目的特化型AIチップの開発、地域で必要とするAIを大学で創る ニューロモーフィック AI チップ

Development of Application-Specific AI Chips: Creating Regionally Needed AI at the University

Neuromorphic AI Chips

BEN ABDALLAH Abderazek

Email: benab@u-aizu.ac.jp

会津大学
THE UNIVERSITY OF AIZU

コンピュータ理工学科 コンピュータ工学部門
Department of Computer Science and Engineering, CE Division
Neuroengineering Laboratory

https://www.u-aizu.ac.jp/misc/neuro-eng/

Neuromorphic AI Computing
ニューロモーフィック AI コンピューティング

- Neuromorphic AI Hardware
  ニューロモーフィック AI ハードウェア

- Neuromorphic AI Hardware Research and Application Examples at our Laboratory
  当研究室におけるニューロモーフィックAIハードウェアの研究と応用例

Neuromorphic AI Computing
ニューロモーフィック AI コンピューティング

- Neuromorphic AI Hardware
  ニューロモーフィック AI ハードウェア

- Neuromorphic AI Hardware Research and Application Examples at our Laboratory
  当研究室におけるニューロモーフィックAIハードウェアの研究と応用例

Hardware Trends
ハードウェアのトレンド

Why do we need accelerators?
なぜアクセラレータが必要なのでしょうか

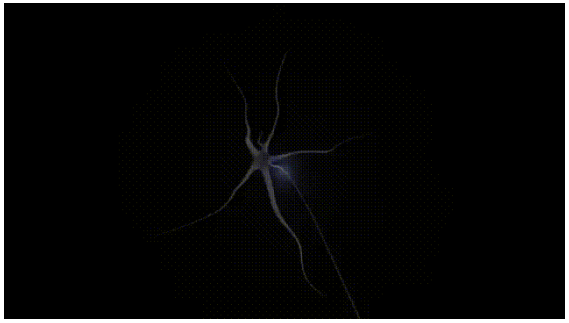
- Challenge of multi-thread performance with general-purpose cores/ 汎用コアによるマルチスレッドパフォーマンスへの挑戦
  - Dark silicon/ダークシリコン→ emphasis on power-efficient throughput
  - End of device and circuit scaling
    デバイスと回路のスケーリングの終わり

- Emergence of machine learning
  機械学習の出現

Commercial ML Hardware
商用 ML ハードウェア

Google TPU (inference and training)
Recent NVIDIA chips (Volta)
Microsoft Brainwave and Catapult
Intel Loihi and Nervana
Cambricon
Graphcore (training)
Cerebras (training)
Groq (inference)
ARM
Tesla (FSD)
Western Digital

Neuromorphic AI Hardware
ニューロモーフィック AI ハードウェア



BENAB-ISTC2023

- Feature: Emulating the brain
  特徴: 脳をエミュレートする
  - Low power – the brain consumes only 20 W
    低電力 – 脳の消費電力はわずか 20 W
  - Fault-tolerant – the brain loses neurons all the time
    フォールトトレラント – 脳は常にニューロンを失います
  - No programming required – the brain learns by itself
    プログラミングは必要ありません – 脳は自ら学習します
  - Potential for high-information content
    情報量の多いコンテンツの可能性

- Many significant efforts/多くの重要な取り組み: HBP (Human Brain Project), DARPA SyNAPSE, etc.

- Implementations/実装: UK SpiNNaker, IBM TrueNorth, Intel Loihi

Sparsity in Neural Networks
ニューラル ネットワークのスパース性

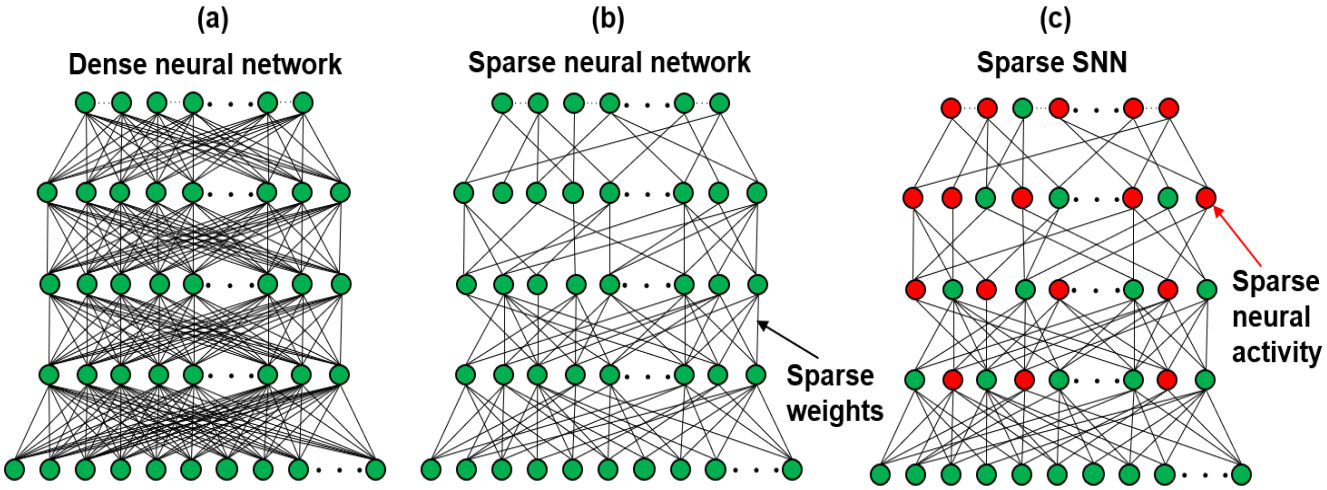
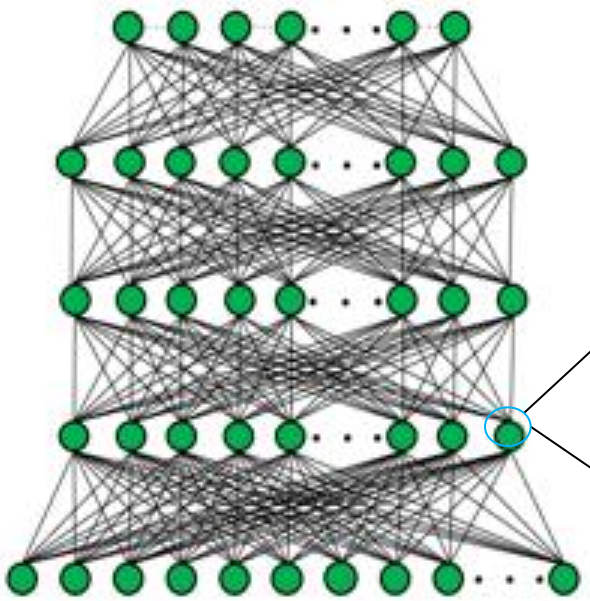


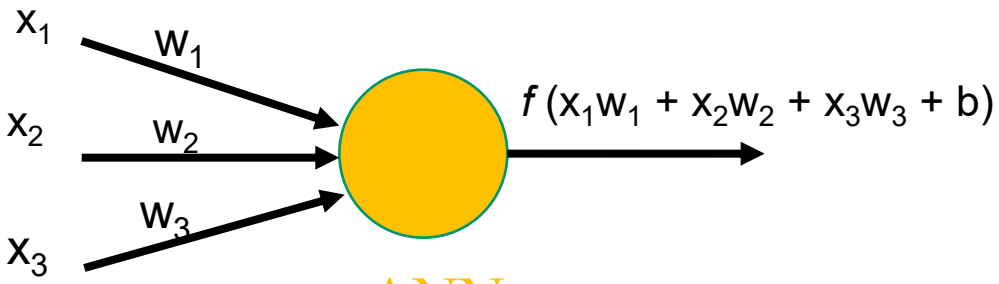
Fig. 1 Illustration of sparsity in neural network

