Geekbang. 极客邦科技 InfoQ

QCon 全球软件开发大会【北京站】2016

平台系统的演变

腾讯 孙子荀

Intemational Software Development Conference

自我介绍

个人信息：

Linux内核，在分布式系统 ，机器学习(并行计算) 上有一定经验。

主要经历：

- 华为/亚信
- 百度
- **2012.** 腾讯

•主导项目： QQ群广告/旋风下载技术2.0/ QQ公众号 后台负责人

1 平台系统

2 防护
接入平面

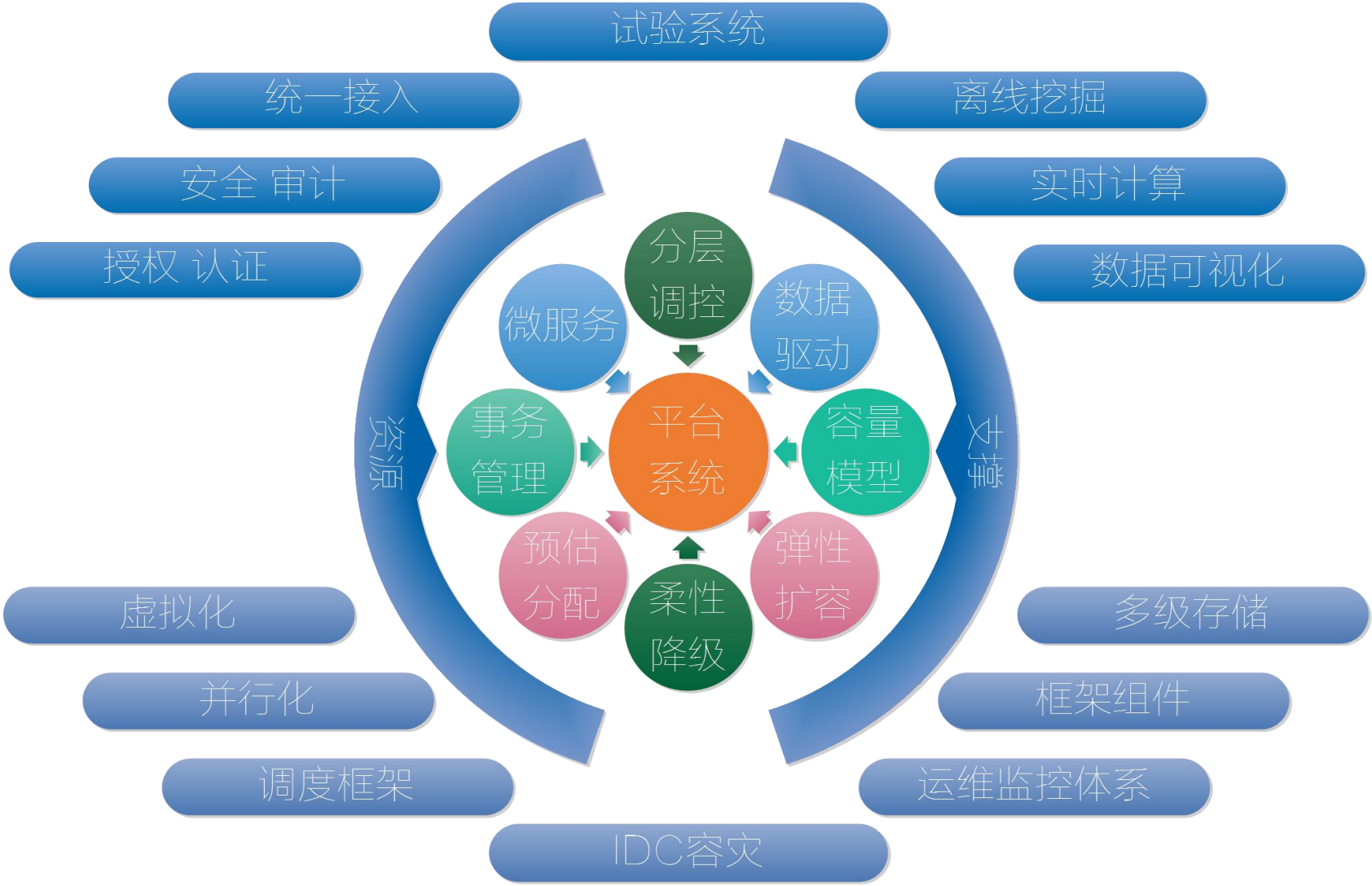
3 建造
框架与调度

4 对抗
容量模型

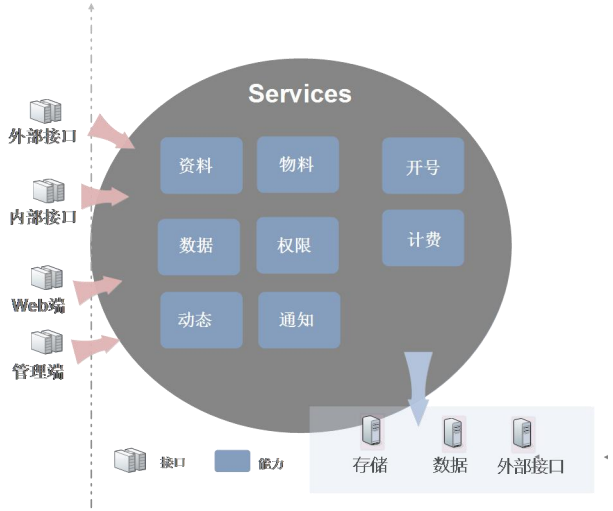
5 扩张
容灾选型

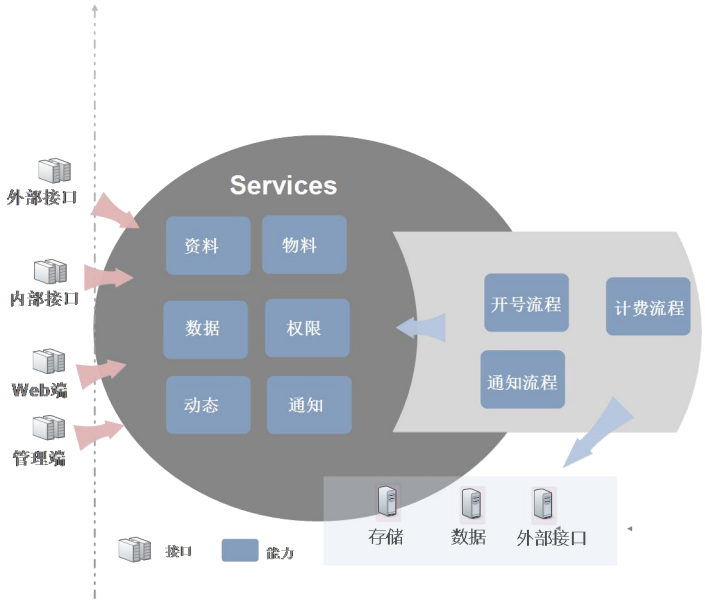
6 循环
数据驱动的闭环设计

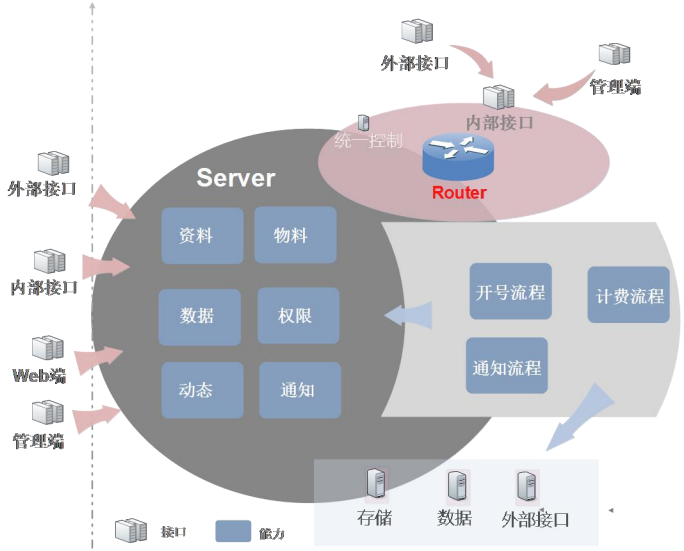
平台系统

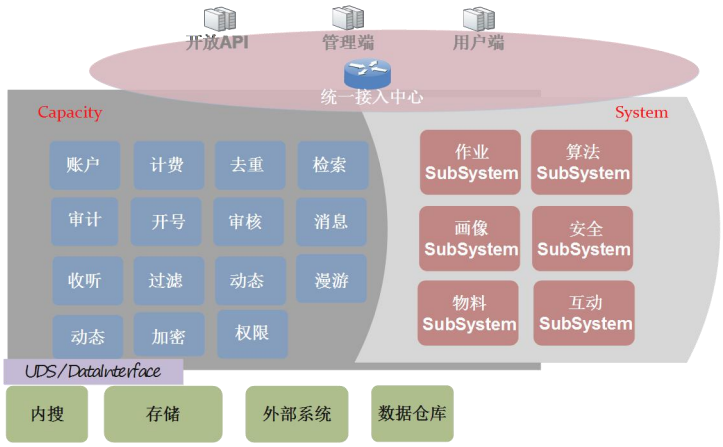

试验系统
统一接入
离线挖掘
安全 审计
实时计算
授权 认证
数据可视化
分层调控
微服务
数据驱动
事务管理
平台系统
容量模型
预估分配
弹性扩容
柔性降级
资源
对外
虚拟化
多级存储
并行化
框架组件
调度框架
运维监控体系
IDC容灾

