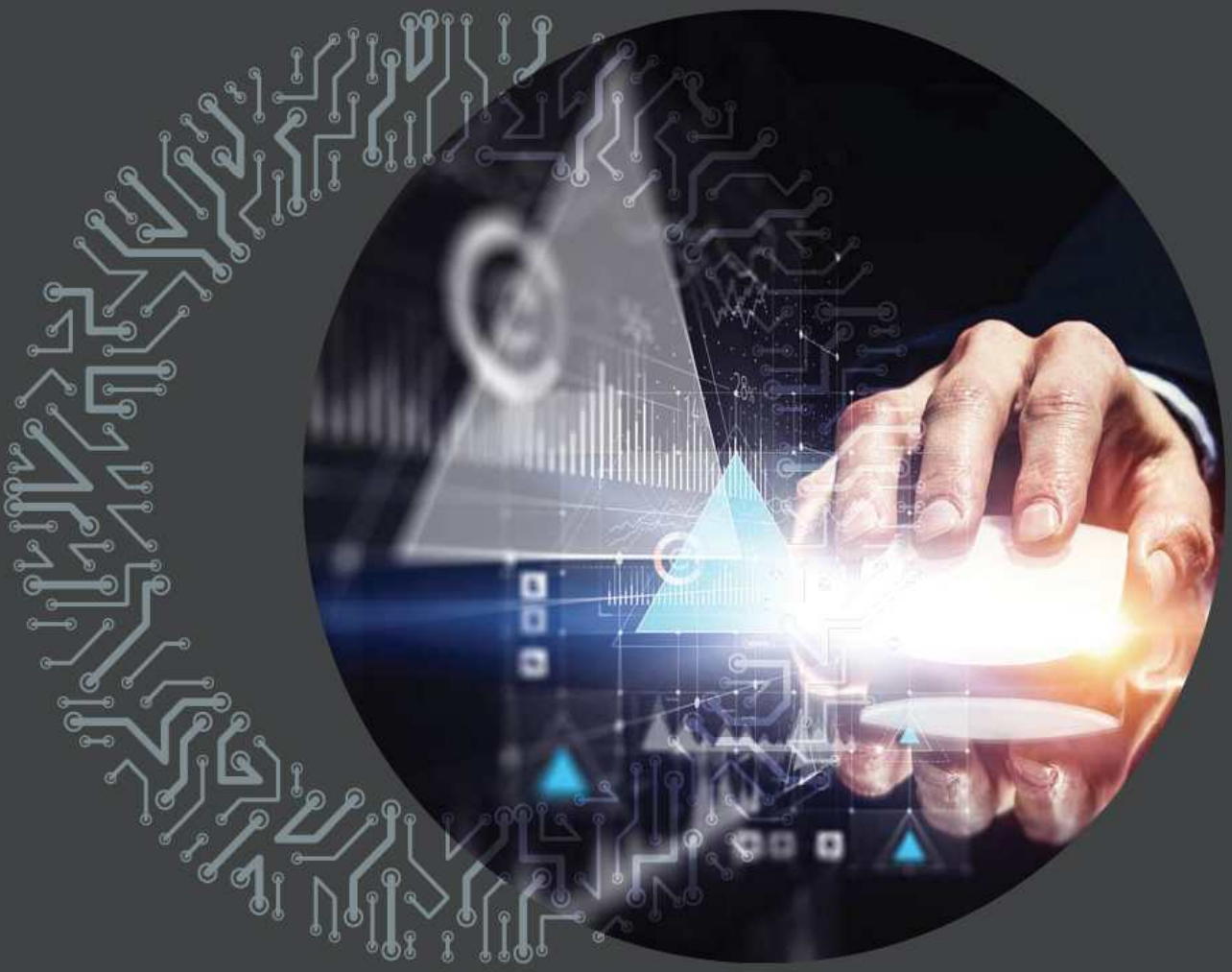
ETRI Insight

Insight Report 2017-37

기술경제연구본부
최새솔

Insight Report

인공지능 반도체 산업동향 및 이슈 분석



ETRI Electronics and Telecommunications Research Institute

※ 본 보고서의 내용은 필자의 개인적인 견해이며, 한국전자통신연구원의 공식 견해가 아님을 알려드립니다.

본 문서에서 음영처리된 부분은 ( ) 정보공개법 제9조의 비공개대상정보와 저작권법 및 그 밖의 다른 법령에서 보호하고 있는 제3자의 권리가 포함된 저작물로 공개대상에서 제외되었습니다.

CONTENTS

요 약 ····· 1

Ⅰ. 인공지능 반도체 개요 ····· 3

Ⅱ. 인공지능 반도체 산업 동향 ····· 12

Ⅲ. 인공지능 반도체 이슈 분석 ····· 19

Ⅳ. 맺음말 ····· 40

참고자료 ····· 42

약어표 ····· 47

참고문헌 ····· 48

요 약

연구배경 및 목적

- 인공지능에 대한 기대가 全사회영역에서 빠르게 고조되면서, 이를 효과적으로 구현하기 위한 인공지능 반도체에 대한 관심과 투자 증대
- 이러한 배경에서 본 연구는 인공지능 반도체 개념과 산업동향을 살피고, 시장, 생태계, 기술 관점에서의 현안을 분석함

인공지능 반도체 개념과 범위

- 인공지능 서비스 구현을 위한 대규모 데이터 처리를 효과적으로 처리할 수 있는 전용연산 컴퓨터 하드웨어구조 필요
- **(기술개념)** 인공지능의 학습과 추론 알고리즘 연산을 최적 구현하기 위한 특화 프로세서
- **(기술범위)** 협의로는 기존 반도체 아키텍처 기반의 인공지능 연산전용 가속 프로세서를 의미, 광의로는 두뇌신경을 모사한 뉴로모픽 칩까지 포함
  - 협의의 인공지능 반도체는 FPGA, ASIC, GPU를 기반한 별도 칩 혹은 SoC 탑재 유닛 형태로 CPU와 연동하며, 현재 개발·판매되는 대다수 칩이 이에 해당

인공지능 반도체 산업동향

- 인공지능 반도체 시장은 초기이나 기술의 진전과 높은 시장수요로 반도체 산업의 새로운 사업기회를 제시할 것으로 전망
- 세계 인공지능 반도체 시장은 2016년 6억 달러 규모에서 2021년 52.4억 달러 규모로 연평균(CAGR) 54.3%의 고성장 전망
- 인공지능 반도체 시장은 전통적인 프로세서 반도체 업체(예: 퀄컴, 엔비디아, 인텔, ARM) 뿐 아니라, ICT업체(예: 구글, 애플, MS, IBM, 화웨이) 등 유수 업체들의 치열한 각축전이 전개 중
  - 모바일向 인공지능 반도체가 2017년 대거 출시, 향후 치열한 제품 경쟁 전망

요 약

인공지능 반도체 이슈분석

이슈 ① 시장분화 : 모바일 엣지 AI의 부상

- 지능화 서비스와 사물인터넷이 확산될수록 단말 자체에서 인공지능 추론연산을 처리하는 모바일 엣지 AI의 중요성이 커질 전망
- 인공지능 반도체 시장은 클라우드向 시장(기존)과 모바일 엣지 向 시장(신규)으로 양분될 것으로 전망되나, 이들은 배타적 관계가 아닌 상호보완적 관계로 진화

이슈 ② 시장경쟁 심화 : 희미해지는 업체 간 영역경계

- 인공지능 반도체의 신규 시장 창출력과 시장 질서를 재편할 파급력이 확인되면서 ICT기업과 반도체 기업의 사업영역이 중첩되기 시작
- 비교적 영역경계 명확했던 모바일과 PC·서버 프로세서 시장의 경계가 옅어지며 산업내부 기업 간 경쟁 역동성이 증대

이슈 ③ 개발자 생태계 강화 : 오픈소스 및 개발자친화 문화

- 개발자 생태계 구축은 인공지능 반도체 시장 주도권을 확보하는 핵심전략이며, 오픈소스와 개발자친화 문화는 인공지능 반도체 벤더들이 반드시 고려해야 하는 생태계 전략도구

이슈 ④ 반도체 산업구조 변화: 가치사슬 변화, 롱테일시장

- 다변화하는 시장 수요 속에서 다품종 소량생산 체계가 정립되고, 이를 위한 디자인 하우스와 스타트업·중소기업 중심의 팹리스가 활성화될 전망
- 개별 산업 내 디지털 전환(DX)이 심화될수록, 범용 프로세스보다 도메인 요구사항에 최적화된 특화 칩 수요가 증가하며 킬러앱이 없는 롱테일 시장으로 산업구조가 개편

이슈 ⑤ 기술발전 가속화: 알고리즘, 반도체 공정, 차세대 인공지능칩

- 초저전력, (특화된)고성능 인공지능 프로세서를 위해, 소프트웨어적인 알고리즘 최적화와 하드웨어 측면의 반도체 공정 고도화가 병행되며, 궁극적으로는 학습과 추론이 온칩(on-chip)으로 모두 가능한 뉴로모픽 아키텍쳐로 발전 전망

I 인공지능 반도체 개요

(1) 발전배경

- 최근 인공지능(딥러닝)은 거의 모든 산업분야에서 활용가치와 파급력을 재평가 받으며, 모든 기술이슈를 잡아먹는 소용돌이

  - 글로벌 ICT 기업들은 발 빠르게 변화에 대응하며 자사의 전략 방향 확립 중

※ MS, 누구나 AI 툴 사용해 개발하는 'AI 민주화' 목표 (EPNC, '16.11.04.)

※ 구글 AI 대원칙을 발표..."모든 제품에 적용한다, 누구나 쓰게 한다." (조선비즈, '17.11.28.)

  - 인공지능은 개인, 산업, 사회에 광범위한 파급 효과를 유발하며, 인간 삶의 방식을 근본적으로 변화시킬 것으로 전망

※ Bank of America는 로봇과 인공지능 확산으로 인해 많은 산업에서 30% 가량 생산성이 향상되고, 제조 노무비가 18~33% 절감될 것이라고 전망

※ '16년 세계경제포럼은 세계 주요 15개국에서 '20년까지 510만 개의 일자리 감소 전망

인공지능 세계시장 규모와 분야별 시장규모

자료 : 디지에코(2017)

- 인공지능(딥러닝)의 3대 핵심요소인 데이터, 알고리즘, 컴퓨팅 파워 중, 가장 큰 제약요소였던 컴퓨팅 파워가 빠른 기술적 진보 이루면서 최근의 인공지능(딥러닝) 부흥을 촉발

  - 인공지능은 그 개념이 발현된지 반세기만에 최근에 빅데이터, 딥러닝 알고리즘, GPU등 연산가속 HW 발달 등으로 이전과 다른 새로운 발전 국면을 맞이

※ 인공지능(Artificial Intelligence)이란 용어는 1956년 존 매카시(John McCarthy)가 '기계가 인간 행동의 지식과 같이 행동하게 만드는 것'으로 최초 정의

- **(알고리즘)** 최근 각광받고 있는 심층신경망(DNN: Deep Neural Network)은 50년이 넘는 오랜 기간 동안 진전되어온 인공신경망 알고리즘에 기초

※ 딥러닝의 시발점을 '06년 G. Hinton 교수논문[1]으로 삼기는 하나, 신경망의 기본개념인 퍼셉트론(Perceptron)은 '50년대 Frank Rosenblatt에 의해 제안되었고, 딥러닝 연산의 핵심인 다층신경망에서의 역전달(back propagation)알고리즘은 80년대 G. Hinton과 그 동료들에 의해 제안

- **(데이터)** '90년대 인터넷 시대가 열리면서 방대한 디지털 데이터의 생산 및 축적이 가능해졌고, 이는 '00년대 이후, ICT산업이 부흥으로 급격히 가속화

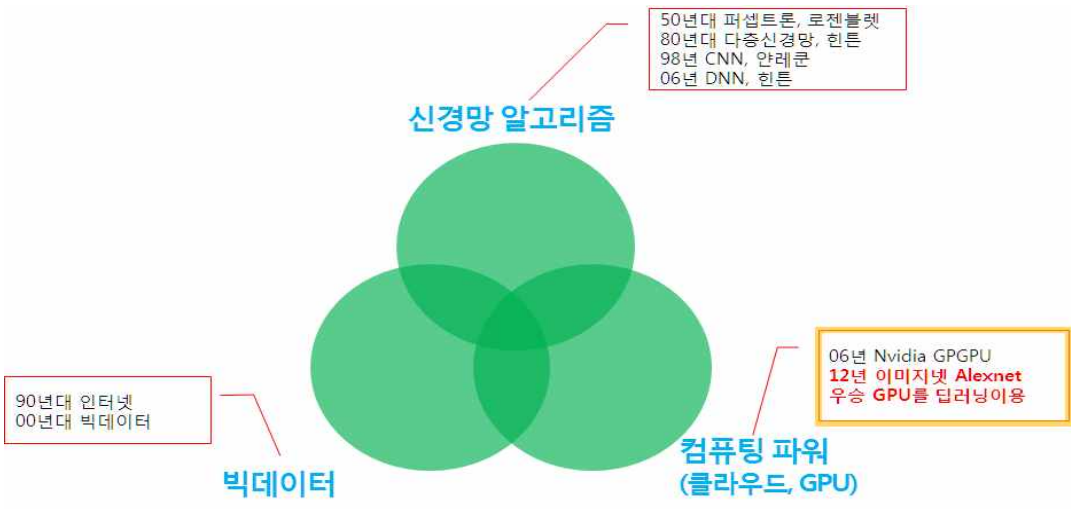
※ 세계 데이터규모는 '16년 16ZB에서 '25년 163ZB로 연평균 29%의 高성장 전망 (IDC,2017)

- **(컴퓨팅 파워)** '06년에 대규모 병렬 연산처리에 효과적인 방안이 제시되면서, 인공지능 발전의 가장 큰 병목이었던 컴퓨팅 파워 문제의 새로운 돌파구가 열림

※ '06년 그래픽연산 프로세서(GPU) 생산업체인 엔비디아는 GPU를 일반적인 병렬 연산 처리에 활용하는 GP-GPU 개념과 개발 툴(CUDA)를 발표하면서, 기존 CPU 중심의 연산처리의 한계를 극복하는 새로운 방향성을 제시

※ '12년 ILSVRC(Imagenet Large Scale Visual Recognition Challenge, 약칭 이미지넷)[2]에서 GPU를 활용한 딥러닝 시스템인 Alexnet이 현격한 성능우위로 우승을 하면서 가능성 확인

최근 딥러닝 발전의 원동력



자료: https://www.slideshare.net/awskorea/amazon-ai-deep-learning-on-aws 참고하여 수정

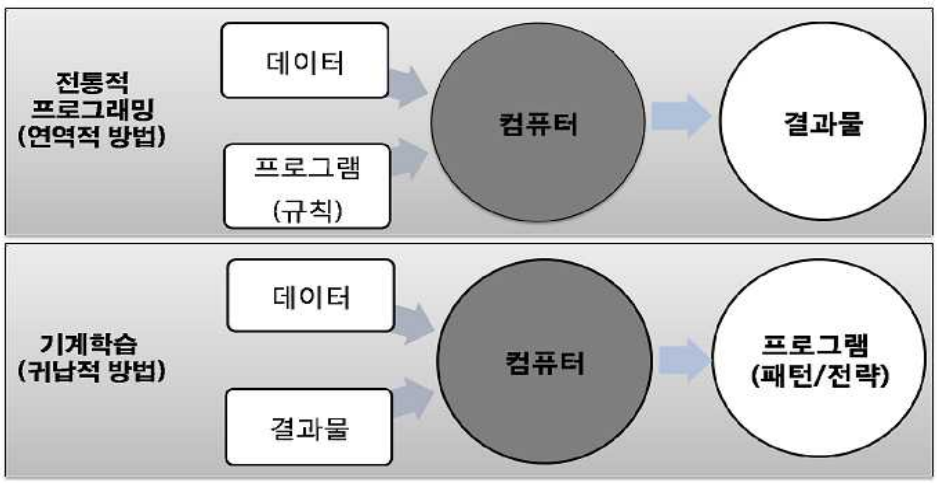
1) GE. Hinton, et.al. (2006), "A Fast Learning Algorithm for Deep Belief Nets", Neural Computation, 18:7, 1527-1554

2) 이미지넷은 1000개가 넘는 카테고리로 분류된 100만개의 이미지를 인식해 그 정확도를 겨루는 대표적인 시각지능 대회

인공지능과 컴퓨팅 패러다임의 변화

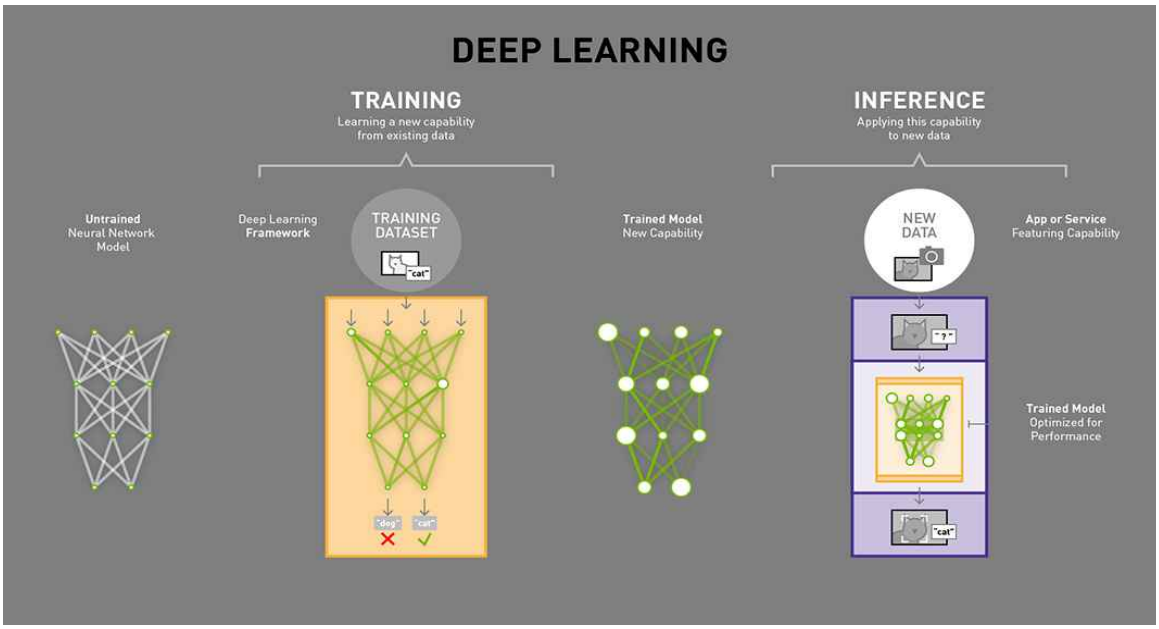
- 인공지능(DNN)의 발전은 기존 컴퓨팅 패러다임의 변화를 야기
  - 기존 컴퓨터가 사람이 정해놓은 규칙을 따라 데이터를 처리, 결과물을 도출했다면, 딥러닝은 입력 데이터와 결과 값이 주어지면, 스스로 규칙을 찾아 학습하는 방식
    ※ 일반적으로 학습을 통해 사람이 만든 모델보다 좋은 결과를 보여주며, 컴퓨터에게 일일이 규칙(프로그램)을 작성해 주지 않아도 되기 때문 효율적

