인공지능프로세서 기술 동향

Trends in AI Processor Technology

이미영 (M.Y. Lee, sharav@etri.re.kr) 인공지능프로세서연구실 책임연구원
정재훈 (J. Chung, jchung@etri.re.kr) 인공지능프로세서연구실 연구원
이주현 (J.H. Lee, juehyun@etri.re.kr) 인공지능프로세서연구실 책임연구원
한진호 (J.H. Han, soc@etri.re.kr) 인공지능프로세서연구실 책임연구원/실장
권영수 (Y.S. Kwon, yskwon@etri.re.kr) 지능형반도체연구본부 책임연구원/본부장

ABSTRACT

As the increasing expectations of a practical AI (Artificial Intelligence) service makes AI algorithms more complicated, an efficient processor to process AI algorithms is required. To meet this requirement, processors optimized for parallel processing, such as GPUs (Graphics Processing Units), have been widely employed. However, the GPU has a generalized structure for various applications, so it is not optimized for the AI algorithm. Therefore, research on the development of AI processors optimized for AI algorithm processing has been actively conducted. This paper briefly introduces an AI processor especially for inference acceleration, developed by the Electronics and Telecommunications Research Institute, South Korea., and other global vendors for mobile and server platforms. However, the GPU has a generalized structure for various applications, so it is not optimized for the AI algorithm. Therefore, research on the development of AI processors optimized for AI algorithm processing has been actively conducted.

KEYWORDS AI Processor, Neural Network Processor, Systolic Array

Ⅰ. 서론

사용자들은 다양한 플랫폼에서 높은 정확도를 가진 AI 서비스(자율주행, 음성인식 등)가 제공되기를 기대한다. 정확도를 올리고 서비스를 고도화하기 위해서 AI 알고리즘의 복잡도는 나날이 증가하고 있어서, 이를 효율적으로 처리할 수 있는 프로세서에 대한 요구가 커지고 있다.

AI 알고리즘은 대량의 연산이 반복되는 구조로, GPU(Graphics Processing Unit)와 같은 병렬처리에 특화된 프로세서가 가속기로 현재 널리 사용되고 있다. 그러나 다양한 응용에 적용할 수 있도록 일

* DOI: https://doi.org/10.22648/ETRI.2020.J.350307
* This work was supported by the ICT R&D program of MSIT/IITP[2018-0-00195, Artificial Intelligence Processor Research Laboratory].

반화된 구조를 가지는 GPU는 AI 알고리즘 처리에 필요하지 않은 블록이나 데이터 패스를 포함하고 있어 성능개선의 여지가 있고 전력 면에서도 불필요한 소모가 있다. 특히 AI 알고리즘은 연산 과정에서 필수적인, 대량의 데이터 전송에 드는 시간이 전체 동작 성능을 좌우해서 AI 알고리즘에 최적화된 고속 데이터 전송 구조 설계가 필수적이다. 따라서 AI 알고리즘 가속에 최적화된 연산 처리와 고속데이터 전송 구조를 가진 전용 인공지능 프로세서 개발이 필요하다.

한편, 플랫폼의 구조적 특성 차이로 성능 최적화 방법이 달라진다. 서버 환경에서는 고성능 요구를 최우선으로 만족해야 하고, 전력이나 폼팩터(form factor)의 제약이 크지 않아 HBM(High Bandwidth Memory) 등 고속데이터 전송 메모리를 도입할 수 있다. 모바일 환경에서는 저전력 요구를 만족해야 하고, 폼팩터의 제약이 크므로 HBM 등을 사용하기 어렵다. AI 알고리즘 자체를 압축하는 방식 등을 통해 메모리 대역폭 요구량 자체를 줄이는 기술, 연산량을 줄이기 위해 희소성(Sparsity)을 활용하는 연산기 기술, 연산기 비트 수를 줄이기 위한 낮은 비트 해상도(Bit precision) 변환 기술 등이 연구되고 있다.

본 고는 AI 알고리즘 중 추론(Inference)의 가속을 위한 인공지능 프로세서들을 위주로 소개한다. Ⅱ장에서는 서버와 모바일 분야의 인공지능 프로세서 기술 동향을 소개하고, Ⅲ장에서 한국전자통신연구원의 VIC(저전력)/AB9(고성능) 칩을 설명한다.

Ⅱ. 인공지능프로세서 기술

AI 알고리즘의 대량의 연산량을 효율적으로 처리하기 위해 기존의 프로세서와는 다른 새로운 구조의 인공지능프로세서 연구가 활발히 진행되고 있다. 모바일/서버 플랫폼에서 각각 저전력/고성능을 타겟으로 한 인공지능프로세서 개발 현황을 소개한다.