平台系统


Services
外部接口
内部接口
Web端
管理端
资料
物料
开号
数据
权限
计费
动态
通知
存储
数据
外部接口
接口
能力


Services
外部接口
内部接口
Web端
管理端
资料
物料
数据
权限
动态
通知
开号流程
计费流程
通知流程
存储
数据
外部接口
接口
能力


外部接口
管理端
内部接口
统一控制
Router
Server
外部接口
内部接口
Web端
管理端
资料
物料
数据
权限
动态
通知
开号流程
计费流程
通知流程
存储
数据
外部接口
接口
能力


开放API
管理端
用户端
统一接入中心
Capacity
System
账户
计费
去重
检索
审计
开号
审核
消息
收听
过滤
动态
漫游
动态
加密
权限
作业 SubSystem
算法 SubSystem
画像 SubSystem
安全 SubSystem
物料 SubSystem
互动 SubSystem
UDS/DataInterface
内搜
存储
外部系统
数据仓库

接入平面

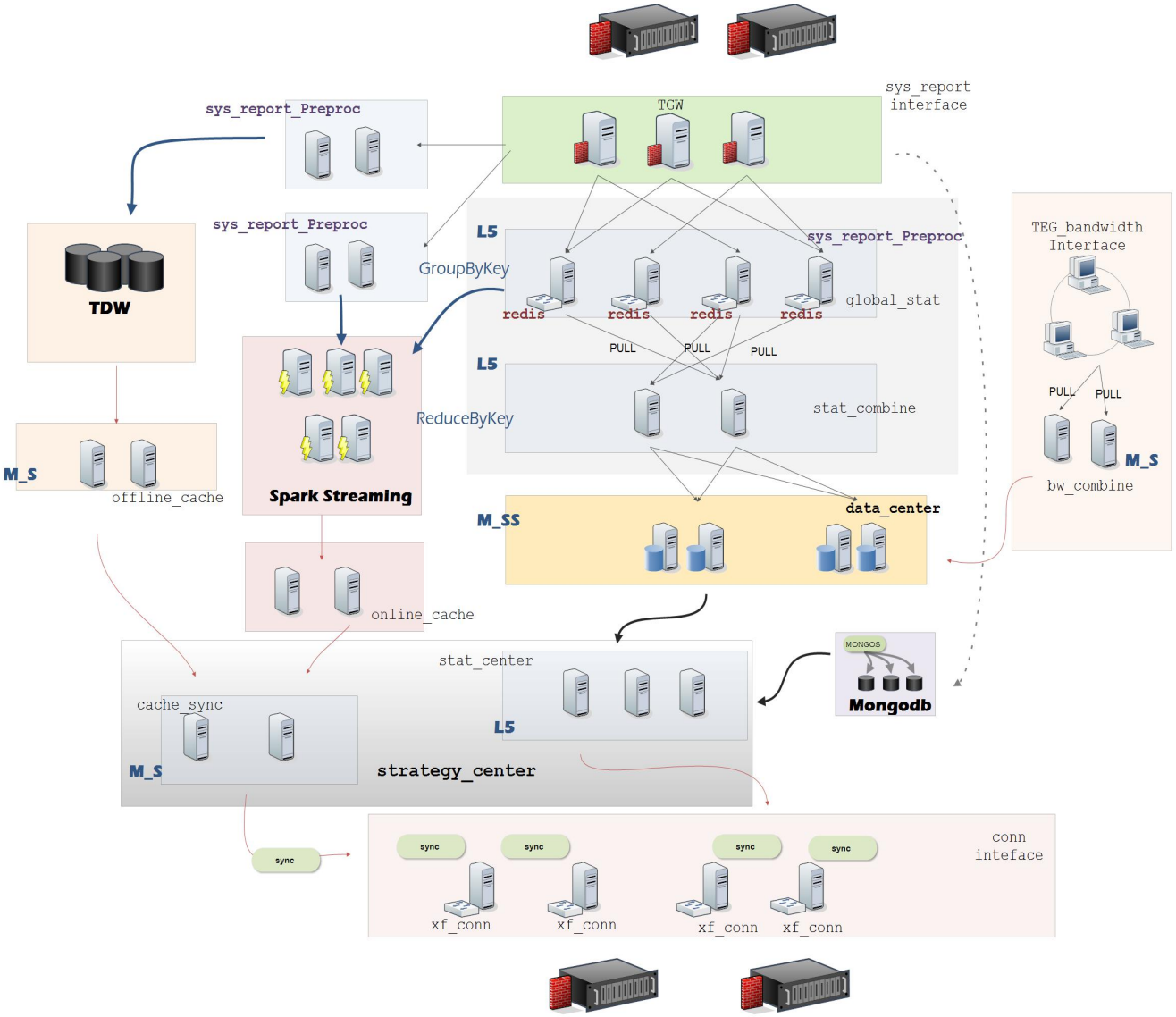
1 接入平面的分类

2 微服务路由

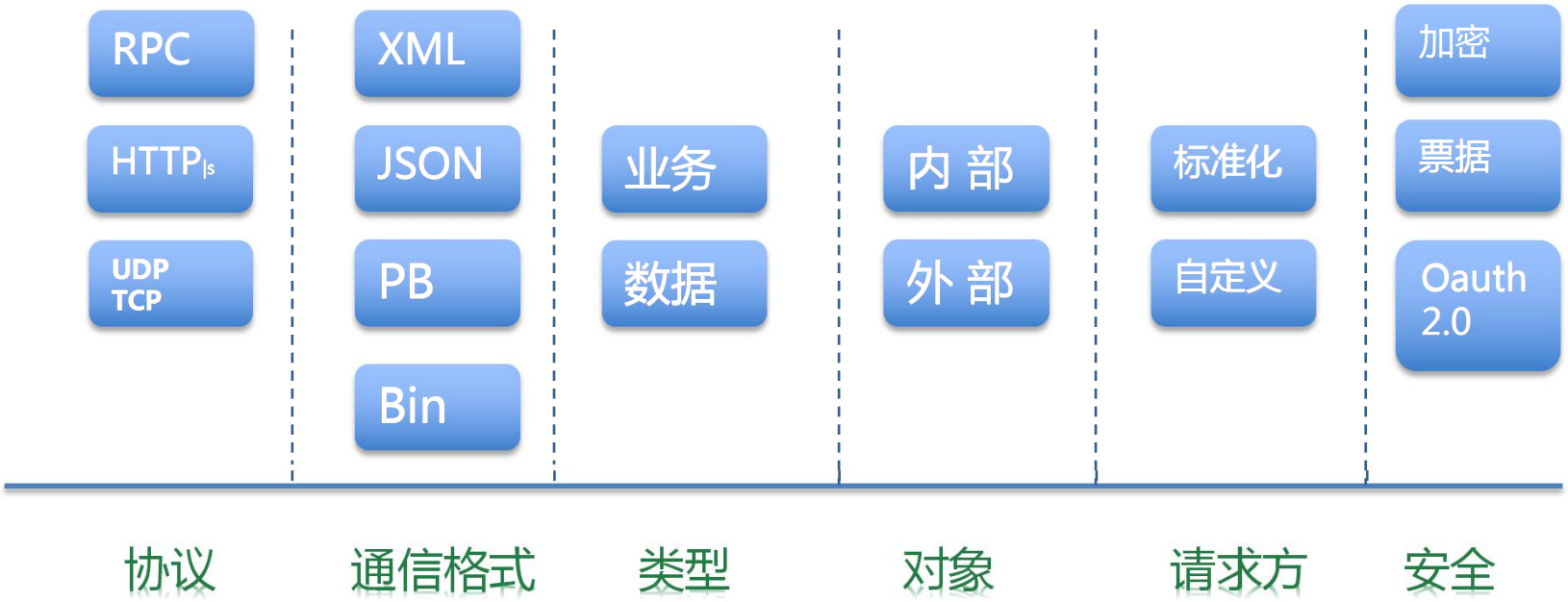
3 权限与频控

4 全网流量调度

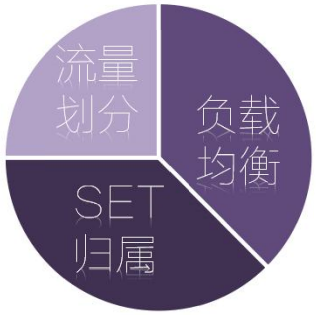
接入平面 | 接入的分类


sys_report_Preproc
TGW
sys_report
interface
sys_report_Preproc
L5
sys_report_Preproc
GroupByKey
TEG_bandwidth
Interface
TDW
redis
redis
redis
redis
global_stat
PULL
PULL
PULL
L5
ReduceByKey
stat_combine
PULL
PULL
M_S
M_S
offline_cache
Spark Streaming
bw_combine
M_SS
data_center
online_cache
MONGOS
stat_center
Mongodb
cache_sync
L5
M_S
strategy_center
sync
sync
sync
sync
sync
conn
inteface
xf_conn
xf_conn
xf_conn
xf_conn

