故障注入技术方案

文档历史

修订日期	修订内容	修订版本	修订人
2022-07-04	创建文档	v0.1	liangming.huang@shopee.com

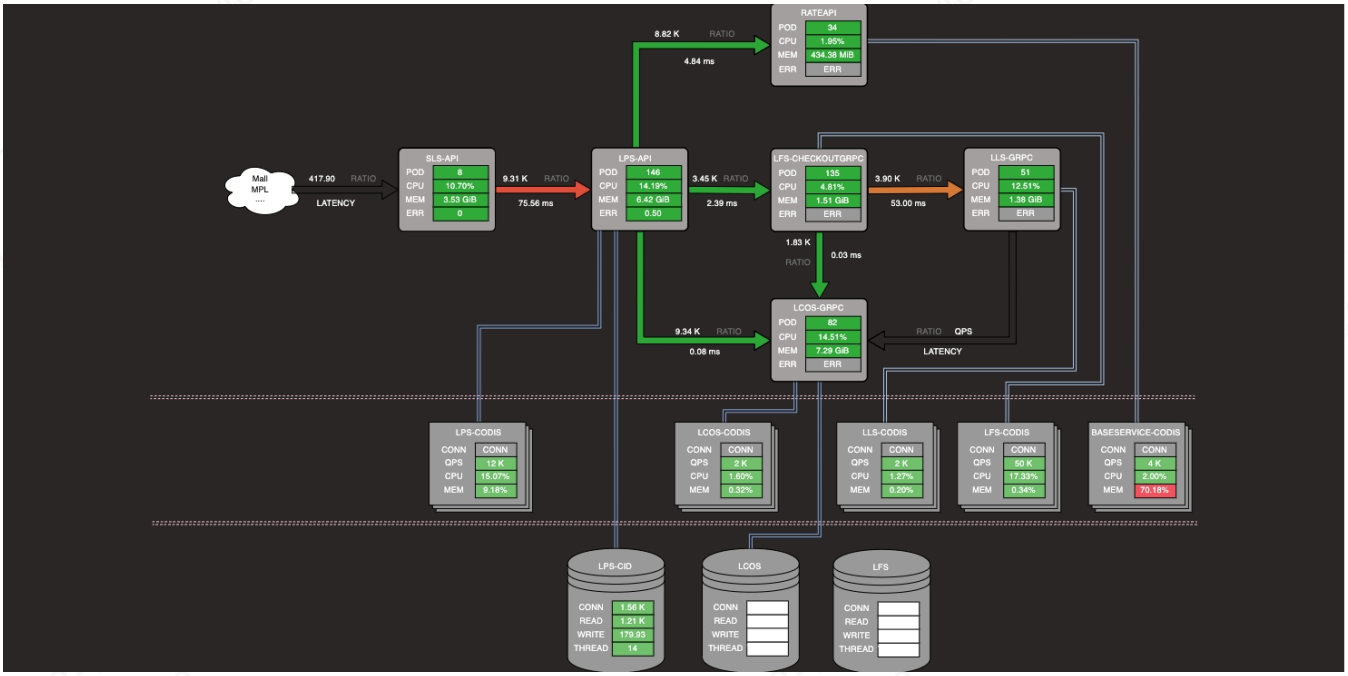
摘要

编写目的

分析故障注入的实现，为故障注入工具提供技术参考。

项目背景

随着服务拆分，系统链路变长，系统可能出错的地方也随之增加。一个常见的系统链路如下：



参考资料

混沌工程(Chaos Engineering) 总结

阿里巴巴混沌测试工具ChaosBlade

字节跳动混沌工程实践总结

混沌工程原则

Golang 常见性能问题总结

在Go中使用Failpoint注入故障

任务概述

对于故障注入，主要需要实现以下功能：

任务	描述
场景节点分析	按场景的维度进行故障注入，需分析出各服务或三方件节点，明确注入目标和故障
chassis框架提供故障注入能力	通过过chassis框架提供通用的服务级的故障注入能力
故障注入admin平台	提供可视化的故障配置和触发界面，并提供实验历史记录查看能力

