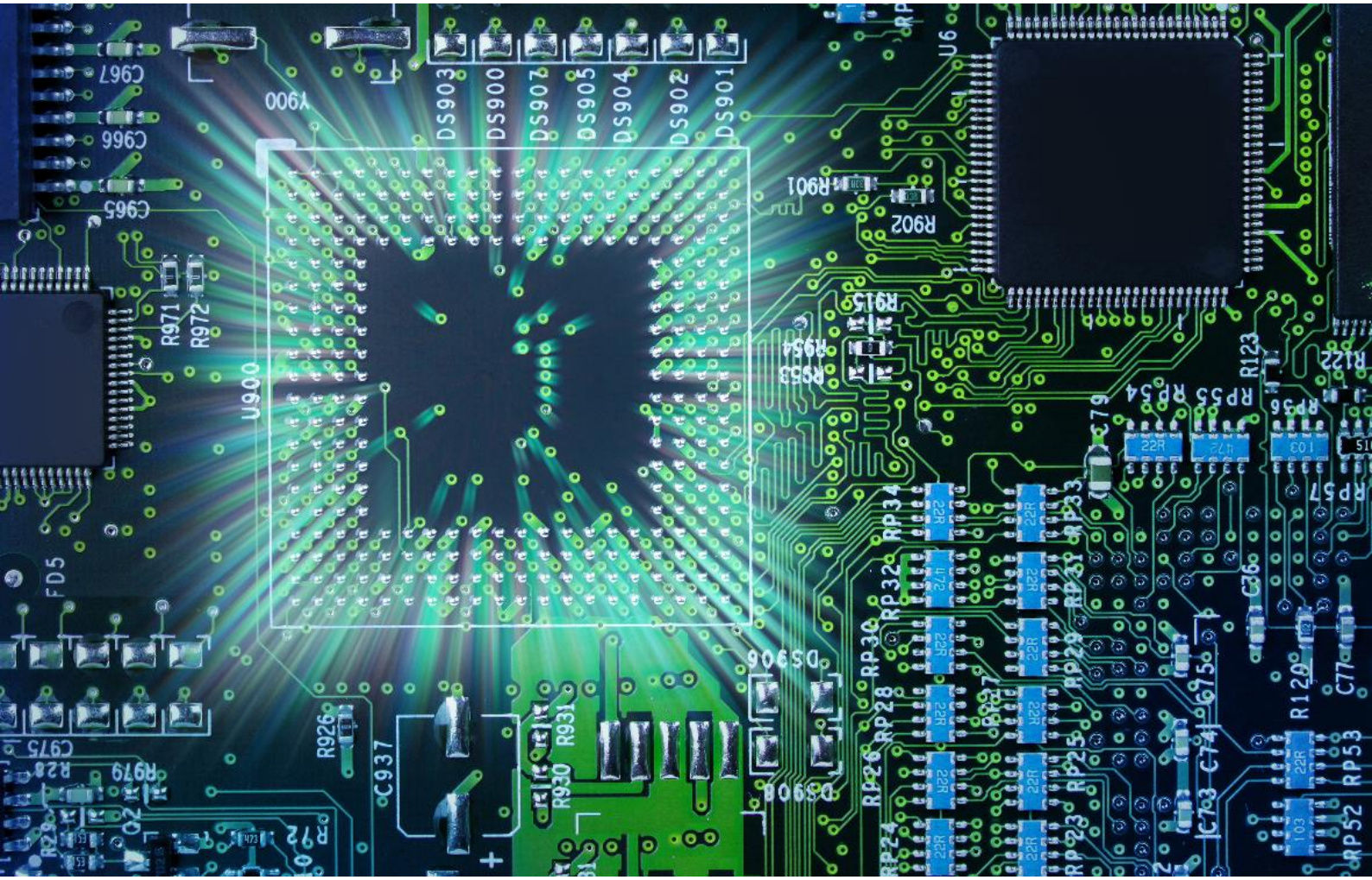
Deloitte.