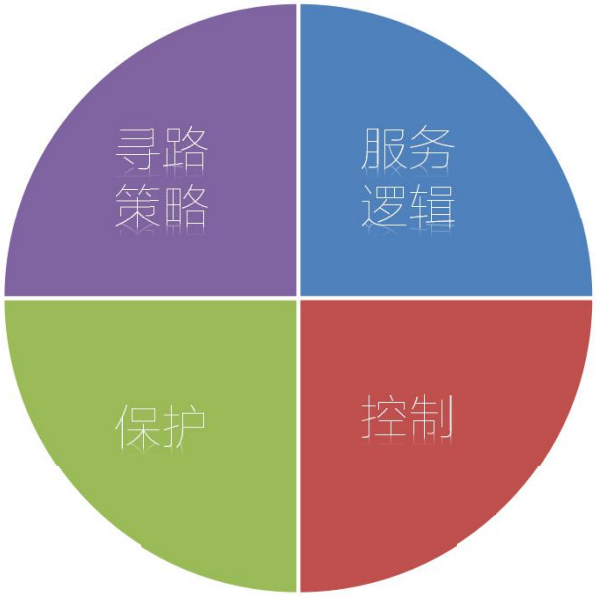
接入平面 | 微服务与路由


RPC
HTTPs
UDP
TCP
XML
JSON
PB
Bin
业务
数据
内 部
外 部
标准化
自定义
加密
票据
Oauth 2.0
协议
通信格式
类型
对象
请求方
安全

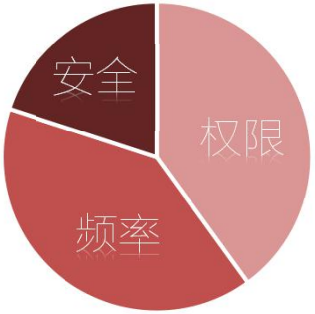
接入平面 | 微服务与路由


流量
划分
负载
均衡
SET
归属

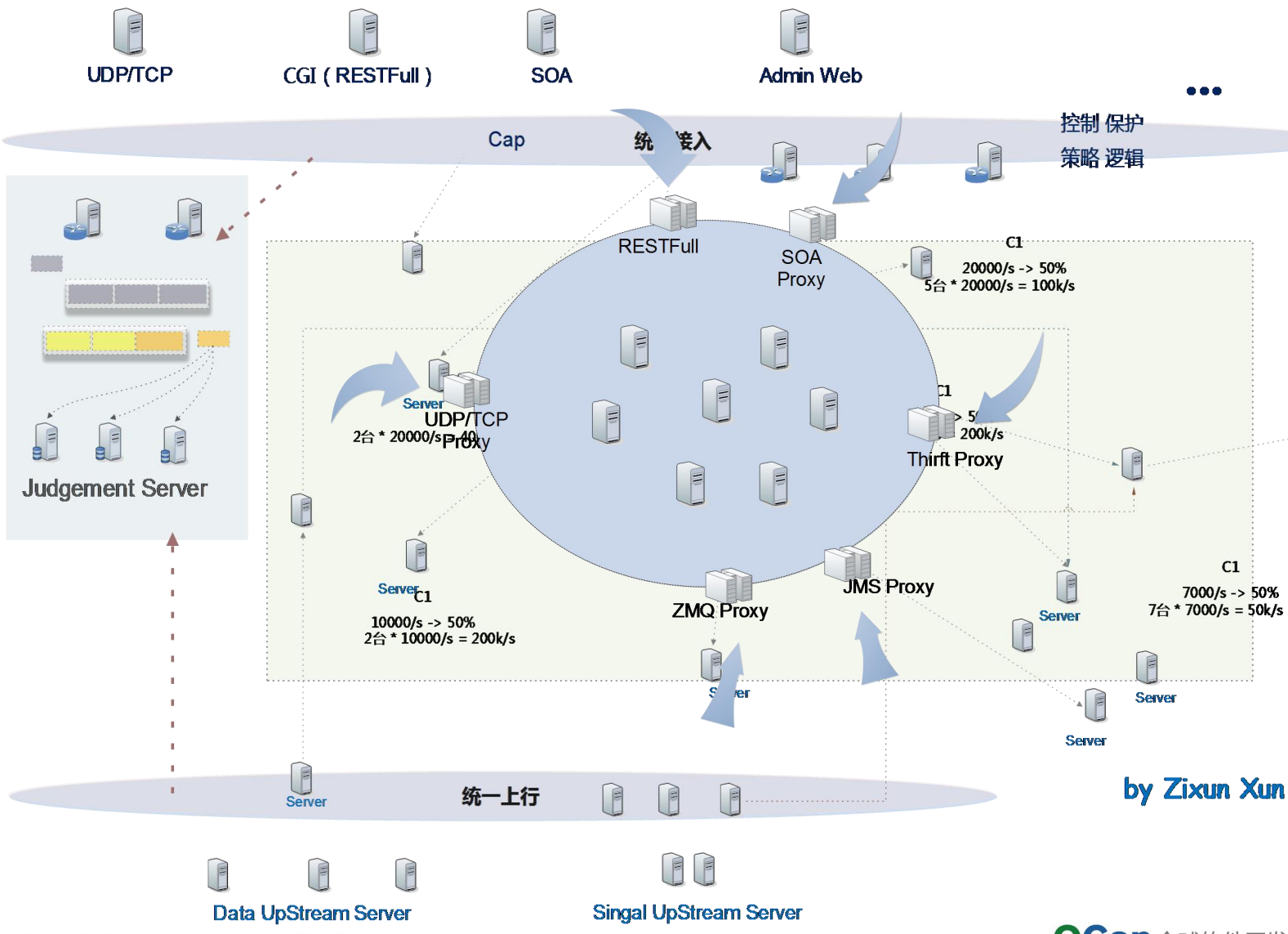

服务
寻址
协议
转换
会话登记


寻路
策略
服务
逻辑
保护
控制


有损
降级
服务
QPS
防雪崩

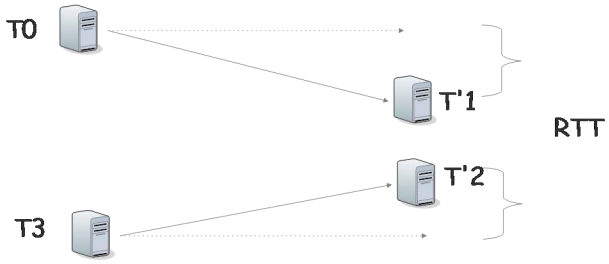

安全
权限
频率

接入平面 | 权限与频控


UDP/TCP
CGI (RESTFull)
SOA
Admin Web
...
Cap
统一接入
控制 保护
策略 逻辑
RESTFull
SOA
Proxy
C1
20000/s -> 50%
5台 * 20000/s = 100k/s
Server
UDP/TCP
Proxy
2台 * 20000/s = 40k/s
Judgement Server
Thirft Proxy
200k/s
Server
C1
10000/s -> 50%
2台 * 10000/s = 200k/s
ZMQ Proxy
JMS Proxy
Server
C1
7000/s -> 50%
7台 * 7000/s = 50k/s
Server
Server
Server
统一上行
by Zixun Xun
Data UpStream Server
Singal UpStream Server