1. 모바일 인공지능프로세서

가. 모바일 인공지능프로세서 개발 현황

스마트폰 회사들이 2019년 발표한 모바일 AP들은 다수의 혼종 CPU 코어와 GPU 이외에 인공지능프로세서인 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 특징을 요약한 테이블은 표 1과 같다.

화웨이 Kirin 칩은 3D 텐서(Tensor) 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 사이클당 1개의 FP16 연산이나 2개의 INT8 연산을 수행한다. DaVinci 코어는 MAC 연산기 큐브 이외에 스칼라 ALU, 벡터 ALU, load/store 유닛 등을 포함한다[8].

삼성 Exynos 990에 적용된 NPU의 구조는 ISSCC2019 논문에서 찾아볼 수 있다. NPU 제어기와 2개의 NPU 코어로 구성되고, NPU 제어기는 CPU, DMA, SRAM, 네트워크 제어기를 포함한다. NPU는 1,024개 MAC 연산기로 구성되며, 웨이트(Weight)의 희소성을 활용하여 필요한 연산만을 수행할 수 있는 NPU 구조를 제안했다. Inception-v3 신경망으로 3.4 TOPS/W 결과를 보였다[9].

표 1 모바일 AI processor 비교

AI chips	Key Features
Apple A13 Bionic	2x Lightning cores @2.66GHz 4x Thunder cores @1.728GHz 4x new GPU NPU: 8-core Neural Engine (*A12 NPU: 100 GOPS/W [1]) 7nm TSMC process LPDDR4X 출시일 09.10.2019 [2]
Huawei Kirin 990 5G	Tri-Cluster Octa-Core 2x Cortex-A76 @2.86GHz 2x Cortex-A76 @2.36GHz 4x Cortex-A55 @1.95GHz Mali-G76 MP16, 700MHz NPU: 2 + 1 DaVinci NPU 〉16 TOPS(Asend 310으로 inference) TSMC 7nm+ EUV FinFET LPDDR4X @ 2133MHz 출시일 09.26.2019 [3-5]
Samsung Exynos 990	2x Mongoose 5th gen 2x Cortex-A76 4x Cortex-A55 Mali-G77 MP11 Dual NPU + DSP 15 TOPS 7nm EUV process LPDDR5 @2750MHz 출시일 10.24.2019 [6,7]
Qualcomm Snapdragon 865	1x 2.84GHz(Cortex A77) 3x 2.4GHz(Cortex A77) 4x 1.8GHz(Cortex-A55) Adreno 650@587MHz Hexagon 698(DSP) Hexagon Tensor Accelerator(DSP) 15 TOPS 7nm process LPDDR5 @2750MHz 출시일 12.04.2019 [7]
MediaTek Dimensity 1000	4x 2.6GHz(Cortex A77) 4x 2GHz(Cortex A55) Mali-G77 MP9 Hexa-core APU (2x heavy cores, 3x medium cores, 1x light core) 4.5 TOPS 7nm process LPDDR4X@1866MHz 출시일 11.26.2019 [7]

나. 저전력 AI 프로세서 개발 현황

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

DSP IP가 주력인 CEVA사는 인공지능프로세서 NeuPro-S를 발표했다. 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[11].

Gyrfalcon사는 매트릭스 연산 전용 엔진을 구현한 Ligthspeeur 2801, 2803을 출시했다. 168×168 MAC 연산기로 구성된 매트릭스 연산 엔진을 포함하며, 300mW의 저전력으로 2.8 TOPS의 성능으로 9.3 TOPS/W 높은 에너지 효율 결과를 발표했다.

표 2 Low Power AI chips

Low power AI chips	Key Features
ARM Ethos-N (37/57/77)	512/1024/2048 MAC(8x8) 1/2/4 TOPS INT8 and INT16 [10]
CEVA NeuPro-S	4096 8x8 MAC 12.5 TOPS INT8 and INT16 Memory: DDR [11]
Gyrfalcon Ligthspeeur 2801s	168x168 MAC 2.8 TOPS, 300mW 9.3 TOPS/W [12]
Hailo Hailo-8	cluster(8 core) 26 TOPS, 1.7W(ResNet-50) INT8 and INT16 Memory: SRAM(32MB)[13,14]

PIM(Processing In Memory) 구조로 설계하여, 전력을 많이 소모하는 외부메모리로부터의 데이터 전송을 없애서 저전력으로 동작할 수 있도록 설계하였다[12].

