阿里云弹性计算智能化监控诊断探索和实践

张尤

阿里巴巴高级技术专家

InfoQ 传媒和整合营销服务

对技术人群极具影响力的新闻网站 / 技术社区

InfoQ 是一家全球性的在线新闻 / 社区网站，创立于 2006 年，创始人是 Floyd Marinescu。目前全球拥有英、法、中、日共五种语言的站点。InfoQ 中国于 2007 年由极客邦科技创始人兼 CEO 霍太稳引入中国。

十五年来，InfoQ 致力于促进软件开发及相关领域知识与创新的传播，凭借在技术服务领域的深耕。

300W+	150W+	100W+	300W+	1600+
InfoQ 网站月访问量	积累公众号粉丝	微博粉丝	覆盖中高端技术开发者	CTO、技术高管

关于我

神龙计算平台-异常调度-监控&诊断
核心研发+诊断平台负责人

ECS数据&稳定性团队-性能基线
核心研发

ECS管控&工程效率
核心研发

· 2015.02 · 2016.10 · 2017.10 · 现在

大纲

• 01 • 概述&背景

- 业务难点&客户诉求
- ECS监控诊断发展历程

• 02 • 业界方案

- 云原生方案
- 其他商业软件方案

• 03 • 我们的方案

- 可靠的数据采集方案
- 根因可解释性如何解决
- 智能运维决策
- DryRun系统&评价体系

• 01 概述&背景

客户对诊断能力诉求

异常特征的训练和学习/不断迭代

客户自愈方案自动推荐

全链路根因诊断异常感知

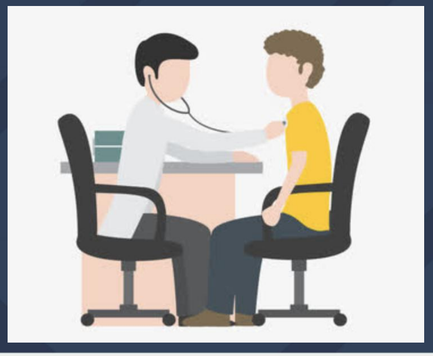
简单直接的可用性判定

类比医疗诊断技术

远程AI诊断/早期诊断

完善的治疗方案

指标量化的复杂体检



简单直接的诊断手段

业务难点 基础设施规模庞大

200+
可用区(AZ)

3000+
网络和CDN节点

4
大洲

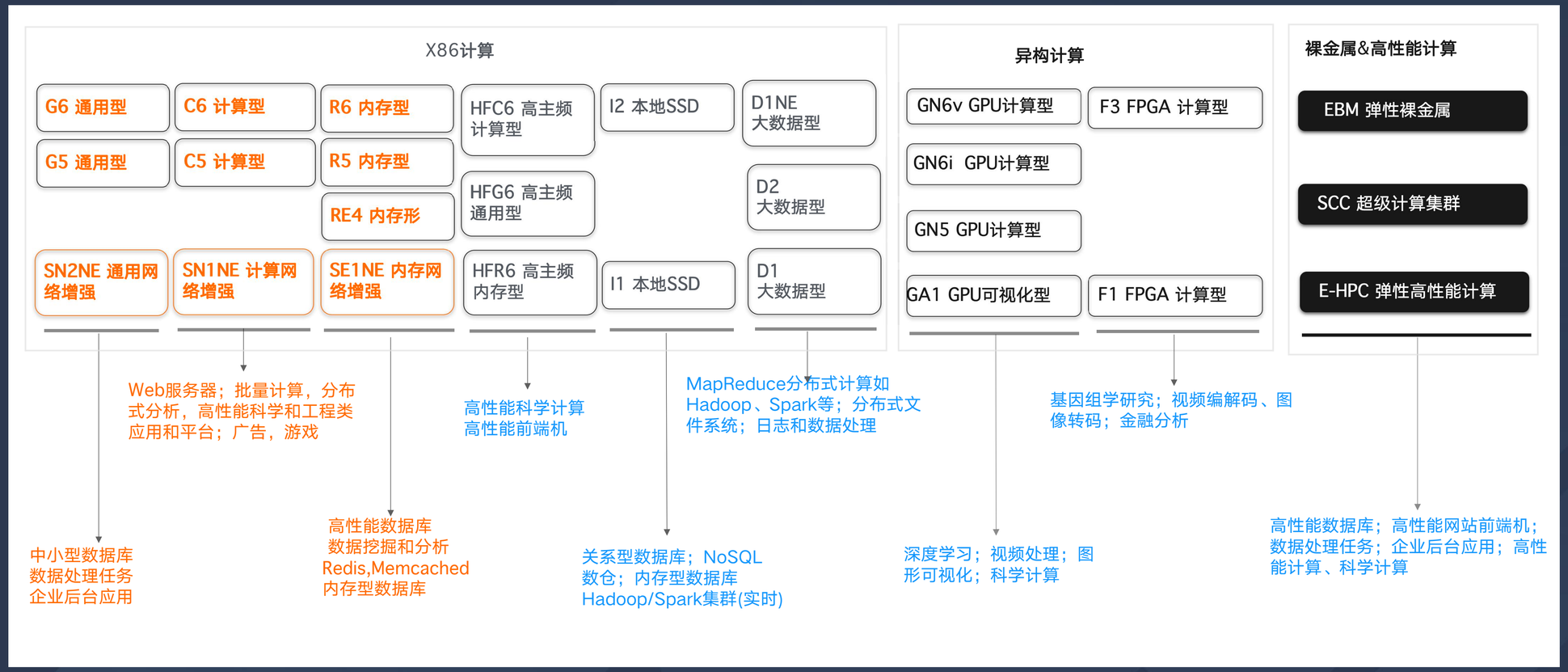
25+
数据中心区域(Region)

~ 100,000,000+
部件
(CPUs, disks etc.)

