算力网络人工智能模型推理算力度量研究报告

中国联通研究院
中讯邮电咨询设计院
联通数字科技有限公司

2025 年 8 月

版权声明

目录

前 言

随着人工智能技术的迅猛发展，模型推理已成为算力需求的核心驱动力。从 AI 搜索、智能体的兴起到多模态内容生成的广泛应用，模型推理的算力需求呈现出前所未有的加速态势。在此背景下，算力网络作为计算与网络深度融合的新型基础设施，为人工智能模型推理提供了灵活、高效的算力支持。然而，如何精准度量模型推理所需的算力资源，并实现算力的高效调度与优化，是当前行业面临的重要挑战。

本报告深入研究了算力网络人工智能模型推理算力度量的理论框架、方法体系及关键技术，并结合典型应用案例验证其有效性。本报告旨在提出一套科学、系统且可落地的算力度量方案，推动人工智能的规模化、普及化应用。

编写组成员（排名不分先后）：

中国联通研究院：曹畅、张岩、刘永生、王施霁、曹云飞、崔煜喆

中讯邮电咨询设计院：刘扬、尼松涛、张奎、裴培、何万县、段谊海、马威、申佳、周旭晖、王迪

联通数字科技有限公司：温源、姜辉、刘点、刘文涛、宋占军

一、研究背景与意义

（一）算力网络概念与架构

算力网络是指在计算能力不断泛在化发展的基础上，通过网络手段将计算、存储等基础资源在云-边-端之间进行有效调配的方式，以此提升业务服务质量和用户的服务体验。

中国联通在探索计算与网络融合思路的基础上，结合业界先进经验，制定了算力网络体系架构，如图 1 所示。在该算力网络架构图中，主要包含服务提供层、服务编排层、网络控制层、 算力管理层和算力资源层/网络转发层等若干功能模块，其中服务提供层主要实现面向用户的服务能力开放；服务编排层负责对虚机、容器等服务资源的纳管、 调度、配给和全生命周期管理；网络控制层主要通过网络控制平面实现算网多维度资源在网络中的关联、寻址、调配、优化与确定性服务；算力管理层解决异构算力资源的建模、纳管与交易等问题；算力资源层和网络转发层扁平化融合，并需要结合网络中计算处理能力与网络转发能力的实际情况和应用效能，实现各类计算、存储资源的高质量传递和流动。

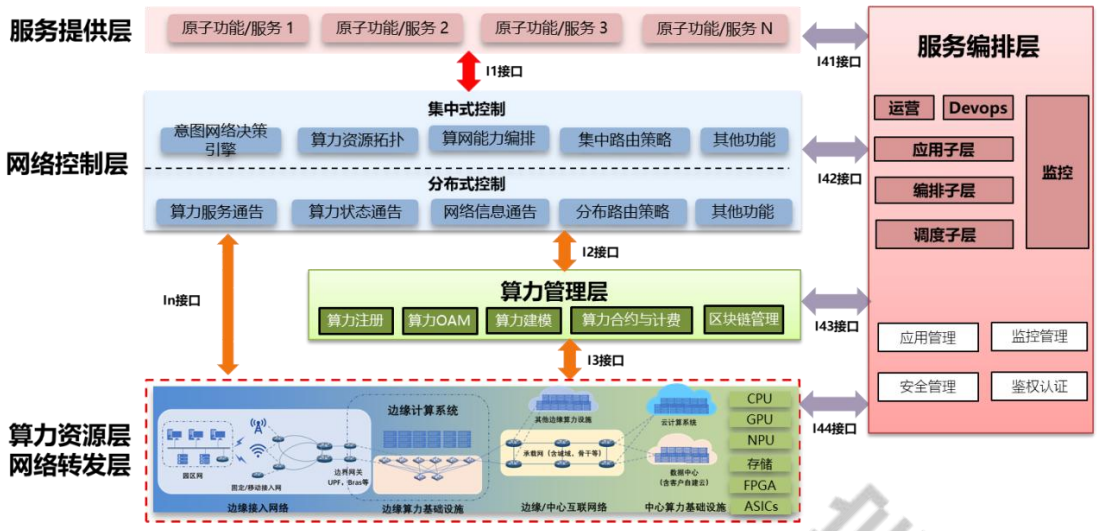

图 1 中国联通算力网络架构

算力网络中的计算资源类型包括通用 CPU、专用 AI 芯片（如 GPU、FPGA、ASIC 等）、以及各类加速卡等。不同类型的计算资源在处理 AI 模型推理任务时，性能表现各异，如 GPU 擅长并行计算，适合大规模矩阵运算为主的深度学习模型推理；而 FPGA 在特定定制化推理任务中具有低功耗、高能效优势。

算力网络被明确定义为提供“整体算力服务”，并强调“按需分配和灵活调度计算资源、存储资源以及网络资源”。这种转变意味着网络的核心价值正从单纯的数据传输演进为集成化的计算能力交付。这不仅是网络技术的升级，更是数字经济中价值创造的新范式，它将网络从被动的传输介质提升为主动的资源编排者，对未来的网络架构、服务模型和商业模式都将产生深远影响。

