# 分布式推理网络(DIN)技术白皮书

（2025 年）

中国移动通信有限公司研究院

# 前　言

2025 年 1 月 20 日，深度求索（DeepSeek）公司自主研发的国产大模型 DeepSeek-R1 模型以极低成本实现了与国际顶尖 AI 模型相当的性能，凭借出色的性能和易用性快速扩张。随之而来的海量并发用户访问请求，造成服务器资源迅速耗尽，DeepSeek 多次出现网页和 API 无法访问的情况，用户在使用其服务时，频繁收到“服务器繁忙，请稍后再试”的提示。同时，DeepSeek 的火爆出圈也吸引了攻击者使用多种网络攻击技术和手段持续进行攻击。随着普惠 AI 推理时代的到来，需要考虑以 AI 模型和 AI 推理为中心构筑互联网，网络也将面临新的网络流量模式的变化。实现普惠 AI 和 AI 推理大规模应用面临 AI 推理基础设施能力不足，AI 推理网络技术待完善，AI 服务网络安全防护能力待提升等挑战。

中国移动提出面向普惠 AI 服务的新型分布式推理网络（Distributed Inference Network，DIN），融合运营商网络协议可编程和流量感知调度能力的优势，支撑中心、边缘或边云协同部署等多种分布式方式的推理架构，挖掘算网一体安全推理、边云协同后训练、模型分层协同、大小模型协同、训推协同进化、PD 分离协同等多种端边云协同模式，研究推理网络保障技术、推理服务调度技术、推理安全防护技术等关键技术，解决大模型集中化部署模式下的大规模并发推理能力不足的问题，构筑多维度安全能力，从而有效应对亿级海量用户并发推理挑战并实现安全高效的 AI 推理服务。

未来中国移动在分布式推理网络方面，将联合产业界重点拓展应用场景，构建融合端、边、网、算的 DIN 技术体系，解决 AI 推理在个人（ToC）、家庭（ToH）及企业（ToB）应用中的成本、效率、安全与场景适配难题，加速全社会普惠 AI 时代到来。

# 目 录

# 1. 业务发展趋势与挑战

## 1.1 AI 大模型发展趋势

2025 年 1 月 20 日，深度求索公司自主研发的DeepSeek-R1 模型震惊世界，以极低的成本实现了与国际顶尖AI模型相当的性能。人工智能大模型技术的飞速发展，正在深刻改变人类社会的生产生活方式，对物理世界、虚拟世界和生命世界带来全方位的影响，加速人类社会从信息社会向智能社会演进。当前出现两个重要趋势：

**趋势一：AI普及速度显著加快，推理成本迅速降低，用户从访问内容向访问AI模型转变。**DeepSeek-R1 大模型的表现达到了行业领先水平，推理速度提升 4 倍，API调用成本仅为GPT-4-Turbo的近百分之一。从DeepSeek发布后不到一个月的时间，日活用户量DAU也在短短一个月的时间内从 100 万迅速突破 3000 万，增长速度刷新了行业纪录。据不完全统计，国内外已有 50+企业宣布接入DeepSeek，涉及网络安全、汽车、智能硬件、金融行业、芯片制造、云服务提供商等各行业，通过与应用深度集成，AI大模型正在从聊天工具向生产生活工具演进，预计会形成不可逆的新业务场景。

**趋势二：AI Agent无处不在，Agent之间的通信会显著增长。**普惠AI推理进一步推动AI智能体需求爆发，逐步演进为具备更高自主性和协作能力的泛在多智能体系统，如Manus发布多AI Agent协作效果视频，OpenAI推出的Operator智能体展现自主执行多任务能力。为协同完成如供应链管理、金融等复杂任务和高效决策，Agent之间会产生大量去中心化的、高度实时、安全敏感的通信流量。

## 1.2 AI 时代网络流量变化

云计算时代以云为中心构建互联网，互联网流量增加东西向流量承载，发展出SRv6 等网络技术。随着AI推理时代的到来，大量应用、IoT设备以及未来AI智能体等交互式访问AI推理服务，以及AI模型分发模式的变化、AI训推一体化等技术的发展，需要重新考虑以AI模型和AI推理为中心构筑互联网，网络将面临新的网络流量流向和流量模式变化，需要应对普惠AI时代新的业务模式，发展新的网

络技术。

**端云多模态交互带来南北流量持续增长。**随着AI智能体、智能终端的普及，以及AI推理服务的开放化和普惠化，多模态交互需求和云端推理需求激增，用户侧与云侧南北向流量将快速增长。根据预测，到 2030 年，仅在中国市场AI Token引发的日均网络流量将达到 500TB，是当前全国移动网络日均总流量的 5.5 倍。

**Agent间交互带来东西向流量持续增长。**随着AI技术的不断进步，多Agent系统、AI模型的分布式训练和推理架构逐渐成为研究和应用的热点，这些系统由多个具有自主决策和交互能力的Agent组成，他们之间需要进行频繁的通信和协作以完成复杂任务。如在金融风控系统中，信用评估Agent需与反欺诈Agent实时交换数据，游戏场景中更强的AI NPC智能体通过Multi-Agent架构实现动态交互。这种多智能体间的交互会产生大量的东西向流量。

**复杂推理任务对时延提出新要求。**在运行复杂推理任务时，往往需要多步骤交互，在通算、智算、存储等系统之间形成高频率流水线调用，对时延提出严苛要求。例如，当用户向一个AI智能体发出“规划一次旅行并预订相关服务”的指令时，AI 智能体首先需要通过通信网络获取用户位置、偏好等信息，然后调用智算资源进行数据分析与旅行方案规划，进一步通过通信网络与各旅游服务供应商的Agent系统交互来完成预订操作。在这一系列过程中，如果时延过高，用户体验将受到极大影响，可能导致用户放弃使用该服务。

## 1.3 AI 普惠时代面临的挑战

DeepSeek访问量峰值超过 4900 万次/日，海量的并发用户访问请求，造成服务器资源迅速耗尽，DeepSeek多次出现网页和API无法访问的情况，用户使用时经常遇到“服务器繁忙，请稍后再试”的问题。2 月 6 日DeepSeek官方宣布，由于服务器资源紧张，已暂时停止API服务的充值功能。

DeepSeek的火爆出圈也吸引了攻击者使用多种攻击技术和手段，持续进行攻击。2025 年 1 月 28 日，DeepSeek发布公告称，其线上服务遭遇大规模恶意攻击，导致平台注册繁忙。网络安全公司奇安信通过监测发现，攻击从最初的放大攻击演变为更难防御的HTTP代理攻击，并在 1 月 30 日凌晨升级为由僵尸网络主导的攻击，攻击烈度相比 1 月 28 日暴增上百倍。

