QCon 全球软件开发大会
INTERNATIONAL SOFTWARE DEVELOPMENT CONFERENCE

BEIJING 2017

# OCTO：千亿规模下的服务治理挑战与实践

美团点评基础架构团队 张熙

主办方 Geekbang. 极客邦科技 InfoQ

# 个人简介

- 12年加入美团，基础架构部服务治理、集群调度团队负责人
    - OCTO：分布式服务通信框架及服务治理系统
    - HULK：容器集群管理及弹性伸缩平台
- 专注于面向服务架构、服务治理、大规模分布式系统、高性能通信框架、容器化、弹性调度等领域

# Agenda

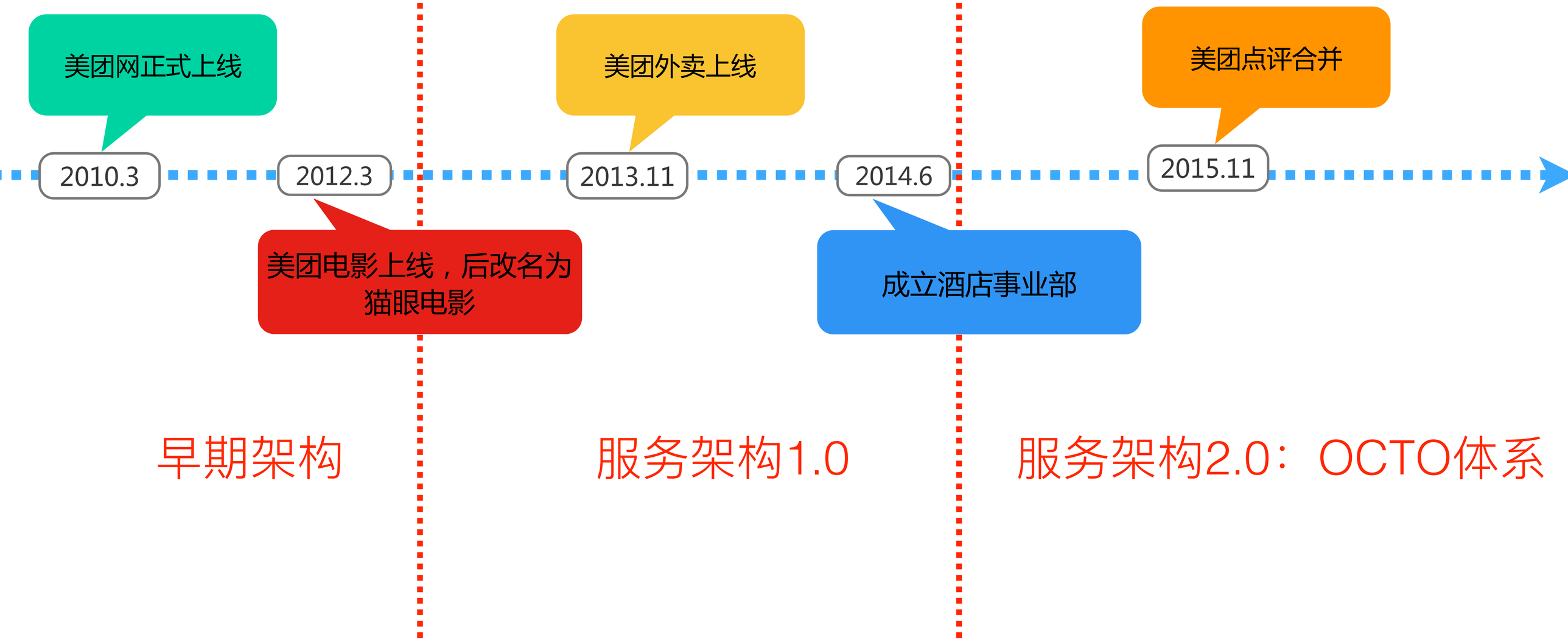
- 美团服务架构演进历程
- OCTO架构设计及研发要点
- OCTO服务治理实践

# 美团点评介绍

- 中国O2O行业最大的公司
- 覆盖全国2800多个市、县、区
- 2016年交易额超2000亿
- 共480万合作商户
- 6亿独立用户数
- 外卖日订单量1000万
- …

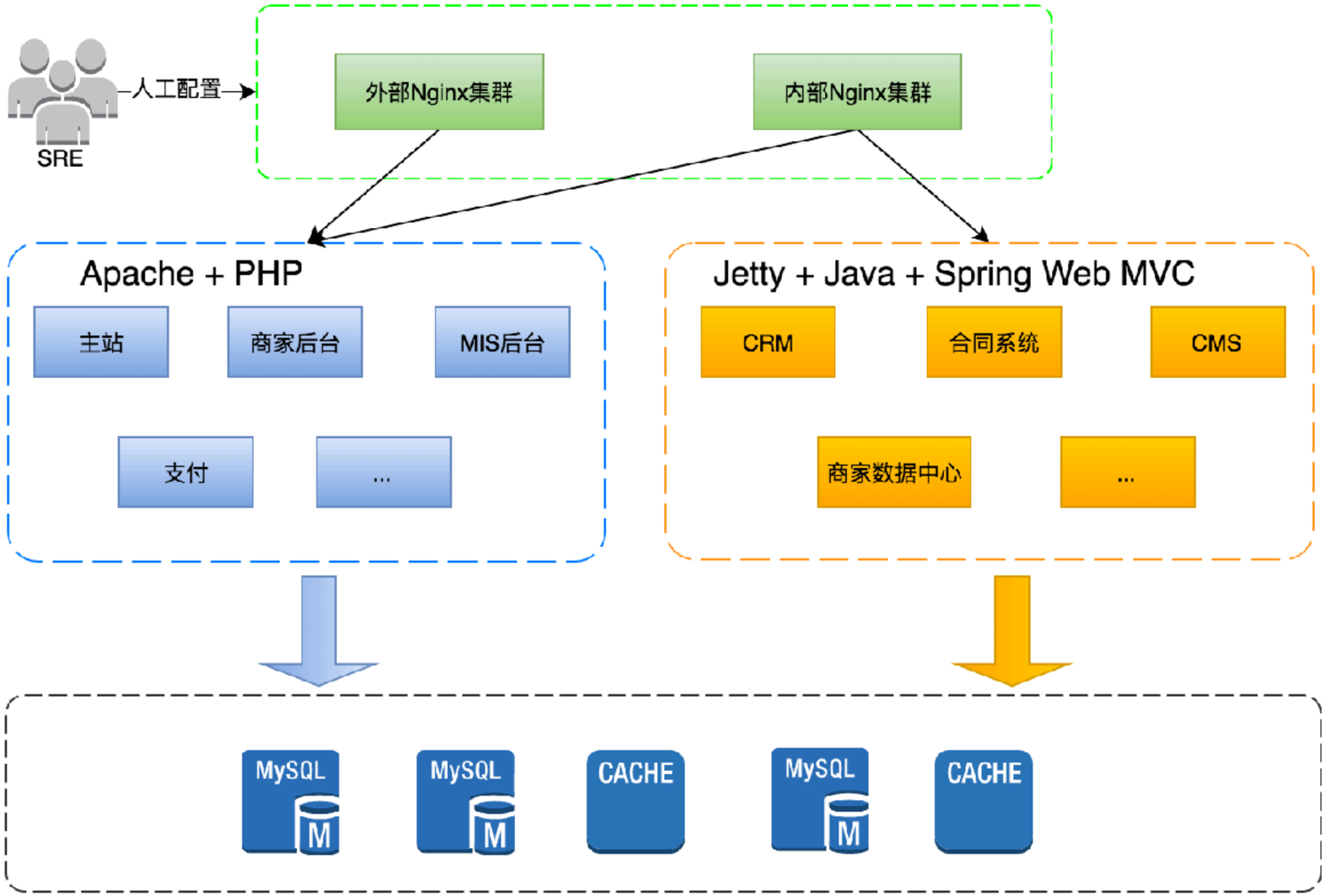
# 美团服务架构演进历程

# 美团服务架构演进


美团网正式上线
2010.3
2012.3
美团电影上线，后改名为猫眼电影
美团外卖上线
2013.11
2014.6
成立酒店事业部
美团点评合并
2015.11
早期架构
服务架构1.0
服务架构2.0：OCTO体系

