Spiking Neural Network(SNN)技術の現状と課題に関する考察

アトム回路は ニューロンか？

岡島　義憲

自己紹介　（岡島義憲）

■ **来歴**

- 1956年： 札幌市生まれ
- 1981年： 京都大学理学部物理学科卒
- 1981年～2019年　　富士通(株)/富士通セミコンダクター(株)で半導体設計
  - > メモリ、マイコン、ASIC、ASSP
  - > 業界戦略（半導体業界団体 in 2015年～2019年)
- 2020年： 曖昧検索回路のネットワークについて考えたい。

目次

1. Neuromorphic と Spiking Neural Network(SNN)
   - SNNの基本構成
     - > 発火モデル、シグナル・モデル、シグナル・コーディング
     - > Winner 生成回路、Local Network
   - Neuromorphic Processors

2. SNN と ANN のベンチマーク議論

3. 考察
   - Atom Circuit と Atom of Information

   （脳の動作の表現戦略における「粒度」について）

Neuromorphic & Spiking Neural Network(SNN)

Neuromorphic / SNN に関する論文動向

(参) Warren McCulloch ; "Embodiments of Mind (1965)"

- 形式ニューロン（ for ANN）⇒ Parallel Distributed Processing (PDP)
- Hard-wired nets that process images or sounds ⇒ 脳科学のテーマ

Carver Mead ; "Neuromorphic Electronic Systems (1990)",

Wolfgang Maass ;

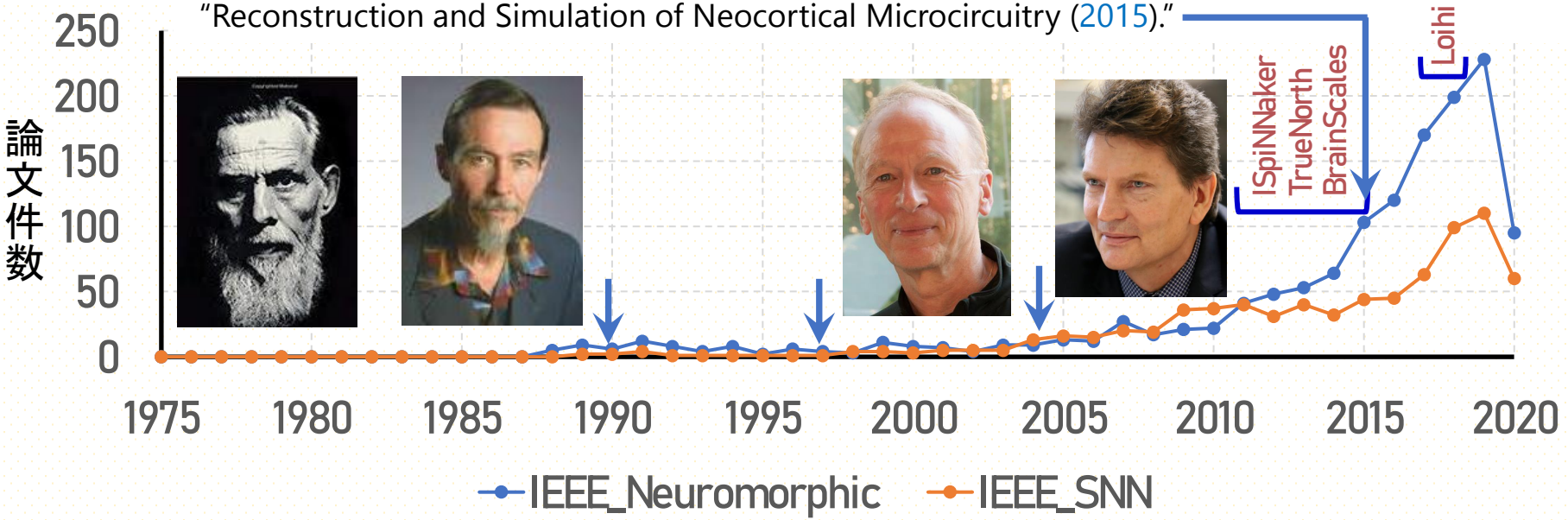
"Networks of Spiking neurons: the third generation of neural network models (1997)"

"On the computational power of circuits of spiking neurons (2004)"

- 第１世代： McCulloch-Pitts Perceptrons
- 第２世代： Deep Neural Network
- 第３世代： Spiking Neural Network

Henry Markram ; Blue Brain project / Human Brain Project / Brain Mind Institute Project

"Reconstruction and Simulation of Neocortical Microcircuitry (2015)."



Recent Hardware Developments for Neuromorphic Computing

Work	Technology	Application	Input data	Neuron model	I&F	学習	Network	SW言語
NeuroGrid_2014 Stanford Univ.	ASIC_180	Neuro Sciences	Spikes	Dimensionless model	Analog	（不明）	Programmable	NGPyton
TrueNorth_2014 IBM	ASIC_28	Classification	Frame-based	LIF	Digital	不可	Conv/FC/RNN	Corelets
SpiNNakerCMP_2014 Manchester Univ.	ASIC_130	Neuro Sciences	Spikes	LIF, IZH, HH	Digital	**Program**	Programmable	PyNN
BrainScales_2017 Heidelberg Univ.	ASIC_180 (Wafer Scale)	Neuro Sciences Classification	Frame-based	exp IF	Analog	**STDP**	Full Connection	PyNN
Loihi_2018 Intel	ASIC_14	Classification	Spikes	CUBA LIF	Digital	**STDP**	Conv/FC/RNN	Loihi API
ODIN_2018 Univ. of Louvain	ASIC_28	Classification	Spikes	Izhikevich	Analog	**SDSP**	Programmable	（不明）
Minitaur_2014 Zurich Univ. & ETH	FPGA	Classification	Frame-based	LIF	Digital	不可	FC	RTL
Fast pipeline_2015 Univ.of Sevilla	FPGA	DVS-based classification	DVS	LIF	Digital	不可	Conv/Pool	RTL
Hfirst_2015 Johns Hopkins Univ.	FPGA	DVS-based object recognition	DVS	Complex IF	Digital	不可	Conv/Pool	RTL
DYNAPS _2017 Zurich Univ. & ETH	FPGA	Classification	DVS	AdExp-IF	Analog	不可	Conv/Pool	CHP language
Conf Conv Node_2018 Univ.of Sevilla	FPGA	DVS-based classification	DVS	LIF	Digital	不可	Conv/Pool	RTL
2019 Cote d'Azur Univ.	FPGA	Embedded-AI classification	Frame-based	IF	Digital	不可	FC	N2D2, TF, Keras

Neuromorphicとは

1. Signaling models (Spiking Signal) ; 量子化された情報を伝送
2. Timing/clock ; (非同期)、 Event-Driven

} 基本

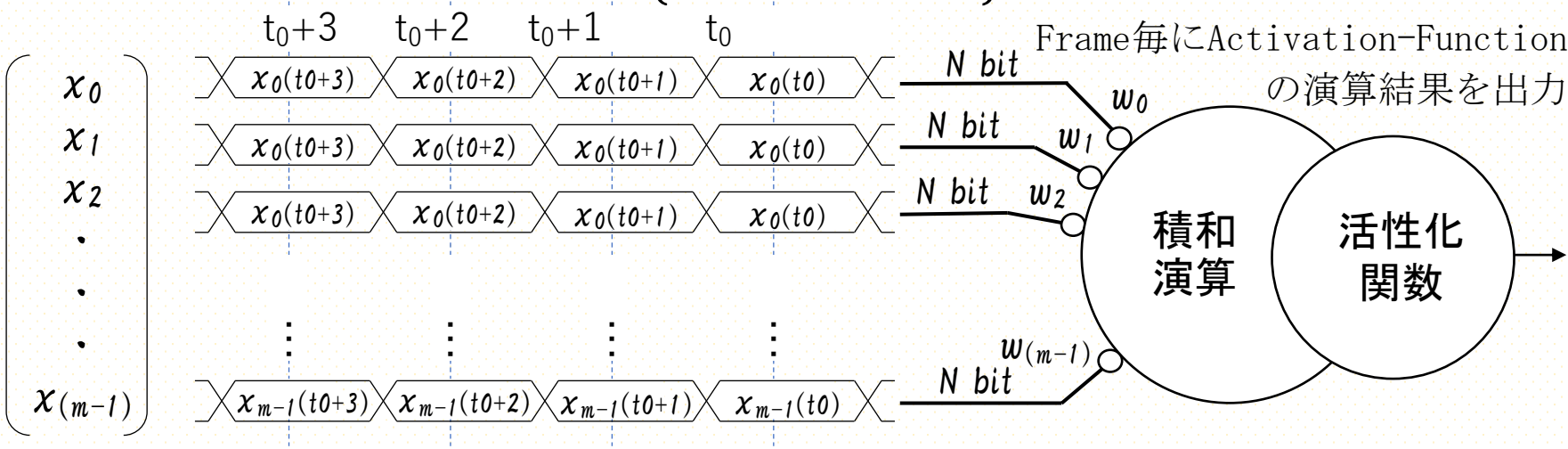
3. Integrated memory and compute ; (ニューロン動作 の ハードウエア化)
4. In-Situ learning ; (Back-GroundでのFeed-Forward学習)
5. Fault tolerance, Noise tolerance ; (多数決動作、確率的動作、近似動作)
6. Nonvolatile memory ; (人エシナプス、新材料メモリ、新製造工程)
7. Architecture ; (脳型Computing、Probabilistic Computing)

Conclusion :
The development of a new brain-like computational system will not evolve in a single step ; it is important to implement well-defined intermediate steps that give useful scientific and technological information.

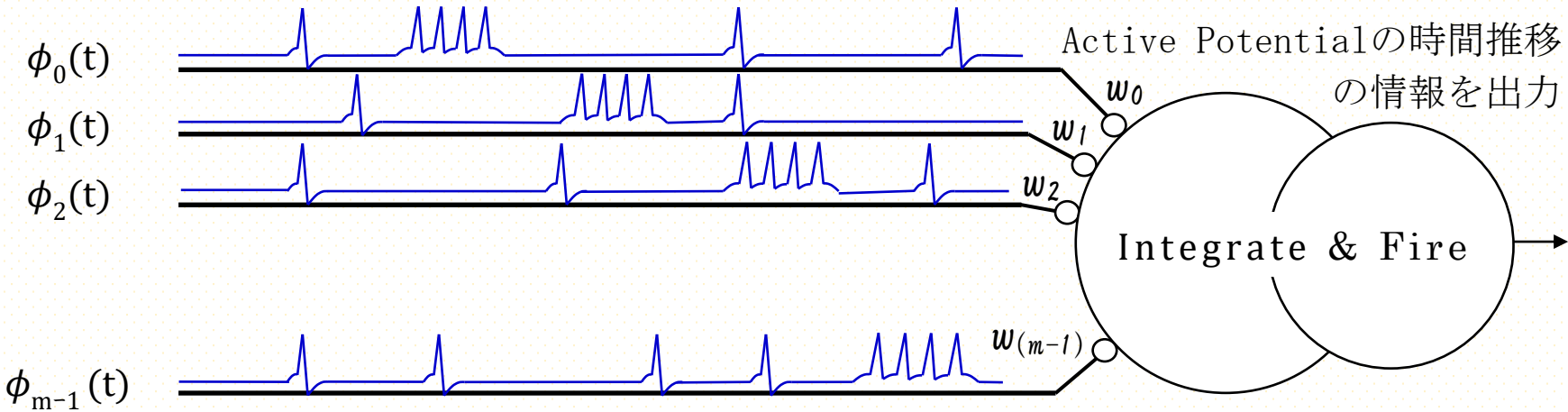
(引用) 米国DOE Report : 2015 ; https://www.osti.gov/biblio/1283147
Neuromorphic Computing – From Materials Research to Systems Architecture

Neuron Model

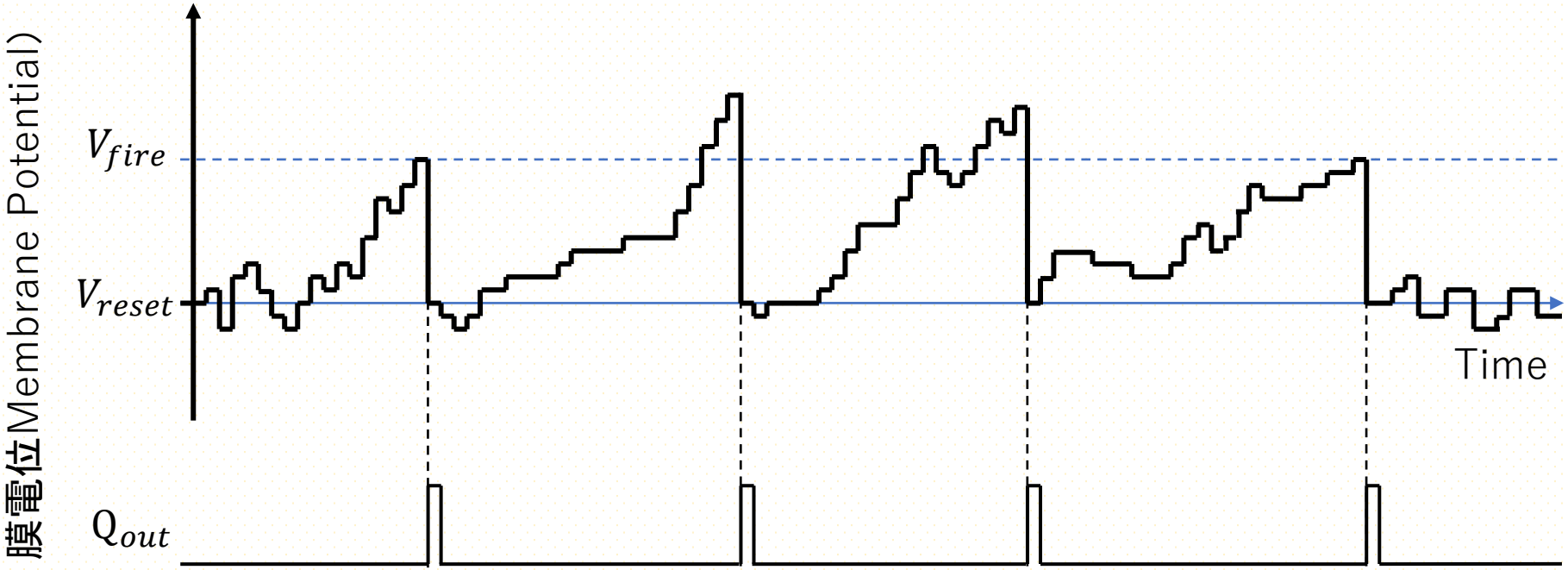
1) Artificial Neuron : Frame-based (or Batch-based)



2) Spiking Neuronの当初のコンセプト；非同期 Event-Driven、 Frame-less



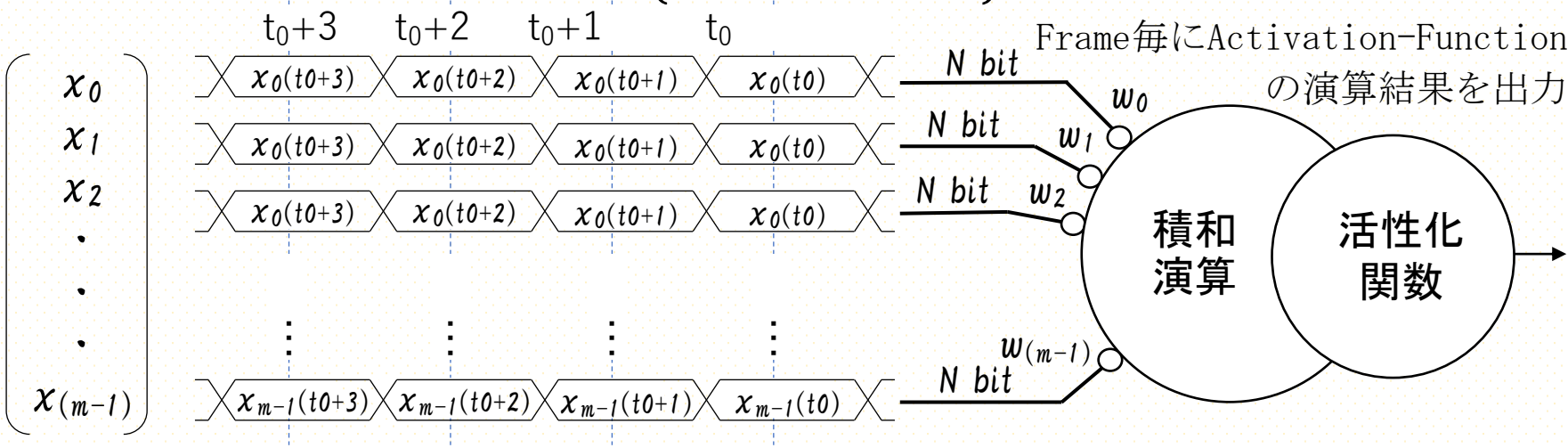
Integrate & Fire の動作例



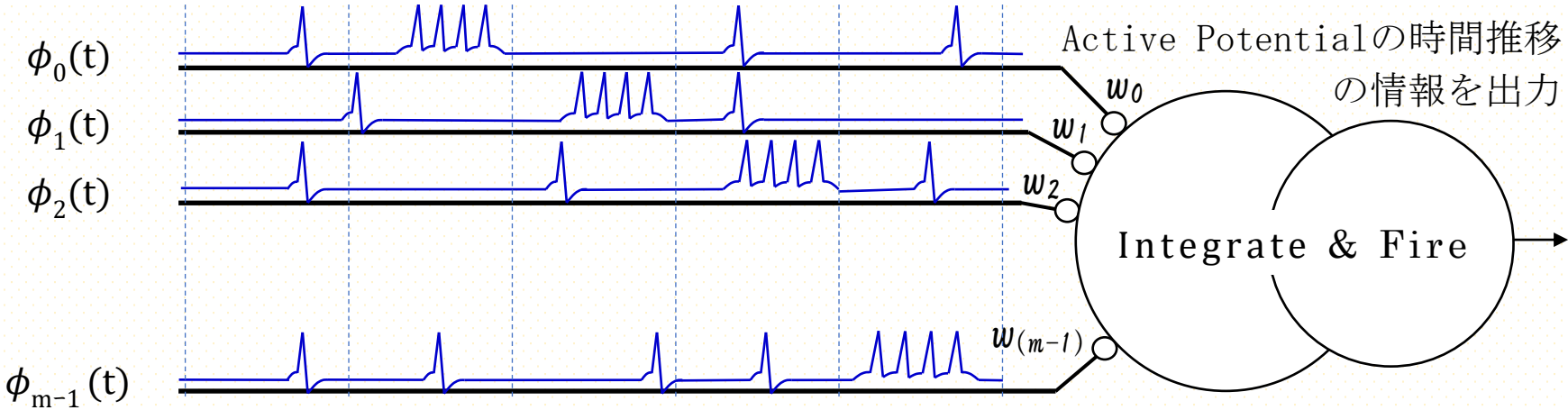
（参考）Pérez-Carrasco, J. A. et al.; Mapping from frame-driven to frame-free event-driven vision systems by low-rate rate coding and coincidence processing—application to feedforward ConvNets., in IEEE Trans. Pattern Anal. Mach. Intell. 35, 2706–2719 (2013).

Neuron Model

1) Artificial Neuron : Frame-based (or Batch-based)

