TinyVers：一种用于极端边缘端机器学习推理的微型多功能片上系统，采用状态保持型嵌入式磁阻式随机存取存储器

Vikram Jai n、Sebastian Giraldo、Jaro De Roose、Linyan Mei、Bert Boons 和 Marian Verhelst

arXiv:2301.03537v1 [cs.AR] 9 Jan 2023

摘要——**极端边缘设备或物联网节点既需要始终开启的超低功耗处理能力，也需要按需采样和处理的能力。此外，支持语音识别、机器监测等物联网应用，还要求能够执行范围广泛的机器学习工作负载。这给硬件设计带来了挑战：如何构建在超低功耗状态下运行的灵活处理器。本文介绍 TinyVers，这是一款微型、多功能、超低功耗机器学习片上系统，旨在增强极端边缘端的智能能力。TinyVers 利用数据流重配置实现多模态支持，并通过积极的片上电源管理实现占空比循环，从而支持智能感知应用。该片上系统集成了一个 RISC-V 主机处理器、一个能效达 17 TOPS/W 的数据流可重配置机器学习加速器、一个功耗为 1.7 μW 的深度睡眠唤醒控制器，以及用于启动代码和机器学习参数保持的嵌入式磁阻式随机存取存储器。该片上系统最高可实现 17.6 每秒十亿次操作，功耗范围为 1.7 μW 至 20 mW。我们将面向多种应用的多个机器学习工作负载映射到该片上系统上，以展示其灵活性和能效。所有模型的能效均达到 1–2 TOPS/W，连续运行时功耗低于230 μW。在机器监测的占空比循环应用场景中，功耗可降至 10 μW 以下。**

索引词——**极端边缘端、微型机器学习、机器学习加速器、超低功耗、片上系统。**

I. 引言

Extreme 边缘设备 [1] 或物联网 (IoT) 节点主要执行非视觉任务，即使采用小型轻量级神经网络（NN）模型，也能达到较高的准确率 [2]。这与更传统的任务形成对比：传统任务旨在处理图像数据，涉及数百万至数十亿个参数和运算，且对硬件资源的需求很高。例如，Google语音助手只需 14 千字节 (kB) 的神经网络参数，就能在边缘设备上运行关键词识别应用 [3]。并非所有应用都需要最高准确率以及大型、复杂的神经网络模型；这一认识催生了机器学习应用开发的新范式，称为微型机器学习（tinyML）或极端边缘端机器学习 [4]。这一趋势的核心驱动力是电池供电且性能和功耗受限的物联网节点所面临的要求。大多数物联网传感器节点由一个微控制器单元 (MCU)、一组传感器、用于存储采集数据的存储器、一个中央处理器以及一个无线数据收发器组成。当这些微控制器用于采集数据时，如果神经网络模型较小，便可在靠近传感器的位置处理数据，避免将原始数据传输至性能更强的边缘端或云端设备所带来的高昂代价。

然而，这种本地机器学习处理带来了若干新的挑战：1.) 由于这些节点由电池供电，系统通常受到严重的功耗或能量限制，因此需要以超低功耗运行，并具备空闲能力。2.) 此外，微控制器的计算能力和存储空间有限，导致模型大小、执行性能和硬件复杂度之间存在关键权衡；3.) 尽管需要高效率，系统还应具备足够的灵活性，以支持不同应用中的不同类别神经网络模型；4.) 系统还应具有较小的占用空间。近期文献中已提出多种机器学习硬件，可分为三大类：1) 为超低功耗运行而设计的高度专用边缘机器学习加速器，几乎没有灵活性且性能较低 [5]–[8], 2; 2) 多模态边缘机器学习加速器，在中等至较高功耗下提供中等程度的灵活性和较高性能 [9]–[13], ；以及 3) 商用现成微控制器，灵活性更高，但性能较低且功耗中等 [14]–[16]。这些硬件设计大多无法满足极端边缘端设备的全部要求。Vega 是一个例外 [17]，它提供了完整的片上系统；然而，Vega 的专用加速器不具备处理所有深度神经网络工作负载的灵活性。因此，需要面向极端边缘部署的新型灵活超低功耗平台。

在此背景下，本研究提出了 TinyVers [18]，这是一种高度自适应的片上系统平台，通过采用以下技术，显著改善了极端边缘设备所需能效与灵活性之间的权衡：A.) 配备灵活机器学习加速器（FlexML）的 RISC-V 处理器，该加速器支持数据流重配置，以适应多种机器学习工作负载，并支持针对块结构化稀疏性和反卷积的高效零值跳过；B.) 用于非易失性存储的嵌入式磁阻式随机存取存储器（eMRAM），支持独立运行并实现高效断电（或空闲）；C.) 可编程唤醒控制器（WuC），支持多种上电和空闲模式，从而同时支持常开推理，以及按需和占空周期式智能感知。

V. Jain、L. Mei 和 M. Verhelst 就职于比利时鲁汶大学电气工程系 MICAS。S. Giraldo 曾就职于比利时鲁汶大学电气工程系 MICAS，现就职于比利时 B12 Consulting。J. De Roose 和 B. Boons 曾就职于比利时鲁汶大学电气工程系 MICAS，现就职于比利时 Magics Technologies。
© 2023年 IEEE。允许个人使用此材料。对于其他所有用途，包括在任何当前或未来媒体中重新印刷/再版此材料用于广告或宣传目的、创建新的汇编作品、转售或重新分发至服务器或列表，或在其他作品中重复使用本作品的任何受版权保护部分，均必须获得 IEEE 的许可。

以及典型微型机器学习物联网应用中的计算。该片上系统不仅让用户能够灵活地为各种微型机器学习应用映射多种机器学习工作负载，还支持占空比循环和智能感知等多种用例。我们通过深度学习和传统机器学习工作负载，展示了 TinyVers 在机器监测、异常检测、音频信号分析和图像分类等多种应用中的能力，以及相较于最先进水平（SotA）的改进。

本文其余部分安排如下：第二节介绍机器学习计算内核的基础知识。第三节概述 TinyVers 的架构，第四节进一步介绍 FlexML 加速器。第五节详细说明如何在 TinyVers 上实现用于机器学习部署的软件栈。随后，第六节展示映射不同工作负载和应用场景的实验结果。最后，第七节将 TinyVers 的性能与相关工作进行比较，第八节总结全文。

II. 算法背景

机器学习应用大量采用深度神经网络（DNN），其中包含传统的卷积网络（CNN）层和全连接（FC）层。然而，各种新型神经网络层拓扑结构不断涌现。例如，时域卷积网络（TCN）用于关键词检测等音频任务 [19]–[21], ；自编码器（AE）则通过卷积与反卷积层对，用于机器监测和异常检测任务 [22]–[24]。此外，极端边缘物联网节点中仍在使用不依赖神经网络层的机器学习模型，例如用于新颖性检测和异常检测应用的支持向量机（SVM）[25]。将所有这些工作负载部署到专用加速器上，可使其执行效率提升数个数量级。然而，计算内核种类繁多，使得难以将其高效映射到单一硬件平台上。以下各小节将讨论不同机器学习运算的特性、它们如何归类为数学运算，以及由此产生的硬件影响。

A. 卷积与密集运算

卷积层和全连接层是深度神经网络中最常用的计算内核，可分解为矩阵乘法（MMM）和矩阵向量

矩阵向量乘法（MVM），依此类推。这两种矩阵运算可以表示为嵌套的*for* 循环，如图 1 所示。大多数机器学习计算内核都可归类为这两种数学运算之一，但有些特殊层需要额外的硬件改动。其中一种核函数是 TCN 层，它可以表示为一维卷积神经网络，并且需要额外支持可编程膨胀，这与卷积中的步幅类似。长短期记忆网络（LSTM）和门控循环单元（GRU）等循环神经网络（RNN）可以分解为矩阵向量乘法（MVM），但需要额外的硬件来支持激活函数。这些硬件改动将在第 IV 节进一步讨论。

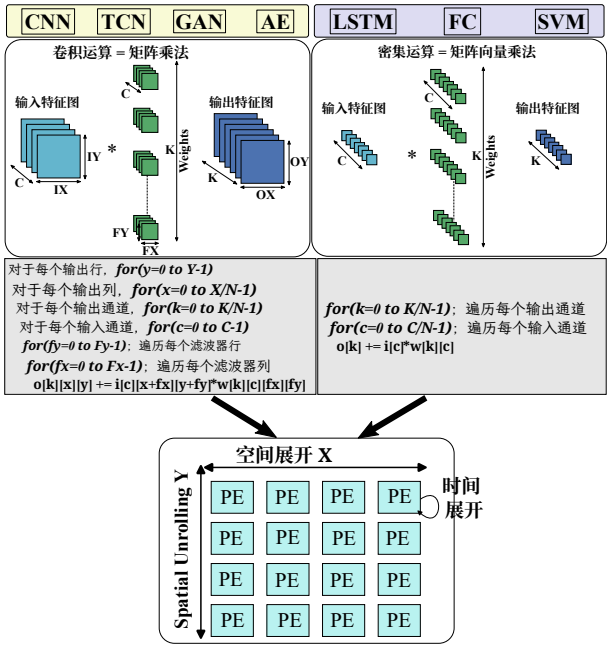


图 1. 不同机器学习模型及其以矩阵乘法（MMM）和矩阵向量乘法（MVM）表示的数学形式。嵌套的*for* 循环表示可通过空间展开和时间展开映射到专用加速器上。

将矩阵乘法和矩阵向量乘法（MVM）映射到专用硬件加速器时，嵌套的*for* 循环可以在空间和时间上展开，这在文献中称为数据流 [26]。在二维处理单元（PE）阵列上，可以对两个*for* 循环进行空间展开，即沿 X 和 Y 维度并行化循环，如图 1 所示。本文其余部分将这种空间展开表示为（空间展开 X）|（空间展开 Y）。其余*for* 循环则进行时间展开，即顺序执行。根据可用的并行性和可复用性，空间展开（X 和 Y）需要可配置，以便高效映射第 IV-B 节详述的所有工作负载。

