ETRI AI 실행전략 2: AI 반도체 및 컴퓨팅시스템 기술경쟁력 강화

ETRI AI Strategy #2: Strengthening Competencies in AI Semiconductor & Computing Technologies

최새솔 (S.S. Choi, saesol.choi@etri.re.kr) 지능화정책연구실 선임연구원
연승준 (S.J. Yeon, sjyeon@etri.re.kr) 지능화정책연구실 책임연구원/실장

ABSTRACT

There is no denying that computing power has been a crucial driving force behind the development of artificial intelligence today. In addition, artificial intelligence (AI) semiconductors and computing systems are perceived to have promising industrial value in the market along with rapid technological advances. Therefore, success in this field is also meaningful to the nation's growth and competitiveness. In this context, ETRI's AI strategy proposes implementation directions and tasks with the aim of strengthening the technological competitiveness of AI semiconductors and computing systems. The paper contains a brief background of ETRI's AI Strategy #2, research and development trends, and key tasks in four major areas: 1) AI processors, 2) AI computing systems, 3) neuromorphic computing, and 4) quantum computing.

KEYWORDS AI processor, AI computing system, Neuromorphic computing, Quantum computing, Hyper performance, Artificial Intelligence

Ⅰ. 서론

1. 배경 및 필요성

빅데이터, 인공지능(AI: Artificial Intelligence), 사물인터넷 등 지능화 기술이 빠르게 발전함에 따라 데이터 처리 수요가 폭증하고 있으며, 이에 따른 고속연산 능력에 대한 수요도 증대되고 있다. 시장조사기관 IDC(2018년)는 전 세계 데이터의 총 규모가 2025년에는 175제타바이트로 연평균 61% 증가할 것으로 예측한다[1]. 또한, AI 연구기업인

* DOI: https://doi.org/10.22648/ETRI.2020.J.350703
* 이 논문은 한국전자통신연구원 연구운영지원사업의 일환으로 수행되었음[20ZR1400, 국가지능화 기술정책 및 표준화 연구].
* 이 논문은 ETRI 기술정책연구본부 주관으로 담당 부서와의 워크숍 및 전문가 심층회의 등을 통해 수립된 'ETRI AI 실행전략'의 동향분석을 중심으로 작성되었다. 이 논문을 쓸 수 있도록 도움을 주신 ETRI 인공지능연구소 초성능컴퓨팅연구본부, 지능형반도체연구본부, ICT창의연구소 양자기술연구단 담당자분들께 감사드린다.

OpenAI는 AI 학습모델에 필요한 컴퓨팅 자원은 3.4개월마다 2배씩 증가한다고 전망한다[2].

지능화 혁명이 과거 정보화 혁명과 다른 점은 데이터의 양적 증가뿐만 아니라 데이터 수요측면에서 거대·비정형·실시간 데이터의 생성과 활용이 증가하고 있다는 점인데, IDC(2018년)는 2025년 전체 생성 데이터의 약 50% 데이터가 퍼블릭 클라우드 환경에 저장되어 활용되고, 생성 데이터 50% 이상을 IoT 디바이스가 담당하며, 총 데이터의 30%는 자율주행차 등과 같은 실시간 응용에서 소비될 것으로 보고 있다[1].

이러한 데이터의 유형 및 활용방식의 다변화에 대응하기 위해 '클라우드컴퓨팅 – 엣지컴퓨팅–디바이스'로 이어지는 데이터 처리 사슬상의 데이터 처리 및 연산기술의 고도화가 필요하나, 기존 반도체와 컴퓨팅기술이 곧 한계에 다다를 것으로 전망되고 있어 기술 혁신이 필요한 상황이다.

반도체와 컴퓨팅구조의 기술적 한계는 두 가지 차원으로 논의되고 있다.

첫 번째는 무어의 법칙(Moore's Law)으로 대변되는 반도체 집적(集積)의 물리적 한계이다. 트랜지스터의 집적도가 고도화될수록 미세공정으로 인한 발열 및 간섭 등 물리적 한계를 극복하기 어려워지기 때문에 반도체 공정의 새로운 돌파(Break–through) 기술이 필요하다.

두 번째는 현재 컴퓨터 시스템 구조의 근간인 폰 노이만(Von–Neumann) 방식이 빅데이터나 AI 등에서 요구되는 고속병렬연산에 효율적이지 못하다는 점이다. 주기억장치, 중앙처리장치, 입출력장치로 이어지는 순차적 처리(Serial Processing)를 기본 골격으로 하는 폰 노이만 구조는 AI가 요구하는 고속 병렬연산 수행 시 심각한 데이터 병목 문제(Data Bottleneck)를 발생시킨다[3,4].

병렬처리를 위해서 다수 프로세서의 저장장치 간, 그리고 각 저장장치 계층 간의 데이터 이동이 필수적이나 이러한 데이터 이동이 많아질수록 CPU(Central Processing Unit)의 처리속도가 아닌 데이터 이동속도가 컴퓨팅 성능 및 에너지 소비에 영향을 미치게 되는 것이다.

이와 같은 문제를 해결하기 위해서는 고속병렬연산 등 AI 처리성능을 대폭 개선하는 고성능컴퓨팅 기술의 향상이 필요하며, 장기적으로는 기술한계를 돌파하기 위한 변혁적 방식의 컴퓨팅시스템 소재, 구조, 계산모델 등의 개발이 필요하다.