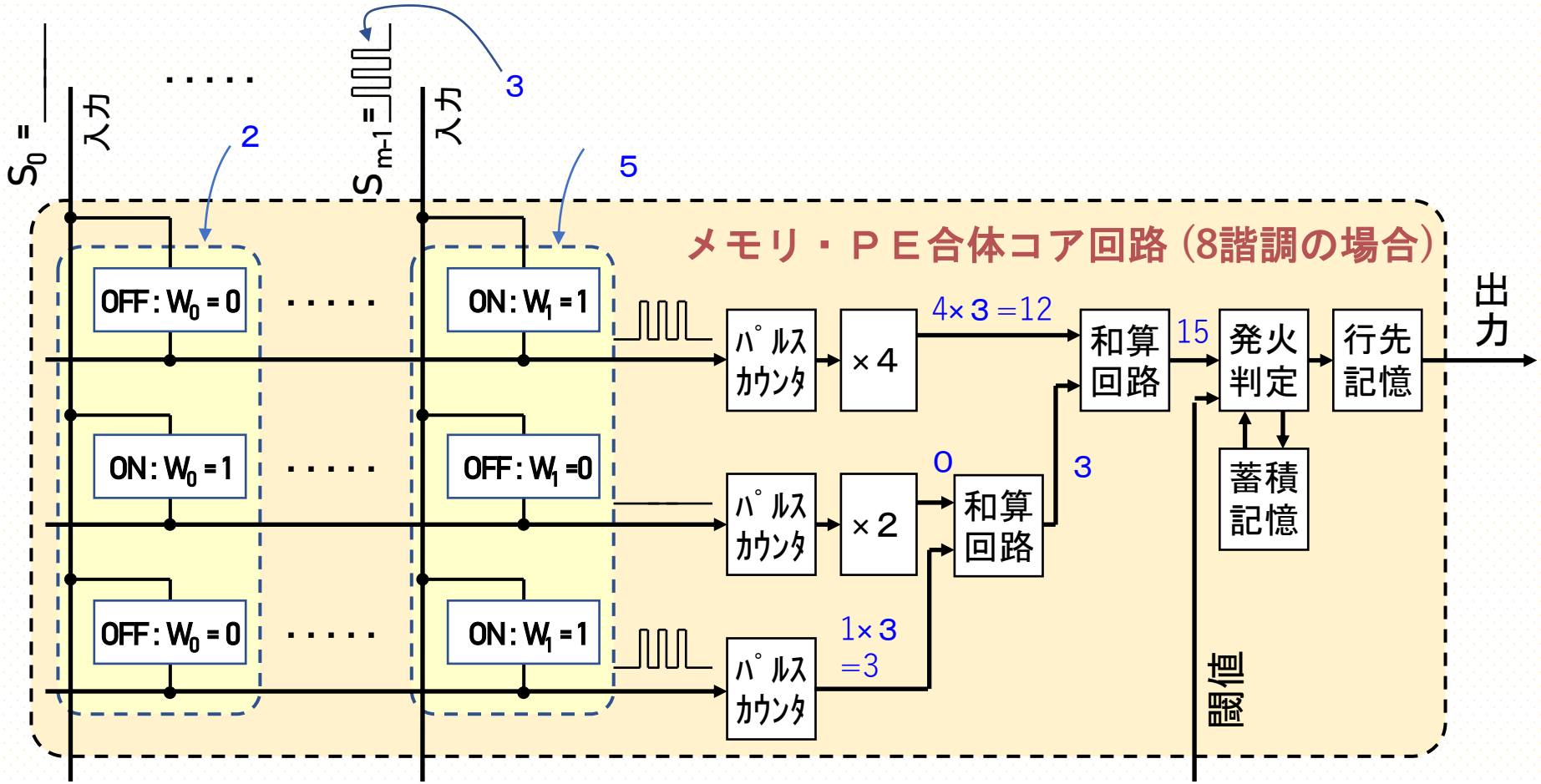


2) Spiking Neuronの当初のコンセプト；非同期 Event-Driven、 Frame-less

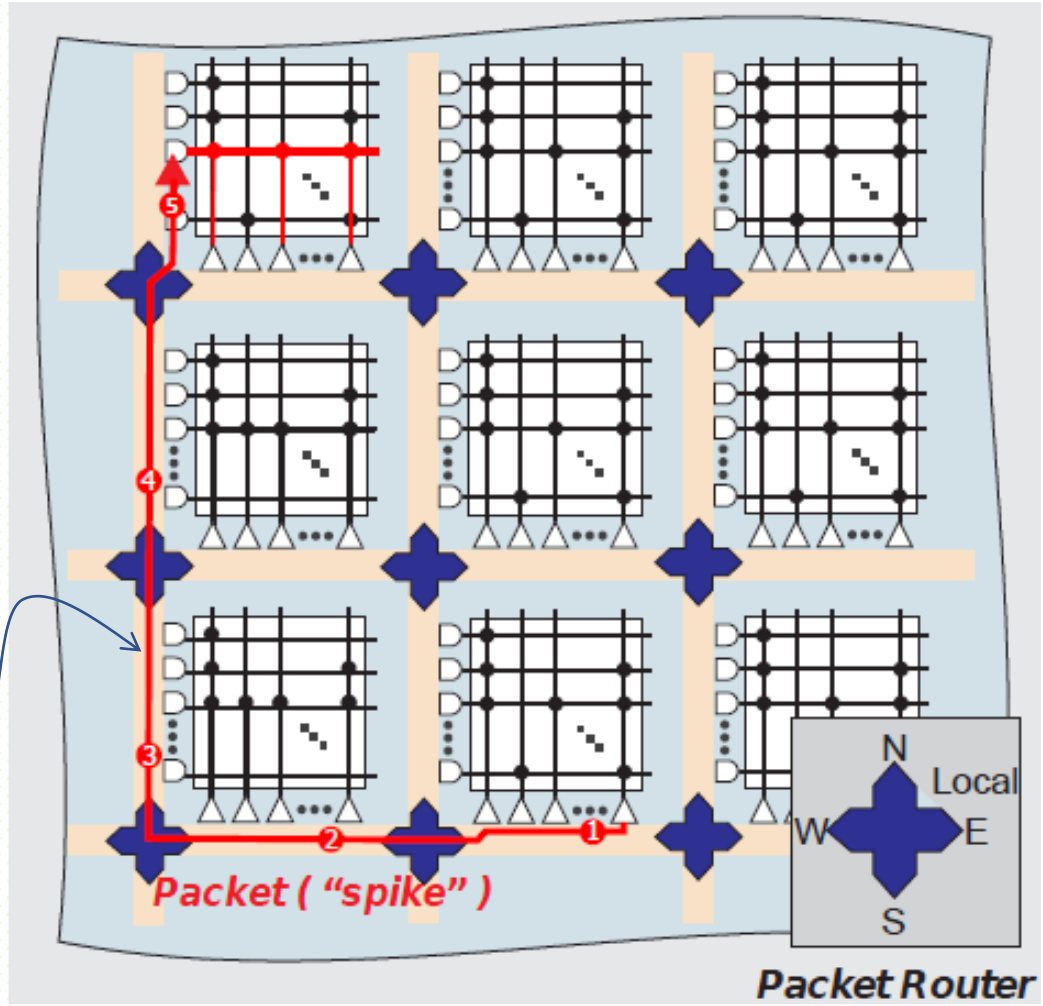


整数方式での積和演算：In-Memory方式

シナプス・データを読み出さずに、メモリ内で２進数のケタ毎の行列積算を行い、その周辺回路にて和算を行い、発火判定し出力する。

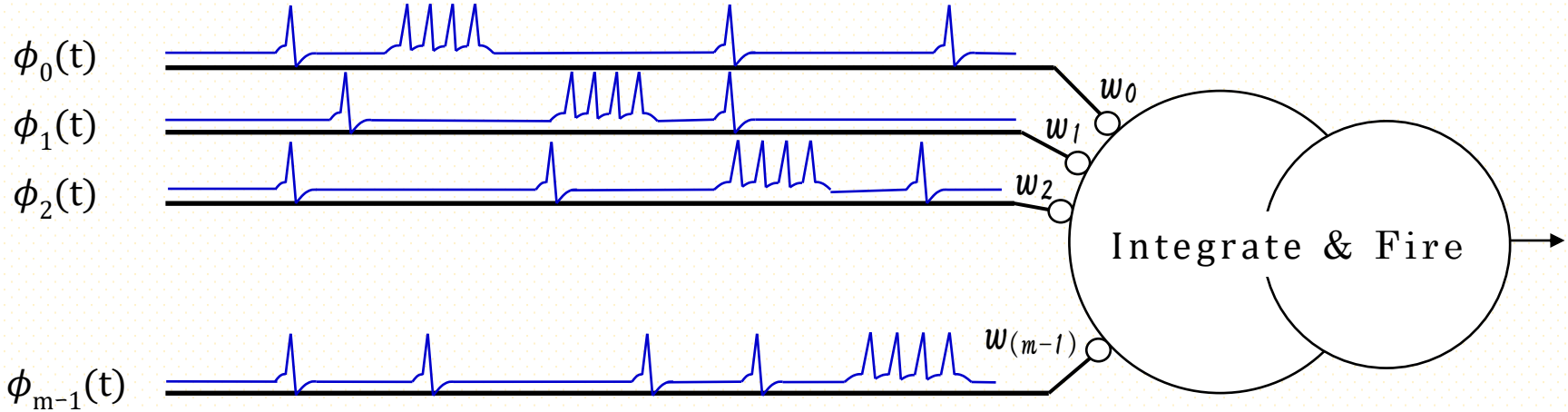


Paul A. Merolla, et al., A Million Spiking-Neuron Integrated Circuit with a Scalable Communication Network and Interface, In Science（2014）

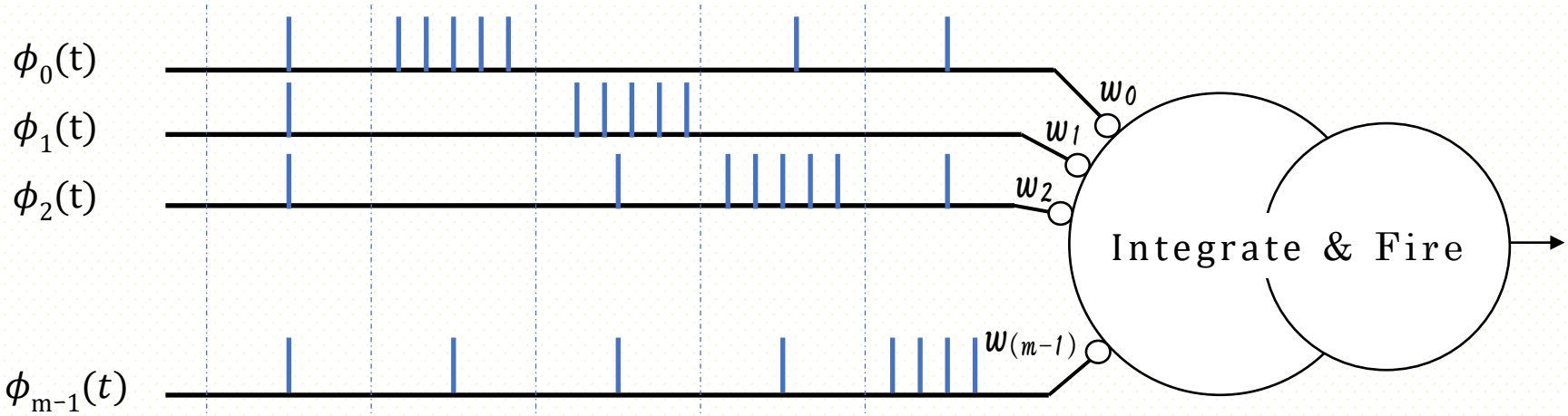


SNNのニューロン・モデル

2) Neuromorphic の 基本コンセプト



3) 現実のSNN表現



Spiking SignalへのEncoding技術：

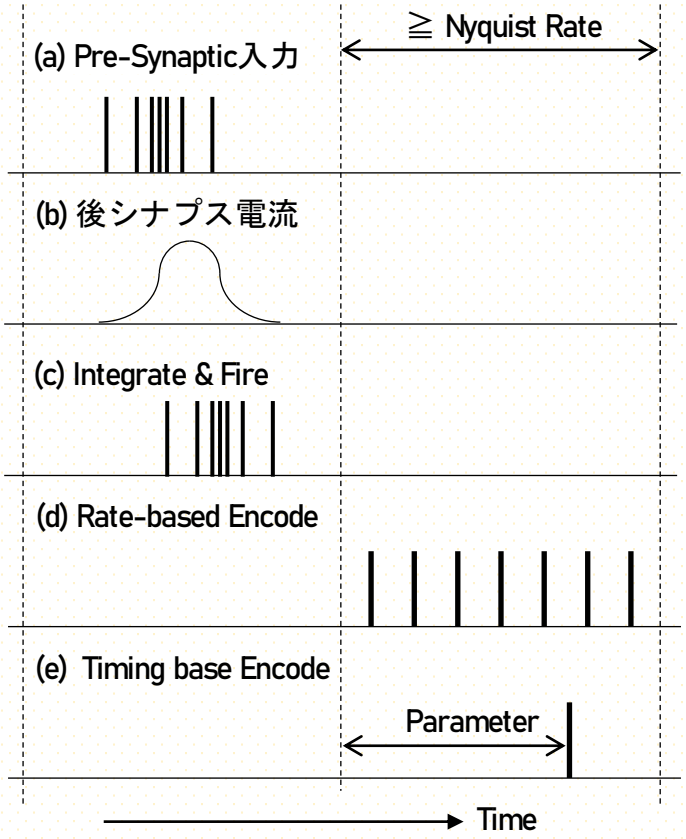
参照） Romain Brette; “Philosophy of the spike: Rate-based vs. spike-based theories of the brain.”, in Frontiers in Systems Neuroscience, Nov 2015,

■ 必要性

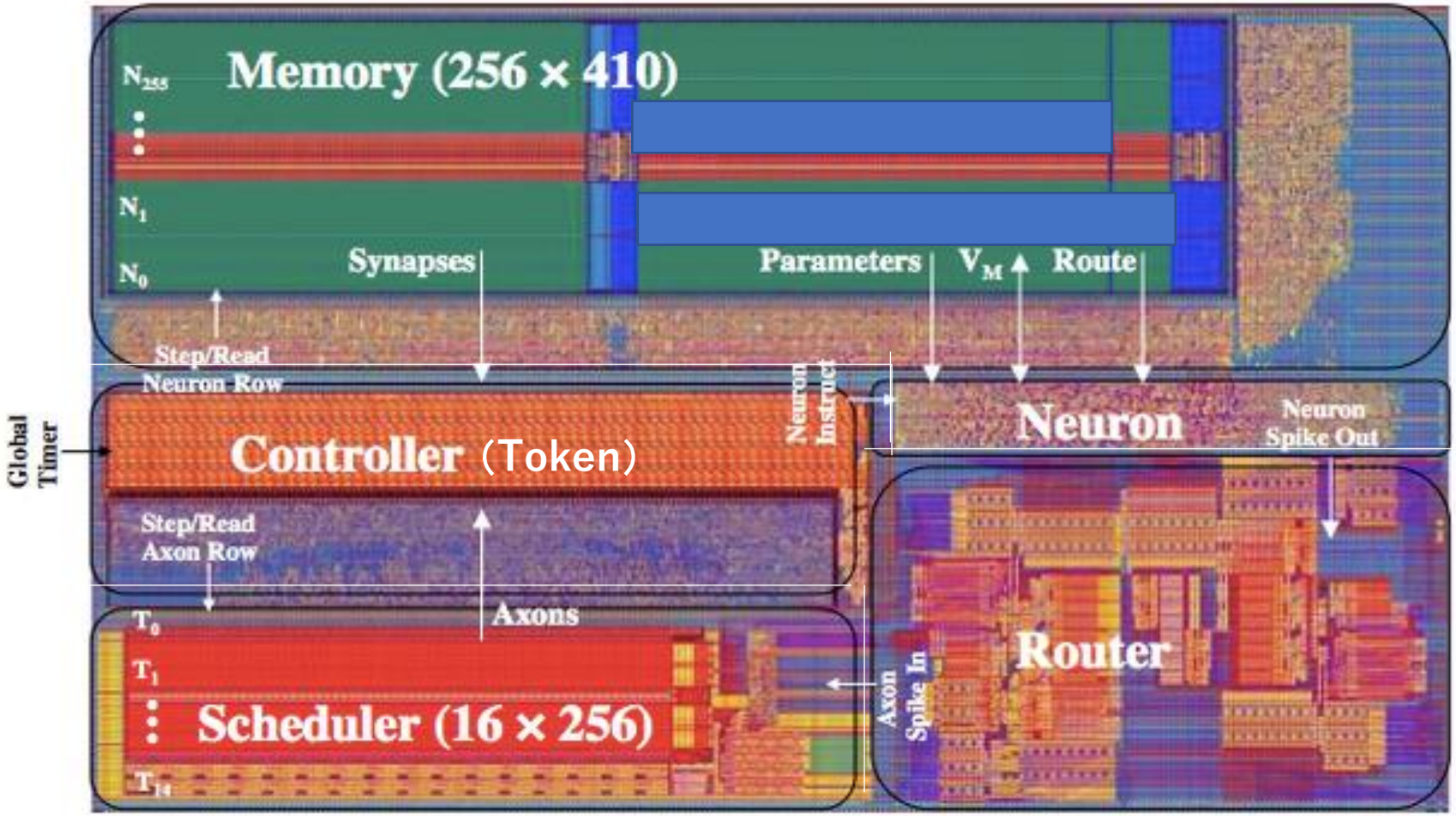
1) **ネットワーク用のパケット・フォーマットとの間の乗り換え**
2) **センサー等の入力デバイスとの接続のため**
3) DNN / DLN**を実装のため**
4) Probabilistic Computing **を行わせるため**
5) **ソフトウエアの見込む結果と合わせるため**

■ 議論：Spikeは、Eventの発生を知らせるのだが、

- 入力の積分値が、閾値を超える毎に発火すれば良い（図c）
- 発火はランダムであり、個別の発火、特に、正確なタイミングは情報とならない。 従って、入力の積分値と、スパイクがバースト発生する時の周波数を関連付ける
  ⇒ Rate-based、図d
- パルス発生タイミングと、基準タイミングの時間差にて、何らかの「量」の情報を伝送する。⇒ Timing-Based、図e
- 近隣のニューロン群との同調性を監視し、発火にタグ付けした情報も送る。

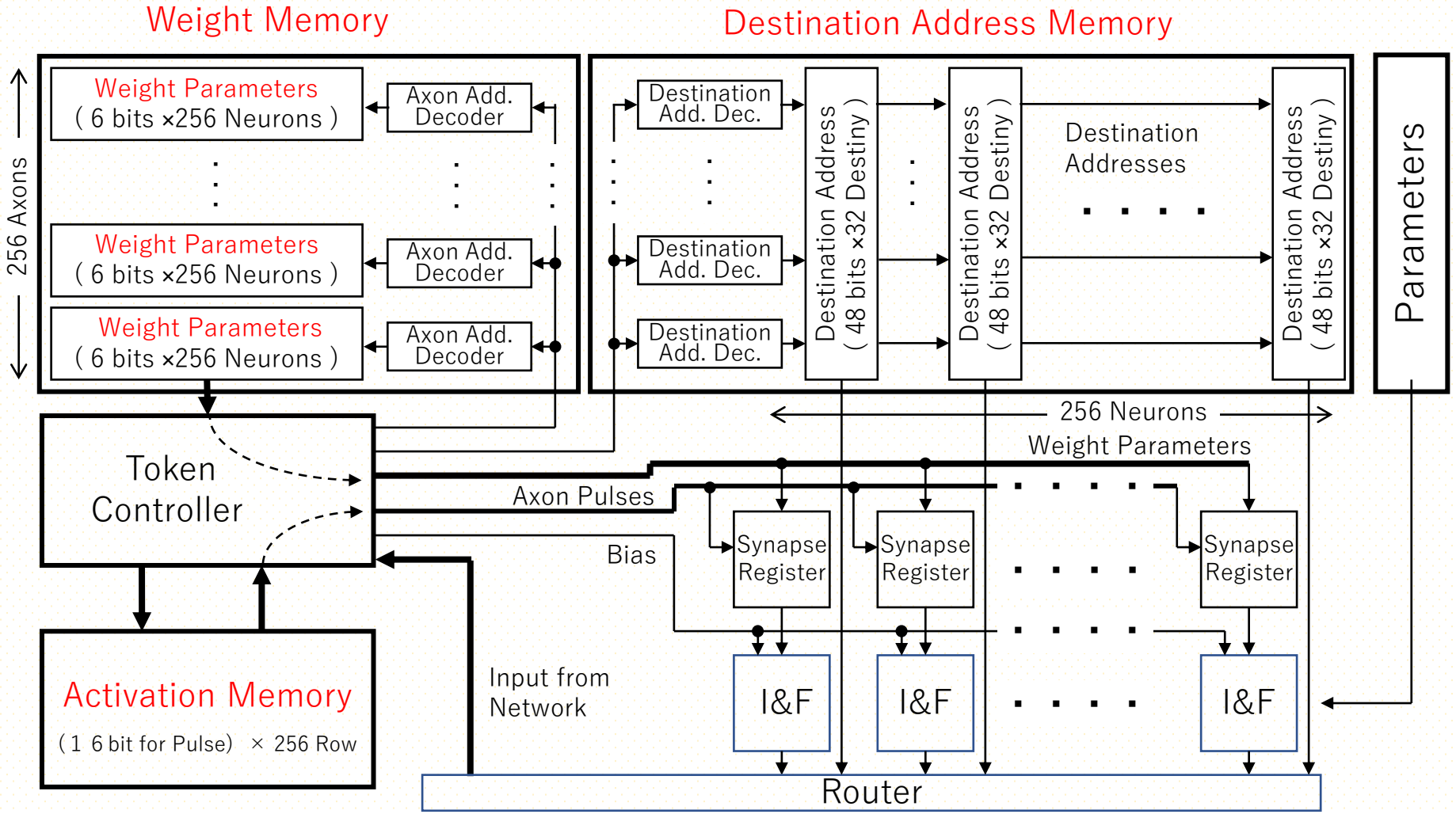


Paul A. Merolla, et al., “A Million Spiking-Neuron Integrated Circuit with a Scalable Communication Network and Interface”, in Science（2014）



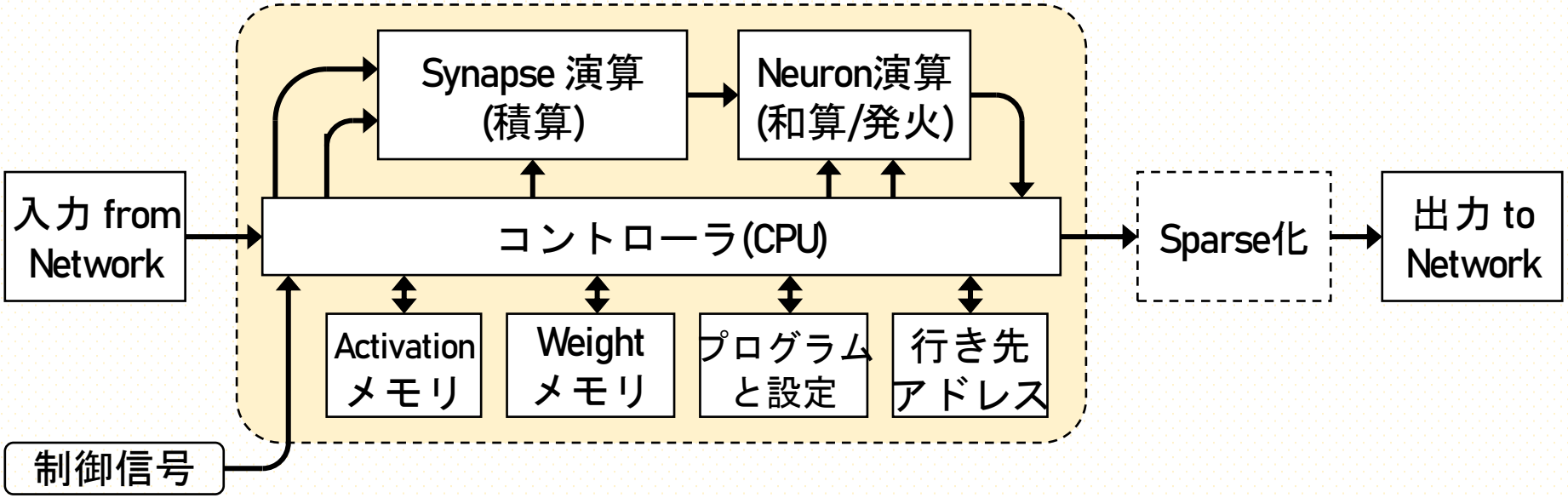
出典：https://electronics360.globalspec.com/article/4445/ibm-seeks-customers-for-neural-network-breakthrough

Paul A. Merolla, et al., “A Million Spiking-Neuron Integrated Circuit with a Scalable Communication Network and Interface(2014)”を読み解き、発表者が作成した図

