2024 International Solid-State Circuits Conference

(ISSCC) Review

UNIST 전기전자공학과/인공지능대학원 이규호 교수

#ML Accelerators and Compute-in-Memory

Session 20 / Machine Learning Accelerators

Session 34 / Compute-In-Memory

Machine Learning Accelerators (**Session 20**)에서는 다양한 애플리케이션에 적용된 8편의 머신 러닝 프로세서가 소개되었다. 이 중, 매우 높은 전력효율을 보인 LiDAR를 사용한 실시간 Semantic SLAM을 위한 인공지능 프로세서와 대형 언어 모델의 상보적 Transformer 프로세서가 돋보였다. Compute-In-Memory (**Session 34**)에서는 다양한 메모리 구조를 채택한 CIM 프로세서 9편이 소개되었다. 이번 후기를 통해 두 세션의 총 9개의 논문에 대해 간략하게 살펴보고자 한다.

#20.1은 MediaTek에서 발표한 Digital-CIM 기반의 고화질 비디오 품질향상 프로세서이다. 고화질 데이터를 위한 12bit 연산을 수행할 수 있는 Digital-CIM Macro로 구성되어 있고, Computation Workload Balancing과 높은 하드웨어 Utilization을 위해 Computing Engine Fusion을 제안한다. Transposed CONV와 Strided CONV 연산에 대한 프로세서의 Idle Time을 줄이기 위해 추가적인 Input/Output Channel을 연산함으로써 프로세서의 활용도를 각각 65%, 58% 만큼 향상시켰다. 제안된 프로세서는 3nm FinFET 공정으로 구현되었으며, 23.2 TOPS/W의 최대 에너지 효율을 달성하였다.

 # 20.3은 Renesas에서 발표한 실시간 로봇 애플리케이션 위한 Embedded MPU이다. 해당 논문에서는 다중작업 시스템에서 100 TOPS 성능을 제공하면서도 10W 미만의 전력 소비를 충족하고, AI와 non-AI 알고리즘을 동시에 수행하기 위한 프로세서의 중요성을 언급한다. 따라서, Flexible Pruning Rate를 지원하는 Fine-grain N:16 Pruning Technology를 통해 8배 속도 향상을 달성하였다. 또한, embedded CPU와 AI accelerator의 multi-thread 그리고 pipelined processing을 지원하여 embedded CPU보다 17배 속도 향상과 12배 높은 전력 효율을 달성하였다.

#20.4는 중국 Northwestern University에서 발표한 물리정보신경망 (Physics-Informed Neural

Network, PINN)과 유한요소법 (Finite Element Method, FEM)을 모두 지원하는 통합 계산과학 (Scientific Computing) 프로세서를 소개한다. 7가지의 특별한 Dataflow를 지원하는 2D Physics Processing Element (PHY-E) 배열구조를 통해 다양한 PINN연산에 대응하며, Conjugate Gradient Method를 통해 Classical FEM 또한 지원한다. 초기 좌표계와 그리드 개수를 이용한 데이터 압축 기법으로 PINN과 FEM연산에 소모되는 전력을 27%~32% 감소시켰다. 본 논문은 2.67 TOPS/W의 최대 에너지 효율을 달성하며, 계산과학 애플리케이션에서 RTX3090 대비 최대 2590배 빠른 속도를 달성한 것이 인상적이다.

