

美团点评餐饮开放平台的高可用实践

餐饮生态技术部－朱洋波

2018.01.13



新到店事业群 | 餐饮生态技术部

自我介绍

2007.7月硕士毕业于北京大学摄影测量与遥感专业。

2015.1月加入美团基础服务组，研发分布式key-value系统，见证了系统从几万到千亿的日均调用量。

2016.5月加入新美大餐饮生态技术部，现负责开放平台和会员系统的研发管理工作。
加入美团之前，分别在腾讯和360任职多年，一直负责高性能的后台服务研发工作。

兴趣爱好：篮球，音乐，旅行

大纲

- 一.高可用的定义
- 二.美团点评餐饮开放平台简介
- 三.高可用之服务容错
- 四.高可用之服务优化
- 五.高可用之监控报警
- 六.建立适当的流程规范

三个故事

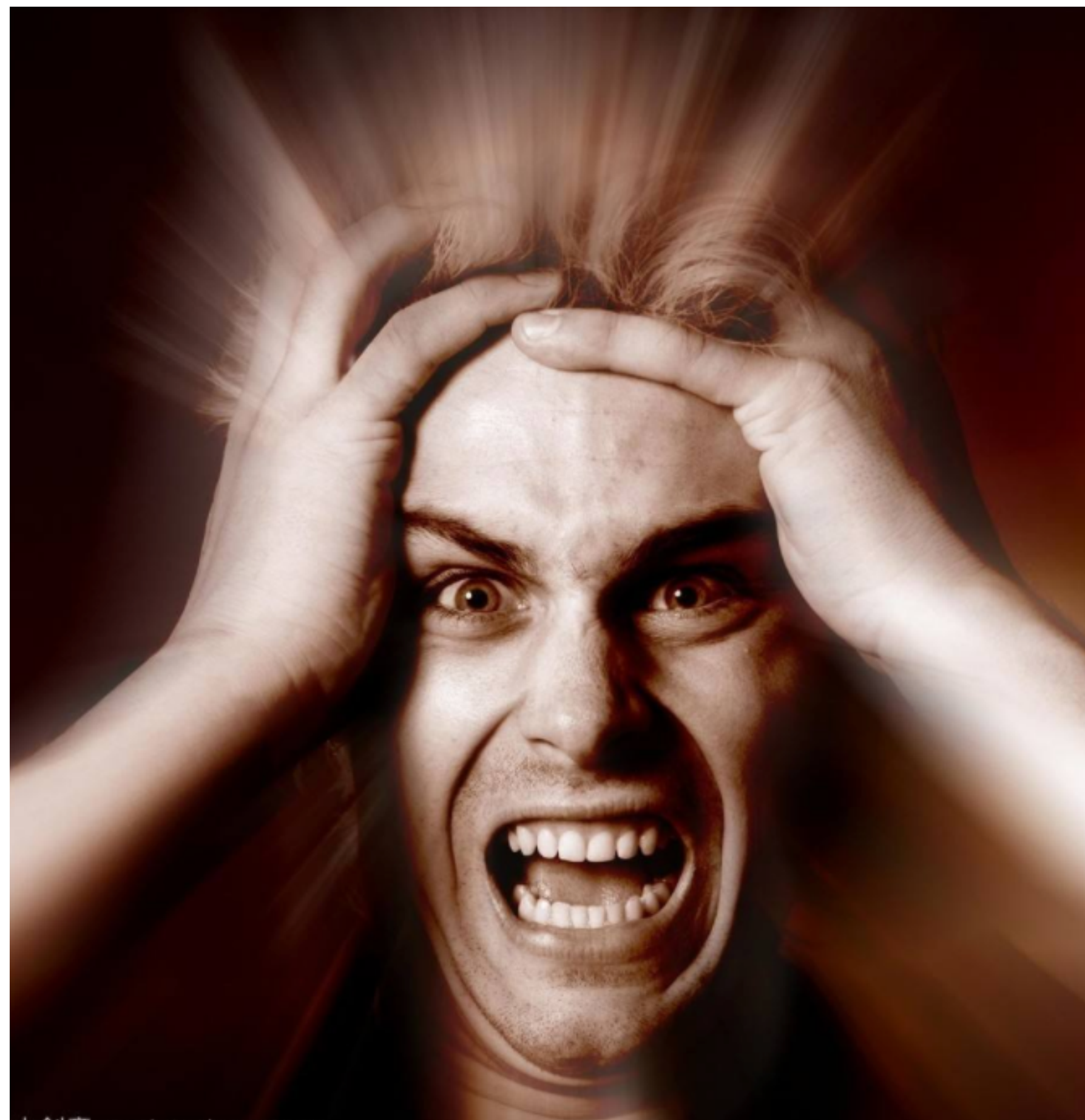
一.发布程序时手动备份程序的故事

二.前端徽章服务同步调用后端的故事

三.营销活动的故事



小问题引起大事故



一.高可用的定义

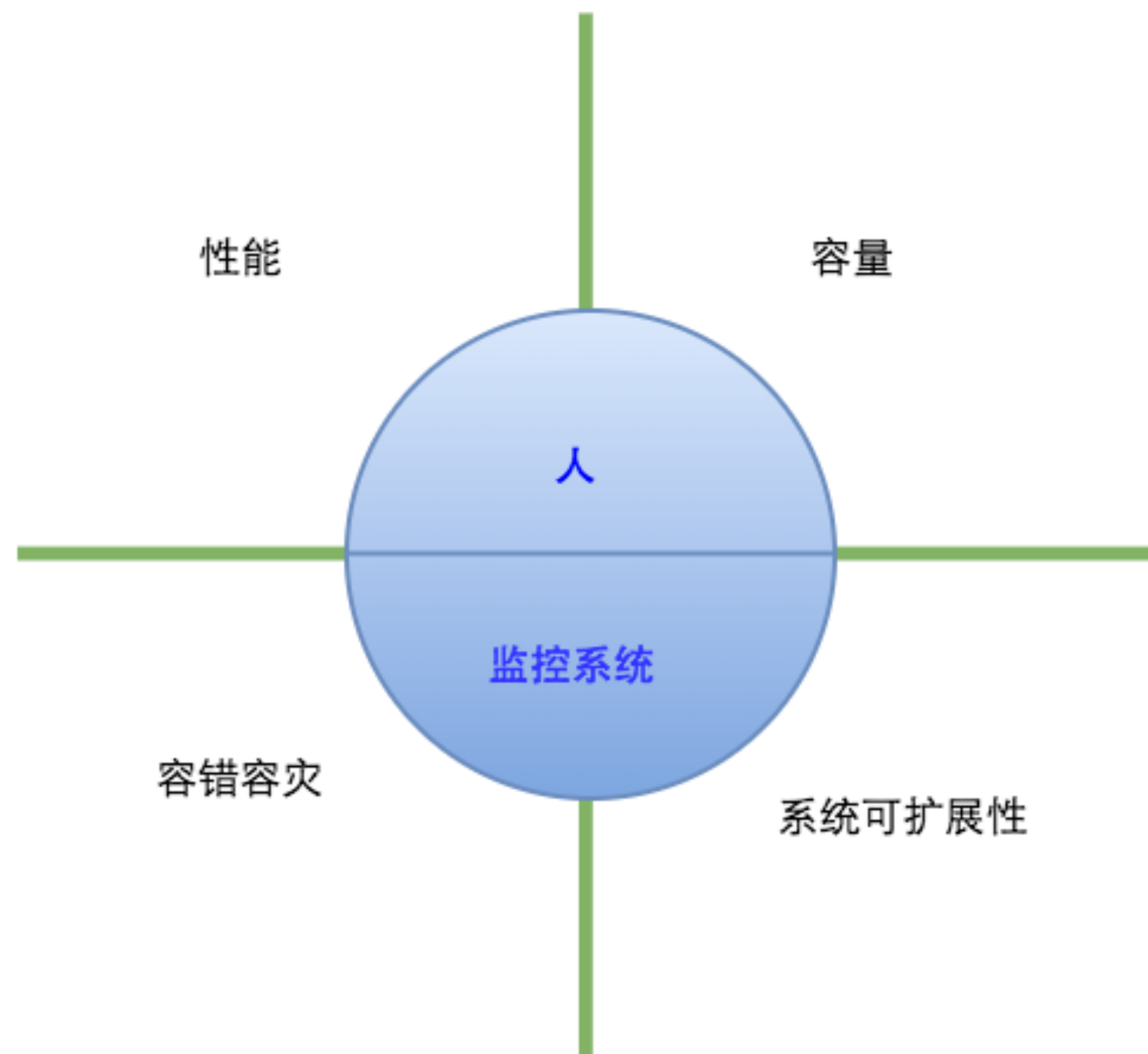
服务的高可用定义

- 平台的可用性：平台按照预先定义好的服务等级协议(SLA) 提供服务的可能性
- 计算公式： $\text{Availability} = \text{MTBF} / (\text{MTBF} + \text{MTTR})$
- MTBF：平均无故障时间； MTTR：平均恢复时间
- 一般我们采用：在一个特定的时间范围内，网站可用时间长度除以总的时间

系统可用性%	宕机实践/年	宕机实践/月	宕机实践/周
90 (1个9)	36.5天	72小时	16.8小时
99 (2个9)	3.65天	7.2小时	1.68小时
99.9 (3个9)	8.76小时	43.8分	10.1分钟
99.99 (4个9)	52.56分	4.38分	1.01分钟
99.999 (5个9)	5.26分	25.9秒	6.05秒

