

2.3. GPU(Graphic Processing Unit): 그래픽 처리를 넘어 인공지능까지, 병렬 연산의 강자

① 왜 나오게 되었나?

GPU는 원래 컴퓨터 게임이나 3D 그래픽 작업에서 화면에 영상을 빠르고 부드럽게 그려내기 위해 개발되었다. 그런데 GPU가 단순히 그래픽 처리에만 특화된 것이 아니라, 대규모 데이터를 동시에 처리하는 병렬 연산에

탁월한 성능을 발휘한다는 것이 알려지면서 인공지능 분야에서도 주목받게 되었다. 특히 딥러닝(Deep Learning) 같은 인공지능 기술이 발전하면서 GPU의 중요성은 더욱 커졌다.

② 쉽게 설명하면?

GPU는 마치 공장에서 여러 개의 조립 라인이 동시에 작동하는 것처럼, 수많은 연산 코어를 동시에 가동해 방대한 양의 데이터를 빠르게 처리할 수 있다. 반면 중앙처리장치(CPU, Central Processing Unit)는 한 번에 하나의 작업을 처리하는데 보다 특화되어 있다. 이런 GPU의 병렬 처리 능력은 3D 그래픽 렌더링 같은 작업에 최적화되어 있지만, 인공지능에 필요한 대규모 데이터 연산에도 안성맞춤이었던 것이다. 이를 통해 GPU는 그래픽 처리를 넘어 인공지능의 핵심 부품으로 자리매김하게 되었다.

③ 예시와 함께 좀 더 자세히 알아볼까?

딥러닝은 인공지능의 한 분야로, 인간의 뇌 구조를 모방한 인공신경망(Artificial Neural Network)을 대규모 데이터로 학습시켜 스스로 문제를 해결하도록 한다. 이 과정에서 수많은 연산이 필요한데, GPU의 병렬 처리 능력이 크게 빛을 발한다.

예를 들어, 자율주행차가 도로를 주행하려면 카메라로 찍은 영상에서 보행자, 신호등, 다른 차량 등을 실시간으로 인식해야 한다. 이때 딥러닝 알고리즘을 사용하는데, GPU를 통해 방대한 영상 데이터를 빠르고 정확하게 분석할 수 있다. 만약 CPU만으로 이 작업을 처리하려고 한다면 엄청난 시간이 걸릴 것이다. 하지만 수천 개의 코어를 가진 GPU를 사용하면 몇 초 안에 처리가 가능하다. 이처럼 GPU는 딥러닝의 실용화에 결정적인 역할을 하고 있다.

④ 왜 중요한가?

GPU는 인공지능, 특히 딥러닝의 발전에 중추적인 역할을 하고 있다. 방대한 데이터를 빠르게 처리할 수 있는 GPU 덕분에 딥러닝 알고리즘의 학습 시간이 크게 단축되고, 더 복잡하고 정교한 인공지능 모델을 만들 수 있게 되었다.

실제로 딥러닝의 대표 격인 알렉스넷(AlexNet)이 2012년 이미지 인식 대회(ILSVRC, ImageNet Large Scale Visual Recognition Challenge)에서 압도적인 성적을 거둔 것도 GPU를 활용했기에 가능했다. 이는 학계와 업계에 큰 충격을 주었고, 이후 GPU를 활용한 딥러닝 연구가 폭발적으로 증가했다. 이처럼 GPU는 인공지능 기술의 상용화와 대중화를 가속화하는 원동력이 되고 있다. 이에 엔비디아(NVIDIA), AMD 등 GPU 제조사들도 인공지능 특화 GPU를 잇달아 내놓으며 치열한 경쟁을 벌이고 있다.

⑤ 어디에 활용되는가?

GPU는 딥러닝을 비롯한 다양한 인공지능 분야에서 활용된다. 자율주행, 의료 진단, 언어 번역, 이미지 및 음성 인식 등 인공지능 기술이 적용되는 거의 모든 영역에서 GPU의 역할이 크다.

