



●이력재 교수

▶ Education

- 1991.8 ~ 1996.12 미 Purdue 대학교, 박사
- 1987.3 ~ 1989.2 서울대학교 전자공학과 석사
- 1983.3 ~ 1987.2 서울대학교 전자공학과 학사

▶ Research Interest

- 컴퓨터 구조 및 병렬처리
- 인공지능 반도체 하드웨어 설계
- 영상처리용 반도체 하드웨어 설계

▶ Career

- 2001.3 ~ 현재: 서울대학교 전기정보공학부 교수/학부장
- 2020.1 ~ 현재: 대한전자공학회 부회장
- 1998.11 ~ 2001.2: 미 Intel, 선임연구원

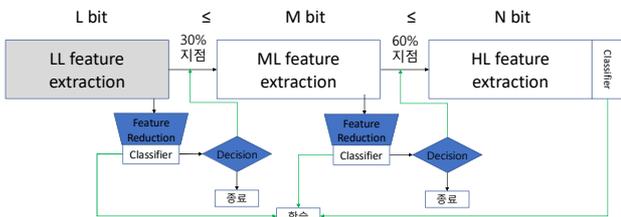
▶ 연구실 현황

- 22 Ph.D. candidates / 18 M.S. candidates
- 홈페이지: <https://capp.snu.ac.kr>
- contact: 02-880-1302 (서울대학교 104-1동 306-4호)

●인공지능 반도체 관련 핵심 보유기술

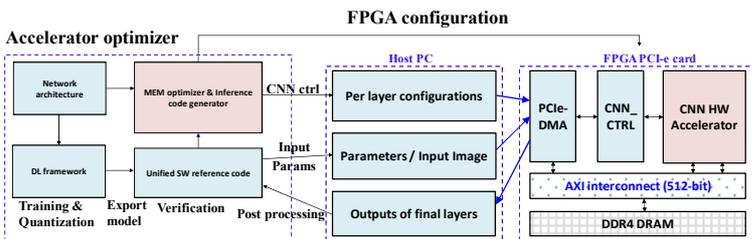
1) 영상 분석을 통해 run-time에 quantization 정밀도를 결정하여 NPU의 소비 전력을 최적화하는 기술

- ▶가변 정밀도 IP 를 포함한 NPU architecture
- ▶Main network를 LL(low level), ML(Midium level), HL(high level) feature extraction으로 분리하고 뒤로 갈수록 높은 정밀도가 할당되며 각 단계가 종료되면 가변 정밀도 IP를 동작시킴
- ▶가변 정밀도의 decision network에서 계산된 uncertainty값에 따라 다음 단계로 넘어갈지/종료할지를 결정하여 interrupt와 정밀도 값을 NPU에 전달



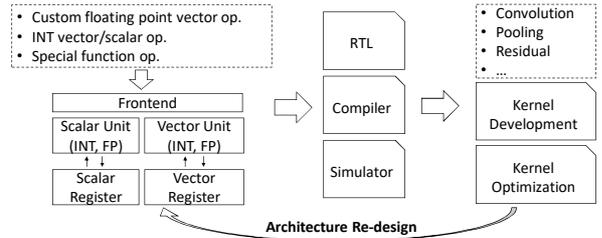
2) CNN을 타겟으로 하는 가속기 설계 기술 및 quantization training, model analysis 등의 pre-processing 기술

- ▶CNN을 타겟으로 하는 가속기 및 quantization training, model analysis 등의 pre-processing을 포함한 framework 보유
- ▶Accelerator optimizer에서 network를 분석하여 CNN control signal을 생성하고, FPGA 상에서의 최적화된 memory access를 제공
- ▶각 layer마다 weight와 activation에 대해 최적의 precision을 갖게 하는 quantization optimization을 제공하여 model size를 성능의 손실없이 줄임



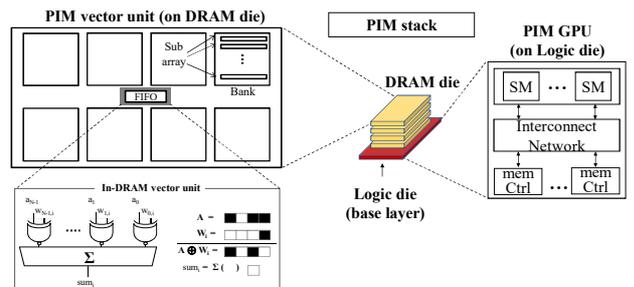
3) Custom floating point 딥러닝 연산을 지원하는 벡터 프로세서의 설계 기술

- ▶Custom floating point 연산을 지원하는 벡터 프로세서의 ISA 및 마이크로 아키텍처
- ▶상용 툴과 연동하여 RTL, 컴파일러, 시뮬레이터 등을 보유



4) Processing In Memory 기반의 CNN 가속 기술

- ▶3D stacked memory의 logic die 및 DRAM die에 위치하여 CNN의 memory bottleneck을 효과적으로 처리할 수 있는 아키텍처 보유
- ▶Logic die에는 GPU 기반의 lightweight한 processor 탑재
- ▶DRAM die에는 Binary Neural Network를 처리할 수 있는 vector unit 탑재



5) Approximate memory 기반의 bit-level precision 조절 기술

- ▶딥러닝 training시 bit-level로 precision을 조절하고, 이를 효율적으로 처리할 수 있는 HBM 기반의 GPU용 memory system 시스템 보유
- ▶8, 16, 32 등이 아닌 precision을 가지는 data를 효과적으로 fetch하는 cache 및 core 구조 보유
- ▶Memory mapping을 변경하여 MSB와 LSB를 서로 다른 bank에 저장
- ▶LSB가 저장된 bank에 대한 refresh를 줄여 전체 refresh 에너지를 감소시킴

