2023 International Solid-State Circuits Conference

(ISSCC) Review

UNIST 전기전자공학과 이규호 교수

Topic : Digital Circuits and ML

Session 2 : Digital Processors
Session 22 : Heterogeneous ML Accelerator

Digital Processors **(Session 2)**에서는 CPU, Gaming, Genome Analysis, Robots, Video Application 등 7편의 디지털 회로가 소개되었으며, Heterogeneous ML Accelerator **(Session 22)**에서는 다양한 애플리케이션에 적용된 머신 러닝 프로세서 9편이 발표되었다. 이 중, 최신 ML 알고리즘인 3D 렌더링을 위한 Neural Radiance Fields 프로세서와 대형 언어 모델의 Transformer 프로세서의 높은 전력효율이 돋보였다. 이번 후기를 통해 두 세션 중 총 11개의 논문에 대해 간략하게 살펴보고자 한다.

#2.1은 AMD에서 발표한 Zen4 아키텍처이다. 지난 Zen3는 동일한 7nm 공정에서 최적화가 이루어졌다면, 올해에는 5nm 공정 스케일링을 통해 성능을 개선하였다. 기존 Zen 시리즈는 L3 캐시를 2배씩 증가시킨 반면 Zen4에서는 L2 캐시를 2배 증가시켜 (1MB) 집적하고 로직을 개선하여 Cache Conflict를 감소시켰다. 그 외에도 Branch Prediction, Execution Engine 등 아키텍처 개선을 통해서 전작 대비 동일 전력에서 최대 34%의 성능 향상을 달성하였다. 또한, 256b Operation으로 구현된 AVX-512 연산을 지원하여 Zen3 대비 ML 연산 처리량을 3배~4.2배 향상시켰다.

#2.2는 MediaTek에서 발표한 4nm FinFET 공정의 고성능 5G 모바일 게이밍 SoC이다. 해당 SoC는 3.35 GHz High-Performance Core 1개, 3.2 GHz Balance-Performance Core 3개, 1.8 GHz High-Efficiency Core 4개를 집적한 Octa-Core CPU로 이루어져 있으며, 955MHz Deca-Core GPU를 집적하여 높은 성능을 제공한다. 온도가 상승해도 안정적인 성능을 유지하도록 발열제어를 위해 ISSCC'22에서 제안되었던 Octa-Core Tri-Gear DVFS CPU에 Energy/Temperature-Aware Scheduling을 통한 Task Reallocation, Throttling 등의 열 관리 시스템을 추가하여 발전시킨 것이 특징이다.

#2.5는 National Taiwan University에서 발표한 논문으로 Autonomous Mobile Robot 동작 제어를 위한 SoC이다. 해당 논문은 7-DoF 로봇 팔의 130 궤적 Time Steps에 대해 최대 4.935kHz, 750 궤적 Time Steps에 대해서는 1kHz 제어 속도를 지원한다. 제안된 궤도 최적화 가속기는 4x4 배열 구성의 Processing Elements (PEs)를 지니며, 각 PE는 Trajectory Pruner (TP)를 통해 cost가 높은 궤적에 대한 연산을 감소시킨다. 궤적 프루닝 차이로 발생하는 PE 간 연산 불균형을 각 행/열에 대한 Arbiter와 Trajectory Buffer로 구성된 NoC를 통해서 Workload를 재분배한다. 제안된 논문은 다른 로봇 동작 제어 논문과 비교하여 66배의 면적 효율, 350배의 효율 상승을 보여준다.

