JDT 京东科技

IAAS网络架构分享

孙希望

IAAS产品研发部 虚拟网络组

2022年08月17日

JDT 京东科技

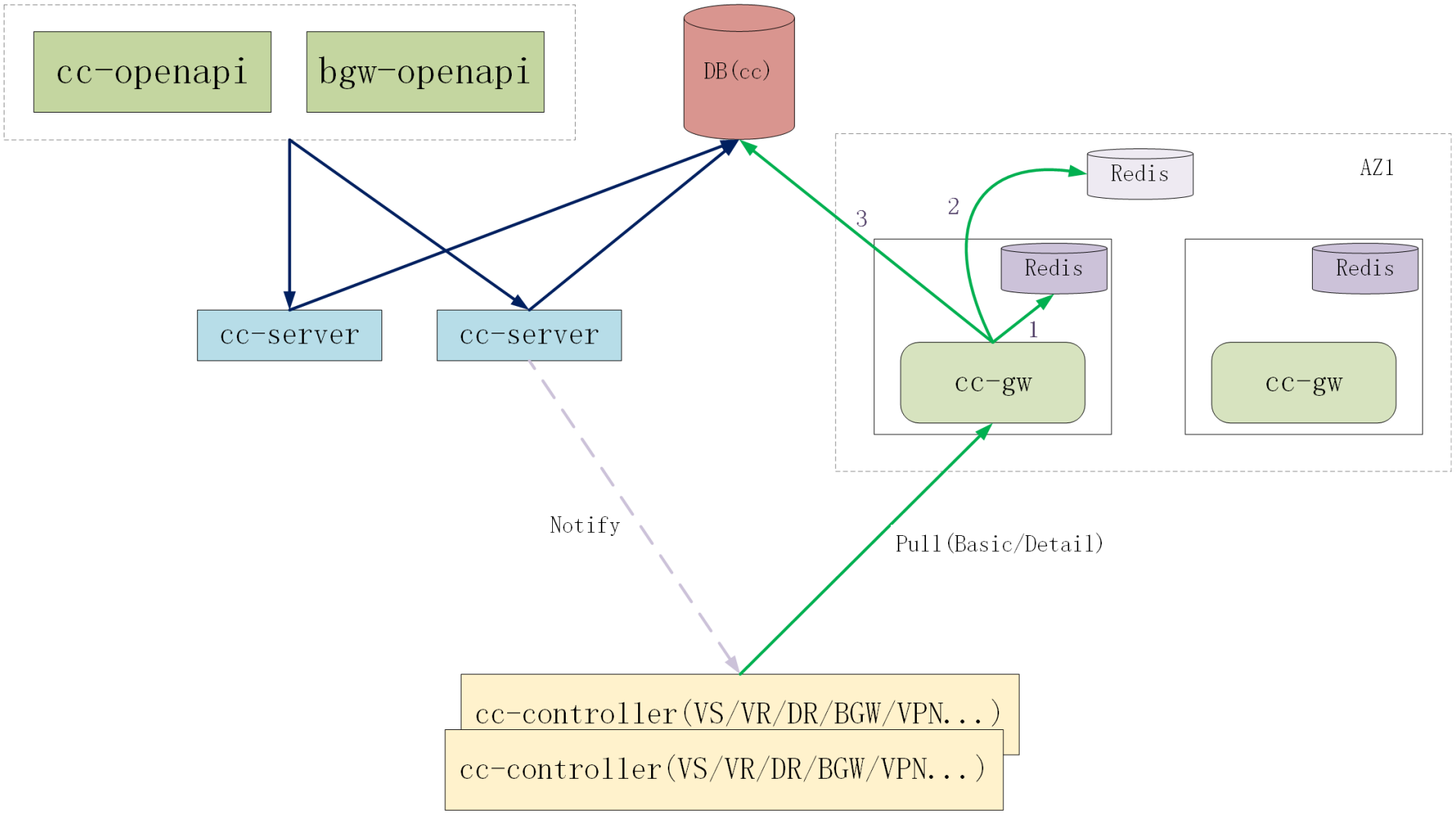
目录

JDT 京东科技

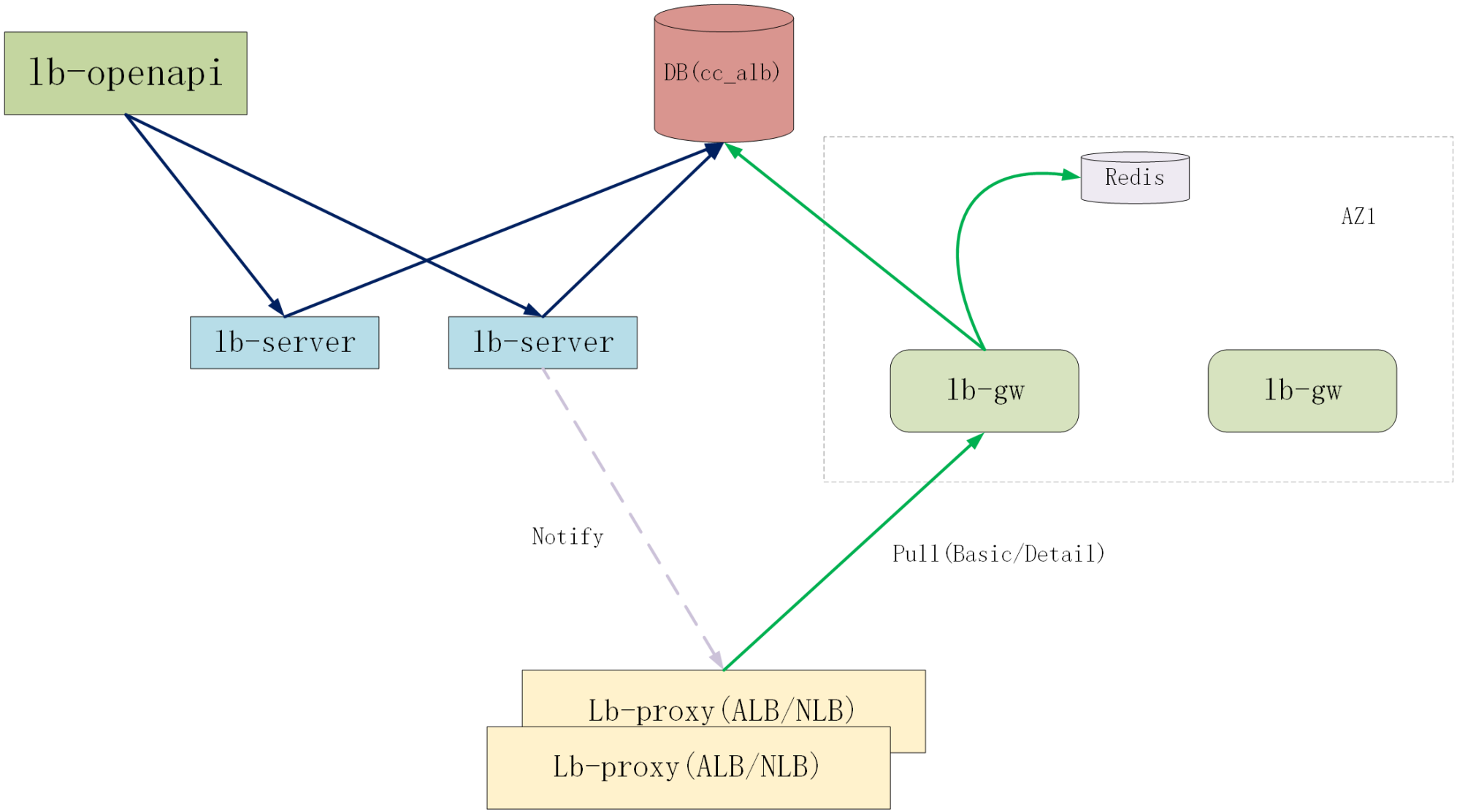
01

网络控制面架构主体逻辑

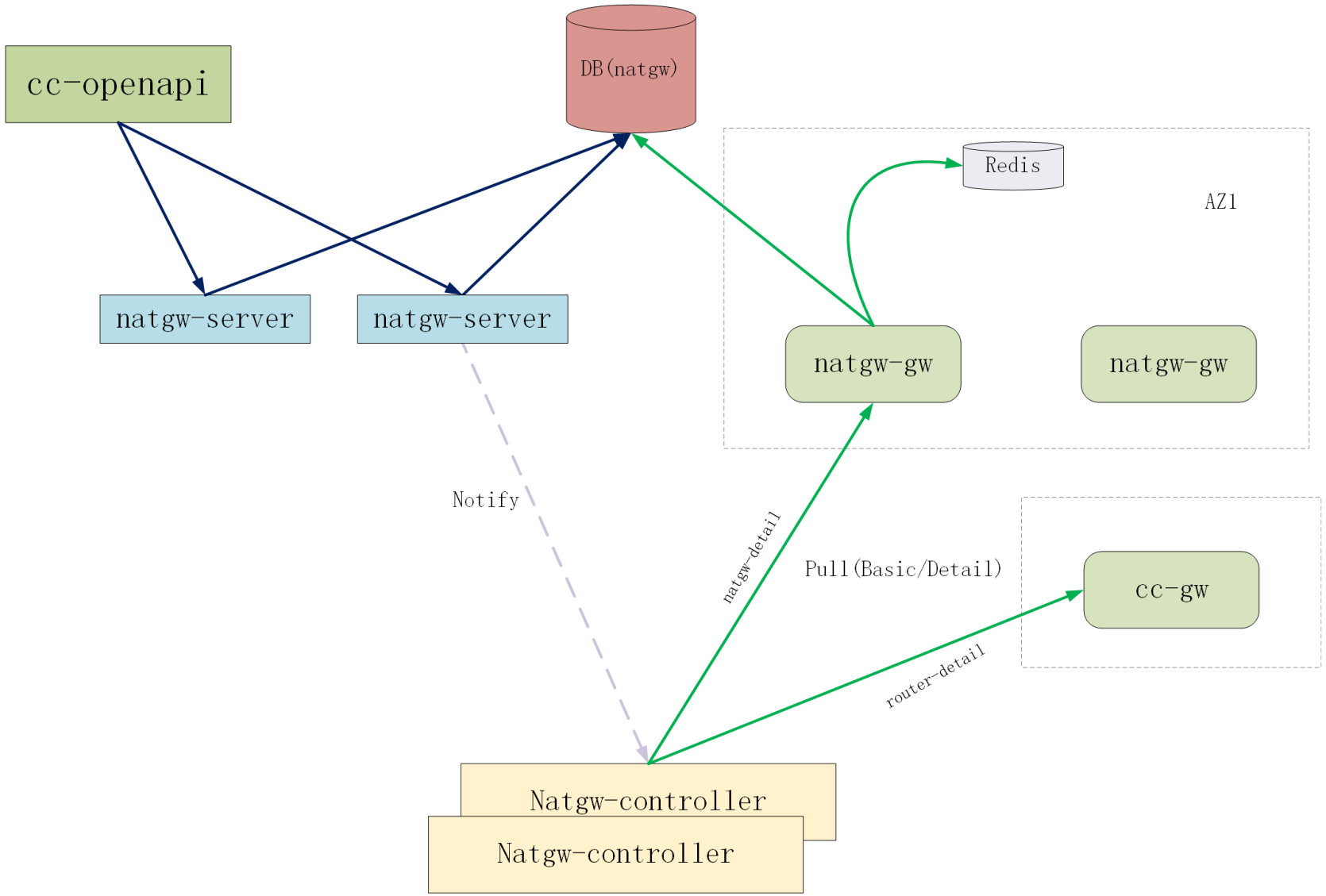
CC侧配置生成以及下发逻辑



LB侧配置生成以及下发逻辑



NATGW侧配置生成以及下发逻辑



JDT 京东科技

