ISSCC2022：A 28nm 15.59μJ/Token Full-Digital Bitline-Transpose CIM-Based Sparse Transformer Accelerator with Pipeline/Parallel Reconfigurable Modes

这篇文章介绍了一款名为TranCIM的计算内存（CIM）加速器，专为Transformer模型设计。Transformer模型在自然语言处理和计算机视觉等领域取得了显著的成果，但其大量的矩阵乘法导致了高延迟和能耗。TranCIM加速器具有以下特点：

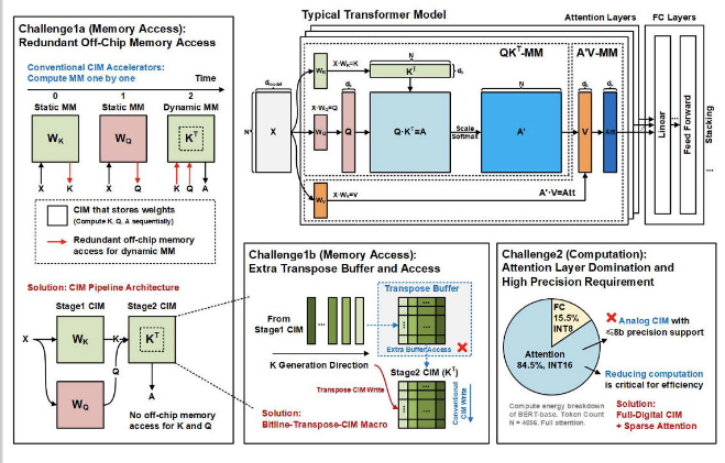
1.通过可重构流网络（RSN）连接静态引擎（SEngine）和动态引擎（DEngine），为Transformer的不同层提供专用模式。对于注意力层，TranCIM配置为管道模式，将第一阶段引擎的输出直接流式传输到第二阶段引擎。对于全连接层，TranCIM配置为并行模式，所有引擎独立工作。

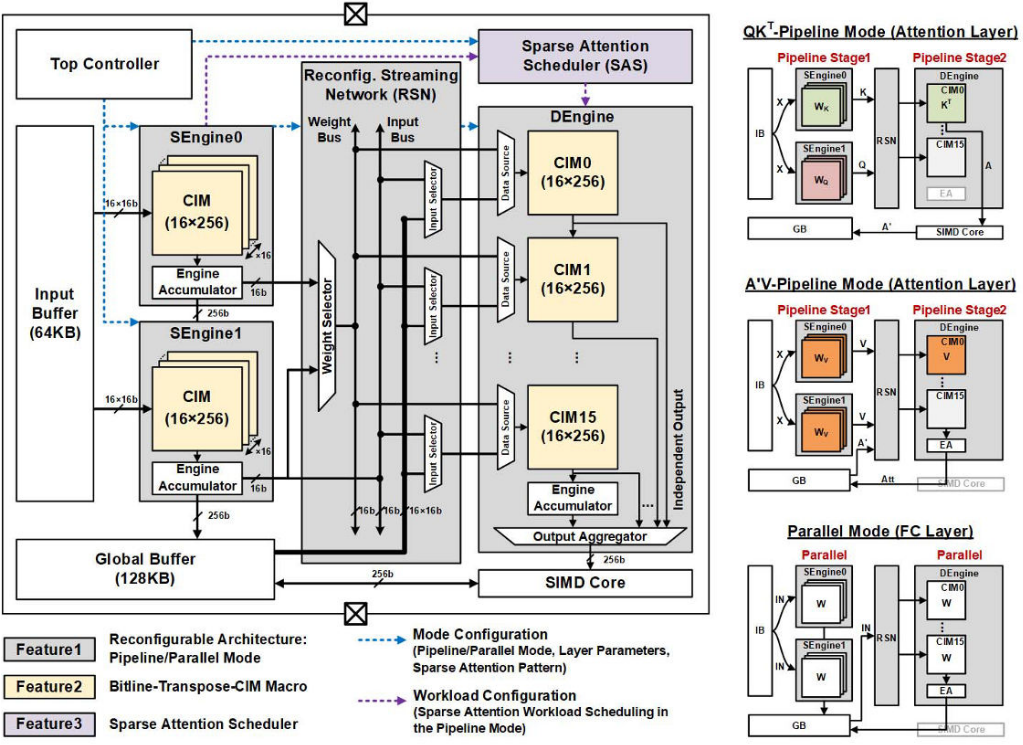
2.与之前的字线方向馈送解决方案不同，TranCIM的CIM宏采用比特线转置结构，使输入馈送和权重写入的方向对齐。这样可以在QKT-pipeline模式下实现无需额外存储和缓冲区访问的K转置。全数字CIM设计支持注意力层的INT16和全连接层的INT8，避免模拟CIM逻辑导致的精度损失。

