반도체 : 고성능 AI 반도체의 기술적 과제

김한준 FuriosaAI CTO

1. 머리말

ChatGPT 등장 이후, 초거대 언어 모델을 중심으로 연구 영역에 있던 인공지능 모델들이 서비스 영역으로 빠르게 확산되고 있다. 구글의 연구[1]에 따르면, 언어 모델이 일정 규모에 이르면 새로운 능력이 창발적으로 발현된다. 초거대 언어 모델의 경우 사용자가 입력한 맥락으로부터 사고하는 능력인 인컨텍스트 러닝(In-Context Learning)이 창발적으로 발현되어, 특정한 태스크만 수행하는 것이 아니라 사용자가 입력한 맥락에 따라 다양한 태스크를 수행할 수 있게 된다. 초거대 언어 모델의 인컨텍스트 러닝 능력 덕분에, 필요로 하는 태스크에 맞게 맥락을 입력하여 학습이나 번역 등 다양한 방식으로 활용할 수 있게 되었다. API를 통해 다양한 애플리케이션에 이 모델을 통합하여 사용하는 것도 가능해졌다. 나아가 최근 주목받고 있는 LangChain 같은 프레임워크는 언어 모델을 기반으로 검색이나 데이터베이스 등을 연결해 높은 수준의 소프트웨어 서비스 개발을 용이하게 하며, HuggingGPT[2] 같은 프로젝트는 GPT 같은 언어 모델을 중심으로 HuggingFace에서 제공하는 다양한 인공지능 모델을 결합하여 보다 복잡한 태스크를 수행할 수 있게 한다. 이런 식으로 기존에는 단순한 태스크에 한정되어 있던 다양한 인공지능 모델들이 이제는 소프트웨어와 결합해 초거대 언어 모델을 인터페이스로 활용하여 보다 복합적이고 실용적인 태스크를 수행할 수 있게 되었다. 이러한 추세는 인공지능이 다양한 서비스 영역으로 확장되는 새로운 시대가 열리고 있음을 보여준다.

딥러닝 모델의 추론 연산량은 인공지능 서비스를 사용하는 활성 사용자 수, 사용자의 서비스 사용 빈도, 그리고 각 서비스에 적용되는 인공지능 모델의 수에 비례하여 증가한다. 현재 ChatGPT는 역대 가장 빠른 시간에 1억 명의 활성 사용자를 달성한 서비스이며, ChatGPT를 포함한 OpenAI[3]의 API들이 연동된 서비스의 수 또한 빠르게 증가하고 있다. 이런 흐름 속에서 데이터센터에서 딥러닝 추론 작업을 하는데 쓰이는 비용이 중요한 이슈로 부각되고 있다. 그 결과 최근에는 초거대 언어 모델의 추론 비용을 줄이는 알고리즘, 소프트웨어 등에 대한 연구가 빠르게 진행되고 있다.

딥러닝 연산 비용을 줄이는 근본적인 방법 중 하나는 AI 반도체를 비롯한 시스템의 개발이다. 구글의 연구[4]에 따르면, 데이터센터의 총 운용비용(Total Cost of Ownership)은 시스템의 설계 전력량(Thermal Design Power)과 높은 상관관계를 갖는다. 따라서 더 높은 에너지 효율로 딥러닝 추론을 처리할 수 있는 AI 반도체는 데이터센터의 핵심 경쟁력 요소가 되리라는 예상이다. 대규모 데이터센터를 운영하는 대표적 기업들인 구글, 아마존, 마이크로소프트, 메타 등은 GPU

에 비해 높은 에너지 효율을 실현하고, 데이터센터의 비용을 줄이기 위해 자체적인 AI 반도체를 이미 사용하고 있거나 개발 중이다. 기업들은 보다 효율적인 시스템 운영을 위해 자체 기술 개발에 주력하고 있다. 이러한 추세는 AI 반도체의 중요성을 잘 보여준다.