# 早期架构

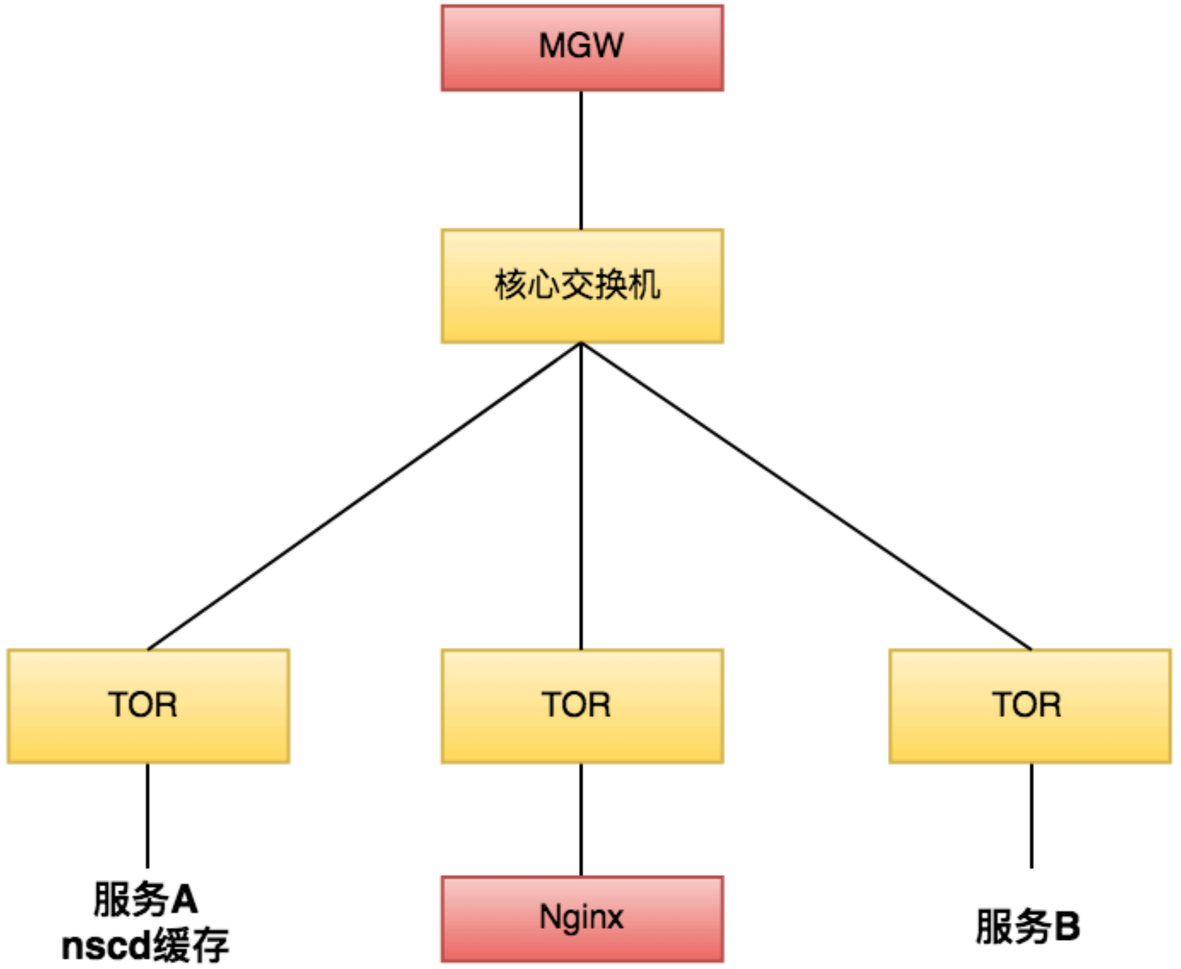
- 垂直应用架构
- LAMP体系
- HTTP + JSON



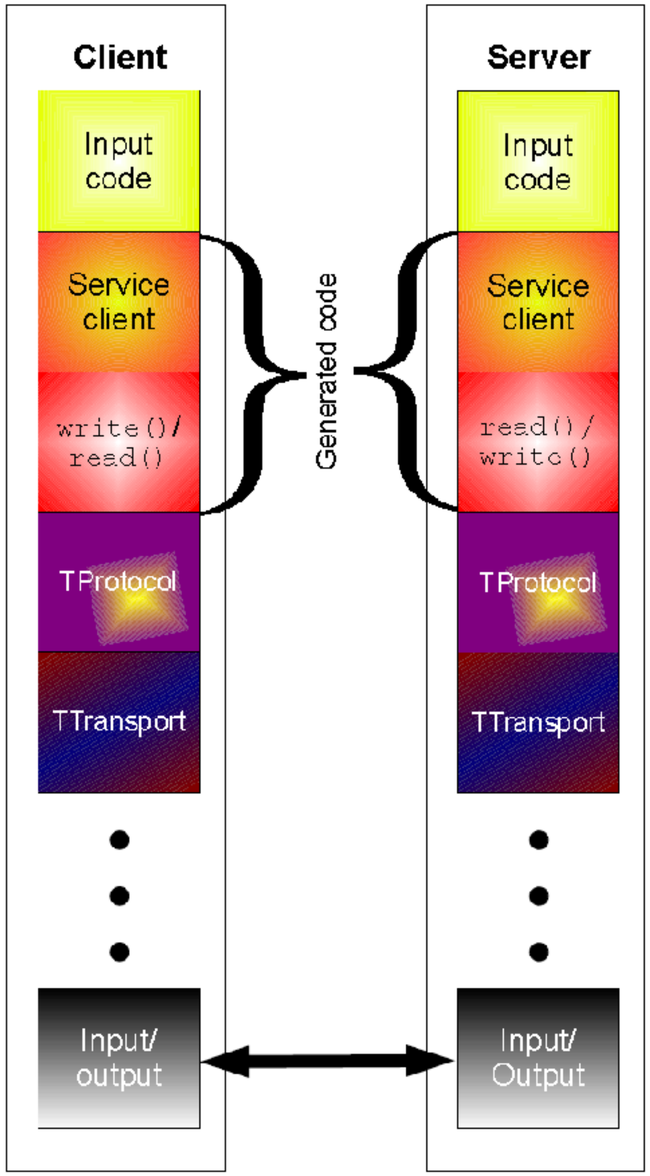
# 早期架构 - 问题与挑战

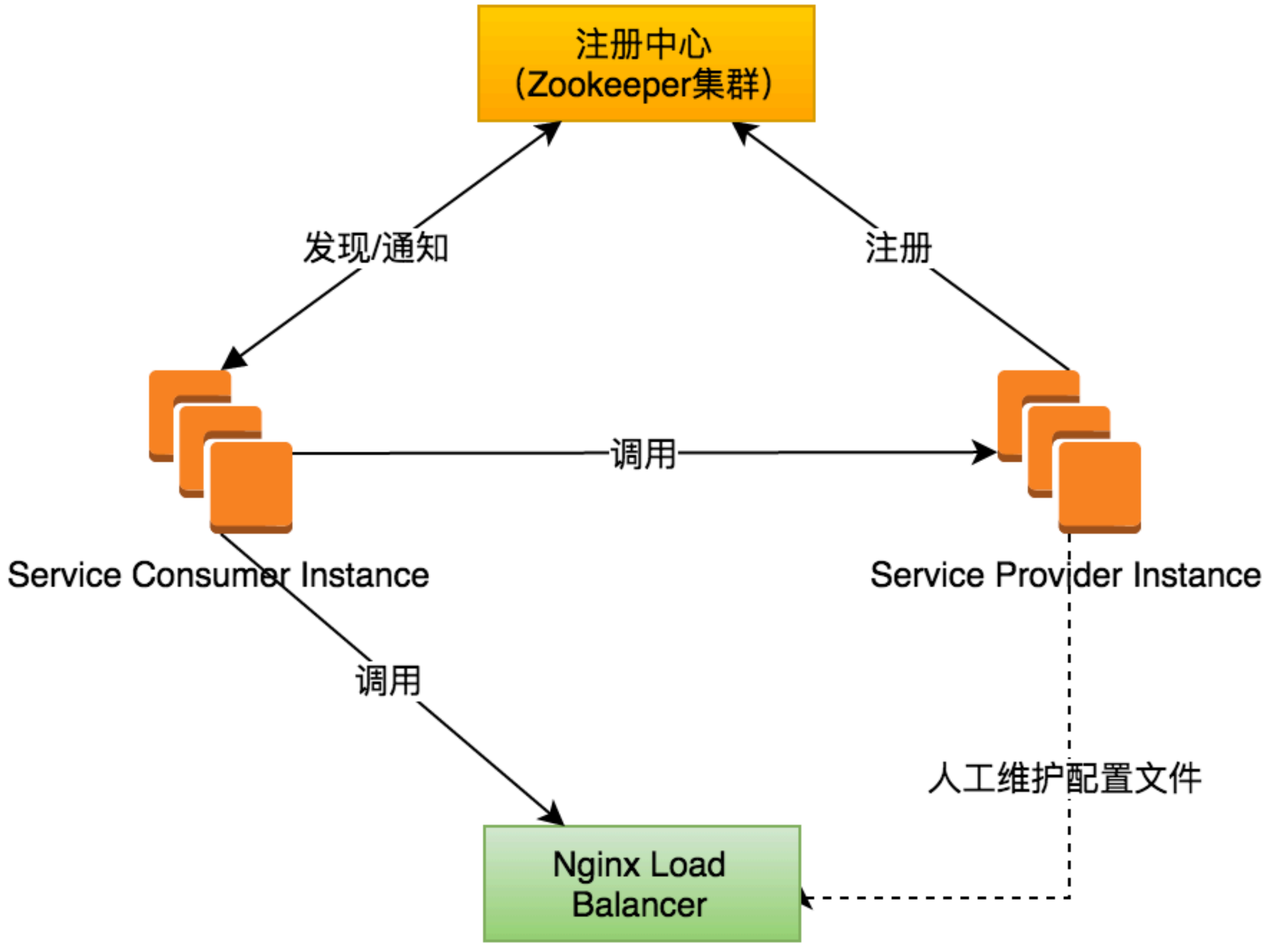
- 接口定义：缺乏强Scheme约束
- HTTP + JSON：开发成本、规范
- HTTP协议：内网链路过长

- 服务化设计、实践不够普及
- 缺乏易用、高性能的RPC通信框架
- 缺乏服务自动注册、发现机制



# 服务架构1.0


Client
Server
Input code
Service client
write()/ read()
Generated code
Input code
Service client
read()/ write()
TProtocol
TTransport
TProtocol
TTransport
Input/ output
