解决方案简介

英特尔可编程解决方案事业部
英特尔® Stratix® 10 FPGA
神经形态计算

在可重配置硬件上进行人类大脑级神经形态计算

对大脑进行逆向工程的艰巨任务进入了依赖英特尔® FPGA 的新时代。

WESTERN SYDNEY UNIVERSITY

国际神经形态系统中心

概要

本文描述了澳大利亚正在进行的一个项目，该项目旨在构建一台大型神经形态计算机，其硬件可使用软件进行重配置。支持此类可配置硬件的主要技术是 FPGA，英特尔® FPGA 产品中包含的其他技术可帮助该项目扩展到与人类皮层中神经元和突触数量相当的规模。

建造一种新型计算机

西悉尼大学（WSU）的国际神经形态系统中心（ICNS）正计划建造一台与众不同的计算机 — 一种可扩展的神经形态计算系统，由通过高性能计算（HPC）网络结构互连的现场可编程门阵列（FPGA）组成。该项目的目标是，通过创建世界上第一台可配置的大脑级神经形态计算机，推动对人工智能（AI）和神经科学新领域的研究。

什么是大脑级计算？此概念旨在创建一个拥有与人类大脑皮层相当数量神经元和突触数量的计算环境。大脑皮层中的神经元数量估计在 100 亿到 200 亿之间，而突触数量更为庞大，在 60 到 240 *万亿*之间。[1] 大脑级计算的难点在于构建数量相当的人工神经元和突触，这一技术有望提供新的方法来研究生物大脑中的信息处理，包括在出现问题时，以及开发更出色的机器学习（ML）和人工智能技术。

神经形态计算是计算机科学的一个分支，专注于在硬件中真切模拟生物大脑的结构和功能，包括神经元、突触连接和尖峰信号。这是一个跨学科领域，需要深入了解生物学、计算机架构和机器学习。

FPGA 是一种特殊的微芯片，其独特性在于支持使用软件在硬件层面进行配置和重配置，这为在同一硬件上实现不同模型以进行人工神经元组织提供了可能性。[2] WSU 团队已经证明，在单个 FPGA 主板上使用 FPGA 进行神经形态计算的概念具备可行性，但这样的项目并没有大脑级实现先例。

一旦 WSU 成功构建了这种可配置的神经形态平台，神经形态硬件设计的研发必将迎来空前机遇，因为该平台支持在没有新硬件的情况下实现新设计，能够加快破解神经形态奥秘的步伐。

三方面的重大进展让这一项目露出曙光。首先是 WSU 进行的开创性研究和概念验证（PoC）。他们先是发现 FPGA 可以配置和可重配，能够模拟大脑皮层中不同种类的生物结构，然后一鼓作气，开发高效的核心与方法，以扩展到大量神经元。[3]

其次，在最新一代的英特尔® Stratix® 10 FPGA 中，高带宽内存集成在 BittWare 520N-MX 主板上，从而提供了一种强大且经济的现成硬件平台。[4]

最后，英特尔的可配置网络协议加速器（COPA）技术提供了一个高带宽和低延迟的接口，能够以高度可扩展的方式将大量 FPGA 连接在一起 — 由于过去没有这种能力，FPGA 无法大规模使用。[5] COPA 可提供从 FPGA 到主机资源（如内存和硬盘）的简单编程访问，从而为将系统扩展至人类大脑大小创造了可能性（尽管会影响实时性能）。

本文讨论了神经形态计算这种新方法在促进 AI 突破和帮助人类深入了解大脑神经计算方面的光明前景。文中解释了 FPGA 的工作原理及其对加速神经形态研究进展的关键作用，描述了 COPA 和其他英特尔技术如何帮助 FPGA 在大脑级神经形态计算中释放强大能力。

跨学科项目

WSU 旨在将神经形态加速器打造成一个迭代式多学科开发平台。机器学习和神经科学领域的 HPC 架构师和专家正在携手合作，共同设计一个通用的现成系统，以加速该领域的发展和探索。

对计算机科学家和神经科学家而言，跨学科探索与合作本身会带来更多机会。在神经科学家看来，神经形态计算系统是一种重要工具，可对难以在活体大脑中测试的理论进行测试。例如，理论上人脑可能使用一种时空编码形式，但今天的实现方案无法大规模复制这种行为。大脑级神经形态计算机支持通过网络探索和实验增强对多种局部学习规则（支持局部自组织）的全局洞察，尤其对交互等棘手课题的洞察。

