ETRI

2018
Electronics and
Telecommunications
Trends

인공지능 프로세서 기술 동향

AI Processor Technology Trends

권영수 (Youngsu Kwon, yskwon@etri.re.kr) 프로세서연구그룹 책임연구원/그룹장

The Von Neumann based architecture of the modern computer has dominated the computing industry for the past 50 years, sparking the digital revolution and propelling us into today's information age. Recent research focus and market trends have shown significant effort toward the advancement and application of artificial intelligence technologies. Although artificial intelligence has been studied for decades since the Turing machine was first introduced, the field has recently emerged into the spotlight thanks to remarkable milestones such as AlexNet-CNN and Alpha-Go, whose neural-network based deep learning methods have achieved a ground-breaking performance superior to existing recognition, classification, and decision algorithms. Unprecedented results in a wide variety of applications (drones, autonomous driving, robots, stock markets, computer vision, voice, and so on) have signaled the beginning of a golden age for artificial intelligence after 40 years of relative dormancy. Algorithmic research continues to progress at a breath-taking pace as evidenced by the rate of new neural networks being announced. However, traditional Von Neumann based architectures have proven to be inadequate in terms of computation power, and inherently inefficient in their processing of vastly parallel computations, which is a characteristic of deep neural networks. Consequently, global conglomerates such as Intel, Huawei, and Google, as well as large domestic corporations and fabless companies are developing dedicated semiconductor chips customized for artificial intelligence computations. The AI Processor Research Laboratory at ETRI is focusing on the research and development of super low-power AI processor chips. In this article, we present the current trends in computation platform, parallel processing, AI processor, and super-threaded AI processor research being conducted at ETRI.

* DOI: 10.22648/ETRI.2018.J.330513

Ⅰ. 서론: 병렬컴퓨팅의 대중화

폰 노이만(John von Neumann, 1903~1957)은 컴퓨터 구조뿐만 아니라 양자역학, 함수 해석학, 통계론 등 다양한 분야에 업적을 남겼다. 폰 노이만이 1945년 핵무기를 개발하던 맨해튼 프로젝트에 참여할 당시 뉴멕시코의 Los Alamos로 돌아가는 기차에서 수기로 작성하였던 리포트에서, 그는 현대 컴퓨터의 아키텍처를 구성하는 근간이 된 von Neumann architecture(폰노이만 아키텍처) 즉, Stored-Program Computer(프로그램 저장형 컴퓨터)를 제안한다[1]. 당시의 ENIAC과 같은 컴퓨터는 Program-Controlled Computer 구조였다. 즉, 컴퓨터가 무슨 일을 할지 알려주기 위한 작업인 프로그래밍(Programming)을 하기 위해서 컴퓨터 내부의 스위치를 일일이 변경하는 작업을 통하여 컴퓨터의 구조를 목적에 맞게 변경했던 것이다. 즉, 프로그래밍을 위해서 컴퓨터의 아키텍처를 변경해야 하는 것이다. 반면 폰 노이만 아키텍처에서는 컴퓨터 프로그래밍을 위한 Instruction(명령어)과 Data(데이터)가 메모리(Memory)에 저장되고, CPU(Central Processing Unit)가 Instruction을 읽어서 어떤 연산을 할지를 결정한다. von Neumann architecture를 구상하면서 목적하였던 것은 High-speed computing architecture였다. 전쟁에서의 승리이든 여하한 목적이든 상관없이 당시에도 빠른 컴퓨팅에 대한 사회적 요구는 존재하였던 것이다. 이후, 수십 년의 시간이 흐르는 동안 인류가 사용하고 있는 거의 모든 컴퓨터는 폰 노이만 구조를 채택하여 설계되었고, 지금도 우리 일상생활에서 필요한 컴퓨팅(Computing)을 실행(Execution)하고 있다. 폰 노이만의 리포트에는 진공관으로 만들어지는 프로세서의 회로 설계, 메모리 설계, Instruction의 구조와 Immediate(특정 값을 가진 숫자), Instruction Branch(명령어 분기)의 개념 등을 제시하고 있는데 실제로 현재의 프로세서 반도체와 컴파일러의 실제적인 개념을 제시한다. 그런데, 폰 노이만 아키텍처의 특징 중 하나는 Sequential computing(순차 처리)이다. 그가 Sequential computing을 굳이 특정하고자 했는지는 확실치 않지만 폰 노이만 아키텍처의 구조적인 특징에 따르면 이것은 메모리에 저장된 Instruction을 CPU가 하나씩 읽어들여 순서대로 처리하는 순차 처리방식인 것이다.

순차처리 방식의 한계는 결국 성능한계이다. High-level programming language인 C나 최근 사용빈도가 높아지고 있는 Python으로 각 연구자의 의도를 기술(Description)하고 나면 컴파일러는 이것을 컴퓨터가 이해할 수 있는 수십 MB(Mega Bytes) 이상의 Instruction의 조합(Sequence)으로 변환한다. CPU는 끝이 없는 In-struction sequence를 한 번에 한 개씩 실행(Execution)하므로 일정시간 동안 실행할 수 있는 Instruction의 개수가 제한될 수밖에 없는 것이다.

병렬 컴퓨팅(Parallel Computing)은 순차 처리의 한계를 극복하기 위한 컴퓨팅 아키텍처이다. 병렬 컴퓨팅은 소프트웨어 또는 컴퓨터 아키텍처의 Parallelism(병렬성)에 그 기반을 둔다. 흥미로운 것은 von Neumann architecture의 리포트의 서두인 4.2절에 이미 컴퓨터 아키텍처의 근본은 Neurons of higher animals와 흡사한 구조를 가지고 있음을 밝히고 있다. 즉, 컴퓨터의 구조가 처음으로 개발될 때부터 컴퓨터라는 개념 자체는 인간의 두뇌와 흡사한 구조를 염두에 두고 설계된 것으로 확대해석할 수도 있다. 인간의 뇌는 이미 Parallelism을 가지고 있다. 우리 뇌의 수십억 개의 뉴런과 수조 개의 시냅스는 시각, 청각, 촉각, 미각, 후각 등의 감각을 동시 처리할 뿐만 아니라, 기억, 사고, 반사작용 등의 고차원적인 컴퓨팅을 병렬적으로 실행하는 능력을 이미 갖추고 있다. 컴퓨터 설계자들은 이미 그 개념을 알고 있었으며, 1958년에 병렬 프로그래밍의 개념이 논문으로 제시된 바 있다[2]. 이후 수십 년간에 걸친 컴퓨터의 역사를 통하여 수많은 연구자가 병렬 컴퓨팅을 연구해 왔다. 즉, 병렬 컴퓨팅은 컴퓨터가 더 빠르게 연산을 실

행하도록 하기 위한 컴퓨터 아키텍처 연구로 이어지는 것이다.