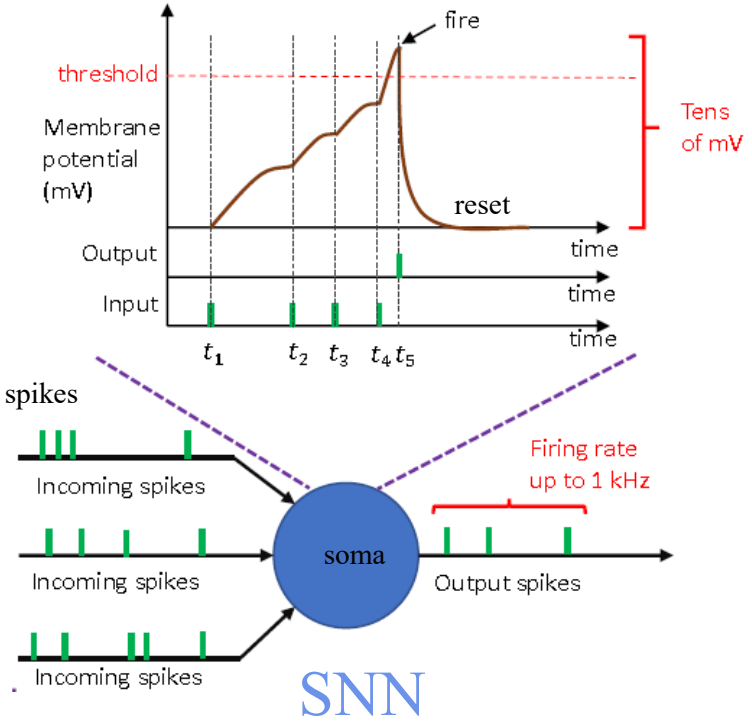
- Only 0.5% to 2% of neurons in the neocortex are active at any time [Lennie 2003]
  新皮質のニューロンのうち常に活動しているのは 0.5% ～ 2% だけです

- Only 1% to 5% of connections exist between two connected layers in the neocortex and 30% of those connections change every few days [Holmgren 2003]
  大脳新皮質の 2 つの接続された層の間には接続の 1% ～ 5% しか存在せず、それらの接続の 30% は数日ごとに変化します

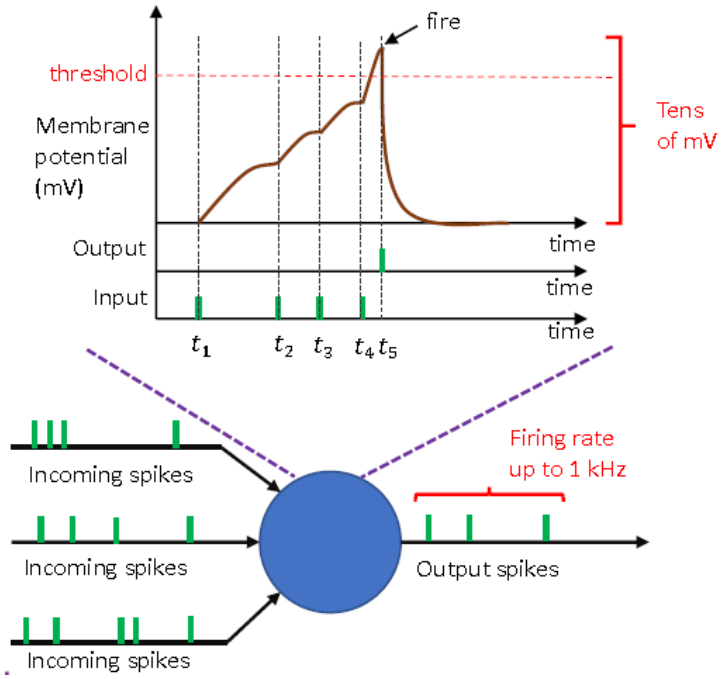
ANN vs. SNN



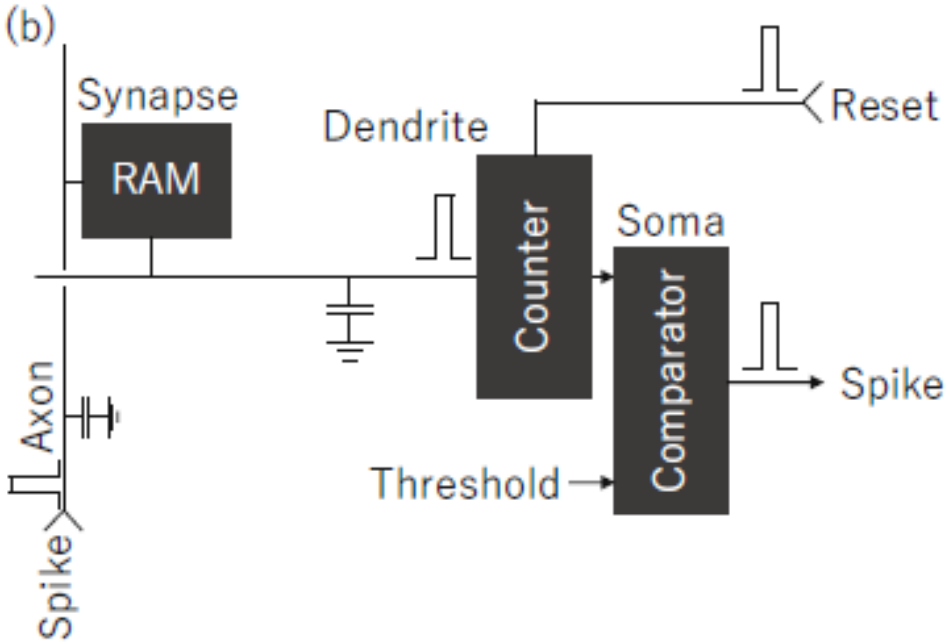

x1
x2
x3
w1
w2
w3
f (x1w1 + x2w2 + x3w3 + b)
ANN


fire
threshold
Membrane potential (mV)
reset
Tens of mV
Output
Input
time
time
time
t1
t2
t3
t4
t5
spikes
Incoming spikes
Incoming spikes
Incoming spikes
soma
Firing rate up to 1 kHz
Output spikes
SNN

