采用硅光子学的2.5D芯粒平台中的机器学习加速器

Febin Sunny、Ebadollah Taheri、Mahdi Nikdast、
Sudeep Pasricha科罗拉多州立大学电气与计算机工程系
美国科罗拉多州柯林斯堡 {febin.sunny, ebad.taheri,
mahdi.nikdast, sudeep}@colostate.edu

摘要——**谷歌的张量处理器和苹果的神经网络引擎等领域专用机器学习（ML）加速器，如今在能效型机器学习处理中已超越中央处理器和图形处理器，占据主导地位。然而，由于单片处理芯片的计算密度有限，且依赖速度较慢的金属互连，电子加速器的发展正面临根本性限制。本文提出一种构想：将光计算与光通信集成到2.5D芯粒平台中，以推动全新一类可持续、可扩展的机器学习硬件加速器的发展。我们阐述了如何通过光器件、电路和架构的跨层设计与制造，以及硬件/软件协同设计，打造高效的基于光子学的2.5D芯粒平台，从而加速新兴机器学习工作负载。**

关键词—***2.5D*** 芯粒平台、机器学习、硅光子学、中介层网络、众核计算

I. Introduction

如今，深度神经网络（DNNs）已广泛应用于各种领域，包括但不限于自动驾驶汽车、医学诊断、网络安全、推荐系统和导航方案[1]-[5]。为满足这些应用的目标，深度神经网络已经变得相当多样化，其家族包括卷积神经网络（CNNs）、循环神经网络（RNNs）、图神经网络（GNNs）、Transformer模型等。这些深度神经网络变体的共同点是模型复杂度不断增加，参数数量呈上升趋势。为满足这些应用对处理能力和延迟的要求，硬件架构也必须在处理能力、片上存储器容量和片上通信能力方面实现扩展。如今，图形处理器（GPUs）通常用于加速深度神经网络的执行，但近年来，GPU架构通用性所带来的若干局限性已逐渐显现。这些局限性包括功耗较高、面积开销不断增加、每瓦性能下降以及内存带宽受限[6]。

图形处理器的局限性凸显了对更高效的领域专用加速器架构的需求。然而，现代深度神经网络不断增长的处理需求并不适合采用单片（单芯片）架构 [7]。领域专用加速器的单片实现可能面临可扩展性、功率密度、制造良率和延迟方面的问题 [8]。为了解决这些问题，并以可扩展的方式有效加速现代深度神经网络，如今人们正积极考虑采用 2.5D架构 [9]。

扩展2.5D架构面临着芯粒间距离不断增加的挑战。在这种情况下，可以看出，芯粒间金属互连会因延迟过高和能耗过大而严重影响系统性能 [10]。鉴于这些限制，必须考虑电互连替代方案，以确保2.5D加速器能够满足深度神经网络加速对低延迟和高能效的需求。

基于硅光子学的光互连能够凭借高带宽通信[11]，、单跳数据传播 [12]，和高能效 [13]等优势，克服金属互连带来的限制。硅光子互连还便于实现广播 [13], [14]，，这对于深度神经网络加速 [15]-[17]而言是一项理想特性。通过同时利用光子学进行计算，还能进一步提升能效并降低延迟。此前的许多研究表明，用于深度神经网络推理加速的光子处理在延迟和能效方面具有显著优势 [18]-[26]。因此，有理由认为，同时利用光子学进行通信和计算，或许能进一步放大上述优势。在这项工作中，我们探讨硅光子芯粒和网络在2.5D芯粒平台中用于深度神经网络加速所带来的优势。

本文其余部分的组织结构如下。第二节概述硅光子学。第三节讨论基于硅光子的深度神经网络加速器。第四节介绍最先进的硅光子中介层网络。第五节介绍我们的基于2.5D芯粒的深度神经网络加速器。第六节展示各种实验结果。最后，第七节总结这一新兴领域并探讨尚待解决的挑战。

II. 硅光子学概述

硅光子学是一种与互补金属氧化物半导体兼容的技术，旨在实现芯片级光通信。为此，采用绝缘体上硅（SOI）波导，其中硅（Si）用作芯层材料，二氧化硅（SiO_2）用作包层和衬底材料。此外，通过使用波分复用（WDM），不同波长上的光信号可以同时通过同一波导传输。随着通信距离的整体尺度不断增大，硅光子学有望实现高能效、高带宽密度和低延迟 [27]。鉴于硅光子学带来的优势，人们对将其用于计算（包括利用光子实现数字逻辑）日益感兴趣 [28], [29]。要实现任何基于光子的计算或网络系统，都需要许多基础组件，接下来将对此进行讨论。

硅光波导类似于电芯片中的金属导线，可实现芯片内的光信号传输和路由。光子波导利用全内反射（TIR）原理来约束和引导光信号 [30]。为确保全内反射，这些波导要求其芯芯材料与包层材料之间具有较高的折射率差（例如绝缘体上硅平台）。