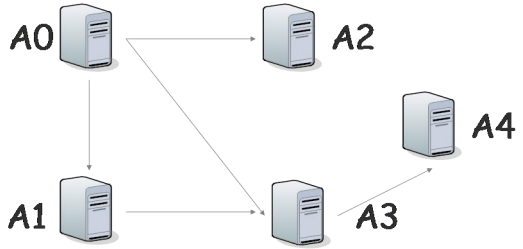
主机时间不同步

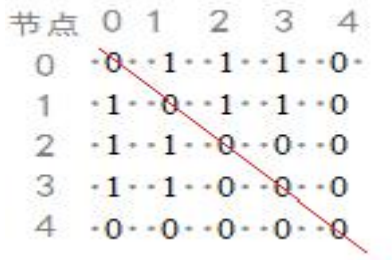
滑动平均
自回归模型
神经网络



网络延迟问题

DC分区问题





	A0	A1	A2	A3	A4
A0	0	1	1	1	0
A1	1	0	0	1	0
A2	1	0	0	0	0
A3	1	1	0	0	1
A4	0	0	0	1	0

3 2 1 3 1

A0

	A0	A1	A2	A3	A4
A0	0	1	1	1	0
A1	0	0	0	0	0
A2	0	0	0	0	0
A3	0	0	0	0	0
A4	0	0	0	1	0

0 1 1 2 0

A0 A3

	A0	A1	A2	A3	A4
A0	0	0	0	0	0
A1	0	0	0	0	0
A2	0	0	0	0	0
A3	0	0	0	0	0
A4	0	0	0	0	0

0 0 0 0 0

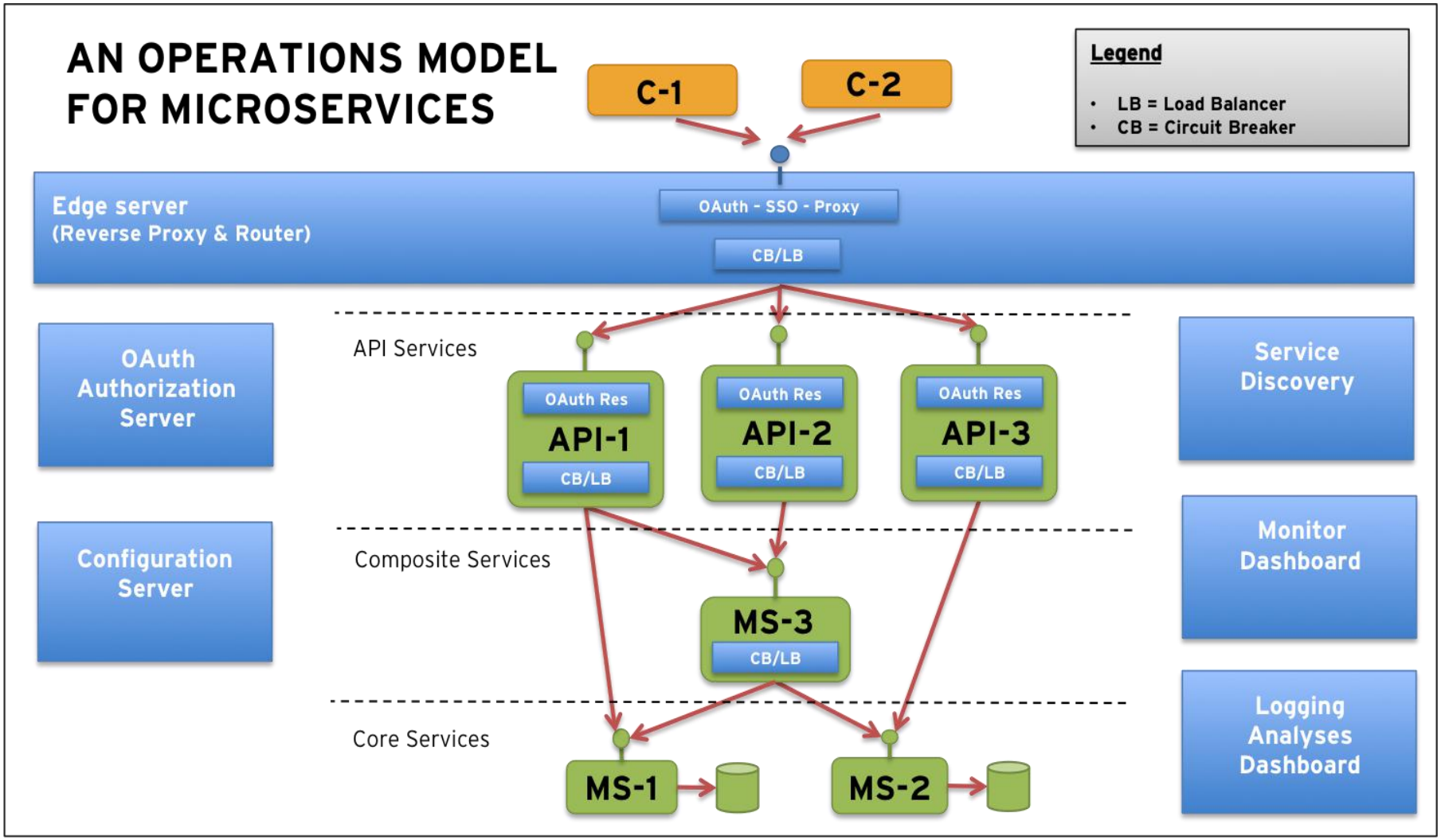
A0 A3

框架与调度

1 服务框架

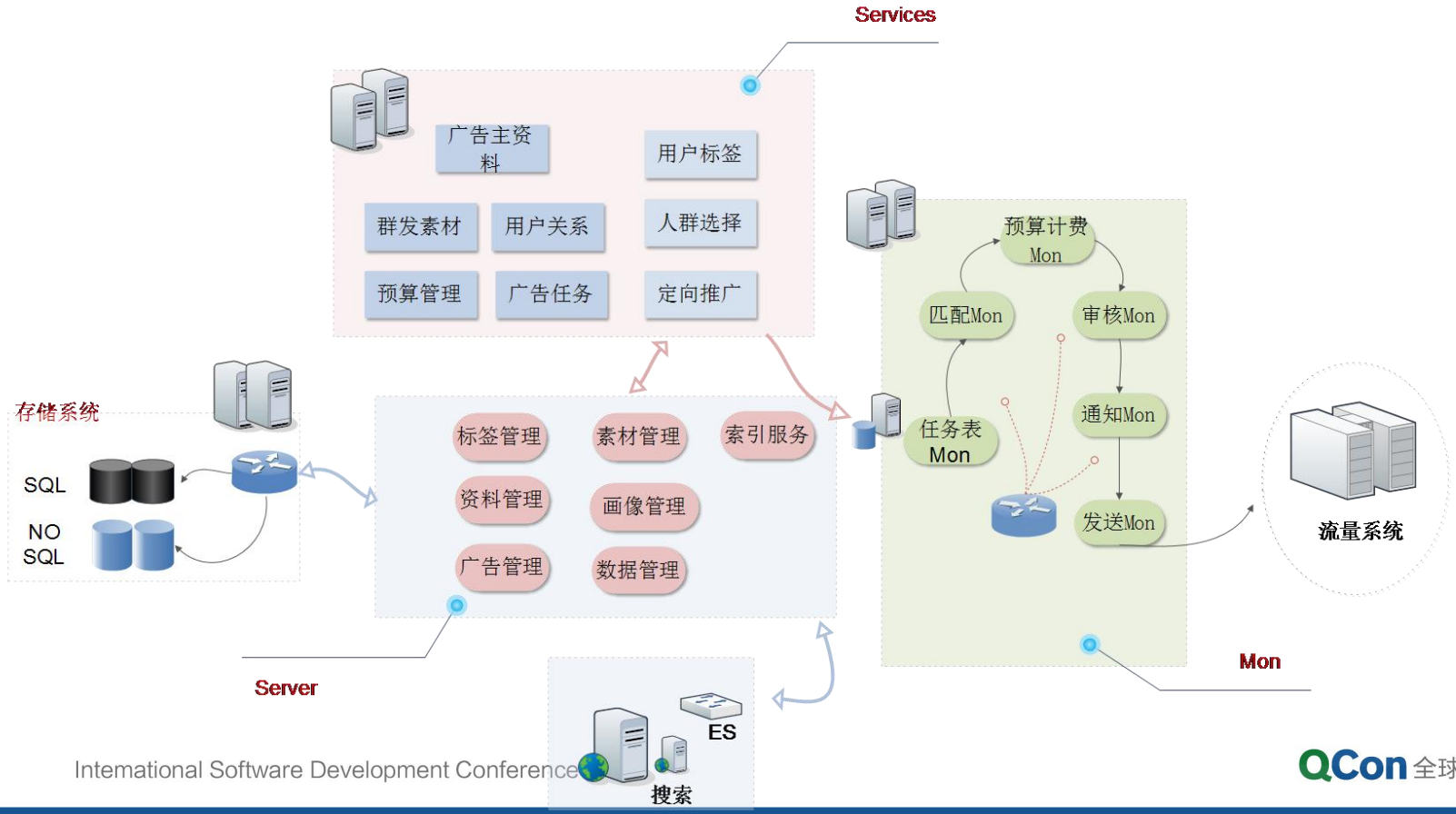
2 调度

3 虚拟化技术


AN OPERATIONS MODEL FOR MICROSERVICES
C-1
C-2
Legend
LB = Load Balancer
CB = Circuit Breaker
Edge server
(Reverse Proxy & Router)
OAuth - SSO - Proxy
CB/LB
OAuth Authorization Server
Configuration Server
API Services
OAuth Res
API-1
CB/LB
OAuth Res
API-2
CB/LB
OAuth Res
API-3
CB/LB
Service Discovery
Monitor Dashboard
Logging Analyses Dashboard
Composite Services
MS-3
CB/LB
Core Services
MS-1
MS-2

调度与分配 | 服务框架

	特点	框架	模型
Service	对外的能力展现	Async / Coroutine	RestFull
Server	内部的能力封装	Leader-Fellower	MicoServer
Mon	作业任务	Reactor/ Proactor	SOA