规范与约定

1. 代码规范参考：代码及开发规范
2. 应用分为接口层、应用层、核心领域层和支撑层，按照CQRS原则可酌情合并应用层和核心领域层
3. 各层间定义DTO进行参数传递和数据返回
4. 数据的持久化必须在Repository中以便后续的数据层重构
5. 对外API返回采用retcode、message、data形式
6. 不允许随意向context添加参数进行隐式传递，必须经过审核阐述必要性

术语与缩略语

缩略语/术语	全称	说明
链路		系统中业务逻辑从开始到结束所需要经过的子系统组成的调用链路
故障目标		进行注入故障的目标IP节点，当前主要是pod节点

系统分析设计

系统设计目标

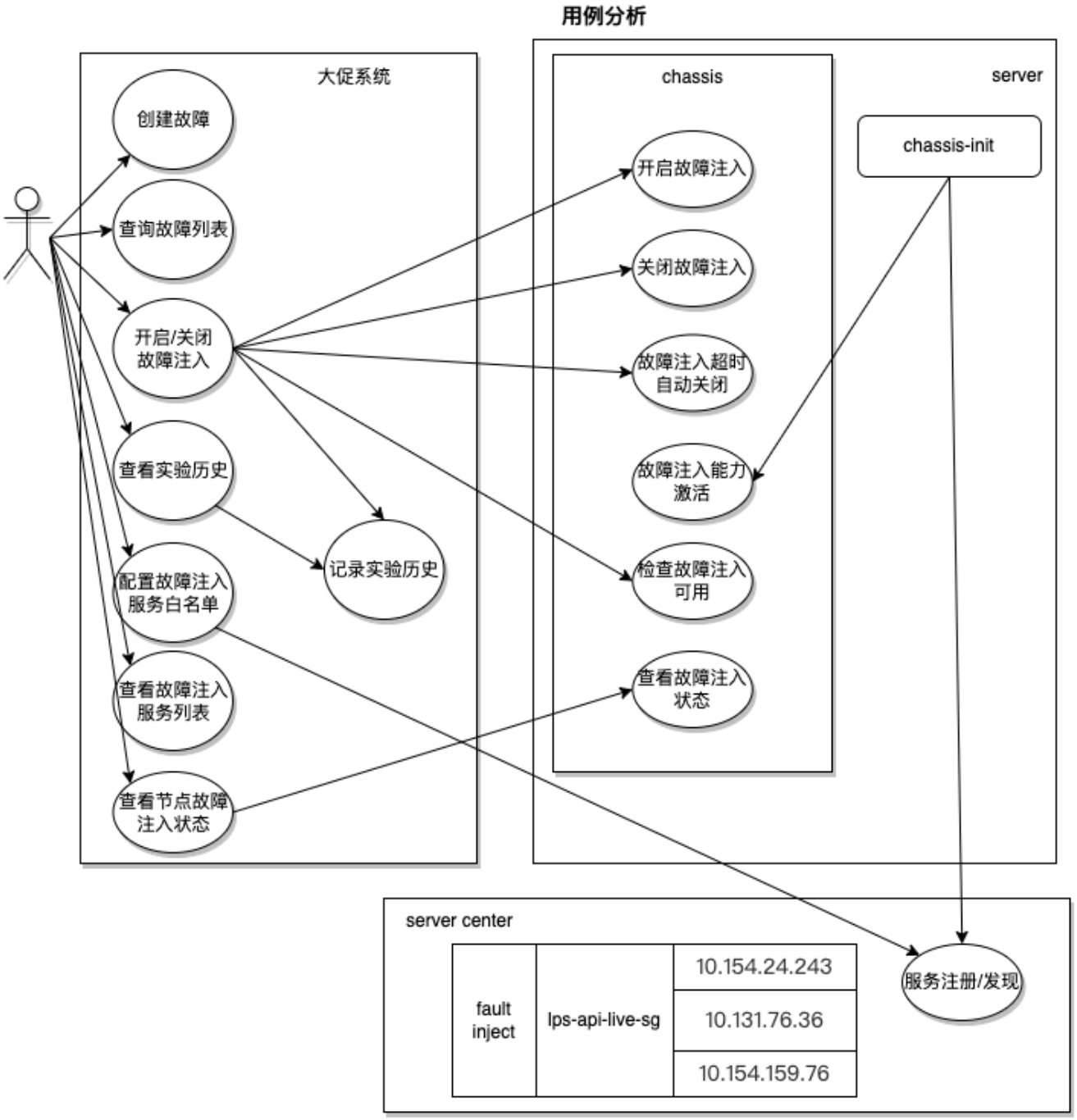
描述系统设计的目标，比如*pv/uv*，*tps*，容量，预估数据量，并发等，系统分析设计将以此作为目标展开。

