

# 초저전력 엣지 지능형반도체 기술 동향

## Trends in Ultra Low Power Intelligent Edge Semiconductor Technology

오광일 (K.I. Oh, kioh@etri.re.kr)

김성은 (S.E. Kim, sekim@etri.re.kr)

배영환 (Y.H. Bae, yhbae@etri.re.kr)

박성모 (S.M. Park, smpark@etri.re.kr)

이재진 (J.J. Lee, ceicarus@etri.re.kr)

강성원 (S.W. Kang, kangsw@etri.re.kr)

SoC 설계연구그룹 선임연구원

SoC 설계연구그룹 선임연구원

SoC 설계연구그룹 책임연구원

SoC 설계연구그룹 책임연구원

SoC 설계연구그룹 책임연구원/그룹장

지능형반도체연구본부 책임연구원/본부장

In the age of IoT, in which everything is connected to a network, there have been increases in the amount of data traffic, latency, and the risk of personal privacy breaches that conventional cloud computing technology cannot cope with. The idea of edge computing has emerged as a solution to these issues, and furthermore, the concept of ultra-low power edge intelligent semiconductors in which the IoT device itself performs intelligent decisions and processes data has been established. The key elements of this function are an intelligent semiconductor based on artificial intelligence, connectivity for the efficient connection of neurons and synapses, and a large-scale spiking neural network simulation framework for the performance prediction of a neural network. This paper covers the current trends in ultra-low power edge intelligent semiconductors including issues regarding their technology and application.

\* DOI: 10.22648/ETRI.2018.J.330603

\*This work was supported by the ICT R&D program of MSIT/IITP. [2018-0-00197, Development of ultra-low power intelligent edge SoC technology based on lightweight RISC-V processor]



본 저작물은 공공누리 제4유형  
출처표시+상업적이용금지+변경금지 조건에 따라 이용할 수 있습니다.

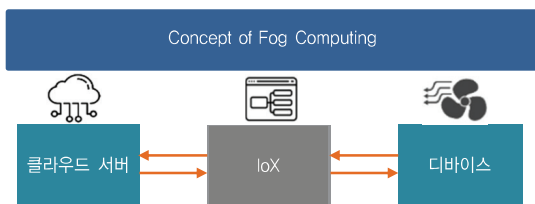
2018  
Electronics and  
Telecommunications  
Trends

최신 반도체, 하드웨어 기술  
동향 특집

- I. 개요
- II. 초저전력 엣지 지능형 반도체 기술 개념
- III. 지능형 반도체 코어 기술 동향
- IV. 지능형 반도체를 위한 커넥티비티 기술 동향
- V. 지능형 반도체를 위한 시뮬레이션 프레임워크 기술 동향
- VI. 결론

## I. 개요

최근 전자기기 시장에서 휴대용 기기 및 스마트 디바이스의 사용이 급격히 증가함에 따라 각 기기들이 서로 유무선의 네트워크로 연결이 되고 있으며 각각 본연의 태스크를 수행하기 위하여 취득된 데이터를 상위 계층의 네트워크 허브로 전달하거나 상위 허브에서 전달된 정보를 받아 사용자에게 보여준다. 사용자들이 각자 스마트 기기를 통하여 인터넷에 접속하여 정보를 공유하는 시대를 뛰어넘어 사용자가 주도적으로 관여하지 않아도 주변의 모든 사물들이 각기의 목적을 위하여 네트워크에 연결되는 IoT 시대가 도래하게 되었다. 다만 폭발적으로 증가되는 IoT 기기의 확산과 이 기기들이 모두 네트워크에 연결된다고 가정하면 트래픽 증가의 트렌드는 현재 예측보다 더 가파른 기울기로 증가할 것이다. 실제로 시스코는 2015년 ‘시스코 2015-2020 글로벌 클라우드 인덱스(The Cisco Global Cloud Index 2015-2020)’ 보고서를 통하여 2020년 500억개 이상의 IoT 디바이스가 인터넷에 연결되어, 클라우드 트래픽은 14.1 제타바이트에 이를 것으로 전망하였다[1]. 이러한 트래픽의 증가에 대비한 하나의 해법으로 통신 기술의 세대를 전환하려는 노력이 진행되어 왔으며 최근 5G 기술 상용화를 통하여 조만간 이러한 기술의 실생활 적용을 체험할 수 있을 것으로 기대하고 있다[2]. 다만 통신의 세대 전환만으로는 그 한계가 명백할 것이기 때문에 IoT 기기의 홍수에 대비하기 위하여 시장은 또 다른 기술의 변화를 요구하게 된다. 2014년 시스코는 (그림 1)의 ‘포그 컴퓨팅’이라는 개념을 발표 하였는데 시스코



(그림 1) 시스코가 제안한 포그 컴퓨팅 개념

IoT 그룹의 로베르토 데라모라(Roberto De La Mora) 수석 이사에 따르면 “포그 컴퓨팅은 IoT 시대에서 필요로 하는 종래와 다른 개념의 분산 컴퓨팅에서 출발한 기술로서 탄력성, 확장성, 이동성, 속도에 중점을 두어 클라우드 컴퓨팅의 개념을 종단 사용자(end-user) 혹은 엣지(edge)까지 확장한 개념”으로 정의하고 있다[3]. 즉, 종래 클라우드에서 집중적으로 처리하던 데이터 프로세싱에서 탈피, 엣지까지 데이터 프로세싱을 분산하고자 하는 기술을 뜻하며 현재의 엣지 컴퓨팅의 모태가 되는 개념이다. 단순히 속도 혹은 저지연에 초점을 둔 통신의 세대 전환과는 다른 개념이며 이는 클라우드 컴퓨팅의 확장된 개념으로서 데이터가 생산되고 소비되는 종단 가까이에서 데이터 처리를 일부 담당하고자 하는 개념으로 정의된다.