병렬컴퓨팅은 소프트웨어 프로그래밍과 컴퓨터 아키텍처의 두 분야로 구분될 수 있다. 사실상 Parallelism이 온전한 성능을 발휘하려면 소프트웨어와 컴퓨터 아키텍처가 동시에 Parallelism을 만족하는 상황에서 가장 높은 성능을 발휘할 수 있다. 그러나, 기술적인 이유로 Parallel Computing을 실현하기 위한 많은 제약이 존재한다. Parallel Computing을 제약하는 요인으로는 대표적으로 Dependency(의존성)를 들 수 있다. 폰 노이만 아키텍처에서 Instruction의 개념을 제안한 것에서 보듯이 인간이 컴퓨터가 특정 연산을 실행하도록 명령할 때 스스로 명령을 내리는 행위 자체를 간단하게 만들기 위하여 순차처리의 개념이 만들어진 것이다. 따라서, 어떤 목적을 실행하기 위하여 A라는 작업과 B라는 작업을 해야 한다고 알려주었다면, 일반적으로 A 작업이 끝난 후에 B 작업을 할 수 있다는 제한조건, 즉, Dependency가 발생한다. 이것은 A 작업과 B 작업이 동시에 실행되는 Parallelism의 조건이 만족되지 않도록 하는 가장 결정적인 문제이다. Dependency와 연결된 문제이긴 하나, 프로세서 간의 Synchronization, Mutual Exclusion 등의 문제, 그리고 여기에서 발생하는 Race condition 등이 Parallelism을 방해하는 소프트웨어적인 요소이다(본 개념은 본 고의 범위를 초과함). 소프트웨어를 만드는 프로그래머, 어플리케이션 개발자와 연구자들은 프로그래밍을 하는데 있어서 순차처리의 개념이 Parallelism을 살리는 것보다 더 용이했다. IBM에서 1990년대 초반 Octopus와 같은 Parallel Compiler를 연구하였고, 2000년대 초에는 OpenMP, Heterogeneous parallel computing을 위한 OpenCL, 좀 더 발전된 개념의 Open-ACC 등의 Open API(Application Programming Interface)를 통하여 Parallel Computing Software를 대중화하고자 하는 여러 가지 시도가 있었지만, 일부 성공적인 시도였음에도 불구하고 패러다임 자체를 변화시킬 정도는 되지 못했다. 소프트웨어 개발자들은 오히려 C, C++에서 Java, C#, Ruby, Python 등으로 이어지는 일련의 노력에서 보듯이 소프트웨어 프로그래머가 방대한 규모의 서비스를 개발할 때 발생하는 개발 시간, Human Resource, 실수에 의한 Bug를 줄여서 생산성을 높이기 위한 연구를 수십 년간 진행하여 왔다.

Parallel Computing을 실현하기 위해서 하드웨어, 즉, 프로세서 반도체 연구자들은 수십 년간 다양한 아키텍처 구조를 제시하였고 성공적으로 상용화하였다. 그들은 생산성을 높인 소프트웨어가 가진 근본적인 Parallelism limitation을 불가항력적으로 수용하기는 하였으나, 기술적인 다른 방법 즉, 프로세서 아키텍처(Processor architecture)의 변화, 반도체 공정(Semiconductor Process)의 미세화를 무기로 비약적인 성능 발전을 이루어냈다. 아키텍처 측면에서 보면 x86 기반의 프로세서와 PC 산업의 성장 시기에는 Instruction pipelining, Out-of-order execution, Register renaming, Reordering buffer, Instruction and data cache, L1, L2, L3 Cache architecture, Branch prediction, Branch Target Buffer 등 폰 노이만 아키텍처에서의 Instruction을 Basic block 단위로 편성하여 Instruction Parallelism을 극대화하기 위한 수많은 연구가 진행되었고, 현재도 사용되고 있다. 최근의 Spectre & Meltdown 보안 오류(Bug)는 이러한 Instruction Parallelism 아키텍처의 근본적인 문제점에서 기인한다. Instruction Parallelism에서 한계를 느낀 연구자들은 좀 더 소프트웨어 수준에 근접한 Processor parallelism에 집중하였고, SMP(Symmetric Multi-Processing), NUMA(Non-Uniform Memory Architecture), SMT(Simultaneous Multi-Threading), SIMD(Single-Instruction Multiple Data), MIMD(Multiple Instruction Multiple Data), VLIW(Very Long Instruction Word) 등의 다양한 개념을 만들어 내었고, 이들 중 대부분은 이미 상용화되어 우리의 스마트폰과 PC에서 현재도 채택되어 생산되고 있

는 프로세서 아키텍처이다. Processor Parallelism 아키텍처 기법에서 보이는 핵심기술은 Thread, 즉, 다수의 소프트웨어 흐름을 한 개의 프로세서 반도체에서 동시 실행하거나, 다수의 데이터를 소수의 Instruction으로 동시처리 하는 것이다. 이러한 기법은 다수의 Thread를 실행하는 운영체제(OS: Operating System, 운영체제) 기반의 프로세서에서는 효율적이지만, 그 자체가 대량의 데이터를 연산해야 하는 구조에는 Parallelism에 의한 성능이 그다지 뛰어나지 못하다. 이것은 Amdahl's Law로 간단하게 설명될 수 있다. 즉, 전체 소프트웨어 중 Parallelism이 적용되는 소프트웨어 구간이 큰 비중을 차지하지 않으면 일반 사용자는 그 성능 증대를 느낄 수 없는 것이다. 2010년 이후 인텔의 Core x86 프로세서 계열의 제품이 지속적으로 출시되고 있기는 하지만, 1990년에서 2000년대 중반까지 일반 사용자가 느끼던 성능 향상은 보이지 않고 있다. 이것은 Processor Parallelism 또는 Parallel Processor architecture 중심의 연구개발이 한계에 달했음을 이미 많은 연구자들이 체감하고 있었음을 짐작케 한다.

Processor Parallelism이 정체되어 있던 기간에도 프로세서 산업의 발전은 지속적으로 이루어졌고, 성능 향상을 위해서는 반도체 공정(Semiconductor Process)의 발전이 큰 비중을 차지하였다. 90년대 중반의 350nm 반도체 공정에서 10MHz 정도의 동작주파수를 가지던 것이, 수십 년간의 공정기술 연구자들의 노력에 힘입어 현재는 14nm, 7nm 수준으로 미세화되면서, 이제는 국내 대기업에서도 2GHz(2,000MHz)의 동작주파수를 가지는 스마트폰 어플리케이션 프로세서를 양산하고 있는 상황이 도래한 것이다. 하지만, 2010년 중반을 넘어가면서 반도체 공정의 물리적 한계(Physical limitation)에 의하여 거의 한계에 이르렀다는 의견이 지배적이며 실제로 원자 수준 이하의 미세화가 가능한 반도체를 생산할 수 있는 기술이 없기 때문에 지난 수십 년간 업계가 기대어 왔던 반도체 공정 미세화에 의한 성능 증대는 한계에 다다를 수밖에 없다는 의견이 지배적이다. 최근에는 FinFET와같이 전력 소모량을 획기적으로 줄이는 새로운 트랜지스터의 구조가 등장하여 14nm 이하의 공정을 제패하고 SGT(Surrounding Gate Transistor)와 같은 구조도 스타트업에 의하여 개발되고 있다. 트랜지스터의 신구조와 달리 공정 미세화는 앞으로 십 년 이상 지속하기는 매우 어려울 것으로 전망되며, 완전히 새로운 형태의 반도체 소재와 디바이스가 등장해야 할 것이다.

그렇다면, 지금 반도체의 미래를 혁신할 기술은 무엇인가? 2011년 이후로 인공지능이 컴퓨터 관련 학계와 IT 서비스를 제공하는 업계를 혁신할 Groundbreaking Technology 로 등장하면서 인공지능이 요구하는 거대한 컴퓨팅 성능(Computing Performance)를 어떻게 해결할 것인지를 많은 컴퓨터 과학자들이 고민하고 있다. 이제 반도체 기술은 반도체의 재료나 물성, 트랜지스터의 구조에 의한 혁신보다는 인공지능 서비스의 대중화 및 고도화에서 시작된 AI Computing(인공지능 컴퓨팅)의 시대로 발전하고 있는 것이다. 인공지능 컴퓨팅은 인공지능이 요구하는 거대한 컴퓨팅 성능을 만족시켜줄 반도체 기술이 현재 존재하지 않는다는 것에서 그 필요성을 절감할 수 있다. 일반적으로 최소 크기(440×440)의 영상에 존재하는 물체의 종류와 위치를 실시간으로 인식하기 위해서는 1.0TFLOPS, 즉, 1초당 1조 개의 부동소수점 연산을 실행해야 할 정도로 많은 연산량을 필요로 한다. 우리가 사용하는 최신의 스마트폰 내의 어플리케이션 프로세서의 연산성능이 10.5GFLOPS 임을 감안하면 인공지능은 무려 1,000배의 연산성능을 요구하는 것이다.

