chapter 2

병렬 컴퓨팅 기반 인공지능 프로세서 기술 동향

한진호 ‖ 한국전자통신연구원 실장
권영수 ‖ 한국전자통신연구원 본부장

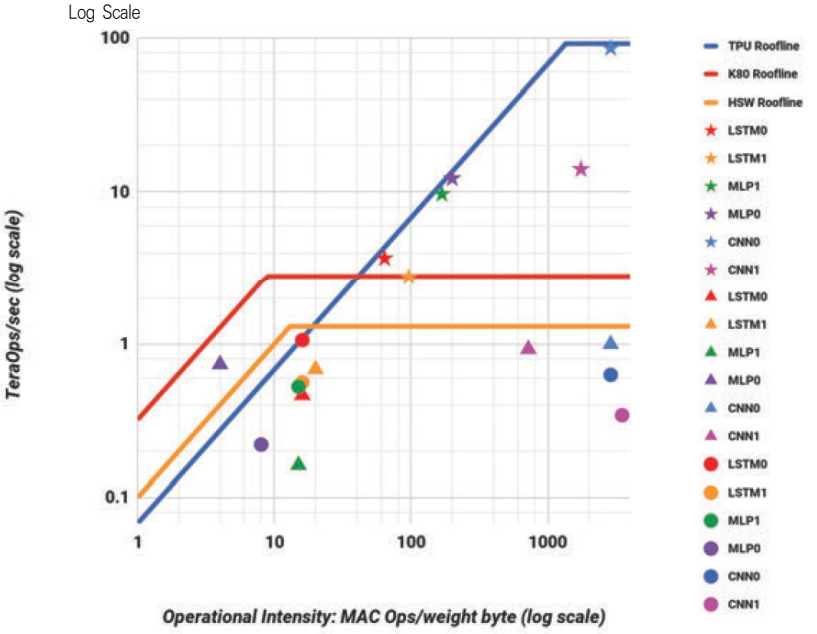
I. 서론: 병렬컴퓨팅의 대중화

컴퓨터의 클럭 속도는 1980년도 중반부터 2004년까지 컴퓨터 성능을 향상시키는 데 가장 영향력 있는 요소였다. 실행시간은 명령어 수를 명령어 당 평균시간(1/IPC)을 곱한 것과 같았는데 클럭 수를 늘리면 명령을 실행하는 평균시간은 짧아진다. 그러나 칩의 전력 소모량은 $P=C\times V^2\times F$의 공식에 따른다. 여기서, C는 전기용량, V는 전압이고, F는 주파수이다. 주파수 수를 높이면 전력 사용량이 늘어나게 되고, 이는 전력의 증가는 동작 시 많은 문제를 일으키고 있다. 그래서 무어의 법칙은 18에서 24개월 동안 집적도가 2배씩 늘어난다는 것을 예측하는 것이지만, 이는 주파수 척도가 아닌 병렬 컴퓨팅에 의해 계속해서 유효하게 된다.

병렬 컴퓨팅은 동시에 많은 계산을 하는 연산 방법으로 크고 복잡한 문제를 작게 나눠 동시에 병렬적으로 해결하는 데에 주로 사용된다. 그러나 병렬화로 인한 속도 향상은 병렬화할 수 없는 작은 부분이 전체적인 병렬화에 영향을 가져온다는 암달의 법칙에 의해 그

* 본 내용은 한진호 실장(☎ 042-860-6558, soc@etri.re.kr)에게 문의하시기 바랍니다.
** 본 내용은 필자의 주관적인 의견이며 IITP의 공식적인 입장이 아님을 밝힙니다.
*** 인공지능 프로세서연구실원: 김병조, 이미영, 정재훈, 김현미, 함제석, 김혜지, 전인산, 조용철, 최민석, 신경선, 여준기, 양정민, 김찬, 석정희, 전영득, 조민형, 박기혁, 김진규, 김주엽, 이주현, 김성민

OPEN 출처표시 상업용금지 공공누리 공공저작물 자유이용허락



〈자료〉 David Patterson, "50 Years of Computer Architecture: From the Mainframe CPU to the Domain-Specific TPU and the Open RISC-V Instruction Set", ISSCC 2018.

[그림 1] Haswell, K80, TPU의 Roofline 성능 모델 비교

한계도 있다. 또한, 병렬화는 자료의 종속성에 의해 병렬화에 한계를 가지기도 한다.

인공지능 알고리즘을 수행하기 위해서는 많은 연산량을 요구하고 있으며, 이러한 연산 성능을 내기 위해서는 병렬 컴퓨팅의 방법을 사용할 수 있다. [그림 1]은 CPU, GPU, Google TPU의 성능을 나타내는 Roofline 성능 모델 그래프이다[1].