计算机科学家可以利用神经科学家对生物大脑的研究成果来设计包含神经元和突触连接的类似人工结构，并测试这些结构的功效，通过迭代改进性能。WSU 就是遵循这一思路进行项目概念验证的，具体步骤包括，根据目前对大脑皮层结构的了解，建构神经元和突触层次结构的模型，并通过模拟具有 1 亿个神经元的简化听觉皮层来验证该模型。[3]

神经形态计算助力人工智能演进

近年来，人工智能领域主要关注深度学习（DL）神经网络。深度学习探索时代已经过去，深度学习产业化时代已经到来。深度学习长期徘徊不前，直到大约五年前生成对抗网络（GAN）横空出世。一直以来，该领域的发展主要体现在规模化方面：越来越大的数据集用于在 CPU 和图形处理单元（GPUs）越来越多的超级计算机上训练更大的神经网络。

规模化效益已经开始萎缩。计算能力并未保持早期的增长速度，而硬件价格和功耗成本日益高企，阻碍了广泛深入的研究工作。[6] 此外，深度学习方法存在固有局限性，即仅限于在狭隘的专业领域创建更准确的预测模型。虽然这种方法很有用，但无助于实现打造更接近生物大脑的通用人工智能或真正人工智能产品的宏伟目标。与深度学习超级计算机不同，两岁儿童的大脑不需要经过 10,000 张图像的训练便可识别猫。下一步，人工智能的研究重点是构建可真切模拟生物大脑结构与功能的硬件，解开神经形态的奥秘。

WSU 将着力使用简单的泄漏积分触发（LIF）神经元模拟大规模且结构上相连的尖峰神经网络（SNN）。这些概念都不是这个项目的新概念或独创概念。LIF 神经元提供了一种成熟的方法，可帮助在芯片中模拟生物神经元的行为，这些神经元可在细胞膜上产生尖锐的电位或 “尖峰”。同样，创建的 SNN 可通过模拟生物大脑中的自然神经网络，模拟大脑的工作方式。SNN 中的每个人工神经元都可以独立于其他神经元触发，向网络中的其他神经元发送脉冲信号，直接改变这些神经元的电动状态。

这个项目使用了独特的技术。过去最成功的 SNN 项目都是建立在应用特定集成电路（ASIC）上的。建立在 CPU 和 GPU 上的 SNN 运行缓慢，因此专门制造的 ASIC（例如英特尔® Loihi 研究芯片）旨在提供大规模神经形态计算所需的性能。

ASIC 的缺点是设计和制造非常耗时且成本高昂。ASIC 芯片一经制成，便不可更改。开发的任何硬件设计都无法修改，不可使用不同的配置。这就是 WSU 项目对该领域的革命性意义 — 它建立在 FPGA 技术之上，这意味着底层硬件结构可重配置，从而将最新研究成果整合到硬件平台中。

利用 FPGA 进行神经形态计算

ASIC 漫长的制造周期和严格的配置阻碍了技术发展。FPGA 支持硬件更新，有助于测试新理论和应用最新的发现成果。目标是创建一个基于现成硬件的开源解决方案，为研究人员和行业专业人士提供一个通用平台，以在更广泛的范围内执行受大脑启发的计算和发现。

FPGA 是可编程逻辑门的集成电路，与 CPU、GPU 或 ASIC 硬件不同，它可以使用软件进行配置和重配置。这很重要，因为人们还未完全揭开大脑神经元的组织奥秘 — 为实现这一目标，我们需要持续开展模拟实验。理论需要测试和完善；测试理论需要能够修改底层硬件。基于 FPGA 的神经形态计算为技术研究和发现提供了一个可配置的平台。