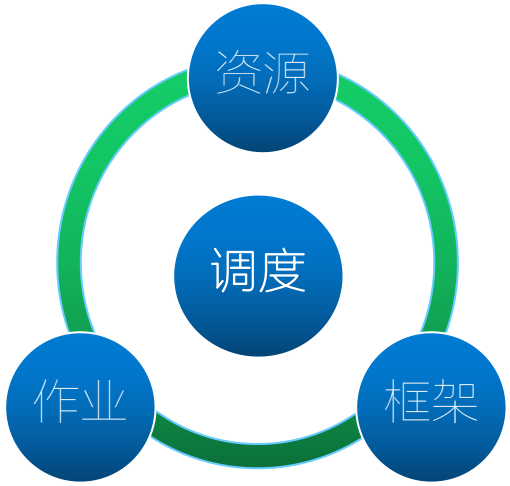


单机调度：

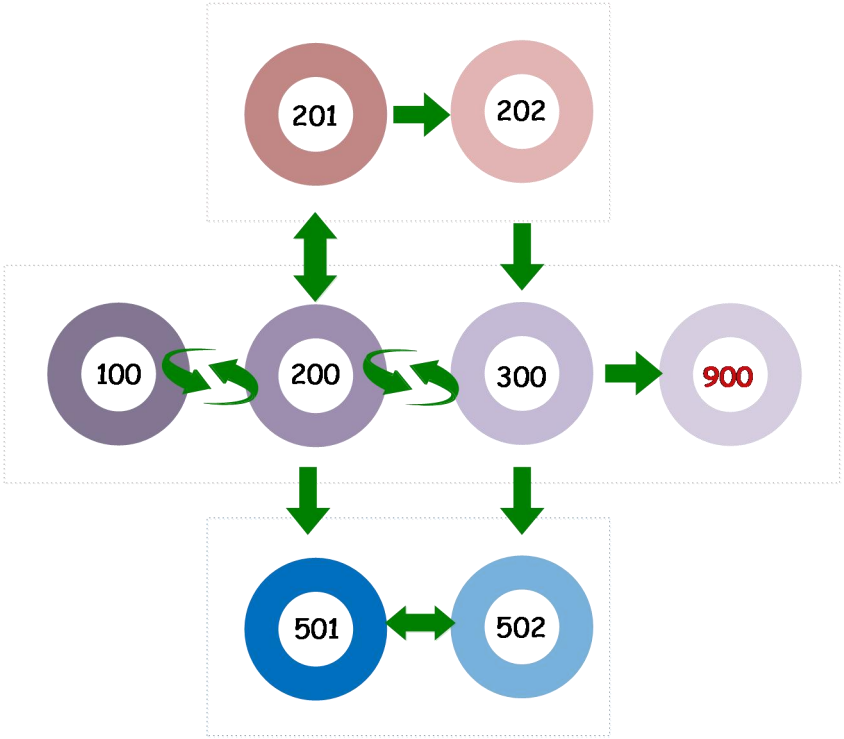
Liunx Process Scheduer
Liunx IO Scheduer

分布式：

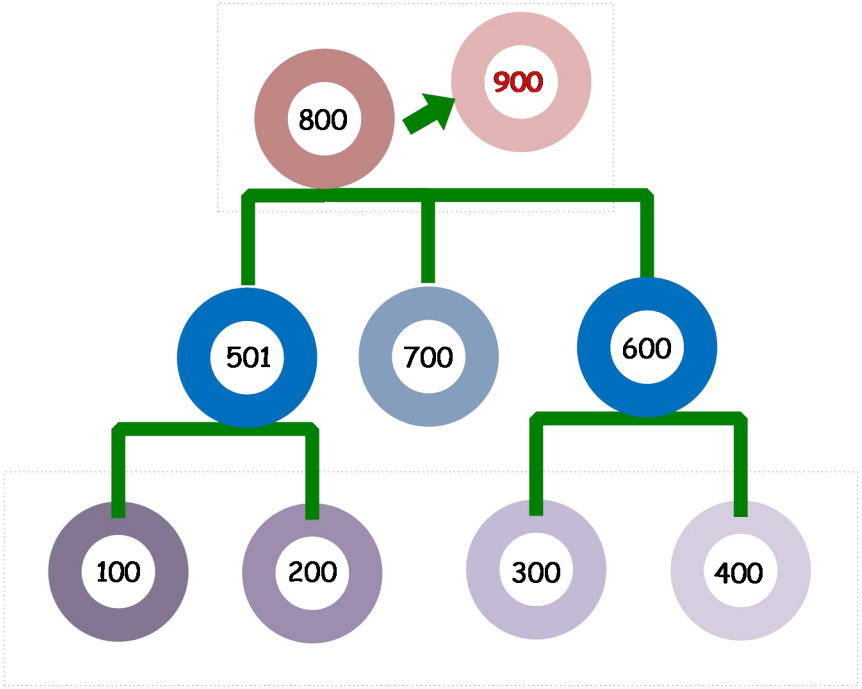
maui torque
mesos
yarn



	资源	框架	作业单元	调度
Linux	CPU时间	Scheduer	Process	CFS
Mesos	集群硬件资源	MR/Spark	并行化算子	DRF
Ads	empression	RTB	Request	eCPM



Filter模式状态机
串行



Switch模式 状态机
并行

资源：

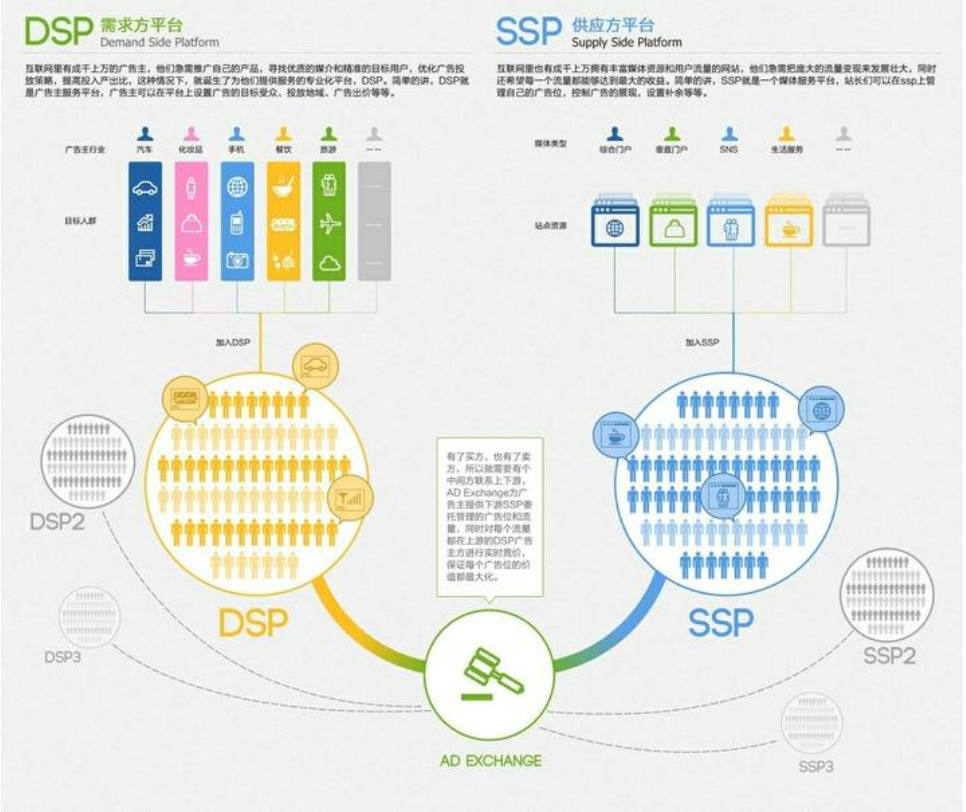
硬件(CPU ,内存,外存, 网络) ->Cloud Compution
带宽 -> CDN
流量 -> A D X

