Wprowadzenie do impulsowych sieci neuronowych (SNN – Spiking Neural Network)

Metody symulacji i implementacji

Prowadzący: dr inż. Andrzej Skoczeń
CERN, TE-MPE-EP
AGH, KOiDC
e-mail: skoczen@fis.agh.edu.pl

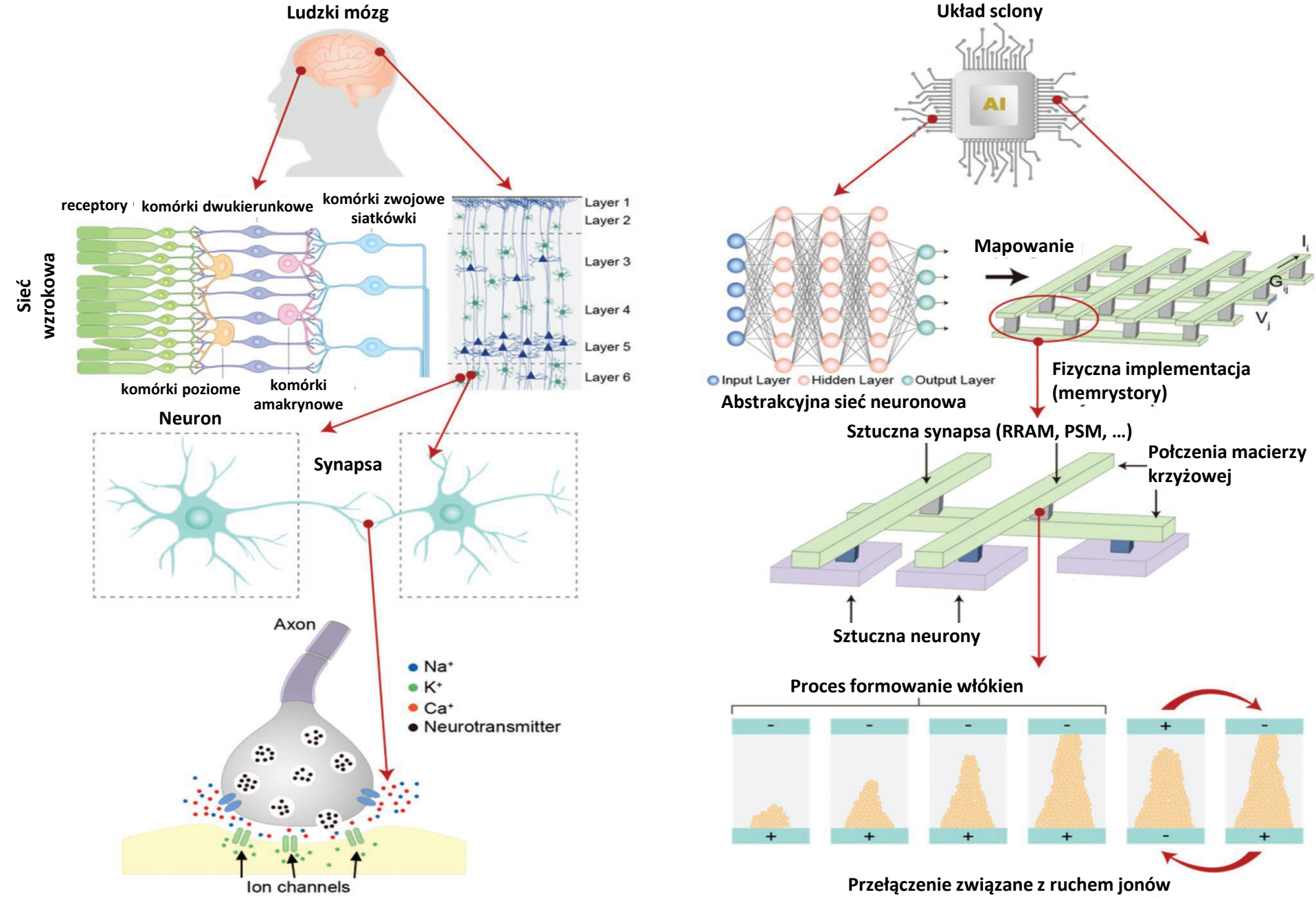
Wykład 1
2020

14 lipiec 2022

Fizjologia – technologia - informatyka:

- Synapsa, dendryt, neuron; potencjał czynnościowy
- Informacja temporalna
- Jak to przenieść do technologii ?
- Wielkie projekty: TrueNorth (IBM), SpiNNaker (Manchaster), ..., Loihi (Intel)
- Memrystory

Wysokopoziomowe porównanie ludzkiego układu nerwowego i sztucznego układu nerwowego zbudowanego z nowych urządzeń neuromorficznych.



Neuromorficzne systemy obliczeniowe

Systemy, które czerpią inspirację na temat architektury, technologii i zasad obliczeniowych bezpośrednio z wiedzy o mózgu biologicznym.

Morfologicznie mózg człowieka składa się z około 10^{11} podstawowych jednostek przetwarzania zwanych neuronami, połączonych masowo plastycznymi, adaptowalnymi połączeniami zwanymi synapsami.

Każdy neuron łączy około 10^3-10^4 innych neuronów poprzez połączenia synaptyczne.

Synapsa posiada około 20 poziomów wartości konduktancji co daje rozdzielczość 4,3 bita.

Neurony są rozmieszczone warstwami, a większość połączeń synaptycznych poświęcona jest łączeniu neuronów należących do kolejnych warstw.

Wysokopoziomowe porównanie ludzkiego układu nerwowego i sztucznego układu nerwowego zbudowanego z nowych urządzeń neuromorficznych.

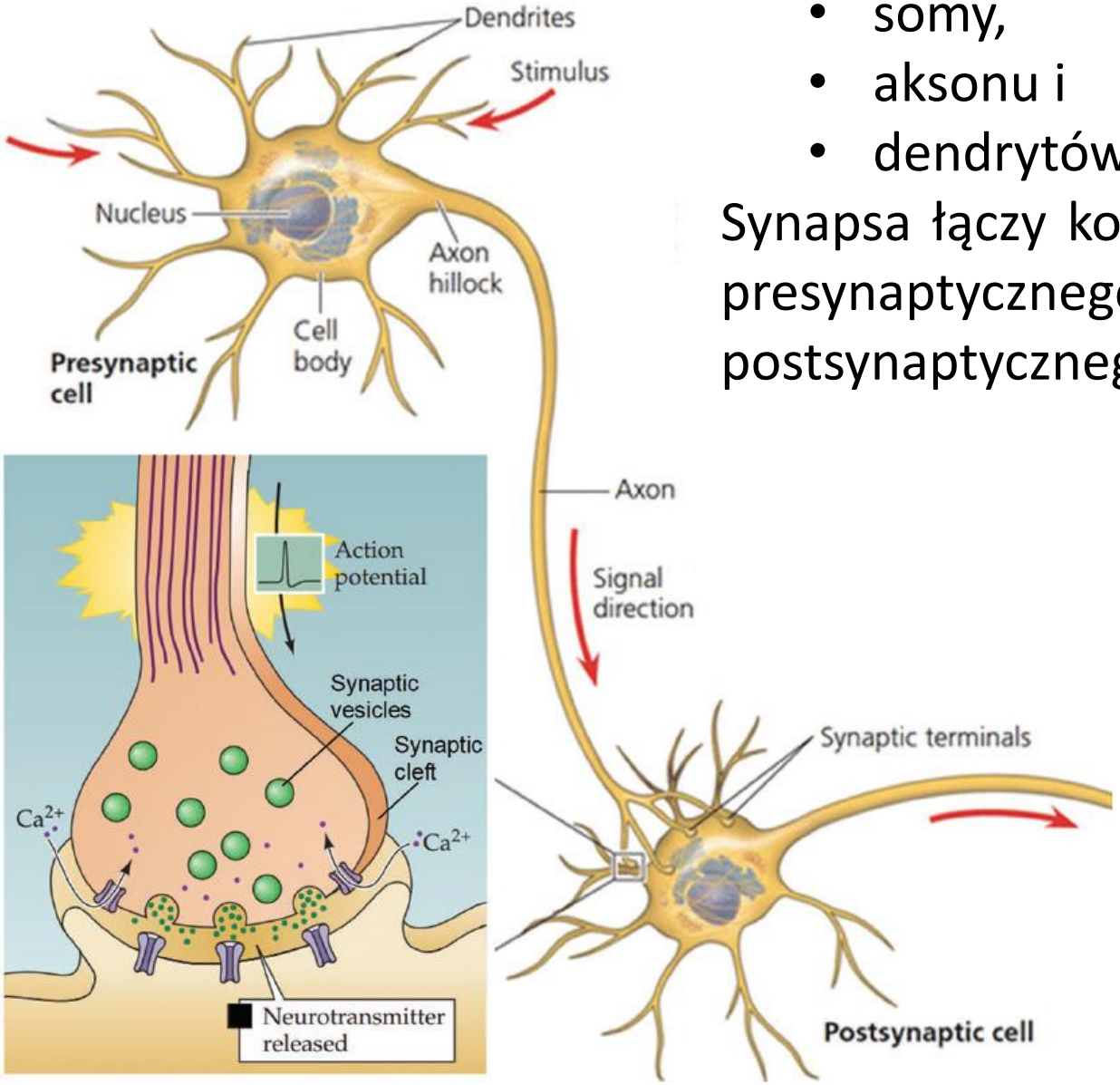
Biologia

Układ nerwowy człowieka ma różne typy sieci neuronowych, których podstawowymi elementami funkcjonalnymi są neurony i synapsy, w których różne typy kanałów jonowych leżą u podstaw elektrycznej aktywności neuronów.

Elektronika

Układ scalony sztucznej inteligencji AI (Artificial Intelligence) składa się z różnych typów ANN, które można mapować za pomocą krzyżowych macierzy sztucznych synaps i neuronów, których mechanizmem działania mogą być przewodzące włókna związane z ruchami jonów indukowanych elektrycznie, jak w przypadku RRAM.

Neuron

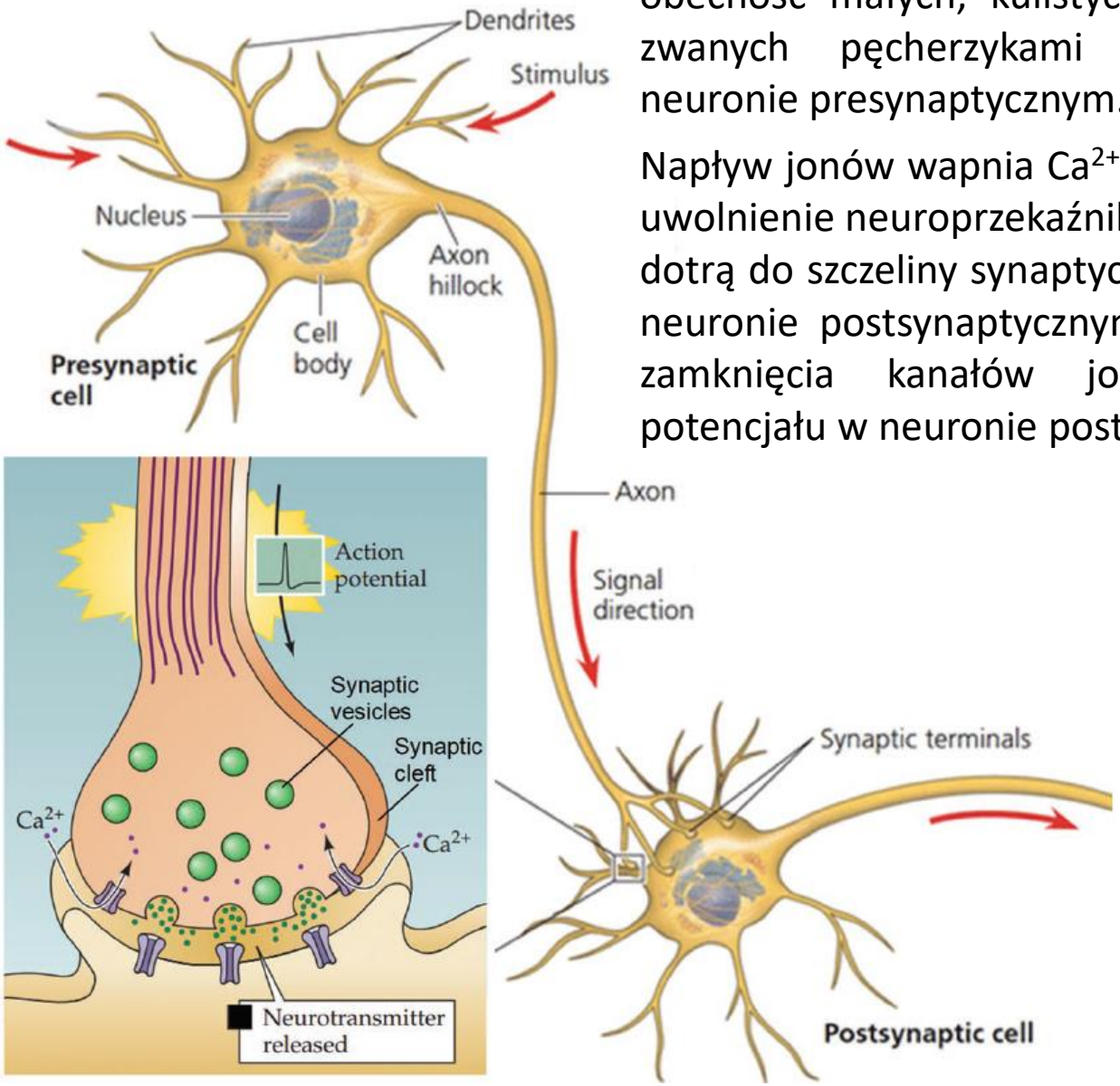


Typowy neuron składa się z:

- somy,
- aksonu i
- dendrytów.

Synapsa łączy końcówkę aksonu neuronu presynaptycznego z dendrytem neuronu postsynaptycznego.

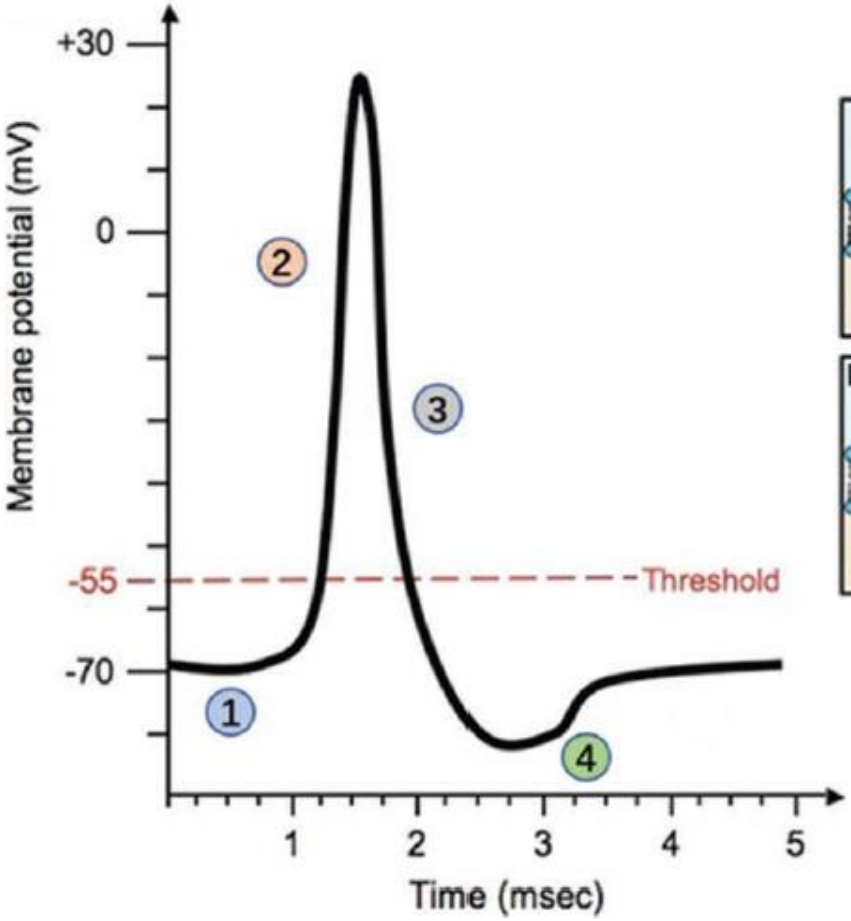
Neuron



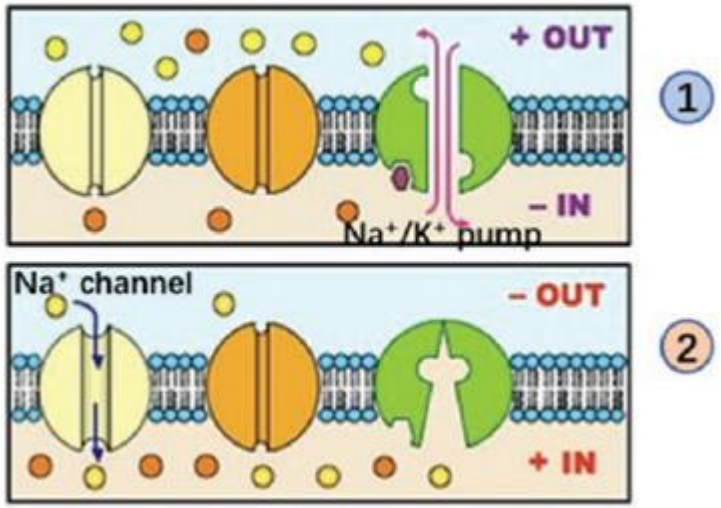
Przestrzeń w synapsach chemicznych nazywana jest szczeliną (cleft) synaptyczną, a istotną jej cechą jest obecność małych, kulistych organelli połączonych błoną zwanych pęcherzykami (vesicles) synaptycznymi w neuronie presynaptycznym.