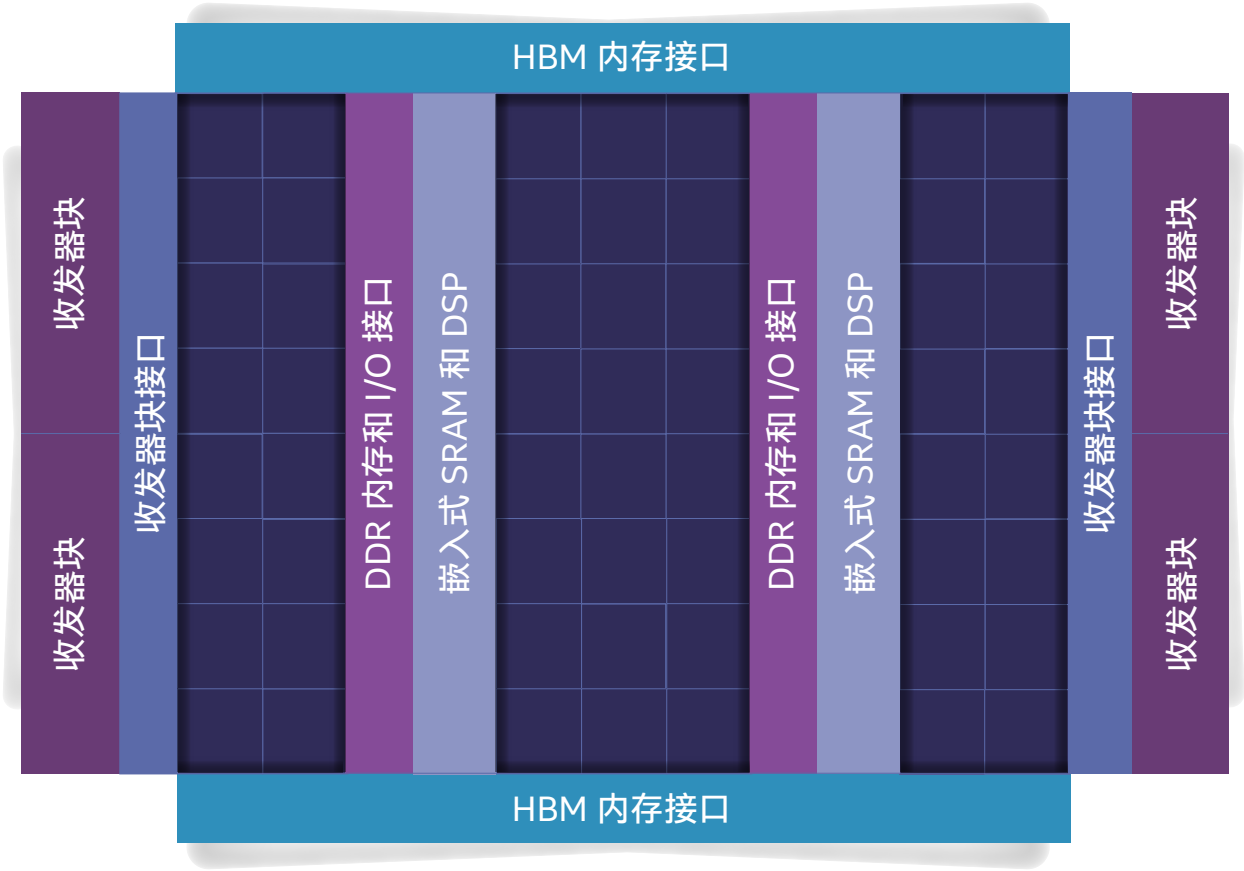

图 1. FPGA 的组件

WSU 完成的概念验证表明，FPGA 硬件的可编程性可用于探索高度复杂的神经形态计算模型。在 Mark Wang 的带领下，该团队使用 FPGA 可编程性模拟人类皮层中突触连接的层次模式，如图 2 所示。

