超大规模弹性计算节点自动化运维稳定性实践

唐磊（小唐）

阿里云 技术专家

InfoQ 企业会员

企业数字化传播一站式服务

InfoQ 企业会员是为满足企业在中国开发者群体中的品牌曝光需求而推出的一款矩阵化资源包。可为企业提供包括“ 企业号服务 ”、“ 企业动态宣发 ”、“ 品牌展示通道 ”在内的多项专属权益与服务，助力企业高效触达开发者群体，提升数字化时代影响力。

企业号服务

深度触达 300 万中高端开发者

企业动态宣发

新媒体矩阵覆盖百万粉丝

品牌展示通道

线上平台 10 万+ 流量曝光

个人简介

姓名：唐磊，花名：小唐，网名：石头

- 2008 ~ 2015：学生@中南/清华，（CG&CAD）
- 2015 ~ 2017：工程师@宜信大数据创新中心，（互金–信贷）
- 2017 ~ 2019：工程师、TL@大疆创新，（社区 SkyPixel）
- 2019 ~ 至今：工程师@阿里云神龙计算平台，（弹性计算–异常调度）

内容提纲

01 概述&背景

- 客户诉求
- 业务难点

02 业界方案

- 业界方案
- 发展趋势

03 我们的方案

- 基于专家规则的自动化运维策略
- 运维评价
- 发布熔断

01 概述&背景

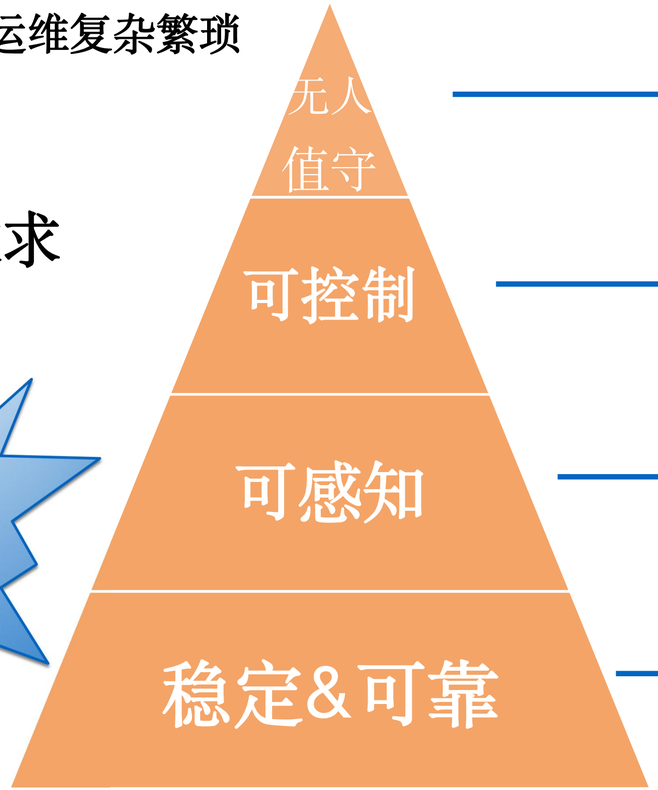
概述&背景 – 客户诉求

❑ 云下环境特点

- 维护成本高
- 资源利用率低
- IT 资源管理和运维复杂繁琐
- …

❑ 客户用云述求

VM:
IaaS产品天然单点



智能自愈：
- 智能编排
- 预测&自愈

可控制：
- 原子操作
- 自动化

异常感知：
- 监控&告警
- 根因诊断

稳定&可靠：
- 稳定性 SLA
- 性能 SLA

幸福感

自动化运维
提升幸福感

已发生的不可用：及时止损
未发生的不可用：提前规避

安全感

业务难点 – 基础设施规模大

200+
可用区(AZ)

3000+
网络和CDN节点

28
数据中心区域(Region)

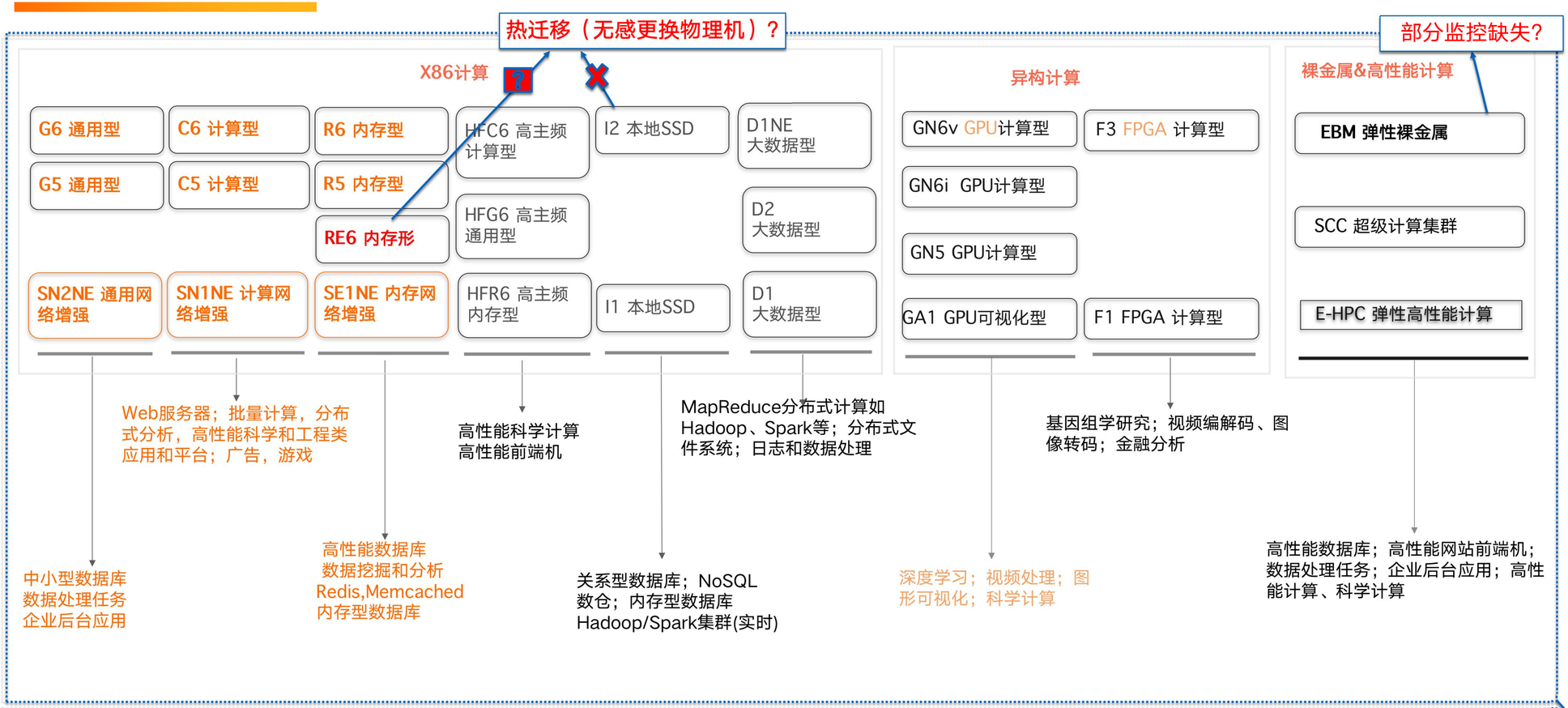
100,000,000+
部件
(CPUs, disks etc.)