초거대 언어 모델의 대중화와 생성형 AI의 발전은 이전보다 더 높은 수준의 기술적 과제들을 요구한다. 이는 딥러닝 모델의 연산이 다음과 같은 특성을 가지기 때문이다. 첫째, 일반적인 다른 알고리즘들에 비해 월등히 높은 수준의 연산 처리량을 필요로 하며, 모델의 크기가 커질수록 더욱 많은 연산 처리량을 요구한다. 둘째, 높은 연산 처리량과 동시에 3~5 TOPS/W 이상의 에너지 효율을 필요로 하며, 이는 최첨단 제조 공정을 필요로 하므로 높은 초기개발비(NRE)를 수반한다. 셋째, 딥러닝 모델에 따라 초거대 언어생성모델과 같이 1TB/s 수준의 메모리 대역폭을 요구하는 메모리 집약적인(Memory-Intensive) 모델이 있는 반면, 많은 수의 연산 집약적인(Compute-Intensive) 모델이 있다. 또 응용에 따라 다양한 텐서의 크기와 형태, 텐서 연산의 조합을 활용한 다양한 모델이 복합적으로 사용된다. 마지막으로, 일반적인 칩 개발 주기에 비해 알고리즘의 발전 속도가 월등히 빠르며, 새로운 알고리즘의 출현에 따라 데이터센터에서 사용되는 딥러닝 모델 분포 역시 빠르게 변화한다[4].

그 결과, 다양하고 복합적인 모델의 사용과 빠른 변화는 AI 반도체의 프로그래머빌리티(Programmability)와 유연성(Flexibility)을 요구한다. 본고에서는 인공지능, 특히 초거대 언어모델의 폭발적인 확산이 딥러닝 추론 비용의 급증을 가져오는 상황에서 데이터센터가 AI 반도체를 필요로 한다는 사실에 주목하면서, AI 반도체가 데이터센터에서 널리 사용되려면 해결해야 하는 도전적인 문제들은 무엇인지, 하드웨어와 소프트웨어 측면에서 AI 반도체가 당면한 기술적 과제들은 무엇인지 기술하고자 한다. 또한, 아직 상용화 단계에 이르지 못했지만 중장기적인 연구 및 개발이 필요한 기술들에 대해서도 살펴보고자 한다.

2. 하드웨어에 대한 기술적 과제

2.1 높은 메모리 대역폭

최근 각광받는 생성형 언어 모델 중 가장 대표적인 모델은 GPT-3로, 이는 사용자가 입력한 문장을 기반으로 다음 토큰(또는 단어)의 분포를 예측한다. GPT-3의 가장 큰 모델은 1750억 개의 파라미터를 가지며, 하나의 토큰을 생성할 때마다 이 모든 파라미터를 사용하여 연산을 한다.

일반적으로 사용자들은 서비스를 이용할 때 어느 정도의 서비스 요구 사항(Service Level Agreement)을 갖는다. 현재 OpenAI가 제공하는 ChatGPT의 서비스 수준이 명확히 명시되어 있지 않지만, 몇몇 분석에 따르면 초당 50토큰 이상을 처리할 수 있는 것으로 알려져 있다. ChatGPT가 사용하는 모델 중 하나인 GPT-3.5가 1750억 개의 파라미터를 갖고, 서버에서는 8비트 정수 양자화를 기반으로 추론을 처리한다는 가정 하에, 서버에서 제공해야 하는 메모리 대역폭은 초당 8.5TB이다. 서버 1개당 8개의 AI 반도체 칩이 사용된다면, 각 AI 반도체는 초당 1TB 이상의 메모리 대역폭을 제공해야 한다. 이런 대역폭을 제공할 수 있는 양산 가능한 메모리는 HBM(High Bandwidth Memory)뿐이므로, 초거대 생성 언어 모델 기반의 서비스를 위한 AI 반도체는 반드시 HBM과 같은 고대역폭 메모리를 사용해야 한다. 그러나 HBM을 사용하려면 실리콘 인터포저를 활용한 패키징 과정이 필요하며, 이는 도입과 양산을 위한 기술적 난이도가

매우 높다. 더구나 이를 개발하고 양산할 수 있는 경험이 있는 반도체 설계 업체는 많지 않다. 또한 HBM을 도입하고 개발하는 비용과 HBM의 높은 단가, 공급보다 더 큰 수요 등이 HBM 도입의 어려움을 더한다.

