A一种面向 Transformer 的 818 TOPS/W、CSNR-31 dB、SQNR-45 dB 10 位电容器重构型存内计算宏，采用软硬件模拟协同设计Kentaro Yoshioka

庆应义塾大学，日本横滨，电子邮箱：kyoshioka47@keio.jp

摘要

Transformer 推理需要很高的计算精度；由于模拟存内计算固有的计算误差，要实现这一目标一直颇具挑战。为克服这一挑战，我们提出电容器重构型存内计算（CR-CIM），通过采用高面积/功率效率的 10 位模数转换器，实现高计算精度的模拟存内计算。CR-CIM 重构其电容器阵列，使其兼用于计算和模数转换，从而显著节省面积。此外，CR-CIM 在操作期间保持信号电荷静止，消除了信号衰减，使比较器能效提升 4 倍。我们还提出一种软硬件模拟协同设计技术，将多数投票集成到 10 位模数转换器中，根据当前运行层动态优化存内计算噪声性能，进一步降低推理功耗。

我们的 CR-CIM 实现了模拟存内计算中最高的计算精度，818 TOPS/W 的能效也具有先进水平的竞争力。此外，综合考虑信号量化噪声比和 CSNR 的性能指标分别比以往工作高出 2.3 倍和 1.5 倍。我们实现了视觉 Transformer（ViT）推理，并在模拟存内计算中实现了最高的 CIFAR10 准确率 95.8%。

引言

存内计算技术有望通过最大限度减少数据传输，大幅降低深度神经网络推理的能耗。尽管许多研究致力于提升模拟存内计算的面积和功率效率 [1-6],，但鲜有研究关注提高存内计算的有效计算精度；存内计算的模拟噪声和量化噪声往往未得到充分考虑。尽管卷积神经网络等相对简单的网络在计算精度较低时仍能达到可接受的整体精度，但要执行更复杂的

如图 1（A）所示，Transformer 等新一代自然语言处理、计算机视觉和音频识别算法应用。因此，模拟存内计算需要更精确的计算方案。

然而，由于固有的模拟计算误差，模拟存内计算要实现高 CSNR 极具挑战性。为了有效执行 Transformer 推理，模拟存内计算必须具备较高的 1）线性度、2）ADC 分辨率和 3）电路抗噪能力，同时将电路开销降至最低。当前的[2] 基于电流和基于时间的存内计算[3] 容易发生失配，难以达到>8位线性度。基于电荷的存内计算 [4, 5] 凭借电容器的高匹配特性，有望满足线性度要求，但由于面积和功耗开销巨大，难以扩展至 Transformer 所需的 10 位模数转换器分辨率，如图 1（B）所示。鉴于这些挑战，据作者所知，迄今尚未有使用模拟存内计算加速 Transformer 的相关报道。尽管数字存内计算不存在模拟误差，但只有采用 5纳米等先进制程节点[6]，其能效才有望与模拟存内计算相竞争。为解决上述问题，我们提出一种新颖的电容重构型模拟存内计算架构，可在基于 CMOS 的边缘端和物联网应用中实现高面积/功耗效率与高 CSNR

ns. CR-CIM 概念

图 2 展示了提出的 CR-CIM 电路，该电路利用可重构电容阵列实现双重用途：1）基于电荷的计算；2）用作 10 位模数转换器操作中的 10位电容式数模转换器。

