초거대 인공지능 프로세서 반도체 기술 개발 동향

Technical Trends in Hyperscale Artificial Intelligence Processors

전원 (W. Jeon, jeonwon@etri.re.kr) 초거대AI반도체연구실 선임연구원
여준기 (C.G. Lyuh, cglyuh@etri.re.kr) 초거대AI반도체연구실 책임연구원/실장

ABSTRACT

The emergence of generative hyperscale artificial intelligence (AI) has enabled new services, such as image-generating AI and conversational AI based on large language models. Such services likely lead to the influx of numerous users, who cannot be handled using conventional AI models. Furthermore, the exponential increase in training data, computations, and high user demand of AI models has led to intensive hardware resource consumption, highlighting the need to develop domain-specific semiconductors for hyperscale AI. In this technical report, we describe development trends in technologies for hyperscale AI processors pursued by domestic and foreign semiconductor companies, such as NVIDIA, Graphcore, Tesla, Google, Meta, SAPEON, FuriosaAI, and Rebellions.

KEYWORDS artificial intelligence, large language model, neural processing unit, transformer

Ⅰ. 서론

작은 이미지 처리 등 한정된 영역에 집중되었던 초기 인공신경망과 달리, 새롭게 등장한 생성형 초거대 인공신경망은 인간의 언어를 이해하고 문장 및 이미지를 생성해내는 능력을 지닌다. 특히, OpenAI의 GPT(Generative Pre-trained Transformer)로 대표되는 트랜스포머 계열의 인공신경망 모델은 다양한 분야에서 압도적인 성능을 보여주고 있으나, 그만큼 모델의 학습에 필요한 연산량과 메모리 용량이 기하급수적으로 증가하고 있다. 2018년에 등장한 GPT-1이 약 1.1억 개의 모델 크기와 1.7×10^{19}의 학습 연산량을 지녔던 반면, 2023년에 서비스 중인 GPT-4의 경우 약 1.7조 개의 모델 크기와 2.1×10^{25}의 학습 연산량을 지닌 것으로 예측되고 있다[1]. 현세대 초거대 인공신경망, GPT-4는 거대 언어모델(LLM: Large Language Model)과 이미지 입력 이해를 주요 목표로 하고 있다. 이후 등장할 GPT-5 등에서는 비디오 입력 등 사용자의 더욱 다양한 경험을 이해할 수 있는 모델이 될 예정이며, 그에 따

* DOI: https://doi.org/10.22648/ETRI.2023.J.380501
* 이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임[No.2022-0-00018, 거대 인공지능 학습을 위한 K-인공두뇌 반도체 개발].

표 1 서비스별 월간 1억 사용자 달성 시간

서비스	출시 월	개월 수
ChatGPT	2022. 11.	2
TikTok	2016. 09.	9
Instagram	2010. 10.	30
Pinterest	2009. 12.	41
Spotify	2008. 10.	55
Telegram	2013. 08.	61
Uber	2010. 06.	70
Google Translate	2006. 04.	78

출처 Reproduced from [2].

라 학습 모델 크기와 연산량은 더욱 기하급수적으로 증가할 것으로 예상된다. 그뿐만 아니라 이미지 생성형 인공지능 모델(e.g., DALL·E, NovelAI, Stable diffusion)과 대화형 인공지능 모델(e.g., ChatGPT, Bing, Bard)에 대한 일반 사용자의 관심도가 매우 높아지고 있다. 기존 단순 이미지 인식 인공지능, 바둑 인공지능 등의 서비스가 한정된 수요만 있었던 반면, 초거대 인공신경망 기반의 생성형 인공지능 서비스는 매우 높은 수요를 보이고 있다. OpenAI의 ChatGPT는 등장과 함께 큰 관심을 받아 출시 2달 만에 월간 1억 사용자를 달성하여 기존 어떤 온라인 서비스보다도 빠른 달성 시간을 보였다(표 1 참고) [2].

기하급수적으로 증가하는 생성형 인공신경망의 학습 모델 크기, 연산량 및 일반 사용자의 수요는 모델을 학습하고 서비스하는 기업이 매우 많은 하드웨어 자원이 필요하게 만들었다. 대량의 병렬 연산에 유리한 인공신경망 연산의 특성에 따라 다양한 병렬 프로세서(Throughput Processor) 구조 및 기술이 등장하고 있다. OpenAI 등 초거대 인공신경망 모델의 학습을 수행하는 기업은 현재 주로 NVIDIA사의 그래픽 처리 장치(GPU: Graphics Processing Unit)를 활용하여 학습 및 서비스를 제공한다. 그러나 그래픽 처리에 특화되어 제작된 GPU의 한계를 탈피하고 인공신경망, 특히 트랜스포머 계열의 초거대 인공신경망 학습에 특화되어 더욱 효율적인 동작을 수행할 수 있는 인공신경망 전용 프로세서 반도체의 필요성이 대두되고 있다. 2020년대에 이르러, 시스템 반도체 생산 및 공급 문제가 심각해지고 NVIDIA GPU의 공급이 여러 기업의 데이터센터에 원활히 이루어지지 않자, 인공지능 서비스를 제공하는 기업에서 자체적인 인공지능 프로세서 반도체를 개발하는 것이 매우 중요해졌다. 해외에서는 Google, Graphcore, Tesla, Meta 등의 기업에서 관련 기술을 개발하고 있으며, 국내에서도 시스템 반도체 기술에 대한 투자의 중요성이 강조됨에 따라 FuriosaAI, SAPEON, Rebellions 등의 기업에서 초거대 인공신경망 학습 및 서비스를 위한 인공지능 프로세서 기술을 개발하고 있다.