1,000,000+
设备

5000+
集群

- 云计算基础设施规模决定了其运维复杂度
- 没有现成的产品可以借鉴，需要探索出自己的道路

业务难点 – 产品形态多、业务领域广



业务领域广

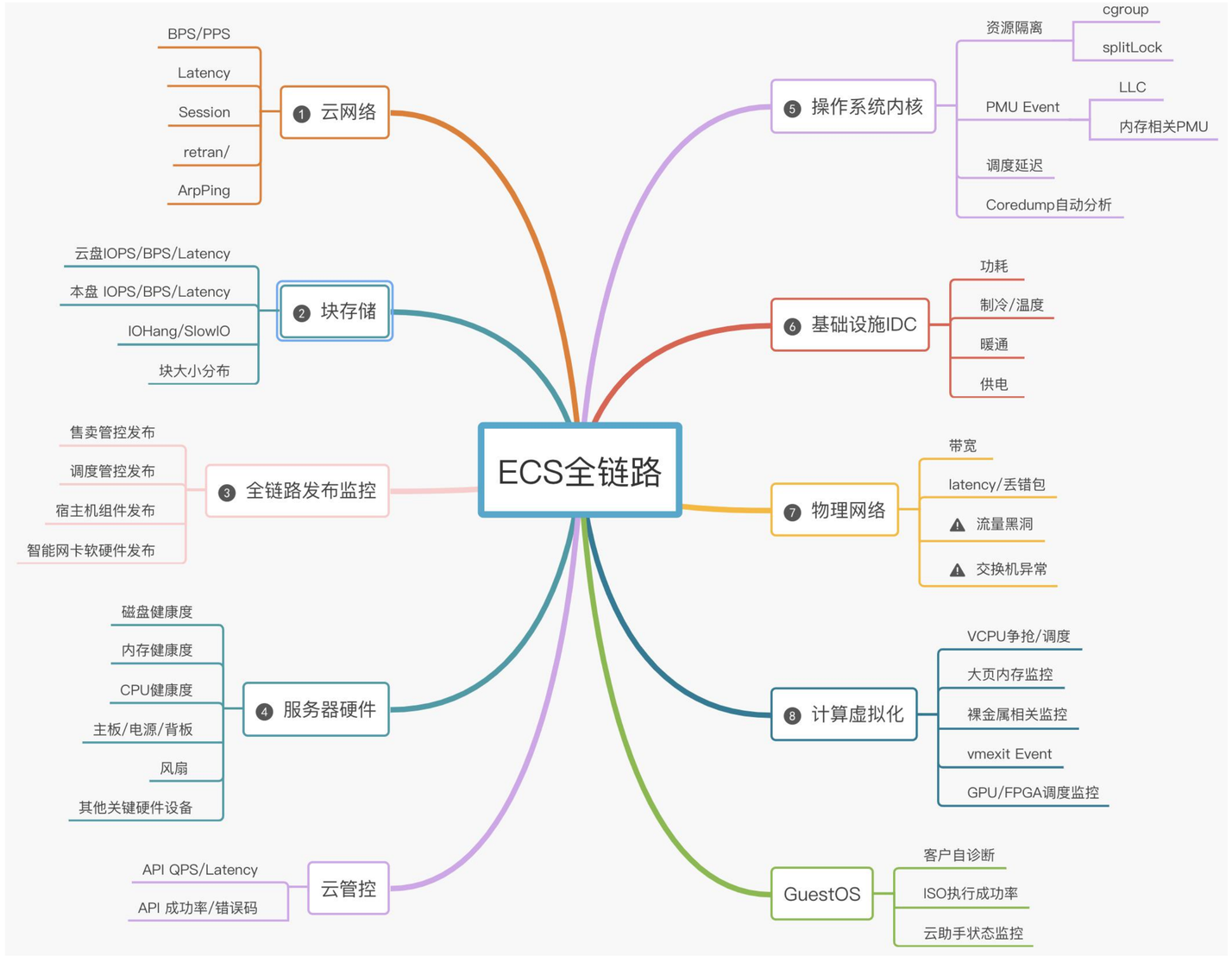
- 通用计算业务
- 高性能高主频业务
- 本地盘存储大数据业务
- 异构GPU业务
- 异构FPGA业务
- SCC超算业务
- RDMA高性能网络

