

【爱卡汽车】回环问题排查

先说结论：

1.北京tke集群

- 1)cls-1buzjf75 使用 “externalTrafficPolicy=local ”有回环问题，并且体现出回环问题
- 2)cls-27hke42h 使用 “externalTrafficPolicy=cluster ”无回环问题
- 3)cls-37au35kt 使用 “externalTrafficPolicy=local ”有回环问题，但没有体现出回环问题，原因是 client pod 和 nginx-ingress 不在同一个节点上

2.广州tke集群:

- 1)cls-cq4e4na6 使用的是公网clb，无回环问题，不通的原因已排查清楚，是因为安全组设置问题

排查问题原理：

如果是内网clb(北京三个集群)：

回环触发关键点：PodA (client) -> CLB(内网) -> 节点nodeport (nginx-ingress nodeport,且节点是 PodA 所在节点)

为啥使用 “externalTrafficPolicy=local ”有回环问题，而使用“externalTrafficPolicy=cluster ”无回环问题？

原因：

iptables 差异：

配置了“externalTrafficPolicy=local ”此配置后自动插入的 iptables 规则(NAT 表)：

```
COMMIT
[lokeyli@M-46-151-centos ~]$ diff -u q1 q2
+++ q2 2021-01-12 10:49:51.426149501 +0800
@@ -1,10 +1,11 @@
+nat
-:PREROUTING ACCEPT [2686:146776]
+:INPUT ACCEPT [1:28]
-:OUTPUT ACCEPT [156:3360]
+:PREROUTING ACCEPT [2148:121546]
+:PREROUTING ACCEPT [795:44760]
+:INPUT ACCEPT [4:176]
+:OUTPUT ACCEPT [15:900]
+:POSTROUTING ACCEPT [847:49384]
+:IP-MASQ-AGENT - [0:0]
+:KUBE-FIREWALL - [0:0]
+:KUBE-KUBELET-CANARY - [0:0]
+:KUBE-LOAD-BALANCER - [0:0]
+:KUBE-MARK-DROP - [0:0]
+:KUBE-MARK-MASQ - [0:0]
@@ -15,13 +16,14 @@
-A OUTPUT -m comment --comment "Kubernetes service portals" -j KUBE-SERVICES
-A POSTROUTING -m comment --comment "Kubernetes postrouting rules" -j KUBE-POSTROUTING
-A IP-MASQ-AGENT -d 172.17.0.0/16 -m comment --comment "ip-masq-agent: ensure nat POSTROUTING directs all non-LOCAL destination traffic to our custom IP-MASQ-AGENT chain" -m addrtype ! --dst-type LOCAL -j IP-MASQ-AGENT
-A IP-MASQ-AGENT -d 172.17.0.0/16 -m comment --comment "ip-masq-agent: cluster-local traffic should not be subject to MASQUERADE" -m addrtype ! --dst-type LOCAL -j RETURN
-A IP-MASQ-AGENT -d 172.18.0.0/16 -m comment --comment "ip-masq-agent: cluster-local traffic should not be subject to MASQUERADE" -m addrtype ! --dst-type LOCAL -j RETURN
-A IP-MASQ-AGENT -m comment --comment "ip-masq-agent: outbound traffic should be subject to MASQUERADE (this match must come after cluster-local CIDR matches)" -m addrtype ! --dst-type LOCAL -j MASQUERADE
-A KUBE-FIREWALL -j KUBE-MARK-DROP
+A KUBE-LOAD-BALANCER -m comment --comment "Kubernetes service load balancer ip + port with externalTrafficPolicy=local" -m set --match-set KUBE-LOAD-BALANCER-LOCAL dst,dst -j RETURN
-A KUBE-LOOP-BACK -j KUBE-MARK-MASQ
-A KUBE-MARK-DROP -j MARK --set-xmark 0x8000/0x8000
-A KUBE-MARK-MASQ -j MARK --set-xmark 0x4000/0x4000
+A KUBE-NODE-PORT -p tcp -m comment --comment "Kubernetes nodeport TCP port with externalTrafficPolicy=local" -m set --match-set KUBE-NODE-PORT-LOCAL-TCP dst -j RETURN
-A KUBE-NODE-PORT -p tcp -m comment --comment "Kubernetes nodeport TCP port for masquerade purpose" -m set --match-set KUBE-NODE-PORT-LOCAL-TCP dst -j KUBE-MARK-MASQ
-A KUBE-POSTROUTING -m comment --comment "Kubernetes service traffic requiring SNAT" -m mark --mark 0x4000/0x4000 -j MASQUERADE
-A KUBE-POSTROUTING -m comment --comment "Kubernetes endpoints dst ip:port, source ip for solving hairpin purpose" -m set --match-set KUBE-LOOP-BACK dst,dst,src -j MASQUERADE
```

做了 snat 的集群

没做 snat 的集群

多出的两条规则

对比不同配置的规则发现，设置了 “externalTrafficPolicy=local ”的节点会多出两个“externalTrafficPolicy local”的iptables 规则。

