

# 基于 eBPF 的大模型性能 分析及慢节点检测

姓名：程书意

工作单位：阿里云



# 目录

## CONTENTS

01

### AI系统性能瓶颈分析

A designer can use default text to simulate what text would look like.

02

### 持续剖析：AI火焰图

A designer can use default text to simulate what text would look like.

03

### 场景：慢节点检测

A designer can use default text to simulate what text would look like.

04

### 未来展望

A designer can use default text to simulate what text would look like.



# 01. AI系统性能瓶颈分析

## 慢节点现象

- 现象：单节点拖慢全局迭代
- 影响：训练耗时变长，算力严重浪费

## GPU利用率低

- 现象：GPU利用率上不去
- 影响：训练/推理效率低

## 性能差异现象

- 现象：相同配置下推理/训练性能数据不一致
- 影响：客户流失



CLK



# 01. 常用的分析方法

分析方法	优势	不足
指标	宏观概览、历史趋势、告警、轻量级	缺乏细节、难定位根因、多个指标关联分析复杂
日志	信息丰富、问题定位	性能问题往往没有日志输出
Continuous-profiling	代码级瓶颈定位、历史回溯	缺乏GPU信息
Nsight-system	GPU深度分析、CPU-GPU协同分析	性能影响大、数据量大



# 01. 智能运维平台 Sysom

**SysOM**：一站式智能运维平台，中心端采用微服务的方式处理来自多个region的Profiling和监控数据；

**SysAK**：系统工具集，处理来自中心端的请求和数据预处理，进行诊断结果汇总和Profiling、监控数据上传；

**Coolbpf**：eBPF开发框架和功能库，提供持续剖析数据采样，网络、IO、内存、调度及GPU指标等功能。





# 01. SysOM AI可观测

## AI Profiling

使用完全无侵入式方式，对AI模型训练场景中资源、调用栈等进行细粒度采集，帮助开发者快速定界异常或性能瓶颈，提升系统整体性能

## AI 诊断

通过对 算子延迟、硬件xid等信息检测，判断GPU等异常情况

## AI 集群拓扑

提供对GPU资源的整体视图；对集群GPU拓扑分布展示；对GPU穷、推理链路延迟高的节点进行异常提示。

## AI 火焰图、热力图

通过对CPU栈和GPU算子等信息进行特征提取，常态化实现慢卡、故障定界的分析。



## 智能工单

使用AI对工单进行初步分类、分析和值班人员关联，使用OS专用大模型继续答疑，诊断智能体通过MCP调用相关诊断工具解决问题

## 智能运维操作

融合AI能力到运维竞争力产品中，如诊断智能体和智能宕机分析工具 在运维产品维护过程中，使用AI辅助告警分类推送和版本自动升级；

## 运维知识沉淀

通过运维处理流程自动沉淀知识库，以MCP、智能机器人等方式输出到一线服务团队，提升问题解决效率

## 02. 智能运维平台 Sysom

**SysOM**：一站式智能运维平台，中心端采用微服务的方式处理来自多个region的Profiling和监控数据；

**SysAK**：系统工具集，处理来自中心端的请求和数据预处理，进行诊断结果汇总和Profiling、监控数据上传；

**Coolbpf**：eBPF开发框架和功能库，提供持续剖析数据采样，网络、IO、内存、调度及GPU指标等功能。





## 02. 持续剖析：AI火焰图

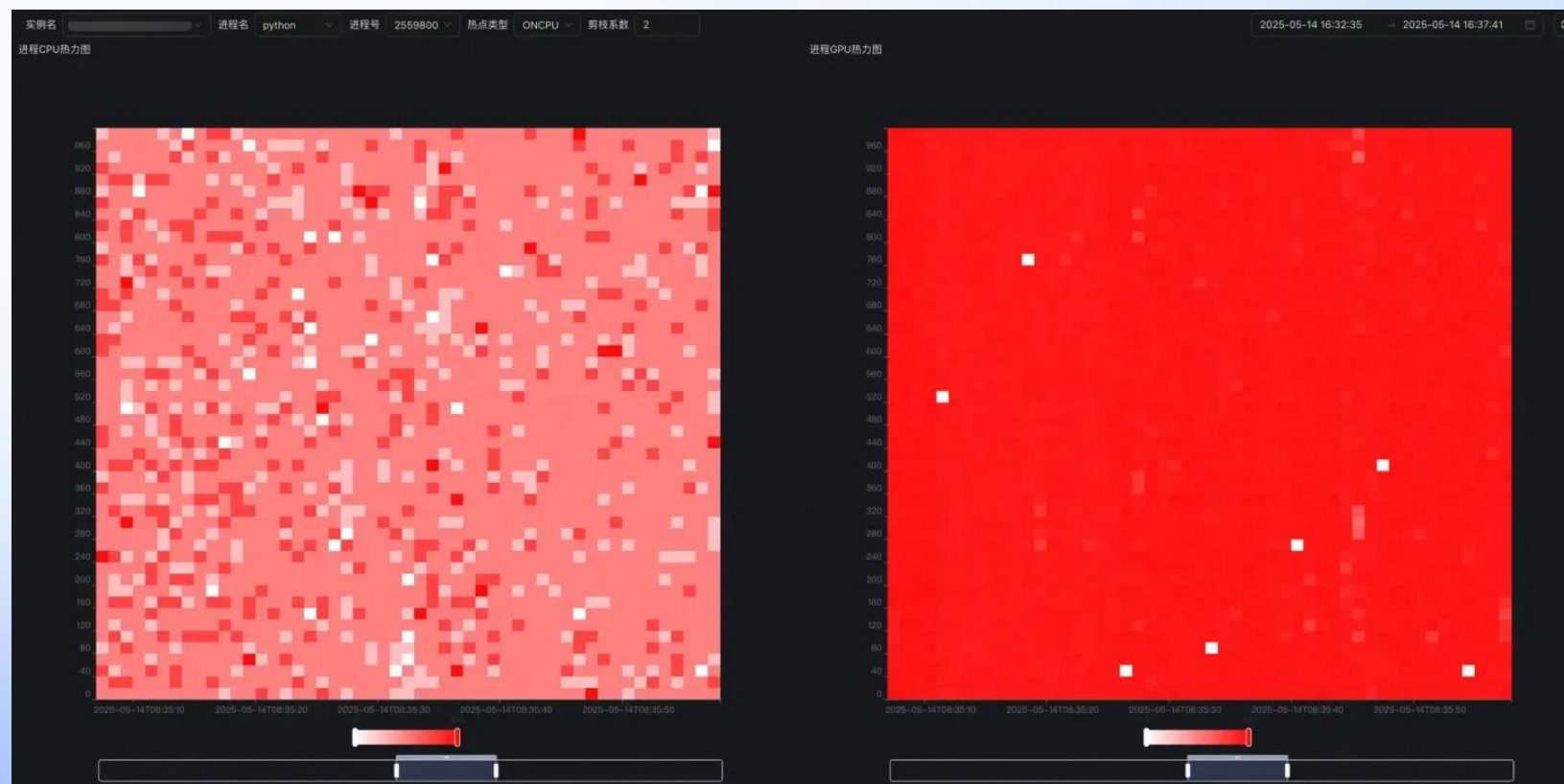
AI火焰图：补齐AI场景下  
Continuous-profiling缺  
乏GPU信息