## II. 초저전력 엣지 지능형 반도체 기술 개념

앞절에서 설명한 포그 컴퓨팅 혹은 엣지 컴퓨팅의 실현을 위하여 반도체 하드웨어에서 바라보는 엣지 컴퓨팅의 개념은 앞절에서 설명한 것에서 좀 더 구체적인 모습을 보인다. 데이터 프로세싱을 종단 사용자 혹은 엣지까지 확장을 하기 위하여 각 IoT 기기들은 단순히 주변의 데이터를 취득하거나 사용자에게 정보를 보여주는 것에서 탈피 각기의 데이터를 처리하기 위한 기능의 탑재를 요구한다. 이러한 개념을 중심으로 2017년 8월경 쉘컴은 산업용 사물 인터넷의 해법으로 ‘엣지 프로세싱’이라는 개념을 시장에 발표하였다[4].

엣지 프로세싱으로 크게 다음의 세 가지 장점이 있을 수 있는데, 트래픽 감소에 따른 비용절감, 엣지에서의 즉각적인 데이터 처리를 통한 저지연 실현, 민감한 개인 데이터를 클라우드까지 보내지 않으므로써 얻을 수 있는 개인 프라이버시 향상 등이 있다. 이러한 데이터 프로세싱은 종래와 같이 사용자가 지정한 일련의 알고리즘대로 태스크를 수행하는 CPU 기반에 의한 프로세싱

일 수도 있으며, 혹은 인공지능에 입각한 학습 결과를 바탕으로 한 지능적 판단 기반 프로세싱일 수도 있다. 초저전력 엣지 지능형 반도체 기술에서 추구하는 개념은 후자로서 지능형 반도체라는 용어처럼 인공지능 기반의 지능적 판단을 수행하는 엣지용 반도체이다. 수많은 IoT 기기들은 수집된 데이터에 대하여 각기 나름의 지능적 판단을 하고 그에 따라 현재 수집된 데이터에 대하여 상위 계층의 네트워크 허브로 전달할지에 대한 결정을 내려 데이터를 생산하거나 소비되는 장소에 더 가까운 곳에서 처리하게 된다. 이로써 불필요한 네트워크 트래픽을 감소하고 클라우드까지 전달되는 시간을 절약하여 즉각적인 데이터 처리가 가능하여 저지연을 실현할 수 있다. 여기서 지능적 데이터 프로세싱이 수행되는 엣지는 네트워크의 게이트 웨이(Gate way)거나 있고 더 확장하여 IoT 기기 그 자체일 수도 있다. 즉 시스코에서 제안한 포그 처럼 사용자 주변의 IoT 기기들이 직접 지능적 데이터 처리를 일부 담당한다.

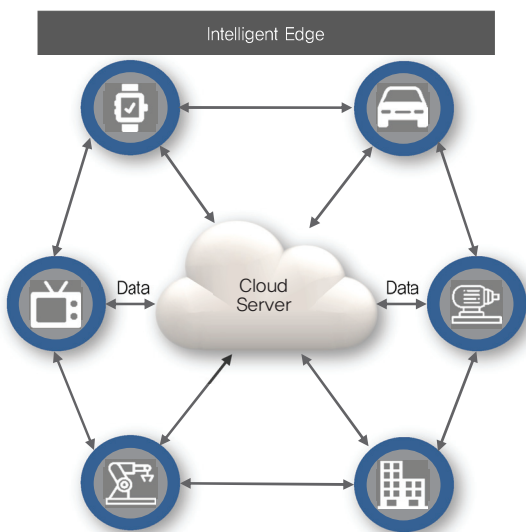
한편 전력소모에 대한 제약조건이 거의 없는 클라우드에서의 데이터 처리와는 달리 게이트 웨이나 IoT 기기들이 에너지 효율적인 지능적 데이터 처리를 위하여 선제되어야 할 것은 최소한의 전력 소모이다. 한정된

에너지원이나 배터리 전력으로 수년 단위의 교체 주기도 극복해야 하는 IoT 기기들이기에 전력소모는 성능보다 더욱 중요한 항목이다. 초저전력을 소모하면서 지능적인 판단을 수행하는 인공지능 반도체를 실현하기 위하여 고성능을 추구하는 종래 딥 뉴럴 네트워크 알고리즘의 반도체 적용은 적절하지 않으며 초저전력 엣지 지능형 반도체를 위한 인공지능 아키텍처가 필수적이다. 현재까지 시도되고 있는 다양한 인공지능 탑재 반도체 들중 초저전력을 추구하는 아키텍처는 생물학적인 뉴로 사이언스 메커니즘에 기반한 스파이킹 뉴럴 네트워크(SNN: Spiking Neural Network)로 알려져 있으며, 스파이크가 발화하고 누적되는 시간적인 정보를 바탕으로 지능적인 정보처리를 수행한다[5].

