ETRI Insight

기술정책 트렌드 2022-11

소프트웨어 관점의 인공지능 반도체 경쟁력

이선재

ETRI 한국전자통신연구원
Electronics and Telecommunications Research Institute

본 보고서는 ETRI 기술정책연구본부 기본사업인
“국가 지능화 기술정책 및 표준화 연구”를 통해 작성된 결과물입니다.

목 차 C O N T E N T S

핵심 요약 ····· i

Ⅰ. 연구 개요 ····· 1

1. 기술의 범위 및 정의 ····· 1

2. 인공지능 반도체의 진화 및 기술유형 ····· 3

Ⅱ. 인공지능 반도체 산업 동향 ····· 5

1. 인공지능 반도체 생태계 및 산업구조 ····· 5

2. 인공지능 반도체 기업 동향 ····· 6

3. 생태계 확장 전략과 소프트웨어의 중요성 ····· 7

Ⅲ. 인공지능 반도체 기업의 경쟁력 분석 ····· 11

1. 해외 반도체 기업의 소프트웨어 강화 및 풀스택 전략 ····· 11

2. 국내 기업의 풀스택 전략 ····· 17

Ⅳ. 국내 인공지능 반도체 기업의 경쟁력 강화 방안 ····· 19

핵심 요약

인공지능 반도체 활용 전략 및 산업 동향

- (인공지능 반도체의 활용 목적) 기업이 인공지능 반도체를 사용하는 최종 목적은 고객을 만족시킬 수 있는 성능이 우수한 AI 서비스를 제공하기 위함에 있음
- (인공지능 반도체의 활용 전략) 기업의 역량 및 서비스 목적에 따라 호환성/안전성 측면에서 이미 검증된 기존 제품을 활용하거나 하드웨어(칩)을 직접 제작(설계)하기도 함
- (개발 동향) 인공지능 기업들의 요구사항을 충족시키기 위해 기존의 반도체 기업(팹리스 포함)은 칩의 성능을 향상시켜 왔으며, 빅테크 기업들은 자사 알고리즘을 최적으로 구현하기 위한 칩을 개발하여 적용하기도 함
  - 반도체 및 팹리스 기업은 기존의 하드웨어 중심의 비즈니스 모델에서 하드웨어를 기반으로 한 AI 서비스까지 제공
  - 빅테크 기업은 인공지능 기반 서비스 구현을 위한 반도체를 직접 설계하고 디바이스까지 출시함
- (경계의 소멸) 인공지능 반도체 관련 기업들의 개발 동향 및 전략 발표를 종합하여 살펴보면, 반도체 기업(팹리스 포함)부터 빅테크 기업까지 풀스택을 갖추고 인공지능 반도체를 성능을 효과적으로 구현하기 위한 생태계를 조성하기 노력중
  - 풀스택은 SW와 HW 전반에 대한 이해도가 높은 개발자를 주로 뜻하는 용어로 사용하지만, 본 연구에서는 하드웨어부터 소프트웨어까지 모든 단계의 제품 및 서비스를 제공하는 형태를 의미함
  - 인공지능 반도체 기업들이 확장의 시작점(하드웨어에서 출발, 서비스에서 출발)은 다르지만 풀스택을 확보하고자 하는 목표는 동일함

연구방법 및 필요성

- (기존 연구와의 차별성) 기존 연구들은 인공지능 반도체 하드웨어의 성능과 개발 동향에만 집중
- (연구 방향성) 인공지능 반도체 선도 기업의 생태계 확장 전략을 살펴보고 국내 인공지능 반도체 경쟁력 강화 및 생태계 활성화 방안을 연구하고자 함
  - AI 지원의 사실상 표준 컴퓨팅 플랫폼으로 자리 잡은 엔비디아, ‘IDM2.0’ 전략 발표를 통해 제조, HW, SW를 모두 아우르려는 인텔, 모바일의 대표 기업 퀄컴의 AI풀스택, 구글의 하드웨어 시장 진출 전략 등을 분석하고 경쟁력 확보를 위한 생태계 확장 전략을 벤치마킹 하고자 함
  - 국내 AI반도체의 경쟁력 확보 전략을 SKT 및 KT 사례를 통해 알아보고 국내 인공지능 반도체 경쟁

력 강화를 위한 전략을 마련하고자 함

소프트웨어 관점의 인공지능 반도체 경쟁력 분석 및 기업 경쟁력 강화 방안

- (소프트웨어 관점의 분석 필요성) 인공지능 서비스는 하드웨어+시스템 SW+응용 SW의 최적화일 때 성능이 극대화하기 때문에 하드웨어 뿐만 아니라 이를 운영하고 응용 소프트웨어를 지원하는 시스템 소프트웨어, AI 모델, 서비스까지 모두 중요함

- (AI 칩 성능의 필요조건) 연산 능력이 뛰어난 하드웨어와 이를 운용하는 소프트웨어에 따라 성능이 결정됨

  - 대규모 학습과 추론 처리 성능을 극대화하기 위해서는 인공지능 프로세서의 자체 연산 성능도 중요하지만, 이를 운용하기 위한 인공지능 컴파일러의 최적화도 함께 요구

- (국내 팹리스의 한계) 국내 팹리스 기업들이 개발한 칩이 별도의 시스템 소프트웨어로 운영된다면 비용 및 시간 측면에서 매우 불리하기 때문에 각 기업이 개발한 칩의 성능을 최적화할 수 있는 통합된 시스템 소프트웨어가 있다면, 이를 기반으로 팹리스는 칩 설계에 집중할 수 있으며 다양한 사업으로 확장할 가능성이 높아짐

- 소프트웨어 관점의 인공지능 반도체 산업 경쟁력 강화를 위한 제언

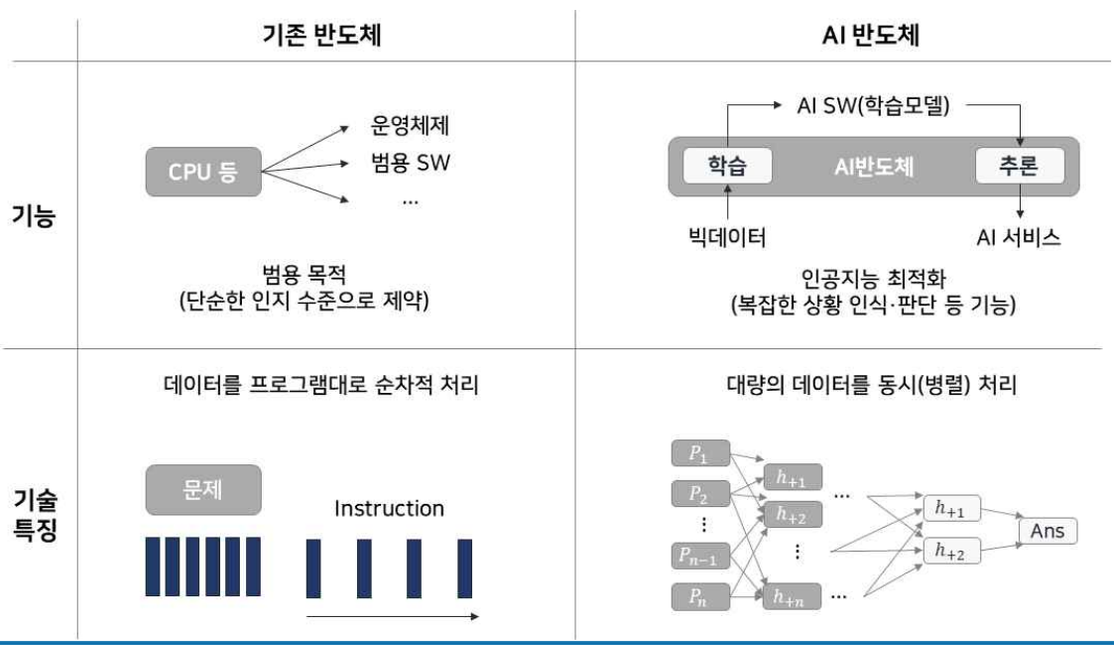
  - 팹리스 기업이 성장하기 위해서는 국가연구기관에서 통합 시스템 소프트웨어를 개발하고 오픈소스로 공개하여 중소 팹리스에 적용 필요