~ 1,000,000+
设备

~ 5000+
集群

业务难点 业务领域覆盖众多



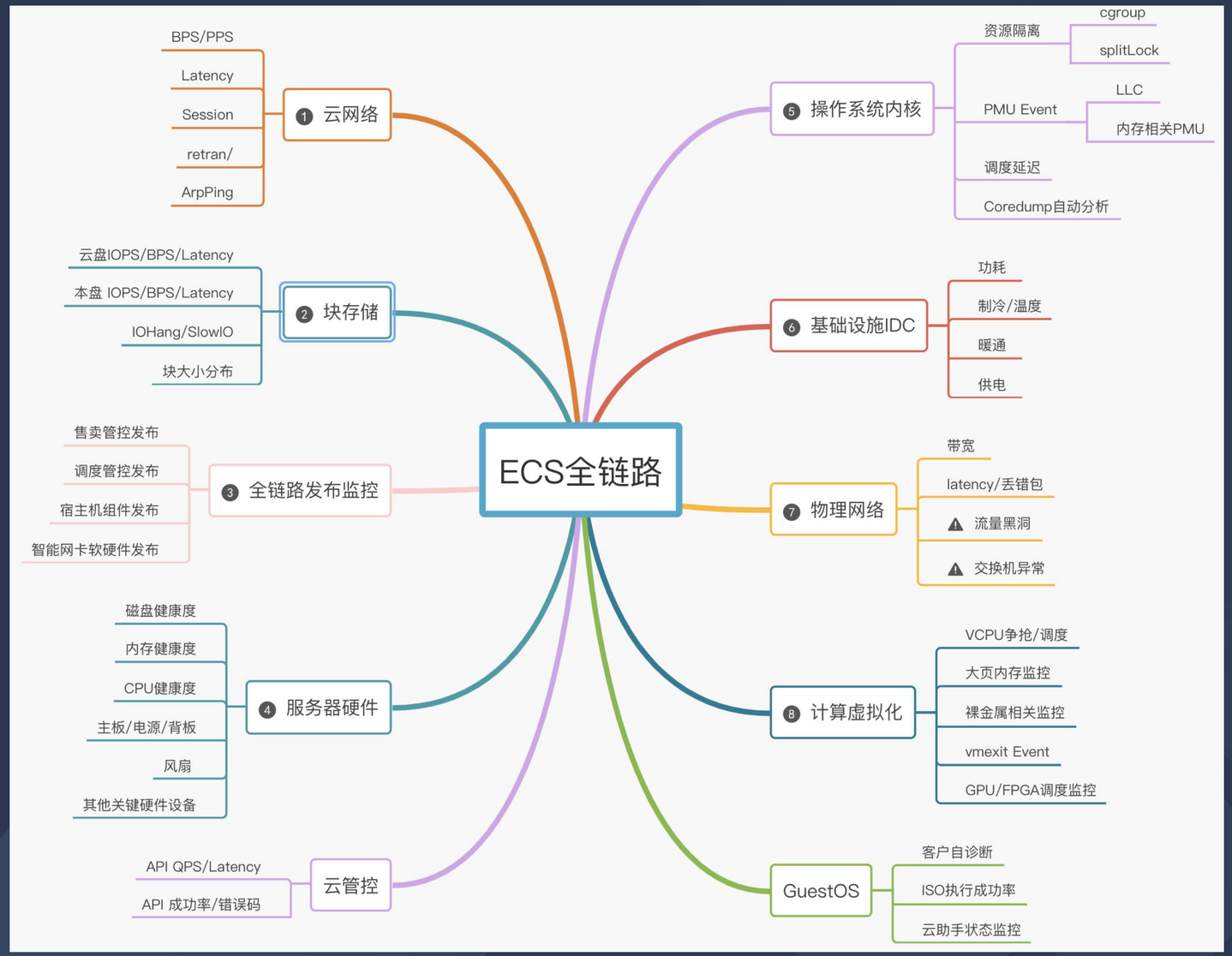
覆盖业务领域广

- 通用计算业务
- 高性能高主频业务
- 本地盘存储大数据业务
- 异构GPU业务
- 异构FPGA业务
- SCC超算业务
- RDMA高性能网络

售卖形态多

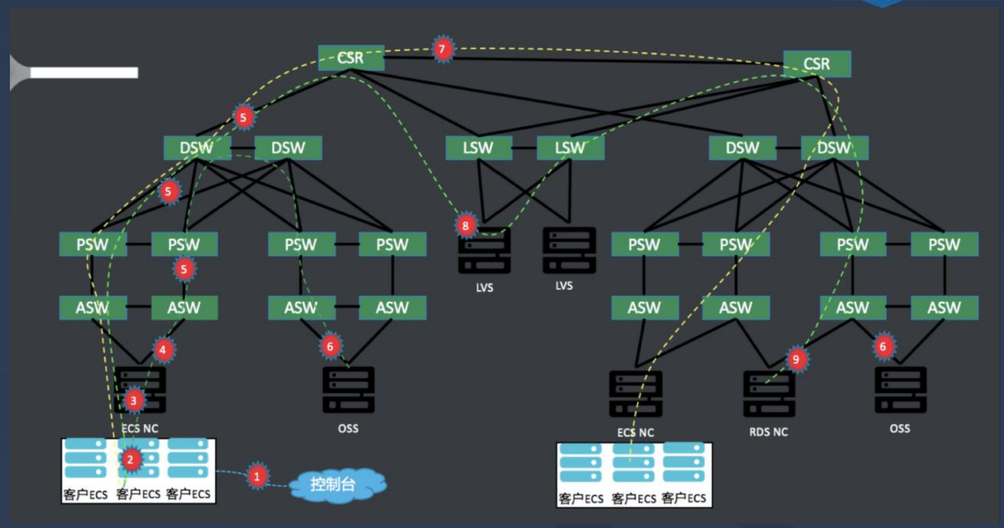
- 预付实例
- 按量实例
- 预留实例
- 竞价实例
- 停机不收费实例
- DDH实例
- 智能机柜实例
- 混合云一体机

业务难点 领域知识覆盖广且技术难度深



覆盖子系统众多

链路长



技术难度深

CPU子系统监控举例

core

- turbo P02
- AVX P1 Freq
- AVX P0n Freq

cache

- LLC访问一致性
- LLC容量
- LLC容量QOS

IMC

- IMC freq
- IMC channel
- memory buffer

ECS监控诊断运维发展历程

工具+人工时代

平台+半自动编排

数据化智能化

飞天底座监控上线

IDC硬件监控上线

组件监控接入Alimonitor

ECS正式售卖

0. / .