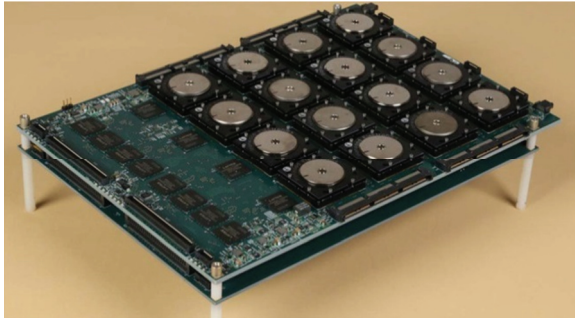
### III. 지능형 반도체 코어 기술 동향

지능형 반도체 코어는 초저전력 엣지에서 수집된 데이터의 효과적인 처리를 위하여 데이터에 대한 일차적인 지능적 분석을 담당하는 부분으로 엣지 프로세싱에 인텔리전스를 부여하기 위한 중추적인 역할을 한다. 아직 엣지 프로세싱의 지능화를 위한 연구가 출발 단계에 있는 만큼 초저전력 인텔리전트 엣지를 위한 전용의 지능형 반도체 코어는 거의 발표된 것이 없으나 기술 발전의 토대를 위한 인공지능 반도체 코어에 대한 연구는 각 분야에 걸쳐 상당 부분 개발이 진행된 상태이다.

인공지능에 관한 연구는 1950년대부터 시작되어 장기간 침체기를 겪은 후 최근 설계 수준의 비약적 향상 및 대용량 데이터 처리 기술의 발달에 따라 다시금 IT 업계 전반의 화두로 등장하게 되었다. 인공지능 반도체 코어란 인간 두뇌의 기능을 모방하여 지능형 서비스를 제공할 수 있는 시스템반도체 융합기술을 의미한다. 2008년 미국 방위고등연구계획국(DARPA)의 SyNAPS 프로그램을 토대로 주요국가 중심으로 관련 기술을 개발하고 있다.



(그림 2) 초저전력 엣지 지능형 반도체 개념



(그림 3) SNN기반 대표적인 인공지능 반도체 코어인 IBM의 TrueNorth 16 chip Board

[출처] By DARPA SyNAPSE [Public domain], via Wikimedia Commons.

2014년 8월 IBM은 SyNAPS 프로그램에 참여하여 트루노스(TrueNorth)라는 SNN 기반 인공지능 코어를 발표하였으며 이는 4,096개의 코어에 100만 개의 뉴런과 2억 5,600만 개의 시냅스를 구현하여 1W 소비전력당 초당 460억회 시냅틱 작동의 전력 효율을 가진다. IBM은 트루노스로 다양한 사물을 식별하는 데 성공하였으며 이는 회로 소자들을 인간의 신경망처럼 연결해 인간의 두뇌와 유사한 활동을 모방했음에 큰 의미를 가진다[5], [(그림 3)참조].

2018년 Intel은 CES 2018을 통해 자가학습(Self-Learning)이 가능한 인간 두뇌 동작을 닮은 로이히(Loihi)라는 SNN 기반 인공지능 코어를 발표하였다. 이는 인간의 문제 해결 능력을 모방하여 13만 개의 뉴런과 1억 3천만개의 시냅스를 구현하였고, 사전에 학습된 결과 기반의 지능적 동작이 가능함 물론 주변 환경으로부터 피드백을 받아 실시간 학습이 가능하도록 다양한 신경망 토폴로지를 지원하는 신경 신호 및 가소성 시냅스를 구현하였다. 로이히는 범용 컴퓨팅대비 1,000배 높은 전력 효율성을 가진다고 알려져 있다[6].

이와 같은 대규모 인공지능망에 대한 응용뿐만 아니라 모바일 AP에서도 인공지능 기능 구현을 위한 NPU 기능을 도입하고 있는데, 2013년 퀄컴은 인간의 지능을 모방하여 인간처럼 사물을 인식하고 주변 환경으로부터

의 학습이 가능한 제로스(Zeroth)라는 NPU를 발표하며 자율 주행 자동차를 대상으로 경쟁력 확보에 노력하였으며 퀄컴의 대표적 모바일 AP인 스냅드래곤 내의 인공지능 처리 엔진의 토대가 되었다[7].

2017년 9월 Huawei사는 Kirin 970 모바일 칩에 인공지능의 주요 성능 중 하나인 추론을 위한 가속 코어인 NPU를 내장함으로써 상용 모바일 AP 최초로 지능형 반도체를 탑재시켰으며, 이를 시작으로 세계 주요 반도체 회사들은 퀄컴의 스냅드래곤, 애플 A11 바이오닉 및 삼성의 엑스노스 9 프로세서 등 각사의 모바일 AP에 인공지능을 위한 NPU를 탑재하여 더욱 정교한 지능적 서비스를 제공하고 있다. (그림 4)는 2018년도에 소개된 주요 모바일 AP를 보여준다. Huawei사는 2018년 8월 독일 베를린에서 개최한 IFA 전시회를 통해 세계 최초 7nm 공정 바탕의 2개의 NPU를 탑재한 Kirin 980을 발표하면서 모바일 AP 시장의 인공지능 기술 경쟁에 속도를 높이고 있으며 애플은 A12 바이오닉을 소개하며 A11 바이오닉에서 출발한 인공지능 수행 능력을 지속적으로 향상시키고 있다. 이제 초저전력 NPU는 모바일 AP의 구성에 있어서 빠질 수 없는 주요 부분으로 자리 잡고 있다[8]-[11].

