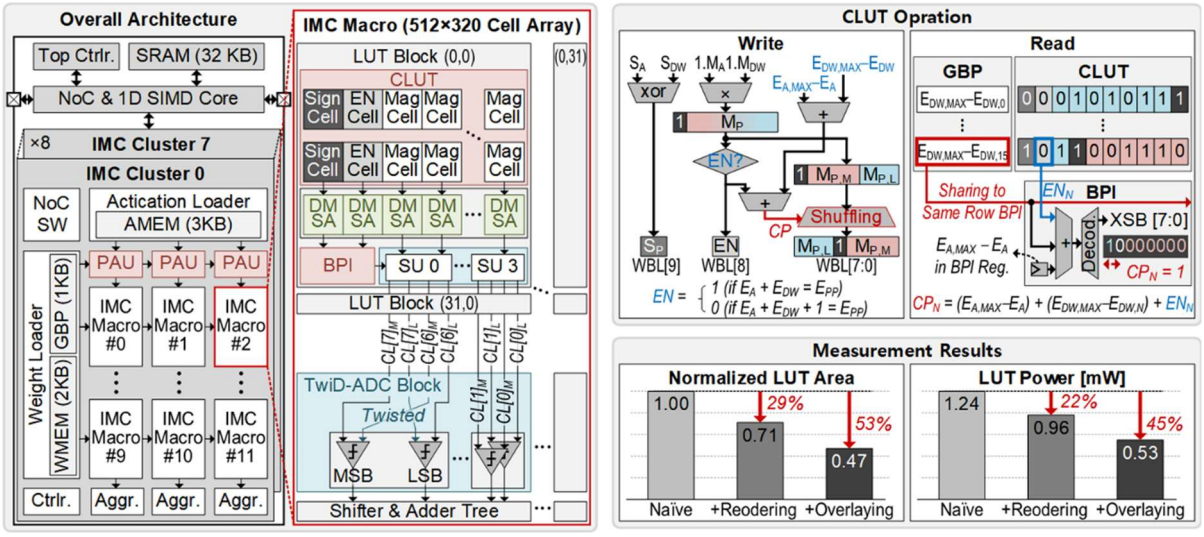
2025 IEEE VLSI Review

포항공과대학교 반도체대학원 박사과정 박은빈

Session 17 CIM-based AI Accelerators

이번 VLSI 2025의 Session C17 CIM-based AI Accelerators는 인메모리 컴퓨팅이 단순한 행렬 곱셈 가속을 넘어 AI 전반과 복잡한 문제 해결까지 영역을 넓혀가고 있음을 보여주었다. 특히 이번 세션의 공통된 흐름은 세 가지로 요약된다. 첫째, DRAM·ReRAM·SRAM 등 다양한 메모리 기술을 혼합 활용해 CIM의 한계를 보완하고 에너지 효율을 끌어올리는 시도가 활발했다. 둘째, Transformer와 LLM과 같은 최신 AI 워크로드에 특화된 아키텍처가 다수 제시되며, 기존 CNN 중심의 설계를 넘어 실질적인 응용지향 최적화를 강조했다. 셋째, AI 가속을 넘어 SAT와 같은 NP-hard 문제 해결로 응용 범위를 확장하는 연구도 등장해, CIM이 범용 연산 가속기에서 특수 목적 연산 플랫폼으로 진화하고 있음을 보여주었다. 종합적으로 이번 세션은 메모리-연산 융합 구조가 단순한 효율 개선을 넘어, 초저전력·실시간·범용성을 동시에 추구하는 차세대 AI 및 최적화 컴퓨팅의 핵심 축으로 자리잡아가는 흐름을 잘 드러냈다.

