浅议类脑计算硬件和器件

——我希望的芯片和计算机

一、 介绍

人类对机器智能的探索最早可以追溯到图灵，在学科建立的初期，图灵就已经高瞻远瞩，对智能科学就提出了很高的要求与长远的愿景——未来的智能系统能够像人一样思考[1]，这或许就是类脑技术最初的萌芽。

人脑是人类进化的结晶，从单层感知机到深度学习，神经网络借鉴了神经元、突触连接的基本概念、采用了人脑类似的复杂神经元网络，一代代科学家尽可能在当前主流计算机系统中搭建出一个近似于人脑的处理架构。但这仅限于概念相似，神经网络的学习方式基于“反向传播”，是依靠数学计算完成的，这与人脑的运作方式大相径庭。同时，目前几乎所有的人工智能系统都需要首先进行人工形式化建模，转化为一类特定的计算问题，这与人脑的运作方式有差异，终归无法胜任大脑般的通用任务。

事实上，类脑问题最终实现是受制于长久以来积累的架构鸿沟——主流计算机的模型主要由图灵机和冯·诺依曼体系结构演化而来，但图灵机的本质是使用预定义的规则对预定义的输入符号进行处理、而冯·诺依曼架构是存储程序式计算，程序也是预先设定好的，无法根据外界的变化和需求的变化进行自我演化。[2][3]或许从一开始，传统计算机就与脑科学走向了不同的方向。

综上，当前利用传统硬件解决类脑问题存在三个硬件瓶颈：

1. 传统 CPU、GPU 受限于冯·诺依曼结构，并行度优势无法完全发挥；
2. 通用传统硬件结构是提前设定好的，无法临时编辑，不够灵活，在选择通用性的同时放弃了定制化的优势；
3. 传统硬件功耗过高。

因此，类脑计算应运而生。它是借鉴生物神经系统信息处理模式和结构的计算理论、体系结构、芯片设计以及应用模型与算法的总称。不同于冯·诺依曼存算分离的特性，类脑计算硬件基于仿生的脉冲神经元实现信息的高效处理，具有低功耗、低延迟的技术优势，是打破“内存墙”的潜在技术之一，其在对功耗、延迟敏感的边缘计算领域具有广泛的应用价值和潜力。同时，由于类脑计算的特殊性，当前主流芯片采用定制化策略，进一步提升了效率，也成功降低了功耗。

本文主要探讨类脑计算机的核心——类脑芯片。

二、 类脑芯片的历史

类脑芯片根据不同目的，借鉴不同层次脑科学成果，提出不同计算架构，大致可分为两条发展路线——侧重脑仿真模拟、侧重智能场景应用。前者可以大规模并行，但是应用场景单一；后者可以应用算法加速计算（SNN 等），但是本身很难实现突破。

NeuroGrid、BrainScales、SpiNNaker 是典型的侧重脑仿真模拟芯片器件。NeuroGrid 采用模拟/数字混合设计，使用亚阈值模拟电路实现神经元和突触动力学，可以模拟 1 百万神经元和 60 亿突触连接，功耗较低；BrainScales 使用高阈值的模拟电路完成模数混合，可模拟 20 万神经元和 4 千万突触，功耗很高；SpiNNaker 是专门设计于脉冲神经网络，基于 ARM 处理器的大规模并行，功耗较低。[4][5][6]

而侧重智能场景应用方面，早在 2011 年，IBM 公司就通过模拟大脑结构，首次研制

出两个具有感知认知能力的硅芯片原型，可以像大脑一样具有学习和处理信息的能力。这便是第一代类脑芯片，它的每个神经元都是交叉连接，具有大规模并行能力。但 IBM 指出：该类芯片“脑容量”仅相当于虫脑的水平。

基于第一代芯片，在 2014 年，IBM 公司推出名为“TrueNorth”的第二代类脑芯片。与第一代类脑芯片相比，“TrueNorth 芯片性能大幅提升。其神经元数量提高 3000 余倍；可编程突触数量提高近千倍；每秒可执行 460 亿次突触运算，总功耗是第一代类脑芯片的百分之一；而且“TrueNorth”处理核体积仅为第一代类脑芯片的十五分之一。[7]