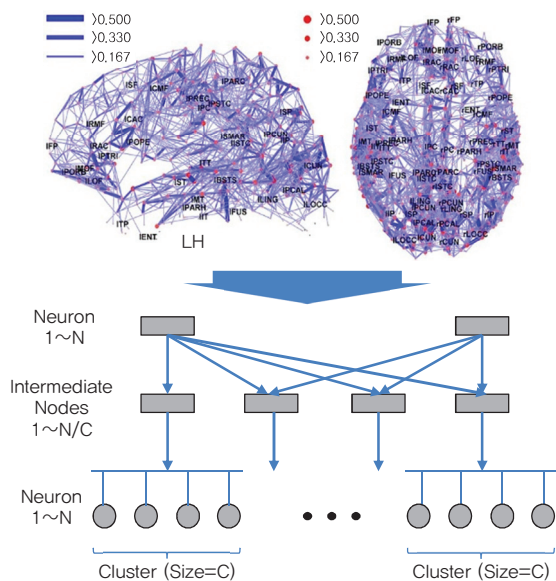
점점 커지는 초저전력 지능형 반도체에 대한 요구에 부응하기 위하여 최근에는 뉴로사이언스 기반 인공지능 알고리즘 기술인 스파이킹 뉴럴 네트워크(SNN)에 대한



(그림 4) 2018년 출시된 주요 모바일 AP

[출처] By Samsung Electronics (<https://twitter.com/samsung-exynos>) [Public domain], via Wikimedia Commons.

[출처] By Rodrigo Garrido [CC BY-SA 4.0 (<https://creativecommons.org/licenses/by-sa/4.0/>)], via Wikimedia Commons.



(그림 5) 뇌 구조 모방 스파이킹 뉴럴 네트워크

관심이 높아지고 있다. (그림 5)에서 볼 수 있듯 스파이킹 뉴럴 네트워크는 기존의 컴퓨터 사이언스에 기반한 선형 병렬 방식이 아닌 자연 상태의 인간 뇌 구조를 모방한 인공 지능 모델로서 기존 모델 대비 높은 에너지 효율과 지능 처리 능력을 보여줄 것으로 기대하고 있다[12].

SNN을 반도체로 구현하기 위해서는 신경 세포의 기초 단위인 뉴런을 모방할 수 있는 하드웨어에 관한 연구가 필요하다. 1952년 Alan Lloyd Hodgkin과 Andrew Fielding Huxley가 제안한 Hodgkin-Huxley 모델의 경우 뉴런 내 이온 전달 매커니즘 분석을 바탕으로 생물학적 뉴런에서 발생할 수 있는 regular spiking, fast spiking, chattering 등 다양한 스파이킹 형태를 모방할 수 있으나 하드웨어로 구현시 너무 높은 복잡성으로 인하여 실제 반도체 구현에는 많이 활용되고 있지는 않다. 오히려 다수의 뉴런을 직접하기 위하여 단순한 스파이킹 기능만을 모사하지만 하드웨어의 구조가 간단한 integrate-and-fire 모델을 중심으로 한 연구가 최근 많이 진행되고 있다. 더불어 뉴런과 뉴런 간의 연결을 위한 시냅스 및 다수의 뉴런 집합체 간의 확장을 위한 비동기식 네트워크 구성의 한 방법인 AER에 기반한 연

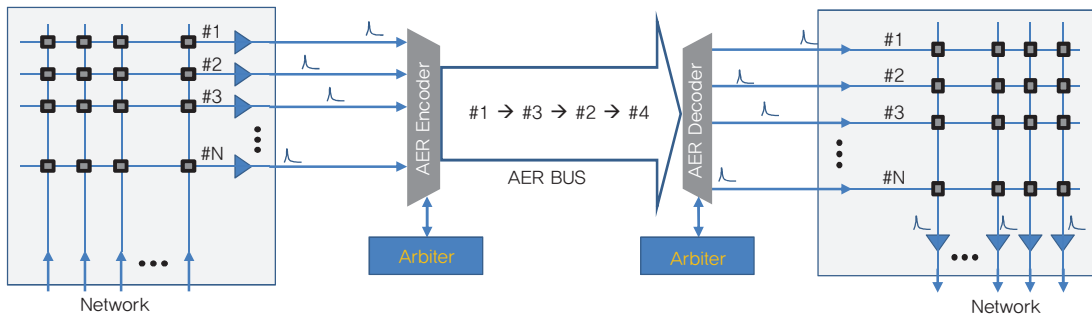
구도 활발히 진행되고 있다.

#### IV. 지능형 반도체를 위한 커넥티비티 기술 동향

대규모 생물학적 신경 세포를 모사한 스파이킹 뉴럴 네트워크 반도체를 구현함에 있어서 유연성과 확장성을 가지는 수많은 뉴런과 시냅스 연결을 2차원 반도체 평면에 그대로 구현하는 것은 매우 어려운 일이다. 예컨대 각 뉴런들은 최소 100에서 1,000개 이상의 뉴런과 시냅스로 연결되는데, 실제 생물학적인 뉴로모픽 시스템을 모방한 SNN에서는 시냅스 연결의 희소성(sparsity)에 따라 뉴런 레이어들 간에 완전 연결 네트워크(Fully-Connected Network)를 사용하지 않고 시냅스 가중치가 일정 값 이상인 경우에만 연결을 허용한다. 이러한 시냅스 연결을 위하여 Dedicated 배선을 사용할 경우 과도한 칩 면적이 필요하고 시스템의 유연성과 확장성도 매우 저하된다. 인간의 생물학적인 뉴런의 발화율(Firing Rate)이 초당 100번 정도이며, 이는 실리콘 칩의 동작 속도에 비하여 상대적으로 매우 느리기 때문에 공유 버스(Bus) 혹은 NoC를 이용하여 구현하는 것이 더 효율적이다.