02

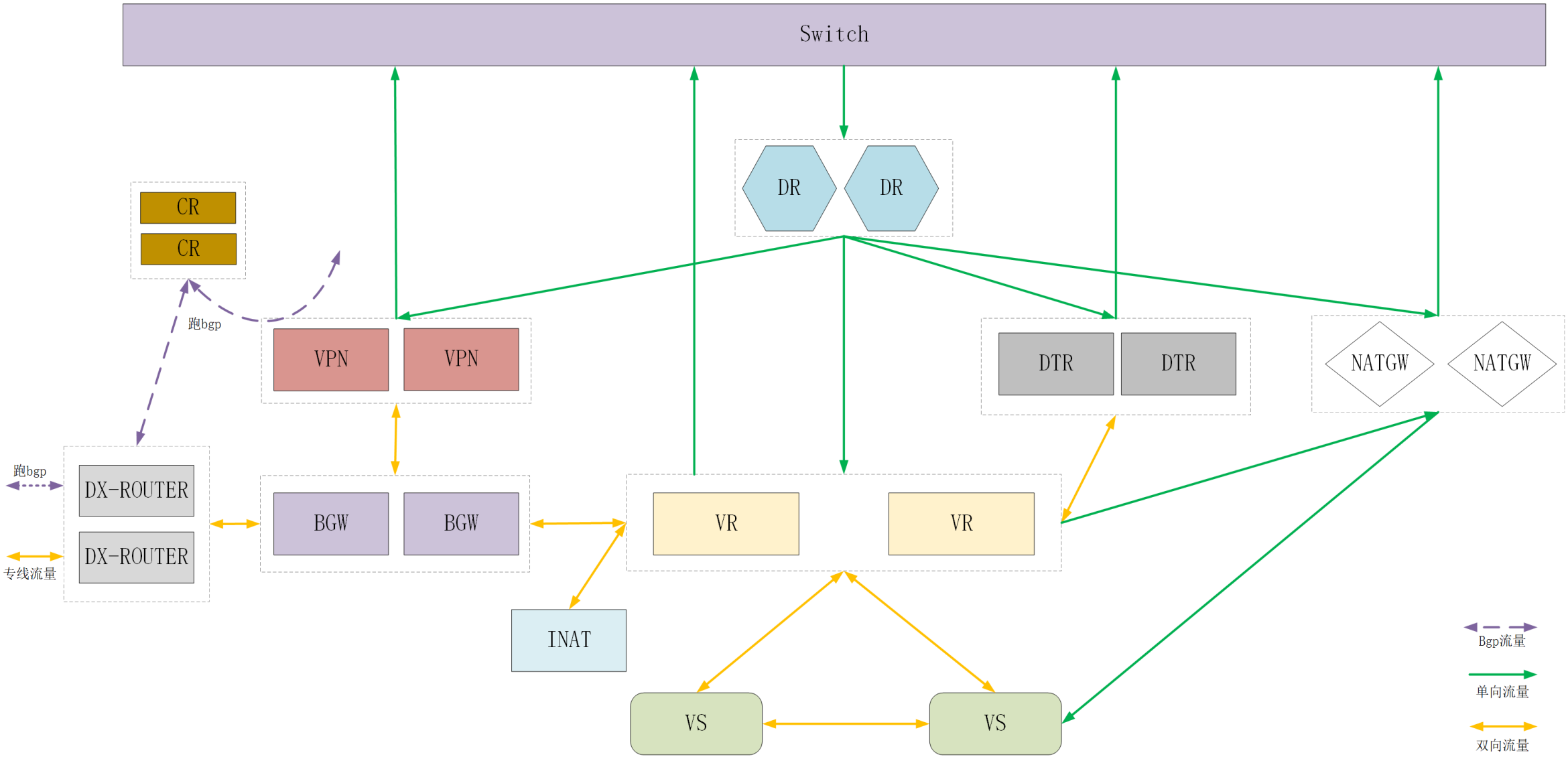
各网络节点功能

各网络节点功能

JDT 京东科技

节点名称	简要介绍	主要功能
VS节点	承载用户虚机、容器等实例的节点	过安全组，过ACL和过BNLB等等
VR节点	用户共享的Router节点，对于一个vpc来说类似一台路由器，可以匹配用户子网路由表规则进行转发流量	匹配子网路由表，进行流量转发；用户访问公网，进行源/目的ip的 1 : 1 Nat
DR节点	云网络的公网入口节点，通过BGP宣告云上公网ip段来从公网引流到京东云，然后再将公网流量导入云内	