

# 大规模云原生应用网络问题定位与排查实践

陈定斌

网易资深云计算解决方案架构师

# 极客邦科技 会议推荐2019

5月

**QCon** 北京

全球软件开发大会

大会: 5月6-8日  
培训: 5月9-10日

**QCon** 广州

全球软件开发大会

培训: 5月25-26日  
大会: 5月27-28日

6月

**GTLC**  
GLOBAL  
TECH LEADERSHIP  
CONFERENCE

上海

技术领导力峰会

时间: 6月14-15日

**GMTC** 北京

全球大前端技术大会

大会: 6月20-21日  
培训: 6月22-23日

**ArchSummit** 深圳

全球架构师峰会

大会: 7月12-13日  
培训: 7月14-15日

7月

**QCon** 上海

全球软件开发大会

大会: 10月17-19日  
培训: 10月20-21日

10月

**ArchSummit** 北京

全球架构师峰会

大会: 12月6-7日  
培训: 12月8-9日

11月

**GMTC** 深圳

全球大前端技术大会

大会: 11月8-9日  
培训: 11月10-11日

**AiCon** 北京

全球人工智能与机器学习大会

大会: 11月21-22日  
培训: 11月23-24日

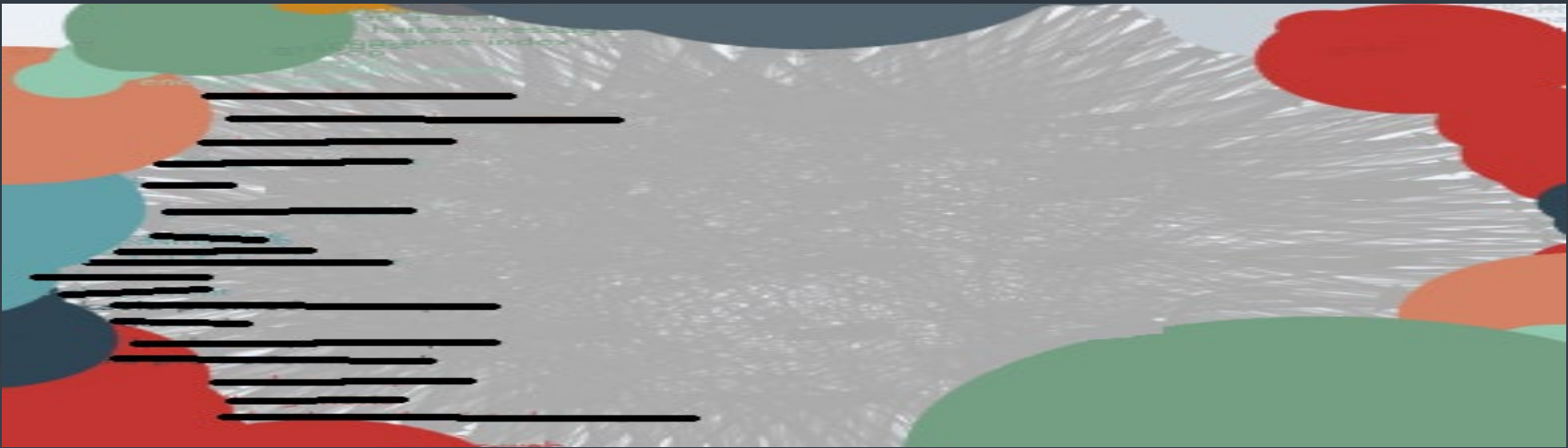
12月

# 自我介绍

网易云资深云计算解决方案架构师，主要负责内部考拉业务上云解决方案。参与考拉整个上云过程中的架构方案、需求分析以及问题定位，帮助考拉顺利进行云上架构演进并且稳定度过双十一大促。在支持帮助大规模应用上云的过程中，积累了丰富的云上问题解决经验。

