

B站运维系统从无到有的演进之路

梁晓聪 B站 devops



About Me



- lxcong 梁晓聪
- 2015年加入B站
- Devops
- 热爱新技术,热爱开源,小宅男



B站



弹幕视频网站
1亿活跃用户
100万活跃UP主

“最大的同性交友网站”



故事的开始



B站炸了



超过1000万人正在使用



今天B站炸了吗

知识就是力量，法国就是培根，B站就是爆炸。

+ 关注

私信



她的主页

她的相册



Lv9



海外 日本



丧偶



2009年6月26日



简介：知识就是力量，法国就是培根，B站就是爆炸。



个性域名：yamanasion



今天B站炸了吗

6月6日 17:48 来自 微博 weibo.com

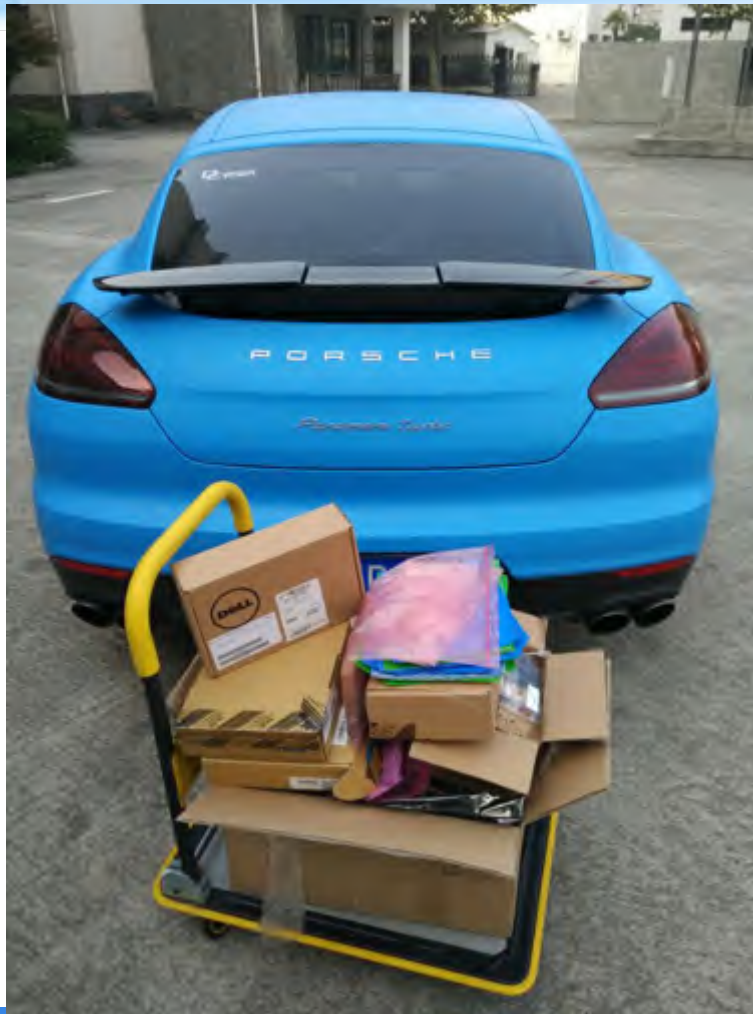
炸了

我的內心毫無波動



私信聊天





We Need



- 标准化
- 高质量交付
- 弹性伸缩





在运维人员紧缺的情况,如何标准化

文档转化成可执行代码





ansible-playbook

面向结果

幂等



```
- hosts: all
  gather_facts: True

vars:
  docker_package: "docker-engine_1.12.0-0~jessie_amd64.deb"
  consul_host: "172.18.21.230"
  consul_port: 8500
  influx_host: "172.18.17.10"
  influx_port: 8086
  influx_db: "monitor"
  influx_user: "monitor"
  influx_pass: "?????"
  mesos_master: "zk://172.18.17.101:2181,172.18.17.102:2181,172.18.17.103:2181/mesos"

bridges:
  - { "dest": "br1060", "src": "eth0.1060" }
  - { "dest": "br1061", "src": "eth0.1061" }
  - { "dest": "br1062", "src": "eth0.1062" }
  - { "dest": "br1063", "src": "eth0.1063" }

docker_networks:
  - { "name": "vlan1060", "subnet": "172.18.60.0/24", "opt": "parent=br1060" }
  - { "name": "vlan1061", "subnet": "172.18.61.0/24", "opt": "parent=br1061" }
  - { "name": "vlan1062", "subnet": "172.18.62.0/24", "opt": "parent=br1062" }
  - { "name": "vlan1063", "subnet": "172.18.63.0/24", "opt": "parent=br1063" }

roles:
  - { role: docker-install, tags: "docker-install" }
  - { role: mesos, tags: "mesos" }
```

TASK: [docker-install | u'创建co目录'] *****

ok: [pd-ops-vms-07]
ok: [pd-ops-vms-09]
ok: [pd-ops-vms-08]
ok: [pd-ops-vms-10]