본고에서는 국내외 반도체 기업에서 개발되고 있는 초거대 인공신경망을 위한 프로세서 반도체의 구조 및 기술에 대한 동향을 소개한다.

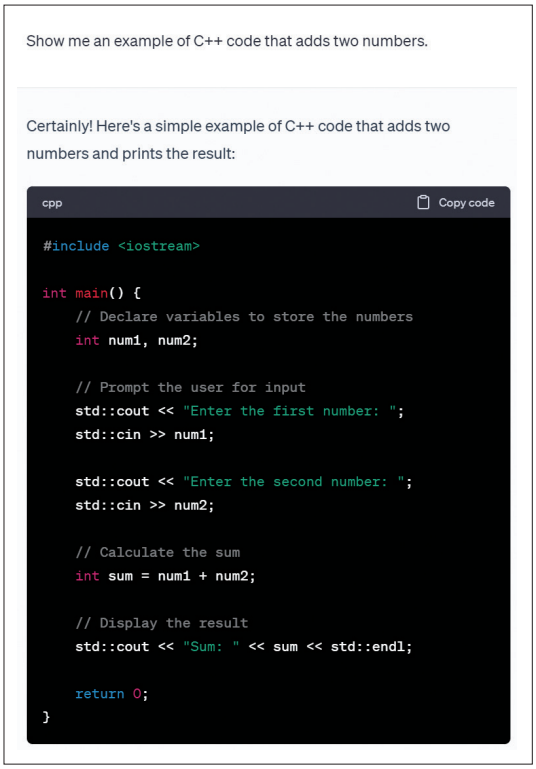
II. 초거대 인공신경망

1. 트랜스포머 모델의 특성과 활용

초기 인공신경망 모델의 발전은 이미지 처리를 위한 합성곱 신경망(CNN: Convolutional Neural Network)과 자연어 처리(NLP: Natural Language Processing)를 위한 순환 신경망(RNN: Recurrent Neural Network) 두 가지 방향으로 발전되어 왔다. 그러나, RNN 형태의 모델은 병렬 연산에 매우 불리한 형태를 지녀 병렬 프로세서에서 대량으로 처리하기에 어려움이 있었다. 트랜스포머 모델은 이러한 RNN의 한계를 극복하며 등장하였고, NLP 분야에서 높은 성능을 보여주었다[3]. 트랜스포머 모델은 크게 self-attention, feed-forward, autoregressive generation 등 다양

한 형태의 연산 방식을 지닌다. 각 연산은 행렬 중심의 compute-intensive 특성이 있기도 하고, 벡터 중심의 data-intensive 특성이 있기도 하여, 트랜스포머 모델을 효과적으로 처리하기 위해서 프로세서 반도체는 높은 연산 처리량과 메모리 대역폭을 모두 지녀야 한다.

BERT(Bidirectional Encoder Representations from Transformers), GPT, HyperCLOVA 등의 초거대 인공신경망 기반 NLP 모델이 트랜스포머를 기반으로 구성되어 있다(GPT-3.5 기반의 ChatGPT 예시, 그림 1)[4]. 초거대 순차 데이터 내부에 존재하는 관계를 추적하고 맥락 및 의미를 학습할 수 있는 트랜스포머 모델은 자연어가 아닌 이미지 분야에도 적용 가능한 특성을 보인다(ViT: Vision Transformer)[5]. 최신의 GPT-4 모델은 이상과 같은 특성을 활용하여 자연어와 이미지를 동시에 처리(Multi-modal)하고 둘의 문맥을 이해함으로써 그림 이해가 필요한 올림피아드 시험 문제를 풀고, 그림의 유머를 이해하는 등 사용자에게 더욱 풍부한 답변을 제공할 수 있게 되었다. 이처럼 복합적인 인공지능 기능은 GPT-4가 초기 일반 인공지능의 수준에 도달했다는 평가를 받도록 하였다[6,7].

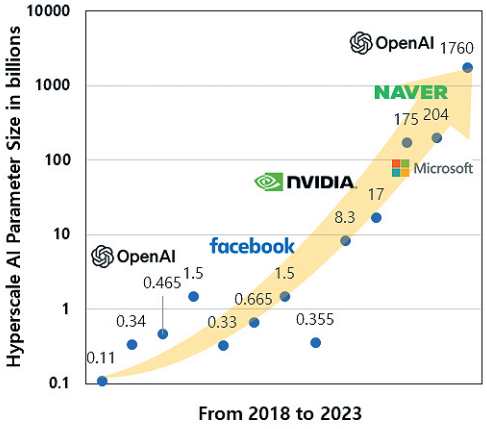


출처 Reproduced from [4].

