2022-9호

AI TREND WATCH

2022. 10. 31.

인공지능 반도체 선도기업 성공요인 분석

한국전자통신연구원 인공지능프로세서/SW연구실 권현정 연구원
정보통신정책연구원 디지털경제연구실 윤성욱 부연구위원

KISDI 정보통신정책연구원
KOREA INFORMATION SOCIETY DEVELOPMENT INSTITUTE

인공지능 반도체 선도기업 성공요인 분석

KISDI

한국전자통신연구원 권현정 연구원
정보통신정책연구원 윤성욱 부연구위원

개 요

◈ **글로벌 인공지능 반도체 선도기업의 대표 인공지능 하드웨어 제품, 소프트웨어 스택을 포함한 주요 기술현황을 분석하고 선도기업의 성공요인을 도출**

▶ 대표적인 글로벌 인공지능 반도체 선도기업에 공통적으로 나타나는 성공요인을 분석하고 이를 바탕으로 대한민국 디지털 전략의 '6대 혁신기술 분야' 중 하나인 인공지능 반도체의 국가 경쟁력 제고를 위한 정책제언을 도출

※ 본고는 정보통신정책연구원(2021)의 "인공지능 반도체 산업 확산 가속화 방안"의 내용 중 "글로벌 인공지능 반도체 선도기업 혁신전략"을 중심으로 요약·정리

주요 내용

[인공지능 반도체 개요]

◈ **(인공지능 반도체) 첨단 시스템반도체의 일종으로 데이터의 수집부터 연산까지 데이터 전주기 전반에 활용되는 시스템반도체와 달리, 인공지능 반도체는 데이터의 학습·추론 등 인공지능 구현의 핵심 연산에 집중한 하드웨어**

▶ 인공지능의 활용이 전산업으로 확대됨에 따라 범용 반도체를 활용한 인공지능 연산의 한계가 명확해졌고, 인공지능 연산을 보다 효율적으로 구현할 수 있는 인공지능 반도체의 중요성이 부각

▶ 인공지능 모델의 연산은 인공지능 반도체에 할당하고, 기존의 범용 프로세서(CPU 등)는 다른 순차적 연산을 담당하여 연산속도를 높이고 소비전력을 관리할 필요

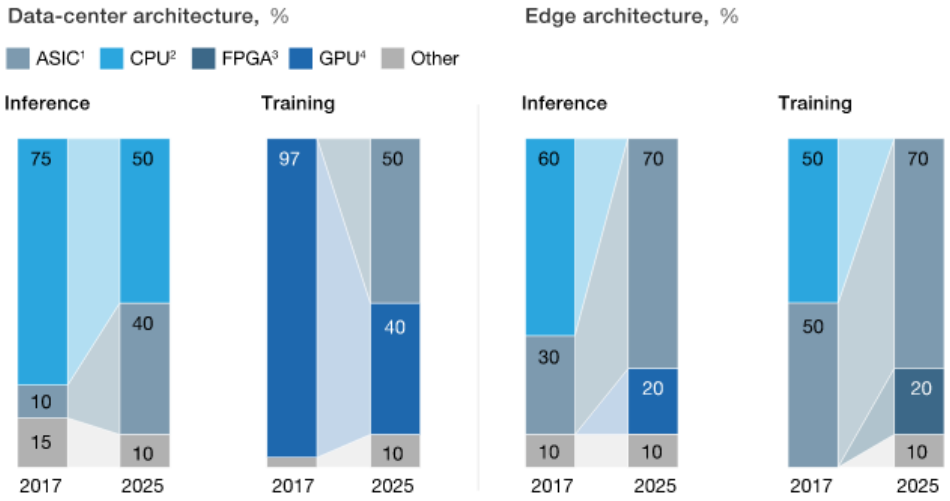
◈ **(인공지능 연산 가능 하드웨어) GPU, FPGA, ASIC(NPU), CPU 등이 인공지능 알고리즘을 구동할 수 있는 하드웨어임**

▶ (GPU) Graphic Processing Unit의 약자인 GPU는 그래픽 처리 장치로써 그래픽 연산을 빠르게 처리하여 결과값을 출력하는 연산장치

※ GPU는 VPU(Visual Processing Unit)라고도 불림

- GPU는 병렬성이 우수하고 초당 처리할 수 있는 명령어의 수가 많아 인공지능 알고리즘 구현 시 높은 성능을 보이고 최근까지도 인공지능 학습을 위해 GPU가 가장 많이 활용
- GPU는 높은 계산 효율을 갖고 있지만, 내부 메모리 용량이 범용 프로세서인 CPU 대비 낮아 내부 메모리 용량을 넘어서는 연산을 수행해야 할 때 외부 메모리 접근에 많은 시간이 소요되어 효율이 감소하는 경향이 있음

〈데이터센터 및 엣지 아키텍쳐에서 학습, 추론에 활용되는 인공지능 반도체 유형〉



자료: McKinsey & Company(2018)

- NVIDIA가 GPU에서 가장 높은 시장점유율을 보유하고 있으며, Intel, ARM, AMD 등 전통적인 CPU 주력기업들도 자사의 GPU를 개발하여 CPU-GPU 및 관련 소프트웨어가 모두 통합된 제품군을 제공

▶ (FPGA) 설계 가능 논리소자와 프로그래밍이 가능한 내부 회로가 포함된 반도체 소자로 Field Programmable Gate Array의 약자

- FPGA는 디자인한 이후에 수정 및 재설계가 가능하여 프로토타입 제작이나 ASIC 설계 테스트용으로 활용할 수 있음
- 인공지능 모델은 애플리케이션마다 최적화에 필요한 하드웨어 구조가 다를 수 있다는 점을 고려하면 FPGA의 유연성은 인공지능 모델의 활용에 큰 장점
- 단위 시간당 처리량과 전력효율이 높고 latency, 제작 시간 및 비용이 낮아 Xilinx, Intel, Microsoft 등을 포함한 많은 기업에서 개발을 진행하고 있음

▶ (ASIC) 특정 용도용 집적회로인 ASIC(Application-Specific Integrated Circuit)은 범용이 아닌 특정 용도에 맞게 맞춤 제작한 반도체로, 용도를 인공지능 연산으로 설정하여 설계한 회로를 일반적으로 NPU(Neural Processing Unit)로 지칭

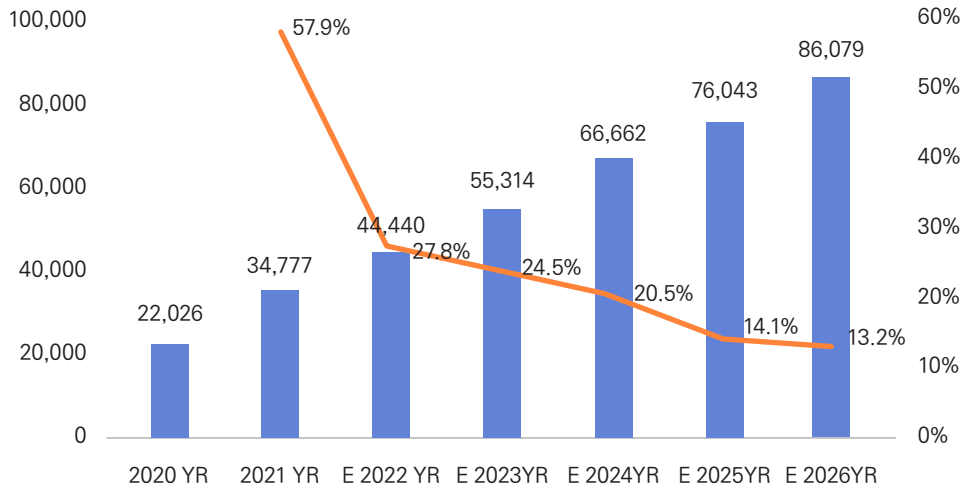
- 현재 인공지능 연산에서 가장 많이 활용되는 GPU가 원래 목적이 인공지능 연산이 아니었기 때문에, 전력효율, 수행시간, 면적 등에서 최적화가 되어있지 않음
- 인공지능 연산에 최적화된 하드웨어를 만들기 위해 Google, Tesla, Amazon, GraphCore, Arm, Intel 등 많은 글로벌 IT, 하드웨어 기업이 NPU를 개발 중

▶ (CPU) 중앙 처리 장치인 CPU(Central Processing Unit)은 흔히 불리는 인공지능 반도체에는 포함되지 않지만, 범용 컴퓨팅 시스템을 통제하고, 프로그램의 연산을 처리하는 가장 핵심적인 컴퓨터의 제어장치로써 인공지능 연산 또한 수행 가능

- CPU는 GPU보다 코어당 성능은 우수하지만, 병렬처리의 한계가 존재하여 인공지능 반도체로 분류하지 않는 경우가 많음
- 하지만 모든 device에 포함되어 있고, 가격이 저렴하여 접근성이 높으며, 내부 메모리 용량이 높아 인공지능 알고리즘 구현에서 주요한 역할을 수행
- 기존 대표적인 CPU 설계 기업인 Intel, ARM, AMD 등이 각 사의 주력 CPU의 성능을 지속적으로 개선하고 있으며, deep learning 구동용 솔루션 또한 개발하여 배포 중

◈ **(인공지능 반도체 시장 전망) 2022년 인공지능 반도체 시장은 전년 대비 27.8% 증가한 444억 달러로 예상되며, 연평균 19.9%씩 증가하여 2026년에는 861억 달러에 이를 것으로 전망됨**

