CNUTCon 2017 全球运维技术大会

机器学习在大规模服务器治理复杂场景的实践

陈立波

阿里巴巴高级技术专家

主办方 Geekbang 极客邦科技 InfoQ

Geekbang.
极客邦科技

极客搜索

全站干货，一键触达，只为技术

s.geekbang.org

扫描二维码立即体验

有没有一种搜索方式，能整合 InfoQ 中文站、极客邦科技旗下12大微信公众号矩阵的全部资源?
极客搜索，这款针对极客邦科技全站内容资源的轻量级搜索引擎，做到了!

扫描上方二维码，极客搜索!

EGO EXTRA GEEKS' ORGANIZATION NETWORKS

这里只有
技术领导者

EGO会员第二季招募季正式开启

报名时间：9月1日-9月15日

扫描添加E小欧，

邀您进入EGO会员预报名群

立即报名

E小欧

TABLE OF **CONTENTS**

AIS为什么要引入机器学习

Case1：统一机型

Case2：批量问题管理系统

Case3：资源调度

如何支持百万级服务器？

更高效
发挥规模优势
不完全依赖人的能力

TABLE OF **CONTENTS**

AIS为什么要引入AI

Case1：统一机型

Case2：批量问题管理系统

Case3：资源调度

为什么要统一机型

现状：

- 曾经集团有数量众多的不同机型，相同机型下又细分众多不同的model
- 集团业务全部需要云化

机型无法收敛带来的问题：

- 资源个性化，碎片化，很难统一调度的
- 增加业务云化的难度
- 议价能力下降
- 稳定性差，故障多发
- 增加引入、测试和运维等隐性成本

期待

- 资源扁平化，简化调度器
- 资源不闲置，充分得到利用
- 成本最优化

算法模型选择

箱形推荐

- 结合业界供应现状，计算可能的CPU/MEM/DISK取值范围

装箱模型

- 在上述范围内进行迭代计算