한편, 기술발전 측면뿐만 아니라 시장 및 산업 경쟁력 차원에서도 AI 반도체 및 컴퓨팅기술은 더욱 중요해지고 있다. AI 응용 산업이 빠르게 확산되면서 AI를 위한 반도체와 컴퓨팅 산업이 새로운 산업 기회 분야로 대두되고 있으며, 주요국들과 세계적 업체들은 경쟁적으로 개발에 뛰어들고 있다.

반면, 국내의 AI 반도체 및 컴퓨팅 관련 기술은 선도국 대비 뒤처진 상황으로, 기술과 제품의 해외 의존도가 높아 외화 유출이 지속해서 발생하고 있다. AI 반도체는 ARM, 인텔, 퀄컴 등 글로벌 비메모리 반도체 회사의 IP 의존 및 수입 비중이 높은 편이다. 고성능컴퓨팅 인프라 분야는 미래 수요 대비 인프라 확보율이 낮을 뿐만 아니라 선진국 대비 현저히 뒤처진 관련 기술력으로 인해 국내 대부분 고성능컴퓨터는 해외 수입에 의존하고 있다.

AI의 안정적, 효율적 성능구현은 반도체 및 컴퓨팅기술이 결정적으로 좌우하기 때문에 이들은 국가 산업 및 과학기술 경쟁력을 결정할 미래 핵심기술로 인식되며, 국가 차원의 기술 확보 노력이 필요하다.

이러한 배경에서 발표된 ETRI의 AI 실행전략 2는 AI 반도체와 고성능컴퓨팅 분야에서 국가 차원의 기반 구축과 원천기술 확보에 이바지할 것이다.

2. 그간의 ETRI AI 연구 성과

ETRI는 그간 4M DRAM(1989년), 행정전산망용 주전산기(1991년)를 국내 최초로 개발하는 등 반도체, 컴퓨팅시스템 분야에서 대한민국의 기술 발전을 이끌어왔으며, 이러한 축적된 기술력은 AI와 접목되면서 다양한 우수 연구성과로 이어지고 있다.

반도체 분야에서 ETRI는 최근 고성능 연산이 필요한 서버·자율이동체 등에 활용 가능한 세계적 수준의 초저전력 AI 반도체를 개발하였다. 2017년 지능형 범용 AI 프로세서 알데바란을 개발하였고, 데이터 센터 내 클라우드 서버에 활용 가능한 실용성(전력·가격·크기)을 갖춘 고성능, 저전력 AI 프로세서인 AB9을 SKT와 국내 최초로 개발하였다.

컴퓨팅시스템 분야에서는 2014년 바이오 특화형 슈퍼컴퓨팅 시스템 마하(MAHA)를 독자 개발하고 2018년 초절전·고집적 마이크로서버 개발에 성공하였으며, 2019년에는 기존 기술 대비 최대 2~4배 고속 학습 지원이 가능한 고성능 분산 딥러닝 플랫폼 기술개발을 완료하는 등 AI에 대응한 기술력을 선제적으로 확보하고 있다.

3. ETRI AI 연구 추진 방향

ETRI는 2030년까지 중점적으로 확보할 원천기술의 로드맵을 수립하고, AI 인프라와 x+AI 서비스 기술경쟁력의 동반 상승을 도모하기 위한 연구전략을 수립하였다. 이를 위해 ETRI의 주요 역할 및 기술력을 고려하여 전문가 회의를 통해 ETRI의 AI 반도체 및 컴퓨팅 시스템 분야의 중점 추진영역을 다음과 같이 도출하였다.

1) AI 프로세서
2) AI 컴퓨팅 시스템
3) 뉴로모픽 프로세서·컴퓨팅
4) 양자소자 및 컴퓨팅 원천기술

Ⅱ장에서는 이 4가지 기술 분야에 대한 기술 개념 및 국내·외 연구 동향을 살펴보고, Ⅲ장에서는 ETRI 추진과제를 소개한다. Ⅳ장에서는 추진 전략의 방향을 간략히 소개한다.

Ⅱ. 기술 개념 및 연구 동향

1. AI 프로세서

가. 개념 및 필요성

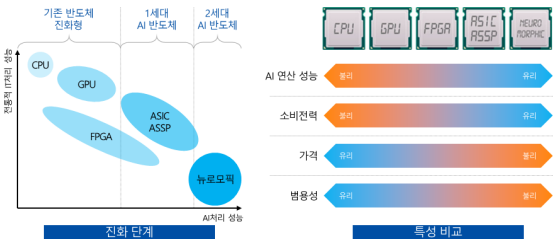
AI 반도체는 대규모 데이터 병렬처리 등 AI 데이터 처리에 최적화된 반도체로 정의된다[3-5]. AI 기술의 발전과 산업 확산에 따라 최적화된 고성능·저전력의 AI 반도체 수요가 급증하고 있으나, 현재의 시스템 반도체는 AI를 구현하는 데 전력 소모 등 성능효율이 떨어져 새로운 AI 전용 반도체가 필요하다.