Napływ jonów wapnia Ca^{2+} przez kanały jonowe powoduje uwolnienie neuroprzekaźników z tych pęcherzyków, a kiedy dotrą do szczeliny synaptycznej, wiążą się z receptorami w neuronie postsynaptycznym, prowadząc do otwarcia lub zamknięcia kanałów jonowych powodując zmianę potencjału w neuronie postsynaptycznym.

Potencjał czynnościowy

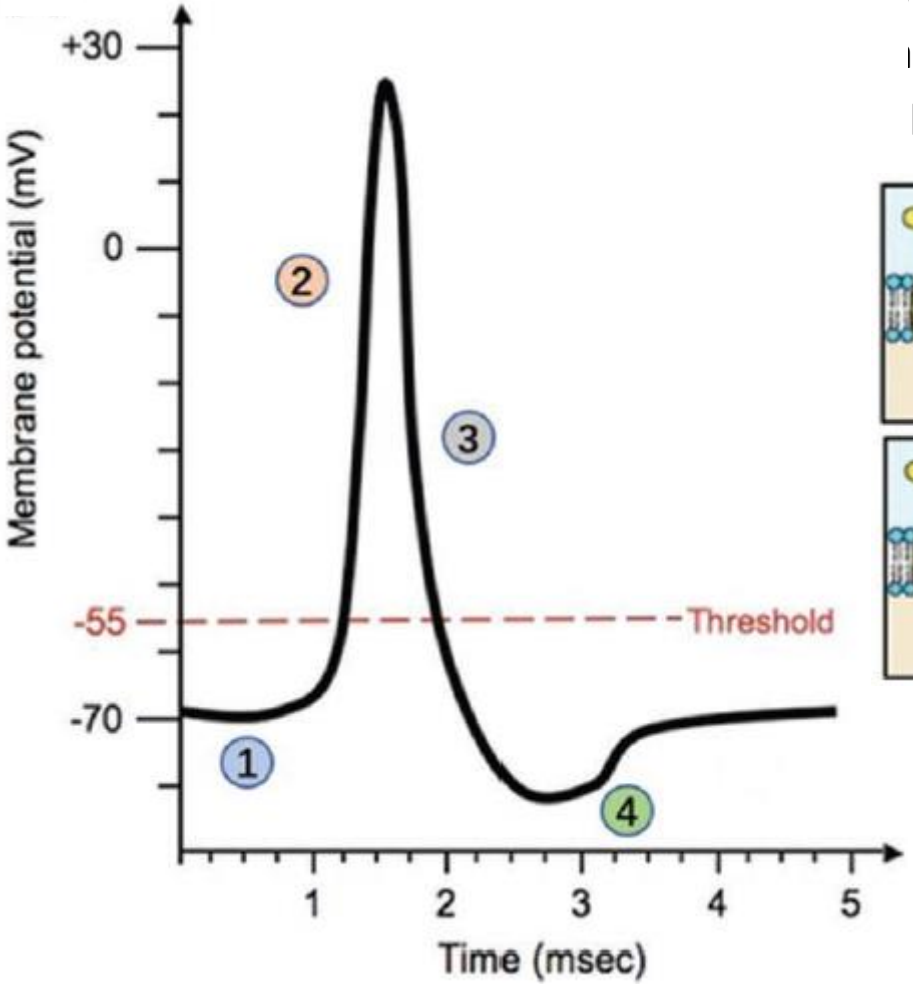


W **stanie spoczynku** pompy Na^+/K^+ transportują Na+ na zewnątrz membrany i K^+ do wnętrza membrany. Gradient stężenia jonów powoduje przepływ K^+ na zewnątrz błony, co skutkuje potencjałem błony o watości około -70 mV - różnica potencjału wewnątrz minus potencjał na zewnątrz błony.

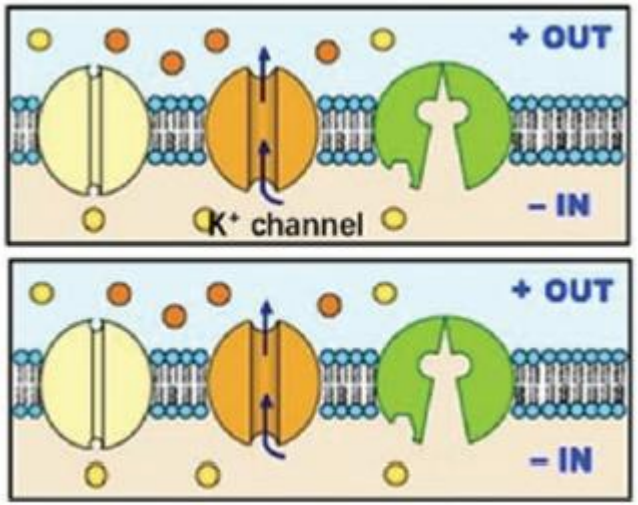


Gdy potencjał błonowy wzrasta do wartości progowej, aktywowane są kanały Na^+ i jony sodu przepływają z zewnątrz błony do wewnątrz, przez co potencjał błonowy gwałtownie rośnie. Ten proces nazywa się **depolaryzacją**.

Potencjał czynnościowy

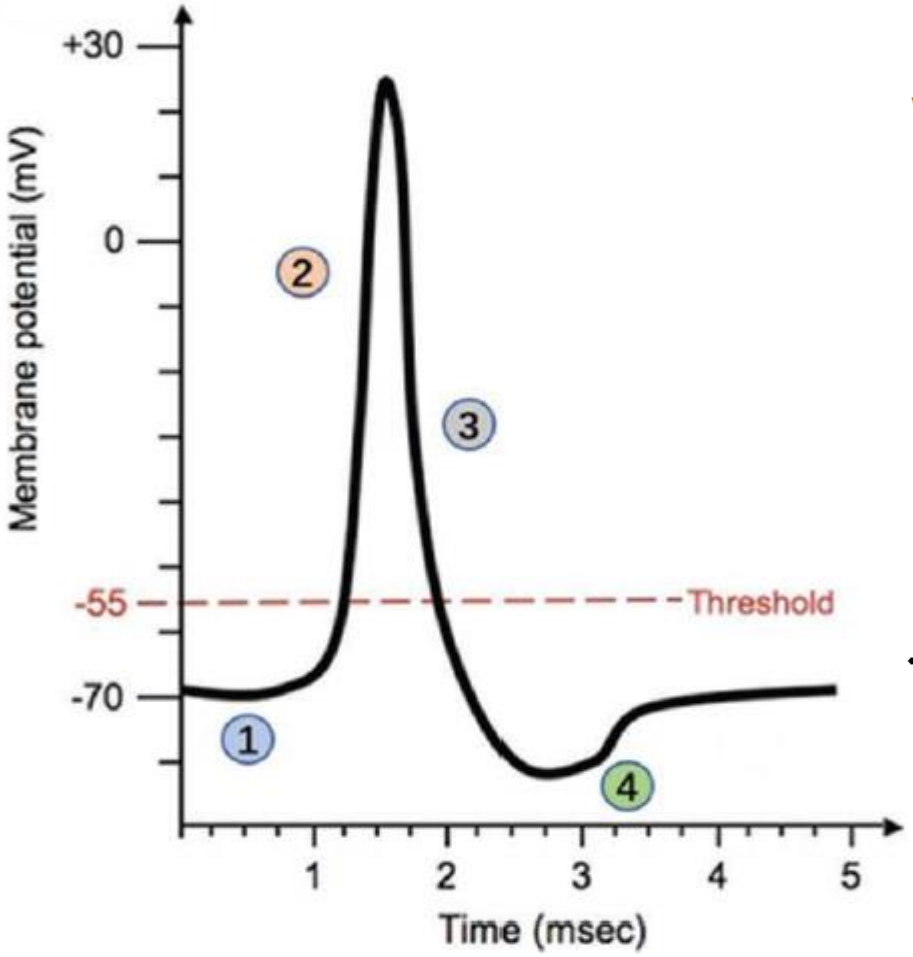


Wraz ze wzrostem potencjału błonowego aktywowane są kanały K^+ i napięcie ponownie pada, co nazywa się **repolaryzacją**.

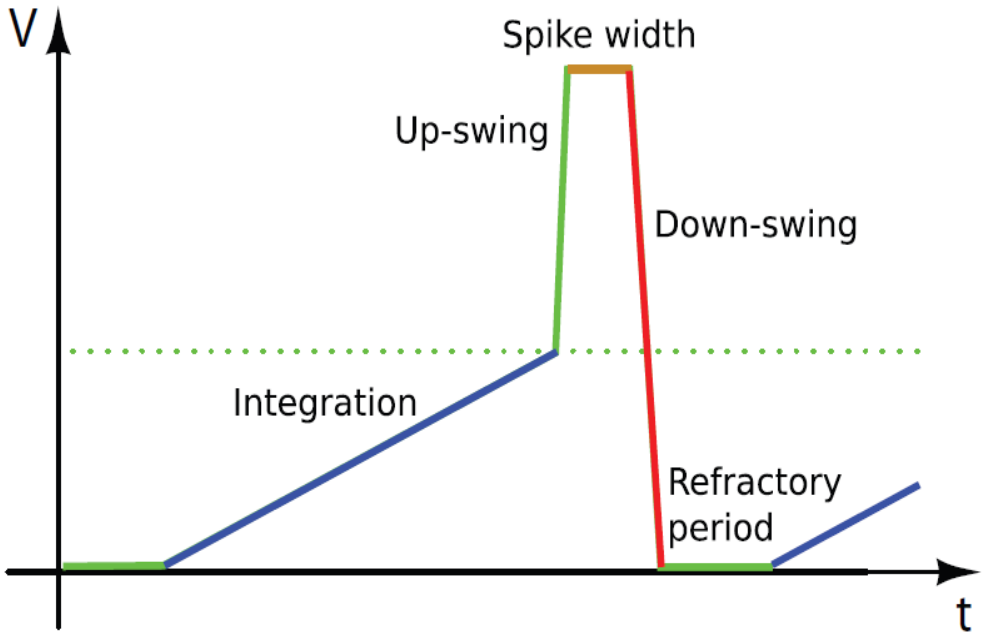


Z powodu wypływu jonów K^+ napięcie zwykle wykazuje „przerzut” poniżej potencjału spoczynkowego. Nazywa się to **hiperpolaryzacją**.

Biologia



Elektronika



Wymagania dla neuromorficzne systemów

Sprzęt neuromorficzny powinien:

- naśladować wysoce równoległą architekturę mózgów biologicznych,
- wykorzystywać architektury obliczeniowe w pamięci (in-memory computing) jako sposób na poprawę wydajności i wydajności energetycznej

Ponadto sprzęt neuromorficzny powinien wykorzystywać:

- kodowanie informacji o sygnale
- zasady obliczeniowe i paradygmaty uczenia się które umożliwiają nawet prostym biologicznym mózgom:
  - zachwycającą wydajność w interakcji,
  - adaptację do złożonych i nieoczekiwanych środowisk,
  - dużą szybkością reakcji i
  - minimalne zużycie energii

To wszysko powinno być osiągalne pomimo polegania na bardzo prostych i wysoce zawodnych jednostkach obliczeniowych.

1940 - Von Neuman - model komputera z dwama niezależnymi zadaniami: przechowywaniem (program i danych), przetwarzanie. Te dwa zadania wykonywane są w odseparowanych obszarach o dużej potrzebie komunikacji danych między nimi.

Fizyczne oddzielenie pamięci i jednostek przetwarzania danych zwiększa koszt ruchów dużych zbiorów danych. Efekt znany jest jako „wąskie gardło von Neumanna” (“von Neumann bottleneck”).

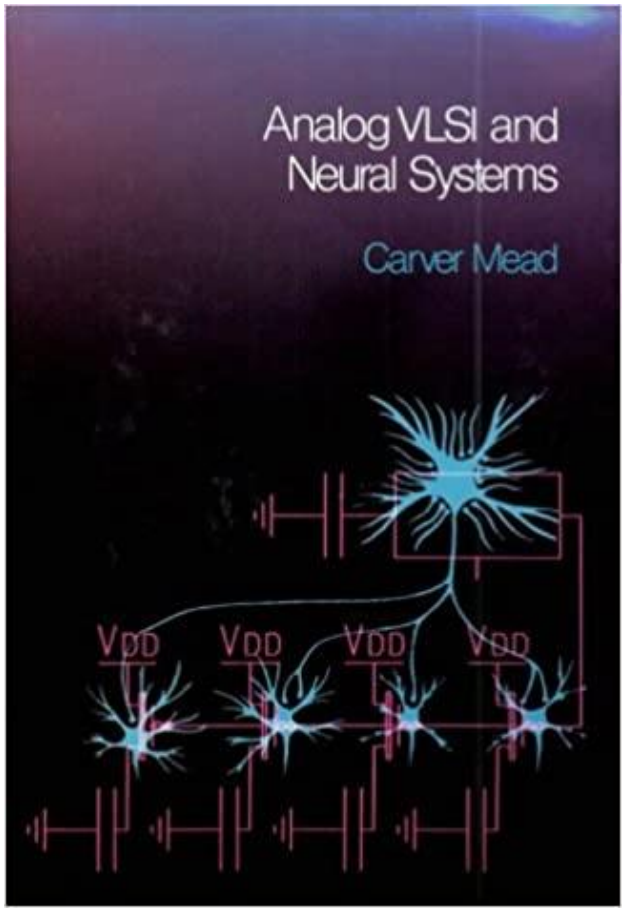
1947 - Brattain, Bardeen, Shockley– skonstruowano i uruchomiono pierwszy działający tranzystor ostrzowy

1958 - Kilby Noyce - zaprojektowali i zbudowali działające modele układów scalonych

1960 - Moore - Wraz z rozwojem układów scalonych liczba tranzystorów w układzie scalonym podwaja się co 18 do 24 miesięcy.

Produkcja półprzewodników zbliża się do granic prawa Moore'a, dlatego zaproponowano różne rozwiązania, aby utrzymać przyszłą ewolucję systemów przetwarzania.

1980 - Carver Mead zwrócił uwagę na analogię między fizyką neuronów biologicznych a zachowaniem tranzystorów w reżimie podprogowym rozwijając sieci neuronowe oparte na obwodach analogowych.



Doprowadziło to do:

- wdrożenia pierwszych siatkówek krzemowych i
- zaproponowania nowego paradygmatu obliczeniowego, w którym dane i zadania przetwarzania są wykonywane przez niepodzielne byty, czerpiąc inspirację z biologicznych systemów neuronowych.

Wyzwaniem dla sprzetu neuromorficznego

Ogromnym wyzwaniem pozostaje realizacja urządzeń , które

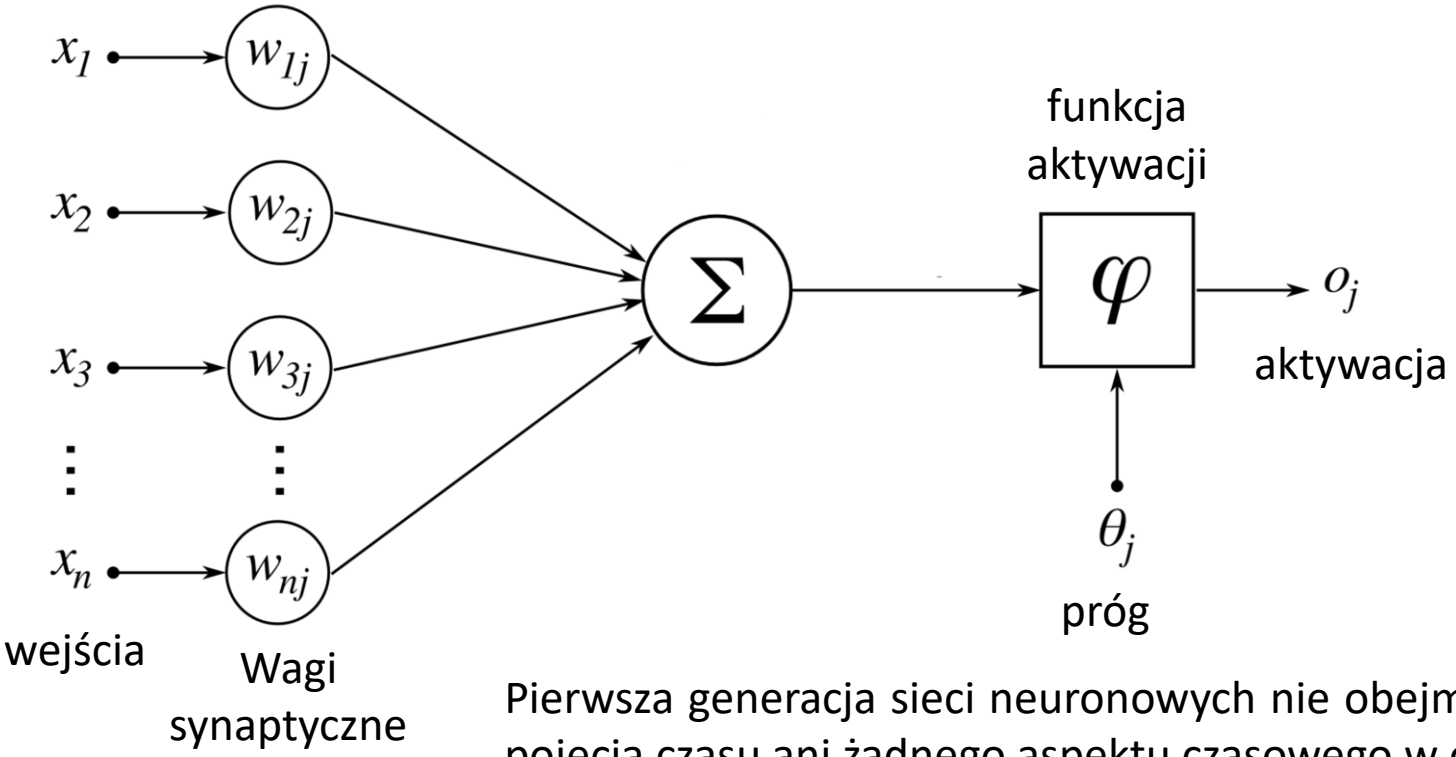
- naśladują dynamikę neuronów i synaps w mózgu,
- naśladują ogromną równoległość i
- posiadają sterowaną zdarzeniami (event-based) architekturę obliczeniową
- posiadają porównywalną złożoność i budżet mocy co mózg,
- i pracują w czasie rzeczywistym.

