QINIU

万亿级流量的大数据平台架构开发实践

七牛云-党合萱

主要内容

- 一个场景
- 产品
- 设计目标与架构
- 挑战与解决方案
- 成果

运维日志分析

Nginx是现代web服务栈中最重要的组件之一，通过对nginx的分析处理可以发现数据更大的价值

日志分析步骤:

- 下载logkit，配置并运行，将数据打入pandora
- 查看日志处理业务逻辑
- 查询日志
- 聚合日志
- 数据回流至平台
- 离线处理
- 实时数据展示与监控
- XSpark处理

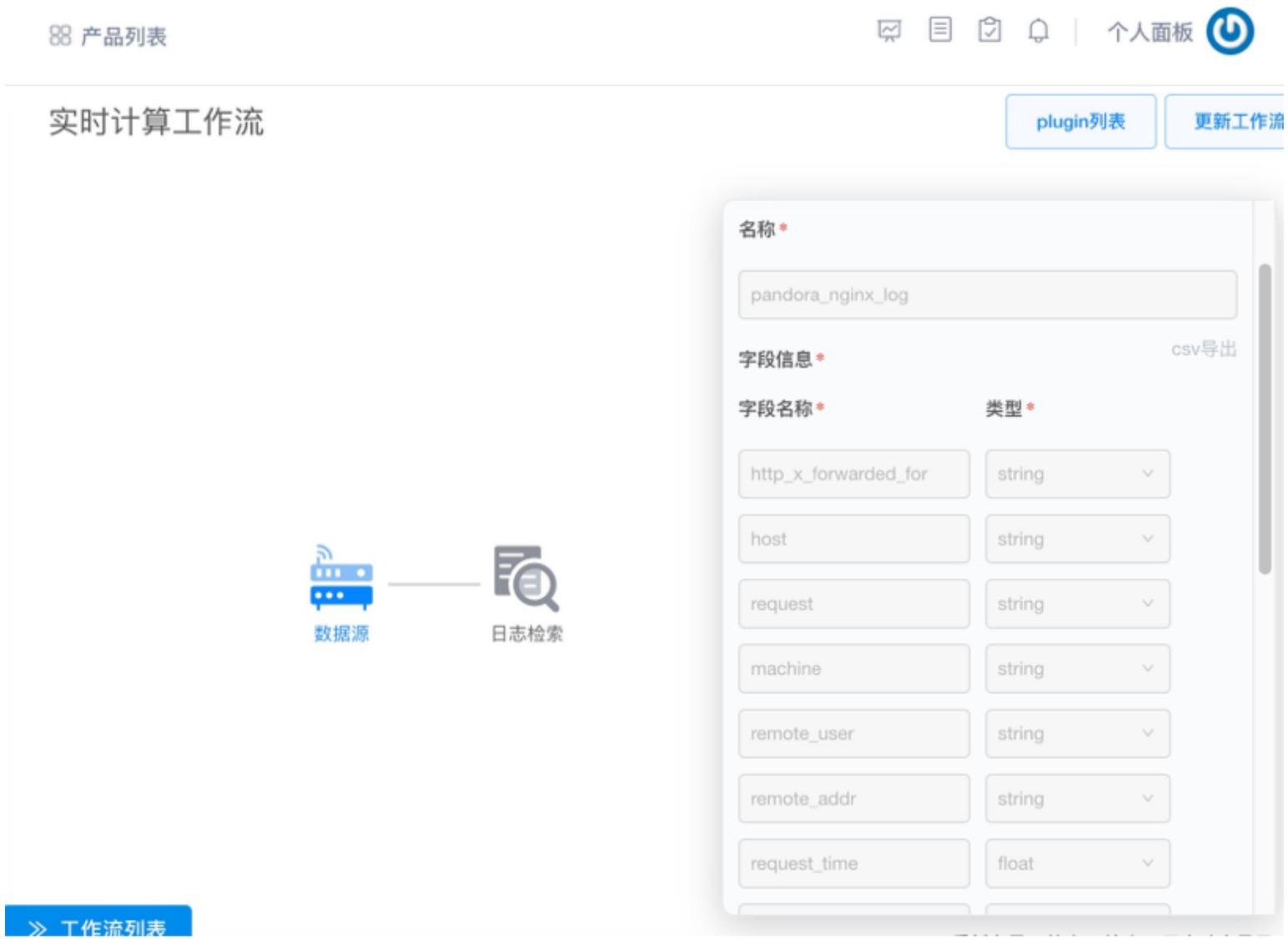
数据接入


'$host $request_time v0';
log_format v1 '$remote_addr - $remote_user [$time_local] "$request" '
'$status $body_bytes_sent "$http_referer" '
'"$http_user_agent" "$http_x_forwarded_for" '
host $request_time "$sent_http_x_reqid" "$upstream_addr" v1';
server {
listen 80;
server_name localhost;
location /nginxstatus {
access_log off;
stub_status on;
allow 127.0.0.1;
deny all;
}
}
log_format main '$remote_addr - $remote_user [$time_local] "$request" '
'$status $bytes_sent $body_bytes_sent "$http_referer" '
'"$http_user_agent" "$http_x_forwarded_for" '
'$upstream_addr $host $sent_http_x_reqid $request_time';
access_log logs/access.log main;
#limit_conn_zone $http_host zone=servicelimit:10m;
#limit_conn_zone $http_host zone=limitspeed:50m;
#limit_conn_log_level error;
#limit_req_zone $binary_remote_addr zone=one:50m rate=2r/s;
sendfile on;
#tcp_nopush on;
server_tokens off;
ssl_protocols TLSv1 TLSv1.1 TLSv1.2;
ssl_ciphers HIGH:!aNULL:!MD5:!DES;
#keepalive_timeout 0;
keepalive_timeout 65;
gzip_min_length 1000;
gzip_comp_level 8;
gzip_proxied any;
gzip_types text/plain text/css text/javascript text/xml application/x-javascript applicatic
ss application/javascript;
gzip off;
日志格式名称
日志格式


proxy_buffering off;
add_header Vary Accept-Encoding;
add_header X-Whom nb2105;
client_max_body_size 1024m;
proxy_set_header X-Forwarded-For $proxy_add_x_forwarded_for;
proxy_set_header Host $http_host;
proxy_set_header X-Real-IP $remote_addr;
proxy_set_header X-Scheme $scheme;
proxy_redirect off;
retry next upstream
proxy_next_upstream error timeout http_570;
access_log /opt/nginx/logs/pipeline_com.log main;
#limit_conn servicelimit 8000;
实际nginx日志格式配置
main就是format名称


{
"name":"nginx_runner",
"reader":{
"mode":"file",
"meta_path":"meta",
"log_path":"/opt/nginx/logs/pipeline.log"
},
"parser":{
"name":"nginx_parser",
"type":"nginx",
"nginx_log_format_path":"/opt/nginx/conf/nginx.conf",
"nginx_log_format_name":"main",
"nginx_schema":"time_local:date, status:long, bytes_sent:long, body_bytes_sent:long,request_time:float"
},
"senders":[{
"name":"pandora_sender",
"sender_type":"pandora",
"pandora_ak":"your_ak",
"pandora_sk":"your_sk",
"pandora_host":"https://pipeline.qiniu.com",
"pandora_repo_name":"nginx_log",
"pandora_region":"nb",
"pandora_schema_free":"true",
"pandora_gzip": "true",
"pandora_enable_logdb":"true",
"fault_tolerant":"true",
"ft_save_log_path":"./ft_log",
"ft_strategy":"always_save",
"ft_procs":"2"
}]
}
nginx 日志路径
您的 nginx 日志格式配置文件
nginx 日志格式名称
需要进行类型转换的字段
填写您的七牛 ak/sk
填写您Pandora的工作流(数据源)名称
填写解析完毕至发送前日志数据临时存放的路径

查看工作流


产品列表
个人面板
实时计算工作流
plugin列表
更新工作流
名称 *
pandora_nginx_log
字段信息 *
csv导出
字段名称 *
类型 *
http_x_forwarded_for
string
host
string
request
string
machine
string
remote_user
string
remote_addr
string
request_time
float
数据源
日志检索
工作流列表

日志检索

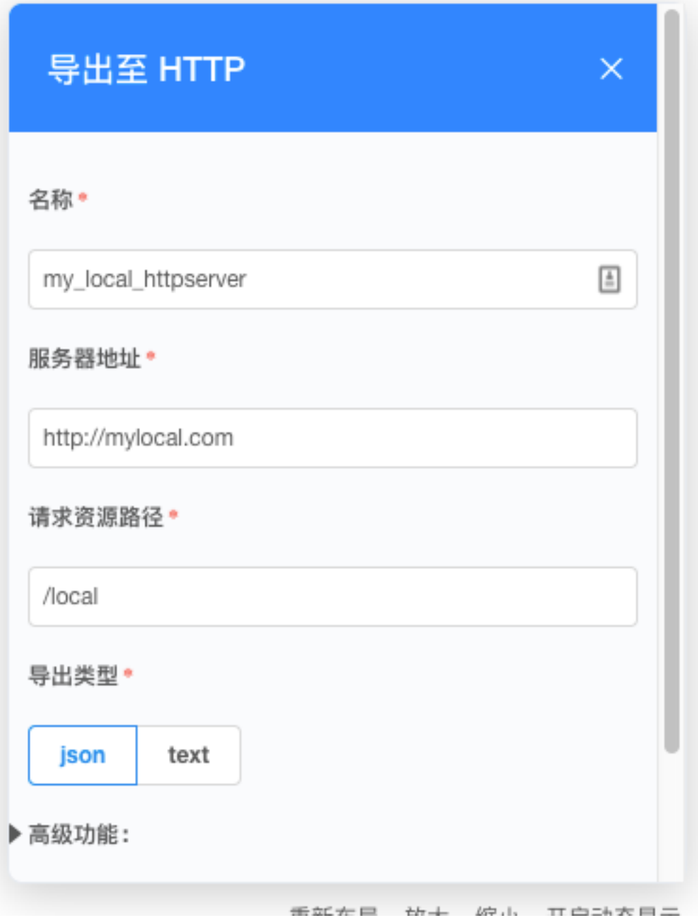

产品列表
数据统计
文档中心
工单
站内信
个人面板
资源主页
个人中心
财务统计
对象存储
大数据工作流引擎
时序数据库
日志检索
融合 CDN
SSL 证书服务
数据处理
直播云服务
容器计算
容器应用市场
仓库列表
日志查询
监控中心
选择创建的数据源名称
选择仓库
pandora_nginx_log
查看字段关
开
切换相关度排序
保存当前配置
根据时间字段排序
时间字段
time_local
最近3天
输入条件
sent_http_x_reqid: QTsAADhBbMw_-NEU
填入搜索条件
快速查询
时限查询
最近5分钟
最近15分钟
最近30分钟
最近1小时
最近3小时
最近6小时
最近12小时
最近1天
最近3天
最近7天
最近15天
最近30天
进入日志检索
您的时间排序字段为: time_local 数据总量为:7 耗时:4735ms 帮助文档
表格
Json数据
时间排序
source

日志聚合


SQL 编辑模式
SELECT count(request) as cnt from stream
Select *
Select all
format
清空
测试
SQL 可正常运行
字段名 cnt
类型 long
是否必填 false
取 消
保 存
计算任务
名称 *
容器类型 *
容器数量 *
1核 (CPU) 2G (内存)
1
自定义计算
SQL 计算
SQL 代码 *
高级编辑模式
运行间隔 *
1m
数据源
日志检索
计算任务
消息队列