인공지능 발전과 컴퓨팅 패러다임의 변화



자료 : STEPI(2016)

- **(딥러닝 연산의 2단계)** 딥러닝 연산은 학습(training)과 추론(inferencing)의 2단계 과정으로 구분
  - **학습(training)** : 딥러닝의 학습은 정의된 신경망에 대규모 데이터를 돌려가며, 신경망의 최적 가중치를 찾아가는 과정으로, 주로 데이터센터 서버에서 진행
    ※ 신경망 복잡도가 높을수록, 학습되는 데이터가 많을수록 분석 정확도 높아지나, 요구 컴퓨팅파워는 높아짐. '12년 이미지넷에서 우승한 AlextNet은 8개 층을 가진 신경망이었으나, '15년 우승한 ResNet은 152개 층을 갖는 심층망으로 구성
  - **추론(inferencing)** : 학습이 끝난 신경망에 신규 데이터를 적용하는 과정으로, 실행(Run, execution)으로도 표현. 신경망 최적화 수준에 따라 품질이 좌우.
    ※ 데이터 센터에서 진행이 되어 사용자 단말로 전달되는 것이 일반적 서비스 제공 방식이었으나, 향후에는 단말에서의 추론처리가 일반화 될 것으로 전망



자료: https://www.bigdataguys.com/deep-learning-training-nyc/ai_difference_between_deep_learning_training_inference/

(2) 기술개념

인공지능 처리를 위한 컴퓨팅 구조 진화

- 딥러닝은 단순한 연산의 수없는 반복으로 구성되어 계산량이 매우 큰 특징. 반면, 現컴퓨터 아키텍쳐는 이를 효과적으로 지원하는 대규모 병렬연산에 한계

  - 두 가지 측면에서 한계가 존재: ① 현재 컴퓨터 구조의 데이터 병목 문제, ② 무어의 법칙으로 대변되는 반도체 집적도의 물리적 한계

  - **폰 노이만(von-Neumann) 구조의 데이터 병목 문제**: 주기억 장치, 중앙 처리 장치, 입·출력 장치로 이어지는 직렬처리(Serial Processing)는 최근 요구되는 고속 병렬 연산에서 심각한 데이터 병목 현상이 발생

    ※ 하나의 CPU가 중앙에서 모든 데이터를 처리·제어하므로, 연산량이 많아질수록 메모리와 CPU사이의 병목현상 심각

  - **무어의 법칙(Moore's Law)[3]의 물리적 한계**: 반도체 집적도를 높임으로서, 데이터 처리 속도를 높일 수 있었으나, 점차 미세공정 고도화의 물리적 한계 봉착 中

    ※ 7nm 이하의 미세공정은 어려울 것으로 전망. 현재의 양산 최고수준은 10nm 급

    ※ 집적도를 높일수록 심화되는 발열과 간섭 등의 문제점도 해결과제임

3) 트랜지스터의 집적도가 18개월 내지 24개월마다 2배씩 늘어나는 것을 의미.

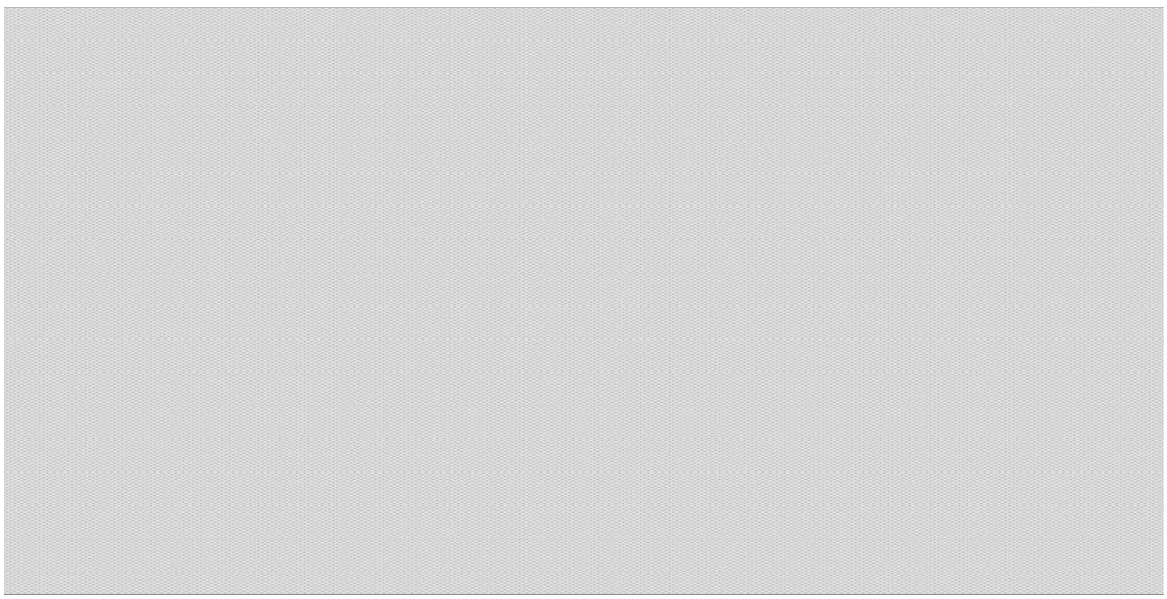
- 딥러닝 기반의 대규모의 데이터 고속처리를 위해서는 이를 효율적으로 감당하기 위한 새로운 컴퓨터 하드웨어 구조가 필요

  ※ “현재의 컴퓨터 아키텍처는 AI에 맞지 않다. CPU와 GPU와 다른 연산 구조가 필요” (손영권 삼성전자 사장, 한국경제, ’16.11.)

- 기존 CPU 중심에서 특정 데이터 연산(주로 딥러닝을 위한 행렬곱 등)에 특화된 가속프로세서를 추가하는 異기종 시스템 구조가 새로운 해결책으로 대두

  ※ HSA(Heterogeneous System Architecture): 異기종 시스템 아키텍쳐. CPU는 시스템 부팅 및 컨트롤 등을 담당하는 호스트 프로세서를 담당하고, 특정 연산에 특화된 GPU 등의 가속 프로세서(accelerator)를 CPU와 메모리를 공유하여 하나의 연산 장치처럼 활용하는 기술

이종 컴퓨팅 시스템 구조(HSA) 개요



자료 : Gartner(2017)

인공지능 가속 프로세서(accelerator)의 주요 기술방식 개요[4)]

- 인공지능 연산을 전문적으로 수행하는 별도 가속 프로세서를 둠으로써 전체 시스템 성능을 높이면서 획기적인 전력소모 감소가 가능

- 인공지능 연산을 위한 가속 프로세서의 주요 기술로는 GPU, FPGA, ASIC이

4) 주요기술의 특징 및 개발 사례는 2장에서 상세히 다룸

대표적이며, 각기 장·단점이 존재하여 경쟁하면서 공존할 것으로 전망

- GPU(Graphical Processing Unit) : GPU는 동시 계산 요구량이 많은 그래픽 영상 처리를 위해 고안된 프로세서로, 수천 개의 코어를 탑재해 대량 연산성능이 CPU 대비 매우 우수
  ※ 상대적으로 낮은 에너지 효율을 보이나, 가장 많은 구축참조 사례 존재

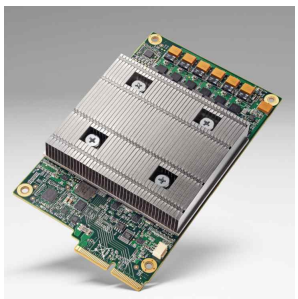
- FPGA(Field-Programmable Gate Arrays) : 회로 프로그래밍과 재구성을 통한 용도에 맞는 최적화와 변경이 가능한 높은 유연성이 특징
  ※ 범용 프로세서 대비 높은 프로그래밍 기술 수준이 필요한 단점 존재

- ASIC(Application Specific Integrated Circuits) : 특정한 용도에 맞도록 제작된 주문형 반도체인 ASIC은 가장 빠른 속도와 높은 에너지 효율이 특징
  ※ 설계비용이 비싸고, 특정 연산에 기능이 한정되어 범용성이 낮은 단점 존재

GPU, FPGA, ASIC 가속프로세서 사례

엔비디아의 최신 GPU 아키텍쳐 Volta를 적용한 Tesla V100

MS 클라우드 서버에 적용한 Altera의 FGPA



(ASIC) 구글 TPU

자료 : 각사 홈페이지

본 연구에서의 인공지능 반도체 개념과 범위

- 인공지능 서비스가 확산되면서 학습과 추론연산을 효율적으로 처리(실시간 처리, 저전력 등)하기 위한 특화 프로세서 유닛이 필요
  - 인공지능 반도체는 인공지능 서비스 상품성과 가격경쟁력 제고의 핵심요소로, 시장 확산의 획기적 기여 전망
  - 지능형 서비스 구현가능성과 개념정립을 넘어, 실제 서비스 출시와 시장수용을 위해서는 특정기능(사물인식, 음성인식 등) 최적화된 인공지능 하드웨어가 필수적

- **(기술개념)** 인공지능의 학습과 추론 알고리즘 연산을 최적 구현하기 위한 특화 프로세서

◐ **(기술범위)** 협의로는 기존 반도체 아키텍쳐 기반의 인공지능 연산전용 가속 프로세서를 의미, 광의로는 두뇌신경을 모사한 뉴로모픽 칩까지 포함

- 본 연구에서는 주로 협의로서 인공지능 반도체를 다루며, 광의의 뉴로모픽 칩은 기술 발전 전망을 다룬 3장에서 일부 논의

표 1 | 인공지능 반도체의 기술 범위

구분	세부 내용
협의	• 기존 반도체 아키텍쳐에 기반한 인공지능 연산 전용 가속 프로세서 - FPGA, ASIC, GPU기술을 기반한 별도칩 혹은 SoC 탑재 유닛형태으로 CPU와 연동하여 작동 - 현재 개발 시판되고 있는 대다수의 인공지능 프로세서가 해당
광의	• 기존 아키텍처를 벗어나 인간두뇌 모사한 신개념 소자 기반의 뉴로모픽 칩 포함 - 초저전력의 대규모 병렬연산 가능, on-chip에서학습/추론 효과적 처리 - 상용화까지 많은 기간/투자 소요 전망, 글로벌 기업 과감한 투자 중

자료 : 저자작성

(3) 시장전망

2~5년 안에 주류시장으로 빠르게 성장 전망

◐ 인공지능 반도체 시장은 초기이나 기술의 진전과 높은 시장수요로 반도체 산업의 새로운 사업기회를 제시할 것으로 전망

◐ Gartner(2017) hype cycle에 따르면, ASIC, GPU, FPGA기반 인공지능 반도체 2~5년 안에 plateau 단계[5] 로 빠르게 이동할 것으로 전망

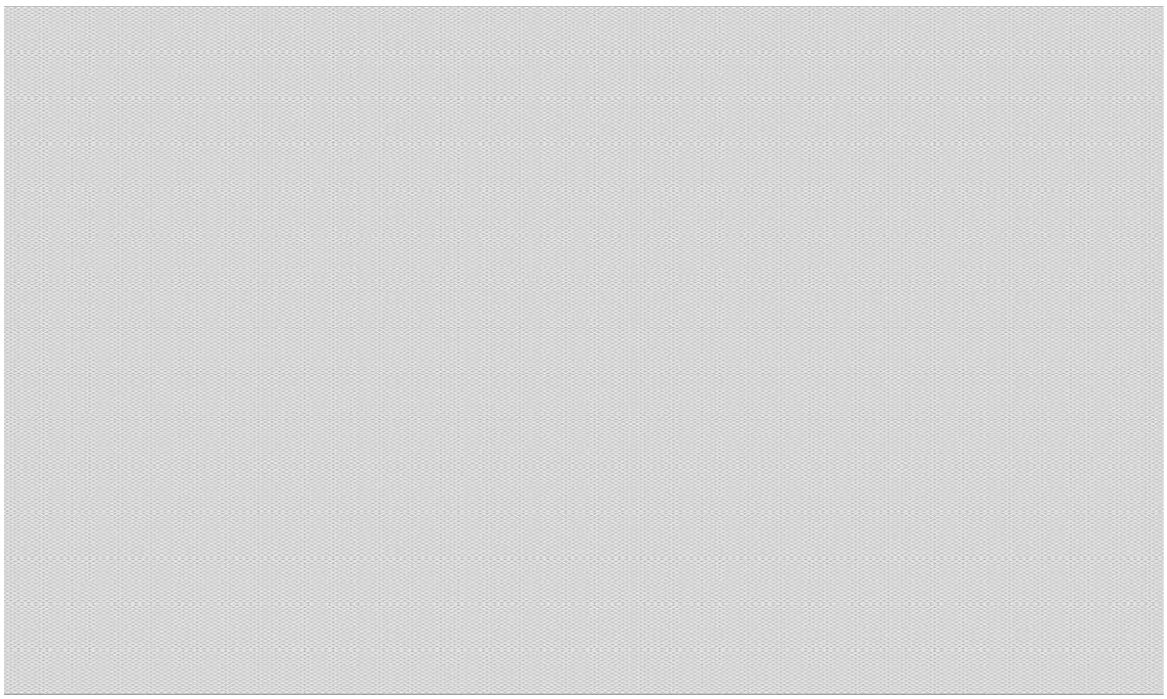
- DNN ASIC은 현재 시장 수용률1% 미만인 innovation trigger의 단계이나 향후 2~5년 사이 20%까지 시장 확산이 빠르게 진행 될 것으로 전망

- 현재 동일한 innovation trigger 단계인 뉴로모픽 하드웨어는 5~10년 이후에나 plateau 단계에 도달할 것으로 전망

- 전년 분석 대비 plateau 도달기간이 단축, 시장관심과 기술개발 가속이 전망

※ 뉴로모픽(10년 이상→ 5-10년), DNN AISC(5-10년 → 2-5년)

5) 이 단계는 technology innovation curve에서 early mainstream에 진입하는 단계로, 대략 20%의 침투율를 보이는 것으로 알려져 있음



자료 : Gartner(2017)

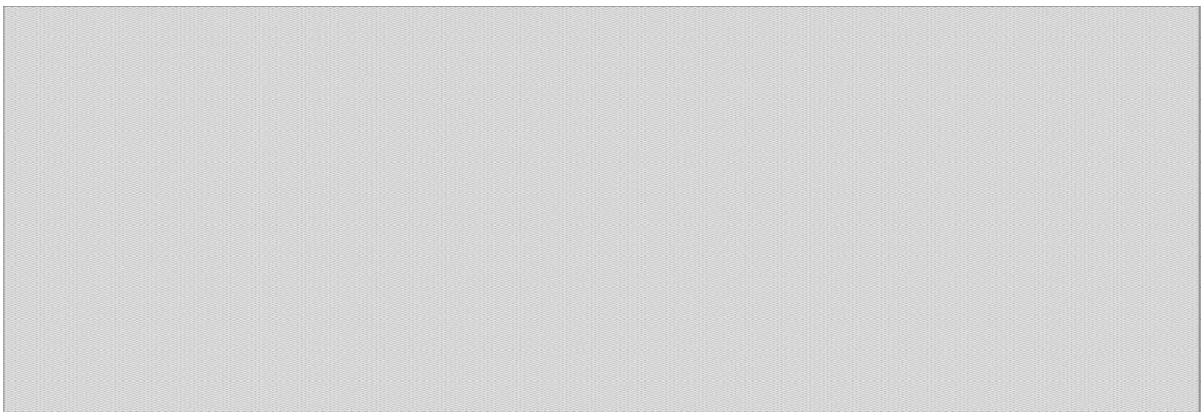
시장규모 : 연평균 54.3%의 고성장 전망

- 인공지능 반도체는 GPU, ASIC, FPGA의 품목형태로 출시되고 있으나, 개화기 시장으로 아직까지 별도의 인공지능 반도체 시장으로 구분하여 관련 시장 통계를 집계하고 있는 메이저 시장조사 기관은 없음[6)]
  - 본 연구는 시장조사기관인 Technavio(2017)의 발표 보고서를 활용
  - 이는 다른 시장조사기관 Tractica의 전망치와 큰 차이가 없음. Tractica는 '16년 시장규모를 5.13억 달러로 추계하여, Technavio의 6억 달러와 큰 차이가 없고, '20년 규모는 약 30억불로, 같은 기간 Technavio의 29.8억불과 대동소이
  - IDC는 ASIC과 ASSP기반의 AI칩 시장규모가 '21년 경 약 20억 달러에 이를 것으로 전망하였는데, 이는 같은 기간 Technavio의 ASIC기반 AI칩 시장전망치와 큰 차이가 없음
  - Technavio의 AI chip 시장 추정치 규모 측면에서도 '16년 전체 세계 반도체 시장에서 차지하는 비중이 0.2%~0.8% 수준으로 과대계상의 위험은 없다고 판단

6) 2017년 11월 현재, 시장조사기관 Technavio와 Tractica가 딥러닝용 인공지능 칩 시장을 별도획정하여 전망치를 발표. 본 연구는 가용한 자료인 Technavio(2017)를 기반으로 작성됨

- Technavio(2017)에 따르면, 세계 인공지능 반도체 시장은 '16년 6억 달러 규모에서 '21년 52.4억 달러 규모로 연평균(CAGR) 54.3%의 고성장 전망
- 구현 기술별 평균 비중('16~'21)을 살펴보면, GPU(35.7%), ASIC(32.3%), FPGA(17.5%), CPU(14.5%) 순으로, 향후 5년 내에 지배적 기술이 존재하지 않고 기술간 공존하며 경쟁할 것으로 전망

표 2 | **세계 AI칩 시장 전망('16-'21)** **단위: 백만$**



자료 : Technavio(2017)

프로세서 반도체 산업의 시장기회, 데이터 센터를 넘어...

