2018

第九期

360互联网技术训练营

360容器技术解密与实践

自我介绍

董永彬：360 netops 部门网络开发工程师，从事奇虎网络自动化开发和基于DPDK相关项目的开发。目前专注研究基于dpdk高性能snat和抗DDoS防护。

360在容器网络部署的实践

一、容器网络方案的选取

二、360网络架构优化满足容器网络的需求

三、360容器网络部署自动化

四、360容器网络配置优化

一、容器网络方案的选取

1、容器网络简介

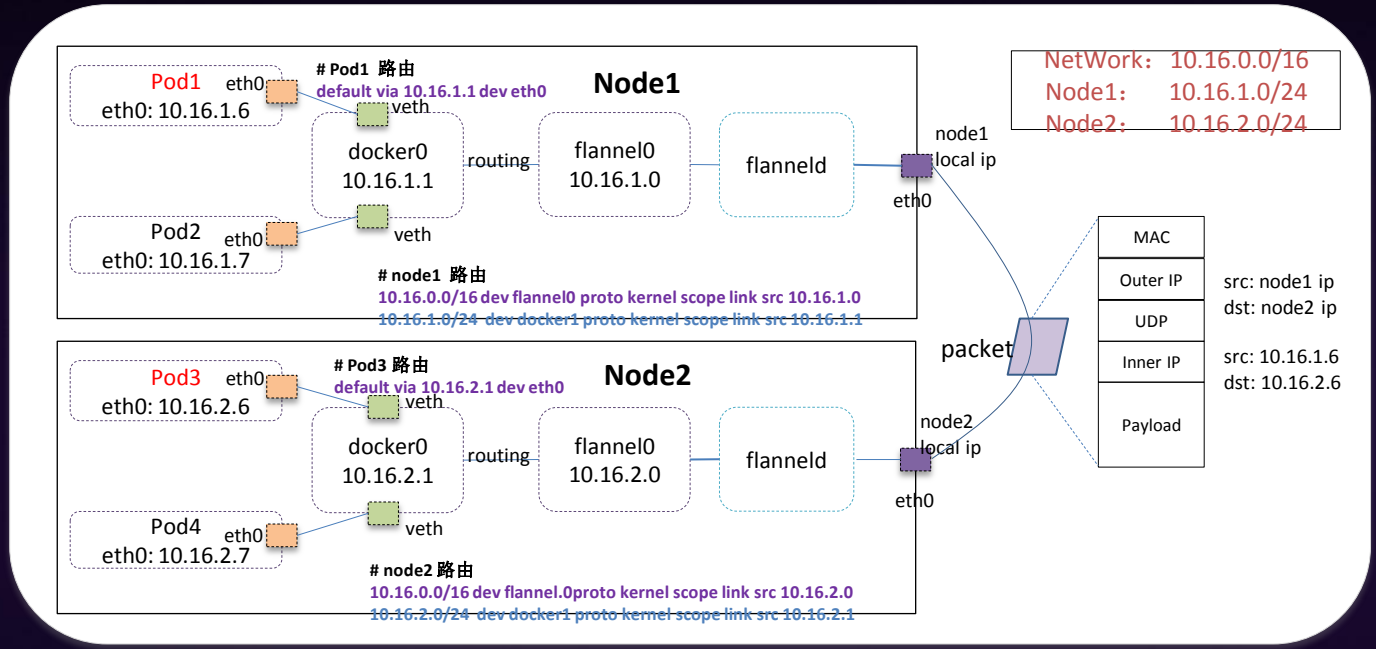
随着容器技术的发展，给传统的网络提出了一些新的挑战，Docker本身的网络方案比较简单，所以围绕Docker产生了很多不同的网络解决方案，下面对比一下当前比较流行的三种方案。

	Calico	Flannel	Weave
组网模型	纯L3 路由方案	Vxlan or UDP 封装	Vxlan or UDP封装
协议支持	TCP,UDP,ICMP	所有协议	所有协议
分布式存储	Etcd分布式存储	Etcd分布式存储	不需要(Rumor协议)
时延测试	最小	次之	最大
BPS测试	最高	次之	最小
CPU使用率	最小	次之	最大