结合我国信息通信行业发展、DeepSeek的情况以及未来智能体等交互式AI推理服务来的流量模型变化等可以看到，要实现AI普惠面临 3 大挑战：

**挑战一：AI推理基础设施能力不足。**我国是全球最大的移动和宽带网络规模的国家。全国移动电话用户、宽带接入用户以及移动物联网用户数分别为 17.9 亿、6.7 亿和 26.56 亿。当十亿级体量的访问并发时，由于AI服务器和网络资源不足，将导致大量推理服务请求失败。

**挑战二：AI推理网络架构及技术待完善。**为解决十亿级用户/API对推理大模型访问问题，需要在企业/个人用户/IoT设备/智能体与算力资源之间，构建一张高效并发访问AI服务的分布式AI推理网络，支持可靠、稳定的确定性体验保障能力，实现“智能无处不在”的发展愿景。

**挑战三：AI服务网络安全防护能力待提升。**针对AI平台、AI模型、AI应用及模型参数、KV Cache等数据在分布式链路传递过程中的安全威胁日益凸显，包括数据安全与隐私泄露、数据篡改、资源滥用与恶意攻击、模型窃取与训练任务中断等风险。尤其是在DDoS攻击频发的情况下，须抵御大规模有组织网络攻击。

# 2. 推理业务服务模式及网络需求

AI 普惠时代激发 2B、2C、2H 领域衍生多样化推理服务范式，需要云边端协同一体实现 AI 推理的低时延、高精度以及泛在接入服务，对网络能力提出了前所未有的挑战。

## 2.1 ToB 推理服务

### 2.1.1 算网一体安全推理服务

面向政务、金融、医疗等对数据安全性要求极高的行业，例如政务行业涉及大量社会经济数据、人口数据、公共安全数据通过模型推理支撑发展战略决策依据，金融行业通过金融大模型实现信贷风险评估与反欺诈分析，医疗行业对影像数据、病历数据等进行分析辅助医生诊断，以及病房、手术室、ICU 等区域开展感染风险防控，需提供数据安全访问能力，防止数据泄漏。

考虑到设备和运维成本、计算资源共享、数据集中管控等因素，在部门/企业总部部署推理一体机/服务器，分支机构可通过互联网、专线等方式安全访问推理服务。由于推理服务涉及上述场景中安全隐私性要求极高的数据，为了防止数据泄露，需要提供分支到总部的高安全互联访问通道，特别是通过互联网访问时需要对通道进行加密。中国移动原创融合 Overlay 和 Underlay 网络的智享 WAN 产品，可通过基于 SRv6 和 IPSec 协议的隧道安全加密技术，快速构建数据安全的模型推理接入网络，打造“推理一体机/服务器+智享 WAN”的安全推理产品，提供算网一体安全推理服务。

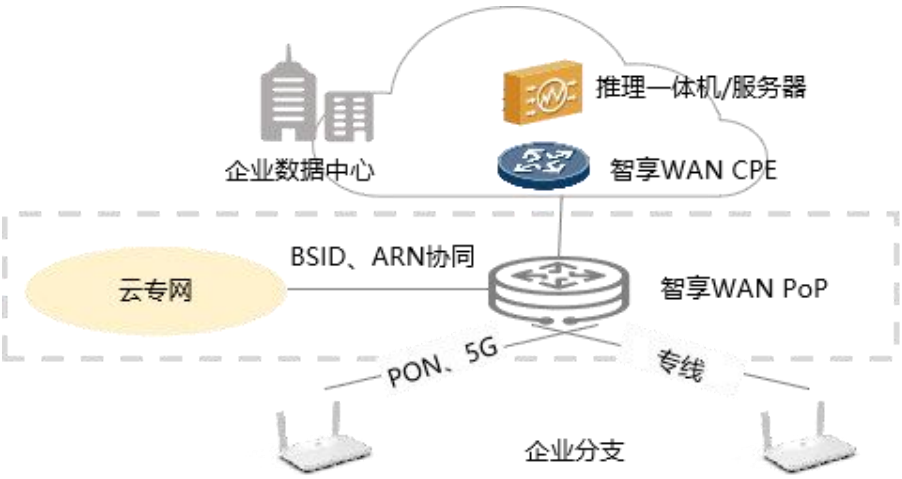


图 1 算网一体安全推理方案

## 2.1.2 边云协同后训练

本地部署算力集群进行模型后训练微调成本过高，为降低企业本地部署成本，可通过以租代买，租用部分云端算力，与本地算力跨广域协同完成模型后训练（Post Training）微调。同时，为保证企业私有数据安全，可将模型的前几层和最后几层部署在边缘节点上，其它层部署在中心节点，原始样本仅存储在园区内；云边间通信采用流水线并行，基于 RDMA 协议通过参数面进行通信；利用计算对通信的掩盖，在有限的训练算效损失(<5%)下完成云边协同后训练。

