开放搜索服务：从系统、平台到服务架构演变

阿里巴巴搜索事业部　　郭瑞杰 (花名：国泊)

2015年4月24日

促进软件开发领域知识与创新的传播

InfoQueue

ArchSummit
全球架构师峰会

[深圳] 2015年7月17日-18日

QCon
全球软件开发大会

[上海] 2015年10月15-17日

关注InfoQ官方微信
及时获取QCon演讲视频信息

阿里搜索系统架构演变

系统、平台、服务

搜索引擎基础数据结构

Inverted Index

Attribute
(FieldData)

Summary
(Stored Field)

Doc	Term
doc0	a, b, c, a
doc1	b, c, d
doc2	a, c, d
doc3	c, e

Term	
a	0<0,3>, 2<0>
b	0<1>,1<0>
c	0<2>,1<1>, 2<1>, 3<0>
d	1<2>,2<2>
e	3<1>

Per Doc:

Field	attribute
field1	1, 2, 4, 9
field2	abc, cde, a
field3	1.02
field4	true

Per Doc:

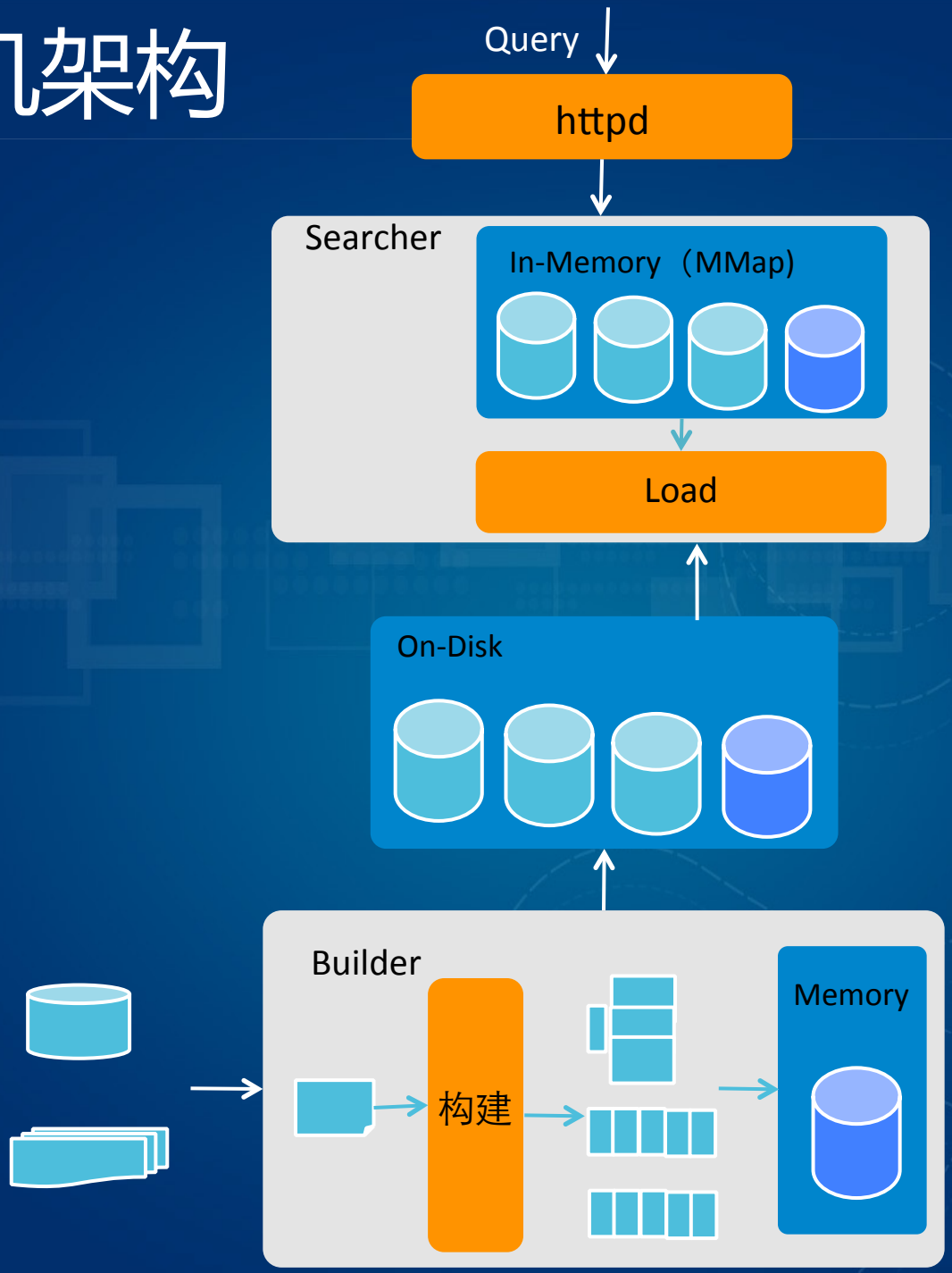
Field	attribute
field1	abc cde a
field2	aa bb cc
field3	a b c
field4	…

使用Inverted Index加速查询

使用Attribute过滤、分组统计、聚合

使用Summary做展示和摘要飘红

单机架构



全量索引构建

- 对所有文档重新构建索引，并替换旧索引数据

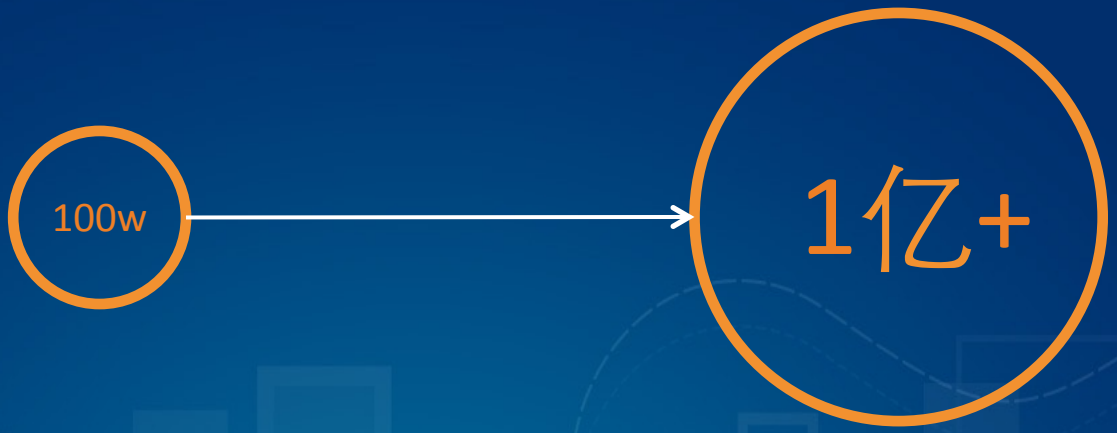
增量索引构建

- 对新增文档批量构建增量索引，合并至索引库中

实时索引更新

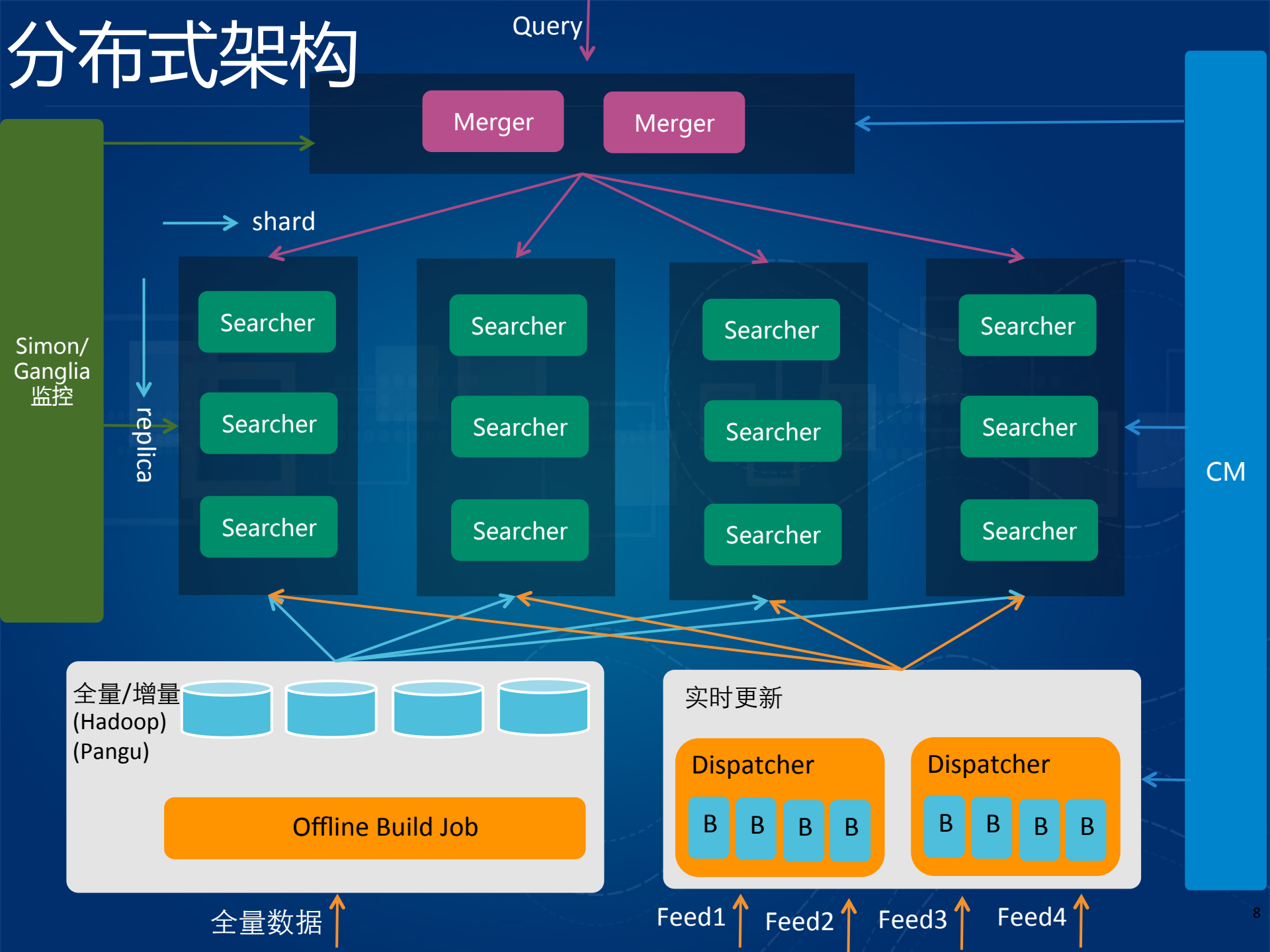
- Builder通过定时flush内存索引至磁盘，Searcher从磁盘Load新索引，实现分钟级或小时级更新时效性
- 性能较差