除了 TrueNorth 外，英特尔 Loihi 芯片、高通 Zeroth 芯片、西井科技 DeepSouth 芯片、浙大“达尔文”类脑芯片、AI-CTX 芯片也都是优良的类脑侧重智能场景应用芯片。

值得作为里程碑的是，2019 年清华大学开发出了全球首款异构融合类脑计算芯片“天机芯”。该芯片结合了类脑计算和基于计算机的机器学习，这种融合技术有望发挥基于计算机科学的人工神经网络和基于神经科学的脉冲神经网络的优势，促进人工通用智能的研究和发展。[8]这种设计是划时代的，它将业界争论不休的 ANN 和 SNN 取长补短，将计算智能和人脑智能结合在一起。

截止如今，类脑芯片的发展取得了非凡的成就，但是依旧处于起始阶段。

三、 类脑芯片的原理

TrueNorth 芯片被设计为适配 SNN 神经网络的硬件，因此先描述 SNN 神经网络的基本原理。与人工神经元不同，脉冲神经元之间的交流通过二进制事件，而不是连续的激活值。如图 1 所示，S_0代表上个神经元传递的一个一个的脉冲，通过突触传递到树突的位置，并且最终由细胞体来处理这些脉冲，细胞体在时间维度上积累脉冲，达到阈值之后输出。

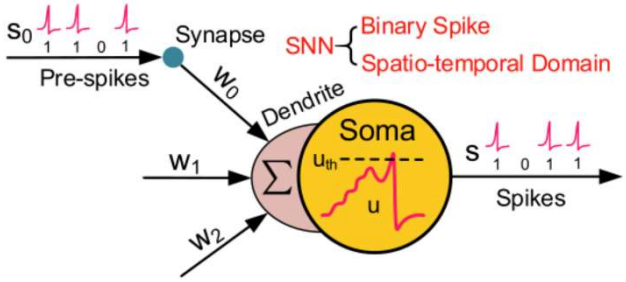

图 1. SNN 结构图

具体数学形式可以由下式得到。

$$\tau\frac{du(t)}{dt} = -[u(t) - u_{r1}] + \sum \omega_j \sum K\left(t - t_j^k\right)$$

$$s(\mathrm{t}) = 1, u(\mathrm{t}) = \mathrm{U_{r2}}, if\ u(\mathrm{t}) \geq u_{th}$$

$$s(\mathrm{t}) = 0, if\ u(\mathrm{t}) \leq u_{th}$$

式中t代表时间步长，u和s代表膜电位和输出峰值，U_{r1}和U_{r2}是静息电位和重置电位，w_j是第j个输入突触的权重，u_{th}是代表神经元是否被激活一次的阈值。[9]

在此基础上，TrueNorth 芯片突破了传统的冯·诺依曼架构，存储和计算相互融合，采用纯数字电路实现了大规模、高集成度、低功耗的 SNN 硬件平台。一个 TrueNorth 芯片包含 4096 个核，每个核都与自己的核以及东、西、北、南方向的四个相邻路由器进行通信，形成一个二维网状网络。同时，相比于全天候工作的数据中心芯片来说，TrueNorth 是由事件驱动的芯片。即假如没有需求，那么它就处于关闭状态。

事实上，其余类脑芯片虽核心思想相似，但亦有其独特的设计之处，不过为了简便起见，本文暂时略去，仅讨论典型的 TrueNorth 芯片——它的内存、CPU 和通信部件是集成在一起的。因此信息的处理完全在本地进行，而且由于数据量并不大，传统计算机内存与 CPU 之间的冯·诺依曼瓶颈便不复存在了。相比于传统硬件，在部署 SNN 神经网络时，神经元之间可以方便快捷地相互沟通——只要接收到其他神经元发过来的脉冲，这些神经元就会同时动作。同时，由于其特殊的事件驱动式构造，芯片保证了极低的功耗。[7]

