

# 인공지능 프로세서 반도체 기술 및 표준화 동향

최민석 한국전자통신연구원 인공지능프로세서연구실 책임연구원

한진호 한국전자통신연구원 인공지능프로세서연구실 책임연구원/실장

## 1. 머리말

제4차 산업혁명의 핵심 기술인 빅데이터, 인공지능 등이 산업에 널리 활용되고 있으며 이를 구현하는 데 필요한 기반기술인 인공지능 반도체가 시스템 반도체의 새로운 기회요인으로 대두되고 있다. 인공지능 반도체는 학습·추론 등 인공지능 구현에 요구되는 대규모 데이터 처리를 위한 기존의 폰노이만 구조 반도체의 한계를 극복하기 위해 고성능·저전력 기술 중심으로 발전을 가속화하고 있다. 본고에서는 2020년 2월 기준으로 인공지능 프로세서 반도체 기술동향 및 국내·국제 표준화 현황에 대해 살펴보고자 한다.

## 2. 인공지능 프로세서 반도체

### 2.1 정의 및 분류

인공지능 반도체란 인공신경망 알고리즘을 효율적으로 계산할 수 있는 반도체로 정의할 수 있다. 인공지능 기술의 핵심기술 중 학습·추론 기술을 구현하기 위한 데이터 연산처리를 저전력, 고속처리하여 효율성에 특화된 반도체를 의미한다.

인공지능 반도체는 인공지능 시스템 구현 목적에 따

라 크게 학습용(Training)과 추론용(Inference)으로 구분할 수 있으며, 두 가지 과정을 반복 실행하여 최적의 답을 찾도록 성능을 강화하는 데 주로 사용한다.

현재 인공지능 학습·추론은 대부분 데이터 센터에서 실행되고 있다. 인공지능 서비스에 요구되는 대규모 연산 처리 성능을 확보하기 위해 인공지능 반도체를 서버에 장착하여 활용하고 있다. 하지만 방대한 양의 데이터 연산 처리로 인한 발열 및 전력소모로 인해 인공지능 반도체의 효율성 개선이 필요하다.

또한, 스마트폰, IoT 기기 등의 보급 확산 및 클라우드 기술의 발전에 따라 디바이스의 추론 기능에 대한 수요가 증가하면서, 데이터 센터 서버(클라우드)와의 연결을 최소화하고 디바이스 자체에서 인공지능 연산이 가능하도록 소형화·저전력·고성능 중심의 인공지능 반도체 기술 개발이 가속화되고 있다. 최근에는 실시간 AI 처리, 네트워크 트래픽 감소, 개인정보 보호 및 클라우드 연결이 불가능한 경우 등에 대비해 에지 디바이스상에서의 온칩 학습(On-chip Learning)에 대한 요구가 증대되고 있다.

인공지능 반도체는 기술적으로 ‘기존 반도체 진화형’, ‘1세대 AI 반도체’, ‘2세대 AI 반도체’의 3가지 유형으로 분류할 수 있다.

### 〈표 1〉 인공지능 반도체의 분류

시스템 구현 목적	학습용	추론용	
서비스 플랫폼	서버용	에지 디바이스용	
기술 구현 방식	기존 반도체 친화형	1세대 AI 반도체	2세대 AI 반도체
	CPU/GPU/FPGA	ASIC/ASSP	뉴로모픽 프로세서

‘기존 반도체 진화형’ 인공지능 반도체는 CPU·GPU·FPGA 등이 해당되며, 상대적으로 가격이 싸고 유연성이 높으나 인공지능 연산 성능과 소비전력 효율이 낮다는 단점이 있다. 대표적인 업체로는 인텔(Intel), 엔비디아(NVIDIA), 자이링스(Xilinx) 등이 있다.

