AI面临的互联网挑战

Dirk KUTSCHER

香港科技大学(广州)
THE HONG KONG
UNIVERSITY OF SCIENCE AND
TECHNOLOGY (GUANGZHOU)

目录

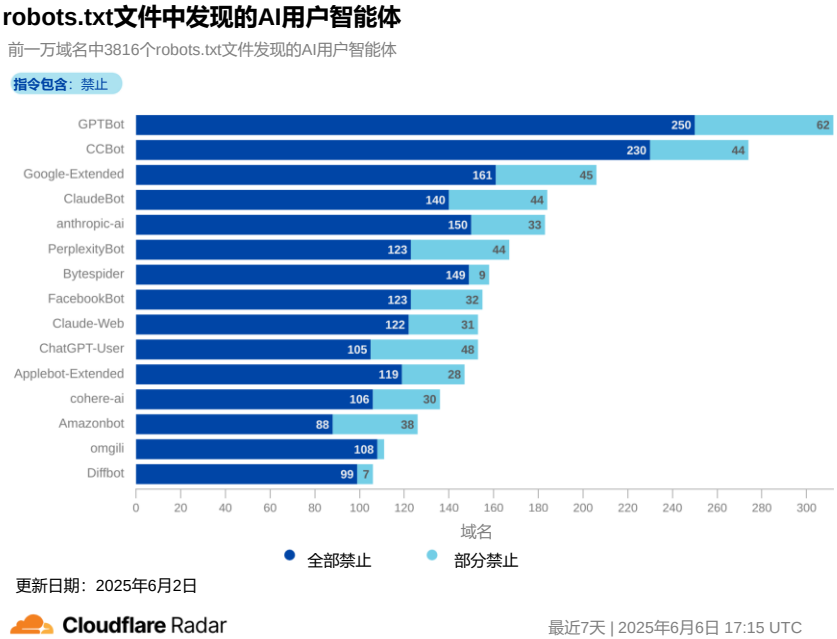
AI爬虫对互联网的影响

- 大语言模型（LLM）服务数量持续攀升
- AI爬虫持续抓取网络数据
- 网络流量和服务器负载日益增长
- 控制AI网络爬虫行为的技术解决方案与标准制定愈发受到关注

协议测量与分析研究组（MAPRG）

- **2025年7月互联网工程任务组（IETF）第123次大会专题聚焦会议：AI流量爬虫影响评估**
  - 评估对网络、服务器、内容分发网络（CDN）的影响
  - 探讨技术思路及当前发展动态
- **https://datatracker.ietf.org/meeting/123/session/maprg**

AI爬虫流量



单个爬虫流量增加，且爬虫总体数量增长

https://blog.cloudflare.com/from-googlebot-to-gptbot-whos-crawling-your-site-in-2025/

