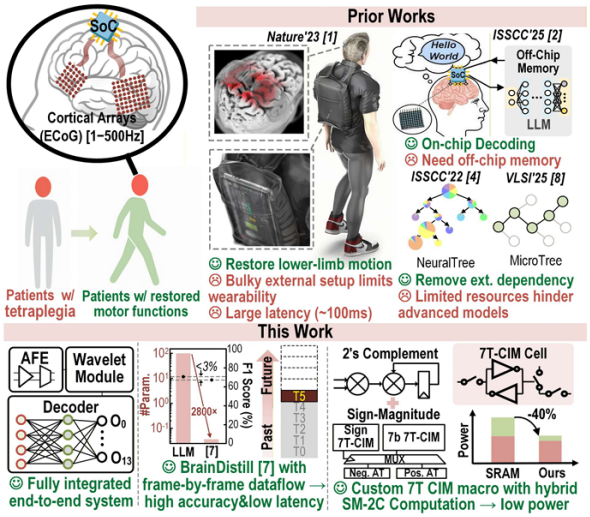


Fig. 1. Prior work (top) and contributions of this work to address shortcomings of prior art (bottom).

[그림 1] 선행연구 대비 본 연구의 contribution 구조도

64 채널 저잡음 AFE, 웨이블릿 기반 특징 추출기, 온칩 백엔드 디코더로 구성된 end-to-end 시스템(Fig. 2)을 제안하며, 핵심은 지식 증류(knowledge distillation)를 통해 약 2800 배 파라미터를 압축한 경량 트랜스포머 BrainDistill 이다. 약 30K 파라미터만으로 소프트웨어 기반 약 100M 파라미터 모델에 근접하는 성능을 유지하며, 10 프레임 윈도우를 한 번에 배치 처리하지 않고 프레임 단위(frame-by-frame)로 순차 처리함으로써 웨이블릿-디코더 간 메모리 접근을 약 90% 절감한다. 모든 동작은 온칩 FSM 으로 오케스트레이션되어 오프칩 메모리 접근 없이 완전한 온칩 처리를 가능하게 한다.

AFE 부분(Fig. 3)은 4 개의 16 채널 모듈로 분할되어 멀티플렉서를 통해 공유 Gm-C 필터 및 10 비트 SAR ADC(64kS/s, 채널당 4kS/s)로 라우팅되는 구조이며, AC-coupled LNA 와 백엔드 자원 공유의 아키텍처 최적화를 통해 기존 대비 30%의 면적 절감을 달성하였다. 측정 결과 채널당 3.56μW 의 전력을 소모하면서 33~66dB 의 프로그래머블 게인, 45.1dB 의 SNDR, 58.1dB 의 SFDR, 1kHz 에서 10.6MΩ 의 입력 임피던스, 3.06μVrms 의 입력 환산 잡음(IRN)을 달성해, 신경 디코딩에 필요한 신호 충실도를 유지하면서도 면적을 줄인 것이 특징이다.

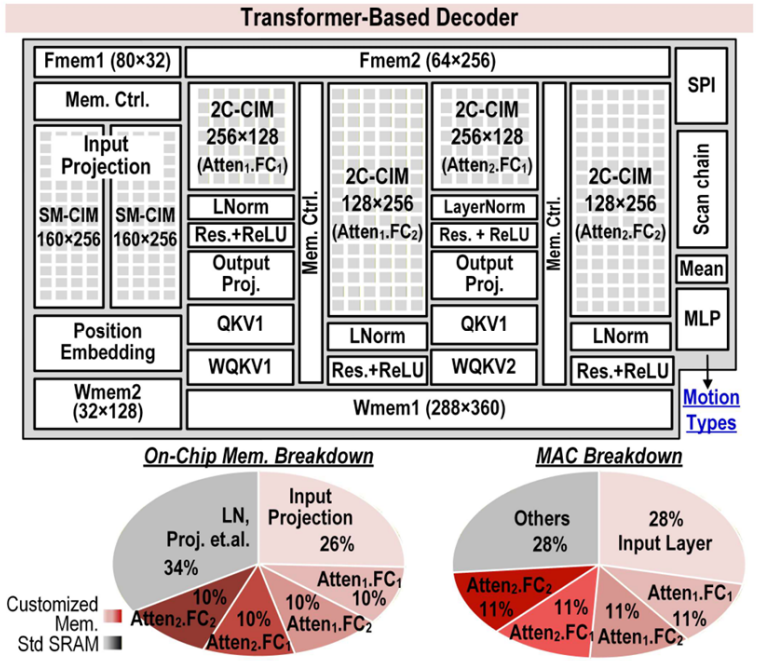


Fig. 4. Architecture of the proposed transformer implementation for motion-intention decoding.

[그림 2] 트랜스포머 기반 디코더의 아키텍처

디코더의 핵심은 커스텀 7T-SRAM CIM 매크로이며, 부호-크기(SM)와 2 의 보수(2C) 표현을 혼합 사용하는 하이브리드 SM-2C 연산 기법을 제안한 점이 이 논문의 가장 독창적인 부분이다. 활성화 값이 0 근처에 밀집되는 입력 프로젝션 단계에는 비트

토글이 적은 SM-CIM 을, 넓은 활성화 범위를 갖는 FC1/FC2 단계에는 2C-CIM 을 적용함으로써 전체 토글율을 30% 감소시킨다(Fig. 6). 8 비트 가중치는 부호·크기 비트로 분해되어 8 개의 수직 적층 7T-CIM 셀에 저장되고, isolated read/multiply(VDDH)와 retention/write(VDDL) 전압 도메인을 분리한 구조(Fig. 5)를 채택해 Monte-Carlo 시뮬레이션 기준 VDDH>0.7V 조건에서 신뢰성 있는 쓰기 동작을 확보하였다. 이 구조를 통해 전체 파라미터의 66%, MAC 연산의 72%가 CIM 어레이로 처리되며, 디코더 전력의 40% 절감 효과를 얻는다. 또한 두 개의 feature SRAM 을 두 선형 어텐션 레이어에 재사용함으로써 feature 메모리 사용량을 61.44kb 에서 14.08kb 로 4.4 배 줄였다.

65nm 공정으로 제작된 칩은 전체 10.1mm²(디코더 4.7mm², 코어 면적 1.87×2.49mm²)를 차지하며, 온칩 SRAM 은 foundry SRAM 9.47kB 와 CIM-SRAM 26kB 로 구성된다(Fig. 8). WIMAGINE 장치로 기록한 6-클래스 상지 운동 의도 데이터셋에서 67.8%의 분류 정확도(F1-score 68%)를 달성하였으며, 손가락 동작 데이터셋에서는 77%(F1-score 75.7%)로 더 높은 정확도를 보였다. 디코더는 47μW(0.224ms 연산 지연, 10.53nJ/classification)만을 소모하며, 이는 68.6 TOPS/W 의 에너지 효율에 해당한다. 전력 breakdown(Fig. 9)을 보면 dynamic power 가 19%, leakage power 가 81%를 차지하고, foundry SRAM 의 leakage 가 전체 leakage 의 81%를 점유해 SRAM 이 여전히 지배적인 전력 소모원임을 알 수 있다. CIM 아키텍처 도입으로 소프트웨어 시뮬레이션 대비 dynamic/leakage 전력이 각각 73%/18% 감소하였고, 6-클래스 confusion matrix 에서도 Rest 클래스가 983 건으로 가장 많은 정분류 건수를 보이는 등 클래스별 편차가 존재함이 확인된다. Fig. 10 에 제시된 실제 예측 확률 궤적에서는 어깨 굴곡, 팔꿈치 신전, 손목 회내, 손 벌림/쥐기 등 6 개 클래스가 안정적으로 구분되는 것을 확인할 수 있다.

Table I 의 비교에서 본 연구는 기존 CNN·DFT 기반 발표작들(ISSCC'24 Zhong, ISSCC'22 Shin, TBioCAS An 등) 대비 가장 낮은 전력을 유지하면서도 트랜스포머라는 더 큰 모델 규모(35KB)와 더 많은 클래스(14 클래스, 상지 동작) 지원을 동시에 달성하여, 손가락 단위의 단순 제스처 디코딩에 머물렀던 선행 연구들과 차별화된다. 특히 0.224ms 의 연산 지연은 이전 SoC 들의 <1000ms 또는 50ms 수준과 비교해 압도적으로 낮은 값이며, 이는 backpack 형 외부 디코딩 시스템(>100ms) 대비 실시간·비침습적 사용성을 크게 개선한 결과로 해석된다. 다만 정확도(67.8%)는 여전히 개선 여지가 있으며, 이는 6-클래스 상지 동작 분류가 기존 손가락 수준 태스크보다 본질적으로 더 어려운 문제라는 점을 감안할 필요가 있다.

#9-3 본 논문은 동일 연구실(EPFL)에서 발표한 발작(seizure) 검출용 초저전력 CIM 프로세서 TARNA 연구이다. 기존 발작 검출 SoC는 환자별 발작 데이터를 이용한 few-shot

학습에 의존하여 실사용이 어렵거나, zero-shot 방식이더라도 전극 위상(topology) 정보와 장기 시간 의존성을 반영하지 못하거나 오프칩 FFT 전처리에 의존해 완전한 온칩 통합이 어려운 한계가 있었다. 저자들은 그래프 합성곱(GCN)과 게이트가 없는 잔차 순환 상태 업데이트를 결합하여, 환자별 발작 데이터 없이도 높은 정확도와 초저전력 동작을 동시에 달성하는 구조를 제안한다.

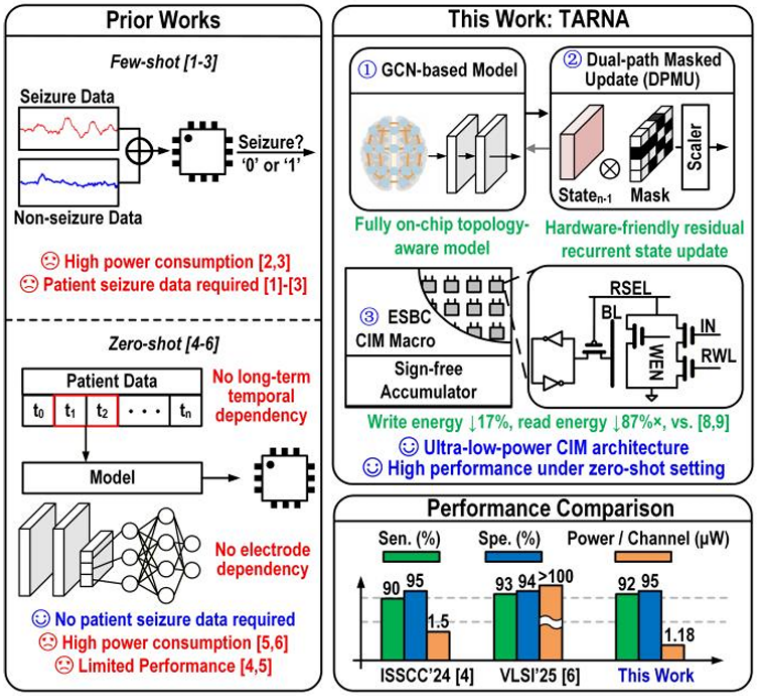


Fig. 1. Limitations of existing designs and contributions of this work.

[그림 3] 선행연구의 한계와 TARNA의 contribution 요약

TARNA 의 모델 파이프라인은 전극별로 32 개 샘플(0.125s, CHB-MIT 기준)을 하나의 그래프 노드로 구성하고, 4 개 주파수 대역의 스펙트럴 에너지(SE)를 밀집층으로 투영한 뒤 이전 상태(Staten-1)와 잔차 결합하여 시간적 문맥을 보존한다(Fig. 2). zero-shot fine-tuning 이나 추가적인 데이터 기반 기법 없이도 기본 모델만으로 83% 민감도, 84% 특이도를 달성해 선행 zero-shot 연구 대비 향상된 일반화 성능을 보인다. 전극의 물리적 배치를 반영한 sparse edge matrix 기반 GCN 레이어(Fig. 3)는 공간적으로 먼 채널 간 연결을 가지치기하여, 예를 들어 첫 채널의 공간적 집계에서 포인트와이즈 컨볼루션이 19 개의 부분곱을 생성하는 것과 달리 인접한 5 개 채널만으로 연산을 수행함으로써 동일 크기의 CNN 레이어 대비 MAC 연산을 76% 줄이면서도 민감도·특이도를 각각 6.6%, 5.2% 개선한다.

상태 업데이트는 dual-path masked update(DPMU) 기법을 사용한다(Fig. 4). 학습 경로에서는 무작위 이진 마스크를 저장된 상태에 적용해 정규화 효과를 얻고, 추론

경로에서는 이 마스크를 우회함으로써 별도의 의사난수 생성기 없이도 비선형 게이팅(GRU 등)이 요구하는 추가 연산과 제어·메모리 오버헤드를 모두 제거한다. 상태 크기를 정렬하기 위한 스케일링 팩터가 잔차 결합 전에 적용되며, 이 gate-free 접근법은 CHB-MIT 데이터셋에서 82.6% 민감도, 84.0% 특이도를 달성해 동일 조건의 GRU 베이스라인 대비 민감도 2.6%p, 특이도 2.8%p 우수한 성능을 보인다.

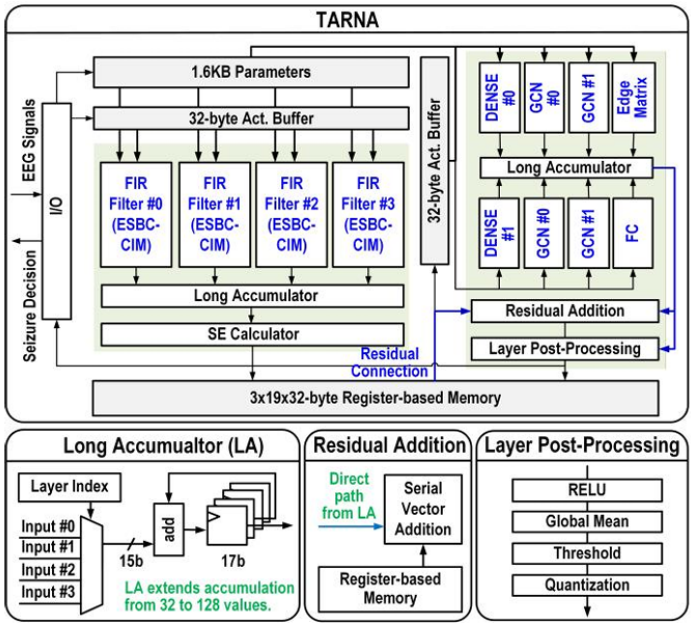


Fig. 5. Hardware architecture of the TARNA.

[그림 4] TARNA의 전체 하드웨어 아키텍처

전력 효율의 핵심은 energy-saving bit cell(ESBC) 기반 CIM 매크로이다(Fig. 6). 매크로는 1k 개의 4 비트 가중치를 저장하며, 4 개의 비트 셀이 하나의 공유 control-and-compute unit(CCU)을 사용해 bit-cell group(BCG)을 구성한다. 매크로 내부에는 8 개의 sign-free accumulator 가 통합되어 있으며, 한 사이클에 32 개 가중치로부터 32 비트를 페치해 부분곱을 생성한 뒤 시프트·누적한다. 부호 있는 곱셈을 정확히 처리하기 위해 이후 사이클에서 최대 4 개의 캐리("+1") 비트를 주입하는 방식을 채택해, 누산기 입력이 비트 반전만으로 동작하도록 하고 누산기당 3 회의 덧셈 연산을 제거하였다. 읽기 시에는 CCU 가 우측 접근 트랜지스터를 차단해 별도의 감지 증폭기 없이 고정 노드 전압을 형성하는 sparsity-aware read pattern(SARP)을 사용하고, 쓰기 시에는 셀을 준안정 상태로 두어 최대 4 배에 달하는 비트라인 충전 에너지를 회피한다. 이를 통해 비트당 평균 27.3%의 읽기 에너지 절감을 달성하며(Fig. 7), CCU 공유 자체만으로도 면적과 누설 전력을 각각 35%, 33.4% 절감한다. 종합적으로 기존 CIM 설계([8],[9]) 대비 쓰기 에너지 1.17 배, 읽기 에너지 1.87 배, 읽기 지연 10.2 배의 개선을 보인다.

65nm 공정으로 제작된 TARNA 는 3mm²(1.0mm×3.0mm) 면적에 5,416 개의 가중치와 5.64kB 의 온칩 SRAM 만으로 구성되며, 레지스터 기반 메모리를 사용해 0.65V 의 서브스레숄드 동작을 지원한다. 22.5μW(1.18μW/channel)의 전력을 소모하며, CHB-MIT 소아 발작 데이터셋 전체 환자를 대상으로 평가한 결과 100%의 이벤트 기반 민감도, 91.6%의 프레임 기반 민감도, 95.3%의 특이도, 0.048/hr 의 위양성률, 0.54s 의 검출 지연을 달성하였다(Fig. 9, Fig. 10). 전력·면적 breakdown 을 보면 분류기(classifier)가 전력의 54%, 특징 추출이 19.2%, 메모리 및 기타가 26.8%를 차지하며, 면적 측면에서는 메모리 및 기타가 36.7%로 가장 큰 비중을 차지해 전력과 면적의 지배 요인이 서로 다름을 보여준다. 또한 TUSZ 데이터셋(315 명 전체 환자) 평가에서도 사전 fine-tuning 없이 0.88 의 AUROC 를 기록해 진정한 zero-shot 일반화 성능을 입증하였다.

Table I 비교에서 TARNA 는 기존 발표작(ISSCC'22, JSSC'22, VLSI'25, ISSCC'23, VLSI'25, ISSCC'24) 대비 채널당 전력이 1.18μW 로 가장 낮으면서도, 학습에 사용한 테스트 환자 데이터의 비율이 0%인 진정한 zero-shot 조건에서 위양성률과 민감도 모두 우수하거나 동등한 성능을 보인다. 특히 few-shot 기반 선행 연구들이 환자 데이터의 20~35%를 학습에 사용한 것과 대비된다는 점에서, 저전력·고정확도·환자 비의존성이라는 세 가지 상충 조건을 동시에 만족시킨 사례로 평가된다.

저자정보



이가은 석사과정 대학원생

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

2026 VLSI Review

포항공과대학교 반도체대학원 박사과정 박은빈

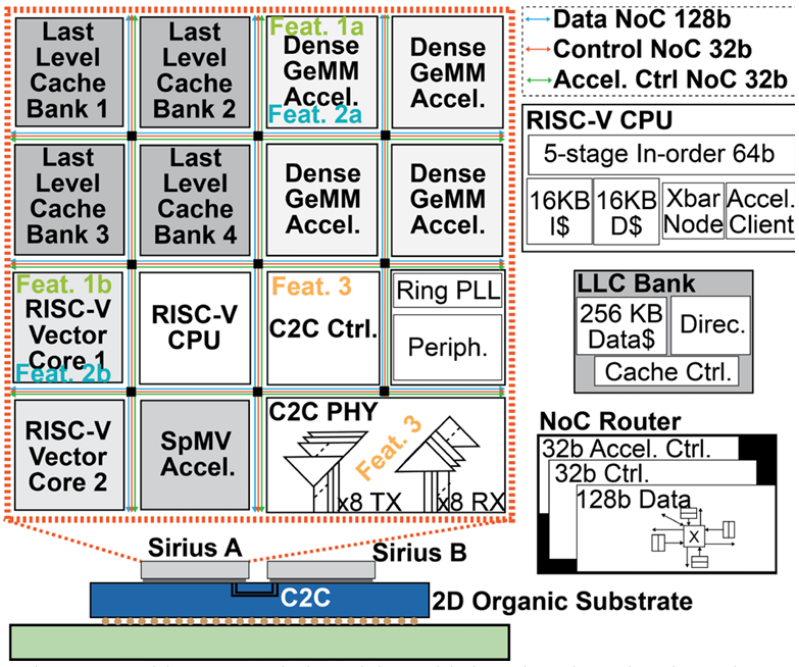
Session 21 SoCs for Reasoning and Robotics

2026 IEEE Symposium on VLSI Technology and Circuits의 Session 21에서는 초저전력 on-device reasoning, edge 기반 agentic AI, multimodal embodied AI, 그리고 Vision-Language-Action(VLA) 기반 실시간 로봇 제어를 중심으로 총 네 편의 논문이 발표되었다. 최근 AI 시스템은 단순한 이미지 인식이나 언어 모델 추론을 넘어, 주변 환경을 지속적으로 인식하고(reasoning), 판단하며, 실제 물리적 행동까지 수행하는 Physical AI 및 embodied AI 방향으로 빠르게 확장되고 있다. 이러한 변화에 따라 AI 프로세서 또한 높은 TOPS를 제공하는 범용 행렬 연산 가속기에서 벗어나, Transformer/vision/robotics workload의 서로 다른 연산 특성을 동시에 지원하고 제한된 메모리 및 전력 환경에서 end-to-end inference를 수행할 수 있는 시스템 수준의 설계가 요구되고 있다. Session 21의 논문들은 이러한 흐름을 반영하여, wearable의 always-on reasoning부터 agentic AI 및 실제 robot perception-to-action까지 다양한 형태의 차세대 edge AI SoC 구현 사례를 제시하고 있다.

#21-3 논문에서는 Transformer 기반 language reasoning, vision backbone, robotics perception과 같이 연산 특성이 서로 다른 multimodal workload를 효율적으로 처리하기 위한 dual-chiplet heterogeneous AI SoC인 Sirius를 제안하였다. 각 chiplet은 4개의 Dense GeMM accelerator, 1개의 SpMV accelerator, 2개의 RISC-V Vector(RVV) core를 통합하며, programmable LUT 기반 non-uniform quantization과 Dense+Sparse decomposition을 통해 dense 연산과 sparse outlier 연산을 각각 적합한 accelerator에서 수행한다. 또한 RVV와 GeMM accelerator의 local memory를 logically shared, physically distributed SPAD로 연결하여 intermediate data가 LLC나 DRAM을 거치지 않고 직접 전달되도록 함으로써 heterogeneous kernel fusion에서 최대 약 3×의 성능 향상을 달성하였다.

Sirius는 두 개의 chiplet을 custom chip-to-chip interconnect로 연결하여 multimodal pipeline을 병렬 실행할 수 있도록 확장성을 확보하였다. Intel 16 nm FinFET 공정으로 제작된 시스템은 Dense GeMM에서 최대 2.24 TFLOPS/W, SpMV에서 7.11 TFLOPS/W를 달성하였으며, 실제 robotics workload인 ViPlanner에서 0.97 TFLOPS/W, Qwen2.5-0.5B와 같은 SLM에서는 최대 16 tokens/s, 23 mJ/token의 성능을 보였다. 본 연구는 단일 AI accelerator의 peak throughput을 높이는 것보다, dense/sparse/irregular workload를

heterogeneous compute engine에 분배하고 chiplet 단위로 확장하는 embodied-AI SoC 구조를 제시했다는 점에서 의미가 있다.

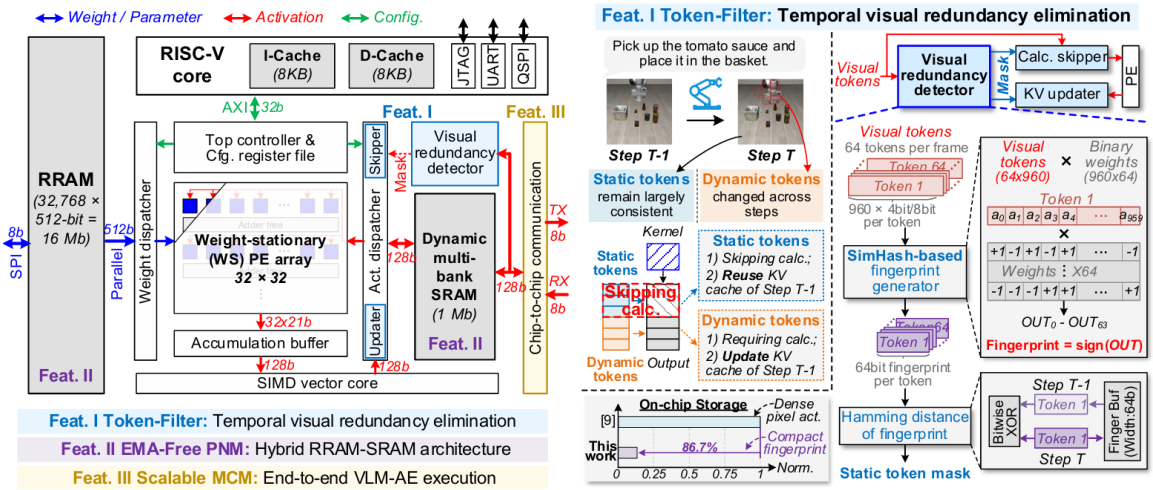


[그림 1] Sirius dual-chiplet SoC 구조 및 heterogeneous accelerator 구성

#21-4 논문에서는 Vision-Language Model(VLM)의 perception과 Diffusion Transformer 기반 Action Expert(AE)를 edge에서 end-to-end로 실행하기 위한 실시간 Vision-Language-Action(VLA) processor를 제안하였다. VLM에서는 연속된 robot action step 사이에 반복되는 static visual token이 많은 점에 착안하여, 960-dimensional token을 64-bit fingerprint로 변환하는 SimHash 기반 temporal token filter를 적용하였다. 이를 통해 static token의 연산과 KV update를 생략하여 평균 59.6%의 연산량을 줄였으며, success rate는 90%에서 87%로 소폭 감소하였다. 반면 memory-bound 특성을 갖는 AE에는 16 Mb RRAM과 1 Mb SRAM을 결합한 hybrid Processing-Near-Memory(PNM) 구조를 적용하여 weight와 activation의 external memory access를 제거하였다.

또한 16-chip MCM 구조를 적용하고 chip 간 activation transfer와 computation을 overlap하여 VLM과 AE의 전체 pipeline latency를 크게 감소시켰다. 55 nm CMOS로 제작된 prototype은 Libero benchmark에서 5.31–6.6ms의 end-to-end perception-to-action latency, 10.46–12.74 mJ/inference, 그리고 최대 150 Hz의 robot control rate를 달성하였다. 특히 기존 accelerator가 VLM 또는 Action Expert 중 일부만 가속하던 것과 달리, 본 연구는 perception부터 action generation까지 전체 VLA pipeline을 실제 silicon에서 구현했다는 점이 특징적이다. 따라서 algorithm-level token filtering, memory-centric PNM, multi-

chip scheduling을 함께 최적화하여 실시간 Physical AI 및 robotics를 위한 end-to-end VLA hardware의 가능성을 보여준 연구로 평가할 수 있다.



[그림 1] SimHash 기반 temporal token filtering 및 VLA processor 전체 구조

저자정보



박은빈 박사과정 대학원생

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

2026 VLSI Review

연세대학교 전기전자공학과 석박통합과정 김동욱

Session 23 ASICs for Emerging Workloads

2026 VLSI의 Session 23은 ASICs for Emerging Workloads 라는 주제로 총 4편의 연구가 발표되었다. 연구들은 소형(compact) video encoder 시스템, spatiotemporal reconstruction, 웨어러블 AI 및 dynamic RF 신호 분석을 대상으로, 각 응용의 연산과 데이터 흐름에 최적화된 하드웨어 솔루션을 제시했다. 본 리뷰에서는 새로운 워크로드의 병목을 해결하는 아키텍처적 기여에 주목하여, spatiotemporal reconstruction을 위한 4DGS 프로세서, 다양한 AI 모델을 지원하는 Hybrid CIM-NPU 프로세서, 그리고 dynamic RF 환경에 대응하는 하드웨어 스케줄링 기반 스트리밍 프로세서를 다룬 3편의 연구를 중심으로 살펴본다.

#23-2 KAIST 연구진이 발표한 본 논문 [Specter: 376 fps 2.67 nJ/pixel 4DGS Processor for Spatiotemporal Reconstruction with Contribution-Ordered Rendering and Deformation-Aware Modeling]은 시간에 따라 변화하는 scene의 재구성을 위한 4D Gaussian Splatting (4DGS) 프로세서인 Specter를 제안했다. 4DGS는 3차원 Gaussian으로 표현한 공간 정보에 시간에 따른 변형을 반영하는 기술이지만, 매 시점 수십만 개의 Gaussian을 처리해야 하므로 외부 메모리 접근과 연산 비용이 크다는 한계가 있었다. 특히, 실제 픽셀 색상에 기여하지 않는 Gaussian까지 타일에 할당하고 처리하는 과정에서 상당한 낭비가 발생했다. 본 연구에서는 이를 해결하기 위해, Gaussian의 불투명도(opacity)와 형상(shape)에 따라 타일 할당 범위를 조정하는 Adaptive Tile Allocation Core (ATAC)를 설계하여 baseline 대비 외부 메모리 접근을 67.3% 줄였다. 또한, Gaussian의 기여도가 높은 픽셀부터 처리하고 이후의 불필요한 연산을 타일 및 서브타일 단위로 생략하는 Contribution-ordered Hierarchical Skipping Core (CHSC)를 도입했다. Modeling 단계에서는 타일별 Gaussian ID 목록의 해시(hash)값을 비교하여 변화가 없는 것으로 판단한 타일의 재계산을 생략함으로써 추가적인 연산 절감을 달성했다. 결과적으로 28nm CMOS 공정으로 제작된 12.96mm² 칩은 1352×1014 해상도에서 최대 376fps의 rendering 성능과 최대 112.3 iterations/s의 modeling 성능을 기록했다. 이는 dynamic scene rendering뿐 아니라 modeling까지 지원하면서, scene의 공간적·시간적 중복성을 하드웨어 수준의 연산 생략으로 연결했다는 점에서 의미가 있다.

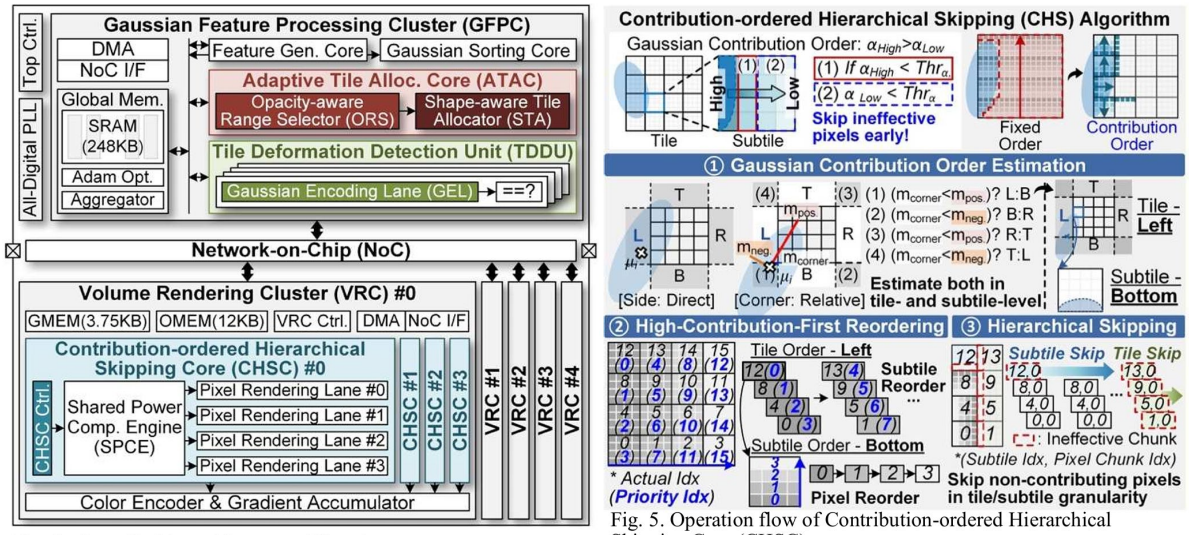


Fig. 3. Overall chip architecture of Specter.

Fig. 5. Operation flow of Contribution-ordered Hierarchical Skipping Core (CHSC).

[그림 1] Specter 전체 구조와 기여도 기반 계층적 렌더링 생략 기법 (원문 Fig. 3, Fig. 5)

#23-3 UNIST와 연세대학교 연구진이 발표한 본 논문 [HCNP: A 70.2 TOPS/W Hybrid CIM-NPU Processor with In-Streaming Processing for Energy-Efficient CNN/Transformer Acceleration in HMD]는 전력 및 발열 제약이 큰 Head-Mounted Display (HMD)를 위한 Hybrid CIM-NPU 프로세서를 제안했다. Compute-in-Memory (CIM)는 고정된 가중치를 사용하는 대규모 행렬 곱셈에 높은 에너지 효율을 제공하지만, 다양한 보조 연산과 데이터 흐름을 단독으로 처리하기에는 제약이 있다. 반면 NPU는 연산의 유연성이 높지만, 두 구조를 별도 칩으로 구성하면 중간 데이터의 외부 메모리 접근이 새로운 병목이 된다. 본 연구에서는 CIM과 NPU를 단일 칩에 통합하고, 데이터가 연산 블록 사이를 이동하는 동안 필요한 처리를 수행하는 In-Streaming Processing Core (ISPC)를 도입했다. ISPC는 nonlinear activation function과 depth-wise convolution 등의 연산을 스트리밍 과정에서 수행하여 layer fusion을 지원하며, 모바일 CNN에서 외부 메모리 접근을 최대 79.7% 줄였다. 또한, attention 기반 중요도에 따라 불필요한 KV cache 항목을 제거하는 Runtime KV Cache Eviction (RKVe)을 NPU와 co-design하여 KV cache 메모리 요구량과 attention MAC 연산량을 각각 최대 79.8%, 32.1% 절감했다. 여기에 flexible-route network-on-chip (FR-NoC)과 single-column-multiple-output (SCMO) CIM mapping 구조를 결합해, 평가한 모바일 CNN 및 LLM에서 평균 CIM 활용도를 각각 77.2%, 81.8%까지 높였다. 결과적으로 28nm CMOS 공정으로 제작된 칩은 8-bit 연산에서 최대 70.2 TOPS/W를 달성했으며, 16-bit MobileNet-v2와 LLaMA2-7B 워크로드에서는 각각 15.0, 17.9 TOPS/W의 효율을 보였다. 이는 CIM 연산기 자체의 효율을 넘어, 이종 연산 블록 사이의 데이터 이동과 활용도까지 함께 최적화하는 co-design을 제시했다는 점에서 주목할 만하다.

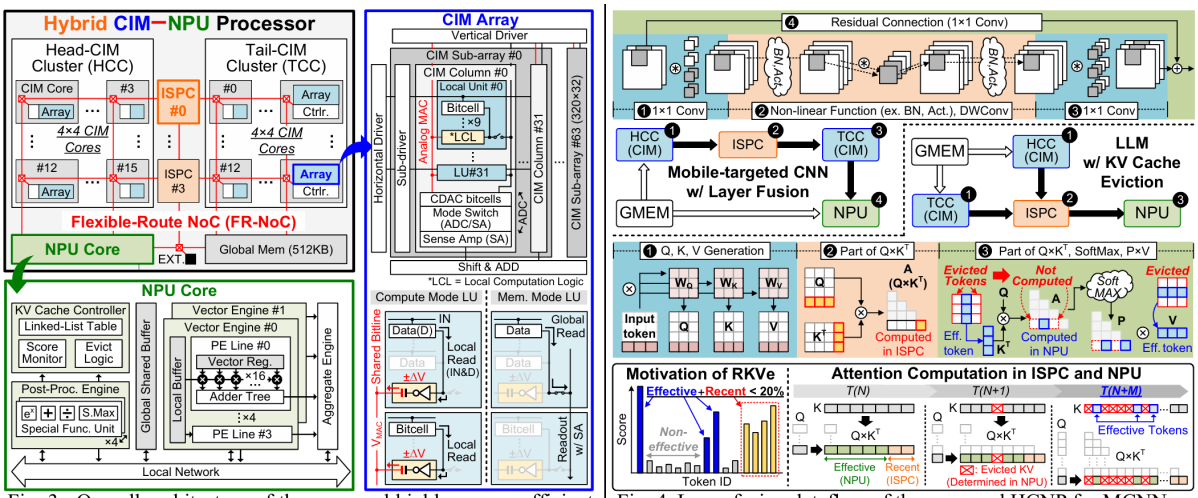


Fig. 3. Overall architecture of the proposed highly energy-efficient Hybrid CIM-NPU Processor for HMD platform.

Fig. 4. Layer fusion dataflow of the proposed HCNP for MCNN and LLM, with the motivation and operation of the proposed RKVe.

[그림 2] HCNP의 Hybrid CIM-NPU 구조와 layer fusion을 위한 데이터 흐름 (원문 Fig. 3, Fig. 4)

#23-4 University of Michigan 연구진이 발표한 본 논문 [A 248.5 GFLOPS/mm², 1.67 TFLOPS/W Streaming Processor with Hardware-Level Scheduling for Advanced Spectrum Sensing]은 무선 환경에서 빠르게 변하는 RF 채널에 대해, 다수 채널을 분석하기 위한 스트리밍 프로세서를 제안했다. Spectrum sensing에서는 현재 신호의 분석 결과에 따라 다음에 실행할 커널이 결정되므로, 실행 순서와 자원 배치를 사전에 스케줄링하기 어렵다. 또한, 개별 커널의 실행 시간이 짧아 host CPU가 매번 커널을 배치하고 loading하는 방식은 큰 오버헤드를 유발한다. 본 연구에서는 이를 해결하기 위해, 77개의 Processing element (PE)와 96개의 Switch로 구성한 2차원 배열에 하드웨어 Kernel Launcher (KL)를 통합했다. KL은 커널 종료 시 실행 그래프의 분기를 판단하고, PE·Switch·DMA 자원의 가용성을 확인하여 다음 커널을 배치한다. 특히, 실행 중인 커널의 후속 분기 후보를 미리 살펴보고 프로그램을 preload 함으로써, host CPU의 개입 없이 최단 10ns의 reconfiguration 시간을 달성했다. PE 내부에는 복소수 MAC 등을 포함한 VLIW 구조를 적용하여 FFT butterfly를 단일 PE에 mapping할 수 있도록 했으며, 사용하지 않는 연산기와 메모리에 대한 세밀한 clock/power gating으로 전력 소모를 줄였다. 결과적으로 28nm 공정으로 제작된 4.78mm² 칩은 FP16 기준 최대 248.5 GFLOPS/mm²의 연산 밀도와 1.67 TFLOPS/W의 에너지 효율을 기록했다. 다중 채널의 동시 실행으로 PE 활용도를 약 5배 높였고, 제안한 스케줄링을 적용하지 않은 baseline 대비 최대 8.3배의 end-to-end 성능 향상을 보였다. 이는 데이터에 따라 실행 경로가 달라지는 워크로드에서, 연산기 성능뿐 아니라 작업 전환과 자원 관리의 지연까지 하드웨어 설계의 핵심 대상으로 삼았다는 점에서 의미가 있다.

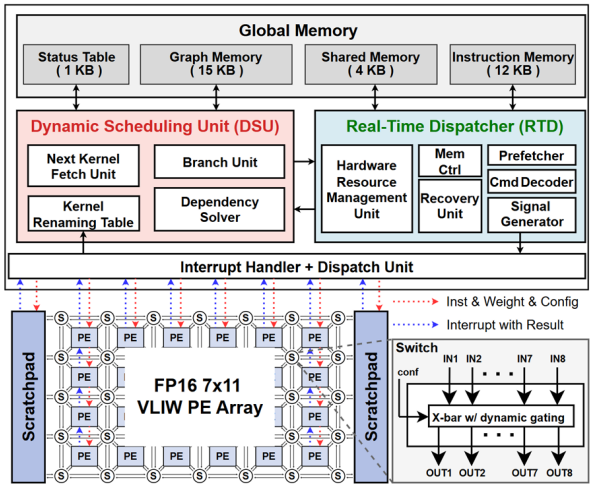


Fig. 3. Overview of proposed architecture. Kernel launcher (KL) is shown in upper part including Global Memory, DSU, RTD, interrupt handler and dispatch unit.

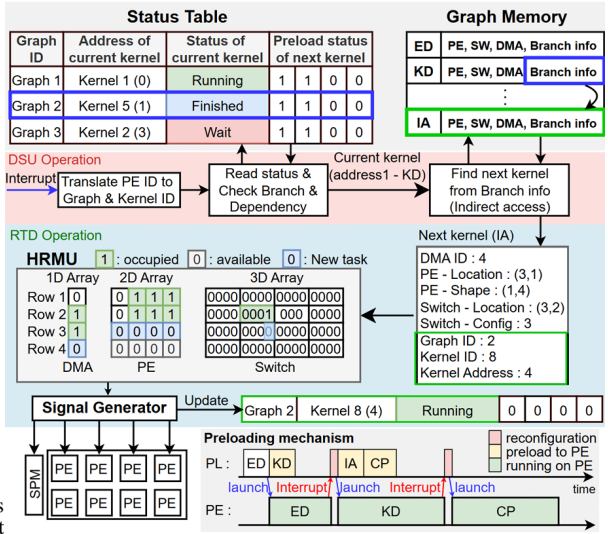


Fig. 5. Detailed explanation of kernel launching protocol

[그림 3] 스트리밍 프로세서 전체 구조와 하드웨어 기반 커널 실행 프로토콜 (원문 Fig. 3, Fig. 5)

저자정보



김동욱 석박통합과정 대학원생

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

2026 VLSI Review

서울대학교 전기정보공학부 박사과정 이재성

Session 27 Next-Generation Wireless Baseband Processors

이번 2026 IEEE/JSAP Symposium on VLSI Technology & Circuits의 Circuits Session 27은 Next-Generation Wireless Baseband Processors를 주제로 구성되었다. 본 글에서는 #27-1 tiled beamspace training/detection processor와 #27-2 multi-subcarrier massive MU-MIMO-OFDM detector를 중심으로 정리한다. 두 연구는 모두 mmWave massive MIMO에서 발생하는 높은 연산량과 메모리·데이터 이동 비용을 줄이기 위해 beamspace의 공간적 sparsity를 이용한다는 공통점이 있다.