Input/ Output


注册中心
(Zookeeper集群)
发现/通知
注册
调用
Service Consumer Instance
Service Provider Instance
调用
人工维护配置文件
Nginx Load Balancer

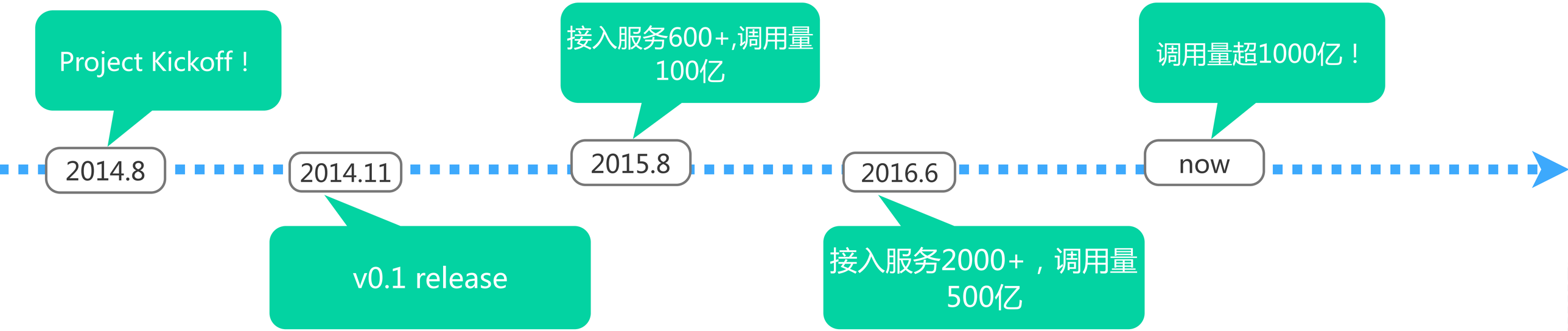
# 服务架构1.0 - 问题与挑战

- 服务注册中心
  - Zookeeper、临时节点、故障隔离
  - 多语言支持
- 服务通信框架
  - 路由、流量策略
  - 强耦合、客户端过重
  - 缺乏数据及监控
- 缺乏服务治理、运营功能

# OCTO：分布式服务通信框架及服务治理系统

- OCTO是什么？

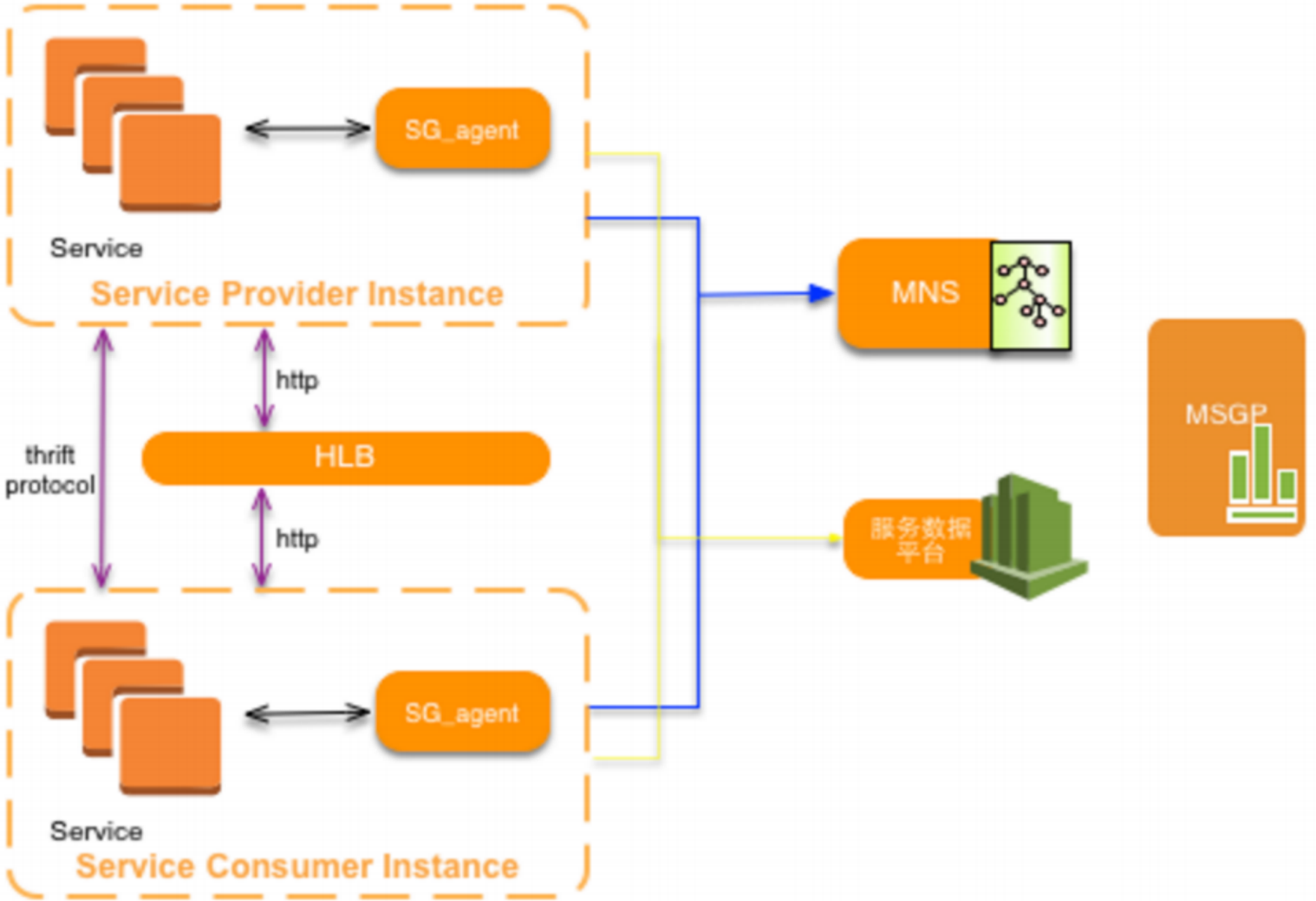
  公司级基础设施，为公司所有业务提供统一的高性能服务通信框架，使业务具备良好的服务运营能力，轻松实现服务注册、服务自动发现、负载均衡、容错、灰度发布、数据可视化、监控告警等功能，提升服务开放效率、可用性及服务运维效率。