그림 1 사용자의 요구에 따라 C++ 코드를 생성하는 ChatGPT 실행 예시

2. 전용 프로세서 반도체의 필요성

트랜스포머, 특히 GPT로 대표되는 초거대 인공신경망 모델은 기능성과 성능이 매우 높은 만큼 학습에 필요한 데이터의 크기와 연산량이 기존 인공신경망 모델 대비 기하급수적으로 증가하였다. 아직 공식 발표되지 않은 GPT-4의 학습 정보에 따르면, GPT-1 대비 약 1,500배 많은 모델 파라미터 크기와 120만 배 많은 학습 연산량을 지닌다[1]. 이처럼 거대한 인공신경망 모델에 대한 학습 연산을 수행하기 위해 OpenAI는 1만 대 이상의 NVIDIA



출처 Reproduced from [1].

그림 2 초거대 인공신경망의 모델 크기 증가

H100 GPU로 구성된 시스템을 활용하고 있으며, 결과적으로 GPT-4를 학습하는 데 필요한 비용이 1억 달러에 달한다고 한다[8].

그림 2에서 볼 수 있듯이, 초거대 인공신경망의 모델 크기는 시간이 지남에 따라 기하급수적으로 증가하고 있다. 이는 곧 높은 학습 비용으로 이어짐에도 불구하고 OpenAI, Meta, Microsoft, NVIDIA, NAVER, AI2 등 다양한 기업이 사용자의 수요에 대응하기 위해 초거대 인공신경망 모델 학습에 투자하고 있다.

현재 대부분 초거대 인공신경망 학습은 OpenAI의 GPT-4 예시와 같이 다수의 NVIDIA GPU상에서 이루어진다. GPU는 일반적인 병렬 연산에 매우 효율적인 프로세서로, 특히 GPGPU(General-Purpose computing on GPU) 개념이 도입되고, 행렬 연산에 특화된 NVIDIA Tensor Core가 개발됨에 따라 인공신경망 연산에 강점을 보인다. 그럼에도 GPU는 그래픽 처리를 위해 만들어진 하드웨어라는 한계와 초거대 인공신경망 연산에 특화된 전용 하드웨어가 아니라는 점에 프로세서 반도체로서 연산 효율성을 더욱 개선할 수 있는 부분이 존재할 수 있다.

GPT 등 초거대 인공신경망 연산을 수행하는 데에 있어 GPU에 존재할 수 있는 동작 비효율성을 해결하여 GPT 모델 학습 비용을 절감하고, 반도체 생산 공급 문제를 완화하기 위해 인공지능 서비스 기업 및 반도체 디자인 기업을 필두로 초거대 인공신경망을 위한 전용 프로세서 반도체 개발의 필요성이 대두되고 있다.

Ⅲ. 초거대 AI 프로세서 기술

1. SAPEON

SAPEON은 국내 반도체 설계 업체 중 최초로 데이터센터 타겟의 인공지능 프로세서를 개발한 기업으로, AI inference를 주요 목표로 하는 X220 프로세서를 2020년에 출시하였다. X220은 AI inference에 집중하기 위해 부동소수점 연산을 지원하지 않으며, 정수형 데이터 타입 연산에 대하여 전력 소모 대비 효율적인 연산 성능을 보인다(표 2 참고).

SAPEON은 2023년 중으로 부동소수점 연산 및 AI 학습을 지원하는 X330 프로세서에 대한 개발을 완료할 것으로 예상되며, 이후에는 고대역폭 메모리(HBM: High Bandwidth Memory)인 HBM3를 탑재한 X430 프로세서를 개발할 예정이다[9]. 특히, SAPEON은 SK Hynix와의 협업을 통해 인공지능 프로세서와 인-메모리 컴퓨팅 기술을 융합한 구조를 개발하고 있다. SK Hynix는 가속기 모듈을 off-chip 메모리 내부에 두어 데이터 이동에 낭비되는 에너지와 성능을 향상할 수 있는 구조인 GDDR6-AiM (Graphics Double Data Rate 6 - accelerator in memory) 제품을 개발하였다[10]. SAPEON은 인공지능 프로세서와 GDDR6-AiM을 하나의 시스템으로 묶어 사용할 수 있는 구조를 개발할 예정이다.

SAPEON의 X220 프로세서는 초거대 인공신경망 학습에 대응할 수 있는 반도체는 아니지만, 향후 공개될 X430 프로세서는 초거대 인공신경망 학습을 목표로 개발되고 있다. 특히 다른 반도체 기술과 차별되는 점으로 AiM 기술을 적용하여 GPT 모델 학습 중 data-intensive 특성을 지니는 연산 구간을 더욱 효율적으로 처리할 수 있을 것으로 기대된다.