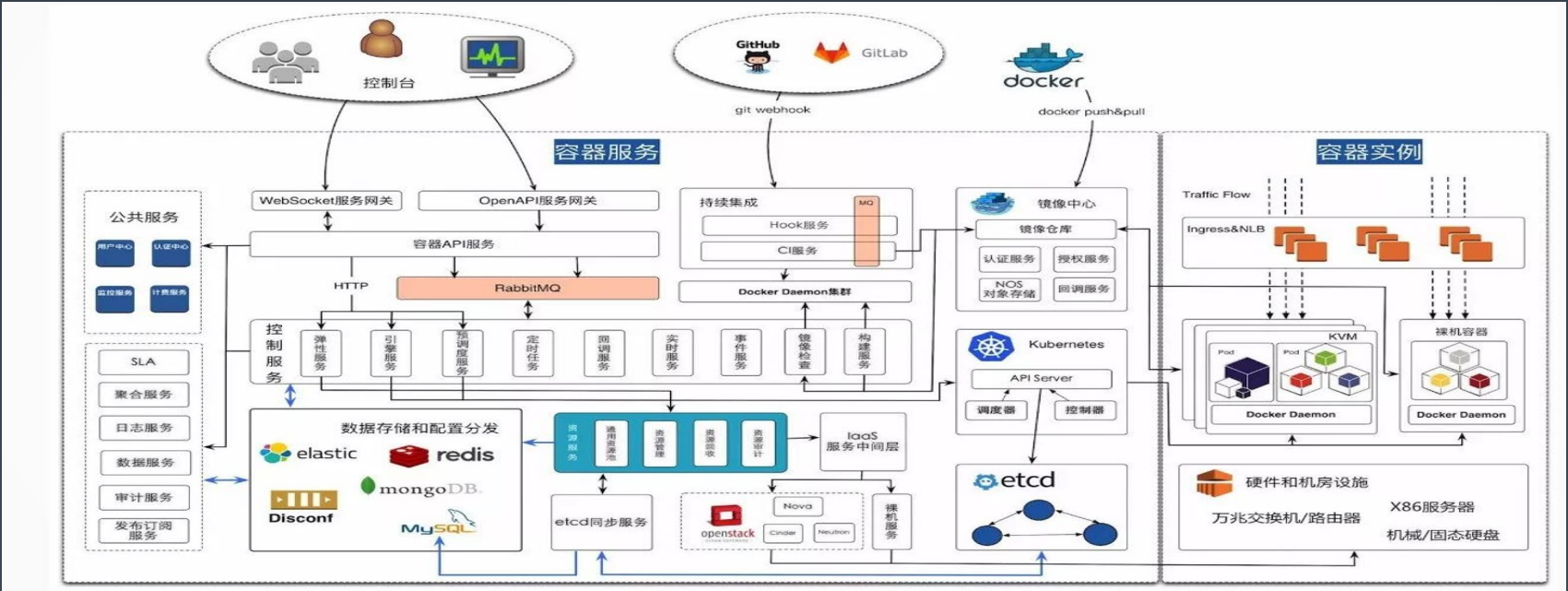
# 服务分层

应用层



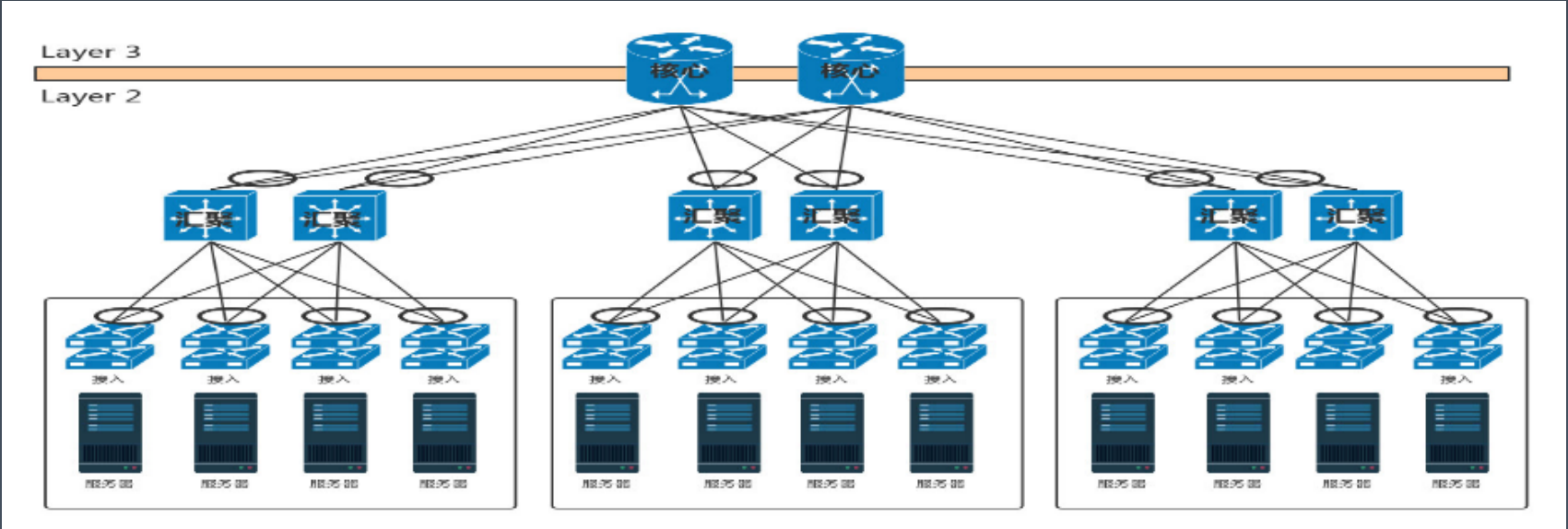
应用服务调用关系

虚拟网络



云计算服务

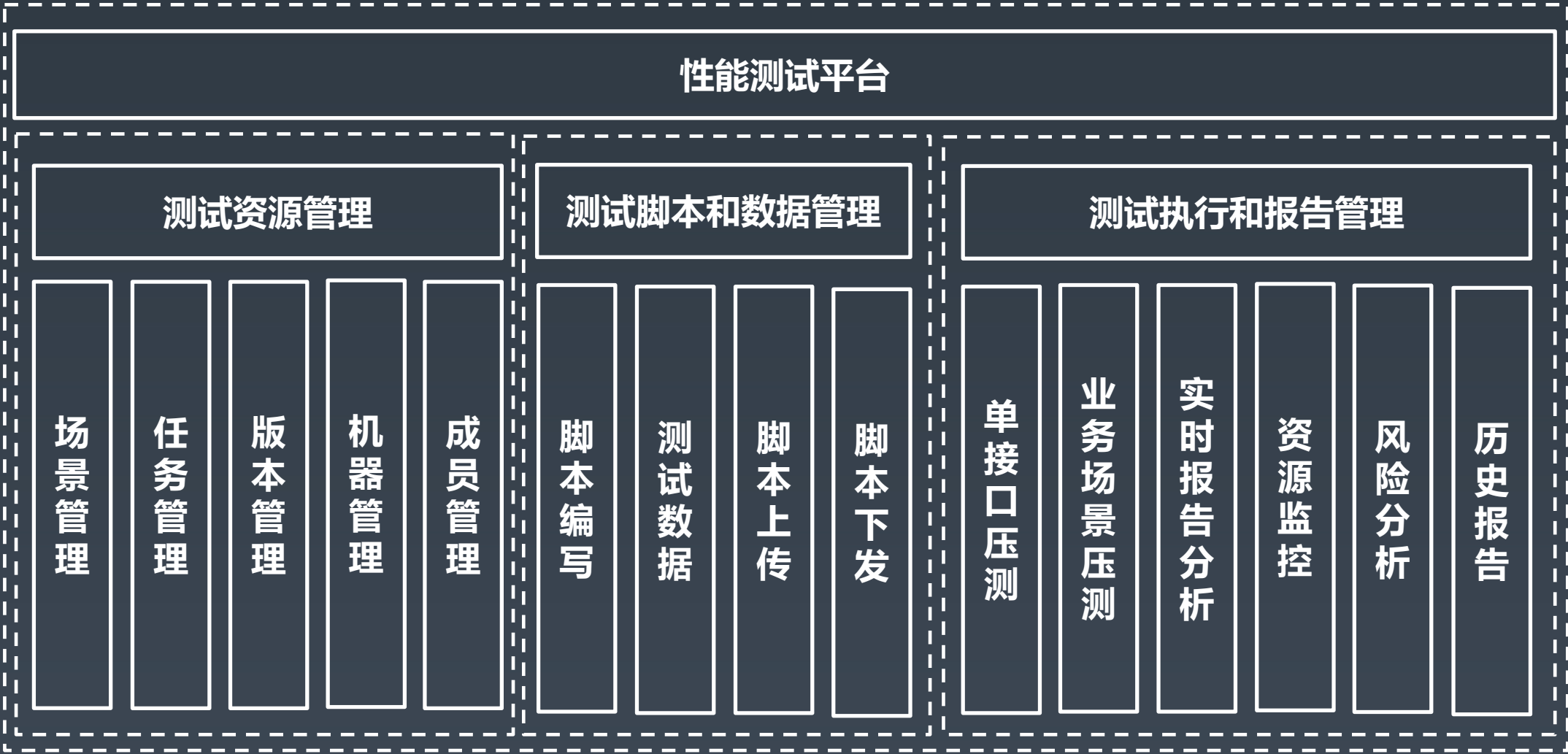
物理网络



多机房，多可用区



# 服务压力测试



## 容量测试

• 采用梯度压力，看服务的性能变化情况，评估出服务的最大容量值

## 摸高压测

• 在达到停止条件之后，继续增加压力，检验服务集群在失效状态下的表现

## 峰值稳定性测试

• 在峰值压力下，保持30分钟（可讨论）稳定

## 秒杀场景测试

• 针对秒杀类业务，制定秒杀测试场景

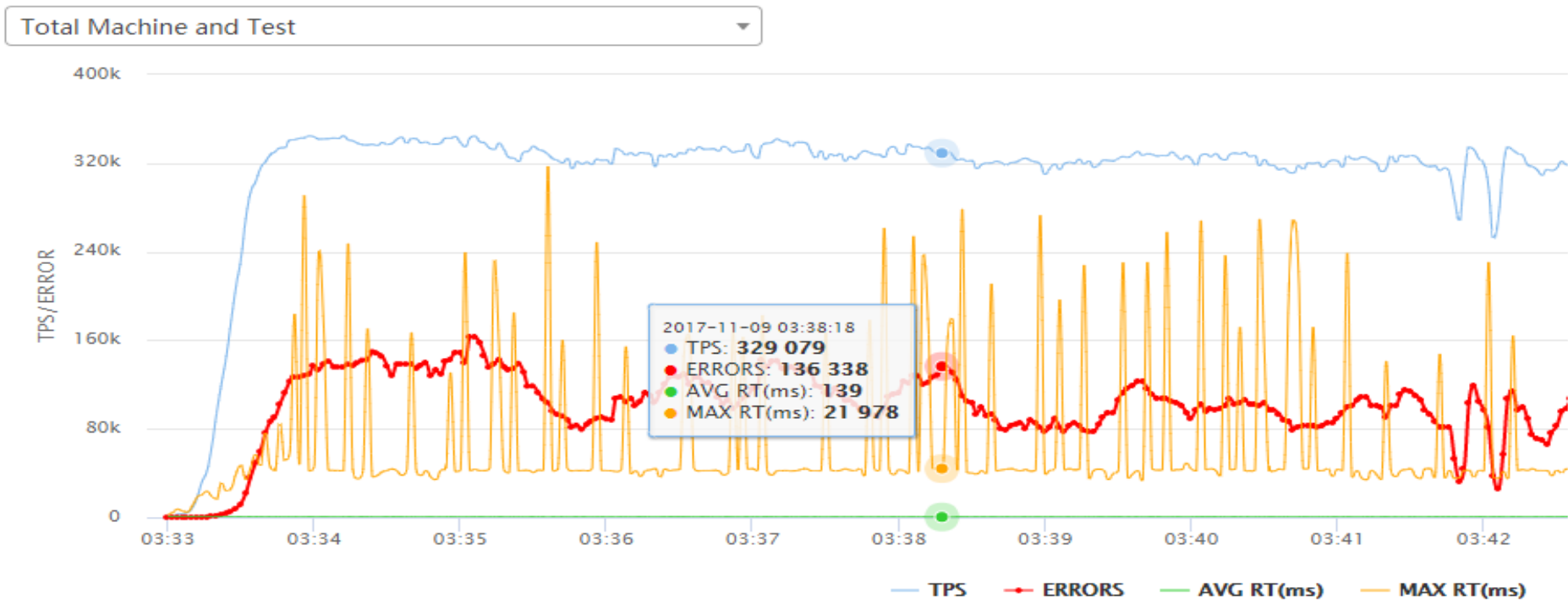
## 限流演练

• 多级限流，保护系统稳定提供服务

## 降级演练

• 非核心业务降级，提升整体服务能力

TPS-RT曲线



HTTP响应字节数（KB/s）曲线



# 系统服务化平台

接口调用总数：12,135,874,404  
失败调用总数：340,129

按应用统计：