宣告本DR组上的所有segment（公网ip段）；对引入的公网流量按配置进行分发到各网络节点
NATGW节点	京东云自研的 N:1 Nat服务节点，可提供内网ip到公网ip的 N:1 Nat，支持主备高可用	负责内网到公网的n:1 nat，并将流量发到公网
INAT节点	京东云为用户vpc默认提供访问100段服务的节点，用namespace+iptables来实现的n:1 nat功能	用户vpc可以用其访问dns、yum源等服务
BGW节点	边界网关节点，承担云内和云外之间路由的功能，支撑专线、VPN等功能的实现	在云内和云外之间转发专线、vpn流量
VPN节点	实现VPN功能的网络节点，可与用户自身的vpn节点建立vpn隧道，用户可用vpn来和京东云打通内网	建立vpn隧道，以及转发vpn流量
CR节点	为BGW和Dx-Router之间，京东云VPN和用户VPN之间跑BGP协议的节点，负责发布以及学习路由到边界网关路由表（Bgw-Route）	和Dx-Router、用户VPN跑bgp宣告并学习路由，并上报cc-server
Dx-Router节点	连接用户IDC和京东云专线的云边界路由器，和用户的边界路由器跑bgp互相宣告网段并传播路由	和用户IDC的边界路由器跑BGP学习和传播路由，并根据路由转发专线流量