Ⅰ 개요

1 인공지능 반도체의 정의와 범위

- (정의) “학습·추론 등 인공지능 서비스 구현에 필요한 대규모 연산을 높은 성능, 높은 전력효율로 실행하는 반도체”[1)]
  - (예시) 인공지능 반도체가 인간의 두뇌처럼 학습을 하면서 아이에서 성인수준으로 인지능력이 향상될 때, 이를 매우 짧은 시간에 효율적(낮은 전력)으로 구현
  - 좁은 의미의 지능형 반도체는 인식·추론·학습·판단 등 인공지능 서비스에 최적화 (지능화, 저전력화, 안정화)된 소프트웨어와 시스템 반도체가 융합된 반도체임
  - 넓은 의미의 지능형 반도체는 개인형 인공지능 디바이스, 자율이동체, 지능형 헬스 케어, 빅데이터 처리, 차세대 통신 서비스, 스마트 시티, 증강현실, 인간형 지능로봇, 신약개발, 에너지 등 사회, 경제, 국방 등 전 분야에 응용되는 지능형 반도체들을 포괄하는 기술로 정의할 수 있음

그림 1 인공지능 반도체 특징



* 출처: 인공지능 반도체 산업 발전전략(관계부처 합동, 2020) 재인용

1) 인공지능 반도체 산업 성장 발전전략(관계부처 합동, 2020)

- (활용 목적) 인공지능 반도체는 인공지능 시스템의 구현 목적에 따라 크게 학습용과 추론용으로 구분할 수 있으며, 두 가지 과정을 반복 실행하여 최적의 답을 찾도록 성능을 강화하는데 주로 사용
  - (학습용) 딥 러닝 등 기계 학습의 특정 작업을 수행하기 위해 방대한 데이터를 통해 반복적으로 지식을 배우는 단계
  - (추론용) 학습을 거친 최적의 모델을 통해 외부 명령을 받거나 상황을 인식하면 학습한 내용을 토대로 가장 적합한 결과를 도출하는 단계
- (사용 환경) 기존 인공지능 시스템은 주로 데이터센터에서 학습과 추론을 병행하여 사용되었으나, 스마트폰 및 IoT 등의 보급 확산, 클라우드 기술 발전과 동시에 디바이스의 추론 기능의 수요가 증가하면서 이를 수행하기 위한 반도체 기술 중심으로 발전
  - (데이터 센터용) 현재 인공지능 학습/추론은 대부분 데이터센터에서 실행되며 일반적인 하드웨어로는 CPU가 담당하고 있지만, 인공지능 서비스에 요구되는 대규모 연산 처리 성능을 위해 인공지능 반도체를 서버에 장착하여 활용
  - 데이터센터 전용 반도체는 방대한 데이터를 처리하기 때문에 발열과 전력 소모로 인한 효율성 개선이 지속적으로 필요
  - (엣지 디바이스용) 데이터센터 서버(클라우드)와 연결을 최소화하고 디바이스 자체에서 인공지능 연산이 수행되는 경우가 점차 확대되면서 소형화·저전력·고성능 중심의 인공지능 반도체 기술 개발이 가속화
  - 온 디바이스 AI는 네트워크 의존도를 낮추고 클라우드보다 빠르며 생체 인식 등 민감한 데이터를 안전하게 활용할 수 있음

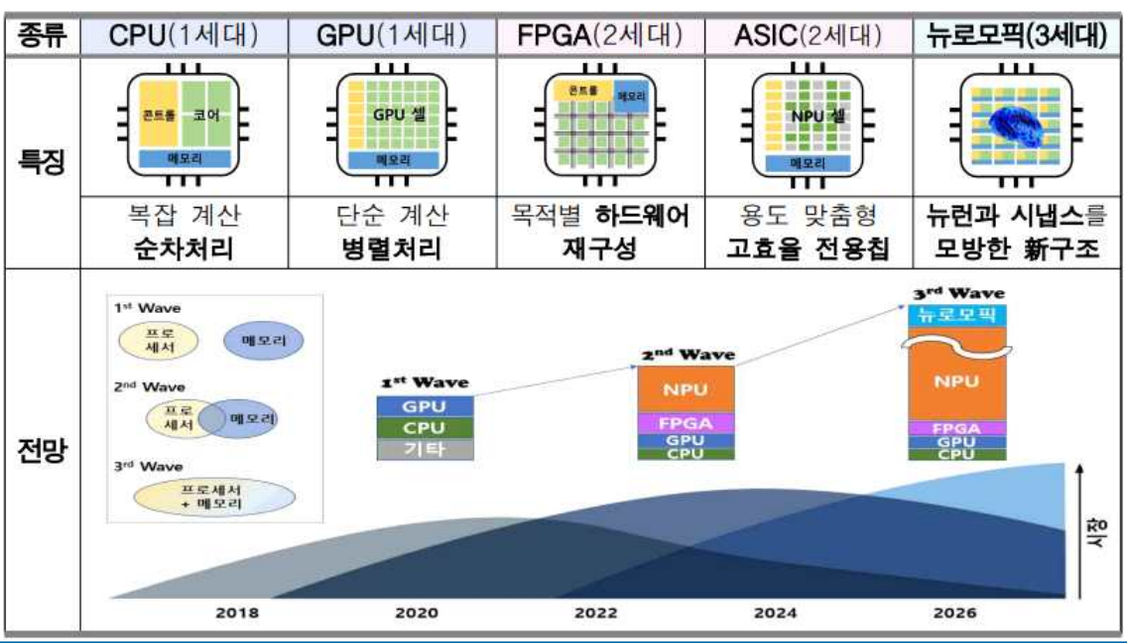
그림 2 사용목적에 따른 인공지능 반도체

* 출처: 엔비디아

2 인공지능 반도체의 진화 및 기술유형

- 인공지능 반도체는 (1세대)CPU+GPU 등 ⇒ (2세대)NPU ⇒ (3세대)뉴로모픽으로 발전하고 있으며, 또한 메모리(기억)·프로세서(연산) 통합으로 미래 반도체 설계 패러다임을 완전히 바꿀 수 있는 신개념 PIM(Processing-In-Memory)기술도 발전([그림 3], [그림 4] 참조)

그림 3 인공지능 반도체 기술 진화

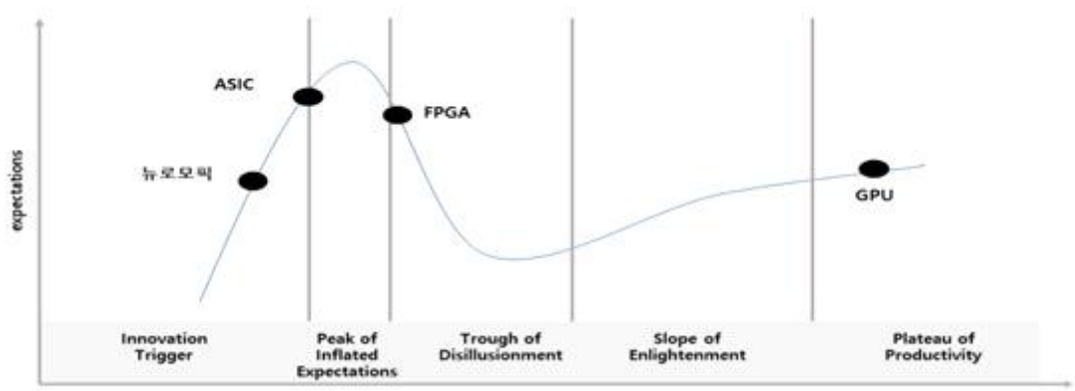


* 출처: 인공지능(AI) 반도체의 산업경쟁력(특허청, 2022)