一、容器网络方案的选取

2、flannel 网络介绍

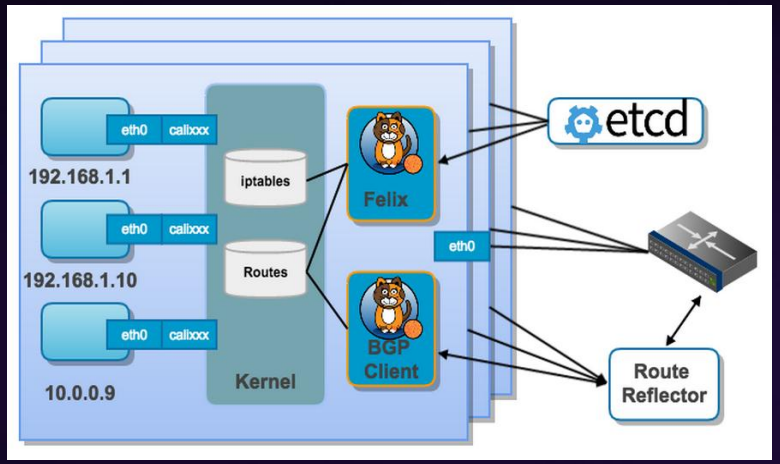
flannel是CoreOS提出用于解决Dokcer集群跨主机通讯的覆盖网络工具。