激光器是任何光子系统的关键组成部分，因为它们是通信和计算所需的光源。所采用的激光源可以是片上或片外的 [31]。片外激光器具有更高的发光效率，但它们

在耦合到片上波导时会面临较高的光功率损耗。片上激光器具有更高的集成密度和更低的光损耗，因为无需耦合光，但其发光效率较低。在大多数光子学系统中，常用的激光器是垂直腔面发射激光器（VCSEL）或微环激光器。如果使用片外激光器，则需要借助耦合器，将来自片外光源的光信号耦合到片上波导。所采用的耦合方案可以是表面光栅耦合器或边缘耦合器 [33]。

微环谐振器（MRs）是广泛用于设计调制器、光开关和光滤波器的光子器件 [34]。在计算系统中，它们可通过振幅调制执行乘法运算 [35]。MRs 由环状硅光波导制成（见图 1）。MR 有谐振和非谐振两种状态，光信号可据此切换至不同端口。通过利用硅的电光（EO）或热光（TO）效应改变器件复合波导的有效折射率，可以调节 MR 的谐振波长。在通信系统中，有源微环谐振器或配有调谐电路的 MRs 可用于滤除对应开关键控（OOK）调制方案中 0 的波长。在四电平脉冲幅度调制（PAM-4）等先进调制方案中 [44], , MRs 可用于将信号幅度调制到四个不同的电平。多个对同一波长敏感的 MRs 可用于连续进行振幅调制，从而实现参数乘法。微盘谐振器 [36] 与 MRs 类似，但采用盘状结构而非环状结构。它们比 MRs 更紧凑，但运行损耗更高。

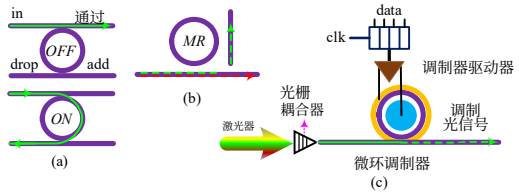

图*1* 微环谐振器（*MR*）：*(a)* 关闭/开启状态，*(b)* 微环滤波器，以及 *(c)* 微环调制器。

马赫-曾德尔干涉仪（MZI）由两个 3-dB定向耦合器和两条带有移相器的波导臂组成。移相器通过电光或热光调谐实现，可改变 MZI 一条或两条波导臂中的光相位，从而在输出端产生相长或相消干涉，使光信号在各输出端口之间切换。MZI 可用于设计光调制器、光开关和滤波器。与 MZI 相比，微环谐振器占用面积更小、功耗更低。不过，MZI 在工作时具有更好的热稳定性，消光比也优于微环谐振器 [20]。

光电探测器（PD）用于将光子信号转换为电信号。PD 的工作可能需要在工作带宽与能效之间进行权衡。高效的 PD 只需较小的输入光信号，就能产生所需的电输出。然而，较小的输入光信号可能导致 PD 的带宽性能较低。要高效地将光子信号转换为电信号，PD 接收到的光子信号强度应大于 PD 的响应度。高带宽 PD 可用于光子计算，对不同波长的信号执行累加运算 [32]。

III. 基于硅光子的深度神经网络加速器

硅光子学有望提升能效并降低延迟，因此在学术界和工业研究领域中，硅光子学在深度神经网络加速方面日益受到关注 [36]。光子学尤其适合加速深度神经网络推理运算，因为这类运算高度依赖固定的矩阵乘法。矩阵乘法涉及的线性变换可以利用光子学在模拟域中高效实现。如下文所述，硅光子深度神经网络加速器可以采用相干架构或非相干架构。

相干架构采用单一波长，并依靠相长干涉和相消干涉来改变相干光束的相对功率水平 [18]-[20]。通过光相位控制，将参数编码到光波信号上。为此，相干架构使用带有嵌入式相位调制器的马赫-曾德尔干涉仪。加权通过与权重值成比例的电场振幅衰减，以及与权重符号成比例的相位调制来实现。级联合束器用于累加，并促进信号之间的相干交互。该领域的许多研究都致力于降低片上实现的深度神经网络的计算复杂度。可通过剪枝方法 [18] 或奇异值分解（SVD）[19], [20]来降低计算复杂度。