序号	应用名称	成功执行总数	前一日成功执行 总数	30天平均执行 总数	平均耗时	前一日平均 耗时	30天平均耗 时	失败次 数	前一日失 败次数	30天平均失 败次数
1		1,507,428,963 ↑	331,465,172	151283070	19.9 ↓	23.35	105.7	--	--	--
2		1,490,688,359 ↑	308,939,666	148249996	1.46 ↓	1.85	3.21	--	--	--
3		1,428,106,100 ↑	285,158,926	139997948	7.01 ↓	12.68	49.49	--	--	--
4		1,353,611,291 ↑	201,975,783	119677214	3.08 ↓	3.76	2.57	--	--	--
5		1,327,544,219 ↑	268,601,492	148823471	6.8 ↓	13.18	27.53	11	--	3.714
6		865,875,013 ↑	186,133,147	89328492	3.94 ↓	4.69	5.46	--	--	--
7		610,625,911 ↑	117,317,810	60264184	1.22 ↓	3.33	13.09	--	--	--
8		376,552,559 ↑	68,427,990	36833367	5.37 ↓	6.06	7	--	--	--
9		353,364,519 ↑	64,671,615	36753281	9.74 ↑	8.1	8.88	--	--	--
10		347,159,416 ↑	202,740,031	113601361	16.13 ↓	19.8	21.49	--	--	--
11		322,878,920 ↑	45,512,053	30721233	3.57 ↓	5.45	4.54	--	--	--

## 内部服务调用统计

# 问题抛出

- 应用报错
- 错误日志搜集、告警
- 系统 QPS 上不去
- 系统处理效率变 “慢”


# Why?

- 通过错误日志追查调用链路的源头集群
- 查看源头集群的错误告警信息
- 是否是处理变慢？！
- 集群流量负载是否均衡
- 分析集群整体负载：CPU、内存使用量、网卡流量、存储 IO



# 应用层初步定位

- 是否有近期变更，发布上线
- 查看日志告警信息，确定报错点
- 分析报错原因：
  - Bug?
  - 死锁?
  - 缓存穿透业务降级?
  - 内存泄露，垃圾回收?
  - 调用超时?
  - .....