各网络关系图

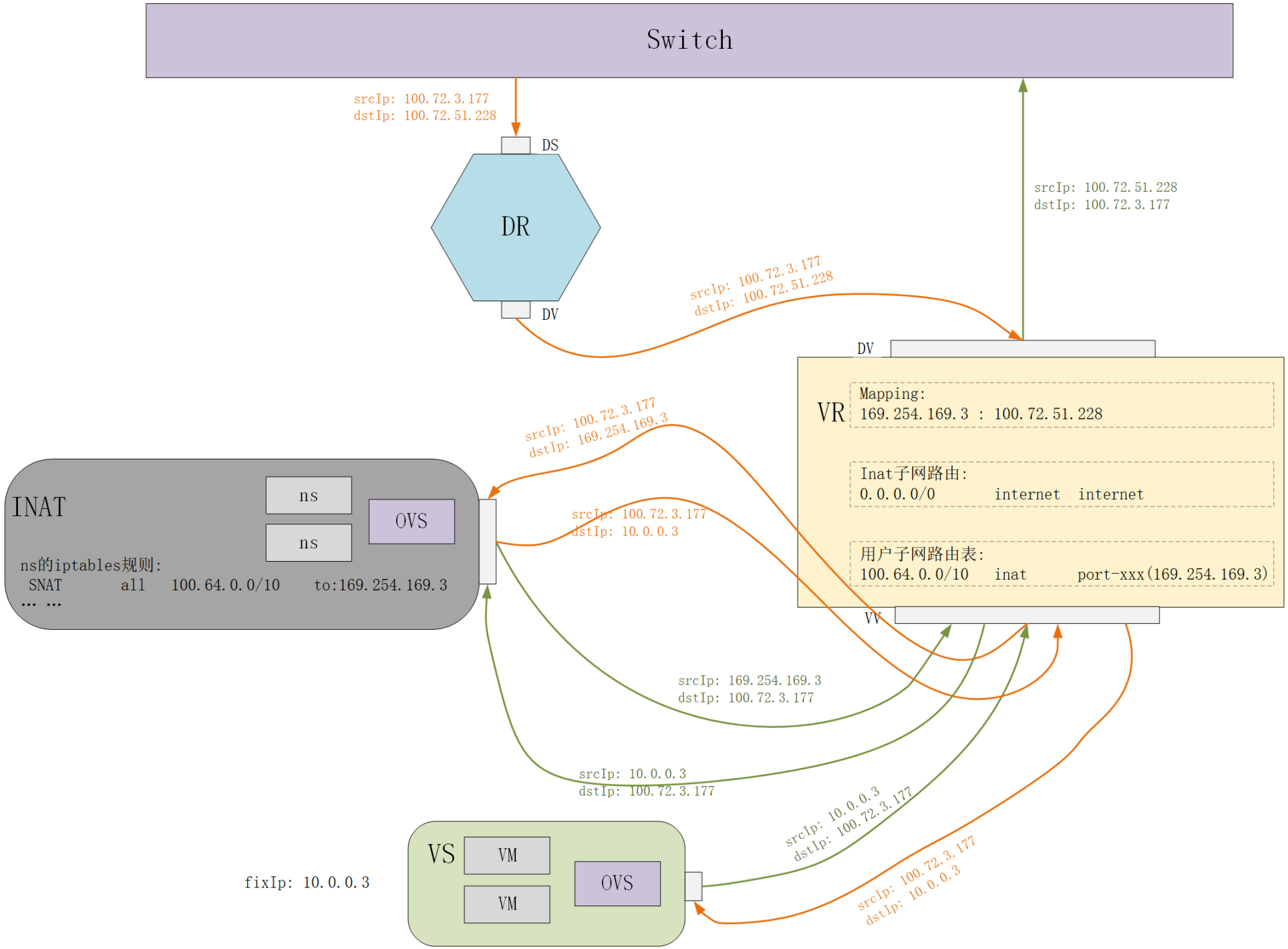

Switch
DR
DR
CR
CR
跑bgp
VPN
VPN
DTR
DTR
NATGW
NATGW
跑bgp
DX-ROUTER
DX-ROUTER
专线流量
BGW
BGW
VR
VR
INAT
VS
VS
Bgp流量
单向流量
双向流量