接入落日弓发布编排

交付效能提升10倍

0. / 1

底层应用接入天基&监控接入XFlush

攻克底座环境问题

0. / 4

ECS全链路诊断平台上线

Cloudops运维平台发布

0. / 5

监控采集Agent重构

监控覆盖度99.995%

云上客户系统事件发布

智能运维决策大脑上线

0. / 6

攻克根因可解释性问题

内存故障预测上线

0. / 7

99.9%异常自动闭环

策略评价体系建设

0. 0. +至今

小结1

① 两个背景-(自身)

- ECS是IAAS级别的云产品，产品上**没有单VM的灾备能力**
- 物理机上的软/硬件从**长时间来看：单机的异常不可避免**
- ECS**产品形态越来越丰富，各个子模块监控技术难度深**

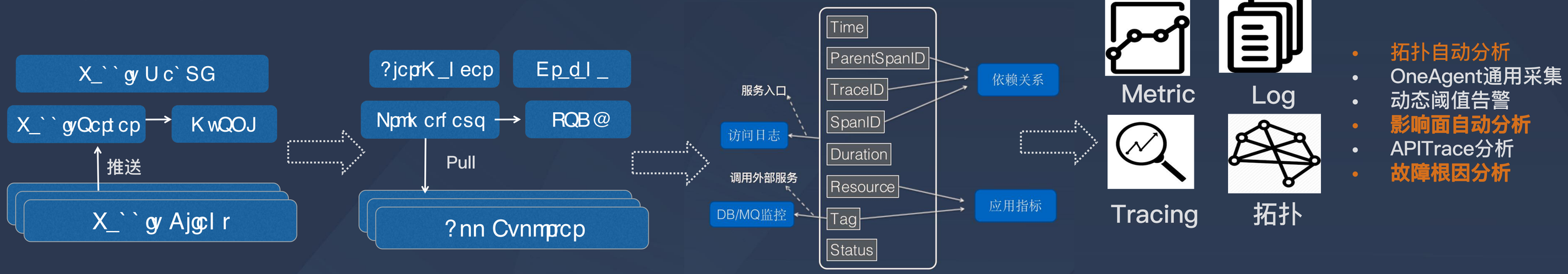
② 两个背景-(客户)

- 未发生的不可用：**提前规避**
- 已发生的不可用：**及时止损**
- 整个处置过程：**信息透明**

三个挑战

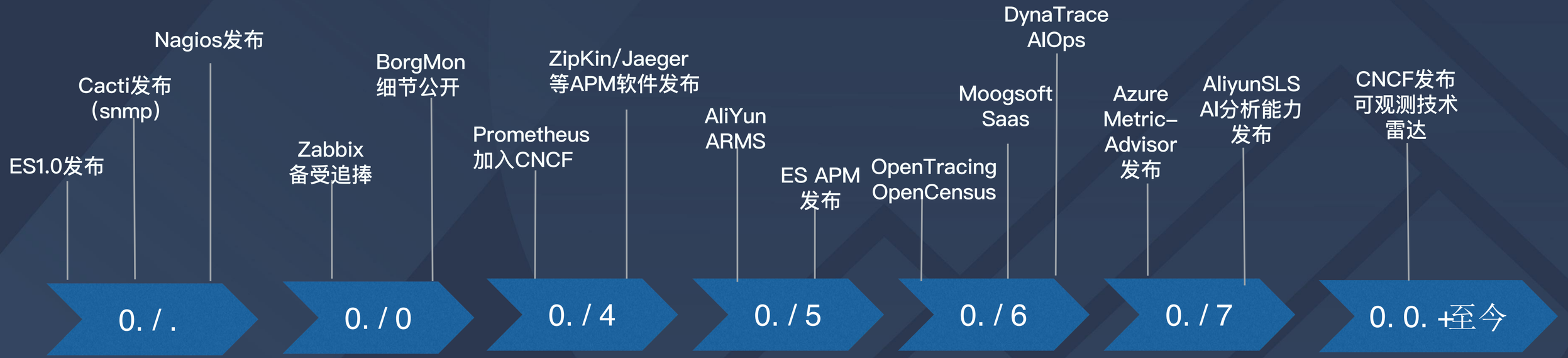
02 业界方案

业界方案及发展



- 拓扑自动分析
- OneAgent通用采集
- 动态阈值告警
- 影响面自动分析
- APITrace分析
- 故障根因分析

传统监控时代 | 新监控标准+APM产品百家齐放 | 数据化/AIOps/可观测



03 我们的方案

ECS监控诊断架构设计-采集端设计

1) 采集的数据回流&消费

服务消费

数据报表 | SLA平台 | 监控诊断 | 故障快恢 | 其他系统 | 采集大盘

回流存储

ODPS(弹内) ← 2小时同步 ← ODPS(弹外)

半小时投递一次

SLS(弹内) ← Blink同步 ← SLS(德国) / SLS(迪拜) / SLS(ó ó)