（二）人工智能模型推理的算力需求加速

2025 年被认为是"推理之年"，AI 模型推理已成为算力需求增长的主要驱动力。根据行业分析，推理算力需求规模"轻松超过去年估

计的 100 倍"。

1）搜索 AI 化转型，如谷歌搜索在今年 5 月 21 日正式迎来 AI 模式，并逐步在美国市场推出，考虑到谷歌搜索全球范围内年搜索量为 5 万亿次+，假设单次回答平均为 2000token，则该功能都将带来日均 27 万亿 token 消耗，类似案例如抖音搜索、微博 AI 智搜，搜索功能开始从普通服务器迁移到 AI 服务器并重塑所有搜索体验；

2）智能体爆发，智能体和深度思考推理的结合，通过两者结合，智能体执行任务准确率大幅提高，智能体执行一次任务平均消耗 token 达到十万亿的量级，大幅超过 AI 搜索单次问答 token 消耗，并且能延伸到更多开放式场景，同时多智能体协作的群体智能也已开始逐步商用化，过去复杂、多步骤的任务可通过智能体实现，智能体的普及将带来推理算力需求的大幅增长；

3）多模态内容生成，随着多模态生成的图片及视频质量今年均显著提升，今年 AI 营销内容占比提升十分明显，根据《2025 中国广告主营销趋势调查报告》显示“超过 50%的广告主，已经在生成创意内容时使用 AIGC，并且 AI 营销内容占比超过 10%”，而一分钟视频的生成 token 消耗基本在 10 万亿 token 量级，目前多模态模型开始步入快速商业化阶段，如快手可灵四五月连续两月付费金额超过 1 亿，多模态的加速渗透带来明显的算力需求提升。

4）大模型推理普及，如 OpenAI o1、DeepSeek R1 等推理模型的广泛应用，国内豆包的 token 消耗数量从 2024 年的 1200 亿增长到

2025 年的 16 万亿，增长大约 160 倍。

图 2 豆包 token 消耗量

（三）算力网络为人工智能模型推理提供算力

算力网络通过整合和优化网络中的各种算力资源，能够为人工智能模型推理提供强大的算力支持。在算力网络中，人工智能模型推理可以根据任务的特点和需求，灵活地分配到不同的计算节点上。例如，对于大规模的任务，可以利用云计算中心的强大算力进行集中处理；而对于实时性要求较高的推理任务，如智能语音助手、智能驾驶中的实时决策等，可以将推理任务卸载到离用户更近的边缘计算节点上，以降低时延，提高响应速度。

算力网络的广泛覆盖和便捷接入，使得人工智能模型能够更快速地应用到各个领域，如医疗、金融、教育、制造业等，为这些领域的数字化转型和智能化升级提供有力支撑。例如，在医疗领域，算力网络可以支持医学影像分析、疾病预测等人工智能应用，帮助医生更准确地诊断疾病；在金融领域，能够实现风险评估、智能投资顾问等功能，提高金融服务的效率和质量。

算力网络还能够实现算力资源的动态调度和共享。当某个地区或

某个时间段内对人工智能模型推理的算力需求激增时，算力网络可以从其他资源闲置的区域或时间段调配算力资源，实现资源的高效利用。这种动态调度和共享机制不仅能够提高算力资源的利用率，还能够降低企业和用户的计算成本，模型推理的成本是选择模型推理的重要考虑因素之一。

二、算力网络人工智能模型推理算力度量

（一）算力度量相关研究进展

算力网络算力度量领域的研究进展主要分为论文研究、白皮书和行业标准三个部分，以全面呈现当前算力评估与建模技术的发展现状。

学术界对算力度量的研究日益深入，尤其是在算力网络和异构计算场景下。研究方向主要聚焦于如何构建科学、统一的度量体系，并将其应用于算力调度与资源管理。在算力网络方面，多篇论文探讨了相关的技术体系和应用。例如，杜宗鹏等人在《中兴通讯技术》中提出了“算力网络四面三级算力度量技术体系”，为算力网络中的度量提供了一个分层、多维度的框架。李一男等人的研究则以“以服务为中心”的视角，对算力网络的度量与建模进行了深入分析。乔楚等人将算力度量与算力资源调度紧密结合，为实际应用提供了思路。针对不同的应用场景，研究者们也提出了相应的算力度量方法。在端边云协同计算的背景下，姜海洋等人和冯汉枣等人分别提出了适用于该场景的度量方法。姜海洋等人侧重于端边云的整体协同，而冯汉枣等人则针对异构混合云服务下的多任务场景。此外，还有研究聚焦于具体的度量指标和特定设备。王磊等人介绍了“BOPs”这一新型算力度量指标，为算力评估提供了新的视角。在方面，祝淑琼等人的研究探讨了如何对物联网端侧设备的算力进行度量。

白皮书为算力度量提供了宏观的指导和框架，国家人工智能标准

化总体组和全国信标委人工智能分委会联合发布的《计算中心有效算力评测体系白皮书（2022）》，系统地构建了计算中心的有效算力评测体系，为评估大规模计算设施的综合算力提供了权威的指导原则。

