Neuromorphic Architecture 및 CAD 연구 동향 -두뇌의 리버스 엔지니어링

조 준 동
성균관대학교

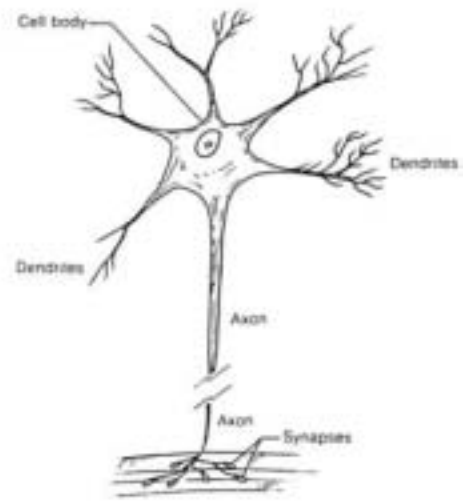
최근 3D 휴먼 팩터가 이슈가 되면서 인간 3D 정보처리 시스템에 대한 관심이 증가하고 있다. 즉, 인간 두뇌의 3D 정보 처리 시스템은 순차적으로 정보를 처리하면서 동시에 병렬적으로 처리한다.

또한 머신 비젼 분야에서 얼굴과 같은 물체를 인식할 때 기계가 하기 어려운 일은 크기의 차이, 방향, 빛, 영상의 복잡도에 따른 변위를 다루는 일이다. 또한 다른 예로는 효과적인 음성인식, 계획, 그리고 깊이, 재질 및 색채 인식이 있다. 머신 비전은 제한적으로 잘 명시된 분야, 예를 들면 광학적 문자 인식 또는 지문 인식등은 잘 인식할 수 있으나, 혼란스러운 영상과 같은 제약되지 않은 문제들에 대해서는 인식하기 어려운 단점이 있다. 한편 사람의 뇌와 같은 생물체의 인식은 시각 데이터로부터 충분한 정보를 추출하여 추론을 통해서 인식하게 된다. 인간의 두뇌에서 시각 정보는 상위 수준 지식과 여러가지 센서 형태들을 결합하여 추론 과정을 통하여 해결 공간상에 제약을 두어 인식을 가능하게 한다.

본 논고는 다음과 같이 구성된다. 1장에는 두뇌의 시각 정보 전달 체계, 2장에는 두뇌 정보 전달 모델링, 3장에는 두뇌 정보 전달 구조에 대해서 알아 보고, 4장에서 결론을 맺는다.

Ⅰ. 두뇌의 시각 정보 전달 체계

뇌의 후두엽은 시각연합영역(visual association area)과 시각피질(Visual Cortex)이라고 하는 시각중추(the visual center)가 있어 시각정보의 처리를 담당한다. 외부로 부터 빛, 자극과 같은 영상 정보는 후두엽에서 초기 시각 정보 특질(feature)을 처리한다. 눈으로 들어온 시각정보가 시각피질에 도착하면 좀 더 복잡한 정보, 즉, 모양과 깊이, 운동, 색등을 처리하는 중간 단계를 거친다. 그리고 이



〈그림 1〉 생물학적인 뉴론 모델 ("Phenomenon of Science" by Valentin Turchin, 1977)

를 과거의 기억속의 특정 대상과의 합치 여부를 따져보는 고차 인지과정이 진행된다. 각각의 단계는 병렬처리하여 짧은 시간에 인식이 가능하게 된다. 우리의 뇌가 이러한 활동을 하는데 있어서 한 개의 신경세포(Neuron)는 수천, 수만 개의 신경세포와 정보를 주고받고 있다.

뉴론은 수상돌기(dentries; inputs)과 축삭돌기(axon; output)으로 구성되어 있으며, axon은 synapse를 통하여 다른 뉴론의 dentries와 연결된다. 각 뉴론은 전기적인 포텐셜을 가지고 있으며, 이 포텐셜이 어느정도의 크기에 다다르면 뉴론은 신호를 연결된 인접 뉴론에 전달한다. 이 연결선을 synapse라고 부른다. synapse는 신호를 수정하여 보낸다. Spiking 뉴론의 Izhikevich model을 사용한다. Spike의 타이밍에 따라 synaptic weight가 줄어들거나 늘어난다. Spike-timing-dependent plasticity[5] 알고리즘을 사용한다. 즉, pre-synaptic과 post-synaptic spike 사이의 시간차에 따라서 synaptic weight가 변한다.

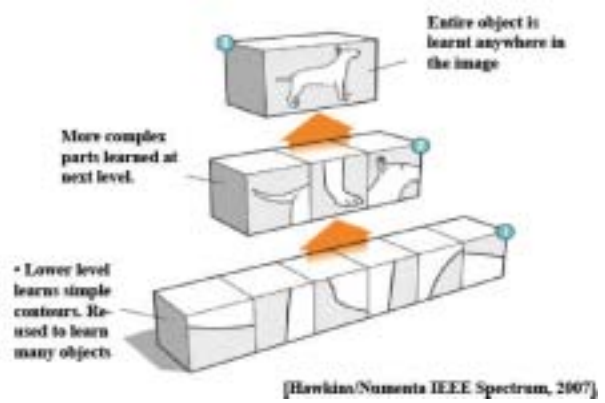
세포 뉴론들은 서로 방향성을 가지고 연결되어 있다. 세포 I에서 세포 j로 연결된 시냅스는 I(presynaptic)가 먼저 발화되고 j(postsynaptic)가 나중에 발화될 때 강화되고, j가 먼저 발화하고 i가 나중에 발화될 때 약화된니다. 여기서 I의 발화시점과 j의 발화시점의 차이가 너무 크면 시냅스 세기에 영향을 주지 못하고 어느정도 가까운 경우에만 영향을 준다고 합니다. 이렇게 시냅스의 세기가 변화하는 것을 시냅스 가소성(synaptic plasticity)이라고 한다. 특히 앞서 언급한 발화시간 의존적인 가소성을 발화시간 기반 가소성(Spike-timing-dependent plasticity)이라고 한다.