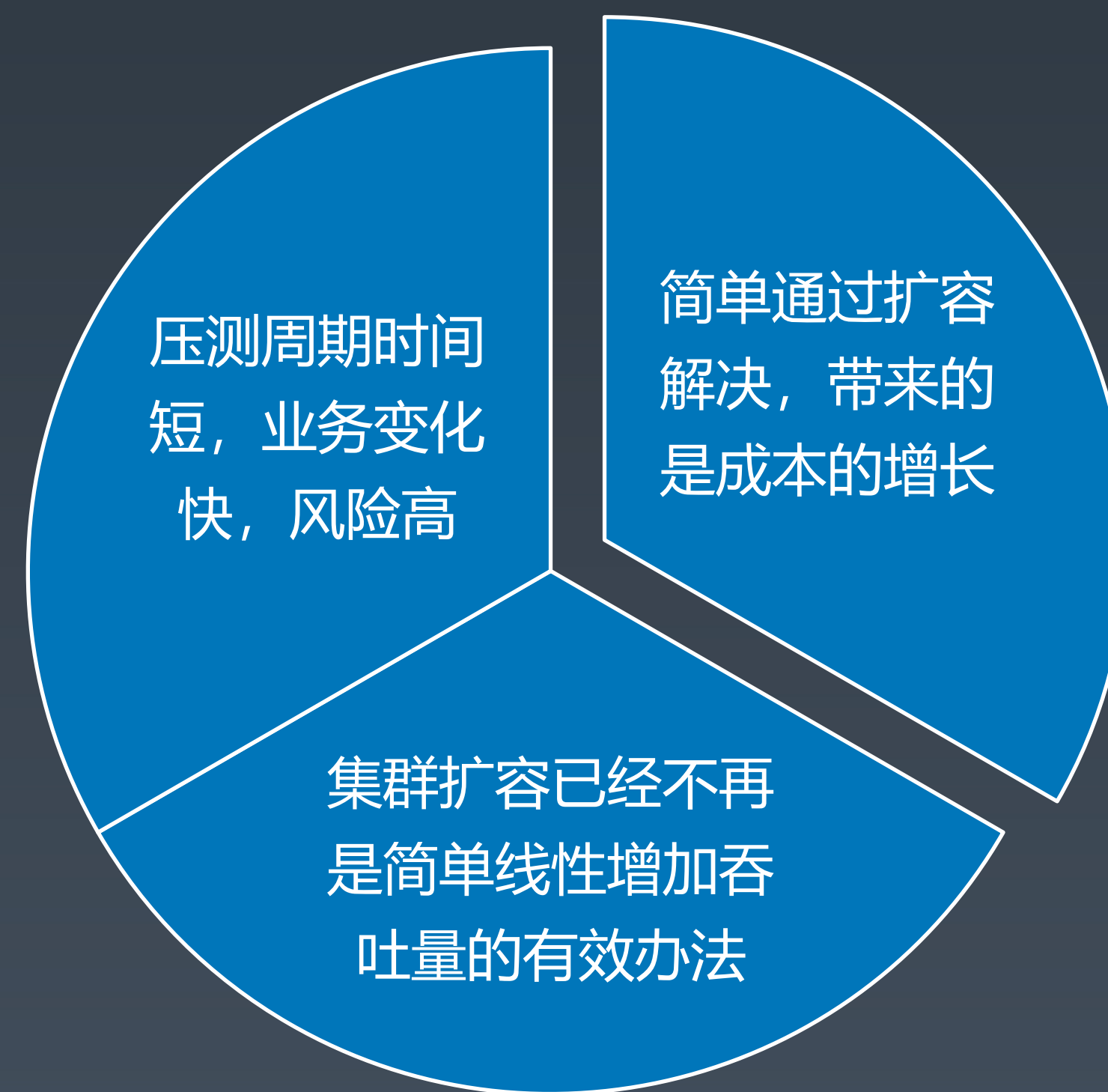


90%问题都与变更有直接关系

# 网络问题应用层表现


- 业务调用出现超时报错
- 集群处理响应 RT 增加
- 集群上下游业务网络重传增加
- 特别是缓存集群

# 为什么一定要解决网络问题！



# 压测出现的一些问题

- Nginx 云主机软中断队列满，出现丢包超时
- 压测云主机客户端到 Nginx 建链慢
- 缓存集群调用 RT 高，集群重传高，商品、库存、交易出现调用缓存超时
- 相比缓存使用云外物理机，云内的缓存云主机集群表现差距较大
- 云主机 Steal 高
- 虚拟网卡丢包
- .....



排查难度大  
压测消耗大



# 服务集群分析常用工具

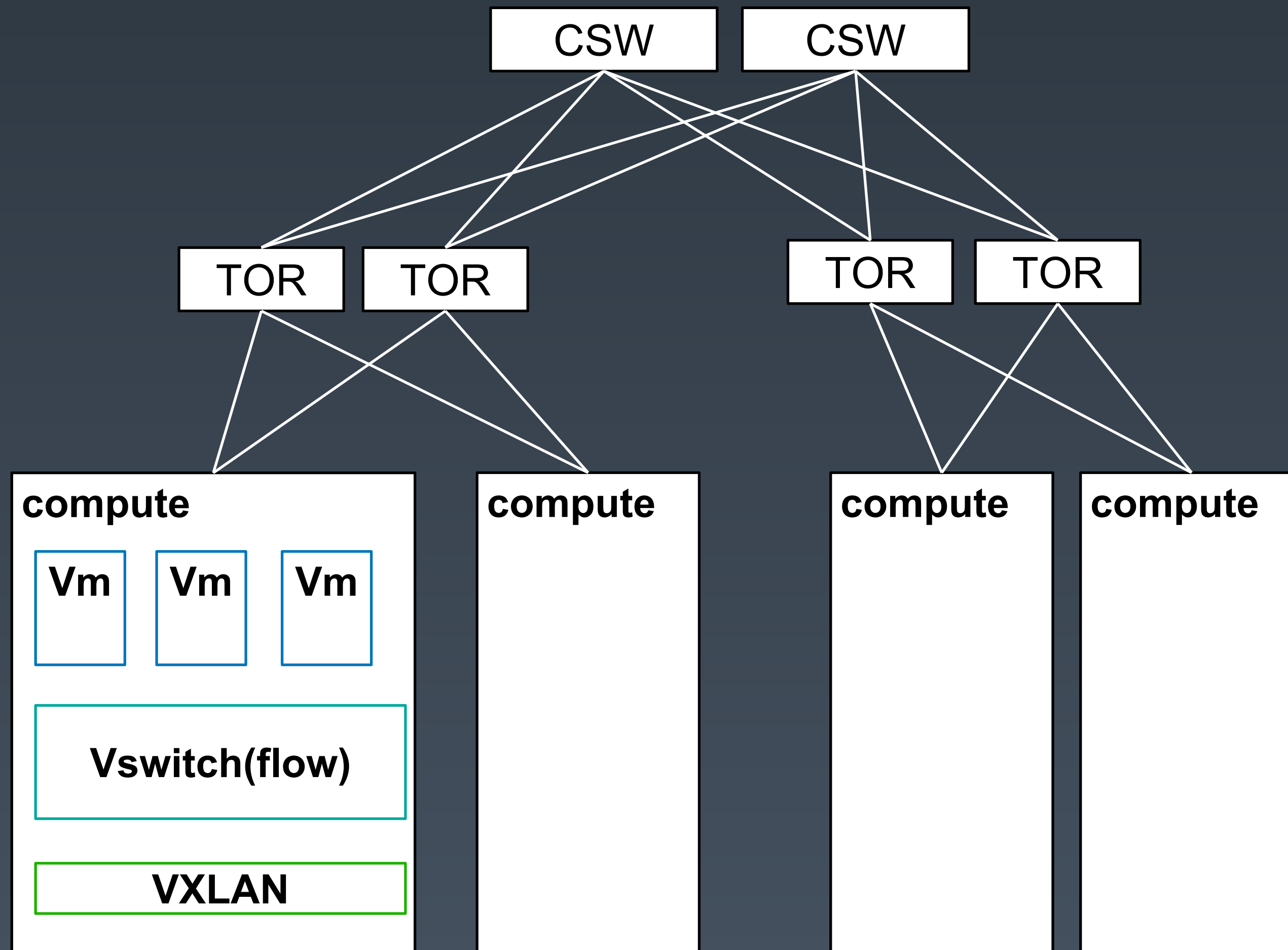
- 监控平台
- 云主机操作系统检查网络状态  
nstat
- 使用 perf 在宿主机上跟踪 KVM  
perf, 火焰图
- 网络虚拟化性能分析对比  
iperf  
sar
- 全链路抓包分析  
tcpdump  
ovs-tcpdump