예를 들어, 의료 분야에서는 자기공명영상(Magnetic Resonance Imaging; MRI), 컴퓨터단층촬영검사(Computed Tomography; CT) 등 의료 영상을 분석하여 질병을 진단하는데 GPU가 활용된다. 방대한 양의 의료 데이터를 GPU로 빠르게 처리함으로써 의사의 진단을 도와 의료의 질을 높일 수 있다. 금융 분야에서는 GPU를 통해 실시간으로 방대한 금융 데이터를 분석하고, 고객의 투자 성향에 맞는 포트폴리오를 추천하는 등의 서비스를 제공한다.

뿐만 아니라 빅데이터 분석, 과학 시뮬레이션 등 대규모 병렬 연산이 필요한 분야에서도 GPU가 널리 사용된다. 근래에는 게임용으로 개발된 고성능 GPU가 암호화폐 채굴에도 활용되면서 수요가 폭발적으로 증가하기도 했다. 이처럼 GPU는 우리 삶의 다양한 영역에서 그 존재감을 발휘하고 있으며, 앞으로도 인공지능 시대를 이끄는 핵심 부품으로서 그 역할이 더욱 커질 전망이다.

2.4. NPU(Neural Processing Unit): 인공지능을 위한 작고 강력한 두뇌⁷⁾

① 왜 나오게 되었나?

AI 기술이 점점 더 우리 삶 속으로 들어오면서 스마트폰, 사물인터넷(IoT, Internet of Things) 기기 등 소형 기기에서도 AI를 구동하는 것이 중요해졌다. 하지만 기존의 중앙처리장치(CPU)나 그래픽처리장치(GPU)는 크기가 크고 전력 소모가 많아 이런 작은 기기에는 적합하지 않았다. 이런 문제를 해결하기 위해, 작고 효율적으로 딥러닝 연산을 처리할 수 있는 NPU가 등장했다.

② 쉽게 설명하면??

NPU는 인간의 뇌 구조를 모방한 인공신경망을 효과적으로 구동하도록 설계된 작은 AI 칩이다. 마치 뇌 속의 생각을 관장하는 작은 신경세포처럼, NPU는 스마트폰이나 IoT 기기 안에서 AI의 역할을 한다.

7) NPU는 AI 연산, 특히 추론(Inference)에 특화된 AI 가속기이다. NPU는 학습(Training) 보다는 추론에 최적화되어 있으며, GPU 대비 저전력 소모와 효율 극대화가 주요 장점이다.

원래 NPU는 활용 분야에 따라 기기(Device)용과 데이터센터(Data Center)용으로 구분할 수 있다. 하지만 최근 업계에서는 데이터센터용 NPU를 각 기업이 고유의 브랜드로 차별화하는 추세이다. 구글의 TPU(Tensor Processing Unit), 테슬라의 Dojo, 그래프코어의 IPU(Intelligence Processing Unit) 등이 대표적인 데이터센터용 NPU 브랜드이다.

이러한 상황을 고려하여, 본 보고서에서는 데이터센터용 NPU와 구분짓기 위해 NPU를 스마트폰, 사물인터넷(IoT) 등 소형 기기에 한정지어 논의하고자 한다

우리 뇌에 수많은 뉴런이 서로 연결되어 복잡한 생각을 처리하듯, NPU 내부에는 대량의 연산 유닛이 병렬로 연결되어 AI 연산을 수행한다. NPU는 GPU나 CPU와 달리 딥러닝에 최적화된 구조를 가지고 있어, 더 적은 전력으로 AI 연산을 빠르게 처리할 수 있다.

③ 예시와 함께 좀 더 자세히 알아볼까?

아침에 스마트폰으로 사진을 찍으면, NPU가 그 사진을 분석해 피사체를 인식하고 최적의 설정으로 보정해준다. 예를 들어, 풍경 사진을 찍었을 때 하늘은 더 파랗게, 녹음은 더 싱그럽게 보정되는 것이다.

또 스마트폰으로 외국어를 번역할 때도, NPU가 복잡한 언어 모델을 구동해 빠른 번역을 가능하게 한다. 한국어를 영어로 번역한다고 가정해보자. NPU는 입력된 한국어 문장을 단어 단위로 분석하고, 이를 영어 단어로 변환한 뒤, 문법과 어순에 맞게 재배열하여 자연스러운 영어 문장을 만들어낸다.