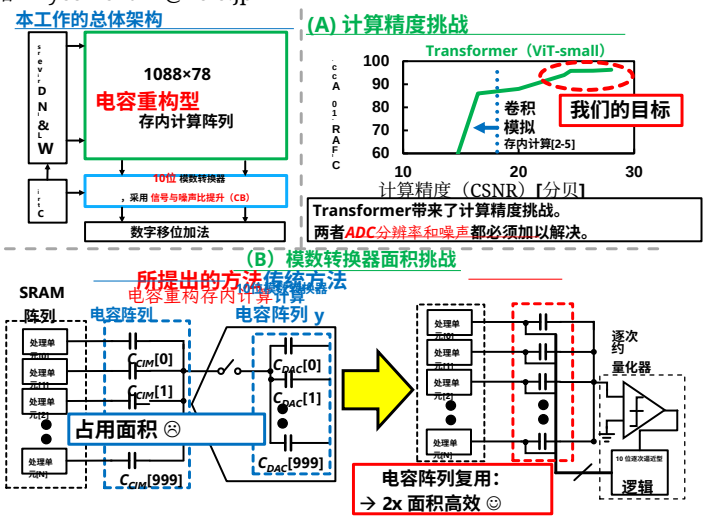

图 1. 高精度存内计算面临的挑战。

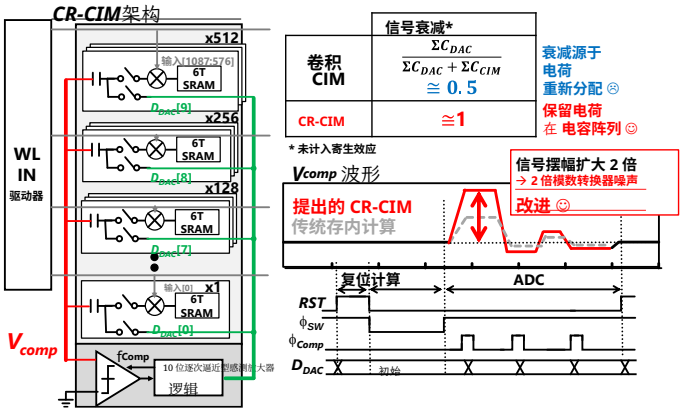

图 2. 提出的 CR-CIM 架构。

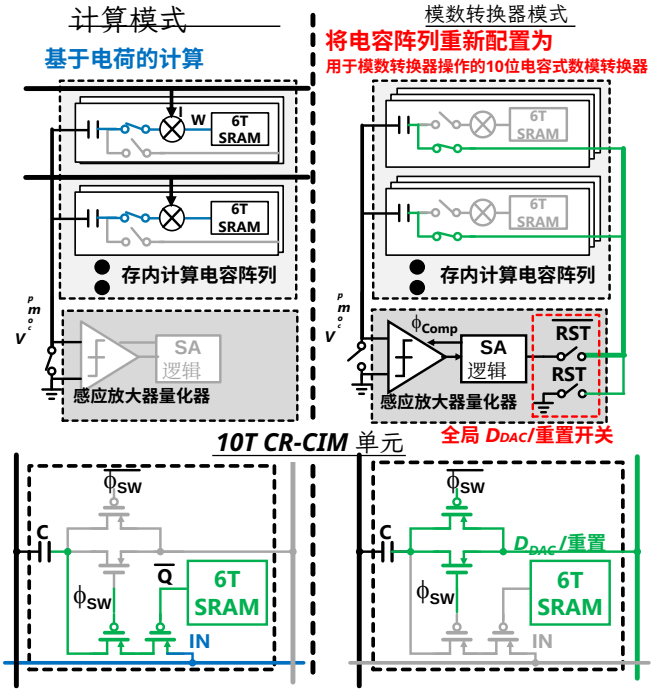

图 3. CR-CIM 操作与单元原理图。

这实现了高分辨率模数转换器的模拟存内计算，同时保持极小的面积。在计算阶段，输入（IN）与存储在 6T静态随机存取存储器中的权重的乘积被送入存内计算单元电容。在模数转换器阶段，存内计算电容连接到数模转换器反馈信号（D_{DAC}[9:0]），以便将其重构为电容式数模转换器阵列。信号 D_{DAC}[9] 连接到 512 个存内计算单元，信号 D_{DAC}[8] 连接到 256 个存内计算单元，以此类推，形成二进制电容式数模转换器阵列。 在传统存内计算中，计算过程中获得的电荷通过电荷再分配传输至模数转换器，导致显著的信号衰减。CR-CIM 通过在整个操作过程中将信号电荷保留在电容器阵列中，解决了这一问题。CR-CIM 电容器阵列的顶板直接连接到比较器，从而可通过 D_{DAC}[9:0] 反馈实现逐次逼近（SA），完成 10 位模数转换。与传统存内计算相比，CR-CIM 的信号摆幅提高了 2 倍，相当于模数转换器噪声改善了 2 倍。 高分辨率逐次逼近型模数转换器的功耗通常主要来自低噪声比较器，而 CR-CIM 可通过放宽比较器噪声要求显著降低功耗。