数据回流

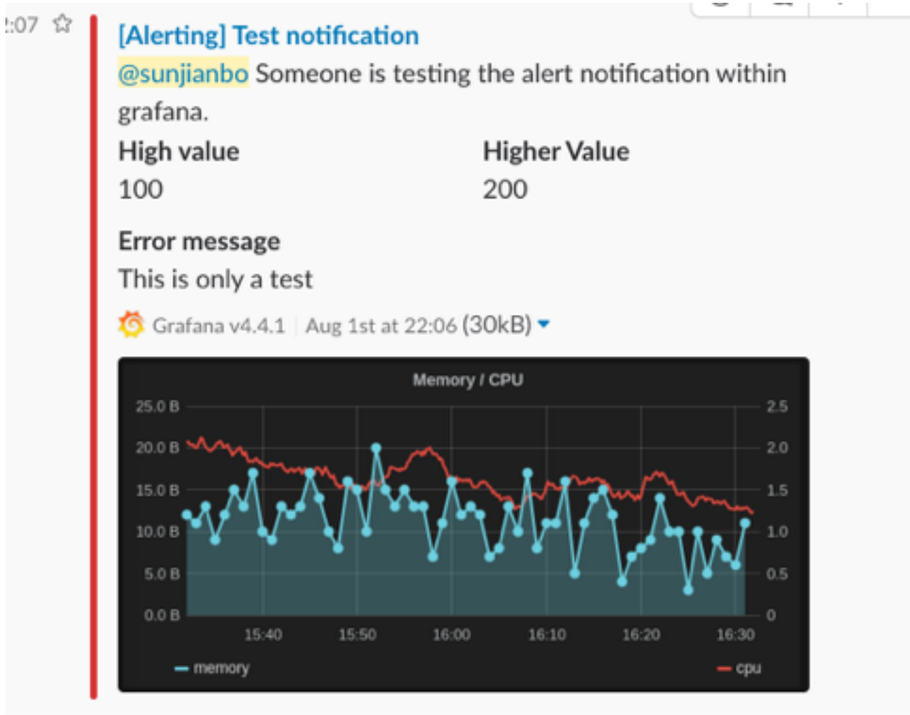

日志检索
数据源
计算任务
消息队列
HTTP


导出至 HTTP
名称 *
my_local_httpserver
服务器地址 *
http://mylocal.com
请求资源路径 *
/local
导出类型 *
json
text
高级功能：

数据展示与监控


产品列表
数据统计
文档中心
工单
站内信
个人面板
资源主页
个人中心
财务统计
融合 CDN
SSL 证书服务
数据处理
直播云服务
通用计算
应用市场
容器应用市场
官方应用
官方应用
第三方应用
更多应用
Grafana
新建部署
查看应用详情
XSpark
Pandora XSpark V1.5.0
Kibana应用
优秀的LogDB可视化工具


Elasticsearch Nginx Logs
visits: 0


[Alerting] Test notification
@sunjianbo Someone is testing the alert notification within grafana.
High value
100
Higher Value
200
Error message
This is only a test
Grafana v4.4.1 | Aug 1st at 22:06 (30kB)
Memory / CPU
25.0 B
20.0 B
15.0 B
10.0 B
5.0 B
0.0 B
2.5
2.0
1.5
1.0
0.5
0
15:40
15:50
16:00
16:10
16:20
16:30
memory
cpu

