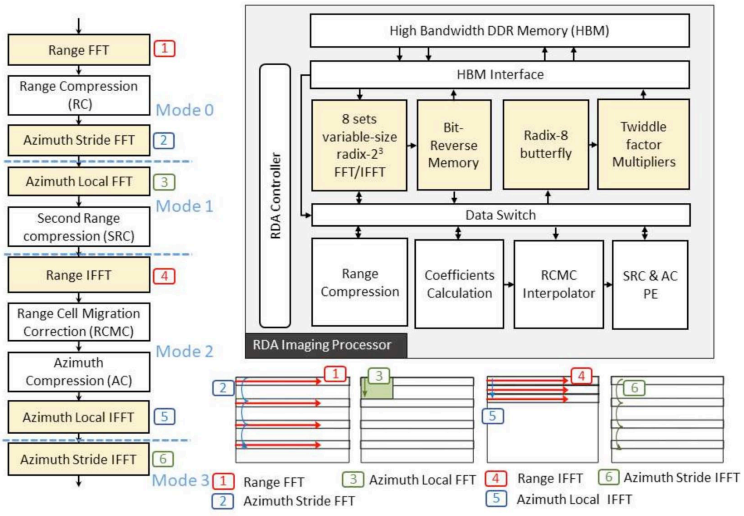
A-SSCC 2022
IEEE Asian Solid-State Circuits Conference

포항공과대학교 전기및전자공학과 박사과정 홍승우

Session 19 Imaging & Machine Learning Processing on FPGA

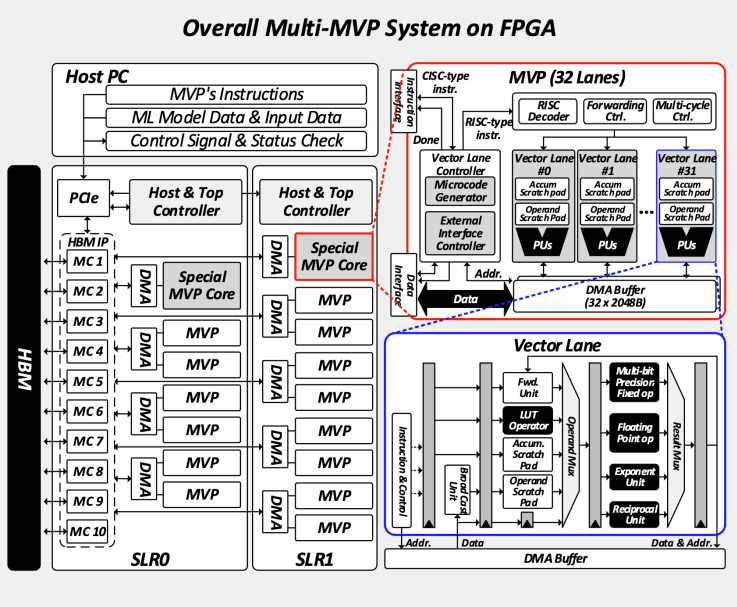
이번 A-SSCC 2022의 Session 19는 Imaging & ML processing on FPGA라는 주제로 총 네편의 논문이 발표되었다. 그 중 세 편의 논문은 ML기반의 Image processing system이었으며, 하드웨어 utilization을 높여 높은 효율을 얻고자 하는 것에 중점을 두었다.

#19-1 논문에서는 실시간 고해상도 Synthetic Aperture Radar (SAR) 이미지 처리를 위한 Imaging processor 구조 및 FPGA 구현 결과를 발표하였다. SAR imaging은 항공기나 지구 궤도상에서 레이더 신호를 이용하여 지상의 이미지를 수집하는 기법으로, 수집되는 신호의 양이 방대하여 실시간 처리가 용이하지 않다. 이를 해결하기 위해 해당 논문에서는 1) data access pattern 최적화 2) hybrid datapath 구현 및 3) multi-segment high-order Taylor series를 통한 연산 근사화를 수행하였다. 거리(range)와 방위(azimuth)의 2D 도메인에서 작은 블록으로 나뉘어 처리되는 FFT/IFFT의 dataflow 최적화를 통해 HBM의 throughput을 최대한 활용하여 8-parallel processing을 가능하게 하였으며, multi-segment high-order Taylor expansion이 double precision division 및 square-root 연산의 99 % 이상을 제거하였다. 또한 throughput 및 SQNR 목표를 만족하는 최적의 hybrid datapath (17-bit custom FP & double-precision operations)를 구현하여 약 40%의 double-precision adder 및 multiplier를 감소하였다. 본 논문은 종래 기술 대비 월등히 빠른 pixel당 5.07 ns 프로세싱 latency를 달성했으며, 1.6초 동안 capture된 8Kx32K SAR image를 1.35초에 처리하여 실시간 처리를 구현해냈다는 점에서 큰 의미가 있다고 할 수 있다.