전접합부(pre-synaptic)의 발화 시간 tpost과 후접합부(post-synaptic)의 발화 시간 tpre 의 시간 차 Δt= tpost - tpre에 따라 시냅스의 크기는 다음과 같이 변화한다:

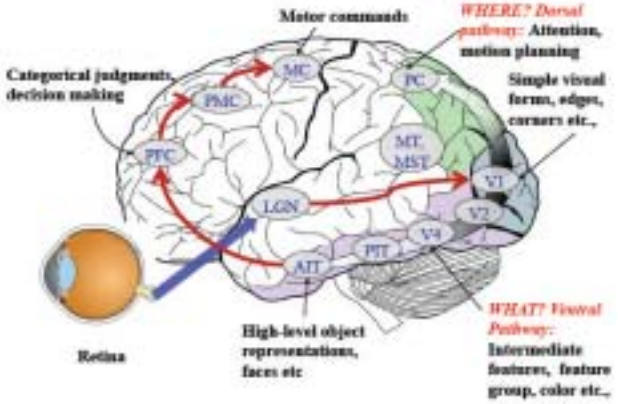
$$F(\Delta t) = \begin{cases} A_+ e^{\frac{\Delta t}{\tau_+}} & \Delta t < 0, \\ -A_- e^{\frac{-\Delta t}{\tau_-}} & \Delta t \geq 0. \end{cases} \quad (1)$$

여기서 A_+는 전접합부의 발화가 후접합부의 발화보다 먼저 일어날 때, 접합부의 최대 강화치를 나타내고, 마찬가지로 A-는 후접합부의 발화가 먼저 일어날 때 최대 약화치를 나타낸다. 신경접합부에서 신경세포 'A'의 발화가 통계적으로 신경세포 'B'의 발화보다 먼저 일어나면 A->B 신경접합부의 크기는 증가하고, B->신경접합부 크기는 점차 감소하는 인과 관계가 형성된다.

세포 뉴런은 크게 흥분성(Excitatory Cortical Cell)과 억제성(Inhibitory Cortical Cell)로 나누어 진다. 흥분성 뉴런의 경우 자신이 받은 전류들을 모아 자신과 연결된 뉴런들에게 전달해주는 역할을 한다. 반면, 억제성 뉴런의 경우에는 자신이 받은 전류들을 자신과 연결된 뉴런들에게 전달해주지 않는다. 뉴런의 점화는 자극성 입력과 억제성 입력의 가중화된 (weighted) 합에 좌우된다. 활성화된 전위차가 발화를 일으키면, 이는 다시 다음 단계의 신경세포로 전달되는 과정을 반복하게 된다. 즉 한 신경세포의 발화는 그 이전 단계에 들어왔던 모든 외부 환경변화를 나타내는 정보가 해당 신경세포에서 성공적으로 처리되어서 또 다른 정보처리 단계로 넘어갈 수 있게 됨을 뜻하고, 만약에



〈그림 2〉 계층적 구조의 인식 체계



〈그림 3〉 두뇌의 시각 정보 전달 경로

발화가 일어나지 않는다면, 해당 신경세포에서 그 이전의 모든 정보 신호가 더 이상 처리되지 않고 잃어버리게 됨을 의미하게 된다.

사람의 뇌는 1000억개의 뉴론으로 되어 있고, 각 뉴론은 수만개의 다른 뉴론에 연결되어 있다. 각 뉴론은 신호 (즉, spike)를 10Hz의 속도로 보낸다. 각 뉴론은 컴퓨터 네트워크에서와 같이 메시지를 송수신하는 객체로써 설계될 수 있다.

이러한 정보 교환을 담당하고 있는 것이 신경전달물질 (neuro transmitter)이다. 신경전달물질로는 도파민(dopamine), 세로토닌(serotonin), 아세틸콜린(acetylcholine) 등이 있다.

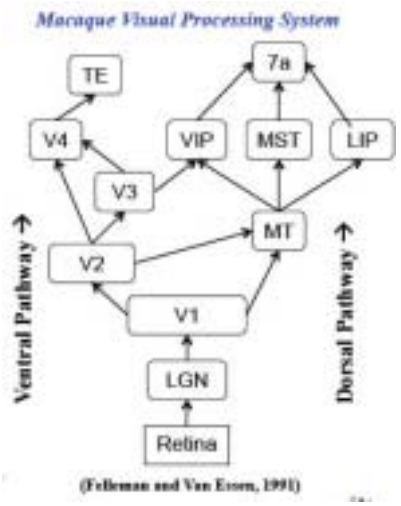
세로토닌은 쾌와 불쾌의 균형을 유지하며 도파민은 각성제(stimulant)이며 쾌의 원인이 된다. 아세틸콜린은 기억과 지적기능을 담당한다.

신경전달 물질은 시냅스(synapse)들 간에 정보를 교환해주는데 전달물질에 따라 수용체가 다르다. 신경전달물질이라고 하는 열쇠가 수용체(receptor)라고 하는 열쇠구멍에 맞게 결합함으로써 다음 신경세포막에 있는 대문이 열려 정보가 전달될 수 있는 것이다. 각각의 신경전달물질들은 각자 특유의 수용체 분자하고만 결합하여 특정정보를 전달한다. 특수 단백질(protein) 분자(molecule)의 결합으로 정신기능에서부터 행동·감정에 이르기까지 모든 것이 결정 되는 것이다. 신경계의 두뇌가 컴퓨터와 다른 점이 있다면 컴퓨터는 하드웨어가 공정되어 있으며, 소프트웨어 프로그램을 교체하여 다른 일을 할 수 있다. 그러나 인간의 의식은 두뇌에 저장 될 때 단순히 신경세포에 저장되는 것이 아니라 신경세포의 시냅스들을 새롭게 연결하면서 새로운 신경회로를 만들어 나간다는 점이 다르다. 인간의 의식이 진화하는 것처럼 컴퓨터를 인간과 같이 만들려면 컴퓨터 소프트웨어가 두뇌의 신경계처럼 스스로 개조해 나갈 수 있어야 한다.

비전 인식체계는 아래 그림과 같이 feed forward 형태로 되어 있다.

계층의 하위수준에 있는 간단한 구조로부터 학습을 통하여 복잡한 구조를 얻는 구조로 되어 있다. 또한 피드백 루프가 존재하지 않는다. 즉, 하위수준에서의 예측 및 판단이 중요하다는 것을 알 수 있다.