算力度量相关的标准正在逐步建立和完善。中国联通牵头的《面向公共通信业务体验的算力量化与建模技术要求》，将算力度量与用户体验紧密结合，强调了算力评估应以服务质量为导向。其他重要标准包括《算力网络 算力度量与算力建模技术要求》、《算力网络异构算力资源计算能力度量指标》和《算力网络算力节点能力度量及评估方法》，这些标准为算力网络中的算力度量和建模提供统一的技术规范，确保不同厂商和平台之间的互操作性和一致性。

（二）算力网络人工智能模型推理算力度量概念

算力网络人工智能模型推理算力度量是指对算力网络执行推理任务所需的算力资源进行量化评估，其核心目标是根据人工智能模型推理特性和规模，预估模型推理所需计算资源，从而为模型部署、算力调度、算力交易提供决策依据，同时能够识别模型推理过程中的瓶颈和故障。

算力网络人工智能模型推理算力度量中模型推理规模占据主导因素，数据的规模和类型也会对算力需求产生显著影响，同时，还需考虑计算资源的实际使用效率，在模型推理过程中，并非所有的计算资源都能被充分利用。还应与算力网络的整体性能指标相结合，网络的带宽、延迟、吞吐量等因素都会对模型推理的实际效果产生影响。

此外，从用户侧来看算力网络人工智能模型推理算力度量是用户使用人工智能模型推理的使用量，直接的体现是使用模型推理的时长，最终模型推理的使用要传导到算力资源的使用量上。

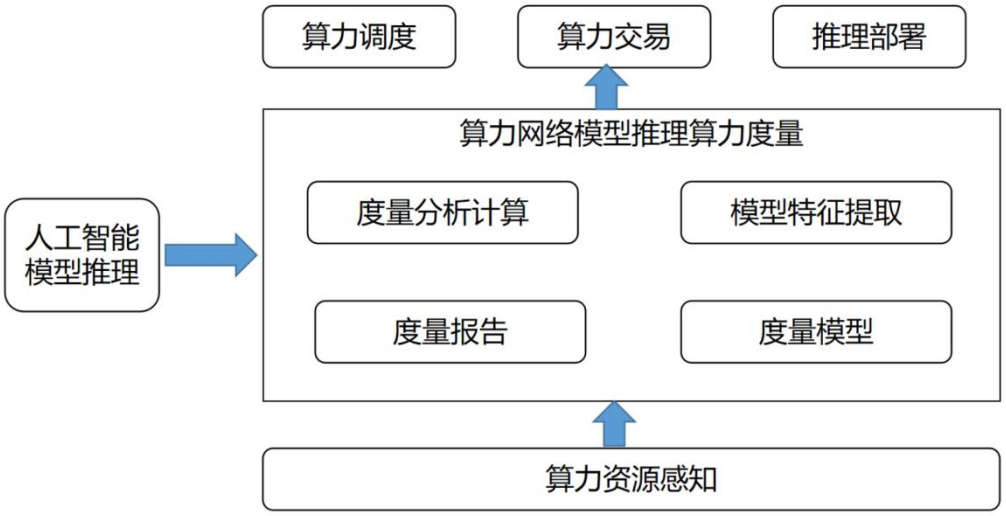

图 3 算力网络人工智能模型推理功能示意图

图 3 展示了算力网络中人工智能模型推理的算力度量功能框架。人工智能模型推理需要在算力网络中运行，算力资源感知将算力资源信息输入给算力度量模块，算力度量结果用于算力调度、算力交易、推理部署的应用。算力度量模块中，度量分析计算模块整合模型特征提取信息，依据度量模型生成完整的度量报告，从而评估推理性能、能效及服务质量。

