四年磨一剑：蚂蚁集团注册中心 SOFARegistry 的开发实践之路

向旭 / 李旭东

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

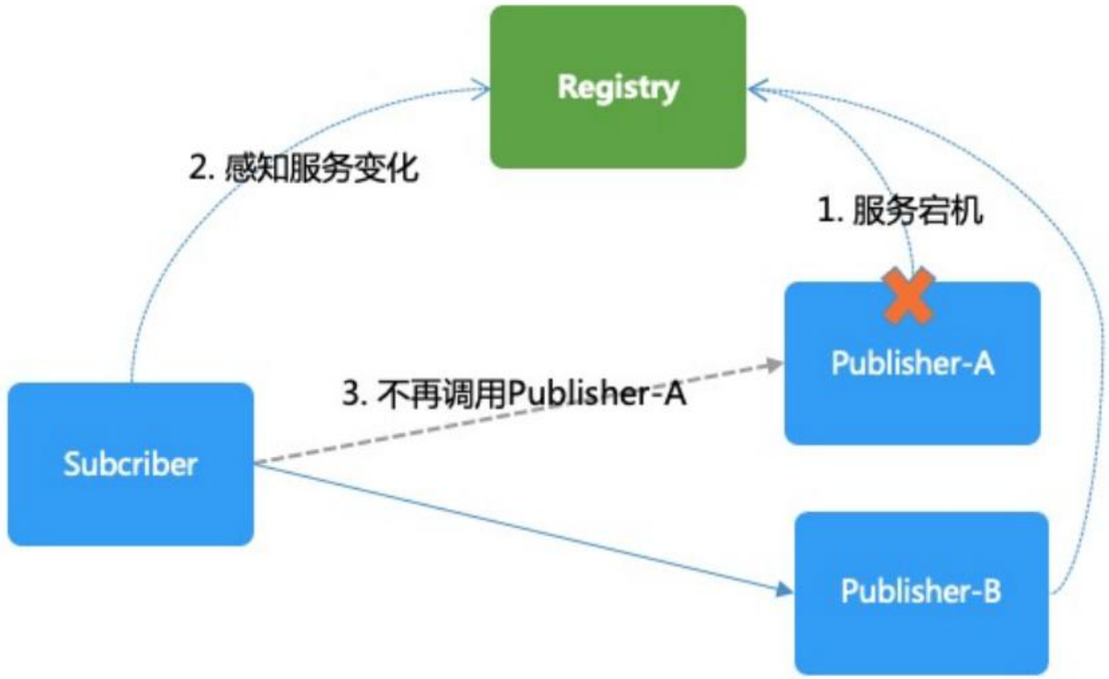
扫码查看大会详情>>

分享内容

- 蚂蚁集团注册中心的10年
- 新挑战：SOFARegistry
  - 架构：超大规模
  - 质量：高效迭代
  - 运维：自动化
- 开源与共赢

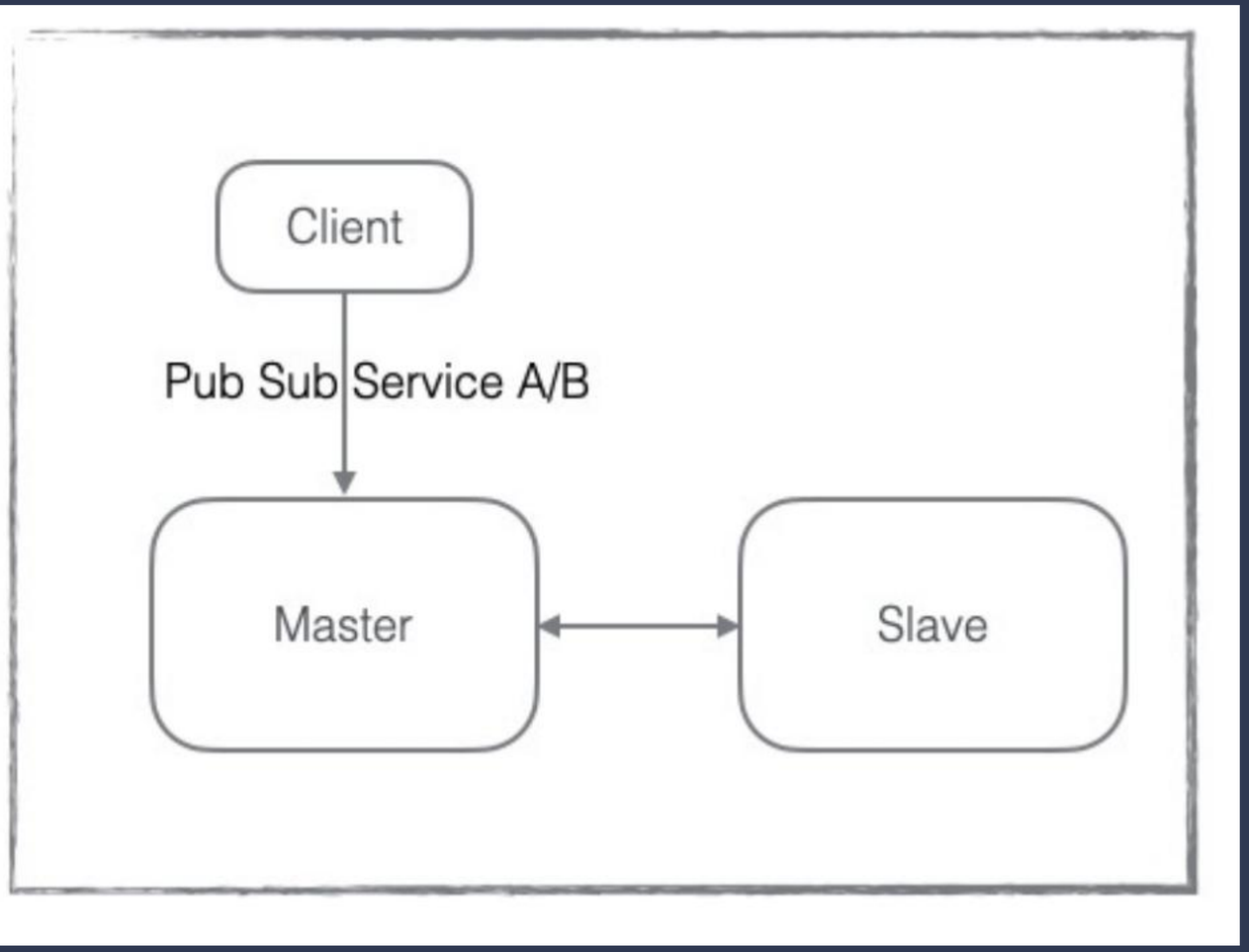
服务发现的核心能力

应用场景： RPC 场景的服务寻址



动态感知服务地址的变化

演进:V1 引进淘宝的 configserver

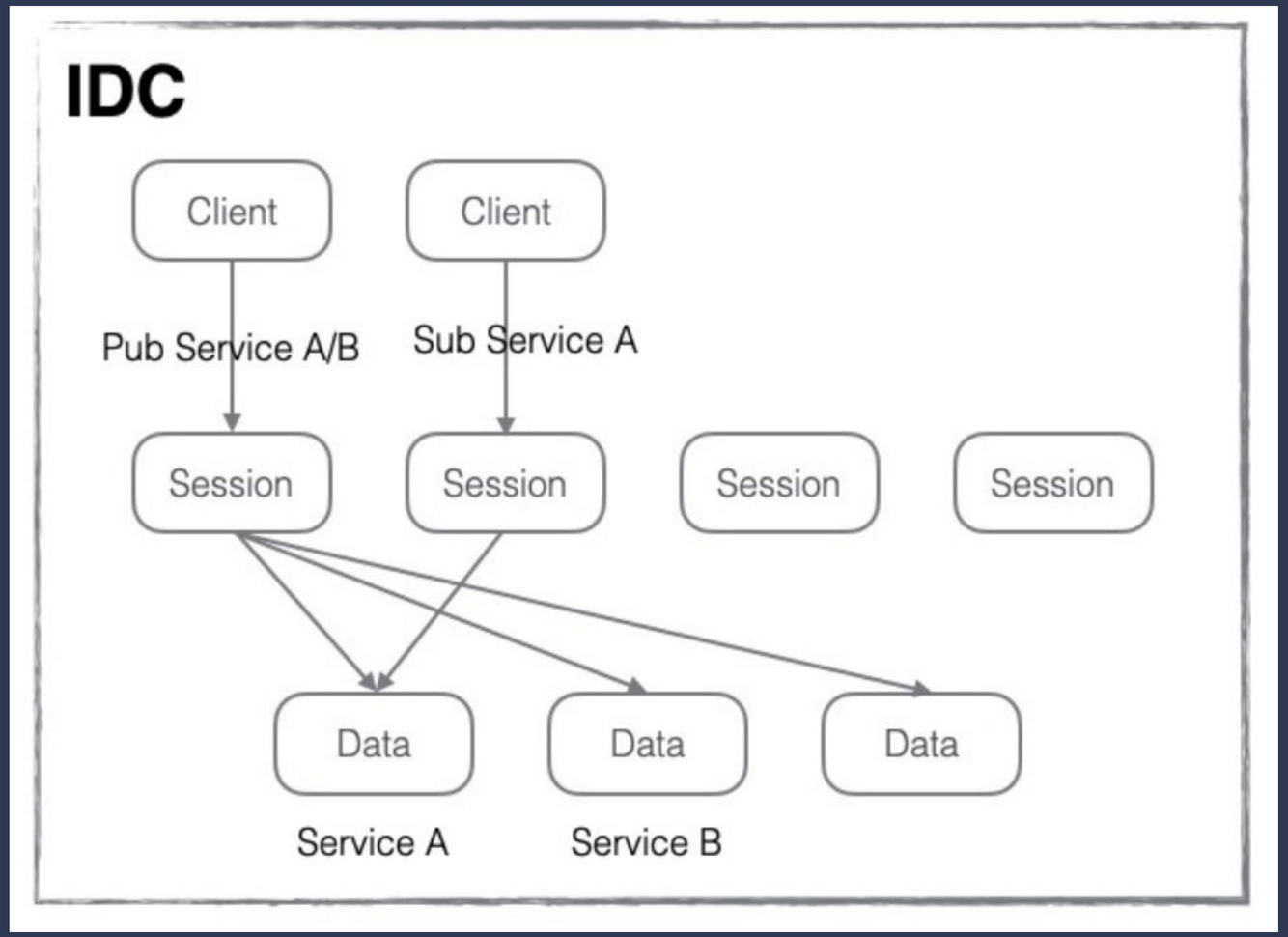


架构:单节点,master/slave 备份

面临问题

- 容量瓶颈
- 容灾风险

演进:V2 横向扩展



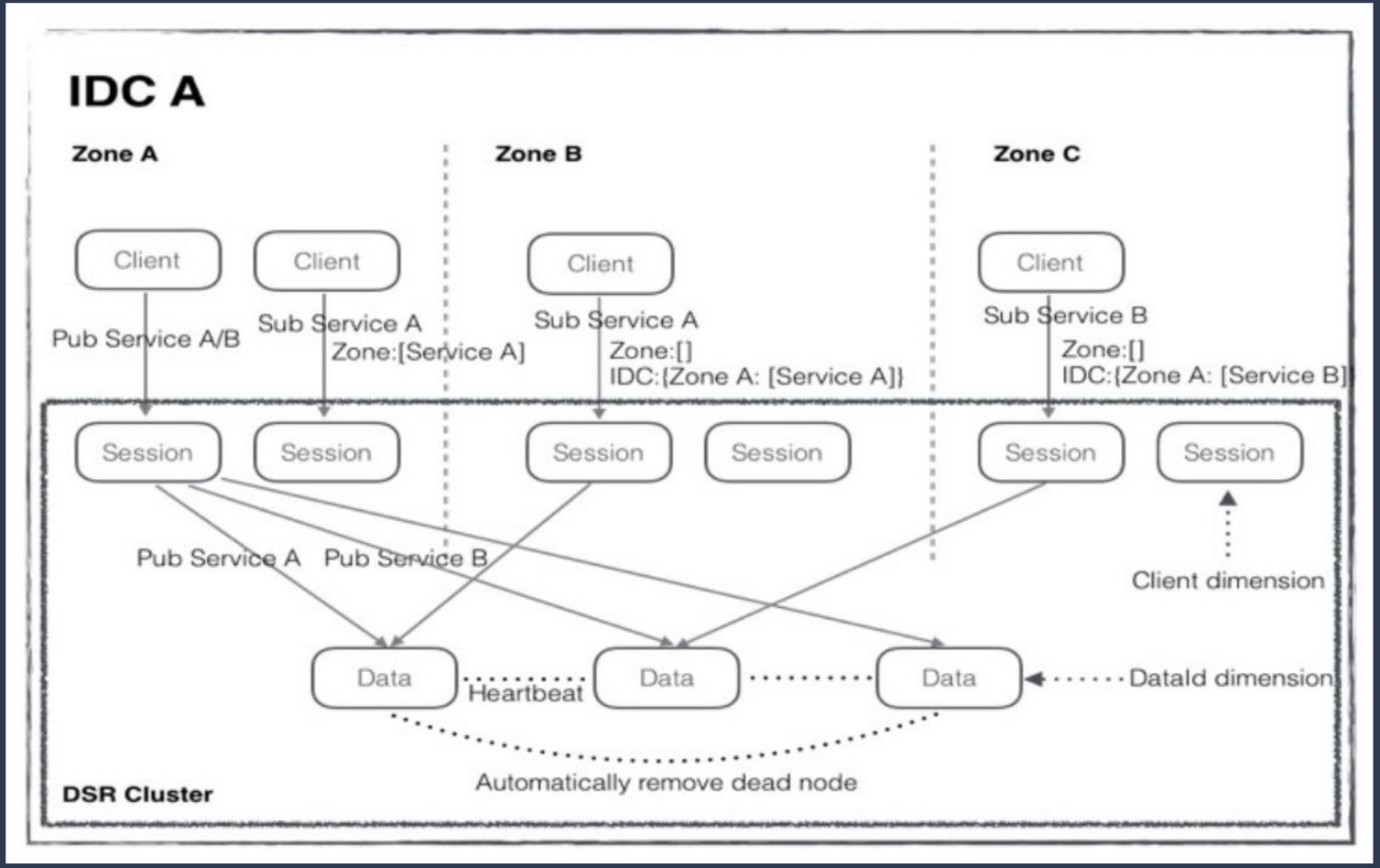
架构:拆分 session/data,水平扩展

- session:处理连接