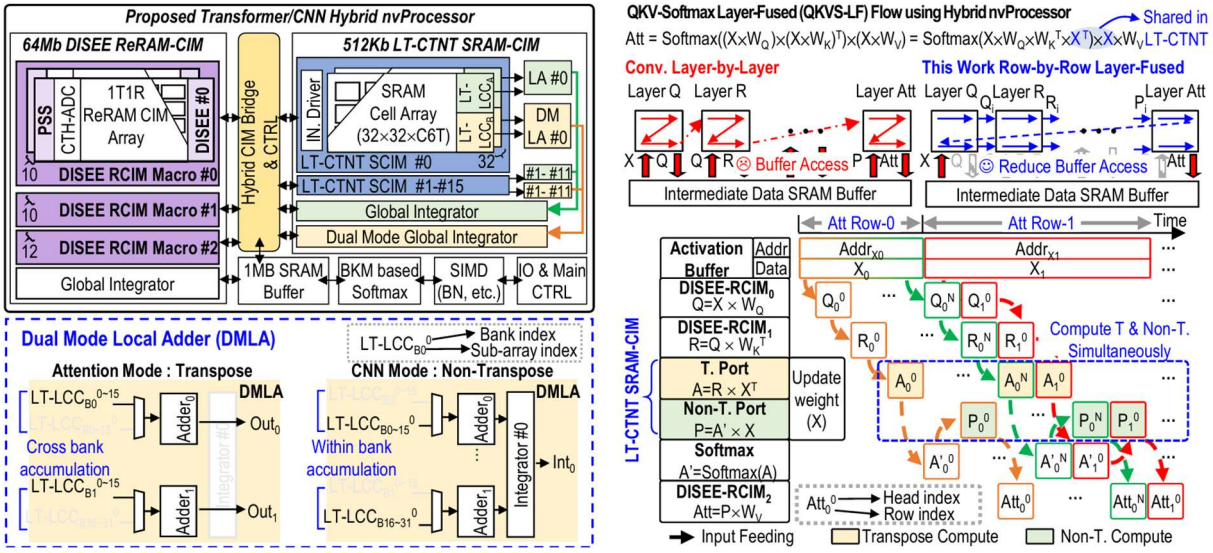
#C17-1은 KAIST와 삼성전자 연구진이 발표한 DIAL (DRAM In-memory Accelerator with LUT and Twisted ADC)로, DRAM 기반 CIM에서 발생하는 LUT와 ADC의 비효율을 동시에 해결한 점이 큰 특징이다. 기존 DRAM CIM은 고정밀 연산을 지원할 때 LUT에 불필요한 0이 다수 포함되어 면적과 전력이 낭비되고, ADC가 병목이 되어 에너지 효율이 급격히 떨어지는 문제가 있었다.



[그림 1] 제안한 DIAL의 전반적인 구조 및 CLUT의 동작 설명과 측정결과

이를 해결하기 위해 DIAL은 Compact LUT (CLUT) 구조를 도입해 불필요한 확장 0을 제거하여 LUT 면적과 전력을 각각 53%, 45% 절감하였고, Twisted Differential ADC (TwiD-ADC)를 설계해 비교기 공유 방식으로 ADC 면적과 전력을 줄이는 동시에 SNDR을 크게 개선하였다. 또한 Dual-Mode Sense Amplifier(DMSA)를 도입해 데이터 패턴에 따라 동작 모드를 전환함으로써 LUT 전력 소모를 추가로 줄였다. 28nm 공정에서 구현된 칩은 GPT-2, ViT 등 고난도 AI 모델에서도 동작이 검증되었으며, 최대 55.4 TFLOPS/W의 에너지 효율을 달성해 기존 DRAM CIM 대비 4.1배 개선된 성능을 보여주었다. 이 논문은 DRAM 기반 CIM이 단순한 이미지 처리 수준을 넘어 LLM과 Transformer와 같은 대규모 AI 모델까지 커버할 수 있는 실질적 확장 가능성을 제시했다는 점에서 의의가 크다.