2. Google

약 2015~2016년의 이른 시기부터 TPU(Tensor Processing Unit)라는 전용 인공지능 프로세서를 개발하여 자체 클라우드 서비스인 Google Cloud에서 사용해온 Google은 2020년에 TPUv4를 발표했다[11]. TPUv4는 트랜스포머 모델 기반의 LLM과 인공지

표 2 국내외 인공지능 프로세서 반도체 개발 현황

HW Spec.	SAPEON X220	Google TPUv4	Graphcore MK2 IPU	Tesla D1	FuriosaAI Warboy	NVIDIA H100 SXM	Rebellions ATOM	Meta MTIA
Tech node	2020, 28nm, –	2020, 7nm, 〈600mm²	2021, 7nm, 832mm²	2021, 7nm, 645mm²	2022, 14nm, 180mm²	2022, 4nm, 814mm²	2023, 5nm, –	2023, 7nm, 373mm²
Data types	INT16, INT8, INT4	BF16, INT8	FP32, FP16	FP32, BF16, FP8, INT8	INT8	**FP64, TF32, FP16, BF16, FP8, INT8**	FP16, INT8, INT4, INT2	FP16, BF16, INT8
16-bit TOPS	–	275 BF16	250 FP16	362 BF16	–	**989.4 FP16/ BF16**	32 FP16	51.2 FP16/ BF16
8-bit TOPS	213 INT8	275 INT8	–	362 FP8	64 INT8	**1978.9 FP8**	128 INT8	102.4 INT8
On-chip memory size(MB)	–	170 (128+ 32+10)	**897**	442.5	32	116 (50+66)	64	136 (128+8)
Off-chip memory size(GB)	–	32, HBM2	–	–	32, LPDDR4X	**80, HBM3**	16, GDDR6	64, LPDDR5
Off-chip memory bandwidth(GB/s)	–	1200	–	–	66	**3352**	256	176
16-bit TOPS/W	–	1.432	0.833	0.905	–	1.413	0.213	**2.048**
8-bit TOPS/W	1.578	1.432	–	0.905	1.067	2.827	0.853	**4.096**
Peak Power(W)	135	192	300	400	60	700	150	25
Major objective	AI vision inference	LLM training, DLRM	LLM training	FSD training	AI vision inference	LLM training	AI vision, NLP inference	DLRM

출처 Reproduced from [9,12-15,17,19,21].

능 기반 추천모델인 DLRM(Deep Learning Recommendation Model)에 대한 연산을 효율적으로 처리하는 것을 주요 목표로 한다. Google에 따르면 2022년 10월 기준, 전체 TPUv4 클라우드 서비스 사용에서 DLRM 학습이 24%, 트랜스포머 학습이 57%의 비중을 지닌다[11].

TPU는 최초 개발된 TPUv1부터 TPUv4까지 시스톨릭 어레이 형태의 행렬곱 가속기 구조를 지녀왔다. TPUv4 프로세서는 128×128 크기의 시스톨릭 어레이 기반 행렬곱 유닛(MXU: Matrix Multiply Unit)을 다수 포함하고 있으며, 벡터 연산을 효율적으로 처리하기 위한 벡터 전용 연산기(VPU: Vector Processing Unit)를 지닌다.

그림 3에 묘사되었듯이, 시스톨릭 어레이 기반의 MXU는 연산기에 필요한 데이터를 전달할 때 모든 연산기가 메모리에 직접 연결되지 않고, 인접한 다른 연산기로부터 피연산 데이터를 전달받는 구조를 지닌다. 이를 통해 데이터 이동에 소모되는 전력 및 하드웨어 자원을 아낄 수 있어 일반 연산기 구조 대비 효율적으로 행렬곱 연산을 수행할 수 있다.

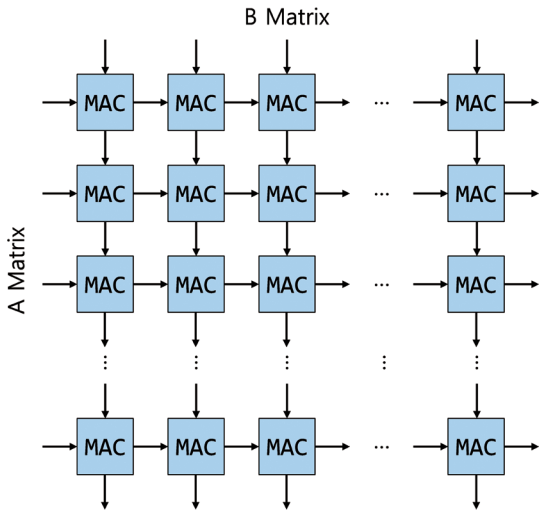


그림 3 시스톨릭 어레이 기반 행렬곱 연산기 구조

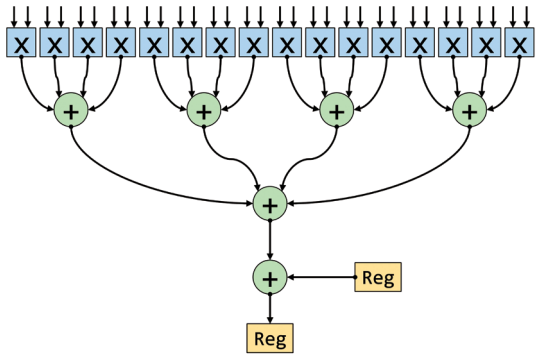


그림 4 스칼라곱 연산 기반 연산기 구조

초거대 인공신경망을 활용한 NLP 및 DLRM의 학습을 수행할 수 있도록 TPUv4는 부동소수점인 BF16(16-bit brain floating point number) 데이터 타입의 연산을 지원한다. 또한, data-intensive 연산 구간에 필요한 데이터를 효율적으로 연산기에 가져올 수 있도록 높은 대역폭을 지니는 HBM2 메모리를 사용한다.