B. 反卷积

许多机器监测应用中使用的自编码器由编码器和解码器组成，旨在重构输入数据。在正常数据上训练后，测试数据中的重构误差可用于指示异常。反卷积或转置卷积应用于这些自编码器中，通过将卷积和上采样合并为单一操作实现。反卷积可映射为卷积（MMM），但为实现高效映射，需要额外的硬件来支持输入的零值跳过。硬件改造可以提高该操作的映射效率，并更好地利用其固有稀疏性，详见第 IV-C 节。

C. 支持向量机（支持向量机）

支持向量机是用于分类和回归任务的机器学习算法。当需要对正常行为与异常之间的输入数据进行分类时，可以使用一种名为单类支持向量机（OC-SVM）的二元分类器

来完成 [27], [28]。使用径向基函数（RBF）核函数的单类支持向量机决策函数由公式（1）给出。对于拉普拉斯核，L2 范数替换为 L1 范数。

$$f(x) = \sum_{i=0}^{N} \alpha_i \cdot \exp^{\frac{-\|x - sv_i\|_2}{2\sigma^2}} - b \tag{1}$$

其中，x 是长度为 D 的输入向量，sv 是长度为 D 的支持向量，N 是支持向量的数量，σ 是标准差，α 是拉格朗日乘子，b 是偏置。在这些工作负载中，支持向量的数量 N 与向量长度 D 相结合后可能会变得很大，使 L1 和 L2 范数计算变得复杂；将其部署在专用加速器上，可使性能提升数个数量级。范数运算的 D 和 N 维度可按与全连接层（矩阵向量乘法）的 C 和 K 维度类似的方式处理，并可在处理单元阵列上进行空间展开。除了展开范数运算外，还需要在每个处理单元中添加用于支持平方、减法、舍入和绝对值运算的额外硬件。随后，中央处理器核心可使用范数计算结果来计算整体核函数。

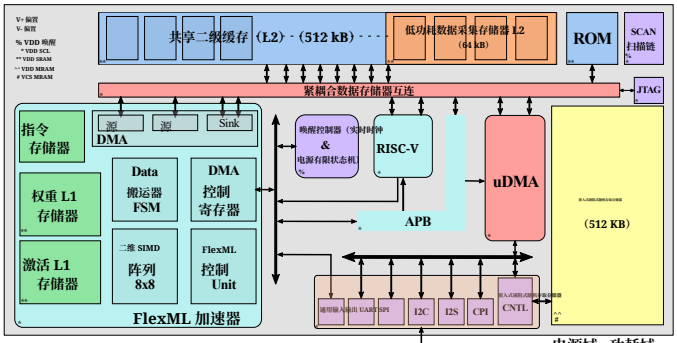
D. 结构化稀疏性

利用深度神经网络中的稀疏性，通过跳过零值并压缩神经网络参数，有助于降低计算复杂度和内存需求。然而，随机剪枝或非结构化稀疏性往往难以高效映射到硬件上，并且需要专门的逻辑来实现零值跳过和负载均衡 [29]–[31]。稀疏结构（剪枝粒度）会显著影响硬件效率和预测准确率。一些研究发现，非结构化稀疏性能够实现比结构化稀疏性更高的预测准确率，但结构化稀疏性更适合硬件实现，并能提高计算效率 [30]。因此，可以通过更多轮次的迭代训练结构化稀疏模型，使其恢复到接近原有的预测准确率，同时获得相近的整体效率/成本。此外，更粗粒度的稀疏性可以减少存储非稀疏数据索引所需的额外内存。

考虑到这些不同的机器学习工作负载及其特性，需要设计一个能够高效映射上述所有工作负载的平台。

III. TINYVERS 硬件架构

如图 2 所示，TinyVers 是一种异构片上系统，由一个单核 RISC-V 处理器、一个名为 *FlexML* 的灵活机器学习加速器、一个 512 kB 共享二级缓存（L2）静态随机存取存储器，以及一个用于数据移动的微型直接存储器访问（uDMA）组成在外设与存储器之间，配置了用于非易失性存储的 512 kB 嵌入式磁阻随机存取存储器，以及用于电源管理的唤醒控制器（WuC）。该片上系统的开发基于 PULPissimo 平台 [32]。其中集成了一个 2 kB 只读存储器（ROM），用作第一阶段引导加载程序（FSBL），同时还控制从 JTAG、外部 SPI 闪存或嵌入式磁阻式随机存取存储器启动。系统采用两种通信总线：1.) 对数互连，可提供紧耦合数据存储器（TCDM），实现单周期访问共享二级缓存；2.) APB标准总线，用于控制不同的存储器映射模块。



功耗模式	逻辑 PD	L1 PD	数据采集存储器 PD	低功耗数据采集存储器	MRAM PD	UDMA PD	常开电源域
Boot	ON	ON	ON	ON	ON	ON	ON
活动	ON	ON	ON	ON	开启/关闭	ON	ON
数据采集	OFF	OFF	ON	ON	OFF	ON	ON
低功耗数据采集	OFF	OFF	OFF	ON	OFF	ON	ON
深度睡眠	OFF	OFF	OFF	OFF	OFF	OFF	ON

图 2. TinyVers 完整片上系统概览，展示了不同的电源域 (PD)、各电源域包含的模块以及所支持的功耗模式。

片上系统与 FlexML 加速器之间的接口基于 [33] 中介绍的 HWPE 框架。借助 [33]，中的流传输器，数据可在 FlexML 的 DMA 引擎帮助下与共享L2内存相互传输。该 DMA 引擎是一个有限状态机，负责控制其私有存储器的数据加载与卸载以及双缓冲操作。片上系统支持多种外设接口协议，包括 UART、SPI、I2C、I2S 和 CPI，此外还配备了 32 个通用输入输出接口（GPIO）。主核逻辑、外设接口和常开域分别使用独立时钟；常开域包含唤醒控制器和输入输出焊盘。

A. 微型机器学习的智能感知模式

物联网微型机器学习应用通常通过传感器阵列在指定时间窗口内采集数据，随后处理采集到的数据并据此做出决策。在许多应用中，开始处理前所需的数据采集时间窗口可能从几 ms 到几 sec 不等。此外，在传感器数据采集期间，由于尚未进行繁重的处理，微控制器的许多模块都不会被使用。这为许多微型机器学习应用节能提供了机会：数据采集期间，只需让将窗口化数据从传感器外设接口传输到存储器所必需的模块保持活动，例如，可以让中央处理器进入睡眠状态。此外，对于音频等处理时间序列数据的应用，窗口化数据所需的存储空间较小（< 64 kB），因此也可以关闭微控制器主存储器的大部分电源，以避免未使用存储器区域产生漏电功耗。

为此，TinyVers 引入了两种针对 tinyML 优化的数据采集电源模式：1.)“数据采集”和 2.)“低功耗数据采集”，如图2所示。数据采集模式面向视觉等样本数据量较大的应用，会保持 uDMA 模块和完整的共享二级缓存（512

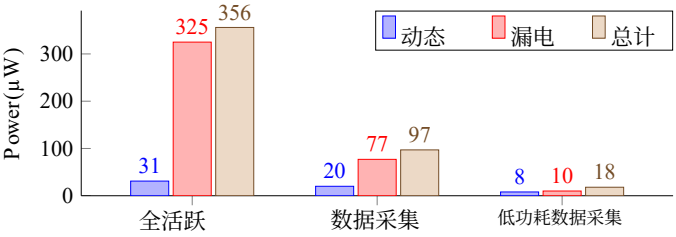


图3. 使用 Cadence Genus 工具对综合后网表进行的三种功耗模式功耗仿真。三种模式下，I2S 数据均以 44.1 kHz 的采样频率采集，持续时间为 2 秒。报告的全活跃功耗包含 RISC-V 核心对微型直接存储器访问的配置以及中断处理过程，此外还包括数据采集。

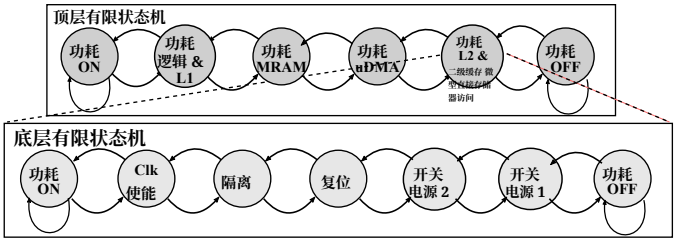


图4. 展示 WuC 中所用层次化有限状态机的流程图。

kB）处于通电状态。与之相对，低功耗数据采集模式除了 uDMA 外，仅保持部分共享二级缓存（64 kB）处于通电状态。此模式专为需要时间序列和音频数据的应用而设计，例如关键词检测、机器监测、生物信号分析等。图3估算了从全活跃模式切换到两种 tinyML 感知模式时可实现的节能效果：从全活跃模式切换到数据采集模式，功耗改善近 3.5× ；从数据采集模式切换到低功耗数据采集模式，功耗改善 5.5× 。