每增加一个9，难度都相当大

如何构建一个高可用的系统

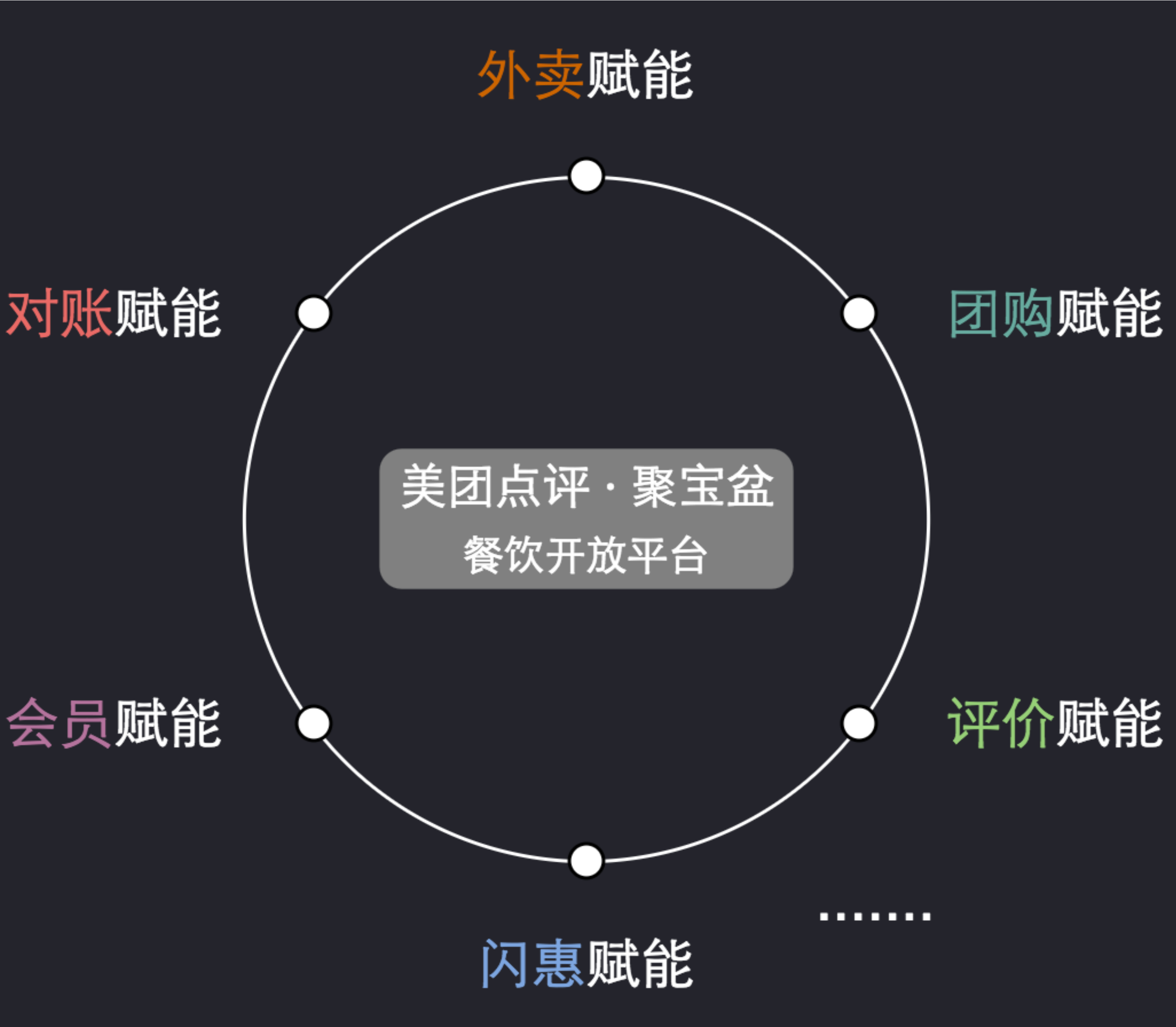


二.美团点评餐饮开放平台简介

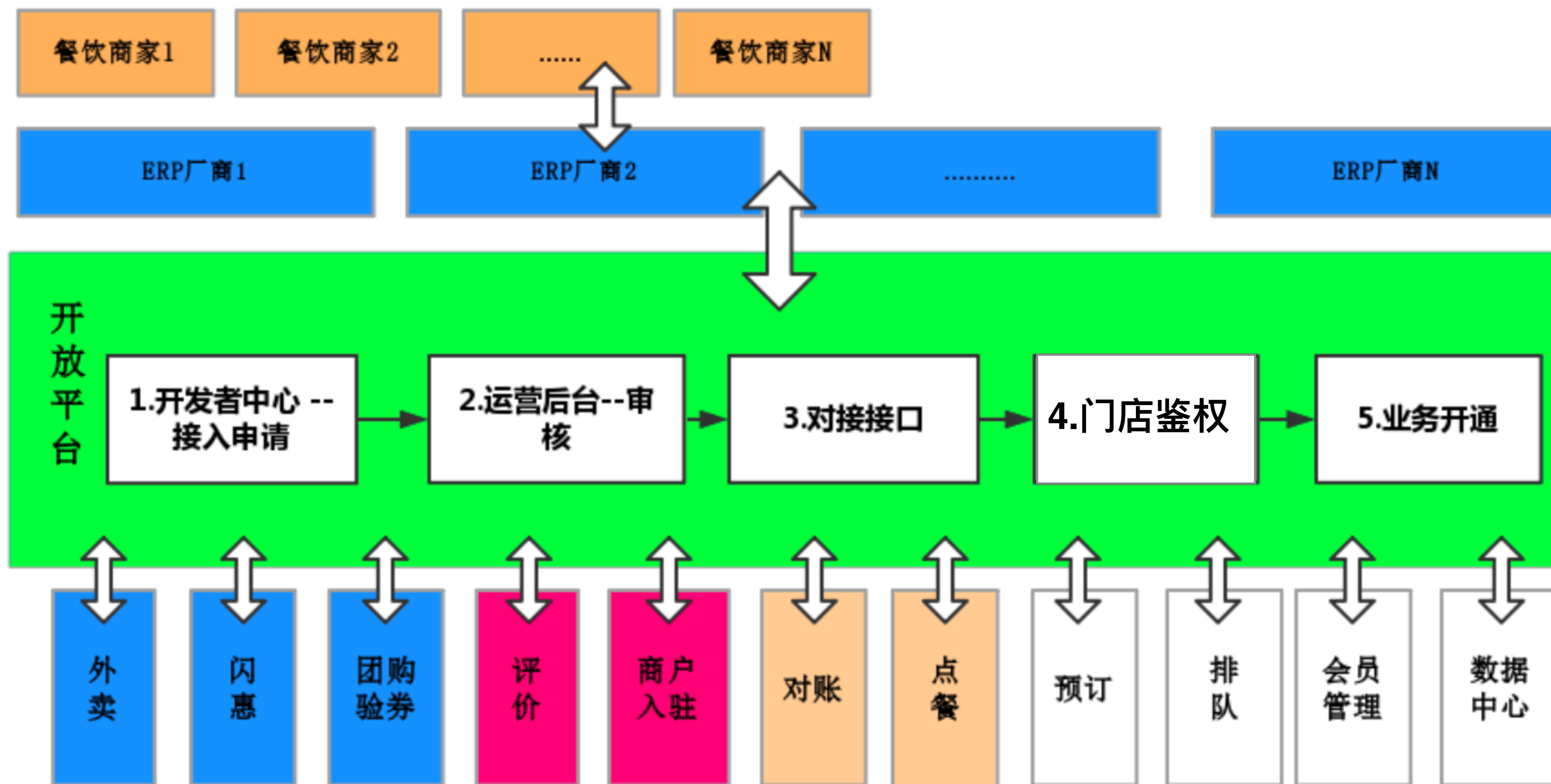
开放平台是什么

新美大餐饮开放平台是基于美大餐饮业务（包括外卖、团购、闪惠、点餐、排队、预订等），为第三方平台及餐饮软件公司，提供更加便捷的操作平台，从而提升商家经营效率，提高消费者的服务体验。

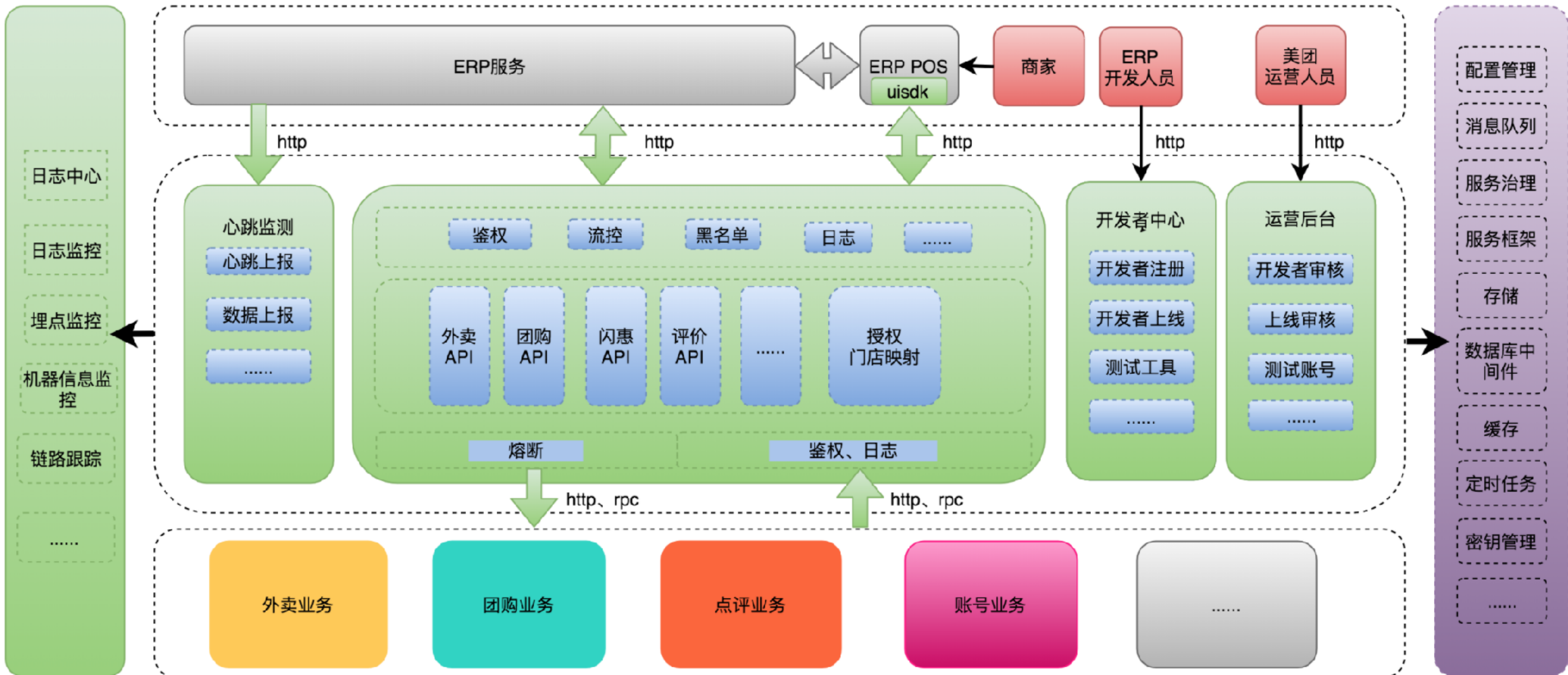
开放平台的赋能



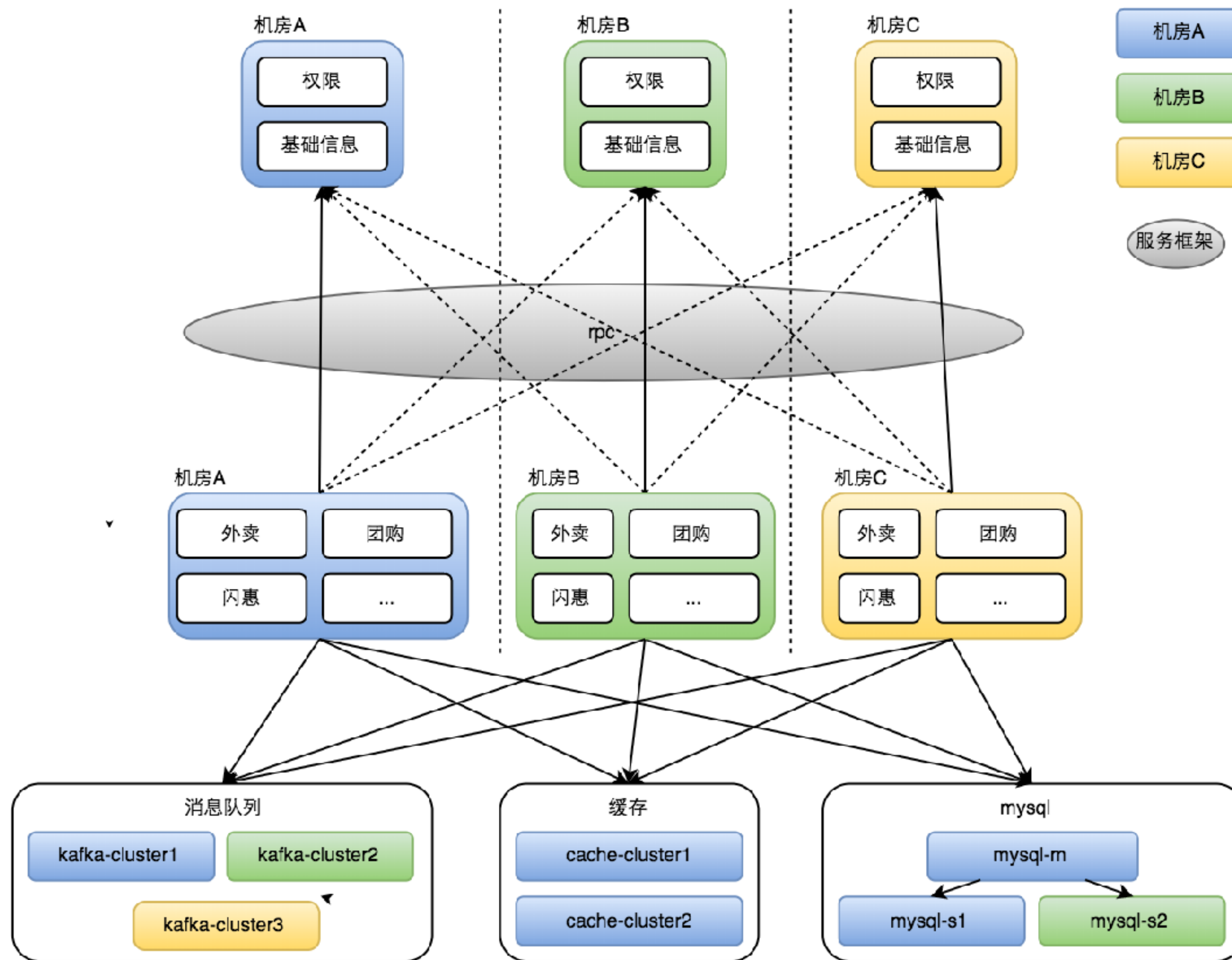
开放平台接入流程



开放平台的架构图



开放平台多机房部署



备注:

- 1、rpc路由策略: 同机房>同城>异地
- 2、42台服务器
- 3、消息队列数据量最大topic共52个分区, 每个分区2个副本

开放平台的关键指标

日期	上线厂商数	绑定门店总数	外卖门店数	团购+闪惠门店数	外卖推单数	外卖确认订单数	团购总验券数	开放平台验券数	新美大闪惠订单数
2017-12-25	357	186,099	167,516	29,438	2,285,331	1,667,738	151,393	59,658	10,994
2018-01-01	362	192,084	172,428	30,446	2,463,954	1,806,275	219,945	86,502	16,839
同比(百分比)	1.4	3.22	2.93	3.42	7.82	8.31	45.28	45	53.17

