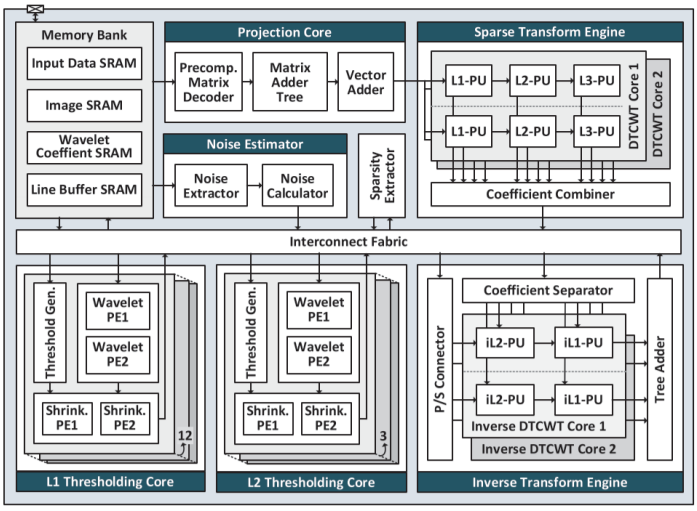
2023 IEEE VLSI Review

포항공과대학교 전자전기공학과 박사과정 홍승우

Session 3 Processors

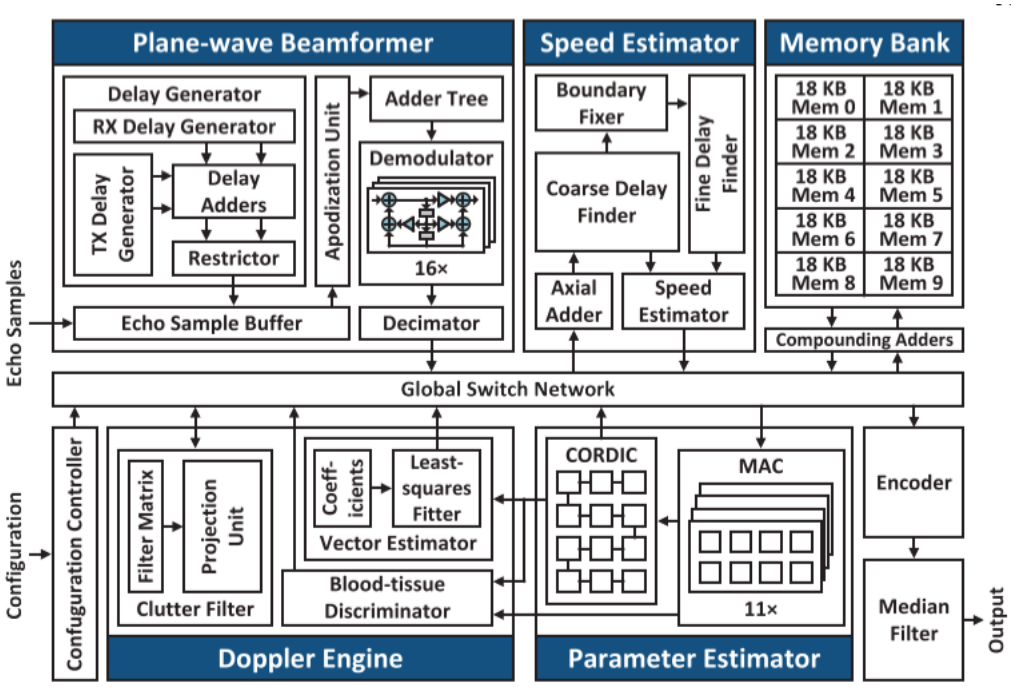
예시) 이번 2023 IEEE VLSI의 Session 3은 Processors라는 주제로 총 4편의 논문이 발표되었다. 이 세션에서는 특정 task를 수행하는 processor의 power 및 delay를 최적화시켜 높은 효율을 보여주는 processor의 implementation에 중점을 두었다.

