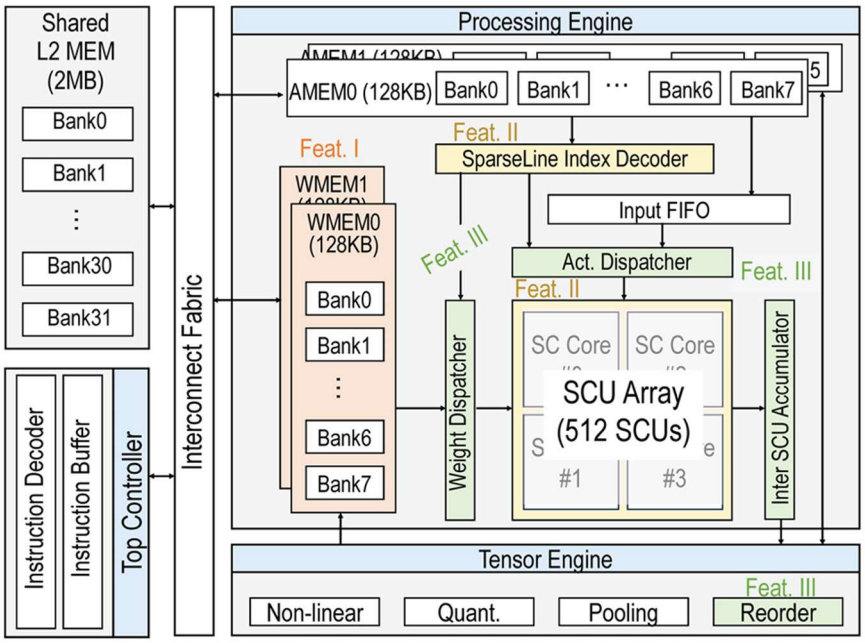
# 2025 IEEE VLSI Review

KAIST 전기및전자공학부 석사과정 박민하

## Session 29 Communication and Processors

세션 29는 'Communication and Processors'라는 주제로, 미래 통신 기술과 이를 지원하는 핵심 프로세서 연구들을 다루고 있다. 이 세션에서는 대규모 다중 사용자 MIMO(Massive MU-MIMO) 시스템의 데이터 처리 효율을 극대화하는 검출기(detector), 고속 광통신을 위한 에너지 효율적인 FEC(Forward Error Correction) 디코더, 그리고 디지털 신호 처리(DSP)에 특화된 고성능 RISC-V 벡터 프로세서 등 다양한 하드웨어 기술들이 발표되었다. 특히, AI 시대에 맞춰 전력 효율을 극대화한 NPU(Neural Processing Unit) 논문도 포함되어, 통신 기술의 발전과 더불어 이를 처리하는 프로세서 기술의 중요성을 보여주고 있다.

