

2026 VLSI Review

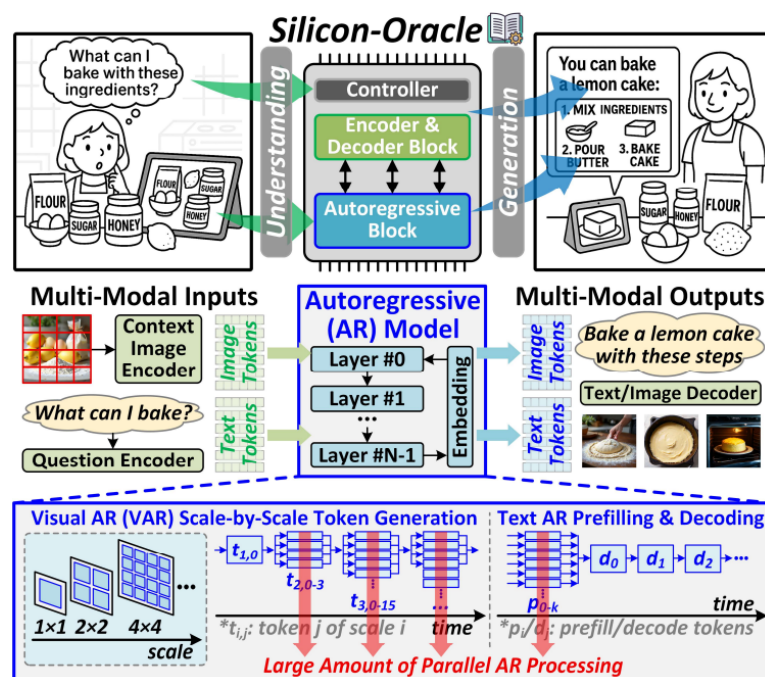
경북대학교 전자전기공학부 박사과정 박승현

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.

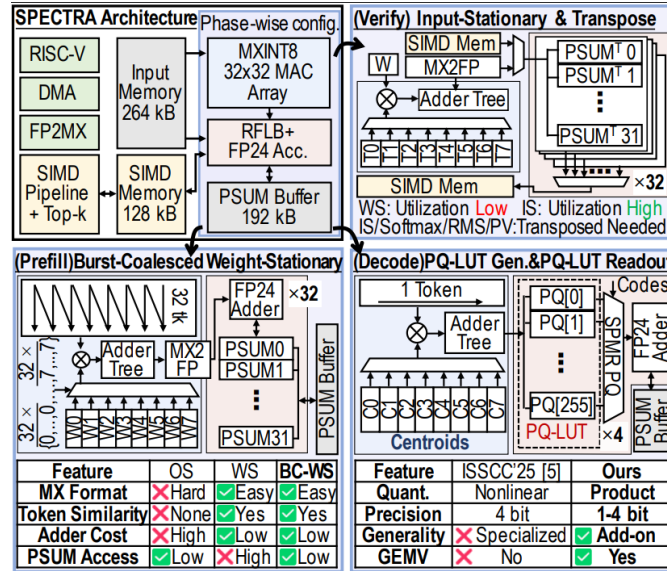
Soracle은 먼저 **Exponent-Tracked LUT-Counter MAC, ET LC-MAC**을 사용해 mixed-precision 연산 비용을 줄인다. 일반적인 multiplier 대신 LUT와 counter를 이용해 연산하고 exponent를 별도로 추적해 서로 다른 precision의 값을 accumulation한다. Canonical term decomposition과 equivalent-term coalescing을 이용해 counter activation도 줄인다. 측정 결과 conventional multiplier-based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

또 다른 핵심은 **row-broadcast/column-unicast(RBCU)** 구조이다. 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 log₂-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 mm² die area를 가지며, Infinity-125M을 이용한 image generation에서 **83.93 mJ/frame**과 **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.

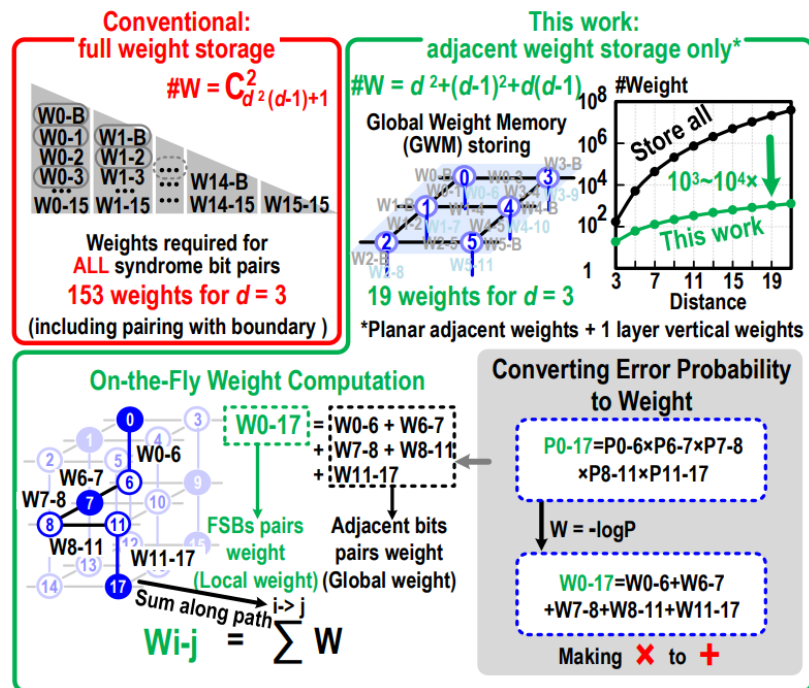
Hardware 측면에서 가장 중요한 부분은 **Reconfigurable Flip-Flop Local Buffer(RFLB)**이다. 동일한 RFLB가 prefill에서는 partial sum buffer, decode에서는 PQ lookup table, verify에서는 transpose buffer로 동작한다. Prefill에서는 burst-coalesced weight-stationary 방식으로 PSUM access를 8배 줄이고, clock gating과 input gating을 통해 RFLB 내부 power를 86% 줄여 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**, **verify**의 서로 다른 연산 특성에 맞추어 **precision**과 **dataflow**, **buffer** 역할을 모두 바꾼다는 점이 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 1.9×10^{-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 tuning**과 **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} LER**과 **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 방향으로 발전하고 있다는 점**이 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/>