下一代AI大模型计算范式洞察

Insights into Computational Paradigm of Next-Generation AI Large Model

熊先奎/XIONG Xiankui，王程晨/WANG Chengchen，
蔡文豪/CAI Wenhao

（中兴通讯股份有限公司，中国 深圳 518057）
(ZTE Corporation, Shenzhen 518057, China)

DOI：10.12142/ZTETJ.202505008
网络出版地址：https://link.cnki.net/urlid/34.1228.TN.20250926.1148.002
网络出版日期：2025-09-26
收稿日期：2025-08-10

摘要：现代大模型规模随扩展定律持续扩大，近万亿的模型参数量带来了算法、硬件、工程领域的一系列困境。Transformer架构固有的计算效率低下问题愈发凸显，引发了研究人员对通用人工智能（AGI）实现路径的深入思考。一方面，针对现有自回归Transformer架构，已形成注意力机制、低精度量化、参数共享等算法改进方向，以及集群系统优化、硬件系统升级等工程改进方向；另一方面，下一代AI大模型计算范式正朝着不以next token prediction为核心的方向演进，具体包括两类路径：一是从更高抽象层次进行预测的扩散和联合嵌入预测架构，二是从物理第一性原理和计算基材特性出发构建的动力学模型、热力学模型和能量模型。同时，新型计算范式与新型计算基材相结合，有望从根本上改变传统AI算法软件与硬件割裂的局面，成为迈向AGI的高效路径。

关键词：大语言模型；计算范式；人工智能

Abstract: The continuous expansion of modern large-scale models, guided by scaling laws, has led to a series of challenges in algorithms, hardware, and engineering due to model parameters approaching the trillion-scale mark. The inherent computational inefficiency of the Transformer architecture has become increasingly evident, prompting in-depth reflection among researchers regarding the path to achieving artificial general intelligence (AGI). On one hand, improvements to the existing autoregressive Transformer architecture are being pursued along two main avenues: algorithmic enhancements such as refined attention mechanisms, low-precision quantization, and parameter sharing, as well as engineering advancements including cluster system optimization and hardware upgrades. On the other hand, the next-generation computational paradigms for AI models are evolving away from the core framework of next token prediction. This shift includes two distinct pathways: first, architectures that operate at higher levels of abstraction, such as diffusion and joint embedding prediction models; and second, approaches grounded in first principles of physics and the characteristics of computational substrates, including dynamic, thermodynamic, and energy-based models. Concurrently, the integration of novel computational paradigms with new computational substrates holds the potential to fundamentally alter the traditional disconnect between AI software and hardware, constituting an efficient pathway toward AGI.

Keywords: large language model; computational paradigm; artificial intelligence

引用格式：熊先奎, 王程晨, 蔡文豪. 下一代AI大模型计算范式洞察 [J]. 中兴通讯技术, 2025, 31(5): 50-56. DOI: 10.12142/ZTETJ.202505008
Citation：XIONG X K, WANG C C, CAI W H. Insights into computational paradigm of next-generation AI large model [J]. ZTE technology journal, 2025, 31(5): 50-56. DOI: 10.12142/ZTETJ.202505008

1 LLM现状及瓶颈

1.1 LLM架构相对固化的现状

2020年，OpenAI揭示了大模型规模扩展定律（Scaling Laws）[1]：大语言模型（LLM）的最终性能取决于计算量、参数量和训练数据量的堆叠扩展[2-8]。拥有175B参数量的GPT-3模型[2]在自然语言理解、知识问答等多项任务中，获得了远超同期模型的性能。近年来，以DeepSeek-V3、GPT-4o、Llama4、Qwen3、Grok4为代表的大模型无不在证明这个定律。

构建一款先进的基础大模型，需要堆叠数十万卡算力、收集数百太字节海量语料，基于自回归（AR）Transformer架构，采用预训练（Pre-training）和后训练（Post-training）等手段，完成其内部近万亿参数量的训练。整个训练过程的沉没成本极为高昂，如X.AI的Grok4模型，在2个150 MW功率的数据中心构建的20万卡分布式集群里，耗时半年才完成预训练。因此，LLM的预训练探索和实践主要在工业界

完成，而学术界只能集中在理论层面的研究和较小规模（参数量<7B）的实践。然而，尽管当前架构仍有一系列算法、硬件、工程、成本等瓶颈问题，为实现通用人工智能（AGI）的愿景并验证Scaling Law的有效性，产业界不断增大投入。模型规模持续增加的趋势短期内难以改变。

