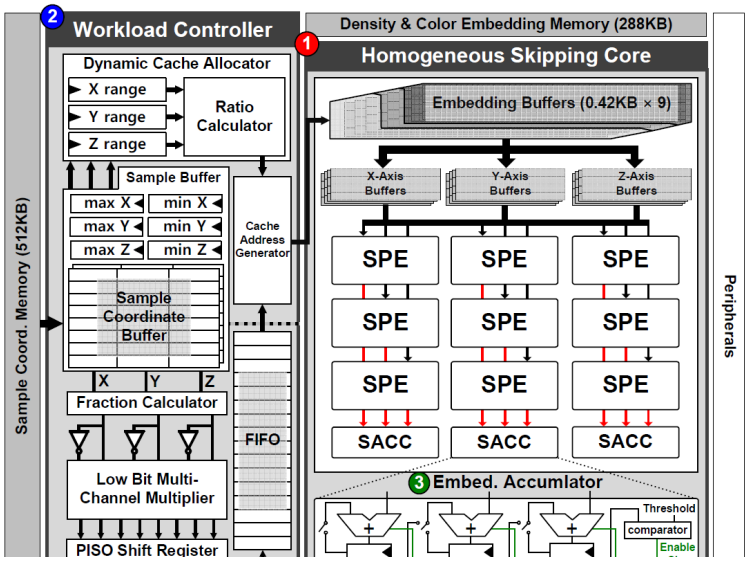
# 2023 IEEE ASSCC Review

KAIST 전기및전자공학부 박사과정 엄소연

## Session 13 AI on FPGA

이번 2023 IEEE ASSCC의 Session 13은 AI on FPGA라는 주제로 총 4편의 논문이 발표되었다. 이 세션에서는 neural rendering, sparse matrix-vector multiplication, transformer, 그리고 super-resolution를 위한 FPGA기반의 가속기가 발표되었다. 이번 후기를 통해 4개의 논문에 대해 간략하게 살펴보고자 한다.

**#13.1**은 KAIST에서 발표한 FPGA 를 이용해 구현한 Neural Rendering Accelerator로, Interpolation 기반의 Explicit Neural Radiance Field (NeRF) 알고리즘을 효율적으로 가속시킬 수 있는 하드웨어 구조를 소개한다. 본 논문은 Explicit NeRF 방식의 계산 부하를 줄이기 위해, Pruning Aware Skipping (PAS), Sample Centric Workload Balancing (SCWB), 그리고 Embedding Accumulation Early Stopping (EAES)의 세 가지 기법을 제안하여 Explicit NeRF 기반의 실시간 렌더링를 가능케 하였다. 렌더링을 위한 색상과 밀도 연산을 모두 지원하는 Switchable PAS Core (SPASC)와 SCWB를 이용한 workload balancer를 이용해 두 연산 사이의 workload를 효율적으로 분배하여 처리량과 에너지 효율을 높인 것이 특징이다. 이를 통해 기존의 논문들보다 최대 75배 에너지 효율적이며 1,680배 빠