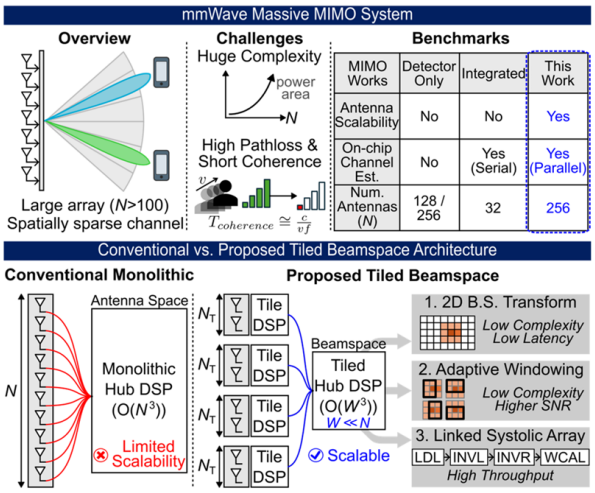
#27-1 본 논문은 mmWave massive MIMO의 training과 detection을 하나의 칩에서 처리하는 tiled beamspace processor를 제안한다. Massive MIMO는 base station에 수백 개 antenna를 사용하므로 array 규모가 커질수록 pilot sequence가 길어지고, channel estimation 이후의 matrix processing 복잡도와 모든 antenna chain을 단일 DSP로 연결하는 routing 부담이 급격히 증가한다. 특히 antenna-space 방식에서는 signal energy가 넓은 antenna dimension에 분산되어 있으므로 channel estimation의 covariance 계산과 weight calculation이 큰 차원의 matrix를 대상으로 수행된다. 저자들은 mmWave channel이 방향성으로 인해 spatially sparse하다는 특성에 주목하고, spatial FFT를 통해 antenna-space signal을 beamspace로 변환하면 에너지가 소수 beam bin에 집중된다는 점을 이용한다.

제안 구조의 핵심은 대규모 256-antenna array를 16개의 tile로 분할하고 각 tile이 16개 antenna를 담당하도록 한 뒤, local tile DSP와 hub DSP에 걸친 2-stage 2D beamspace transform을 수행하는 것이다. 첫 번째 intra-tile FFT 이후 각 tile에서는 폭 W=4의 window만으로도 90% 이상의 signal energy를 포착할 수 있고, 두 번째 inter-tile FFT를 거치면 전체 2D beam map에서 에너지가 다시 집중되어 W1×W2=3×3, 즉 D=9의 reduced dimension으로 축소된다. 이후 adaptive trimming을 통해 이를 2×3 또는 3×2로 추가 줄일 수 있다. 이에 따라 원래 N=256 차원에서 수행되던 training과 detection을 훨씬 작은 D 차원으로 축소할 수 있으며, channel estimation의 p와 R 계산은 각각 reduced dimension에서 진행되고 weight w=R^-1p 계산의 O(D^3) 부담도 크게 낮아진다.

이 접근의 의미는 단순히 FFT를 앞단에 추가하는 것이 아니라, 이후의 matrix inversion과

vector-matrix multiplication이 다루는 problem size 자체를 축소한다는 데 있다. 논문은 full antenna-space baseline과 비교했을 때 2D beamspace transform이 training area를 약 3 orders of magnitude, detection area를 약 1.5 orders of magnitude 줄일 수 있음을 보인다. 즉, beamspace는 통신 성능을 유지하면서도 후단 digital baseband가 실제로 처리해야 하는 state dimension을 줄여주는 algorithm-hardware co-design의 핵심 수단으로 사용된다.

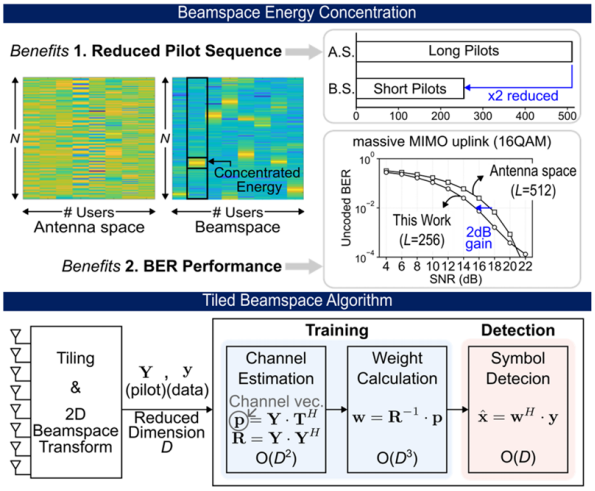
Reduced dimension을 얻은 뒤에는 multi-user channel estimation을 효율화하기 위한 추가 최적화가 적용된다. Covariance matrix R은 Hermitian symmetric이므로 저자들은 lower triangular element만 계산하여 estimator area를 44% 줄였고, ±1 Walsh-Hadamard pilot sequence를 이용하는 p estimator에서는 일반 multiplier를 mux-add 구조로 대체하여 추가적으로 54%의 area를 절감하였다. 또한 3×3 beam window를 항상 유지하지 않고 p의 energy distribution을 이용해 2×3 또는 3×2로 trimming하는 adaptive windowing을 적용한다. 이 방식은 fixed 3×3 window 대비 area를 34% 줄이면서, 단순 fixed 2×3 trimming과 비교하면 dominant beam 방향을 따라 window를 선택할 수 있어 약 1.8 dB의 SNR 이득을 제공한다.



[그림 1] mmWave massive MIMO 처리 문제와 proposed tiled beamspace architecture

그림 1은 저자들이 해결하려는 scalability 문제와 이를 위한 tiled beamspace 구조를 보여준다. 기존 antenna-space 기반 monolithic architecture에서는 수백 개 antenna의 데이터를 하나의 hub DSP로 모아 처리해야 하므로 routing과 matrix-processing complexity가 array 크기에 따라 빠르게 증가한다. 제안 구조는 256개 antenna를 16개의 tile로 분할하고, 각 tile의 local DSP에서 먼저 spatial FFT를 수행한 뒤 hub DSP에서 inter-tile processing을 이어가는 방식으로 timing-critical data movement를 지역화한다. 이처럼 physical array tiling과 beamspace 변환을 함께 설계함으로써 단일 대형 DSP에 집중되는 데이터 경로를 분산시키고, 후단 연산이 처리해야 하는 유효 dimension도 함께 줄인다. 특히 channel estimation뿐 아니라 weight calculation과 symbol detection까지 on-chip에

통합하여 training과 detection을 하나의 scalable dataflow로 연결한 점이 기존 detector-only 구조와 구별된다.



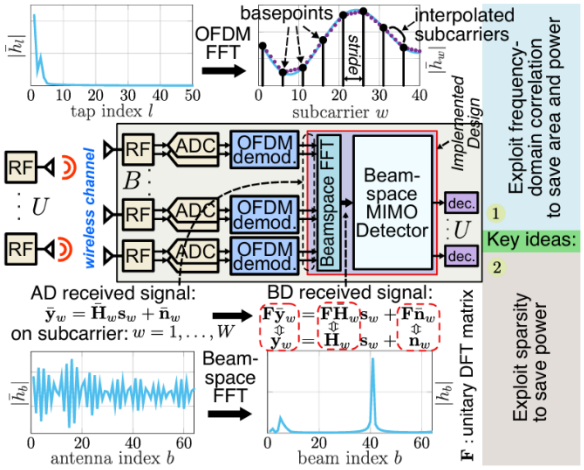
[그림 2] beamspace energy concentration과 reduced-dimension training/detection algorithm

그림 2에서는 beamspace 변환이 실제 algorithm complexity 감소로 이어지는 과정을 확인할 수 있다. Antenna-space에서는 signal energy가 넓은 antenna dimension에 분산되지만 spatial FFT 이후에는 소수 beam bin에 집중되므로, 더 짧은 pilot sequence로도 유효 channel 정보를 확보할 수 있고 BER 측면에서도 유리하다. 해당 논문에서는 intra-tile FFT 후 각 tile에서 W=4 window를 선택하고, hub DSP의 두 번째 inter-tile FFT를 통해 최종 2D beam map에서 W1×W2=3×3, 즉 D=9의 작은 영역만 남긴다. 이에 따라 channel estimation의 p와 R 계산, O(D^3)의 weight calculation, 그리고 payload symbol detection을 모두 reduced dimension에서 수행할 수 있다. 논문이 제시한 결과에서는 full antenna-space processing 대비 training area가 약 3 orders of magnitude, detection area가 약 1.5 orders of magnitude 감소하며, beamspace의 energy concentration 자체가 후단 hardware 규모를 줄이는 핵심 역할을 한다.

이 reduced-dimensional training flow에서는 p estimator와 covariance R calculator가 pilot sequence를 병렬 처리한다. R은 Hermitian symmetric이므로 lower triangular element만 계산하여 estimator area를 44% 줄이고, ±1 Walsh-Hadamard pilot을 이용하는 p estimator에서는 multiplier를 mux-add 구조로 대체하여 추가적으로 54%의 area를 절감한다. 또한 그림 2의 3×3 window를 항상 고정하지 않고 p의 energy distribution을 이용해 2×3 또는 3×2로 trimming하는 adaptive windowing을 적용한다. 이 방식은 fixed 3×3 window 대비 area를 34% 줄이면서 fixed 2×3 trimming보다 약 1.8 dB 높은 SNR을 제공한다. 이후 가장 큰 hardware bottleneck인 weight calculation은 R=LDL^T decomposition, triangular matrix inversion L^-1, inverse matrix assembly R^-1=(L^-1)^T D^-1 L^-1, WCAL의 네 단계로 구성된다. 저자들은 stage 사이에 전체 matrix를 저장하는 alignment buffer

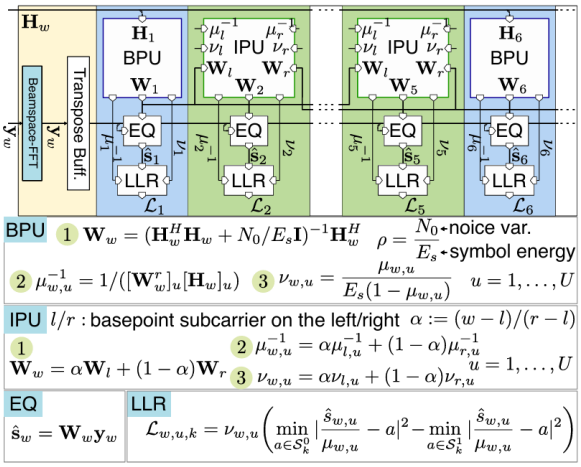
를 없애고 LDL-INVL-INVR-WCAL을 직접 연결하는 linked systolic dataflow를 사용하여 intermediate storage와 stall을 동시에 줄였다.
이 linked systolic array는 단순한 PE count 감소보다 data alignment와 intermediate storage를 없애는 데 초점이 있다. LDL-INVL direct linking만으로 baseline 대비 area 23%, latency 22%를 줄였으며, INVR와 WCAL까지 포함한 end-to-end linked organization에서는 baseline 대비 area 66%, power 76% 감소와 20% latency 개선을 달성하였다. 즉, beamspace가 algorithm dimension을 줄이고 linked systolic dataflow가 그 작은 matrix를 불필요한 buffering 없이 처리하도록 구성되어, algorithmic sparsity와 architecture-level streaming이 서로 보완되는 형태이다.
측정 칩은 Intel16 공정에서 3.17 mm^2의 core area로 제작되었으며 16개의 tile DSP와 hub DSP를 포함해 256×8 uplink를 지원한다. 0.9 V, 550 MHz에서 16QAM detection throughput 17.6 Gb/s를 달성하고 687.6 mW를 소비하여 39.1 pJ/b의 energy efficiency를 보였다. 8명의 user에 대한 channel estimation latency는 461 ns이며, 기존 beamspace ASIC 대비 더 많은 antenna를 지원하면서 weight calculation까지 통합한다. 저자들이 비교한 최근 MIMO detector 대비 detection throughput은 최대 9.76배 높고, abstract 기준으로 area efficiency는 최대 6배 높으며 training latency는 1.78배 낮다. 이 결과는 2D beamspace와 tiled implementation이 대규모 array의 routing/scaling 문제까지 함께 다룬다는 점에서 의미가 있다.

#27-2 본 논문은 massive MU-MIMO-OFDM uplink에서 여러 subcarrier를 동시에 처리하는 soft-output detector를 제안한다. OFDM에서는 수천 개의 subcarrier마다 equalization을 수행해야 하므로 각 subcarrier에 완전한 preprocessing hardware를 복제하면 area와 power가 커지고, 반대로 serial processing을 택하면 latency와 buffering이 증가한다. 저자들은 이 문제를 해결하기 위해 channel impulse response가 소수 dominant tap에 집중된다는 frequency-domain correlation과, mmWave channel이 beamspace에서 sparse하다는 spatial property를 동시에 이용한다. 즉, 모든 subcarrier에서 LMMSE equalization matrix를 독립적으로 계산하지 않고 일부 basepoint에서만 정확한 preprocessing을 수행한 뒤 나머지는 interpolation으로 생성한다.



[그림 3] massive MU-MIMO-OFDM uplink와 frequency-domain correlation/beamspace sparsity 활용 개념

그림 3은 해당 논문이 동시에 활용하는 두 종류의 redundancy를 정리한다. 첫째, channel impulse response가 소수 dominant tap에 집중되기 때문에 인접 OFDM subcarrier의 channel과 equalization matrix가 높은 frequency-domain correlation을 갖는다. 따라서 모든 subcarrier에서 LMMSE preprocessing을 반복하는 대신 일정 stride마다 basepoint만 직접 계산하고 중간 subcarrier의 parameter는 interpolation으로 생성할 수 있다. 둘째, mmWave channel은 beamspace에서 sparse하므로 antenna-domain received signal을 spatial FFT로 변환하면 signal과 equalization matrix의 energy가 소수 beam에 집중된다. 저자들은 첫 번째 특성을 이용해 preprocessing hardware의 복제 비용을 줄이고, 두 번째 특성을 이용해 실제 MAC switching activity를 줄이는 방식으로 area와 power를 각각 다른 경로에서 절감한다.



[그림 4] 16-subcarrier soft-output detector top-level architecture와 BPU/IPU/EQ/LLR dataflow

그림 4의 fabricated ASIC은 16개의 parallel soft-output detector core를 두어 16개 subcarrier를 동시에 처리한다. 예시에서는 subcarrier 1, 6, 11, 16이 basepoint이며, 각

basepoint의 BPU가 LMMSE equalization matrix W와 LLR 계산에 필요한 inverse bias term 및 post-equalization SINR을 생성한다. 그 사이 non-basepoint subcarrier는 IPU가 좌우 basepoint의 parameter를 위치에 따라 linear interpolation한다. 이후 EQ가 symbol estimate를 계산하고 LLR unit이 max-log approximation 기반의 soft output을 생성한다. 즉 expensive matrix preprocessing은 일부 basepoint에만 집중시키면서도 모든 subcarrier에서 soft-output detection을 유지하도록 architecture가 구성되어 있으며, IPU가 matrix를 column-by-column으로 처리하기 때문에 interpolated subcarrier를 위한 full matrix storage도 피할 수 있다.

BPU의 중심은 U×(U+1)/2개의 PE로 구성된 triangular systolic array이며 U=8이므로 36개의 PE가 Gram computation, Cholesky decomposition, forward substitution, backward substitution의 네 phase를 순차적으로 수행한다. Phase마다 별도 accelerator를 두지 않고 동일 PE에 여러 data port와 multiplexer를 배치하여 hardware를 재사용하며 controller가 dataflow를 전환한다. Beamspace sparsity는 EQ와 BPU 내부의 signal silencing으로 실제 dynamic power 절감에 연결된다. 두 input magnitude가 threshold 이하인 complex MAC에서는 register update를 막아 연산을 skip하고, beamspace 신호의 증가된 fixed-point bit-width 문제는 variable-point(VP) number format과 FXP2VP/VP2FXP conversion을 통해 완화한다. VP 적용만으로 fully fixed-point design 대비 약 8%의 area가 줄어든다.

Implementation은 GlobalFoundries 22FDX에서 이루어졌고 core cell area는 1.3 mm^2이다. 64×8 MIMO, QPSK부터 256-QAM까지 지원하며 0.8 V에서 286 mW를 소비하면서 7.3 Gb/s throughput과 39 pJ/b의 energy efficiency를 달성한다. Nominal 0.8 V에서 최대 clock frequency는 약 755 MHz이고, forward body bias를 사용하면 900 MHz 이상으로 상승해 8.8 Gb/s를 넘는 throughput을 얻을 수 있으나 leakage power가 증가하는 trade-off가 있다. Table I에서 preprocessing을 포함하는 비교 대상들 중에서는 throughput과 energy efficiency가 가장 우수한 수준으로 보고된다.

정확도 측면에서 linear interpolation은 단순한 근사로 끝나지 않는다. 측정된 coded BLER는 antenna-domain(AD) mode와 beamspace-domain(BD) mode 모두 floating-point brute-force baseline과 대부분 유사하며, 가장 어려운 non-LoS 256-QAM 조건에서만 5G target인 10% BLER 지점에서 약 1 dB의 손실이 나타난다. Power 측정에서는 모든 subcarrier를 basepoint로 처리하는 brute-force baseline 대비 interpolation이 2.1배 power reduction과 3배 area reduction을 제공하고, beamspace sparsity를 이용한 silencing이 추가로 최대 1.4배 power reduction을 제공하여 전체적으로 약 3배의 power 절감 효과를 만든다. 즉, frequency-domain redundancy는 preprocessing 복제 비용을 줄이고 spatial sparsity는 실제 MAC switching activity를 줄이는 서로 다른 역할을 수행한다.

저자정보



이재성 박사과정 대학원생

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

2026 VLSI Review

고려대학교 전기전자공학과 석사과정 심승우

Session 1 High-Frequency Arrays for 6G and SATCOM

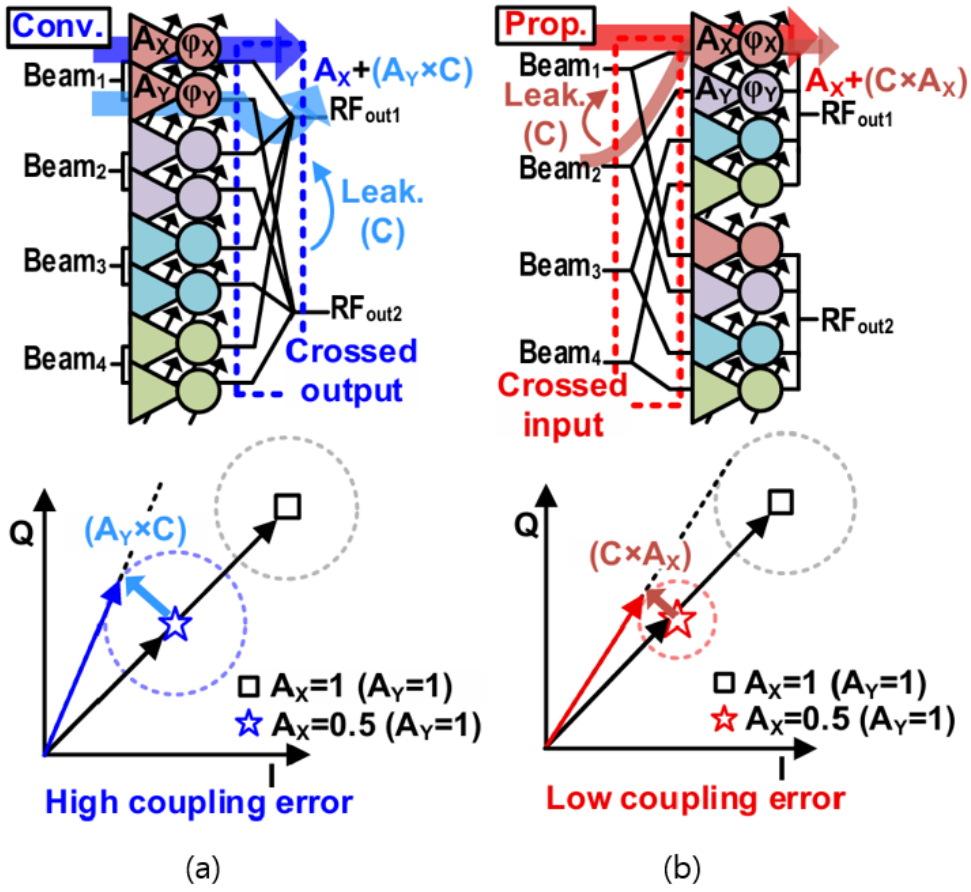
이번 2026 IEEE VLSI에서는 Session 1 "High-Frequency Arrays for 6G and SATCOM"과 Session 14 "Adaptive and Agile Radio Architectures"에서 각각 5편과 4편의 논문이 발표되었습니다. 두 세션에서는 6G 및 SATCOM을 위한 다양한 주파수 대역의 transceiver와 phased-array 기술이 소개되었으며, 이번 글에서는 그중 흥미로운 아이디어를 제안한 Q-band transceiver와 V-band transceiver 논문을 각각 한 편씩 소개하고자 합니다.

#1-1 이번에 소개할 논문은 전남대학교 박진석 교수님 연구팀에서 발표한 "An Ultra-Compact 4-Element 4-Beam Q-band Multibeamformer IC with IQ Pre-weighted Phase Shifter for VLEO SATCOM"입니다. 최근 low-latency SATCOM에 대한 수요가 증가하면서 very low earth orbit(VLEO) 위성통신이 주목받고 있으며, 높은 data capacity 확보를 위한 Q/V-band 활용도 활발히 연구되고 있습니다. 본 논문에서는 이러한 VLEO SATCOM을 위한 4-element 4-beam Q-band multibeamformer IC를 제안하였습니다.

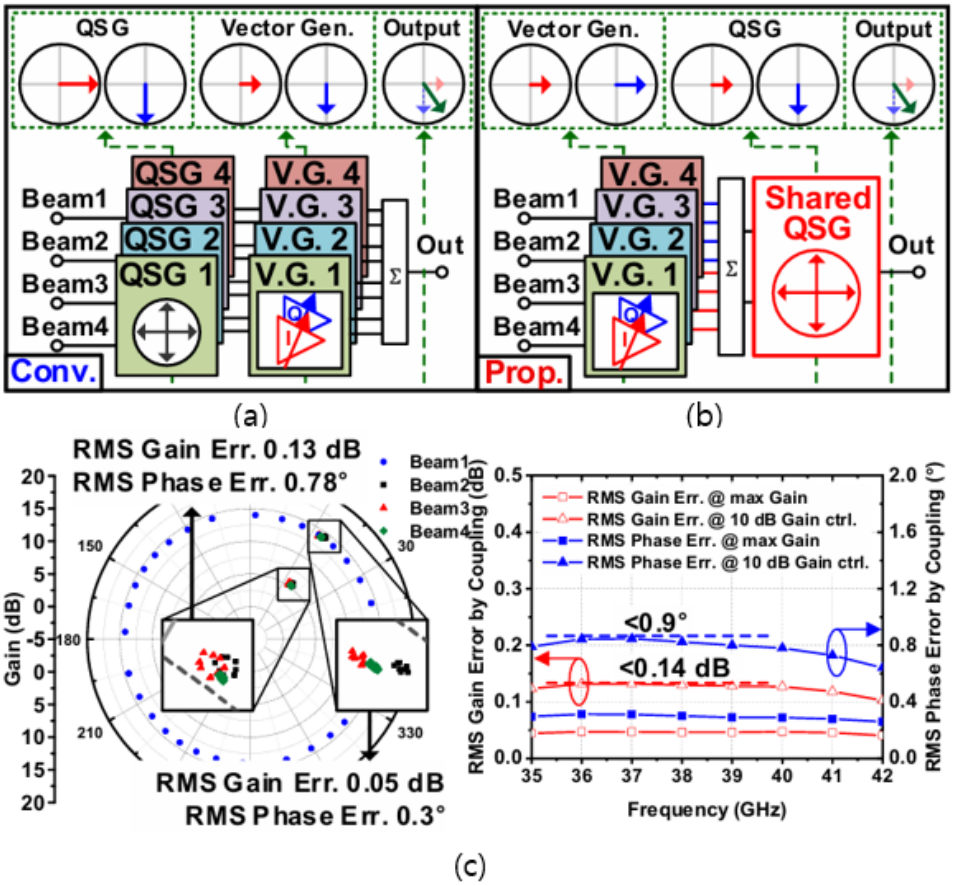
Fig. 1(a)의 conventional fully-connected multibeamformer에서는 divider/combiner network가 gain/phase control 이후에 배치됩니다. 이 경우 다른 beam에서 coupling된 신호가 desired signal과 서로 다른 gain/phase weighting을 받기 때문에, beam의 gain을 조절할수록 coupling에 의한 gain/phase error가 증가하게 됩니다. 이를 해결하기 위해 본 논문에서는 Fig. 1(b)와 같이 divider/combiner network와 gain/phase control block의 순서를 변경하였습니다. 이에 따라 desired signal과 coupled signal이 동일한 gain/phase control을 받게 되어, gain control에 따른 coupling error를 효과적으로 줄일 수 있습니다. 하지만 이러한 구조에서는 하나의 antenna element가 여러 beam 신호를 동시에 처리해야 하므로 multi-beam phase shifter가 필요합니다. Fig. 2(a)의 conventional phase shifter에서는 각 beam마다 별도의 quadrature signal generator(QSG)가 필요하기 때문에 beam 수가 증가할수록 회로 면적 또한 크게 증가합니다. 이를 해결하기 위해 본 논문에서는 IQ pre-weighted phase shifter를 제안하였습니다. Fig. 2(b)와 같이 기존의 QSG와 vector generator의 순서를 변경하여 vector generator를 먼저 배치하고 하나의 QSG를 공유하도록 구성하였습니다. 각 beam에서 I/Q 성분이 될 신호의 amplitude를 먼저 pre-weighting한 뒤 shared QSG에서 quadrature phase를 생성함으로써, multi-beam phase shifter를 보다 compact하게 구현할 수 있습니다.

즉, 이 논문의 핵심은 beam-to-beam coupling을 줄이기 위해 divider/combiner network

와 gain/phase control의 배치 순서를 변경하고, 이로 인해 발생하는 multi-beam phase shifter의 area penalty를 IQ pre-weighting 기반의 shared-QSG 구조로 해결한 것입니다. 제안된 구조는 simulation에서 10-dB gain control 시에도 coupling에 의한 RMS gain 및 phase error를 각각 0.14 dB와 0.9° 이하로 유지하였습니다. 특히 기존 회로의 성능을 단순히 개선하는 것이 아니라, block 배치 자체를 재구성하고 그에 따른 penalty를 새로운 phase-shifter architecture로 해결했다는 점에서 흥미로운 접근이라고 볼 수 있습니다.



[Fig 1] Coupling error accrording to the location of the multibeam divider/combiner in a multibeamformin IC (a) Conventional, (b) Proposed structure.



[Fig 2] (a) Conventional PS structure, (b) Proposed IQ pre-weighted PS structure, (c) Simulated results about coupling effect.

저자정보



심승우 석사과정 대학원생

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

2026 VLSI Review

고려대학교 전기전자공학과 석사과정 심승우

Session 14 Adaptive and Agile Radio Architectures

이번 2026 IEEE VLSI에서는 Session 1 "High-Frequency Arrays for 6G and SATCOM"과 Session 14 "Adaptive and Agile Radio Architectures"에서 각각 5편과 4편의 논문이 발표되었습니다. 두 세션에서는 6G 및 SATCOM을 위한 다양한 주파수 대역의 transceiver와 phased-array 기술이 소개되었으며, 이번 글에서는 그중 흥미로운 아이디어를 제안한 Q-band transceiver와 V-band transceiver 논문을 각각 한 편씩 소개하고자 합니다.

#14.3 이번에 소개할 논문은 Institute of Science Tokyo의 Kenichi Okada 교수님 연구팀에서 발표한 "A 57–71-GHz CMOS Phased-Array Transceiver with Aperture-Tuning Antenna Achieving 47.9–62.2% EIRP Efficiency Improvement"입니다. 5G NR FR2-2의 n263 band(57–71 GHz) 는 넓은 frequency coverage를 요구하지만, antenna와 transceiver가 전 대역에서 항상 높은 성능을 유지하도록 broadband하게 설계하는 것은 쉽지 않습니다. 일반적으로 높은 EIRP를 확보하기 위해 PA output power를 증가시킬 수 있지만, 이 경우 DC power consumption이 증가하고 EIRP efficiency가 낮아질 수 있습니다.

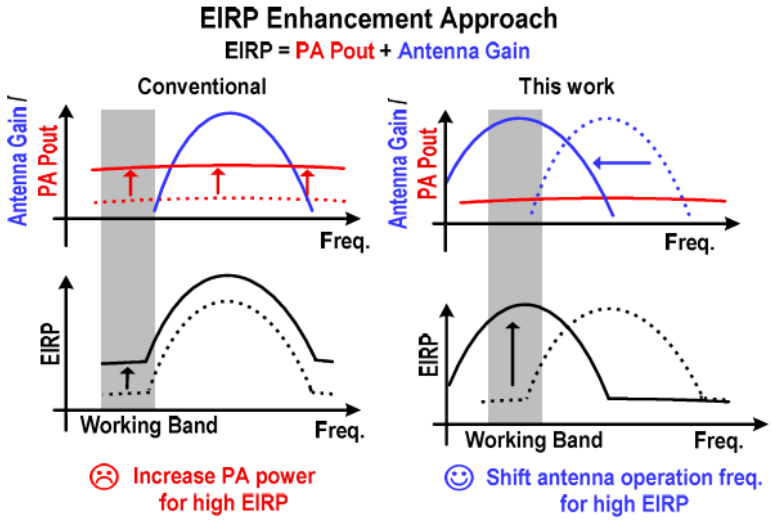
Fig. 3(a)의 conventional approach에서는 고정된 antenna gain characteristic에서 높은 EIRP를 얻기 위해 PA output power를 증가시키는 반면, Fig. 3(b)의 제안된 구조에서는 distributed tuner의 impedance를 조절하여 antenna의 current distribution과 effective radiation aperture를 변화시킵니다. 이를 통해 antenna gain의 peak frequency를 원하는 operating band로 이동시켜, PA output power를 추가로 증가시키지 않고도 높은 EIRP를 확보할 수 있는 reconfigurable aperture-tuning 방식을 제안하였습니다.

제안된 aperture tuner는 단순히 antenna 외부에 추가되는 구조가 아니라, PA와 LNA의 matching network 내부에 포함된 IC–antenna co-design 형태로 구현되었습니다. 이를 통해 tuner가 impedance transformation의 일부로 동작하는 동시에 antenna의 radiation mode를 제어할 수 있도록 하였습니다. 즉, conventional feed-port impedance tuning을 넘어 RF front-end와 antenna를 함께 설계하여 operating frequency에 따라 antenna의 radiation characteristic을 reconfigure할 수 있도록 하였습니다.

Fig. 4는 이러한 aperture tuning의 측정 결과를 보여줍니다. Fig. 4(a)와 (b)의 measured beam pattern에서 57 GHz와 71 GHz 모두 main beam 방향을 안정적으로 유지하는 것을 확인할 수 있으며, 각 frequency에 적합한 aperture-tuning mode를 통해 높은 antenna

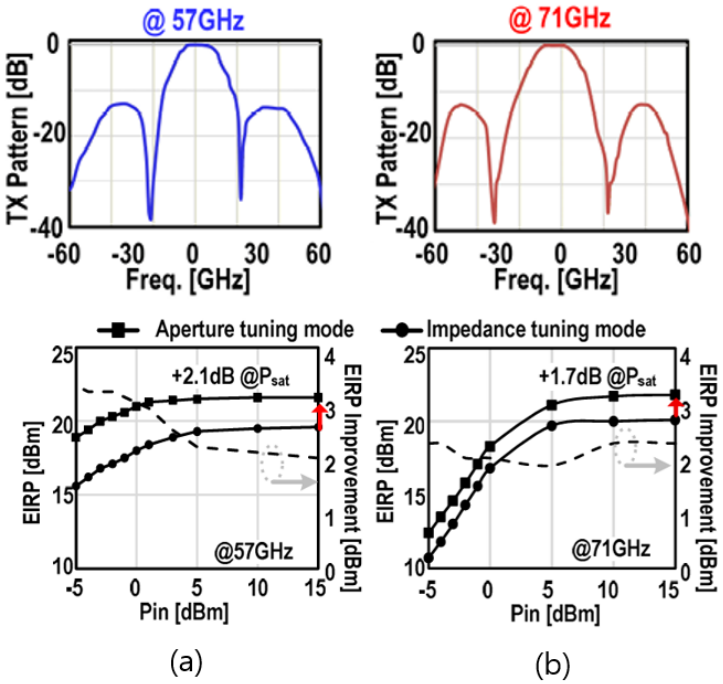
gain을 확보합니다. 또한 measured EIRP 결과에서는 conventional impedance-tuning mode 대비 57 GHz에서 2.1 dB, 71 GHz에서 1.7 dB의 EIRP improvement를 달성하였습니다. 이는 각각 62.2%와 47.9%의 EIRP efficiency improvement에 해당하며, additional power consumption 없이 얻어진 결과입니다.

즉, 이 논문의 핵심은 넓은 frequency coverage를 RF circuit과 antenna의 broadband 설계만으로 해결하는 대신, aperture tuning을 통해 antenna gain peak를 현재 사용하는 frequency band에 맞게 이동시켜 system-level EIRP를 높인 것입니다. 특히 PA output power를 추가로 증가시키는 conventional한 접근에서 벗어나, antenna의 radiation characteristic 자체를 reconfigure하여 EIRP efficiency를 향상시켰다는 점에서 흥미로운 접근이라고 볼 수 있습니다.



	Conventional	This work
PA Pout	Increased	Invariant
Antenna Gain	Fixed	Shifted
EIRP	Increased	Increased
EIRP Efficiency	Degraded	Improved

[Fig 3] Operating principle of the proposed aperture-tuning antenna compared with the conventional approach.



[Fig 4] measured radiation patterns and EIRP performance at different frequencies.

저자정보



심승우 석사과정 대학원생

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

2026 VLSI Review

KAIST 전기및전자공학부 석박사통합과정 이동윤

Session 26 UWB, IoT, RF Enablers

2026 IEEE Symposium on VLSI Technology and Circuits의 Session 26은 UWB, IoT, RF Enablers라는 주제로 총 4편의 논문이 발표되었다. 이 세션의 논문들은 저전력 무선 연결과 그 구성 회로를 다루며, 스마트폰이나 wearable 기기와 직접 통신해야 하는 IoT 태그, VR/AR 기기와 뇌-컴퓨터 인터페이스 (BCI, brain-computer interface)를 위한 초저전력 고속 링크, 차량용 UWB 레이더처럼 기존 무선 규격 위에 새로운 기능을 얹는 응용까지 폭이 넓다. 26-3은 switched-capacitor 디지털 PA의 캐패시터 어레이 사이에 인덕터를 삽입해 고조파 충전 전류를 억제하는 SLCPA 구조로, 6 dB power back-off에서 Class-B 대비 1.29배의 효율을 얻고 DPD 없이 100 MSymbol/s 256-QAM (800 Mb/s) 신호를 지원한다. 26-4는 IEEE 802.15.4z 호환 IR-UWB SoC에 FIR 필터 기반 self-interference cancellation (평균 27 dB)을 넣어 TX/RX가 안테나를 공유하는 full-duplex 레이더 동작을 가능하게 하고, 실제 차량에서 아동 감지 (child presence detection)를 시연했다. 본 기사에서는 이 가운데 기존 변조 방식과 통신 구조의 전제 자체를 바꾼 두 편, 즉 펄스의 상승 에지를 위상 기준으로 삼아 PLL과 preamble 없이 PSK를 IR-UWB에 결합한 26-1과, 급전 주파수와 backscatter 주파수를 다른 대역으로 분리해 self-jammer를 제거하고 배터리 없는 AoA 위치 추정과 Wi-Fi 표준 mask 준수를 동시에 달성한 26-2에 집중하여 서술한다.

#26-1 VR/AR wearable과 implantable BCI는 수 Gb/s의 데이터율을 수 mW 이하의 전력으로 전송해야 하는 상충된 요구를 갖는다. Impulse-radio UWB (IR-UWB)는 짧은 펄스만 간헐적으로 내보내는 방식이라 이 조건에 적합하며, 최근에는 펄스 하나에 pulse-position modulation (PPM), pulse-width modulation (PWM), frequency-shift keying (FSK) 등 여러 변조를 동시에 실어 변조 차수를 높이는 hybrid modulation이 데이터율과 에너지 효율을 함께 끌어올리는 방향으로 자리 잡았다. 그런데 변조 방식 가운데 phase-shift keying (PSK)만은 hybrid에 결합하기 어려웠다. 기존 PSK는 연속파 시스템을 전제로 만들어져 위상 기준을 유지하기 위한 PLL과 상시 동작하는 LO, 그리고 preamble 기반 동기가 필요하고, 펄스가 간헐적으로만 존재하는 IR-UWB 환경에서는 펄스 사이에 위상 잡음이 누적되어 동기가 자주 깨진다. 게다가 위상이 바뀌면 펄스 파형 자체가 달라지므로 PPM이나 PWM과 직교하게 결합하기도 어렵다.
논문이 제안하는 edge-reference differential PSK (ER-DPSK)는 이 문제를 펄스 자체에 기준을 심는 방식으로 푼다. 각 펄스는 먼저 동일한 형태의 기준 에지 (reference edge)를

만들고, 그 뒤에 원하는 위상만큼 지연된 꼬리를 붙인다. 하나의 펄스 안에 기준 정보와 위상 정보가 함께 담기므로 별도의 위상 기준을 외부에서 유지할 필요가 없다. 에지의 타이밍 정확도가 곧 성능이므로 펄스는 급한 상승과 완만한 꼬리를 갖는 비대칭 형태로 설계되어 에지 위치가 안정적으로 잡히도록 했다. 데이터는 인접한 두 펄스의 위상 차이에 실리는 차동 구조라, 에지 검출 과정에서 생기는 타이밍 오프셋이 자동으로 상쇄되고 펄스마다 위상이 새로 정렬되므로 위상 잡음이 누적되지 않는다. 기준 에지와 포락선이 위상과 무관하게 동일하다는 점 때문에 PPM, PWM과의 구조적 호환성도 확보되어, PSK를 hybrid modulation의 한 축으로 넣을 수 있게 된다.

구현에서는 ER-DPSK를 직교 위상 4개를 쓰는 ER-DQPSK로 만들어 펄스당 2비트를 싣고, 여기에 digitalized multi-pulse-position modulation (DMPPM), PWM, FSK를 결합했다. 250 MHz 클럭 한 주기 (4 ns)에 ER-DQPSK 4비트, DMPPM 3비트, PWM 4비트, FSK 2비트, 합계 13비트가 실려 3.25 Gb/s가 된다. 핵심 회로는 ring oscillator를 pre-charge하여 위상 노드를 정렬해 두었다가 먼저 기준 에지를 내보내고 그다음 선택된 위상의 꼬리가 이어지도록 하는 ER-DQPSK 컨트롤러와 pulse shaper이며, 어떤 위상을 고르든 항상 같은 기준 에지가 나오도록 보장한다.

[그림 1]은 기존 UWB PSK와 제안 ER-DPSK의 개념을 비교한 것이다. 기존 방식에서는 펄스 열 안에 무엇이 0°인지 알려 주는 기준이 없어 수신단이 preamble과 PLL에 의존해야 하지만, 제안 방식에서는 모든 펄스가 자신의 기준 에지를 갖고 그 뒤에 0°, 90°, 180°, 270°의 위상 정보가 따라온다. 수신단은 에지를 검출하고 그 에지에 맞춰 하향 변환한 뒤 인접 펄스 간 위상 차이를 미분하는 것만으로 데이터를 복원할 수 있다.

송신기는 28 nm FD-SOI로 구현되어 0.018 mm^2의 면적을 차지하며, 6.1~8.6 GHz 대역에서 FCC spectral mask를 만족한다. 3.25 Gb/s에서 소비 전력은 2.41 mW로 에너지 효율은 0.74 pJ/bit이고, 논문은 이를 3 Gb/s급 IR-UWB 송신기로는 처음으로 1 pJ/bit 아래로 내려간 결과로 제시한다. 비교 대상인 최근 hybrid IR-UWB 송신기들이 1.66~2.54 Gb/s에서 1.2~5.8 pJ/bit인 것과 대비된다. 3 dBi 안테나와 오실로스코프로 무선 측정한 결과 3.25 Gb/s에서 BER 10^{-3} 기준 50 cm의 통신 거리를 얻었고, 에지 검출 문턱을 피크 진폭의 5% 만큼 바꿀 때 비대칭 펄스의 에지 타이밍 편차는 DMPPM에서 4.7 ps, PWM에서 9.7 ps로 삼각 펄스 (14.0 ps, 29.0 ps)의 약 3분의 1이었다. 연속파를 전제로 한 PSK를 펄스 기반 시스템에 맞게 다시 정의하여 PLL과 preamble을 없앤 발상은, BCI와 wearable을 위한 초저전력 고속 링크에서 변조 차수를 높이는 새로운 길을 열었다고 볼 수 있다.

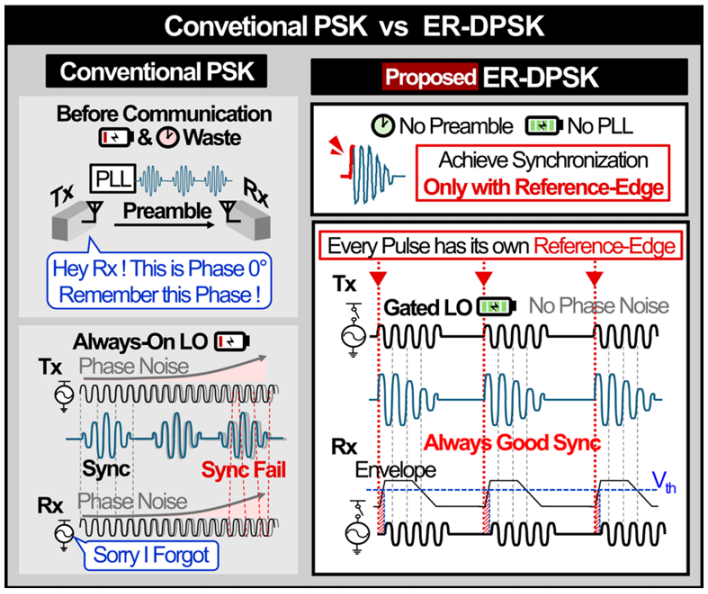


Fig. 3. Conventional PSK vs. ER-DPSK: Preamble-Based PLL Synchronization with Phase Drift vs. Reference-Edge-Based Per-Pulse Synchronization, Providing Always-Good Synchronization

[그림 1] 논문 26-1에서 제안하는 edge-reference DPSK의 개념: 위상 기준이 없는 기존 UWB PSK와, 모든 펄스가 기준 에지와 위상 정보를 함께 갖는 ER-DPSK의 비교

#26-2 스마트폰과 직접 통신하는 배터리 없는 IoT 태그는 두 가지 방향으로 연구되어 왔다. 무선 전력으로 충전한 뒤 능동 송신기로 표준 패킷을 보내는 방식은 Wi-Fi/BLE 표준을 그대로 만족하지만 mW급 전력 때문에 충전 시간이 길어 패킷 전송률이 매우 낮고, 들어오는 반송파를 반사시켜 변조하는 backscatter 방식은 μW급 전력으로 높은 패킷 전송률을 내지만 세 가지 한계를 안고 있었다. 반사 신호와 급전 반송파가 같은 대역에 있어 강한 self-jammer가 수신 품질을 떨어뜨리고, 위치 추정은 multipath에 취약한 RSSI에만 의존하며, 반사기 스위치 어레이의 비선형성 때문에 스펙트럼이 넓게 퍼져 BLE mask는 만족해도 Wi-Fi 표준 mask는 만족하기 어려웠다.