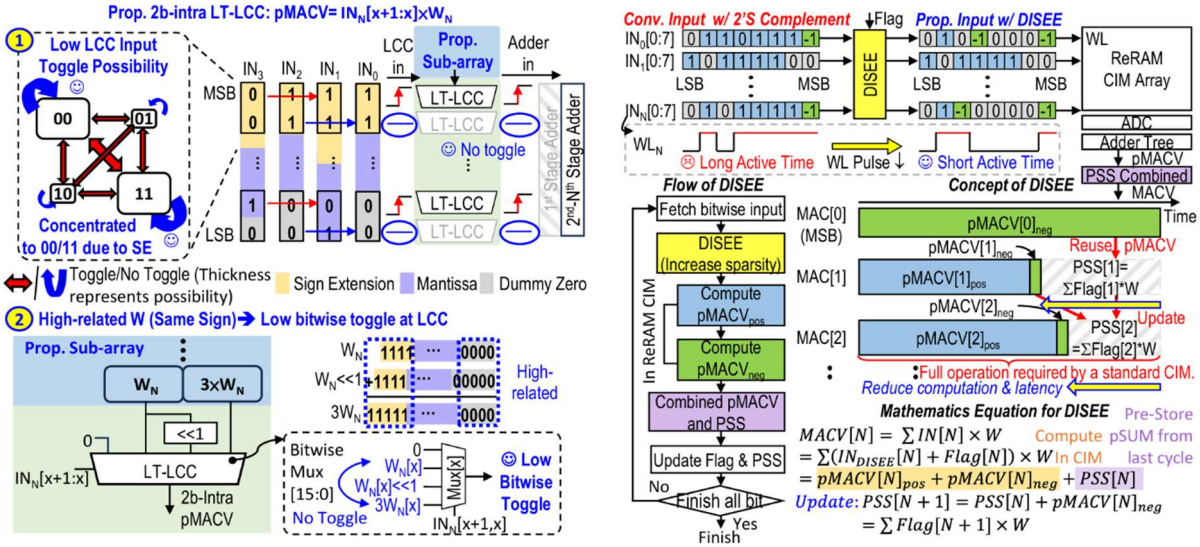
#C17-2는 국립칭화대학교와 TSMC가 공동으로 발표한 22nm Hybrid ReRAM–SRAM CIM 기반 AI-edge Processor로, Transformer와 CNN을 동시에 지원하는 비휘발성 CIM 아키텍처를 제안한 것이 특징이다. 기존 nvCIM은 CNN 가속에는 효과적이지만, Transformer에서는 QKV 연산과 Softmax 같은 동적 가중치 기반 연산 때문에 에너지 효율이 크게 저하되는 문제가 있었다. 이를 해결하기 위해 이 연구는 QKV-Softmax Layer-Fused (QKVS-LF) 구조를 도입하여 레이어 간 중간 데이터 저장을 최소화했고, Concurrent Transpose/Non-Transpose SRAM-CIM (LT-CTNT)으로 전치·비전치 연산을 동시에 수행해 지연과 메모리 접근을 줄였다.



[그림 1] 제안한 hybrid nvProcessor의 전반적인 구조 및 QKVS LF의 동작 설명

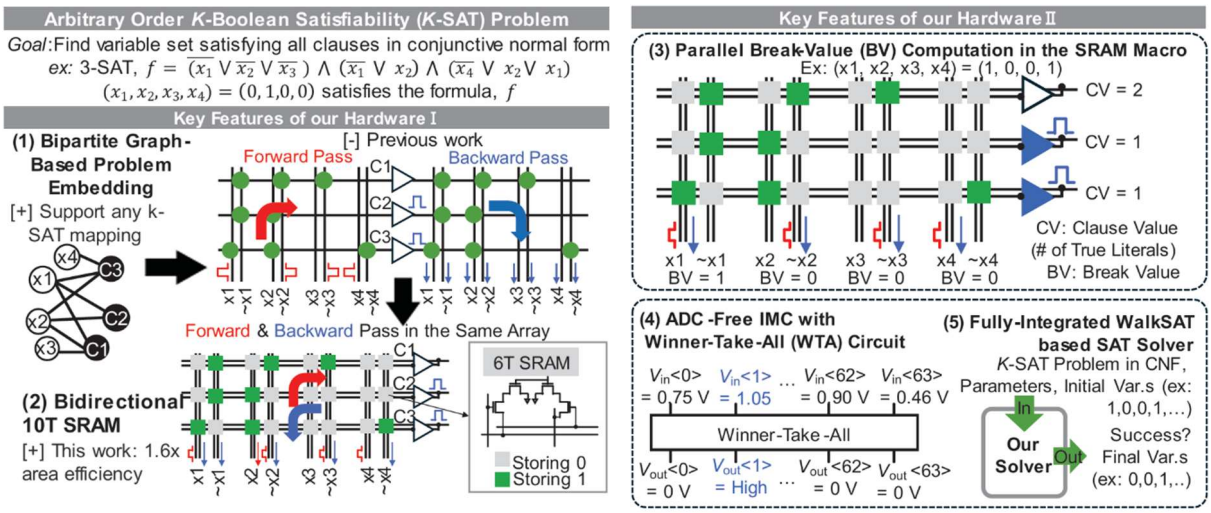
또한 ReRAM-CIM에는 Dynamic Input Sign Extension Encoding (DISEE) 기법을 적용해 입력 희소성을 높이고 불필요한 연산을 줄임으로써 nvCIM의 높은 에너지 소모를 완화하였다. 22nm 공정으로 제작된 칩은 MobileViT 및 MobileNet 벤치마크에서 시스템 수준