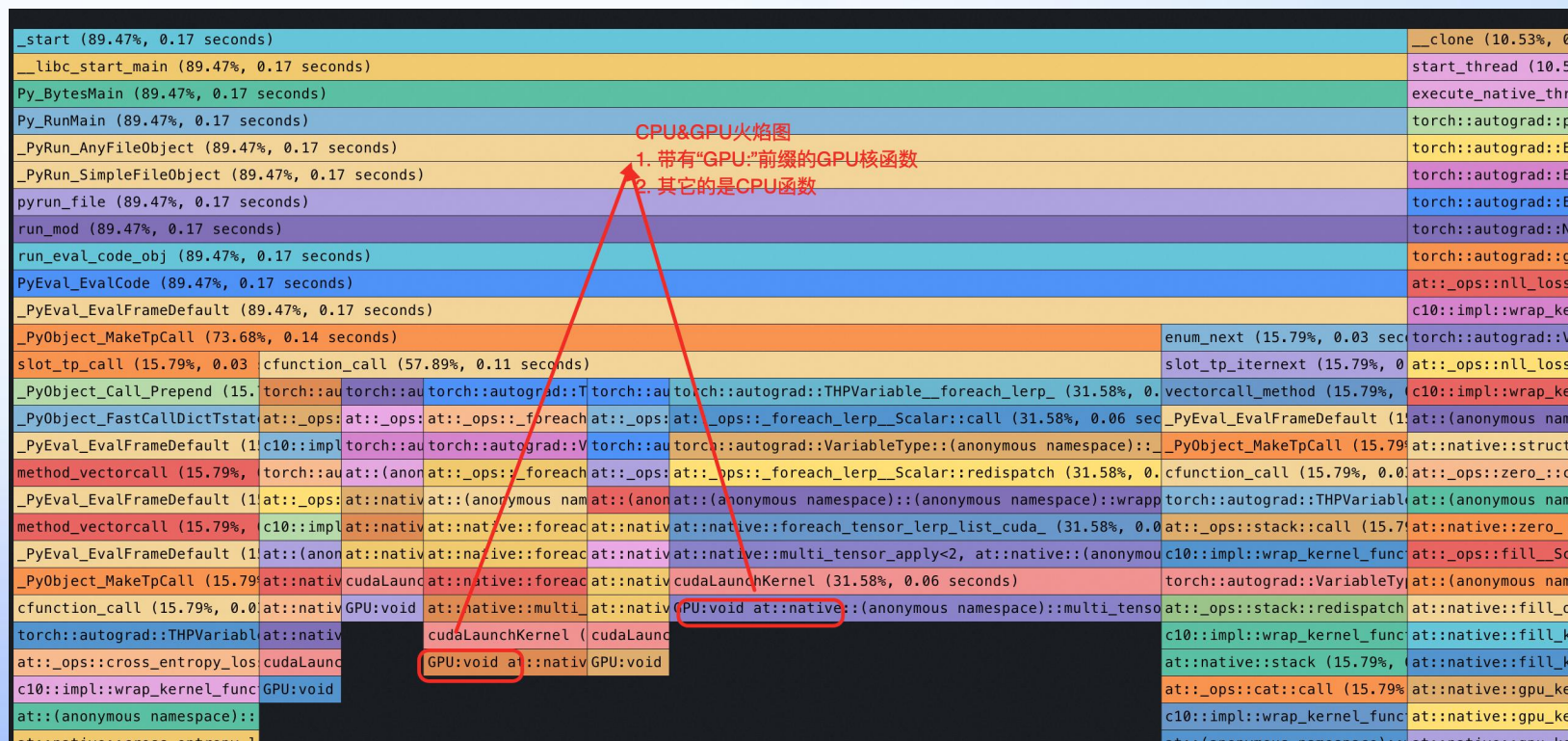
## 02. 热力图

- 每一列表示1秒钟的时间
- 每列由50个小方格组成，每个放个代表20ms的时间区间
- 方格颜色深浅表示归一化后的采样事件数量
- 颜色越深，表示该时间段热点越高