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(656 FPS, 32W) 대비 1/15 전력으로 동등한 신경망 수행능력을 보였다[14].

2. 서버용 인공지능프로세서

가. NVIDIA GPU

GPU는 Streaming Processor(SP)를 SIMT(Single Instruction Multiple Thread) 구조로 엮은 SM들을 다시 n개씩 묶어 TPC(Texture/Processor Cluster)로 구성한다[15,16]. TPC, SM들을 구성하는 방식 및 개수에 따라 GPU 모델이 결정된다.

각각의 SP는 32, 64비트 floating-point와 integer 연산 등을 지원한다. 또한, 분기(Branch)로 인한 성능 저하 완화를 위해 분기 예측기 등이 포함되어 있으며, 계층적인 cache 구조에 따른 coherence protocol이 필요하다.

하지만 CNN, MLP 등의 AI 알고리즘은 일정한 데이터 흐름을 가지고 있어 분기 예측기나 cache를 사용함에 따른 효용성이 미약하며, 추론을 위한 연산자들은 단일 데이터 타입이 사용되므로 AI 알고리즘 가속에 있어 GPU는 성능 및 전력 소모 측면에서 비효율적이다.

나. Google TPU

2017년 ISCA 학회에서 Google은 추론 가속을 주목표로 개발한 TPU v1을 발표하였다. 이후 학습까지 가속할 수 있는 TPU v2와 v3가 순차적으로 공개되었다. 각 버전에 따른 주요 차이점은 표 3과 같다. 본 고는 추론 가속용 인공지능 프로세서인 TPU v1 위주로 소개하고자 한다.

표 3 TPU 버전에 따른 특징 비교[17]

Feature	TPUv1	TPUv2	TPUv3
Floating-Point Unit	X	O	O
Clock(MHz)	700	700	940
TDP(Watts)	75	280	450
Die Size(mm^2)	〈331	〈661	〈648
Chip Technology	28nm	〉12nm	〉12nm
On-chip memory	28MB	37MB	37MB
주 가속 목표	추론	학습	학습

TPU v1은 AI 알고리즘 연산에 필수적인 로직들만 설계하고, 데이터 흐름을 고려한 연산 유닛 배치를 통해 GPU 대비 30배 이상의 TOPS/W를 달성하였다.

TPU v1의 구조는 Google에서 발표한 논문에서 확인할 수 있다[18]. 추론을 위한 연산 유닛들로 Matrix Multiply Unit, Activation, Accumulator, Normalize/Pool 유닛을 포함한다. 피쳐맵(Feature Map)과 웨이트를 저장하는 내부 메모리와 외부 IP와의 통신 인터페이스, 내부 제어 로직 등으로 구성된다.

Matrix Multiply Unit(MMU)은 700MHz의 클럭 주기로 8비트 행렬(Matrix) MAC 연산을 수행할 수 있는 256×256개의 Processing Element(PE)들이 systolic array 형태로 배치되어 있다. MMU의 인접한 위치에 내부 메모리가 배치되어 있어 피쳐맵 및 웨이트를 PE들에 매 사이클(cycle) 공급한다. PE들은 주어진 연산자에 대해 연산을 수행함과 동시에 바로 인접한 PE들에게 그 연산자들을 넘겨주게 된다. 이때 서로 다른 연산자가 각각 전달되는 방

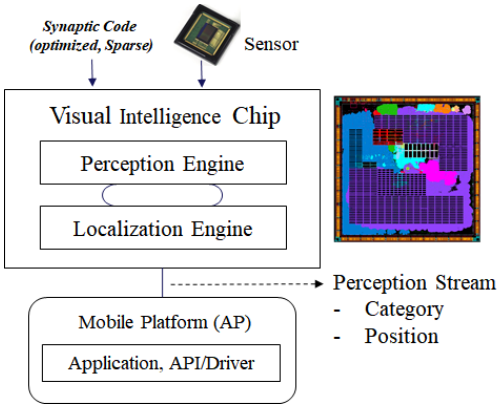


그림 1 시각지능칩 개요

향은 고정되어 있다(예를 들어 피쳐맵은 왼쪽에서 오른쪽, 웨이트는 위에서 아래 방향으로만 전달). 연산이 끝나면 그 결과물은 아래 방향의 TPU의 Accumulator로 전달되어 순차적으로 더해지며, AI 알고리즘에 따라 ReLU, Max Pooling 등의 함수 연산이 TPU의 Activation, Normalize/Pool에서 수행된 후, 최종 결과물이 다시 Unified Buffer에 저장된다.

