基于存算一体集成芯片的大语言模型专用硬件架构

Large Language Model Specific Hardware Architecture Based on Integrated Compute-in-Memory Chips

何斯琪/HE Siqi，穆琛/MU Chen，陈迟晓/CHEN Chixiao
（复旦大学，中国 上海 200433）
(Fudan University, Shanghai 200433, China)

DOI：10.12142/ZTETJ.202402006
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.tn.20240407.1932.006.html
网络出版日期：2024-04-09
收稿日期：2024-02-25

摘要：目前以ChatGPT为代表的人工智能（AI）大模型在参数规模和系统算力需求上呈现指数级的增长趋势。深入研究了大型模型专用硬件架构，详细分析了大模型在部署过程中面临的带宽问题，以及这些问题对当前数据中心的重大影响。提出采用存算一体集成芯片架构的解决方案，旨在缓解数据传输压力，同时提高大模型推理的能量效率。此外，还深入研究了在存算一体架构下轻量化-存内压缩协同设计的可能性，以实现稀疏网络在存算一体硬件上的稠密映射，从而显著提高存储密度和计算能效。

关键词：大语言模型；存算一体；集成芯粒；存内压缩

Abstract: Artificial intelligent (AI) models represented by ChatGPT are showing an exponential growth trend in parameter size and system computing power requirements. The dedicated hardware architecture for large models is studied, and a detailed analysis of the bandwidth bottleneck issues faced by large models during deployment is provided, as well as the significant impact of this challenge on current data centers. To address this issue, a solution of using integrated compute-in-memory chiplets has been proposed, aiming to alleviate data transmission pressure and improve the energy efficiency of large-scale model inference. In addition, the possibility of lightweight in-memory compression collaborative design under the in-memory computing architecture is studied, in order to achieve dense mapping of sparse networks on the integrated in-memory computing architecture hardware, thereby significantly improving storage density and computational energy efficiency.

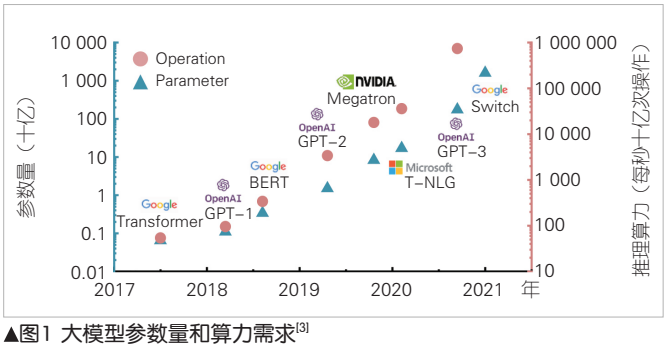
Keywords: large language model; compute-in-memory; chiplet; in-memory compression

引用格式：何斯琪, 穆琛, 陈迟晓. 基于存算一体集成芯片的大语言模型专用硬件架构 [J]. 中兴通讯技术, 2024, 30(2): 37-42. DOI: 10.12142/ZTETJ.202402006

Citation: HE S Q, MU C, CHEN C X. Large language model specific hardware architecture based on integrated compute-in-memory chips [J]. ZTE technology journal, 2024, 30(2): 37-42. DOI: 10.12142/ZTETJ.202402006

近年来，基于注意力机制的大语言模型（LLM）[1]取得了显著成功。与此同时，模型尺寸在迅速增长，如图1所示，每两年模型尺寸增长240倍，而相应的算力需求则增长近750倍。与此同时，硬件每两年3.1倍的发展速度[2]已逼近物理极限，进入了技术发展的瓶颈期。传统的超大规模和超大面积的单芯片系统级芯片（SoC）方案面临着利用率低、良率低、验证复杂度高、设计成本激增等一系列问题，同时集成电路制造已经达到了光刻掩膜版的最大面积上限。因此，大型模型的推理变得异常复杂且成本高昂，这成为当前研究和实际应用中需要解决的关键问题。