- 申请入驻开发者1200， 上线360
- 接入门店数将近20w， 周增6k
- 团购验券数9w,周增1w
- 每天亿级别的接口访问量， qps峰值2.5k， tp99耗时<250ms
- 推送订单数250w， 周增20w
- 订单日志记录数， 日增2000w， 数据量日增35g+

开放平台的挑战

- 外部依赖接口多，大约200+
- 各厂商回调接口的质量参差不齐
- 订单相关接口每天写db 2000万，每天新增35G
- 开发者和接口粒度的监控报警
- 接入的业务众多，团购，闪惠，外卖，评价，预定，会员

如何提高系统的可用性？

三.高可用之服务容错

容错—Hystrix容错组件

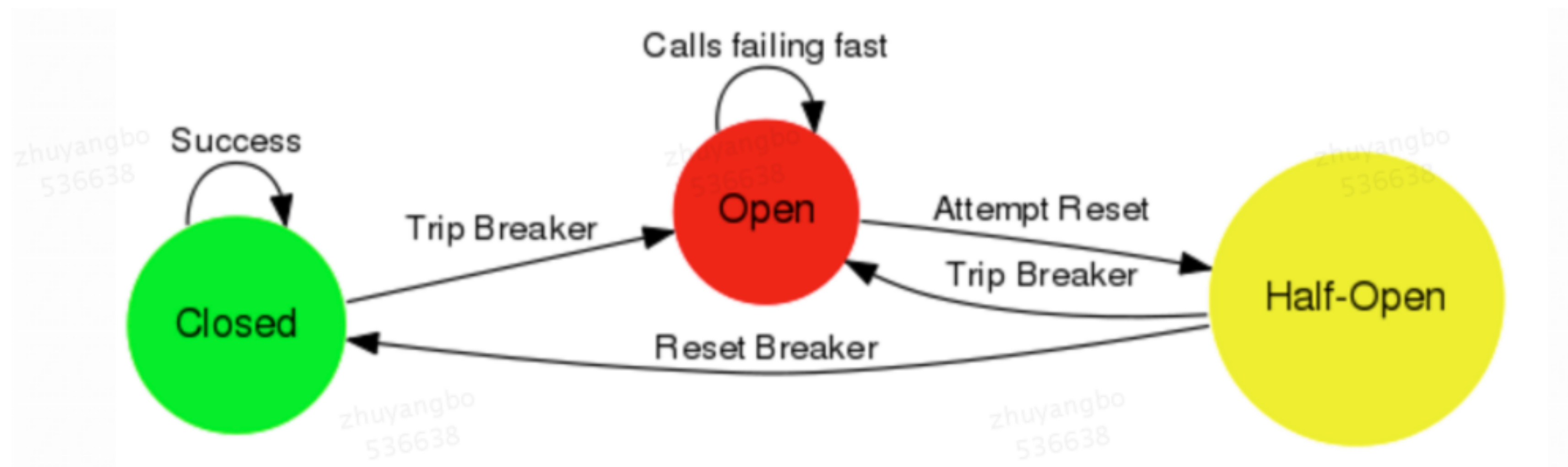
- Hystrix对应的中文名字是“豪猪”，豪猪周身体长满了刺，能保护自己不受天敌的伤害，代表了一种防御机制，Netflix团队将该框架命名为Hystrix，并使用了对应的卡通形象做作为logo。
- Hystrix是Netflix针对微服务分布式系统的熔断保护中间件；它是一个jar包，通过在不同的分布式系统中增加容错逻辑，来提高服务的可用性。



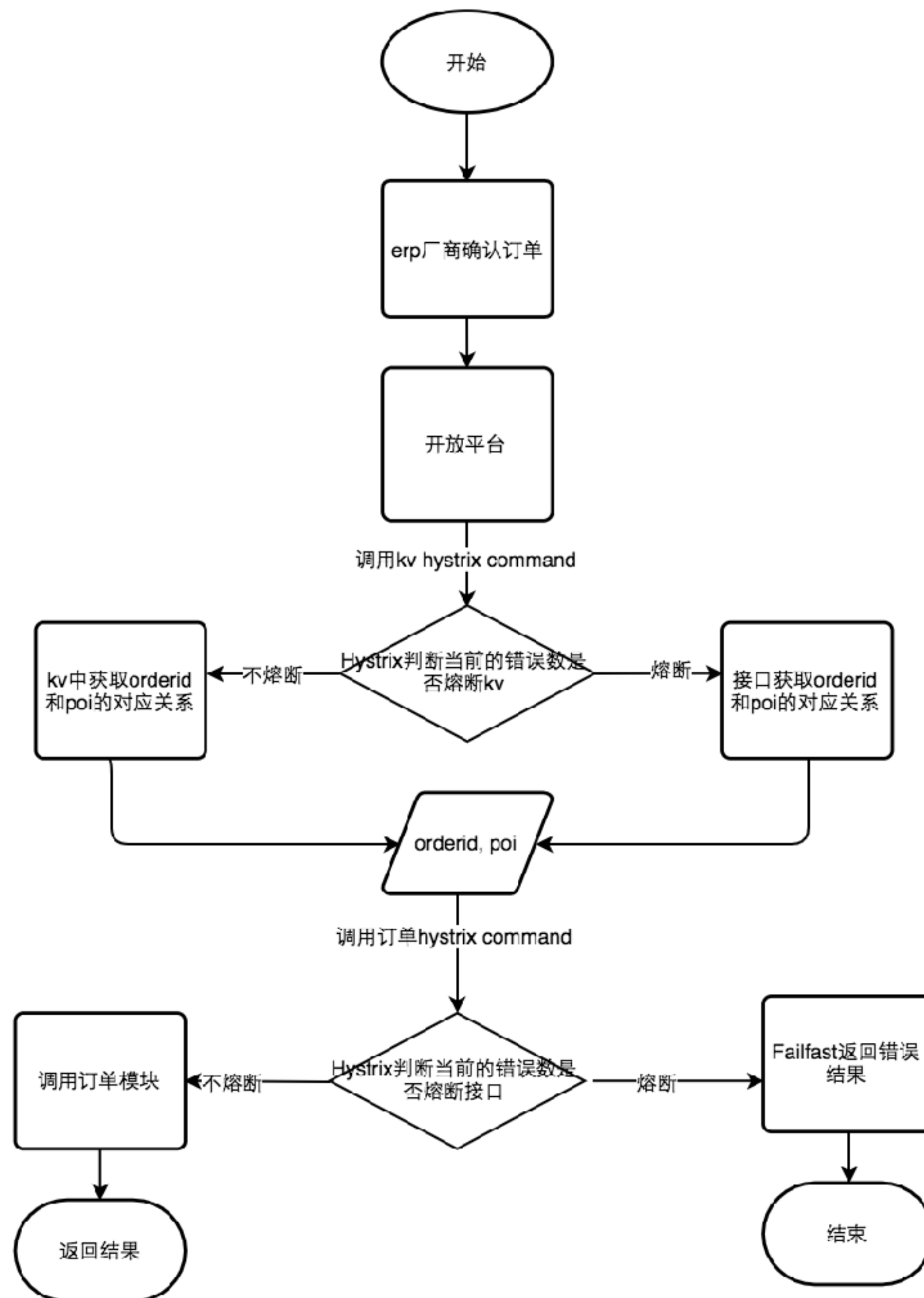
容错—Hystrix容错组件

- 电路熔断器模式(Circuit Breaker Pattern)
 - 舱壁隔离模式(Bulkhead Isolation Pattern)
 - 回退(fallback)
-
- 限流(Rate Limiting/Load Shedder)

