**类脑计算芯片突触实现关键技术**

类脑计算芯片旨在模仿生物大脑的结构和功能，以实现高效、低功耗的并行计算。突触作为神经元之间连接的桥梁，其功能和实现方式是类脑计算芯片设计的核心。近几年，类脑计算芯片在突触实现的关键技术上取得了显著进展，接下来主要围绕突触存储器的类型、突触权重存储机制、突触权重精度以及突触可塑性这四个方面展开。

一、突触存储器类型

突触存储器是类脑芯片中用于存储突触权重的核心器件。其选择直接影响芯片的功耗、面积、速度和可扩展性。近几年，非易失性存储器因其在存储密度、功耗和与CMOS工艺兼容性方面的优势，成为主流研究方向。

电阻式随机存取存储器 (RRAM/ReRAM) 仍然是研究热点。RRAM通过施加电压脉冲改变材料的电阻状态来存储信息，其电阻值可以模拟突触权重。其优势在于结构简单、尺寸小、读写速度快、非易失性以及与CMOS工艺良好的兼容性。【1】提出了一个基于CMOS-RRAM的nvTCAM电路，它将RRAM与传统CMOS工艺结合使用，相较于现有的CMOS/NVM设计，使得读取延迟改善了36%。

铁电场效应晶体管 (FeFET) 作为一种新兴的非易失性存储器，近年来也受到广泛关注。FeFET利用铁电材料的剩余极化来调制晶体管的沟道电导，从而实现非易失性存储。FeFET具有低功耗、高读写速度、良好的耐久性和与CMOS工艺的兼容性等优点，特别适合作为模拟突触。目前有很多工作【2】【3】，利用FeFET提高了运算速度和精度、降低功耗和面积，展示了其在高性能神经形态计算发展方面的巨大潜力。

尽管非易失性存储器是主流，但传统的CMOS技术仍然在某些类脑芯片中发挥作用，尤其是在需要高速度和灵活性的场景。例如，SRAM或DRAM结合数模转换器（DAC）可以实现突触权重存储。比较著名的几款神经拟态芯片，例如Loihi、天机芯、Darwin【4】等，仍使用的是传统SRAM或DRAM存储器。

二、突触权重存储机制

突触权重存储机制是指突触器件如何将神经网络中的权重值映射并保持在其物理状态中。这通常涉及到模拟存储和数字存储两种方式，而模拟存储是非易失性存储器实现类脑计算的关键优势。

模拟权重存储，这是非易失性存储器（如RRAM、FeFET）的核心优势。这些器件能够将突触权重直接存储为其物理参数（如电阻、电导或电荷）的连续或多级模拟值。【】

虽然模拟存储是类脑计算的理想目标，但在一些混合信号或全数字类脑芯片中，仍然会采用数字方式存储权重。权重以数字形式存储在SRAM或DRAM中。在计算时，这些数字权重通过数模转换器（DAC）转换为模拟电压或电流，然后与输入信号进行模拟乘法运算。  
为了兼顾模拟存储的能效和数字存储的精度，一些研究探索了混合权重存储机制。例如，将权重的高位（MSB）以数字形式存储，而低位（LSB）以模拟形式存储在非易失性存储器中。或者，在训练阶段使用高精度数字权重，而在推理阶段将权重量化并存储到低精度的模拟器件中。Polaris23芯片【】便是以类似的方式采用分档位存储策略：将权重划分为4-bit数字值（基础权重）叠加模拟型增量（Δ±5%），兼顾稳定性和精细调节能力。

三、突触权重精度

突触权重精度是指单个突触能够存储和表示的权重值的精细程度。在传统的数字计算中，权重通常以32位或16位浮点数表示，具有很高的精度。然而，在类脑计算中，权重精度通常较低。

商用芯片如IBM TrueNorth【】采用低比特量化技术，使用1–4bit权重，在损失一定权重精度情况下，降低运行功耗。新型类脑芯片（如Loihi 2）【】在突触后端引入树突区脉冲积分单元，通过时空累积效应补偿单个突触的低精度缺陷，该方案可支持百亿参数类脑树突大模型训练。