存内计算既需要计算机制，也需要复位机制，但 CR-CIM 还需要传输 D_{DAC} 信号，这会带来额外开销。为尽量减少这些开销，我们提出了一种 D_{DAC}/复位共享结构（图3）。D_{DAC} 通路也用于复位电荷，从而无需在单元内设置复位开关，并使 CR-CIM 能够通过小型 10T 单元实现。共享的 D_{DAC}/复位[9:0] 节点在复位期间连接至复位电压，在模数转换器阶段则通过全局开关切换至 D_{DAC}[9:0] 。存内计算单元电容采用定制的 1.5fF 边缘电容实现，并经后布局仿真确认具有 10 位线性度。按照 65纳米逻辑工艺规则，CR-CIM 单元面积为 2.3um^2，约为典型 6T静态随机存取存储器单元的 2 倍。

图 4 展示了我们通过软硬件模拟协同设计（SAC）提升 Transformer 推理效率的策略。Transformer 可分为两个模块：用于混合图像块特征向量的注意力模块，以及逐向量多层感知器模块。我们利用了以下观察结果：注意力层所需的 CSNR 比多层感知器层低 10 分贝。我们提出了一种信号与噪声比提升（CB）技术，可在读出精度与功率之间进行权衡。根据当前运行的 Transformer 模块自适应地启用 CB，可将推理效率提升至 2.1 倍。启用 CB 时，最后 3 次 SA 逐次逼近比较采用 6路多数表决（MV），使 CSNR 提高 5.5 分贝，但相应地带来 1.9 倍的功率开销和 2.5 倍的转换时间开销。尽管多数投票是逐次逼近型模数转换器常用的降噪技术，但这是首个根据当前运行的深度神经网络层自适应使用多数投票的 SAC。此外，虽然在阵列间距较窄的情况下难以实现低噪声比较器，CB 降低了噪声要求，同时提升了布局可行性。

测量结果

CR-CIM 采用 65纳米 CMOS 工艺的 1088x78 阵列进行了原型验证。图 5 总结了 CR-CIM 测得的列特性。尽管 10 位读出颇具挑战性，测量结果仍显示出良好的 CIM传输特性，INL 误差在 <2 个最低有效位以内。采用 CB 技术时，所有码值的平均噪声为 0.58 LSB；禁用 CB 时，噪声增加 2 倍。考虑电路噪声的信号量化噪声比 [4] 和 CSNR [1]，分别为 45 分贝和 31 分贝。这些指标分别比先前报道的基于电荷的存内计算 [4,5]，高 23 分贝和 14 分贝，表明 CR-CIM 和 CB 技术可实现较高的计算精度。

图 6 总结了 CR-CIM 的性能。为展示对 Transformer 的加速效果，使用具有 12 个

堆叠的 Transformer 层。存内计算负责计算线性层，其中多层感知器层采用 CB，位精度为 6 位，而注意力层则不采用 CB，输入和权重的位精度分别为 4 位。因此，Top-1 准确率达到 95.8%，与理想推理的 96.8% 相当。仿真分析表明，采用所提出的 SAC 和位精度优化后，Transformer 推理效率提升了 2.1 倍。与综合考虑信号量化噪声比和 CSNR 的 SQNR 评价指标及 CSNR 评价指标相比，本工作分别比以往的模拟存内计算高出 2.3 倍和 1.5 倍。

致谢 本研究得到 JST CREST 计划 JPMJCR21D2 和双叶基金会的支持。

参考文献

[1] S. Gonugondla 等人，“存内架构精度的基本极限”，国际计算机辅助设计会议，2020 年。[2] Q. Dong 等人，“用于机器学习应用的 7纳米 FinFET CMOS 中每瓦 351TOPS、372.4GOPS 存内计算 SRAM 宏”， 国际固态电路会议，2020 年。[3] P. Wu 等人，“用于边缘人工智能设备的 28纳米 1Mb 时域存内计算 6T-SRAM 宏，支持 8 位乘加运算，延迟为 6.6ns，性能达 1241GOPS 和 37.01TOPS/W”，国际固态电路会议，2022 年。[4] H. Jia 等人，“基于位可扩展存内计算的可编程异构微处理器”，IEEE 固态电路期刊，55(9)，2609-2621，2020 年。[5] J. Lee 等人，“采用异构计算架构与指数存内计算的 13.7 TFLOPS/W 浮点深度神经网络处理器”，VLSI 技术研讨会，2021 年。[6] H. Fujiwara 等人，“支持宽范围动态电压频率调节的 5 纳米 254-TOPS/W、221-TOPS/mm2 全数字存内计算宏”，国际固态电路会议，2022 年。