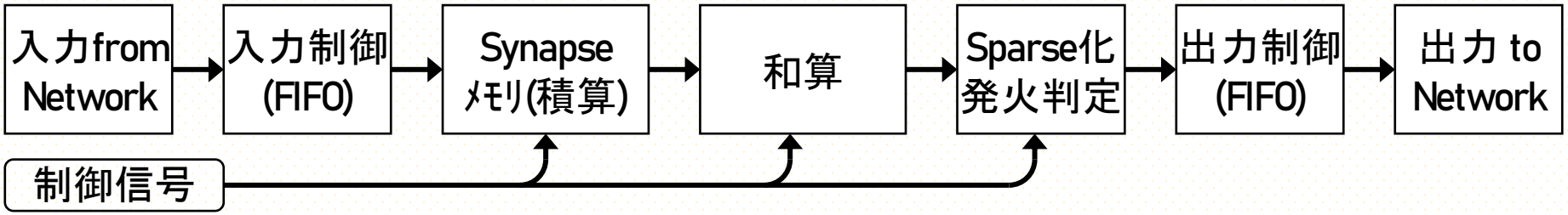


SNNのニューロン・コアの電子回路表現

(A) 仮想ニューロン方式

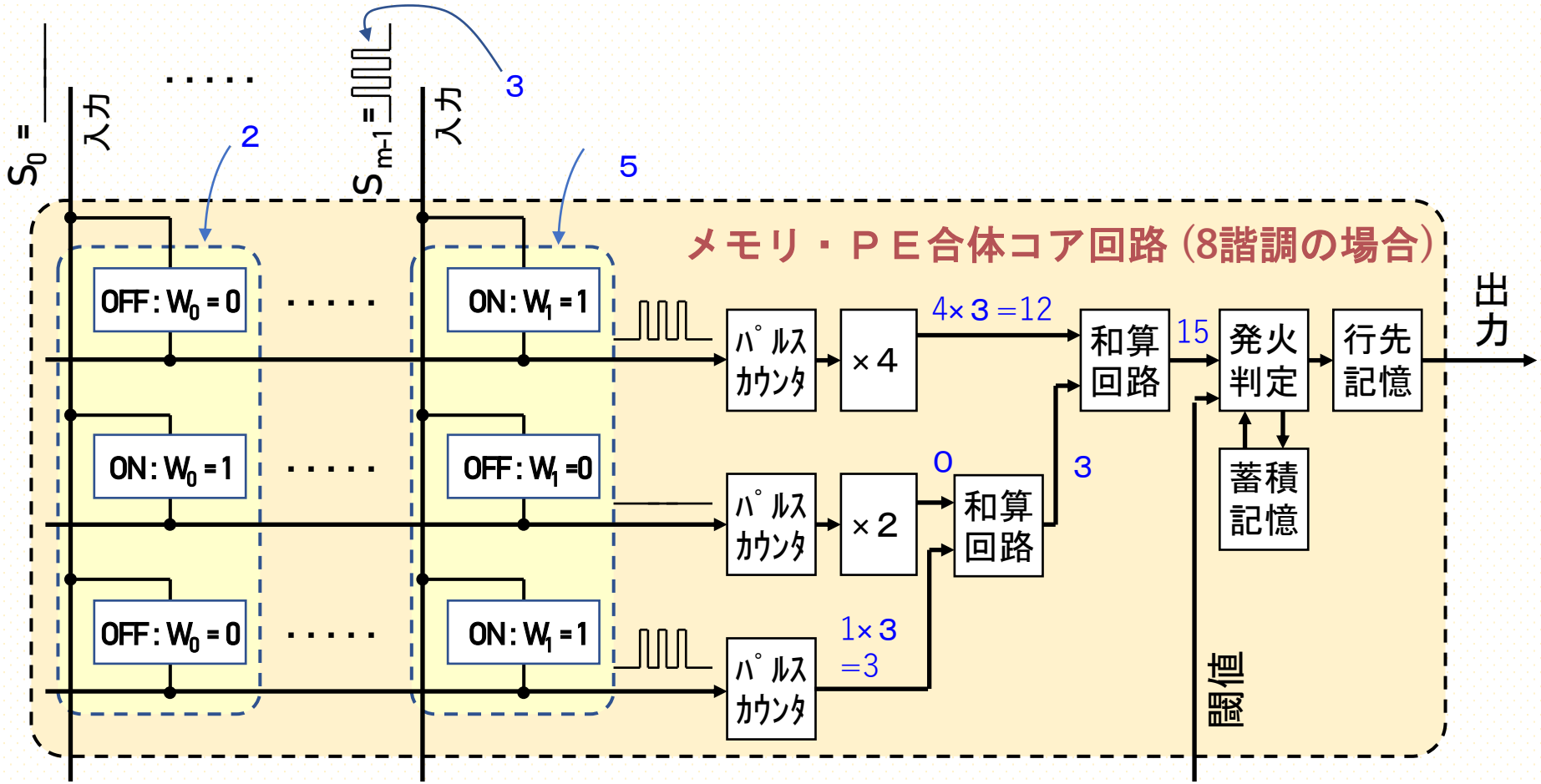


(B) Fully Parallel 方式 / In-Memory方式

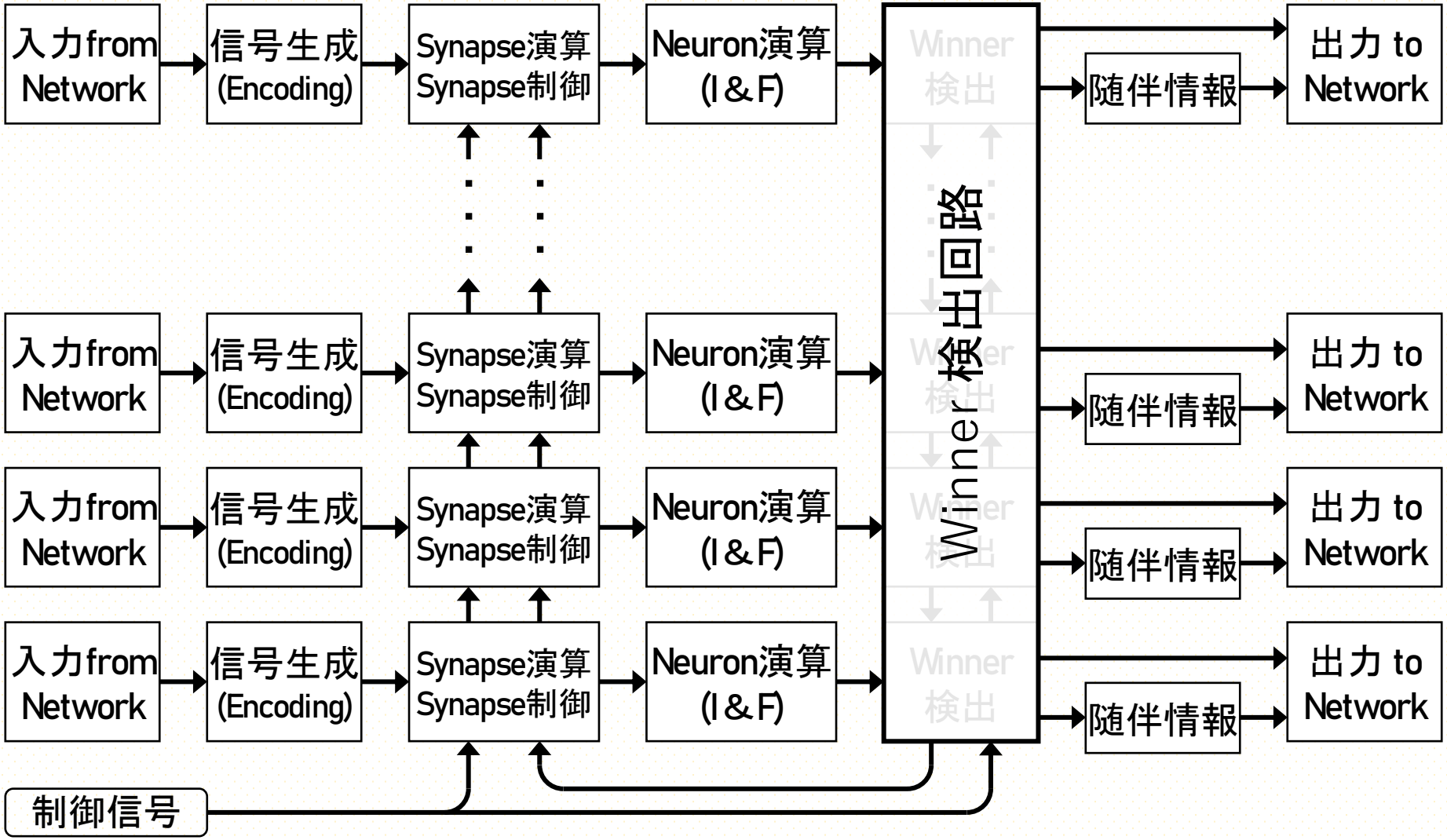


整数方式での積和演算：In-Memory方式

シナプス・データを読み出さずに、メモリ内で２進数のケタ毎の行列積算を行い、その周辺回路にて和算を行い、発火判定し出力する。



(参考) Tomoki Fukai, and Shigeru Tanaka; A Simple Neural Network Exhibiting Selective Activation of Neuronal Ensembles: From Winner-Take-All to Winners-Share-All, in *Neural Computation* (1997)



SNN-ANNのベンチマーク議論

SNN-ANN ベンチマーク

SNNは、本当に、ANNの次の世代の技術なのか？

1）ベンチマーク方法

- 共通のデータ・セット
- 共通のワークロード

2） 性能指標か何か？

- 認識精度
- 消費電力
- レイレンシ
- 回路規模（製造コスト）

3）棲み分けるのか？

- アプリケーション
- 市場ニーズ

Rethinking the performance comparison between SNNs and ANNs.

・SNNの認識精度はANNを上回ってはいない。

それは、SNNのメリットを理解したベンチマーク・ワーク・ロードが無く、ANNベンチマーク用の静止画セットをSpiking用に変換してトレーニングし評価するという状況が続いていたためである。

・既に、ANNの認識精度は人間を上回っているのに、Brain-Likeなメカニズムを導入して、ANNの性能を上回ろうとするには無理がある。

・ANNの成功は、成熟した学習モデル、多彩なベンチマークインフラ、プロセッサの性能向上が支えているが、SNNにはそれらの環境に劣る。

Rethinking the performance comparison between SNNs and ANNs.

Accuracy on ANN-oriented workloads (Frame-based datasets)

Dataset	Network	Model	Accuracy
MNIST	MLP	Model-1 (Natural ANN ※1)	98.60%
		Model-2 (Converted SNN ※2)	98.51%
		Model-3 (Enforced SNN ※3)	98.31%
	CNN	Model-1 (Natural ANN ※1)	99.31%
		Model-2 (Converted SNN ※2)	99.07%
		Model-3 (Enforced SNN ※3)	99.22%

※１） ANN training & ANN inference

※２） Converted SNN with ANN training and SNN inference;

※３） Enforced SNN with SNN training and SNN inference.

SNN model is directly trained from scratch on the converted spike datasets, using BP-inspired supervised algorithms for training SNNs, and tested in the same domain.

Rethinking the performance comparison between SNNs and ANNs.

Accuracy on SNN-oriented workloads (Frame-free spike datasets)

Dataset	Network	Model	Accuracy
N-MNIST	MLP	Model-4 (Enforced Binary ANN ※4)	97.24%
		Model-5 (Enforced Intensity ANN ※5)	97.63%
		Model-6 (Natural SNN ※6)	98.61%
	CNN	Model-4 (Enforced Binary ANN ※4)	99.08%
		Model-5 (Enforced Intensity ANN ※5)	98.63%
		Model-6 (Natural SNN ※6)	99.42%

※４） This uses converted binary for ANN training and ANN inference

※５） This uses intensity images respectively, for ANN training and ANN inference

※６） SNN training and SNN inference.

Towards spike-based machine intelligence with neuromorphic computing

・Encoding技術によって、 従来のデータ・セットを使ってSNNを評価できるようになったが、それではSNNのメリットを引き出せない。

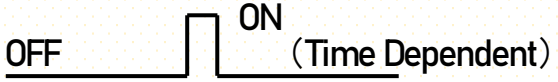
・SNNの実用的な価値については、長い間議論が続いているが、その為に、ニューロモーフィック・コンピューティングの開発は遅れており、ディープラーニングの急速な進歩によって、状況は悪化している。

・SNNでは、強化学習の実装ができていない. 結果、ANNでトレーニングを行い得たシナプス係数を変換してSNNに適用しするというConversion-based アプローチがとられているが、そのようにしてANNと同等の精度を得たとしても、SNNの入力信号にはEncoding時間が追加となるので、Inferenceのレイテンシが伸び、エネルギー効率も下がる。

・SNNに、ANNの学習方法(BP等)を実装し、ANN用のワークロードでトレーニングし、ANNベースの評価を行うというのでは、ハードウエアの進歩とネットワークサイズの拡張が著しいANNに追いつくことはできない。

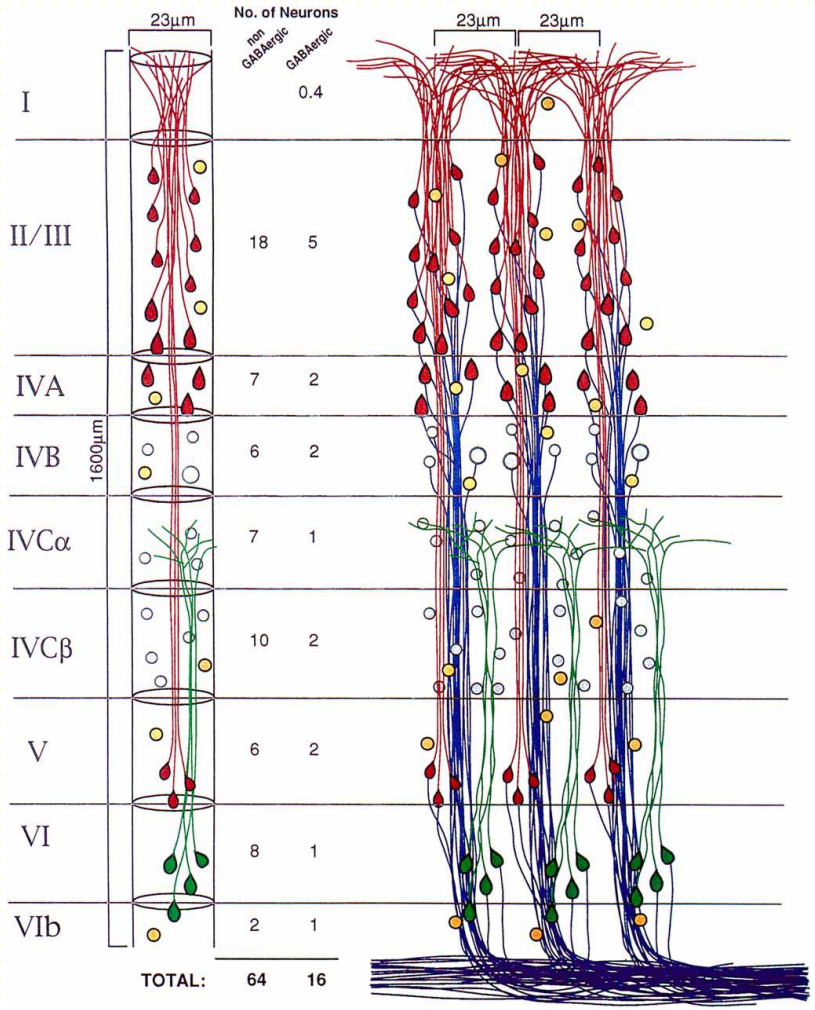
考察

SNN-ANN比較のまとめ

	SNN	ANN
Data	整数 と分数	浮動小数点数値
Signal in Hardware Layer		
Neuron Model	Integrate & Fire	McCulloch-Pitts の 形式ニューロン （活性化関数）
Synapse Model	Dynamic Synapse (Spike Time Dependent Plasticity)	Static Synapse (Updated through Off-Lined Training)
State Description	(Liquid State)	Static State
演算器	面積小（In-Memory 演算）	汎用ALU / Accelerator
消費電力	～ 1/10000	1 (Ref.)
分類精度	< 1.0	1 (Ref.)
課題	Network-Topology Network技術 Spike Encoding	Computing Power削減 学習の高速化

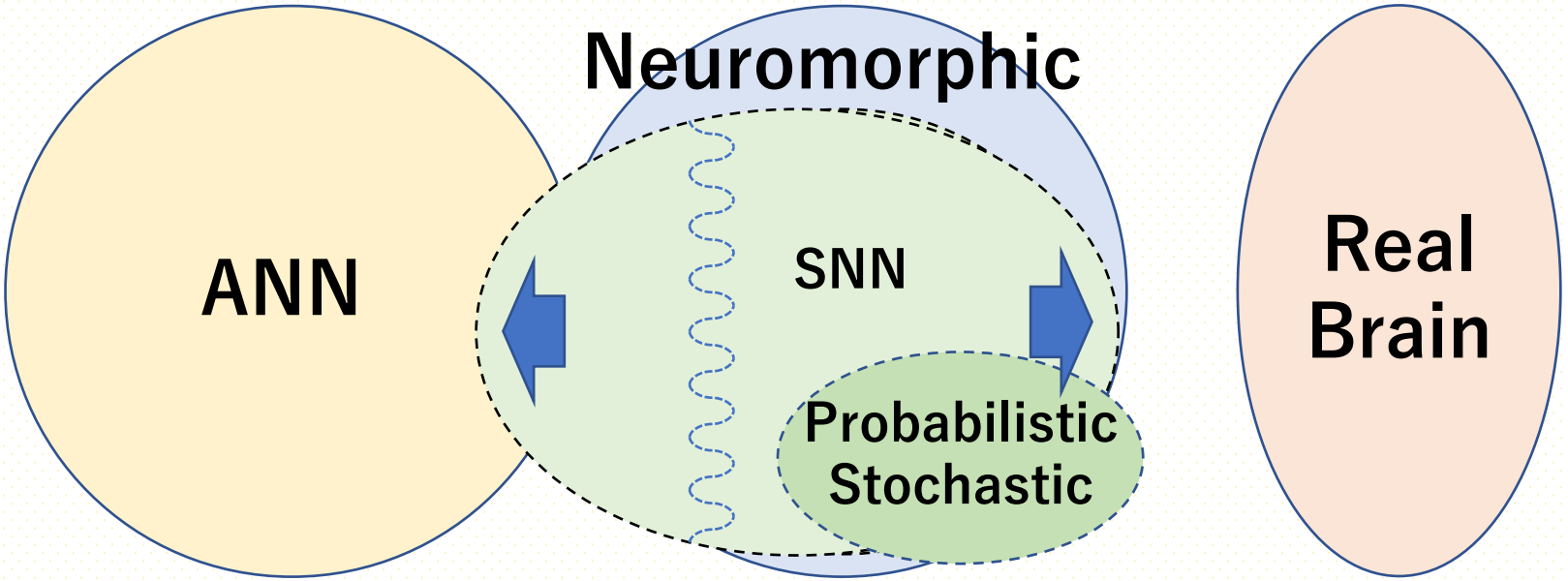
アトム回路はニューロンか？ ： ミニコラムへの挑戦

- 大脳皮質の局所神経回路
- ３系統以上のネットワーク接続を持つノード回路（**少なくとも２層はGlobal**）
  - **近隣とのLocal 接続**
    - ＞ **畳み込み層に似る？**
    - ＞ **分類学習に関連？**
    - ＞ **多数決制御に関連？**
    - ＞ **LSMの単位？**
  - **他の領野との接続**
  - **大脳辺縁系や視床との接続**



［出典］ A Peters, C Sethares; Myelinated axons and the pyramidal cell modules in monkey primary visual cortex, in Journal of Comparative Neurology (1996)

SNNのアプリケーション



- ANN用のハードウエアのエネルギー効率を改善
- Edge-AI (スマホ/自動車)

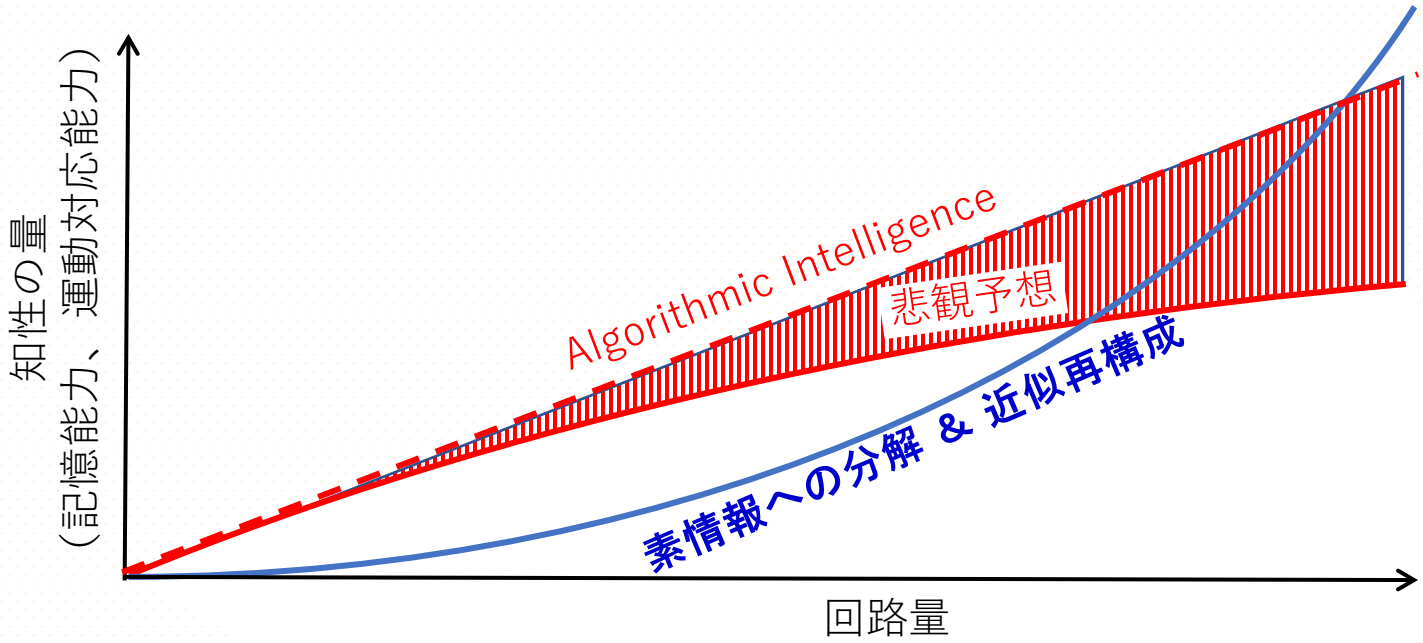
サイエンス志向 + Roadmap

⇒ AGI

ご清聴ありがとうございました。

仮説提案 ： 素情報の近似再構成ネットワークの能力

- Algorithmic Intelligence **の能力は、回路量に比例してしか拡大せず、** resilience**に欠ける。（ANN戦略）**
- **「素情報への分解 ＆ 近似再構成戦略」の能力は、回路量の２乗で拡大しうる。（メカトーフの法則の拡張解釈）**



SNNの表現戦略（私案）

1） 要素ゲート回路(ニューロン)

- 外部入力（電気インパルス）による活動電位上昇が閾値以上となった時に発火
- 学習機能とFeed-Back機能

2） 脳内ネットワークの最小ノード単位(ミニコラム)

- 機能
  - ① 外部からの信号に対するFilter機能（存在理由の主張）
  - ② 近隣ニューロンの動作の監視制御機能（量子化/整数化の閾値の制御）
  - ③ 外部からのQueryの受信機能
  - ④ 外部への応答機能
  - ⑤ 上記機能に周辺ニューロン群を参加させる機能
- 外部との情報表現（発火個数や発火間隔の意味）
  - ① 母集団に関する情報（母集団の関する情報）
  - ② Event発生のタイミングに関する情報
  - ③ Eventの量に関する整数情報（もしくは、発火周期に関する情報）
  - ④ Event発生時の誤差、もしくは差分に関する情報
- 数値の演算ルール

量子化表現と確率過程の表現

もう一つの問題点

1） AIにおけるハードウエアの役割り

2） ＡＩの性能と、ハードウエアの性能の関係

3） ネットワークへの要求要件は？

Neuron演算器の性能を表現するパラメータ

- 演算器の専有面積を小さくし、演算個数($\mathcal{N}$)の値を増大させる。
- 演算動作の動作サイクル(f)を高める。

とすると、チップのニューロン演算性能 ＝ $f \times \mathcal{N}$

The computation is usually much simpler in spiking neurons than in formal neurons. Even though several models have been identified in neuroscience studies, in a machine learning context, spiking neurons are most often based on a simple (Leaky) Integrate and Fire (IF) model.

by Nassim Abderrahmane, Edgar Lemaire, Benoît Miramond; “**Design Space Exploration of Hardware Spiking Neurons for Embedded Artificial Intelligence**”, in Neural Networks Volume 121, January 2020, Pages 366-386.

By Clemens JS Schaefer, and S. Joshi;

Quantizing Spiking Neural Networks with Integers,

in International Conference on Neuromorphic Systems (2020).