논문의 첫 번째 기법은 급전과 backscatter를 다른 대역으로 분리하는 cross-band multi-carrier fusion이다. 태그는 915 MHz ISM 대역의 급전 신호로 에너지를 수확하고, 같은 신호에 실린 5G-NR 포락선을 Manchester 디코딩하여 채널 할당과 태그 ID를 받는다. 업링크에서는 수확한 915 MHz 기준 주파수로부터 sesqui scaler와 carrier regenerator로 2/3 f_{WPT}와 f_{WPT}를 만들고, 이를 baseband와 함께 multi-carrier mixer에 넣어 약 1.5 GHz의 IF 신호를 합성한다. 반사기 어레이가 입사한 915 MHz 반송파를 이 IF로 변조해 반사하면 2.4 GHz 대역에 BLE/Wi-Fi 업링크가 만들어지고, 약 610 MHz에 생기는 image와 불필요한 혼변조 성분은 이중 대역 안테나의 정합 회로가 걸러낸다. 입사파와 반사파가 서로 다른 대역에 놓이므로 고 Q 필터 없이 self-jammer를 제거할 수 있고, 인덕터 같은 큰 수동 소자 없이 단측파대 (SSB) backscatter가 구현되며, 반사 신호가 깨끗해진 덕분에 배터리 없는 태그로는 처음으로 angle-of-arrival (AoA) 위치 추정이 가능해진다.

두 번째 기법은 Wi-Fi 표준 mask를 맞추기 위한 polar-modulated pre-distortion mask shaping이다. baseband 데이터를 디지털로 파형 정형한 뒤 위상 (PM) 성분과 진폭 (AM) 성분으로 나누고, PM 성분은 5/3 f_{WPT} 반송파와 섞어 IF 신호를 만들며, AM 성분은 디지털 영역에서 미리 왜곡을 보상한 뒤 thermometer 코드로 변환해 32비트 선형 반사기 어레이를 구동한다. 반사기 어레이에서 PM과 AM이 다시 합쳐지면 정형된 스펙트럼의 backscatter 신호가 나오고, 이 방식으로 업링크의 고조파가 억제되어 Wi-Fi와 BLE 두 표준의 mask를 모두 만족한다.

[그림 2]는 cross-band multi-carrier fusion의 구성과 효과를 보인다. 915 MHz 급전 톤에서 주파수 복원 경로가 IF 반송파를 만들고 backscatter 경로가 이를 반사하는 구조로, 대역 밖의 급전 톤이 안테나 단에서 거부되어 self-jammer 없이 AoA 위치 추정이 가능해지는 것과, 원치 않는 측파대가 안테나에서 걸러져 전력 분배기 없이 SSB 업링크가 되는 것을 함께 보여 준다.

태그 칩은 65 nm CMOS로 구현되어 0.798 mm^2를 차지한다. 915 MHz 30 dBm 급전원 아래에서 9 m 떨어진 스마트폰이 태그가 backscatter한 Wi-Fi와 BLE 패킷을 수신해 화면에 표시하는 것을 시연했고, 스마트폰 측에서 over-the-air로 측정한 2.4 GHz 반송파의 위상 잡음은 1 MHz 오프셋에서 −111 dBc/Hz였다. cross-band 반사 스펙트럼에서 원하는 측파대는 image보다 32 dB 이상 강해 32 dB 이상의 image rejection을 보였고, 세 개의 태그를 무선 급전한 상태에서 AoA 기지국이 BLE 패킷으로 ±10 cm 이내의 실시간 위치 추정과 이동 추적을 수행했다. polar pre-distortion을 적용하면 backscatter 계수의 INL이 6.70 LSB에서 2.92 LSB로 줄고 Wi-Fi 스펙트럼이 표준 mask 안에 들어온다. 전력은 수확 단계 27 nW, 다운링크 단계 1.2 μW, 업링크 단계 215 μW이며 다운링크 54 kb/s, 업링크 1 Mb/s이다. 비교 대상인 기존 스마트폰 호환 배터리 없는 태그들이 모두 in-band backscatter로 image rejection이 사실상 0 dB이고 self-jammer가 존재하며 Wi-Fi mask를 만족하지 못하거나 검증하지 않은 것과 비교하면, cross-band 구조와 AoA 위치 추정, Wi-Fi mask 준수는 이 논문이 처음 보인 항목들이다. 급전은 스마트폰이 아닌 별도의 915 MHz 급전원이 담당하므로 실제 배치에서는 급전 인프라가 필요하다는 점을 감안해야 하지만, 급전과 통신의 대역을 분리한다는 하나의 결정이 self-jammer, 위치 추정, 표준 mask라는 서로 다른 세 문제를 동시에 풀어낸 점에서 backscatter 태그의 설계 공간을 넓힌 논문이라 볼 수 있다.

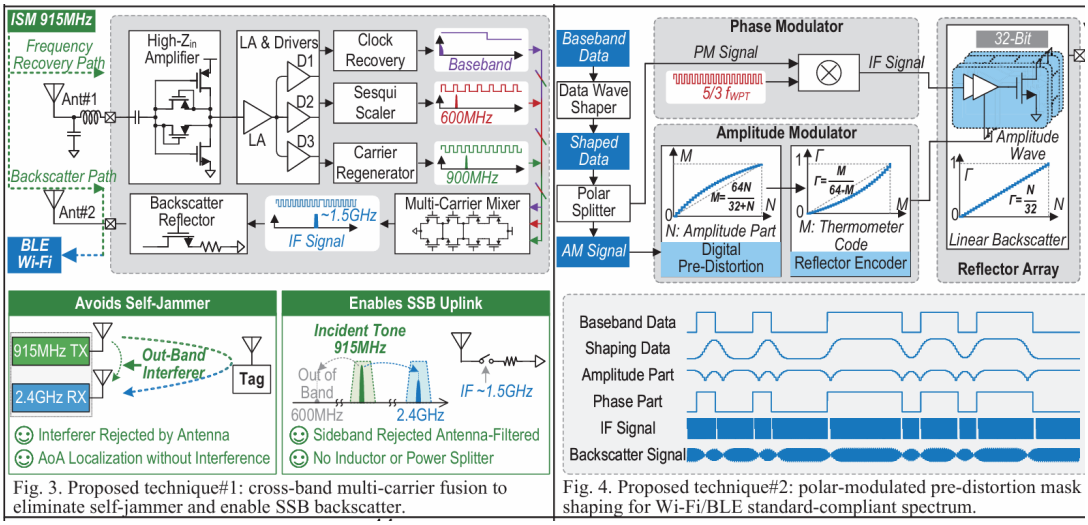


Fig. 3. Proposed technique#1: cross-band multi-carrier fusion to eliminate self-jammer and enable SSB backscatter.

Fig. 4. Proposed technique#2: polar-modulated pre-distortion mask shaping for Wi-Fi/BLE standard-compliant spectrum.

[그림 2] 논문 26-2에서 제안하는 cross-band multi-carrier fusion 기법: 915 MHz 급전 톤과 2.4 GHz backscatter 업링크의 대역 분리를 통한 self-jammer 제거와 SSB 업링크 구현

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

VLSI 2026 Review

한국과학기술원 바이오및뇌공학과 박사과정 석동열

Session C13 In Body Bioelectronic Systems

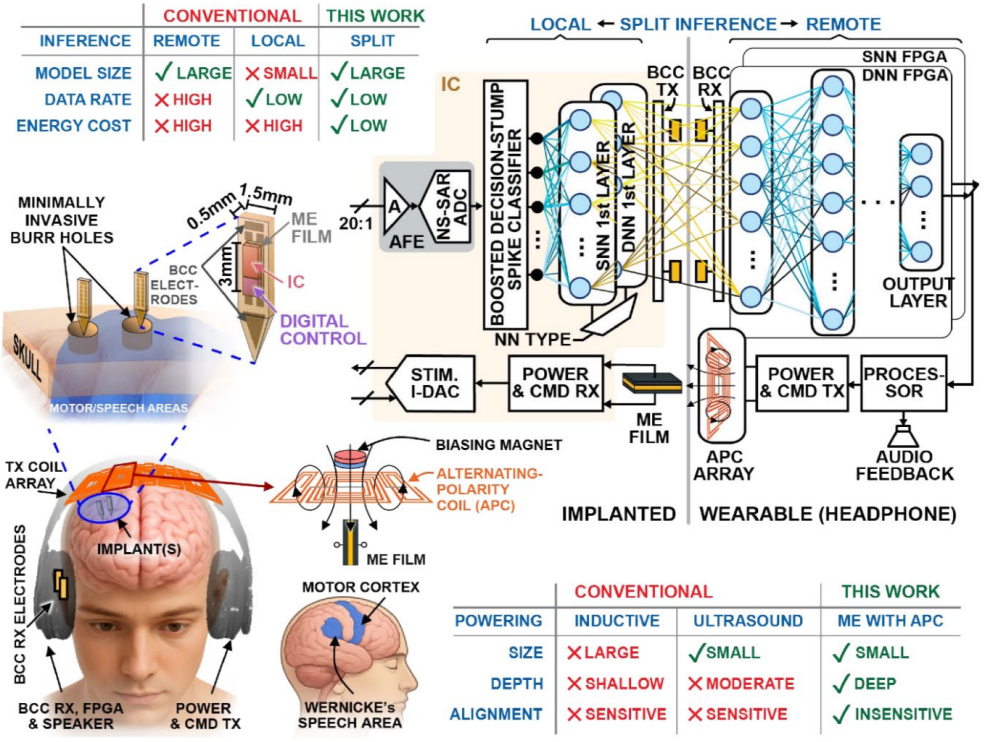
이번 2026 IEEE/JSAP Symposium on VLSI Technology and Circuits의 Session C13, In-Body Bioelectronic Systems에서는 체내 이식형 바이오전자 시스템에 적용되는 최신 회로설계 기술이 소개되었다. 최소침습형 신경 인터페이스를 구현하기 위해 요구되는 초소형화와 초저전력화, 체내 무선 전력 및 데이터 전송, 온칩 인공지능을 활용한 신경신호 처리 등 동시대의 관심사를 반영한 기술들이 발표되어 눈길을 끌고 있다. 특히 기록 채널이 증가할수록 연산량과 송신 데이터가 함께 커지는 문제를 시스템 수준에서 해결하려는 접근이 돋보인다.

#C13.1 본 논문은 토론토 대학교(University of Toronto)와 라이스 대학교(Rice University) 연구진의 공동 연구 결과로, 배터리 없이 구동되는 밀리미터 크기의 뇌 임플란트에서 신경신호 측정과 인공지능 추론을 수행하는 집적회로를 보여준다. 임플란트 내부의 제한된 전력과 통신 대역폭을 고려하여 신경망 연산을 임플란트와 외부 웨어러블 장치로 나누는 분할 추론(split inference) 구조를 제안한 점이 특징이다. 32개의 기록 전극 중 20개 입력을 시분할 방식으로 처리하는 저잡음 아날로그 전단부와 Noise-shaping SAR ADC, 스파이크 분류기, SNN 또는 DNN의 첫 번째 계층, 신체 채널 통신(Body-coupled communication) 송신기를 하나의 칩에 집적하였다. 유연한 마이크로전극에는 양면으로 16×2 전극을 배치하였으며, 고정형 전극보다 조직 움직임에 따른 artifact를 줄이는 구성을 지향한다.

온칩 추론은 지수이동평균(exponential moving average, EMA) 필터로 스파이크 파형의 시간적 특징을 추출한 뒤, 비교기 기반의 boosted decision-stump 모델로 스파이크를 분류하는 방식으로 이루어진다. 두 EMA 필터의 서로 다른 시간상수와 peak 및 trough 임계값, 시간 지연 정보를 활용하므로 곱셈기 없이 shift-and-add 연산과 비교만으로 채널당 최대 4개의 뉴런을 구분할 수 있다. 이어서 SNN의 leaky integrate-and-fire 뉴런 또는 DNN의 첫 번째 계층만 임플란트에서 계산하고, 나머지 계층은 웨어러블 FPGA에서 수행하도록 하여 최대 80개의 생물학적 뉴런에서 얻은 정보를 낮은 연산량으로 처리한다. EMA 시간상수와 임계값은 외부에서 재학습하여 다시 설정할 수 있어 피험자와 과제의 변화에 대응할 수 있다.

통신 전력은 첫 번째 신경망 계층의 활성값이 발생할 때만 짧은 펄스를 보내는 event-based body-coupled communication 방식을 적용하여 줄였다. 여러 임플란트는 TDMA로 웨어러블 수신기를 공유하며 최대 8개의 장치를 지원하고, 연속적인 원시 파형 대신 의미 있는 중간 활성값만 전달한다. 전력 공급에는 저주파 자기장을 전기에너지로 변환하는 3×1×0.3mm³ 크기의 magnetoelectric 필름과 alternating-polarity coil 배열을 사용하였다. 머리 모형 실험에서 20mm 깊이까지 전력을 전달하고 축이 어긋난 경우에도 전력 전달이 가능함을 보여, 배터리와 대형 수신 코일 없이 깊은 피질 영역에 전력을 공급할 수 있는 가능성을 보여주었다.

회로는 65nm 공정으로 구현되었으며, 신경 증폭기는 10.3nV/√Hz의 입력 환산 잡음밀도와 1.5의 NEF를 달성하고, 2차 Noise-shaping SAR ADC는 8kHz 대역에서 15.8bit ENOB와 181dB Schreier FoM을 기록하였다. 기존 데이터셋을 이용한 분할 추론 평가에서는 원숭이의 세 가지 운동 의도를 90.4%의 교차검증 정확도로 분류하면서 데이터율을 1,066배 줄였고, 사람의 8개 내면언어 클래스는 96.1%의 클래스별 정확도와 16,732배의 데이터율 감소를 보였다. 전체 소비전력은 추론 시 입력당 6.07μW이며, 온칩 연산과 통신에 추가되는 비용은 각각 입력당 53nW와 1.23nW이다. 원시 데이터를 모두 외부로 보내거나 모든 계층을 임플란트에서 계산하는 두 극단 사이에서 통신과 연산 에너지를 함께 낮추어, 초소형 임플란트에서도 깊은 신경망 기반 해독을 수행할 수 있음을 보여주었다.



[그림 1] 배터리 없는 밀리미터 크기 뇌 임플란트와 웨어러블 헤드셋으로 구성된 분할 추론 시스템: 임플란트에서는 신경신호 측정, 스파이크 분류 및 SNN/DNN 첫 번째 계층을 처리하고, 희소

한 중간 활성값만 BCC로 전송하여 웨어러블 FPGA에서 나머지 계층을 수행함. ME 필름과 alternating-polarity coil 배열을 이용하여 깊은 피질 영역에 무선 전력을 공급함.

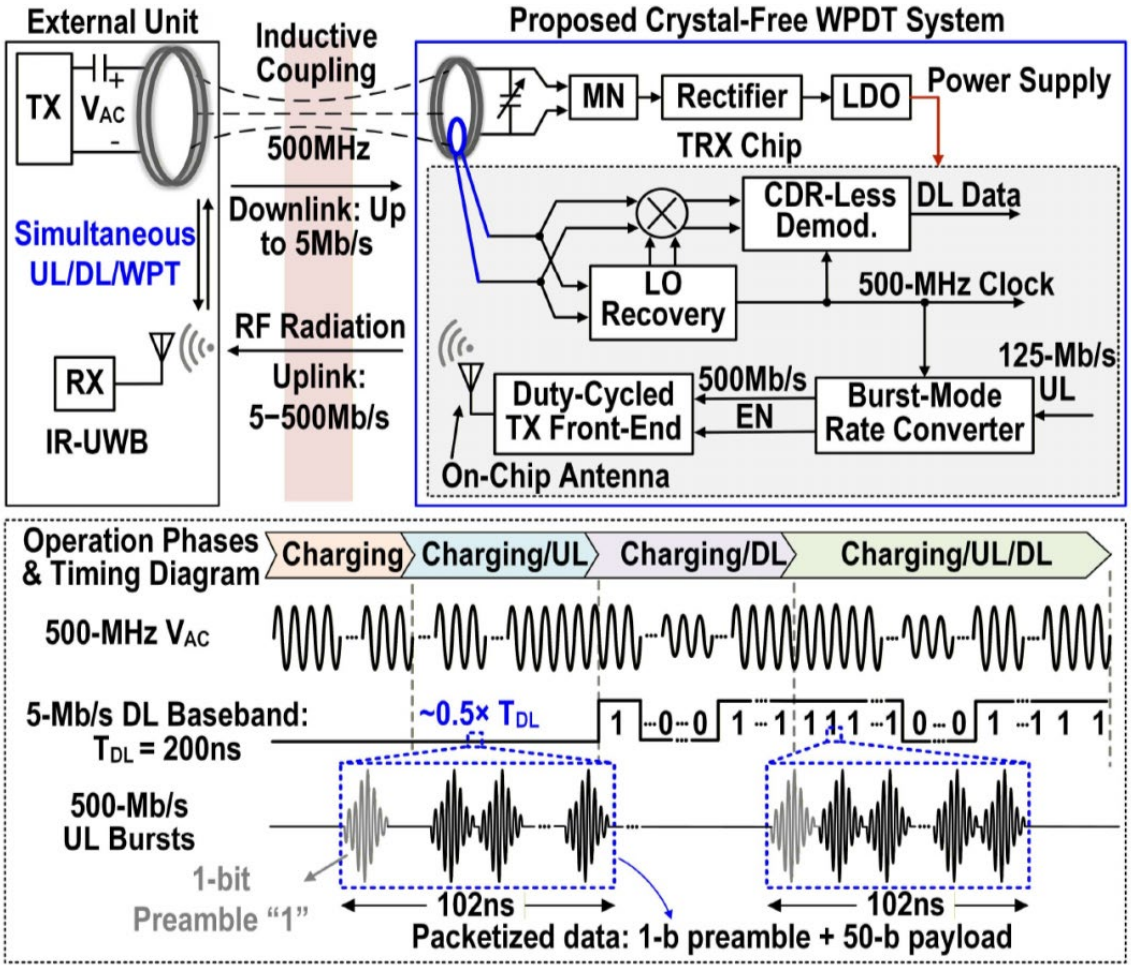
#C13.4 체내 신경 인터페이스는 외부 장치로부터 안정적으로 전력을 공급받는 동시에 제어 명령을 수신하고, 다채널 신경신호를 높은 전송률로 외부에 보내야 한다. 그러나 기존의 저주파 유도결합 방식은 전력전송효율과 데이터율 사이에 trade-off가 있고, 전력과 통신에 서로 다른 주파수를 사용하는 구조는 결합계수와 코일 정렬 변화에 민감하다. 또한 고속 uplink를 위한 별도의 crystal oscillator와 clock-data recovery 회로는 임플란트의 면적과 대기전력을 증가시킨다. 본 세션의 C13.4 논문은 하나의 전력 carrier를 시스템 clock으로 재사용하여 이러한 제약을 함께 해결한다.

본 논문은 스위스 로잔연방공과대학교(EPFL), 취리히연방공과대학교(ETH Zurich), IT'IS Foundation 및 유타 대학교(University of Utah) 연구진의 공동 연구 결과로, 50mm² 면적 안에 loop antenna와 칩을 포함한 crystal-free 양방향 wireless power and data transfer 시스템을 제안한다. 500MHz 전력 carrier를 임플란트의 유일한 기준 clock으로 사용하며, 전력 공급 중에도 5Mb/s ASK downlink와 5~500Mb/s IR-UWB uplink를 동시에 지원한다. 별도의 crystal oscillator 없이 전력 carrier로 downlink 복조 timing과 multi-GHz uplink PLL을 모두 동기화한 것이 핵심이다.

전력 수신과 downlink 데이터 복조를 분리하기 위해 6mm 외곽 RX loop와 그 내부의 data loop로 구성된 dual-loop 구조를 사용한다. 외곽 loop는 matching network와 rectifier 및 LDO를 통해 전력을 공급하고, 약하게 결합된 내부 loop는 공진 tank를 거의 loading하지 않는 고입력 임피던스 voltage-mode mixer를 직접 구동한다. 수신된 500MHz carrier는 coherent ASK 복조기의 local oscillator로 재사용되므로 좁은 대역의 간섭을 억제하면서 전력전송효율을 유지할 수 있다. 외부 10mm TX loop와 임플란트 RX loop 사이의 |S12|는 9mm 간격에서 -5.2dB로 측정되었다.

Downlink에는 전통적인 CDR loop 대신 복원한 carrier를 분주하여 retiming clock을 만드는 CDR-less 구조가 적용되었다. 여러 위상의 sampling edge를 생성하는 delay-buffered oversampling chain에서 eye center에 가장 가까운 위상을 선택함으로써 phase quantization 오차를 줄이며, symbol rate보다 2배 또는 5배 빠른 sampling도 지원한다. Uplink에서는 저속 입력 데이터를 짧은 500Mb/s burst로 묶는 burst-mode rate converter를 사용한다. Type-I PLL은 500MHz 기준 clock에 50주기 이내로 lock되며, 각 burst 약 100ns 전에만 켜졌다가 전송이 끝나면 꺼져 낮은 데이터율에서 PLL의 상시 전력을 제거한다.

IR-UWB 송신기는 두 개의 current-starved delay chain으로 RF pulse와 PSK 제어 신호의 timing을 조절하고, switched-capacitor power amplifier와 온칩 folded-dipole antenna를 통해 데이터를 방사한다. PSK scrambling은 OOK에서 나타나는 discrete tone을 억제하여 spectrum을 분산시킨다. 65nm 칩은 4.4×1.05mm²이며 downlink RX는 5Mb/s에서 208μW를 소비하고 10mm 거리에서 BER 10^{-6}을 달성하였다. Uplink 전력은 5Mb/s에서 0.23mW, 500Mb/s에서 4.93mW로 데이터율에 따라 조절되며, 500Mb/s에서 20cm 전송을 측정하고 최대 40cm에서 BER 10^{-3}을 확인하였다. 3mm의 x 및 y축 오정렬에서도 결합 손실 증가가 3dB 미만이었고, 20dBm 입력에서 SAR(10g)는 1.27W/kg으로 2W/kg 안전기준 이하를 보였다. 고속·장거리 통신과 지속적인 무선 전력 공급을 하나의 carrier 기반 clock 체계로 통합한 점에서 완성도가 높은 시스템이다.



[그림 2] 500MHz 전력 carrier를 공통 clock으로 재사용하는 crystal-free 양방향 WPDT 시스템: dual-loop inductive link를 통해 전력과 5Mb/s ASK downlink를 전달하고, 복원된 carrier로 CDR-less 복조기와 burst-mode IR-UWB uplink PLL을 동기화함. 저속 입력 데이터는 500Mb/s의 짧은 burst로 변환하여 송신기와 PLL의 동작시간을 줄임.

저자정보



석동열 박사과정 대학원생

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

VLSI 2026 Review

고려대학교 전기및전자공학부 박사과정 안재웅

Session 19 Bio-Sensing and Molecular Interfaces

이번 2026 VLSI의 Session 19에서는 총 5편의 논문이 발표되었다. 이 중 #19-1은 300-channel photometry neural probe에 adaptive-noise ΔΣ modulator를 적용하여 입력 신호가 작아질수록 noise floor가 감소하도록 하였으며, 16.8 fA/√Hz의 noise와 88.7 dB의 dynamic range를 달성하였다. #19-3은 FSCV 기반 neurotransmitter sensing에서 반복적으로 발생하는 non-faradaic current를 template으로 추적하는 dynamic track-and-zoom IADC를 제안하여, sensor drift에 대응하면서 83.5 dB의 dynamic range와 73.8 pA_{RMS}의 noise floor를 달성하였다. 마지막으로 #19-5는 time-gated fluorescence와 pseudo-differential sensing을 이용한 filter-free CMOS biochip으로, 9.9 fA의 minimum detection limit과 129 dB의 dynamic range, 133 dB 이상의 background suppression을 달성하고 실제 PCR 및 DNA melt-curve 분석까지 검증하였다.

#19-1 논문은 뇌 신경 활동에 따른 형광 신호를 고해상도로 측정하기 위한 300-channel photometry neural probe에 관한 내용이다. 기존의 fiber photometry는 광섬유와 외부 photodetector를 이용하기 때문에 여러 세포 영역을 동시에 세밀하게 측정하기 어렵고, 공간 해상도에도 한계가 있다. 이를 개선하기 위해 CMOS 기반 optical neural interface가 연구되고 있지만, SPAD 기반 구조는 15 V 이상의 높은 bias voltage와 비교적 큰 전력 소모가 필요하다. 본 논문에서는 이러한 문제를 해결하기 위해 photodiode (PD)와 저잡음 readout circuit을 probe shank 내부에 직접 집적한 300-channel photometry neural probe를 제안하였다.

제안된 probe는 shank를 따라 300개의 photometry channel을 배치하고, 각 channel마다 PD와 in-pixel readout을 함께 구성하였다. 기존 centralized readout 구조에서는 sub-nA 수준의 작은 photodiode current를 긴 analog line을 통해 전달해야 하기 때문에 noise에 취약하지만, 본 논문은 PD 바로 옆에서 신호를 변환함으로써 이러한 문제를 줄였다. 다만 좁은 probe shank에 수백 개 channel을 배치해야 하기 때문에 각 readout의 면적과 전력을 최소화하면서도 낮은 noise를 확보해야 하며, 이를 위해 adaptive-noise ΔΣ modulator 구조를 사용하였다.

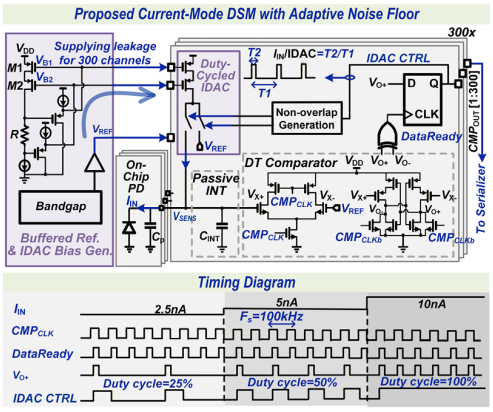
기존 current-mode ΔΣ modulator는 noise shaping을 통해 낮은 noise를 구현할 수 있지

만, active integrator의 면적이 크고 complementary current DAC에서 입력 크기와 관계없이 지속적으로 noise가 발생한다. 이에 본 논문에서는 입력 photocurrent의 polarity가 정해져 있다는 점을 이용하여 unipolar P-side IDAC과 passive integrator를 사용하였다. 특히 IDAC을 항상 동작시키는 것이 아니라 입력 신호에 비례하는 duty cycle로 동작시키기 때문에, 입력 photocurrent가 작아질수록 IDAC이 켜져 있는 시간이 감소하고 이에 따라 noise contribution도 함께 줄어든다. 결과적으로 낮은 입력 신호에서 noise floor가 자동으로 감소하며, 기존 구조와 비교해 약 12 dB의 noise reduction을 추가적인 power나 area overhead 없이 얻을 수 있다.

또한 각 channel은 1-bit ΔΣ output을 생성하며, shift-register 기반 serializer를 이용하여 300개 channel의 데이터를 단 두 개의 wire로 전송한다. 이를 통해 제한된 shank 내부의 routing area를 줄이고 bias와 power distribution을 위한 공간을 확보하였다. 각 channel의 photodiode parasitic capacitance와 top-metal MIM capacitor를 함께 이용해 passive integrator를 구성하고, IDAC과 dynamic comparator는 그 아래에 배치하여 면적을 최소화하였다. 또한 IDAC bias는 probe base에서 생성하여 모든 channel이 공유하도록 구성해, channel별로 별도의 current routing을 사용하는 것을 피하였다.

제안된 neural probe는 130-nm CMOS 공정으로 제작되었으며, 각 channel은 25 μm pitch와 0.0015 mm^2의 면적을 갖고 약 70 nW/channel의 낮은 전력을 소모한다. 측정 결과 입력 신호가 감소함에 따라 실제 noise floor도 함께 낮아지는 adaptive-noise 특성이 확인되었으며, 전체 시스템에서 88.7 dB의 dynamic range와 76.5 dB의 peak SNR을 달성하였다. Dark condition에서 측정한 3개 representative channel의 평균 input-referred noise density는 16.8 fA/√Hz로 나타났다. 특히 adaptive-noise 특성으로 인해 signal power가 6 dB 감소할 때 in-band noise가 약 3 dB 감소하여 넓은 입력 범위에서 sensitivity를 확보하였다.

마지막으로 실제 fluorescence sensing 가능성을 확인하기 위해 fluorescent bead를 probe의 일부 channel 위에 선택적으로 배치하고 470 nm light를 조사하였다. 그 결과 bead가 존재하는 위치의 channel에서만 명확한 fluorescence peak가 측정되어, 제안한 probe가 probe shank 위치에 따른 형광 신호를 공간적으로 구분하여 측정할 수 있음을 검증하였다. 이를 통해 높은 channel density와 낮은 noise 및 power를 동시에 확보한 photometry neural probe의 가능성을 보여주었다.



[그림 1] #19-1에서 제안된 in-pixel current-mode readout과 timing diagram

#19-3 논문은 뇌에서 dopamine과 같은 neurotransmitter를 측정하기 위한 high-dynamic-range sensor interface에 관한 내용이다. Neurotransmitter 측정에는 fast-scan cyclic voltammetry (FSCV)가 널리 사용되는데, 약 1.7 Vpp의 빠른 scan waveform을 electrode에 인가하면 측정하고자 하는 faradaic current는 nA 수준으로 매우 작은 반면, electrode capacitance에 의해 최대 1 μA 수준의 non-faradaic (NF) current가 함께 발생한다. 따라서 작은 neurotransmitter 신호를 큰 background current 속에서 구분하기 위해 80 dB 이상의 dynamic range와 낮은 noise floor가 요구된다. 기존 방식들은 NF current를 ADC 앞에서 제거하기 위해 DC 또는 DAC 기반 subtraction을 사용하거나 추가 electrode를 이용하지만, subtraction circuit의 non-ideality로 SNR이 저하되거나 sensor drift에 따라 주기적인 recalibration이 필요하고, 추가 electrode로 인해 implantation complexity가 증가하는 문제가 있다.

이를 해결하기 위해 본 논문에서는 dynamic template-tracking direct-digitization incremental ADC (IADC)를 제안하였다. FSCV에서 큰 비중을 차지하는 NF current가 scan 마다 거의 반복되는 quasi-repetitive 특성을 갖는다는 점을 이용하여, 처음에는 NF signal의 대략적인 형태를 template으로 저장하고 이후 ADC가 해당 template 주변의 작은 신호만 고해상도로 측정하도록 한다. 이를 통해 별도의 analog current subtraction 없이도 large background signal과 small faradaic signal을 동시에 처리할 수 있으며, electrode current를 ADC로 직접 전달하여 중간 analog stage에서 추가되는 noise도 줄였다.

제안된 ADC는 template acquisition (TA)과 dynamic track-and-zoom (DTZ)의 두 가지 mode로 동작한다. 시스템 시작 시 TA mode에서는 ADC의 input range를 넓히고 resolution을 낮춰 NF current의 coarse template을 획득하며, FSCV 한 scan에 해당하는 84개 sample을 5-bit로 저장한다. 이후 DTZ mode에서는 저장된 template이 예측하는 범위로 ADC conversion range를 좁혀 작은 faradaic current를 높은 resolution으로 측정한다.

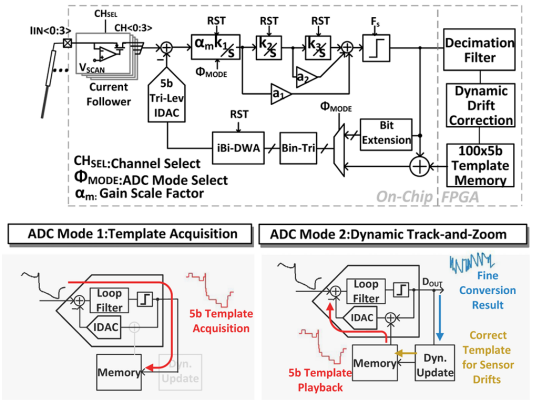
즉, 전체 μA 범위를 항상 높은 resolution으로 변환하는 대신 이미 알고 있는 반복 background 주변만 zoom-in하여 변환함으로써 높은 DR을 확보하는 방식이다.

또한 electrochemical sensor의 특성상 시간이 지나면서 NF current의 형태가 변하는 sensor drift가 발생할 수 있는데, 고정된 template을 사용하면 실제 signal이 ADC conversion range를 벗어날 수 있다. 본 논문에서는 DTZ mode에서 얻어진 fine conversion 결과가 upper/lower threshold를 벗어나는지를 판단하고, 이에 따라 해당 template 값을 증가 또는 감소시키는 dynamic drift correction을 적용하였다. 따라서 기존 방식처럼 측정을 중단하고 주기적으로 recalibration을 수행하지 않아도 지속적으로 template을 실제 sensor 상태에 맞게 업데이트할 수 있다.

회로적으로는 high-side-regulated current follower를 사용하여 1.7 Vpp의 FSCV scan waveform을 electrode에 인가하는 동시에 sensor current를 ADC 입력으로 직접 전달한다. Current follower는 3.3 V에서 동작하며 ADC의 1.5 V input range로 신호를 level shift 하여 별도의 current mirror나 추가 level-shifting reference를 사용하지 않도록 구성하였다. ADC는 3rd-order CIFF incremental ΔΣ 구조를 사용하며, template acquisition과 fine conversion에서 동일한 analog hardware를 재사용하여 두 동작 사이의 mismatch를 줄였다. 또한 tri-level current DAC과 incremental bidirectional data-weighted averaging (iBi-DWA)을 적용하여 DAC noise와 mismatch 영향을 줄였으며, FSCV가 주기적으로 짧게 동작한다는 특성을 이용한 power gating으로 약 91%의 전력 절감을 구현하였다.

제안된 sensor interface는 180-nm CMOS 공정으로 제작되었으며, on-chip power consumption은 6.7 μW, drift correction과 memory의 추정 소비전력을 포함하면 약 6.9 μW/channel이다. Dynamic track-and-zoom을 적용한 측정 결과 83.5 dB의 dynamic range, 73.8 pA_{RMS}의 noise floor, 70.6 dB의 peak SNR을 달성하였으며, 기존 구조 대비 DR은 22.4 dB, peak SNR은 22.0 dB 향상되었다. 또한 carbon fiber electrode를 사용한 wet-bench 실험에서 sub-100 nM 수준의 dopamine detection이 가능함을 확인하였다.

마지막으로 실제 in-vivo 실험을 통해 회로의 적용 가능성을 검증하였다. 제작된 sensor interface와 carbon fiber electrode를 rat brain에 적용하고 ventral tegmental area (VTA)를 전기적으로 자극한 결과, 약 300 nM의 dopamine response를 측정하였다. 즉 단순한 electrical characterization에 그치지 않고 실제 생체 환경에서도 neurotransmitter 신호를 검출했으며, 추가 electrode 없이 높은 DR을 확보했다는 점이 본 논문의 주요 특징이다.



[그림 2] #19-3에서 제안된 overall system block diagram과 두 ADC mode의 작동 원리

#19-5 논문은 point-of-care (PoC) DNA 검사를 위한 filter-free time-gated fluorescence CMOS biochip에 관한 내용이다. 기존 CMOS fluorescence biochip은 형광 신호보다 훨씬 강한 excitation light와 autofluorescence를 제거하기 위해 optical emission filter와 같은 추가 post-CMOS 공정이 필요한 경우가 많다. 본 논문에서는 이러한 추가 optical filter 없이도 DNA fluorescence를 검출할 수 있도록 time-gated fluorescence (TGF) 기법을 적용한 16×24 CMOS biosensor array를 제안하였으며, chip 내부에 heater와 temperature sensor를 함께 집적하여 PCR과 melt-curve analysis까지 수행할 수 있도록 구성하였다.

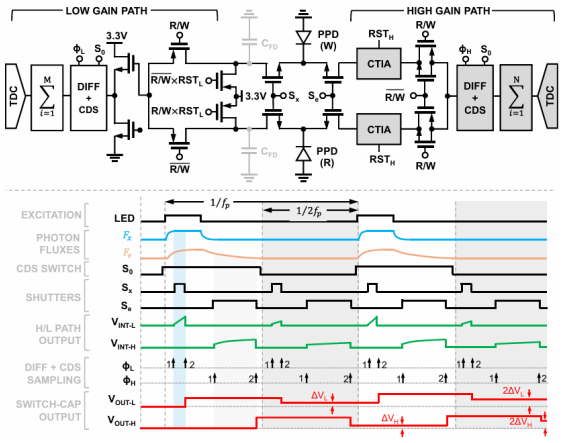
각 pixel은 working (W) photodetector와 reference (R) photodetector로 구성되며, working detector는 실제 fluorescence를 측정하고 reference detector는 optical interference를 측정하여 pseudo-differential 방식으로 두 신호를 비교한다. Excitation LED가 꺼진 이후 약 10~100 μs의 delay를 두고 fluorescence decay만 integration하는 time-gating을 적용하여 excitation light의 영향을 줄이며, bright/dark sample을 correlated double sampling (CDS)하여 dark current와 offset 성분도 제거한다. 또한 fluorescence와 excitation signal 사이의 큰 dynamic range를 처리하기 위해 high-gain과 low-gain readout path를 별도로 구성하였다.

High-gain path에서는 pinned photodiode (PPD)와 CTIA를 이용하여 작은 fluorescence photocurrent를 측정한다. CTIA에는 low-leakage reset switch와 feedback level shifter를 적용하였으며, 이후 switched-capacitor integrator에서 bright와 dark 신호를 CDS 방식으로 차분하고 여러 cycle 동안 누적하여 SNR을 높인다. 최종 신호는 20-bit TDC를 이용하여 digital value로 변환된다.

제안된 biochip은 0.18-μm CIS 공정으로 제작되었으며, high-gain path에서 9.9 fA의 minimum detection limit과 두 gain path를 이용한 129 dB dynamic range를 달성하였다. In-pixel CDS를 통해 offset을 49 dB 감소시키고 LED variation correction으로 spatial

nonuniformity를 37%에서 5.2% 이하로 감소시켰으며, pseudo-differential sensing과 time-gating을 함께 적용하여 excitation 및 autofluorescence background를 133 dB 이상 억제하였다. 또한 integrated heater를 이용해 18 °C/s 이상의 heating/cooling rate를 달성하여 빠른 PCR thermal cycling이 가능함을 확인하였다.

마지막으로 europium-chelate fluorescence label과 DNA probe를 chip 표면에 적용하여 실제 PCR 및 melt-curve 실험을 수행하였으며, 서로 다른 target concentration에서 target DNA의 존재 여부를 구분할 수 있음을 확인하였다. 이를 통해 제안된 biochip은 optical filter와 복잡한 post-CMOS processing을 최소화하면서 fluorescence sensing, PCR thermal cycling, DNA melt-curve analysis를 하나의 CMOS platform에서 수행할 수 있음을 보여주었다.



[그림 3] #19-5에서 제안된 functional block diagram과 timing diagram

저자정보



안재웅 박사과정 대학원생

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

VLSI 2026 Review

고려대학교 전기전자공학과 박사과정 이윤호

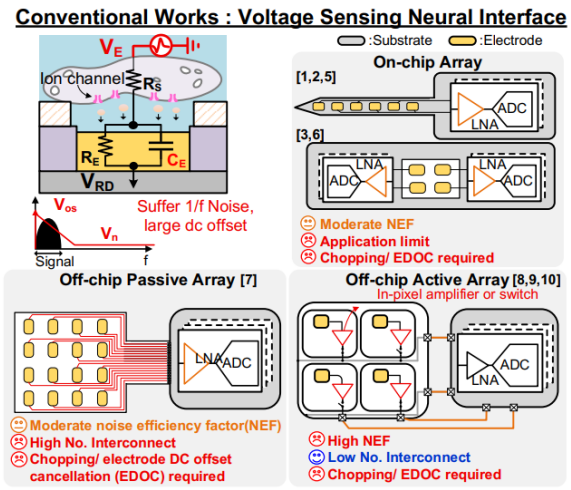
Session 24 Circuits for Bio-Signal Acquisition and Neural Interfaces

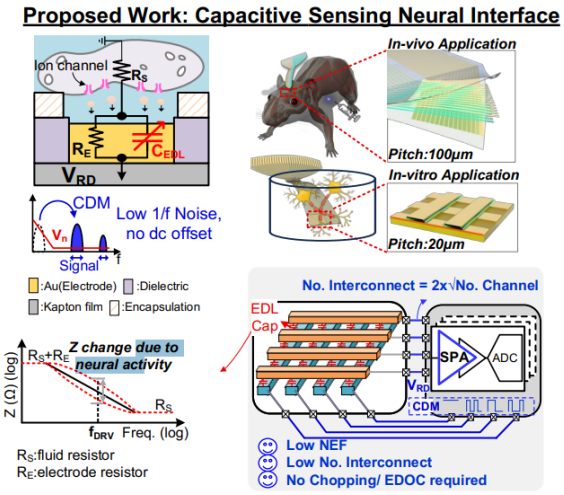
이번 VLSI 2026 Session 24에서는 wearable 및 implantable biomedical system에서 생체 신호를 정밀하게 측정하고 신경을 효과적으로 자극하기 위한 저전력 neural interface 기술을 다룬 총 5편의 논문이 발표되었다. 본 리뷰에서는 이 가운데 capacitive sensing과 code-domain multiplexing(CDM)을 이용하여 32개의 interconnect만으로 256-channel neural signal을 측정하는 neural interface (24-2)와, 전극에 필요한 자극전압에 따라 high-voltage supply를 동적으로 조절하고 정확한 charge balancing을 제공하는 temporal interference stimulation(TIS) ASIC (24-5)을 중심으로 살펴보고자 한다.

