云网分析与可视化

发掘网络数据的真正价值

向阳@云杉网络

QCon2017@北京

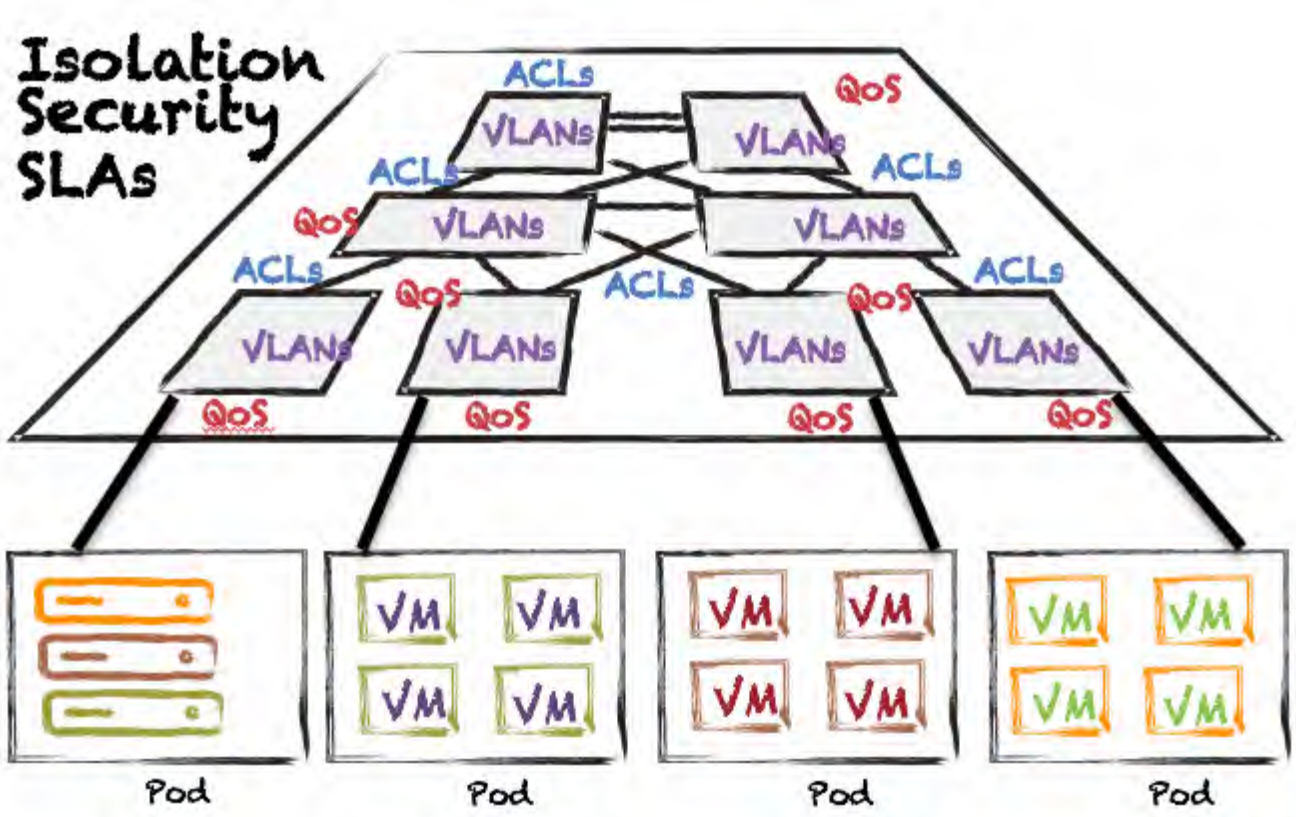
云网：云数据中心网络

- 云网的变迁
- 云网运维的痛
- 摆脱云网运维的困境
  - 数据采集
  - 数据分析
  - 数据存储

云网的变迁

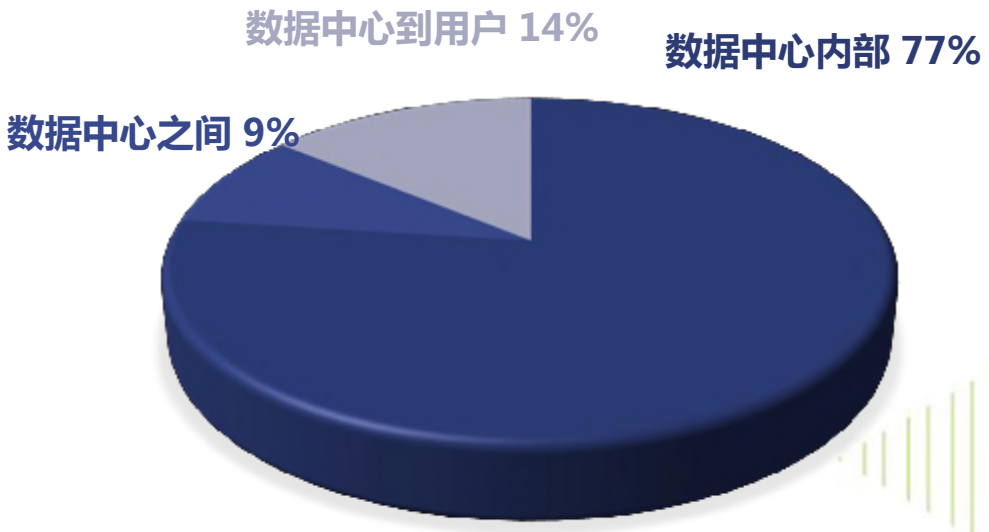
由实到虚

- 从烟囱式到集中化
  - 单租户独享 → 多租户共享
- 从Underlay到Overlay
  - VLAN
  - VXLAN
  - EVPN



由外到内

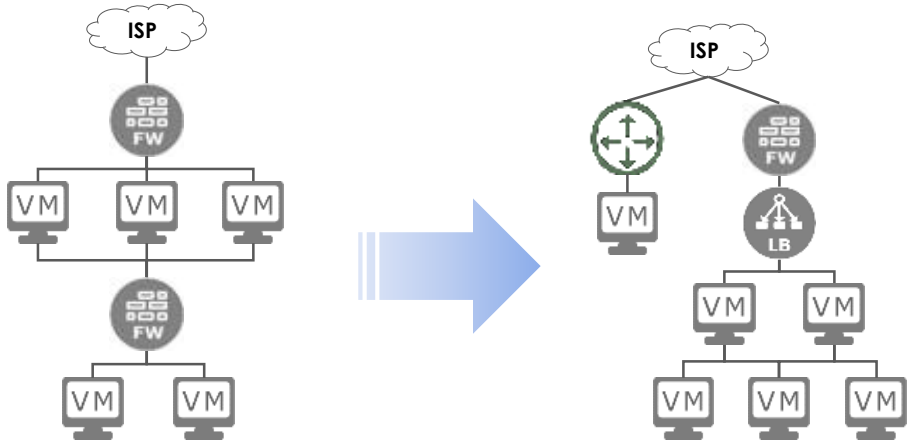
- 2020年，云数据中心将处理92%的工作负载
- 数据中心东西向流量将占据总流量的77%
  - 多租户共享资源池
  - 分布式系统
  - 数据备份

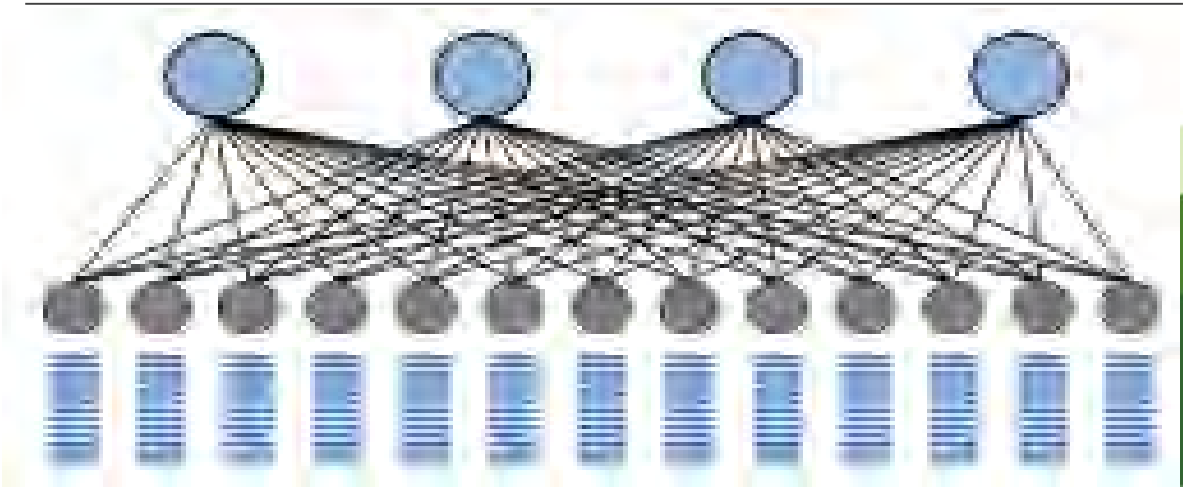


来源：思科云产业调研报告（2015-2020）

由静到动

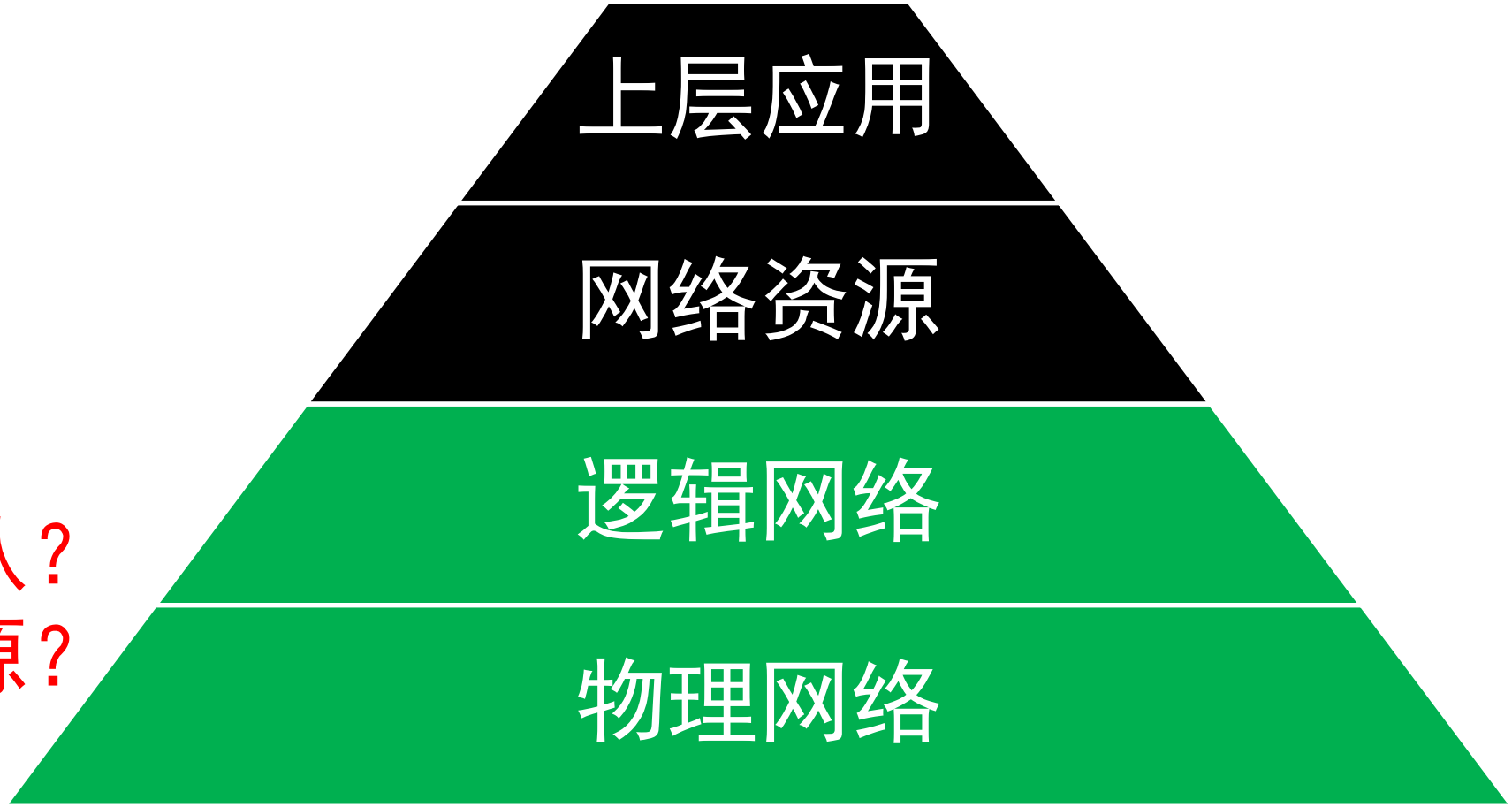
- 业务负载弹性伸缩，网络随之动态配置
- 物理基础网络（Fabric）极少变更
  - 虚拟网络需要随业务秒级变更