非相干架构（如 [21]-[26], ）使用多个波长，每个波长都可用于并行执行计算。在这些架构中，参数通过波长选择器件（如微环谐振器）加载到信号幅度上。如上所述，已有多项研究探索了基于非相干光子学原理的深度神经网络加速。在 [21], 中，有研究提出了一种基于微环谐振器的深度神经网络加速器架构，该架构采用模块化向量点积单元，并通过优化微环谐振器设计和调谐电路优化来提高能效和吞吐效率。为了进一步优化非相干加速器的功耗和能耗，尤其是在电-光子接口处，[22] 采用了异构量化（即每个深度神经网络层可使用不同的参数位宽），并结合软硬件协同优化。为降低面积和功耗，[23] 使用微盘取代微环谐振器。为了进一步降低电光接口处的功耗，可以考虑采用二值化神经网络。[24] 提出了一种面向全二值化深度神经网络（权重和激活参数均为单比特）的基于微盘的光子加速器。尽管全二值化神经网络在存储和功耗方面效率更高，但其可实现的准确率可能不足。为满足部分二值化神经网络的加速器需求，[25] 提出了一种基于微环谐振器的部分二值化深度神经网络加速器。事实证明，非相干架构对于循环神经网络加速也很有效，如 [26] 所示。在 [26], 中，研究利用光子加速器基底的运行速度，执行不同类型循环神经网络所需的大规模矩阵运算，包括深层长短期记忆网络（LSTM）和门控循环单元（GRU）模型。

IV. Silicon Photonic Interposer Networks

传统上，芯粒系统采用无源[39] 和有源[40] 电子中介层进行封装。与无源电子中介层相比，有源电子中介层采用包含逻辑元件的互连架构，而非仅使用无源金属互连，因此具有更好的通信可扩展性。然而，由于金属互连固有的局限性——长互连延迟高且每条互连的带宽低——有源和无源电子中介层都无法高效支持包含大量芯粒的系统。

另一方面，由于光互连具有低延迟和高带宽，光子互连架构可成为中介层设计的一种很有前景的解决方案。因此，光子中介层近年来在芯粒系统中受到广泛关注 [10], [11]。例如， [10] 采用高带宽阵列波导光栅路由器（AWGR），以克服芯粒系统中传统电子中介层的高延迟和低带宽问题。除了长距离互连具备高带宽和低延迟之外，硅光子中介层还天然支持动态调节芯粒间带宽。PROWAVES [11] 介绍了一种光子中介层网络，它根据流量负载调整激活的波长数量，从而动态管理芯粒间带宽。在流量负载较低、所需带宽较低时，PROWAVES会使用较少的波长，并关闭离片激光器中未使用的波长，以节省功耗。另一方面，在流量负载较高时，PROWAVES会激活更多波长，以提供更高带宽，从而提升芯粒间通信的性能，但代价是功耗更高。为应对 PROWAVES 中的高流量负载，每个芯粒都使用一个高带宽网关（即具有大量波长的网关）。然而，在这种架构中，由于同一芯粒上的所有节点都通过同一个网关与中介层通信，高带宽网关可能导致芯粒上发生拥塞。此外，各芯粒上的节点无法以公平的方式访问网关。例如，距离网关较远的节点可能需要经历极高的延迟才能到达网关。

ReSiPI [37] 通过在波长数量相对较少的芯粒上采用多个网关，对 PROWAVES 设计进行了改进。此外，ReSiPI 在考虑在线流量的同时，通过调整启用的网关数量（而非启用的波长数量）来管理芯粒间带宽。在 ReSiPI 架构中，芯粒间通信的流量负载按时间周期进行监测，并根据所需的芯粒间带宽确定启用的网关数量。网关的启用或停用通过基于相变材料的耦合器（PCMC）实现，其设计依据为 [38] 中的耦合器设计。如图 2 所示，PCMC 可处于三种状态：1）晶态，将输入光引导至直通端（B）输出；2）部分晶态，将部分输入光引导至交叉端（C）输出，其余部分引导至直通端输出；3）非晶态，将光引导至交叉端输出。CL_{am} 和 CL_{cr} 分别是非晶态和晶态的耦合长度。通过调整CL_{am} 与CL_{cr}的比值，可以调节从光学激光器输入的适当光功率以传送至写入网关。通常，在硅光子网络中，写入网关会将数据调制到光信号上，供读取网关接收；无源分光器则用于将来自光学激光器的光信号分配并传送给各个写入器。然而，无源分光器使网络无法动态停用写入网关，而 PCMC 能够调节每个写入器的光输入，从而实现网关的动态启用和停用。借助 PCMC，ReSiPI 中介层经过设计，可重新配置启用的网关数量并降低网络功耗。控制器会根据各芯粒的芯粒间流量，调整每个芯粒中启用的网关数量。根据启用的网关数量，PCMC 会经过调节，向每个网关提供适当的光功率。除了调节 PCMC 外，控制器还会相应地调节激光功率，以降低激光器的功耗。