#24-2 본 논문은 flexible microelectrode array(MEA)의 전극 수가 증가할 때 발생하는 배선 및 전력 문제를 해결하기 위해, 32개의 interconnect만으로 256개의 전극을 동시에 측정하는 capacitive neural interface를 제안하였다. 뇌의 신경활동을 높은 해상도로 관찰하려면 좁은 영역에 많은 전극을 배치해야 하지만, 전극마다 별도의 신호선을 연결하면 시스템의 크기와 복잡도가 크게 증가한다. 아래 그림 6-(a)의 우 상단 그림과 같이 전극과 readout circuit을 하나의 CMOS chip에 집적하면 배선을 줄일 수 있지만, 단단한 silicon substrate로 인해 뇌 표면에 밀착시키기 어렵다. 반면 유연한 외부 전극을 사용하면 생체 표면에 쉽게 부착할 수 있으나, passive MEA(좌 하단)는 많은 배선을 필요로 하고 active MEA(우 하단)는 좁은 전극 면적에 amplifier와 switch를 구현해야 하므로 noise와 전력소모가 증가한다. 또한 기존 voltage-sensing 방식은 작은 neural signal과 함께 큰 electrode offset 및 1/f noise도 증폭되기 때문에 이를 제거하기 위한 별도의 보정 회로가 필요하다는 한계이 있다.

본 논문은 이러한 한계를 개선하기 위해 그림 6-(b)와 같이 neural voltage 대신 EDL(Electrical Double-Layer) capacitance의 변화를 측정한다. EDL capacitance는 전극과 체액의 경계면에 형성되는 capacitor 성분으로, 신경세포가 활성화되어 주변의 이온 농도가 변하면 그 값도 함께 변한다. 제안한 16×16 MEA에서는 16개의 column에 서로 다른 Walsh code를 인가하고, 각 row에서 16개 전극의 신호를 하나로 합쳐 읽는다. 이때 Walsh code가 각 column을 구분하는 표식 역할을 하므로, 하나로 합쳐진 신호도 ADC 변환 후 digital demodulation을 통해 다시 16개의 개별 신호로 분리할 수 있다. 따라서

16개의 column drive line과 16개의 row readout line만으로 총 256개의 전극을 측정할 수 있다. 또한 neural signal이 Walsh code에 의해 높은 주파수 영역으로 이동하므로, 저주파 영역에 집중된 electrode offset과 1/f noise의 영향도 함께 줄일 수 있어 기존의 문제를 개선할 수 있다.

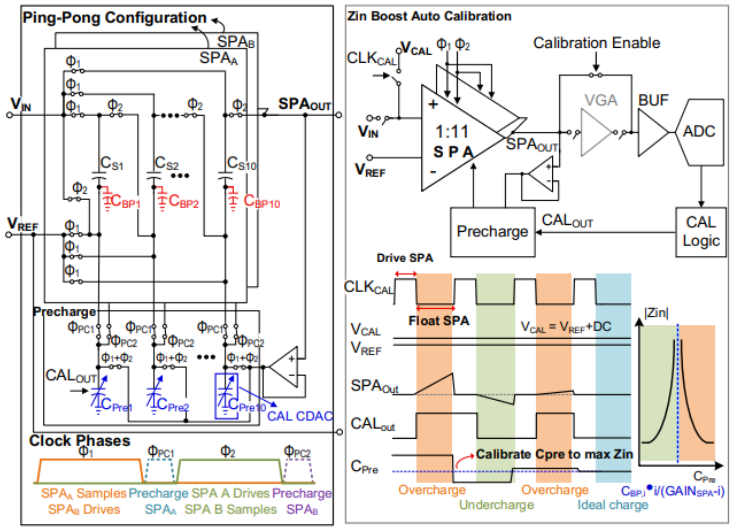




[그림 6-(a)] [그림 6-(b)]

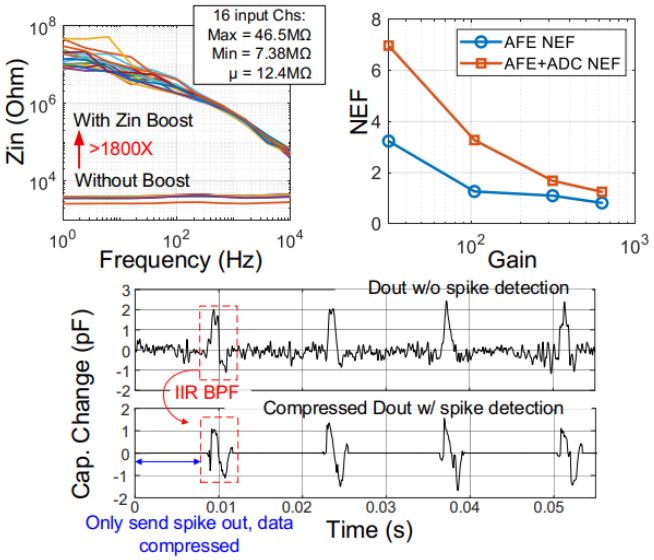
[그림 6] 기존 neural interface의 구조 및 한계점 (a)과 본 논문에서 제안하는 capacitive-sensing 기반의 neural interface (b).

아래 그림 7은 각 row에서 합쳐진 작은 신호를 증폭하기 위한 본 논문의 SPA(Series-Parallel Amplifier)와 input-impedance boosting 기법을 보여준다. SPA는 sampling 구간에서 여러 capacitor를 입력에 병렬로 연결하여 신호를 저장한 뒤, amplification 구간에서 이를 직렬로 재구성하여 약 11배의 voltage gain을 얻는다. 두 개의 SPA가 번갈아 sampling과 amplification을 수행하므로 신호를 연속적으로 처리할 수 있지만, sampling capacitor가 전극에 연결될 때 전극으로부터 charge를 가져가면 입력신호가 감소하고 input impedance가 낮아지는 문제가 발생한다. 이를 방지하기 위해 제안한 회로는 sampling capacitor를 입력과 가까운 전압으로 미리 충전하여, 전극과 연결될 때 이동하는 charge를 최소화한다. 이를 통해 추가적인 amplifier 없이도 전력소모를 줄이면서 작은 neural signal을 충분히 증폭시켜 전극의 신호 감소를 방지할 수 있다.



[그림 7] 본 논문에서 제안하는 SPA의 동작 원리와 auto-calibrated input-impedance boosting 기법.

아래 그림 8은 본 논문에서 제안한 input-impedance boosting의 효과와 on-chip spike detection의 측정 결과를 보여준다. Auto-calibration을 적용한 AFE의 input impedance는 1Hz에서 평균 12.4MΩ으로 측정되어, boosting을 적용하지 않은 경우보다 1,800배 이상 향상되었다. 전력소모 대비 noise 성능을 나타내는 NEF는 AFE 단독의 경우 0.81, ADC를 포함한 전체 readout에서 1.25를 달성하였다. 또한 on-chip spike detector는 연속적으로 발생하는 전체 데이터를 전송하는 대신, neural spike가 검출된 구간의 파형과 발생 시점만 출력하여 외부로 전송되는 데이터의 양을 줄였다. 본 논문의 neural interface는 22nm FDSOI 공정에서 1.76mm²의 면적과 0.484mW의 총 전력소모로 구현되었으며, channel당 1.89μW의 전력소모와 4.36nV/√Hz의 input-referred noise density를 달성하였다. 마지막으로 20μm-pitch MEA를 이용한 배양 neuron 측정과 100μm-pitch flexible MEA를 이용한 mouse 측정을 통해, 적은 수의 interconnect로 256-channel neural activity를 낮은 noise와 낮은 전력으로 측정할 수 있음을 입증하였다.



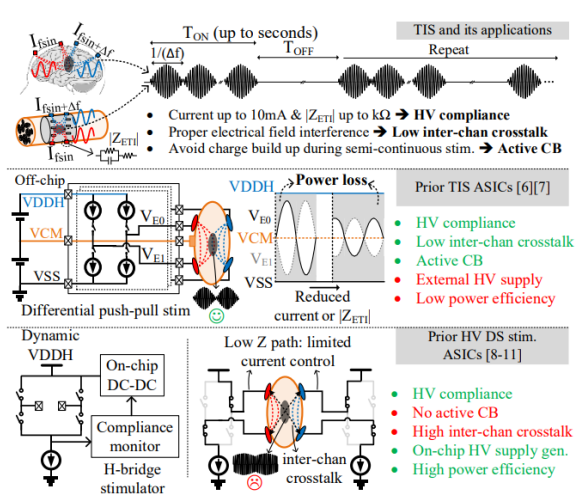
[그림 8] AFE의 Input-impedance 및 NEF 측정 및 on-chip spike detection 측정 결과.

#24-5 본 논문은 기존의 temporal interference stimulation(TIS)에서 고정된 고전압 전원으로 인해 발생하는 전력 손실과, dynamic supply 적용 시 어려워지는 charge balancing 문제를 해결하기 위해 on-chip dynamic supply를 갖는 새로운 TIS 구조를 제안하였다. TIS는 서로 다른 주파수를 갖는 두 쌍의 sinusoidal current를 인체에 인가하고, 두 전류가 중첩되면서 형성되는 저주파에 대한 envelope를 이용하여 원하는 신경 부위를 자극하는 방식이다. 두 stimulation 쌍의 전류 크기 비율을 조절하면 자극이 집중되는 위치도 변경될 수 있으므로, 수술 없이 더 깊은 신경을 선택적으로 자극할 수 있다는 장점이 있다. 그러나 최대 10mA의 전류를 수 kΩ에 이르는 electrode-tissue impedance에 공급해야 하므로 10V 이상의 높은 compliance voltage가 필요하며, stimulation이 수초 동안 지속되는 경우 source와 sink current 사이에 작은 오차가 발생하면 전극과 조직 계면에 charge가 축적되어 source와 sink current에 대해 정확한 charge balancing이 요구된다.

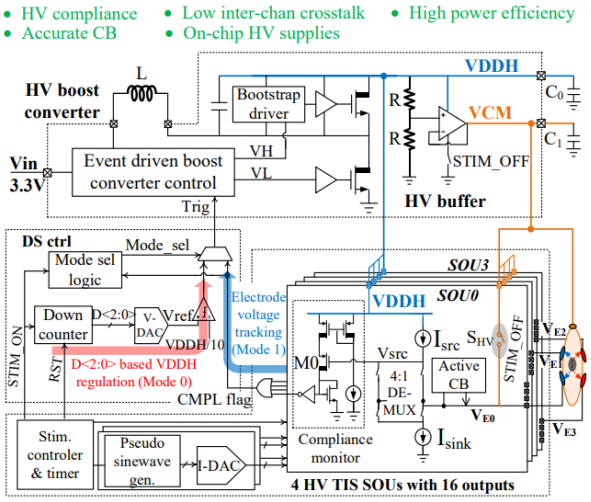
기존 TIS 방식은 아래 그림 9-(a)와 같이 고전압과 안정적인 charge balancing을 확보하기 위해 외부의 고정된 고전압을 사용하였다. 이 구조는 전극전압이 낮은 경우에도 항상 최대 전원전압을 유지하므로, stimulator 내부에 상당한 전력 손실이 발생하는 한계가 있다. 기존 dynamic supply stimulator의 경우는 전극전압에 맞춰 전원전압을 낮출 수 있지만, 주로 biphasic waveform을 생성하는 H-bridge 구조를 사용하기 때문에 여러 채널이 동시에 동작하는 TIS에서는 channel 사이에 cross-talk 문제가 발생할 수 있으며 active charge balancing도 적용하기 어렵다.

본 논문에서는 이러한 기존의 문제점들을 개선하기 위해 그림 9-(b)와 같이 4개의 push-pull SOU(Stimulation Output Unit)를 두 개의 differential pair로 구성하고, 각 SOU를 4개의 전극 중 하나에 연결하여 총 16개의 출력을 선택할 수 있는 TIS 구조를 제안하였다. 내부 고전압 boost converter가 각 SOU에서 요구되는 가장 높은 전압을 기준으로 공

통 고전압 전원 VDDH를 생성한다. 각 SOU에는 active 및 passive charge-balancing 회로가 포함되어 있어, 전원전압을 낮게 유지하면서도 자극전류의 왜곡과 전극과 조직 계면의 charge 축적을 방지하여 기존의 문제를 해결하였다.



[그림 9-(a)]

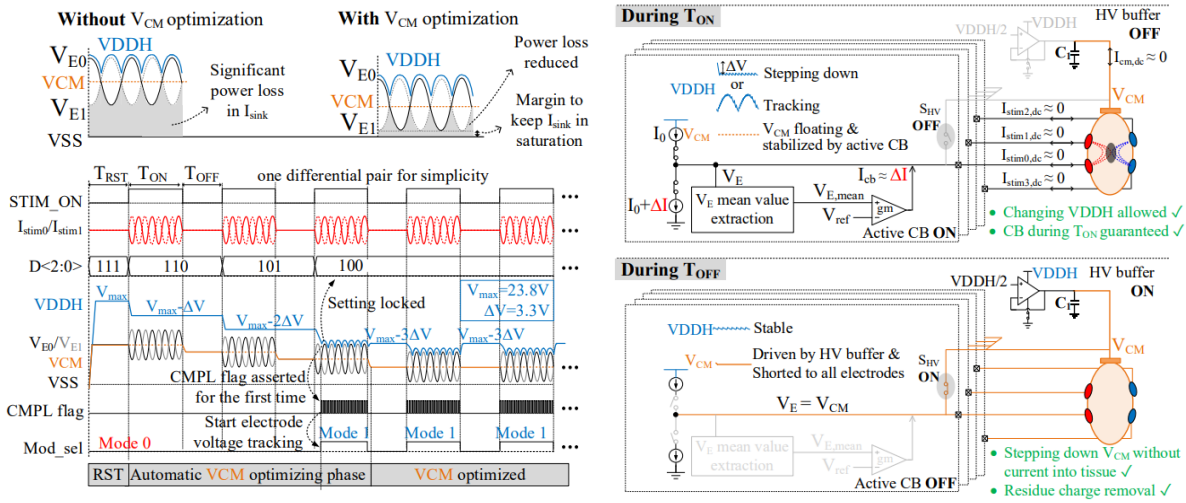


[그림 9-(b)]

[그림 9] TIS의 동작 원리와 기존 구조의 한계 (a) 및 본 논문에서 제안하는 dynamic-supply TIS ASIC의 구조 (b).

아래 그림 10-(a)는 전력 손실을 최소화하기 위해 본 논문에서 제안하는 common-mode voltage(VCM) 최적화에 대한 과정을 보여준다. VCM이 지나치게 높으면 VDDH가 전극전압을 따라 낮아지더라도 sink 전류원에 큰 전압이 걸려 전력 절감 효과가 제한된다. 따라서 stimulation이 시작되면 mode 0에서 VDDH와 VCM을 최대값으로 설정한 후, 3-bit code를 이용하여 두 전압을 단계적으로 낮춘다. 전극전압이 SOU의 동작 한계에 가까워져 최초의 flag가 발생하면 해당 VCM을 최적 값으로 결정한다. 이후 mode 1에서는 VCM을 유지하면서 VDDH가 4개 SOU 중 가장 높은 전극전압을 필요한 headroom만큼만 여유를 두고 따라가도록 한다.

하지만, VDDH가 계속 변하면 charge-balancing 회로의 기준이 되는 VCM도 흔들릴 수 있다. 이를 방지하기 위해 그림 10-(b)와 같이 stimulation이 인가되는 TON 동안에는 HV buffer와 passive switch를 끄고 VCM을 capacitor에 유지한다. 이때 active charge-balancing loop가 전극전압의 평균값을 감지하여 source와 sink current의 차이를 보상하므로, VCM을 저장하는 capacitor는 거의 일정하게 유지된다. 반대로 TOFF 동안에는 HV buffer와 passive switch를 켜서 모든 전극을 VCM에 연결함으로써 남아 있는 charge를 제거하고, 다음 stimulation에 필요한 VCM으로 갱신한다. 이를 통해 dynamic supply가 동작하는 동안에도 안정적인 VCM과 정확한 charge balancing을 동시에 유지할 수 있다.

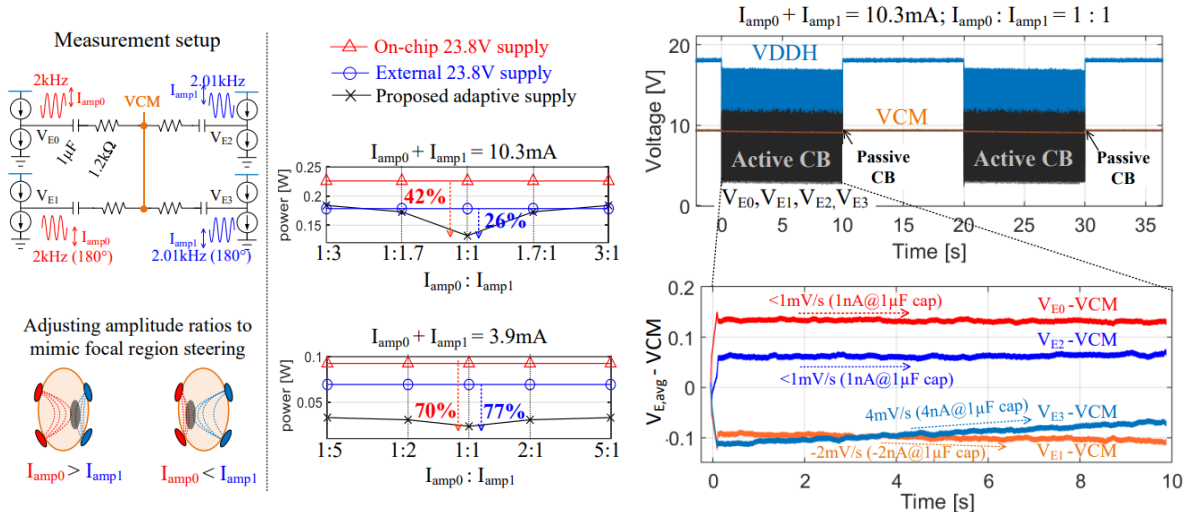


[그림 10-(a)] [그림 10-(b)]

[그림 10] 본 논문의 VCM 및 dynamic-supply 공급 동작에 대한 타이밍 다이어그램 (a) 및 active/passive 방식의 charge-balancing 방법 (b).

아래 그림 11-(a)는 두 stimulation pair를 2kHz와 2.01kHz로 구동한 뒤, 두 전류의 크기 비율을 변경하면서 측정한 TIS 동작 결과를 보여준다. 두 stimulation pair의 전체 전류가 10.3mA일 때는 23.8V의 고정 전원을 사용하는 on-chip supply 대비 최대 42%, 이상적인 외부 고정 전원 대비 최대 26%의 전력을 절감하였다. 전체 전류가 3.9mA로 감소하면 불필요한 voltage headroom이 더욱 커지기 때문에 전력 절감 효과도 증가하여, 각각 최대 77%와 70%의 전력 감소를 보이는 것을 확인할 수 있다.

그림 11-(b)에서는 5mA의 stimulation current에 약 20μArms의 의도적인 source-sink mismatch를 인가할 때, active charge balancing을 통해 잔류 오차를 4nA 이하로 크게 낮추는 것을 확인할 수 있다. 본 논문에서 제안한 TIS는 130nm BCD 공정에서 4×5mm²의 면적으로 구현되었으며, 3.3V 입력에서 최대 21V의 compliance voltage와 SOU당 최대 10.3mA의 stimulation current를 제공한다. Boost converter는 4V 출력에서 최대 93%, 23.8V에서 80%의 효율을 달성하였고, 전체 stimulator의 최대 효율은 57%로 측정되었다. 또한 saline 및 고밀도 nerve-cuff electrode를 이용한 실험을 통해 원하는 TIS electric field를 형성하면서 전력소모를 약 175–176mW에서 72–73mW로 줄일 수 있음을 확인하였다. 이를 통해 본 논문은 TIS용 multi-channel stimulator에 dynamic supply와 정확한 charge balancing을 동시에 적용하여, 높은 stimulation voltage와 안전성을 유지하면서 최대 70%의 전력을 절감할 수 있음을 입증하였다.



[그림 11-(a)] [그림 11-(b)]

[그림 11] 전류 크기 및 비율에 따른 TIS 동작의 전력소모 측정 (a) 및 dynamic supply 동작 중 charge-balancing 성능 측정 (b).

저자정보



이윤호 박사과정 대학원생

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

2026 VLSI Review

KAIST 인공지능반도체대학원 박사과정 하상우

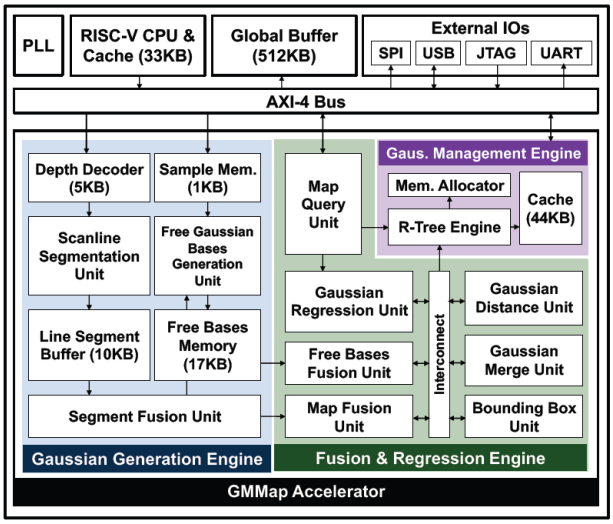
Session 3 Accelerators: Spatial and Immersive Visual Computing

이번 2026 IEEE/JSAP Symposium on VLSI Technology & Circuits의 Session C3에서는 spatial computing 및 immersive visual computing을 위한 accelerator SoC를 주제로 총 5편의 논문이 발표되었다. 최근 AR/VR, 로봇, embodied agent와 같은 edge device에서 3D 공간을 실시간으로 인지하고 재구성하려는 수요가 커지면서, 3D Gaussian 기반 표현, neural video codec, vision-language navigation, event 기반 센싱과 같은 새로운 workload가 등장하고 있다. 이들 workload는 연산량 뿐 아니라 불규칙한 자료구조 접근, sorting, entropy coding과 같은 non-GEMM 연산이 병목이 되는 경우가 많아, 알고리즘 특성을 회로 구조에 직접 반영하는 algorithm-hardware co-design이 핵심 설계 이슈로 부각되고 있다. 본 리뷰에서는 세션에서 발표된 다섯 편의 논문 #C3-1부터 #C3-5까지를 차례로 살펴보고자 한다.

#C3-1 본 논문은 MIT 연구진이 발표한 것으로, Gaussian 기반 3D occupancy map인 GMMap을 위한 6 mW급 SoC인 Gleanmer를 제안하였다. 3D occupancy map은 공간을 occupied, free, unexplored 영역으로 구분하여 안전한 자율 주행 및 AR/VR 상호작용을 가능하게 하지만, OctoMap과 같은 기존 방식은 하드웨어 가속을 적용하더라도 100 mW 이상을 소모하여 edge 전력 예산을 초과한다. GMMap은 occupied 및 free 영역을 3D Gaussian ellipsoid로 압축 표현하지만, 기존 embedded software 구현은 2 W 이상을 소모한다는 한계가 있었다.

이를 해결하기 위해 Gleanmer는 세 가지 co-optimization을 적용하였다. 첫째, 전체 map construction 에너지의 최대 65%를 차지하던 free Gaussian bases 생성을 수천 개의 line segment가 아닌 occupied Gaussian으로부터 직접 계산하여 construction 에너지를 22~63% 절감하였다. 둘째, trajectory 상의 인접한 query coordinate들이 유사한 R-Tree 경로를 순회한다는 점에 착안하여, 16개 coordinate를 하나의 bounding box로 묶는 batch query를 통해 query throughput을 4~10배 높이고 query 에너지를 74~81% 절감하였다. 셋째, scanline segmentation의 slope를 1-cycle로 근사하고 Gaussian precision을 32-bit에서 19-bit로 낮추는 approximate computing으로 accelerator 면적을 38% 줄였다.

16nm CMOS로 제작된 칩은 640×480 depth image를 88 fps 이상으로 처리하며, 6 mW 이하의 전력으로 NVIDIA Jetson TX2 대비 341배, OctoMap 가속기인 OMU 대비 44배 낮은 전력 소모를 달성하였다.

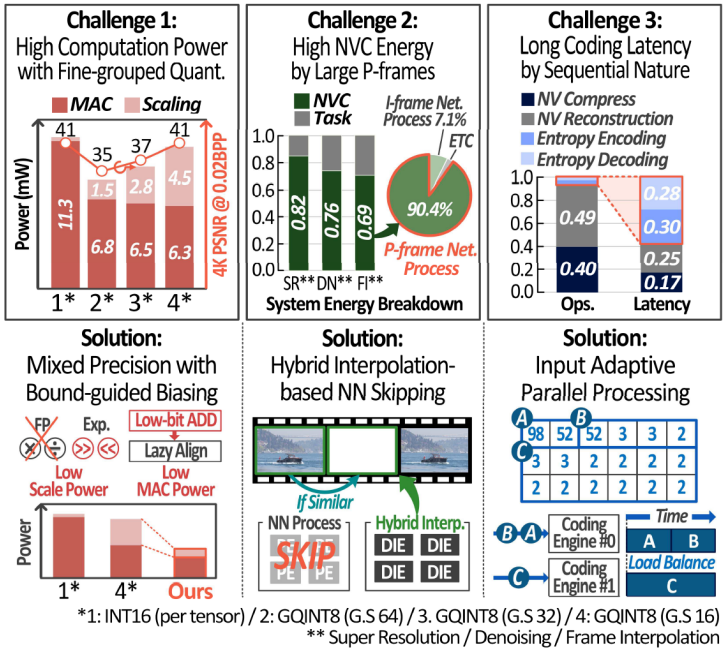


[그림 1] Gleanmer SoC의 System architecture

#C3-2 본 논문은 KAIST 연구진이 발표한 NOVA로, 단일 칩에서 decoding-processing-encoding을 end-to-end로 수행하는 최초의 fully integrated neural video SoC를 제안하였다. 기존 neural video processor들은 super resolution, denoising과 같은 video task 가속에만 집중하여 압축 및 전송을 위한 별도의 codec 하드웨어가 필요했고, implicit NVC는 video별 학습에 서버급 GPU로 1~2일이 소요되어 decode-only에 머물렀다. 단일 forward pass로 인코딩이 가능한 explicit NVC의 등장으로 단일 칩 통합의 가능성이 열렸지만, quantization 민감성, P-frame 처리가 시스템 에너지의 69~82%를 차지하는 문제, symbol 단위 순차 연산으로 인한 entropy coding latency가 새로운 병목으로 부각되었다.

NOVA는 이를 세 가지 기법으로 해결하였다. 첫째, bound-guided mixed quantization(BMQ)은 inlier를 INT4로 표현하면서 shared exponent bias를 shift 연산으로 대체하여 FP scaling 전력을 91.9% 줄이고, shift-ordered adder tree를 갖춘 bound-guided unified(BU) core와 함께 연산 전력을 41.7% 절감하였다. 둘째, temporal interpolation unit(TIU)은 NVC 초기 layer의 flow 정보를 활용해 temporally predictable한 frame의 neural processing을 선택적으로 생략하고, tiered DMA를 통해 TIU와 core의 병렬 동작을 지원하여 encoding/decoding 에너지를 각각 71.2%/71.8% 절감하였다. 셋째, adaptive entropy coding unit(AECU)은 entropy 분포에 따라 chunk 경계를 동적으로 결정하고 load balancing controller로 병렬 engine에 분배하여 hardware utilization을 91.6%까지 높

이고 coding latency를 최대 79.2% 줄였다. 28nm CMOS로 제작된 칩은 4K UVG benchmark에서 31.4 TOPS/W의 에너지 효율을 달성하였으며, encoding은 5.3 mJ/frame과 13.7 ms/frame으로 edge GPU 대비 76.7배 빠르고, decoding은 6.3 mJ/frame으로 4K 70.8 FPS를 달성하였다.

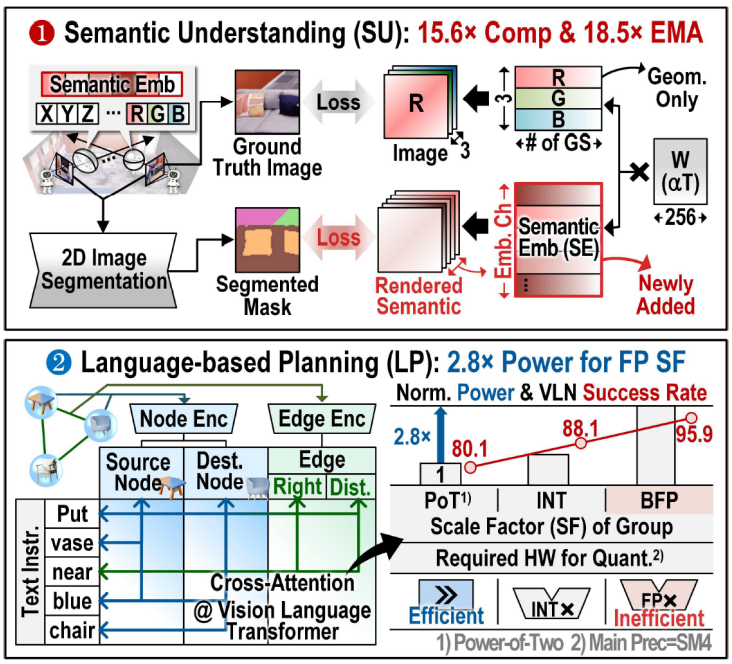


[그림 2] Video Codec HW의 문제점과 논문 #C3-2의 해결법

#C3-3 본 논문 역시 KAIST 연구진이 발표한 것으로, embodied agent를 위한 spatial reasoning 기반 vision language navigation(VLN) 가속기인 SR-VLNA를 제안하였다. 기존 point planning 가속기들은 미리 정의된 목표 좌표를 향한 occupancy 기반 경로 계획만 수행하여 동적인 주변 환경과 언어 명령을 이해하지 못한다. 제안된 VLN 시스템은 3D Gaussian splatting 기반 semantic understanding(SU)과 language-based planning(LP)으로 구성되지만, edge GPU에서는 semantic embedding 추가로 인한 15.6배의 연산량과 18.5배의 external memory access, cross-attention의 넓은 activation 범위로 인한 FP group scaling 전력, SU와 LP의 상이한 sparsity 패턴이 병목이 된다.

SR-VLNA는 네 가지 기법을 제안하였다. Embedding skipping unit(ESU)은 가려진 객체의 semantic embedding과 중요도가 낮은 channel을 제거하여 연산량을 65.6%, external memory access를 77.7% 줄였고, mapping region decision unit(MRDU)은 balanced KD-tree 기반 bounding box 추출로 정적 객체의 Gaussian redundancy를 제거하여 SU latency를 93.4% 절감하였다. Context-aware group quantization unit(CAGQU)은

instruction 유형과 edge 수에 따라 power-of-two와 INT/FP scale factor를 선택적으로 사용하여 FP SF 사용을 92.9% 줄였으며, heterogeneous sparsity-aware adder tree(HSAT)는 sparse한 데이터에는 carry-suppressed adder tree를, dense한 데이터에는 zero count adder tree를 할당하여 에너지 효율을 2.3~2.7배 개선하였다. 28nm FDSOI로 제작된 칩은 SU에서 0.004~0.021 μJ/point의 state-of-the-art 성능을 보였고, LP를 최초로 시연하여 0.25~0.96 mJ/task를 달성하였으며, 시스템 레벨에서 5.0~23.9 mJ/meter로 edge GPU 대비 압도적으로 낮은 에너지를 기록하였다.



[그림 3] Spatial reasoning 기반 VLN의 문제점

#C3-5 본 논문은 Tsinghua University 연구진이 발표한 것으로, 3D/4D Gaussian splatting 실시간 렌더링을 위한 12nm spatial video rendering processor인 Uni4D를 제안하였다. 4DGS는 Hex-plane encoding과 MLP 기반 deformation으로 3DGS를 동적 장면으로 확장하지만, 하드웨어 관점에서는 K-plane의 3D indexing에 필요한 대형 2D register array, tile 기반 preprocessing에서의 순차적인 Gaussian box-bounding, software 방식 radix sorting으로 인한 frame당 수백 MB의 external memory access가 병목이 된다.

Uni4D는 K-plane sampling과 MLP deformation을 단일 datapath로 통합한 UniMAC deformation engine과, 64×64 grid를 16×16 sub-grid로 분할하는 K-plane weights segmentation(KWS)을 통해 K-plane buffer 면적을 84% 절감하였다. Non-blocking KVP generation engine은 boundary tile filtering unit으로 tile 범위를 shift와 decoder만으로 도출하고 다수의 key-value generation unit에 병렬 분배하여 preprocessing 시간을 74% 단

축하였으며, hierarchical radix sorting engine(HRSE)은 global tile sorting과 on-chip depth sorting을 분리하고 4-bit partial sub-sorting과 rescheduled parallel sorting workload를 적용하여 external memory access를 frame당 464 MB에서 126 MB로 줄였다. 12nm CMOS로 제작된 5.29 mm^2 칩은 0.16~0.52 mJ/frame의 에너지로 HyperNeRF 시퀀스를 74~76 fps, synthetic D-NeRF 장면을 최대 232 fps로 렌더링하며, 기존 NeRF/3DGS 프로세서 대비 한 자릿수 이상 낮은 frame당 에너지를 달성하였다.

종합적으로 Session C3는 3D Gaussian이 occupancy mapping(#C3-1), semantic understanding(#C3-3), spatial video rendering(#C3-5)에 이르기까지 edge spatial computing의 공통 표현으로 자리잡고 있으며, 이를 뒷받침하는 전용 하드웨어가 mW급 전력 예산에서 실시간 성능을 달성하는 단계에 도달했음을 보여주었다. 또한 R-Tree 순회, entropy coding, KD-tree 탐색, radix sorting, graph 구성과 같이 GPU가 비효율적인 불규칙 연산을 전용 engine으로 가속하고, sparsity와 temporal locality 같은 workload 특성을 quantization 및 adder tree 수준까지 반영한 점이 세션 전반의 공통된 설계 철학으로 읽혔다.

저자정보



하상우 박사과정 대학원생

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

2026 VLSI Review

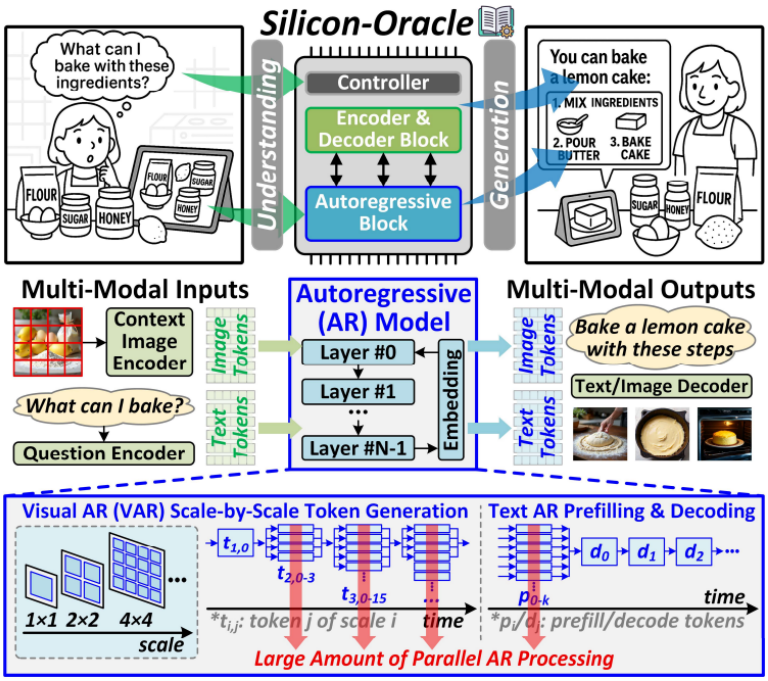
경북대학교 전자전기공학부 박사과정 박승현

Session 7 Accelerators: Generative Models and Specialized Decoding

이번 Session C7의 세 논문은 multi-modal autoregressive model, LLM speculative decoding, quantum error correction처럼 서로 다른 응용을 다루지만, 공통적으로 계산 자체보다 **memory access와 data movement에서 발생하는 병목을 줄이는 방향**에 초점을 맞추고 있다. 첫 번째와 두 번째 논문은 autoregressive model과 LLM의 external memory access를 줄이고, 세 번째 논문은 quantum error correction에서 weight storage와 matching complexity를 줄인다.

특히 세 논문 모두 단순히 연산기의 peak performance를 높이는 방향보다는 workload의 특성을 이용해 필요한 데이터만 남기거나, dataflow를 바꾸거나, 필요한 순간에만 계산하도록 한다는 점이 인상적이다. 최근 accelerator 설계에서 실제 application의 execution flow와 memory behavior를 얼마나 잘 반영하는지가 중요해지고 있다는 점을 잘 보여주는 세션이라고 생각한다.

#7-1 Silicon-Oracle (Soracle): A Multi-Modal Autoregressive Model Accelerator for Context-Aware Assistance on Mobile Platform



[그림 1] Soracle의 head-local RBCU AR core와 EMA-free attention execution 구조

첫 번째 논문은 mobile 환경에서 image와 text를 함께 처리하는 multi-modal autoregressive accelerator인 **Soracle**을 제안한다. 일반적인 LLM accelerator와 달리 text generation뿐 아니라 visual autoregressive model을 이용한 image generation까지 하나의 accelerator에서 처리한다. 논문에서는 mixed-precision MAC의 높은 hardware cost, weight와 K/V의 반복적인 external memory access, 그리고 낮은 probability token까지 모두 계산하면서 발생하는 불필요한 연산을 주요 문제로 본다.

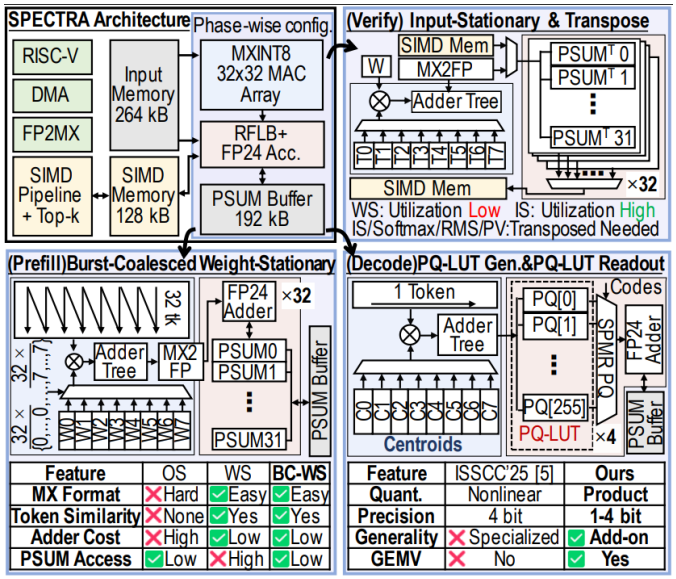
Soracle은 먼저 **Exponent-Tracked LUT-Counter MAC, ET LC-MAC**을 사용해 mixed-precision 연산 비용을 줄인다. 일반적인 multiplier 대신 LUT와 counter를 이용해 연산하고 exponent를 별도로 추적해 서로 다른 precision의 값을 accumulation한다. Canonical term decomposition과 equivalent-term coalescing을 이용해 counter activation도 줄인다. 측정 결과 conventional multiplier-based MAC과 비교했을 때 power를 71.7%, area를 53.7% 줄였다.

또 다른 핵심은 **row-broadcast/column-unicast(RBCU) 구조**이다. Query, score, probability처럼 짧게 사용되는 데이터는 row 방향으로 broadcast하여 spatial reuse하고, weight와 K/V처럼 반복 사용되는 데이터는 각 memory에 저장한 뒤 column 방향으로 전달해 temporal reuse한다. 이를 통해 attention head 내부에서 데이터를 최대한 재사용하며, attention execution의 EMA를 97.9%, data loading과 storing latency를 59.1% 줄였다.

이와 함께 slice-shared $\log_2$-partition quantization과 bandwidth-aligned streaming을 적용해 weight storage와 refill을 줄이고, 이전 generation의 dominant token 위치가 다음 generation에서도 이어지는 특성을 이용한 **dominance-trackable attention**으로 $Q \times K^T$와 $P \times V$ 연산량을 줄였다. QKV projection과 FFN weight는 각각 76.7%, 86.2% 감소하고 refill request 역시 각각 88.9%, 93.6% 감소했다.

28 nm FDSOI에서 제작된 Soracle은 20.25 mm^2 die area를 가지며, Infinity-125M을 이용한 image generation에서 **83.93 mJ/frame과 0.671 s/frame**을 달성한다. 기존 generation accelerator와 비교하면 image generation energy를 52.6%, latency를 36.7% 줄였으며 text generation도 함께 지원한다. MAC 구조부터 dataflow, quantization, attention sparsity까지 여러 단계의 최적화를 결합해 multi-modal autoregressive generation을 하나의 accelerator에서 효율적으로 처리했다는 점이 인상적이다.

#7-2 A SPECTRA: An Asymmetric-Precision Speculative Decoding LLM Accelerator with Product Quantization and Reconfigurable Flip-Flop Buffers in 28nm CMOS



[그림 2] Prefill-Decode-Verify에 따라 dataflow가 바뀌는 SPECTRA의 reconfigurable 아키텍처

두 번째 논문인 SPECTRA는 LLM의 **speculative decoding**을 hardware architecture에 직접 반영한 연구이다. LLM inference에서 prefill은 data reuse가 높은 GEMM이지만, decode는 한 token씩 GEMV를 반복하므로 external memory access의 영향을 크게 받는다. 논문에서는 speculative decoding의 특성을 이용해 정확도가 중요한 prefill과 verify에는 MXINT8을 사용하고, 오류가 발생해도 target model에서 다시 검증할 수 있는 draft decode에는 2-bit product quantization(PQ)을 사용한다.

