在生产环境管理一个快速增长的OpenStack云

ByteDance

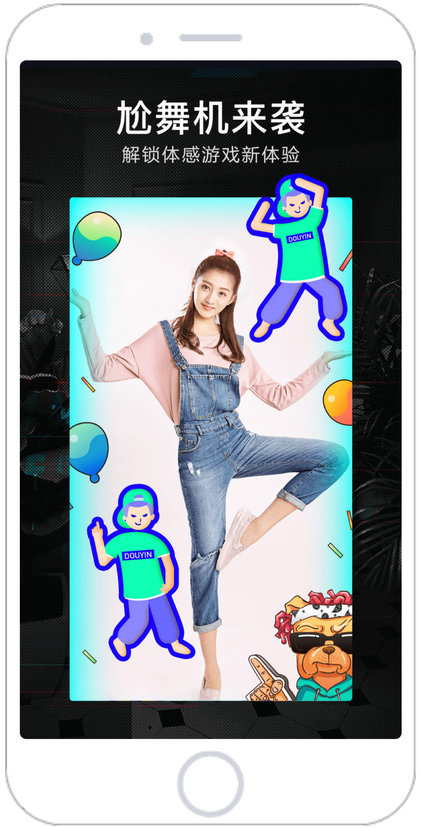
Who We Are

ByteDance

一款个性化资讯推荐引擎产品，致力于连接人与信息，让优质、丰富的信息得到高效、精准的分发，为用户创造价值。

抖音短视频

一个帮助大众用户表达自我，记录美好生活的短视频平台

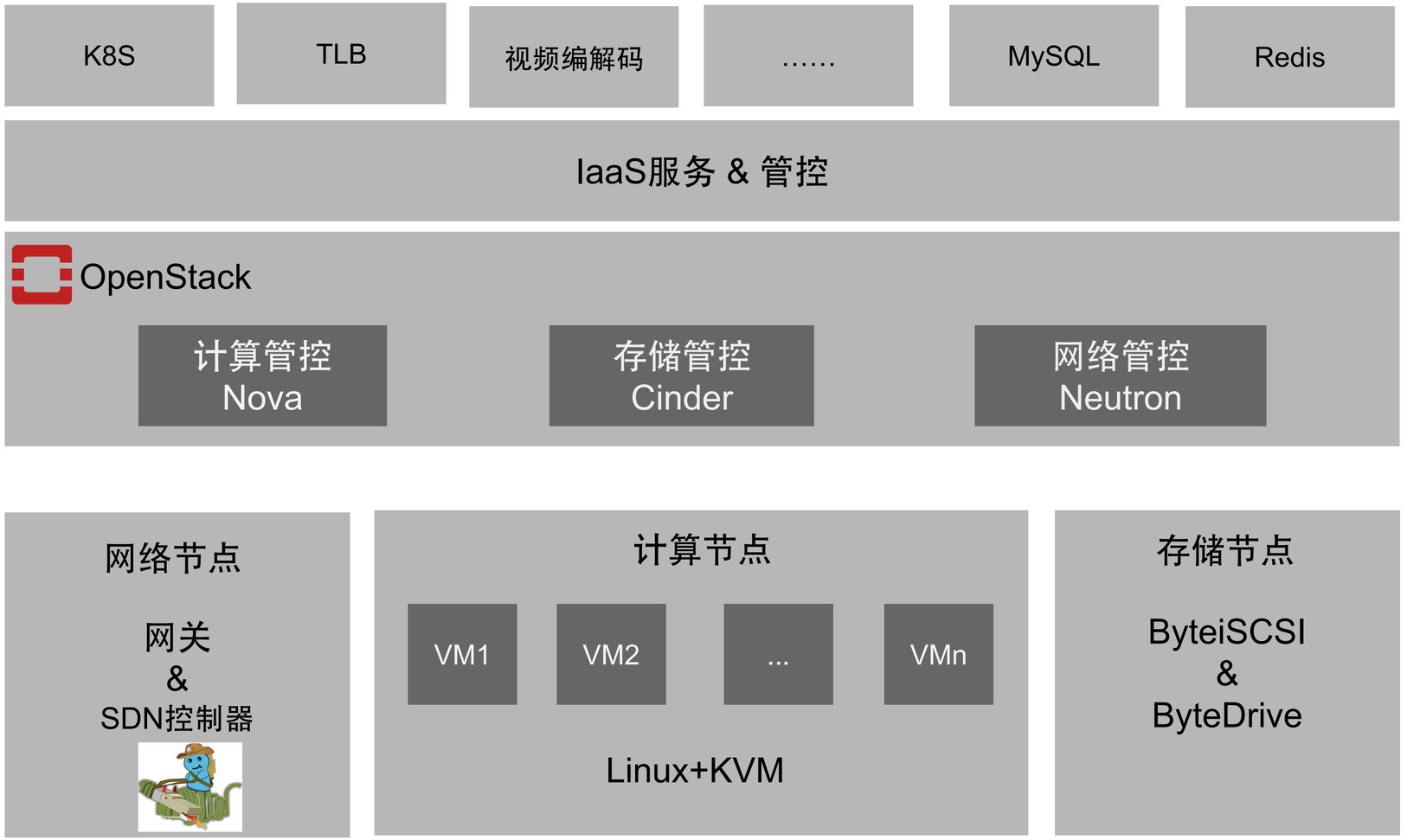


不止是今日头条和抖音短视频，**我们还有更多产品**

IaaS@Bytedance

ByteDance

字节跳动IaaS Stack



虚拟化

零损耗VM

sriov，hugetlbfs, cpu隔离，smart idle polling...

服务管控

生命周期

启动，停止，重启，创建，删除，重建...

镜像

Debian，Cirros, CentOS, Ubuntu..

SDN

Overlay

支持虚拟机, IP地址漂移

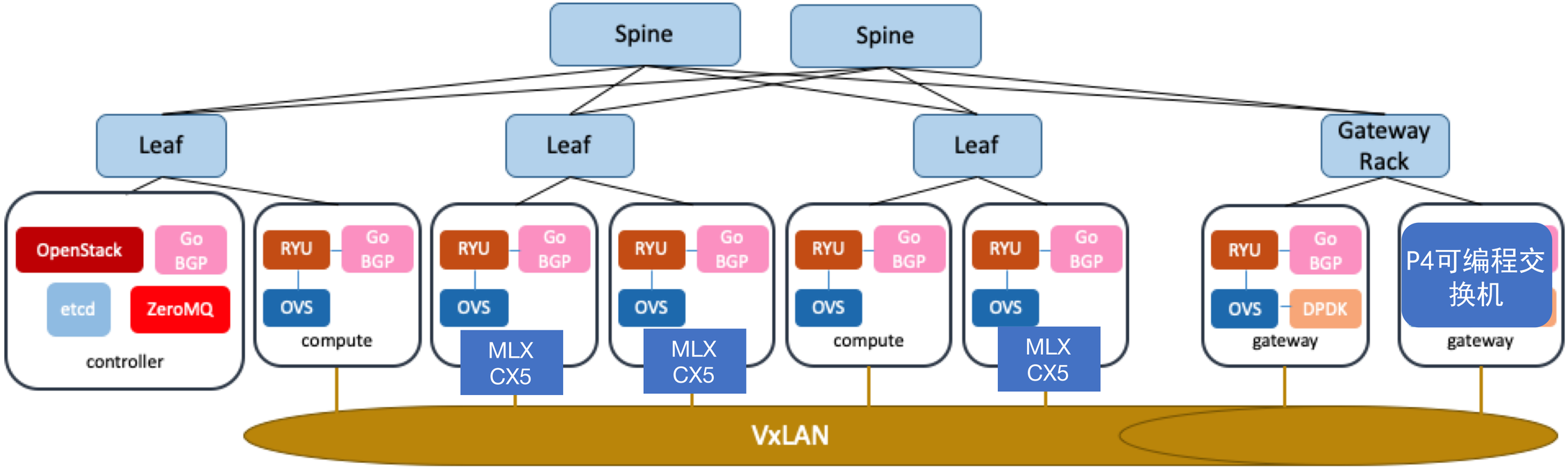
网络服务

DNS，DHCP，vRouter，Firewall

虚拟存储

本地磁盘，ByeiSCSI(p2p)，ByteDrive(集群)