- data:数据分片存储

面临问题

- 运维成本:serverlist 维护

演进:V3/V4 LDC 支持和容灾

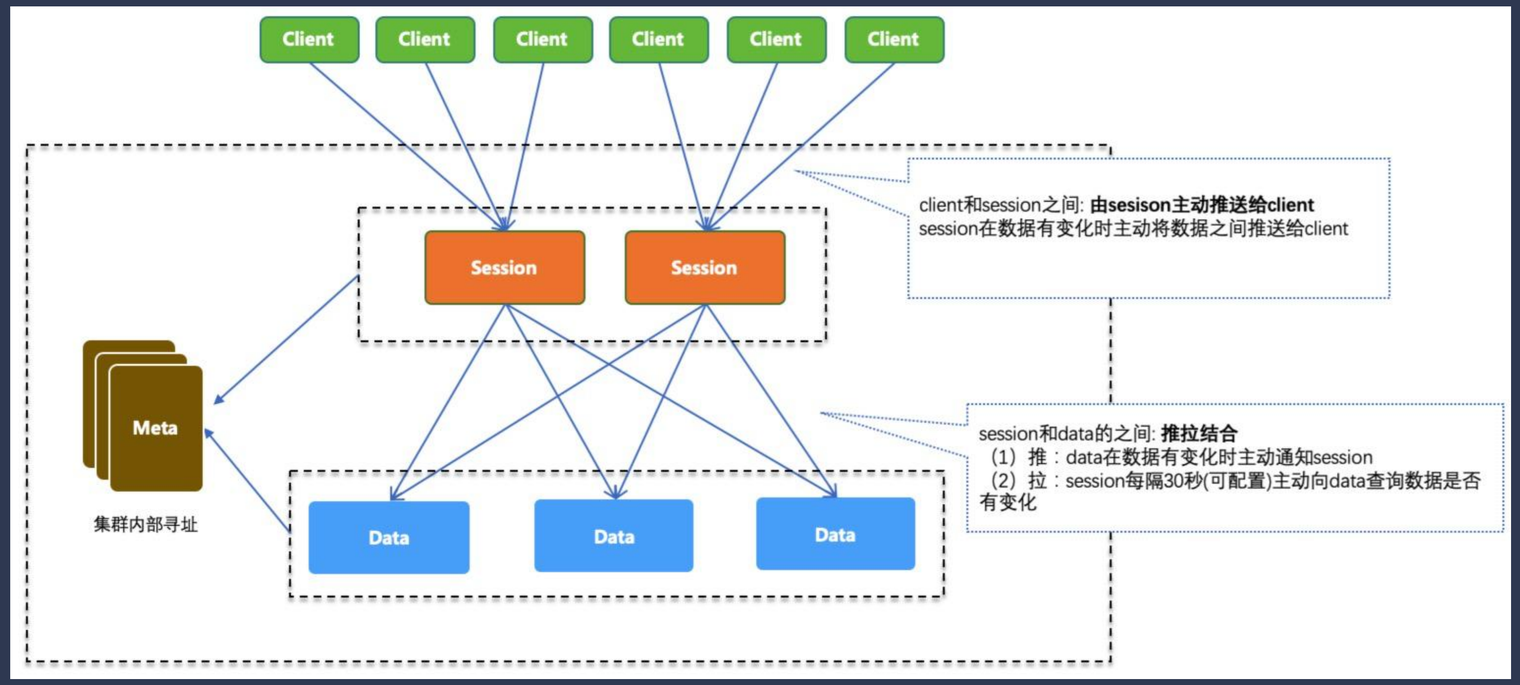


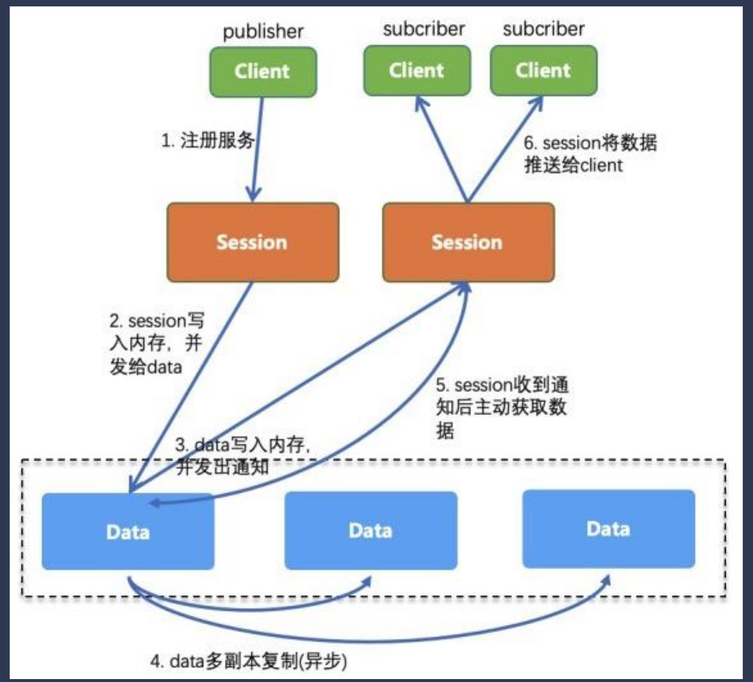
架构:单元化支持

面临问题

- 运维成本:serverlist 维护
- 跨集群服务发现

演进:V5（SOFARegistry）

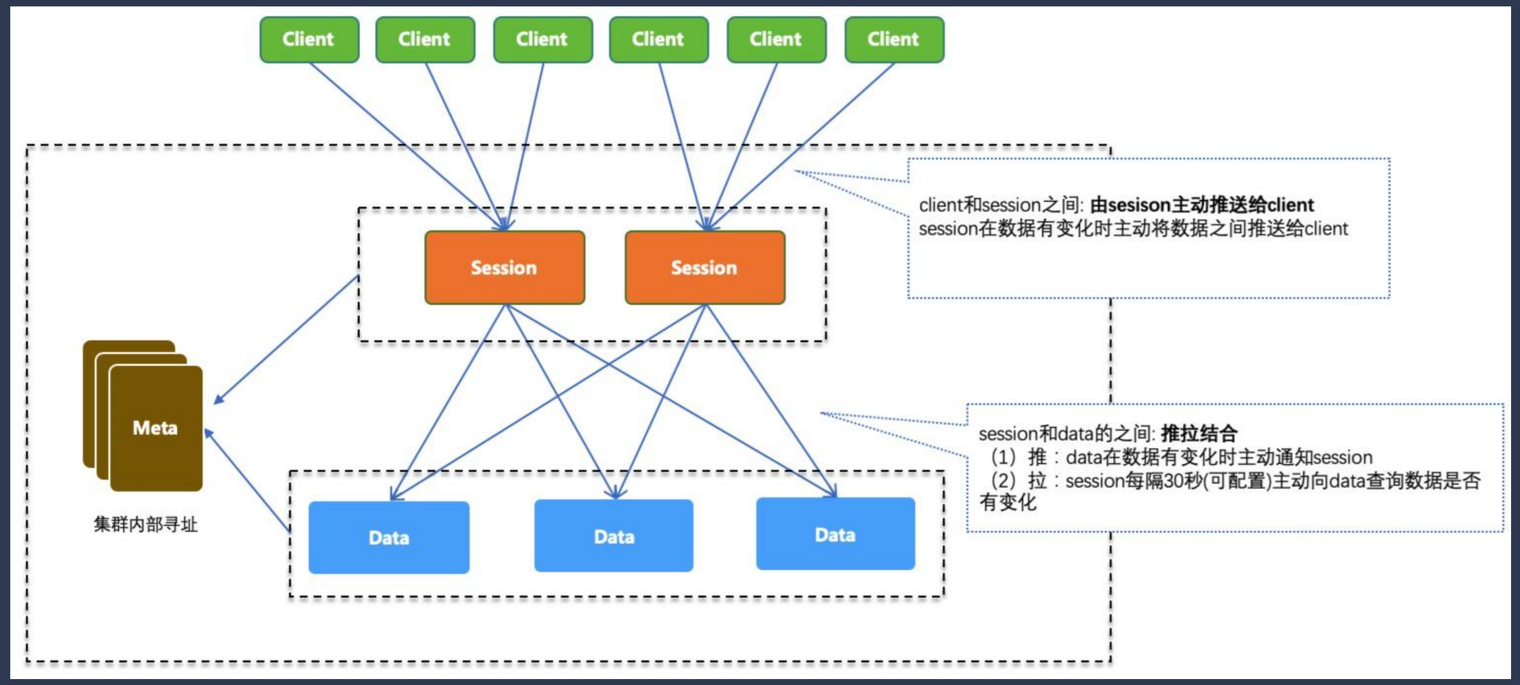


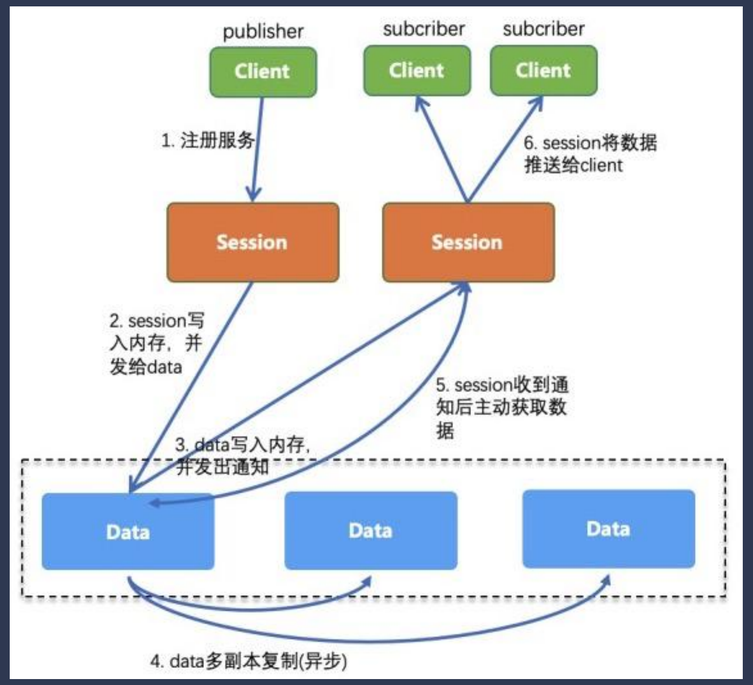


架构:

- 新增 meta(raft):serverlist 维护
- 数据分片：一致性 hash
- 数据多副本容灾

演进:V6（SOFARegistry）





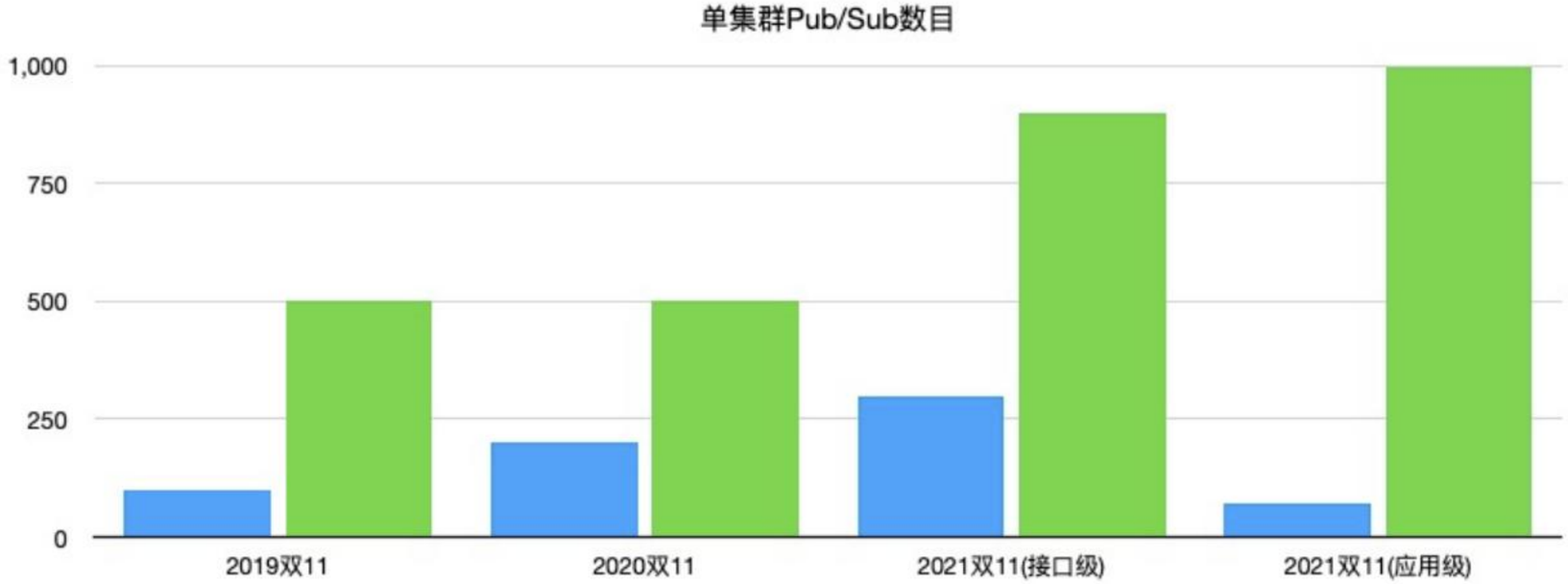
架构:

- meta去强一致性依赖
- 数据分片：SlotTable
- 数据多副本容灾：diff sync

特性对比

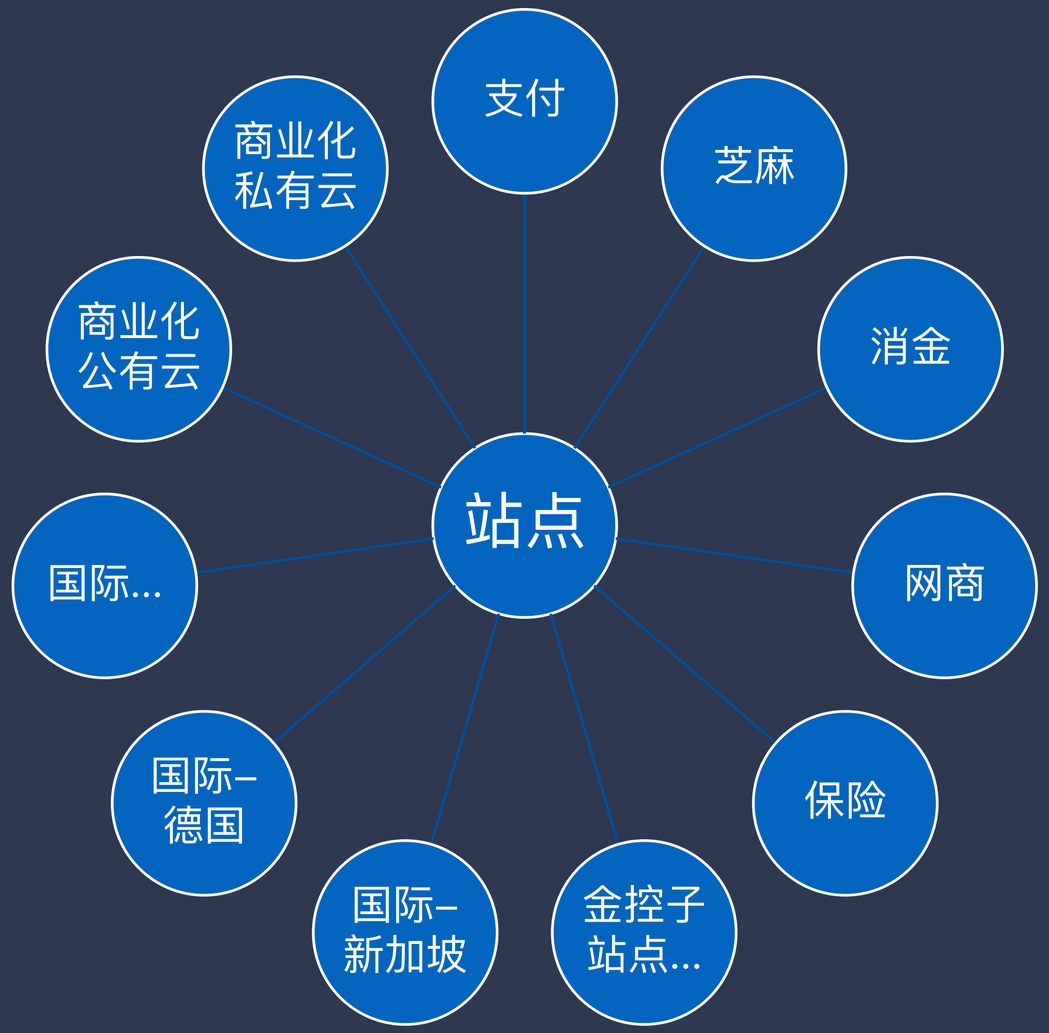
Feature	Consul	Zookeeper	Etcd	Eureka	SOFARegistry
服务健康检查	定期healthcheck (http/tcp/script/docker)	定期心跳保持会话 (session) + TTL	定期 refresh(http)+TTL	定期心跳+TTL;支持自定义healthCheck	定期连接心跳 + 断链敏感
Kv存储服务	支持	支持	支持	–	–
一致性	raft	ZAB	raft	最终一致性	最终一致性
cap	cp	cap	cp	ap	ap
使用接口（多语言能力）	支持http和dns	客户端	http/grpc	客户端/http	客户端（java）
watch支持	全量/支持long polling	支持	支持long polling	不支持(client定期fetch)	支持（服务端推送）
安全	acl/https	acl	https支持	–	acl
spring cloud集成	支持	支持	支持	支持	支持

挑战:数据规模增长(扩展能力)



	Pub	Sub
2019双11	100	500
2020双11	200	500
2021双11(接口级)	300	900
2021双11(应用级)	70	1,000