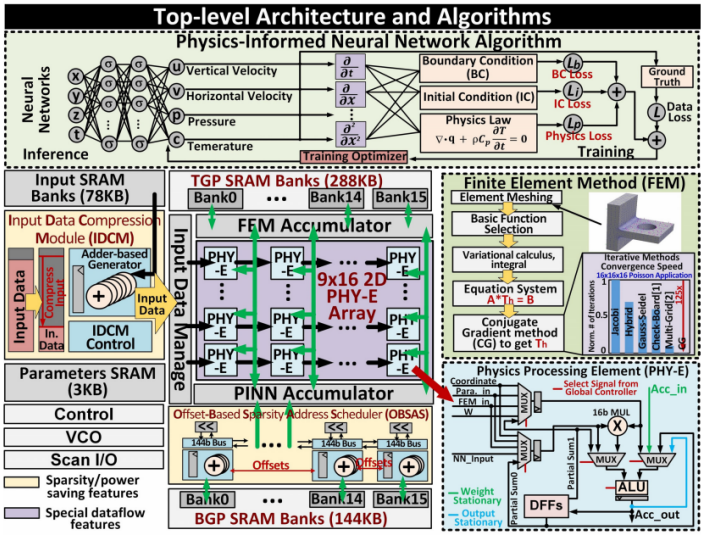

그림 1. PEY-E 배열의 계산과학 프로세서 구조

#20.6 은 UNIST에서 발표한 실시간 Semantic LiDAR-SLAM 프로세서이다. 본 논문은 기존 SLAM의 한계점을 지적하며, Point Neural Network (PNN)과 SLAM을 결합한 Semantic SLAM인 LiDAR-PNN-SLAM (LP-SLAM) 알고리즘과 함께 k-Nearest-Neighbor (kNN)과 Multi-layer Perceptron (MLP), Keypoint Extraction, Levenberg-Marquardt optimization (LMO)의 연산으로 이루어진 각 알고리즘에 최적화된 이종코어 아키텍처를 제안한 것이 특징이다. LiDAR의 시간적/공간적 연관 특성을 활용하여, 효율적인 kNN을 위한 2D/3D 구좌표계 탐색기법과 해시 페이지기반의 동적 메모리 할당 기법과 함께, MLP의 연산 스케줄링을 위한 2단계 워크로드 균형화기법을 제안하였다. 또한, LMO의 2-Phase 특성을 지원하기 위해 재구성가능한 명령어 모드를 제안하였다. 본 논문은 기존 GPU 대비 6.83배 빠른 처리량과 10349배 높은 에너지 효율을 달성하였으며, 자율주행로봇을 위한 실시간 Semantic LiDAR-SLAM이 가능함을 보인 것이 인상적이다.

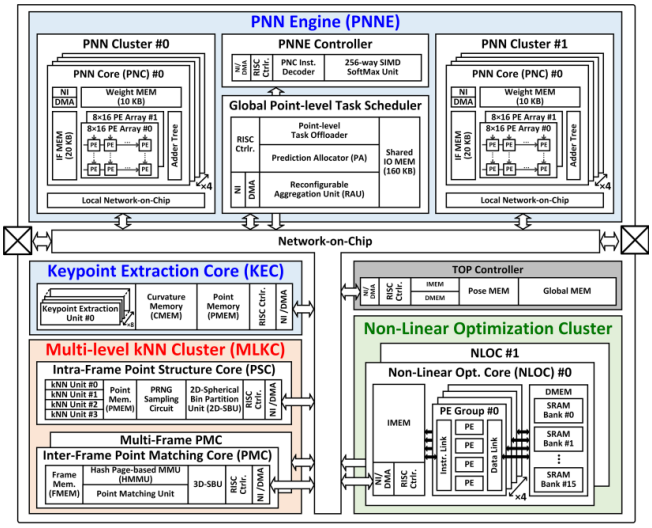

그림 2. Semantic LiDAR SLAM SoC 하드웨어 구조

#20.7은 KAIST에서 발표한 해시기반 Neural Radiance Field (NeRF) 프로세서로 실시간 3D 모델링 및 렌더링을 지원한다. 해시테이블의 외부데이터접근 문제를 해결하기 위해 해시테이블 분할 기법과 Sub-block 단위의 Sample 관리 기법을 제안하였다. 거리 기반의 Attention으로 연산을 Skip 하고 Cache 적중률에 따라 다른 연산 방식을 채택함으로써 처리량을 증가시키고 파워 소모를 줄였다. 또한, 입력데이터의 희소성과 유사도를 활용한 연산 Skip 기법을 제안하여 에너지 효율을 높였다. 본 논문은 3D Rendering 뿐만 아니라 ASIC 최초로 Modeling도 지원하며, Edge GPU 대비 18배 적은 시간과 231배 높은 에너지 효율을 달성하였다.