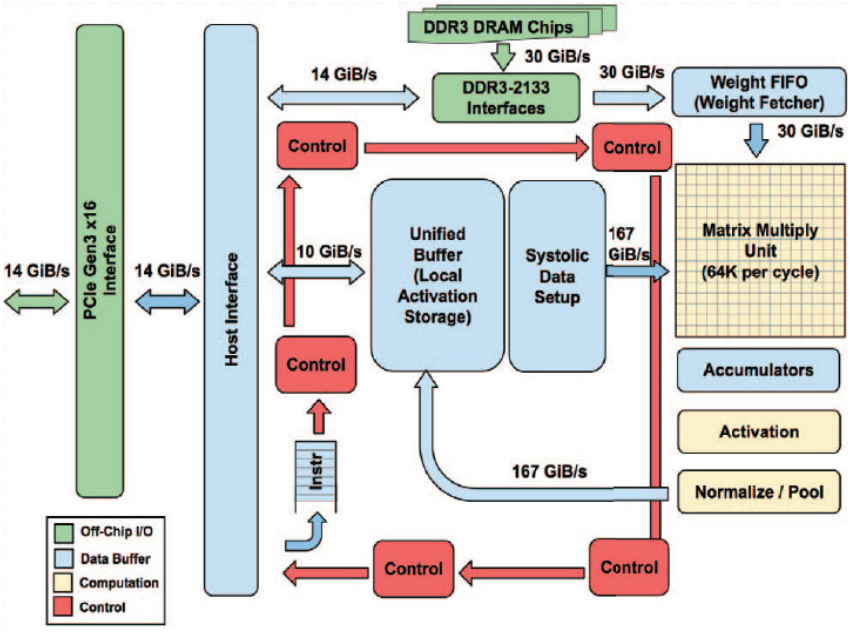
nVidia K80 GPU는 SIMT 구조로 Intel Haswell CPU보다 병렬 컴퓨팅을 이용하여 성능을 높이고 있다. 그리고, Google TPU AI Processor는 K80보다 더 높은 97TOP/sec의 성능을 내고 있고, Operational Intensity인 하나의 weight를 읽어 와서 더 많은 연산을 수행할 수 있도록 하고 있다. 즉, 읽어온 오퍼랜드로 더 많은 연산을 수행할 수 있도록 병렬 컴퓨팅 성능을 높이고 있다.

인공지능 프로세서는 이렇게 인공지능 알고리즘을 빠르게 수행하기 위해 병렬 컴퓨팅 성능을 극대화하고 있으며, 외부 메모리 대역폭 한계를 극복하기 위해 읽어온 Weight 값을 최대한 재활용하여 Operational Intensity를 높이고 있다.

II. 인공지능 프로세서

인공지능 프로세서인 Google TPU는 다음과 같은 구조로 일반적인 CPU, GPU보다 30~50배 높은 에너지 효율성으로 Deep Neural Network 연산의 15~30배의 추론 성능을 높이고 있다[2].

[그림 2]에서 오른쪽 중앙에 있는 Matrix Multiply Unit은 매 사이클 당 256×256의 8비트 곱셈 및 덧셈을 할 수 있는 MAC을 포함하고 있다. 그리고 16비트 곱셈 결과는 4MiB Accumulator에 저장이 된다. 그리고 Matrix Multiply Unit은 매 사이클 당 256개의 partial product 값을 출력하고 이를 Accumulator에 의해 누적한다. 또한, Weight FIFO에 의해 30GiB/s의 대역폭으로 Weight 값을 공급한다. 연산을 위한 입력 값은 14GiB/s의 대역폭을 가지는 PCIe Gen3×16 인터페이스를 통해 Unified Buffer에 전송되고, 24MiB 용량을 가지는 Unified Buffer는 167GiB/s의 대역폭으로 Matrix Multiply Unit에 공급된다.



〈자료〉 Norman P. Jouppi et al., "In-Datacenter Performance Analysis of a Tensor Processing Unit," In Proceedings of the 44th Annual International Symposium on Computer Architecture, 2017.

[그림 2] 구글 TPU v1 아키텍처

TPU는 18 코어로 이루어진 Intel Haswell CPU 또는 Nvidia Kepler K80 GPU 면적의 약 절반을 가지고, 절반의 전력을 소모하지만, 25배의 MAC 연산 성능을 내고, 3.5배의 온칩 메모리를 가지고 있다.

III. 인공지능 프로세서 개발 동향

1. 모바일 인공지능 프로세서

스마트폰 회사들이 2019년 발표한 모바일 AP들은 다수의 혼종 CPU 코어와 GPU 이외에 인공지능 프로세서인 NPU(Neural Processing Unit)를 대부분 포함하고 있는 구조이다. 퀄컴은 이와는 다르게 전용 NPU를 적용하지 않고 텐서(tensor) 가속기로 DSP를 채용하고 있다. 애플, 화웨이, 삼성, 퀄컴, 미디어텍의 2019년 발표된 모바일 AI 프로세서의 특징은 다음과 같다[3]-[9].