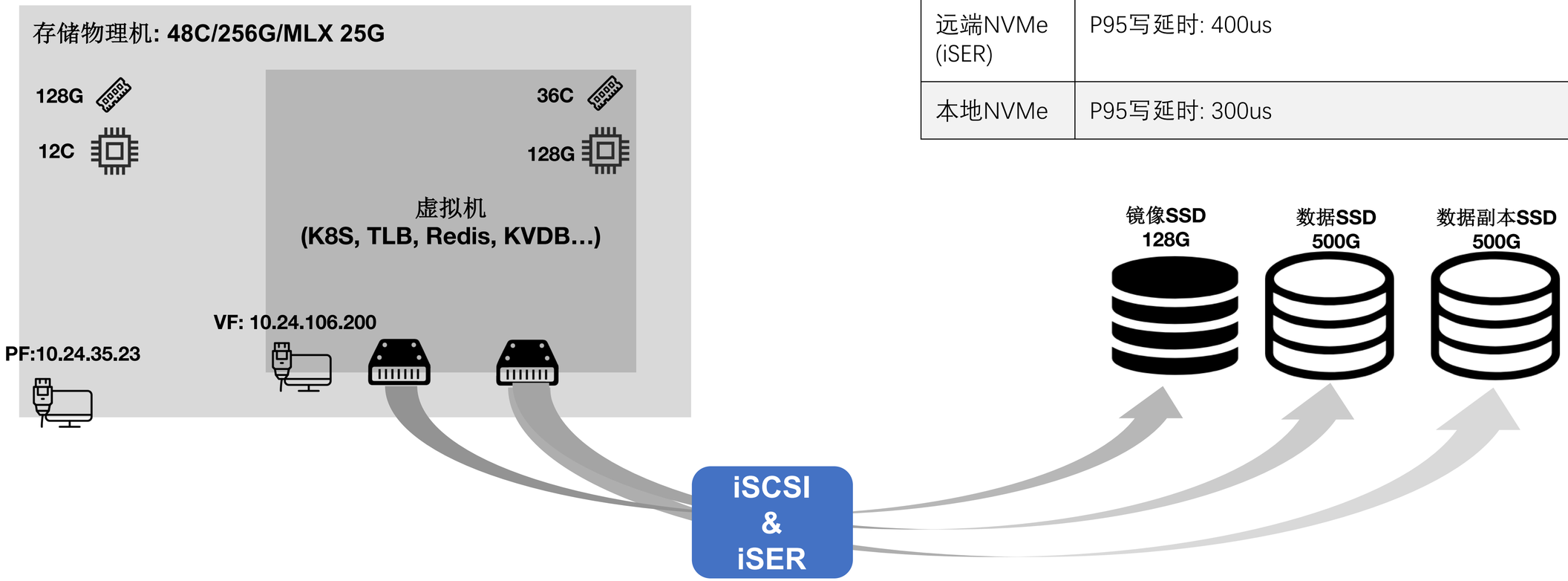
基于OpenStack的大规模Overlay网络集群



PPS@64B, 16队列	10.8M
延时(东西向)	19.5us(VM) v.s. 17.5us (物理机)
带宽(单流+LRO)	22.4Gb
Host CPU Overhead	<0.5C

基于OpenStack的计算存储大规模混部集群

CPU	PVIPI， Idle Polling List,
内存	虚拟机hugetlbfs, 避免受主机内存回收干扰
网络	基于网卡的带宽控制, PF: 10G, VF: 15G
远端SSD (iSCSI)	IOPS: 49K@8k 100%随机写, Depth=128; P95写延时: 1200us
远端NVMe (iSER)	P95写延时: 400us
本地NVMe	P95写延时: 300us

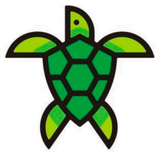


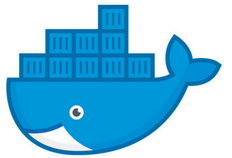
OpenStack@Bytedance

ByteDance

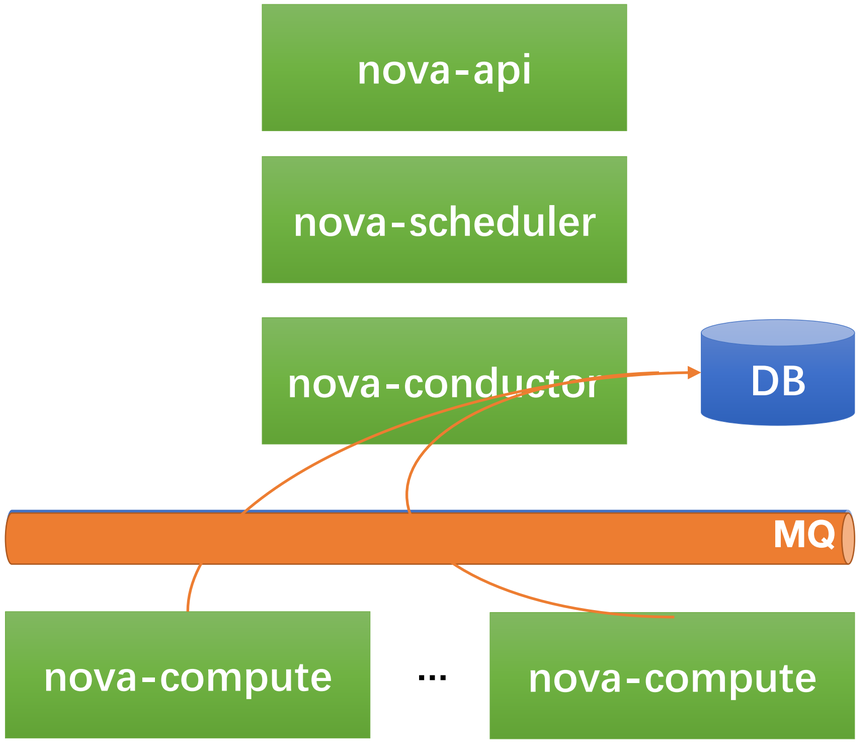
OpenStack大规模

- 规模
  - •单集群10K主机管理
- 部署
  - •OpenStack组件容器化部署
- 运维
  - •通过Ansible部署运维
- 监控
  - •基于influxdb+grafana框架