우리 뇌는 시각정보를 처리할 때 "물체가 어디에 있는가"와 "물체가 무슨 물체인가"를 각각 따로 처리한다. 각각, 뇌의 위쪽 경로를 따라 처리되는 "어디에"의 경로를 where stream 또는 dosal stream, 또는 parietal stream이라 하고 "무엇?"을 알아내는 경로를 what stream 또는 ventral stream 또는 temporal stream이라 한다. Primary Visual Cortex인 V1은 시각 정보를 dorsal 과 ventral의 두가지 경로에 전달한다. Where stream은 motion을 분석하거나 물체의 위치를 분석하는 경로, what stream은 물체인식을 수행하는 경로이다. 각각의 최종노드는 dosal(where) stream의 경우 정수리께의 MT, ventral stream의 경



〈그림 4〉 Macaque 원숭이의 시각 처리 시스템

우는 관자놀이의 IT(Inferotemporal) cortex이다. 이 둘은 당연히 서로 정보를 교환하며 작동되기는 한데 where가 먼저 처리된다. 이러한 인식 작업은 12.5ms에서 100ms내에서 처리된다.

뇌는 massive parallelism (1011 neurons), massive connectivity (1015 synapses), 전력 효율이 탁월하며 (~ 20 W for 1016 flops), 저성능의 컴포넌트 (~ 100 Hz), 저속도의 통신 (~ meters/sec), 저 정밀도 synaptic 연결, 확률적인 반응과 fault-tolerant하며 자발적인 학습이 가능한 구조로 되어 있다.

〈표1〉에 의하면 거칠은 정도의 통계적 생물신경시스템은 현대 전자공학에 의해서 구현되는 것이 가능하다고 생각할 수 있다.

〈표2〉의 비교를 보면, 뇌는 Watson에 비해서 $1.25x10^6$배 전력대 성능비가 우수하다. 그 이유는 대부분이 와이어의 전력소모를 줄이기 위해서 글로벌 와이어를 사용하지 않고 로컬 와이어를 사용하여 계산과 통신을 하기 때문이다. 또한 정보 계산 처리는 물리적 프리미티브를 사용하여 아날로그 방식을 사용하며, 시그널 복원을 위해서 sparse한 디지털 통신 방식을 사용한다.

〈표 1〉 Human NeoCortex와 Neuromorphic Electronics

Human NeoCortex	Neuromorphic Electronics
$\sim 10^{10}$ synapses/cm^2	10^{10} intersection/cm^2 in crossbar arrays w/ 100 nm pitch
$\sim 10^6$ Neurons/cm^2	$\sim 5x10^8$ transistors/cm^2 in state of the art CMOS
$\sim 5 \times 10^8$ long range axons @ ~1 Hz	~30 Gbit/sec multiplexed digital addressing

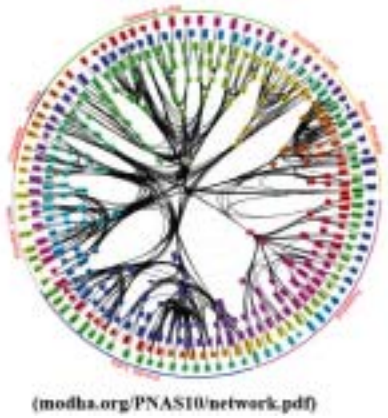
Ⅱ. 두뇌의 정보 전달 모델링

라지 스케일 인지 시뮬레이션은 최근 들어 연산 뇌과학, 시뮬레이션 방법론 및 슈퍼컴퓨팅 분야의 학제간 연구분야이다. 두뇌 인지 컴퓨터 분야에서 인지 시뮬레이터는 두뇌구조, 동작 및 기능에 대한 가설 실험을 가능하게 하는 중요한 기술이다. 시뮬레이션은 또한 새로운 시냅스 나노 디바이스를 사용한 컴팩트 저전력 뉴로모픽 시냅트로닉 칩들의 혁신적 시스템을 불러일으키고자 하는 야심 찬 목표를 갖고 있는 DARPA's SyNAPSE (Systems of Neuromorphic Adaptive Plastic Scalable Electronics) 프로그램 같은 최첨단 연구 항목들의 집합체이다.

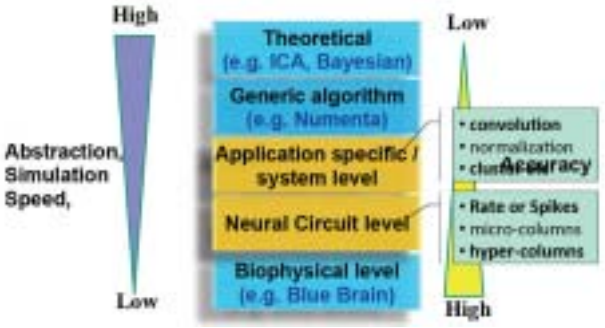
아래 그림은 "The Mandala of the Mind"라고 불리

〈표 2〉 뇌와 컴퓨터의 비교

	Brain	Watson(Morphy the robot)	Road-runner (IBM)
복잡도	100 Billion neurons	10 racks of IBM Power 750 servers with 2,880 processor cores	6,912 AMDx2 12,960 IBM CELL
성능	1 KHz (synaptic rate) 100 Peta FLOPS	80 Tera FLOPS	1.7 Peta FLOPS
전력 소모	20 Watts for 10PF	200 KW for 1PF	3.9 MW Power for 1PF
메모리	750 gigabytes to 6.4 terabytes	15 terabytes of RAM	107 TB memory



〈그림 5〉 Macaque 원숭이 뇌의 네트워크 모델



〈그림 6〉 뇌의 검증 모델 수준 계층도

우는 Macaque 원숭이의 뇌에 대한 network 모델[1]이며 cortex, thalamus, basal ganglia에 걸친, 383개의 vertex는 뇌의 영역을 표현하며, 6,602개의 edges는 각 영역을 연결하는 긴 거리 연결선을 의미한다. 계층적으로 Cluster를 구성한 것을 알 수 있다.

원숭이의 경우 층당 10 ms으로 10개의 층에 대해서 80-100 ms의 지연시간을 갖는다. 발화율은 0-100이며, 네트워크가 넓게 펴져 있으므로 Massive parallel processing이 가능하며, Wave pipelining과 같은 형태로 빠른 시간에 spikes를 전파할 수 있다.

뇌의 검증 모델은 다음과 같은 다섯 가지 수준으로 분류된다.