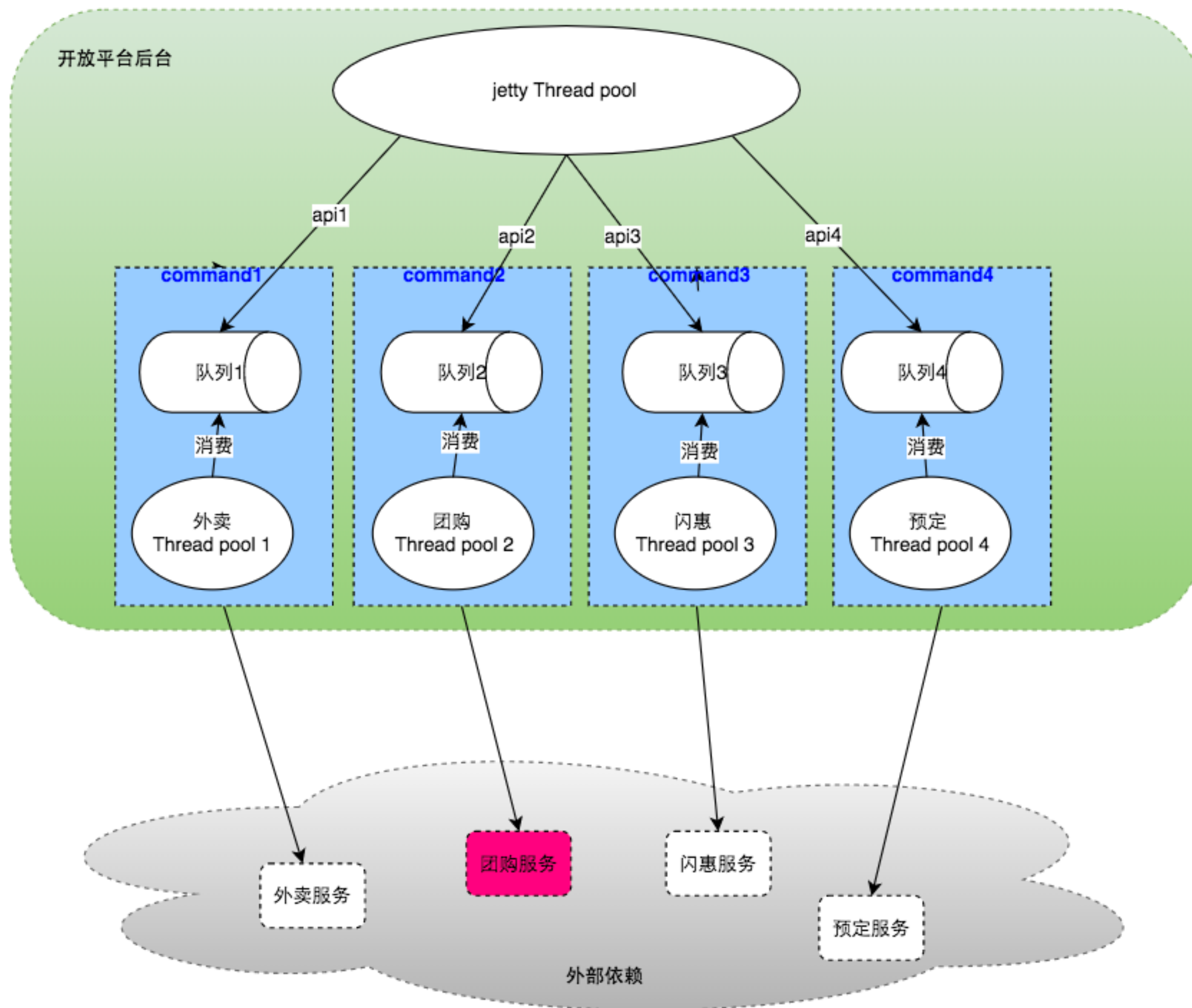
Hystrix容错组件－电路熔断器模式



应用电路熔断器模式



Hystrix容错组件－舱壁隔离模式

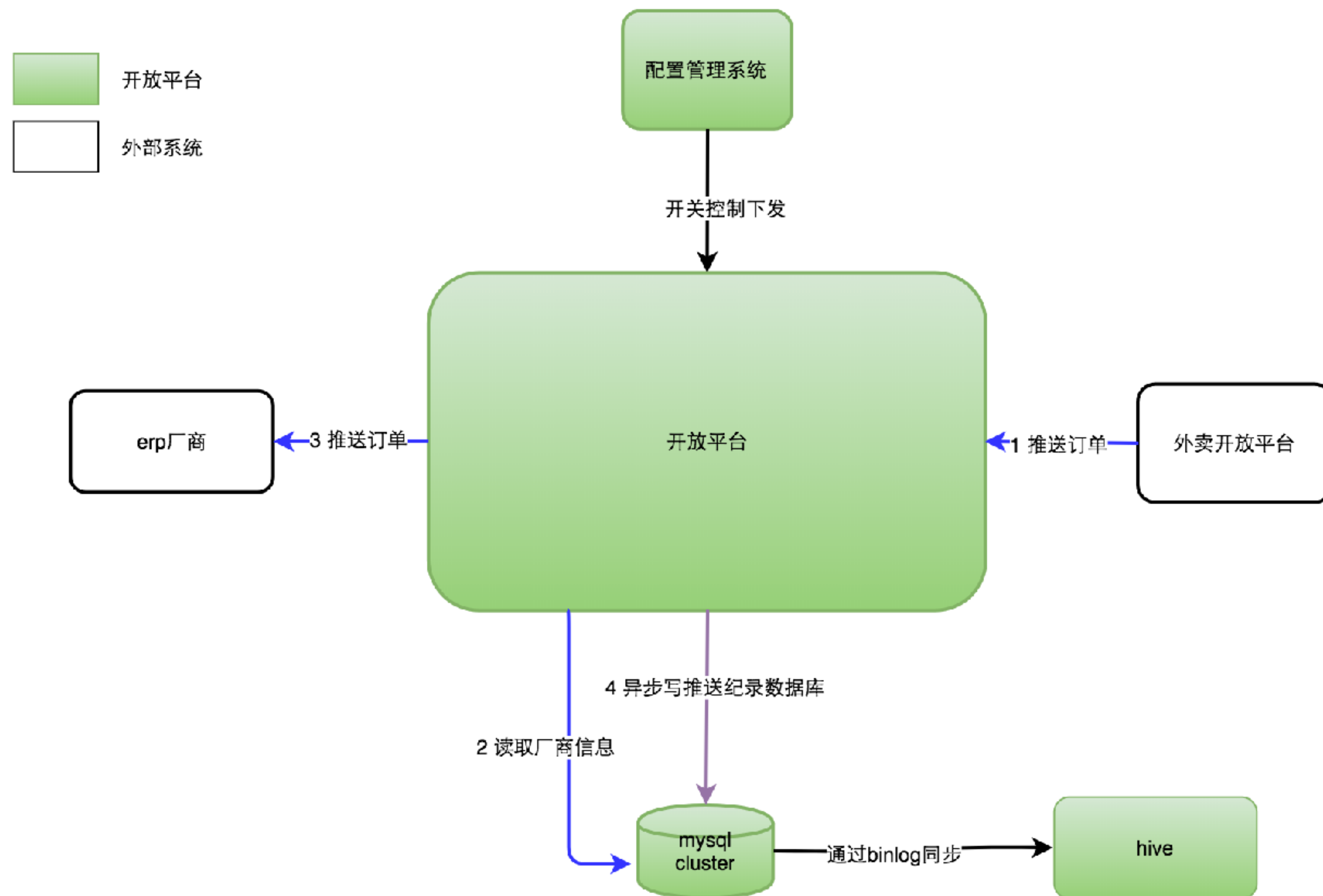


Hystrix实践经验

- 把Hystrix监控纳入公司系统
- 参数配置热加载
- 慎用限流功能
- 确保有fallback函数
- 捕获被调用端的异常，并自定义

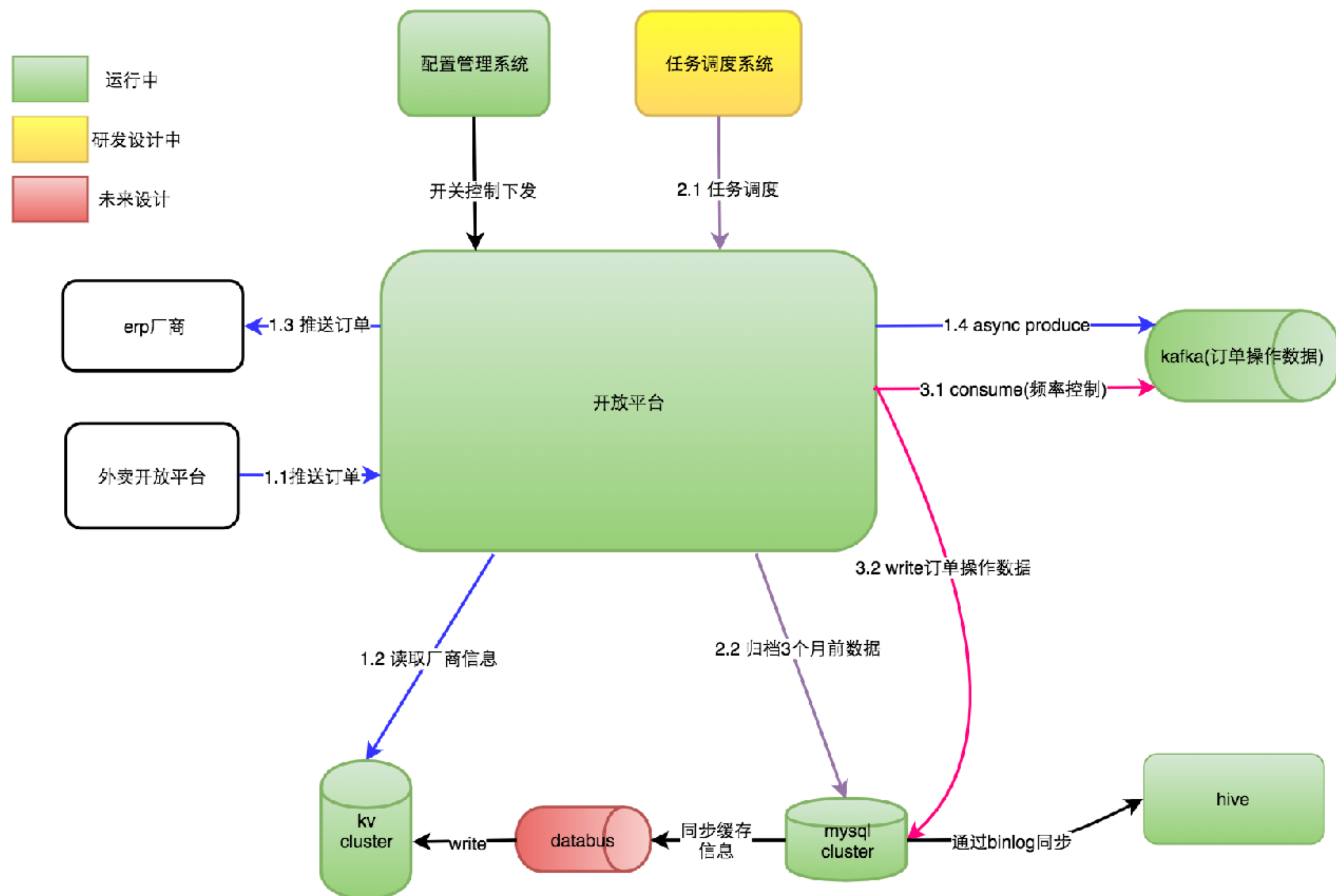
四.高可用之服务优化

外卖推单优化-优化前

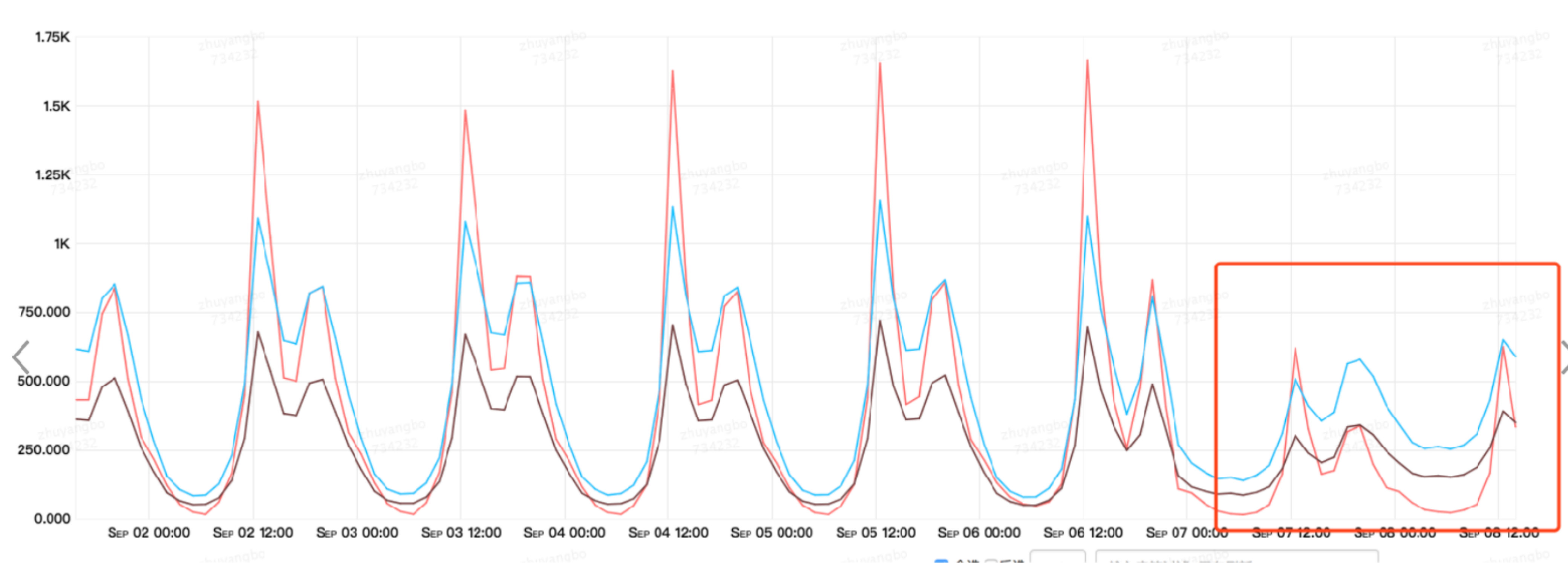


- 数据库写入数据量大
- 高峰数据库访问延迟高
- 系统的耦合高

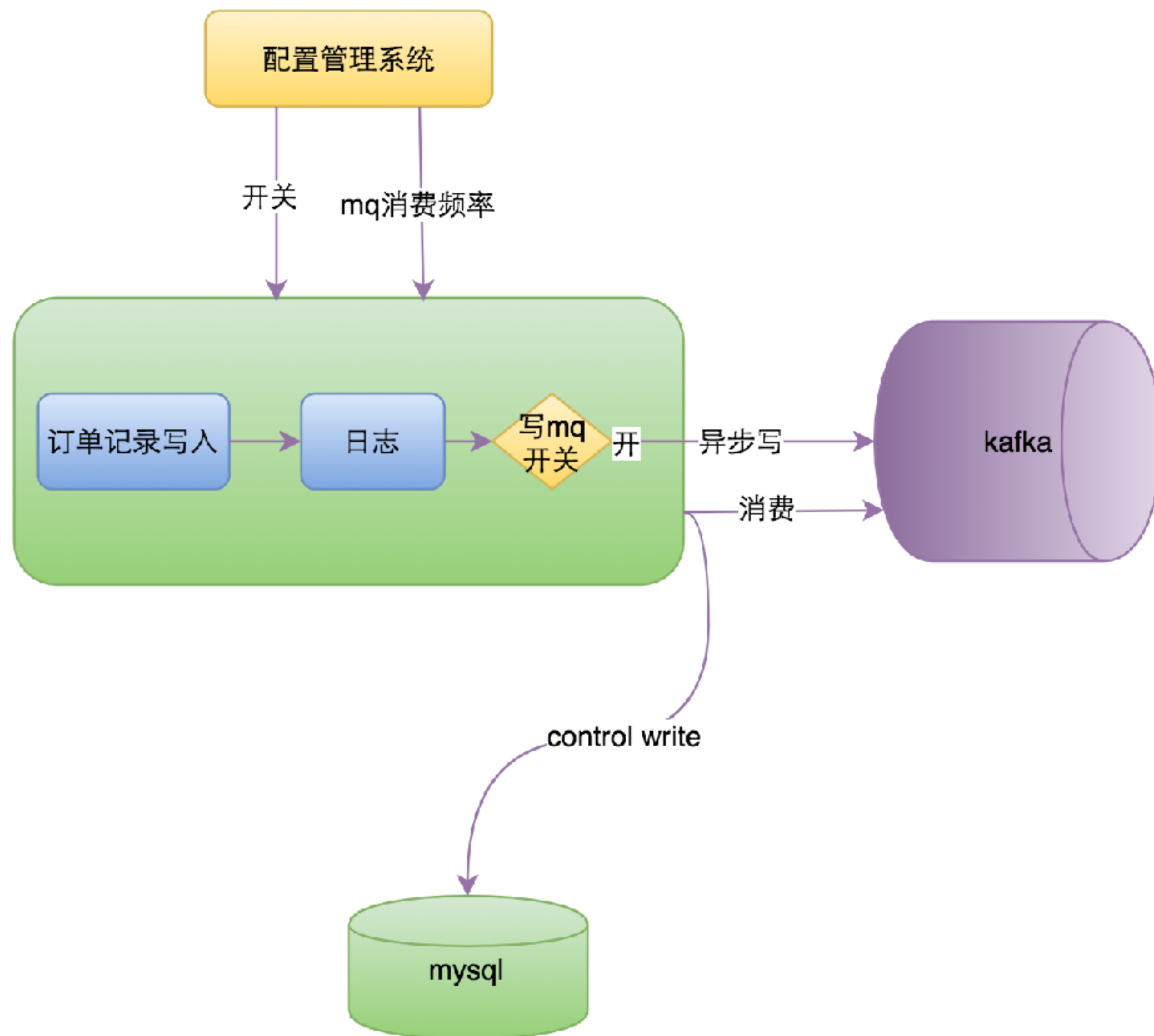
外卖推单优化-优化后



关键路径优化-读mysql下降明显