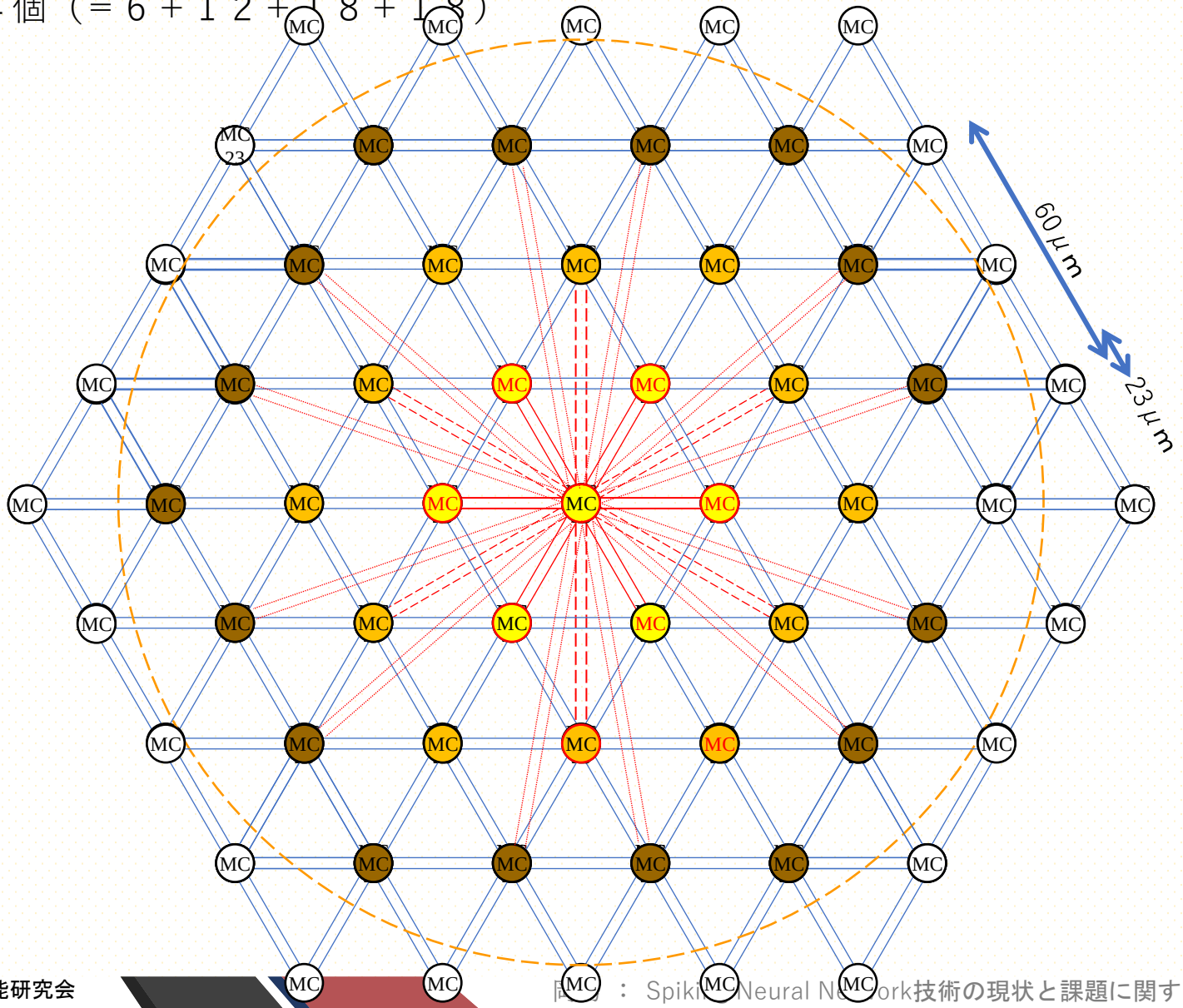
Our results show that **SNNs trained using only integer fixed-point representations** can still retain their accuracy ・・・(中略)・・・ the memory usage of SNNs trained with reduced precision weights, errors, gradients and neural dynamics can be **downsized by 73.78% at the cost of 1.04% test error increase** on the DVS gesture data set.

（整数ベースの演算とすることで、メモリ容量を73.78%減らしても、精度の劣化を1.04% に抑えることができた）

By 深井 朋樹 (理化学研究所)；
研究領域『脳を創る』の2003年研究報告(R013000246)
研究課題名『**時間的情報処理の神経基盤のモデル化**』

我々が提案するメカニズムに依れば、神経回路内で緩やかに結合したニューロン集団の統計的な振る舞いによって、秒レベルの時間スケールでゆっくり変化するニューロン集団のスパイク発火活動が実現され、タイミング情報が生成される。　この仕組みが機能するためには、ニューロンが双安定な内部状態変化（Down状態→UP状態、またはその逆の遷移）を示すことが必要であるが、多くの生理実験の結果がこの仮定を支持している。　仕組みを簡単に説明すると、**各ニューロンは既にUP活動にあるニューロンが自分の周囲にどのぐらい存在するかをリカレント入力の大きさから「判断」し、自分自身がUP状態に移る頃合を見計らう。**比喩を用いるならば、**多数決において他人の挙動を日和見的に観察して賛成票を投じる行為に似ている。**

マイクロカラムは、樹状突起の長さ＜１００μｍ の範囲内ののマイクロカラムと接合を持つ。
（３６～５４個（＝６＋１２＋１８＋１８）



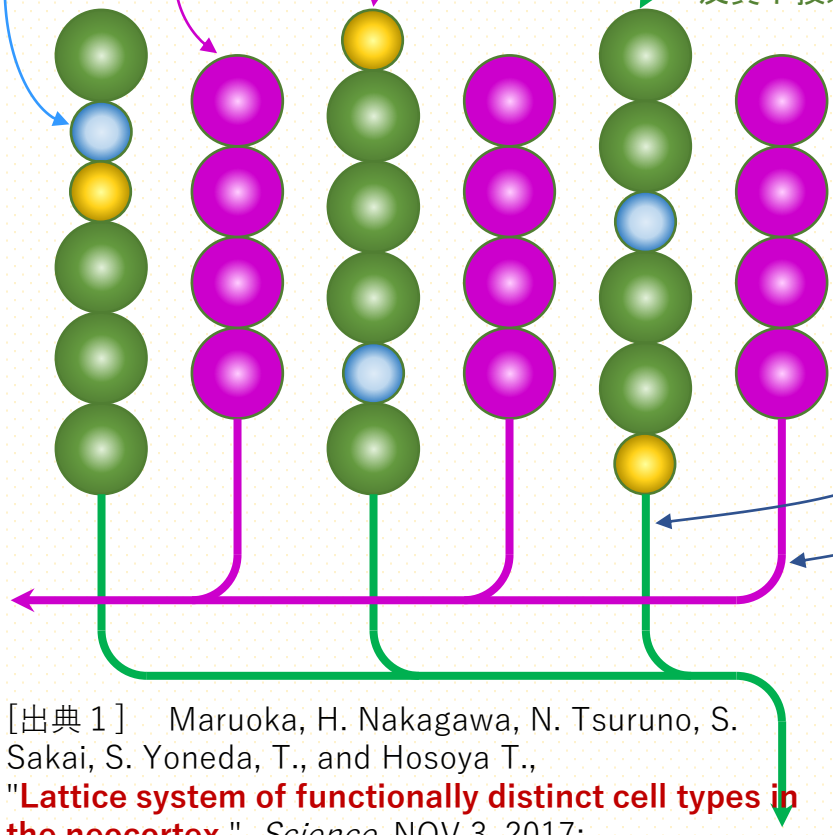
第５層の六方格子状の皮質下投射細胞ミニコラム

パルブアルブモン(ＰＶ)発言細胞(抑制型)： 近傍細胞の樹状突起・スパイン・細胞体・活動電位起始部に結合

皮質投射細胞(ＣＰＮ、活性型)

ソマトスタチン(ＳＯＭ)発言細胞(抑制型)： 近傍細胞の樹状突起やスパインに結合

皮質下投射細胞(ＳＣＰＮ、活性型)



■ 入力

① 興奮性神経線維 from 皮質の他の領域

② 興奮性神経線維 from 視床

③ セロトニン線維 from 中脳の縫線核（Raphe Nucleus）

④ ドーパミン繊維 from 腹側被蓋野（Ventral Tegmental Area、系統発生的に古い中脳の一領域）

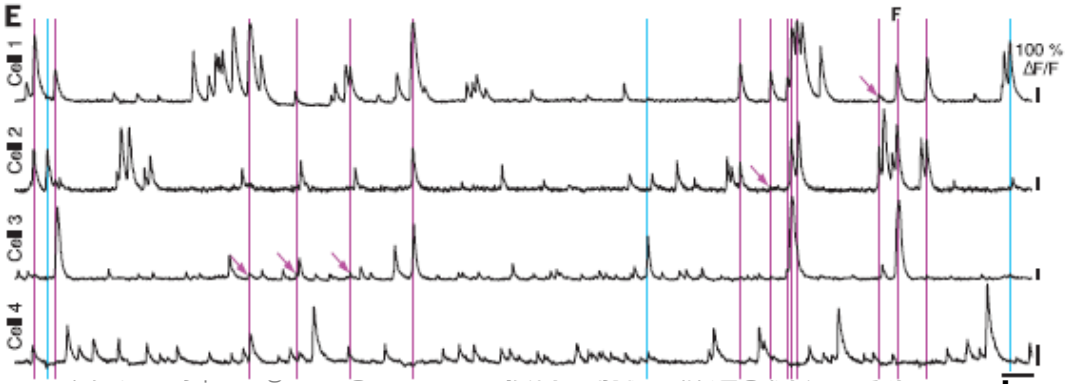
⑤ アセチルコリン繊維 from 大脳基底核（Basal Nucleus）

■ 出力

① 皮質下投射細胞(ＳＣＰＮ、活性型)

② 皮質投射細胞(ＣＰＮ、活性型)

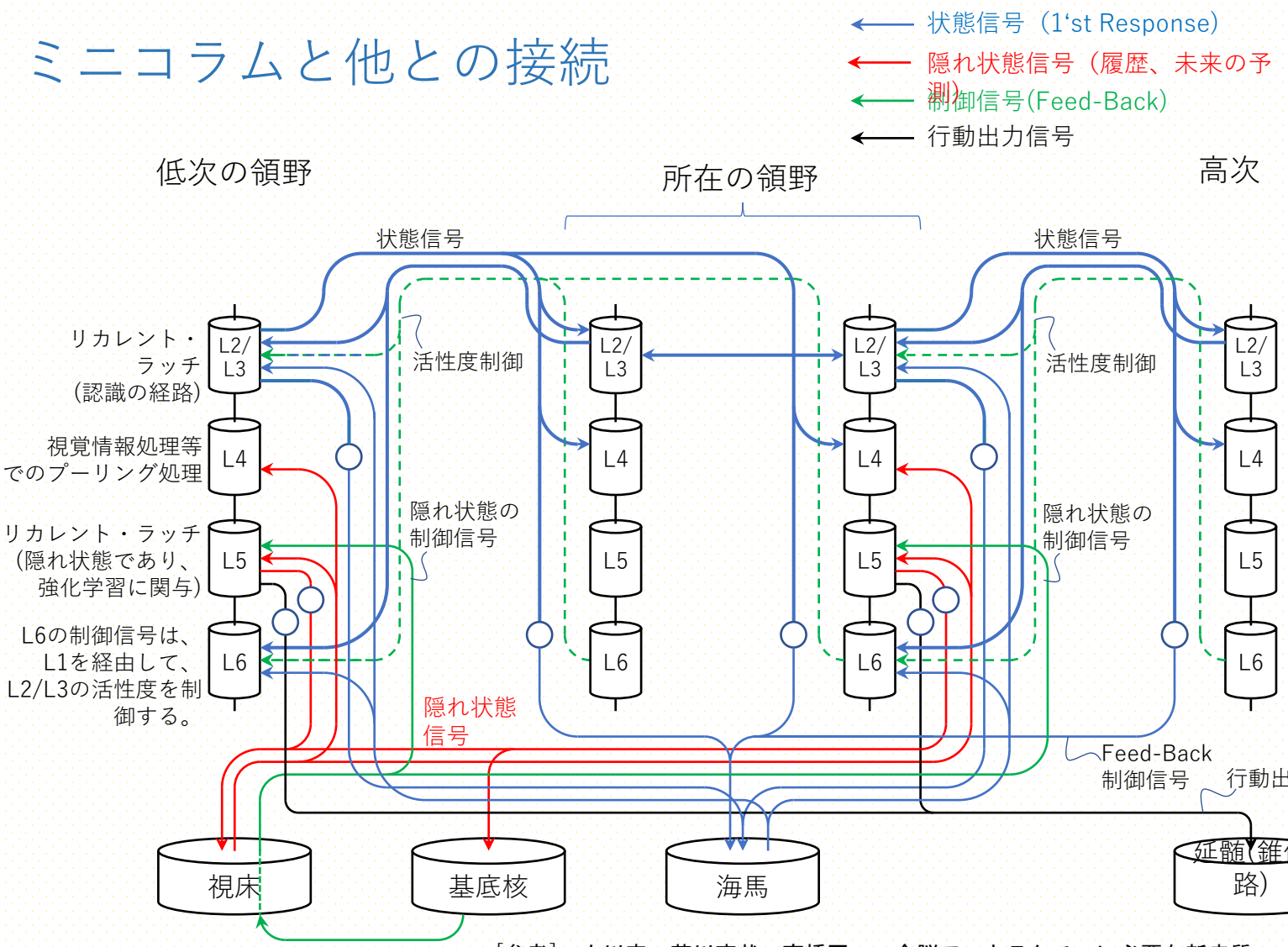
■ 投射細胞（）



[出典１] Maruoka, H. Nakagawa, N. Tsuruno, S. Sakai, S. Yoneda, T., and Hosoya T., "**Lattice system of functionally distinct cell types in the neocortex.**", *Science*, NOV 3, 2017; doi: 10.1126/science.aam6125

[出典２] 窪田芳之、「大脳皮質の神経細胞と局所神経細胞」、日本神経回路学会誌、Vol.21,No.3, 2014, pp122-

ミニコラムと他との接続



Ｌ１層(分子層)では、主に皮質深層の錐体細胞や紡錘細胞の樹状突起、Martinotti細胞の軸索が複雑に絡み合い、水平方向に走る（切線線維）。

Ｌ２層(外顆粒層)では樹状突起は分子層に向かい、軸索は深層に向かう。（皮質内の連絡）。

Ｌ３層(外錐体層)は中型の錐体細胞からなり、皮質内の連絡を行う。

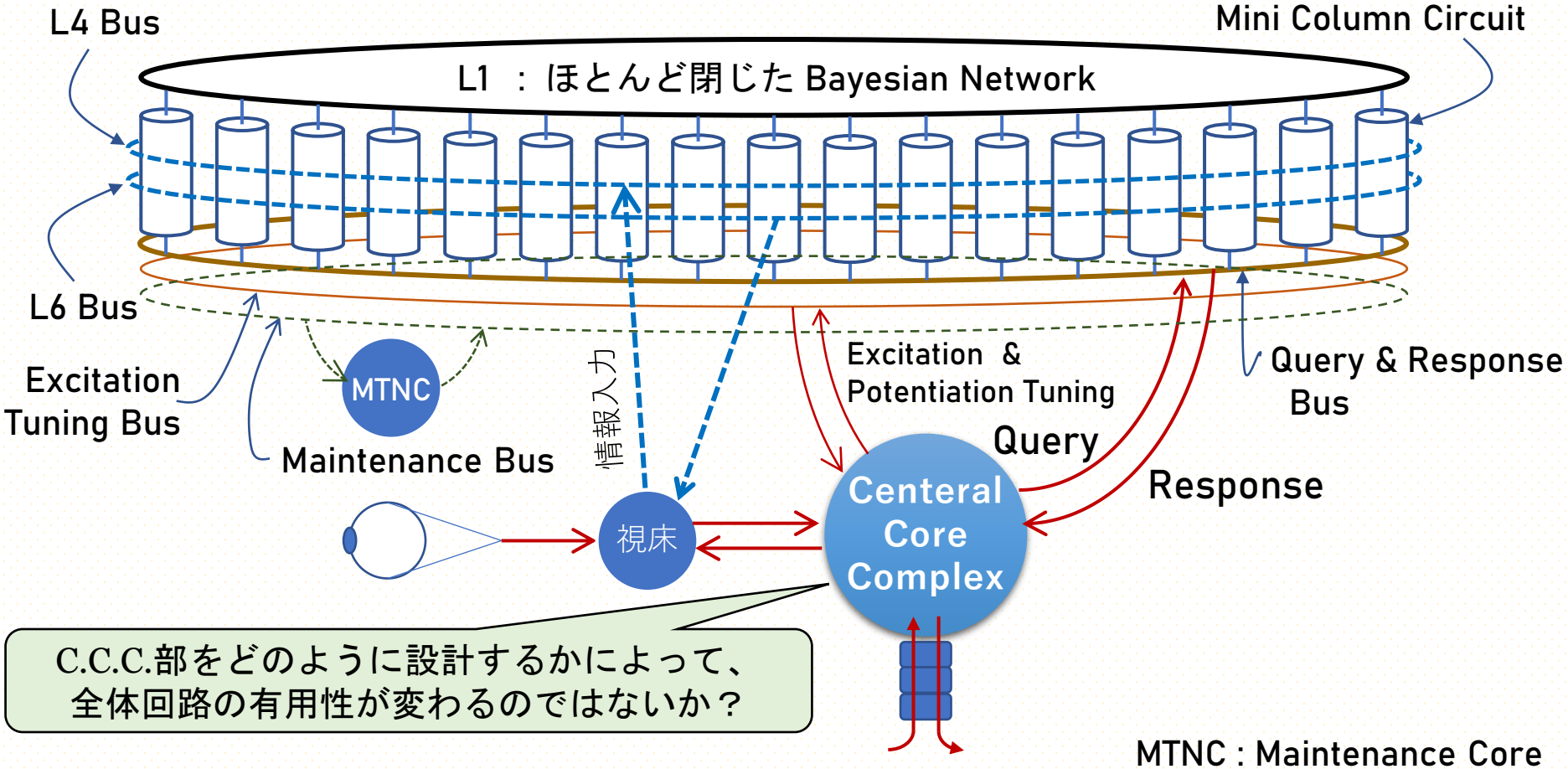
Ｌ４層(内顆粒層)は顆粒細胞が密集し、主に末梢からの入力を受ける

Ｌ５層(内錐体層)の錐体細胞の軸索は投射線維または連合線維として髄質に向かう。（投射：末梢への出力）また、少数の顆粒細胞とMartinotti細胞

Ｌ６層(多形細胞層)の錐体細胞の軸索は投射線維または連合線維として髄質に向かう。（投射：末梢への出力）

［参考］ 山川宏、荒川直哉、高橋亘一；全脳アーキテクチャに必要な新皮質マスターアルゴリズムの検討
The 31st Annual Conference of the Japanese Society for Artificial Intelligence, 2017

