趣丸SLA质量体系下的业务高可用建设

2024年10月

SREelite

https://sre-elite.com

个人介绍

李楠

业务SRE负责人

工作经历：

2020 -至今，趣丸网络科技

个人简介：

- 目前主要负责趣丸业务稳定性相关工作：SLA体系建设、业务高可用和容灾建设、动态风险治理、运维大模型助手探索等。
- 从事业务稳定性保障、虚拟化、容器化、Devops平台开发和运维数字人探索等工作。
- 有10年大规模业务稳定性运维管理经验。

目录

CONTENT

SREelite

01

第一部分 SLA质量体系与业务高可用的关系

SREelite

1.SLA质量体系与业务高可用的关系

SREelite

用户体验SLA是目标牵引，高可用建设是落地手段

这个关系是怎么总结出来的?

如何度量高可用成效，优化措施对用户体验是否提升

如何避免过度设计，过度冗余，过渡收缩?

如何设计、优化ROI最佳。

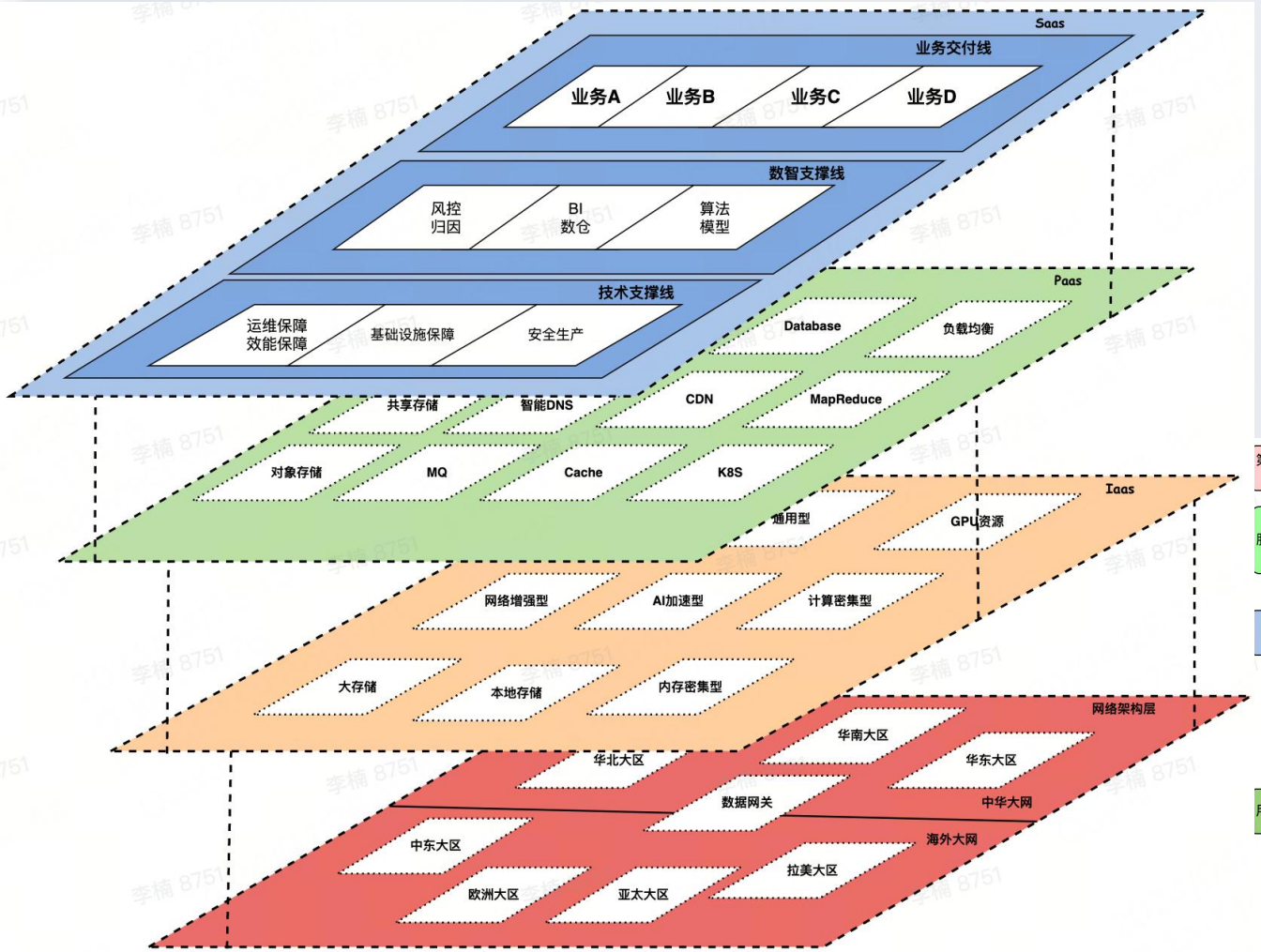
如何在业务动态高峰中，持续保障业务达标?