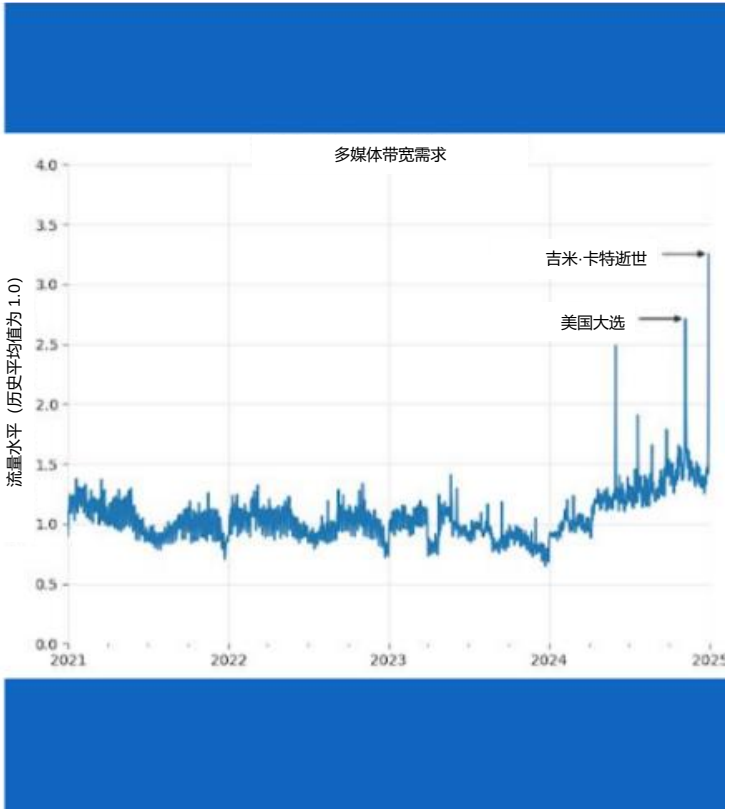
协议测量与分析研究组

- **观察要点**
  - 自动化流量占据全球网络流量的30%至50%
  - AI爬虫是近期流量激增的主要推动力
  - 过去，爬虫主要用于索引内容并引导用户，互惠互利
  - 如今，爬虫抓取行为对基础设施造成沉重负担，却未能提供相应回报
  - 爬虫机器人被视为资源掠夺者，普遍受到排斥

案例：维基媒体基金会（Wikimedia）

爬虫机器人的隐患

- 自2024年1月以来，维基媒体基金会的带宽用量增长了50%，主要由机器人流量引起。
- 爬虫不仅针对维基媒体项目，还会针对基础设施中任何URL进行访问，比如：缺陷跟踪系统、代码审查平台等。
- 由于大量多媒体内容受到新增关注和访问，支撑这些访问所需的资源已显著增加。



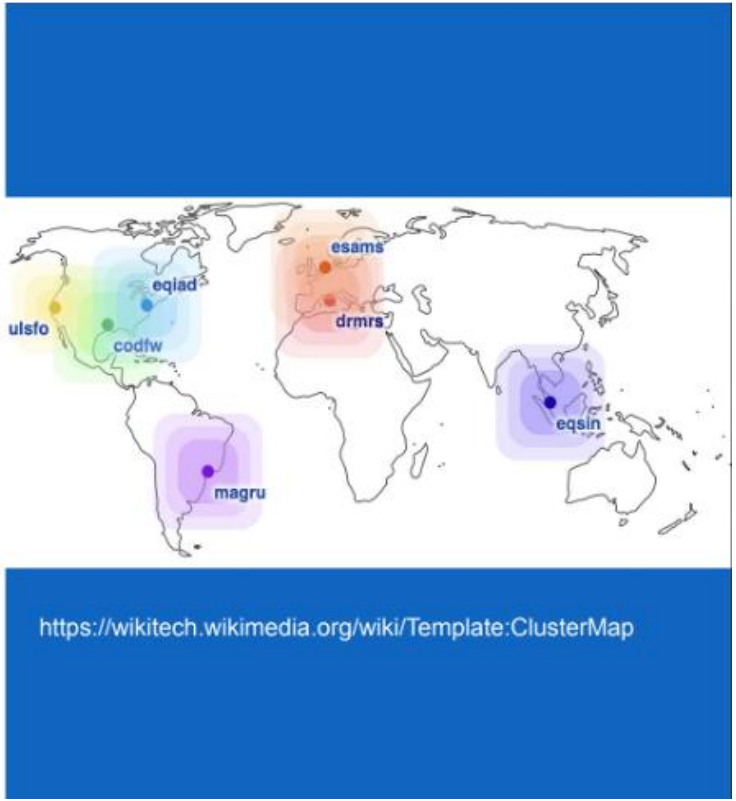
https://datatracker.ietf.org/meeting/123/materials/slides-123-maprg-bot-traffic-at-wikimedia-measurement-identification-and-response-00