SNN에서는 뉴런의 스파이크가 발생하는 타이밍이 중요한 정보 인코딩의 수단이므로 시냅스 연결 네트워크는 스파이크를 빠른 시간 내에 목적지 뉴런으로 전달하여야 하며, 시스템 내의 여러 뉴런이 동시다발적으로 스파이크를 발생시키므로 이를 제한된 시간 내에 전달할 수 있는 충분한 통신 대역폭(Bandwidth)을 지원해야 한다. 또한, 처리하고자 하는 문제들에 따라서 달라지는 뉴런 간의 연결 관계를 위하여 SNN의 네트워크는 유연성과 확장성이 요구되며, 생성된 스파이크가 다수의 뉴런으로 전달되도록 멀티 캐스팅(Multi-Casting) 기능이 제공되어야 한다.

Sivilotti와 Mahowald는 서로 다른 칩의 뉴런 간에 스파이크를 전송하기 위해 AER(Address-Event Repre-



(그림 6) Address-Event Representation

sentation) 통신 방식을 제안했다[13], [14]. SNN에서는 스파이크의 타이밍, 시냅스 연결 관계 및 시냅스 가중치를 이용하여 정보를 처리하기 때문에 시냅스 네트워크가 전달해야 할 정보는 스파이크의 타이밍과 스파이크 생성 뉴런의 주소이다. AER은 (그림 6)과 같이 스파이크 발생 뉴런의 주소가 공유 버스를 통해 비동기적으로 다른 배열에 있는 특정 뉴런으로 전달되는 point-to-point 통신 방식이다. 출발지 뉴런의 주소를 목적지 뉴런의 주소로 변환하여 주는 SRT(Synapse Routing Table)를 통해 출발지 뉴런의 주소를 목적지 뉴런의 주소로 변환 함으로써 재구성 가능한 시냅스 연결을 구현할 수 있다[15]. 그러나 대규모 뉴로모픽 시스템에서는 단일 AER 버스의 대역폭에 제한이 있어, 2차원 메쉬(Mesh) 또는 트리(Tree) 등 여러 토폴로지를 복합한 계층적 NoC를 사용하는 경향이 있다.

2차원 메쉬는 높은 통신 대역폭을 제공하지만, 통신 대기 시간(Latency)이 긴 단점이 있다. 트리 토폴로지는 최대 대역폭이 낮은 반면 통신 대기 시간이 짧고 Dead-Lock이 없는 멀티 캐스팅이 가능하다[16]. 각 NoC 토폴로지의 특성이 다르기 때문에 대부분의 시스템에서는 다양한 NoC 토폴로지를 계층적으로 통합하여 효율성을 극대화한다.

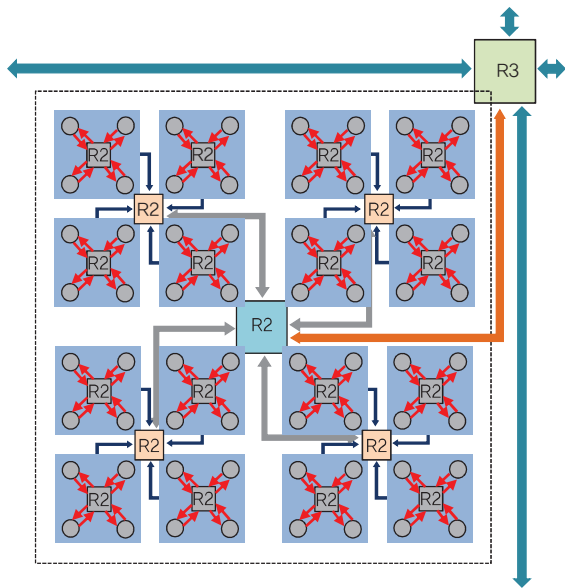
이러한 시냅스 연결 구조의 예로서 H-NoC를 들 수 있는데 이는 스타(Star) 구조와 2차원 메쉬 구조를 계층적으로 결합하여 멀티 캐스팅 통신에 효율적인 NoC 구조이다[17]. 계층 구조상 최하위에는 10개의 뉴런들을

스타 구조로 연결한 Neuron Facility가 있고, 다시 10개의 Neuron Facility들을 스타 구조로 연결한 Tile Facility, 최상위에는 4개의 Tile Facility를 스타 구조로 연결한 Cluster Facility가 있다. 최상위 수준에서는 Cluster Facility들은 2차원 메쉬 NoC로 연결된다. 또한 하나의 Neuron Facility 내에 위치하는 뉴런들로 멀티 캐스팅되는 스파이크들을 하나의 패킷 내에서 비트마스킹(Bitmask)를 이용하여 압축하여 네트워크의 트래픽을 줄일 수 있는 방식을 제안하였다.

맨체스터 대학은 2013년 SNN을 빠르게 시뮬레이션하기 위한 목적으로 SpiNNaker라는 시스템을 발표하였는데, 이는 하나의 칩이 18개의 ARM968 코어들과 라우터 및 주변 회로들이 NoC로 연결된 CMP(Chip Multi-Processor) 구조이며 각 18개의 코어 die는 128MB SDRAM과 die 상태로 스택킹 된 구조를 가진다 [18]. 각 ARM968 코어는 1,000개의 뉴런을 시뮬레이션할 수 있으며, 2차원 토러스(Torus) 네트워크를 사용하여 최대 65,536개의 칩을 연결하여 통합 시뮬레이션 수행이 가능하다.

IBM의 TrueNorth에서는 하나의 코어에 256개의 뉴런이 256×256 크로스바(Crossbar)를 통하여 256개의 입력 시냅스에 연결될 수 있도록 구현되었으며, 하나의 칩에는 64×64의 2차원 메쉬 NoC로 연결된 4,096개의 코어가 구현되어 있다[5]. PCB 상에서는 4×4 메쉬 형태로 최대 16개 칩까지 연결 확장 가능하다.