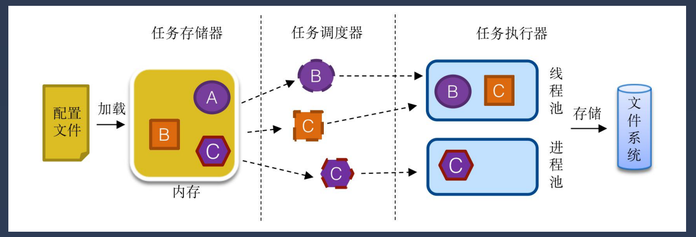
节点采集

- controller/ecs_alarm_agent(Cgroup)
  - EcsAlarmAgent(Daemon)
    - 配置管理 | 监控管理 | 监控打散
    - 监控调度 | 监控观测 | 结果存储
    - ClusterInfo | NcInfo | VmInfo

天基Tianji底座

2) 数据采集的异常治理闭环

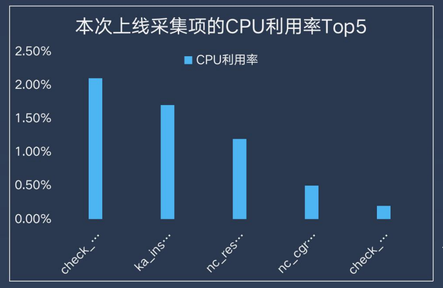
- 调度器轻量化改造
- 采集之间数据共享



架构改造

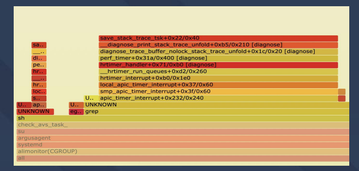
采集准入

