美团集群调度系统HULK技术演进

涂扬

美团点评基础架构部

Geekbang 极客邦科技 | TGO 鲲鹏會

TGO鲲鹏会

汇聚全球科技领导者的高端社群

全球12大城市

850+高端科技领导者

使命 Mission

为社会输送更多优秀的科技领导者

愿景 Vision

构建全球领先的有技术背景优秀人才的学习成长平台

扫描二维码，了解更多内容

自我介绍

- 自14年加入美团点评基础架构部以来，曾负责过基础架构部统一密钥管理服务Kms、分布式调用链CMtrace、消息中间件Mafka、容器集群调度系统HULK等项目的设计和落地，目前负责HULK弹性策略团队。

目录

- HULK架构演进
- 调度系统痛点、解法
- 弹性伸缩痛点、解法
- 经验总结

HULK项目

缘起：

容器实践：统一运行环境，提升交付效率。

弹性调度：提升业务的资源利用率。

命名由来：漫威里面的HULK在发怒的时候会变成绿巨人，这点和我们容器的“弹性伸缩”比较Match。

HULK的演进

线上9000+应用，
70000+容器

HULK 1.0

基于OpenStack演进

打通CMDB、服务治理、
发布平台、监控平台等，
验证容器的可行性

HULK 2.0

基于Kubernets演进

打磨弹性策略、调度系统
建设容器运营平台
基础系统软件加强
自研内核，提升安全隔离

HULK2.0架构图


安全
服务治理
发布平台
测试部署
OPS
监控管理平台
容器弹性
服务画像
容器编排
镜像管理
Docker
容器运行时
HULK agent

目录

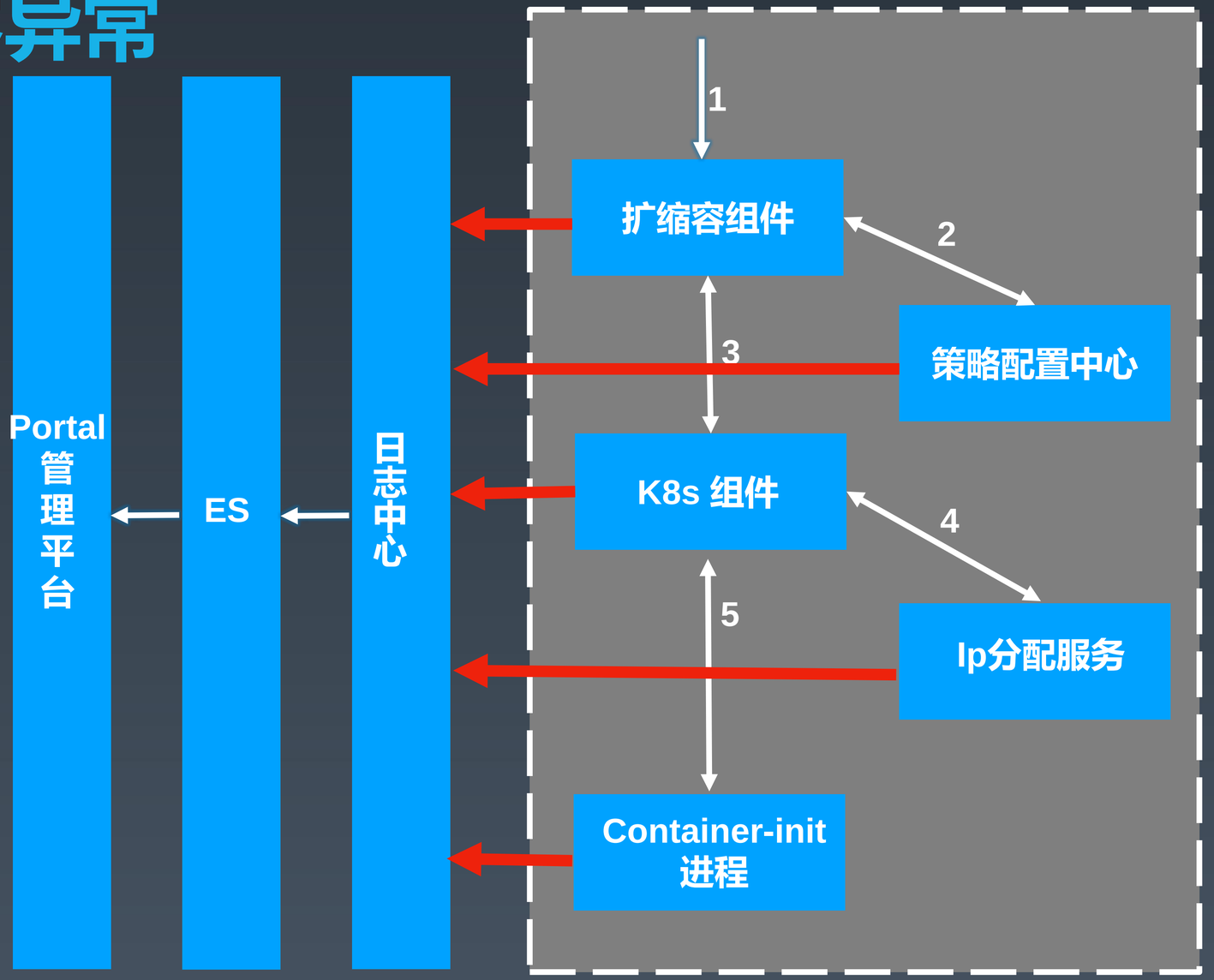
- HULK架构演进
- 调度系统痛点、解法
- 弹性伸缩痛点、解法
- 经验总结