数据

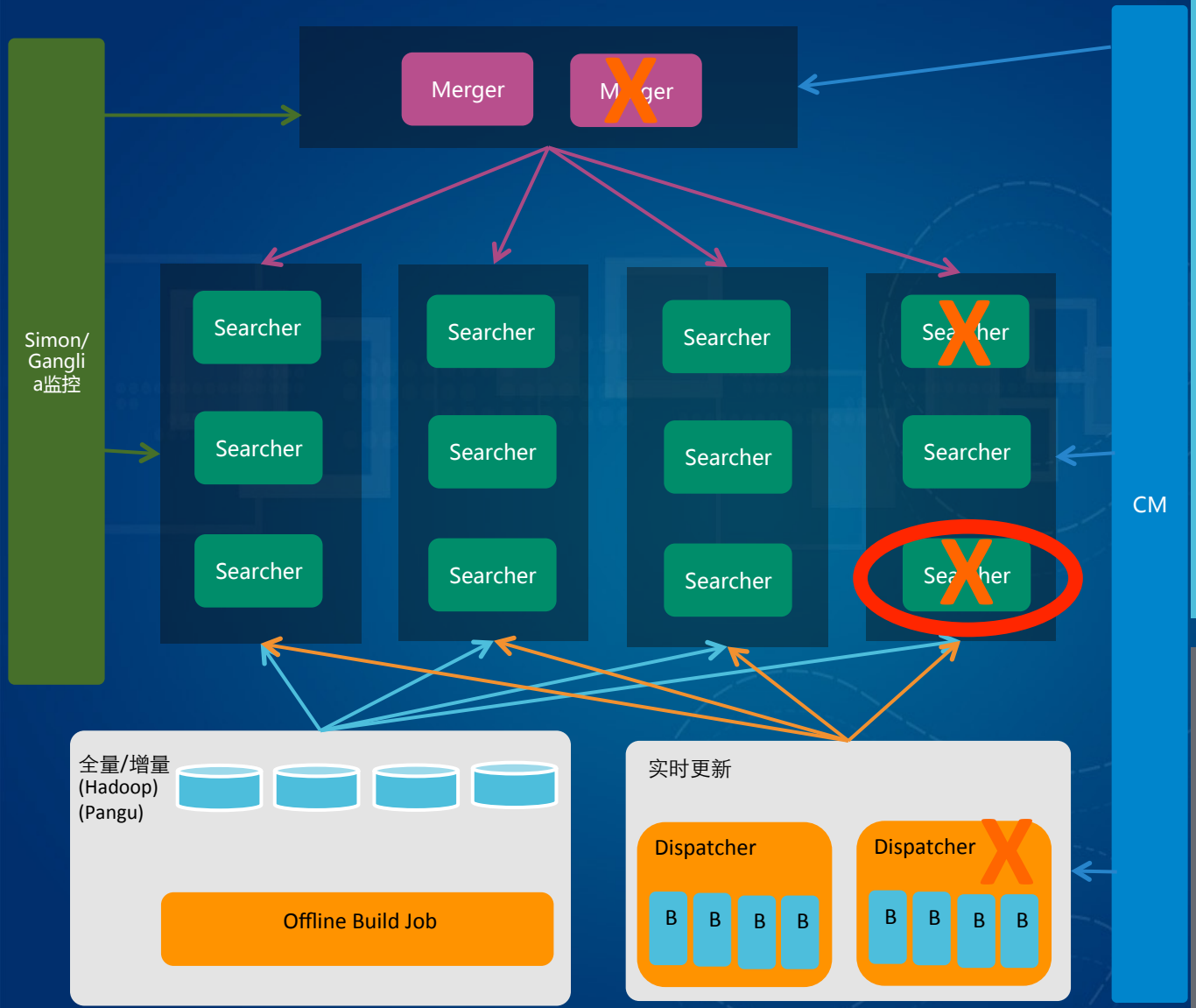


PV

分布式架构


Query
Merger
Merger
shard
Searcher
Searcher
Searcher
Searcher
Searcher
Searcher
Searcher
Searcher
Searcher
Searcher
Searcher
Searcher
Simon/
Ganglia
监控
replica
CM
全量/增量
(Hadoop)
(Pangu)
Offline Build Job
实时更新
Dispatcher
B
B
B
B
Dispatcher
B
B
B
B
全量数据
Feed1
Feed2
Feed3
Feed4