B. 电源管理

TinyVers 在标准低功耗设计的基础上，采用了积极的电源管理策略。如图2所示，片上系统被划分为 6 个可切换电源域和 1 个常开域（始终开启）。每个可切换电源域由多个电源门控开关组成，这些开关将该电源域的 VDD 与全局VDD电源隔离开来。这些电源门控开关由常开域中的唤醒控制器发出的控制信号控制。电源域之间的所有互连跨越均配备了双向电平转换器和隔离单元，因此可以独立控制各域的供电电压。

智能 WuC 负责这项电源管理控制，并依靠实时时钟计数器（RTC）运行。该计数器可由 RISC-V 核心以毫秒为粒度进行编程。RISC-V 核心可指示 WuC 使片上系统进入图2所示的五种支持的功耗模式之一。为此，WuC 包含由 RTC 驱动的分层有限状态机（FSM），如图4所示，用于控制整个片上系统及各个电源域的上电和断电。顶层有限状态机控制各电源域上电/断电的时序，而底层有限状态机则控制各个电源域隔离单元和电源门控开关的启用/停用精细时序。

新兴存储器（如阻变存储器、磁阻随机存取存储器、铁电存储器、相变存储器等） [34], [35], 在构建经济高效的嵌入式非易失性存储器（NVM）方面展现出良好前景，可面向汽车或工业 4.0 边缘计算应用。NVM 存储器可用作启动代码及其他需要保存的参数的存储空间。这带来两项优势：1.) 对于无需始终开启运行的应用，可采用占空比循环来降低功耗；2.) 片上系统在电源循环后，无需访问中央云服务器即可获取启动代码和神经网络参数。此外，片上嵌入 NVM 后，无需从片外获取数据，从而避免了高昂的能耗。

磁阻随机存取存储器被誉为通用存储器，它利用位单元中的磁极性来存储数据 [36]。它具有非易失性，密度几乎与传统静态随机存取存储器一样高，因此非常适合用于极端边缘片上系统的微型机器学习应用。有鉴于此，TinyVers 在芯片上集成了 512 kB 嵌入式磁阻式随机存取存储器，可为智能感知和按需计算实现极端电源管理策略。在片上系统中，嵌入式磁阻式随机存取存储器作为非易失性存储，用于存放 RISC-V 唤醒并开始处理所需的启动代码，也可存储映射后的机器学习工作负载的神经网络参数。此外，嵌入式磁阻式随机存取存储器还可用作非易失性暂存区空间，以存储智能感知应用中的窗口化数据。嵌入式磁阻式随机存取存储器与共享L2内存之间的接口使用微型直接存储器访问单元，该设计基于 [17] 的工作。

IV. FLEXML 加速器

本节首先介绍 FlexML 加速器的架构概述，随后介绍用于灵活映射的数据流重配置、用于反卷积和结构化稀疏性的高效零值跳过，最后介绍支持 SVM 的硬件，正如第二节中简要讨论的那样。

A. FlexML 架构概述

FlexML 加速器是 TinyVers 专用的多功能硬件加速器。FlexML 旨在高效支持微型机器学习应用中多种多样的机器学习工作负载，同时利用各层特性中存在的数据复用。这通过零延迟运行时数据流重配置实现，详见第四节 B。如图 5 所示，FlexML 包含一个 8×8 单指令多数据（SIMD）处理单元（PE）阵列，其中每个处理单元由一个精度可扩展的乘累加（MAC）单元组成，支持 INT 8/4/2 [37],，如图 6 所示。得益于精度可扩展性，SIMD 阵列可重配置为8×8/16/32 个 INT8/4/2 MAC 单元的阵列。每个处理单元根据所选精度（INT8/4/2）每周期执行 1/2/4 次乘加运算，并将结果累加到一个 32 位寄存器中，同时采用完全或部分输出驻留，以减少大位宽部分和的数据移动开销。最终输出会经过 ReLU 函数（若已使能），随后重新量化至所选精度并写入

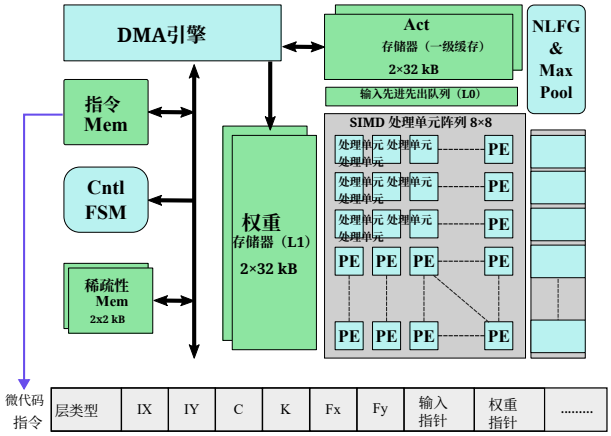


图 5. FlexML 加速器架构概览及微码指令。

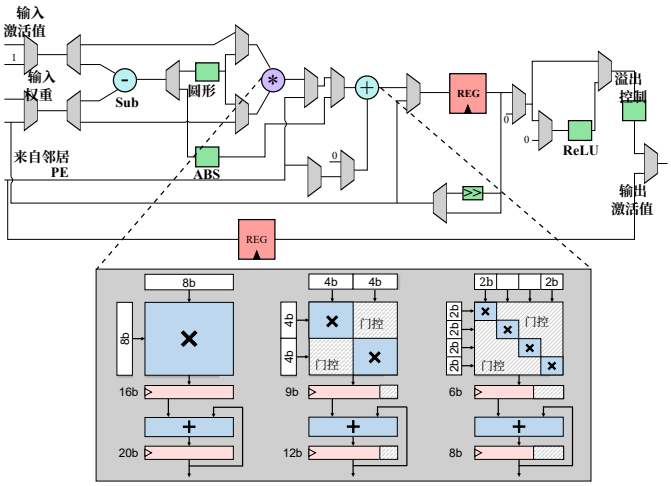


图 6. FlexML 加速器中所用处理单元的框图，展示了精度可扩展乘加单元以及用于支持 SVM 的额外硬件。

回激活 L1。混合精度量化有助于在精度低于 8 位时提升深度神经网络模型的性能。然而，由于带宽变化和串行数据流 [38]，混合精度的硬件开销可能会降低处理单元的整体效率。因此，FlexML 仅支持权重和激活值采用对称精度。此外，输出归一化采用简单的移位和 ReLU，也使硬件开销保持较低。为保持模型准确率，我们采用了第五节中提到的硬件感知训练框架。

为 SIMD 处理单元阵列提供支持的是基于 SRAM 的专用一级缓存（L1）存储器，用于存储权重（64 kB）和激活值（64 kB）。权重 L1 和激活 L1 均由两个 32 kB 存储体组成，以乒乓方式运行，使数据写入和读取能够重叠，从而提升整体性能。在激活 L1 和处理单元阵列之间还设有中间存储层 L0。该 L0 存储器是一个大小为 16×8 位的 FIFO 缓冲区，用于在卷积中执行滑动窗口操作时提升数据局部性。此外，还配备了独立的非线性函数发生器（NLFG）和最大池化单元。

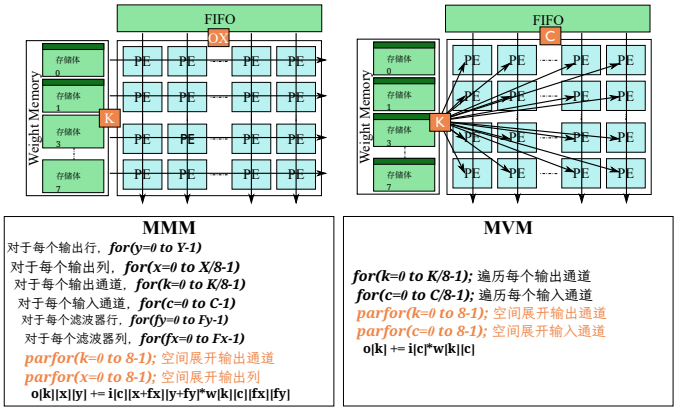


图 7. 展示用于切换数据流的图示：从用于矩阵乘法的 OX|K 数据流（左侧）切换为用于矩阵向量乘法的 C|K 数据流。下方嵌套的 *for* 循环展示了为实现空间展开而添加的并行 for 循环。

NLFG 使用基于查找表的线性近似，生成神经网络模型中使用的各种激活函数（ReLU 除外），例如双曲正切、Sigmoid 等。为控制加速器内部的数据流和控制流，配备了一个包含有限状态机的控制单元，用于从指令存储器中获取微代码指令、解码指令，并通过更新控制信号和跟踪工作负载的计数器，在处理单元阵列上部署相应的层。微代码指令由以 Python 编写的伪编译器生成（第五节），其形式为包含超参数的类 CISC 的逐层长指令，如图 5 所示。控制单元还进行了扩展，以支持在反卷积时高效跳过激活值中的零，并结合稀疏索引存储器跳过剪枝权重（第四节 C 部分）。

