### Plně hardwarová neuronová síť v podobě křemíkového čipu od IBM



**V minulém roce se společnosti IBM povedlo prakticky vyrobit a odzkoušet první opravdu komerčně snadno realizovatelný čip realizující hardwarovou neuronovou síť. Pojďme se blíže podívat na strukturu a funkci čipu a proč je tento projekt tak významný a přelomový pro budoucnost výroby inteligentních systémů.**

Doposud většina počítačových systémů, procesorů a mikrokontrolérů vyznává centralizovanou architekturu založenou na jednom či dnes i více výkonných výpočetních jádrech, jedné velké paměti dat a jednoho společného rozhraní vstupů a výstupů. I když struktury s více jádry i vícenásobným přístupem do paměti již umožňuji jakousi malou možnost paralelního zpracování dat, stále však současné procesory řízená zařízení zpracovávají a vyhodnocují úlohu či řídí proces prostřednictvím sekvenčně vykonávaného programu, tedy postupně příkaz po příkazu, kde doposud získaná či vypočítaná data dosazují do dalších výpočtů.

Náš mozek však pracuje odlišně. Ten tvořený rozsáhlou sítí neuronů vzájemně propojených prostřednictvím synapsí naopak vykazuje velmi distribuovanou strukturu, kde každý jednotlivý neuron sice tvoří jen malou a poměrně jednoduchou výpočetní jednotku a s ním související synapse jen malou paměť. Jejich vzájemné propojení v počtu mnoha miliard však dává v celku velmi výkonnou a efektivní řídící, vyhodnocovací i paměťovou jednotku. Protože každý neuron pracuje samostatně, vykazuje celá jejich síť velmi vysoký paralelizmus zpracování dat, který je prakticky omezován jen přísunem informací od snímacích prvků, tedy u živočichů od smyslových orgánů. Díky tomu pak v kritické situaci mozek dokáže velmi rychle reagovat a vyhodnotit možnosti nebo současně řídit více činností i při „pomalejším taktování“. Rychlost práce jednotlivých neuronů se odehrává v řádu jednotek až stovek ms.

Další rozdíl našeho mozku od počítačů je fakt, že jeho funkce je prakticky založena jen na hardwaru, který se svojí funkcí (učením) sám mění. Mění se stav a význam propojení neuronů a prostě celá síť se neustále modifikuje. „Hardware“ jde zde tedy velmi dynamická jednotka, zatímco hardware současných počítačů je až na paměť prakticky statický s pevně danou strukturou, kde každá cesta má svojí pevně danou funkci. Propojení a důležitost jednotlivých bloků je pevně „zadrátovaná“ a je tedy nutné použít nadřazený program, který definuje, jak se tato struktura využívá. Samotný počítačový hardware jsou vlastně jen předměty, které se využívají na základě stavu vykonávání příkazů softwaru vytvořeného naším mozkem, zatímco neuronová síť je sama o sobě hardwarem i programem v jednom. A přesto je velmi univerzální.

### Obecný popis funkce neuronové sítě

Neuronová síť, ať již v přírodní podobě mozku, nebo uměle realizovaná, je tvořena strukturou vzájemně propojených „řídících jednotek“ zvaných neurony a husté sítě propojení zvané synapse. Ty zároveň i realizují funkci distribuované paměti.

Neurony mezi sebou komunikují prostřednictvím předávání pulsních zpráv zvaných „spikes“, typické generované s frekvencí v řádu kHz. „Spike“ se vysílá na výstup (axon) generujícího neuronu a následně přes synapsi na vstup (dendrit) následujícího napojeného neuronu. Každý neuron má tzv. membránový potenciál, jehož hodnota se mění (zvýší nebo sníží) na základě příchodů spike signálů a zároveň jeho hodnota klesá s časem (funkce zapomínání). Jestliže hodnota membránového potenciálu překročí určitou hodnotu, neuron vygeneruje a vyšle svůj spike signál dalším neuronům k vyhodnocení a jeho membránový potenciál se „vynuluje“. Zároveň spike signál slouží k úpravě (adaptaci) úrovně synapsí prostřednictvím režimu učení.



