

QCon
全球软件开发大会

重塑可观测边界：小红书在大模型时代的稳定性工程实践

王亚普

小红书可观测团队负责人



📍 北京

QCon

全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AI Ops
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

📍 上海

QCon

全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月

5月

6月

8月

10月

12月

📍 上海

AiCon

全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

📍 深圳

AiCon

全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

📍 北京

AiCon

全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LM Ops
- 具身智能

极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播



参会咨询



查看会议

目录

- 01 小红书可观测现状以及 AI 时代面临的挑战
- 02 AI Infra 可观测：面向训推服务的稳定性体系建设
- 03 面向稳定性提效的 AI Agent 场景建设与探索
- 04 未来规划

01

小红书可观测现状以及 AI 时代面临的挑战

小红书可观测体系在稳定性工程中的定位



小红书可观测在 AI 时代的变化

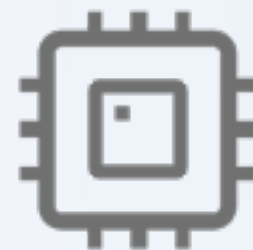
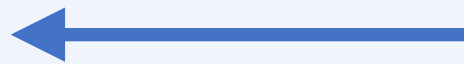
观测对象变化

系统关系变化

运行行为变化

表达方式变化

可观测边界变化



GPU 集群



大模型



训练/推理

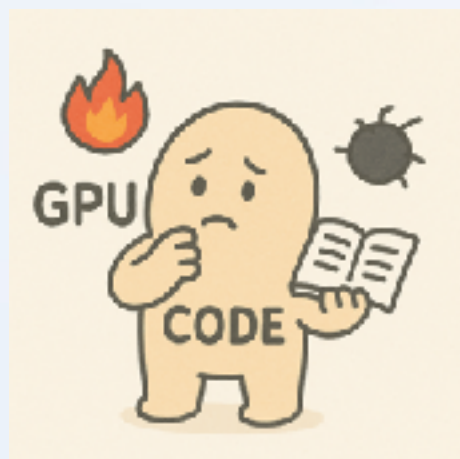


Agent

小红书可观测在 AI 时代面临的挑战



黑盒困局加剧



软硬件的问题多样性



观测对象变化带来的复杂性



故障放大效应

为什么需要 AI 可观测

- AI 训练稳定性管理已经成为智能时代的精密工程，尤其是在千卡甚至万卡以上规模，整个任务会发生各种故障，导致资源利用率不高或任务中断。
- 基础设施交付和运行中的质量保障是稳定性基础，GPU 高昂的成本使得业务对 GPU 交付质量、节点故障感知、处置时效性都要非常高的要求。
- AI 应用需兼顾可用性、性能与效果体验的多重质量目标，使得业务连续性保障面临更大挑战。

02

AI Infra 可观测：面向训推服务的稳定性建设

业务痛点



故障率高

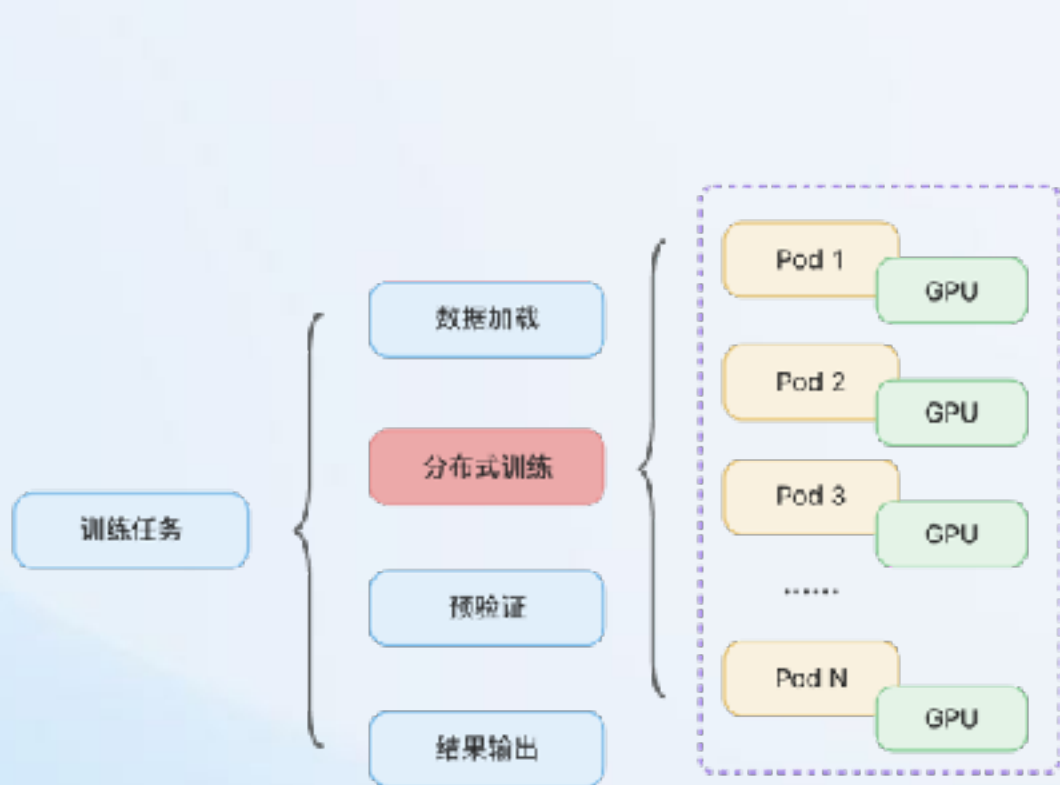


环境异构挑战



故障发现与定位难

训练任务痛点问题分析

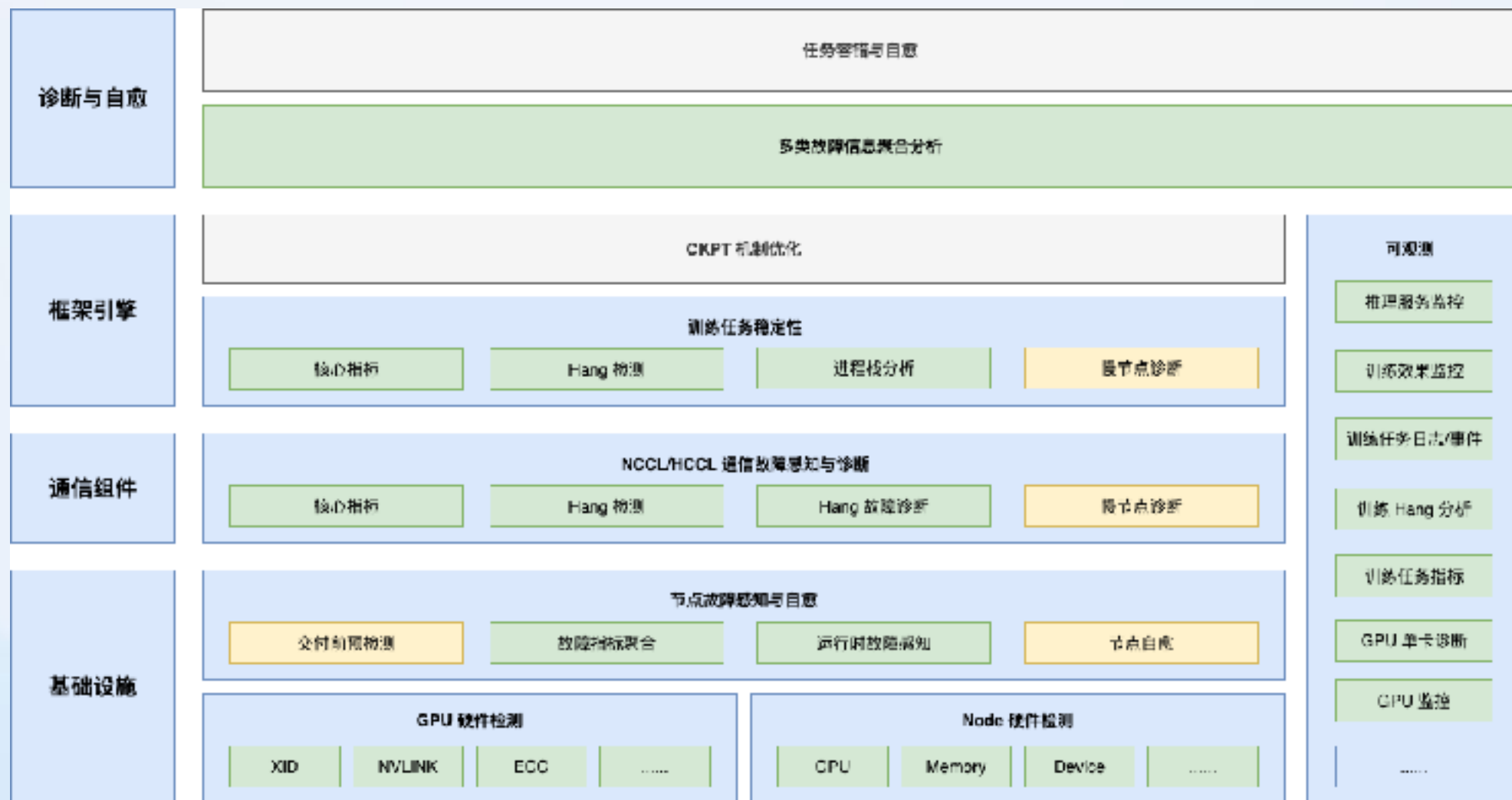


$MFU = \text{Model FLOPs Per Iteration} / (\text{GPU 单卡算力} * \text{卡数} * \text{一次迭代时间})$