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

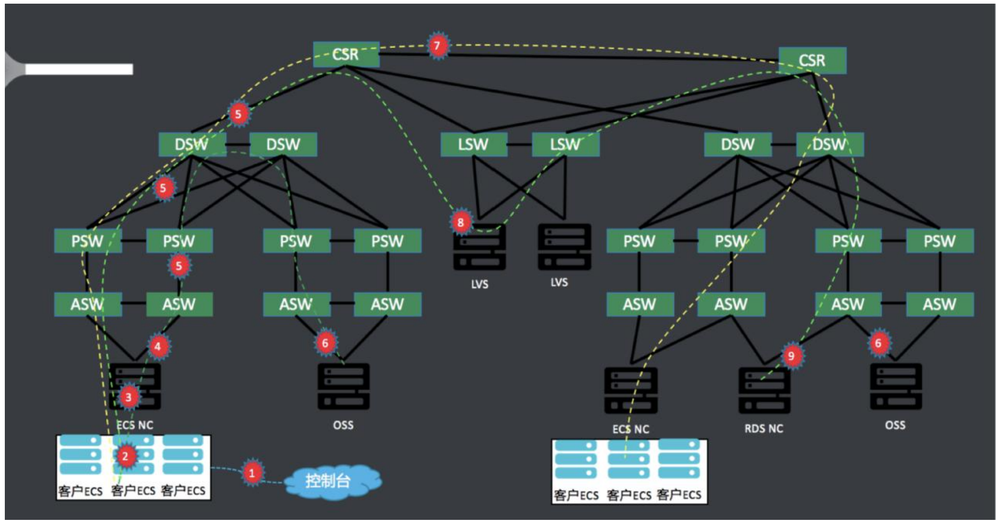
监控诊断运维手段有差异

业务难点 - 知识面广&技术难度深



覆盖子系统众多

链路长



技术难度深

CPU子系统举例

core

- Intel: SKL/ICL/Atom
- AMD:MiLan/ROME
- YiTian710
- kunpeng

cache

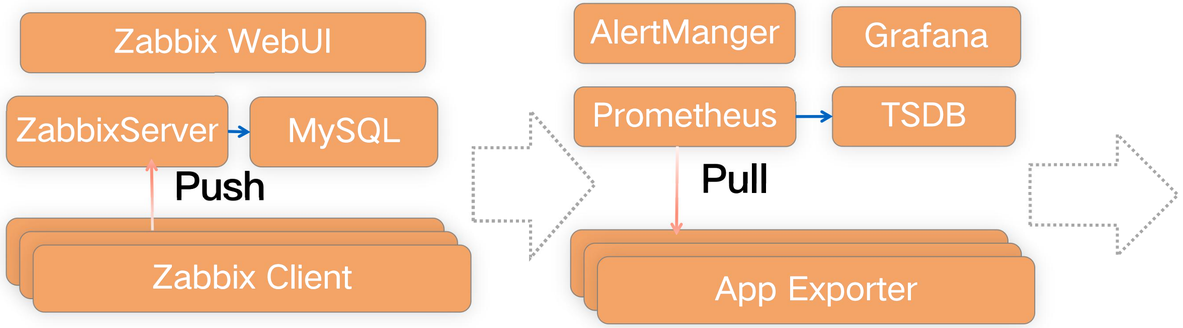
- LLC 一致性
- LLC 容量 QoS
- LLC 争抢

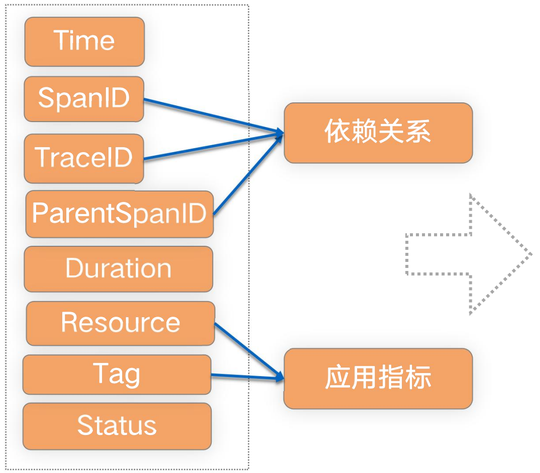
IMC

- IMC freq
- IMC channel
- memory buffer

02 业界方案

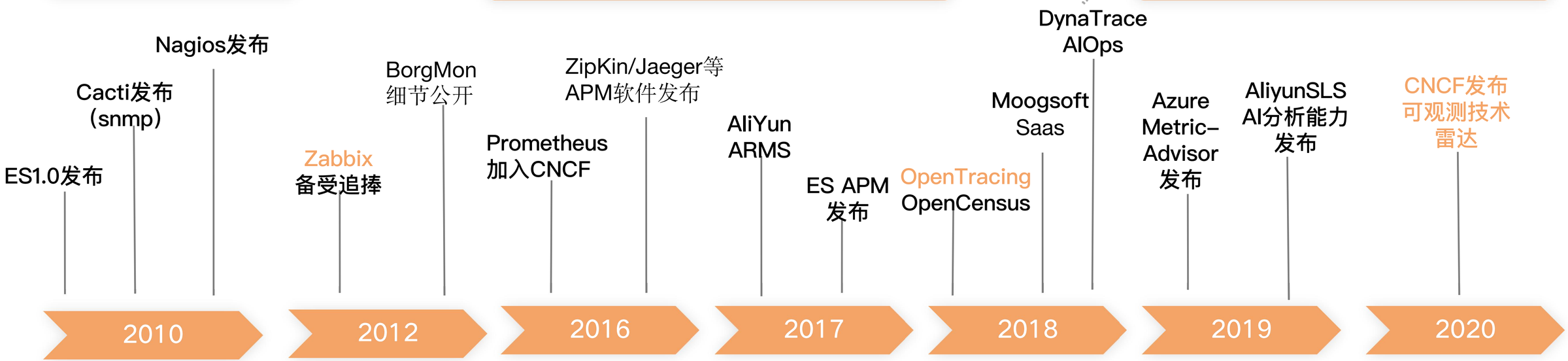
业界方案及发展





- 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



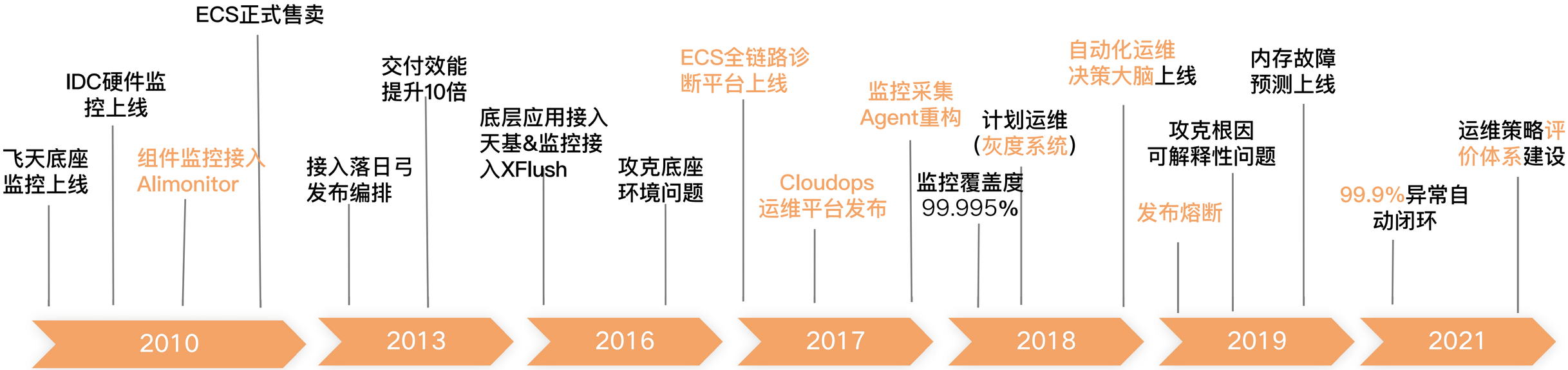
03 我们的方案

我们的方案 – ECS 监控运维体系 发展历程

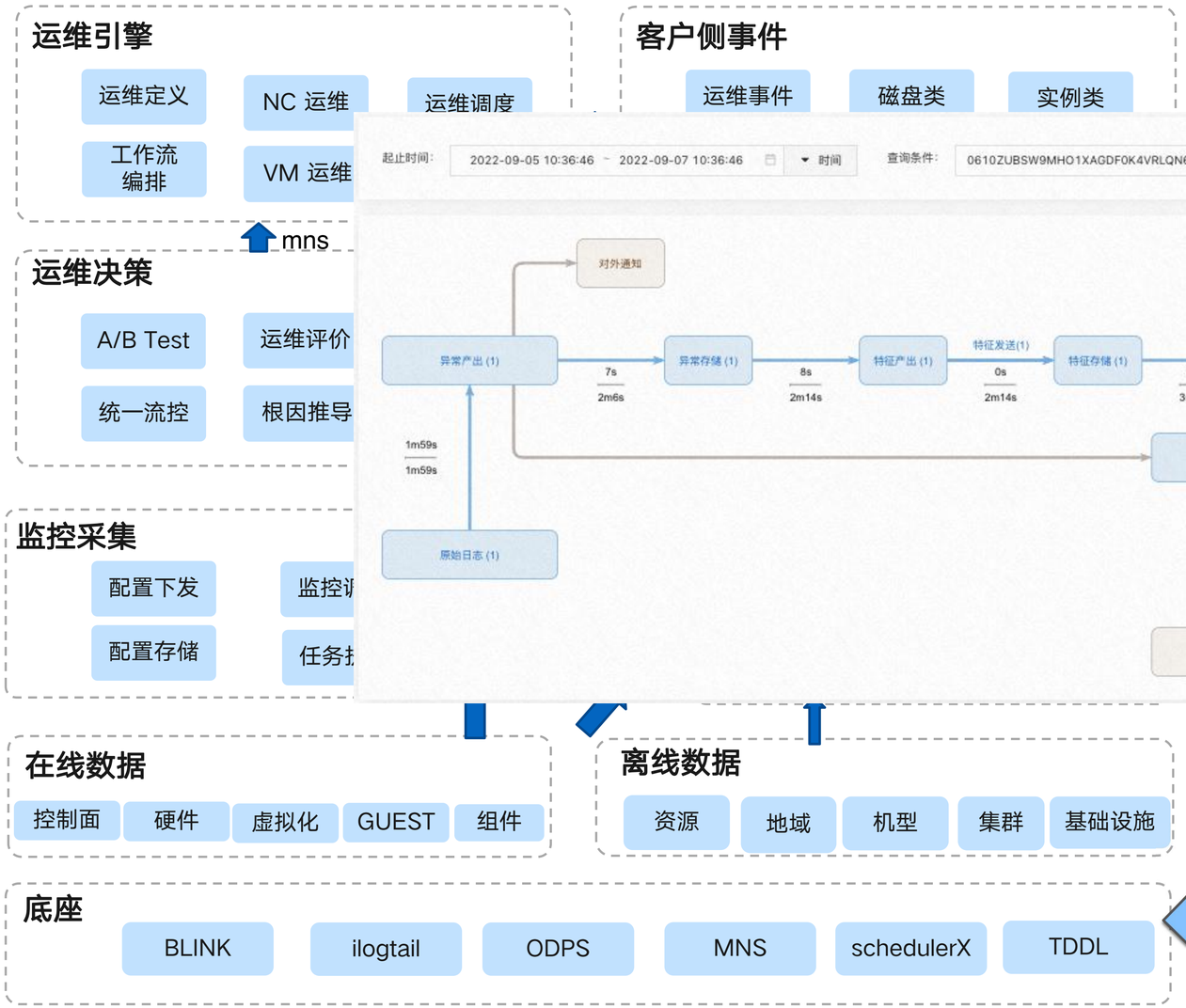
工具+人工时代

平台+半自动编排

数据化智能化



我们的方案 – 整体架构