#3-1 본 논문에서는 compressive sensing (CS)를 활용한 real-time image reconstruction processor를 처음으로 제안하였다. CS는 signal을 sparse representation으로 sampling하는 기술이며, CS encoding을 활용할 경우 sampling rate을 Nyquist theorem에서 요구하는 sampling rate보다 낮게 가져갈 수 있다는 점에서 장점이 있다. 본 논문에서는 sparse transform에 dual-tree complex wavelet transform (DTCWT)를 활용하여 traditional한 DWT 대비 image 품질은 유지하면서 실행 시간을 79% 단축시켰다. computation 단계를 reordering함을 통해 matrix inversion에서의 연산 복잡도를 99% 감소시켰으며, Iterative한 algorithm의 특성을 활용한 hardware sharing과 coefficient selection을 통해 hardware utilization을 3배 향상시켰다. Real-time streaming을 target하여 40nm CMOS technology를 통해 fabricate되었다. 48 MHz, 0.73V supply 환경에서 26.4mW를 소모하며 VGA (640x480) image를 60fps 로 real-time 처리하였으며 종래 1D CS reconstruction processor에 비해 27.2배와 19.6배 높은 energy efficiency와 area efficiency를 달성하였다.



[그림 1] Compressive sensing reconstruction processor 구조

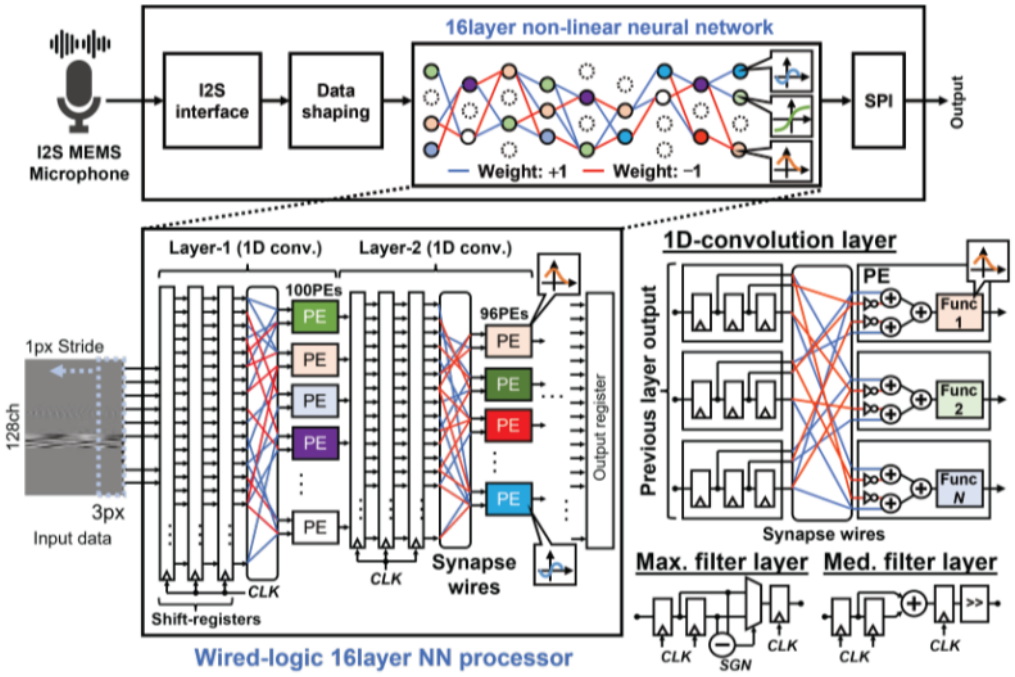
#3-2 본 논문에서는 portable ultrasound imaging의 standard, advance mode를 모두 지원하는 processor를 처음으로 제안하였다. 본 processor는 plane-wave beamforming, doppler processing 등 ultra sound imaging의 다양한 workflow를 지원한다. Look-up table을 활용한 TX, RX delay를 도출, beamline-to-point delay를 reconstruct하였으며 stored delay value 수를 99.8% 감소시켰다. 추가적으로, nappe-based beamforming을 통해 echo sample에 활용되는 storage size를 93.8% 감소시켰다. 또한, interleaved firing을 통하여 다양한 pulse repetition frequency에 대해 frame rate를 향상시켰다. 32 frame data에 대해 종래 기술 대비 전체 storage size를 78.1% 감소시켰으며 Catmull-Rom interpolation의 적용을 통해 interpolator의 면적을 79% 줄였다. 기존 standard mode만 지원하는 processor 대비 최대 36.6배 낮은 power를 소모하며 20.3배 빠른 beamforming rate을 보여주었으며, gate count 기준 60% 감소된 hardware complexity를 도출하였다.