减少训练受影响的时间

训练的过程复杂，整个训练过程如果存在一个 GPU 卡异常，整个训练任务就会失败

整体建设思路



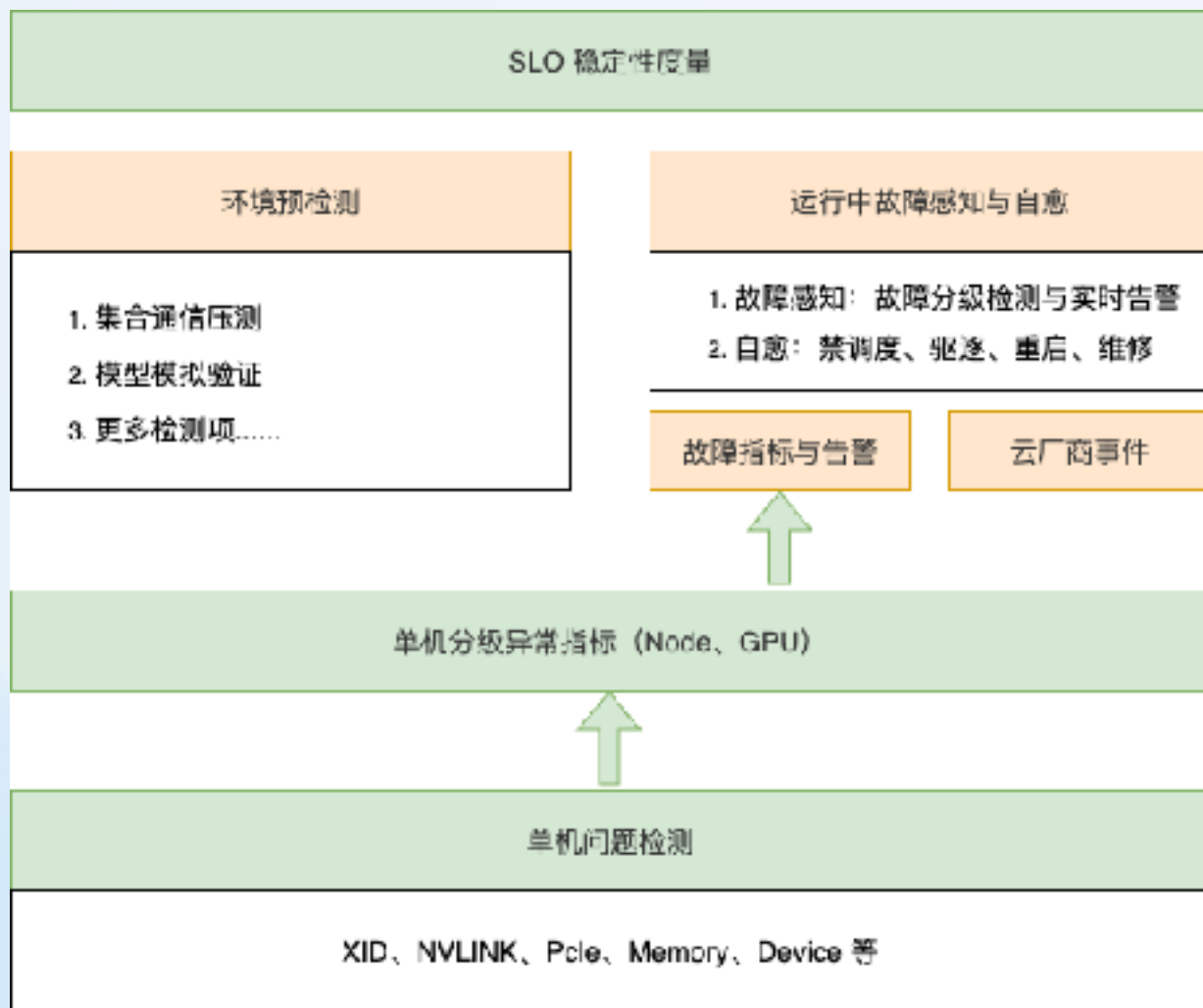
2-1

GPU 资源稳定性保障

GPU 故障等级定义

故障等级	对 GPU 的影响	示例
1 Notice	不影响正常运行	如 Corrected ECC Error、Pending Recovery
2 Warning	可以使用，存在一定风险	如 Uncorrected ECC Error 累计值高
3 Error	无法使用，需要重启恢复	如 SRAM UCE、掉卡、驱动卡死
4 Fatal	无法使用，需要替换/维修	如 Row Remapping Failure