- (1세대 AI반도체) CPU를 포함하여 현재 인공지능 시장을 지배하는 있는 GPU가 이에 해당하며 해당 분야의 선도기업으로 인텔 · 엔비디아 등이 있음
  - 호환성 높은 것이 장점이나, 인공지능 연산 성능(Performance)과 소비전력 효율(Power Efficiency)이 낮다는 것이 단점
- (2세대 AI반도체) 인공지능 연산 고속화를 위해 반도체 구성을 최적화시킨 ASIC/ASSP가 이에 해당되며, 다양한 팹리스 스타트업이 NPU를 상용화하는 시점으로써 시장을 선점하기 위해 노력하고 있음
  - NPU는 인공신경망 구조에 특화되어 GPU 대비 저렴하고 전력효율·발열 성능이 우수하여 GPU와 경쟁이 시작되는 단계
  - 스타트업(그래프코어(英) 등) 중심으로 2세대 NPU 개발이 활발, 빅테크 기업은 자체 칩 개발·활용 단계(구글 데이터센터 전용 TPU, 테슬라 전기차용 칩 D1, 애플 M2 등)

- (3세대 AI반도체) 인간의 뇌를 모방한 非폰노이만 방식의 뉴로모픽 반도체가 현재까지는 가장 진보된 형태의 AI 반도체로 평가받고 있으며 학계·선도기업 중심으로 기초연구·기술력 축적중
  - 인공신경망 연산 성능과 소비전력 효율은 AI 반도체 가운데 가장 뛰어나지만, 아직은 기술성숙도가 낮고 폰 노이만 구조를 사용하지 않기 때문에 범용성이 낮은 것이 단점
  - (인텔) 연구용 칩 '로이히' 발표('21) / (KAIST) 뉴로모픽 소자 후보기술인 멤리스터 연구('22)
- 인공지능 반도체는 기술성숙도, 사용환경 및 기능에 따라 기술 간 서로 대체되어 사용되기도 하나 공통적으로 초고성능·초저전력 중심으로 진화할 전망

그림 4 인공지능 반도체 Hyper Cycle



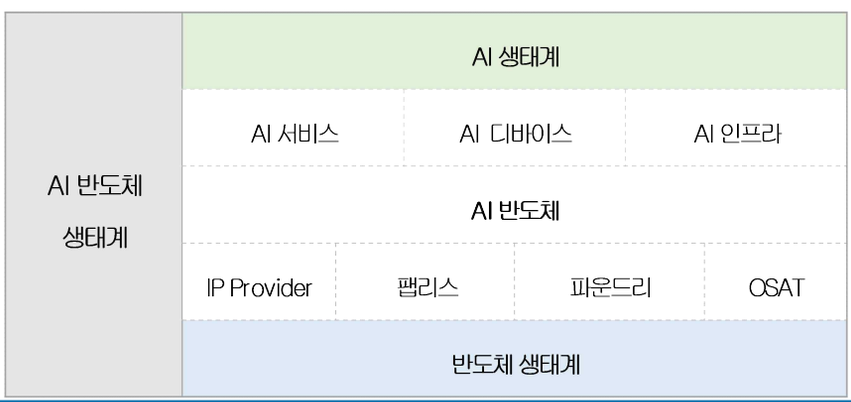
* 출처: Gartner(2019)

Ⅱ 인공지능 반도체 산업 동향

1 인공지능 반도체 생태계 및 산업구조

- 좁은 의미의 인공지능 반도체 생태계는 특정 목적에 맞는 하드웨어를 개발하는 시스템 반도체의 일부이며 넓은 의미로는 기존 시스템 반도체 생태계와 인공지능 생태계가 융합된 새로운 산업구조로 바라볼 수 있음

그림 5 인공지능 반도체 생태계



* 출처: 인공지능 반도체 산업 발전전략(관계부처 합동, 2020), 저자 재작성

- (반도체 생태계 관점) 인공지능 반도체는 시스템 반도체의 부분 집합으로써 시스템 반도체 산업구조를 따르며 크게 종합 반도체 회사인 IDM, IP 기업, 팹리스, 파운드리로 분류를 나누어 볼 수 있음
  - (종합 반도체 기업(IDM)) 회로 설계부터 판매까지 전 과정 총괄하는 기업으로 삼성전자, 인텔 등이 있음
  - (IP 기업) 셀 라이브러리 및 특정 설계 블록을 제공하고 IP 사용에 따른 라이센스로 수익을 내는 기업으로 ARM 등이 있음
  - (팹리스) Fab(반도체 생산설비)을 소유하지 않고 회로설계와 판매만 담당하는 설계 전문 업체로써 엔비디아, 퀄컴 등이 있음
  - (파운드리) 위탁받은 설계대로 반도체 생산을 담당하는 수탁 제조 업체로써 삼성전자, TSMC 등이 있음
- (인공지능 생태계 관점) 인공지능 반도체 산업과 관련된 인공지능 생태계는 인공지능 칩을 기반으로 AI 제품 및 서비스를 제공하는 기업, AI 디바이스 제조사, 클라우드와 같은 AI 인프라 기업으로 구성되며 주요 기업으로는 구글, 애플, 메타, 테슬라, 마이크로 소프트, 아마존 등이 있음

2 인공지능 반도체 기업 동향

- 반도체와 인공지능 생태계가 융합된 특성으로 인해 주요 기업들은 주요 사업영역을 기반으로 융합시장 대상으로 확장 방향성을 가짐

- (기존 반도체 기업) 엔비디아, 인텔, 퀄컴 등의 반도체 기업들의 칩은 지속적으로 성능을 향상시키면서 호환성을 장점으로 다양한 산업 영역에서 활용되고 있으며 최근에는 데이터센터용 인공지능 반도체 개발에 초점을 맞추고 있음

- (AI 서비스 기업) 초기에는 호환성 측면에서 엔비디아, 인텔, 퀄컴 등의 반도체 기업들의 제품을 활용하였으나, 각 기업의 AI 서비스는 고도의 AI 알고리즘을 수행하는데 최적화된 반도체가 뒷받침 되어야하기 때문에 자체 칩을 제작하여 적용하고 있음

- 그 외에도 한정된 응용 분야에 특화되지 않고 다양한 용도로 사용할 수 있는 가속기를 개발해 엔비디아에 대항하려는 여러 팹리스 스타트업들도 등장함

  - 해외 주요 기업으로는 세레브라스, 그래프코어, 그로크, 하일로 등이 있으며 국내 기업은 퓨리오사, 리벨리온, 딥엑스 등이 인공지능 반도체 시장을 선점하기 위해 치열하게 경쟁하고 있음

- (생태계 변화-경계의 소멸) 인공지능 반도체 관련 기업들의 개발 동향 및 전략 발표를 종합하여 살펴보면, 반도체 기업(팹리스 포함)부터 빅테크 기업까지 풀스택을 갖추고 인공지능 반도체를 성능을 효과적으로 구현하기 위한 생태계를 조성하기 노력 중

- (풀스택이란) SW와 HW 전반에 대한 이해도가 높은 개발자를 주로 뜻하는 용어로 사용하지만, 본 연구에서는 하드웨어부터 소프트웨어까지 모든 단계의 제품 및 서비스를 제공하는 형태를 의미함

  - 인공지능 반도체 기업들이 확장의 시작점(하드웨어에서 출발, 서비스에서 출발)은 다르지만 풀스택을 확보하고자 하는 목표는 동일함

- (풀스택의 배경) 인공지능 반도체, 머신러닝을 기반으로 인공지능 성능의 진화가 유례없이 빨라지고, 5G 네트워크를 통해 인프라부터 엣지까지 데이터가 이동하면서 특정 기술 카테고리에 집중하는 전략만으로는 전체 구조를 조망하고 적재적소에 맞는 솔루션을 제공하기 힘들어짐