[그림 2] Ultrasound imaging processor의 구조

#3-3 논문에서는 Voice Command Recognition (VCR)의 효율적인 처리를 위한 DNN processor의 구조 및 implementation 결과를 발표하였다. VCR은 음성 데이터에서 특정 voice command를 감지하는 기술이며, IoT device에서 주로 활용된다. 많은 단어를 recognition해야 한다는 점과 함께 always-on operation이 요구된다는 점에서 low-power implementation이 필요한 분야이다. 이를 해결하기 위해 해당 논문에서는 1) wired-logic non-linear neural network (NNN) 활용 2) Neural network 경량화 기법을 제안 및 활용하였다. 논문에서 활용한 wired-logic architecture로 구성되어 있는 DNN으로 95% 이상 pruning을 통한 높은 area efficiency를 제공하며, non-linear activation function과 neuron을 개별적으로 최적화하는 방법을 이용하여 sparse network에서도 좋은 성능을 보여준다.

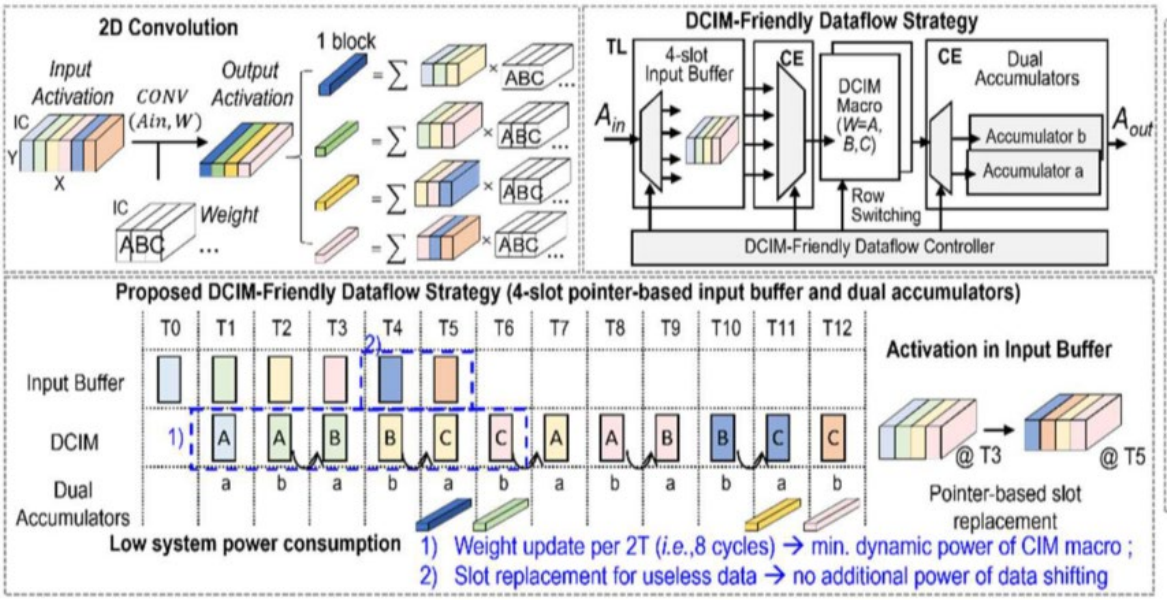
또한, binary weight coefficient를 활용하여 multiplication 과정을 제거하였으며, 이를 three-stage shift register를 통해 연산을 진행하여 DRAM/SRAM access를 제거하는 효과를 거두었다. Neural network에 logically compressed non-linear function (LC-NLF), logical-compression-aware retraining (LC-RT) 기법을 추가적으로 적용하여 hardware resource 사용량을 감소시켰다. 본 논문은 종래 기술 대비 많은 수의 command를 처리할 수 있으면서 183.4nJ/inf의 power efficiency를 달성하였다. Inference latency는 1.2ms로 analog VCR processor에 비해 10배 빠른 모습을 보여주어 low-power digital VCR을 구현했다는 점에서 큰 의미가 있다고 할 수 있다.