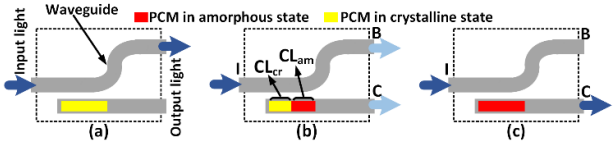

图 *2 ReSiPI* 中使用的基于相变材料的耦合器（*PCMC*）的三种状态：*(a)* 晶态，*(b)* 部分晶态，*(c)* 非晶态。

V. 硅ilicon 光子hotonic 2.5D 深度神经网络 加速器ccelerators

为探究在 2.5D 中介层平台上加速深度神经网络的影响，我们提出一项案例研究，将 CrossLight [21] 光子深度神经网络加速器扩展至 2.5D 芯粒平台。

CrossLight 是一种神经网络加速器，旨在光子域中执行高速乘加运算（MAC）。然而，原始的单片 CrossLight 架构可扩展性较差，能效也相对较低。我们提出采用基于 ReSiPI 的光子中介层架构，设计一种可扩展性更强、能效更高的 2.5D CrossLight 实现。图 3 展示了采用光子中介层的芯粒式 2.5D CrossLight 加速器的高层概览。多个芯粒封装在硅光子中介层基板上。我们将不同类型的芯粒纳入异构架构中考虑。这种异构设计使片上系统（SoC）设计人员能够使用合适的现成芯粒，并创建多种 2.5D 封装，以满足其设计目标 [8]。

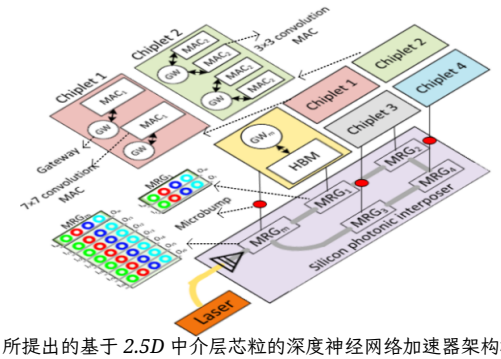

图 *3* 所提出的基于 *2.5D* 中介层芯粒的深度神经网络加速器架构概览。

所提出架构中的芯粒包括各种计算芯粒和存储芯粒。其中一个或多个

芯粒采用光互连接内存架构，例如高带宽内存（HBM；如图 3 所示），并配有专用网关，用于与系统其余部分通信。每个计算芯粒（例如图 3 中的芯粒 1–4）都搭载多个光子乘加单元，并设有本地网关，用于通过中介层网络从存储芯粒读取数据并向其写入数据。每个网关主要由两部分组成：芯粒上的电子电路，以及中介层上的微环谐振器组（MRG）。网关的电子部分通过微凸点技术与 MRG 中的微环相连。

光子乘加单元利用非相干光子学对参数执行乘法运算，并使用光电探测器获取乘积之和。权重和激活值通过特定波长微环滤波器组加载到波长上。该加载过程遵循 [35] 中所述的广播加权协议。尽管所提出的设计采用光子通信来传输数据，但要以稳健的方式传输多比特幅度调制数据仍颇具挑战。因此，中介层采用基于开关键控（OOK）的数据传输，并在网关处进行中间光电转换，同时在乘加单元中缓冲参数。缓冲的数据用于调谐相应的微环谐振器（MR），从而通过波长振幅表示参数值。调谐微环谐振器时采用电光调谐。所提出的架构在不同芯粒上采用异构乘加单元尺寸（此处的尺寸指可部署向量的大小），以适配不同的卷积核尺寸，并处理全连接层所需的大规模乘加运算。例如，在图3中，芯粒1包含 3×3 个卷积乘加单元，而芯粒2包含 7×7 个卷积乘加单元。此外，由于不同尺寸的乘加单元占用面积各异，每个芯粒上的乘加单元数量也可能不同。图4展示了一个乘加单元的概览。