TASK: [docker-install | u'创建需要的目录结构'] *****

changed: [pd-ops-vms-08] => (item=/docker)
changed: [pd-ops-vms-07] => (item=/docker)
changed: [pd-ops-vms-09] => (item=/docker)
changed: [pd-ops-vms-10] => (item=/docker)

TASK: [docker-install | u'创建/root/.docker'] *****

changed: [pd-ops-vms-09]
changed: [pd-ops-vms-07]
changed: [pd-ops-vms-08]
changed: [pd-ops-vms-10]

TASK: [docker-install | u'查看/mnt/storage00是否存在'] *****

ok: [pd-ops-vms-10]
ok: [pd-ops-vms-07]
ok: [pd-ops-vms-09]
ok: [pd-ops-vms-08]

TASK: [docker-install | u'bind /mnt/storage00 to /docker'] *****

skipping: [pd-ops-vms-07]
skipping: [pd-ops-vms-08]
skipping: [pd-ops-vms-09]
skipping: [pd-ops-vms-10]

TASK: [docker-install | u'上传相关文件'] *****

ok: [pd-ops-vms-09] => (item={'dest': 'u'/root/docker-engine_1.12.0-0-jessie_amd64.deb', 'src': 'u'docker-engine_1.12.0-0-jessie_amd64.deb'})
ok: [pd-ops-vms-10] => (item={'dest': 'u'/root/docker-engine_1.12.0-0-jessie_amd64.deb', 'src': 'u'docker-engine_1.12.0-0-jessie_amd64.deb'})
ok: [pd-ops-vms-09] => (item={'dest': '/etc/pki/CA/servercert.pem', 'src': 'servercert.pem'})
ok: [pd-ops-vms-10] => (item={'dest': '/etc/pki/CA/servercert.pem', 'src': 'servercert.pem'})
ok: [pd-ops-vms-09] => (item={'dest': '/etc/pki/CA/serverkey.pem', 'src': 'serverkey.pem'})
ok: [pd-ops-vms-10] => (item={'dest': '/etc/pki/CA/serverkey.pem', 'src': 'serverkey.pem'})
ok: [pd-ops-vms-09] => (item={'dest': '/etc/pki/CA/cacert.pem', 'src': 'cacert.pem'})
ok: [pd-ops-vms-10] => (item={'dest': '/etc/pki/CA/cacert.pem', 'src': 'cacert.pem'})
ok: [pd-ops-vms-09] => (item={'dest': '/opt/iptables-switch.sh', 'src': 'iptables-switch.sh'})