本文尝试从企业视角，剖析大模型架构的关键因素、发展契机和潜在技术路径。

1.2 LLM架构的关键瓶颈

Transformer架构的计算效率低，访存需求大[9]。特别是基于Decode-only的自回归结构算术强度仅为2，即每读取1字节数据只能完成2次计算。卷积神经网络（CNN）高达数百的算术强度，其高数据复用率可充分满足图形处理器（GPU）/特定领域架构（DSA）的矩阵乘加单元需求；而Transformer架构因数据搬移开销较大，导致模型算力利用率（MFU）较低。同时，当前硬件难以并行运算Transformer架构中的Softmax、Layer-norm、Swish等特殊非线性算子。总之，LLM架构对先进工艺和高带宽存储器（HBM）的依赖大、工程成本高，这是阻碍其规模应用、性能进一步提升的关键瓶颈。

未来，随着基础模型参数量的持续增加、推理模型长思维链输出上下文长度的飙升，以及以生物制药为代表的AI for Science等新型高性能计算应用的普及，Transformer架构瓶颈将愈发突出，这与摩尔定律放缓的趋势相矛盾。依赖先进工艺提升算力和能效的技术路径将遭遇“功耗墙”“内存墙”等问题。计算和存储分离的冯·诺依曼架构在大模型规模和算力不断增长的需求下将面临严峻挑战。

1.3 迈向AGI的LLM发展路线

当前LLM在实践过程中或多或少存在幻觉、可解释性差等问题，这些问题在Scaling Law不断提升模型能力的过程中被掩盖。但Transformer自回归架构的核心是“next token prediction”，导致部分AI科学家如LECUN等认为，从稀疏编码和等价映射原理看，现有LLM难以真正理解物理世界。因此，关于物理世界映射、世界模型构建的路线，在学术界仍有很大争议。

从工业界角度看，Scaling Law路线仍然需要进一步探索，因为平台期过后可能存在指数上升的拐点。这种路线的核心是商业闭环下的工程优化能力，同时需探索非AR模式乃至非Transformer模式的全新计算范式和算法[10]。未来AGI的发展路线，大概是开发能“感知”、能“物理思考”、能“实践”的认知大模型与具身大模型，这类模型需直接对齐可解释组件，并能通过实践反馈机制形成所谓的自主意识[11]。因此，高能效端侧硬件、高效率算法将成为探索具身大模型工程化的关键。

2 LLM自回归模式的工程改进和优化

针对前文所述问题，学术界和工业界基于自回归LLM开展了一系列算法、系统、硬件的改进和优化工作。

2.1 算法改进和优化

2.1.1 注意力机制优化

文档理解、代码分析、检索增强生成（RAG）等应用场景要求模型支持长上下文输入，而以DeepSeek-R1为代表的推理模型又要求模型支持长思维链输出。序列长度增加会导致自注意力机制计算复杂度呈$O(N^2)$上升。因此，分组查询注意力（GQA）、多头潜在注意力（MLA）等注意力机制的改进，以及以Flash-Attention为代表的算子优化，已被广泛采用，Linear-attention、RWKV、Mamba等线性注意力机制展现出巨大应用潜力。此外，旋转位置编码（RoPE）插值方案被进一步优化，部分注意力机制如原生稀疏注意力（NSA）、混合块注意力（MoBA），以及针对多卡场景的长上下文推理框架（如Ring-attention、Tree-attention）也被用来降低计算量。

2.1.2 低精度量化

Decode-Only架构中典型的运算过程是矩阵向量乘法（GEMV），该运算数据搬移频繁、计算效率低，既消耗算力，又占用带宽。

利用硬件原生FP8、FP4、MXFP等低精度数据类型进行模型量化，既能够有效减少内存带宽需求，又可以等效增加芯片算力利用率。现有研究证明，4 bit量化拥有相对最优扩展率[12]，在推理场景中已得到实际应用。然而，量化引入的误差，难免导致模型能力下降，同时非线性层的量化/反量化操作也有额外开销。因此，量化技术只能缓解计算和带宽瓶颈。