大数据平台-Pandora

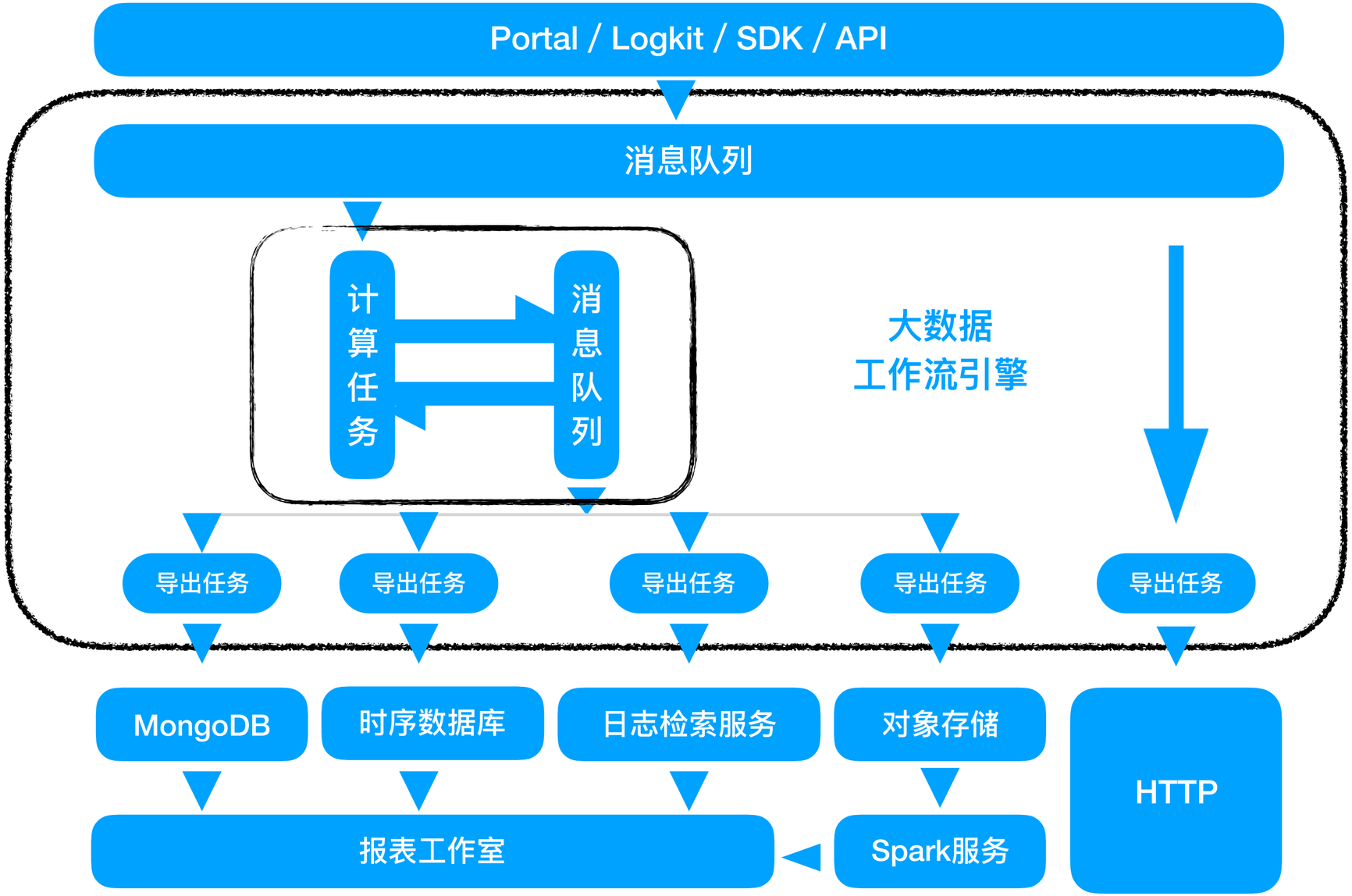
Pandora 是七牛云的大数据平台，提供简单、高效、开放的一站式大数据服务。

简单

高效

开放

大数据平台-Pandora架构图



Pipeline设计目标与技术选型

设计目标

- 高速数据写入，高吞吐量与低延迟
- 海量用户、消息队列支持
- 提供易用的实时计算与离线计算框架
- 可视化操作

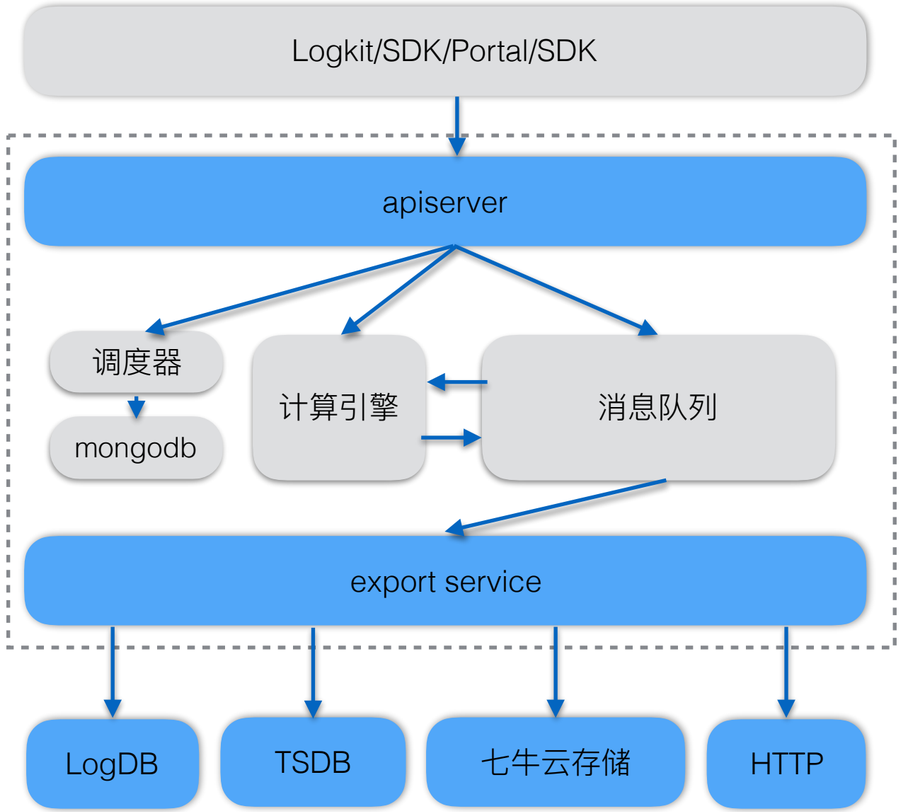
选型原则

- 具备高吞吐能力的存储系统
- 强大灵活的大数据处理引擎
- 可以快速开发迭代

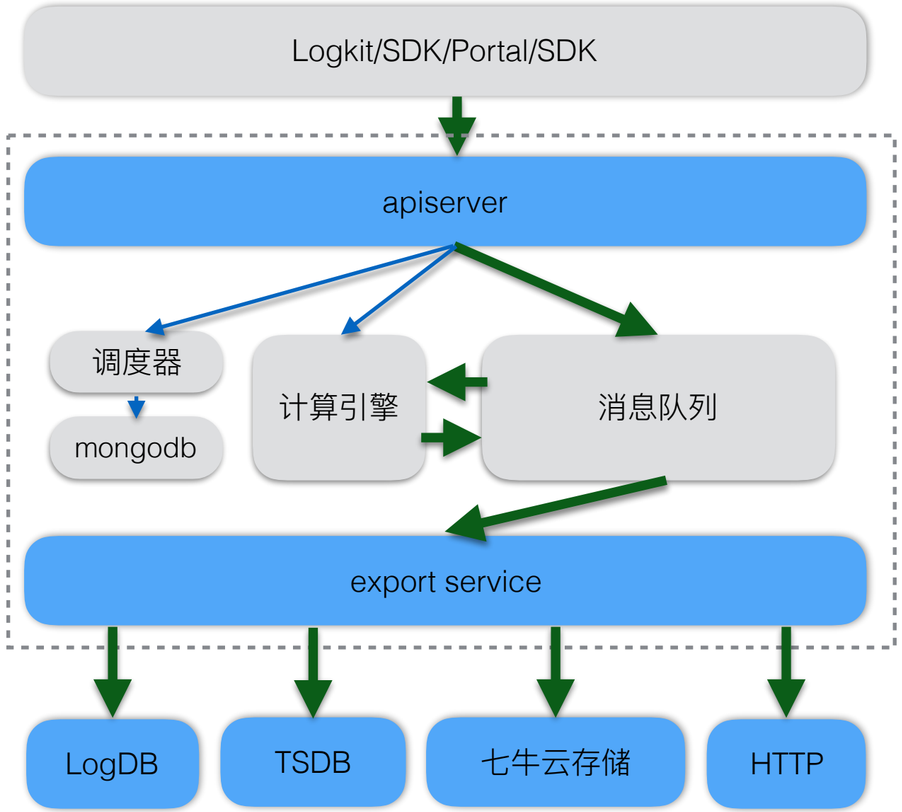
技术选型

- Kafka
- Spark streaming
- Golang

Pipeline架构图



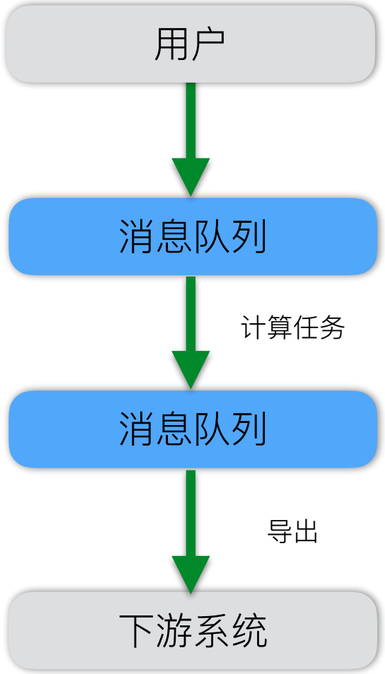
Pipeline架构图



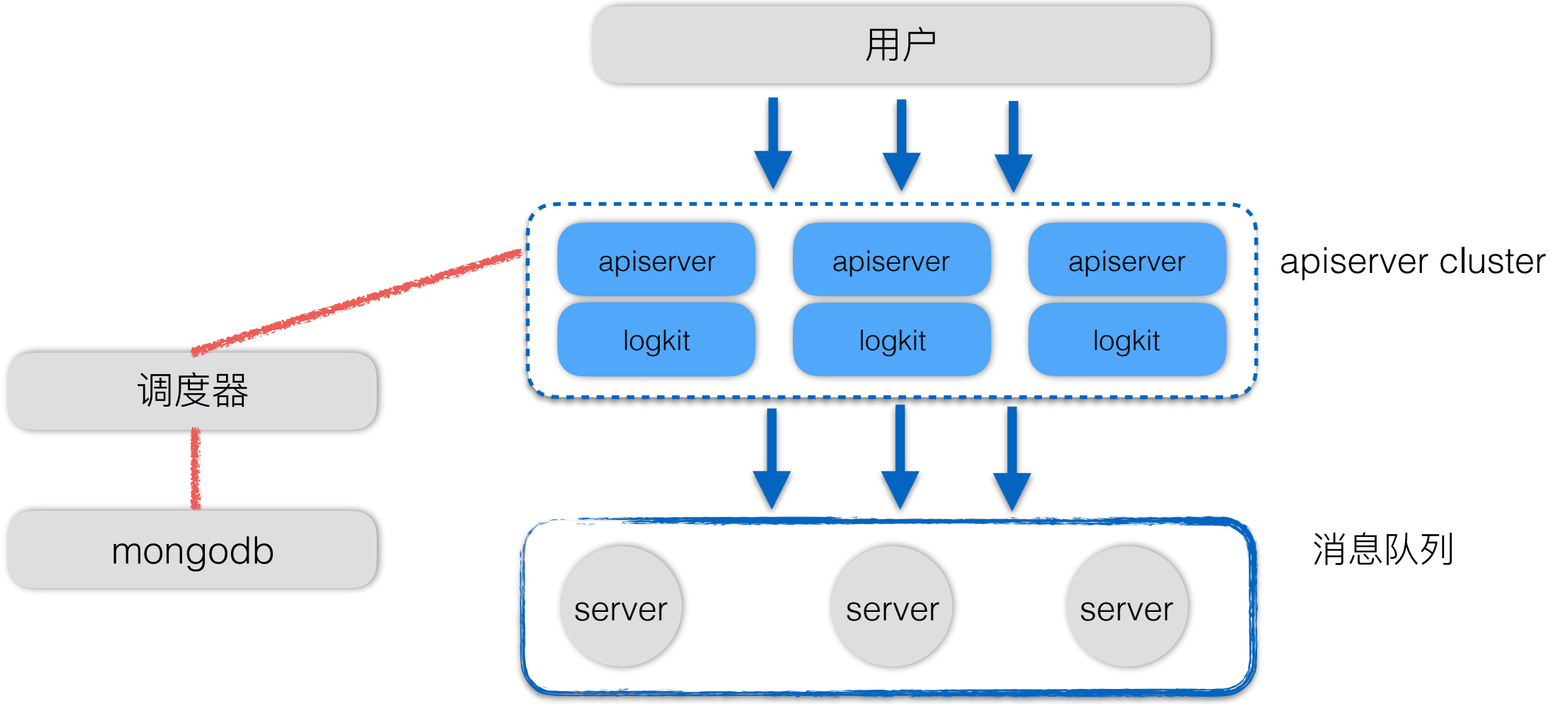
数据流剖析

一般影响因素

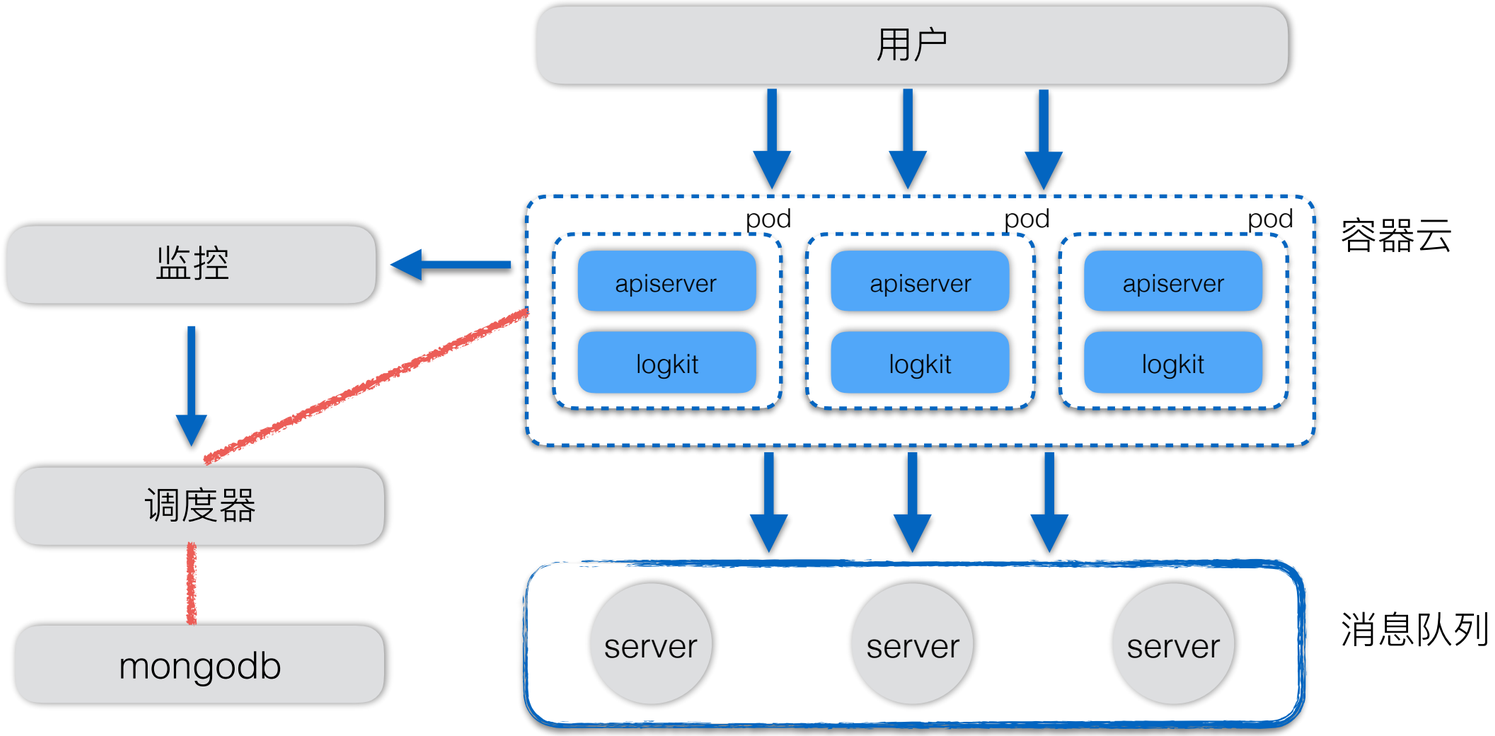
- 资源利用率
- 处理效率
- 木桶效应
- 链路损耗
- 其他



数据接入层

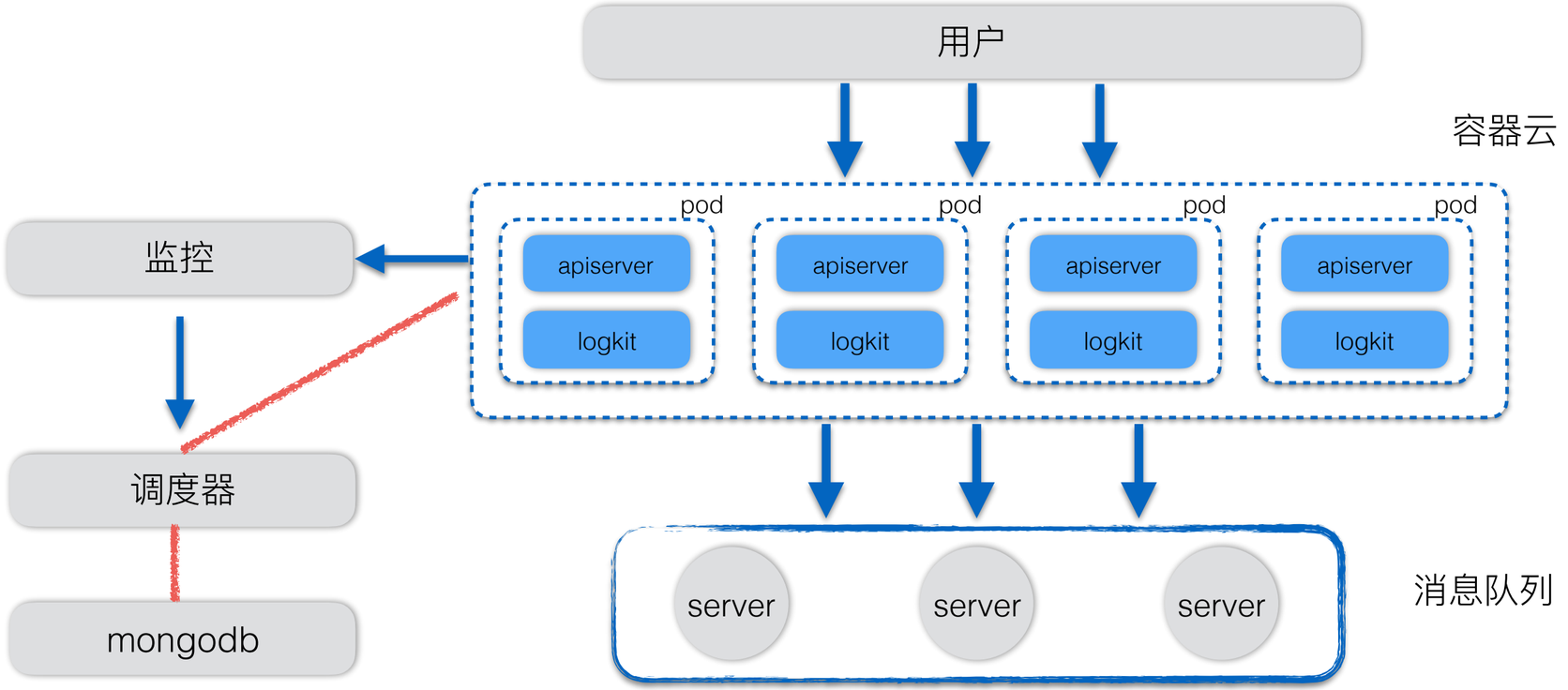

用户
apiserver
apiserver
apiserver
logkit
logkit
logkit
apiserver cluster
调度器
mongodb
server
server
server
消息队列