调度系统-业务扩缩容异常

痛点：运维成本高

解法：
1. 全链路监控
2. 建设可视化平台Hulk-Portal

成效：
1.问题排查提效：多人联合花大半个小时到单人分钟级搞定
2.系统瓶颈可视化

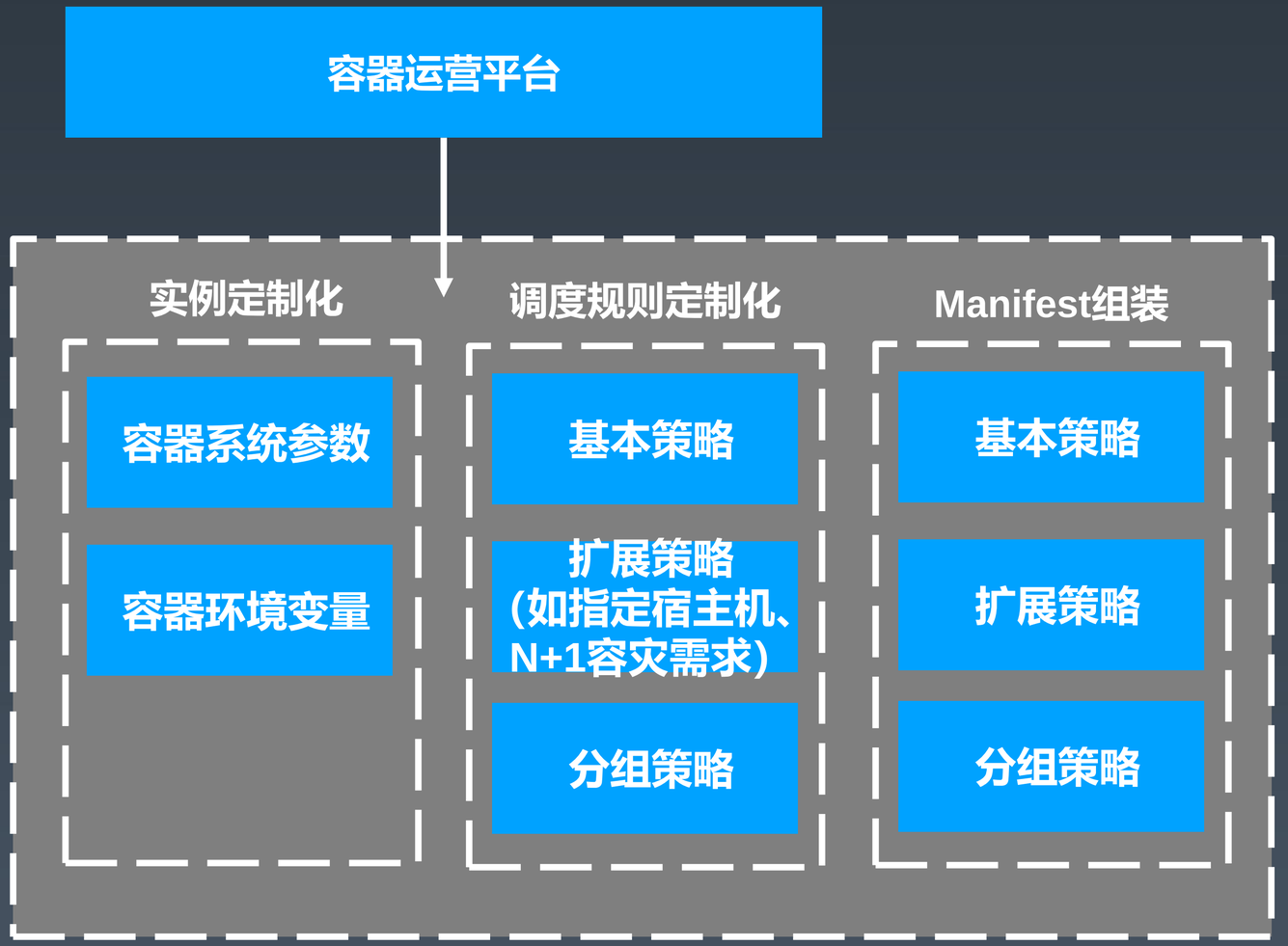


调度系统-业务定制化需求

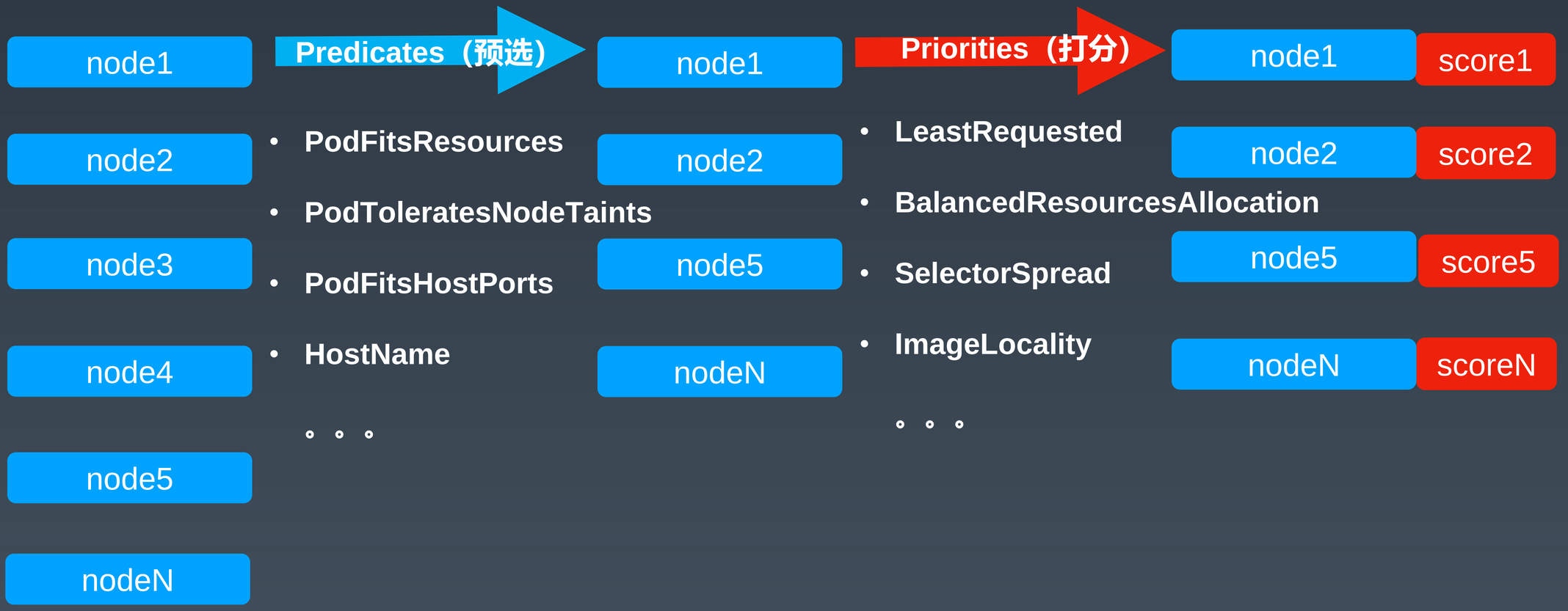
痛点：更改核心链路代码，灵活性不够

解法：建设一体化配置平台

成效：迈向自动化配置，解放运维人员。



调度系统-调度器策略


node1
node2
node3
node4
node5
nodeN
Predicates（预选）
PodFitsResources
PodTolerateNodeTaints
PodFitsHostPorts
HostName
。。。
node1
node2
node5
nodeN
Priorities（打分）
LeastRequested
BalancedResourcesAllocation
SelectorSpread
ImageLocality
。。。
node1
score1
node2
score2
node5
score5
nodeN
scoreN