Spiking Neuron
スパイキングニューロン



A simple spiking neuron model

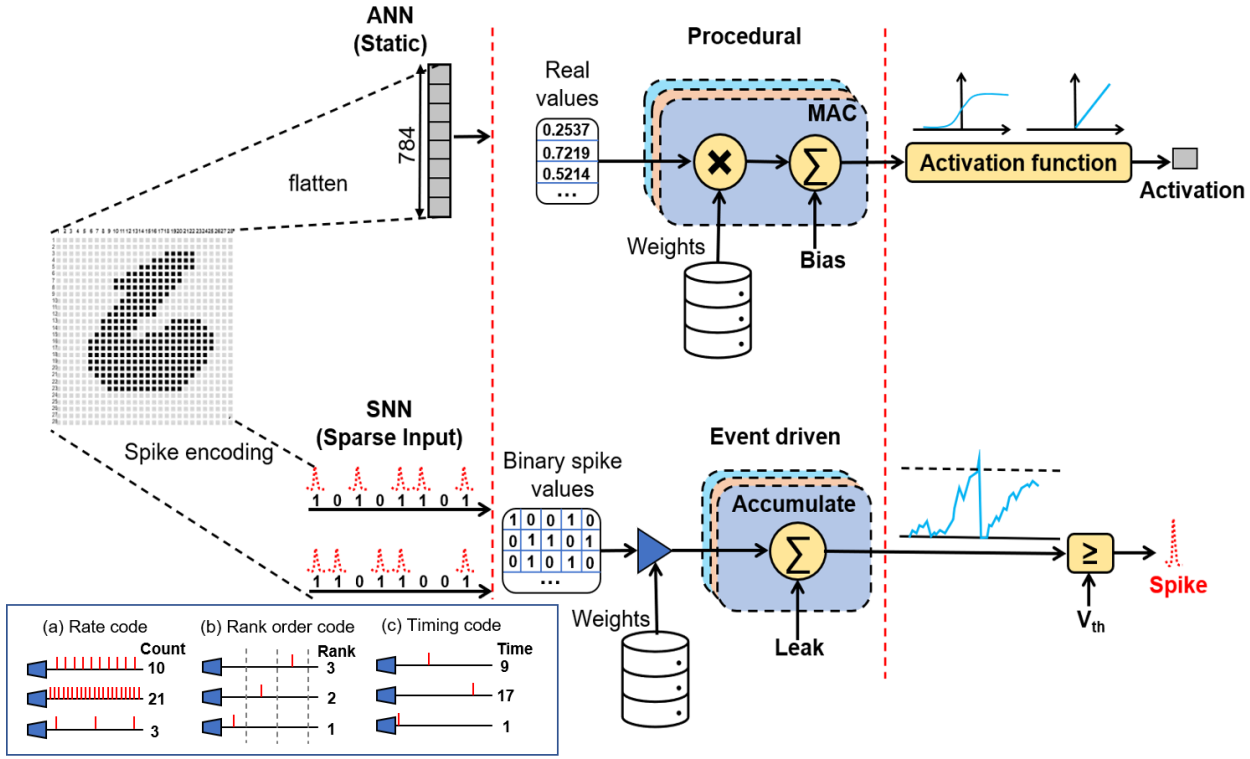


Digital silicon neuron implementation

- Digital implementation: a counter is incremented (dendrite) each time a 1 is read out of a bit cell (synapse), triggered by the incoming spike (axon).
- The counter's output is compared (soma) with a digitally stored threshold and a spike is triggered when it is supra threshold. The counter is then reset and the cycle starts over.

ANN vs ANN – Example/例

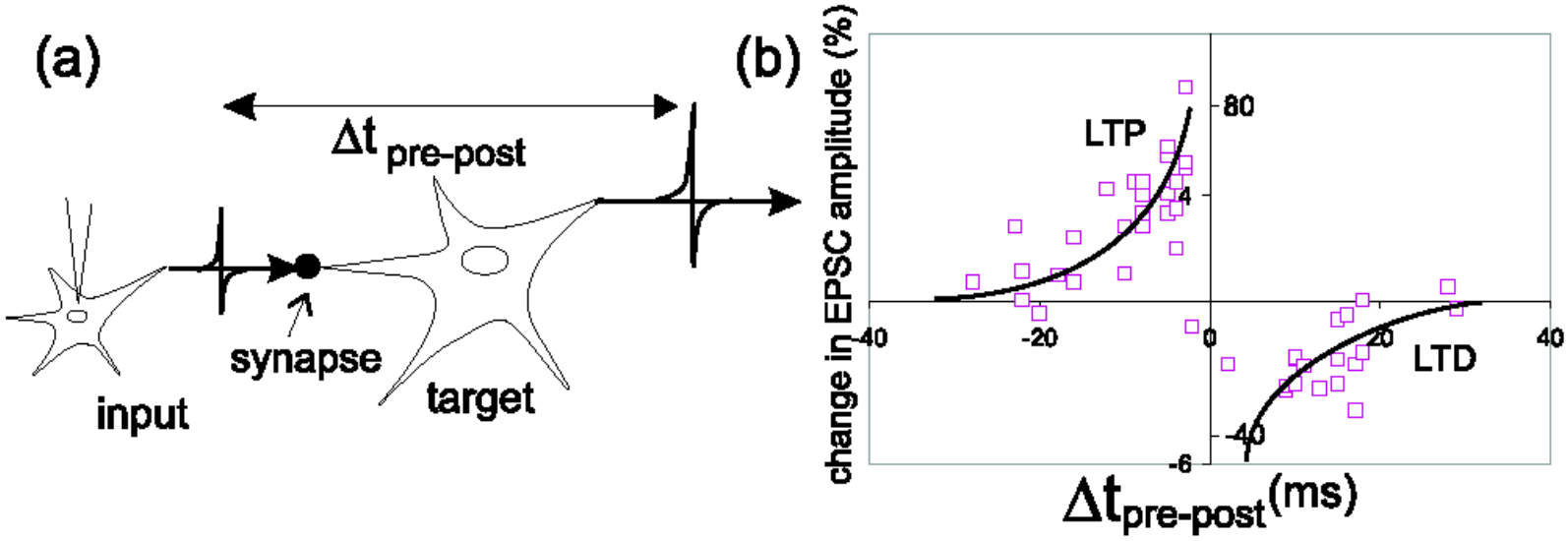
Illustration of conventional ANN and SNN operations



- Sparse input in SNN means sparse memory use.
  SNN におけるスパース入力は、スパースなメモリ使用を意味します。
- Spike communication means minimal power per event signal
  スパイク通信により、イベント信号あたりの電力が最小限に抑えられます
- Event-based processing in SNN also contributes to low power.
  SNN のイベントベースの処理も低電力化に貢献します。

Neuron Learning/ニューロン学習

- Learning rules based on STDP specify changes in synaptic strength depending on the time interval between each pair of presynaptic and postsynaptic events.

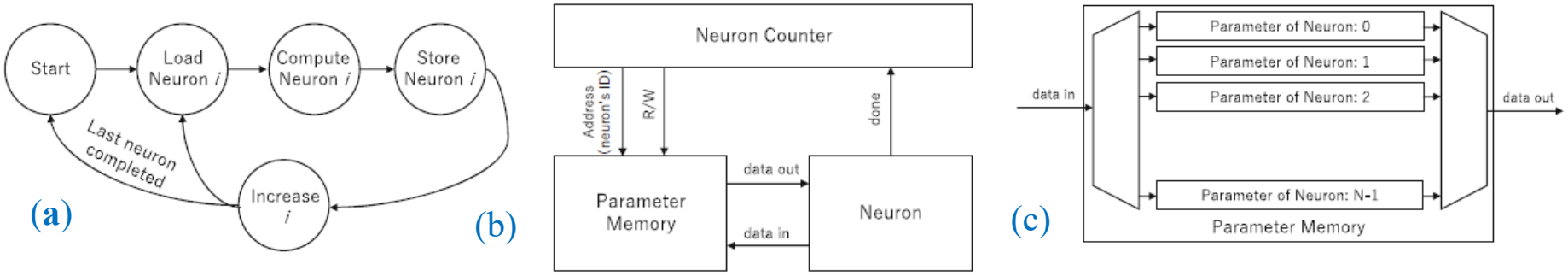


Spike-timing-dependent plasticity (STDP)