한편, 스마트 스피커에 “오늘 날씨 어때?”라고 물으면 NPU가 그 음성을 분석해 사용자의 의도를 파악하고 날씨 정보를 알려준다. 이처럼 NPU는 우리 주변의 다양한 기기에서 AI 기능을 구현하는 데 핵심적인 역할을 하고 있다.

④ 왜 중요한가?

NPU는 AI 기술의 대중화와 실생활 적용에 핵심적인 역할을 한다. 작고 효율적인 NPU 덕분에 스마트폰이나 IoT 기기 같은 소형 기기에서도 강력한 AI 기능을 사용할 수 있게 되었다.

과거에는 AI 연산을 위해 기기에서 클라우드 서버로 데이터를 보내고, 연산 결과를 다시 받아오는 과정이 필요했다. 하지만 NPU는 기기 내에서 직접 AI 연산을 처리하므로, 클라우드 서버에 데이터를 보내지 않아도 된다. 이는 사용자의 프라이버시 보호에도 도움이 된다.

또한 NPU는 저전력으로 동작하기 때문에, 배터리 사용 시간이 길어지는 효과도 있다. 스마트폰에서 AI 기능을 사용해도 하루 종일 배터리가 갈 수 있게 된 것이다. NPU의 발전은 온디바이스AI 발전을 강화하며, 우리 일상 속 AI 활용을 더욱 가속화할 것으로 기대된다.

⑤ 어디에 활용되는가?

NPU는 스마트폰, 태블릿, 스마트 스피커, 드론 등 다양한 모바일 및 IoT 기기에 사용된다. NPU 덕분에 이런 기기들은 사용자의 음성 명령을 이해하고, 사진과 동영상을 스마트하게 촬영 및 편집하며, 얼굴을 인식해 기기를 잠금 해제하는 등 다양한 AI 기능을 제공할 수 있게 되었다.

예를 들어, 최신 스마트폰들은 NPU를 활용해 야간 모드, 인물 사진 모드, 실시간 번역 등 다양한 AI 카메라 기능을 제공한다. 스마트 스피커는 NPU로 사용자의 음성 명령을 이해하고, 음악 재생, 스마트 홈 제어 등을 수행한다.

또 자율주행차의 객체 인식, 스마트 공장의 이상 징후 감지 등 산업 현장에서도 NPU의 역할이 점차 커지고 있다. 앞으로 NPU는 우리 삶 곳곳에서 AI의 힘을 발휘하는 ‘작지만 강한 두뇌’로 활약할 전망이다.

2.5. TPU(Tensor Processing Unit): 딥러닝 가속을 위한 구글의 특별한 칩

① 왜 나오게 되었나?

구글은 딥러닝(Deep Learning)을 활용한 다양한 서비스를 제공하면서, 기존의 중앙처리장치(CPU, Central Processing Unit)나 그래픽 처리 장치(GPU, Graphics Processing Unit)로는 감당하기 어려운 방대한 연산량과 데이터 처리의 한계에 부딪혔다. 이에 구글은 딥러닝에 최적화된 자체 칩을 개발하기로 결정하고, TPU를 만들어냈다. TPU는 딥러닝의 핵심 연산인 행렬 연산(Tensor 연산)을 더욱 빠르고 효율적으로 처리할 수 있도록 특별히 설계되었다.

② 쉽게 설명하면?

TPU는 딥러닝을 위해 태어난 구글의 특별한 칩이다. 마치 특정 요리를 위해 개발된 전용 주방 도구처럼, TPU는 딥러닝이라는 특별한 작업을 위해 만들어졌다. 딥러닝은 수많은 데이터를 처리하고 복잡한 수학 연산을 수행해야 하는데, TPU는 이런 작업을 더 빠르고 효율적으로 처리할 수 있도록 도와준다.