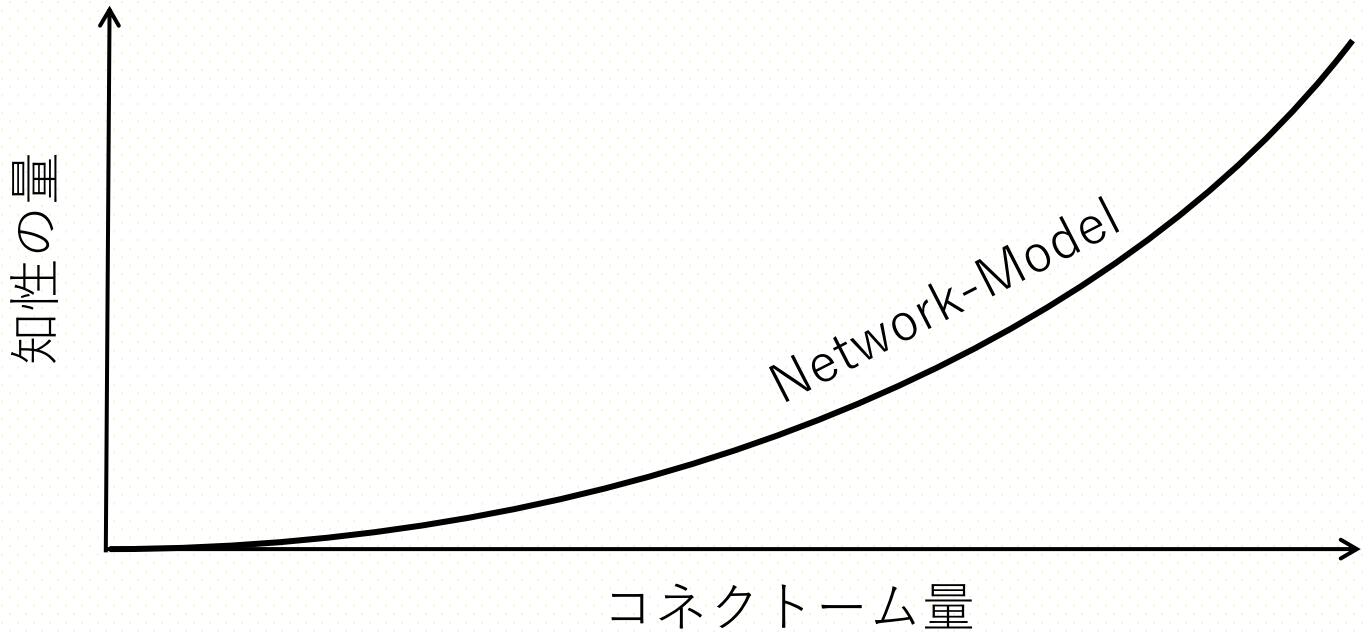
Artificial Brain を載せる Platform（直感的イメージ）



SNNの究極の目標 ： Universalなデータ分解＆論理分解手段の提供

- 近くデータ管理限界に達するので、データを標準素情報（新概念）に分解し、再構成することで、管理要素を減らす必要が生ずる。

- 近く「アルゴリズム管理限界」に達するので、アルゴリズムを標準素アルゴリズムに分解し、再構成することで管理要素を減らす必要が生ずる。

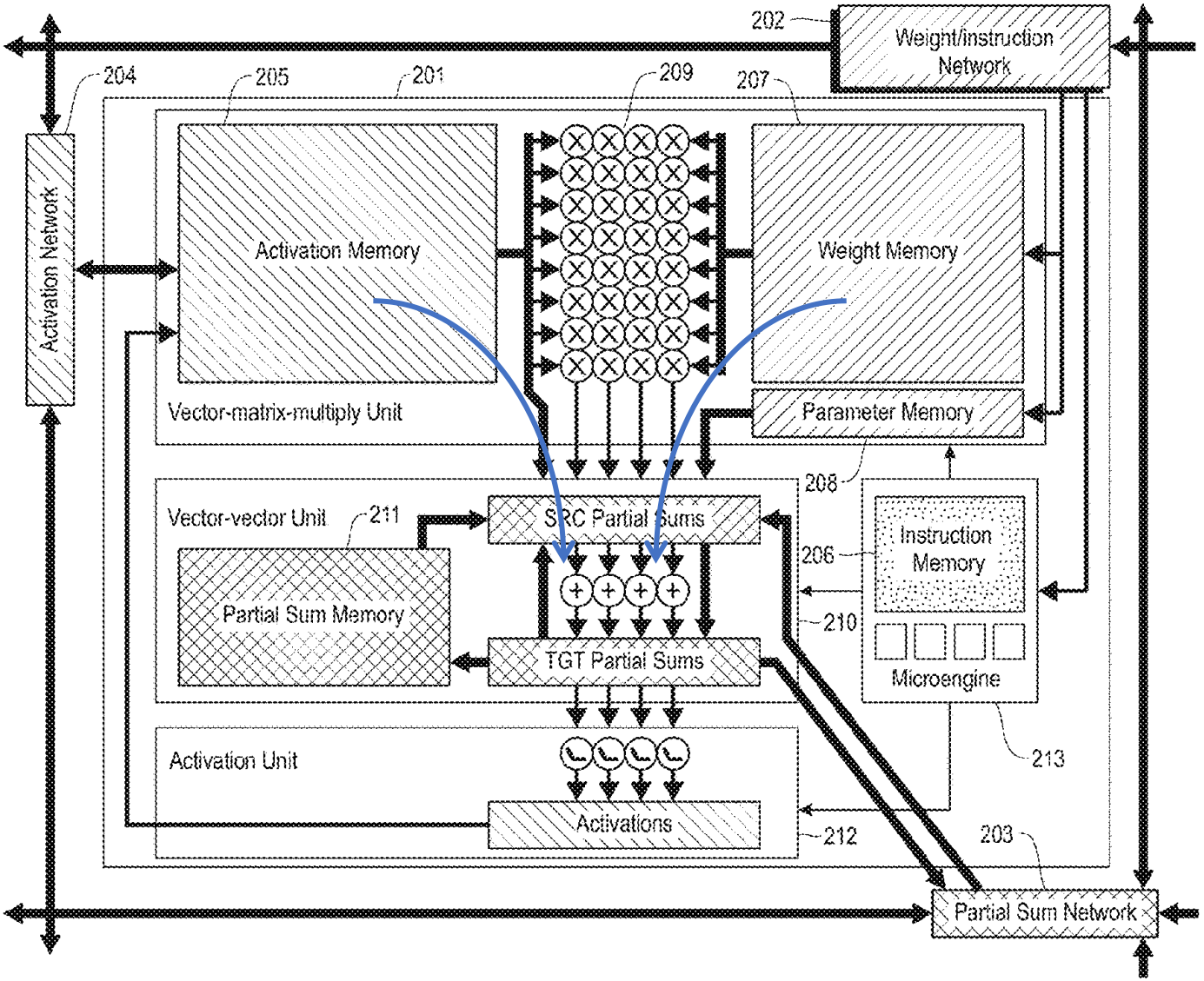


ニューロモーフィックの構成要素

ニューロモーフィックの構成要素

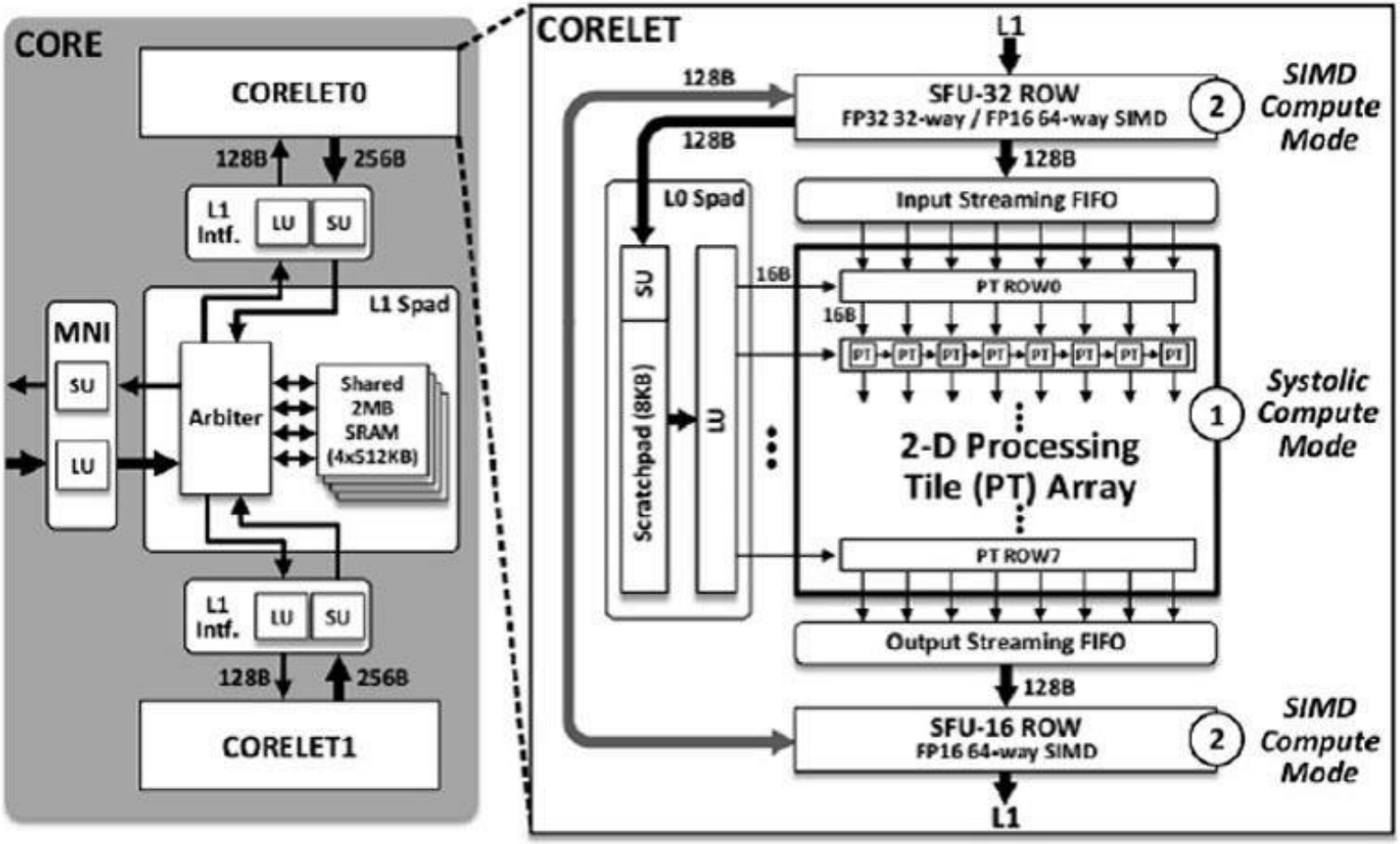
True Northコア

出典：https://electronics360.globalspec.com/article/4445/ibm-seeks-customers-for-neural-network-breakthrough



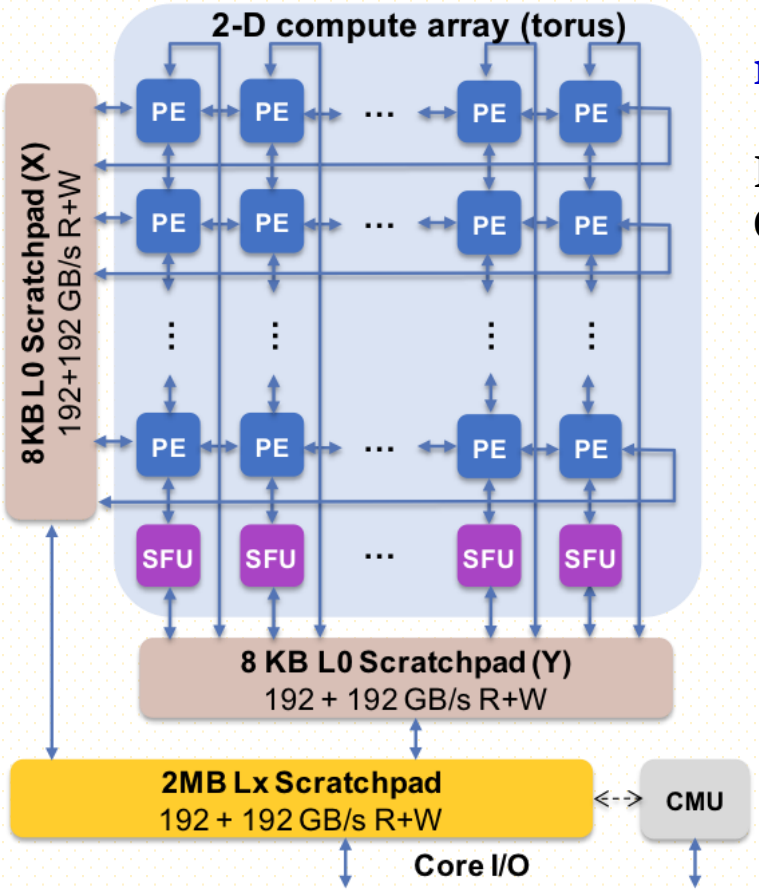
出典 ： https://www.ibm.com/blogs/research/2020/06/vlsi-2020-ibm-research-highlights/

Jinwook Oh, et al., “**A 3.0 TFLOPS 0.62V Scalable Processor Core for High Compute Utilization AI Training and Inference**”, in 2020 VLSI in Symposium on Technology



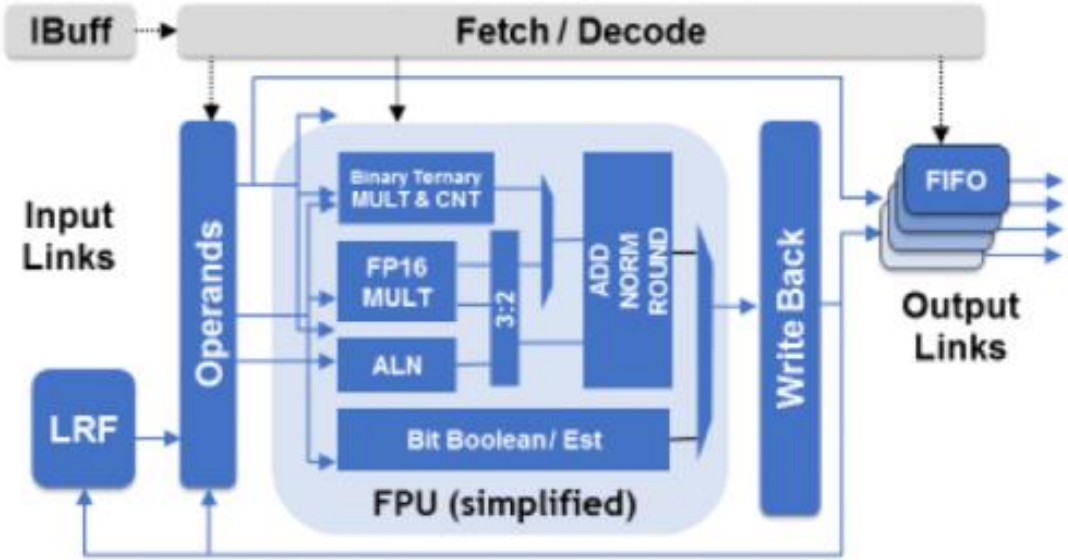
出典　https://www.ibm.com/blogs/research/2018/06/approximate-computing-ai-acceleration/

Bruce Fleischer, Sunil Shukla, et al. **A Scalable Multi-TeraOPS Deep Learning Processor Core for AI Training and Inference,** in 32nd IEEE Symposium on VLSI Circuits, 2018



・Pipeline of AI accelerator for matrix multiplication

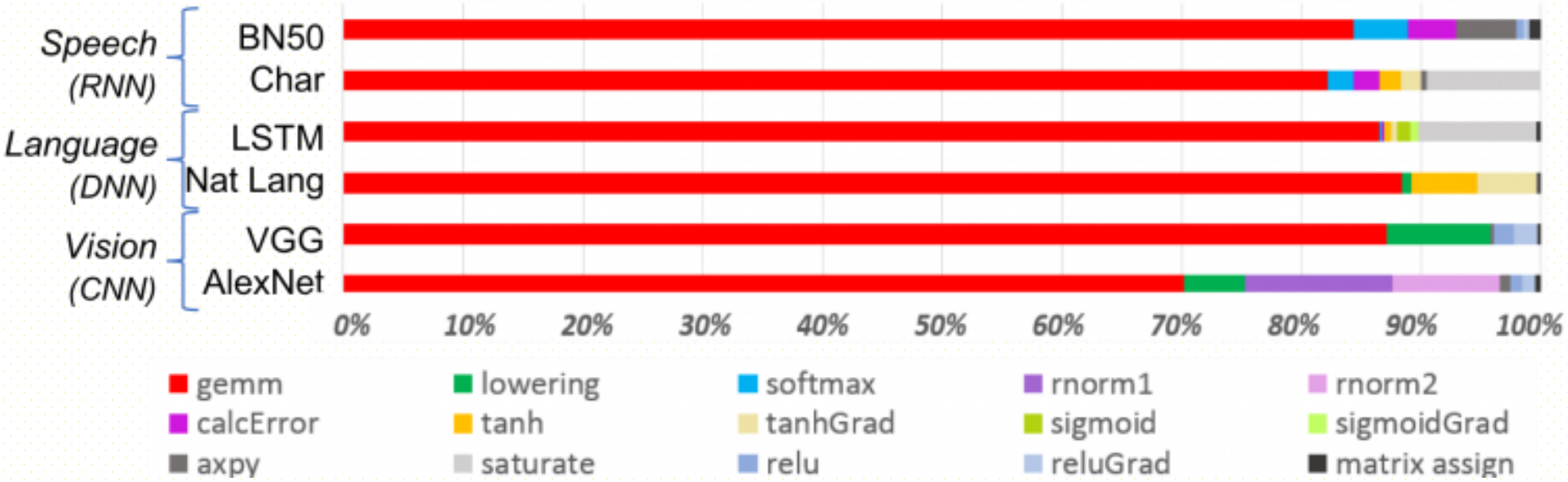
・Processing Element, 16 bit FPU, for Matrix Multiplication, Activation function, an Boolean Operations



出典　https://www.ibm.com/blogs/research/2018/06/approximate-computing-ai-acceleration/

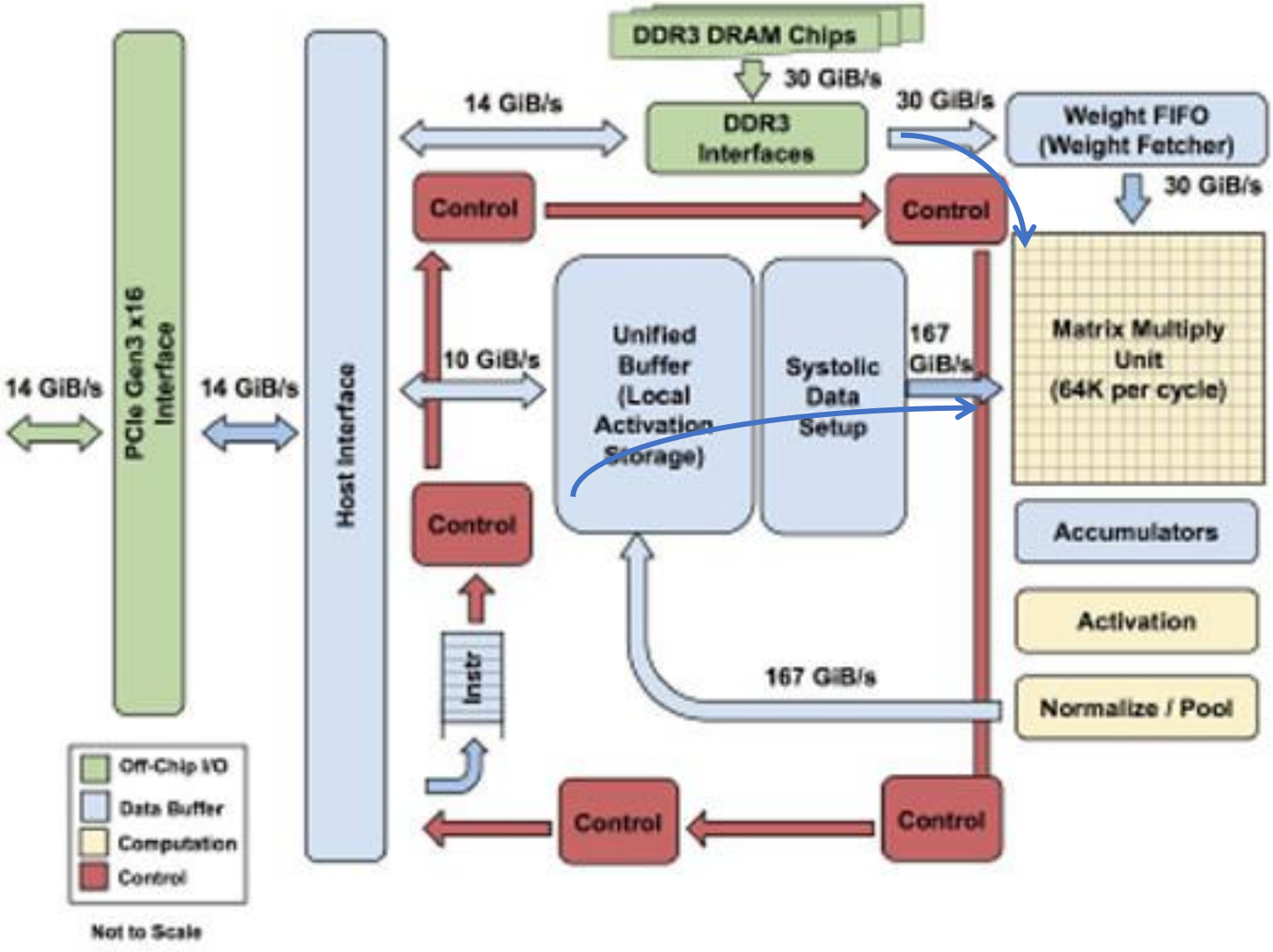
Variation of deep learning functions

・ Deep learning algorithms are comprised of dominant matrix multiplications, dominant, optimizing performance efficiency while maintaining accuracy requires the core architecture to efficiently support all of the auxiliary functions.

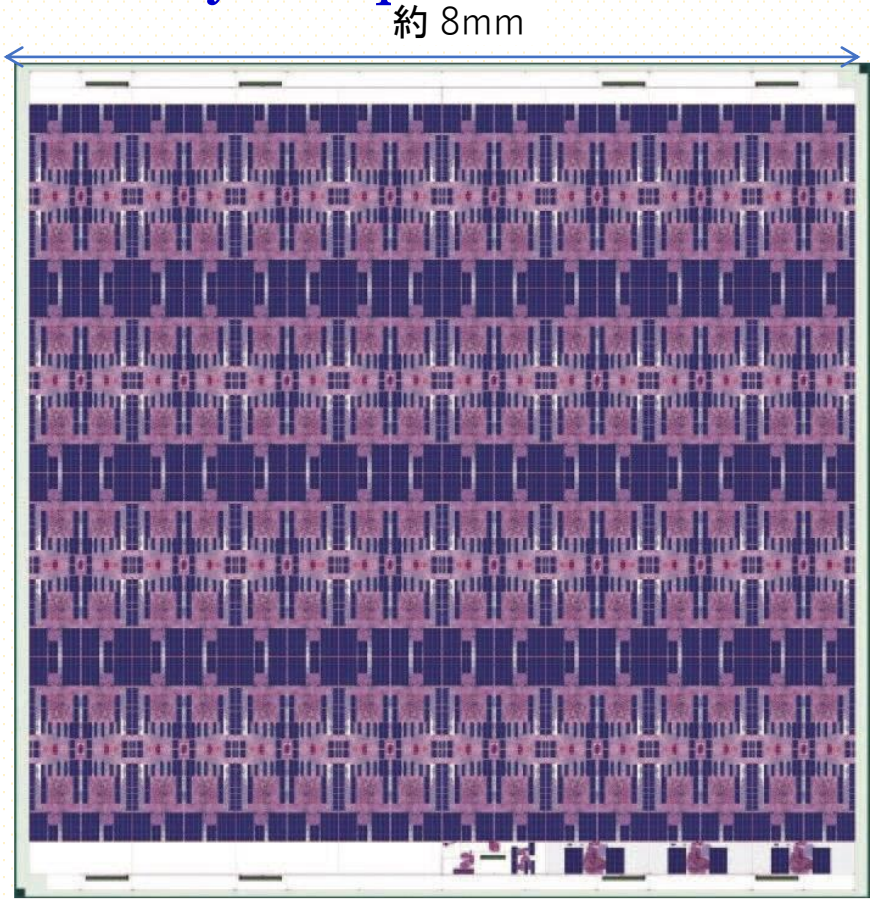


（出典）https://cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu

TPU, Google

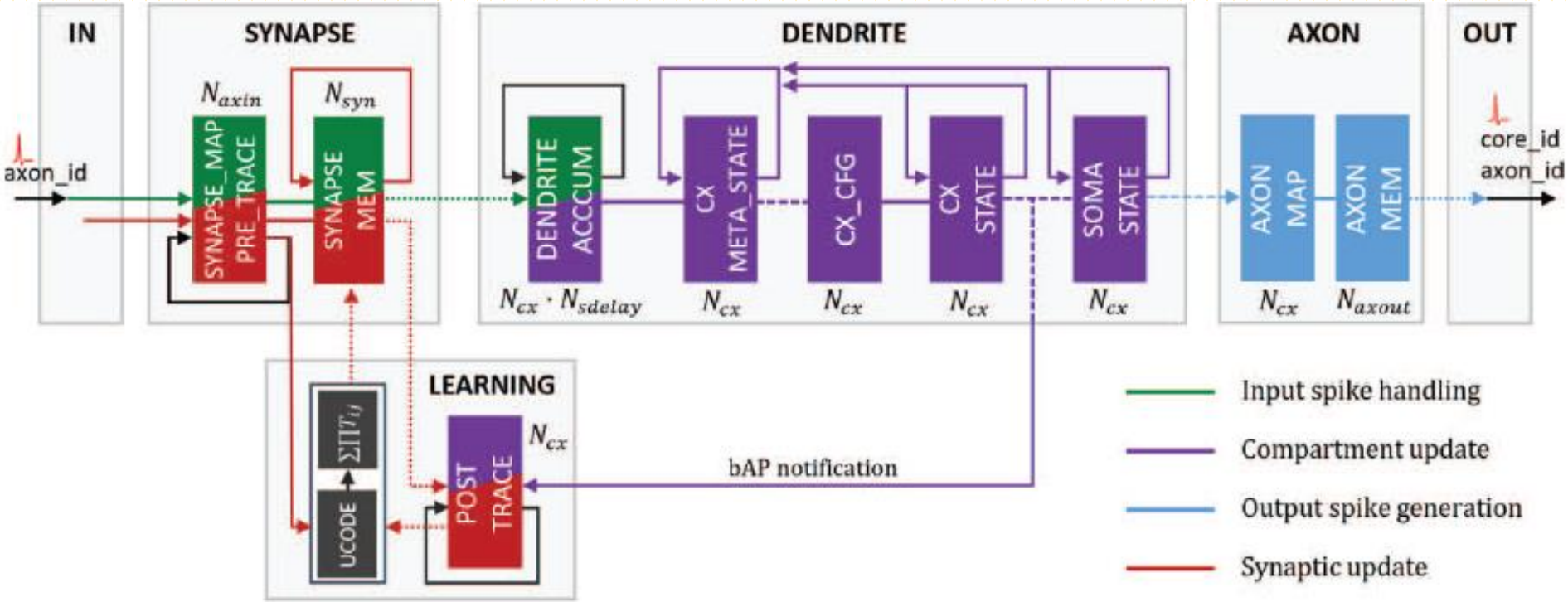


Davies, et al.; Loihi: A neuromorphic manycore processor with on-chip learning, in IEEE Micro, 2018.



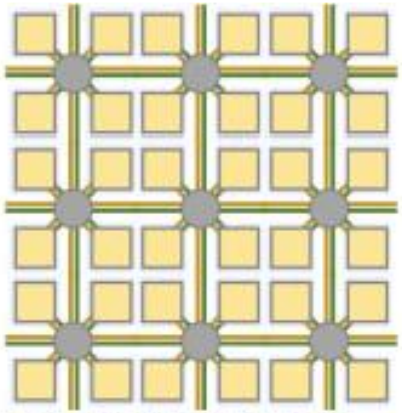
- Intel's 5th generation chip for neuromorphic
- 20.億トランジスタ＋33 MB SRAM
  (Synaps Memory = 128Byte /Neuron)
  (AXON memory 等 =128Byte/Neuron)
- 128 cores/chip × 1024 Neurons/ core
  = 128K Neurons/Chip
- Network Expansion ≦ 16K Chip
  （最大構成時のNetwork = 2G Neurons）
- Neuron演算速度 ≦ 8.4×256 ns？
  ≦ 2.1 μ sec (約400KHz)
- Programmable Synaptic Learning
  > Differential Hebbian Learning by measuring perturbations in spike patterns
  > Bienenstock-Cooper-Munro Learning using triplet STDP
  > Reinforcement Learning
  > Special Reward Spikes for Reward & Punishment.