错误恢复



Merger进程挂掉

- CM发现后屏蔽
- 进程重启，向CM注册，恢复服务

Searcher进程挂掉

- CM发现后屏蔽
- 进程重启，向CM注册
- Dispatcher获取断点信息，续传实时更新的文档

Dispatcher进程挂掉

- CM发现后屏蔽
- 进程重启，向CM注册，恢复服务

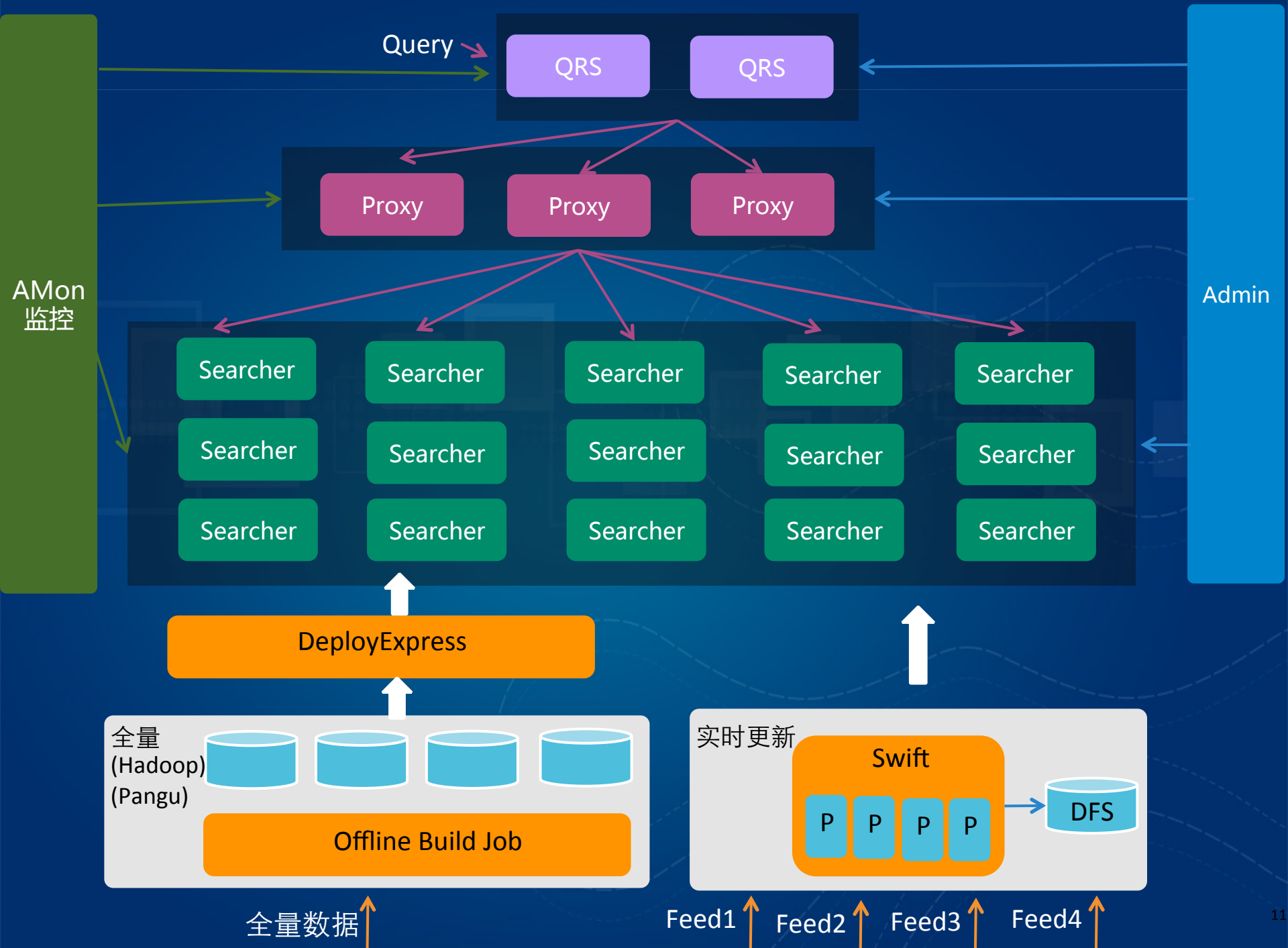
机器挂掉

- CM发现后自动屏蔽
- 每天晚上全量索引完成后换机器
- 不支持动态切换机器

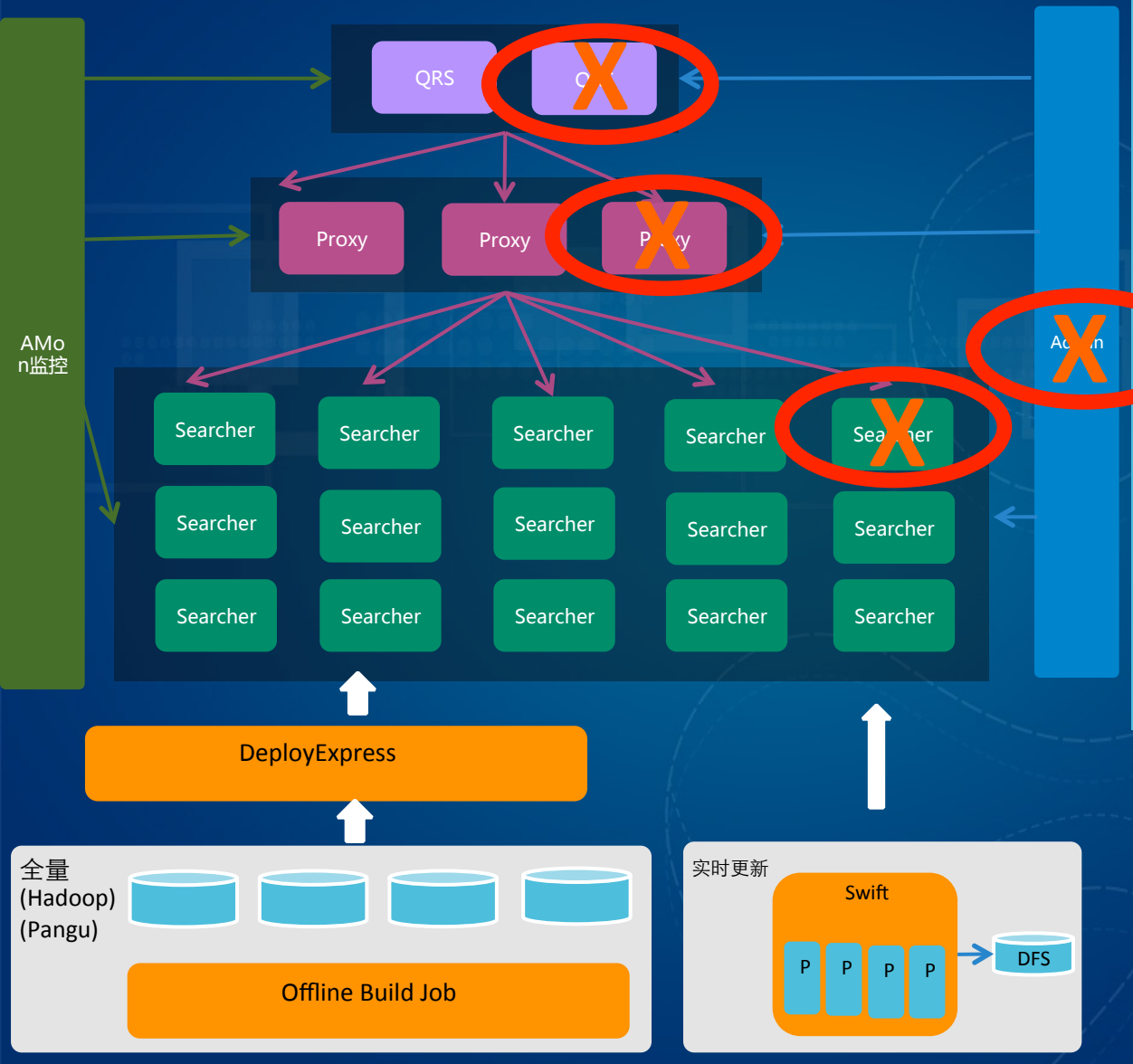
索引损坏

- 同上


机器
50
500+
PE


Query
QRS
QRS
Proxy
Proxy
Proxy
AMon
监控
Admin
Searcher
DeployExpress
全量
(Hadoop)
(Pangu)
Offline Build Job
实时更新
Swift
P
DFS
全量数据
Feed1
Feed2
Feed3
Feed4

错误恢复



QRS/Proxy机器挂掉

- Admin从Buffer机器列表中选择1台机器
- 启动进程，替代挂掉机器提供服务

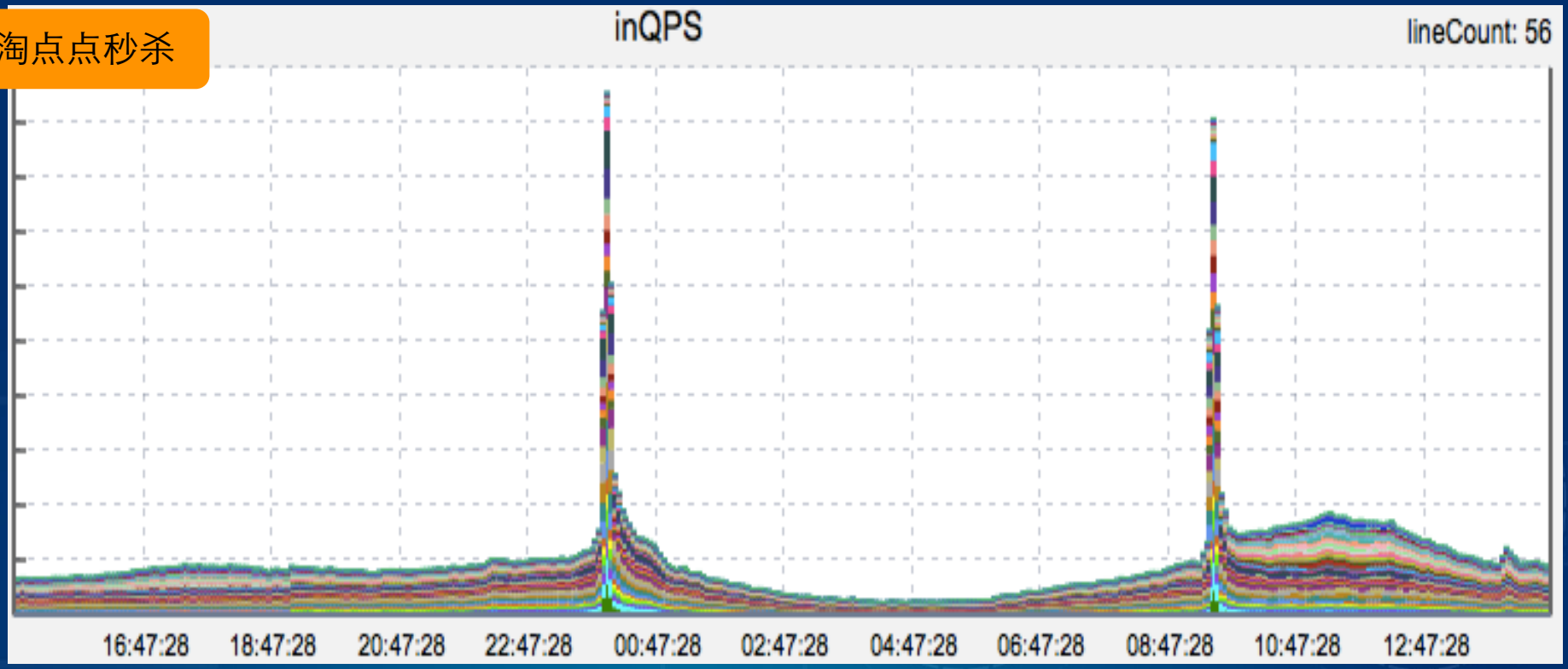
Searcher机器挂掉

- Admin从Buffer机器列表中选择1台机器
- 启动进程
- Admin发送命令给DeployExpress，拉取全量索引数据
- 从Swift拉取实时更新数据，数据完整后提供服务