容器化

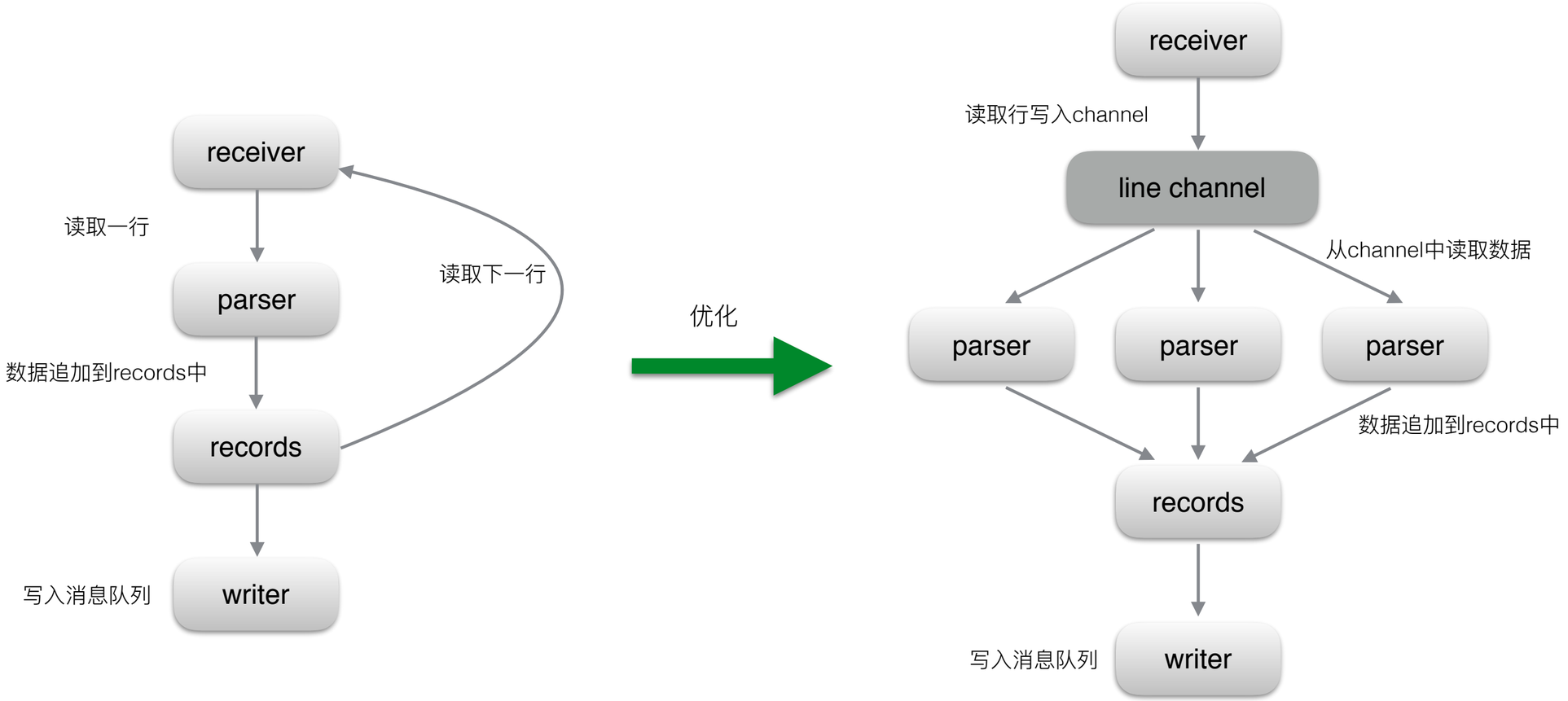

用户
pod
pod
pod
apiserver
apiserver
apiserver
logkit
logkit
logkit
容器云
监控
调度器
mongodb
server
server
server
消息队列

动态扩容

- 基于时序数据的监控
- 基于监控数据的扩容缩容



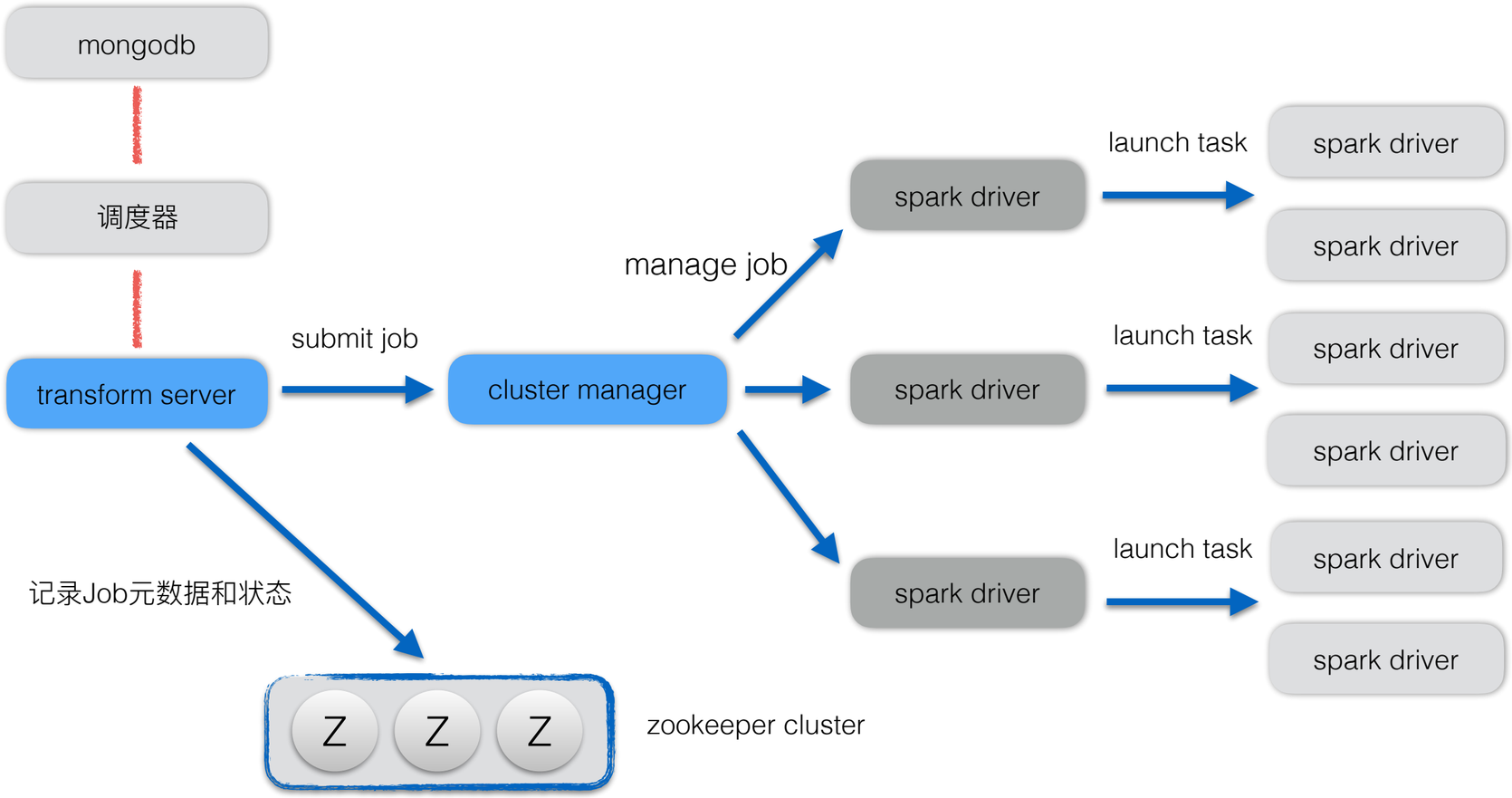
数据写入优化


receiver
读取一行
parser
读取下一行
数据追加到records中
records
写入消息队列
writer
优化
receiver
读取行写入channel
line channel
从channel中读取数据
parser
parser
parser
数据追加到records中
records
写入消息队列
writer

计算

- 计算引擎基于spark
- 提供SQL计算
- 屏蔽底层实现细节
- 支持海量用户

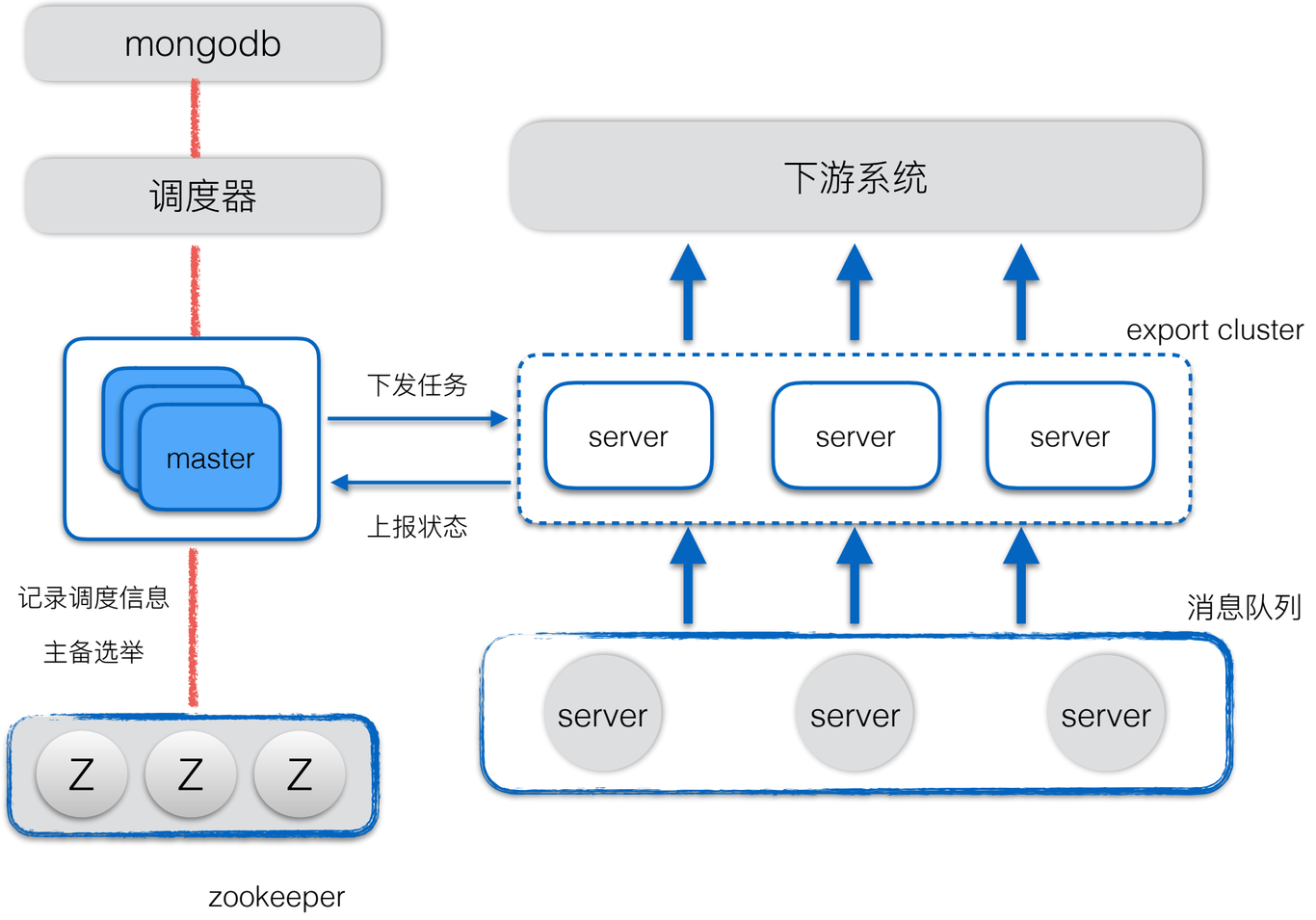
计算


mongodb
调度器
transform server
submit job
cluster manager
manage job
spark driver
spark driver
spark driver
launch task
launch task
launch task
spark driver
spark driver
spark driver
spark driver
spark driver
记录Job元数据和状态
Z
Z
Z
zookeeper cluster