또한, AI 반도체 산업은 지식재산(IP)과 전문적 설계역량이 중요한 기술집약적 산업이나 아직 지배적 강자가 없는 초기 시장으로 기술 혁신을 통해 시장 선점을 추구할 수 있는 시장이기도 하다[5].

나. R&D 방향

Ⅰ장에서 살펴본 반도체 기술의 한계 극복을 위한 연구개발은 두 가지 방향에서 진행되고 있다.

첫 번째 방향은 딥러닝 등 AI 알고리즘 처리에 특화된 연산패턴을 지원하는 AI 전용 프로세서의 개발이다. 이는 프로세서와 메모리 사이의 초 병렬화된 회로 구현을 통해 병렬처리 성능과 지연시간,



출처 김용균, "반도체 산업의 차세대 성장엔진 AI반도체 동향과 시사점," S18-01, 2018. 공공누리 제2유형: 출처표시 + 상업적 이용금지.

그림 1 AI 반도체 진화 단계 및 특성 비교

전력효율을 향상시킨다.

그림 1과 같이 AI 전용 프로세서는 용도에 따라 학습형, 추론형으로 구분되며, 구현기술에 따라 각각의 성능 특성을 보인다[6].

두 번째 방향은 메모리와 프로세서를 같은 칩에 패키징하는 HBM(High Bandwidth Memory) 기반의 PIM(Processing in Memory) 프로세서의 개발이다[7]. 프로세서를 메모리칩 혹은 적층 메모리 셀에 같이 패키징하여, 메모리에서 프로세서까지 데이터 이동의 비효율성을 극복하는 것으로, CPU와 메모리 간 대역폭 차이에 따른 병목 문제와 데이터 이동에 따른 에너지 소모를 대폭 단축할 수 있다.

다. 국내·외 동향

먼저 AI 가속기(AI 전용 프로세서) 시장 동향을 살펴보면, 엔비디아의 독주가 이어지고 있다. 엔비디아의 GPU(Graphics Processing Unit)는 4대 클라우드 서비스(아마존, 마이크로소프트, 구글, 알리바바)에 사용되고 있는 AI 가속기의 97%를 점유하고 있다[8].

엔비디아는 2020년 5월 새로운 '암페어' 아키텍처 기반의 A100 GPU를 공개했는데, A100은 540억 개의 트랜지스터를 집적한 것과 같은 성능을 지닌 GPU로, 기존 '볼타' 기반의 V100과 비교해 20배 뛰어난 성능을 지녔다[8].

구글은 2016년부터 인공신경망 추론 및 학습용 프로세서인 TPU(Tensor Processing Unit) 시리즈를 개발하고 있다. TPU는 AI 알고리즘 전용 가속 구조를 채택함으로써 GPU 대비 30배 이상 에너지 효율을 높이는 것으로 알려져 있으며, 현재는 100페타플롭스 성능의 TPU까지 개발되어 구글 클라우드에서 제공되고 있다[9].

새로운 반도체 설계기술 분야에서는 데이터 처리에 있어 심각한 비효율성을 드러내는 메모리 병목을 해소하고 메모리 대역폭을 개선하기 위한 메모리 적층기술 연구가 활발히 전개되고 있다. 대표적 기술로는 마이크론에서 선도하는 HMC(Hybrid Memory Cube)와 삼성전자와 SK하이닉스에서 개발 중인 HBM이 있다[10].

2. AI 컴퓨팅시스템

가. 개념 및 필요성

AI 컴퓨팅시스템은 대용량 데이터를 초고속으로 생산·처리·활용 가능한 고성능컴퓨터로 정의된다. AI 기술은 높은 정확도를 구현할수록 더 많고, 더 높은 해상도의 학습 데이터를 요구하므로 기하급수적인 계산량 증가를 수반한다. 이 때문에 고성능컴퓨터는 AI 시대의 산업발전과 과학기술 진보를 위한 필수적인 인프라이며, 이를 통해 제약 및 바이오, 신소재, 항공우주 등 고난도 연산이 요구되는 과학·산업 분야에서의 획기적인 효율 향상을 기대할 수 있다[11].

표 1에서와 같이 세계 주요국은 엑사스케일급의 고성능컴퓨터 개발에 집중적 투자를 진행 중이다. 미국, 중국, EU는 시스템당 3억에서 6억 달러의 대규모 투자를 진행 중이며, 일본 역시 조 단위 예산을 개발과 구축에 투입하고 있다[11].

표 1 주요국의 엑사스케일 슈퍼컴퓨터 투자 동향

구분	세부 내용
미국	• 2021년부터 엑사-스케일 컴퓨터 구축을 목표로 시스템당 6억 달러 수준의 투자를 2023년까지 계획
유럽	• 2024년을 목표로 시스템당 3.5억 달러 수준의 투자 계획
중국	• 2020년 엑사-스케일 컴퓨터 구축을 목표로 시스템당 3.5억~5억 달러 투자를 진행
일본	• 2021년까지 10억 달러를 투입하여 기계학습, 딥러닝에 특화된 엑사-스케일 컴퓨터 개발을 목표

출처 한국연구재단, "슈퍼컴퓨터 개발 선도사업 신규과제 사전공고," 2020. 저자 재정리

국내의 고성능컴퓨팅 핵심기술은 주요국 대비 뒤처진 상황이나 글로벌 수준의 경쟁력 확보를 위해 CPU를 포함한 독자적 슈퍼컴퓨터 기술의 자체 개발을 추진하고 있다[11].