案例：维基媒体基金会

内容分发网络：

以真实用户为核心进行优化

- 配有2个主数据中心和5个缓存中心
- 基于地理位置的数据中心路由机制
- 每月页面浏览量约为250亿次
- 能够承受突发的大规模流量激增
- 人类用户请求通常可在缓存层直接响应
- 机器人请求被转发到主数据中心的概率更高，因为爬虫会访问任意URL，包括不常受访问的页面

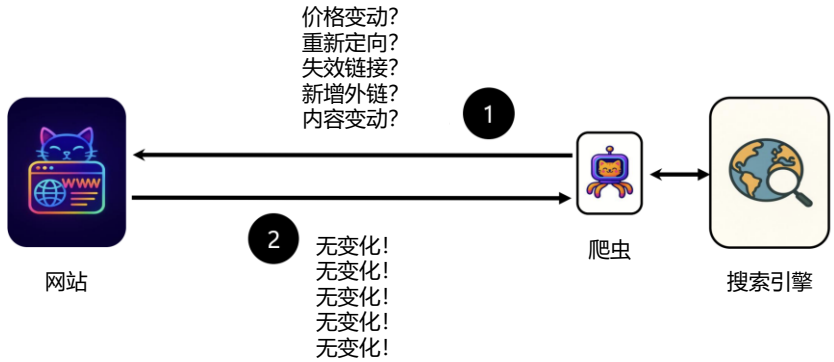


https://datatracker.ietf.org/meeting/123/materials/slides-123-maprg-bot-traffic-at-wikimedia-measurement-identification-and-response-00

爬取效率低下

- **爬取效率：抓取有价值且新鲜内容（包括新增、更新或删除的内容）在总抓取活动中的比例**
  - 当前效率较低
  - 大量资源重复请求
- **新进展**
  - 网站内容更新时通知搜索引擎
  - 爬虫随后抓取更新资源
  - 示例：IndexNow

传统爬取模型亟需创新



爬虫反复访问几乎没有变化的页面!

重复爬取浪费带宽和计算资源，且不环保!

长尾网站包含数十亿条更新频率极低的URL，爬取难以扩展。

目录

IETF AI偏好工作组（AIPREF）

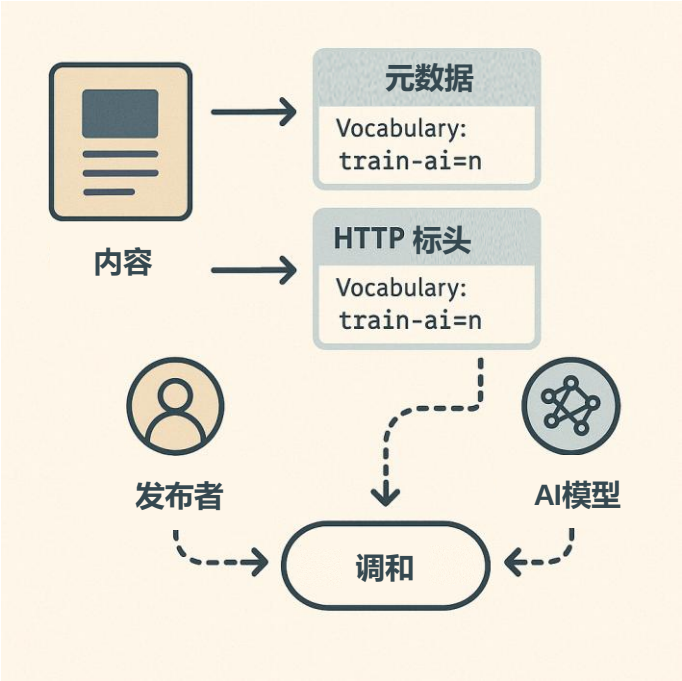
- **基于海量网络数据集训练的AI系统快速发展，以下矛盾日益突出：**
  - 出版商、作者及版权方：
    希望表达其内容是否（以及如何）用于AI训练、推理或相关用途
  - AI开发者与爬虫系统：
    目前缺乏在大规模场景中发现并解读偏好的标准化方式
- **现状：缺乏具备互操作性、机器可读的机制**
  - 无法以一致的方式、跨网络资源传递偏好