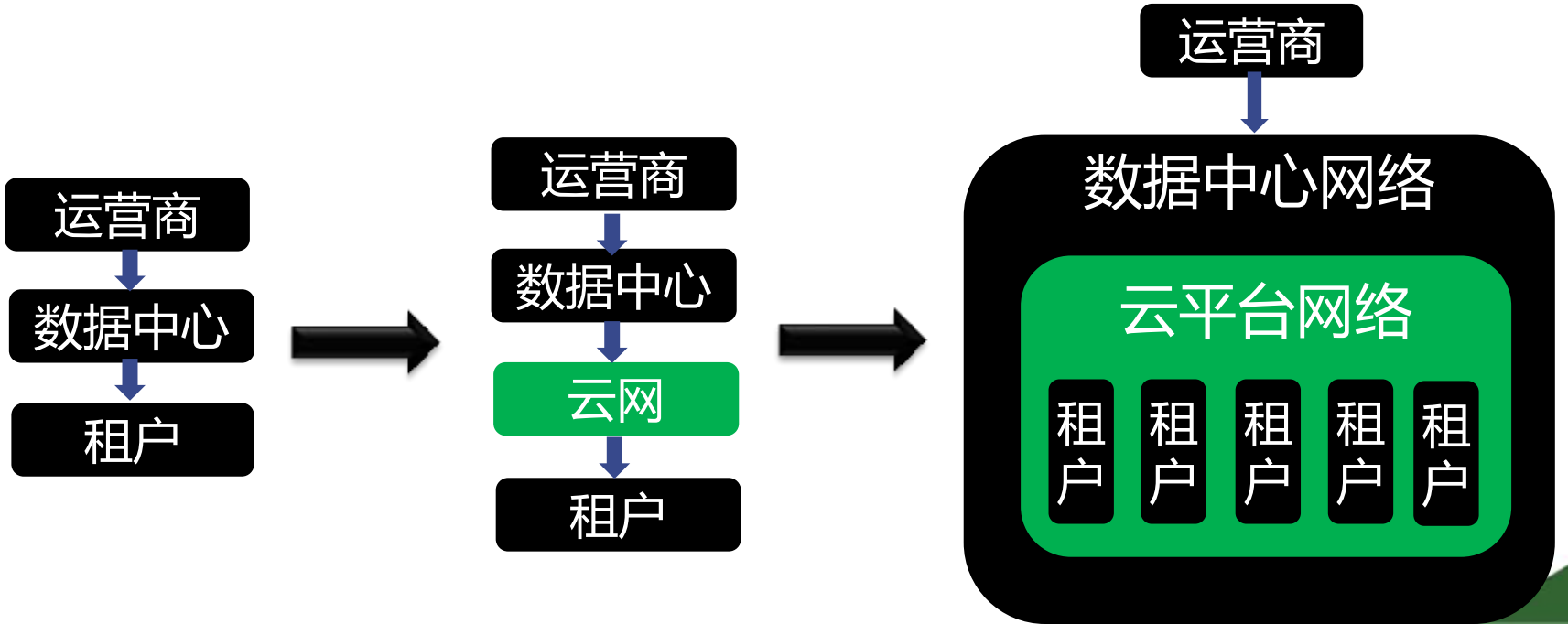
云网运维的痛

物理网络和逻辑网络的边界



两个团队？
两套资源？

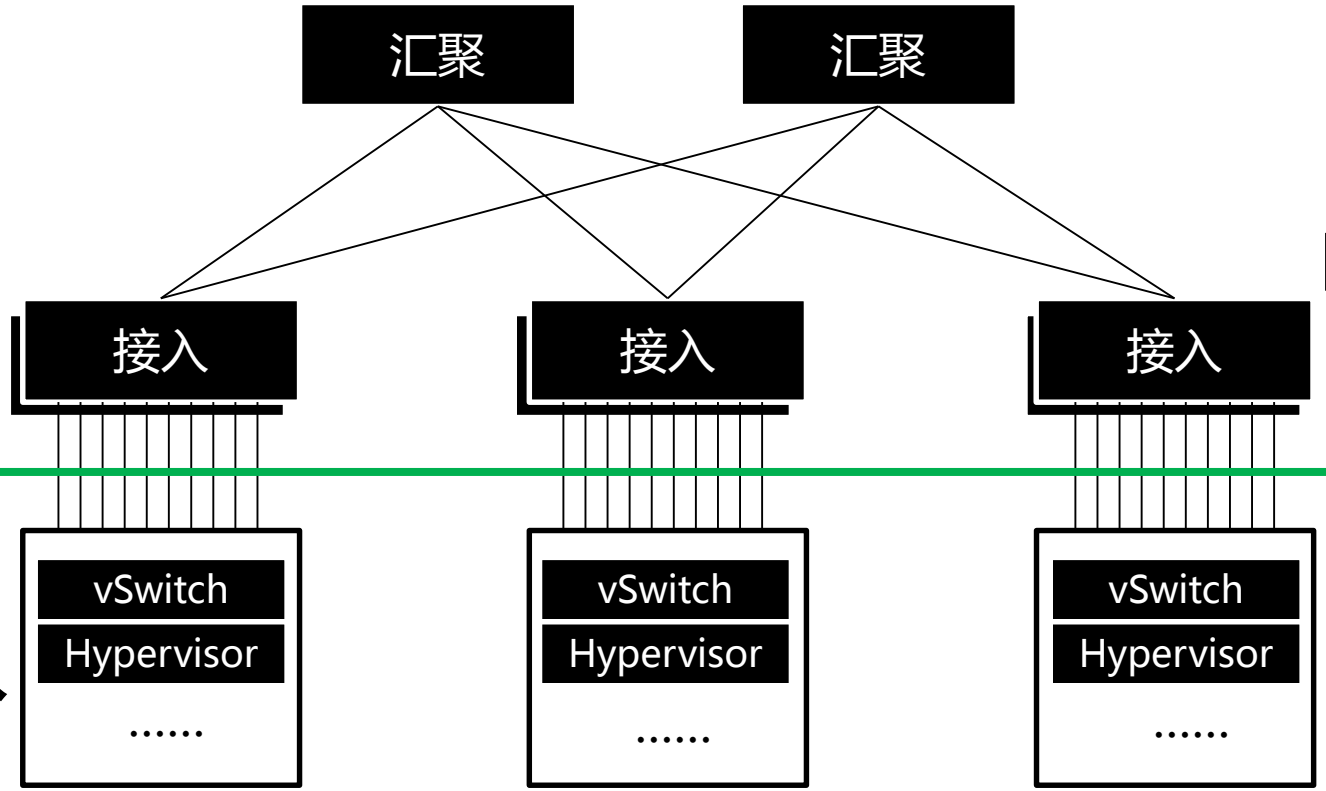
故障总是被用户先发现



服务器和网络团队相关吗

网络团队

服务器团队

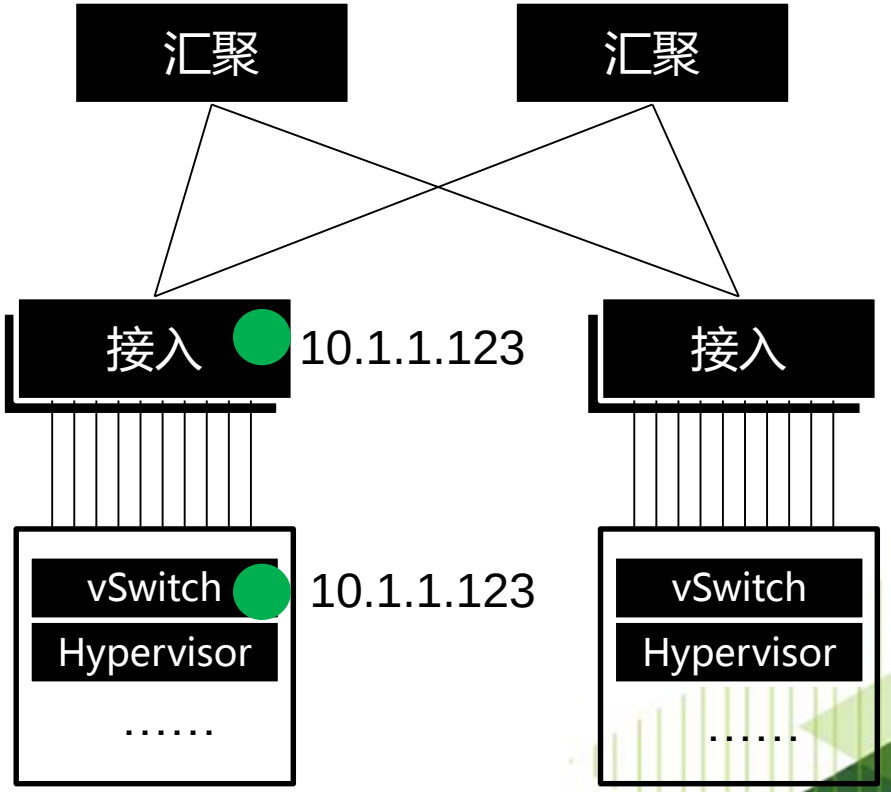


IP地址重叠
隧道封装（UDP 4789）

HTTP / DNS / ICMP / ARP …

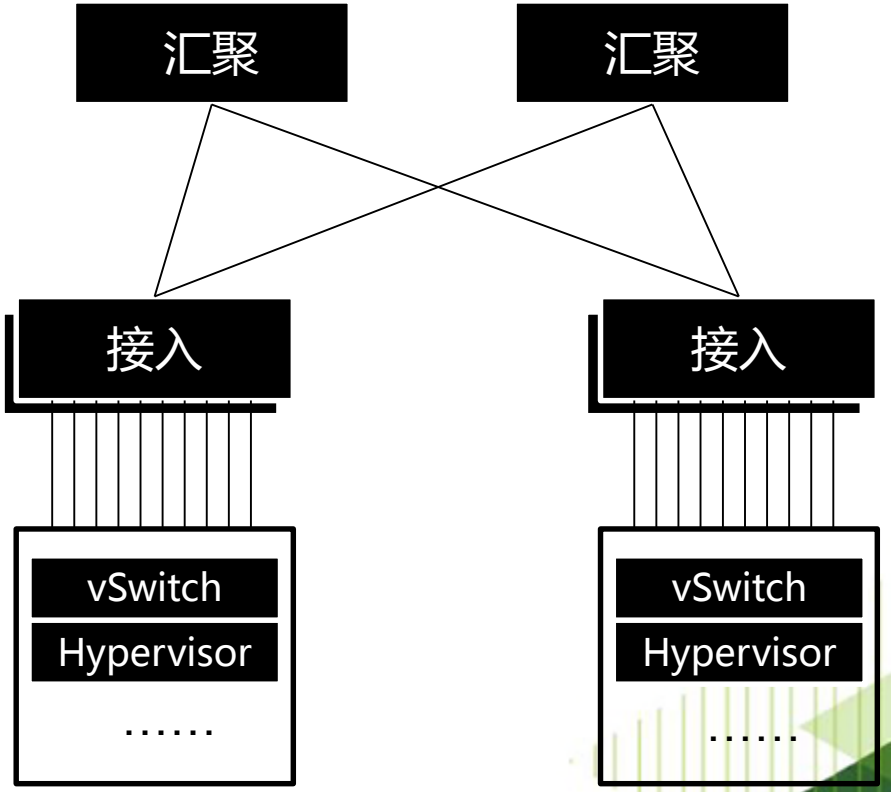
物理网络和逻辑网络的隔离和协作

- 那些Debug以后留下的坑
- 在交换机上创建了一个SVI
  - 影响了租户的虚拟机
- 只在一个MLAG Peer交换机上放行了VLAN