나. R&D 방향

폰 노이만 구조, 즉 프로세서 중심 컴퓨팅은 대용량 데이터 처리 시 CPU의 처리성능보다 데이터 이동속도가 컴퓨팅 성능 및 에너지 소비에 결정적 영향을 미치는 문제를 초래하기 때문에 AI 응용과 같은 대규모 병렬처리에 적합하지 않다.

대부분 슈퍼컴퓨터는 전력 소모가 많고 에너지 효율성이 낮은데, 일례로 미국 1위 슈퍼컴퓨터 서밋(Summit)은 초당 148,000조 회의 연산을 수행하며 최대 13MW 전력을 소비하는데, 이는 미국 가정 8,000세대의 조명을 동시에 켤 수 있는 전력량에 해당하는 수준이다[12].

따라서 최근의 고성능컴퓨팅 연구는 이러한 한계를 극복하기 위해 메모리 중심 컴퓨팅의 방식을 채택하고 있다.

메모리 중심 컴퓨팅이란 프로세서 간 데이터 이동을 최소화하기 위해 컴퓨터 구조를 프로세서가 아닌 메모리 중심으로 재편하는 컴퓨팅 모델이다. 다수의 메모리 노드를 고속의 패브릭(Fabric)으로 연결하여 거대한 공유메모리 풀(Pool)을 구성하고, 이를 통해 복수의 컴퓨팅 노드가 각자 데이터를 병렬적으로 처리할 수 있게 함으로써 정보처리 성능을 획기적으로 향상시키는 것이 특징이다[10,13].

다. 국내·외 동향

메모리 중심 컴퓨팅의 근간인 고속 인터커넥트 및 비휘발성 메모리 기술과 함께 이를 활용하는 컴퓨팅구조, 시스템 SW 등의 기술개발과 생태계 조성을 위한 기업 간의 합종연횡이 활발히 진행되고 있다.

2016년부터 메모리 중심 컴퓨팅을 위한 차세대 연결망의 산업 표준을 위해 AMD, 자일링스를 중심으로 한 CCIX(Cache Coherent Interconnect for Accelerators), Dell EMC, HPE를 중심으로 한 Gen-Z, IBM을 중심으로 한 Open CAPI 컨소시엄이 구성되어 활동하고 있으며, 2019년 3월에는 인텔이 CXL(Compute Express Link) 컨소시엄을 발족하여 고대역폭에서 가속기와 CPU 간의 메모리 공유를 지원하는 인터커넥트 기술인 CXL 1.0을 규격 발표하였다.

국내에서는 슈퍼컴퓨터를 개발하기 위한 연구가 시작되고 있는데, 광주과학기술원은 광주 AI 집적단지에 2022년 약 110.5페타플롭스급 AI 슈퍼컴퓨터를 도입할 예정이며, KISTI에서는 '차세대 초고성능컴퓨터를 위한 이기종 매니코어 HW 시스템개발' 사업을 통해 CPU를 제외한 컴퓨팅 플랫폼 및 HPC 시스템을 개발을 추진 중이다. 또한, 메모리 중심 컴퓨팅을 위한 고속 비휘발성 메모리, 인터커넥트, 운영체제, 응용수준의 연구가 삼성, 하이닉스, 알티베이스, 선재 소프트, ETRI 등을 중심으로 진행되고 있다.

3. 뉴로모픽 프로세서 및 컴퓨팅

가. 개념 및 필요성

뉴로모픽(Neuromorphic) 반도체는 뉴런과 시냅스가 생물학적 뇌 내에서 기능하는 방식(뉴런-시냅스 구조)을 모사하는 SNN(Spiking Neural Network) 기술을 사용한 대표적인 非 폰 노이만 방식의 반도체이다[14].

뉴로모픽 반도체 코어에는 트랜지스터와 메모리를 비롯한 몇 가지의 전자 소자들이 탑재되어 있으며, 코어의 일부 소자는 뇌의 뉴런의 역할을 담당하고, 메모리 반도체는 뉴런과 뉴런 사이를 이어주는 시냅스 역할을 담당한다[14].

뉴로모픽 반도체의 장점은 적은 전력만으로 많은 양의 데이터 처리가 가능하며, 높은 집적용량으로 인간의 뇌처럼 학습할 수 있어 연산 성능이 대폭 향상된다는 점이다. 따라서 기존의 딥러닝 방식과 유사한 성능구현은 물론 높은 전력효율을 달성할 수 있어, 특히 제한된 전력 자원을 갖는 모바일 시스템의 성능을 획기적으로 개선할 수 있다.

가트너에 따르면, 뉴로모픽 반도체는 향후 10여 년 내에 시장에 본격 진입하여 자율주행차, 지능형 로봇 및 각종 AI기기에 탑재될 것으로 전망하고 있다[15].