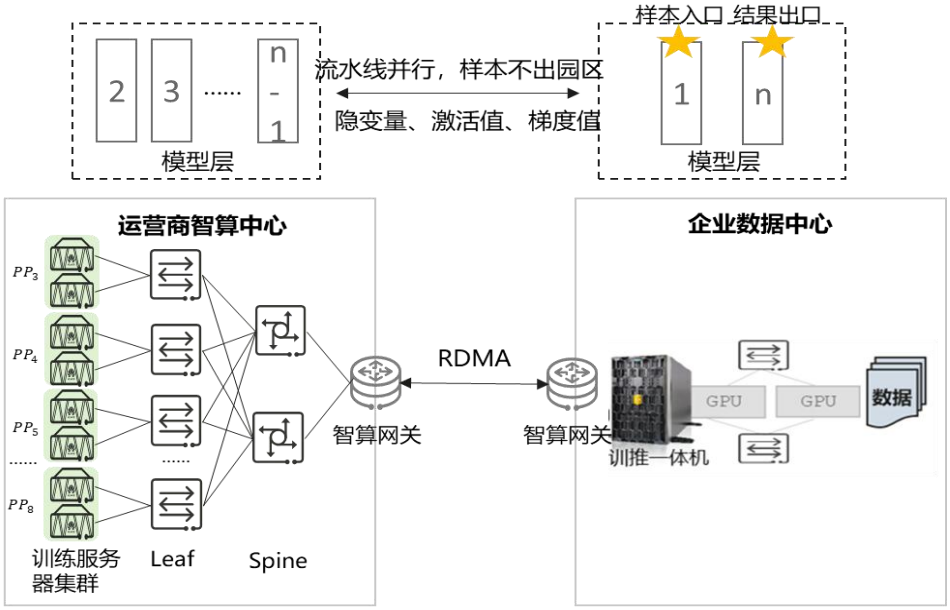


图 2 边云协同后训练（PostTraining）方案

以 Qwen72B 微调为例，纯本地部署需建立 128 卡集群，一次性投入建设成本对某些企业会难以承受；而采用以租代买的边云协同后训练部署，仅需本地算力 8 卡+云侧租用算力 120 卡，将显著降低综合成本。云端算力资源支持按需弹性扩缩容，根据实际使用时长计费，可有效避免闲时浪费。同时，微调过程中企业原始数据将始终保留在私有域，满足其数据安全要求。

在模型训练过程中，前向计算需在云边间传输中间计算结果隐变量，反向计算则需传输反向误差梯度。实测显示，即使 $10^{-5}$ 的低丢包率，也会造成协同训练算效损失超 8%。因此，网络必须具备 RDMA 高效传输能力，将丢包率严格控制在 $10^{-9}$ 以下。此外，流水线并行训练的效率保障依赖于减少传输气泡、以计算掩盖通信。研究表明，数十毫秒级时延或十毫秒级抖动，会使协同训练吞吐量下降超 50%。为满足上述严苛要求，网络基础设施需运用微流级流控、层次化细粒度切片、PHYSec 等核心技术，打造超大带宽、超低时延、丢包率达 $10^{-9}$ 的高安全网络环境。

### 2.1.3 模型分层协同

高质量推理模型（如 Deepseek、Qwen-2.5 系列等）促进企业推理业务的迅速增长，为企业带来了本地自建算力集群物理扩容难题：（1）频繁扩容，硬装周期长；（2）自建机房能力限制，不支持液冷服务器部署；（3）机房电力受限，电路改造成本高。通过以租代买，边云分布式协同推理服务部署，可支持企业弹性扩容、按需分时租用诉求，降低综合成本。同时为满足企业数据安全诉求，保障推理过程中原始样本数据与输出结果不出园区，可采用 Split Learning 技术对模型进行切分，将模型的前几层与后几层部署在企业边缘节点上，其他层则部署在云端，实现数据隐私保障的边云协同推理。同时，Split Learning 技术可进一步结合以 KV Cache 为中心调度的 PD 分离式架构，通过合理的 Prefill、Decode 阶段算力配比，适配推理服务实时负载，最大化算力&显存资源利用率，提升系统吞吐。例如将所有 Prefill 模型以及 Decode 模型的前几层与最后几层部署在边缘，将 Decode 模型的其他层部署在云端，边缘支持 KV Cache 复用管理降低 Prefill 算力开销，云侧弹性扩展 Decode 算力资源支持大并发。

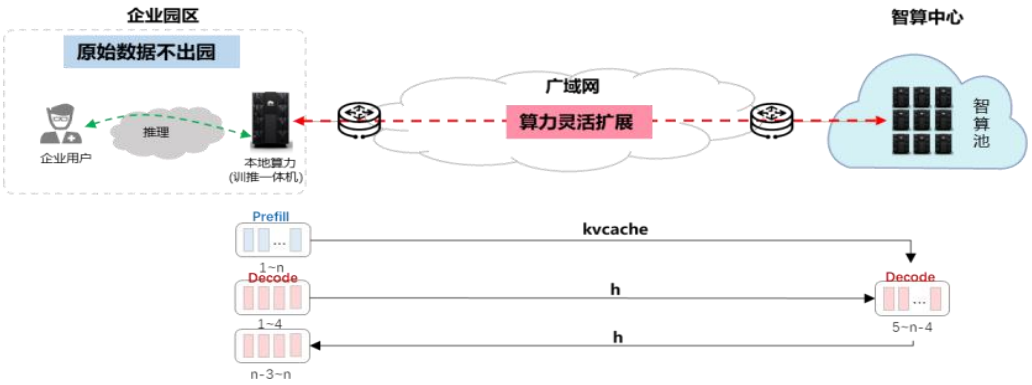


图 3 模型分布式推理方案

以实际应用为例，在金融风控与客服应用中，银行利用隐私计算能力，在本地部署金融大模型，实现信贷风险评估与反欺诈分析，确保企业数据不出域。同时，AI 客服系统通过动态调度技术，在高峰期自动扩展云端计算资源，服务响应效率提升 50%。

对于模型分布式推理方案，云、边之间通过隐变量进行协同，隐变量通常从几 KB 到几十 KB，为保障 Token 间极低时延（Time-Between-Tokens, TBT）和亚毫秒级抖动的指标 SLA，网络基础设施需具备超低时延、零丢包、极低抖动特性，通过确定性切片、微流级流控等技术手段，构建端到端低延迟通信链路，保障模型分布式推理的实时性与连贯性。

## 2.2 ToC/ToH 协同推理服务

### 2.2.1 大小模型协同

大模型与小模型的协同工作可以兼顾推理精度和效率，并提升系统的智能弹性。在工业场景中，面向具体行业的工业大模型具备强大的通用认知和推理能力，但推理开销较大；小模型则通常是针对特定任务精心优化的轻量模型，速度快且专用性强。

大小模型协同策略包括以下几种：一是能力调度，将大模型作为大脑统筹，小模型作为工具。在复杂业务场景中，大模型可以根据任务需要调用相应的小模型来执行专业子任务。例如在一个智能巡检应用中，大模型负责理解巡检报告的语义，并决定何时调用瑕疵检测的小模型来识别具体缺陷，这种大模型调度小模

型能力的方式实现了通用智能与专业智能的结合。二是推理加速，即利用小模型为大模型提速。由于小模型在边缘侧推理延迟很低，可让小模型先行给出初步结果，再由云端大模型对结果进行校正提高精度。三是场景自适应策略，根据应用场景和资源条件在大小模型间动态切换。例如在网络良好且对精度要求极高时，可将请求转发至云端的大模型处理；而在边缘需要即时响应时，则优先采用本地小模型完成推理，待联机后再由大模型异步验证或改进结果。