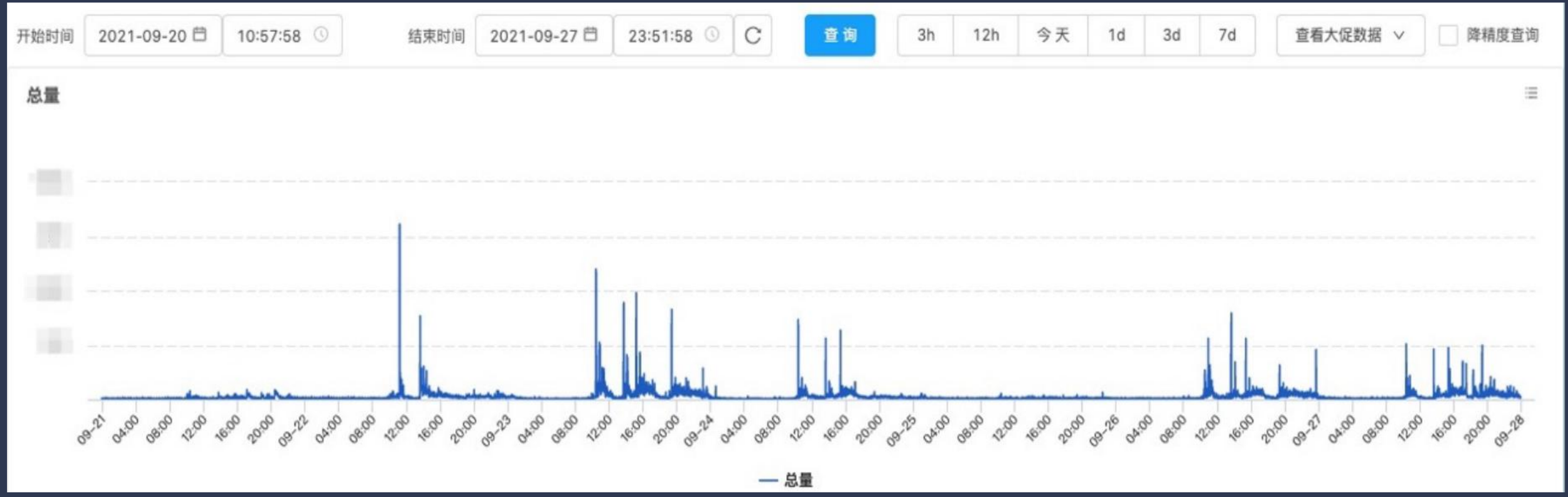
挑战:集群数增长(运维成本)



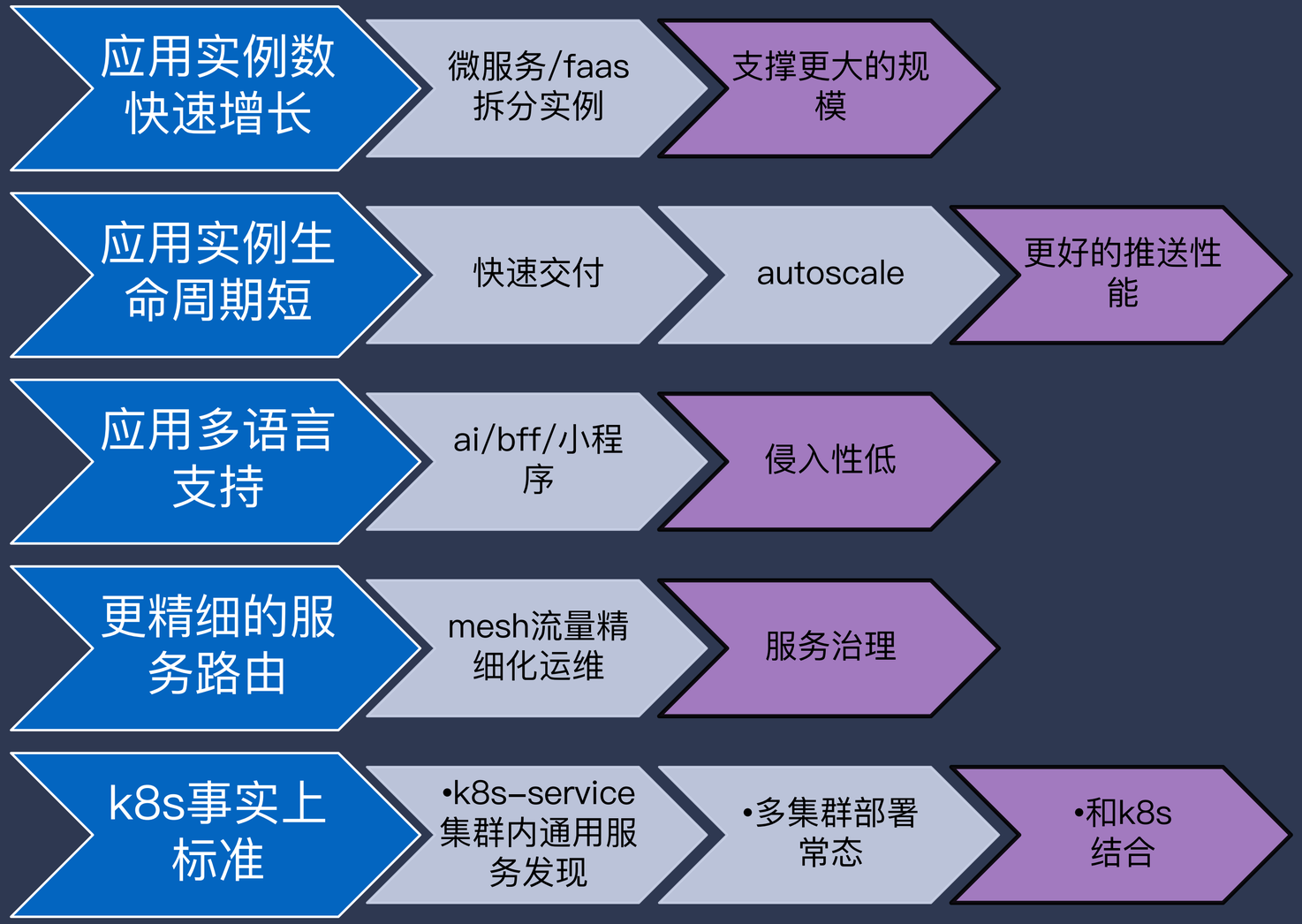
} JOIN {

挑战:业务 7*24 运维(可用性)

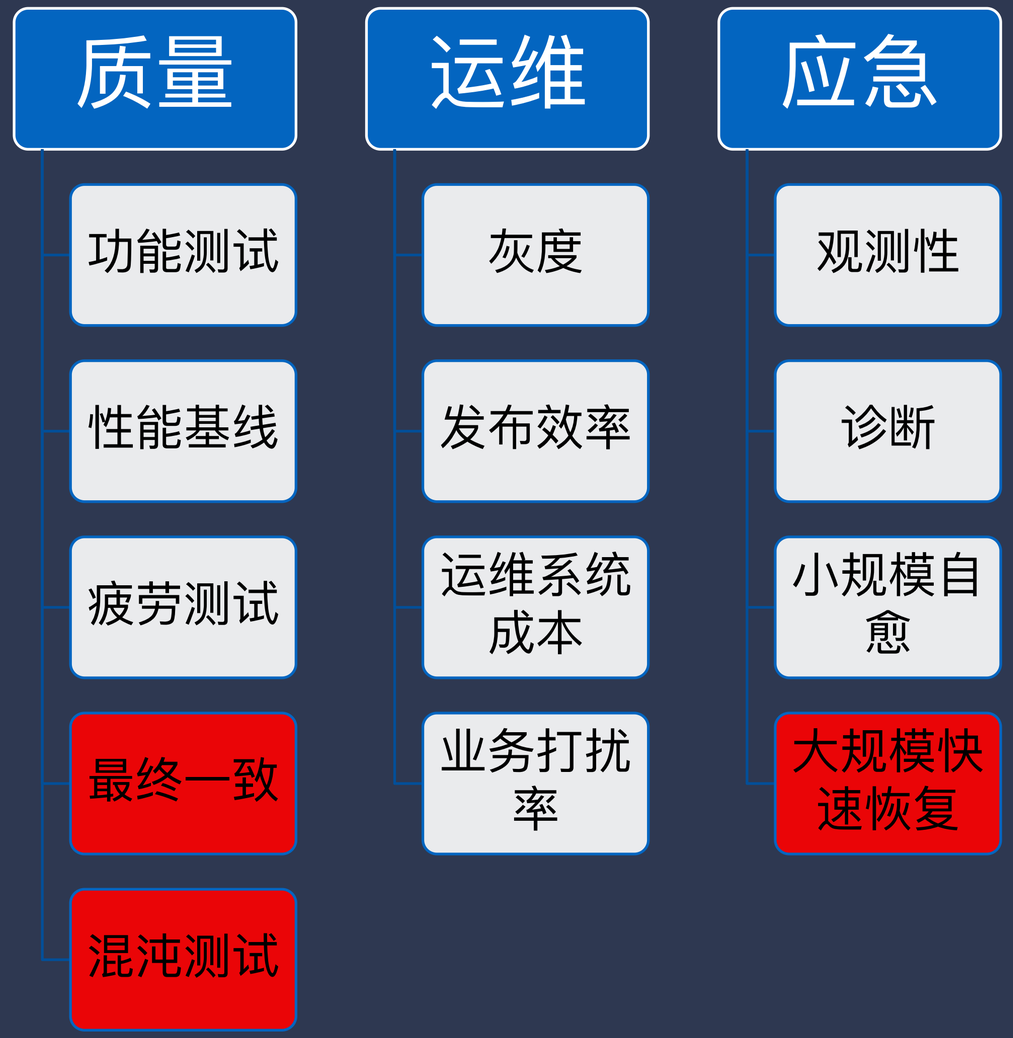
业务运维



云原生:naming 挑战



SOFARegistry6.0 目标:面向效能



SOFARegistry6.0 目标:面向效能

架构

- 应用级服务发现
  - 降数据规模
- 移除raft
  - 节点无状态
- slot&&slotTable
  - 增强数据片管控能力
- 优化数据通信性能
  - 支撑更大规模
- 容灾备份集群
  - 2分钟逃逸
  - 跨版本容灾