类别：

实时决策 (L2R HWM 贪心)
离线决策 (运筹学)
分类，预测

决策依据：

三方最优
平台， 需求方，资源供给方



容量模型

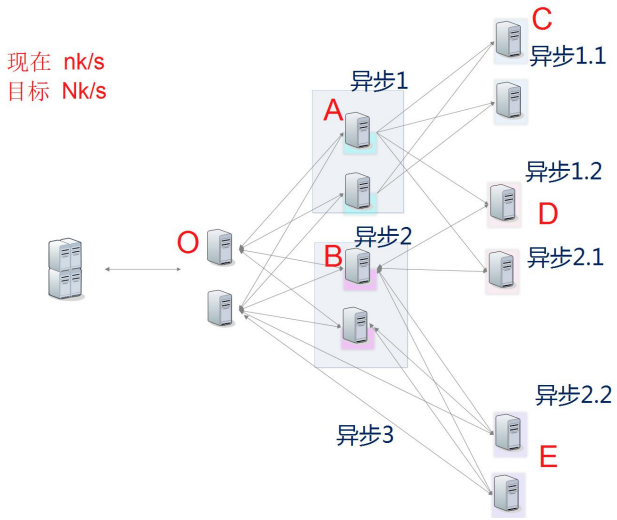
1 一切尽在掌控

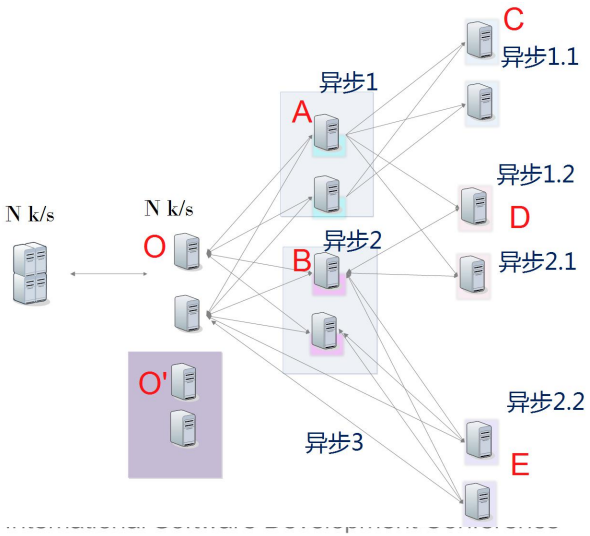
2 压测与扩容

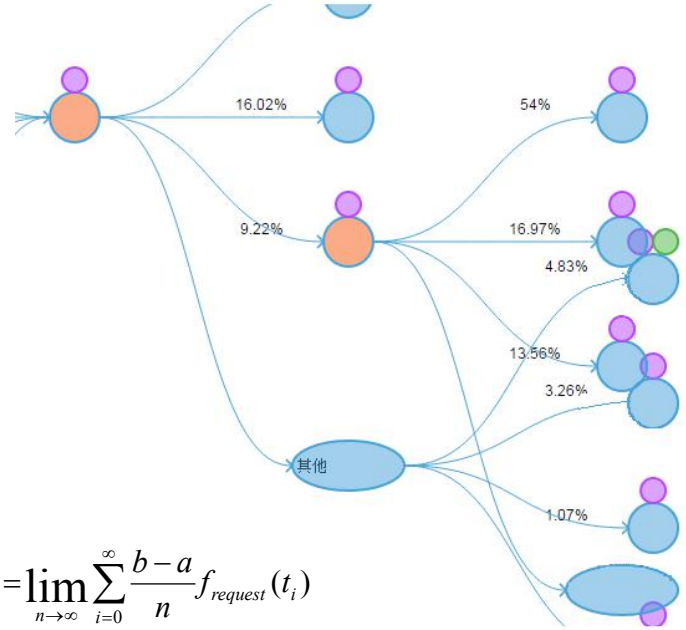
3 试验系统

4 SET化部署

容量模型 | 压测与扩容







$$\int_a^b f_{request}(x)dx = \lim_{n\to\infty}\sum_{i=0}^{\infty}\frac{b-a}{n}f_{request}(t_i)$$

$$\int_{a+t}^{b+t} f_{response}(x)dx = \lim_{n\to\infty}\sum_{i=0}^{\infty}\frac{b-a}{n}f_{response}(t_i)$$