어려운 수학 문제를 푸는데 있어, 일반 계산기로 풀 수도 있지만 공학용 계산기를 사용하면 훨씬 쉽고 빠르게 풀 수 있는 것처럼, TPU는 딥러닝을 위한 특화된 ‘공학용 계산기’라고 볼 수 있다. 구글은 TPU를 자사의 딥러닝 서비스에 적극 활용하여 성능을 크게 향상시켰다.

③ 예시와 함께 좀 더 자세히 알아볼까?

우리가 구글 검색을 하면, 검색어와 관련된 수많은 웹페이지 정보를 순식간에 분석하여 가장 적절한 결과를 보여준다. 또 구글 번역기는 단어와 문장의 의미를 이해하고 다른 언어로 자연스럽게 옮겨준다.

이런 서비스의 핵심에는 딥러닝 기술이 있고, TPU가 그 딥러닝 알고리즘을 빠르고 정확하게 구동하는 역할을 한다. 마치 자동차 엔진이 휘발유를 연소시켜 동력을 만들어내듯, TPU는 데이터를 연산하여 딥러닝에 필요한 동력을 공급하는 셈이다. 실제로 TPU는 GPU 대비 15~30배 이상의 속도로 딥러닝 연산을 처리할 수 있다고 한다.

④ 왜 중요한가?

TPU는 딥러닝 기술의 상용화와 대중화에 중요한 역할을 하고 있다. 빠르고 효율적인 TPU 덕분에 대규모 딥러닝 모델을 실제 서비스에 적용하는 것이 가능해졌다. 또 TPU를 통해 딥러닝 연구와 개발에 필요한 시간과 비용을 크게 줄일 수 있게 되었다.

구글이 딥러닝의 핵심 기술 개발에 많은 기여를 한 것을 감안한다면, 이를 뒷받침한 TPU의 역할이 컸다는 점에 반발을 제기할 사람은 없을 것이다. 더 나아가 TPU는 단순히 기술 개발에 그치지 않고, 딥러닝이 우리의 일상생활에 성큼 다가설 수 있게 만드는 데에도 일조하였다.. 구글의 TPU 성공 사례는 다른 기업들도 자체 AI 칩 개발에 뛰어들게 하는 계기가 되었다. 앞으로 TPU와 같은 전용 칩은 AI 기술 발전의 핵심 동력이 될 전망이다.

⑤ 어디에 활용되는가?

TPU는 구글 검색, 번역, 지도, 포토 등 구글의 다양한 서비스에서 딥러닝 기술을 구동하는 데 활용된다. 우리가 구글 서비스를 사용할 때마다, 보이지 않는 곳에서 TPU가 열심히 일하고 있는 셈이다.

또한 구글 클라우드를 통해 외부 개발자와 기업에게도 TPU가 제공되어, 다양한 딥러닝 프로젝트에 사용되고 있다. 의료 영상 분석, 자율주행차, 음성 인식 등 딥러닝이 적용되는 다양한 분야에서 TPU의 활약을 기대할 수 있다. 예를 들어, 의료 분야에서는 TPU를 활용해 방대한 의료 이미지 데이터를 분석하여 질병을 조기에 진단하는 시스템을 만들 수 있다.

한편, TPU의 등장으로 AI 칩 시장이 더욱 활기를 띠게 되었고, 다른 기업들도 경쟁적으로 AI 칩을 개발하고 있어 앞으로 TPU와 같은 특화 칩이 더욱 다양해질 전망이다. 이는 AI 기술의 대중화와 산업 전반의 혁신을 가속화할 것이다.

2.6. HBM (High Bandwidth Memory): 데이터 고속도로, 인공지능의 날개를 달아주는 메모리

① 왜 나오게 되었는가?

인공지능, 빅데이터, 고성능 컴퓨팅 등의 기술이 발전하면서, 기존 메모리의 데이터 처리 속도는 한계에 부딪혔다. 특히, 인공지능 모델 학습과 추론에는 대량의 데이터를 빠르게 처리해야 하는데, 기존 메모리로는 병목 현상이 발생하여 성능 저하를 일으켰다. 이를 해결하기 위해 데이터 전송 속도를 획기적으로 높인 새로운 형태의 메모리가 필요했는데, 이것이 바로 HBM의 등장 배경이다.