  - 租户网络时通时断



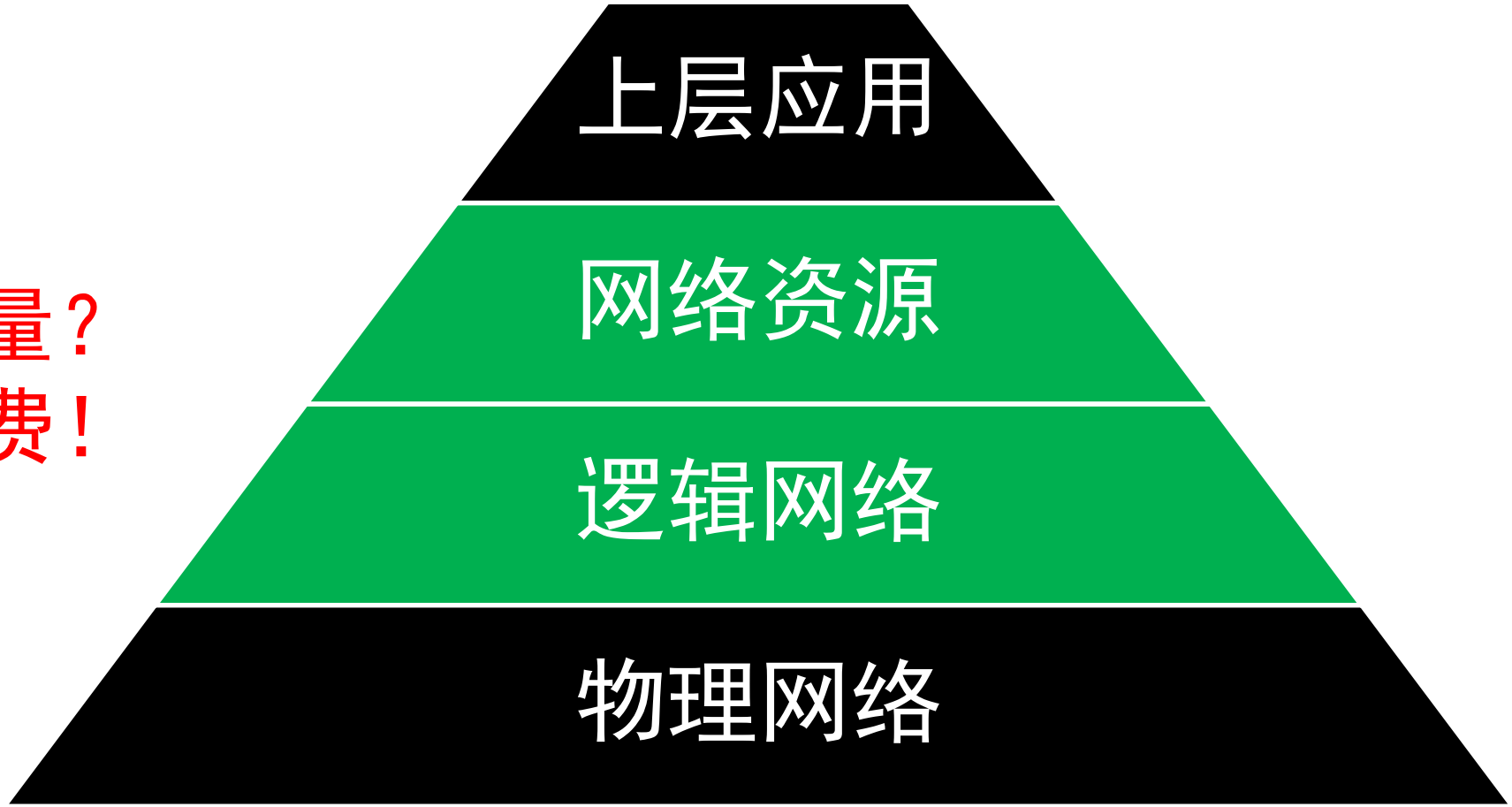
如何设置风暴抑制

- 硬件交换机不怕风暴
  - 硬件交换机擅长转发
  - 设置明确的BUM风暴速率
- 复制是虚拟交换机的灾难
  - 硬件交换机放行的风暴流量能被虚拟交换机放大几十倍



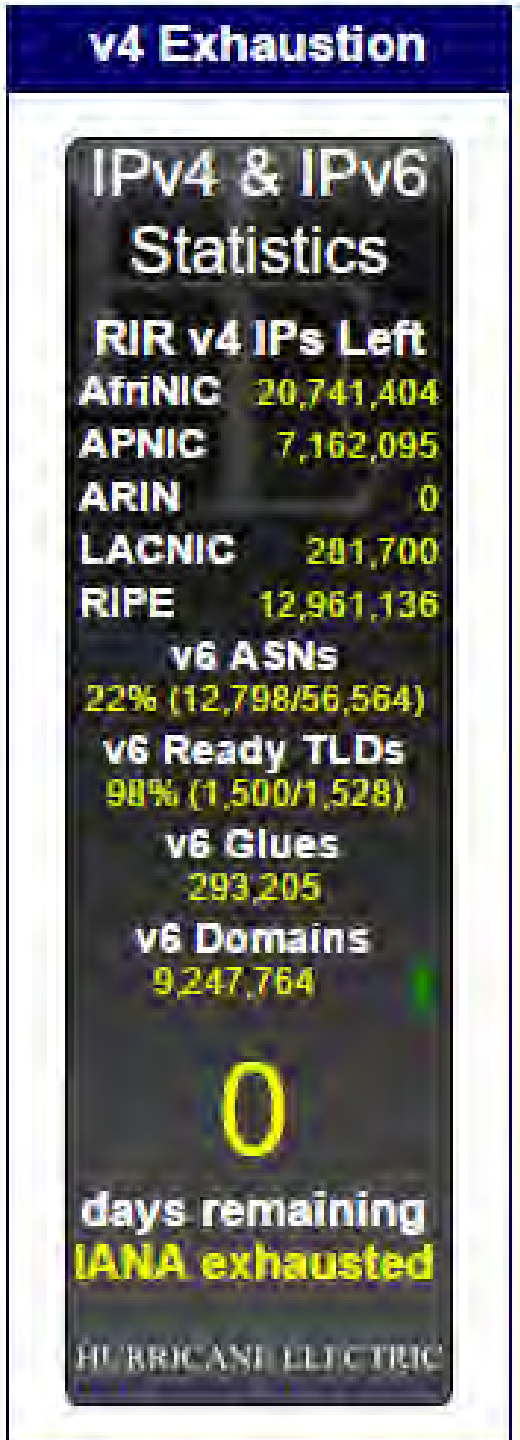
网络资源的监控

没有计量？
就有浪费！



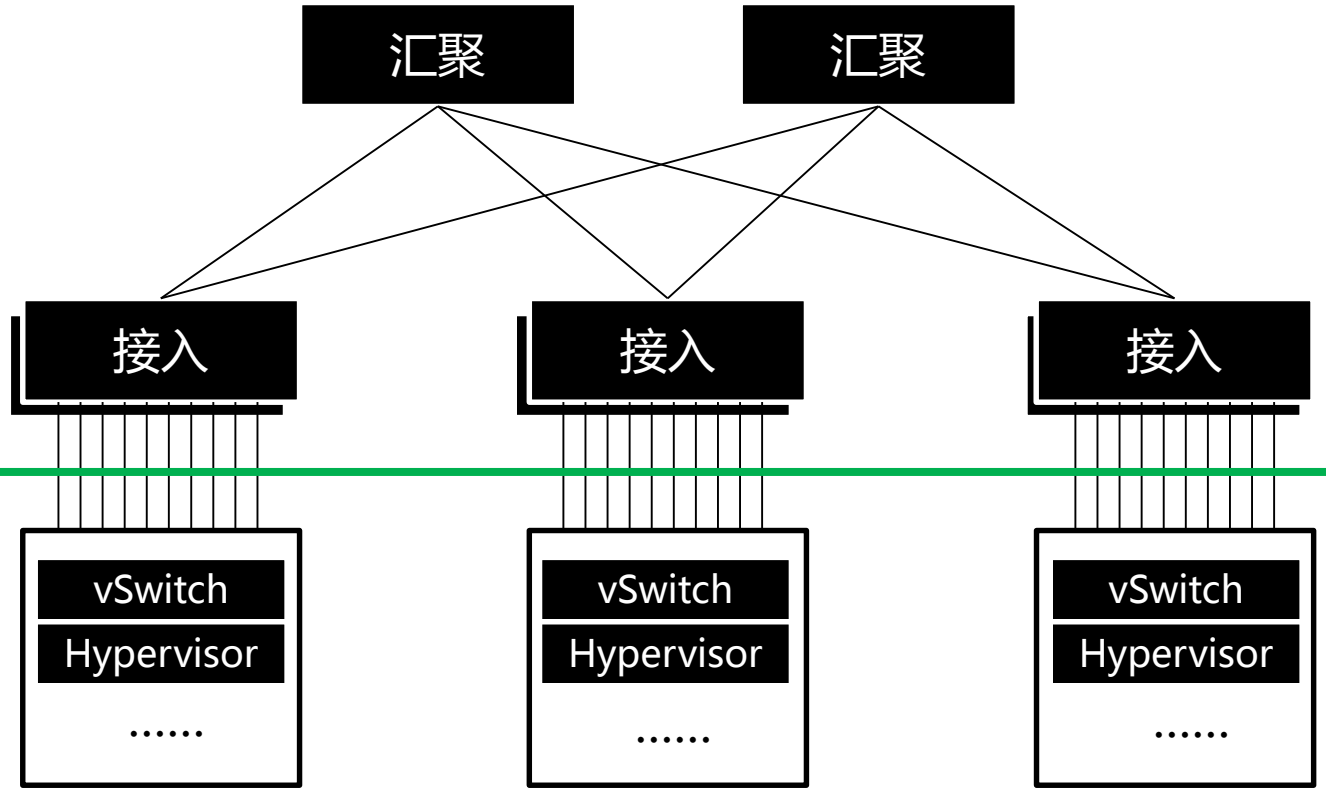
IP地址不够用

- 租户自由申请
  - IP地址自动分配
- 如何分配
  - 网关下沉：地址碎片化
  - 内网地址：体验差
  - /32地址：运维复杂
- 分配 == 使用？
  - 如何发现Dark Host