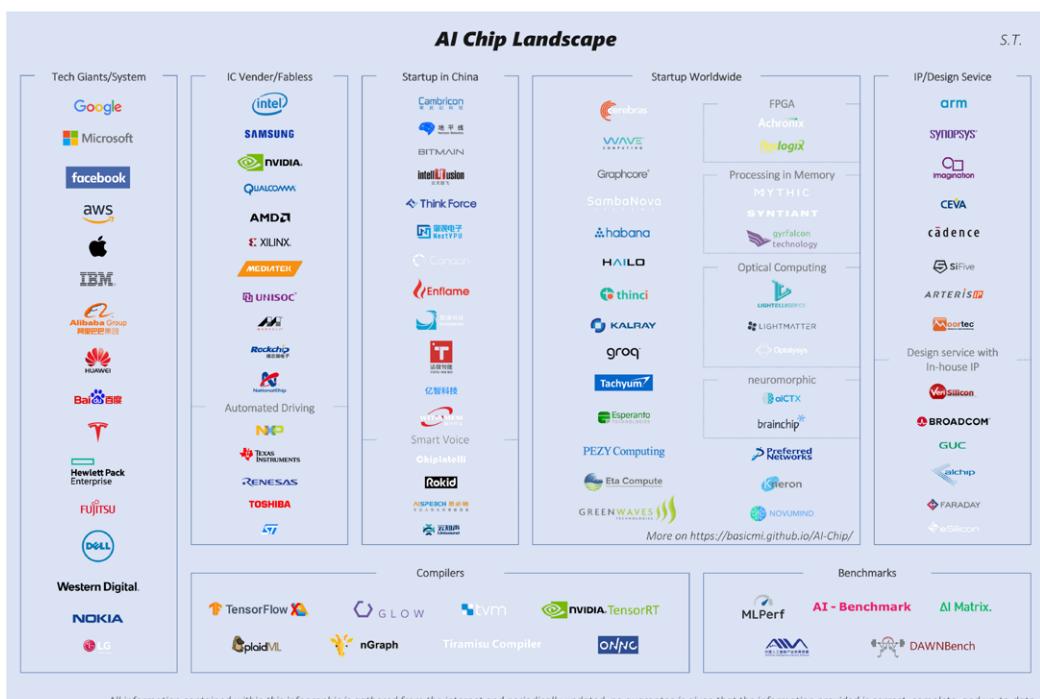
'1세대 AI 반도체'는 인공지능 연산 고속화를 위해 반도체 구성을 최적화시킨 ASIC/ASSP가 해당되며, 인공지능 연산 성능과 소비전력 효율이 높지만, 가격이 비싸고 유연성이 낮아 디자인된 알고리즘으로만 사용할 수밖에 없다는 단점이 있다. 대표적인 업체로는 구글, 인텔(모빌아이, 모비디우스) 등이 있다.

‘2세대 AI 반도체’는 인간의 뇌를 모방한 非폰노이만

방식의 뉴로모픽 반도체로 인공지능 연산 성능과 소비 전력 효율은 인공지능 반도체 가운데 가장 진화했지만 아직 기술 성숙도가 낮고 폰노이만 구조를 사용하지 않아 범용성이 낮은 특징이 있다. IBM의 TrueNorth의 경우, 초당 46억 회 실행되는 시냅스의 동작을 70~200mW 수준 소비전력으로 수행할 수 있다.

## 2.2 시장 전망

제4차 산업혁명의 핵심동력인 AI·자율주행·가상현실·드론 등 최첨단 기술이 빠르게 발전하면서 이를 안정적으로 구현할 수 있는 인공지능 반도체가 IT 산업의 핵심기술로 부상하고 있다. 이에 따라 기존의 반



*All information contained within this infographic is gathered from the internet and periodically updated, no guarantee is given that the information provided is correct, complete, and up-to-date.*

※ 출처: AI Chips (ICs and IPs), 2019.12. <https://basicmi.github.io/AI-Chip/>

[그림 1] 인공지능 반도체 생태계 현황

도체 업체는 물론, 인터넷, 스마트폰 등 다양한 분야의 업체들도 AI 반도체 개발에 뛰어들고 있다. 시장조사기관인 Allied Market Research에 따르면, 글로벌 인공지능 반도체 시장은 2018년 66억 3,800만 달러(약 7조 700억 원)에서 2025년 911억 8,500만 달러(약 106조 원)에 이르며, 동 기간 45.5%의 CAGR(연평균 증가율)을 기록할 것으로 전망하고 있다.

