面向分布式推理的大模型键值缓存分发网络

Large Model Key-Value Cache Distribution Network for Distributed Inference Services

宋俊平/Song Junping[1]，黄天一/Huang Tianyi[1,2]，
周旭/Zhou Xu[1]，李智/Li Zhi[3]，吴茜/Wu Qian[3]
（1. 中国科学院计算机网络信息中心，中国 北京 100083；
2. 中国科学院大学，中国 北京 100190；
3. 中国移动通信集团有限公司，中国 北京 100032）
(1. Computer Network Information Center, Chinese Academy of Sciences, Beijing 100083, China;
2. University of Chinese Academy of Sciences, Beijing 100190, China;
3. China Mobile Communications Group Co., Ltd., Beijing 100032, China)

DOI：10.12142/ZTETJ.202603009
网络出版地址：https://link.cnki.net/urlid/34.1228.TN.20260629.1815.007
网络出版日期：2026-06-26
收稿日期：2026-04-25

摘要：大模型推理已成为智能算力体系的核心，传统云端集中式推理面临带宽压力大、推理延迟高、算力资源紧张等瓶颈。现有云边协同方案在通信开销与键值缓存（KV-Cache）复用效率方面仍存在明显不足。为此，提出一种基于内容分发网络（CDN）的大模型KV-Cache分发网络（KVCDN），将传统CDN升级为具备张量感知与状态管理能力的分布式缓存架构。该架构采用端侧语义规范化、边缘热点分级缓存与核心全局去重编排3层协同结构，并结合提示词（Prompt）标准化、分级存储预放置、全局去重融合与语义路由4项关键技术，从而显著提升KV-Cache复用率，降低冗余计算与推理延迟。

关键词：大模型推理；键值缓存(KV-Cache)；内容分发网络(CDN)；云边协同；缓存复用

Abstract: Large language model inference has become a core component of intelligent computing systems. Traditional cloud-centric inference approaches face critical bottlenecks such as high bandwidth consumption, excessive latency, and intensive computational demands. Existing cloud-edge collaborative schemes still suffer from significant communication overhead and low cache reuse efficiency of Key-Value Caches (KV-Cache). To address these issues, a KV-Cache distribution network based on content delivery network (CDN), termed KVCDN, is proposed. In this architecture, conventional CDN nodes are upgraded into a distributed caching system with tensor-aware perception and state management capabilities. A three-layer collaborative structure is adopted, consisting of semantic normalization at the end side, hotspot-aware hierarchical caching at the edge, and global deduplication with orchestration at the core. Combined with four key enabling techniques, i.e., prompt standardization, tiered storage and proactive pre-placement, global deduplication and fusion, and semantic-aware routing, the KVCDN significantly improves KV-Cache reuse ratio, while reducing redundant computations and inference latency.

Keywords: large model inference; Key-Value Caches (KV-Cache); content delivery network (CDN); cloud-edge collaboration; cache reuse

引用格式：宋俊平, 黄天一, 周旭, 等. 面向分布式推理的大模型键值缓存分发网络 [J]. 中兴通讯技术, 2026, 32(3): 45-50. DOI: 10.12142/ZTETJ.202603009
Citation: Song J P, Huang T Y, Zhou X, et al. Large model key-value cache distribution network for distributed inference services [J]. ZTE technology journal, 2026, 32(3): 45-50. DOI: 10.12142/ZTETJ.202603009

1 研究背景

随着大模型从实验室走向千行百业，人工智能（AI）技术正从以训练为核心转向以推理为核心。国际数据公司（IDC）在其发布的《中国人工智能计算力发展评估报告》中指出，到2027年，中国推理算力占整体智能算力的比例将突破72.6%，占据绝对主导地位。

第十四届全国人民代表大会第四次会议和中国人民政治协商会议第十四届全国委员会第四次会议的相关提案指出，中国当前面向海量推理任务的算力集群和网络调度能力仍存在显著缺口。与此同时，通用大语言模型的广泛部署与海量并发应用带来了极高的算力和访存挑战。传统高度集中的云端单点推理模式不仅给国家骨干网络带来了巨大的带宽压