〈인공지능 반도체 시장 전망〉



자료: Gartner(2022.5)

▶ 활용 분야별로 구분하여보면, 통신기기에 활용되는 인공지능 반도체는 2022년 기준 전년 대비 18.7% 성장한 256억 달러이고, 이는 인공지능 반도체의 57.7%에 해당함

〈활용분야별 인공지능 반도체 시장 전망〉

(단위: $M, %)

	'20	'21	'22 E	'23 E	'24 E	'25 E	'26 E	CAGR ('21–'26)
자동차	1,740	2,746	3,422	4,460	6,171	7,811	9,534	28.3%
통신	14,994	21,597	25,630	28,750	31,491	31,948	32,645	8.6%
컴퓨팅	5,119	9,943	14,124	19,261	24,747	30,089	35,353	28.9%
소비자	87	262	728	1,554	2,131	2,932	3,922	71.7%
산업	84	221	519	1,244	2,035	3,133	4,431	82.1%
저장	3	7	17	44	86	130	194	95.8%
합계	22,026	34,777	44,440	55,314	66,662	76,043	86,079	19.9%

자료: Gartner('22, May)

▶장비 유형별로 구분하면 '22년 통합 베이스밴드/애플리케이션 프로세서(Integrated Baseband/Application Processor)가 전년 대비 39.9% 증가한 173.1억 달러를 기록할 것으로 예상되며, 이는 전체 인공지능 반도체의 40.0% 수준

〈장비 유형별 인공지능 반도체 시장 전망〉

(단위: $M, %)

Device	'20	'21	'22 E	'23 E	'24 E	'25 E	'26 E	CAGR ('21–'26)
디지털 신호 처리기	6	14	32	69	102	152	216	71.6%
애플리케이션/멀티미디어 프로세서	10,771	14,440	15,725	18,423	18,935	18,366	19,371	6.1%
그래픽 처리 장치	2,609	4,786	5,869	7,231	8,897	10,559	12,166	20.5%
FPGA	115	241	442	912	1,439	1,870	2,398	58.4%
통합 베이스밴드/애플리케이션 프로세서	7,002	12,371	17,312	19,806	24,199	27,589	29,270	18.8%
마이크로컨트롤러	19	49	106	212	286	473	755	73.0%
마이크로프로세서-Compute	1,128	2,107	3,510	5,836	7,669	9,455	11,360	40.1%
마이크로프로세서-Embedded	40	86	156	291	475	683	881	59.2%
기타 특정 분야	337	684	1,288	2,535	4,660	6,895	9,663	69.8%
총합계	22,026	34,777	44,440	55,314	66,662	76,043	86,079	19.9%

자료: Gartner(2022.5)

[글로벌 인공지능 반도체 선도기업 분석]

◈ (선도기업 분석) 대표적인 글로벌 인공지능 반도체 선도기업의 대표 인공지능 제품 및 주요 기술 현황 분석을 통해 선도기업의 성공요인을 도출

▶ 선도기업을 기업 개요, 주요 제품, 소프트웨어 스택으로 구분하여 분석

※ 인공지능 하드웨어 제품은 인공지능 모델 구동 목적에 따라 추론용인 Edge용과 학습용인 Cloud용으로도 구분 가능

- (소프트웨어 스택) 인공지능 반도체 개발 후 이를 구동하고 사용자들에게 서비스를 제공하는 역할을 수행하며, Cloud Service, Deep Learning Platform, Deep Learning Complier, Runtime Library & Driver 등으로 구분

〈소프트웨어 스택〉

구분	역할
Cloud Service	대량의 인공지능 반도체 컴퓨팅 리소스를 여러 사용자가 나누어 이용할 수 있도록 하는 서비스로, 사용자가 하드웨어를 갖추고 있지 않아도 이용 가능하며, 리소스의 균등한 분배와 통신의 안정성, 보안성이 중요
Deep Learning Platform	딥러닝 모델 개발 및 추론, 학습을 위한 다양한 함수를 라이브러리화하는 플랫폼으로, 주로 파이썬 환경의 텐서플로우, 파이토치, 카페 등의 딥러닝 플랫폼이 널리 사용됨
Deep Learning Complier	딥러닝 모델을 실제 하드웨어에서 연산할 수 있도록 연산 명령어를 출력해주는 소프트웨어 스택으로, 딥러닝 모델의 연산 종류 및 연결 상태를 읽은 후, 연산기에서 수행 가능한 명령어의 집합으로 해석하는 과정을 포함
Runtime Library & Driver	컴파일러에 의해 생성된 명령어를 하드웨어의 메모리에 저장하고, 저장된 명령어를 여러 연산기 리소스에 분배하여 차례로 수행할 수 있도록 해주는 가장 low-level의 소프트웨어 스택

- 사용자가 딥러닝 하드웨어를 보유하고 있어도, 딥러닝 플랫폼 언어를 하드웨어 명령어로 변환해주는 딥러닝 컴파일러와 런타임 라이브러리, 하드웨어 드라이버가 없으면 인공지능 모델을 구동할 수 없음

- 하드웨어를 직접 보유하지 않더라도, 하드웨어와 SW 스택을 묶어서 활용할 수 있는 Cloud Service가 발전하고 있으며, 이는 딥러닝 개발자들에게 새로운 딥러닝 모델을 타깃 하드웨어 상에서 개발할 수 있는 기회를 제공하고, 딥러닝 하드웨어 개발 기업에게는 새로운 딥러닝 하드웨어 이용자 수를 늘리는 효율적인 방법 중 하나로 부각

〈인공지능 반도체 선도기업의 대표 제품, 유형 및 목적〉

기업	HW 분류	제품명	하드웨어 목적
Intel (Habana Labs)	CPU	Xeon Scalable Processors	Cloud/Edge
	GPU	Iris Xe MAX	Edge
	FPGA	Stratix 10 NX FPGA	Edge
	NPU	Movidius VPU	Edge
		Gaudi	Cloud
		Goya	Edge
ARM	CPU	Cortex-X2	-
		Neoverse-N2	Cloud/Edge
	GPU	MALI-G710	Cloud/Edge
	NPU	Ethos-N78	Inference
AMD	CPU	EPYC-7003	Cloud
	GPU	Instinct MI100 Accelerator	-
NVIDIA	CPU	A100	Cloud
	GPU	Xavier	Edge
Xilinx	FPGA	Vitis AI	Inference
Sambanova	FPGA	Reconfigurable Dataflow Unit	Cloud/Edge
Tesla	NPU	D1	Cloud/Edge
Google	NPU	TPU	Cloud
Amazon	NPU	Inferentia	Edge
GraphCore	NPU	IPU	Cloud/Edge
Kneron	NPU	KL720 AI SoC	Edge
Cerebras	NPU	WSE-2	Cloud
Hailo	NPU	Hailo-8TM AI Processor	Edge

1. Intel, Habana Labs

◈ **(기업 개요) 인텔은 1968년 설립한 반도체 설계·제조 기업으로 CPU, 메모리, 컨트롤러, 칩셋 등 다양한 유형의 반도체를 생산하고 있으며 자체 R&D 뿐만 아니라 기술력 있는 기업을 인수하는 방식으로 인공지능 반도체 개발에 착수**

▸ (Habana Labs) 인텔이 2019년 20억 달러에 인수한 Habana Labs는, '24년까지 25억 달러 이상의 가용시장 가치가 전망되는 전도유망한 인공지능 반도체 전문 설계 업체

▸ (Mobileye) 이스라엘의 자율주행 자동차 전문 기업으로 인텔이 153억에 인수하였으며, '21년 상반기에만 3억 270만 달러의 매출을 기록하여 전년 대비 124%의 매출 상승을 기록

▸ (그 외) 이외에도 Anodot(자율주행 및 이상감지), KFBIO(의료 영상 처리), MemVerge(In-memory 처리 application 개발), Xsight Labs(클라우드 기반 데이터 처리 가속화) 등에 투자

◈ **(대표 제품) Xeon Scalable Processor(인공지능용 CPU), Stratix 10 NX FPGAs(FPGA), Habana Labs의 Goya, Gaudi(NPU) 등 다양한 종류의 인공지능 반도체 라인업을 보유**

▶ (Xeon Scalable Processor) Cloud용 AI CPU로써 알리바바, 바이두, 마이크로소프트, 오라클 등 세계적인 클라우드 서비스 사업자가 인텔의 Xeon Scalable processor를 이용하여 서비스를 제공

- 다양한 core 개수, 동작 frequency, power level을 통해 폭넓은 성능의 제품을 시장에 제공하여 클라우드, 슈퍼컴퓨터, IoT, 에너지디바이스 등에도 활용

▶ (Stratix 10 NX FPGAs) AI에 최적화된 FPGA로 높은 bandwidth, 낮은 latency를 제공하여 인공지능 가속 애플리케이션에 적합한 성능을 보임

- INT8 기준 143TOPS를 달성하였으며 3D stacked high-bandwidth DRAM을 포함

▶ (Gaudi/Goya) Habana Labs가 출시한 인공지능 학습용/추론용 NPU

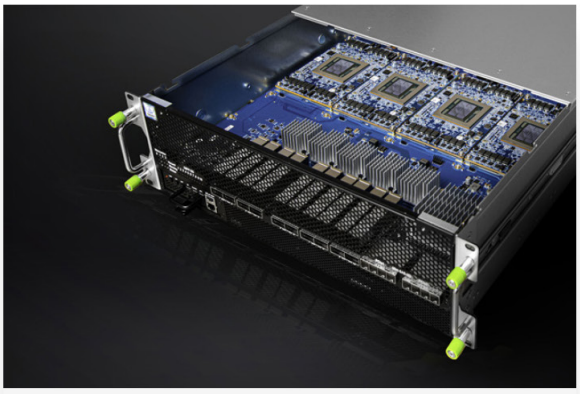
- (Gaudi) 대형 슈퍼컴퓨터를 위해 설계된 NPU로써, RoCE RDMA v2의 100GbE port 10개를 칩에 통합하여 높은 Scalability 및 쓰루풋을 달성하는 등 우수한 성능을 보임