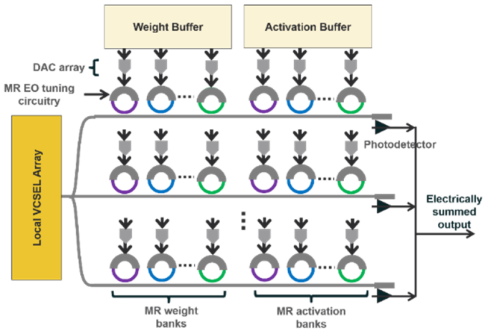

图*4* 乘加运算单元架构（数模转换器：数模转换器）

图 5 展示了该架构中中介层上的光接口与通信示例。在此示例中，乘加单元正从另一芯粒上的高带宽内存读取数据。为实现成功通信，高带宽内存芯粒中使用了一个写入网关，其中包含用于存储和转发数据的缓冲区；同样，带有乘加单元的芯粒上也使用了一个读取网关。写入网关缓冲区中存储的数据会被调制到由片外激光器生成的光信号上。不同颜色的调制器表示它们用于调制不同波长的不同光信号。如前所述，采用多个不同波长的光信号，使我们的网络能够同时传输更多数据在同一波导上，从而提高通信带宽。读取网关还连接了多个 MR滤波器。每个 MR滤波器都调谐至特定波长，以滤出并分离指定的光信号。此后，光信号通过光电二极管转换为电子信号，再通过微凸点传送至读取器芯粒。读取网关将电子信号转换为数字数据，并将接收到的数据存储在其缓冲区中。最后，数据将转发至乘加单元。这种光通信协议中，读取器通过波导从写入器接收数据，称为单写入器单读取器（SWSR）协议。同样，如果多个读取器通过波导从一个写入器接收数据，则该协议称为单写入器多读取器（SWMR）协议。

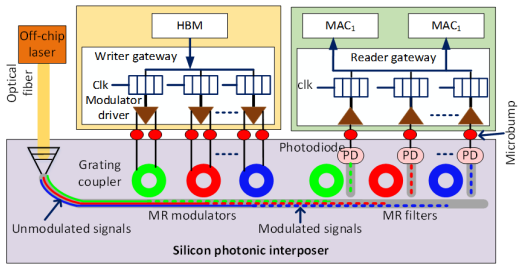

图*5* 中介层上的光通信示例：乘加单元正在从存储器读取数据。

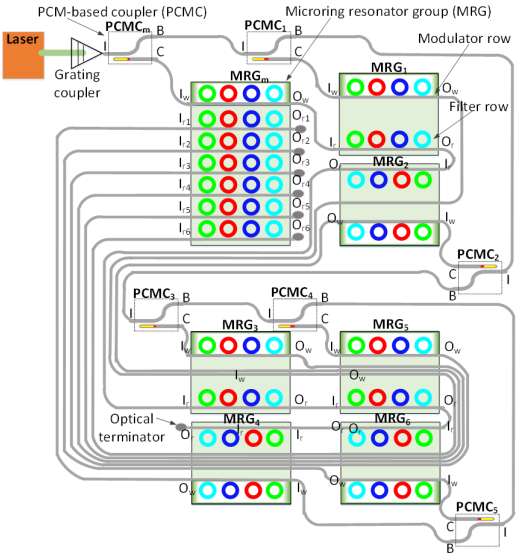

图*6* 我们的基于*2.5D*芯粒的深度神经网络加速器中的硅光子网络。每个微环谐振器组都连接到一个芯粒上的网关。

在我们的架构中，芯粒之间有两种类型的流量：1）从存储器读取乘加单元所需的权重和输入；2）将乘加运算结果写入存储器。因此，从存储芯粒到计算芯粒，我们采用 SWMR 协议从存储器读取数据。此外，从计算芯粒到存储器，我们采用 SWSR 协议。因此，存储芯粒的 MRG 需要若干组 MR滤波器（每组 MR滤波器对应 MRG 中的一行，如图3所示），以接收来自计算芯粒的数据。另一方面，计算芯粒只需要一组 MR滤波器，因为它只接收来自存储器的数据。计算芯粒和存储芯粒都需要一组 MR调制器来发送数据。

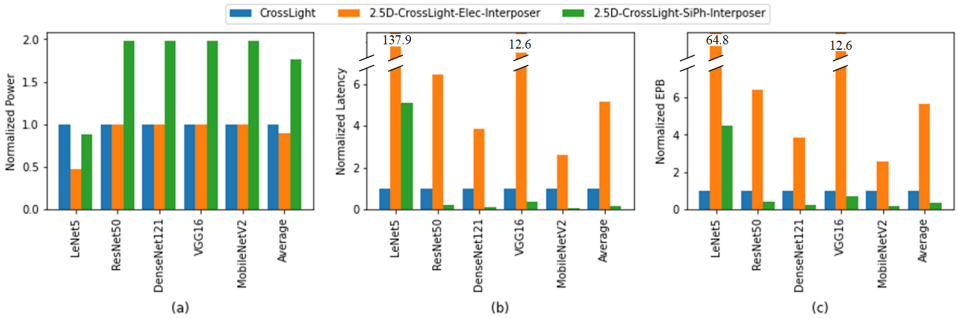

图 7 *CrossLight*、采用电子中介层的 *2.5D-CrossLight* 和采用硅光子中介层的 *2.5D-CrossLight* 的性能分析：*(a)* 归一化功耗，*(b)* 归一化总延迟，以及 *(c)* 归一化比特能耗

图6展示了集成 ReSiPI 中介层的 2.5D CrossLight 示例。虽然该示例展示了六个网关（分别关联一个存储芯粒和五个计算芯粒），但中介层设计可以扩展到更大的系统，且不失一般性。如图6所示，存储芯粒的 MRG（MRG_m）有六行滤波器，用于接收来自计算芯粒六个网关（MRG_1 – MRG_6）的数据；而 MRG_m 有一行调制器，用于向所有网关发送数据。光子中介层网络架构采用无源网络以节省能耗。这意味着有一条特定波导用于将数据从每个写入网关传输到每个读取网关，且该路由（波导）不会改变。