Admin机器挂掉

- 启动备份Admin
- 恢复配置，接管服务

淘点点秒杀

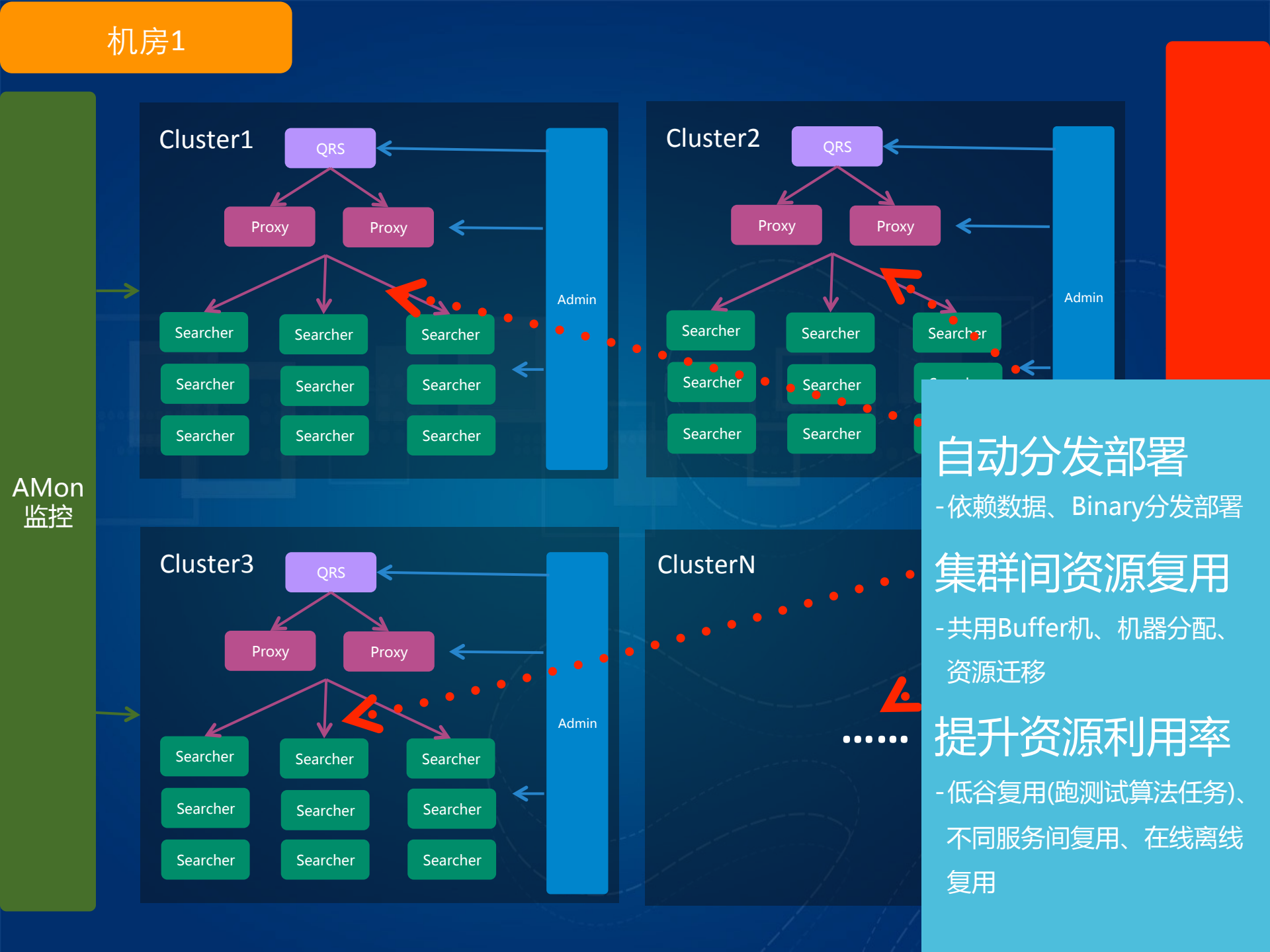


资源问题

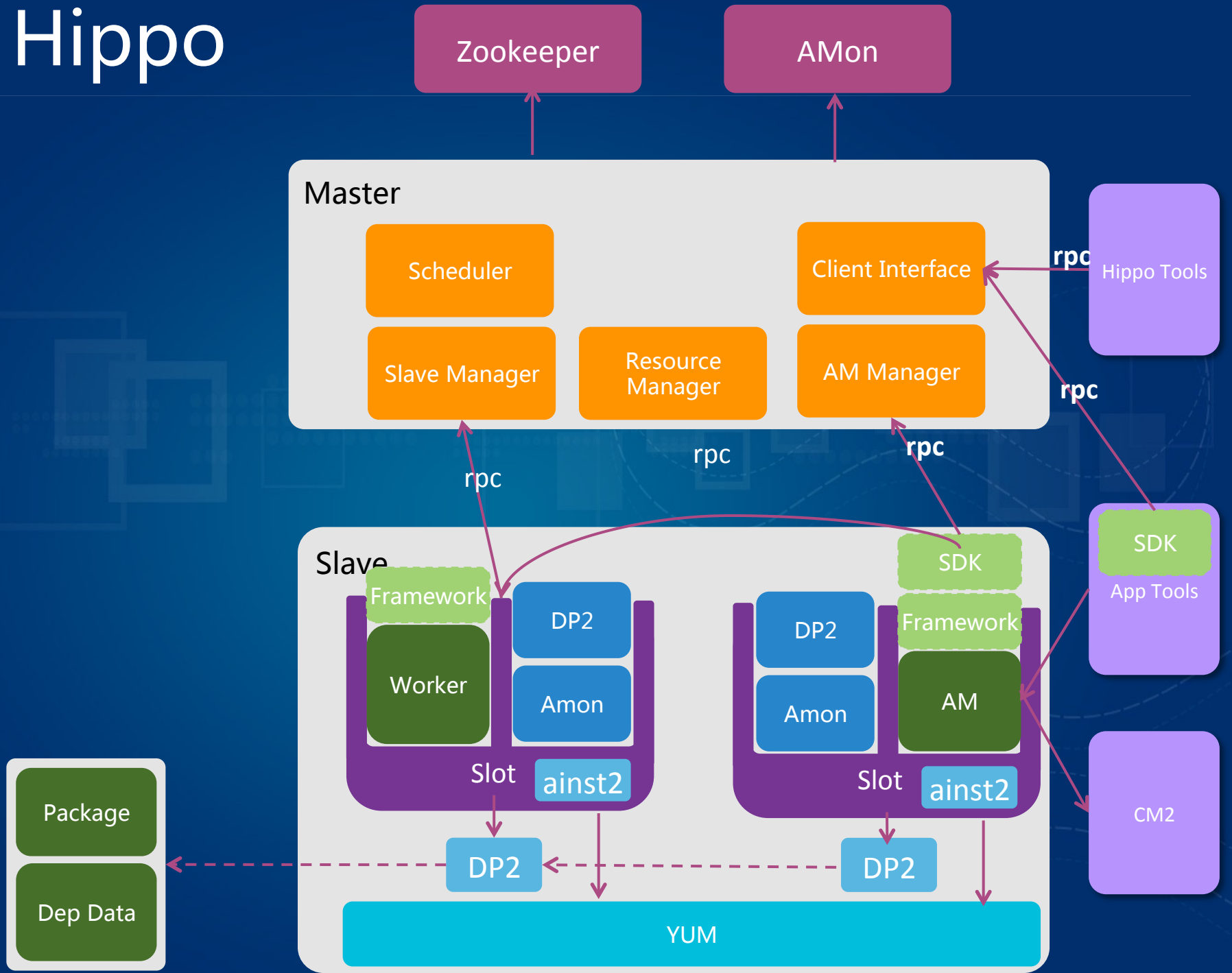
-流量波峰波谷、多业务间、在线离线资源无法复用

运维问题

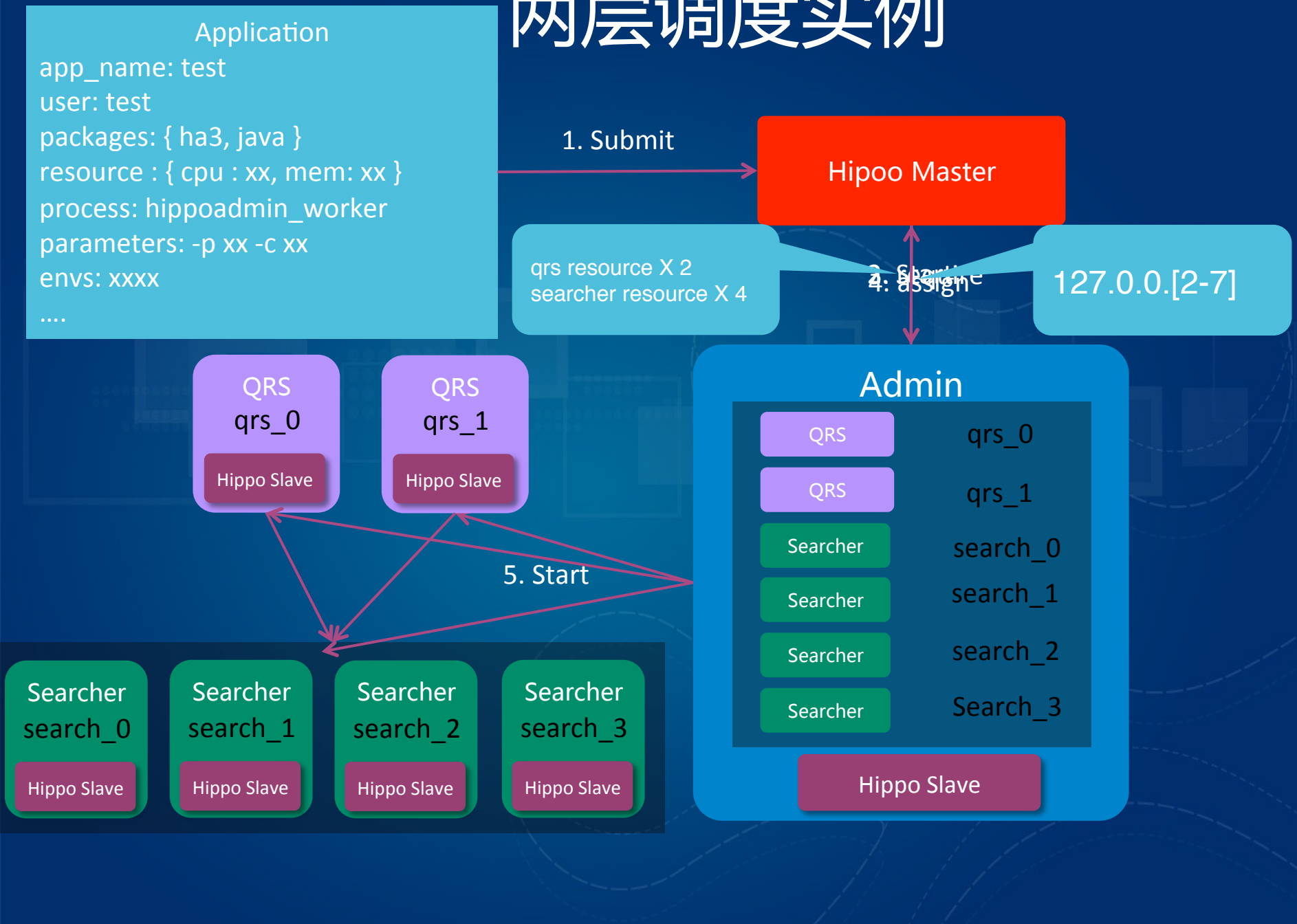
-业务多、机器多、峰值突增，扩容和机器迁移耗时耗力


机房1
AMon
监控
Cluster1
QRS
Proxy
Proxy
Searcher
Admin
Cluster2
Cluster3
ClusterN
自动分发部署
-依赖数据、Binary分发部署
集群间资源复用
-共用Buffer机、机器分配、资源迁移
提升资源利用率
-低谷复用(跑测试算法任务)、不同服务间复用、在线离线复用

Hippo


Zookeeper
AMon
Master
Scheduler
Client Interface
Slave Manager
Resource Manager
AM Manager
rpc
Hippo Tools
rpc
rpc
rpc
rpc
SDK
App Tools
Slave
Framework
DP2
Worker
Amon
Slot
ainst2
SDK
DP2
Framework
Amon
AM
Slot
ainst2
CM2
Package
Dep Data
DP2
DP2
YUM

两层调度实例


Application
app_name: test
user: test
packages: { ha3, java }
resource : { cpu : xx, mem: xx }
process: hippoadmin_worker
parameters: -p xx -c xx
envs: xxxx
....
1. Submit
Hipoo Master
qrs resource X 2
searcher resource X 4
4: assign
127.0.0.[2-7]
QRS
qrs_0
Hippo Slave
QRS
qrs_1
Hippo Slave
Admin
QRS
qrs_0
QRS
qrs_1
Searcher
search_0
Searcher
search_1
Searcher
search_2
Searcher
Search_3
Hippo Slave
5. Start
Searcher
search_0
Hippo Slave
Searcher
search_1
Hippo Slave
Searcher
search_2
Hippo Slave
Searcher
search_3
Hippo Slave