화웨이 Kirin 칩은 3D 텐서 계산 구조에서 착안한 DaVinci 아키텍처를 신경망 연산 코어로 적용했다. DaVinci 코어는 3D 텐서 계산 방식에 맞게 구조화된 16×16×16 MAC 연산기 큐브(cube)를 포함하며, 각 MAC 연산기는 사이클 당 1개의 FP16 연산이나 2개의 INT8 연산을 수행한다. DaVinci 코어는 MAC 연산기 큐브 이외에 스칼라 ALU, 벡터 ALU, load/store 유닛 등을 포함한다[7].

삼성 Exynos 990에 적용된 NPU의 구조는 NPU 제어기와 2개의 NPU 코어로 구성되고, NPU 제어기는 CPU, DMA, SRAM, 네트워크 제어기를 포함한다[6]. NPU는 1,024개 MAC 연산기로 구성되며, Weight의 희소성을 활용하여 필요한 연산만을 수행할 수 있는 NPU 구조를 제안했다. Inception-v3 신경망으로 3.4 TOPS/W 결과를 보였다.

모바일용 저전력 CPU, GPU IP를 주력으로 하는 ARM사는 다양한 AI 응용에 적용할 수 있도록 3가지 사양의 Ethos-N NPU를 발표했다[10]. Ethos-N37, N57, N77은 각각 512개, 1,024개, 2,048개의 8×8 MAC 연산기로 구성되며 1~4 TOPS 성능을 보인다.

DSP IP가 주력인 CEVA사는 인공지능 프로세서 NeuPro-S를 발표했다[11]. AI 엔진인 NeuPro-S 엔진과 벡터연산용 CEVA-XM DSP로 구성되어 있다. NeuPro-S 엔진은 신경망의 대표적 레이어들인 콘볼루션(convolution), 액티베이션(activation), 풀링(pooling)

레이어 처리 기능을 내부에 포함하고 있으며, 12.5 TOPS 처리 성능 결과를 발표했다.

Gyrfalcon사는 매트릭스 연산 전용 엔진을 구현한 Ligthspeeur 2801, 2803을 출시했다[12]. 168×168 MAC 연산기로 구성된 매트릭스 연산 엔진을 포함하며, 300mW의 저전력으로 2.8 TOPS의 성능으로 9.3 TOPS/W의 높은 에너지 효율 결과를 발표했다. PIM(Procssing In Memory) 구조로 설계하여, 전력을 많이 소모하는 외부 메모리로부터의 데이터 전송을 없애서 저전력으로 동작할 수 있도록 설계하였다.

이스라엘 스타트업 Hailo사는 자체개발 코어 8개로 구성된 Hailo-8로 CES 2020 Innovation Award를 수상했다[13]. 5W 이하의 전력으로 26 TOPS의 높은 성능을 발표했다. ResNet-50(224×224) 신경망에 대해 672 FPS, 1.7W로 NVIDIA Xavier 대비 1/15 전력으로 동등한 신경망 수행능력을 보였다.

2. 서버 인공지능 프로세서

NVIDIA는 1990년대 인텔에 맞서서 CPU를 개발하기 위해 설립된 기업이다. CPU 시장에서 x86을 내세운 인텔의 시장 지배자로서의 위치를 확인한 후 2000년대 초에 GPU에서 Geometry Processing과 Pixel Processing을 통합한 최초의 GPU를 출시하면서 그래픽스 시장의 최강자로 자리 잡는다. 이후 그래픽스 카드 시장이 정체되면서 NVIDIA는 GPU를 Parallel Processing을 위한 칩으로 이용하는 GPGPU라는 개념을 내어놓는다. GPGPU의 근본 구조는 Stream Processor(SP)를 기반으로 하는 Single Instruction Multiple Thread(SIMT) 구조의 프로세서로 구성되어 있다는 점에서 NVIDIA는 병렬컴퓨팅을 개발하는 회사로 급속히 성장하기 시작했고, 아키텍처의 구조를 변화, 향상시키면서 Tesla(2007년), Fermi (2010년), Kepler(2012년), Maxwell(2014), Pascal(2016), Volta(2017), Ampere(2020)라는 코드명을 붙이면서 발전해 왔다[14]-[18].

2020년 5월 GPU Technology Conference(GTC) 2020에서 차세대 GPU 아키텍처인 Ampere 기반의 데이터센터용 AI 프로세서 A100을 공개하였다[19]. A100에는 8개의 GPU processing cluster(GPC), GPC당 8 Texture processing cluster(TPC), 그리고 TPC 별로 2개의 SM으로 구성되어 총 128개 SM이 집적되어 있다.