B. 数据流重配置

为高效映射种类繁多的机器学习工作负载，FlexML 加速器支持运行时数据流重配置，且不会带来延迟开销。该可配置性可高效映射以下两类工作负载：1.) 用于卷积神经网络、反卷积和时间卷积网络的矩阵乘法，在具有输出驻留的数据流 OX|K 下利用输入和权重的空间数据复用；2.) 用于批大小为 1 的全连接、循环神经网络以及支持向量机范数计算的矩阵向量乘法，在具有部分输出驻留的数据流 C|K 下利用可用的输入空间数据复用。此前已有多项研究提出在硬件中进行数据流重配置，以实现不同工作负载的最佳映射 [39]–[41]。然而，这些研究在支持多种数据流时存在较大的硬件开销和延迟，不适用于极端边缘设备。本研究将数据流限制为两种最优映射方案，从而降低硬件和功耗开销。此外，此前没有研究探讨在同一硬件加速器上映射时间卷积网络、自编码器和支持向量机。图 7 展示了 OX|K（左）和 C|K 数据流（右）及其硬件实现。在 OX|K 数据流中，对嵌套*for* 循环的 OX 和 K 维度进行空间展开，将展开后的 OX 维度分配到 SIMD 处理单元阵列的列上，并将展开后的 K 维度分配到其行上，阵列维度为 8×8。

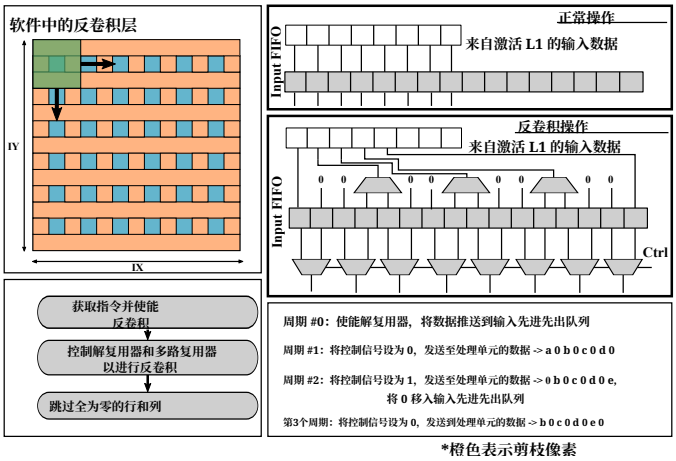


Fig. 8. 软件中反卷积层的表示（左上）、执行零跳过操作的控制单元（左下）、支持反卷积所需的 L0 FIFO 架构变更（右上），以及 FIFO 和处理单元逐周期的运行情况（右下）。

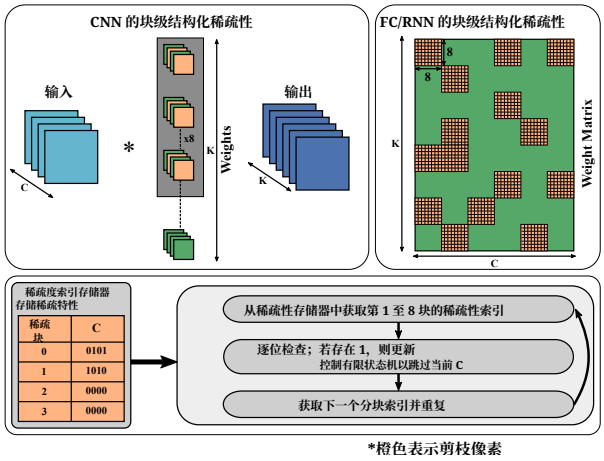


图 9. 应用于卷积神经网络和全连接层的块状结构化稀疏性（上）；控制单元与稀疏性索引存储器协同工作以支持零值跳过（下）。

其余 for 循环按图 7 中嵌套的*for* 循环所示进行时间展开，从而形成输出驻留数据流。在这种数据流模式下，激活 L1 内存将输入激活数据沿垂直方向多播至 L0 FIFO 内存；在第一个周期提取 8 个字，随后在滑动窗口操作期间每次提取一个字，从而利用复用机会降低内存带宽需求和内存提取次数。权重 L1 内存沿水平方向提供数据，通过 2 个内部存储体提供 8 个字，并将每个字沿行多播。由于采用输出驻留，累加会持续进行，直至生成最终输出；随后，输出以脉动方式沿垂直方向移出至激活 L1，完成输出回写需要 8 个周期。L0 FIFO 内部的输入数据移位可编程，以支持时间卷积网络中使用的可变膨胀率，或一般情况下的可变步幅。

对于 MVM，采用另一种 C|K 数据流，因为由于权重缺乏可复用性，此工作负载无法高效利用 OX|K 数据流。在这种数据流下，C 维沿垂直列方向进行空间展开，K 维沿水平行方向展开。激活值 L1 内存沿垂直方向多播 8 个输入激活值字，同时绕过 L0 FIFO 内存。批大小为 1 时，权重无法复用，因此每个处理单元每个周期都需要一个新权重。为满足这一要求，权重内存使用全部 8 个存储体，向处理单元单播 64 个不同的权重字。处理单元的各行对同一输出通道 (K) 的不同输入通道 (C) 进行运算。因此，每个处理单元完成所需的乘加运算后，同一行中各处理单元的输出通过加法树累加，每行的一个最终输出再移位至激活值存储器。

C. 反卷积与块状结构化稀疏性的高效零值跳过

FlexML 加速器支持对反卷积工作负载进行高效的零值跳过。如图 8 所示，输入先进先出队列经过专门设计，在反卷积模式下只获取一组字，并通过零填充对其进行重排。控制单元会跳过包含零值、会导致冗余计算的行和列，与在卷积模式下通过上采样运行反卷积相比，性能最高可提升 $2\times$ 。

TinyVers 还支持结构化稀疏性，更具体地说，支持卷积层和全连接层中的块状核级稀疏性（2D） [29], [31]。如图 9 所示，在此方案中，滤波器核的完整输入通道会被剪枝，并要求 8 个滤波器核（$K = 8$）以相同方式剪枝。块大小由处理单元阵列的维度以及 K 沿 2D 处理单元阵列水平方向的空间展开方式决定。在我们的实现中，所选块大小使数据流和控制流更易于控制。对处理单元阵列上并行映射的全部 8 个滤波器核应用相同的通道剪枝，可提高映射效率，因为所有行仍可共用控制逻辑运行；这不仅能够节能，还能提升吞吐量。为此，FlexML 加速器配备了专用的稀疏性索引存储器，用于存储剪枝通道组的位编码索引。图 9 展示了稀疏性索引存储器，以及控制单元中使用的控制流逻辑。每次滤波器核块递增之前，控制单元都会读取一个索引存储器字，并随着输入通道递增逐位检查数据，以确定稀疏状态。如果检测到稀疏通道，则跳过该通道的全部计算，从而避免对零值进行计算。

D. 支持向量机

OC-SVM 的 L1 和 L2 范数需要修改处理单元，以便使用相同的硬件映射工作负载。如图 6 所示，每个处理单元都扩展了一个减法模块、一个绝对值单元和一个舍入单元，并对乘法器进行修改，使其也能执行平方运算，从而在处理单元阵列中计算范数。输入数据向量 x 和支持向量 sv_i 的维度为 D，支持向量的数量为 N。在使用 C|K 数据流时，

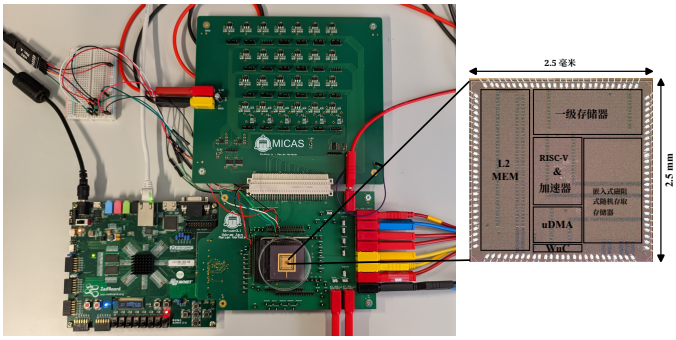


图 10. 测量设置和芯片显微照片。

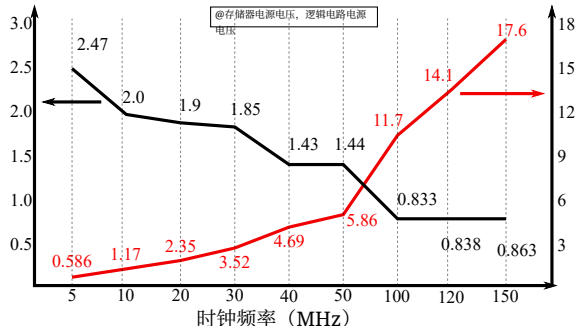


图 11. 卷积神经网络3×3 层的峰值性能分析。

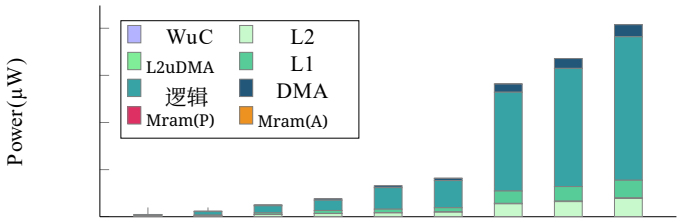


图 12. 使用卷积神经网络3×3进行峰值性能分析时的功耗分布。活动模式下，MRAM处于关闭状态，因此其功耗可以忽略不计。MRAM(A) 和 MRAM(P) 分别表示 MRAM 阵列和 MRAM 外围电路。

D 维输入数据向量 x 沿处理单元阵列垂直展开并进行多播（C），而支持向量 sv_i 的 N 维则沿水平方向展开并进行单播（K）。处理单元中计算出的 N 范数结果随后发送到共享二级缓存，再由 RISC-V 核心使用 GNU C 内置的指数函数进行后处理，与 α 相乘，并对 N 求和，最终生成方程（1）所示的输出。