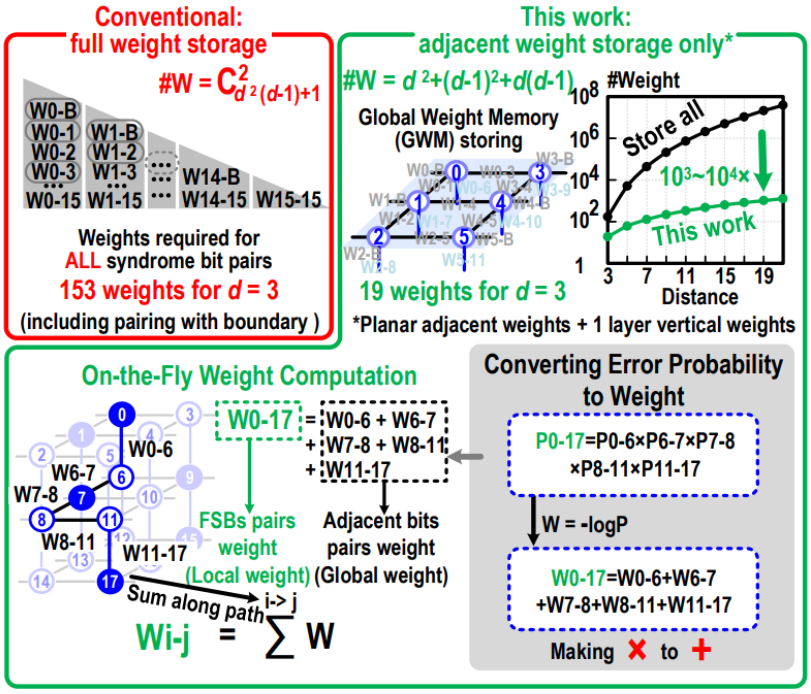
Hardware 측면에서 가장 중요한 부분은 **Reconfigurable Flip-Flop Local Buffer(RFLB)**이다. 동일한 RFLB가 prefill에서는 partial sum buffer, decode에서는 PQ lookup table, verify에서는 transpose buffer로 동작한다. Prefill에서는 burst-coalesced weight-stationary 방식으로 PSUM access를 8배 줄이고, clock gating과 input gating을 통해 RFLB 내부 power를 86% 줄여 전체 prefill stage energy를 24% 감소시켰다.

Decode에서는 RFLB를 **PQ-LUT**으로 재구성하고, 여러 PE가 동시에 LUT를 읽을 때 발생하는 area overhead를 줄이기 위해 **Shared-Prefix Multi-Readout(SPMR)** 구조를 사용한다. 256-entry LUT를 4개의 bank로 나누고 prefix decoder를 공유함으로써 높은 read throughput을 작은 area로 구현한다. 실제 AQLM과 PQCache workload에서 각각 8.6배, 13.3배 높은 throughput을 보였고, AQLM 기준 MXINT8 MAC array보다 energy는 5.9배, GEMV throughput은 7.7배 개선되었다.

Verify 단계에서는 RFLB의 transpose 기능을 이용해 input-stationary 방식으로 dataflow를 바꾸어 MAC utilization을 1.9~2.4배 높인다. 또한 PQCache를 이용해 long-context에서 중요한 KV entry만 가져오며, 12k token 조건에서 KV-cache energy와 latency를 각각 12.7배와 9.1배 줄였다.

SPECTRA는 28 nm CMOS에서 구현되었으며 148~635 MHz에서 동작한다. GSM8K, GPQA, PIQA, MATH500을 이용한 Llama-3.1-8B target / Llama-3.2-1B draft model 평가에서 accuracy loss 없이 MXINT8 baseline 대비 EMA 4.1배, latency 3.7배를 줄이고, FP16 baseline 대비 각각 7.7배와 7.3배를 줄였다. 하나의 accelerator를 모든 phase에 동일하게 사용하는 것이 아니라, **prefill, decode, verify의 서로 다른 연산 특성에 맞추어 precision과 dataflow, buffer 역할을 모두 바꾼다는 점**이 이 논문의 가장 큰 특징이라고 생각한다.

#7-3 A Full-MWPM Surface Code Decoder with On-the-Fly Weight Computation and Cross-Platform Adaptability Achieving 1.9X10^{-6} LER and 20.8-ns Decode Time at 4K



[그림 3] On-the-fly weight computation, lock-pair evaluation, parallel MWPM engine으로 구성된 cryo-CMOS decoder

세 번째 논문은 **quantum error correction(QEC)**을 위한 cryo-CMOS decoder이다. Surface code에서는 반복적인 parity check를 통해 flipped syndrome bit(FSB)를 검출하고, 이 FSB를 서로 matching하여 발생한 error를 추정한다. 정확도가 높은 대표적인 방법이 Minimum Weight Perfect Matching(MWPM)이지만, FSB 수가 증가할수록 가능한 matching 조합과 필요한 weight storage가 급격히 증가한다. 이 논문은 이러한 문제를 해결하면서 full MWPM을 4 K 환경에서 직접 수행하는 것을 목표로 한다.

첫 번째 핵심 기술은 **On-the-Fly Weight Computation(OFWC)**이다. 모든 FSB pair 사이의 weight를 memory에 저장하는 대신 인접한 syndrome bit 사이의 weight만 저장하고,

멀리 떨어진 FSB 사이의 weight는 decoding 시 해당 path의 adjacent weight를 더해서 계산한다. Error probability의 곱이 −log 변환 이후 weight의 합으로 바뀌는 특성을 이용한 것이다. 최대 6개의 local weight를 한 cycle에 계산하며, 이를 통해 weight memory area를 989배 줄이면서 distance 21까지 arbitrary weight를 지원한다.

두 번째 핵심은 **lock-pair-based speed tuning과 parallel MWPM engine**이다. 가장 작은 weight를 갖는 FSB pair를 일정 개수 먼저 lock하고 관련된 다른 pair를 search 대상에서 제거함으로써 남은 FSB 수를 줄인다. Lock-pair 수는 configurable하여 서로 다른 qubit platform의 coherence time에 맞추어 decoding speed를 조절할 수 있다. 이후 남은 FSB는 depth-first search 기반 parallel MWPM engine에서 처리하고, unmatched FSB가 6개 이하가 되면 7개의 match core가 서로 다른 후보를 동시에 계산한다.

40 nm test chip은 실제 4 K cryostat에서 측정되었으며, surface code distance d=9, PER=10^{-3} 조건에서 **1.9×10^{-6} LER과 20.8 ns decode time**, 2.46 nJ/decode를 달성했다. Full MWPM decoding threshold는 약 0.047이며, d=11, PER=10^{-3} 조건에서는 4개의 pair를 lock했을 때 average decode time이 17,870배 감소한다.

기존 surface-code decoder가 speed를 위해 approximate algorithm을 사용하는 경우가 많은 것과 달리, 이 논문은 **weight storage와 matching architecture 자체를 바꾸어 full MWPM을 hardware로 구현했다는 점**이 인상적이다. 또한 decoder를 4 K stage에 직접 배치하고 lock-pair 수를 조절해 서로 다른 qubit platform에 대응하도록 했다는 점에서 algorithm뿐 아니라 실제 quantum computing system과의 integration까지 고려한 설계라고 생각한다.

이번 세 논문을 같이 보면 accelerator와 dedicated processor 연구가 단순히 연산량을 높이는 방향에서 벗어나 **workload가 실제로 갖는 구조적인 특성을 hardware에서 직접 활용하는 방향**으로 발전하고 있다는 점이 잘 보인다. Soracle은 attention 내부의 reuse와 dominant token을 활용하고, SPECTRA는 prefill·decode·verify의 서로 다른 특성에 따라 hardware를 재구성하며, MWPM decoder는 모든 weight를 저장하지 않고 필요한 순간에 계산한다.

결국 세 논문 모두 "무엇을 더 빠르게 계산할 것인가"보다 **무엇을 굳이 저장하거나 이동하거나 계산하지 않아도 되는가**를 먼저 찾고, 이를 회로와 architecture로 구현한다는 공통점이 있다. Memory access와 data movement가 점점 더 큰 system bottleneck이 되는 상황에서 이러한 workload-aware optimization이 앞으로도 중요한 설계 방향이 될 것으로 보인다.

저자정보



박승현 박사과정 대학원생

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

VLSI 2026 Review

울산과학기술원 전기및전자공학부 석박통합과정 홍기업

Session 6: Intelligent Vision and Perception Sensors

2026 IEEE/JSAP Symposium on VLSI Technology & Circuits의 Session C6는 Intelligent Vision and Perception Sensors라는 주제로 총 4편의 논문이 발표되었다. Hybrid processing-in-sensor를 이용한 filter-less color vision(C6.1), triple conversion gain과 LOFIC을 결합한 100 dB single-exposure HDR CIS(C6.2), clocked-recharging SPAD와 TTFS encoding을 이용한 neuromorphic image sensor(C6.3), scene-adaptive power saving과 object-of-interest 추출을 결합한 all-digital EVS(C6.4)가 소개되었다. 각 논문은 서로 다른 pixel 및 readout 구조를 사용하지만, sensor 내부에서 데이터 표현을 단순화하고 필요한 정보만 선별하여 전송함으로써 high dynamic range, low power, high throughput 및 edge intelligence를 동시에 확보하려는 공통된 방향성을 보인다.

#6.1 Carnegie Mellon University에서 발표한 논문으로, Computational Mosaic Array(CMA)와 residual-enhanced analog skip connection을 결합한 hybrid processing-in-sensor(PIS)를 제안한다. 기존 processing-in-pixel(PIP)은 높은 병렬성을 제공하지만 pixel area와 weight resolution이 제한되고, processing-in-column(PIC)은 복잡한 연산이 가능하지만 2-D convolution을 위한 row buffer와 memory 부담이 크다. 이를 해결하기 위해 3×3 CMA에서 일부 pixel만 triple-well photodiode와 analog front end로 사용하고 중앙 pixel에 convolution 결과를 저장함으로써 photodiode area를 55% 절감하면서 4-b in-pixel convolution을 구현하였다. Triple-well photodiode는 별도의 color filter 없이 R/G/B 성분을 동시에 획득하며, time-domain AFE가 3-b magnitude와 1-b sign으로 구성된 weight를 current accumulation에 반영한다.

CONV1 출력은 column-wise median 및 max pooling을 통해 2-D feature map에서 21×1 time-domain representation으로 변환되며, 이에 따라 데이터 양을 98.8% 줄이고 row-buffer memory를 제거한다. 이후 3×1 CONV2의 출력은 upscaling된 뒤 CONV1 결과와 analog skip connection으로 합산되고, CONV3와 TDC readout을 거쳐 off-chip classifier로 전달된다. 특히 residual-enhanced network는 analog weight mismatch와 noise가 증가할

때 plain network보다 작은 정확도 저하를 보여 analog PIS의 비이상성에 대한 내성을 향상시켰다. 180-nm CMOS에서 제작된 126×126 sensor array는 40 fps에서 62.08 μW를 소모하고 97.76 pJ/pixel·frame의 iFoM 및 97.25%의 face-detection accuracy를 달성하였다.

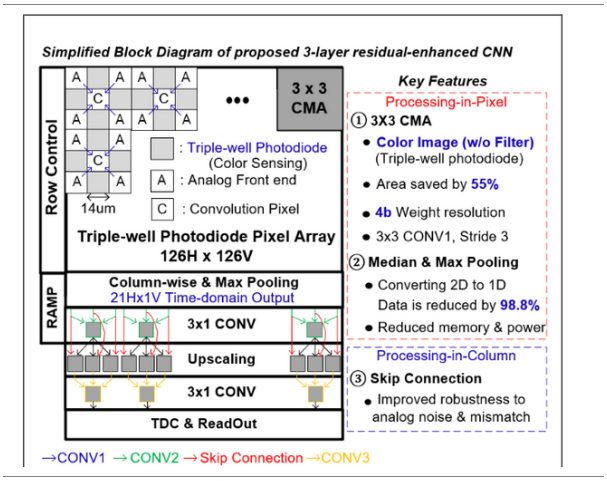


그림. 제안된 3-layer residual-enhanced PIS 전체구조

#6.2 Sony Semiconductor Solutions와 Sony Semiconductor Manufacturing에서 발표한 논문으로, 1.22 μm pixel pitch에서 triple conversion gain과 transfer-gate(TRG)-type LOFIC을 결합한 50-Mpixel single-exposure HDR image sensor를 제안한다. 기존 sub-pixel HDR 구조는 하나의 pixel에 두 개의 photodiode를 배치하므로 2 μm 이하 pitch에서 면적 확보가 어렵다. 제안 pixel은 4개의 photodiode가 FD1을 공유하며, FDG를 통해 metal capacitance가 추가된 FD2를 형성하고 FCG를 통해 대용량 3-D MIM LOFIC capacitor에 연결된다. 조도가 증가하면 charge가 PD에서 FD1, FD2, LOFIC capacitor 순서로 overflow되므로 1st gain, 2nd gain, LOFIC gain의 세 구간에서 넓은 입력 범위를 읽을 수 있다.

TRG 아래에 overflow path를 형성하면 PD의 charge-storage capacity가 감소할 수 있으므로, 본 설계는 PD maximum potential과 FD reset potential을 함께 높였다. 이때 RST_VDD와 source-follower의 SF_VDD를 분리하여 reset headroom은 확보하면서 source-follower의 전력 증가는 방지하였다. 또한 exposure 동안 MIM_VDD를 3.3 V에서 2.4 V로 낮춰 LOFIC storage node의 electric field를 완화하고 FD dark current를 5,000 e^-/s에서 3,700 e^-/s로 감소시켰다. 90-nm CIS와 22-nm ASIC으로 구현된 sensor는 최대 analog gain에서 0.73 e^-rms read noise, LOFIC gain에서 250 ke^- full-well capacity, 그리고 100 dB 이상의 single-exposure dynamic range를 달성하였다.

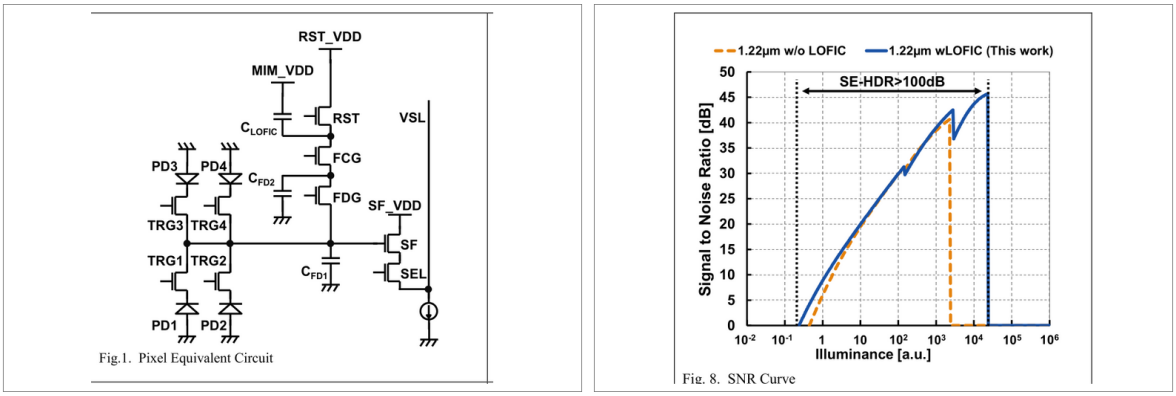


그림. Triple conversion gain과 TRG-type LOFIC을 구현한 pixel equivalent circuit(좌)과 기존 DCG sensor 대비 100 dB 이상의 single-exposure dynamic range를 보이는 SNR curve(우)

#6.3 UNIST와 Sogang University에서 발표한 논문으로, clocked-recharging(CR) SPAD pixel이 광세기를 time-to-first-spike(TTFS)로 직접 encoding하는 특성을 이용한 neuromorphic image sensor를 제안한다. 기존 CNN 기반 PIS는 dense multi-bit data와 반복적인 accumulation으로 인해 데이터 이동과 연산 에너지의 부담이 남아 있는 반면, TTFS 기반 SNN은 하나의 spike 발생 시점만으로 입력 크기를 표현하므로 높은 sparsity와 낮은 latency를 제공한다. 96×96 pixel array의 각 pixel은 CR SPAD, 4-b TTFS memory, 8-b exponential counter, 5-b SRAM 및 pulse generator를 포함하며, 3×3 sub-pixel-array(SPA) 단위로 comparator와 output memory를 공유하여 array-parallel convolution을 수행한다.

Intensity mode에서는 recharging cycle 내 첫 spike를 TS0–TS3의 one-hot code로 저장하고, spike timing에 따라 8-b counter에 +8, +4, +2, +1을 누적한다. 16회의 averaging으로 dark-count 영향을 줄인 뒤, convolution mode에서는 counter의 상위 4-b로 dominant spike timing을 decode하고 SRAM의 5-b weight를 pulse 수와 polarity로 변환한다. Chopping analog counter의 4-b CDAC은 TS0–TS3에 각각 8C, 4C, 2C, 1C의 temporal weight를 부여하며, threshold를 negative weight로 함께 주입해 단일 comparator로 integrate-and-fire를 구현한다. Input-stationary dataflow를 통해 동일한 spike input을 재사용하므로 filter마다 다시 exposure하거나 encoding할 필요가 없다. 110-nm BSI 공정으로 제작된 prototype은 150 μs frame time에서 6.67 kfps로 동작하며, convolution mode에서 5.12 mW를 소모한다. 측정된 compute power는 0.5 mW이고, 2.6 pJ/pixel·fps·filter의 FoM, 3.86 TOPS/W의 energy efficiency 및 91.5% face-detection accuracy를 달성하였다.

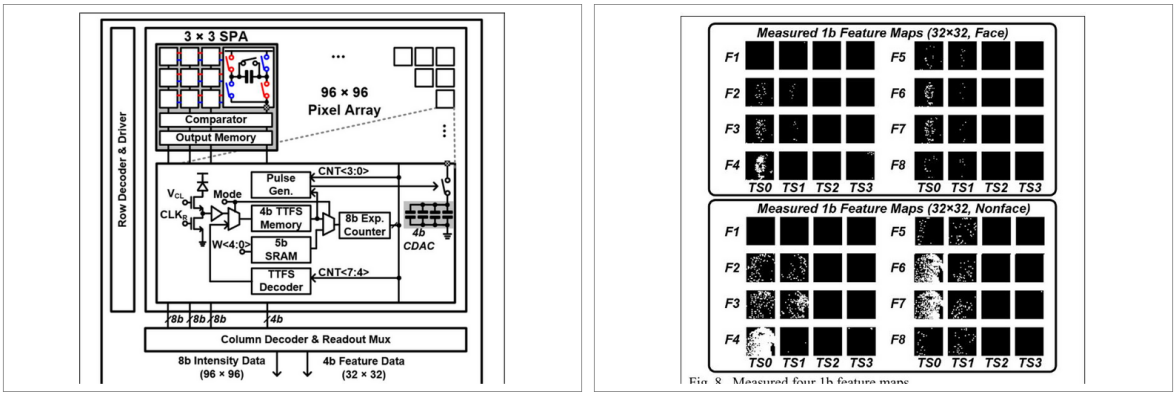


그림. CR SPAD, TTFS memory, exponential counter 및 SPA-level IF 회로로 구성된 overall sensor architecture(좌)와 face/non-face 입력에 대해 측정된 1-b feature maps(우)

#6.4 Sungkyunkwan University와 SolidVue에서 발표한 논문으로, scene-adaptive pixel power saving과 object-of-interest(OOI) selective readout을 결합한 all-digital synchronous event-based vision sensor(EVS)를 제안한다. 200×128 SPAD array는 saturation-based(SAT-PS), temporal(T-PS), spatial(S-PS)의 세 가지 power-saving 기법을 사용한다. SAT-PS는 pixel output이 saturation에 도달하면 이후의 불필요한 counting을 중단하고, T-PS는 짧은 sub-window에서 급격한 intensity change를 검출한 뒤 pixel을 비활성화한다. S-PS는 3개의 primitive pixel과 1개의 reference pixel로 구성된 2×2 macro-pixel에서 temporal intensity trend를 비교하고, 인접 pixel이 유사할 경우 primitive pixel을 끈 뒤 reference event로 overwrite하여 공간적으로 중복된 동작을 줄인다.

Event decision 및 interpolation logic에서 생성된 event는 48-row kernel buffer에 저장되는 동시에 3-layer hybrid SNN–BNN으로 전달된다. 첫 번째 SNN layer가 spatial-temporal filtering을 수행하고, 두 개의 BNN layer가 1-b weight와 activation을 이용해 8×8 tile 단위의 OOI confidence map을 생성한다. 이 OOI mask로 readout을 gating함으로써 관심 영역의 event만 외부로 전송하고, 선택된 event 수를 이용해 다음 frame의 threshold도 scene-adaptive하게 갱신한다. 90-nm 공정의 prototype은 200 lux에서 9.85 mW를 소모하며 pixel 및 core power를 각각 47.6%와 21.6% 절감하였다. S-PS 단독 조건에서는 최대 53.77%, 세 기법이 동시에 동작하는 composite scene에서는 51.4%의 pixel power saving을 보였다. 측정 throughput은 DAQ에 의해 제한된 5.58 kfps 및 714 Meps이다. On-chip OOI block은 IR-drop으로 목표 100 MHz에서 안정적으로 동작하지 않아 동일 network를 FPGA로 구현해 검증하였으며, sensor-captured event image에서 0.7452의 Dice score를 달성하였다.

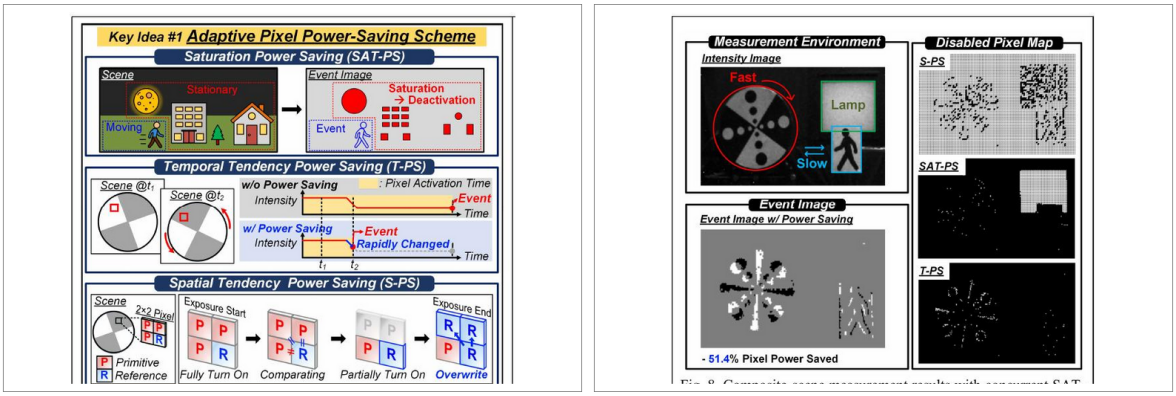


그림. SAT-PS, T-PS 및 S-PS로 구성된 scene-adaptive pixel power-saving 개념(좌)과 세 기법을 동시에 적용한 composite-scene 측정 결과(우)

저자정보



홍기업 석박사통합과정 대학원생

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

2026 VLSI Review

한국과학기술원 전기및전자공학부 박사과정 윤웅노

Session 10 Building Blocks for Multimodal Sensors

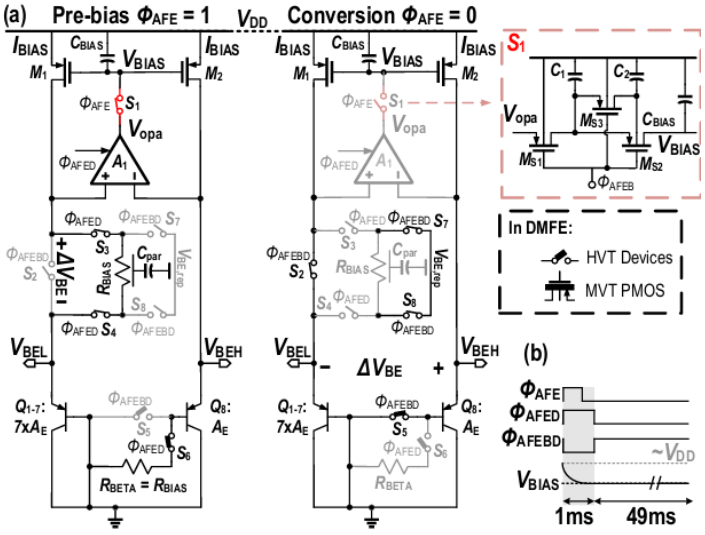
이번 2026 VLSI의 Session 10은 Sensors, Imagers, IoT, MEMS, Display Circuits 주제의 하위 분류로 총 5편의 논문이 발표되었다. 세부적으로는 온도 센서와 MEMS 타입의 멀티모달 센서를 ΔΣ 변조기로 디지털화하는 10.1, 10.2번 논문, coarse-fine 방식으로 강한 배경 조명 조건에서도 측정을 원활히 할 수 있도록 한 LiDAR 센서에 관한 10.3번 논문, 그리고 기업에서 PMIC를 위한 전류 센서(10.4)와 3D IC 내 국소 위치별 온도-전압 센싱을 다룬 연구(10.5)가 선정되었다.

본 리뷰에서는 이들 중 온도 센싱에 관하여 더 자세히 살펴보고자 한다. 온도 센서는 목표로 하는 어플리케이션에 따라 정확도, 분해능, 전력, 변환 시간, 그리고 양산 단계에서의 트리밍/캘리브레이션 비용 중 무엇을 우선할 것인지가 크게 달라지며, 그 선택이 곧 회로 구조와 테크닉을 결정한다. 따라서 각 논문이 어떤 지표를 확보하기 위해 무엇을 포기하였는지를 비교하며 읽어볼 만하다. 이에 always-on 어플리케이션을 겨냥하여 100nW 이하의 전력에서 정밀도를 확보한 10.1번 논문과 AI 시스템을 위한 연산기에서 온도와 전압을 짧은 시간 내에 센싱하는 10.5번 논문을 살펴본다.

#10-1 온도 센서는 센싱 소자에 따라 크게 BJT, MOS, 저항 기반으로 나뉜다. 이 중 BJT 기반 센서는 공정 산포에 의한 오차의 대부분을 1-point trim만으로 0.1℃ 수준의 정확도로 보정할 수 있어, 오랜 기간 정밀 온도 센서의 표준으로 자리해 왔다 [1]. 다만 BJT를 바이어스하는 전류원과 이를 정의하는 앰프가 상시 동작하여야 하므로 소비 전력이 100nW를 넘는 것이 일반적이었고, 이는 배터리나 에너지 하베스팅으로 동작하는 어플리케이션에서의 채용을 어렵게 하는 요인이었다.

10.1번 논문은 BJT 기반 온도 센서로, always-on IoT와 바이오메디컬 어플리케이션을 겨냥하였다. 180nm 공정에서 1V 공급 전압으로 75nW만을 소모하면서 -40°C에서 85°C 범위에 대해 1-point trim 후 ±0.1°C (3σ)의 정확도를 달성하였는데, 이는 동급 정확도를 갖는 기존 정밀 BJT 기반 센서 대비 3배 낮은 전력이며 동시에 저전력 MOS 기반 센서 대비 2~3배 높은 정확도이다. 온도 범위로 정규화한 상대 부정확도(R-IA)로 보면 0.16%로, sub-μW 영역에서 최고 수준에 해당한다.

논문에서 제안하는 핵심 아이디어는 dual-mode front-end (DMFE)로, 프론트엔드를 pre-bias phase와 conversion phase의 두 상태로 나누어 동작시키는 것이다. Pre-bias phase에서는 일반적인 BJT-기반 reference에서와 같이 앰프를 통해 저항 R_{BIAS} 양단의 ΔV_{BE}를 강제하여 PTAT 전류 I_{BIAS}를 생성하고, 프론트엔드가 정착한 뒤 M1,2의 게이트 전압 V_{BIAS}를 5pF 크기의 cap에 저장한다. 이어지는 conversion phase에서는 A1을 완전히 꺼버리고, 프론트엔드를 CTAT 전압 V_{BEH}와 PTAT 전압 ΔV_{BE}를 생성하는 구성으로 재배치하여 ΔΣ 변조기가 이를 샘플링하도록 한다. 그림 1(b)의 타이밍에서 볼 수 있듯 pre-bias는 1ms, conversion은 49ms이므로 20nA를 소모하는 앰프가 켜져 있는 시간은 전체의 2%에 지나지 않는다. 커패시터에 바이어스 상태를 저장해 두고 전력 소모가 큰 블록을 꺼두는 duty cycling 자체는 초저전력 아날로그 회로에서 널리 쓰이는 테크닉이지만, 이를 0.1°C급 정밀 온도 센서의 프론트엔드에 적용하면서도 정확도를 잃지 않도록 만든 점이 본 논문의 기여라 할 수 있다.

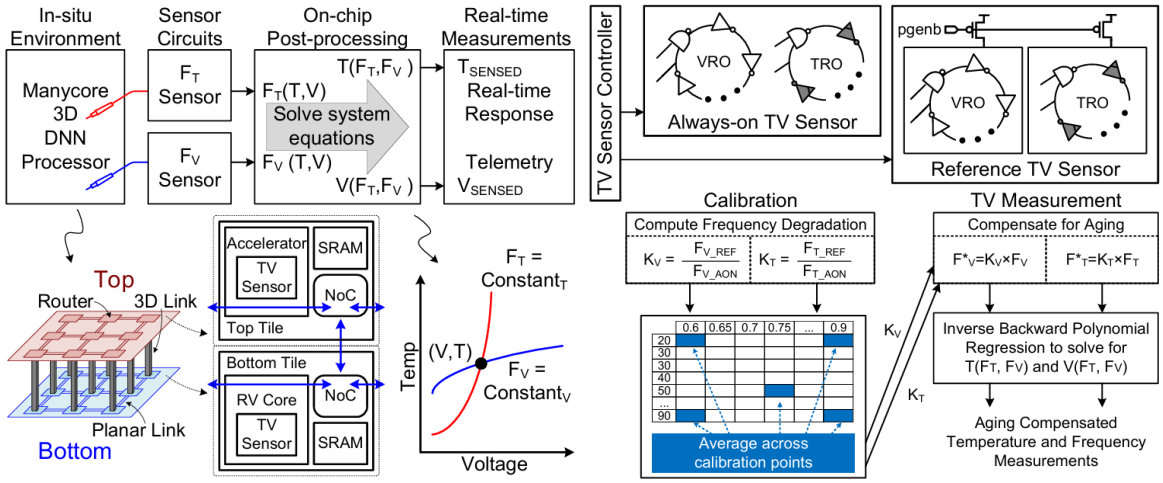


[그림 1] 제안된 PNP 기반 dual-mode front-end (DMFE)의 2-phase 동작 (a)과 해당 타이밍 다이어그램 (b) (원문 Fig. 1)

#10-5 AI 시스템에서는 집적된 디지털 회로의 동작에 따라 국소 면적의 온도와 전압이 크게 흔들리고, 이는 곧 연산 속도와 신뢰성에 영향을 준다. 따라서 온도-전압 센서를 칩 전면에 어레이로 배치하여 모니터링하면 핫스팟이 생긴 코어에만 선택적으로 주파수를 낮추는 동작이 가능해지고, 모든 코어를 최악 조건에 맞추어 묶어 두는 가드밴드 또한 줄일 수 있다. 이를 위해 센서는 (1) 어레이로 이용할 만큼 작고, (2) 워크로드에 따른 변동을 따라갈 만큼 빠르며, (3) 모니터링 대상 로직과 동일한 디지털 공급 전압에서 외부 기준 없이 동작하여야 한다. 위의 목적을 달성하기 위하여 standard 셀로 링 오실레이터를 구성하여 이를 센서로 이용하였다. 실제로 본 논문의 센서는 하이브리드 본딩된 3D

IC DNN 프로세서 [2]의 상·하부 다이에 216μm 이하의 간격으로 배치되었으며, Intel 3와 Intel 18A 두 공정에 각각 구현되었다.

본 논문은 동일 그룹이 2025 ISSCC에서 발표한 연구를 공정을 옮기며 확장한 것이다 [3]. 선행 연구의 핵심 아이디어는 sensor fusion으로, 온도에만 혹은 전압에만 반응하는 이상적인 센서를 회로로 만들려 애쓰는 대신 두 변수 모두에 반응하되 감도가 서로 다른 두 개의 오실레이터를 두고, 그 주파수 쌍 (FT, FV)이 (T, V) 평면에서 단 하나의 점만을 지정한다는 성질을 이용하여 온도와 전압을 동시에 푸는 것이다 [그림 2]. 센서마다 다른 공정 산포는 회로가 아니라 센서별 역다항식 회귀 모델이 흡수하며, 소수의 캘리브레이션 포인트만으로 그 모델을 세우기 위해 계수 벡터가 저차원 부분 공간에 놓인다는 점을 이용한 주성분 분석(PCA)이 사용된다.



[그림 2] TV 센서의 동작 개요와 3D IC 내 배치 (좌, 원문 Fig. 2) Always-on 센서와 파워 게이팅된 reference 센서의 주파수 비를 이용한 노화 보상 방법 (우, 원문 Fig. 5)

본 논문이 여기에 더한 것은 크게 네 가지이다. 첫째, 서브 센서가 단순해졌다. 온도 센서용 오실레이터는 3T 기준 전압과 current-starved 구조를 쓰던 것에서 diode-connected pull-up을 포함한 지연 셀로, 전압 센서용 오실레이터는 standard 인버터만으로 구성되어 완전한 디지털 표준 셀 설계가 되었으며 면적도 줄었다. 둘째, 응답이 빨라져 온도는 2μs, 전압은 100ns의 시간 분해능을 갖는다. 셋째, 계산의 위치가 바뀌었다. 측정이 끝나면 센서가 인터럽트로 RISC-V 코어를 깨우고, 코어가 역다항식을 계산하여 온도 맵을 만들거나 스로틀링 신호를 내보낸다. 넷째, 그리고 가장 눈여겨볼 부분으로 reference 센서를 이용한 노화(aging) 보상이 추가되었다. 상시 동작하는 always-on 센서와 평소 파워 게이팅되어 거의 열화되지 않은 reference 센서의 주파수 비를 캘리브레이션 포인트에서 평균하여 스케일 계수로 삼고, 측정된 카운트에 이를 곱한 값을 원래의 역다항식에 그대로 넣는다. 즉, 모델을 다시 세우는 대신 입력 자체를 보상하는 방식이다.

어플리케이션 측면에서의 성능으로는, 3D IC DNN 프로세서에서 실제 DNN 워크로드를 구동하며 상부 다이 56개 코어의 실시간 온도 맵을 얻었고, 이를 RISC-V 코어가 처리하여 목표 온도를 넘는 코어의 가속기에 스로틀링을 걸었다. 700MHz에서 스로틀링을 켠 경우 최대 온도를 목표 이하로 유지하면서 정규화 지연은 0.76으로, 외부 주파수 스케일링 (400MHz 고정, 정규화 지연 1.0)과 비교하여 24% 낮은 지연과 더 낮은 최대 온도를 동시에 달성하였다.

결론적으로 본 논문은 센서 회로 자체의 정밀도를 끌어올리는 대신, 비이상적인 두 오실레이터의 정보를 계산으로 합치고 양산 캘리브레이션부터 수명 동안의 열화까지 흐름 전체를 설계한 연구로 읽힌다. 같은 세션의 10.1번 논문이 아날로그 회로 테크닉으로 오차를 제거하여 100ms의 변환 시간을 대가로 ±0.1°C를 얻었다면, 10.5번 논문은 오차를 모델과 계산으로 흡수하여 수십 μm²의 면적과 μs 단위의 응답을 얻고 °C 단위의 정확도를 감수하는 트레이드오프를 취했다.

참고문헌

[1] K. A. A. Makinwa, "Smart Temperature Sensor Survey," [Online].

[2] P. Knag et al., ISSCC, 2026.

[3] H. S. Ozturk et al., "A Dual VDD-Temperature Sensor Employing Sensor Fusion with 2.4°C, 9mV (±3σ) Inaccuracy in 65nm CMOS," ISSCC, pp. 170-172, 2025.

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/

VLSI 2026 Review

고려대학교 전기및전자공학부 박사과정 안재웅

Session 18 Display Interfaces and Data Conversion Circuits

이번 2026 VLSI의 Session 18에서는 총 5편의 논문이 발표되었다. 이 중 #18-1은 2880-channel OLED DDI에 Miller-charge cancellation (MCC)과 fast-slew input (FSINP)을 적용하여, 모든 channel이 동시에 switching하는 조건에서도 13.3 V/μs의 slew rate를 달성한 내용이다. #18-4는 OLEDoS용 11-bit source-driver IC에 3-stage resistor-resistor-embedded (RRE) DAC과 multi-step offset calibration을 적용하여 1,723 μm²/channel의 면적과 1.6 mV의 DVO, 0.69 μs의 settling time을 달성하였다. 마지막으로 #18-5는 모든 current source를 활용하는 cross-point (XP) DAC과 improved correlated double integration (I-CDI)을 제안하여 mismatch와 amplifier offset을 줄이고, 1,223 μm²/channel의 면적과 4.3 mV의 DVO를 달성한 내용이다.

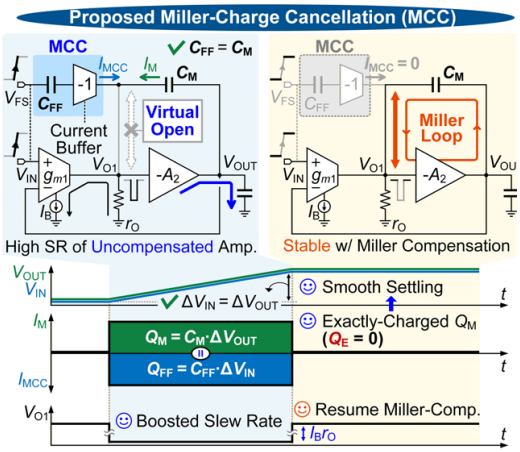
#18-1 논문은 고해상도·고주사율 OLED display에서 source driver의 구동 속도를 높이기 위한 display driver IC (DDI)에 관한 내용이다. 최근 OLED display의 해상도와 frame rate가 증가하면서 한 개의 row에 데이터를 기록할 수 있는 1-H time이 짧아지고 있으며, 이에 따라 source amplifier가 큰 panel load를 빠르게 구동해야 한다. 특히 기존 Miller-compensated amplifier는 큰 입력 신호가 인가될 때 Miller capacitor (C_M)을 bias current로 충전해야 하기 때문에 slew rate가 (I_B/C_M)으로 제한된다. 또한 수천 개의 source channel이 동시에 동일한 reference voltage를 선택하면 global resistor string에 큰 dynamic load가 발생하여 reference voltage가 흔들리고 settling time이 증가하는 문제가 있다.

이를 해결하기 위해 본 논문에서는 Miller-charge cancellation (MCC) 기법을 제안하였다. MCC는 source amplifier의 입력과 내부 current buffer 사이에 feedforward capacitor (C_{FF})를 추가하고, $C_{FF}=C_M$으로 설정하여 입력 전압이 변할 때 발생하는 displacement current를 이용해 Miller capacitor에 필요한 charge를 직접 공급한다. 따라서 slew 구간에서는 Miller capacitor의 영향을 사실상 제거한 amplifier처럼 빠르게 동작하고, 입력 transition이 끝나면 C_{FF}를 통한 전류가 자연스럽게 사라지면서 기존 Miller compensation이 다시 동작하여 안정적인 settling을 유지한다. 기존 comparator 기반의 slew-rate enhancement와 달리 별도의 입력 offset이나 delay를 이용하여 slew 종료 시점을 판단하지 않기 때문

에 under-charge 또는 over-charge에 의한 settling error도 줄일 수 있다. 또한 기존 folded-cascode summing circuit을 current buffer로 재사용하여 추가 transistor 없이 MCC를 구현함으로써 area 및 static power overhead를 최소화하였다.

한편, 모든 channel이 동시에 동작할 때 global R-string에 집중되는 capacitive load를 줄이기 위해 fast-slew input (FSINP) 기법도 함께 제안하였다. MCC의 입력을 실제 선택된 reference node (V_R[k])에 직접 연결하는 대신 인접한 (V_R[k-N])에 연결하여 수천 개 channel의 capacitive load를 여러 R-string tap으로 분산한다. 이때 MCC는 입력의 DC level이 아니라 voltage transition 크기에 의해 동작하기 때문에 서로 다른 reference tap을 사용하더라도 Miller charge cancellation 동작은 유지할 수 있다.

제안된 DDI는 28-nm CMOS 공정으로 제작되었으며, 2880개의 source channel을 이용하여 3200×2880 해상도의 6.9-inch OLED panel을 120 Hz에서 구동하였다. 16개 channel을 동작시킨 경우 MCC를 적용하여 slew rate가 기존 4.8 V/μs에서 15.7 V/μs로 약 3.3배 향상되었으며, 2880개 channel이 동시에 transition하는 worst-case 조건에서도 MCC와 FSINP를 함께 적용하여 13.3 V/μs의 slew rate를 달성하였다. 또한 전체 2880개 channel의 slew-rate variation은 σ_{SR}=7.5%로 측정되었다. 실제 OLED panel에서도 기존 구조에서 나타나던 color imbalance와 smearing이 감소했으며, luminance dispersion이 약 10%에서 2%로 감소하여 고해상도 OLED display에서 제안 기법의 효과를 실제 panel 수준에서 검증하였다.



[그림 1] #18-1에서 제안된 Miller-charge cancellation (MCC) 기법

#18-4 논문은 초고해상도 OLED-on-Silicon (OLEDoS) display를 위한 11-bit source-driver IC에 관한 내용이다. OLEDoS는 높은 pixel density를 구현해야 하기 때문에 source-driver의 channel 면적을 매우 작게 만들어야 하는 동시에, 높은 frame rate를 위해 빠른 settling이 요구된다. 또한 OLED pixel은 작은 전류로 동작하기 때문에 channel 간 output voltage 편차인 deviation of voltage output (DVO)에 민감하며, 이는 display uniformity 저

하로 이어질 수 있다. 기존 multi-stage DAC이나 low-voltage DAC 구조는 면적을 줄일 수 있지만, sequential switching에 따른 속도 저하나 channel 수가 증가했을 때 reference-line IR drop으로 DVO가 커지는 문제가 존재한다. 본 논문에서는 이러한 area, speed, DVO 문제를 동시에 해결하기 위해 3-stage resistor-resistor-embedded (RRE) DAC 구조를 제안하였다.