质量效能

- SOFARegistryChaos自动化测试
  - 功能回归
  - 性能测试
  - 疲劳测试
  - 混沌测试
- 常规化线上故障攻防演练

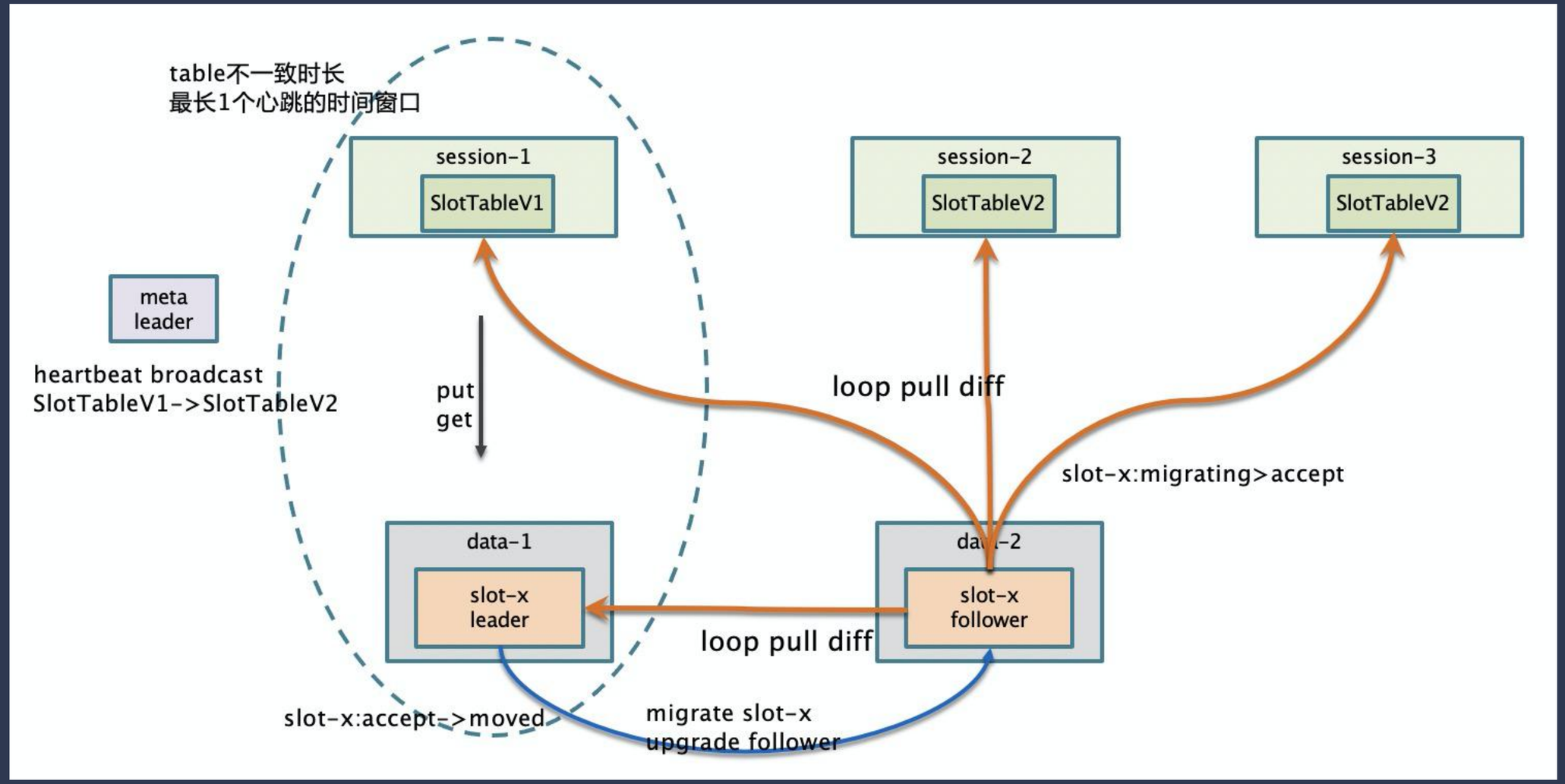
运维效能

- nightly build
  - 灰度环境以下自动发布
- 运维标准化
  - 交付成本低
- 可观测
- 自愈
  - 小规模应急自动化

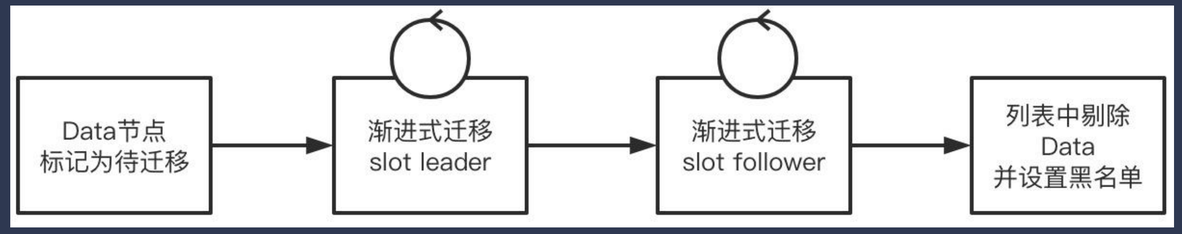
SOFARegistry6.0 架构原则

meta 一致性	•解决强一致的最好方法是不要依赖强一致 •脑裂时 data 节点具备能力获取完整数据
推送正确的数据	•最终一致明确可预期，最终：时间延迟，一致：数据完整性
横向扩展	•数据分片存储，避免单节点存全部数据的约束

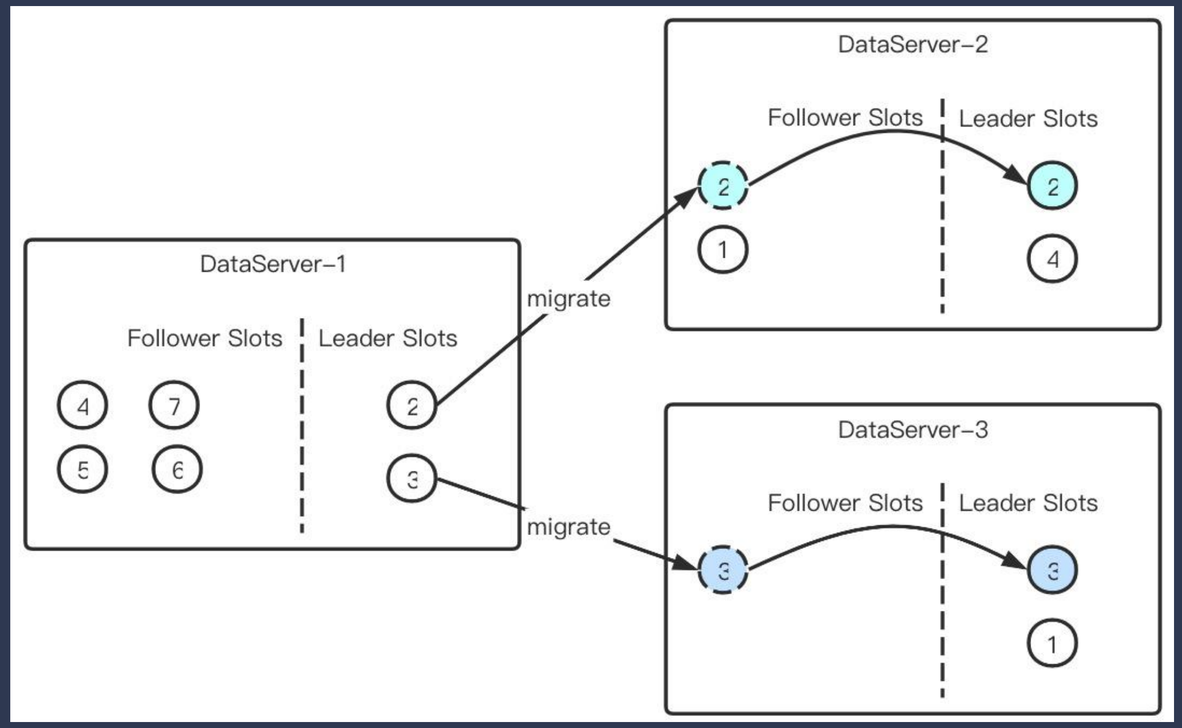
数据最终一致性


table不一致时长
最长1个心跳的时间窗口
session-1
SlotTableV1
session-2
SlotTableV2
session-3
SlotTableV2
meta
leader
heartbeat broadcast
SlotTableV1->SlotTableV2
put
get
loop pull diff
slot-x:migrating>accept
data-1
slot-x
leader
data-2
slot-x
follower
loop pull diff
slot-x:accept->moved
migrate slot-x
upgrade follower

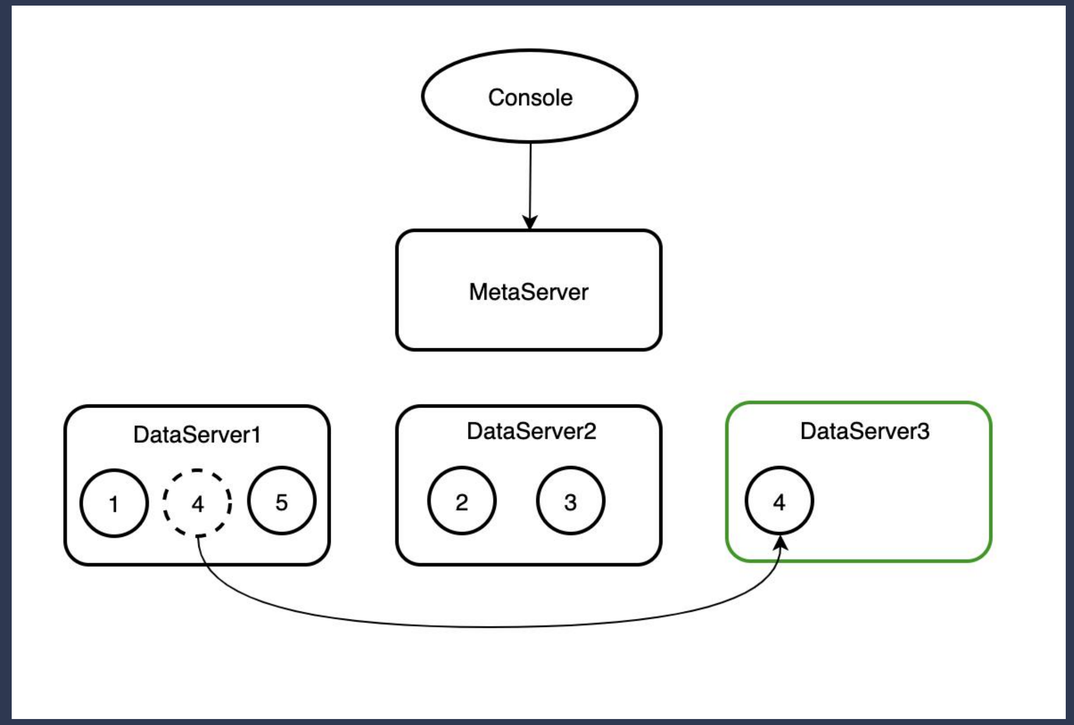
Slot 调度



无损迁移

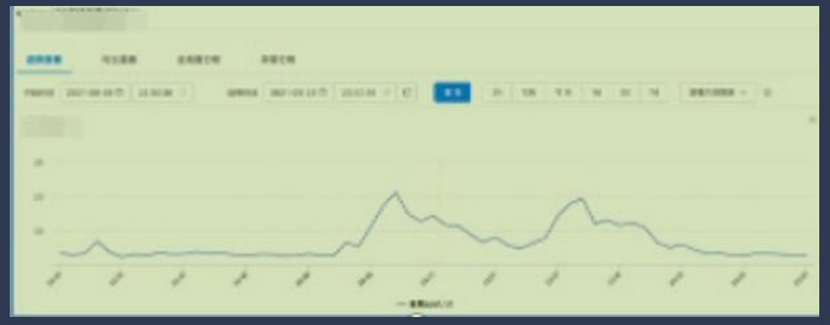


data灰度



性能

- 规模
  - session:240
  - data:50
  - client:10w+
- 地址列表
  - IP数1K~7.2K
  - 包大小 250K~1.8M



应用级–数据模型

- 服务级Publisher实例

JSON 复制代码

```
com.alipay.testapp1.FooService:1.0@DEFAULT
11.34.200.88:8080?v=4.0&_TIMEOUT=3000&_HOSTNAME=testapp1-85-5595&p1&app_name=testapp1&_SERIALIZETYPE=4&tls=false&mosn_version_none&mosn=true

com.alipay.testapp1.BarService:1.0@DEFAULT
11.34.200.88:8080?v=4.0&_TIMEOUT=3000&_HOSTNAME=testapp1-85-5595&p1&app_name=testapp1&_SERIALIZETYPE=4&tls=false&mosn_version=version_none&mosn=true
```