  - 엣지부터 클라우드, 인프라까지 모든 곳에서 동시에 구동되고 서로 상호작용하기 때문에 이를 지원하는 컴퓨팅 또한 클라우드, 엣지, IoT 및 자동차, 고성능컴퓨팅, 마이크로프로세서, 가속 컴퓨팅, SW 등 모든 분야를 커버해야 함

3 생태계 확장 전략과 소프트웨어의 중요성

- 엔비디아나 인텔과 같은 해외 반도체 기업은 반도체 하드웨어 성능 기반의 생태계를 유지하면서 소프트웨어를 강화하는 전략으로 확장하고 있으며 후발주자인 국내 기업은 자체 수직계열화나 연합모델을 통한 전략으로 생태계를 확장하고 있음
- 인공지능의 성능은 하드웨어+시스템 SW+응용 SW의 최적화일 때 극대화하기 때문에 하드웨어뿐만 아니라 이를 운영하고 응용 소프트웨어를 지원하는 시스템 소프트웨어, AI 모델, 서비스까지 모두 중요함
- 인공지능 반도체의 SW 기술 스택은 크게 인공지능 반도체를 구동할 수 있는 시스템 소프트웨어와 AI 서비스를 위한 응용 소프트웨어로 구분함

그림 6 AI반도체 Full Stack

AI SW	AI 서비스 및 앱
	AI 플랫폼
	AI frameworks, AI 라이브러리
	시스템 SW (AI 컴파일러 등)
AI HW	AI 반도체

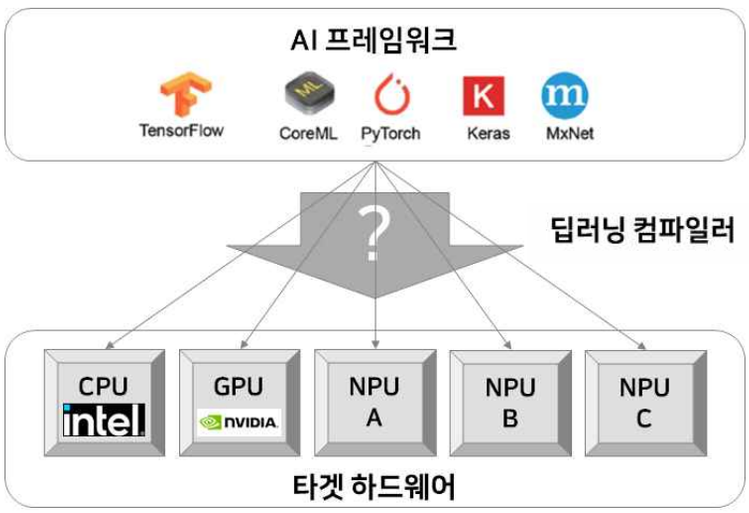
* 출처: 저자작성

- (서비스 및 앱) 비전 처리, 고객 지원 챗봇, 인공지능 비서, 알고리즘 주식 거래 등 AI 지원 SW 애플리케이션
- (AI 플랫폼) AI 애플리케이션을 쉽게 구현할 수 있도록 지원하기 위해 다양한 AI 기능을 조합한 플랫폼
- (AI frameworks) 머신러닝(ML) 알고리즘을 사용하여 특정 애플리케이션을 위한 딥러닝 모델을 설계, 구현, 훈련시키는 도구 및 프레임워크로써 오픈소스 형태로 광범위하게 제공
- (AI 라이브러리) 특정 반도체 칩셋에 AI 프레임워크(AI 모델 및 알고리즘)를 구동시키고 최적화하기 위한 낮은 수준의 SW 함수(low-level software functions)
- (AI 컴파일러) 딥러닝 모델을 특정 디바이스 상에서 효율적으로 동작시키기 위해서는 동작시킬 딥러닝 모델을 타겟 디바이스에서 최적의 속도와 정확도를 낼 수 있는 머신 코드로 자동 변환 작업을 지원해주는 도구
- (AI 하드웨어) AI 작업 및 연산을 수행을 가속화하기 위해 설계된 최적화된 프로세서 유닛

및 반도체 로직 회로

- (시스템 소프트웨어의 중요성) 같이 대규모 학습과 추론 처리 성능을 극대화하기 위해서는 인공지능 프로세서의 자체 연산 성능도 중요하지만, 이를 운용하기 위한 인공지능 컴파일러의 최적화도 함께 요구됨
  - (시스템 소프트웨어) 컴퓨터 시스템의 개별 하드웨어(HW) 요소들을 직접 제어, 통합, 관리하는 소프트웨어로서 사용자가 컴퓨터 하드웨어의 물리적인 특성이나 구조를 전부 알지 못하더라도 컴퓨터 시스템을 사용할 수 있게 도와주는 역할을 하는 소프트웨어
  - (컴파일러) 고급 프로그래밍 언어로 작성된 원시 코드를 컴퓨터 내부에서 사용가능한 언어인 기계어로 번역하고, 실행을 가능하게 해주는 프로그램
- (딥러닝 컴파일러) 딥러닝 모델을 특정 디바이스 상에서 효율적으로 동작시키기 위해서는, 동작시킬 딥러닝 모델을, 타겟 디바이스에서 최적의 속도와 정확도를 낼 수 있는 머신 코드로 변환해야 하며 이러한 코드 변환 작업을 자동으로 지원해주는 도구([그림 15] 참고)
  - 딥러닝 컴파일러는, 다양한 딥러닝 플랫폼에서 학습된 딥러닝 모델을 입력으로 받아, 특정 하드웨어에서 동작 가능한 머신 코드(또는, 백엔드 코드)를 자동으로 생성함
  - 예를 들어 XLA, TVM, Glow와 같은 딥러닝 컴파일러들은 TensorFlow, Pytorch, MxNet, ONNX 등의 프레임워크로 작성된 모델을 입력으로 하여, CPU 및 GPU 용 백엔드 코드를 생성함
- 인공지능 반도체 기업들의 생태계 확장 전략에서 중요한 요소는 하드웨어의 성능 향상은 기본이며, 인공지능 반도체는 인공지능 모델과 컴파일러 환경이 없으면 무용지물이기 때문에 SW 플랫폼이 가장 중요함
  - (엔비디아 사례) 엔비디아의 CEO 젠슨 황은 엔비디아는 반도체 회사가 아니고 소프트웨어 회사라고 언급했으며 이는 엔비디아 GPU는 '쿠다(CUDA)'라는 플랫폼을 갖고 있기 때문에 가치가 있기 때문임
  - (구글 사례) 구글 반도체의 성능 때문에 구글의 텐서플로(TensorFlow) 사용하는 것이 아니라 SW 컴파일러 환경과 그에 따른 사용자가 많기 때문에 사용함
  - 다만 프로그램 개발자가 항상 사용하면서도 너무나 자연스럽고 쉽게 사용하고 있기 때문에 반도체 개발자나 SW모델 개발자는 사실 컴파일러 존재 및 중요성을 인식하지 못함
- 따라서 인공지능 반도체 기업들은 [표 1]과 같이 자사의 칩에 맞는 시스템 소프트웨어를 개발하고 사용자에게 활용하도록 함으로써 소프트웨어 개발자들이 자사의 생태계에 묶어두는 락인효과를 기반으로 생태계 확장 및 활성화 효과를 기대함

그림 7 AI반도체를 위한 시스템 SW



* 출처: 저자작성

표 1 해외 AI 반도체 기업의 시스템 소프트웨어

기업명	엔비디아	인텔	AMD	구글	메타
시스템 소프트웨어	NVIDIA CUDA	intel oneAPI BASE TOOLKIT	AMD ROCm	XLA	Glow
AI 하드웨어 (주요 제품)	GPU (H100)	GPU (ARC A350M)	GPU (Radeon RX 7900)	TPU V4	오큘러스용 칩

* 출처: 저자작성

- (락인효과) 특정 재화 혹은 서비스를 한 번 이용하면 다른 재화 혹은 서비스를 소비하기 어려워져 기존의 것을 계속 이용하는 효과 혹은 현상으로써 현재 이용하고 있는 특정 재화 또는 서비스가 다른 재화 혹은 서비스의 선택을 제한하여 기존에 이용하던 것을 계속 선택하게 되는 현상을 말함