# 服务集群分析策略

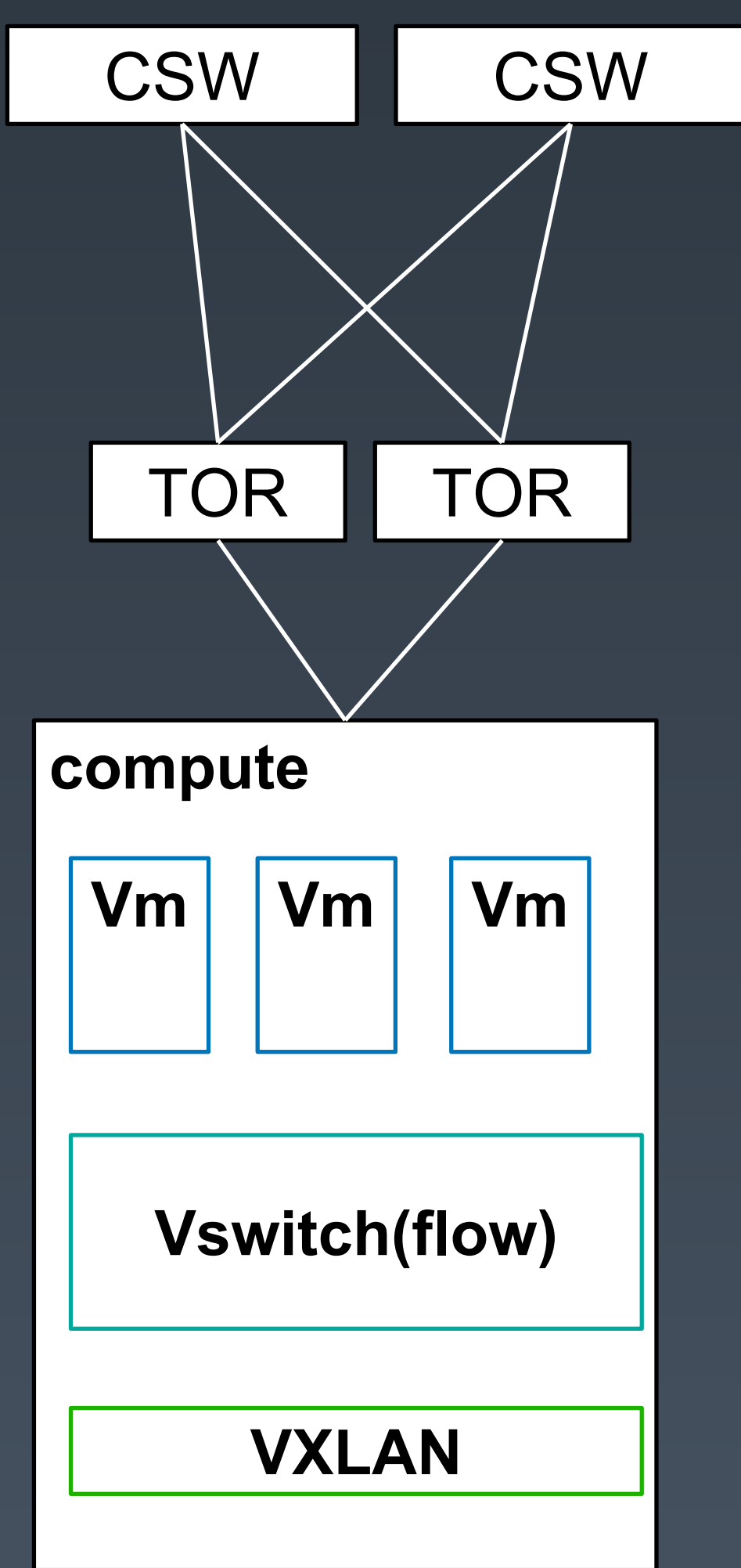
- 全链路抓包分析
- tcpdump 参数选择, 减少 pcap 大小  
使用 VXLAN 内层包 IP 过滤  
使用 -s 70 (vm) , -s 130 (宿主机)
- 程序分析  
找出 TCP 的数据包与 ACK 相差超过 10ms+ 的包
- 人工检查  
根据抓包的时戳、TCP 序列号、TCP Timestamp、IP 的 ID 域分析比对  
GSO, GRO, 不能一一对应

# 云内业务调用



- 云内服务调用不出 VPC
- 流量线路可以在同一台计算节点内、流量绕 TOR 到另一台计算节点、绕 CSW 到另一组 TOR 下
- 机房服务器放置的规划，一般有存储区、网关区、计算区、公共服务区

# 云内服务集群流量分析



VPC核心

- 监控协议调用频率过高影响交换机转发性能
- 网卡流量过载
- 线路故障, CRC 校验异常

接入交换机

GuestOS

- 应用层限流
- 虚拟网卡丢包
- 连接队列溢出
- 网卡流量超过 QOS

虚拟交换

- DPDK PMD CPU 占用率过高
- 安全组配置
- 网卡流量超过 QOS
- 流表配置

物理网络

- 网段路由配置
- 网卡 bond 状态



# 基础监控

服务器性能监控

CPU、内存、网卡、磁盘、TCP.....

应用性能监控

JVM堆内存、GC、Thread、CPU利用率、Method性能.....

业务指标监控

下单数、支付数、购物车请求数.....

调用链路监控

RT、TPS、Exception.....

底层组件监控

RT、TPS、连接数、连接状态、消息积压、zk节点数.....

系统异常监控

流量尖刺、Exception log、服务线程数、异常报警.....

# 云内服务监控



视图


状态

配置

报警

拓扑

# 机房物理网络监控



链路监控

☆

ISP端口报警(网值50%或者DOWN) ▾

No data to show ?