总体架构分析

故障注入工具是大促系统性能分析工具的一部分，大促系统架构如下：


已有功能
建设中
待建设
可视化
链路基线
链路放大
链路时延
链路深度
核心链路
性能基线
TPS
RT
CPU
请求成功率
Mem
探测分析
强弱依赖分析
依赖合理性
降级有效性
预案有效性
监控有效性
代码分析
DeepCopy
协程泄漏
reflect反射
内存逃逸
锁
.....
基础组件使用分析
Mysql
Codis
KafKa
ES
Apollo
K8S Pod
大促走查
大促预测
流量预估
链路预估
资源预估
常规项检查
存储/读写容量
池化参数检查
索引/大key
....
中间件检查
容量/性能
灾备
告警
大促预案
降级
限流
熔断
大促压测
小规模日常压测
模拟压测
全链路压测
大促演练
剧本
故障注入
演练
模型
性能模型
TPS-CPU
TPS-T/D
TPS-Mem
流量模型
PCU-TPS
链路模型
TPS-放大TPS
沙盘模型
性能
容量
流量
预案
链路
工具
压测工具
流量采集
按比例加压
报表生成
自动压测
链路
链路采集
链路聚合
混沌工程
场景构建
故障注入
全链路压测工具
场景构造
影子链路
模型工具
机器学习
深度学习