3.基于最近的稀疏Transformer算法，注意力层可以训练具有块稀疏模式，以减少计算量，同时保持高准确性。设计了一个稀疏注意力调度器（SAS），以动态配置CIM工作负载，以适应不同的稀疏注意力模式，将计算复杂度从O(N^2)降低到O(N)，其中N为Token计数。

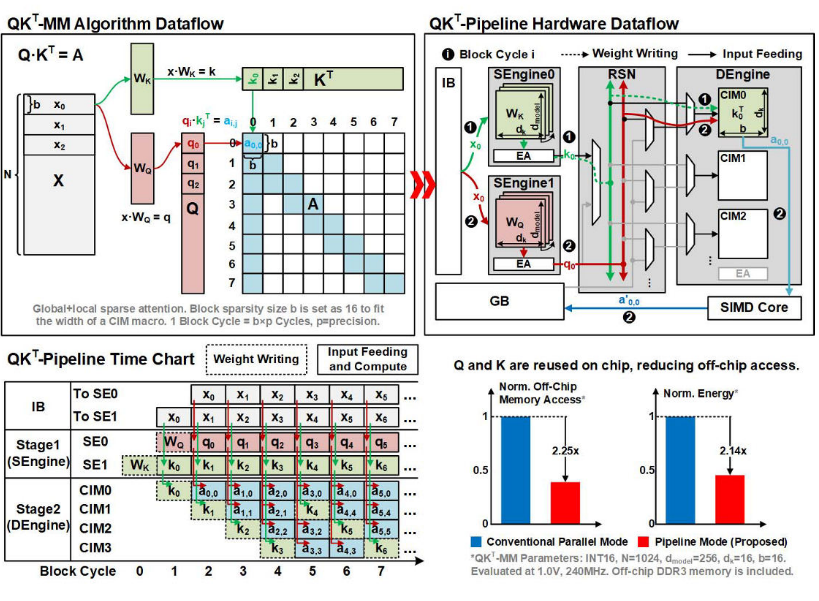
实验结果表明，TranCIM在28nm CMOS工艺下实现了高性能和低能耗。与现有的CIM加速器相比，TranCIM在BERT-base模型的注意力层上实现了5.22倍的速度提升和7.99倍的能耗降低。

文章首先提出挑战：1.attention层的权值和激活都是动态变化的，导致重复的片上数据搬移。一个CIM流水线体系结构能减轻这个问题，但是K必须经过转置存入下一级CIM，就需要一个大的buffer来进行转置处理。2.attention层需要的精度大于8bit，因此先前的模拟CIM（精度<=8b）并不能直接被利用。