AI 칩을 향한 무한경쟁, 시작되다

딜로이트 안진회계법인
Audit TMT
손재호 파트너 (Technology, Media & Telecom Industry Leader)
이환수 SM



September 2019

AI 칩을 향한 무한경쟁, 시작되다

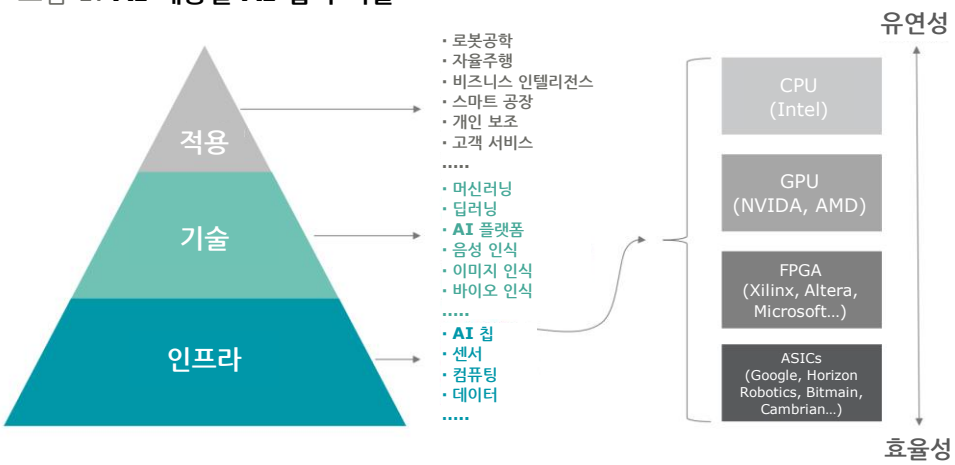
AI 칩 시장 선점을 위한 노력

AI 칩 시장 선점

인공지능 프레임워크(Framework, 소프트웨어의 개발을 돕기 위한 환경 조성 과정)는 크게 세 계층으로 구분 가능하다. 핵심 AI 칩과 빅 데이터는 기반 계층에 속하며, 그로부터 영향을 받는 인지/감지 연산 능력은 기술 계층에 속한다. 가장 높은 계층인 응용 계층에는 자율주행, 인공지능 로봇, 인공지능 보안 그리고 AI 비서 등의 서비스들이 해당된다. AI 칩은 이와 같은 인공지능 기술의 핵심이며, AI 알고리즘, 특히 딥 뉴럴 네트워크(DNN, Deep Neural Networks)에 중추적 역할을 수행한다.

"뎁스"란 뉴런 네트워크 모델의 레이어들과, 그 접점들을 의미한다. 최근 몇 년간 뎁스의 복잡성은 기하급수적으로 증가해왔다. 이는 연산 수행 능력에도 큰 문제를 가져왔다. 기존의 중앙처리장치(CPU, Central Processing Units)는 일정한 규칙이 있는 많은 양의 정보 처리에 강점이 있다. 그러나 CPU는 최근 부상하는 AI 알고리즘 처리에서 나타나는 병렬 처리 문제(Parallelism) 해결에는 역부족이다.

그림 1. **AI 계층별 AI 칩의 역할**



* 출처: 딜로이트 분석

이 병렬 처리 문제를 해결하는 방법은 두 가지가 있다. 첫 번째는 현존하는 컴퓨터 구조에 맞는 전담 가속 장치(Dedicated accelerator)를 추가하는 것이다. 두 번째는 인간 두뇌의 뉴런 네트워크를 모방한 완전히 새로운 구조를 개발하는 것이다. 이 중 두 번째 방법은 아직 개발 초기 단계이며, 따라서 상용화에는 아직 더 많은 시간이 필요하다. 그러므로 첫 번째 방법이 현실적으로 가능한 해결 방법이다. GPU(Graphics Processing Unit)를 필두로 한 주류 칩, FPGA(Field-Programmable Gate Arrays)와 ASIC(Application Specific Integrated Circuits) 그리고 TPU, NPU, VPU와 BPU 등으로 이어지는 여러 종류의 AI 칩들이 가속 장치로 사용될 수 있다. 모두 각각 다른 장단점이 있다.