用例分析

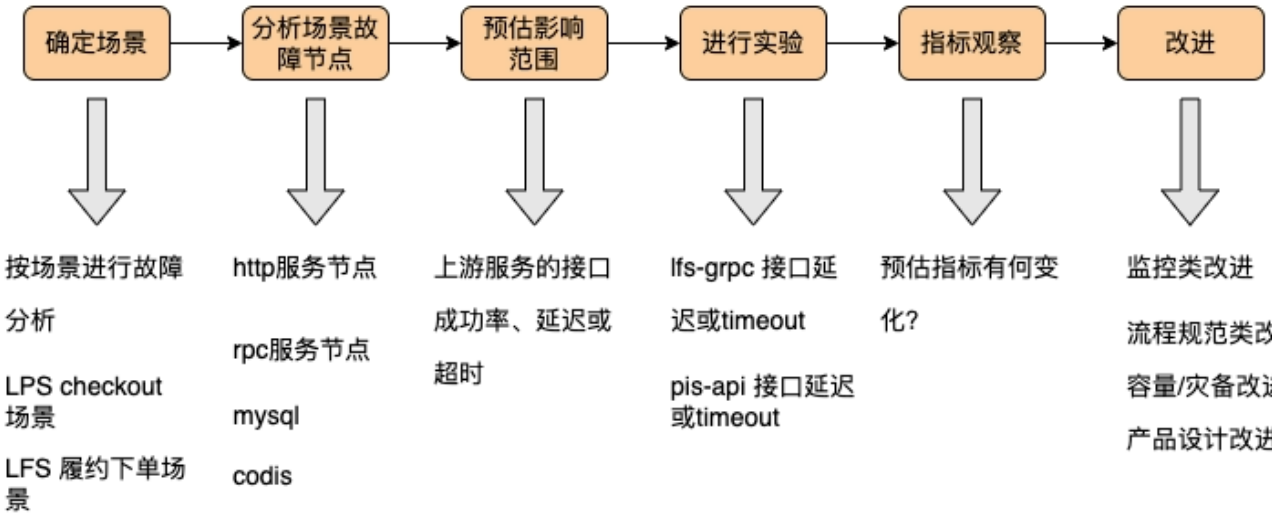


如上图所示，故障注入主要分三个功能模块：大促系统的admin模块、chassis故障注入实现模块和server center 的服务注册发现。

核心业务规则

下面具体描述故障注入全生命周期。

故障注入流程



故障维护

故障维护	故障名	lps_checkout接口延迟200ms
	目标服务	lps-api-live-sg
	故障节点类型	http_interface
