Contribution ID: **149** Type: 口头报告

面向结构性网络退化的解耦式大语言模型推理系统可靠性调度方法

摘要

随着大语言模型推理服务从单体式架构转向 Prefill-Decode 解耦架构，跨节点 KV Cache 传输逐渐成为影响系统可靠性的关键因素。长上下文请求会产生大规模状态迁移，使推理服务的成功完成不仅依赖 GPU 计算能力和显存容量，也依赖数据中心网络路径的可用性。现有调度方法多依赖本地队列长度、连接数或显存余量进行逐请求路由，能够较好处理局部计算扰动，但在结构性网络退化下可能产生误判。当核心链路带宽下降时，请求可能在 KV Cache 传输阶段阻塞并最终超时，而该过程并不一定及时反映到目标节点的本地队列指标中，从而诱发持续的超时级联失效。

本文提出 OTF-R，一种面向解耦式大语言模型推理系统的双环可靠性调度方法。该方法通过慢环聚合全局拓扑与资源状态，将网络退化和资源拥塞转换为 pod 级风险价格信号；快环在逐请求尺度融合本地队列、显存状态与全局风险价格，实现故障感知路由。基于 SimPy 离散事件仿真，本文在 64 节点 Fat-Tree 推理集群中注入计算退化和核心链路退化故障，并采用失败率、退化可用性、韧性指数、超时失败数和失败构成等指标进行评估。结果表明，局部计算退化主要可由本地队列感知调度处理，而结构性网络退化需要全局风险价格抑制超时级联。在最严重网络退化场景下，OTF-R 将平均失败率由 20.1% 降低至 5.9%，将退化可用性由 0.32 提升至 0.99，显著提升了系统在退化工况下的可靠性与韧性。

关键词

大语言模型推理；解耦式服务；可靠性调度；结构性网络退化；级联失效；韧性控制

Abstract

As large language model (LLM) inference systems move from monolithic serving architectures to prefill-decode disaggregation, cross-node key-value (KV) cache transfer becomes a critical factor affecting system reliability. Long-context requests generate large state transfers, making successful inference dependent not only on GPU compute capacity and memory availability, but also on the availability of data-center network paths. Existing schedulers usually rely on local queue length, active connections, or memory availability for per-request routing. While such local policies are effective for component-level compute perturbations, they may be misleading under structural network degradation. When core-link bandwidth is reduced, requests can stall during KV-cache transfer and eventually time out, without being promptly reflected in the target node's local queue indicators. This can lead to timeout-driven cascading failures.

This paper proposes OTF-R, a dual-loop reliability-aware scheduling method for disaggregated LLM inference systems. The slow loop aggregates global topology and resource states and converts network degradation and resource congestion into a pod-level risk price. The fast loop then combines local queue information, memory state, and the global risk price to perform fault-aware request routing. Using a SimPy-based discrete-event simulator, we evaluate OTF-R on a 64-node Fat-Tree inference cluster with injected compute degradation and core-link network degradation. Reliability is assessed using failure rate, degraded availability, resilience index, timeout failures, and failure composition. The results show that local queue-aware scheduling is sufficient for component-level compute degradation, whereas structural network degradation requires global risk pricing to suppress timeout cascades. Under the most severe network degradation scenario, OTF-R reduces the mean failure rate from 20.1% to 5.9% and improves degraded availability from 0.32 to 0.99, significantly enhancing system reliability and resilience under degraded operation.

Keywords

large language model inference; disaggregated serving; reliability-aware scheduling; structural network degradation; cascading failure; resilience control

Author: HUANG, Feng (Tsinghua University)

Presenter: HUANG, Feng (Tsinghua University)

Session Classification: 人工智能

Track Classification: 口头报告: 人工智能