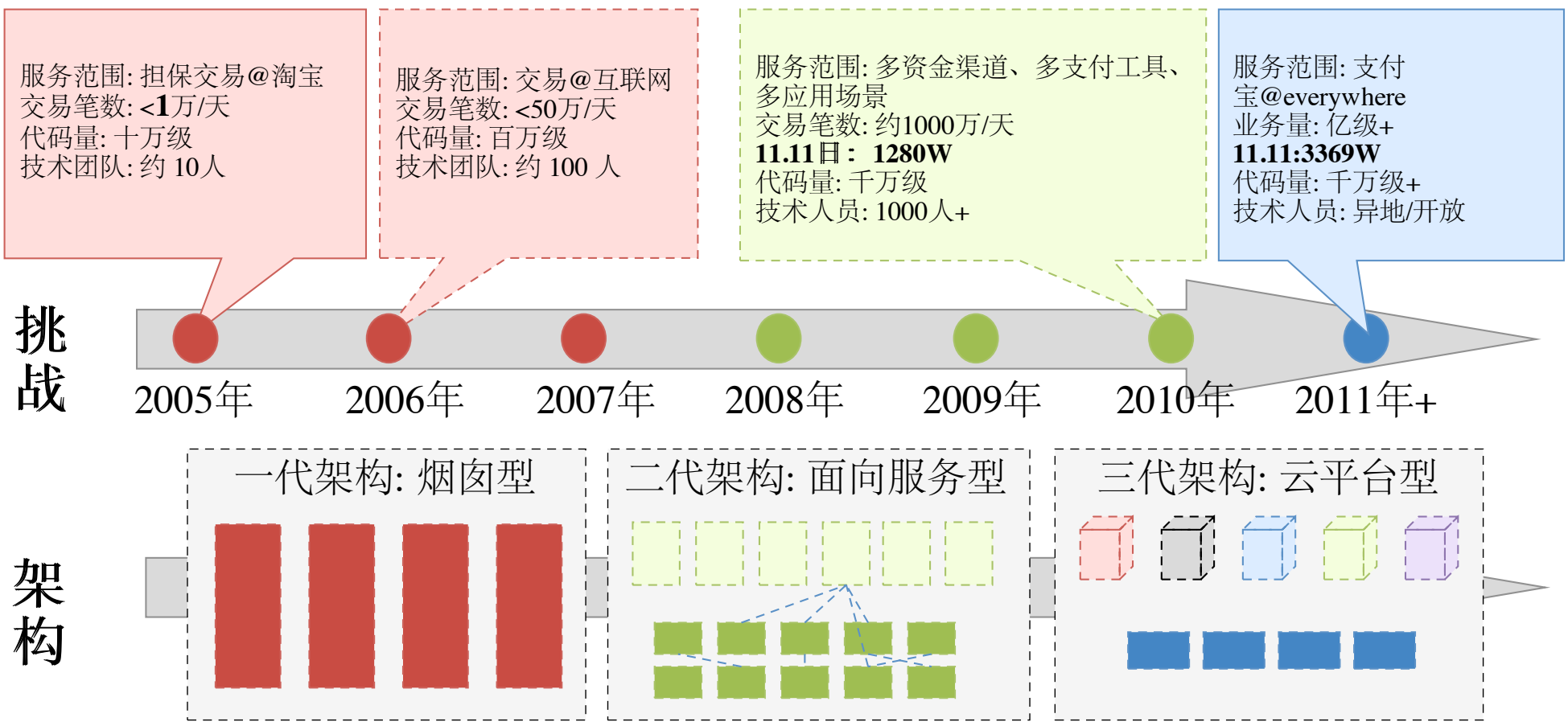
支付宝弹性计算架构

阿玺
支付宝-技术部
Mail: xi.hux@alipay.com
新浪微博: 支付宝_阿玺

个人介绍

- 胡喜，花名阿玺，2007年加入支付宝，主持支付平台基础技术的架构设计与研发工作，并且参与支付宝核心支付平台的架构设计和系统升级。

支付宝系统发展历程


服务范围: 担保交易@淘宝
交易笔数: <1万/天
代码量: 十万级
技术团队: 约 10人
服务范围: 交易@互联网
交易笔数: <50万/天
代码量: 百万级
技术团队: 约 100 人
服务范围: 多资金渠道、多支付工具、多应用场景
交易笔数: 约1000万/天
11.11日： 1280W
代码量: 千万级
技术人员: 1000人+
服务范围: 支付宝@everywhere
业务量: 亿级+
11.11:3369W
代码量: 千万级+
技术人员: 异地/开放
挑战
2005年
2006年
2007年
2008年
2009年
2010年
2011年+
架构
一代架构: 烟囱型
二代架构: 面向服务型
三代架构: 云平台型

2012.11.11系统必须具备交易处理能力：

1亿+

80亿+数据库事务
500亿+的SQL执行
1000亿+服务调用
500+个应用协同完成

我们需要什么样的架构

底层计算资源（IAAS）做到弹性是否满足?

应用层面如何做到可伸缩性?

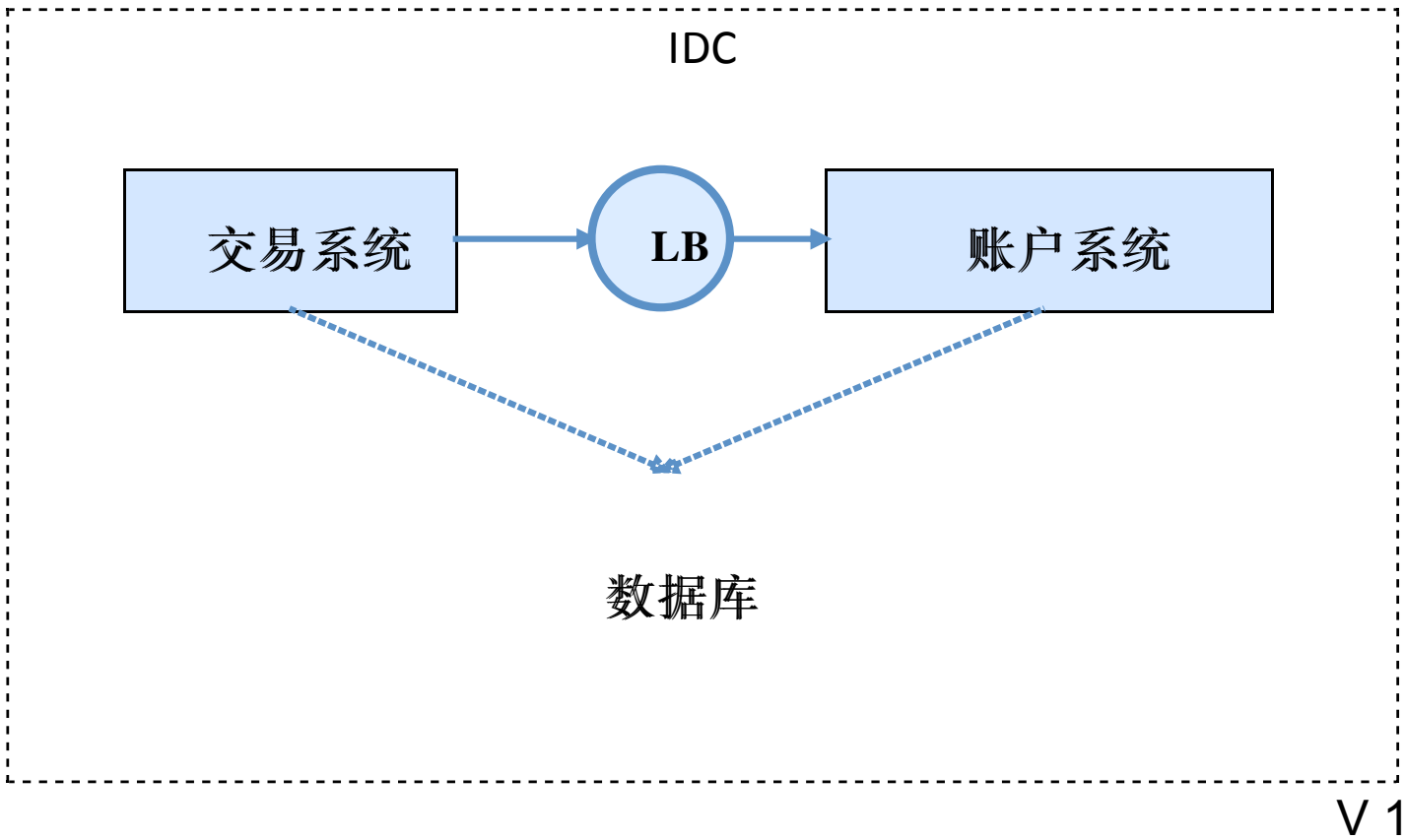
出现故障后是否能够做到快速恢复?

这一切是否能够做到自动化控制?