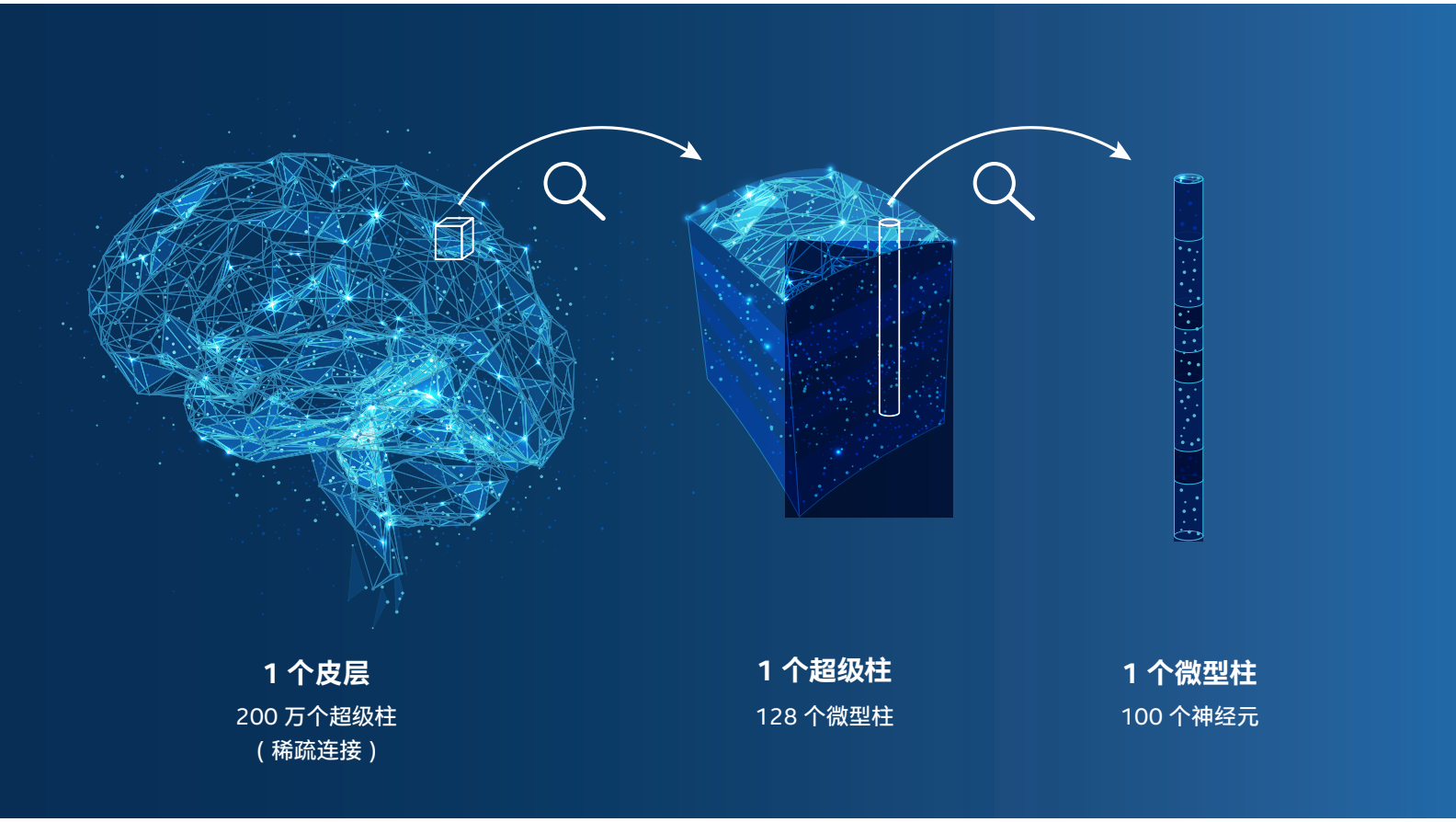

图 2. 神经形态计算系统的功能构建模块是一个稀疏连接的超级柱网络，每个超级柱由多达 128 个连接的微型柱组成，每个微型柱由 100 个紧密连接的神经元组成[3,7]

Wang 工作的独到之处在于将神经形态架构抽象化为由微型柱和超级柱表示的集群（如图 2 所示），类似于神经生物学中的基本结构单元。层次通信方案允许一个神经元从多达 200,000 个神经元中扇出，也就是说，一个神经元的输出可以作为多达 200,000 个其他神经元的输入。

> “如果没有这种方法，模拟完全连接的大规模网络就需要非常大的内存来存储点对点连接的查找表。我们使用基于新皮层结构连接的新型架构，所有需要的参数和连接都可存储在片上内存中。通过对内存进行编程，皮层模拟器可以轻松地重配置，以模拟不同的神经网络，而无需更改硬件结构。”[3]