2.1.3 循环递归参数复用

循环式Transformer架构[13]，例如Universal Transformer、混合专家Universal Transformer（MoEUT）等，通过跨层共享参数实现深度递归。这类架构引入循环神经网络的递归表达能力后，通过参数共享使权重可支持多次计算，从而有效提升算术强度，在内存带宽受限时提升系统性能。然而，当前这种架构的实验规模较小，其扩展后的表达能力和稳定性尚不明确。

2.2 集群系统改进

传统CNN（如ResNet、Yolo）的网络参数量和计算量只在MB和GOPS（10亿次每秒）量级，在当前百TOPS级别算力（能效比2TOPS/W）的算力单元中，通常单卡/单机即可工作。而现代LLM由于巨大的参数量和计算量，会不可避免地引入多卡/多机的集群系统，通过张量并行（TP）、数据并行（DP）、流水线并行（PP）和专家并行（EP）等并行计算范式，加速训练和推理过程。

基于MoE的分布式计算范式可以降低超大参数规模模型的训练强度。其核心原理是每次前向计算时仅激活top-K个专家，从而降低算力需求。例如，Deepseek V3便通过这种方式将前馈神经网络（FFN）的计算量缩减为原来的1/32[14]。

P/D分离的部署可以利用Prefill/Decode在计算和带宽需求上的差异：Prefill阶段是计算密集型，追求首个token生成时间（TTFT）；Decode阶段是访存密集型，追求单个token生成时间（TPOT）。二者分离部署，不仅互不影响[15]，还能充分利用硬件利用率。

云端AI系统能够协同解决端侧算力资源受限情况下的大模型部署问题。端侧部署参数量较小的模型，可实现本地实时推理。对于复杂任务的拆解和深度思考任务，可通过云端部署参数量较大的模型来完成。分析结果将被反馈至端侧，从而通过端云AI协同搭建“快慢思考”系统[16]。

2.3 硬件工程优化

LLM集群借用了传统高性能计算（HPC）集群工程经验来优化当前计算范式，具有以下工程化技术创新：

1）微架构DSA化：在通用图形处理器（GPGPU）中，引入了更多DSA领域采用的专用架构设计。如Nvidia GPU Tensor Core引入异步数据搬移模式以及混合精度训练，借鉴数据流计算范式的相关经验。

2）互联优化：通过将集群划分为Scale Up和Scale Out域，引入匹配计算范式的互联技术。Scale Up作为高带宽域，使用总线类技术（如Nvlink），提供200 ns超低延迟、数千节点高并行度、原生内存语义的超节点连接，以摆脱Amdahl's law扩展率的约束。而Scale Out则借用远程直接内存访问（RDMA）类技术支持通用扩展，复用HPC集合通信原语（如NCCL），建立并行计算软件模型。

3）光电混合集群：在当前国产化算力能力受限情况下，基于硅光工艺以及晶圆级扩展的“小电算、大光联”软硬件架构有望成为构建万卡、10万卡以上集群的关键技术。

4）新型计算范式：在解决带宽问题的过程中，“存算一体”“Data-Centric”等突破冯氏架构“内存墙”“功耗墙”限制的一些新型计算范式也得到了高度关注。

5）算网存仿真平台：万卡以上超大规模集群部署的寻优问题，需要通过仿真平台对算、网、存系统进行算力部署和工作流的优化。构建高准确率、高时效性的仿真架构是亟待研究的问题。

当前，有两个前瞻性硬件工程技术至关重要：

1）基于光IO技术重构先进计算体系结构，是优化LLM计算范式的关键技术。可助力Scale Up百纳秒级超低延迟的超节点连接、内存池化和拉远等架构级创新。

2）基于3D动态随机存取存储器（DRAM）和无电容DRAM提供大容量、高带宽的内存，并结合LLM计算范式“读多写少”“顺序多于随机”等访存特点，采取异构介质（如高带宽闪存）、层次化缓存、压缩计算、存算一体等架构设计，构建超越高带宽内存（HBM）的新型内存体系。

3 下一代AI大模型计算范式演进和展望

通过Scaling Laws持续扩展超大参数模型实现AGI的路线，受到算力、带宽、能耗、语料多方面的限制。AGI的实现需要进行根本性变革，如将基于物理第一性原理的算法模型与计算基材硬件工程相结合。