网络边缘部署小模型需保障大小模型协同流畅，通过网络优化加速推理。可通过网络设备内置异构 AI 算力板卡（算力达几 TFlops 到几百 TFlops），集成 NPU、NP 和 CPU 形成“计算 - 传输 - 控制”闭环，其中 NPU 提供 AI 计算能力，NP 构建高速数据平面，CPU 负责模型管理与智能服务。推动网络设备从单纯转发走向“连接→感知→推理”的能力跃迁，结合传统路由器协议可编程和流量感知调度能力，进一步提供差异化网络及智能应用服务。

### 2.2.2 训推协同进化

云端模型通过蒸馏得到多个小模型并部署到不同边缘节点，各边缘节点使用小模型进行推理，提供低时延实时响应。但由于小模型泛化能力较弱，当场景发生变化时推理精度会出现骤降，此时边缘小模型可将推理过程收集的高价值数据，如不确定性较高的推理结果等，反馈给中心大模型进行训练调整优化。大模型基于边缘小模型上传的数据，重新进行知识蒸馏，将蒸馏后获取的小模型参数更新到相应边缘节点。

例如，在家庭具身智能应用中，机器人凭借其感知、互动、行动和学习能力，为家庭成员提供家务清洁、儿童陪伴、老人陪护、家居管理等多种服务；并且机器人需根据家庭环境、生活习惯等具体场景进行动态调整适配。为具有更好的泛化能力，机器人在提供家庭服务过程中可通过采用分级过滤方式获取不确定性数据(从案例到 token 粒度)，并上传云端；云端则基于不确定性数据，采用基于适配器的知识蒸馏，微调跨模态转化和大模型的适配器，下发小模型参数更新。

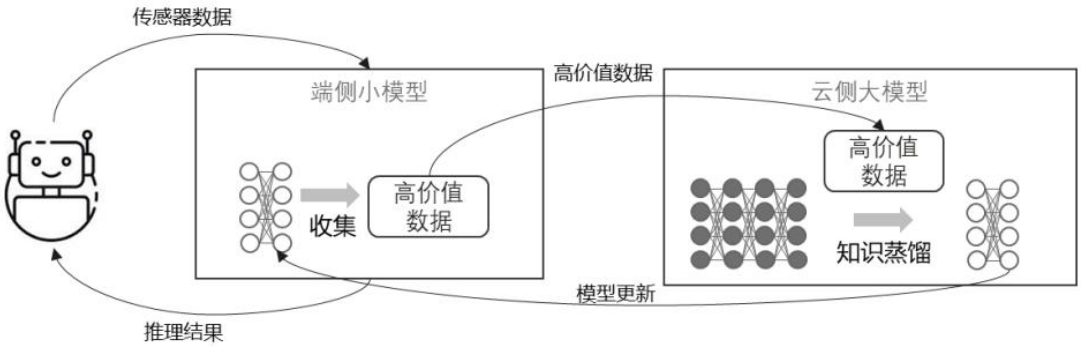


图 4“中训边推”协同进化过程

上述“中训边推”协同进化过程中，边缘节点向云侧上传收集的样本数据，云侧下发模型参数。一方面高价值样本数据可能涉及用户敏感信息，另一方面模型参数下发过程中可能被窃取或攻击篡改，影响推理服务正常运行。因此，需要对云边交互通道进行加密，使用例如 PHYSec 等技术，通过智享 WAN 新平面，快速构建数据安全的模型分发接入网络。

### 2.2.3 PD 分离协同

对于延迟敏感型推理服务，如家庭具身智能、车路云协同等，在集中部署时，云上推理输出返回时会受到广域网时延抖动影响，导致端侧服务无法接收稳定的指令，影响体验甚至产生事故。可将计算密集型的 Prefill 任务放在中心节点执行，尝试将访存密集型的 Decode 任务部署在边缘节点上。用户发出的推理问题在中心节点执行 Prefill 计算并将计算得到的 KV Cache 下发到边缘节点，边缘节点可以通过扩展存储用于保存 KV Cache 数据，存储单用户的对话上下文信息和跨用户的高热度信息，有助于降低后续对话计算代价、提升精度，提升边缘节点的推理服务体验；边缘节点继续执行 Decode 并将结果以流式方式发送给用户。

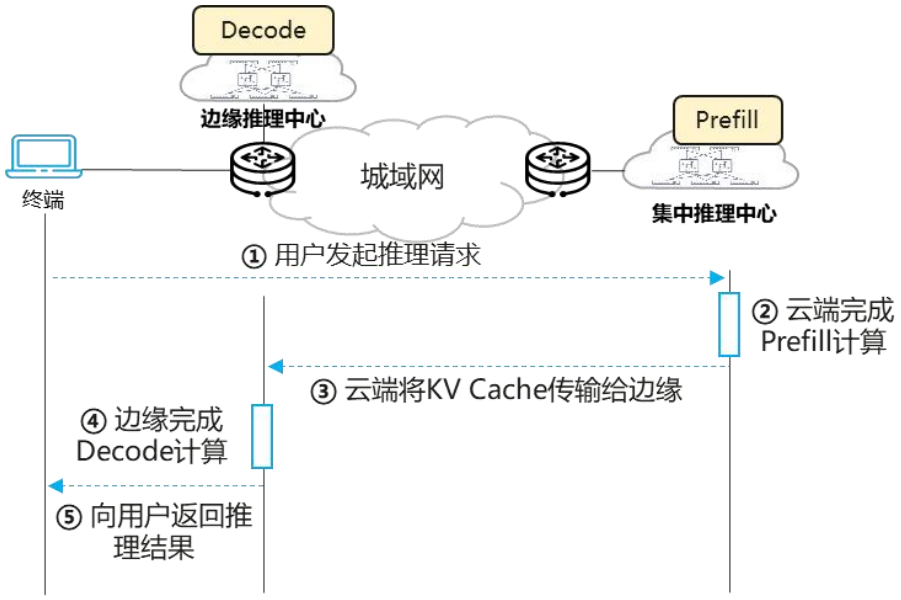


图 5 PD 分离协同推理

对于 PD 分离协同方案，云、边之间通过 KV Cache 进行协同，KV Cache 通常从 MB 到 GB 级，为了保障首 token 的时延要求，需要在百毫秒内完成传输，因此广域网需要具备稳定的大带宽。由于 PD 间采用 RDMA 协议，丢包会破坏计算传输流水，导致计算长时间等待，经实测 $10^{-5}$ 丢包率即导致推理服务中断，因此广域网需具备高带宽、低丢包承载能力。