- 오늘날, 현재의 CPU중심의 컴퓨터 아키텍쳐가 효율적인 인공지능 연산을 지원하지 못하고 있는 점은 반도체 벤더들의 새로운 기회요소로 작용
- 인공지능 반도체의 수요는 단순 데이터 센터向 딥러닝 프로세서 시장을 넘어, 지능정보사회의 다양한 연관 시장에서도 매우 크게 나타날 것으로 전망
  - IDC(2017)는 '21년 경 자율주행 ADAS(35억 달러), 공장 자동화(13억 달러), 스마트 빌딩(7.8억 달러), 스마트 교통(9억 달러), 스마트홈(8.6억 달러), 통신&네트워크(39억 달러) 등의 분야에서 지능서비스를 위한 프로세서 반도체 시장 수요가 증대될 것으로 전망

※ CPU, MCU, SoCs 등 프로세서 반도체 시장만 집계한 경우임

Ⅱ 인공지능 반도체 산업 동향

(1) 반도체 산업 일반특성

높은 진입장벽이 존재하는 산업

- 매년 대규모 시설투자(CAPEX)및 R&D 투자를 진행하는 하이테크 장치 산업

  - 2017년도 반도체 산업 CAPEX 규모(MPU/MCU 분야) : 116억 달러

2017년 전 세계 반도체 분야별 시설투자(CAPEX) 규모

2017 전 세계 반도체 분야별 시설투자(CAPEX) 규모

분야	올해 투자규모	비중(%)	전년대비 증가율(%)
마이크로컨트롤러/ 마이크로프로세서 유닛(MCU/MPU)	11.6	14%	16%
로직(Logic)	7.6	9%	11%
파운드리(Foundry)	22.8	28%	4%
D램·S램(DRAM·SRAM)	13	16%	53%
낸드플래시(Flash/NV)	19	24%	33%
아날로그/기타(Analog/Other)	6.9	9%	21%
전체	80.9	100%	20%

<자료:IC인사이츠, 단위:십억달러, %>

자료: http://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=101&oid=119&aid=0002208019

  - Top 10 반도체 기업의 2016년도 R&D 투자규모 : 350억 달러

Top 10 반도체 기업의 2016년도 R&D 투자규모

Top Semiconductor R&D Spenders
(Companies with ≥$1.0B in Spending)

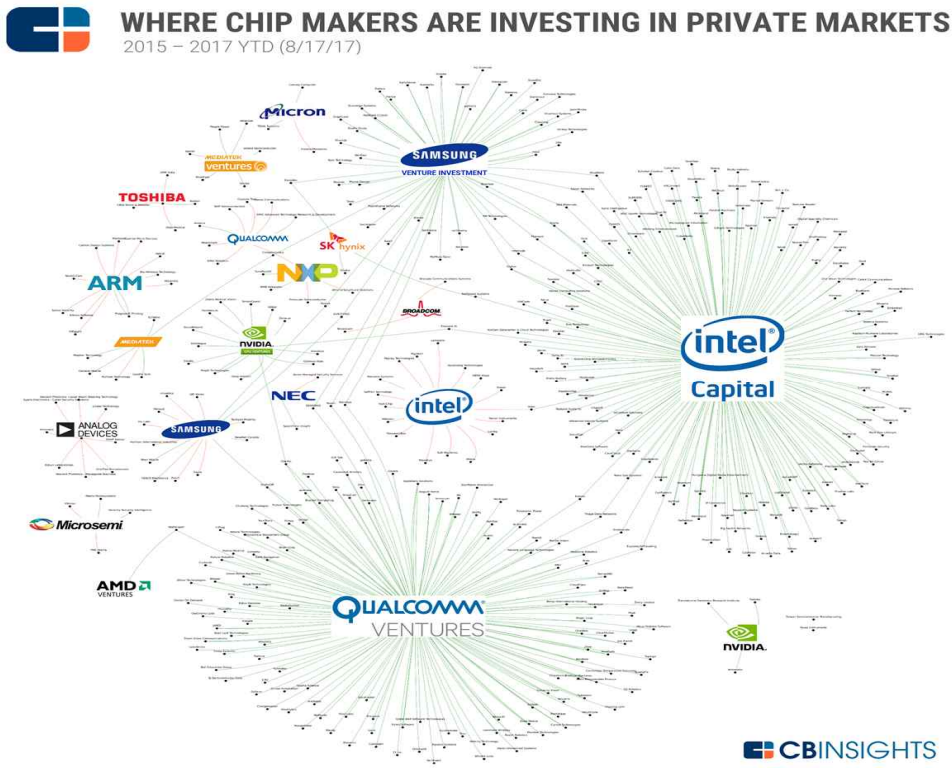
2016 Rank	Company	R&D Exp ($M)	R&D/Sales %	16/15 % Chg in R&D
1	Intel	12,740	22.4%	5%
2	Qualcomm	5,109	33.1%	-7%
3	Broadcomm Ltd.	3,188	20.5%	-4%
4	Samsung	2,881	6.5%	11%
5	Toshiba	2,777	27.6%	-5%
6	TSMC	2,215	7.5%	7%
7	MediaTek	1,730	20.2%	13%
8	Micron	1,681	11.1%	5%
9	NXP	1,560	16.4%	-6%
10	SK Hynix	1,514	10.2%	9%
	Top 10 Total	**35,395**		
11	Nvidia	1,463	22.0%	10%
12	TI	1,370	11.0%	7%
13	ST	1,336	19.3%	-6%

Source: Company reports, IC Insights' *Strategic Reviews* database

자료: https://m.blog.naver.com/PostView.nhn?blogId=narabaljeon&logNo=220940297622&isFromSearchAddView=true

- 전·후방산업의 업황에 민감한 중간재, 부품산업의 특성
  - PC/스마트폰의 경기나, 주요한 원자재인 희토류의 主산지인 중국의 규제 등에 민감
- 주요기업들은 독자 벤처캐피탈을 구축하여 내부 R&D 뿐 아니라, 외부 혁신도 흡수하는 전략 구사 중
  - '15-17년 반도체 스타트업 투자 건수[7] : 인텔, 퀄컴, 삼성 370여 건/IoT 79건/AI관련 50건

'15-17년 반도체 스타트업 투자 현황



자료 CBinsight(2017)

(2) 제품출시 동향[8]

- 인공지능 반도체 시장은 전통적인 프로세서 반도체 업체(예: 퀄컴, 엔비디아, 인텔, ARM) 뿐 아니라, ICT업체(예: 구글, 애플, MS, IBM, 화웨이) 등 유수업체들의 치열한 각축전이 전개 중

7) https://www.cbinsights.com/research/chip-semiconductor-startup-investments/
8) 기업별 상세 동향은 참고자료 확인

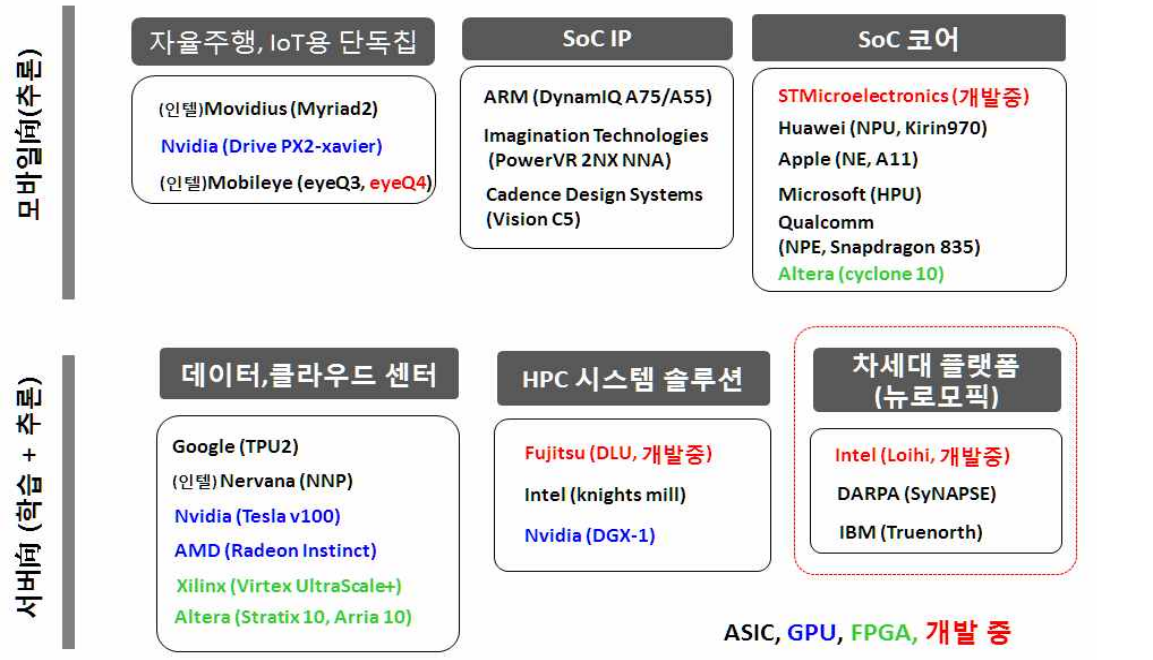
표 3 주요 기술 분야별 인공지능 반도체 출시 기업과 제품명

구분	기업 명	제품 명	출시일	비고
GPU	Nvidia	Tesla v100	2017.5	데이터 센터/ 학습 및 추론
		Jetson TX2	2017.3	임베디드 시스템/추론
	AMD	Radeon Instinct	2016.12	데이터 센터/ 학습 및 추론
FPGA	Xilinx	Virtex UltraScale+	2016.11	데이터 센터/ 학습 및 추론
	Altera(現 Intel)	Stratix 10	2017.11	데이터 센터/ 학습 및 추론
AP/DSP core unit	STMicroelectronics	미상	개발 중	초저전력 DCNN, 추론
	Cadence Design Systems	Vision C5	2017.5	모바일/추론, DSP IP core
	Huawei	Kirin 970 내 NPU	2017.9	모바일/ 추론
	Apple	A11 Bionic 내 Neural Engine	2017.9	모바일/ 추론
	Microsoft	HPU	2017.7	모바일/ 추론
	ARM	DynamIQ(A75, A55)	2017.5	모바일/ 추론, AP IP Core
	Qualcomm	Snapdragon 내 NPE(Nural Processing Engine)	2017.7	물리적 칩은 미 출시
	Imagination Technologies	PowerVR 2NX NNA	2017.9	모바일/ 추론, AP IP Core
ASIC (Stand alone)	Nervana(現 Intel)	NNP (Neural Network Processor)	2017.10	데이터 센터/ 학습 및 추론
	Google	TPU2	2017.5	데이터 센터/ 학습 및 추론
	Movidius(現 Intel)	Myriad2	2016.1	모바일/ 컴퓨터 비전 추론
신규 AI 특화 HPC 시스템 구조	Fujitsu	DLU(Deep Learning Unit)	개발중	딥러닝 HPC 시스템/학습 및 추론
	Nvidia	DGX-1	2016.4	GPU기반의 HPC 시스템/학습 및 추론
	Intel	knights mill	개발 중	인텔 제온 프로세서 기반 HPC
차량 특화	Mobileye (現 Intel)	EyeQ3	2014	자율주차량ADAS
		EyeQ4	개발중	
	Nvidia	Drive PX2 (xavier)	2017.1	AI Car 수퍼컴 시스템
뉴로모픽 (SNN[9])	Intel	Loihi	개발중	뉴로모픽칩, 학습 및 추론
	IBM	Truenorth	2014	뉴로모픽칩, 학습 및 추론

자료 : 각사 홈페이지 및 기사를 바탕으로 자체 작성

9) Spiking Neural Network

인공지능 반도체 시장 세분화



자료 :저자작성

- ‘17년 인공지능 반도체 全 분야에서 많은 신제품이 출시
  - 조사된 24개의 제품 중 개발 중인 제품을 제외한 기 출시된 제품 20개 중 14개가 ‘17년 출시
  - 모바일 디바이스의 AP/DSP SoC 유닛형태의 제품군은 모두 ‘17년에 출시되어 모바일 추론용 인공지능 반도체의 원년으로 삼을 만하며, 향후 치열한 경쟁이 전망
- 기업별로는 반도체산업 종합 1위 기업인 인텔과 인공지능 수혜를 가장 많이 받고 있는 엔비디아의 경쟁을 눈여겨볼 필요
  ※ 두 기업은 거의 전 시장영역에서 경쟁 중

표 4 인텔과 엔비디아의 인공지능 반도체 분야별 경쟁제품 현황

분야	인텔	엔비디아
모바일추론 분야	Myriad2, MyriadX, Cyclone	Jetson TX2, Tegra
데이터 센터용 학습 및 추론용 분야	NNP, Stratix, Arria	Tesla v100
HPC 시스템 분야	knights mill (개발중)	DGX-1
자율주행 자동차 분야	EyeQ3 (2014)	Drive PX2
뉴로모픽 분야	Loihi (개발중)	-

(3) 기술별 동향

전력효율 : 연산성능과 더불어 기술결정의 핵심 요인

- 데이터센터에서 가속화 시스템(ex. 인공지능 반도체)을 도입 할 때, 연산성능 뿐 아니라 전력효율 측면도 매우 중요한 고려요소로 작용

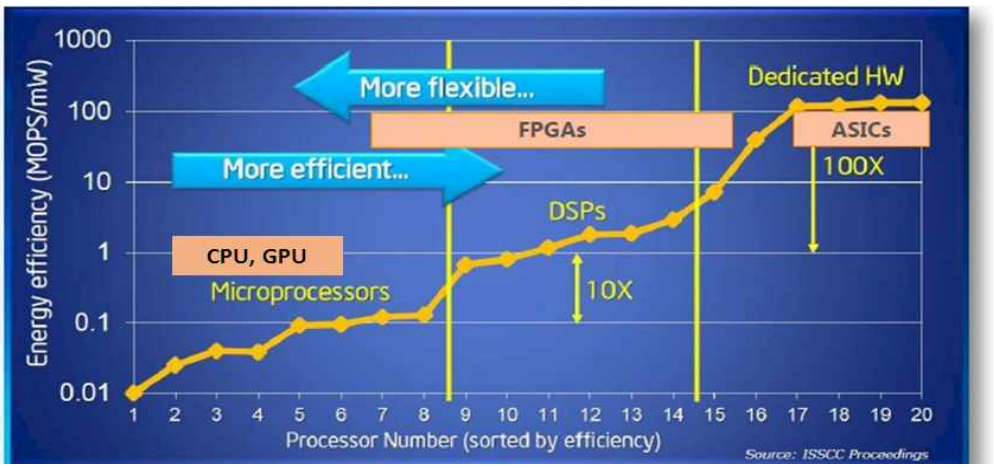
  - 24시간 운용되는 데이터센터는 많은 전력소모와 발열을 동반하기 때문에, 적절하게 냉각시키는 건 데이터센터의 안정성을 보장하는 중요한 요소

    ※ 마이크로소프트는 데이터센터의 냉각 성능을 높이기 위해 바다 속, 수중 데이터센터 구축을 위한 프로젝트(Natick)를 실험하고 있는 중

- 높은 병렬연산 성능과 많은 구축사례에도 GPU기반 시스템이 완전한 해결책이 되지 못하는 이유가 다른 대안 대비 전력효율이 떨어지기 때문

  - CPU와 GPU와 같은 범용 프로세서는 활용에 유연성이 있는 반면, 상대적으로 복잡한 명령어집합(ISA)을 가지고 있어 단순반복 연산이 많은 딥러닝 처리에는 비효율적인 측면 존재(단위 작업 당 전력소모가 높음)

범용프로세서(CPU, GPU), FPGA, ASIC기술 간 에너지 효율차이



자료: https://www.enterprisetech.com/2014/09/03/microsoft-using-fpgas-speed-bing-search/

- 이러한 맥락에서 전력소모를 과도하게 증가시키지 않으면서도 인공지능 알고리즘을 가속화할 수 있는 프로세서 반도체에 대한 필요성이 증대되는 상황

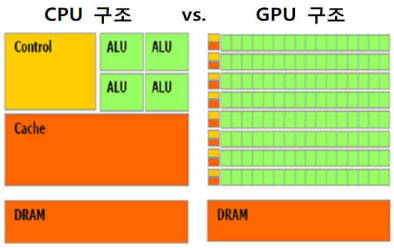
  - GPU기반 시스템에 비해 작동용이성이 떨어지는 FPGA나 제작비용이 높은 ASIC이 대안이 될 수 있는 것은 이들이 빠른 속도와 높은 에너지 효율의 특성을 갖기 때문