※ 아마존 웹서비스(AWS)가 AWS BC2 인스턴스에 Gaudi 가속기 8개를 사용(2020년 8월)하는 등 시장에서 좋은 반응을 보임

- (Goya) BERT(자연어처리 모델)를 초당 최대 2,774문장을 처리하며, ResNet50(이미지 처리 모델)을 초당 최대 45,448장 처리하는 등 높은 성능을 보임

※ UCSD(미국 캘리포니아 대학교 샌디에이고) 슈퍼컴퓨터 센터의 보이저 슈퍼컴퓨터에 336개의 Gaudi, 16개의 Goya 탑재

〈Gaudi(왼쪽), Goya(오른쪽) 〉



자료: habana.ai 웹페이지, https://habana.ai/

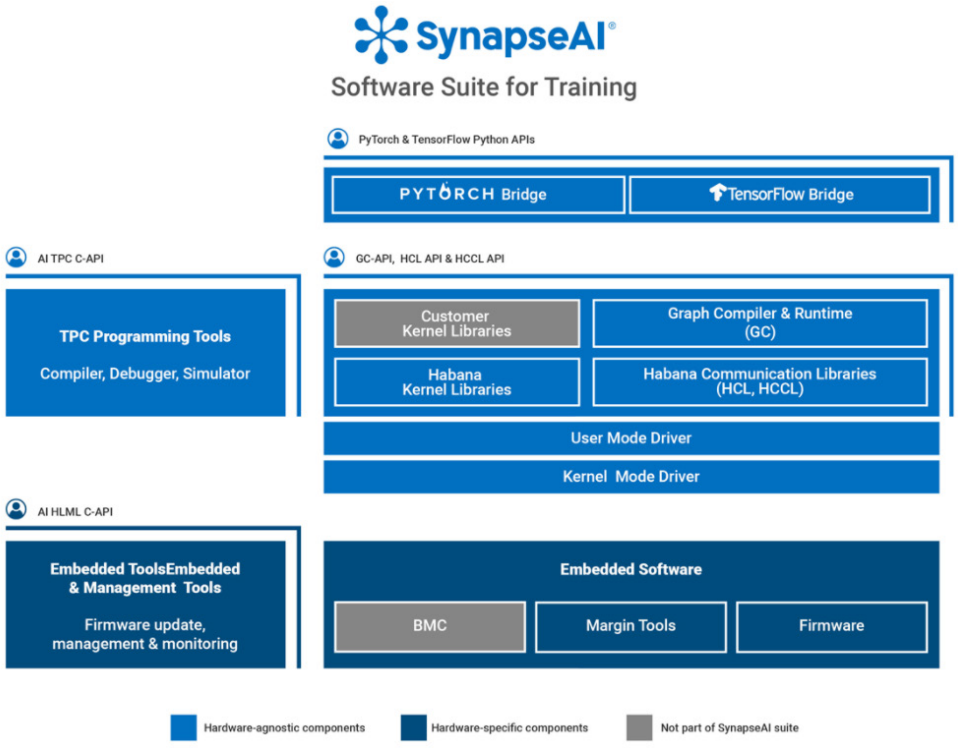
◈ **(소프트웨어 스택) Edge device와 cloud computer 간의 통신보안을 위해 다양한 소프트웨어 스택을 제공**

▸ (Intel) Intel Software Guard Extension(Intel SGX), Crypto Acceleration, Quick Assist Technology, Total Memory Encryption, Platform Firmware Resilience 서비스 등을 제공

▸ (Habana Labs) SynapseAI라는 자사의 NPU를 구동하기 위한 전체 소프트웨어 스택 제공

– 사용자가 개발한 Pytorch, Tensorflow 모델이 SynapseAI내 컴파일러를 통해 머신코드로 변환되어 Habana의 하드웨어를 구동

〈Habana Labs의 소프트웨어 스택 SynapseAI〉



자료: habana.ai 웹페이지, https://habana.ai/training-software/

2. NVIDIA

◈ **(기업 개요) 1984년에 설립된 GPU기업으로 CPU와 같은 범용 프로세서의 한계인 가속 컴퓨팅, 그래픽 컴퓨팅 문제를 해결하는 GPU를 다음 세대의 컴퓨팅 유닛으로 판단하고, 이를 비디오게임에 적용**

▸ (투자) 2018년 ABEJA, 2019년 Weka.IO, 2020년 Youvize에 투자하였으며, 2021년 추가로 4개의 기업에 투자

〈2021년 NVIDIA의 투자 요약〉

투자 일시	투자처	분야
2021년 02월	Rescale	클라우드 플랫폼
2021년 05월	VAST Data	데이터 저장소
2021년 09월	Activ Surgical	디지털 의료
2021년 10월	Domino Data Lab	인공지능 데이터 처리

자료: Crunchbase 웹페이지, NVIDIA, https://www.crunchbase.com/organization/nvidia

◈ **(대표 제품) Ampere 아키텍쳐 기반의 A100, Volta 아키텍쳐 기반의 V100 등이 대표적인 NVIDIA의 딥러닝 특화 GPU**

▶ (A100) 딥러닝 연산을 효율적으로 수행하기 위해 텐서 코어를 장착하였으며, FP16 연산 시 기존 2배 이상의 성능 향상

- 2,048개의 A100 GPU 이용 시 초대형 모델인 BERT나 DLRM을 SOTA 수준으로 빠르게 추론 가능
- FP32에서 INT4에 이르기까지, 다양한 정밀도로 딥러닝 연산을 수행하여 연산 속도를 가속화하는 기능 보유

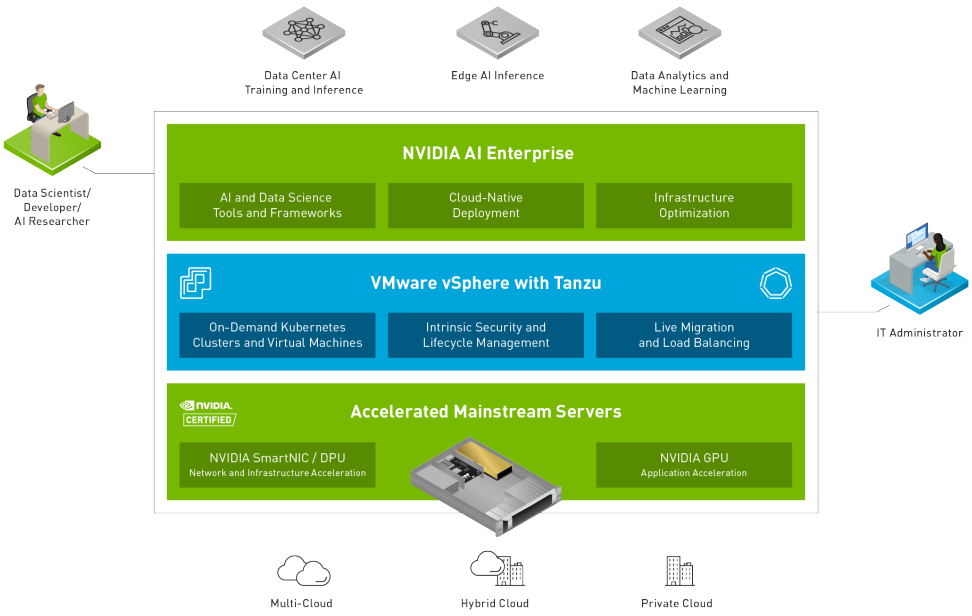
▶ (V100) Volta 아키텍쳐를 탑재한 제품군으로 단일 GPU만으로도 CPU 100개의 성능을 보임

- 단일 V100은 640개의 텐서 코어를 장착하여 100TFLOPS의 성능을 보이며 NVIDIA의 NVLink를 이용하여 최대 300GB/s의 대역폭으로 여러 V100을 연결할 수 있어 데이터 센터 구축에 적합

◈ **(소프트웨어 스택) VMWare과의 협력을 통해 AI 지원 플랫폼 서비스를 제공하고 있으며 라이브러리 컬렉션 또한 CUDA라는 이름으로 유저들에게 제공**

▶ (NVIDIA+VMWare AI 지원 플랫폼) AI workload 최적화를 위한 end-to-end 서비스 제공

〈NVIDIA+VMWare AI 지원 플랫폼〉



자료: NVidia 웹페이지, https://blogs.nvidia.com/blog/2021/03/09/vmware-ai-ready-enterprise-platform/

▶ (CUDA) NVIDIA 라이브러리 컬렉션으로 기본적인 딥러닝 연산 기능을 수행하는 cuDNN, 머신러닝 알고리즘 가속화를 위한 cuML, 추론용 모델 최적화를 위한 TensorRT, 그래프 분석을 수행하는 cuGraph 알고리즘 등이 CUDA에 포함