*Ukázka živé neuronové sítě a jejich vzájemné předávání zpráv pomocí impulsu přes synapsi.*

### Hardwarová neuronová síť od IBM

Do tohoto úkolu se s velkým úsilím vrhla společnost IBM ve spoluprácí se společností DARPA. Aktuálním výsledkem je jeden z prvních opravdu prakticky fungujících a použitelných čipů realizující hardwarovou neuronovou síť. Proti softwarově realizovaným neuronovým sítím běžícím na dnes běžné počítačové architektuře typu „Von Neumann“, které vyžadují velmi vysoký výpočetní výkon procesoru i velmi rychlý přístup do paměti, hardwarově realizovaná síť si vystačí s výrazně menšími výpočetními prostředky, spotřebou el. energie a výrazně menší rychlostí hardwaru. Přesto dokáže vyřešit stejně náročné úlohy. Dnes se velmi triviální softwarové neuronové sítě (v porovnání se sítí mozku i nejmenších živočichů) využívají pro rozpoznávání ručně psaných znaků, předpovědi trendů v bankovnictví, automatické dolaďování řídích jednotek automobilů apod. A právě velká výpočetní náročnost softwarového řešení neuronových sítí doposud znemožňovala jejich výraznější masové použití např. v řízení výroby, regulaci strojů apod. 

Navíc IBM vyvinutá struktura čipu je vyrobitelná plně jen z křemíku a stejnými technologiemi, jako se využívá pro výrobu běžných současných procesorů a integrovaných součástek. Konkrétně 45 nm technologii SOI (Silicon On Insulator) a elektronické komponenty na čipu využívají strukturu obvodů CMOS. Navíc funkce čipu je prakticky celá digitální (číslicová), na rozdíl od předchozích pokusů, které se snažily až příliš kopírovat běžný mozek a kvůli jeho analogové funkci kombinovaly na jednom čipu analogové a digitální systémy, což není z pohledu rychlosti funkce, spotřeby el. energie ani velikosti čipu příliš výhodné řešení.

Oficiálně se součástka nazývá „kognitivní čip (cognitive chip)“ a aktuálně první realizované vzorky obsahují 256 neuronů v poli 16 x 16 neuronů a až 256 tisíc učených synapsí. To vše při celkové velikosti čipu 3 x 2 mm a velikosti jednoho neuronu 35 x 95 mikrometrů.



*IBM vyrobený reálný křemíkový neuronový čip byl již i úspěšně prakticky použit a testován na různých aplikacích přes USB připojení na PC.*

To je však jen začátek. Cílem společnosti je totiž vytvoření systému s 10 miliardami neuronů a až 1000 miliardami synapsí, který spotřebuje jen 1 kWh elektřiny. Takový systém by již byl schopen realizovat i složitější funkce a aplikace, které je schopen aktuálně vykonat jen mozek. I když stále je to málo například proti 109 neuronům a 1013 synapsím kortexu mozku kočky.

### Struktura kognitivního čipu IBM

Základem čipu jsou „výpočetní centra“ nazývaná „neurosynaptická jádra (neurosynaptic core)“. Každé jádro obecně obsahuje *N* počet neuronů, *K* počet axonů a tím související *KxN* počet synapsí. Každé jádro je pak obecně schopné realizovat různé známé algoritmy neuronových sítí pomocí různého vzájemného propojení výstupů neuronů se vstupy synapsí a následně vstupy dalších neuronů. Například dosud největší IBM vyrobený čip obsahuje 256 neuronů a 1024 axonů, tedy počet synapsí má rozměr 1024 x 256.



*Bloková struktura neurosynaptického jádra a 256 neurony a synaptickou maticí rozměru 256 x 256.*

Synapse jsou zde realizovány jako speciální matice paměťových buněk typu SRAM. Protože však každý řádek reprezentuje jeden axon předchozího napojeného neuronu a každý sloupec jeden dendrit následujícího napojeného neuronu, musí mít paměť nezávislý „oboustranný“ přístup, tedy umožňuje nezávisle adresovat jednotlivá paměťová místa (paměťové buňky) ze strany řádků i ze strany sloupců. Každá paměťová buňka je tak tvořena strukturou s 8 tranzistory – viz obrázek níže. Na základě učícího algoritmu jsou pak stavy „1“ a „0“ jednotlivých buněk měněny tak, aby ve výsledku počet aktivních synapsí na každém dendritu (vstupu neuronu) násobený hodnotou velikosti kroku dával potřebnou analogovou hodnotu v neuronu, tzv. potenciál či napětí membrány (dle skutečného napětí, které se nachází v živém neuronu).

 

*Synaptická matice se vždy nachází mezi axony předcházejících neuronů a dendrity následujících neuronů a podle jejich počtu pak definuje rozměr matice. Jednotlivé body matice tvoří paměťové buňky typu SRAM, které však svojí konstrukcí umožňují přístup a zápis ze směru řádků i sloupců. V případě uspořádání 256 x 256 buněk může každý vstup neuronu nabývat hodnoty v rozsahu 0 až +/- 256 (hodnoty typu axonu 1 nebo 2 definuje kladnou či zápornou hodnotu).*

Jednotlivé neurony jsou pak tvořeny následujícím blokových schématem postaveným na funkčním modelu označovaném jako „integrate-and-fire“. Jeho první část prakticky hlavně slouží k převodu binárního kódování synapse prostřednictvím hodnoty kladného nebo záporného kroku *s+* a *s-* na analogovou hodnotu v mozku odpovídající „napětí membrány“, která dále slouží pro porovnání s nastavenou rozhodovací úrovní. Výsledek porovnání pak definuje, zda se jedná o zpracování nového podnětu a tedy má se informace předat učícímu algoritmu, nebo ne.