V. 神经网络在 TinyV 上的神经网络部署 ERS

用于机器学习应用的硬件还需要一个用户可编程全栈，能够将机器学习算法直接从现有的机器学习训练和推理框架（如 TensorFlow、Keras、PyTorch 等）转换而来。这使得各种机器学习工作负载能够快速、轻松地部署到现有硬件上成为可能。我们为 TinyVers 创建了一个基于 Python 的伪编译器框架，并考虑了其异构性。首先，使用 QKeras 框架 [42] 进行量化感知训练，将机器学习算法量化到选定精度。量化感知训练框架会考虑硬件约束，例如对称量化，以及加速器处理单元中基于移位的输出缩放。随后，量化模型会被传入基于 Python 的神经网络编译器；该编译器接收硬件描述，并为 RISC-V 核心提供一组基于 C 的头文件，其中包含加速器的微代码指令、神经网络参数，以及用于验证映射后工作负载的黄金模型。

VI. 芯片实现与测量

图 10 所示的 TinyVers 芯片显微照片是在 GlobalFoundries 22FDX™ 工艺中实现并制造的。图中展示了片上系统中使用的各个子模块，这些模块已在前文详细介绍。图 10 还展示了用于测量和基准测试的实验室设置。以下各小节将详细介绍针对片上系统开展的功耗、能效和性能测量与基准测试。

A. 峰值性能分析

首先，使用一个包含 32 个输入通道、32 个输出通道和 3×3滤波器内核的卷积神经网络层进行峰值性能分析。峰值性能分析选用该层，是因为带有 3×3 滤波器内核的卷积层是现代深度神经网络模型中最常用的层。卷积神经网络层的超参数选择受以下条件驱动：处理单元阵列的利用率达到最大，以及加速器私有 L1 存储器的容量。使用 Google 语音数据集进行关键词检测，并通过编译器框架生成该卷积神经网络的 8 位量化激活值和非稀疏（结构化）权重 [43]，再与黄金模型进行比对，以验证功能正确性。

图 11 展示了在对逻辑和存储器的电源电压进行扫描时，基准测试卷积神经网络层的峰值能效和吞吐量随时钟频率的变化情况。为与其他最先进水平的芯片进行公平比较，未施加体偏置。图 12 展示了运行基准测试层时各个模块的功耗分布。该片上系统具有很大的性能调节范围：当时钟频率为 5 MHz、逻辑电压为 0.4 V、存储器电压为 0.5 V 时，可实现 2.5 TOPS/W 的高能效和 586 每秒百万次操作的低吞吐量；而当时钟频率为 150 MHz、逻辑和存储器电压均为 0.8 V 时，则可实现 0.8 TOPS/W 的低能效和 17.6 每秒十亿次操作的高吞吐量。这为极端边缘端的微型机器学习应用提供了广泛的运行范围，可在速度与能效之间进行权衡。

B. 工作负载基准测试

采用第 VI-A 节中的峰值能效工作点（5MHz，逻辑电压 0.4 V，存储器电压 0.5 V），对不同的合成基准测试和实际实时基准测试进行了进一步的性能分析。表 I 展示了片上系统通过映射不同机器学习层和完整

表 I 工作负载基准测试

	工作负载	Acc.	功耗 (μW)	Peak 性能。（每秒十亿次操作）	Peak (有效非零值) 能效 (TOPS/W)
Synthetic	卷积神经网络@8 位	-	237	0.586	2.47(2.47)
	卷积神经网络@4 位	-	197	1.17	5.94(5.94)
	卷积神经网络@2b	-	197	2.35	11.9(11.9)
	卷积神经网络@8 位，50% 稀疏	-	239	1.03	4.31(2.46)
	卷积神经网络@8 位，87.5% 稀疏	-	212	3.64	17.1(2.76)
	全连接层/循环神经网络/支持向量机，批次=16	-	140	0.116	0.829(0.829)
	反卷积@8 位	-	235	1.36	5.78(2.49)
Real-time	时间卷积网络（关键词识别）	93.3%*	193	0.204	1.05(1.05)
	CAE	-	209	0.442	2.11(1.27)
	ResNet-8	82%+	228	0.267	1.17(1.17)
	单类支持向量机	-	129	0.126	0.972(0.972)

* 12分类任务，基线=93.46%，+ 基线=85%

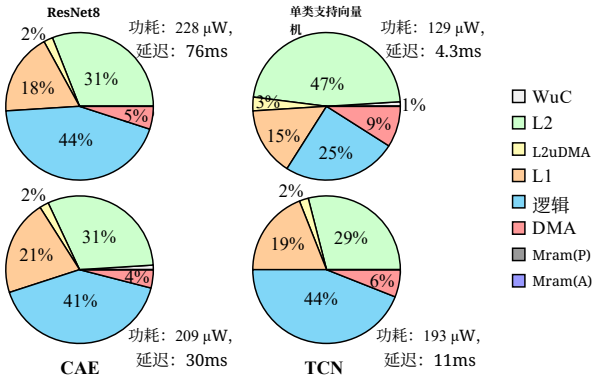


图 13. 能耗分解，展示了在 FlexML 和 RISC-V 上运行四种实时工作负载的单次推理时芯片模块的实测能耗分布，输入数据已预先存储在 L2 存储器中。功耗和延迟测量从 RISC-V 设置加速器参数开始，涵盖从 L2 到 L1 的数据移动、推理计算，并在 RISC-V 核心完成后处理后结束。由于磁阻随机存取存储器在活动模式下处于关闭状态，其功耗可以忽略不计。

工作负载所体现的灵活性。对第 VI-A 节中的卷积神经网络层进行了扩展，并测量了不同精度和块状结构化稀疏性（BSS）等级下的表现。精度降至 INT-4 和 INT-2 时，峰值吞吐量分别提升 2× 和 4×，峰值能效则分别提升 2.4× 和 4.8×，在 INT-2 下达到最高 11.9 TOPS/W。如表 I 所示，在 8 位精度、50% BSS（剪除 32 个输入通道中的 16 个）时，性能提升约 1.7× ；在 87.5% BSS（剪除 32 个输入通道中的 28 个）时，性能提升约 6.9×。进一步降低精度可提升性能，但低精度与高 BSS 等级结合可能导致准确率大幅下降，因此本次基准测试未对此进行探索。还对其他合成基准测试进行了研究，包括全连接、循环神经网络、支持向量机，以及一个在超参数方面与该卷积神经网络层相似的反卷积层，结果如表所示。对于全连接层，采用 16 的批次大小。

最后，采用 4 个实时应用基准测试来

表 II 不同低功耗模式的测量结果。

功耗模式	AON 频率 (kHz)	Core 频率 (MHz)	功耗 (μW)	唤醒延迟 (μs)
深度睡眠	33	-	1.7	788
低功耗数据采集*	33	5	23.6	788
数据采集*	33	5	67	788

* @F_s=44.1 千赫兹

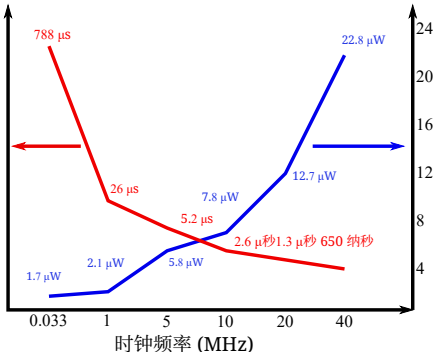


图14. 深度睡眠功耗-延迟-频率权衡。

展示该 SoC 的能力：1.) 使用 TCN 模型在 Google语音数据集上进行关键词识别（KWS） [21], [44]；2.) 使用卷积自编码器（CAE） [24] 在 MIMII数据集 [45]，3上进行连续机器监测；3.) 在 MLPerf™ 微型基准测试 [46],中使用 CIFAR-10 进行 ResNet-8 图像分类；4.) 使用单类支持向量机（OC-SVM） [47]进行新颖性检测。表 I 展示了这些基准测试在该 SoC 上的峰值性能特征，更具体地说，是 RISC-V 核心和 FlexML 在 8位精度、单次推理，并假设所有输入数据均位于共享L2内存中的性能特征。对于 TCN 和 ResNet-8，采用了硬件感知量化，并测量了能耗和性能指标；而对于 CAE 和 OC-SVM 工作负载，则使用了随机输入和权重。全部 4 个工作负载均可在低于 230 μW 的连续实时功耗下部署，并在 1-2 TOPS/W 之间实现峰值能效。这意味着，该 SoC 能够以低于 mW 级的功耗提供高度灵活的工作负载映射能力，从而在极端边缘设备上实现真正高能效的微型机器学习应用。图 13 展示了 4 个实时工作负载的功耗分布。对于 OC-SVM（密集运算），由于权重缺乏可复用性，导致需要进行更多数据读取，因此存储器的功耗占主导地位。另一方面，基于 CNN 的工作负载（TCN、ResNet8 和 CAE）的功耗分布显示，存储器与逻辑的功耗基本均等，这是因为数据流充分利用了最大可复用性。

C. 电源管理

表 II 展示了第 III-B 节所述片上系统各低功耗模式的实时功耗测量结果。在深度睡眠模式下，片上系统以 33 kHz 的始终开启时钟频率运行。在此模式下，只有由唤醒控制器和控制输入输出焊盘的逻辑组成的始终开启域保持供电