참고자료: Semiconductors – the Next Wave: Opportunities and Winning Strategies for Semiconductor Companies (Deloitte Taiwan, April 2019)

먼저 게임 등 그래픽 작업에 많이 활용되는 GPU는 병렬 처리 문제를 고려하여 만들어진 칩이다. GPU는 높은 수행 능력을 가지고 있기 때문에 병렬 처리 문제가 많이 발생하는 AI 딥 러닝 알고리즘에 적합하다. 따라서 AI 하드웨어에 적합한 이 특성 덕에 GPU는 AI 훈련을 위한 클라우드와 데이터 저장소 등에 많이 쓰인다. 이 외에 차량이나 보안 영역에도 널리 쓰이는 GPU 칩은 현재 가장 널리 상용화된, 가장 다재다능한 AI 칩이다.

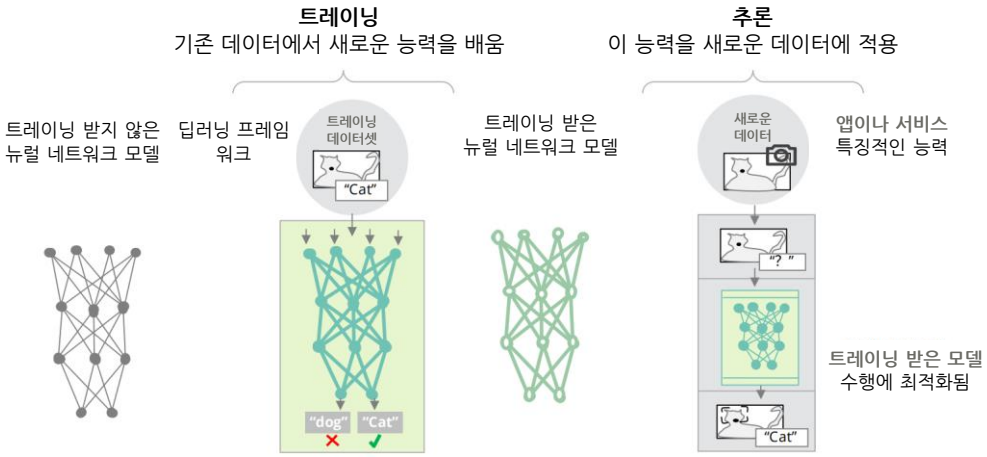
FPGA는 고객의 필요에 따라 칩 배열을 설정할 수 있는 칩이다. FPGA의 특성은 ASIC에 비해 빠른 개발 사이클과 GPU에 비해 낮은 전력 소모율이다. 그러나 그 높은 자유도 때문에 가격 또한 상대적으로 높은 편이다. FPGA는 효율성과 유연성 사이의 적당한 타협점으로 볼 수 있는데, 이 장점은 AI 알고리즘이 아직 확정되지 않은 상황에서 더욱 빛을 발한다. FPGA 사용자들은 ASIC 사용에서 발생할 수 있는 비용적, 기술적 한계들을 회피함과 동시에 칩을 각자 적용 방법에 맞게 커스터마이즈할 수 있다.

ASIC 칩은 그와 반대로, AI 적용을 위한 전용 아키텍처를 보유하고 있다. ASIC는 많은 종류가 있는데, TPU, NPU, VPU, BPU 등이 모두 그 일부다. 이 칩들은 공통적으로 다양한, 고연산의, 규칙이 정해진 정보 처리를 효율적이고 빠르게 처리하면서 동시에 CPU의 범용성까지 잡는 것을 목표로 한다. 기본적으로, ASIC 칩은 더 효율적이고, 다이 사이즈(die size)도 더 작으며, GPU나 FPGA에 비해 전력 소모도 적다. 그러나 그 개발 사이클이 훨씬 길고 유연성도 낮기 때문에, 상용화 정도는 낮다.

딥 러닝에서의 AI 배치(Deployment) 순서엔 두 가지가 있는데, 훈련 과정과 추론 과정이다. AI는 빅 데이터를 뉴런 네트워크 모델을 "훈련"하기 위한 기반으로 사용하는데, 그중에서도 훈련용 데이터 묶음을 다수 활용한다. 이렇게 형성된 뉴런 모델은 새로운 데이터 묶음들을 통해 학습하여 "추론" 능력을 얻게 된다. 훈련 과정에서는 매우 높은 수준의 연산 처리 능력이 필요한데, 엄청난 양의 데이터 묶음을 뉴런 네트워크 모델에 입력해야 하기 때문이다. 즉 방대한 양의 평행한(Parallel) 데이터 묶음을 처리할 수 있는, 병렬 처리 능력이 뛰어난 최첨단 서버가 요구된다. 따라서, 훈련 과정은 보통 중앙 클라우드의 하드웨어에서 처리된다. 반면 추론 과정은 중앙 클라우드에서도 진행 가능하고, 외곽의 기기들에서도 진행 가능하다. 훈련 과정을 거친 칩에 비해, 추론 과정을 거친 칩들은 더 신중한 전력, 비용, 그리고 지연 속도(Latency, 컴퓨터의 입력에서부터 결과가 제공되기까지 걸리는 시간) 배분이 요구된다.