三、模型推理算力度量方法

（一）算力度量模型

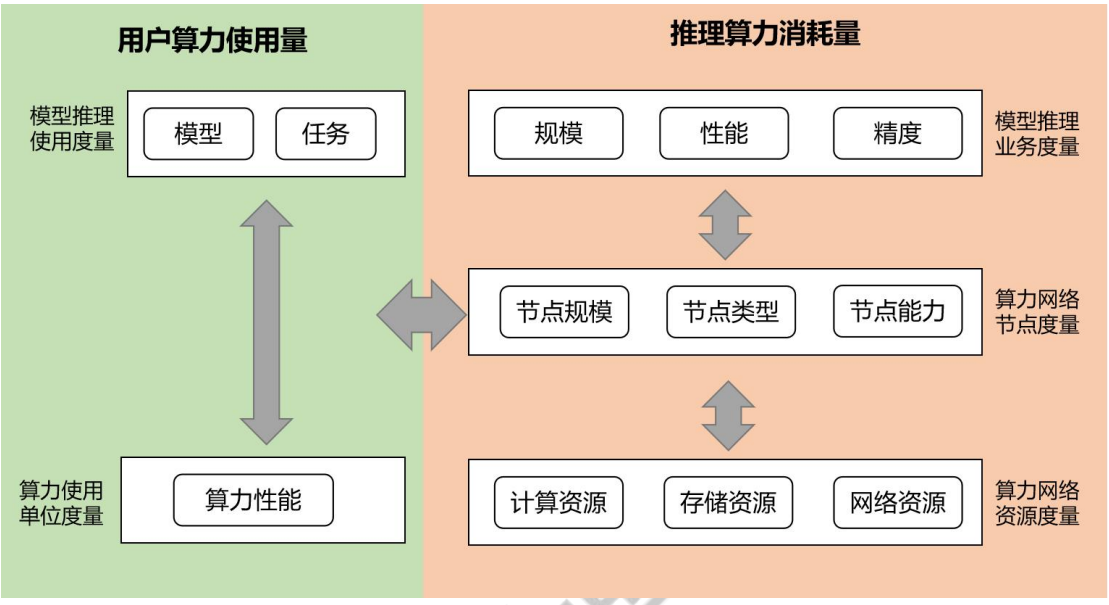

图 4 模型推理算力度量模型

上图显示了人工智能模型推理算力度量模型，分为算力网络算力消耗量度量和用户算力使用量度量两部分。算力网络算力消耗量度量通过分析模型推理执行时所需的算力资源总量，从而量化模型推理所需要的算力资源。算力网络用户算力使用量度量分析模型推理执行时所使用的算力资源总量。

- 算力消耗量度量模型。图 4 的右侧显示算力网络算力消耗量度量模型，从上到下分为模型推理业务度量、算力网络节点度量和算力网络资源度量。模型推理业务度量，描述算力网络中模型推理业务特征，包括参数量、数据精度等；算力网络节点度量，描述算力网络中模型推理执行时，节

点的计算、通信、存储的总量；算力网络资源度量，描述算力网络中模型推理执行时所需要算力资源的总量，包括计算资源、存储资源、网络资源。

- 算力使用量度量模型。图 4 左侧为算力使用量度量模型，从上到下分为 2 层：模型推理使用度量和单位算力使用度量。模型推理使用度量，从用户角度描述完成使用模型推理完成一定工作量所需要使用的算力；单位算力使用度量，从用户角度描述需要的处理速度和单位时间内使用的算力。

（二）算力消耗量度量方法

模型推理算力消耗量度量方法采用三层级架构，从业务需求出发，逐步分解为理论算力需求，最终映射为实际资源需求。该方法通过分层度量与转换，实现模型推理业务与算力资源的高效映射。

首先是模型推理业务度量，聚焦模型推理业务任务，采集模型规模（如参数数量、模型文件大小）、模型结构（如 CNN、Transformer）、性能（支持并发用户数、推理延迟要求）、数据精度（如 FP32/FP16）等核心需求，明确模型推理对算力度量的业务导向。

其次进行算力网络节点度量，基于业务度量结果，通过理论分析与推导，将业务需求转化为算力网络节点的量化指标，计算推理过程理论计算量（如浮点运算次数），统计所需存储量，分析节点内通信量，建立业务需求到节点算力消耗的关联。

最后实施算力网络资源度量，感知算力网络中计算资源、存储资

源、网络资源的种类与参数，基于节点度量的计算量、存储量、通信量，通过映射算法将理论需求转化为实际资源需求，最终形成模型推理业务到算力资源的映射。

（三）算力使用量度量方法

模型推理算力使用量度量方法将用户业务需求直接映射为可量化使用的算力，核心思路分为两个关键步骤：模型推理使用度量（明确用户需求量）和算力使用单位度量（定义单位度量值）。

模型推理使用度量聚焦用户需求采集，识别用户使用的模型类型（如 CNN、DeepSeek 等），明确对应业务场景（图像识别、问题解答等 ），并统计业务总量（如识别图像数量、问答请求数），构建包含模型类型、业务规模等的量化需求，最终清晰界定算力服务对象与任务量级。

算力使用单位度量围绕单位时间算力展开。基于模型推理业务，度量单位时间内业务处理量（如每秒识别图像张数、每分钟问答请求处理数 ），同步统计单位时间算力消耗量，最终建立 “模型推理业务量-单位时间业务处理速度-单位算力消耗量”的量化映射。