# 3. 分布式推理网络（DIN）架构及设计目标

随着 AI 普惠时代的到来，个人、家庭、企业用户与智能之间的连接会越来越紧密。流量模型包括访问推理服务的南北向流量及边云协同的东西向流量，用户多模态推理服务访问导致南北向流量模型向高安全、高频交互、快速膨胀的方向演进，对抖动极为敏感，而推理计算复用、模型调度等多边云协同需求将推动东西向流量向大带宽、高突发的方向变化，对丢包高度敏感。

DIN（Distributed Inference Network）的技术架构是支撑高效推理服务的网络基础设施，采用分布式架构，实现端、边、网、算的有效协同，提供差异化、高安全、高并发、高频率、高突发的网络连接服务保障。通过微流级流控、层次化细粒度切片及以太网相干 PHYSec 技术达成广域丢包率小于 $10^{-9}$、微秒级抖动、端到端安全等目标，实现总体效能最优、推理体验可保障。

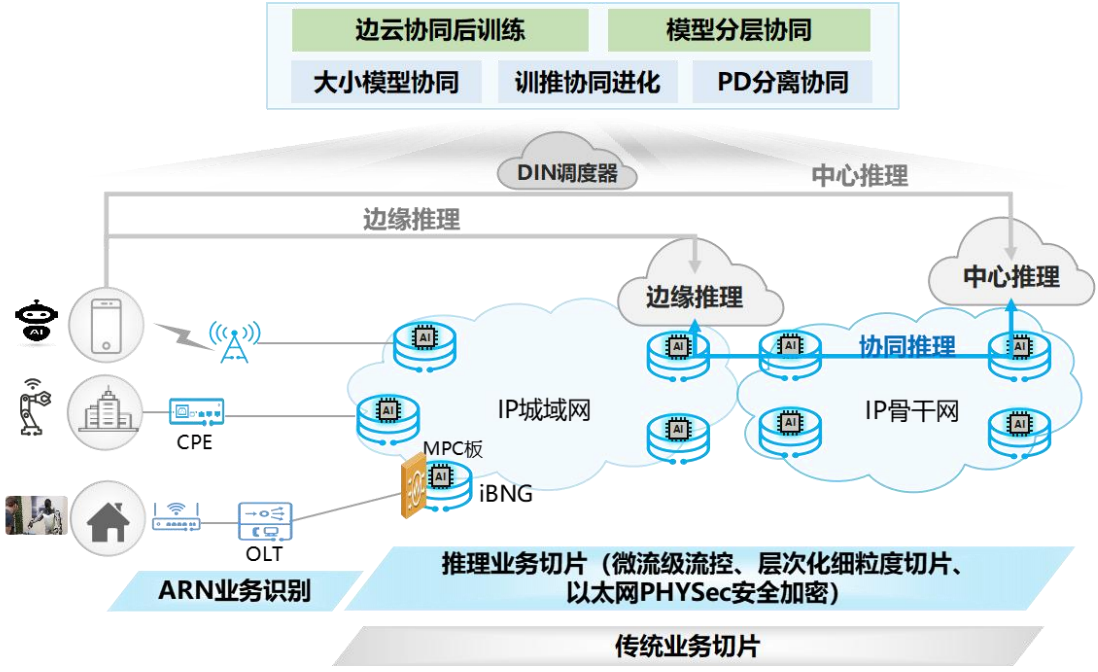


图 6 DIN 架构

DIN 主要的设计目标包括：

1、可扩展架构：推理业务爆发驱动网络流量快速增长，加速网络向边缘进一步延伸，网络需具备弹性扩展、高效交互、智能调度等能力，为用户提供无处不在的低时延推理服务，保障多推理节点间频繁通信与协作的流畅性。

2、确定性连接：面向不同模式的云边端协同推理场景，网络需提供确定性、低时延、高并发的极致性能保障，实现广域丢包率小于 $10^{-9}$、微秒级抖动的数据传输能力，满足算效和推理体验的双重优化。

3、多层次安全：为应对 AI 平台、AI 模型、AI 应用、KV Cache 等数据在传递过程中的安全威胁，网络需具备多层次安全防护能力，实现数据全链路加密与隐私保护，动态抵御网络攻击，保障 AI 服务安全稳定运行。

# 4. 分布式推理网络（DIN）关键技术

面向 AI 普惠推理时代的推理业务场景需求和流量模型变化，DIN 需要具备推理业务质量保障、推理服务调度、推理安全防护等关键能力。

## 4.1 节点间互联质量保障技术

### 4.1.1 微流级流控技术

在边云协同后训练、模型分层协同、PD 分离协同等多个推理服务场景中，训练阶段的前向计算中间结果（如隐变量）与反向梯度、推理阶段的 KV Cache 等关键数据需要在云边之间高效传输，AI 训推的性能边界与网络丢包率存在强耦合关系，需要网络将丢包率控制在极低水平，才能避免算效损失与服务中断。

针对此需求，中国移动创新提出微流级流控技术，通过网络设备中部署租户级和业务流级的队列隔离，为每个用户队列设定反压阈值，并能实时感知网络的拥塞状况。一旦队列使用的缓存超出预设的反压阈值，设备会迅速生成精确的流控反压报文，并通过反向路径通知上游设备暂停该用户队列的数据传输。当用户队列已使用缓存回落至停止门限阈值以下时，拥塞设备解除拥塞状态，并停止向上游设备发送精准流控反压报文，上游设备重新发包，实现用户队列的零丢包弹性传输能力。在网络边缘节点，微流级流控技术协同数据中心内的 PFC 技术实现弹性速率控制，使能网络级缓存协同，实现端侧协同降速。

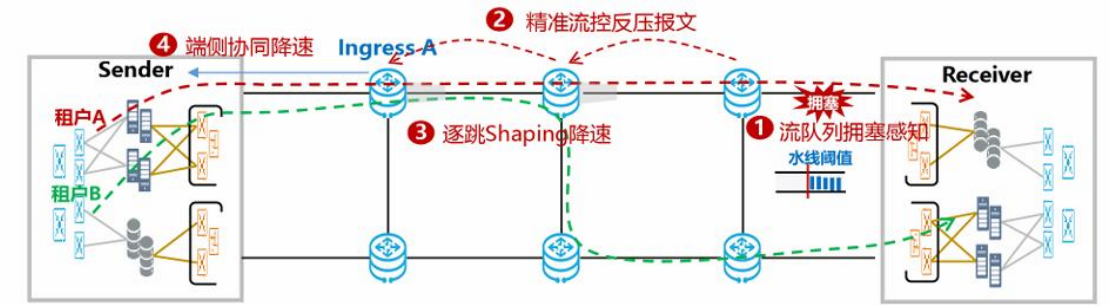


图 7 微流级流控反压

后续中国移动将进一步结合 AI 预测流量模型等技术，在主动拥塞避免等方面深入探索研究，助力构建突破性能、极低时延的 DIN 网络。

### 4.1.2 层次化细粒度切片技术

为满足推理业务的低时延、高可靠、大带宽等需求，网络需通过 G-SRv6 融合网络切片、随流检测技术，构建端到端确定性质量保障体系，实现差异化业务 SLA（时延、抖动、带宽、丢包率等）的精准承诺与动态优化。