[그림 3] wired-logic VCR processor 구조

#3-4 본 논문에서는 end-to-end Always-on Vision (AoV) task를 위한 ultra-low power DCIM-based DL system (DCIM-DLS)을 발표했다. AoV task는 image와 관련하여 특정 물체를 detect 또는 recognize하는 작업을 지속적으로 수행하는 것을 말한다. AoV의 경우 computation과 memory access가 지속적으로 이루어지기 때문에 높은 전력을 요구하는 task이다. 본 논문에서 제안하는 DCIM-DLS는 DCIM-based DL accelerator (DCIM-DLA), RISC-V microprocessor, 512KB level-1 memory, JTAG controller, 그리고 SPI controller를 포함한다. 해당 DCIM-DLA에서는 convolution, depthwise convolution, fully connected re-quantization, ReLU, 기록 pooling operation을 지원하며, 2개의 64x16 MAC array와 8-KB memory를 포함하고 있다. 각 compute in memory에서는 1024 "4 activation 4 weight" multiplication을 수행한다. Weight의 reuse를 높이기 위해 DCIM-DLA에서는 dual accumulator를 활용하여 인접한 두 point에 대한 연산을 interleaving 방식으로 수행한다.

또한, RC loading을 줄여 power와 timing 측면에서 이득을 보기 위해 folded kernel selector (FKS)를 제안하였다. FKS는 DCIM-DLA의 ping-pong operation을 지원하기 위해 kernel을 weight group으로 나누며, dynamic circuit topology를 활용한 adder tree 구성을 통해 전체 input에 대한 연산 결과를 serial하게 generate하게 된다. Uni-precision 구성보다 좋은 성능을 보여주는 mixed-precision MAC computation을 위해 DCIM-DLA를 reconfigurable하게 구성하였다. 제안하는 DCIM-DLS는 12nm CMOS technology로 fabricate 되었으며 57TOPS/W의 energy efficiency를 달성하였고, 1.61mW의 power로 초당 15 frame을 처리하여 sub-mW ultra-low power AoV requirement를 제시하였다.



[그림 4] 2D convolution 에서의 DCIM-friendly dataflow

저자정보

명예기자 홍승우

- 소 속 : 포항공과대학교 전기및전자공학과 박사과정
- 연구분야 : DSP Architecture, ASIC/FPGA design
- 이 메 일 : seungwoohong@postech.ac.kr
- 홈페이지 : https://sites.google.com/view/epiclab