3.1 下一代AI大模型发展趋势

产业界正在探索不以next token prediction为核心的下一代AI大模型范式。基于能量、动力学等第一性原理的模型由于能有效表述各种分布并在物理系统中自然演化，有望成为下一代AI大模型的核心架构。例如，由Hinton提出的玻尔兹曼机，受统计物理中伊辛模型和玻尔兹曼分布的启发，引入了随机、递归的神经网络，能够学习数据的潜在分布，解决复杂组合优化问题。后续的受限玻尔兹曼机和深度置信网络，促进了人工智能技术的快速发展，并促进了生成式模型在图像生成、自然语言处理和强化学习等领域中的广泛应用。

然而，这些基于能量、动力学原理的模型在现有冯·诺依曼计算机上运行时，其能耗和计算效率仍面临显著挑战。这是因为，基于布尔逻辑的确定性计算架构，在处理基于统计和概率的生成式模型时面临以下两个关键问题：其一，互补金属氧化物半导体（CMOS）器件的物理特性限制了其在随机过程模拟方面的硬件实现能力；其二，在面对自然语言处理中的语义模糊性、动态环境下的实时决策等非确定性需求时，现有计算范式效率显著下降。这一瓶颈催生了面向统计和概率等新型计算范式的需求：通过算法和硬件联合设

计，打破存储器与运算器分离的传统流程。这有望大幅提升能效比和计算性能，为突破当前AI算力瓶颈提供全新思路。

3.2 未来模型发展方向

目前，针对下一代AI模型设计主要有以下两种思路：

其一，可能仍是Transformer，但不再是next token prediction自回归。从更高抽象空间、更强表达能力、长期学习能力的目标出发，设计新一代模型结构，代表工作包括：1）Diffusion LLM架构[17]，代表模型包括LLaDA、Mercury等，通过扩散方法将自回归模型串行化生成过程，改进为从粗粒度到细粒度的并行化生成过程。在相同计算资源和模型规模下，这种架构能够提升10倍以上的推理吞吐量，将计算能耗减少到原架构的1/10，同时提升模型的逆向推理能力和上下文关注长度等指标性能；2）联合嵌入预测架构[18]，代表模型包括联合嵌入预测模型（JEPA）、大型概念模型（LCM）等，通过将语言、图像、视频等数据编码到高层潜空间中，学习世界模型级别的抽象表示，并在表示空间中通过基于能量的模型替代基于概率的模型进行预测，从而有效提升模型的表达效果与规划能力。

其二，基于物理第一性原理，从计算基材特性出发，根据物理过程的动力学特性、能量变化趋势设计模型架构和数据流，代表工作包括：1）液态神经模型（LFM）[19]，代表模型包括液态结构状态空间模型（LSSM），其核心原理是液态时间常数（LTCN）模型：

$$\frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t}=-\frac{\boldsymbol{x}(t)}{\tau}+f\left(\boldsymbol{x}(t),\boldsymbol{I}(t),t,\theta\right)\times\left(A-\boldsymbol{x}(t)\right) \quad (1)。$$

LFM是一种由小型生物神经动力学模型启发的新型时间连续循环神经网络（RNN），可以通过反向传播进行训练，并在时间序列预测任务中表现出良好的边界和稳定的动态特性、卓越的表达能力和较高的内存效率[20]。2）以Hopfield网络、受限玻尔兹曼机（RBM）、深度置信网络（DBN）等为代表的基于能量的模型（EBM），为概率密度估计和表示学习提供了一种统一的框架。这类模型的理论基础都可追溯到统计物理中的自旋玻璃模型。EBM通过定义能量函数来表示所希望学习的概率分布，因而也可作为生成模型来学习数据分布并生成与训练数据类似的新样本。与显式定义概率分布的模型相比，EBM具有更大的灵活性，能够建模更加复杂的依赖关系。近年来，基于能量的模型理论仍在不断发展，同时也面临不少挑战。其中，配分函数的计算和采样效率问题仍是制约模型应用的主要瓶颈。此外，能量函数的设计缺乏系统的指导原则，往往需要依赖经验和启发式方法。同时，模型的表达能力、泛化性能等仍缺乏更深入的研究。