### 3. 인공지능 프로세서 반도체 기술 동향

#### 3.1 인공지능 반도체의 기술 배경

빅데이터를 통한 인공지능 학습 데이터 확보, 딥러닝 알고리즘 개발에 따른 인공지능 정확도의 비약적 향상과 고속 병렬처리가 가능한 GPU 기반 HW 컴퓨팅 파워에 힘입어 인공지능 기술이 급진전하고 인공지능의 상용화가 시작됐다. 그러나 인공지능 서비스를 제공하는 데 필요한 대량의 학습·추론 데이터 연산 처리에 따른 발열과 전력소모로 인해 인공지능 서비스 환경은 주로 데이터 센터 형태의 클라우드 시스

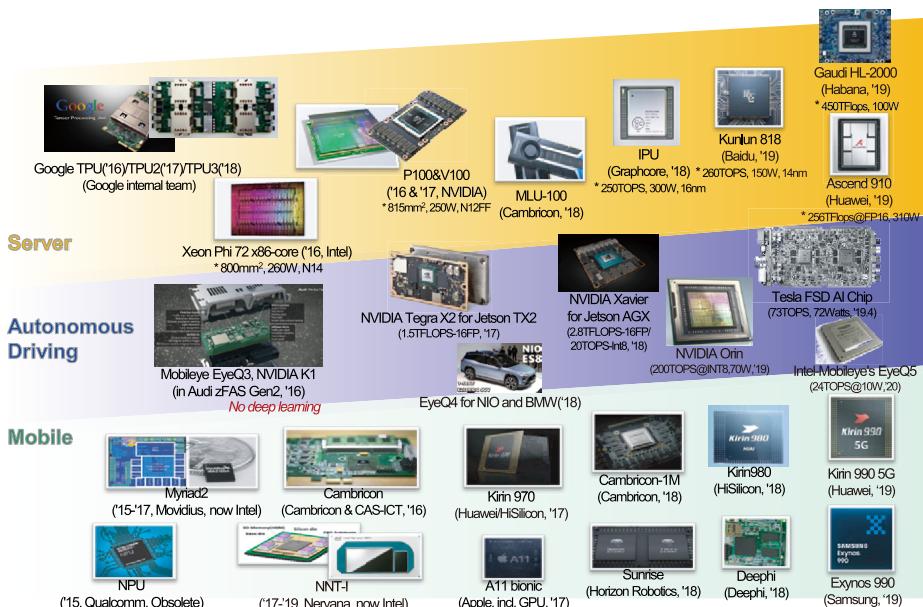
템으로 국한되었다. 이에 인공지능 반도체의 에너지 효율 향상과 예지 AI 컴퓨팅의 필요성이 커짐에 따라 인공지능 반도체 기술 혁신을 촉구하게 되었다.

#### 3.2 글로벌 ICT 기업의 인공지능 반도체 개발 동향

인텔, 페이스북, 화웨이, 삼성전자 등 글로벌 ICT 기업들은 차세대 인공지능 반도체 개발을 위한 기술 개발 및 투자 확대를 통해 독자적인 AI 생태계를 구축해나가고 있다.

구글의 경우, AI 연구책임자 제프 딘(Jeff Dean)이 샌프란시스코에서 열린 'ISSCC 2020' 국제학회에서 'AI를 활용한 AI 반도체 개발'을 시도하고 있다고 발표하였는데, AI를 활용해 AI 반도체 성능을 개선시키면 개선된 AI 반도체가 SW·HW 성능을 향상시키고 다시 AI의 성능을 높이는 선순환이 가능하다고 설명했다.