OpenStack大规模 – 静态负载优化



静态负载来源：主要来自周期任务

优化点：

分类	优化方式	优化点
周期任务优化	旁路周期任务	心跳上报采用memcached组件
	调整周期	调整资源上报周期任务周期
	合并RPC请求	合并多个周期任务中同一RPC请求
	关闭不必要周期任务	比如延迟删除虚拟机周期任务等
	REST调用优化	关闭heal_instance_info_cache周期任务，避免频繁调用Neutron 接口
MQ优化	连接数调整	OpenStack组件RPC连接数调整
	MQ调优	Rabbitmq Runtime Tuning
慢SQL优化	索引优化	
	SQL语句优化	

优化结果：RabbitMQ CPU使用率：38C - > 10C　　RabbitMQ消息量：8k/s -> 1.5k/s

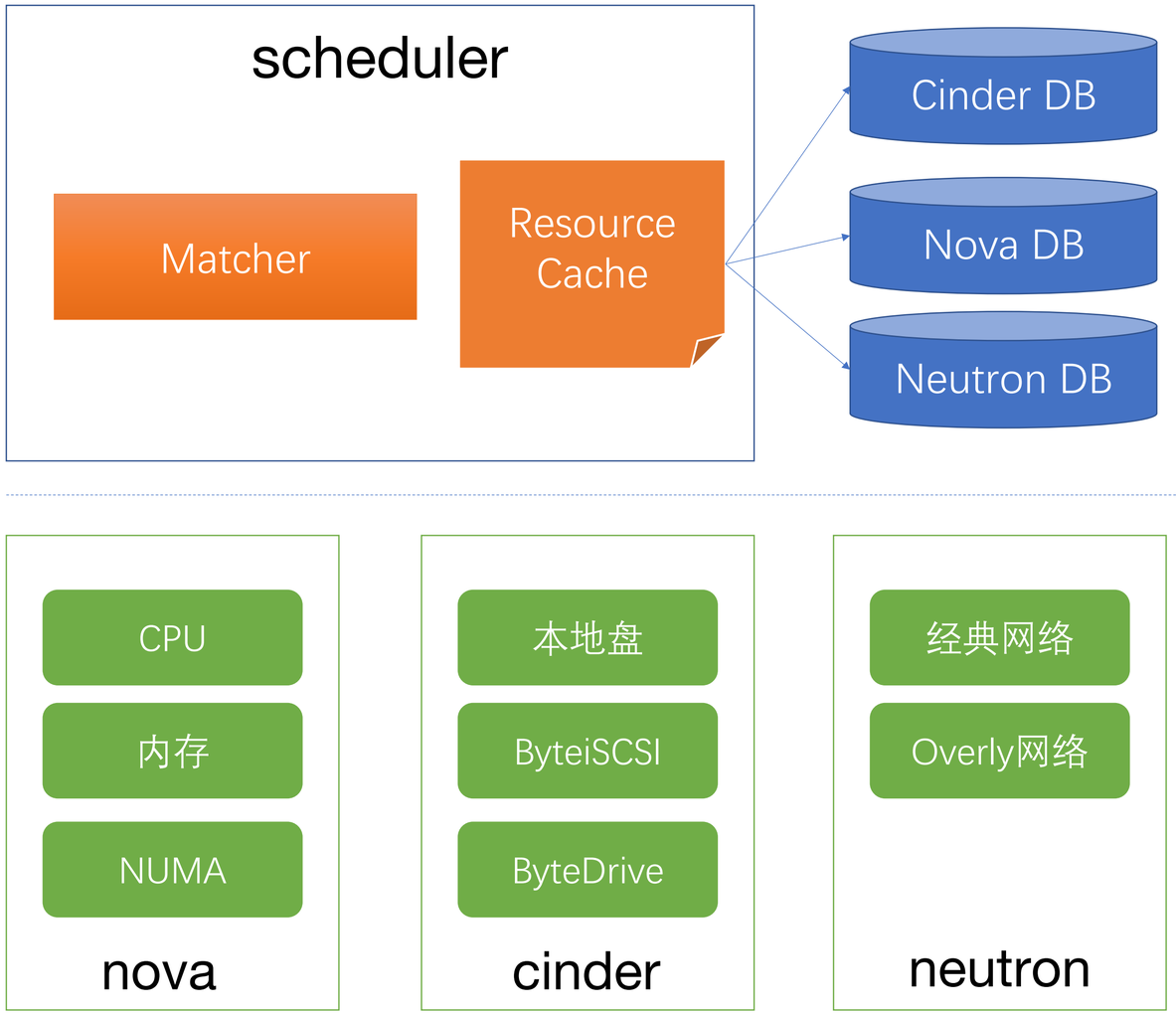
OpenStack大规模 – 并发创建虚拟机优化

优化点：

分类	优化方式	优化点
进程数调整	进程数量调整	调整OpenStack各组件worker数量，提升并发能力
	支持多进程	nova-scheduler、cinder-scheduler支持多进程
MQ优化	调整MQ连接池大小	组件RPC连接池大小调整，提升并发度
	MQ分库	nova、cinder、neutron采用不同的MQ，支持业务并发
锁优化	Quota锁优化	实现NoopQuotaDriver关闭Quota功能
调度器优化	调度器缓存	nova-scheduler采用缓存调度器，缩短虚拟机调度性能
组件间REST调用优化	Token Cache	调整token cache过期时间，减少keystone压力
	减少REST请求	nova-metadata-api缓存port相关信息，减少neutron压力
	命令调用方式优化	开启rootwrap daemon机制，减少每次命令调用耗时
Neutron控制面广播优化	ETCD + EVPN	

200并发创建虚拟机优化至20s

OpenStack大规模 – 调度



- 支持存储、计算、网络混合调度
- 采用golang语言实现
- 资源视图内存cache
- 多schduler主从模式

Networking@Bytedance

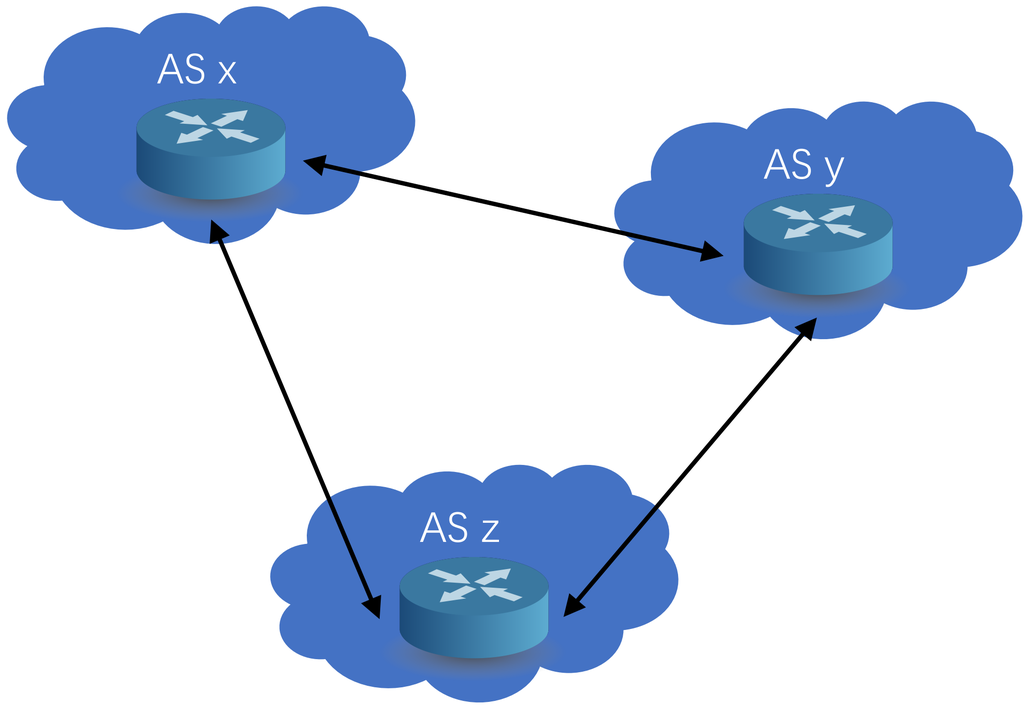
ByteDance

OpenStack Neutron在大规模下的问题

- 中心节点 — neutron server 必须参与所有网络相关的计算
- RPC — RabbitMQ 不能够支撑大规模环境下，控制面消息的分发
- SQL 数据模型关系过于复杂 — 导致高并发读写时效率较低
- 基于namespace的网络服务 — 增加了没有必要的网络路径