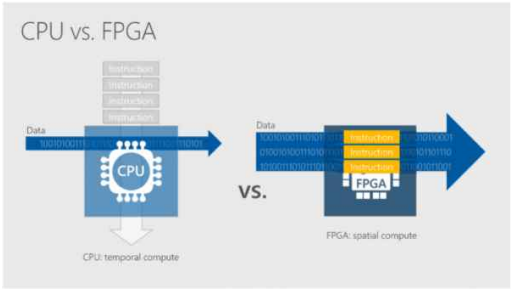
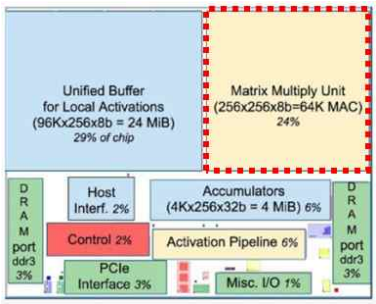
    ※ 마이크로소프트는 자사 데이터 센터 서버에 FPGA를 구축하는 프로젝트 캐터펄트를 통해, 일반 CPU 대비 비용절감 효과 30%, 전력효율성은 10% 향상

기술별 장·단점과 적용 사례

- GPU, FPGA, ASIC는 각각의 특징과 장·단점을 가지고 있으며, 기업들은 처한 상황과 필요에 따라 전략적으로 해당 기술을 활용하고 제품을 출시
- 개별 기술별 장·단점과 각각의 사례를 정리하면 다음과 같음

표 5 GPU, FPGA, ASIC 기술별 인공지능 반도체 특징과 적용 사례

유형	장·단점과 사례
GPU (GPGPU)	**장점** - GPU는 병렬처리에 최적화된 프로세서로, CPU에 비해 빠른 가속능력 - 엔비디아 CUDA 등 개발자환경이 잘 갖춰져 있고, 개별 도메인에 대한 적용 사례가 많아 지원받기 용이  출처 : http://khanrc.tistory.com/entry/GPU%EC%99%80-CPU **단점** - FPGA, ASIC 대비 낮은 전력 효율 - 기존 x86 시스템에 추가구축 시, 확장성과 호환성에 한계 (예) 데이터 전송 병목문제(PCI-e일 경우), 시스템 호환문제(NV link) 등 **주요 플레이어** - 엔비디아, AMD **응용사례** - **(페이스북, 바이두, IBM, 아마존, MS)** 클라우드나 데이터 센터에 엔비디아 테슬라 P100 적용 ※ 페이스북: 차세대 AI서버 '빅 바신(Big Basin), MS: 애저 클라우드, 아마존: AWS 등 - **(테슬라, 보쉬)** 테슬라는 자사의 모든 차에 엔비디아 제품 탑재계획, 보쉬(Bosch)는 엔비디아와의 양산용 자동차를 위한 인공지능 자율주행 시스템 개발 협력 발표
FPGA	**장점** - ASIC보다 초기 개발 비용이 저렴 - CPU와 병렬 작동이 용이하여 전체 시스템 병목현상 발생 없음 - 회로 재구성이 가능, 발전 중인 AI 알고리즘을 유연하게 적용 가능한 구조 (예) A라는 업무에 최적화해 사용하다가 칩 회로 구성을 다시 설정해 B라는 업무에 맞춰 사용 가능 **단점** - ASIC보다 느리고 CPU나 GPU 같은 범용 프로세서 대비 더 높은 프로그래밍 기술 수준을 요함

유형	장·단점과 사례
	- 8/16비트 수준의 낮은 정확성 알고리즘에 적합, 현재 주류 신경망 알고리즘 트렌드와 괴리  출처: https://thenewstack.io/developers-fpgas-cloud/ **주요플레이어** 알테라(Altera), 자일링스(Xilinix), 래티스(lattice) **응용사례** - **(마이크로소프트)** FPGA를 가장 활발히 데이터센터에 사용하고 있는 기업 ※ 수년간 FPGA 칩을 이용해 자사의 검색엔진 빙(bing)과 애저(Azure)의 성능을 향상 ※ 모든 분산형 데이터센터 서버에 FPGA를 적용하여 가속능력을 향상할 계획 - **(아마존)** 아마존 웹서비스에서 FPGA 기반의 가속 기능 및 개발도구 제공 - **(바이두)** 머신러닝 가속을 위해 자일링스 울트라스케일 FPGA사용을 발표 - **(퀄컴/IBM)** 데이터센터 가속을 위해 자일링스와 전략적협력계획 발표
ASIC	**장점** - GPU, FPGA 대비 매우 빠른 속도와 우수한 전력효율  출처 https://www.anandtech.com/show/11749/hot-chips-google-tpu-performance-analysis-live-blog-3pm-pt-10pm-utc **단점** - 매우 비싼초기 제작비용, 장시간의 개발소요시간 - 특정 연산에 최적화 되었기 때문에 응용분야가 한정 ※ (예) DNN전용 ASIC일 경우, 행렬 곱셈 계산에 최적화된 코어 설계 (구글 TPU는 24%가 행렬곱 연산) **주요플레이어 및 응용사례** - **(구글)** 구글의 ASIC칩인 TPU는 알파고 구현의 클라우드 컴퓨팅 플랫폼(GCP)의 관리 통제를 위해 활용 ※ 구글은 자사의 데이터 센터의 팬, 냉각 시스템, 창문 등 약 100여 개 장비와 시설의 통제와 관리를 TPU기반 인공지능을 통해 최적화 - **(모빌아이, 現인텔)** 모빌아이의 eyeQ 시리즈는 대표적인 ADAS용 ASIC 프로세서 - **(너바나, 現인텔)** 인텔이 인수한 Nervana의 NNP는 데이터 센터용 AI 전용 프로세서, Movidius의 Myriad2는 모바일 시각추론 전용 ASIC 칩 - **(애플, 화웨이, MS, Cadence 등)** 모바일 딥러닝 추론을 위한 ASIC기반의 AP/DSP SoC 내 프로세서 유닛이 최근 경쟁적으로 출시 중

III 인공지능 반도체 이슈분석

이슈분석 개요

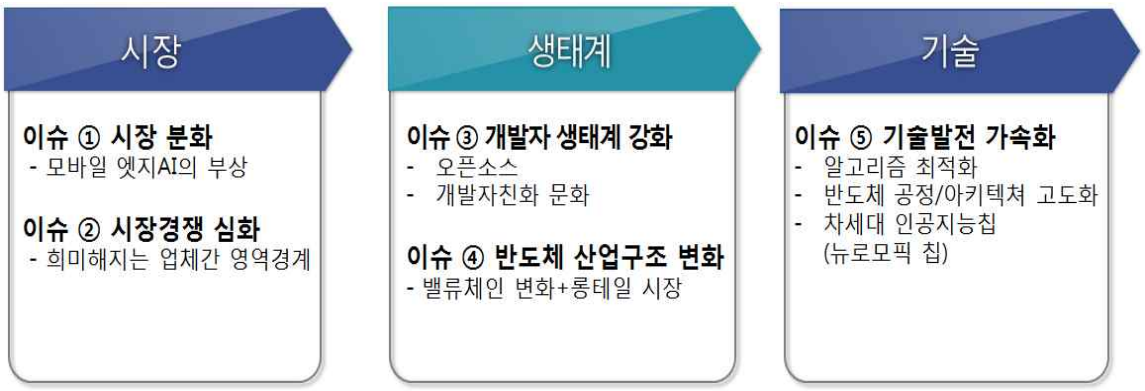
- 본 장에서는 인공지능 반도체 등장 이후, 시장, 생태계, 기술의 3가지 관점에서 반도체 산업의 쟁점사항을 분석

- **시장관점**
  - 이슈 ① 시장분화 : 모바일 엣지 AI의 부상
  - 이슈 ② 시장경쟁 심화 : 희미해지는 업체간 영역경계

- **생태계관점**
  - 이슈 ③ 개발자 생태계 강화 : 오픈소스 및 개발자친화 문화
  - 이슈 ④ 반도체 산업구조 변화: 밸류체인 변화, 롱테일 시장

- **기술관점**
  - 이슈 ⑤ 기술발전 가속화: 알고리즘, 반도체 공정/아키텍쳐 차세대 인공지능칩

이슈분석의 개요

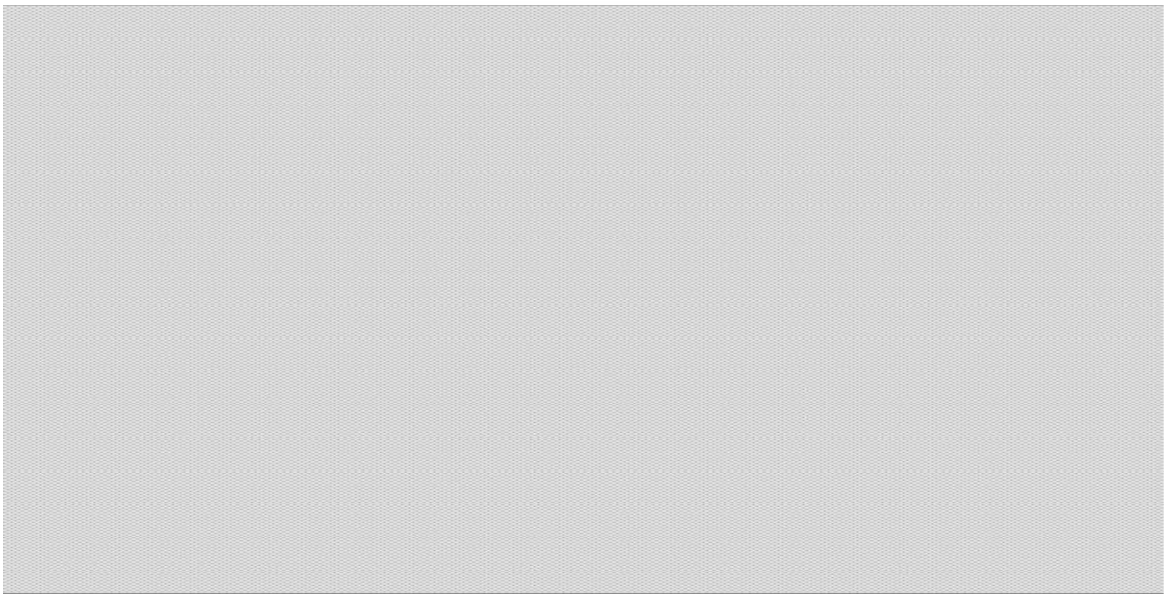


(1) 이슈 1: 시장분화

모바일 엣지 AI 등장배경과 특징

- 기존 인공지능 서비스는 클라우드 서버가 인공지능 학습(Training)과 추론(Inferencing)을 모두 담당하여 데이터를 분석한 후, 그 결과를 네트워크로 사용자에게 전송하는 방식이나,
- 이는 스마트폰이나 IoT 단말에 적용하기에는 서비스 실시간성이 떨어지고, 사용자 데이터수집에 대한 사생활침해 우려 등의 문제 존재
  - 데이터를 서버로 전송해 처리하는 방식은 연산마다 1~2초씩 지연시간이 발생하며, 네트워크가 연결되지 못하면 실행 자체가 불가능. 또한 사용자 민감 개인데이터 수집에 대한 우려 상존
- 이의 해결방식으로, 클라우드 서버에 의존하지 않고 기기 자체가 인공지능 연산(추론)을 처리하는 모바일 엣지 AI[10]가 대두

모바일 엣지 AI 개념



출처 : Gartner(2017)

- **(장·단점)** 모바일 엣지 AI는 서비스의 실시간성과 처리속도를 보장하고, 네트워크에 연결되는 못하는 상황에서도 서비스가 가능하고, 사용자 프라이버시를 보호

10) Mobile egde AI라고도 하고, On-device AI라고도 함

하며, 클라우드 서버 부하를 절감하는 할 수 있음

- 반면, 평소에 비해 배터리 소모량이 증가하게 되고, 모바일환경에 따른 컴퓨팅 파워의 제약은 인공지능 응용서비스의 수준을 제한

표 6 모바일 엣지 AI 반도체의 장·단점

구분		세부 내용
장점	저지연성	자율주행과 같은 실시간성이 요구되는 응용에 효과적
	사용성	네트워크에 연결되어 있지 않아도 서비스 구동 가능
	처리속도	AI서비스를 신속하게 제공 가능
	프라이버시	개인 민감 데이터를 클라우드로 전송 불필요
	비용	클라우드서버의 워크로드를 많이 경감하여 서버투자·운용 비용절감
단점	배터리 소모	인공지능 연산을 하게 되면 평소보다 많은 전력 소모 불가피
	제한된 응용	모바일 환경으로 인해 제한된 수준의 컴퓨팅만 가능

자료 : https://developer.android.com/ndk/guides/neuralnetworks/index.html 내용 정리

모바일 엣지 AI 반도체와 작동과정

- **(개념)** 모바일 엣지 AI를 구현하기 위한 전용 반도체로, CPU와 GPU의 부하를 줄이고, 모바일 서비스 대상 인공지능 연산(대부분 추론)을 담당하는 프로세서

  - 대부분의 딥러닝 추론은 PC 메니코어 CPU 수준으로 실행이 가능하나, 모바일의 경우, 발열, 배터리, 폼팩터 등의 제한된 환경으로 인해 추론연산에도 별도 가속장치 필요

  ※ 모바일 엣지용 인공지능 반도체는 다중 행렬 곱셈 계산과 같이 딥러닝을 효율적으로 처리하기 위한 연산에 특화

- **(형태)** ASIC칩 형태의 단독 칩(stand alone) 형태 혹은 SoC 내에서 CPU, GPU를 잇는 제 3의 코어 유닛 형태

- **(작동과정)** 최적화된(trained)알고리즘을 내부 메모리에 탑재(porting)하여 추론연산을 가속화

  1. 개발자 툴킷(toolkit)을 통해 신경망 훈련(Training)이 최적화된 알고리즘을 로딩
  2. 툴킷의 model optimizer가 모바일 엣지 AI반도체의 하드웨어 스펙에 알맞게 알고리즘을 변환

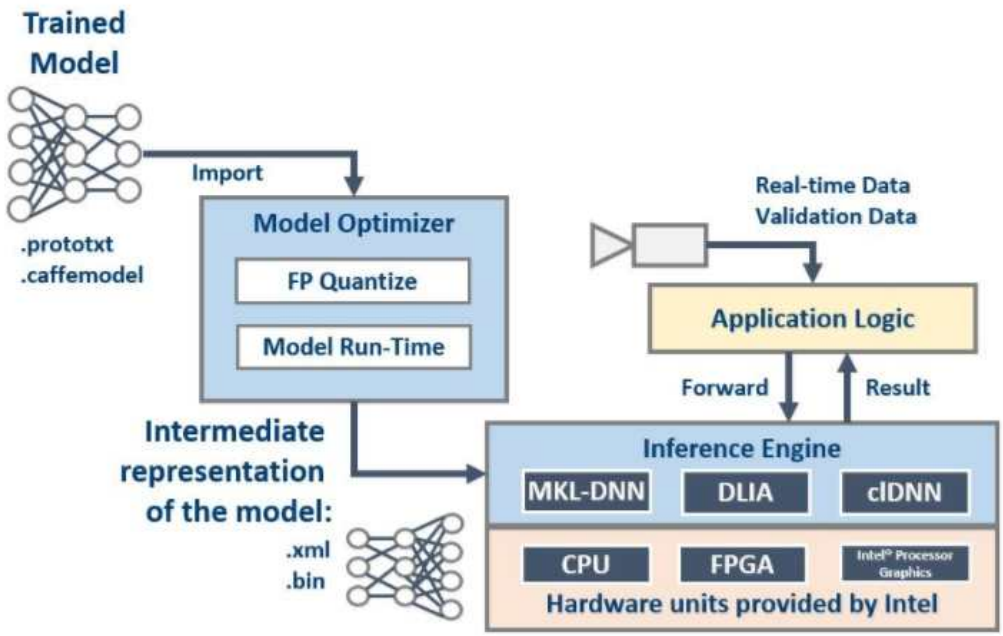
  ※ 디바이스 스펙에 따라 지원 불가한 경우도 존재

3. 변환된 알고리즘을 모바일 엣지 AI 반도체 내의 추론엔진에 로딩

4. 새로운 사물이나 상황 데이터를 실시간으로 입력받아 그것이 무엇을 의미하는지 추론(Inferencing)과정을 수행

모바일 엣지 AI 반도체 작동과정(알고리즘 입력에서부터 수행까지)

Figure 1: Model flow through the Deep Learning Deployment Toolkit



자료 : https://software.intel.com/en-us/articles/accelerating-deep-learning-inference-with-intel-processor-graphics

인공지능 반도체 시장분화와 지향가치

- **(시장분화)** 모바일 엣지 AI의 수요가 커짐에 따라, 인공지능 반도체 시장도 기존 클라우드 AI 중심 시장에서 클라우드 시장과 모바일 엣지 시장으로 양분되는 상황

  - (엣지AI 활용) 배터리 관리와 실시간성이 중요한 드론, 로봇, 자율주행차의 사물인식, 충돌회피, 스마트폰에서의 카메라 기능향상, 음성인식, 얼굴/지문 보안, AR/VR에서의 아이트레킹 등