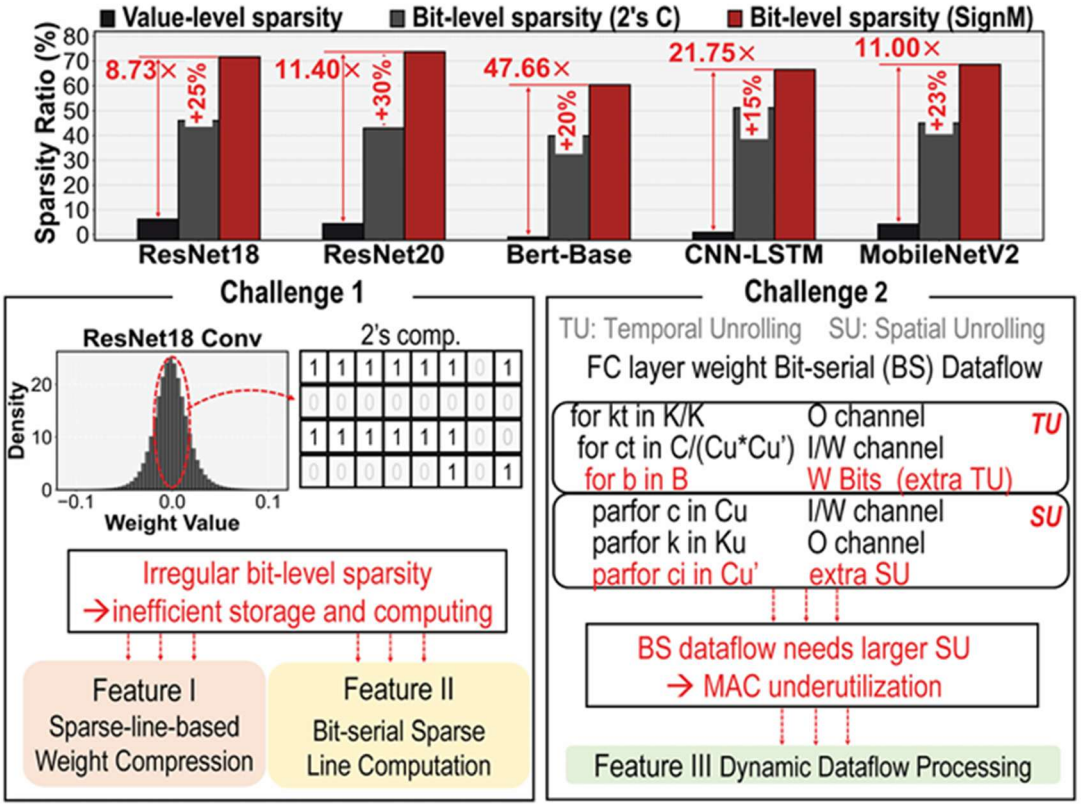
**#29-4** KU Leuven과 NXP Semiconductor가 공동 발표한 이 논문은 훈련 없이도 높은 에너지 효율을 달성하는 신경망 처리 장치(NPU)를 다루고 있다. 이 칩은 신경망 및 트랜스포머 워크로드에서 공격적인 가중치 압축과 높은 연산 활용도를 동시에 달성하기 위해 구조화된 비트 단위 희소성(structured bit-level sparsity)과 동적 데이터 흐름(dynamic dataflows)을 활용하는 희소 신경망 프로세서이다.




[그림 1] 테스트 칩의 아키텍처 (3가지 특징)

위의 그림은 논문이 제안하는 NPU(Neural Processing Unit)의 전체적인 하드웨어 아키텍처를 보여준다. 단순한 블록 다이어그램을 넘어, 논문의 핵심적인 세 가지 기술적 특징인 훈련 없는 비트 단위 희소성(training-free structured bit-level sparsity), 동적 데이터 흐름(dynamic dataflows), 그리고 이를 통합하는 명령어 기반의 프로세서(instruction-programmable processor)가 어떻게 유기적으로 연결되어 있는지를 알 수 있다. 특히, 데이터가 메모리(L2, L1)에서 나와 동적 디스패처를 거쳐 512개의 SCU(Sparse Line Computation Unit) 배열로 전달되는 흐름을 보여줌으로써, 효율적인 데이터 이동과 연산이 이 칩의 핵심임을 알 수 있다.

이 논문은 AI 하드웨어 설계의 두 가지 핵심 난제를 해결하려는 시도로 볼 수 있다. 첫째, AI 모델의 비대화에 따른 연산량 및 메모리 요구량 증가, 둘째, 이를 효율적으로 처리해야 하는 전력 효율성 문제이다. 특히, 모바일 및 엣지 디바이스와 같이 전력 제약이 심한 환경에서는 이 두 가지 과제를 동시에 해결하는 것이 필수적이다. 저자들은 이 문제를 해결하기 위해 가중치의 비트 단위 희소성이라는 새로운 관점에 주목했다.