Agenda

- 可伸缩性：提升容量百万级到亿级
- 故障容忍：99.9%到99.99%+
- 弹性控制：人工控制到秒级自动调度

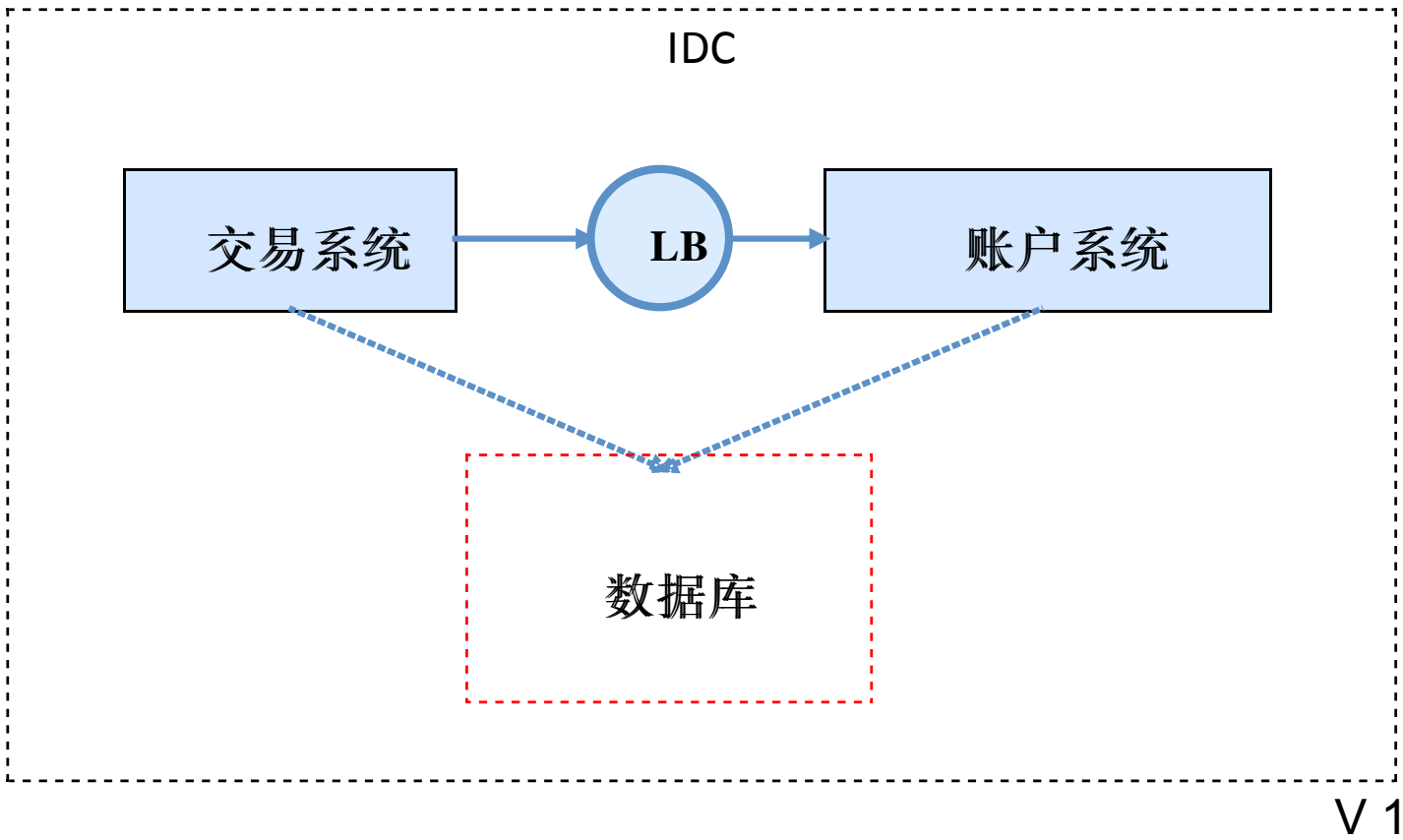
一个简化的支付宝系统模型



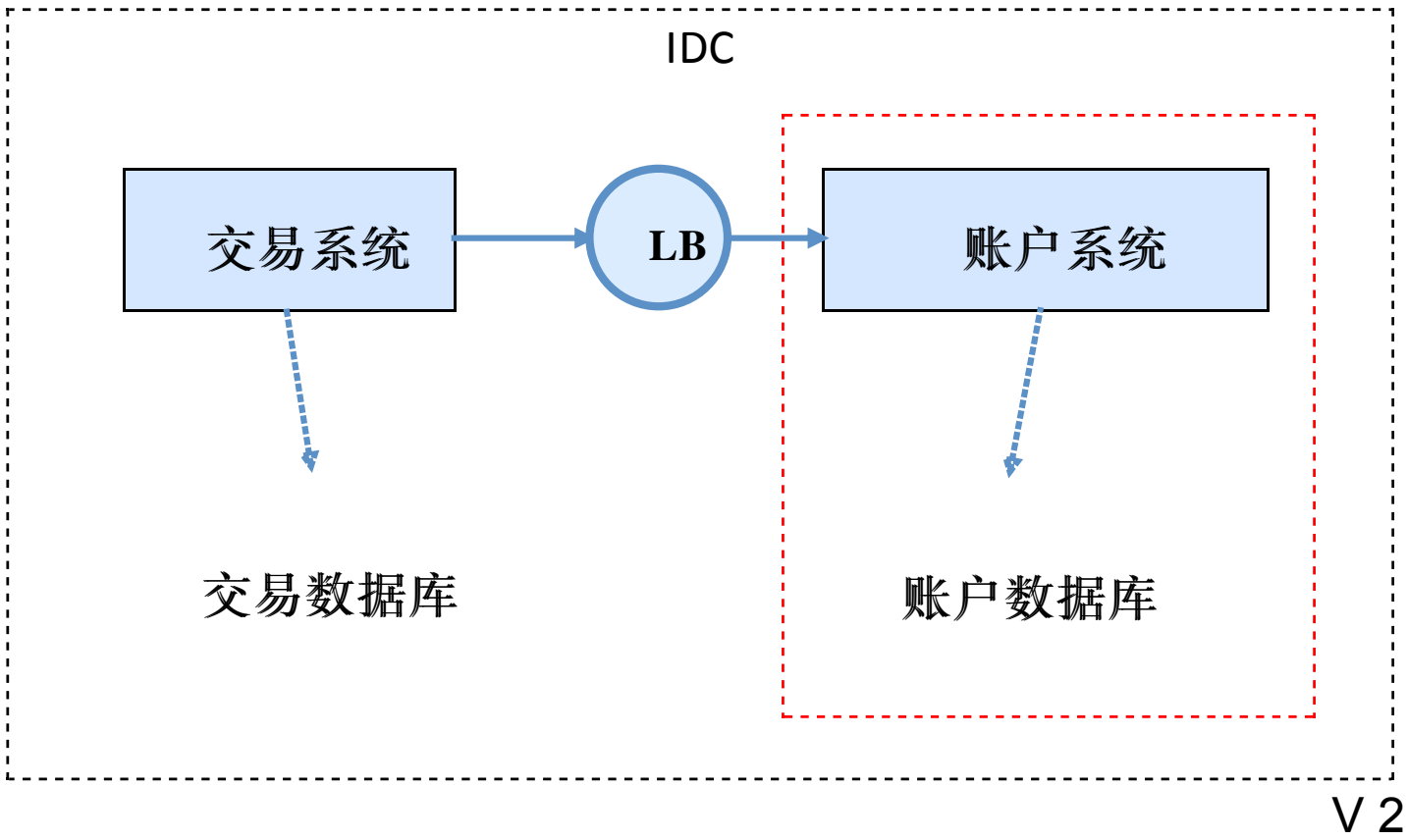
提升容量百万级到亿级

可伸缩

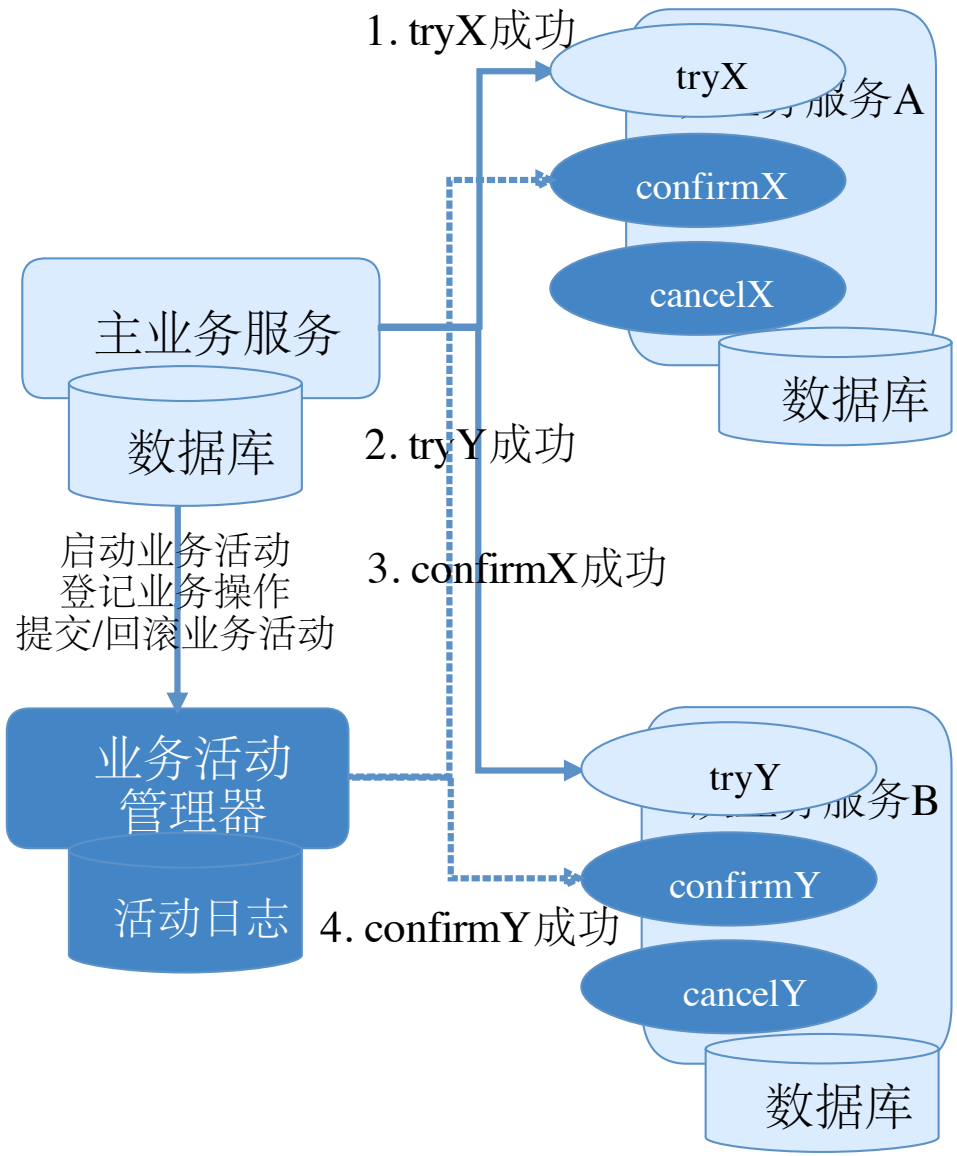
数据库的瓶颈



一致性瓶颈



业务一致性：service层的分布事务



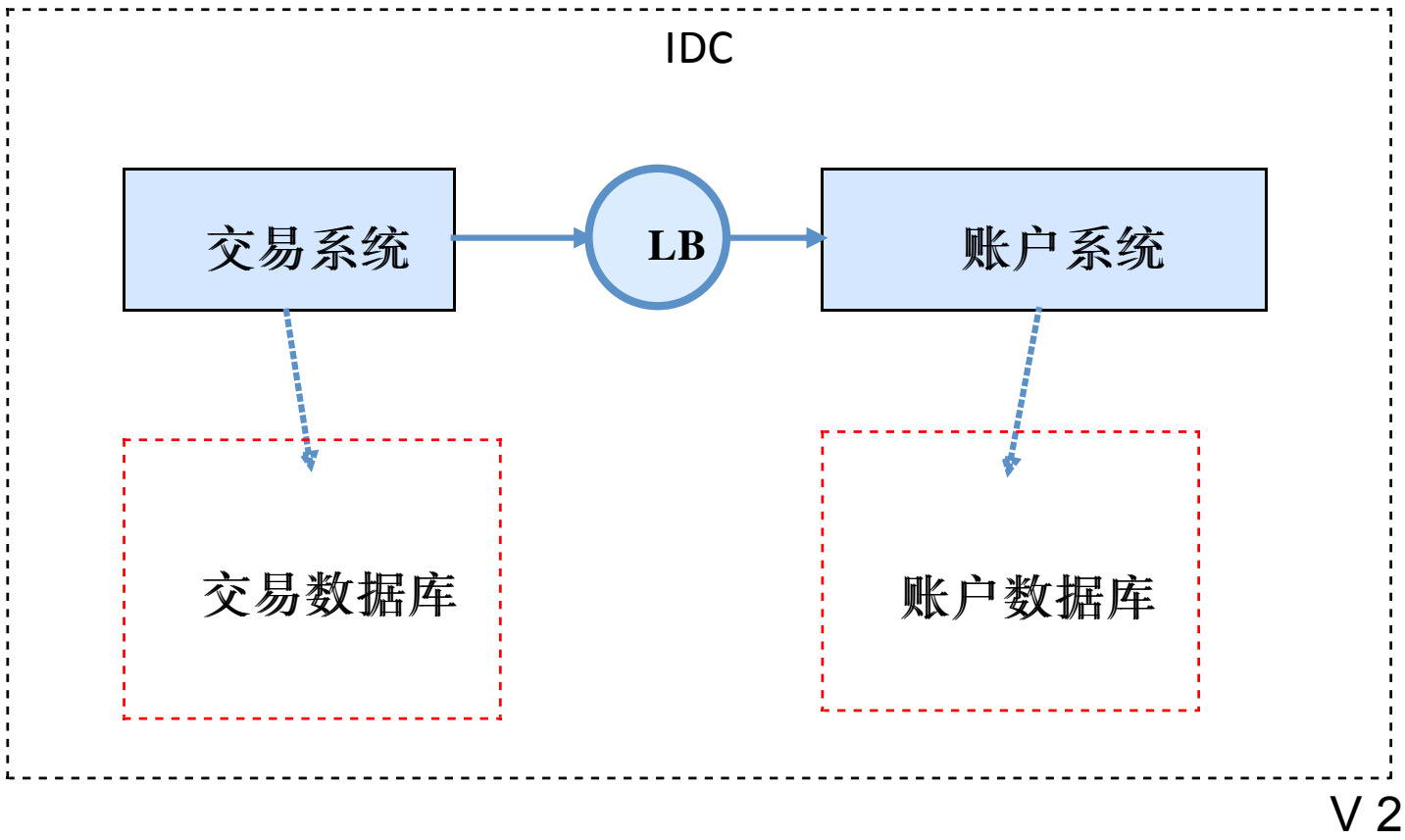
实现

- 一个完整的业务活动由一个主业务服务与若干从业务服务组成
- 主业务服务负责发起并完成整个业务活动
- 从业务服务提供TCC型业务操作
- 业务活动管理器控制业务活动的一致性，它登记业务活动中的操作，并在业务活动提交时确认所有的TCC型操作的confirm操作，在业务活动取消时调用所有TCC型操作的cancel操作