1) Mathematical/Theoretical Analysis (e.g., Bayesian coding hypothesis/probability density functions, ICA, Sparse coding, Neural Darwinism) 2) brain-inspired generic algorithm for computer vision, robotics and learning (e.g., Numenta) 3) neurobiological application specific system level for motor control, vision, audition (e.g., MIT H-Max), 4) Neural circuit level 5) Brain-inspired biophysical/cellular level (e.g., EPFL Blue Brain).

그러한 뇌의 네트워크에 대한 모델은 추상 레벨이 높은 상위 수준일수록 검증시간은 빠르게 되는 반면 검증이 정확도를 떨어지게 된다.

현재 블루 브레인은 세포 수준에서 정확한 뉴런을 10만 개 정도 모은 수준까지 이르렀다. 앞으로 컴퓨터의 계산 능력이 100만 배 더 증가하게 되면 인간의 두뇌 전체를 시뮬레이션 할 수 있는 파워를 갖게 된다. Neural circuit 시뮬레이션 도구로는 IBM cortical simulator, UCI GPU-SNN등이 있다.

Ⅲ. 두뇌의 정보 전달 하드웨어 구현

1. Spiking Neural Network architecture (SpiNNaker)

ARM 프로세서 코어의 선구자인 맨체스터 대학의 Steve Furber 교수는 2005년에 아날로그 회로를 대신해서 프로그래밍이 가능한 일반 목적 디지털 컴퓨터를 가지고 인간의 뇌를 모델링함으로써 디지털과 아날로그 접근 방식의 단점들을 보상하는 과제에 착수하였다. 영국 맨체스터 · 캠브리지 · 셰필드 대학 등은 800만달러의 정부지원을 받아 Spiking Neural Network architecture (SpiNNaker)[3] 프로젝트를 수행해왔다. 뇌과학의 연구 방향으로는, 1) 뇌의 생물학적인 구조 자체를 모방하거나 (to model the biological structure of the brain - hard AI); 2)뇌의 문제 해결 방법을 모사해서 진행하는 방법 (to

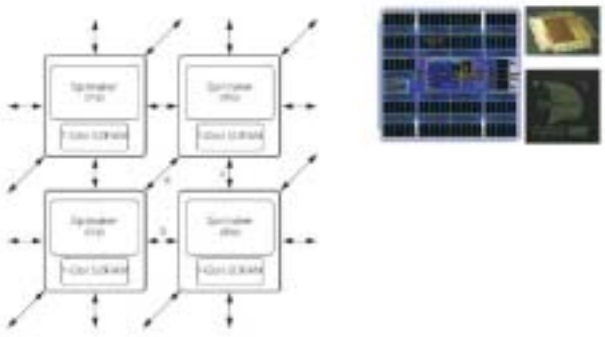
model the problem solving process -soft AI)가 있다. SpiNNaker project는 전자에 해당한다. 즉, SpiNNaker는 event 중심의 계산 방식을 따르는 실시간 범용 신경회로망을 시뮬레이션 하는 하드웨어 기반의 시뮬레이터이다.

1. 기본 병렬: 각 신경 세포는 대량 병렬 시스템 안의 원시적인 논리 계산을 할 수 있다. 이와 마찬가지로, SpiNNaker도 병렬 계산을 사용한다.
2. Spiking Communication: 생물학에서는 신경세포들은 스파이크 (40bit)를 통해서 통신한다. SpiNNaker는 뉴런 신호와 동일하게 전송하기 위해 소스 기반의 Address Event Representation (AER) packet을 사용한다.
3. 이벤트 중심의 행동: 신경세포는 매우 효율적이고 현대 하드웨어보다 적은 전력을 소비한다. 전력 소모를 줄이기 위해 이벤트를 기다리는 "수면" 상태를 하드웨어에 넣었다.
4. 분산 메모리: 생물학에서는, 뉴런은 입력 자극을 처리하기 위해 지역 정보만을 사용한다. SpiNNaker 아키텍처에서는 각 코어에는 지역 메모리를 사용하고, 각 칩에는 SDRAM을 local로 사용한다.
5. 재구성: 생물학에서는, 시냅스는 가소성이 좋다. 이 것은 신경 연결 형태와 연결 강도에서 변화할 수 있음을 뜻한다. SpiNNaker 아키텍처는 작동중에 재구성을 허락한다.

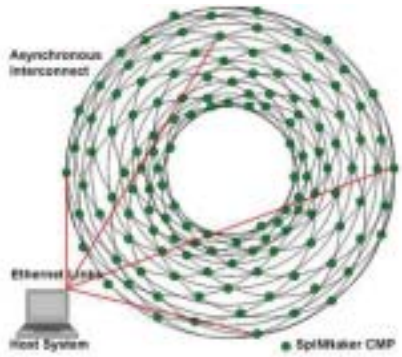
SpiNNaker 프로젝트는 뇌 신경세포인 뉴런과 같은 구조를 만들기 위하여 2011년에 72개의 프로세서 코어를 사용한 바 있으며, 2013년까지 57600개의 칩 [10] (칩 하나에 18개의 ARM9 코어를 포함) 기술을 이용해 만들고 있다. 한 개의 칩 중앙에는 한 코어들로부터 패킷을 받기 위한 링크를 설정하기 위하여 라우터가 배치 되어 있으며, 1천6백만개의 시냅스 연결을 위한 정보를 가지고 있는 마이크론의 1기가비트(Gb) DDR SDRAM이 사용됐다. 사람의 뇌는 1천억 개의 뉴런이 서로 1천조개의 연결을 이룬다. SpiNNaker는 신경과학자, 심리학자 및 의사들이 복잡한 두뇌 부상, 질병 및 상태를 이해하는데 도움을 줄 수 있는 중요한 툴이 될 것이다. 그래서 가장 효율적인 치료법을 알아낼 수 있도록 도와줄 수 있을 것이다. 하지만 100만개의 ARM 프로세서 코어를 사용한다고 하도라도 겨우 사람 뇌의 1% 정도밖에 다루지 못한다.