Davies, et al.; Loihi: A neuromorphic manycore processor with on-chip learning, in IEEE Micro, 2018.

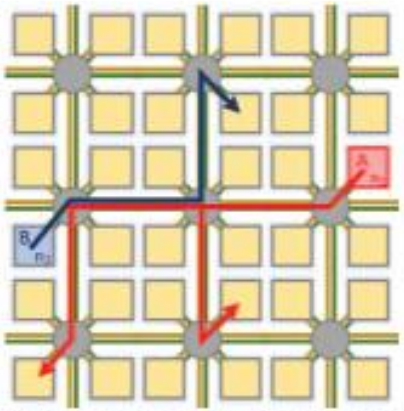


Davies, et al.; Loihi: A neuromorphic manycore processor with on-chip learning, in IEEE Micro, 2018.

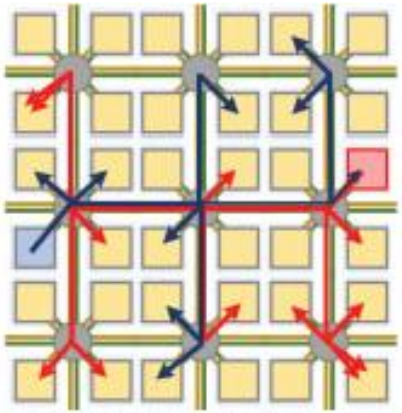
Flexible and well provisioned SNN connectivity features are crucial for supporting a broad range of workloads. Some desirable networks may call for dense, all-to-all connectivity while others may call for sparse connectivity; some may have uniform graph degree distributions, others power law distributions; some may require high precision synaptic weights, *e.g.* to support learning, while others can make do with binary connections. As a rule, algorithmic performance scales with increasing network size, measured not only by neuron counts but especially neuron-toneuron fanout degrees. We see this rule holding all the way to biological levels (1:10,000). Due to the $O(N2)$ scaling of connectivity state in the number of fanouts, it becomes an enormous challenge to support networks with high connectivity using today's integrated circuit technology. To address this challenge, Loihi supports a range of features to relax the sometimes severe constraints that other neuromorphic designs have imposed on the programmer:



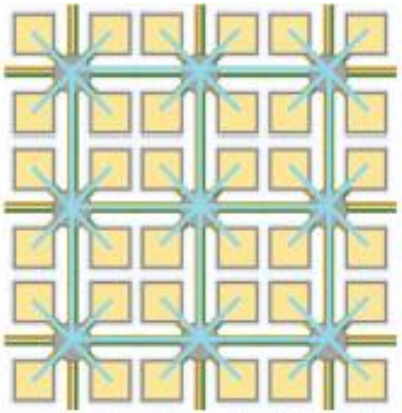
(a) Initial idle state for timestep t. Each square represents a core in the mesh containing multiple neurons



(b) Neurons n_1 and n_2 in cores A and B fire and generate spike messages



(c) Spikes from all other neurons firing on timestep t in cores A and B are distributed to their destination cores



(d) Each core advances its algorithmic timestep to t+1 as it handshakes with its neighbors via barrier synchronization messages

Brain ScaleS

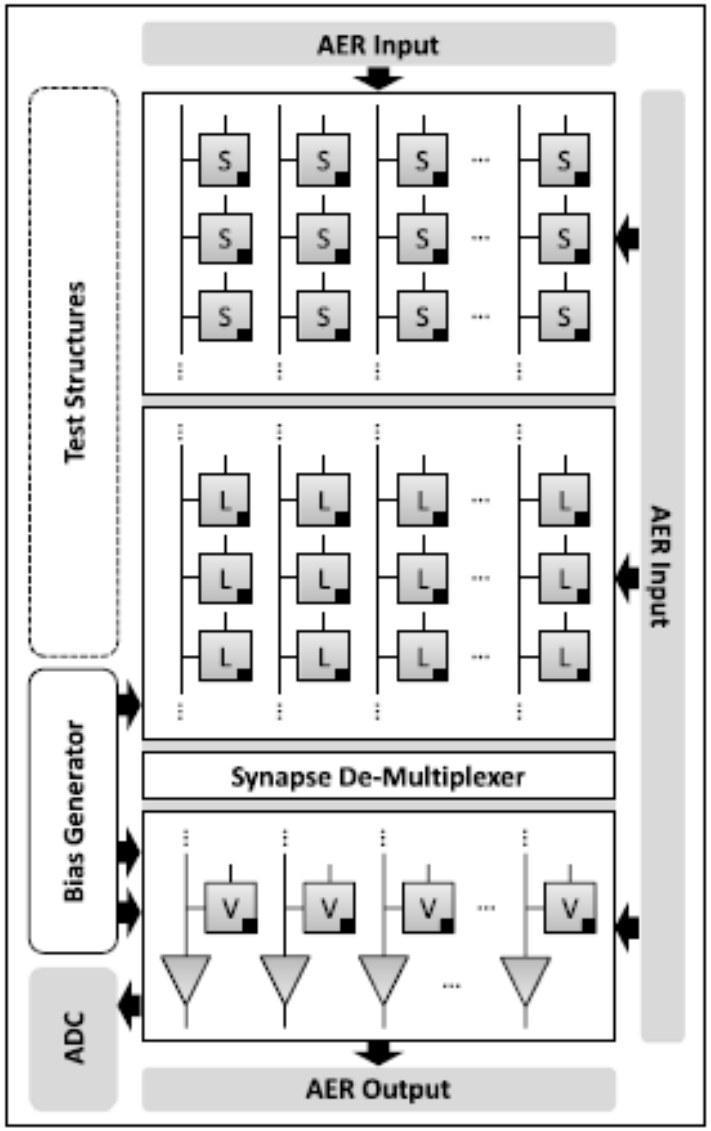
Fig. 6. Schematic overview over the ROLLS neuromorphic processor (Qiao et al., 2015).

The chip contains two 256 × 256 grids of synapses for short-term plasticity (STP) and long-term plasticity (LTP) (see Section 4.3).

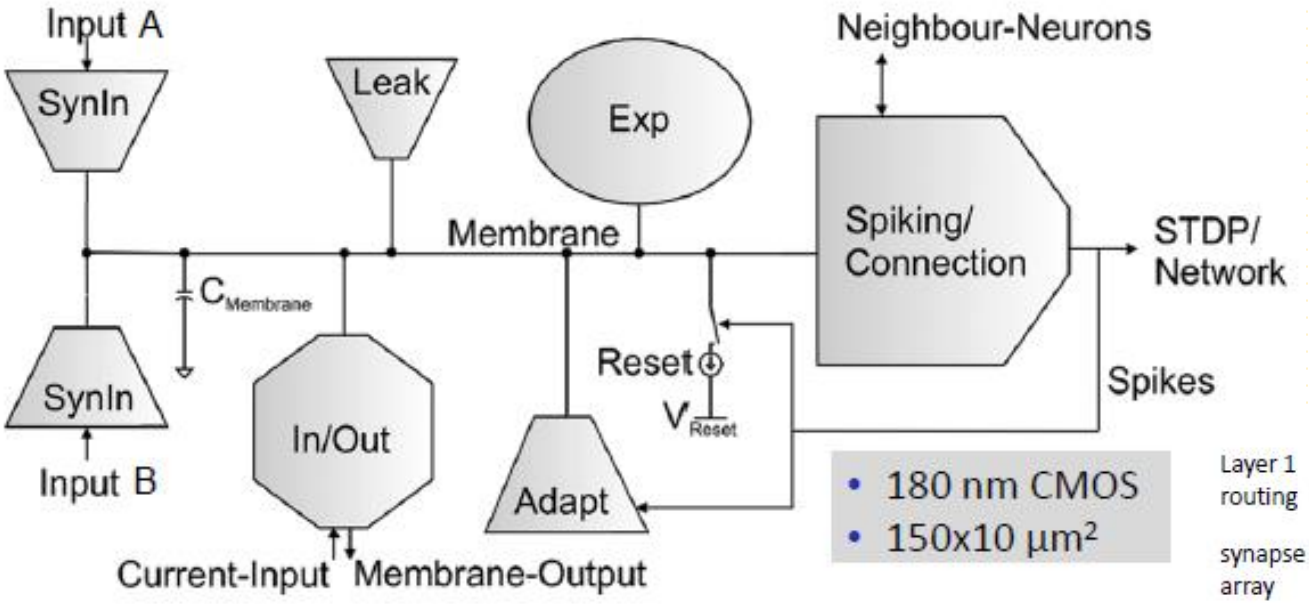
A synapse de-multiplexer can assign several rows of synapses to a single silicon neuron.

The additional virtual synapses above the neuron circuits can simulate background activity of the neural network. A bias generator stores global network parameters.

The analog digital converter (ADC) can be used to read analog state information from neural and synaptic circuits. The test structures are irrelevant for the neural simulation



Johannes Schemmel , et al., A Wafer-Scale Neuromorphic Hardware System for Large-Scale Neural Modeling, in 2010 IEEE International Symposium on Circuits and Systems (ISCAS)



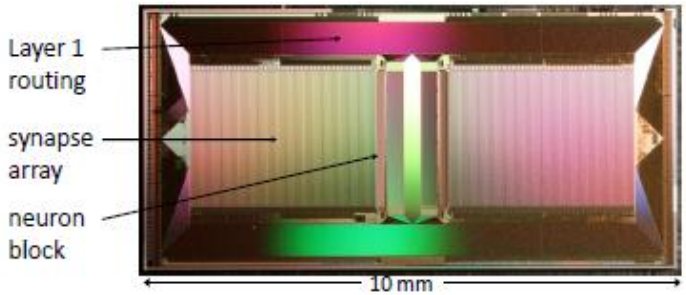
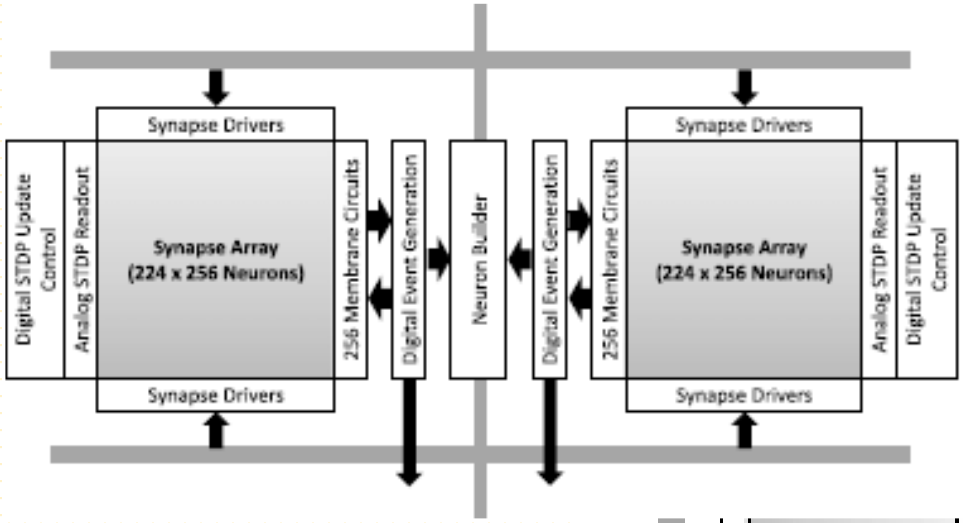
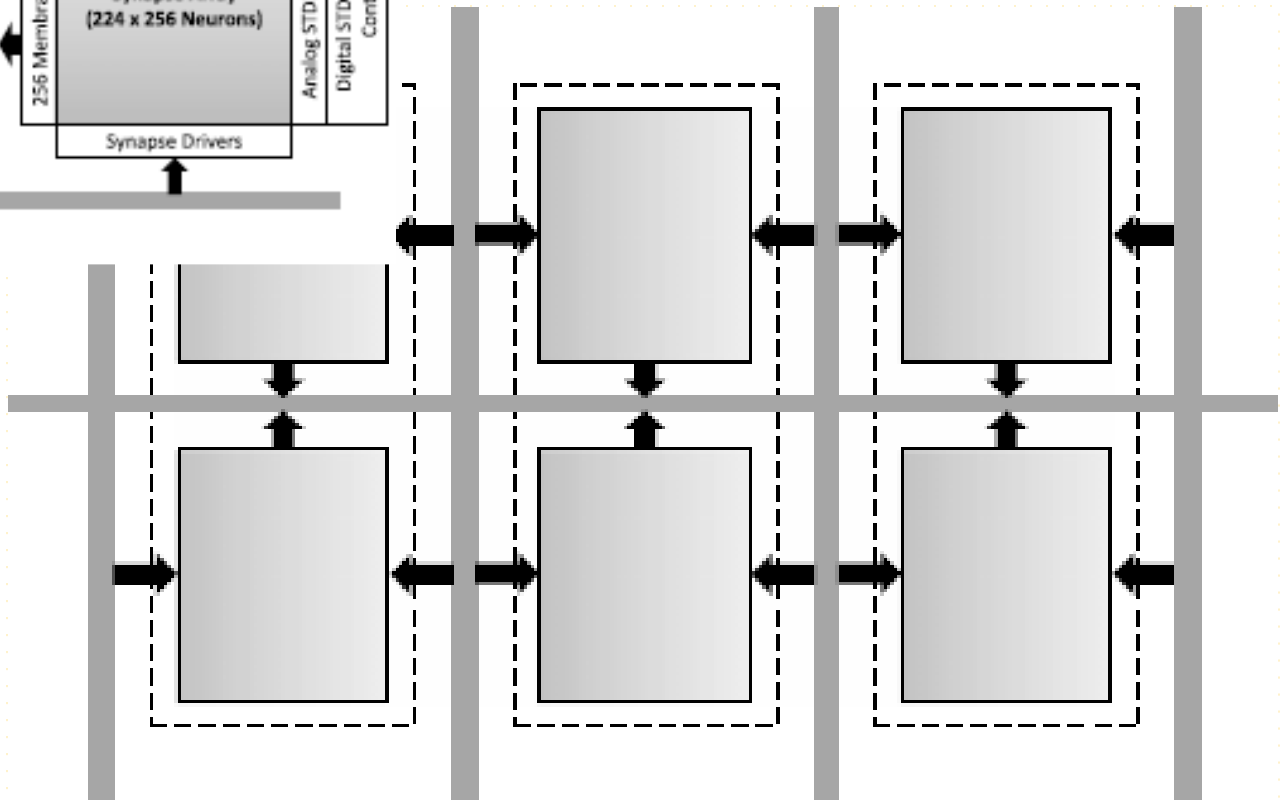


Fig. 7. Photograph of the HICANN die.


Synapse Drivers
Digital STDP Update Control
Analog STDP Readout
Synapse Array
(224 x 256 Neurons)
256 Membrane Circuits
Digital Event Generation
Neuron Builder
Digital Event Generation
256 Membrane Circuits
Synapse Array
(224 x 256 Neurons)
Analog STDP Readout
Digital STDP Update Control
Synapse Drivers
Synapse Drivers
Synapse Drivers

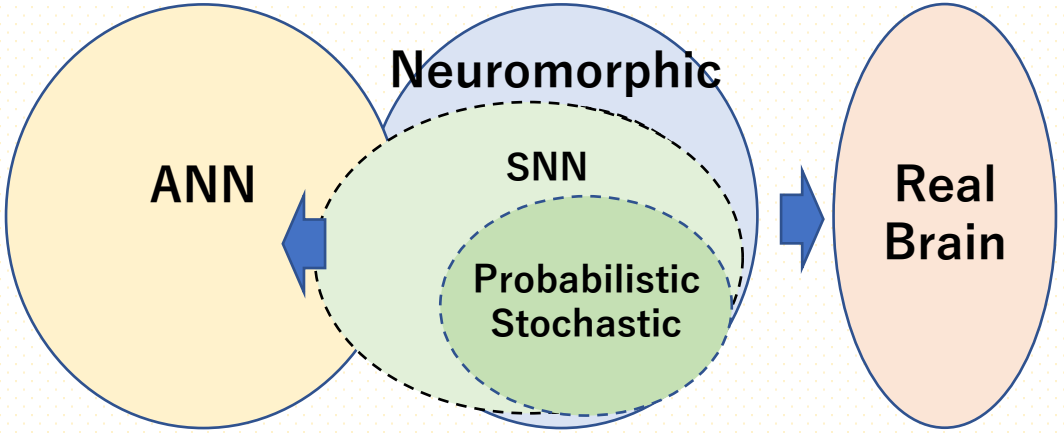


Lei Deng , et al., Rethinking the performance comparison between SNNs and ANNs, In Neural Networks 121 (2020)

ANNの成功要因

① Mature models
② Various benchmarks
③ Open-source datasets
④ Powerful computing platforms

⇒ SIG



SNNの状況

① Long-time ongoing debates
② Skepticisms about the value of SNNs in practical applications, except for the low power attribute benefit.
③ SNNs usually perform worse than ANNs especially in terms of the application accuracy.

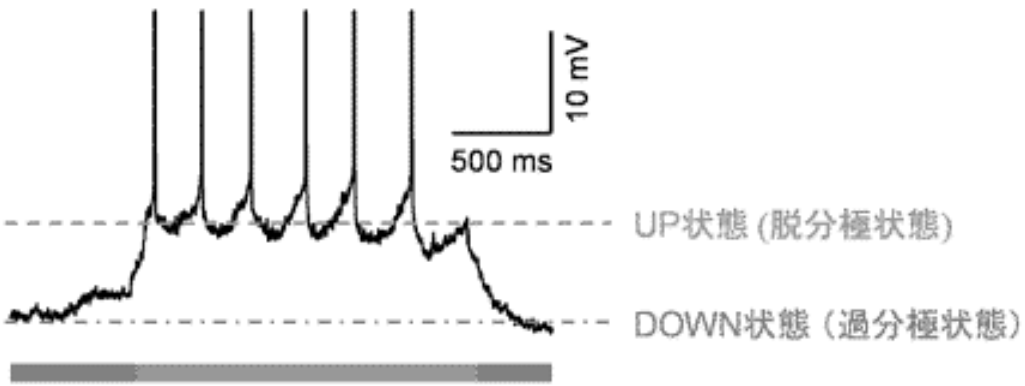
・ SNNにANNの学習方法(BP等)を実装し、ANN用のワークロードでトレーニングし、ANNベースの評価 ⇒ ハードウエアの進歩とネットワーク拡張が著しいANNに追いつけないだろう。

⇒ AGI

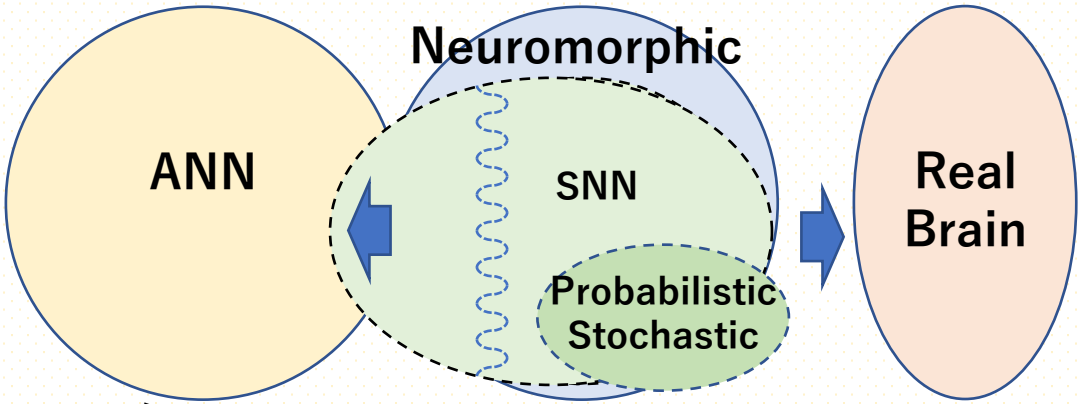
（図）発火の図は、池谷裕二氏のHPより掲載：http://gaya.jp/research/spontaneous_activity.htm

発火のコンピューティング原理

1．Spiking Neural Networkの
コンセプトの概要
- 発火モデル（Integrate & Fire）
- Signal ModelとEncoding
- Winner-Takes-All
（ニューロン母集団）
- 量子化と発火頻度と発火確率



2．Neuromorphic Processors
- コア回路
- ANNとの
ベンチマーク議論



3．考察
- Atom of Informationはニューロンか？
- 脳の動作の表現戦略における「粒度」について

Nassim Abderrahmanea, et, al., Design Space Exploration of Hardware Spiking Neurons for Embedded Artificial Intelligence, in Neural Networks Volume 121, January 2020.

・ FPGAへの実装にて比較

・ SNNs cost about 50% less in terms of hardware, while having approximately the same accuracy compared to ANNs. (Khacef, Abderrahmane, & Miramond, 2018)

・Mapping a traditional neural network to a spiking one does not severely impact the recognition rate, and results in more economical hardware.

(Diehl et al., 2015、Perez-Carrasco et al., 2013)

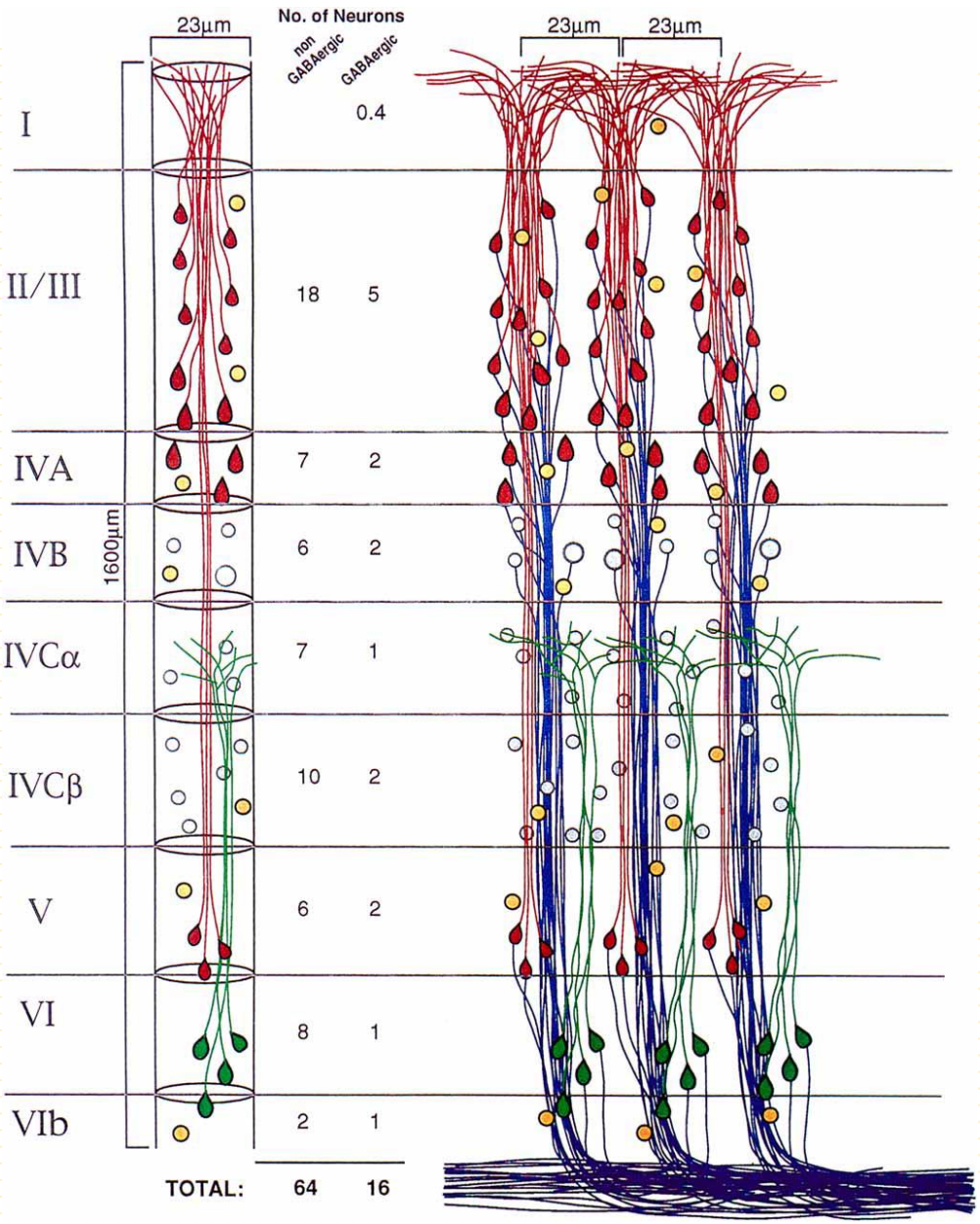
1) 同じNetwork-TopologyにANNでトレーニングを行い、そのシナプス係数をSNNに使用
（Back-Propagate Learningを行って、 SNNを Supervised Feed-Forward動作させる）

・Sze, V., Chen, Y., Yang, T., & Emer, J. S. (2017).; "Efficient processing of deep neural networks: A tutorial and surve.", Proceedings of the IEEE, 105(12), 2295–2329.