力，也越来越难以满足海量用户对低延迟交互的严苛要求。

作为互联网内容分发的重要基础设施，内容分发网络（CDN）凭借其广泛的省市级边缘节点部署，天然具备为用户提供低延迟服务的优势。然而，传统CDN主要面向静态文件或流媒体内容进行缓存，直接将大模型部署于CDN之上，在创新性和资源利用率方面均存在明显不足。大模型推理的核心瓶颈在于处理自回归生成过程中产生的海量键值缓存（KV-Cache）。随着大模型对超长上下文处理需求的爆发式增长，中间态张量KV-Cache的管理正从单机显存优化，逐步向广域分布式解耦与跨节点协同架构演进。如何突破单机物理硬件的局限，利用网络拓扑的空间分布特性实现KV-Cache的高效分级与调度，已成为当前网络与系统交叉领域的前沿热点。

2 研究现状

在大模型算力需求爆发的初期，学术界主要沿用了传统深度学习中的模型切分思路，提出了一种基于算力空间转移的云边协同架构[1]。该架构通过在物理空间上对大语言模型（LLM）的前向计算图进行基于层级或注意力头的静态切分[2]，将计算负载较低的嵌入层及浅层Transformer模块部署于边缘节点，用于提取初级语义特征、承接初始流量并兼顾用户隐私保护[3]；而将对显存需求极高的深层模块与全量权重保留在云端智算中心，以确保复杂推理的精度[4]。这一策略旨在通过层级间的流水线并行，缓解集中式数据中心在处理海量并发请求时面临的瞬时算力压力[5]。

然而，这种静态切分方案在LLM特有的自回归推理场景下面临严峻的通信挑战。

由于LLM中间层输出的张量维度远超原始文本，且解码（Decode）阶段频繁的跨节点同步容易受到广域网带宽波动的影响，导致推理时延显著增加。近期研究（如CE-CoLLM[6]）进一步证实，云边协同中LLM上下文数据的跨节点传输已成为核心通信瓶颈，并初步探索了延迟感知的云端上下文管理机制。

为突破静态切分方案带来的“通信墙”瓶颈，学术界开始探索具备环境感知能力的动态协同机制。Splitwise系统[7-8]（ISCA 2024、UCC 2025）提出了一种基于Lyapunov优化理论的动态协同框架。该研究摒弃了粗粒度的层间切割方式，转而通过深度强化学习实时感知网络带宽与节点负载，在注意力头和前馈神经网络（FFN）等细粒度层级上自适应调整云边切分边界。这一突破将云边协同从静态物理划分提升至动态资源调度层面，显著增强了系统在不稳定网络环境下的推理鲁棒性。

在此基础上，Jupiter系统[9]（INFOCOM 2025）进一步针对LLM推理中预填充（Prefill）与Decode阶段的计算不对称性进行了优化。该研究指出，传统方案往往仅注重优化预填充阶段，而严重忽视自回归解码阶段，导致解码过程中因“逐层、逐Token”的串行依赖而使边缘资源长时间空转。为此，Jupiter通过对边缘硬件进行精细化的负载轮廓化建模，提出了一种计算与通信深度重叠的流水线解码机制，即利用云端处理当前Token深层权重所耗费的时间窗口，驱动边缘端同步预执行下一个Token的浅层计算，从而有效消除了流水线中的“计算气泡”。这种从空间切分向时间维度相位解耦的演进，为实现低延迟、高吞吐的生成式服务提供了关键支撑。