iptables 的差异如何影响 出节点 包的转发：

从iptables 规则可以看出，当要出节点时，都要先命中如下 OUTPUT chain，

-A OUTPUT -m comment --comment "kubernetes service portals" -j KUBE-SERVICES

然后一路找转发规则就好了：

默认 “externalTrafficPolicy=cluster” 的节点，做了SNAT 的情况：

```
[root@tx-ms-k8s-node-32160 172.18.32.160 ~]# iptables-save | grep KUBE-SERVICES
:KUBE-SERVICES - [0:0]
-A PREROUTING -m comment --comment "kubernetes service portals" -j KUBE-SERVICES
-A OUTPUT -m comment --comment "kubernetes service portals" -j KUBE-SERVICES
-A KUBE-SERVICES -m comment --comment "Kubernetes service lb portal" -m set --match-set KUBE-LOAD-BALANCER dst,dst -j KUBE-LOAD-BALANCER
-A KUBE-SERVICES -m comment --comment "Kubernetes service cluster ip + port for masquerade purpose" -m set --match-set KUBE-CLUSTER-IP src,dst -j KUBE-MARK-MASQ
-A KUBE-SERVICES -m addtype --dst-type LOCAL -j KUBE-NODE-PORT
-A KUBE-SERVICES -m set --match-set KUBE-LOAD-BALANCER dst,dst -j ACCEPT
-A KUBE-SERVICES -m set --match-set KUBE-LOAD-BALANCER dst,dst -j ACCEPT
[root@tx-ms-k8s-node-32160 172.18.32.160 ~]# ipset list KUBE-LOAD-BALANCER | grep 172.18.33.49
172.18.33.49,tcp:443
172.18.33.49,tcp:80
[root@tx-ms-k8s-node-32160 172.18.32.160 ~]# iptables-save | grep KUBE-LOAD-BALANCER
:KUBE-LOAD-BALANCER - [0:0]
-A KUBE-LOAD-BALANCER -j KUBE-MARK-MASQ
-A KUBE-SERVICES -m comment --comment "Kubernetes service lb portal" -m set --match-set KUBE-LOAD-BALANCER dst,dst -j KUBE-LOAD-BALANCER
-A KUBE-SERVICES -m set --match-set KUBE-LOAD-BALANCER dst,dst -j ACCEPT
[root@tx-ms-k8s-node-32160 172.18.32.160 ~]
```

1

因为lb ip 在 set KUBE-LOAD-BALANCER，所以会匹配 上面 1 规则

最终到达 KUBE-MARK-MASQ 规则，出节点会做 SNAT

设置了 “externalTrafficPolicy=local” 的节点，不做SNAT 的情况：

```
[root@VM-96-12-centos ~]# iptables-save | grep KUBE-SERVICES
:KUBE-SERVICES - [0:0]
-A PREROUTING -m comment --comment "kubernetes service portals" -j KUBE-SERVICES
-A OUTPUT -m comment --comment "kubernetes service portals" -j KUBE-SERVICES
-A KUBE-SERVICES -m comment --comment "Kubernetes service lb portal" -m set --match-set KUBE-LOAD-BALANCER dst,dst -j KUBE-LOAD-BALANCER
-A KUBE-SERVICES -m comment --comment "Kubernetes service cluster ip + port for masquerade purpose" -m set --match-set KUBE-CLUSTER-IP src,dst -j KUBE-MARK-MASQ
-A KUBE-SERVICES -m addtype --dst-type LOCAL -j KUBE-NODE-PORT
-A KUBE-SERVICES -m set --match-set KUBE-LOAD-BALANCER dst,dst -j ACCEPT
-A KUBE-SERVICES -m set --match-set KUBE-LOAD-BALANCER dst,dst -j ACCEPT
[root@VM-96-12-centos ~]# ipset list KUBE-LOAD-BALANCER
Name: KUBE-LOAD-BALANCER
Type: hash:ip,port
Revision: 5
Header: family inet hashsize 1024 maxelem 65536
Size in memory: 344
References: 2
Number of entries: 4
Members:
172.18.96.4,tcp:443
120.53.128.6,tcp:443
172.18.96.6,tcp:443
172.18.96.4,tcp:80
[root@VM-96-12-centos ~]# iptables-save | grep KUBE-LOAD-BALANCER
:KUBE-LOAD-BALANCER - [0:0]
-A KUBE-LOAD-BALANCER -m comment --comment "Kubernetes service load balancer ip + port with externalTrafficPolicy=local" -m set --match-set KUBE-LOAD-BALANCER-LOCAL dst,dst -j RETURN
-A KUBE-LOAD-BALANCER -m comment --comment "Kubernetes service lb portal" -m set --match-set KUBE-LOAD-BALANCER dst,dst -j KUBE-LOAD-BALANCER
-A KUBE-SERVICES -m comment --comment "Kubernetes service lb portal" -m set --match-set KUBE-LOAD-BALANCER dst,dst -j KUBE-LOAD-BALANCER
-A KUBE-SERVICES -m set --match-set KUBE-LOAD-BALANCER dst,dst -j ACCEPT
[root@VM-96-12-centos ~]# ipset list KUBE-LOAD-BALANCER-LOCAL
Name: KUBE-LOAD-BALANCER-LOCAL
Type: hash:ip,port
Revision: 5
Header: family inet hashsize 1024 maxelem 65536
Size in memory: 216
References: 1
Number of entries: 2
Members:
172.18.96.4,tcp:443
172.18.96.4,tcp:80
```

1

LB ip 在上面 1 规则的 set 中，所以匹配 1

重点，由于 多出的 LOCAL 策略放在了 KUBE-MARK-MASQ 之前

并且 LB 的 ip 在 set KUBE-LOAD-BALANCER 中，所以匹配多出的 LOCAL 策略

如果是外网clb(广州tke集群)：

本身么有回环问题，原因见材料：

[CLB 回环问题总结.pdf](#)

描述截图：

为什么公网 CLB 没这个问题？

使用公网 Ingress 和 LoadBalancer 类型公网 Service 没有回环问题，我的理解主要是公网 CLB 收到的报文源 IP 是子机的出口公网 IP，而子机内部感知不到自己的公网 IP，当报文转发回子机时，不认为公网源 IP 是本机 IP，也就不存在回环。