JDT 京东科技

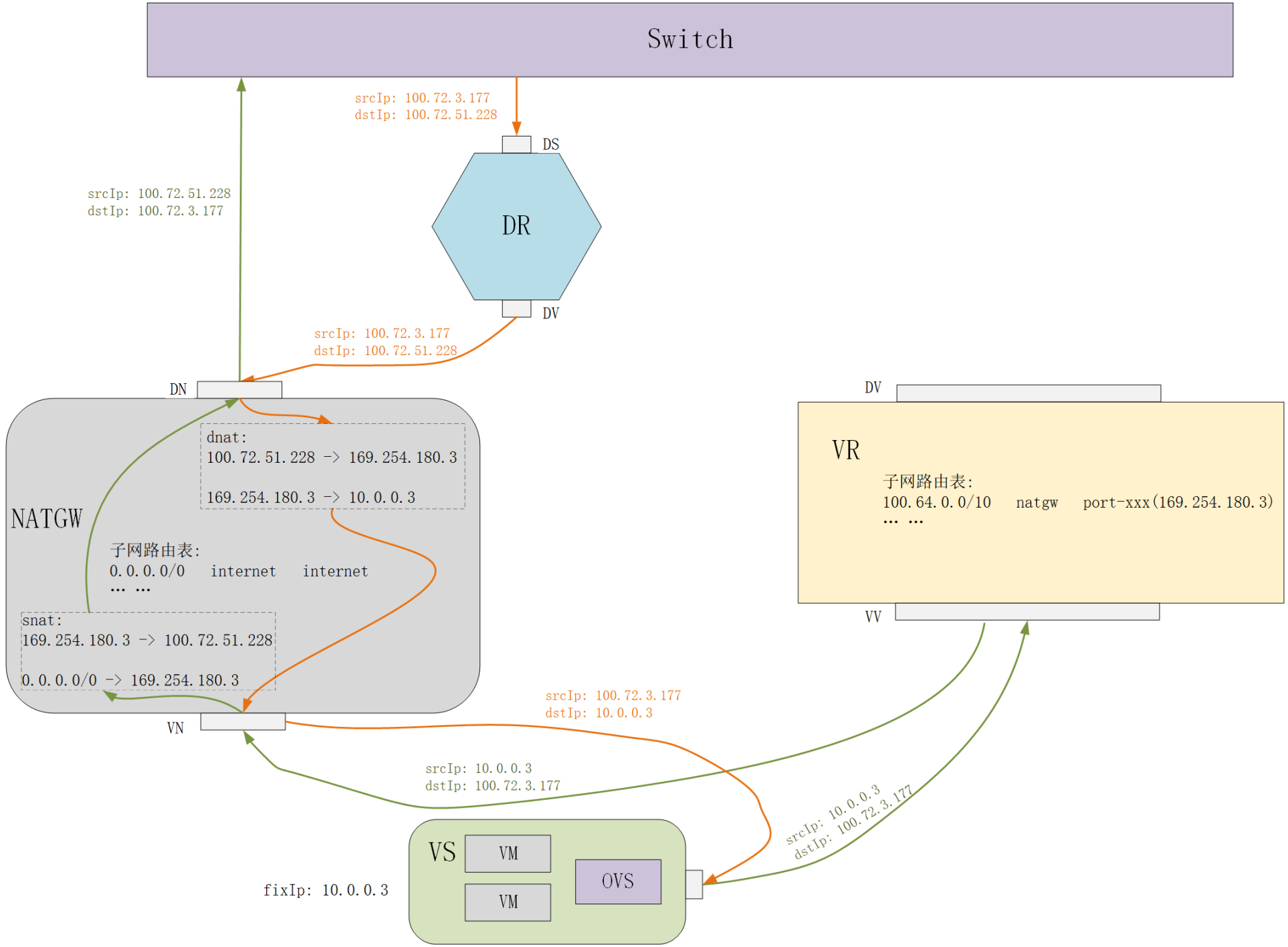
03

典型流量路径

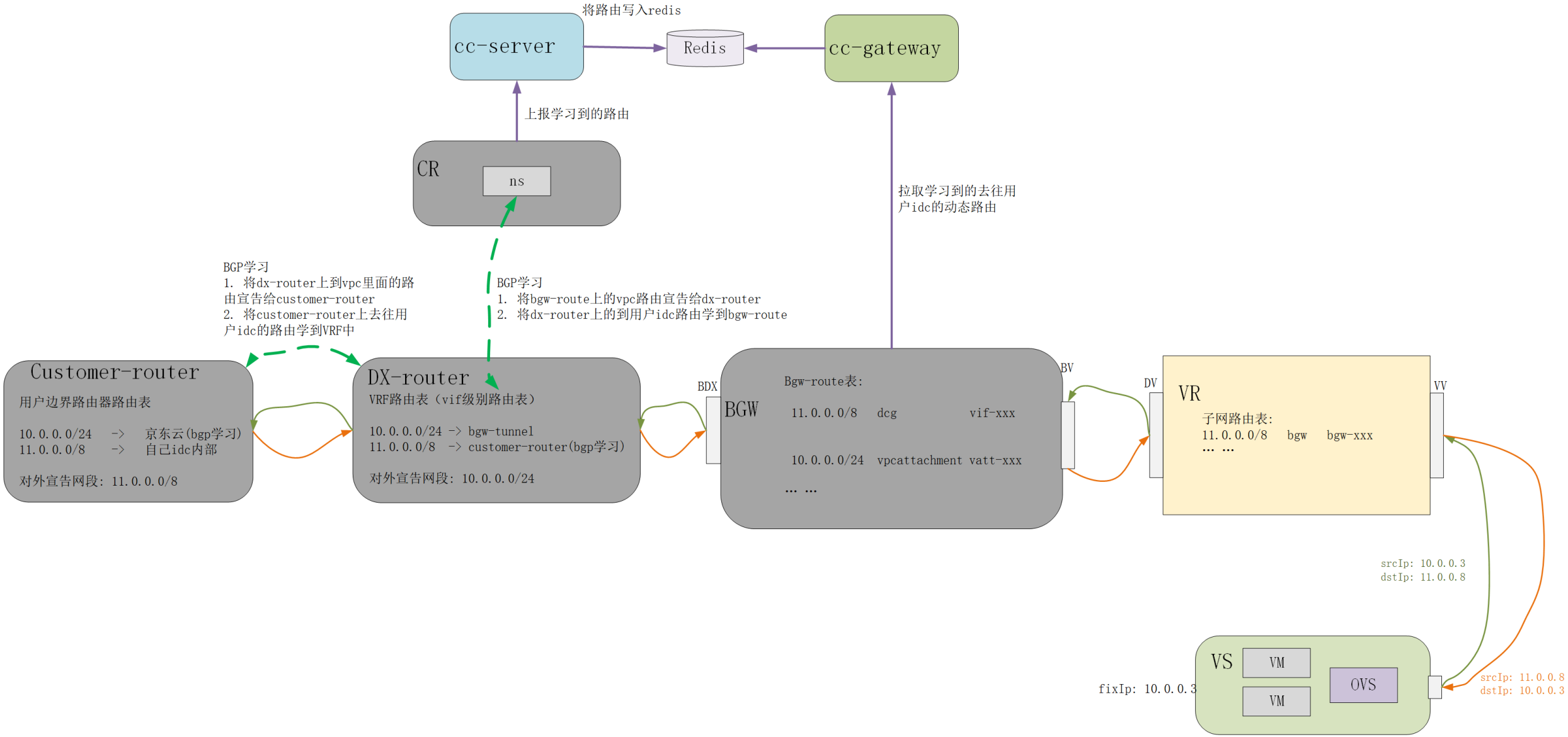
老版INAT流量


Switch
srcIp: 100.72.3.177
dstIp: 100.72.51.228
DS
DR
DV
srcIp: 100.72.51.228
dstIp: 100.72.3.177
srcIp: 100.72.3.177
dstIp: 100.72.51.228
DV
VR
Mapping:
169.254.169.3 : 100.72.51.228
Inat子网路由:
0.0.0.0/0 internet internet
用户子网路由表:
100.64.0.0/10 inat port-xxx(169.254.169.3)
VV
srcIp: 100.72.3.177
dstIp: 169.254.169.3
INAT
ns
ns
OVS
ns的iptables规则:
SNAT all 100.64.0.0/10 to:169.254.169.3
... ...
srcIp: 100.72.3.177
dstIp: 10.0.0.3
srcIp: 169.254.169.3
dstIp: 100.72.3.177
srcIp: 10.0.0.3
dstIp: 100.72.3.177
srcIp: 10.0.0.3
dstIp: 100.72.3.177
srcIp: 100.72.3.177
dstIp: 10.0.0.3
VS
VM
VM
OVS
fixIp: 10.0.0.3

新版INAT流量


Switch
srcIp: 100.72.3.177
dstIp: 100.72.51.228
DS
DR
DV
srcIp: 100.72.51.228
dstIp: 100.72.3.177
srcIp: 100.72.3.177
dstIp: 100.72.51.228
DN
dnat:
100.72.51.228 -> 169.254.180.3
169.254.180.3 -> 10.0.0.3
NATGW
子网路由表:
0.0.0.0/0 internet internet
... ...
snat:
169.254.180.3 -> 100.72.51.228
0.0.0.0/0 -> 169.254.180.3
VN
DV
VR
子网路由表:
100.64.0.0/10 natgw port-xxx(169.254.180.3)
... ...
VV
srcIp: 100.72.3.177
dstIp: 10.0.0.3
srcIp: 10.0.0.3
dstIp: 100.72.3.177
srcIp: 10.0.0.3
dstIp: 100.72.3.177
VS
VM
VM
OVS
fixIp: 10.0.0.3