두뇌에서 뉴런은 어떤 자극이 전기신호로 변환해 이동하는 통로 역할을 한다. 이 시스템에서, 뉴런들은 매우 작은 전기신호를 전달하는 스파이크를 방출하게 된다. 이러한 전기 신호는 뉴런 역할을 하는 ARM 프로세서 코어들 사이에 스파이크 이벤트를 통하여 전달되며, NoC 구조를 이용하여 연결을 최소화하면서 반응속도를 실제 두뇌의 뉴런 간 전달 속도만큼 높이도록 설계했다. 각 임펄스는 SpiNNaker에서 데이터의 패킷으로서 모델화되었다. 패킷은 소스 뉴론의 주소를 포함하며 연결된 다른 뉴런들에게 보내지게 된다. 칩 사이의 링크에 문제가 있을때는 인접 노드를 거쳐서 패킷을 재 라우팅하는 fault tolernance 기능을 가지고 있다. 또한 전력소모를 줄이기 위해서 spike가 입력되지 않을 때는 코어를 sleep모드로 전환시킨다. 신호처리는 neuron과 dendritic tree에 의해서 수행되고, 신호통신은 스파이크를 사용하여 전달하며, 발화율, 발화순서, 또는 시간적인 코드를 이용하여 부호화된다. 메모리는 Synapse와 axon (delay lines)이 있으며 axon은 무손실의 케이블이며, 고정된 지연시간을 가지고 있으며 지연시간의 범위는 1~20ms이다. Axon은 메시지 또는 정보 패킷을 저장하는 메모리 라인의 역할을 한다. Billion 단위의 Neural Network 모델을 Simulation하기 위해서는 엄청난 Processing Power가 필요한데, 이를 위한 Massively Parallel한 NoC 구조가 필요하다.

SpiNNaker 칩은 두 개의 NoC를 가지고 있다. 시스템 NoC는 20개의 ARM 코어에 의해서 공유되는 off-chip 1Gbit SDRAM이다. Low-power processor인 ARM968 Core (200MHz)는 1000개



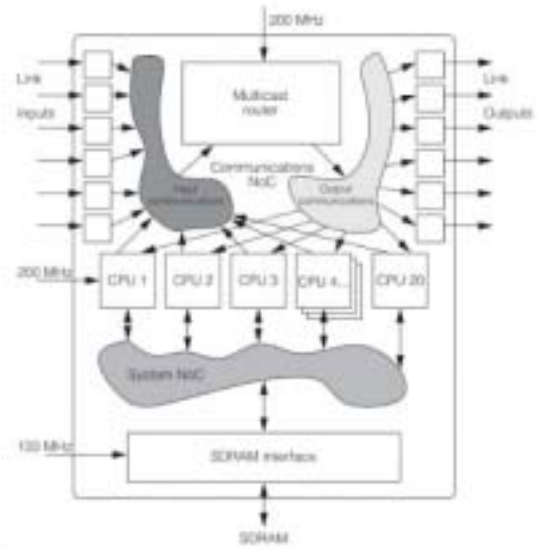
〈그림 7〉 SpiNNaker 시스템의 노드들 (Chip+RAM), 그리고 SpiNNaker Chip Layout (Die Size는 10x10 mm)



〈그림 8〉 SpiNNaker의 Toroidal interconnection mesh

의 뉴론을 실시간에 모델링한다. 20개의 코어 중 19개가 각각 1000개의 뉴론을 모델링할 수 있다. 나머지 한 개의 코어는 관리 작업을 수행한다. 각 코어는 지역 명령과 data가 들어있는 96 KB TCM(Tightly Coupled Memory)를 가지고 있다. 거기다, 공유되는 128M SDRAM은 Chip의 모든 코어들이 사용할 수 있다. 시스템 NoC는 SDRAM을 접근하기 위해서 8Gbps의 밴도폭을 가진다. 또한 이더넷 인터페이스가 외부의 컴퓨터가 SpiNNaker 네트워크에 접근하기 위해서 필요하며 라우팅 테이블을 초기화하고, 뉴럴 네트워크를 모니터하고 구성하는데 사용된다.

또 다른 NoC는 통신 NoC이며 각 코어가 다른 on chip 또는 off-chip 코어에 연결되도록 하는데 사용된다. 이 칩은 GALS (Globally Asynchronous, Locally Synchronous) Timing model을 사용하며, Delay Insensitive 통신 방식을 사용한다. GALS 비동기로 동작하기 때문에 각 프로세서는 각각의 클록으로 동작한다. 즉, Globally synchronization 과정이 필요 없이 각자의 core가 자유롭게 동작하므로 synchronization에 필요한 circuit 절약할 수 있다. Routing table을 이용한 NoC communication는 on-chip 그리고 chip 간의 통신을 할 수 있는 멀티 캐스트 packet router를 사용하여, 주요 작업은 source neuron의 식별자를 기반으로 목적 코어에게 네트워크 패킷을 전달하는 것이다. 신경 세포가 발생하면, packet이 생성되고 네트워크를 통해 전달된다. packet에 포함된 정보는 발사한 신경 세포의 식별자이다. 해당 목적지로 packet을 전달하는데 필요한 정보는 각 router의 routing 테이블 존재에 포함되어 있다. router가 packet을 수신할 때, 다음 단계의 방향을 결정하는 것은 routing 테이블에 보인다. 각 routing 항목은 세 레지스터, Key, Mask, direction에 의해 형성된다. Key field는 선택될 수 있는 항목에 대해 일치하는 routing 키를 식별한다. Mask field는 선택될 수 있는 항목에 대해 받은 packet의 routing 키와 일치해야 하는 키의 bit들을 식별한다. 만약 masked된 항목이 받은 routing 키와 일치하는 경우, direction 레지스터는 packet을 전파하는 방향을 포함한다. 만약 받은 routing 키와 일치하지 않는 경우, "기본 routing"을 위한 메커니즘은 실행되고, packet을 받았던 쪽에서 반대편 링크로 전



〈그림 9〉 System NoC와 Communication NoC 구조

달된다.

사람 두뇌의 1%를 시뮬레이션하기 위해서는 10억 개의 뉴론을 시뮬레이션할 수 있어야 한다. 따라서 한 개의 칩이 19000개를 동시에 시뮬레이션 할 수 있으므로 5만에서 6만개 정도의 SpiNNaker 칩이 필요하게 된다.