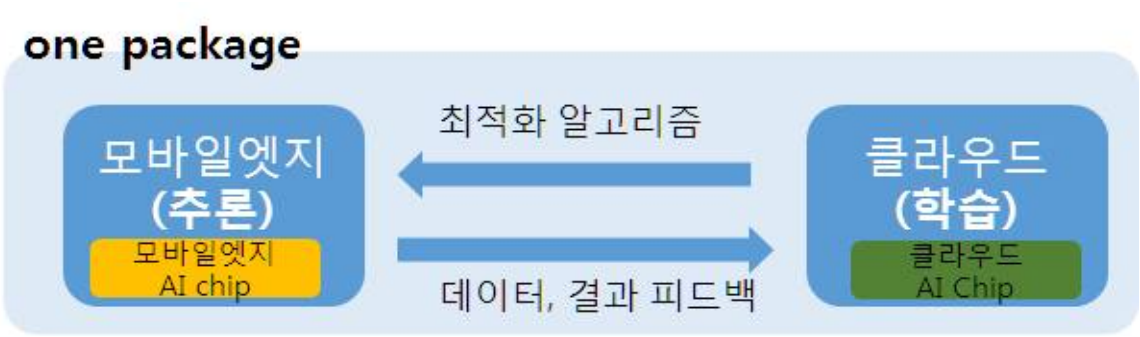
  - 글로벌 업체들은 이러한 모바일 엣지 AI솔루션에 주목하고, 이를 위한 별도의 맞춤형 프레임워크 구축 중

  ※ 구글 텐서플로우 라이트, 페이스북의 Caffe2go, 아마존의 AI 프레임워크, 퀄컴 뉴럴

프로세싱 엔진(NPE) 등

- 각 시장에서 인공지능 반도체가 지향하는 가치는 상이하나, 배타적 관계가 아닌 상호보완적 관계에서 하나의 서비스 package로 운용될 전망
  - **(클라우드向)** 딥러닝 학습을 위한 보다 강력한 병렬연산 능력과, 수요변화에 대응하기 위한 컴퓨팅파워 구성의 확장성과 유연성, 운용비용(OPEX) 관점에서의 전력효율 등
  - **(모바일 엣지向)** 개별서비스에 특화된 연산가속, 배터리, 폼팩터 등의 제한 환경 극복을 위한 초저전력, 경량화, 제조원가절감(low cost) 등

모바일 엣지 AI와 클라우드의 관계



자료 : 저자작성

(2) 이슈 2: 시장경쟁 심화

신규 진입자의 등장 : 글로벌 ICT 기업의 반도체 역량 강화

- 글로벌 ICT 기업들은 IoT, 인공지능, 빅 데이터, 자율주행 자동차, 로봇 등 차세대 유망 산업에서 반도체가 서비스와 시스템 성능에 결정적 영향을 미친다는 것을 이미 인식
- 이들은 자체 R&D와 함께 적극적인 유망 기업 인수를 통해 반도체 역량 확보에 매진

표 7 글로벌 ICT 기업의 반도체 역량 확보를 위한 M&A

업체	M&A 주요 내용
소프트뱅크	- '16년 7월, 모바일 반도체 IP 1위 기업인 'ARM'을 약 310억 달러에 인수
구글	- '알파고'를 위한 TPU(Tensorflow Processing Unit) 개발하기 위해 Gecko(2014.8), Agnilux(2010.4) 등을 인수활용
아마존	- '15년 1월, 이스라엘 반도체 회사인 안나푸르나랩스를3.75억 달러에 인수, ARM기반 칩인 알파인 칩을 출시하여 사물인터넷, 클라우드 시장 공략을 본격화
마이크로소프트	- '15년 2월, 이스라엘의 모바일 입력단말 업체인 n-trig를 2억달러에 인수 - '17년 8월, 미국의 Cloud HPC 업체인 cycle computing사를 인수
시스코	- '16년 3월에 네트워크 장비용 반도체설계회사인 이스라엘의 리아베(Leaba)를 3.2억달러에 인수 - '17년 10월, 미국의 머신러닝 업체 Perspica를 인수
애플	- AP를 자체 개발하기 위해 '00년대 중반부터 P.A.(2008.4), Anobit(2011.12), Passif(2013.8.) 등을 인수하고 스마트카, 착용형 스마트 디바이스 진출 - '16년 8월, 미국의 머신러닝업체Turi를2억달러에 인수 - '17년 5월, 미국의 AI업체 Lattice Data를 2억달러에 인수

자료 : 각사 홈페이지 및 기사를 바탕으로 자체 작성

- 이들은 더 나아가 자사의 인공지능 서비스 구현 성능 높이기 위해 자체 인공지능 반도체 자체 개발 中

  - 자사 서비스 성능개선이나 서버 운용효율을 위한 용도이나, 이러한 내부조달(in-house sourcing)은 인텔, 엔비디아 등 프로세서 벤더들의 사업에는 부정적 영향

표 8 글로벌 ICT 기업의 자체 인공지능 반도체 개발 사례

업체	개발 사례
마이크로소프트	- '17년 7월, 자사의 증강현실 고글인홀로렌즈2를 공개. 이 제품에는 MS가 개발한 언어와 영상 인식 기술을 구동하는딥 러닝 기술에 특화된 전용 칩(HPU)이 탑재
구글	- '17년 5월, 모바일 기기 환경에 최적화된 머신러닝 프레임워크 버전인 TensorFlow Lite와 45 테라플롭 급 성능의 'TPU2 칩' 공개
애플	- '17년 8월, 애플은 자사의 최신 AP인 'Apple A11 Bionic APL1W72'을 아이폰8에 탑재하며 최초 공개. A11은 뉴럴엔진이라 명명된 하드웨어기반 AI전용 프로세서를 탑재
화웨이	- '17년 9월 화웨이는AI전용 프로세서 유닛인NPU를 탑재한 최신 AP 기린970을발표 - NPU는 중국의 AI칩 스타트업인 캠브리콘의IP를 기술 이전한 결과
테슬라	- 자사의 차량에 들어갈 자율주행 전용 프로세서의 자체개발 계획을 발표

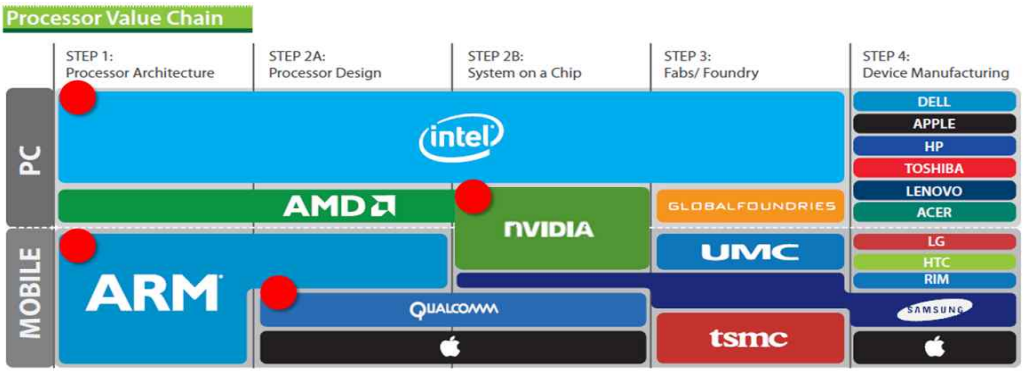
자료 : 각사 홈페이지 및 기사를 바탕으로 자체 작성

반도체 산업 내부 경쟁 양상 변화

- 인공지능이 활성되기 전에는 프로세서 반도체 벤더들간의 영역이 비교적 명확하였음 (아래 그림 참고)

  - PC·서버 분야에서 인텔이 압도적 지배력을, 모바일분야에서는 ARM의 아키텍쳐가 90%이상을 점유, 퀄컴은 모바일 AP와 통신반도체 시장을 석권, 엔비디아는 PC용 그래픽카드 분야에서는 1위 벤더였고, 이들 간의 영역경계는 명확

인공지능 활성화 前 프로세서 반도체 밸류체인과 해당기업들



자료 : http://iveybusinessreview.ca/cms/1070/intel-outside-breaking-into-mobile-3/ 재수정

- 그러나 최근 인공지능 반도체 시장이 급성장하면서 이들 기업 간의 새로운 질서와 대응 양상이 발생

  - **(엔비디아)** 반도체 매출규모로는 20위권('15년 기준 22위)도 안 되는 엔비디아는 단순 하드웨어 칩셋회사에서 가장 주목받는 인공지능 솔루션 업체로 변모 중

    ※ 모든 인공지능 반도체의 라인업을 갖추면서 경쟁력 강화

  - **(인텔)** 인공지능 반도체 분야의 투자가 뒤늦었던 부동의 반도체 1위 기업 인텔은 모바일, 차량반도체, 서버·HPC 영역까지 전방위적 M&A를 통해 인공지능 반도체 역량을 강화 중

  - **(ARM)** 모바일 엣지 AI 수요가 증대되면서 모바일 프로세서 core IP의 90%이상을 점유하고 있는 ARM 역시 AI에 특화된 다이나미크(DynamiQ) 기술을 적용한 첫 제품 발표

  - **(퀄컴)** 모바일 AP와 Connectivity 반도체 시장을 석권하고 있는 퀄컴 역시 세계 최대 자동차 반도체회사인 NXP를 인수하고, 모바일 AI 전용 프레임워크를 발표하는 등 AI 주도권을 놓지 않기 위해 노력 중

엔비디아의 부상

- 엔비디아의 지난 4분기(2016년 11월~2017년 1월) 매출은 전년 동기 대비 55% 늘어난 22억 달러이며, 주가는 1년간 4배가 증가

- 엔디비아는 모바일 엣지시장, 데이터 센터용, HPC 시스템, 자율주행차량 프로세서 등 모든 인공지능 반도체의 라인업을 갖추면서 경쟁력을 강화 하고 있음

- 이는 구글, 아마존, 마이크로소프트 등 인터넷 기업이 붐을 일으킨 인공지능 산업 내 공급망 내 부품 공급자 역할을 뛰어넘어 구글, 아마존 등과 직접 경쟁하는 AI 솔루션 사업자로 부상하고 있음을 의미

인텔의 전략과 행보

- 인텔은 인공지능시장이 개화하고 나서도 ASIC, GPU, FPGA 제작 보다는 자사의 CPU 를 더욱 강력하게 만드는 전략을 취해왔으나, 이는 결과적으로 환경변화에 적절히 대응하지 못한 결과 초래

- 인텔은 인수합병(M&A)을 통해 뒤늦은 만회를 시도 중. M&A를 통해 빠른 시간 내에 모든 인공지능 반도체의 라인업을 갖추면서 경쟁력을 높이는 중

표 9 | 인텔의 인공지능 반도체 관련 M&A 사례

피 인수기업	M&A 사례
알테라 (Altera)	- '15년 6월, FPGA 제조업체인 알테라를 169억달러에 인수 - 인텔은 추론용 딥러닝 프로세서는 인수한 알테라의 FPGA를 중심으로 제품 라인을 가져가고 있음. 알테라의 FGPA솔루션은 MS의 cloud Azure에 광범위하게 활용되고 있음
너바나 (Nervana)	- '16년 8월, 너바나 시스템즈를 4억달러에 인수 - '17년 10월, 인텔은 인수한 너버나 시스템즈의 제품에 기반하여, 자사 최초의 딥러닝용 ASIC 기반 상용칩셋인 NNP (Neural Network Processor)를 발표
모빌아이 (mobileye)	- '17년 8월, 이스라엘기업 모바일아이를 153억달러에 인수 - 모빌아이는 ADAS 솔루션 분야에서 독점적 지위를 구축하며 300여종의 차량에 자사 솔루션을 공급 · 자체 개발한 영상신호 처리 알고리즘을 칩으로 구현한 시스템온칩(SoC) 기술이 모빌아이의 차별화된 기술력
모비디우스 (Movidius)	- '16년 8월, 모비디우스를 인수 - 모비디우스는 구글 Tango 3D 센서 기술에 하드웨어를 공급하고, Lenovo, DJI 등의 회사에 비전 프로세서를 공급한 저전력 고성능 컴퓨터 비전 SoC 칩으로 유명한 스타트업으로, 딥러닝을 소형칩 상에서 저전력으로 구현한 Myriad2를 개발

자료 : 홈페이지 및 기사를 바탕으로 자체 작성

표 10 | 인텔의 인공지능 참여시장, 제품명, 기술역량 원천

참여 시장	제품 명	기술역량 원천
모바일 엣지 시장	MyriadX Myriad2	모비디우스(M&A)
	Cyclone	알테라(M&A)
데이터 센터시장	NNP	너바나(M&A)
	Stratix, Arria	알테라(M&A)
HPC 시스템 시장	knights mill (개발중)	기존CPU역량 + 너바나(M&A)
자율주행차 시장	EyeQ3	모빌아이(M&A)

ARM의 행보

- '17년 5월, 전세계 스마트폰 AP의 IP core의 약 90% 가량을 차지하는 ARM은 모바일 인공지능에 특화된 다이나미크(DynamIQ)기술을 적용한 첫 번째 코어 IP(A75, A55) 공개

- ARM은 직접 반도체를 생산하지 않는 IP회사이며, 이미 모바일 저전력 프로세서 아키텍쳐 시장에서 90% 상당의 점유율을 확보하고 있음. 따라서 향후에도 큰 경쟁 없이 IoT 및 모바일 엣지 디바이스 대상의 AP 시장을 선점할 것으로 전망

  ※ 여전히 모바일에서는 개별 디바이스와 전체 시스템의 전력 소모를 최소화하는 것이 핵심기술이 될 전망

퀄컴의 전략과 행보

- '17년 7월, 퀄컴은 모바일 AP 자체적인 AI처리가 가능하도록 한 모바일 기반 인공지능 프레임워크인 뉴럴 프로세싱 엔진(NPE)를 발표

- NPE는 기존의 머신러닝 솔루션인 제로스를 발전시킨 것으로, AI 어플리케이션 상황에 맞게 자사 AP인 스냅드래곤의 CPU, GPU, DSP 자원을 최적 운용할 수 있도록 지원

- '16년 10월, 퀄컴은 MCU 및 차량반도체의 선두기업인 NXP를 인수. Connectivity와 AP 시장에서의 장점을 활용하여, 드론, 웨어러블, 스마트홈 보안 카메라 시장 등을 선점하려 전략 추구

희미해지는 업체간 영역경계와 주도권 경쟁

- 하드웨어(반도체)와 소프트웨어(알고리즘)이 접점에 위치하는 인공지능 반도체는 인공지능 서비스 성공을 위한 필수적 요소로 인식
- 인공지능 반도체의 신규 시장 창출력과 시장 질서를 재편할 파급력이 확인되면서 ICT기업과 반도체 기업은 영역이 중첩되기 시작
- ICT 기업들(인터넷서비스, 단말, 클라우드 등)은 자사의 AI 서비스 품질강화와 효율적 운용을 위해 반도체 역량을 강화하고 자체 AI 반도체 제작까지 진행
- 또한, 반도체 산업 내에서도 비교적 영역경계 명확했던 모바일과 PC·서버 프로세서 시장의 경계가 희미해지며, PC용 그래픽 카드 시장 영역에 머물던 엔비디아가 인공지능 반도체 全 영역에서 가장 유망한 사업자로 부상하기도 함
- 신흥 강자 엔비디아의 행보와 기존 역량 있는 벤더들(인텔, 퀄컴, ARM 등)의 대응 주목
- 인공지능 반도체시장이 초기 시장인 만큼 산업 신규 진입자의 등장과 산업 내부 기업 간 경쟁양상 변화와 같은 시장역동성은 지속될 것으로 판단

(3) 이슈 3: 개발자 중심 생태계 강화

오픈소스와 개발자 생태계 조성

- 최근 인공지능(딥러닝) 발전에는 오픈소스 라이브러리와 이를 구심점으로 한 개발자 커뮤니티가 큰 기여
- 글로벌 ICT 기업들은 많은 투자를 통해 다양한 형태의 AI 라이브러리를 개발하고 무료로 공개하며 더 많은 개발자를 모집 중

  ※ MS, 누구나 AI 툴 사용해 개발하는 'AI 민주화' 목표 (EPNC, '16.11.04.)
  ※ 구글 AI 대원칙을 발표..."모든 제품에 적용한다, 누구나 쓰게 한다." (조선비즈, '17.11.28.)
  ※ 아마존, '인공지능의 민주화'로 누구나 쓰는 AI 만든다. (IT동아, 17.11.30.)