四、模型推理算力度量指标

（一）模型推理算力度量指标体系

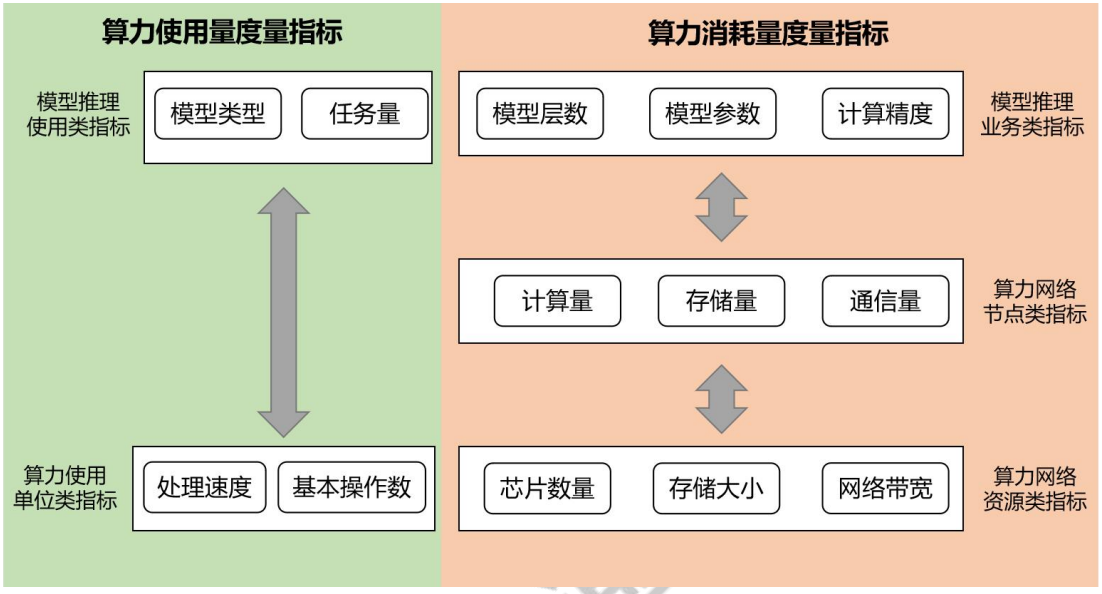

图 5 模型推理算力度量指标体系

基于模型推理算力度量模型构建模型推理算力度量指标体系，旨在全面、精准地量化模型推理过程中的算力消耗与使用情况，为算力资源调度、计量计费提供支撑。

该指标体系分为两大维度：算力消耗量度量指标和算力使用量度量指标。算力消耗量度量指标对应三层级架构（业务、节点、资源），从模型推理运行出发，逐步深入到节点及物理资源层面，实现从高层业务到底层资源的精准映射；算力使用量度量指标聚焦用户视角，量化用户使用模型推理时的算力需求量和使用量。

（二）算力消耗量度量指标

表 1 算力消耗量度量指标

类别	指标名称	单位	含义
模型推理业务类指标	参数数量	个	模型包含的参数总数
	模型大小	GB	模型存储文件的占用空间
	模型结构	无	模型的网络架构类型（如 CNN、Transformer）
	并发用户数	个	同时支持的最大用户数量
	推理延迟	ms	模型从接收输入到返回输出的时间
	首 Token 时间	ms	模型从接收输入到生成第一个 token 的时间
	精度	浮点/整形	模型计算时采用的数值精度标准
算力网络节点类指标	计算量	FLOP	节点完成计算所需的运算总数
	计算能力	FLOPS	峰值计算速率，比如 1PFLOPS
	通信量	GB	节点间传输数据的总容量
	通信速率	Gbps	节点内外通信带宽
	存储量	GB	节点存储的模型参数及中间结果的总容量
	存储速率	GB/s	节点存储读写速率
算力网络资源类指标	CPU 型号	无	CPU 的具体型号
	CPU 性能	TOPS	单个 CPU 的每秒整形运算能力
	CPU 数量	个	模型推理执行需要的 CPU 总数
	AI 芯片型号	无	人工智能加速芯片的具体型号（如 GPU、NPU）
	AI 芯片性能	FLOPS	单个 AI 芯片的每秒浮点运算能力
	AI 芯片数量	个	模型推理执行需要的 AI 芯片总数
	AI 芯片内存	GB	AI 芯片内存大小

类别	指标名称	单位	含义
	容量		
	AI 芯片内存速率	GB/s	AI 芯片每秒数据读写速度
	网络带宽	Gbps	模型推理执行时需要的网络带宽
	网络时延	ms	模型推理执行时能够容忍的时延
	拓扑结构	无	节点内外的网络结构（如星型、网状）
	网络协议	无	节点内外采用的协议（如 RDMA）
	存储容量	GB	模型推理需要硬盘的总容量
	存储速率	GB/s	模型推理需要硬盘的读写速度

（三）算力使用量度量指标

表 2 算力使用量指标汇总

类别	指标名称	单位	含义
模型推理使用类指标	模型类型	无	用户选择的模型推理，比如 ResNet，Yolo 等
	任务量	张/次/token	用户需要处理的总数量，比如照片、语音、文字
	目标精度	%	完成任务要求的精度，比如图像识别准确率 95%
算力使用单位类指标	处理速度	张/秒、token/分钟	单位时间处理数量
	基本操作数	无	每秒算力使用量，计算操作和数据移动操作的集合

五、模型推理算力度量关键技术

（一）模型剖析技术

模型剖析技术是深入理解模型运行机制，精准评估其推理算力需求的基础。通过对这些模型结构和参数的综合剖析，可以精确计算出模型推理运行时的关键数据。

在计算量方面，以矩阵乘法为例，深度学习模型中的大量运算都涉及到矩阵乘法操作。对于一个具有 m 个输入神经元和 n 个输出神经元的全连接层，其矩阵乘法的计算量为 2mn-n 次浮点运算。当模型包含多个这样的全连接层以及其他复杂的运算层时，通过对每层结构和参数的分析，可以累加得出整个模型的计算量。

在存储量方面，模型参数的存储是主要部分，不同的数据精度决定了每个参数占用的存储空间，结合模型的参数量就能计算出模型参数的存储需求。同时，模型运行过程中产生的中间结果，如激活值等，也需要占用一定的存储空间，通过对模型结构和运算流程的分析，可以估算出这些中间结果的存储量。