HiAER는 AER 선형 그리드를 기반으로 계층적 트리



(그림 7) DyNAPs의 트리/메쉬 복합 NoC의 구조

구조의 NoC 네트워크를 사용한다[19]. 이전의 플랫폼 기반 접근법의 한계를 극복하고, 유연성과 확장성을 높이기 위하여 장거리 시냅스 연결을 위한 다차원 트리 기반의 확장된 계층적 AER 시냅스 배선 기법을 제안하였다. 또한 시냅스의 연결 강도뿐만 아니라 소스에서 대상으로 배선되는 이벤트의 타이밍도 프로그래밍 가능하다.

Neurogrid에서는 기본 블록인 Neurocore 내에  $256 \times 256$ 개의 뉴런이 2차원 배열로 구현되고, 각 뉴런은 블록 전체로 브로드캐스팅 가능한 AER 버스를 통해 연결되어 있다[16]. 이러한 Neurocore 들은 다시 상위 수준에서 16개의 뉴런이 트리 네트워크로 연결되는데, 각 Neurocore 내부에는 트리 네트워크를 위한 라우터 및 배선 테이블이 내장되어 있다.

yNAPs의 각 코어 내의 뉴런들은 브로드 캐스팅되는 버스를 통하여 트리 NoC의 라우팅 스위치에 연결되어 있고, 각 코어는 (그림 7)에서와 같이 3단계의 quad-tree 구조와 최상위 수준에서의 2차원 메쉬 구조를 갖는 복합적 계층 구조의 NoC 구조로 통합되어 있다[12]. 특히 단순한 AER 주소 대신에 2단계의 태그(Tag)에 기반한 배선 방식을 제안하여 획기적으로 감소된 배선 테이블

을 하나의 칩 내에 실장 함으로서 외부 메모리 참조로 인한 성능 감소 및 전력 증가를 최소화하였다.

## V. 지능형 반도체를 위한 시뮬레이션 프레임워크 기술 동향

스파이킹 뉴럴 네트워크(SNN) 시뮬레이션 프레임워크는 뉴럴 회로 생성 및 최적화를 위한 이론적인 모델, 실험적인 데이터, 생물학적인 모델을 시뮬레이션하기 위한 도구로서, 사용자가 생성한 스파이크 뉴럴 네트워크를 범용 컴퓨팅 환경에서 시뮬레이션할 수 있도록 다양한 기능들을 제공한다. SNN 시뮬레이션 프레임워크는 뉴로 사이언스, 실시간 응용 로봇, 뉴로모픽 엔지니어링 분야 등에서 활용이 되고 있으며 최근에는 뉴로모픽 하드웨어 분야에 상위 수준 시뮬레이터로 활용이 되고 있다[20]. <표 1>은 대표적인 SNN 시뮬레이션 프레임워크의 종류와 특징을 정리한 표로서, 각 프레임워크가 지원하는 뉴런/시냅스 모델, 시냅틱 가소성, 구현 언어 및 플랫폼 등에 대한 정보를 나타낸다.

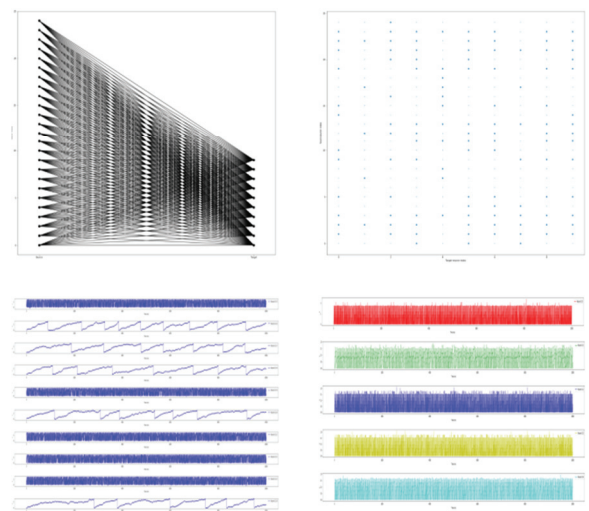
CARLsim은 GPU 가속 라이브러리 기반의 SNN 시뮬레이션 프레임워크이며 생물학적 디테일이 높은 대규모의 스파이크 뉴럴 네트워크를 목표로하는 프레임워크이다. CARLsim을 활용하는 경우, 다수개의 GPU와 x86 CPU를 이용한 시스템 구성이 가능하며, 상기 설명한 바와 같이 대규모 SNN 모델을 시뮬레이션이 가능하다는 장점을 갖는다. 또한, C/C++ 프로그래밍 인터페이스를 제공함으로써 뉴런, 시냅스 및 네트워크 레벨에서 세부정보와 파라미터를 지정할 수 있다. CARLsim의 최신 버전은 2018년 7월 발표된 CARLsim4.0 버전으로, 기존 버전 대비 다중 GPU/CPU 기반의 하이브리드 시뮬레이션을 지원하고, 지원하는 뉴런 모델의 수가 증가하였다.