나. R&D 방향

뉴로모픽 반도체는 기존 시스템 반도체보다 효과적이고 전력 소모가 적어 광범위한 AI 활용을 가능하게 할 것으로 전망되나 소자 및 소재 등의 선결과제가 많다. 소프트웨어 기반 뉴로모픽 시스템은 뇌의 시냅스와 뉴런의 기능을 수식적으로 정의하고, 실제 연산은 기존 컴퓨터 시스템을 통해 진행하기 때문에 궁극적으로는 성능, 학습시간, 소비전력 면에서 한계를 보일 수밖에 없다[16,17].

결국, 이를 해결하기 위해서는 뇌의 시냅스와 뉴런의 기능을 모방한 하드웨어 기반 뉴로모픽 반도체가 필요하며, 현재는 최적화된 신소재 및 이에 기반한 연산 구조 등의 초기연구가 시작된 상황이다.

인텔, 퀄컴, IBM에서 상용화가 가능한 수준의 시제품을 내놓고 있으나, 뉴로모픽 컴퓨팅은 다양한 분야의 학문이 함께 연구되어야 하는 대표적인 융합기술로 인간 뇌의 동작 방식에 관한 연구, 뉴로모픽 소자 연구, 뉴로모픽 소자로 구성된 반도체 연구, 뉴로모픽 반도체로 구성된 컴퓨터 연구, 뉴로모픽 컴퓨터를 동작하게 하는 소프트웨어 연구 등 다양한 분야에서의 기술발전이 요구된다.

다. 국내·외 동향

인간 뇌를 모사한 뉴로모픽 반도체 연구는 유럽과 미국을 중심으로 2000년대 중반부터 국가 주도 R&D 사업으로 진행되고 있다[18].

EU는 2013년부터 10년간 10억 유로를 투자하여, Human Brain Project(HBP)라는 인간 두뇌에 관한 대규모 원천연구 프로젝트를 진행하고 있으며, 미국 역시 BRAIN Initiative를 2013년부터 수립하여 인간 두뇌에 대한 광범위한 기술개발을 추진 중이다[18].

IBM은 2008년부터 美 국방부 산하 DARPA가 주도하는 시냅스(SyNAPSE) 프로젝트에 참여하여, 2014년 TrueNorth라는 뉴로모픽 칩을 개발하였다. 이 칩은 기존 프로세서의 1/10,000 수준의 전력소모량으로 초당 1,200~2,600프레임의 이미지 분류가 가능하다[14,18].

2012년에 영국 맨체스터 대학이 중심이 되어 개발한 SpiNNaker는 스파이크 신경망을 실시간으로 모델링할 수 있도록 대규모 병렬처리 뉴로모픽 컴퓨터로 10억 개 뉴런의 시뮬레이션이 가능하다[14,18].

인텔은 2019년 학습이 가능한 스파이킹 뉴로모

픽 칩인 로이히(Loihi)를 발표하였는데, 이는 DNN (Deep Neural Network) 대비 100만 배 빠른 학습 및 실행 성능을 증명하였다[14,18].

4. 양자 소자 및 컴퓨팅 원천기술

가. 개념 및 필요성

양자컴퓨팅은 기존 컴퓨팅과 작동원리가 완전히 다른 양자역학에 기반을 둔 차세대 컴퓨팅기술로, ICT 생태계의 패러다임을 바꿀 수 있는 핵심기술로 주목받고 있다. 양자 상태의 중첩과 얽힘 현상을 이용하는 정보처리의 단위 큐비트(Qubit)와 양자 병렬성을 이용할 수 있는 양자 알고리즘을 통해 연산 속도를 비약적으로 높일 수 있다[19].

최근 제반 기술이 급속히 발달하면서 데이터 검색, 소인수 분해 등 몇몇 특정 연산에서는 기존 슈퍼컴퓨터보다 월등한 성능을 보일 만큼 발전속도가 빨라지고 있다. 최근 20년간 레이저 및 나노 광학기술, 극저온·전자기 환경구축 기술, 신소재·초미세 소자 공정기술 및 초정밀·초고속 계측기술이 모두 비약적 발전을 이루어 양자컴퓨터의 현실화 가능성이 대폭 상승하게 된 것이다. 특히, 구글이 자사의 54큐비트의 시카모어(Sycamore) 프로세서로 양자 우위를 보임으로써 향후 양자컴퓨팅 주도권을 두고 경쟁이 더욱 치열해질 전망이다.

양자컴퓨팅이 실용화되면 현재 보안체계가 무력화되고 주요 양자기술의 수출 승인이 제한될 수 있으므로 양자 분야 전반에 대하여 핵심원천 기술 확보가 필수적이다. 현재 산업·군사·금융 등 주요 보안체계에서 사용 중인 RSA 공개키 방식의 암호가 양자 컴퓨터를 이용하게 되면 해독 가능성이 커지기 때문에 독자기술 확보가 필요하다[20].

이미 개발 양자 알고리즘을 적용한 병렬연산만으로도 빅데이터 검색, 양자화학, 기계학습, 해석학과 이미지 분석 등의 분야에 큰 혁신이 가능할 것으로 기대된다[21].

나. R&D 방향

양자컴퓨터는 크게 양자 시뮬레이터와 범용 양자컴퓨터로 분류할 수 있다.

양자 시뮬레이터는 초깃값을 넣고 최적화된 결과값을 도출하는 방법으로, 어닐링 방식(D-wave社 주도), 디지털 어닐링(후지쯔社 주도) 방식 등 구현 모델을 이미 실용화하여 사용하고 있으나 제어할 수 없고 활용처가 제한적이며 오류보정이 불가능한 단점이 존재한다[20].