2.2 높은 연산 능력

최근에는 초거대 언어 모델의 컨텍스트 길이가 모델의 성능에 영향을 미치는 중요한 요소로 인식되고 있다. 사용자가 인컨텍스트 러닝 기능을 적극적으로 활용하는 것이 중요해짐에 따라, 초기 컨텍스트 계산 시간은 전체 서비스 관점에서 중요한 요구사항이 되고 있다. ChatGPT의 초기 버전인 GPT3.5 버전에서는 최대 문맥의 길이로 4000 토큰을 지원하였던 반면, 최근 버전인 GPT4에서는 최대 문맥의 길이를 3만 2000개 수준으로 크게 증가시켜 약 50페이지 분량의 글을 문맥으로 사용할 수 있다. 이후 다른 언어 모델들도 경쟁적으로 더 긴 문맥의 길이를 지원하고 있음을 발표하고 있다. 초거대 언어 모델이 기반을 둔 트랜스포머 모델들의 셀프 어텐션 계산량이 문맥 길이의 제곱에 비례하여 증가하므로, 이를 충족시키기 위해서는 현재 서비스가 수행되고 있는 GPU와 경쟁할 수 있는 수준인 최소 수백 TOPS 수준의 연산 처리능력이 필요하다.

2.3 다양한 숫자 형식 지원

초거대 언어 모델의 메모리 전송량과 연산 비용을 줄이기 위해 다양한 양자화 기법이 연구되고 있다. 현재 딥러닝 모델의 학습에 사용되는 가장 일반적 숫자 형식은 FP32이다. 하지만 다양한 연구와 프레임워크에 의해 BF16/FP16, INT8 등의 숫자 형식들이 모델의 정확도 손상을 최소화하면서도 쉽게 도입되어 사용될 수 있도록 발전하여 왔으며, 추론 비용을 낮추기 위해 널리 사용되고 있다. 나아가 FP8, INT4 혹은 4비트 이하의 양자화 기법들에 대한 연구도 적극적으로 진행되고 있으며, 성과가 점차 드러나고 있다.

데이터센터에 사용되는 AI 반도체는 사용자가 정확도를 손상시키지 않으면서, 추가적인 모델의 변경없이 사용할 수 있는 FP32 혹은 BF16 연산을 지원하는 것이 최우선적으로 고려되어야 한다. 그러나 동시에 숫자 형식에 의한 효율성을 극대화하기 위해서는 기본적으로 INT8을 지원하고, 추가적으로 FP8, INT4 혹은 4비트 이하의 양자화 기법을 지원하여 모델에 따라 추가적인 성능과에너지 효율성 향상을 보일 수 있어야 한다. 이를 통해 GPU에 대해 경쟁력을 가질 수 있다.

2.4 다양한 모델 지원 가능한 유연성

AI 반도체는 높은 메모리 대역폭, 성능, 다양한 숫자 형식에 대한 연산 유닛을 갖춘 것은 물론, 최신 GPU보다 더 높은 에너지 효율성을 보여야 한다. 이를 위해서는 최첨단 공정을 사용해야 하며, 이러한 하드웨어 요구사항은 높은 NRE 비용을 필요로 한다. 반도체 산업의 특성상, 이러한 NRE 비용을 정당화하려면 충분히 큰 시장 규모가 필요하다. 그러므로 데이터센터용 AI 반도체는 특정한 모델에 과도하게 특화되지 않고, 최소한 일정 범위 이상의 초대형 언어 모델을 지원할 수 있는 프로그래밍 가능한 구조가 요구된다.

2.5 이종 AI 반도체를 포함한 시스템

초대형 언어 모델의 추론은 고대역폭 메모리를 필요로 하지만, 모든 딥러닝 모델이 동일한 요구조건을 가진 것은 아니다. HuggingGPT와 같은 프레임워크를 예로 들면, 다양한 태스크에 적합한 다양한 모델이 복합적으로 사용되기 때문에, 고대역폭 메모리가 필요 없고 연산이 중심인 모델에 대해서는 HBM 대신 저전력 DDR(LPDDR) 혹은 그래픽 DDR(GDDR)을 사용하는 비용 효율적 AI 반도체가 적합할 수 있다. 따라서, 복합적인 모델이 사용되는 데이터센터에서는 다양한 종류의 AI 반도체들이 시스템 레벨에서 함께 사용될 것으로 예상된다.