目标函数设定

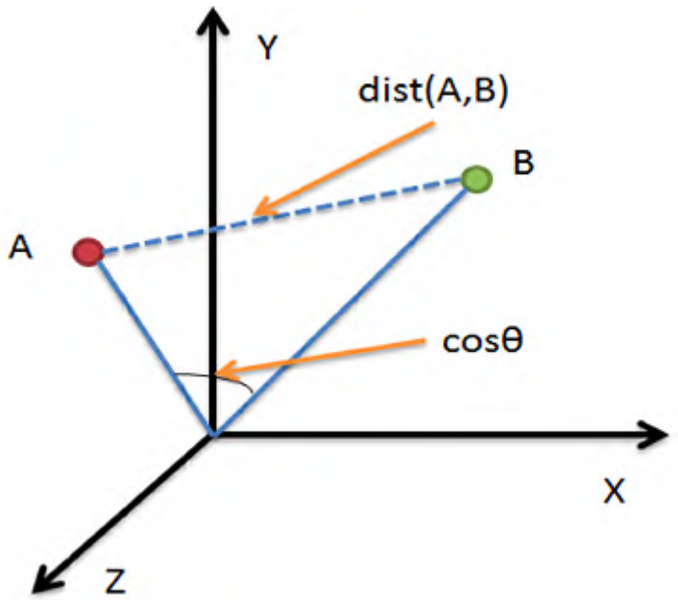
min Target = Num*TCO

机型装满所有容器需要的最小机器数量

机型的单机TCO成本

启发式优化算法

- 为虚拟机找到最适合的机器



- 寻找最佳的虚拟机部署顺序

落地还需要考虑的

- 真正需要计算的业务范围
- 已云化和未云化的资源使用情况
- 计算存储分离的影响
- 在离线混部的影响

得出满足业务需要的配置

TCO模型

- 机器采购成本
- 网络端口成本
- 机架租用成本
- 电力成本

冰山之下的工作

- 海量的数据获取
- 大量的数据校准
- 大量的算法及业务讨论

TABLE OF **CONTENTS**

AIS为什么要引入AI

Case1：统一机型

Case2：批量问题管理系统

Case3：资源调度

背景介绍

问题发现

用户反馈质量问题

问题分析

- 报修单查询
- 宕机日志查询
- 故障率查询
- 同类机型多维度查询
- 厂家FA

跟踪修复

- 运营修复
- 业务自己修复
- 无需修复
- 时间窗口不一致
- 修复进度

- 不能及时发现疑似批量问题隐患
- 人肉查询过程较多，影响效率(涉及系统:idcfree/report/os-log/armory)
- 维度灾难，机型*厂家*机房*业务*批次组合数量约125000000种，无法聚焦
- 容易被“脏数据”误导
- 无高质量样本
- 线上环境复杂，修复周期缓慢，新老问题交织

产品提供的能力

主动问题发现：

- 由被动处理改为主动发现
- 系统自动定位批量问题
- 给出真实和潜在影响范围
- 给出问题分析辅助决策

修复过程管理：

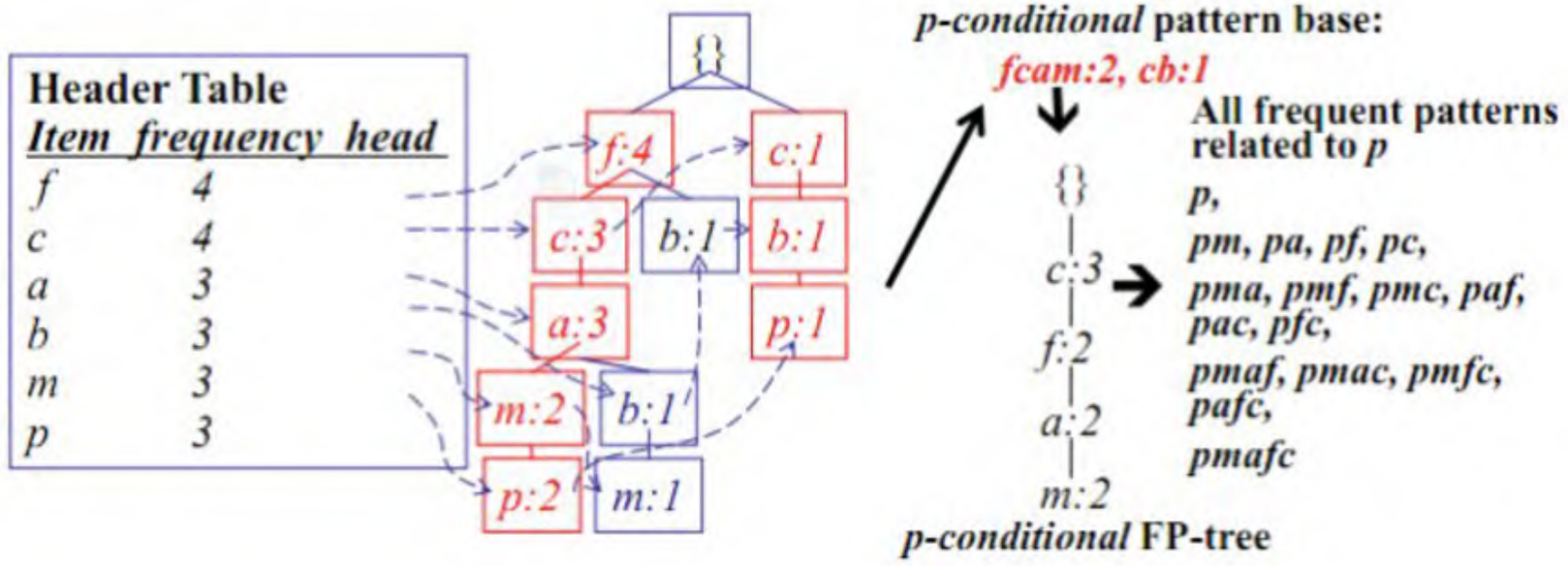
- 变更状态监控
- 用户通知和订阅
- 支持灰度和批量升级
- 进度管理
- 测试用例管理