- 实例级Publisher实例

JSON 复制代码

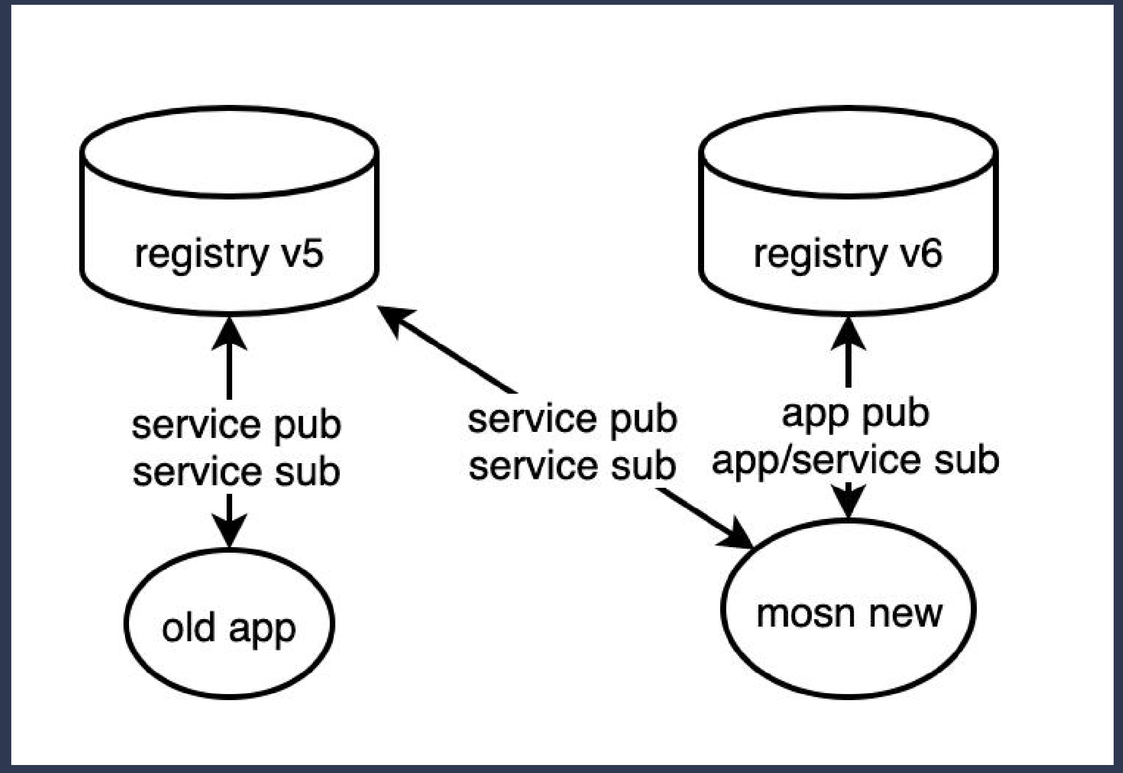
```
{
  "addr": "[redacted]:8080",
  "rev": "testapp1-594f803b380a41396ed63dca39503542",
  "mv": "v1.1.0",
  "bps": {
    "_HOSTNAME": ["testapp1-85-5595"],
    "mosn_version":{"values":["version_none"]},
    "mosn":{"values":["true"]},
  }
}
```

- 元数据

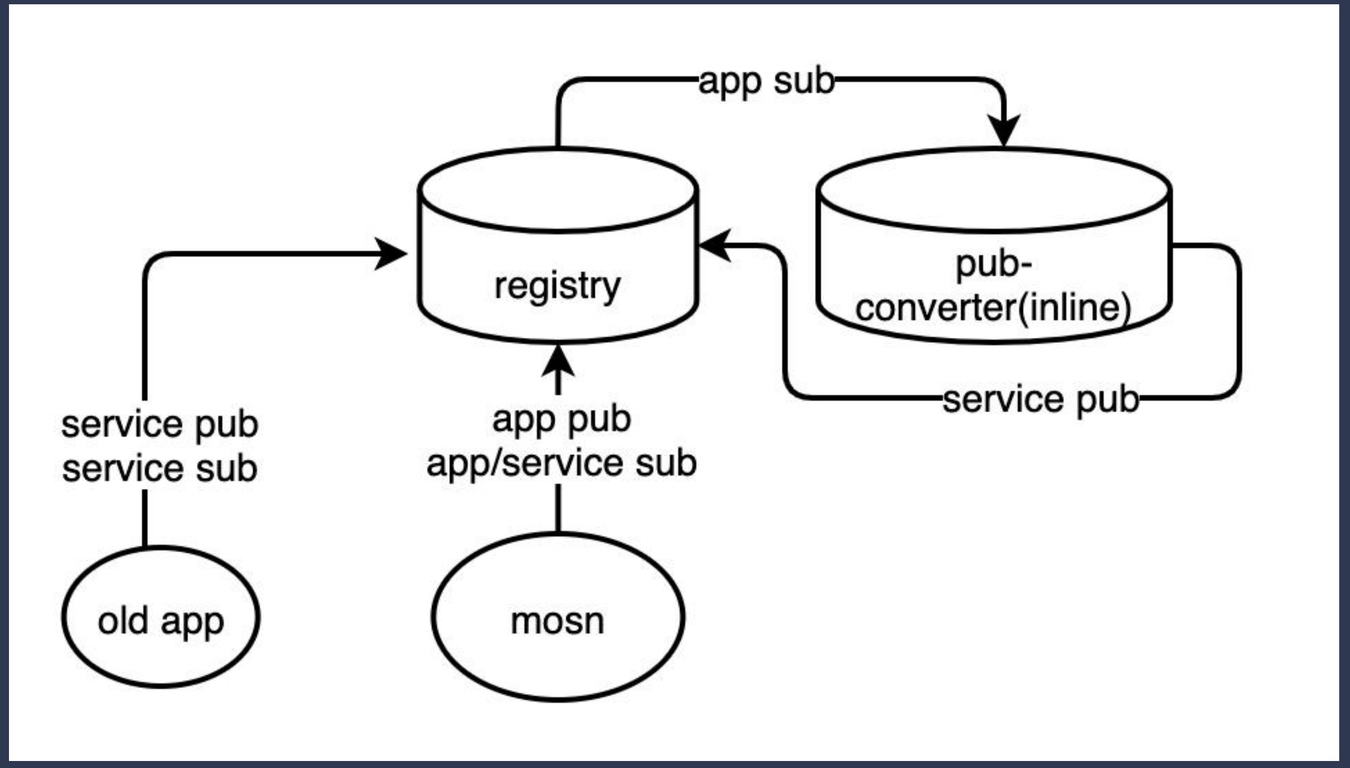
```
{
  "application": "testapp1",
  "revision": "testapp1-594f803b380a41396ed63dca39503542",
  "clientVersion": "v1.1.0",
  "baseParams": {
    "__SERIALIZETYPE":{"values":["4"]},
    "app_name":{"values":["testapp1"]},
    "_TIMEOUT":{"values":["3000"]},
    "tls":{"values":["false"]},
  },
  "services": {
    "com.alipay.testapp1.FooService:1.0@DEFAULT": {
      "id": "0",
    },
    "com.alipay.testapp1.BarService:1.0@DEFAULT": {
      "id": "1",
    }
  }