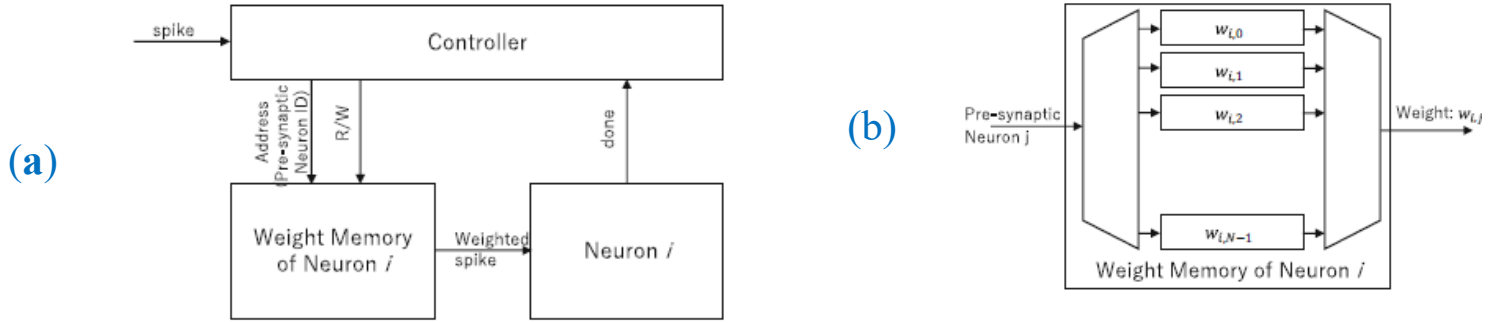
- If the presynaptic neuron fire **before** the postsynaptic neuron within a preceding 20ms, LTP occurs

- If the presynaptic neuron fire **after** the postsynaptic neuron within the following 20ms, LTD occurs

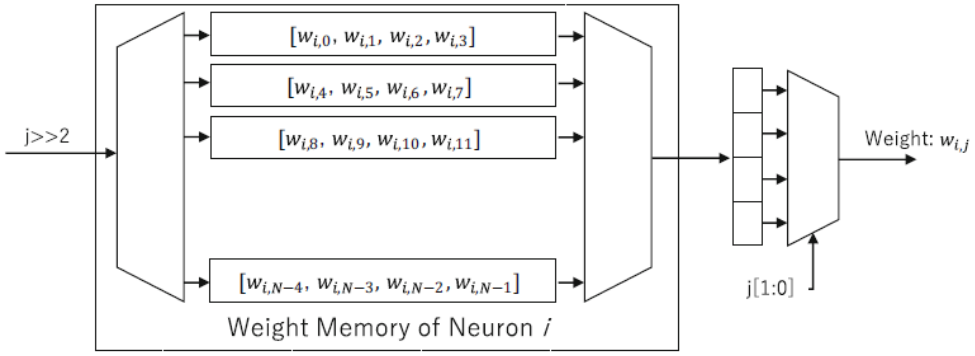
Synapse Memory Design/シナプスメモリの設計



The *serial* neuron model. (a) The model architecture. (**b**) The finite state machine. (**c**) The parameter structure

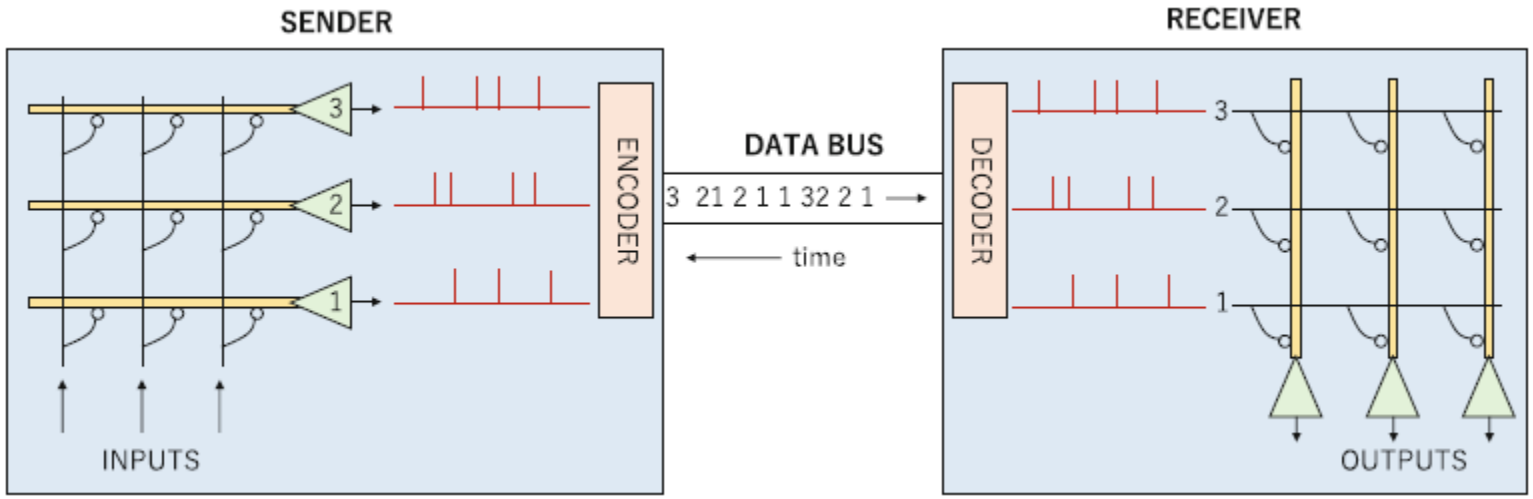


The *parallel* neuron weight model. (**a**) The model architecture. (**b**) The weight structure

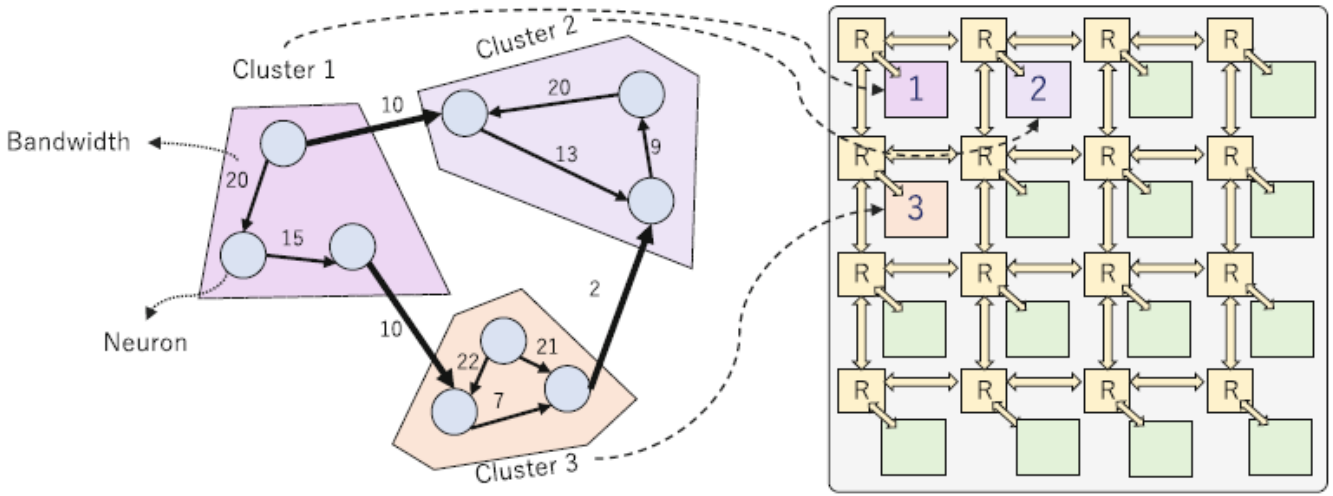


The *parallel* neuron weight memory with merged four weights in a memory row

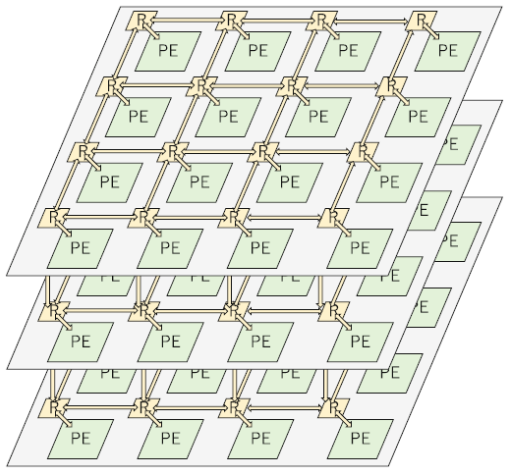
HW Mapping of SNN/SNNのハードウェアマッピング



Address-event representation (AER) protocol



2D-NoC



3D-NoC

Hardware Mapping of Spiking Neural Network on Network-on-Chip (2D/3D)