Sztuczna sieć neuronowa

ANN (Artificial Neural Network)

1943 - Mc Culloch i Pitts - zaproponowali pierwsze modele obliczeniowe neuronów biologicznych

1958 - Rosenblatt - zaproponował perceptron

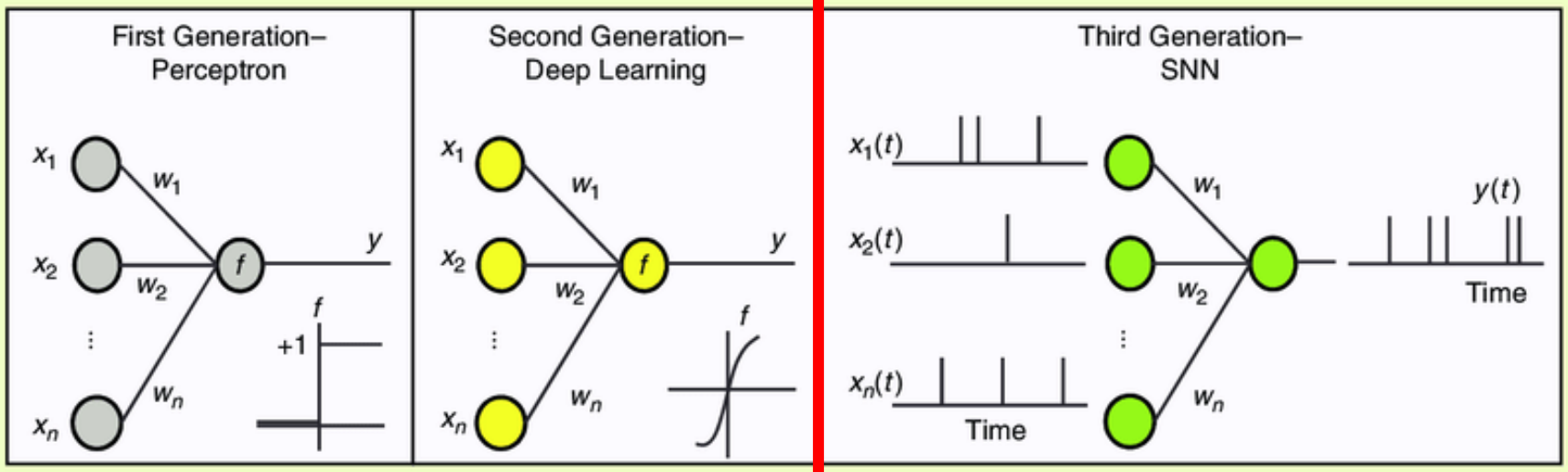


Pierwsza generacja sieci neuronowych nie obejmowała żadnego pojęcia czasu ani żadnego aspektu czasowego w obliczeniach.

McCulloch,W.S.; Pitts,W. A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. **1943**, 5, 115–133

Trzy generacje sztucznych sieci neuronowych

Pierwsze generacje sieci neuronowych nie obejmowały żadnego pojęcia czasu ani żadnego aspektu **czasowego** w obliczeniach.



Funkcję aktywacji progu zastąpiono ciągłą taką jak

- gładki sigmoid,
- radialna funkcja bazowa lub
- ciągła funkcja odcinkowo liniowa,
- prostująca nieliniowa funkcja aktywacji ReLU

DNN – Deep Neural Network
RNN – Recurent Neural Network

SNN działają przy użyciu impulsów w podobny sposób jak neurony biologiczne.

Oprócz stanu neuronu i wag synaptycznych, SNN wprowadzają również pojęcie **czasu** do swojego modelu działania.

Druga generacja

Algorytmy uczenia oparte na metodzie gradientowej (gradient descent) można zastosować do optymalizacji wag sieci.

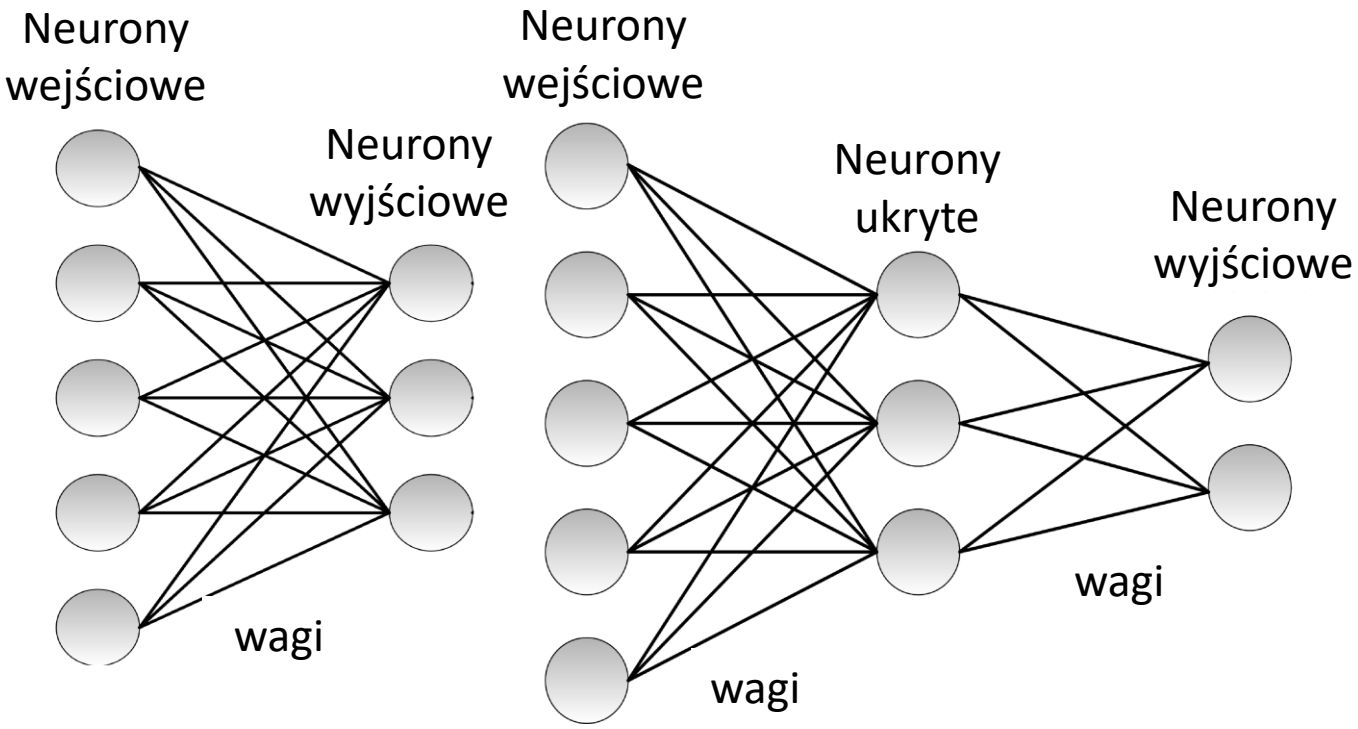
Alternatywne reguły uczenia się zostały zaproponowane jako reguła delta oparta na algorytmie najmniejszych średnich kwadratów (LSM) opublikowanym przez Widrow'a.

Druga generacja okazała się metodą tworzenia uniwersalnych aproksymatorów dla dowolnej analogowej funkcji ciągłej,

czyli każda analogowa funkcja ciągła może być aproksymowana przez sieć tego typu z pojedynczą jednostką ukrytą.

Algorytm wstecznej propagacji błędów BP (BackPropagation) rozszerzył zastosowanie techniki gradientowej na sieci z dowolną liczbą N warstw ukrytych, popularnie zwane głębokimi sieciami neuronowymi DNN. W przypadku z N=3 warstwami:

- pierwsza warstwa to warstwa neuronów wejściowych,
- druga warstwa to warstwa neuronów ukrytych i
- trzecia warstwa to warstwa neuronów wyjściowych.



Pokazane architektury **ANN** są architekturami sprzężonymi wyłącznie do przodu (feedforward), tzn. sygnał propaguje się od wejścia do wyjścia w sposób jednokierunkowy.

Zaproponowano inne architektury, znane jako rekurencyjne sieci neuronowe **RNN** (Recurrent Neural Network), w których istnieją połączenia zwrotne z następnych warstw w architekturze do poprzednich warstw. Do pionierskich można zaliczyć architectury:

- adaptacyjnej teorii rezonansu (ART) Grossberga,
- mapy samoorganizujące się Kohonena czy
- modele Hopfielda.

Obecnie dominuje idea LSTM (Long Short-Term Memory) pochodząca od Hochreiter'a (1991)

Sieci ANN drugiej generacji zostały opracowane w oprogramowaniu i przeszkolone w trybie offline. Uczenie sieci DNN wymaga ogromnej ilości danych z adnotacjami, aby poprawnie uogólnić problem bez nadmiernego dopasowania (overfitting) i intensywnych zasobów obliczeniowych.

Wzrost możliwości obliczeniowych nowoczesnych komputerów i dostępność ogromnych ilości informacji sprawiły, że DNN stało się bardzo popularne, umożliwiając rozwój wielu aplikacji, które wykorzystują złożone architektury, takie jak:

- LeNet do rozpoznawanie cyfr pisanych odręcznie,
- system rozpoznawania mowy Microsoftu czy
- AlexNet do rozpoznawania obrazu.

W konsekwencji byliśmy świadkami eksplozji sieci DNN i uczenia maszynowego.

Wydajność DNN w porównaniu z ludzkim mózgiem:

- pod względem szybkości i zużycia energii jest za niska,
- pod względem ilości potrzebnych zasobów sprzętowych jest za wysoka.

Czyżby podobieństwo do ludzkiego mózgu pod względem kodowania informacji było zbyt małe?

W konwencjonalnych sieciach DNN:

- procesują sekwencję statycznych ramek,
- wyjścia różnych warstw neuronowych są obliczane w sposób sekwencyjny,
- każda warstwa musi czekać, aż wyjście z poprzedniej warstwy zostanie obliczone, aby wykonać swoje obliczenia, co powoduje znaczne opóźnienie (latency) rozpoznawania w sieci.

W mózgu biologicznym:

- informacja jest przetwarzana w sposób ciągły w czasie,
- neurony biologiczne przekazują informacje do kolejnych warstw neuronalnych w postaci impulsów (spikes),
- za każdym razem, gdy neuron emituje impuls, jest on przekazywany do połączonych neuronów i przetwarzany z opóźnieniem jedynie połączenia synaptycznego.

1996 - Thorpe zademonstrował, że ludzki mózg był w stanie rozpoznać znajomy wzrokowo obiekt w czasie, gdy tylko jeden impuls rozchodzi się we wszystkich warstwach kory wzrokowej.

1994 - Rolls - podobne szybkości przetwarzania wizualnego zostały zmierzone u makaków.

Obliczenia pojedynczego obszaru korowego są zakończone w ciągu 10–20 ms, podczas gdy szybkość wyzwalania neuronów biorących udział w obliczeniach jest poniżej f=100 Hz (T=10ms), co nie pozwala na obliczenia oparte na kodowaniu zmiennych analogowych jako szybkość (częstotliwość) “odpalania” neuronów.

Trzecia generacja sieci neuronowych, impulsowe sieci neuronowe (SNN), ma na celu:

- wypełnienie luki między neuronauką a uczeniem maszynowym, wykorzystując biologicznie **realistyczne modele neuronów** do kodowania informacji i obliczeń,
- próba pełniejszego wykorzystania **wydajności sygnału** przestrzenno-czasowego w kodowaniu i przetwarzaniu oraz
- osiągnięcie **wydajności energetycznej** obserwowanej w mózgach biologicznych.

Impuls (spike):

- jest główną formą transmisji informacji w SNN, I
- jest często uważany za zdarzenie (event).

SNN jest sterowane zdarzeniami (event-driven), czyli neurony są aktywowane tylko wtedy, gdy generowany jest spike.

Gdy spike zostanie wprowadzony do SNN, neurony wybudzają się krok po kroku do obliczeń, podczas gdy neurony, które nie są przebudzone, nie będą uczestniczyć w procesie rozpoznawania. Dlatego też impulsowa sieć neuronowa jest odpowiednia dla tych aplikacji, które wymagają dużego zużycia energii i są przez większość czasu uśpione.

Dzięki zdarzeniowej charakterystyce SNN, te neurony, które nie są aktywowane, nie uczestniczyą w rzeczywistej aktywności, oszczędzając w ten sposób zasoby enrgetyczne, co jest bardzo odpowiednie do zastowań o małej mocy na dedykowanych chipach.

Metody kodowania ciągów impulsów

Zakodowanie poziomu aktywności neuronów jako **częstotliwości ich odpalania nie** jest prawidłowe gdyż ten typ kodowania nie korzysta z rzadkości (sparsity) impulsów. Sparsity powinna charakteryzować przetwarzanie SNN.

W odniesieniu do możliwości szybkich obliczeń oczekiwanych od SNN, to kodowanie szybkości wyzwalania wprowadza opóźnienie w obliczaniu wyjściowej szybkości wyzwalania. Co więcej, nie jest to biologicznie wiarygodne, o czym świadczą eksperymenty Thorpe'a i Rollsa.

Inne schematy kodowania to:

- czas między spajkami
- opóźnienie względem danego czasu synchronizacji, znane również jako kodowanie czasu do pierwszego impuls TFS (Time to First Spike)
- po prostu kodowanie wartości w kolejności spajków co jest znane jako kodowanie kolejności rang lub
- kodowanie detekcji synchronicznej

1951 – Hodgkin, Huxley - Klasyczny model jest biofizycznym modelem czwartego rzędu, który opisuje zachowanie prądów wpływających do neuronowych kanałów jonowych w biologicznie realistyczny sposób. **The Nobel Prize in Physiology or Medicine 1963**

1961 – FitzHugh, Nagumo

1981 – Morris, Lecar

- zaproponowano różne uproszczone modele drugiego rzędu,

2003 - Izhikevich

2005 - Brette, Gerstner - AdEx (Adaptive Exponential Integrate and Fire)

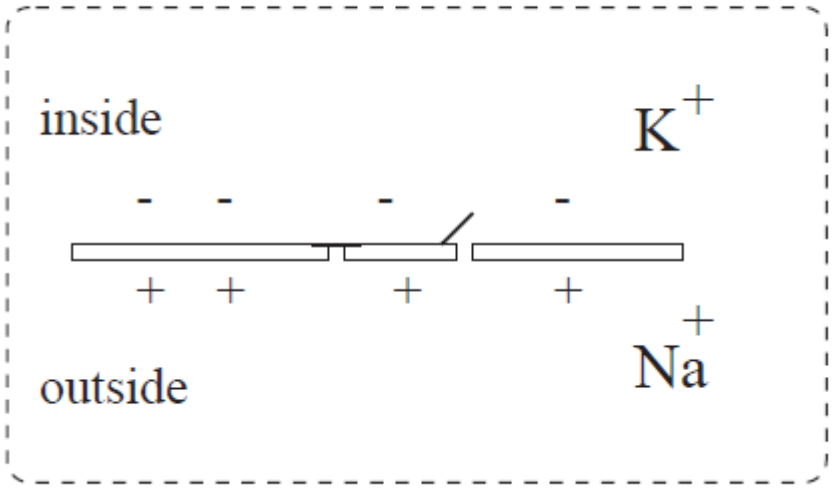
Modele zdolne do odtworzenia wielu różnych reżimów impulsów obserwowanych w neuronach biologicznych poprzez zmianę zredukowanej liczby parametrów.