- 代码Review性能评审
- Gamma环境性能评估



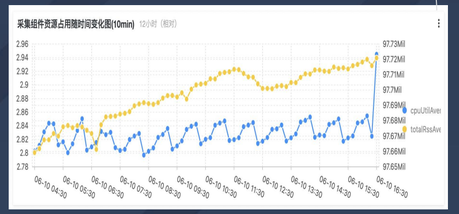
异常分析

- 内存泄露分析系统化
- 火焰图找瓶颈点

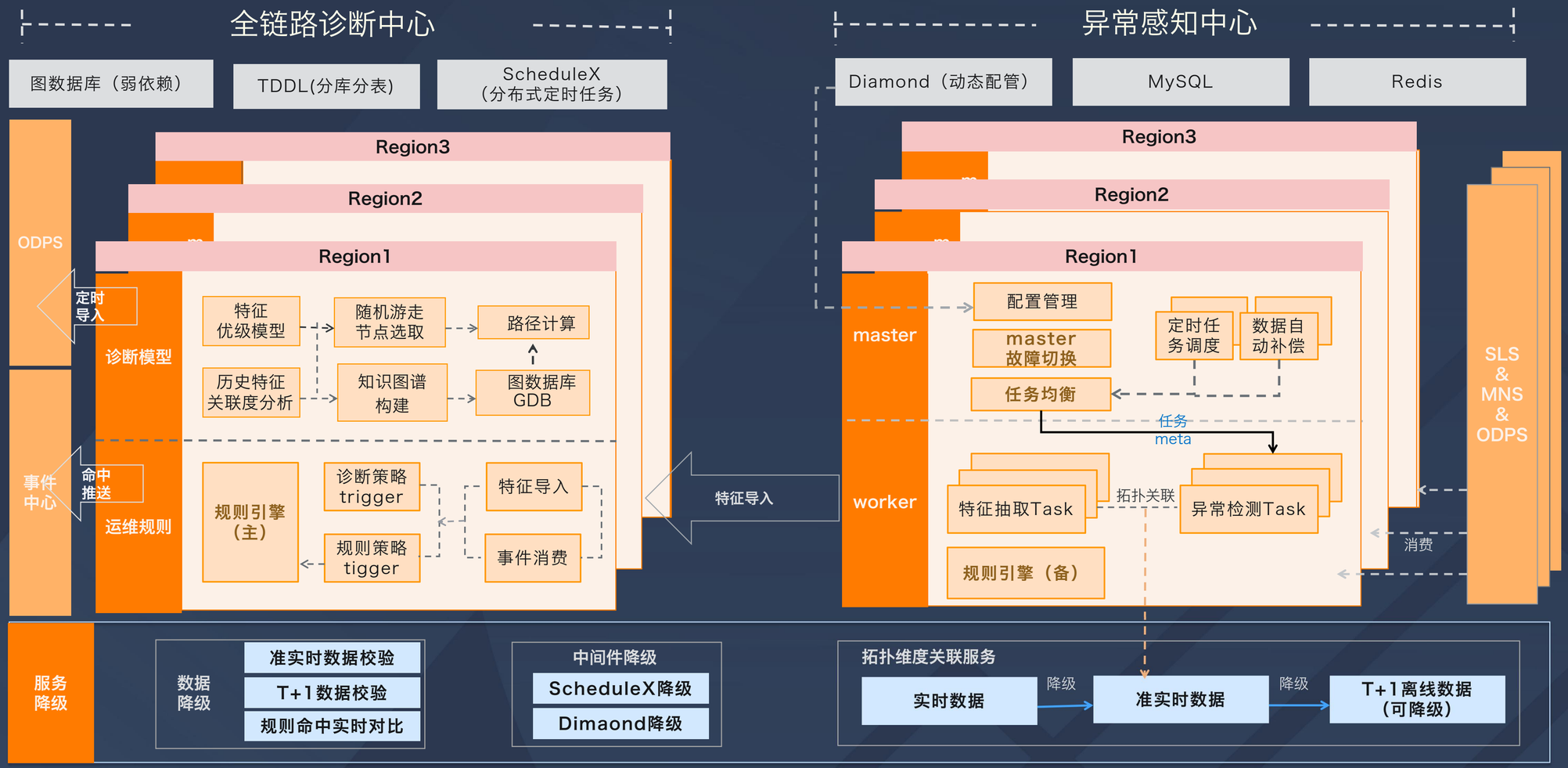


日常巡检

- 异常占用实时告警
- 日常报表持续治理



ECS监控诊断架构设计-服务端设计



ECS异常根因诊断的解决方案

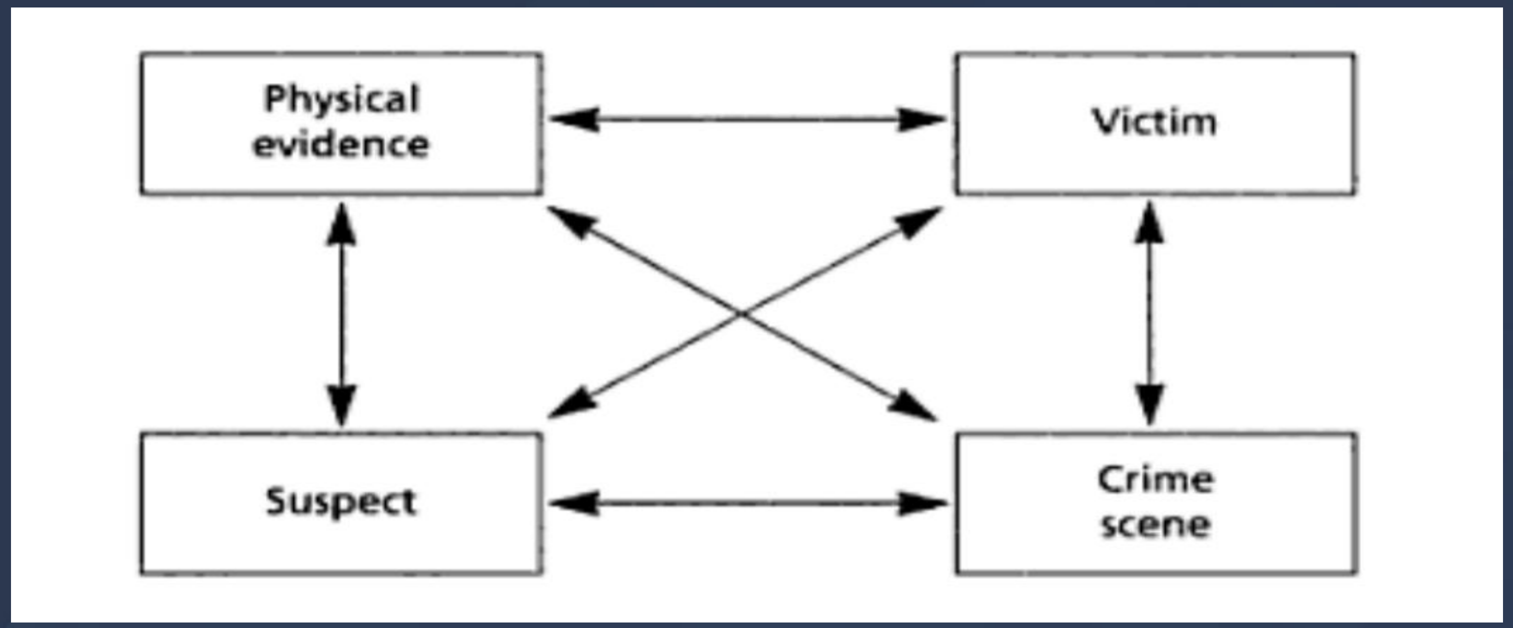
准确性

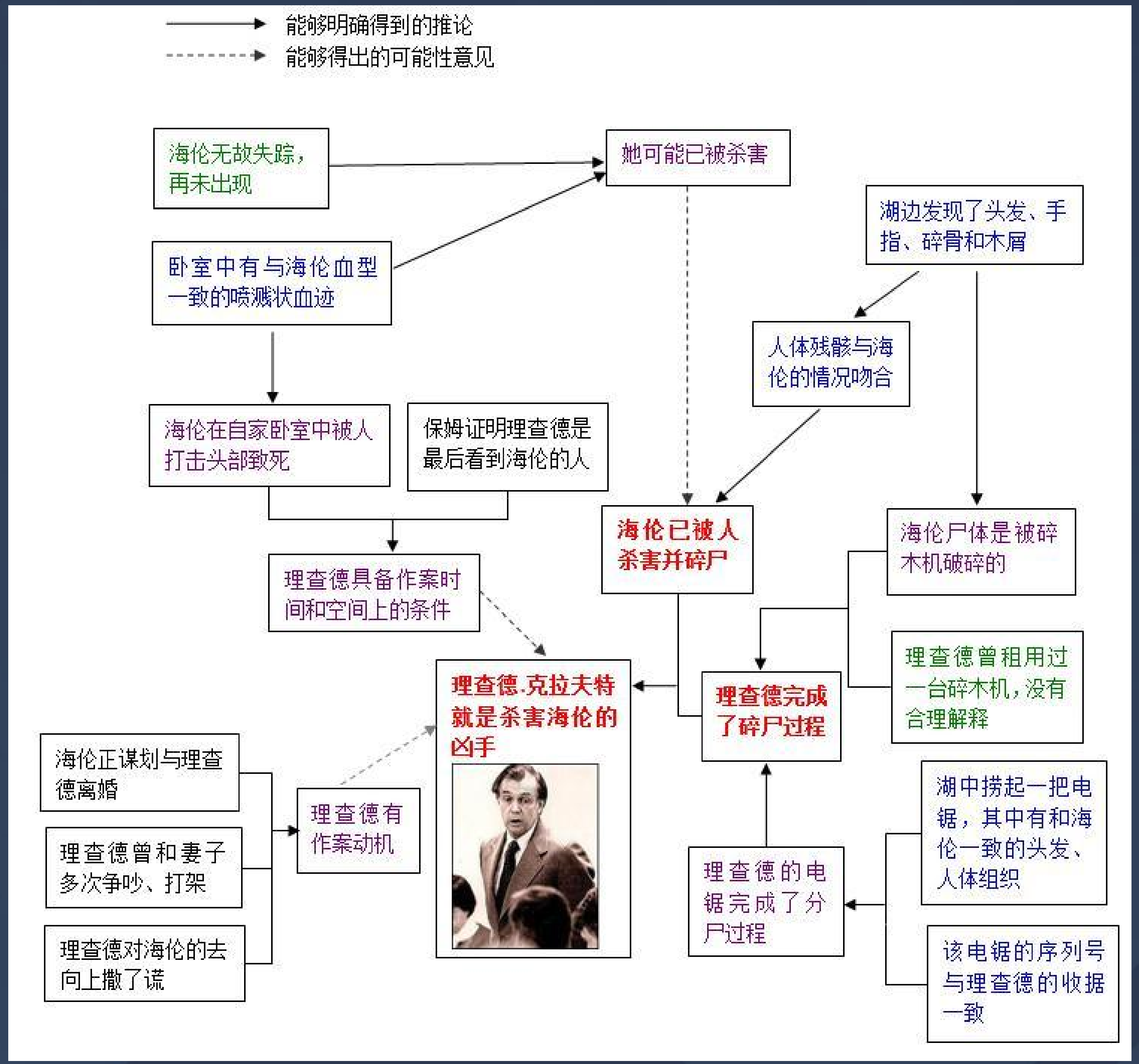
- 故障时刻异常太多，哪一个才是问题的根因？
- 传统根据决策树诊断的根因没有可解释性。