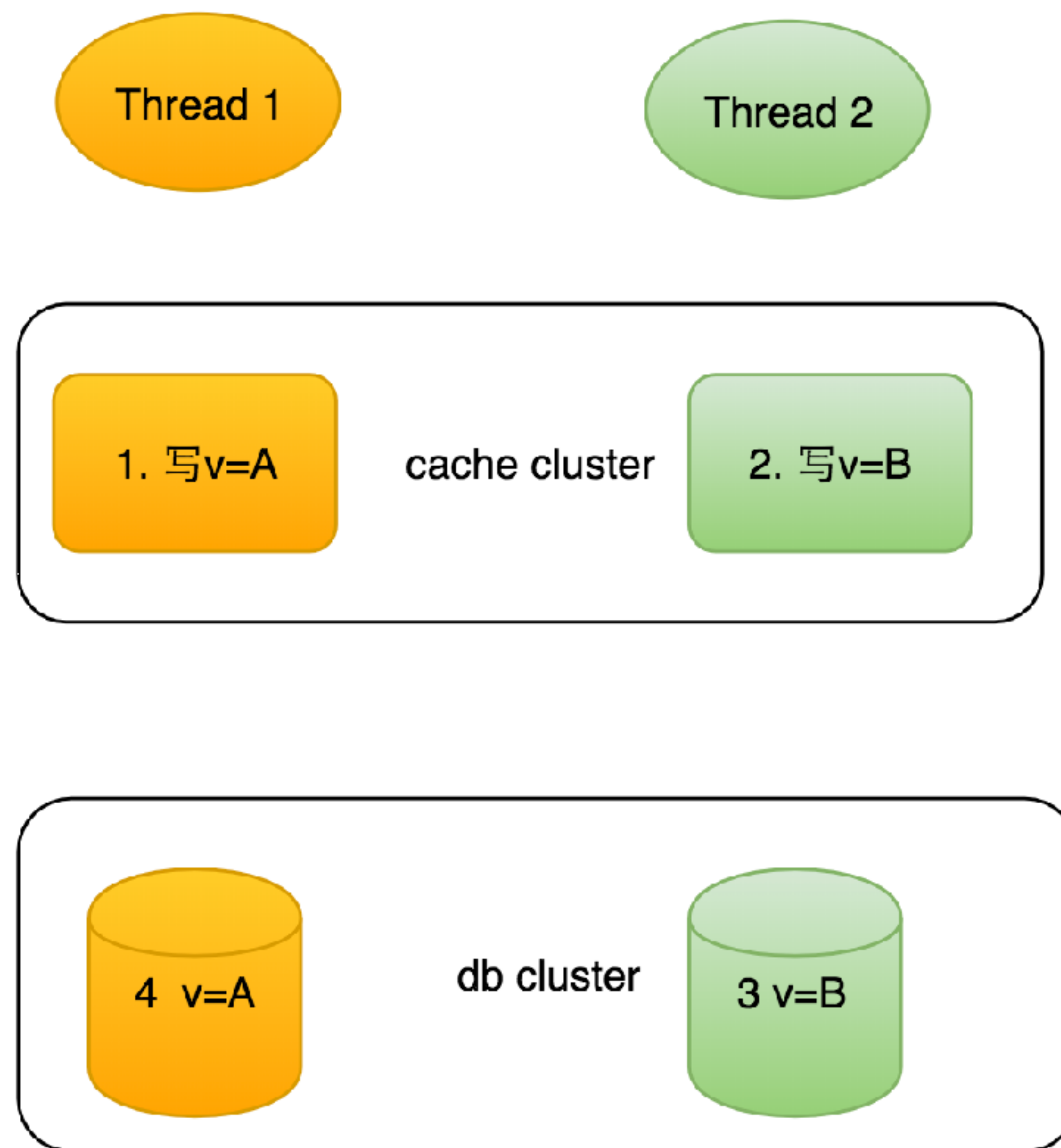


关键路径降级



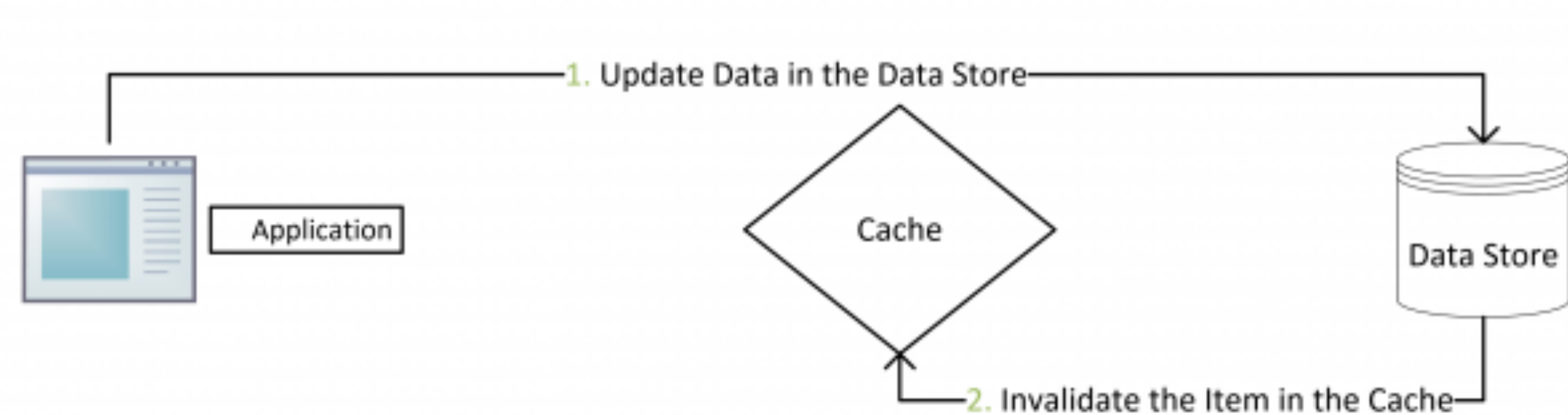
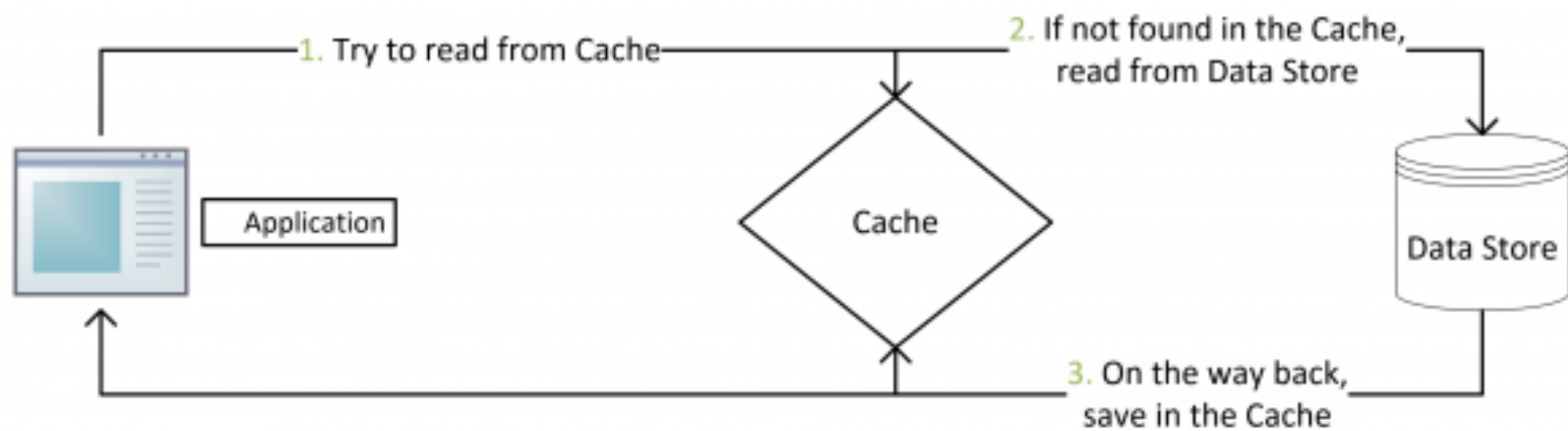
- kafka消费速度过快，配置管理系统限制消费速率
- kafka服务不可用，可实时开关切换mysql
- mysql负载很高时，写程序日志，后期脚本处理倒入数据库
- kv多集群互为备份切换，切换时容忍少量不一致

缓存的一致性的问题



缓存的一致性 – Cache Aside Pattern

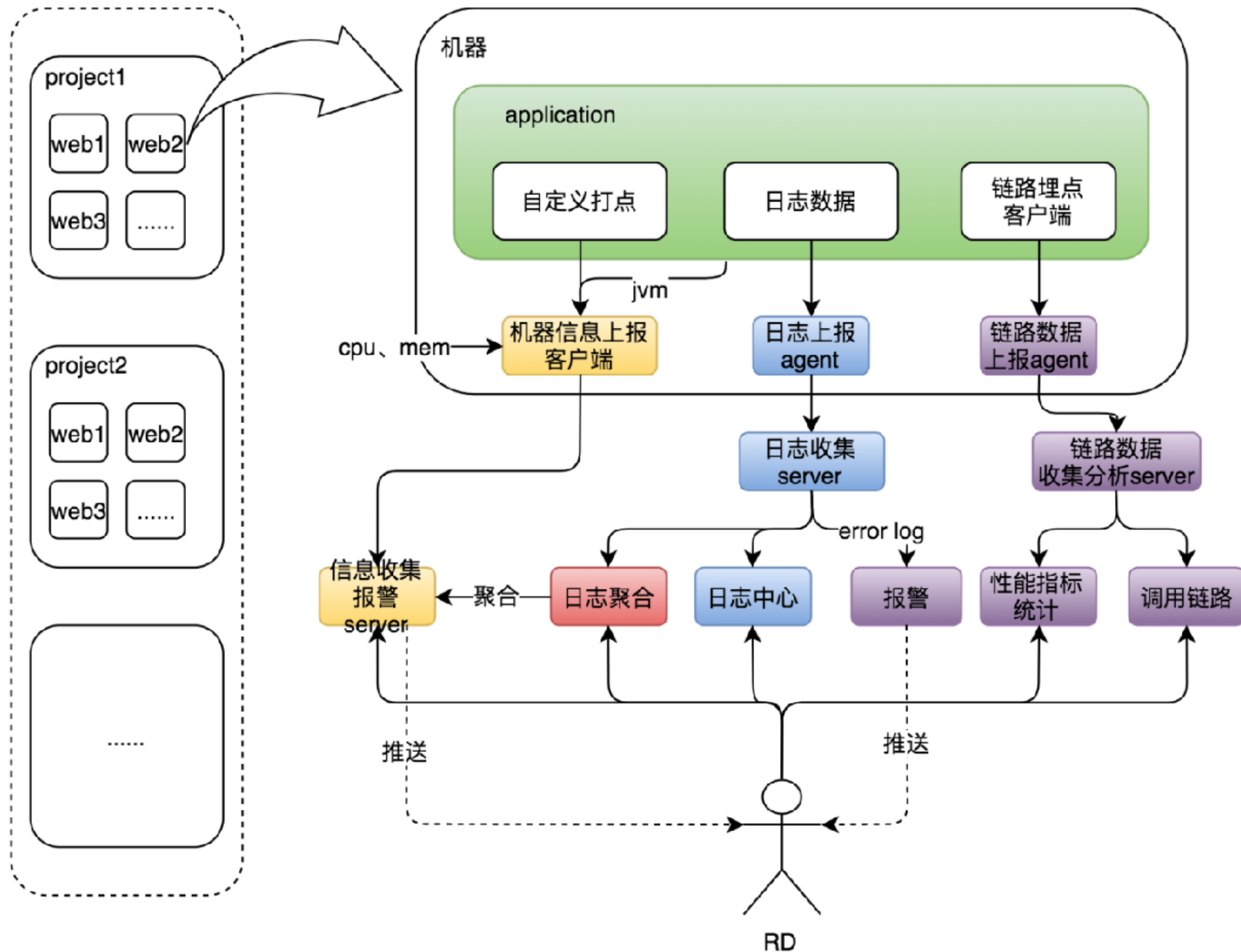
- 失效：应用程序先从cache取数据，没有得到，则从数据库中取数据，成功后，放到缓存中。
- 命中：应用程序从cache中取数据，取到后返回。
- 更新：先把数据存到数据库中，成功后，再让缓存失效。
- 读多写少的缓存设置一定的超时时间。