为了突破存储单元和计算单元之间的数据搬移的瓶颈，提高计算芯片能效，存算一体的专用芯片架构逐渐成为了神



▲图1 大模型参数量和算力需求[3]

基金项目：国家自然科学基金项目（62322404）；复旦大学-中兴通讯强计算架构研究联合实验室“存算一体架构研究项目”

经网络计算芯片研究和大模型实际部署的重要前进方向。通过电路与架构的协同创新，存算一体架构试图打破存储器和计算器之间的壁垒，实现数据搬移效率的提高或数据搬移次数的减少，从而提高芯片的计算能效。

然而，目前已有的神经网络计算芯片可扩展性欠佳，无法完全适应大模型的推理需求。在上述背景下，处理器领域的巨头已经将目光投向了集成芯粒（Chiplet）这一新兴技术。集成芯粒技术最早由加利福尼亚大学圣塔芭芭拉分校（UCSB）大学的谢源教授于2017年国际计算机辅助设计会议（ICCAD）上提出[4]。与单芯片SoC方案不同，集成芯粒方案先将多个小颗粒芯片独立设计并实现，然后通过先进封装技术重新组装，从而完成系统上的功能集成。美国Intel公司、AMD公司、英伟达公司的服务器/数据中心芯片都已开始广泛采用集成芯粒方案[5-7]。这些方案将高性能计算核心设计为模块化芯片，通过2.5D/3D封装技术、高速片间互联技术和有源基板技术将计算核心芯片模块集成。在不明显增加设计复杂度的前提下，保证芯片的良率，延续了后摩尔时代芯片算力提升。这一趋势为硬件设计提供了更为灵活和高效的解决方案，以适应不断增长的大型模型算力需求。

1 大模型对数据中心的挑战

1.1 集成芯片技术

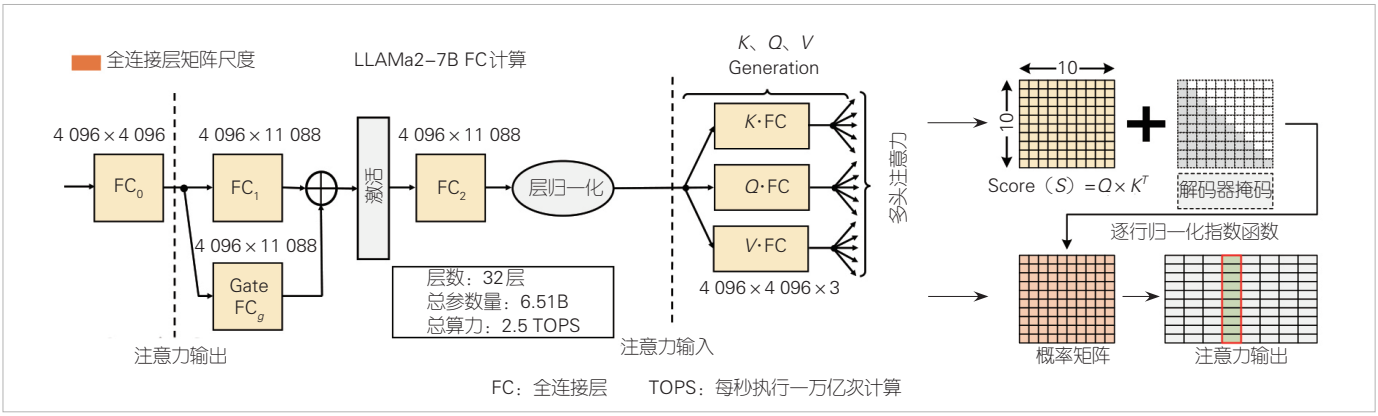
以ChatGPT为代表的人工智能（AI）大模型在参数规模和系统算力需求上呈现出指数级的增长趋势。当前，能够支持大型模型的数据中心和超级计算机普遍采用以xPU+主机内存缓冲器（HBM）集成芯片为核心的高性能处理器芯片系统。如图2所示，这些大算力芯片具备PFLOPS级算力和100 GB级存储性能，例如Nvidia H100图形处理器（GPU）拥有2 PFLOPS（每秒执行1 000万亿次浮点运算）的算力，AMD Instinct MI300拥有383 TFLOPS（每秒执行1万亿次浮点运算）的算力，华为昇腾910 B则具备256 TFLOPS算力等。传统的超大规模和超大面积的单芯片SoC方案已经面临着诸多问题，包括利用率低、良率低、验证复杂度高以及设计成本激增等。同时，集成电路制造已经达到了光刻掩膜版的最大面积上限，而30.48 cm（12英寸）晶圆的掩膜也在光刻机的要求下存在上限，最大芯片设计面积为858 mm^2。在这样的背景下，单芯片SoC的算力进一步扩充空间受到限制，潜在的良率问题和面积限制使得算力的提升变得更加困难。同时，自2023年起，美国进一步加强了针对中国芯片产业的出口限制，对总处理性能和算力密度超过超过规定的芯片实施了更加严格的管制。