Konkrétně s každým časovým krokem (taktování funkce neuronové sítě) se k předchozí hodnotě „napětí membrány“ ve sčítačce přičte či odečte nová hodnota odpovídající aktuálnímu nastavení synapsí a buď kladného nebo záporného kroku *s+* a *s-* a pak také hodnota „*lambda*“ reprezentující funkci zapomínání. Vzniklá nová hodnota „napětí membrány“ se v komparátoru porovnává s nastavenou rozhodovací úrovní. Pokud je vyšší, generuje se logický signál ve struktuře čipu označovaný jako „spike“ (špička), který se vysílá buď na výstup z čipu nebo do dalšího napojeného jádra a pak také zároveň na vstup navazujícího bloku učení, který je součástí každého neuronu. Pokud se spike signál vygeneruje, vynuluje se hodnota vnitřního „napětí membrány“, tedy vynuluje se stav sčítačky neuronu. Hodnoty kroků *s+*, *s-*, *lambda*a rozhodovací úroveň jsou parametry nastavené a ovlivňované uživatelem čipu („programátorem“).



*Blokové schéma neuronu typu "Integrate and Fire" i s blokem automatického učení "Learning" typu STDP. Prostřednictvím vstupního nastavitelného kladného a záporného kroku (s+ a s-) a sčítačky vzniká signál "napětí membrány", který se porovnává s přednastavenou rozhodovací úrovní (Threshold). Pokud je úroveň překročena, generuje se výstupní signál neuronu.*

Blok učení, který bezprostředně následuje za funkční části neuronu, realizuje časově závislý učící algoritmus STDP (Spike-Timing Dependent Plasticity). Ten na základě sledování změny hodnoty „napětí membrány“ v čase, přesněji řečeno délku časových úseků mezi jednotlivými spike signály, mění stavy ve sloupcích synaptické matice před neuronem a také v řádcích matice následující hned za neuronem, na který je axon (výstup neuronu) napojen (tzv. pre-synaptic a post-synaptic array). Proto je u paměťové matice synapsí nutný již výše uvedený „oboustranný“ přístup ze strany dendritů (pro pre-synaptic matici) i axonů (pro post-synaptic matici). Algoritmus STDP podle nejnovějších výzkumů velmi dobře modeluje funkci učení mozku.

Konkrétně je blok učení každého neuronu složen ze dvou 8bitových čítačů časových kroků (taktů) *C+* a *C-* pro nezávislé řízení pre-synaptic a post-synaptic matice, které počítají čas od poslední vygenerované špičky (spike), tedy od posledního překročení hodnoty „napětí membrány“. Při každém časovém taktu se pak jejich hodnota porovnává s hodnotu z bloku generátoru pseudonáhodných čísel realizovaného zpětnovazebním posuvným registrem LFSR (Linear Feedback Shift Register) a na základě výsledku se aktualizují stavy sloupců pre-synaptické, resp. řádků post-synaptické paměťové matice. Tím se kognitivní čip sám učí (mění obsah své paměti na základě zjištění nových událostí).

### Celkový popis funkce jádra

Na vstup každého jádra se postupně v časových krocích (taktování jádra) v řádu jednotek kHz přivádějí ze vstupů čipu či neuronů z dalších jader na čipu stále nové informace v podobě tzv. spike zpráv (spike messages). Ty jsou dekódovány ve podobě AER hodnoty (Address-Event Representation), která přenáší informaci o aktivních oblastech v synaptické matici a typu axonu (typ excitatory aktivují v neurony kladný krok s+, typ inhibitory aktivuje záporný krok *s-*). Tyto zprávy tedy musí být nejdříve dekódovány a pak jsou aktivovány příslušné axony (řádky synaptické matice). Na základě rozložení stavů synapsí v maticové struktuře a jejich počtu v jednotlivých řádcích a sloupcích se na dendrity napojených neuronů generuje vstupní signálů neuronů.