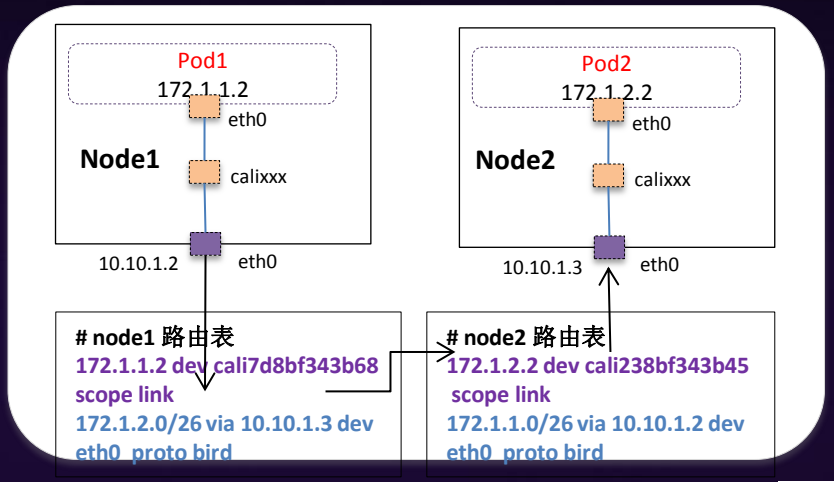
一、容器网络方案的选取

3、Calico 网络介绍

Calico 是一个三层的数据中心网络方案，能够为容器之间提供高效可控通信。



Calico的核心组件图

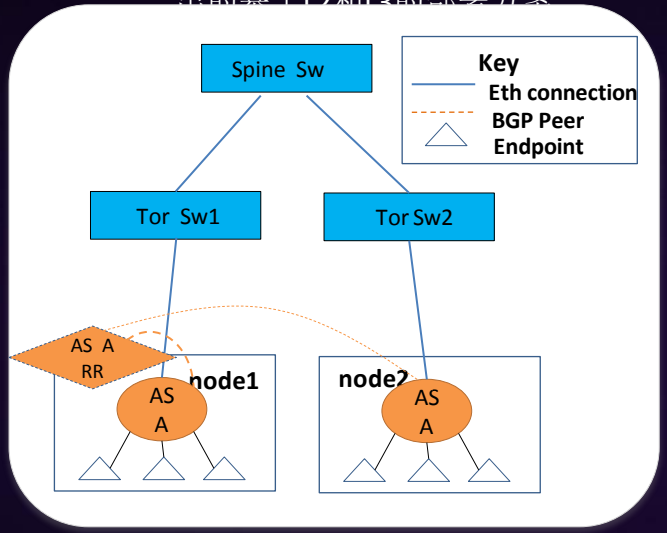


容器跨主机通信图

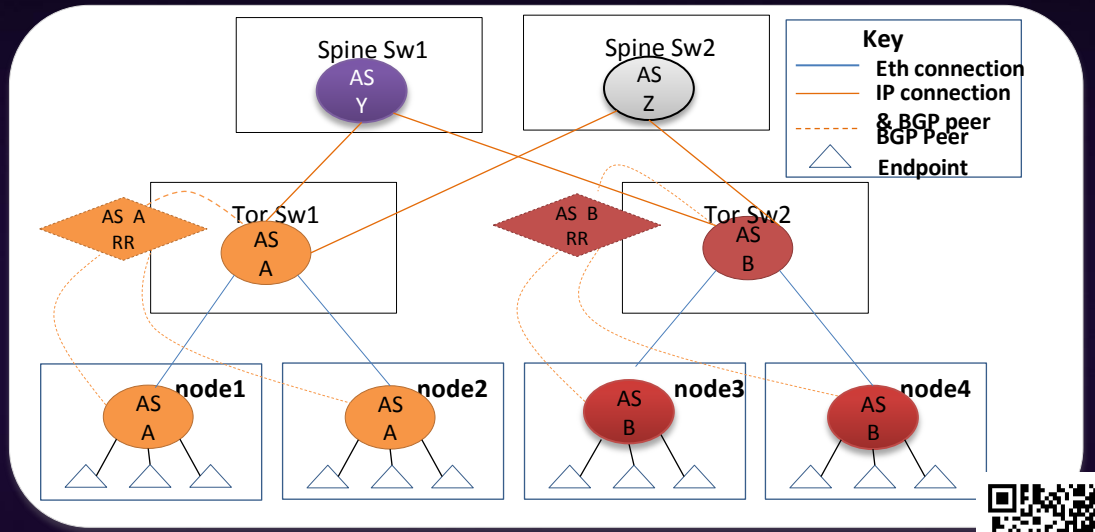
一、容器网络方案的选取

calico网络的部署方案

Calico部署方式分为L2 Fabric和L3 Fabric部署，在 L2中node间都是二层可达，node不需要把交换机当做下一跳；如果是L3 Fabric，物理交换机要存容器的32位路由，会给交换机带来压力，下面介绍一下Calico官网给出的基于L2和L3的部署方案



L2 层部署calico网络



L3层部署calico网络

一、容器网络方案的选取

Calico 优化：“Downward Default model” 减少需要记录的路由

在上面的L3网络的组网方式中，所有的Node、Tor交换机和Spine交换机都需要记录全网路由。

“Downward Default model”模式中:

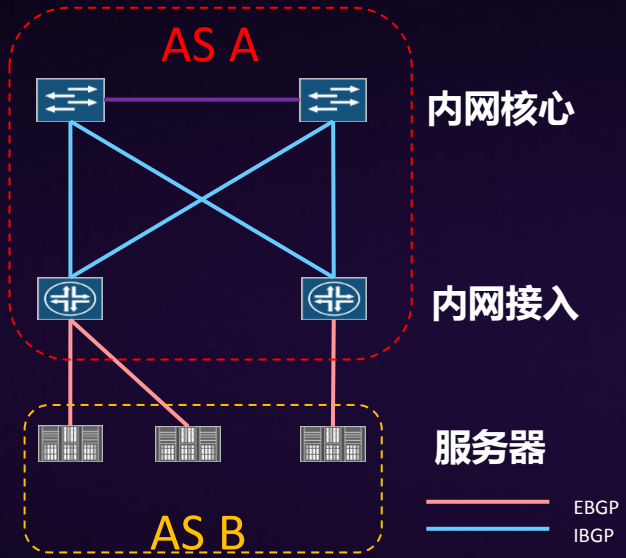
1、每个Node向上Tor交换机通告所有路由信息，而Tor向Node只通告一条默认路由

2、每个Tor交换机向上Spine交换机通告所有路由，Spine交换机向Tor交换机只通告一条 默认路由

这种模式减少了Tor交换机和node上的路由数量，但缺点是，Node发送的无效IP的流量必须到达Spine交换机以后，才能被确定为无效。

二、360网络架构优化满足容器网络的需求

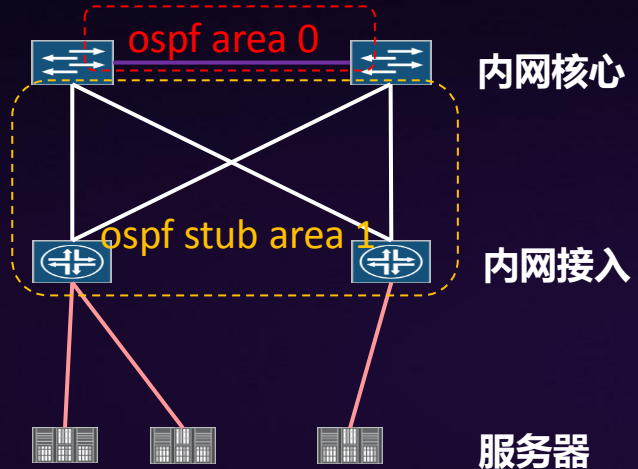
1、BGP机房改造方案



- **服务器与内网接入建立EBGP邻居**
  1、服务器发网段路由给内网接入
  2、内网接入给服务器发默认路由
- **内网核心与内网接入建立IBGP邻居**
  1、内网接入发网段路由给内网核心
  2、内网核心给内网接入发默认路由

