大家好，我是高鹤箫。今天向大家介绍的是这篇论文：Demonstrating the Advantages of Analog Wafer-Scale Neuromorphic Hardware 生物神经网络模拟的挑战与类脑硬件的优势。接下来我将逐渐向大家解释什么是“类脑硬件以及其优势”。

AI 已经渗透到我们生活的方方面面，但创造和使用 AI 都会需要大量的算力，其训练和推理离不开计算硬件的支持。目前业界主流的 AI 推理硬件分为右边这三种，CPU，GPU 和 TPU。随着 AI 技术的不断发展和算力需求的不断提高，AI 的计算平台也在不断改变。

首先介绍 CPU：最初，AI 规模并不大，其计算主要依赖 CPU。CPU 作为计算机的大脑，擅长处理顺序任务和复杂的操作。但在大规模数据和并行计算任务中，CPU 的表现不佳，导致其训练速度慢，推理延迟高。

后来，大家发现 GPU 可以通过其图形渲染管线的数百到数千个核心实现并行计算，显著加速 AI 训练和推理。其专门用来进行矩阵乘法运算的集成电路使其比 CPU 快几个数量级，但其仍为通用计算硬件，并没有完全针对 AI 任务优化。

CPU 和 GPU 都没有专门对 AI 进行优化，谷歌看到了这一点，创造了 TPU。TPU 作为专为 AI 任务设计的硬件，针对大量的“张量”计算优化了计算架构和能效。其理论性能是 GPU 的数十倍，还有着更低的功耗，是目前 AI 计算的最佳选择。

但是这些硬件都有一个共同的问题：他们都是在内存中使用数据结构构造神经网络，并通过数字电路进行浮点运算来模拟神经网络的运行，这样的计算会消耗大量的在数据的搬运和处理上，使得其运行及其低效。不仅浪费电力资源，还会产生大量的废热，使得散热系统的设计和构建成为难题，潜在的加剧了全球变暖的进程。

BrainScale S-1 应运而生。它解决这一问题的思路是抛开数字电路的束缚，尝试使用模拟人脑的电路直接构建神经网络。

这，是浮动栅极晶体管，一般被用来制造固态硬盘和手机里的内存颗粒。当你在控制极施加足够高的电压时，电子将通过隧穿效应通过绝缘层进入浮动栅极，这些电子能在浮动栅极中保存数十年。这时可以通过测量其导通电压来读取数据。如果导通电压高，则为 0，电压低则为一。BrainScale 的想法是不将这些晶体管视作开关，而是滑动变阻器。他们在浮动栅极中注入不同数量的电子。电子数量越多，等效电阻就越高。如果事后你再往上加一个比较小的电压，电流就等于电压除以电阻。你也可以把它当成电压与电导的乘积，电导就是电阻的倒数。这样，一个浮栅晶体管就变成了两输入的乘法器，即电压乘以电导，用这样的晶体管，我们就能设计出类脑硬件。

要在这样的芯片上运行人工智能，首先要将网络的所有的权重作为电导，写入闪存。然后将激活值通过数模转换器转换为电压施加在芯片的输入端，所得到的结果就是电压与电导的乘积，也就是激活函数的激活值乘以权重。这些单元相互连接，可以对每一处乘法的结果进行求和，最后通过模数转换器读取电流值，这样就实现了矩阵乘法。

BrainScale S-1 一经推出技惊四座。其相比传统的计算硬件不仅性能得到了极大的提升，而且功耗也低的几乎不在一个数量级，似乎 BrainScale 就是 AI 计算的未来，但问题

在于 BrainScale S-1 和其他设备不同。其他设备都是通用计算设备，而 BrainScale S-1 则是一款推理芯片。所以你只能在 S1 上运行提前制作好的人工智能，而训练的工作则仍然需要传统的通用计算设备承担，这也就是为什么这几年英伟达的股票节节攀升，显卡的价格居高不下。

总而言之，我们能看到 BranScale S-1 这种类脑硬件潜在的巨大优势，但将其从实验室中带出来带有很长的路要走。如何提高其热稳定性（刚才忘了说了，在高温环境下神经网络中会产生很强的噪声，必须要将其置于超低温环境下才能稳定运算），还有如何缩小体积，如何降低成本，以及有没有可能让其功能不局限于推理，也能参与到神经网络训练中，这届不仅是其目前的局限性，更是其未来可能的发展前景。我们期待着在可以预见的未来，更低能耗和更高性能的芯片能够在我们手中飞入寻常百姓家。我是高鹤箫，谢谢