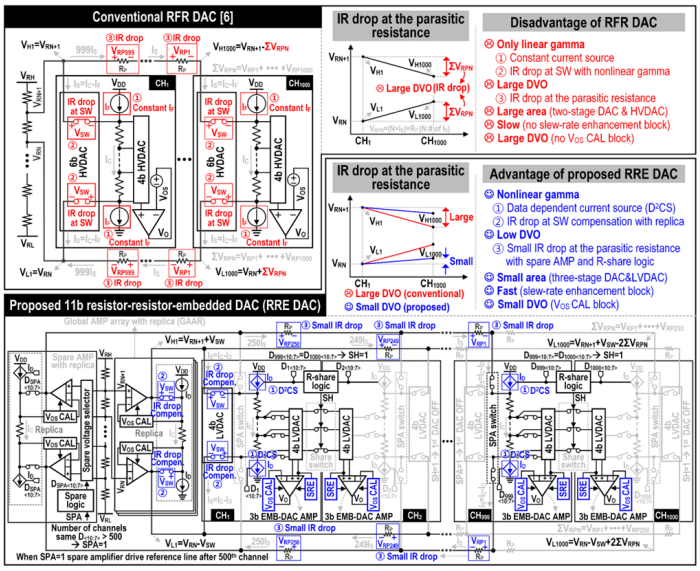
제안된 11-bit RRE DAC은 4b+4b+3b의 3-stage 구조로 구성되며, low-voltage DAC을 함께 사용하여 transistor 수와 device size를 줄인다. 기존 resistor-floating-resistor (RFR) DAC에서는 nonlinear gamma voltage를 생성할 때 coarse DAC에서 선택된 (V_H), (V_L)에 따라 R-string current가 달라지고, 이로 인해 mismatch current가 coarse switch와 reference line을 흐르면서 IR drop이 발생한다. 특히 수백~수천 개의 channel이 동일한 MSB code를 선택하면 reference-line resistance를 통해 흐르는 전류가 누적되어 DVO가 크게 증가할 수 있다. 이를 해결하기 위해 본 논문은 input data에 따라 전류를 조절하는 data-dependent current source (D^2CS)를 적용하여 mismatch current를 줄였으며, replica coarse switch와 fine R-string을 global amplifier에 구성하여 compensating voltage를 생성함으로써 nonlinear gamma 동작에서 V_H와 V_L의 정확도를 높였다.

또한 대규모 channel이 동시에 동작할 때 발생하는 reference-line IR drop을 줄이기 위해 spare amplifier와 R-share logic을 제안하였다. 500개 이상의 channel이 동시에 활성화될 경우 spare amplifier가 기존 amplifier의 driving load를 분담하며, 인접 channel이 동일한 MSB 값을 선택하는 경우에는 R-share logic을 통해 동일한 R-string voltage를 공유한다. 이를 통해 reference line을 흐르는 전류를 줄여 1000-channel load 조건에서도 DVO 증가를 억제한다. 한편 global amplifier의 offset은 서로 다른 sourcing/sinking amplifier 사이의 voltage 차이를 유발하여 DAC monotonicity를 깨뜨릴 수 있기 때문에, multi-step offset calibration unit (MOCU)을 적용하였다. MOCU는 5-bit coarse calibration 이후 5-bit fine calibration을 수행하여 residual offset을 줄이면서 calibration 회로의 면적과 시간을 줄였다.

고속 동작을 위해서는 high-pass-filter-assisted dynamic current source (HPDCS)를 이용한 slew-rate enhancement 기법을 적용하였다. HPDCS는 dynamic-bias source follower와 high-pass filter, dynamic current source로 구성되며, 입력이 transition할 때만 dynamic current source가 활성화되어 amplifier의 slew rate를 증가시킨다. 정상 상태에서는 dynamic current source가 꺼지기 때문에 추가적인 static current를 소모하지 않는다. 측정 결과 최대 12 V/μs의 slew rate와 0.69 μs의 settling time을 달성하였다.

제안된 source-driver IC는 65-nm CMOS 공정으로 제작되었다. DAC의 1 LSB는 0.49 mV이며, 최대 DNL과 INL은 각각 0.30 LSB와 0.55 LSB로 측정되었다. MOCU 적용 전 최대 26.2 mV였던 DVO는 1.3 mV까지 감소하였으며, 1000-channel load 조건에서도 spare

amplifier와 R-share logic을 적용하여 최대 DVO를 1.6 mV로 제한하였다. 최종적으로 11-bit resolution을 지원하면서 1,723 μm²/channel의 channel area, 1.6 mV의 DVO, 0.69 μs의 settling time을 달성하여 기존 연구 대비 작은 channel area와 낮은 DVO, 빠른 settling 성능을 동시에 확보하였다.



[그림 2] #18-4에서 제안된 11b resistor-resistor-embedded DAC (RRE DAC)

#18-5 논문은 mobile OLED display용 10-bit source-driver IC의 channel 면적과 channel 간 출력 전압 편차(DVO)를 동시에 줄이기 위한 DAC 구조에 관한 내용이다. 고해상도 OLED에서는 많은 source channel을 한 chip에 집적해야 하기 때문에 channel 면적을 줄이는 것이 중요하며, 특히 DAC가 channel 면적에서 큰 비중을 차지한다. 기존 segmented DAC는 global resistor-string에서 생성한 V_H, V_L 사이를 in-channel sub-DAC가 세부적으로 보간하는 구조를 사용하지만, sub-DAC 면적을 줄일수록 transistor mismatch가 증가하여 channel 간 deviation of voltage output (DVO)가 커지는 trade-off가 존재한다. 또한 기존 clock-based sub-DAC는 동작 속도가 느리고, charge-modulation QM DAC는 비교적 빠르게 동작하지만 current-source mismatch에 민감하여 device size를 충분히 줄이기 어렵다는 문제가 있다. 본 논문에서는 이를 해결하기 위해 cross-point (XP) DAC와 improved correlated double integration (I-CDI) 기법을 제안하였다.

기존 QM-DAC에서는 선택된 binary-weighted current source를 이용해 amplifier output의 slope를 만들고, 별도의 full-scale current source로 reference capacitor를 방전하여 일정한 integration time (T_C)를 생성한다. 따라서 output voltage는 선택된 current의 slope와 고정된 T_C의 곱으로 결정되는데, 이 구조는 T_C 생성을 위한 추가적인 full-scale current source가 필요하고, 선택되지 않은 current source들은 conversion 과정에서 사용되지 않는다.

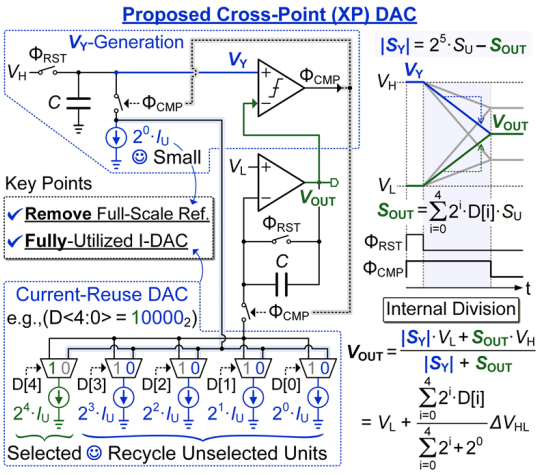
또한 T_C를 만드는 current와 output slope를 만드는 current 사이의 mismatch가 직접 output error로 나타나는 문제가 있다.

이를 개선하기 위해 제안된 XP-DAC는 선택되지 않은 current source까지 모두 재사용한다. 입력 code에 의해 선택된 current들은 V_{OUT}을 상승시키는 slope (S_{OUT})을 만들고, 선택되지 않은 current들은 반대쪽 node (V_Y)를 하강시키는 slope (S_Y)를 형성한다. Comparator는 두 전압 V_{OUT}과 V_Y가 만나는 cross-point를 검출하며, 이 시점에서 두 integration을 동시에 종료한다. 따라서 output은 고정된 integration time이 아니라 두 slope의 상대적인 크기에 의해 결정되며, 모든 binary-weighted current source가 매 conversion마다 사용된다. 이러한 internal division 구조를 통해 local current-source mismatch를 평균화할 수 있으며, 특히 code에 따른 output deviation이 한쪽 방향으로 누적되지 않고 symmetric-centroid 형태로 분포하게 된다.

Mismatch 분석 결과, XP-DAC는 기존 QM-DAC보다 output variation에 대한 민감도가 낮으며, worst-case code에서 약 2배 작은 deviation을 갖는다. 논문에서는 Pelgrom's law를 기준으로 이 mismatch 감소를 transistor 면적 감소로 연결할 경우 current-source transistor size를 약 4배 줄일 수 있으며, 기존 QM-DAC에서 필요했던 별도의 full-scale current source도 제거할 수 있기 때문에 결과적으로 DAC area를 최대 8배까지 줄일 수 있다고 설명한다. 또한 XP-DAC의 output은 V_H와 V_L 사이의 internal division으로 결정되기 때문에 output이 reference 범위를 벗어나지 않아 monotonicity 확보에도 유리하다.

한편 cross-point를 검출하는 comparator의 propagation delay로 인해 발생하는 error와 source amplifier의 offset을 제거하기 위해 improved correlated double integration (I-CDI) 기법을 적용하였다. 기존 CDI는 comparator의 cross-point detection error를 sampling하여 보정할 수 있지만 output amplifier의 offset까지 제거하지는 못한다. 제안된 I-CDI는 기존 error-sampling 과정에서 사용되는 feedback capacitor를 amplifier offset auto-zero에도 재사용하며, 이를 위해 MUX 하나만 추가한다. 따라서 별도의 큰 calibration circuit 없이 comparator-induced error와 amplifier offset을 동시에 줄일 수 있다.

제안된 source-driver IC는 180-nm CMOS 공정으로 제작되었으며, 5-bit global R-DAC와 5-bit XP-DAC를 결합하여 10-bit grayscale을 구현하였다. 30 kΩ과 30 pF의 panel load 조건에서 최소 1-H time 1.5 μs로 10-bit settling accuracy를 만족하였다. I-CDI를 적용하면 16개 channel의 output dispersion이 기존 약 27 mV에서 3 mV로 감소하였으며, 전체 grayscale에서 최대 DVO 4.3 mV를 달성하였다. 또한 측정된 maximum DNL과 INL은 각각 0.36 LSB와 1.02 LSB였으며, 최종 channel size는 107.6 × 11.3 μm²를 달성하였다. 이를 통해 빠른 1-H scan time을 유지하면서 작은 channel area와 낮은 DVO를 동시에 확보하였다.



[그림 3] #18-5에서 제안된 cross-point (XP) DAC based internal division 메커니즘

저자정보



안재웅 박사과정 대학원생

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

2026 VLSI Review

단국대학교 파운드리공학과 석사과정 임재영

Session 4 Advanced High-Frequency Synthesizers

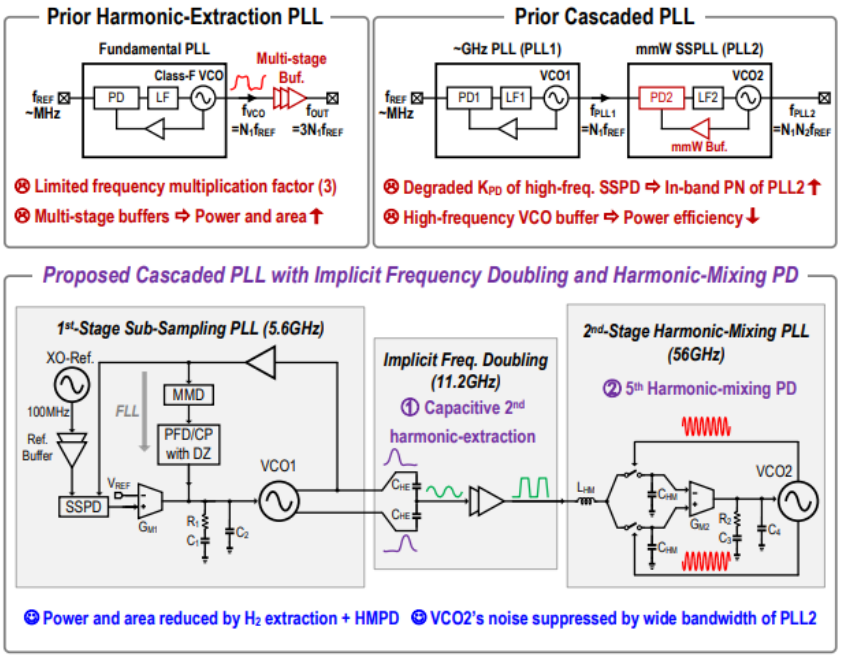
이번 VLSI 2026 Session 4에서는 초고속 유·무선 통신 시스템에서 요구되는 저지터 고주파 클록을 생성하기 위한 다양한 frequency synthesizer 기술들이 소개되었다. 특히 mmWave 대역에서 발생하는 VCO의 phase noise와 주파수 합성 과정에서의 전력 및 면적 증가를 개선하기 위한 새로운 VCO 및 PLL 구조들이 발표되었다. 본 리뷰에서는 이 중 harmonic extraction과 harmonic-mixing phase detection을 결합하여 낮은 jitter와 높은 전력 효율을 달성한 #4-3 논문을 중심으로 그 구조와 핵심 아이디어를 살펴본다.

#4-3 본 논문은 중국과학원 마이크로전자연구소 연구진을 중심으로 발표한 연구로, 낮은 jitter와 전력 소모를 동시에 달성하기 위한 2nd harmonic extraction과 5th harmonic-mixing phase detector(HMPD)를 적용한 54~60 GHz cascaded PLL 구조를 제안하였다. 100 Gb/s 이상의 초고속 wireline transceiver에서는 낮은 jitter를 갖는 mmWave clock이 요구되지만, 동작 주파수가 높아질수록 passive component의 제한된 quality factor로 인해 VCO의 phase noise가 증가하여 전체 PLL의 jitter 성능을 제한한다. 기존에는 낮은 주파수 VCO의 harmonic을 추출하거나 두 개의 PLL을 연결하는 cascaded 구조를 사용하였으나, harmonic-extraction 방식은 frequency multiplication 범위가 제한되고 여러 단계의 buffer로 인해 전력과 면적이 증가한다. 또한 기존 cascaded PLL의 2nd-stage에서는 high-frequency isolation buffer와 낮은 phase-detector gain으로 인해 전력 소모와 in-band phase noise가 증가하는 한계가 있다. 이를 해결하기 위해 제안된 cascaded PLL은 5.6 GHz의 1st-stage sub-sampling PLL과 56 GHz의 2nd-stage harmonic-mixing PLL로 구성된다. 먼저 1st-stage의 Class-F_2 VCO에서 발생하는 2nd harmonic을 직접 추출하여 11.2 GHz 신호를 생성한다. 이러한 implicit frequency doubling은 별도의 active frequency doubler를 사용하지 않고 VCO 내부에서 생성되는 harmonic을 이용함으로써 추가적인 전력 및 면적 소모를 줄인다. 또한 추출된 2nd-harmonic 신호는 process 및 temperature 변화에서도 충분한 amplitude를 유지하여 2nd-stage PLL의 고주파 reference로 사용된다. 2nd-stage PLL에서는 기존 high-frequency sub-sampling 방식의 한계를 개선하기 위해 5th harmonic-mixing phase detector(HMPD)를 사용한다. 11.2 GHz 신호를 inverter-based buffer를 통해 harmonic 성분이 포함된 quasi-square waveform으로 변환하고, 이 신호의 5th harmonic과 56 GHz VCO 출력을 직접 mixing하여 phase error를 검출한다. 이를 통해 별도의 mmWave VCO isolation buffer와 복잡한 sampling clock을 제

거하고, 기존 voltage-sampling 방식에서 sampling capacitance에 의해 phase-detector gain이 제한되는 문제를 완화하였다. 또한 작은 series-peaking inductor를 이용하여 5th-harmonic 성분을 강화함으로써 phase-detector gain을 향상시켰으며, harmonic-generation buffer를 포함한 HMPD는 1.3 mW의 낮은 전력을 소비한다. 결과적으로 5.6 GHz의 1st-stage PLL 출력은 2nd harmonic extraction과 5th harmonic mixing을 통해 총 10배의 frequency multiplication을 거쳐 56 GHz까지 확장된다.

제안된 cascaded 구조는 2nd-stage PLL의 넓은 loop bandwidth를 이용하여 56 GHz VCO의 phase noise를 효과적으로 억제한다. 고주파 VCO 자체는 낮은 quality factor로 인해 상대적으로 높은 phase noise를 가지지만, 약 100 MHz의 넓은 PLL2 bandwidth를 통해 저주파 영역에서는 1st-stage에서 전달된 저지터 reference를 따라가도록 한다. 이를 통해 우수한 phase noise 특성을 갖는 multi-GHz PLL의 성능을 활용하면서 mmWave 대역까지 주파수를 확장하고, 추가적인 jitter 증가를 최소화하였다.

측정 결과, 28 nm CMOS 공정으로 제작된 PLL은 0.23 mm²의 core area에서 54~60 GHz의 tuning range를 가지며, 56 GHz에서 53.8 fs의 RMS jitter와 14.9 mW의 소비전력을 달성하였다. 또한 측정된 3개의 chip 모두 전체 tuning range에서 60 fs 이하의 RMS jitter를 유지하였다. 결과적으로 본 연구는 implicit 2nd harmonic extraction과 5th harmonic-mixing phase detection을 cascaded PLL에 결합하여 고주파 buffer에 의한 전력 소모를 줄이고, 낮은 jitter를 유지하면서 56 GHz까지 효율적인 frequency multiplication을 구현하였다. 이를 통해 초고속 wireline transceiver에서 요구되는 저전력·저지터 mmWave clock generation을 위한 새로운 frequency synthesizer 설계 방향을 제시하였다.



[그림 1] 기존 고주파 PLL 구조와 제안된 Cascaded PLL의 전체 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

2026 VLSI Review

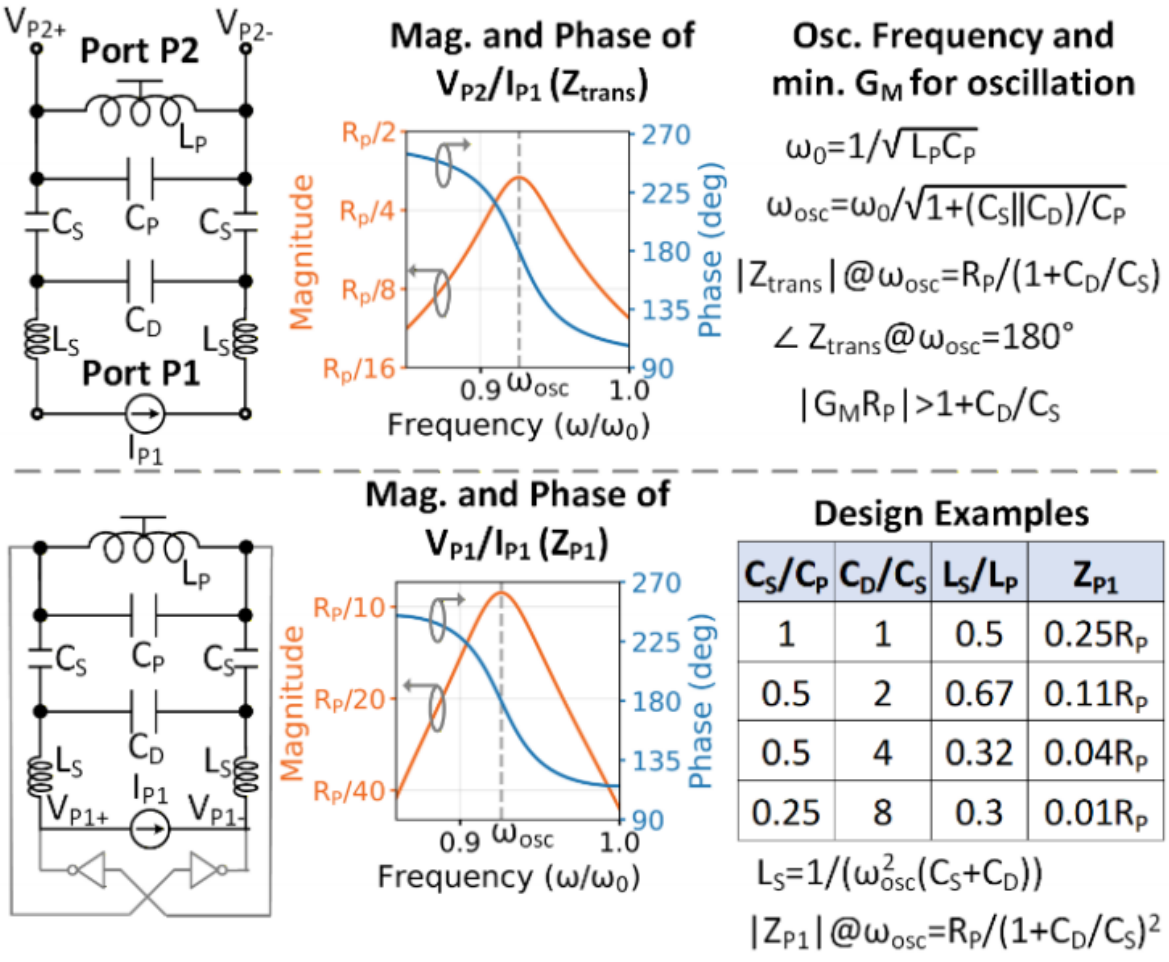
단국대학교 파운드리공학과 석사과정 임재영

Session 16 Low-Power Oscillators and Digital Techniques

이번 VLSI 2026 Session 16에서는 저전력 클록 생성과 디지털 회로의 전력 및 면적 효율을 향상시키기 위한 다양한 연구들이 소개되었다. 특히 낮은 위상 잡음을 달성하기 위한 고성능 발진기와 저전력 디지털 회로를 중심으로, 제한된 전력과 면적에서 회로 성능을 향상시키기 위한 새로운 구조들이 발표되었다. 본 리뷰에서는 새로운 공진 네트워크를 통해 낮은 위상 잡음과 높은 전력 효율을 달성한 #16-1 CSA-PR VCO를 중심으로 그 구조와 핵심 아이디어를 살펴본다.

#16-1 본 논문은 스위스 ETH Zurich에서 발표한 연구로, 높은 전력 및 면적 효율을 유지하면서 낮은 위상 잡음(Phase Noise, PN)을 달성하기 위한 Capacitively-Coupled Series-Assisted Parallel-Resonance(CSA-PR) VCO 구조를 제안하였다. 고속 데이터 변환기와 무선·유선 통신 시스템에서는 저지터 클록 생성을 위해 VCO의 위상 잡음을 줄이는 것이 중요하다. 기존에는 여러 VCO 코어를 병렬로 연결하거나 series-resonance 구조를 이용하여 유효 탱크 임피던스를 낮추고 공진기에 더 많은 전력을 공급함으로써 위상 잡음을 개선하였다. 그러나 multi-core 구조는 큰 면적과 코어 간 mismatch 문제를 가지며, 기존 series-resonance 구조는 높은 소비전력을 요구하는 한계가 있다. 이를 해결하기 위해 제안된 CSA-PR VCO는 parallel LC resonator에 C_s, C_D, L_s로 구성된 보조 공진 네트워크를 capacitive coupling하는 구조를 사용한다. C_D/C_s 비율을 조절하면 active circuit에서 바라보는 유효 탱크 임피던스를 변화시킬 수 있으며, 이를 통해 제한된 voltage swing에서도 공진기에 더 많은 전력을 공급하여 위상 잡음을 감소시킬 수 있다. 또한 capacitor ratio를 통해 탱크 임피던스를 조절할 수 있어 요구되는 소비전력과 위상 잡음에 따라 동작점을 유연하게 설정할 수 있다. 제안된 공진 네트워크는 passive voltage gain과 waveform shaping을 통해 위상 잡음을 추가적으로 개선한다. LC network에서 발생하는 passive voltage gain을 이용하여 cross-coupled inverter의 gate에 더 큰 signal swing을 전달하며, 보조 인덕터 L_S는 C_s와 C_D에 의해 발생하는 capacitive reactance를 상쇄한다. 이를 통해 gate와 drain 신호 사이의 위상 차이를 줄이고 rising 및 falling transition의 대칭성을 향상시켜 transistor의 flicker noise가 위상 잡음으로 변환되는 flicker-noise upconversion을 억제하였다. 측정결과, 22 nm FDSOI CMOS 공정으로 제작된 VCO는 0.14mm^2의 core area에서 5.7–6.8 GHz의 주파수 tuning range와 10.1-11.8mW의 소비전력을 달성하였다. 또한 10MHz offset에서 최대 -148.8dBc/Hz의 위상 잡음과 193.5dB의 FoM을 기록하였다. C_D 조절을 통해 약 3.3dB의 위상 잡음 조절 범위를 확보하면서도 FoM 변화는 1dB 이내로

유지하여, 소비전력과 위상 잡음 사이의 유연한 조절 가능성을 측정을 통해 확인하였다. 결과적으로 본 연구는 capacitive coupling을 이용한 임피던스 조절과 passive voltage gain, waveform shaping을 하나의 공진 네트워크에서 구현함으로써 낮은 위상 잡음과 높은 전력 및 면적 효율을 동시에 확보하였다. 이를 통해 저지터 PLL과 고성능 클록 생성 회로에서 요구되는 소비전력과 위상 잡음을 유연하게 최적화할 수 있는 새로운 VCO 설계 방향을 제시하였다.



[그림 1] 제안된 CSA-PR VCO의 공진 네트워크 및 동작 원리

저자정보



임재영 석사과정 대학원생

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

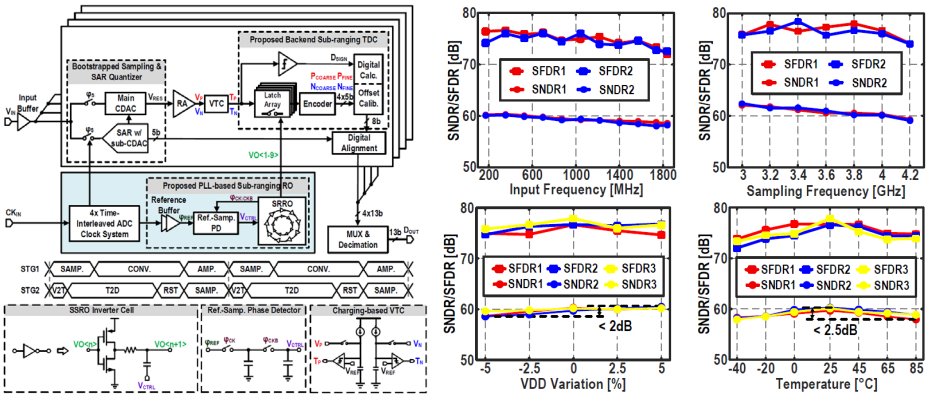
VLSI 2026 Review

울산과학기술원 전기및전자공학부 석박통합과정 홍기업

Session 17: High Speed ADCs

올해 2026 IEEE VLSI Symposium on Technology and Circuits의 Session C17은 High-Speed ADCs라는 주제로 총 5편의 논문이 발표되었다. 본 세션에서는 time-interleaved pipelined ADC, noise-shaping pipelined ADC, TI-SAR ADC, pipelined-SAR ADC 등 다양한 고속 ADC 구조가 다루어졌으며 수 GS/s 수준의 높은 sampling rate를 확보하면서도 해상도와 전력 효율을 유지하기 위한 회로 기법들이 제안되었다. 이를 통해 최근 high-speed ADC 연구가 단순히 sampling rate를 높이는 것에 그치지 않고, time-domain processing과 parallelization을 적극 활용하여 conversion 및 residue-processing latency를 줄이는 동시에 background calibration과 PVT-robust 회로 기법을 통해 고속 동작에서의 정확도 저하를 최소화하고, residue amplification 및 reference 회로의 전력 부담까지 줄이는 방향으로 발전하고 있음을 확인할 수 있다.

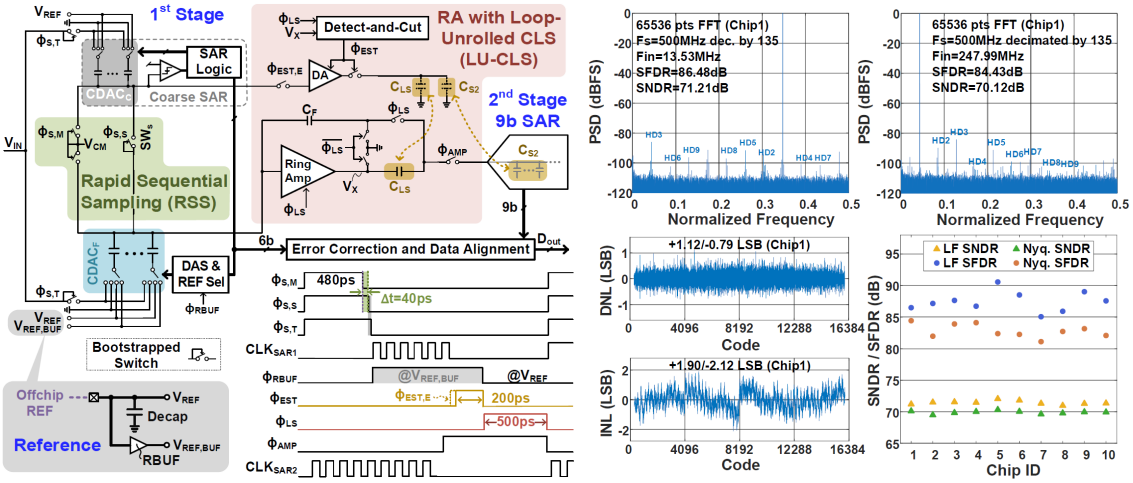
#17.1 University of Macau와 UM Hetao IC Research Institute에서 발표한 논문으로 PLL-based sub-ranging RO-TDC를 backend로 사용하는 12-bit 4-GS/s time-interleaved pipelined ADC를 제안하였다. 기존 RO-based TDC는 높은 hardware efficiency를 가지지만 oscillator frequency가 PVT variation에 직접 영향을 받아 time resolution이 변화하고 resolution과 dynamic range 사이에 trade-off가 존재한다. 본 연구에서는 2-GHz coarse RO와 12-GHz fine RO를 결합한 sub-ranging 구조를 사용하여 coarse RO로 넓은 input range를 확보하고 fine RO를 통해 4.63-ps의 높은 time resolution을 구현하였다. 또한 두 RO를 reference-sampling PLL 내부에서 동작시켜 oscillator frequency를 ADC reference clock에 지속적으로 locking함으로써 PVT 변화에 따른 TDC resolution 변화를 억제하였다.



[그림1] System architecture of proposed ADC (top) and measured SNDR and SFDR performance

전체 ADC는 4-channel time-interleaved pipelined-SAR 구조로 구성된다. 1st-stage의 5-bit ping-pong SAR quantizer가 입력을 우선 양자화한 뒤 CDAC에서 voltage residue를 생성하고, 이를 floating-inverter amplifier로 증폭한 후 VTC를 통해 time-domain signal로 변환하여 8-bit sub-ranging TDC에서 나머지 양자화를 수행한다. 또한 coarse/fine RO 사이의 phase offset 및 cross-boundary error는 background digital calibration을 통해 보정하여 별도의 calibration cycle 없이 고속 동작을 유지한다. 결과적으로 본 연구는 PLL locking과 sub-ranging time-domain quantization을 결합하여 고속 ADC의 PVT robustness와 resolution을 동시에 확보한 구조라고 볼 수 있다.

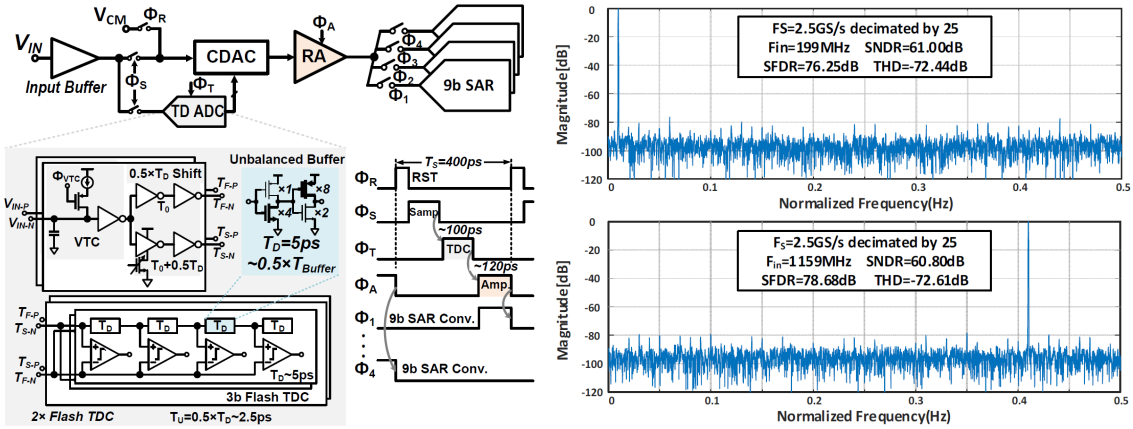
#17.4 University of Macau에서 발표한 논문으로 loop-unrolled correlated level shifting (LU-CLS)과 rapid sequential sampling (RSS)을 적용한 14-bit 500-MS/s pipelined-SAR ADC를 제안하였다. Pipelined-SAR ADC는 높은 energy efficiency를 갖지만 high-speed 동작에서는 1st-stage SAR conversion에 많은 시간이 소요되어 residue amplifier의 amplification 및 settling time이 크게 제한된다. 특히 기존 correlated level shifting (CLS)은 residue를 먼저 estimation하는 EST phase와 실제 level shifting을 수행하는 LS phase가 순차적으로 동작하기 때문에 amplification time이 추가로 필요하다. 본 연구에서는 coarse SAR의 residue를 미리 활용하는 별도의 estimation path를 구성하여 EST를 기존 amplification interval에서 분리한 LU-CLS 구조를 제안하였다. 이를 통해 CDACF settling과 EST를 병렬로 수행할 수 있으며, 충분한 EST time을 확보하면서도 LS time을 감소시키지 않아 기존 CLS 대비 8 dB 이상의 SDR 향상을 달성한다. 또한 detect-and-cut loop를 적용하여 estimation gain error를 보정함으로써 별도의 background gain calibration 없이 높은 residue amplification accuracy를 확보한다.



[그림2] Proposed Pipelined-SAR ADC architecture and LU-CLS/RSS operation [좌] Measured dynamic performance at 500 MS/s [우]

한편 coarse SAR의 CDACC와 fine CDACF 사이에는 sampling timing과 bandwidth mismatch에 의해 sampling error가 발생하며, 이는 high-frequency input에서 residue swing을 증가시켜 EST accuracy를 저하시킬 수 있다. 이를 해결하기 위해 RSS를 적용하여 sampling 직후 두 CDAC 사이에서 rapid charge sharing을 수행함으로써 coarse/fine sampling mismatch를 10배 이상 감소시켰다. 또한 ring amplifier에는 interstage negative capacitor (I-NC)를 적용하여 non-dominant pole을 높은 주파수로 이동시키고 settling speed와 phase margin을 개선하였으며, worst PVT corner에서 settling time을 300 ps 이상 단축하였다. Reference 회로에는 dynamic-buffer-assisted reference (DBAR)를 적용하여 quantization과 CDACF settling 동안에만 reference buffer를 동작시킴으로써 reference accuracy를 유지하면서 buffer power와 decap size를 감소시켰다. 결과적으로 본 연구는 LU-CLS를 통해 residue amplification의 time overhead를 줄이고, RSS·I-NC·DBAR를 함께 적용하여 sampling mismatch, amplifier settling 및 reference 부담을 개선함으로써 high-resolution과 high-speed 동작을 동시에 구현한 pipelined-SAR ADC라고 정리할 수 있다.

#17.5 University of Macau와 UM Hetao IC Research Institute에서 발표한 논문으로 dual-zero-compensation ring amplifier를 적용한 12-bit 2.5-GS/s pipelined-SAR ADC를 제안하였다. High-speed pipelined ADC의 residue amplifier에서는 높은 gain과 large output swing을 유지하면서 bandwidth를 증가시키는 것이 어렵다. 기존 deadzone 기반 ring amplifier는 높은 gain과 안정성을 확보할 수 있지만 bandwidth가 제한되며, critically damped 구조는 bandwidth를 높이는 대신 gain과 linearity가 저하되는 trade-off가 존재한다. 본 연구에서는 이러한 문제를 해결하기 위해 두 개의 feedforward path를 이용하여 left-half-plane zero를 형성하는 dual-zero-compensation 구조를 제안하였다.



[그림3] Two-stage Pipelined ADC architecture and time domain sub-ADC [좌]
Measured dynamic performance at 2.5GS/s [우]

제안된 방식은 pole 위치 자체를 크게 변경하기보다 zero를 이용해 phase margin을 보상하여 ring amplifier의 높은 gain과 output swing을 유지하면서 bandwidth를 향상시킨다. 또한 pole과 zero가 PVT 변화에 따라 함께 이동하도록 설계하여 compensation robustness를 확보하였으며, 제안된 amplifier는 약 120 ps의 settling time에서 8× amplification, 40 dB DC gain 및 55 GHz UGB를 달성하였다. 전체 ADC는 two-stage pipelined-SAR 구조로 구성된다. 결과적으로 본 연구는 zero compensation을 통해 ring amplifier의 gain–bandwidth trade-off를 완화하고 이를 고속 pipelined ADC에 적용한 연구라고 정리할 수 있다.

저자정보



홍기업 석박사통합과정 대학원생

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

2026 VLSI Review

KAIST 전기 및 전자공학부 박사과정 박수언

Session 28 High-Resolution Data Converters

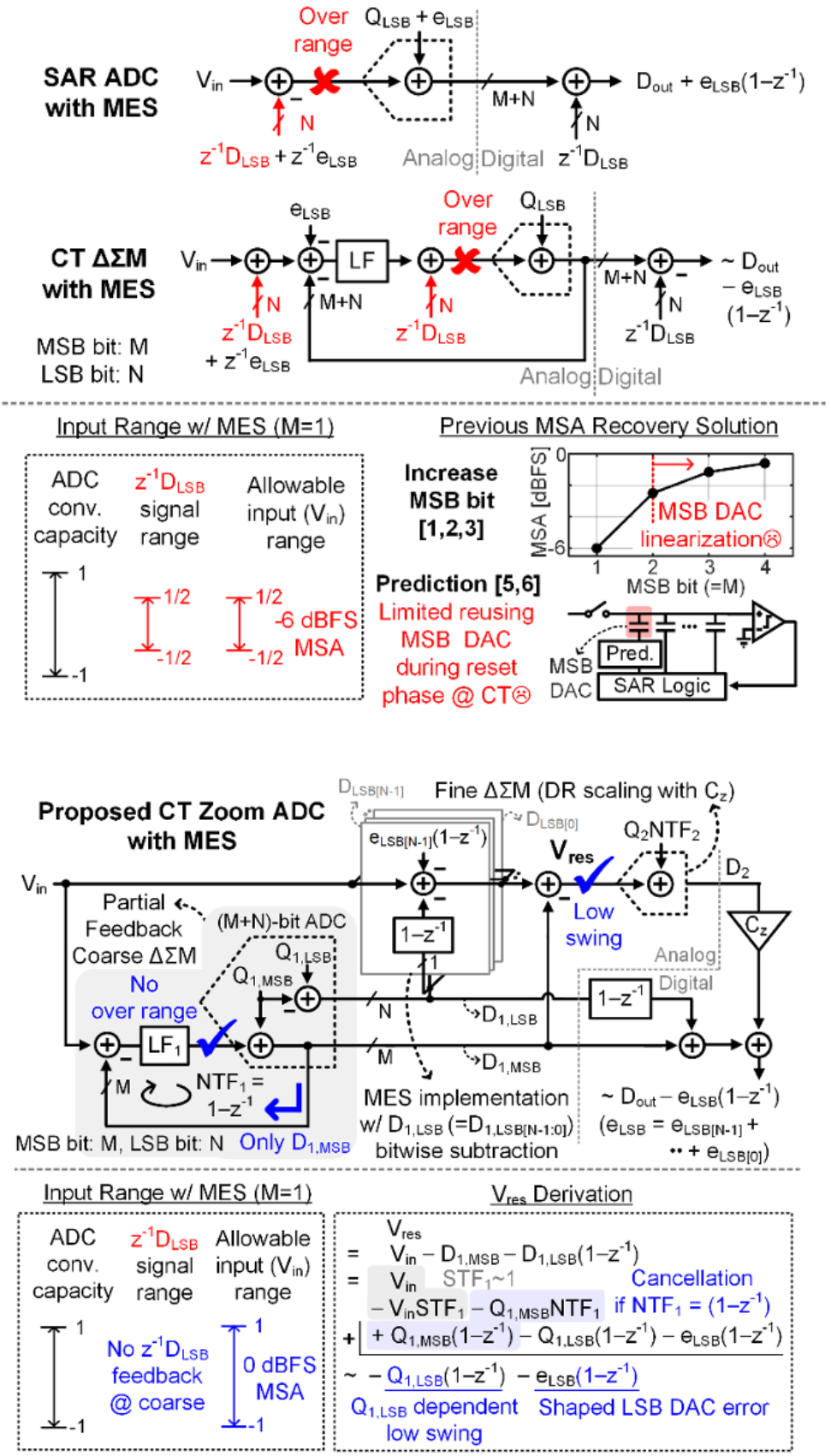
Session 28은 고해상도 Data Converter의 성능 한계를 극복하기 위한 회로 기술들이 제시되었다. 본 글에서는 연세대학교에서 발표한 CT Zoom ADC (C28.2)에 주목하여, 기존 Mismatch Error Shaping(MES) 기법의 한계를 극복한 아키텍처를 다뤄보고자 한다.

#28-2 논문은 연속 시간(Continuous-Time, CT) Zoom ADC에서 DAC mismatch 요소를 효율적으로 제어하면서도 입력 신호의 Full-Scale 범위를 유지하는 Mismatch Error Shaping(MES) 기술을 제안한다.