41.8 TFLOPS/W의 에너지 효율과 ImageNet 정확도 74.5%를 달성했으며, 기존 nvProcessor 대비 1.9배 높은 성능 지표 (FoM)를 기록하였다. 이 논문은 작은 용량의 엣지 디바이스에서 Transformer를 실용적으로 가속할 수 있는 nvCIM 설계 전략을 제시했다는 점에서 중요한 의미를 가진다.



[그림 2] 제안한 2b-intra LT-LCC 와 DISEE 설명

#C17-4는 UCSB와 1QBit이 공동으로 발표한 혼성 신호(Mixed-Signal) CIM 기반 SAT Solver로, 기존 3-SAT 전용 하드웨어를 넘어 임의 차수의 K-SAT 문제까지 직접적으로 해결할 수 있는 구조를 제시한 것이 가장 큰 특징이다.



[그림 1] K-SAT 문제 정의와 이를 해결하기 위해 제안한 work의 주요 feature들

핵심은 10T 양방향 SRAM 셀을 이용해 한 배열 내에서 절 (clause) 평가와 gradient/break-value 계산을 동시에 수행할 수 있도록 하였고, 이를 바탕으로 원샷(one-shot) 방식의 병렬 break-value 연산을 구현해 WalkSAT 알고리즘을 고속으로 실행할 수

있게 했다. 또한 아날로그 Winner-Take-All 회로를 도입해 promising 변수를 ADC 없이 선택 가능하도록 설계, 면적과 전력 효율을 높였다. 55nm 공정으로 제작된 칩은 최대 64 변수 문제까지 검증되었으며, 무작위 3-SAT 문제에서는 기존 ASIC 솔버 대비 약 10배, 4-SAT·5-SAT와 같은 고차수 문제에서는 최대 200배 빠른 연산 속도를 기록했다. 이 논문은 기존 디지털 혹은 3-SAT 전용 가속기의 한계를 넘어, 고차수 NP-hard 문제를 네이티브하게 처리할 수 있는 범용 SAT 하드웨어 솔버로 CIM을 확장했다는 점에서 중요한 의의를 가진다.

저자정보

박은빈 박사과정 대학원생

- 소속 : 포항공과대학교
- 연구분야 : 임베디드 시스템 및 지능형 반도체
- 이메일 : eunbin@postech.ac.kr
- 홈페이지 : https://sites.google.com/view/epiclab/member/ebpark