- 인공지능 반도체 역시 이러한 거대한 소프트웨어 플랫폼과 상호영향을 교차하며 진화 中
  - 반도체 기업은 단순히 하드웨어 칩셋 출시 뿐 아니라, 이러한 오픈소스 AI 라이브러리를 지원하는 API를 지속적이고 신속히 지원해야 함

표 11 대표적인 오픈소스 AI 프레임워크

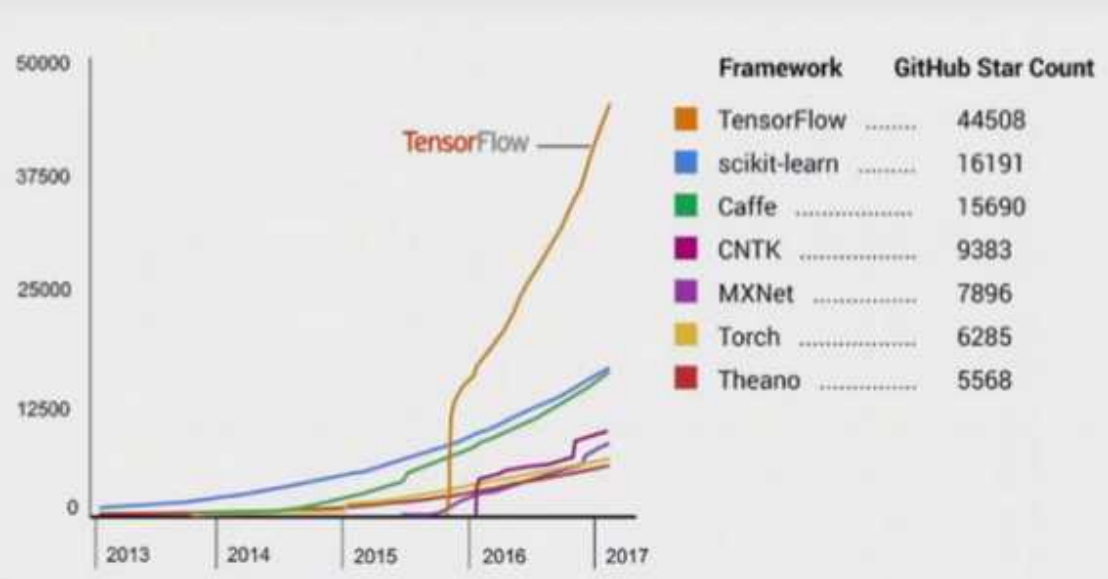
프레임워크	주요내용
TensorFlow Google	- 가장 인기있는 딥러닝 라이브러리 중 하나인 Tensorflow는 Google Brain 팀에서 개발했으며 2015년 오픈소스로 공개 - '2세대 머신러닝 시스템'으로 불리는 Tensorflow는 Python 기반 라이브러리로, 여러 CPU 및 GPU와 모든 플랫폼, 데스크톱 및 모바일에서 사용 가능 - C++ 및 R과 같은 언어도 지원하며 딥러닝모델을 직접작성가능
Microsoft CNTK Microsoft	- CNTK라는 약어로 알려져있는 Microsoft Cognitive Toolkit은 딥러닝 모델을 교육하기 위한 오픈소스 딥러닝 도구 - 고도로 최적화되었으며 Python 및 C++와 같은 언어를 지원 - Cognitive Toolkit을 사용하여 강화학습 모델 또는 Generative Adversarial Networks (GAN)를 쉽게 구현 가능 - 높은 확장성과 성능을 발휘하도록 설계되었으며 여러 시스템에서 실행될 때 Theano 및 Tensorflow와 같은 다른 툴킷과 비교할 때 높은 성능을 제공
Caffe2 Facebook	- 표현, 속도 및 모듈성을 염두에 두고 개발된 Caffe는 Berkeley Vision and Learning Center (BVLC)에서 주로 개발한 최초의 딥러닝 라이브러리 중 하나 - Caffe를 페이스북은 최근 고성능 개방형 학습 모델을 구축 할 수 있는 유연성을 제공하는 새로운 가벼운 모듈 식 딥러닝 프레임 워크인 Caffe2를 공개

자료 : http://blog.daum.net/lonemoon/87 내용을 발췌·정리

- 가장 대표적인 딥러닝 오픈소스 라이브러리는 구글의 텐서플로우로, 텐서플로우 생태계를 통해 구글은 다양한 이점을 확보
  - **(수 많은 개발 데이터 확보)** 자사 직원 뿐 아니라 외부의 전문가의 개발 코드와 훈련 데이터 확보 가능
  - **(혁신 가속화)** 인공지능 분야는 기술 변화가 매우 빠르기 때문에 학계와 업계의 협업과 오픈 이노베이션 문화가 자리 잡고 있음. 구글은 텐서플로우의 개방적인 커뮤니티 활용하여 기술 혁신속도를 가속
  - **(인재 채용 채널)** 구글 같은 기술업체는 인재가 가장 중요한 자산이며, 오픈소스 커뮤니티는 최고의 인재를 채용하는 가장 좋은 경로 중 하나
- 구글은 안드로이드와 같은 모바일의 플랫폼과 같이 텐서플로우를 인공지능의 플랫폼으로 활용할 것으로 전망
  - 아래그림은 깃허브(GitHub)에서 텐서플로우가 출시 이후 엄청나게 빠르게 확산, 개발자 사이에서 활용되고 있음을 보여짐

- 구글은 모바일, IoT환경에 적합한 프레임워크인 텐서플로우 라이트를 출시하여 영향력을 더욱 강화

구글 텐서플로우의 개발자 생태계 구축 속도



자료 : IDG(2017)

개발자친화 문화 : 엔비디아 사례

- 엔비디아는 대규모 개발자 이벤트 등을 주최하고, 최신기술 교육, 스타트업 지원, 사회공헌 활동, 개발소스 제공 등을 통해 지속적으로 개발자와 소통하는 개발자친화 문화 전략으로 자체 생태계를 곤고히 조성

- **(GPU Technology Conference개최)** 2009년부터 매년 개최하고 있는 GTC는 인공지능, 가상현실, 자율주행차 분야의 개발자, 데이터 과학자 등이 참석하는 세계 최대 규모의 개발자 행사 중 하나

  - 500개 이상의 발표 세션 및 150개 이상의 전시가 마련되며, 딥 러닝과 인공지능, 데이터센터 및 클라우드 컴퓨팅, 애널리틱스, 헬스케어, 자율주행 및 인공지능 차량, 생명과학, 방위산업, 가상/증강현실 등 업계 내 가장 주목 받고 있는 주요 주제들을 다룸

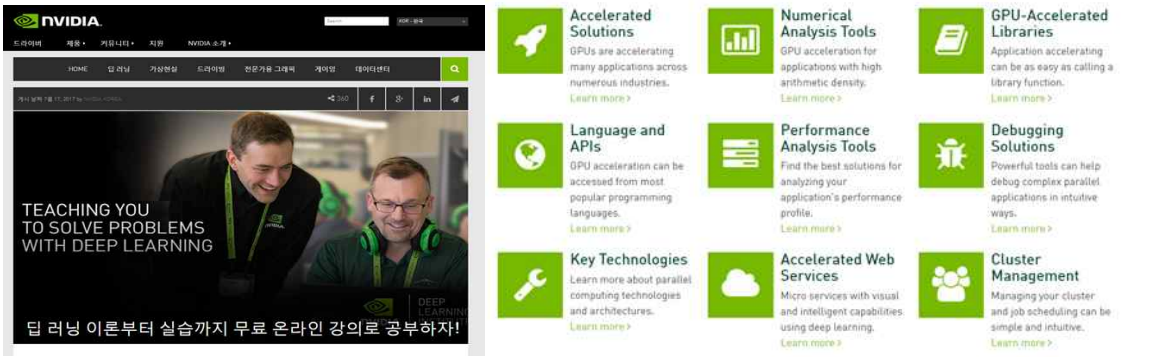
- **(개발자 교육지원)** 엔비디아 딥 러닝 인스티튜트는 개발자 및 데이터 과학자들이 각자의 전문 기술을 더욱 학습할 수 있는 기회 제공

  - 최신 AI 프레임워크 및 SDK를 활용해 자율주행차량 기술, CUDA, 딥 러닝,

임베디드 애플리케이션, 가상현실 등의 다양한 분야에 대해 학습자 수준에 맞춘 교육을 제공

- 엔비디아 딥 러닝 인스티튜트에서 제공하는 강의를 오프라인에서 뿐 아니라 온라인에서도 학습 가능

엔비디아 딥러닝 인스티튜트 홈페이지(좌)와 개발자 제공 솔루션들(우)



자료 : http://blogs.nvidia.co.kr/2017/07/17/dli-self-paced-lab/

- **(스타트업 인큐베이팅)** 엔비디아는 AI 스타트업을 지원하기 위한 인셉션 프로그램(Inception program)를 운영

  - 최신 딥러닝 기술/장비 지원, 최신 전문지식 교육, 글로벌 네트워킹, 자금 지원 등

- **(사회적 공헌)** 매년 진행되는 '글로벌 임팩트 어워드'는 AI 시스템을 통해, 중요한 사회 및 인도주의적 문제 해결에 앞장서는 연구단체를 선정, 15만 달러의 상금을 제공

  - '17년 수상작은 '인공지능 기술로 뇌종양 치료법을 개선시킨 메이오클리닉'

- **(개발소스 제공)** 엔비디아는 오픈소스 운동에 적극 동참하면서 자신들의 하드웨어 노하우의 상당 수를 NVDLA.org사이트에 오픈소스로 공개

  - NVIDIA Deep Learning Accelerator (NVDLA) 소프트웨어, Xavier 기반의 추론 칩 하드웨어 디자인과 소스 등 공개

생태계 전략 : 인공지능 반도체 벤더의 필수요건

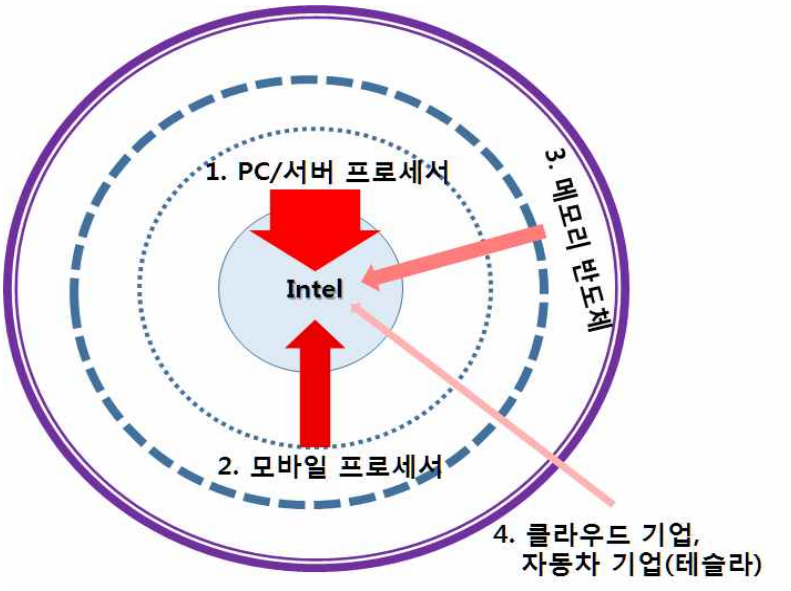
- 개발자 생태계 구축은 인공지능 반도체 시장주도권을 선점하는 핵심요인
  - 인공지능 반도체는 알고리즘 개발자와 반도체 공급자가 중, 어느 한쪽이 일방적인 지배력을 갖는 것이 아니라 상호작용을 통해 진화
- 오픈소스와 개발자친화 문화는 인공지능 반도체 기업의 개발자 생태계 구축의 중요한 전략도구
  - 개발자들이 많이 쓰면 쓸수록 혁신의 속도와 다양성에 대응하기 용이
- 인공지능 반도체 벤더들은 단순 하드웨어 칩셋 뿐 아니라, 개발자 생태계를 고려하여 오픈소스 프레임워크와 라이브러리를 지원하는 API나 SDK를 제공하는 등 포괄적인 개발 환경 지원을 기본적으로 고려할 필요
- 아울러, 개발자친화 문화를 바탕으로 외부 개발자들과 다양한 소통채널을 확보하고, 이들의 생태계 참여를 유도하는 방안 강구 필요

(4) 이슈 4 : 반도체 산업구조 변화

주도기업이 받는 경쟁위협

- 인공지능 시장이 성장하면서, 반도체 산업 부동의 1위 기업, 인텔은 산업 안팎에서 다양하고 강한 경쟁위협을 겪는 중
- 먼저, 반도체 산업 내 업체 경쟁위협이 다변화
  - **PC/서버 프로세서 벤더** : GPU 벤더 엔디비아는 가장 강력한 인공지능 HW 솔루션 업체로 위상강화, 인텔에 가장 위협을 주는 기업
  - **모바일 프로세서 벤더** : 퀄컴, ARM 등 모바일 프로세서의 강자들과 모바일 엣지 AI 시장을 중심으로 경쟁 심화가 전망
  - **메모리 반도체 벤더** : 기술진화 방향(프로세서+메모리)에 따라, 향후 개별 시장에서 존재하던 메모리(삼성, 하이닉스 등)벤더와 프로세서 벤더들은 상호경쟁관계나 가치사슬 상 상·하의 위상을 갖는 등 상당한 산업구조 변화가 전망
- 산업 외부 경쟁자 진입 허용
  - **클라우드 기업, 자동차기업** : 구글, 마이크로소프트 등 클라우드 서비스 사업자들나 테슬라 같은 자동차 기업이 자체 반도체 역량을 갖추거나, 직접제작하면서 인텔의 가장 큰 수익시장이 일부 축소가 불가피

인공지능 시대, 주도기업(Intel)이 받고 있는 경쟁위협

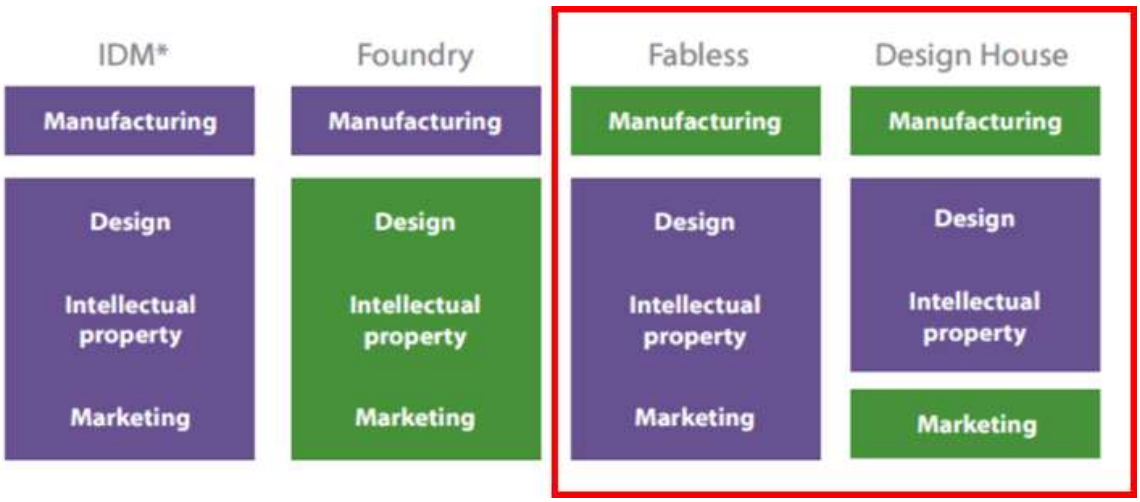


자료 : 저자작성, 주) 화살표 굵기는 위협크기를 의미

시장수요 다변화와 가치사슬 변화

- 인공지능의 발전은 반도체 산업의 새로운 시장과 고객을 형성하므로, 프로세서 벤더들은 다변화된 시장 수요환경을 맞을 것으로 전망
  - 개별 산업도메인(vertical industry)에서 디지털 전환(DX)가 본격 진행될수록 프로세서 반도체에 대한 신규수요가 증대
  - 특히 디지털 전환이 심화될수록 범용 프로세스보다 도메인 요구사항에 최적화된 특화 칩의 수요가 증가할 것
- 시장수요 다변화 속에서 다품종 소량생산 체계가 정립되고, 이를 위한 디자인하우스와 스타트업·중소기업을 중심으로 한 팹리스가 활성화될 전망
  - 시장 수요가 다변화 됨에 따라 현재의 분업화를 통한 소품종 대량생산 체제는 비효과적 임
  - 디자인하우스 : 설계 서비스 전문회사로 설계기업의 다품종·소규모 물량을 취합하여 파운드리 기업의 대량생산체계와 연계 기능 담당
  - 팹리스(fabless): 제조 설비를 뜻하는 패브리케이션(fabrication)과 리스(less)의 합성어

반도체 제조기능별 업체 유형



자료 : https://promwad.com/publications/article-pakholkov-toj-electronics-development-outsourcing-idh

프로세서 다양성 시대, 핵심경쟁력과 스타트업

- **(프로세서의 다양성 시대)** AI 시대에는 서로 다른 종류의 프로세서들이 시장에서 가질 수 있는 기회가 많아질 것으로 전망. 즉, 범용 프로세스보다 도메인 요구사항에 최적화된 특화프로세서의 수요가 증가, 킬러앱이 없는 롱테일 시장으로 산업구조가 개편 전망

  - 규모와 복잡성에 상관없이 모든 연산을 하나의 거대한 CPU가 처리하던 시대 종료

- **(IP가 핵심경쟁력)** 인공지능 반도체개발에서 설계기술(IP; 설계자산)이 핵심 부가가치 원천이 될 전망으로 기술력을 확보한 스타트업도 경쟁력 확보 가능

  - 미국경우, AI반도체 스타트업 창업이 활발히 진행되고 있고, 상당수의 유망 기업은 성공적 exit를 진행

  - 우리보다 반도체 기술력이 떨어진다고 인식되는 중국의 경우도 다수의 AI 반도체 스타트업을 보유. 이들은 상당수준의 기술력을 인정받아, 자국 뿐 아니라 미국 VC 투자도 유치

    ※ 캠브리콘은 알리바바의 투자를 지원 받고 있으며, 중국 AI 반도체 최초로 기업 가치가 10억불을 상회. Cambricon-1A의 화웨이의 최신 AP Kirin 970에 탑재

◐ 반면, 국내 인공지능반도체 관련 스타트업은 찾아보기 어려우며, 연구기관도 ETRI(넥스트칩에 기술이전)와 KAIST와 같은 출연연 기관정도임

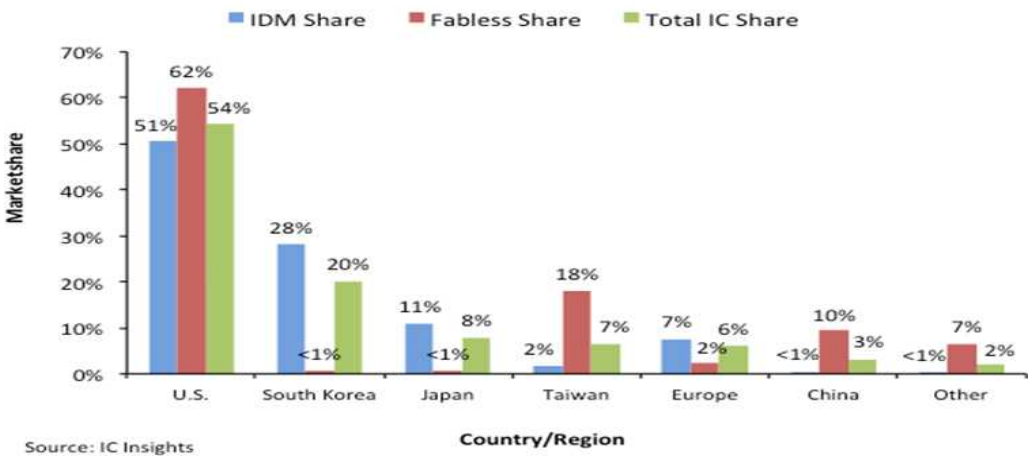
인공지능 반도체 주요 스타트업 현황

자료: https://basicmi.github.io/Deep-Learning-Processor-List/#Horizon_Robotics 내용 정리