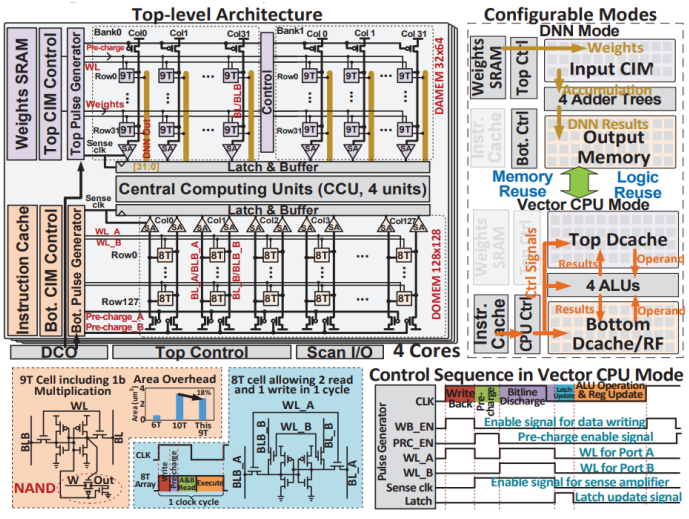
2023 IEEE VLSI Review

성균관대학교 전자전기컴퓨터공학과 박사과정 조종완

Session 21 PIM CIM Systems

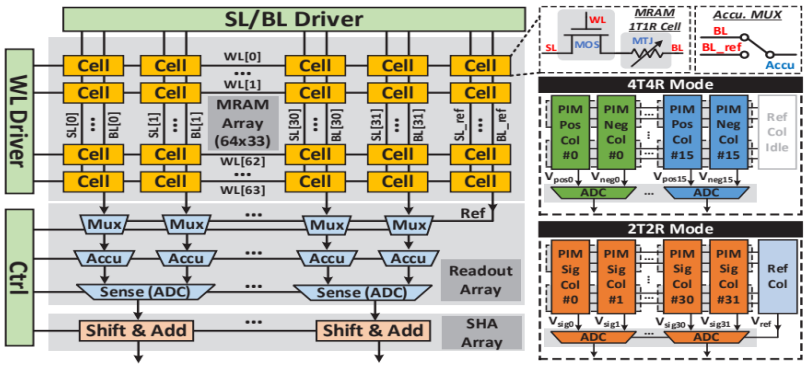
이번 2022 IEEE CICC의 Session 21은 PIM CIM Systems라는 주제로 총 5편의 논문이 발표되었다. 이 세션에서는 Neural Network에서 높은 성능을 가질 수 있는 PIM(Process-In-Memory) 또는 CIM(Compute-In-Memory) 설계를 통해 Deep Learning, 3D Modeling 등의 다양한 분야에서 높은 성능을 가지는 IP를 설계하였으며, MRAM, SRAM 등의 다양한 메모리를 사용하여 시스템을 적절하게 구성하였습니다. PIM과 CIM System의 많은 연구를 통해 저장장치와 처리장치 사이의 데이터 병목 현상을 최소화 할 수 있을 것으로 기대됩니다.

(#21-1. Northwestern University, USA) 본 논문에서 제안하는 CIM의 경우 최근 CPU 관련하여 개선된 많은 부분 중에 데이터 처리 및 전송에 관한 연구를 진행하였으며, CPU와 DNN 작업을 진행하는 과정에서 기존의 아키텍처의 CPU와 Configurable 가능한 CIM을 이용하여, DNN 및 CPU의 성능을 개선시켰다. 그림 1에서 볼 수 있는 Chip architecture 처럼 제안하는 논문에서는 CPU의 mode에 따라서 CIM에 Mode도 같이 동기화하여 사용하는 방법을 통해 높은 데이터 처리 효율 결과를 만들었다.



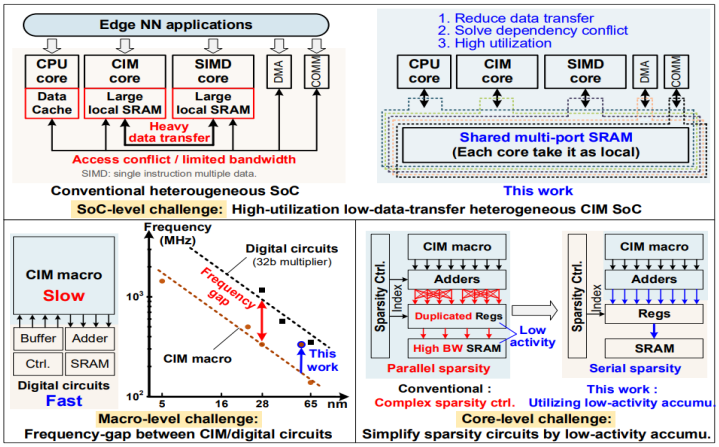
[그림 1] 논문 21-1에서 제안하는 Chip architecture 및 reconfiguration mode and CIM cells

(#21-2. KAIST, Republic of Korea) 본 논문에서는 MRAM PIM의 정확도를 높이기 위해 새로운 구조의 double differential 4T4R MRAM과 2T2R을 Hybrid로 사용하는 PIM을 제안하였다. 그림 2에서 볼 수 있는 것처럼 제안하는 논문에서는 4T4R만 사용하게 되었을 때의 Latency 부담과 높은 에너지 효율문제를 Hybrid 구조를 사용하여 극복하였으며, 높은 정확도가 필요할 경우에는 4T4R 사용하면서 Condition에 따라서 Mode를 변경할 수 있다. 기존 단일 구조를 사용한 구조보다 최적의 상황을 선택할 수 있는 Mode 변경을 통해 에너지 효율과 Inference Acc.를 모두 개선시킬 수 있는 방법입니다.