- 추가로, 데이터 분석 및 AI 모델링을 위한 가속화 라이브러리로 CUDA-X AI를 공개하였으며, 이는 사용자들이 주로 활용하는 대다수의 딥러닝 프레임워크와 통합되어 Amazon, Microsoft, Google Cloud 등의 클라우드 플랫폼에 활용

3. SambaNova

◈ **(기업 개요) "Enabling the Future of AI Today"라는 슬로건을 바탕으로 Kunle Olukotun, Rodrigo Liang, Christopher Re가 2017년에 설립한 반도체 회사로 소프트웨어와 하드웨어 설계 전문가를 중심으로 인력 구성**

▶ (투자유치) 상용화된 제품이 존재하지 않지만, 총 투자금액이 11억 달러를 넘어서는 등 시장의 큰 기대를 받고 있음

- SambaNova의 기업가치는 약 50억 달러로 추산되며, 2018년 6,330만 달러의 시리즈 A 투자유치부터 꾸준하게 증가하여, 2021년 4월 6억 7,600만 달러 규모의 시리즈 D 투자유치

〈SambaNova 투자 유치 내역〉

투자 일시	투자 라운드	투자 유치 금액	주요 투자자
2017년 11월	Seed 라운드	200만 달러	Celesta Capital
2018년 05월	Series A	5,600만 달러	GV, Walden International
2018년 08월	Series A	730만 달러	Celesta Capital
2019년 04월	Series B	1억 5,000만 달러	Intel Capital
2020년 02월	Series C	2억 5,000만 달러	BlackRock
2021년 04월	Series D	6억 7,600만 달러	SoftBank Vision Fund

자료: Crunchbase SambaNova 웹페이지, https://www.crunchbase.com/organization/sambanova-systems

◈ **(대표 제품) 재구성 가능한 처리/기억 유닛인 Reconfigurable Dataflow Unit(RDU)이라는 이름의 FPGA 제품 설계에 집중하고 있으며, RDU는 기억 유닛인 PMU와 처리 유닛인 PCU로 구분**

▶ (PMU) Pattern Memory Unit(PMU)은 높은 on-chip 메모리와 더불어 일부 연산 기능까지 탑재하여 데이터 이동시간을 감소시켜 latency 감소 및 bandwidth 증가

- PMU를 활용하면 1,000개 이상의 GPU가 필요한 것으로 알려진 대형 인공지능 모델을 하나의 RDU chip에서 수행 가능

▶ (PCU) Pattern Compute Unit(PCU)은 Loop level 및 pipeline 병렬 처리를 통해 연산속도를 대폭 증가시킴

◈ **(소프트웨어 스택) SambaFlow라는 이름의 소프트웨어 스택을 설계하여 RDU 하드웨어의 성능 및 장점을 극대화**

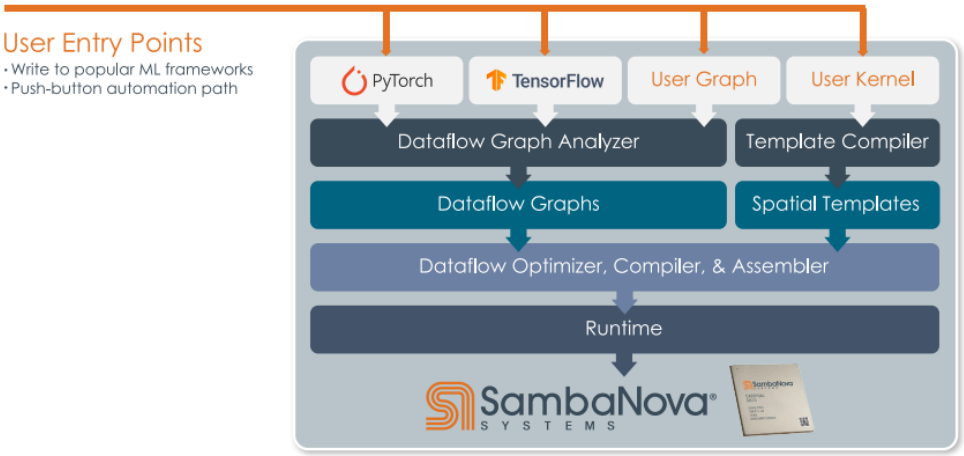
▶ (연산 과정) 많은 인공지능 개발자들이 활용하는 Pytorch, Tensorflow상에서의 모델을 바탕으로 dataflow 그래프 생성 후 최적화 과정을 거쳐 RDU 상에서 효율성을 극대화하여 구동

▶ (병렬화) 다수의 RDU를 통해 분산 처리할 수 있도록 모델·데이터 병렬화 지원

▶ (재구성) 인공지능 모형, 데이터, 배치 사이즈 수정 등으로 인해 RDU 재구성(reconfiguration)이 필요할 경우 수십 마이크로초(μs) 안에 수행

▶ (Template Compiler) 사용자의 모형에 지원하지 않는 연산자가 있는 경우, high-level에서 새로운 연산자의 동작을 서술할 수 있는 API 제공

〈SambFlow 구조〉



자료: SambaNova(2021.6)

4. Tesla

◈ **(기업 개요) 미국 캘리포니아 팔로알토에 기반을 둔 전기자동차 회사로 2003년 Martin Eberhard, Marc Tarpenning이 창업하였으며, 2021년 8월 Tesla AI day를 계기로 전기차 이외에도 다양한 AI 부문에 사업 확대를 선언**

▸(하드웨어) 기존의 강점을 가진 소프트웨어뿐만 아니라 학습, 추론을 위한 AI/딥러닝 전용 하드웨어 개발을 진행 중이며, 개발 중인 하드웨어는 자율주행과 휴머노이드 로봇(Tesla Bot)에 활용 예정

▸(기술개발·사업확장을 위한 인수·투자) 자율주행 분야 기술개발을 위해 AI 기반 자율주행 스타트업 DeepScale을 2019년 10월에 인수하였으며, 그 외 독일 배터리 기업인 ATW, 캐나다 배터리 재료기업 Springpower를 인수

◈ **(대표 제품) 인공지능 하드웨어 D1 Chip과 이를 기반으로 만든 Dojo 컴퓨터를 Tesla AI day에서 공개**

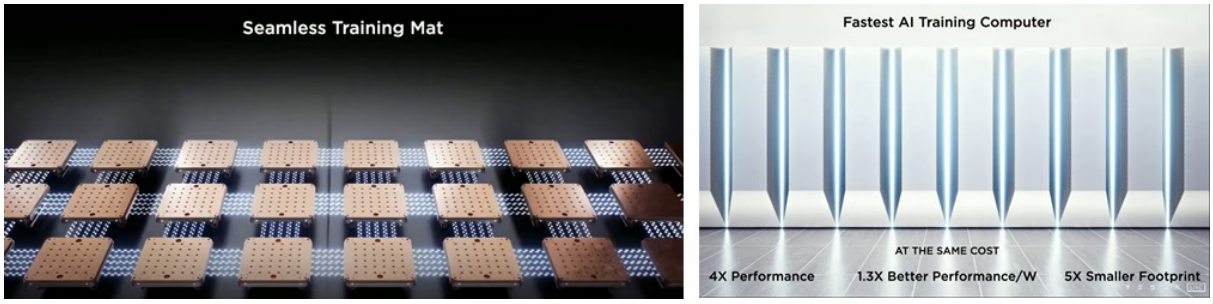
▸(D1) Tesla에서 개발 중인 하드웨어로 높은 확장성을 확보하기 위해 대역폭에서의 손실을 최소화하기 위한 고대역폭 고밀도 커넥터도 함께 개발 중

- 트레이닝 노드라는 가장 작은 연산 단위를 354개 활용하여 하나의 D1 chip을 구성하며, D1 chip 25개로 9페타플롭의 연산을 수행하는 하나의 타일을 구성
- 12개의 타일로 하나의 캐비넷을 구성하며, 11개의 캐비넷을 연결하여 1.1 ExaFLOPS을 달성하는 하나의 ExaPod을 구성

※ Tesla AI day에서 하나의 타일에 대한 실물과 training 장면을 시연

▶ (Dojo) 초대형 딥러닝 모델을 위한 슈퍼컴퓨터 개발 프로젝트로, 확장성 높은 D1 칩을 활용하여 구축되는 것이 특징이며, Dojo를 위한 분산 컴퓨팅 아키텍쳐 또한 함께 개발

〈Tesla의 D1을 활용한 타일(좌)과 이를 기반으로 한 Dojo(우)〉



자료: Tesla AI day(2021.8)

◈ **(소프트웨어 스택) 유저들이 코드 활용을 최소화하는 노코드(No-code) 기조를 지향하여 컴파일러, 드라이버, 인터페이스 등 소프트웨어 스택의 구성요소를 개발 중**

▶ (Pytorch 확장판) 유저들이 평소 개발하는 환경과 동일한 환경에서의 개발을 지원하기 위해 Pytorch 확장판을 구축하였으며, 컴파일러는 이를 읽어 나중에 재사용할 수 있는 코드를 생성

▶ (Compiler) 하드웨어용 바이너리를 생성하는 LLVM 백엔드를 포함하며, 초대형 딥러닝을 위한 체이닝, 데이터·모델·그래프 병렬 등의 병렬 처리 기술과 함께 재계산을 지원

▶ (드라이브 스택) 다수의 호스트에서 데이터를 처리할 수 있도록 멀티 파티셔닝 기능 지원

5. Google

◈ **(기업 개요) 세계 최대 검색엔진 기업으로 2011년 설립한 딥러닝 인공지능 연구팀 구글 브레인을 필두로 이미지 처리, 번역, 로봇틱스 등 다양한 분야에 인공지능을 적용하는 글로벌 ICT 기업**