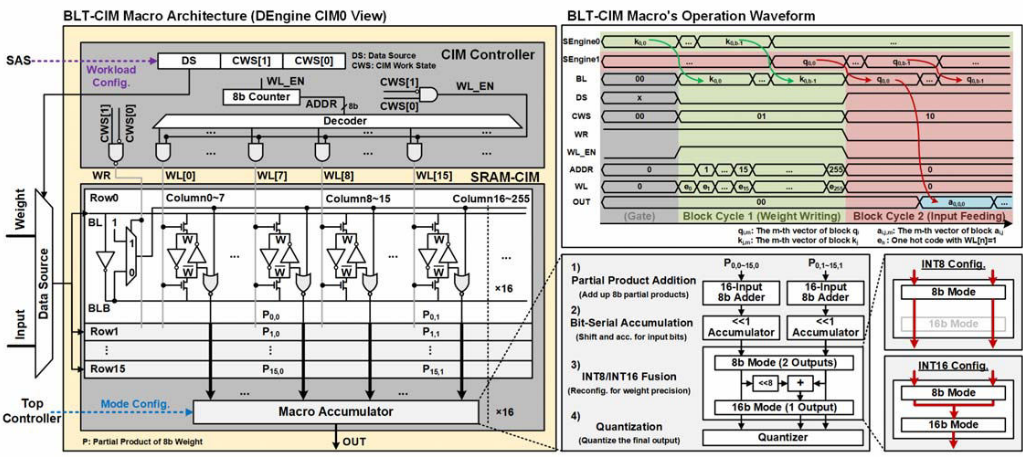
  
本文提出TranCIM：1）可重构流网络，对于attention层，它采用管道模式，直接将一级引擎接到二级引擎上，对于FC层，所有引擎并行工作。2）CIM宏设计有位线转置结构，直接将转置后的K存入CUN内部，不需要额外的存储和缓冲访问。3）基于最近的稀疏transformer算法，作者使用块稀疏模式来训练注意力层，减少计算提高精度。文章设计了一个个稀疏attention调度器来配置CIM工作负载，复杂度从O(N2)降低到O(N)。N是token个数。



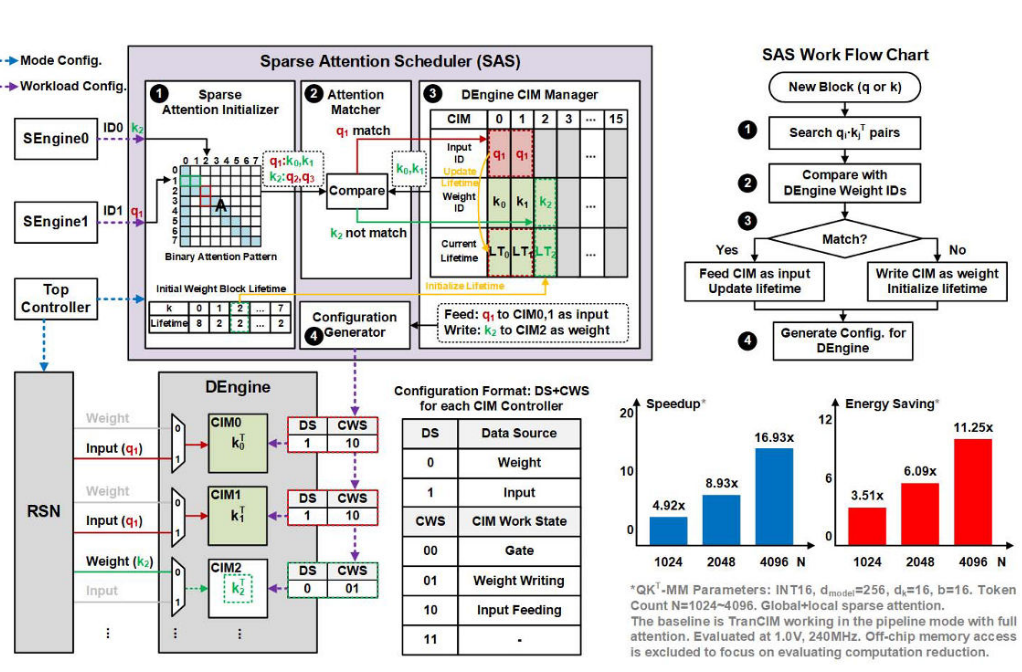
体系结构。



作者将块稀疏大小设置为b，则输入序列每b个token都被一起处理，用来计算A的bxb块。RSN激活用来构建两级流水线，SE0放WK，SE1放WQ。Wk的计算结果K存入DEngine中。



DEngine示意图，电路采用6T SRAM单元和4T NOR门进行位加权存储和乘法运算。阵列的每16列具有4个阶段的宏累加器：1）部分积加法，2）位串行累加，3）INT 8/INT 16融合和4）量化。