- (락인효과 비즈니스 사례) 마이크로소프트 사의 운영 체제인 윈도우를 이용하는 소비자들이 대부분 동일 회사에서 개발한 웹브라우저를 이용하는 현상, 이동통신사들이

유선통신(초고속 인터넷), IPTV(인터넷TV) 상품까지 포함된 결합상품을 판매하고 이 중 하나를 타사 서비스의 것으로 소비할 경우, 더 높은 비용을 부과하여 소비자들을 묶어 두려는 현상 등이 있음

- (락인효과 기술사례) 한 산업에서 어떠한 기술이 표준으로 자리 잡으면 다른 기술들이 그 기술을 기반으로 하여 발전하는 현상이 관찰되는데, 이 경우 해당 표준 기술은 새로운 기술의 등장을 저해할 수 있음
  - 표준 기술의 확산으로 소비자들이 해당 기술에 충분히 익숙해져 있다면, 새로운 기술을 개발하는 기업의 입장에서는 기존 기술을 이용하는 소비자들의 저항에 부딪히며 이 경우 기업들은 새로운 기술 개발에 적극적이지 않을 수 있음

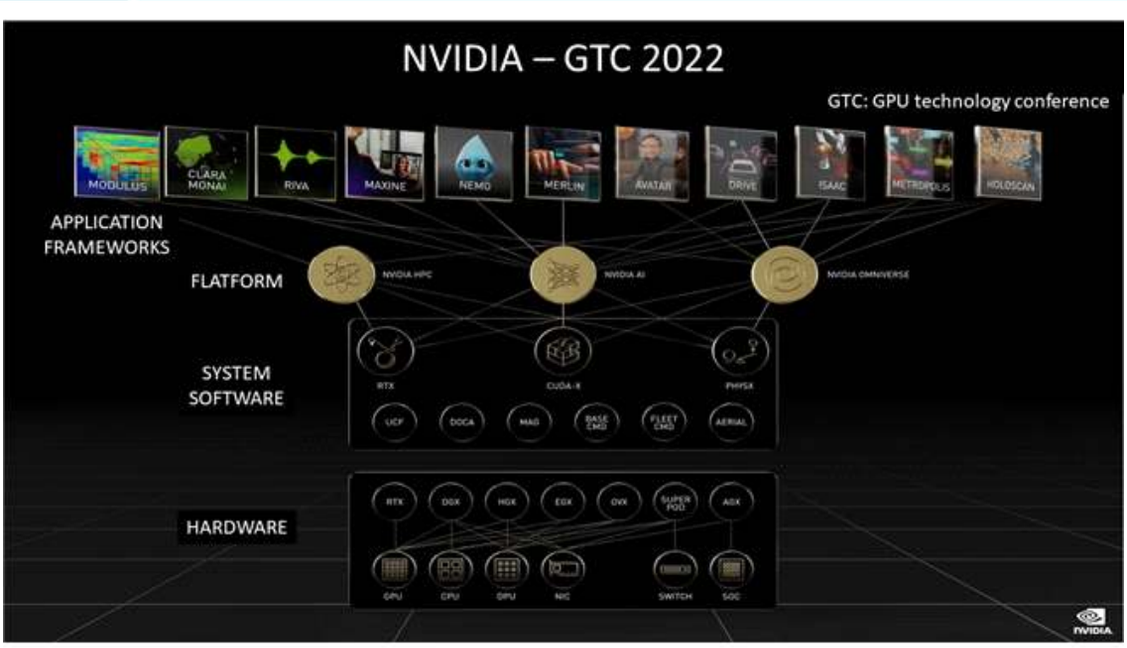
Ⅲ 인공지능 반도체 기업의 경쟁력 분석

1 해외 반도체 기업의 소프트웨어 강화 및 풀스택 전략

가. 엔비디아

- (확장 전략) 엔비디아는 하드웨어인 GPU의 성능 개선 및 데이터센터용 CPU도 개발하고 있으며, GPU 및 시스템 소프트웨어를 기반으로 다양한 AI 서비스를 제공하고 있음
  - (하드웨어) AI 처리를 위한 GPU의 성능 향상을 위해 2018년 GPU에 Tensor core 적용하였으며, GTC 2022에서 공개한 H100 GPU는 대규모의 통합 GPU 클러스터까지 확장 가능한 환경을 제공하기 때문에 차세대 엑사스케일 고성능 컴퓨팅(HPC)과 매개 변수가 조 단위인 AI를 이용할 수 있음
  - (AI 서비스) GPU의 호환성을 기반으로 건축, 엔지니어링 및 건설, 소비자 인터넷, 사이버 보안, 에너지, 금융 서비스, 게임 개발, 헬스케어 및 생명 과학, 고등 교육 및 연구, 제조, 미디어 및 엔터테인먼트, 공공 부문, 레스토랑, 소매, 스마트 시티, 슈퍼컴퓨팅, 통신, 운송 등과 같이 대부분의 산업 분야에 활용되고 있으며 엔비디아가 제공하는 플랫폼을 활용하여 다양한 응용 서비스를 개발할 수 있는 환경을 제공함

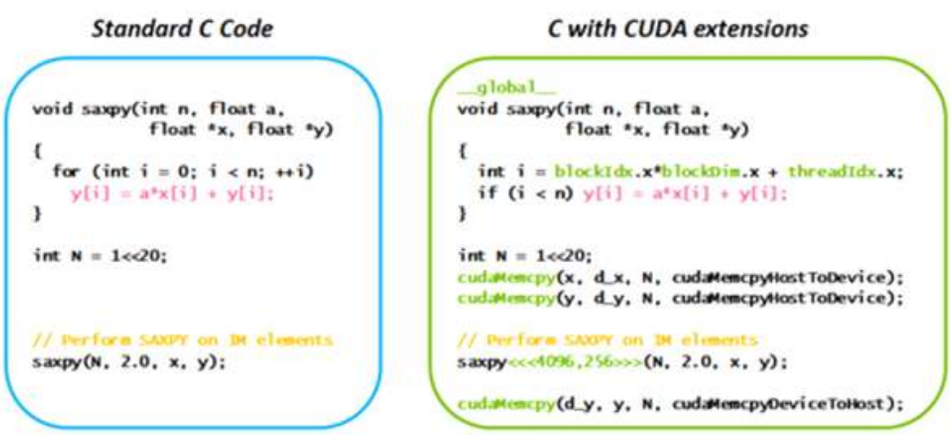
그림 8 엔비디아 풀스택



* 출처: 엔비디아

- (CUDA 기반 생태계) CUDA(Computed Unified Device Architecture)는 GPU 컴퓨팅에서 일종의 컴파일러 역할을 수행하는 도구로써 엔비디아 GPU가 인공지능(AI) 혁명을 견인한 원동력이자 고성능 컴퓨팅(HPC) 분야의 필수적인 솔루션으로 빠르게 확산된 요인은 CUDA에 있다고도 할 수 있음
  - 엔비디아 GPU는 대규모 데이터를 효과적으로 처리할 수 있는 다중 연산 및 초고속 병렬 연산 능력에도 불구하고, 제한적인 범위에서만 활용됨
  - 그래픽스 처리 장치를 이용한 범용 프로그래밍(General Purpose GPU, GPGPU)이 고급 그래픽 프로그래밍 기술 범주에 속했고, 그래픽 애플리케이션 프로그래밍 인터페이스(Application Programming Interface, API)에 익숙하지 않은 일반 개발자들은 GPU를 이용하기가 어려움
  - CUDA가 GPU에서 수행하는 병렬 처리 알고리즘을 C언어 등을 비롯한 산업 표준언어를 이용해 작성할 수 있도록 함으로써 CUDA를 통해 그래픽 API를 알지 못하는 개발자들도 GPU를 활용할 수 있게 됨
  - 딥러닝 기술이 전 산업 분야에서 주목 받으면서 다양한 딥러닝 알고리즘들을 사용하기 쉽게 해주는 여러 오픈 소스 소프트웨어(OSS)들이 개발, 상용화되고 있으며, 대부분의 소프트웨어는 엔비디아가 제공하는 CUDA 라이브러리를 기반으로 하고 있기 때문에 딥러닝 연구진 및 프레임워크 개발자들이 엔비디아 GPU를 선호함