内部链路故障(不含GDC/HADOOP)

两端设备	设备IP	原始链路数量	当前链路数量	上次链路数量	挂掉端口	链路报警	当前IN带宽	当前OUT带宽	总带宽	流入使用率	流出使用率	流量报警
nr		57	1	1		0	4.86 Gbps	1.30 Gbps	10.00 Gbps	48.57%	13.04%	1
nr		58	1	1		0	4.35 Gbps	1.60 Gbps	10.00 Gbps	43.47%	15.99%	1
hc		49	4	4		0	2.28 Gbps	1.71 Gbps	4.00 Gbps	57.11%	42.78%	1
hc		202	1	1		0	5.24 Gbps	1.76 Gbps	10.00 Gbps	52.40%	17.58%	1
bj		5	1	1		0	9.13 Gbps	42.94 Mbps	10.00 Gbps	91.28%	0.43%	1
hc		6.105	2	1	FortyGigE3/5/0/46	1	2.90 kbps	1.22 kbps	40.00 Gbps	0%	0%	0
nr		0.176.193	2	0		1	0 bps	0 bps	20.00 Gbps	0%	0%	0

# 交换机监控现状

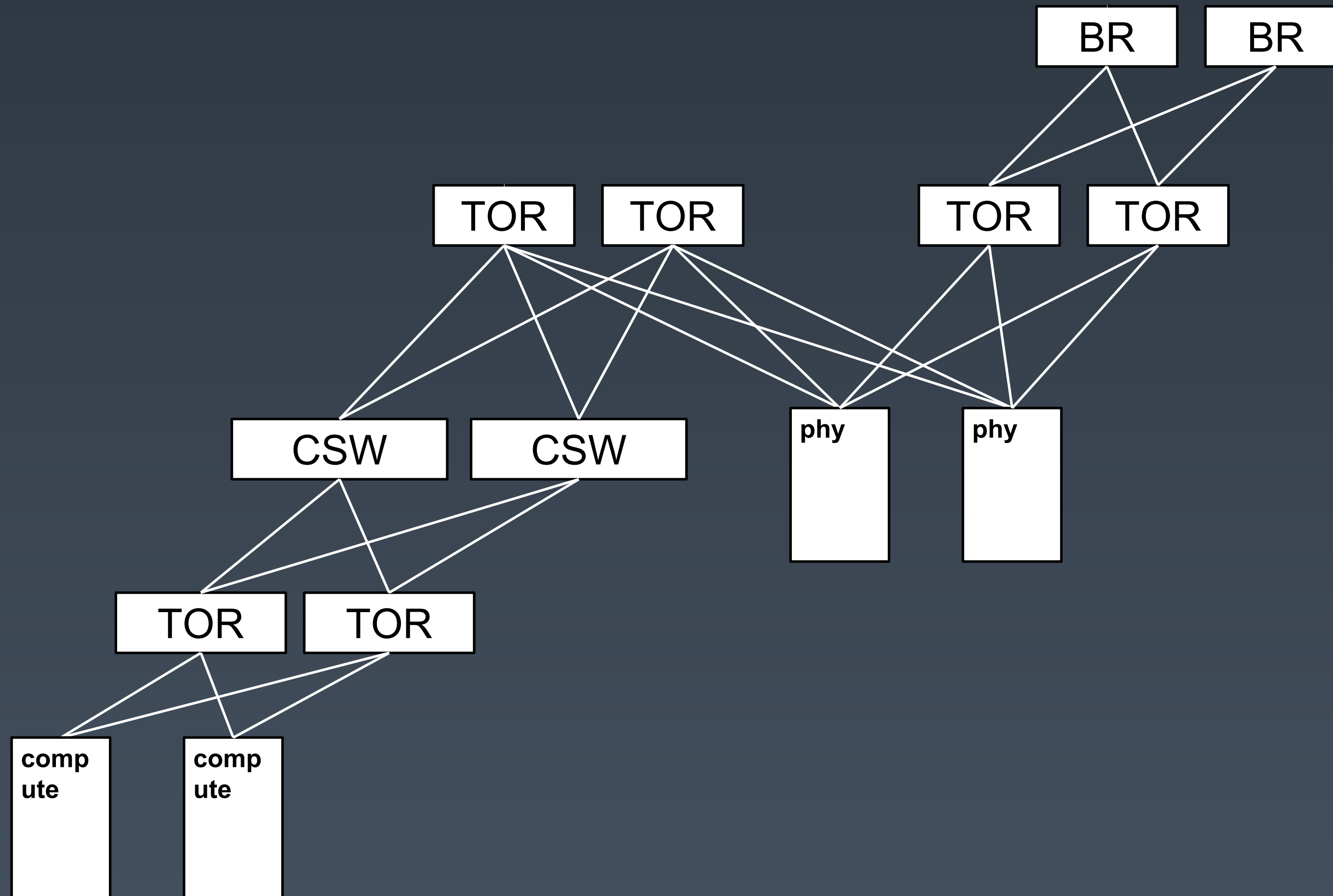
- 早期交换机监控通过 SNMP 协议，交换机监控粒度维持在分钟级
- 新一代基于 gRPC 协议的监控协议，由交换机主动 push 监控数据到监控平台，可支持秒级
- 物理网络层面带宽流量的监控，由于粒度在分钟级，所以压测峰值流量很可能无法发现物理网络瓶颈
- 新的支持秒级监控的机器需要有厂商定制，改造成本和采购成本很高