完整链路追踪

找到每一条异常的源头，通过各种数据清洗、加工、抽象后，基础的数据源

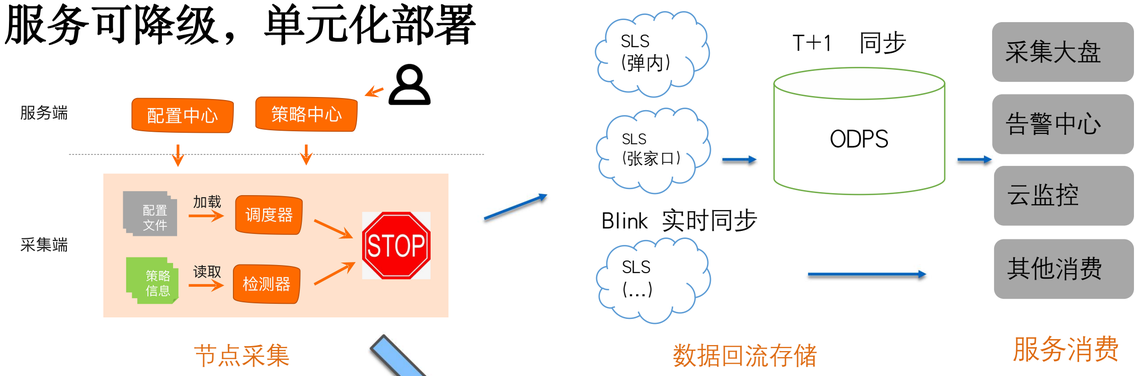
- **数据处理：** 承载百万级物理机的诊断能力
  ODPS 承载百TB级数据传输； （hadoop）
  BLINK 确保每日同等数量级的 数据实时性 (apache-flink)
- **日志服务：** 类似 ELK 产品栈，提供更完整的数据处理能力（数据投递、日志聚类），以及强大的计算聚合函数 (map, json, lamda 函数等）
- **MNS：** 高效、可靠、安全、便捷、可弹性扩展的分布式消息通知服务
- **TDDL：** 分库分表解决方案 （sharding-jdbc/MyCat）
- **SchedulerX：** 基于Akka架构的分布式任务调度平台 （XXL-JOB/ElasticJob）

我们的方案 – 监控数据采集

❑ 规模效应

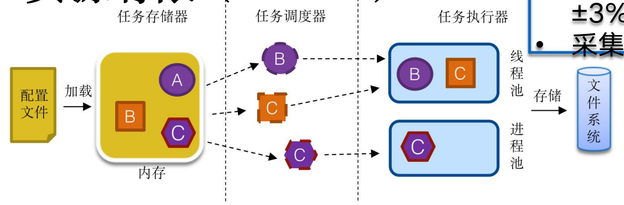
对比项	AlarmAgent	Prometheus	Nightingale
来源	自研	K8s开源	滴滴开源
触发方式	采集侧 服务侧	服务侧	采集侧
采集策略	周期采集 按需采集	周期采集	周期采集
支持规模	百万级节点	千级别	万级别
结果完整度	高 采集回流分离	中 结果实时回流	中 结果实时回流
采集自监控	完善 报表和告警	有限 节点丢失提醒	有限 节点丢失提醒
灰度发布	支持	不支持	不支持