概念验证以前在单个英特尔 Stratix V FPGA 上实现，它能够实时模拟具有多达 26 亿个 LIF 神经元的尖峰神经网络。当前项目将该平台扩展至三个服务器机架的 168 个英特尔 Stratix 10 MX FPGA，并将支持以每秒 86 万亿次突触操作的理论峰值性能模拟 SNN — 该速度与人类皮层每秒 20-118 万亿次突触操作的估计速度相当。[8] 这种规模依靠的不仅仅是 FPGA，而是众多先进的英特尔技术。

关键的英特尔技术

FPGA 是这种可扩展神经形态计算机背后的核心技术。虽然可配置性具有革命性意义，但其他 FPGA 功能对项目也很重要。一方面，FPGA 计算速度很快 — 虽然不如 ASIC 快，但在将数十亿个人工神经元进行复杂的相互连接方面，FPGA 比 CPU 或 GPU 快得多。FPGA 还提供了比 ASIC 更高的密度，因为它们可以使用更小的先进技术。作为现成的商用产品，FPGA 的经济性远超 ASIC。

两项英特尔技术对该项目尤为重要：FPGA 芯片上的高带宽内存（HBM）和 COPA 技术。

高带宽存储

HBM 在神经形态计算机中发挥着重要作用。它是一种封装内存解决方案，是较低容量片上内存和较低带宽外部内存之间的一种理想介质。HBM 提供了足够的容量来实现 210 万个查找表，这些查找表定义了大脑级计算机中神经元、微型柱和超级柱之间的点对点连接模式。它提供了足够的带宽帮助快速访问这些查找表 — HBM 块的带宽大约是传统 DDR4 DIMM 的 10 倍。[9] 带有两个 HBM 块的英特尔 Stratix 10 MX FPGA 封装提供高达 512 GB/秒的内存带宽。这些内存带宽通常不易用完，WSU 开发的 IP 核心旨在充分利用高带宽。

在外部内存上存储和访问 210 万个查找表会非常缓慢。使用紧邻逻辑电路旁 FPGA 芯片的更快 HBM 可提供这种规模的计算所需的速度。

COPA

FPGA 通常不用作分布式计算场景中的自主节点，这在很大程度上是因为此类网络没有标准化的硬件/软件基础设施。为满足这一需求，英特尔的 COPA 技术提供了一种可定制的框架，以使用专用的高速以太网网络将 FPGA 连接到大型集群中。COPA 技术是构建大规模 FPGA 系统的关键系统构建块，可管理 FPGA 节点以及 FPGA 和主机之间的复杂功能。

COPA 技术提供的可扩展性对脑机项目至关重要，因为它支持任何大型集群处理大规模的神经形态工作负载。COPA 提供的不仅仅是网络连接功能。它将加速功能紧密集成到网络中，支持在传输和接收期间直接对数据执行数据转换和过滤操作。该框架高度可定制，通过定制可实现最佳性能。该软件框架支持一种开放标准应用编程接口（API），即 OpenFabrics 接口（OFI），提供了在网络通信中调用加速器功能的扩展。

WSU 神经形态计算机组件

- WSU 机器将由三个全高机架组成，每个机架都有一个 64 端口 100 Gb HPC 交换机（用于连接到每个支持 COPA 的 FPGA），以及一个 24 端口 10 Gb 网络交换机。
- 这三个机架将容纳 42 台搭载第三代智能英特尔® 至强® 可扩展处理器的 2U 主机服务器，每台服务器总共具有 128 GB 的主内存和 2 TB 的固态盘存储。
- 42 台服务器均采用 4 块 BittWare 520N-MX 主板，其配备了英特尔® Stratix® 10 MX2100 FPGA 和 16 GB 集成 HBM2 内存。
- 满载时的预计热设计功耗（TDP）为 38.8 kW。

可扩展、可访问的大脑启发式研究的新时代

WSU 团队正与英特尔合作开发世界上第一台使用英特尔 FPGA 的大脑级神经形态计算机。这种高度可扩展的神经形态计算系统将是独一无二的，但绝不具有独占性 — 整个项目都强调可重复性和可访问性。

一台这样的机器不足以满足全球的访问需求。正因如此，WSU 机器可在其他地方复制。环境是普通的数据中心，硬件是现成的，软件堆栈是开源的。因此，WSU 机器为其他大学和其他数据中心进行技术实例化提供了蓝图和软件堆栈。