II. 인공지능 반도체의 개념

1. 딥러닝, 병렬컴퓨팅, 인공지능 반도체

2010년 이후, 인공지능이 정보통신 산업 전체를 관통

하는 큰 반향을 얻게 된 것은 딥러닝의 영향이 매우 크다. 딥러닝(Deep Learning)은 다층(Multi-Layer)을 학습(Training)시키는 인공신경망(Artificial Neural Net) 구조를 의미한다. 딥러닝의 다층 구조는 인간의 뇌가 그러한 구조로 이루어져 있을 것이라는 추측에서 유래한다. 인간과 똑같은 생각을 하고 행동하는 컴퓨터를 만들고자 하는 것은 컴퓨터의 개념이 처음 만들어질 때부터 등장한 개념이었다. 앨런 튜링(1912~1954)은 그의 저서에서 현명한 기계(Intelligent machine)를 정의하는 Turing test를 제안하였다[3]. 인간의 뇌를 모사한 퍼셉트론(Perceptron)을 기본으로 인공신경망을 구성하는 연구가 1960~1970년대에 수행된 이후 인공지능 기술은 오랜 시간 동안 침체기를 겪었다. 침체기의 근본적인 원인은 알려진 바와 같이 인간이 쉽게 풀 수 있는 매우 간단한 문제라 하더라도 인공지능으로 같은 문제를 해결하려면 매우 많은 컴퓨팅 자원(Computing resource)이 필요함을 알게 되었기 때문이다.

인공지능이 최근 각광을 받게 된 것은 GPU(Graphics Processing Unit)의 역할이 컸다. NVIDIA는 90년대 초반부터 CPU를 개발하다가 인텔과의 기술 격차를 극복할 수 없음을 확인한 후, 게임 사용자를 대상으로 하는 GPU 시장을 성공적으로 공략하였다. GPGPU(General Purpose GPU)는 동일한 GPU 반도체를 이용하되 그래픽스의 각 Vertex와 Pixel 컴퓨팅을 담당하는 수천 개의 Shader라는 소형 프로세서 코어를 이용하여 과학계산이나 물리 시뮬레이션과 같은 일반적인 컴퓨팅 어플리케이션을 위하여 개발된 반도체와 소프트웨어인 것이다. 간단한 인공신경망은 영상인식 관련 학회 등에서 지속적으로 연구, 개발되고 있었지만 2011년까지 인류가 개발한 컴퓨터 기반 알고리즘의 영상인식 성능은 74% 수준으로 알려져 있었다. 2011년 ILSVRC 학회의 영상인식 대회에서 딥러닝 뉴럴넷 기반의 영상인식 알고리즘을 GPU에 구현한 논문으로 토론토 대학의 Alex Krizhevsky는 그의 지도교수 Geoffrey Hinton과 함께 수십 년간 조금씩 향상되고 있던 영상인식 성능을 단번에 89.3%로 올려놓으면서 큰 주목을 받게 된다. 지난 2016년에는 구글의 알파고(AlphaGo)가 우리나라에서 바둑 게임을 펼치면서 가까운 미래에 인공지능이 세상을 지배할지도 모른다는 수많은 기사가 쏟아지면서 이슈를 이끌었다. 알파고의 핵심 알고리즘은 강화학습(Reinforcement Learning)이며 이것은 행위(Action)에 대한 가치(Value)와 최종적으로 얻게 된 이익(Reward)에 따라서 딥러닝, 즉, 인공신경망을 학습하는 것이다[4]. 알파고의 경우 두 개의 인공신경망, 즉, value network와 policy network로 구성되며 value network는 각 행위가 가지게 될 가치를 예측하고, policy network는 Action을 결정하기 위한 것이다. 알파고 역시 딥러닝 알고리즘으로 학습한 인공신경망을 활용한 알고리즘에 불과하며, 바둑 게임에서는 일정한 시간 내에 인공신경망 연산을 종료하여야 다음의 수, 즉, Action을 할 수 있기 때문에 이 알고리즘을 수행하기 위해서 수백 개의 CPU와 GPU가 사용되었다. 당시 Google이 개발한 TPU라는 반도체가 이용되었다고 알려져 있는데, 어느 버전의 반도체가 어떤 연산 기능 구현을 위하여 사용되었는지는 확실치 않다. 다만, 딥러닝과 인공신경망이 인공지능의 발전을 이끌었고 알파고 역시 인공신경망에 근간을 둔 알고리즘이며 인공신경망 계산을 위하여 거대한 규모와 엄청난 전력을 소모하는 슈퍼컴퓨터를 사용하였다는 것이 주목해야 할 사실이다.

인공신경망과 딥러닝이 갑작스런 주목을 받게 된 것은 결국 GPU와 NVIDIA가 제공하는 CUDA라는 GPGPU를 위한 소프트웨어 API가 큰 영향을 주었다는 것에는 반론의 여지가 없다. 근본적으로 일반인이나 학생이 쉽게 구매하여 접근 가능한 GPU 반도체에 수천 개의 프로세서 코어가 동시 동작할 수 있는 병렬컴퓨팅(Parallel computing)의 미래를 예상하고 시장 출시가 가능한 상태로 미리 구현해 두었던 영향이 매우 컸던 것이다. 무엇보다도 2011년의 AlexNet(Alex Krizhevsky

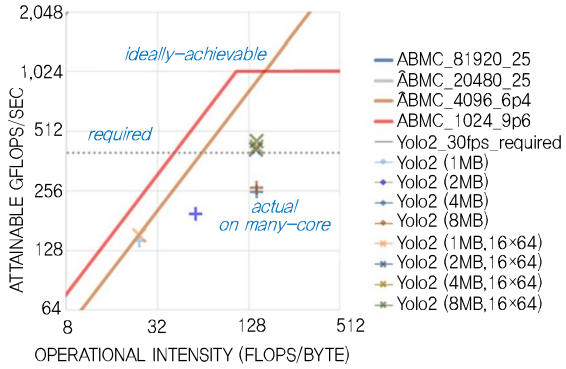
가 개발한 인공신경망)의 의미가 큰 것은 상당히 깊은 레이어를 가지는 인공신경망을 기반으로 구성한 CNN (Convolutional Neural Network)에서 특정 학습(Training) 알고리즘들, 특히, SGD(Stochastic Gradient Descent), Normalization, Overfitting prevention, Activation, ReLU 등의 알고리즘적인 기법을 동원하면 인간의 뇌를 모델링한 인공신경망이 실제로 인간처럼 동작할 수 있는 가능성이 있음을 발견한 것이며, 그러한 알고리즘이 실제로 동작가능하며 의미 있는 인공신경망으로 동작할 수 있다는 것을 실현해준 것은 바로 반도체의 역할이 결정적이라고 볼 수 있다.