批量问题的特点？

- 影响机器数量多
- 显著的设计或制造缺陷
- 特定场景下触发
- 常规容灾手段无法应对
- 业务可感知

模型选择

- 机型、机房、型号、批次、厂家、业务维度描述
- 最大频繁项：



影响范围评估

- 通过聚类出来的工单确定实际影响范围
- 通过聚类出来的维度确认潜在影响范围

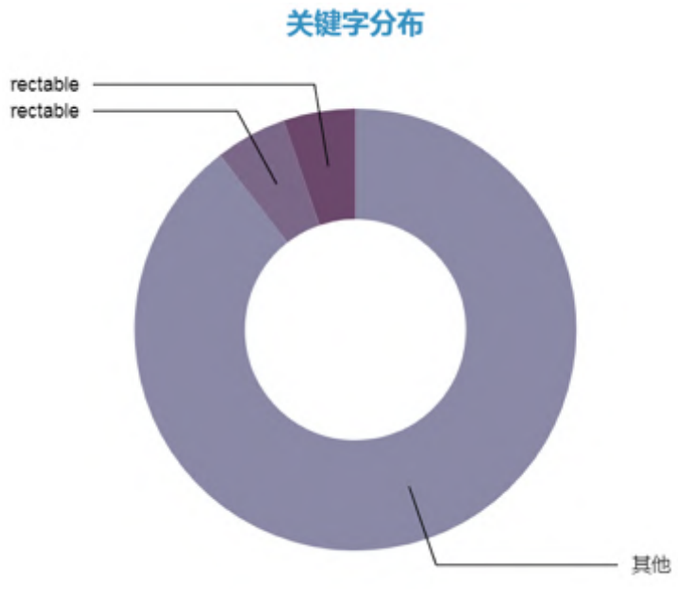
具体案例：

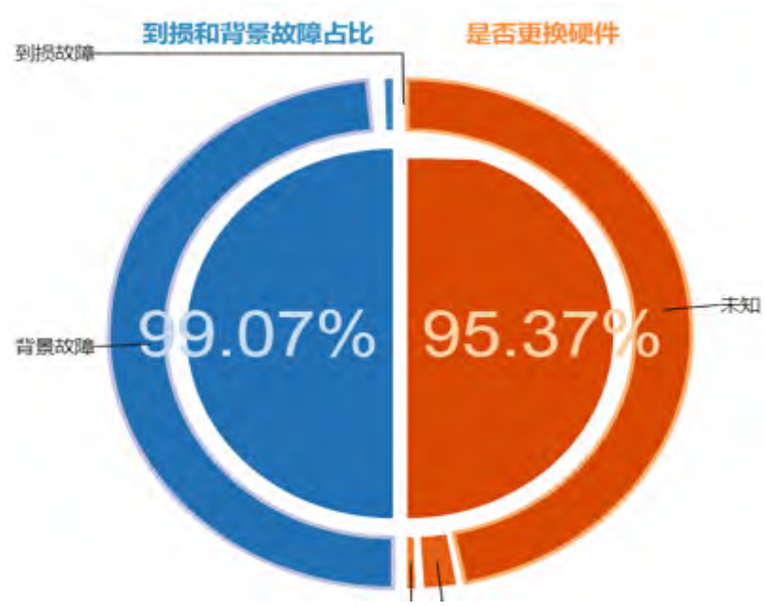
xx机型厂商xx机型xx部件厂商xx部件型号xx介质xxFWxx业务