종래의 MES 기법은 DAC 오차를 추출하기 위해 ADC 입력단에 지연된 LSB 성분을 가감하는 방식을 사용하여, 입력 신호의 최대 허용 범위(Maximum Stable Amplitude, MSA)가 줄어드는 치명적인 구조적 단점이 존재했다. 제안한 논문에서는 Partial-Feedback Coarse ΔΣM 구조를 도입하여, Coarse 루프 내에는 LSB 성분만을 피드백하고 Fine 단계로 Residue Voltage(V_{res})를 넘겨줌으로써 MSA 손실 없이 0 dBFS의 Full-Scale 입력 범위를 보존하였다.

또한, 구현 면적 및 오버헤드를 줄이기 위해 Bitwise Subtraction 기법 기반의 Tri-level RDAC을 제시하였다. Bit-Subtraction을 통해 기존 비트당 2개의 Bi-level DAC 스위치를 1개의 Tri-level DAC 스위치로 대체하여 하드웨어 복잡도를 대폭 개선하였으며, 이때 발생하는 DAC Gain error를 1차 형태로 자연스럽게 Shaping하도록 설계하였다. 더불어 Tri-level RDAC 사용 시 발생할 수 있는 Signal-dependent reference modulation 문제를 해결하기 위해 Dummy DAC을 배치하여 3차 고주파 왜곡 성분(HD3)을 15.5 dB 크게 개선하였다.

65nm CMOS 공정으로 제작된 제안 IC는 1MHz 대역폭(BW)에서 1.7mW만을 소비하며, 104.8dB SFDR, 90.1dB SNDR, 92.5dB DR의 우수한 성능을 달성하였다. 이는 500kHz 이상의 BW를 갖는 CT ADC 중 최고 수준인 180.2dB의 FoM_{DR}을 달성한 결과로, 전력 효율과 선형성을 동시에 극대화하였다.



[그림 2] 기존 MES over-ranging 문제 (top)와 제안하는 MES – CT zoom ADC 구조 (bottom).

저자정보



박수언 박사과정 대학원생

- 소속 : KAIST
- 연구분야 : Power Management IC 설계
- 이메일 : tndjs0905@gmail.com
- 홈페이지 : https://icdesignlab.net/

2026 VLSI Review

한양대학교 박사후연구원 송충석

Session 20 Next-Generation Optical Transceivers

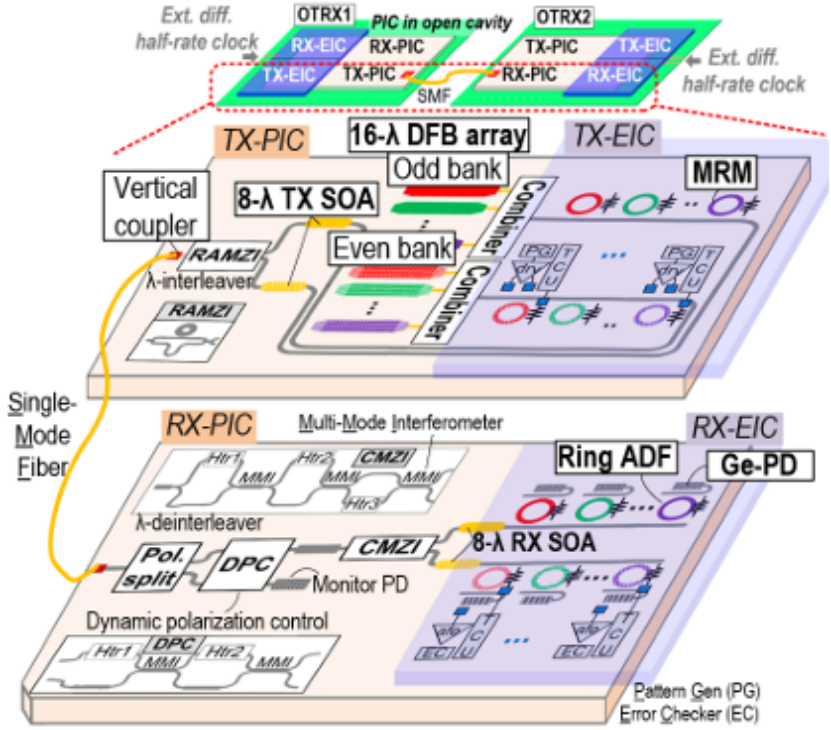
이번 2026 VLSI의 Session 20은 Next-Generation Optical Transceivers라는 주제로 총 4편의 논문이 발표되었다. 본 리뷰에서는 CMOS 전자 IC(EIC)와 실리콘 포토닉스 IC(PIC)를 3D 적층, 하이브리드 본딩으로 이종집적하여 회로-소자 공동설계를 통해 광 인터커넥트의 채널당 데이터율과 비트당 에너지 효율을 동시에 개선하고자 한 3편의 논문을 리뷰하고자 한다.

#20-1 논문에서는 open-cavity package 기반 800 Gbps/fiber 실리콘 포토닉 마이크로링 dense wavelength division multiplexing (DWDM) 트랜시버를 발표하였다. 고성능 컴퓨팅(high performance computing) 스케일업 패브릭의 대역폭 요구를 충족하기 위해 실리콘 포토닉스 기반 co-packaged optics (CPO)가 주목받고 있으며 마이크로링 공진기 기반 광 송수신기는 WDM을 통해 파이버당 대역폭을 동일 면적 및 동일 에너지로 확장할 수 있다는 장점이 있다. 그러나 기존 마이크로링 공진기 기반 DWDM 트랜시버는 몇 가지 한계를 가지는데, (1) 인터포저를 매개로 EIC-PIC를 연결하는 패키징은 고속 경로에 큰 기생 성분을 유발해 대역폭을 제안하고, (2) 마이크로링 공진 구동용 단일종단 스택형 드라이버는 높은 출력 임피던스로 인해 링 변조기 특유의 큰 스윙을 고속으로 인가하기 어려우며, (3) DC 커플링 시 저주파 베이스라인 완더(BLW)로 eye opening 현상이 저하되며, (4) 또한 채널 수 증가에 따라 인접 파장 간 크로스토크와 각 링을 레이저 파장에 개별 록킹시키는 열 제어의 복잡도가 커지게 된다.

이를 해결하기 위해 본 논문은 22nm CMOS EIC와 Intel Fab11x O-band PIC를 인터포저 없이 식각 캐비티 기판에 3D 플립칩으로 직접 실장해 기생 성분을 최소화하고 차동 인버터 기반 TX 드라이버로 고 임피던스 문제없이 전체 스윙을 인가하며, AC/DC 피드포워드-피드백 구조로 BLW를 보상한다.

칩 측정 결과는 200 GHz 간격의 16개 파장을 O-band에서 각 50Gbps NRZ로 동시 변조해 파이버당 총 800Gbps를 달성했으며 짝수, 홀수 8 레인씩 총 16레인 모두에서 BER<1e-9를 만족하였다. 50Gbps PRBS-15 아이 다이어그램에서 5dB 이상의 소광비(extinction ratio)가 측정됐고, BLW 보정 회로 적용 시 8Gbps PRBS-31 조건에서 eye opening 현상이 0.35dB 개선되었다. 에너지 효율은 22nm TX EIC가 0.92pJ/b, RX EIC가

3.1pJ/b, PIC가 1.69pJ/b로 측정되었다. 논문에서는 고집적 O-band PIC, 고대역폭 EIC 회로 기법, 인퍼토저 없는 저기생 패키지의 결합이 파이버당 800Gbps급 데이터율을 에너지 효율적으로 달성한 핵심 요인이라고 언급하였다.



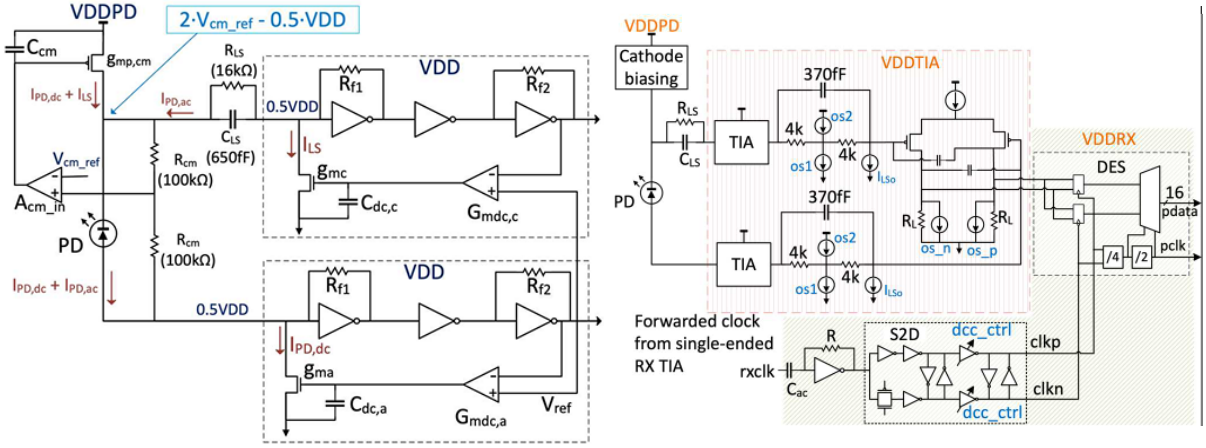
[그림 1] 논문에서 제안한 DWDM 트랜시버

#20-2 논문에서는 3D 적층 실리콘 포토닉스 플랫폼에서 차동 TIA를 이용해 -17.3dBm 수신 감도를 달성한 32Gb/s 광 수신기를 발표하였다. 데이터센터 direct detection 광링크에서 TIA 감도는 링크 budget과 에너지 효율에 직결되나 동기검출(coherent detection)과 달리 direct detection은 포토다이오드(PD)가 본질적으로 단일종단 소자이기 때문에 대부분 단일종단 TIA를 사용해 왔다. PD 캐소드 전류까지 활용하면 출력 신호가 2배가 되고 SNR이 $\sqrt{2}$배 개선되지만 고속 동작을 위한 PD의 높은 역바이어스 전압 때문에 캐소드 단자를 활용하기 어렵다는 문제가 있었다. 기존 방식은 애노드 노드에서 단일종단-차동 변환을 수행해 오히려 SNR이 저하되거나, 진성 차동 TIA에서 캐소드 노드를 AC 커플링하여 바이어스 불일치를 해결하였었는데, 이는 저주파 컷오프를 심하게 제한하며 커플링 커패시턴스를 키우면 기생 성분이 차동 구조의 이득을 상쇄한다는 한계가 있었다. 스택형 차동 TIA는 이를 해결가능하지만 여전히 2개의 레귤레이터와 캐소드측 TIA용 deep n-well이 필요하다는 문제점이 있었다.

본 논문은 PD 캐소드 바이어싱 회로가 전 주파수 대역에서 캐소드측 TIA 입력 임피던스보다 높은 임피던스를 갖도록 설계하고 R_{LS}, C_{LS} 병렬망 기반 광대역 레벨시프터로 캐

소드 전압을 TIA가 받아들일 수 있는 레벨로 낮춤으로써 단 650fF의 작은 커패시터만으로 낮은 저주파 컷오프를 구현했다. 또한 두 TIA 경로간 고주파 손실 차이를 캐소드측 Rf를 5% 높여 보상하고 DC 캔슬레이션 루프로 레벨시프트 전류를 자동 설정해 별도 레귤레이터나 deep n-well 없이 완전 차동 TIA를 구현하였다.

7nm FinFET CMOS EIC를 65nm 실리콘 포토닉스 PIC위에 Cu-Cu 하이브리드 본딩으로 3D 적층한 칩으로 측정한 결과, BER 1e-12 기준 수신 감도는 28Gbps에서 -18.9dBm, 32Gb/s에서 -17.3dBm OMA를 기록하였다. 수신기는 1.05V 공급전압에서 14.45mA, PD용 2V 공급에서 0.14mA를 소모해 32Gbps에서 0.484pJ/b의 에너지 효율을 달성했으며 수신기 면적은 3700um^2으로 측정되었다.



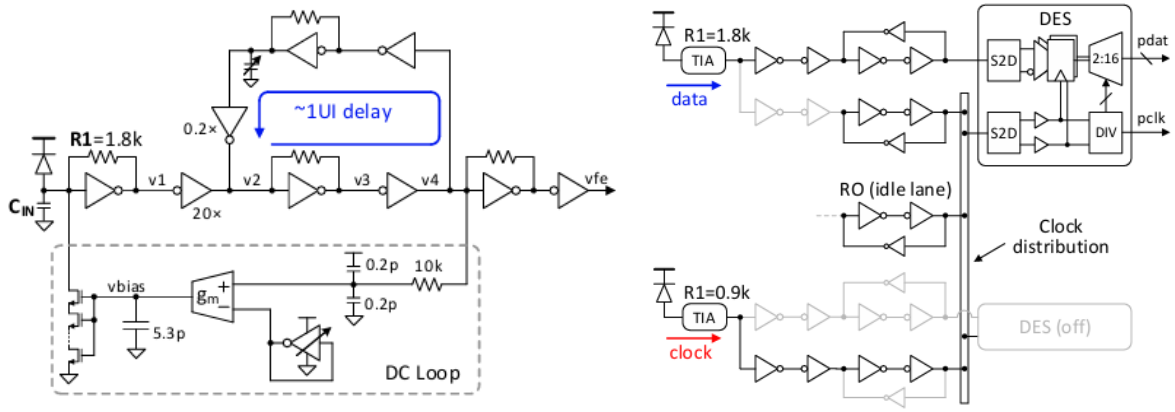
[그림 2] 논문에서 제안한 차동 TIA 회로(좌), 전체 수신기 회로(우)

#20-4 논문에서는 self-timed decision feedback equalization (STDFE)를 이용해 -18.5dBm OMA 감도와 0.416pJ/b의 에너지 효율을 동시에 달성한 32Gb/s 3D 적층 광 NRZ 수신기를 발표하였다. 광 링크의 에너지 효율 설계에서 RX OMA 감도는 레이저 파워를 직접 좌우하는 핵심 지표이며 기존 연구들은 이를 개선하기 위해 RX 입력단 기생 커패시턴스 축소, 트랜스임피던스 저항 증대를 통한 열잡음, 저항 잡음 감소 및 게인 향상, 샘플러 입력 스윙 확대, 데이터-클럭 경로 간 딜레이 매칭 등의 기법을 개별적으로 발전시켜 왔다. 그러나 트랜스임피던스 저항을 키워 노이즈와 게인을 개선할수록 대역폭이 줄어들고 이를 보상하기 위해서 별도의 타이밍 회로와 하드웨어 오버헤드가 추가되는 트레이드오프가 존재했다.

본 논문에서는 클럭포워딩 구조에서 딜레이 매칭을 위해 TIA 출력을 rail-to-rail로 증폭해야 한다는 기존 설계에서 착안하여 이 rail-to-rail 출력을 잡음이 억제된 신호로 재활용하는 STDFE를 제안한다. 즉 별도의 클럭 없이 데이터 신호 자체에서 피드백 타이밍을

얻어 트랜스임피던스 저항을 크게 하여 얻는 저잡음, 높은 게인 이득을 유지하면서 대역폭 저하에 따른 추가 회로 부담 또한 완화시켰다.

7nm FinFET EIC와 65nm SOI PIC를 하이브리드 본딩으로 3D 적층한 칩으로 측정한 결과 32Gbps에서 BER 1e-12 미만 기준 OMA 감도는 PRBS31에서 -18.5dBm, ORBS7에서 -18.7dBm으로 STDFE를 적용하지 않은 베이스라인 RX 대비 2dB 개선되었으며 이때 추가 전류 소모는 1mA 미만으로 측정되었다. 수신기는 1.01V 공급에서 13.2mA를 소비해 32Gbps에서 0.416pJ/b의 에너지 효율을 달성했으며 면적은 3360um^2으로 측정되었다.



[그림 3] 논문에서 제안한 STDFE를 활용한 TIA 회로(좌), 간소화한 수신기 회로(우)

저자정보



송충석 박사 후 연구원

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

2026 VLSI Review

KAIST 전기및전자공학부 석박사통합과정 이동윤

Session 25 Energy Efficient Wireline Transceivers and Clocking Circuits

2026 IEEE VLSI의 Session 25는 Energy Efficient Wireline Transceivers and Clocking Circuits라는 주제로 총 4편의 논문이 발표되었다. AI 학습/추론 클러스터에서 GPU 간 scale-up 인터커넥트, 데이터센터 스위치 ASIC의 I/O, 1.6 Tb/s Ethernet용 전기/광 링크, 패키지 안의 HBM 및 chiplet 간 die-to-die 인터페이스처럼 대량의 데이터 이동을 요구하는 응용이 늘면서 산업계와 학계 모두 high-speed wireline 통신에 주목하고 있다. 이에 따라 224 Gb/s/lane급 Ethernet 전기 링크(OIF CEI-224G, IEEE 802.3dj), HBM PHY와 UCIe 같은 short-reach 병렬 인터페이스, SRIS 환경의 PCIe 6.x 등 서로 다른 reach와 표준을 겨냥한 연구가 함께 발표되고 있다. 기존 wireline TRX는 CTLE, FFE, DFE 등의 equalization으로 채널 손실을 보상해 reach를 늘려 왔고, PAM-4 같은 multi-level signaling으로 심볼당 비트 수를, multi-lane 구성으로 총 데이터율을 높여 왔다. 100 Gb/s/lane을 넘는 long-reach 링크에서는 수십 dB의 손실을 아날로그 등화만으로 감당하기 어려워 ADC-DSP 기반 RX가 주류가 되었다. 25-1은 3nm FinFET에서 4-lane 212.5 Gb/s PAM-4 DAC 기반 TX와 ADC-DSP RX로 54.7 dB reach와 2.05 pJ/bit를 보고했고, 25-4는 같은 3nm에서 224 Gb/s PAM-4 ADC-DSP RX로 38 dB 채널에서 1.39 pJ/bit를 달성했다. 두 편이 200G급 ADC-DSP의 현재 수준을 보여준다면, 나머지 두 편은 구조적으로 다른 접근을 택했다. 본 기사에서는 4개의 transmission line을 하나의 그룹으로 묶어 전류 영역에서 PAM-4 심볼을 전송하는 25-2와, phase interpolator 없이 주파수 오프셋을 직접 추적하고 SSC 프로파일을 복제해 상쇄하는 multi-lane용 디지털 CDR을 제안한 25-3에 집중하여 서술한다.

#25-2 Voltage based 통신에서는 lane의 기생 저항과 캐패시턴스가 만드는 RC 시정수가 대역폭을 제한하고, 수신단이 판정할 수 있는 전압까지 lane 전체를 충방전해야 하므로 데이터율이 오를수록 이 병목이 커진다. 수신단 입력 임피던스를 낮게 잡아 lane 전압을 고정하고 전류만 읽으면 충방전 자체를 피할 수 있다는 발상은 여기서 출발했으며, 캐패시턴스가 큰 SRAM 비트라인을 저임피던스 current sense amplifier로 읽어 속도를 높인 연구 [1]와 같은 원리를 온칩 global interconnect에 적용한 연구 [2]가 그 뿌리에 해당한다. 한편 HBM이나 die-to-die 인터페이스처럼 수천 개의 lane이 동시에 스위칭하는

wide-parallel I/O가 늘면서, lane당 충방전 에너지와 데이터 의존 스위칭 전류가 만드는 supply noise (SSN, simultaneous switching noise)가 시스템 스케일링을 막는 요인으로 새롭게 부상했다. 본 논문은 이 두 문제, 즉 RC 대역폭 제한과 wide-parallel에서의 에너지/SSN을 current based 통신으로 함께 다루려는 시도이다.

논문은 기존 방식을 두 갈래로 정리한다. Voltage based SST (source-series-terminated) driver는 lane을 VDD/VSS 사이에서 스윙시키므로 CV^2에 비례하는 에너지와 큰 데이터 의존 스위칭 전류를 갖는다. CML과 같은 current-mode driver는 lane 스윙을 줄이지만 수신단의 종단 저항에서 전류를 전압으로 바꿔(I-V conversion) 판정하므로, 수신단 입력은 여전히 전압 영역이고 lane 캐패시턴스의 영향에서 벗어나지 못한다. 논문이 겨냥한 지점은 이 I-V conversion을 없애고 수신단까지 전류 영역을 유지하는 것이다.

이를 위해 논문은 4개의 TL (transmission line)을 하나의 그룹으로 묶고, 각 TL에 PAM-4의 네 전류 레벨 {−3, −1, +1, +3} 중 하나를 실어 그룹 단위로 심볼을 정의한다. TX는 큰 직렬 저항(10~20 kΩ)을 통해 VDD 또는 VSS를 lane에 연결하는 저항 구동기이고, RX는 current conveyor (CC)가 lane 전압을 공통 전압에 클램프한 채 유입 전류를 미러링하여 전류 비교기로 넘긴다. Lane 전압 스윙은 수 mV에 머물고, 판정에 기준 전압이 필요 없으며, 에너지 효율이 lane 캐패시턴스에 둔감해진다. 논문은 이 구조의 속도 한계가 lane의 RC가 아니라 CC의 전류 피드백 응답이라고 설명한다.

[그림 1]은 그룹 심볼을 선택하는 과정이다. TL당 4레벨이므로 4개 TL의 조합은 256가지인데, 논문은 세 가지 제약으로 이를 24개로 줄인다. 첫째는 DC-balance로, 네 TL 전류의 합이 항상 0이 되도록 한다. 그룹 전체가 supply에서 끌어오는 전류가 데이터와 무관하게 일정해지므로 데이터 의존 SSN이 억제된다. 둘째는 adjacent-TL contrast로, 물리적으로 인접한 TL이 같은 레벨을 갖지 못하게 하여 인접 쌍이 항상 0이 아닌 차동 신호를 갖도록 한다. 셋째는 differential distinguishability로, RX가 관측하는 것은 6개의 pairwise 대소 비교 [A>B, B>C, C>D, D>A, A>C, B>D]뿐이므로, 어느 두 TL도 같은 레벨을 갖지 않고 이 6비트 패턴이 심볼마다 고유해야 한다. 세 제약을 만족하는 집합은 네 레벨을 서로 다르게 배치한 4! = 24개 순열이며, [그림 1] 하단의 A0~A11, B0~B11이 그것이다. 데이터는 심볼 자체가 아니라 연속한 심볼 간의 전이에 실린다. 24개 심볼 중 IDLE과 END를 제외한 전이 조합이 데이터에 대응하고, IDLE조차 심볼을 바꾸는 전이로 정의되어 매 클럭마다 반드시 상태가 변하므로 별도의 클럭 채널 없이 심볼 전이에서 타이밍을 복원할 수 있다.

논문은 제안 구조를 14 nm CMOS로 구현하고 실리콘 인터포저의 RLCG 파라미터(총 R 19.2 Ω, C 1.0 pF)를 갖도록 칩 위에 배선한 3.3 mm 선로에서 검증했다. 4-TL 그룹 단위로 동작하며 TL당 22.9 Gb/s의 데이터율에서 0.06 pJ/bit의 에너지 효율과 0.006 mm^2의 면적을 달성했고, 동작 중 lane 전압 스윙은 7.07 mVpp에 머물렀다. DC-balanced 인코딩의 효과로 0.8 V 공급에서 측정한 peak SSN은 VDD/VSS 각각 3.77 mV/4.62 mV로, 같은 칩

에 구현한 2-TL NRZ SST 대비 약 9배 낮았다. 비교 대상인 최근 die-to-die/HBM급 PHY들이 0.30~3.20 pJ/bit 범위인 것과 비교하면 한 자릿수 이상의 차이이며, equalization 없이 2 UI의 latency로 동작한다는 점도 short-reach 인터페이스에서 의미가 있다. 논문은 lane 전압을 고정하는 구조 덕에 에너지 효율이 load capacitance에 둔감하고, 속도 한계가 lane의 RC가 아닌 CC의 전류 피드백 응답에 있으므로 공정 미세화에 따라 면적과 데이터율 모두 유리해질 것으로 전망하는데, 이 전망 중 14 nm는 실측이고 180~28 nm는 PDK 시뮬레이션, 7 nm 이하는 파라미터 기반 추정값이다. 수신단까지 전류 영역을 유지하는 발상을 permutation 인코딩과 결합해, RC가 지배하는 wide-parallel 채널에서 에너지와 SSN 문제를 동시에 다룰 수 있음을 보인 점이 이 논문의 기여이며, HBM PHY나 chiplet 인터커넥트처럼 lane 수가 많고 거리가 짧은 응용에서 전압 영역 SST의 대안으로 검토될 만한 방향을 제시한 논문이라 생각된다.

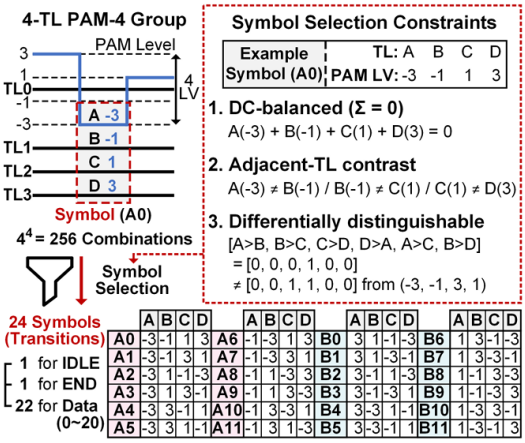


Fig. 4. Symbol selection process. Selected symbols satisfy design constraints for SSN suppression and decoding. The illustrated set corresponds to the 4-TL PAM-4 group used in this work.

[그림 1] 논문 25-2에서 제안하는 4-TL PAM-4 그룹의 심볼 선택 과정 (256개 조합에서 세 가지 제약을 거쳐 24개 순열 심볼 선택)

#25-3 AI 인프라의 multi-lane 인터커넥트는 clocking 관점에서 세 가지 요구를 동시에 받는다. 첫째, 여러 프로토콜을 수용하고 전력과 대역폭 활용을 최적화하기 위해 lane마다 서로 다른 데이터율로 동작할 수 있어야 한다. 둘째, EMI 규제를 통과하기 위해 송신 클럭에 spread-spectrum clocking (SSC)을 적용하는데, 송수신이 독립된 기준 클럭을 쓰는 SRIS (separate reference clock with independent SSC) 환경에서는 수신단이 상대 쪽 SSC의 삼각파 주파수 변조(PCIe의 경우 최대 −0.5%, 30~33 kHz)를 전부 주파수 오프셋으로 보게 된다. 셋째, 기준 클럭 주파수와 lane 수가 늘면서 클럭 분배 전력을 줄이기 위해 global LC PLL이 저주파 기준을 분배하고 각 lane의 ring oscillator (RO) PLL이 고주파 다상 클럭을 만드는 cascaded clocking이 선호된다. 본 논문은 이 세 요구를 만족하는 lane 단위 디지털 CDR을 제안한다.

기존 구조에서는 RO PLL을 정수 N으로 기준 클럭에 잠그고, 데이터와의 주파수 오프셋

f_{OS}는 PLL 되먹임 경로에 놓인 phase interpolator (PI)가 위상을 계속 회전시켜 흡수한다. 이 간접 주파수 추적에는 두 종류의 오차가 남는다. 하나는 PI가 이산 단계로 갱신되고 회전할 때마다 자신의 비선형성을 드러내기 때문에 f_{OS}에 비례해 커지는 주기적 위상 오차(cyclic error)로, PI 기반 CDR의 양자화 지터와 주파수 허용 범위 사이의 근본적 trade-off로 알려져 있다 [3], [4]. 다른 하나는 SSC가 주파수 ramp이므로 위상으로는 포물선이고, 적분기 두 개의 type-II 루프가 이를 오차 없이 따라갈 수 없어 남는 잔류 위상 오차(residual error)이다. SSC 깊이가 큰 SRIS에서는 f_{OS}의 순간값이 수천 ppm에 이르러 두 오차가 모두 커지고, 그 결과 jitter tolerance (JTOL)가 저하된다. 여기에 PI가 CDR 클럭 주파수에서 동작하는 아날로그 회로라는 전력 부담이 더해진다.

논문의 제안은 두 부분이다. 첫째, direct frequency tracking (DFT)이다. PI를 제거하고 RO PLL의 되먹임 분주기를 ΔΣ modulator가 제어하는 multi-modulus divider로 바꾼 뒤, CDR 루프 필터의 출력을 분주비의 분수부로 직접 넣는다. 데이터가 요구하는 주파수 오프셋만큼 PLL이 스스로 f_{CDR}을 옮기므로 PI 회전이 필요 없고, lane마다 정수부 D_{INT}와 분수 오프셋 D_{FOF}를 다르게 설정하면 같은 1.05 GHz 기준 클럭에서 lane별로 임의의 데이터율을 만들 수 있다. ΔΣM이 만드는 양자화 오차는 예측 가능하므로 DTC가 기준 클럭 경로에서 상쇄한다. 둘째, SSC profile replicator (SSC-PR)이다. 들어오는 SSC의 삼각파를 디지털로 재생성하여 분주비 제어에 feed-forward로 더함으로써, 루프가 ramp를 따라가는 대신 미리 빼 주어 type-II 루프의 잔류 오차를 제거한다. 삼각파의 주파수와 위상은 frequency-phase adaptor가 CDR 비례 경로 출력과 재생 삼각파의 상관이 0이 되도록 맞추고, 깊이는 depth adaptor가 삼각파 꼭짓점에서 적분 경로 출력을 비교해 맞춘다. 두 적응기 모두 루프가 이미 갖고 있는 신호만으로 SSC의 파라미터를 추정하며, 위상 정렬이 먼저 수렴하도록 대역폭에 차이를 두었다.

[그림 2]는 기존 PI 기반 구조와 제안 DFT 구조를 비교한 것이다. 왼쪽의 기존 구조에서는 PI가 되먹임 경로에서 위상을 밀고 그 코드가 0에서 2π까지 톱니 모양으로 반복되며, 아래 파형에 cyclic 오차와 residual 오차가 겹쳐 나타난다. 오른쪽의 제안 구조에서는 ΔΣM과 MMD가 그 자리를 대신하고 SSC-PR의 출력 D_{SSC}가 분주비 제어에 더해지며, 두 오차가 모두 사라진 파형이 보인다. 시뮬레이션에서 30 kHz, 5000 ppm SSC 조건의 입력-출력 위상 오차는 PI 기반 0.544 UIpp에서 DFT 적용 시 0.450 UIpp로, SSC-PR까지 적용하면 0.322 UIpp로 줄어든다.

논문은 제안 CDR을 28 nm CMOS로 구현하여 0.029 mm^2의 면적에서 0.7 V 공급으로 동작시켰고, 32 Gb/s에서 29.7 mW, 즉 0.93 pJ/bit의 에너지 효율을 얻었다. 고정된 1.05 GHz 기준 클럭에서 32.0, 29.33, 18.24 Gb/s 등 임의의 lane 데이터율을 생성함을 스펙트럼으로 보였고, 3000, 5000, 7000 ppm 깊이의 SSC를 추적하는 것을 확인하였다. SSC-PR를 켜면 5000 ppm SSC에서 입력-출력 위상 오차가 0.509에서 0.379 UIpp로 줄고, 100 MHz JTOL은 3000 ppm에서 0.180에서 0.260 UIpp로, 5000 ppm에서 0.128에서 0.252

UIpp로 개선되었다. PI 기반의 최근 CDR들과 비교하면 lane별 데이터율 가변, SRIS 환경의 위상 오차 제거, SSC 추적 실측 세 항목을 모두 갖추었다는 점에서 차별점이 있다. 다만 이 논문은 CDR에 초점을 둔 것으로 채널 손실 10.1 dB에 CTLE만을 사용한 NRZ 구현이므로, 같은 세션의 ADC-DSP 수신기들과 링크 성능으로 직접 비교할 대상은 아니다. PI 없이 0.7 V에서 동작하는 디지털 CDR로 SRIS의 깊은 SSC를 실측으로 추적한 점은, 다중 프로토콜과 EMI 요구를 동시에 받는 multi-lane 인터커넥트의 lane 단위 clocking에 실용적인 방향을 제시했다고 볼 수 있다.

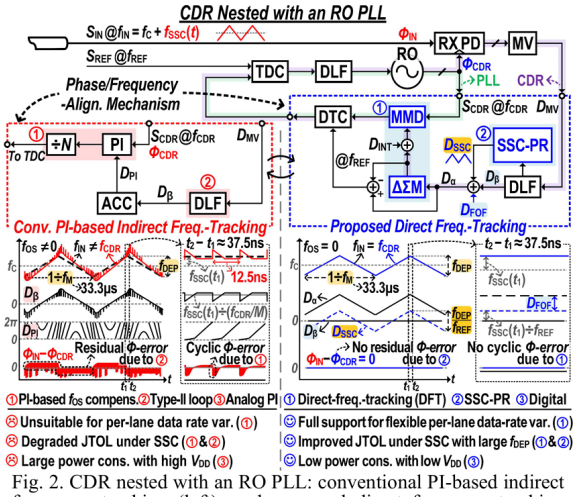


Fig. 2. CDR nested with an RO PLL: conventional PI-based indirect frequency tracking (left); and proposed direct frequency-tracking (DFT) (right).

[그림 2] 논문 25-3에서 제안하는 direct frequency-tracking CDR (오른쪽)과 기존 PI 기반 CDR (왼쪽)의 구조 및 SSC 조건에서의 위상 오차 비교

참고문헌

[1] E. Seevinck, P. J. van Beers and H. Ontrop, "Current-mode techniques for high-speed VLSI circuits with application to current sense amplifier for CMOS SRAM's," in IEEE Journal of Solid-State Circuits, vol. 26, no. 4, pp. 525-536, April 1991.

[2] R. Bashirullah, Wentai Liu and R. K. Cavin, "Current-mode signaling in deep submicrometer global interconnects," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 11, no. 3, pp. 406-417, June 2003.

[3] B. Casper and F. O'Mahony, "Clocking Analysis, Implementation and Measurement Techniques for High-Speed Data Links—A Tutorial," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 56, no. 1, pp. 17-39, Jan. 2009.

[4] J. Kim et al., "A 4 × 32 Gb/s 1.8 pJ/bit Collaborative Baud-Rate CDR With Background Eye-Climbing Algorithm and Low-Power Global Clock Distribution," in IEEE Journal of Solid-State Circuits, vol. 60, no. 8, pp. 2751-2764, Aug. 2025.

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

2026 VLSI Review

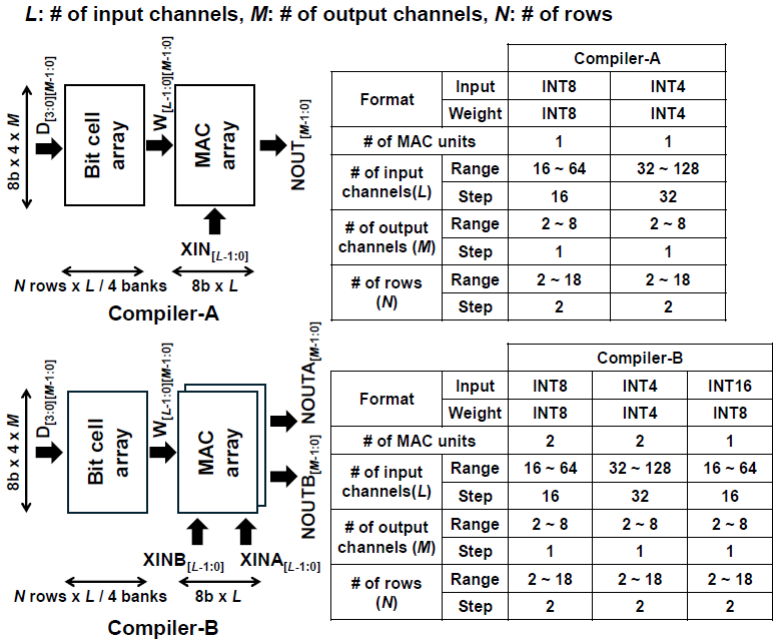
연세대학교 전기전자공학부 박사과정 여민준

Session 8 Compute-in-Memory

이번 2026 VLSI Symposium Session C8은 Compute-in-Memory(CIM)를 주제로, 2nm 첨단 공정의 디지털 CIM 컴파일러부터 부동소수점 CIM 매크로, ROM 기반 초저전력 키워드 인식 SoC, 스핀트로닉스 소자 기반 뉴로모픽 매크로에 이르기까지 메모리 내 연산의 스펙트럼 전반을 다루었다. 발표된 연구들은 다양한 데이터 포맷과 GEMM 형상에 대한 유연성, 트랜스포머 활성값의 outlier로 인한 정밀도 손실, 가산기 트리의 면적 부담, 자율주행 수준의 기능 안전 요구 등 CIM이 실제 응용에 배치될 때 마주하는 문제들을 회로-아키텍처-알고리즘 공동 설계로 해결하고자 하였다. 본 리뷰에서는 그 중 특징적인 네 편의 논문을 선정하여 핵심 아이디어와 성과를 정리한다.

#C8-1 TSMC에서 발표한 본 논문은 2nm 나노시트 공정에서 구현된 SRAM 기반 디지털 CIM(DCIM) 컴파일러로, 다양한 정수 데이터 포맷과 가중치당 복수의 MAC 유닛을 지원하여 면적 효율(TOPS/mm2)과 GEMM 활용률을 동시에 높이는 것을 목표로 한다. 기존 DCIM 매크로들이 단일 포맷과 고정된 MAC 구조에 묶여 최신 신경망의 혼합 정밀도 양자화와 다양한 GEMM 형상에 대응하기 어려웠던 한계를 극복하기 위해, 본 연구는 하나의 가중치를 두 개의 MAC 유닛이 공유하는 Compiler-B 구조를 제안하였다. 비트셀 어레이에서 읽은 가중치를 두 MAC 유닛이 공유하고 각각 독립적인 입력 활성값(XINA/XINB)을 받아 별도의 출력을 생성함으로써 가중치 재사용을 극대화하며, 행 전환(row switch) 시 로컬 SRAM 읽기 에너지가 두 MAC에 분산되어 매 사이클 행 전환 조건에서 단일 MAC 구조 대비 2% 높은 에너지 효율을 보인다. MAC 유닛은 4b×4b 곱셈기 네 개(MUL0–MUL3)와 두 개의 가산기 트리로 구성되어 INT8/INT4의 signed·unsigned를 모두 지원하며, INT4×INT4 모드에서는 MUL1·MUL2를 게이팅하고 시프트 없이 두 트리의 부분합을 누적하여 입력 채널 수를 두 배로 확장하고, NOUTA/NOUTB를 8비트 시프트 후 결합하면 INT16 입력도 처리한다. 0.0437μm2의 래치 기반 12T 비트셀에 인터리브 WL 구조를 적용하여 4개의 가중치를 한 사이클에 갱신하는 높은 쓰기 대역폭도 확보하였다. 테스트 칩 측정 결과 -40°C에서 95% 수율로 VMIN 0.38V 이하, 1.2V에서 FMAX 2.0GHz(Compiler-A)를 달성하였으며, 0.5V에서 INT8 88.8 TOPS/W, INT4 234.4 TOPS/W, 1.2V에서 INT4 511.9 TOPS/mm2의 효율을 기록하여 3nm 선행 컴파일러 대비 면적 효율

을 크게 개선하였다.