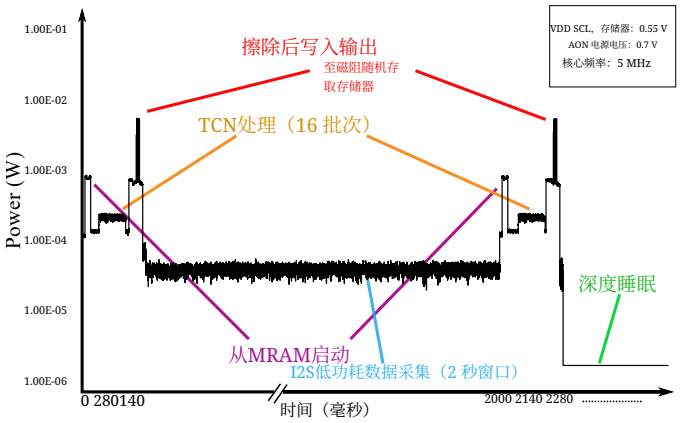


图 15. 瞬时功耗波形，展示了关键词识别应用场景：完成一个完整的智能感知和TCN处理周期后进入空闲状态。

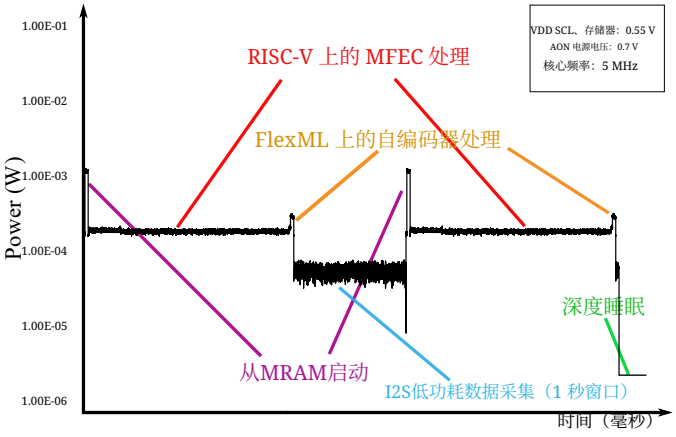


图 16. 瞬时功耗曲线，展示机器监测应用场景中的一个周期，包括智能感知、特征提取和卷积自编码器处理，随后进入空闲状态。

开启。在 0.7 V 电源电压下运行时，测得的深度睡眠功耗为 1.7 μW。与卷积神经网络层测得的峰值功耗相比，深度睡眠功耗低 12,000× 。测得的片上系统从深度睡眠模式唤醒至活动模式的延迟为 788 μs。通过调整始终开启时钟频率，可以在唤醒延迟与深度睡眠功耗之间进行权衡。图14展示了调整始终开启时钟频率时，测得的功耗与唤醒延迟之间的这种权衡。对低延迟有要求的应用可将始终开启时钟设为 40 MHz，从而以 22.8 μW 的实时功耗实现 650 ns 的唤醒延迟。

表 II 还展示了两种面向微型机器学习优化的功耗模式——数据采集和低功耗数据采集——的实测功耗。这些功耗模式是在基于 I²S 协议的窗口化测试向量采集中测得的，其中始终开启时钟频率为 33 kHz，核心和外设时钟频率为 5 MHz。片上系统经过编程，可通过其微型直接存储器访问以 44.1 kHz 的采样频率和 2 秒的采样窗口采集 I²S 音频数据。采样时钟由片上系统使用 5 MHz 时钟生成，并持续整个采样窗口。随后启动数据采集或低功耗数据采集模式并测量功耗。低功耗数据采集和数据采集的实测功耗分别为 23.6 μW 和 67 μW，相比峰值功耗分别降低了 850× 和 300× ；在此情况下，核心和外设频率可动态降低至 5 MHz。

D. 瞬时功耗曲线

为了展示可部署于片上系统上的完整端到端应用，并体现片上系统的全部机器学习功能、占空比循环和电源管理特性，我们在异构片上系统上映射了两个应用，并在低功耗数据采集模式下进行窗口化数据采集：采用时间卷积网络模型进行关键词检测，持续运行 [21]；以及一个机器监测用例，该用例基于梅尔频率能量系数（MFEC）进行特征提取，并在占空比循环模式下使用卷积自编码器 [24]。

1）关键词识别应用：第一个应用场景是使用 TCN 模型进行关键词识别。在此应用场景中，使用 I2S 外设接口协议采集麦克风的音频数据，窗口大小为 2 秒（16 个批次），采样频率为 44.1 kHz；采集的数据在片上系统处于低功耗数据采集模式时，通过片上系统的微型直接存储器访问，同时存入专用的二级缓存微型直接存储器访问存储器中。2 秒后，片上系统唤醒并进入活动模式，随后使用第六节 B 部分所述的 TCN 模型处理采集的数据。TCN 处理的输出随后存储到磁阻随机存取存储器中，以供后续处理或传输；与此同时，片上系统可以进入深度睡眠模式，也可以采集新的窗口采样数据。图 15 展示了关键词识别应用场景的完整瞬时功耗曲线。在这种占空周期运行模式下，整个应用的平均功耗为 173μW。未进行感测或计算期间，使用片上系统的深度睡眠功耗模式可进一步将功耗降低至 10-20 μW。

2）机器监测应用：用于预测性维护的机器监测是选定的第二个应用场景。在此场景中，使用 I2S 外设接口协议从麦克风采集音频数据，窗口大小为 1 秒，采样频率为 16 kHz。I2S 音频数据的采集在片上系统的低功耗数据采集模式下进行。收集完完整的窗口化数据后，片上系统切换至活动模式，使用 RISC-V 核心进行基于 MFEC 的特征提取，随后在加速器上运行 CAE。图 16 绘制了运行机器监测应用时的瞬时功耗曲线。与前一个处理原始音频数据的应用不同，CAE 模型需要对 MFEC 数据进行预处理。由于加速器不支持 MFEC 算法，因此使用 INT16 精度而非 INT32 或 FP32，在 RISC-V 核心上执行该算法以降低功耗 [52]。功耗曲线显示，在 RISC-V 上进行大规模特征提取并不节能，因为单核心运行需要较长时间才能完成。连续运行时的平均功耗低于164 μW，但在此用例中，占空比为 0.05 时功耗为 9.5 μW。RISC-V 上的 MFEC 执行

表 III 与最先进水平的性能比较。

	[48]	[17]	[49]	TinyVers	[50]	[51]
	极端边缘片上系统				边缘机器学习加速器	
工艺技术	28 纳米 FDSOI	22FDX	55nm	22FDX	28nm	65nm
芯片面积 (mm^2)	4.5	12	10	6.25	0.55	16
应用	物联网高斯过程、深度神经网络、NSA	物联网高斯过程、深度神经网络、NSA	物联网高斯过程、深度神经网络、NSA	物联网高斯过程、深度神经网络+、传统机器学习、NSA	始终开启的关键词检测	DNN
支持机器学习层	CNN、全连接层/循环神经网络	CNN、全连接层/循环神经网络	CNN、全连接层/循环神经网络	CNN、全连接层/循环神经网络、生成对抗网络、AE、TCN、SVM	深度可分离卷积神经网络	CNN、全连接层/循环神经网络
架构	1×RI5CY+机器学习加速器	10×RI5CY+机器学习加速器	9×RI5CY	1×RI5CY+FlexML 加速器	深度可分离卷积神经网络加速器	DNN 加速器
SRAM（状态保持）	464 kB（40 kB）	128 kB（一级缓存）（16-1600 kB（二级缓存））	64 kB（一级缓存）（512 kB（二级缓存）	132 kB（一级缓存）（64/512 kB（二级缓存））	2 kB	256 kB
eNVM	-	4 MB 磁阻随机存取存储器	-	512 kB 磁阻随机存取存储器	-	-
深度睡眠功耗（μW）	-	1.7	3.6	1.7	-	-
静态随机存取存储器保持睡眠功耗（μW）	6.4	2.8-123.7	30	23.6-67	-	-
整数精度（位）	8, 16, 32	8, 16, 32	8, 16, 32	2, 4, 8	8	1-16
电源电压 (V)	0.45-0.9	0.5-0.8	1-1.2	0.4-0.9	0.41	0.63-1.1
最高频率（MHz）	350	450	250	150	0.04	200
功率范围	6.4μW-96mW	1.7μW-49.4mW	3.6μW-75mW	1.7μW-20mW	0.51μW	3.2-297mW
最佳机器学习性能	36 每秒十亿次操作 @8b*	32.2 每秒十亿次操作 @8b*	12 每秒十亿次操作 @8b*	17.6 每秒十亿次操作 @8 位**	2.3 每秒百万次操作+ @8 位**	691.2 每秒十亿次操作 @8 位 **
最佳机器学习效率 @性能	1.3 TOPS/W@ 2.8 GOPS ，8b*	1.3 TOPS/W@ 15.6 GOPS ，8b*	200 GOPS/W@ 7 每秒十亿次操作，8b*	2.47 TOPS/W@ 0.58 每秒十亿次操作，8b** 11.9 TOPS/W@ 2.4 每秒十亿次操作，2b**	4.5 TOPS/W@ 2.3 每秒百万次操作，8b**	5.57 TOPS/W，8b** 11.6 TOPS/W，4b**