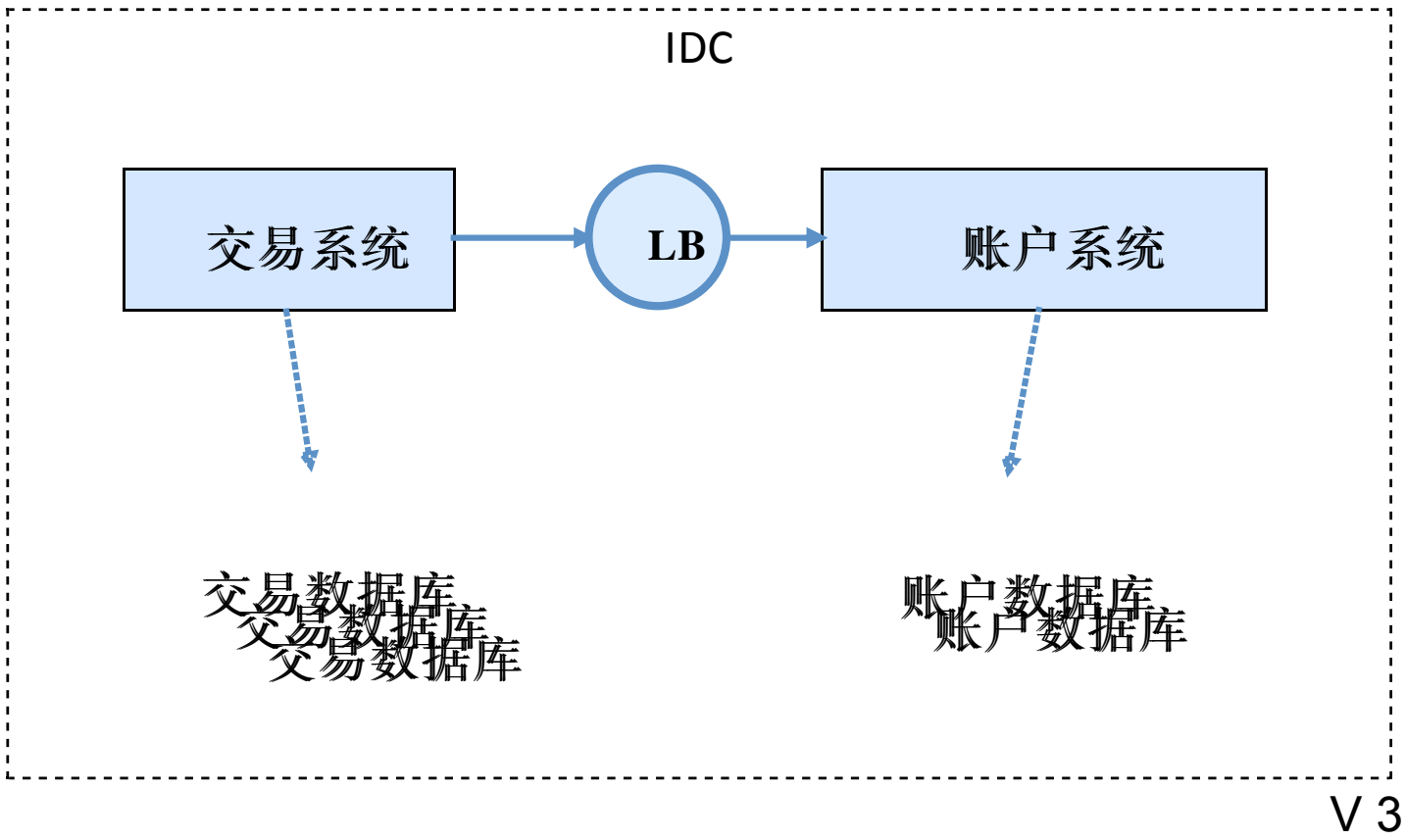
与2PC协议比较

- 没有单独的Prepare阶段，降低协议成本
- 系统故障容忍度高，恢复简单

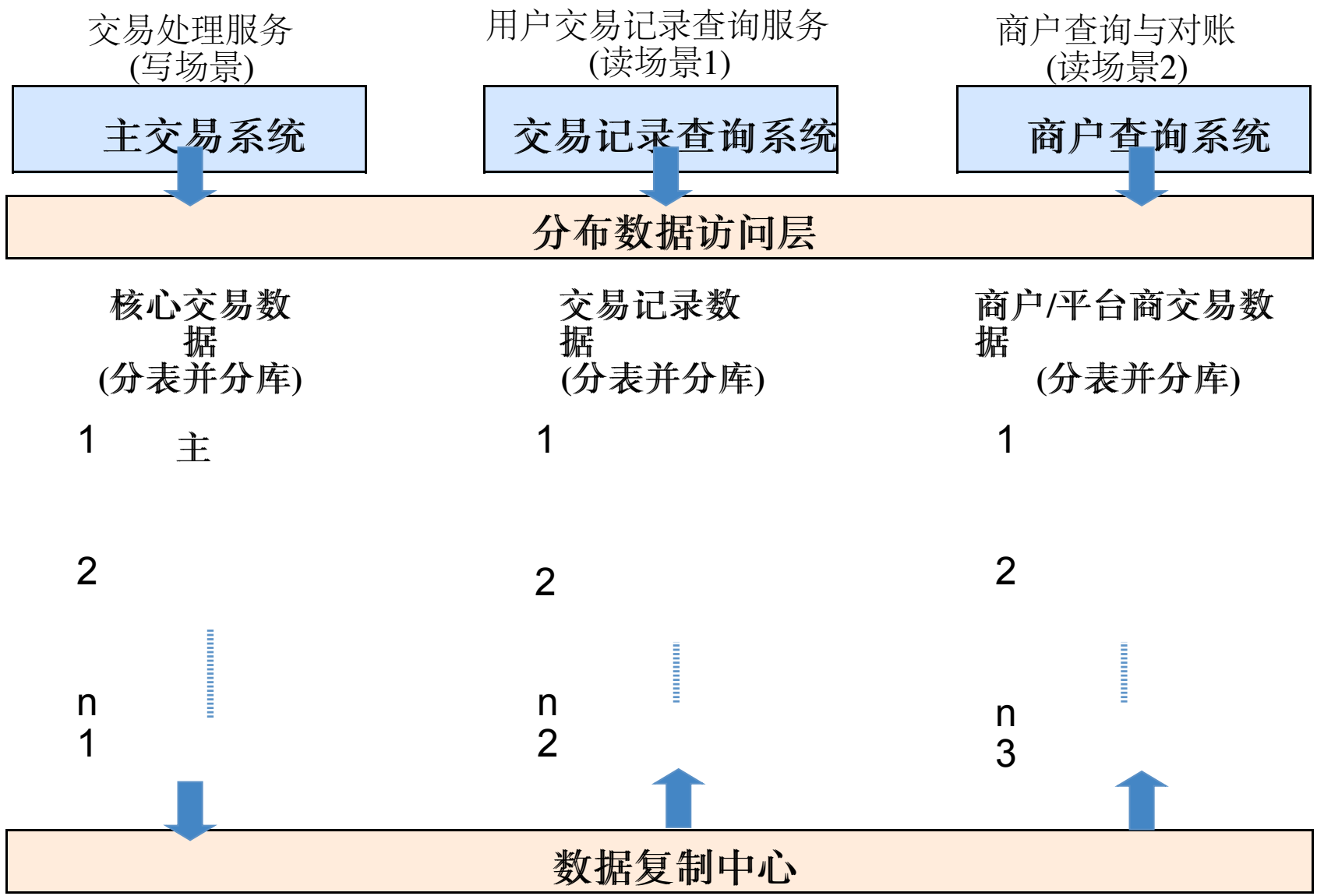
单个库的瓶颈



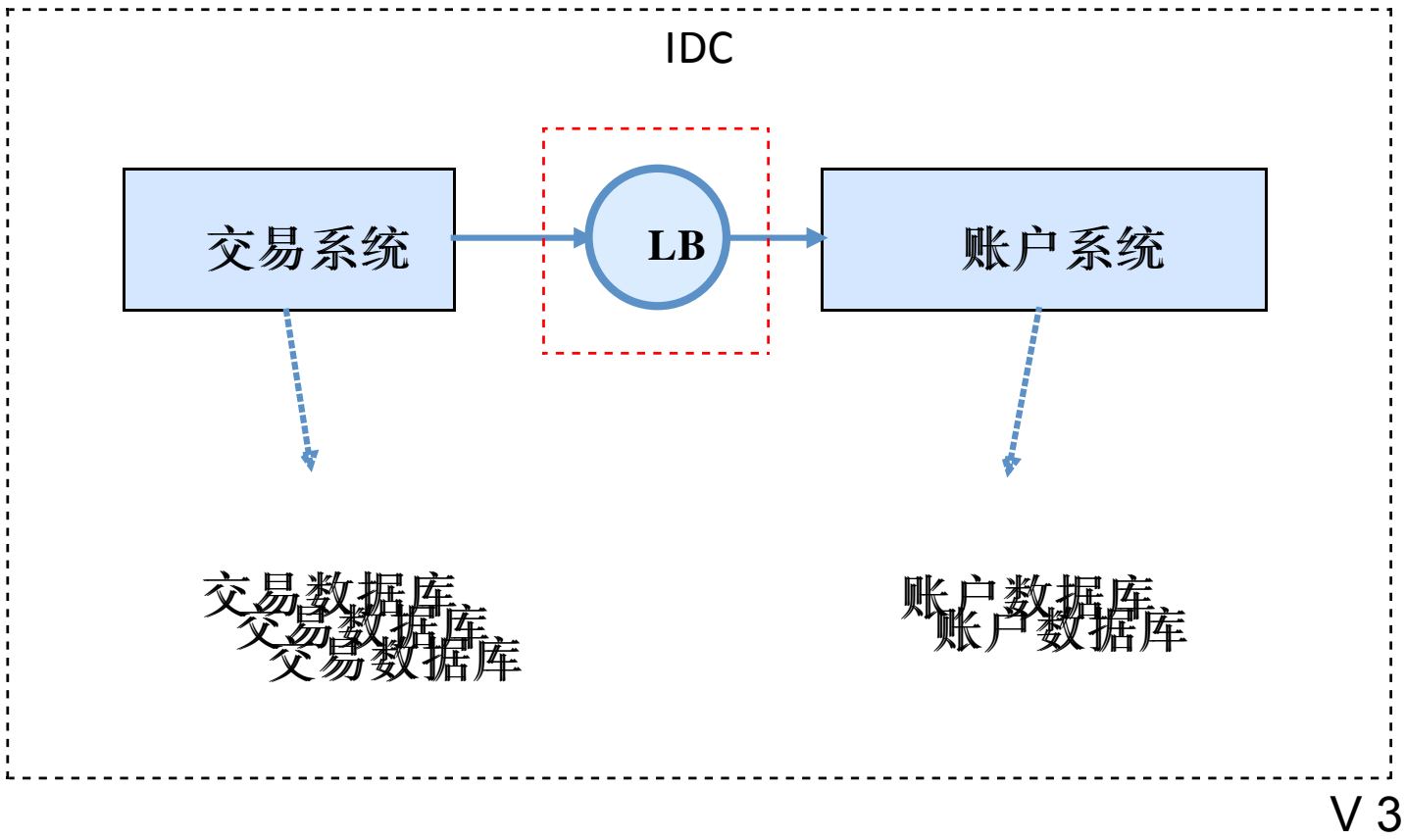
数据可伸缩性：数据水平拆分与复制



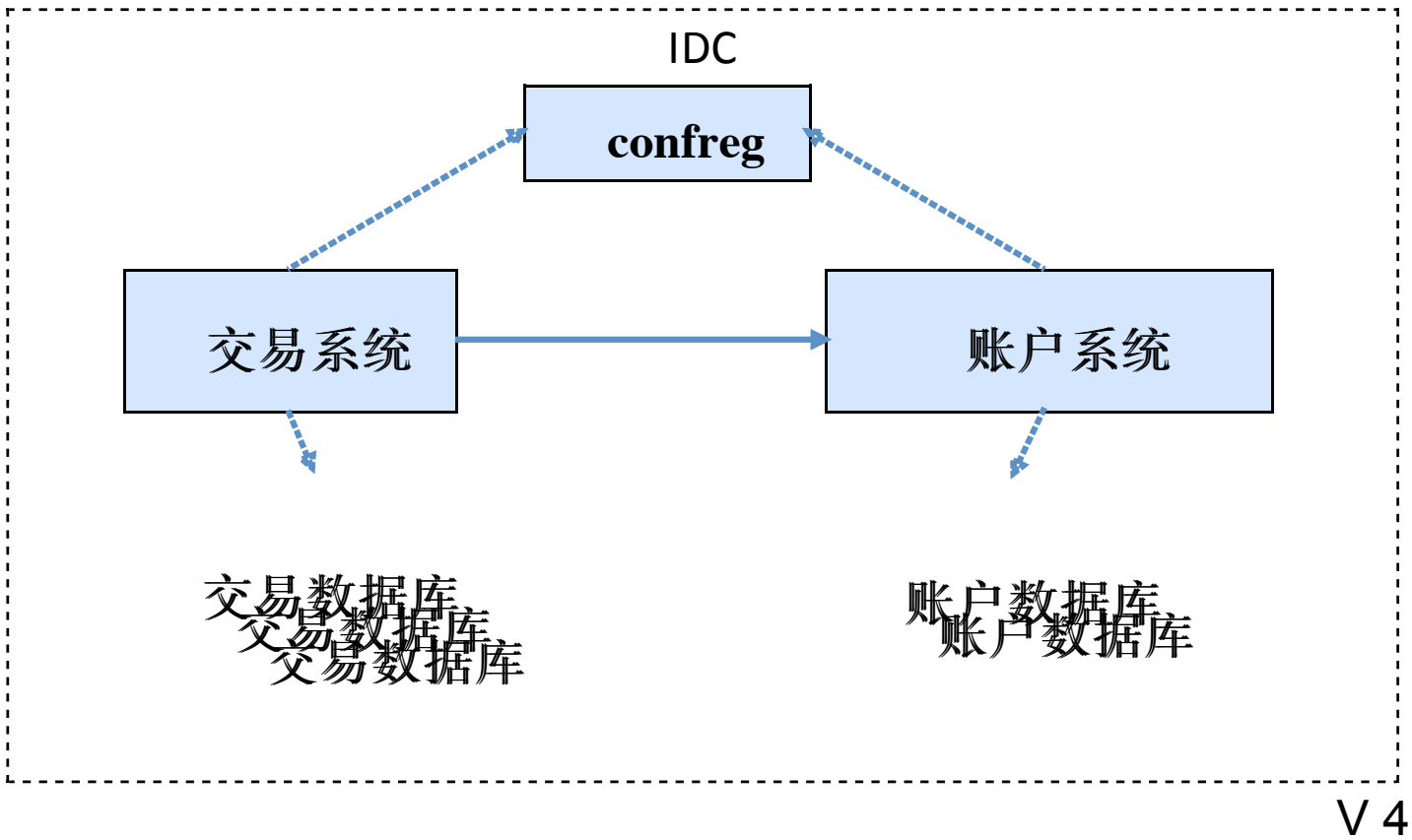
数据可伸缩性：交易数据拆分



网络伸缩瓶颈

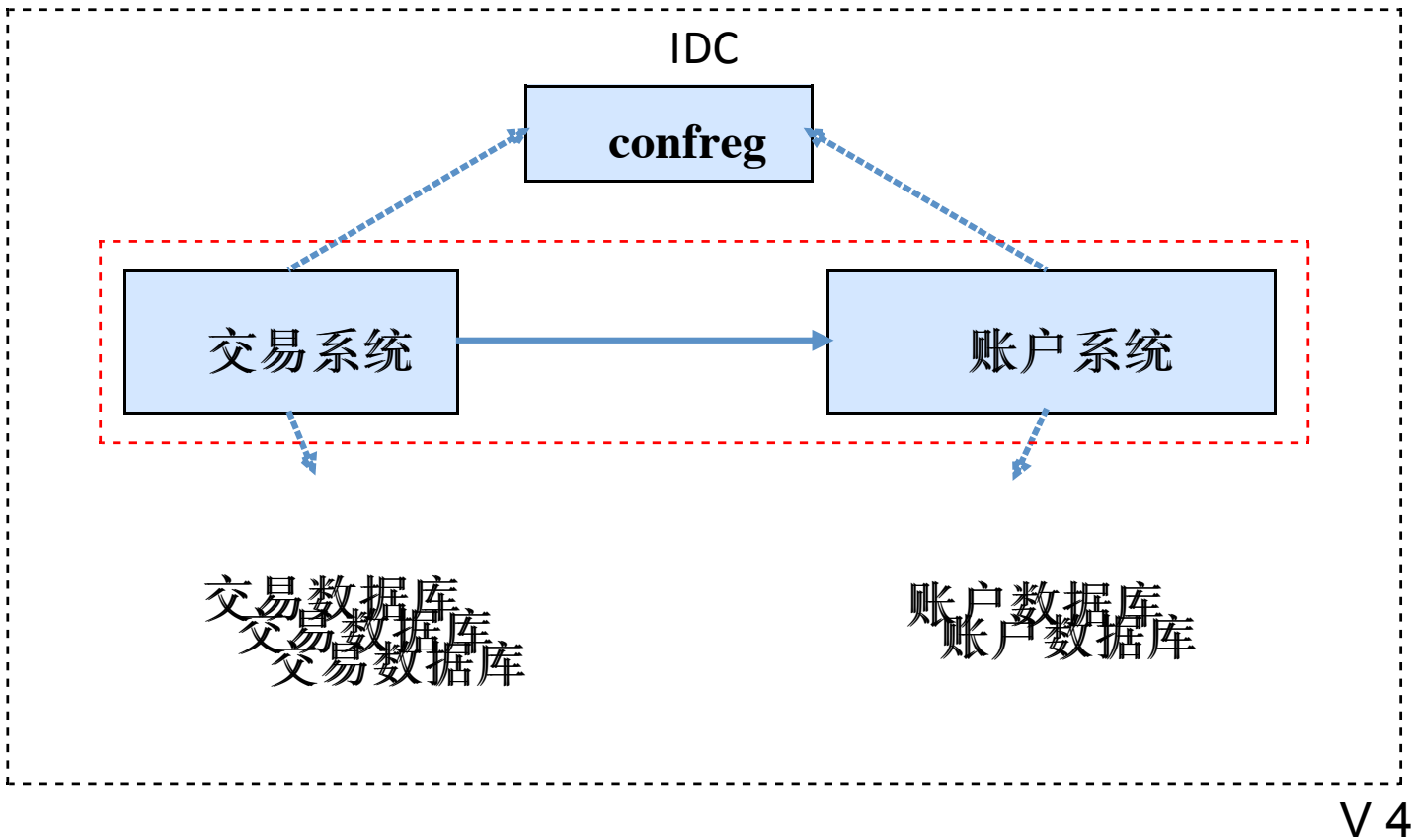


网络可伸缩性：消除网络设备瓶颈



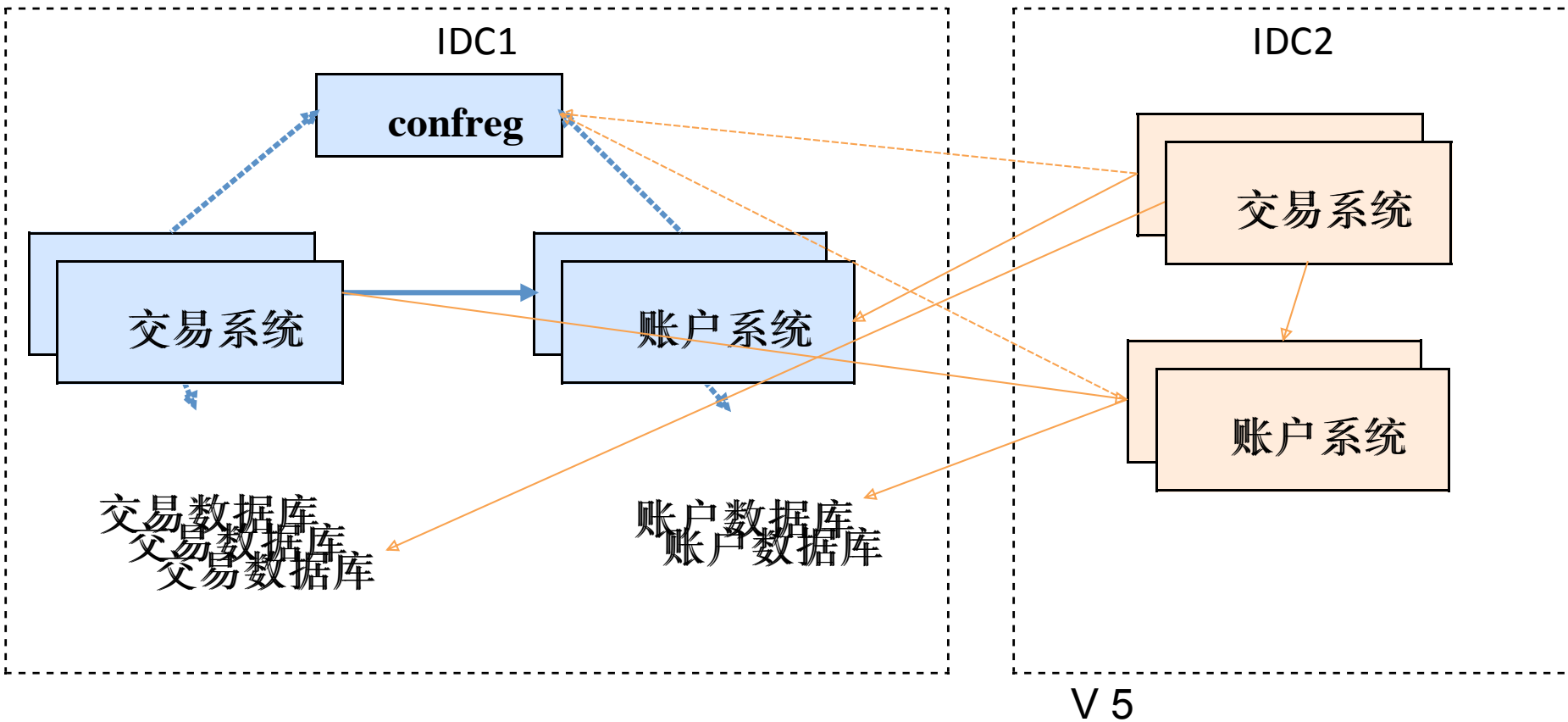
V 4

服务器伸缩瓶颈