3. Graphcore

Graphcore는 2018년에 Colossus MK1 GC2 IPU (Intelligence Processing Unit) 프로세서를 개발하였으며, 후속 프로세서인 Colossus MK2 GC200 IPU를 2021년에 개발하였다[12]. Graphcore는 개발된 IPU를 활용하여 GPT와 같은 초거대 인공신경망이 적용된 NLP 학습 및 서비스를 제공한다.

IPU는 내부 연산기 간 데이터 공유를 위한 메모리를 갖지 않고, 총 1,472개의 IPU-Tile을 분산하여 구현하였다. 하나의 IPU-Tile에는 인공신경망에 특화된 연산을 수행할 수 있는 IPU-Core와 in-processor-memory 형태로 분산되어 구현된 온칩 메모리가 존재한다. IPU-Tile은 행렬 및 벡터 연산을 효율적으로 가속하기 위한 전용 연산 하드웨어 구조인 AMP(Accumulating Matrix Product) 모듈을 포함한다. AMP 모듈은 그림 4에서 묘사된 스칼라곱 연산을 수행하는 DPU(Dot Product Unit)의 특성을 보인다. 특히, Graphcore IPU의 AMP 모듈은 FP16(16-bit half precision floating point number) 입력의 곱셈 결과를 기존 레지스터에 저장된 FP32(32-bit single precision floating point number) 값과 더할 수 있는 mixed-precision 산술연산 기능을 지원한다. 하나의 AMP 모듈에서 FP16 곱셈 결과를 FP16 값과 더할 경우, 한 번에 64개의 MAC(Multiply Accumulate) 연산을 수행할 수 있고, FP16 곱셈 결과를 FP32 값과 더할 경우, 32개의 MAC 연산을 수행할 수 있다.

다른 인공지능 프로세서 반도체와 비교했을 때 Graphcore IPU의 가장 큰 특징은 대량의 온칩 메모리 용량에 있다. IPU-Tile에 분산되어 구현된 온칩 메모리의 총용량은 897MB에 달하며, DRAM (Dynamic Random Access Memory) 대비 매우 높은 대역폭을 제공할 수 있는 온칩 SRAM(Static Random Access Memory)을 중점적으로 활용할 수 있는 구조를 지녀, 초거대 인공신경망 학습 중 data-intensive 연

산 구간에서 강점을 가질 수 있다. IPU 프로세서는 칩마다 별도로 할당된 오프칩 DRAM을 지니지 않으며, 다수의 IPU 프로세서를 묶어 하나의 컴퓨팅 시스템으로 구성할 때 Streaming Memory라는 오프칩 DRAM 모듈을 포함하여 구성된다. Graphcore에서 자체 구성한 Bow-2000 컴퓨팅 시스템의 예시에서는 4개의 IPU 프로세서 칩과 256GB의 오프칩 메모리로 구성되었다.

4. Tesla

자동차의 자율주행 기능, 특히 완전자율주행(FSD: Full Self Driving) 기술에 필요한 인공신경망 모델을 학습하고 서비스하기 위해 Tesla에서는 2021년 인공지능 프로세서 반도체 D1을 개발하였다[13]. Tesla의 D1의 주요 실행 목표는 GPT 등 LLM 학습을 목표로 하는 초거대 인공신경망이 아닌 FSD 등 자율주행과 관련된 인공신경망 모델이다. 해당 인공지능 학습에는 주로 비디오 형태로 주어지는 시각 학습 정보, 주행 경로 계획, 자동 물체 라벨링 등의 기능이 필요하다.

D1 프로세서는 354개의 DOJO 학습 노드를 온칩 네트워크로 연결한 구조를 지닌다. DOJO 학습 노드는 독립된 연산기 코어로, 내부에 BF16, FP8(8-bit floating point number), INT8(8-bit integer number) 등의 데이터 타입에 대한 연산을 수행할 수 있는 행렬곱 전용 연산기와 벡터 연산기 구조를 포함하고 있다. 하나의 DOJO 학습 노드는 한 번에 $(8\times8)\times(8\times8)$의 행렬 입력에 대한 행렬곱 연산을 수행할 수 있는 유닛 4개를 가지며, 각 유닛은 행렬 입력 데이터에 대한 padding, transpose, decompress 등의 데이터 가공 동작을 수행할 수 있다.

온칩 메모리로는 노드 내부에 피연산 행렬 및 벡터 데이터를 저장하기 위한 1.25MB 용량의 SRAM을 지니고 있다. 또한 DOJO는 내부에 온칩 네트워크의 라우터를 지니고 있어, 총 144개의 직렬 링크를 통해 연결된 다른 DOJO 학습 노드와 데이터를 주고받을 수 있다. 결과적으로 D1 프로세서는 총 442.5MB의 대용량 온칩 메모리를 지닌다. 더 나아가, Tesla는 25개의 D1 프로세서를 5×5 형태로 묶어 하나의 모듈로 만들어 학습 연산 노드의 규모를 확장하는 구조를 개발하였다.