通信量方面，在分布式推理场景中，不同计算节点之间需要传输模型参数、中间结果等数据，通过分析模型的并行策略和数据传输需求，可以计算出节点间的通信量。例如，在模型并行中，不同节点负责处理模型的不同部分，节点间需要频繁传输中间结果以完成整个模型的推理，通过对模型结构和并行划分的分析，能够确定数据传输的

规模和频率，从而计算出通信量。

（二）模型推理并行技术

模型推理加速中并行技术发挥着至关重要的作用，通过将推理任务分解并同时执行，显著提升推理效率。常见的并行技术包括数据并行、模型并行和流水并行，它们各自从不同角度对推理过程进行优化。

数据并行是将输入数据分成多个部分，分别在不同的计算单元上进行处理，最后将结果汇总。以一个简单的图像分类模型为例，假设有一批包含 100 张图片的输入数据，采用数据并行技术，可将这 100 张图片平均分成 4 份，每份 25 张图片，分别由 4 个 GPU 进行处理，模型的参数在每个 GPU 上都有完整的副本。

模型并行则是将模型的不同部分分配到不同的 GPU 上进行计算。对于一些大型的神经网络模型，如具有多层结构的 Transformer 模型，其参数规模巨大，单个 GPU 可能无法容纳整个模型，模型并行技术将 Transformer 模型的不同层分配到不同的 GPU 上。具体分为如下方式。

- 张量并行旨在将模型的单个层分配到不同的 GPU 上，每个 GPU 处理一部分模型参数，然后通过 GPU 间通信将输出结果进行合成。张量并行适用于模型参数非常大的情况。
- 流水线并行旨在将模型的不同层分配到多个 GPU 上，在 Transformer 模型中，将连续的层加载到同一 GPU 上，以减少在 GPU 之间传输隐藏状态。流水线并行适用于模型结构复杂且计算密集的情况。

- 专家并行是一种针对混合专家模型(MoE)的高效分布式计算技术，将 MoE 模型中的不同专家分配到不同 GPU 上，每个 GPU 仅负责部分专家的计算，动态分配输入数据到相关专家，仅激活部分专家进行计算，显著降低计算负载。

（三）基本操作数测量技术

BOP 的获取有两种主要途径，分别为分析代码获取和使用 CPU 等设备的性能计数器机制获取。

其中分析代码的方法需要分析代码逻辑并在相应位置插入计算代码，这种方法的优点是准确但是需要源代码。

性能计数器是特殊的硬件寄存器，在大多数现代 CPU/加速设备上都可以使用且不会降低应用程序的速度，实现非入侵分析。对于 CPU 平台，指令可以分为 Load 指令、Store 指令、分支指令、整型指令、浮点指令五大类。因此通过 CPU 硬件计数器获取总指令数（$\mathrm{ins}_{\mathrm{total}}$），分支指令($\mathrm{ins}_{\mathrm{branch}}$)，Load 指令（$\mathrm{ins}_{\mathrm{load}}$），Store 指令($\mathrm{ins}_{\mathrm{store}}$)，使用总指令减去分支指令、Load 指令、Store 指令可得到所有的浮点和整型操作指令，从而获取 BOPs 对应的指令（整型、浮点）数量。这种方法存在体系结构相关的问题且 BOPs 属于近似估计，存在一定程度的误差。为此，我们在 Intel X86_64 平台进行了多次多应用的对比实验，实验结果显示采用硬件计数器方法获取的 BOPs 和应用实际的 BOPs 之间的误差可以控制在 8%以内。同样原理，CPU 平台的获取方法可以映射到加速设备的度量。如 Nvidia 的 GPU

设备可利用 Nvidia 在 CUDA 工具包中提供的 nvprof 工具进行算力消耗的度量。

$$\mathrm{BOPs} \cong \mathrm{ins}_{\mathrm{total}} - \mathrm{ins}_{\mathrm{load}} - \mathrm{ins}_{\mathrm{store}} - \mathrm{ins}_{\mathrm{branch}}$$

六、模型推理算力度量案例

本章结合典型 AI 模型推理场景，从算力消耗量度量、算力使用量度量及商业化计费三个维度，通过实际案例展示算力度量在资源调度、成本核算及服务定价中的应用价值。

（一）模型推理算力消耗量度量案例

算力消耗量度量的核心价值在于实现业务需求与算力资源的精准映射，为资源配置与调度提供量化依据，以下结合计算机视觉与自然语言处理领域的典型模型进行具体说明。

- ResNet50 模型推理案例

ResNet50 作为深度学习领域广泛应用的卷积神经网络架构，其核心通过残差模块与跳跃连接解决深层网络退化问题，适用于图像分类、目标检测等计算机视觉任务。基于算力消耗量三层级度量方法其度量结果如下。

模型推理的业务类指标：参数量约 2560 万，层数 50 层，数据精度 FP16。推理延迟要求为 30ms，并发用户数 10，最大处理能力 QPS = 1000ms ／ 30ms × 10 并发≈333 推理/秒