错误恢复

QRS/Proxy机器挂掉

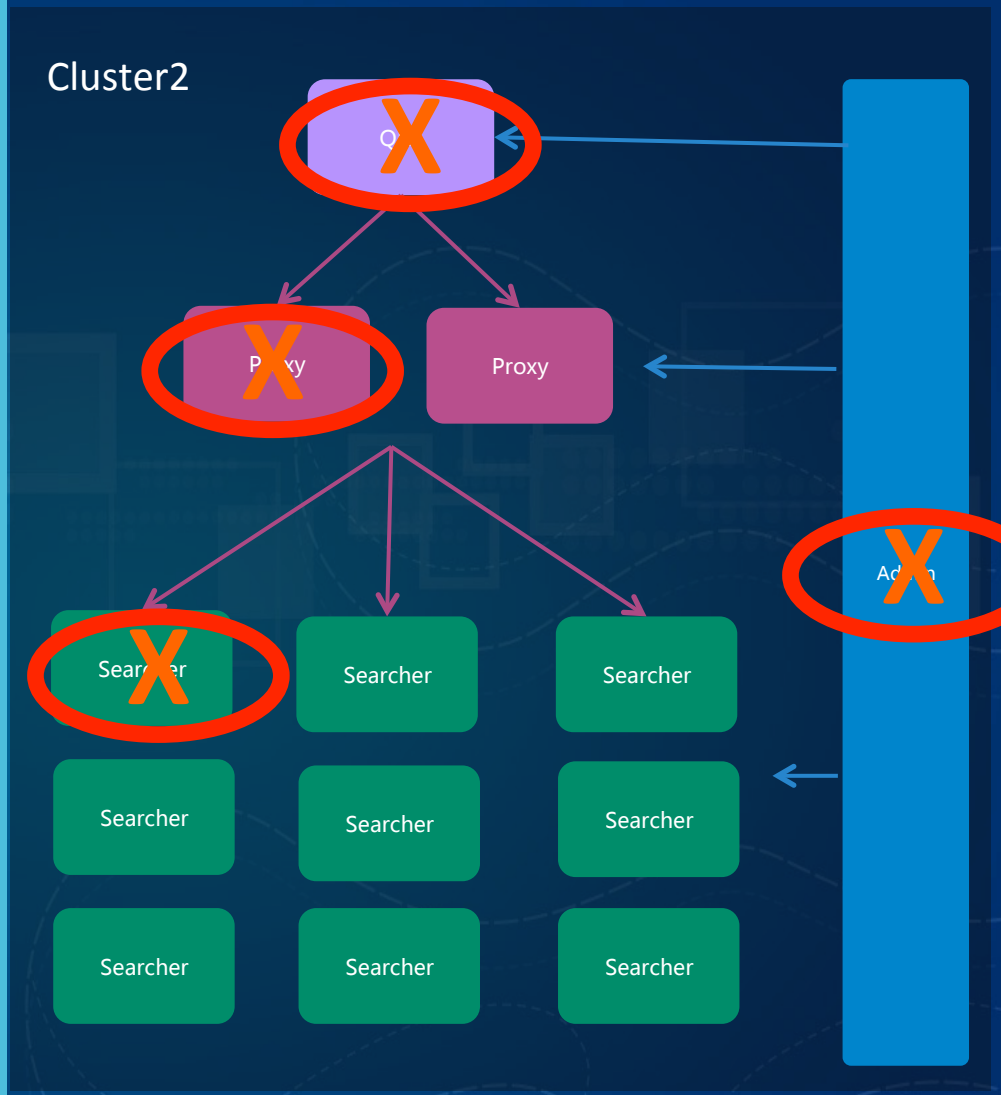
- Admin向Hippo申请机器资源
- Hippo分发Binary包和数据，启动进程，替代挂掉机器提供服务

Searcher机器挂掉

- Admin向Hippo申请机器资源
- Hippo分发Binary包和数据，启动进程
- Admin发送命令给DeployExpress，拉取全量索引数据
- 从Swift拉取实时更新数据，数据完整后提供服务