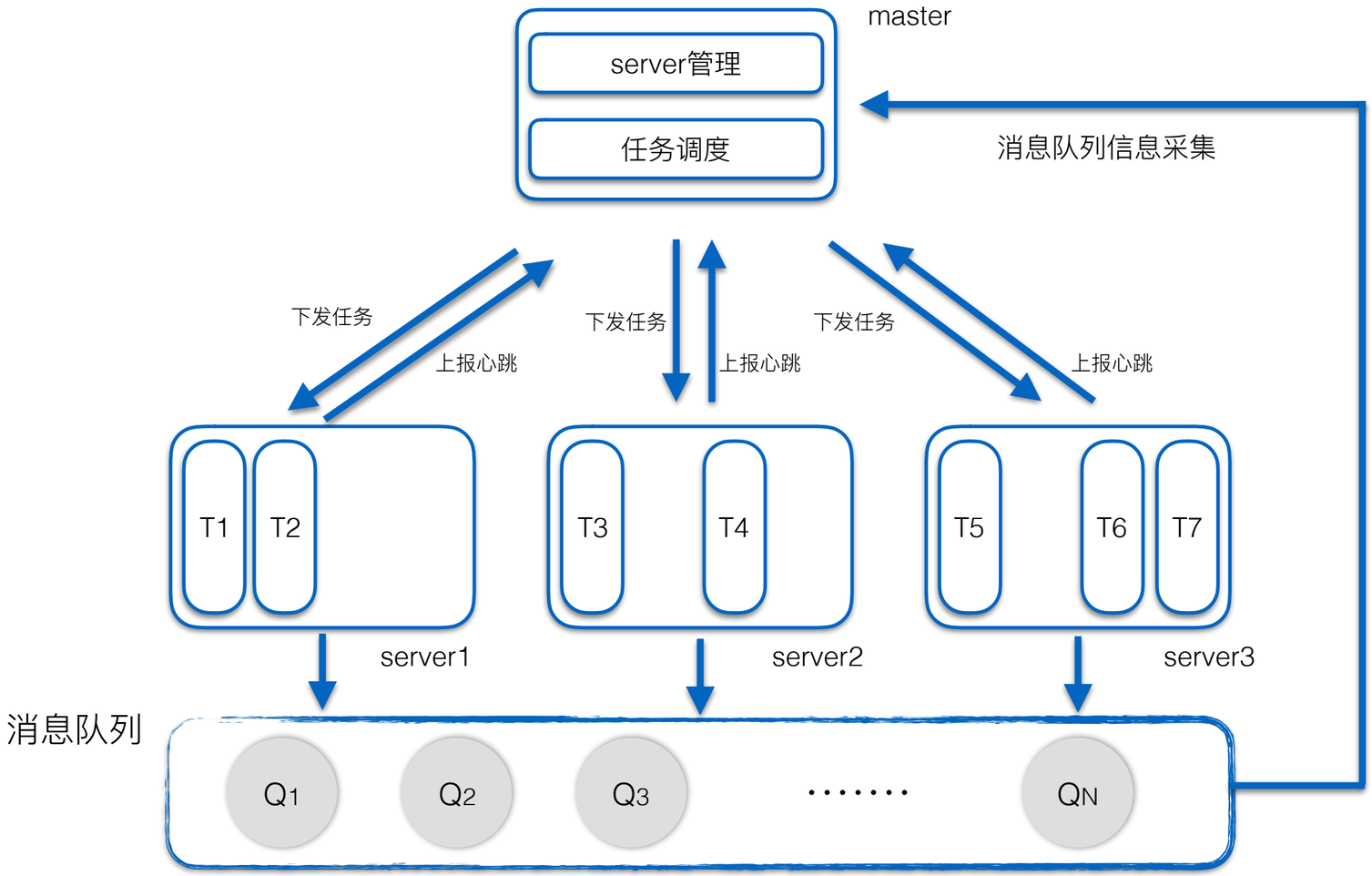
导出

连接上下游

- 任务切分
- 调度
- 任务自动均衡
- 水平扩展
- 资源隔离
- 高可用



任务切分与管理

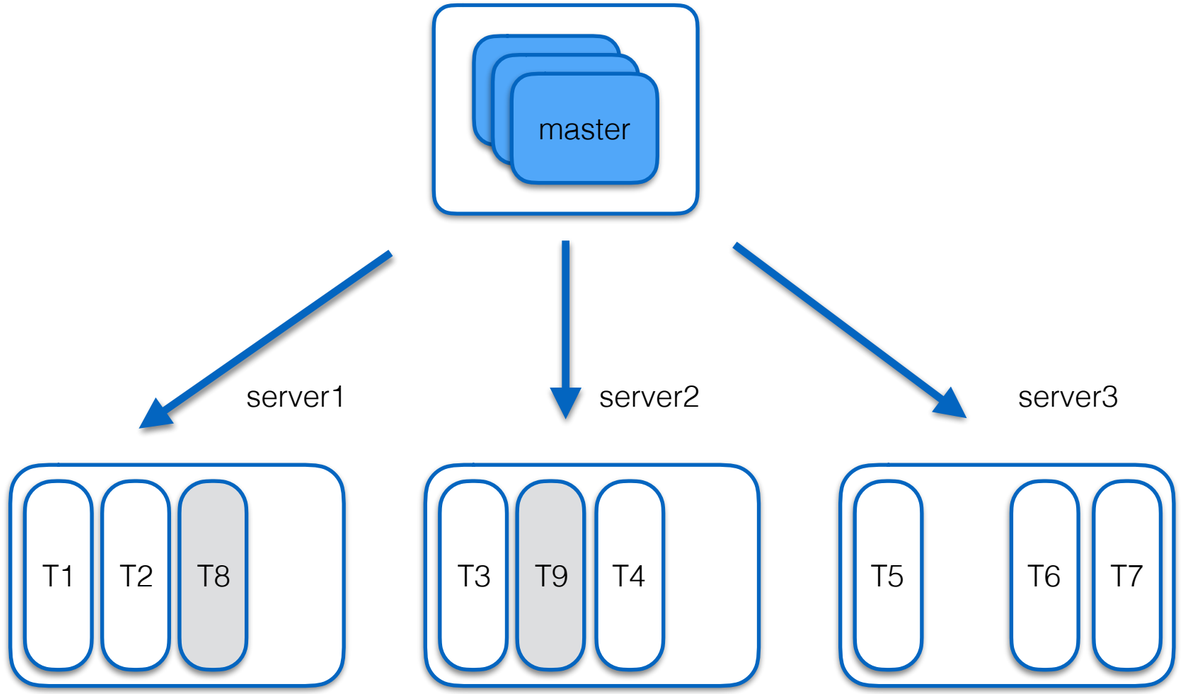

master
server管理
任务调度
消息队列信息采集
下发任务
上报心跳
下发任务
上报心跳
下发任务
上报心跳
T1
T2
T3
T4
T5
T6
T7
server1
server2
server3
消息队列
Q1
Q2
Q3
.....
QN