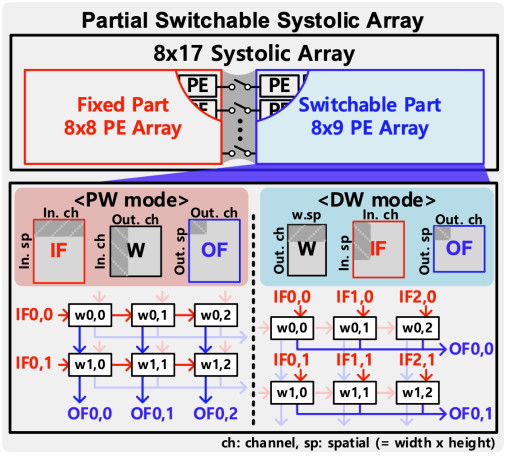
[그림 1] RDA imaging processor 구조 및 2D-FFT processing flow

#19-2 논문에서는 mixed-precision vector processor (MVP) 및 이를 사용한 general-purpose ML 가속기를 발표했다. 본 논문에서 발표된 MVP 및 multi-core system은 1) fixed/floating point 데이터 타입을 동시에 지원하며 2) ML model mapping을 간단하게 해주는 two-level ISA 와 microcode generator, 그리고 3) multiple MVP를 효율적으로 다루기 위한 software stack을 제공한다. MVP의 vector lane을 구성하는 mixed-precision arithmetic unit은 8-, 16-bit fixed point 연산 및 16-bit BFP 연산을 동시에 지원하며, nonlinear operation을 위한 LUT operator 및 interpolation 정확도를 향상시키기 위한 configuration table이 함께 존재한다. Software stack은 ML model을 single MVP에서 수행될 subtask로 쪼개어 scheduling하며, CISC-like ML-friendly ISA로 program된 code는 MPV의 microcode generator에서 RISC-like vector ISA로 변환되어 구동된다. 이러한 구조를 통해 다양한 data type의 general purpose ML model을 손쉽게 프로그래밍하여 가속할 수 있으며, 이전에 발표된 ML 가속기의 성능을 능가하는 409.6 GOPS/204.8 GFLOPS를 달성하였다.



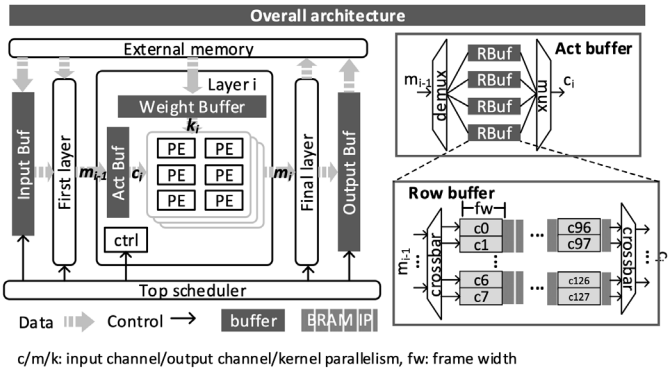
[그림 2] Multi MVP로 구성된 general ML accelerator 구조

#19-3 논문에서는 Unsupervised learning (UL) 기반의 monocular depth estimation (MDE) processor를 발표하였다. MDE란 하나의 카메라만 사용하여 추가적인 라벨이나 preprocessing, RGB-D sensor 없이 3D depth image를 생성하는 것이다. UL 기반 MDE에서는 compute intensive 한 pixel-wise (PW) conv 연산과 memory-intensive한 depth-wise (DW) conv 연산의 balance를 맞춰 전체 시스템의 efficiency를 높이는 것이 중요한데, 본 논문에서는 이를 위하여 PW/DW를 layer를 fusion하여 처리하는 multi-path simultaneous processing (MPSP) 구조를 제안하여 중복된 input feature map에 의한 extra external memory access를 16.8% 줄였으며, paired 및 unpaired processing을 동시에 효율적으로 지원하기 위해 partial-switchable systolic array (PSSA) 구조를 설계하여 workload imbalance를 해결함과 동시에 encoder와 decoder의 latency를 각각 46%, 57% 줄였다. 추가로 static situation에서 camera pose가 거의 변하지 않는 특성을 사용하여 동적으로 모델을 다운샘플링하는 기법을 통해 에너지 효율을 48~59% 줄였다. 본 논문은 workload imbalance가 존재하는 UL-MDE의 연산을 효율적으로 처리하기 위한 기법 및 FPGA를 통해 on-device training의 실시간 처리를 보였다는 점에서 주목할만하다고 보인다.



[그림 3] Partial switchable systolic array 구조 및 PW/DW convolution mode

#19-4 논문에서는 Learned image compression (LIC)를 위한 fully-pipelined 가속기를 발표하였다. Autoencoder 기반 LIC 인코딩/디코딩을 효율적으로 처리하기위해 deconvolution의 zero-skipping 기법을 포함하는 fixed point, fine-grained pipelined 구조를 제안하였다. 제안하는 구조는 kernel-width 방향의 parallelism 향상을 위한 Cascaded DSP를 사용하였고, channel 방향으로의 parallel 구현을 위해 multiple BRAM과 crossbar를 포함하는 ping-pong 구조의 activation buffer를 포함하여 종래 LIC 구현에 비해 훨씬 우수한 40.69/35.77 fps 인코딩/디코딩 속도를 보였다 (720p 기준).



[그림 4] Fine-grained pipelined LIC accelerator 구조

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