Neuromorphic AI Computing
ニューロモーフィック AI コンピューティング

- Neuromorphic AI Hardware
  ニューロモーフィック AI ハードウェア

- Neuromorphic AI Hardware Research and Application Examples at our Laboratory
  当研究室におけるニューロモーフィックAIハードウェアの研究と応用例

A1

Low-power Neuromorphic SoC
低消費電力ニューロモーフィック SoC

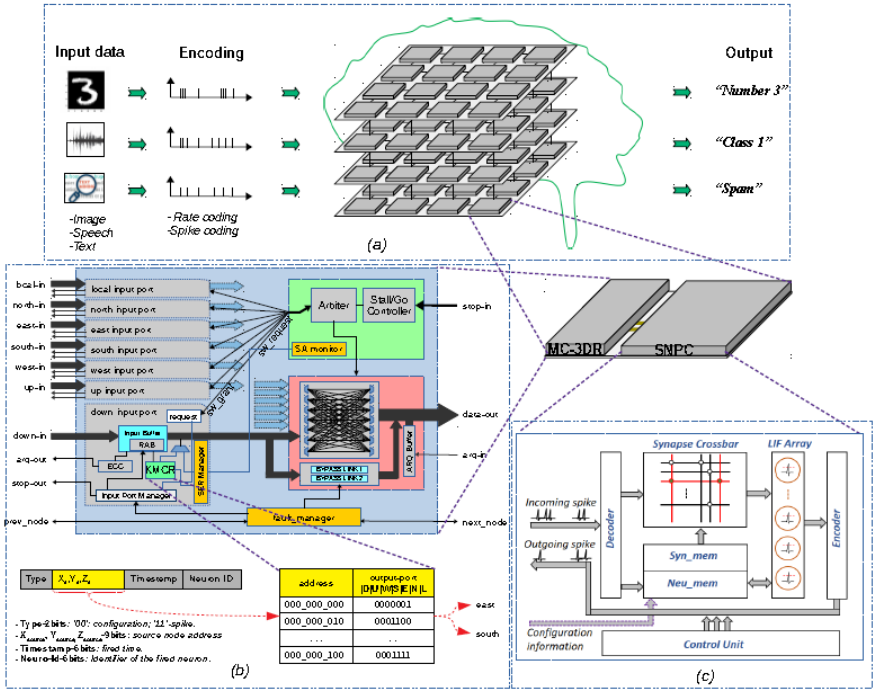
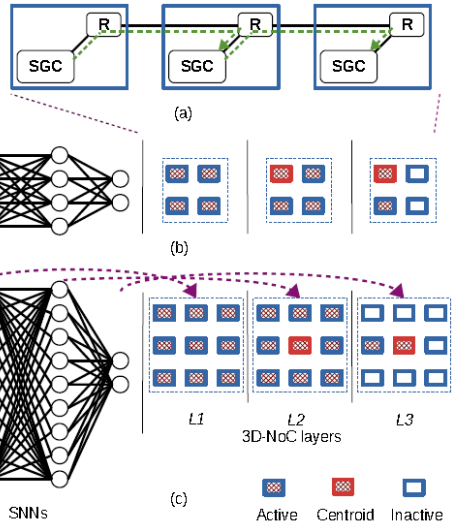
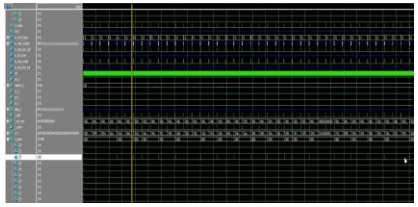
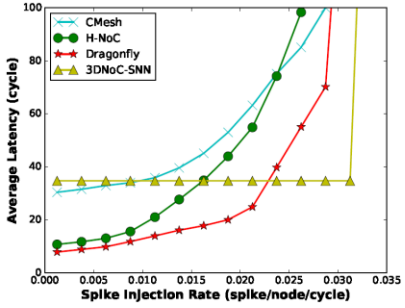


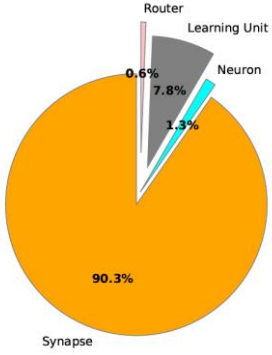
Fig. 5: System architecture: (a) 3DNoC-SNN organization, (b) Multicast router architecture (MC-3DR), (c) Spiking neuron processing core (SNPC).

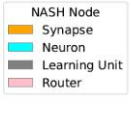


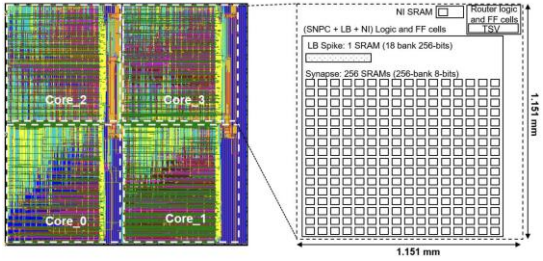




Evaluation Result, Average latency evaluation, and comparison over various SIRs.







Area analysis of a NASH node

Parameter/System	XY-UB	XYZ-UB	SP-KMCR		FTSP-KMCR	
Architecture	Baseline	NASH	Baseline	NASH	Baseline	NASH
Area (mm^2)	1.312	1.316	1.322	1.322	1.320	1.325
Power (mW)	66.16	66.63	66.50	66.84	68.22	70.10