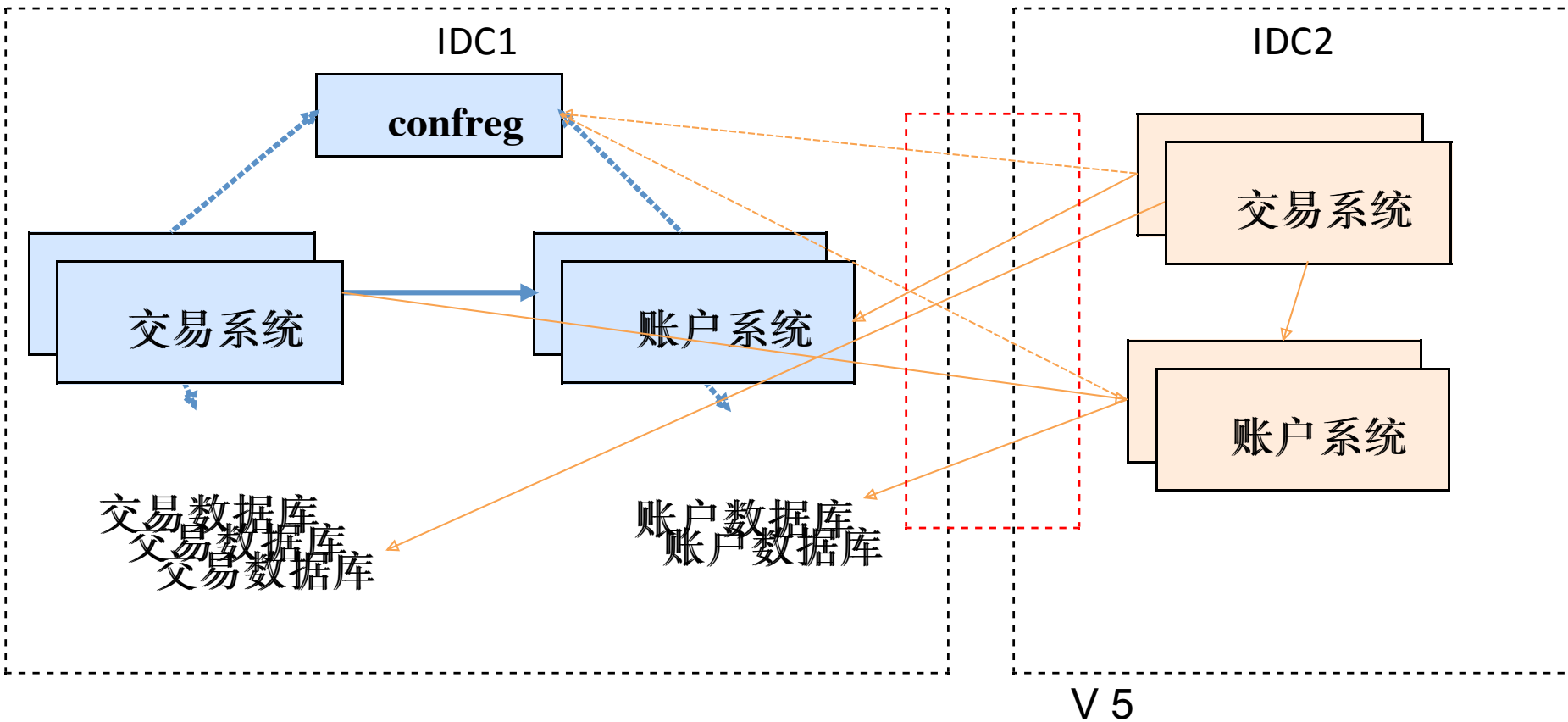
V 4

服务器伸缩：服务器扩展到多个IDC

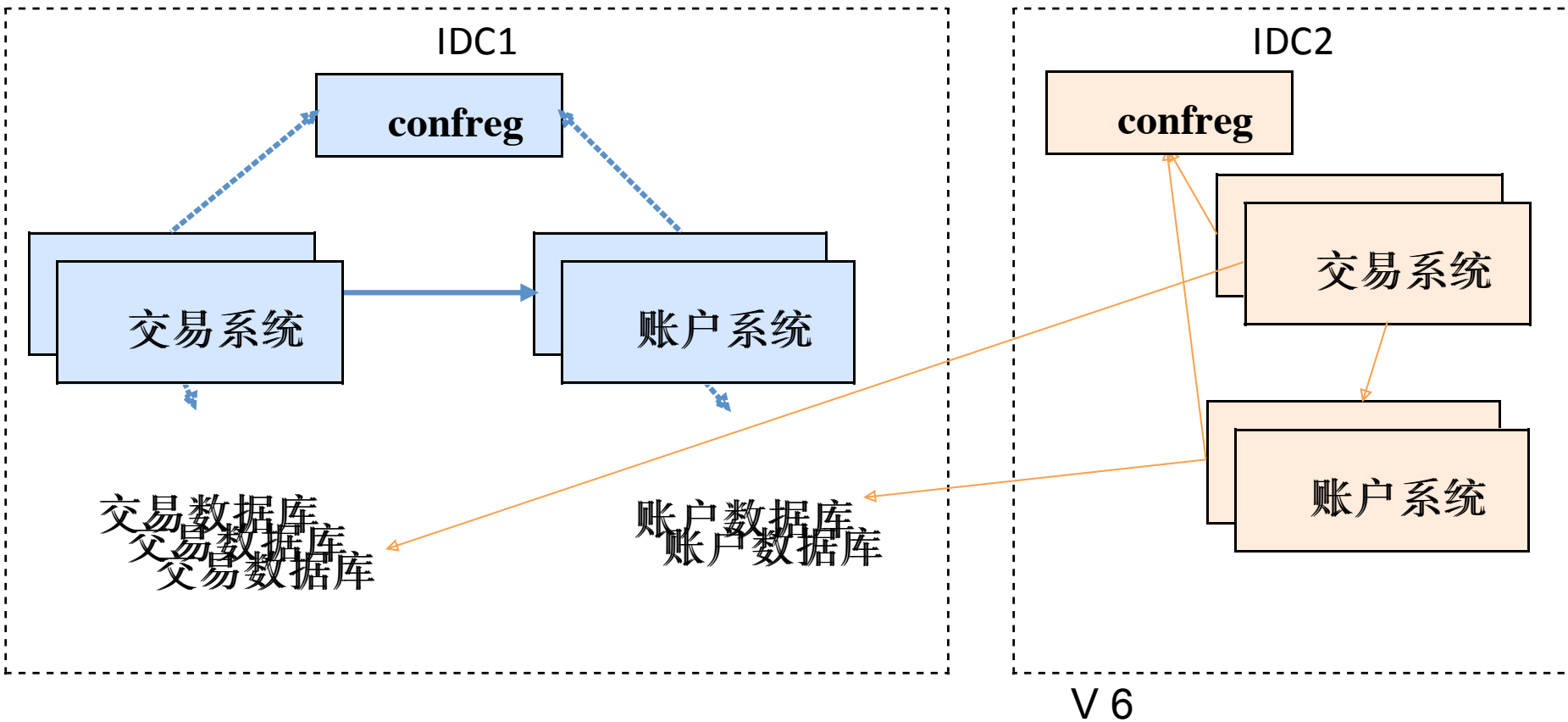


V 5

跨机房通讯的瓶颈


IDC1
IDC2
confreg
交易系统
账户系统
交易系统
账户系统
交易数据库
交易数据库
交易数据库
账户数据库
账户数据库
V 5

IDC伸缩：部分独立IDC

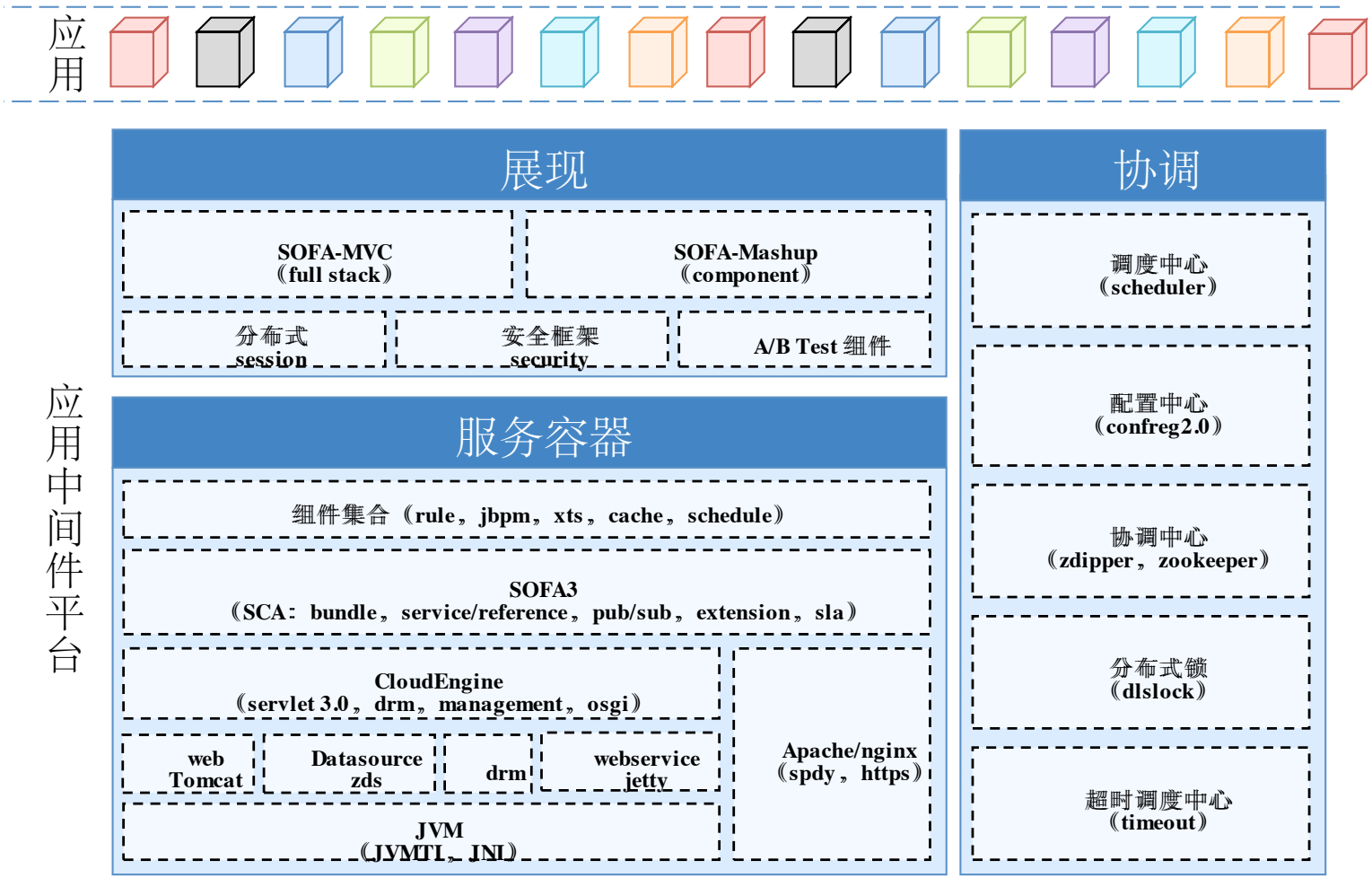


V 6

小结：提升容量百万级到亿级

- 数据的可伸缩性
  - ✓ 垂直，水平拆分，复制，分布式事务
- 网络可伸缩性
- IDC可伸缩性

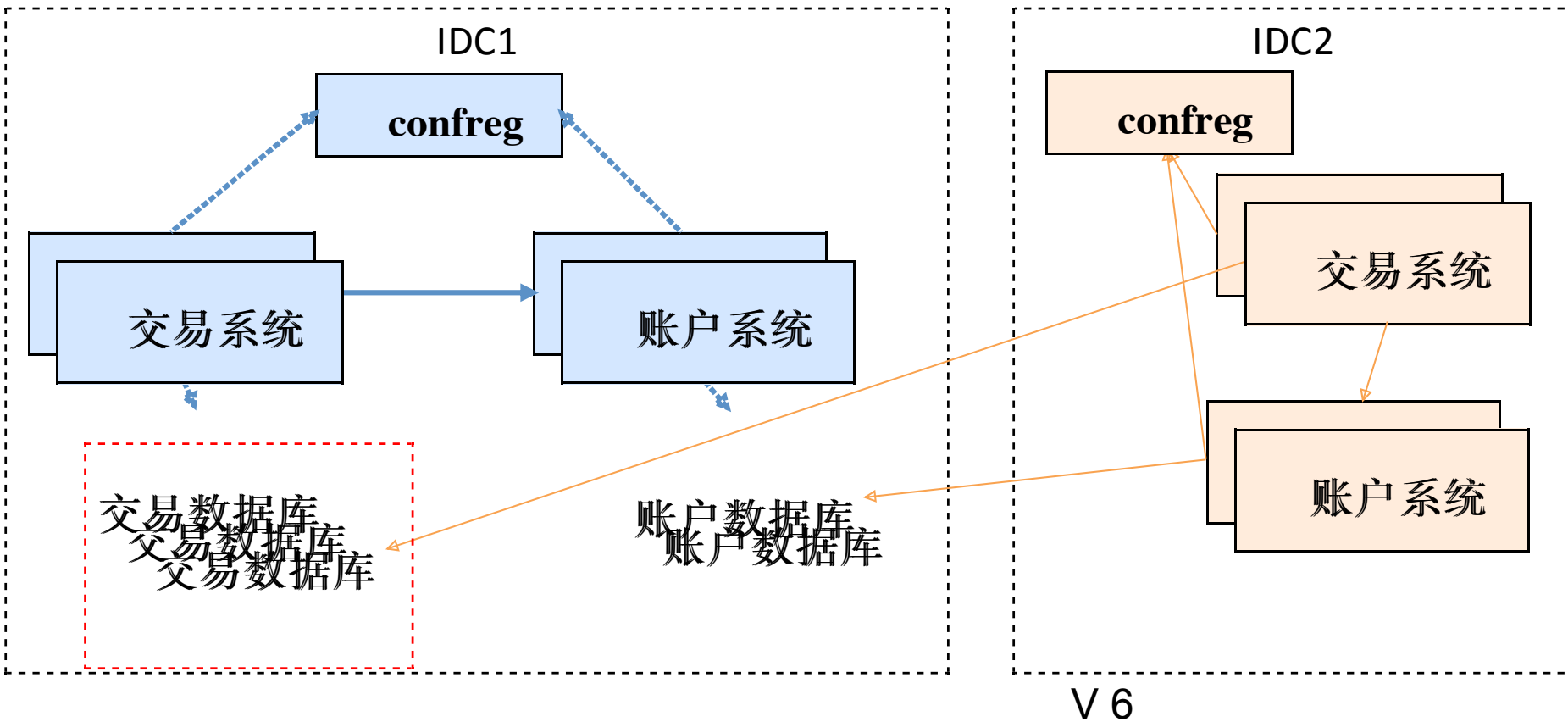
应用中间件技术架构


应用
展现
SOFA-MVC
《full stack》
SOFA-Mashup
《component》
分布式
session
安全框架
security
A/B Test 组件
协调
调度中心
《scheduler》
配置中心
《confreg2.0》
协调中心
《zdipper，zookeeper》
分布式锁
《dlslock》
超时调度中心
《timeout》
应用中间件平台
服务容器
组件集合《rule，jbpm，xts，cache，schedule》
SOFA3
《SCA：bundle，service/reference，pub/sub，extension，sla》
CloudEngine
《servlet 3.0，drm，management，osgi》
web
Tomcat
Datasource
zds
drm
webservice
jetty
JVM
《JVMTI，JNI》
Apache/nginx
《spdy，https》