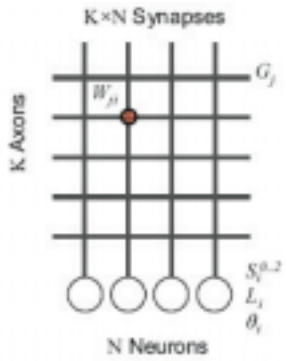
2. Digital Neurosynaptic Core

IBM과 코넬대학은 "A Digital Neurosynaptic Core Using Embedded Crossbar Memory with 45pJ per Spike in 45nm"을 발표하였다. 이 칩은 256개의 뉴론과 1024개의 개별적으로 어드레싱이 가능한 axon, 그리고 SRAM crossbar array로 구현된 1024 × 256 프로그래버블 바이너리 시냅스로 구성된다.

K개의 axon이 N개의 뉴론에 K × N 바이너리 값을 가지고 있는 시냅스를 통하여 연결된다. Gj는 0,1,2 세가지 타입이 있으며 SjGj는 Gj 타입을 가진 axon j로 부터의 입력에 대한 가중치를 의미한다. 그래서 뉴론 i 는 axon j로부터 다음을 입력 받게 된다.

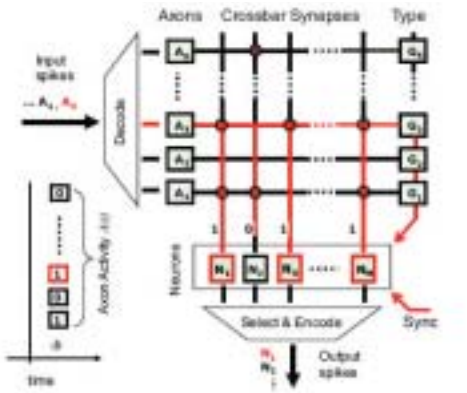
$$A_j(t) \times W_{ji} \times S_i^{G_j}$$

Aj(t)는 j번째 뉴론이 타임 t에 에너지가 충전되어 발화(fire)된 것을 의미한다. 뉴론이 발화한다는 것은 그 뉴런의 입력의 합계가 문턱값을 넘어서 다른 뉴론으로



Name	Description	Range
W_{ji}	Connection between axon j and neuron i	0,1
G_j	Axon type	0,1, 2
$S_i^{0..2}$	Synapse values	-256 to 255
L_i	Leak	-256 to 255
θ_i	Threshold	1 to 256

〈그림 10〉 Neurosynaptic Core [IBM & 코넬대학]



〈그림 11〉 Neurosynaptic Core의 처리 과정

활동 전위가 전송되는 것을 말한다. 뉴론 i의 세포막 전위는 각 타임스템에서 다음과 같이 업데이트된다.

$$V_i(t+1)=V_i(t)+L_i+\sum_{j=1}^{K}\left[A_j(t)\times W_{ji}\times S_i^{G_j}\right]$$

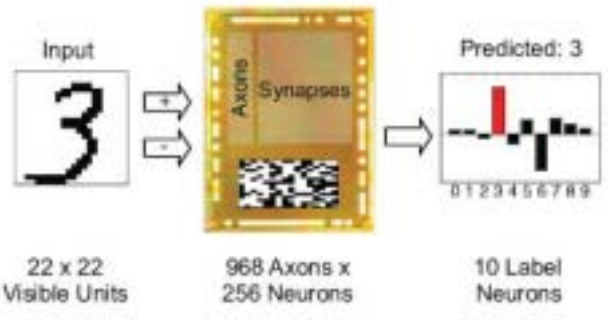
V(t)가 문턱값을 넘게 되면, 뉴론은 스파이크를 발생하고 그 뉴론의 전위값은 0로 리셋된다. 여기서 Li는 누설을 의미한다.

위 그림의 경우, 세 번째 axon이 활성화되고 그 axon에연결된 뉴론 1,2,M의 값이 업데이트된다. 즉, 이러한 신경망 방식에서는 입력된 자료를 토대로 어떤 물체인지를 판단하는 것이 아니라 특정한 영상이 입력되면 학습에 의해서 특정한 결과물로 연결시키도록 가중치를 조절한다.

두가지 단계로 수행된다. 첫 번째 단계에서는 어드레스 이벤트가 한 개씩 코어에 보내진다. 이 이벤트는 적합한 axon block에 복호화된다. 이벤트를 받자 마자, axon은 SRAM 열을 활성화시키고 axon의 모든 연결과 타입을 읽는다. 예를 들면 그림에서 "1"의 값에 해당하는 모든 연결을 통하여 뉴론에 업데이트 된 값이 전달하게 된다. 두 번째 단계는 수 미리세컨드내에 수행되며 동기 이벤트가 모든 뉴론에 전달된다. 이 동기 신호를 받자 마자, 각 뉴론은 세포막 변위가 문턱값을 넘었는지 확인하고 그렇다면 스파이크를 발생시키고 그 변위를 0로 리셋한다. 이 스파이크가 부호화되어 어드레스 이벤트로 순차 형태로 보내진다.

스파이크를 검사한 후에 누설이 적용된다.

코어는 IBM의 45nm SOI 공정을 이용하여 설계되었다. 380만개의 트랜지스터가 4.2 mm^2 면적에 집적되었으며 트랜지스터는 높은 문턱전압을 사용하여 누설을 줄였다. 공급전압이 0.85v일때 코어의 전력소모는 45pJ/spike이다.



〈그림 12〉 Boltzman machine과 neuromorphic architecture를 이용한 숫자 인식 과정

이 그림은 숫자 인식을 할 수있는 뉴럴 알고리즘인 Boltzman machine을 구현한 것이다. 볼쯔만 머신은 지도학습 (supervised learning) 알고리즘으로 신경망과 시뮬레이티드 어닐링으로부터의 흥미로운 성질들을 결합시킨 모델인데 대규모 병렬처리를 이용하는 강력한 계산 장치이다. 볼쯔만 머신은 1984년 Geoffrey E. Hinton 과 Terrence J. Sejnowski 에 의해 도입되었다.이것은 홉필드 모델의 일반화로 여겨질 수 있는데 홉필드 네트워크의 동작 규칙을 확률적인 동작 규칙으로 확장시킨 것으로 생각될 수 있다. 홉필드 네트워크의 동작 규칙에서는 네트워크의 상태를 에너지를 감소시키는 방향으로만 변화시키지만, 볼쯔만 머신에서는 에너지가 증가하는 상태의 전이에 대해서도 작은 확률로나마 허용하는 동작규칙을 사용한다.