[그림 2] 신경망 프로세서에서 가중치-비트 희소성(weight-bit sparsity)을 활용하기 위한 핵심 과제

전통적인 신경망 희소화(sparsity) 기법은 주로 가중치의 '값(value)'이 0에 가까운 것을 제거하는 방식이었으며, 전통적인 신경망 희소화(sparsity) 기법은 주로 가중치의 '값'이 0에 가까운 것을 제거하는 방식으로, 이는 모델의 크기를 줄이는 효과가 있었다. 그러나 이를 위해서는 모델을 희소성에 맞게 re-training해야 하는 복잡한 과정을 거쳐야 했다.

이 논문은 이러한 한계를 극복하기 위해 비트 단위 희소성(Bit-level Sparsity, BLS)이라는 혁신적인 접근법을 제안한다. 즉, BLS는 가중치를 이진수로 표현했을 때, '0' 비트가 많은 현상을 이용하는 것이다. 특히 Sign-Magnitude(SignM) 인코딩을 사용하면 '0' 비트가 더 많이 생성되는데, 이 논문은 이 '0' 비트가 있는 연산을 건너뛰는 방식으로 하드웨어 효율을 극대화한다.

첫 번째는 메모리 효율적 비트라인 압축 기술이다. 메모리 효율적 비트라인 압축: 가중치를 여러 그룹으로 묶고, 각 그룹 내에서 동일한 유효성을 가진 비트 라인을 식별한다. 이 중 '0' 비트로만 이루어진 라인(Zero-lines)은 제거하고, 비제로 라인(Non-Zero-Lines, NZLs)만 인코딩하여 저장한다. 이를 통해 메모리 사용량을 획기적으로 줄일 수 있다.

두 번째는 Zero-bit Line Boost (ZLBoost) 기술이다. 이는 가중치의 '0' 비트 위치를 유연하게 조절하는 기술로, 기존의 양자화 방법들이 비제로 비트가 연속적으로 위치해야 한다는 제약을 가졌던 것과 달리, 손실이 있는 SignM 인코딩을 통해 더 높은 압축률을 달성한다. 예를 들어, BERT-Base 모델에서 ZLBoost는 0.5% 미만의 정확도 손실로 2.47배의 가중치 압축을 달성했다. 이는 기존 4-bit 양자화 기법(PTQ)이 약 10%의 정확도 손실을 보였던 것과 비교해 압도적인 성능이다. 이 기술은 재훈련 없이도 높은 압축률과 정확도를 동시에 보장함으로써, AI 모델 경량화의 새로운 가능성을 열었다.

| | DA-SignM ISSCC'23 [2] | USCA ISSCC'23 [12] | HUAA ISSCC'23 [13] | Onyx VLSI'24 [14] | This work |
|---|---|---|---|---|---|
| **Technology** | 28nm | 28nm | 28nm | 12nm | **16nm** |
| **Area [mm$^2$]** | 7.75 | 2.69 | 7.81 | 23 | **6.5** |
| **Frequency [MHz]** | 55-285 | 60-400 | 100-500 | 500-980 | **50-280** |
| **Voltage [V]** | 0.65-0.9 | 0.48-0.92 | 0.66-1.3 | 0.6-1.0 | **0.565-0.93** |
| **Number of MACs** | 1024 | 256 | 1024 | 1536 | **512** |
| **On-Chip Mem. [KB]** | 1074 | 176 | 1120 | 4500 | **2560** |
| **Sparsity Support** | W bit | A value | None | W/A value | **W bit** |
| **Flexible Dataflow** | None | None | Yes | None | **Yes** |
| **Power [mW]** | 6.6-179.4 | 366 | 17-174 | - | **6.13-102.4** |
| **Peak Performance [BTOPS]*1** | 27.97@0.9V<br>5.38@0.65V | - | - | 73.09 | **54.72@0.93V*2**<br>**11.66@0.6V*3** |
| **Effective Peak EE [E-BTOPS/W]*1** | - | 99.2@0.85V<br>264.9@0.48V | 480-716.8 | - | **190.71@0.93V*4**<br>**645.12@0.6V*5** |
| **Sparse Peak EE [S-BTOPS/W]*1** | 268.8@0.9V<br>517.7@0.65V | 179.2@0.85V<br>767.3@0.48V | - | 193.5@0.66V, 500MHz *6 | **550.4@0.93V*2**<br>**1320.8@0.6V*3** |
| **Area Efficiency [BTOPS/mm$^2$]*1** | 3.61@0.9V<br>0.69@0.65V | - | - | 1.59 | **8.41@0.93V*2**<br>**1.79@0.6V*3** |