▲图2 超算中心总算力和集成芯片数

为了缩小智能计算和处理器芯片技术上的差距，采用微纳架构工艺将多个芯片（粒）集成已经成为克服单芯片制造最大面积极限和芯片电路规模瓶颈的重要手段。不同于单芯片方案，集成芯片方案通过使用先进封装技术将多个小颗粒芯片组件组装在一起，实现了系统上的功能集成。这种方法将大型昂贵的SoC分解为体积更小、良率更高且更具成本效益的单芯片，同时也有助于缩短设计周期，降低成本。集成芯片技术已成为高性能处理器不可或缺的组成部分，而它正朝着3D多层堆叠、更多种类的芯片以及更大规模集成的方向发展。这一发展趋势的目标是进一步满足大型模型对硬件性能的不断增长的需求，适应日益增加的计算和处理任务。

1.2 大模型部署的带宽瓶颈

以图3中展示的拥有70亿参数的大型模型（LLAMa2-7B）为例，该大型模型的每一层多头注意力都包括多个连续前馈（FCL）计算。与此相关的单层参数量达到2.03亿，而32层的参数总量达到65亿，占用整体系数和计算的85%以上，远超过单一互补金属氧化物半导体（CMOS）芯片的片上存储空间。注意力模块的计算存储要求则相对较低，CPU/中等性能网络处理器（NPU）即可完成。在大型模型推理中，如要满足每秒1万个令牌的实时要求，即令牌速率为10 000个/秒，对GPU的带宽需求将达到64 TB/s，而当前的HBM3带宽仅为0.8 GB/s。因此，对于十亿级以上规模的大型模型网络应用场景，现有的GPU/TPU+DRAM分离计算架构难以满足不断增长的模型参数传输带宽需求。



▲图3 LLAMa-7B模型全连接层和注意力模块参数维度示意图

这种情况表明，随着大型模型的不断发展和应用场景的扩大，现有的硬件架构在满足大规模模型计算需求方面面临着巨大的挑战。具体而言，参数量巨大且算力要求高的大模型导致了计算和存储资源高需求的问题，而当前的GPU/TPU+DRAM结构的带宽限制使得数据传输方面的瓶颈日益显现。因此，未来的硬件设计和架构需不断创新，以适应快速增长的大型模型计算需求，提供更高效的数据传输和处理解决方案。