# 分钟级监控下的机器布置策略

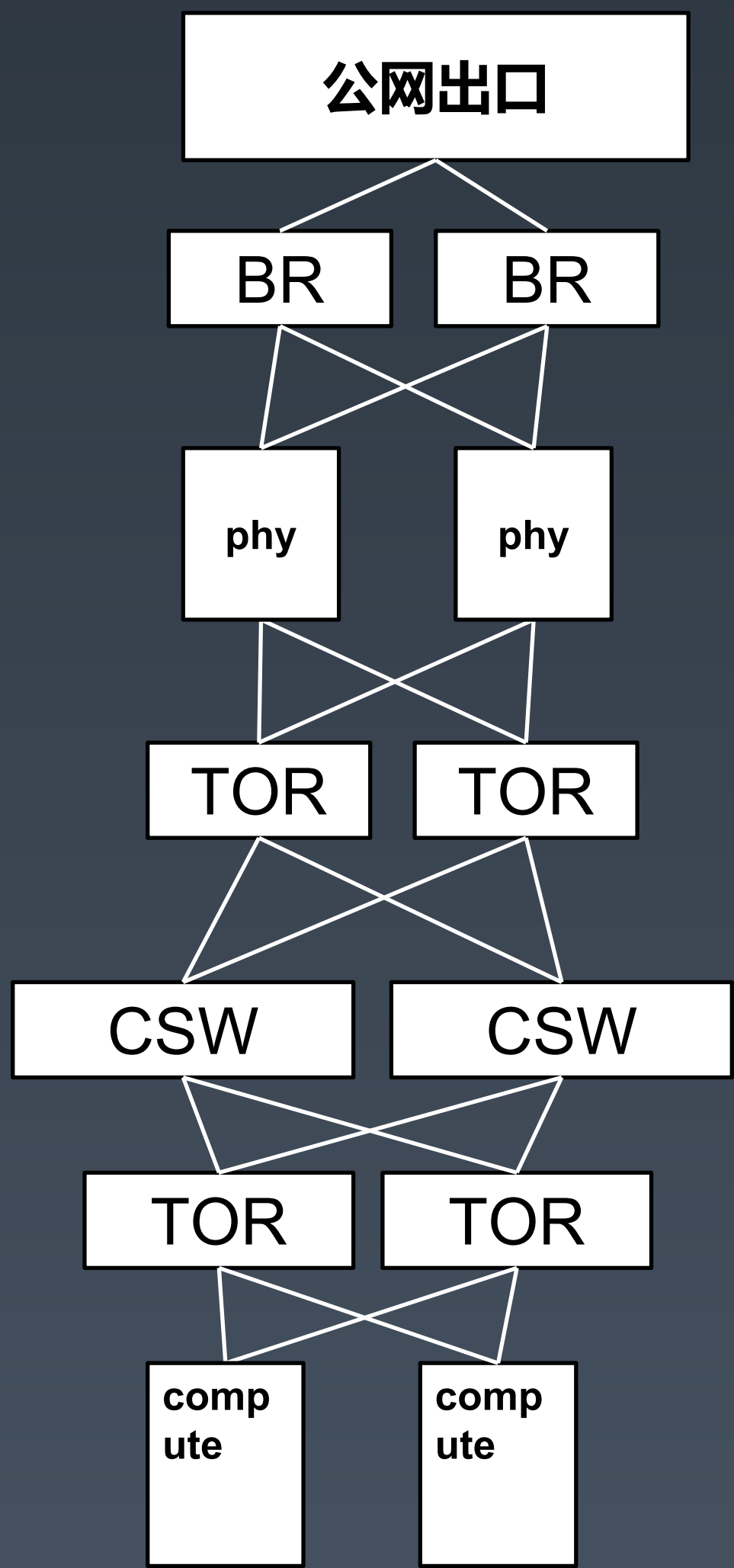
- 交换机上联口到核心带宽假如打满，可能导致丢包
- 物理交换机 CPU 比较弱，监控无法做到秒级，因此没有发现问题
- 物理层面容量规划，打散高转发流量的物理机器（Nginx、缓存等）
- 一组 TOR 下不超过 3 台高性能机器

# 云内与云外网络调用



- 出云的网络流量，必须会经过网关服务器做 NAT
- 网关服务器提供包括 DNAT 映射、SNAT 等
- 公网网络出口：BGP、联通、电信
- 出云访问可以通过 VPC SNAT 访问或者绑定公网地址访问
- 外部访问应用系统主要通过 NLB 服务访问

# 云内云外服务网络流量分析



公网出口、运营商网络

- 运营商线路状态
- 运营商割接
- 公网网络质量监控

机房核心

- 路由配置
- 出口带宽

网关节点

- Vswitch 流表配置问题
- PMD CPU 占用率过高
- 网卡带宽过高导致丢包
- SNAT 新建连接性能瓶颈

机房网络

- 同云内服务集群流量分析

计算节点

# 网络质量监控



- 全国范围，覆盖所有省份
- 24 小时不间断
- 各省份用颜色标明网络质量
- 大数据分析处理，实时掌握全网质量
- 异常告警



# VPC网关及负载均衡

- BGP ECMP  
网络高可用  
水平扩展及负载均衡
- OVS-DPDK  
虚拟交换机, DPDK 驱动提升网络转发性能
- SSL Offload 加速  
SSL 握手异步化 + Intel QAT 资源池  
压力大时优化明显、压力小时反而增加延时  
CPU 计算消耗减少 50%

# 总结

- 清晰的应用逻辑调用关系
- 明确的网络链路拓扑
- 完善的监控系统



全球技术领导力峰会

Geekbang> | TGO 鲲鹏会  
极客邦科技

# 500+ 高端科技领导者与你一起探讨 技术、管理与商业那些事儿



🕒 2019年6月14-15日 | 📍 上海圣诺亚皇冠假日酒店



扫码了解更多信息

THANKS! | QCon <sup>th</sup>