2025 IEEE VLSI Review

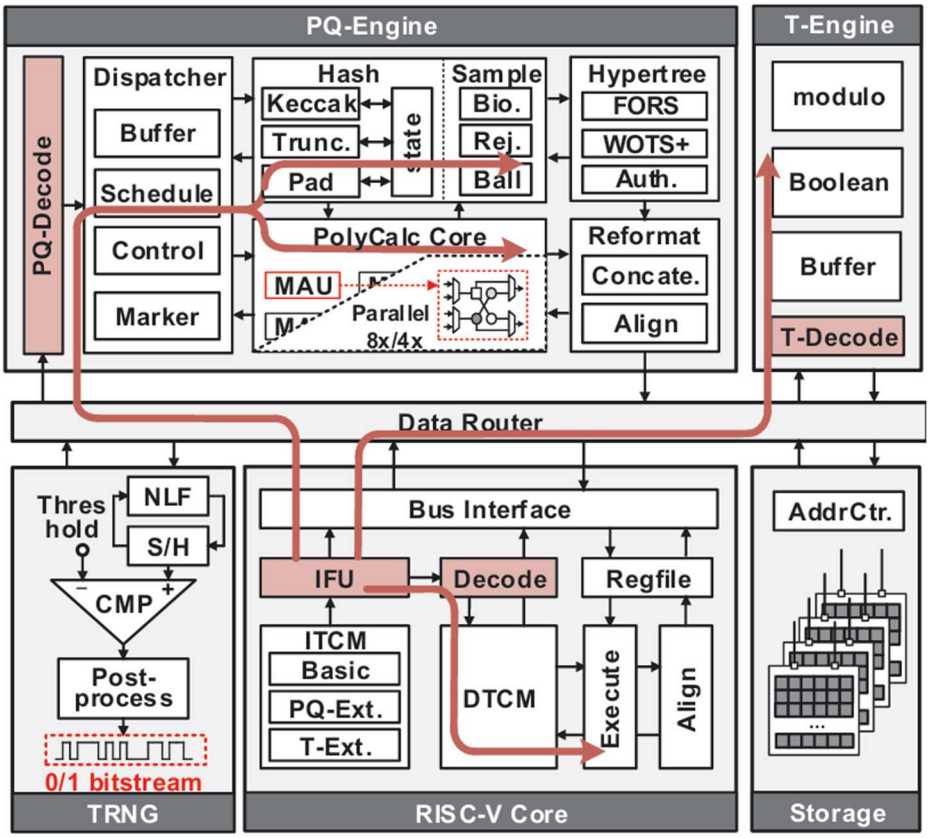
한국과학기술원 전기및전자공학부 석사과정 권재훈

Session C23 Innovatie Computing Systems

이번 2025 IEEE VLSI의 Session C23은 Innovatie Computing Systems라는 주제로 총 4편의 논문이 발표되었다. 이 세션에서는 양자암호용 SoC, multi-node system 보호용 TIME architecture가 제안되었고, privacy와 security용 하드웨어에 중점을 두었다.

#C23-3 본 논문은 NIST의 최신 Federal Information Processing Standards (FIPS)를 지원하는 최초의 RISC-V 기반 양자암호용 SoC를 제안한다. SoC의 특징은 크게 3가지가 있고, 첫 번째는 RISC-V core와 deep-coupling된 post-quantum cryptographic (PQC) engine을 두고, dual-rail parallel scheduling으로 hash와 polynomial 연산 같은, 종류가 다른 primitive를 동시에 처리하여 pipeline활용을 극대화한 것이다. 두 번째로 Vectorized instruction-driven workflow를 통해 RISC-V custom instruction이 primitive를 직접 구동하는 것이다. 세 번째는 modulo add/sub/mult, Number theoretic transform(NTT)등의 primitive 연산을 fine-grained operator로 구현하여, 필요에 따라 reconstruction함으로써 여러 PQC scheme을 하나의 하드웨어로 지원하는 것이다.

그림 1에 전체 System architecture의 Top-level diagram이 나와있으며, RISC-V core, PQC engine(PQ-engine), traditional cryptographic engine(T-engine), TRNG, memory blocks 등으로 구성된 것을 확인할 수 있다. System은 그림에 적색으로 표시된 3가지 instruction pipeline execution path를 갖는데, 1번째 path는 Core path로서 RISC-V core 자체 실행 경로이다. control flow, parameter 초기화, pre/post-processing을 담당하며 2개의 64 KB globally-shared TCM을 사용하고 일반 RISC-V instruction으로 수행된다. 2번째 path는 PQ-engine path로서 Post-Quantum 전용 경로이다. PQC primitive인 hash/Keccak(SHAKE), sampler, polynomial/NTT, hypertree 등을 고속으로 처리하는 path이며, SRAM과 TRNG를 사용하고, 일반 instruction이 아닌 RISC-V custom extension instructions를 통해 수행된다. 3번째 path는 T-engine path로서 Traditional cryptographic path이다. SM2/SM3 등의 traditional한 algorithm을 처리하는 역할을 하며, core에서 신호를 주면 core와 병렬로 실행되는 2번째 path와 달리 3번째 pipeline에서 독립적으로 실행된다.