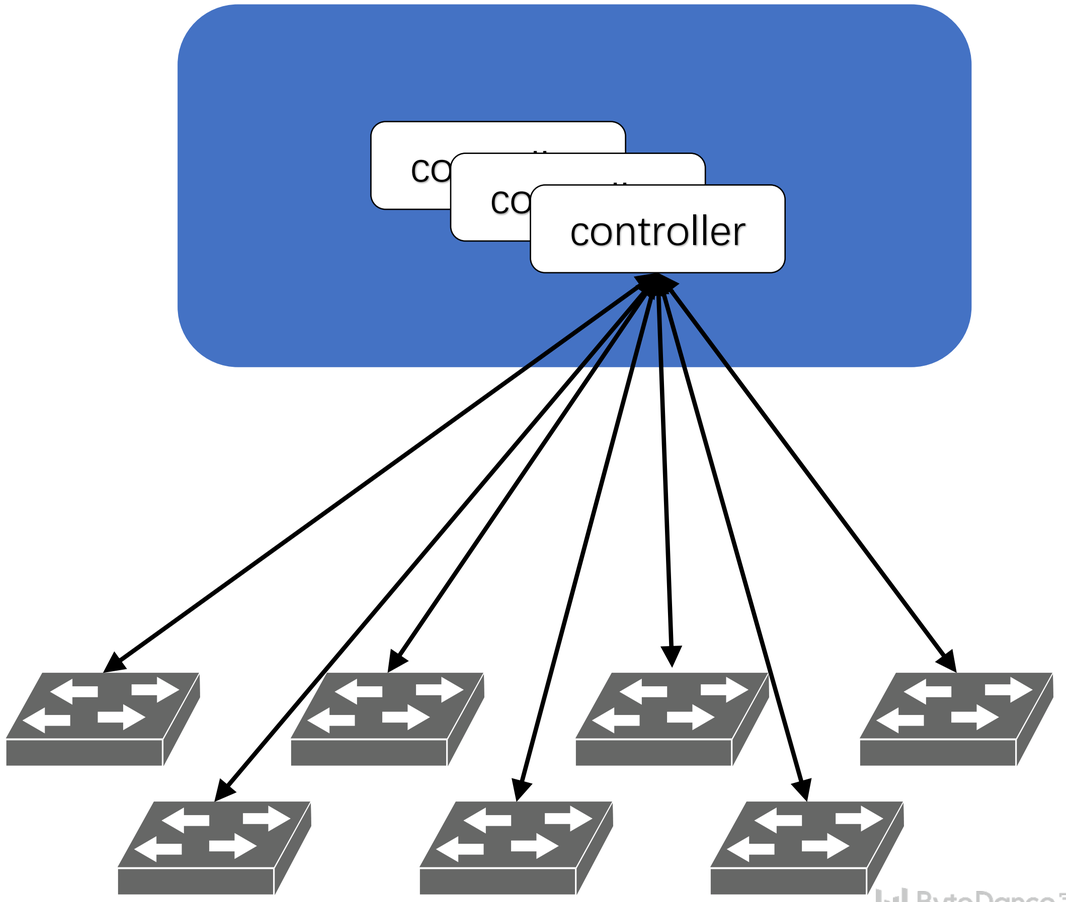
网络控制面

BGP

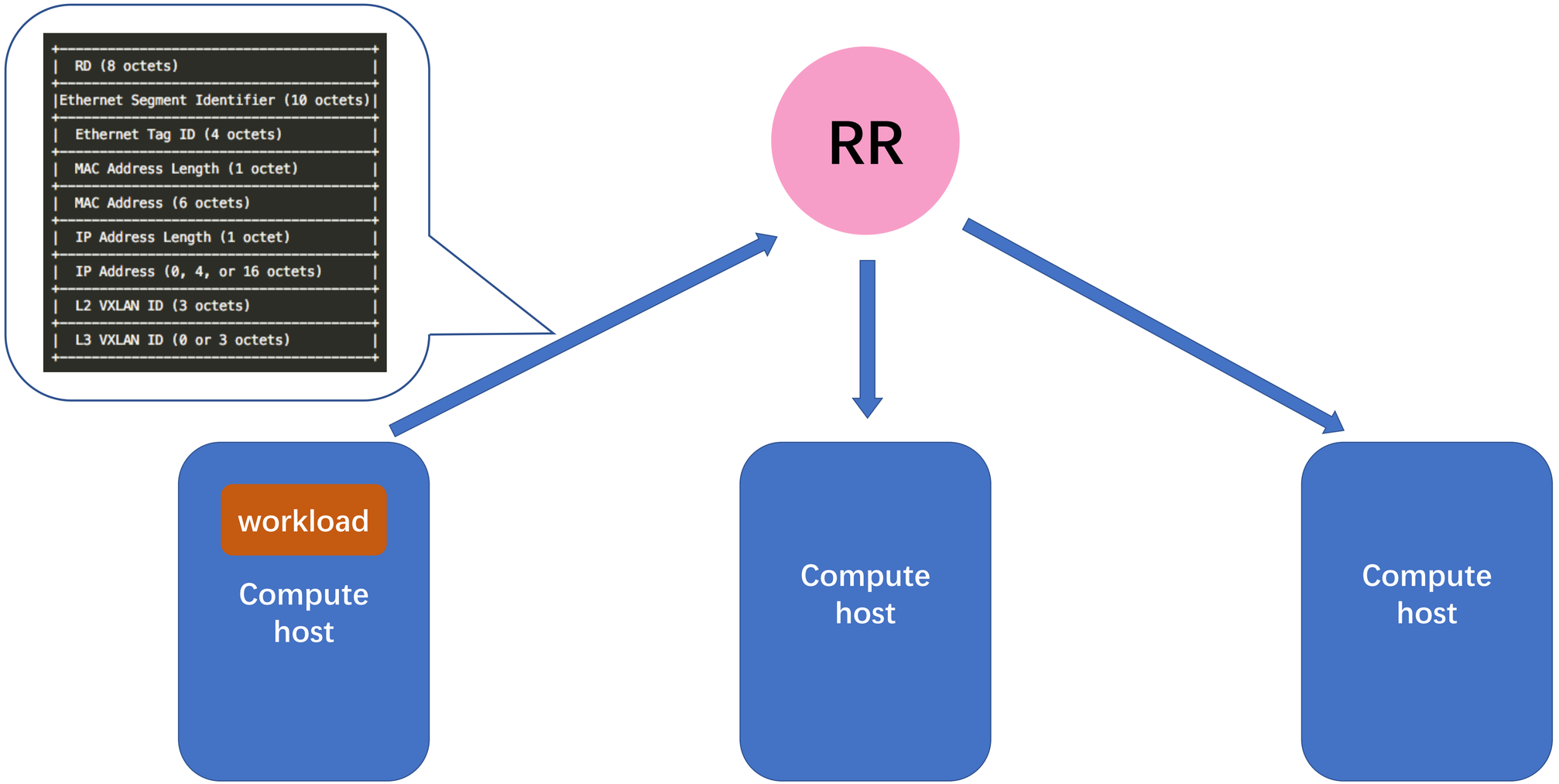


+

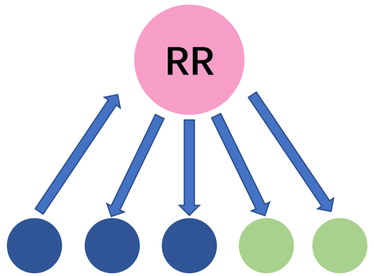
SDN

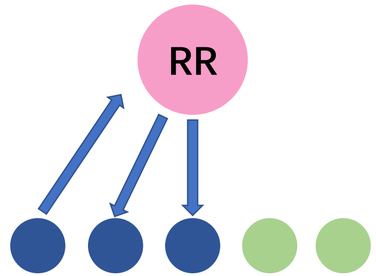


EVPN (Ethernet VPN)



RTC (Route Target Constraint)

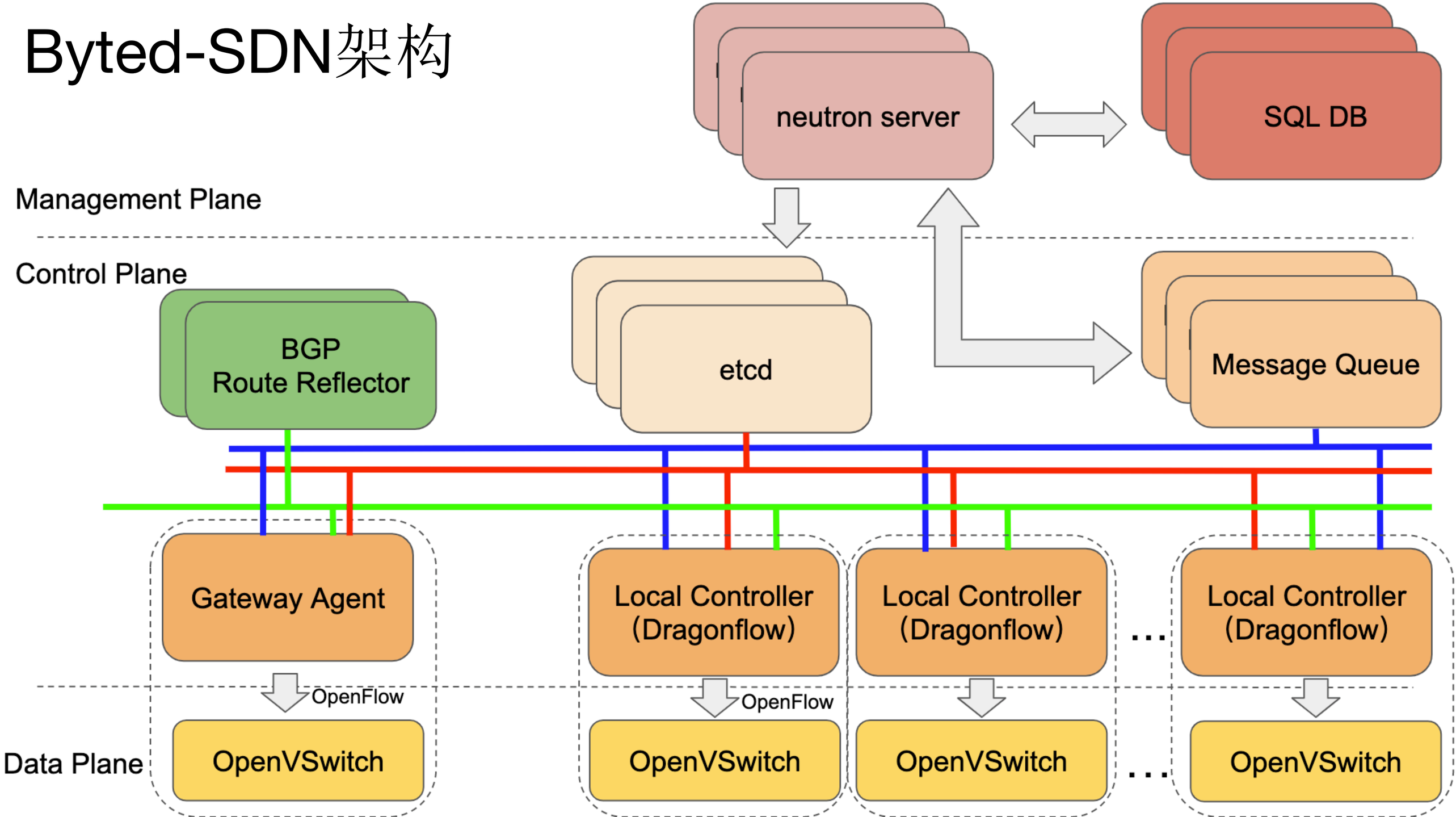




Route Target Constraint

SDN (Software Defined Networking)