尽管上述基于流水线切分的云边协同研究在理论上极具吸引力，但在实际部署于跨省、跨市的广域网环境中时，其内在的结构性缺陷依然明显。首先，通信开销的本质瓶颈尚未消除。模型切分技术在推理运行期必然于切割层产生海量的高维隐状态张量。这类中间计算结果的跨网流转对物理带宽的绝对容量与抖动极其敏感，在网络长尾延迟的影响下，极易引发首字延迟（TTFT）严重超时，难以满足实时交互业务的苛刻需求。其次，更为致命的症结在于，这类架构将LLM推理视作一条无记忆的刚性单向流水线。现有研究普遍忽视了真实业务场景中海量用户提示词（Prompt）之间在语义内容上的强相关性、上下文的重叠度以及计算状态的可复用性，导致严重的重复计算与存储冗余。

为了突破上述物理切分方案带来的跨网张量通信瓶颈，近期的协同推理研究重心正逐渐从单一的“模型空间切割”转向非对称的“云-端协同”与“边-端异步协同”。针对终端算力逐渐增强且广域网条件存在波动的场景，学术界广泛引入了基于投机解码的协同架构[10-11]。该机制的核心在于，将传统大语言模型逐字生成的串行模式重构为“端侧起草—云/边侧验证”的并行范式：首先，由部署在用户终端或边缘节点的轻量级小模型凭借其极速响应能力，一次性“投机”起草一段候选文本；随后，由算力充沛的云端大模型对该段草稿进行一次性的并行校验。这种“端侧小模型试错、云端大模型定论”的非对称计算结构，从根本上改变了自回归生成中“算一个词、传一个词”的低效模式，同时也降低了大模型推理对网络实时性的严格依赖。该方案在保证输出精度与原生大模型相当的前提下，显著降低了云端交互频率，并提升了端到端的响应速度。

在此背景下，分布式拆分投机解码（DSSD）系统[12]（ICML 2025）针对分布式投机解码中“验证数据量过大”这一瓶颈，展现了较强的创新性。传统方案中，终端小模型每

生成一段草稿，均需向边缘或云端上传庞大的词表概率分布以供校验，导致较高的上行通信开销。DSSD将原本由边缘节点承担的复杂计算逻辑进行逆向重构，使原本需要海量上传的概率数据转化为极其轻量化的单次下行确认信号，从而实现了通信延迟的数量级压缩。这一成果标志着分布式推理从单纯压缩数据体积向重构通信逻辑的重要跨越。

虽然投机解码大幅优化了单次任务的执行效率，但其底层机制仍未有效解决大模型推理的核心难题——推理状态的持久化与语义复用。

在现有系统中，动态上下文KV-Cache虽然占据着庞大的显存空间，却在每次验证或请求结束后被视作无保留价值的临时变量而直接清空。这种无记忆的资源处理模式，导致边缘侧在处理具有高度语义重叠的并发任务时（如多轮对话、共用参考资料）陷入“计算孤岛”困境——每条语义关联的请求都被迫从零开始进行重复的预填充与特征提取，从而造成了严重的计算冗余与存储浪费。因此，如何从“瞬态通信优化”转向“稳态状态管理”，构建具备跨任务记忆能力的云边协同架构，已成为当前制约大语言模型迈向大规模工业化应用的核心科学问题。

随着超长文本推理与检索增强生成（RAG）技术的普及，学术界逐渐形成共识：大模型推理产生的中间态数据亟需引入一种具备持久化、模块化与空间分布式特性的新型基础设施。这一共识直接催生了将传统CDN向“大模型知识网络”演进的理论构想。

芝加哥大学系统研究团队在操作系统原理会议（SOSP）2024及2026年的后续工作中，首次明确提出了“知识分发网络（KDN）”的概念[13]，并开源了核心原型系统LM-Cache[14]。该研究明确指出，传统的模型微调虽能内化知识，但会严重破坏模型的模块化与响应速度；而常规的上下文学习则会在预填充阶段产生超线性增长的显存与计算成本。为此，KDN提倡将大模型基于外部知识源预计算生成的KV-Cache进行物理剥离，将其视作互联网中的“静态视频或网页资源”进行独立管理与分发。在KDN架构中，系统通过独立的存储模块，以及支持多源KV-Cache动态拼接组合的智能混合模块，使大模型能够像传统CDN分发多媒体流一样，在地理分布的存储节点之间实现推理状态的动态组装。