인공지능 컴퓨팅을 위한 반도체, 즉, 인공지능 반도체가 수십 년간 연구되어오던 인공신경망과 인공지능 알고리즘을 상용화가 가능한 본 궤도로 올려놓았다. 2017년부터 전세계 글로벌 기업 대부분이 인공지능이 앞으로의 정보통신 산업 전반을 바꾸어놓을 것이라는 인식하에 인공지능 반도체를 개발하고 있으며 기존의 반도체 기업뿐만 아니라 서버기반의 인터넷 기업도 인공지능 반도체 개발 사업에 뛰어들기 시작하였고, 2018년에는 상당한 수준의 컴퓨팅 성능을 갖춘 인공지능 반도체들이 이미 발표되고 있다. 하지만, 앞서 말한 바와 같이 최종적으로 인간이 바라는 수준, 즉, 인간과 비슷한 수준의 사고가 가능한 인공신경망을 우리 손 위에 올려놓을 만한 작은 반도체에서 구현하기 위해서는 아직 넘어야 할 기술적 장애물이 분명히 존재하고 있기 때문에, 앞으로 수년간은 낮은 수준의 인공지능 반도체의 상용화가 본격적으로 이루어지고 동시에 고차원의 인공지능 컴퓨팅을 위한 반도체 기술 개발이 지속적으로 이루어지면서 아마도 수년 후에는 SF영화에서 볼 수 있던 휴머노이드를 현실에서 볼 날이 올 것이다.

2. 인공신경망 반도체와 뉴런-시냅스 반도체

인공신경망 반도체는 거대한 다층 레이어(Multi-Layer)로 구성된 인공신경망을 학습(Training)하거나 학습된 결과로부터 영상, 음성 인식 등 특정 기능을 수행(Inference)하는 반도체이다. 딥뉴럴넷(Deep Neural Net)은 다수의 퍼셉트론(Perceptron)으로 구성되어 있다. 딥뉴럴넷은 실재하는 것이 아니라 알고리즘상에서 존재하는 개념이며 퍼셉트론은 간단히 말하면 비중을 가진 입력값(Weighted inputs)들의 합(Summation)을 구하는 노드(Node)이다. 각 레이어는 수십 개에서 수천, 수만 개 이상의 노드로 구성될 수 있으며, 이것은 목적지향적으로 구성된 인공신경망의 구조에 따라서 매우 다양한 형태를 가진다. 2015년 전후로 전세계적으로 인공지능 관련 학회가 매우 큰 인기를 끌고 있는데, 딥러닝, 인공신경망의 구조, 인공신경망의 학습 방법, One-shot training과 같이 1일 또는 1주 이상 걸리는 학습시간을 단축하여 단시간에 인공신경망을 학습시키는 방법, 인공신경망의 학습 결과 분석, 로봇을 위한 전용 인공신경망 구조 연구에 전세계의 수많은 연구자가 뛰어들고 있음을 확인할 수 있다. 특이한 것은, 인공신경망의 학습 결과 분석(Analysis of trained network)연구에서 보듯 인공신경망을 SGD와 같은 방법으로 학습(Training)을 하여 도출한 Trained network weight가 어떻게 영상의 픽셀 데이터로부터 물체의 종류에 대한 확률값을 출력하여 주는지 확실히 밝혀진 이론이 없다는 점이다. 인공신경망은 인간 두뇌의 예상 구조를 구현한 알고리즘이고 학습 알고리즘 역시 추측에 의해서 만들어진 것이기 때문에 그러한 학습 방법을 사용하면 왜 이런 결과가 나오는지 명확한 설명이 불가하다는 점은 매우 흥미롭다.

인공신경망에서 레이어가 이전 레이어의 출력에 weight를 곱하여 합(Summation)한 각 출력은 다음 레이어의 입력으로 사용된다. 이러한 레이어들이 수십 개 이상이 모이면 다층의 레이어가 구성되고 최종 레이어의 출력은 특정한 의미의 값을 가진다. CNN의 경우를 예로 들면 다음과 같다. 주어진 영상의 모든 픽셀값을 일렬로 배열된 행렬 내의 column으로 배치한다음 딥뉴



(그림 1) ETRI의 인공신경망 프로세서 아키텍처 상에서 분석한 인공지능 알고리즘의 성능

[출처] 한국전자통신연구원, 프로세서연구그룹작성

럴넷의 입력으로 사용한다. 딥뉴럴넷의 퍼셉트론 연산을 영상의 픽셀 데이터부터 시작하여 순서대로 계산을 마치면 최종 레이어의 출력값이 나오는데 최종 레이어의 각 퍼셉트론의 출력값은 부동소수점, 예를 들면 0.991, 0.872, …와 같은 값이 출력된다. 뉴럴넷의 최종단의 퍼셉트론이 나타내는 출력값은 현재의 주어진 영상이 그 퍼셉트론에 할당된 물체일 확률(Probability)을 의미한다. 예를 들면, 영상이 자동차일 확률이 99.1%, 자전거일 확률이 87.2%와 같은 식이다.

인공신경망 그리고 딥러닝에 기반한 인공지능이 그 알고리즘의 효율성으로 인하여 최근 각광을 받고 있다면, 인공지능 반도체에서의 도전 포인트는 결국 그 복잡도(Complexity)이다. 인공신경망 알고리즘의 복잡도를 분석한 예를 (그림 1)에서 확인할 수 있다. 병렬 컴퓨팅(Parallel computing) 반도체의 성능과 성능에 영향을 미치는 요인을 분석하기 위하여 Roofline model이 자주 사용된다. 가로 길이가 440pixel인 인공지능 영상인식 알고리즘이 요구하는(Required) 성능과 한국전자통신연구원에서 2017년 개발한 초기 단계의 인공지능 프로세서인 AB7에서 제공하는 아키텍처 성능(Ideally-achievable)이 표시되어 있고, 반도체 칩 내에 구현한 메모리 크기와 Super-threaded array의 Dimension에 따라서 확보 가능한 성능이 표시되어 있다. 이 인공신경망 프로세서는 인공신경망에서 필요로 하는 데이터의 Flow, Amount를 최적화하여 Computing efficiency를 극대화하기 위한 새로운 구조를 채택하였는데, 이때 영상인식 알고리즘이 필요로 하는 성능은 1초에 요구하는 성능이 300GFLOPS 정도이다. 이것은 현재의 가장 빠른 스마트폰이 제공하는 성능 대비 30배에 해당한다. 즉, 30개의 스마트폰 어플리케이션 프로세서를 한 개의 칩에 넣어야만 확보 가능한 성능이라는 뜻인데, 현재의 스마트폰 어플리케이션 프로세서 중 1개의 양산품을 개발하기 위해서 이미 국내 대기업에서는 수백 명의 인력을 1.5년가량 투입하여 개발하고 있음을 고려해 보면, 앞으로 인공지능 컴퓨팅 반도체가 요구할 Resource와 그 시장성장의 폭에 대하여 예상해 볼 수 있다.