GPU 稳定运营与故障感知



场景支撑

整体异常分析

单节点问题排障

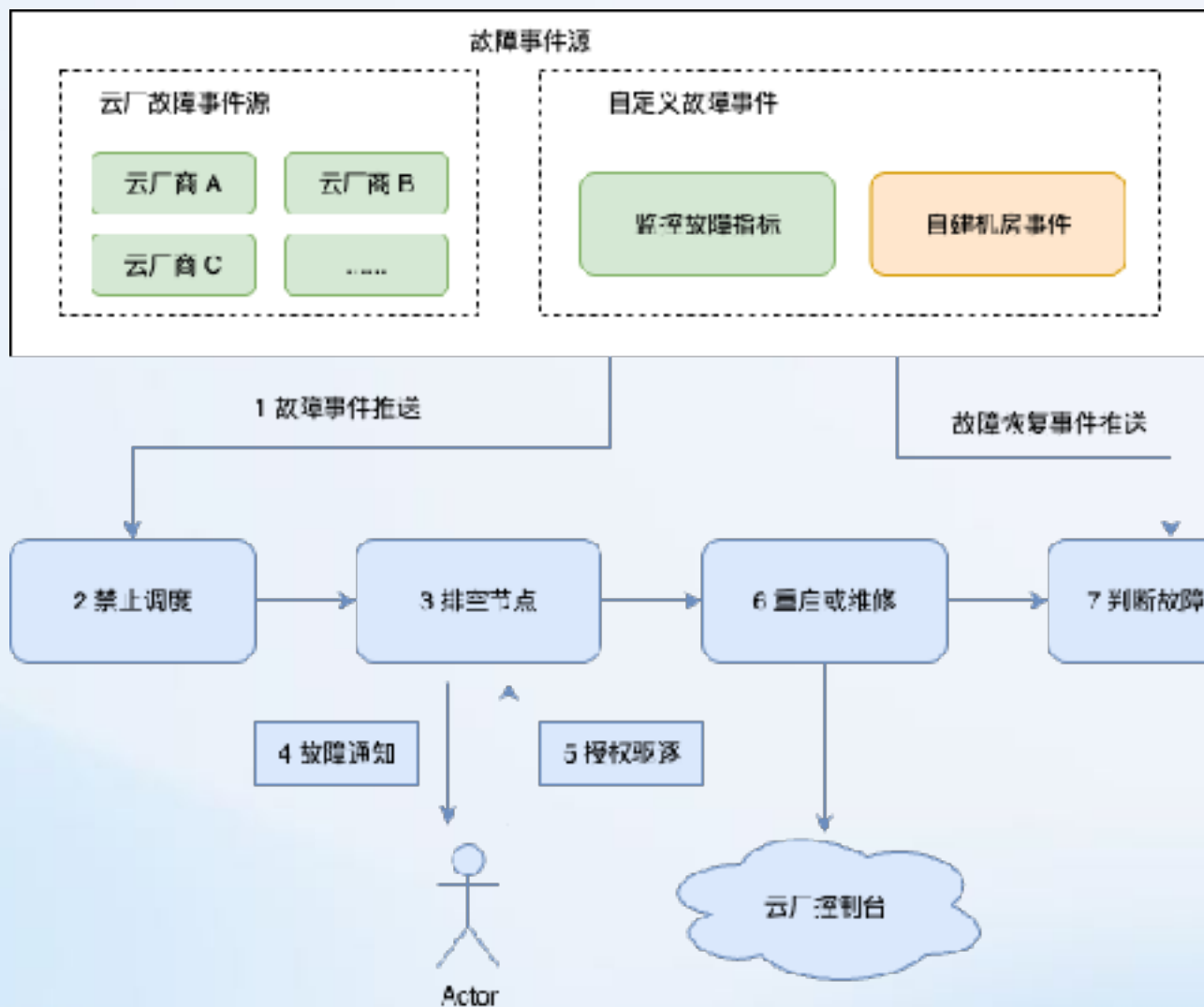
GPU 可用性 SLO

节点/GPU 可用性状态

多维异常度量

单卡诊断/监控下钻

GPU 节点自愈

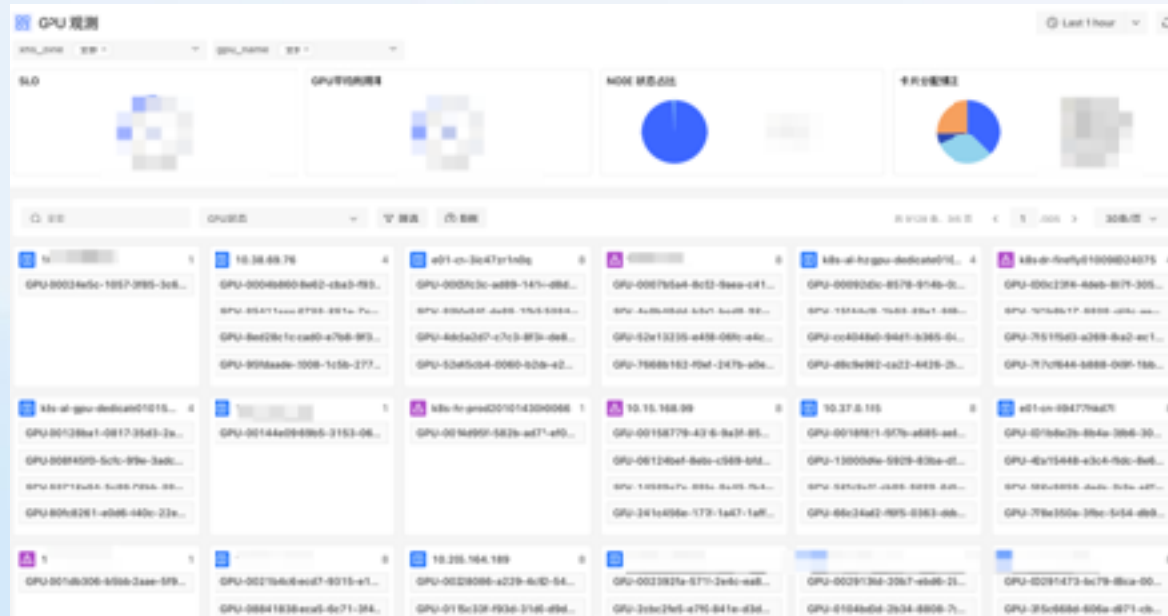


日均触发 5+ 自愈流程

GPU 观测落地成果与最佳实践

提供 GPU 全局监控能力以及单机故障诊断能力，支持 20+ 故障指标识别，可解决 80%+ GPU 硬件故障诊断定位。

整体异常分析



单节点问题排障



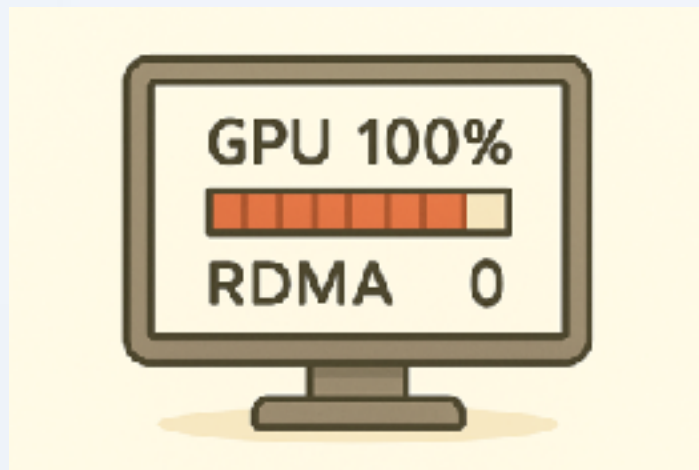
2-2

训练任务 Hang 的发现与故障定位

● 训练任务发生 Hang 可能的表现



所有任务程序日志不输出

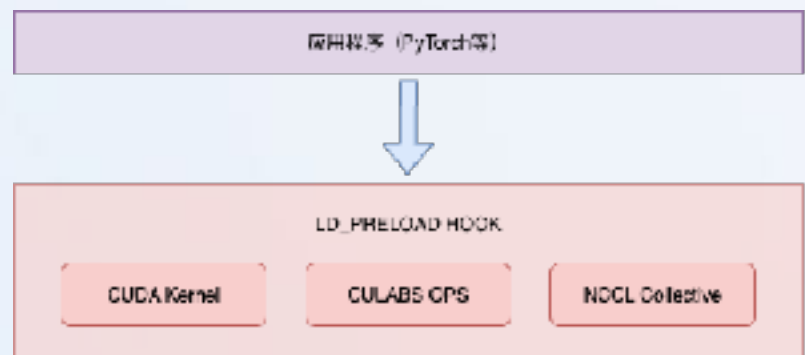


指标突变并持续维持某个状态

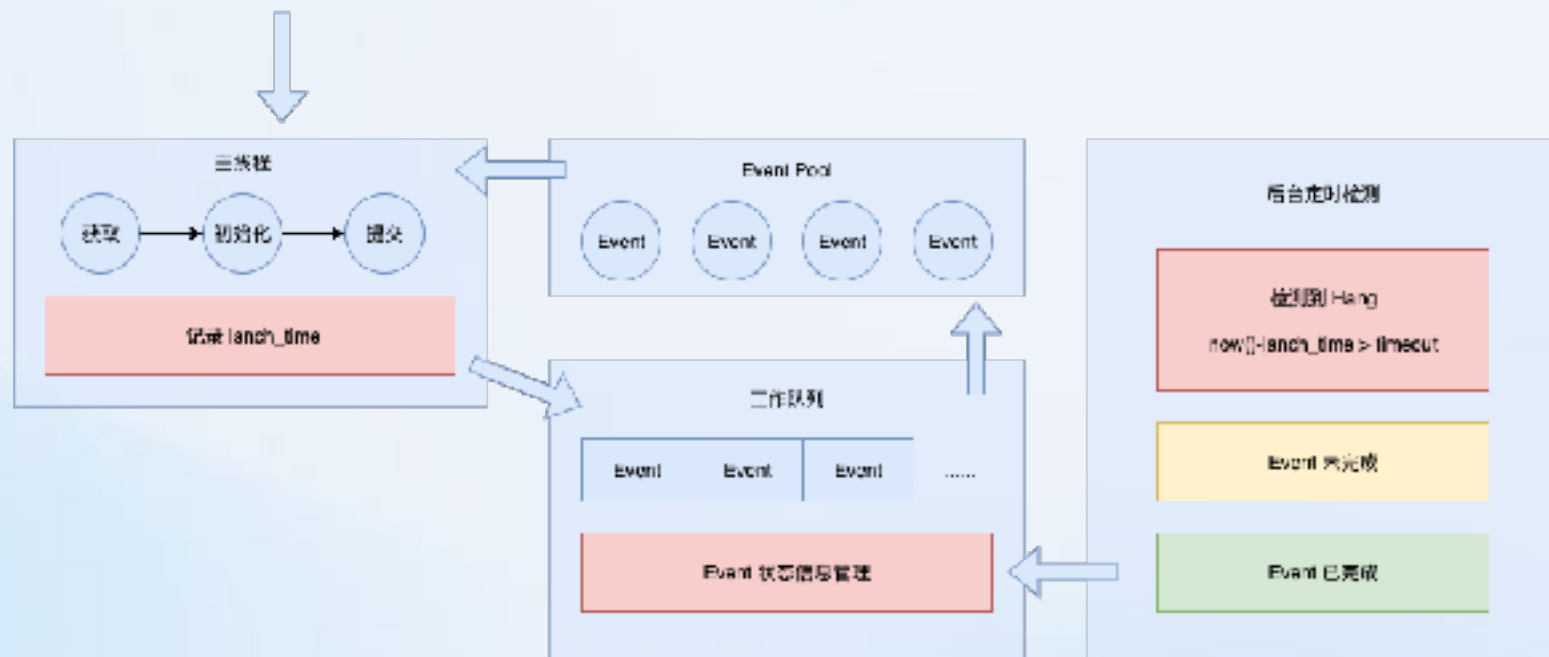


进程调用栈不再变化

问题分析与设计思路



设计的依据：NCCL 和矩阵乘是分布式训练的核心操作，以此为锚点构建相关核心指标统计，不仅可以实现较精确的 Hang 检测，而且可以有效辅助故障定位和性能分析。



判定 Hang 的核心思路：对矩阵计算、NCCL 调用等核心函数进行 Hook 拦截，监控对应的操作记录；当事件超过指定的超时时间未拿到返回结果，判断为 Hang。

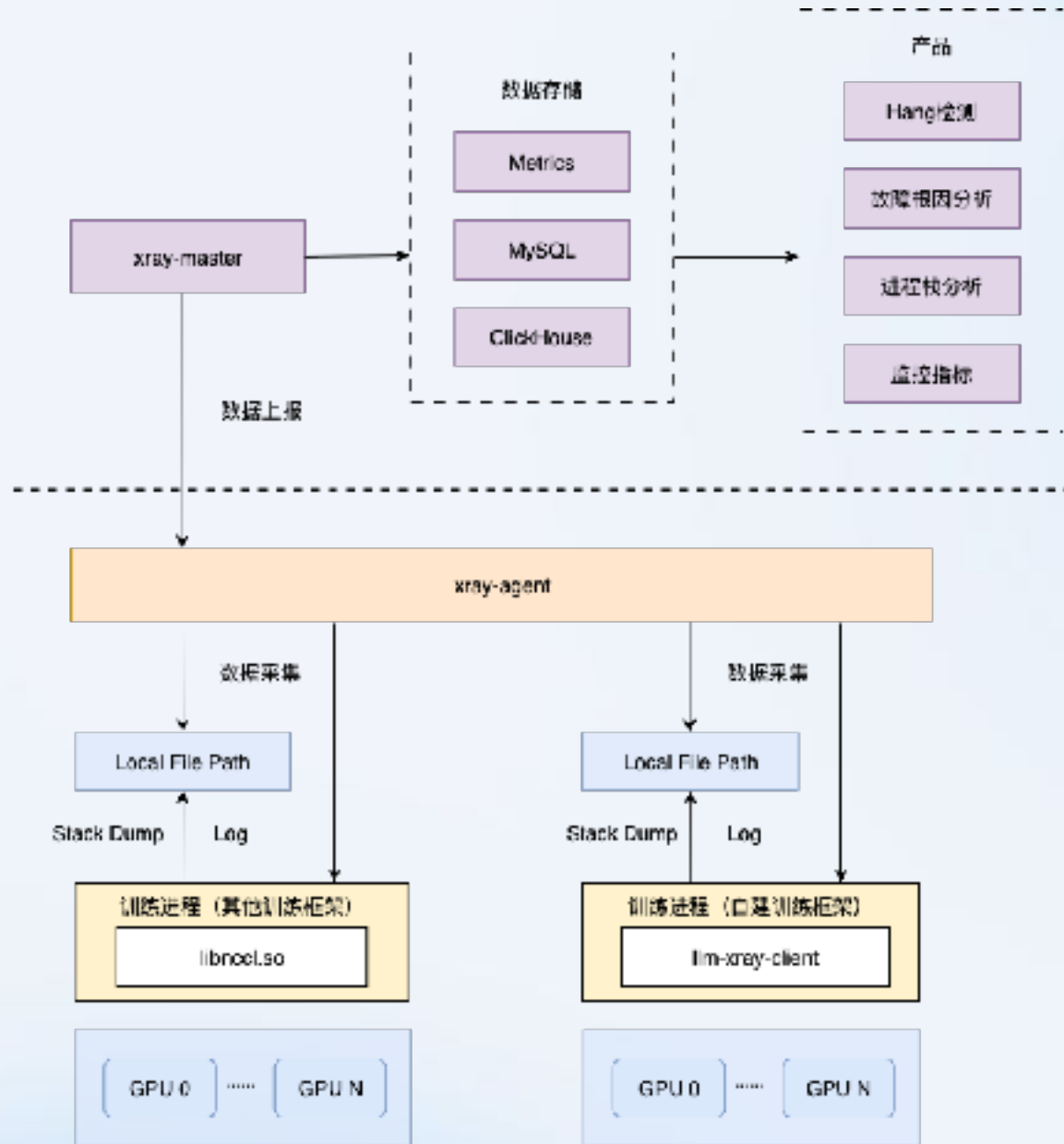
整体技术方案

技术方案差异化选型:

- 自建训练框架: xray-llm-client 低侵入方式, 提供更全面的功能
- 其他训练框架: 仅提供 NCCL 层面的监控和检测能力, 与框架无关

轻依赖的数据采集:

- 保持 xray-llm-client 足够轻量, 复杂功能尽可能上移
- Unix Domain Socket
- DaemonSet/HostPID/进程PID映射关系



故障定位的核心思路（基础场景）

故障表现：

- 受影响节点：表现基本一致
- 故障节点：与其他节点有明显表现差异

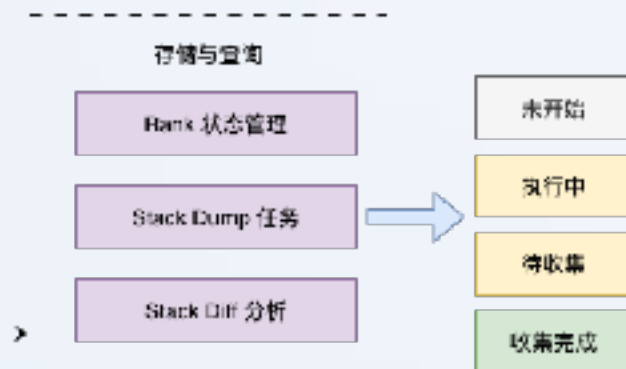
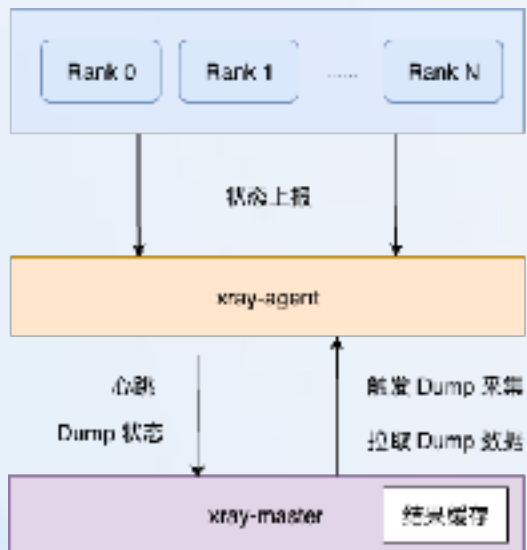