A100은 3세대 Tensor core 기술로서 FP32데이터 가속용 TensorFloat-32(TF32) Tensor core, HPC용 IEEE 호환 FP64 Tensor Core, FP16과 동일한 처리량을 가지는

BF16 Tensor core와 INT8/INT4 및 Binary 등의 모든 데이터 유형에 대한 가속을 지원하면서 희소성 연산 기능을 제공한다. Tensor Core의 TF32연산은 V100의 FP32 FMA보다 10배 빠르며 희소성 연산에서는 20배 빠른 가속 성능을 나타낸다. FP16/FP32 혼합 정밀 딥러닝 연산에서는 V100보다 2.5배 높은 성능을, 희소성 연산에서는 5배 높은 성능을 보인다. 그리고 A100은 40GB HBM2 메모리를 적용하여 V100보다 1.7배 이상의 메모리 대역폭을 지원하고 3세대 NVLIN와 NWSwith 기술로 600GB/s 대역폭을 구현하였으며, Multi GPU, Multi node 및 Multi-Instance GPU(MIG)를 통해 다중 GPU 시스템 연결을 위한 확장성을 제공한다.

GTC 2020에서는 A100 프로세서를 기반으로 한 5페타플롭급의 DGX A100 데이터센터용 플랫폼과 700페타플롭의 140개 DGX A100 시스템으로 구성된 차세대 DGX 슈퍼 POD(DGX SuperPOD)를 함께 공개하였다. DGX A100은 8개의 A100으로 구성되어 6개의 NV 스위치와 NV 링크 기술을 통해 초당 4.8TB의 양방향 대역폭을 지원하여 Mellanox사의 네트워킹 기술과 함께 데이터센터 확장에 편리하도록 설계되었다. GTC 2020 키노트를 통해 Nvidia는 Ampere 아키텍처를 바탕으로 공통 GPU 아키텍처를 개발하고 HPC부터 엣지까지 다양한 제품군에 공통 적용하는 전략을 펼치고 있음을 알 수 있다.

퀄컴은 온 디바이스 AI 기술력을 바탕으로 5세대 AI 엔진을 탑재한 서버 전용 AI 가속기 솔루션인 추론용 cloud AI 100을 2019년 4월에 개발하였고, AI 반도체용 SW 및 개발 툴인 Qualcomm Neural Processing SDK를 함께 제공하였다. 퀄컴은 데이터센터용 AI 프로세서 맞춤용 라이브러리, 컴파일러 등 SW 통합 개발 환경을 제공하여 서버용 AI 생태계에서도 시장 주도권을 장악하려고 노력하고 있다. Cloud AI 100은 350 TOPS 이상의 연산 성능과 경쟁 AI 추론 솔루션 기술 대비 10배 이상의 와트 당 성능을 가지고 있다고 발표하였다. 2020 CES에서는 Cloud AI 100을 기반으로 한 첫 제품인 Smart Edge Box를 2020년 하반기 대규모 상용을 목표로 개발 중에 있다고 발표하였다.

중국은 프로세서 기술 개발을 위해서 CAS(Chinese Academy of Science, 중국의 정부출연연구소)를 통해 상당한 투자를 해 왔다[16]. 자체 개발한 프로세서인 SW26010을 격자구조의 대규모 멀티프로세서로 구성한 Taihulight라는 슈퍼컴(Top 500 Supercomputer list에서 1위를 차지)을 개발하였으며, CAS의 ICT(Instutite of Computing Technology)

에서는 인퍼런스 가속기인 DianNao, DaDianNao, ShiDianNao를 개발하였다. 중국의 Cambricon Technologies 스타트업이 개발한 Cambricon-X는 6.38㎟의 반도체 면적에서 544GOPS의 성능을 내고[5], Sparse matrix 가속 성능이 있다. 중국 Huawei의 스마트폰 내에 있는 Kirin 970 프로세서 내에서 Cambricon-X는 NPU IP로 활용되고 있다.

인텔은 인공지능을 위한 반도체 개발을 위해 인수합병을 통해 매우 다양한 시도를 하고 있다. Movidius의 Myriad, Nervana 학습용 AI 프로세서(NNP-T)와 추론용 AI 프로세서(NNP-I) 2종의 Nervana AI 프로세서를 공개하였다[20]. NNP-T는 Nervana가 2년간 'Lake Crest'라는 코드명으로 개발한 1세대 기술 후속인 스프링 크레스트(Spring Crest) 기반으로 TSMC의 16nm 공정에서 설계되었다. NNP-T는 Bfloat16 데이터 유형을 지원하여 최대 108TOPS 성능을 보여주었고, 4개의 32GiB급 HBM2 스택과 함께 PCIe Gen3 및 OCP OAM 가속기 카드 2가지의 폼팩터를 제공하였다.