3.3 下一代计算范式展望

在未来AI计算中，相较于算力，能耗将成为更为根本的限制。现有AI计算低效的根本原因是，神经网络的实现依赖于传统冯·诺依曼计算架构通过二进制操作“模拟”神经网络的计算。这种方法实质上是使用高精度的逻辑计算来处理仅需低精度的人工智能任务，大量能量被用于数据搬移和纠错，导致资源的低效利用。为了在进一步提高计算性能的同时降低计算能耗，研究者们探索了多种新型计算范式，其主要思想是采用非冯·诺依曼计算结构和存算一体。目前比较重要和热点的研究包括如下路线：

3.3.1 物理原理启发的计算架构

物理神经网络（PNN）是利用物理第一性原理构建人工智能的技术路径。现有技术路线包括光计算、量子计算、电磁计算等。

光计算是一种利用光子作为信息载体进行计算和传输的计算模式，具有超高速度、超高带宽、低延迟、高并行等优势。光计算利用光干涉、衍射、强度/相位调制等物理特性直接在模拟域执行特定的计算任务，尤其在AI计算中展现出颠覆性潜力。例如，清华研究团队推出了太极系列光计算系统，利用空间对称和互易特性实现了训推一体的光神经网络（ONN）[21]。但光计算目前仍面临集成度、器件性能、系统复杂度、精度、软件生态等多重严峻挑战，成熟度仍然较低。

量子计算是一种遵循量子力学规律调控量子信息单元进行计算的新型计算模式。现有的量子算法和量子神经网络框架需在有限的量子比特和较大的计算错误率约束条件下运行。例如，使用量子加权张量混合网络（QWTHN）实现大模型微调[22]，将FFN训练转化为二次无约束二次规划问题（QUBO）并通过量子Ising机求解，利用量子位构建储层并实现储备池计算等。然而，量子计算目前由于技术路线未收敛、量子比特位数量有限、工作环境苛刻等问题，暂时难以实现广泛应用。

电磁计算直接利用电磁波（微波、毫米波、太赫兹波）的特性进行信息处理，而非依赖传统的电子开关状态。其核心优势包括超高速操作、高并行性、低传输损耗等。计算实现形式主要分为微波/毫米波模拟计算、可编程电磁处理[23]以及电磁存内计算。电磁计算通过物理定律直接映射数学运算，在特定领域（线性变换、实时处理）展现出应用潜力，当前仍处于实验室阶段。

3.3.2 基于材料特性的模拟计算架构

研究者们正探索多种神经形态器件，这些器件利用材料

的本征物理现象模拟生物系统的复杂行为，通过特定的连接方式，构建单元间相互耦合的系统，能够利用系统自身演化特性替代传统计算过程。因此，利用材料的本征特性，推动算法、软件与硬件的联合设计，有望根本性地改变传统AI算法软件与硬件割裂的局面，从而实现软硬件的协同优化。现有技术路线包括概率计算、吸引子网络、热力学计算等。

概率计算系统依赖具有真随机特性的概率比特单元（p-bit），它是位于量子计算和数字逻辑之间的中间计算范式，能够比传统计算机更好地利用自然和概率的潜在属性，在组合优化、因式分解、密钥生成、马尔可夫链蒙特卡洛（MCMC）采样等应用场景中均有较大优势。此外，概率计算系统还能够训练随机神经网络和深度生成模型，例如深度玻尔兹曼机[24]。

吸引子是动力系统中不同初始条件下趋向的一组数值，可以在动力学系统中实现记忆功能。2024年，LI等利用可变电阻式存储器（RRAM）器件的双向阻变特性实现回滞型神经元[25]，并据此构建了一种双极性忆阻器电路涌现的循环神经网络，相比于传统Hopfield网络具有硬件高效、记忆容量大等优势。

热力学计算基于热力学原理，利用自然界固有的计算能力，开发新的信息处理网络的设计原则，应用于未来计算系统。Normal Computing通过构建具有精确表达的状态空间、表现力丰富的非线性函数以及可扩展能力的硬件单元，从而高效地从复杂分布中进行采样，解决物理仿真和机器学习任务中的计算瓶颈问题。

3.3.3 生物启发的计算架构

生物启发计算通过模拟自然系统的信息处理机制重构计算架构，突破传统冯·诺依曼瓶颈。目前主流的研究方向包括类脑计算和DNA计算等。

类脑计算泛指一类受脑启发的新型信息处理架构，这类架构依托大规模并行计算平台，有望突破存储与计算分离的冯·诺依曼架构瓶颈，为通用智能问题提供高能效解决方案。

DNA计算是一种利用分子的生化特性进行信息存储与处理的新型计算范式，具有高存储密度、低功耗等优势。未来DNA计算将通过硅基和生物混合计算，赋能AI时代数据处理。

生物启发计算架构正从专用加速器向通用计算范式跃迁。短期看，类脑计算芯片在边缘智能领域将率先爆发；中长期则将形成“硅基+生物群体协同”的融合架构，最终实现生物级能效的智能计算系统。