02

第二部分 如何通过SLA体系驱动高可用建设

SREelite

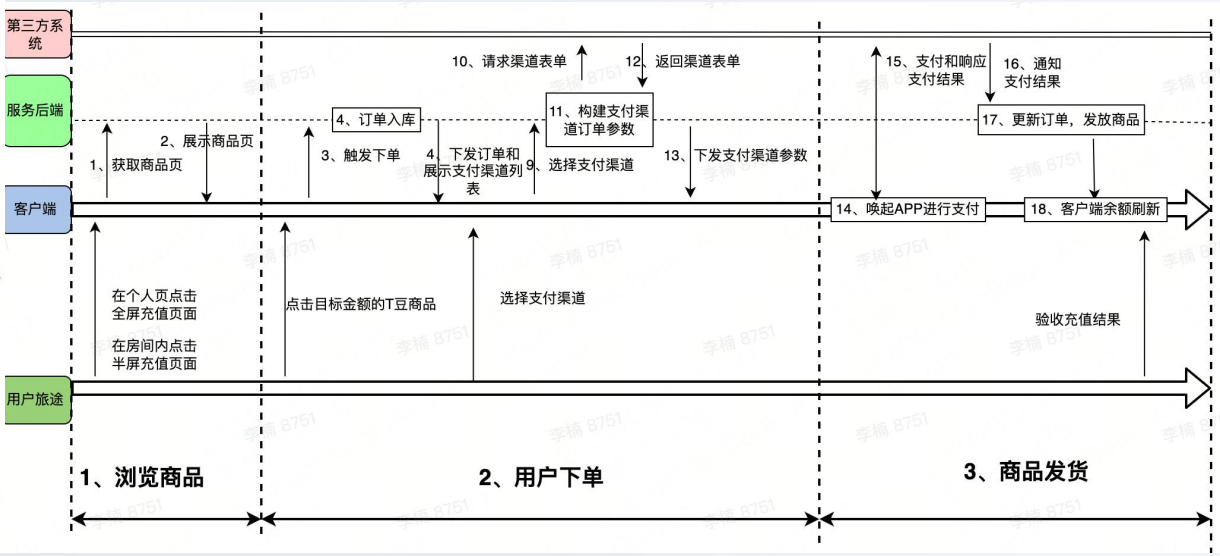
2.如何通过SLA体系驱动高可用建设



1、确定业务平面--最前端的用户体验SLA目标

2、基于用户旅途梳理核心链路和关键应用功能

3、通过业务分层架构进行横向扩展、纵向下钻，确认每个业务组件功能的SLA



2.如何通过SLA体系驱动高可用建设

SREelite

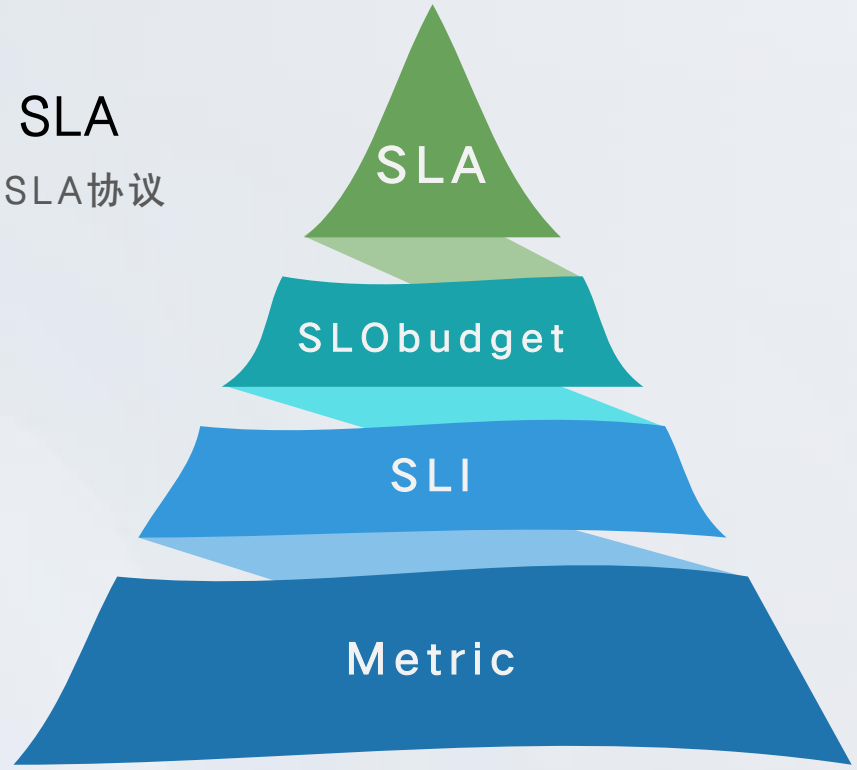
SLA

对客户方承诺的SLA协议

SLI黄金指标

基于SLO档案，梳理核心SLI

治理SLI告警有效性



SLO budget

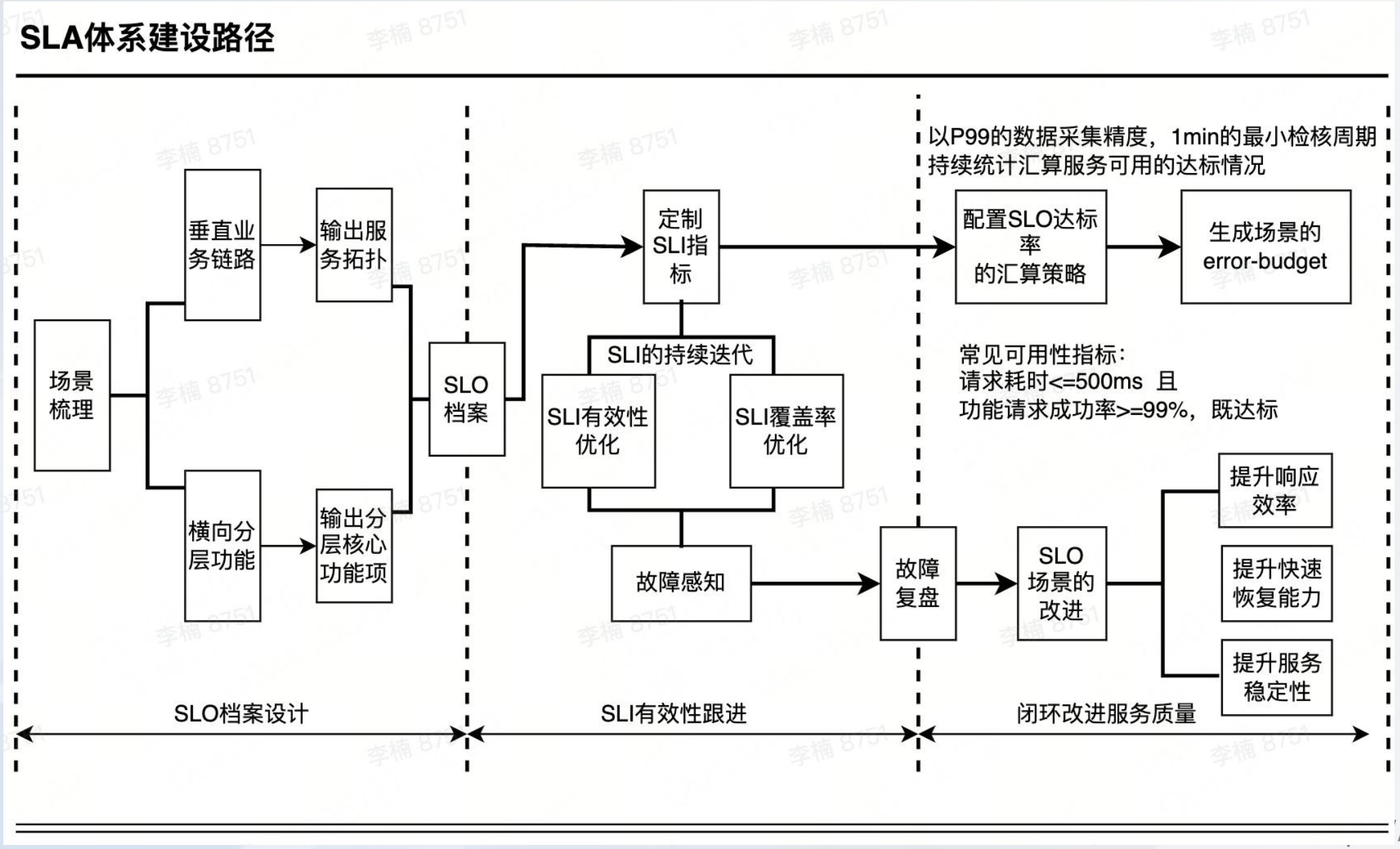
基于SLA产生目标、时间、范围的约束 产生SLO故障预算

Metric

构建可观测能力，持续保障高可用建设的成效

2.如何通过SLA体系驱动高可用建设

SREelite



03

第三部分 趣丸科技的高可用建设实践和成效

3.趣丸科技的高可用建设实践和成效

SREelite

如何定义“业务具备高可用了”

如何逐层分析高可用建设方案

如何应对动态的业务高峰，保障业务SLO持续达标

3.1如何定义“业务具备了高可用”

“业务具备了高可用能力”