2 存算一体集成芯片的优势

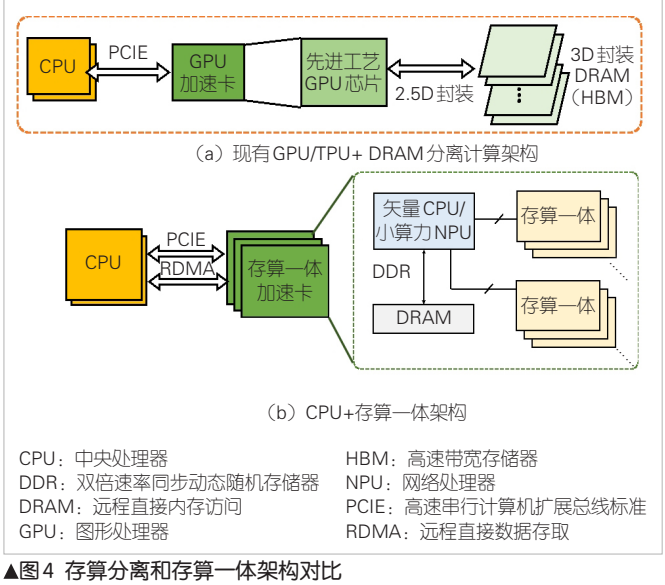
2.1 缓解带宽瓶颈

经典存算一体的设计基于交叉阵列。根据欧姆定律和基尔霍夫定律，输入特征用存储阵列的字线上的电压表示，输出特征会表示为位线上的电流大小，因此能够一次性完成矩阵乘加操作。同时，由于在计算过程中仅进行输入输出的搬运，权重系数一直固定在存储阵列中，所以能够显著减少数据搬移开销。我们发现，如果采用CPU+存算一体的组合的架构，相较于现有的GPU/TPU+DRAM分离计算架构（如图4所示），能够在相同的令牌速率和算力下，实现带宽的显著节约，达到xPU+HBM架构下1 000+倍的水平。举例来说，当采用和第1节相同的令牌速率（10 000个/s）时，存算一体架构仅需32～64 Gbit/s的带宽，就能节省超过1 000倍的带宽。

另一方面，当单颗芯粒的算力达10 TOPS，存储容量达到200 MB时，根据12/14 nm工艺估算，芯粒的计算电路面积约为8 mm^2，存储面积约为300 mm^2，此时实际的算力密度大约为0.032 5 TOPS/mm^2。因此，存算一体集成芯片架构相对于传统的数据中心系统不仅在性能上取得了显著的提升，还在所需的单芯粒接口速度远低于现有管控指标的前提下，为大规模模型的计算提供了更为可行的解决方案。

2.2 存边架构高并行度数据流

以图3所示LLAMa模型为例，我们对大模型全连接层算力和存储容量进行分析，其三层连续的全连接层网络的算力需求为：（4 096×11 008+11 008×4 096+4 096×4 096）×32×10 000 ×2≈68 TOPS；存储容量为：（4 096×11 008+11 008×4 096+4 096×4 096）×32≈3.4 GB，即模型的算力需求与存储容量的比值为目标令牌速率，与网络大小无关。在数据中心中，令牌速率约为1～10 000个/秒，经典的卷积神经网络模型ResNet-50的算力与存储比为4.1×帧率GOPS/25 MB=164×帧率（GOPS/kB)），因此大模型的算力存储比远低于以卷积神经网络（CNN）为主的传统深度神经网络（DNN）模型的算力存储比。传统交叉阵列架构算力存储比为时钟频率×2。为适应大模型的算力存储比，我们提出了存边计算架构（COMB），即将乘加计算逻辑分布在片上权重缓冲静态随机存储器（SRAM）的边缘，算力存储比



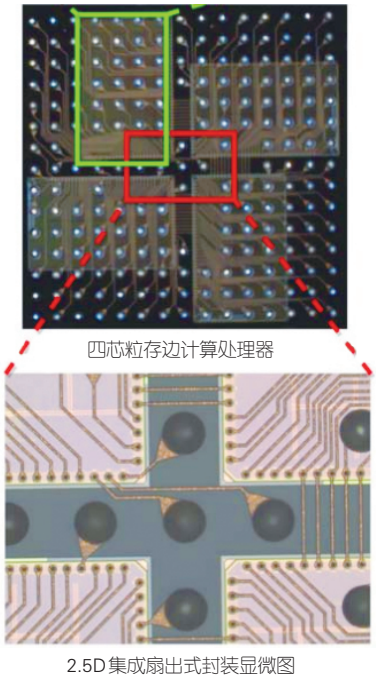
▲图4 存算分离和存算一体架构对比

为时钟频率×2/存储深度。近存计算架构中广泛使用的数据流映射方法完全可以运用在存边计算架构中，权重在计算开始前预先加载在COMB宏中，权重沿输入通道方向切块后，可以展平存入COMB宏的不同列。同时我们可以利用多个COMB Marco电路提高输出通道方向的并行度，完成空间并行计算。