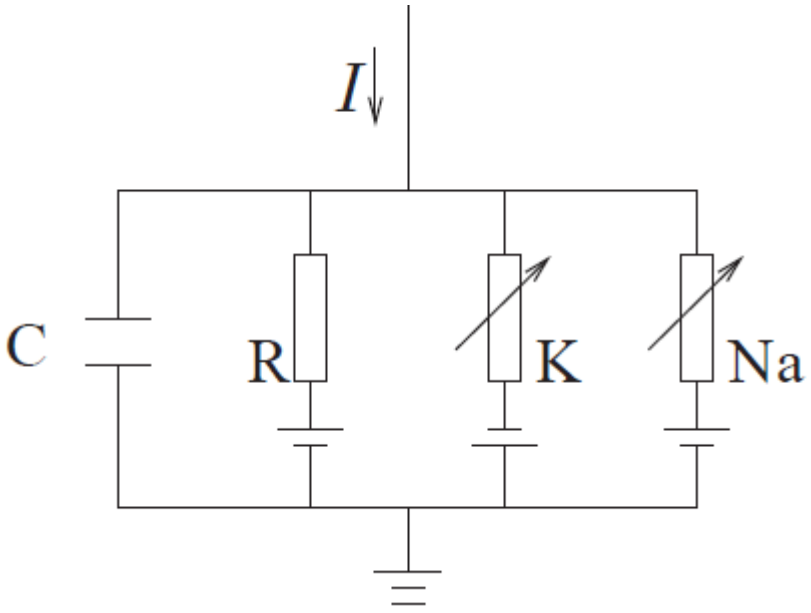
Trudne do analizy obliczeniowej i nieprzyjazne dla implementacji sprzętowych.
Do celów obliczeniowych stosuje się proste modele.

1907 – Lapicque - LIF (Leaky-Integrate-and-Fire) – model fenomenologiczny pierwszego rzędu.

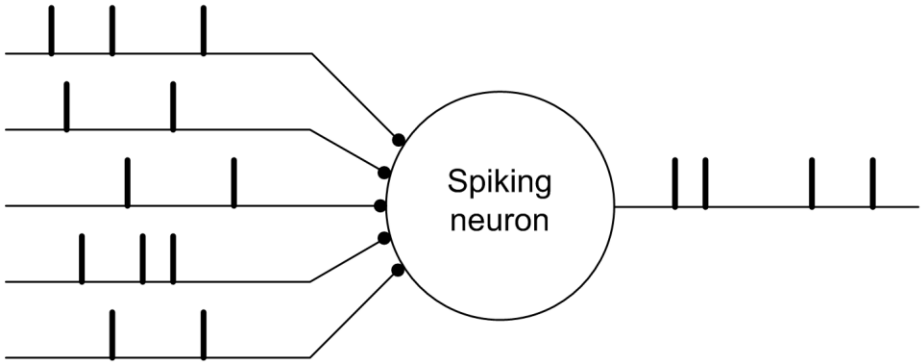
Lapicque, L. Recherches quantitatives sur l'excitation e´lectrique des nerfs traite´e comme une polarization. J. Physiol. Pathol. Gen. 9:620–635; 1907.

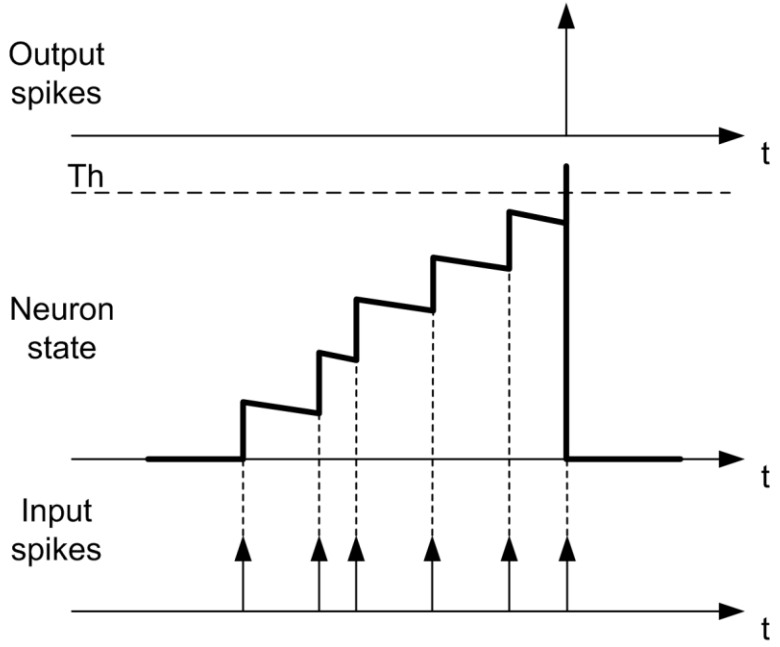
Model Hodgkin'a-Huxley'a





Działanie modelu LIF

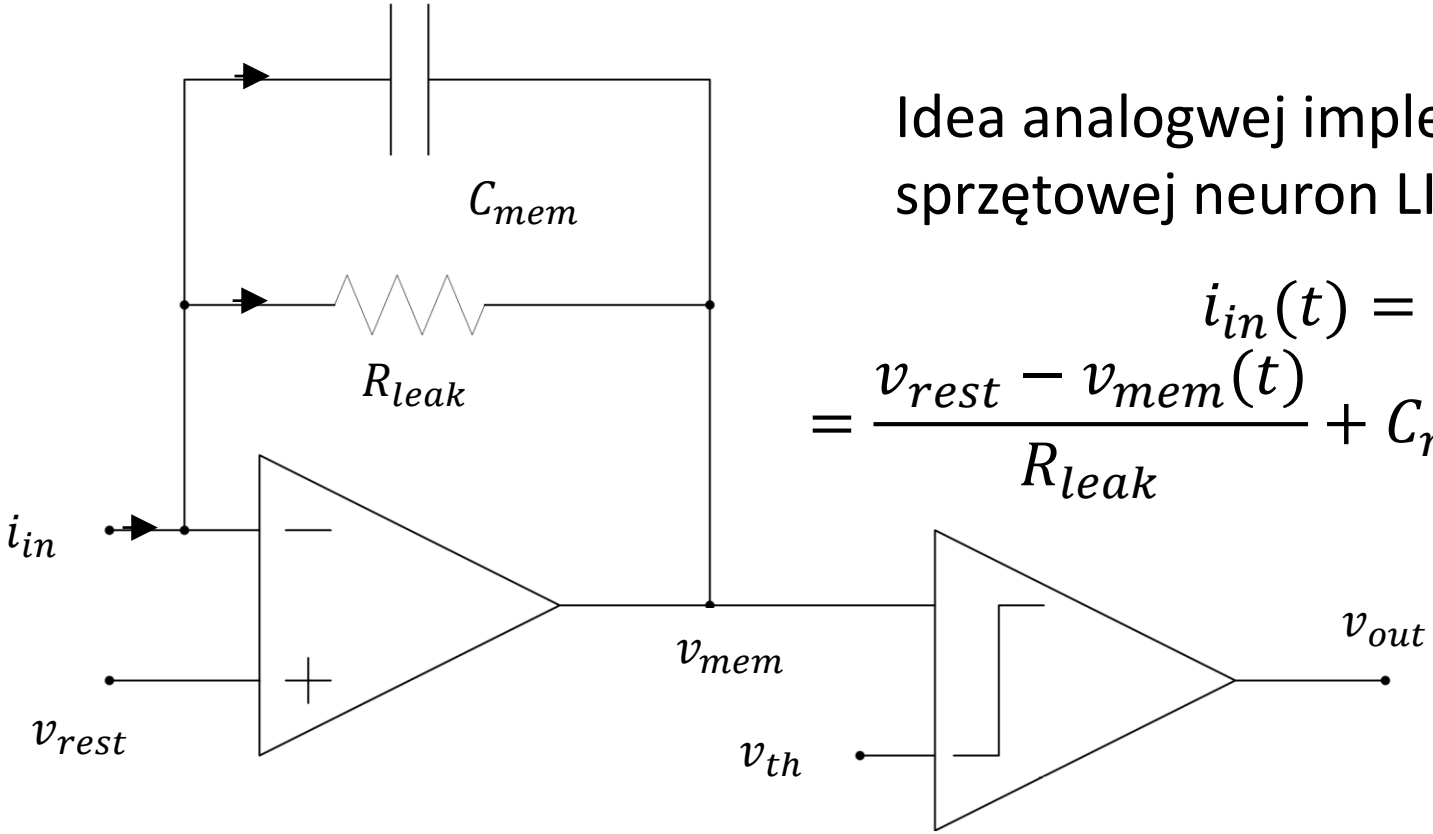




Neuron “spikujący” otrzymuje impulsy z kilku wejść (dendrytów), przetwarza je i generuje impulsy wyjściowe (do axonu) ze swojego węzła wyjściowego.

Ewolucja czasowa stanu neuronu podczas odbierania impulsów wejściowych. Po osiągnięciu progu generuje impuls wyjściowy.

Działanie modelu LIF



Idea analogwej implementacji sprzętowej neuron LIF

$$i_{in}(t) = \frac{v_{rest} - v_{mem}(t)}{R_{leak}} + C_{mem}\frac{dv_{mem}(t)}{dt}$$

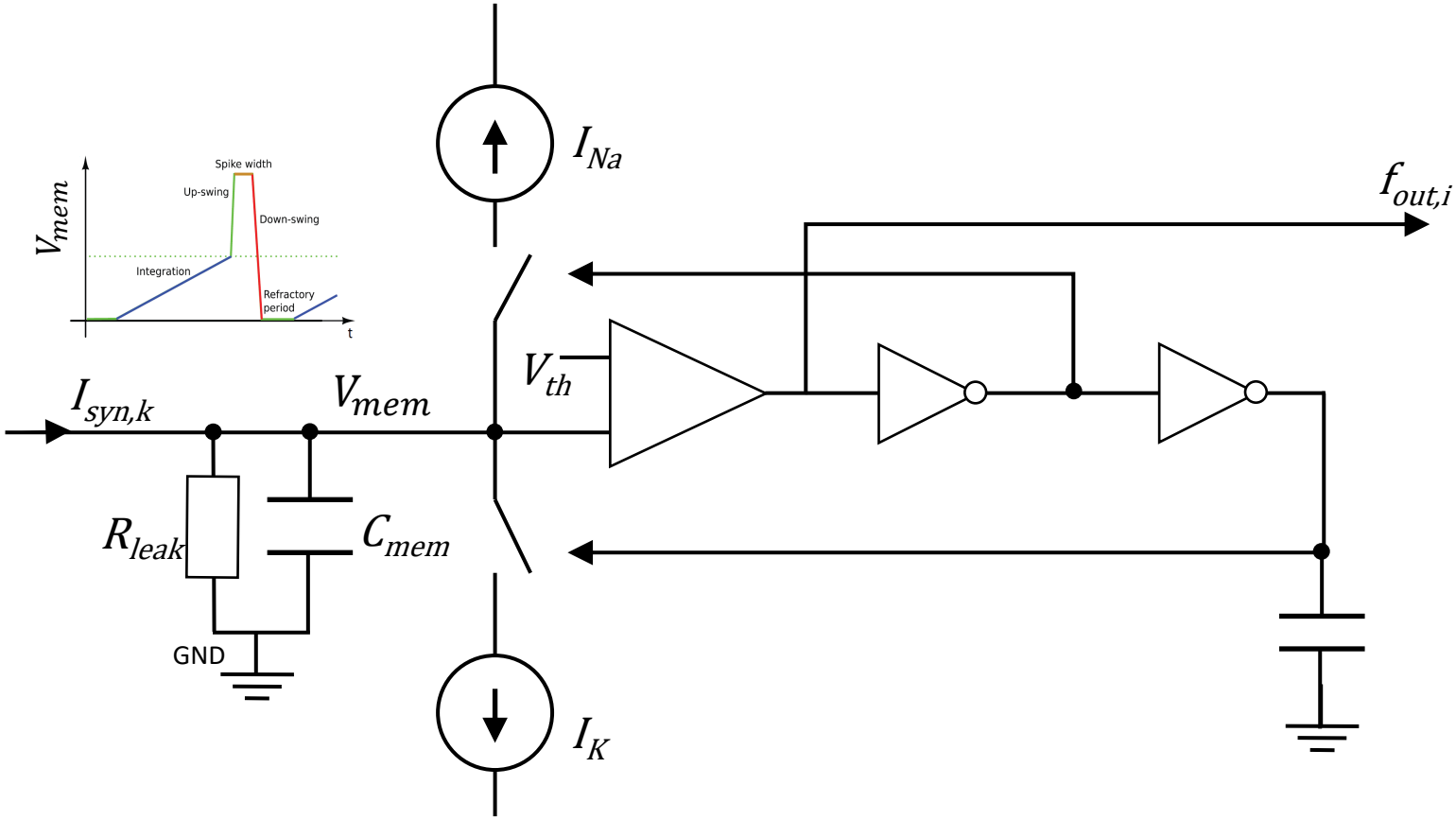
i_{in} - Prąd synaptyczny

v_{mem} - Potencjał membrany

C_{mem} - Pojemność membrany

R_{leak} - Upływ

Analogowy model neuronu LIF



Własność	**ANN**	**SNN**
Przetwarzanie danych	Na ramkch (frame-based, clock-based)	Na impulsach (spike-based, event-based)
Opóźnienie (latency)	Duże	Małe
Rozdzielczość czasowa	Mała	Duża Zachowanie korelacji czasoprzestrzennej
Przetwarzanie w czasie	Próbkowanie	Ciągły
Złożność model neuronu	Mała	Duża
Dokładność rozpoznania	Większa	Mniejsza
Przełączanie sprzętu	Możliwy	Nie możliwy
Skalowalność układu	Ad hoc	Dodając moduły
Szybkość rozpoznawania	Mała Niezależność od bodźca wejściowego Zależne od zasobów sprzętowych W zależności od złożoności system	Duża W zależności od statystyk wejściowych Nie zależy od złożoności systemu
Zużycie mocy	Określane przez moc procesora i pobieranie z pamięci Niezależność od bodźca wejściowego	Określone przez mocy na przetwarzanie jednego zdarzenia w module W zależności od statystyk bodźców
Rekurencyjność topologi	Konieczność iterowania aż do zbieżności	Natychmiastowa

Implementacja połączeń synaptycznych

Kluczowy problemów przy próbach implementacji sprzętowej dużych macierzy populacji neuronowych:

- biologiczne neurony są rozmieszczone w 3D i masowo połączone między populacjami, ale
- w równoległym sprzęcie 2D fizyczne okablowanie pozwala na zaimplementowanie połączeń między sąsiadującymi neuronami.

AER Address-Event-Representation - asynchroniczny protokół komunikacyjny, który został pomyślany w celu masowego łączenia populacji neuronów, które mogą być zlokalizowane w tych samych lub różnych chipach systemu „wirtualnego okablowania”.

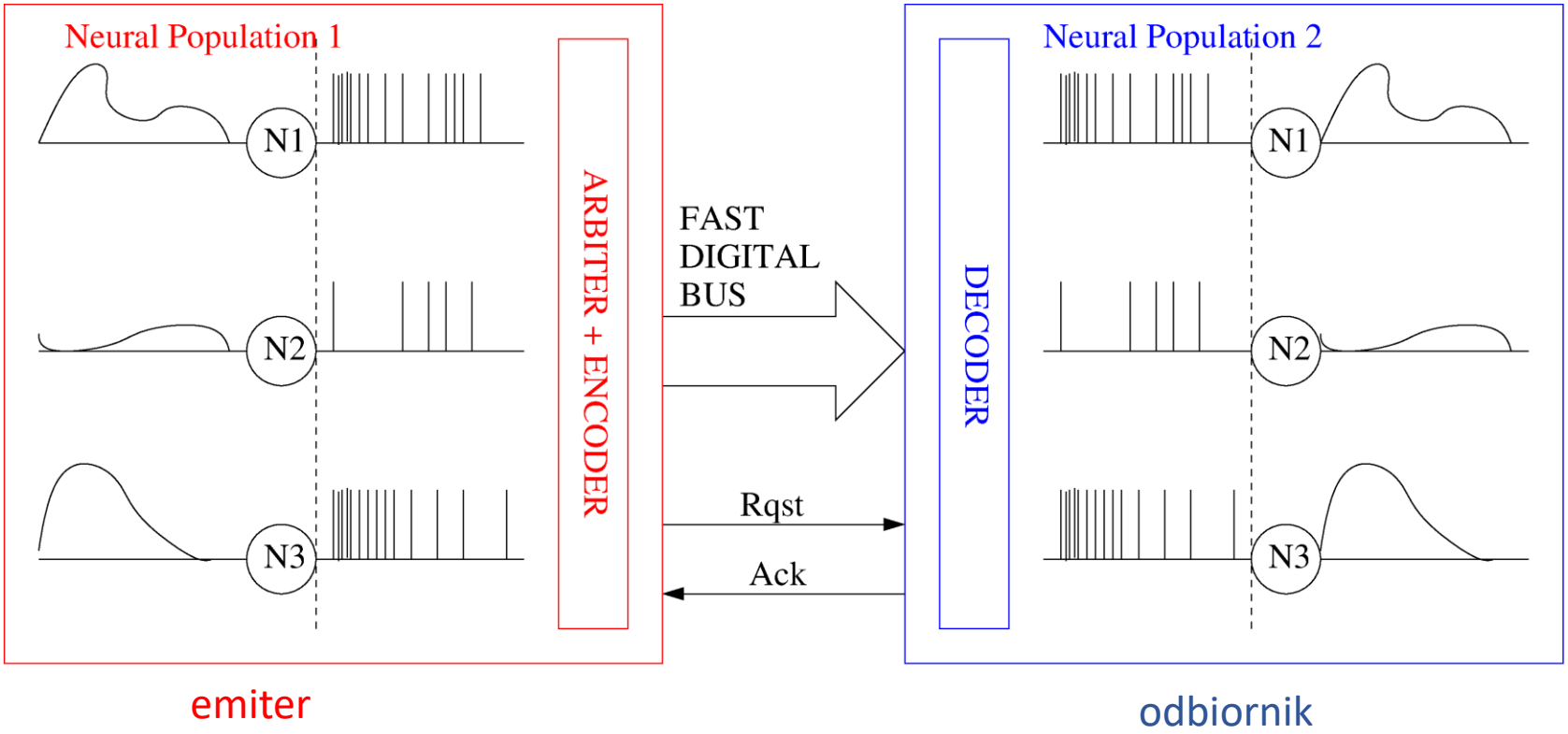
Implementacja połączeń synaptycznych

Implementacja **adaptowalnych** połączeń synaptycznych w technologii CMOS jest kosztowna. Wymaga:

- użycia dużej ilości obwodów dla pamięci analogowej lub bloków pamięci cyfrowej, które są kosztowne pod względem powierzchni i zapotrzebowania na energię, i
- ponadto należy zaimplementować zasady uczenia się, aby zaktualizować te urządzenia pamięci synaptycznej.