고도의 Parallel computing 성능을 요구하는 인공지능 반도체 분야는 최근의 시장 요구 및 인공지능 시장의 성장에 발맞추어 지속적으로 차세대 반도체의 혁신을 이끌 것으로 보인다. 인공지능 반도체의 관심과 더불어 컴퓨팅 성능을 비약적으로 발전시키려는 연구분야 이외에도, 인간의 두뇌 속에 존재하는 뉴런(Neuron)과 시냅스(Synapse)와 흡사한 기능을 반도체 소자(Device)에 구현하려는 연구분야도 있다. 예를 들면, 전하(Charge)와 자속 플럭스(Magnetic flux) 사이의 비선형적 관계를 특성으로 가지는 멤리스터(Memristor)를 이용하여 뉴런의 기억 특성을 구현하고자 하는 것이다. 멤리스터는 그 명확한 존재 유무가 1971년의 개념 정립 이후 아직도 불명확할 정도이지만 RRAM(Resistive RAM)의 개념과 연결되었다. RRAM은 단지 가소성(Plasticity) 저항(Resistance)을 가지는 소자일 뿐이므로 멤리스터와 직접적으로 연관되지는 않는다. RRAM은 그 특성상 가소성과 저항값에 의한 MAC(Multiply-and-ACcumulate) 연산이 가능하므로 인공지능 컴퓨팅을 위한 새로운 반도체 소자 기술로 관심을 받고 있다. 2017년 전후로 인공지능 반도체 관련 관심이 증폭되면서 알고리즘, 컴퓨팅 반도체, 반도체 소자에 이르기까지 인공

지능 분야로 진입하기 위한 다양한 노력이 전개되고 있고, 인공지능 반도체 소자 관련 연구도 다양하게 진행되고 있다. 인공지능 소자의 경우 뉴런과 시냅스를 그대로 모델링 하는 연구와 인공지능 컴퓨팅의 성능을 증대하기 위하여 면적과 전력을 최적화하기 위한 소자 관련 연구가 진행되고 있다. 다만, 전자의 경우 우리 뇌의 기억 구조나 행동에 대한 기제가 생물학적으로 완벽히 검증되지 않았으므로 좀 더 심화된 연구가 필요할 것이다. 인공지능 컴퓨팅의 새로운 방향성을 제시한다는 측면에서는 현재 알려진 인공신경망과 딥러닝 기반의 인공지능 컴퓨팅을 위한 대용량의 데이터를 칩내에 저장할 수 있는 새로운 반도체 소자의 개발이 필요하다.

Ⅲ. 인공지능 반도체 개발 동향: 세계와 국내

1. 세계 글로벌 기업의 인공지능 반도체 주도권 확보를 위한 노력

NVIDIA는 1990년대 인텔에 맞서서 CPU를 개발하기 위하여 설립된 기업이다. CPU 시장에서 x86을 내세운 인텔의 시장 지배자로서의 위치를 확인한 후 2000년대 초에 GPU에서 Geometry Processing과 Pixel Proess-ing을 통합한 최초의 GPU를 출시하면서 Voodoo, ATI 등 당시 Pixel Processing에 집중하고 있던 다른 기업들을 제치고 단번에 그래픽스 시장의 최강자로 자리 잡는다. 이후 그래픽스 카드 시장이 정체되면서 NVIDIA는 GPU를 Parallel Processing을 위한 칩으로 이용하는 GPGPU라는 개념을 내어놓는다. GPGPU가 2010년 전후로 상당한 인기를 끌기는 했지만, 그동안 구축해 온 게임을 위한 GPU 시장을 능가할 정도는 아니었다. GPGPU와 CUDA가 나온 이후에도 NVIDIA의 매출 상당량은 GPU이며, 최근 비트코인 붐에서 입은 혜택도 저가형 GPU들이었다.

NVIDIA는 GPU와 GPGPU를 위한 별도의 반도체 제품을 출시하고 있지만, GPU에 Texture mapping 등 그래픽스를 위한 특별한 설계가 일부 포함되어 있는 것을 제외하면 GPU와 GPGPU의 근본 구조는 SP(Stream Processor)를 기반으로 하는 SIMT(Single Instruction Multiple Thread)구조의 프로세서로 구성되어 있다는 점에서 그 구조가 동일하다. 2000년대 초반부터 GPU가 Unified Shader라고 부르는 Programmable processor 구조로 통일되면서 NVIDIA는 병렬컴퓨팅 아키텍처를 개발하는 회사로 급속히 성장하기 시작했고, 아키텍처의 구조를 변화, 향상 시키면서 Tesla(2007년), Fermi (2010년), Kepler(2012년), Maxwell(2014), Pascal (2016), Volta(2017)라는 코드명을 붙이면서 발전해 왔다. Tesla는 Unified Shader 구조로 만들어진 최초의 GPU 아키텍처라고 볼 수 있다. Tesla는 기본 연산 유닛인 SP와 SP를 연결한 SM(Stream Multiprocessor), 그리고 8개의 SM이 한 개의 반도체 칩을 구성한다. 각각의 Program flow를 가지는 Multiple Thread에서 각 Thread 별 Register를 저장하기 위한 대용량의 Register file set을 가지고 있으며, Instruction Cache, Data Cache, Multiple thread execution unit 등을 가진 대용량 데이터 중심의 프로세서 코어라고 볼 수 있다. SP를 8개 모아서 한 개의 SM이 되고 SM은 무려 768개 Thread의 동시 처리가 가능한데, Tesla 1개의 반도체 프로세서 칩에 8개 이상의 SM이 집적되므로 6,144개의 Thread, 즉, 6,000개 이상의 독립적인 Program flow가 동시에 실행되는 것이다. 인공지능 시장에 업계가 큰 관심을 보인 것이 Pascal 아키텍처가 등장하던 당시이다. Pascal에 와서는 한 개의 SM 내에 기존의 SP가 이름을 변경하여 발전한 CUDA Core를 64개 집적하고 있고, DP, 즉, Double precision을 위한 전용 코어를 16개, SFU(Special Function Unit)를 8개 포함한다. 병렬의 메모리 액세스를 위한 LD/ST(Load Store Unit)도 8개를 포함하고 있다. 반도체 제작 공정도 28nm에서 16nm로 발전하면서 SM을 무려 56개 탑재하여 3,584개의 Core가 한 개의 칩에 집적되어 있다. 동시에 칩의

면적은 무려 610mm^{2}에 달한다. 가로와 세로가 각각 25mm에 달하는데 일반적인 PC의 인텔 칩보다도 더 큰 면적을 가지는 칩이다. 가장 최근의 Volta 아키텍처에서 NVIDIA는 AI 전용의 Tensor core를 SM내에 집적하였다. 즉, Tensor core 128개, CUDA Core가 이름을 변경한 FP32 Core가 8개, INT(Integer unit, 정수연산 전용)가 16개, 64-bit Double precision 부동소수점을 위한 FP64 Core 8개의 Set을 총 4개 연결하여 한 개의 SM 내에 집적하고 있다. SM 구조를 여전히 유지하고 Warp라고 하는 Group of thread가 SM 내에서 한 번에 실행되는 구조를 여전히 유지하고 있기 때문에 Tesla 아키텍처에서부터 시작된 Multiprocessing의 개념은 그대로 유지하고 있다고 볼 수 있다.

NVIDIA의 Parallel computing 아키텍처는, 요컨대 NVIDIA의 SM과 CUDA Core로 이루어진 Parallel computing 구조는 기본적으로 Multi-thread Parallel processor의 구조를 갖추고 있고, 이것은 마치 x86-64 기반의 인텔 Xeon 프로세서가 수만 개 모여서 건물 크기로 만들어진 슈퍼컴퓨터를 손바닥 크기의 거대한 칩에 모아놓은 것과 같다. 다만, NVIDIA의 SP, SM은 x86과 같은 어떠한 소프트웨어도 실행 가능한 General-Purpose CPU가 아니고 그래픽스의 4-point floating-point computation에 최적화된 매우 작은 Core를 아주 많이 모아서 한 개의 칩에 구현한 그야말로 Parallel computing을 위해서 만들어진 반도체라고 볼 수 있다.