② 쉽게 설명하면?

HBM은 여러 개의 메모리 칩을 수직으로 쌓아 올려 데이터 전송 속도를 획기적으로 높인 고대역폭 메모리이다. 마치 단층 주택과 고층 아파트의 차이처럼, HBM은 같은 면적에 더 많은 메모리 칩을 집적할 수 있다. 또한, 메모리 칩 사이를 연결하는 선을 대폭 넓혀, 마치 차선이 많은 고속도로처럼 더 많은 데이터를 동시에 주고받을 수 있다. 이를 통해 HBM은 기존 메모리보다 훨씬 빠른 속도로 대량의 데이터를 처리할 수 있게 되었다.

③ 예시와 함께 좀 더 자세히 알아볼까?

HBM은 여러 개의 메모리 칩을 쌓아 올린 후, TSV(Through Silicon Via)라는 기술을 사용하여 칩 간에 수직으로 연결한다. TSV는 마치 칩 사이에 엘리베이터를 설치하는 것처럼, 데이터가 칩 간을 빠르게 이동할 수 있도록 돕는다.

이해를 돕기 위해 우리 일상에서 찾아볼 수 있는 비유를 들어보자. 마치 단층 주택에서 엘리베이터 없이 계단으로만 이동할 수 있는 것처럼, 기존의 메모리는 데이터가 칩 사이를 이동하는 데 제한이 있었다. 하지만 HBM은 TSV라는 엘리베이터를 설치하여, 데이터가 마치 고층 아파트의 엘리베이터를 타고 이동하듯 빠르게 칩 간을 오갈 수 있게 되었다. 이러한 구조 덕분에 HBM은 기존 메모리보다 훨씬 넓은 데이터 전송 대역폭을 제공하며, 전력 소비도 줄일 수 있다.

④ 왜 중요한가?

HBM은 인공지능, 빅데이터, 고성능 컴퓨팅 등 데이터 처리 속도가 중요한 분야에서 핵심적인 역할을 한다. 빠른 데이터 전송 속도는 인공지능 모델의 학습 및 추론 시간을 단축하고, 빅데이터 분석 속도를 향상시키며, 고성능 컴퓨팅 시스템의 성능을 극대화한다.

예를 들어, 자율주행차는 수많은 센서에서 실시간으로 들어오는 대량의 데이터를 빠르게 처리해야 한다. HBM이 이러한 데이터 처리 속도를 높여주어, 자율주행차가 더 안전하고 신뢰성 있게 주행할 수 있도록 돕는다. 또한, 우리가 인터넷 검색을 할 때도 대량의 데이터 속에서 필요한 정보를 빠르게 찾아내야 하는데, HBM이 이러한 빅데이터 처리 속도를 높여 사용자 경험을 개선한다. 이처럼 HBM은 우리 삶의 다양한 영역에서 데이터 처리 속도를 높이는 핵심 요소로 자리매김하고 있다.

⑤ 어디에 활용되는가?

HBM은 인공지능 가속기, 고성능 컴퓨팅 시스템, 그래픽 카드, 네트워크 장비 등 다양한 분야에서 활용된다. 특히, 인공지능 분야에서는 GPU(Graphics Processing Unit)와 함께 사용되어 딥러닝 모델 학습 및 추론 속도를 크게 향상시킨다.

실제로 많은 AI 연구자들과 개발자들이 엔비디아의 테슬라(Tesla) GPU와 같은 HBM 탑재 가속기를 사용하여 더 빠르고 효율적으로 AI 모델을 개발하고 있다. 또한, HBM은 슈퍼컴퓨터, 데이터 센터 서버 등 대규모 데이터 처리가 필요한 시스템에서도 핵심적인 역할을 한다. 세계 최고 성능의 슈퍼컴퓨터들은 대부분 HBM을 사용하여 엄청난 양의 데이터를 빠르게 처리한다. 앞으로도 HBM은 데이터 중심 시대에 필수적인 기술로서, 우리 삶의 더 많은 영역에서 활약할 것으로 기대된다.