MMU 구조상 같은 행, 열에 위치한 PE들은 동일한 연산자에 대해 데이터 복사(Copy) 없이 MAC 연산을 수행할 수 있으므로 operational intensity를 높이고, 내부 메모리의 용량(Capacity)을 최대한 활용할 수 있다. 또한, 오직 integer 연산만을 지원하기 때문에 FPU(Floating-Point Unit)로만 구성된 로직과 비교해 상대적으로 낮은 전력과 면적으로 높은 성능을 보일 수 있다. 하지만 대부분의 AI 알고리즘은 floating-point 타입으로 학습(Training)된다. Floating-point 타입으로 학습된 AI 알고리즘의 추론을 TPU상에서 가속하기 위해선 양자화를 통해 integer형으로의 변환이 필요한데, 이 과정에서 정확도 손실(Accuracy loss)이 발생한다. 이는 자율주행 등 신뢰도가 중요시되는 영역에서 치명적인 약점이 될 수 있다.

III. ETRI 저전력 인공지능프로세서

1. 시각지능칩 개요

한국전자통신연구원(ETRI)의 인공지능프로세서 연구실에서는 저전력 시각지능 기반 고속 추론을 위한 전용 칩 개발 과제를 수행하여 2019년 12월 VIC(Visual Intelligence Chip)을 발표했다. 센서나 경량단말에서의 사람 수준의 시각지능 추론 기능을 담당할 수 있는 칩을 목표로 개발된 칩으로, 범용 객체에 대한 인지(Recognition)를 수행하고 위치추출 결과(Localization)를 출력할 수 있다.

2. 시냅스컴파일러

VIC은 저전력 고속 동작을 위해서 신경망의 웨이트 중 제로(Zero) 값을 연산에서 제외할 수 있는 기능이 있다. 이를 최대로 활용하기 위해서 신경망을 프루닝(Pruning)하고 재학습하는 과정으로 희소성(Sparsity)을 이끌어내는 시냅스컴파일러(Synapse Compiler)를 개발했다. 시냅스컴파일러는 신경망

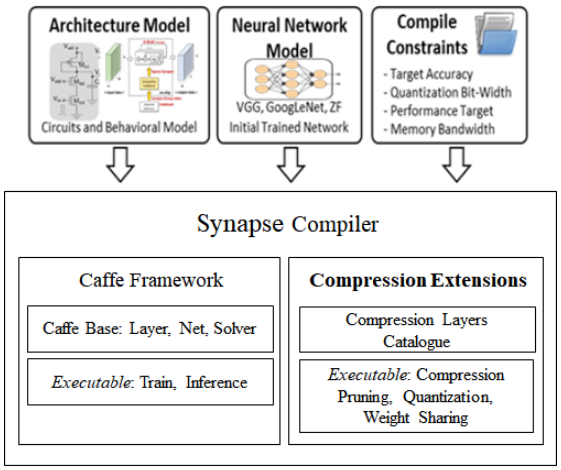


그림 2 시냅스컴파일러 구성

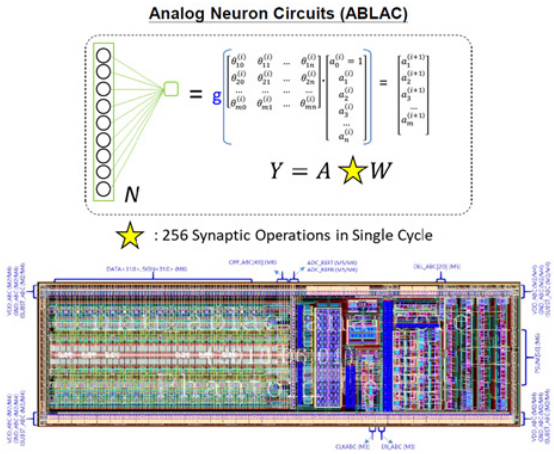


그림 3 ABLAC core

압축 효율을 올리기 위해 양자화, 웨이트그룹화, 레이어 퓨징(Fusing) 기능 등을 포함한다. 오픈소스 딥러닝 프레임워크인 Caffe[19] 프레임워크를 기반으로 개발했다.

3. ABLAC: 아날로그 신경망 연산기

신경망 연산의 핵심인 MAC 연산기를 저전력으로 설계하기 위해서 아날로그 신경망 연산기인 ABLAC(Analog Basic Linear Algebra Circuit)을 개발했다. ABLAC은 2.36pJ(1.21mW, 512MSOP/s)의 저전력 동작 성능을 보인다. VIC은 저전력 ABLAC 연산기를 포함하여, 고속 동작을 위한 디지털 MAC 연산기를 적용한 아날로그/디지털 혼종의 MAC 연산기 구조를 채택했다.