通过对比差异快速定位故障源头



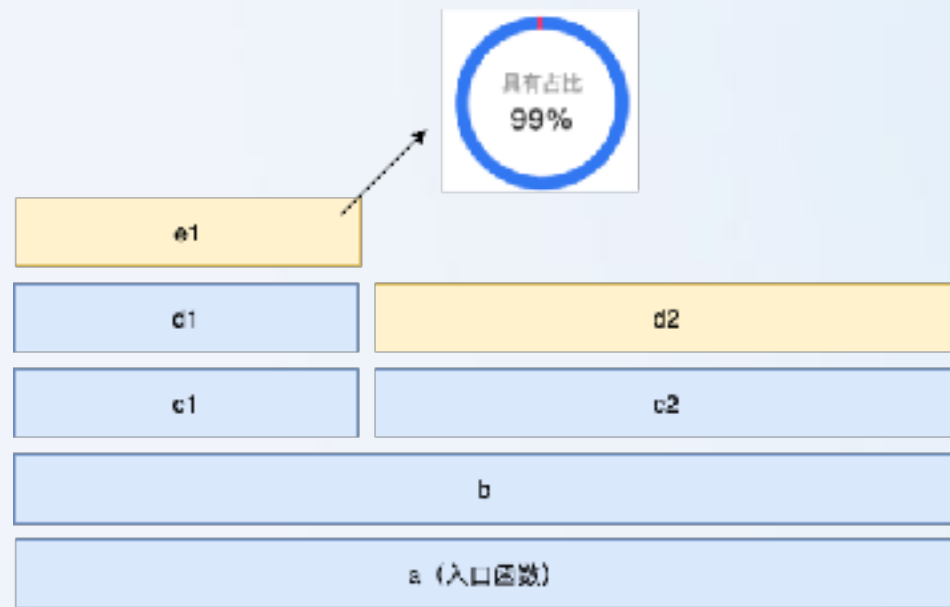
进程栈采集与聚合

栈采集



只要有一个 rank hang，认为这个训练任务 hang 住，
触发一次 stack dump

栈聚合



- 相同前缀、线程名聚合
- 寻找分叉点

故障定位的核心思路（复杂场景）

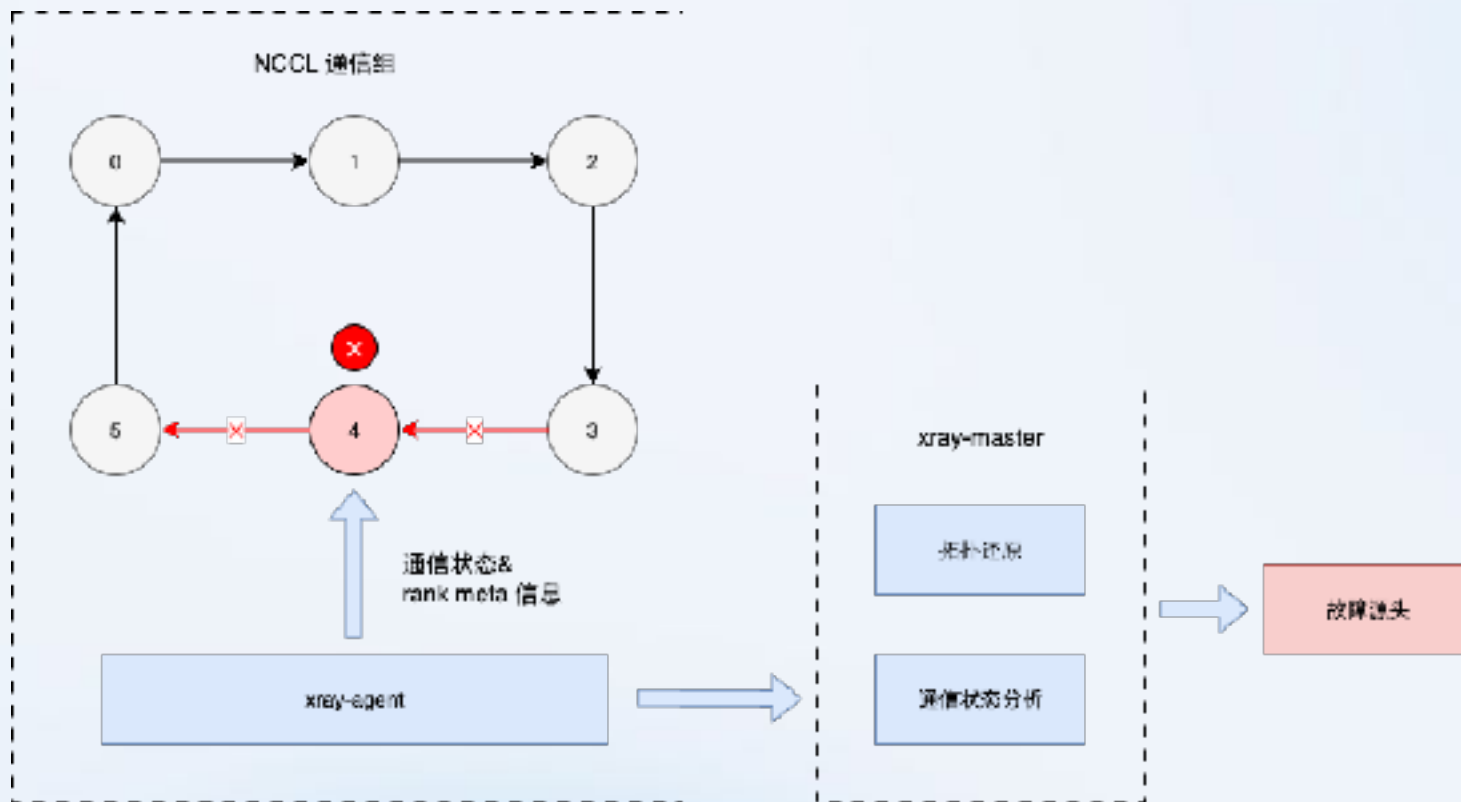
故障源头 rank 和其他 rank，在 NCCL 的通信状态上有差异

故障表现

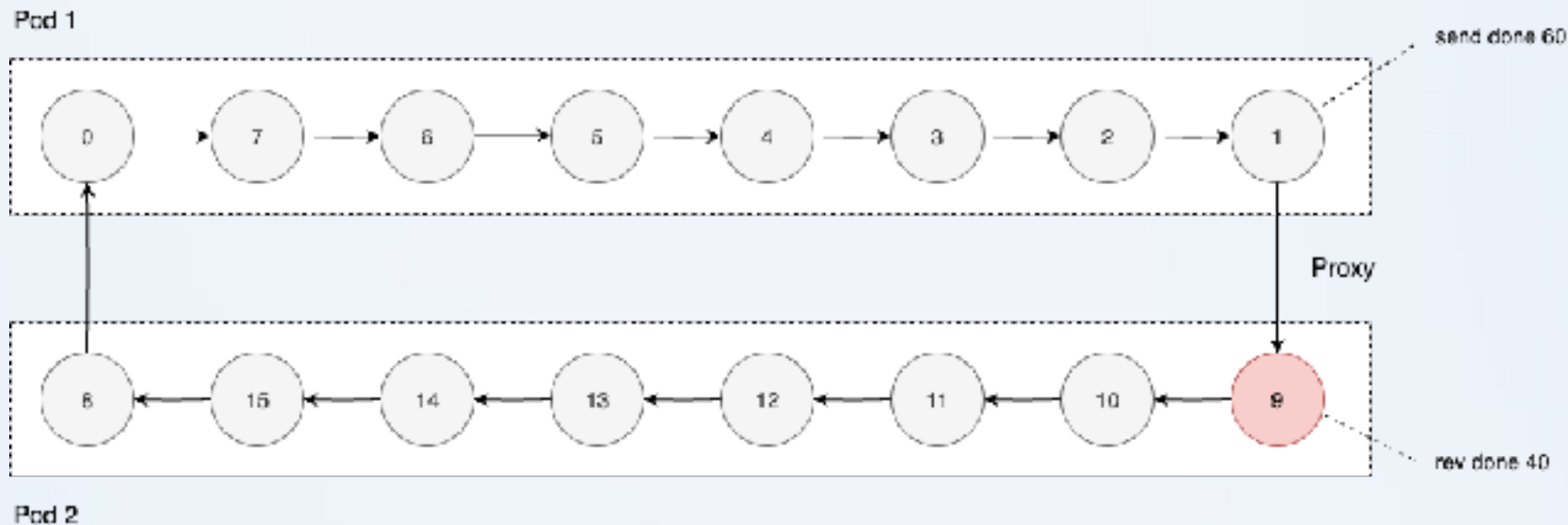
✗ 所有节点表现一致

- GPU全部打满 100%
- 无法从指标区分
- 应用进程栈无差异

⚡ 常规手段失效，需要深入NCCL层面分析



NCCL 网络通信故障的定位与拓扑还原



Send: posted→transmitted→done

故障1: done < transmitted

状态: Sending on network

含义: GPU已生产数据, 网络已提交WR, 但发送尚未完成

Network (RDMA) 故障

故障2: transmitted < posted

状态: Waiting on GPU

含义: CPU准备好了buffer, 但GPU还没提供数据

GPU Device 故障

Rev: done←transmitted←Received←posted

故障1: received < posted

状态: Receiving from network

含义: CPU准备好buffer, 但数据还在网络传输中

Network 故障

故障2: done < transmitted

状态: Waiting on GPU

含义: 已flush完成, 确保GPU收到数据, 但GPU处理延迟

GPU Device 故障

业务实战案例分享：监控指标

训练总览



基础指标



NCCL 通信指标

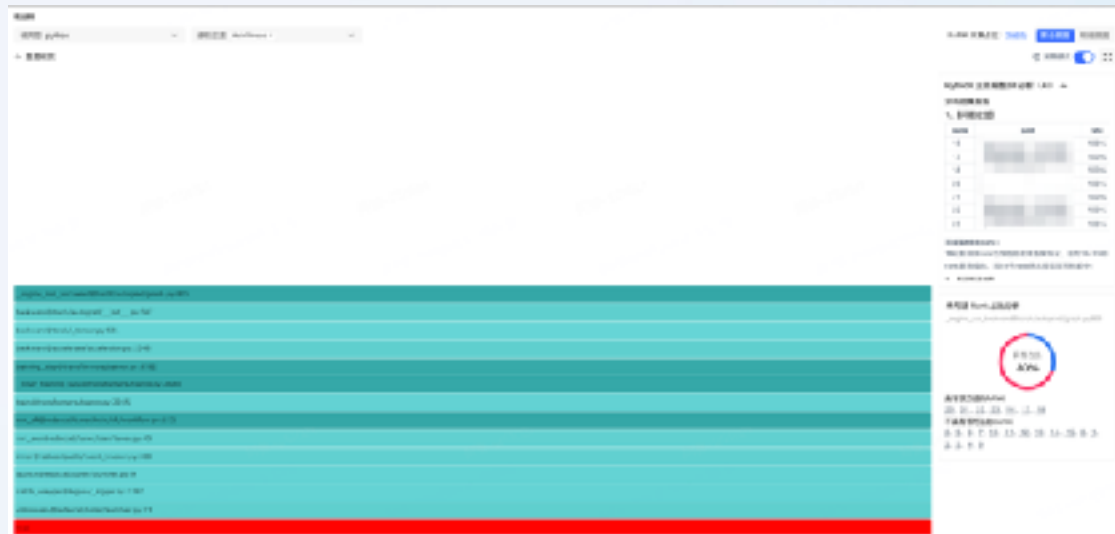


矩阵乘指标





栈分析

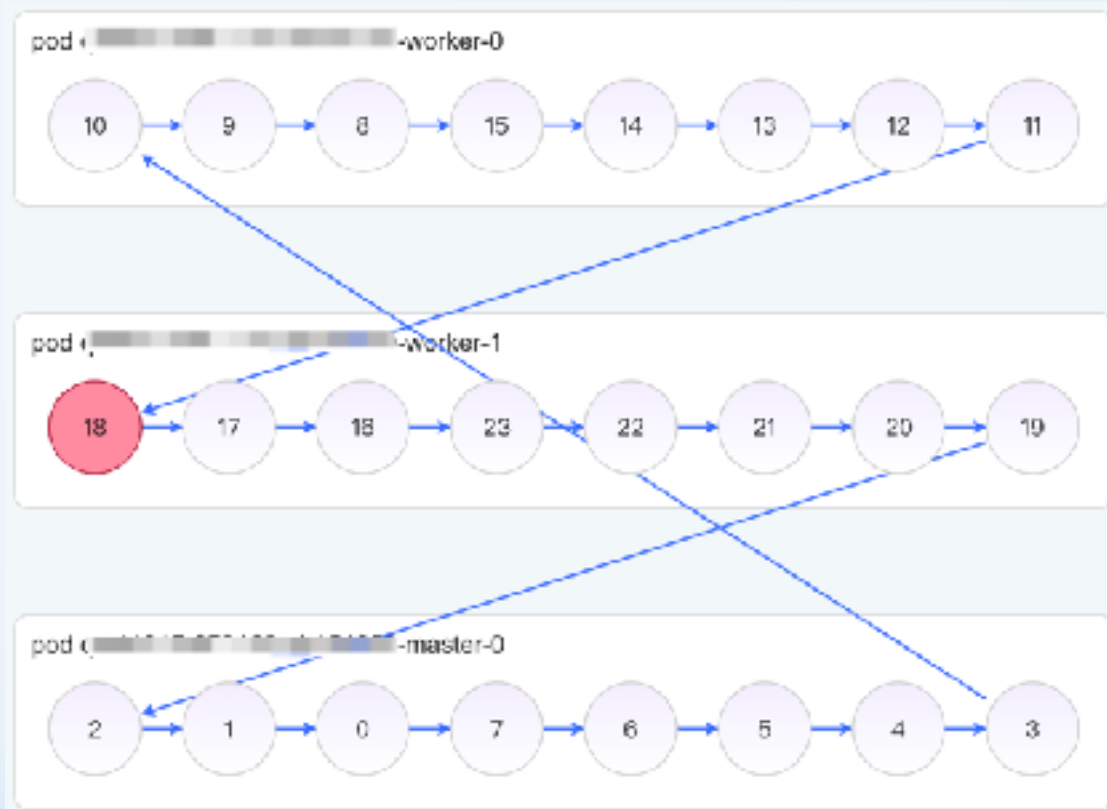


业务实战案例分享：NCCL 网络通信故障定位

tree



ring



03