	故障类型	delay
	故障对象	/api/v3/logistics/checkout/integration/
	故障比例	100
	故障默认配置	200

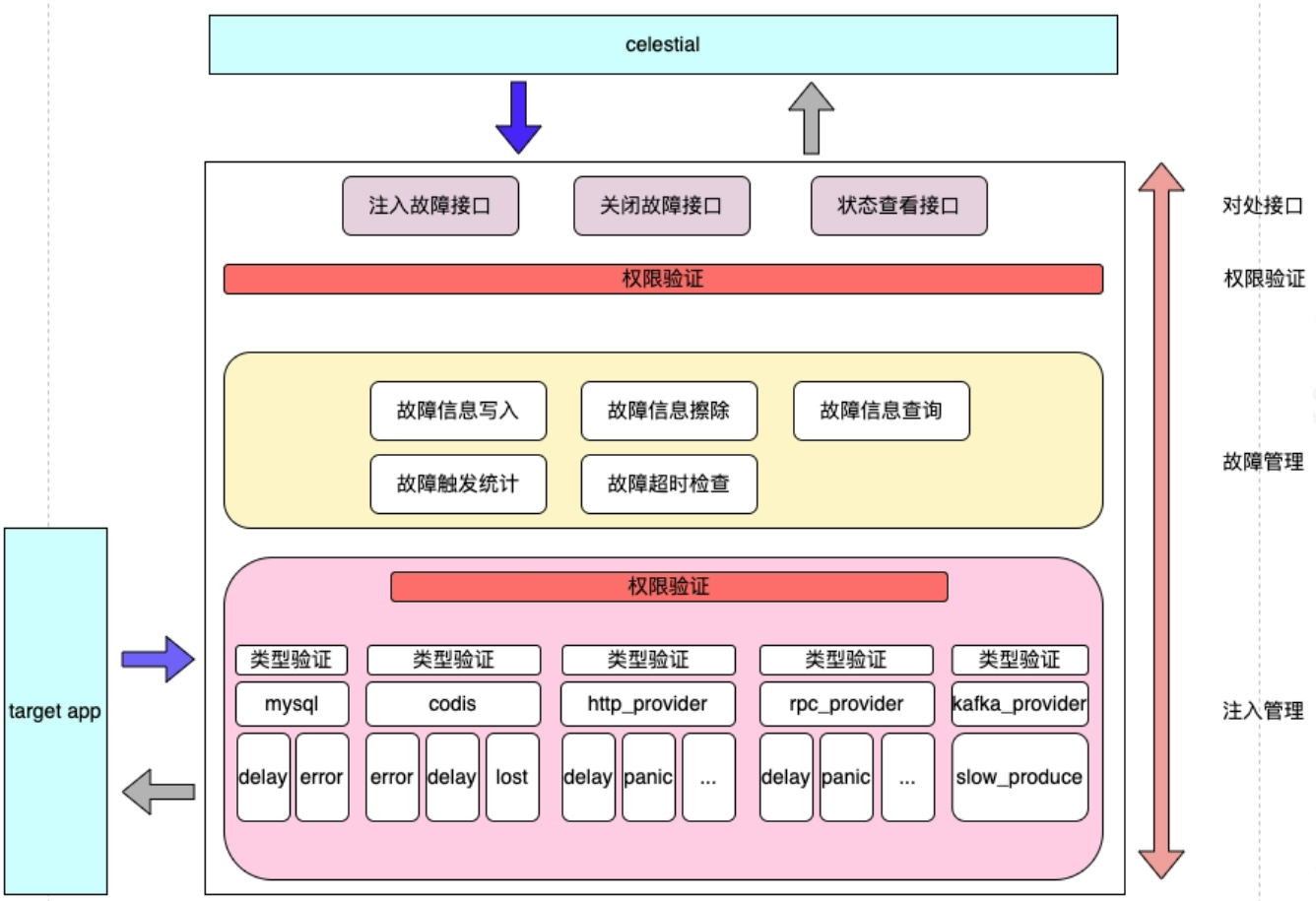
按每一个故障点作为一个对象来存储。其中”故障节点类型“、”故障类型“为预定义常量；”故障对象“根据不同的”故障节点类型“有不同的特定格式；”故障(默认)配置“根据不同的”故障类型“有不同的特定格式。