专线流量

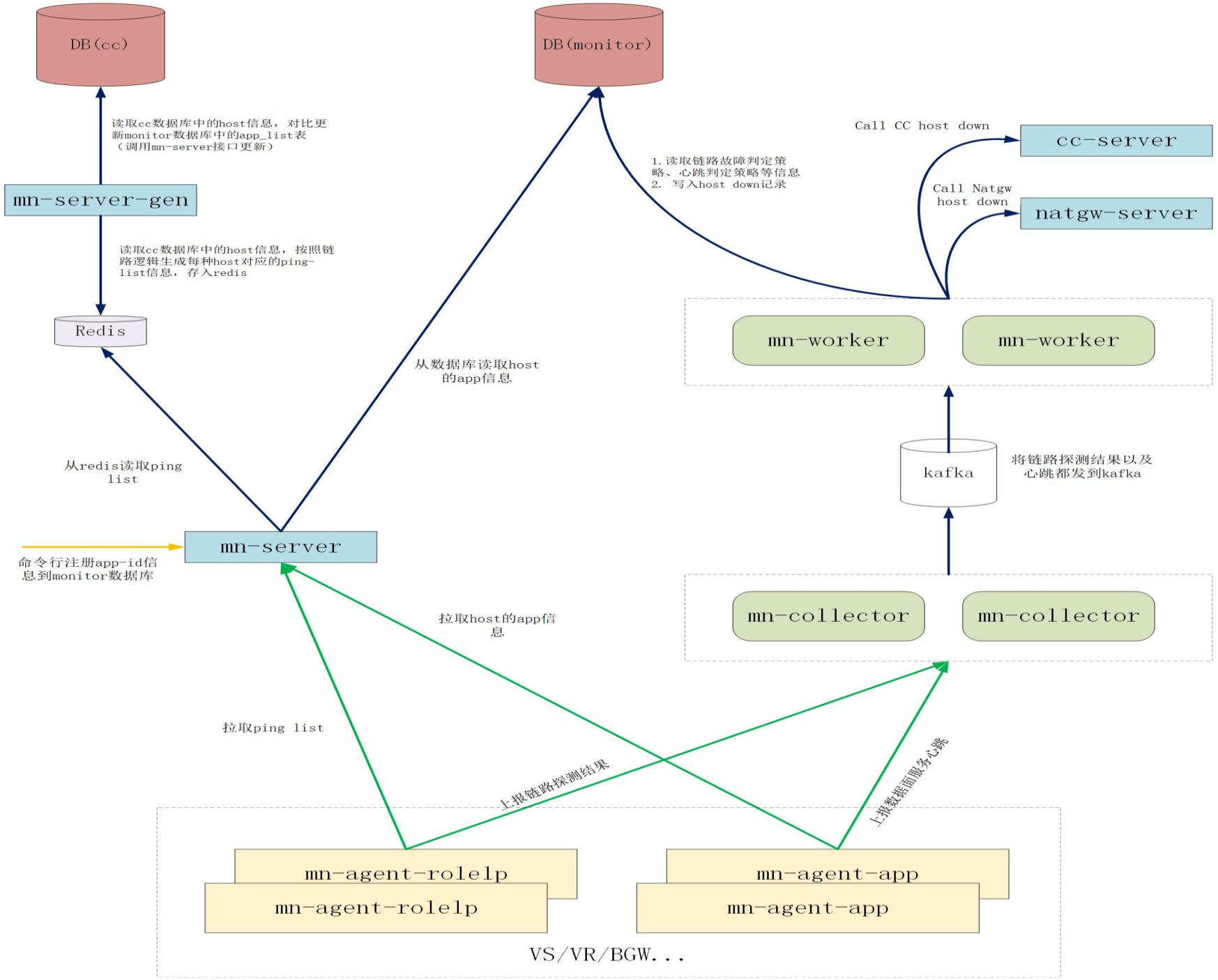

将路由写入redis
cc-server
Redis
cc-gateway
上报学习到的路由
CR
ns
拉取学习到的去往用户idc的动态路由
BGP学习
1. 将dx-router上到vpc里面的路由宣告给customer-router
2. 将customer-router上去往用户idc的路由学到VRF中
BGP学习
1. 将bgw-route上的vpc路由宣告给dx-router
2. 将dx-router上的到用户idc路由学到bgw-route
Customer-router
用户边界路由器路由表
10.0.0.0/24 -> 京东云(bgp学习)
11.0.0.0/8 -> 自己idc内部
对外宣告网段: 11.0.0.0/8
DX-router
VRF路由表（vif级别路由表）
10.0.0.0/24 -> bgw-tunnel
11.0.0.0/8 -> customer-router(bgp学习)
对外宣告网段: 10.0.0.0/24
BDX
BGW
Bgw-route表:
11.0.0.0/8 dcg vif-xxx
10.0.0.0/24 vpcattachment vatt-xxx
... ...
BV
DV
VR
子网路由表:
11.0.0.0/8 bgw bgw-xxx
... ...
VV
srcIp: 10.0.0.3
dstIp: 11.0.0.8
VS
VM
VM
OVS
fixIp: 10.0.0.3
srcIp: 11.0.0.8
dstIp: 10.0.0.3

JDT 京东科技

04

网络节点高可用机制

Monitor服务组件关系图



Monitor高可用机制介绍

JDT 京东科技

心跳摘除机制

1、各网络节点上部署mn-agent-app服务，该服务会监听一个udp端口，负责接收数据面上报上来的心跳

2、收到心跳时会记录一次timestamp，定时将心跳信息上报给mn-collector（上报时会再取一次当前系统时间戳，也就是上报上去的有两个时间）

3、上报到mn-collector的心跳数据，mn-collector会通过kafka发送给mn-worker

4、mn-worker收到心跳数据后，会判断其心跳是否超时（计算两个时间戳的差值是否超时），如果超时则会触发摘除机器（host status down）

链路探测摘除机制

1、各网络节点上部署mn-agent-role-lp服务，该服务会定时拉取最新的ping-list，然后去定时ping这些ip，然后将ping结果上报给mn-collector

2、mn-collector会将各个mn-agent-role-lp的探测结果按照ip做汇聚发到kafka（用ip做hash key），保证同一个ip的探测结果只会被一个mn-worker实例收到

3、mn-worker收到ping探测结果数据后，会判断ping失败率是否达到策略阈值，如果到达阈值则会触发摘除机器（host status down）

机器故障了又没有自动摘除的几种情况

JDT 京东科技

1、没有部署mn-agent

2、mn-agent拉不到正确的ping-list

3、kafka有问题，探测结果数据送不到决策节点mn-worker

4、没有设置down机器的策略

<1> 使用mns role-lp-policy-list 命令查看是否由role-lp探测置down的策略是否存在

<2> 使用mns app-alive-policy-list 命令查看是否有app心跳超时置down的策略是否存在

如果没有策略是不会进行host置down的

5、mn-worker收到的探测结果票数不够，导致不进行决策

mn-worker有个配置项：Judge.LeastVoteRateToJudge // 需要至少百分之多少的投票者才能开始决策

5、同一时间窗口内down太多机器触发熔断

mn-worker有两个配置项：

DownHost.Interval // 时间窗口，单位：秒

DownHost.DownThreshold // 时间窗口内down掉机器个数的阈值

down掉的机器存储在monitor数据库的 down_host 表中

6、cc或者natgw那边进行了熔断，例如VR是一个hg在每个AZ至少剩1台

JDT 京东科技

05

常用排障手段以及命令

JDT 京东科技

Thanks