# OCTO架构设计及研发要点

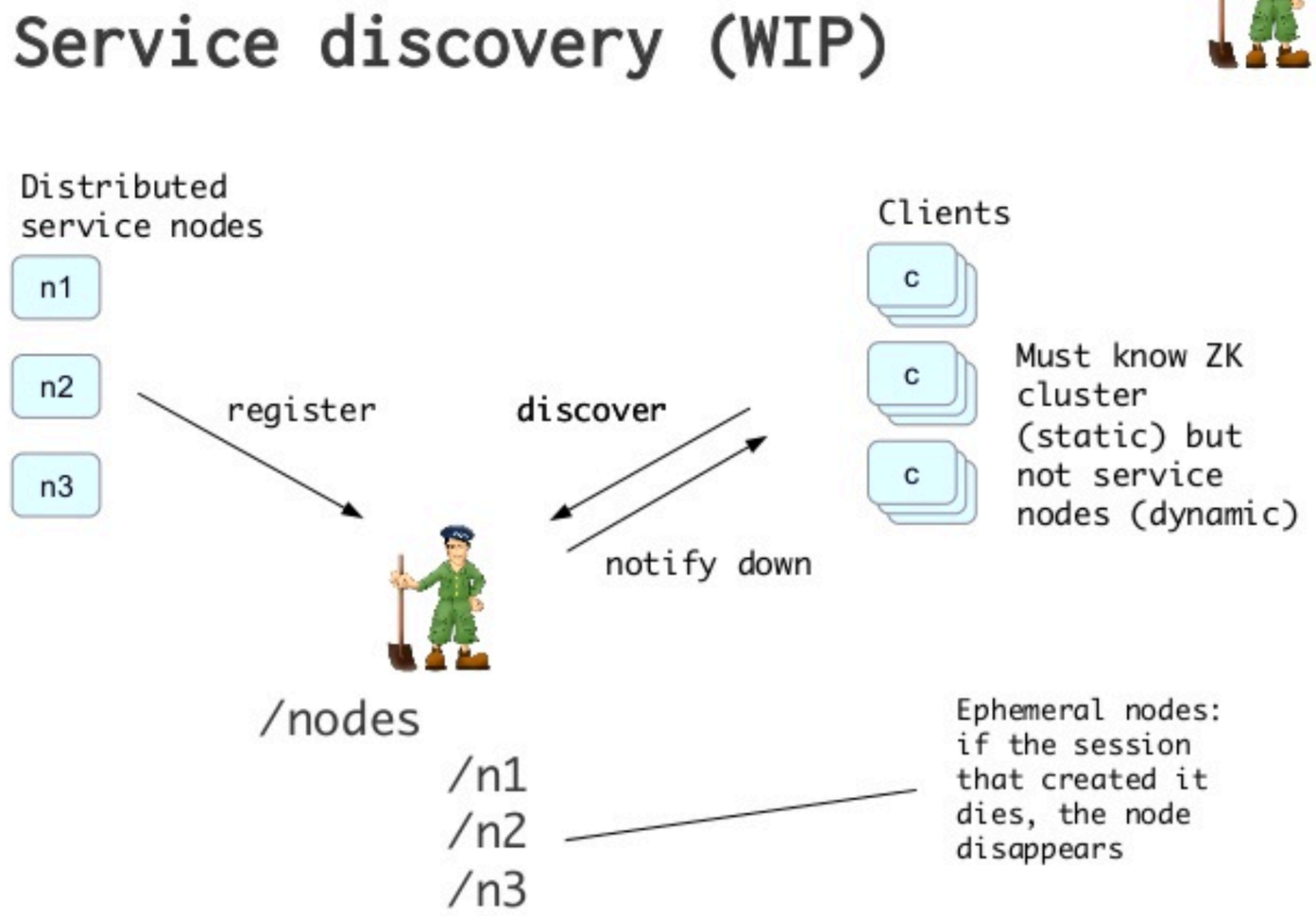
# OCTO - 整体架构

- MNS：命名服务
- SG_Agent：服务治理代理
- MTransport：服务通信框架
- HLB：弹性负载均衡器
- MSGP：服务治理平台



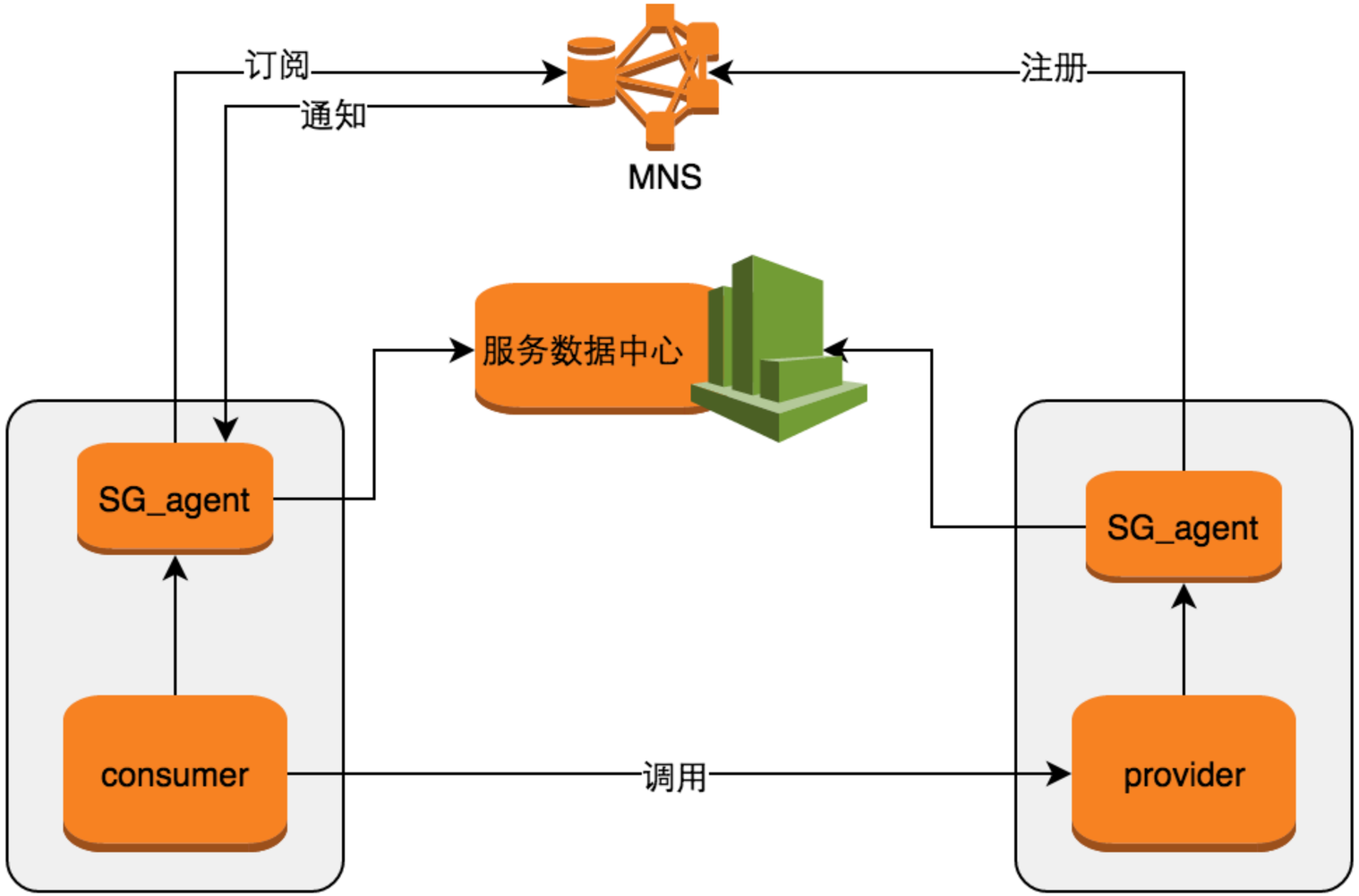
# 服务注册发现 - 传统实现

- ZooKeeper方式
  - 临时节点
  - 框架直连
  - 触发or轮询更新
- 问题与挑战
  - session timeout
  - ZooKeeper ACL
  - 紧耦合、运维影响
  - 集群不稳定、故障隔离