❑ 服务可降级，单元化部署



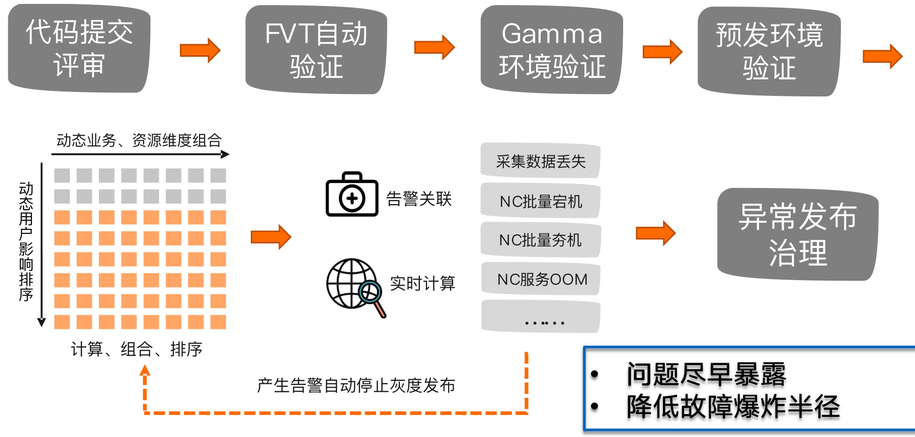
❑ 资源最大化利用：资源有限（1-2HT）

- 调度器轻量化
- 采集之间数据共享



- 覆盖率：99.9%
- P95 CPU 利用率：±3%
- 采集数据：100T/D

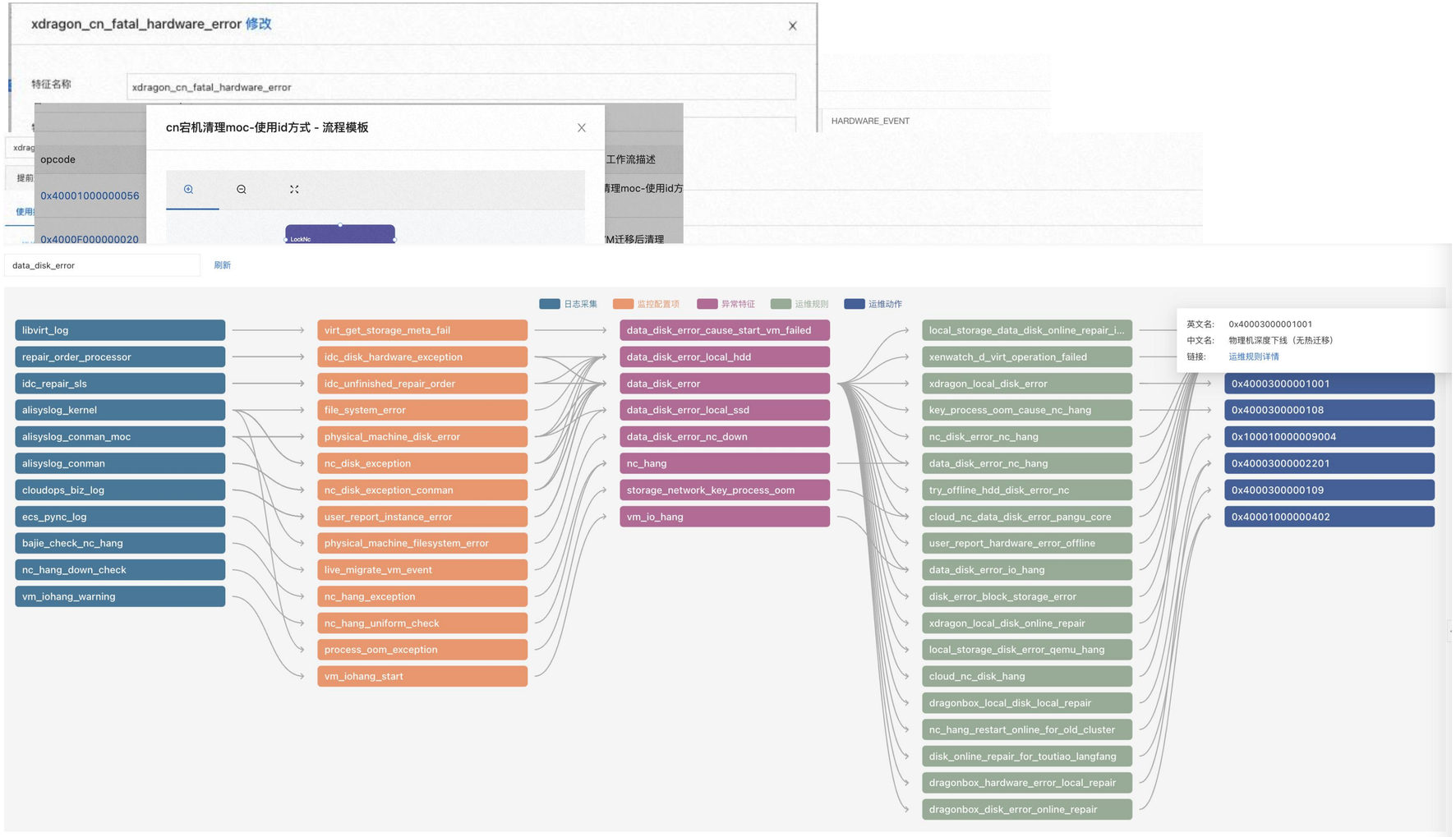
❑ 代码发布灰度可控



- 问题尽早暴露
- 降低故障爆炸半径

我们的方案 – 运维策略相关定义

- 相关定义：
  - 监控异常
  - 特征定义
  - 运维规则
  - 运维动作



我们的方案 – 运维评价（理论）

❑ 产生背景：

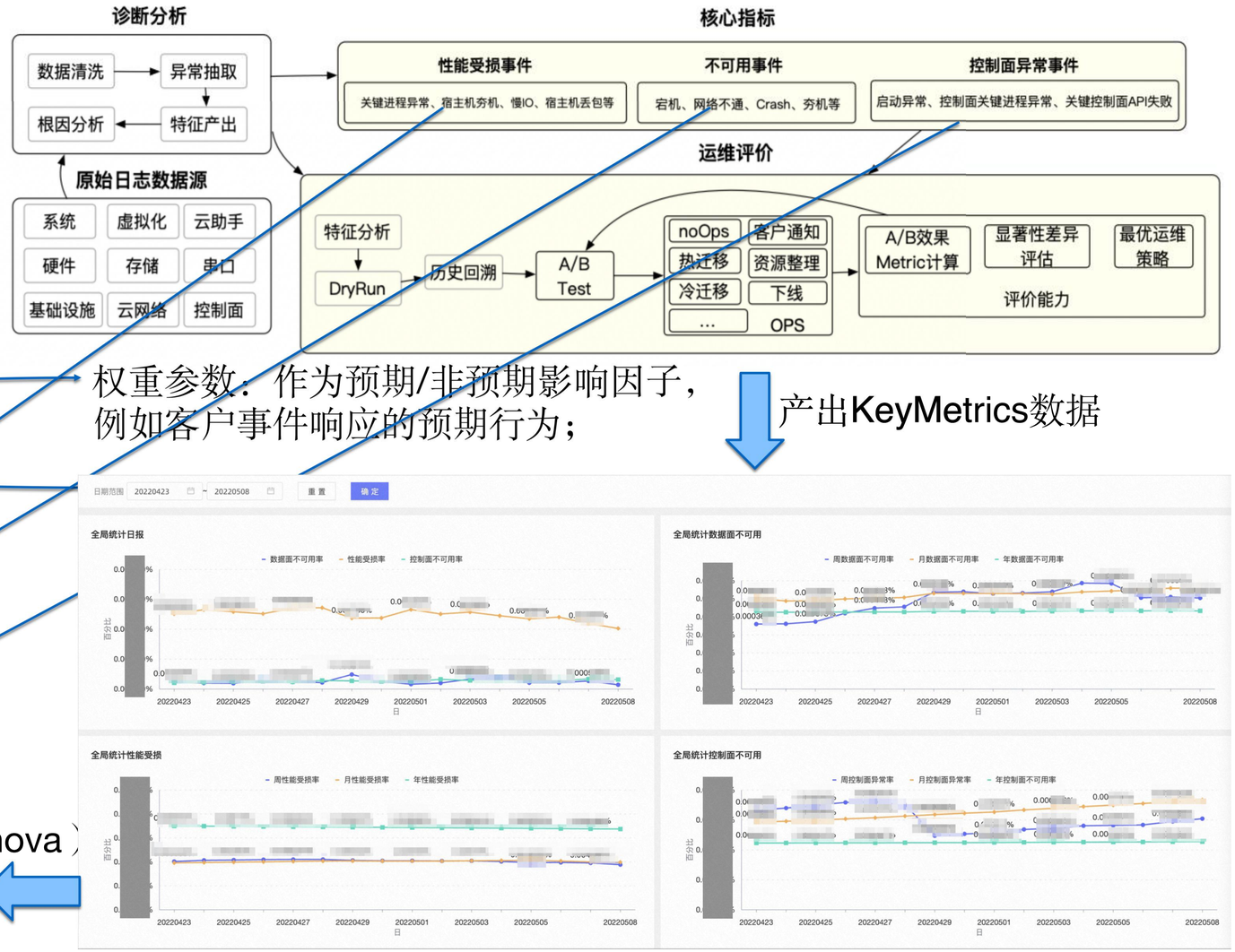
- 区分运维动作的“好坏优劣”？锁机器、热迁移、下线
- 是否存在过度运维的问题？
- 对客户真实体感是什么？

❑ 评价度量*：

- 性能度量 $$KeyMetric_{perf} = \frac{\sum_{l=1}^{l=n} F_l \sum_{i=1}^{i=n} F_e * Vm_i Perf_T}{\sum_{i=1}^{i=n} Vm_i TotalLifeTime_T}$$
- 不可用度量 $$KeyMetric_{down} = \frac{\sum_{i=1}^{i=n} F_e * Vm_i Down_T}{\sum_{i=1}^{i=n} Vm_i TotalLifeTime_T}$$
- 控制面度量 $$KeyMetric_{ctrl} = \frac{\sum_{l=1}^{l=n} F_l \sum_{i=1}^{i=n} F_e * Vm_i Ctrl_T}{\sum_{i=1}^{i=n} Vm_i TotalLifeTime_T}$$