Design complexity comparison of NASH and Baseline nodes

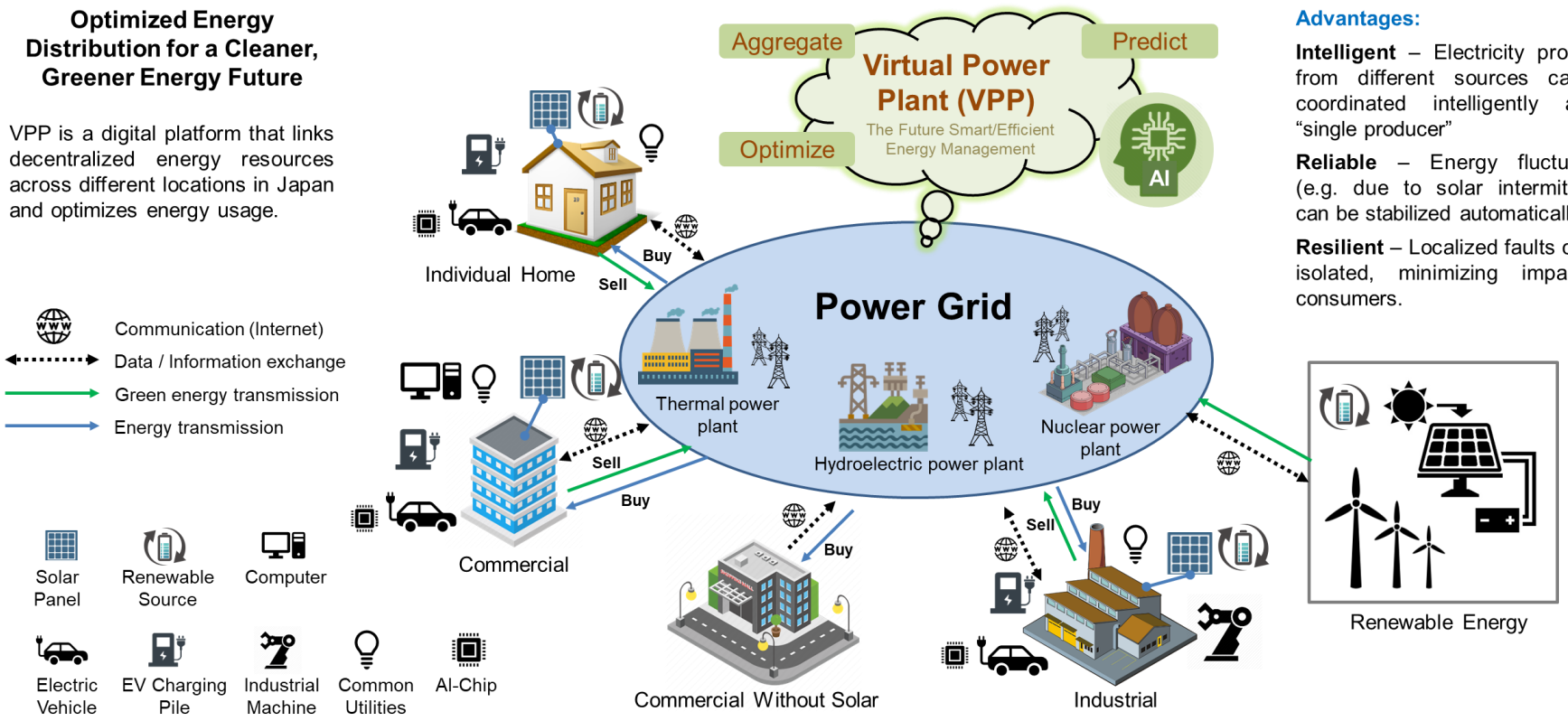
O. M. Ikechukwu, K. N. Dang and A. Ben Abdallah, "On the Design of a Fault-Tolerant Scalable Three Dimensional NoC-Based Digital Neuromorphic System With On-Chip Learning," IEEE Access, vol. 9, pp. 64331-64345, 2021, doi: 10.1109/ACCESS.2021.3071089

A2

Off-Grid Energy Storage Solar Carport
オフグリッドエネルギー貯蔵ソーラーカーポート

Optimized Energy Distribution for a Cleaner, Greener Energy Future

VPP is a digital platform that links decentralized energy resources across different locations in Japan and optimizes energy usage.



Advantages:

Intelligent – Electricity produced from different sources can be coordinated intelligently as a "single producer"

Reliable – Energy fluctuations (e.g. due to solar intermittency) can be stabilized automatically.

Resilient – Localized faults can be isolated, minimizing impact to consumers.

Z. Wang, M. Ogbodo, H. Huang, C. Qiu, M. Hisada, A. Ben Abdallah, "AEBIS: AI-Enabled Blockchain-based Electric Vehicle Integration System for Power Management in Smart Grid Platform," IEEE Access, vol. 8, pp. 226409-226421, 2020, doi:10.1109/ACCESS.2020.3044612.

A2

Off-Grid Energy Storage Solar Carport
オフグリッドエネルギー貯蔵ソーラーカーポート

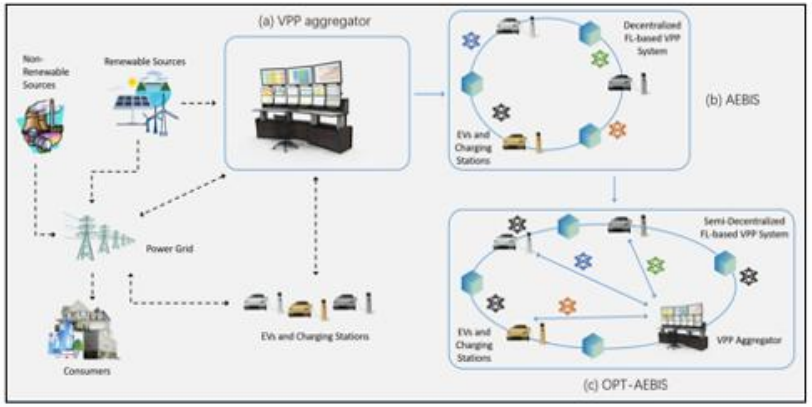


Fig. 1. Virtual Power Plant (VPP): (a) conventional VPP aggregator, (b) AEBIS, (c) optimized AEBIS (O-AEBIS).

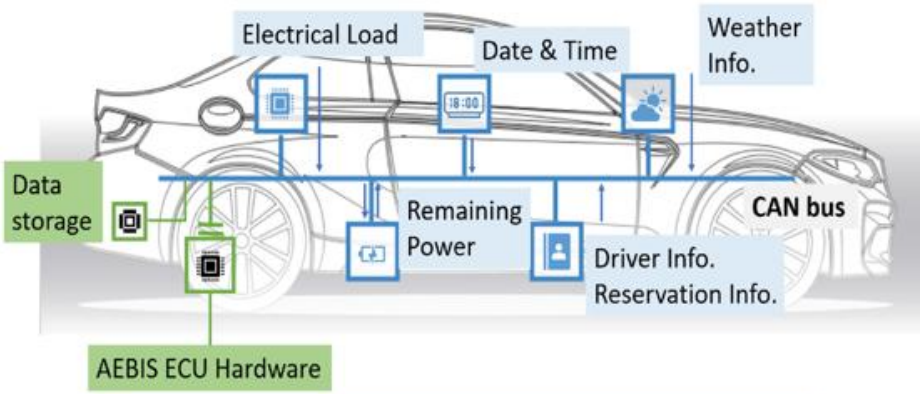
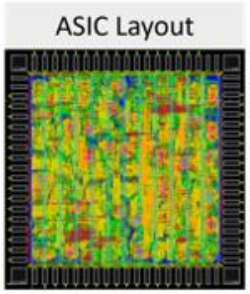


Fig. 2. Neural Network for Power Consumption Prediction of Electric Vehicle (EV).