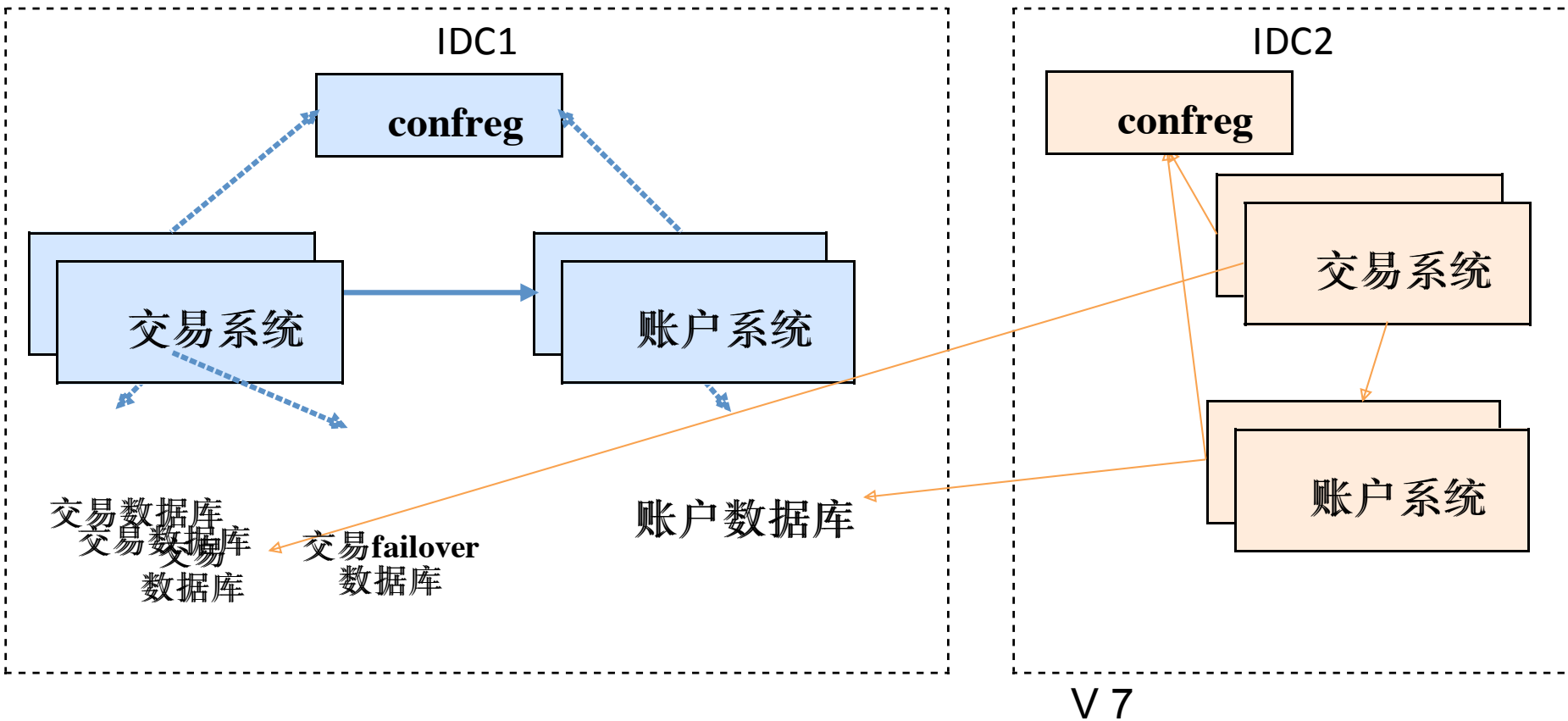
99.9%到99.99%+

故障容忍

数据库单点故障

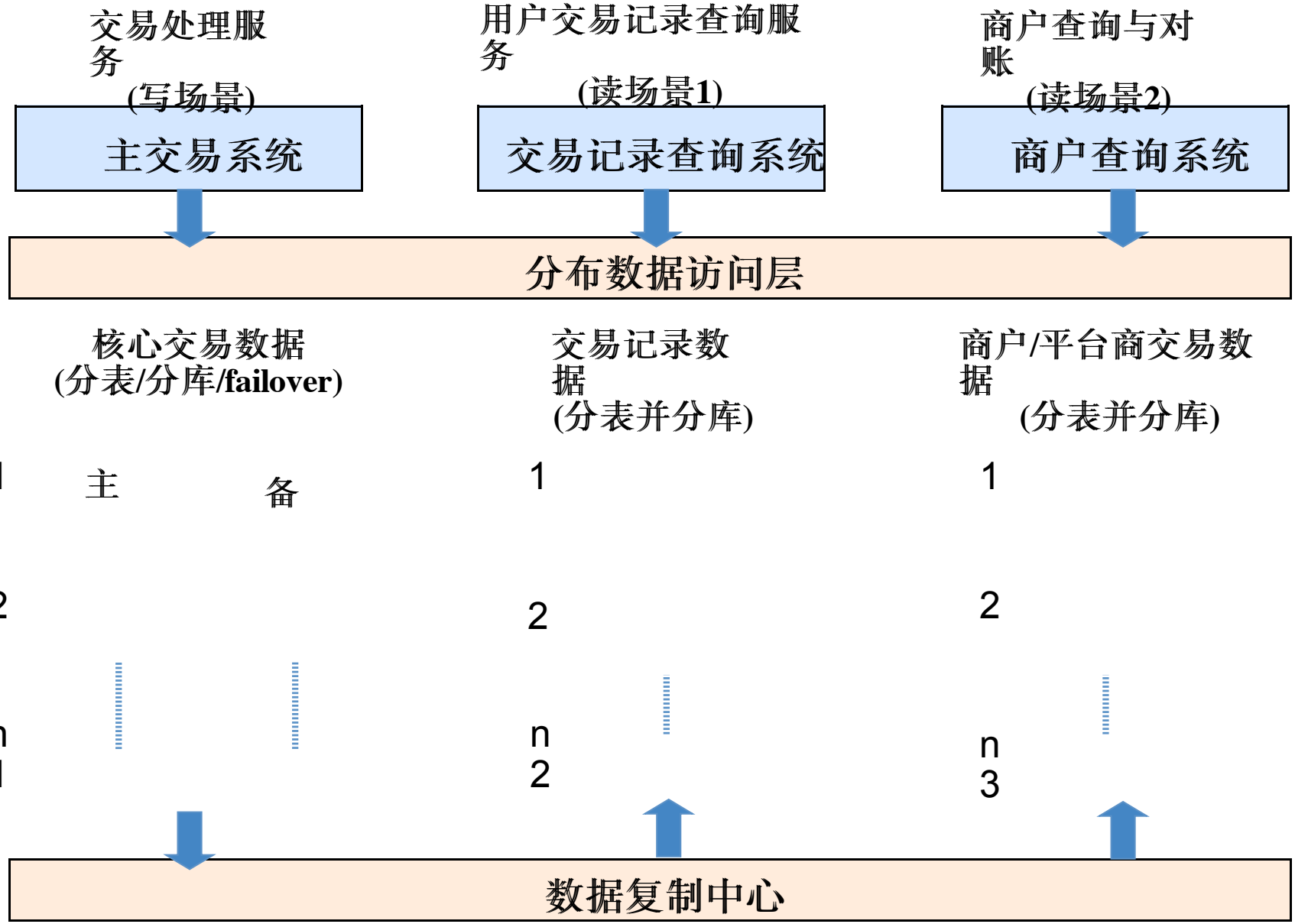

IDC1
confreg
交易系统
账户系统
交易数据库
交易数据库
交易数据库
账户数据库
账户数据库
IDC2
confreg
交易系统
账户系统
V 6

故障容忍-消除数据库单点

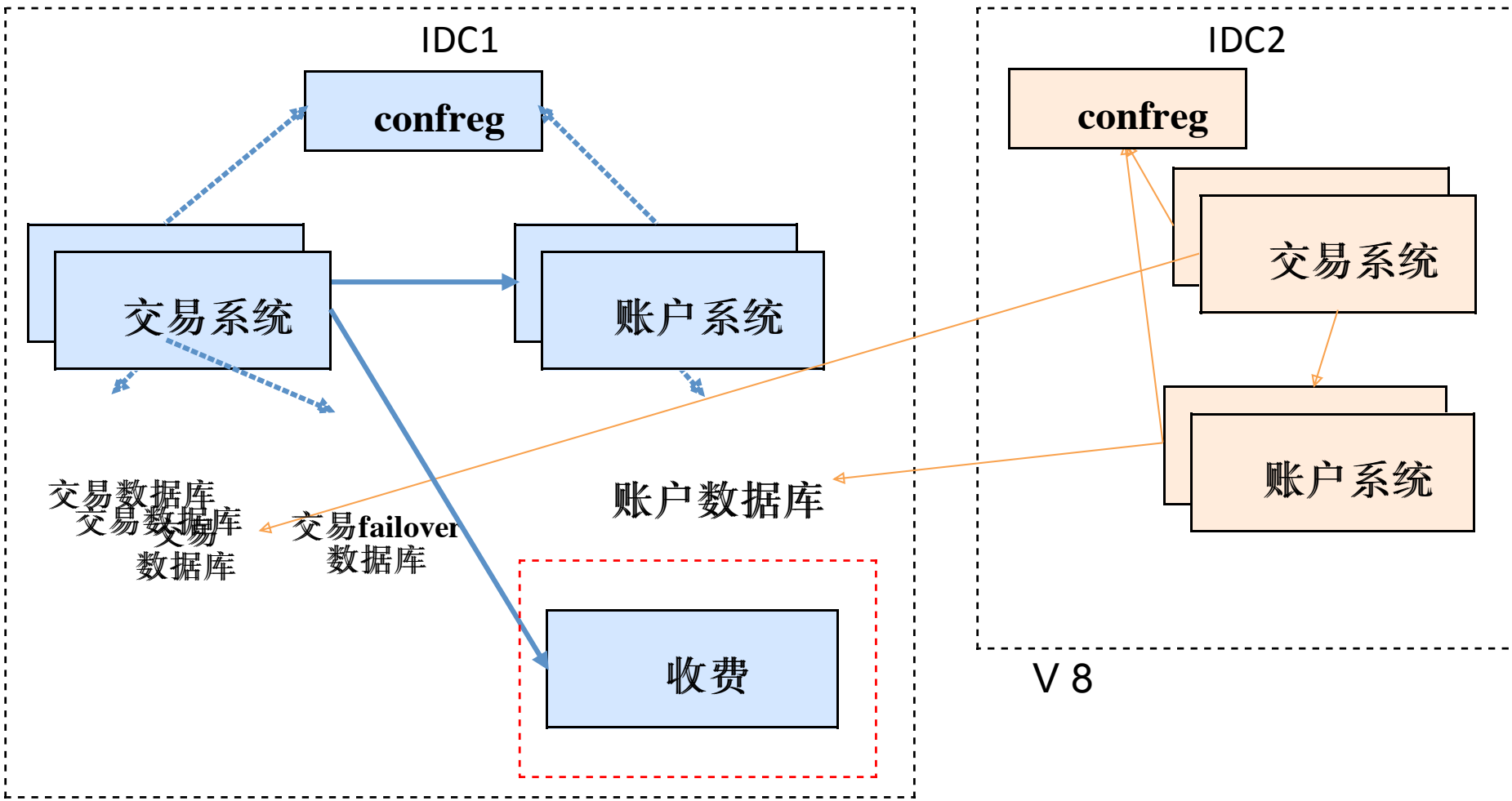


V 7

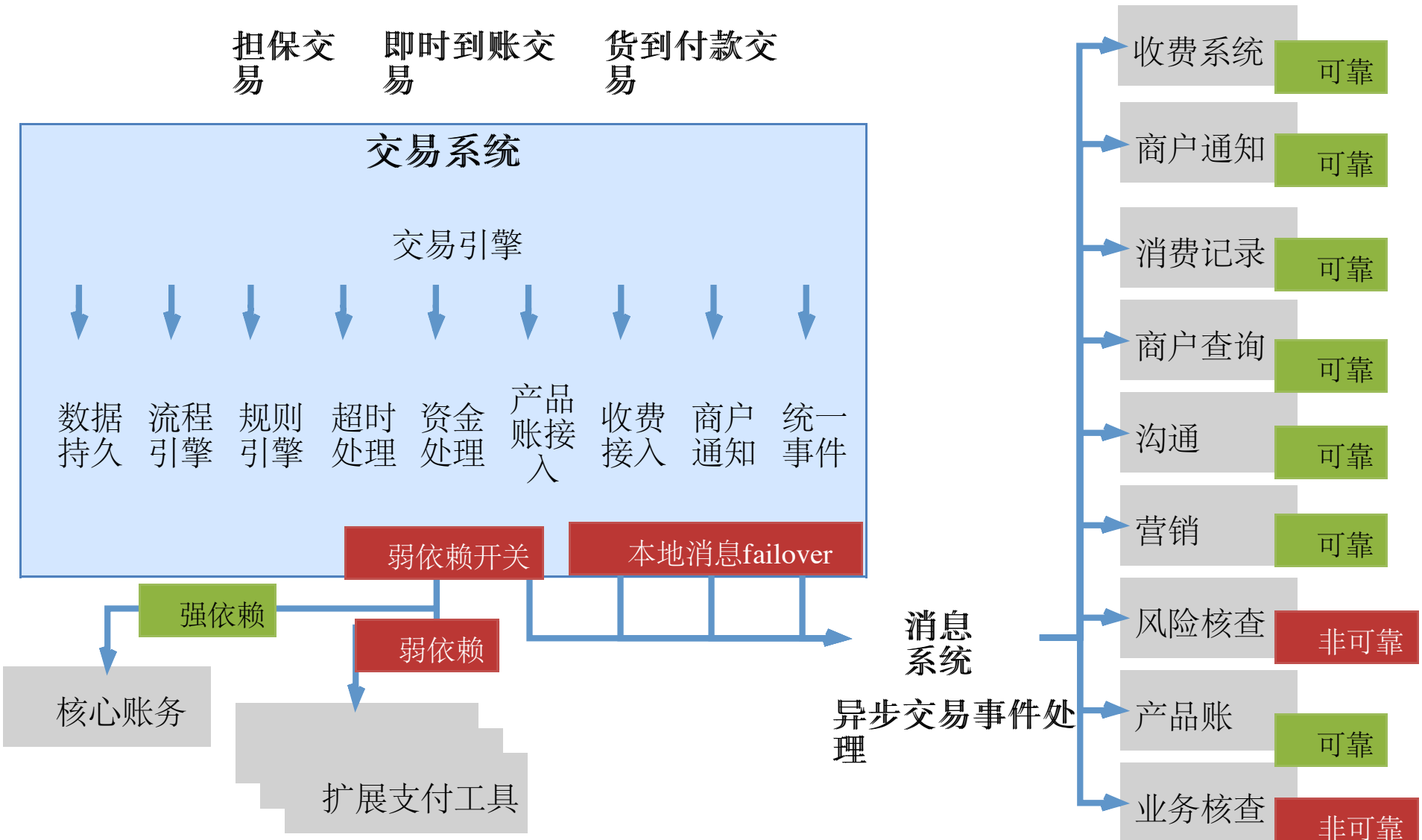
故障容忍-数据库的failover



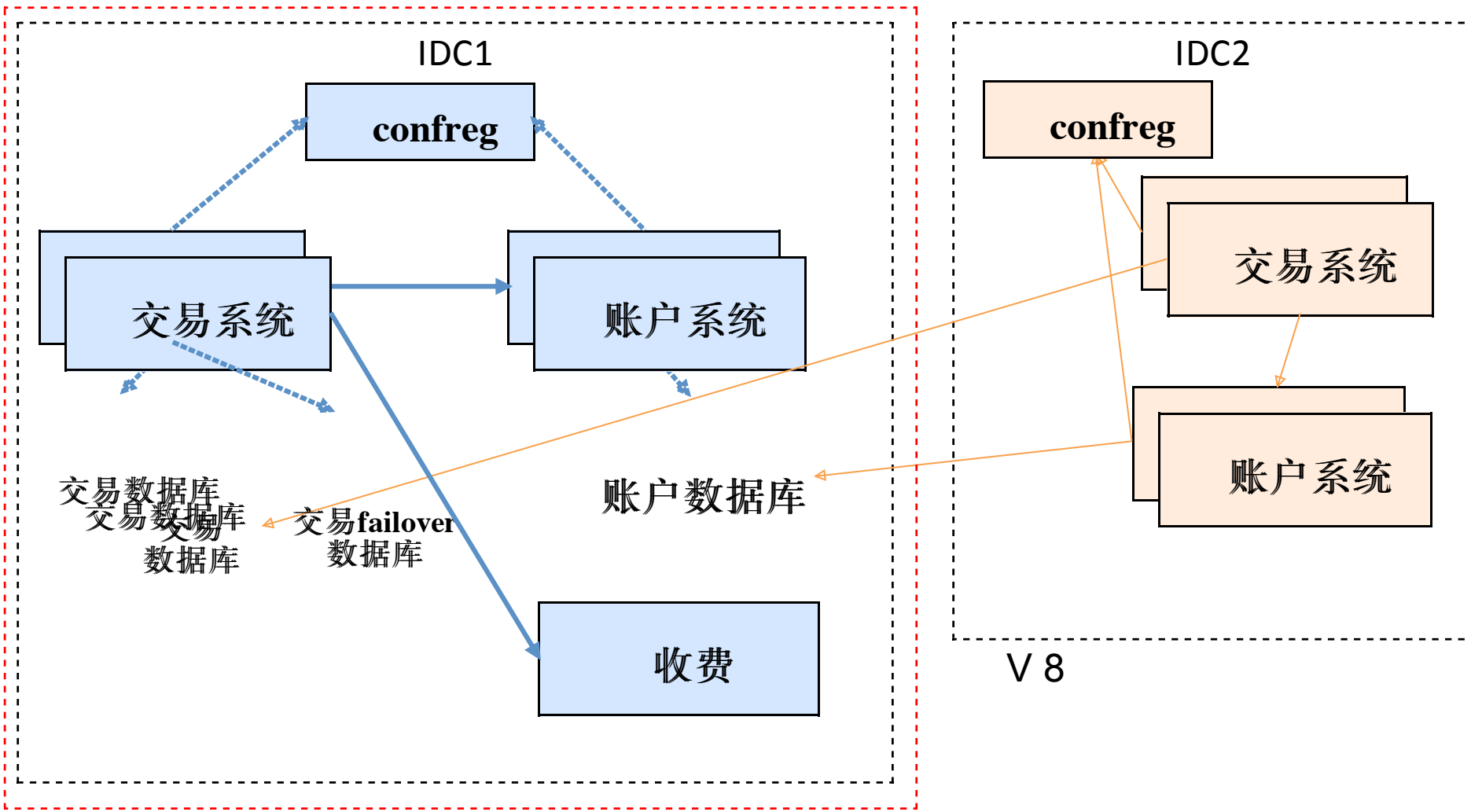
服务依赖故障


IDC1
confreg
交易系统
账户系统
交易数据库
交易数据库
交易
数据库
交易failover
数据库
账户数据库
收费
IDC2
confreg
交易系统
账户系统
V 8