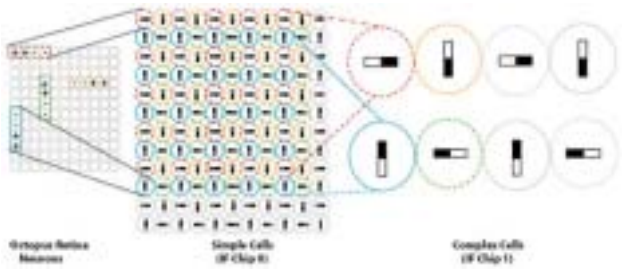
왼쪽의 픽셀은 시각 단위를 표현한다. 256개의 뉴론 중 스파이크가 검은색으로 표시되며 feature로 부호화된다. off-chip 리니어 분류기가 그 feature에 대한 학습 훈련을 통하여 3이 가장 확률이 높은 숫자로 6가 가장 확률이 낮은 숫자로 나타나게 된다.

3. HMAX

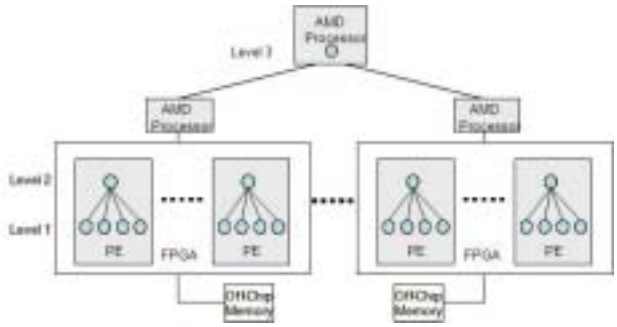
HMAX는 개체 인식을 위하여 생물, 생리 및 행동학적인 알고리즘을 사용한 예이다. 이 모델은 S(simple) 레이어에 있는 Gaussian-like tuning operation과 C(complex) 레이어에 있는 Nonlinear MAX-like operation을 사용한다.

S1 뉴론은 지역 변화를 감지하기 위한 공간 필터에 적용되며 S1 셀은 4 by 1 크기의 망막 수신 영역으로부터 입력을 통합한다. C1 뉴론은 공간 영역에서 유사한 방향성을 갖는 C 셀들의 MAX를 취한다. C1 셀은 5 by 5 크기의 유사한 방향성을 갖는 S1 셀들로 부터의 입력을 통합한다.

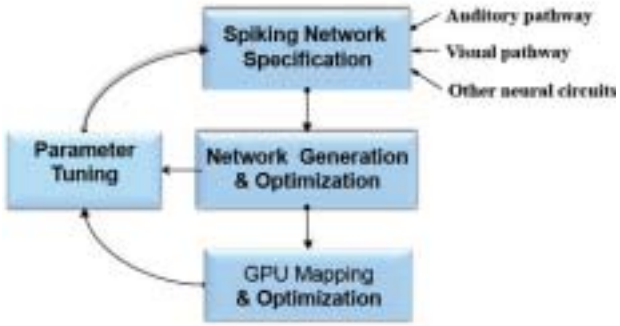
HMAX의 정확한 성능과 견고함에도 불구하고 범용 프로세서로 구현할 때는 제한된 컴퓨터의 자원들과 파워 제약이 따른다. 따라서 임베디드 시스템을 타겟으로 한 알고리즘일 때엔, 느린 실행시간이 알고리즘의 실시간 실현을 위하여 FPGA를 사용한다. 최근에 발표된 28nm node인 FPGA들은 1,955,000 로직 셀들을 포함하며, 최소한의 파워 공간을 유지하면서 전례없는 계산 능력을 제공한다. 사용가능한 FPGA 자원



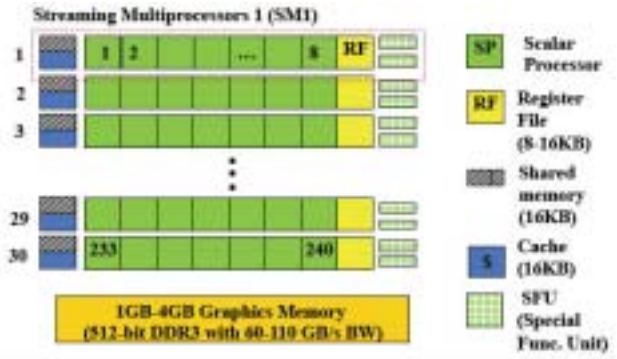
〈그림 13〉 HMAX의 인식과정



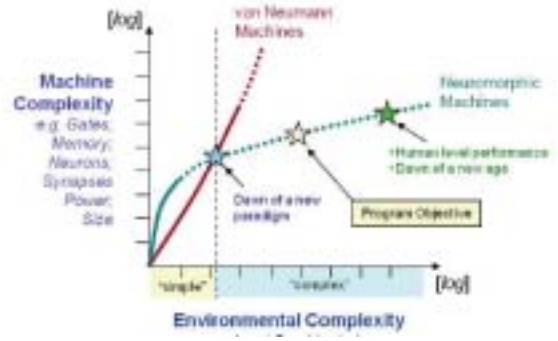
〈그림 14〉 HTM on FPGAs



〈그림 16〉 Framework for Spiking Neural Network



〈그림 15〉 NVIDIA GPU Hardware



〈그림 17〉 Neuromorphic machine의 새로운 파라다임과 시대

증가로 neuromorphic 비전 알고리즘을 가속하기 위해 FPGA들의 사용이 증가하고 있다.

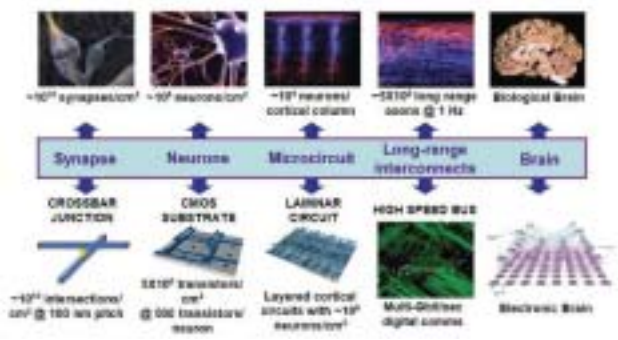
4. 기타 구조

응용특화된 구조인 SpinNNAker[3,4] 이외의 하드웨어 플랫폼에는 슈퍼컴퓨터(예, HTM - Cray XD1), 재구성 하드웨어, neuromorphic 구조 (예, 스탠포드의 Neurogrid), 고성능 GPU (예: NVIDIA GPUS) [6]등이 있다.