堡垒机执行

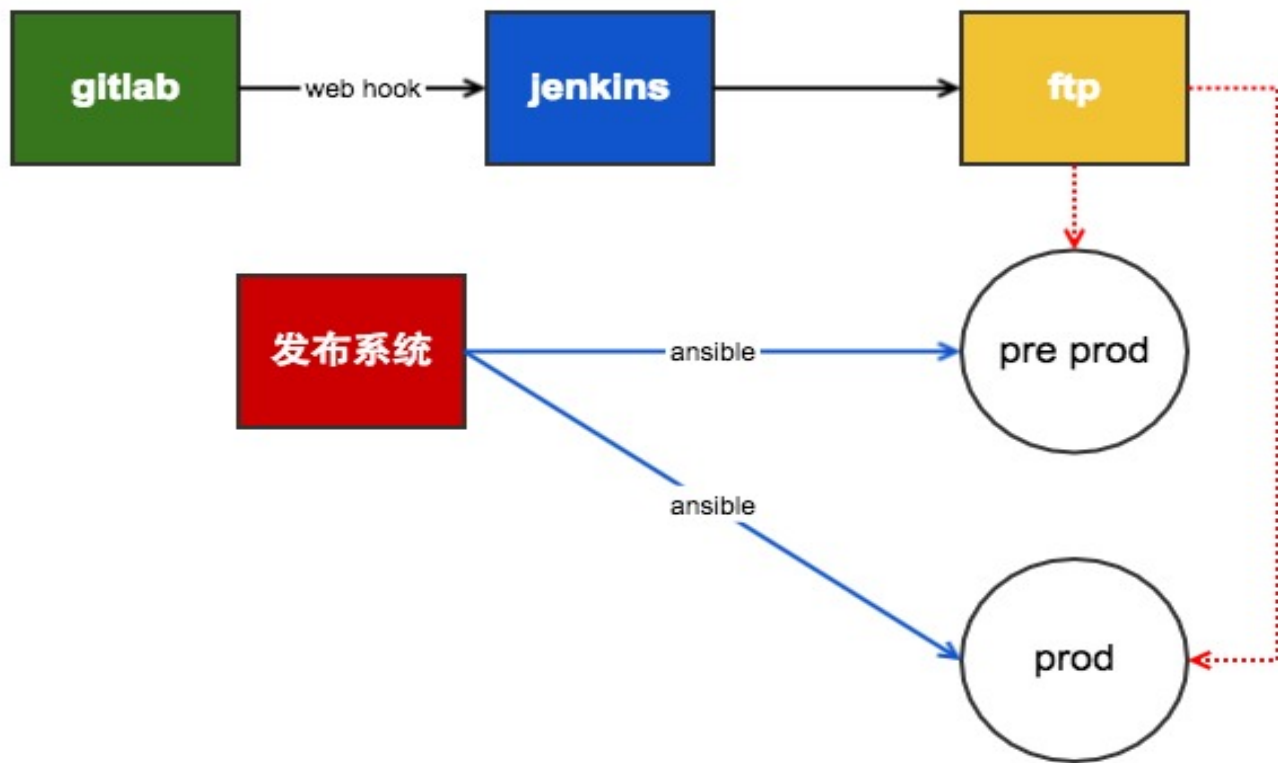
as docker_sh docker/site.yml



如何提高交付质量

标准化发布,快速反馈





开发者视图

发布

故障报告

自助http监控

白山云CDN监控

阿里云管理

Publish / 正式环境发布

发布信息

发布项目

app-api

发布目录

/data/app/app-api

Jenkins工程

app-api

项目域名

ENV环境变量

程序启动命令

发布前执行shell

cd {{download_base}}/{{ project }}/{{ build_num }}
chown -R nobody:nogroup ./
chmod a+x ./*

发布后执行shell

supervisorctl restart app-api
mv /data/www/appcheck.bak
/data/www/appcheck.html

开始发布

选择需要发布的服务器
(默认全部)

search...

shd-pf-app-01

shd-pf-app-02

shd-blizone-01

shd-blizone-02

shd-blizone-03

shd-blizone-04

search...