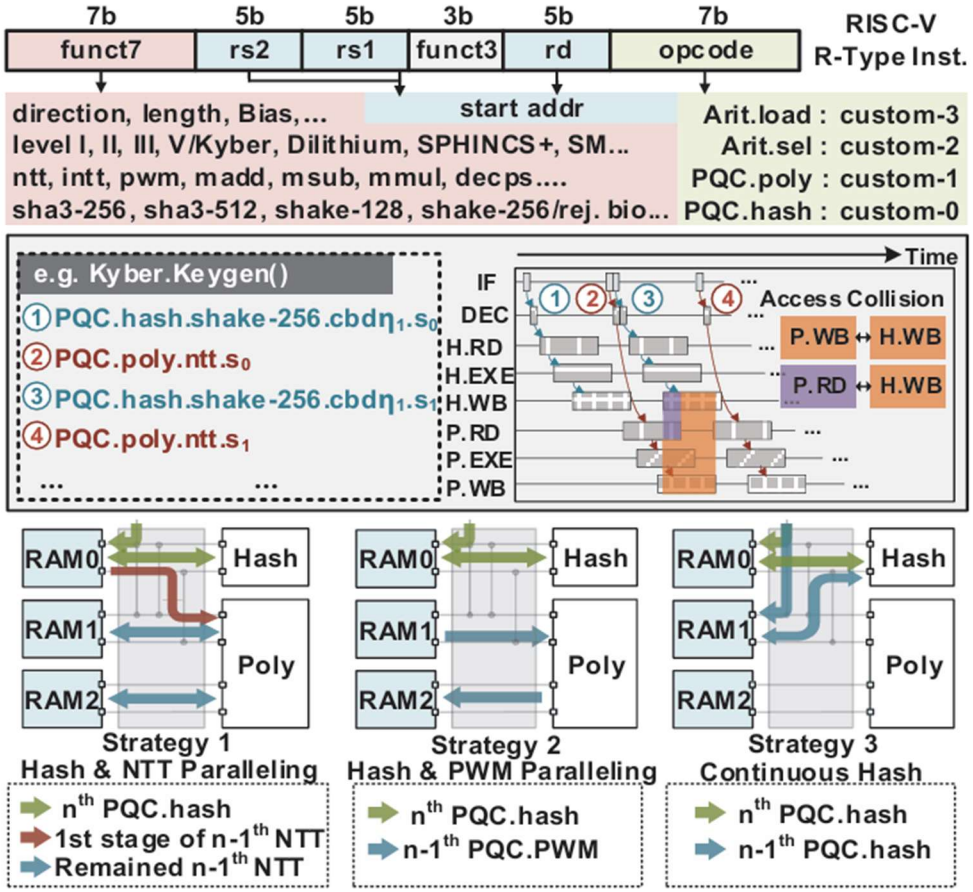


[그림 1] System architecture의 Top-level block diagram

SoC의 특징에 대해 더 자세히 설명하면, 일단 PQC algorithm은 여러 primitive로 분해될 수 있고, 이를 RISC-V vector extension instruction로 제어하는 것이 핵심이다. 이때 R-type custom instruction을 opcode expansion을 통해 확장하여 4가지 class로 분류하는데, 1번째로 algorithm, security level을 선택하여 scheme과 parameter set을 지정하는 역할을 하는 Arit.sel, 2번째로 core와 engine간 또는 memory와 register간의 data transmission을 담당하는 Arit.load, 3번째로 NTT/INTT, modular mul/add 등의 polynomial 계산을 하는 PQC.poly, 마지막으로 hash와 sampling을 담당하는 PQC.hash이다. 이렇게 분류한 instruction들을 rearrange하면 run-time 중에도 다양한 PQC scheme으로 유연하게 reconfiguration할 수 있다. 특히 SoC의 1번째 특징으로 언급한 deep-coupling은 consecutive하게 배치된 PQC.poly와 PQC.hash를 parallel하게 실행한다는 뜻이다. parallel processing을 지원할 때 주의할 점 중 하나는, 하나의 resource를 동시에 접근할 수 없다는 점인데, 본 논문에서는 SRAM bank 분리 등의 data routing strategy를 적용하여 PQC.poly와 PQC.hash간의 time-overlapping execution에서 발생할 수 있는 문제를 해결했다. 이는 그림 2에서 Strategy 1, 2, 3로 표시돼있고, memory access collision을 효과적으로 방지한다는 것을 쉽게 알 수 있다.