SAS稀疏注意调度模块：在模式配置期间，顶部控制器将二进制注意模式和权重块生命周期写入稀疏注意初始化器。DCengine CIM管理器维护CIM工作负载信息：输入ID、权重ID和当前生存期。在这个例子里，q1负责k0和k1，因此作为CIM0和1的输入，k2没有匹配的块，因此它被设置为CIM 2的新权重。如果有新的输入，CIM的weight寿命会缩短。

ISSCC2023：A 28nm 53.8TOPS/W 8b Sparse Transformer Accelerator with In-Memory Butterfly Zero Skipper for Unstructured-Pruned NN and CIM-Based Local-Attention-Reusable Engine

这篇文章介绍了一款针对Transformer网络的高效加速器，专为处理稀疏性优化和局部注意力而设计。该加速器在28nm CMOS工艺下实现了53.83TOPS/W的能量效率和0.85TOPS/mm²的面积效率。主要贡献包括：

1.基于蝴蝶网络的稀疏感知前馈计算架构，实现宏内部的不规则分布式零权重跳过。

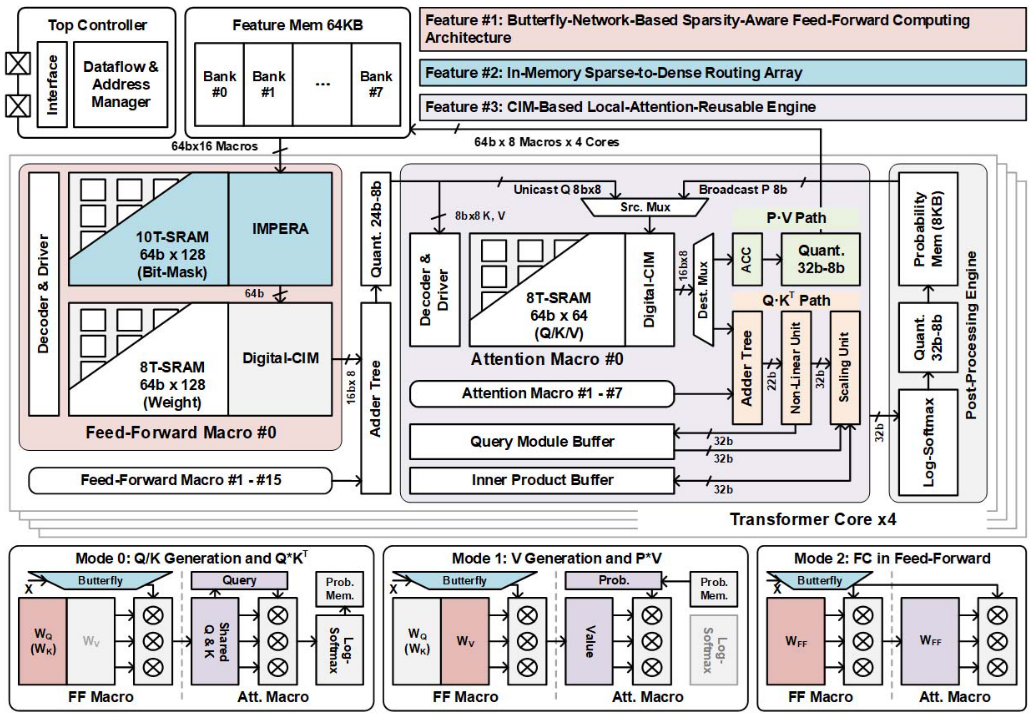
2.一种32-8内存稀疏到密集路由阵列（IMPERA），根据稀疏位掩码提取待计算的输入。