内网带宽共享

服务质量
如何监控

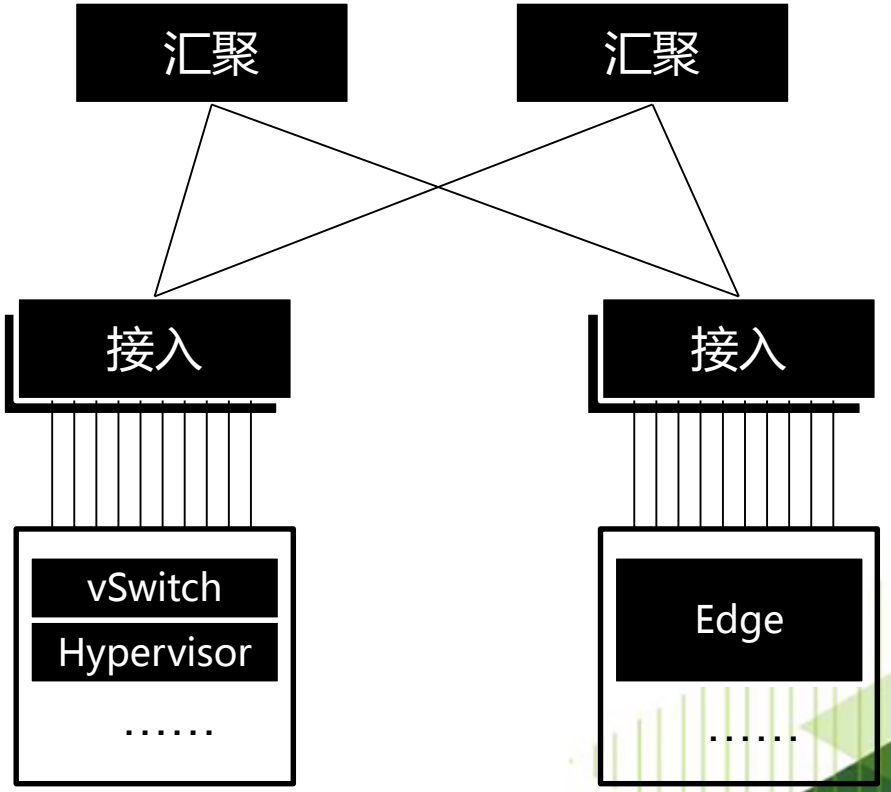


IP地址重叠
隧道封装（UDP 4789）

HTTP / DNS / ICMP / ARP …

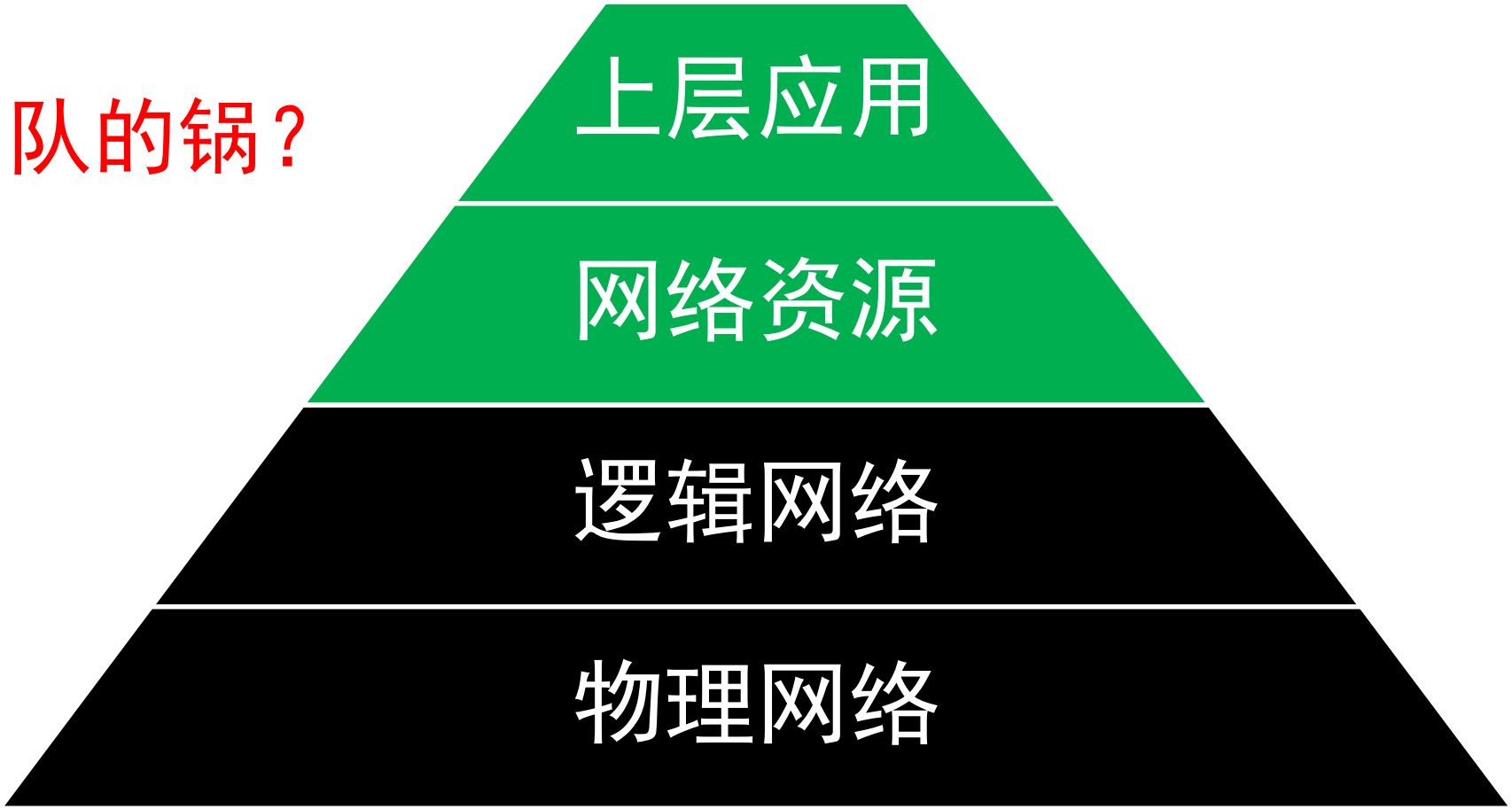
外网带宽计量

- 多维度
  - IP地址
  - L4 Port：租户
  - VLAN：不同运营商线路
  - VXLAN VNI：DCI虚机专线
- 细粒度
  - 95峰
  - 第三峰
  - …
- Netflow?
  - 数据糙
  - 负载高



是应用，还是网络？

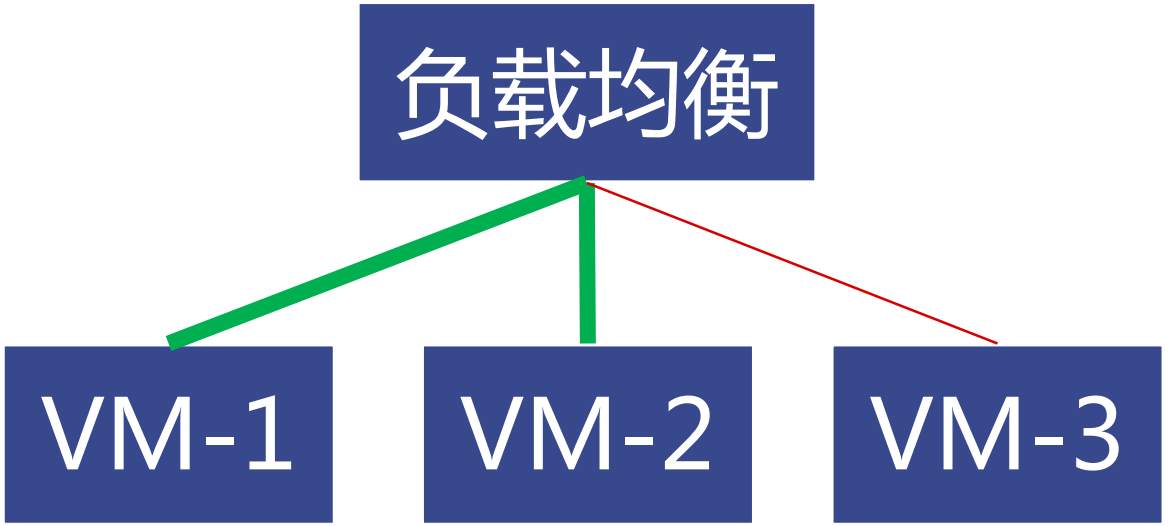
哪个团队的锅？



是应用的问题还是网络的问题

一个抽风的后端主机

- 哪个阶段出了问题
  - TCP三次握手
  - HTTP Response
  - TCP Reset
- 其他配置是否正常
  - VIP/RIP、MAC地址是否正常
  - DNS配置是否正常