+ 按 MAC 利用率为 90% 估算，* 矩阵乘法，** 卷积神经网络，1 MAC = 2 次运算

可利用 PULP 库提供的专用 DSP 扩展进行优化，这将留待未来工作。

VII. 与最先进水平的比较

表 III 从两个方面展示了我们的片上系统与最先进水平的对比：一方面与现有的极端边缘片上系统进行比较（左侧），另一方面与边缘端机器学习加速器进行比较（右侧）。与左侧的极端边缘片上系统相比，我们的片上系统在应用映射方面具有相近或更高的灵活性，同时能效和功耗表现也有显著提升。TinyVers 不仅支持物联网通用处理（GP）、深度神经网络和近传感器分析（NSA），如 [17], [48], [49],，还支持时间卷积网络（TCN）和自编码器（AE）等深度神经网络，以及支持向量机（SVM）等传统机器学习；凭借高效的映射，所有这些工作负载都能实现更高的能效。TinyVers 运行卷积神经网络层时，能效最高可达 2.47 TOPS/W，这充分体现了这一点。将卷积神经网络工作负载量化为 2 位整数后，能效可进一步提升至 11.9 TOPS/W。与极端边缘片上系统相比，TinyVers 支持精度可扩展性，因此能够利用量化提升性能。此外，借助对块结构化稀疏性的支持，TinyVers 在 8 位卷积神经网络层上的峰值性能可达 17 TOPS/W。这远高于 [17], [48], [49] 报告的能效。

与右侧的边缘机器学习加速器相比，TinyVers 在能效和功耗方面的性能指标相当，但灵活性高得多。边缘机器学习加速器只能极高效地支持单个或少数几种模型，但这种方案在部署到极端边缘设备时存在缺点。例如， [50] 只能使用深度可分离卷积神经网络执行关键词识别，其性能远低于 TinyVers，而能效相当。UNPU [51],只能支持卷积神经网络和全连接层/循环神经网络层，而且也不是完整的独立片上系统，这会影响系统级能效。此外，由于缺乏电源管理和保持存储器支持，这些边缘机器学习加速器无法支持任何形式的占空比循环。TinyVers 能以相近的 TOPS/W 能效满足极端边缘设备的多模态需求。此外，对于占空比循环应用场景，它还支持进入极低功耗空闲状态，从而实现 $< 10\mu$W 运行；第六节 D 部分对此进行了实证展示。总之，TinyVers 集极端边缘处理器和边缘机器学习加速器的优势于一身。

VIII. 结论

极端边缘端的微型机器学习应用不仅需要具备灵活加速器、能够支持多种工作负载的异构片上系统，还需要针对不同占空比循环操作进行自适应电源管理。此外，要实现此类自适应电源管理，还需要嵌入式非易失性存储器。TinyVers 在 RISC-V 核心的基础上扩展了灵活的机器学习加速器，可支持多种机器学习工作负载映射，涵盖不同的计算内核、精度和结构化稀疏性条件。此外，加入唤醒控制器和嵌入式磁阻式随机存取存储器，实现了许多占空比循环应用场景所需的自适应电源管理。测量结果表明，该芯片在 0.58 每秒十亿次操作至 17.6 每秒十亿次操作的吞吐量下，能效范围为 0.8-17 TOPS/W。不同的低功耗模式使芯片的功率范围达到 1.7μW-20 mW。机器监测应用利用了

深度睡眠模式，在占空比为 0.05 时仅消耗 9.5μW 的功耗。因此，TinyVers 朝着打造新一类超低功耗极端边缘片上系统迈出了一步。

致谢

作者谨此感谢 ETHZ 对 PULP 平台的支持，以及 GlobalFoundries 对 22FDX 流片工作的支持。本研究在与 Magics Technologies 合作的 ISAAC 项目（FODEconomie Belgium Energietransitiefonds（第二轮征集））框架下开展，并获得佛兰德政府（人工智能研究计划）的资助。

参考文献

[1] J. Portilla、G. Mujica、J.-S. Lee 和 T. Riesgo，“物联网底层的极端边缘端：综述”，IEEE 传感器期刊，第 19 卷，第 9 期，第 3179–3190 页，2019 年。[2] Y. Zhang、N. Suda、L. Lai 和 V. Chandra，“你好，边缘端：微控制器上的关键词检测”，CoRR，第 abs/1711.07128 卷，2017 年。

[3] [ ] 17~~11 071~~28 在线。网址：http://arxiv.org/abs/。P. Warden 和 D. Situnayake，《TinyML：使用 TensorFlow Lite 在 Arduino 和超低功耗微控制器上进行机器学习》。O’Reilly Media，2019 年。[4] TinyML。 [在线]。网址：https://www.tinyml.org/ [5] P. Jokic、E. Azarkhish、R. Cattenoz、E. T¨uretken、L. Benini 和 S. Emery，“用于边缘端端到端人脸分析的亚毫瓦双引擎机器学习推理片上系统”，发表于 2021 年超大规模集成电路会议，2021 年，第 1–2 页。[6] P. P. Bernardo、C. Gerum、A. Frischknecht、K. L¨ubeck 和 O. Bringmann，“UltraTrail：一种可配置的超低功耗 TC-ResNet 人工智能加速器，用于高效关键词检测 ”，IEEE 集成电路与系统计算机辅助设计汇刊，第 39 卷，第 11 期，第 4240–4251 页，2020 年。[7] J. S. P. Giraldo、S. Lauwereins、K. Badami 和 M. Verhelst，“Vocell：一款用于 10-μW 关键词检测和说话人验证的 65 纳米语音触发唤醒片上系统”，IEEE固态电路期刊，第 55 卷，第 4 期，第 868–878 页，2020 年。[8] J. Giraldo 和 M. Verhelst，“Laika：一款用于常开关键词检测的 5uW 可编程长短期记忆网络加速器，采用 65 纳米互补金属氧化物半导体工艺”，发表于欧洲固态电路会议 2018——IEEE 第 44 届欧洲固态电路会议（ESSCIRC），2018 年，第 166–169 页。[9] D. Im、G. Park、Z. Li、J. Ryu、S. Kang、D. Han、J. Lee 和 H.-J. Yoo，“DSPU：一款 281.6mW 实时深度信号处理单元，用于移动平台上基于深度学习的密集 RGB-D 数据采集、深度融合和三维边界框提取”，发表于 2022 年 IEEE 国际固态电路会议（ISSCC），第 65 卷，2022 年，510–512页。–。[10] Y. Ju 和 J. Gu，“一款 65 纳米脉动式神经网络 CPU 处理器，可结合深度学习与通用计算，具备 95% 的处理单元利用率、高数据局部性和增强的端到端性能”，发表于 2022 年 IEEE 国际固态电路会议（ISSCC），第 65 卷，2022 年，第 1–3 页。[11] H. Mo、W. Zhu、W. Hu、G. Wang、Q. Li、A. Li、S. Yin、S. Wei 和 L. Liu，“9.2 一款采用有效权重卷积和基于误差补偿的预测的 28 纳米、12.1TOPS/W 双模式卷积神经网络处理器”，发表于 2021 年 IEEE 国际固态电路会议（ISSCC），第 64 卷，2021 年，第 146–148 页。[12] B. Moons、R. Uytterhoeven、W. Dehaene 和 M. Verhelst，“14.5 en-vision：一款采用 28 纳米 FDSOI 工艺、0.26 至 10tops/w、子字并行、动态电压准确率频率可扩展的卷积神经网络处理器”，发表于 2017 年 IEEE 国际固态电路会议（ISSCC）。IEEE，2017 年，第 246–247 页。[13] Y.-H. Chen、T. Krishna、J. S. Emer 和 V. Sze，“Eyeriss：一种用于深度卷积神经网络的节能型可重构加速器”，IEEE固态电路期刊，第 52 卷，138，2016 年 1，127第。–。[14] GAPuino。 [在线]。网址：https://greenwaves-technologies.com/product/gapuino/ [15] Arduino Nano BLE 33。 [在线]。网址：https://docs.arduino.cc/hardware/nano-33-ble/[16] Silicon Labs Thunderboard。 [在线]。网址：https://www.silabs.com/development-tools/thunderboard/thunderboard-sense-two-kit