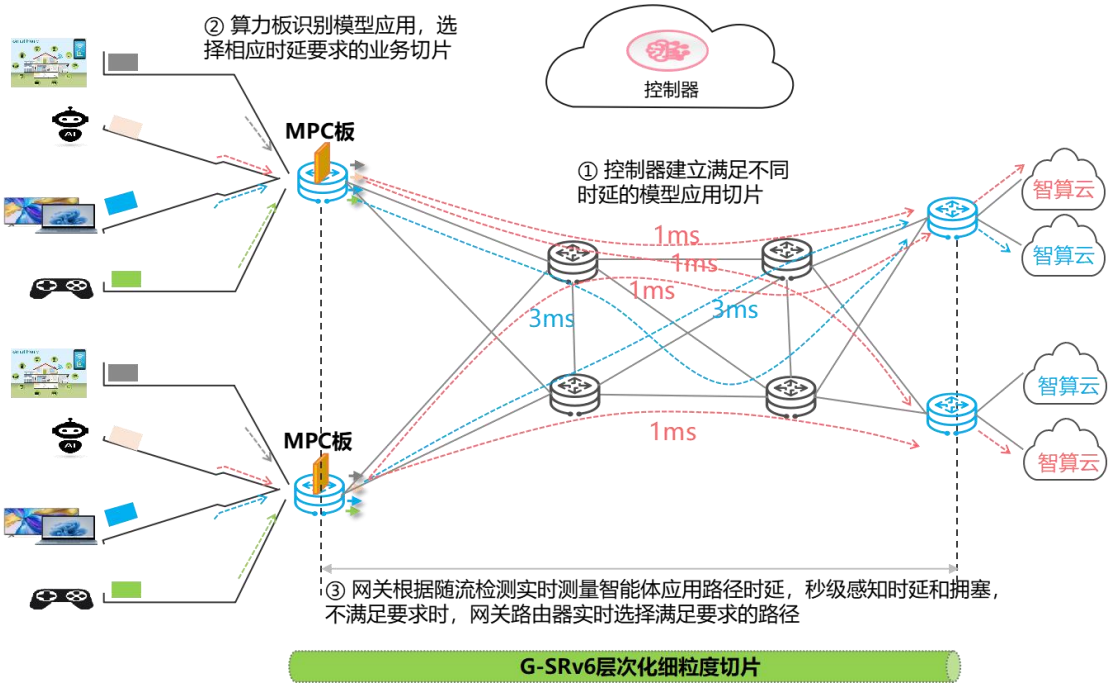


图 8 层次化细粒度切片

G-SRv6 层次化细粒度切片技术可将物理网络按需划分，如：超低时延型（例

如时延<1ms）、带宽保障型（例如带宽>10Gbps）、高可靠型（例如丢包率<0.01%）。边缘推理接入设备对应用进行识别后，将业务流引入不同 G-SRv6 切片网络中，通过 G-SRv6 提供高效可编程路径结合层次化细粒度切片实现物理资源硬隔离，为不同切片分配独立通道，提供确定性质量保障。

同时，可通过随流检测提供网络质量的实时监控，在用户业务流中嵌入检测标记（包含流标识、时间戳、序列号），逐跳采集时延、抖动、丢包等数据。控制器基于采集数据可构建全网质量指标地图，当检测到 SLA 偏差时，触发 SRv6 路径重优化（如切换至低拥塞的 Underlay 路径）。通过监控数据驱动 SRv6 路径秒级调整，实现从“尽力而为”到“确定性”质量保障。

### 4.1.3 推理业务识别技术

接入网络的业务类型多样，不同推理业务的不同用户所需的推理资源可能各不相同。因此，为了优化资源分配和提升服务质量，有必要对业务或应用进行精确区分，识别出需要推理服务的具体应用，基于应用识别的结果为业务提供满足诉求的推理服务。

端侧设备根据推理诉求携带标识，可通过 ARN（Application Responsive Network，应用响应网络）技术实现。ARN 通过在应用和网络之间增加一个中间层，基于数据面进行网络能力开放可编程，让应用像调用操作系统一样调用网络资源。利用 IPv6 报文自带的可编程空间，将应用标识信息携带进入网络，网络入口根据 ARN 标识对应用提供差异化的推理网络服务。ARN 将地址和服务解耦，以及网络和应用解耦，网络不需要直接感知应用，屏蔽了应用的多样性，同时也防止应用直接访问网络能力。通过对网络和应用各自进行封装，实现应用隐私及网络内部信息的隐藏。

端侧设备不具备 ARN 标记能力时，需要网络侧具备 AI 推理应用识别能力，为其封装 ARN 标识并提供对应等级服务。网络侧面向不同模型推理类型（归纳总结、多轮对话、多模态推理、Agents 交互等）的流量行为数据，通过决策树、神经网络、知识推理增强、多专家识别等方式实现推理模式的精准智能分类，给出快速识别结果以及识别置信度，提供高性能、高精度的推理模式识别能力。

## 4.2 推理服务的调度技术

模型调度是 DIN 的主要功能之一，旨在根据用户需求、设备性能、网络状态等多维度因素，动态分配和优化模型资源，以实现高效、低延迟的模型推理服务。模型调度的核心目标是在保证服务质量的前提下，最大化资源利用率并降低运营成本。业务或用户请求推理服务时，边缘节点根据业务的推理诉求进行推理资源的分布式资源调度，满足多用户的推理诉求。

运营商可将已有的网络调度技术能力优势扩展到推理模型调度，通过高效的模型调度，DIN 能够在复杂多变的环境中实现资源的智能分配，为用户提供低延迟、高可用的 AI 服务，同时降低运营成本。