4 中兴通讯面向下一代AI大模型计算范式的探索与实践

4.1 存内计算架构

中兴通讯利用8T SRAM数字存内计算技术实现了12.31 TOPS/W@INT8的高能效AI加速器，同时也在进行xpu-pim异构架构探索，如图1所示。该架构基于压缩和量化实现端

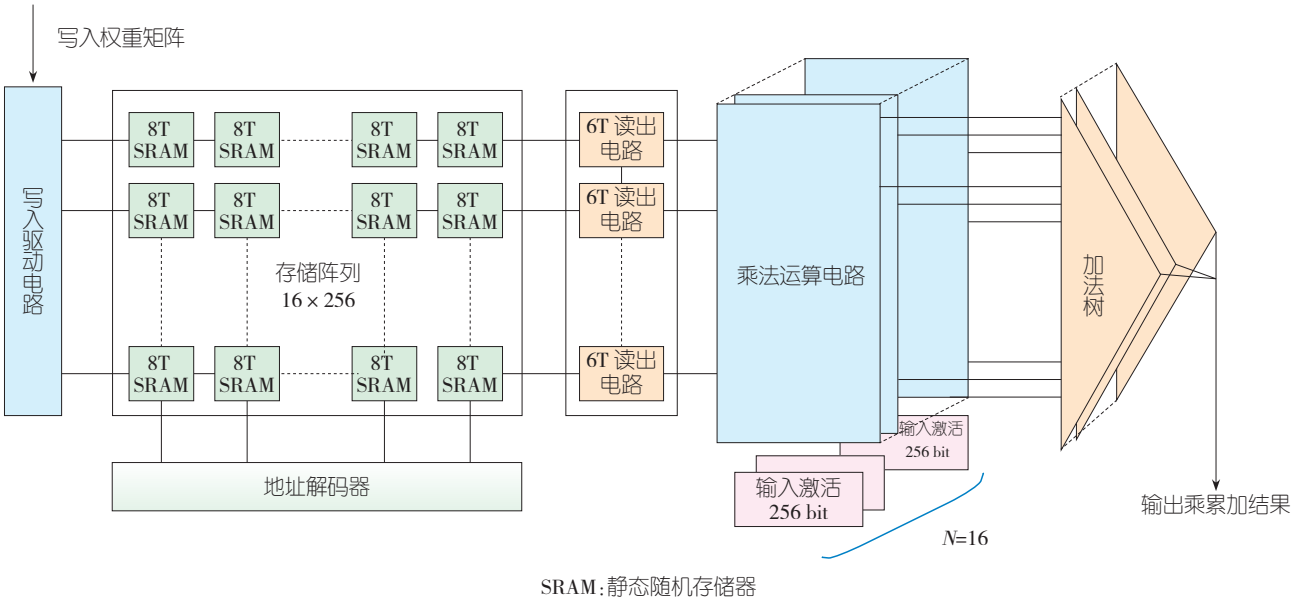

图1　存内计算架构示意图

侧大模型加速，在能效和吞吐量上具有数量级提升，近存架构将在端侧场景下发挥显著能效优势。

4.2 新型AI算法和硬件实现

中兴通讯在新型AI算法和硬件实现方面，探索了从物理第一性原理出发的新型技术路线。例如，基于循环式Transformer架构的高效参数共享特性，中兴通讯探索了其在替代多层Transformer架构上的能力。使用GPT-2 small的单个Transformer层作为模型“基块”，可以在减少超过50%参数量的同时保持模型的表达能力不下降。随着基块结构的改进，基块层数和循环次数可以进一步降低，如表1所示。

同时，稀疏玻尔兹曼机（DBM）架构由于其稀疏特性和基于最小化能量的推理目标，特别适合利用非易失性存储器执行端侧低功耗任务。DBM的能量方程可以描述为：

$$E = -\sum_{i<j} \boldsymbol{J}_{ij} m_i m_j + \sum \boldsymbol{h}_i m_i \quad (2),$$