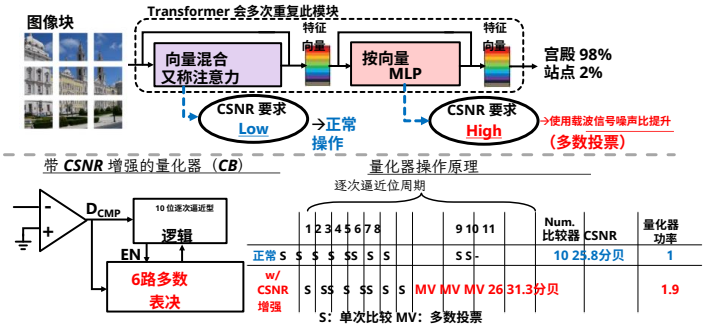

图4 Transformer 软件-模拟协同设计（SAC）。

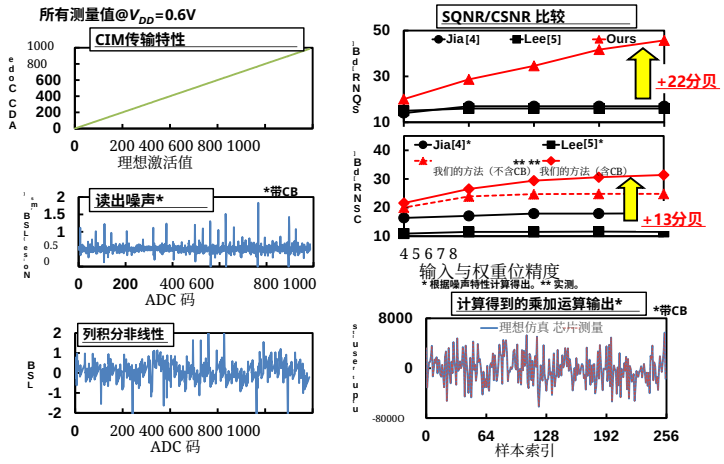

图5 实测 CR-CIM 列特性。

	本工作[4] IEEE固态电路期刊 2020 [5] 超大规模集成电路 2021			[2] 国际固态电路会议 2020
存内计算类型	电荷	电荷	电荷	电流
工艺	65nm	65nm	28nm	7nm
电源电压[V]	0.6-1.1	0.85-1.2	0.6-0.9	0.8
阵列大小	10KB	72KB	36KB	0.5KB
位精度 (激活值/权重)	6b/6b (可配置)	8b/8b	5b/1b (可配置)	4b/4b
峰值 TOPS 归一化至 1 位	1.2	2.1	6.1	5.9
峰值 TOPS/mm^2 归一化至 1 位	2.5	0.6	12	112
峰值 TOPS/W 归一化至 1 位	818	400	5796	5616
ADC 位数	10	8	8	4
峰值信噪失真比 [分贝]	45.3	22	17.5	21
峰值电路信噪比 [分贝]	31.3	17	10.5	N.A.
SQNR 性能指标*	118841 (2.3x)	4113	33512	51466
CSNR 性能指标**	24541 (1.5x)	2449	15855	N.A.
支持 Transformer?	Yes	No	No	No
CIFAR-10 准确率 y	95.8	92.4	91.1	MNIST

* SQNR 性能指标=每瓦万亿次运算*2^信号量化噪声比位数 ** CSNR 性能指标=每瓦万亿次运算*2^信号相关噪声比位数
*** 信号量化噪声比位数 = (信号量化噪声比[分贝]-1.76)*6.02 信号相关噪声比位数 = (CSNR[分贝]-1.76)*6.02

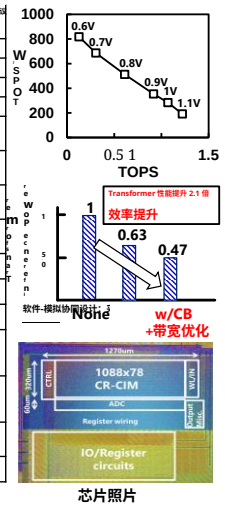

图6 性能总 ary.