五.高可用之监控报警

监控报警的故事

开放平台监控架构图

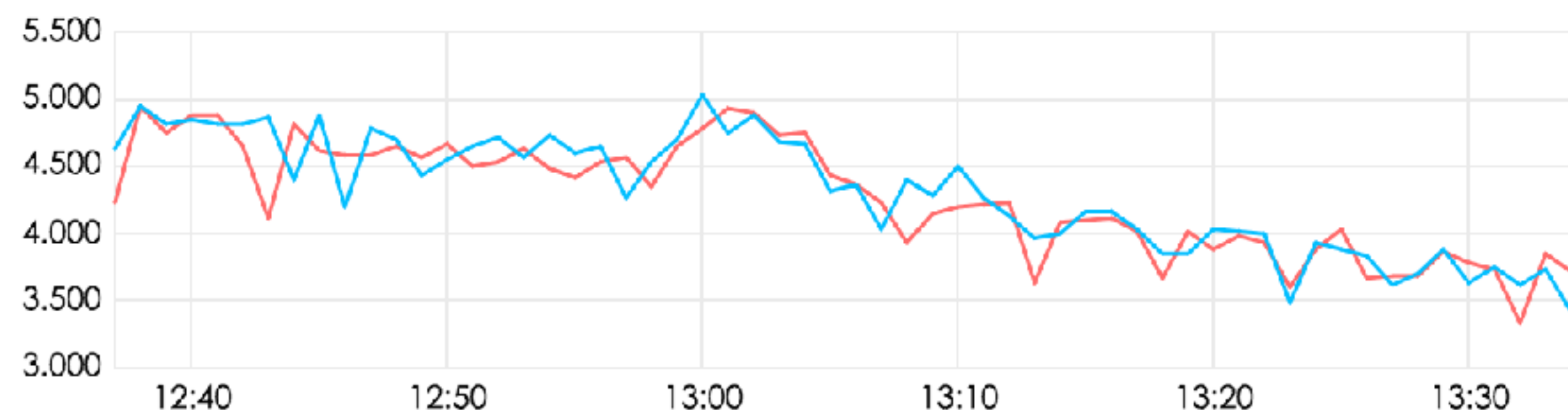


机器负载的监控

网卡入流量



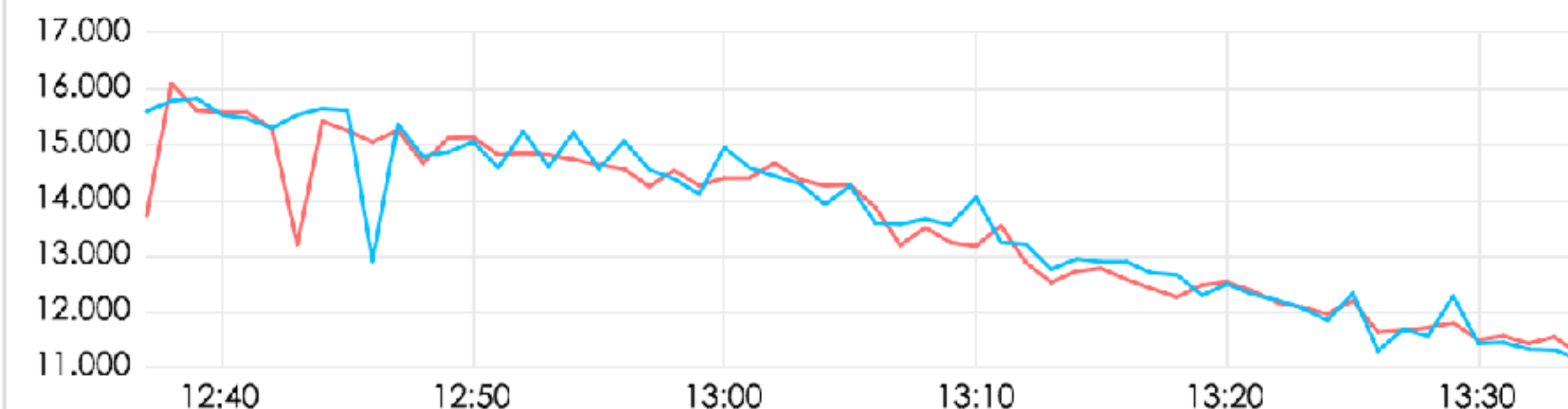
net.if.in.Mbps/iface=eth0,nettype=10g



网卡出流量



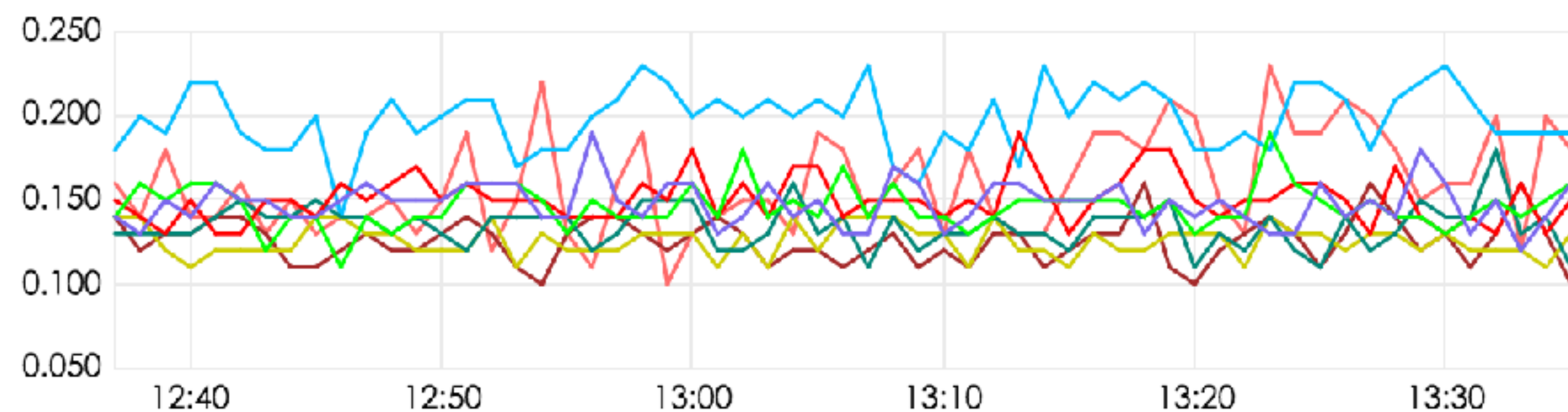
net.if.out.Mbps/iface=eth0,nettype=10g



ping响应时间



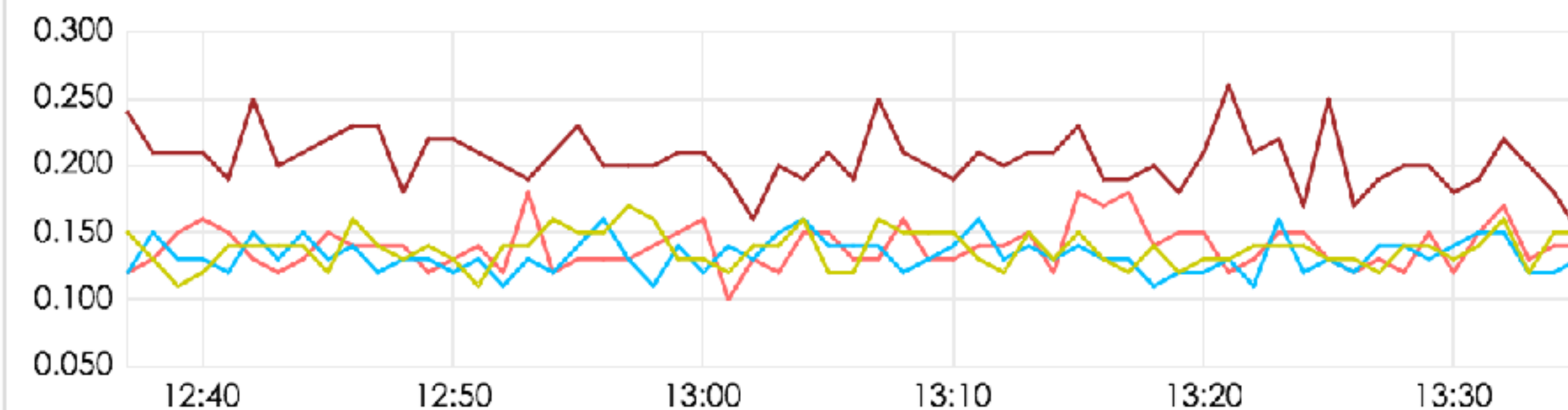
icmp.ping.msec/from=dx



ping响应时间



icmp.ping.msec/from=yf

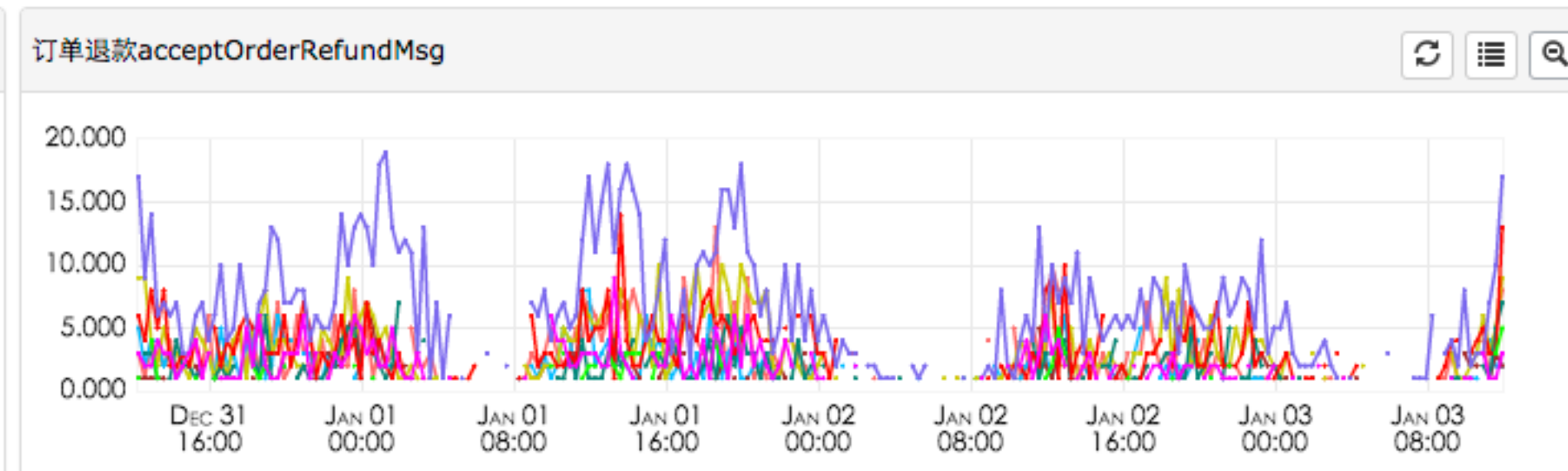
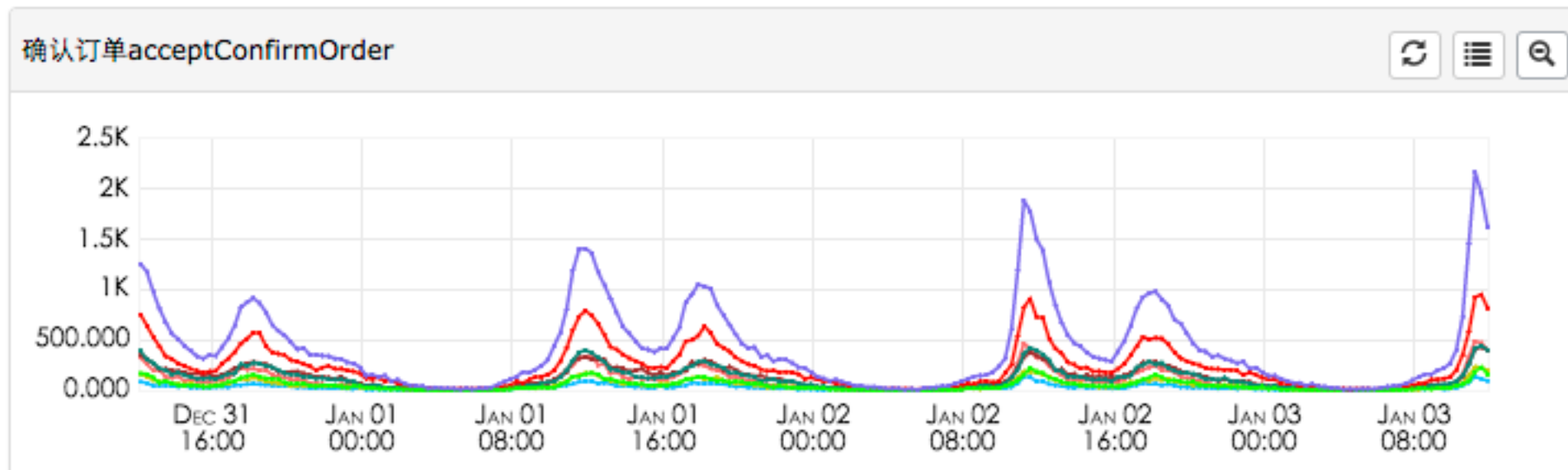
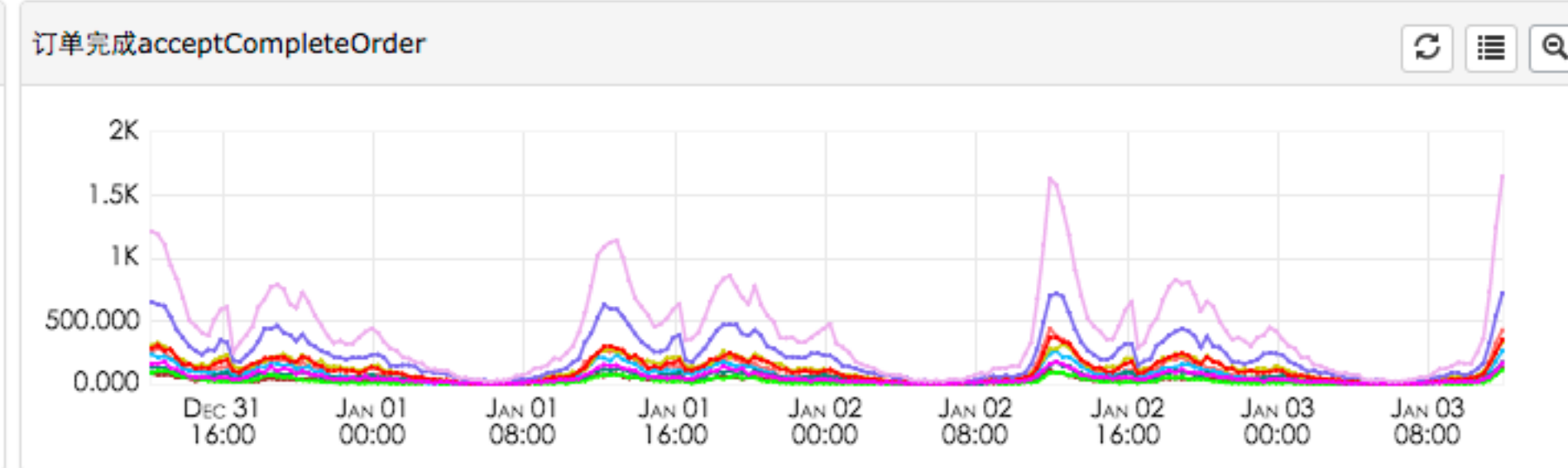