2.6 선제적 하드웨어 사양 결정

위에서 언급한 여러가지 하드웨어 요구사항들이 알고리즘의 빠른 발전에 의해 변화한다는 점이 하드웨어 개발을 위한 사양의 결정에 어려움을 더한다. 반도체 개발은 일반적으로 최소 2년이상의 기간을 필요로 하므로, AI 반도체는 현재의 대표적인 딥러닝 모델들을 지원해야할 뿐 아니라, AI 반도체가 출시될 시점의 딥러닝 모델들도 지원할 수 있어야 한다. GPT-3와 같은 초거대 언어 모델이 발표되고 상용화된 후에도 초거대 언어 모델을 추론하기 위한 AI 반도체에 대한 요구사항이 높지 않았던 반면, ChatGPT 등장 이후에는 몇 개월 만에 초거대 언어 모델을 추론하기 위한 높은 수준의 하드웨어 성능이 요구되는 현재의 상황도 반도체 시장의 이런 특성을 잘 보여준다. 이 때문에 AI 반도체가 미래의 딥러닝 모델을 지원할 수 있도록 프로그래머빌리티와 유연성을 갖는 것이 필요한 동시에, 다양한 알고리즘연구와 산업 변화를 주의 깊게 관찰하고, 그에 따른 큰 흐름을 예측하여 미래의 알고리즘에 적합한 하드웨어 사양을 결정하는 능력이 중요하다.

3. 소프트웨어에 대한 기술적 과제

3.1 CUDA 수준의 사용성

하드웨어 스펙이 뛰어난 AI 반도체가 GPU와 경쟁하는데 필수적이지만, 널리 사용되기 위한 충분조건은 아니다. 실제로 많은 AI 반도체 업체들이 GPU에 비해 높은 성능, 우수한 에너지 효율, 인상적인 벤치마크 결과를 보여주었고, 대규모 데이터센터도 AI 반도체 업체들의 제품들을 도입하기 위해 노력하였으나 여전히 널리 사용되고 있지 않다. 이러한 현상의 주요 원인 중 하나로 소프트웨어 스택의 사용성 부족이 지적된다.

많은 사람들이 NVIDIA GPU의 CUDA를 사용성 문제의 핵심으로 지목한다. 실제로 딥러닝의 급격한 발전에는 GPU와 CUDA의 기여가 크며, 초기 연구자들과 프레임워크들은 CUDA에 밀접하게 통합되어 있었다. 그러나 최근 발표된 PyTorch 2.0의 주요 변경사항 중 일부는 CUDA에 대한 의존성을 줄이고 다양한 AI 반도체가 PyTorch 2.0의 백엔드에 연결될 수 있도록 권장하고 있다. TensorFlow나 PyTorch와 같은 프레임워크를 개발하는 주요 업체들은 자체 개발한 AI 반도체뿐 아니라 다양한 AI 반도체를 사용할 수 있도록 지원하는 것이 바람직하기 때문이다. 현재 대다수의 연구자들은 PyTorch나 TensorFlow의 연산자를 사용하여 모델을 개발하고 있으며, 직접 CUDA 프로그래밍을 하여 개발하는 경우는 드물다. CUDA를 직접적으로 사용하는 것은 아니므로 CUDA에 록인(Lock-In)되어 있는 것은 아니지만, PyTorch나 TensorFlow가 CUDA를 통

해 GPU를 사용하므로, AI 반도체가 널리 사용되기 위해서는 CUDA 수준의 사용성 혹은 프로그래머빌리티를 제공하는 것이 중요하다.

3.2 AI 반도체에 최적화된 컴파일러

AI 반도체가 딥러닝 모델의 연산을 수행하고 결과를 반환하기 위해선 디바이스 드라이버, 컴파일러, 런타임 소프트웨어 등 여러 소프트웨어 스택들이 필요하다. 이 중에서 사용자의 모델이 실행 가능하거나 경쟁력 있는 성능으로 실행될 수 있는지를 결정하는 핵심 소프트웨어는 컴파일러이다. 컴파일러는 다양한 프레임워크에서 기술된 사용자의 모델을 입력으로 받아, AI 반도체가 실행할 수 있는 바이너리로 변환하는 역할을 담당한다. 이 과정에서 AI 반도체의 연산 유닛들과 메모리에 사용자의 모델이 포함한 연산, 파라미터, 텐서들을 스케줄링하고 자원을 할당하는 최적화 작업을 수행한다. 컴파일러가 사용자 모델에 포함된 연산을 수행할 수 있는 적절한 연산 유닛을 찾지 못한다면, 실행에 실패하거나 가속되지 않은 성능으로 CPU 등에서 연산이 수행될 수 있다.