인텔은 2019년 12월에 데이터센터용 AI 프로세서 강화를 위해 이스라엘 AI 반도체 스타트업인 Habana Labs를 인수하여 AI 추론 프로세서 고야(Goya) 및 학습용 AI 프로세서 가우디(Gaudi)를 출시하였다[21]. 가우디 칩은 3배 뛰어난 학습 연산 성능을 보였고 640개 가우디 프로세서 기반 서버 시스템은 ResNet-50 학습 연산처리 기준으로 640개의 Nvidia V100 기반 시스템보다 3.8배 높은 처리성능을 나타냈다. 이는 대규모 HLS-1 기반 클러스터가 Nvidia의 DGX-1 AI 서버 시스템의 처리량보다 3.8배 높은 성능을 보여줌을 말한다.

가우디는 텐서 프로세서 코어(TPC), GEMM 및 DMA의 3가지 이기종 컴퓨팅 구조기반으로 동작을 하며, FP32, BF16, INT32, INT16, INT8, UINT32, UINT16 및 UINT8 등 다양한 혼합 정밀 데이터 유형을 지원하여 높은 연산 성능을 보여준다. 데이터센터 확장성을 위해 8개의 100GB 이더넷을 지원하면서 이더넷 네트워크를 통한 원격 직접 메모리 액세스 기술인 RDMA over Converged Ethernet(ROCE V2)을 지원하였다. RoCE는 학습과정에서 필요한 프로세서 간 통신에서 최대 2Tb/s의 양방향 처리량을 지원한다.

인텔은 결국 2020년 2월에 Nervana AI 프로세서 개발 중단을 발표하였고, 가우디, 고야로 데이터센터와 클라우드 시장에 집중하려는 전략을 펼치고 있다.

Xeon Phi 등과 같이 68개의 x86 CPU를 한 개의 반도체 칩에 집적한 제품을 개발하였지만, 300Watt 이상의 소모전력으로 많은 활용처를 찾지 못하고 있다.

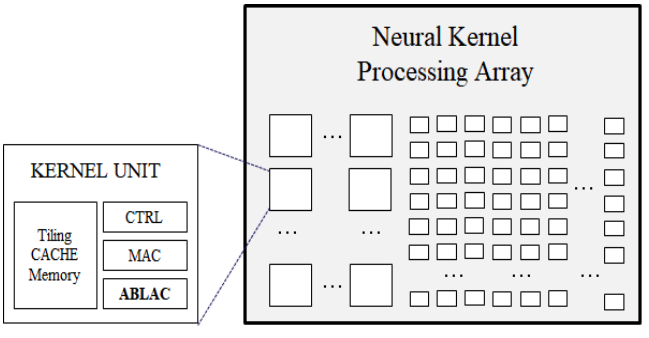
인공지능 반도체에 대한 관심이 증대되면서 글로벌 기업들은 매우 다양한 제품들을 발표하였고, 국내에는 UX factory, Furiosa A.I., Mobiliant 등의 스타트업이 인공지능 알고리즘이 요구하는 높은 연산성능을 내기 위한 독창적인 구조로 인공지능 반도체를 개발하고 있다.

IV. 국내 인공지능 프로세서 개발

1. VIC

VIC은 한국전자통신연구원에서 시각지능 AI 알고리즘의 저전력 고속처리를 위해 개발되었다. 저전력 동작을 위한 아날로그 맥(MAC) 연산기를 포함한 Neural Network Processing 부분, AI Algorithm 처리 부분, 외부장치 연결제어 및 애플리케이션 처리 부분으로 구성된다.

신경망의 주요 연산을 담당하는 Neural Network Processing 부분은 [그림 3]과 같이 신경망의 대량 커널 연산을 수행하는 병렬 어레이 구조인 Neural Kernel Processing Array(Kernel PA)를 기반으로 한다. AI Algorithm 처리 부분은 로열티 프리인 RISC-V CPU로 AI 알고리즘의 다양한 변종을 처리할 수 있도록 구성하고, 애플리케이션 처리