调度系统-调度器性能问题

痛点：3000台集群规模，一次Pod调度耗时5s左右（k8s 1.6版本）

解法：预选失败中断机制

成效：生产环境验证，提升性能40%。（PR 56926，社区1.10版本作为默认调度策略）

调度系统-调度器性能问题

痛点：BestFit代价高

解法：局部最优

成效：大大减少调度时间，同时对调度结果未产生较大影响。（和社区合作共同完成，PR 66733/67555，社区1.12版本作为默认调度策略）

调度系统- kubelet的自决策问题

痛点：

1. 容器重启/迁移问题：
   1.1. 容器和系统盘的信息丢失。
   1.2. 容器的IP也变更了。

2. 驱逐策略问题：
   Kubelet会自动杀死一些违例容器，但是有可能这个业务是非常核心的业务。

解法：

1. 容器重启/迁移
   1.1. 新增Reuse策略，保留原生重启策略（Rebuild）。
   1.2. 自研CNI插件，基于Pod标识申请和复用IP。

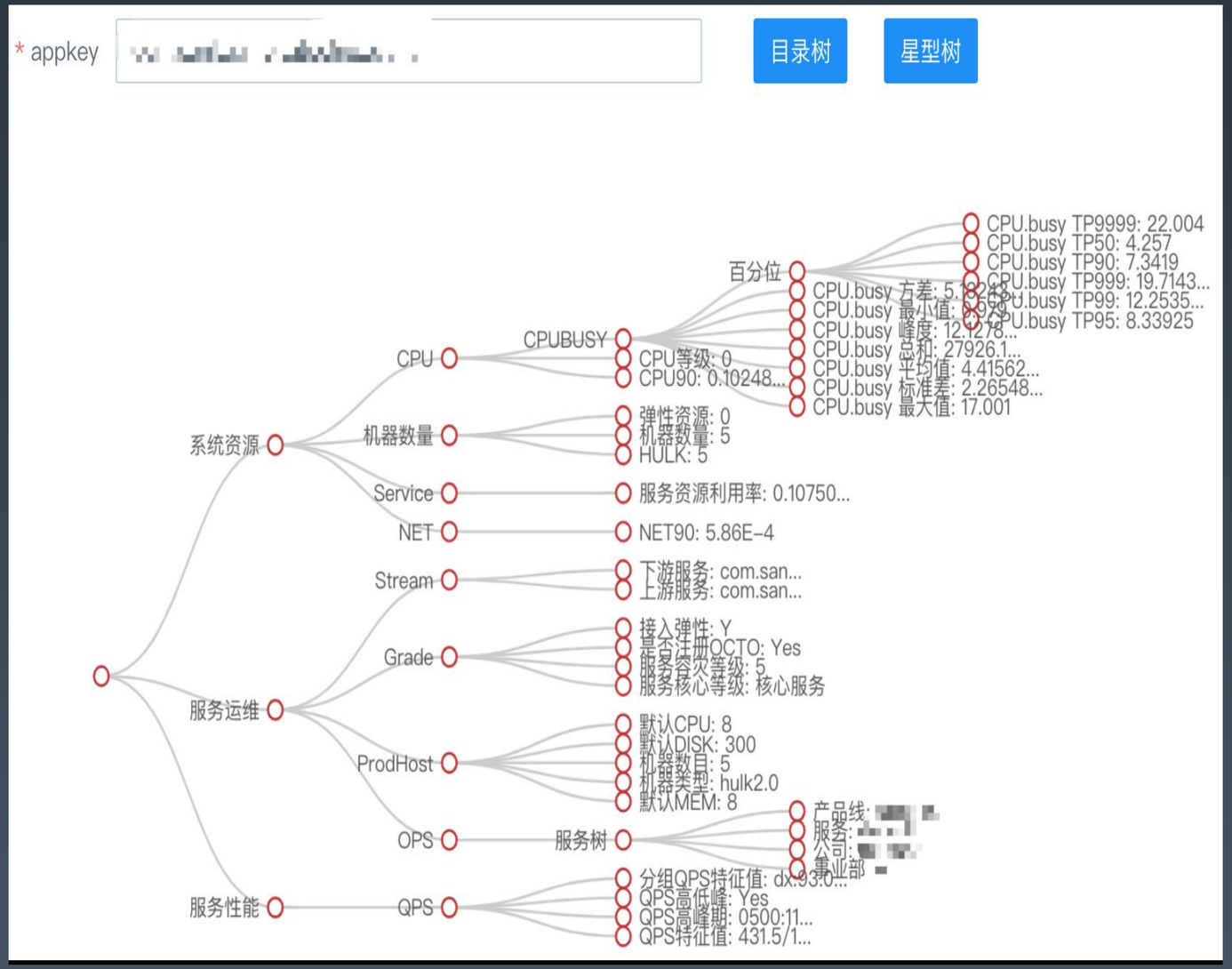
2. 限制原生的驱逐策略

调度系统-调度决策难题

痛点：资源最大化和SLA保障

解法：服务画像，供能于调度前决策、调度后决策。

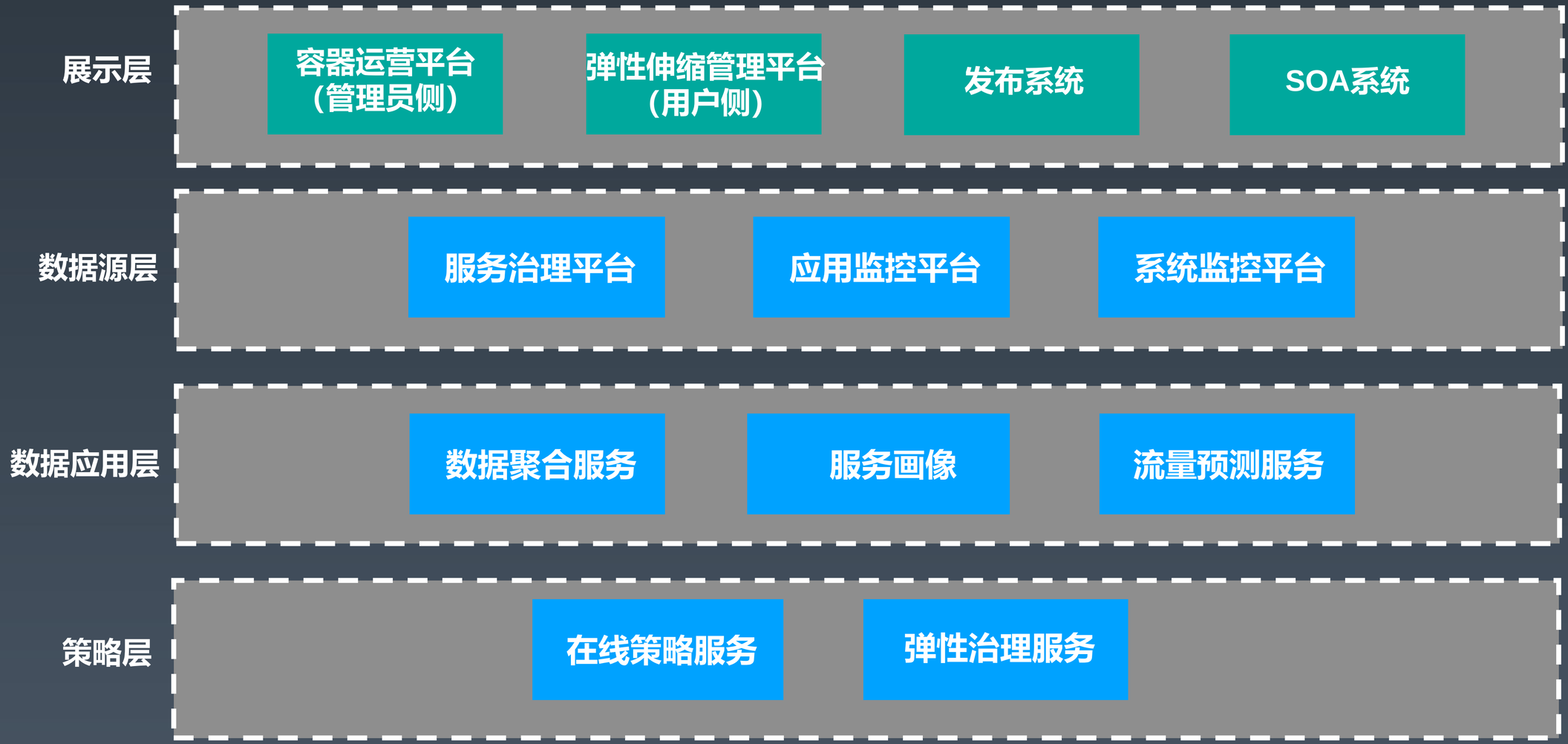
成效： 40+基础标签，N+聚合标签，其中不少标签已经成为调度决策的重要因素。



目录

- HULK架构演进
- 调度系统痛点、解法
- 弹性伸缩痛点、解法
- 经验总结

弹性伸缩平台架构图


展示层
容器运营平台（管理员侧）
弹性伸缩管理平台（用户侧）
发布系统
SOA系统
数据源层
服务治理平台
应用监控平台
系统监控平台
数据应用层
数据聚合服务
服务画像
流量预测服务
策略层
在线策略服务
弹性治理服务