2.3 存算一体技术分类

目前业界已有一些存储颗粒形态的存边计算商业实现方案：海力士（SK Hynix）提出的AiM的每颗DRAM芯粒含有0.5 GB的存储和512 GFLOPS的算力；三星提出的LPDDR5-PIM（存内计算）颗粒的峰值算力可达102.4 GFLOPS。与NPU相比，该设计提升了4.5倍的算力，并节省了72%的功耗。然而，高密度DRAM的工艺专用性强，与CMOS逻辑制造工艺的兼容性较差，且受制于读破坏和电荷泄漏，需要定期刷新存储。

传统嵌入式存储介质SRAM工艺下的微缩比例远远低于逻辑微缩比例。考虑到光刻极限，单芯片的最大SRAM在100 MB量级，且难以发生剧变。因此在过去，集成度一直限制了SRAM存算一体的发展。但随着2.5D/3D堆叠技术的发展，代工厂有望在SRAM上实现更高的集成密度，实现投影面积上等效晶体管密度的提升。如图5所示，我们基于集成扇出（FanOut）工艺，将4颗65 nm SRAM存边计算芯粒集成为一体，实现了SRAM存边计算架构算力和存储容量的显著提升。对于超过4颗芯粒集成的情况，映射方法尚需优化以实现算力随着芯粒数量的线性增长。除此之外，另一种存储颗粒形态的存边计算实现方案是阻变存储器（RRAM）。RRAM是一种能够通过改变二端器件的阻值来存储信息的技术，具有与CMOS工艺兼容性高、非易失、低读取功耗等特点。基于1TnR的RRAM存储器阵列通过三维堆叠技术，能够实现接近DRAM的高密度存储。这一技术趋势为存算一体提供了更为灵活和高效的解决方案。



▲图5 四芯粒2.5D集成Fanout封装[8]