범용 양자컴퓨터는 범용 게이트를 제어하여 알고리즘을 직접 연산하는 것으로, 기술발전에 따라 양자계 검증(10개 이하 물리 큐비트) 단계, NISQ(50~100개 물리 큐비트로 구성된 노이즈가 있는 중간 형태의 양자 컴퓨팅) 단계, 오류보정 범용 양자컴퓨터(50개 이상 논리 큐비트)로 나뉜다[20]. 범용 게이트는 논리 큐비트를 만들 수 있을 만큼 물리 큐비트의 에러율이 아주 낮은 안정적 상태가 필요하며, 하나의 논리 큐비트를 구현하기 위해서는 여러 개의 물리 큐비트가 필요하므로 기술적 도전이 큰 분야이다.

양자기술 발전 수준을 고려할 때, 양자컴퓨팅은 단기적으로는 극저온 환경에서 여러 개의 큐비트를 매우 제한된 연산 응용에서 활용되는 수준으로 개발되고, 중기적으로는 비트 기반 컴퓨팅의 계산 한계를 돌파할 것으로 예상하며, 장기적으로는 상온 수준에서 동작하며 AI 분야에서 활용될 것으로 전망된다.

다. 국내·외 동향

최근 선도기업들은 양자 프로세서를 포함한 양자 컴퓨팅기술을 경쟁적으로 발표하고 있다.

IBM과 구글은 초전도 기반의 양자컴퓨터 개발에 주력하고 있으며, IBM은 2019년 53큐비트 양자 프로세서(Hummingbird)를 이용하여 'IBM Q 53' 양자컴퓨터를, 구글은 2018년 72큐비트 양자 프로세서(Bristlecorn)와 2019년 53큐비트 양자 프로세서(Sycamore)를 이용한 양자컴퓨터를 각각 발표하였다[22].

IBM은 2020년 9월, 다중 양자 보드 적층기술을 활용한 1백만 개 이상의 큐비트 실현이 가능한 양자컴퓨팅 로드맵을 발표하면서, 매년 전년보다 2~3배 큐비트 수를 갖는 양자머신 개발계획을 제시하였다[23].

인텔은 전자스핀이라는 양자 특성을 활용하는 실리콘 기반의 양자컴퓨터 개발에 집중하고 있으며, 2018년 49큐비트 양자 프로세서(Tangle Lake)를 이용한 양자컴퓨터를 발표한 바 있다[22].

국내에서는 과학기술정보통신부가 2019년부터 양자컴퓨팅 핵심원천기술 확보 및 국내 연구생태계 조성 사업을 통해 향후 5년간 미래 유망분야에 총 445억 원의 투자계획을 발표하였고[24], 민간 차원에서는 2017년 삼성전자가 양자컴퓨팅이 반도체에 미치는 영향에 대한 국제 공동 연구를 진행하였다. SK텔레콤은 양자암호통신 분야 1위 기업으로 평가받는 IDQ를 2018년에 인수하였다.

Ⅲ. ETRI 추진과제

1. AI 프로세서 개발

ETRI에서 추진하는 'AI 프로세서 개발'은 ETRI의 기 확보한 차별화된 설계기술을 바탕으로, AI 응용을 위한 고성능 연산능력과 높은 에너지 효율 달성이 가능한 AI 프로세서 개발을 목표로 한다.

추진과제의 세부목표로는 (1) 페타플롭스급 AI 프로세서 개발, (2) 저전력 모바일 시각지능 AI 프로세서 개발, (3) HBM을 활용한 차세대 PIM 프로세서 개발, (4) 인메모리 AI 프로세서 개발이 있다.

(1) 페타플롭스급 AI 프로세서에서는 이미 개발한 알데바란 CPU를 쿼드코어로 구성하여 기존 프로세서의 10만 배 성능(40테라플롭스)급의 객체인식 및 자율주행에 특화된 서버용 AI 프로세서의 개발을 목표로 한다. (2) 저전력 모바일 시각지능 AI 프로세서는 연산량 감축을 통해 낮은 전력에서도 사람 수준의 사물 인식 정확도를 갖는 고효율 프로세서 설계기술 개발을 목표한다. (3) HBM을 활용한 PIM 프로세서는 반도체 적층 설계를 통해 메모리 병목 문제 해소와 저전력 소비를 동시에 달성하는 혁신적인 AI 프로세서 설계기술 확보를 목표로 한다. (4) 인메모리 AI 프로세서는 비휘발성 메모리 기반의 인메모리 프로세싱 기술과 최적화된 AI 신경망 기술을 통해 연산과 저장이 통합된 AI 프로세서의 개발을 목표로 한다.

2. AI 컴퓨팅시스템 개발

ETRI가 추진하는 'AI 컴퓨팅시스템 개발'은 데이터 접근 병목 문제해결을 위한 메모리 중심 컴퓨팅의 핵심요소 및 시스템 기술을 목표로 한다. 이를 위한 세부 목표기술로는 (1) AI 슈퍼컴퓨터 시스템 구현기술, (2) 메모리 중심 컴퓨팅 서버 기술, (3) 차세대 혼성 메모리 지원 시스템 기술이 있다.