❑ 差异化分析：

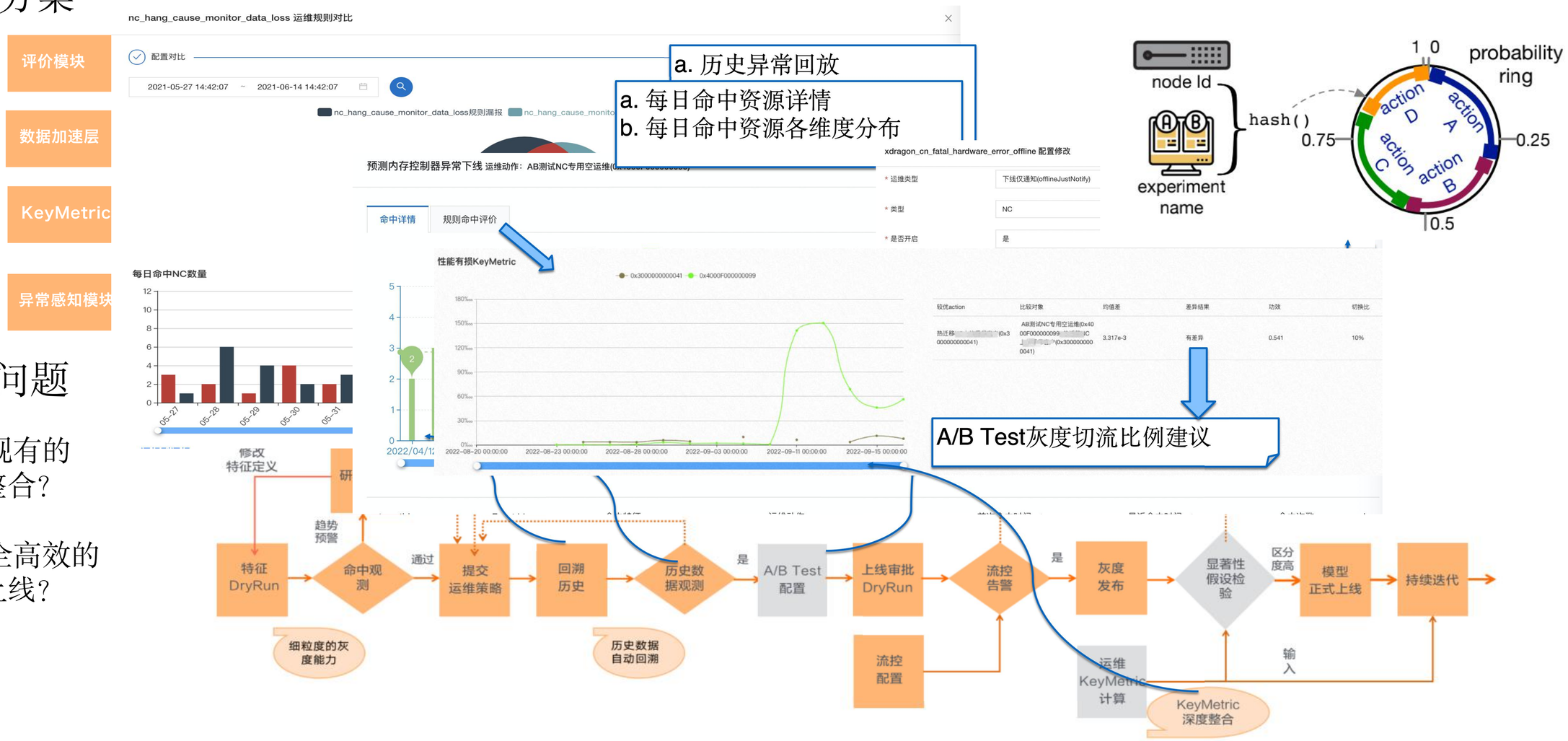
- 显著性差异检验 - 单因素方差分析F检验（Welch's anova）
- 精准控制切流比例 - 功效分析（Cohen’s f）



[*] Levy S, et al. Predictive and Adaptive Failure Mitigation to Avert Production Cloud VM Interruptions.[C]// Operating Systems Design and Implementation. 2020.

我们的方案 – 运维评价（工程落地）

❑ 落地方案



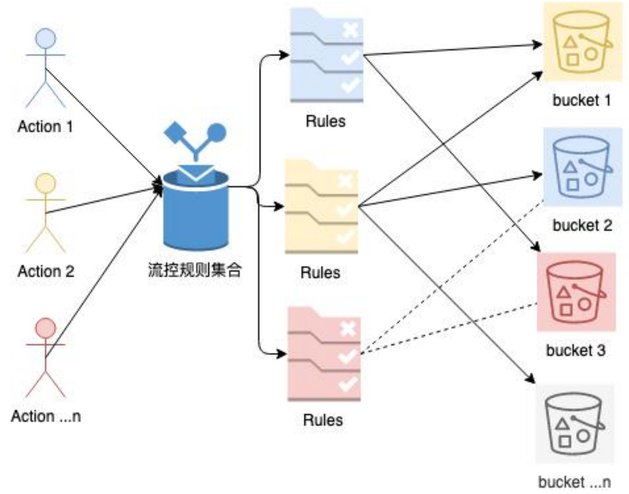
❑ 面临问题

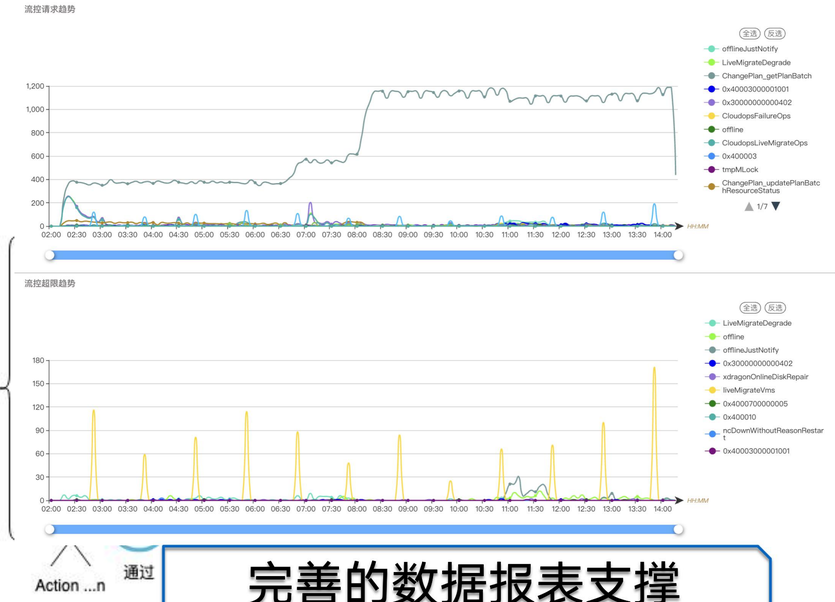
- 如何与现有的运维体系整合?
- 如何安全高效的灰度发布上线?