综上所述，尽管投机解码等技术显著优化了云边端协同推理中单次任务的执行效率，现有系统仍将大模型推理视为无记忆的流水线，未能解决推理状态持久化与语义复用这一核心难题。同时，随着超长文本推理及RAG技术的普及，学术界逐渐形成共识：大模型的中间态数据亟需引入一种具备持久化与空间分布式特性的新型基础设施[15]。正如KDN概念所倡导的，将大模型预计算生成的KV-Cache视作互联网中的“静态网页资源”进行独立管理与物理剥离，已成为突破“计算孤岛”困境的关键路径。

3 基于CDN的大模型键值缓存分发网络

在广域算网融合与分布式AI技术高速演进的背景下，大模型推理服务正经历从“云端集中式单点处理”向“云边端多级协同”的深刻范式转移。传统CDN主要面向静态文件、图片或流媒体视频的分发而设计，其底层逻辑基于统一资源定位符（URL）精确哈希与就近物理路由。然而，在大模型协同推理服务中，核心性能瓶颈已转移至自回归生成过程中产生的海量高维动态特征张量——KV-Cache。同时，Prompt具有高度非结构化、强口语化及上下文依赖的特征，使得传统CDN的静态文件缓存机制在此类高维、动态、碎片化的智能算力场景下完全失效。

为了彻底打破“缓存击穿”与“管理粗放”的瓶颈，本文提出一种基于CDN的大模型键值缓存分发网络（KVCDN）。该方案在不改变现有算网物理拓扑的前提下，将传统数据分发节点全面升级为具备张量感知与状态管理能力的智能存储单元。整体技术路线依托“端侧规范化重塑—边缘领域热点分级—核心全局去重”3层协同机制，深度融合计算感知路由与可编程网络能力，从而形成一套端到端的大模型缓存优化解决方案。

如图1所示，KVCDN架构在逻辑上划分为3个核心功能层：

1）端侧/接入侧：作为防范“缓存击穿”的第一道防线，该层通过部署轻量级模型，对异构的口语化输入进行意图对齐、无损压缩与规范化重写，从数学源头上提升请求特征的结构化程度，并生成全局一致的语义指纹。

2）边缘下级CDN侧：作为“存”的核心承载层，该设计摒弃单机内部的图形处理器-中央处理器-固态硬盘（GPU-CPU-SSD）硬件级存储思维，创新性地提出一种面向“物理空间—社会领域”的多级CDN网络拓扑分级存储架构。结合计算感知理念，利用特定地理区域的社会属性（如专科医院、科研院所），精准定位垂直领域的KV-Cache热点，并实现动态预放置。

3）核心上级CDN侧：作为全局资源统筹枢纽，该层针对广域网中海量且时空碎片化的冗余张量，构建基于前缀感知的全局字典树索引。通过特征匹配与选择性重计算技术，突破严格前缀匹配的空间限制，实现跨区域的物理去重与融合共享。