◐ 국내의 프로세서 설계역량과 팹리스 경쟁력은 매우 취약한 편으로 이를 개선하기 위한 정책적 노력이 필요

- 국내 팹리스 기업이 전세계 시장에서 차지하는 비중은 '16년에는 1% 이하

국내 팹리스 산업 경쟁력(점유율)



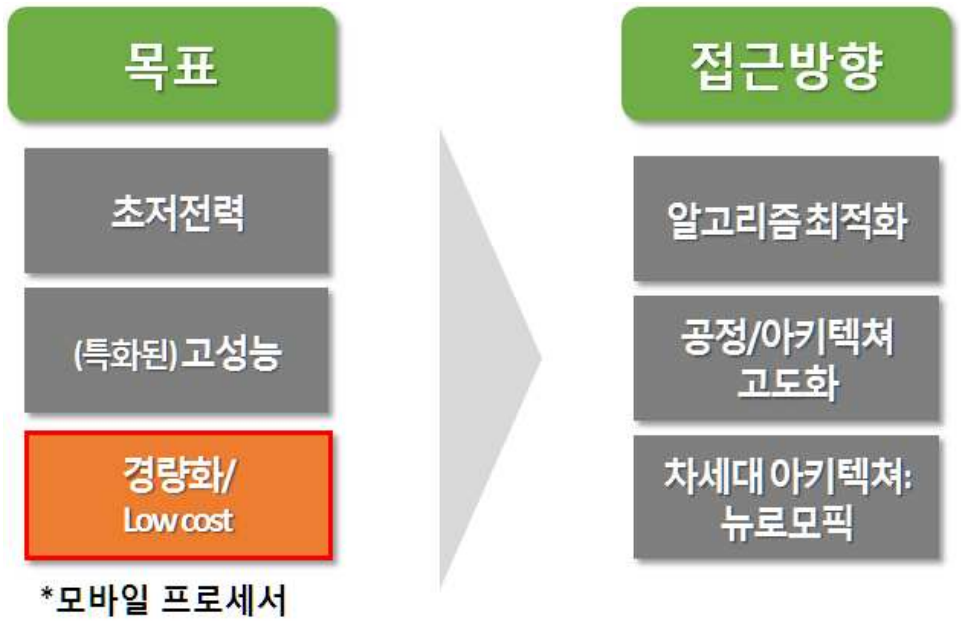
자료 : https://www.electronicsweekly.com/news/business/us-companies-have-54-of-ic-market-says-ic-insights-2016-04/

(5) 이슈 5 : 기술 발전 가속화

인공지능 반도체 기술발전 방향성

- 인공지능 반도체의 기술진화 방향은 초저전력, (특화된)고성능 반도체로 대별가능
  - 모바일 프로세서의 경우, 경량화와 원가절감(low cost)이 추가적 목표
- 이를 위한 접근방향으로, 소프트웨어적인 알고리즘 최적화와 하드웨어 측면의 반도체 공정과 아키텍쳐 고도화가 병행되고 있음
  - 궁극적으로는 학습과 추론이 온칩(on-chip)으로 모두 가능한 뉴로모픽 칩으로 발전

인공지능 반도체 기술발전 방향

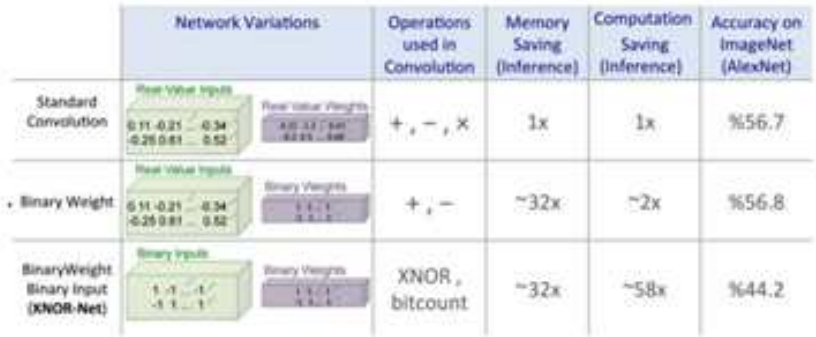
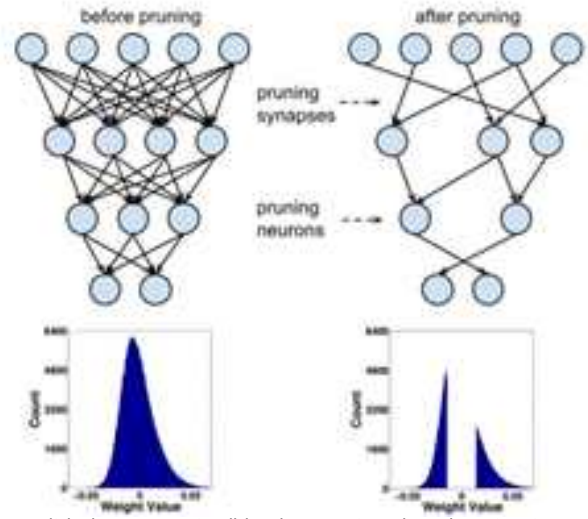


자료 : 저자작성

인공지능 알고리즘 최적화

- 프로세싱 파워와 전력 사용량을 줄이기 위해서 신경망 알고리즘의 수치 정밀도 조정, 신경망 가지치기, 최적화된 신경망 압축하기 등의 방안이 알고리즘 최적화 차원에서 시도

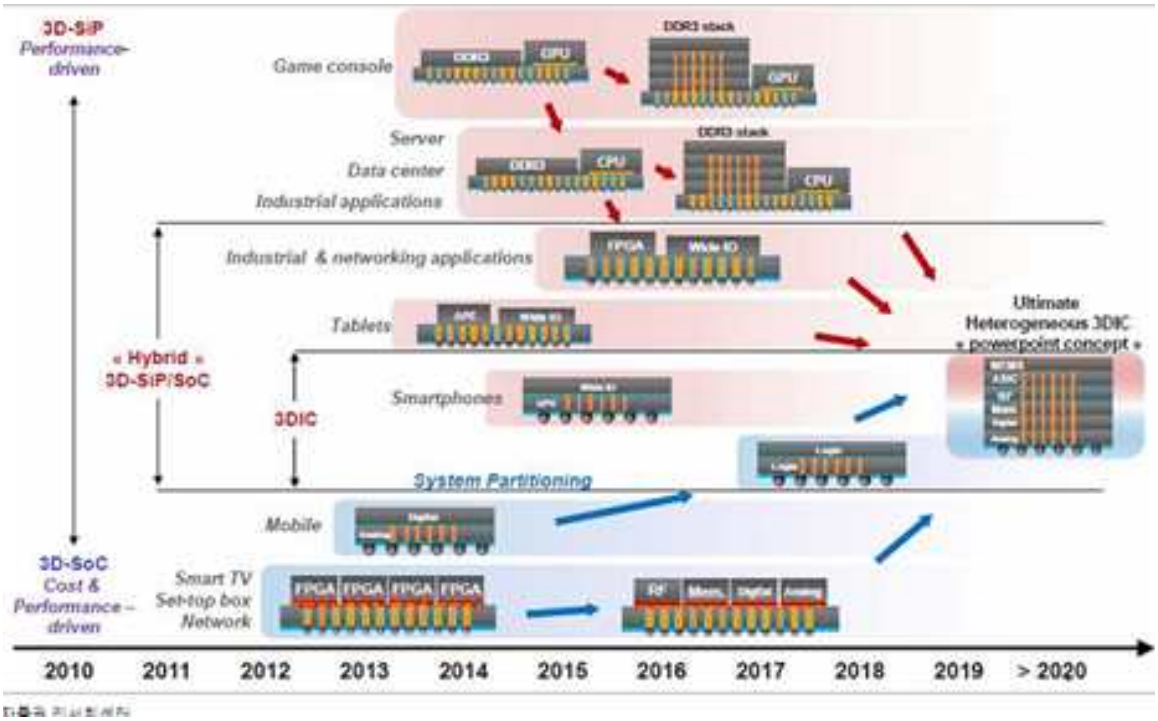
표 12 | 인공지능 알고리즘 최적화

접근 방향	내용
신경망 데이터변수 정밀도 조절	- 어느 정도 데이터변수 정확도를 떨어트려도 실제 추론 정밀도는 크게 하락하지 않는 것에 착안 - 추론에선 학습보다 데이터 정밀도를 더 낮춰도 되기에 16비트 부동소수점/정수에서 8비트 정수, 4비트 정수, 더욱 극단적으로는 바이너리, 정밀도까지 시도되고 있음  자료: https://www.slideshare.net/anirudhkoul/squeezing-deep-learning-into-mobile-phones
신경망 가지치기	- 추론 신경망 가지치기(Pruning): 추론을 위해 학습된 신경망에서 중요도가 떨어지는 연결을 삭제. - 가중치가 '0'에 가까워 크게 의미가 없는 연결을 삭제하는 방식. 이로서 신경망의 가중치 데이터 용량이 극적으로 줄어듦  자료: https://www.slideshare.net/anirudhkoul/squeezing-deep-learning-into-mobile-phones
최적화된 신경망 압축	- 학습된 신경망 데이터 자체를 압축하는 방안 - 모바일 NPU의 경우, 신경망 데이터 자체를 온칩 메모리에 넣을 수 있어, 전력소모와 속도 성능 극적 향상 가능

반도체 공정 및 아키텍쳐 고도화

- **(데이터 병목 경감 기술)** 대부분의 컴퓨터 데이터 병목현상은 메모리와 프로세서 사이에 발생. 메모리와 프로세서를 보다 근접하게 하는 방안 진행中
  - 데이터 병목을 줄이기 위한 메모리 기술로 HBM2 등이 현재 사용되고, 프로세서-메모리-스토리지 통합설계나 맴리스터(메모리와 프로세서 원칩화)기술이 연구개발 中
- **(이기종 컴퓨팅 아키텍쳐)** 호스트 프로세서로서 CPU와 특정 도메인 연산 특화된 가속코어가 함께 탑재 되는 구조가 일반화
- **(원칩 패키징 솔루션)** 메모리-프로세서 통합, 이종 프로세서간의 통합을 위한 반도체 원칩 패키징 솔루션 고도화

반도체 원칩화 로드맵



자료: NH투자증권 리서치센터

- 다양한 형태의 알고리즘에 최적화된 반도체의 아키텍쳐 개발 필요
  - CNN외 generative adversarial networks (GANs)이나 one-shot learning는 기존과는 다른 하드웨어 아키텍쳐를 요구[11]

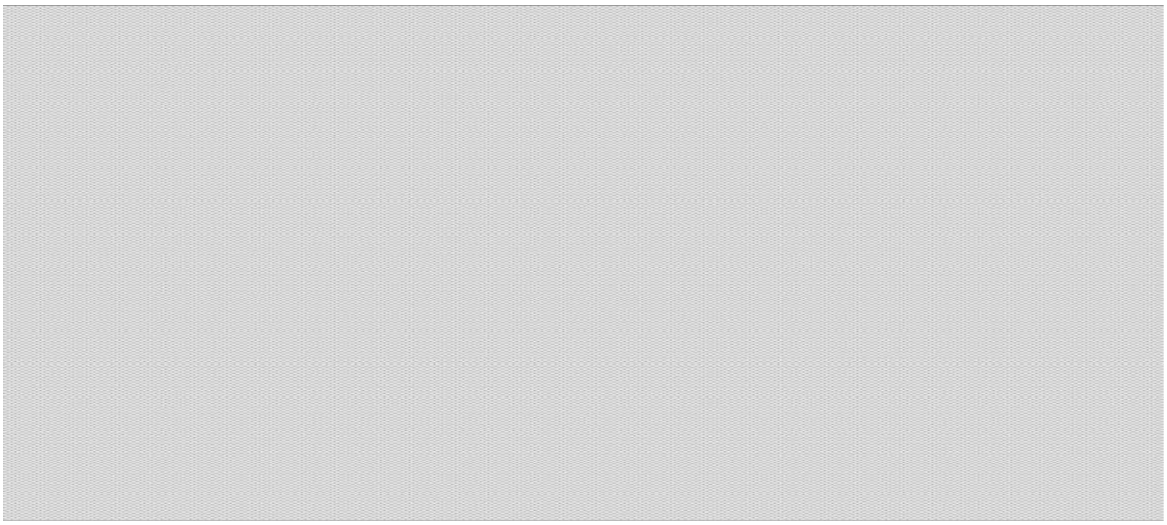
차세대 아키텍쳐 : 뉴로모픽 칩

- **(뉴로모픽 칩)** 폰 노이만 컴퓨팅 구조의 한계 극복, 인공지능 알고리즘 구현에 적합한 인간 두뇌 신경 모방형 프로세서인 초저전력으로 온칩(on-chip)에서 학습과 추론이 모두 가능

  - 뉴로모픽칩은 뉴런시냅스 모사 소자와 프로세싱 구조를 가지며, 프로세싱과 메모리가 동일소자에서 진행되는 특징을 갖음

  ※ 뉴로모픽 후보소자로는 EEPROM이나 Memristor(memory+resistor)가 고려되고 있음

뉴로모픽칩 특징



자료 : Frost & Sullivan(2016).

- **(기술수준)** 다양한 소재 및 구조를 갖는 뉴로모픽 칩 연구가 전 세계적으로 진행되고 있으나 아직까지 실용성이 담보된 고집적 신경망 기술은 없으며, 대부분 초기수준으로 판단

- **(기술발전 가속)** 가트너는 뉴로모픽 AI 하드웨어의 시장 안정화 도달시기를 '16년 10년 이상에서 '17년 5~10년으로 변경

  - 글로벌 기업은 뉴로모픽 칩에 대한 지속적인 R&D를 진행 중

  ※ IBM(TrueNorth;2014), Intel(Loihi, 개발중)

- 따라서, 고집적·고효율 뉴로모픽 프로세서와 같은 변혁적 기술에 국가적 차원의 R&D가 선제적으로 투입될 필요

11)http://www.forbes.com/sites/quora/2016/08/05/this-is-the-cutting-edge-of-deep-learning-research/#5bcbceaa27cb

IV 맺음말

- 인공지능 시장이 성숙할수록 인공지능 반도체에 중요성은 더욱 강조될 전망
  - 인공지능 반도체는 인공지능 서비스가 실체적인 존재로서 가능하게 하는 핵심 요소
  - 서비스 개념정립과 구현가능성은 클라우드 인공지능 플랫폼으로 가능하나, 시장 수용성(경제성, 안정성)을 높이기 위해서는 최적화된 H/W 프로세서가 뒷받침되어야 함
- 본 연구는 이런 배경과 인식 하에 인공지능 반도체의 배경과 현황, 그리고 시장·생태계·기술관점에서 현안사항을 논의하였고, 주요내용은 다음과 같음
- 첫째, 인공지능 시장이 클라우드 시장과 모바일 엣지 시장으로 분화하면서, AI 추론을 영역을 담당하는 모바일 엣지 AI가 새롭게 부상
  - 지능화 서비스와 사물인터넷이 확산될수록 단말 자체에서 인공지능 추론 연산을 처리하는 모바일 엣지 AI의 중요성이 커질 전망
  - 인공지능반도체 시장은 클라우드向 시장(기존)과 모바일 엣지向 시장(신규)으로 양분될 것으로 전망되나, 이들은 배타적 관계가 아닌 상호보완적 관계로 진화
- 둘째, 인공지능 반도체의 시장경쟁이 심화. 기존 반도체 시장영역경계가 희미해지고, 주도기업이 변화 중
  - 인공지능 반도체의 신규 시장 창출력과 시장 질서를 재편할 파급력이 확인되면서 ICT기업과 반도체 기업의 사업영역이 중첩되기 시작
  - 비교적 영역경계 명확했던 모바일과 PC·서버 프로세서 시장의 경계가 옅어지며 산업내부 기업 간 경쟁 역동성이 증대
- 셋째, 개발자 생태계 구축은 인공지능 반도체 시장 주도권을 확보하는 핵심전략이며, 오픈소스와 개발자친화 문화는 인공지능 반도체 벤더들이 반드시 고려해야 하는 생태계 전략도구
- 넷째, 인공지능의 진전으로, 반도체산업 시장수요 다변화와 가치사슬 변화가 진행되며, 이는 반도체 산업구조 변화가능성을 높임
  - 다변화하는 시장 수요 속에서 다품종 소량생산 체계가 정립되고, 이를 위한 디자인하우스와 스타트업·중소기업 중심의 팹리스가 활성화될 전망

- 개별 산업 내 디지털 전환(DX)이 심화될수록, 범용 프로세스보다 도메인 요구사항에 최적화된 특화 칩 수요가 증가하며 킬러앱이 없는 롱테일 시장으로 산업구조가 개편

- 끝으로, 인공지능 반도체의 기술발전은 가속화될 전망. 초저전력, (특화된) 고성능 인공지능 프로세서를 위해, 소프트웨어적인 알고리즘 최적화와 하드웨어 측면의 반도체 공정 고도화가 병행되며, 궁극적으로는 학습과 추론이 온칩(on-chip)으로 모두 가능한 뉴로모픽 아키텍쳐로 발전 전망