如果历史可以重来

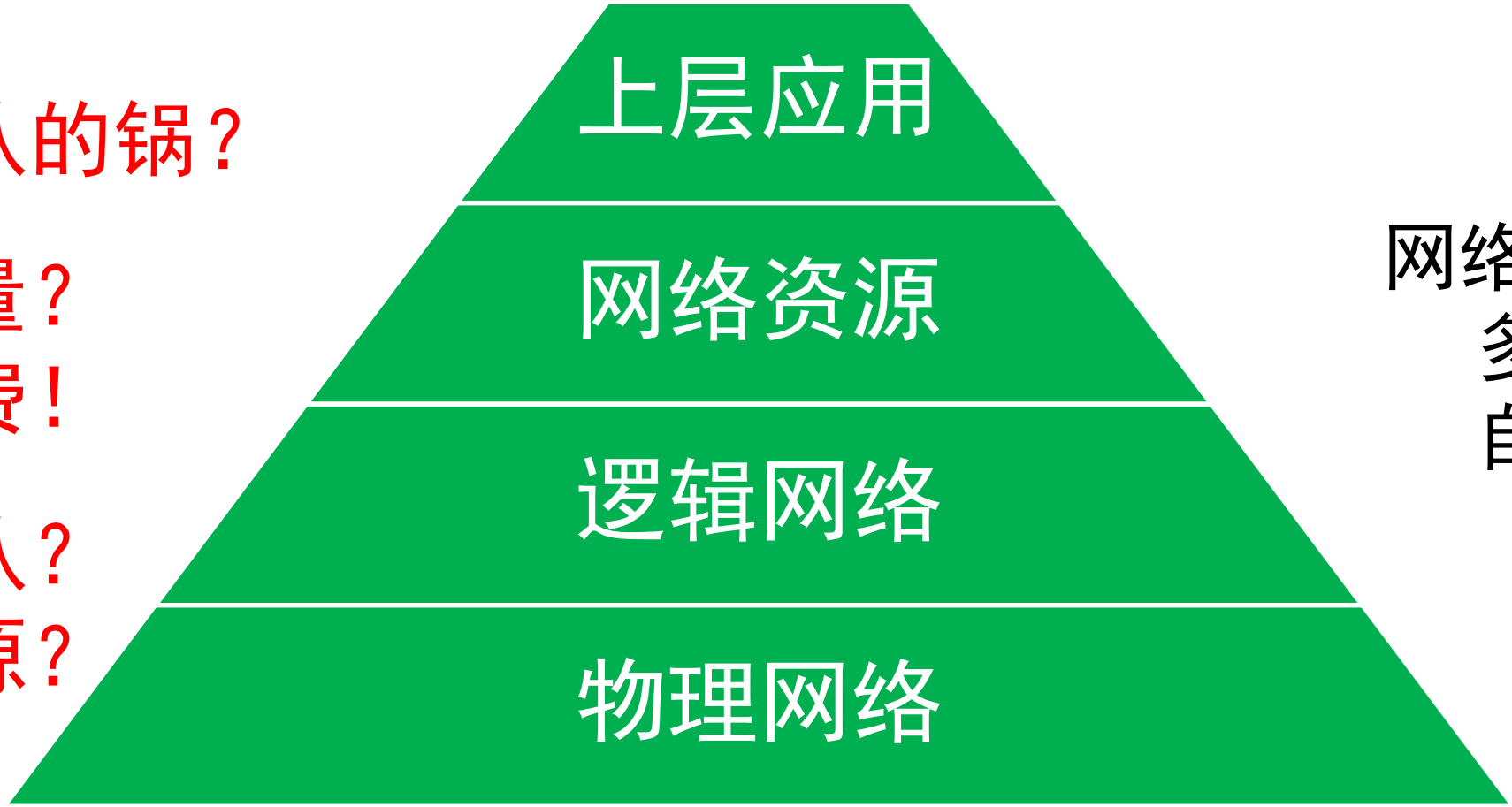
- 排查网络故障的时效性
  - 那些莫名其妙恢复的故障
- 历史证据
  - 昨天？上周？上个月

云网运维的痛

哪个团队的锅？

没有计量？
就有浪费！

两个团队？
两套资源？



网络Overlay
多租户
自服务

如何解决这些问题

网络运维内外交困

老问题灭不完

应用问题 or 网络问题
内部网络问题 or 运营商网络问题

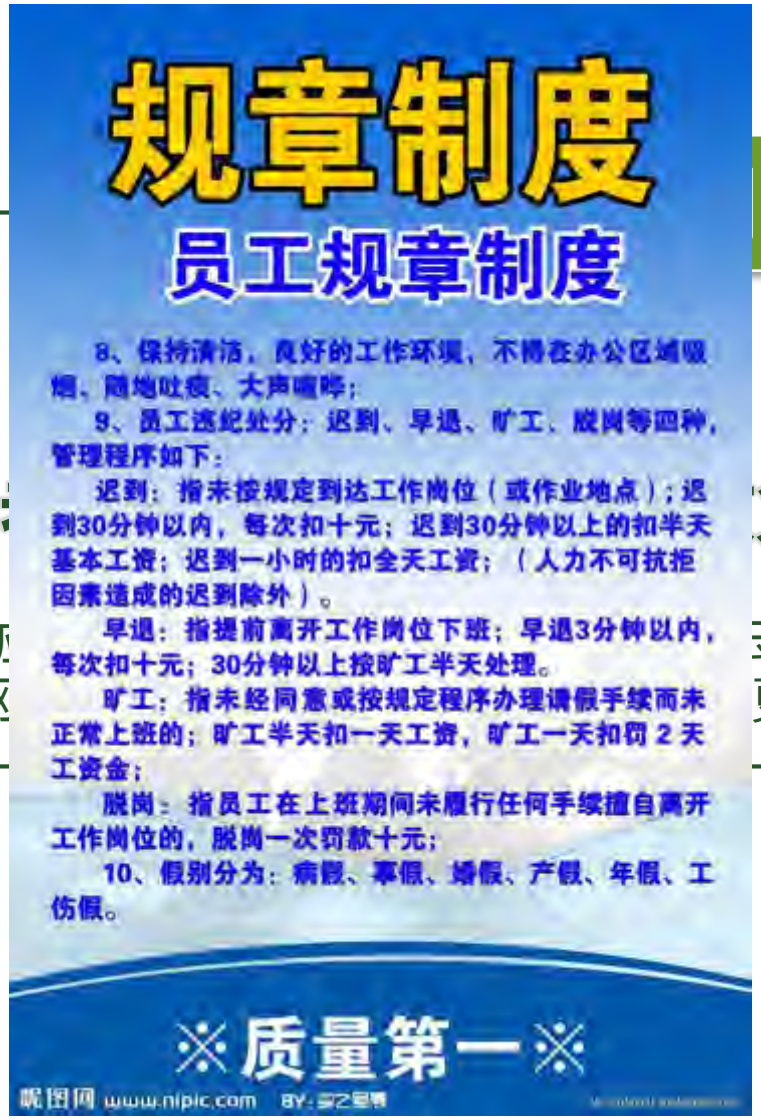
救火经验和本领失效

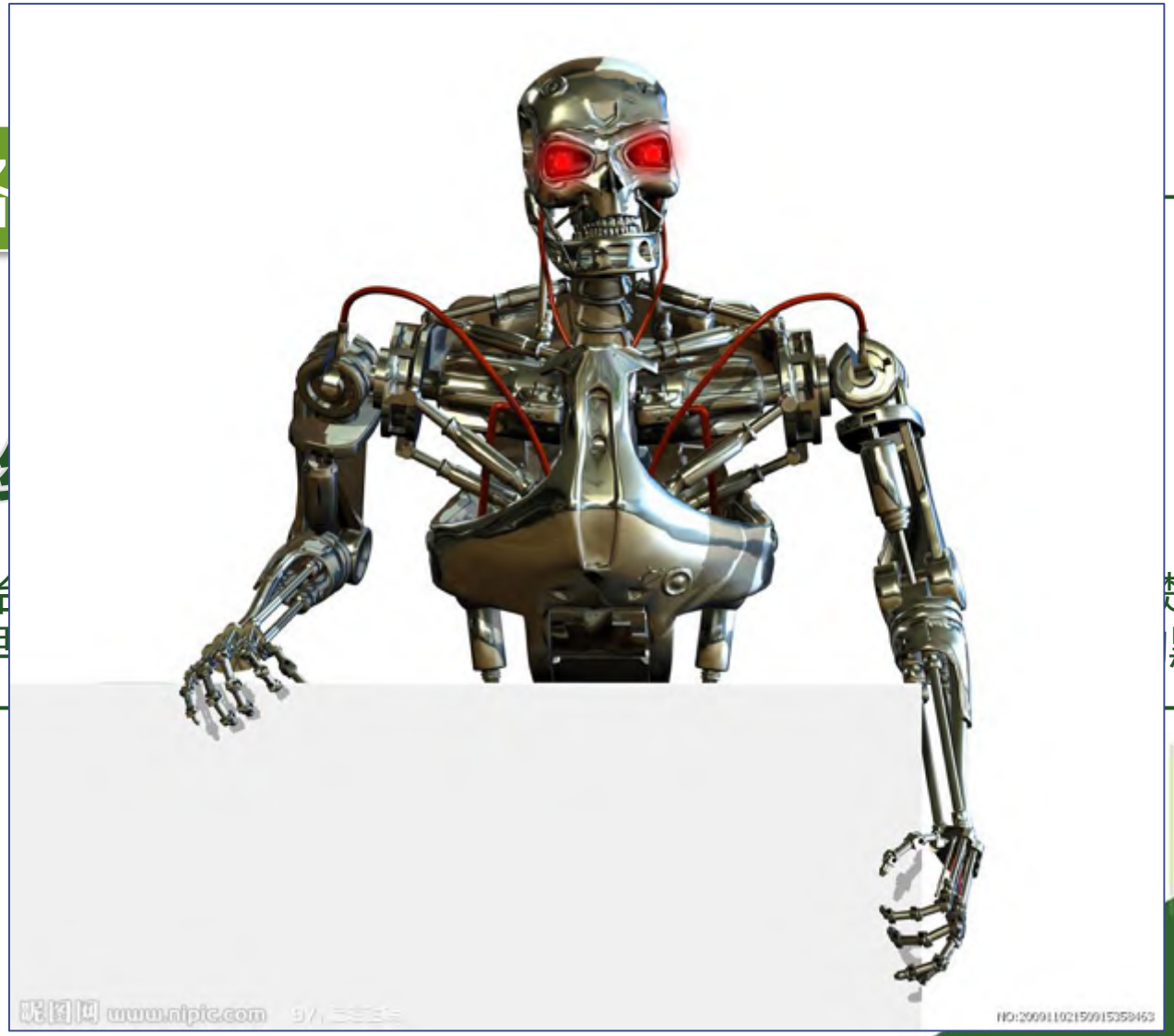
云平台无报警，但用户报障
更早的网络故障能追溯吗

新问题层出不穷

全是UDP4789，傻傻分不清楚
虚拟网络故障怕不怕，看不见的黑洞

如何解决这些问题


规章制度
员工规章制度
8、保持清洁，良好的工作环境，不得在办公区域吸烟、随地吐痰、大声喧哗；
9、员工违纪处分：迟到、早退、旷工、脱岗等四种，管理程序如下：
迟到：指未按规定到达工作岗位（或作业地点）；迟到30分钟以内，每次扣十元；迟到30分钟以上的扣半天基本工资；迟到一小时的扣全天工资；（人力不可抗拒因素造成的迟到除外）。
早退：指提前离开工作岗位下班；早退3分钟以内，每次扣十元；30分钟以上按旷工半天处理。
旷工：指未经同意或按规定程序办理请假手续而未正常上班的；旷工半天扣一天工资，旷工一天扣罚2天工资金；
脱岗：指员工在上班期间未履行任何手续擅自离开工作岗位的，脱岗一次罚款十元；
10、假别分为：病假、事假、婚假、产假、年假、工伤假。
※质量第一※
昵图网 www.nipic.com


昵图网 www.nipic.com

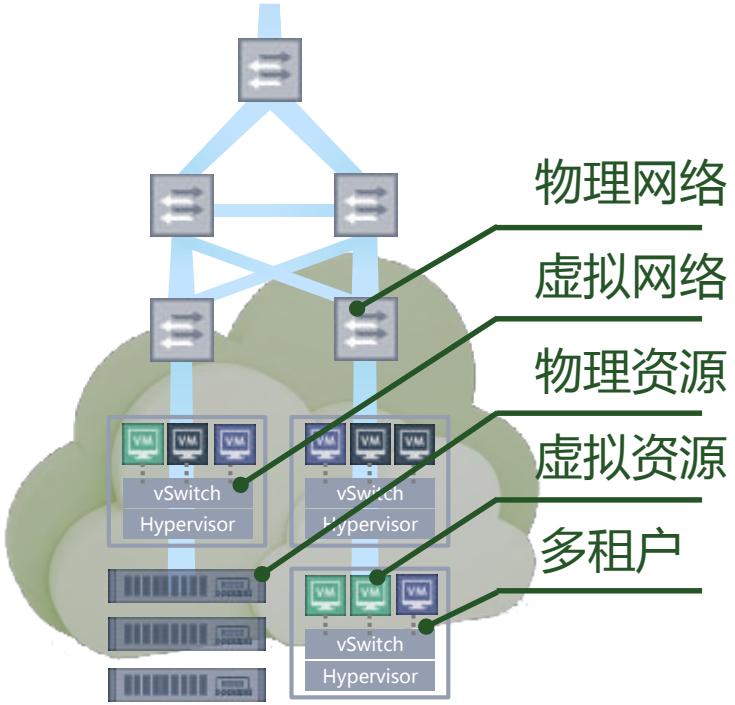
打造云网大数据平台
摆脱云网运维困境

技术上的挑战

- 数据采集
  - 虚拟交换机：与服务器/云平台运维团队协作
  - 全网全时：全量采集的能力 + 细粒度策略控制
- 数据分析
  - 海量流量：高性能协议栈、分布式处理
  - 多租户网络：VLAN/VXLAN识别，流量和资源/租户进行映射
  - 数据高效存储：聚合、压缩
- 数据展现
  - 数据关联，多维度展现
  - 实时报警：发现毫秒级毛刺、与业务关联