表1. 建模参数

参数		值
光链路数据速率（每个波长）		12 Gb/s
网关频率		2 GHz
电气片上网络链路位宽		128 位
电气片上网络频率		2 GHz
波长数量		64
存储芯粒数量		1
计算芯粒数量		8
100 个单位密集乘加运算	芯粒数量	2
	每个芯粒的乘加运算数	4
	每个网关的乘加运算数	1
7×7 卷积 MAC	芯粒数量	1
	每个芯粒的乘加运算数	8
	每个网关的乘加运算数	2
5×5 卷积 MAC	芯粒数量	2
	每个芯粒的乘加运算数	16
	每个网关的乘加运算数	4
3×3 卷积 MAC	芯粒数量	3
	每个芯粒的乘加运算数	44
	每个网关的乘加运算数	11

表 2. 评估中考虑的深度神经网络模型。

模型	卷积层	全连接层	参数
LeNet5	3	2	62,006
ResNet50	53	1	25,636,712
DenseNet121	120	1	8,062,504
VGG16	13	3	138,357,544
MobileNetV2	52	1	3,538,984

VI. 实XPERIMENTAL 验ESULTS

我们设计了两种 2.5D CrossLight 架构变体：一种采用基于 ReSiPI 的中介层 [37] (*2.5D- CrossLight-SiPh-Interposer*)，另一种采用电气网格中介层 [40] (*2.5D-CrossLight-Elec* 中介层)。我们还从功耗、延迟和能效方面，将这两种 2.5D CrossLight 变体与原始的单片（单芯片）CrossLight 架构进行比较。本研究采用的模型参数汇总于表1。我们还采用了 [11] 中使用的功耗模型和功耗参数和 [37]。我们考虑一个存储芯粒和八个计算芯粒，其中两个芯粒包含全连接层 MAC，另外六个包含卷积层 MAC（3×3、 5×5 和 7×7 卷积乘加单元）。我们使用表2中汇总的多种深度神经网络模型进行评估。

性能结果如图 7 所示。总体而言，*2.5D-CrossLight-SiPh-Interposer* 能够在几乎所有模型上实现更高的能效和更低的延迟，但非常小的模型（例如 LeNet5）除外。异构芯粒和高带宽芯粒间光子网络，使深度神经网络能够比单片式 *CrossLight*架构实现更加高效的能耗执行。

*2.5D-CrossLight-SiPh-Interposer*会带来不可忽略的功耗开销，因为其光子网络在通信过程中比电子网络消耗更高的功耗。然而，对于较小的深度神经网络模型（例如 LeNet5），*2.5D-CrossLight-SiPh-Interposer*的功耗相对较低，这是因为 ReSiPI 控制器会重新配置光子中介层，并停用不必要的网关。尽管如此，对于较小的模型（LeNet5）而言，由于每一层仅占用总体计算面积中的一小部分，*2.5D-CrossLight-SiPh-Interposer*的开销会变得十分显著，并对能效产生不利影响。

对于将多个层映射到芯粒上的较大模型，*2.5D-CrossLight-SiPh-Interposer*在功耗方面的开销可通过这些映射得到更好的摊销。控制器还会在大型模型中启用网关，以应对高流量，这有助于降低芯粒间延迟。尽管*2.5D-CrossLight- Elec-Interposer*的功耗较低，但由于金属互连的延迟显著更高，尤其是在大型中介层上进行较长距离传输时，其性能会受到影响。

表 3. 电子和光子深度神经网络加速器平台的平均功耗、延迟和每比特能耗。

	功耗（W）	延迟（ms）	每比特能耗（nJ/bit）
CrossLight [21]	50.8	8	3.6
2.5D-CrossLight-Elec	45.3	41.4	20.5
2.5D-CrossLight-SiPh	89.7	1.21	1.3
Nvidia P100 GPU	250	13.1	12.3
Intel 9282 CPU	400	86.5	64.4
AMD 3970 CPU	280	141.3	73.7
Edge TPU	2	2366.4	17.6
Null Hop [42]	2.3	8049.3	68.9
Deap_CNN [43]	122	619.01	1959.4
HolyLight [23]	66.5	86.4	40.3

平均而言，与单片式*CrossLight*相比，*2.5D-CrossLight-SiPh-Interposer*的延迟降低了 6.6×，每比特能耗（EPB）也因此降低了 2.8× 。与*2.5D-CrossLight-Elec-Interposer*相比，*2.5D- CrossLight-SiPh-Interposer*的延迟降低了 34×，并且