인텔은 모비디우스(Movidius), 너바나(Nervana), 알테라(Altera), 모빌아이(Mobileye)에 이어 이스라엘 AI 반도체 스타트업 '하바나랩스(Habana Labs)'를 20억 달러에 인수(2019.12)하는 등 AI 반도체 시



[그림 2] 인공지능 프로세서 반도체 개발 현황

장 확장을 가속화하고 있는데, 하바나랩스는 엔비디아보다 3배 뛰어난 성능을 지닌 추론 프로세싱 AI칩 ‘고야(Goya)’와 데이터 센터 적용을 위한 서버용 AI 훈련용 반도체 ‘가우디(Gaudi)’를 개발하였다.

페이스북은 AI 반도체 개발 조직을 설립(2019.04)하고 AI 프로그램을 지원하기 위한 자체 SoC, ASIC 개발에 착수하였다.

화웨이는 AI 모델의 훈련속도를 향상시킨 AI 반도체 ‘어센드(Ascend) 910’을 출시(2019.08)하고 경쟁사 제품 대비 25%가량 성능이 뛰어난 ‘쿤펑(Kunpeng) 920’을 공개(2019.01)하는 등 독자적인 AI 생태계를 구축하고 있다. 스마트폰에 탑재하는 자체 AI 반도체 ‘기린(Kirin)’ 시리즈도 지속적으로 업데이트하여 2020년 상반기 ‘기린1000(Kirin1000)’을 출시할 계획이며, 궁극적으로 AI 반도체 기술력을 확보해 스마트

폰과 소프트웨어에 최적화된 기술을 구현하여 AI뿐만 아니라 SW·HW에 이르는 생태계 강화 전략을 추진 중이다.

삼성전자는 신경망 처리 장치(NPU)를 탑재한 ‘엑시노스(Exynos)’ 시리즈를 개발하는 등 독자적인 AI 반도체 개발 행보를 지속하고 있으며 2019년 10월에는 2세대 자체 NPU 코어 2개와 DSP를 탑재해 10TOPS(초당 1조 회) 이상의 AI 연산 성능을 확보한 모바일 AP ‘엑시노스 990(Exynos 990)’을 공개하였다.

### 3.3 AI 스타트업 기업의 인공지능 반도체 개발 동향

최근 AI 솔루션 스택 전반에 걸쳐 AI 스타트업의 수가 급격히 증가하고 있는데, 이들은 특정 AI 활용에 중점을 둔 확장 가능한 플랫폼을 구축하고 있다. 벤처 캐피탈 자금 순위 상위 19개 AI 반도

**(표 2) 주요 AI 스타트업의 현황**

스타트업	설립	본사	전략적 투자자들	AI 기술
Horizontal Robotics	2015	Beijing, China	SK Hynix, SK China	Vision DSP
Graphcore	2016	Bristol, UK	Microsoft, BMW	Deep-learning processor
Cambricon Technologies	2016	Beijing, China	SDIC	Deep-learning processor
Wave Computing	2010	Campbell, CA	Samsung	Deep-learning processor
Vicarious	2010	San Francisco, CA	Samsung	Neuromorphic processor
Rigetti Computing	2013	Berkeley, CA		Optical/quantum AI computing
Cerebras	2016	Los Altos, CA		Deep-learning processor
Vayyar	2011	Yehud, Israel		Vision DSP
ThinkForce	2017	Shanghai, China		AI acceleration engine
Movidius	2006	San Mateo, CA	Intel (Acquired)	Neural compute engine accelerator (Appl: Vision DSP)
Mythic	2012	Redwood City, CA Austin, TX		Neuromorphic processor
KnuEdge	2005	San Diego, CA		Neuromorphic processor
Nervana	2014	San Diego, CA	Intel (Acquired)	Deep-learning processor
Xanadu	2016	Toronto, Canada		Optical/quantum AI computing
Reduced Energy Microsystems	2014	San Francisco, CA		Deep-learning processor
LightOn	2016	Paris, France		Optical/quantum AI computing
CyberSwarm	2017	San Mateo, CA		AI-assisted cybersecurity CPU
Tenstorrent	2016	Toronto, Canada		Deep-learning processor
Vathys	2015	Portland, OR		Deep-learning processor