5. FuriosaAI

FuriosaAI는 국내 반도체 설계 업체로, 객체 감지, 얼굴 인식 등 CNN 계열의 AI inference를 효율적으로 가속할 수 있는 Warboy 프로세서 반도체를 2022년 발표하였다. 인공신경망 학습 기능은 목표로 하지 않아 부동소수점 연산은 지원하지 않으며, 8-bit 정수형 데이터 타입에 대한 연산을 수행할 수 있다[14,15]. YOLO, VGG 등 다양한 비전 관련 인공지능 모델에 대한 가속을 지원하며, 양자화(Quantization) 기능 등을 지원하는 소프트웨어 스택을 제공한다. 그를 통해 MLPerf의 이미지 분류 및 사물 인식 분야 등에서 NVIDIA T4 프로세서보다 높은 성능을 보인 바 있다[14,15]. Warboy 프로세서는 카카오 사의 클라우드 서비스인 Kakao i cloud에 적용되어 일반 사용자도 쉽게 접근할 수 있는 서비스를 제공 중이다[16].

FuriosaAI는 현재 2세대 인공지능 프로세서를 2024년 완료를 목표로 개발하고 있으며, 해당 칩은 HBM3 등 고대역폭 메모리를 포함하는 등 초거대 인공신경망 학습이 가능한 수준의 고성능 프로세서일 것으로 예상된다. 특히, GPT-3 등의 LLM뿐만 아니라 ViT, DALL·E, Stable diffusion, DLRM 등 다양한 초거대 인공신경망 모델에 대한 학습을 목표로 하고 있다.

6. NVIDIA

기존 GPU는 컴퓨터 시스템에서 그래픽 연산을 처리하기 위한 장치였으나, GPGPU 개념이 도입된 후 병렬 처리를 통한 연산에 유리한 어플리케이션을 중심으로 GPU의 주요 목표가 되었으며, 인공신경망 연산이 주목받기 시작하였다. 2017년 NVIDIA는 인공신경망의 핵심 연산인 행렬 연산을 기존 CUDA (Compute Unified Device Architecture) 코어보다 효율적으로 가속하기 위해 Tensor Core라는 새로운 연산기 구조를 NVIDIA Volta GPU에 도입하였다. Tensor Core는 다수의 GPU 세대를 거치며 진화하여, 2022년 발표된 NVIDIA Hopper H100 GPU에서는 4세대의 Tensor Core가 사용되고 있다[17]. H100 GPU는 다른 인공지능 프로세서 대비 매우 높은 부동소수점 처리 성능을 보이며, 다양한 데이터 타입에 대한 연산 지원, HBM3 활용을 통한 높은 데이터 대역폭 및 메모리 용량 등을 제공한다. 그를 통해, NVIDIA는 클라우드 기반 AI workload 시장에서 약 85%의 점유율을 갖고 있다[18].

H100 GPU에서 1개의 SM(Streaming Multiprocessor)은 4개 Tensor Core를 지니며, 전체 프로세서는 총 132개의 SM을 지녀, 1개 GPU당 528개의 Tensor Core를 지닌다. 하나의 Tensor Core는 FP16/BF16 데이터 타입의 행렬 기준, 한 번에 (4×16)×(16×8) 행렬곱 연산을 수행하여 (4×8) 형태의 행렬곱 연산 결과를 만들 수 있다. FP8 행렬에 대해서는 한 번에 (4×32)×(32×8) 형태의 행렬곱 연산을 수행한다. 곱셈기 및 덧셈기의 구조는 그림 4와 같은 DPU 형태를 지닌다.

Graphcore MK2 IPU, Telsa D1, Google TPUv4 등과 비교했을 때, H100 GPU의 온칩 메모리 용량은 크지 않은 경향을 보인다. 50MB의 L2 캐시 메모리와 각 SM에 분산되어 존재하는 L1 캐시 메모리, scratchpad 메모리, 및 레지스터 파일의 용량을 모두 더하여도 116MB의 온칩 메모리를 지닌다(표 2 참고). 다만, 80GB의 대용량 및 3,352GB/s의 고대역폭 HBM3 메모리를 사용함으로써 온칩 메모리의 용량이 작아서 발생할 수 있는 성능 저하를 최소화할 수 있는 것으로 보인다.

NVIDIA의 이전 세대 GPU인 Ampere A100 GPU와 비교하여, H100 GPU에는 트랜스포머 모델 가속에 특화된 기능을 제공하기 위해 트랜스포머 엔진이라는 새로운 모듈을 개발하였다. 트랜스포머 엔진은 트랜스포머 모델의 레이어마다 어떤 데이터 타입으로 해당 레이어에 대한 연산을 수행할지 실시간으로 결정지을 수 있도록 도와주는 모듈이다. 해당 모듈은 Tensor Core 연산에서 통계를 받아 부동소수점 연산 결과의 오차 등의 데이터를 분석하고, 다음 레이어에서 FP8 등의 작은 포맷을 사용할 수 있는지, FP16, FP32 등 높은 정확도의 포맷을 사용해야 하는지 결정한다. H100 GPU에서 FP8 포맷은 FP16, FP32 대비 높은 연산 성능을 지니고 더 작은 메모리 공간을 차지하지만, 트랜스포머 모델의 학습 및 inference 과정에서 각 레이어에서 필요한 부동소수점 정확도의 최저 조건을 맞추지 못할 수 있다. 트랜스포머 엔진은 실시간 분석을 통해 연산 성능과 정확도에서 최적의 선택을 할 수 있도록 도와주어, FP8 포맷의 높은 연산 성능의 장점을 활용할 수 있게 한다.