弹性伸缩痛点

多策略
决策不一致

扩缩不幂等

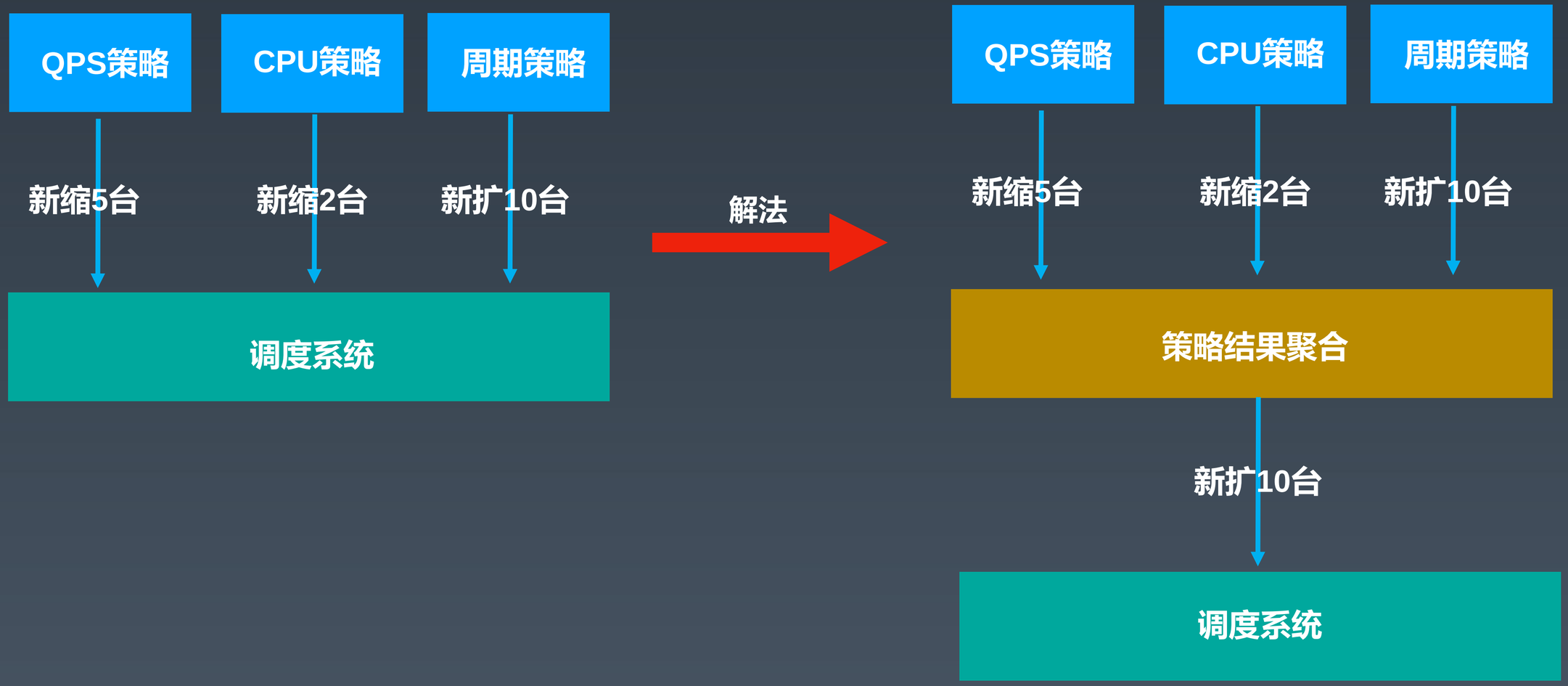
线上
代码多版本

资源
保障问题

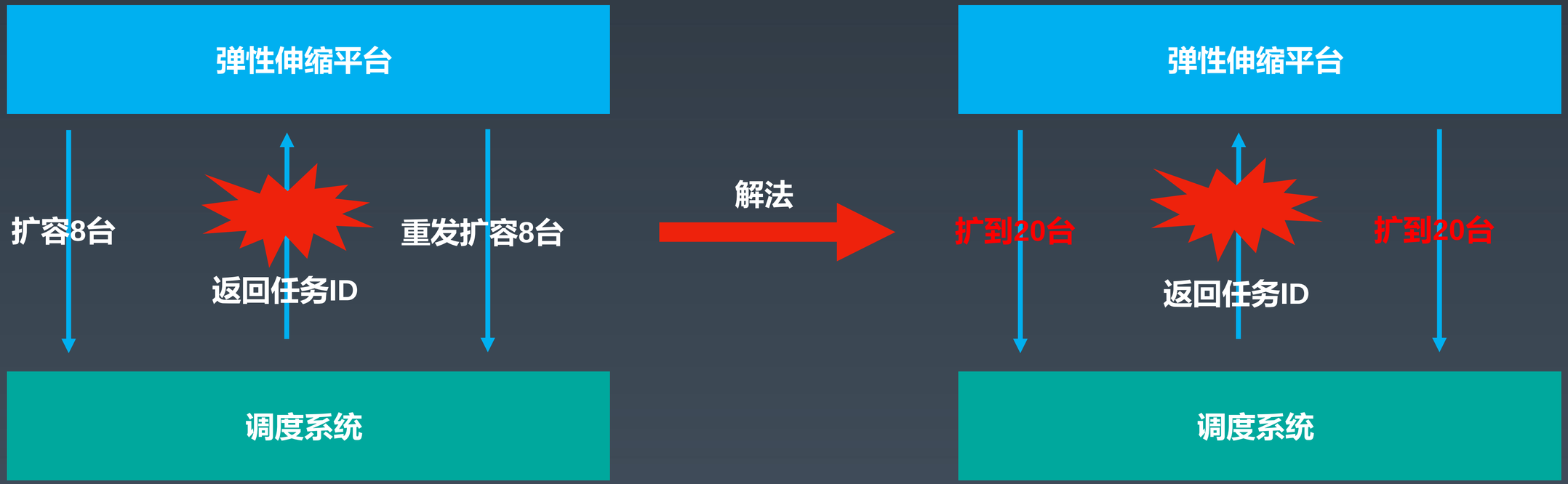
端到端时效

问题实例隔离

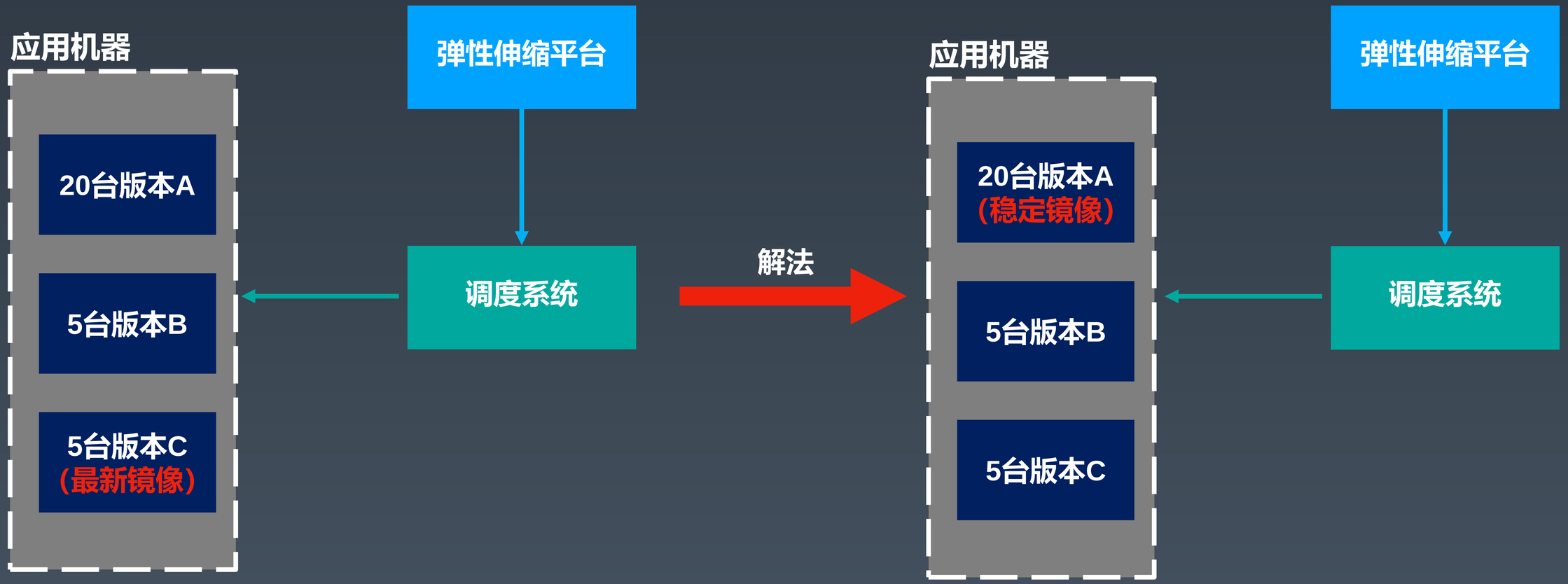
弹性伸缩痛点-多策略决策不一致



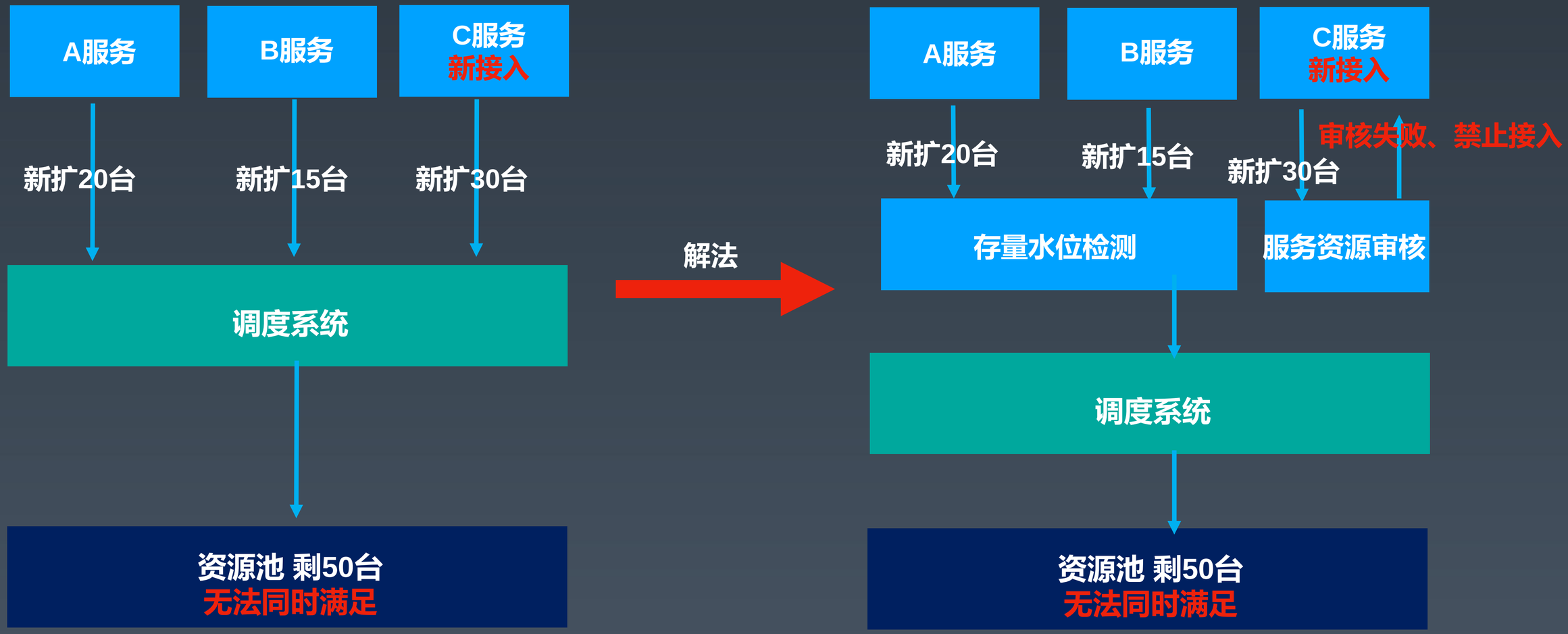
弹性伸缩痛点-扩缩不幂等