面向大模型特征张量分发设计的KVCDN，从端侧请求

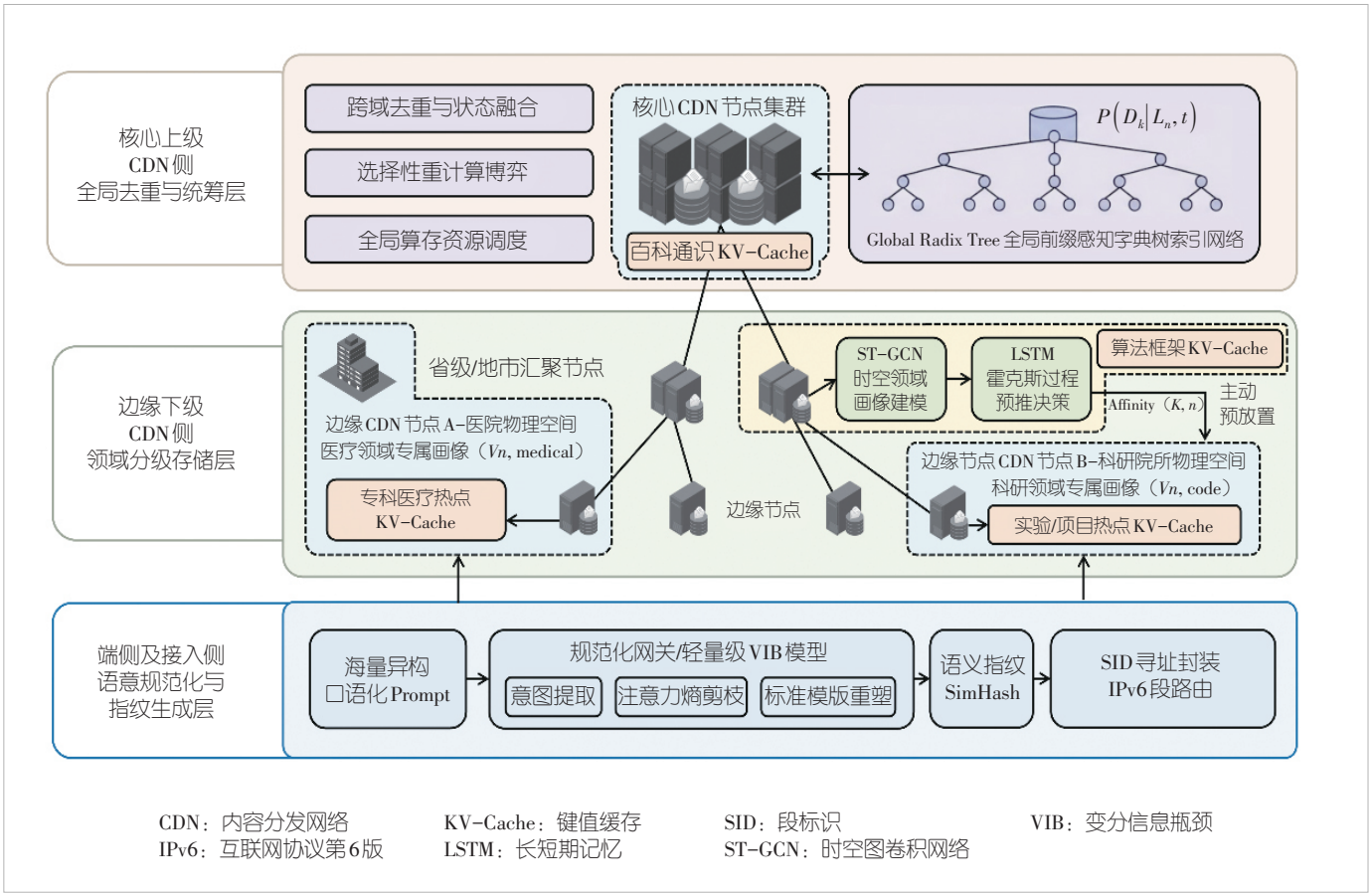


图1 一种基于CDN的大模型键值缓存分发网络架构

重塑到多级张量编排调度，提供了一整套系统性的缓存优化解决方案。针对高度异构且口语化的端侧大模型请求，该系统能够感知并提取核心语义意图，剖析非结构化输入与标准化提示词之间的映射关系，进而精准刻画并重塑具备统一前缀的标准化请求，为后续KV-Cache的高效聚合与全网命中提供源头基础。面对大模型长上下文所带来的海量显存消耗与动态多层级特征，该系统从局部高频响应与全局特征复用的角度，对KV-Cache张量进行编排管理，充分利用边缘与核心等多级算存资源，提升分布式特征张量的处理与调度效率。同时，结合广域网中多并发请求的高度时空变化特征，系统能够感知并预测局部长上下文热点态势，并根据冗余态势与前缀相似度，动态调整多层级间的存储与去重融合方案。