3.一种数字CIM（计算内存）基于的局部注意力可重用引擎，支持可变的稀疏注意力跨度和QK共享。

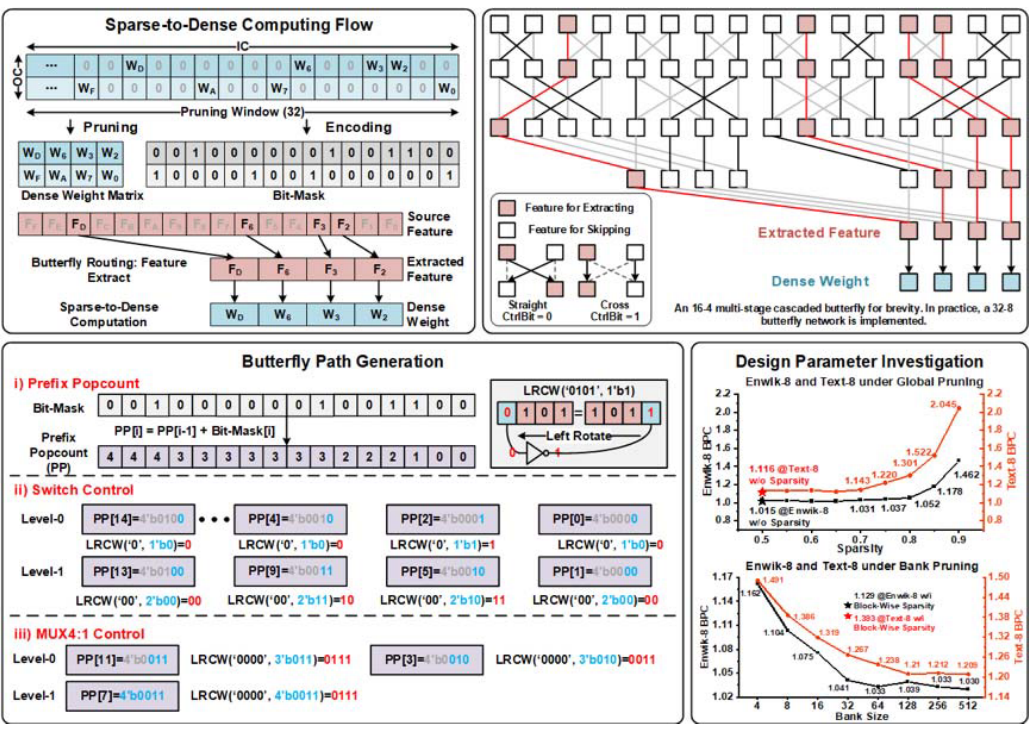
文章详细介绍了蝴蝶网络的实现、IMPERA电路设计以及基于CIM的局部注意力可重用引擎。实验结果表明，该加速器在保持准确性的同时，实现了显著的能量效率和面积效率提升。与现有的稀疏性优化技术相比，该加速器在Transformer模型的计算和存储方面表现出了显著的优势。

首先提出挑战：由于网络的扩展以及能耗增加，块稀疏体系结构被提出。但是这种结构化稀疏是以显著的准确性损失为代价的。因此，基于CIM的加速器面临的挑战是如何处理非结构化修剪NN，同时保持高效率。这些非结构化模式可以表示为：1）矩阵内不规则分布的零权重，以及2）不同attention头对应不同局部attention spans。

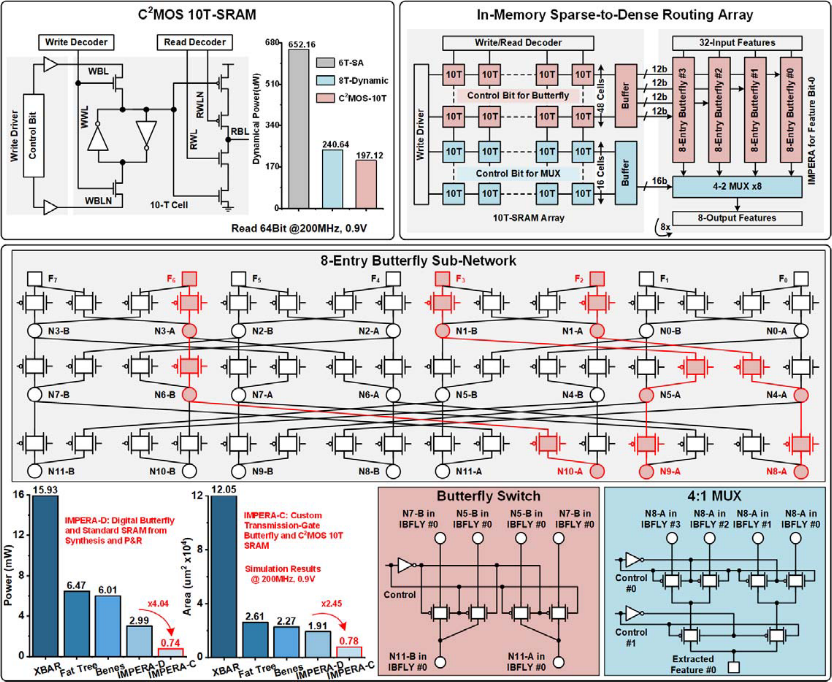
该工作提出一个高效的Transformer加速器，具有自定义内存路由和计算宏，用于上述非结构化稀疏性。主要贡献是：1）基于butterfly网络的稀疏感知前馈计算架构，支持宏中不规则分布的零权重跳过，2）32-8内存中稀疏到密集路由阵列（IMPERA），根据稀疏位掩码提取要计算的输入，以及3）基于数字CIM的本地注意力可重用引擎，促进可变稀疏注意力跨度和QK共享。