四、 我希望的类脑芯片与计算机

类脑芯片的性能受到准确率、吞吐率、延迟、能效、功耗、硬件成本影响，因此我希望从这些方面展开我的构想。

首先对于准确率而言，这是一个芯片最根本的特性，但是当前的各芯片都有所保障，我认为可以借鉴他们的方法。不过准确率方面，浮点数运算是一个很关键的环节，在设计类脑芯片时或许可以考虑更多浮点数精度问题。

对于吞吐率和延迟，其主要限制在于冯·诺依曼架构下，内存与 CPU 交互过于频繁，且速度受到“内存墙”的限制。因此我认为需要采取存算一体的新架构，使得输入电压立即输出电流，一瞬间完成矩阵向量乘法，同时提高吞吐率、降低延迟。事实上，这种思想在相机的成像原理上就有应用，即 CCD 和 CMOS。如图 2 所示，前者先将光电子储存起来，一步步输出；后者接收到光电子后直接输出电压，计算速度更快。

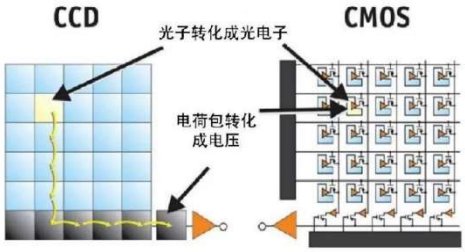

图 2.CCD 和 CMOS 的区别

但是这种实现并非易事，甚至使用场景存在局限性，因此还可以采用众核去中心化数据流架构，并行多个核心且每个核有自己独立的存储单元，降低核心与内存的交互时间，同时使用流水线技术，虽然不可以降低延迟，但是可以提高处理吞吐量。

能效、功耗也是芯片设计的一个重点。除了降低核心与内存的交互次数，还可以采用更先进的芯片制造技术。光电智能芯片也许是一种新的突破，Nature 杂志 2023 年最新发布了有关光电混合模拟计算芯片架构的文章[10]，并且登上了当期封面。它的思想与“天机芯”是不谋而合的——即将传统与创新结合，“天机芯”将传统的人工神经网络和最新的脉冲神经网络结合在一起，反应产生了更好的效果；而如图 3 所示，这个光电混合模拟计算芯片以图像处理为例，从光计算特征提取出发，省略了大量卷积网络，再辅以电计算，完成了一整套图像分类问题。论文数据表明，这个芯片突破光电计算领域多年瓶颈难题，在实验性能媲美电子卷积神经网络的同时，端到端算力相比顶尖显卡 A100 提升三千倍，能效提升百万倍，同时具备可重构自校准能力。

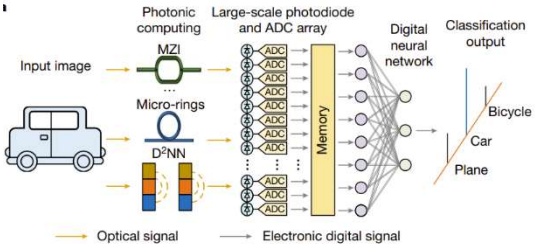

图 3. 光电混合模拟计算芯片原理

因此，我想将光电混合芯片与“天机芯”相结合，或许在一定任务上可以取得更优的效果。做些类比的畅想，如果人作为碳基生物，脑用以碳做载体的电信号进行计算，那么计算机芯片作为一种硅基“生物”，它的“脑”或许更适用于以二氧化硅、硅晶为载体的光信号计算。

硬件成本似乎与存算一体有些许冲突。芯片面积越大，制造成本越高，但是若要将芯片封装成小区块，则又陷入了类似“内存墙”的通信限制中。这需要做一些取舍，但我认为技术的发展是互相促进的，随着光刻技术的不断发展，芯片小型化会愈加显著，成本会进一步降低。

综上，从计算、存储、通信的角度来思考我希望的芯片。计算方面，我认为可以采用光电结合、人工-脉冲结合的方式，以更专一的方法去处理特定的任务。储存方面，我认为可以采用存算一体的结构，减少计算过程中与内存的交互。通信角度，我认为可以采用去中心化路由网络进行通信，在特定区块还可以辅以光信号交互通信。