其中，$\boldsymbol{J}_{ij}$是耦合矩阵，$\boldsymbol{h}_i$是偏置向量，m_i表示每个神经元的状态[18]。各个神经元串行更新并达到玻尔兹曼平衡的过程可以表示为：

$$m_i(t) = \mathrm{sgn}(\tanh[-\beta I_i(t)] - \mathrm{rand}_{\mathrm{U},[-1,1]}) \quad (3),$$

$$I_i(t+\Delta t) = \sum \boldsymbol{J}_{ij} m_j(t) + \boldsymbol{h}_i \quad (4)。$$

在数千神经元的规模下，利用GPU完成单batch训练需要超过10 h。而基于FPGA的DBM的快速计算单元，采用概率计算范式，通过例化数千个神经元及它们之间的稀疏连接，从而将单batch的训练时间缩短至5 min，实现了超过2个数量级的加速效果。未来，使用RRAM、MRAM等非易失性存储器件，能够进一步降低计算开销，提升推理速度，以满足DBM在端侧推理场景的广泛应用需求。

此外，在光连接、新型内存等支撑性工程技术，以及计算存储分离的数据池化系统、内存语义互联系统、大规模仿真平台等架构技术方面，中兴通讯也展开了一系列前瞻性研究。

5 结束语

现代LLM基于Scaling Laws持续扩展，参数量接近万亿。巨大的模型规模引发了跨越算法设计、硬件架构和系统工程的多方面挑战。基于二次注意力复杂度的Transformer架构的内在计算效率瓶颈越来越显著，而这也推动了对通用人工智能可行途径的思考。

一方面，对当前流行的自回归变压器范式的增量改进，集中在算法改进（稀疏注意力机制、低精度量化、参数共享）和工程优化（集群系统、硬件工程）上。另一方面，越来越多的研究正在探索超越next token prediction的计算范式，代表性方向包括：1）基于扩散和联合嵌入预测架构的更高抽象层次模型；2）从物理学和底层计算基础得出的第一性原理模型，具体包括动力学模型、热力学模型和能量模型。

至关重要的是，这些新兴的算法范式与新型计算基材（如神经形态、光子和模拟内存加速器）的融合，为统一的硬件-算法协同设计框架提供了前景，有望成为通往AGI的高效路径。

表1　基于GPT-2、Qwen3基块构建循环神经网络的训练结果

基块类型	GPT-2	GPT-2	GPT-2	GPT-2	GPT-2	Qwen3
基块层数	12	1	6	3	2	1
循环次数	1	12	6	12	18	6
最佳损失	3.35	3.65	3.30	3.35	3.45	3.41

参考文献

[1] KAPLAN J, MCCANDLISH S, HENIGHAN T, et al. Scaling laws for neural language models [EB/OL]. (2020-01-23) [2025-08-15]. https://arxiv.org/abs/2001.08361

[2] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners [EB/OL]. (2020-05-28) [2025-08-15]. https://arxiv.org/abs/2005.14165

[3] 田海东, 张明政, 常锐, 等. 大模型训练技术综述 [J]. 中兴通讯技术, 2024, 30(2): 21-28. DOI: 10.12142/ZTETJ.202402004

[4] 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

[5] 冯文佼, 李宗航, 虞红芳. 低资源集群中的大语言模型分布式推理技术 [J]. 中兴通讯技术, 2024, 30(2): 43-49. DOI: 10.12142/ZTETJ.202402007

[6] REN T Q, LI R P, ZHAO M M, et al. Separate source channel coding is still what you need: an LLM-based rethinking [J]. ZTE communications, 2025, 23(1): 30-44. DOI: 10.12142/ZTECOM.202501005

[7] 裴丹, 张圣林, 孙永谦, 等. 大语言模型时代的智能运维 [J]. 中兴通讯技术, 2024, 30(2): 56-62. DOI: 10.12142/ZTETJ.202402009

[8] 韩炳涛, 刘涛. 大模型关键技术与应用 [J]. 中兴通讯技术, 2024, 30(2): 76-88. DOI: 10.12142/ZTETJ.202402012

[9] 朱炫鹏, 姚海东, 刘隽, 等. 大语言模型算法演进综述 [J]. 中兴通讯技术, 2024, 30(2): 9-20. DOI: 10.12142/ZTETJ.202402003