可解释性

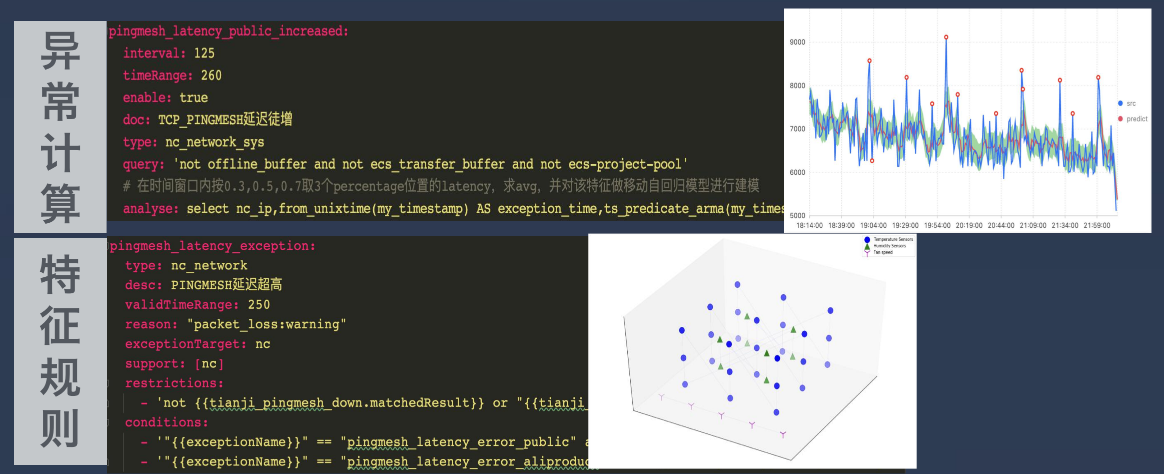
- 异常推导链怎么得出？
- 如何量化推导链的置信度 or 支持度？
- 随着业务的发展，如何自适应迭代？

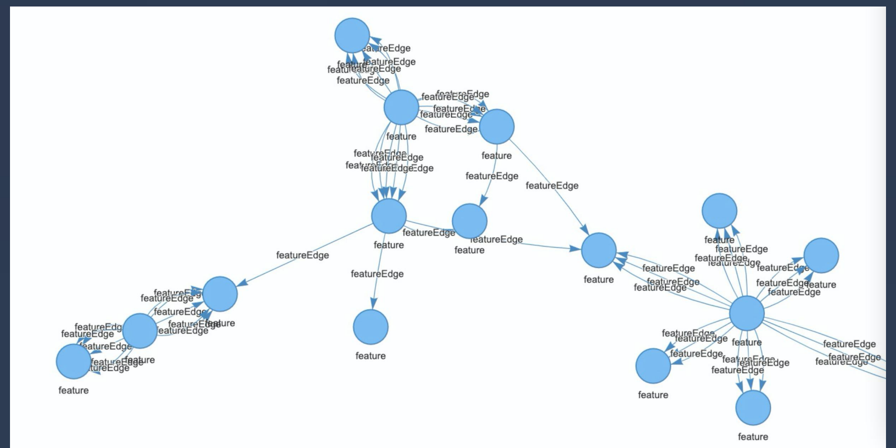
1 现实中的“根因诊断”