▶ (인공지능 관련 인수 및 투자) 런던 AI 기업 DeepMind Technology(6억 2,500만 달러)를 포함하여 Jetpac, Dark Blue Labs & Vision Factory, Halli Labs, AI Matter 등 다양한 인공지능 관련 기업을 인수 중

- 2021년 4월엔 카카오 모빌리티 리드 투자자로 565억원을 투자하였고, 9월엔 클라우드 가속화 기업 Oatfin에 10만 달러 투자

◈ **(대표 제품) 2016년 Tensor Processing Unit(TPU) v1 공개를 시작으로 2021년 TPU v4.0을 공개하였으며, 상용제품군으로는 Edge TPU가 있음**

▸(TPU v4.0) 가장 최신 버전의 TPU는 v4로 2021년 5월 Google I/O 온라인 학회에서 공개하였으며, v3과 비교하여 약 두 배의 칩당 성능을 구현

– (TPU POD) 4,096개의 TPU Chip을 모아 하나의 TPU POD를 구성하며, 1 ExaFLOPS의 성능을 보임

▸(Edge TPU) Tensorflow Lite 기반으로 구동되며 엣지 컴퓨팅용으로 제작되어 구글 데이터센터용에 비해 크기가 작고 더 적은 전력을 소모

※ (성능) 2W를 사용하며 초당 4조 번의 연산을 수행

〈Google TPU 성능 비교〉

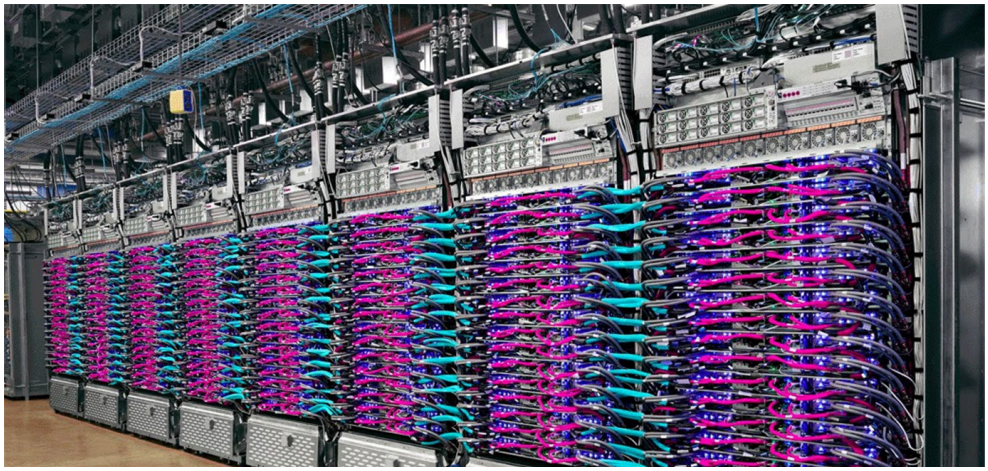
	TPU v1	TPU v2	TPU v3	TPU v4	Edge TPU
공개 일시	2016	2017	2018	2021	2018
노드	28 nm	16 nm	16 nm	7 nm	–
온칩 메모리	28 MB	32 MB	32 MB	144 MB	–
TOPS	23	45	90	–	4

자료: Jouppi, N.P. 외(2021)

◈ **(소프트웨어 스택) 딥러닝 프레임워크 뿐만 아니라, 딥러닝 알고리즘, 클라우드 서비스 등 다방면의 소프트웨어 서비스를 함께 제공하여 TPU의 활용성을 제고**

▸(Cloud TPU) Cloud TPU라는 이름의 클라우드 서비스를 제공하여 실제 TPU를 구매하지 않더라도 이용 시간에 따라 금액 지불 후 딥러닝 알고리즘을 TPU에서 구동하는 환경을 제공

〈Google Cloud TPU〉



자료: Google Cloud 웹페이지, http://cloud.google.com/tpu

▶ (Accelerated Linear Algebra: XLA) 딥러닝 알고리즘을 TPU 상에서 구현하기 위해 명령어를 생성하는 컴파일러인 Accelerated Linear Algebra(XLA)를 오픈소스로 제공

- 구글의 Tensorflow를 TPU에서 해석하는 소프트웨어로, XLA와 관련된 아키텍쳐, 문법, 중간표현, 새로운 하드웨어 도입을 위한 백엔드 개발에 관한 정보 제공

6. GraphCore

◈ **(기업 개요) 영국의 인공지능 반도체 설계 전문 기업으로 2016년 설립되었으며 Intelligence Processing Unit(IPU)라는 이름의 하드웨어 개발에 주력**

▶ (투자 유치) 2016년 10월 Amadeus Capital Partners, Robert Bosch Venture, Capital Samsung Strategy and Innovation Center의 리드 투자로 이어진 3,000만 달러 시리즈 A 투자 유치 이후 2020년 12월까지 약 6억 8,200만 달러의 투자유치를 기록

※ GraphCore의 기업가치는 27억 달러 이상으로 추산(테크크런치, 2020.12)

〈GraphCore 투자유치 내역〉

투자 일시	투자 라운드	투자 유치 금액	리드 투자자
2016년 10월	Series A	3,000만 달러	Amadeus Capital Partners, Robert Bosch Venture Capital, Samsung Strategy and Innovation Center
2017년 07월	Series B	3,000만 달러	Atomico
2017년 11월	Series C	5,000만 달러	Sequoia Capital
2018년 12월	Series D	2억 달러	BMW i Ventures, Microsoft
2020년 02월	Series D	1억 5,000만 달러	Mayfair Equity Partners
2020년 12월	Series E	2억 2,200만 달러	Ontario Teachers' Pension Plan

자료: Crunchbase 웹페이지, GraphCore, https://www.crunchbase.com/organization/graphcore

◈ **(대표 제품) 현재까지 두 세대에 걸친 IPU 프로세서를 출시하였고 이를 이용한 플랫폼 또한 공개**

▶ (IPU) 1세대 MK1-GC2 IPU 대비 8배의 성능 향상을 기록한 2세대 IPU MK2 GC2000 IPU를 출시하였으며, 이는 마이크로소프트 클라우드 Azure에서 인공지능 반도체로 납품

〈GraphCore IPU Processor 제품 비교〉

제품명	노드	코어 개수	Thread 개수	In-Processor-Memory 용량	Bandwidth
ColossusTM MK1-GC2 IPU	16 nm	1216	7296	300 MB	45 TB/s
ColossusTM MK2 GC2000 IPU	7 nm	1472	8832	900 MB	47.5 TB/s

자료: GraphCore 웹페이지, https://www.graphcore.ai/

▶ (플랫폼) IPU 프로세서 기반 플랫폼을 통해 클라우드 서비스 제공

〈GraphCore IPU Processor 활용한 플랫폼〉

제품명	특징
IPU-M2000	4개의 ColossusTM MK2 GC2000 IPU 프로세서 탑재, 1 PetaFLOPS 달성
IPU-POD16	4개의 IPU-M2000 연결, 4 PetaFLOPS 달성
IPU-POD64	16 PetaFLOPS 달성
GraphCloud	IPU-POD16 및 IPU-POD64를 클라우드 형태로 제공하며 사용량에 따라 요금을 부과

자료: GraphCore 웹페이지, https://www.graphcore.ai/

◈ **(소프트웨어 스택) Poplar라는 이름의 소프트웨어 스택을 개발, 제공함으로써 IPU의 활용성을 높임**

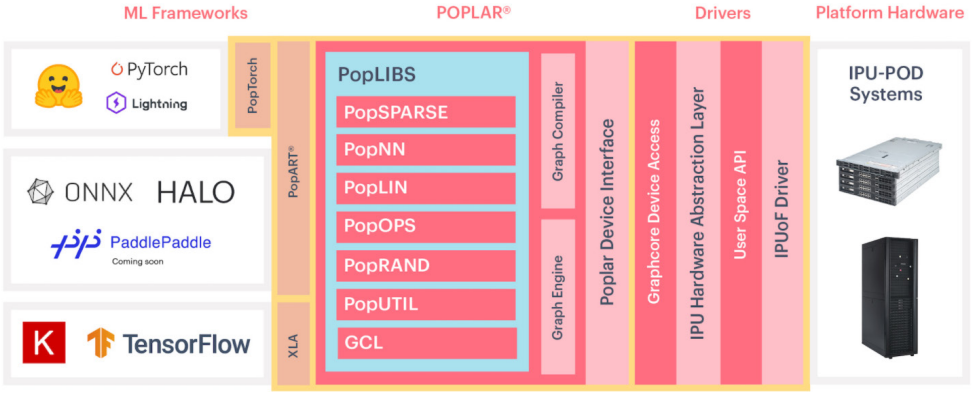
▶ (Poplar) 인공지능을 활용하는 개발자들이 일반적으로 사용하는 딥러닝 프레임워크를 다수 지원하는 소프트웨어 스택

- Tensorflow 1.0, 2.0을 모두 지원하며, Pytorch 사용 시에는 ATEN backend를 사용하여 타깃 하드웨어를 IPU로 지정
- Poplar 중 PopART 활용 시 ONNX 모델을 인식하여 Python/C++을 통한 모델 구축을 지원

▶ (PopLibs) GraphCore에서 제공하는 그래프 라이브러리로 여러 머신러닝 모델에 공통적으로 활용되는 50개 이상의 함수를 최적화하여 제공