〈자료〉 한국전자통신연구원 자체 작성

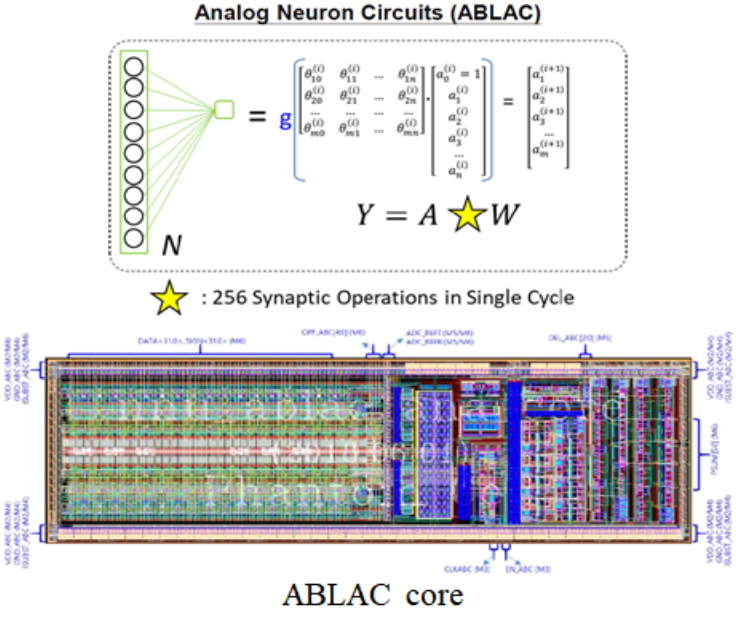
[그림 3] Neural Kernel PA 구조

부분도 RISC-V CPU로 구성하였다. VIC 칩은 고속 USB3 인터페이스, 대용량 외부 메모리와의 고속 병렬 인터페이스, 칩 제어를 위한 I2C, UART 인터페이스 등을 지원한다.

신경망 연산의 핵심을 담당하는 Neural Network Processing 부분을 좀 더 자세히 들여다보면, Kernel PA, 메모리와 메모리 컨트롤러, 고속 데이터 전송을 담당하는 Neural Network Direct Memory Access Controller(NDMAC)와 신경망 연산기 전용 캐시 기능을 담당하는 NCU(Neural Cache Unit)를 포함한다.

Kernel PA는 신경망의 대량 커널 연산을 처리하는 병렬 Kernel Unit들로 이루어져 있다. Kernel Unit은 로컬 메모리인 Tiling Cache Memory와 MAC 연산기들로 구성되어있다. 저전력 MAC 연산 동작을 위해 [그림 4]와 같이 아날로그 신경망 연산기인 Analog Basic Linear Algebra Circuit(ABLAC)을 개발하였다. ABLAC는 2.36pJ(1.21mW, 512MSOP/s)의 저전력 동작 성능을 보인다. 저전력 동작을 위한 ABLAC 연산기와 고속 모드를 위한 디지털 MAC 연산기를 공통으로 적용한 아날로그/디지털 혼종의 MAC 연산기 구조를 최종 채택하여 Kernel Unit을 설계했다.

Kernel Unit의 MAC은 저전력 동작을 위해 Sparse 신경망 처리 기능을 지원한다.



〈자료〉 한국전자통신연구원 자체 작성

[그림 4] ABLAC core

신경망의 웨이트 중 제로(“0”)인 웨이트에 대한 연산을 회피하여, 고속 저전력 MAC 연산을 가능하게 하는 Sparse 처리 기능을 MAC 연산기에서 제공한다. 모든 웨이트 연산을 처리하는 ‘Dense’ 연산 방식이 유리한 신경망에 대비하여 Sparse/Dense 연산을 동시에 지원하는 MAC 연산기를 개발하였다. 신경망에 따라 유리한 MAC 연산 모드로 고속, 저전력으로 동작시킬 수 있는 장점이 있다.

〈자료〉 한국전자통신연구원 자체 작성

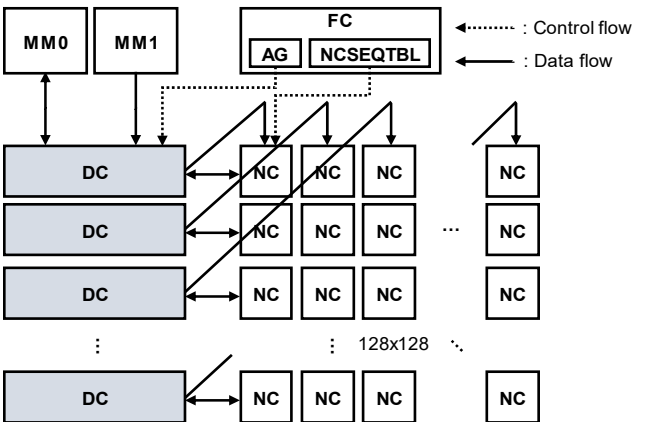
[그림 5] VIC

VIC 칩은 신경망의 웨이트나 입출력 데이터의 일반적 구조인 3D 텐서 형태의 데이터를 외부 메모리로부터 고속 전송하는 NDMAC를 포함한다. 3D 텐서 구조의 데이터는 보통 연속된 메모리 주소에 위치하지 않는데, 이를 각각 분리된 메모리 전송 명령으로 처리하는 일반적 DMAC로 처리할 경우, 연속된 데이터 전송과 비교하면, 대역폭이 현저히 떨어진다. 이를 해결하기 위해 3D 텐서 구조의 데이터에 대한 전송을 일괄처리할 수 있는 NDMAC를 개발하여 데이터 전송 대역폭을 개선하였다. 메모리의 데이터 전송 병목 현상을 해결하기 위한 신경망 연산기 전용 캐시 기능을 담당하는 Neural Cache Unit을 포함한다. 최신 신경망의 다양한 콘볼루션 커널에 대한 처리를 검증하기 위해 SSD, ResNet, MobileNet, Inception, MobileNet, GoogLeNet 등 다양한 신경망으로 VIC 칩을 검증했다.