面向稳定性提效的 AI Agent 场景建设与探索

从可观测 AI 助手开始

XPilot Ai

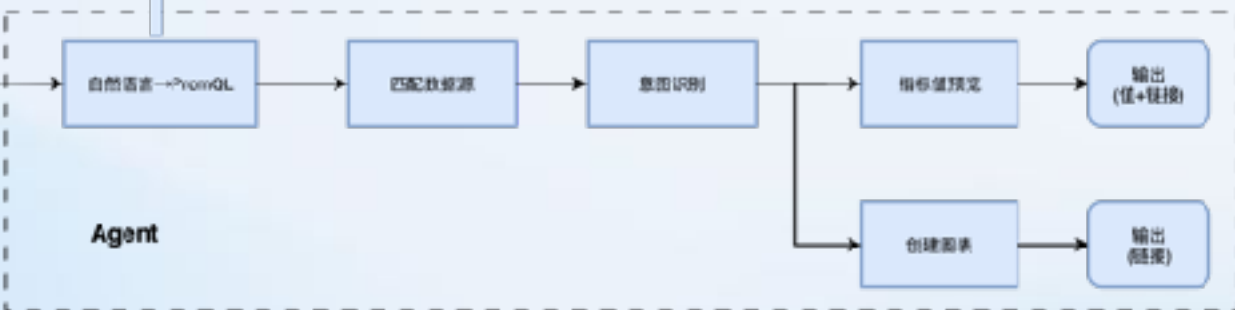
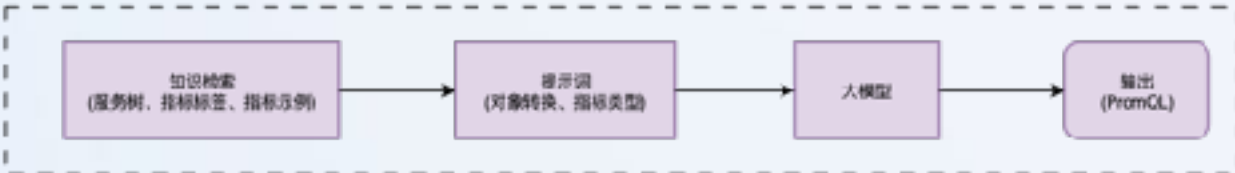
自然语言&配置自动化

数据分析&根因诊断

最佳实践 Workflow

业务自定义场景

Workflow

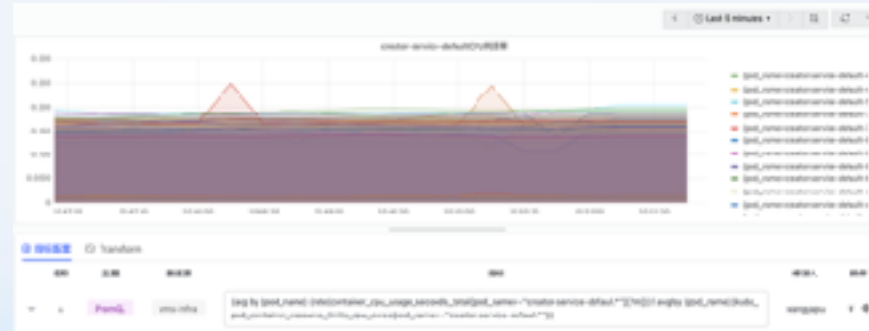


帮我查询creator-service-default服务的CPU利用率



```
[avg by (pod_name) (rate(container_cpu_usage_seconds_total{pod_name="--creator-service-default.%s"}) / avg by (pod_name) (kubernetes_resource_limits_cpu_cores{pod_name="--creator-service-default.%s"}))]
```

Run query



产品定位与目标

以 AI 驱动协助研发与 SRE 完成稳定性全生命周期工作，提升用户效率和体验，降低使用门槛



稳定性 AI 工作台产品大图

角色驱动/高频通用

统一交互/技能封装

团队协同/自主规划与协作

数据标准化与开放

场景层

变更助手

发布助手

容量顾问

可玩性专家

智能诊断

风险专家

知识答疑

更多场景探索

产品层

自然语言交互
(检索/解耦可视化)

AI 工作台
(统一平台)

AI 助手
(随时唤包)

意图识别
多轮对话

AI 分身
(SHE 能力)

AI 技能包
(Agent/MCP/Workflow)

Agent 层

Multi-Agent
(XPilot)

运维助手 Agent

可观测 Agent

智能容量 Agent

配置执行 Agent

模型 & Prompt

策略 & 评测

智能诊断 Agent

知识答疑 Agent

变更/发布 Agent

台鉴 Agent

Prompt 管理

效果评测

模型与算法

小模型+大模型智能底座

数据与知识

数据 (Map Server 标准化)

可玩性

故障平台

HA 可用性

自愈平台

发布策略

容量平台

变更管理

.....