故障注入

故障注入	故障名	lps_checkout接口延迟200ms
	目标服务	lps-api-live-sg
	故障节点类型	http_interface
	故障类型	delay
	故障对象	/api/v3/logistics/checkout/integration/
	故障比例	70
	故障配置	150
	注入故障节点	10.154.24.243;10.131.76.36
	未注入故障节点	10.154.159.76
	故障开始时间	2022-07-04 16:41:31
	故障结束时间	
	状态	open

故障注入时，从维护的故障列表中选择一个故障点，可以输入自定义的“故障比例"和“故障配置”即可开启故障注入，中途支持调整“故障比例"和“故障配置”。

chassis内分模块设计如下：



目录结构设计

目录设计

chassis

--faultinject

--openapi

--faultinjectapi.go

--validate

--faultinjectvalidate.go

--faultmanagement

--faultmanagement.go

--injectmanagement

--faultinjectmysql.go

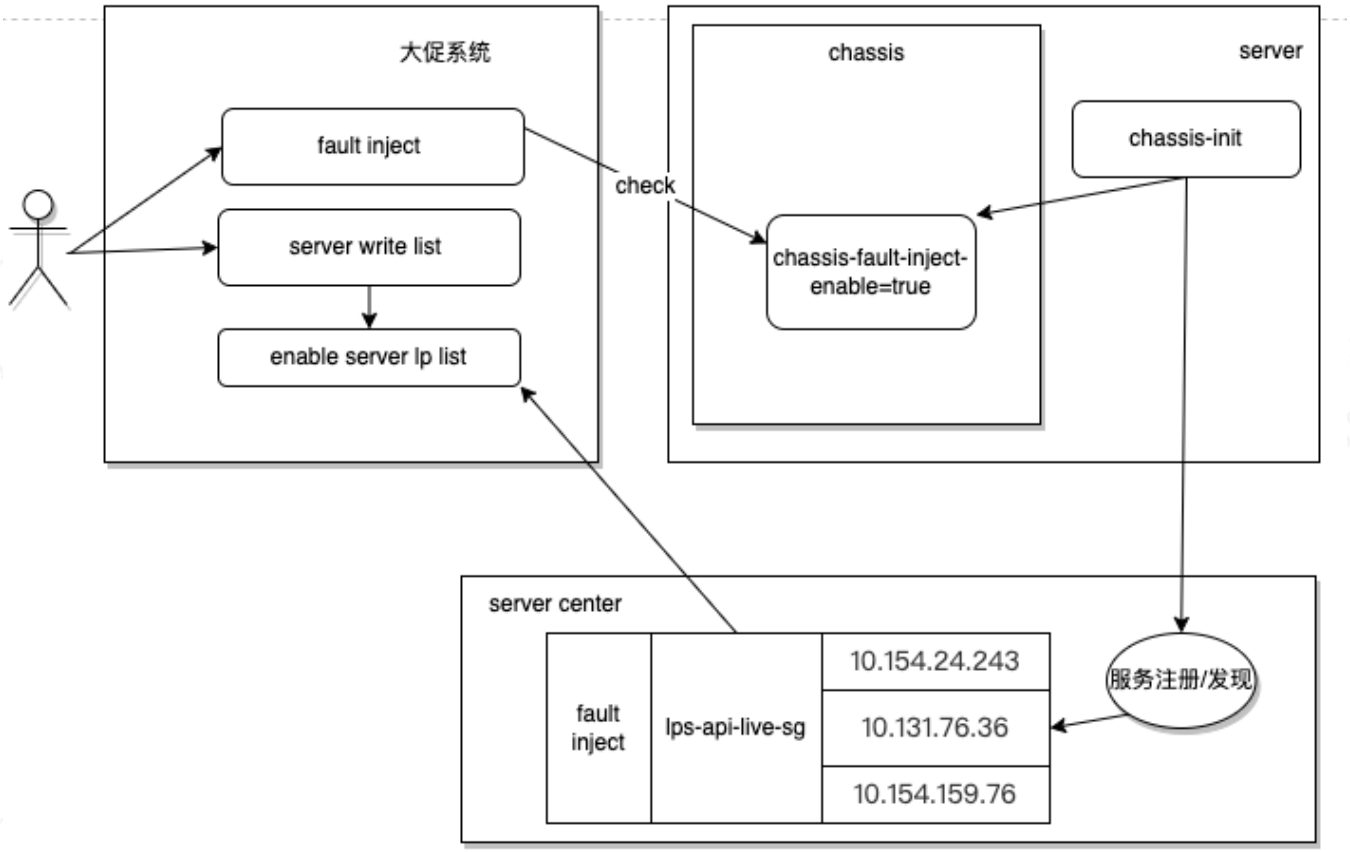
--faultinjectcodis.go

--faultinjecthttpprovider.go

...

--faultinfo.go

权限管理



故障注入对系统具有破坏性，因些必需对其权限作限制。首先，服务自身必需开启故障注入，开启故障注入的服务才能被大促系统发现；大保系统在进行故障注入时也会调用服务提供的check 接口对故障注入权限进行验证；故障注入的接口只给授权的服务(大促系统)使用。

故障盘点

http/rpc interface	延迟(delay)	细粒度控制。针对一个服务实例，可做到不同比例的请求延迟不同。例如 90% 延迟200ms；10%延迟500ms；策略了解？
	异常(exception)	网络错误;服务panic；业务错误(特定错误码)
mysql	延迟(delay)	
	异常(error)	网络错误;特定数据库error
codis	延迟(delay)	
	未命中(lost)	指定key返回获取不到信息
kafka	慢消费 (slow_consume)	消费速度慢，模拟消费能力差
	慢生产 (slow_produce)	生产速度慢，模拟网络能力差
business function	大缓存加载、业务变更	触发特定业务某一功能

具体定义如下：

故障节点类型	故障对象	故障类型	故障配置	备注
http_interface_provider/ rpc_interface_provider	请求endpoint	delay	{延迟的值(单位ms)}	endpoint=/api/v3/logistics/checkout/integration/

		business_exception	{retcode}	retcode=5003
		panic		
http_interface_consumer/ rpc_interface_consumer	请求endpoint	delay	{延迟的值(单位ms)}	endpoint=/api/v3/logistics/checkout/integration/
		net_err	{status_code}	status_code=404/503
		business_exception	{retcode}	retcode=5003
mysql	库名+表名+操作(select/update/insert/delete)	delay	{延迟的值(单位ms)}	200
		error	err_code	
codis	库名+操作 + key	delay	{延迟的值(单位ms)}	200
		error	err_code	
		lost		
kafka_consumer	topic	slow_consume	{消费速度}	
kafka_provider	topic	slow_produce	{生产速度}	