7. Rebellions

Rebellions는 2021년 4 TFLOPS(Tera Floating Operations Per Second)의 FP16 연산 성능을 지닌 AI 기반 금융 분석 반도체인 ION 프로세서를 발표한 대한민국의 반도체 설계 기업이다. 엣지, 금융 및 이미지 처리에 집중되었던 ION의 차세대 반도체로, Rebel-

lions는 ATOM 프로세서를 발표하였다[19]. ATOM은 ION 대비 8배 증가한 32 TFLOPS의 FP16 성능을 기반으로, BERT, T5, GPT 등 트랜스포머 기반 인공신경망 모델에 대한 가속을 지원한다. FLOPS 성능 대비 높은 용량의 온칩 SRAM 용량을 지니며, GDDR6의 오프칩 메모리를 채용하여 높은 메모리 대역폭을 확보하였다.

ATOM 프로세서는 앞서 설명된 성능 및 기능을 통해 KT 사의 초거대 AI 모델인 믿음(Mi:dm) 서비스에 활용되고 있다[20]. 초거대 AI 모델 믿음은 한국어에 특화된 사투리 변환, 문법 교정 등 LLM 기반 NLP 인공지능 서비스와 ION 프로세서부터 제공되었던 금융 관련 솔루션 등을 제공한다.

8. Meta

Meta는 다양한 콘텐츠 및 아이템 추천을 위해 DLRM 인공지능 모델을 적극적으로 사용하고 있다. 기존 CPU(Central Processing Unit) 및 GPU의 사용을 탈피하기 위해 개발된 MTIA(Meta Training Inference Accelerator)를 2023년 발표하였다[21]. MTIA의 핵심 연산기는 8×8 형태로 구성된 64개의 PE(Processing Elements)를 지니며, 각 PE는 행렬곱 연산을 가속하기 위해 그림 4와 유사한 연산을 수행할 수 있는 DPE(Dot-Product Engine)와 벡터 연산을 처리할 수 있는 SIMD(Single Instruction Multiple Data) engine 등을 지닌다. 하나의 DPE는 한 번에 (32×32) 형태의 INT8 또는 (32×16) 형태의 FP16/BF16 행렬에 대한 곱셈을 수행할 수 있다. 각 PE는 2개의 RISC-V 프로세서를 포함하고 있어, 사용자의 어플리케이션을 실행하면서 DPE, SIMD engine 등에 연산 명령을 전달하는 역할을 수행한다. 이와 같은 하드웨어 구조를 통해 MTIA는 본고에서 분석된 인공지능 프로세서 중 전력 소모 대비 가장 높은 16-bit/8-bit 성능을 보였다(표 2 참고).

각 PE 내부에는 RISC-V 프로세서, DPE 및 SIMD engine 등에서 접근할 수 있는 128KB의 로컬 메모리를 지닌다. 64개의 PE 외부에는 총 128MB의 온칩 SRAM이 분산된 모듈로 구현되었으며, 해당 메모리는 scratchpad 메모리의 형태로 사용된다. 오프칩 DRAM으로는 16채널의 LPDDR(Low Power Double Data Rate)메모리를 사용한다. MTIA는 DLRM 레이어 중 Fully-connected 레이어(Compute-intensive)와 Table Batched Embedding 레이어(Data-intensive) 등을 중심적으로 가속하였으며, 특정 DLRM 모델 등에서 GPU 대비 향상된 전력 소모 대비 성능을 보였다.

현재 발표된 MTIA의 내용으로는 GPT 등 초거대 인공신경망 기반 NLP보다는 DLRM 기반 추천 인공지능을 주요 가속 목표로 하는 것으로 보인다. 향후, Meta의 NLP 모델인 LLaMA(Large Language model Meta AI)를 MTIA 아키텍처 기반 반도체를 통해 가속할 수 있을 것으로 기대된다.

Ⅳ. 결론

인공신경망 모델의 크기와 복잡성이 급속도로 증가함에 따라 트랜스포머 모델 기반의 초거대 인공신경망이 등장하고 수요가 급증하고 있다. 그에 따라, 기존 CPU 및 일반 그래픽용 병렬 코어 기반 GPU를 활용하여 모델을 학습하는 것에 한계가 도달하였다. 본고에서는 인공신경망 모델의 학습 또는 inference를 효율적으로 처리할 수 있으며 실제 반도체 칩 제작 및 정보가 공개된 인공지능 프로세서 8종에 대한 정리 및 관련 배경에 대하여 살펴보았다. 현세대의 인공지능 프로세서 중에서는 NVIDIA의 H100 Tensor Core GPU가 가장 높은 부동소수점 성능 및 메모리 시스템 성능을 보인다.