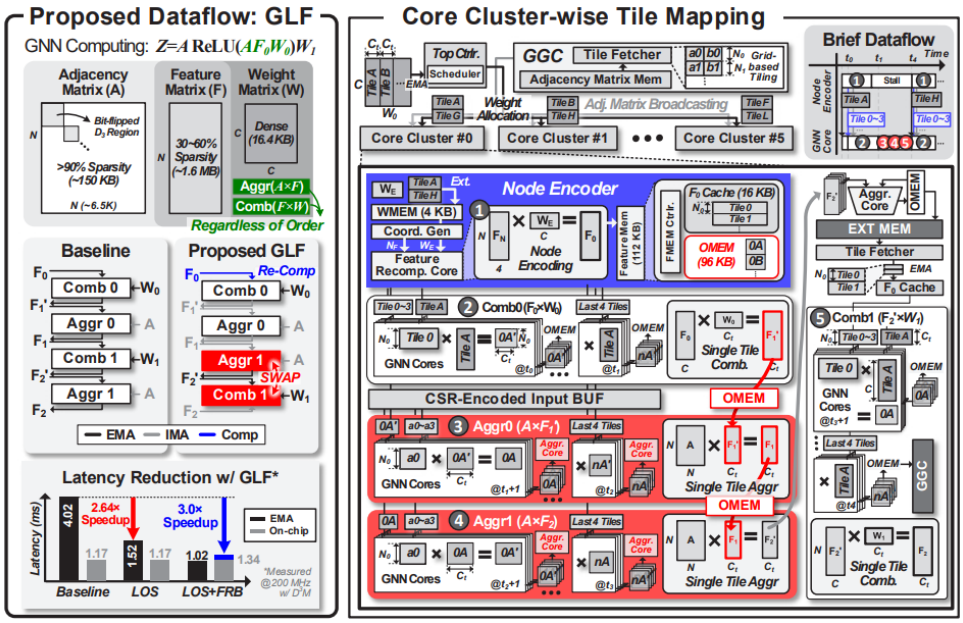
[그림 2] 논문 21-2에서 제안하는 MRAM PIM macro structure and two working modes

(#21-3. Tsinghua University, China) 본 논문에서는 그림 3에서 볼 수 있는 것처럼 Multi-core SoC에서 CIM을 간단하게 구현하여 높은 에너지 효율을 달성하기 위해 CIM의 SRAM에서 Multi-port로 서로의 데이터 입출력을 공유하고 CIM macro bandwidth 자체를 높이면서 Producer-consumer inst dependency ctrl. 을 통해 데이터 전송을 최적화하였다.



[그림 3] 논문 21-3에서 제안하는 SoC/macro/core-level challenges in a heterogeneous CIM SoC

(#21-4. KAIST, Korea) 본 논문에서는 그림 4에서 볼 수 있는 것처럼 GLF-based GNN을 사용하여 Dataflow를 최적화하여, 기존 연구 대비 Processing Latency를 최적화할 수 있었으며, 낮은 전력 소모와 데이터 처리 기준으로 높은 에너지를 소모하는 구조를 연구하였다.



[그림 4] 논문 21-4에서 제안하는 GLF-based GNN Core Cluster Dataflow Optimization

(#21-5. KAIST, Korea) 본 논문에서는 8개의 PIM을 그룹으로 사용하여, 3D 모바일 NeRF PIM processor를 재활용하여, 대량의 메모리 액세스 에너지를 효과적으로 활용할 수 있는 방법을 제안하였습니다. A ray-wise dynamic block reuse를 최대화하여, PIM의 throughput을 개선시켰습니다.

저자정보

조종완 박사과정 대학원생

- 소　　속 : 성균관대학교 전자전기컴퓨터공학과 박사과정
- 연구분야 : Energy Harvesting, PMIC
- 이 메 일 : jw1114@skku.edu
- 홈페이지 : http://iclab.co.kr