4 关键技术

为了提升基于CDN的大模型键值缓存分发网络的大规模分布式推理性能，解决其面临的云边端推理协同困难、缓存管理粗放、计算性能低下等问题，我们提出以下核心关键技术思路：

1）语义等价性驱动的端侧Prompt规范化与无损压缩机制

广域网中的海量大模型推理请求普遍存在极强的个体表达异构性。即便多个用户在核心逻辑与所需知识库上完全一致，诸如语气词、停用词或句法结构等方面的微小差异，也会导致CDN节点在进行底层特征张量匹配时产生严重偏差。这种输入异构性直接使本可复用的KV-Cache被系统判定为“未命中”，从而引发显著的算力浪费与冗余传输。

在端侧（或接入侧网关）引入基于语义等价性的规范化机制：基于信息瓶颈（IB）理论的轻量级序列提纯算法，构建轻量级编码器，将用户的原始口语化提示词序列 $X=[x_1,x_2,\cdots,x_N]$ 映射为紧凑表示 Z，在压缩原始序列表示（最小化 $I(X;Z)$）的同时，最大程度保留与核心意图相关的信息（最大化 $I(Z;Y)$）。在提纯意图的基础上，进一步对Prompt进行字粒度的无损压缩，利用预置在端侧的小型语言模型（如极小参数量的蒸馏模型）的注意力掩码机制，计算

每个Token的信息熵；利用局部敏感哈希（LSH）机制，将高维的语义嵌入向量映射为低维特征空间中的紧凑比特串$F(X_{std}) \in \{0,1\}^d$，作为后续网络路由层在IPv6扩展报头中封装的标识符，从源头确保具备相同语义的请求能够被精准汇聚。

2）基于时空热度与空间拓扑感知的KV-Cache分级存储与动态预放置方法

大模型推理所产生的上下文数据体量庞大，且不同地理区域的用户请求呈现显著的地域特征与“语义热点”集聚现象。例如，中关村地区的科技类Prompt较为集中，而关于某些全球性社会议题的询问则分布极广。在此背景下，如何在CDN分级存储架构下，根据区域特征实现分布式KV-Cache的精准预热与动态替换，并进而基于语义与缓存热点完成精准路由适配，以最大化缓存命中率与复用率，已成为亟待解决的复杂网络调度问题。

考虑到广域网中大模型推理请求的地域聚集性与圈层性特征，以及多级CDN网络中不同层级节点在覆盖范围、并发承载力与存储资源等方面的差异，需建立基于时空特征的动态张量热度刻画模型。进一步地，从网络拓扑与用户分布的角度出发，设计面向KV-Cache的“拓扑分级”存储策略：将具备强地域性且局部极热的KV-Cache精准下沉并锚定至最接近端侧的边缘节点；将访问用户分散、通用性强的系统级KV-Cache上移至上级汇聚节点存储。同时，在基于KV-Cache的结构化合并树（如LSM-tree）底层引擎之上，结合局部区域的热度演化预测，通过跨边缘节点的热点主动预放置与负载均衡备份机制，消除突发高并发场景下的单点性能瓶颈，从而实现局部高通量推理请求的就近极速响应。

3）基于前缀感知与选择性重计算的全局编排与去重融合方法

在广域网环境中，大量冗余张量并非完全一致。在多轮智能对话、RAG或多文档阅读等典型应用场景中，用户常基于相同的背景知识库（如一份长篇学术报告）插入不同的提问引导语，或对文档序列进行不同的拼接组合。这些碎片化操作打破了自回归模型所依赖的“严格绝对位置前缀匹配”规则，导致传统哈希缓存直接判定缓存失效，进而使不同CDN节点内堆积了大量实际高度同源却因非严格前缀限制而彼此独立的冗余KV张量碎片。

提出跨节点的KV-Cache深层解耦与动态去重融合方法。该方法在上级核心CDN节点构建全局前缀字典树索引目录，用于感知跨域节点间的冗余张量态势。针对非严格前缀序列，引入“选择性重计算”机制以恢复交叉注意力，从而根据全局冗余态势做出即时合并反馈，并迅速消除各节点间的碎片化张量，最终实现全局空间去重与资源公用，最大化广域算存资源的统筹与利用率。