推理服务的调度技术分为集中式和分布式两种方案。其中集中式方案由用户发出推理业务的访问请求到 DNS，DNS 域名解析先解析到 DIN 调度器，DIN 调度器再根据边缘节点的负载、用户 IP 地址就近、网络质量、流量均衡等原则分配最佳 DIN 节点 IP 地址，并把 IP 地址返回 DNS 以及进一步返回给用户。用户再次发起访问请求直接访问最优 DIN 边缘节点，边缘节点根据处理能力，直接返回推理结果，或与中心节点协同推理后返回推理结果，为用户提供差异化的推理应用体验；分布式采用算力路由方案，算力路由支持算力感知、通告、联合路由功能，打破了传统互联网路由方式，基于“算力+网络”的多因子联合调度算法，通过对算力资源/服务的部署位置、实时状态、负载信息的感知，以及对推理业务需求的感知，按需动态生成业务调度策略，将业务沿最佳网络路径调度到目的推理服务节点。相比于集中式的调度方案，算力路由技术通过在 DIN 边缘节点处查找算力路由信息表，并确定最佳服务节点，为用户提供算力+网络融合的最优服务。

## 4.3 模型推理安全防护技术

### 4.3.1 以太网相干 PHYSec 技术

DIN 通过“算-网-边-端”协同架构实现模型的分布式部署与动态调度，其边云协同后训练（PostTraining）、PD 分离协同等推理服务均存在长距链路传递过程中的数据机密性与完整性安全需求。传统安全技术如 MACSec 应用于 DIN 网络链

路时，存在安全信息开销大、降低网络带宽利用率，无法掩盖用户业务流量特征等限制。相干以太网相干 PHYSec 技术适用于以太网长距相干链路，将现有密码学方法与以太网物理层技术融合，具有保护用户全栈信息、掩盖流量特征的高安全链路防护能力。其基于原生 PAD 域承载安全协议不引入加密开销，不影响链路带宽利用率。

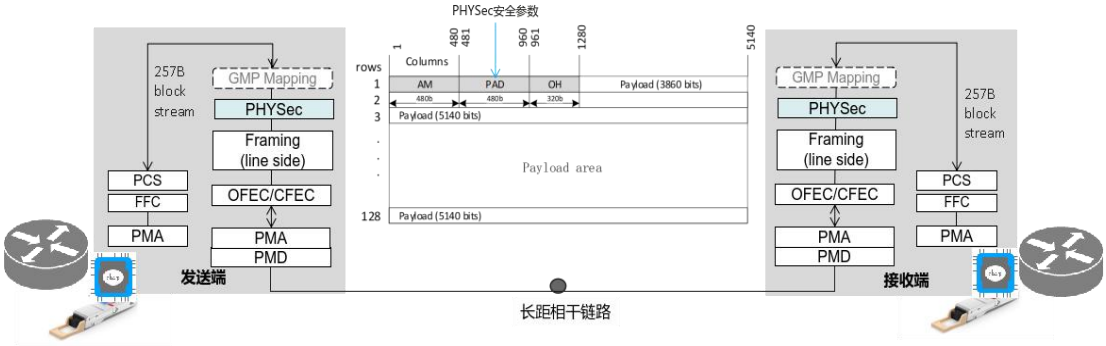


图 9 以太网相干 PHYSec 技术架构

以太网相干 PHYSec 技术架构如图所示，创新提出新架构、新算法、新协议三大核心机制。新架构将安全协议下沉到物理层，对比特流实施全加密。新算法可以提前计算解密参数 Ek、H，实现低时延解密和完整性校验。新协议基于原生机制 PAD 区域携带安全协议，无额外开销。其核心技术流程为：发送端的用户业务数据经物理层处理还原为 257B 码块，对比特流进行 GMP 映射，插入 AM/PAD/OH 形成相干复帧，同时对帧内比特流进行加密或完整性校验，经 PAD 域承载安全参数。接收端接收到相干复帧后，对帧内比特流进行解密或完整性校验，GMP 解映射后还原为以太网物理层的原生 257B 码块。

以太网相干 PHYSec 通过物理层 257B 码块进行加解密，可以做到对上层业务和协议进行透明加解密，即无需修改上层协议栈即可对模型分发、KV Cache 等过程数据进行加密，防止链路传递过程中被中间节点窃取或篡改。另外，以太网相干 PHYSec 的安全参数由相干复帧的原生 PAD 域承载，实现零加密开销，且兼容已有 FlexOsec 技术，可高效保护 DIN 模型分发流量，避免传统安全加密导致的带宽资源浪费。

### 4.3.2 拒绝服务流量防护

DIN 利用嵌入式 AI 技术，在检测与处置两方面提供抗 DDoS 服务，与安全平台协同，做到恶意流量快速感知、自动阻断。

传统 DDoS 攻击检测采用抽样检测方式，耗时一般在分钟级，影响对新型秒级加速、分钟级持续的瞬时泛洪攻击的防护效果。DIN 嵌入的智能流识别算法，在数据面实现了百毫秒周期多维流行为 1：1 统计数据采集，在控制面流式学习并维护每 IP 粒度的业务流量模型，协同监控异常的流速突升、报文成分或报文长度变化情况，识别出攻击关键信息并输出到安全平台，将感知 DDoS 攻击的时间从分钟级缩短到秒级，从而协同清洗中心有效应对快速泛洪攻击。

对于 DDoS 攻击的处置同样面临新的挑战。近年来，全球记录的最大攻击达到 Tbps 量级，而清洗带宽的增加依赖多方面投入，难以匹配攻击带宽的爆炸增长。DIN 嵌入的 DDoS 攻击溯源能力，基于多维报文特征建立 AI 模型，长期监控业务流量趋势与攻击期间的特征异变情况，从而精准识别攻击源，对常见网络层攻击起到前置过滤的效果，线速处理流量型攻击向量，缓解清洗池带宽压力。