*1: One operation (OP) is one mult. or one add. *2: Measured at the highest performance point. 0.93V for logic, 0.92V for mem., 280MHz, 6 sparse lines. Activation with 50% sparsity.
*3: Measured at the highest efficiency point 0.6V for logic, 0.55V for mem., 71.4MHz, 6 ZLs.
*4 *5: Peak EE excluding skipped sparse computations w/o ZLBoost. *6: Reported TOPS/W was not clearly declared to E-BTOPS/W or S-BTOPS/W, used as S-BTOPS/W as sparsity is exploited.

[그림 3] 최신 기술 비교 표

이 표는 논문에서 제안하는 NPU의 성능이 기존의 최신 기술과 비교했을 때 어느 정도 수준인지 객관적으로 보여주는 자료이다. 공정 기술, 칩 면적, 연산 능력, 전력 소모, 그리고 에너지 효율과 같은 핵심 지표들을 다른 주요 NPU들과 비교하였다. 특히 S-BTOPS/W(Sparse BTOPS/W)라는 희소 연산 효율 지표에서 압도적인 우위를 보여주며, 논문이 주장하는 비트 단위 희소성 기술이 기존의 어떤 기술보다도 효과적임을 입증한다.

희소성 기술을 적용하면 연산량이 불규칙해지므로, 하드웨어 연산기의 활용도를 유지하기 어렵다는 문제가 발생한다. 이 논문은 이를 해결하기 위해 동적 데이터 흐름을 적용한다. 512개의 SCU(Sparse Line Computation Unit)로 구성된 연산 배열은 특정 작업에 고정된 데이터 흐름을 사용하는 것이 아니라, 현재 처리하는 레이어의 특성(차원)에 맞춰 최적의 데이터 이동 경로를 선택한다.

이 논문은 단순히 AI 하드웨어의 성능을 개선하는 것을 넘어, 새로운 설계 패러다임을 제시했다는 점에서 매우 중요한 의미를 가진다. '훈련 없이 비트 단위 희소성을 활용'하는 방식은 기존의 희소화 연구와 차별화되는 고유한 아이디어이며, 이는 하드웨어 설계 단계에서 효율을 극대화하면서도 소프트웨어적인 복잡성(재훈련)을 제거하여, 실제 상용화에 유리한 이점을 제공한다.

또한 550~1320 BTOPS/W라는 높은 에너지 효율은 모바일, 엣지 컴퓨팅, 사물 인터넷(IoT) 등 전력 제약이 심한 환경에서 AI를 구동하는 데 필수적인 성능이다. 이 칩의 효율은 기존 최고 기술 대비 최대 6.82배나 높아, 차세대 AI 가속기 시장에서 강력한 경쟁력을 가질 수 있음을 시사한다.

그리고, CNN, RNN, 트랜스포머 등 다양한 AI 모델을 효과적으로 처리할 수 있는 동적 데이터 흐름 지원은 이 칩이 특정 애플리케이션에 국한되지 않고 광범위하게 활용될 수 있음을 보여준다. 이는 현재의 AI 기술이 특정 모델에 의존하기보다 다양한 모델이 공존하는 방향으로 나아가고 있다는 점을 고려할 때 매우 중요한 특징이다.

종합적으로 볼 때, 훈련 없는 희소성, 동적 데이터 흐름과 같은 혁신적인 아이디어를 성공적으로 구현함으로써, AI 하드웨어 연구의 새로운 방향을 제시하고 고성능-고효율 AI 시스템 개발에 기여를 하였다.

## 저자정보


**박민하 석사과정 대학원생**

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