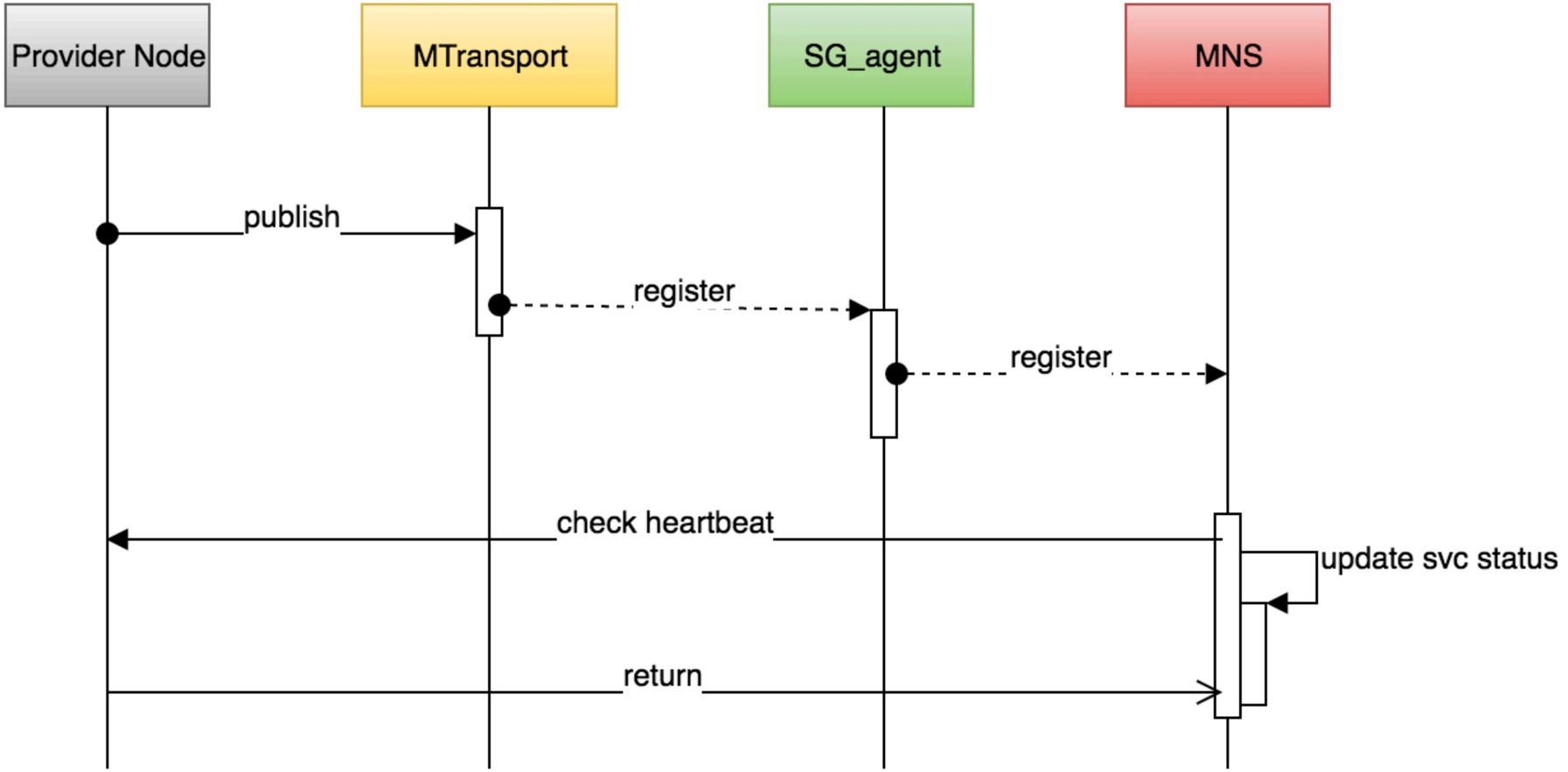
# 服务注册发现 - 代理模式

- SG_Agent：服务治理代理
  - 本地进程
  - 标准化接口
  - 策略热更新
- 特点
  - 高可用
  - 低消耗
  - 标准化部署



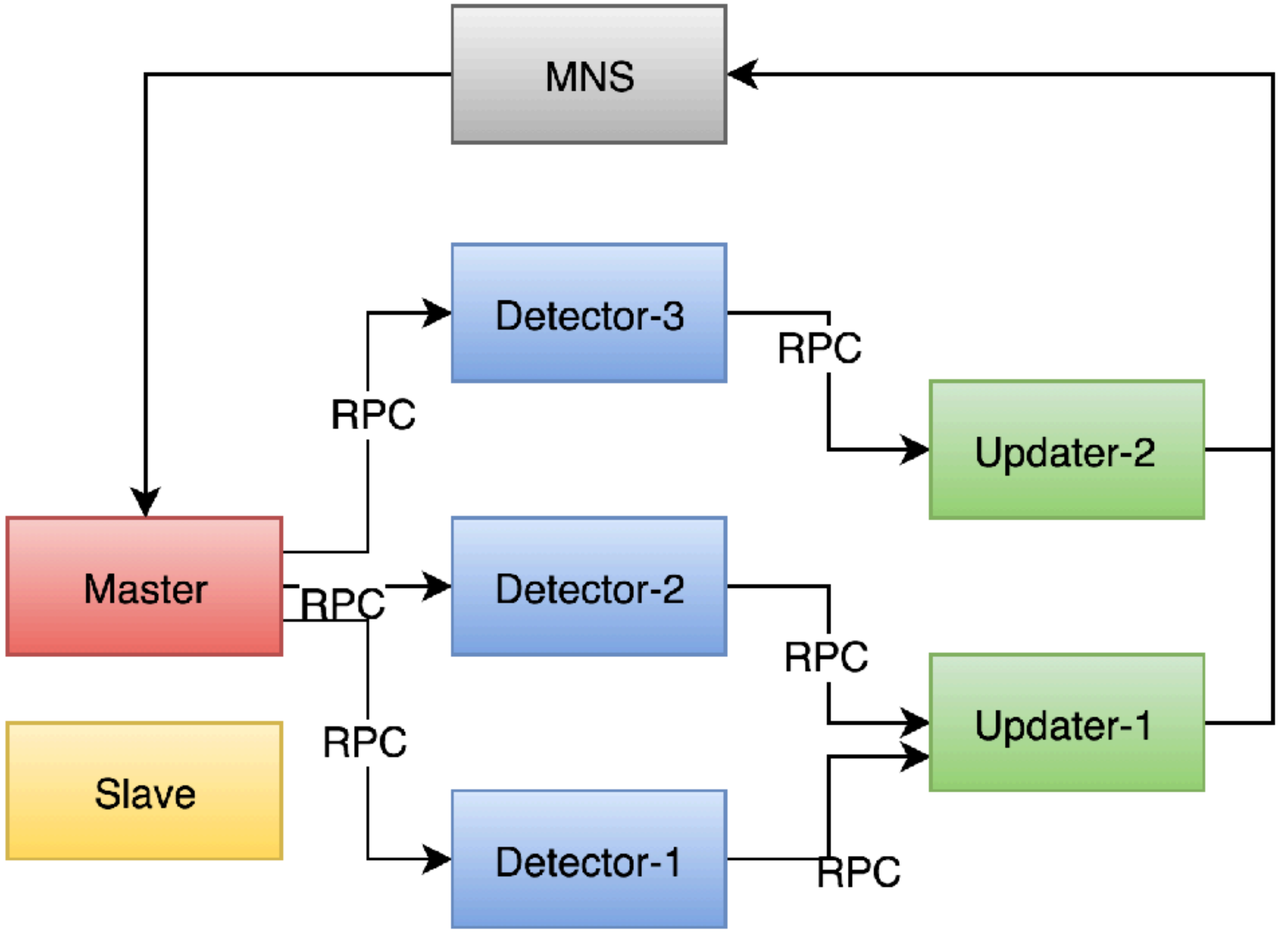
# 代理模式 - 服务注册流程

- 框架启动时注册
- 委托代理执行
- MNS探测状态



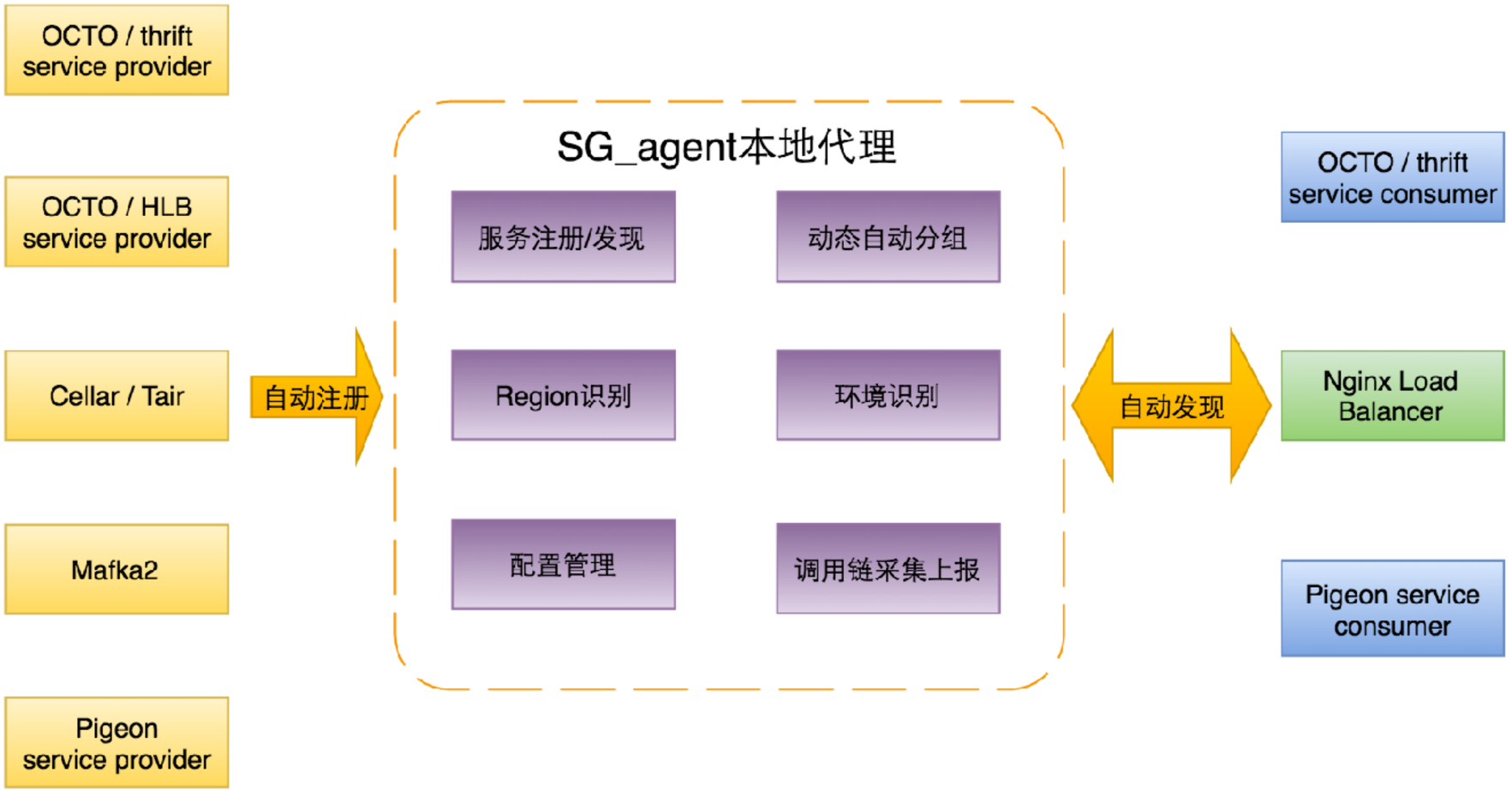
# 命名服务 - 状态检查

- 一致性：集中式检查、全局视图
- 高可用：热备、集群化、熔断
- 高准确：double check
- 高效率：Akka Actor、水平扩展