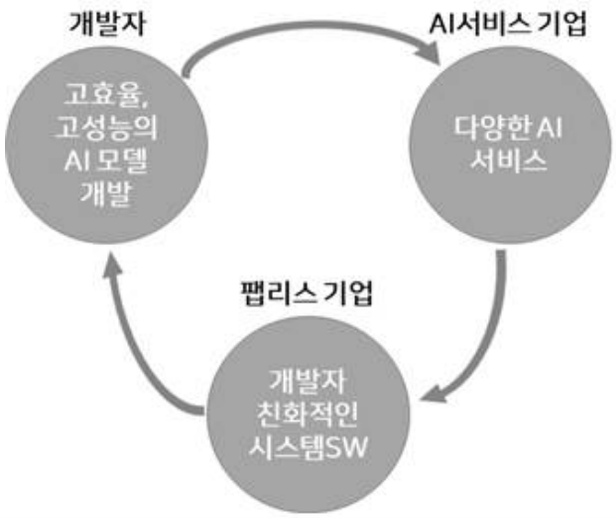
그림 9 엔비디아 CUDA



* 출처: 엔비디아

- (CUDA 기반 생태계 확장 전략) AI 지원의 사실상 표준 컴퓨팅 플랫폼으로써 개발자에게 CUDA 생태계에 속할 수 밖에 없는 환경을 제공
  - 가장 빠르게 새로운 딥러닝 모델을 지원하면서 모델 개발자가 자발적으로 개발하고 관련된 SW 개발자가 다시 참여하는 선순환 구조
  - 라이브러리, 도메인 플랫폼, SW 플랫폼과 같은 서비스 지원을 위한 SW를 개발하여 제공

그림 10 엔비디아 선순환 생태계



* 출처: 저자작성

나. 인텔

- (확장 전략) 인텔은 하드웨어 위에 구동할 소프트웨어 역량 강화 추진 전략으로써 슈퍼컴퓨팅을 구축하고 서비스형 소프트웨어(SaaS, Software as a Services)를 제공함[2)]
  - 고성능 하드웨어를 잘 만드는 것 이상으로 이를 잘 활용할 수 있도록 하는 소프트웨어까지도 제공해 고객들이 복잡한 환경에서 디지털 전환을 가속화할 수 있도록 하겠다는 전략
  - (하드웨어) 12세대 인텔 코어 HX 프로세서, 데이터센터 워크로드용 트레이닝을 위한 인텔 하바나 가우디2 AI 프로세서, 데이터센터용 GPU(코드명 아틱사운드 M), 4세대 인텔 제온 스케일러블 프로세서(코드명 사파이어 래피즈), 인텔 인프라 처리장치(IPU) 로드맵의 상세 내용을 공개
  - (AI 서비스) 앤드게임 프로젝트는 클라우드를 통해 네트워크 내 다른 장치에서 사용 가능한 컴퓨팅 리소스로 사용자가 활용할 수 있으며, 액센추어와의 아폴로 프로젝트는 기업을 위한 30개가 넘는 오픈소스 인공지능 레퍼런스 키트를 제공하기 때문에 클라우드와 온프레미스, 엣지에서의 인공지능의 접근성을 높임

2) 인텔비전 2022에서 공개

그림 11 인텔 oneAPI

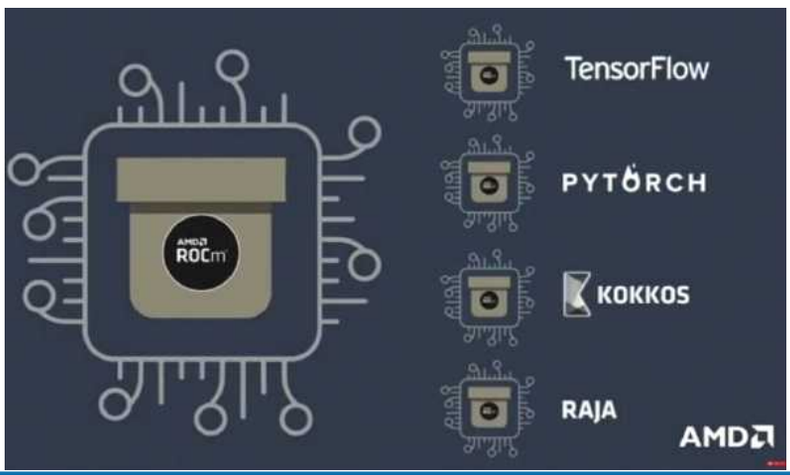
* 출처: 인텔

- (oneAPI) 인텔은 지난 2020년 12월에 oneAPI 툴킷의 출시를 발표했으며 oneAPI 툴킷을 이용하여 개발자는 CPU, GPU, FPGA(통칭 XPU)를 활용한 고성능 교차 아키텍처 애플리케이션을 개발할 수 있음
  - API는 쉽게 말해 소프트웨어·하드웨어가 서로 호환돼 작업할 수 있도록 하는 연결고리
  - 개방형 표준 기반의 통일된 교차 아키텍처 프로그래밍 모델인 oneAPI를 이용해 개발자는 가속화된 컴퓨팅을 위한 최적의 하드웨어를 자유롭게 선택가능
  - 원API 핵심은 소프트웨어를 통해 이종 디바이스들을 하나의 매개체로 지원하는 것

다. AMD

- (확장 전략) AMD는 2022년 3월 새로운 AMD 인스팅트 MI210(AMD Instinct MI210) 액셀러레이터와 ROCm 5 소프트웨어를 발표하며 AMD 인스팅트 생태계를 다시 한번 확장하고자 함
  - 에이수스, 델, 기가바이트, 휴렛 팩커드, 레노버 및 슈퍼마이크로 등 주요 파트너사의 시스템과 폭넓게 호환되며 고성능 컴퓨팅과 인공지능 워크로드에서 고객에게 엑사스케일급 성능을 제공한다고 발표
  - 이전 세대 액셀러레이터 대비 약 2배 확장된 시스템 호환성, HPC 및 AI 애플리케이션 고객 확대, 주요 워크로드에 대한 상용 독립 소프트웨어 업체(ISV) 지원을 통해 AMD 인스팅트 생태계를 지속적으로 확장하고자 함

그림 12 AMD ROCm



* 출처: AMD

라. 퀄컴

- (확장 전략) 엔드-투-엔드 AI 소프트웨어, 개발자 경험 개선 위해 단일 통합 소프트웨어 스택으로 퀄컴의 AI 소프트웨어 제공
  - 퀄컴 AI 스택은 텐서플로(Tensorflow), 파이토치(PyTorch), ONNX 등 다양한 AI 프레임워크와 더불어 일반적으로 사용되는 런타임, 개발자 라이브러리 및 서비스, 시스템 소프트웨어, 툴, 컴파일러를 지원해 특정 기기에 맞게 개발된 AI 기능을 다른 기기에도 쉽게 적용 가능함
  - 제조사와 개발자는 퀄컴 AI 스택을 통해 스마트폰, IoT, 자동차, XR(확장현실), 클라우드, 모바일 PC 등 커넥티드 지능형 엣지 제품의 AI 역량을 극대화하고, 이를 통해 단일 AI 소프트웨어 패키지로 향상된 성능 제공