※ 출처: PwC research, 2019, 자료 편집, 벤처 캐피탈 투자금액 순서로 정렬

체 스타트업 중 11개 업체가 미국에 본사를 두고 있으며 대부분 다양한 AI 및 딥러닝 워크로드에 맞게 설계된 특별한 프로세서 아키텍처를 개발하고 있다. 그중 호라이즌로보틱스(Horizon Robotics), 그래프코어(Graphcore), 캄브리콘(Cambricon Technologies), 웨이브컴퓨팅(Wave Computing), 세레브라스(Cerebras), 모비디우스(Movidius), 너바나(Nervana) 등의 9개 업체는 딥러닝 프로세서를 개발하고 있으며, 비카리우스(Vicarious), 미씨(Mythic), 누엣지(KnuEdge)는 인간 뇌 기능을 모사하기 위한 새로운 뉴로모픽 프로세서 아키텍처를 개발 중이다. 모비디우스와 너바나는 이미 인텔에 인수되었으며, 삼성전자의 경우 웨이브컴퓨팅, 비카리우스에, SK하이닉스의 경우, 호라이즌로보틱스에 전략적 투자를 아끼지 않고 있다.

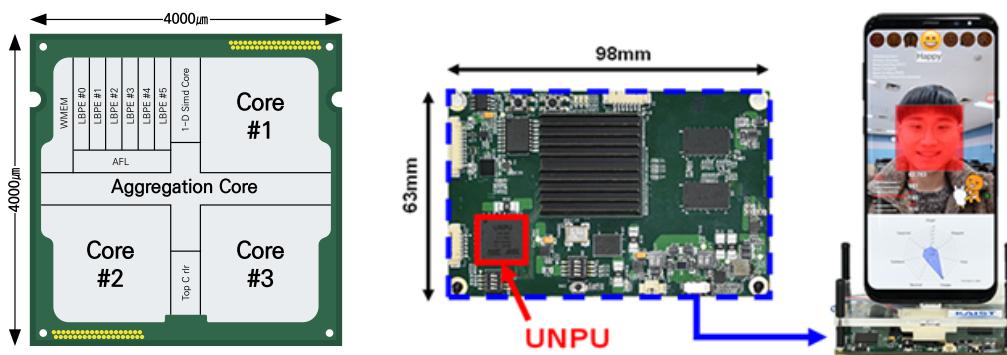
### 3.4 국내 인공지능 반도체 개발 동향

국내의 경우, AI 반도체를 개발 중인 기업 및 기관이 10여 곳으로 미국·중국과 비교할 때 산업 저변이 매우 열악한 상황이다. 국내 기업 대부분 예지 디바이스용 반도체를 개발하고 있으며, ASIC/ASSP 개발 비중이 높은 미국·중국과 달리 뉴로모픽 반도체 개발 사례가 많다는 점이 특징이다. 대기업 중에서는 삼성전자가 스마트폰용 일체형 AI 반도체 상용화에

이어 삼성종합기술원을 중심으로 뉴로모픽 AI 반도체 선행연구를 추진 중이며, 국내 반도체 패키징 업체인 네페스는 미국 뉴로모픽칩 개발업체 제너럴비전(General Vision)과 생산·판매에 관한 글로벌 독점권 계약을 체결, 2018년 1월부터 뉴로모픽칩 양산을 시작하였다. 팹리스업체인 네셀은 셀(XELL)로 명명한 AI 반도체 설계자산(IP)을 개발 중이며 셀 NPU의 성능은 엔비디아 Tesla K80(8.74TFLOPS, 150W)이나 구글 TPU2(45TFLOPS, 75W)보다 우수한 수준이라고 평가된다. 학계에서는 서울대, KAIST, POSTECH, UNIST가 공동으로 뉴럴프로세싱연구센터(NPRC)를 개설하여 뉴로모픽 반도체 개발을 추진하고 있다.

KAIST 유희준 교수 연구팀은 팹리스 스타트업 유엑스팩토리와 공동으로 회선 신경망(CNN)과 재귀 신경망(RNN)을 동시에 처리할 수 있고, 인식 대상별로 에너지 효율과 정확도를 다르게 설정할 수 있는 모바일용 AI 반도체를 개발했다고 발표했다. 세계 최고 수준 모바일용 AI 반도체 대비 CNN과 RNN 연산 성능이 각각 1.15배, 13.8배이며, 에너지 효율도 40% 높다고 설명한다.