WSU 机器为研究领域提供了一种通用神经形态计算平台的原型。借助现成的组件方法和开放的软件堆栈，世界任何地方能够以任何规模复制类似的机器，以满足计算需求并整合新的发现成果。如果一切顺利，在不久的将来，这些机器将在云端运行，从而为神经形态计算即服务打开方便之门。届时，世界各地的神经科学家和神经形态工程师可通过申请计算时间（而不是硬件）来执行大脑级神经网络模拟。

该项目有望显著加速神经形态计算领域的发展。像这样的神经形态计算机 - 有些可能更小，有些可能更大 - 将在全球遍地开花，神经网络模拟将广泛普及，经济适用。跨学科研究人员的发现成果将在彼此基于 FPGA 的计算系统（大同小异）上得到共享和验证。在人工智能和神经科学研究的初始阶段之后，可能诞生具有广泛商业用途的新型人工智能方法。这种方法应该成为各行各业的机器学习和数据分析工具。

有一点毋庸置疑，基于 FPGA 的机器绝不是神经形态计算的发展终点，而只是漫漫长征的起点。这些机器将成为技术研究和发现的重要工具，可帮助进行理论测试和完善模型，助力人们在大脑工作方式和模拟技巧探索方面取得突破。新发现的模型在经过测试和改进后，可用于构建具有更出色性能特征的专业集成处理器。

现在，我们需要一个可重配置的硬件平台，该平台支持快速迭代，能够整合并推广新的研究成果。WSU 正在构建基于 FPGA 的神经形态计算系统，该系统将成为首个满足该需求的平台，并将成为行业追随的原型。

了解更多信息

敬请访问国际神经形态系统中心的网站：
westernsydney.edu.au/icns

了解英特尔 FPGA 产品和技术的更多信息：
intel.com/content/www/us/en/products/programmable.html

详细了解 BittWare 520N-MX FPGA 卡：
bittware.com/fpga/520n-mx/

了解可配置网络协议加速器（COPA）技术：
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9280342

[1] 较低的估计值来自：G.M. Shepherd。*大脑的突触组织*。1998。较高的估计值来自：C. Koch。*计算生物物理学：单个神经元的信息处理*。1999。

[2] 有关 FPGA 技术的详细介绍，请参见：英特尔。“Architecture All Access：现代 FPGA 架构 | 英特尔技术。” 2021 年 5 月。https://youtu.be/EVy4KEj9kZg。视频 12:56 处开始介绍的城市规划比喻特别形象：FPGA 开始就像一个城市骨架，只有空荡荡的建筑物和街道，您可以决定建筑物的用途和交通布局。

[3] R.M. Wang，C.S. Thakur，和 A. van Schaik。“基于 FPGA 的大规模并行神经形态皮层模拟器。” *神经科学的前沿发展*。2018 年 4 月。frontiersin.org/articles/10.3389/fnins.2018.00213/full

[4] 英特尔。“英特尔 Stratix 10 FPGA 和 SoC FPGA。” intel.com/content/www/us/en/products/details/fpga/stratix/10.html

[5] Venkata Krishnan、Olivier Serres 和 Michael Blocksome。“可配置网络协议加速器（COPA）：集成网络/加速器硬件/软件框架。” *IEEE 高性能互连研讨会*。2020 年 8 月。https://ieeexplore.ieee.org/document/9188286

[6] 例如，训练 GPT-3 的单次成本据说为 1200 万美元。资料来源：Ala Shaabana。“人工智能的未来系于去中心化。” 2021 年 2 月。https://towardsdatascience.com/the-future-of-ai-is-decentralized-848d4931a29a

[7] Matthieu Thiboust。*来自大脑的洞察：机器智能之路*。2020 年 4 月 insightsfromthebrain.com/

[8] Peter Lennie。“皮层计算的成本。” *现代生物学*。2003 年 3 月 sciencedirect.com/science/article/pii/S0960982203001350

[9] 传统的 DDR4 DIMM 带宽大约为 21 Gb/秒，而 1 个 HBM2 区块高达 256 GB/秒。资料来源：英特尔。“英特尔 Stratix 10 MX FPGA。” intel.com/content/www/us/en/products/details/fpga/stratix/10/mx.html

中国印制　0821/SS/PRW/PDF　请注意环保　347936-001CN