## 02. 火焰图

- 带“GPU:”前缀的条目表示GPU核函数
- 方格宽度代表GPU核函数执行时间
- 通过bpftime完成用户态函数高性能跟踪

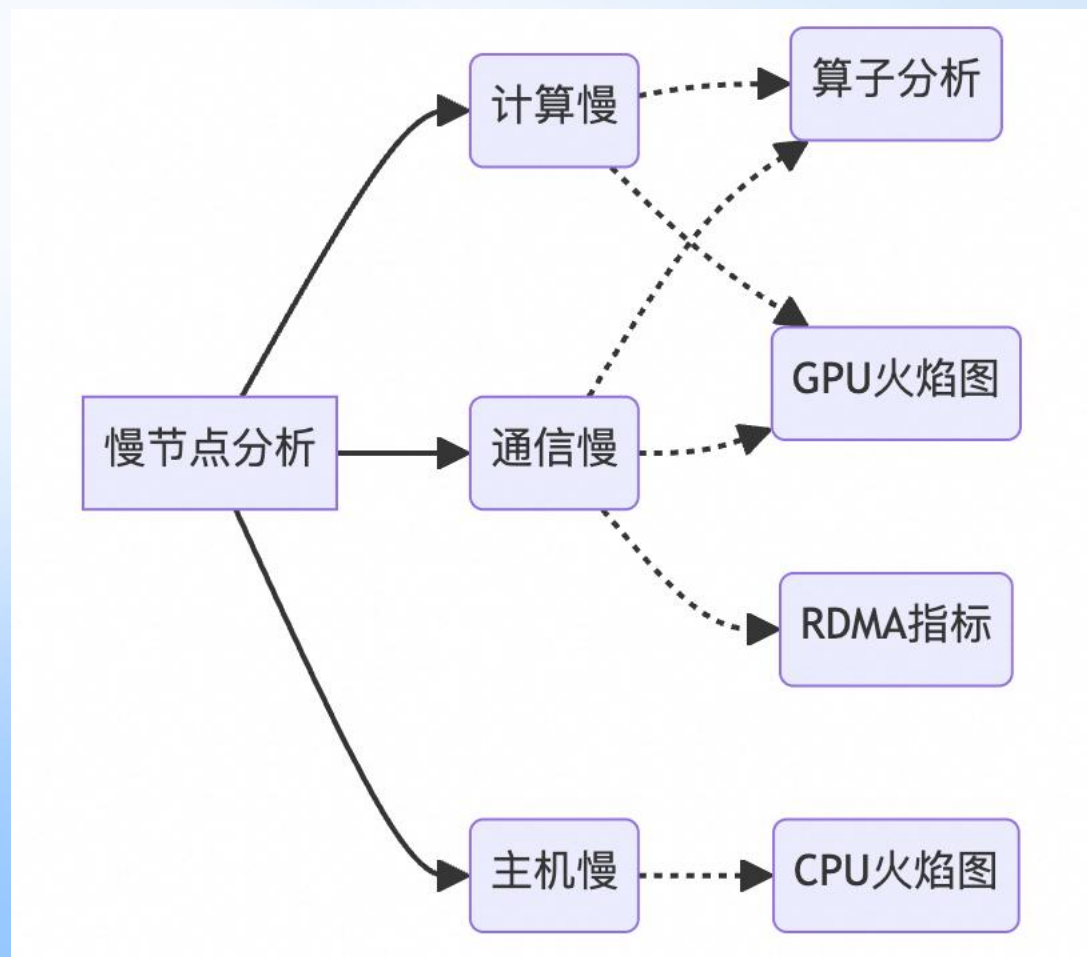






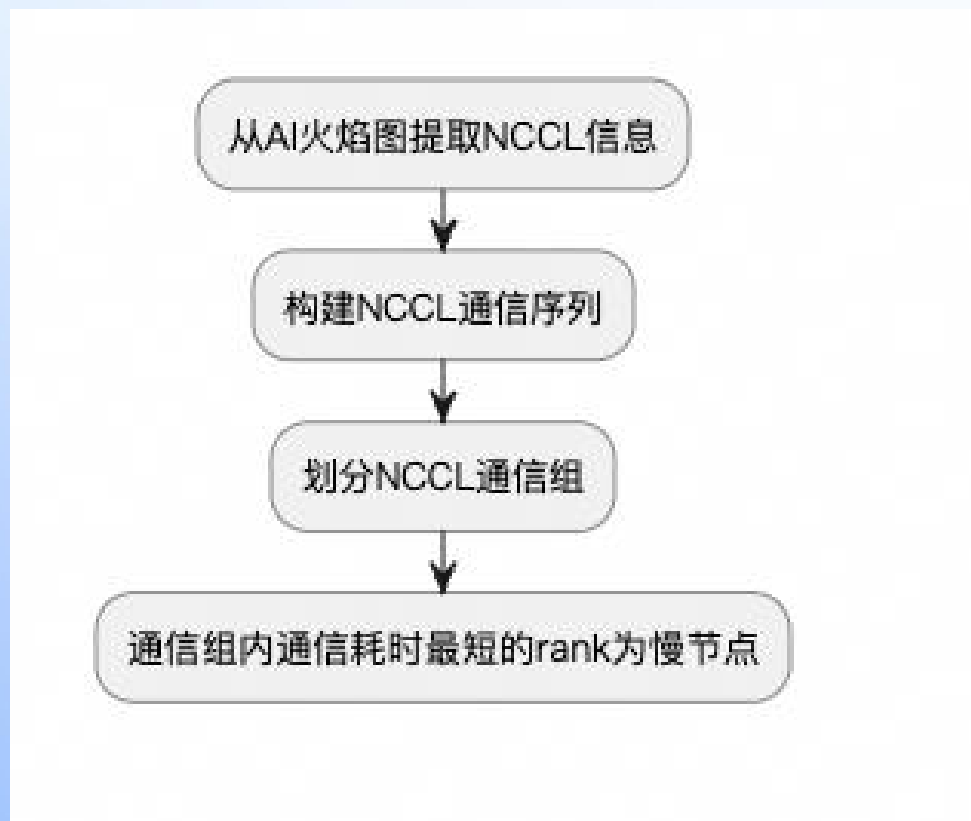
## 03. 场景：慢节点检测

- 计算慢：由于节点自身计算性能下降或负载异常，导致其算子执行时间显著延长，无法及时完成本地计算并进入 NCCL 通信同步阶段，从而拖慢全局训练进度。
- 通信慢：因节点自身通信性能受限或负载异常，致使其在 NCCL 通信同步阶段耗时显著增加，导致全局通信延迟，进而拖慢整体训练进度
- 主机慢：因主机端（CPU）性能瓶颈或异常，导致计算与通信算子无法及时下发至 GPU，从而拖慢整体训练进度。