한국전자통신연구원(ETRI)에서는 세계 최고 수준인 40TFLOPS급의 고성능·저전력 시스톨릭 어레이 구조의 매니코어 아키텍처 기반 복합 인공지능 프로



※ 출처: KAIST, 모바일 기기용 딥러닝 인공지능 칩(UNPU) 개발, 인공지능신문, 2018.02.

[그림3] KAIST 유희준 교수 연구팀이 개발한 UNPU 칩과 이를 이용한 감정인식시스템

세서 반도체 칩인 AB9과 0.45W급의 초저전력 인공지능 프로세서 반도체 칩 VIC2를 개발하였다.

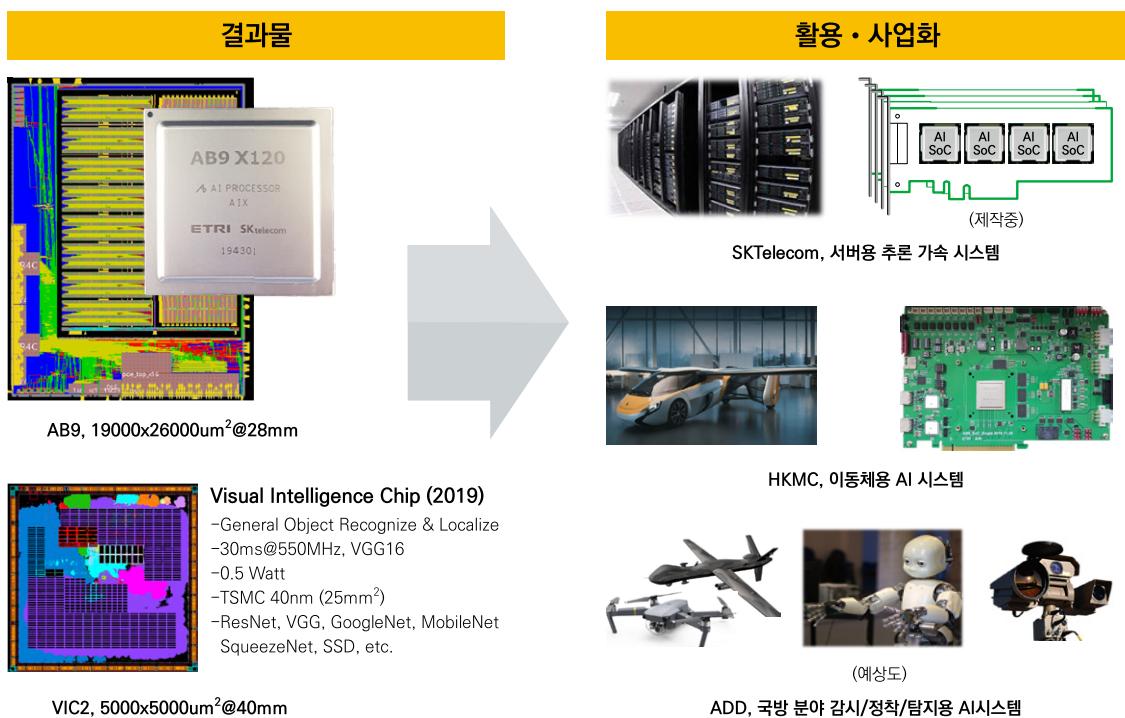
영상인식, 음성인식을 포함하는 고성능 딥러닝 인공지능 연산 반도체인 AB9는 범용성을 가진 인공지능 서버와 자율주행 차량 기술의 핵심 부품으로 사용 가능하다. 16,484개의 연산기로 구성된 AB9는 딥러닝 추론(inference) 연산 가속을 위한 고성능 인공지능 반도체로서 28nm 공정으로 제작되어 1.2GHz의 동작 주파수와 40TFLOPS의 성능을 보인다. 인간의 시각을 모사한 VIC2 칩은 단 0.45W의 저전력만 소비하면서도 4TOPS급 성능을 발휘하는, 에너지 효율이 높은 모바일 인공지능 반도체다. 작은 부피나 낮은 소비전력이 필요한 로봇, 드론, CCTV 등의 모바일 장치나 특별한 영상을 판별해야 하는 국방 분야 등에 적용되어 보다 편리하고 효율적인 인공지능 시스템 구현이 가능하다.