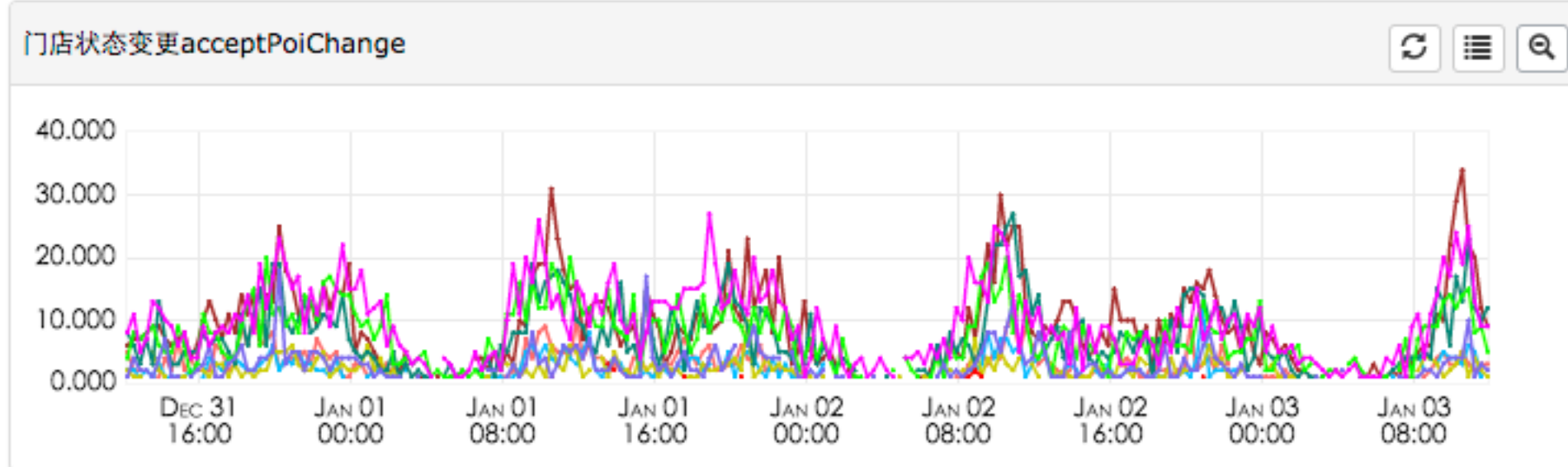
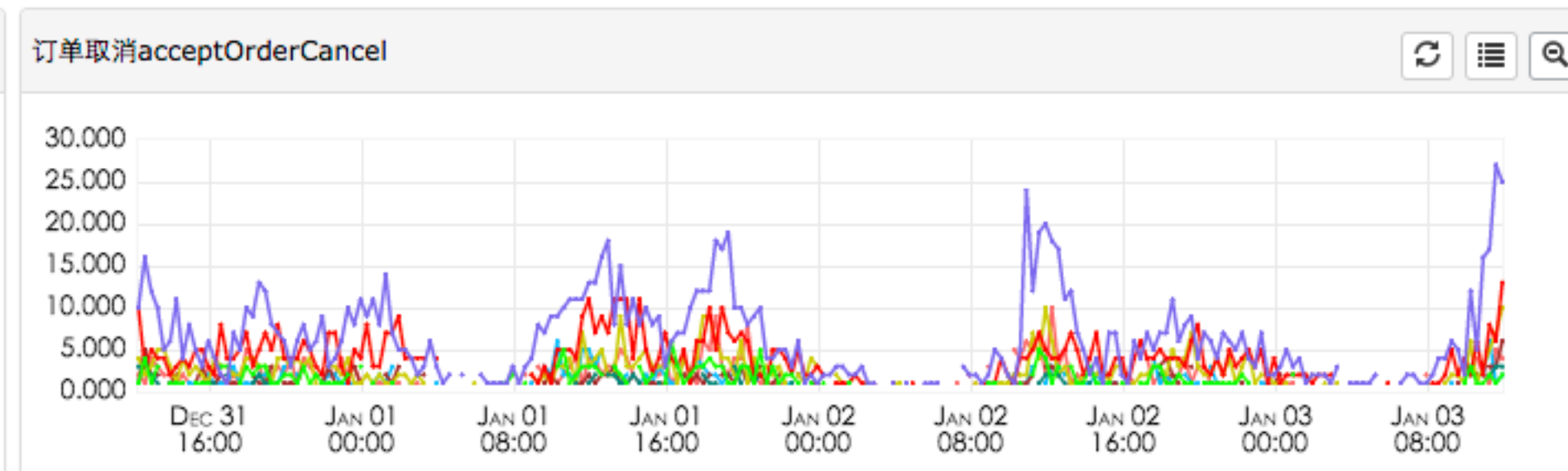
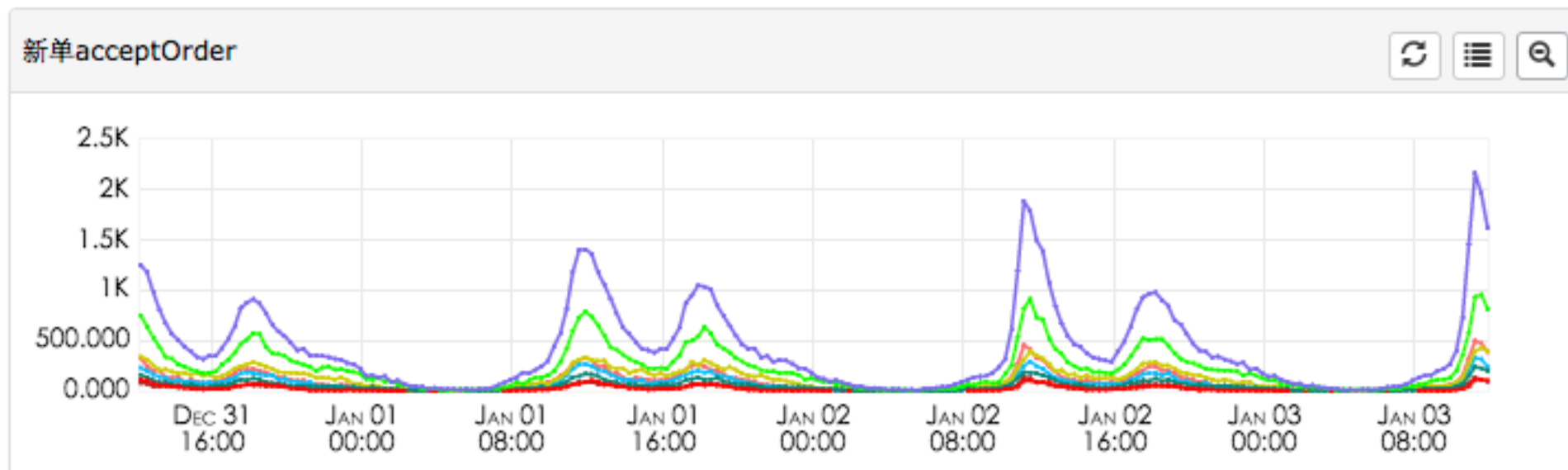


TOP 10 厂商监控

Mt-Falcon-Screen

[Dashboard](#) [报警红盘](#) [使用手册](#) [常用监控项](#) [索引自维护](#) [链接](#)

餐饮生态平台 / 开放平台TOP10厂商每分钟推单数 / 布局 / 时间段 / 采样方法 / 刷新频率 / [编辑](#) / [批量编辑](#) / [删除](#) / [克隆](#) / [❤](#) / [+graph](#) / [+显示legend](#)



问题厂商监控

美团点评
技术团队

IT大咖说
知识共享平台

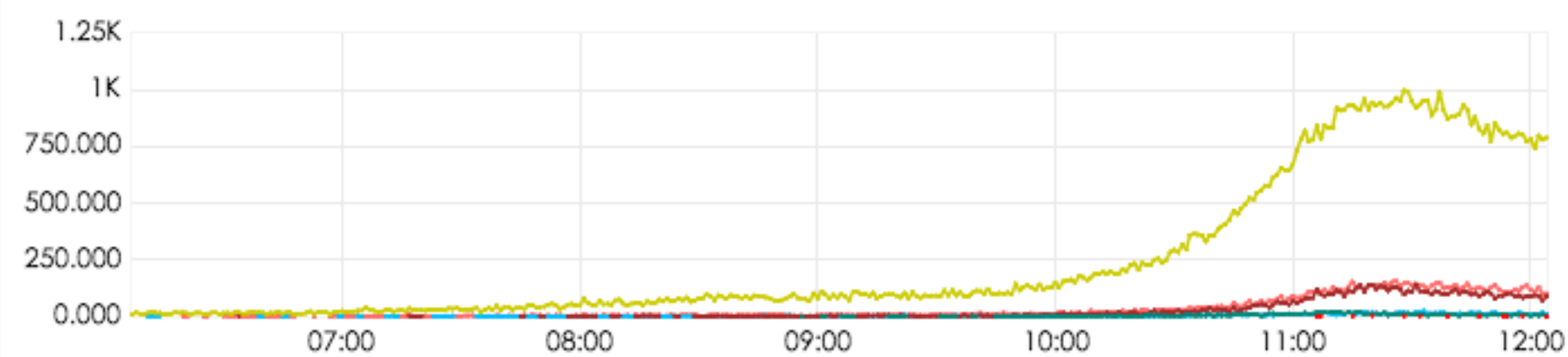
Mt-Falcon-Screen

[Dashboard](#) [报警红盘](#) [使用手册](#) [常用监控项](#) [索引自维护](#) [链接](#)