我们的方案 – 业务流控

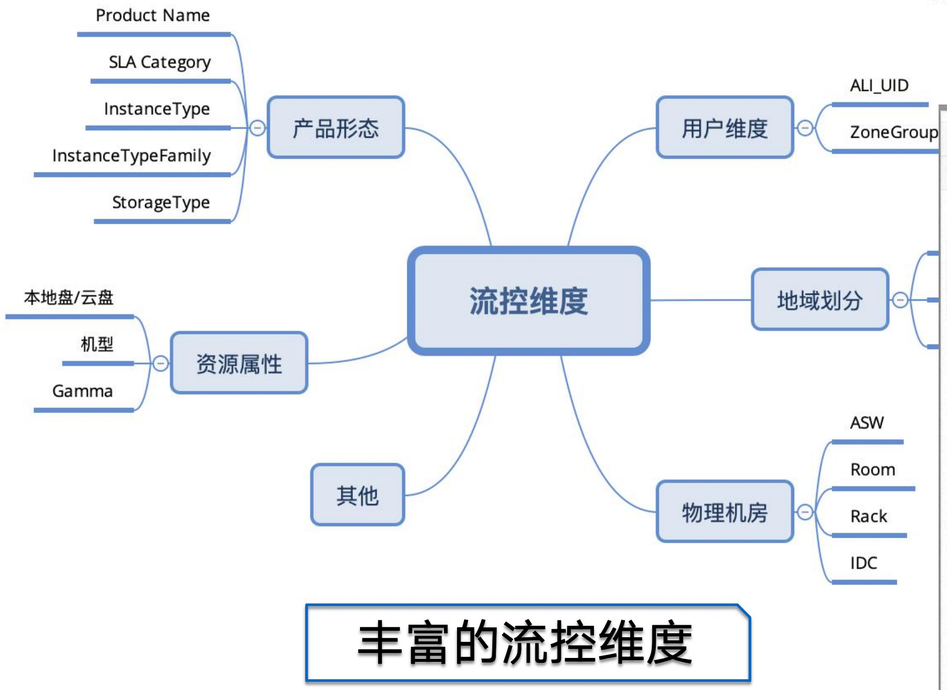
❑ 业务流控：

- 自动运维维持正常水位
- 有效阻止故障发生

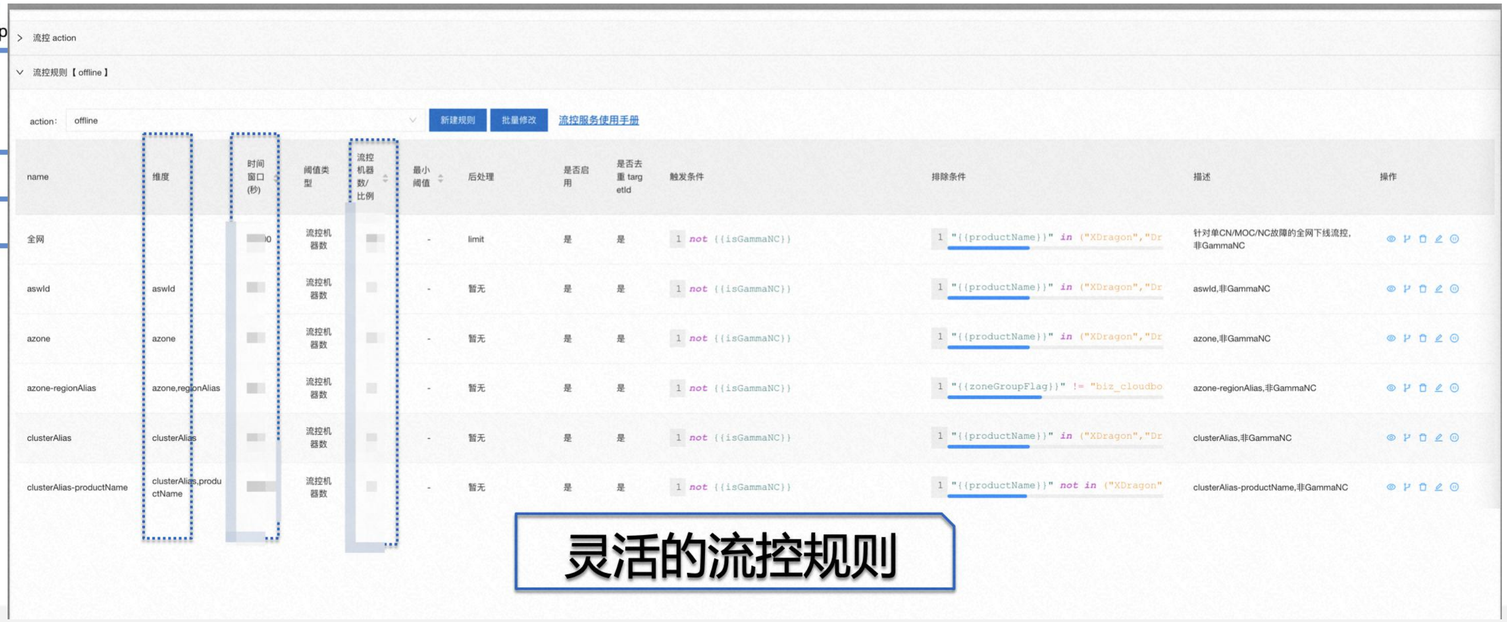




完善的数据报表支撑



丰富的流控维度



灵活的流控规则

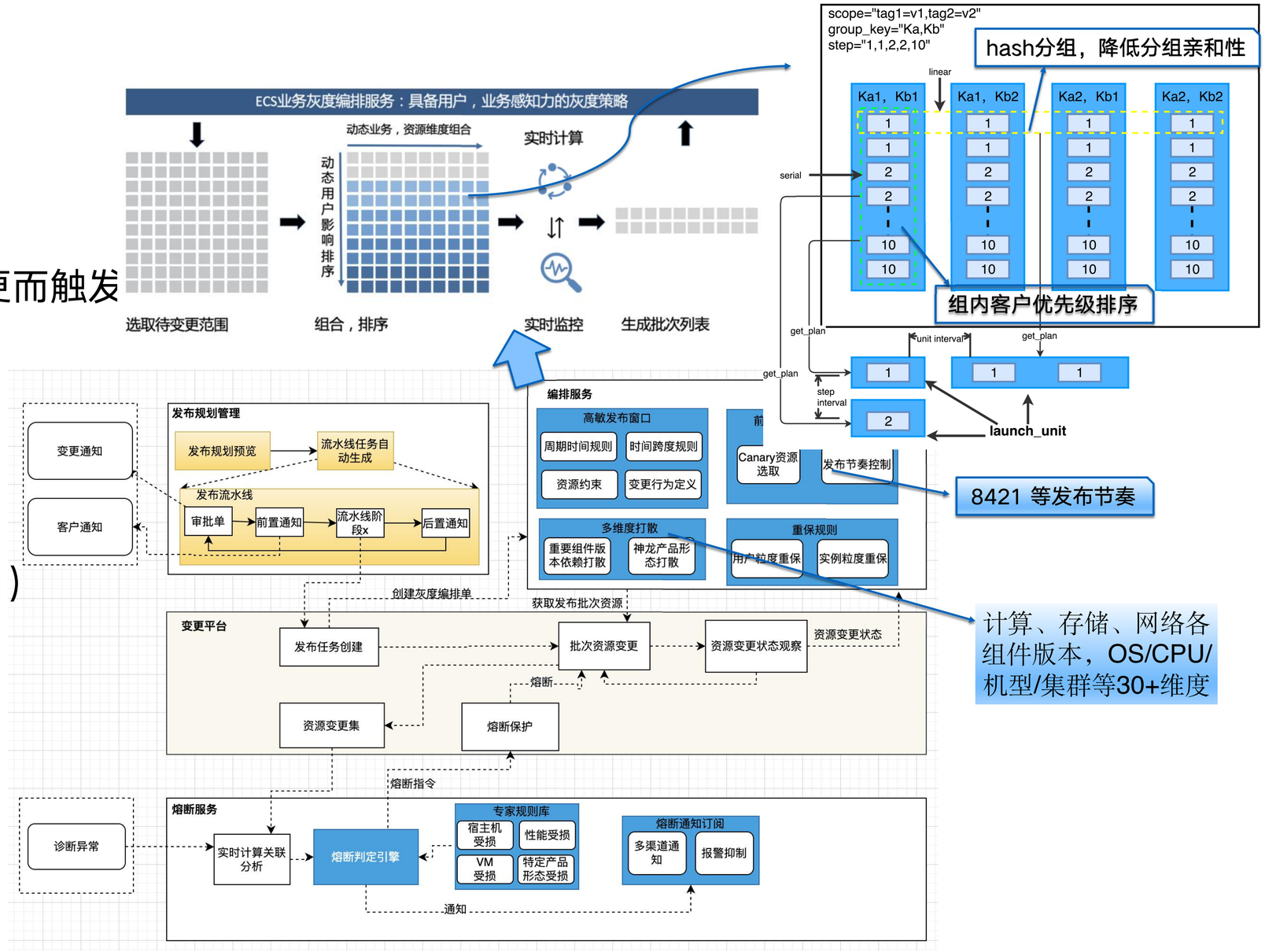
我们的方案 – 灰度发布

❑ 背景：

- 业界大概 XX % 的生产事故由变更而触发
- 集团全部故障中 xx %+和变更相关

❑ 业务：

- 支撑百万级资源的发布（千万级 VM）
- 发布业务方数百
- 累计变更次数 n 亿
- 发布次数 n 万



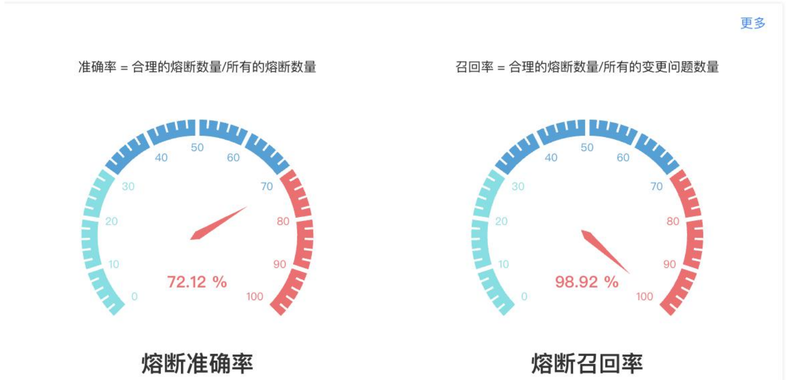
我们的方案 – 灰度发布&熔断

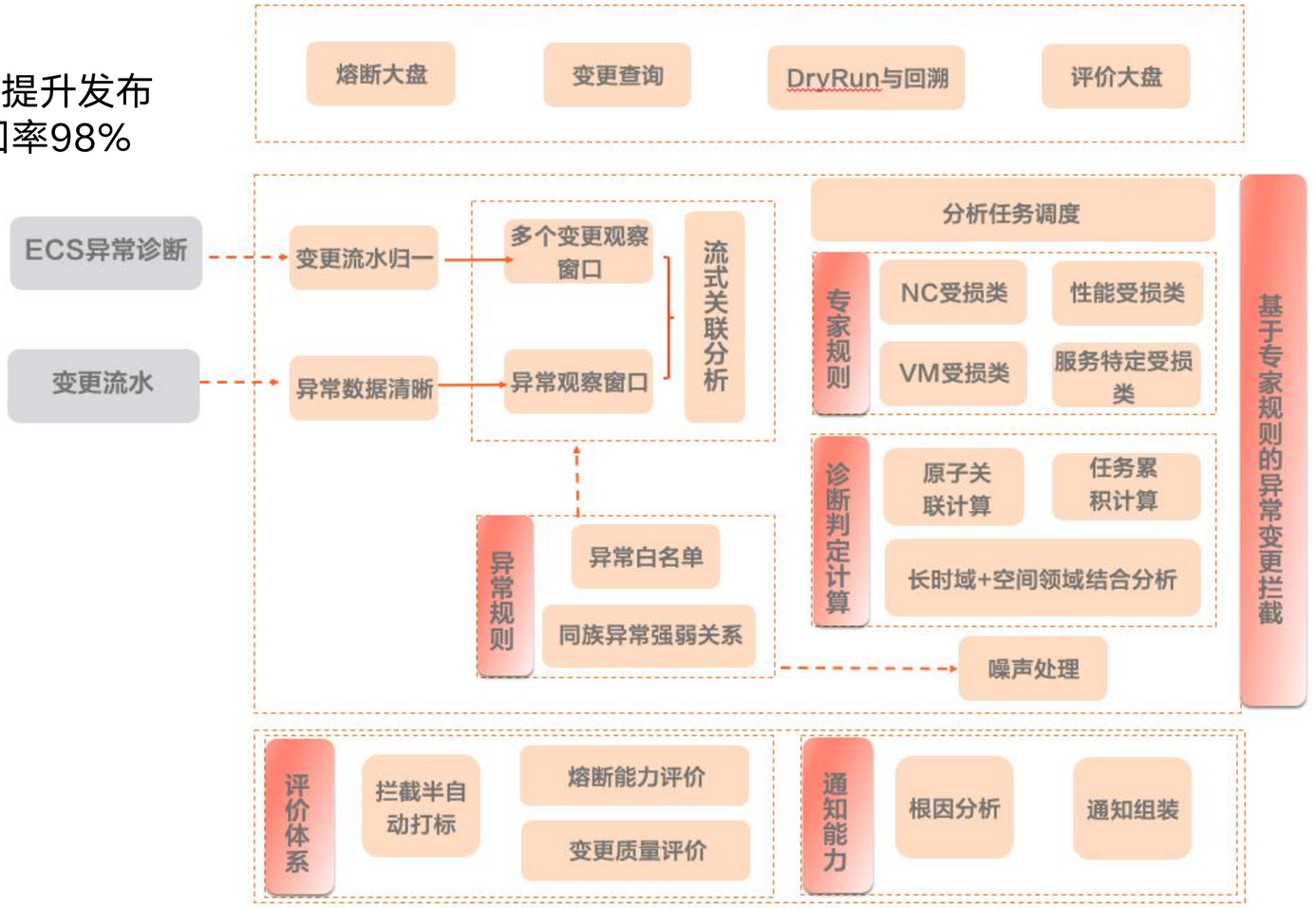
❑ 熔断能力

发布过程 ，通过诊断识别异常或隐患，主动阻断发布，是提升发布稳定性的利器，XX 期间，发布有效熔断近百次，有效召回率98%

$$召回率 = \frac{熔断次数}{熔断次数 + 漏召回次数} * 100\%$$

$$准确率 = \frac{有效熔断}{熔断次数 + 漏召回次数} * 100\%$$





[*] Li Z , et al. Gandalf: An Intelligent, End-To-End Analytics Service for Safe Deployment in Large-Scale Cloud Infrastructure[C]// Networked Systems Design and Implementation. 2020.