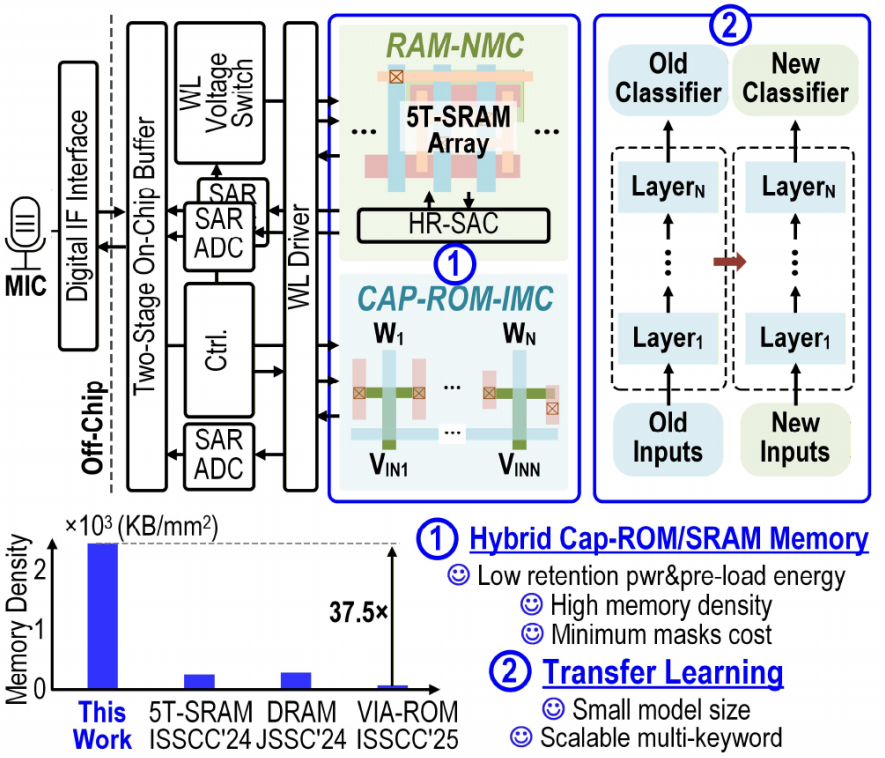
		Compiler-A	
Format	Input	INT8	INT4
	Weight	INT8	INT4
# of MAC units		1	1
# of input channels(*L*)	Range	16 ~ 64	32 ~ 128
	Step	16	32
# of output channels (*M*)	Range	2 ~ 8	2 ~ 8
	Step	1	1
# of rows (*N*)	Range	2 ~ 18	2 ~ 18
	Step	2	2

		Compiler-B		
Format	Input	INT8	INT4	INT16
	Weight	INT8	INT4	INT8
# of MAC units		2	2	1
# of input channels(*L*)	Range	16 ~ 64	32 ~ 128	16 ~ 64
	Step	16	32	16
# of output channels (*M*)	Range	2 ~ 8	2 ~ 8	2 ~ 8
	Step	1	1	1
# of rows (*N*)	Range	2 ~ 18	2 ~ 18	2 ~ 18
	Step	2	2	2

[그림 1] "A 2nm 234.4 TOPS/W and 511.9 TOPS/mm2 Digital Computing-in-Memory Compiler with Multiple MAC Units per Weight and Multiple Data Format Support" 2nm CIM 컴파일러 (Compiler-A/B)의 구조 및 지원 구성 범위

#C8-2 마카오 대학교와 EPFL이 공동 발표한 본 논문은 28nm 공정에서 구현된 0.0257mm2, 1.16μW의 항상 켜짐(always-on) 키워드 인식(KWS) SoC로, 어휘 수가 늘어날수록 모델과 메모리 요구량이 급증하여 전력이 수백 μW 수준으로 치솟는 문제를 해결하는 것을 목표로 한다. 기존 10개 이상 키워드를 지원하는 SoC들이 820–949μW를 소모하고 온칩 SRAM 부족으로 오프칩 플래시 접근 에너지가 추론 에너지의 20배를 넘었던 한계에 대해, 본 연구는 Cap-ROM 기반 인메모리 연산(Cap-ROM-IMC)과 5T-SRAM 기반 근접메모리 연산(SRAM-NMC)을 결합한 하이브리드 구조를 제안하였다. 모델 가중치의 92.7%는 금속 배선의 기생 커패시턴스로 고정 저장되는 트랜지스터 없는 Cap-ROM 셀에 담겨 전하 공유 방식의 비트 직렬 MAC을 수행하며, 나머지 7.3%만 1.16KB의 상시 유지 5T-SRAM에 저장된다. Virtuoso로 커패시턴스-가중치 매핑을 추출하고 M1–M6 금속 형상을 자동 생성하는 설계 흐름을 통해 신경망에서 물리 어레이로의 오류 없는 매핑을 구현하였고, 최소 2장의 마스크만으로도 VIA-ROM 대비 5.4배, 전체로는 37.5배 높은 메모리 밀도(2372.6KB/mm2)를 달성하면서 정적 전력을 거의 0으로 유지하여 시스템 전력을 44% 절감하였다. 또한 세 개의 공유 합성곱 계층과 키워드별 분류기로 구성된 전이 학습 기반 1D-CNN을 도입하여 20개 키워드 분류기를 모두 온칩에 저장하고 재학습 없이 임의의 5개 키워드를 동시에 활성화할 수 있게 하였으며, 클래스당 파라미터를

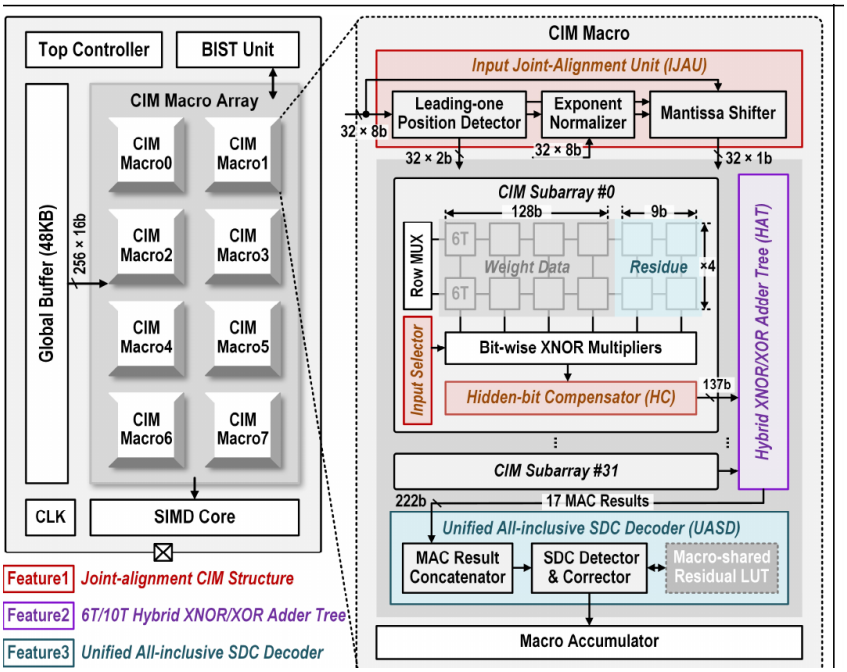
1.45KB로 줄였다. 측정 결과 GSCD-v2에서 92.1%의 정확도와 2ms 지연을 유지하면서 10개 이상 키워드 SoC 대비 700배 이상 낮은 전력을 기록하였다.



[그림 2] "A 0.0257 mm2, 1.16 μW, Mask-Efficient KWS SoC with Hybrid Cap-ROM-IMC/SRAM-NMC and Transfer Learning for Flexible Multi-Keyword Recognition" 제안 KWS SoC의 블록도와 두 가지 핵심 해법(하이브리드 Cap-ROM/SRAM 메모리, 전이학습)

#C8-3 칭화대학교와 홍콩과기대학교가 공동 발표한 본 논문은 28nm 공정에서 구현된 자율주행용 부동소수점 CIM(FP-CIM) 매크로로, 정확도·성능·신뢰성 요구가 모두 엄격한 안전 필수 응용에 FP-CIM을 적용하는 것을 목표로 한다. 기존 FP-CIM이 지수만 정렬(exponent-only alignment)한 뒤 절단 오차를 줄이기 위해 가수 비트폭을 확장하여 비트 직렬 연산 지연이 22–44% 늘어나고, 가산기 트리가 매크로 면적의 34–71%를 차지하며, 메모리와 연산 로직을 ECC와 TMR로 따로 보호하느라 과도한 비용이 드는 세 가지 한계를 지적하였다. 이를 해결하기 위해 첫째, 히든 비트와 가수의 선행 1 사이의 본질적으로 무효인 비트 공간을 유효 비트로 재활용하는 지수-가수 joint-alignment 구조를 제안하여, 비트폭 확장 없이 1.62%의 추가 면적만으로 정확도 손실을 3.97배 줄이고 31.24%의 면적 효율 손실을 회피하였다. 둘째, Sum과 Cout에 공통된 XOR 항을 미리 계산해 재사용하는 하이브리드 XNOR/XOR 가산기 트리(HAT)를 도입하여 1단은 6T XNOR 기반, 그 외 단은 10T XOR 기반 전가산기로 구성함으로써 28T 가산기 트리 대비 면적을 71.35% 절감하였다. 셋째, 가중치를 잔여 코드(residue code)로 인코딩해 저장하고 최종 MAC 결과의 나머지 검사로 오류를 검출·복원하는 통합 SDC 디코더(UASD)를 통해 8.41%의 면적

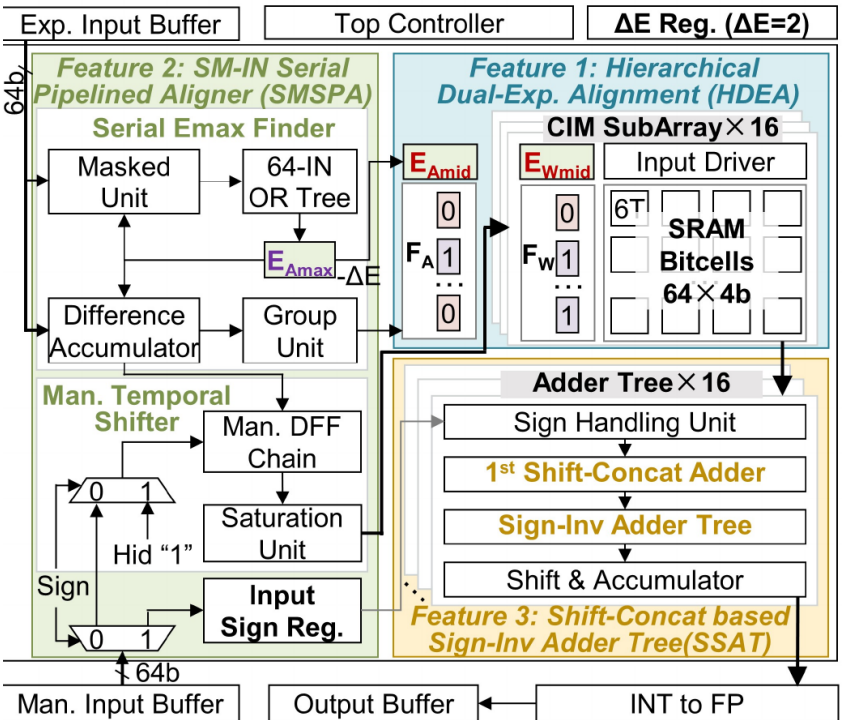
비용만으로 0.50%의 SDC율과 ISO 26262 ASIL-D 안전 등급을 달성하였다. 측정 결과 BF16 기준 44.15 TFLOPS/W, 1.57 TFLOPS/mm2로 선행 FP-CIM 대비 각각 1.51–1.62배, 1.37배의 개선을 보였다.



[그림 3] "A 28nm 44.15TFLOPS/W and 1.57TFLOPS/mm2 Joint-Alignment Floating-point CIM Macro with ISO26262 ASIL-D Safety Level for Autonomous Driving" 제안 FP-CIM의 전체 아키텍처와 세 가지 핵심 기능

#C8-5 화중과기대학교와 홍콩과기대학교가 공동 발표한 본 논문은 40nm 공정에서 구현된 FP8 디지털 CIM 매크로로, 트랜스포머 활성값과 가중치의 outlier 분포로 인해 INT 양자화가 심각한 클리핑·반올림 오차를 겪는 에지 트랜스포머 가속을 목표로 한다. 기존 FP-CIM은 공유 지수로 사전 정렬(pre-alignment)하면 정확도가 떨어지고, 사후 정렬(post-alignment)은 대규모 시프터와 19비트 폭의 가산기 트리로 면적·에너지 부담이 큰 상충 관계에 갇혀 있었다. 본 연구는 이를 세 가지 기법으로 해결하였다. 첫째, 계층적 이중 지수 정렬(HDEA)은 입력을 동적으로 분류하여 outlier는 Emax에, 정상값은 Emid=Emax-ΔE에 정렬함으로써 좌측 시프트량을 제한하고 가산기 트리 폭을 8비트로 줄이며, 설계 공간 탐색을 통해 ΔE=2를 정확도와 면적의 최적점으로 도출하였다. 둘째, 부호-크기 입력 직렬 파이프라인 정렬기(SMSPA)는 마스크 기반 직렬 Emax 탐색, 포화 검사로 불필요한 시프트를 게이팅하는 차분 누산기, DFF 체인을 이용한 시간 영역 시프터로 병렬 정렬과 비트 직렬 MAC 간의 파이프라인 정체를 제거하여 지연·면적·에너지를 각각 20%, 41.1%, 34.1% 절감하였다. 셋째, Shift-Concat 기반 부호 반전 가산기 트리(SSAT)는 첫 단에서 겹치는 상위 4비트만 더하고 하위 4비트는 연결하며, 이후 단에서는 부호 비트를 반전시켜 부호 있는 덧셈을 부호 없는 덧셈으로 바꿔 불필요한 부호 확장과

스위칭을 제거하여 면적 25.9%, 에너지 18.2%를 절감하였다. 그 결과 0.8–1.1V, 100–326MHz에서 24.85–80.59 TFLOPS/W를 달성하여 최신 사후 정렬 FP-CIM 대비 2.81–6.66배 높은 에너지 효율을 기록하였다.



[그림 4] "A 40nm 24.85-to-80.59 TFLOPS/W FP8 Computing-in-Memory Macro with Outlier-Aware Hierarchical-Alignment for Edge Transformers" 제안 CIM 매크로의 전체 아키텍처(SMSPA, HDEA, SSAT)

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/

2026 VLSI Review

고려대학교 전기전자공학과 박사과정 한창우

Session 29 Advanced Memory Design

이번 2026 VLSI의 Session 29에서는 Advanced Memory Design을 주제로 총 5편의 논문이 발표되었다. 본 세션에서는 SRAM, DRAM, Mask ROM, Anti-Fuse OTP 및 STT-MRAM 등 다양한 메모리를 대상으로, 선단 공정에서의 고집적화와 저전력·저전압 동작, 동작 안정성 및 신뢰성 향상을 위한 회로 및 아키텍처 기술이 소개되었다. 특히 공정 미세화와 공급 전압 감소에 따라 발생하는 동작 마진 저하와 전력 소모 문제를 해결하면서도 높은 성능과 집적도를 확보하기 위한 다양한 설계 기법이 제안되었다.

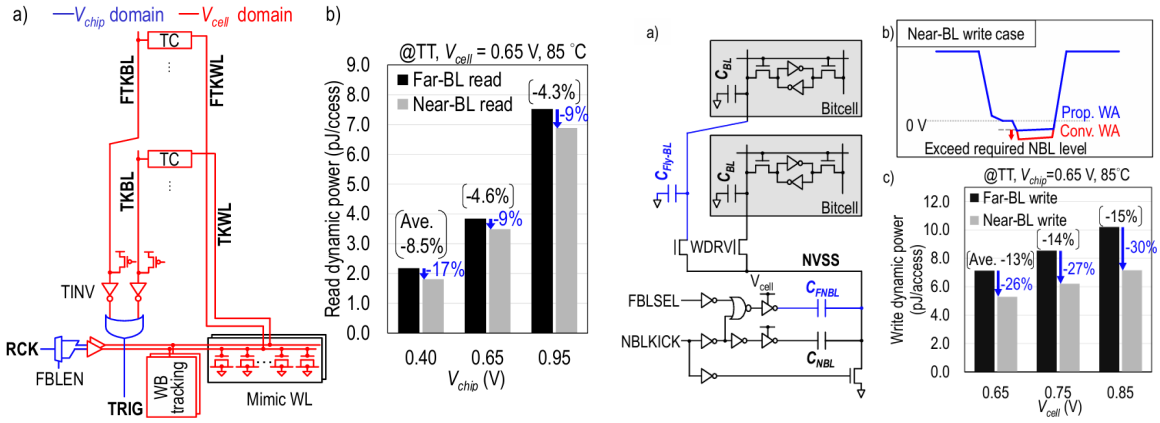
본 리뷰에서는 2 nm Nanosheet SRAM의 에너지 효율 향상 기술, 0.75 V 저전압 DRAM의 sensing 신뢰성 개선 기술, 그리고 PVT variation에 강인한 STT-MRAM의 dual-reference sensing 기술을 다룬 3편의 논문을 중심으로 살펴보고자 한다.

#29-1 본 논문은 TSMC에서 발표한 논문으로, 2 nm Nanosheet 공정 기반의 고밀도 6T SRAM에서 read/write 에너지 효율을 향상시키기 위한 Row-Access Aware Read Tracking 및 Write Assist(WA) 회로를 제안한다. 기존 dual-rail SRAM에서는 array와 peripheral의 공급 전압(V_{cell}, V_{chip}) 차이로 인해 read tracking과 실제 bit-line(BL) 방전 시간 사이에 오차가 발생하여 불필요한 동적 전력이 소모될 수 있다. 또한 기존 Negative Bit-Line(NBL) Write Assist는 worst-case bitcell을 기준으로 일정한 boost capacitance를 사용하기 때문에, 부하가 작은 near-array에서는 과도한 NBL boost로 인해 write power가 증가하는 문제가 존재하였다.

이를 해결하기 위해 near/far-array의 접근 위치를 고려한 read tracking 회로를 적용하고, tracking BL과 inverter를 memory array와 동일한 V_{cell} domain에서 동작시켜 실제 BL에 맞춰 Sense Amplifier Enable(SAE)과 WL 비활성화 시점을 조절하였다. 그 결과 worst-case에서 불필요한 BL differential swing을 45% 감소시키고, read dynamic power를 최대 8.7%, 전체 row access 평균 8.5% 절감하였다.

Write 동작에서는 접근 위치에 따라 NBL boost capacitance를 조절하는 Row-Access Aware Write Assist를 적용하여 near-array의 NBL over-boost를 억제하였다. 이를 통해 near-array의 write power를 26~30%, 전체 row access 평균 최대 15% 절감하였다. 최종적으로 2 nm 공정에서 제작된 539-kbit dual-rail SRAM macro는 37.42 Mbit/mm^{2}의 bit

density와 125 °C에서 0.35~1.10 V의 동작 범위, 2.28 pJ/access의 read/write 에너지 효율을 달성하였다.

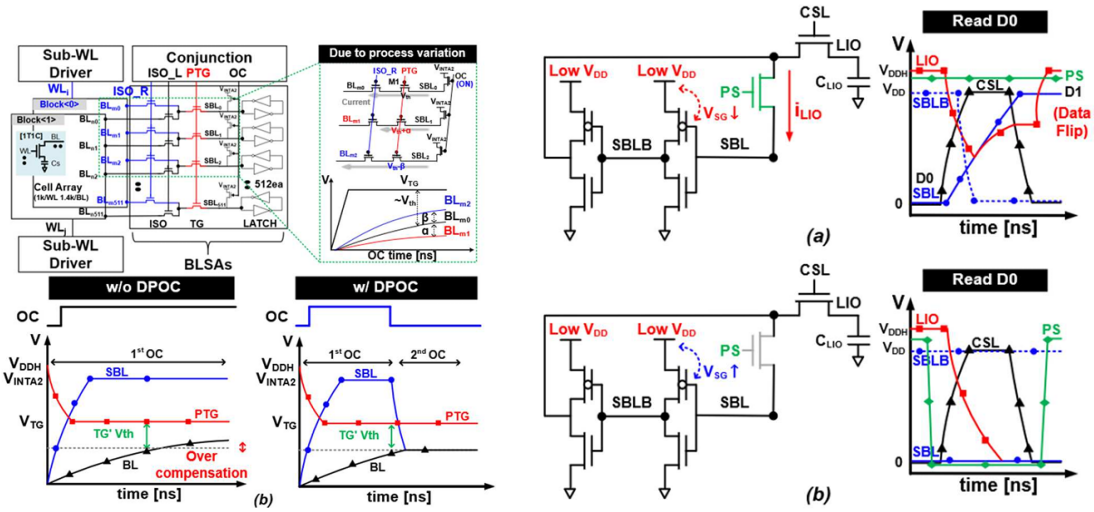


[그림 1] 제안된 Row-Access Aware Read Tracking 및 Write Assist 회로와 전력 절감 효과

#29-2 본 논문은 삼성전자에서 발표한 논문으로, 저전압 DRAM에서 sensing 신뢰성과 전력 효율을 향상시키기 위한 Fully Single-Ended Core Data-Line Architecture를 제안한다. 기존 DRAM은 공급 전압이 낮아질수록 bit-line의 charge-sharing voltage와 sense amplifier의 voltage headroom이 감소하여 안정적인 sensing이 어려워지며, single-ended charge-transfer sensing에서는 offset compensation 과정의 over-compensation과 CSL 활성화에 따른 latch disturb 문제가 발생할 수 있다.

이를 해결하기 위해 본 논문에서는 offset compensation을 두 단계로 분리하는 Dual-Phase Offset Compensation(DPOC) 기법을 적용하여 charge-transfer 이전의 over-compensation을 억제하였으며, 이를 통해 1 ns offset compensation window에서 Fail Bit Count(FBC)를 99.9% 감소시켰다. 또한 CSL pulse 동안 latch를 일시적으로 격리하는 Conditional Latch Gating(CLG) 기법을 적용하여 저전압에서 발생하는 data flip을 방지하였으며, V_{DD}=0.75 V에서 FBC를 99.4% 감소시켰다.

또한 BLSA부터 GIO까지 data path를 fully single-ended 구조로 구성하여 기존 differential 구조에서 요구되던 GIO metal layer 하나를 제거함으로써 약 2%의 backend cost를 절감하였다. 여기에 Data-State Inversion(DSI) 기법을 적용하여 동작 전류를 줄였으며, 기존 구조 대비 IDD0를 평균 27% 감소시켰다. 최종적으로 10-nm-class DRAM 공정에서 제작된 test chip을 통해 0.75 V의 저전압에서도 높은 sensing 신뢰성과 전력 효율을 동시에 확보할 수 있음을 실증하였다.

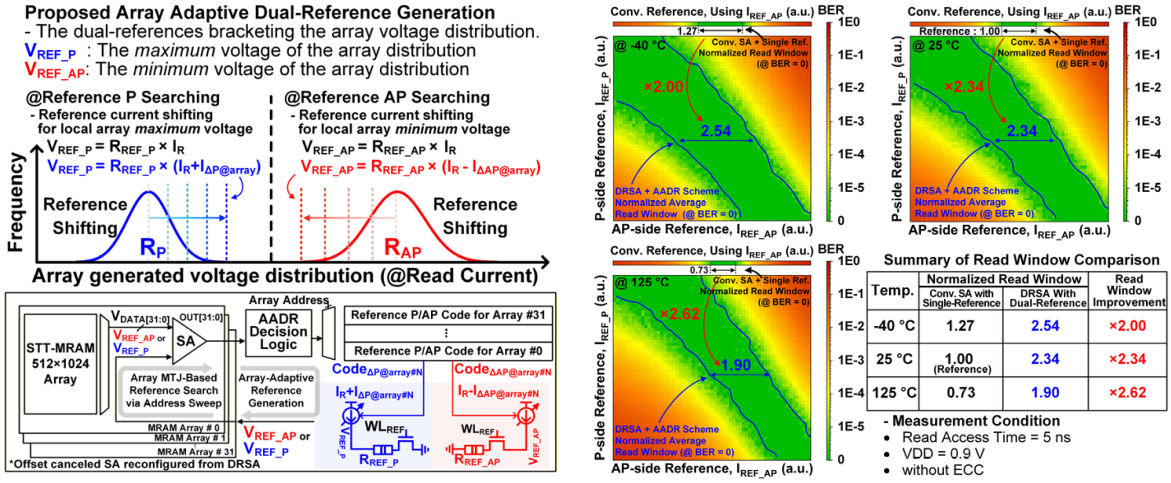


[그림 1] 제안된 DPOC 및 CLG 기법을 이용한 저전압 DRAM sensing 신뢰성 향상 구조

#29-5 본 논문은 성균관대학교와 삼성전자 공동 연구팀에서 발표한 논문으로, 28 nm FD-SOI 공정 기반 16 Mb STT-MRAM에서 PVT variation에 따른 read margin 저하를 개선하기 위한 Dual-Reference Sense Amplifier(DRSA)와 Array-Adaptive Dual-Reference(AADR) 기법을 제안한다. STT-MRAM은 높은 집적도와 비휘발성 특성을 갖지만, MTJ의 낮은 Tunnel Magnetoresistance(TMR)와 공정·전압·온도 변화로 인해 P/AP 상태의 저항 분포가 변하면서 read sensing margin과 access speed가 저하될 수 있다.

이를 해결하기 위해 본 논문에서는 하나의 reference를 사용하는 기존 방식 대신, P-state와 AP-state에 각각 대응하는 두 개의 reference를 사용하는 DRSA를 적용하였다. 또한 AADR을 통해 각 array의 저항 분포를 탐색하여 V_{REF_P}는 P-state 분포의 상단에, V_{REF_AP}는 AP-state 분포의 하단에 위치하도록 설정함으로써 local variation에 따라 reference를 보정하였다. 해당 calibration은 power-on 시 한 번 수행되고, 이후에는 저장된 digital code를 이용하여 동작하므로 runtime overhead를 최소화하였다.

그 결과 제안된 DRSA와 AADR 구조는 기존 single-reference 방식 대비 zero-BER read window를 -40°C, 25°C, 125°C에서 각각 2.00배, 2.34배, 2.62배 확대하였다. 또한 0.9 V에서 25°C 기준 2.0 ns, 125°C 기준 3.3 ns의 read access time을 달성하였으며, 25°C에서 0.6 pJ/bit의 read energy를 보였다. 이를 통해 넓은 온도 범위와 PVT variation 조건에서도 안정적이고 빠른 STT-MRAM read 동작이 가능함을 실리콘 측정을 통해 입증하였다.



[그림 1] 제안된 Array-Adaptive Dual-Reference 생성 방식 및
PVT 조건에서의 read window 향상 결과

저자정보



한창우 박사과정 대학원생

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

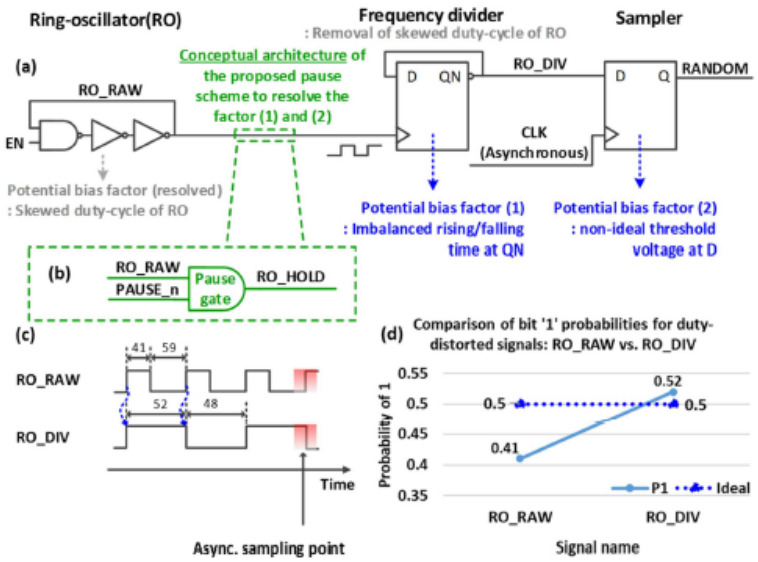
2026 VLSI Review

한국과학기술원 전기및전자공학부 석사과정 박성용

Session 12 True Random Number Generators

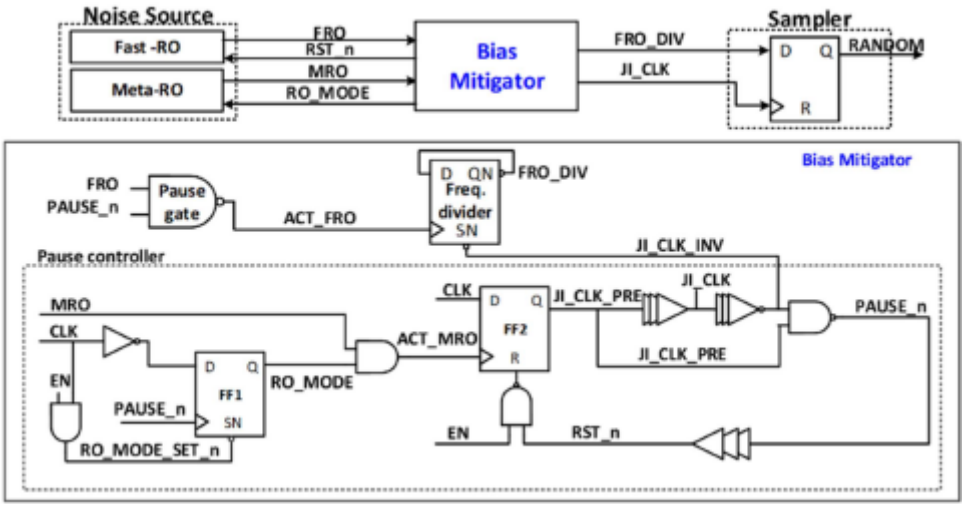
이번 Session 12에서는 총 4편의 TRNG(True Random Number Generator) 관련 논문이 발표되었다. TRNG는 반도체 회로에서 발생하는 물리적 noise를 이용해 0 또는 1의 랜덤 비트를 생성하는 회로이다. 아래의 논문들은 각각 advanced-node에서 기존 RO(Ring Oscillator) 기반 TRNG의 bias를 직접 제거하는 방법과, differential CMOS thyristor를 이용하여 후처리 없이 높은 엔트로피를 얻는 방법을 통해 TRNG 하드웨어의 문제를 해결하는 데에 중점을 두었다.

#12-1 삼성전자의 논문은 4nm FinFET 공정에서 RO 기반 TRNG의 bias를 줄이면서 기존 RO 구조의 단순성과 작은 면적을 유지하고자 하였다. RO 기반 TRNG는 oscillator의 phase jitter를 비동기적으로 샘플링하여 랜덤 비트를 생성하므로 구조가 단순하지만, RO의 duty-cycle이 50%에서 벗어나면 0과 1의 샘플링 확률이 달라진다. 이를 보정하기 위해 frequency divider를 사용하더라도 선단 공정에서는 divider의 상승 및 하강 트랜지션 시간 차이와 샘플러 D 입력의 비이상적 threshold가 새로운 bias 원인이 될 수 있다. 본 연구는 이러한 bias가 RO 자체뿐 아니라 랜덤 비트를 받아들이는 플립플롭과 샘플러에서도 추가로 발생한다는 점에 주목하고, 샘플링 직전에 RO 트랜지션을 일시적으로 차단하는 Bias Mitigator를 제안하였다.



[그림 1] 기존 RO 기반 TRNG의 bias 원인과 pause-gate 기반 bias mitigation

그림 1은 기존 RO 기반 TRNG에서 bias가 발생하는 원인과 제안된 회로의 기본 아이디어를 보여준다. 일반적인 구조에서는 frequency divider를 사용하면 RO의 duty-cycle 불균형을 줄일 수 있다. 선단 공정에서는 divider 내부의 상승·하강 전이 특성이나 샘플러의 비이상적인 입력 threshold로 인해 0과 1이 유지되는 시간이 달라질 수 있다. 이 경우 비동기적으로 샘플링되는 0과 1의 확률이 달라져, TRNG 출력 비트가 1일 확률인 P(1)이 이상적인 값인 0.5에서 벗어나게 된다. 제안된 구조는 divider 앞에 pause gate를 추가하여, 샘플링 직전에 새로운 트랜지션을 차단함으로써, 플립플롭의 트랜지션 불균형과 샘플러 threshold가 샘플링 결과에 미치는 영향을 줄일 수 있다. 이를 통해 복잡한 calibration 없이, 현재 값이 0인지 1인지에 관계없이 동일한 조건에서 샘플링되도록 하여 divider와 샘플러의 비대칭이 최종 랜덤 비트의 bias로 이어지는 것을 억제한다.



[그림 2] Bias Mitigator가 포함된 dual-RO TRNG의 전체 구조

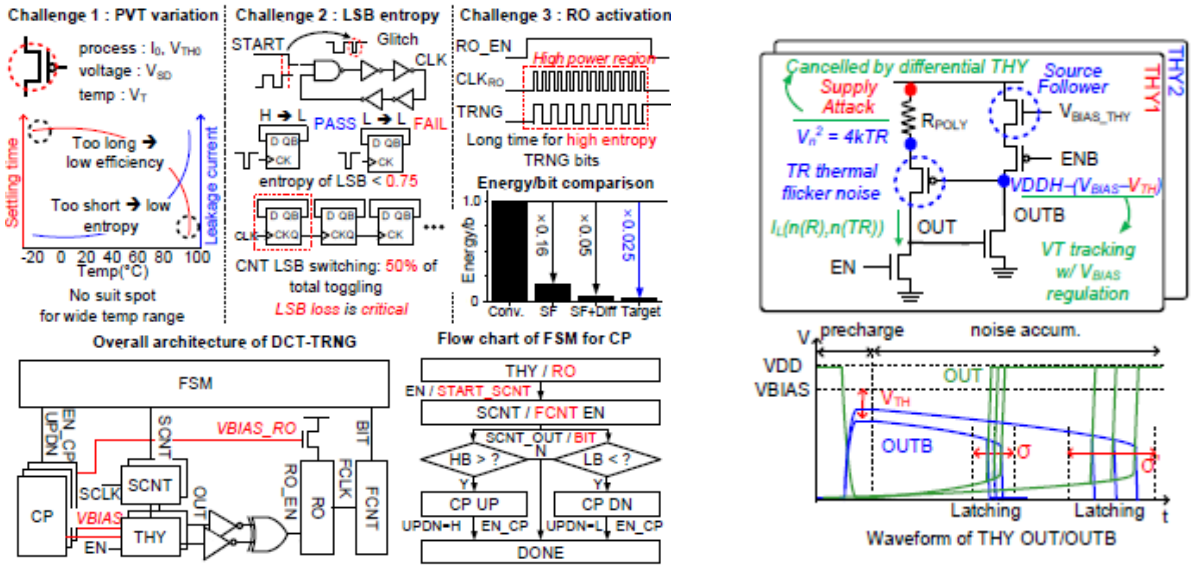
그림 2는 제안된 TRNG의 전체 구조이다. Noise source는 Fast-RO와 Meta-RO 두 경로로 구성된다. Meta-RO에서는 metastability에 의한 초기 위상 jitter와 이후 RO의 jitter를 함께 이용한다. Fast-RO는 독립적으로 jitter를 누적하며, frequency divider를 거친 FRO_DIV가 최종 샘플링 신호가 된다. Bias Mitigator는 pause gate, frequency divider, pause controller로 구성되며, Meta-RO의 첫 상승 엣지를 이용해 샘플링보다 먼저 PAUSE_n을 생성한다. 따라서 샘플링 직전의 변화를 정지시키고, Meta-RO의 무작위성 timing을 이용해 샘플러를 동작시킨다. 랜덤 비트 생성 이후 내부 상태를 리셋하여 이전 비트의 상관관계도 줄인다. 이를 통해 기존 dual-RO 구조의 jitter 기반 엔트로피 생성 방식은 유지하면서, divider와 샘플러의 mismatch에서 추가적으로 발생하는 bias를 억제하였다.

기존 RO 기반 TRNG에서는 RO에서 충분한 jitter를 생성하는 것뿐만 아니라, 생성된 무작위성을 최종 디지털 비트로 변환하는 샘플링 과정 또한 중요하다. 특히 선단 공정에

서는 소자의 mismatch가 증가하면서 frequency divider와 샘플러와 같은 회로의 작은 비대칭도 반복적인 샘플링 과정에서 bias로 이어질 수 있다. 이러한 문제를 별도의 복잡한 보정 회로나 후처리로 해결하는 대신, 샘플링 직전에 신호의 변화를 일시적으로 정지시키는 간단한 회로를 추가하였다.

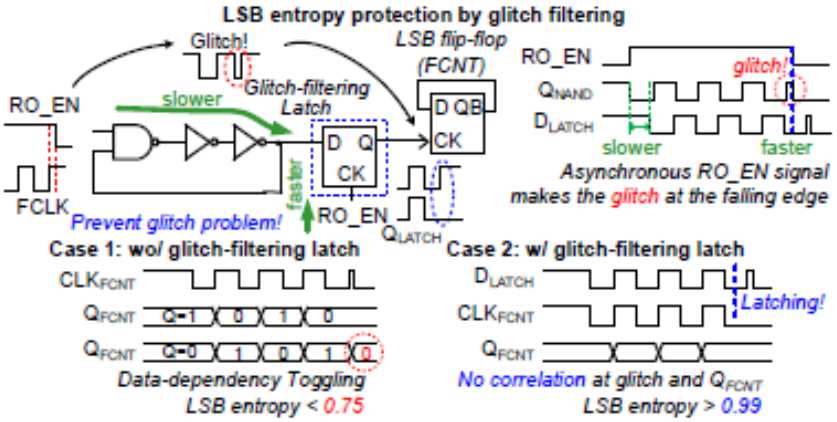
제안된 기법은 5개의 process corner에서 총 290개의 칩을 대상으로 0.675~0.935V, -40~150℃ 범위에서 측정되었다. 대표 조건에서 Bias Mitigator를 사용하지 않은 P(1)은 0.4471이었으나 제안된 pause 동작을 적용하면 0.5013으로 개선되어 97.54%의 bias reduction을 보였다. NIST SP 800-90B 평가에서는 별도의 15개 칩을 사용하였으며 IID track에서 평균 0.986 bit/bit의 최소 엔트로피를 얻었다. 스루풋은 4Mbps, TRNG core 면적은 80.198μm², 소비전력은 2.16mW이다. 따라서 해당 연구는 미세공정에서 커지는 mismatch로 인한 bias를 간단한 샘플링 제어 회로로 억제하여 bias 개선 효과와 높은 최소 엔트로피를 확보하였으며, 이를 80.198μm²의 작은 RO 기반 구조로 구현하여 선단 공정에서의 실용적인 TRNG 구현 가능성을 보여주었다.

#12-2 성균관대학교의 논문은 differential CMOS thyristor의 latching-time jitter을 엔트로피 소스로 사용하여 PVT 변화에 강하면서도 낮은 비트당 에너지를 갖는 후처리 없는 TRNG를 구현하고자 하였다. 기존의 시간 기반 TRNG는 noise를 충분히 누적하기 위해 RO를 오랫동안 동작시켜 에너지 소비가 크고, metastability 기반 TRNG는 mismatch와 PVT 변화에 민감하여 calibration이나 XOR 후처리를 사용하는 경우가 많다. 특히 XOR 방식은 일부 엔트로피 소스가 fault나 attack으로 deterministic하게 변하더라도 최종 출력이 랜덤하게 보일 수 있어 source failure가 가려질 수 있다. 이를 해결하기 위해 1) thyristor의 noise 누적 시간을 일정하게 유지하는 Bias Self-Regulation, 2) 비동기 RO disable에서 발생하는 glitch를 제거하는 Glitch-Filtering Latch, 3) 두 thyristor의 latching time 차이 동안에만 RO를 활성화하는 Differential Short-Window Activation을 포함한 DCT-TRNG 구조를 제안하였다.



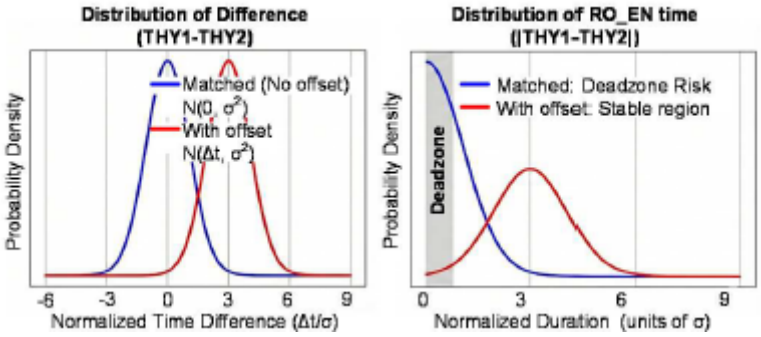
[그림 3] DCT-TRNG의 전체 구조와 CMOS thyristor의 noise 누적 원리

그림 3은 DCT-TRNG의 전체 구조와 엔트로피 소스의 동작을 보여준다. Thyristor는 누설전류에 의해 내부 노드가 천천히 변화하다가 피드백 조건에 도달하면 상태가 뒤집히며 latch되는 구조를 이용해 무작위성을 만들며, precharge부터 latching까지의 시간을 T_{acc}로 정의한다. T_{acc}에는 누설전류와 noise에 의한 jitter가 함께 포함되므로 시간 자체가 무작위성을 띈다. 그러나 누설전류는 PVT에 민감하여 고정된 bias에서는 저온에서 T_{acc}가 너무 길어 에너지를 낭비하고, 고온에서는 너무 짧아 noise 누적이 부족할 수 있다. 이를 해결하기 위해 소스 팔로워로 PMOS를 weak inversion 영역에서 동작시키고, 시스템 클럭 카운터(SCNT)가 실제 T_{acc}를 측정하여 FSM과 charge pump가 bias 전압을 조절한다. 동시에 주파수 카운터(FCNT)는 RO 주파수를 측정하여 time-to-digital 변환의 양자화 단계도 PVT에 따라 변하지 않도록 한다. Differential thyristor의 common source에 삽입된 폴리실리콘 저항은 common-mode supply noise를 줄이면서 differential noise 성분을 이용하도록 한다.



[그림 4] Glitch-filtering latch를 이용한 LSB entropy protection

카운터의 LSB는 가장 빈번하게 변화하는 비트이므로, LSB의 엔트로피가 낮아 이를 버리게 되면 비트당 에너지 효율이 크게 저하될 수 있다. 그림 4는 비동기 RO 활성 신호에서 발생하는 glitch가 LSB의 엔트로피를 낮추는 원인과 이를 해결하는 구조를 보여준다. 두 thyristor의 latching event로 만들어진 RO_EN은 시스템 클럭과 비동기이므로 하강 엣지가 카운터 클럭과 가까우면 짧은 glitch가 발생할 수 있다. 이 glitch가 카운터의 LSB 플립플롭에 직접 전달되면 이전 Q state에 따라 setup/hold violation의 결과가 달라져 데이터 의존 bias가 생기며, 논문에서는 LSB 최소 엔트로피가 0.75 이하로 감소할 수 있음을 보였다. 제안된 구조는 RO_EN 경로에 transparent latch를 삽입하여 glitch와 현재 Q state 사이의 상관관계를 제거한다. 이를 통해 LSB 최소 엔트로피를 0.99 이상으로 회복하여 가장 많이 toggle하는 LSB를 버리지 않고 사용할 수 있다.



[그림 5] Differential latching-time의 entropy deadzone

한편, 두 thyristor의 latching time을 최대한 같게 만드는 것이 항상 유리한 것은 아니다. 본 구조에서는 THY1과 THY2가 latch되는 시간 차이 동안만 RO가 활성화되므로 두 시간이 지나치게 가까우면 RO 활성 시간이 거의 0이 된다. 이 경우 |THY1-THY2|의 분포가 0에서 folding되면서 부호 정보를 잃고, RO가 충분한 엣지를 만들기 전에 비활성되어 zero-count가 반복될 수 있다. 저자들은 이를 엔트로피 데드존(entropy deadzone)으로 정의하였으며, RO 활성 시간이 약 9 ns 이하인 영역에서 NIST Runs test의 성공률이 급격히 저하됨을 확인하였다. 이를 해결하기 위해 두 thyristor의 평균 latching time에 약 45ns의 오프셋을 두었다. 이 오프셋은 무작위성을 만드는 값이 아니라 분포를 데드존 밖으로 이동시키는 역할이며, 사이클 간 편차는 여전히 누적된 noise에 의해 결정된다. 따라서 두 경로를 완전히 매칭하는 대신 최소 RO 활성 시간을 보장하여 엔트로피와 효율성 사이의 trade-off를 해결하였다.

28nm CMOS 공정에서 제작된 DCT-TRNG는 0.8V에서 13.3Mb/s의 스루풋과 0.33pJ/bit의 에너지 효율을 달성하였다. 5개 칩에 대해 0.8~1.1V, -40~125℃에서 측정한 섀넌 엔트로피는 0.99999 이상으로 유지되었으며, NIST SP 800-22/90B test와 power-supply injection 조건에서도 raw random bit을 검증하였다. 제안된 구조는 XOR 후처리 없이 Bias Self-Regulation으로 PVT에 따른 noise 누적 시간을 유지하고, Glitch-Filtering

Latch로 LSB 엔트로피 손실을 막으며, 두 thyristor 사이에 오프셋을 두어 엔트로피 데드존에서 벗어났다. 따라서 엔트로피가 생성되어 디지털 비트로 변환되는 전체 경로의 실패 원인을 회로적으로 제거하여 낮은 비트당 에너지와 높은 엔트로피를 동시에 달성하였다.

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/