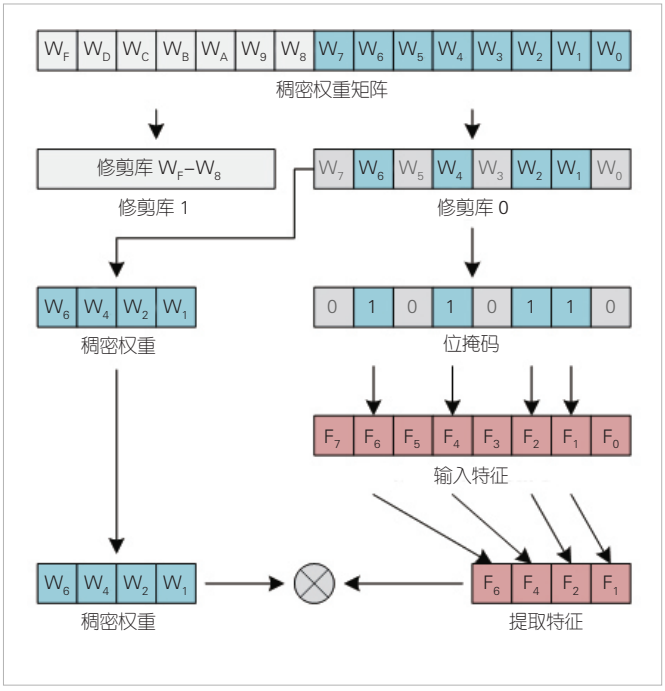
3 轻量化-存内压缩的协同设计

3.1 稀疏网络在存算一体上的部署挑战

随着参数和算力需求的不断增加，大型模型网络的存算一体架构的部署面临更多的挑战。幸运的是，稀疏技术为这一问题提供了一种软硬件协同设计的解决方案。首先，通过对大型模型网络的全连接层进行权重修剪，能够明显减少在生成查询、键和值矩阵时的参数存储需求。其次，大型模型网络所特有的注意力稀疏性进一步减少了自我注意机制的计算和存储需求。然而，在加速稀疏模型的存算一体架构中，仍然存在一些问题。传统的存算一体架构通常以一个交叉杆的形式组织来支持阵列级的计算并行性。在将非结构化剪枝的权重矩阵映射到交叉杆时，存储单元仍然需要保留零值权重，以维持计算的同步性。相较之下，结构化剪枝技术与并行处理更为兼容，但这会降低网络准确性。为了应对这些挑战，我们提出了一种存内稠密权重系数存储方案和基于蝶形网络的存算一体稀疏提取的激活拓扑网络。

3.2 存内稠密权重系数存储

图6展示了模型权重系数稀疏化和稠密存储方案的流程。首先，权重向量被划分为不同的剪枝子组，每个子组具有相同的大小，并按照预定义的稀疏度进行修剪。为了确定稀疏率和剪枝子组的大小，我们在Enwik-8和Text-8任务上使用12层注意力模型。在通过全局修剪对网络进行稀疏化时，我们发现在剪枝子组大小为32、修剪3/4的权重时，网络性能保持不变。因此，我们将剪枝子组大小设置为32，稀疏率设置为75%，以进行稀疏前馈计算。随后，剪枝后的权重被压缩为密集向量和用二进制编码表示的比特掩码，后者可以指示稠密权重的原始位置。最后，根据比特掩码的信息，我们需要从原始输入中提取和路由那些未跳过的输入特征。这一过程实现了对稀疏权重的有效处理。最终的乘积是通过将这两个稠密向量相乘得到的。整个流程的顺序性和稳