算力网络节点类指标：单次推理计算量≈4.12 GFLOP，总的计算量≈1.37 TFLOP，计算性能为 4.12 GFLOP÷0.03 s≈137.33 GFLOPS。ResNet50 的参数存储量约为 25.6 MB，在 FP16 数据精度下，每个参数占用 2 字节存储空间，所以存储该模型需要的空间为 51.2 MB。因

采用单机推理模式，节点间无额外通信需求。

算力网络资源类指标：若采用单核心性能为 100 GFLOPS 的 CPU，所需核心数为 1.37 TFLOP÷（100 GFLOPS*1 s）≈ 13.7 核。存储资源需同时考虑模型参数与中间结果，按中间数据量为模型大小 2 倍计算，总存储需求为 153.6MB。

- DeepSeek R1 模型推理案例

DeepSeek R1 是面向长文本场景的大语言推理模型，支持 128K 输入上下文长度，采用多头潜在注意力（MLA）与混合专家（MoE）机制，适用于复杂问答、长文本生成等任务。其算力消耗量度量结果如下。

模型推理的业务类指标：参数量约 6710 亿个，层数 61 层，数据精度 FP16。推理延迟要求为 30ms，并发用户数 100，QPS = 1000ms / 30ms × 100 并发≈3333 推理/秒，平均输入长度 150 token，平均输出长度 50 token。

算力网络节点类指标：MoE 架构下仅激活 5.5%参数，激活的参数量为≈370 亿，单次推理计算量 2×370 亿×200token≈14.8 TFLOP；总计算量 14.8 TFLOP×3333 ≈ 49.26 PFLOP；计算性能 14.8 TFLOP ÷0.03s≈493.3 TFLOPS；模型存储空间 6710 亿×2 字节=1342 GB；在张量并行时 GPU 之间通信带宽为 200 token×7168 维×2 字节×3333 推理/秒≈ 9.56GB/s。

算力资源类指标：NVIDIA H100 PCIe 性能为 FP16 计算 1513

TFLOPS,内存80GB,内存带宽2TB/s,卡间互联带宽PCIe5.0下128GB/s和NVLINK下为600GB/s。从计算角度需要的GPU数量49.3 PFLOP÷1513 TFLOPS≈32.6块，取整为33块；从内存角度需要的GPU数量1342 GB÷80GB≈16.78块，取整为17块，即单台服务器可以满足要求。GPU之间的带宽能满足通信要求。

（二）模型推理算力使用量度量案例

模型推理算力使用量是算力度量在用户层面的直接应用，其核心是通过基本操作数（BOP）量化用户的实际算力使用。基本上BOP包括浮点/整型计算操作（体现计算操作）和地址操作（体现数据的移动操作）。

BOP的分析代码的方法需要分析代码逻辑并在相应位置插入计算代码，这种方法的优点是准确但是需要源代码。比如：Sort应用采用C++语言实现指定规模（10E8）的整型数组的快速排序，Sort应用的整型计算数量为142×109，地址计算量为387×109，浮点计算数量趋于0,可知Sort应用的BOP为529 GBOP。Stream应用源自内存带宽测试的微基准实现浮点数组中元素的复制和简单数学运算，其基本操作数为144 GBOP。MatMul应用完成浮点矩阵的分块乘运算，其基本操作数为681 GBOP。

（三）联通云计量计费案例

模型推理算力度量为人工智能服务的商业化定价提供了可量化

的成本依据，使服务定价更具透明度与合理性。

中国联通旗下联通云在大模型 AI 推理服务领域已经构建了一个全面且完善的支撑体系，充分利用其独特的“云网融合”优势以及全国一体化的算力布局，为用户提供从大模型部署、推理到应用落地的一站式解决方案。这一举措不仅加速了 AI 技术的应用推广，也大大增强了各行各业对复杂数据分析和智能决策的能力。目前，联通云将 AI 推理服务封装成标准化的云产品，其中 V1 版本已经在联通行业云、政企创新头条以及私有云等多个平台成功上线。这款云产品支持创建对话（Chat）功能，使用户能够轻松发起交互，实现与大模型的直接沟通，无论是进行文本生成、问答还是编程等任务，都能高效完成。此外，该版本还提供了 API Key 管理功能，允许用户创建 API Key，并将其与具体的服务关联起来，确保数据和服务的安全性和灵活性。尤为值得一提的是，联通云的 AI 推理服务已集成了包括 DeepSeek、Qwen、Meta-Llama 等在内的多种主流大模型，覆盖近 20 个不同的版本。这种广泛的模型兼容性使得用户可以根据自己的具体需求选择最适合的大模型，享受定制化的 AI 服务。不论是需要高度精准的语言处理能力，还是复杂的图像识别任务，联通云都能提供强有力的技术支持和服务保障。

对于 AI 推理服务提供的几款主流大模型，提供按量计费与 Token 资源包两种模式：按量计费是一种基于实际使用量的后付费模式，用户按需付费，无需预付。其特点是灵活性高、成本透明，可有效避免