표 4 ABLAC 주요 성능

ABLAC core	Specification
# of input	~256(Feature & Weight)
# of output	1
# of ABLAC	~512
Data Format	Input: 8bits(signed) Output: 24bits(signed)
Energy/SOP	2.36pJ(1.21mW, 512MSOPS/s) (IBM TrueNorth: 26pJ/SOP)
Area	0.005mm^2
Precision	~5%(Signed Arithmetic)
Technology	40nm

4. VIC 아키텍처

VIC의 내부 아키텍처 구조는 크게 3부분으로 구성되는데, Neural Network Processing 부분과 AI Algorithm 처리 부분과 외부장치 연결제어 및 부트(Boot) 제어를 담당하는 Application 부분이다. RISC-V 기반의 CPU 서브 시스템으로 Application 부분을 구성하고, 고속 데이터 전송을 위해 256비트, 32비트 계층의 버스구조를 사용했다. 신경망의 연산을 담당하는 Kernel PA(Neural Kernel Processing Array), 메모리 및 메모리컨트롤러, 메모리로부터의 고속데이터전송을 담당하는 DMAC(Direct Memory Access Controller)와 신경망 연산기 전용 캐시 기능을 담당하는 NCU(Neural Cache Unit)로 구성된다.

Kernel PA는 대량 병렬 Kernel Unit들로 이루어져 있다. 로컬 메모리인 Tiling Cache Memory와 ABLAC을 포함하는 아날로그/디지털 혼종 MAC 연산기로 구성되어 있다.

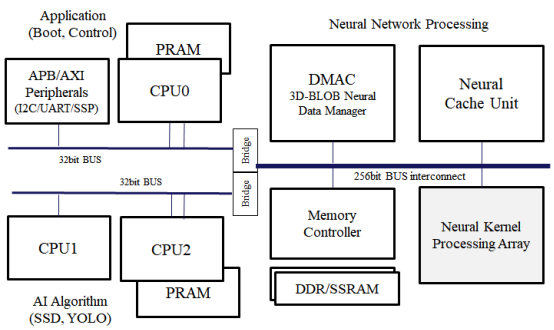
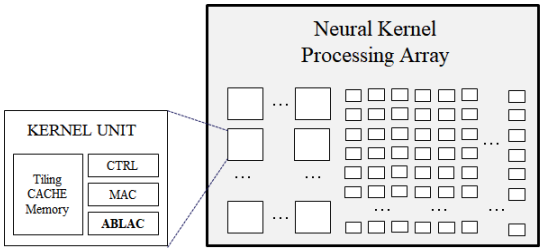


그림 4 VIC 내부 아키텍처 구조



Neural Kernel PA 특징
• zero skip을 포함한, sparse/dense 연산을 모두 지원하는 신경 연산기 Kernel Unit • 16 쓰레드(thread) 연산을 지원하는 병렬 대량 연산 Kernel PA • 메모리에 가상 3D tensor 구조 접근을 지원하는 DMAC, 최대 8개 채널 동시 지원 • 메모리의 데이터 전송 병목(bottleneck) 현상을 해결하는 전용 NCU

그림 5 ABLAC core Neural Kernel PA 구조

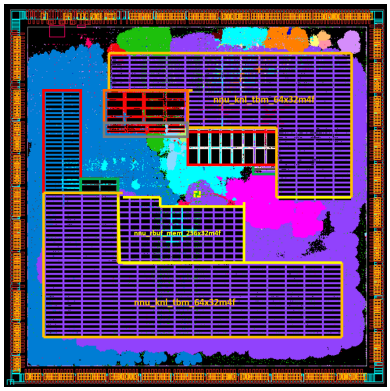


그림 6 VIC 칩 제작 결과

5. VIC 칩 제작 결과

시각지능칩을 TSMC 40nm LP 공정에서 칩 면적 5mm×5mm으로 제작하였다.

ABLAC을 포함하여 자체 설계한 PLL 회로를 포함하고, 외부 입력 클럭은 50MHz, 내부 동작 주파수는 최소 400MHz으로 제작되었다.

VIC칩의 동작 특징을 요약한 테이블은 표 5와 같다. SSD, ResNet, MobileNet, Inception, MobileNet, GoogLeNet 등 다양한 신경망으로 VIC칩을 검증했다. 표 5에 나열한 검증 신경망을 구성하는 다양한 레이어와 연산 커널 구조를 모두 지원한다.