Konieczna alternatywna nanotechnologia zgodna z CMOS dla stworzenia kompaktowego, adaptowalnego urządzenia przestrzegającego biologicznych zasad uczenia.

Address-Event-Representation



Ilustracja przedstawiająca dwie populacje neuronowe komunikujące się za pośrednictwem magistrali AER typu punkt-punkt. Każdy neuron w populacji emitera może być praktycznie połączony z każdym neuronem w populacji odbiornika.

Schematy routingu typu impulsów umożliwiły:

- wdrożenie wysoce równoległych, masowo połączonych sieci neuronowych typu impulsowego oraz
- wielochipową integrację sprzętu SNN przeznaczonego do realizacji różnych określonych części funkcji poznawczych, w tym
- integrację **czujników** i **procesorów neuronowych** opartych na spajkach.

Czujniki SNN

Istnieje wiele różnych czujników wizualnych AER, które wykorzystują różne podejścia do kodowania luminancji:

- proste czujniki luminancji do transformacji częstotliwości [87],
- czujniki kodujące czas do pierwszego impulse TFS (Time-to-First-Spike),
- sensory zazębione,
- sensory kodujące kontrast przestrzenny,
- sensory filtrujące przestrzenne i czasowe, które dostosowują się do oświetlenia i kontrastu czasoprzestrzennego oraz
- czasowe detektory przejść - Dynamic Vision Sensors DVS

Dynamic Vision Sensors

Produkują one jako dane wyjściowe strumień zdarzeń asynchronicznych, w których każdy piksel koduje czasową zmienność oświetlenia wchodzącego w piksel.

Zalet czujnika DVS:

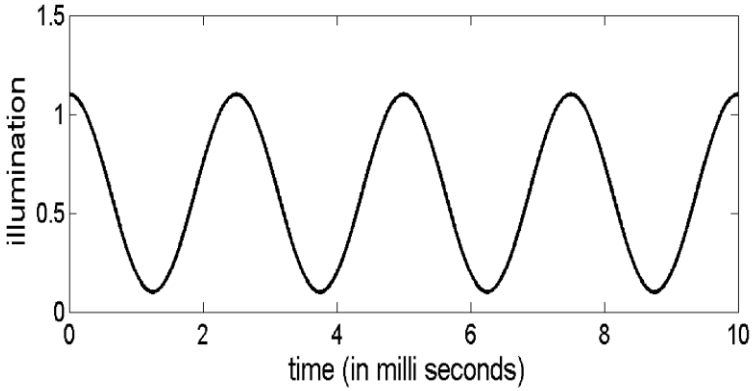
- koduje informacje w sposób skompresowany, wysyłając tylko impulsy w przypadku istotnej zmiany oświetlenia i tym samym usuwając z poruszającego się obiektu statyczne cechy tła sceny.
- wszystkie dokładne informacje przestrzenno-czasowe obiektu są zachowywane z raportowaną dokładnością w czasach impulsu rzędu 10 µ.

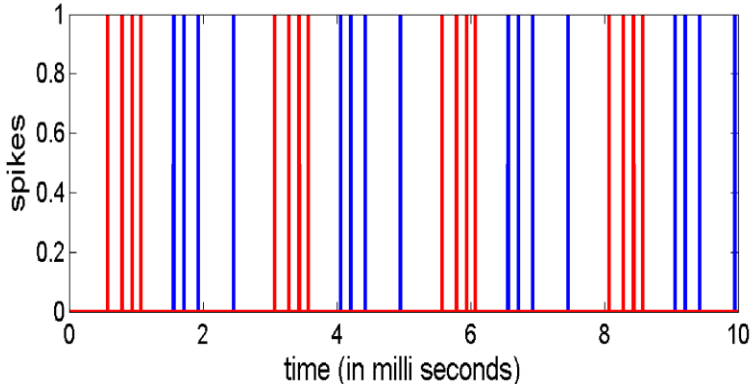
DVS są idealne do szybkich systemów przetwarzania i rozpoznawania.

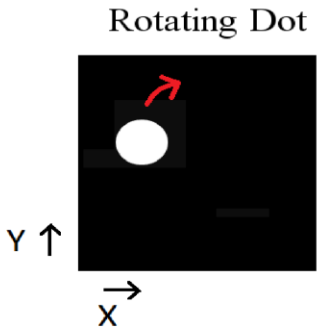
Istnieją komercyjne prototypy kamer DVS o wysokiej rozdzielczości do opracowania szybkich autonomicznych inteligentnych systemów wizyjnych: iniVation, Insightness, Samsung, CelePixel, Prophesee.

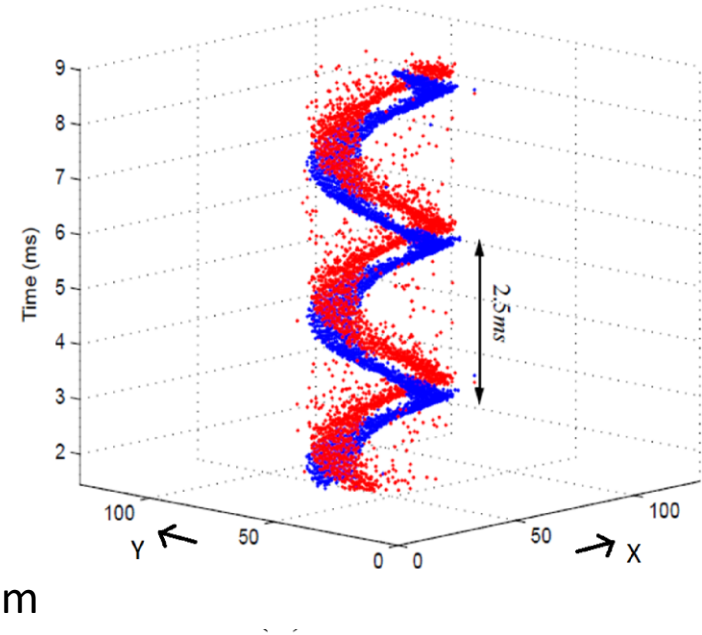
Dynamic Vision Sensors

oświetlenie padające na piksel, który zmienia się jako sinusoidalny kształt fali w czasie z okresem 2,5 ms









skoki wyjściowe generowane przez odpowiedni piksel DVS

Niebieskie **impulsy** odpowiadają dodatnim pikom wyjściowym, które są generowane, gdy oświetlenie wzrasta, podczas gdy czerwone **impulsy** ilustrują ujemne piki ze znakiem generowanym przez malejące oświetlenie w czasie

Wielkoskalowe systemy neuromorficzne

W odniesieniu do neuromorficznego sprzętu do przetwarzania należy odróżnić:

- sprzęt realizujący określone funkcjonalności poznawcze od
- platform sprzętowych SNN ogólnego przeznaczenia, przeznaczonych do emulacji masywnych macierzy neuronowych.

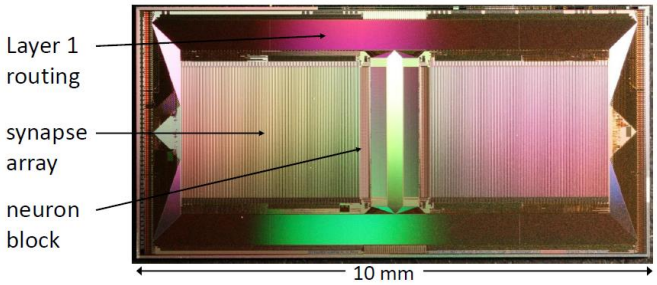
Wielkoskalowe systemy neuromorficzne

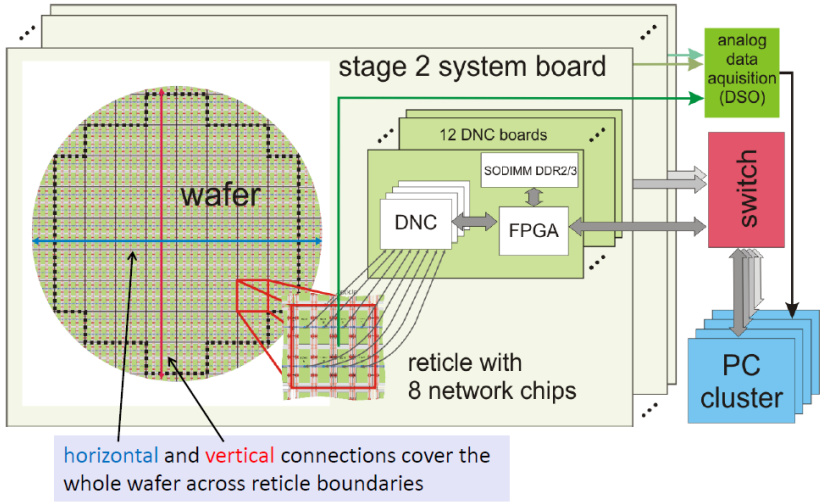
HBP – Human Brain Project – umożliwia prowadzenie biologicznych symulacji w czasie rzeczywistym dla miliona neuronów i ich połączeń synaptycznych

https://www.humanbrainproject.eu/en/

Projekt			Opis
HICANN (FACETS) BrainScaleS	University of Heidelberg	2010	Technologia 180 nm CMOS. Osiem pojedynczych chipów, nazwanych HICANN (High Input Count Analog Neural Network), zawierających po jednym ANC każdy wraz z repeaterami L1 oraz niezbędnymi obwodami pomocniczymi, stanowi jedną siatkę. Ta drobna ziarnistość pozwala na produkcję prototypów w serii MPW. 44 siatki z 352 chipami HICANN mieszczą się na 20 cm płytce krzemu. W siatce poziome i pionowe kanały L1 łączą układy HICANN. Dodatkowa warstwa metalu jest osadzana na wierzchu płytki w etapie obróbki końcowej, który umożliwia wzajemne połączenie poszczególnych siatek o skoku metalu znacznie poniżej 10 µm.
NeuroGrid	Stanford University	2014	12 x14 mm^2, technologiai 180-nm CMOS Neurocore ma 256 krzemowe-neuronowy, nadajnik, odbiornik, router i dwie pamięci RAM. Neuron ma obwody somy, dendrytu, czterech zmiennych bramkujących i cztery obwody populacji synaps (czyli wspólną synapsę i dendryt). Sprzęt składa się z Cypress EZ-USB FX2LP, Lattice ispMACH CPLD, płyty rozszerzenia i 16 Neurocores połączonych w drzewo binarne. FX2 obsługuje komunikację USB. CPLD łączy się między FX2 i Neurocores. Płyta zależna realizuje główne rozgałęzienie aksonów przy użyciu układu FPGA Xilinx Spartan-3E i ośmiu pamięci SRAM Cypress 4MB.
SpiNNaker	University of Manchester	2013	składa się z matrycy 18 identycznych rdzeni ARM968, komunikujących się za pośrednictwem pakietów przenoszonych przez niestandardową strukturę połączeń. Technologia UMC w 130-nm procesie CMOS,
TrueNorth	IBM	2014	28nm CMOS; wymiary jednego rdzenia: 240x390µm^2; 5.4-billion-tranzystorów, 4096 rdzeni neurosynaptycznych połączonych poprzez sieć wewnątrz układu, która łaczy 1 million programowalnych neuronów impulsowych (spiking) i 256 millionów konfigurowalnych synaps
Loihi	Intel	2018	chip składa się z siatki 128 rdzeni neuromorficznych ze zintegrowanym silnikiem uczącym na chipie
Darwin	Zhejiang University, Hangzhou	2017	Darwin Neural Processing Unit to sprzętowy koprocesor z cyfrową logiką zaprojektowany specjalnie do wbudowanych aplikacji o ograniczonych zasobach
ROLLS	ETHZ-INI	2015	256 neuronów i 128 k uczenia on-linesynapsy
DYNAPs	ETHZ-INI	2018	Dynamiczny neuromorficzny procesor asynchroniczny (DYNAP) z 1 tys. neuronów i 64 tys. synaps uczenia się on-line
ODIN	University of Leuven	2019	Cyfrowa realizacja chipa neuromorficznego (ODIN) zawierającego 256 neuronów i 64 K 4-bitowych synaps wykazujących sterowaną kolcami plastyczność synaptyczną w technologii FDSOI 28 nm

HICANN (FACETS)

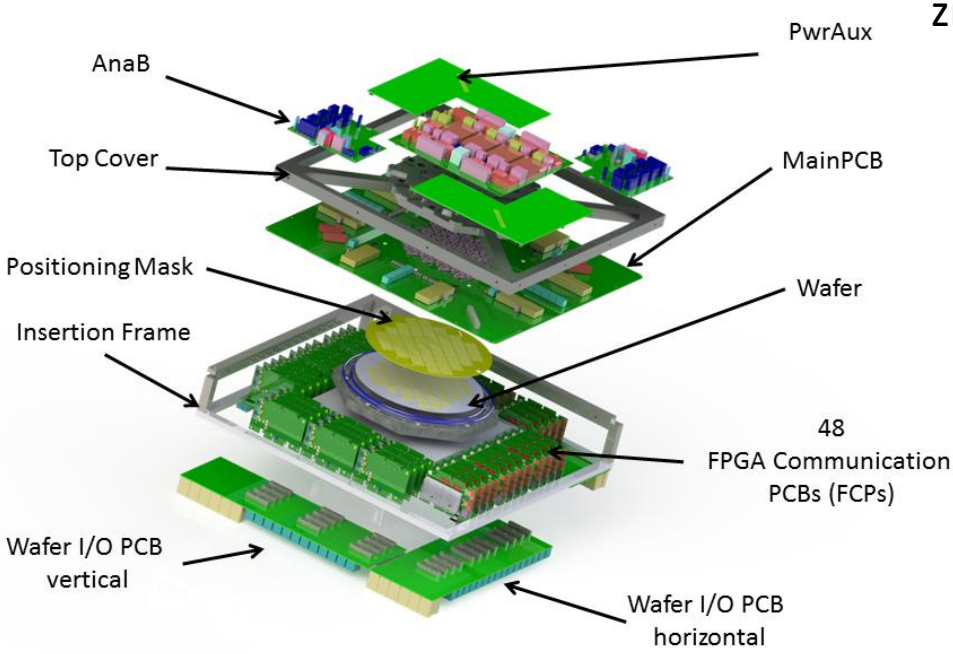




J. Schemmel et al., “A Wafer-Scale Neuromorphic Hardware System for Large-Scale Neural Modeling,” *Proc. IEEE Int’l Symp. Circuits and Systems* (ISCAS 10), 2010, pp. 1947–1950.

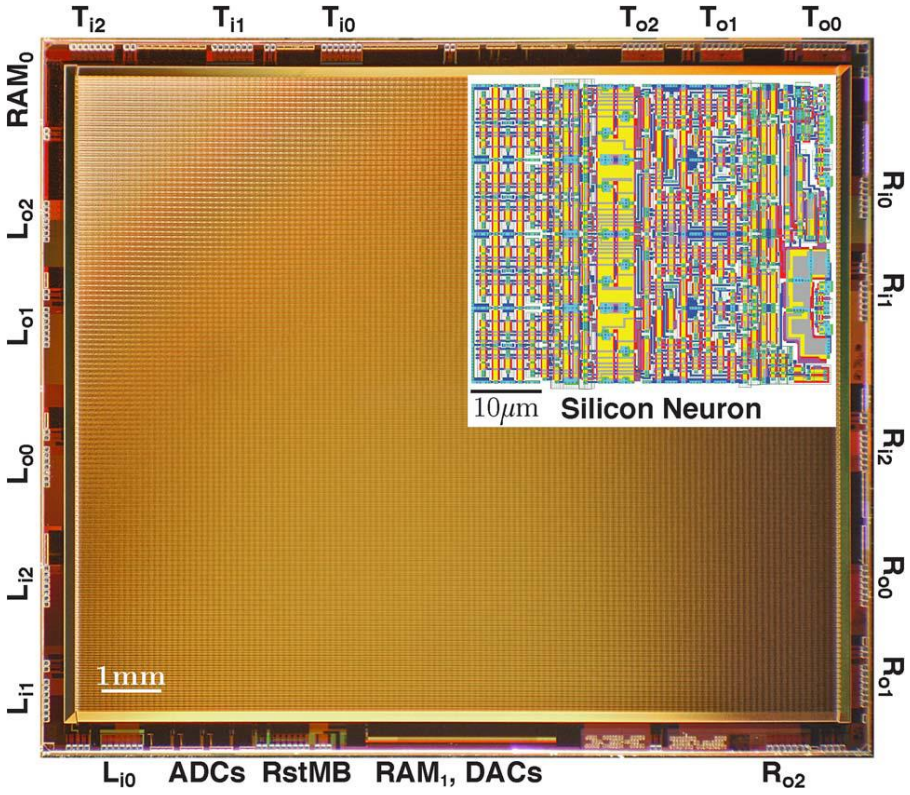
BrainScaleS

Heidelberg, Germany
“BrainScaleS” system,
znany też jako “physical model” lub PM system



System sprzętowy BrainScaleS w skali płytki krzemu: *Wafer* zawierający bloki konstrukcyjne HICANN i infrastrukturę komunikacji *on-wafer*, infrastrukturę mechaniczną (górna pokrywa i ramka do wsuwania), analogowe karty odczytowe (AnaB), zasilacz i cyfrowe moduły komunikacji w obrębie płytki krzemu.