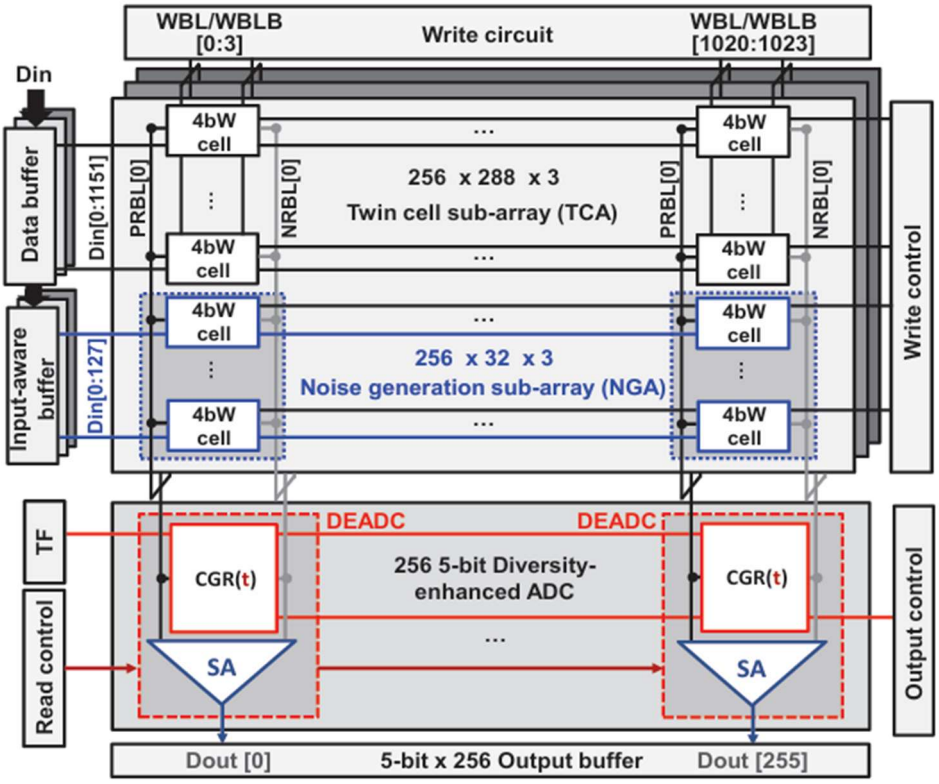
SoC는 3가지의 PQC scheme(Kyber, Dilithium, SPHINCS+)에 대해 throughput과 energy efficiency를 측정하는 것으로 평가되었고, Kyber-512에서 최고 throughput 84.9 KOPS, 에

너지 1.82 μJ/op를 달성하였다. SoC의 spec은 28nm 공정으로 만들었을 때, core area가 2.11 mm^2이었고, gate count로 환산하였을 때 1.5M개 이고, 0.65–1.1 V 범위에서 50~800 MHz로 동작할 수 있었다.



[그림 2] RISC-V의 extension instruction과 data routing strategy

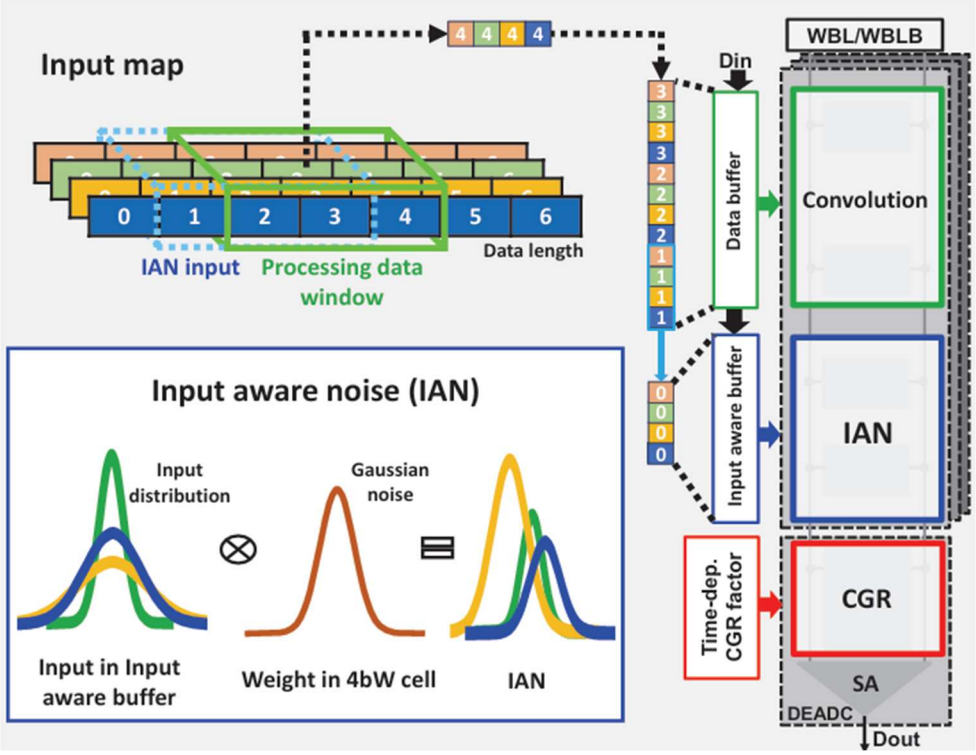
#C23-4 본 논문은 twin in-memory encryption/processing(TIME) macro를 설계해 Group Differential Privacy(GDP) + Spatial/Temporal Ensemble(STE) 를 하나의 memory macro에 통합하였다. 용어를 정리하면, GDP는 여러 node/data 간의 correlation을 고려하여 noise를 강하게 injection하는 DP의 변형된 형태이다. 이것은 AIoT 네트워크나 federated learning 등 여러 node가 있는 collective AI에서, node간의 correlation을 악용하는 공격으로부터 보호할 때 유용하다. 기존의 local differential privacy (LDP)는 single-node system에서는 유효하지만, 여러 node가 있는 system에서는 취약하므로 본 논문에서 GDP를 도입한 것이다. STE는 여러 model에서 추론한 결과인 Spatial ensemble과, 동일 입력에 대해 시간적으로 여러 번 추론한 결과인 Temporal emsemble을 majority voting을 통해 결합하여, encryption/noise로 인한 정확도 손실을 완화하는 ensemble이다. ensemble은 여러 decision making model의 출력을 결합하여 하나의 예측을 만드는 방법이다.