发布版本

构建号:236 // 构建时间:2016-08-18 01:05:24 //

并发部署数

10

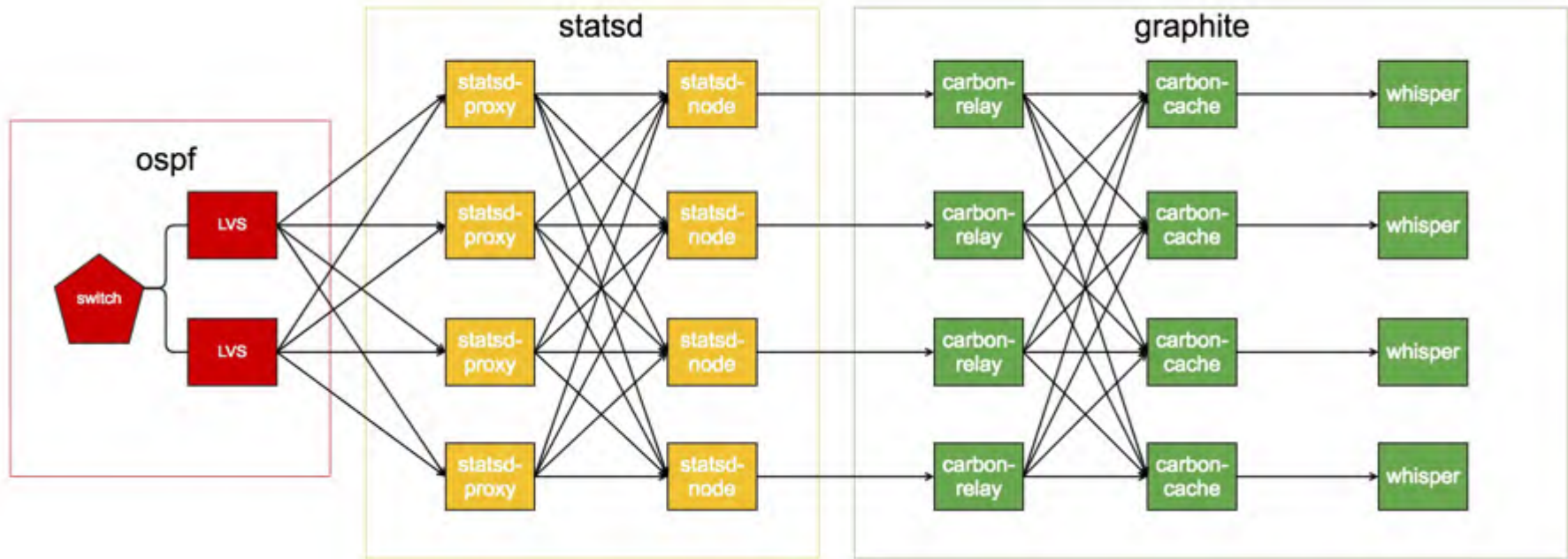
正式发布



测量 VS 监控

you can't optimise what you can't measure







```
start = time.time()  
// code which connects and sends email  
end = time.time()  
  
cost = int(end - start)  
statsd.timing("api.sendEmailResponseTime", cost);
```



测量系统场景



- 接口访问量,平均访问耗时
- 最近成功支付的数量
- 采样统计cache的key大小
- 等等等....





bilizone接口访问量



Zoom Out

6 hours ago to a few seconds ago



server: All method: return_code: 0

All

bilizone_health_check

bilizone_monitor_ping

bilizone_reload

localdomain

x_ad_loc

x_ad_video

x_admin_account_refresh

x_admin_ad_load

x_admin_archive_bangumi

x_admin_archive_bp

x_admin_archive_cache_del

x_admin_archive_delay

x_admin_archive_download



T-bilizone-app-1
app-test
bilizone-pre
pd-mainsite-app-01
shd-bilizone-01
shd-bilizone-02
shd-bilizone-03
shd-bilizone-04