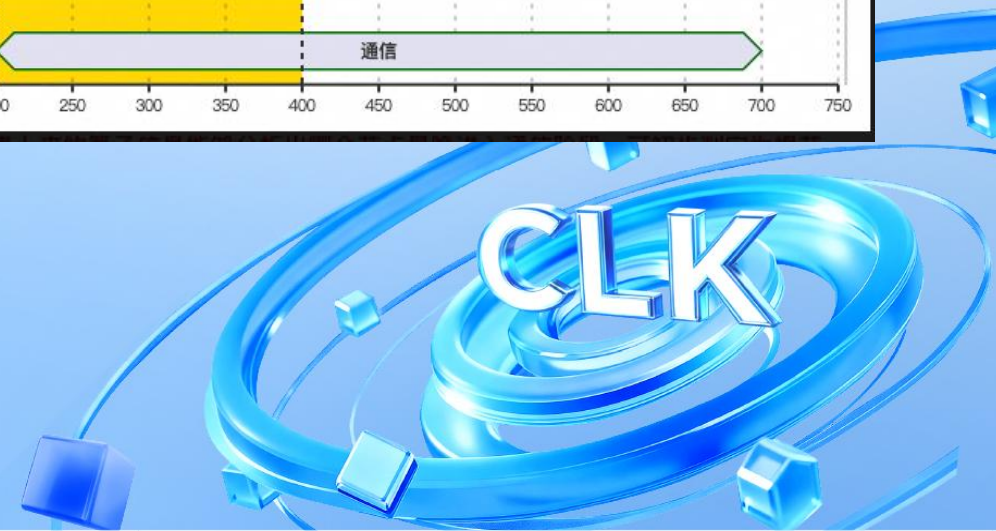
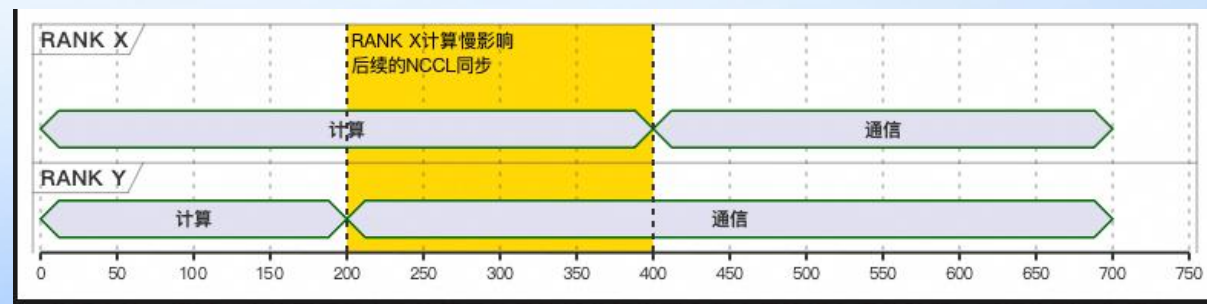
## 03. 分析流程