한편, AI 반도체가 충분히 프로그래머블하고 유연하게 설계되어 있음에도 불구하고, 소프트웨어 복잡도를 간과하여 다양한 모델에 대해 경쟁력있는 성능으로 컴파일하는 데 실패하는 경우도 많이 있다. 이렇게 되면 AI 반도체가 포함하고 있는 연산 유닛과 메모리에 충분히 높은 성능과 효율로 할당하는 최적화 문제가 지나치게 복잡하여, 컴파일 시간 안에 최적의 할당에 실패하여 GPU 대비 낮은 성능으로 연산을 수행하게 된다. 고성능 AI 반도체는 수십만 개의 연산 유닛들이 로컬 메모리와 함께 분산되어 온칩 네트워크로 연결된 구조를 갖는 경우가 일반적이다. 이런 대규모 연산 유닛들에 대해 높은 이용률(Utilization)을 유지할 수 있도록 끊임없이 데이터가 공급되기 위해서는 분산되어 있는 로컬 메모리들에 데이터를 위치시키고 온칩 네트워크를 통해 이동시키는 최적화 문제를 동시에 풀어야 한다. 이는 일반적으로 연산 유닛 스케쥴링 최적화 문제에 비해 더 높은 차원의 복잡도를 갖는다[5]. 결과적으로 널리 쓰이는 간단한 벤치마크 모델들에 대해서는 하드웨어 설계 단계에서 스케쥴링과 메모리 할당과 이동 문제를 충분히 고려하여 설계가 이뤄져서 메뉴얼하게 컴파일된 프로그램이 경쟁력 있는 결과를 보여줄 수 있는 반면, 임의의 사용자의 복잡한 모델에 대해서는 최적화에 실패하는 경우가 발생할 수 있다.

AI 반도체에 대한 컴파일러 문제를 해결하기 위해 MLIR[6]과 TVM[7]과 같은 범용 컴파일러 연구가 진행되고 있다. MLIR은 구글에서 개발되어 LLVM과 같이 다양한 계층의 중간표현(IR)을 지원하고, AI 반도체를 포함한 다양한 프로세서에 대한 백엔드를 지원할 수 있는 프레임워크를 제공한다. 이는 프로세서에 독립적인 공통 모델 최적화를 공유하며, 각 프로세서에 적합한 최적화를 IR에 추가하거나 확장할 수 있도록 설계되었다. TVM은 다양한 AI 반도체에 대한 스케줄링 및 최적화를 수행할 수 있는 범용 컴파일러를 제공한다. TVM은 상당히 넓은 범위의 AI 반도체 최적화를 지원하며, 실제로 다양한 AI 반도체에 대한 컴파일러로 사용될 수 있다. 그러나 AI 반도체 내부 연산뿐만 아니라 메모리 계층 구조, 온칩 네트워크의 구조와 특성 등이 모두 다르므로 모든 AI 반도체에 대한 최적화 문제를 해결하는 것은 어렵다. 실제로 일부 연구 논문에서는 GPU에서의 TVM 컴파일 결과가 CUDA에서 제공하는 라이브러리에 비해 상당히 낮은 성능을 보여주는 것으로 나타났다[5].

따라서 AI 반도체 개발에서는 하드웨어의 성능과 에너지 효율뿐만 아니라, 컴파일러의 복잡도를 고려한 프로그래밍 모델을 제공하는 것도 필요하다. AI 반도체는 딥러닝 알고리즘에 대한 깊은 이해를 바탕으로 하드웨어와 소프트웨어가 통합 설계되어야 하며, 이는 딥러닝 연산이라는 특정 도메인에 특화된 프로그래밍 모델을 형성하는데 중요하다.