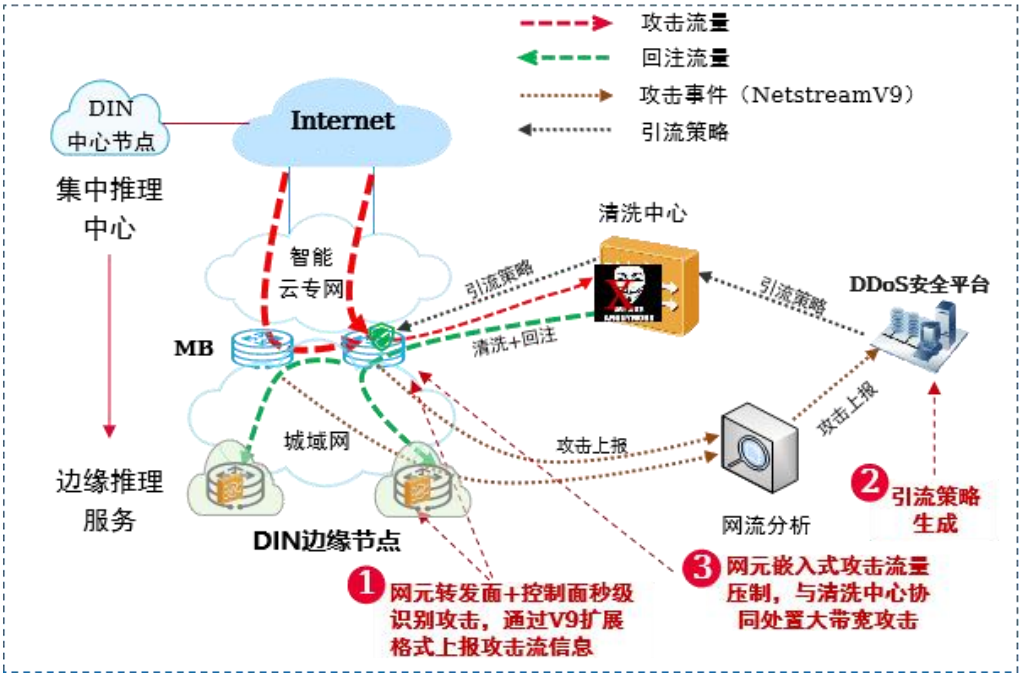


图 10 DDoS 防护

### 4.3.3 基础设施轻量化 APT 监测能力

人工智能大模型已经渗透到各行各业中，DeepSeek 在应用中遭受的攻击促使我们关注人工智能时代的网络安全问题。DIN 需要具备入侵检测及防御能力，可基于“白+黑”模式的轻量化入侵检测技术，通过结合路由器的正常业务行为（如合法操作、白名单规则）与 APT（Advanced Persistent Threat，高级持续性威胁）攻击知识库（包括样本特征和攻击行为模式），构建多维攻击检测系统。通过监控内核关键数据对象、系统文件、系统账号和配置等，识别非法篡改行为，并监测异常进程活动，从而实现对隐蔽且持久的 APT 劫持攻击的全面威胁感知，保证基础设施安全。

# 5. 总结与展望

当前 AI 大模型的发展迎来重要变化，一是 AI 普及速度显著加快，用户从访问内容向访问 AI 模型转变，二是 AI Agent 无处不在，Agent 之间的通信会显著增长。与之相对应，AI 时代网络流量特征的变化，以及亿级海量用户并发以及随之而来的安全威胁，对 AI 推理基础设施、AI 推理技术架构、AI 服务安全防护能力都带来了新的挑战。

面向推理服务商“模型随需部署、随需更新、高效应对海量并发、弹性高效调度”，推理用户“推理无处不在、智能触手可及、服务体验最优”的 AI 普惠目标，中国移动提出新型分布式推理网络 DIN，融合了运营商网络协议可编程、流量感知调度能力、确定性体验保障以及安全防护能力等优势，通过算网一体安全推理、边云协同后训练、模型分层协同、大小模型协同、训推协同进化、PD 分离协同等多种端边云分布式协同模式，解决大模型集中化部署模式下的大规模并发推理能力不足的问题，有效应对亿级海量用户并发挑战，同时通过构筑多维度安全能力，实现了供需双方的安全高效。

展望未来，新技术发展日新月异。多 Agent 技术将让 AI 能够以更自然、更智能的方式与用户交流，通过多智能体的协同，精准理解和满足用户需求，并提供更加个性化的服务体验；具身智能技术将使得人形机器人能够更好地感知环境、理解自然语言指令、完成复杂的任务操作；IoT 与 AI 深度融合，利用 AI 技术对物联网数据进行深度分析和挖掘，实现设备的智能预测性维护、能源管理等，为各行业创造更大的价值。

面向未来，中国移动将继续发挥自身在算力网络领域的领先优势，联合产业界重点构建和完善融合端、边、网、算的 DIN 技术体系和标准体系，解决大模型在个人、家庭及企业应用中的成本、效率与场景适配难题，并在实践中与合作伙伴形成 AI 推理时代的新商业模式，助力 AI 推理普惠化发展，为中国以及全球的 AI 产业发展注入新动力，加速迈向全面智能社会。

# 6. 缩略语

| 缩写 | 全称 | 说明 |
|---|---|---|
| AI | Artificial Intelligence | 人工智能 |
| API | Application Programming Interface | 应用程序编程接口 |
| APT | Advanced Persistent Threat | 高级持续性威胁 |
| ARN | Application Responsive Network | 应用响应网络 |
| CNP | Congestion Notification Packet | 拥塞通知报文 |
| CPE | Customer Premise Equipment | 客户前置设备 |
| CPU | Central Processing Unit | 中央处理器 |
| DAU | Daily Active User | 日活跃用户 |
| DDoS | Distributed Denial of Service | 分布式拒绝服务攻击 |
| DIN | Distributed Inference Network | 分布式推理网络 |
| eMBB | Enhanced Mobile Broadband | 增强移动宽带 |
| GPU | Graphics Processing Unit | 图形处理器 |
| iBNG | Intelligent Broadband Network Gateway | 智能化宽带网络网关 |
| ID | Identification | 标识 |
| IIoT | Intelligent Internet of Things | 智能物联网 |
| KV | Key Value | 键值 |
| DIN | Model Distribution Network | 模型分发网络 |
| MoE | Mixed Expert Models | 混合专家模型 |

| | | |
|---|---|---|
| MPC | Multi-Purpose Computing | 多维算力单板 |
| NP | Network Processor | 网络处理器 |
| NPU | Neural network Processing Unit | 神经网络处理器 |
| PD | Prefill-Decode | 预填充-解码 |
| PoP | Point of Presence | 网络接入点 |
| RDMA | Remote Direct Memory Access | 远程直接内存访问 |
| SDK | Software Development Kit | 软件开发工具包 |
| SLA | Service Level Agreement | 服务等级协议 |
| SRv6 | Segment Routing over IPv6 | IPv6 段路由 |
| TBT | Time-Between-Tokens | Token 间时延 |
| TTFT | Time to First Token | 输出第一个词元时间 |
| uRLLC | Ultra-Reliable Low Latency Communications | 超可靠低时延通信 |