※ 참고자료 : 인공지능 반도체 개발 출시 동향

◐ 모바일 엣지 시장 1 : Stand alone 칩

기업	AI칩 개발 관련 동향	비고
Movidius (現 Intel)	- '17년 8월 인텔은 딥러닝 처리를 위한 전용 프로세서 미리아드X를 발표 - 미리아드X는 인텔이 '16년 인수한 모비디우스가 개발한 칩셋으로 VPU(Vision Processing Unit) 프로세서에 영상과 신경망 처리를 가능하도록 설계 - 인텔은 소형 드론, 로봇, 보안카메라 등 높은 수준의 자율 행동이 요구되는 소형 디바이스에 적용예정 - 미라이드X는 센서와 카메라를 통해 주변 환경을 시각적으로 인식하고 스스로 상황 분석과 판단 가능 - 크기는 8.7×8.5㎜에 불과	stand alone ASIC 자율주행 드론, 로봇 비전 인식
Mobileye[12](現 Intel)	- 모빌아이 EyeQ 시리즈는 자율주행에 특화된 비전 인식 프로세서로 BMW, GM, Opel, Volvo등 메이저 자동차 업체 차량에 탑재 - EyeQ3는 2014년에 출시되었고 2018년에 eyeQ4 출시 예정	stand alone ASIC 자율주행차량 비전인식
Nvidia Jetson TX2	- '17년 3월, 임베디드 시스템에서 인공지능 컴퓨팅을 구현하는 Jetson TX2 공개 - 신용카드 크기의 플랫폼인 젯슨 TX2는 고도의 인텔리전스를 갖춘 공장 로봇, 상업용 드론, 인공지능 도시를 위한 스마트 카메라 등을 구현하는 데 활용 - 텐서 RT 1.0를 지원. - 텐서 RT는 딥 러닝 애플리케이션의 제품 구축을 위한 고성능 뉴럴 네트워크 추론 엔진	GPU기반 IoT응용 지원

12) https://www.mobileye.com/our-technology/evolution-eyeq-chip/

모바일 엣지 시장 2 : SoC 코어

기업	AI칩 개발 관련 동향	비고
STMicroelectronics[13]	- 심층CNN 가속 프로세서를 담은 28nm급 SoC개발 중 - 초저전력 SoC CNN 가속이 특징	SoC 개발중
Huawei	- '17년 9월 화웨이는 세계최초로 모바일용 AI 전용칩셋 (NPU)를 탑재한 자사의 최신 AP, Kirin 970을 출시 - 중국의 스타트업 캠브리콘의 IP를 기술 이전하여 출시한 것이 특징	SoC core 모바일 추론
Apple	- 17년 9월, 뉴럴 엔진이 탑재된 A11 바이오닉을 공개 - 뉴럴엔진은 아이폰에서 머신 러닝과 안면인식 등을 할 때 사용 - 애플은 뉴얼 엔진 정확도를 높이기 위해 헐리우드 스튜디오에서 제작한 10억개 이상의 안면을 이용해 훈련한 것으로 알려짐	SOC core 모바일 추론
MS[14]	- Hololens2는 MS가 개발한 스마트 스마트 고글로 AR 지원함. 이를 위해 Holographic Processing Unit(HPU)를 탑재함.(2017.7) - HPU는 이미지 처리를 맡아 배터리를 전력소모를 줄이고 CPU부하를 줄일것으로 기대.	이미지처리 SoC 모바일추론
Qualcomm	- '17년 7월, 모바일 AP 자체적인 AI처리가 가능하도록 한 모바일 기반 AI 프레임워크인 뉴럴프로세싱엔진(NPE)를 발표 - NPE는 기존의 머신러닝 솔루션인 제로스를 발전시킨 것으로, AI 어플리케이션 상황에 맞게 자사 AP인 스냅드래곤의 CPU, GPU, DSP 자원을 최적운용	SW 프레임워크로 물리적인 AI가속 코어(칩) 아님
Altera(現 Intel)	- 2월 인텔은 산업용 IoT와 자동차 시장을 겨냥한 FPGA '사이클론 10 시리즈' 10GX와 10LP를 출시	FPGA기반 IoT, 자율주행용

13) https://reconfigdeeplearning.files.wordpress.com/2017/02/isscc2017-14-1visuals.pdf
14)https://www.microsoft.com/en-us/research/blog/second-version-hololens-hpu-will-incorporate-ai-coprocessor-implementing-dnns/

모바일 엣지 시장 2 : SoC IP

기업	AI칩 개발 관련 동향	비고
ARM	- ARM은 2017년 5월, AI에 특화된 다이나미크(DynamIQ) 기술을 적용한 첫 번째 코어 프로세서(A75, A55)를 공개 - ARM은 A75와 A55프로세서가 향후 3년에서 5년 내 AI 성능을 약 50배 가량 향상시킬 것으로 전망	IP core 모바일 추론
Imagination Technologies	- '17년 9월 이매지네이션은 PowerVR 2NX NNA (Neural Net Accelerator)을 출시 - flexible bit depth 지원: 16bit에서 4bit로 신경망 솔루션 정밀도 조정가능하여 연산속도 향상	IP core 모바일 추론
Cadence Design Systems[15]	- '17년 5월, Cadence은 신경망처리에 최적화된 DSP IP인 Vision C5를 출시 - C5는 비전, rader/lidar 데이터 처리에 최적화된 신경망을 자체 탑재한 최초의 DSP IP core로 알려짐 - 회사는 AI 가속기에서 처리한 데이터를 다시 DSP로 이동해서 처리하는 불필요한 작업 없이 DSP에서 AI전체를 다 처리할 수 있음으로 효율적이라 주장 - 회사는 상용 GPU 보다, AlexNet CNN 처리에 6배 정도 빠른 처리성능을 보인다고 주장 - 자율주행, 지능형 CCTV, 드론, 웨어러블 등의 응용에 활용될 것으로 전망	DSP IP 모바일추론

15)https://www.cadence.com/content/cadence-www/global/en_US/home/company/newsroom/press-releases/pr/2017/cadence-unveils-industrys-first-neural-network-dsp-ip-for-automo.html

◐ 데이터 센터 시장 : GPU/FPGA/ASIC

기업	AI칩 개발 관련 동향	비고
Google	- '17년 5월, TPU2를 발표 - AI SW framework인 텐서플로우의 최적화된 ASIC 기반의 클라우드(서버) AI 가속기. - 구글 AI서비스에 CNN 응용에 전반적으로 활용	ASIC 데이터센터 학습+추론
Nervana[16](現 Intel)	- Intel Nervana NNP (Neural Network Processor)는 코드명 Lake Crest라고도 불리움. - 2017년 10월에 발표한 이 칩은 인텔의 딥러닝을 위한 최초의 ASIC 기반 상용칩셋으로 인텔이 인수한 너버나 시스템즈의 제품에 기반하고 있음 - CPU, GPU에 비해 행렬곱연산에 특화	ASIC 데이터센터 학습+추론
Nvidia Volta platform	- 17년 5월 개최된 엔비디아의 GPU 테크놀로지 컨퍼런스(GPU Technology Conference, 이하 GTC)에서 첫 공개된 엔비디아의 최신 GPU플랫폼	GPU 데이터센터 학습+추론
AMD[17]	- 라데온 인스팅트(Radeon Instinct)는 AMD의 딥 러닝 지향 GPU 브랜드로, 딥 러닝, 인공신경망, 슈퍼컴퓨터/GPGPU 응용을 가속화하기 위해 고안 - 16년 12월 첫 출시	GPU 데이터센터 학습+추론
Xilinx Virtex UltraScale+[18]	- 자일링스의 버텍스(virtex) 울트라스케일(UltraScale™) 디바이스 - New 16nm Virtex FPGA로 구동되는 가속 스택(Reconfigurable Acceleration Stack) 'reVISION'은 개발자보드와 FPGA, 프레임워크, 라이브러리, 오픈스택 등이 포함돼 있는 것이 특징	FPGA 데이터센터 학습+추론
Altera(現 Intel)	-인텔의 스트라틱스(Stratix)는 FPGA 라인업 중 가장빠르고 파워풀한 제품으로 2013년 출시된 stratix 10이 가장 최신 제품 - 주로 데이터 센터 및 클라우드 센터의 HPC 서버 등에서 활용 - 인텔의 '아리아(ARRIA) 10 FPGA'는 데이터를 수집하고 전송하는 한편, 다양한 IoT기기에서 수집된 정보를 기반으로 실시간 결정을 내릴 수 있도록 지원	FGPA 데이터센터 학습+추론

16)https://www.intelnervana.com/intel-nervana-neural-network-processors-nnp-redefine-ai-silicon/?_ga=2.62312428.1380200850.1508486032-2008757629.1504021982

17)https://ko.wikipedia.org/wiki/%EB%9D%BC%EB%8D%B0%EC%98%A8_%EC%9D%B8%EC%8A%A4%ED%8C%85%ED%8A%B8

18) http://www.eewebinar.co.kr/xilinx/technical_view.asp?g=3&idx=302

HPC 시스템 솔루션 시장

기업	AI칩 개발 관련 동향	비고
Nvidia DGX 시리즈[19]	- '17년5월, Volta 플랫폼 기반의 Tesla V100 가속기로 제작된 시스템 출시 - DGX-1는 기존 서버 250대와 맞먹는 성능을 제공하는, 딥 러닝과 AI 분석 가속화를 위한 목적으로 구축된 세계 최초의 시스템	딥러닝용 수퍼컴퓨터 학습, 추론 데이터센터
Intel[20]	- 인텔최초 HPC용 딥러닝 학습용 프로세서 솔루션인 khights mill의 상세 로드맵을 2017년 8월 공개 - knights mill(코드네임: 나이트 밀) 차세대 Xeon Phi(제온 파이) 기반 딥러닝 솔루션 공개 (2017년) - SoC 형태로 제온 파이를 호스트 프로세서로 하고 AI 가속기인 lake crest, knights crest도 co-processor로 탑재 될 전망 - 인텔은 추론용 딥러닝 프로세서는 인수한 알테라의 FPGA를 중심으로 제품라인을 가져가고 있음. 이 알테라의 FGPA솔루션은 MS의 cloud Azure에 광범위하게 활용되고 있음	CPU(메니코어) 인텔 제온 프로세서 라인 기반 슈퍼컴퓨터 (HPC)용
후지쯔[21]	- 17년8월, 후지쯔는 2018년까지 HPC용 자체 AI 프로세서인 DLU(Deep learning Unit)을 제작한다고 밝힘. - 후지쯔는 자체 processor 에 대한 자체제작 능력 가진 일본의 대표적 슈퍼컴퓨터 제작사. - Spark64이란 자체 아키텍쳐 기반의 processor 개발 활용, 슈퍼컴1위를 K computer 제작	HPC용 학습용

차세대 플랫폼 (뉴로모픽 시장)

기업	AI칩 개발 관련 동향	비고
IBM	- IBM이 개발한 TrueNorth는 DARPA 의 SyNAPSE 프로그램에서 지원을 받은 인공 지능 연구 - 하나의 칩에 64 X 64 코어가 탑재되어 하나의 코어마다 SRAM,뉴론등 각각의 프로세스를 빠르게 처리할수 있도록 구현한 것이 특징 - '17년 6월, IBM과 미 공군 연구소는 64개의 '트루노스 뉴로시냅틱' 칩을 이용한 AI 컴퓨터 시스템을 개발한다고 발표	뉴로모픽칩 개발 중
Intel[22]	- 17년 10월, 인텔은 스스로 학습할 수 있는 자율 학습형 뉴로모픽칩, 로이히(Loihi)의 테스트하고 있다고 밝힘. 로이히는 13만개의 뉴런과 1억3000만개의 시냅스로 구성	뉴로모픽칩 개발중

19) http://blogs.nvidia.co.kr/2016/07/14/dgx-1_story/
20) https://www.top500.org/news/intel-spills-details-on-knights-mill-processor/
21) https://www.nextplatform.com/2017/08/09/fujitsu-bets-deep-leaning-hpc-divergence/
22) https://newsroom.intel.com/editorials/intels-new-self-learning-chip-promises-accelerate-artificial-intelligence/

※ 약어표

AI	Artificial Intelligence
ANN	Artificial Neural Networks
API	Application Program Interface
ASIC	Application-Specific Integrated Circuit
CPU	Central Processing Unit
DSP	Digital Signal Processors
FPGA	Field ProGrammable Array
GAN	Generative Adversarial Networks
GPU	Graphics Processing Unit
IC	Integrated Circuit
IoT	Internet of Things
IP	Intellectual Property
MCU	Micro Controller Unit
MPU	Micro Processor Unit
RNN	Recurrent Neural Networks
SDK	Software Development Kit
SNN	Spiking Neural Network
TPU	Tensor Processing Unit

※ 참고문헌

Frost & Sullivan. (2016). “Future of Artificial Intelligence Investigating the Hardware and Machine Learning Technologies that would Realize the AI Agents of the Future”

Gartner. (2016). “Find the Right Accelerator for Your High-Performance Computing Needs.”

Gartner. (2016). “Market Insight: Disruptive Macro Trends for 2025 Personal Tech Market – Artificial Intelligence – Me, Myselves and AI.”

Gartner. (2017). “AI Creates New Semiconductor Business Opportunities.”

Gartner. (2017). “AI on the Edge: Fusing Artificial Intelligence and IoT Will Catalyze New Digital Value Creation.”

Gartner. (2017). “Forecast Analysis: Electronics and Semiconductors, Worldwide, 1Q17 Update.”

Gartner. (2017). “Forecast Overview: Industrial Electronics and Semiconductors, Worldwide, 2017 Update.”

Gartner. (2017). “Hype Cycle for Artificial Intelligence, 2017.”

Gartner. (2017). “Market Trends: AI in Edge Devices Create Opportunities for Device Manufacturers.”

Gartner. (2017). “Market Trends: Application Processor Vendors Need Tailored Strategies for Fragmented IoT Applications.”

Gartner. (2017). “SWOT: Xilinx, PLD/FPGA Market, Worldwide.”

IDC. (2017). “GPUs, FPGAs, ASICs, or Many-Core Processors_ Which Acceleration Technology Do Data centers Need.”

IDC. (2017). “IDC's Worldwide Accelerated Compute Taxonomy, 2017.”

IDC. (2017). “Mapping the Future of Silicon for AI.”

IDC. (2017). “Worldwide Accelerated Server Infrastructure Forecast,

2017-2021.”

IDG. (2017). “구글 텐서플로우부터 MS CNTK까지 딥러닝/머신러닝 프레임워크 6종 비교 분석.”

KIET. (2017). “정책과 이슈 : 정책과 이슈 한국 반도체산업의 4.0시대 전략.”

NH투자증권 리서치센터. “반도체 구조 변화.”

STEPI. (2016), “신기술 발전에 따른 산업 지형의 변화 전망과 대응 전략.”

Technavio. (2017). “GLOBAL ARTIFICIAL INTELLIGENCE CHIPS MARKET 2017-2021.”

디지에코. (2017). “4차 산업혁명을 이끄는 인공지능 - 딥러닝을 중심으로”

웹사이트

http://biz.chosun.com/site/data/html_dir/2017/11/07/2017110702183.html , 최종접속 : 2017.11.27.

http://blogs.nvidia.co.kr/2017/07/17/dli-self-paced-lab/, 최종접속 : 2017.11.27.

http://v.media.daum.net/v/20171025092548315 , 최종접속 : 2017.11.27.

http://www.ciokorea.com/print/35017 , 최종접속 : 2017.11.27.

http://www.cnbc.com/2017/07/24/microsoft-creating-ai-chip-for-hololens.html, 최종접속 : 2017.11.27.

http://www.etnews.com/20171020000044?mc=ns_003_00003 , 최종접속 : 2017.11.27.

http://www.hellot.net/new_hellot/magazine/magazine_read.html?code=202&sub=004&idx=29474 , 최종접속 : 2017.11.27.

http://www.zdnet.co.kr/news/news_view.asp?artice_id=20170802154810 , 최종접속 : 2017.11.27.

https://basicmi.github.io/Deep-Learning-Processor-List/#Horizon_Robotics , 최종접속 : 2017.11.27.

https://byline.network/2017/07/24-2/ , 최종접속 : 2017.11.27.

https://developer.qualcomm.com/software/snapdragon-neural-processing-engine , 최종접속 : 2017.11.27.

https://docs.microsoft.com/en-us/cognitive-toolkit/reasons-to-switch-from-

tensorflow-to-cntk , 최종접속 : 2017.11.27.

https://www.nanalyze.com/2017/05/12-ai-hardware-startups-new-ai-chips/, 최종접속 : 2017.11.27.

https://www.nanalyze.com/2017/05/12-ai-hardware-startups-new-ai-chips/, 최종접속 : 2017.11.27.

https://www.qualcomm.com/news/onq/2017/08/16/we-are-making-device-ai-ubiquitous , 최종접속 : 2017.11.27.

https://www.theverge.com/2017/5/29/15707606/arm-cortex-a75-a55-mali-g72-specs-announced , 최종접속 : 2017.11.27.

https://www.wired.com/2017/05/google-rattles-tech-world-new-ai-chip/ , 최종접속 : 2017.11.27.

https://www.slideshare.net/anirudhkoul/squeezing-deep-learning-into-mobile-phones 최종접속 : 2017.11.27.

저자소개

최 새 솔 ETRI 미래전략연구소 기술경제연구본부 선임연구원
e-mail: saesol.choi@etri.re.kr Tel. 042-860-1803

인공지능 반도체 산업동향 및 이슈 분석

발 행 인 : 한 성 수

발 행 처 : 한국전자통신연구원 미래전략연구소 기술경제연구본부

발 행 일 : 2017년 12월

ETRI 한국전자통신연구원
미 래 전 략 연 구 소

34129 대전광역시 유성구 가정로 218
전화 : (042) 860-3874, 팩스 : (042) 860-6504

ETRI 미래전략연구소 www.etri.re.kr