知识库

风险元数据

最佳实践

稳定性规范

问答策略

历史数据库

产品文档

大模型的优势与劣势（现阶段）

大模型适合做什么

语言理解与生成

知识整合与逻辑推理

语义推理与解释

启发式或探索式任务

大模型不适合做什么

精确计算任务

大数据量处理

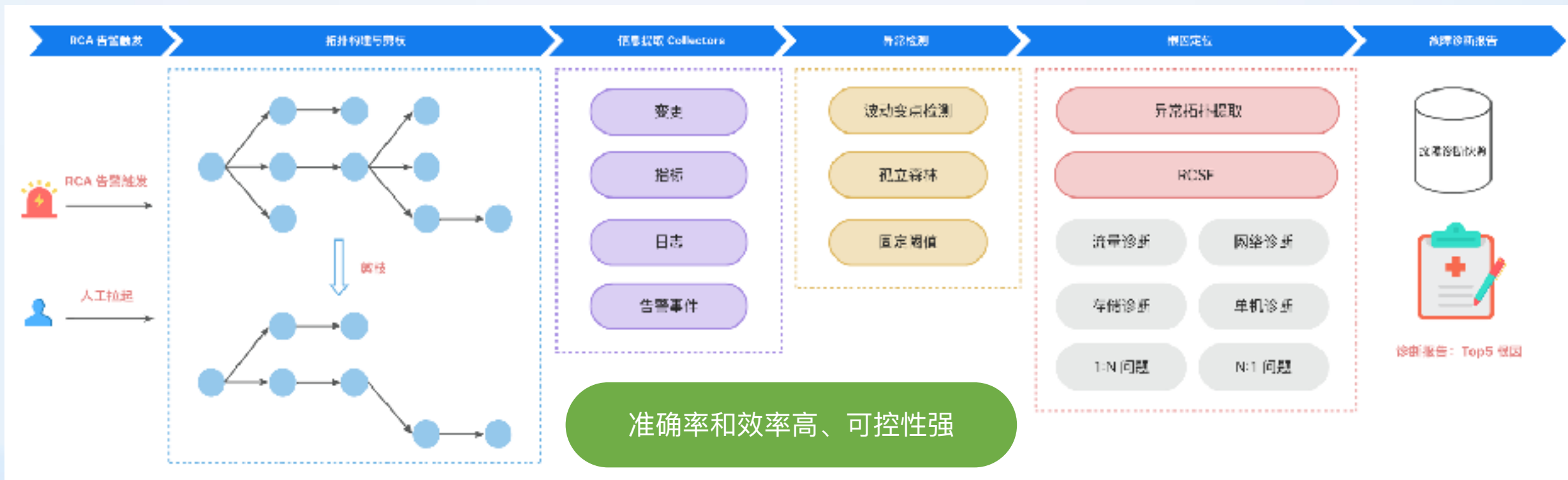
多维聚合与下钻

用户友好的可视化表达

基建角度：需要解决的核心问题

- 原始数据量大，大模型缺乏高效的数据处理能力，如可观测领域
- 跨团队之间的数据交互问题，各团队 Agent 建设处于百花齐放的阶段
- 缺乏对数据上下文的记忆管理能力，无法提供很好的自由对话效果
- 缺少工具动手能力以及用户友好的可视化解释方式

传统 AIOps 的故障定位能力



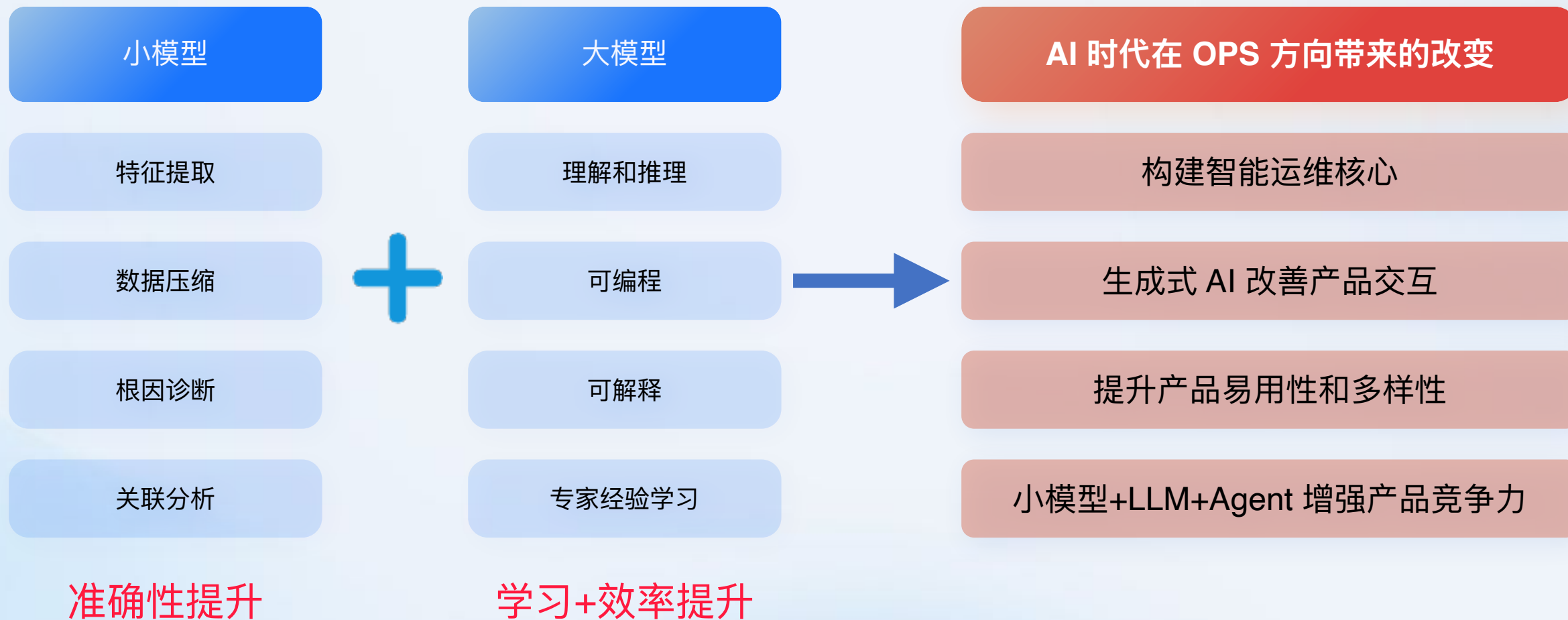
局限性

扩展性有限

专家经验无法复用

可解释性和用户交互不友好

大模型和小模型协同发展



智能诊断底座



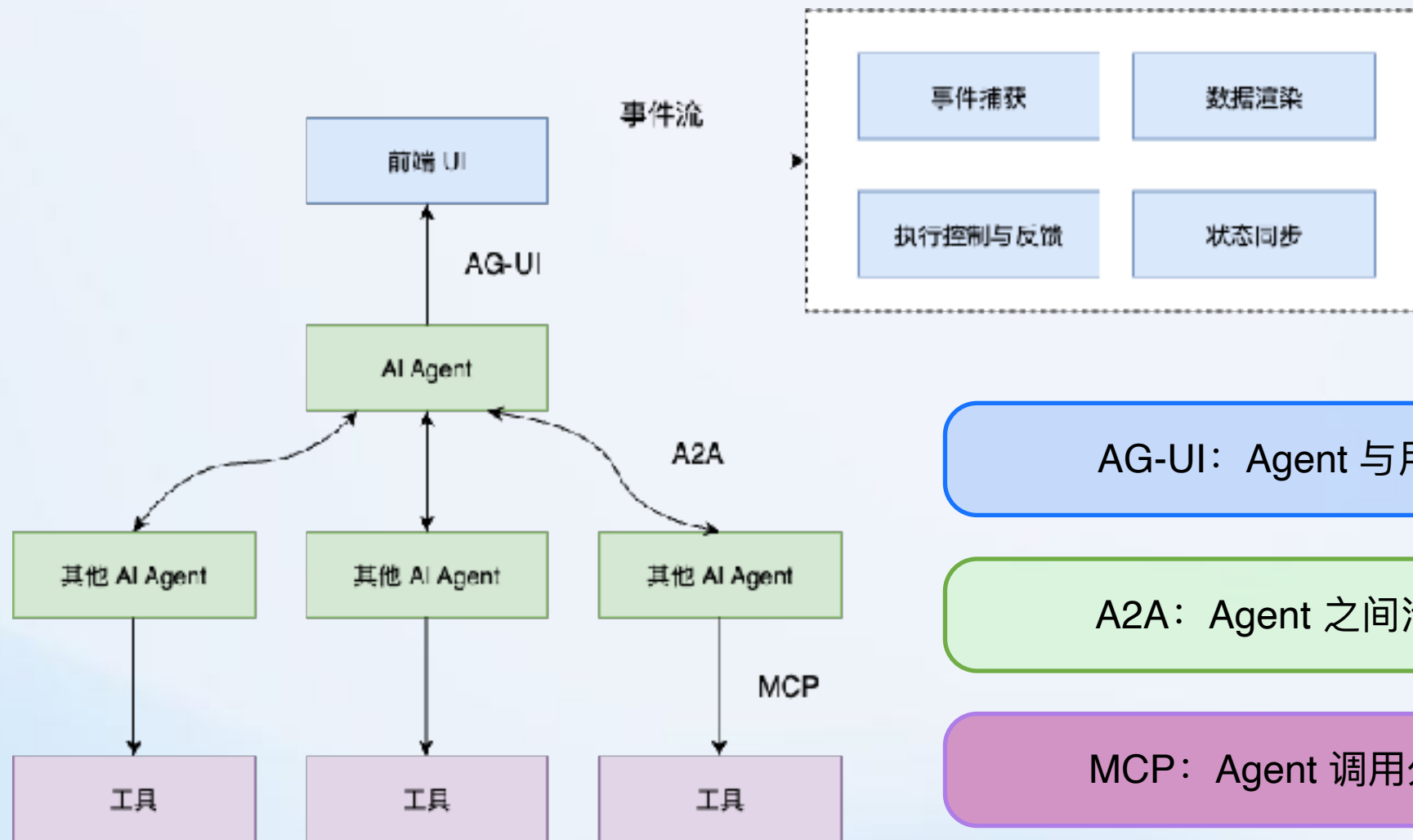
传统 AI 分析与 LLM 深度结合

灵活可插拔的工具调用能力

可复用的智能诊断 Agent

赋能多个技术平台产品

Agent 协议“三剑客”