回顾开篇所提到的三个类脑计算的问题，存算一体结构可以突破冯·诺依曼架构的壁垒、专用的芯片设计可以突破不够灵活的劣势、而我光电结合、人工-脉冲结合的方式，理论上可以大幅降低某些专门任务的功耗。因此可以认为是一个合理的设计。

事实上，计算机的核心在于芯片，我对计算机的畅想大致便是我对类脑芯片的畅想。不过我认为，计算机应当处理更多通用化的内容，而我之前希望的类脑芯片则更趋向于专一化。但我想，这并不完全受制于硬件，合理的算法也是密不可分的一部分。相信在不久的将来，通用的类脑芯片或许可以承“天机芯”之威，继续在融合的道路上不断突破，把芯片与传统计算机的 SSD 等硬件更好地结合在一起，真正实现一个高效、低耗能、通用的类脑计算机。

五、总结

类脑芯片的优势在于，它一定程度上突破了冯·诺依曼架构下的“内存墙”壁垒，是人类在追求算力极限时的一条“救生艇”。本文从传统芯片的问题入手，回顾了类脑芯片的历史，简单介绍了部分类脑芯片的原理，提出了我对类脑芯片以及类脑计算机的遐思——即追求融合，在信号介质层面融合、在硬件算法导向上融合、在多个新兴成熟芯片间融合。没有任何一款计算机是完美的，但是我相信在科学家的不断努力下，我们能不断消除不完美，一步一步奔向下一个局部最优。

展望未来，在追求算力极限时，除了类脑计算机，量子计算机也可以一定程度上实现这一目标，因此或许将类脑计算机和量子计算机融合创新也并不是一个坏选择，但我的类脑计算机故事讲完了，这就让无数躬耕的天才去继续谱写吧。

六、后记

在本学期的《类脑计算和类脑计算系统技术》课程中，我收获颇丰，在课上的深入浅出和课后的阅读钻研中，我了解到了很多计算科学的前沿知识，也将许多之前了解过的零散知识串联了起来。非常感谢施老师、赵老师、邓老师的耐心指导，也非常感谢助教团队的辛勤付出。希望今后有机会再与你们相遇！

参考文献

[1] Turing A M .Computing Machinery and Intelligence[J].American Association for Artificial Intelligence, 1995.DOI:10.1007/978-1-4020-6710-5_3.

[2] Turing A .On Computable Numbers, with an Application to the Entscheidungs problem[J].Alan Turing His Work & Impact, 1937, s2-42(1):13-115.DOI:10.1112/plms/s2-42.1.230.

[3] Von Neumann J. The Computerand the Brain. New Haven, USA:YaleUniversityPress,1958

[4] Khodagholy D, Gelinas J N, Thesen T, et al. NeuroGrid: recording action potentials from the surface of the brain[J]. Nature neuroscience, 2015, 18(2): 310-315.

[5] Schmitt S, Klähn J, Bellec G, et al. Neuromorphic hardware in the loop: Training a deep spiking network on the brainscales wafer-scale system[C]//2017 international joint conference on neural networks (IJCNN). IEEE, 2017: 2227-2234.

[6] Furber S B, Galluppi F, Temple S, et al. The spinnaker project[J]. Proceedings of the IEEE, 2014, 102(5): 652-665.

[7] Akopyan F, Sawada J, Cassidy A, et al. Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip[J]. IEEE transactions on computer-aided design of integrated circuits and systems, 2015, 34(10): 1537-1557.

[8] Pei J, Deng L, Song S, et al. Towards artificial general intelligence with hybrid Tianjic chip architecture[J]. Nature, 2019, 572(7767): 106-111.

[9] Tavanaei A, Ghodrati M, Kheradpisheh S R, et al. Deep learning in spiking neural networks[J]. Neural networks, 2019, 111: 47-63.

[10] Chen Y, Nazhamaiti M, Xu H, et al. All-analog photoelectronic chip for high-speed vision tasks[J]. Nature, 2023, 623(7985): 48-57.