중국은 최근 미국과의 무역장벽 사태, 특히 ZTE 사태 등과 더불어 반도체, 그중에서도 어플리케이션 프로세서나 CPU가 기술의 핵심임을 간파하고 국가적인 관점에서 말 그대로 엄청난 투자를 하고 있다. 중국은 프로세서 기술 개발을 위해서 CAS(Chinese Academy of Science, 중국의 정부출연연구소)를 통하여 상당한 투자를 해 왔다. 2000년대 초에 MIPS를 기반으로 하는 Godson을 개발하고 연구소 기업 형태로 Loongson 이라는 회사를 만들었으나 큰 성공을 거두지는 못했다. 최근, Taihulight라는 Top 500 Supercomputer list에서 1위를 차지한 슈퍼컴을 만들면서 자체 개발한 프로세서인 SW26010를 격자구조의 대규모 Multiprocessor로 구성하여 핵심기술을 확보하였다. CAS의 ICT(Instutite of Computing Technology) 연구소를 중심으로 x86 Xeon CPU를 위한 인공지능 학습 및 인퍼런스 가속기인 DianNao, DaDianNao, ShiDianNao를 개발해 왔다.

중국의 Cambricon Technologies 스타트업은 CAS의 연구원들이 창업한 회사로 보이며 CAS의 ICT 내의 연구원과 Cambricon Technologies가 작성한 논문에 의하면 Cambricon-X는 6.38mm^{2}의 반도체 면적에서 544GOPS, 1초당 5,000억 개의 연산을 수행할 수 있는 구조를 갖추고 있다[5]. Cambricon의 경우 인공신경망의 필요 연산량을 감소시킬 수 있는 기술 중 하나인 Sparse matrix에 주목하고 있다. Sparse matrix는 인공신경망에서 학습 후에 나타난 Weight 값들 중에서 '0' 또는 '0'에 가까운 값을 가진 Weight는 실제로 곱셈연산을 할 필요가 없다는 사실을 응용하여 연산량을 줄이는 것이 기본 아이디어이다. Cambricon은 Sparse matrix를 효율적으로 이용하기 위한, 특히 연산이 필요치 않은 Weight의 Indexing을 위한 구조를 연구하여 개발한 가속기 구조이다. 그들이 개발한 인공지능 반도체는 IP형태로 제품화되어 중국 Huawei의 스마트폰인 Mate 계열의 제품을 구동하는 Kirin 970 프로세서 내에서 인공지능 컴퓨팅을 담당하는 NPU(Neural Processing Unit)로 사용된 것으로 알려져 있다.

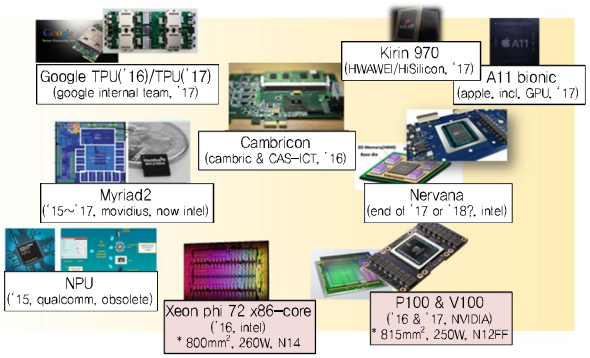
Google의 TPU는 알파고에 사용된 반도체라고 알려진 칩이다[6], [7]. TPU의 개발에 있어서 특이한 점은 Google은 반도체 기업이 아니라는 점이라는 것과 동시에 TPU 발표 당시에는 적어도 인공신경망에 있어서는 반도체 전문 기업보다도 더 우수한 성능을 보였으며, TPU2 칩을 개발하면서 부동소수점(Floating-point) 연산을 집중적으로 구현하여 학습과 인퍼런스 모두를 지원하는 반도체를 선보인 점이다. TPU(Tensor Pro-

cessing Unit)의 기본 개념은 슈퍼컴퓨터 또는 서버에서 흔히 채용하는 인텔의 Xeon 프로세서의 Floating-point extension이다. 즉, Xeon에서 부족한 부동소수점의 연산 성능을 높이기 위한 일종의 Co-processor라는 것이다. TPU2 칩은 TPU1의 구조에서 추측해보면 Xeon에서 데이터를 PCI express를 통하여 TPU2로 전송하고 이를 TPU2내의 메모리 버퍼에 저장한 후, 부동소수점으로 표현되어 있는 Feature data와 Weight data가 좌측과 상단에서 순차적으로 Floating-point systolic array로 입력되고 순대로 array를 지나가면서 연산이 이루어진다. 결과는 다시 TPU2 내부의 버퍼에 저장되고 Xeon은 PCI express를 통하여 연산된 결과를 CPU의 Main memory로 읽어오는 구조이다. 성능을 높이기 위하여 다수의 DMA를 TPU2 내에 구현했을 가능성이 있다. TPU2는 반도체 칩 당 300Watt 정도의 전력을 소모하는 것으로 알려져 있다.

TPU2로 구성된 시스템 보드는 4개의 칩이 한 개의 보드에 구성되어 있으며 BlueLink로 TPU2 칩간의 데이터 전송을 하고, OPA(Omni-Path Architecture) 인터페이스로 Xeon과 TPU2 사이의 데이터 공유를 하는 것으로 보인다. TPU2의 4개 칩 시스템 보드를 다수 적층하여 슈퍼컴퓨터를 구성하였다. 해당 컴퓨터는 현재 구글에서 인공지능 알고리즘 개발용으로 개발자들에게 공개하고 있다. Google의 TPU는 기본적으로 인터넷 기업인 Google의 인공지능 어플리케이션 및 알고리즘 개발을 위한 칩으로 보이며, Cambricon Technologies의 NPU나 NVIDIA의 Jetson TX보다 훨씬 큰 규모의 슈퍼컴퓨터를 위한 반도체이며, 특히 Xeon의 Floating-point 연산성능을 증대시키기 위한 Co-processor이다.

인텔은 인공지능을 위한 반도체 개발을 위하여 매우 다양한 시도를 하고 있는데 Movidius를 인수한 것이 첫 번째 시도라고 할 수 있다. Movidius는 스타트업으로서 2014년의 논문에 의하면 공유메모리형의 Application-Specific Parallel Processor로 볼 수 있다. Myriad 계열의 칩의 핵심구조는 SIPP(Streaming Image Processing Pipeline)라는 영상처리를 위한 Programmable accelerator array이고, 이들은 2MBytes에 해당하는 대용량의 온칩 메모리를 공유하고 있다. 일종의 영상처리 알고리즘에 최적화된 병렬처리 프로세서라고 할 수 있는데, 2014년에 발표된 만큼 인공지능 또는 딥러닝을 위한 인공신경망 처리에 최적화된 구조라기보다는 영상처리 및 인식 관련하여 알려져 있는 기존의 알고리즘들의 상당 부분을 구현할 수 있는 특수한 구조의 프로세서가 다수 배치되고 이들이 메모리를 공유하고 있다. 인공신경망 구조가 아닌 정해진 영상처리에 집중한 설계이다보니 오히려 면적이나 전력을 적게 소모하는 구조를 만들 수 있었고, Myriad 계열의 칩은 USB 스틱 형태의 제품으로 출시되어 인공지능 컴퓨트 스틱이라는 개념으로 상용화되었다.