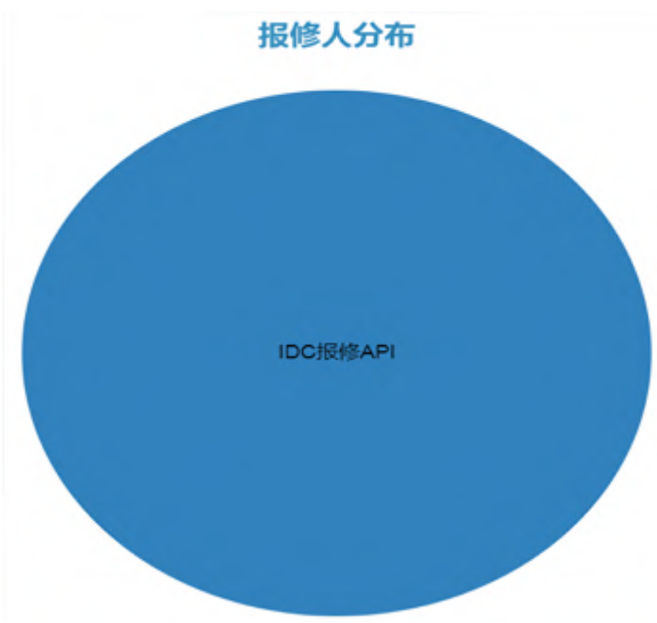
让数据可解释

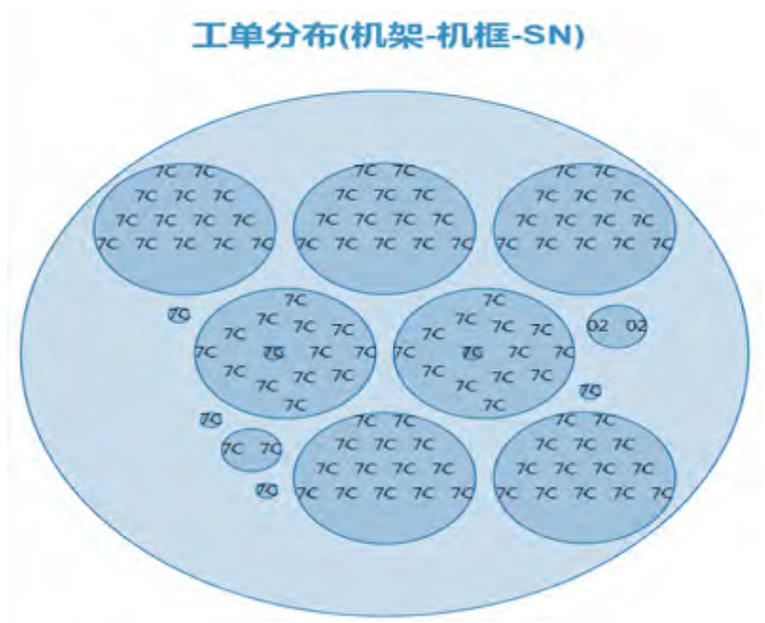
让数据可解释：

- 机器的分布
- 表现特征
- 故障来源
- 问题发生的阶段
- 历史表现
- 故障率



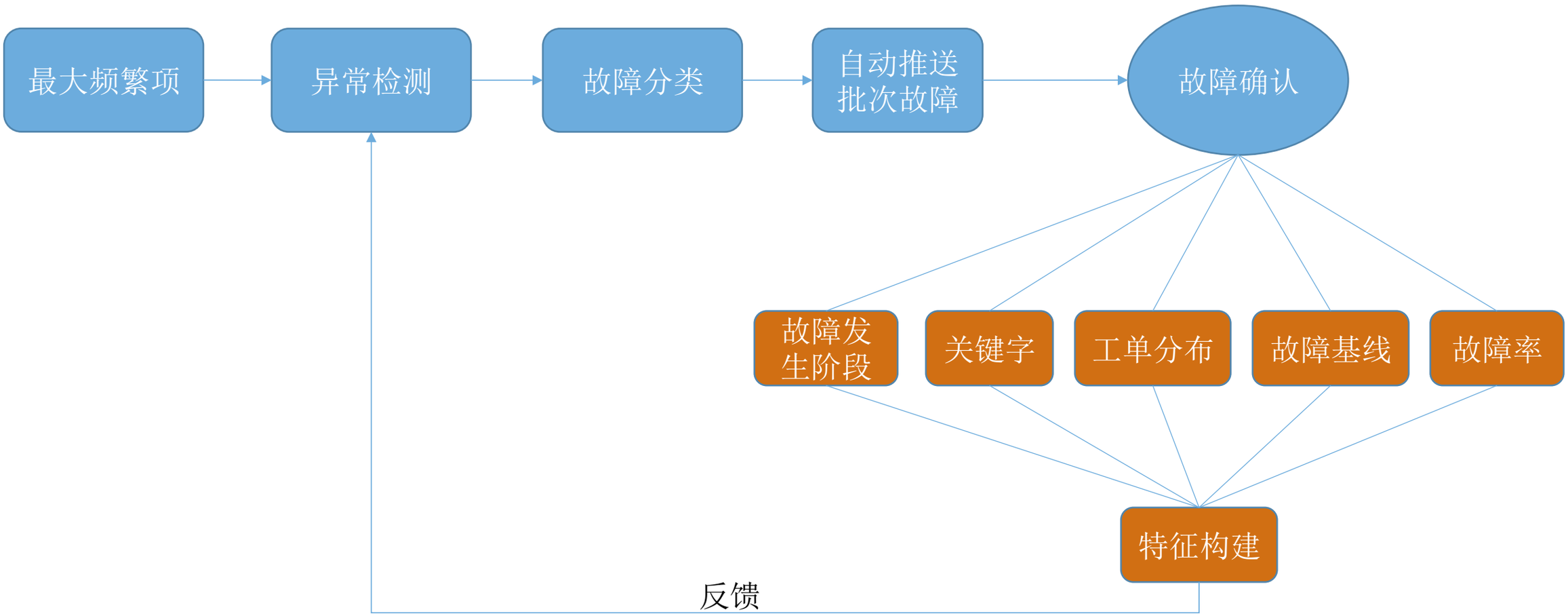






构建闭环

- 异常检测和分类



故障管理系统


服务器故障管理系统
数据视图
自动诊断视图(全面性/准确性/及时性)
批量问题处理状态视图
观察期视图
修复状态视图
智能化故障诊断
批量问题自动诊断
基于工单批量问题诊断
基于日志批量问题诊断
单机问题自动诊断(HWmonitor)
硬件传感器监控
日志监控
故障处理跟踪
修复状态跟踪
匹配规则管理
修复任务管理
自动升级/修复
故障观察期/关闭
观察期管理
单机故障跟踪与关闭（idcfree）
工单管理
更换/维修
维修服务审计
AVMS
测试用例
Bug管理
OS Portal
日志
通道
Aone
kelude
硬件决策平台
Firmware引入管理
固件管理
HWQC管理