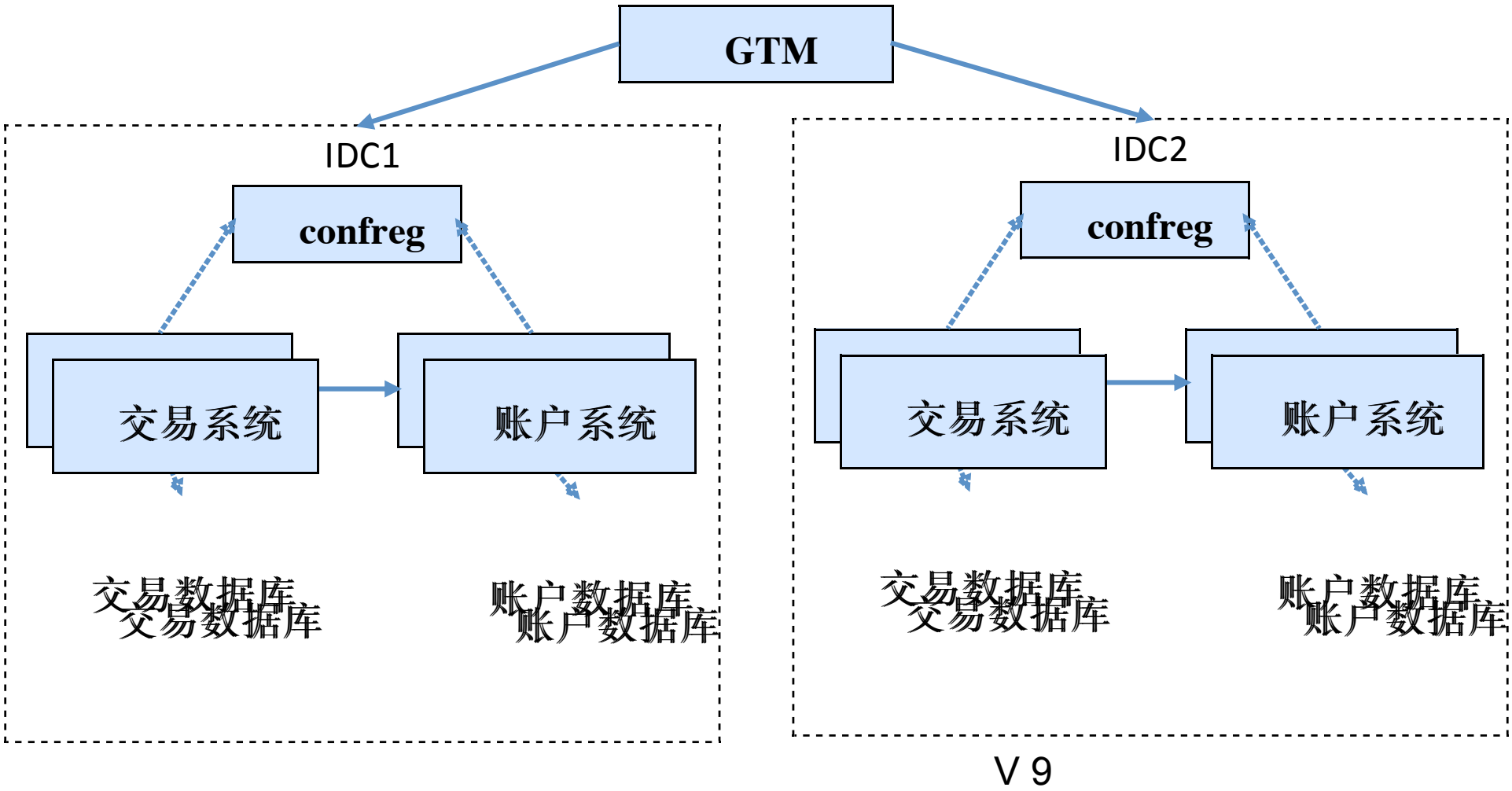
故障容忍-控制服务依赖


担保交易
即时到账交易
货到付款交易
交易系统
交易引擎
数据持久
流程引擎
规则引擎
超时处理
资金处理
产品账接入
收费接入
商户通知
统一事件
弱依赖开关
本地消息failover
强依赖
弱依赖
核心账务
扩展支付工具
消息系统
异步交易事件处理
收费系统
可靠
商户通知
可靠
消费记录
可靠
商户查询
可靠
沟通
可靠
营销
可靠
风险核查
非可靠
产品账
可靠
业务核查
非可靠

IDC故障


IDC1
confreg
交易系统
账户系统
交易数据库
交易数据库
交易
数据库
交易failover
数据库
账户数据库
收费
IDC2
confreg
交易系统
账户系统
V 8

故障容忍-完全独立IDC



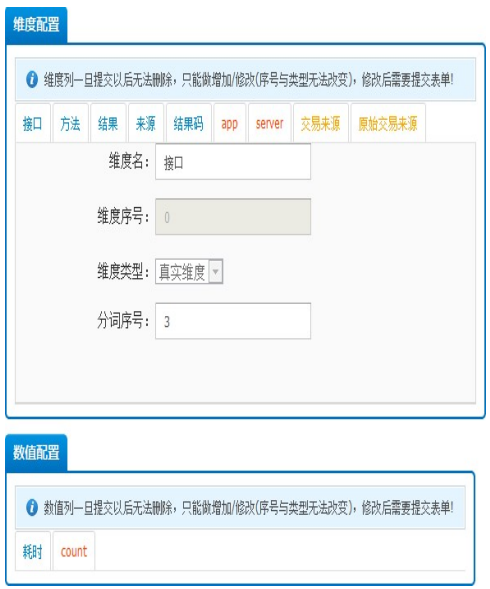
小结：99.9%到99.99%+

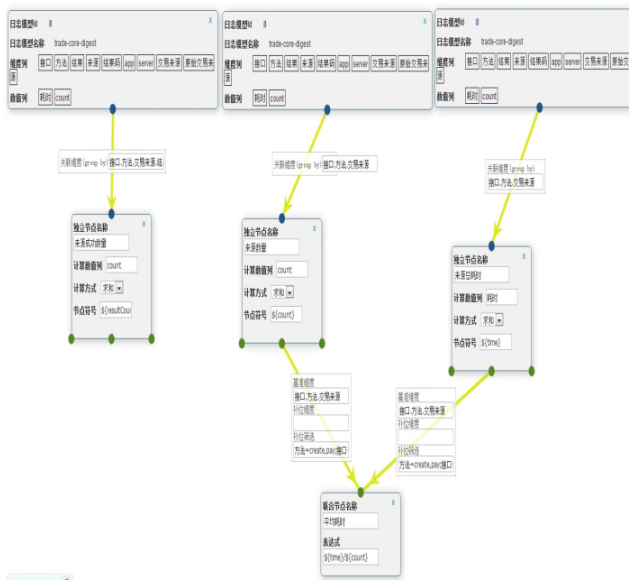
- 消除任何数据库单点
- 控制服务依赖
- 完全独立的IDC

人工控制到秒级自动调度

弹性控制

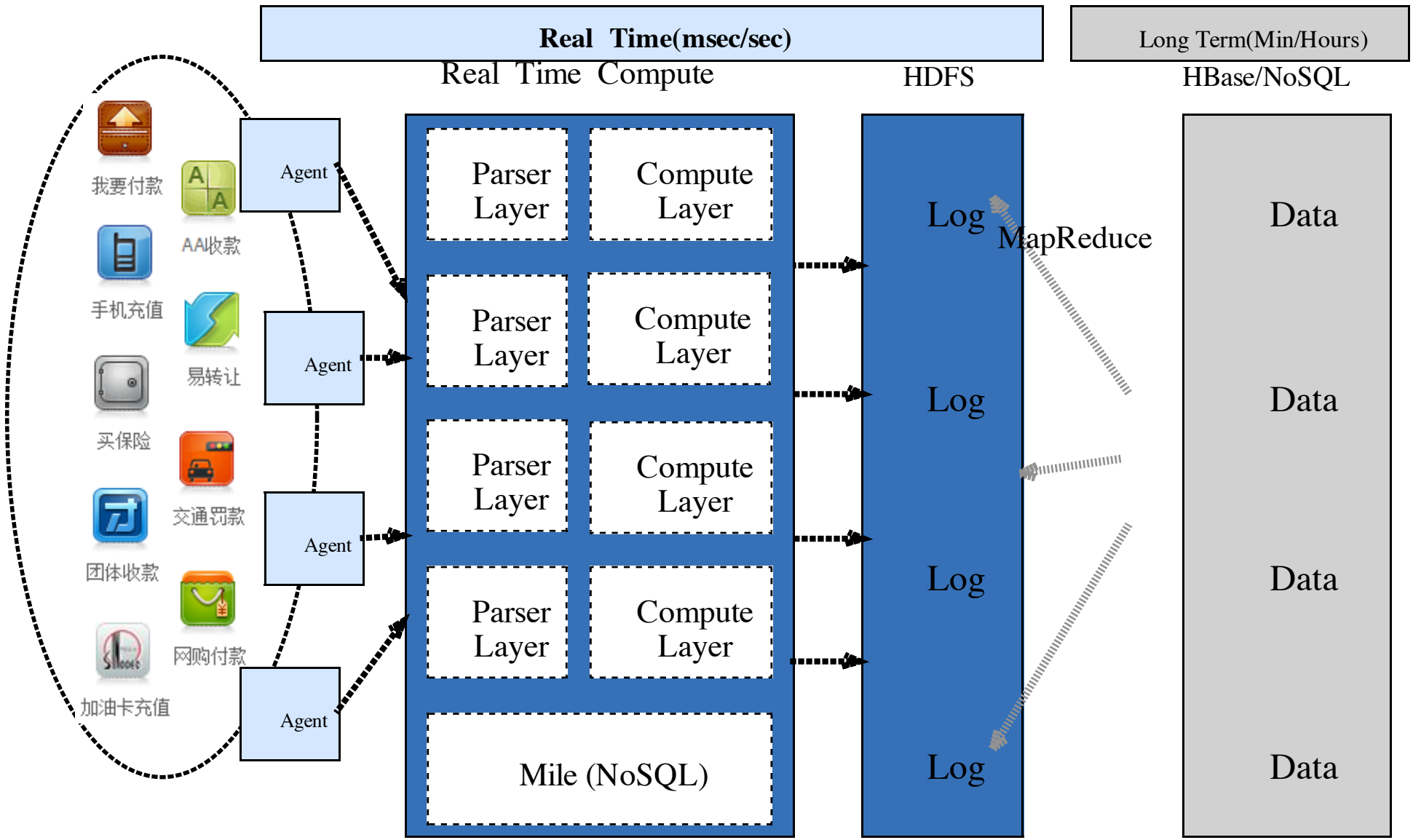
弹性能力-监控平台





时间	来源成功数量	基线
09:42	12510	12348
09:41	12528	12290
09:40	12277	12229
09:39	12240	12166
09:38	12327	12101
09:37	12217	12035
09:36	12195	11968
09:35	11963	11900
09:34	11979	11832
09:33	11954	11764

弹性能力-秒级监控系统

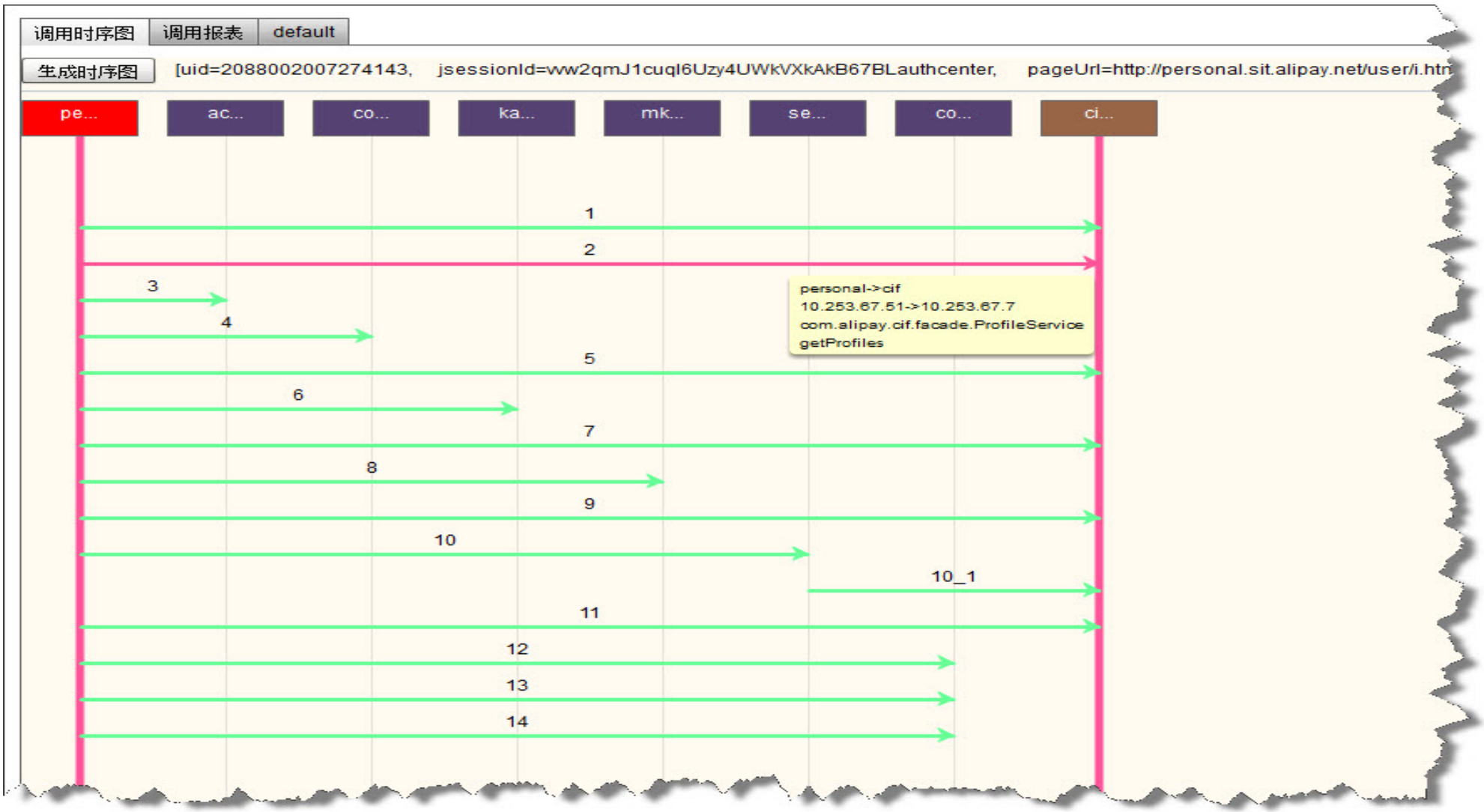

Real Time(msec/sec)
Long Term(Min/Hours)
Real Time Compute
HDFS
HBase/NoSQL
我要付款
AA收款
手机充值
易转让
买保险
交通罚款
团体收款
网购付款
加油卡充值
Agent
Agent
Agent
Agent
Parser Layer
Compute Layer
Parser Layer
Compute Layer
Parser Layer
Compute Layer
Parser Layer
Compute Layer
Mile (NoSQL)
Log
Log
Log
Log
MapReduce
Data
Data
Data
Data