표 5 VIC 동작 특징 요약

VIC	Key Features
Precision	INT8, INT16, ...
MAC	Analog/Digital Hybrid
Sparsity	Sparse/Dense Dual mode
Kernel PA	16–Thread parallel processing Kernel PA
CNN Kernel Processing	Various Kernel/Stride/Pad Size, Shape, Dilation
Neural Networks	VGG16, ResNet–101, ResNet–50, MobileNet, GoogLeNet, Inception–v3, VGG16–SSD300 (PASCAL VOC, MS–COCO), SqueezeNet

IV. ETRI 고성능 인공지능프로세서

1. AB9 개요

AB9은 AI 알고리즘을 가속하기 위한 플랫폼으로써 정규 알고리즘 처리용 프로세서인 STC와 그 외 비정규 알고리즘을 처리하기 위한 SPARC 아키텍처의 알데바란 CPU 코어가 쿼드코어로 구성된다[20]. STC에 충분한 메모리 대역폭을 제공해 주기 위해 170Gbps의 LPDDR4 채널 2개가 사용되며, 호스트 시스템과의 통신을 위한 PCIe x16 Gen3 인터페이스를 제공한다. 그 외 CNN 기반 AI 알고리즘을 위한 HDMI비디오 입출력, 데이터 저장을 위한 저장장치 및 제어기, 자율주행용 CAN 인터페이스 등의 주변 장치들로 구성된다.

2. STC 아키텍처

STC의 기본적인 구조는 그림 7과 같다. 크게

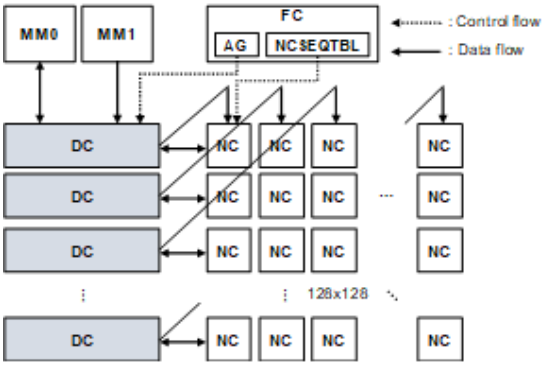


그림 7 STC 아키텍처

연산 유닛인 SA(Systolic Array), 데이터 저장을 위한 DC(Data Control), DMA를 위한 MM(Memory Mover)과 내부 제어를 위한 FC(Flow Control)로 구성되어 있다.

SA는 1GHz로 동작하는 16-bit floating-point MAC 연산용 NC(Nano-Core)들이 128×128의 systolic array 구조로 배치되어 있고, 총 32 TFLOPS의 성능을 가진다. 각 NC는 16-bit floating-point 곱셈(multiply), 덧셈기(add), 비교(compare), 최대값(max) 연산을 지원한다. 또한 SA가 idle 상태일 때

```
A_ADDR = S_ADDR
for(A=0; A<A_LOOP; A++)
    B_ADDR = A_ADDR;
    A_ADDR += A_INC;
    for(B=0; B<B_LOOP; B++)
        C_ADDR = B_ADDR;
        B_ADDR += B_INC;
        for(C=0; C<C_LOOP; C++)
            D_ADDR = C_ADDR;
            C_ADDR += C_INC;
            for(D=0; D<D_LOOP; D++)
                E_ADDR = D_ADDR;
                D_ADDR += D_INC;
                for(E=0; E<E_LOOP; E++)
                    F_ADDR = E_ADDR;
                    E_ADDR += E_INC;
                    for(F=0; F<F_LOOP; F++)
                        G_ADDR = F_ADDR;
                        F_ADDR += F_INC;
                        for(G=0; G<G_LOOP; G++)
                            ADDRESS = G_ADDR;
                            G_ADDR += G_INC;
```

그림 8 AG의 7차원 네스트 루프 예시

누설전력 차단을 위해 power gating 기능을 지원한다. 이때, SA의 power domain을 16개로 나누어 병렬적인 PG 제어가 가능하도록 설계함으로써 전력 공급/차단 시의 지연시간을 최소화한다.