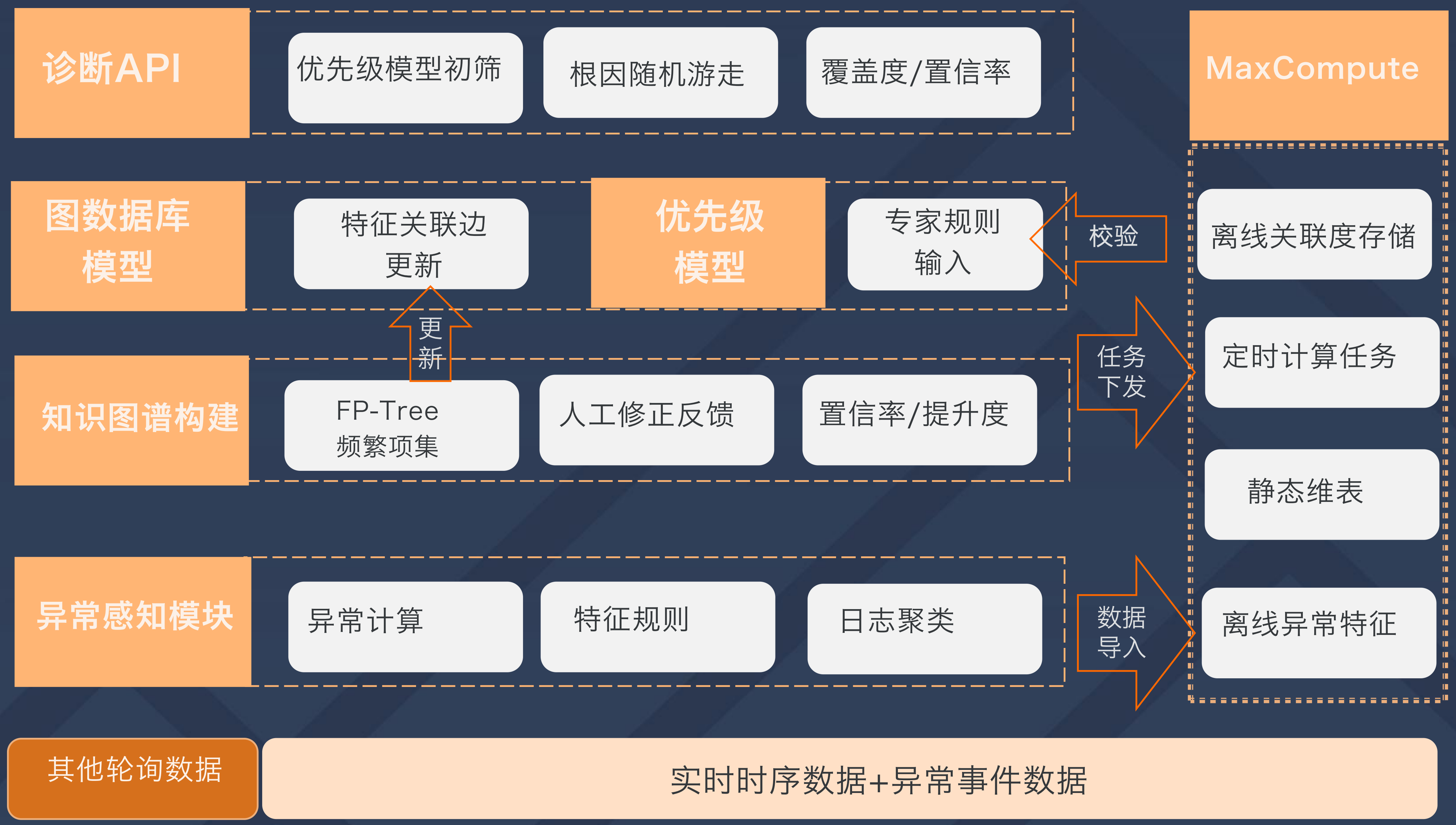




2 阿里云ECS根因诊断的总体方案







ECS智能运维决策系统--概览

面临问题

- 单一维度判定执行运维风险高。
- 异常恢复手段单一，没有结合客户通知。
- 流控维度单一不支持表达式配置。
- 规则变更无法回溯历史，全凭经验

a. 每日命中Resource详情
b. 每日命中资源地域集群分布
c. A/B Test灰度切流比例建议

a. 历史异常回放
b. 提前命中程度统计分析
c. 漏召回详情分析
d. 新增命中资源详情分析

ECS智能运维决策系统--评价体系（理论基础）

有效性问题

- 运维动作&运维模型是否有正向效果？
- 如何评价并选取最佳运维动作？

评价指标

- ECS的稳定核心指标（KeyMetric）的应该考虑哪些因素？
- 采用何种假设检验算法体现不同运维的差异性？

1) ECS主动运维构成的要素

- 预测的模型特征 + 运维的动作
- 运维的观察窗口，以及观察窗口过程中的不可用时长
- 运维过程中是否有提前通知客户行为，提前时长是多久？

2) KeyMetric计算

- VM总活跃时长：$VmTotalLifeTime_T = \sum_{vm_1}^{vm_n}(LifeTime_T + Down_T)$
- 异常严重程度系数： F_i (越严重的异常系数越大)
- 主动通知后计算的系数加权： F_e（提前量越多系数越小）

3) 评价--假设检验算法等

- 显著性差异检验-单因素方差分析F检验（Welch's anova）
- 精准控制切流比例--功效分析（Cohen’s f）

数据面

性能维度

控制面

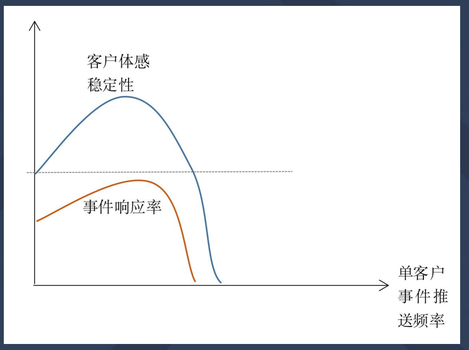
a. 数据面不可用服从正态分布

b. 性能/控制面可能是多峰分布

c. 样本组的方差齐性大概率不满足

d. 切流比例

$$\text{Cohen's} f_p = \sqrt{\frac{\eta_p^2}{1-\eta_p^2}} = \sqrt{\frac{SS_{effect}}{SS_{error}}}$$