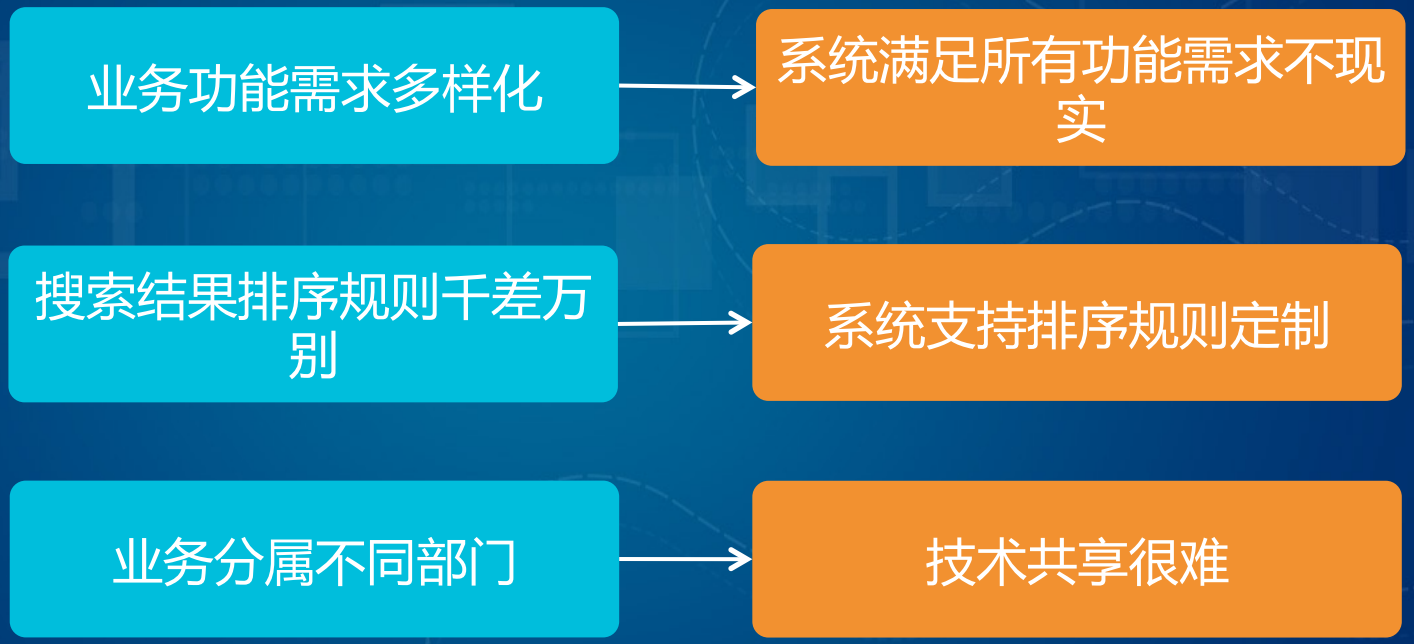
Admin机器挂掉

- 启动备份Admin
- 恢复配置，接管服务



系统、平台、服务

搜索业务和系统


业务功能需求多样化
系统满足所有功能需求不现实
搜索结果排序规则千差万别
系统支持排序规则定制
业务分属不同部门
技术共享很难

一套系统、多个代码分支、多套部署

代码不能复用维护困难

系统提升无法共享

随着业务数量增长效率急速下降

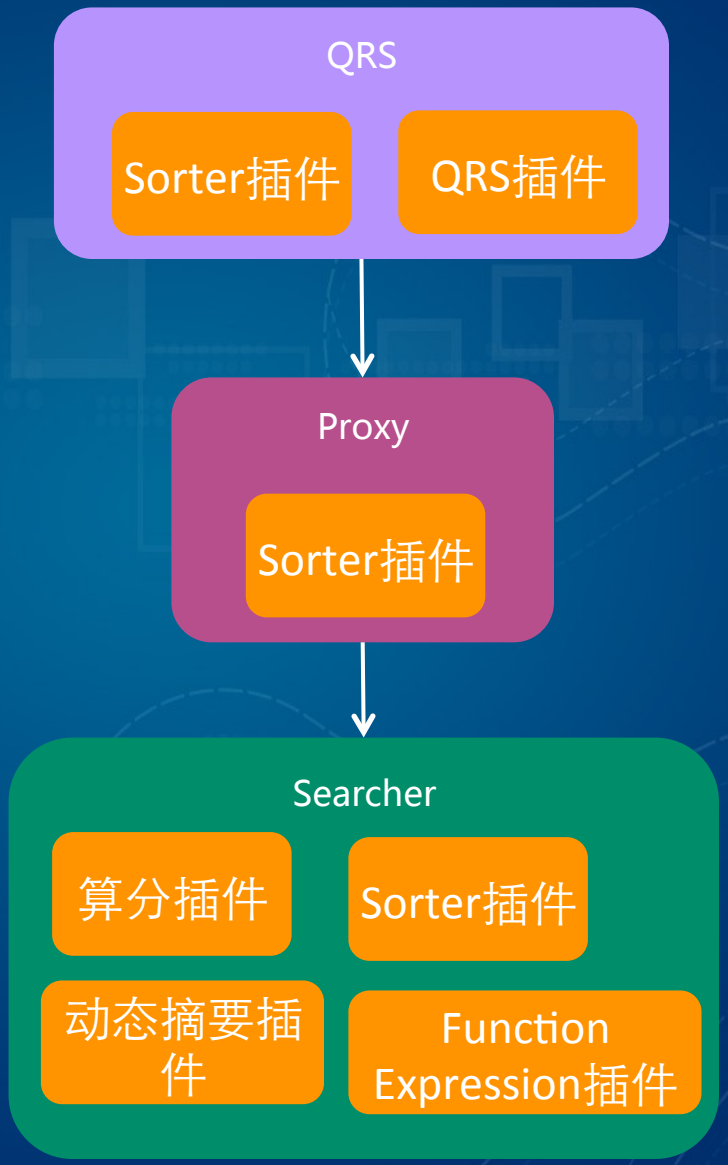
解决方案：平台化

自平台化的挑战

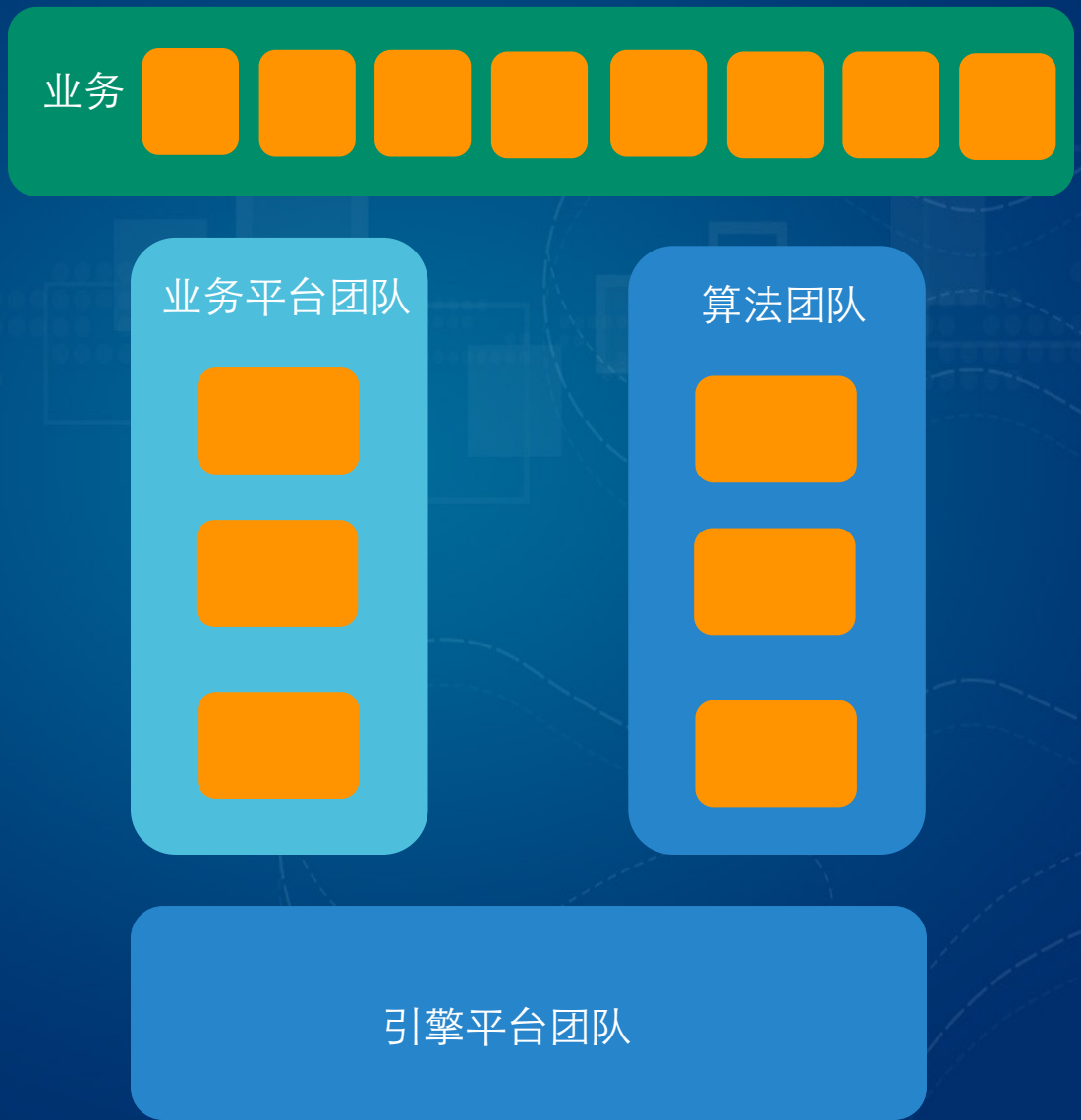
- 多样化功能需求如何满足
- 搜索结果排序规则如何灵活定制
- 业务需求如何快速响应

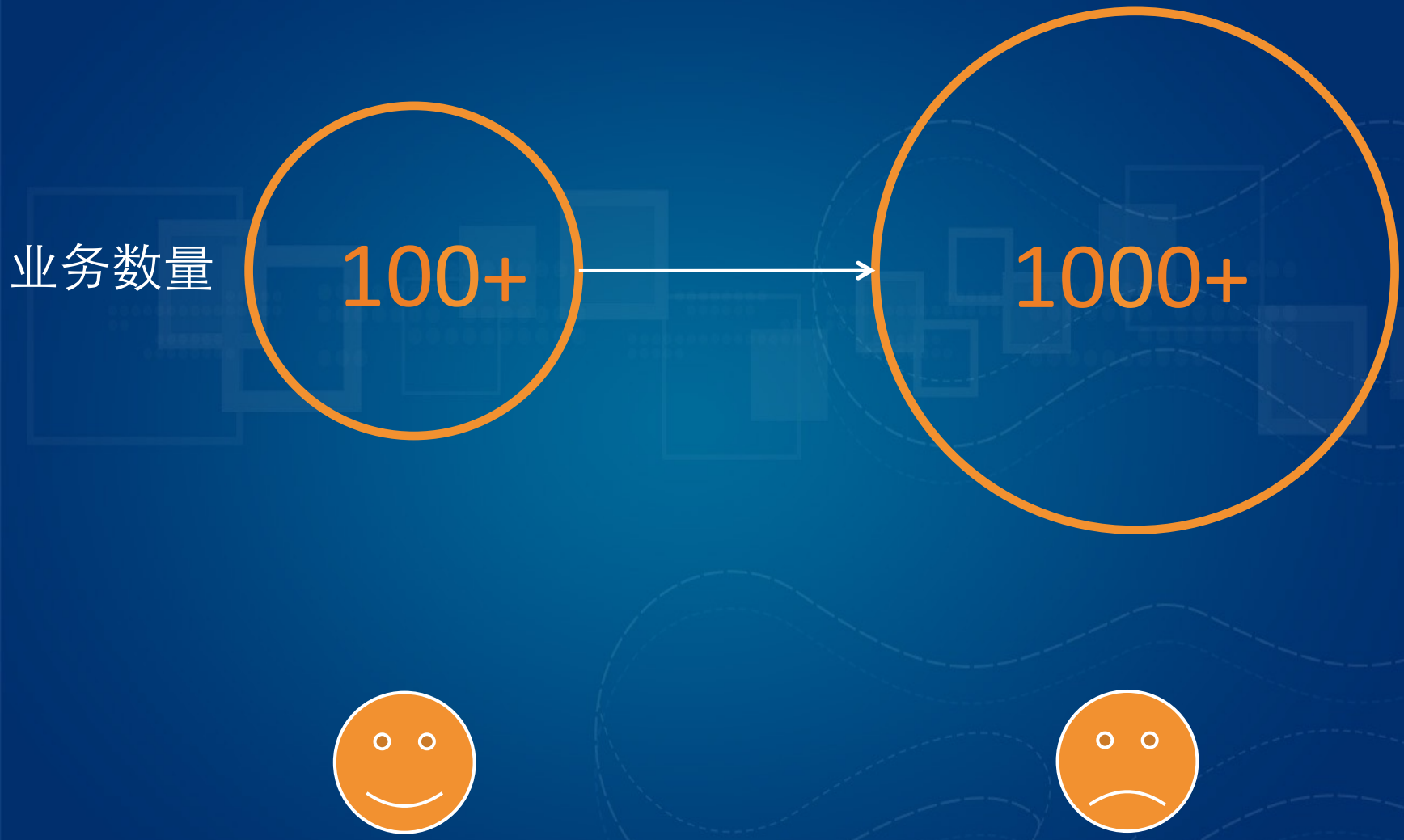
系统插件化

- 算分插件
- QRS插件
- 文档处理插件
- 动态摘要插件
- Function Expression插件
- Sorter插件



团队组织架构保障




业务数量
100+
1000+

服务化

不让用户写插件

屏蔽系统细节

自助式定制

OpenSearch：自助式云搜索服务

UI & API

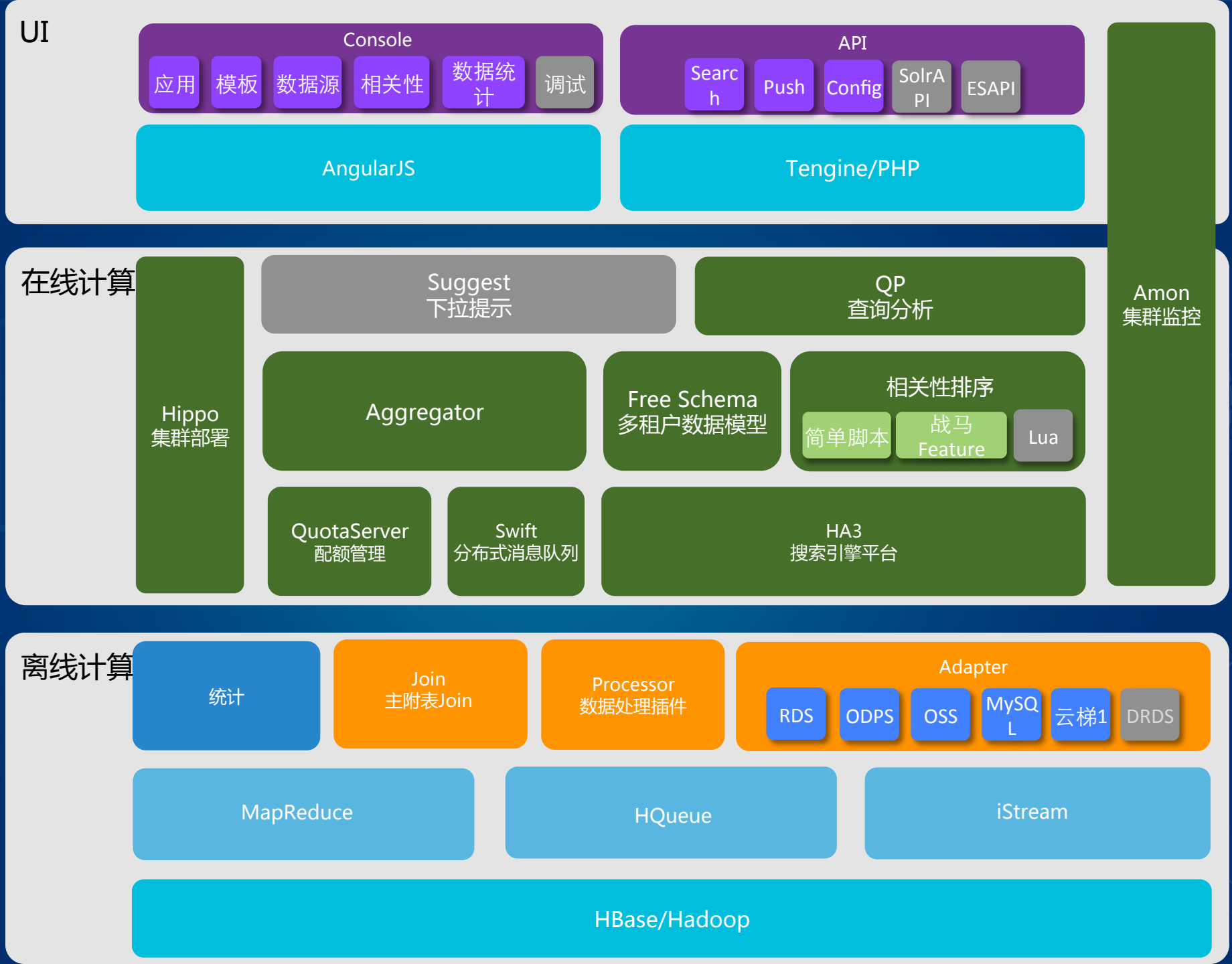
- 所有功能可以在控制台中操作完成
- 所有主要功能可通过API操作实现
- 按需使用服务，按需调整配额