四、突触可塑性

突触可塑性是生物大脑学习和记忆的基础，它指的是突触连接强度（权重）根据神经元活动模式而改变的能力。在类脑计算芯片中，实现高效、生物启发的突触可塑性是实现片上学习和自适应能力的关键。

近几年，研究主要集中在以下几种突触可塑性规则的硬件实现。

① 脉冲时序依赖可塑性 (STDP)： STDP是生物学中最常见的突触可塑性形式之一。它根据突触前后神经元脉冲的相对时序来调整突触权重。如果突触前神经元在突触后神经元之前发放脉冲，则权重增强（LTP）；反之，如果突触后神经元在突触前神经元之前发放脉冲，则权重减弱（LTD）。【5】提出了一种用于大规模SNN神经形态处理器的基于地址的数字STDP单元。该STDP单元消除了不必要的解码过程，并通过使用发射神经元的地址降低了硬件成本。在非易失性存储器（如RRAM、FeFET）上，STDP的硬件实现通常是利用其电导调制特性。通过施加特定时序和幅度的电压脉冲对（模拟突触前和突触后脉冲），来改变器件的电阻/电导。例如，RRAM可以通过正负电压脉冲的组合来模拟LTP和LTD。FeFET则可以通过门电压和源漏电压的组合来调节沟道电导【】。

②

* 速率依赖可塑性 (Rate-Dependent Plasticity, RDP)： RDP根据突触前后神经元的平均发放速率来调整权重。例如，如果两个神经元都以高频率发放脉冲，它们的连接可能会增强。
  + 硬件实现： RDP通常通过积分电路来累积神经元的活动，然后根据累积的活动量来更新突触权重。这可以通过模拟电路或混合信号电路来实现。
* 监督学习规则： 除了无监督的STDP和RDP，类脑芯片也需要支持监督学习规则，如反向传播（Backpropagation, BP）算法。
  + 硬件实现： 直接在模拟域实现完整的反向传播非常困难，因为它需要精确的梯度计算和权重更新。目前主流的方法是：
    - 混合信号方法： 在片上进行前向传播和部分反向传播（如局部梯度计算），而全局梯度计算和权重更新则在数字域或外部处理器中完成。
    - 近似反向传播： 开发更适合硬件实现的近似反向传播算法，例如基于局部信息的权重更新规则，或者利用器件的物理特性来近似梯度计算。
    - 基于忆阻器的BP： 利用忆阻器阵列进行矩阵向量乘法，并探索其在反向传播中计算梯度乘积的潜力。

片上学习与训练：【6】  
近两年，片上学习（On-chip Learning）是类脑计算领域的一个重要趋势。这意味着神经网络的训练过程直接在芯片上进行，而不是在外部高性能计算设备上完成训练后再将权重部署到芯片。片上学习可以显著减少数据传输，降低功耗，并实现真正的自适应和在线学习能力，这对于边缘计算、机器人和物联网设备等应用至关重要。

* 挑战： 片上学习对硬件的要求更高，因为它需要支持双向的权重更新（读写操作），并且要处理器件的非理想性对训练收敛性的影响。此外，实现大规模网络的片上训练仍然面临功耗和面积的挑战。

新型可塑性机制：  
除了经典的STDP，一些研究也探索了更复杂的生物启发可塑性机制，如突触标签（synaptic tagging）、突触稳态可塑性（homeostatic plasticity）等，以提高网络的学习效率和稳定性。这些机制的硬件实现通常需要更复杂的电路设计和对器件特性的更精细控制。

[参考文献]

【1】CMOS-RRAM Based Non-Volatile Ternary Content Addressable Memory (nvTCAM)

【2】Enhanced FeFET Performance for Energy Efficient Neuromorphic Computing at Cryogenic Conditions

【3】FeFET-Based Neuromorphic Architecture with On-Device Feedback Alignment Training

【4】**Darwin3: A large-scale neuromorphic chip with a Novel ISA and On-Chip Learning**

【5】Efficient Hardware Implementation of STDP for AER Based Large-Scale SNN Neuromorphic System

【6】Recent Advances in on-Chip Learning with Organic Neuromorphic Circuits

Neuromorphic