调度方法

- 面向资源
- 充分利用异构机器
- 自动调整

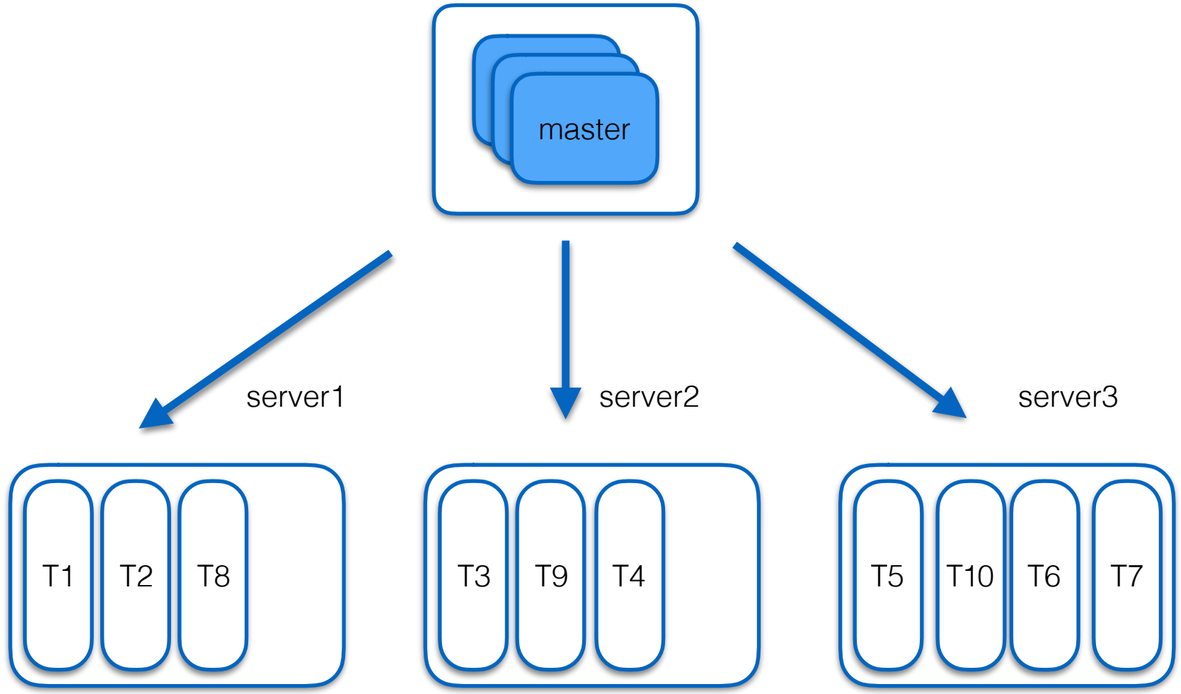
任务分配

- 任务均匀分配在server上
- T8和T9加入



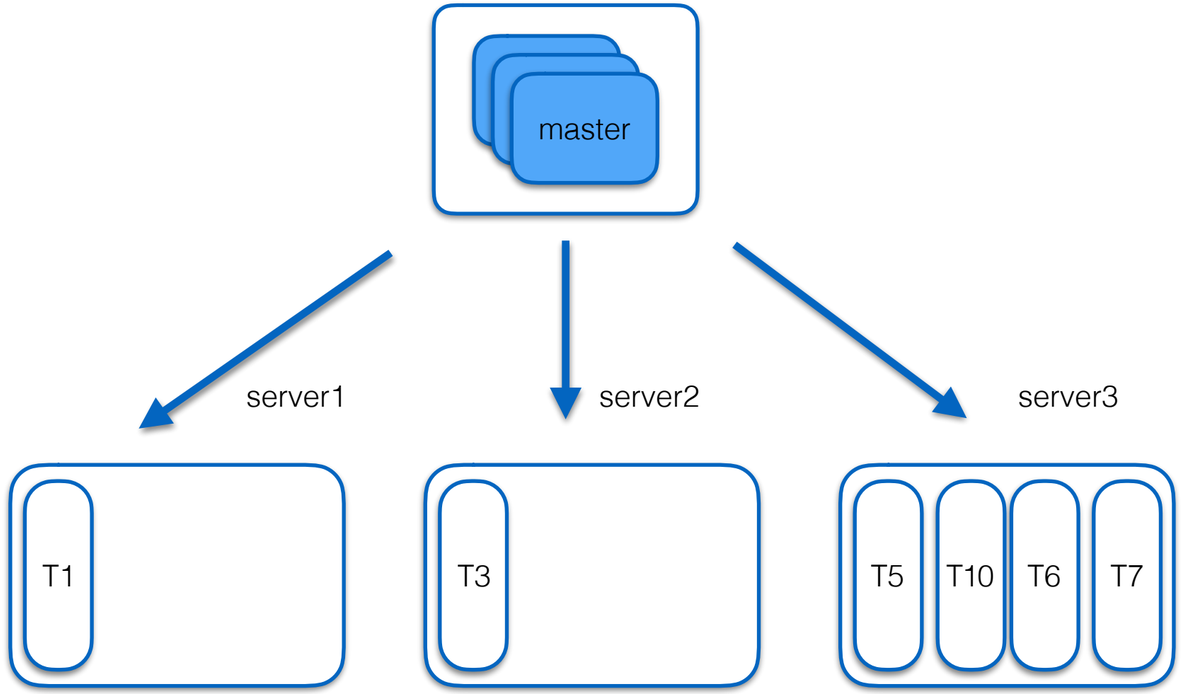
自动调整

- 任务均匀分配在3台server上



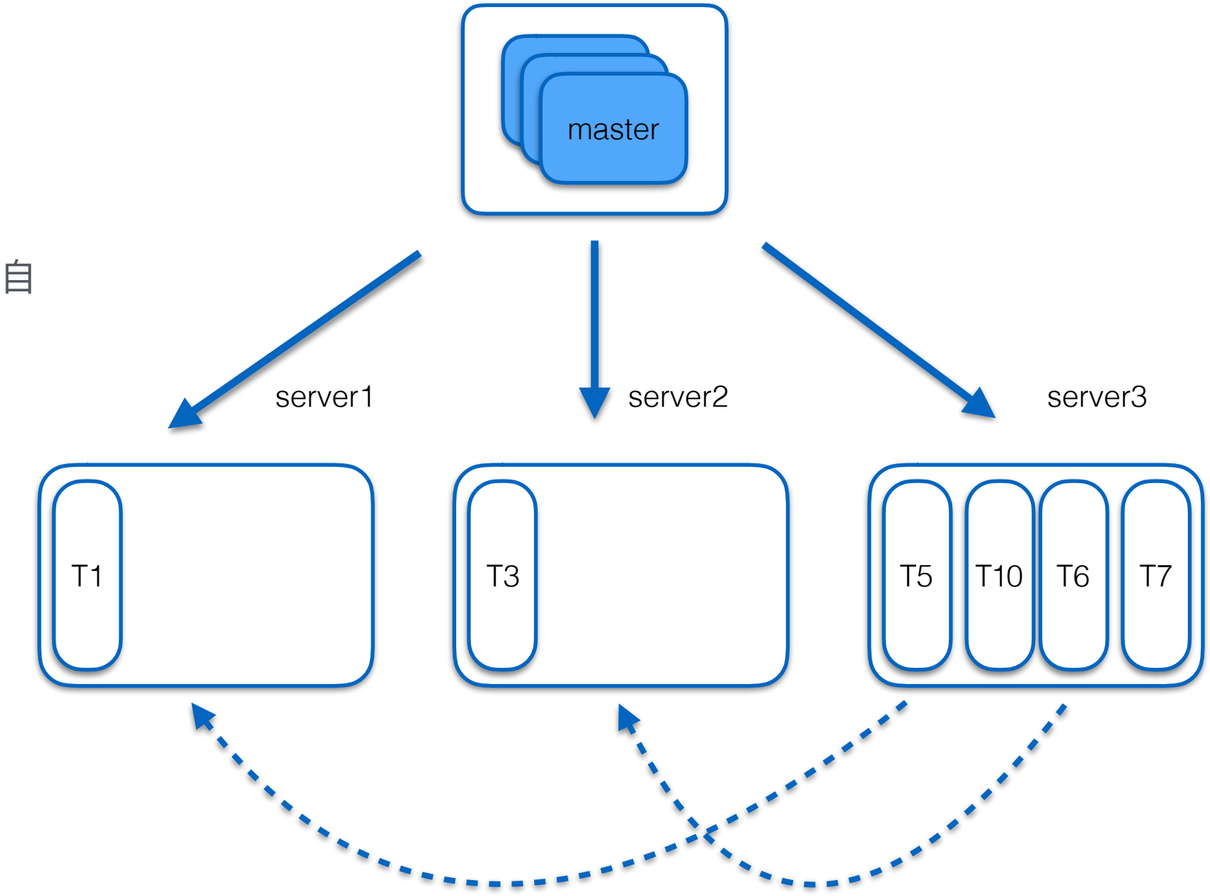
自动调整

- 任务均匀分配在3台server上
- T2、T8、T4、T9被删除



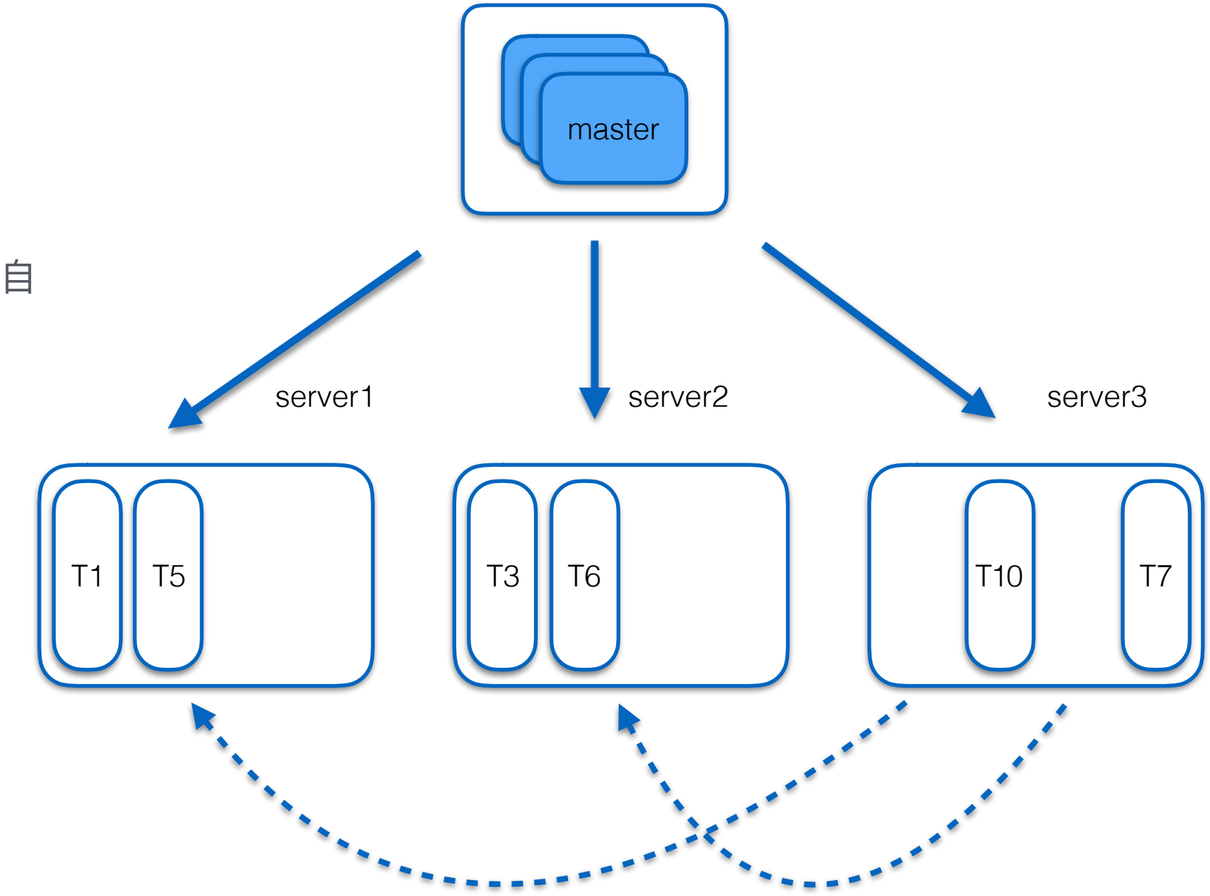
自动调整

- 任务均匀分配在3台server上
- T2、T8、T4、T9被删除
- 资源出现不均衡的情况，触发任务自动均衡



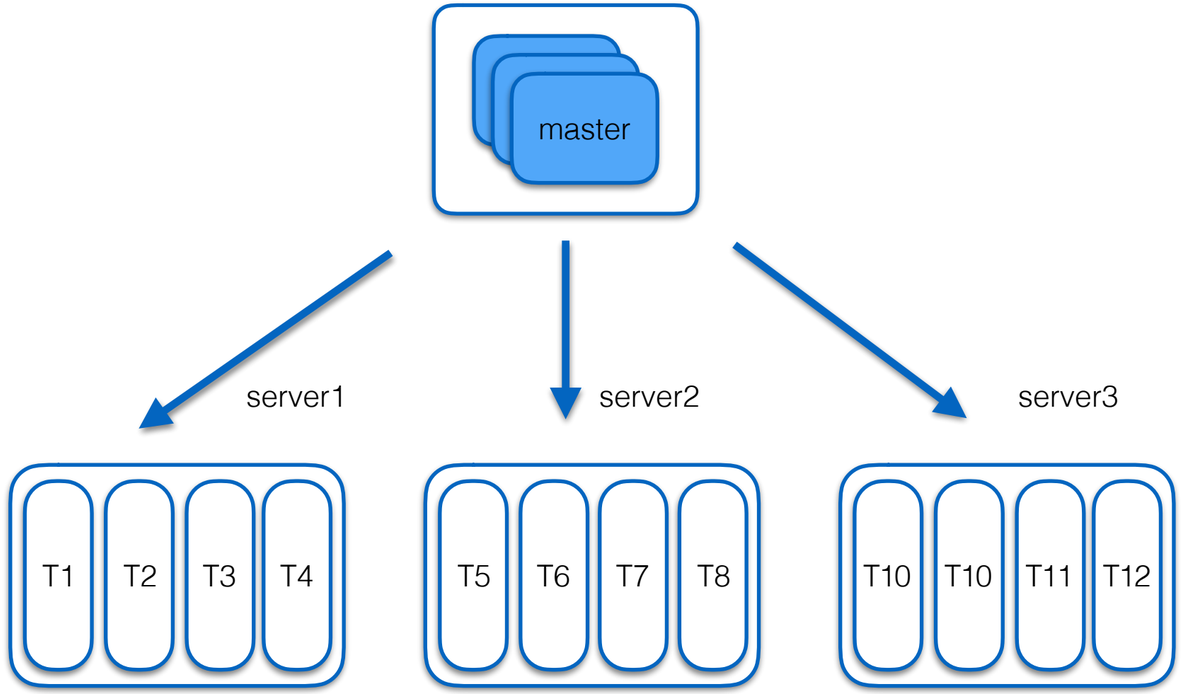
自动调整

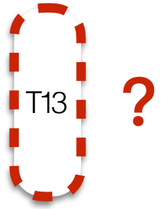
- 任务均匀分配在3台server上
- T2、T8、T4、T9被删除
- 资源出现不均衡的情况，触发任务自动均衡
- 调度任务至空闲机器



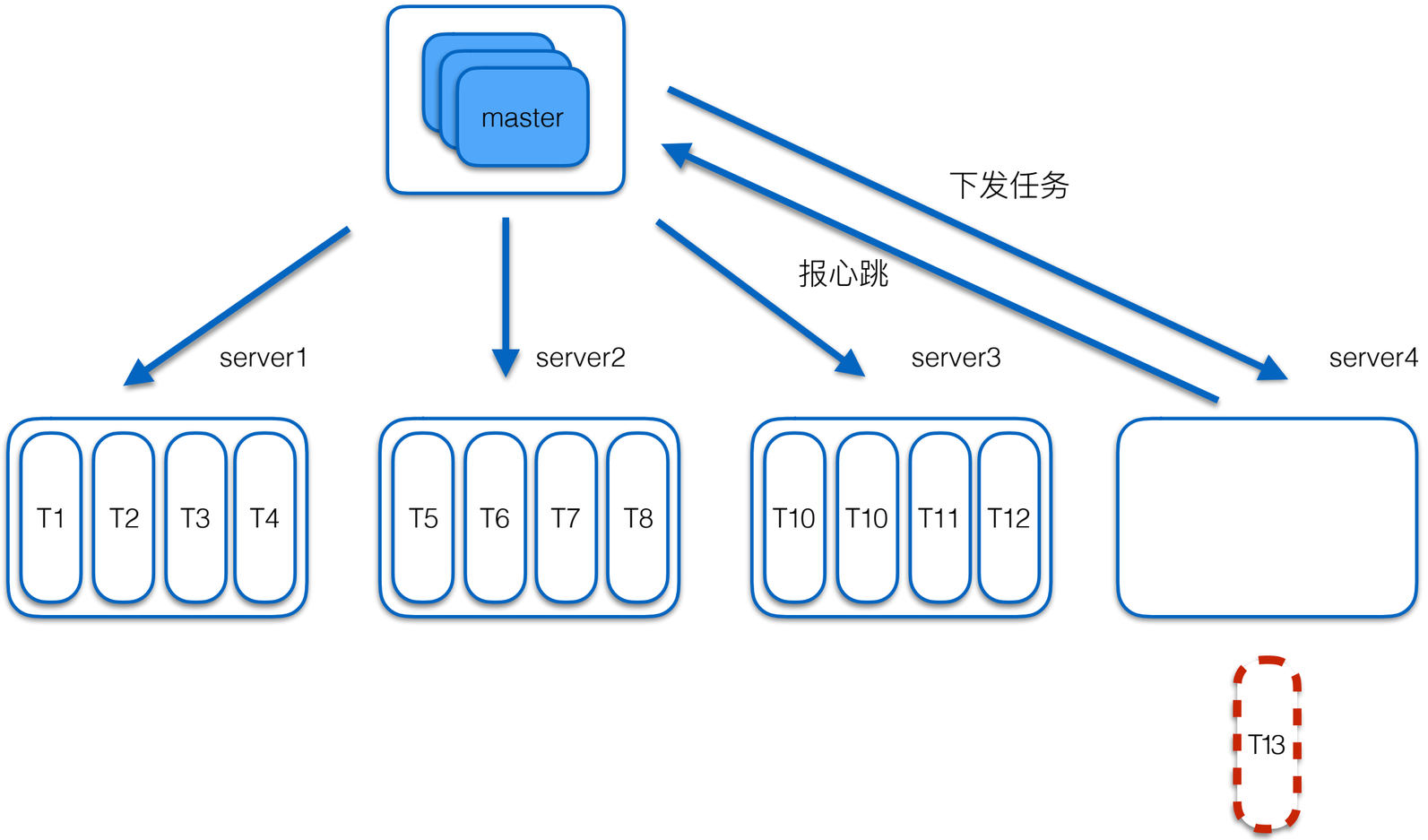
水平扩展

- 3台server已经全部处于满负载情况
- 新加入的任务T13无法被有效处理



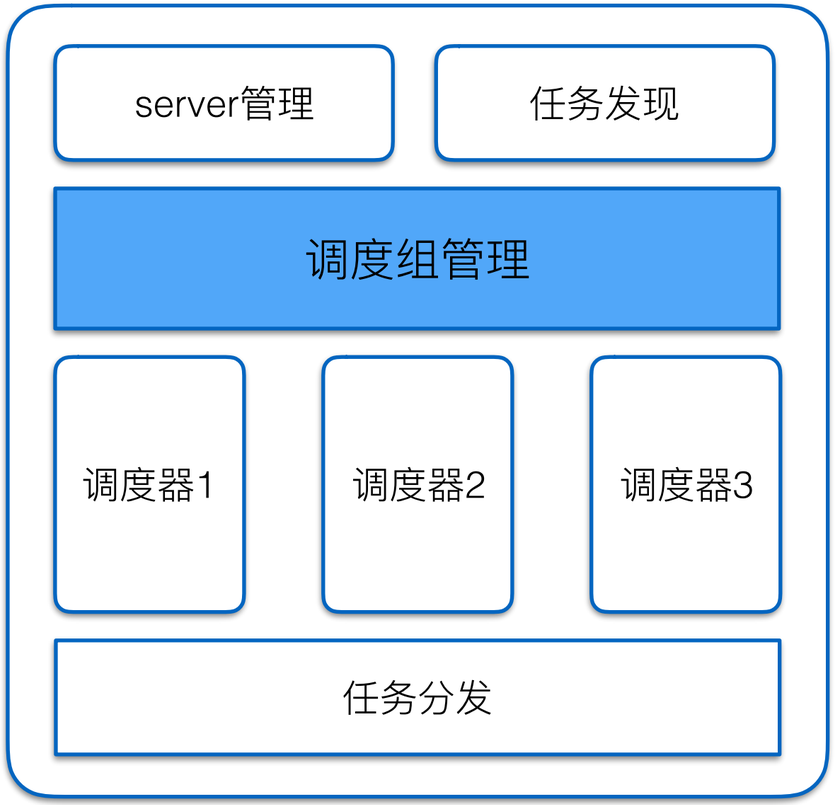


水平扩展


master
下发任务
报心跳
server1
server2
server3
server4
T1
T2
T3
T4
T5
T6
T7
T8
T10
T10
T11
T12
T13