앞서 언급하였듯이, NVIDIA의 H100은 이전 세

대 A100 GPU 등을 포함하여 AI 데이터센터 시장에서 약 85%의 독점에 가까운 점유율을 보인다[18]. 최근 마이크로소프트의 연례보고서에 따르면, AI 프로세서 반도체의 공급 부족이 심각하며, 이에 따라 클라우드 및 데이터센터 운영이 중단될 수도 있다고 한다[22]. NVIDIA의 높은 선도성에도 불구하고, 지나치게 단일 기업의 반도체에 의존적인 상황을 타파하고, 전 세계 AI 데이터센터에 필요한 인공지능 프로세서를 공급을 원활히 하기 위해 초거대 인공지능 프로세서 기술에 대한 연구개발은 계속 이루어질 것으로 예상된다. 다양한 인공지능 프로세서가 목표로 하는 성능 및 기능이 어떤 방향성을 지니는지, 앞으로 더욱 거대해질 초거대 인공신경망 모델의 학습에 성공하기 위해서 프로세서 반도체가 가야 할 방향은 어떤 것인지 통찰이 필요한 시점이다.

약어 정리

AiM	Accelerator in Memory
AMP	Accumulating Matrix Product
BERT	Bidirectional Encoder Representations from Transformers
BF16	16-bit Brain Floating point number
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CUDA	Compute Unified Device Architecture
DLRM	Deep Learning Recommendation Model
DPE	Dot-Product Engine
DPU	Dot-Product Unit
DRAM	Dynamic Random Access Memory
FP	Floating Point number
FSD	Full Self Driving
GDDR	Graphics Double Data Rate
GPGPU	General-Purpose computing on Graphics Processing Unit
GPT	Generative Pre-trained Transformer
GPU	Graphics Processing Unit
HBM	High Bandwidth Memory
INT	Integer
IPU	Intelligence Processing Unit
LLaMA	Large Language model Meta AI
LLM	Large Language Model
LPDDR	Low Power Double Data Rate
MAC	Multiply ACcumulate
MTIA	Meta Training Inference Accelerator
MXU	MatriX multiply Unit
NLP	Natural Language Processing
PE	Processing Elements
RNN	Recurrent Neural Network
SIMD	Single Instruction Multiple Data
SRAM	Static Random Access Memory
TFLOPS	Tera FLoating Operations Per Second
TOPS	Tera Operations Per Second
TOPS/W	Tera Operations Per Second per Watt
TPU	Tensor Processing Unit
ViT	Vision Transformer
VPU	Vector Processing Unit

참고문헌

[1] Epoch, Parameter, Compute and Data Trends in Machine Learning, 2023. 8. 21., Retrieved from https://epochai.org/mlinputs/visualization

[2] Yahoo!finance, "ChatGPT on track to surpass 100 million users faster than TikTok or Instagram: UBS," 2023. 2. 3.

[3] A. Vaswani et al., "Attention is all you need," in Proc. Int. Conf. Neural Inform. Process. Syst., (Long Beach, CA, USA), Dec. 2017, pp. 6000-6010.

[4] https://chat.openai.com/

[5] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint, CoRR, 2020, arXiv: 2010.11929.

[6] OpenAI, "GPT-4 technical report," arXiv preprint, CoRR, 2023, arXiv: 2303.08774.

[7] S. Bubeck et al., "Sparks of artificial general intelligence: Early experiments with gpt-4," arXiv preprint, CoRR, 2023, arXiv: 2303.12712.

[8] Wired, "OpenAI's CEO Says the Age of Giant AI Models Is Already Over," 2023. 4. 17.

[9] https://www.sapeon.com/

[10] Korea IT News, "Sapeon 'Enhancing omputation with next-generation memory artificial brain 'CIM'',," 2022. 11. 17.

[11] N. Jouppi et al., "Tpu v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings," in Proc, ISCA 2023, (Orlando, FL, USA), June 2023, pp. 1–14.

[12] Graphcore, "Graphcore documents," accessed online at https://docs.graphcore.ai/en/latest/.

[13] E. Talpes et al., "The microarchitecture of dojo, tesla's exa-scale computer," IEEE Micro, vol. 43, no. 3, 2023, pp. 31–39.

[14] FuriosaAI, FuriosaAI WARBOY: High performance inference chip for the most advanced vision applications.

[15] FuriosaAI, FuriosaAI NPU & SDK 0.10.0 Documents, Accessed online at https://furiosa-ai.github.io/docs/latest/en/

[16] https://kakaoicloud.co.kr/service/detail/1-44

[17] NVIDIA, NVIDIA H100 Tensor Core GPU Architecture: EXCEPTIONAL PERFORMANCE, SCALABILITY, AND SECURITY FOR THE DATA CENTER v1.04, 2023.

[18] Head Topics, "Nvidia's AI Chips Are Pulling Ahead in the Cloud," 2023. 8. 18.

[19] Rebellions, "ATOM: 5nm Versatile Inference SoC," 2023.

[20] https://enterprise.kt.com/pd/P_PD_NE_00_316.do

[21] A. Firoozshahian et al., "MTIA: First generation silicon targeting Meta's recommendation systems," in Proc. ISCA 2023, (Orlando, FL, USA), June 2023, pp. 1–13.

[22] CNN, "The big bottleneck for AI: A shortage of powerful chips," 2023. 8. 6.