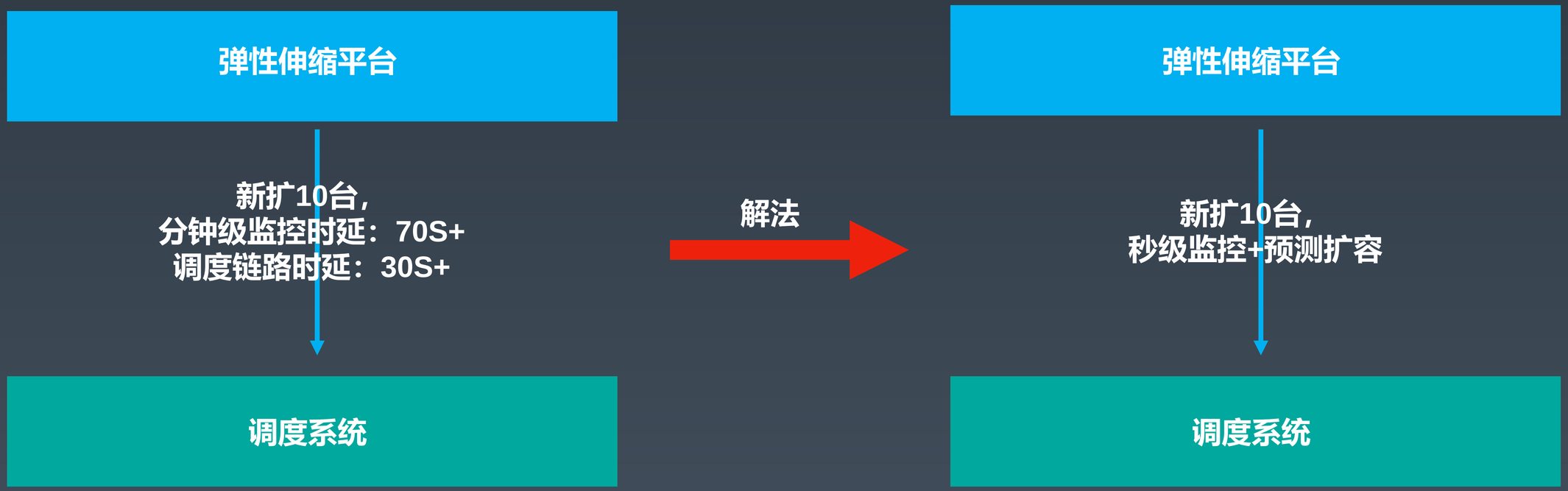
弹性伸缩痛点-线上代码多版本



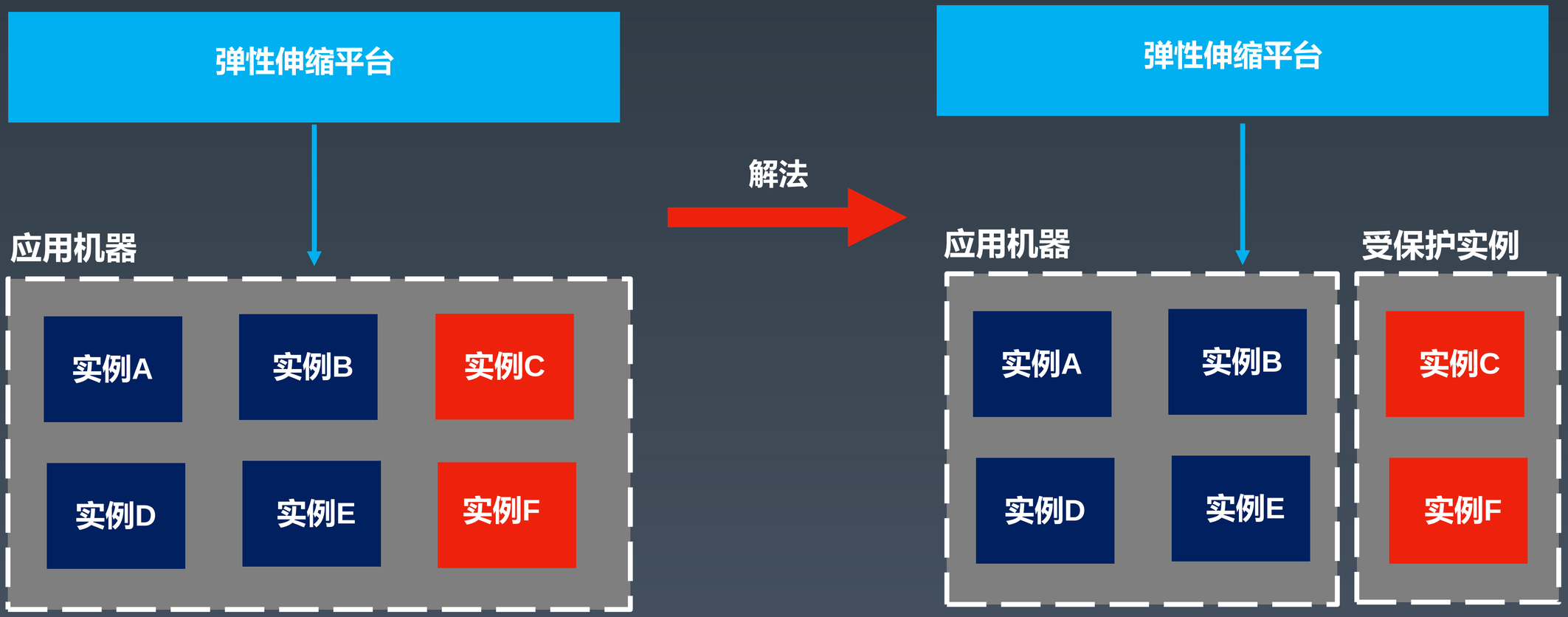
弹性伸缩痛点-资源保障问题



弹性伸缩痛点-端到端时效



弹性伸缩痛点-问题实例隔离



目录

- HULK架构演进
- 调度系统痛点、解法
- 弹性伸缩痛点、解法
- 经验总结

经验总结

- 开源产品“本土化”： 服务树、发布系统、服务治理平台、监控系统等的融合
- 来自业务的挑战：稳定性 && 人力投入