大范围和细粒度数据采集

多维度采集



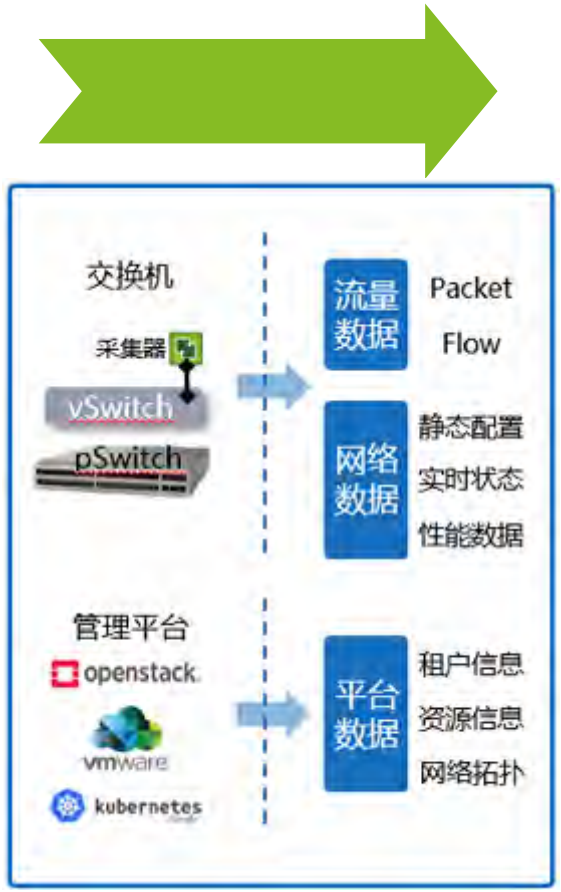
物理网络

虚拟网络

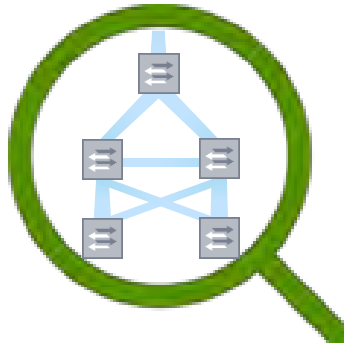
物理资源

虚拟资源

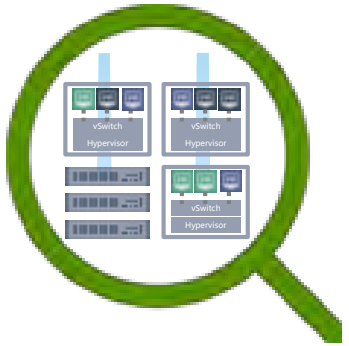
多租户



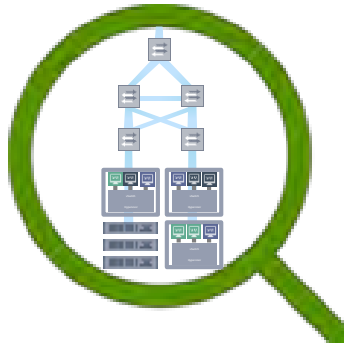
物理网络



虚拟网络

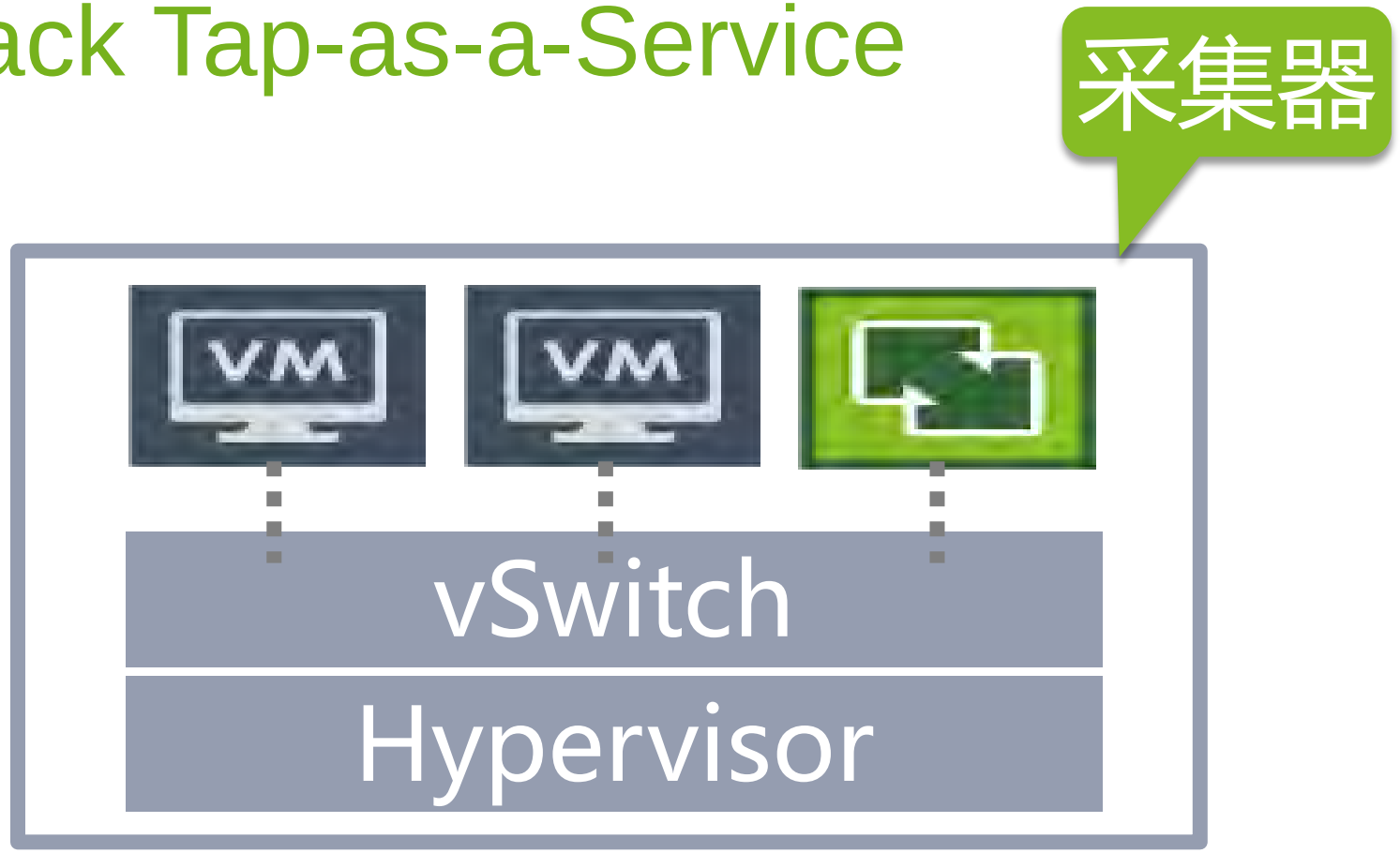


全网混合网络



采集虚拟网络

- OpenStack Tap-as-a-Service



控制投入

- 生产网络和监控网络的投入比例
  - 包处理的瓶颈：Linux内核 1.488 Mpps
  - 数据存储的瓶颈：CPU/内存/硬盘的速度

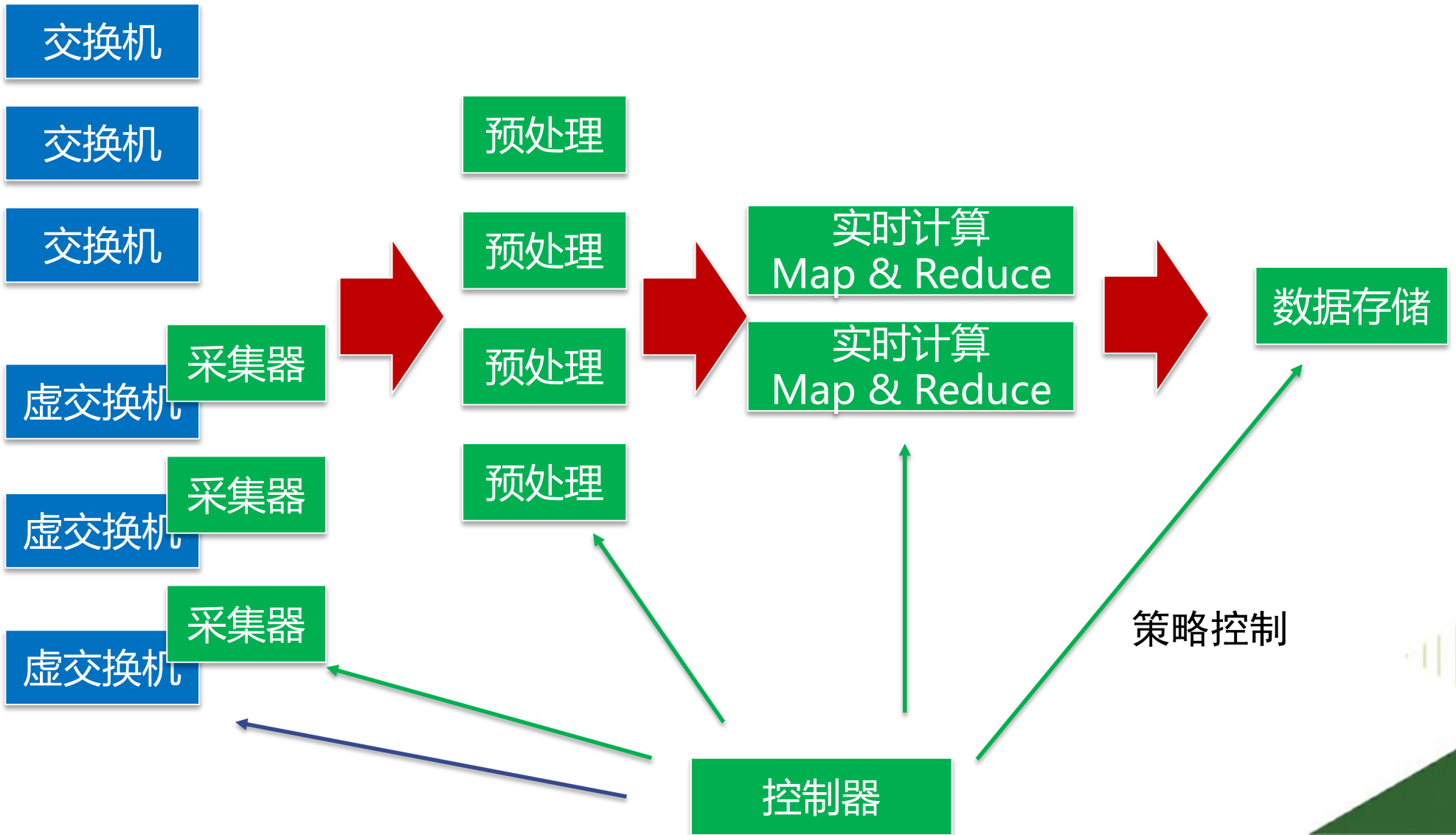
- 全量采集的能力
  - 采集点全覆盖
- 细粒度策略控制
  - 物理交换机：ACL过滤流量
  - 虚拟交换机：OpenFlow过滤流量