그림 13 퀄컴 AI Stack

* 출처: 퀄컴

마. 구글

- (확장 전략) '구글I/O 2022'에서 스마트워치부터 태블릿PC, 무선이어폰에 이르기까지 다양한 신제품들을 선보이며 하드웨어 시장 진출을 공식화함으로써 기존 강세를 보였던 소프트웨어(SW) 시장을 넘어 '하드웨어(HW)' 시장으로의 진출을 본격화함
  - 특히 구글이 중점적으로 사업을 추진할 것으로 예상되는 하드웨어 부문은 '스마트워치'이며 애플과 삼성전자, 화웨이 등 글로벌 IT기업들이 경쟁중인 시장임
  - 구글I/O 2022 컨퍼런스에서 공개한 자사의 첫 번째 스마트워치는 '픽셀 워치'이며 그 외 중요 하드웨어 사업 부문은 '스마트폰'과 '태블릿PC', '무선 이어폰'
- (하드웨어 시장 진출 배경) 기업의 차세대 전략 방향인 독자적 모바일 하드웨어 생태계를 확장하고자 하며 전략의 배경에는 애플의 SW+HW의 시너지 효과가 존재함
  - 단순 OS부문만 본다면 구글의 안드로이드가 애플의 macOS보다 강력한 것은 맞지만 하드웨어와의 시너지 효과 면에서는 애플이 훨씬 강력함
  - 애플은 아이폰, 맥북, 애플워치, 아이패드 등 하드웨어들은 무조건 자사의 OS인 macOS를 자동으로 사용하기 때문에 한 번 애플 제품을 사용해 macOS에 익숙해진 고객들은 다음 제품 역시 애플을 사게 되는 락인효과가 존재함

그림 14 구글 H/W

* 출처: 구글

2 국내 기업의 풀스택 전략

가. SKT

- (풀스택 범위) SK텔레콤에서 분사한 AI 반도체 전문 기업 사피온(SAPEON)과 협력해 AI 반도체 '사피온'을 개발하고, 이를 데이터센터에 사용할 예정이며, AI반도체 칩 기반 하드웨어부터 AI알고리즘, API 등 소프트웨어까지 AI서비스 제공에 필요한 통합 솔루션을 제공하는 풀스택 전략을 계획하고 진행중
  - (하드웨어) 그래픽카드(GPU)의 80% 수준의 전력을 사용하고 연산처리 속도는 1.5배 향상되었으며 가격은 절반 수준으로 소개하였으며 PIM 기술이 적용된 SK하이닉스의 반도체(GDDR6-AiM)와 '사피온'이 결합된 제품 개발 로드맵 발표
  - 글로벌 빅테크 기업을 주요 고객사로 삼아 AI반도체 사업 확장을 목표로 함

그림 15 SKT AI Full Stack

* 출처: SKT

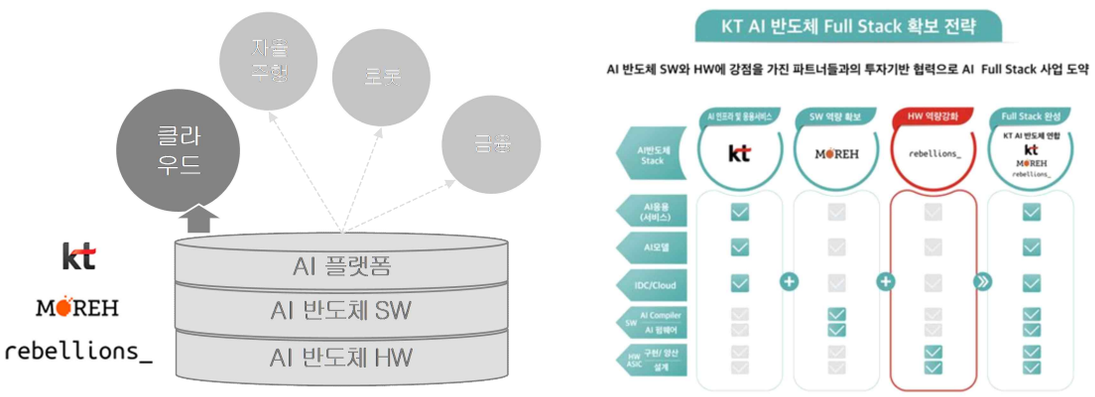
나. KT

- (풀스택 범위) KT는 국내 팹리스 기업과 협력으로 AI 반도체 분야에 본격 진입해 디지코 성장을 가속화하고 미래 사업 기회를 확보하기 위해 2021년 AI인프라 솔루션 전문 기업 '모레', 2022년에는 국내 인공지능 반도체 팹리스 기업인 리벨리온에 300억원 규모의 전략 투자를 진행하고 사업을 협력하기 하기로 함
  - KT는 리벨리온과 차세대 AI 반도체 설계와 검증, 대용량 언어모델 협업 등 AI 반도체 사업의 컨트롤 타워 역할을 수행해 나갈 예정
  - KT는 올해 KT그룹의 AI 인프라와 응용서비스, 모레의 AI 반도체 구동 소프트웨어, 리벨리온의 AI 반도체 역량을 융합해 그래픽처리장치(GPU) 수천 장 규모에 달하는 초대규모

'GPU팜'을 구축 완료할 예정

- 2023년에는 해당 GPU팜에 하이퍼스케일 AI컴퓨팅 전용으로 자체 개발한 AI 반도체를 접목할 계획

그림 16 KT AI반도체 Full Stack



* 출처: KT 보도자료, 저자 재작성

Ⅳ 국내 인공지능 반도체 기업의 경쟁력 강화 방안

- (소프트웨어 관점의 팹리스 한계) 팹리스 기업은 AI 반도체, 시스템 소프트웨어, 응용프로그램을 함께 개발해 판매하는데. 그간 중소기업과 스타트업은 반도체 설계에 역량을 집중하기 어려웠음
  - 시스템 소프트웨어, 응용프로그램 개발 및 최적화에 적지 않은 시간과 비용이 투입됨
- 대형 제조사가 제공하는 시스템 소프트웨어는 자사 칩에 최적화돼 있고 비공개로 개발돼 적용에 한계가 있음
  - 특히, 컴파일러는 칩 종류와 AI 응용프로그램에 따라 일일이 다르게 개발해야 하는 번거로움도 존재하며 상호 호환성과 새로운 영역으로의 확장성에 한계가 있음
- 국내 팹리스 스타트업은 대기업의 수직계열화, 대기업과의 연합모델이 아닌 독자 생존을 위해서는 엔비디아의 GPU 벽을 넘어야 하며, 칩이 경쟁력을 갖기 위해서는 GUDA 기반의 락인 효과 극복할 방안을 찾아야 함
  - (인텔의 oneAPI 사례) 인텔이 2021년 첫 서버용 GPU를 출시하며 XPU 전략을 공개했으며 XPU 전략 핵심은 'oneAPI'
  - CPU・GPU・FPGA까지 인텔만의 기기 호환성을 강화해 다양한 고객군을 인텔 생태계에 묶어두려는 전략(락인 전략)이며 인텔 조차도 엔비디아의 벽을 넘기 위해서 '장기전'으로 예상
  - 국내 팹리스의 역량으로는 벤치마킹하기 어려운 전략이기 때문에 국가 연구기관의 역할이 필요
- 인공지능 반도체 시장은 아직 지배적 강자가 존재하지 않는 초기 단계로, 지금부터의 국가적 대응 노력이 글로벌 주도권 경쟁의 성패를 좌우
- 우리나라도 인공지능 반도체 기술개발을 위한 대규모 재원을 확보하고 투자를 시작('20)한 결과 국내 팹리스 기업을 중심으로 국산 인공지능 반도체가 출시되어 상용화 준비단계
  - 퓨리오사, 사피온, 리벨리온, 딥엑스, ETRI, KAIST 등에서 제품 출시
- 인공지능 하드웨어는 구동하는 소프트웨어가 얼마나 최적화돼 있는지에 따라 그 성능이 크게 좌우됨
  - 수천, 수만 개의 연산 회로를 시스톨릭 배열(Systolic Array)을 통해 동시에 구동하고 그 결과를 효율적으로 취합하는 일은 고도의 계산에 따라 조직화(Coordination)하는 과정이 필요한 작업