[그림 3] TIME의 전체 architecture

그림 3에 TIME의 architecture가 나와있고, 1Mb SRAM array, twin cell sub-array (TCA), noise generation sub-array (NGA), diversity-enhanced ADC (DEADC)와 peripheral circuit으로 구성된 것을 확인할 수 있다. TCA는 weight를 저장하면서 동시에 in-memory multiply를 수행하는 연산용 subarrya이다. 더 구체적으로는, 3개의 model fusion group으로 구성되어 있고, 각 group은 256×288개의 SRAM unit cell에 signed 4-bit weights(4bW) 를 저장하고, 각 4bW unit cell은 3×4 subthreshold-MOS transistor array로, 4-bit input(4bIN) 과 4bW의 multiply를 수행한다. NGA는 간단하게 GDP용 noise를 생성하는 역할을 하며, TCA와 동일한 unit cell을 쓰지만 256×32개의 4-bit random weights를 저장해 noise generation에 활용한다. DEADC는 bit-line current 누적, SA, comparator로 이어지는 in-memory ADC이다. 즉, 같은 bit-line에 흐르는 전류를 누적하고, current mirror-based sense amplifier(SA)의 adjustable current gain ratio(CGR)로 scailling한 뒤 comparator로 양자화하는 역할을 한다. CGR을 사용하는 이유는 dynamic range/robustness를 확보하기 위해서이다.

TIME의 architecture의 mode는 크게 GDP encryption mode, STE mode가 있고 둘은 consecutive한 pipeline 단계를 거친다. GDP encryption mode에서는 Gaussian noise를 크고 sparse하게 만들기 위해 두 가지 방법을 사용한다. 1번째로, Input-Aware Noise(IAN)로 NGA에 저장된 random weights와 input-aware buffer의 input 일부를 mult&add하여 추가 RNG 없이 Gaussian에 가까운 decorrelated noise를 생성한 것이다.



[그림 4] GDP encryption mode의 datapath

2번째로, time-controlled CGR로 DEADC의 CGR을 time축으로 바꿔가며 5-bit output encoding 중 noise 변동성을 더 키운 것이다. 두 방법을 결합하여 SA만 쓸 때보다 Gaussian noise distribution 범위를 약 120%늘렸다. 구체적인 GDP encryption mode의 datapath는 그림 4에 나와있다. STE mode는 정확도를 회복하기 위해 처리하고, 추론하는 mode이다. 3개의 TCA가 하나의 DEADC로 연결되어 Spatial ensemble을 구현하고, CGR을 같은 입력에 대해 여러 번 처리하여 Temporal ensemble을 구현했으며 이 결과를 voting과 averaging을 통해 결합했다. 마지막으로 TIME architecture의 전체적인 흐름은 다음과 같다. Input - GDP noise injection(IAN + CGR) - In-Memory 연산(TCA/DEADC) - STE(Spatial+Temporal) - Voting – Output

저자정보

권재훈 석사과정 대학원생

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