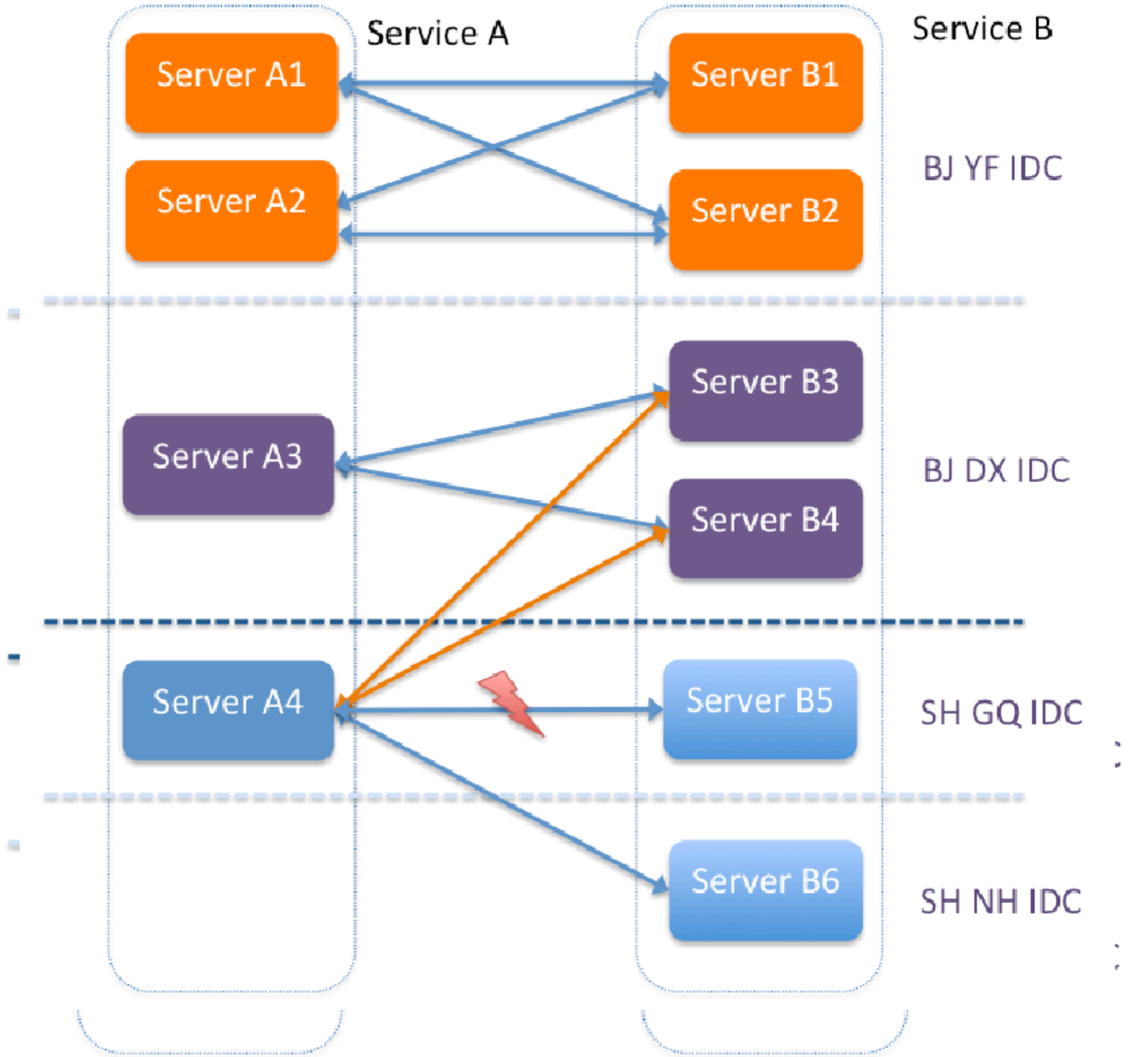
# 服务通信框架 - 策略下移

- 客户端做薄
  - 注册、发现逻辑
  - 多机房、地域
  - 路由策略
  - 环境识别
  - 慢启动策略
  - 节点注册限制



# 策略下移 - 路由策略

- 路由策略
  - 机房
  - 地域
  - 自定义路由
  - 泳道
  - 多中心



# 服务通信框架 - 网络内核

老内核

- 不支持链接复用、异步化支持不友好
- 性能稍差，4核4G、1K数据QPS 8w+
- 强依赖，模块紧耦合，不方便做自定义扩展

新内核

- 支持链接复用，原生异步支持
- 4核4G、1K数据 QPS 11w+
- 弱依赖，方便扩展，如服务鉴权、链接保护