인텔은 2018년 1월에 인공신경망 컴퓨팅을 위한 반도체 회사인 Nervana의 인수를 완료하였다. Nervana는 인공신경망 컴퓨팅을 위하여 Flexible precision을 가지는 Floating-point 연산에 관련된 특허를 보유하고 있다. 인공신경망 컴퓨팅이 앞서 언급한 바와 같이 현재 반도체가 제공할 수 있는 연산량보다 300~1,000배의 연산 성능을 요구하기 때문에 반도체 설계자 입장에서는 Sparse matrix 또는 32-bit integer, 16-bit integer 등을 사용하려 연산량을 줄이는 연구를 하였다. Flexible precision Floating-point는 정확도를 높이기 위해서 Floating-point를 사용하면서도 연산량 또는 반도체 구조의 복잡도를 줄이기 위한 시도라고 할 수 있다. Nervana는 2년간 'Lake Crest'라는 코드명으로 제품을 개발해 오다가 최근 제품 출시를 앞두고 있다. 고속 대용량의 데이터를 위하여 HBM3를 Package 내에 2~4개 장착하고 1.2Tbps의 반도체 칩과 칩 사이의 인터페이스를 위하여 SerDes를 갖추고 있는 것으로 알려져 있다. 현재까지 완성된 제품이 공개된 적이 없으므로 향후 추이를 지켜볼 필요가 있다.



(그림 2) 인공지능 컴퓨팅을 위하여 2017년 1년간 출시된 반도체와 성능

Xeon Phi 등과 같이 68개의 x86 CPU를 한 개의 반도체 칩에 집적한 제품을 인텔이 생산하고 있지만, 소모 전력이 300Watt 이상이며 인텔의 가장 큰 지적재산인 x86 코어를 장착하고 있는 Multi-processor 제품들은 인공지능을 위한 것이라 하더라도 서버용 고성능 CPU로 정의할 수 있다. 즉, 소형의 이동성이 있는 인공지능 어플리케이션을 위한 제품이라고는 볼 수 없고 대형의 서버, 빅데이터 서버 또는 슈퍼컴퓨터 구성을 위한 고성능 CPU라고 볼 수 있는 것이다.

인공지능 반도체에 대한 관심이 증대되면서 2017년 한 해 동안 글로벌 기업들은 매우 다양한 제품들을 발표[(그림 2) 참조]하였고, 2018년 현재까지 미국, 중국, 우리나라 등에서 다양한 스타트업이 창립되어 가장 효율적인 인공지능 컴퓨팅을 위한 다양한 제품들을 개발 중이다. Graphcore, Cerebras Systems, Sambanova, Deephi, Cambricon Technologies, Horizon Robotics, UX factory(한국), Furiosa A.I.(한국) 등이 인공지능 반도체를 위하여 설립된 스타트업이다. 최근의 스타트업으로의 투자와 창립, 제품 출시까지의 경향을 보면 스타트업 창립 후 상당 기간 개발모드(Stealth mode)를 거치며 이 기간에는 제품 출시나 외부 홍보를 하지 않고 아이디어와 기술에 기반한 제품 개발에 집중하는 형태로 시장이 형성되고 있다.

Graphcore는 2016년에 창업한 회사로 2018년을 목표로 16nm의 Massively parallel mixed-precision floating point processor 개발을 진행 중인 것으로 알려져 있다. Horizon Robotics는 이미 국내에서 자율주행차 관련 연구자들에게 널리 알려져 있는데 Sunrise, Journey 시리즈의 반도체 칩을 발표하고 있으며 Journey의 경우 자동차 영상인식에서 수십 개의 주행 중의 물체를 동시에 인식할 수 있으며, 1.5Watt에서 1Tera OPS (Operations per Second)의 성능을 보이는 것으로 알려져 있다. 다만 Horizon Robotics의 Journey 칩은 중국의 도로 사정에 최적화하여 만들어져 있으므로 중국 시장을 목표로 먼저 시작하는 것으로 알려져 있다.

2. 자율주행차와 인공지능 반도체

4차 산업혁명이 우리 사회와 정보통신 산업을 혁신할 키워드로 받아들여지면서 가장 대표적으로 등장할 서비스로 자율주행차가 주목받고 있다. 자율주행차에 관련된 반도체 기술은 대단히 다양하고 광범위하며, 산업적으로 자율주행차에 사용하는 센서도 영상 CIS(CMOS Image Sensor), Radar, Lidar, 초음파, GPS 등 매우 다양하다. 그러나, 최근에 와서는 자율주행 자동차의 센서는 전기자동차로 세계적 반열에 오른 Tesla의 발표와 같이 영상 CIS와 Radar로 굳어지는 상황이다.

자율주행 자동차의 운행 기술을 보면 Global Routing과 Local Driving으로 구분된다. Global Routing은 사용자가 목적지를 알려주면 목적지까지의 경로분석, 경로설정, High Traffic, Lane Merge 등 목적지까지 가기 위한 전략(Tactics)을 정하는 것이다. Local Driving은 Global Routing에 의하여 결정된 경로를 고해상도 맵과 GPS를 이용하여 운행하는 도중 전후방의 차량, 보행자, 신호등, 공사구간, 사고회피 등 주행 당시의 상황에 대하여 동적으로 대응하기 위한 Driving 제어를 의미한다. 자율주행차에서 요구하는 Local Driving에서 영상 CIS로 입력되는 영상을 분석하기 위해서는 최종적으

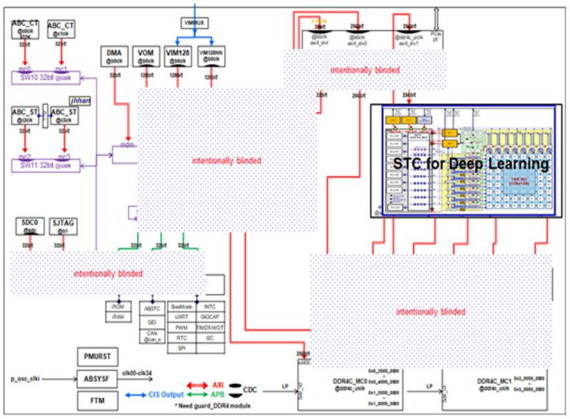
로 인공지능 반도체가 사용되어야만 하는 것으로 보인다. 자동차 메이커(OEM)의 경우 2016년 이전까지는 Mobileye(당시 이스라엘 기업)의 칩을 도입하여 컨트롤 SW를 개발하는 경우가 대부분이었으나, 2017년 이후 자사가 필요로 하는 반도체를 직접 개발하려는 노력을 시도하는 기업이 많아졌는데, 현재 인공지능 반도체가 자율주행차의 Local Driving 측면에서도 완벽한 기술을 가진 반도체 칩이 없기 때문이다.