任务进度设计

新建修复任务
辅助说明

影响范围圈定
辅助说明

方案制定
辅助说明

业务同步
辅助说明

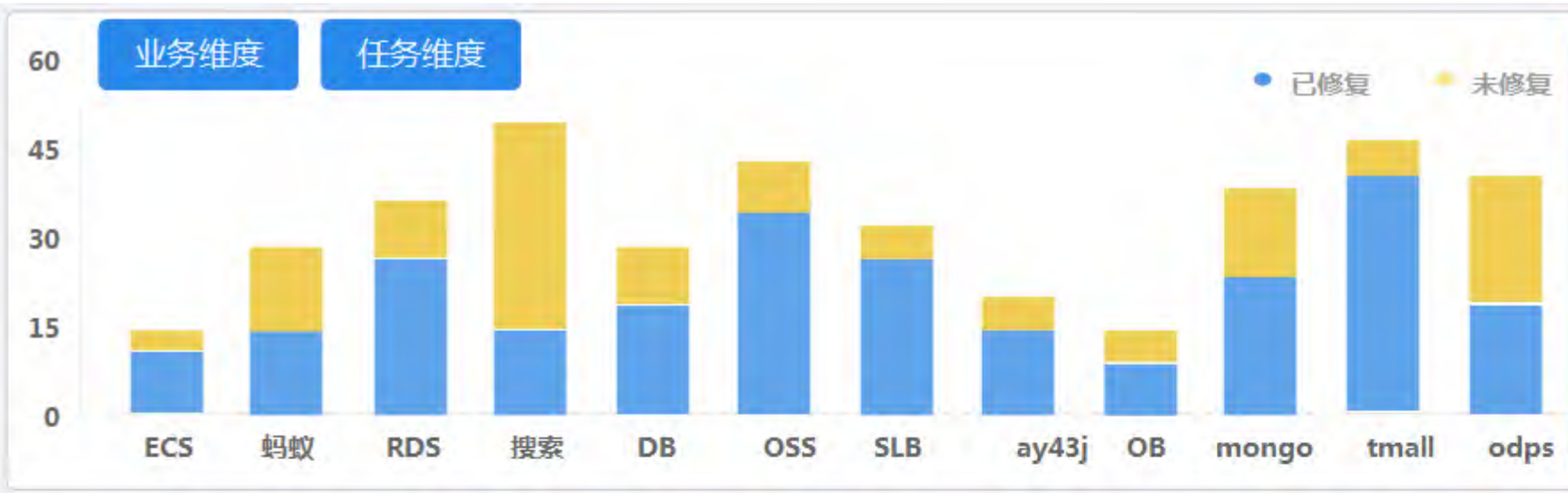
灰度方案实施
辅助说明

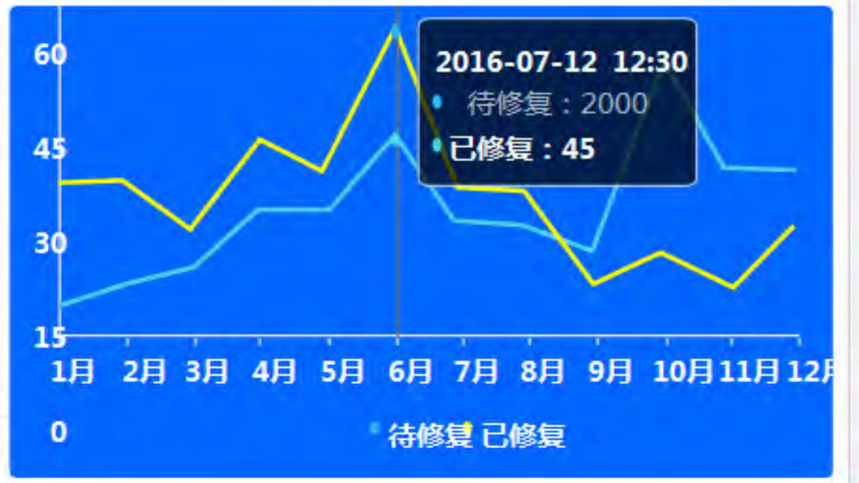
方案实施
辅助说明

观察期
辅助说明

关闭
辅助说明

灰度和批量升级





修复方案描述：

自定义监控指令： xxx　　升级检测指令： xxx　　升级检测

任务ID	任务名称	任务描述	故障类型	分组PE	状态	提交时间	服务器待修复/总	升级时间	任务类型	change free工单	操作
1022	a'li'yun-ay3d3 SSD 驱动升级	服务器厂商(INSPUR) 厂商机型(SA5248M4) 阿里机型(F41)	SAS/RAID卡故障	小明	升级中	2018-01-10 12:00:50	10023/10023	NA	灰度	工单	升级详情 | 导出sn | 删除
1023	a'li'yun-ay34rt SSD 驱动升级	服务器厂商(INSPUR) 厂商机型(SA5248M4) 阿里机型(F41)	机框故障	小明	完成升级	2018-01-10 12:00:50	32/4412	2018-3-10	批量	工单	自动升级 | 详情 | 删除
1024	OB-ay3d3 SSD 驱动升级	硬盘介质(HDD) 硬盘Connector(SATA) 硬盘FW(SN04) 硬盘厂商(SEAGATE)	带外故障	百事	升级中	2018-01-10 12:00:50	0/1000	2018-3-10	批量	24423	自动升级 | 详情 | 删除
1025	a'li'yun-ay3d3 SSD 驱动升级	服务器厂商(INSPUR) 厂商机型(SA5212M4)	CPU故障	马哥	观察中	2018-01-10 12:00:50	0/1000	2018-3-10	批量	13233	详情 | 删除
1026	蚂蚁OB-ay3d3 SSD 驱动升级	服务器厂商(INVENTEC) 厂商机型(K800G3-10G) 阿里机型(G42)	电源故障	芳芳	完成升级	2018-01-10 12:00:50	0/100	2018-3-10	批量	24423	详情 | 删除