EPB 降低了 15.8× 。这一显著改进得益于 *2.5D-CrossLight-SiPh-Interposer*能够为每个深度神经网络模型的各层选择合适的芯粒进行映射，并相应地调整所需的芯粒间带宽。我们还从平均功耗、延迟（各层的总延迟）和 EPB 方面，将*2.5D-CrossLight-SiPh-Interposer* 加速器与先进的加速器进行了比较（表 3）。由于*2.5D-CrossLight-SiPh-Interposer*在较大模型上表现出色，且在延迟和 EPB 方面优于先进的电子和光子加速器，这种基于光子的 2.5D 芯粒平台在加速新兴的大型深度神经网络模型方面展现出巨大潜力。

VII. 结ONCLUSIONS 论PEN 与HALLENGES开放性挑战

本文提出了一种基于 2.5D 芯粒平台的光子深度神经网络加速器，其中中介层上的通信和芯粒上的计算均采用硅光子学。与单片光子加速器相比，基于芯粒的方案不仅能够提高制造良率并降低成本，还能利用中介层上的高带宽光子网络降低延迟。此外，芯粒可以采用异构设计，并且能够在 2.5D 封装中集成现成芯粒，从而构建具有不同计算功耗预算和能力的多种系统。

要设计更高效的硅光子学深度神经网络加速器，该领域仍面临若干开放性挑战：1）当前最先进的光子器件功耗相对较高，因此需要开展器件级改进，以设计低功耗器件；2）设计高效的电子控制器对于以低延迟高效控制通信和计算操作至关重要；3）硅光子学 2.5D 深度神经网络加速器架构需要进行设计空间探索（例如，波长数量、每个芯粒的网关数量以及每个芯粒的乘加运算数），以构建针对目标深度神经网络优化的架构。

REFERENCES

[1] V. K. Kukkala、J. Tunnell、S. Pasricha 和 T. Bradley，《高级驾驶辅助系统：迈向自动驾驶汽车之路》，*IEEE Consum. Electron. Mag.*，第 7 卷，第 5 期，第 18–25 页，2018 年。[2] O. Ronneberger 等人，《U-net：用于生物医学图像分割的卷积网络》，*MICCAL*，2015 年。[3] V. K. Kukkala、S. V. Thiruloga 和 S. Pasricha，《INDRA：汽车嵌入式系统中基于循环自编码器的入侵检测》，*IEEE TCAD*，第 39 卷第 11 期，2020 年 11 月。[4] S. Jaiswal 和 T. Jaiswal，《基于深度学习网络的推荐系统综述》，*Artificial Intelligence Evolution*，2020 年。[5] S. Pasricha、V. Ugave、Q. Han 和 C. Anderson，《LearnLoc：面向嵌入式移动设备的智能室内定位框架》，*ACM/IEEE CODES+ISSS*，2015 年 10 月。[6] N. Jouppi 等人，《首款张量处理单元的设计动机与评估》。*IEEE Micro*，第 38 卷，第 3 期，第 10–19 页，2018 年。[7] V. Sundaram 等人，《低成本、高性能、高可靠性的 2.5D 硅中介层》，*IEEE ECTC*，2013 年。[8] J. Kim 等人，《面向中介层的 2.5D 芯粒集成的架构、芯片与封装协同设计流程，实现异构 IP 复用》，*IEEE TVLSI*，第 28 卷，第 11 期，第 2424–2437 页，2020 年。[9] N. Beck 等人，《Zeppelin：面向多芯片架构的片上系统》，*IEEE ISSCC*，2018 年。[10] P. Fotouhi 等人，《基于阵列波导光栅路由器的 2.5D互连网络助力构建可扩展的解耦计算系统》，*Journal of Optical Communications and Networking*，第 11 卷，第 7 期，2019 年。[11] A. Narayan 等人，《PROWAVES：面向节能光子片上网络的主动式运行时波长选择》，*IEEE TCAD*，2020 年。[12] S. Bahirat、S. Pasricha，《探索面向新兴芯片多处理器的混合光子片上网络》，*IEEE/ACM CODES+ISSS*，2009 年