▲图6 稀疏–密集计算流程

健性保证了功能的正确性和高效性。

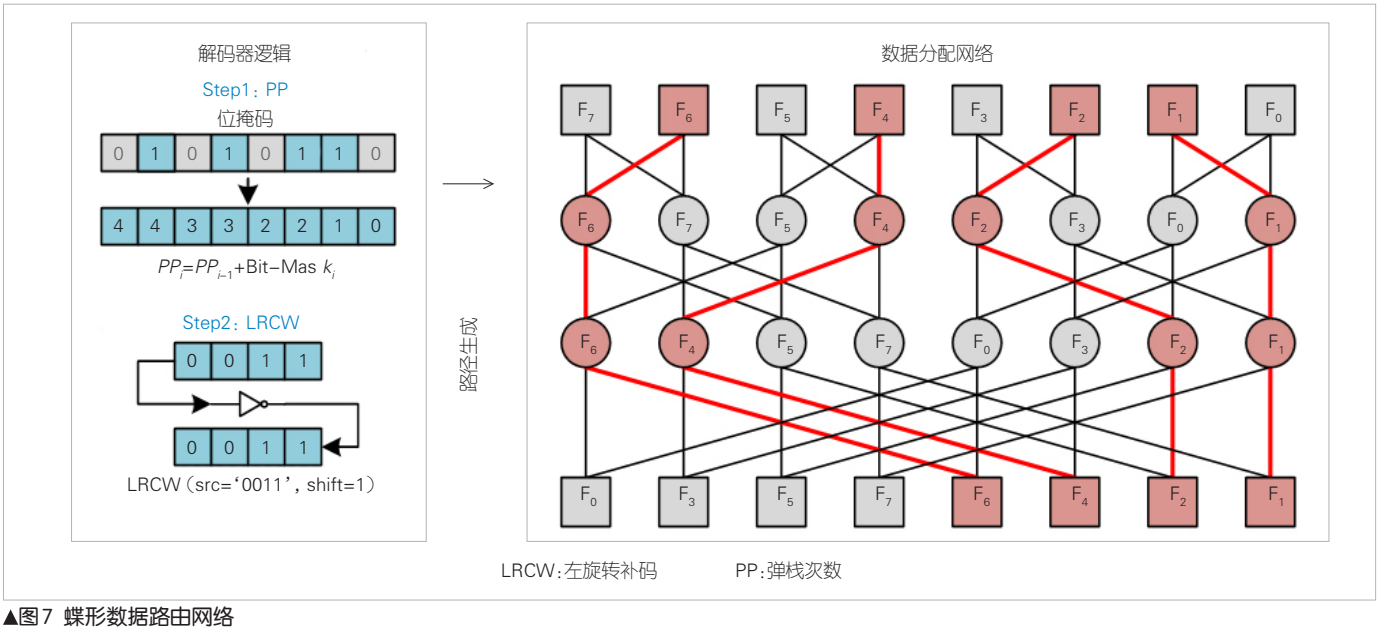
3.3 基于蝶形网络的稀疏提取拓扑

我们运用蝶形网络来提取压缩后的稠密权重所对应的输入激活特征。如图7所示，红色的特征经过蝶形的拓扑网络后，被路由至右侧。这个蝶形网络基于传输管的实现，而传输管的控制信号由解码器实时产生。解码器逻辑接收稠密权重的比特掩码，然后生成控制比特以配置蝶形网络中数据分发的路径。解码机制主要包含两个操作，即前缀pop计数和左旋转补码（LRCW）。前缀pop计数扫描位掩码的序列，并输出当前位置之前1的总数。LRCW是一个标准的左旋转，其唯一不同之处在于移位在任何时候都以补码形式表示。通过这样的操作，我们能够有效地处理比特掩码，从而实现对蝶形网络的灵活配置和输入特征的提取。

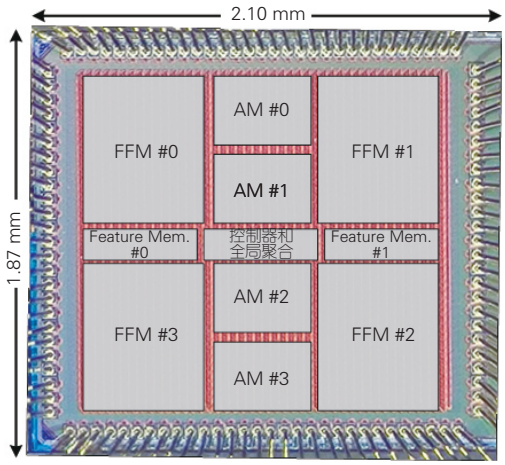
图8显示了采用28 nm CMOS工艺制造的芯片，该芯片工作频率高达320 MHz，总功耗为250.65 mW。考虑到网络稀疏性，该芯片峰值性能为3.3 TOPS。芯片面积3.93 mm^2，面积效率为0.85 TOPS/m^2。该芯片在生成查询、键和值矩阵和整体注意力方面分别实现了高达11.83/25.22 TOPS/W的系统能效。上述轻量化–存内压缩协同设计方案实现了稀疏网络在存算一体硬件上的稠密映射，显著提高存储密度和计算能效。

4 结束语

针对十亿级以上规模的大模型网络应用场景，目前的GPU/TPU+DRAM分离计算架构难以满足不断增长的系数数据传输带宽需求。为了缓解这一问题，存算一体的解决方案，特别是存边计算型的存储颗粒尤为重要，它们有望有效提高带宽。DRAM存算因具有高密度的特点，SRAM和RRAM因其具有高能效特点而备受关注。同时，存内压缩技术的应用可以实现稀疏网络在存算一体硬件上的稠密映射，从而同时提高存储密度和计算的能量效率。因此，在未来的



▲图7 蝶形数据路由网络



工艺节点	28 nm	
精度	INT8	
定制单元晶体管数目	8 T	80 kB
	10 T	64 kB
数字静态随机存储器	96 kB	
芯片面积/mm^2	3.93	
电压/V	0.64~1.03	
频率/MHz	20~320	
功耗/mW	8.27~250.65	
延迟/ms	Enwik8	841
	Text8	896
峰值性能/TOPS	3.33@320 MHz，1.03 V	
能量效率/(TOPS/W)	宏单元	2.52~32.17
	系统	1.96~25.22
面积效率(TOPS/mm^2)	0.85	