4）基于KV-Cache缓存的语义路由技术

在现有CDN边缘网关处引入轻量级语义解析与动态调度能力，将高维自然语言请求转化为网络路由友好的低维标识，并融合全网算力与存储状态进行多目标联合寻优调度。

关键技术思路包含以下核心维度：

（1）高效注意力裁剪抽取方法。面向大模型前缀匹配与长文本检索场景，设计高效注意力裁剪与抽取方法，将冗长的Prompt转化为紧凑的特征向量，在保障语义精度的同时，满足微秒级网络路由的时延要求。

（2）分布式语义路由表。以语义向量和LSH为核心，构建分布式语义路由表。该表将稠密语义向量降维映射为定长（如64位或128位）的二进制哈希签名，并通过计算海明距离来刻画语义相似度。

（3）云边协同分级路由与联合调度架构。在云边协同环境下，设计全局与局部路由表分级架构。综合考量KV-Cache缓存复用率（即预填充计算降低率）、CDN边缘节点实时算力队列负载以及云边网络传输时延，构建联合调度模型。利用深度强化学习或在线凸优化算法，实现高并发场景下的动态路由决策策略，有效防止热点缓存引发的局部流量风暴。

5 结束语

面向大模型推理从云端集中式向云边端协同演进的趋势，以及KV-Cache复用不足、广域通信开销大、推理延迟偏高的核心痛点，本文提出基于CDN的大KVCDN。通过将传统CDN升级为具备张量感知、状态管理与语义调度能力的智能分发架构，该方案实现了从“单机硬件调度”向“广域网络拓扑协同”的演进，以及从“被动等待命中”向“主动语义重塑”的转变。本文系统梳理了所提架构面临的核心关键挑战及可行的技术解决思路，旨在为大模型时代新型网络基础设施的发展提供参考方向。

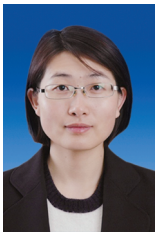
参考文献

[1] Qu G Q, Chen Q Y, Wei W, et al. Mobile edge intelligence for large language models: a contemporary survey [PP/OL]. arXiv (2025-03-20)[2026-04-20]. https://arxiv.org/abs/2407.18921

[2] Kafetzis D, Khalili R, Koutsopoulos I. Large language model partitioning for low-latency inference at the edge [C]// Proceedings of the 23rd International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks