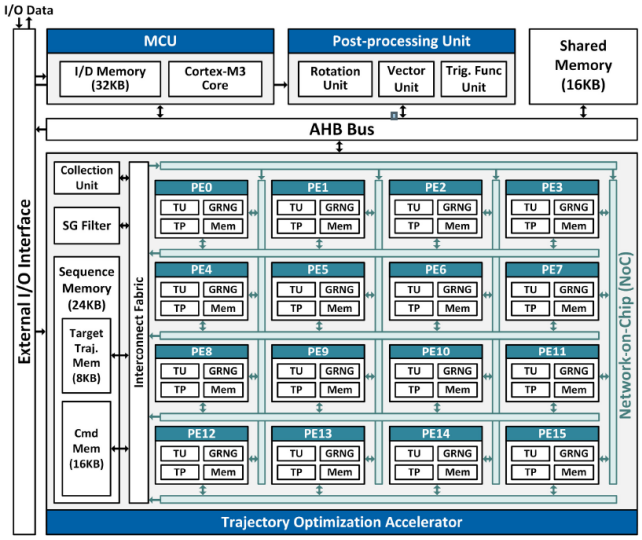

그림1. PE 배열의 모션 제어 SoC 구조

#2.6은 National Tsing Hua University에서 제안한 Video CNN 가속기로 Video Data Super Resolution (VSR)과 Video Frame Interpolation (VFI)을 수행한다. Video/Image Generation 네트워크는 큰 Feature Map 사이즈로 인해 많은 양의 External Memory Access (EMA) 와 높은 연산 복잡도를 수반한다. 이를 완화하기 위해 기존 2차원 단일 이미지를 이용한 CNN 가속기에서 사용했던 Depth-First Layer Fusion을 발전시켜 Depth-Time-Block 순으로 Layer Fusion 하는 Cuboid-Based Layer Fusion을 제안하여 33-53%의 EMA를 감소하고, 19-42%의 연산 복잡도를 완화하였다. 또한 연속된 프레임 처리 시 Reference Frame을 먼저 처리하여 On-Chip Storage를 25% 감소시켰다. 또한 Deformable Convolution 연산 시 타일 단위로 Offset Field에 Confinement를 두어 필요한 Line Buffer의 크기를 획기적으로 줄임으로써 VSR에 대해서는 50fps, VFI에 대해서는 38fps의 실시간 퍼포먼스를 달성하였다.

#2.7은 KAIST에서 발표한 Neural Radiance Fields (NeRF) 프로세서로서 구두 발표와 데모 세션에서 많은 이목을 끌었다. 본 논문은 NeRF의 DNN 계산 부하를 줄이기 위해, Sampling (Spatial Attention)과 Reusing result (Temporal Familiarity), DNN Skipping (Top-Down Attention)의 3단계 기법을 제안하여 소프트웨어-아키텍처-마이크로아키텍처 전 레벨에서 획기적인 아이디어를 제시하였다. 1D-2D Neural Engine 구조와 CS-DNNA 코어를 통해 Workload를 예측하고 효율적으로 분배하여 처리량과 에너지 효율을 높인 것이 특징이다. 이를 통해 기존 GPU 대비 911배 빠른 처리 속도를 달성하였으며, 작은 시스템 보드로 구현한 데모를 통해서 저전력 모바일/엣지 디바이스에서 활용 가능함을 선보인 것이 인상적이다.

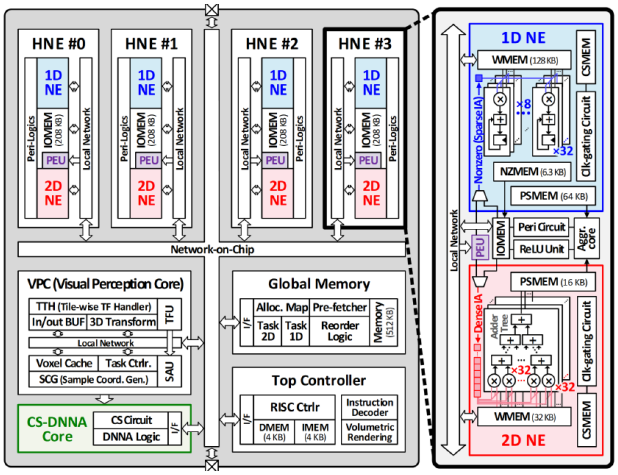

그림 2. Neural Radiance Field 가속기 구조

#22.1은 University of Bologna에서 발표한 All-Digital AI-IoT Accelerator이다. 목표 애플리케이션에 따라 요구되는 성능과 에너지 소모량을 유동적으로 지원하기 위해 Reconfigurable Binary Engine과 Adaptive Body Bias Generator를 사용한 것이 특징이다. 본 논문은 Reconfigurable Binary Engine에 의해 Mixed-Precision 연산을 지원함으로써 Weight와 Input의 Bit Precision을 조합하여 넓은 범위의 연산 처리량을 지원하고 있다. 또한, 다양한 Workload에 의해 발생하는 타이밍 문제를 On-Chip 모니터링을 통해 관측하고, 동적 Adaptive Body Bias Generator가 적합한 VBB 값으로 조절함으로써 성능을 향상시켰다. Mixed-Signal AIMC 기반의 프로세서보다 소비 전력과 성능이 뛰어나지는 않지만 지금까지의 디지털 AI-IoT Accelerator과 비교했을 때 가장 높은 범용성을 지원하는 것이 장점이다.

#22.2는 Tsinghua University에서 발표한 Large-scale Point Cloud 데이터를 활용한 2D/3D Unified Accelerator이다. Point Cloud 데이터는 불규칙한 EMA와 불균등한 Workload 등 데이터 자체의 특성으로 인해 발생하는 다양한 문제점들이 존재한다. 해당 논문에서는 불규칙한 EMA를 해결하기 위해 Voxel 포인트를 Block-Partitioning 하여 메모리에 저장하였으나, 이로 인해 불균등한 Workload가 발생된다. 따라서, 레이어 별로 입출력 채널수가 고정된 특징을 활용한 Asynchronous/Synchronous Hybrid Schedular를 제안함으로써 EMA Overhead를 52%가량 낮추었다. 또한, Priority-Code-Based Neighbor Search 방식을 통해 Sparse Convolution 연산과 병렬적으로 처리함으로써 연산 시간을 44%가량 줄였다는 장점이 있다.