AM：注意力单元　FFM：前馈单元

▲图8 芯片照片和汇总表

发展中，矢量计算CPU与存算颗粒的结合有望成为大模型专用的硬件架构。这样的整合能够更好地应对大模型的计算需求，为数据中心芯片带来更为可持续和高效的解决方案。

参考文献

[1] JIAO Y, HAN L, JIN R, et al. 7.2 A 12nm programmable convolution-efficient neural-processing-unit chip achieving 825TOPS [C]// 2020 IEEE International Solid- State Circuits Conference - (ISSCC). IEEE, 2020: 136-140. DOI: 10.1109/ISSCC19947.2020.9062984
[2] DEAN J. 1.1 The deep learning revolution and its implications for computer architecture and chip design [C]//2020 IEEE International Solid- State Circuits Conference - (ISSCC). IEEE, 2020: 8-14. DOI: 10.1109/ISSCC19947.2020.9063049
[3] LIU S W, LI P Z, ZHANG J S, et al. 16.2 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 [C]//Proceedings of IEEE International Solid- State Circuits Conference (ISSCC). IEEE, 2023: 250-252. DOI: 10.1109/isscc42615.2023.10067360
[4] STOW D, XIE Y, SIDDIQUA T, et al. Cost-effective design of scalable high-performance systems using active and passive interposers [C]//Proceedings of IEEE/ACM International Conference on Computer-Aided Design (ICCAD). IEEE, 2017: 728-735. DOI: 10.1109/iccad.2017.8203849
[5] GOMES W, KHUSHU S, INGERLY B D, et al. 8.1 Lakefield and mobility compute: a 3D stacked 10nm and 22FFL hybrid processor system in 12×12mm2, 1mm package-on-package [C]// 2020 IEEE International Solid- State Circuits Conference - (ISSCC). IEEE, 2020: 144-146. DOI: 10.1109/ISSCC19947.2020.9062957
[6] NAFFZIGER S, LEPAK K, PARASCHOU M, et al. 2.2 AMD chiplet architecture for high-performance server and desktop products [C]//Proceedings of IEEE International Solid- State Circuits Conference - (ISSCC). IEEE, 2020: 44-45. DOI: 10.1109/isscc19947.2020.9063103
[7] SHAO Y S, CLEMONS J, VENKATESAN R, et al. Simba: scaling deep-learning inference with multi-chip-module-based architecture [C]//Proceedings of the 52nd Annual IEEE/ACM International Symposium on Microarchitecture. ACM, 2019: 44-45. DOI: 10.1145/3352460.3358302
[8] ZHU H Z, JIAO B, ZHANG J S, et al. COMB-MCM: computing-on-memory-boundary NN processor with bipolar bitwise sparsity optimization for scalable multi-chiplet-module edge machine learning [C]//2022 IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2022: 1-3. DOI: 10.1109/ISSCC42614.2022.9731657

作　者　简　介

何斯琪，复旦大学集成芯片与系统全国重点实验室在读硕士研究生；主要研究方向为面向大模型的存算一体SOC研究、深度学习的算法硬件协同设计；发表论文6篇。

穆琛，复旦大学集成芯片与系统全国重点实验室在读博士研究生；主要研究方向为基于易失性、非易失性存储器混合的存算一体SOC研究，通过算法架构电路协同的方式进行功耗及性能优化；发表论文4篇，申请专利2项。

陈迟晓，复旦大学芯片与系统前沿技术研究院副研究员、集成芯片与系统全国重点实验室集成芯片创新中心主任、国家优青、上海市青年科技启明星；研究方向包括人工智能芯片与系统、数模混合集成电路EDA以及先进封装、Chiplet集成；主持多个国家自然科学基金委面上项目；获上海市技术进步奖一等奖；发表论文50余篇，授权中国发明专利9项。