索引结构定制

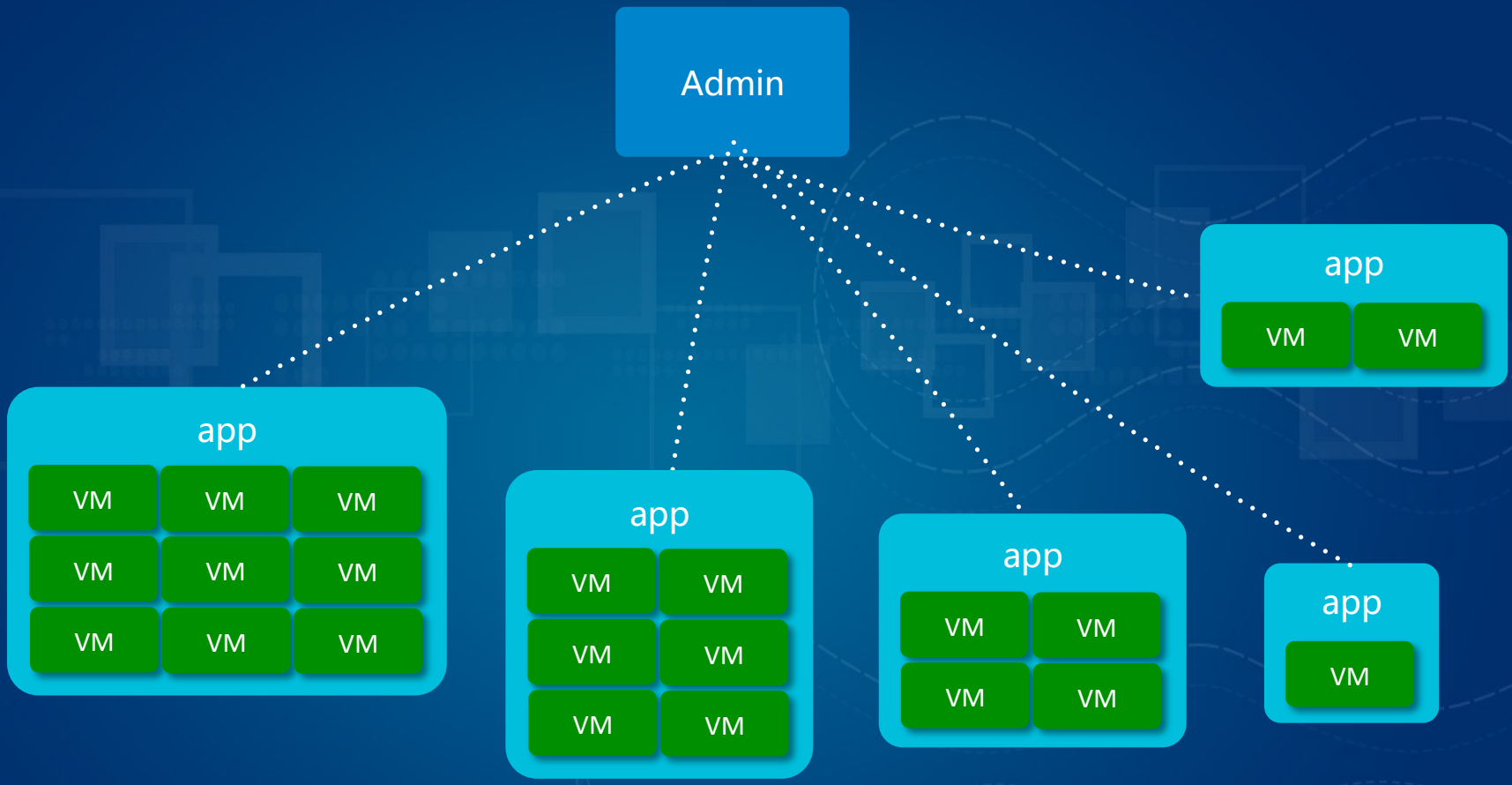
- 文档字段、字段类型、索引方式可定制
- 运行中的服务实例，索引结构可动态修改

相关性定制

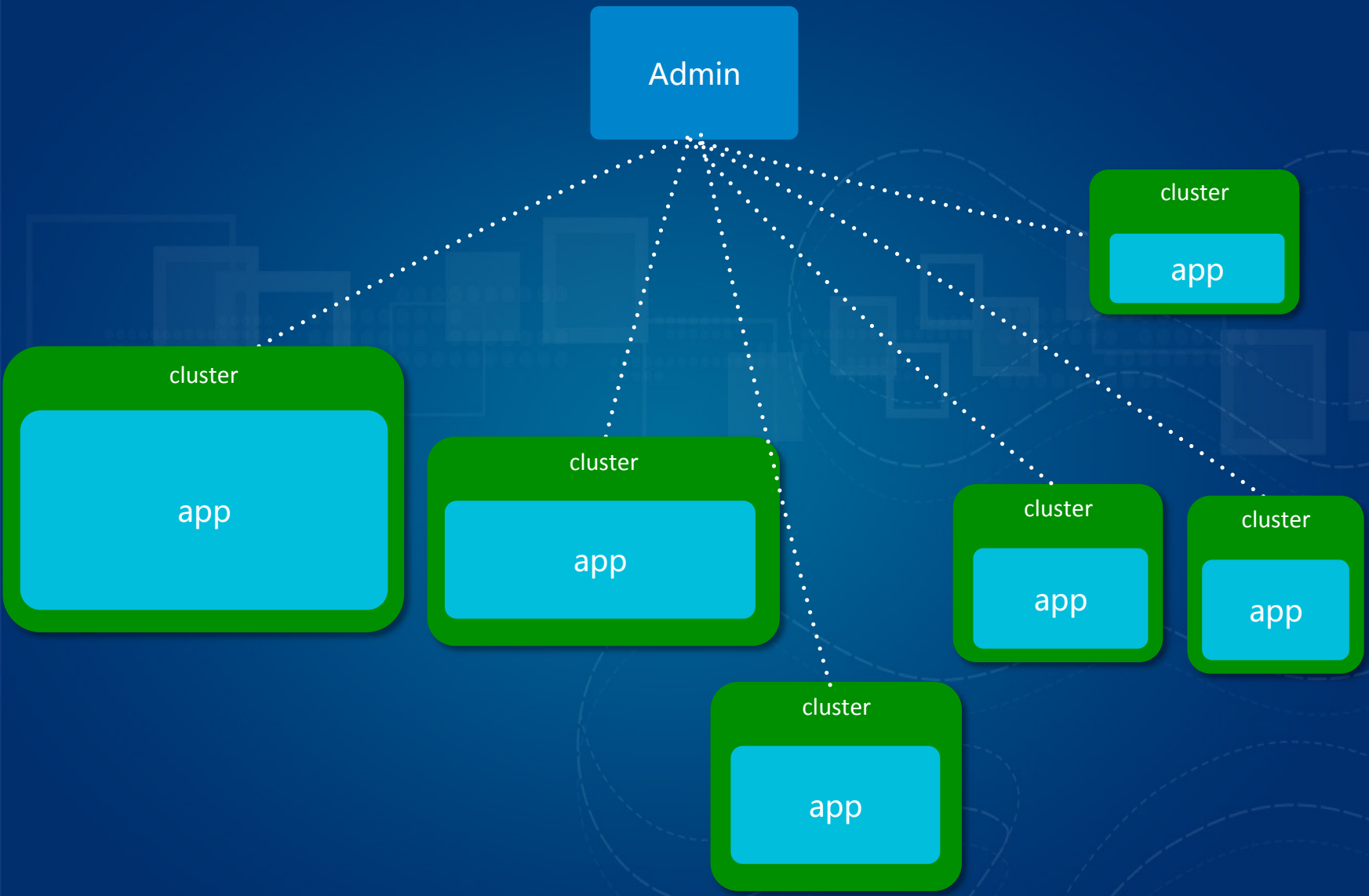
- 通过表达式定制搜索结果排序逻辑（后续将支持Lua）
- 内置文本、地理位置相关等算分特征函数
- 使用查询分析服务更好理解用户query


UI
Console
应用
模板
数据源
相关性
数据统计
调试
API
Search
Push
Config
SolrAPI
ESAPI
AngularJS
Tengine/PHP
Amon
集群监控
在线计算
Hippo
集群部署
Suggest
下拉提示
QP
查询分析
Aggregator
Free Schema
多租户数据模型
相关性排序
简单脚本
战马
Feature
Lua
QuotaServer
配额管理
Swift
分布式消息队列
HA3
搜索引擎平台
离线计算
统计
Join
主附表Join
Processor
数据处理插件
Adapter
RDS
ODPS
OSS
MySQL
云梯1
DRDS
MapReduce
HQueue
iStream
HBase/Hadoop