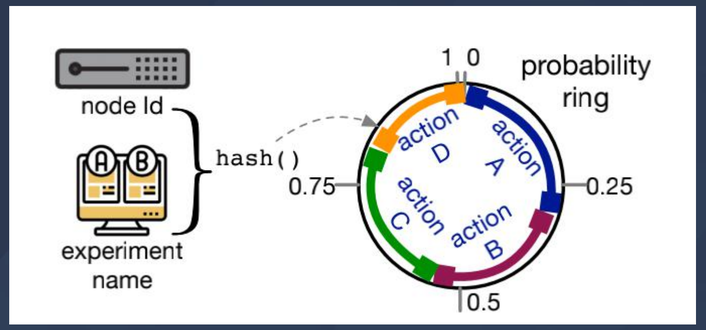


ECS智能运维决策系统--评价体系（工程落地）

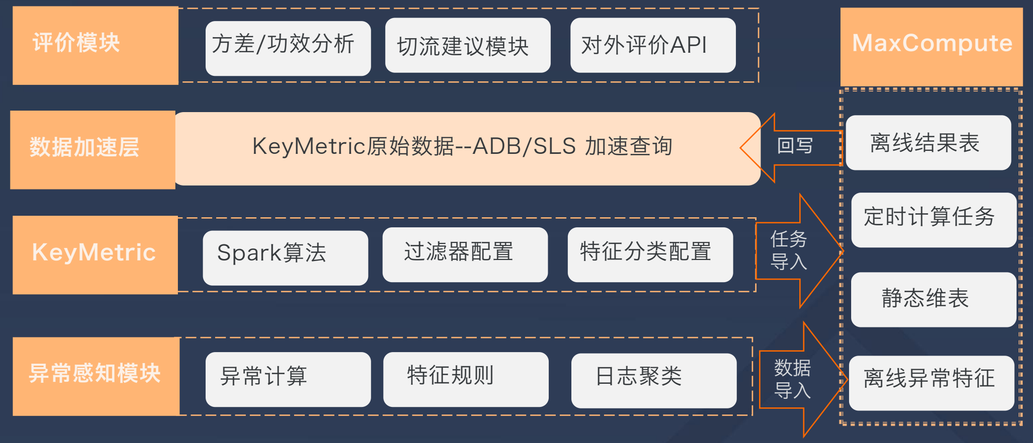
面临问题

- **对照组选取**的不确定性问题？
- 对照组样本太少怎么办？

1 对照组选取（HashRing）



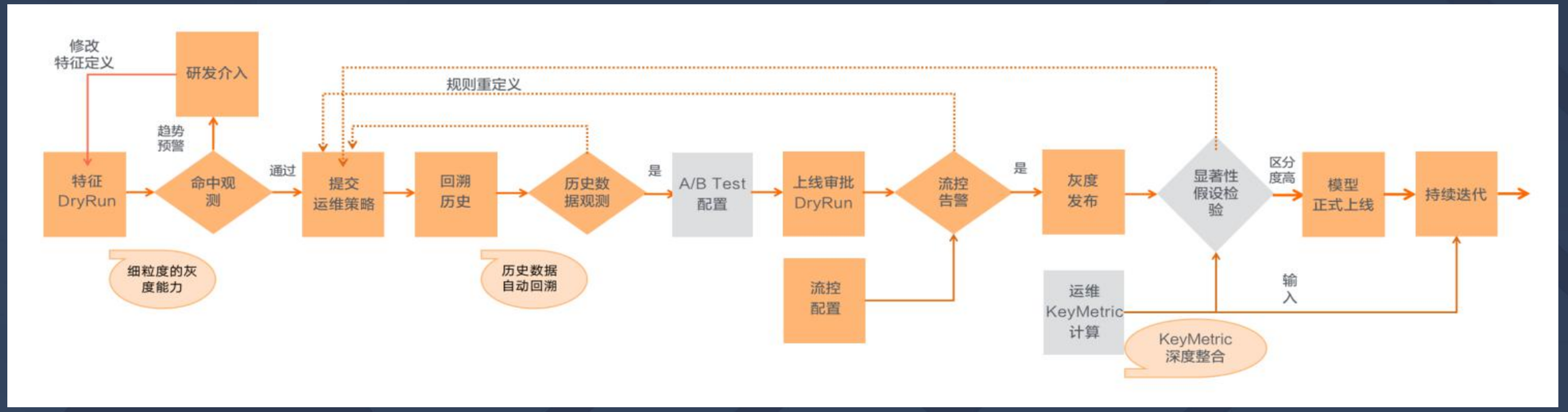
2 总体工程落地的方案



面临问题

- 如何于现有的**运维体系整合**？
- 如何安全高效的灰度**发布上线**？

3 运维策略A/B Test上线流程



感谢

【1】Localizing Failure Root Causes in a Microservice through Causality Inference link
【2】Henry Lee. Henry Lees Crime Scene Handbook [M] .London:Academic Press, 2001
【3】Failures in Large Scale Systems: Long-term Measurement, Analysis, and Implications. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis 2017
【4】Kernel Methods in Machine Learning
【5】Openstack Ceilometer / prometheus 社区相关文档
【6】Ganter for Spring 2021 AIOps G2CR-Dynatrace-AIOps-Platforms-Spring2021.pdf
【7】Critical Capabilities for Application Performance Monitoring https://www.gartner.com/doc/reprint
【8】DynaTrace Problem Detection and Analysis https://www.dynatrace.com/support/h
【9】(达摩院时序智能团队+大数据基础工程与技术团队) Chaoli Zhang^, Zhiqiang Zhou^, Yingying Zhang^, Linxiao Yang^, Kai He^, Qingsong Wen^, Liang Sun^ (^Equally Contributed), "NetRCA: An Effective Network Fault Cause Localization Algorithm," in Proc. IEEE 47th International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2022), Singapore, May 2022.
【10】(大数据基础工程与技术团队+达摩院时序智能团队) Yingying Zhang, Zhengxiong Guan, Huajie Qian, Leili Xu, Hengbo Liu, Qingsong Wen, Liang Sun, Junwei Jiang, Lunting Fan and Min Ke, "CloudRCA: A Root Cause Analysis Framework for Cloud Computing Platforms," in Proc. 30th ACM International Conference on Information and Knowledge Management (CIKM 2021), Queensland, Australia, Nov. 2021