・Tavanaei, A., Ghodrati, M., Kheradpisheh, S. R., Masquelier, T., & Maida, A. (2019); "Deep learning in spiking neural networks.", Neural Networks, 111, 47–63.

２）SNNにて学習（SpikeProp or STDP）

Kheradpisheh et al., 2018、 Mostafa, 2018、Thiele et al., 2018
Mozafari, Ganjtabesh, Nowzari-Dalini, Thorpe, & Masquelier, 2018



[出典] A Peters, C Sethares; Myelinated axons and the pyramidal cell modules in monkey primary visual cortex, in ournal of Comparative Neurology, Volume 365, Issue 2 (1996)

Atom回路に関する問題 ： ミニコラム

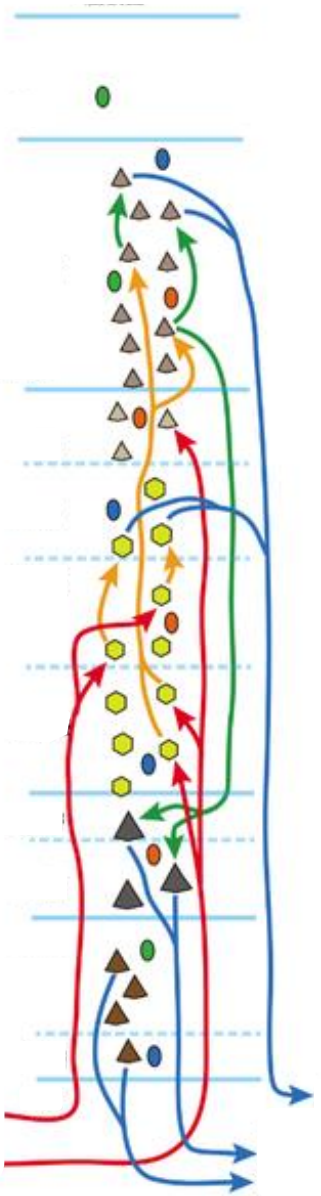
サルの大脳皮質の一次視覚野では、ひとつのミニ円柱構造の中に、18個の2/3層の錐体細胞、**6個の5層の錐体細胞**、10個の6層の錐体細胞の尖端樹状突起とその軸索（axon）が束をなし、上述の錐体細胞を含んだ64個の興奮性の神経細胞（錐体細胞、Spiny stellate 細胞等）と16個の抑制性の非錐体細胞がその周りを取り囲む様に分布し、それらが互いにシナプス結合で連絡し合い

大脳新皮質の神経細胞は、脳表面に沿った方向の数十μm程度までの範囲に存在する神経細胞と特に強く相互作用することが知られています。このため、このような範囲に存在する神経細胞群は互いに密な情報交換を行い、それらが作る回路は情報処理にとって重要な役割を果たしている。 （Micro Columnar Circuit）

[引用元] 細谷 俊彦；「高度な機能を司る大脳新皮質には、機能ごとの小規模な構造単位が存在－大脳新皮質を小規模な回路の繰り返しとして解析する道を開く－」。
2011年12月14日、独立行政法人理化学研究所

大脳皮質には、80個程度の神経細胞で構成されるミニ円柱構造と呼ばれる最小単位の局所神経回路があり、その単位神経回路が多数並列的に存在する事で、大脳皮質の神経回路が作られている。

[引用] 窪田芳之
大脳皮質の神経細胞と局所神経回路 ；
日本神経回路学会誌, 2014 - jstage.jst.go.jp
https://www.jstage.jst.go.jp/article/jnns/21/3/21_122/_article/-char/ja/



ミニコラムに関する仮説

- 1本のミニコラムの応答には、そのミニ・コラムが属するローカル・ネットワーク内の複数段のニューロンが行う確率プロセスがが関与している。

- ミニ・コラムの応答は、「興奮時の入力刺激量を記憶したTableの曖昧検索」と似る。

- 大脳皮質のL1 Networkは、閉じているベイジアン・ネットワークに似るが、そのベイ ジアン・ネットワークは、完全には閉じておらず、L4/L6から入力される外部情報や、脳の中心部分（大脳辺縁系、視床、等）との通信・応答（Dialogue）に干渉され、様々に、興奮・抑制の制御を受けている。

＜取り組み方針＞

- 1本のミニ・コラムは、複数のポートからのアクセスを受けていおり、実際の動作を単一のアルゴリズムで表現することは困難。 ハードウエアによる模倣をPaper－Workする。

- 価値ある商用装置を構想するのは、それ自体が大きな課題
（脳の研究と、価値ある商用装置の構想は、互いに独立させて進めるべき）

- 脳の研究、数理研究、価値ある商用装置研究（コンピュータ化）のコミュニティが必要。

“Reconstruction and Simulation of Neocortical Microcircuitry”

by Henry Markram, et al., in 2015 cell 163, 456-492, October 8, 2015.

生体と電子回路の信号伝搬速度比較

1） 配線を伝搬中の信号伝送速度

参考）清水 崇弘 、池中 一裕、
Web上の「脳科学辞典(https://bsd.neuroinf.jp/wiki/)」の「有髄線維」の項より

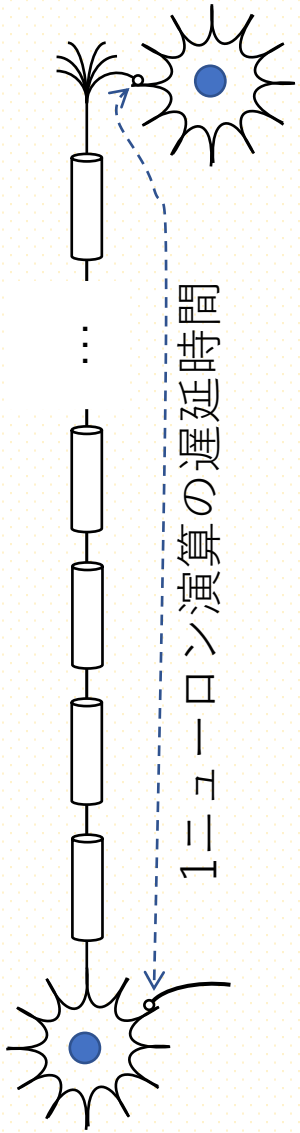
	径（μm）	伝導速度(cm/ms)	10cm伝送に要する時間
生体内（有髄神経）	12～20	7.2～12.0	0.8～1.4 ms
	6～12	3.6～7.2	1.4～2.8 ms
	1～6	0.4～3.6	2.8 ～ 25 ms
	≦3	0.3～1.5	6.6 ～ 33 ms
伝送線路(Micro Strip Line)	500	$15～20\times10^{6}$	0.5 ～ 0.7 ns
チップ内配線	0.2	$1～10\times10^{5}$	10～100 ns

2） ルータ回路の通過時間　～ １０数ns程度

- 生体内では不要
- 電子回路としては、３～５段程度、ルーターを経由する必要あり。

3） チップのI/O回路の通過時間 ～ ルータ通過由回数の２倍程度

ニューロン動作の遅延時間モデル



1ニューロン当たりの総遅延時間

＝ 入力信号のＥｎｃｏｄｅ時間

＋ 積算演算

＋ 和算演算

＋ 発火判定（＆リセット）

＋ 出力先アドレス読出し

＋ 出力パケット生成

＋ ローカル・ネットワーク

＋ ルータ

＋ グローバル・ネット

Ｄｅｎｄｒｉｔｅ＋Ｓｏｍａの遅延相当

生体： 十ｍｓ以上

ＳＮＮ： 数μｓ以下

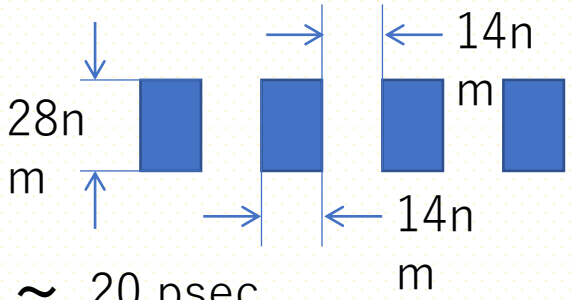
ＡＸＯＮ遅延相当

生体：数ｍｓ～数十ｍｓ

ＳＮＮ： １μｓ以下

電子回路内の信号伝達速度

1）**集積回路**



Relative Dielectric Constance : 2.2

C = 0.13 fF/um、 R = 2.6 Ohm/um

RC Delay ～ 50 fsec/um2、 Circuit Delay ～ 20 psec

とすると、10cmの信号伝達の理論最小値は、0.2 μ sec。

現実的には、Fan-Outを考慮する必要があり、0.5 ～ 1 μ sec 。

[参照] Ivan Ciofi, et al, “Impact of Wire Geometry on Interconnect RC and Circuit Delay”, Article *in* IEEE Transactions on Electron Devices · June 2016

2) Micro Strip Line （**ボード上**） :

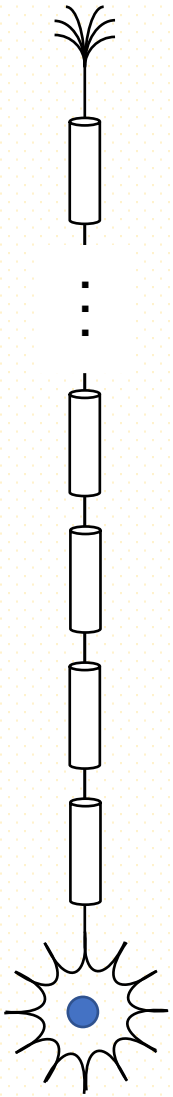
配線幅（W)＝0.2mm、**絶縁層**（H)=0.2mm、**誘電率**(ε) ＝4.3 **の時、**

$$\varepsilon_{eff} = \frac{(\varepsilon+1)}{2} + \frac{(\varepsilon-1)}{2} \times \left(1 + \frac{10H}{W}\right) = 3.4$$

遅延＝ $\frac{\sqrt{\varepsilon_{eff}}}{(3\times10^8)} = 6.15\ {}^{ns}/_{m}$

であり、10cm**伝導するのに、約**0.6 nsec。

神経細胞の活動電位の伝播速度



1) **有髄線維の跳躍伝導(saltatory conduction)の信号伝送速度**

・**直径15μmの有髄神経繊維は約25m/sの伝導速度**

（これは、10cm伝導するのに 4msec ）

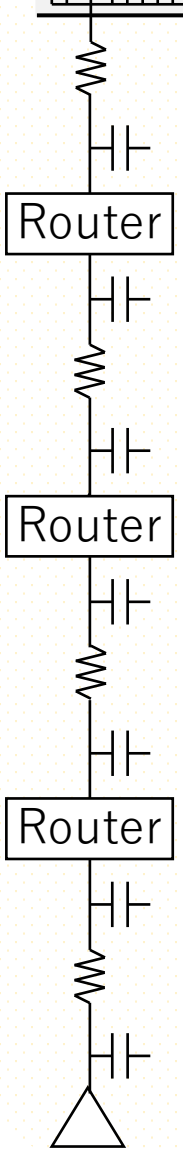
[出典] 「**有髄線維**」、**脳科学辞典**、2013年8月21日更新、
https://bsd.neuroinf.jp/wiki/%E6%9C%89%E9%AB%84%E7%B7%9A%E7%B6%AD

2) **神経線維の太さによる分類**（Erlanger and Gasser）

分類	直径	スピード	備考
Aα	15μ	100m/s	骨格筋運動線維、筋紡錘求心線維
Aβ	8μ	50m/s	皮膚触覚、皮膚圧覚
Aγ	5μ	20m/s	筋紡錘運動線維
Aδ	3μ	15m/s	皮膚温度感覚、皮膚痛覚
B	3μ	7m/s	交感神経節前線維
C	0.5μ	1m/s	皮膚痛覚、交感神経

[出典] H. S. Gasser, and Joseph Erlanger
, “**A STUDY OF THE ACTION CURRENTS OF NERVE WITH THE CATHODE RAY OSCILLOGRAPH**01 Nov 1922, ”American Journal of Physiology 62, 496—524, 1922），
https://doi.org/10.1152/ajplegacy.1922.62.3.496

機能ブロック：約６０個、　コラム　：約９０万、　マイクロカラム（機能単位）　：２～２．５億個



・電子回路にFlexibilityを持たせるには、神経回路間の配線を固定することはできないので、「Router
のように接続を学習する回路」を挿入する必要
（一般的な既存のRouter回路は、Hash-Tableを持ち、検索エンジン、または、連想メモリによって、
内部に搭載するTash-Tableを検索して、信号の送り先を決定する）

・Router回路は、行き先となりうる140億個の全神経細胞にアドレスを振り、そのアドレスをHash-Table
に登録する必要ある。　その場、　最低でも34ビットアドレスが必要

＞　神経細胞の総個数（140億Entry必要）や、マイクロカラム（1億Ｅｎｔｒｙ必要）を１度に検索しようとすると、
Hash-Tableは大きくなり過ぎる。　必ずしも、製造技術の微細化限界がネックではない。
=>　**Valid（Fullに確定的）な「1億分の１検索」は、古典力学的には非現実的。（エネルギーが膨大）**
検索精度を落とすことで、動作時の消費エネルギーを下げ、検索Tableの規模を大きくする仕組
みが必要だろう。　「この点が、技術Roadmapを考える上での要点の一つ」と思われる。

＞　物理アドレスは１億個分必要としても、「因果関係が独立なアドレスの対応関係」はもっと少ないので
はないか？　=>　**Ｒｏｕｔｅｒには、ＭｕｌｔｉＣａｓｔ機能が必要。**　ＭｕｌｔｉＣａｓｔ数の最大値を決める必要ある。

＞　Hash-Tableの検索はＶａｌｉｄなプロセスにはならない。　入力される信号パターンのアナログ的な類似
性を元に、最適な行き先を選び出す機能が必要　=>　そのためには　**情報のタグ付け**

2．1　Algorithmic Outlook ： Spiking neural networks

・ Maass, W.[1997年]の”Networks of spiking neurons: the third generation of neural network models.”は、SNNを第三世代Neural Network技術と初めて位置付けた。
　＞第１世代： McCulloch–Pitt perceptrons
　＞第２世代： Nonlinearity upgrade (微分可能な活性化関数の導入) ＆DLL化＆ BP
　＞第３世代： Spiking Neurons (integrate-and-fire、時間方向にSpikeカウントを蓄積)

・第２世代と第３世代の最大の違いは、
　＞第２世代：real-valued computation (say, the amplitude of the signal),
　＞ ：using timing of the signals (or the spikes) to process information.

[岡島注]　従来のComputingのデータは、数値Codeであり、論理層(アルゴリズム)上の概念。物理世界の信号とは関係を持たない。データはの値１と値０は、基本的には対称であり、それ自体では意味を持たない。 Spike信号は、物理層(ハードウエア)上の概念であり、
- 1本の信号で、**タイミング**と**強度情報**と**素情報のStatus**を3情報を伝送する。
- 値1と値0が各々の別意味を持ち、値1は「活性状態にある」、値0は「非活性状態にある」と、**そのニューロンが担う素情報のStatusとエネルギー供給(電源ON、電源OFF)を意味しうる**。
- **本質的にSparse Codeである**（活性化状態の素情報は、そのニューロンが属するドメインの中で非常にマイナーな存在であるため。また、そのために本質的に低消費電力である）
- そのニューロンが属するドメイン中の各ニューロンのパルスをカウントすることで、**各ニューロンが担う素情報の活性化度（もしくは、素情報の活性化確率）を相対的に意味しうる**。
- AND論理を取ると積算、OR論理を取ると和算となり、**積和演算を取るのが簡易**。
- ドメインの中で、最初に閾値以上に発火したニューロンを**Winnerと扱える**。
- Spikeの反復に、**差分検出の意味**を持たせることができる可能性がある。

と、**信号それ自体が意味を担うで、本質的にGroundingされている**。

Spiking Signal方式とは

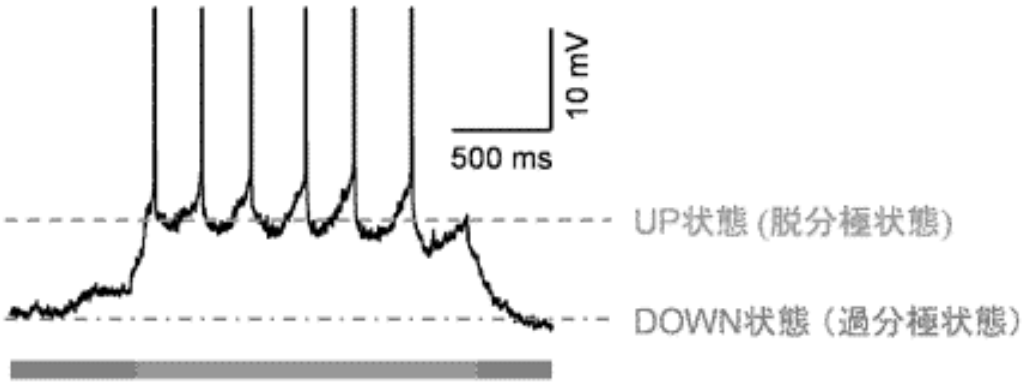
■ 神経回路中の発火現象

・信号伝送方式としては

1) 電圧レベル ・・・ LSI内部（伝送路）
2) AM/FM/PM
3) パルス変調

etc.

（図）池谷裕二氏のHPより掲載
http://gaya.jp/research/spontaneous_activity.htm



■ 回路テクニックとしてのSpiking 方式 ： 積和演算とWTAの表現が容易

・ パルス個数による「数」を伝送

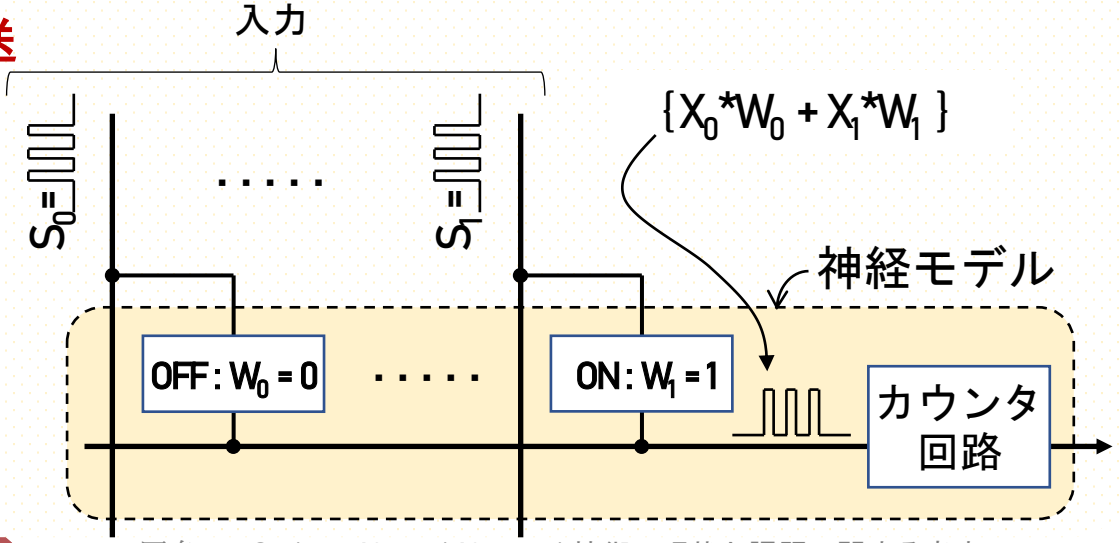
＞ 刺激の強度 / 発火頻度（発火のバースト回数）

＞ 発火確率

・ 「タイミング」を伝送

＞ 発火タイミング

（伝送時間の考慮必要）



Liquid State

by W. Maass, T. Natschlaeger, and H. Markram; "Real-time computing without stable states: A new framework for neural computation based on perturbations.", in Neural Computation, 14(11):2531-2560, 2002.

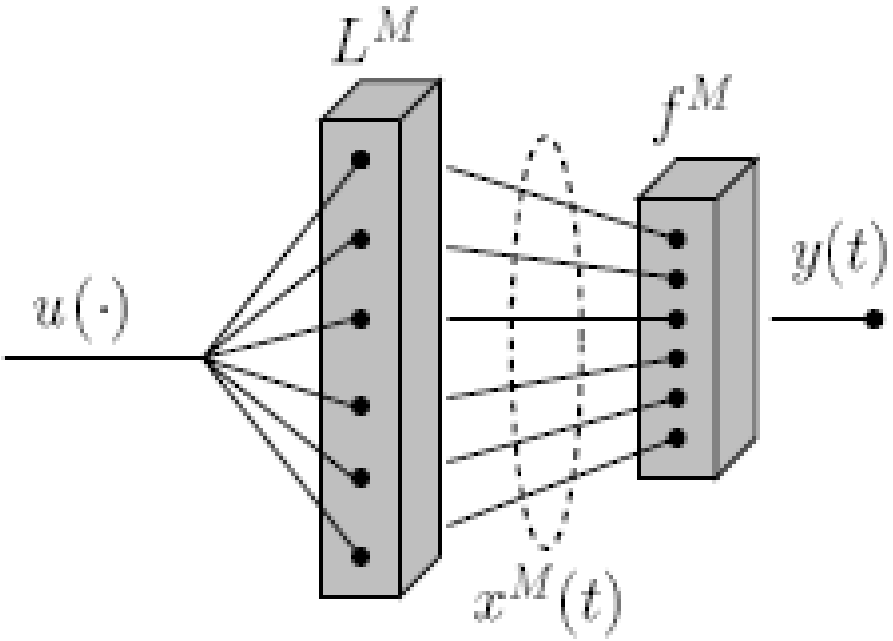


Figure 1: Architecture of an LSM.

A function of time (time series) $u(\cdot)$ is injected as input into the liquid filter L^M, creating at time t the *liquid state* $x^M(t)$, which is transformed by a memoryless readout map f^M to generate an output $y(t)$.