Tesla의 Autopilot 개발 과정을 살펴보면 자율주행차에서의 인공지능 반도체의 중요성을 쉽게 파악해 볼 수 있다. Tesla의 경우 2016년경 Model S에 장착된 'Tesla autopilot* version 1'은 Mobileye 반도체에서는 영상인식을 한 결과를 출력하고, Tesla가 자체 개발한 SW를 Infineon 칩에 올려서 컨트롤하였다. 당시 Tesla의 Autopilot은 자동차의 주행 중 특정 구간에서 자율적으로 제어해주는 옵션이었고, 완전자율주행 옵션은 향후에 등장한다. 즉 기술이 완벽하지 못했던 것이다. 2016년 6월, Autopilot 모드이던 Model S가 옆차로를 달리던 트럭을 인지하지 못하여, 운전자가 사망하는 사고가 발생한 후, Tesla는 Mobileye와 법정다툼 중이며 Mobileye와 공식적으로 사업적인 결별을 선언하였다. Tesla는 NVIDIA의 GPU칩과 몇 가지의 프로세서 반도체를 이용하여 'Tesla autopilot* version 2'를 개발하였다. 이 시기에 Elon Musk CEO는 자체 자율주행 프로세서 반도체를 개발하기 위한 회사의 의지를 천명하였으며, 국내 대기업의 파운드리 사업부와 국내 연구소 및 실리콘밸리 반도체 기업에 자율주행차를 위한 영상인식 반도체 개발 프로젝트를 맡겨서 개발을 시작한 것으로 보인다. 현재까지의 결과를 보면 당시 개발 시도는 실제적으로는 실패로 끝나지 않았나 추측된다. Tesla는 AMD의 유명 x86 프로세서 개발자인 Jim Keller를 영입하여 자체개발을 시도하나 여의치 않았고, Jim Keller는 2018년 4월에 Intel로 자리를 옮기기에 이른다. AutoPilot Version 2는 8개의 카메라로부터 입력된 데이터를 기반으로 딥러닝 알고리즘과 컨트롤 알고리즘을 융합하여 차량을 제어한다. AutoPilot Version 2는 Tesla의 기존 사용자들로부터 혹평을 받았는데 기존의 Autopilot보다 성능이 더 떨어진다거나 차량 주행거리가 10% 정도 줄어든다는 등의 평을 받은 것이다. 2018년 1월 Tesla Model X가 주행 중 차선 변경 조건을 제대로 인지하지 못하고 임시 가설물에 충돌하여 운전자가 사망하는 사고가 발생하였다. 이후 CEO가 Stan-ford의 AI 개발자인 Andrew Karpathy를 AutoPilot 총책임자로 영입하여 딥러닝 알고리즘 개발을 가속하기 위한 노력을 진행중에 있다. Kaparthy는 Stanford의 박사과정이나 AI 관련 협력에 능한 사람으로 알려져 있는 반면, Tesla가 Mobileye를 인수한 Intel과 협력, 자율주행 프로세서 반도체를 개발하고 있다는 추정도 있다.

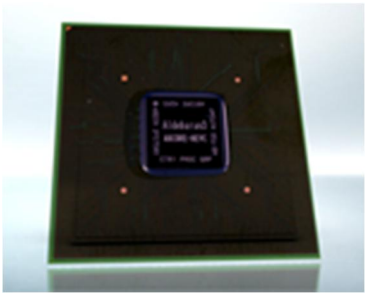
현재의 자율주행자동차의 기술 상황을 보면 완벽한 영상인식 인공지능 반도체 기술이 실재하지 않는 것으로 판단된다. 자율주행자동차, 자율비행 드론을 넘어서 인간과 동일한 또는 우수한 인지능력을 가진 로봇을 우리 현실에서 만나기 위해서는 인공지능 알고리즘의 지속적인 혁신이 필요하며, 동시에 매우 복잡한 구조와 연산능력을 요구할 인공지능 알고리즘의 등장에 대비하여 새로운 기술을 가진 인공지능 반도체 개발을 서둘러야 한다는 결론을 내릴 수 있다. 장기적으로는 모든 조건을 갖춘 인공지능 알고리즘과 인공지능 반도체를 개발해야 하겠지만, 단기적으로는 시장 진입 및 형성을 위해서 자율주행차의 도로상의 물체 또는 자율 비행 드론을 위한 비행 중에 만나게 되는 물체 인식 등을 위한 전용 인공지능 반도체가 필요할 것이다.

3. 우리나라의 인공지능 반도체

한국전자통신연구원에서는 40TFLOPS 급의 초고성능 인공지능 반도체 개발을 위하여 자체적인 기술개발을 2018년 현재 완료한 상태이며, 현재 성숙되지 않은 인공지능 반도체 시장의 상황을 고려하여 가장 먼저 진



(a)



(b)

(그림 3) (a) 40Tera FLOPS급의 인공지능 반도체 아키텍처와 (b) 자율주행자동차용 칩

[출처] 한국전자통신연구원 프로세서연구그룹 작성

입 가능한 시장인 자율주행 자동차를 목표로 자체기술 100%로 완성한 인공지능 반도체 아키텍처 개발을 완료하였다[(그림 3) 참조]. 현재 자율주행 자동차 및 중대형 서버용 가속기를 목표로 반도체 칩 개발 작업을 진행 중이다. 동시에 앞으로 등장할 인공지능, 즉, 강화학습, 비지도 학습 등을 위한 새로운 구조의 인공지능 프로세서 아키텍처 개발 및 대용량의 메모리와 거대한 반도체 칩 면적 및 전력 소모량 해결을 위한 연구를 병렬 진행하고 있다.

국내에도 인공지능 반도체 관련 스타트업이 4~5개 정도 있는 것으로 판단되고 있으나, 현재까지 복잡도, 시장상황 등 여러 가지 상황을 볼 때 본격적으로 양산을 시작한 곳은 없는 것으로 보인다. 일부 기업은 최근 투자를 받아서 개발을 진행 중인 단계로 판단된다. 한국전자통신연구원은 향후 4차 산업혁명과 사회혁신을 이끌 중요한 요소가 인공지능 기술이며 이를 뒷받침하기 위한 인공지능 반도체 기술을 국내에서 확보하는 것이 앞으로 수십 년간의 반도체 시장 및 정보통신 기술 주도권을 확보할 기회로 보고 있다. 이에 따라서 우리 기술로 100% 설계 기술을 확보한 인공지능 반도체를 개발 중이며 2019년 초에 자체 개발한 반도체 칩을 발표할 예정이다.

Ⅳ. 결론 및 시사점

병렬 컴퓨팅은 수십 년간 컴퓨팅과 컴퓨터 연구자들의 관심사였지만 인공지능의 등장과 전 산업으로의 영역 확장이 전개되면서 병렬컴퓨팅이 컴퓨팅의 패러다임을 다시 정의함과 동시에 정보통신 시장의 흐름을 변화시키고 있다. 인공지능 반도체는 인공지능에 최적화된 병렬 컴퓨팅 능력을 우리 눈앞에 실현해주는 유일한 솔루션이다. 서버 수준의 인공지능 음성 서비스에 머물러 있는 시대는 곧 저물어 갈 것이며, 자율주행 자동차, 자율비행 드론 그리고 인간처럼 생각하고 사고하는 휴머노이드에 이르는 새로운 인공지능 시대가 열릴 때 인공지능 반도체는 모든 정보통신 디바이스 내에 자리 잡게 될 것이다. 우리나라도 메모리 반도체로 세계를 제패한 경험을 되살려 미래혁신을 선도할 인공지능 반도체의 기술을 온전히 우리기술로 자립할 수 있는 기반을 마련하기 위한 노력을 전개해야 할 것이다.

참고문헌

[1] J. von Neumann, “First Draft of a Report on EDVAC,” Contract No. W-670-ORD-4926 between US Army and Univ. of Pennsylvania, June 30, 1945.

[2] S. Gill, “Parammel Programming,” *Comput. J.*, vol. 1, no. 1, Apr. 1958, pp. 2-10.

[3] A. Turing, “Computing Machinery and Intelligence,” *Mind*, vol. 59, no. 236, Oct. 1950, pp. 433-460.

[4] D. Silver et al., “Matering the game of Go with deep neural

networks and tree search," *Nature*, 529, Jan. 2016, pp. 484-489.

[5] S. Zhang et al., "Cambricon-X: An Accelerator for Sparse Neural Networks," in *Annu. IEEE/ACM Int. Symp. Microarchitec.*, Taipei, Tawan, Oct. 15-19, 2016, pp. 1-12.

[6] N.P. Jouppi et al., "In-Datacenter Performance Analysis of a Tensor Processing Unit," *Proc. Auun. Int. Symp. Comput. Architec.*, Toronto, Canada, June 24-28, 2017, pp. 1-12.

[7] P. Teich, "Under the Hood of Google's TPU2 Machine Learning Clusters," THENEXTPLATFORM, May 22, 2017. Available: http://www.nextplatform.com