AI 칩의 발전은 이제 막 기지개를 켰고, 기업들은 이러한 발전 양상에 대해 각기 색다른 접근을 하고 있다. 구글은 ASIC 칩을 활용하고 있으며, Microsoft의 경우 FPGA를 사용하면 더 나은 결과를 도출할 수 있다고 주장한다. Xilinx, Baidu 그리고 Amazon은 FPGA 칩에 대한 진입 장벽을 낮추는 데 주력하고 있다.

그림 2. **딥 러닝의 두 가지 순서**



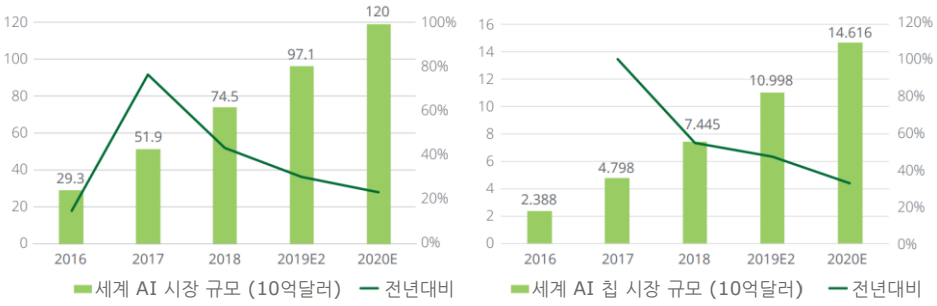
* 출처: nVidia

AI 칩의 폭발적인 성장

AI 칩 시장의 폭발적인 성장 및 다양한 활용방안

AI 칩 시장은 2022년까지 약 54%의 연평균 성장률(CAGR, Compound Annual Growth Rate)로 전체 AI 시장의 12% 이상을 차지할 것으로 분석되고 있다. 미주 국가들이 시장을 이끌 것이라고 예상되며, 그 뒤를 EMEA(유럽, 중동, 아프리카), 그 뒤는 APAC(아시아, 태평양 국가들)가 뒤이을 것으로 보인다. 2022년에는 미주 국가가 시장을 지배할 것으로 예상된다.

그림 3. **세계 AI와 AI 칩 시장(2022)**

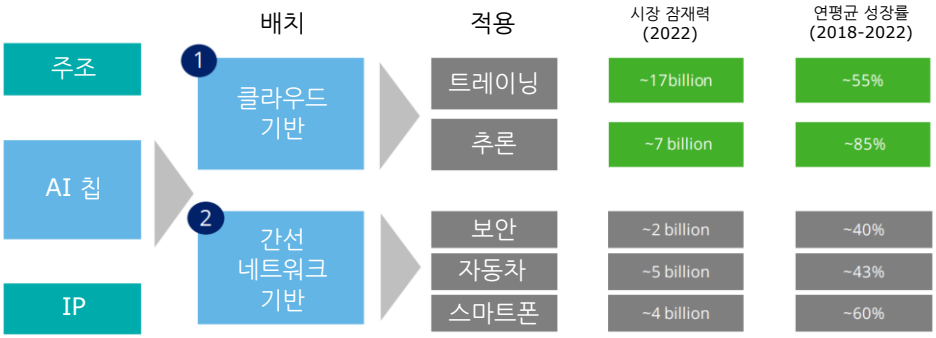


* 출처: CDIC

클라우드 기반 **AI** 칩의 밝은 미래

AI 칩 시장은 배치(Deployment) 방식에 따라 두 가지 분류로 나뉜다: 클라우드 기반과 간선 네트워크 기반(Network Edge, 네트워크의 중심부가 아닌, 외곽의 네트워크).