VIC 칩은 TSMC 40㎚ 공정으로 제작하였다. 커스텀 레이아웃 설계한 PLL, ALBAC을 포함하여 전체 크기는 5.5×5.5㎟이다. 전체 게이트 카운트는 17,551,342 규모이고, 소비전력 최적화를 위해 Multi-VT 기술을 적용하여 제작하였다([그림 5) 참조)].

2. AB9

AB9은 한국전자통신연구원에서 개발한 AI 알고리즘을 가속하기 위한 인공지능 프로세서로서, Convolutionary Layer의 처리 속도를 향상시키기 위한 Matrix 연산 가속기인 Super Thread Core(STC)와 이를 제어하고 Pre-processing, Post-processing을 위한 SPARC Instruction Set Architecture기반의 General Purpose CPU인 쿼드코어로



〈자료〉 한국전자통신연구원 자체 작성

[그림 6] STC 아키텍처

구성된 알데바란 프로세서로 구성된다[19]. STC는 [그림 6]과 같이 32MB의 Data Control(DC) Memory와 Nano Core(NC)로 구성된 Systolic Array(SA)로 구성되어 있다. SA는 128×128의 NC로 구성이 되어 Deep Neural Network를 위한 병렬 연산을 하게 되고, SA를 위한 웨이트와 IFM(Input Feature Matric)을 공급하는 역할을 DC 메모리가 담당하게 된다. 그리고 연산된 결과는 다시 32MB의 DC 메모리에 저장을 하게 된다. 그리고 MM0, MM1은 외부 메모리로부터 필요한 웨이트와 Input Feature Matric (IFM)을 읽어오고, 연산된 결과인 Output Feature Matric(OFM)을 저장하는 역할을 한다. 그리고 Flow Control(FC)은 외부 메모리에 저장된 STC를 위한 명령어를 읽어와 웨이트, IFM, 그리고 OFM을 위한 저장 주소를 제어하거나, NC를 위한 명령어를 NC Sequence Table(NCSEQTBL)에 저장하고, 이를 NC에 전송하는 역할을 하게 된다.

SA를 구성하는 NC는 최대 1.25GHz로 동작하며, 16-bit floating-point Data Type으로 연산을 한다. 이러한 SA는 128×128 NC로 구성이 되어 있고, 모두 동작을 할 경우 최대 40TFLOPS의 성능을 가진다. NC는 16-bit floating-point multiply, add, comparison, max 연산을 지원한다. 또한, SA는 동작하지 않을 때는 Power Gating (PG) 기능을 통해 대기 전력 소모를 차단한다. 이때, SA의 Power Domain을 16개로 나누어 병렬적인 PG 제어가 가능하도록 설계함으로써 전력 공급/차단 시의 지연시간을 최소화한다.

DC Memory는 32MB 크기의 내부 SRAM과 이의 제어를 위한 로직들로 구성되어 있다. SA의 행 개수와 동일하게 128개의 행으로 이루어져 있고, 각 행은 8개의 독립적인 256KB SRAM 뱅크들로 구성되어 있어, 128×8개의 읽기/쓰기를 병렬적으로 수행할 수 있다. FC의 Address Generation(AG)으로부터 IFM과 웨이트 주소가 전달되면, 모든 행의 DC 메모리는 해당 위치의 데이터를 NC들에게 공급한다. DC로부터 읽혀 나온 IFM이 좌하향의 NC에 전달된다면, 웨이트는 feed-through path를 거쳐 우상향의 NC에 전달된다. IFM과 웨이트가 모두 단일 DC에 저장되므로 효율적으로 사용할 수 있다.

MM0와 MM1은 256비트의 읽기/쓰기를 지원하는 Direct Memory Access 기능을 함으로써 외부 LPDDR4/PCIe와 DC 간 인터페이스를 담당한다. MM0는 외부로부터 읽어 들인 IFM을 DC에 저장하거나, DC에 저장된 출력 데이터를 다시 외부로 전송한다. MM1은 외부로부터 웨이트만을 읽어 들여 DC에 저장한다.

FC는 NCSEQTBL과 AG로 구성된다. NCSEQTBL(NC Sequence Table)은 32비트의 NC 명령어를 1,024개까지 저장할 수 있는 FIFO 구조로 이루어져 있으며, NC 명령어는 NC까지 5단 파이프라인(pipeline)을 거쳐 전달된다. NC 명령어를 통해 각 NC의 다양한 연산기를 재구성할 수 있어 CNN(Convolutionary Neural Network), FCN(Fully-Connected Network), LSTM 등 다양한 종류의 Deep Neural Network에 필요한 연산을 가속할 수 있다. DC 주소는 DNN의 Tiled 연산을 위해 다양한 Dimension 연산을 지원하며, 폭(Width), 높이(Height), 깊이(Depth)에 대한 총 7가지 조합을 지원한다. 이러한 구성은 AG(Address Generation)를 통해 7차원의 네스트 루프(nested loop)로 구성될 수 있으며, 시작 주소, offset, 루프 수행 횟수 등의 parameter에 의해 주소가 생성된다.