Brian은 무료 오픈 소스 SNN 시뮬레이션 프레임워크로서 사용자가 뉴런, 시냅스 모델 및 네트워크를 생성하

〈표 1〉 스파이크 뉴럴 네트워크 시뮬레이터의 종류 및 특징 (○: 지원, △: 부분 지원)

|                     |                          | CARLsim<br>4 | Brian<br>2 | NEURON<br>7.5 | GeNN<br>3 | NCS<br>6 | Nemo<br>0.7 | Nengo<br>2.6 | NEST<br>2.14 | PCSIM<br>0.5 |
|---------------------|--------------------------|--------------|------------|---------------|-----------|----------|-------------|--------------|--------------|--------------|
| Neuron model        | Leaky integrate-and-fire | ○            | ○          | ○             | △         | ○        |             | ○            | ○            | ○            |
|                     | Izhikevich 4-param       | ○            | ○          |               | ○         | ○        | ○           | ○            | ○            | ○            |
|                     | Izhikevich 9-param       | ○            | ○          |               | △         |          |             |              |              |              |
|                     | Multi-compartment        | ○            | ○          | ○             |           |          |             |              | ○            |              |
|                     | Hodgkin-huxley           |              | ○          | ○             | ○         | ○        |             | ○            | ○            | ○            |
| Synapse model       | Current-based            | ○            | ○          | ○             | ○         | ○        |             | ○            | ○            | ○            |
|                     | Conductance-based        | ○            | ○          | ○             | ○         | ○        | ○           | ○            | ○            | ○            |
|                     | AMPA, NMDA, GABA         | ○            | ○          | ○             | ○         | △        |             | ○            | ○            | ○            |
|                     | Neuromodulation          | △            | △          | ○             | △         |          | △           | ○            | △            | △            |
| Synaptic plasticity | Short-term plasticity    | ○            | ○          | ○             | △         | ○        | ○           | ○            | ○            | ○            |
|                     | E-STDP                   | ○            | ○          | ○             | ○         | ○        | ○           | △            | ○            | ○            |
|                     | I-STDP                   | ○            | △          | ○             | △         |          | ○           |              | ○            |              |
|                     | DA-STDP                  | ○            | △          | ○             | △         |          |             | △            | ○            | ○            |
|                     | Homeostasis              | ○            | △          | ○             | △         |          |             | ○            | ○            | ○            |
| tools               | Parameter tuning         | ○            | △          |               |           |          |             | ○            |              |              |
|                     | Analysis/visualization   | ○            | △          | ○             |           | △        | ○           | ○            |              | △            |
|                     | Regression suite         | ○            | ○          |               | ○         |          | ○           | ○            | ○            | ○            |
| Integration methods | First-order/exponential  | ○            | ○          | ○             | ○         | ○        | ○           |              | ○            | ○            |
|                     | Exact/crank-nicholson    |              | ○          | ○             |           |          |             |              | ○            |              |
|                     | Runge-kutta              | ○            | ○          |               |           |          |             | ○            | ○            |              |
| Computing hardware  | Single-threaded CPU      | ○            |            | ○             | ○         | ○        | ○           | ○            | ○            | ○            |
|                     | Multi-threaded CPU       | ○            |            | ○             |           | ○        | ○           | ○            | ○            | ○            |
|                     | Distributed              |              | △          | ○             |           | ○        |             | ○            | ○            | ○            |
|                     | Single GPU               | ○            | △          |               | ○         | ○        | ○           | ○            |              |              |
|                     | Multi-GPU                | ○            |            |               |           | ○        |             |              |              |              |
|                     | Hybrid(Multi-GPU/GPU)    | ○            |            |               |           | ○        |             |              |              |              |

고 이를 일정한 시간의 흐름에 따라 시뮬레이션할 수 있다[21]. Brian은 다양한 뉴런 및 시냅스 모델을 지원하는 특징을 가지고 있으며, Python 프로그래밍 언어로 작성되어 구현 및 테스트가 용이하다는 장점을 갖는다. 또한 Linux, Mac OS, Windows 등의 주요한 범용 컴퓨팅 플랫폼을 모두 지원하고 있다. Brian은 Romain Brette(Institut de la Vision, France), Dan Goodman (Imperial College, United Kingdom), Marcel Stimberg (Institut de la Vision, France) 등이 공동으로 참여하여 개발하였으며 2018년 6월에 최신 버전인 Brian 2.1.3 버전을 발표하였다. (그림 8)은 Brian을 이용하여 다양한 형태로 시뮬레이션 결과를 출력한 예제이다[22].



(그림 8) Brian을 이용한 다양한 결과 출력 예제

## VI. 결론

인터넷의 시대를 뛰어넘어 모든 사물이 네트워크에 연결되는 IoT의 시대가 도래함에 따라 종래 클라우드 컴퓨팅의 기술로는 감당하기 어려웠던 데이터 트래픽과 지연(latency)의 증가, 그리고 개인 프라이버시 침해 리스크 증가에 대한 솔루션 요구가 나타나게 되었다. 이러한 요구에 대한 해법으로 엣지 컴퓨팅 개념이 등장하게 되었으며 더 나아가 IoT 디바이스 자체가 스스로 지능적 판단을 수행하고 데이터를 처리하는 초저전력 엣지 지능형 반도체에 대한 개념이 수립되었다. 이러한 지능수행의 핵심 요소로 인공지능 기술 기반의 지능형 반도체가 있으며 스파이킹 뉴럴 네트워크를 기반으로 한 지능형 반도체 코어 기술, 코어 내 뉴런과 시냅스의 효율적인 연결을 위한 커넥티비티 기술, 대규모의 스파이킹 뉴럴 네트워크의 성능 예측을 위한 시뮬레이터 기술 등이 해당 기술의 기반을 형성한다. 아직은 출발점에서 시작하고 있는 초저전력 엣지 지능형 반도체 기술이지만 각 기술에 대한 최신 동향을 보건데 향후 도래하는 IoT 기기의 시장에서 인공지능 기술 기반의 지능형 반도체 기술은 여타 일반적인 IoT 기기와는 다른 뚜렷한 차별화를 도출할 수 있는 핵심 기술로 자리 잡을 것으로 예측된다.