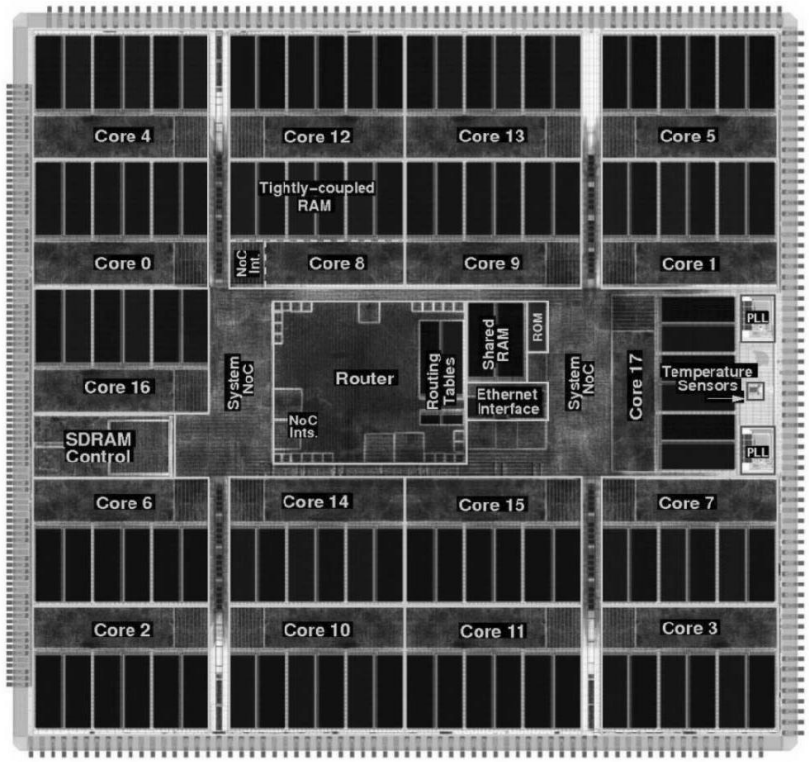
NeuroGrid



12 x14 mm^2,
23 M tranzystorów i 180 padów,
technologiai 180-nm CMOS
Wstawka: topografia jednego neuron - 337 tranzystorów

B.V. Benjamin et al., “Neurogrid: A Mixed-Analog-Digital Multichip System for Large-Scale Neural Simulations,” *Proc IEEE*, vol. 102, no. 5, 2014, pp. 699–716.

SpiNNaker

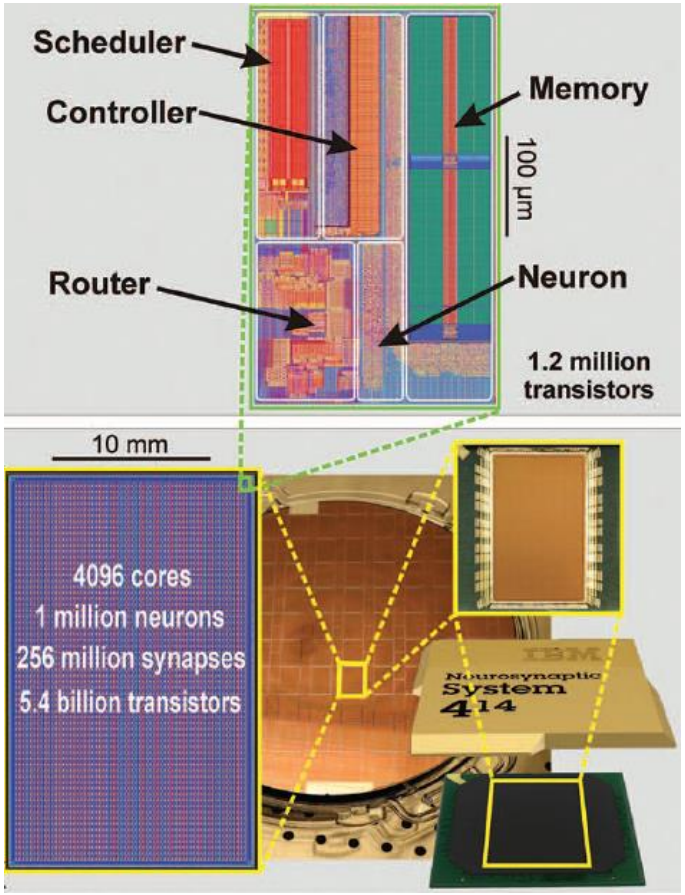


Manchester, United Kingdom
"SpiNNaker" system,
Znany też jako "many core" lub MC system

500.000 core system (NM-MC1) in Manchester on 25 March 2016

S. Furber et al., "Overview of the Spinnaker System Architecture," *IEEE Trans. Computers*, vol. 62, no. 12, 2013, pp. 2454–2467.

TrueNorth



Topografia jednego rdzenia 28nm CMOS mieści się w 240µm x 390µm

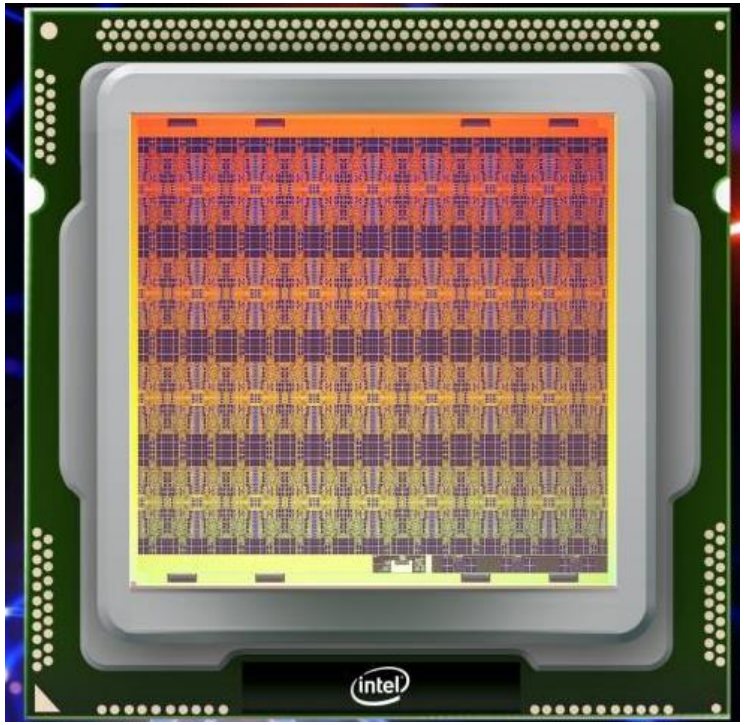
Pamięć SRAM przechowuje wszystkie dane dla każdego neuron;
multipleksowany w czasie obwód neuronu aktualizuje potencjały błon neuronów;
planista (scheduler) buforuje przychodzące zdarzenia szczytowe (spike) w celu implementacji opóźnień aksonalnych;
router przekazuje impulsy;
sterownik (controller) organizuje działanie rdzenia.

Cały układ scalony – matryca 64 na 64,
Płytka krzemu
obudowa

P. Merolla et al., "A Million Spiking-Neuron Integrated Circuit with a Scalable Communication Network and Interface," *Science*, vol. 345, no. 6197, 2014, pp. 668–673.

Loihi - wielordzeniowy procesor neuromorficzny

z funkcją układowego uczenia się



Loihi

wymowa “low-EE-hee”

Chip został nazwany na cześć wulkanu Loihi jako gra słów - Loihi to wyłaniający się hawajski wulkan podmorski, który pewnego dnia wypłynie na powierzchnię.

Parametr	Wartość
Technologia:	Intel 14nm CMOS proces
Powierzchnia całkowita:	$60mm^2$
Powierzchnia matrycy neuro-rdzeni:	$53mm^2$
Liczba tranzystorów:	$2{,}07 \cdot 10^9$ tranzystorów
Pamięć SRAM	33MB
Napięcie zasilania (nominalne):	0.75V
Częstotliwość pracy (rutery w sieci)	860MHz
Częstotliwość aktualizacji neuronów	100MHz

Środowiska programowania sprzętu neuromorficznego

(software frameworks for neuromorphic hardware)

Środowisko	Opis	Strona
PyNN HBP, 2020	Wspólny interfejs programistyczny dla wielu symulatorów zmniejsza lub eliminuje problemy związane z różnorodnością symulatorów, przy jednoczesnym zachowaniu korzyści. PyNN jest interfejsem, umożliwiającym jednorazowe napisanie skryptu symulacyjnego przy użyciu języka programowania Python i uruchomienie go bez modyfikacji na dowolnym obsługiwanym symulatorze takim jak NEURON, NEST, PCSIM, Brian lub na sprzęcie neuromorficzny VLSI BrainScaleS (Heidelberg) lub SpiNNaker (Menchaster) poprzez HBP.	http://neuralensemble.org/PyNN https://flagship.kip.uni-heidelberg.de/jss/FileExchange/HBPNeuromorphicComputingPlatformGuidebook.pdf?fID=1504&s=qqdXDg6HuX3&uID=65
Nengo	Nengo Brain Maker to pakiet Pythona do budowania, testowania i wdrażania sieci neuronowych, który pomaga rozwiązywać problemy przy użyciu najbardziej wydajnego sprzętu dostępnego dla tego problemu. Nengo jest wysoce rozszerzalny i elastyczny. To potężne środowisko programistyczne na każdą skalę: FPGA, Loihi, SpiNNaker, …	https://www.nengo.ai/
TrueNorth Corelets IBM, 2013	Paradygmat programowania kognitywnego: (a) abstrakcja dla programu TrueNorth o nazwie Corelet, reprezentująca sieć rdzeni neurosynaptycznych, która zawiera wszystkie szczegóły z wyjątkiem zewnętrznych wejść i wyjść; (b) zorientowany obiektowo język Corelet do tworzenia, komponowania i dekomponowania coreletów; (c) Biblioteka Corelet, która działa jako stale powiększające się repozytorium rdzeni wielokrotnego użytku, z których programiści tworzą nowe corelety; (d) kompleksowe laboratorium Corelet, które jest środowiskiem programistycznym, które integruje się z symulatorem architektury TrueNorth, Compass, w celu obsługi wszystkich aspektów cyklu programowania, od projektowania, przez programowanie, debugowanie i wdrażanie.	http://meseec.ce.rit.edu/722-projects/spring2015/2-1.pdf

Czwarty element obwodu elektrycznego

Aby przezwyciężyć ograniczenia wynikające z prawa Moore'a i architektury von Neuman'a zaproponowano technologie wykraczające poza CMOS.

1971 – Leon Chua - pierwszy teoretyczny opis (postulat istnienia) **memrystora** jako czwartego elementu pasywnego ustalającego związek między ładunkiem elektrycznym a strumieniem magnetycznym.

Chua, L.O. Memristor—*The Missing Circuit Element*. IEEE Trans. Circuit Theory **1971**, 18, 507–519

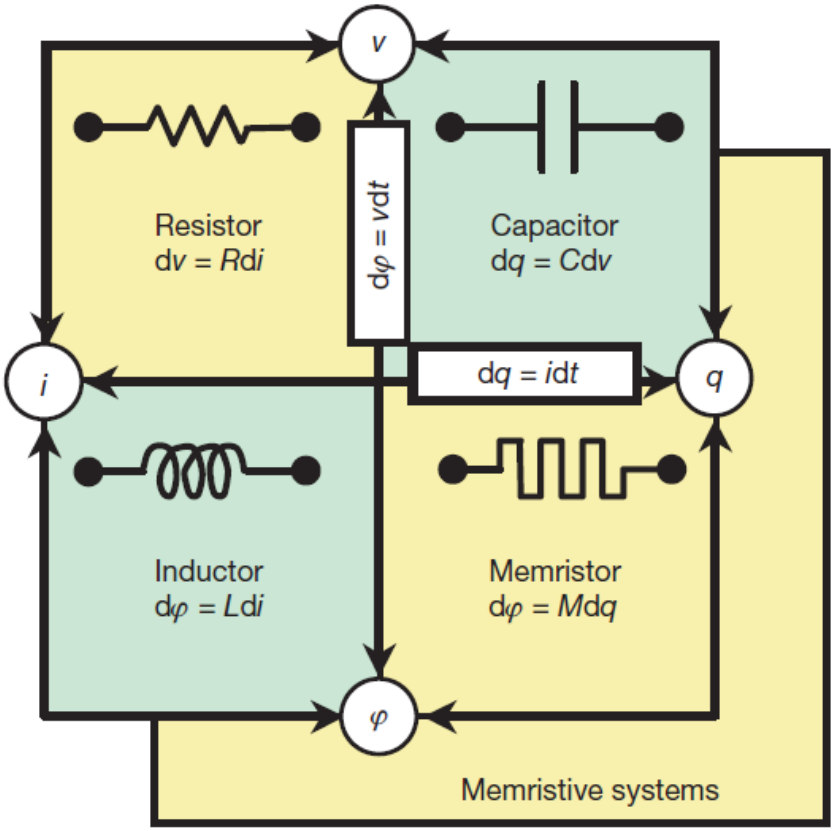
2008 - zespół z HP Labs - pierwszy eksperymentalny memrystor oparty na cienkiej warstwie tlenku tytanu TiO_2

Strukov, D.B.; Snider, G.S.; Stewart, D.R.; Williams, R.S. The missing memristor found. Nature **2008**, 453, 80–83

Memrystor

Funkcjonalności tego pasywnego urządzenia nie można odtworzyć przez jakąkolwiek kombinację podstawowych elementów obwodu dwuzaciskowego - rezystorów R, kondensatorów C i cewek indukcyjnych L.

Dlatego zostało ono oznaczone jako „brakujący element obwodu”.



Memrystancja M jest pochodną strumienia magnetycznego φ po ładunkiem q:

$$M = \frac{d\varphi}{dq}$$

Memrystor

To 2-zaciskowe urządzenie zachowuje się jak rezystor zmienny, którego wartość można modyfikować poprzez przyłożenie określonych napięć lub prądów.

Strukturą tego urządzenia jest połączenie metal-dielektryk-metal MIM, w którym warstwa dielektryczna ma grubość kilku nanometrów.

Przyłożenie pola elektrycznego i kontrolowanych prądów przez dielektryk powoduje zmianę jego rezystancji poprzez wytworzenie włókien przewodzących, które w innym stanie polaryzacji mogą zostać z powrotem usunięte.

Obecnie dostępne memrystory są w większości urządzeniami binarnymi, ponieważ mogą przełączać się między dwiema wartościami rezystancji: stanem wysokiej rezystancji HRS (High-Resistance State) i stan niskiej rezystancji LRS (Low-Resistance State).

Memrystor

Opisano dynamikę uogólnionego systemu pamięciowego matematycznie jak:

$$i = G_M(\vec{x}, i) \cdot v$$

$$\frac{dx}{dt} = f(\vec{x}, i)$$

Gdzie:

v to napięcie,

i to prąd, a

$\vec{x} = (x_1; x_2; :::; x_n)$ to wewnętrzny wektor stanu, który składa się z $n \geq 1$ składników, zwanych zmiennymi stanu. Reprezentują wewnętrzne parametry fizyczne, takie jak temperatura, ciśnienie, stężenie zanieczyszczeń, skład chemiczne itp.

Funkcja f jest ciągłą n-wymiarową funkcją wektorową.

Pierwsze równanie to zależne od stanu, zależne od prądu prawo Ohma.

Drugie równanie to równanie stanu układu dynamicznego.