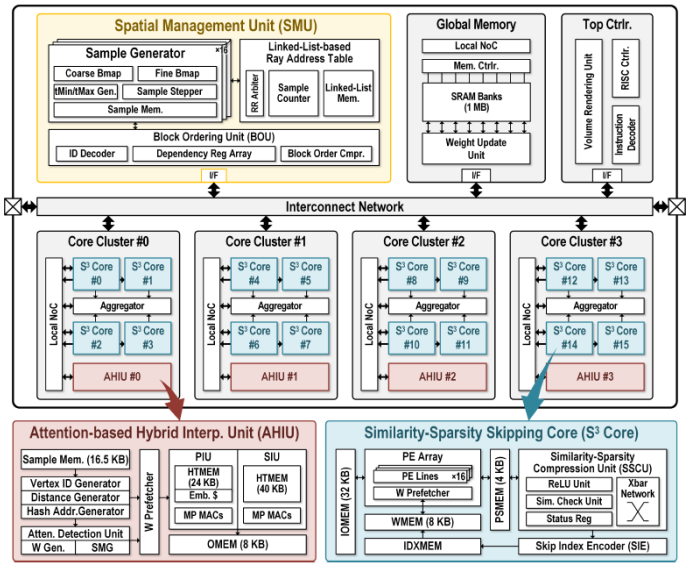

그림 3. NeuGPU 하드웨어 구조

#34.2는 TSMC에서 발표한 논문으로, Integer (INT) 기반 인공신경망과 Floating Point (FP) 기반 인공신경망 모두를 가속하기 위한 Computing-in-Memory (CIM) Processor이다. 본 논문은 Dual-mode Local-computing-cell (DM-LCC) 구조를 제안하여 INT Mode와 FP Mode에서 높은 면적 효율을 구현하였다. 또한, Zone-based Input Processing Scheme (ZB-IPS)를 제안함으로써 Exponent SUB 연산을 제거해 Energy & Area Efficient 한 연산을 구현하였다. 더불어, Two-port Gain Cell (GC) Array를 제작해 Data Update와 연산을 동시에 진행하여 System Latency를 감소하였다. 제안된 Processor는 최대 163.3 TOPS/W, 11.07 TOPS/mm^{2}의 높은 에너지/면적 효율을 달성하였다.

#34.4는 TSMC에서 발표한 논문으로 3 nm FinFET 공정을 사용한 SRAM 기반 Digital Computing-in-Memory (DCIM) 구조이다. 단위 면적 당 연산 효율과 bit density (Mb/mm^2)를 높이기 위해 Flying Bit-line (BL) Architecture를 제안한다. BL을 TOP, BOTTOM으로 나눔으로써 라인이 길어져 발생하는 신호 지연과 노이즈 문제를 최소화하였고, 레이아웃 최적화를 통해 전체 Macro Area가 약 5% 감소한다. 또한 짧아진 BL 덕분에 방전 시간이 짧아졌을 뿐만 아니라, Weight Memory Storage를 18 Segments로 나눔으로 줄어든 Row 개수 덕분에 Sense Amplifier를 제거할 수 있었다. 직렬 MAC에 비해 Shift & Add 연산이 없고 Toggle 횟수가 적어서, 병렬 MAC 연산을 사용하여 연산 처리량과 에너지 측면에서 이득을 달성하였으며, Dynamic Power를 줄이기 위해 Look-up Table (LUT) 기반의 MAC 연산을 제안하는 등의 기법들을 통해 32.5 TOPS/W, 55.0 TOPS/mm^2 효율과 3.78 Mb/mm^{2}의 Bit Density를 달성했다.