AI偏好工作组

- **robots.txt或元数据标签等机制最初为搜索引擎爬虫设计**
  - 不适用于AI特定用途
- **各类网站尝试自定义标头（Header）或格式，导致碎片化问题**
- **AI爬虫缺乏统一的词汇来解读偏好，如：**
  - “允许索引，但禁止用于训练”
  - “允许生成摘要，但禁止在推理中重用”
- **缺乏标准化，内容所有者难以掌控内容使用**
  - AI爬虫在合规性方面面临诸多不确定性

AI偏好工作组目标

- **AI偏好工作组构建通用框架**
  - 定义一套用于表达AI相关内容使用偏好的词汇
  - 制定HTTP标头、robots.txt文件、元数据等标准附加方式，传达相关偏好信息
  - 提供处理多重偏好表达冲突的调和机制



https://datatracker.ietf.org/wg/aipref/about/

目录

代理式AI

- **赋予基础大模型目标导向行为能力**
  - 规划
  - 工具使用/执行操作
  - 记忆
  - 自我监控
- **在最少提示下主动追求目标，并与其他智能体或人协调配合**
  - 具备主动性并以工作流程为中心

```
graph LR
    subgraph "Specialized Agents"
        FBA[✈️ Flight Booking Agent]
        HRA[🏨 Hotel Reservation Agent]
        CCA[💱 Currency Conversion Agent]
        LTA[🚌 Local Tours Agent]
    end

    AI_Assistant[🤖 AI Assistant] --> FBA
    AI_Assistant --> HRA
    AI_Assistant --> CCA
    AI_Assistant --> LTA
```

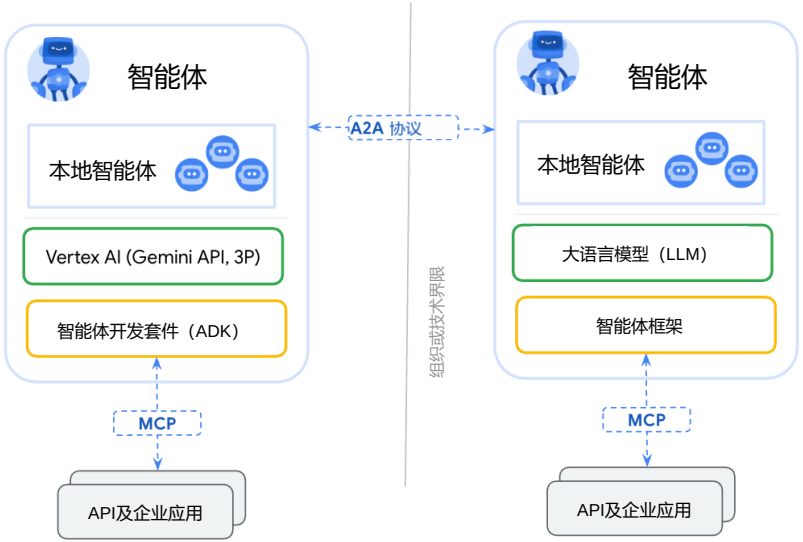
https://a2a-protocol.org/latest/topics/what-is-a2a/#the-interoperability-challenge

代理式AI的挑战

- 发现智能体/API
- 鉴权与身份验证
- 授权与委托
- 生命周期/会话管理
- 安全、保障、责任
- 协议语义/互操作性
- 模态与能力协商

行业精选动态

- **模型上下文协议（MCP）**
  - 连接应用程序与智能体，实现上下文共享
  - https://modelcontextprotocol.io/
- **Agent2Agent（A2A）协议**
  - 互联网中智能体互联互通
  - https://github.com/a2aproject/A2A
- **AGNTCY**
  - 智能体协作的基础设施栈
  - https://agntcy.org/