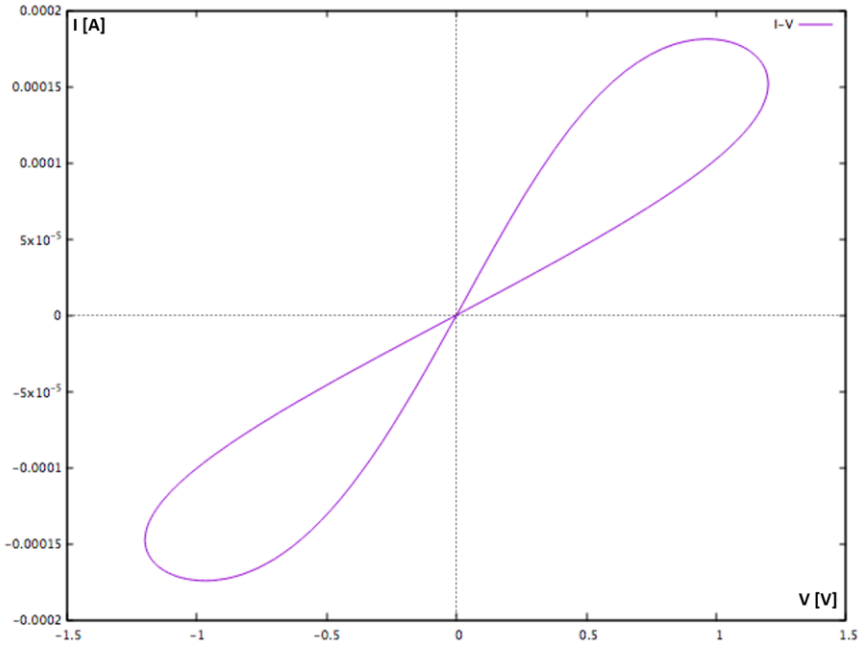
Memrystor

Memrystor powinien wykazywać trzy charakterystyczne cechy:

• zaciśnięta pętla histerezy w płaszczyźnie napięcia v w funkcji prądu i zawsze przechodzi przez początek dowolnego dwubiegunowego układu współrzędnych dla dowolnego okresowego przebiegu $v(t)$,

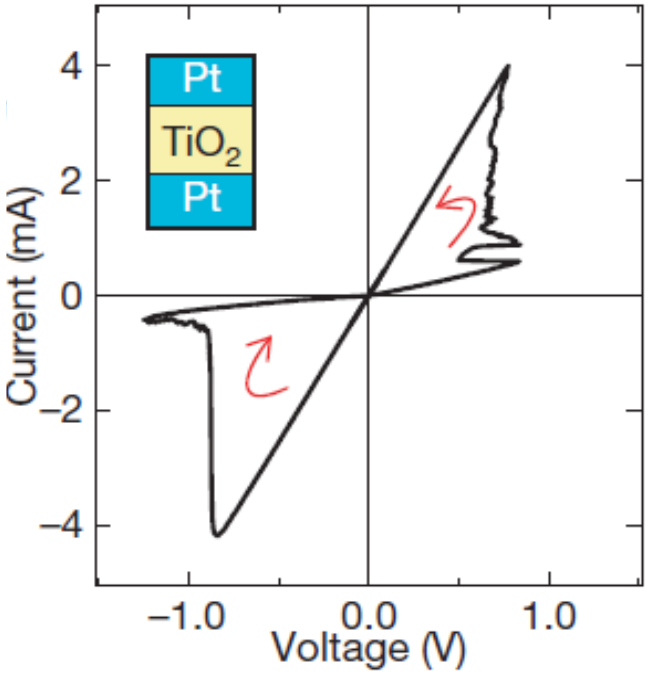
• powyżej krytycznej częstotliwości f_c, obszar zaciśniętego listka histerezy maleje monotonicznie wraz ze wzrostem częstotliwości okresowego sygnału wejściowego,

• kształt zaciśniętej pętli histerezy zmienia się wraz z częstotliwością f i kurczy się do linii prostej, gdy częstotliwość zmierza do nieskończoności.



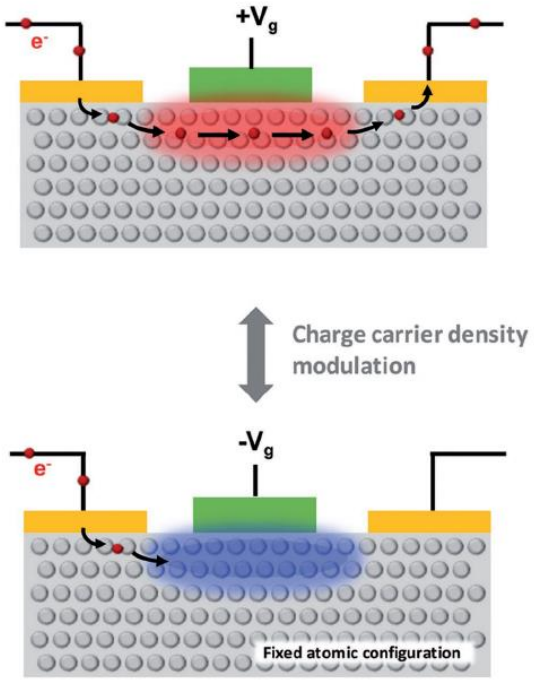
Pierwszy memrystor

Widok przekroju przedstawia trzy warstwy urządzenia: warstwę „magazynującą” wykonaną z dwutlenku tytanu umieszczoną pomiędzy dwiema platynowymi elektrodami. Ta wewnętrzna warstwa pamięci może być dynamicznie rekonfigurowana poprzez stymulację elektryczną, a ta rekonfiguracja tworzy efekt pamięciowy, w którym rezystancja urządzenia zależy od historii przepływającego przez nią prądu. Co najważniejsze, ten zaprogramowany stan nie jest tracony po odłączeniu zasilania.

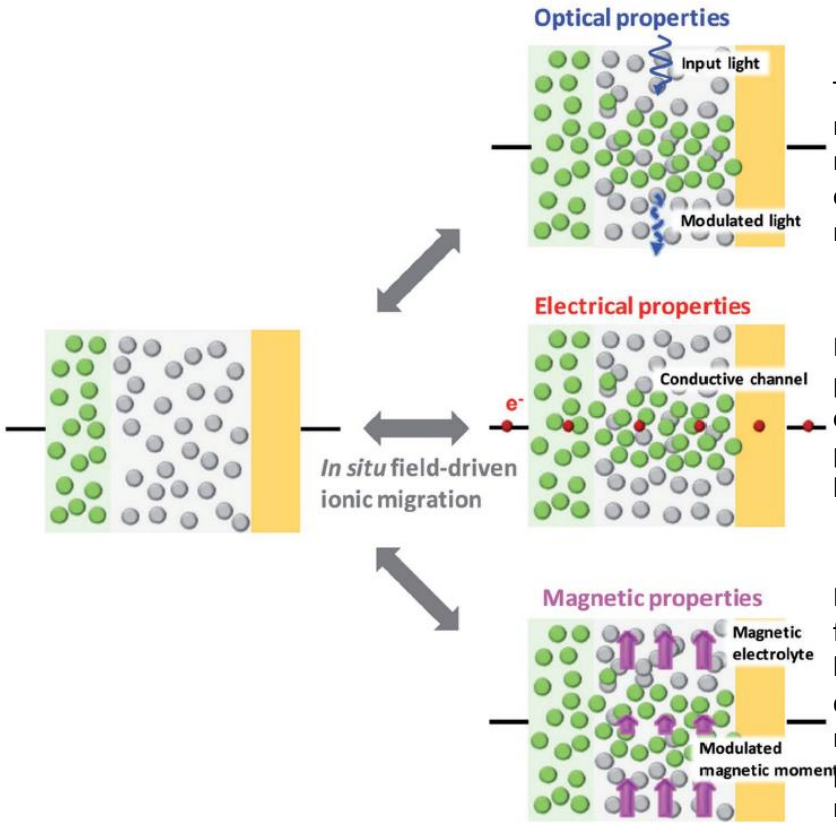


Memrystor

Rekonfiguracja funkcjonalna



Rekonfiguracja fizyczna



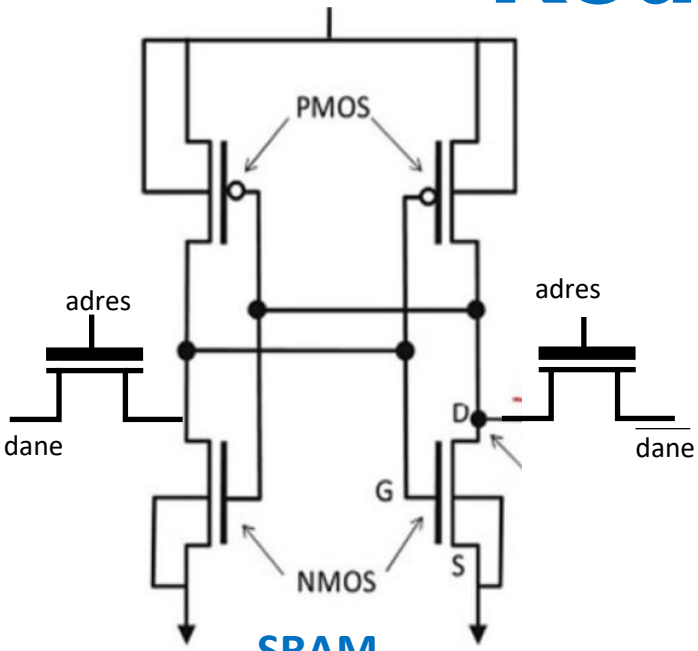
Transmisja przez falowód oparty na dielektryku może być modulowana przez utworzenie elektrycznie uformowanego metalowego włókna

Rezystancyjne zachowania przełączania są powszechnie obserwowane w urządzeniach pamięci opartych na anionach i kationach

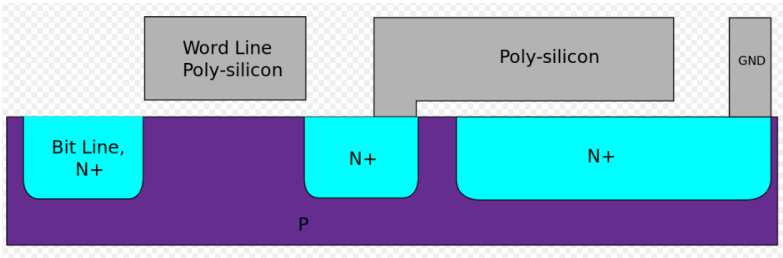
Namagnesowanie materiału ferromagnetycznego może być kontrolowane przez interkalację / deinterkalację jonów, po których następują reakcje redoks, prowadzące do zmian momentu magnetycznego

Działanie **sprzężonych urządzeń jonowo / elektronicznych** wiąże się ze zmianami fizycznymi **opartymi na migracji jonów i reakcjach chemicznych**, umożliwiając in situ kontrolę właściwości elektronicznych, optycznych i magnetycznych materiałów i urządzeń in situ.

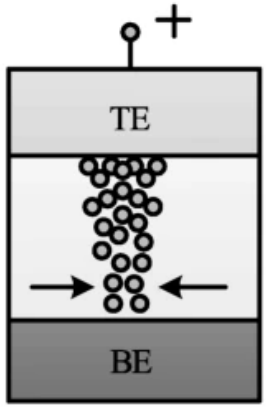
Rodzaje pamięci



SRAM

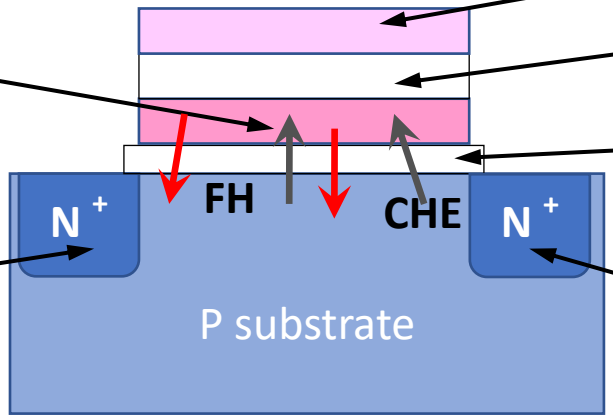


DRAM



RRAM

Falsh



Bramka sterująca, polikrzem

Izolator **blokujący** pomiędzy polikrzemami

Cienki, **tunelowy** tlenek bramkowy

Polikrzemowa, przewodząca warstwa gromadząca ładunek – pływająca bramka

źródło

dren

zapis

wymazywanie

CHE – channel hot electrons
FH – Fowler-Nordheim tunneling

Memrystory oferują:

- doskonałą skalowalność rozmiaru do 2 nm,
- przełączanie szybciej niż 100 ps,
- ilość energii na aktualizację przewodnictwa poniżej 3 fJ,
- długa retencja,
- duży stosunek ON / OFF,
- wielopoziomowe działanie komórek,
- nieniszczące czytanie,
- prosta konstrukcja,
- prawie nieskończona możliwość układania w stosy,
- niska cena,
- świetna kompatybilność z CMOS i możliwość produkcji.

https://knowm.com/

KNOWM

Po co nam memrystor?

- zaproponowano wiele rodzin logicznych opartych na memrystorach do obliczeń cyfrowych
- wykazano ich potencjał dla technologii cyfrowej pamięci o długotrwałej nieulotnej NVM (Non-Volatile Memory)
- obiecująco wygląda również zastosowanie jako urządzeń bio-czujnikowych.

W dziedzinie inżynierii neuromorficznej:

memrystory posiadają szczególną plastyczność, która przypomina reguły adaptacyjne obserwowane w synapsach biologicznych.

memrystory mogą z czasem dostosowywać się i zmieniać swoje zachowanie w odpowiedzi na różne wzorce stymulacji, tak jak dzieje się to w ludzkim mózgu.

Memrystory wykazują biologicznie inspirowane zasady uczenia się przypominające plastyczność zależną od synchronizacji impulsów STDP (Spike-Timing-Dependent Plasticity) obserwowaną w neuronach biologicznych. Dlatego memrystory uznano za sztuczne synapsy nieorganiczne.

Memrystor

Wykazują pewne właściwości szczególnie cenne jako elektroniczne elementy synaptyczne:

- Memrystory można skalować do wielkości poniżej 10 nm.
- Mogą zachowywać stany pamięci przez lata.
- Mogą przełączać się w nanosekundowych skalach czasowych.
- Przechodzą uczenie oparte na skokach w czasie rzeczywistym zgodnie z biologicznie inspirowanymi zasadami uczenia się jak plastyczność zależna od czasu skoku (STDP)

Memrystor

Różne mechanizmy przełączania konduktancji:

- pamięć zmiany fazy **PCM** (Phase Change Memory),
- magnetyczna pamięć o dostępie swobodnym ze spinowym przenoszeniem momentu obrotowego **STT-MRAM** (Spin-Transfer Torque Magnetic Random Access Memory),
- pamięć z mostkiem przewodzącym **CBRAM** (Conductive Bridge Memory),
- pamięć ferroelektryczna **FeRAM** (Ferroelectric Random Access Memory),
- pamięć o dostępie swobodnym z przełaczaniem rezystancyjnym **RRAM** (Resistive switching Random Access Memory),
- organiczne urządzenia pamięciowe **OMD** (Organic Memristive Devices),
- **OxRAM**

Memrystor

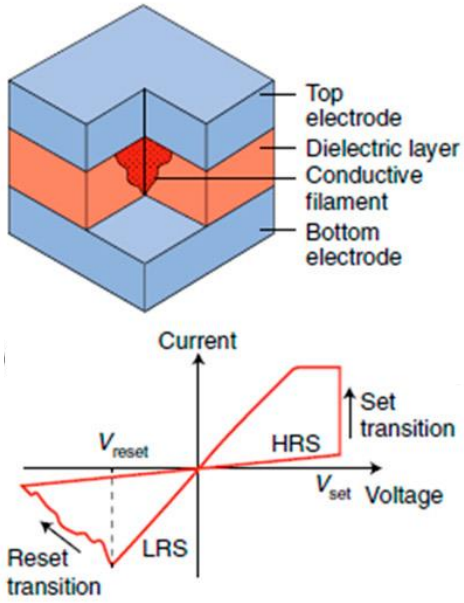
Każde z proponowanych urządzeń ma inną charakterystykę pod względem:

- zwartości,
- niezawodności,
- wytrzymałości,
- czasu przechowywania pamięci,
- stanów programowalnych i
- efektywności energetycznej

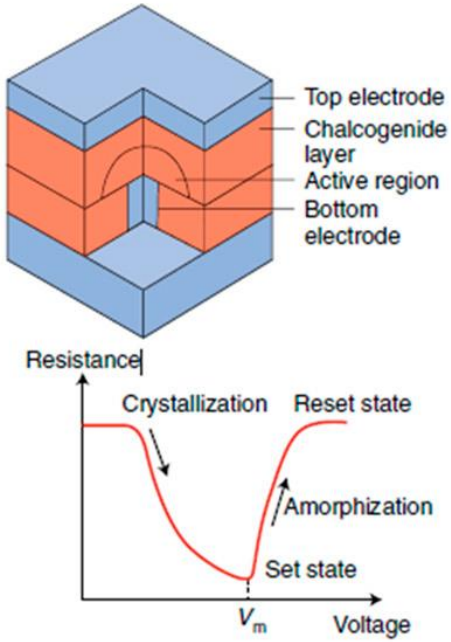
	Memristor	DRAM	SRAM	Flash(NAND)
Cell Density(F^2)	4	6-30	140	4-6
Cell Element	1R/1T1R	1T1C	6T	1T
Retention	>10years	>64ms	Weeks?	>10years
Endurance	$>10^{12}$ cycles	$>10^{16}$ cycles	$>10^{16}$ cycles	$>10^{5}$ cycles
Read Time	<2ns	2ns	0.2ns	0.1ms
Write Time	<10ns	10~50ns	10~100ns	200us

Memrystor

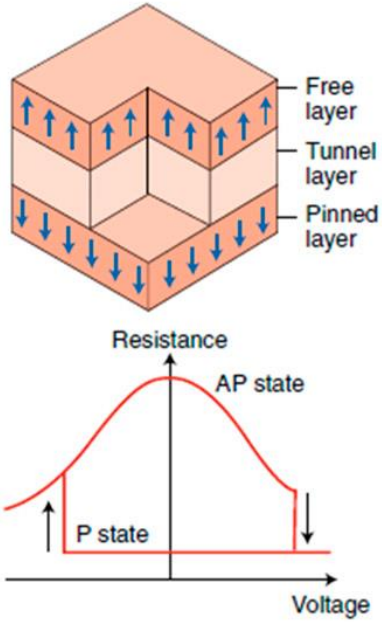
Resistive switching Random Access Memory RRAM