### ● 工作流程



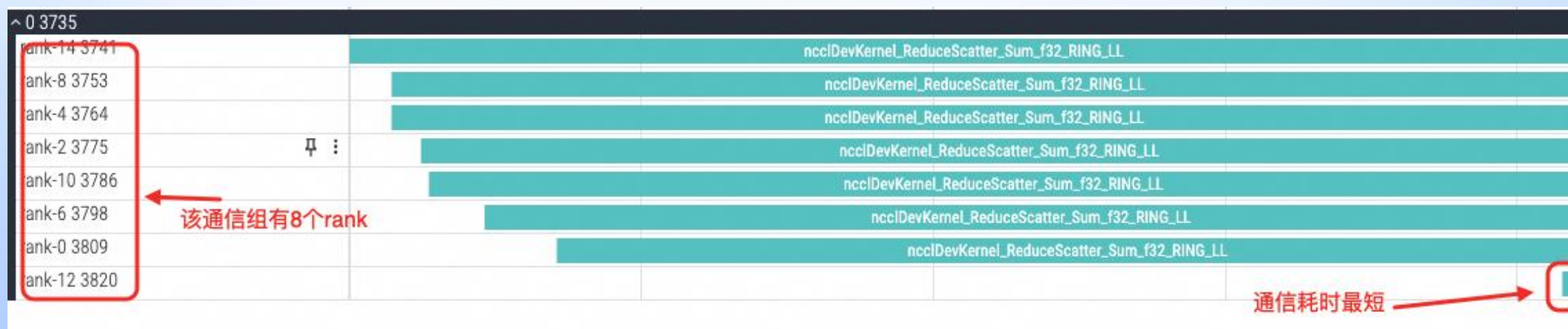
### ● 为什么通信耗时最短为慢节点？

RANK X最晚进入NCCL通信。其中黄色区间是RANK Y等待RANK X进入NCCL通信的等待时间



## 03. 实践案例

结论：rank-12最晚进入NCCL通信，通信耗时最短，其为慢节点





## 04. 未来展望

- 慢节点检测只能定位到节点，无法定位到根因。

后续规划：通过正常rank和异常rank的AI火焰图差分分析来做进一步定位。

- 火焰图分析需要专家知识，门槛高。

后续规划：大模型辅助分析，建立知识库，相同问题匹配。

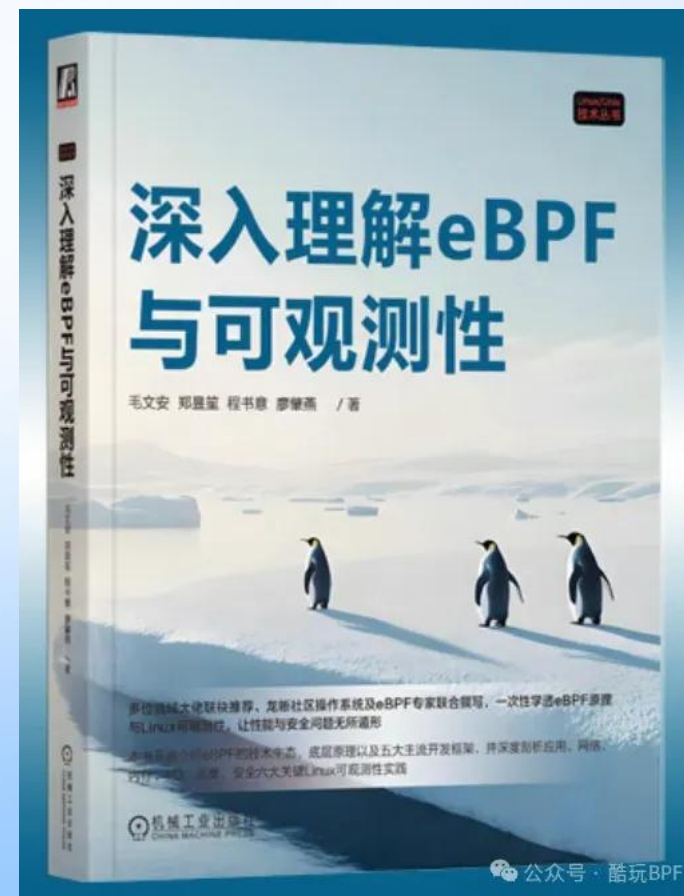


# .好书推荐

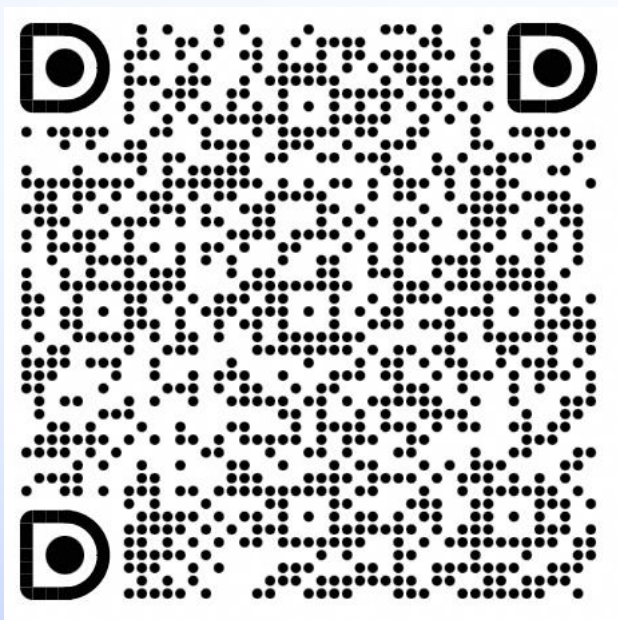


酷玩 BPF

微信扫描二维码，关注我的帐号



# 项目推荐



阿里云操作系统控制台用户交流群  
( 钉钉群号: 94405014449 )

项目链接:

SysAK

<https://gitee.com/anolis/sysak>,

SysOM

<https://gitee.com/anolis/sysom>

Coolbpf

<https://gitee.com/anolis/coolbpf>

阿里云操作系统控制台

<https://alinux.console.aliyun.com/>





THANK YOU

感谢您的观看