- (Graph Compiler) 병렬화를 위한 스케쥴링, 메모리 컨트롤, partition에 관한 작업을 지원하며, LLVM을 활용하여 코드를 생성하는 역할을 수행
- (Graph Engine) 인공지능 모델을 처리하기 위한 런타임으로, IPU와 호스트간의 interface의 최적화를 통해 효율적 데이터 이동을 지원하며, 네트워크 확인을 병행하여 디바이스를 관리

〈GraphCore 소프트웨어 스택 Poplar 구조〉



자료: GraphCore 웹페이지, https://www.graphcore.ai/

7. Kneron

◈ **(기업 개요) 2015년에 Albert Liu에 의해 설립된 샌디에이고 기반 AI Edge 디바이스 솔루션 기업으로 자동차, 보안 등의 IoT application에 활용되는 Edge용 하드웨어와 소프트웨어 스택 개발에 주력**

▸ 하드웨어뿐만 아니라 소프트웨어 스택과 하드웨어에 적합한 이미지 인식 모델을 같이 개발함으로써 하드웨어 성능의 최적화를 추구

▸ (이미지 센싱) 멀티미디어 SoC 솔루션 업체 Vatics 인수(2021.5)를 통해 자사의 하드웨어에 이미지 센싱 기능 탑재 추진

▸ (자율주행) 반도체 설계 기업 Weltrend, 휴먼 인터페이스 솔루션 Elan/Avisonic과 협력하여 자율주행 자동차 기술개발 진행

- ADAS는 Kneron의 인공지능 하드웨어, Weltrend사의 스마트 카메라용 하드웨어, Elan/Avisonic사의 객체 감지 알고리즘 등 협력사의 기술을 활용하여 기술개발을 진행

▸ (투자) Alibaba Enterpreneurs Fund, CDIB Capital 등의 시리즈 A 투자 이후 Delta Electronics 등의 투자로 총 1억만 달러 이상의 투자유치 달성(MeetGlobal, 2021)

〈Kneron 투자유치 내역〉

투자 일시	투자 라운드	투자 유치 금액	리드 투자자
2017년 11월	Series A	1,500만 달러	Alibaba Entrepreneurs Fund, CDIB Capital
2018년 05월	Series A	1,800만 달러	Horizons Ventures
2020년 01월	Series A	4,000만 달러	Horizons Ventures
2021년 05월	Corporate 라운드	1,700만 달러	Delta Electronics

자료: Crunchbase 웹페이지, Kneron, https://www.crunchbase.com/organization/kneron

◈ (대표 제품) 개발 중인 on-device edge AI를 성능, 전력 소모, 비용 부문에서 효율적인 구현을 위한 SoC(System on Chip) 개발 및 출시

▶ (SoC) 고성능 제품인 KL720 AI SoC와 저전력 제품인 KL520 AI SoC로 구분하여 출시

- (KL720) 사이클당 1,024 MAC 연산을 하는 자사의 NPU와 Arm의 Cortex M4 CPU, Tensilica DSP, 128MB SDRAM 메모리, 128MB 플래시 메모리로 구성된 고성능 SoC
- (KL520) 사이클당 576 MAC 연산을 하는 자사의 NPU와 AI coprocessor용 2개의 Arm의 Cortex M4 CPU, 32/64MB SRAM, 64MB 플래시 메모리로 구성된 저전력 SoC으로 8개의 AA 배터리만으로 디바이스를 약 15개월 동안 구동 가능

〈Kneron 솔루션 제품 목록〉

카테고리	제품명	특징
SoC	KL720 AI SoC	사이클당 1024 MAC 계산 4K 이미지, Full HD 비디오, 3D 센싱 등에 최적화
	KL520 AI SoC	사이클당 576 MAC 계산 스마트 도어락, 로봇 청소기 등 스마트 홈 디바이스 등에 최적화
	NPU	Kneron의 NPU를 IP형태로 제공
딥러닝	Kneron-003	이미지 인식 모델로써, 미국 NIST 2019 FRV test에서 64MB 이하 모델 중 가장 높은 점수를 기록
모듈	Kneo Stem	AI앱을 개발하고 KNEO AI 앱스토어에 업로드할 수 있는 USB 형태의 개발 kit 블록체인으로 데이터를 보호

자료: Kneron 웹페이지, https://www.kneron.com/

◈ (소프트웨어) 저전력 및 이미지 인식의 정확도 향상을 위한 소프트웨어 개발

▶ (Kneron-003) 50MB 이하의 모델로 6개월 이상 동작하는 edge AI 모델 애플리케이션 알고리즘으로 저전력의 특성을 활용하여 배터리로 동작하는 드론, 로봇, 스마트벨, 도어락, AI 캠 등에 적용

- 미국 National Institute of Standards and Technology(NIST)의 Face Recognition Vendor Test(FRVT)에서 6개의 경량 모델 중 가장 높은 점수를 취득(2019.7)

※ Kneron 모델은 57MB의 용량이며, 경량화 이외 모든 모델을 포함하였을 때 VISA 카테고리에서 5번째, Mugshot 카테고리에서 2번째 고득점을 기록

8. Cerebras

◈ **(기업 개요) Andrew Feldman, Gary Lauterbach, Michael James, Sean Lie, Jean Philippe Friker가 설립한 미국의 AI용 컴퓨팅 시스템 기업으로 자사의 컴퓨터에 탑재하기 위한 인공지능 하드웨어를 자체 개발**

※ (SeaMicro) Cerebras의 공동창업자들은 Andrew Feldman, Gary Lauterbach가 설립한 SeaMicro에서 인연을 맺은 사이로, SeaMicro는 2012년 3억 3,400만 달러에 AMD에 인수됨

▶ (Wafer Scale Engine) 2019년 8월 1조 2천억개의 transistor로 구성된 초거대 AI 반도체(Wafer Scale Engine: WSE)를 발표하였는데, 이 칩은 TSMC 16nm 공정으로 생산되었으며, 40만 개의 코어로 구성되어 9PByte/s의 메모리 bandwidth를 달성하였고, 칩의 크기가 46,225㎟로 웨이퍼 사이즈와 유사함

▶ (투자유치) 2019년 2억 7천만 달러가 넘는 시리즈 E 투자를 유치하여 24억 달러 이상의 기업가치로 평가받고 있음

〈Cerebras 투자 유치 내역〉

투자 일시	투자 라운드	투자 유치 금액	리드 투자자
2016년 05월	Series A	2,700만 달러	Foundation Capital,
2016년 12월	Series B	2,500만 달러	Benchmark
2017년 01월	Series C	6,500만 달러	-
2018년 11월	Series D	8,100만 달러	-
2019년 11월	Series E	2억 7,200만 달러	-

자료: Crunchbase 웹페이지, Cerebras Systems, https://www.crunchbase.com/organization/cerebras-systems

◈ **(대표 제품) Wafer Scale Engine (WSE)와 WSE 기반의 Cerebras System(CS)가 주요 제품군**

▶ (WSE) 컴퓨팅, 메모리, interconnect로 이루어져 있으며, 웨이퍼 하나와 유사한 크기를 가지며, 2019년 8월 1세대 WSE-1, 2021년 3사분기 2세대인 WSE-2를 출시

〈Cerebras의 WSE 세대별 비교〉

	WSE-1	WSE-2	Increase
AI 코어 개수	400,000	850,000	2.13x
Manufacturing	TSMC 16nm	TSMC 7nm	-
출시일	2019년 8월	2021년 Q3	-
다이 크기	46,225 mm2	46,225 mm2	-
트랜지스터 개수	1조 2,000억개	2조 6천억개	2.17x
집적도	25.96 mTr/mm2	56.246 mTr/mm2	2.17x
On-board SRAM	18 GB	40 GB	2.22x
Memory Bandwidth	9 PB/s	20 Pb/s	2.22x
Fabric Bandwidth	100 Pb/s	220 Pb/s	2.22x

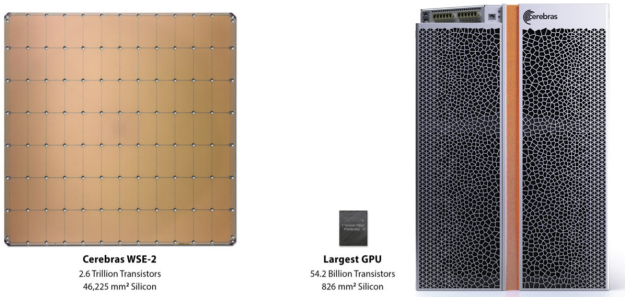
자료: Anandtech('21.4)

▶ (Cerebras System) WSE를 탑재한 Cerebras의 AI 컴퓨터

- (CS-2) 2세대 Cerebras System으로 26인치의 높이를 보유하여 표준 데이터센터 렉 규격의 1/3 크기이고, 12개의 100Gbit 이더넷 lane을 탑재하고 있으며, 40GB의 on-chip memory를 통해 시간당 높은 연산량을 달성

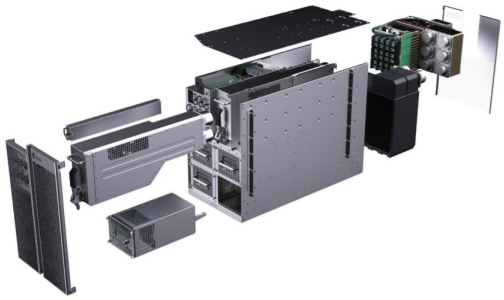
※ CS-2는 단순한 구조로 조립이 간편하게 디자인되어 서버 설치 과정이 수 분 내에 가능