整体架构：4个transformer内核，一个控制器和一个64kb输入缓冲。每个内核包括16个前馈宏（FFM）、8个数字attention宏（AM）和用于计算基于log-softmax的概率的后处理引擎（PPE）。这些模块可以配置成三种模式：在第一种模式中，FFM顺序计算Q和K，AM执行Q·KT，PPE将AM的结果转换为流水线中的概率。在第二种模式中，FFM计算V，AM获得P·V的自注意结果。最后一种模式负责最终的全连接层，其中FFM和AM都被设置为完成矩阵乘法。



基于蝶形网络的稀疏感知前馈跳零方案：对于给定稀疏权重向量，该方案首先将其压缩成密集向量和二进制编码的位掩码。非跳过的输入特征需要根据位掩码从原始输入中提取和路由。通过将两个稠密向量相乘来获得乘积。实验结果表明，在两个常见的NLP数据集（Enwik-8和Text-8）中，网络的准确性直到75%的权重被裁剪都会保持不变。同样，与全局剪枝（N=512）相比，（N=32）在上述任务中的精度损失并不大，并且明显优于N=16的情况。所以作者采用N=32，S=0.25作为该蝶形网络的参数。

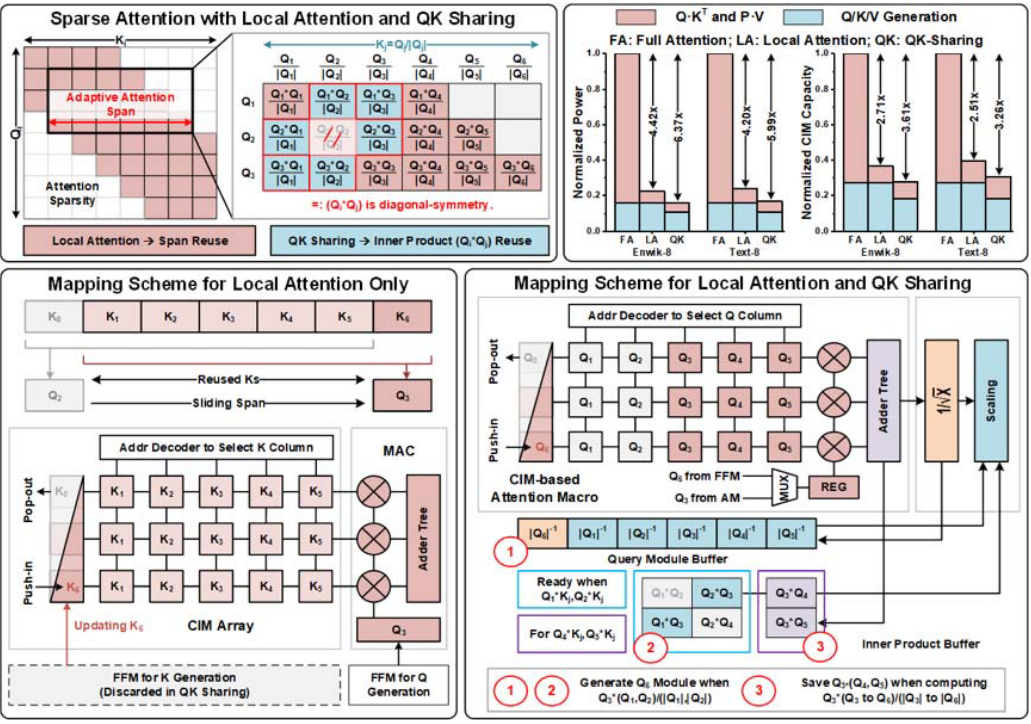


IMPERA电路实现：

这里作者设计了一个10T的SRAM，用来输出Features的路径选择信号。而后通过传输门来实现蝴蝶网络以及4x1多路选择器。

最后是AM设计以及attention数据重用：

这一块没太看懂



ISSCC2023：MulTCIM: A 28nm 2.24μJ/Token Attention-Token-Bit Hybrid Sparse Digital CIM-Based Accelerator for Multimodal Transformers

这篇文章介绍了一款名为MulTCIM的数字计算内存（CIM）加速器，专为多模态Transformer（MulT）模型设计。多模态Transformer模型在处理多种模态（如视觉、自然语言、语音等）的人工智能任务方面取得了优异的成果。文章提出了三种针对混合稀疏性（注意力稀疏、标记稀疏和位稀疏）的优化方法：

1.长复用消除调度器（LRES）：通过动态调整注意力矩阵的稀疏模式，提高计算内存（CIM）的利用率。