高性能
数据分析

发挥最大的能力

- 不同的流量，需要不同的处理
  - 任意流量：基于Flow元组的分析
  - TCP流量：基于四层包头的分析
  - HTTP流量：基于HTTP头的分析
  - 私有协议流量：基于前N个字节的分析
  - 重要流量：全量记录，用于审核

策略控制


交换机
交换机
交换机
虚交换机
虚交换机
虚交换机
采集器
采集器
采集器
预处理
预处理
预处理
预处理
实时计算
Map & Reduce
实时计算
Map & Reduce
数据存储
策略控制
控制器

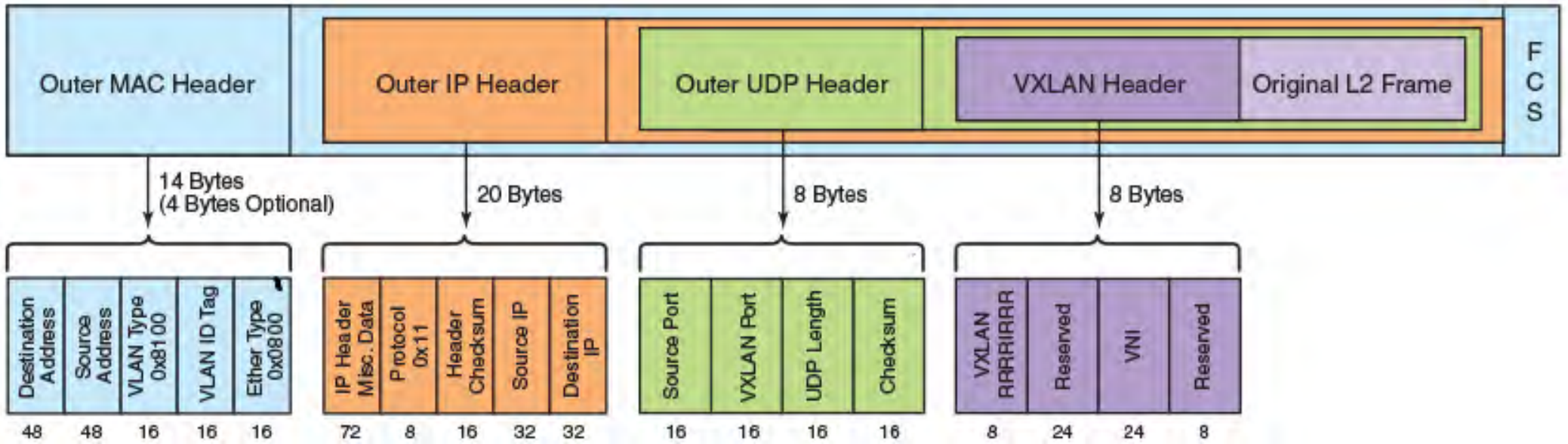
高性能协议栈：网络层

- 聚合Packet为Flow元组
- 保留所有4层包头

- 处理性能
  - Linux内核：1.488 Mpps
  - DPDK 14.88 Mpps 起
- 存储性能
  - 包头差量压缩，90%压缩率
  - 一个惊人的事实：包头存储空间 ~ Flow元组

识别多租户网络

- VLAN - 租户映射关系
- VXLAN VNI - 租户映射关系：网包重新解析



高性能协议栈：应用层

- 常见应用的元数据提取
  - HTTP头：非侵入式Web Server日志
  - DNS请求日志

多维度
数据展现

网络配置的Git log

- 记录配置历史
- 追溯人为的失误

```
[root@NSP_Controller_Server ~]# mt history.switch mip=172.17.1.102
--- 35b4895e733d1c004eb961c5dd609c35 SEQ-28 2016-12-28 13:44:21
+++ 95dc542ab31049a6baf8ba12a39b0aaa SEQ-29 2016-12-29 15:00:27
@@ -102,10 +102,10 @@
   policer-aggregate ingress_qos_eth-0-2
 !
 interface eth-0-1
- description VIF933_IFINDEX1_Nets
  jumboframe enable
  ip arp inspection trust
  switchport access allowed vlan add 1024,1025
+ shutdown
  lldp enable txrx
 !
 interface eth-0-2

--- 0a103cbd52092c53e5a110b8607d0702 SEQ-27 2016-12-27 23:43:07
+++ 35b4895e733d1c004eb961c5dd609c35 SEQ-28 2016-12-28 13:44:21
@@ -105,7 +105,7 @@
  description VIF933_IFINDEX1_Nets
  jumboframe enable
  ip arp inspection trust
- switchport access allowed vlan add 1024,4090
+ switchport access allowed vlan add 1024,1025
  lldp enable txrx
 !
 interface eth-0-2
```