#### 4. 인공지능 프로세서 반도체 관련 표준화 현황

##### 4.1 인공지능 프로세서 반도체 관련 국내 표준화 현황

TTA TC 4 PG 417(지능형반도체 프로젝트그룹)에서는 인공지능 알고리즘을 효율적으로 수행하는 데 필요한 지능형반도체 분야 표준화를 추진하고 있다. 인공지능 알고리즘 수행 시 기능안전 문제를 해결하기 위해 ‘인공지능 프로세서의 기능안전 인터페이스’, ‘다중 칩 구성이 가능한 인공지능 프로세서 인터페이스’ 등 ISO-26262 및 ISO/PAS-21448에 준하는 지능형반도체 기능안전 표준이 제정되고 있다.

반도체공학회에서는 지능형반도체 포럼을 운영함으로써 프로세서 구조, 인터페이스, 스파이킹 뉴럴넷 하드웨어 분야의 지능형반도체 기술 표준화와 신경망 압축의 지능형반도체 신경망 기술 표준화를 추진하고 있다.



[그림 4] ETRI 인공지능 프로세서 반도체인 AB9, VIC2

#### 4.2 인공지능 프로세서 반도체 관련 국제 표준화 현황

전자부품의 기능안전성은 IEC 61508에 정의되어 있으며, 현재 IEC TC 47에서도 전자부품 및 반도체와 관련한 표준을 제정하고 있다.

전기·전자시스템의 기능안전성(Functional Safety)을 위한 국제표준은 ISO/TC 22/SC 32/WG 8에서 기존에 도출되어 있는 기능안전성 표준 내용을 망라하여 자동차 전장시스템에 적용하기 위해 ISO 26262 표준을 제정(2011)하였고, ISO 26262 2nd Edition(2018)에서는 반도체에 대한 기능안전성을 규정하기 위한 표준이 추가 제안되어 자동차용 지능형반도체 모두에 적용될 예정이다.

또한, ISO 26262 표준에서 다루지 않았던 의도된 기능의 설계 자체의 성능 한계에 대한 안전을 확보하기 위해 SAE 자율주행 레벨 1-2의 ADAS 및 자율주행을 위한 ‘의도된 기능에 대한 기능안전성(Safety of the Intended Functionality. SOTIF)’에 대한 별도의 표준 ISO/PAS 21448을 제정(2019)하였다.

ISO/PAS 21448 표준에 SAE 자율주행 레벨 3~5에 필요한 SOTIF의 기술적 내용을 다루기 위해 ‘Machine Learning’, ‘HD맵’, ‘AI 요구사항’, ‘Driving

Policy’ 등의 내용을 추가하여 정식 ISO 21448 표준으로 제정하기 위한 위원회 초안(CD, Committee Draft) 문서가 2020년 3월 투표를 통하여 국제표준안(DIS, Draft International Standard)을 제정하는 절차가 진행 중이다.

#### 5. 맷음말

인공지능 기술은 2010년을 전후로 빅데이터, 딥러닝 알고리즘, 반도체 하드웨어 및 컴퓨팅 기술의 급진 전에 힘입어 인공지능 상용화가 시작되어 그 활용 범위가 더욱 확대되고 있다. 최근 AI 기술의 진화는 고용량·고대역폭·고연산처리속도·저전력소모에 최적화된 인공지능 반도체 개발을 촉진하고 있다. 이러한 추세는 AI 생태계에서 반도체가 차지하는 중요성이 커지고 있음을 보여주고 있으며, 동시에 기존 반도체 생태계에서도 인공지능 반도체는 새로운 성장 동력으로 작용할 전망이다. 이러한 관점에서 AI 생태계와 반도체 생태계가 만나는 접점에는 AI 하드웨어인 AI 반도체가 존재하고, AI 반도체의 역할이 더욱 중요해지고 있다. 