$1s \Rightarrow L$

$1s = \sum T_{Brust}$

$J \Rightarrow T_{Brust}$

$All_{except} = J \bullet 1/T_{Brust}$

$$L = All_{reality} = J \bullet 1/(T_{Brust} + 2T_{rtt/2} + T_1 + T_2 + T..)$$

$$T_1 = f(Num_1, 1/P_1, T_{rtt})$$

$$Num_i = f(J/NodeNum, k)$$

$Num_i \propto k$

$$\sum_i^k Num_i = J$$

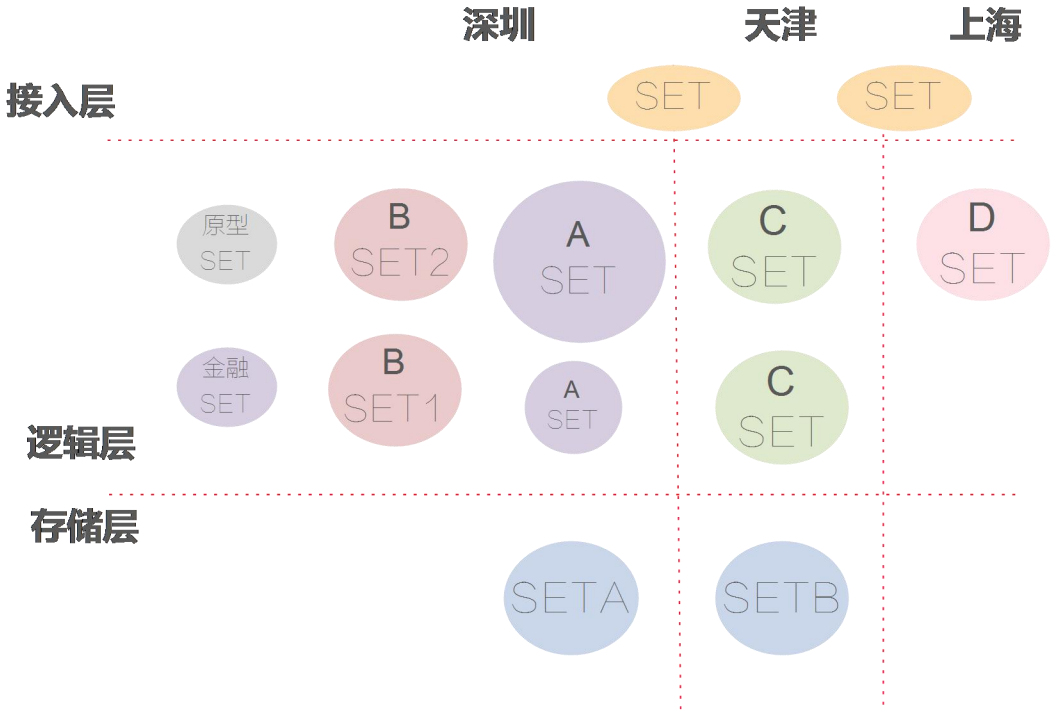
容量模型 | SET化部署

要求：

标准化、规模化、模块化

粒度：

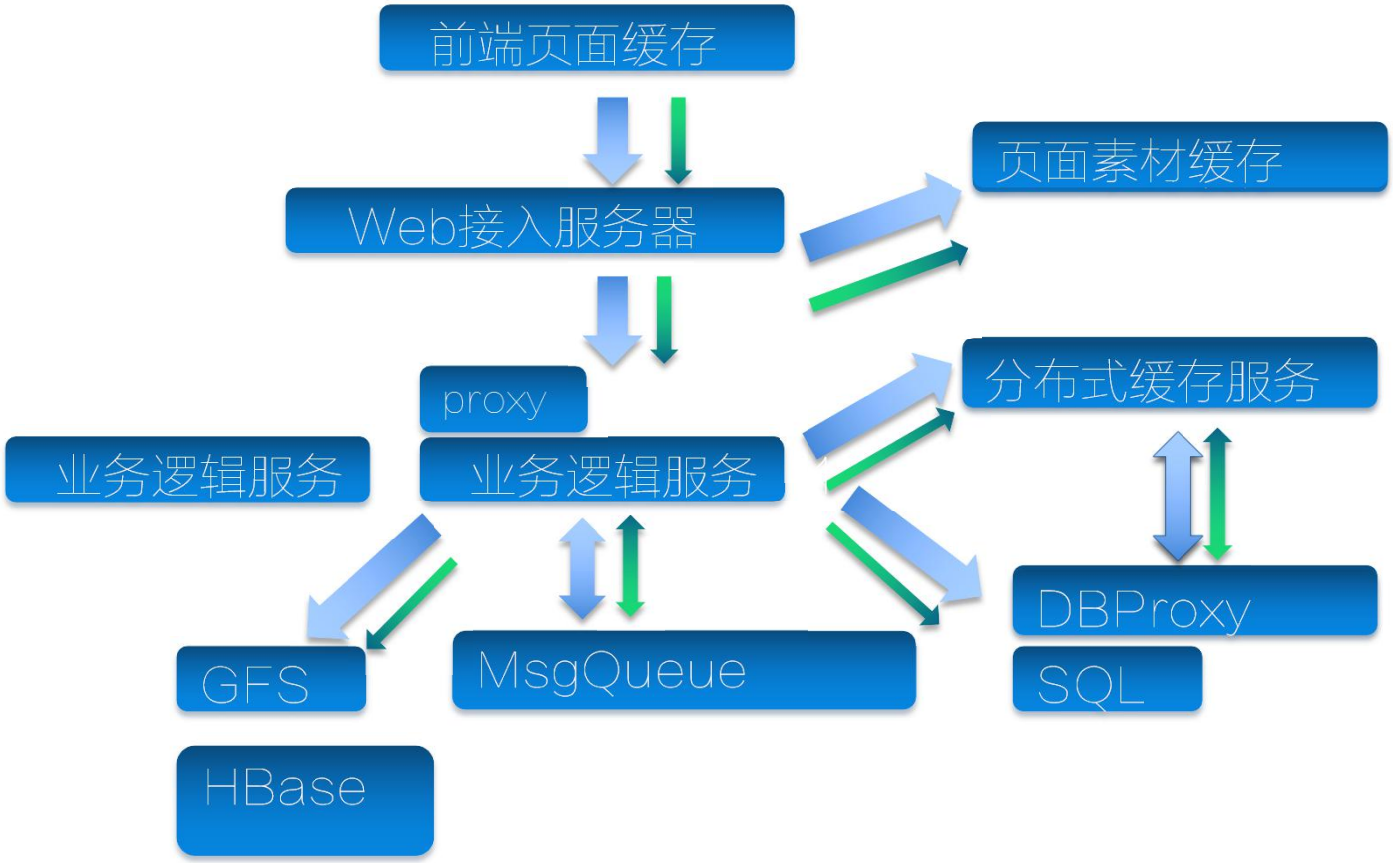
1） 承载 容量
2） 打包 业务
3） 独立子系统



地域	层级	功能	序号
天津	逻辑层	基础信息	C

步骤：

- 灰度策略
- 流量切分
- 部署升级
- 观察验证
- 全量上线

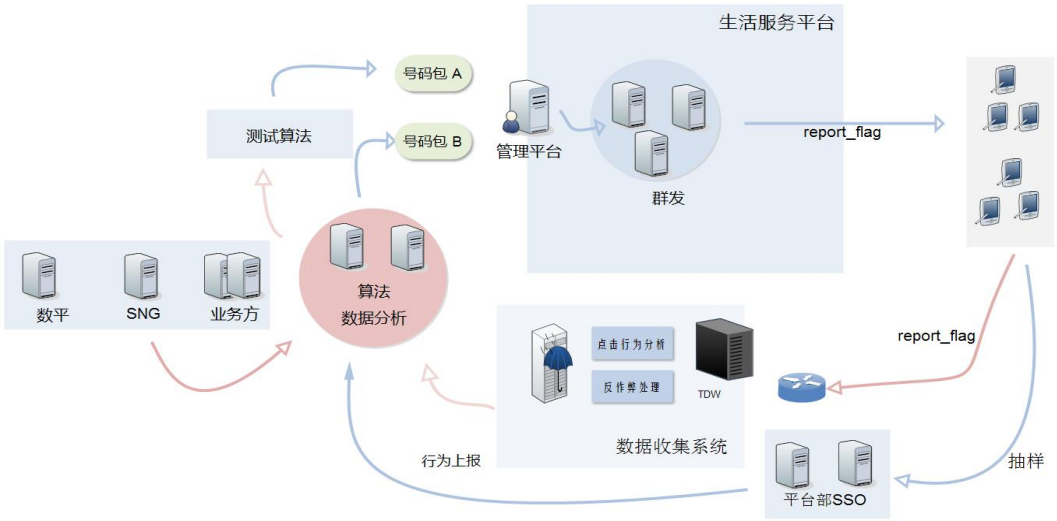


系统层面：

分层流量标记
动态策略

数据层面：

统计学效果检验
多因素方法分析
置信度抽样数
辛普森悖论



容灾

方案比较

容灾选型 | 方案比较

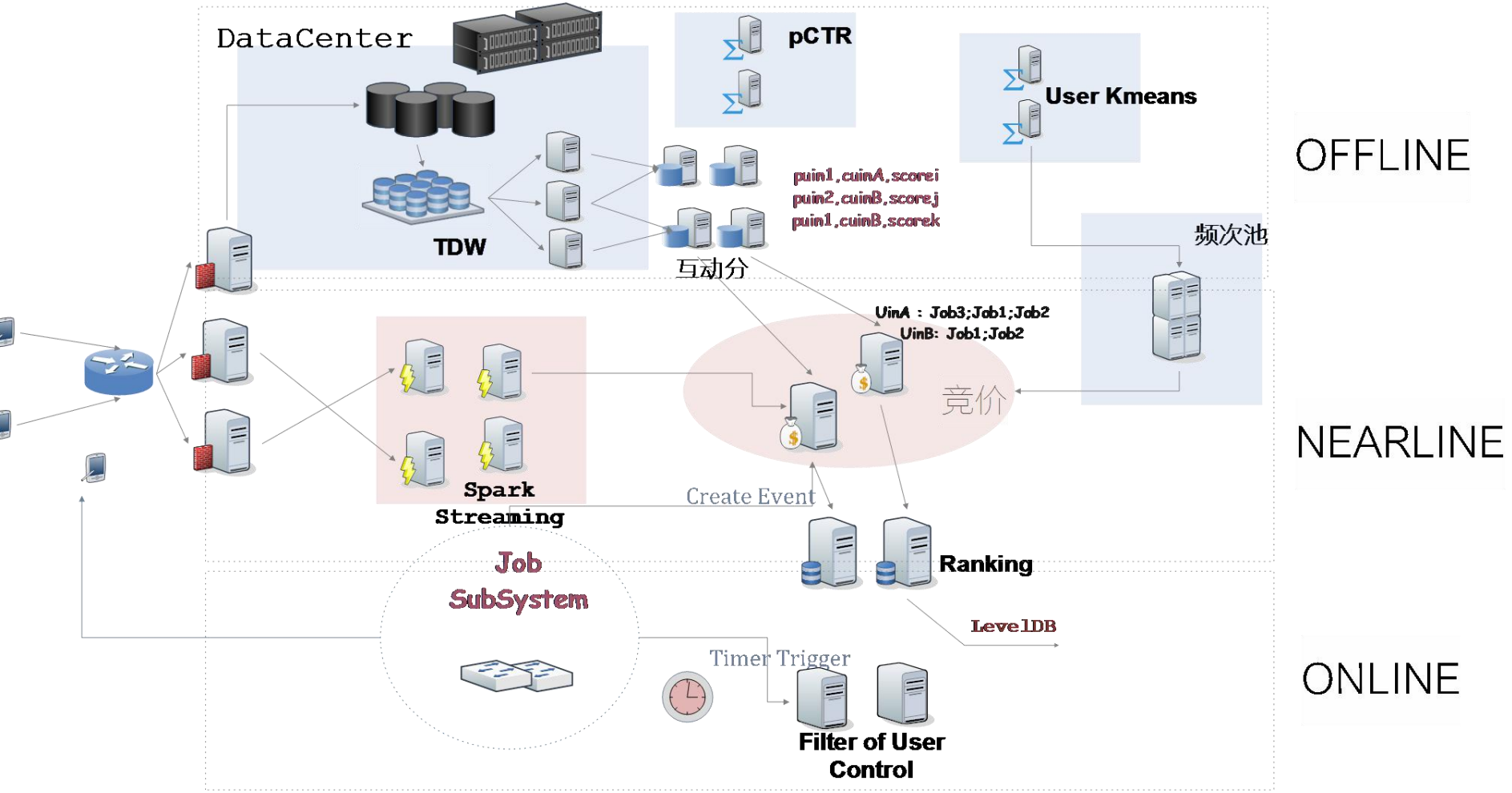
系统模型	接入/逻辑层	存储层	方法	标准
作业平台系统	Check Point	分布式文件系统 P2P网络	回退恢复：(backward recovery) 系统从当前错误的状态回到先前正确的状态	提交作业成功执行
存储平台系统	接入层容灾	多副本机制	一致性模型	C A P
推送平台系统	分层可靠协议	多副本机制	协议+备份	短信平台Mo Mt 入网标准 72小时 内有接受条件到达率： 99.99% 及时率 >= 95%
金融平台系统	事务性操作 幂等	可靠存储	多级流水对账 严格事务性	可溯源 回退 有据可查 ACID