Ty jej na základě nastavených řídících parametrů (kroků, zapomínání, rozhodovací úroveň atd.) vyhodnotí a buď vygenerují nebo nevygenerují další řídící spike signál (hodnota 0 či 1). Pokud se spike signál vygeneruje, tak se pak společně s výstupy z ostatních neuronů opět zakóduje do AER spike zprávy, která tvoří výstup z jádra pro axony dalších navazujících jader. Výstupní signál každého neuronu se také zároveň zavádí do bloku učení, který je součástí každého neuronu. Ten následně provádí aktualizaci / úpravu stavů příslušného sloupce paměťové matice přecházející danému neuronu a příslušného řádku matice napojené na výstup stejného neuronu.

### Výsledky provedených testů

První vyrobené vzorky byly samozřejmě hned podrobeny různému testování, například na aplikaci rozeznání zdeformovaných čísel nebo silně poškozených geometrických útvarů. 

V prvním případě byla z neurosynaptických jader vytvořena dvouvrstvá neuronová siť typu RBM (Restricted Boltzmann Machine), která byla natrénovaná na ručně psané číslice 0 až 9 v podobě digitalizovaného obrazu. Konkrétně na široké sadě reprezentované 60 tisíci předložených obrazů. Následné testování předložením 10 tisíc různých náhodně vybraných obrazů číslic prokázalo 94% úspěšnost správného určení číslice.

V druhém případě byly nejdříve neuronové síti pro naučení předkládány celé různé základní geometrické obrazce typu obdélník, kříž, čtverec a poté následně předkládány stejné obrazce, ale různě zdeformované či neúplné. Z výstupu sítě pak „padaly“ správně rozlišené celé obrazce.



*Synaptická matice naučené neuronové sítě na tvary obrazců je plná různých konfigurací možných obrazců s nimiž se následně porovnává předložený neznámý (poškozený) obrazec a hledá se co nejvíce společných znaků.*

### Závěr

Hardwarové neuronové sítě v podobě kognitivních čipů jistě mají světlou budoucnost. Rozhodně efektivnost jejich funkce je výrazně větší než realizace softwarových neuronových sítí na klasické architektuře běžných CPU. Dle mého názoru budou v budoucnu spolu koexistovat „současná“ provedení procesorů s architekturou typu Von Neumann i kognitivní čipy a v mnoha zařízeních budou spolu vzájemně spolupracovat či budou dokonce spolu integrovány v jedné součástce. Obě architektury totiž mají svoje výhody (rychlé předem definované zpracování velkých objemů dat vs. schopnost rychlé reakce a učení) a hodí se pro jiné typy aplikací. Tedy obě technologie si přímo nekonkurují, ale naopak se mohou vhodně doplňovat. Mohou tak vytvořit budoucí moderní superprocesor. Pro realizaci budoucích skutečně inteligentní strojů a zařízení se pak bude jistě hodit spojení výhod obou systémů.

### Odkazy

* Paul Merolla, John Arthur a další: "A Digital Neurosynaptic Core Using Embedded Crossbar Memory with 45pJ per Spike in 45nm", IBM Research - Almaden, Cornell University
* Jae-sun Seo, Bernard Brezzo, Yong Liu a další: "A 45nm CMOS Neuromorphic Chip with a Scalable Architecture for Learning in Networks of Spiking Neurons", IBM T. J. Watson Research Center, IBM Research - Almaden
* <http://www.popsci.com/technology/article/2009-11/digital-cat-brain-runs-blue-gene-supercomputer>
* <http://www.popsci.com/technology/article/2011-08/first-generation-cognitive-chips-based-brain-architecture-will-revolutionize-computing-ibm-says>
* <http://www.popsci.com/technology/article/2011-01/how-engineers-build-computers-work-mind>
* <http://lowercolumbia.edu/students/academics/facultyPages/rhode-cary/intro-neural-net.htm>
* <http://pcworld.cz/hardware/nahlednete-s-nami-do-budoucnosti-procesoru-3-dil-21031>
* <http://www.lohninger.com/kohonen.html>
* <http://signalnews.com/ibm-cognitive-computing-chip-brain-688>
* <http://www.ibm.com/smarterplanet/us/en/business_analytics/article/cognitive_computing.html>

### Články o samoučících algoritmech na HW serveru

* Článek o GMBH sítích: [Článek o GMBH sítích: http://automatizace.hw.cz/gmdh-neuronova-sit-filtr-pro-digitalni-zpracovani-signalu](http://automatizace.hw.cz/gmdh-neuronova-sit-filtr-pro-digitalni-zpracovani-signalu)
* Článek o LMS a ALMS učících algoritmech: <http://automatizace.hw.cz/clanek/2005053001>
* Článek o Kalmanově filtraci: <http://automatizace.hw.cz/clanek/2007042901>
* Článek o neuronových sítích typu Kohonenovy mapy: <http://automatizace.hw.cz/clanek/2006051401>