3.3 데이터센터 스케일로 확장 가능한 소프트웨어 계층지원

데이터센터와 클라우드는 기업이나 개인이 자신의 컴퓨팅 작업을 처리하기 위해 필요한 자원을 공급하는 중요한 인프라스트럭처이다. 가장 큰 장점은 사용자들이 동일한 자원을 공유함으로써 자원 효율성을 높이고 가용성을 높일 수 있다는 점이다. 가상머신은 물리적 하드웨어를 여러 가상의 컴퓨터로 나누는 기술로, 하나의 서버에서 여러 개의 독립된 가상 서버를 운영할 수 있게 해준다. 따라서 AI 반도체가 데이터센터에서 사용될 경우 가상머신을 지원하여 다수의 사용자가 이를 공유할 수 있도록 해야 한다.

또한 데이터센터에서는 소프트웨어의 확장성을 보장하기 위해 쿠버네티스와 같은 컨테이너 오케스트레이션 플랫폼이 광범위하게 사용되고 있다. 쿠버네티스는 컨테이너화된 애플리케이션을 자동으로 배포, 확장, 관리해주는 오픈소스 플랫폼으로, 데이터센터의 효율성과 신뢰성 제고에 중요한 역할을 한다. AI 반도체가 쿠버네티스에 쉽게 통합되려면 쿠버네티스 플러그인 및 모니터링 도구를 지원해야 한다. 이렇게 함으로써 AI 반도체는 데이터센터의 기존 인프라스트럭처에 매끄럽게 통합되어 최적 성능을 발휘하고 사용자에게 더 나은 서비스를 제공할 수 있게 된다.

4. 중장기적 기술 과제

4.1 컨피덴셜 컴퓨팅

최근 인공지능 기반 챗봇 ChatGPT가 회사 기밀 정보를 외부로 유출하는 문제들이 기사화된 바 있다. 이 사건 이후 삼성, 애플 등 여러 기술 기업들이 내부 정보의 보안을 위해 ChatGPT 사용을 제한하거나 금지하는 등의 대응을 보였다. 이와 같은 문제를 해결하기 위해 회사들이 자체 데이터센터를 활용하여 독립적으로 이러한 서비스를 운영하는 움직임이 생길 것으로 예상되는 한편, ChatGPT를 서비스하는 API 업체들도 회사의 기밀이 보안적으로 안전하게 반도체 내에서만 프로세싱되도록 하려는 움직임이 증가할 전망이다.

기업뿐 아니라, 개인 사용자에게도 이러한 보안 이슈와 함께 프라이버시 문제는 중요한 이슈로 부상할 것이다. AI 서비스의 사용 범위가 확장됨에 따라 개인 정보가 이러한 서비스에 의해 더욱 폭넓게 활용될 것이다. 이런 요구사항에 따라 개인 디바이스 수준에서 AI 연산능력을 강화하려는 시도가 늘어날 것이며, 이는 에지 컴퓨팅 영역에서 AI 반도체 수요 증가로 이어질 전망이다. 그러나 에지 컴퓨팅 영역의 불충분한 연산 능력이나 배터리 용량 등의 제약 때문에 데이터센터에서 보다 안전하고 효율적인 방식으로 AI 서비스를 제공할 필요성 역시 커질 것으로 예상된다.

인공지능 서비스 사용자가 아닌 인공지능 서비스와 모델을 개발하는 기업들 관점에서 보면, 이들에게 인공지능 모델들은 매우 중요한 자산이며, 따라서 이들 역시 인공지능 모델이 공개 클라우드 환경에서 서비스될 때 인공지능 모델에 대한 정보가 클라우드 업체 혹은 클라우드 업체

를 사용하는 다른 사용자에게 접근되거나 조작될 수 있다는 우려가 있다.
이런 문제를 해결하기 위한 한 가지 방법이 바로 컨피덴셜 컴퓨팅이다. 컨피덴셜 컴퓨팅은 하드웨어 기반의 신뢰할 수 있는 실행 환경(TEE, Trusted Execution Environment)을 이용하여 사용 중인 데이터를 보호하는 기술이다. TEE는 사용 중인 애플리케이션과 데이터에 대해 승인되지 않은 액세스나 수정을 방지하는 보안 격리 환경을 제공한다. 이런 보안 표준은 컨피덴셜 컴퓨팅 컨소시엄에서 정의하고 있다[8].
따라서 AI 반도체 제조사들은 컨피덴셜 컴퓨팅을 지원하는 기능을 제공함으로써 데이터센터의 보안 및 프라이버시 이슈를 해결하는 데 기여할 수 있다. 또한 이런 기능은 AI 반도체가 더욱 넓은 범위의 사용자와 기업에게 제공되도록 해 중장기적으로 시장에서 더 큰 성공을 거둘 수 있게 돕는다.