明确高可用需要具备的要素，在这个方向的牵引下进行建设，
使业务系统达到SLA的预定目标，并在合理的ROI预期内。

3.1趣丸在实践中总结的高可用十要素：

具备以下策略和流程，确保服务连续性和稳定性达到承诺目标，业务才算具备高可用性。

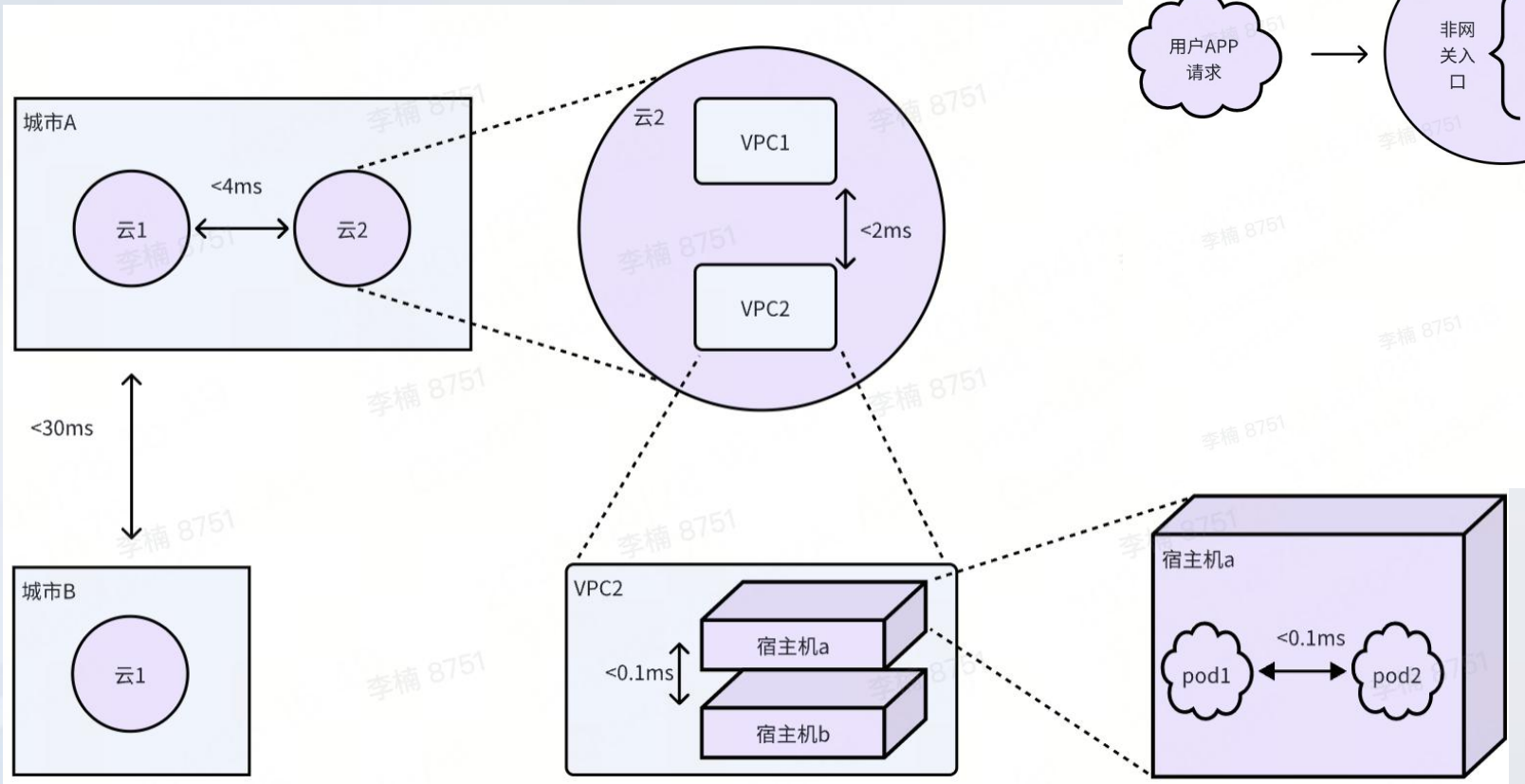
服务等级协议（SLA）：明确的可用性目标，如99.9%。
冗余设计：关键组件有备份，防止单点故障。
容错能力：部分故障不影响整体服务。
灾难恢复和备份：快速恢复服务的策略和流程。
自动故障转移：故障时自动切换到正常组件。
性能和可伸缩性：负载增加时保持性能，自动扩展资源。
监控和报警：跟踪服务健康和及时报警。
持续测试：定期验证高可用性策略。
维护和更新：定期维护和避免高峰时更新。
文档和培训：记录系统信息和培训人员。