그림 4. **클라우드 기반 AI 칩의 밝은 미래**



* 출처: CICC, 딜로이트 분석

클라우드 기반 시장은 AI 칩의 최대 시장인데, 이는 AI 칩을 데이터 센터에 적용했을 때 효율성 증가, 작동 비용의 감소와 시설관리가 향상된다는 장점 때문이다. 특히 AI 훈련 시장은 약 170억 달러, 클라우드의 추론형 칩들은 70억 달러에 육박한다. 상품 종류로만 보면 GPU 칩이 시장 전체의 30% 이상을 차지하며 주류 칩으로 자리 잡고 있다.

간선 네트워크 **AI** 칩의 부상

AI 칩의 활용은 클라우드에만 국한되는 것이 아니라, 스마트폰, 자율주행 자동차, 감시카메라 등 간선 네트워크 기기들에도 적용된다. 간선 네트워크에 사용되는 대부분의 AI 칩은 추론형 칩이며, 이들은 갈수록 전문화되어가고 있다. AI 추론형 칩 시장은 약 40%의 연평균성장률이 예상되며, 2022년에는 20억 달러의 시장 규모를 가질 것으로 예상된다.

스마트폰 ASP(application service provider)를 주도하는 AI 칩

AI 칩 공급 업체들은 칩들의 가격이 오름에 따라 이익을 본다. 예를 들어, 애플의 A11 칩의 가격이 27.50달러로 올랐다. AI 칩의 가격이 오르면 자연스럽게 이를 탑재한 스마트폰 가격도 오르게 되고, 스마트폰 생산자들은 이로 인해 이익을 보게 된다. AI 칩 또한 최고사양 휴대폰에만 탑재되다가 중간사양 휴대폰까지 사용 범위가 확대되었기에, 스마트폰 생산자들에게 더 큰 수익을 가져오게 된다. 스마트폰에 탑재되는 AI 추론형 칩은 애플, 삼성, 화웨이 등 스마트폰 생산 업체, Qualcomm과 MediaTek과 같은 독자적 칩 생산 업체, ARM과 Synopsys와 같은 IP 라이선스 제공 업체 사이의 삼자 대결이다. 스마트폰 생산 업체들의 AI 칩들은 자사 휴대폰과 칩의 최적화를 통해 휴대폰 사용자들에게 더 나은 성능과 UI를 제공하기 위해 노력한다. 한편 독자적 칩 생산 업체는 스마트폰 시장을 제외한 나머지 시장에 좀 더 알맞은 스펙으로 칩을 제공한다.

자율주행, AI 칩 적용의 이상향

자율주행은 인공지능 적용에 있어 까다로운 과제임과 동시에 중요한 과제이기도 하다. 자율주행은 AI 추론형 칩들의 주요 개발 원인이며, 50억 달러의 시장 가치와 40%의 연평균 성장률을 가질 것으로 예상된다.

인지하기, 윤곽 잡기, 의사 결정하기는 자율주행에 필요한 세 가지 순서이며, 추론형 칩은 이 모든 과정에 관여하게 된다. 자율주행 중 주변 환경 인지나 장애물 회피 등의 다양한 상황에 AI 칩의 높은 전산처리능력이 요구된다.

지연 속도의 한계가 있기 때문에, 자율주행에서의 전산처리는 클라우드보다는 간선 네트워크에서 이루어지는 것이 실시간 정보통신이 요구되는 자율주행에 더 알맞다. Toyota에 따르면, L5 자율주행을 위해서는 12 TOPS(Tera Operations Per Second)의 정보처리능력이 필요한데, 현재 대부분의 칩들은 2~3 TOPS 정도밖에 되지 않는다. 따라서 이러한 방대한 양의 연산 처리를 클라우드에서 하는 것보다는, 간선 네트워크에서 하는 것이 필요하다.