(WiOpt). IEEE, 2025: 1–8
[3] Cunningham M. Privacy–aware split inference with speculative decoding for large language models over wide–area networks [PP/OL]. arXiv (2026–02–18) [2026–04–22]. https://arxiv. org/abs/2602.16760
[4] Sung M, Palakonda V, Im S, et al. Memory– and latency–constrained inference of large language models via adaptive split computing [PP/OL]. arXiv (2025–11–06) [2026–04–20]. https://arxiv.org/abs/2511.04002
[5] Feng Z D, Lu L, Li Q, et al. Distributed inference optimization for large language model in edge–cloud collaborative networks [C]// Proceedings of ICC 2025 – IEEE International Conference on Communications. IEEE, 2025: 6161–6166. DOI: 10.1109/icc52391.2025.11160773
[6] Jin H P, Wu Y Z. CE–CoLLM: efficient and adaptive large language models through cloud–edge collaboration [C]// Proceedings of IEEE International Conference on Web Services (ICWS). IEEE, 2025: 316–323. DOI: 10.1109/icws67624.2025.00046
[7] Patel P, Choukse E, Zhang C J, et al. Splitwise: efficient generative LLM inference using phase splitting [C]//Proceedings of ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA). IEEE, 2024: 118–132. DOI: 10.1109/ISCA59077.2024.00019
[8] Younesi A, Shabrang M A, Oustad E, et al. Splitwise: collaborative edge – cloud inference for LLMs via Lyapunov–assisted DRL [C]// Proceedings of the 18th IEEE/ACM International Conference on Utility and Cloud Computing. ACM, 2025: 1–11. DOI: 10.1145/3773274.3774267
[9] Ye S Y, Ouyang B, Zeng L K, et al. Jupiter: fast and resource–efficient collaborative inference of generative LLMs on edge devices [C]//Proceedings of IEEE INFOCOM 2025 – IEEE Conference on Computer Communications. IEEE, 2025: 1–10. DOI: 10.1109/infocom55648.2025.11044734
[10] Chen C, Borgeaud S, Irving G, et al. Accelerating large language model decoding with speculative sampling [PP/OL]. arXiv (2023–02–02)[2026–04–18]. https://arxiv.org/abs/2302.01318
[11] Leviathan Y, Kalman M, Matias Y. Fast inference from transformers via speculative decoding [C]//Proceedings of the 40th International Conference on Machine Learning. ACM, 2023: 19274–19286. DOI: 10.5555/3618408.3619203
[12] Ning J, Zheng C, Yang T. DSSD: Efficient edge–device LLM deployment and collaborative inference via distributed split speculative decoding [PP/OL]. arXiv (2025–07–16) [2026–04–18]. https://arxiv.org/abs/2507.12000
[13] Cheng Y, Du K, Yao J, et al. Do large language models need a content delivery network [PP/OL]. arXiv (2024–09–16)[2026–04–15]. https://arxiv.org/abs/2409.13761
[14] Liu Y, Cheng Y, Yao J, et al. LMCache: an efficient KV cache layer for enterprise–scale LLM inference [PP/OL]. arXiv (2025–12–05)[2026–04–18]. https://arxiv.org/abs/2510.09665
[15] Team TAib. AIBrix: towards scalable, cost–effective large language model inference infrastructure [PP/OL]. arXiv (2025–02–22)[2026–04–18]. https://arxiv.org/html/2504.03648v1

作 者 简 介

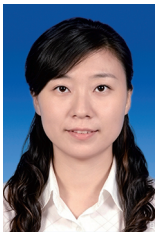


宋俊平，中国科学院计算机网络信息中心高级工程师、中科院青年创新促进会会员、中国通信标准化协会TC630语义智简通信技术与标准推进委员会资源及模型管理组副组长；主要研究方向为算力网络、人工智能；获省部级奖励1项；发表论文30余篇，申请国家发明专利24项。

黄天一，中国科学院计算机网络信息中心在读硕士研究生；主要研究方向为分布式推理、算网融合。

周旭，中国科学院计算机网络信息中心研究员，担任中国通信标准化协会TC614标准推进委员会副主席、中国通信学会算力网络委员会委员、中国通信学会信息通信网络技术委员会委员、重庆大学通信学院兼职教授/博导等学术职务，国家重点研发计划项目首席科学家，国家级人才计划入选者；主要研究方向为计算机网络体系结构、人工智能、5G/6G移动网络；近年来作为项目负责人完成重大项目10余项，获得部级科学技术奖一等奖4次、二等奖2次；发表论文80余篇，申请专利60余项（其中已授权25项），撰写专著1本，主持/参与制定国际标准/行业标准15项。

李智，中国移动通信有限公司网络事业部高级技术专家，高级工程师；主要研究方向为互联网技术、CDN、算力网络、边缘计算等；获得北京市科技进步奖1项，中国通信学会科学技术奖1项，工业和信息化部中国信息通信研究院科技进步奖9项，互联网、CDN、云、人工智能产业协会等科技进步奖4项，中国移动科技进步奖2项；发表论文18篇，拥有专利36项，参与研发标准24项。



吴茜，中国移动通信有限公司网络事业部高级工程师；主要研究方向为CDN的云化、智能化关键技术；获2020年中国通信学会科学技术奖三等奖、2023年工业和信息化部中国信息通信研究院“华彩杯”算力应用创新大赛北区总决赛一等奖、2024年全国信息通信领域安全生产和网络运行安全成果一等奖等。