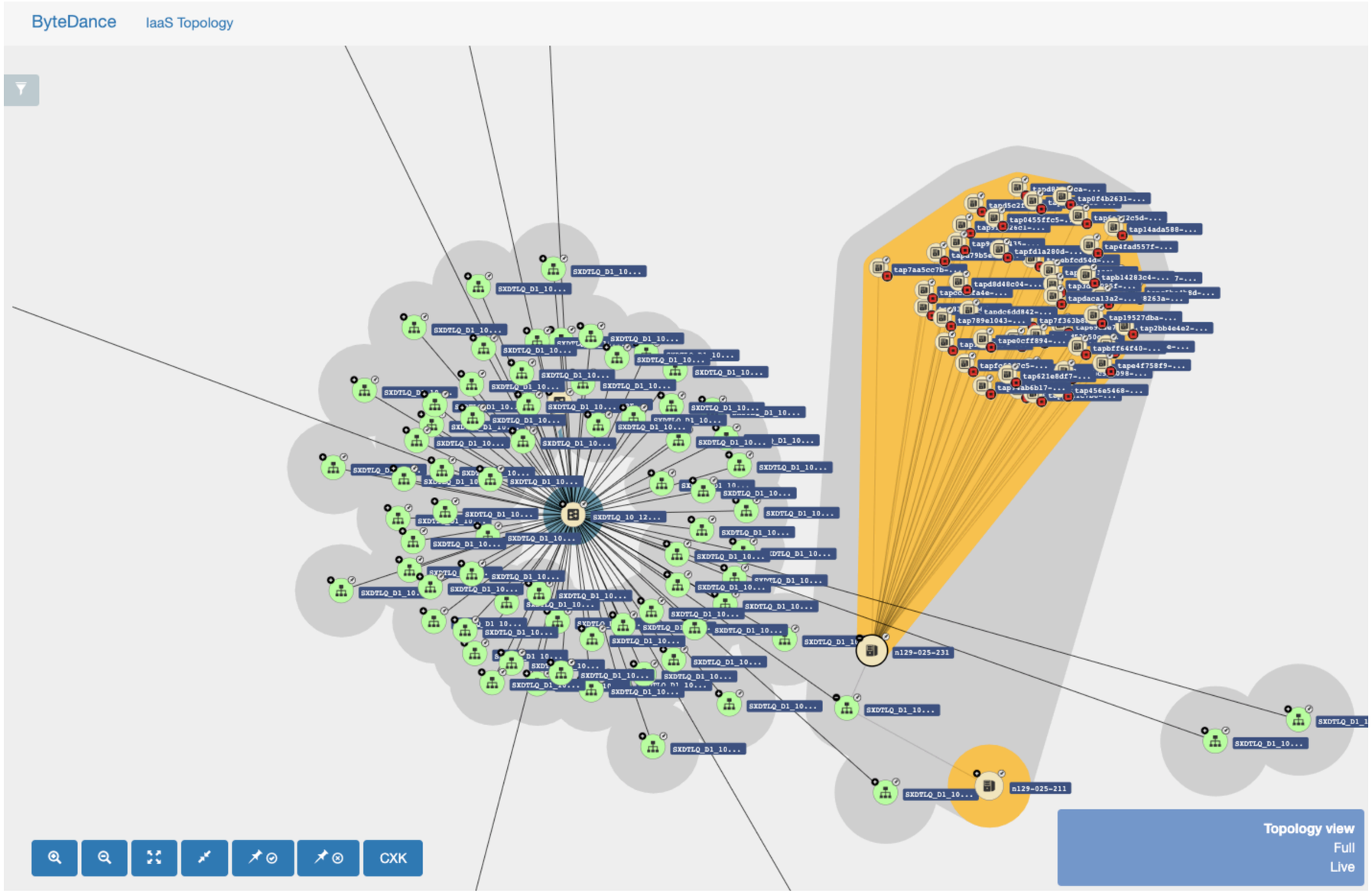
Byted-SDN架构


neutron server
SQL DB
Management Plane
Control Plane
BGP
Route Reflector
etcd
Message Queue
Gateway Agent
Local Controller
(Dragonflow)
Local Controller
(Dragonflow)
Local Controller
(Dragonflow)
OpenFlow
OpenFlow
Data Plane
OpenVSwitch
OpenVSwitch
OpenVSwitch
OpenVSwitch