〈Cerebras WSE-2(좌), CS-2(우)〉



자료: Cerebras (2021.4)

〈Cerebras WSE-2(좌), CS-2(우)〉



자료: Cerebras (2021.4)

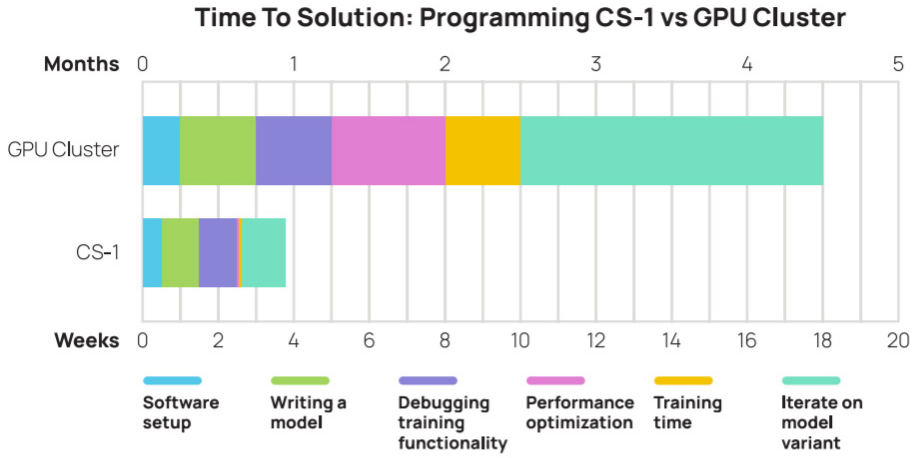
◈ **(소프트웨어 스택) Cirrascale사와 클라우드 서비스를 공동으로 제공하여 하드웨어 구매 없이 클라우드 상에서 딥러닝 모델을 구동 가능**

▶ (Cerebras Graph Compiler) Tensorflow, Pytorch 플랫폼의 코드를 최소한의 수정만으로 CS-2 상에서 딥러닝 모델을 구동할 수 있도록 하는 Cerebras Graph Compiler(CGC) 제공

- Tensorflow, Pytorch 플랫폼의 코드를 기반으로 CS-2에서 딥러닝 모델을 구동 가능한 명령어를 생성하는데, 사용자의 모델을 최적의 단위로 파티셔닝 및 kernel size 조정 등을 통해 많은 코어에서 동시에 연산을 수행하도록 하여 높은 병렬성 확보
- CGC를 통해 개발 시에 많은 시간이 소요되는 성능 최적화(performance optimization) 및 트레이닝(training)에 소요되는 시간을 획기적으로 줄였으며, 이와 더불어 모델의 미세조정 시간(iterate on model variant) 또한 감소

※ 고객사의 BERT 자연어 처리 모델 개발 완성에 소요되는 시간을 4개월에서 4주 이내로 감소(Cerebras "Programming at Scale" White paper)

〈CS와 GPU를 활용한 BERT 모델 개발 시간 비교〉



자료: Cerebras(2021.6)

9. Hailo

◈ **(기업 개요) 이스라엘의 edge AI 디바이스 개발 업체로 자사의 디바이스에 활용하기 위한 AI 하드웨어 개발 또한 활발하게 진행**

▶ (투자 유치) 2021년 6월 1억 달러 규모의 시리즈 C 투자유치를 통해 이스라엘의 유니콘 기업으로 발돋움

- 현재까지 21개 이상의 투자자에게서 투자 유치한 금액이 1억 8천만 달러가 넘으며, 기업가치는 10억 달러를 넘는 것으로 평가

〈Hailo 투자 유치 내역〉

투자 일시	투자 라운드	투자 유치 금액	리드 투자자
2017년 06월	Seed Round	350만 달러	N/A
2018년 06월	Series A	1,250만 달러	N/A
2019년 01월	Grant	300만 유로	N/A
2019년 01월	Series A	850만 달러	Glory Ventures
2020년 05월	Series B	6,000만 달러	N/A
2021년 06월	Series C	1억 달러	N/A

자료: Crunchbase 웹페이지, Hailo, https://www.crunchbase.com/organization/hailo-technologies

◈ (대표 제품) 26TOPS의 연산 능력으로 다른 경쟁사의 엣지 프로세서 대비 월등한 성능을 갖는 Hailo-8 제품군이 있음

▶ NVIDIA의 엣지 프로세서 Xavier와 Jetson Nano와 비교해도 연산 능력(FPS) 및 전력 소모당 연산 능력(FPS/W)에서도 우위를 보임

〈Hailo-8 제품군〉

제품명	특징
Hailo-8™ AI Processor	Full HD 실시간 연산 가능, 초당 26 테라 연산 Typical 모드 전력 소모 2.5 W
Hailo-8™ M.2 AI Acceleration Module	PCI 인터페이스를 통해 Hailo-8 Processor를 사용할 수 있는 가속 모듈, NGFF M.2 폼팩터와 호환, 초당 26 테라 연산
Hailo-8™ Mini PCIe AI Acceleration Module	PCI Express Mini 폼팩터와 호환, 초당 13 테라 연산
Hailo-8™ Evaluation Board	Hailo-8 AI 프로세서를 개발, 테스팅, 디버깅할 수 있는 모듈

자료: Hailo 웹페이지, https://hailo.ai/

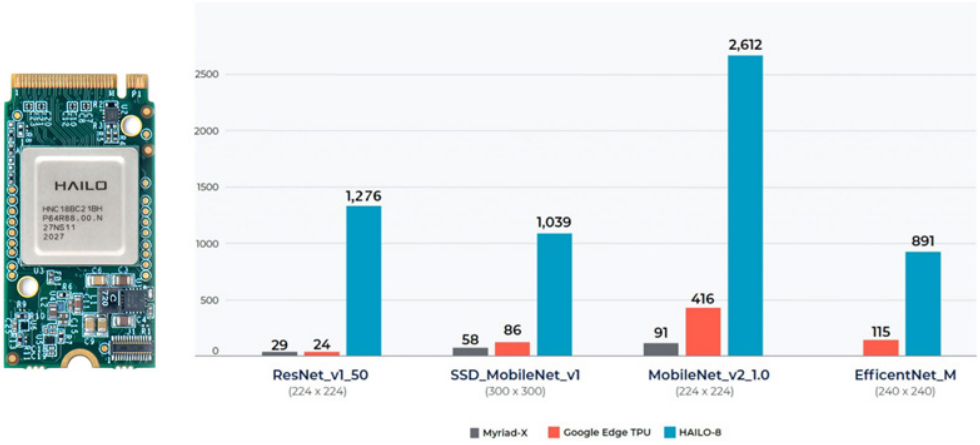
〈Hailo-8 AI 프로세서(좌), NVIDIA Edge 프로세서 Xavier NX, Jetson Nano와의 성능 비교(우)〉

자료: Hailo 웹페이지, https://hailo.ai/

▶ 이외에도 활용성을 위해 Hailo-8 프로세서를 PCI를 통해 연결하는 두 가지 모듈 M.2 AI Acceleration Module과 Mini PCIe AI Acceleration Module 출시

- (M.2 AI Acceleration Module) 경쟁사의 제품군인 Myriad-X(인텔), Edge TPU(구글) 대비 연산 능력(FPS)에서 우위

〈Hailo M.2 AI 가속 모듈(좌), Myriad-X(인텔), Edge TPU(구글)와의 성능 비교(우)〉



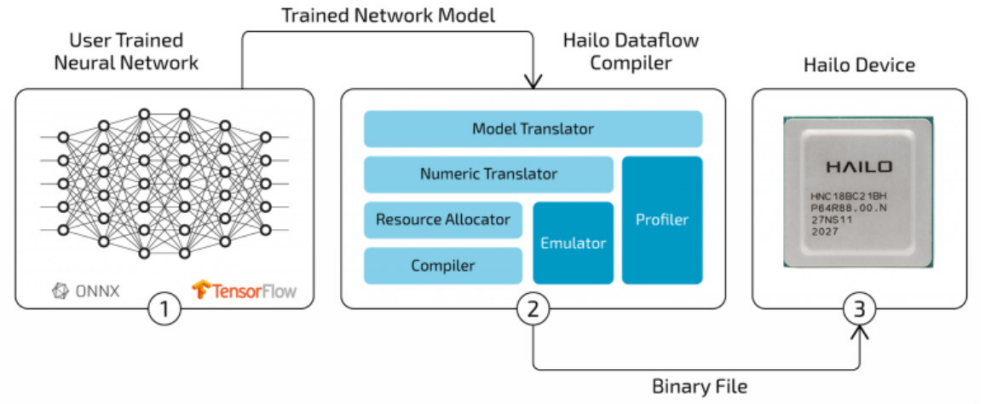
자료: Hailo 웹페이지, https://hailo.ai/

◈ **(소프트웨어 스택) Hailo-8 프로세서 구동을 위한 Dataflow Compiler 등의 AI SDK tool chain 제공**

▶ (Dataflow Compiler) 개발자들이 주로 사용하는 Tensorflow와 ONNX를 지원하는 컴파일러

- (Model translator) ONNX, Tensorflow 상의 AI 모델을 Hailo 구동 형식으로 변환
- (Numeric Translation) 딥러닝 오퍼레이터를 Hailo 상의 표현으로 변환하기 위해 최신 quantization algorithm 적용
- (Resource Allocation) Hailo device의 물리적 리소스에 사용자의 모델을 할당
- (이외) 분석툴과 디버그 툴도 Compiler에 포함되며, 이 중 Emulator는 chip 구동을 위한 bit 단위 에뮬레이션을 제공하며, Profiler는 FPS, power, latency 등 chip의 성능을 시뮬레이션하여 결과를 제공하는 역할을 수행