TABLE OF **CONTENTS**

阿里为什么要引入AI

Case1：统一机型

Case2：批量问题管理系统

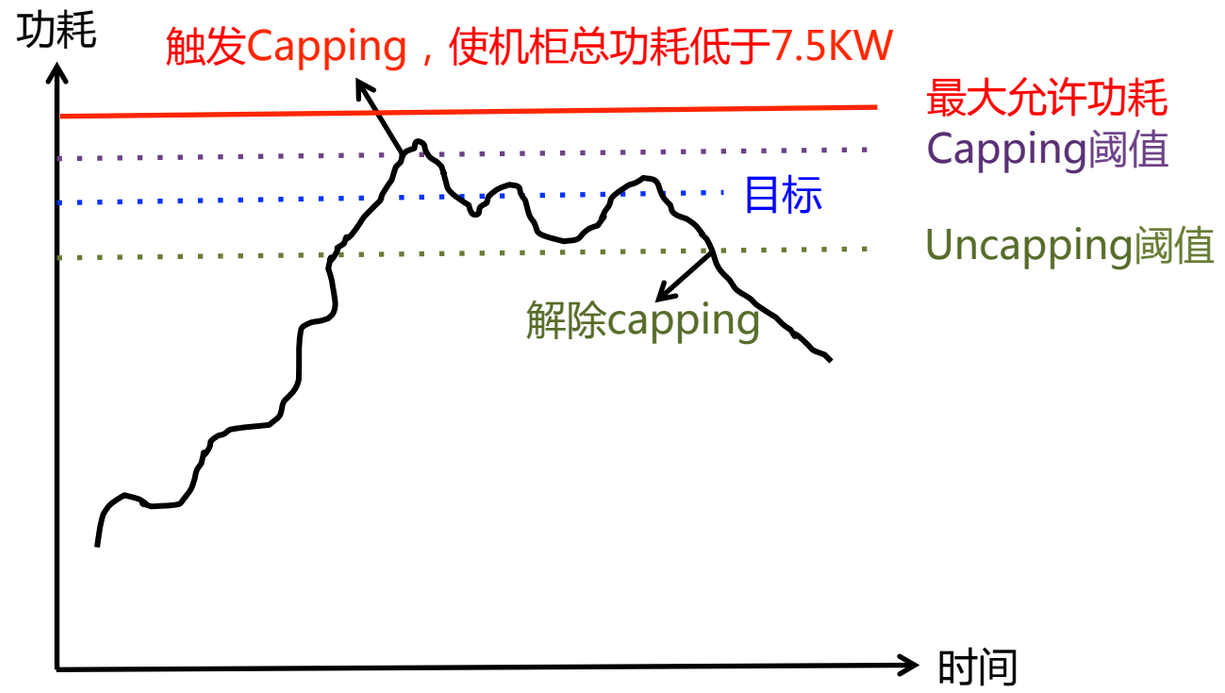
Case3：资源调度

资源调度的目标

- 计算、存储和网络资源能够最大化利用
- 综合成本最优
- 不牺牲稳定性
- 充分榨干水分

部分项目介绍

- 基于利用率的调度
- 基于功耗的调度
- 基于磨损调度



THANKS!

智 能 时 代 的 新 运 维

CNUTCon 2017