avg current

26 3

136.0 K 83.5 K

133.1 K 113.4 K

133.0 K 122.9 K

137.4 K 69.1 K

method:All return = 0



avg current

386.1 K 280.7 K

接口访问耗时/ms

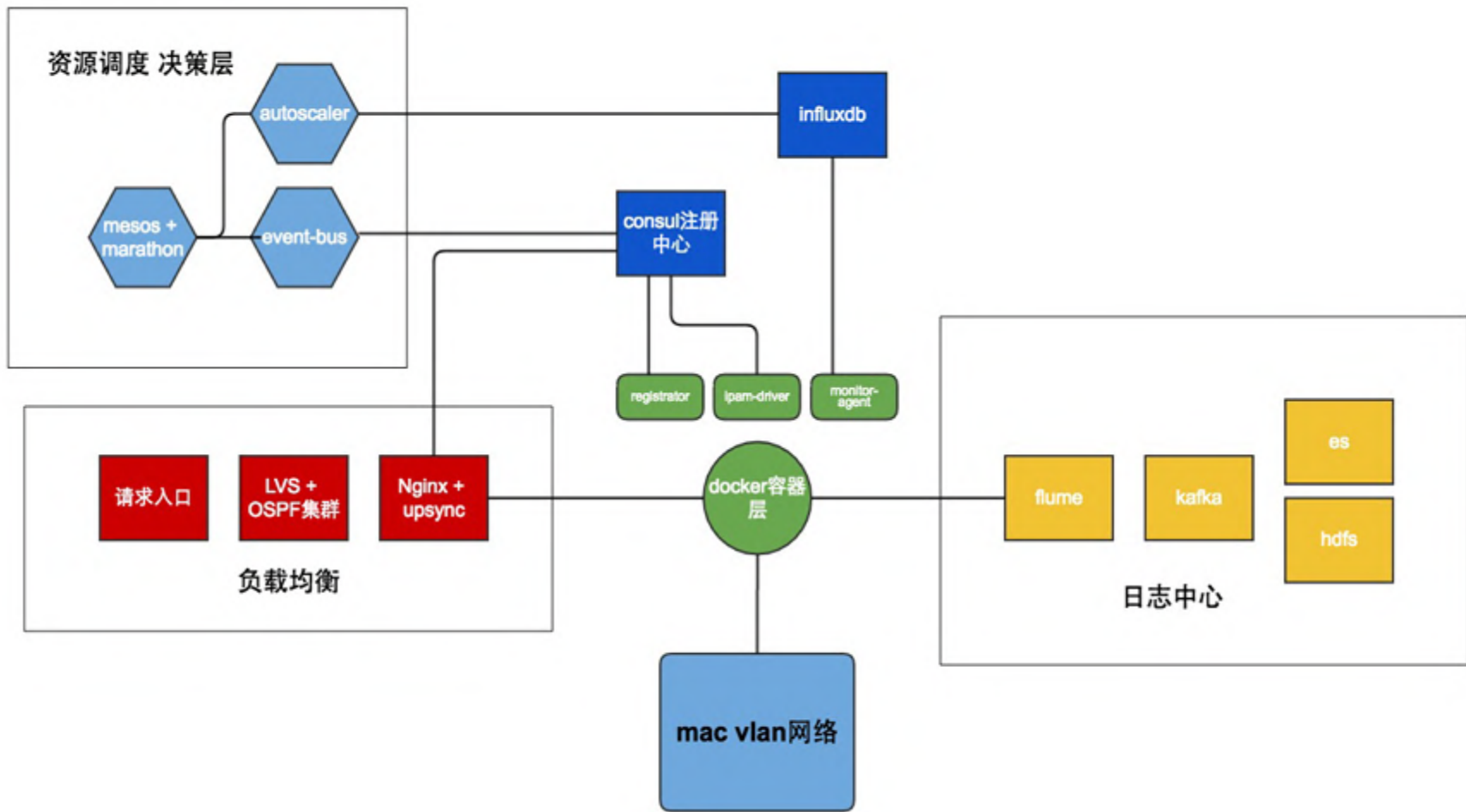




如何弹性伸缩

基于docker的可弹性伸缩web集群







负载均衡

- 难点
 - 自动伸缩/故障转移/代码迭代造成的ip变更
 - nginx reload的性能影响
- 解决
 - nginx+ upsync + consul