OEM들은 가장 적합한 구매자를 찾기 위해 생산자들의 칩을 테스트한다. 대형 OEM들은 자신들만의 자율 주행 플랫폼을 새로 건설하고 AI 칩을 별도로 구매하는 것을 선호하지만, 신진 OEM들은 이미 완성된 자율 주행 플랫폼들을 구매하는 것을 선호한다. 시간이 지날수록 간선 처리(Local Processing, 중앙 클라우드를 거치지 않고 외곽의 간선 네트워크만 활용하여 데이터 처리를 하는 것)를 통해 이득을 볼 수 있는 AI 적용 사례는 증가할 것인데, 그 대표적인 예가 애플의 Face ID이다.

스마트 보안 시스템의 수요 증가하다

가정용 보안 감시 시스템 또한 AI의 개발로 인해 스마트해지고 있다. 근 10년 동안 보안 시스템은 세 단계의 진화를 겪어 왔다. 첫번째는 HD 세대이며, 이 때 고화질의 영상을 촬영할 수 있게 되었다. 두번째는 네트워크 세대인데, 네트워킹과 상호 연결 능력이 추가된 것이 이 시기이다. AI의 추가는 세번째 세대라고 할 수 있다. AI 추론형 칩이 간선 네트워크의 카메라에 추가되면서, 영상 정보를 실시간으로 처리할 수 있게 되었다. 이는 클라우드의 저장 공간을 보존할 수 있으며, 간선 네트워크상에서 매일 발생하는 엄청난 양의 정보를 즉시 처리하여 보안 시스템의 성능을 향상시킨다.

AI 칩의 떠오르는 핫스팟, 중국

중국에서 AI 칩 펀딩은 매우 활발하며, M&A 또한 증가하고 있다. 한 대표적인 예는 자동학습, 고도압축, 프루닝(Pruning, 결정 트리에서 생략 가능한 규칙을 결정해 탐색 공간을 줄일 수 있게 하는 작업), 뉴런 네트워크의 시스템 최적화로 유명한 스타트업 회사인 DeePhi가 초거물회사 Xilinx에 합병된 것이다. Alibaba, Baidu, Huawei 등의 IT 거물들 또한 속속들이 참전 중이다. 특히, Huawei는 스마트폰 분야에서 AI 칩 경쟁을 주도하고 있다. 일부 비트코인 채굴 장비 생산자들도 AI 최적화 경쟁에 뛰어들었다.

중국의 AI 회사들은 비즈니스모델의 혁신이나 즉시 적용 가능한 사업 기회에 항상 주의를 기울이고 있다. 그러나 현재 중국 내의 AI 연구개발이 전산화된, 새로운 AI 프레임워크보다는 기존의 모델들을 수정 보완하는 데에 집중함에 따라 중국의 실질적인 Original AI 모델 개발 능력은 떨어진다. 또한, 중국은 AI 전담 훈련 분야도 미국과 같은 국가들에 비해 많이 부족하다.

AI 시장을 살펴보자

인공지능의 부상은 반도체 기기, 특히 AI 칩들에게 큰 기회가 되었다. AI 시장에 들어왔거나 들어올 예정인 반도체 회사들은 다음 사항들을 따르는 것을 추천한다.

특성화가 AI 칩의 핵심이다

AI 칩의 특성화 및 반도체 처리 기술

미래에는 AI 칩 회사들은 단순히 하드웨어 제작 회사에서 그치는 것이 아니라, 고객들의 요구를 깊이 이해하고 그에 맞는 제품을 생산할 수 있어야 한다. 현재 고객들은 일반 목적의 칩에 약간의 인공지능을 추가한 상품 정도를 원하지 않는다. 그들이 원하는 것은 합리적인 가격에 사업적 필수조건들을 충족할 수 있는 상품이다. 고객들은 전력, 수행능력 그리고 비용을 적당히 조절해야 한다. 따라서 컴퓨터적 밀도(Computational Density, computational Power per unit of power consumed: 소모한 전력 대비 계산력)가 AI 칩 생산자들에게는 핵심 경쟁력이 될 것이다.