MAC震荡和ARP欺骗

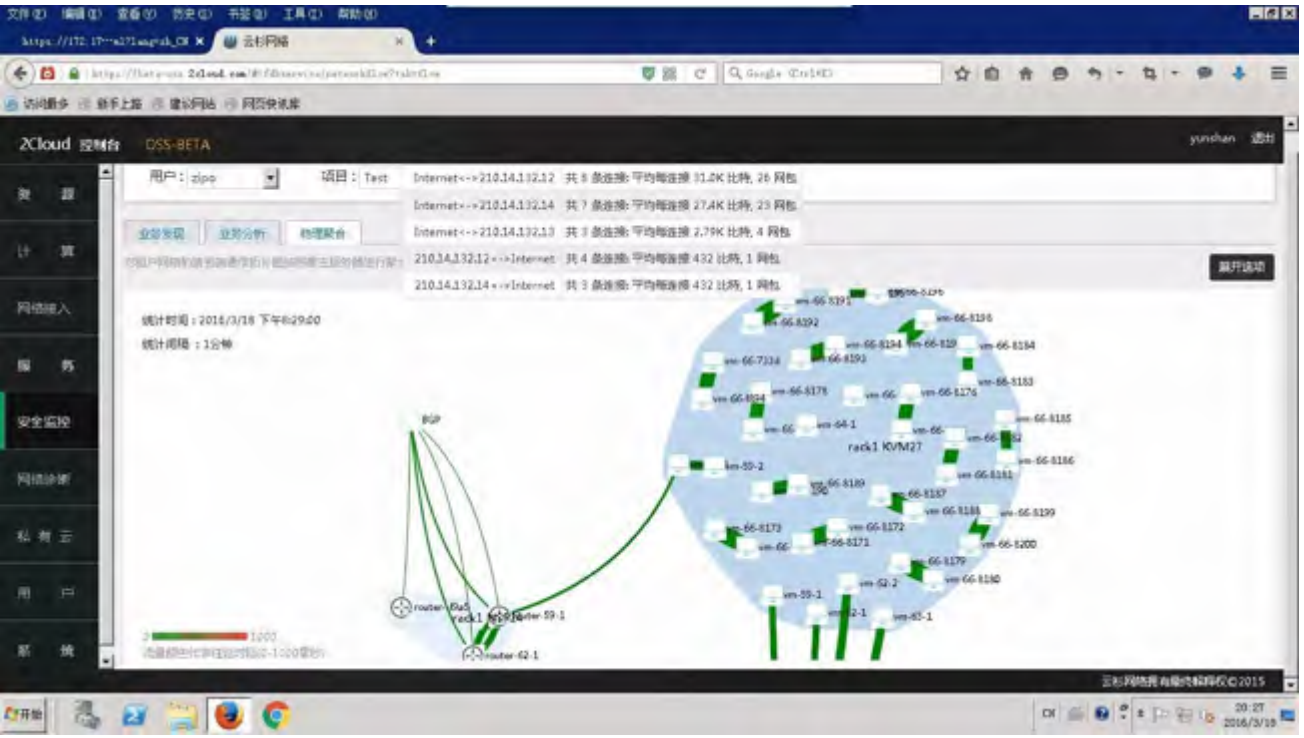
- 覆盖物理、虚拟网络
  - ARP流量分析
  - 流量中的MAC入口分析

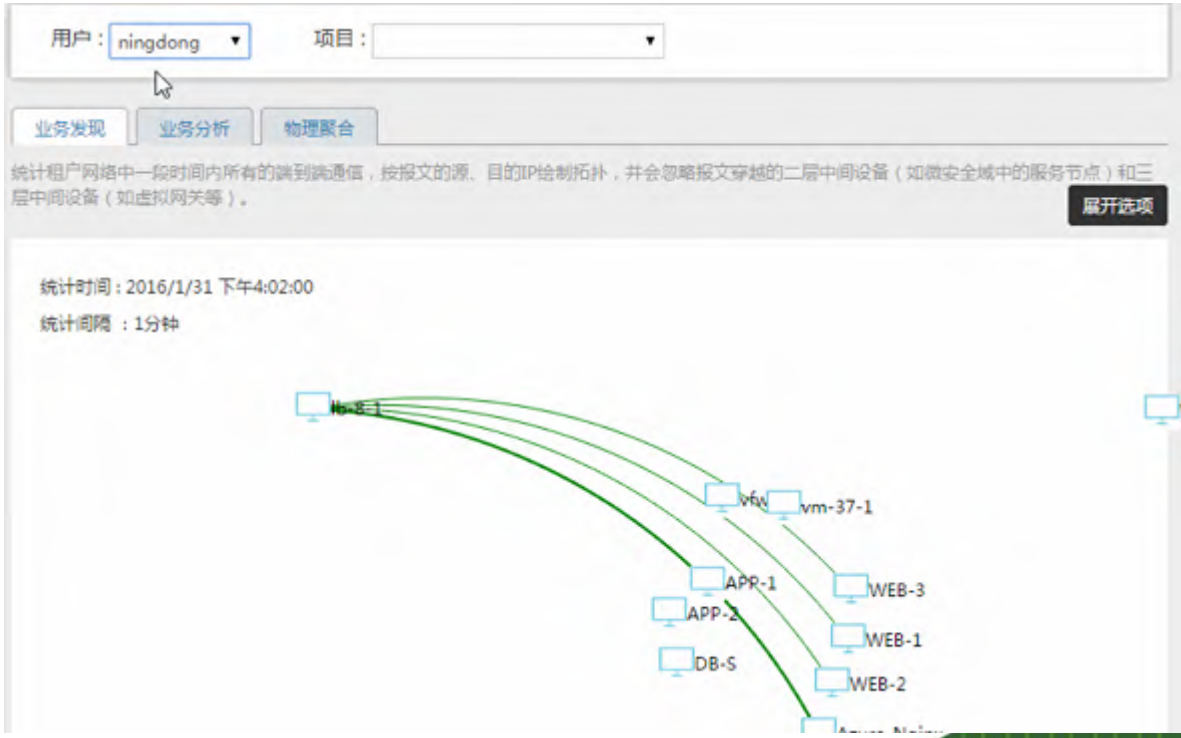
抽取服务日志

- Telnet/SSH/RDP日志：发现暴力破解
- DHCP/DNS：发现集群配置差异

是网络的问题还是应用的问题

- 租户流量拓扑





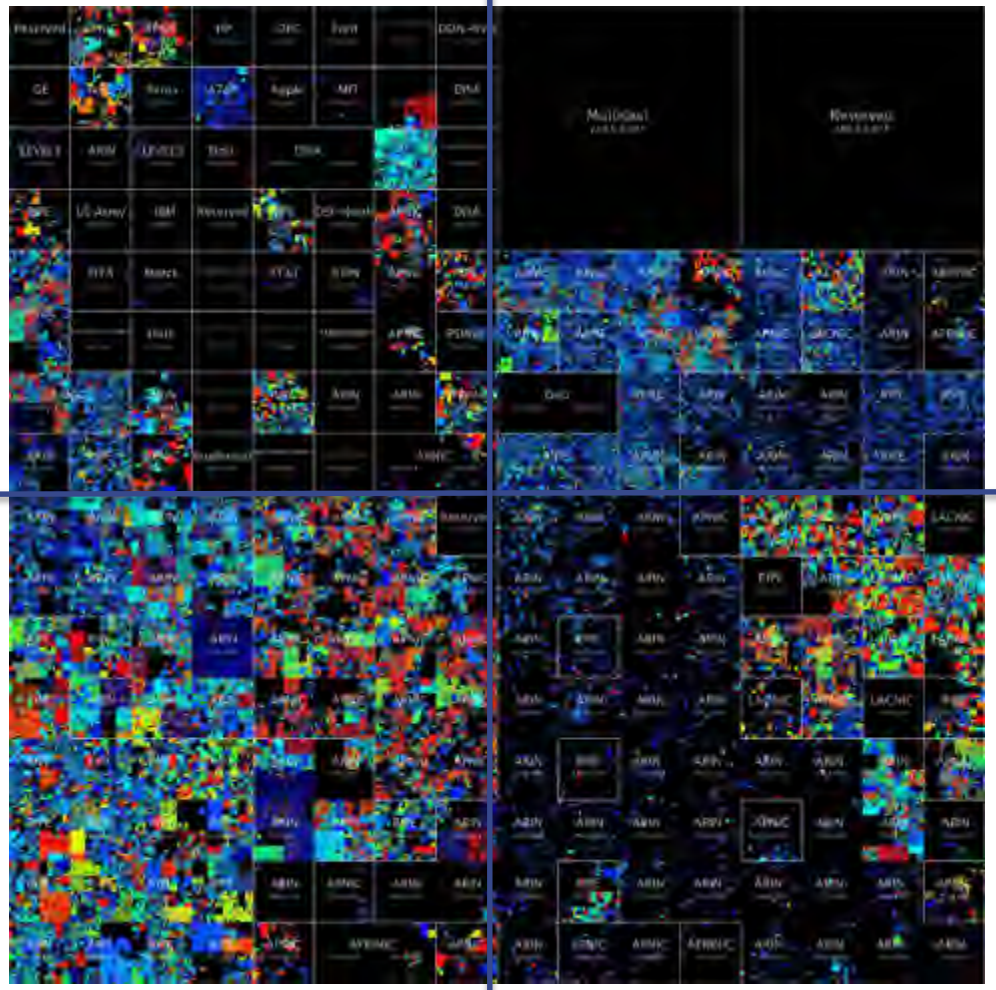
细粒度计量

- 自定义计量的聚合字段
  - IP地址、VLAN、端口号、…
- 基于Apache Storm的实时流式计算

到底哪些IP在用

0.0.0.0/2

64.0.0.0/2

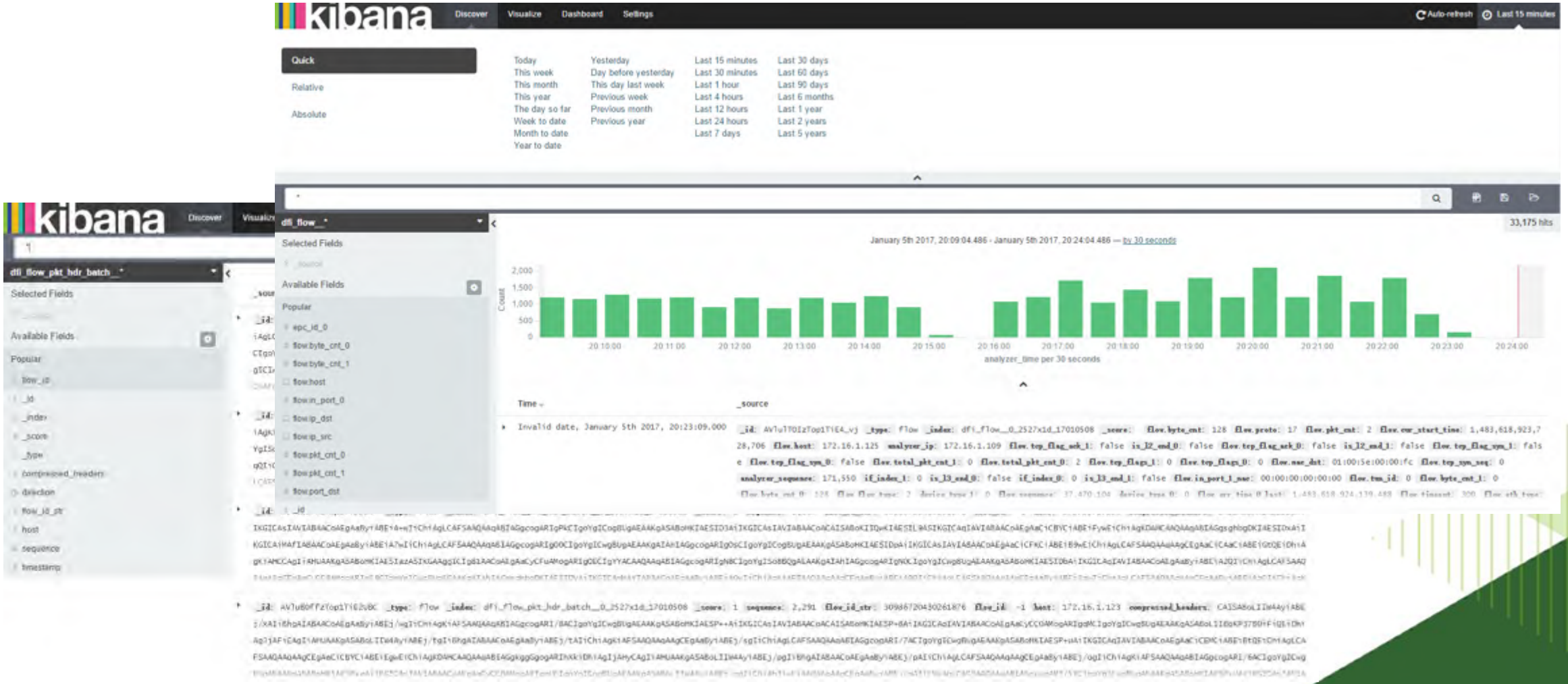


192.0.0.0/2

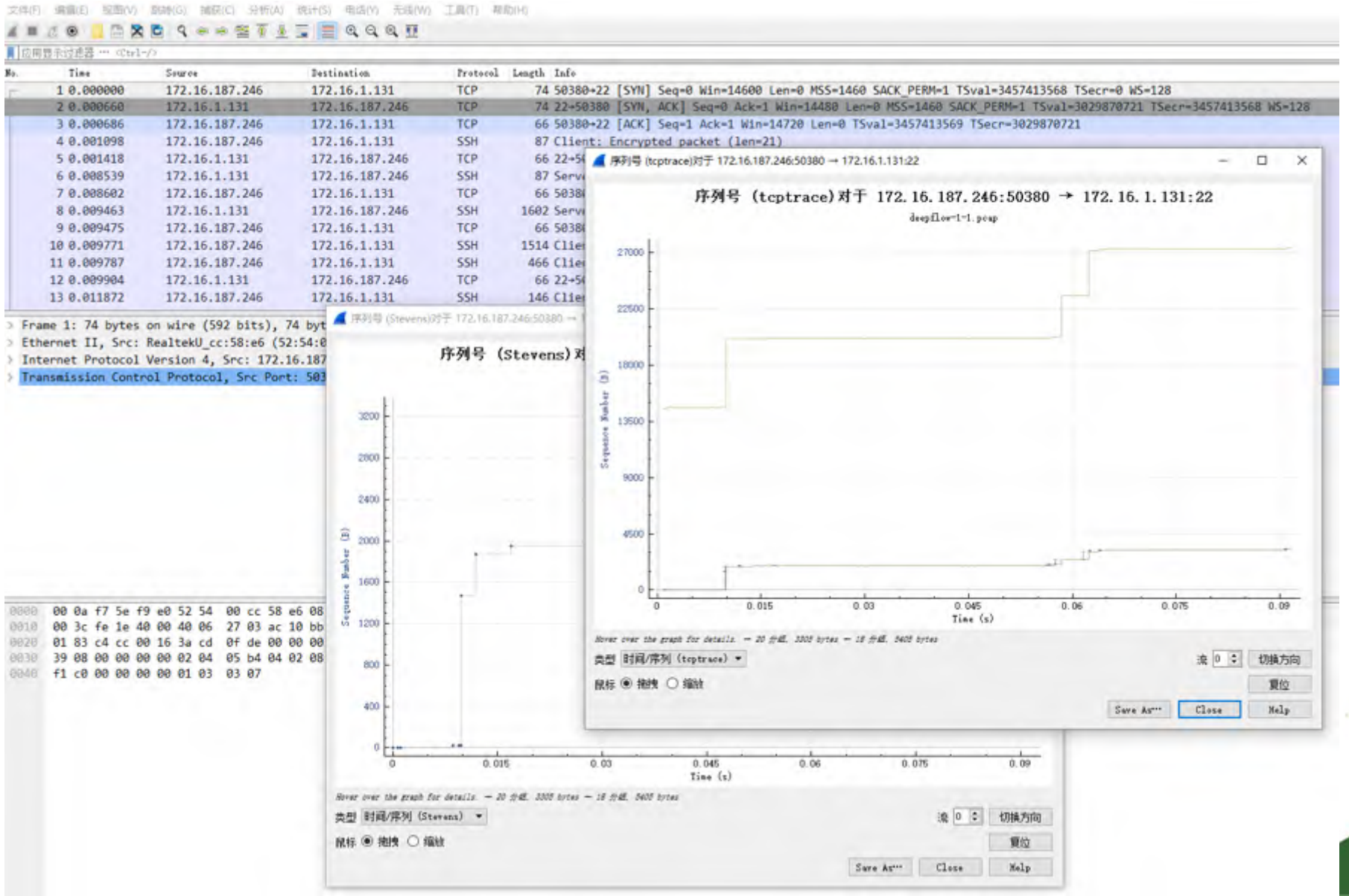
128.0.0.0/2

数据存储：全网全时Flow检索

- ElasticSearch & Kibana



数据存储：用包头还原历史会话



我们的技术栈

展现 Vue.js D3 ECHARTS OpenResty

数据存储 elastic APACHE HBASE InfluxDB

数据分析 JStorm Alibaba Spark

运维 SALTSTACK

数据传输 kafka A distributed streaming platform ØMQ

策略控制 OPENDAYLIGHT

采集/预处理 Telegraf DPDK DATA PLANE DEVELOPMENT KIT .io

数据源 KVM OvS Open vSwitch openstack kubernetes Google

期待你的加入

云杉网络
Yunshan Networks

技术创造价值

www.yunshan.net.cn
info@yunshan.net.cn