[13] D. A. B. Miller，《硅芯片光互连的器件要求》，IEEE会议录，第 97 卷，第 7 期，第 1166–1185 页，2009 年。[14] S. V. R. Chittamuru、S. Desai、S. Pasricha，《SWIFTNoC：支持多播和信道共享、面向多核架构的可重构硅光子网络》，ACM JETC，2017 年。[15] H. Kwon 等人，《MAERI：通过可重构互连实现深度神经网络加速器上的灵活数据流映射》，ASPLOS，2018 年。[16] Y. Li 等人，《利用基于芯粒的架构和光子互连扩展深度学习推理》，ACM/IEEE DAC，2021 年。[17] Y. Li 等人，《SPRINT：面向卷积神经网络推理、采用光子互连的高性能、高能效、可扩展基于芯粒的加速器》，IEEE TPDS，第 33 卷，第 10 期，第 2332–2345 页，2022 年。[18] S. Banerjee、M. Nikdast、S. Pasricha 和 K. Chakrabarty，《剪枝相干集成光子神经网络》，IEEE JSTQE，2022 年。[19] J. Gu 等人，《迈向面积高效的光学神经网络：一种基于快速傅里叶变换的方法》，IEEE/ACM ASP-DAC，2020 年。[20] C. Demirkiran 等人，《用于加速深度神经网络的电光系统》，arXiv: *2109 01126 2 [ ]*, 2022 。v cs.AR

[21] F. Sunny、A. Mirza、M. Nikdast 和 S. Pasricha，《CrossLight：跨层优化的硅光子神经网络加速器》，IEEE/ACM DAC，2021 年。[22] F. Sunny、M. Nikdast、S. Pasricha，《采用异构量化的卷积神经网络硅光子加速器》，ACM GLSVLSI，2022 年。[23] W. Liu 等人，《HolyLight：用于数据中心深度学习的纳米光子加速器》，IEEE DATE，2019 年。[24] F. Zokae 等人，《LightBulb：用于二值化卷积神经网络的基于光子非易失性存储器的加速器》，IEEE/ACM DATE，2020 F. Sunny、A. Mirza、M. Nikdast 和 S. Pasricha，

[25] 《ROBIN：稳健的光学二值神经网络加速器》，ACM TECS，2021 年。[26] F. Sunny、M. Nikdast、S. Pasricha，《RecLight：集成硅光子学的循环神经网络加速器》，ISVLSI，2022 年。[27] Y. Xu、S. Pasricha，《面向未来多核架构的硅纳米光子学：机遇与挑战》，IEEE Design & Test，2014 年。[28] F.K. Law 等人，《基于光子微环谐振器的数字位幅值比较器演示》，Optical and Quantum Electronics，2019 年。[29] J. Huang 等人，《数字纳米光子学：通往亚波长尺度光子学集成的高速公路》，《Nanophotonics》，第 10 卷，第 3 期，第 1011–1030 页，2021 年。[30] M. Tran 等人，《用于异质集成硅/III-V 光子学的超低损耗硅波导》，《Applied Sciences》，2018 年。[31] Z. Zhou 等人，《硅光子学的片上光源》，《Nature》，《Light: Science & Applications》，2015 S. Pasricha 和 M. Nikdast，《能

[32] 效型众核计算的硅光子学综述》，IEEE Design & Test，第 37 卷，第 4 期，2021 年。[33] S. Nambiar 等人，《绝缘体上硅光子电路的光栅辅助光纤到芯片耦合》，《Applied Sciences》，2018 年。[34] W. Bogaerts 等人，《硅微环谐振器》，《L&P Reviews》，第 6 卷，第 1 期，2012 年。[35] A. N. Tait 等人，《微环权重组》，IEEE JSQE，第 22 卷，第 6 期，第 312–325 页，2016 年。[36] F. Sunny、E. Taheri、M. Nikdast 和 S. Pasricha，《面向深度学习的硅光子学综述》，ACM JETC，第 17 卷，第 61 期，2021 年。[37] E. Taheri、S. Pasricha 和 M. Nikdast，《ReSiPI：采用相变材料、用于高能效中介层通信的可重构硅光子2.5D芯粒网络》，IEEE/ACM ICCAD，2022 年。[38] T. Y. Teo 等人，《基于相变材料的可重构硅光子定向耦合器的比较与分析》，《Optical Materials Express》，第 12 卷，第 2 期，2022 年。[39] P. Ehrett 等人，《SiPterposer：面向灵活系统级封装设计的容错基板》，IEEE/ACM DATE，2019 年。[40] E. Taheri、S. Pasricha、M. Nikdast，《DeFT：面向2.5D芯粒网络的无死锁容错路由算法》，IEEE/ACM DATE，2022 年。[41] S. V. R. Chittamuru、S. Pasricha，《高基数、低直径光子片上网络架构的串扰抑制》，IEEE Design and Test (D&T)，第 32 卷，第 3 期，第 29–39 页，2015 年 6 月。[42] M. Capra 等人，《用于加速深度卷积神经网络的高效硬件架构最新综述》，《Future Internet》，2020 年。[43] V. Bangari 等人，《面向卷积神经网络的数字电子学与模拟光子学（DEAP-CNN）》，IEEE JQE，2020 I.

[44] Thakkar、S. V. R. Chittamuru、S. Pasricha，《利用多级信令提高高带宽光子片上网络架构的可靠性和能效》，IEEE/ACM NOCS，2017 年。