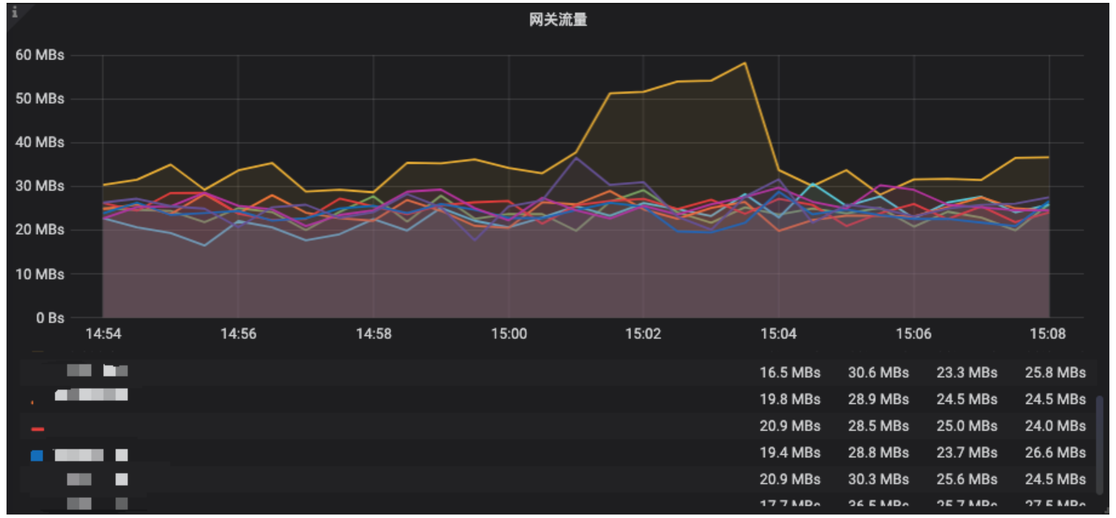
网络监控

- 数据记录
  - 查询当前状态
  - 对比版本差异
  - 回溯历史问题
- 监控告警
  - 异常通知
  - 自动修复

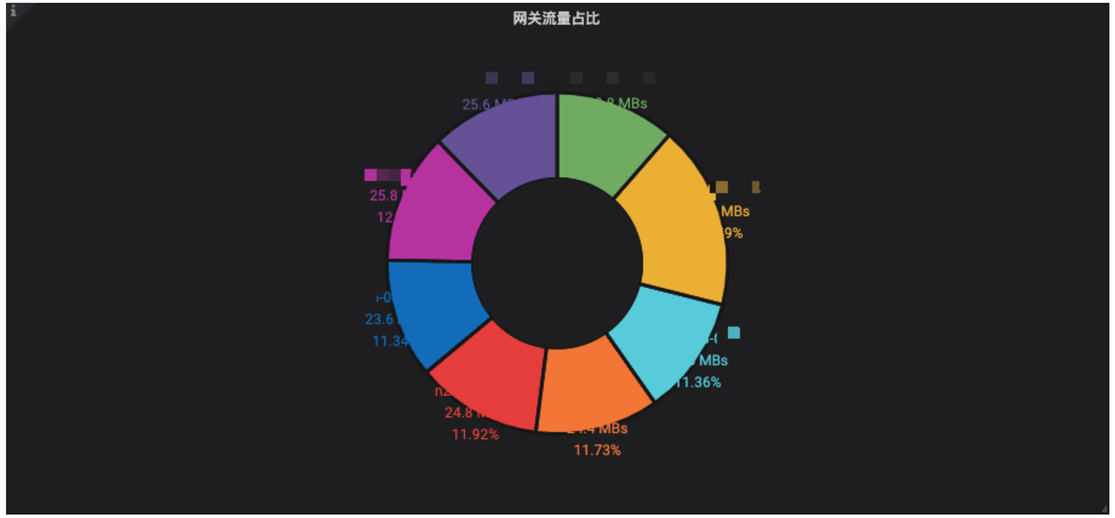
Skydive - 实时网络拓扑与监控



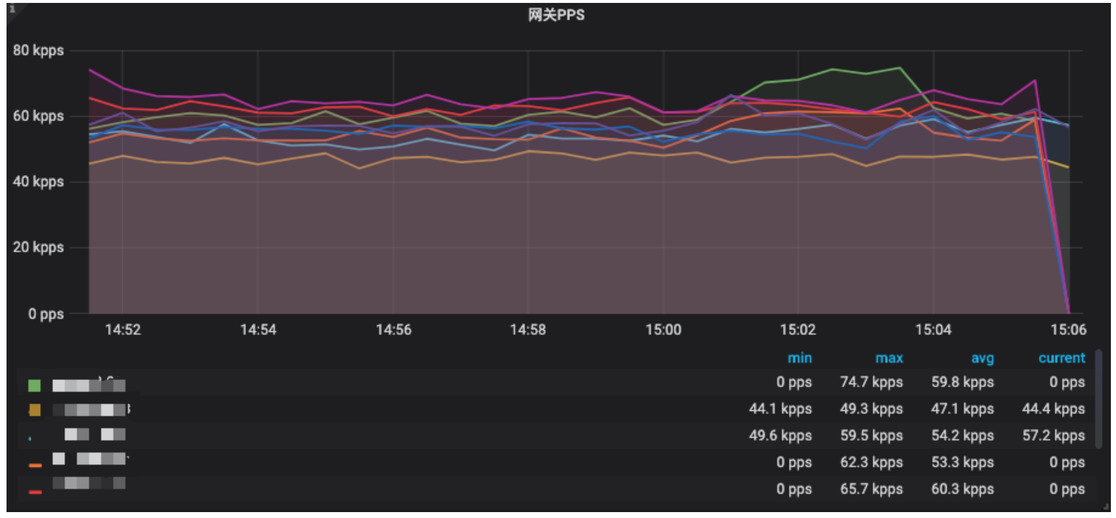