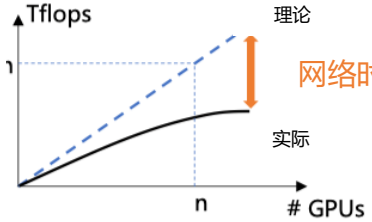
IETF现状

- 探索领域
- 2025年7月IETF第123次会议期间举办智能体通信会议
- 尚未就成立工作组达成共识
- 首批个人提交的互联网草案
  - https://www.ietf.org/id/draft-rosenberg-ai-protocols-00
  - https://datatracker.ietf.org/doc/draft-stephan-ai-agent-6g/

目录

分布式机器学习与网络

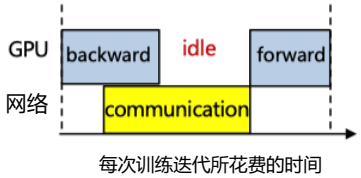
大规模集群 ≠ 有效FLOPS



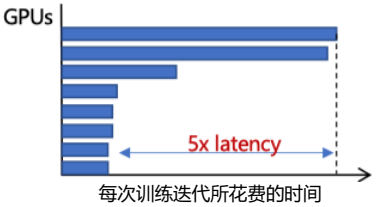
网络时延影响 > 50%

网络性能限制GPU集群的有效计算能力（FLOPS）

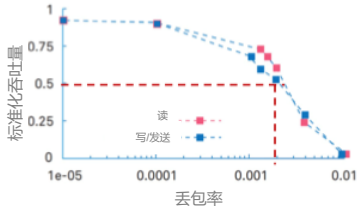
带宽影响（TB）



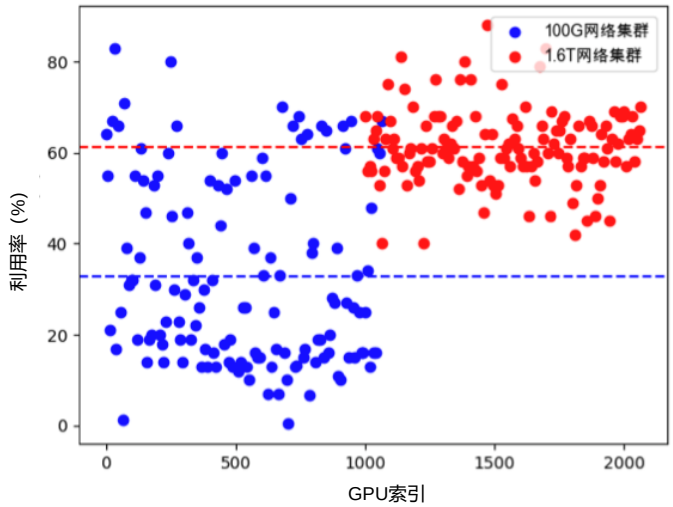
时延影响（5x）



丢包影响（0.1%丢包导致吞吐率下降50%）

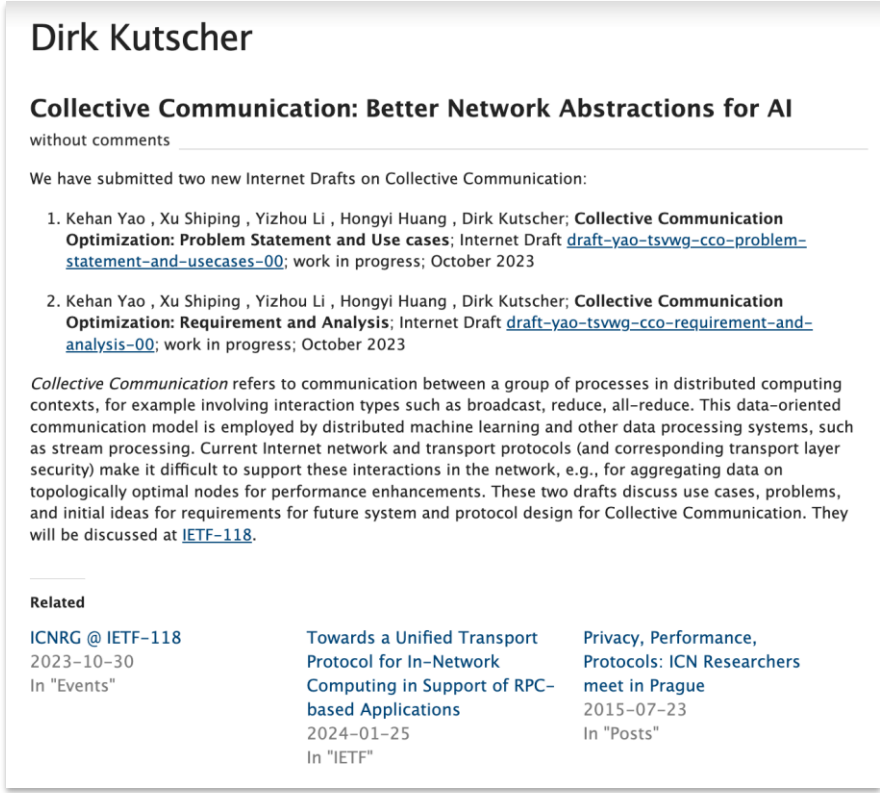


GPU 使用 32%→61%



研究挑战

- **互联网和数据中心传输**
  - 可靠性：底层网络缺乏通信可靠性
  - 基于应用数据单元而非数据包
  - 应对拥塞的最佳策略是什么？
- **网络内聚合**
  - 减轻服务器与网络负载
- **线速网络计算**
  - 机器学习训练和数据聚合对延迟有最大时限要求
  - 严苛的性能要求



Dirk Kutscher

Collective Communication: Better Network Abstractions for AI

without comments

We have submitted two new Internet Drafts on Collective Communication:

1. Kehan Yao , Xu Shiping , Yizhou Li , Hongyi Huang , Dirk Kutscher; **Collective Communication Optimization: Problem Statement and Use cases**; Internet Draft draft-yao-tsvwg-cco-problem-statement-and-usecases-00; work in progress; October 2023
2. Kehan Yao , Xu Shiping , Yizhou Li , Hongyi Huang , Dirk Kutscher; **Collective Communication Optimization: Requirement and Analysis**; Internet Draft draft-yao-tsvwg-cco-requirement-and-analysis-00; work in progress; October 2023