#22.3은 POSTECH에서 제안한 Scalable Bit Precision과 다양한 Quantization Method를 지원하는 하드웨어 가속기이다. 해당 논문에서는 Arbitrary Quantization과 Arbitrary Basis를 각각 LUT Reconfiguration과 Bit-Serial Processing을 통해 연산하여 로직의 낭비를 줄인다. 본 연구에서는 메모리에서 0의 값이 포함된 Raw Data를 Zero Eliminator 모듈을 통해 압축시킨다. 또한 Runtime Density Detector를 통해서 임계 값에 따라 높은 Sparsity 데이터는 RLC Data Format으로 저장하고, 낮은 Sparsity 데이터는 Raw Data Format으로 Load 한다. Data Sparsity를 고려함으로 높은 Sparsity에 대해서는 12.7배 높은 퍼포먼스를 달성하며 높은 PE Utilization을 달성할 수 있다. 해당 연구는 Sparsity를 고려하며 Weight와 Activation에 대해 다른 Quantization Method를 적용하여 연산하며 Multi-Scale Precision을 지원한다는 점이 인상적이었다.

#22.5는 KAIST에서 발표한 논문으로, Heterogeneous CNN/SNN Core Architecture를 갖는 Complementary-Deep-Neural-Network (C-DNN) 프로세서를 소개한다. 본 논문은 저전력 네트워크인 SNN과 높은 이미지 분류 능력을 지니는 CNN을 융합하였다. 제안된 C-DNN 프로세서는 SNN과 CNN의 강점만을 살려서 저전력 추론을 위해 SNN 혹은 CNN 코어로 Workload를 분배한다. 또한, 에너지 효율적인 훈련을 위해 SNN을 활용하여 Forward Gradient Sparsity를 증가시켜 CNN의 Back Propagation 동작을 줄였다. 제안된 프로세서는 State-Of-The-Art 논문에 비해 0.28 ~ 0.39배의 전력을 소비하며 71.2% (ResNet-18) ~ 77.1% (ResNet-50)의 높은 이미지 분류 정확도를 지니는 등, 저전력 동작을 위한 SNN과 CNN의 융합이 돋보인다.

#22.8은 서울대학교에서 발표한 저전력 및 소형 Speech Enhancement 프로세서이다. 해당 논문에서는 사람의 음성이 주로 저주파수 영역에 있다는 점에 기인하여, Speech Enhancement 네트워크의 Gated Recurrent Unit에서 지배적인 주파수 영역의 채널을 단계적으로 확장함으로써 1.8%만의 정확도 손실로 29.7%의 연산량을 감소시켰다. 네트워크의 Depthwise/Pointwise Convolution의 순차적인 연산을 위해 Partial Sum을 저장하는 Intermediate Buffer를 사용함으로써 Unified Buffer만을 두는 것에 비해 5.34배 낮은 Memory Access Energy를 달성하였다. 또한 총 4개의 PE Array 중 3개를 CORDIC 알고리즘과 Neural Network 모두를 가속할 수 있는 구조로 제작하여 상호 간의 연결을 통해 연산함으로써 이종 코어를 사용했을 때와 대비하여 21.5%의 면적을 감소시켰다. 해당 연구는 귀에 탈부착할 수 있도록 0.81mm2의 작은 칩 사이즈와 0.74mW의 낮은 전력소모를 보인다는 점이 매우 돋보인다.

#22.9는 Harvard University에서 발표한 Transformer 프로세서이다. 본 논문은 엣지 플랫폼에서 Transformer 모델을 적용하기 위해 Entropy 기반의 알고리즘부터 회로까지 공동 설계에 대한 아이디어를 제시하였다. Transformer 계층마다 Entropy 기반 조기종료 메커니즘을 적용하여 연산을 제거함으로써 Latency를 증가시켰다. 또한, Entropy 정보를 기반으로 연산의 정밀도와 Dynamic Voltage/Frequency Scaling 기법을 예측/적용하여 처리량과 에너지 소비량을 개선하였다. FP4/FP8의 혼합 정밀도를 지원하는 데이터 패스의 연산 유닛을 제안하여 예측된 정밀도에 따른 연산을 지원하였다. 이러한 복합적인 설계를 통해 Entropy-Controlled Prevision Selection과 FP4 Per-Vector Scaling을 적용하여 단지 1.5%의 정확도 손실에도 불구하고 6배 빠른 Latency를 달성하며, 엣지 디바이스에서도 에너지 효율적인 Transformer 구현이 가능함을 보인 것이 돋보인다.

저자정보

이규호 교수

- 소　　속 : UNIST 전기전자공학과 / 인공지능대학원
- 연구분야 : Machine Learning Processor
- 이 메 일 : kyuho.jsn.lee@unist.ac.kr
- 홈페이지 : https://isl.unist.ac.kr/