3.2逐层分析设计高可用建设方案

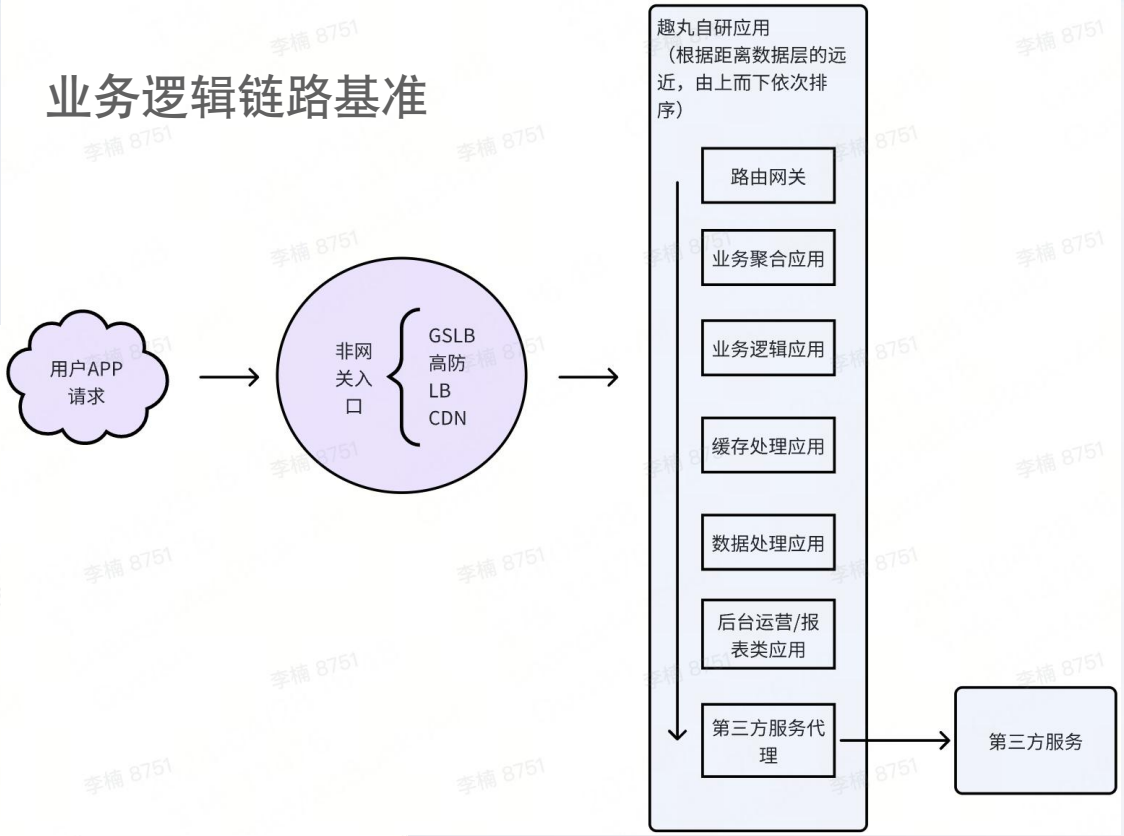
SREelite

分层分解链路，定义指标基准

物理环境基准



业务逻辑链路基准



3.2分层分解链路，定义指标基准

SREelite

分类	特征说明（指标都以P99数据为准）	优秀指标	良好指标	及格指标	劣质指标
路由网关	纯做转发，极少逻辑，比较看重效率	<100us	<500us	<1ms	<2ms
业务聚合应用 （业务客户端网关）	具有流量调度、请求汇聚功能的应用	<200ms	<500ms	<2s	<5s
业务逻辑应用	mg/kafka纯消费者、定时任务、活动开关等	<5ms	<10ms	<30ms	<100ms
缓存处理应用(R&W)	具备处理缓存中间件数据的应用	<0.5ms	<1ms	<2ms	<20ms
数据处理应用(R&W)	具备处理数据库数据的应用	<5ms	<10ms	<30ms	<100ms
后台运营&报表类应用	主要功能后台运营管理，报表输出等	<1s	<3s	<5s	<8s
第三方代理服务	需要跨公网调用	<200ms	<500ms	<1s	<3s
推荐类聚合业务	需要处理大量数据和复杂排序的应用	<150 ms	<400 ms	<750 ms	<1000 ms
GPU类服务-推理	音频：rtf（实时率）来衡量， 10s音频合成需要多少s	<60ms	<100ms	<300ms	<500ms
	图像: 512*512 , webui, steps=25	<2s	<3s	<5s	<8s

指标分级介绍：

优秀：行业内优秀。

良好：公司内优秀，质量治理到该水位后，可以持续保持，可以不做进一步要求。

及格：SLA签订的指标阈值最低线。

劣质：服务方与客户方谈论SLA协议的门槛。

3.2根据评估结论制定ROI合理的方案

SREelite

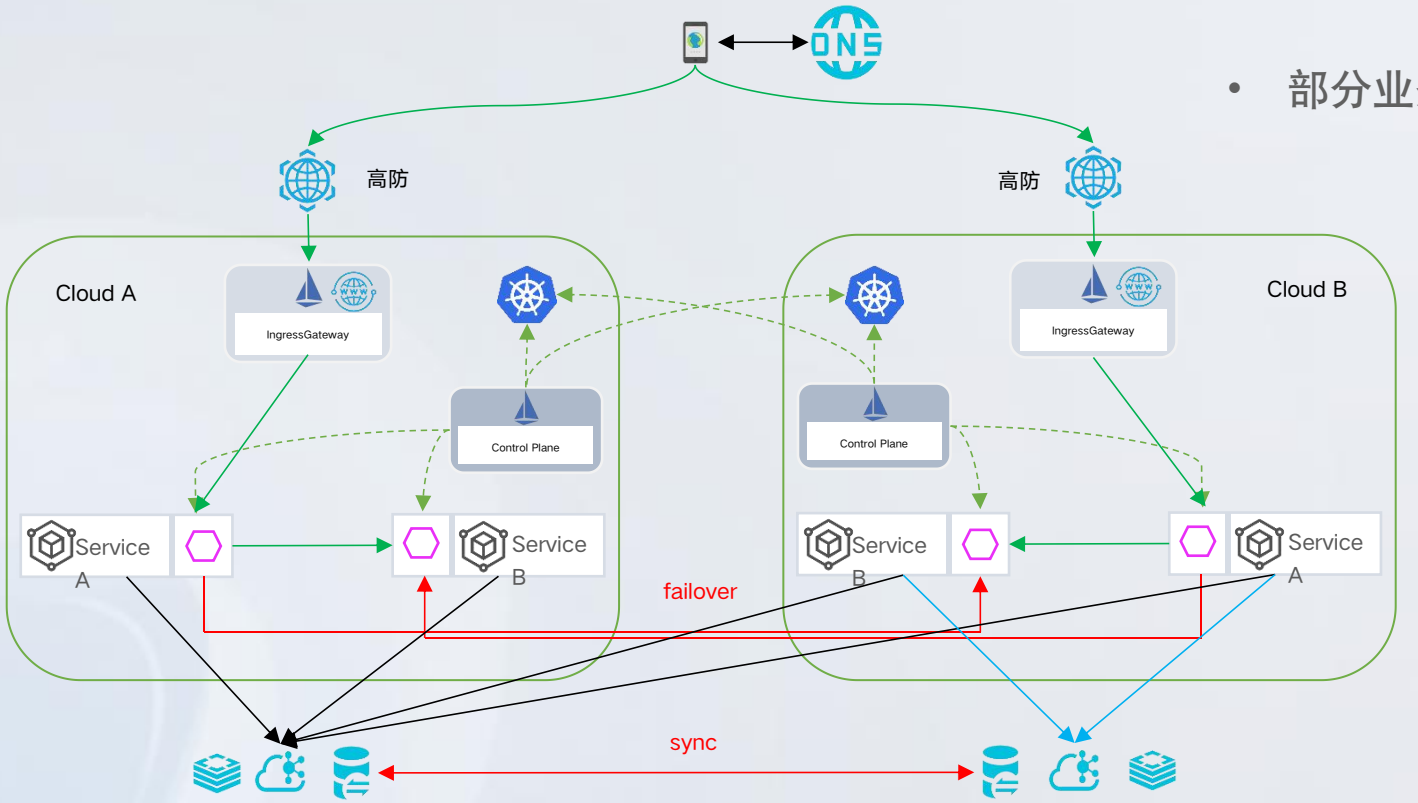
层级		故障场景	故障影响面（按最严重的评估）	发生频率	RTO（现状能力）			RPO（数据类-现状能力）			ROI评估（根据当前现状，规划做到下一个级别能力所需的投
					冷备	热备	多活	冷备	热备	多活	
					小时级	分钟级	秒级	小时级	分钟级	秒级	
应用		[illegible]Region故障	[illegible]业务不可用	[illegible]		√					[illegible]
	业务逻辑服务	韧性故障	个别服务不可用	每月1次		√					边际成本适中，边际效益适中
		单元故障		-							边际成本适中，边际效益适中
		单可用区故障	部分服务不可用	每年1次		√					边际成本适中，边际效益适中
		单Region故障	整体业务不可用	每2年1次		√					边际成本适中，边际效益适中
基础架构组件	Kubernetes	镜像仓库不可用	个别服务不可用	-		√					边际成本适中，边际效益适中
		CoreDNS不可用	核心业务不可用	每2年1次		√					边际成本适中，边际效益适中
		组件异常-Pod无法部署	部分服务不可用	每2年1次		√					边际成本极高，边际效益极低
		资源容量不足-pod pending	个别服务不可用	每季度1次		√					边际成本适中，边际效益适中
		单节点故障	个别服务不可用	每季度1次		√					边际成本极高，边际效益极低
	istiod	性能不足-xds下发异常	部分服务不可用	每年2-3次							边际成本极高，边际效益极低
		网关性能不足-服务延时高	部分服务不可用	每年1次		√					边际成本极低，边际效益极高
		证书不可用	整体业务不可用	-							无需进一步提升能力
	中间件	单节点故障	个别服务不可用	每季度2-3次			√			√	无需进一步提升能力
		集群不可用	部分服务不可用	每年1次	√						边际成本适中，边际效益适中
		集群容量性能不足	个别服务不可用	每年2-3次		√				√	边际成本极高，边际效益极低
		单可用区所有中间服务都不可用	个别服务不可用	每2年1次			√			√	无需进一步提升能力
		单Region所有中间服务不可用	整体业务不可用	每2年1次	√						边际成本极高，边际效益极低
		单节点故障	个别服务不可用	每月2-3次			√			√	无需进一步提升能力
		集群不可用	部分服务不可用	每年2-3次		√					边际成本适中，边际效益适中