DC(Data Control)는 피쳐맵과 웨이트를 저장하기 위한 32MB 크기의 내부 SRAM과 이의 제어를 위한 로직들로 구성되어 있다. SA의 행 개수와 동일하게 128개의 행으로 이루어져 있고, 각 행은 8개의 독립적인 256KB SRAM 뱅크(bank)들로 구성되어 있어, 서로 다른 뱅크에 대해 읽기/쓰기(read/write)를 병렬적으로 수행할 수 있다. FC로부터 피쳐맵과 웨이트 주소가 전달되면, 모든 행의 DC들은 해당 위치의 데이터를 NC들에 공급한다. 이때, 그림 7과 같이 각 DC마다 두 종류의 데이터가 공급되는데, 종류에 따라 전달되는 방향이 달라진다. 예를 들어 n번째 행의 DC로부터 읽혀 나온 피쳐맵이 (n, 1)에 위치한 NC에 전달된다면, 웨이트는 우상향의 feed-through path를 거쳐 (1, n)에 위치한 NC에 전달된다. 피쳐맵과 웨이트가 모두 단일 DC에 저장되므로 피쳐맵과 웨이트 간 data size 비율과 무관하게 내부 메모리를 효율적으로 사용할 수 있다.

MM은 256비트의 읽기/쓰기를 지원하는 DMA(Direct Memory Access) module로써 외부 LPDDR4/PCIe와 DC 간 interface를 담당한다. MM0는 외부로부터 읽어 들인 피쳐맵을 DC에 저장하거나, DC에 저장된 출력 데이터를 다시 외부로 전송한다. MM1은 외부로부터 웨이트만을 읽어 들여 DC에 저장한다.

FC는 외부로부터 32비트 읽기 전용 인터페이스를 통해 전달받은 NC 명령어를 저장하기 위한 NCSEQTBL과 DC 읽기/쓰기를 위한 주소를 생성해 주는 AG로 구성된다. NCSEQTBL은 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 내부 라우팅 경로(routing path)를 재구성할 수 있어 다양한 종류의 AI 알고리즘(convolution, fully-connected, LSTM 등)에 대응할 수 있다. 콘볼루션 연산에 필요한 DC 주소의 종류는 입출력 피쳐맵, 웨이트 각각의 폭(Width), 높이(Height), 깊이(Depth)를 고려하면 총 7가지이다. 이를 고려하여 AG는 그림 8과 같이 7차원의 네스트 루프(Nested loop)로 구성되며, 루프들이 순차적으로 수행되면서 7종류의 주소가 매 사이클 동시에 생성된다. 각 차원에서는 3가지 변수(시작 주소, offset, 루프 수행 횟수)에 의해 주소 생성 패턴이 결정되는데, 각 변수는 모두 programmable하기 때문에 현존하는 CNN 기반 AI 알고리즘 대부분을 지원할 수 있다.

3. AB9 칩 제작 결과

TSMC 28nm 공정에서 칩을 제작하였으며, die의 모습은 그림 9와 같다. 패키지 크기는 33×33mm^{2}이다(die 면적은 19×26mm^2). AB9의 총 gate count 수는 약 2.85억 개, 동작 전압은 1V, 동작 온도는 −40~125℃이며, power/ground를 포함한 패키지의 primary IO pin의 총 개수는 1,599개이다.

V. 결론

높은 정확도에 대한 요구로 복잡해지고 있는 AI 알고리즘의 거대한 연산량을 빠르면서도 효율적으로 처리하기 위한 프로세서의 필요성이 증대하고 있다. 이를 위해 모바일/서버 등 AI 서비스가 적용되는 플랫폼의 특성을 고려한 인공지능프로세서 연구가 활발히 진행되고 있다.

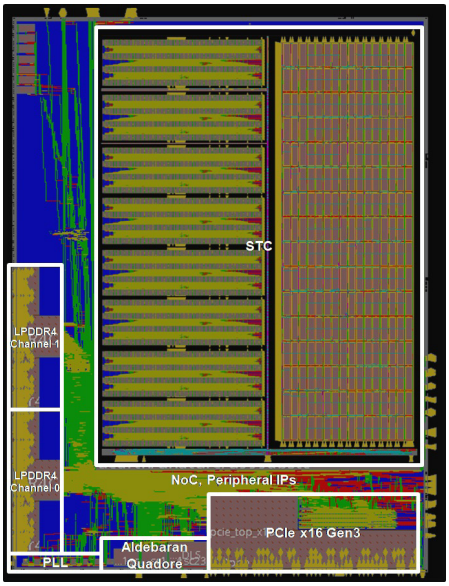


그림 9 AB9 die microphotograph

최근 모바일용 AP는 CPU와 GPU 이외에 인공지능프로세서인 NPU(Neural Processing Unit)를 내재하는 경우가 많아지고 있다. 각 NPU는 저전력으로 MAC 연산을 가속할 수 있는 구조를 가진다. 그에 더하여 ETRI의 VIC 칩은 정확도 손실을 최소화한 AI 알고리즘 압축 기술 및 아날로그 신경망 가속기인 ABLAC을 통해 2.4pJ/SOP의 저전력 동작 성능을 보인다.