- 给到业务的收益：效率 && 成本节省
- 调度决策：规范增量 && 重调度存量
- 弹性伸缩：扩容成功率 && 时延

极客时间 | 企业服务

这是一封正经“福利”贴

想做团队的领跑者 需要迈过这些“槛”

成长型企业，易忽视人才体系化培养 企业转型加快，团队能力又跟不上	VS	从基础到进阶，超100+一线实战 技术专家带你系统化学习成长
团队成员技能水平不一， 难以一“敌”百人需求	VS	解决从小白到资深技术人所遇到 80%的问题
寻求外部培训，奈何价更高且 集中式学习	VS	多样、灵活的学习方式，包括 音频、图文 和视频
学习效果难以统计，产生不良循环	VS	获取员工学习报告，查看学习 进度，形成闭环

课程顾问「橘子」

回复「QCon」
免费获取
学习解决方案

 # 极客时间企业账号 # 解决技术人成长路上的学习问题

GTLC
GLOBAL
TECH LEADERSHIP
CONFERENCE
全球技术领导力峰会

Geekbang. 极客邦科技 | TGO 鲲鹏會

500+高端科技领导者与你一起探讨
技术、管理与商业那些事儿

2019年6月14–15日 | 上海圣诺亚皇冠假日酒店

扫码了解更多信息

THANKS!

QCon 10th