3.2最终落地效果

SREelite

趣丸多云架构现状

- 业务多云多活，数据层部分多活
- 部分业务实现单边写，就近读



3.3如何应对动态的业务高峰

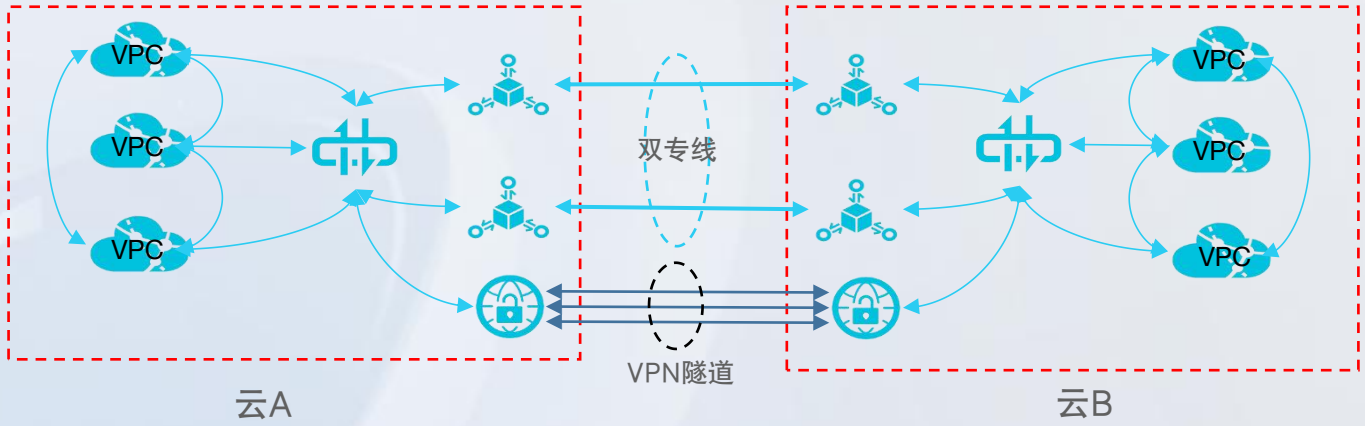
3.3.1快速止损，优先恢复业务

SREelite

- 灾难冗余设计
- 自动故障转移
- 性能和可伸缩性
- 监控告警

网关和专线的容灾冗余设计

- 云原生高防：接入简单、延迟低、防护性能灵活选择
- WAF：基于 Istio IngressGateway ，接入简单、灵活定制



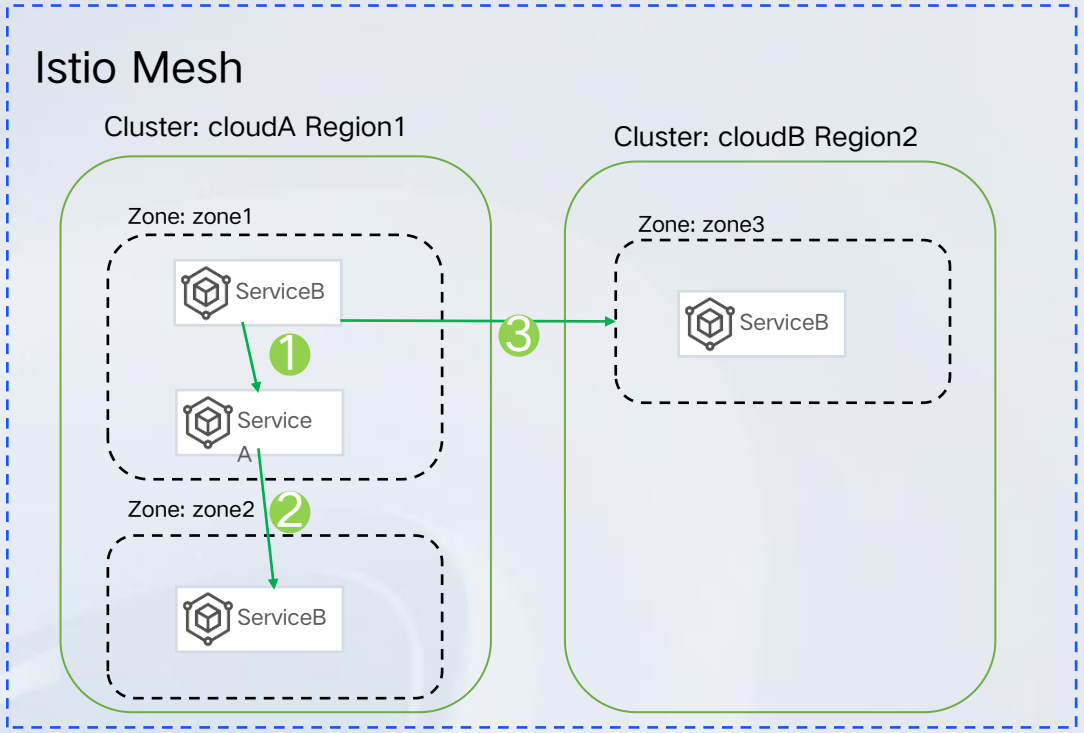
双专线+VPN热备（BGP-ECMP + BFD）

- 在双专线中断时，VPN接管流量保障核心业务不中断
- 多条VPN共担流量，保障带宽充裕

业务网关的故障转移

SREelite

引入Istio的流量路由方案

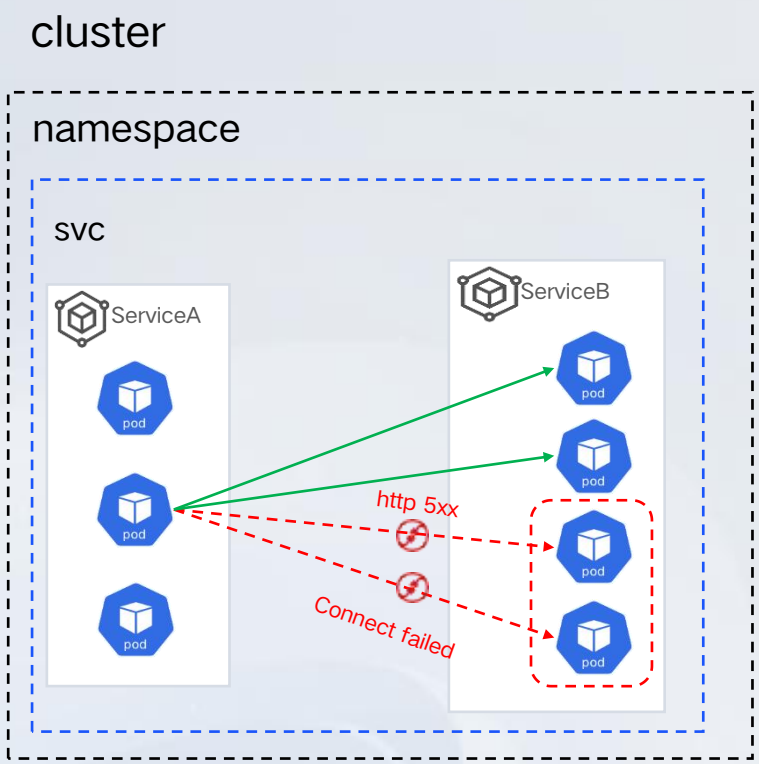


流量管理策略：本地优先

- 优先访问本Region，本Zone
- 本Zone失效，优先访问本Region其他Zone
- 本Region失效，访问其他Region的Zone

```
loadBalancer:
    localityLbSetting:
      enabled: true
      failoverPriority:
      - topology.istio.io/network
      - topology.kubernetes.io/region
      - topology.kubernetes.io/zone
```