(1) AI 슈퍼컴퓨터 시스템 구현기술에서는 기존보다 현저히 적은 전력소모량과 고집적도를 보장하는 분산 딥러닝 고속처리에 특화된 슈퍼컴퓨터 시스템 구현을 목표로 한다. (2) 메모리 중심 컴퓨팅 서버 기술은 다수의 CPU와 대규모 메모리 간의 확장성과 고속 I/O를 보장하는 연결망 구성 및 자원 관리의 원천기술 확보를 목표한다. (3) 차세대 혼성 메모리 지원 시스템 기술은 휘발성/비휘발성의 혼성 메모리로 구성된 대규모 메모리 공간에서

기존 방식이 초래하는 메모리 접근 성능 저하를 방지하여 고속으로 거대 데이터 분석을 지원하는 운영체제 기술 개발을 목표한다.

3. 뉴로모픽 프로세서 및 컴퓨팅 개발

ETRI는 자율학습이 가능한 인간 뇌의 신경망 구조 및 작동원리를 모방한 AI '뉴로모픽 프로세서 및 컴퓨팅 개발'을 추진한다. 이를 위한 세부목표는 (1) 뉴로모픽 프로세서 핵심기술개발과 (2) 뉴로모픽 시스템 SW 개발이다.

(1) 뉴로모픽 프로세서 핵심기술은 저전력으로 인간과 유사한 인지학습이 가능하게 하는 것으로, 병렬 어레이 연산 구조를 이용하여 학습기능과 결정기능을 갖춘 뉴로모픽에 최적화된 하드웨어 아키텍처를 구현하고 IP를 확보하는 것을 목표로 한다. (2) 뉴로모픽 시스템 SW 기술은 인간의 뇌 신호 처리 방식에 관한 모델 연구와 이 모델에 기반한 뉴로모픽 프로그래밍 언어 및 응용 SW로, 컴파일러 및 응용 SW를 효율적 지원하기 위한 OS를 포함한다.

4. 양자 소자 및 컴퓨팅 원천기술 개발

ETRI는 큐비트 확장과 고신뢰 연산자원이 가능한 양자컴퓨팅 분야의 원천기술 확보를 목표로 한다. 이를 위해 (1) 양자 프로세서 소자 구현기술 개발, (2) 양자 운영체제 기술 연구개발을 추진한다.

(1) 양자 프로세서 소자 구현기술은 양자 클라우드 네트워크를 이용하여 큐비트 확장이 가능한 광집적회로 및 반도체 집적회로 기반의 양자 프로세서 소자 개발을 목표로 한다. (2) 양자 운영체제 기술은 결함을 허용하는 논리 양자큐빗을 제공하기 위한 것으로, 시스템 차원의 실시간 양자오류 정정 기능과 연산 시퀀스의 최적 스케줄링을 제공하는 2세대 양자컴퓨팅 운영체제의 원천기술을 확보하는 것을 목표로 한다.

Ⅳ. 결론

AI 기술과 산업이 급격히 성장하면서 반도체 및 컴퓨팅 분야 역시 새로운 전기를 맞고 있다. 세계 주요국과 주요 기업은 앞 다투어 AI를 위한 고성능 AI 반도체 및 컴퓨팅 핵심기술 확보와 폰 노이만 구조를 극복하는 뉴로모픽, 양자기술과 같은 변혁적 컴퓨팅 시대를 준비하고 있다. 우리 정부도 AI 시대의 기술경쟁력을 확보하기 위해 최근 인공지능 국가전략[25], AI 반도체 산업발전 전략[5], 국가 슈퍼컴퓨팅 선도 사업[11], 양자컴퓨팅 기술개발사업 추진계획[24] 등 이 분야에 대한 투자와 정책을 늘려가고 있다.

국내 반도체 및 컴퓨팅 분야에서 그동안 많은 기술력을 축적해 온 ETRI는 이러한 배경에서 AI 실행전략을 통해 AI 반도체 및 컴퓨팅 분야의 경쟁력 강화를 위한 추진과제를 도출하였다. ETRI는 이를 실현하기 위하여 다음의 방향으로 경쟁력을 확보할 계획이다.

첫째, ETRI는 고성능 AI 컴퓨팅 기술개발에 있어, 'AI 반도체 – 컴퓨팅 프레임워크 – 시스템'으로 이어지는 통합 개발을 통해 'ETRI AI 컴퓨팅 플랫폼'을 제시할 계획이다. ETRI는 AI 반도체, 컴퓨팅 프레임워크 및 시스템 분야에서 이미 세계 최고 수준의 개별 R&D 성공 경험을 보유하고 있는 바, 이들 간의 유기적 통합을 통해 시너지를 창출할 수 있을 것으로 기대한다. 이는 AI SW와 시스템을 동시 개발하는 방식으로 전환하여 AI 반도체 R&D를 혁신하고자 하는 정부 정책과도 일맥상통한다[25]. 이러한 통합전략은 외산 장비 국산화와 국가 AI 거점 컴퓨팅 인프라 확충을 기대할 수 있

고, 중장기적으로는 뉴로모픽 및 양자 컴퓨팅과 같은 차세대 컴퓨팅시스템 개발의 참고모델(Reference Model)로 활용될 수 있을 것이다.