非功能特性设计

可靠性

描述系统在可靠性上的分析和设计，系统在容错性上是如何处理的，故障恢复、服务降级、熔断等的处理机制，以及在数据可靠性方面的考虑等。

可运维

描述在提升系统运维方面的分析及设计，好的系统应该是尽可能自动化的完成业务，自身应该具备相当能力的容错处理，不需要运维人员介入处理；同时也尽可能是标准化的，方便运维人员部署，或者提供可视化的运维操作。

1. 故障注入能力为chassis能力的一种，随应用服务部署运维。

安全性

配合安全要求，阐述系统安全方面（例如XSS，SQL注入，DDOS，数据安全等）的设计。

1. 采用livetest环境来模拟live环境，不影响生产;
2. 故障注入工具需要手动激活，并且在大促系统配置白名单方可使用；
3. 每个注入的故障都有相应的终止操作，并且有超时自动终止；
4. 同步live数据做为实验请求数据，给数据添加故障注入的标识，以防数据意外发送其他live服务无法识别。

可测试性

描述系统各个业务在可测试性上的分析，发布应该满足灰度要求，系统应该具备在生产环境的可测试性。必须等到某个时间点、修改操作系统时间，或者直接不可测，都不应该发生。

1. 每一项故障注入都可实现；
2. 每一项稳态指标都有监控或告警；
3. 实验环境选择在livetest环境，apollo中的服务配置信息可通过开关同步live环境的配置。测试数据为live 的数据回放。

监控

1. 当前各项目都有相应监控告警配置。

其他

接口

chassis 提供的接口	故障注入(支持超时自动关闭)

	关闭故障注入
	查看故障注入状态
	检查故障注入是否可用(或是根据故障注入接口的调用结果决定？)
server center 提供的接口	可注入故障服务发现
大促系统提供的接口	创建故障
	查询定义的故障列表
	查询定义的故障详情
	开启故障注入
	关闭故障注入
	查看实验历史
	配置故障注入白名单
	查看发现的故障注入服务列表
	查看节点故障注入状态

注入实现排期

一期

故障节点类型	故障对象	故障类型	故障配置	备注
http_interface_provider/ rpc_interface_provider	请求endpoint	delay	{延迟的值(单位ms)}	
		panic		
rpc_interface_consumer	请求endpoint	delay	{延迟的值(单位ms)}	同http_interface_*
		net_err	{status_code}	支持： connection refused connect timeout no such host
mysql	库名+表名+操作(select/update/insert/delete)	delay	{延迟的值(单位ms)}	
		error	err_code	支持： 未知 MySQL 错误 不能创建 TCP/IP 接字 查询过程中丢失了与 SQL 服务器的连接 未知的 MySQL 服务器主机 等常见的至少3种错误

二期

故障节点类型	故障对象	故障类型	故障配置	备注
http_interface_provider/ rpc_interface_provider	请求endpoint	business_exception	{retcode}	业务需标准化返回信息
http_interface_consumer/ rpc_interface_consumer	请求endpoint	business_exception	{retcode}	业务需标准化返回信息
http_interface_consumer	请求endpoint	delay	{延迟的值(单位ms)}	同http_interface_*
		net_err	{status_code}	支持： connection refused connect timeout no such host
codis	库名+操作+key	delay	{延迟的值(单位ms)}	
		error	err_code	
		lost		

kafka_consumer	topic	slow_consume	{消费速度}	
kafka_provider	topic	slow_produce	{生产速度}	

Q&A

Q：一个实例上，是否支持同时注入多种故障
A：按理说为了精确定位到问题，每次实验时都注入一种故障。原故障未关闭但有新故障注入时，直接覆盖原故障。每次仅有一种故障被注入
Q：如何知道某一实例上是否有被故障注入
A：可以通过故障注入查看状态的接口获取当前实例被注入了何种故障或没被注入故障