资源浪费，特别适合需求波动大或处于测试开发阶段的用户，便于以低成本快速验证和迭代 AI 应用。Token 资源包是一种面向高频或稳定使用场景的预付费模式，用户可提前购买一定数量的 Tokens，享受比按量计费更优惠的价格，有效控制成本并简化预算管理，适合对 AI 服务有持续调用需求的企业或开发者，同时 Tokens 通常设有使用有效期，过期未用部分将失效。用户在使用推理服务时产生的 Tokens 用量，将由计费系统按小时进行统计汇总，并遵循“先到期先抵扣”原则，优先扣除已购量包中的额度。当量包耗尽或过期后，系统将自动切换为按量付费模式，确保服务连续性。无论是选择 Token 资源包还是按量计费，这两种计费方式都旨在满足不同类型用户的个性化需求，并确保他们能够以最经济有效的方式利用强大的 AI 推理服务。

七、总结

本报告围绕算力网络中人工智能模型推理的算力度量展开系统性研究，旨在解决模型推理算力需求激增背景下的算力资源精准评估的问题，通过理论构建、技术剖析与案例验证进行体系化描述。

首先，明确了算力网络人工智能模型推理算力度量的核心概念，即通过量化评估推理任务的算力资源需求，为模型部署、算力调度与交易提供决策支撑。然后，构建了“算力消耗量”“算力使用量”双维度度量模型。在度量模型的基础上，建立了包含模型参数、计算量、存储需求、处理速度等在内的多维度指标体系。最后，结合 ResNet50、DeepSeek R1 等典型模型案例，验证了度量方法的可行性与有效性。

未来，随着 AI 模型的持续迭代和算力网络的继续演进，需进一步完善算力度量技术，推动算力网络与人工智能的深度融合。

参考文献

1. 中国联通研究院，中国联通算力网络白皮书，2019
2. 中国联通研究院，算力网络架构与技术体系白皮书，2020
3. 中国联通研究院. 算力网络可编程服务(SIDaaS)白皮书，2022
4. 中信建投，AI 新纪元：砥砺开疆·智火燎原，2025
5. CCSA，YD/T 6044-2024，算力网络 算力度量与算力建模技术要求
6. CCSA，YD/T 4255-2023，算力网络 总体技术要求
7. 祝淑琼，徐青青，李小涛，等. 算力度量与任务调度：物联网端侧设备策略研究. 电信科学，2024，40(3)：83-91.
8. 杜宗鹏，李志强，陆璐，等. 算力网络四面三级算力度量技术体系. 中兴通讯技术，2023，29(4)：37-43.
9. 乔楚. 算力度量与算网资源调度思路分析. 通信技术，2022，55(9)：1085-1090.
10. 王磊，孙凝晖. BOPs：一种算力度量指标. 中国计算机学会通讯，2024(1)：40-45.
11. 冯汉枣，黎元宝，刘运奇. 异构混合云服务下的多任务算力度量方法. 计算技术与自动化，2023，42(4)：101-106.
12. 姜海洋，李勇. 端边云场景下的算力度量方法. 电信工程技术与标准化，2023，36(7)：60-65.
13. 李一男，唐琴琴，彭开来，等. 以服务为中心的算力网络度量与建模研究. 信息通信技术与政策，2023(5)：52-57.

14. 国家人工智能标准化总体组，全国信标委人工智能分委会. 计算中心有效算力评测体系白皮书（2022）. 2022.

15. CCSA YD/T 6044-2024. 算力网络 算力度量与算力建模技术要求. 2024.

16. CCSA 2023-0191T-YD. 算力网络异构算力资源计算能力度量指标. 2023

17. CCSA 2023-0190T-YD. 算力网络算力节点能力度量及评估方法. 2023

18. CCSA 2022-0755T-YD 面向公共通信业务体验的算力量化与建模技术要求. 2022

中国联通研究院根植于联通集团（中国联通直属二级机构），作为中国联通科技创新专业子公司，自2022年起，挂牌成立中国网络安全研究院、下一代互联网宽带业务应用国家工程研究中心及首个国家级网络安全产业知识产权运营中心，形成“两院两中心”发展格局，开创了研究院高质量发展的新局面。

中国联通研究院作为服务于国家战略、行业发展、企业生产的战略决策参谋者、技术发展的引领者、产业发展的助推者，坚持以高水平科技自立自强为使命担当，聚焦网络强国、数字中国主责，拓展联网通信、算网数智业务，形成“态度、速度、气度、有情怀、有格局、有担当”的企业文化，以下一代互联网、光网络、5G-A/6G、网络安全、数智网优、低空智能网联和新型智库研究七个领域为主攻方向，坚持“四个聚焦”，开展关键核心技术攻关、科创力量建设、专业技术人才队伍建设、创新成果转化等工作，着力实现价值创造提升、战略科技力量提升、专精特新能力成色提升，争做通信行业科技创新主力军，努力建设成为“国家信赖、行业领先、集团倚重、员工自豪”现代化一流研究院。

战略决策的参谋者
技术发展的引领者
产业发展的助推者

态度、速度、气度

有情怀、有格局、有担当

中国联合网络通信有限公司研究院
地址：北京市亦庄经济技术开发区北环东路1号
电话：010-87926100
邮编：100176