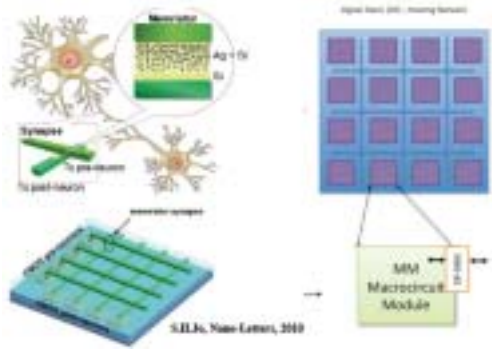
HTM은 Cray XD1에 기반하여 구현되었다.

Neurogrid는 스탠포드 대학의 Kwabena Boahen에 의해서 개발된 멀티 칩 시스템이며 4 by 4 neurocore의 어레이로 구성되며, 각 코어는 256 by 256 뉴론 어레이로 구성되며 6000개의 니냅스 연결이 가능하다. NVIDIA의 GPU를 이용하면, 수천개의 뉴론을 병렬처리가 가능하다. 또한 메모리 밴드폭이 보통 프로세서 보다 5배 이상 넓다. EU FACET 프로젝트는 2005년에 개발되었으며 200,000 뉴론이 5천만개의 시냅스에 연결된 구조이다. 빠른 transient states를 사용하는 Analog Computing을 사용한다. Waferscale integration을 이용하여 대규모 analog spiking neural network hardware를 개발하며, 통신 병목 현상을 줄이기 위해서 Temporal and spatial multiplexing을 사용한다. 목표는 fault tolerance (all levels), low power consumption, 20cm wafer with ~40 Million synapses이다. Torres-Huitzil(2005년)은 128 by 128 이미지에 대해서 pentium 4에 비해 100배의 속도 향상을 보였다.

인간의 두뇌는 1000억개의 뉴론과 1015 개의 synaptic 연결로 되어 있다. 쥐는 1백만개의 뉴론과 1010 개의 synaptic 연결로 되어 있다. DARPA는 SyNAPSES 프로젝트 HRL Team에서는 단기 목표로 쥐 모델을 사용한다. UCI에서는 십만개의 뉴론과 108



〈그림 18〉 DARPA SyNAPSES – HRL team



〈그림 19〉 Memristors 기반의 neuromorphic chips

개의 synapses을 초기 목표로 잡아 연구하고 있다.

그림의 GPU 하드웨어는 GTX 280 Card이며, 240 X 1 GHz scalar processors (CUDA 1.3 device)로 구성되어 100K에서 225K의 뉴론을 검증할 수 있으며, 뉴론당 Synapses 수는 100에서 500개이다.

NeMo는 CUDA GPU를 이용하여 수십만개의 spiking 뉴론 (Izhikevich neurons)의 네트워크를 검증하는 고성능 실시간 시뮬레이터이다.

Ⅳ. 결론

뇌의 리버스 엔지니어링은 많은 CAD 연구를 필요로 한다. 즉, 모델링, sw/hw 설계, 시뮬레이션, 알고리즘 하드웨어 매핑, 재사용 라이브러리, 구조 합성, 최적화 알고리즘 (예: 파라메터 튜닝), 성능 분석, HW platform이 필요하게 된다. 이러한 도구의 개발을 위해서는 도메인 지식이 필요하며, 이를 위하여 인지과학, 신경과학, 전산학의 인터렉션 융합 연구가 필요하다.

감사의 글: 2012년 2월 한국정보산업연합회 주관 IT 교수 역량 지원 사업의 일환으로 진행된 Professor Nikill Dutt's Advanced System on Chip Design Methodologies의 강의에 참가하여 Neuromorphic Architecture의 간략한 최신 동향을 듣게 되었습니다. 감사드립니다.

참고문헌

[1] Dharmendra S. Modha and Raghavendra Singh (IBM), "Network architecture of the long-distance pathways in the Macaque brain", Proceedings of the National Academy of Sciences

[2] S. Furber and A. Brown. Biologically-inspired massively-parallel architectures - computing beyond a million processors. In Proc. 9th International Conference on the Application of Concurrency to System Design, pages 3~12. ACSD'09, 2009.

[3] X. Jin, M. Lujan, L. A. Plana, S. Davies, S. Temple, and S. B. Furber. Modeling spiking neural networks on spinnaker. IEEE Computing in Science and Engineering, September/October 2010 (vol. 12 no. 5), pages 91-97, 2010.

[4] X. Jin, M. Lujan, L. A. Plana, A. D. Rast, S. R. Welbourne, and S. B. Furber. Efficient parallel implementation of multilayer backpropagation networks on spinnaker. In CF '10: Proceedings of the 7th ACM international conference on Computing frontiers, pages 89-90, New York, NY, USA, 2010.

[5] X. Jin, A. Rast, F. Galluppi, S. Davies, and S. Furber. Implementing spike-timing-dependent plasticity on spinnaker neuromorphic hardware. WCCI 2010 IEEE World Congress on Computational Intelligence, pages 2302 - 2309, July 2010.

[6] J. M. Nageswaran, N. Dutt, J. L. Krichmar, A. Nicolau, and A. Veidenbaum. Efficient simulation of large-scale spiking neural networks using cuda graphics processors. In IJCNN'09: Proceedings of the 2009 international joint conference on Neural Networks, pages 3201~3208, Piscataway, NJ, USA, 2009.

[7] R. Rubenstein. Linking arms to make a brain. New Electronics, July 2010, pages 16~18, 2010.

[8] Professor Nikil Dutt's Lecture Notes, Feb. 2012.

[9] http://www.ine-web.org/

[10] Steve Furber, To Build a Brain, pp.39 - 43, IEEE Spectrum, 8. 2012

조 준 동

1980년 3월 성균관대 전자공학과 학사
1989년 8월 (미)폴리테크닉 대학 전산학 석사
1993년 6월 (미)노스웨스턴 대학 전산학 박사
1983년 7월~1995년 2월 삼성전자(주) 반도체연구소 연구원
1995년 3월~현재 성균관대학교 전자공학과 교수
2000년 8월~2001년 7월 IBM T.J.Watson Research Center 연구원

〈관심분야〉
Mobile SoC Applications