# 服务通信框架 - 统一协议

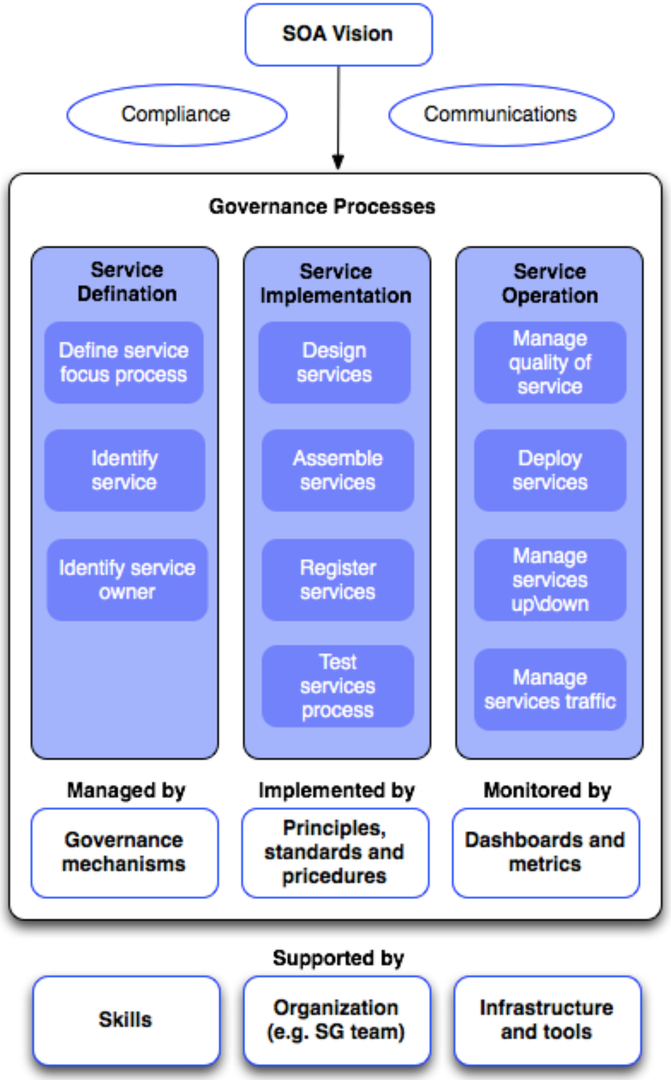
- 美团点评合并：hessian vs thrift，通讯框架互调互通
- 自定义二进制协议：兼容支持原生Thrift协议
- 自定义协议头：携带调用链信息、上下文
- 支持全链路参数传递，集成支持全链路压测需求
- 其它：支持gzip、snappy、checksum

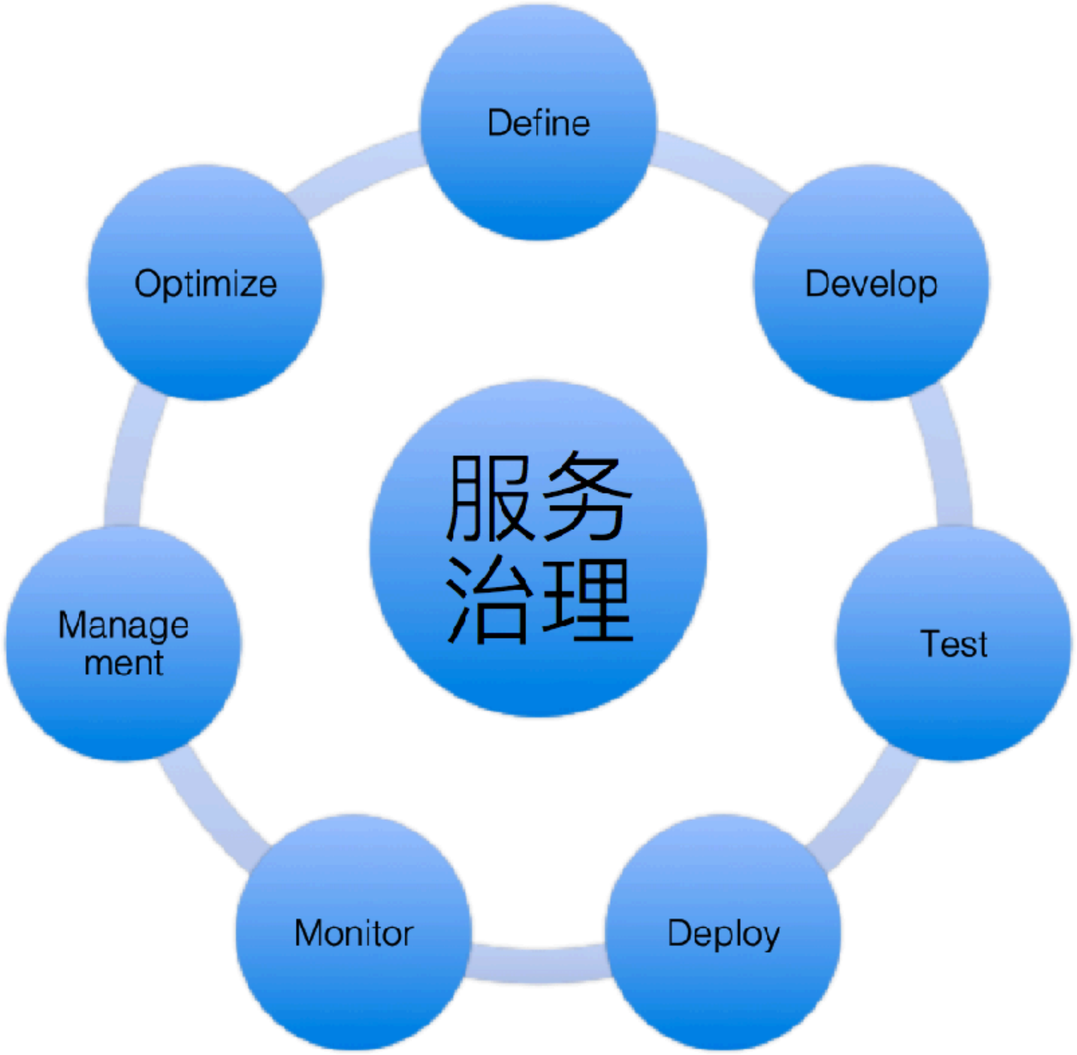
# 服务通信框架 - 容灾降级

- 过载保护：线程池过载保护；按服务、接口设置配额
- 链接保护：服务端可支持的最大链接数
- 服务隔离：不同服务可分别配置线程池；快慢请求隔离
- 一键截流：按调用端、接口等一键掐流量
- 容错处理：快速降权，快速恢复