Area	0.265 mm²
Voltage	1.1 V
Power	318.224 mw
Temperature	25°C

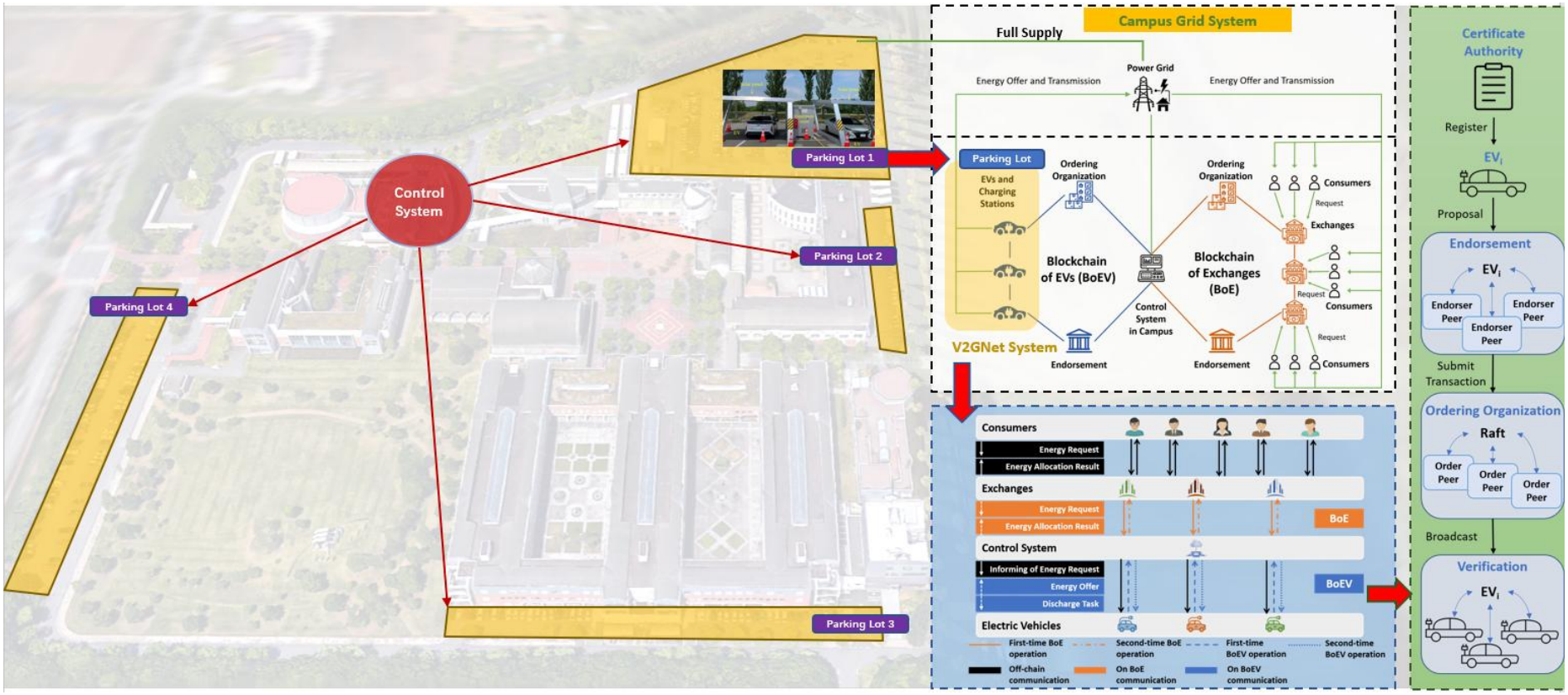
Fig. 5. A demonstration of the energy management system based on our system named AEBIS and its optimized version O-AEBIS.

Name	BRAM_18K	DSP48E	FF	LUT
Expression	-	-	0	493
Instance	-	5	414	950
Memory	2	-	320	20
Multiplexer	-	-	-	627
Register	-	-	454	-
Total	2	5	1188	2090
Available	120	80	35200	17600
Utilization (%)	1	6	3	11
Weights	**Memory required**			
Weights	568 Bytes			
Biases	60 Bytes			
Inputs	44 Bytes			
Total	672 Bytes			

Fig. 6. Hardware complexity of power consumption prediction system on the Zynq-7010 FPGA. The system utilized 3% of the FF, 11% of the LUT, 6% of the DSP48, and approximately 1% 18k BRAM.

A3

Off-Grid Energy Storage Solar Carport
オフグリッドエネルギー貯蔵ソーラーカーポート

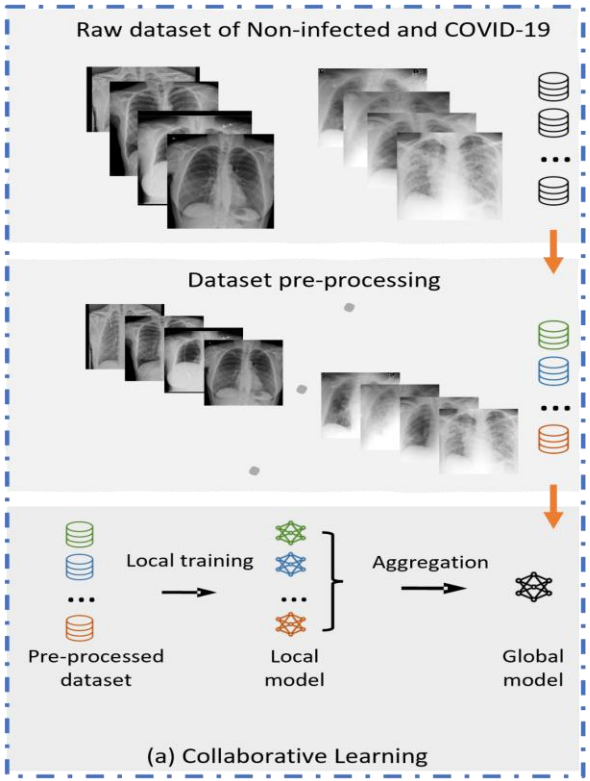


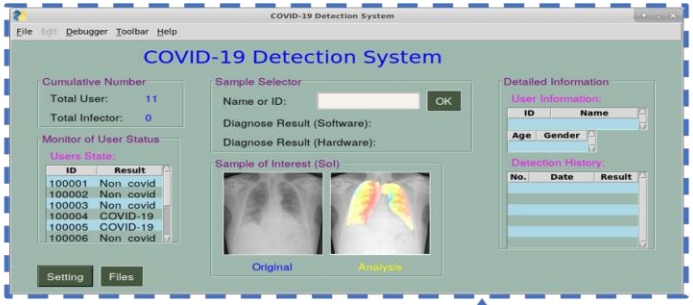
[特許第6804072号] (2020.12.04) ベンアブダラ　ア ブデラゼク (Abderazek Ben Abdallah), 久田雅之, "Virtual Power Platform Control System [仮想発電所制御システム]", 特願2020-033678号 (2020.02.28)

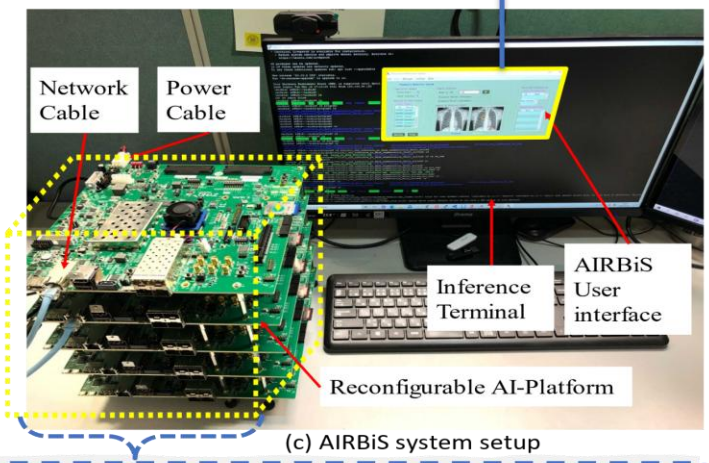
ネルギーの取引方法とシステム/Energy Trading Method and System
Abderazek Ben Abdallah, Wang Zhishang, Masayuki Hisada, ‘’An electricity trading system and an electricity trading method [電力取引システム及び電力取引方法に関する], 特願2022-022472 Power Consumption Prediction Method and System for Power Management in Smart Grid/ スマートグリッドにおける電力管理のためのEV消 費電力予測 方法とシステム

Abderazek Ben Abdallah,Wang Zhishang, Khanh N. Dang, Masayuki Hisada, "EV Power Consumption Prediction Method and System for Power Management in Smart Grid [スマートグリッドにおける電力管理のためのEV消 費電力予測 方法とシステム]", 特願2022-022472

A4 AI-Powered Hardware-Software Platform for Pneumonia Detection/ 肺炎検出のための AI を活用したハードウェアとソフトウェアのプラットフォーム