Turing Machine と Liquid State machine (LSM)

■ **Turing machine (a Finite State Machine.)** ： PDP / Von-Neuman type Computing
for off-line computation　　プログラム・フロー型（データ蓄積型）

- データ/パラメータとプログラムをメモリに蓄積し、カウンタを更新して、演算器との間を往復し、新しいデータを生成(generate)する。
- 特に定義しない限り、データやパラメータを構成する２進数（”０”と”１”）には、本質的な意味はない。（データには意味が在り得るが、信号には意味はない）
- プログラムをRealtimeに更新することは困難　⇒　　オフライン計算用

■ **Liquid State Machine (LSM)**　　： Asynchronous Spiking Signal Architecture
for real-time computing　　データ・フロー型（設定値をRealtimeに更新）

- プログラムやパラメータは演算器の状態(性格)を定義する設定値
- 演算器に外部からの信号を送り、演算器内の状態(性格)を定義する設定値を更新（On-Situ Trainingを実行するのが本来の姿）
- 基本的には外部からの情報は廃棄。 演算器が生成(Encode)したコードを出力する。
- 信号やパラメータ（通常は2値又は3値を取る）の”０”には「非活性化状態、又は、定常状態」との意味がある。(”1(-1)”には「活性化状態、又は、過渡状態」の意）
- 演算器にはデータの概念はない。但し、信号の”０”と”１(-1)”の間は非対称であり、演算器の属する集合全体からの出力（スパースなコード）の一部となり得る。

ミニコラムは、“The Atom of Information(素情報)”の接地回路か？

- **ミニコラムを、入力パターンを解釈し、QualiaをGroundingする装置と仮説設定して、当座は考えたい。**
- **ミニコラムのネットワークによる「教師無し学習」により、新しい素情報（新語、新概念に相当）を生み出す方法を考えたい。**
  - ＞ ミニコラムは、類似のパターン入力があると、発火し、記憶を強化する。
  - ＞「素情報間のグラフ（関連付け）を読み取ることにより、論理を抽出する」方法を考えたい。
- **大脳皮質の L1 Networkを閉じたベイジアン・ネットワークとして、無意識を起こしたい。**
  - ＞ 但し、L1 Networkは完全には閉じておらず、L4/L6から入力される外部情報や、大脳辺縁系、視床、等との 通信・応答に干渉され、興奮・抑制の制御を受け、状況認識を刻々と無意識に似ると思います。
  - ＞ その無意識と人間が対話するインターフェースが重要と思っています。

＜必要条件＞

- **その装置のソフトウエアの在り方を議論するコミュニティが必要**
  （必要な関数に関する議論、等）
- **価値ある商用装置の構想には、先ず、脳以上に高速で無意識する電子回路が必要**
  （価値ある商用装置の構想はそれ自体が大きな課題であり、いずれ、電子回路研究から独立させたい。）

Neural Computation as Perturbations

By Wolfgang Maass, Thomas Natschlager, Henry Markram
"Real-Time Computing Without Stable States: A New Framework for Neural Computation Based on Perturbations", in *Neural Computation* 14, 2531–2560 (2002), MIT.

Dynamic Synapse

Juan A. Varela, et al., "A Quantitative Description of Short-Term Plasticity at Excitatory Synapses in Layer 2/3 of Rat Primary Visual Cortex", in The Journal of Neuroscience, October 15, 1997, *17*(20):7926–7940.

・直前のスパイク信号列によるシナプス係数のFeed-forwardな変動は、次のスパイク信号列での予測精度を高める役割を持つ。（特に、V1系含めたSensory Response に関して）

・Short-termの変動には、3種類の現象が複合している。
① EPSC (Excitatory Post-Synaptic Current）の促進（Facilitation）
② EPSCの減衰(その１） ：数百ミリ秒の時定数で指数的に減衰する
③ EPSCの減衰 (その２） ：数秒の時定数で指数的に減衰する
⇒ EPSCの変動は、刺激に対するDamperのように作用している。

・スパイク信号列によるEPSCの促進現象は加算的に起こるが、減衰現象は、乗算的に起こる。（ANNのシナプスには減衰が無い）

Juan Alberto Varela

Fraction Leaky Integrate & Fire Model

Wondimu Teka, et al.; "Neuronal Spike Timing Adaptation Described with a Fractional Leaky Integrate-and-Fire Model", PLOS Computational Biology, March 1, 2014, Volume 10, Issue 3

生体と電子回路の演算時間比較

１） ニューロン演算に要する時間＃１ ： Loihi

- １チップ当たりのコア数 ： １２８コア
- １コア当たりのニューロン演算器数 ： １０２４ニューロン
- チップ面積 ： ６０mm2
-

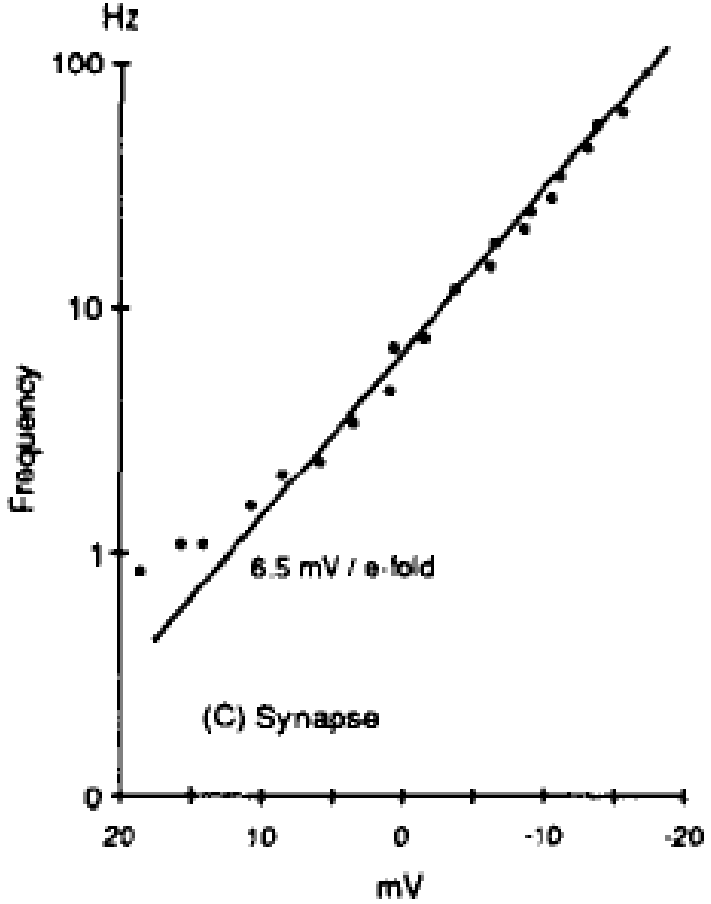
２） ニューロン演算に要する時間＃２ ： Loihi

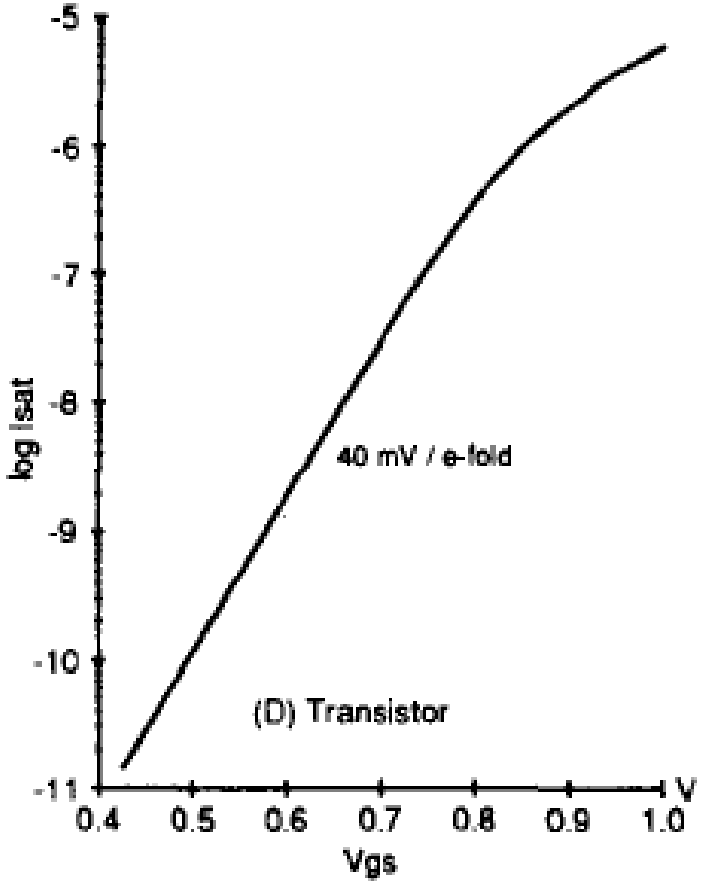
Paul A. Merolla, et al., “A Million Spiking-Neuron Integrated Circuit with a Scalable Communication Network and Interface”, in Science, vol. 345, no. 6197, pp. 668–673

Davies, et al.; Loihi: A neuromorphic manycore processor with on-chip learning, in IEEE Micro, 2018.

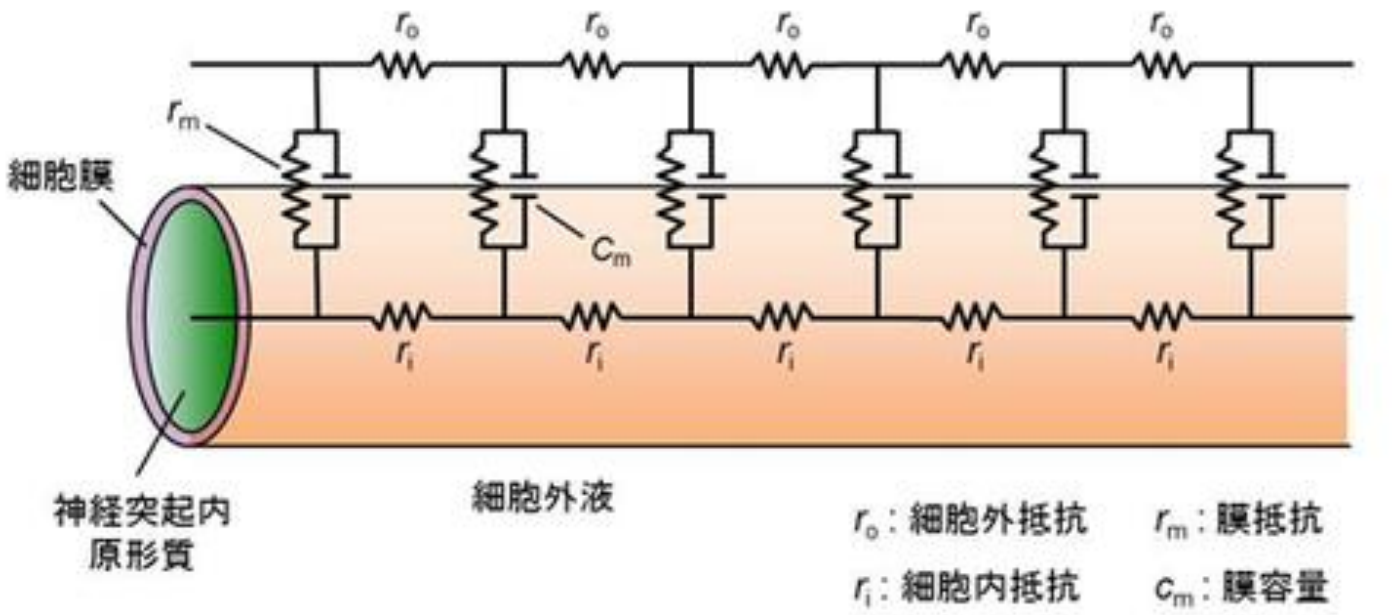
素子としての違い

・シリコン・トランジスタの電流増幅率　～　60mV/div
・シナプス電流　～　８．５mV/div





$$V = V_0 \, e^{-t/\tau}$$



引用）山崎良彦　藤井聡、
Web上の「脳科学辞典(https://bsd.neuroinf.jp/wiki/)」の「**ケーブル理論**（*山形大学医学部生理学講座*）」の項より

Henry Markram

Professor of Neuroscience, Swiss Federal Institute of Technology (EPFL), Lausanne.

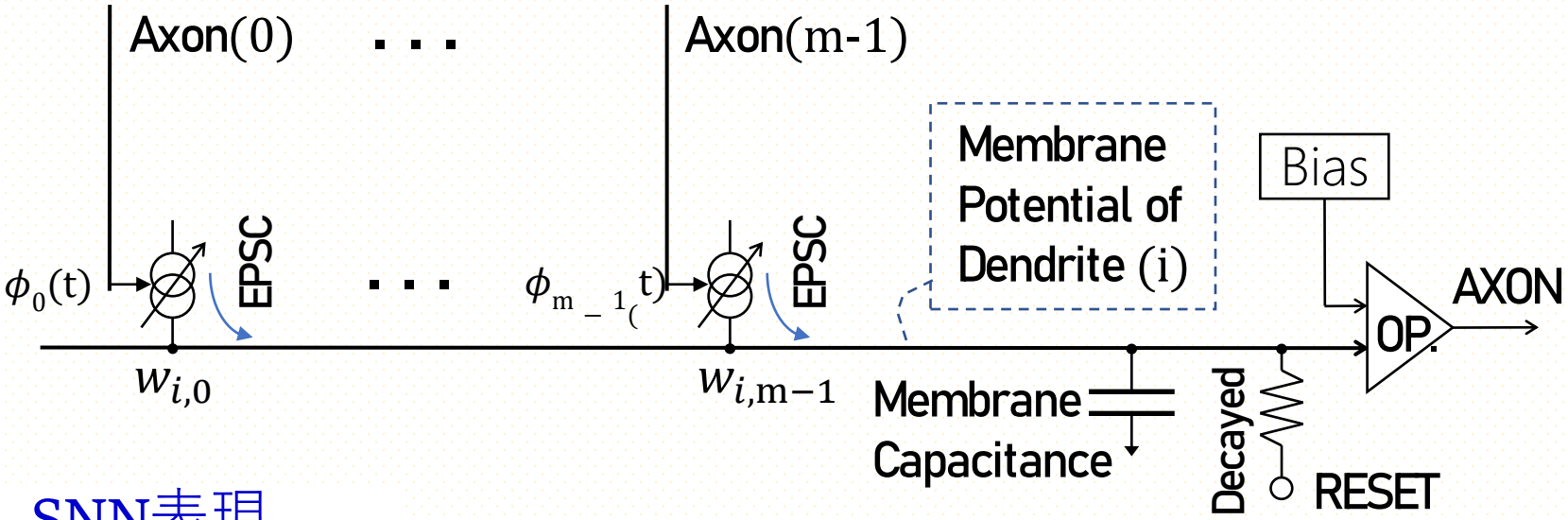
出典：World Economic Forum (https://jp.weforum.org/people/henry-markram)

- **Founder, Brain Mind Institute;**
- **Founder and Director, Blue Brain Project;**
  a supercomputing project that can model components of the mammalian brain
- **Coordinator, Human Brain Project (2013）**
  HBP involves researchers in 80 institutions across Europe.
- **Co-Founder, Frontiers (frontiersin.org)**
  A community-driven open-access academic publisher and social
- **TED** : https://www.ted.com/talks/henry_markram_a_brain_in_a_supercomputer?language=ja

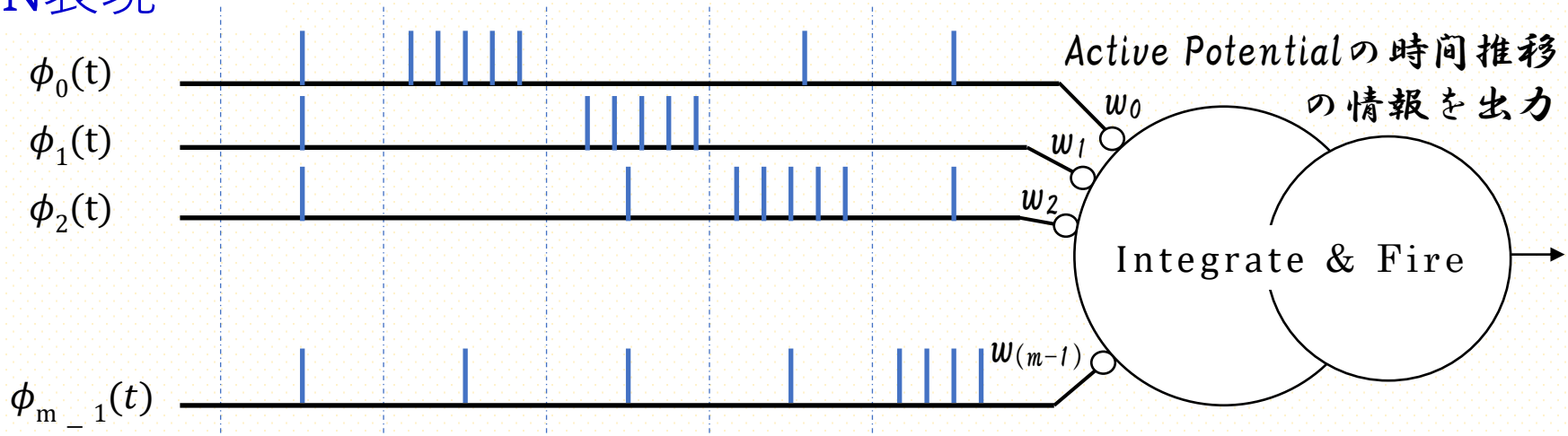
Wolfgang Maaβ Professor, Technische Universität Graz

- **Editor of Machine Learning,** 1995 - 1997
- **Member of the Editorial Board of Machine Learning,** 1998 - 2000
- **Editor of Archive for Mathematical Logic,** 1987 - 2000
- **Associate Editor of the Journal of Computer and System Sciences,** 1992 - 2014
- **Member of the Editorial Board of Neurocomputing, 1994 - 2007**
- **Member of the Editorial Board of Cognitive Neurodynamics,** 2006 - present
- **Editor of Biological Cybernetics,** 2006 - present

Neuron Modelの電子回路表現

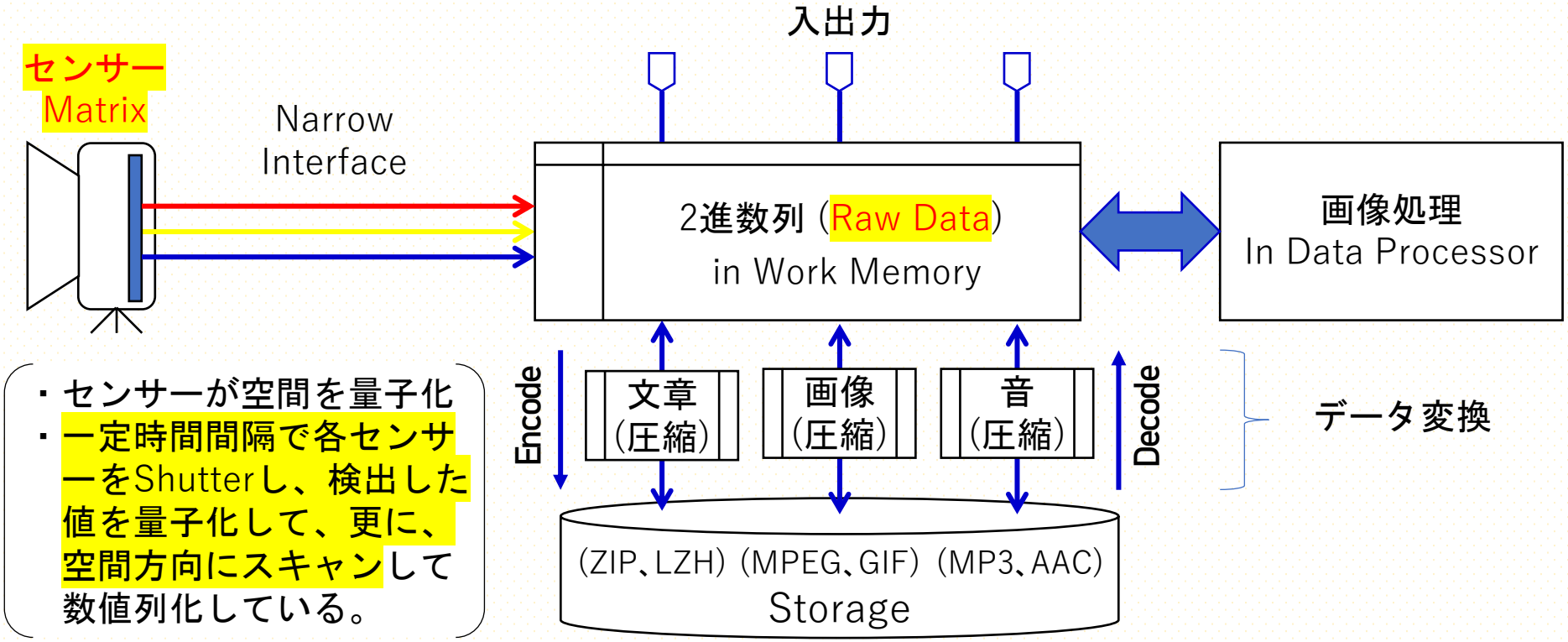


SNN表現



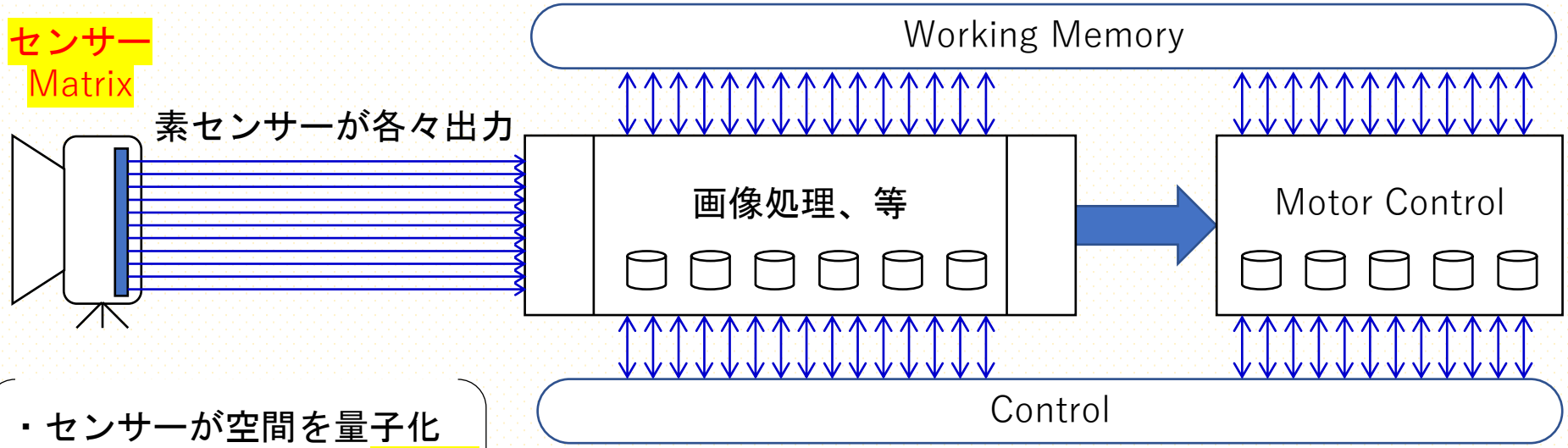
従来型コンピュータへの入力 ： データ

例 ： ２次元空間情報（静止画、Raw-Data）を時間方向にShutterで切り出す。



Neuromorphicへの入力 ： Spacio-Temporal情報

特徴 ： 空間を量子化した素センサーに生じたイベント（時間情報）が伝達される。



・センサーが空間を量子化
・物理アドレス毎に、Event-Drivenで出力
(Time-Slice無し)
（空間方向のスキャン無し）

・＃１ ： 「データ」の概念がない！

・＃２ ： 画像処理結果は出力されない！

両者のベンチマーク方法