```

应用级–兼容性:平滑/遗留应用

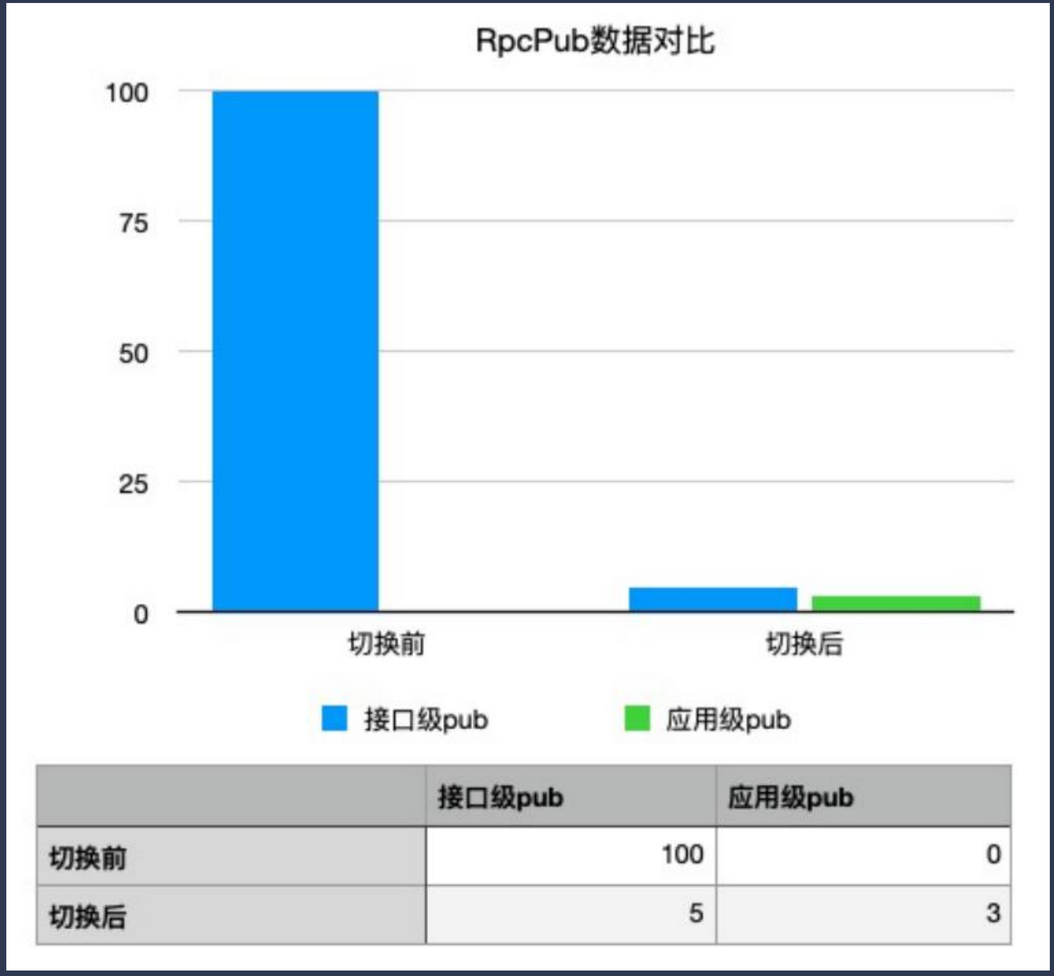
双集群双订阅发布切换



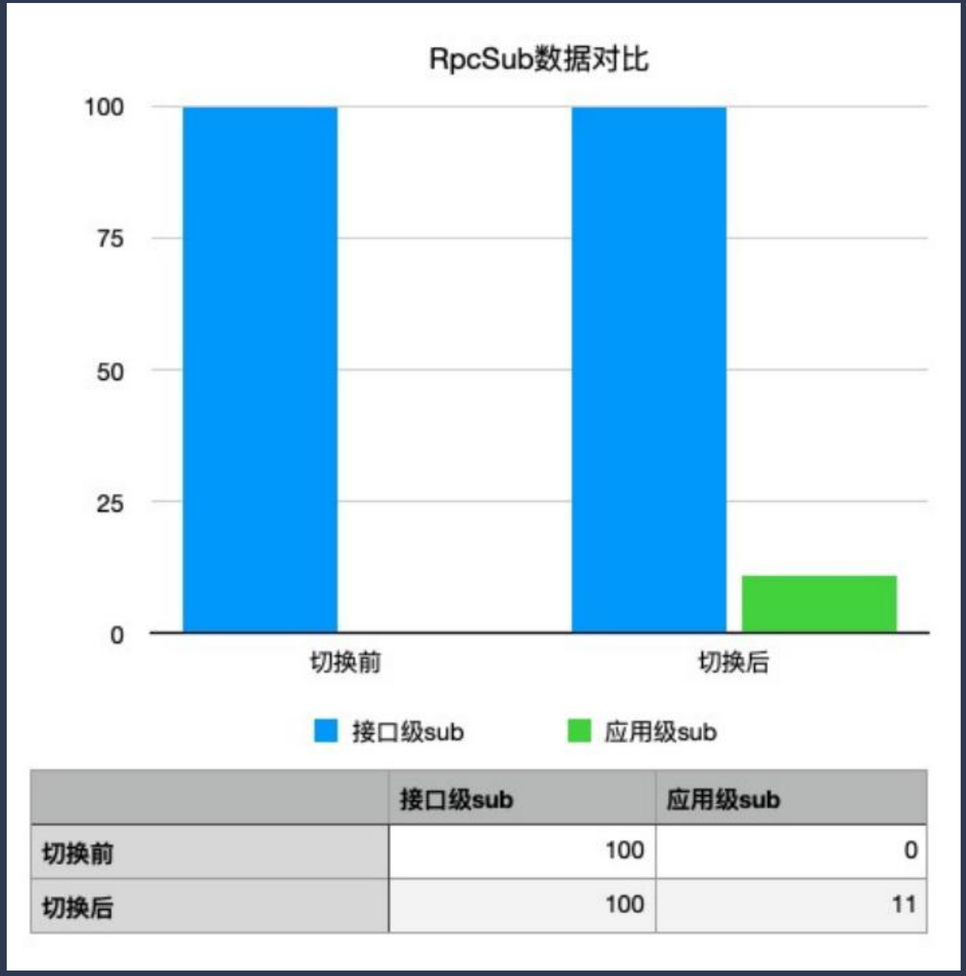
兼容无法升级的遗留应用



应用级–效果:数据下降 1 个数量级



	接口级pub	应用级pub
切换前	100	0
切换后	5	3



	接口级sub	应用级sub
切换前	100	0
切换后	100	11

SOFARegistryChaos:自动化测试

功能

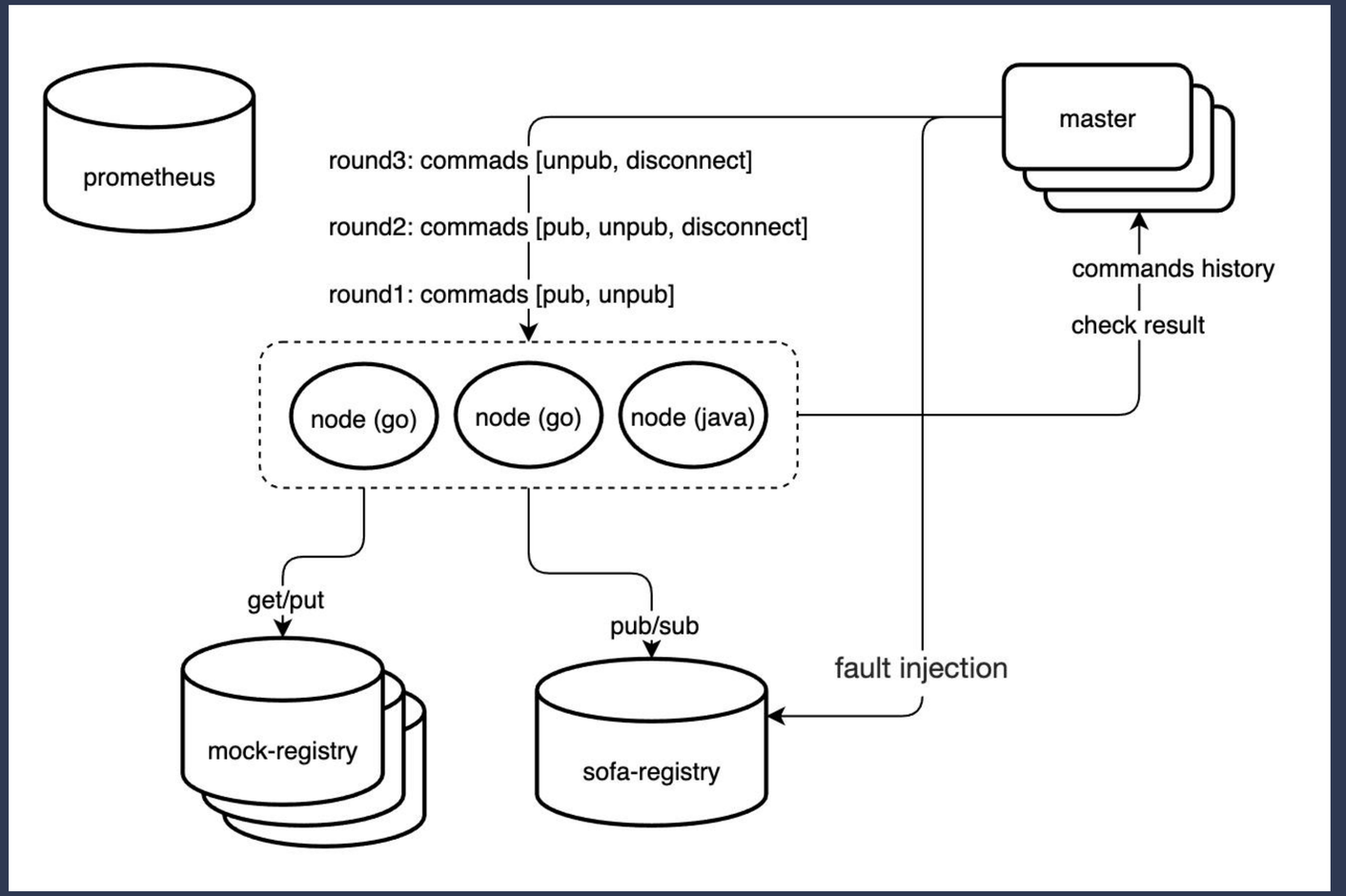
- 基于k8s一键部署
- 横向扩展支持大规模压测
- client编排:pub/sub/disconnect
- 故障编排

观测性

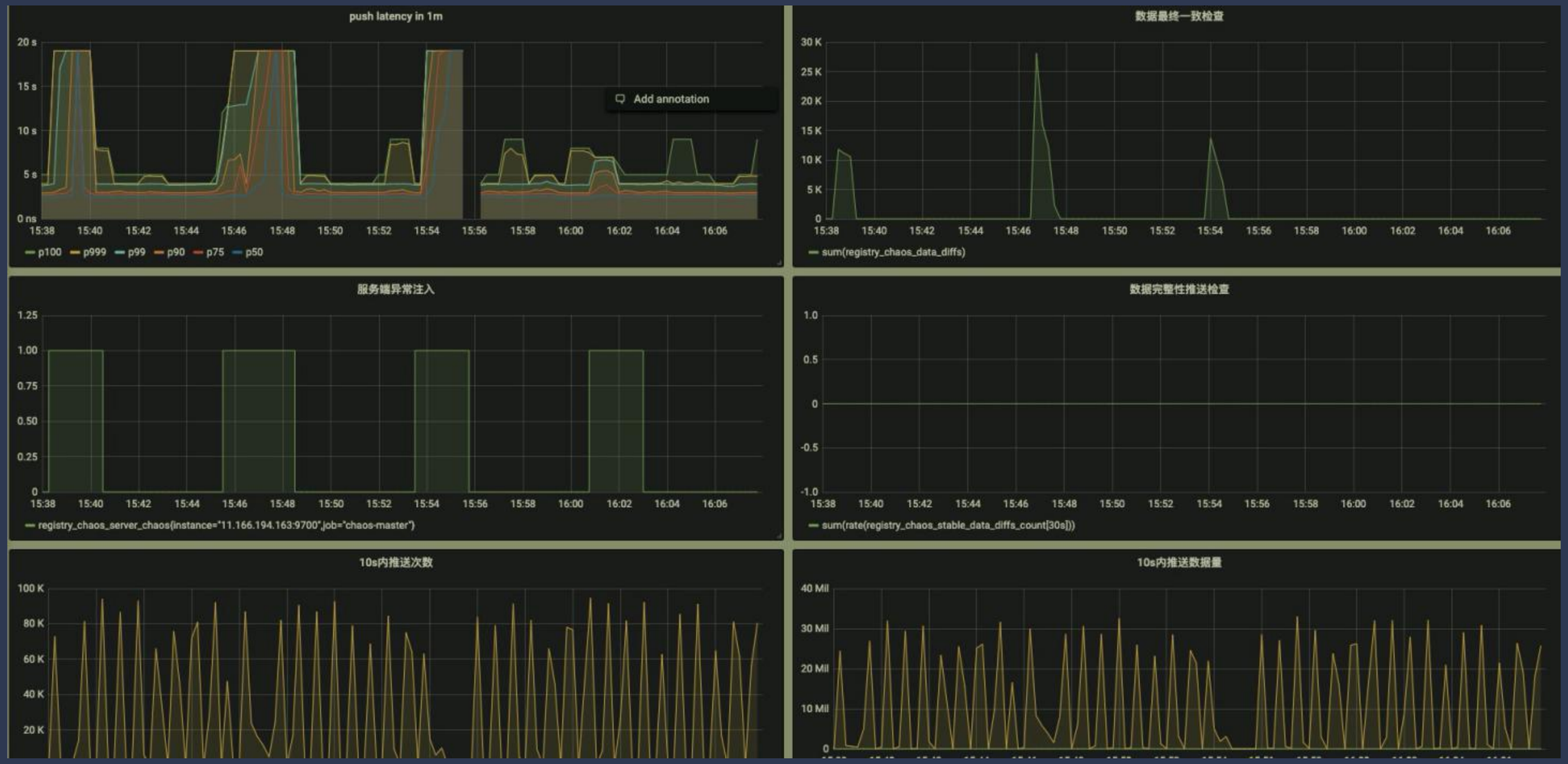
- 端到端推送延迟的分步
- 数据完整性：推少/推错
- 故障注入校验正确性/恢复时间
- 线上小流量部署预警

Trace

- 定位有问题的订阅端/发布端
- 异常期间各个client的状态以及操作历史



SOFARegistryChaos:观测性



SOFARegistryChaos:失败 case

nodeId	total diffs
[redacted]-1	1
[redacted]-30	1
[redacted]-32	1
[redacted]-15	2
[redacted]-28	1