## 약어 정리

|       |   |
|-------|---|
| AER   | Address-Event Representation                |
| AP    | Application Processor                       |
| CES   | The International Consumer Electronics Show |
| CMP   | Chip Multi-Processor                        |
| CPU   | Central processing unit                     |
| DARPA | Defense Advanced Research Projects Agency   |
| GPU   | Graphics processing unit                    |
| IoT   | Internet of Things                          |
| NoC   | Network-on-Chip                             |
| NPU   | Neural Processing Unit                      |

|        |  |
|--------|--|
| PCB    | Printed Circuit Boards                                       |
| SDRAM  | Synchronous DRAM   |
| SNN    | Spiking Neural Network                                       |
| SRT    | Synapse Routing Table  |
| SyNAPS | System of Neuromorphic Adaptive Plastic Scalable Electronics |

## 참고문헌

- [1] T. Barnett Jr. et al., "Cisco Global Cloud Index 2015-2020," Cisco Knowledge Network (CKN) Session, Nove. 2016. [https://www.cisco.com/c/dam/m/en\\_us/service-provider/ciscoknowledgenetwork/files/622\\_11\\_15-16-Cisco\\_GCI\\_CKN\\_2015-2020\\_AMER\\_EMEAR\\_NOV2016.pdf](https://www.cisco.com/c/dam/m/en_us/service-provider/ciscoknowledgenetwork/files/622_11_15-16-Cisco_GCI_CKN_2015-2020_AMER_EMEAR_NOV2016.pdf)
- [2] Qualcomm, "5G - Vision for the Next Generation of Connectivity," Mar. 2015. <https://www.qualcomm.com/media/documents/files/whitepaper-5g-vision-for-the-next-generation-of-connectivity.pdf>
- [3] R. De La Mora et al., "Cisco IOx: Making Fog Real for IoT," Feb., 2014. <https://blogs.cisco.com/digital/cisco-iox-making-fog-real-for-iot>.
- [4] R. Samuel et al., "Edge Processing: How Qualcomm is Helping Build the Industrial IoT," June 2017. <https://www.qualcomm.com/news/onq/2017/06/28/edge-processing-how-qualcomm-helping-build-industrial-iot>
- [5] F. Akopyan et al., "TrueNorth: Design and Tool Flow of a 65mW 1 Million Neuron Programmable Neurosynaptic Chip," *IEEE Trans.Comput.-Aided Des. Integr. Circuits Syst.*, vol. 34, no. 10, Oct. 2015, pp. 1537-1557.
- [6] M. Davies et al., "Loihi: A Neuromorphic Manycore Processor with On-Chip Learning," *IEEE Micro*, vol. 38, no. 1, Jan. Feb. 2018, pp. 82-99.
- [7] Qualcomm, "Introducing Qualcomm Zeroth Processors: Brain-Inspired Computing," OnQ Blog, Oct. 10, 2013. <https://www.qualcomm.com/news/onq/2013/10/10/introducing-qualcomm-zeroth-processors-brain-inspired-computing>
- [8] Qualcomm, <https://www.qualcomm.com/products/mobile-processors>
- [9] Samsung, <https://www.samsung.com/semiconductor/minisite/exynos/>
- [10] Huawei, <https://consumer.huawei.com/en/campaign/kirin980/>
- [11] Apple, "A12 Bionic," <https://www.apple.com/kr/iphone->

xs/a12-bionic/

- [12] S. Moradi et al., "A Scalable Multicore Architecture with Heterogeneous Memory Structures for Dynamic Neuromorphic Asynchronous Processors (DYNAPs)," *IEEE Trans. Biomedical Circuits Syst.*, vol. 12, no. 1, Feb. 2018, pp. 106-122.
- [13] M. Sivilotti, "Wiring Considerations in Analog VLSI Systems with Application to Field-Programmable Networks," Ph.D. dissertation, Computation and Neural Systems, California Inst. Technol., Pasadena, CA, USA, 1991.
- [14] M. Mahowald, "VLSI Analogs of Neuronal Visual Processing: A Synthesis of Form and Function," Ph.D. dissertation, Computation and Neural Systems, California Inst. Technol., Pasadena, CA, USA, 1992.
- [15] G. Indiveri et al., "A Reconfigurable Neuromorphic VLSI Multi-chip System Applied to Visual Motion Computation," in Proc. Int. Conf. *Microelectron. Neural, Fuzzy Bio-Inspired Syst.*, Granada, Spain, Apr. 1999, pp. 37-44.
- [16] B.V. Benjamin et al., "Neurogrid: A Mixed-Analog-Digital Multichip System for Large-Scale Neural Simulations," *Proc. IEEE*, vol. 102, no. 5, May 2014, pp. 699-716.
- [17] S. Carrillo et al., "Scalable Hierarchical Network-on-Chip Architecture for Spiking Neural Network Hardware Implementations," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 12, Dec. 2013, pp. 2451-246.
- [18] E. Painkras et al., "SpiNNaker: A 1-W 18-Core System-on-chip for Massively-Parallel Neural Network Simulation," *IEEE J. Solid-State Circuits*, vol. 48, no. 8, Aug. 2013, pp. 1943-1953.
- [19] J. Park et al., "Hierarchical address event routing for reconfigurable large-scale neuromorphic systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, Oct. 2017, pp. 2408-2422.
- [20] Cognitive Anteatr Robotics Laboratory, "CARLsim: a GPU-Accelerated SNN Simulator," Mar. 2017. <http://www.socsci.uci.edu/~jkrichma/CARLsim/>
- [21] The Neural Simulation Technology Initiative, <http://nest-simulator.org/>
- [22] Brian, "The Brian Spiking Neural Network Simulator," <http://briansimulator.org/>