### 참고문헌

- [1] 권영수, 전자통신기술동향, 인공지능프로세서 기술 동향(33권 5호), 2018.10.
- [2] ICT Spot Issue (2018-01호) 반도체 산업의 차세대 성장엔진, AI 반도체 동향과 시사점, 2018.03.
- [3] Global Artificial Intelligence Chip Market - Opportunities and Forecasts, 2019-2025, Allied Market Research, 2019.05.
- [4] Opportunities for the global semiconductor market - Growing market share by embracing AI, PwC, 2019.04.
- [5] 인공지능 기술의 진화와 AI 반도체·컴퓨팅의 변화에 대한 이해, 지능정보산업협회, 2020.02.
- [6] 글로벌 ICT 기업, 인공지능(AI) 반도체 개발 경쟁 활기, ICT Brief, 2020.02.
- [7] [Insight Report] 인공지능 반도체 산업 동향 및 이슈 분석, 한국전자통신연구원, 2017.12.
- [8] ISO 26262, 2nd Edition 'Road vehicles - Functional Safety', 2018.
- [9] ISO/PAS 21448, 'Safety of The Intended Functionality', 2019.
- [10] ISO/TC 22/SC 32/WG 8, ISO/CD 21448:2019, Road vehicles - Safety of the Intended Functionality, 2019.12.
- [11] ISO/TC 22/SC 32/WG 8 N 755-Result of Vote ISO CD 21448, 2020.03.

### 주요 용어 풀이

- AI(Artificial Intelligence): 인공지능
- IoT(Internet of Things): 사물인터넷
- Training: 학습, 딥러닝 등 기계 학습의 특정 작업을 수행하기 위해 방대한 데이터를 통해 반복적으로 지식을 배우는 단계
- Inference: 추론. 학습을 거친 최적의 모델을 통해 외부 명령을 받거나 상황을 인식하면 학습한 내용을 토대로 가장 적합한 결과를 도출하는 단계
- GPU(Graphics Processing Unit): 컴퓨터에서 그래픽스 연산 처리를 전담하는 반도체 코어 칩, 또는 장치
- FPGA(Field Programmable Gate Array): 이미 설계된 하드웨어를 반도체로 생산하기 직전 최종적으로 하드웨어의 동작 및 성능을 검증하기 위해 제작하는 중간 개발물 형태의 IC
- ASIC(Application Specific Integrated Circuit): 특정 용도의 대규모 집적 회로(LSI)
- ASSP(Application Specific Standard Product): 특정 용도 또는 응용을 대상으로 하는 전용 표준 제품의 집적 회로(IC)나 대규모 집적 회로(LSI)를 충칭하는 용어
- NPU(Neural Processing Unit): 신경망 처리 장치
- CNN(Convolutional Neural Network): 콘볼루션 신경망, 콘볼루션 계층(convolutional layer)과 통합 계층(pooling layer), 완전하게 연결된 계층(fully connected layer)들로 구성된 신경망
- RNN(Recurrent Neural Network): 순환 신경망, 시계열 데이터(time-series data)와 같이 시간의 흐름에 따라 변화하는 데이터를 학습하기 위한 심층 기계 학습(Deep learning) 모델
- 뉴로모픽(Neuromorphic) 반도체: 인간의 뇌 신경 구조를 모방하여 인간의 사고 과정과 유사하게 정보를 처리하는 반도체 소자
- ADAS(Advanced Driver Assistance Systems): 첨단 운전자 지원시스템
- SAE(Society of Automotive Engineers): 미국 자동차 기술자 협회
- SOTIF(Safety of The Intended Functionality): 의도된 기능에 대한 기능안전성으로 ADAS 시스템에서 센서의 인식 성능 한계에 따른 오작동에 대한 안전성이 대표적 예시