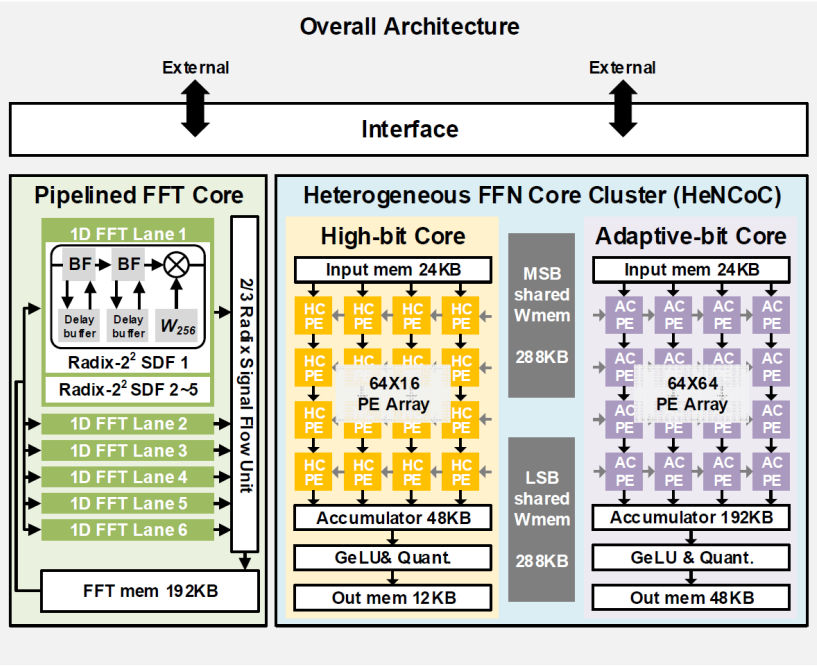



[그림 1] #13.1에서 제안한 Real-time Neural Rendering 가속기 구조

른 렌더링 속도를 달성하였으며, FPGA로 구현한 데모를 통해서 엣지 디바이스에 활용이 가능함을 보였다.

**#13.2**은 이화여대에서 발표한 논문으로, DRAM 대역폭 확장이 가능한 sparse matrix-vector multiplication (SpMV) FPGA processor를 소개한다. SpMV 연산의 메모리 집약적 특성과 불규칙한 메모리 접근 패턴, 압축 형식에서 비롯된 제한된 데이터 재사용으로 인해 성능 향상을 달성하기 어렵다. 이 연구에서는 DRAM 대역폭과 확장 가능한 SpMV 가속기를 제시하여, 큰 희소 행렬을 처리할 때 성능 저하를 피하면서 최대 오프 칩 메모리 대역폭을 효율적으로 활용한다. 제안된 SpMV 가속기는 데이터 분배기, 다중 처리 요소(PE) 라인, 출력 업데이터를 포함하는 구조를 가집니다. 오프라인 전 처리 단계에서 행렬 분할 및 블록 내 재배열이 이루어진다. 이는 bank 충돌과 MAC 단위의 동시 접근 문제를 해결하는데 중요합니다. 최종적으로, 제안된 SpMV 가속기는 평균 89%의 대역폭 활용 효율성을 달성하며, 이전 연구 대비 최대 43.8%의 성능 향상을 보여준다.

**#13.3**은 KAIST에서 발표한 논문으로, Fourier transform의 특성을 이용하여 transformer를 가속하는 프로세서를 소개한다. 본 논문은 self-attention을 Fourier transform으로 대체한 특별한 transformer를 가속하며, 높은 계산 비중을 가지는 FFN 부분을 가속하기 위해 frequency 별로 다른 bit-precision을 사용하는 것을 제안하였다. 또한, 고정된 high-bit workload 비중은 DSP based 2D PE array, low-bit, high-bit 변화하는 workload는 adaptive LUT 2D PE array를 사용하는 것을 통해 연산 과정 중 utilization을 최대로 유지하고, FPGA의 모든 연산기를 활용하였다. 이를 통해 기존 State-Of-The-Art에 비해 0.3배의 전력을




[그림 2] #13.3에서 제안한 Fourier Transform의 특성을 이용한 가속기 구조

소비하고, 한번의 inference에서 0.82배의 에너지만을 소비하는 등, 저전력 모바일/엣지 디바이스에서 transformer의 활용 가능함을 선보인 것이 인상적이다.

**#13.4**는 KAIST에서 발표한 논문으로, super-resolution을 위한 heterogeneous CNN/SNN core architecture를 갖는 FPGA processor를 소개한다. 본 논문은 기존 FPGA 기반 CNN processor와 SNN processor가 super-resolution을 처리하는데 있어서 가지는 제한점을 해결하기 위해 DSP-dominant한 특성을 가지는 CNN core와 LUT-dominant한 특성을 가지는 SNN core를 heterogeneous하게 사용하였고, 이를 통해 FPGA의 resource 활용을 극대화하였다. 또한, Bit-precision 기반 workload allocation과 operand-switchable SNN core를 함께 활용하여 CNN core와 SNN core간의 workload imbalancing 문제를 해결함으로써 throughput을 극대화한 것이 특징이다. 이를 통해 기존 FPGA 기반 CNN processor에 비해 3.55배 빠른 처리 속도를 달성하였으며, FPGA에서의 CNN과 SNN의 융합을 성공적으로 구현한 것이 인상적이다.

**저자정보**


**엄소연 박사과정 대학원생**

- 소속 : KAIST 전기및전자공학부
- 연구분야 : Computing-In-Memory Processor
- 이메일 : soyeon.um@kaist.ac.kr
- 홈페이지 : https://ssl.kaist.ac.kr/