弹性能力 – 容器级的精细化控制

- **help** 显示所有的指令；
- **showall** 以树型显示所有的资源属性。
- **setapp** <appname> 设置app的上下文，如果只有一个app默认选择这个app，这样可以简化整个命令行的操作；
- Display Osgi Bundles Status
  - ✓ **bundle <appname> <id>**显示对应应用的ace中的bundle资源，如果只有一个app显示默认的app中的bundle信息；
- Display Base Config
  - ✓ **config <appname>**显示某个app；
  - ✓ **app [appname]**显示所有app的信息，包含这个app的基本属性（如：有多少个服务，多少个应用，多少个datasouce，多少个drm资源）；
  - ✓ **drm <appname>**显示所有drm的信息;
- Display Service Component Runtime （服务组件模型）
  - ✓ **service <appname> [servicename 注意：可以通配符][-b [jvm|tr|ws|http 注意：可以多选] <id>** 显示所有service信息，如果只有一个app显示默认的app中的service信息；
  - ✓ **reference <appname> [servicename 注意：可以通配符][-b [jvm|tr|ws|http 注意：可以多想] <id>** 显示所有reference信息，如果只有一个app显示默认的app中的reference信息；
  - ✓ **context <appname> <id>** 显示所有的上线问信息，如果只有一个app显示默认的app中的context信息；
  - ✓ **consumer <appname> <id>** 显示所有的上线问信息，如果只有一个app显示默认的app中的consumer信息；
  - ✓ **publisher <appname> <id>** 显示所有的上线问信息，如果只有一个app显示默认的app中的publisher信息；
- Display Web Status
  - ✓ **tomcat <appname>** 显示tomcat的运行关键属性，如果只有app显示默认的app中的bundle信息；
  - ✓ **mvc** mvc框架属性；
- Display DataLevel Status
  - ✓ **datasource (缩写 ds)**
  - ✓ **tddl**
- Display Transports Status
  - ✓ **ws**
  - ✓ **tr**
- Display Performance Status
  - ✓ **ipstats (ip stats)**
  - ✓ **qps (query per second)**
  - ✓ **pv (web)**
  - ✓ **tpr (time per request)**
  - ✓ **thread 线程运行状态**
- Display Misc Status
  - ✓ **cron (schedule)**

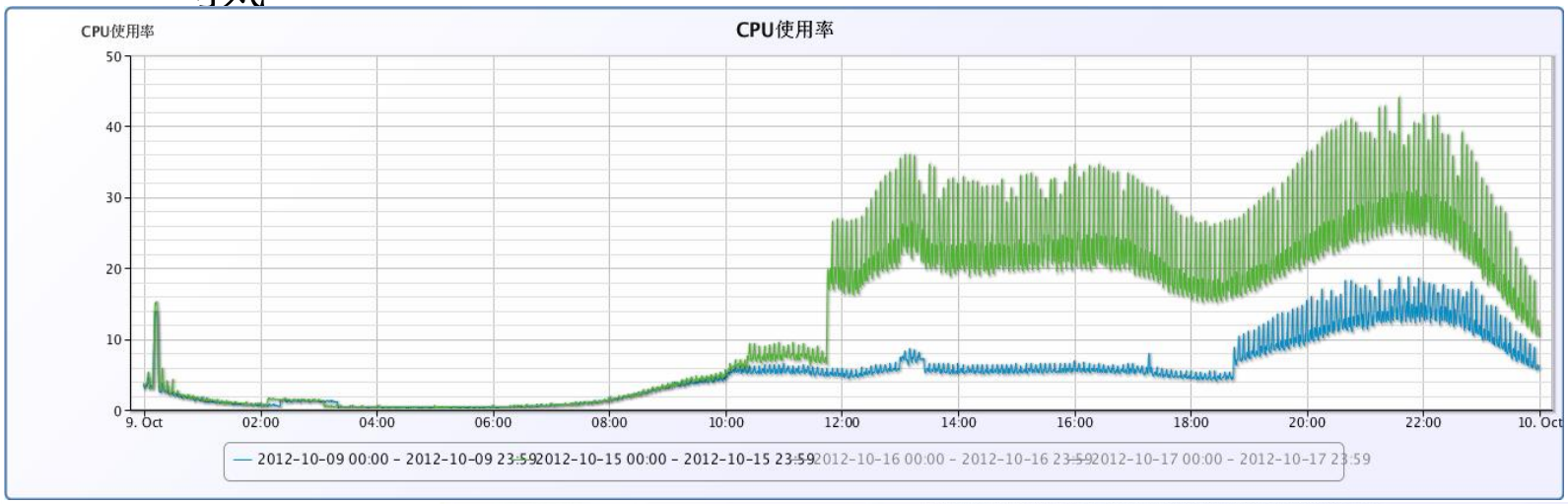
弹性能力-自动化的调度

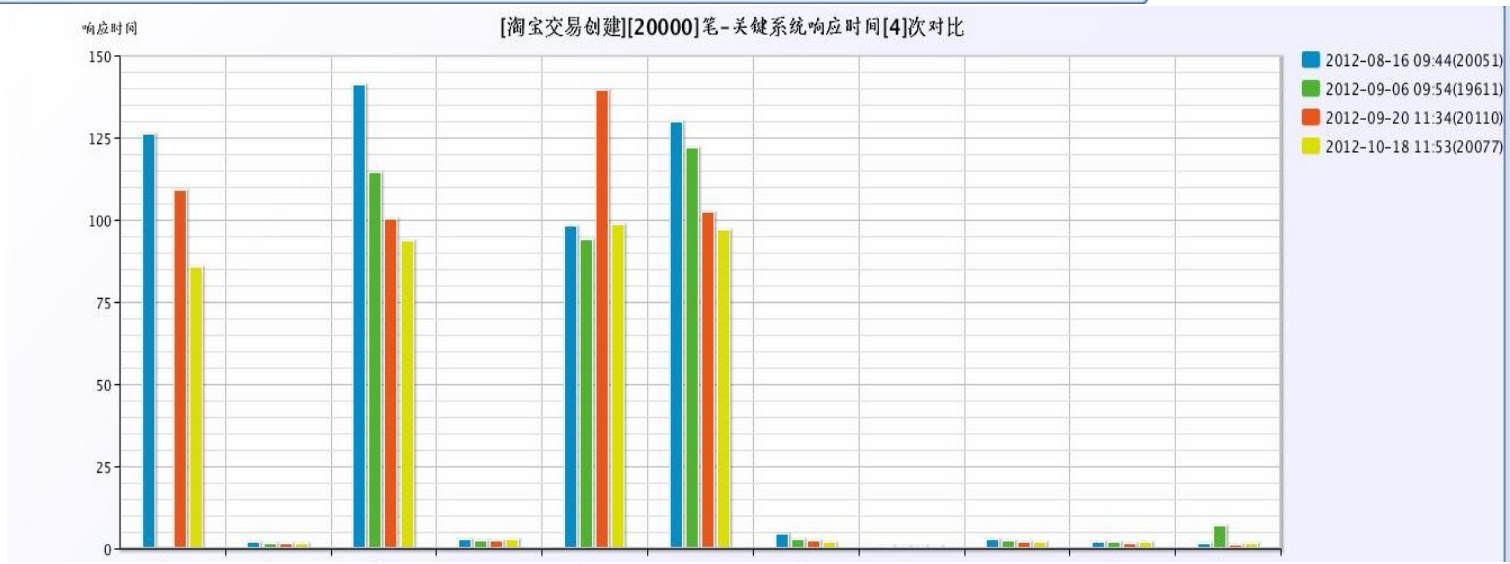
治理 – SOA调用


调用时序图
调用报表
default
生成时序图
[uid=2088002007274143, jsessionId=vvw2qmJ1cuql6Uzy4UWkVXkAkB67BLauthcenter, pageUrl=http://personal.sit.alipay.net/user/i.htm
pe...
ac...
co...
ka...
mk...
se...
co...
ci...
1
2
3
4
5
6
7
8
9
10
10_1
11
12
13
14
personal->cif
10.253.67.51->10.253.67.7
com.alipay.cif.facade.ProfileService
getProfiles

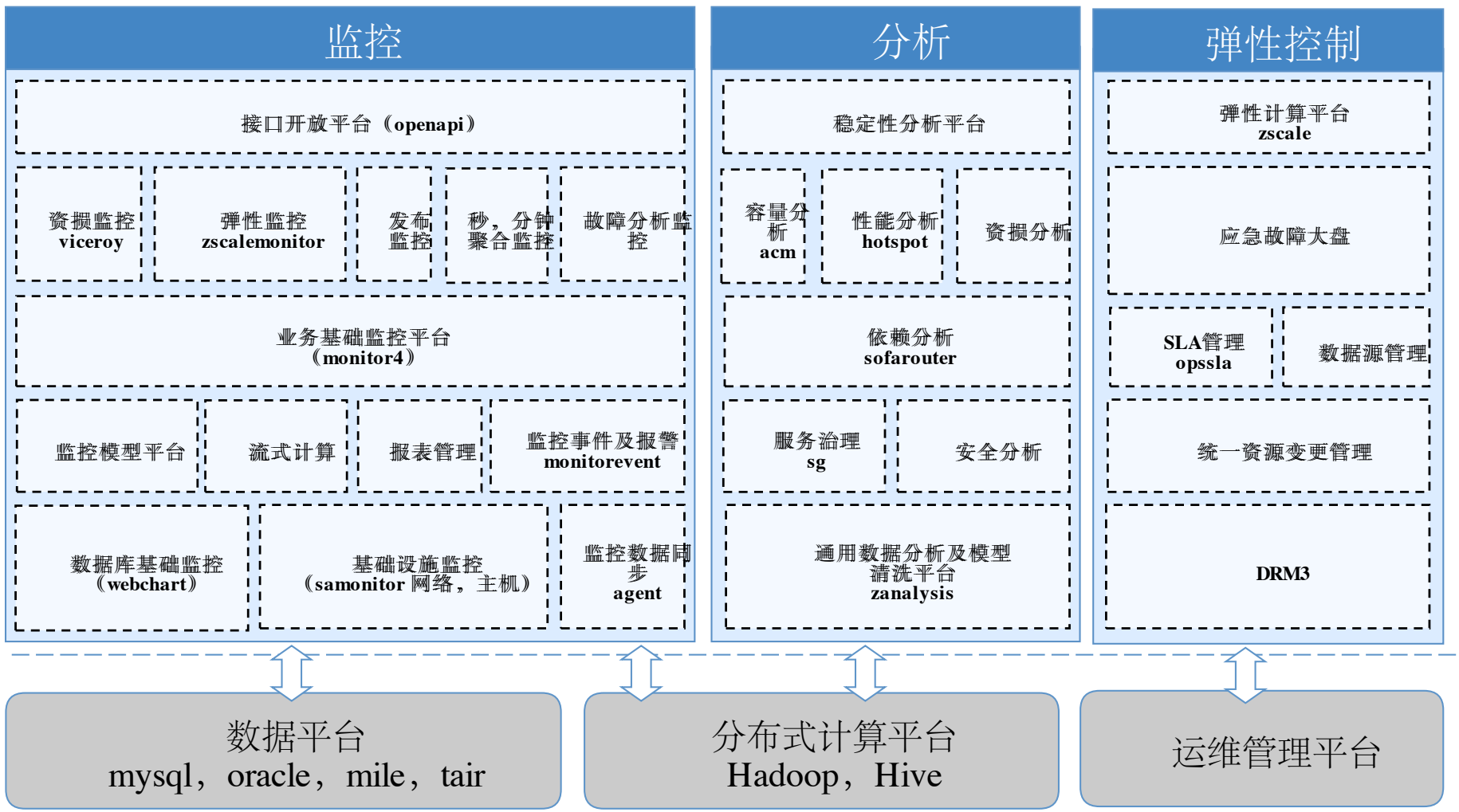
治理 – 性能分析

- 在线上提供性能历史对比图，可以对比每周发布之后的关键链路应用系统的性能变化和增长趋势





弹性控制平台


监控
接口开放平台（openapi）
资损监控
viceroy
弹性监控
zscalemonitor
发布
监控
秒，分钟
聚合监控
故障分析监
控
业务基础监控平台
（monitor4）
监控模型平台
流式计算
报表管理
监控事件及报警
monitorevent
数据库基础监控
（webchart）
基础设施监控
（samonitor 网络，主机）
监控数据同
步
agent
分析
稳定性分析平台
容量分
析
acm
性能分析
hotspot
资损分析
依赖分析
sofarouter
服务治理
sg
安全分析
通用数据分析及模型
清洗平台
zanalysis
弹性控制
弹性计算平台
zscale
应急故障大盘
SLA管理
opssla
数据源管理
统一资源变更管理
DRM3
数据平台
mysql，oracle，mile，tair
分布式计算平台
Hadoop，Hive
运维管理平台

小结：人工控制到秒级自动调度

- 实时的系统监控能力
- 快速自动化的系统调度能力
- 精细化的系统治理能力