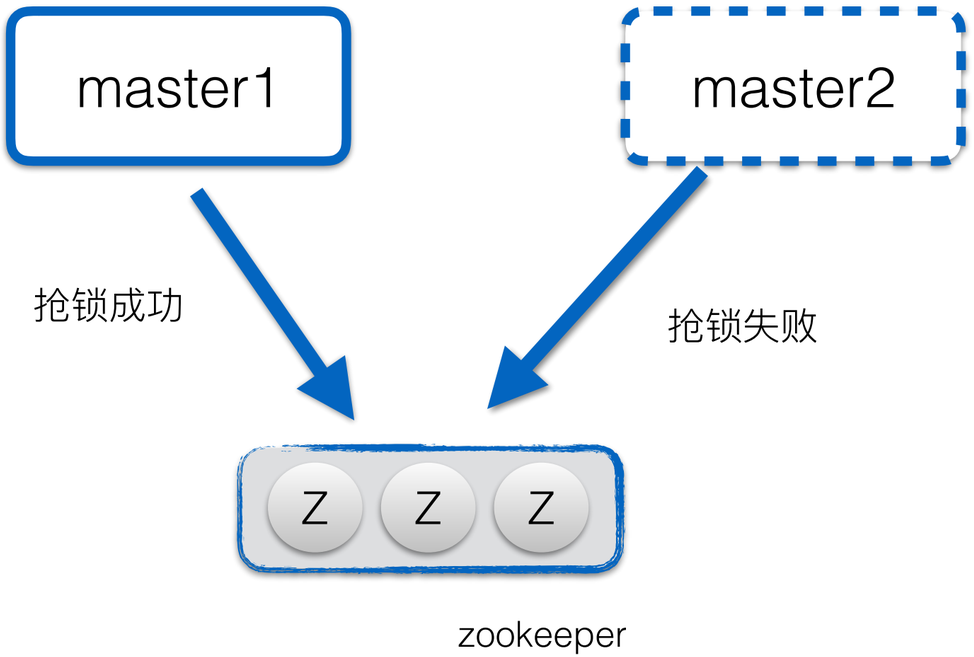
资源隔离

- 隔离特殊类型**任务**
- 利用特殊硬件**资源**
- 保证重要任务平稳



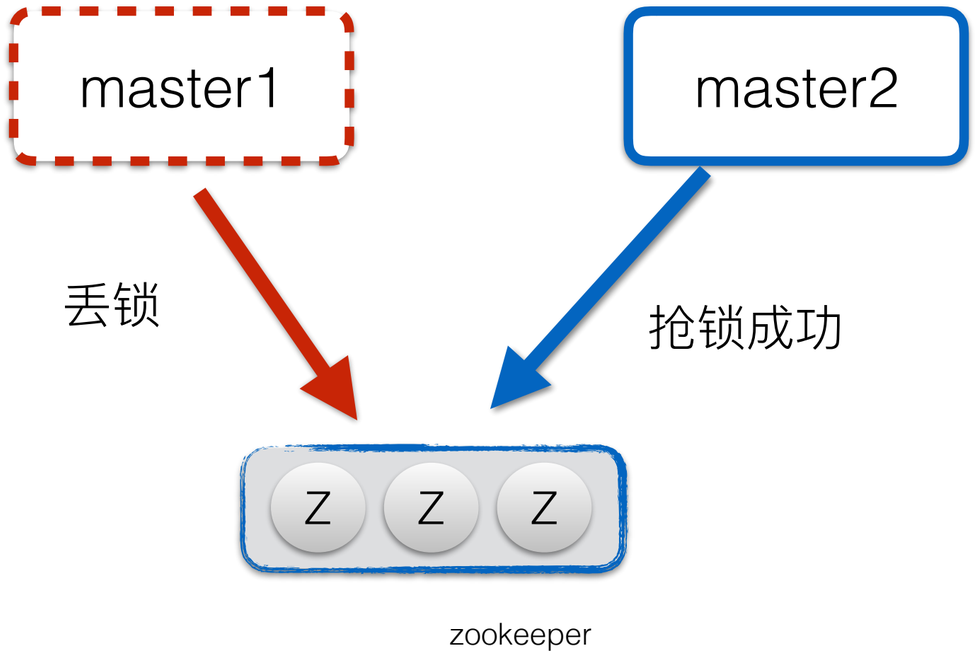
master高可用

- master通过抢锁来决定主和备
- 主master注册自己的身份到zk

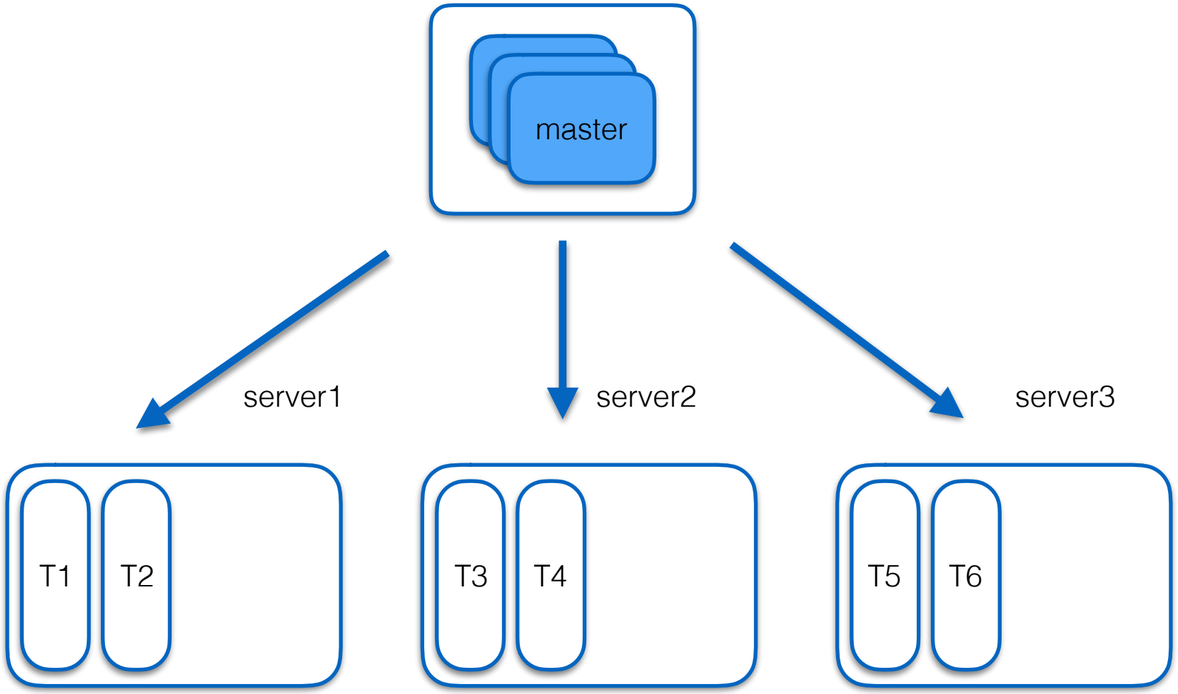


master高可用

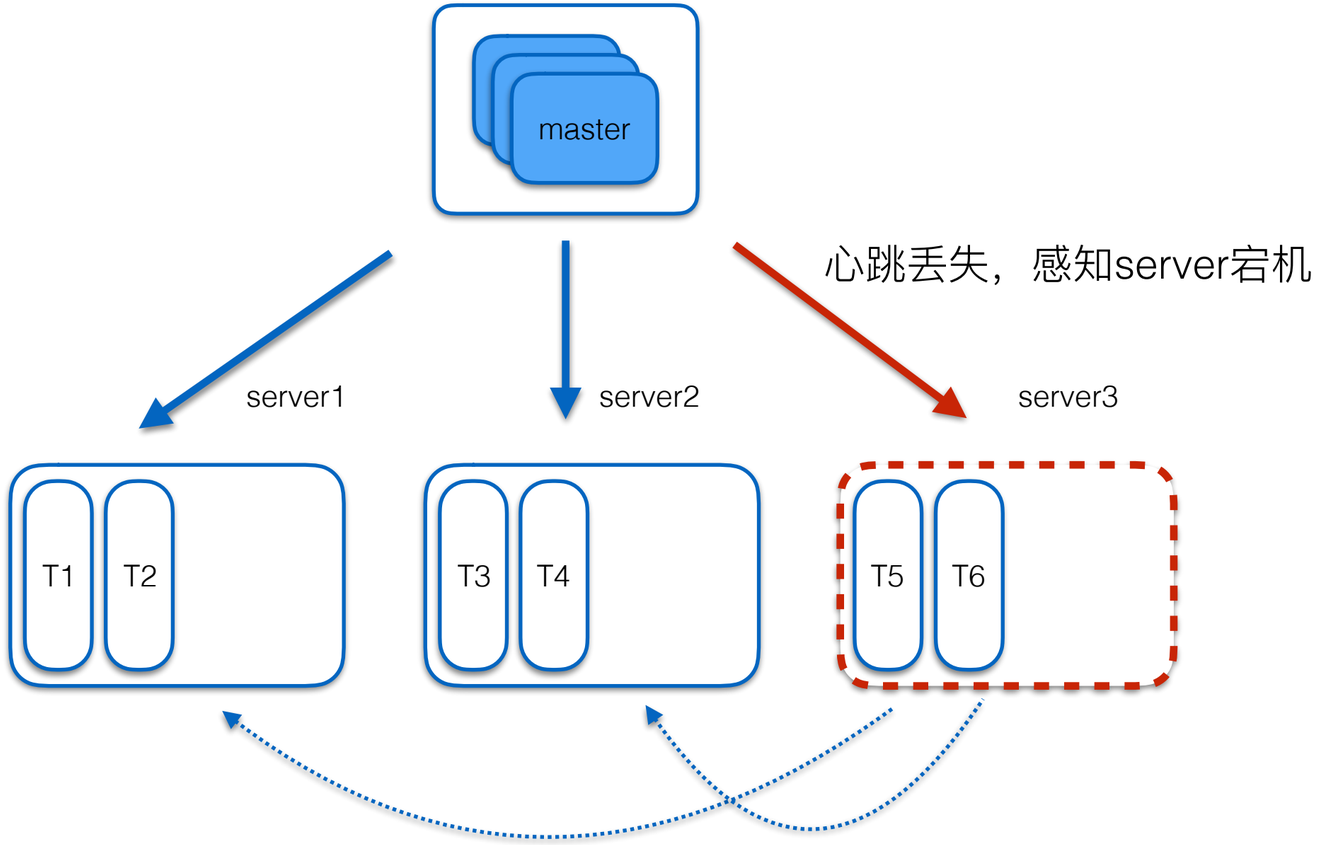
- master通过抢锁来决定主和备
- 主master注册自己的身份到zk



server高可用

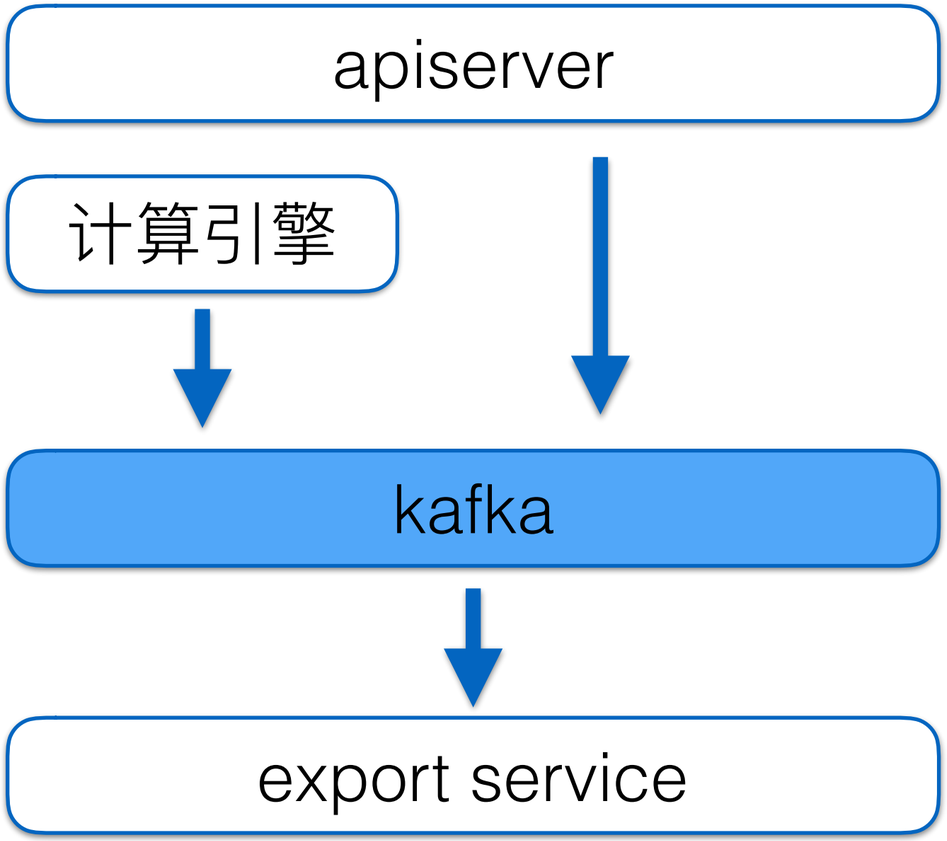


server高可用


master
心跳丢失，感知server宕机
server1
server2
server3
T1
T2
T3
T4
T5
T6

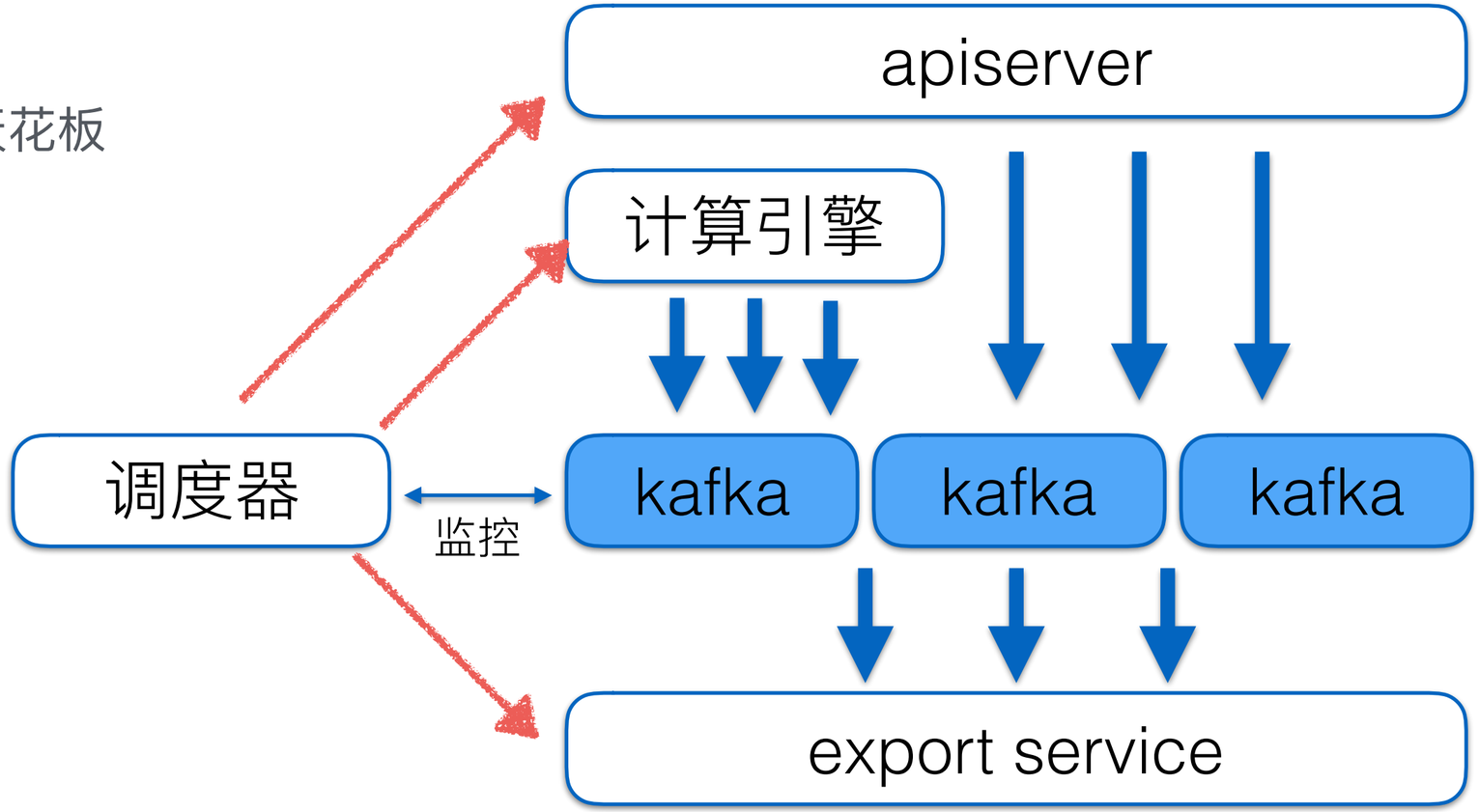
系统级水平扩展

- kafka单集群天花板
- 集群级scale



系统级水平扩展

- kafka单集群天花板
- 集群级scale



上下游协议优化

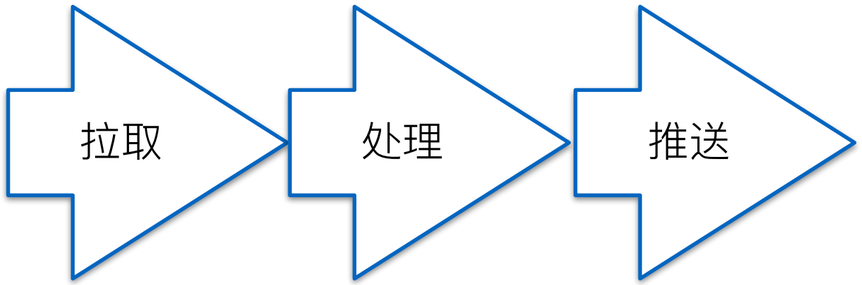
- **Json** vs **Protobuf**

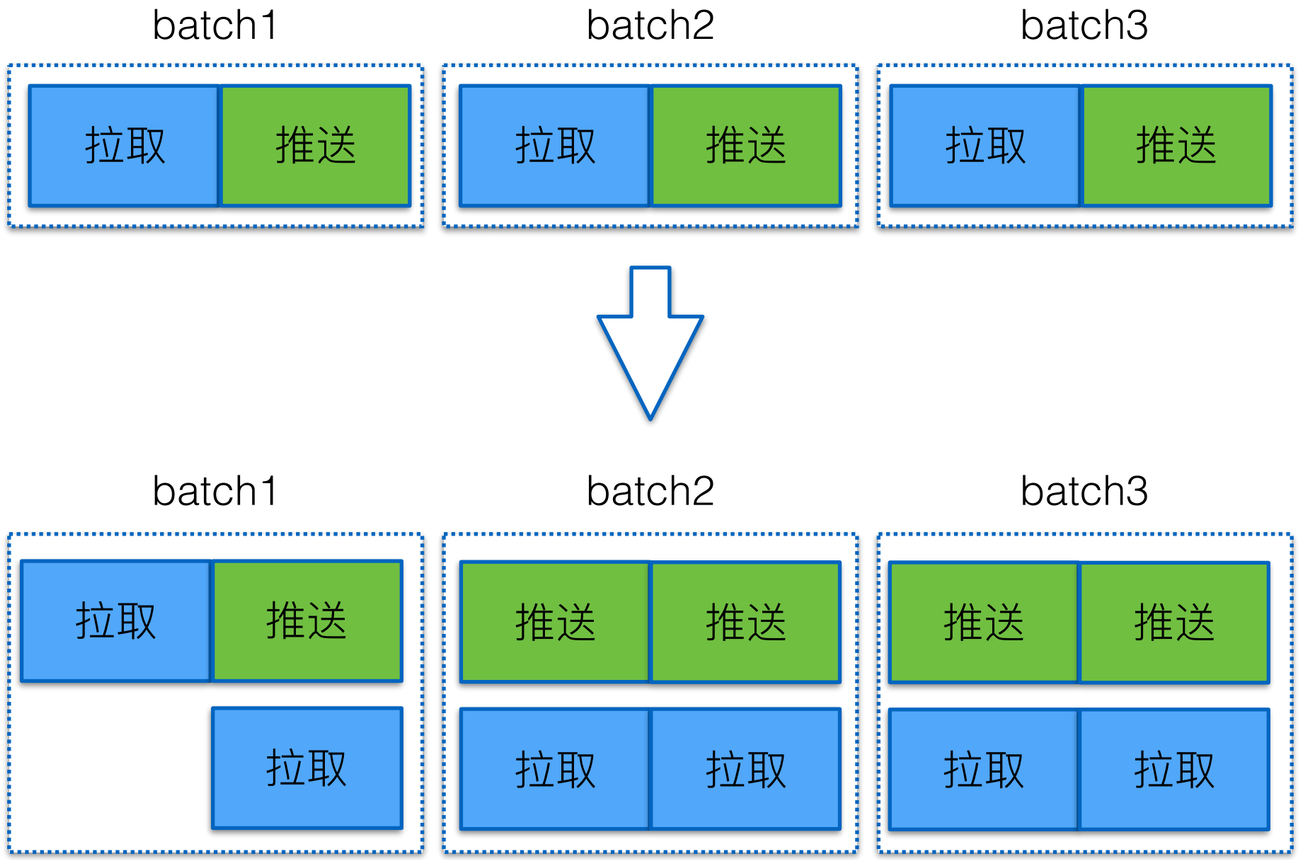
```
type Test struct {
    Uid             string `json:"uid"`
    BatchSize       int64  `json:"batchSize"`
    Hostname        string `json:"hostname"`
    Method          string `json:"method"`
    Operation       string `json:"operation"`
    Instance        string `json:"instance"`
    ReqBodyLength   int64  `json:"reqBodyLength"`
    ReqId           string `json:"reqId"`
    RespBodyLength  int64  `json:"respBodyLength"`
    RespCode        int64  `json:"respCode"`
    RespTime        int64  `json:"respTime"`
    Timestamp       int64  `json:"timestamp"`
}
```

项目	Json	Protobuf
序列化 (ns/op)	82161	67833
反序列化 (ns/op)	36380	7705
序列化长度 (byte)	259	100

流水线处理

- 导出处理模型
- 流水线并行处理





Golang GC

- stop the world
- sync.Pool
- 重用对象
- Golang版本升级

有限资源假设

- 单位资源服务能力
- 资源使用评估
- 资源规划

成果

- 每天支撑万亿级数据点、数百TB级数据量
- 支持海量用户
- 极低的系统延迟
- 自动化运维
- 可用性达到99.9%

Thank you!

七牛云 QINIU IT大咖说 知识分享平台

七牛云
QINIU.COM

简单·可信赖