应用层故障转移



故障转移策略

- 在应用层面降低故障的影响
- 故障转移策略越接近应用优先级越高

```
outlierDetection:
    baseEjectionTime: 60s
    consecutiveGatewayErrors: 10
    consecutiveLocalOriginFailures: 10
    interval: 10s
    maxEjectionPercent: 60
    splitExternalLocalOriginErrors: true
```

```
trafficPolicy:
  connectionPool:
    tcp:
      connectTimeout: 200ms
```

多层级、多维度的弹性策略

SREelite

```yaml
spec:
  advanced:
    horizontalPodAutoscalerConfig:
      behavior:
        scaleDown:
          policies:
          - periodSeconds: 300
            type: Percent
            value: 10
          stabilizationWindowSeconds: 300
  maxReplicaCount: 50
  minReplicaCount: 5
  scaleTargetRef:
    name: push-notification-v2
  triggers:
  - metadata:
      desiredReplicas: "25"
      end: 30 23 * * *
      start: 50 19 * * *
      timezone: Asia/Shanghai
    type: cron
  - metadata:
      desiredReplicas: "35"
      end: 30 00 * * *
      start: 30 23 * * *
      timezone: Asia/Shanghai
    type: cron
  - metadata:
      value: "85"
    metricType: Utilization
    type: cpu
```

Pod HPA (Horizontal Pod Autoscaler)

- 按照资源利用率扩容
- 业务高峰期提前扩容
- 梯度缩扩容缓解 xDS 下发压力

充分利用云资源弹性优势，保障稳定性的同时降本

业务高峰期SLO达标，容器集群7天平均CPU利用率达到30%

规则类型	触发条件	执行动作	操作
指标触发	CPU 分配率 > 85%	关联的节点池都增加 5 个节点	删除 编辑
周期触发	在19:45	关联的节点池都增加 5 个节点	删除
周期触发	在21:45	关联的节点池都增加 5 个节点	删除
周期触发	在23:45	关联的节点池都增加 5 个节点	删除

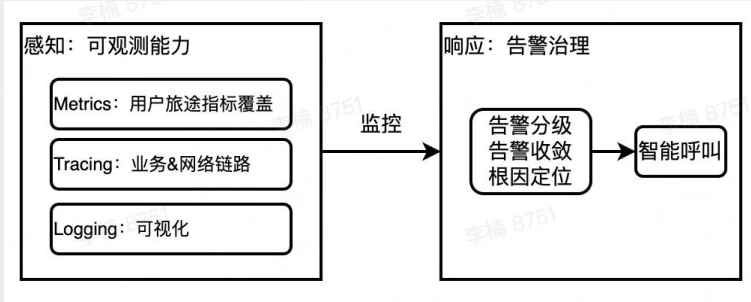
K8S 节点 CA (Cluster Autoscaler)

- 按照资源利用率扩容
- 业务高峰期提前扩容

面向应用的可观测建设和告警治理

将所有资源全上链至应用

围绕应用建设可观测和治理告警



告警有效性治理原则：
第一：我们将告警反应的问题处理了，告警恢复。
第二：告警无效，通过各种手段将无效变为有效。

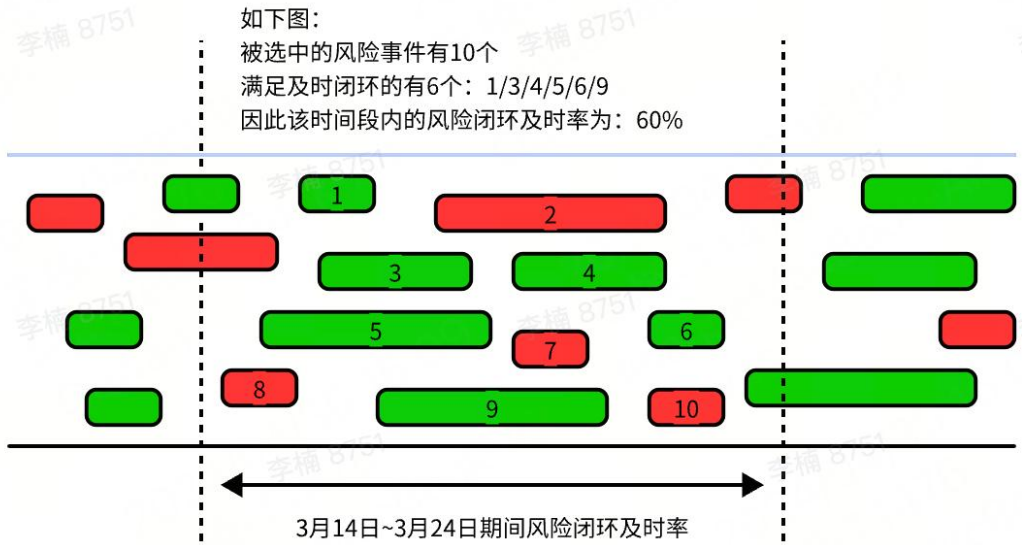
3.3.2持续测试，验收整改成效，杜绝高可用策略失效

SREelite

```yaml
# 端到端 接口级别的网络延迟故障注入
apiVersion: networking.istio.io/v1alpha3
kind: VirtualService
metadata:
  name:                    r-return-503
spec:
  hosts:
  - xxx.xxx.svc.cluster.local
  http:
  - match:
    - uri:
        exact: /xx/xxx
    fault:
      delay:
11      percentage:
12        value: 100
13      fixedDelay: 5s
    route:
    - destination:
        host: xxxx.xxx.svc.cluster.local
  - route:
    - destination:
        host: xx.xx.svc.cluster.local
```

演练方案	演练进度	演练场景	是否符合预期	演练等级	年份	半年	季度
【演练报告】2024年3月6日	已完成		符合预期	P1	2024年	2024年H1	Q1
【演练报告】2024年3月6日	已完成		部分符合预期	P1	2024年	2024年H1	Q1
【演练报告】2024年3月14日	已完成		符合预期	P1	2024年	2024年H1	Q1
【演练报告】2024年3月14日	已完成		符合预期	P1	2024年	2024年H1	Q1
【演练报告】2024年3月20日	已完成		不符合预期	P1	2024年	2024年H1	Q1
【演练报告】2024年3月20日	已完成		符合预期	P1	2024年	2024年H1	Q1
【演练报告】2024年3月20日	已完成		符合预期	P1	2024年	2024年H1	Q1
【演练报告】2023年11月30日	已完成				2023年	2023年H2	Q4
【演练报告】2023年11月17日	已完成		不符合预期	P1	2023年	2023年H2	Q4
【演练报告】2023年11月28日	已完成		部分符合预期	P1	2023年	2023年H2	Q4
【演练报告】2023年08月23日	已完成		符合预期	P3	2023年	2023年H2	Q3
【演练报告】2023年8月30日	已完成		符合预期	P3	2023年	2023年H2	Q3
【演练报告】2023年07月19日	已完成		符合预期	P2	2023年	2023年H2	Q3
【演练报告】2023年06月21日	已完成		符合预期	P3	2023年	2023年H1	Q2
【演练报告】2023年06月19日	已完成		部分符合预期	P2	2023年	2023年H1	Q2