〈Hailo Dataflow Compiler〉



자료: Hailo 웹페이지, https://hailo.ai/

[선도기업 성공 요인 분석]

◈ **(Intel/Habana Labs) 인공지능 모형의 학습과 추론에 통일된 아키텍쳐 지원**

▸ 인텔은 기존에 개발하던 너바나 AI 하드웨어의 개발을 중단하고 Habana Labs의 AI 반도체의 개발에 집중(2020년 2월)

– (너바나 칩 단종) 2016년 너바나 인수 후 Neural Network Processor(NNP) 학습용 칩(NNP-T)과 추론용 칩(NNP-I)을 공개하였으나(2019년 11월), 같은 해에 인수한 하바나 칩 대비 성능이 떨어지고, 아키텍쳐의 확장성에 한계가 있어 출시 3개월 후 단종

▸ (하바나 제품군) 하바나 제품군은 학습, 추론에 통일된 아키텍쳐를 채택하여, 단일 하드웨어 아키텍쳐와 소프트웨어 스택만으로도 학습, 추론이 가능하며 추후 소프트웨어 스택 등의 수정 등이 필요할 때 딥러닝 개발 엔지니어들의 요구에 빠르게 대처 가능

▸ (AWS, 보이저 슈퍼컴퓨터) 인텔의 하바나 제품군으로의 선택과 집중으로 인해 아마존 웹 서비스(AWS), 샌디에이고 슈퍼컴퓨터 센터의 보이저 슈퍼컴퓨터에 하바나 인공지능 반도체가 채택되는 성공을 거둠

◈ **(SambaNova, GraphCore, Google) 완성도 높은 하드웨어-소프트웨어 생태계 지원**

▸ (소프트웨어 스택) 인공지능 반도체의 실수요자들인 인공지능 애플리케이션 개발자 및 관련 기업들은 하드웨어뿐만 아니라 하드웨어의 구동을 지원하는 소프트웨어 스택이 필요

– (높은 완성도의 소프트웨어 스택) SambaNova, GraphCore는 독자적으로 보유한 소프트웨어 스택인 SambaFlow, Poplar를 바탕으로 수요자들의 활용성을 높여 시장에서 좋은 평가를 받고 있음

- (Google) 구글은 Tensorflow와 연동된 XLA 컴파일러, HLO 중간 표현 등 소프트웨어 스택을 지속적으로 개발 중이며, 이와 더불어 클라우드 서비스를 통해 TPU의 접근성을 제고

◈ **(Tesla) 확장성 높은 아키텍쳐**

▶ (초대형 인공지능 모델의 수요 증가) 트랜스포머 등 대형 딥러닝 모델의 성능이 입증되며, 초대형 모델의 개발이 지속적으로 추진되고 있으며, 이는 높은 확장성을 가진 아키텍쳐에 대한 수요를 증가시킬 것으로 전망

▶ (Dojo) Tesla의 Dojo는 높은 확장성으로 사용자가 원하는 수준의 연산량을 유연하게 달성 가능

- (구성) 가장 작은 트레이닝노드 354개로 하나의 D1 칩을 구성하며, 25개의 D1 칩으로 하나의 타일을, 12개의 타일로 하나의 캐비넷을, 11개의 캐비넷으로 1.1 ExaFLOPS의 연산량을 달성

시사점

◈ **인공지능 모델의 활용 범위가 전산업으로 확장됨에 따라 인공지능 반도체는 세계 최고의 디지털 역량을 갖춘 국가로 도약하기 위한 핵심 범용기술로 부각**

※ (대한민국 디지털 전략) 대한민국 정부는 AI 반도체를 '6대 디지털 혁신기술' 중 하나로 선정하고, R&D 자원의 집중 투자를 천명(2022.9)

▶ 인공지능 반도체의 2022년 글로벌 시장규모는 전년대비 27.8% 증가한 444억 달러로 예상되며, 연평균 19.9%씩 증가하여 2026년에는 861억 달러에 이를 것으로 전망됨

- 기술발전의 포화에 따라 가격 경쟁 단계로 접어들어 시장규모가 축소하고 있는 메모리 반도체와 달리, 시스템반도체는 방대한 데이터의 활용이 기존 PC 등에서 자동차, 모바일, 생활가전, 산업가전 등 전 범위의 전자제품으로 확장되면서 수요처 다변화와 동시에 시장규모 또한 지속적으로 증가 전망

- 시스템반도체 중 인공지능 반도체의 비중 또한 2020년 8.5%에서 2025년에는 19.5%로 확대 전망(Gartner, 2021.6)

▶ 주요국이 자국 위주의 반도체 공급망 구축과 인공지능 반도체 산업 육성을 위해 국가적 노력을 기울이는 상황에서 글로벌 인공지능 반도체 선도기업의 성공 요인을 벤치마킹하여 국내 인공지능 반도체 전문 기업 육성을 위한 전략을 마련해야 함

- (하드웨어 – 소프트웨어 기업 협력 지원) 반도체 설계를 전문으로 하는 기업과 컴파일러, 드라이버, 라이브러리 등의 기술을 갖춘 소프트웨어 기업 간의 협업을 지원할 필요

  ※ NVIDIA GPU를 위한 라이브러리의 출현으로 인해 소프트웨어 전문 기업은 컴파일러 등의 소프트웨어 스택 개발 시에 NVIDIA GPU를 딥러닝 가속기 하드웨어로 타겟팅하여 개발에 착수하였고, 이렇게 개발된 소프트웨어 스택은 NVIDIA GPU의 인공지능 반도체 내의 시장점유를 증가시킴

- (소프트웨어 스택 산학연 지원) 소프트웨어 스택 부문 국내 경쟁력이 미국 등의 선도국 대비 낮으므로 해당 분야의 기술력 증진을 위한 산학연 지원체계를 구축

- (대형 인공지능 모델을 위한 반도체 R&D 지원) 추후 대형 인공지능 모델의 활용성이 높아질 것으로 전망됨에 따라 AI 반도체의 확장성 확보를 위한 R&D를 지원하여 대형 인공지능 모델의 수요에 대비

참고문헌

[국내문헌]

정보통신정책연구원(2021), “인공지능 반도체 산업 확산 가속화 방안”

[해외문헌]

Anandtech(2021. 4. 20). “Cerebras Unveils Wafer Scale Engine Two (WSE2): 2.6 Trillion Transistors, 100% Yield”, https://www.anandtech.com/show/16626/cerebras-unveils-wafer-scale-engine-two-wse2-26-trillion-transistors-100-yield.

Cerebras(2021. 4). “Cerebras CS-2 Whitepaper”, https://cerebras.net/wp-content/uploads/2021/04/Cerebras-CS-2-Whitepaper.pdf.

Cerebras(2021. 6). “Deep Learning Programming at Scale Whitepaper”, https://f.hubspotusercontent30.net/hubfs/8968533/Cerebras-Whitepaper_ProgrammingAtScale_V3.1.pdf.

Gartner(2022.05), “Forecast:AI Semiconductors, Worldwide, 2020-2026”

Jouppi, N. P., Yoon, D. H., Ashcraft, M., Gottscho, M., Jablin, T. B., Kurian, G., ... & Patterson, D. (2021, June). Ten lessons from three generations shaped Google's TPUv4i: Industrial product. In2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)(pp. 1-14). IEEE.

McKinsey (2018), Artificial-intelligence hardware: New opportunities for semiconductor companies, https://www.mckinsey.com/~/media/McKinsey/Industries/Semiconductors/Our%20Insights/Artificial%20intelligence%20hardware%20New%20

opportunities%20for%20semiconductor%20companies/Artificial-intelligence-hardware.ashx

SambaNova(2021. 6). “SambaNova RDA Whitepaper”, https://sambanova.ai/wp-content/uploads/2021/06/SambaNova_RDA_Whitepaper_English.pdf.

[웹페이지]

Crunchbase, Cerebras Systems, https://www.crunchbase.com/organization/cerebras-systems

Crunchbase, GraphCore, https://www.crunchbase.com/organization/graphcore

Crunchbase, Hailo, https://www.crunchbase.com/organization/hailo-technologies

Crunchbase, Kneron, https://www.crunchbase.com/organization/kneron

Crunchbase, NVIDIA https://www.crunchbase.com/organization/nvidia/company_financials

Crunchbase, SambaNova Systems, https://www.crunchbase.com/organization/sambanova-systems

Graphcore, https://www.graphcore.ai/

Google Cloud TPU, http://cloud.google.com/tpu

Kneron, https://www.kneron.com/

Habana, An Intel Company, https://habana.ai/

Habana, An Intel Company, Gaudi Synapse AI Software Suite, https://https://habana.ai/training-software/

Hailo, https://hailo.ai/

NVIDIA, How Suite It is: Nvidia and VMware Deliver AI-Ready Enterprise Platform, https://blogs.nvidia.com/blog/2021/03/09/vmware-ai-ready-enterprise-platform/

SambaNova Systems, https://sambanova.ai/

Tesla AI Day, https://www.youtube.com/watch?v=j0z4FweCy4M