网络节点监控



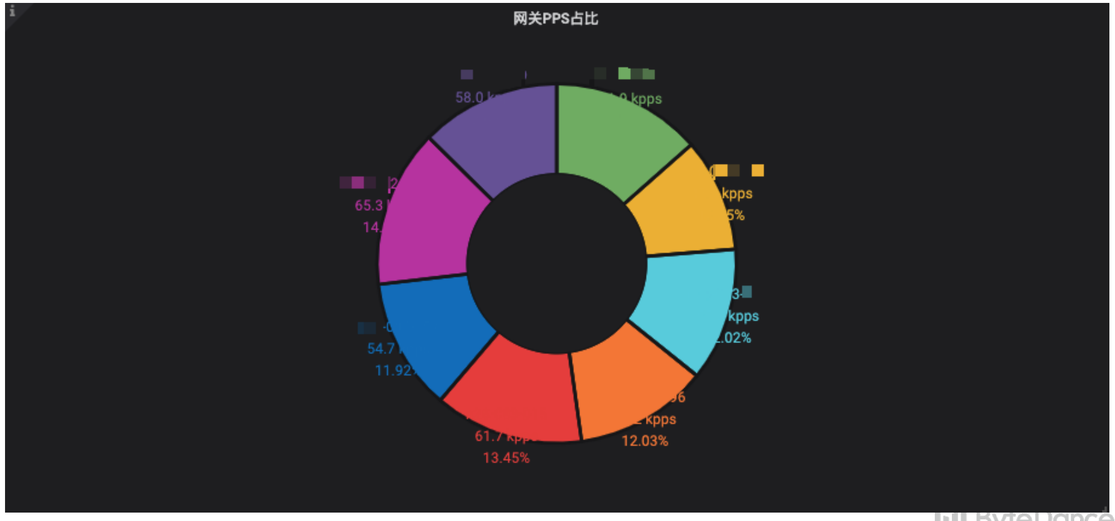
各个网络节点实时流量监控



各个网络节点流量占比

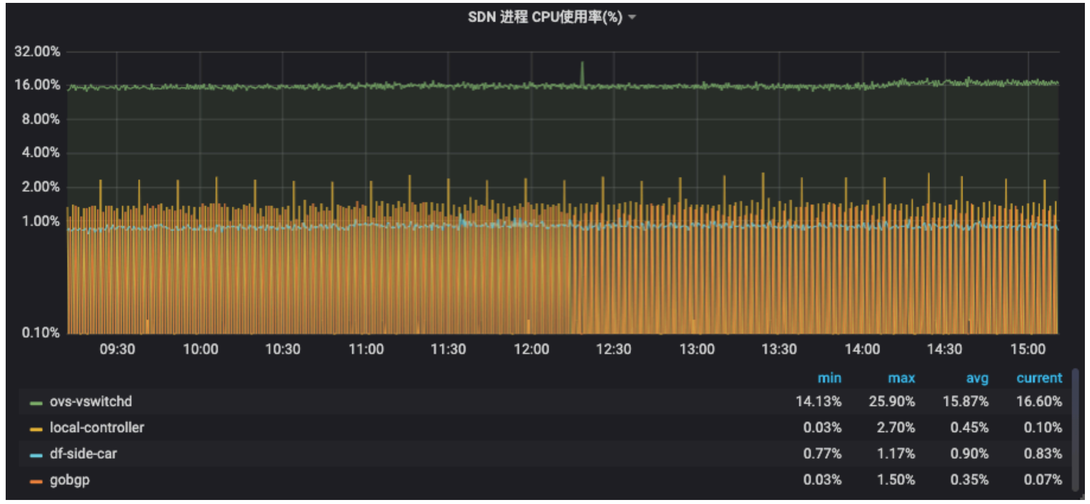


各个网络节点实时PPS监控

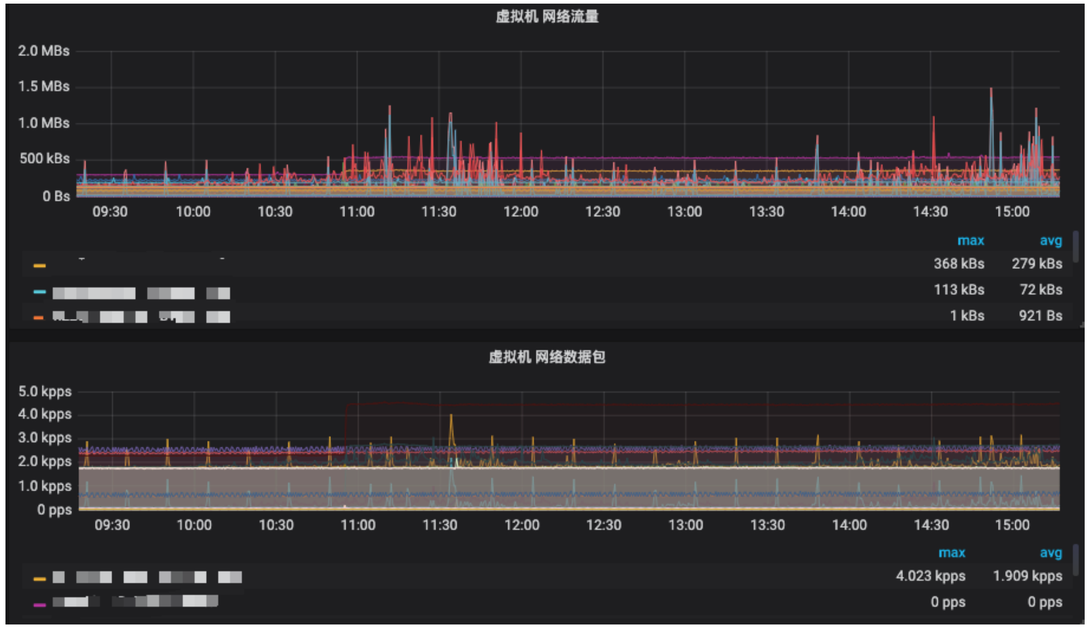


各个网络节点PPS占比