서버 환경에서는 NVIDIA GPU의 구조적인 비효율성을 개선하고자 systolic array 구조 기반의 AI 알고리즘 가속기들이 개발되고 있다. 특히 ETRI AB9의 인공지능프로세서는 integer unit을 사용함에 따른 정확도 손실을 방지하고자 모두 16비트 floating-point 유닛으로만 구성되어 있으며, unified buffer 구조의 DC 및 7차원 네스트 루프 구조의 AG를 통해 AI 알고리즘의 구조와 무관하게 일정한 성능을 보일 수 있다.

AI 알고리즘의 학습을 가속할 수 있는 고성능 인공지능프로세서 및 on-chip 학습을 위한 저전력

인공지능프로세서에 대한 요구가 증대됨에 따라, 향후 ETRI는 학습 가속용 인공지능 프로세서에 대한 연구를 진행하고자 한다.

용어해설

Processing In Memory 연산유닛과 메모리 사이의 병목현상(bottleneck)을 개선하기 위해서 연산유닛을 최대한 메모리에 근접하게 배치하는 설계 구조

Operational intensity memory로부터 읽어온 operand가 연산 유닛에서 재사용되는 빈도

약어 정리

AI	Artificial Intelligence
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DSP	Digital Signal Processing
FLOPS	Floating-point Operations Per Second
GPU	Graphics Processing Unit
LSTM	Long Short-Term Memory
MAC	Multiply Accumulate Operation
SOP	Synaptic Operations

참고문헌

[1] A. Reuther et al., "Survey and benchmarking of machine learning accelerators," arXiv preprint arXiv:1908.11348, 2019.

[2] A. Frumusanu, "The Apple iPhone 11, 11 Pro & 11 Pro Max Review: Performance, Battery, & Camera Elevated," Anandtech, Oct. 16, 2019.

[3] A. Ignatov et al., "AI Benchmark: All About Deep Learning on Smartphones in 2019," arXiv preprint arXiv:1910.06663, 2019.

[4] Huawei, consumer.huawei.com/en/campaign/kirin-990-series/

[5] H. Liao, "DaVinci: A Scalable Architecture for Neural Network Computing," 2018, www.hotchips.org/hc31/HC31_1.11_Huawei.Davinci.HengLiao_v4.0.pdf

[6] Samsung, www.samsung.com/semiconductor/minisite/exynos/products/mobileprocessor/exynos-990/

[7] S. Windsor, "Snapdragon 865 vs Kirin 990 5G vs Exynos 990 (Exynos 9830) vs MediaTek Dimensity 1000 (MT6889): which one is the best 5G processor?" www.gearbest.com, Dec. 10, 2019.

[8] H. Liao et al., "DaVinci: A Scalable Architecture for Neural Network Computing," in Proc. Hot Chips 31 Symp., Cupertino, CA, USA, Aug. 18-20, 2019, doi: 10.1109/HOTCHIPS.2019.8875654.

[9] J. Song 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," in Proc. IEEE Int. Solid-State Circuits Conf.-(ISSCC), San Francisco, CA, USA, Feb. 17-21, 2019, doi: 10.1109/ISSCC.2019.8662476.

[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] www.ces.tech/Innovation-Awards/Honorees/2020/Honorees/H/Hailo-8.aspx

[14] H. Orr Danon, "Introducing Hailo-8: The Most Efficient Deep Learning Processor for Edge Devices," 2019 Embedded Vision Summit, May 2019.

[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, "Nvidia Tesla V100 GPU Architecture," WP-08608-001_v1.1, 2017, https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf.

[17] D. Patterson, "Domain Specific Architectures for Deep Neural Networks: Three Generations of Tensor Processing Units (TPUs)," Allen School Distinguished Lecture: David Patterson (UC Berkeley/Google)

[18] N. P. Jouppi et al., "In-Datacenter Performance Analysis of a Tensor Processing Unit," in Proc. Annu. Int. Symp. Comput. Architect., Toronto, Canada, June 2017doi: 10.1145/3079856.3080246.

[19] caffe.berkeleyvision.org

[20] Y. Kwon et al., "Function-Safe Vehicular AI Processor with Nano Core-In-Memory Architecture," in Proc. Annu. Int. Conf. Artif. Intell. Circuits Syst., Hsinchu, Taiwan, Mar. 2019, doi: 10.1109/AICAS.2019.8771603 .