- 특히 제작된 AI 칩에 있는 수많은 연산 회로가 쉬지 않고 번갈아 동작하도록 데이터의 공급 순서를 정하고 계산 결과를 다음 단계로 보내는 일은 특화된 저장장치(Library)를 통해 이뤄져야 해, 효율적인 저장장치와 컴파일러(Compiler)를 개발하는 것이 하드웨어 설계 못지않게 중요함
- 엔비디아의 GPU도 그래픽 엔진에서 출발했지만 "쿠다(CUDA)"라는 개발 환경을 통해 사용자가 쉽게 프로그램을 작성하고 GPU 위에서 효율적으로 작업을 수행할 수 있도록 해, AI 관련 커뮤니티에서 널리 사용될 수 있었음

○ 국내의 인공지능 컴파일러 현황은 대학 및 출연연구소의 공동 연구를 통해 개발이 진행되고 있으나 국내 SW 컴파일러의 문제는 개발 후 비즈니스가 어렵다는 점에 있음

- 인공지능 반도체를 판매하기 위해 꼭 필요한 기술이지만 컴파일러는 여러 종류의 NPU 반도체에도 적용할 수 있기 때문에 일반적인 팹리스 회사는 당장 물리적으로 보이는 NPU 개발에 많은 비용을 사용함
- 인공지능 반도체를 개발해 양산하려면 인공지능 반도체 고객은 사용할 인공지능 모델이나 관련 SW SDK(소프트웨어 개발 키트)가 없기 때문에 양산과 판매를가 불가함

○ 국내의 상황을 방지하기 위해 인공지능 반도체가 개발되기 전에 인공지능 프로그램과 파이토치(Pytorch), 텐서플로(TensorFlow) 같은 인공지능 플랫폼의 AI 모델을 그대로 번역해 사용하기 위한 컴파일러 기술이 선행돼야 함

- 엔비디아는 물론 ARM, 구글 등 모든 세계적으로 유명한 인공지능 회사는 자체 컴파일러를 보유하고 있으며, 국내에서는 컴파일러 기술을 개발한 인력이 드물기 때문에 한국의 인공지능 반도체 회사들이 공유할 수 있는 공개 컴파일러를 공동으로 개발해야 함

○ (연구기관의 역할) 통합 시스템 소프트웨어 개발을 통해 응용프로그램 개발과 최적화 시간을 단축하고 시스템 소프트웨어 개발을 용이하게 할 수 있으므로 반도체 생산과 판매비용 절감효과와도 연계됨

- CPU, GPU뿐 아니라 NPU 프로세서까지 모두 호환되며 특히, 지원해야 할 AI 응용프로그램과 칩의 종류가 늘어나면 효과 증가
- 기존에는 '딥러닝 플랫폼 종류 칩 종류'만큼 컴파일러를 개발해야 했으나, 이제는 범용성 높은 '네스트 컴파일러' 하나가 대신할 수 있음

○ 통합시스템 소프트웨어인 '네스트 컴파일러'를 오픈소스로 공개하면서 관련 산업 생태계 활성화 목적으로 자체 개발한 AI 반도체에 네스트 컴파일러를 적용한 참조 모델까지 함께 공개함

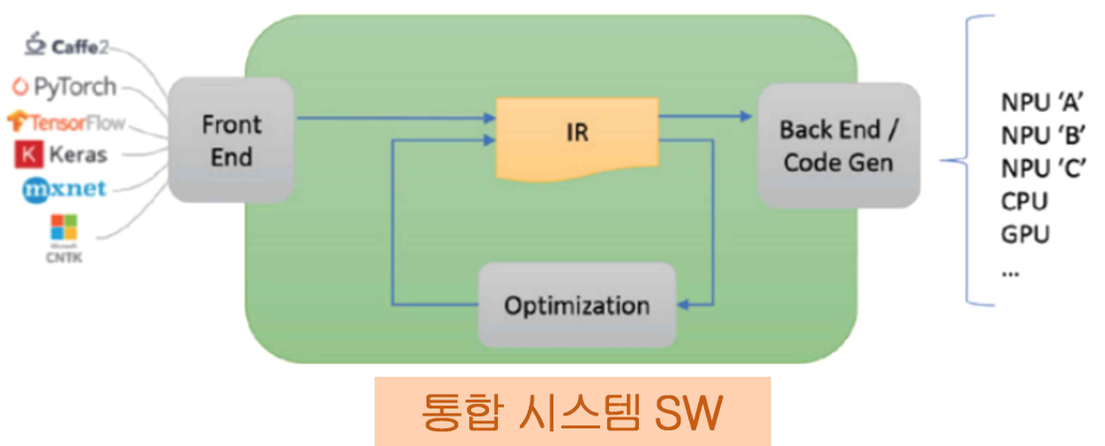
○ (생태계 조성) AI 기술이 발전하면서 자체 개발한 AI반도체로는 한계가 있기 때문에 시장선점이나 기술 경쟁보다는 생태계 마련이 더 중요

- 기존의 CPU, GPU에서의 SW 개발환경이 다르기에 NPU 기반 SW 및 SW 개발환경 생태계 조성이 필요

- 국내 AI 반도체 스타트업이 자체 개발한 고성능 AI 반도체에도 통합 시스템 소프트웨어를 적용

  - 관련 기업과 협업해 딥러닝 컴파일러 지원 범위를 넓힐 계획이며, SW 전문 기업을 통해 AI 반도체 적용 서비스 사업화도 추진 중
  - AI 응용 서비스 개발에 필요한 성능과 편의성을 향상시켜 새로운 서비스 창출에도 기여 할 계획

그림 17 ETRI NEST-C



* 출처: ETRI

- 국내 인공지능 반도체 산업이 경쟁력 확보를 위해서 NPU 반도체를 오픈소스로 공유해 개발하고 SW 컴파일러, SW SDK까지 공유하는 현실에서 인공지능 반도체 회사는 다양한 인공지능 반도체 솔루션에서 경쟁력 확보 방안을 모색해야 함

참고문헌

국내자료

관계부처합동 (2020), 인공지능 반도체 산업 발전전략

김진규 외 (2021), “인공지능 프로세서 컴파일러 개발 동향”, 「전자통신동향분석」, 제36권 제2호, pp.14-22, ETRI.

박영준 (2020), AI 반도체 시장 동향 및 우리나라 경쟁력 분석, ETRI.

오광일 외 (2020), “인공지능 뉴로모픽 반도체 기술 동향”, 「전자통신동향분석」, 제 35권 제3호, pp.76-84, ETRI.

유미선 외 (2018), “딥 뉴럴 네트워크 지원을 위한 뉴로모픽 소프트웨어 플랫폼 기술 동향”, 「전자통신동향분석」, 제33권 제4호, pp.32-42, ETRI.

국외자료

Going vertical:A new integration era in the semiconductor industry, accenture

SML-DRAFT-Microlectronics-Strategy-For-Public-Comment

웹사이트

한국정보통신기술협회 용어사전, http://terms.tta.or.kr/dictionary

KT 홈페이지 홍보센터-보도자료, https://corp.kt.com

신문기사

IT조선(2022.06.20.), 인텔이 소프트웨어 개발자 1만7000명 둔 이유 (https://it.chosun.com/site/data/html_dir/2022/06/20/2022062000296.html)

저자소개

이선재 ETRI 지능화융합연구소 기술정책연구본부 기술경영연구실 박사후연구원
e-mail: lseonj@etri.re.kr Tel. 042-860-1017

기술정책연구본부 기술정책 트렌드

발행인 이 지 형
발행처 한국전자통신연구원 지능화융합연구소 기술정책연구본부
발행일 2022년 12월 31일

www.etri.re.kr

ETRI Electronics and Telecommunications Research Institute

34129 대전광역시 유성구 가정로 218
TEL.(042) 860-6114 FAX.(042) 860-6504