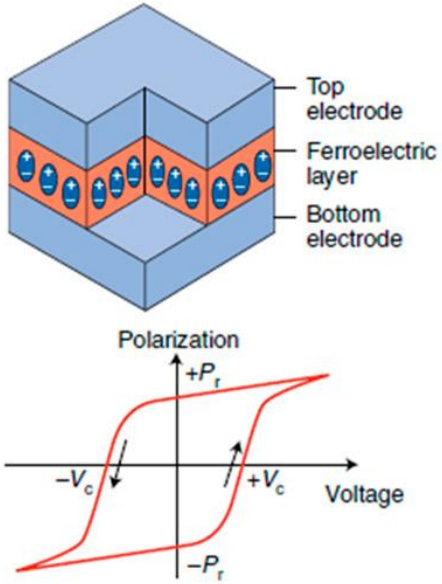
Phase Change Memory PCM



Spin-Transfer Torque Magnetic Random Access Memory STT-MRAM



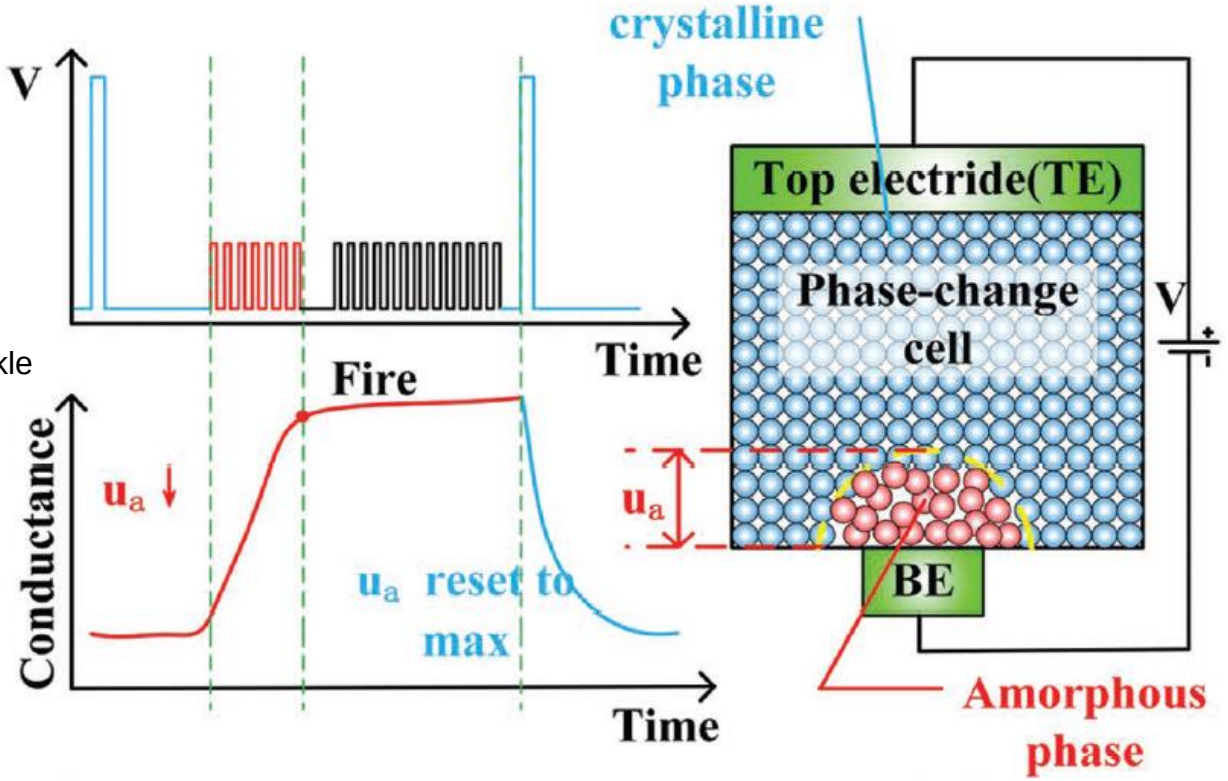
Ferroelectric Random Access Memory FeRAM



Mechanizmy sztucznych neuronów i synaps

PCM - sztuczny neuron

Chalkogenki – nieorganiczne związki chemiczne, w których anionami są chalkogeny (16 grupa układu okresowego), tj. siarczki, selenki i tellurki (zwykle poza tlenkami)

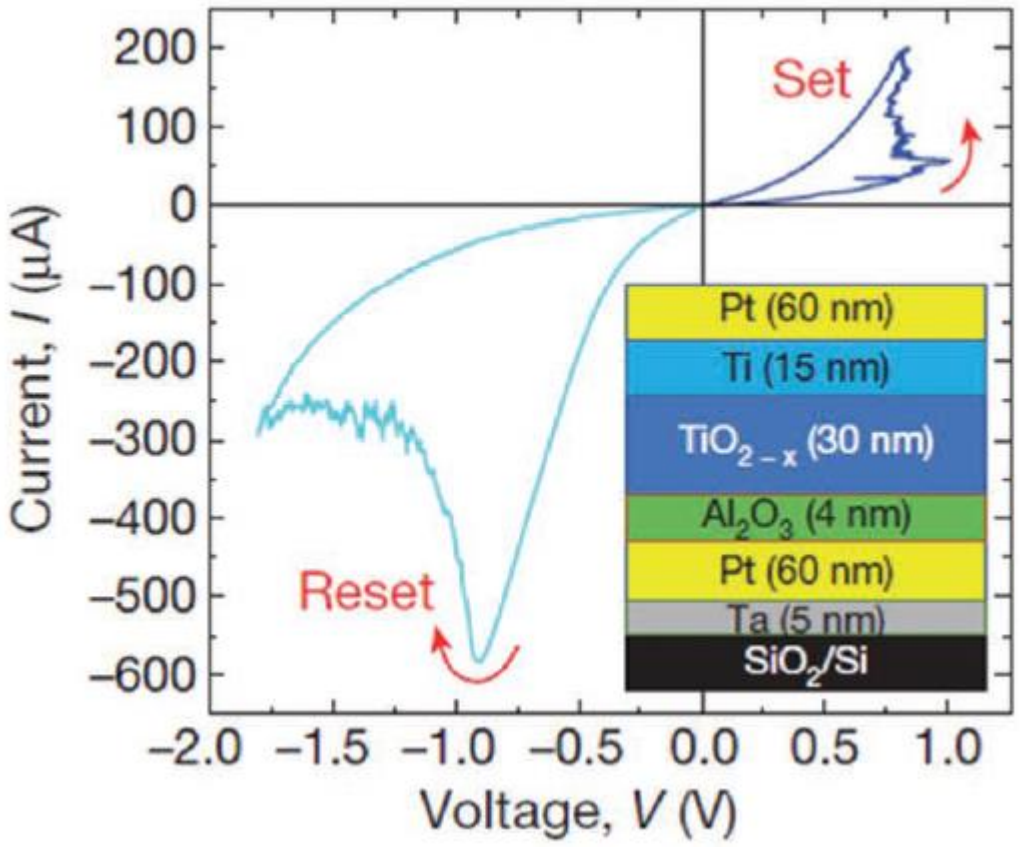


W PCM potencjał membranowy jest prezentowany przez konfigurację fazową warstwy chalkogenidowej.

Integrację czasową potencjałów postsynaptycznych można osiągnąć w skali nanosekundowej dzięki odwracalnym przemianom fazowym z fazy amorficznej do krystalicznej.

Mechanizmy sztucznych neuronów i synaps

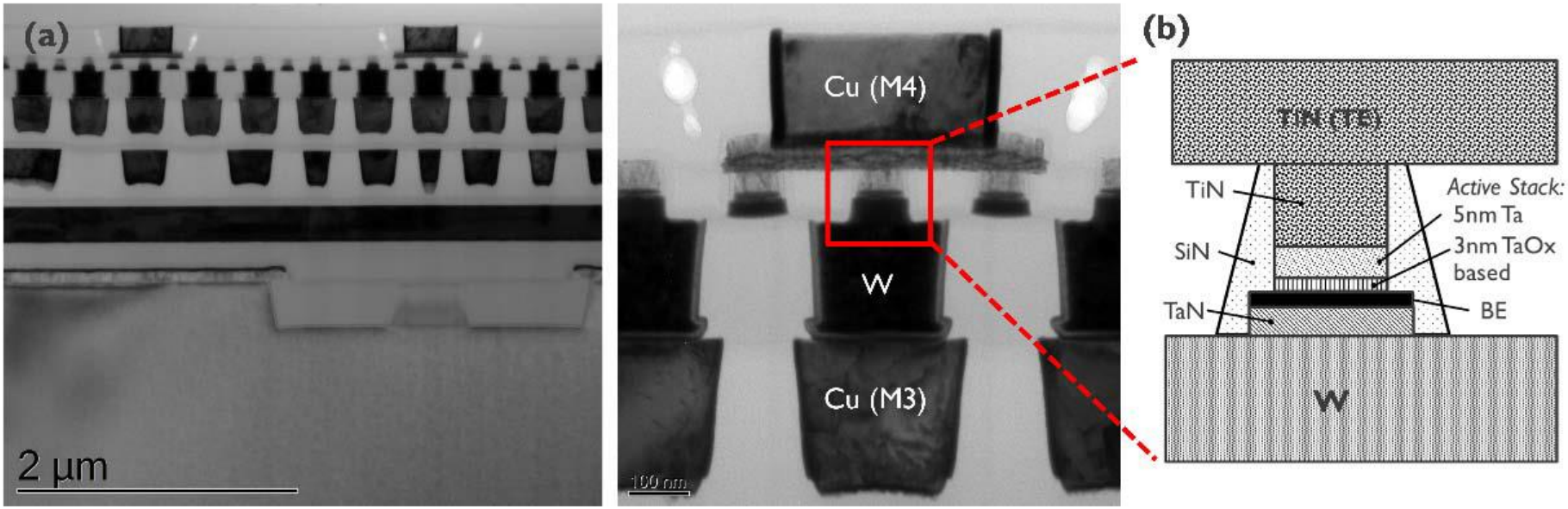
RRAM – sztuczna synapsa



Memrystory RRAM oparte na włóknach tlenowych, wykorzystują tlenki metali, HfO_x, TiO_x, WO_X, TaO_x oraz mieszaniny lub stosy warstw, z których prawie wszystkie są kompatybilne z CMOS.

Mechanizmy sztucznych neuronów i synaps

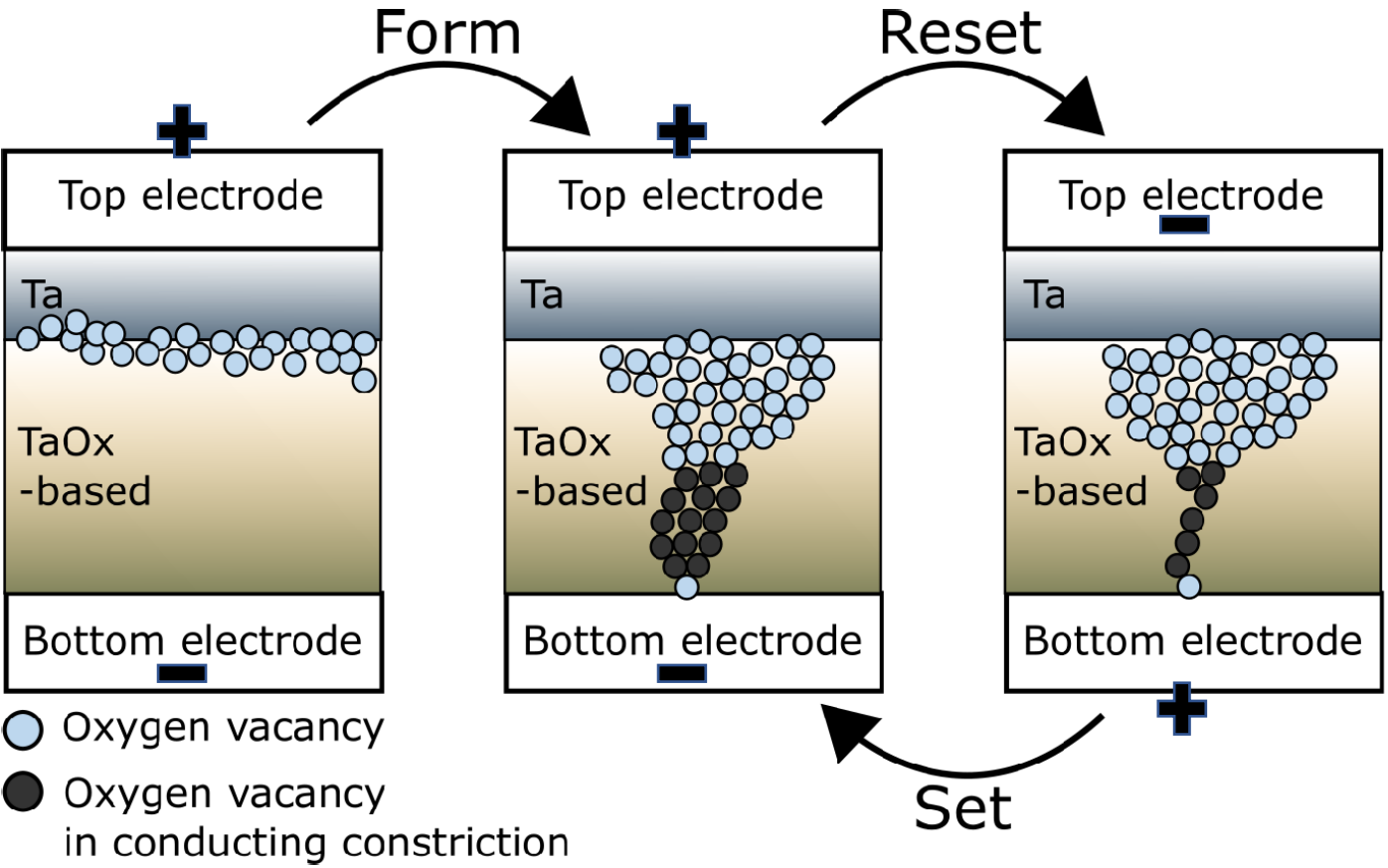
1-Mb macierzy OxRRAM z tlenku tantalu (TaO) w technologii 65-nm CMOS .
Urządzenie OxRRAM zostało zintegrowane pomiędzy miedzianą warstwą 3 i miedzianą warstwą 4.
Aktywny stos składa się z warstwy TaO o grubości 3 nm i warstwy Ta o grubości 5 nm.
Ten aktywny stos znajduje się pomiędzy elektrodą dolną (BE), a elektrodą górną (TE).



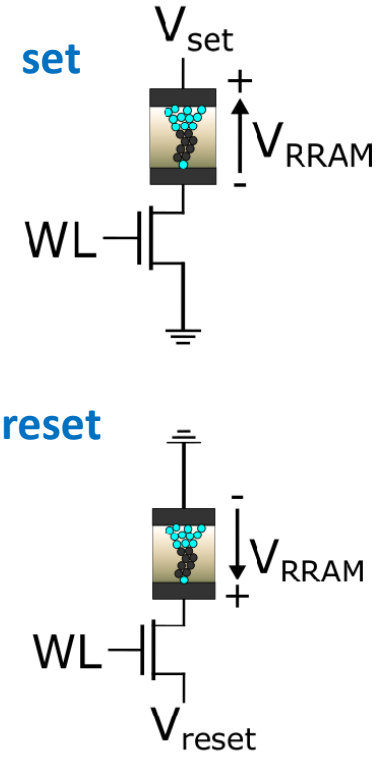
Przekrój TEM części macierzy 1Mbit OxRRAM opartej na TaOx, pokazujący aktywny element pamięci 60 nm i nieaktywne elektrycznie elementy ślepe w odstępie 200 nm

Mechanizmy sztucznych neuronów i synaps

OxRAM – sztuczna synapsa



Warunki pracy układu 1T1R



Zasada działania urządzenia OxRRAM:

po utworzeniu przewodzącego włókna wakansji tlenowych urządzenie można zresetować do HRS i ponownie ustawić do LRS.

Po wyprodukowaniu urządzenia są nieprzewodzące. Aby utworzyć ścieżkę przewodzącą, urządzenia są kształtowane galwanicznie przez przyłożenie impulsu wysokiego napięcia (3,3 V) o długości 100 ns do stosu MIM. Ze względu na wysokie napięcie defekty tworzą przewodzące włókno między TE i BE.

Ważne jest narzucenie ograniczenia prądowego przez ograniczenie napięcia bramki VG tranzystora w strukturze 1T1R podczas formowania w celu ograniczenia szerokości przewodzącego włókna. Bez tego ograniczenia tworzy się szerokie i wysoce przewodzące włókno, które trudno zresetować.

Po uformowaniu za pomocą odpowiedniego prądu, urządzenie jest kasowane RESET przez zastosowanie ujemnej polaryzacji (-1,5V), która usuwa wakansje tlenowe z przewodzącego włókna do górnego zbiornika wakansji. Powoduje to wyższą rezystancję zwaną stanem wysokiej rezystancji (HRS). Na koniec, urządzenie można ponownie ustawić SET do stanu niskiej rezystancji (LRS) przez przyłożenie dodatniego napięcia (1,5V), niższego niż napięcie formowania, które ponownie wprowadza wakansje z górnego zbiornika wakansji do przewodzącego włókna. Wszystkie operacje SET i RESET są wykonywane przez zastosowanie impulsów 100 ns. Ważne aby tą samą wartość prąd stosować podczas formowania i podczas operacji SET, aby zapewnić, że nie zostaną utworzone żadne dodatkowe wakansje.