监控

- 开源方案
 - cadvisor
 - prometheus
- 考虑
 - 接入成本大
 - 新坑
- 最终
 - 自研监控agent



监控



influxdb

monitor agent in docker

metric

tags

cpu

mem

disk io

network

mesos id

container
ip

container
name

labels



注册中心

consul

register info

ip pool

tengine upstream

服务发现

ipam plugin

upsync

container
id

ip

env

...

可用地
址

已用地
址

网关

ip:port

max fails

timeout

...

网络

需求:

- ip per container
- 业务网络隔离
- 尽量简单

可以排除bridge,host方式



网络

站队:

- cni (Container Network Interface)
- cnm (Container Network Model)
 - 亲儿子



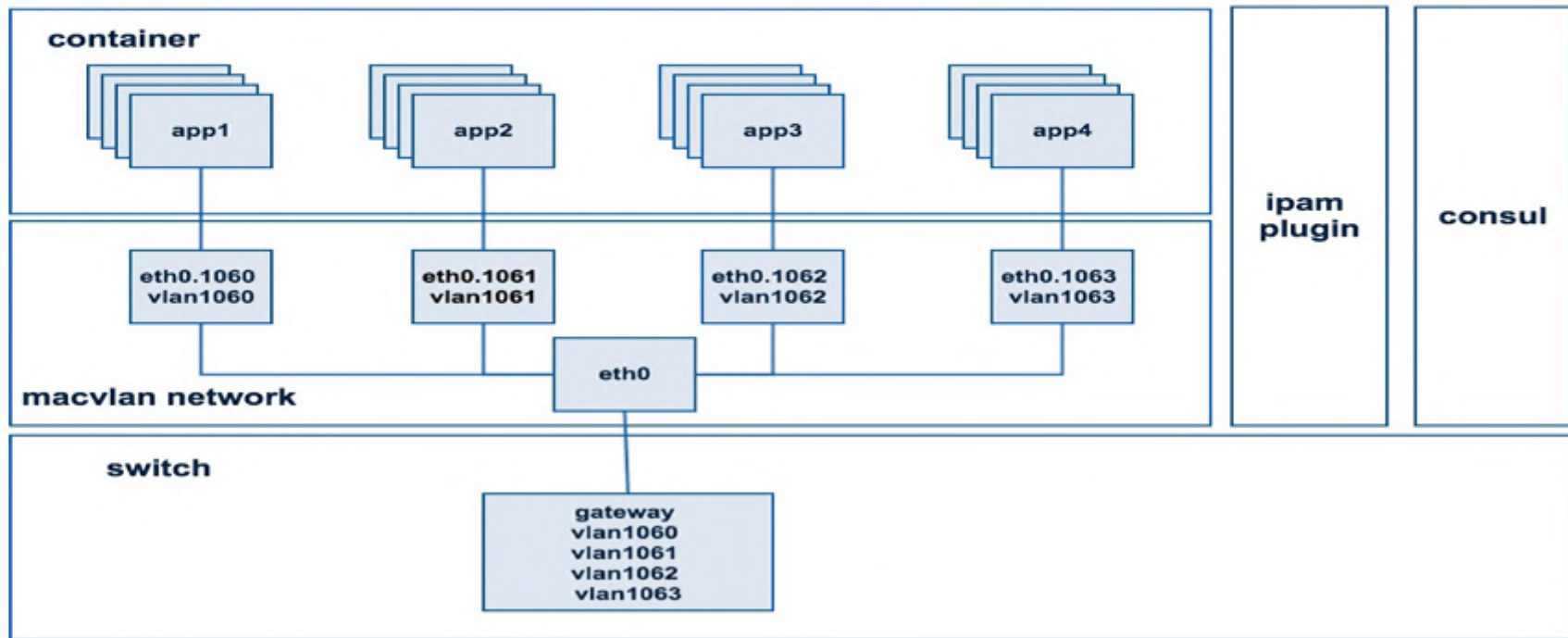
网络



方案选型:

- xvlan技术 overlay
 - 没有多租户隔离需求
 - 排障成本大
 - 没有性能优化能力
- 3层路由 calico
 - node-mesh方式 导致大量路由
 - node-router方式 核心设备bgp peer太多
- 2层隔离 macvlan
 - 简单
 - 性能损耗低
 - 合并到docker 1.12 release





网络



ipam-plugin:

- consul
- flask api
 - RequestAddress
 - ReleaseAddress
 - RequestPool
 - ReleasePool



日志中心



- log driver syslog
- flume
- kafka
- hdfs/elasticsearch

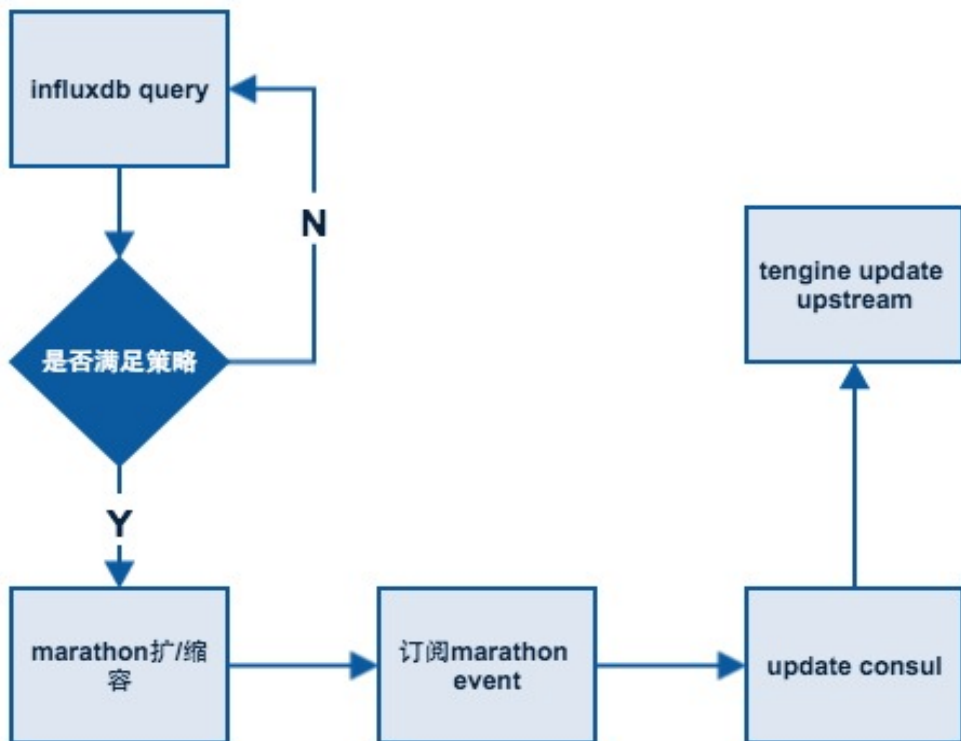


自动伸缩



- mesos + marathon
- autoscaler
- marathon hook event bus





策略

```
{  
  "app" : "/web/app_name",  
  "autoscale_multiplier" : 0.2,  
  "limit" : "mean_cpu_usage > 300 or  
mean_memory_percent > 80",  
  "max_instances" : 15,  
  "min_instances" : 6,  
  "times" : 3 ,  
  "counter" : 0  
}
```



四 todo



- swarm mesos结合
- 更细化的autoscaling策略
- CI/CD的完整流程





- 强大的内心
- 寻找切入点
- 小步快跑
- 不要用战术上的勤奋掩盖战略上的懒惰



THANKS

哔哩哔哩 - (° - °)つ口 乾杯~ - bilibili