둘째, ETRI는 국내 기술개발의 후발성을 고려하여 국제적 생태계를 보유하고 있는 기술과 연계하고 국내·외 협력 연구 및 글로벌 컨소시엄에 적극적으로 참여한다는 계획이다. 즉, 생태계가 잘 형성되어 있는 x86, ARM, RISC-V와 같은 기존의 CPU 기술을 기반으로 연구개발 수행하여 효율성을 높이고, 글로벌 선도기관과의 국제협력 및 인적 교류를 통하여 세계 수준의 고성능컴퓨팅기술을 조기 확보한다는 계획이다.

용어해설

혼성 메모리 휘발/비휘발성 메모리가 동일한 메모리 계층에서 함께 사용되는 차세대 메모리 구조

양자얽힘 한 개의 큐비트가 고유 상태(0 또는 1)뿐만 아니라 그 특정 상태들이 확률적으로 공존한 상태(0과 1이 각 50%씩 공존)도 가능한 성질

양자중첩 두 개 이상의 큐비트가 서로 연관성을 가져 한 큐비트의 상태가 다른 큐비트와 독립되지 않는 성질

약어 정리

CPU	Central Processing Unit
DNN	Deep Neural Network
GPU	Graphics Processing Unit
HBM	High Bandwidth Memory
HMC	Hybrid Memory Cube
NISQ	Noisy Intermediate-Scale Quantum
PIM	Processing in Memory
SNN	Spiking Neural Network
TPU	Tensor Processing Unit

참고문헌

[1] IDC, "data age 2025," 2018.

[2] OpenAI, "AI and Compute," 2018. 5. 16., https://openai.com/blog/ai-and-compute/

[3] 최새솔, "인공 지능 반도체 산업 동향 및 이슈 분석," ETRI Insight Report, 2017-37, 2017.

[4] 나영식, 조재혁, "인공지능(반도체)," KISTEP 기술동향브리프," 2019-01호, 2019.

[5] 과학기술정보통신부, "인공지능(AI) 반도체 산업 발전전략," 보도자료, 2020. 10. 12.

[6] 김용균, "반도체 산업의 차세대 성장엔진 AI반도체 동향과 시사점," ICT Spot Issue, 2018-01호, 2018.

[7] 연세대 ICSL, https://web.yonsei.ac.kr/icsl/research.html

[8] 양대규, "AI 가속 GPU 전쟁'…엔비디아 아성에 도전하는 AMD," AI타임스, 2020. 8. 14., http://www.aitimes.com/news/articleView.html?idxno=131512

[9] MZ megazone, "CLOUD TPU가 뭐길래, AI 전문가들이 열광할까?," https://gc.hosting.kr/blog-cloud-tpu/

[10] 오명훈 외, "메모리 중심 컴퓨팅 기술동향," 융합연구리뷰 6권 3호, 2020, pp. 4-29.

[11] 한국연구재단, "슈퍼컴퓨터 개발 선도사업 신규과제 사전공고 내역," 2020.

[12] G. Dvorsky, "The World's Most Powerful Supercomputer Is an Absolute Beas," 2018. 6. 8., https://gizmodo.com/the-world-s-most-powerful-supercomputer-is-an-absolute-1826679256

[13] 황민규, "5G로 폭증하는 데이터 트래픽… 메모리 중심 컴퓨팅 부상," 조선비즈, 2020. 5. 4., https://biz.chosun.com/site/data/html_dir/2020/05/03/2020050301280.html

[14] 박성모 외, "스파이킹 신경망 기반 뉴로모픽 기술 동향," TTA Journal, 188호, 2020, pp. 28-33.

[15] Gartner(2018), "Hype Cycle for Emerging Technologies," 2018.

[16] 김판길, 배준호, "지능형 반도체 신소자 기술 동향," TTA Journal, 188호, 2020, pp. 34-43.

[17] 이승훈, "저전력 뉴로모픽 프로세서 설계 기술," 주간기술동향, 1924호, 2019. 11. 27.

[18] 오광일 외, "인공지능 뉴로모픽 반도체 기술 동향," 전자통신동향분석, 35권 3호, 2020, pp. 76-84.

[19] 한상욱 외, "양자통신 및 양자컴퓨팅 분야 소개 및 연구동향," 융합연구리뷰, 6권 3호, 2020, pp. 32-58.

[20] 임승혁, "범용 양자컴퓨터," KISTEP 기술동향브리프, 2019-19호, 2019.

[21] 조성선, "양자컴퓨터 개발 동향과 시사점," ICT Spot Issue, 2018-02호, 2018.

[22] 정지형, 최병철, "빛의 속도로 계산하는 꿈의 컴퓨터, 양자컴퓨터," KISTEP Issue Paper, 2019-07, 2019.

[23] 김우용, "IBM, 클라우드 기술로 64 양자볼륨 달성," ZDNET Korea, 2020. 9. 8., https://zdnet.co.kr/view/?no=20200908095603.

[24] 과학기술정보통신부, "양자컴퓨팅 기술개발사업 추진계획," 보도자료, 2019. 1. 31.

[25] 범부처 합동, "인공지능 국가전략," 2019. 12.