4.2 칩렛(Chiplet) 기술과 칩간 연결(Chip-to-Chip Interconnect)

AI 반도체에 요구되는 연산 처리량이 증가함에 따라 AI 반도체의 칩 크기는 반도체 생산에 사용되는 포토 마스크에 의한 최대 크기 한계치인 약 800㎟에 근접하고 있으며, 칩의 대형화는 한계에 도달하였다. 또한 칩의 크기가 커질수록 웨이퍼 결함으로 인한 수율 저하가 발생할 확률이 높아지며, 수율이 떨어지면 제조원가가 높아진다.
칩렛은 다수의 칩을 고속의 칩 간 연결을 사용하여 연결함으로써 반도체 단일 칩 크기의 한계를 극복하고, 수율을 높임으로써 반도체 제조 원가를 낮출 수 있다. 또한 서로 다른 공정을 적용한 칩을 연결할 수 있으므로, IP에 따라 적합한 공정을 적용할 수 있다. 예를 들어, IO를 담당하는 칩은 기존 공정으로 생산된 칩을 그대로 사용하면서, 딥러닝 연산을 담당하는 연산 유닛에 해당하는 칩에는 최신 공정을 적용하여 칩의 개발단가를 낮추면서도 더 높은 에너지 효율을 달성할 수 있다.
또한 인공지능 모델의 가속을 담당하는 AI 반도체와 일반 서비스 로직을 담당하는 CPU를 칩렛으로 연결하여 이종의 프로세서를 연결함으로써, 전체 추론 서비스를 효율적으로 처리하도록 할 수 있다. 딥러닝의 추론 관점에서는 인공지능 모델이 단독으로 사용되기 보다는, 서비스의 처리 과정에서 다수의 인공지능 모델의 추론이 일반 서비스 로직과 함께 복합적으로 실행되기 때문이다.
다수의 칩이 연결되기 위해서는 칩 간 연결의 성능과 에너지 효율이 중요하다. 공통 규격과 프로토콜로 연결될 경우, 다양한 회사의 제품들이 칩렛 형태로 연결될 수 있다는 장점이 있다. 그에 따라 인텔이 개발해온 기술을 바탕으로 칩간 연결의 표준인 UCIe가 발표되어 칩렛 다이가 상호통신하기 위한 전기 신호 규격이나 물리적 레인수, 범프 피치, 상호 통신 프로토콜을 규격화하고 있다. 인텔, AMD, ARM 등의 기업이 참여하고 있다. NVIDIA는 별도로 NVLink-C2C라는 자체 칩간 연결 기술을 사용하여 CPU와 GPU를 연결한 제품을 출시하고 있다.
칩렛 기술과 칩 간 연결은 AI 반도체의 여러 기술적 한계와 비용 문제를 극복할 수 있는 기반 기술로 여러 반도체 기업들에 의해 빠르게 발전 및 도입되고 있다. 추후 AI 반도체에 도입되어 사용됨으로써 AI 반도체의 성능과 효율을 높이고, 비용을 절감하며, 추론 서비스에서 AI 반도체의 활용도를 높일 것으로 기대된다.

4.3 프로세싱 인 메모리

초거대 언어 모델은 AI 반도체의 DRAM으로부터 거대한 모델 파라미터를 매번 추론마다 필요로 한다. 이에 따라 데이터 이동에 상당한 에너지를 소모하게 되는데, 이는 AI 반도체의 효율성을 저하시키는 주요 요인 중 하나이다. 이러한 문제를 해결하기 위해 데이터 이동에 필요한 메모리 대역폭을 줄이고, 메모리 병목 현상을 완화하면서 동시에 에너지 소모를 줄일 방법이 필요하다. 해결책의 하나로 제시되는 프로세싱 인 메모리(PIM, Processing In Memory) 기술은 데이터 이동에 필요한 메모리 대역폭을 줄이고, 성능을 향상시키며, 에너지 소모를 줄이는 데 도움을 줄 수 있다.