2.运行时标记修剪器（RTP）和模态自适应CIM网络（MACN）：RTP用于去除不重要的标记，MACN则负责将CIM核心动态划分为两个流水线阶段，分别用于静态矩阵乘法（MM）和动态MM的注意力计算。此外，MACN还利用模态对称性降低跨模态切换的延迟。

3．有效位宽平衡CIM（EBB-CIM）架构：通过检测和平衡输入位宽，减少计算时间。

实验结果表明，MulTCIM在28nm工艺下实现了高效的性能。在两个典型的多模态Transformer模型（ViLBERT-base和ViLBERT-large）上，混合稀疏技术在注意力层实现了9.47倍的速度提升和8.11倍的能量节省。与现有的Transformer和CIM加速器相比，MulTCIM在ViLBERT-base模型上实现了2.24μJ/Token的能耗，分别比先前的技术提高了5.91倍和5.61倍的能效。

首先提出挑战：文章切入点为多模态Transformer模型。在利用attention-token-bit混合稀疏性时，存在三个挑战：1）对于注意力稀疏性，不规则的模式导致长的重用距离，这要求CIM保持不经常使用的权重，降低CIM利用率。2）虽然token稀疏性减少了计算，但MulT的跨模态注意力处理来自两种模态的token，具有不同的令牌长度（N）和嵌入维度（dm），导致跨模态切换的高延迟。3)在位级别上，由于token稀疏性降低了值的局部性，因此CIM宏对于同一组输入的有效位宽具有更大的差异。

文章提出了一个基于数字CIM的MulT模型加速器MulTCIM，它具有三个特征来解决混合稀疏性挑战：1）实现了一个长重用消除器（LRES），而不是按顺序生成Q，K，V令牌，以动态地重塑attention矩阵，就像全局+局部稀疏模式一样。通过这种方法，可以更频繁地重用CIM中存储的权重，以提高CIM利用率。2)设计了一个RTP令牌修剪器（RTP）来删除无关紧要的token和一个模态自适应CIM网络（MACN）动态地将所有CIM核心分为两个流水线阶段，StageS用于Q，K和V生成中的静态矩阵乘法（MM），StageD用于注意力计算的动态MM。在跨模态切换中，MACN进一步利用模态对称性来重叠Q、K生成以降低延迟。3)有效位宽平衡CIM（EBB-CIM）架构被设计为通过执行有效的位宽检测和位均衡来平衡跨存储器MAC的输入位，从而减少计算时间。