중심(클라우드)에서 외곽(간선 네트워크)으로

간선 네트워크에서 다양한 사업 기회들이 생겨났고 많은 기업들이 클라우드에서 간선 네트워크로 이동하기 시작했다. 그 기업들은 통합적인 AI 솔루션을 제시하여 훈련부터 추론까지 전체적인 AI 스펙트럼을 넓혔다. 최근 AI 시스템은 본 노이만의 구조를-즉 프로세싱와 기억 장치가 별개인-따르고 있다는 점을 기억할 필요가 있다. 그로 인해 AI가 전력을 더 많이 요구하게 되었고, 결국 뉴런 네트워크들은 클라우드에 남겨지게 되었다. 최근에는 프로세서와 기억 장치가 좀 더 긴밀하게 이어져 에너지 효율과 능률을 끌어올리는 새로운 구조를 만들려는 노력이 이어지고 있다. 그 골자는 메모리에 새로운 기능을 추가해서 프로세서를 대체하지 않고 기기가 더 스마트해진다는 것이다. 반도체 시장은 이런 구조를 받아들여서 AI가 클라우드에서 벗어나 간선 네트워크로 나오도록 해야 한다.

적합한 반도체 처리 기술을 골라라

무어의 법칙에 따르면 항상 가장 진보된 프로세싱 기술이 필요한 CPU와 달리, AI에는 병렬 처리 문제가 있기 때문에 가장 최신식 기술만이 좋은 것은 아니다. 예를 들어 40nm 공정과 28nm 공정은 1 TOPS의 계산력(Computational Power)을 충분히 제공한다. 그 이전의 프로세스들은 발달한 도구 세트와 구성 요소에 접근이 가능하다. 많은 반도체 생산 공장들은 전력&성능에 따라 28nm에서 7nm까지 넓은 범위의 첨단 프로세싱 기술을 제공한다. 반도체 업체들이 적절한 반도체 처리 기술을 고르는 적합한 기준은 컴퓨터적 능력(Computational capability)과 전력 소모, 그리고 폼 팩터(Form Factor, 컴퓨터 하드웨어의 크기, 구성, 물리적 배열)이다.

소프트웨어 지원은 필수적이다

반도체 회사에서 표준형 오픈소스 소프트웨어 체제를 얼마나 지원하는지가 AI 시장에서의 경쟁력을 결정한다. 이는 후발주자들에게는 더 중요한데, 선두주자들은 이미 반도체 칩에 대한 딥 러닝 소프트웨어와 도구들을 모두 지원 중이기 때문이다. 시장에서 영향력 있는 경쟁자가 되려면, 반도체 생산 업체는 최소한 주류 오픈소스 소프트웨어 체제인 TensorFlow, Caffe2, Theano, CNTK, MXNet, Torch 등은 지원해야 한다. 개발자들이 애플리케이션을 개발하기 위한 도구들 또한 필요하다. 반도체 업체들은 소프트웨어에 투자하고, 소프트웨어 개발자들과 협업을 통해 AI 기기 구성에 접근해야 한다. 뉴런 네트워크를 프로세싱하기 위한 소프트웨어 프레임워크는 점점 증가하고 있다. 이후 몇 년간 이 중 많은 양이 개발되고 출시될 것이기 때문에, 아직 신흥 주자들이 발전할 자리는 충분하다.

AI 칩 너머 기회를 보라

인공지능은 AI 칩으로만 연산 처리를 하는 것은 아니다. AI의 발전에 있어 메모리도 상당히 중요한 요소인데, 높은 정보량의 병렬 처리는 메모리 시스템상의 데이터 대역폭에 많은 부담을 주기 때문이다. AI 시스템 메모리에 대한 수요는 메모리 생산 업체들에게는 기회다. 또, AI 시스템이 커짐에 따라 기기와 서브시스템들 사이에는 병목 현상이 일어날 것이다. 따라서 이 많은 양의 정보 처리를 위한 초고속 연결 개발도 반도체 업체들에게 기회가 될 수 있다. 이와 더불어 최근의 AI 칩들은 병렬 처리를 극대화하기 위해 많은 수의 프로세서를 탑재할 수 있고, 이는 매우 큰 다이 사이즈를 의미한다. 이는 고온과 고압 관리가 필요함을 의미하며, 따라서 자체적 냉각 구조가 필요할 것이다. 이는 포장 업체들이 더 얇은 폼 팩터와, 더 적은 열 손실을 가진 가성비 좋은 상품을 개발할 기회다.

Contact

손재호 파트너
Technology, Media & Telecom Industry Leader
jaehoson@deloitte.com

Contact

이환수 Senior Manager
Audit TMT
hwalee@deloitte.com

Deloitte.