3.3.3 持续闭环，引入动态风险治理



用例说明：

处理有效期为1天的风险事件

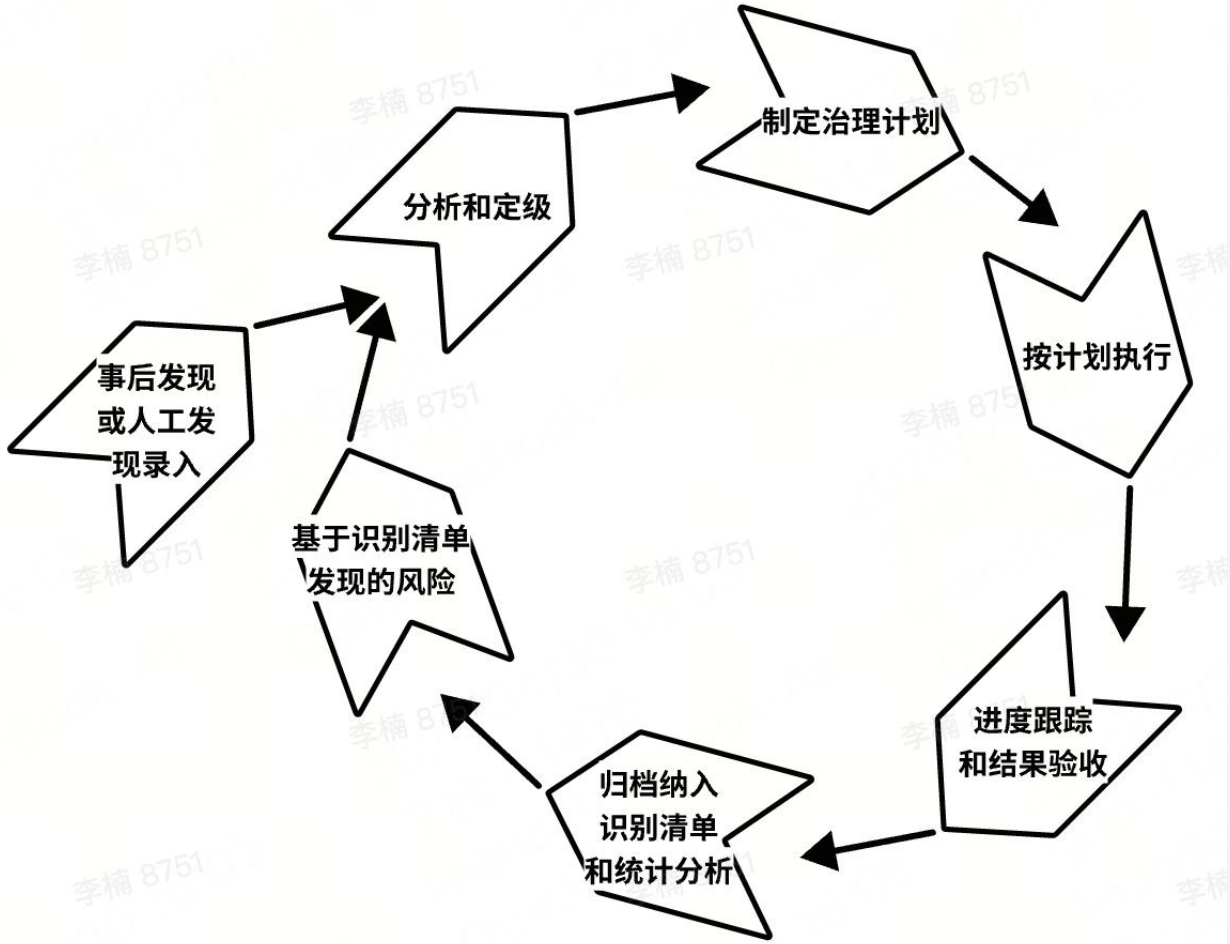
处理有效期为3天的风险事件

处理有效期为7天的风险事件

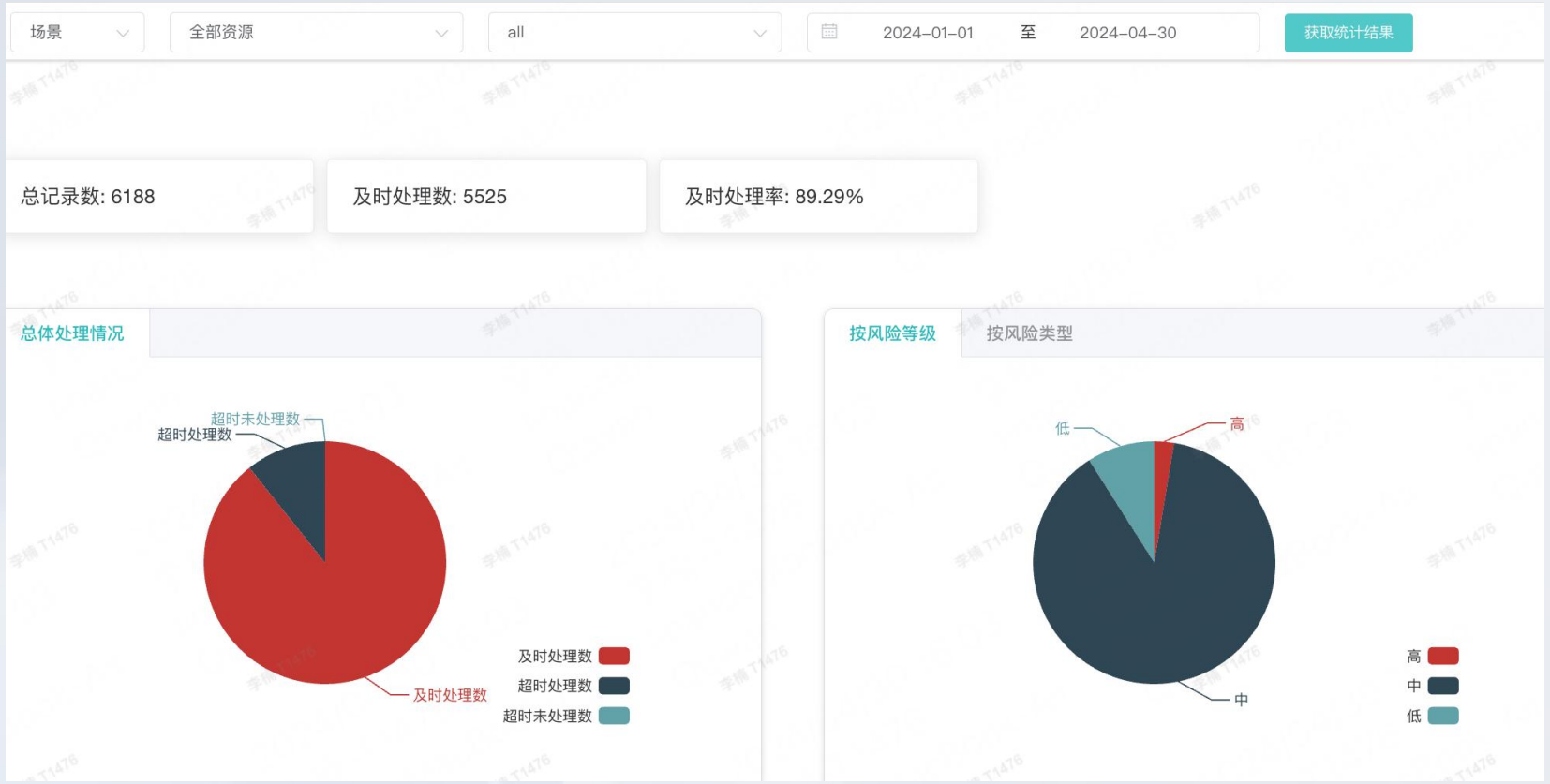
已解除的风险事件标记为绿色

未解除的风险事件标记为红色

加粗边框为被选定的事件



3.3.3 对各层的高可用状态持续巡检和改进任务追踪



巡检内容包括：

- 网络层的拨测
- 入口的冗余容灾
- 云商底层资源的反亲和巡检
- 应用pod的跨region、跨zone部署巡检
- 存储层、数据层的多可用区巡检

04

总结&展望

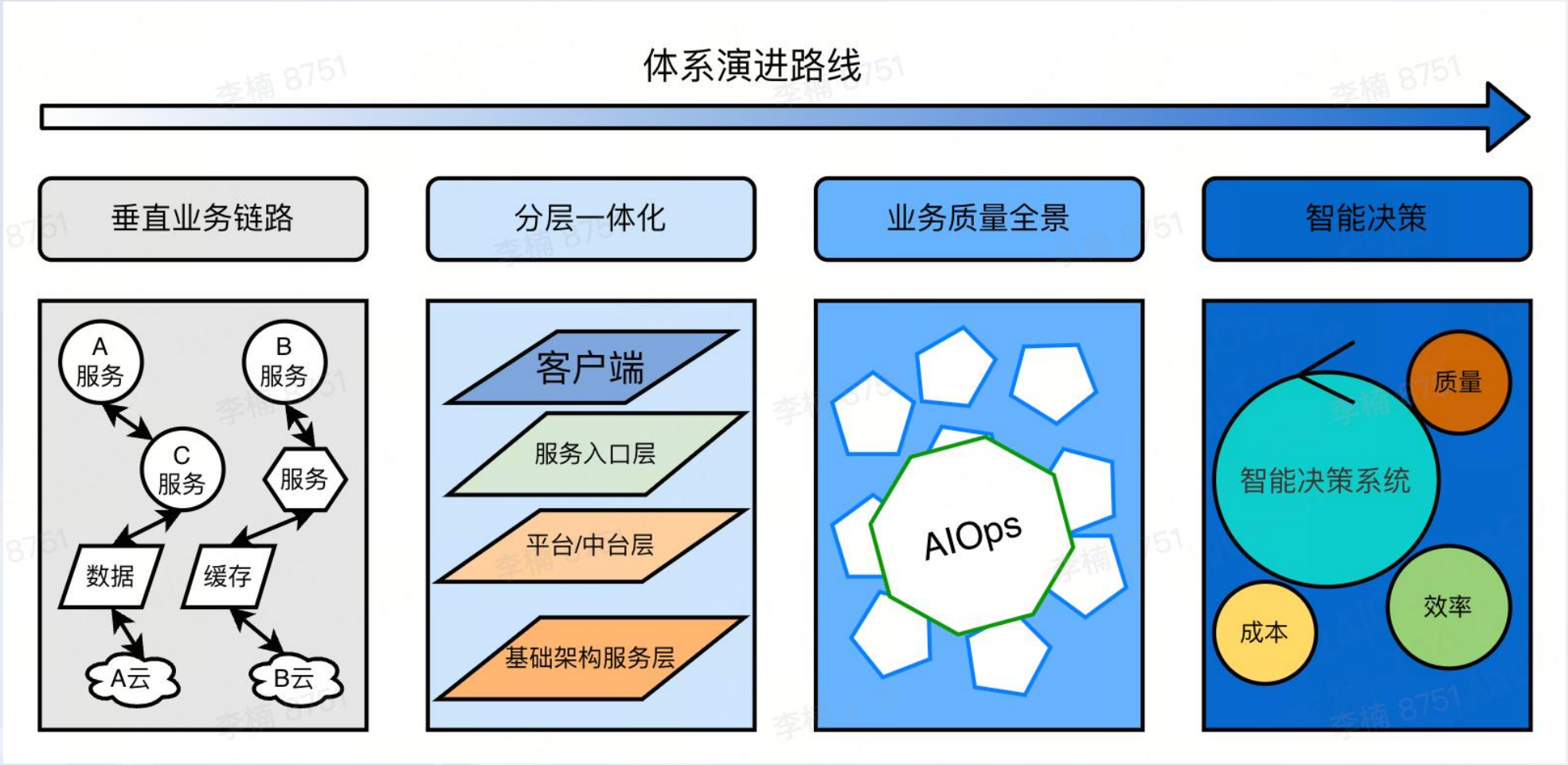
SREelite

4.1总结：SLA 牵引下的高可用建设四原则

SREelite

- **ROI合理性**：质量提升到一定水平，边际收益很低，业务服务质量保障要考虑投入产出比，不会不计成本的提升业务服务质量
- **故障难以完全避免**：业务处于不断发展变化中，组织人员也在不断变化，复杂性不断提升，故障始终难以完全避免
- **区分重点和优先级**：业务服务质量保障和业务交付效率保障，二者存在一定的资源竞争关系，阶段性的改进目标要区分重点和优先级
- **主动降级非核心服务**：为保障核心业务流程的高可用，会设计服务降级机制，可能导致非核心业务体验受损

4.2展望：基于大模型的智能感知、精准定位、主动运维



当前初步完成分层一体化的构建过程

下一阶段：构建业务质量全景

在此基础上引入大模型实现：智能巡检，风险精准定位，主动跟踪进闭环。

Q&A

SREelite

https://sre-elite.com

联盟介绍

社区、白皮书、研讨

SREelite

https://sre-elite.com

“SRE精英联盟”概述

SREelite

2003年
Google 成立 SRE 团队

2016年


孙宇聪翻译出版首部 SRE 著作

2020年

赵成牵头成立 SRE研讨社群

2023年11月
SREelite
SRE 精英联盟正式成立

2024年2月
《SRE 实践白皮书》首次正式发布


2024年2月29 日
联盟首次直播


2024年4月13 日
联盟首次线下 Meetup


《SRE实践白皮书》

SREelite

- 经历数年，20 多位一线专家协作编写。
- 扫码下载 v1.0.4 。版本持续更新迭代中。
- 在官网 https://sre-elite.com/notice/下载最新版。

关注我们

SREelite

公众号

视频号

B 站

YouTube