TSMC 28nm 공정에서 제작된 칩의 Layout은 [그림 7]과 같이 면적은 19×26㎟이며, Gate Count 수는 약 2.85억 개에 달한다. 1V 동작전압, -40~125도 동작 온도에서 최대 1.25GHz로 동작하며, power/ground를 포함하여, 1,599개

〈자료〉 한국전자통신연구원 자체 작성

[그림 7] AB9

의 IO Pin을 가지는 칩이다([그림 7] 참조).

V. 결론

병렬 컴퓨팅은 암달의 법칙에 의한 한계가 있지만, 인공지능 알고리즘에서 요구하는 높은 연산 성능을 달성하기 위한 인공지능 프로세서의 기본 설계 방향이 되고 있으며, SIMT(Single Instruction Multi Thread) 기반의 구조와 달리 Systolic Array의 구조로 Operational Intensity를 높여 주어진 외부 메모리 대역폭에서 높은 연산성능을 내는 구조를 달성하고 있다. 차세대 인공지능 프로세서는 더 높은 연산 성능을 요구하는 학습 연산 성능 향상을 위한 구조로 연구가 되고 있고, 학습을 위한 연산 성능을 달성하기 위해서는 반도체의 한계로 인해 단위 전력 당 더 높은 연산 성능을 요구하는 구조로 가야 할 것이다.

[참고문헌]

[1] David Patterson, “50 Years of Computer Architecture: From the Mainframe CPU to the Domain-Specific TPU and the Open RISC-V Instruction Set,” ISSCC 2018.
[2] Norman P. Jouppi et al., “In-Datacenter Performance Analysis of a Tensor Processing Unit,” In Proceedings of the 44th Annual International Symposium on Computer Architecture, 2017.
[3] Andrei Frumusanu, “The Apple iPhone 11, 11 Pro & 11 Pro Max Review: Performance, Battery, & Camera Elevated,” anandtech.com, October 16, 2019,
[4] Ignatov, Andrey, et al. “AI Benchmark: All About Deep Learning on Smartphones in 2019,” arXiv preprint arXiv:1910.06663(2019).
[5] www.samsung.com/semiconductor/minisite/exynos/products/mobileprocessor/exynos-990/
[6] Song, Jinook, et al. “7.1 An 11.5 TOPS/W 1024-MAC butterfly structure dual-core sparsity-aware neural processing unit in 8nm flagship mobile SoC,” 2019 IEEE International Solid-State Circuits Conference-(ISSCC). IEEE, 2019.
[7] consumer.huawei.com/en/campaign/kirin-990-series/

* 본 논문은 과학기술정보통신부, IITP에 의해 지원받은 인공지능프로세서 전문연구실(과제번호 2018-0-00195) 과제를 통해 이루어 졌습니다.
** This research was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2018-0-00195, Artificial Intelligence Processor Research Laboratory)

[8] Heng Liao et al., “DaVinci: A Scalable Architecture for Neural Network Computing,” Hot Chips Conference 2019.
[9] Sophia Windsor, “Snapdragon 865 vs Kirin 990 5G vs Exynos 990(Exynos 9830) vs MediaTek Dimensity 1000(MT6889): which one is the best 5G processor?,” Dec. 10. 2019.
[10] www.arm.com/products/silicon-ip-cpu/ethos/ ethos-n77, n57, n37
[11] www.ceva-dsp.com/product/ceva-neupro/
[12] www.gyrfalcontech.ai/solutions/2801s, 2801s
[13] Orr Danon, “Introducing Hailo-8: The Most Efficient Deep Learning Processor for Edge Devices,” 2019 Embedded Vision Summit, May 2019.
[14] 권영수, “인공지능 프로세서 기술 동향“, ETRI, 전자통신동향분석 33권 5호, pp.121-134.
[15] E. Lindholm et al., “NVIDIA Tesla: A Unified Graphics and Computing Architecture,” IEEE Micro, Vol.28, No.2, 2008, pp.39-55.
[16] nvidia.com.
[17] Andrew Yang, “Deep Learning Training At Scale Spring Crest Deep Learning Accelerator (Intel Nervana NNP-T),” Hot Chips Conference 2019.
[18] Eitan Medina, “habana,” Hot Chips Conference 2019.
[19] Y. Kwon et al., “Function-Safe Vehicular AI Processor with Nano Core-In-Memory Architecture,” In Proceedings of the 1st Annual International Conference on Artificial Intelligence Circuits and Systems, 2019.