AG-UI: Agent 与用户沟通的桥梁

A2A: Agent 之间沟通协作的标准

MCP: Agent 调用外部工具的标准

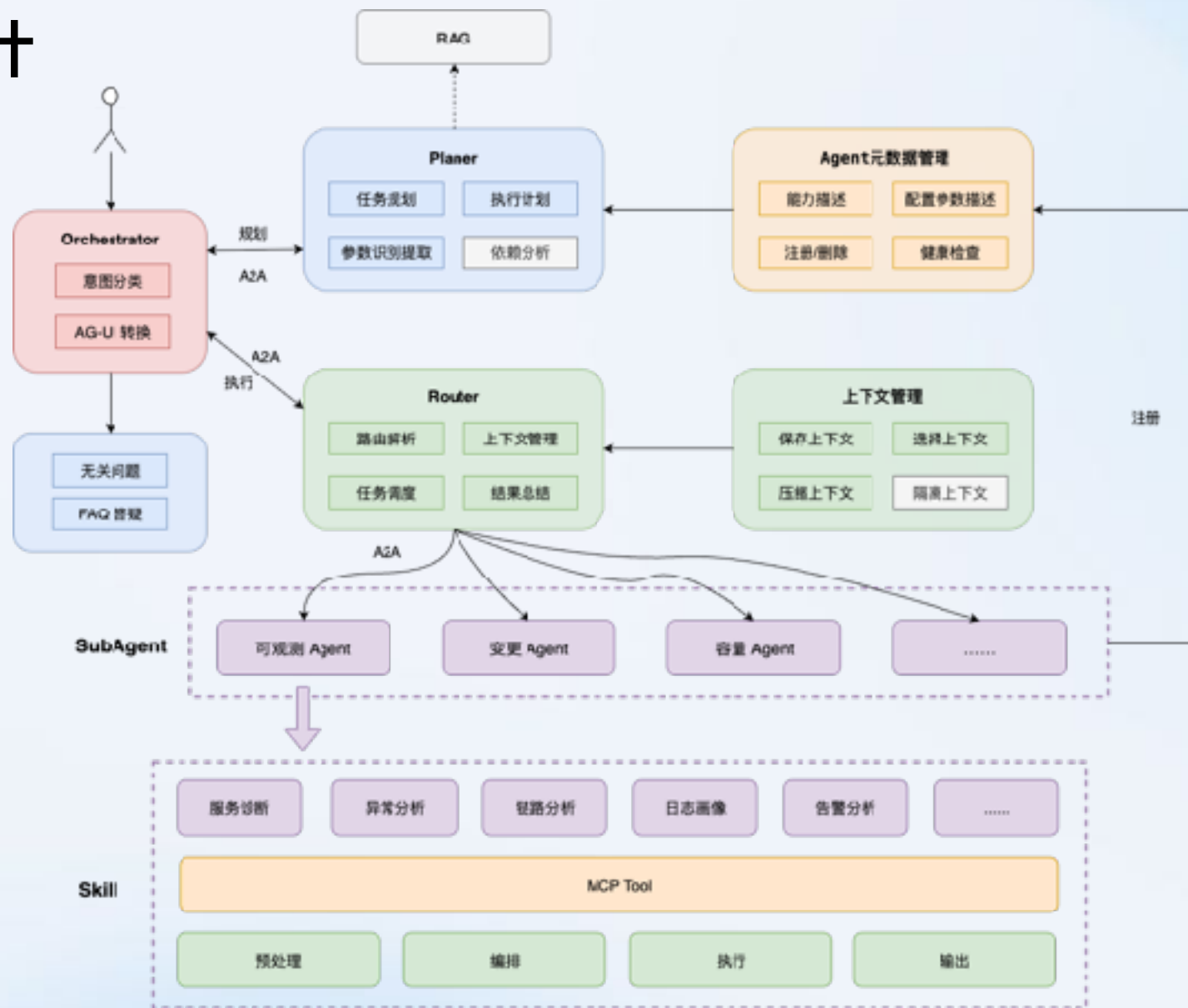
Multi Agent 技术架构设计

中央集权

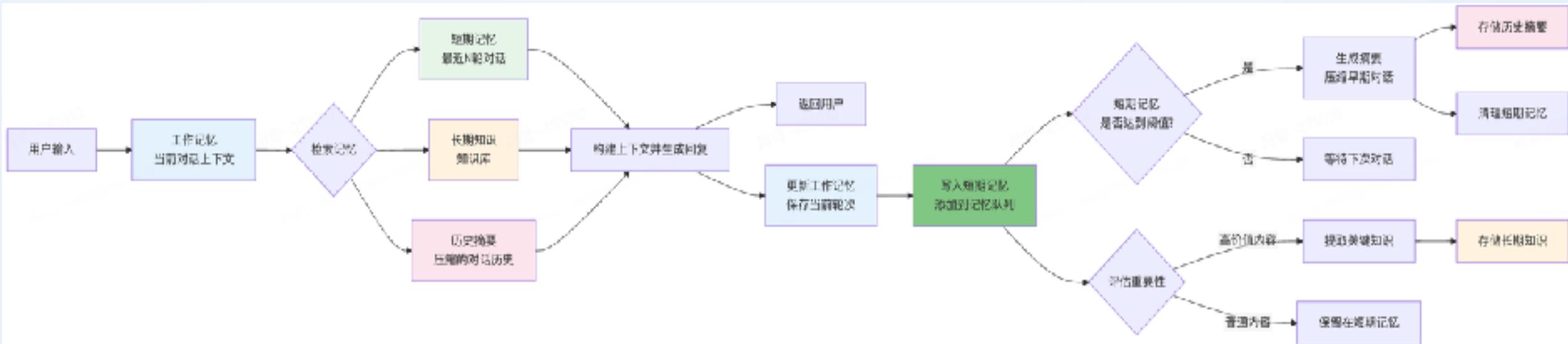
分层委派

分而治之

模块化管理



多层次记忆模块设计



记忆的价值不在于量，而在于信息价值的最大化

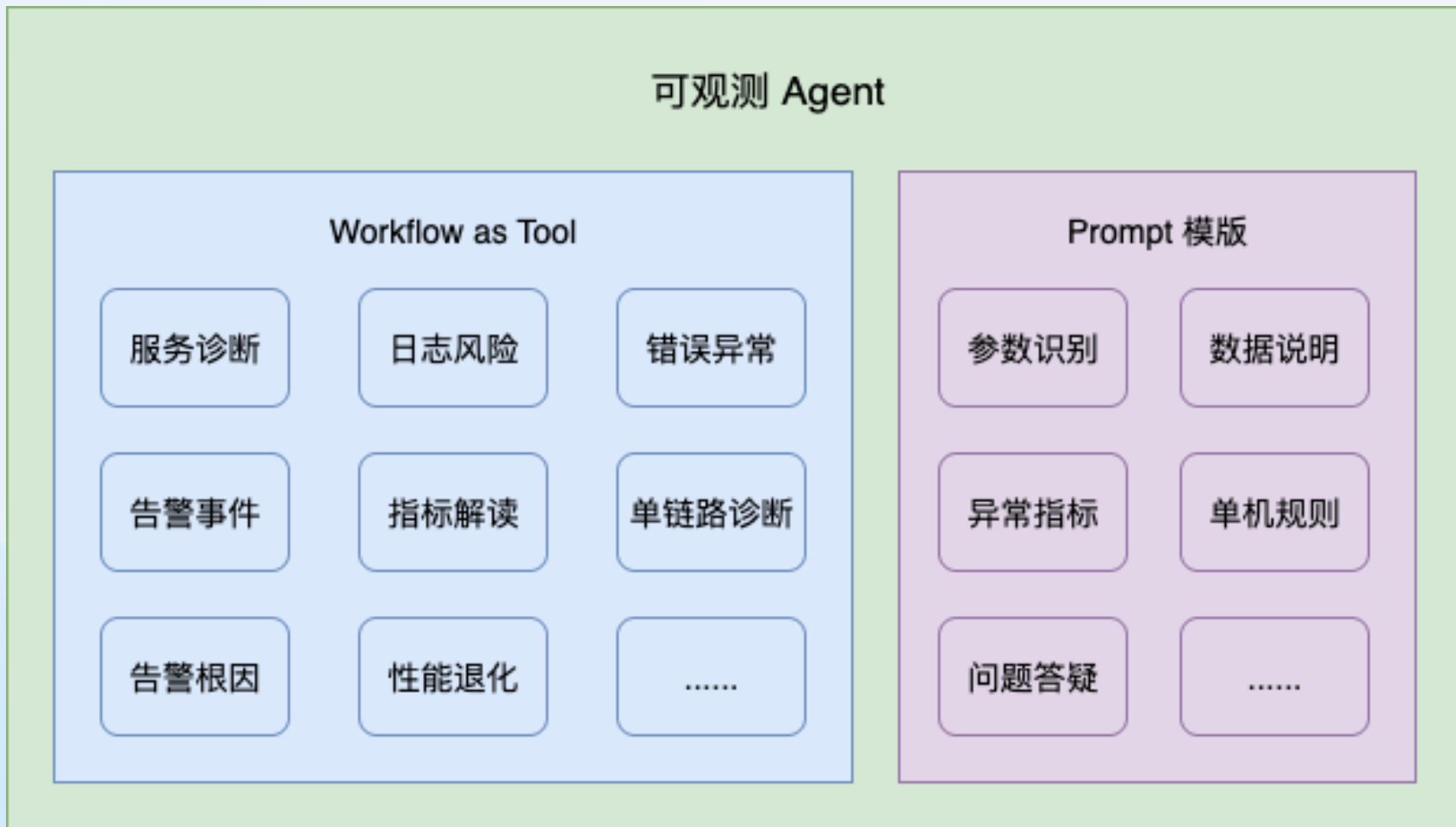
工作记忆

短期记忆

长期知识

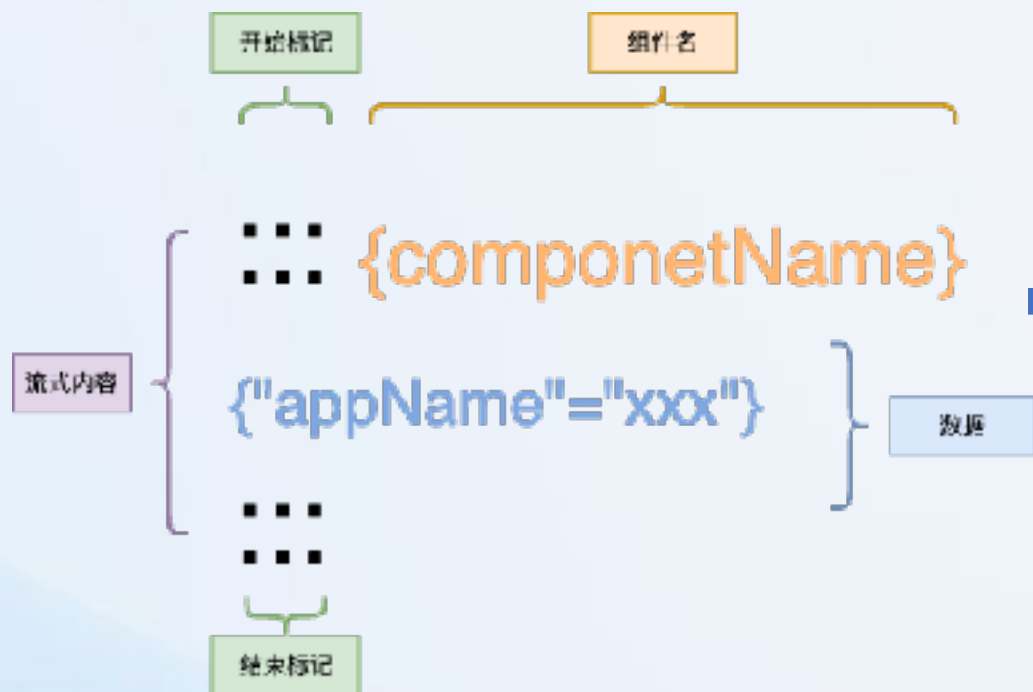
历史摘要

Flow Engineering



- **复杂任务**：拆解，LLM + function 组合，推荐 LangGraph
- 模型在一个受控的、多步骤的流程里工作，而不是一次性“黑箱”输出
- **简单任务**：配置文件，write prompt, not code
- prompt 按场景抽象**可复用模板**