# OCTO服务治理实践

# 服务治理 - 全生命周期


SOA Vision
Compliance
Communications
Governance Processes
Service Defination
Define service focus process
Identify service
Identify service owner
Service Implementation
Design services
Assemble services
Register services
Test services process
Service Operation
Manage quality of service
Deploy services
Manage services up\down
Manage services traffic
Managed by
Implemented by
Monitored by
Governance mechanisms
Principles, standards and pricedures
Dashboards and metrics
Supported by
Skills
Organization (e.g. SG team)
Infrastructure and tools


Define
Develop
Test
Deploy
Monitor
Manage ment
Optimize
服务治理

# 全生命周期

Define

- 服务注册、负责人与SRE服务树/CMDB系统打通

Develop

- 支持IDL定义、代码在线生成

Test

- 测试环境隔离，泳道特性
- 打通性能测试平台，支持引流压测，获取容量数据

Deploy

- 打通发布系统，无损平滑发布
- 打通服务树，限制节点注册

# 全生命周期

## Monitor：采集数据上报监控平台，状态、QPS/TP、异常

Dashboard　业务指标 New!　性能指标　上下游分析 New!　来源分析　主机分析　去向分析　调用链分析 New!

角色：来源 去向　环境：prod stage test　日期：2016-06-22　查询　可用性指标　TP耗时数据　同比环比

| 接口 | 调用总量 | 成功数/百分比 | 异常数/百分比 | 过载数/百分比 | QPS(次/秒),环比,同比 | TP50耗时(毫秒) | TP90耗时(毫秒),环比,同比 | TP95耗时(毫秒) | TP99耗时(毫秒) |
|---|---|---|---|---|---|---|---|---|---|
| all | 2447391900 | 2447391594, 100.0000% | 306, 0.0000% | 0, 0.0000% | 28326.295,-4%,8% | 2 | 4,0%,33% | 6 | 29 |
| [redacted] | 1666041915 | 1666041719, 100.0000% | 196, 0.0000% | 0, 0.0000% | 19282.893,-2%,8% | 2 | 3,0%,0% | 4 | 9 |
| [redacted] | 294961094 | 294961062, 100.0000% | 32, 0.0000% | 0, 0.0000% | 3413.902,-2%,9% | 2 | 4,0%,33% | 6 | 13 |
| [redacted] | 177186291 | 177186264, 100.0000% | 27, 0.0000% | 0, 0.0000% | 2050.767,-13%,11% | 2 | 4,0%,0% | 6 | 11 |
| [redacted] | 54548506 | 54548502, 100.0000% | 4, 0.0000% | 0, 0.0000% | 631.348,-8%,13% | 2 | 4,0%,0% | 6 | 11 |

# 全生命周期

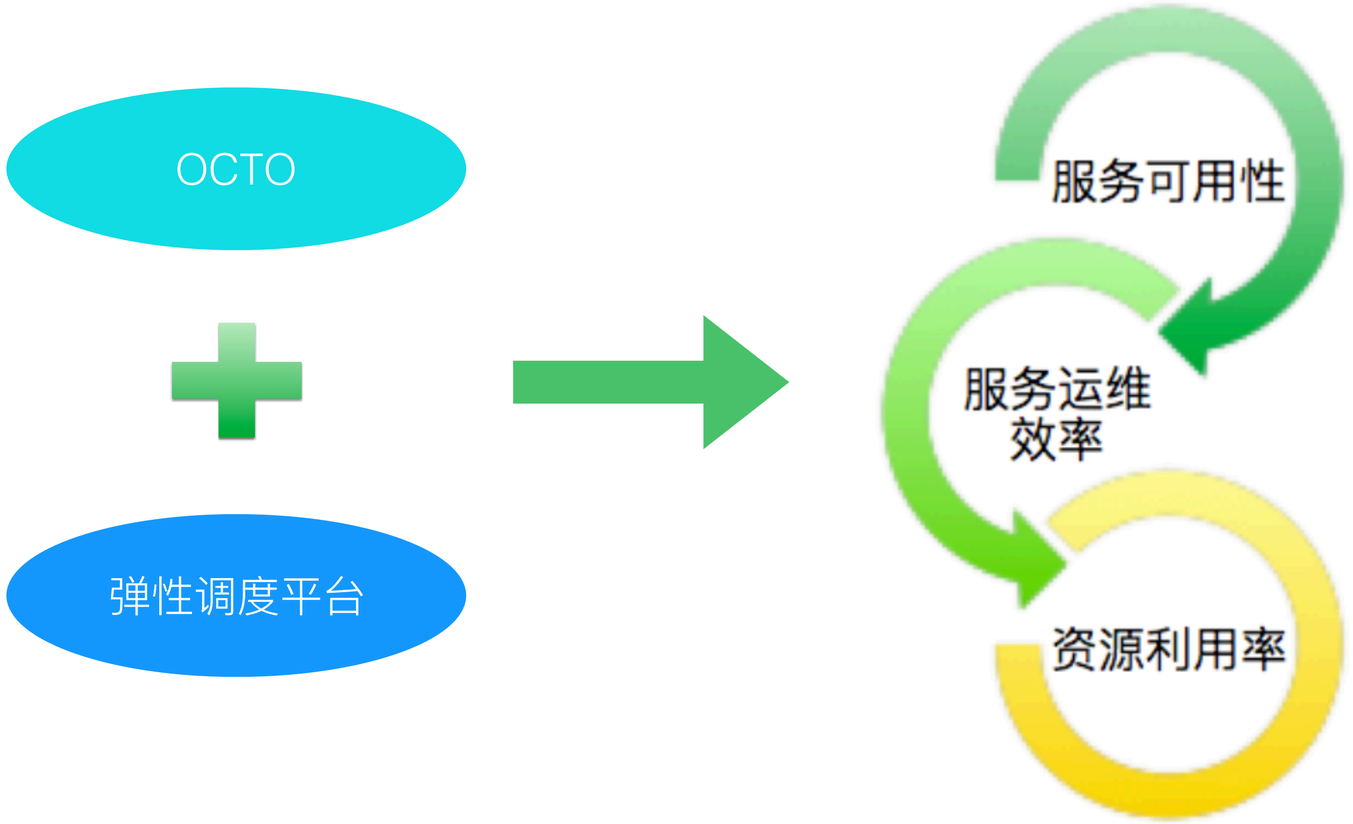
## Management：支持RD自助管理服务，节点权重、路由分组、配置管理

# 全生命周期

Optimize：以报表形式输出，SLA指标、可用性、资源分布及利用率

| Appkey | 服务状态 | 服务可用性 | 节点存活率(thrit,http) | QPS/同比/环比 | TP90/同比/环比 |
|---|---|---|---|---|---|
| com.sankuai.inf.logCollector | GOOD | 100% | 97.3209%, NaN | 20641 / -1% / -4% | 0 / 0% / 0% |
| com.sankuai.inf.data.statistic | GOOD | 100% | 98.5392%, NaN | 9136 / 5% / -3% | 1 / 0% / 0% |
| com.sankuai.inf.mafka.castlehotel.heartbeat | WARN | 99.9999% | NaN, NaN | 7468 / 100% / -1% | 0 / 0% / 0% |
| com.sankuai.inf.mnsc | GOOD | 99.9997% | 99.9752%, 100% | 6200 / -6% / -0% | 5 / 25% / 0% |
| com.sankuai.inf.mafka.castleocto.heartbeat | GOOD | 100% | NaN, NaN | 5907 / -1% / 1% | 0 / 0% / 0% |

# OCTO - 未来演进



# 回顾与总结

- 美团服务架构演进历程
- OCTO研发要点：代理模式、状态检查、策略下移、框架优化
- OCTO治理实践：全生命周期

关注QCon微信公众号，
获得更多干货！

# Thanks!

INTERNATIONAL SOFTWARE DEVELOPMENT CONFERENCE

主办方 Geekbang. 极客邦科技 InfoQ