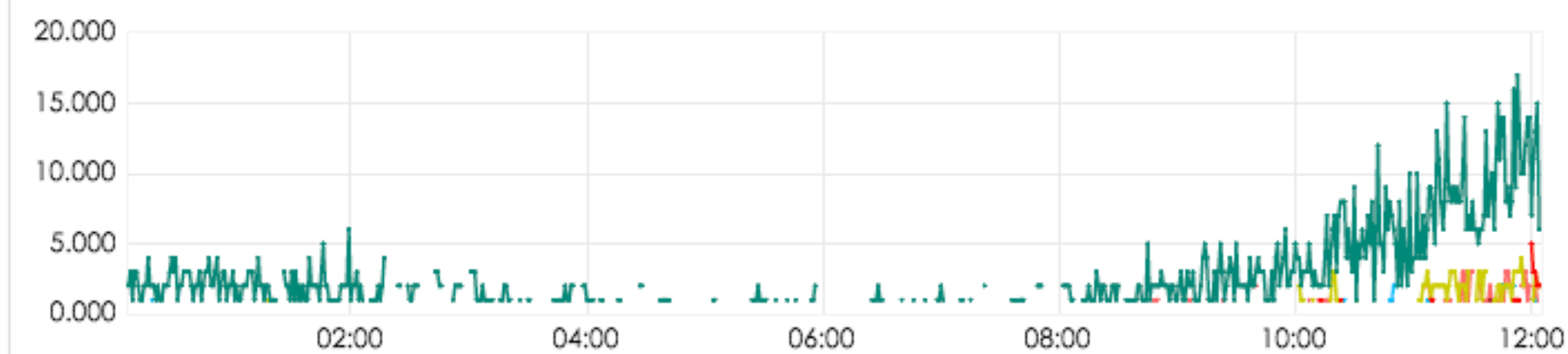
餐饮生态平台 / 开放平台重点监控厂商 / 布局 / 时间段 / 采样方法 / 刷新频率 / 编辑 / 批量编辑 / 删除 / 克隆 / 显示图例 / 显示图例

输入host

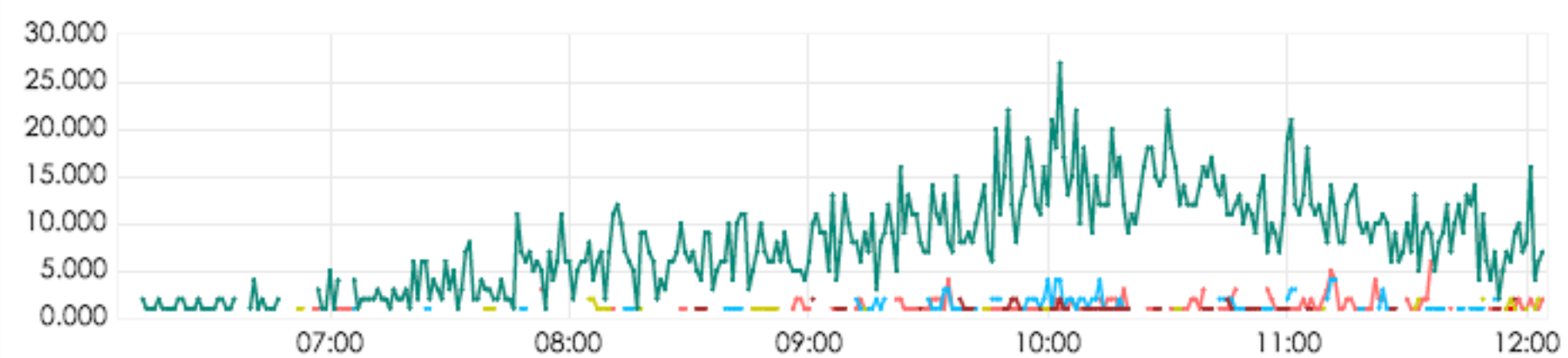
新单acceptOrder



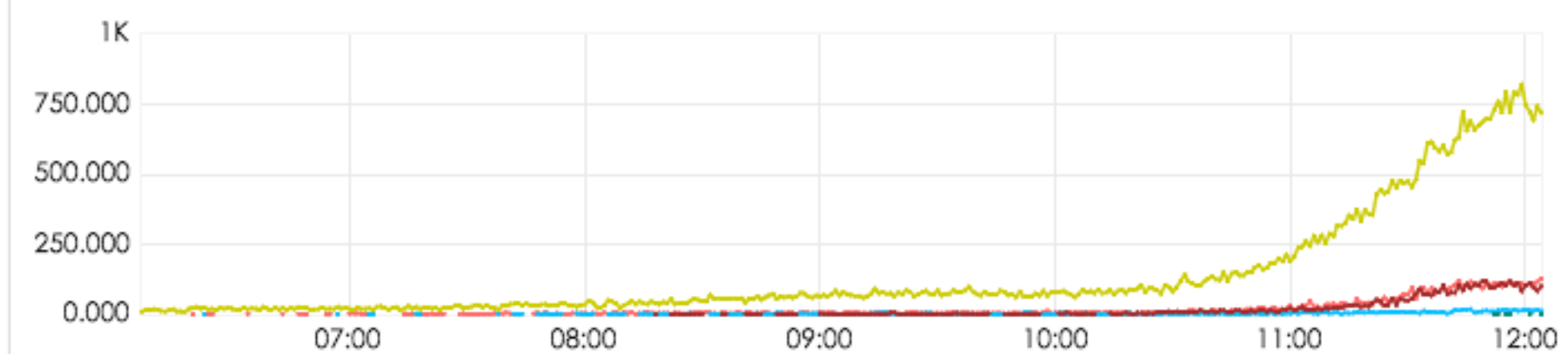
订单取消acceptOrderCancel



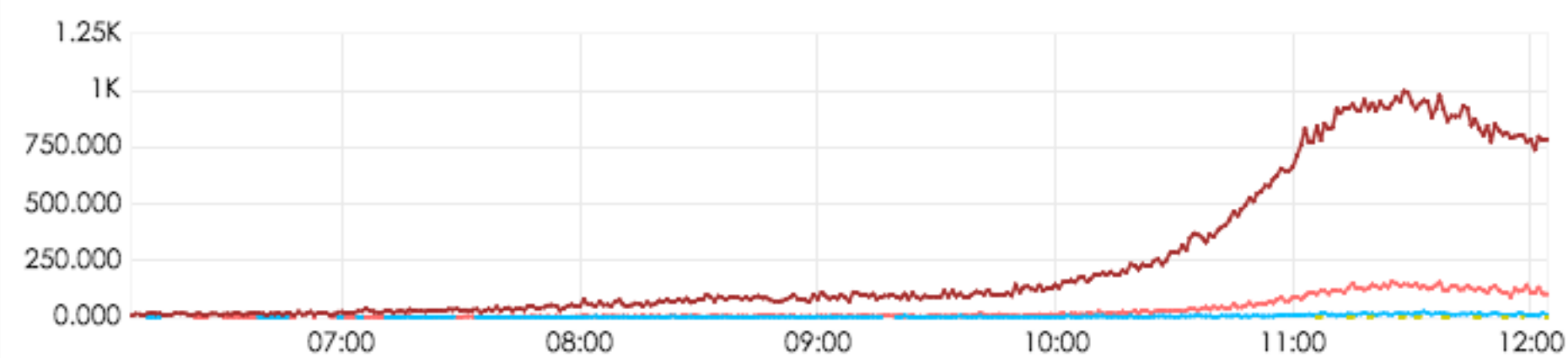
门店状态变更acceptPoiChange



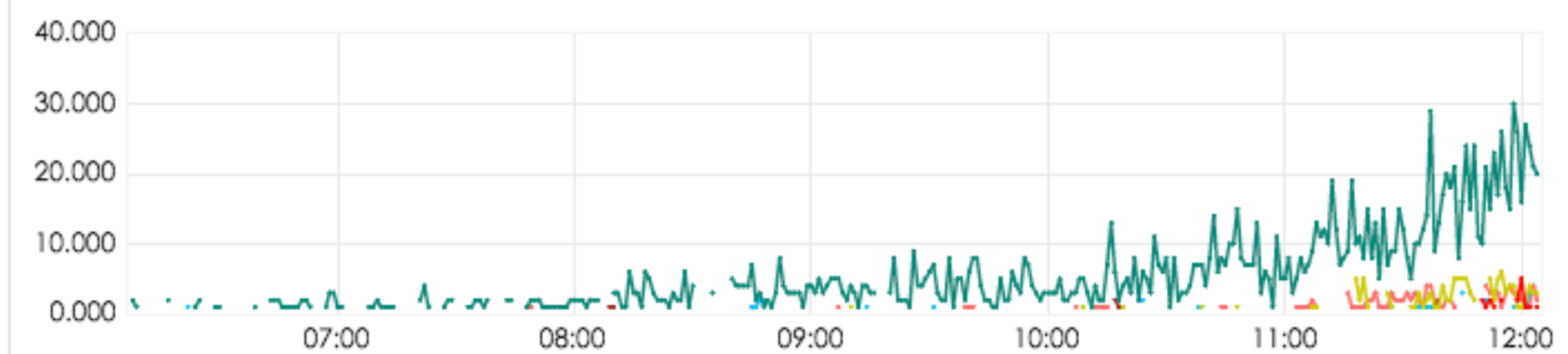
订单完成acceptCompleteOrder



确认订单acceptConfirmOrder



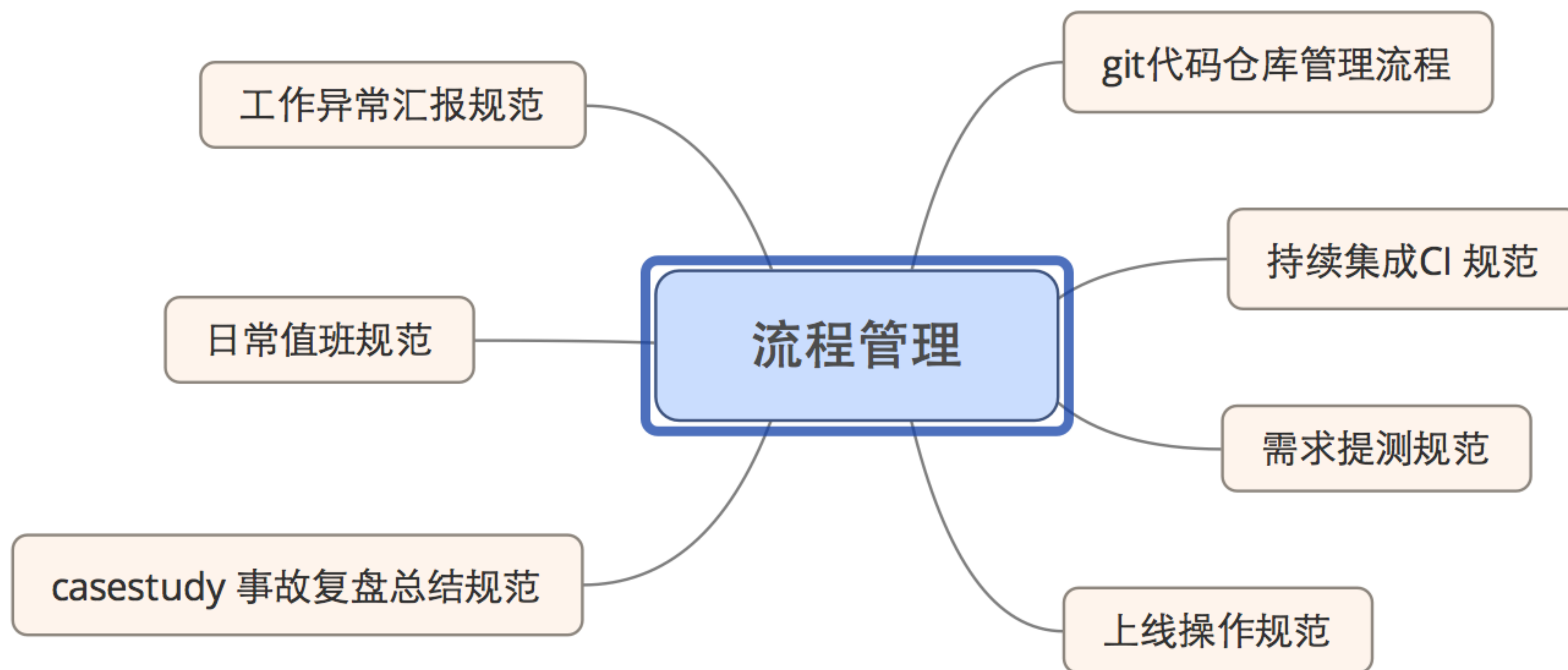
订单退款acceptOrderRefundMsg



配送状态acceptShippingStatus

结算信息acceptTradeDetail

六.建立适当的流程规范



总结：高可用架构的设计原则

- 尽可能自动化
- 考虑程序的异步化
- 为故障进行架构的设计
- 监控系统所有的资源
- 提升程序的性能
- 建立简单易执行的流程规范

Q&A

Thanks

Eat better, Live better.