交互与输出范式



参数确认

请确认以下参数信息，您可以编辑参数值：

应用名: [输入框]

开始时间: 2025-10-22 11:10:57

结束时间: 2025-10-23 11:12:57

确认

参数确认



指标诊断

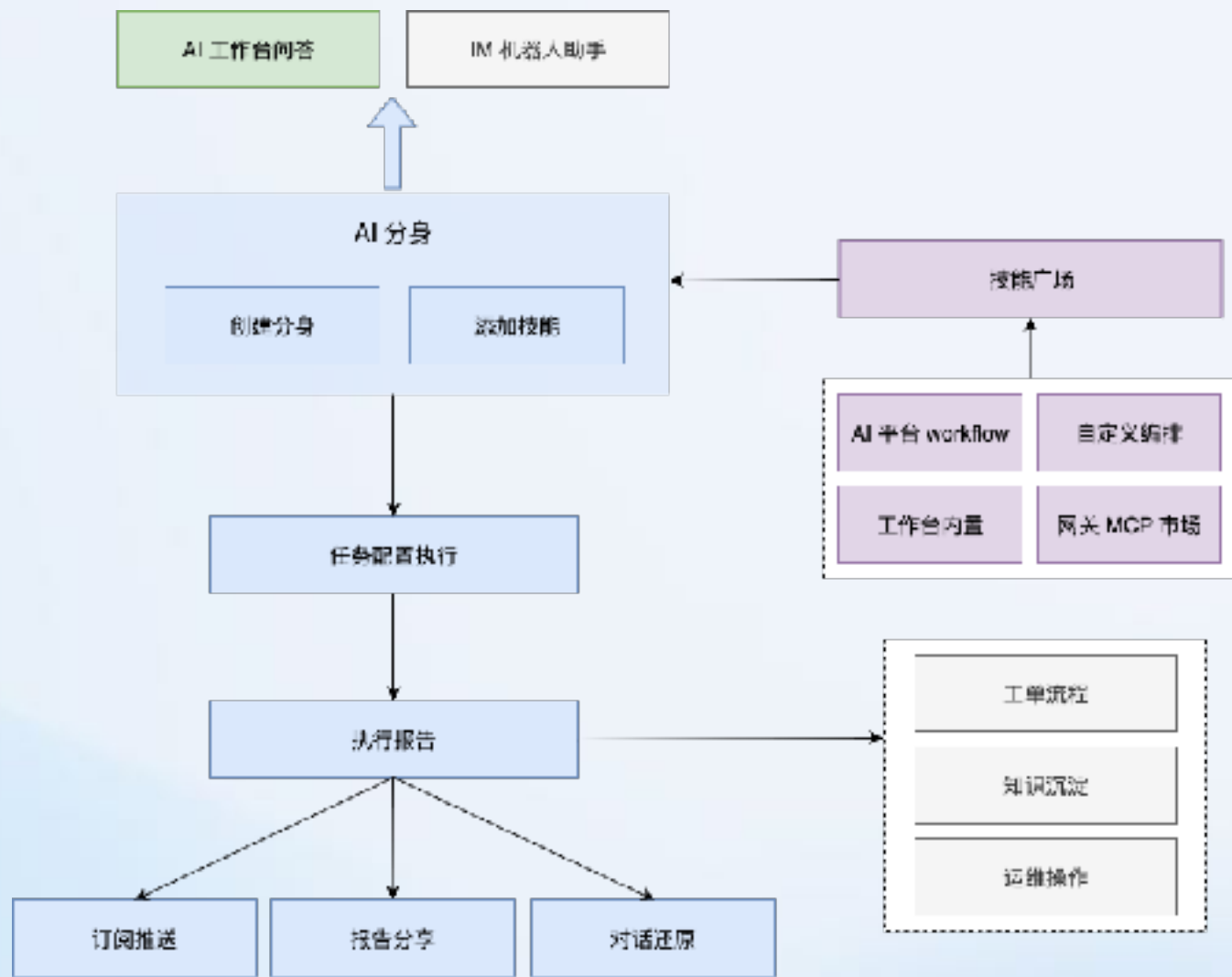
思考过程

工具调用展示

数据下载

.....

业务提效赋能：AI 分身



专家经验转化为可复用 Agent

7*24 小时的专属分身

自然语言+工具构建个人技能

日常工作助手



最佳实践分享：风险治理

日志风险运营

日志风险通知

日志现状摘要

- 服务名称: x
- 风险指标:
 - 当前日志风险速率: 135-60 MB/s
 - 单pod日志基速率: 0.1 MB/s
 - 七层速率: ~2.51%
 - 日志阻塞率: 0%
 - huge级别日志量(一天): 1,472,701 条 (>128MB)
 - big级别日志量(一天): 1,770,994 条 (2889-128KB)
- 治理建议:
 - 每日凌晨的日志采集速率和pod平均采集速率均处于合理范围内，但huge级别的日志量过大，占主资源，以可能导致磁盘成本过高的风险。
 - 日志阻塞率为0%，表明有少部分日志因为超过了单pod的日志速率而被丢弃，虽然目前未达到容量阈值90%，但随着日志量的增长，未来可能会面临更高的磁盘风险。

治理建议

1. 日志内容精简方案

- 建议一：去除冗余信息，定期审查日志内容，移除不必要的调试信息和冗余数据。例如，可以考虑减少或删除重复的日志记录，不必要的堆栈跟踪信息等。
- 建议二：引入日志级别策略，根据日志的重要性调整其级别。对于非关键性的操作日志，可以降低其日志级别，以减少日志的生成和存储。同时，对于关键性的操作日志，保持较高的日志级别，以确保重要信息的记录。

2. 大日志优化方案

- 建议：分析并优化大日志，利用平台提供的Kubernetes日志管理工具，定期清理过期的日志内容。特别是那些超过128MB的日志记录，通过这种方式，可以有效减少磁盘空间的占用，降低磁盘成本。同时，对于关键性的操作日志，保持较高的日志级别，以确保重要信息的记录。

相关资源

- 服务日志规范
- 日志采集规范

风险工单分析

XPolar 00

1173 该条风险工单是什么类型

1. 基本信息

该风险工单 ID: 111726 名称为“集群日志采集速率异常”的规则 触发。结果判定为“失败”，来源于服务“xray-dog-default”，风险类型是“集群日志采集速率异常”。

该风险工单 ID: 111726 名称为“集群日志采集速率异常”的规则 触发。结果判定为“失败”，来源于服务“xray-dog-default”，风险类型是“集群日志采集速率异常”。

2 风险详情

该风险的核心问题在于特定MySQL集群的日志采集速率未达到预期标准。根据数据库容量分析结果，CC级集群容量等级为4（支持同时多机房部署，具备手动/自动切换能力），而当前与两个集群不达标。

当前容量等级为3级，仅满足单机房内部署，无跨机房部署能力，无法应对机房故障（如物理机宕机），风险等级为2025-06-08。

当前容量等级为1级，属于完全单点运行，无任何冗余能力，风险等级为2025-06-13。

请对日志遇到的问题描述...

基本度量

26 KUBERNETES

指标名称 集群日志采集速率异常

度量值

详情

指标类型

风险

风险等级

风险状态

风险结果

负责人

风险描述

风险描述

该风险工单的风险等级为高风险，风险等级为高风险。

风险描述

该风险工单的风险等级为高风险，风险等级为高风险。

治理方案

治理方案

风险详情

当前值

风险等级

当前值

风险等级

当前值



最佳实践分享：诊断分析

服务诊断分析



异常告警根因分析





最佳实践分享：AI 分身

技能与任务管理

工作日志

则平台 服务观测 AI 观测 日志中心 大盘 SLQ稳定性 告警中心 变更事件



工作日志 任务管理

任务名称	任务状态	技能
日志风险巡检	☒	log_govern
告警事件分析	☒	query_alarm

编辑任务

名称 * 日志风险巡检

技能 * 服务E 流量分析

参数 *

参数名	参数值
服务名	

执行周期 *

☒ 每天 ☐ 每周 ☐ 每月

14:20

分析结果

告警规则热度

告警规则	告警事件数
	7
	21
	26
	6
	6
	2

核心发现

CPU使用率和内存使用率的告警规则触发次数较高：[PromQL告警][VictoriaMetrics cpu 使用率]。触发告警的告警规则：[PromQL告警]。触发次数较多，分别为7次和1次。可能存在配置不合理或服务不稳定的问题。

告警接收人分布

	告警事件数	占比 (告警接收人的告警事件数/告警总数)
	9	12.63%
	9	12.63%
	9	12.63%
	3	4.22%

RL + Observability

Observability for RL

行为与轨迹观测

长尾 Request 分析

分布式 Profiling

双向赋能

RL for Observability

智能告警

异常检测（专家规则）

时序理解与推理

未来规划

AI 时代
Ops 新的范式会出现

个性化运维

运维智能化

全域数据

专家算法

风险模型

智能决策

📍 北京

QCon

全球软件开发大会

会议时间：4月10-12日

- 大模型赋能 AI Ops
- 越挫越勇的大前端
- 云上业务架构演进
- 多模态大模型及应用
- 海外 AI 应用创新实践

📍 北京

AICon

全球人工智能开发与应用大会

会议时间：6月27-28日

- 端侧智能
- AI Agent
- 多模态大模型
- 金融+大模型

📍 上海

QCon

全球软件开发大会

会议时间：10月23-25日

- AI Agent
- 大模型训练推理
- 端侧 AI
- 搜推深度融合
- 数智金融

4月

5月

6月

8月

10月

12月

📍 上海

AICon

全球人工智能开发与应用大会

会议时间：5月23-24日

- 通用大模型
- AI Agent
- 垂直领域应用
- 模型可解释性

📍 深圳

AICon

全球人工智能开发与应用大会

会议时间：8月22-23日

- 模型效率与部署
- 多模态大模型
- 大模型安全
- 智能硬件

📍 北京

AICon

全球人工智能开发与应用大会

会议时间：12月19-20日

- 通用大模型
- 智能硬件
- LM Ops
- 具身智能

极客邦科技 2025 年会议规划

促进软件开发及相关领域知识与创新的传播



参会咨询



查看会议

THANKS

大模型正在重新定义软件

Large Language Model Is Redefining The Software