数据闭环

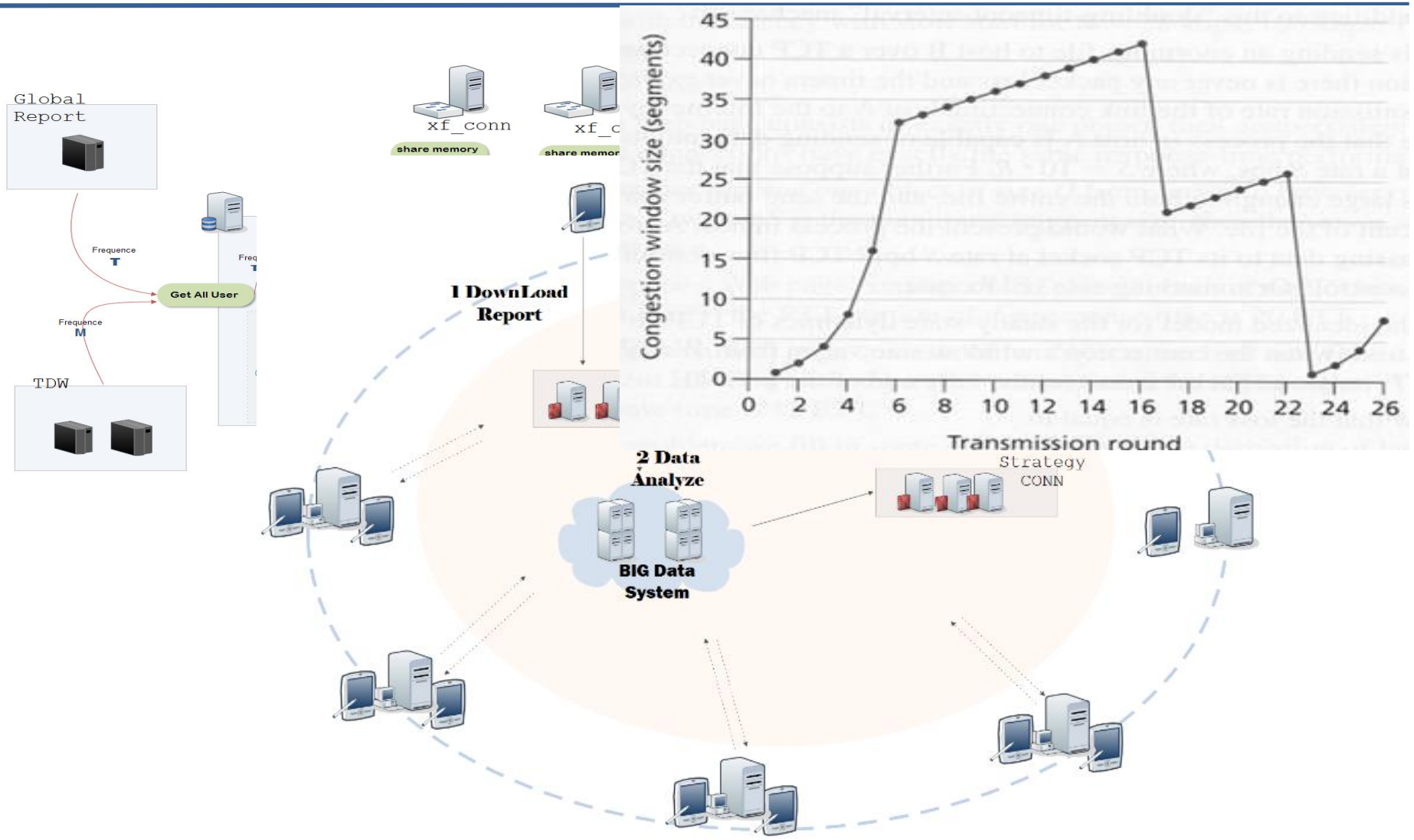
1 数据分层

2 数据驱动

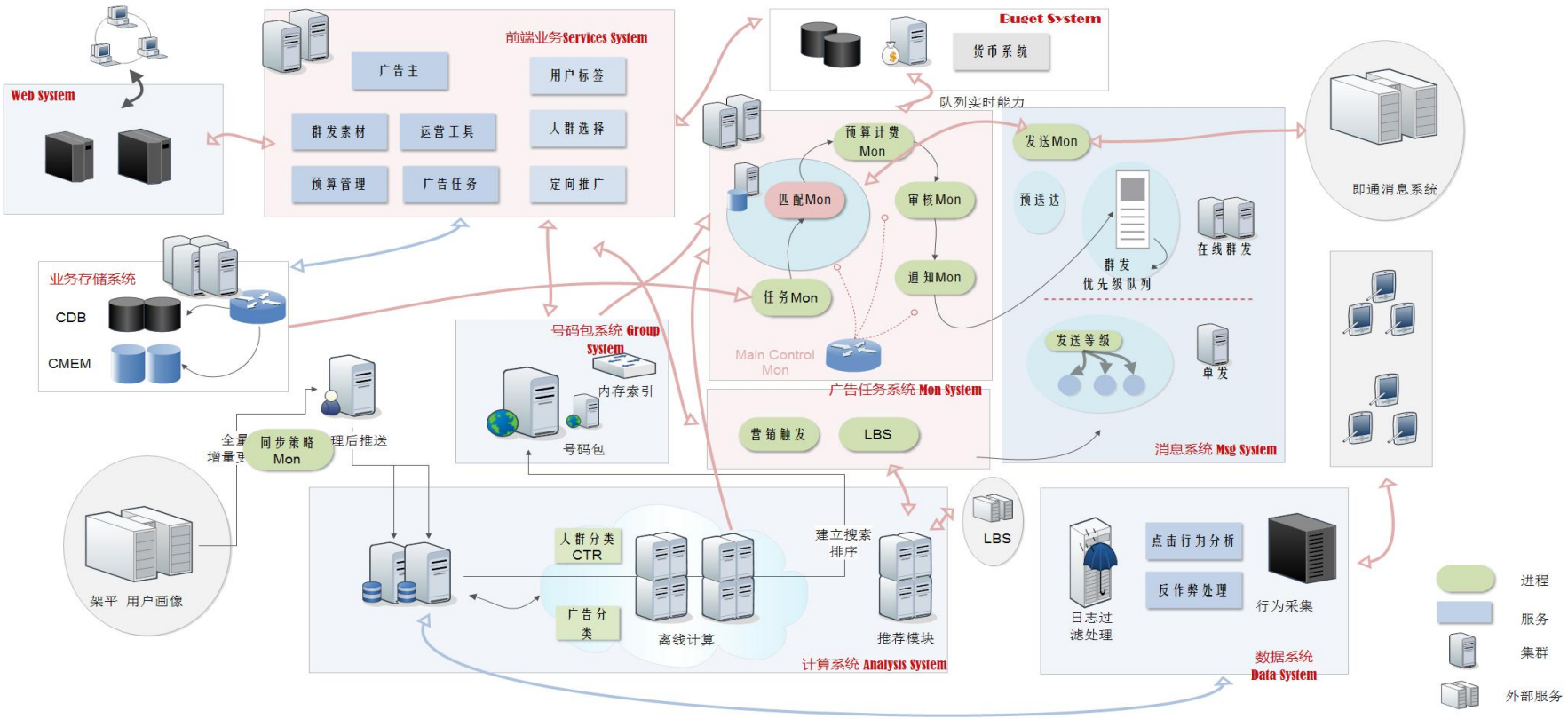
数据闭环 | 数据分层


DataCenter
pCTR
User Kmeans
OFFLINE
TDW
puin1,cuinA,scorei
puin2,cuinB,scorej
puin1,cuinB,scorek
互动分
频次池
UinA : Job3;Job1;Job2
UinB: Job1;Job2
竞价
NEARLINE
Spark
Streaming
Create Event
Job
SubSystem
Ranking
LevelDB
Timer Trigger
ONLINE
Filter of User
Control

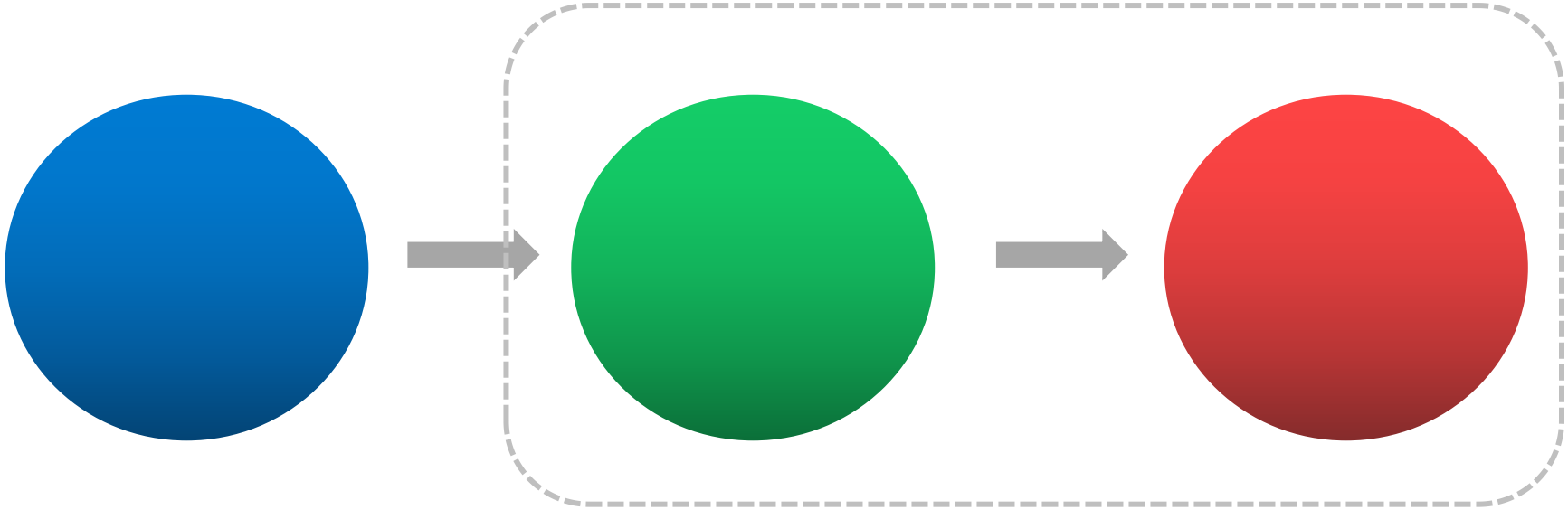
数据闭环 | 数据驱动


Global
Report
Frequence
T
Get All User
Frequence
M
TDW
xf_conn
share memory
xf_c
share memor
1 DownLoad
Report
2 Data
Analyze
BIG Data
System
Strategy
CONN
Congestion window size (segments)
Transmission round
0
5
10
15
20
25
30
35
40
45
0
2
4
6
8
10
12
14
16
18
20
22
24
26

数据闭环 | 数据驱动


Web System
前端业务Services System
广告主
用户标签
群发素材
运营工具
人群选择
预算管理
广告任务
定向推广
Buget System
货币系统
队列实时能力
即通消息系统
预算计费 Mon
匹配Mon
审核Mon
通知Mon
任务Mon
Main Control Mon
广告任务系统 Mon System
发送Mon
预送达
群发 优先级队列
在线群发
发送等级
单发
消息系统 Msg System
业务存储系统
CDB
CMEM
号码包系统 Group System
内存索引
号码包
营销触发
LBS
全量 增量更
同步策略 Mon
理后推送
架平 用户画像
人群分类 CTR
广告分类
离线计算
建立搜索排序
推荐模块
计算系统 Analysis System
LBS
日志过滤处理
点击行为分析
反作弊处理
行为采集
数据系统 Data System
进程
服务
集群
外部服务

平台系统



Geekbang 极客邦科技 InfoQ

THANKS!