Collective Communication refers to communication between a group of processes in distributed computing contexts, for example involving interaction types such as broadcast, reduce, all-reduce. This data-oriented communication model is employed by distributed machine learning and other data processing systems, such as stream processing. Current Internet network and transport protocols (and corresponding transport layer security) make it difficult to support these interactions in the network, e.g., for aggregating data on topologically optimal nodes for performance enhancements. These two drafts discuss use cases, problems, and initial ideas for requirements for future system and protocol design for Collective Communication. They will be discussed at IETF-118.

Related

ICNRG @ IETF-118
2023-10-30
In "Events"

Towards a Unified Transport Protocol for In-Network Computing in Support of RPC-based Applications
2024-01-25
In "IETF"

Privacy, Performance, Protocols: ICN Researchers meet in Prague
2015-07-23
In "Posts"

总结

	趋势	IETF/IRTF
AI爬虫影响	• AI爬虫导致流量增加和服务器负载加重 • 网站尝试控制流量和爬虫行为	IRTF研究
AI偏好	• 构建指定并传输网站偏好的框架 • 功能超越robots.txt	IETF标准开发
代理式AI	• 智能体间通信呈上升趋势 • 首批行业联盟与协议规范	IETF探索
AI基础设施	• 分布式机器学习规模部署的挑战 • 数据中心和互联网传输 • 协议开发与基础设施支持	IETF和IRTF探索

Dirk KUTSCHER
德克·库彻

感谢聆听!

香港科技大学(广州)
THE HONG KONG
UNIVERSITY OF SCIENCE AND
TECHNOLOGY (GUANGZHOU)

dirk-kutscher.info
dku@hkust-gz.edu.cn