[17] D. Rossi、F. Conti、M. Eggiman、A. D. Mauro、G. Tagliavini、S. Mach、M. Guermandi、A. Pullini、I. Loi、J. Chen、E. Flamand 和 L. Benini，“Vega：一款用于物联网终端节点、具备深度神经网络加速和基于磁阻式随机存取存储器的状态保持睡眠模式认知唤醒功能的十核片上系统”，IEEE固态电路期刊，第 57 卷，第 1 期，第 127–139 页，2022年。[18] V. Jain、S. Giraldo、J. D. Roose、B. Boons、L. Mei 和 M. Verhelst，“TinyVers：一款面向极端边缘端机器学习推理、具有状态保持型嵌入式磁阻式随机存取存储器的 0.8-17 TOPS/W、1.7 μW-20 mW 微型通用片上系统”，载于 2022年 IEEE超大规模集成电路技术与电路研讨会（VLSI Technology and Circuits），2022年，第 20–21 页。[19] S. Bai、J. Z. Kolter 和 V. Koltun，“用于序列建模的通用卷积网络和循环网络的实证评估”，CoRR，第 abs/1803.01271 卷，2018年。 [在线]。获取地址：http://arxiv.org/abs/1803 01271 。S. Choi、S. Seo、B. Shin、H. Byun、M. Kersner、B. Kim、D. Kim 和 S. Ha，“移动设备实时关键词检测的时域卷积”，CoRR，第 abs/1904.03814 卷，2019年。 [在线]。获取地址：http://arxiv.org/abs/1904 03814 。E. A. Ibrahim、J. Huisken、H. Fatemi 和 J. Pineda de Gyvez，“在时域卷积网络中使用时域特征进行关键词检测”，载于 2019年 第 22 届 Euromicro数字系统设计会议（DSD），2019年，第 313–319 页。[22] J. K. Chow、Z. Su、J. Wu、P. S. Tan、X. Mao 和 Y.-H. Wang，“利用卷积自编码器检测混凝土结构缺陷”，《先进工程信息学》，第 45 卷，文章编号 101105，2020年。[23] T. B. Duman、B. Bayram 和 G. Ince，“在工业流程中使用卷积自编码器进行声学异常检测”，载于《工业与环境应用软计算模型国际研讨会》。施普林格，2019年，第 432–442 页。[24] G. Coelho、P. Pereira、L. Matos、A. Ribeiro、E. C. Nunes、A. Ferreira、P. Cortez 和 A. Pilastri，“用于机器声学异常检测的深层密集型与卷积自编码器”，载于《人工智能应用与创新》，I. Maglogiannis、J. Macintyre 和 L. Iliadis 编。施普林格国际出版，2021年，第 337–348 页。[25] C. Cortes 和 V. Vapnik，“支持向量网络”，《机器学习》，第 20 卷，第 3 期，第 273–297 页，1995年。[26] L. Mei、P. Houshmand、V. Jain、S. Giraldo 和 M. Verhelst，“ZigZag：扩展深度神经网络加速器的联合架构映射设计空间探索”，IEEE计算机汇刊，第 70 卷，第 8 期，第 1160–1174 页，2021年。[27] Z. Noumir、P. Honeine 和 C. Richard，“关于简单的单类分类方法”，载于 2012年 IEEE信息论国际研讨会论文集。IEEE，2012年，第 2022–2026 页。[28] P. Perera、P. Oza 和 V. M. Patel，“单类分类：综述”，CoRR，第 abs/2101.03064 卷，2021年。 [在线]。获取地址：https://arxiv.org/abs/2101 03064 。H. Mao、S. Han、J. Pool、W. Li、X. Liu、Y. Wang 和 W. J. Dally，“探索卷积神经网络中稀疏结构的规律性”，CoRR，第 abs/1705.08922 卷，2017年。 [在线]。获取地址：http://arxiv.org/abs/1705 08922 。T. Hoefler、D. Alistarh、T. Ben-Nun、N. Dryden 和 A. Peste，“深度学习中的稀疏性：通过剪枝与增长实现神经网络的高效推理和训练”，《机器学习研究期刊》，第 22 卷，第 1 期，2021年1月。[31] D. Kadetotad、S. Yin、V. Berisha、C. Chakrabarti 和 J.-s. Seo，“一种用于设备端语音识别的 8.93 TOPS/W 长短期记忆循环神经网络加速器，具备分层粗粒度稀疏性”，IEEE固态电路期刊，第 55 卷，第 1887 页，2020年7月，第 1877 页–。[32] P. D. Schiavone、D. Rossi、A. Pullini、A. Di Mauro、F. Conti 和 L. Benini，“Quentin：采用 22纳米全耗尽型绝缘体上硅工艺的超低功耗 PULPissimo片上系统”，载于 2018年 IEEE绝缘体上硅、三维及亚阈值微电子技术联合会议（S3S），2018年，第 1–3 页。[33] F. Conti、P. D. Schiavone 和 L. Benini，“同或神经引擎：一种用于二值神经网络推理的硬件加速器知识产权核，能效为 21.6-fJ/op”，IEEE 集成电路与系统计算机辅助设计汇刊，第 37 卷，第 11 期，第 2940–2951 页，2018年。[34] J. Boukhobza、S. Rubini、R. Chen 和 Z. Shao，“新兴非易失性存储器：架构集成与研究挑战综述”，ACM Trans. Des. Autom. Electron. Syst.，第 23 卷，第 2 期，2017年11月。 [在线]。获取地址：https://doi.org/10 1145 3131848 。/ S. Yu 和 P.-Y. Chen，“新兴存储器技术：近期趋势与展望”，IEEE固态电路杂志，第 8 卷，第 2 期，第 43–56 页，2016年。[36] E. K¨ult¨ursay、M. Kandemir、A. Sivasubramaniam 和 O. Mutlu，“评估自旋转移力矩随机存取存储器作为高能效主存储器替代方案

[35]

,”

载于 2013年 IEEE系统与软件性能分析国际研讨会（ISPASS），2013年，第 256–267 页。[37] L. Mei、M. Dandekar、D. Rodopoulos、J. Constantin、P. Debacker、R. Lauwereins 和 M. Verhelst，“用于高效嵌入式深度神经网络推理的子字并行精度可扩展乘累加引擎”，载于 2019年 IEEE人工智能电路与系统国际会议（AICAS），2019年，第 6–10 页。[38] V. Camus、L. Mei、C. Enz 和 M. Verhelst，“嵌入式神经网络处理的精度可扩展乘累加单元架构评述与基准测试”，IEEE电路与系统新兴及精选主题期刊，第 9 卷，第 711 页，2019年4月，第 697 页–。[39] W. Lu、G. Yan、J. Li、S. Gong、Y. Han 和 X. Li，“FlexFlow：一种面向卷积神经网络的灵活数据流加速器架构”，载于 2017年 IEEE高性能计算机体系结构国际研讨会（HPCA），2017年，第 553–564 页。[40] H. Kwon、A. Samajdar 和 T. Krishna，“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”，ACM SIGPLAN Notices，第 53 卷，第 2 期，第 461–475 页，2018年。[41] Y.-H. Chen、T.-J. Yang、J. Emer 和 V. Sze，“Eyeriss v2：面向移动设备中新兴深度神经网络的灵活加速器”，IEEE电路与系统新兴及精选主题期刊，第 9 卷，第 2 期，第 292–308 页，2019年。[42] C. N. Coelho、A. Kuusela、S. Li、H. Zhuang、T. Aarrestad、V. Loncar、J. Ngadiuba、M. Pierini、A. A. Pol 和 S. Summers，“面向粒子探测器边缘低延迟推理的深度神经网络自动异构量化”，2020年。 [在线]。获取地址：https://arxiv.org/abs/2006 10159 。[43] Speech Commands：用于单词语音识别的公共数据集。 [在线]。获取地址：：http://download.tensorflow.org/data/ speech commands v0/[44] J. S. P. Giraldo、V. Jain 和 M. Verhelst，“嵌入式关键词检测中时序卷积网络的高效执行”，IEEE超大规模集成电路系统汇刊，第 29 卷，第 12 期，第 2220–2228 页，2021年。[45] H. Purohit、R. Tanabe、K. Ichige、T. Endo、Y. Nikaido、K. Suefusa 和 Y. Kawaguchi，“MIMII数据集：用于工业故障机器调查与检查的声音数据集”，CoRR，第 abs/1909.09347 卷，2019年。 [在线]。获取地址：http://arxiv.org/abs/1909 09347 。[46] C. R. Banbury、V. J. Reddi、P. Torelli、J. Holleman、N. Jeffries、C. Kir´aly、P. Montino、D. Kanter、S. Ahmed、D. Pau、U. Thakker、A. Torrini、P. Warden、J. Cordaro、G. D. Guglielmo、J. M. Duarte、S. Gibellini、V. Parekh、H. Tran、N. Tran、W. Niu 和 X. Xu，“MLPerf Tiny基准测试”，CoRR，第 abs/2106.07597 卷，2021年。 [在线]。获取地址：https://arxiv.org/abs/2106 07597 。[47] K. Yang、S. Kpotufe 和 N. Feamster，“一种用于物联网异常检测的高效单类支持向量机”，CoRR，第 abs/2104.11146 卷，2021年。 [在线]。获取地址：https://arxiv.org/abs/2104. 11146[48] I. Miro-Panades、B. Tain、J.-F. Christmann、D. Coriat、R. Lemaire、C. Jany、B. Martineau、F. Chaix、A. Quelen、E. Pluchart、J.-P. Noel、R. Boumchedda、A. Makosiej、M. Montoya、S. Bacles-Min、D. Briand、J.-M. Philippe、A. Valentian、F. Heitzmann、E. Beigne 和 F. Cler-midy，“SamurAI：一款自适应通用物联网节点，具备 1.7MOPS-36GOPS 性能、15,000× 峰值至空闲功耗降低、207ns 唤醒时间和 1.3TOPS/W 机器学习能效”，载于 2020年 IEEE超大规模集成电路研讨会，2020年，第 1–2 页。[49] E. Flamand、D. Rossi、F. Conti、I. Loi、A. Pullini、F. Rotenberg 和 L. Benini，“GAP-8：用于物联网边缘人工智能的 RISC-V 片上系统”，载于 2018年 IEEE第 29 届专用系统、架构与处理器国际会议（ASAP），2018年，第 1–4 页。[50] W. Shan、M. Yang、T. Wang、Y. Lu、H. Cai、L. Zhu、J. Xu、C. Wu、L. Shi 和 J. Yang，“一种采用串行 FFT 梅尔频率倒谱系数和二值化深度可分离卷积神经网络的 510-nW 唤醒式关键词检测芯片，基于28纳米CMOS 工艺”，IEEE固态电路期刊，第 56 卷，第 1 期，第 151–164 页，2021年。[51] J. Lee、C. Kim、S. Kang、D. Shin、S. Kim 和 H.-J. Yoo，“UNPU：一种具备完全可变权重位精度的高能效深度神经网络加速器”，IEEE固态电路期刊，第 54 卷，第 1 期，第 173–185 页，2019年。[52] M. Fariselli、M. Rusci、J. Cambonie 和 E. Flamand，“用于多核微控制器超低功耗音频神经网络处理的纯整数近似梅尔频率倒谱系数”，载于 2021年 IEEE第 3 届人工智能电路与系统国际会议（AICAS），2021年，第 1–4 页。