二、360网络架构优化满足容器网络的需求

2、OSPF机房原始结构



OSPF机房结构图

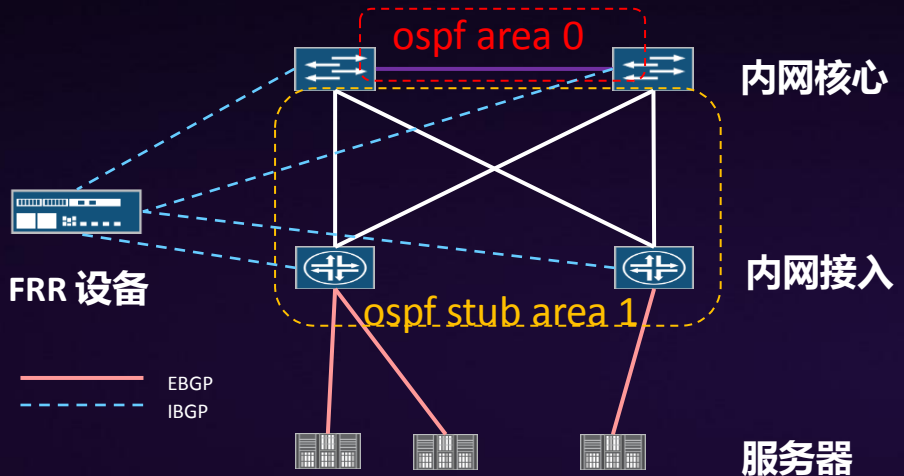
- **内网接入在ospf的stub区域**

  1、Stub区域是ospf特定的区域，该区域只能将ospf的路由传递到本区域，不会引入自治系统外部路由。

  2、这样当服务器与内网接入建立EBGP的话，容器的网段路由是不能发到内网核心的，因此，必须对网络结构进行改造。

二、360网络架构优化满足容器网络的需求

3、OSPF机房改造方案一

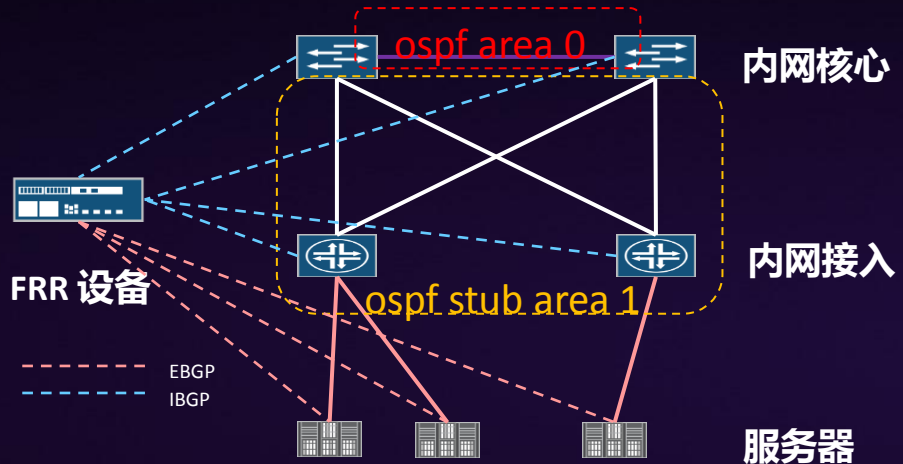


OSPF机房改造图一

- **服务器与内网接入建立EBGP邻居**
  1、服务器发网段路由给内网接入
  2、内网接入给服务器发默认路由
- **FRR分别与内网核心和内网接入建立IBGP邻居**
  1、FRR配成RR，这样内网接入和内网核心就学习到了容器全网的路由

二、360网络架构优化满足容器网络的需求

4、OSPF机房改造方案二



OSPF机房改造图二

- **FRR与服务器建立EBGP邻居**
  1、服务器给FRR发送网段路由
- **FRR分别与内网核心和内网接入建立IBGP邻居**
  1、FRR配成RR，把从服务器收到的网段路由发送到内网接入和内网核心