[10] ZHAO H, WU H Q, YANG D J, et al. BriLLM: brain-inspired large language model [EB/OL]. (2025-03-14) [2025-08-15]. https://arxiv.org/abs/2503.11299

[11] PILOTO L S, WEINSTEIN A, BATTAGLIA P, et al. Intuitive physics learning in a deep-learning model inspired by developmental psychology [J]. Nature human behaviour, 2022, 6(9): 1257-1267. DOI: 10.1038/s41562-022-01394-8

[12] OUYANG X, GE T, HARTVIGSEN T, et al. Low-bit quantization favors undertrained LLMs: scaling laws for quantized LLMs with 100T training tokens [EB/OL]. (2024-11-26) [2025-08-15]. https://arxiv.org/abs/2411.17691

[13] DEHGHANI M, GOUWS S, VINYALS O, et al. Universal

transformers [EB/OL]. (2025-03-17)[2025-08-15]. https://arxiv.org/abs/1807.03819v3
[14] DAI D D, DENG C Q, ZHAO C G, et al. DeepSeekMoE: towards ultimate expert specialization in mixture-of-experts language models [EB/OL]. (2024-01-11) [2025-08-15]. https://arxiv. org/abs/2401.06066
[15] ZHONG Y M, LIU S Y, CHEN J D, etc. DistServe: disaggregating prefill and decoding for goodput-optimized large language model serving [EB/OL]. (2024-06-06) [2025-08-15]. https://arxiv.org/pdf/2401.09670
[16] TIAN X Y, GU J R, LI B L, etc. DriveVLM: the convergence of autonomous driving and large vision-language models [EB/OL]. (2024-06-25)[2025-08-15]. https://arxiv.org/pdf/2402.12289
[17] NIE S, ZHU F Q, YOU Z B, etc. Large language diffusion models [EB/OL]. (2025-02-14) [2025-08-15]. https://arxiv. org/abs/2502.09992
[18] ASSRAN M, DUVAL Q, MISRA I, et al. Self-supervised learning from images with a joint-embedding predictive architecture [C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023: 15619-15629. DOI: 10.1109/CVPR52729.2023.01499
[19] HASANI R, LECHNER M, AMINI A, et al. Liquid time-constant networks [J]. Proceedings of the AAAI conference on artificial intelligence, 2021, 35(9): 7657-7666. DOI: 10.1609/aaai.v35i9.16936
[20] HASANI R M, LECHNER M, WANG T H, et al. Liquid structural state-space models [EB/OL]. (2022-09-26) [2025-08-15]. https://arxiv.org/abs/2209.12951
[21] XUE Z W, ZHOU T K, XU Z H, et al. Fully forward mode training for optical neural networks [J]. Nature, 2024, 632: 280-286. DOI: 10.1038/s41586-024-07687-4
[22] KONG X F, LI L, DOU M H, etc. Quantum-enhanced LLM efficient fine tuning [EB/OL]. (2025-03-17)[2025-08-15]. https://arxiv.org/abs/2503.12790v1
[23] LIU C, MA Q, LUO Z J, et al. A programmable diffractive deep neural network based on a digital-coding metasurface array [J]. Nature electronics, 2022, 5: 113-122. DOI: 10.1038/s41928-022-00719-9
[24] NIAZI S, CHOWDHURY S, AADIT N A, et al. Training deep Boltzmann networks with sparse Ising machines [J]. Nature electronics, 2024, 7: 610-619. DOI: 10.1038/s41928-024-01182-4
[25] LI Y X, WANG S Q, YANG K, et al. An emergent attractor network in a passive resistive switching circuit [J]. Nature communications, 2024, 15(1): 7683. DOI: 10.1038/s41467-024-52132-9

作 者 简 介

熊先奎，中兴通讯股份有限公司无线首席架构师、智算技术委员会前瞻组组长；长期从事计算系统和体系结构、先进计算范式以及异构计算加速器研究工作；曾主导过中兴通讯 ATCA 先进电信计算平台、服务器存储平台、智能网卡和 AI 加速器等系统架构设计。

王程晨，中兴通讯股份有限公司技术预研工程师；主要研究方向为大模型软硬件协同设计、先进计算范式等。

蔡文豪，中兴通讯股份有限公司技术预研工程师；主要研究方向包括深度学习算法、大语言模型、模拟计算、无线通信系统等。