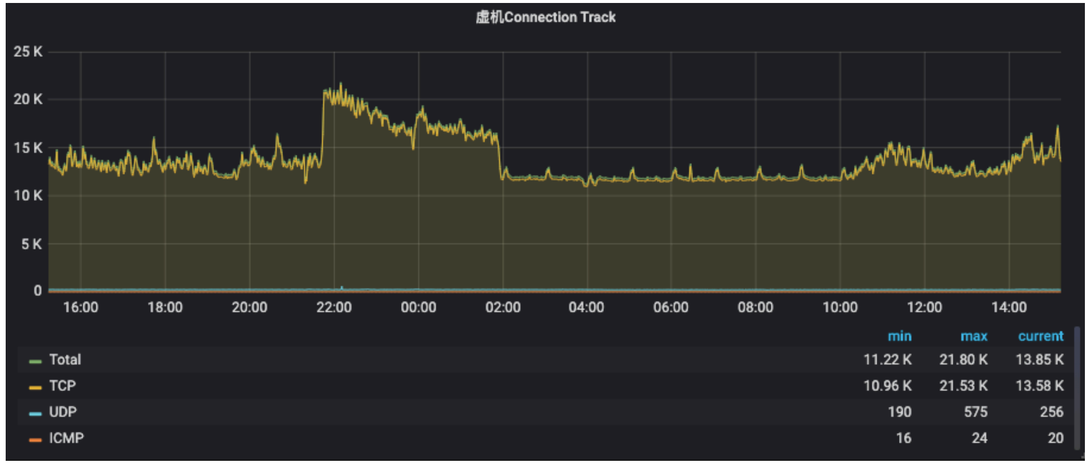
计算节点监控



各个进程CPU使用率监控



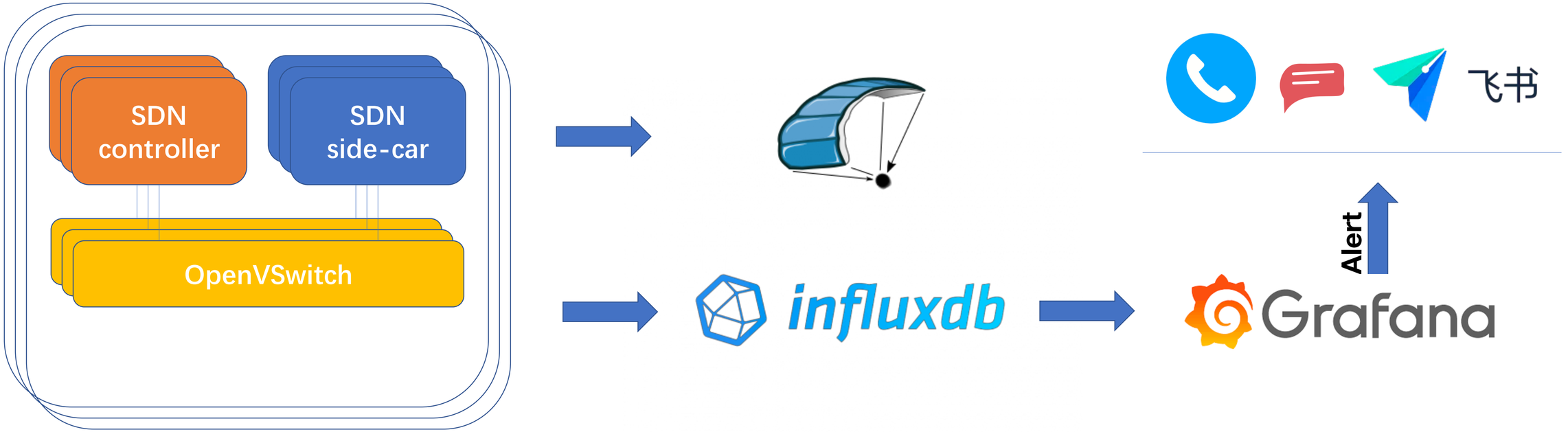
虚机网络流量监控

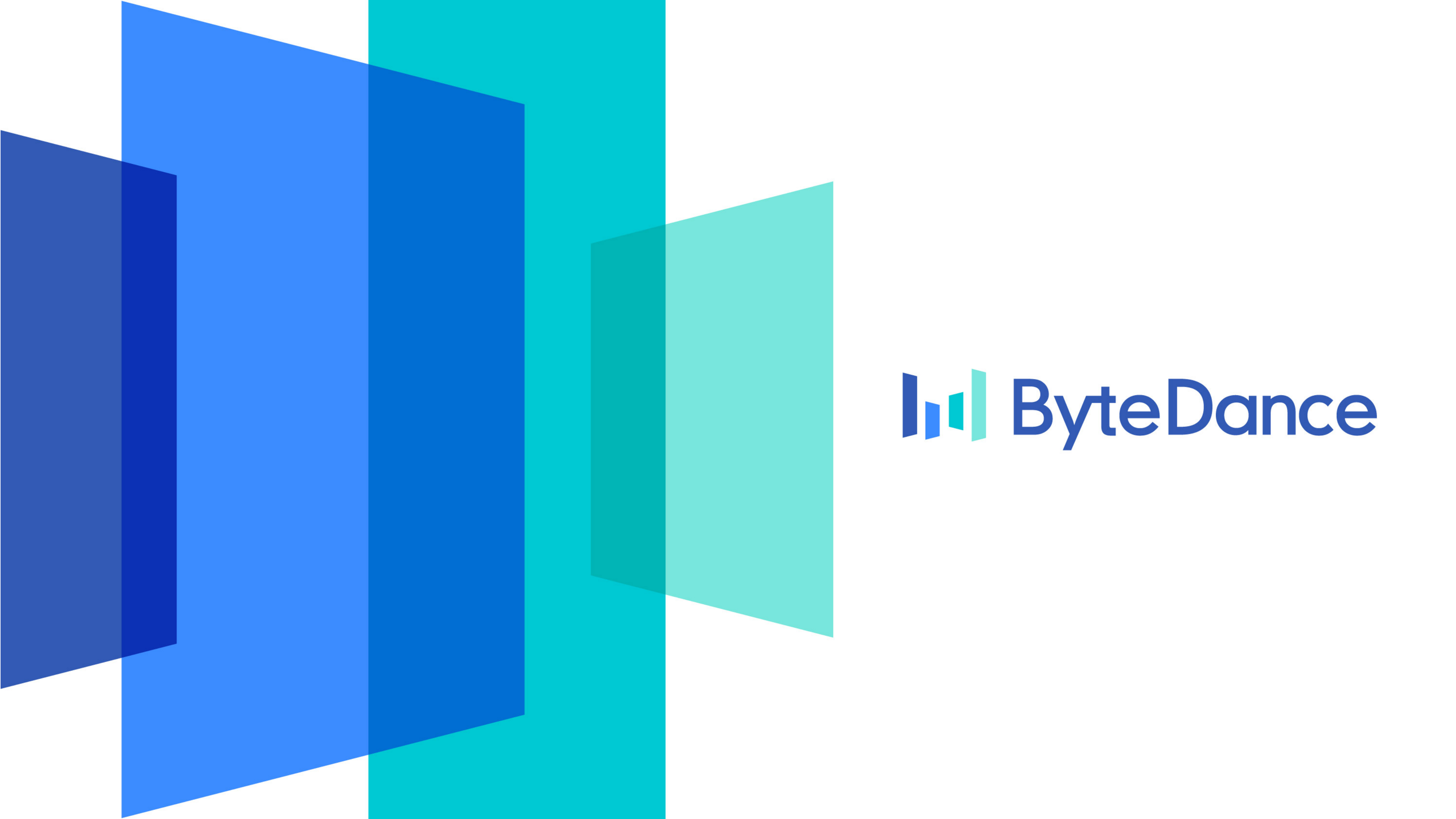


虚机网络连接数监控

OpenVSwitch上送用户态比例监控

监控系统架构




ByteDance