当前数据不一致详细

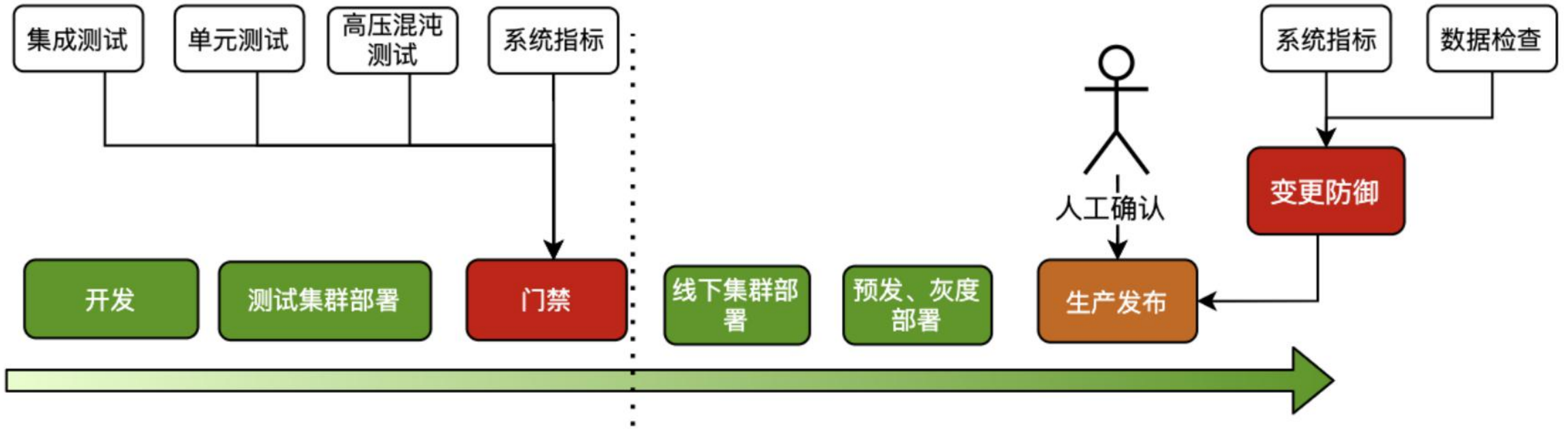
▼ [2021-09-29T03:58:09.547081] [redacted] sub com.alipay.xiangxu.chaos.dev12-24-12@XFIRE

time	2021-09-29T03:58:09.547081
nodeId	[redacted]-1
localAddr	[redacted]
sessionAddr	[redacted]
localZone	[redacted]
dataId	com.alipay.xiangxu.chaos.dev12-24-12@XFIRE
failedZone	[redacted]
scope	dataCenter
expected	[]...(0)
less	[]...(0)
more	[redacted] ndEpoch=625&app_name=app-24]...(1)

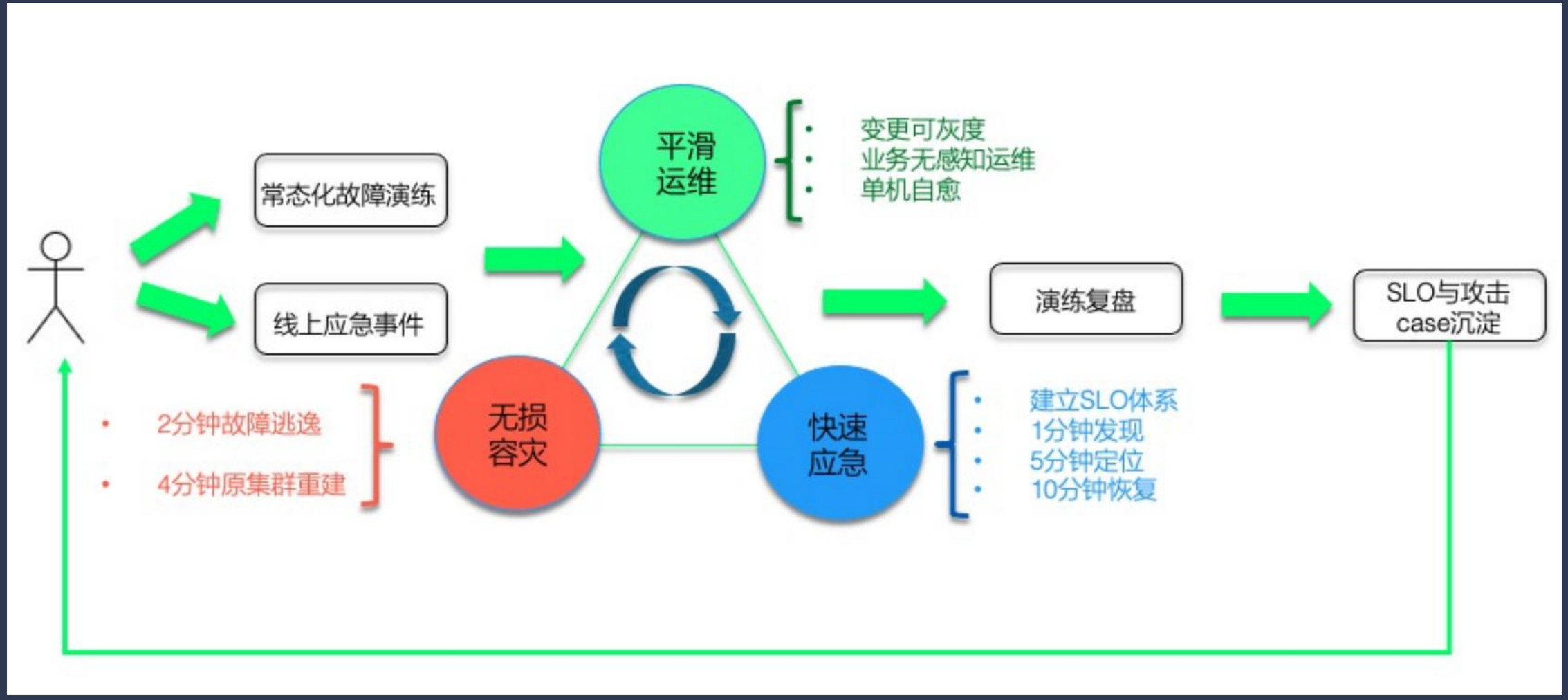
▼ dataId: com.alipay.xiangxu.chaos.dev12-24-12@XFIRE (GZ77F)

```
2021-09-29T03:57:00.117062 [commandEpoch 624] [nod[redacted]sconnect
2021-09-29T03:57:51.333885 [commandEpoch 625] [nod[redacted]0] publish
com.alipay.xiangxu.chaos.dev12-24-12@XFIR[redacted]
2021-09-29T03:57:51.843846 [commandEpoch 625[redacted]5001 disconnect
2021-09-29T03:57:58.930619 [commandEpoch 6[redacted]publish
com.alipay.xiangxu.chaos.dev12-24-12@XFIRE
```

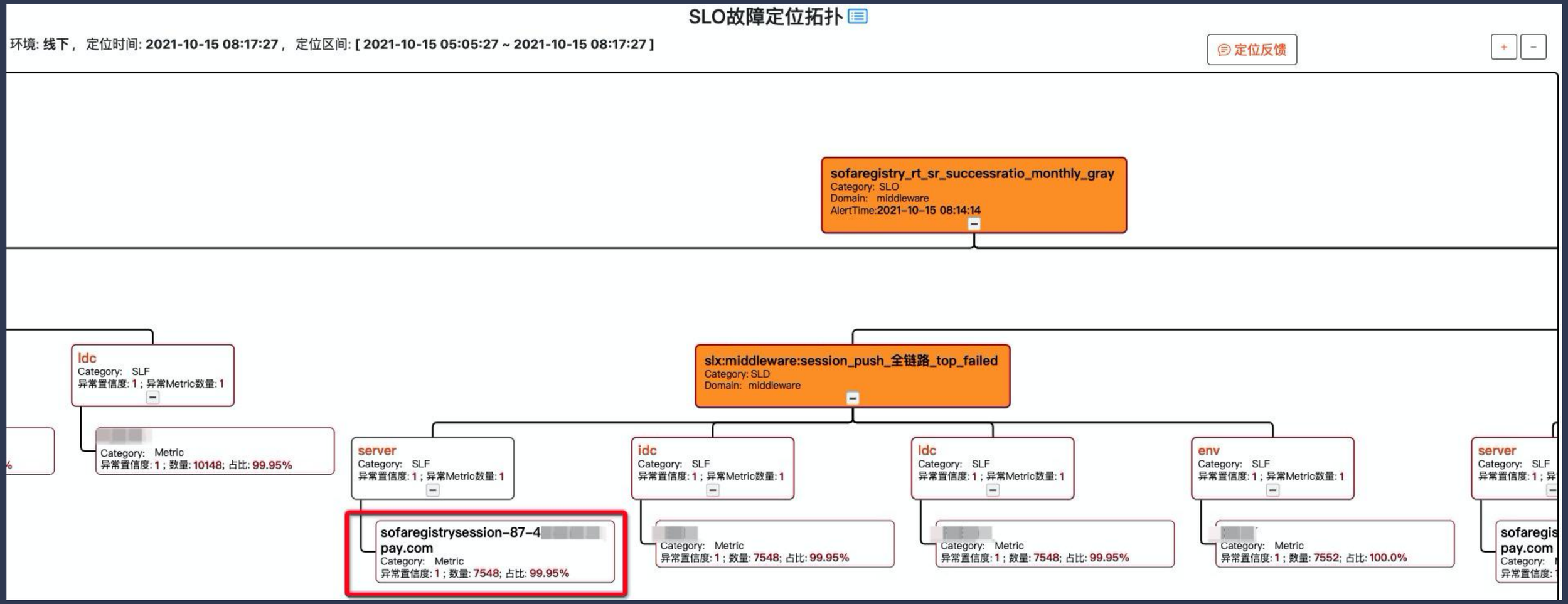
运维效能:nightly build



运维效能:故障演练


常态化故障演练
线上应急事件
平滑
运维
变更可灰度
业务无感知运维
单机自愈
演练复盘
SLO与攻击
case沉淀
无损
容灾
2分钟故障逃逸
4分钟原集群重建
快速
应急
建立SLO体系
1分钟发现
5分钟定位
10分钟恢复

运维效能:定位诊断


SLO故障定位拓扑
环境: 线下， 定位时间: 2021-10-15 08:17:27， 定位区间: [2021-10-15 05:05:27 ~ 2021-10-15 08:17:27]
定位反馈
sofaregistry_rt_sr_successratio_monthly_gray
Category: SLO
Domain: middleware
AlertTime:2021-10-15 08:14:14
idc
Category: SLF
异常置信度: 1 ; 异常Metric数量: 1
Category: Metric
异常置信度: 1 ; 数量: 10148; 占比: 99.95%
slx:middleware:session_push_全链路_top_failed
Category: SLD
Domain: middleware
server
Category: SLF
异常置信度: 1 ; 异常Metric数量: 1
sofaregistrysession-87-4
pay.com
Category: Metric
异常置信度: 1 ; 数量: 7548; 占比: 99.95%
idc
Category: SLF
异常置信度: 1 ; 异常Metric数量: 1
Category: Metric
异常置信度: 1 ; 数量: 7548; 占比: 99.95%
ldc
Category: SLF
异常置信度: 1 ; 异常Metric数量: 1
Category: Metric
异常置信度: 1 ; 数量: 7548; 占比: 99.95%
env
Category: SLF
异常置信度: 1 ; 异常Metric数量: 1
Category: Metric
异常置信度: 1 ; 数量: 7552; 占比: 100.0%
server
Category: SLF
异常置信度: 1 ; 异
sofaregis
pay.com
Category:
异常置信度:

开源和共赢

一个人可以走得很快，但一群人可以走的更远

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

四年磨一剑：蚂蚁集团注册中心 SOFARegistry 的开发实践之路

向旭 / 李旭东

THANKS