#34.7은 Tsinghua University에서 발표한 논문으로 eDRAM-LUT 기반의 Digital CIM Macro이다. 제안된 eDRAM LUT Adder는 160b 용량의 메모리와 4×8b-weight MAC을 수행하는 연산기로 재구성 가능하다. In-memory Refresh and Encode Port (IMREP)는 메모리 동작 시 Readout Port로 동작하며, CIM 동작시에는 Adder Tree 및 Shift Adder를 통하여 Successive-Accumulation 회로로 동작한다. 또한, eDRAM의 짧은 Retention Time을 보완하기 위하여, Write-Back 경로를 통하여 IMREP를 Refresh 할 수 있다. 해당 논문은 적은 면적을 요구하는 eDRAM Bitcell을 활용함과 동시에, LUT 기반의 MAC 연산을 수행하여 기존의 Digital CIM에서 발생하는 주변회로의 면적 부담을 줄일 수 있었으며, 이를 통해 2.4 Mb/mm^{2}의 메모리 집적도와 18.1 TOPS/W의 에너지 효율을 달성하였다.

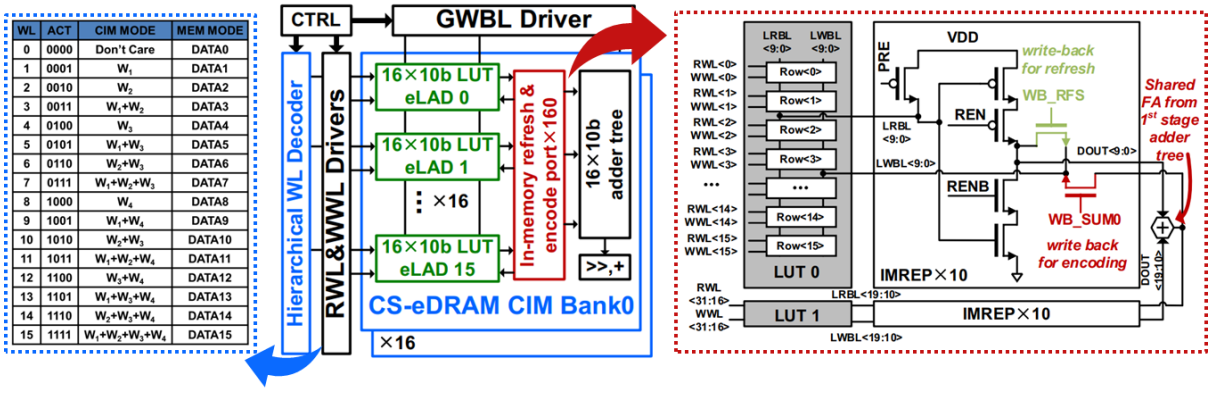

그림 4. Computation-Storage Dual-mode Reconfigurable eDRAM-CIM 구조

#34.8은 National Tsing Hua University에서 발표한 논문으로 22nm 공정의 1T1R ReRAM 메모리를 활용하여, FP16 및 BF16 정밀도를 지원하는 16Mb ReRAM-nvCIM Macro를 제안하였다. 데이터 전처리 중 발생할 수 있는 정확도 손실을 줄이기 위한 커널 별 가중치 사전 정렬 방식, 손실 없는 압축을 통해 MAC 연산의 에너지 소비와 Latency를 줄이는 Rescheduled 멀티 비트 입력 압축 방법, 그리고 ReRAM 어레이의 전류 소비를 줄이기 위한 HRS-flavored Dual-Sign-Bit (HF-DSB) 가중치 인코딩 방식을 제안함으로써, 28.7 TFLOPS/W와 31.2 TFLOPS/W의 효율을 달성하였다. Non-Volatile Memory인 ReRAM을 사용하여 칩으로 제작한 점이 돋보인다.

저자정보

이규호 교수

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