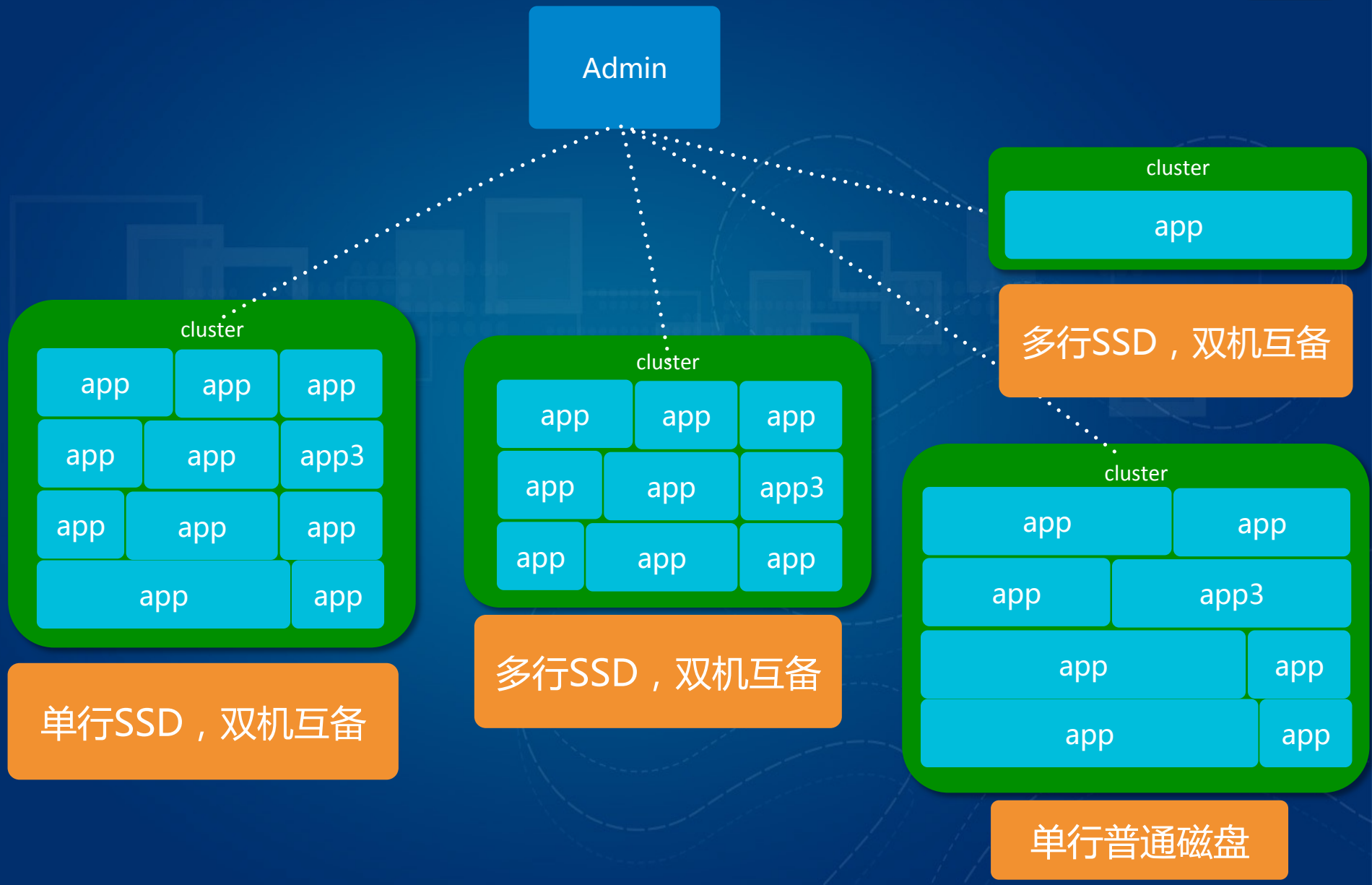
多租户数据模型：虚拟机


Admin
app
VM
VM
VM
VM
VM
VM
VM
VM
VM
app
VM
VM
VM
VM
VM
VM
app
VM
VM
VM
VM
app
VM
VM
VM

多租户数据模型：海量cluster



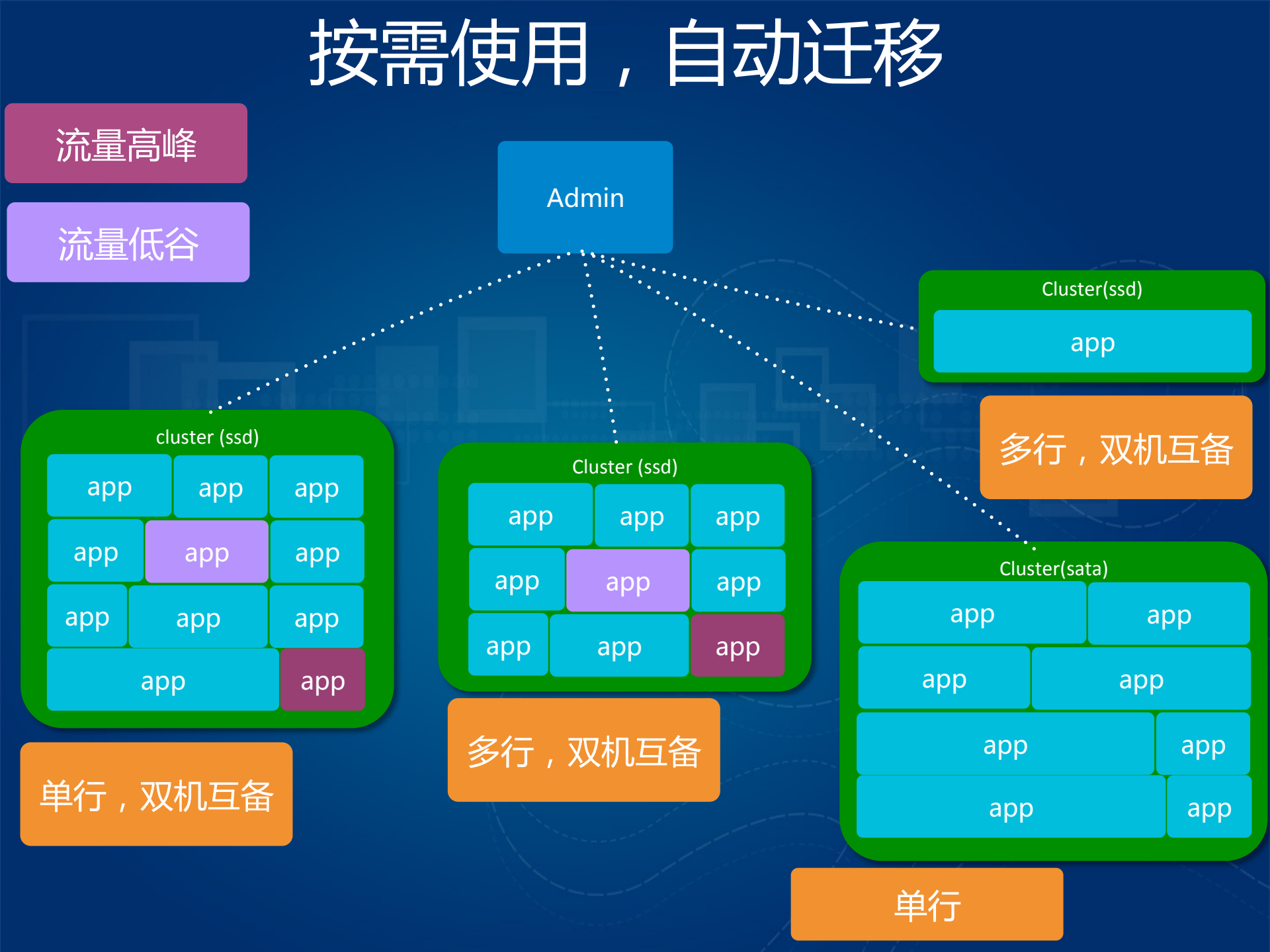
多租户数据模型：混合


Admin
cluster
app
app
app
app
app
app3
app
app
app
app
app
cluster
app
app
app
app
app
app3
app
app
app
cluster
app
app
app
app
app
app3
app
app
app
app
单行SSD，双机互备
多行SSD，双机互备
多行SSD，双机互备
单行普通磁盘

按需使用，自动迁移

流量高峰

流量低谷



开放搜索服务OpenSearch

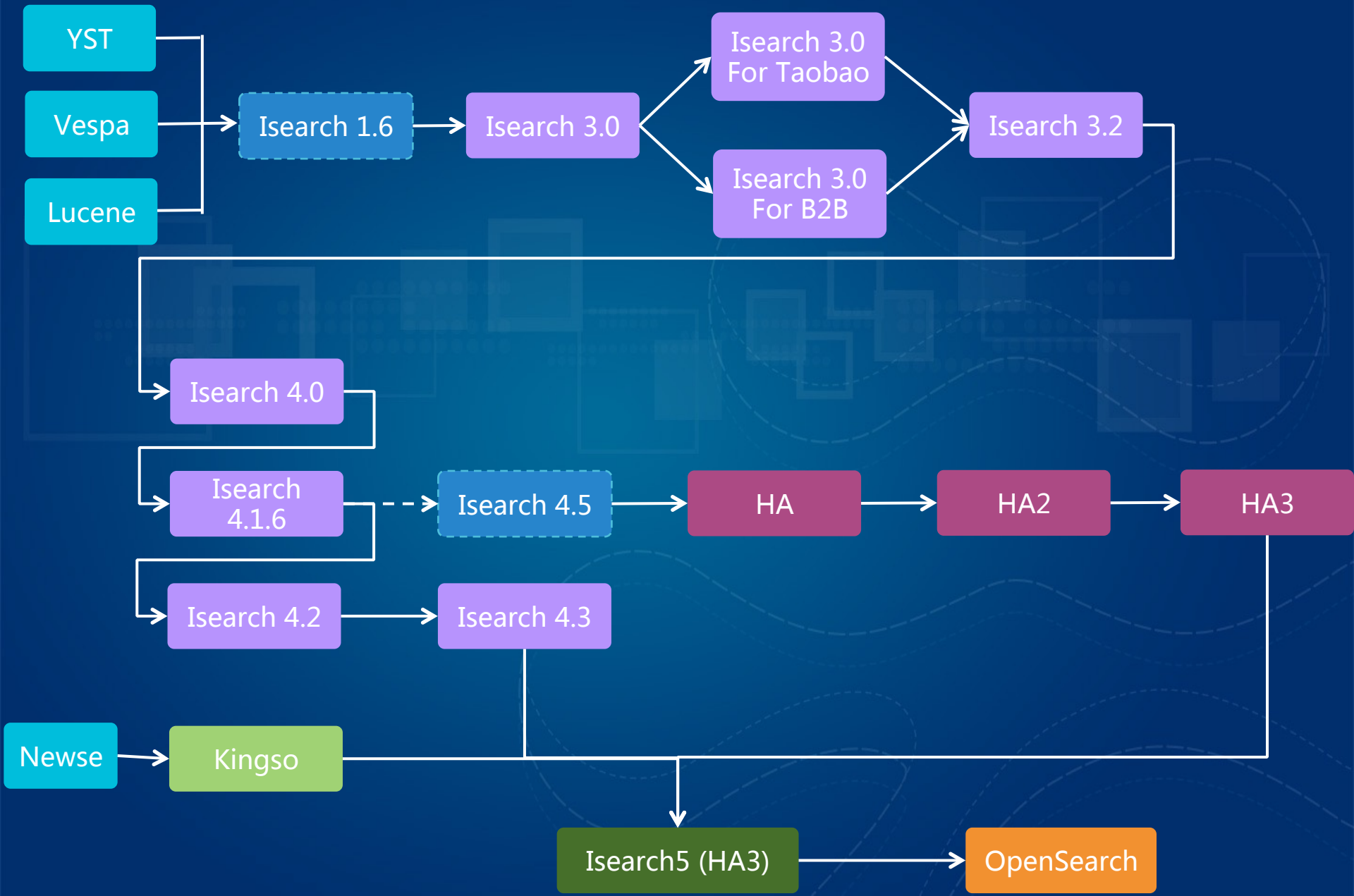
www.aliyun.com/product/opensearch

微博@阿里云开放搜索服务

旺旺群@1318169830

QQ群@370015616

阿里搜索引擎平台演变历史



开源: 2015


HA3(搜索引擎平台)
LocalAgent (守护进程)
Fslib (文件IO)
ALog (日志)
DeployExpress（链式数据分发）
Anet (网络通信)
Swift (消息队列)
Indexlib (索引)
Amonitor (监控)
Hippo (资源管理)
AInst (安装部署)

Thanks

微博@ruijieguo

微信@jared_guo