프로세싱 인 메모리가 상용화를 통해 주요 AI 반도체에 도입되기 위해서는 몇 가지 해결해야 할 과제들이 있다. 메모리 칩에 연산 유닛을 집적함에 따라 DRAM 용량이 감소하는 문제가 있다. 또 일반적으로 데이터가 여러 DRAM 칩에 분산되어 저장되므로, 사용자가 필요로 하는 연산을 하기 위해 여러 DRAM 칩 간에 데이터를 효율적으로 이동하고 처리하기 위한 새로운 프로그래밍 모델이 필요하다. 프로세싱 인 메모리 기술은 AI 반도체의 효율성을 크게 향상시킬 수 있는 기술로 여겨지고 있어, 당면한 여러 난관들을 해결한다면 중장기적으로 AI 반도체를 발전시킬 기술이 될 것으로 기대된다.

5. 맺음말

AI 반도체의 등장은 더 빠르고 효율적인 추론을 가능하게 하면서 기계 학습과 AI 산업에 엄청난 변화를 일으켰다. ChatGPT와 같은 초거대 언어 모델의 등장과 폭발적인 확산으로 추론 인프라스트럭처의 중요성이 대두되고 있고, GPU 대비 높은 에너지 효율을 갖는 AI 반도체의 중요성을 더욱 두드러지게 만들었다. 그러나 AI 반도체 개발은 많은 기술적 어려움에 직면해 있다. 높은 연산 능력과 메모리 대역폭, 높은 에너지 효율이 필요하지만, 이는 최첨단 공정을 써야 해 개발 비용이 증가한다. 또한 AI 반도체는 다양한 시장에서 활용될 수 있어야 하고 빠르게 발전하고 있는 미래의 딥러닝 모델들도 지원해야 하므로, 프로그래머들이 쉽게 활용할 수 있도록 프로그래머빌리티가 필수적이다. 이러한 하드웨어를 최적화하고 사용성을 제공할 컴파일러를 비롯한 소프트웨어 스택도 AI 반도체가 데이터센터에서 사용되기 위한 중요한 필요조건이다. 더불어 클라우드 기반의 대규모 인프라에서 사용될 수 있도록 가상머신이나 쿠버네티스와 같은 컨테이너 오케스트레이션 플랫폼에 대한 지원이 필요하다.

이러한 도전 과제들은 데이터센터를 운영하는 기업들이 비용 절감을 위해 새로운 시도를 하고 있음을 의미한다. 자체 개발을 포함하여 다양한 대안을 모색하고 있으며, 다양한 프레임워크를 통해 지원하려는 노력도 기울이고 있다. 이러한 필요성과 생태계의 흐름 속에서 많은 AI 반도체 개발자들은 현재의 기술적 어려움을 극복해 나가고 있다. AI 반도체는 폭발적으로 성장하고 있는 인공지능 서비스 인프라스트럭처의 핵심 요소중 하나로 인공지능 서비스의 대중화에 크게 기여할 것으로 기대된다.

[참고문헌]

[1] WEI, J. et al. (2023), Larger language models do in-context learning differently. arXiv preprint arXiv:2303.03846.

[2] SHEN, Y. et al. (2023), Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. arXiv preprint arXiv:2303.17580.

[3] OpenAI, https://openai.com/

[4] Jouppi, N. et al. (2021), Ten Lessons From Three Generations Shaped Google's TPUv4i : Industrial Product.

[5] BARHAM, P. et al. (2019), Machine learning systems are stuck in a rut. In: Proceedings of the Workshop on Hot Topics in Operating Systems.

[6] LATTNER, C. et al. (2020) MLIR: A compiler infrastructure for the end of Moore's law. arXiv preprint arXiv:2002.11054.

[7] CHEN, T. et al. (2018), TVM: An automated end-to-end optimizing compiler for deep learning. arXiv preprint arXiv:1802.04799.

[8] Confidential Computing Consortium, https://confidentialcomputing.io/

※ 출처: TTA 저널 제207호