总结

❑ 概述

- 客户上云需求&稳定性述求
- 解决问题的难点

❑ 业界方案

- 监控告警体系演进

欢迎交流

❑ 自动化运维解决方案

- 监控采集：解耦、可降级、单元化、灰度
- 运维策略：全链路 Trace
- 运维评价：过度运维
- 稳定性利器：DryRun、A/B Test，流控、灰度、熔断

❑ 推荐更多内容

- Predictive and Adaptive Failure Mitigation to Avert Production Cloud VM Interruptions
- Gandalf: An Intelligent, End-To-End Analytics Service for Safe Deployment in Large-Scale Cloud Infrastructure
- Localizing Failure Root Causes in a Microservice through Causality Inference
- Predicting Node Failures in an Ultra-large-scale Cloud Computing Platform- an AIOps Solution

InfoQ

精彩继续！
更多一线大厂前沿技术案例

北京站

GMTC
全球大前端技术大会

时间：10月30-31日
地点：北京·国际会议中心

扫码查看大会详情>>

北京站

QCon
全球软件开发大会

时间：10月30-11月1日
地点：北京·国际会议中心

扫码查看大会详情>>

上海站

QCon
全球软件开发大会

时间：11月25-26日
地点：上海·宏安瑞士大酒店

扫码查看大会详情>>

想一想，我该如何把这些技术应用在工作实践中？

THANKS