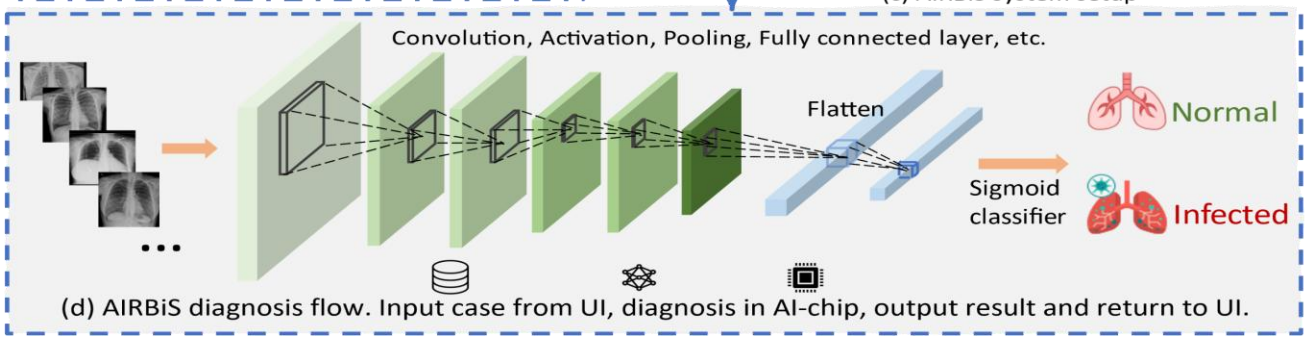


Table 7.3. FPGA Resource Utilization Estimates.

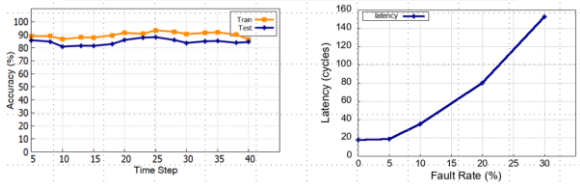
Resource	Utilization		Available	Utilization (%)	
	ANN	SNN		ANN	SNN
LUT	54,585	27,288	274,080	19.9	9.9
LUTRAM	3668	2048	144,000	2.5	1.28
FF	53,035	37,098	548,160	9.7	6.77
BRAM	824	0	912	90.4	0
DSP	35	0	2520	1.4	0
BUFG	4	18	404	1.0	4.45
MMCM	1	0	4	25	0

Table 7.4. Hardware Complexity.

Core/Parameter	Area (mm^2)		Power (mW)	
	SNN	ANN	SNN	ANN
Convolution core	0.0748	0.0755	0.007	0.011

Table 7.2. Dataset description.

Label	Class	Train	Test
COVID	COVID	2870	700
	COVID(Augmented)	14,349	-
Non-COVID	Normal	9791	400
	Lung_Opacity	5762	250
	Viral_Pneumonia	1288	50
Sum		34,060	1400

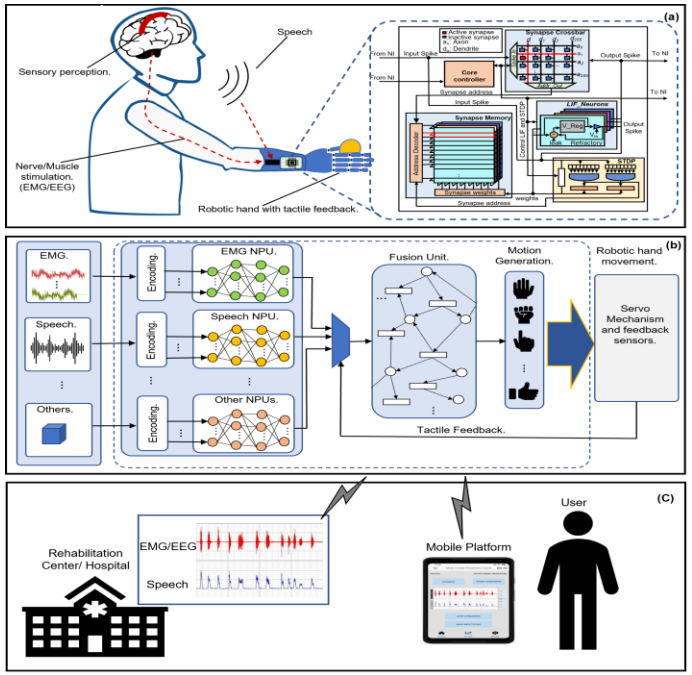


(a) Detection accuracy over various time-steps (b) Detection Latency over various fault-rate

Figure 7.6. Accuracy and fault-rate evaluation result

A5

Neuromorphic Robot Arm and Prostheses
ニューロモーフィック ロボット アームと義足



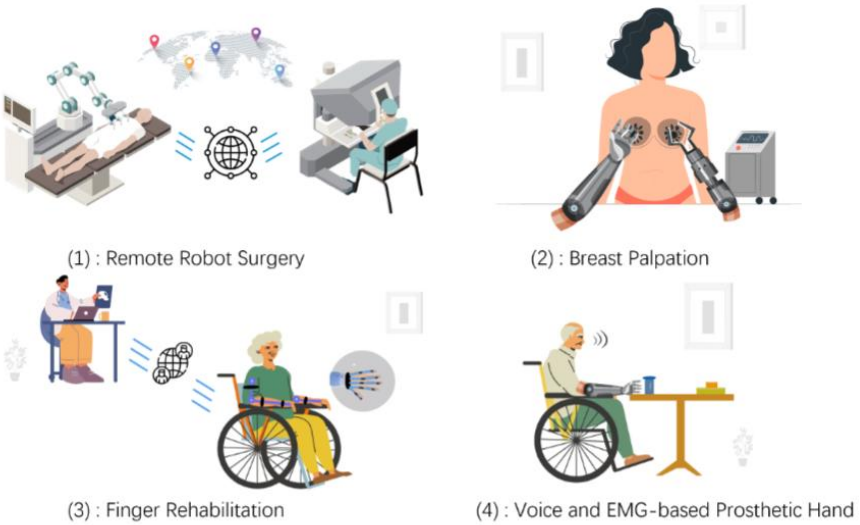


FIGURE 4. The Neuromorphic AIzuHand target applications. (1) Remote Robot Surgery, (2) Breast Palpation, (3) Finger Rehabilitation, (4) Voice and EMG-based Prosthetic Hand.

Abderazek Ben Abdallah, Huankun Huang, Nam Khanh Dang, Jiangning Song, "AIプ ロセッサ [AI Processor]," *特願2020-194733 (2020 年11月24日)*

AIzuHand I, July 2022

- Device Name: AIzuHand I
- Total Weight: 422g (276g without controller)
- Control: sEMG
- DoF: 5
- Feedback: No
- Related patent: 特 願2019-124541
- Contact "benab(at)u-aizu.ac.jp"

https://www.u-aizu.ac.jp/misc/neuro-eng/aizuhand.html

Conclusions/結論

- **Neuromorphic AI-chip/ニューロモーフィック AI チップ**
  - Low power, Fault-tolerant, high-information content

    低電力、フォールトトレラント、高情報コンテンツ
- **Memory access in AI-Chip is the bottleneck**
  AIチップのメモリアクセスがボトルネック
  - Possible HW/SW techniques to cope with the memory access problem/ メモリアクセスの問題に対処するために考えられる HW/SW テクニック:
    - Advanced Storage Technology/高度なストレージテクノロジー
      - ✓ Embedded DRAM (eDRAM) → Increase on-chip storage capacity
      - ✓ 3D Stacked DRAM → Increase memory bandwidth
      - ✓ Use memristors as programmable weights (resistance)
    - Reduce the size of operands for storage/compute/ストレージ/コンピューティングのオペランドのサイズを削減する
      - ✓ Floating point → Fixed point
      - ✓ Bit-width reduction
    - Reduce the number of operations for storage/compute/ ストレージ/コンピューティングの操作数を削減する
      - ✓ Network Pruning; Compact Network Architectures