三、 360容器网络部署自动化

1、面临的问题

当业务部门想在某台服务器上开启Calico BGP时，必须发邮件告知我们部门，收到邮件之后，在该服务器所连接的内网接入交换机上人为添加相应的配置。

（1）没有数据库记录，带来后期维护的复杂性

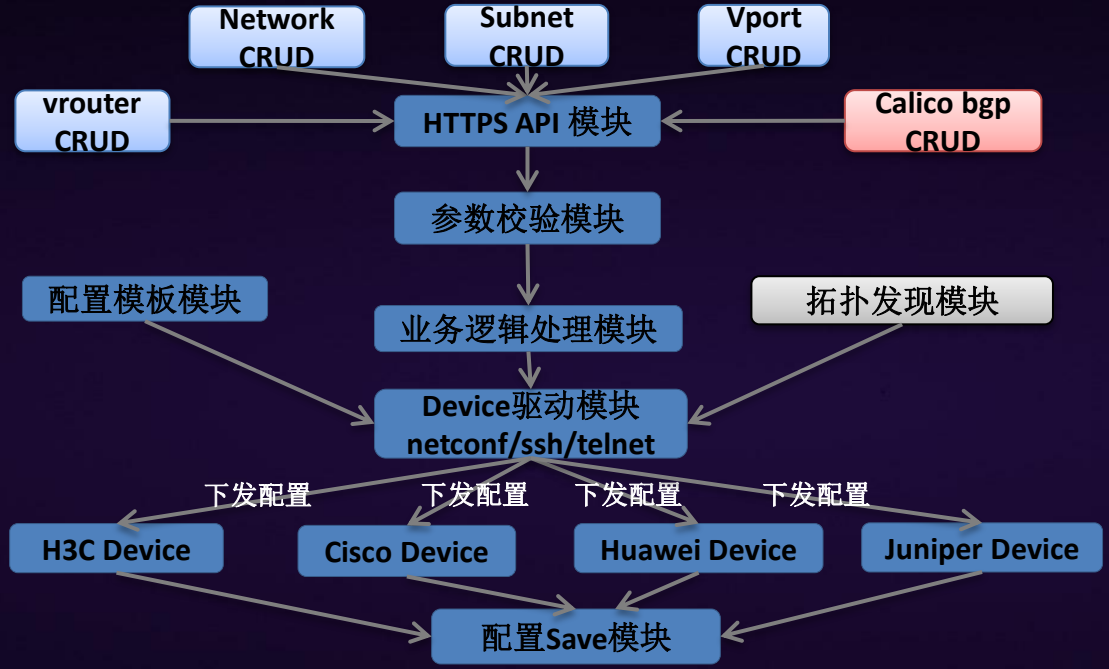
（2）人员排查的网络问题较多，来不及及时处理

（3）配置的标准化问题以及 double check 时间

三、 360容器网络部署自动化

2、NOSA自动化平台

NOSA（network operation standard API）：网络操作标准接口，主要对内或对外提供开放的HTTPS接口操作网络设备，最终目标是实现奇虎网络设备配置的自动化。



三、 360容器网络部署自动化

3、Calico BGP API介绍

（1）在网络设备上创建bgp邻居：

curl -XPOST -d '{"idc":“bjdt", "ip": host_ip, "mac": mac}' https://hostname/nosa/docker_bgp_peer/

（2）查询网络设备上bgp邻居信息：

curl -XGET -d '{"idc":“bjdt", "ip": host_ip, "mac": mac}' https://hostname/nosa/docker_bgp_peer/

（3）删除网络设备上的bgp邻居信息：

curl -XDELETE -d '{"idc":“bjdt", "ip": host_ip, "mac": mac}' https://hostname/nosa/docker_bgp_peer/

四、 360容器网络配置优化

1、Calico线上遇到的问题

Calico需要更新和维护，当restart Calico时，不同node之间的容器互访不通，主要因为BGP邻居关闭，到达对方的路由消失；为了使不同node容器能够互通，保证原通告的路由信息不被清除，我们对BGP的配置进行了优化，加入了 BGP Graceful Restart 配置。

2、Graceful Restart简介

BGP GR（Graceful Restart，平滑重启）是一种在主备倒换或**BGP协议重启时**保证转发业务不中断的机制，GR是BGP协议的一个特性，BGP 邻居之间必须同时配置GR，BGP GR服务才能生效。

四、 360容器网络配置优化

3、BGP GR 配置

交换机端配置：

```
router bgp 65999
   bgp router-id 1.1.1.1
   bgp graceful-restart restart-time 120
   bgp graceful-restart stalepath-time 360
   bgp graceful-restart
   neighbor 1.1.1.12 remote-as 65998
   address-family ipv4
      neighbor 1.1.1.12 activate
```

Calico端配置：

```
protocol bgp {
     description “1.1.1.1";
     local as 65998;
     neighbor 1.1.1.1 as 65999;
     multihop;
```

谢谢

HULK一线技术杂谈

360技术