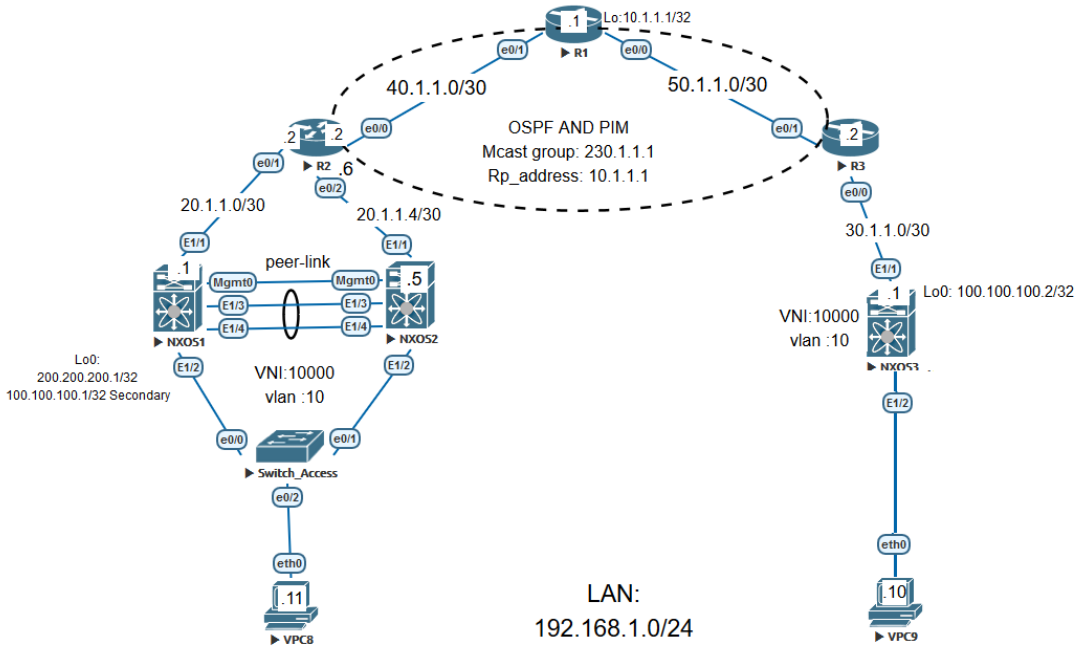
Lab – VxLAN trên cluster Nexus VPC.

1. Sơ đồ



Bài Lab này được dựng trên lab ảo hoá sử dụng các qemu switch nexus nxosv9k-9.3.1, các IOL Router L3-ADVENTERPRISEK9-M-15.4-2T và các vPC.

Khi triển khai VxLAN trên một cluster vPC, chúng ta cần quan tâm đến một số lưu ý sau:

- Dùng một địa chỉ loopback cho NVE. Địa chỉ này là riêng biệt với các địa chỉ loopback đang được sử dụng cho các dịch vụ khác, ví dụ loopback cho OSPF, BGP.
- Tăng thời gian khôi phục của vPC (delay restore interface vlan) nếu số lượng các interface vlan SVI là nhiều. Ví dụ nếu chúng ta có 1000 VNI với 1000 SVI thì thông số thời gian có thể tăng lên 45 giây.
- Thiết lập cổng loopback cho NVE dùng hai địa chỉ. Địa chỉ IP chính (Primary IP Address) và địa chỉ IP phụ (Secondary IP Address). Địa chỉ IP thứ hai sẽ sử dụng cho tất cả lưu lượng VxLAN bao gồm cả Multicast và Unicast.
- Các VPC Peer cần phải được cấu hình giống nhau: VLAN phải được ánh xạ (mapping) thông qua VN-Segment. NVE1 phải được liên kết đến cổng Loopback giống nhau. Sử dụng cùng địa chỉ IP phụ (Secondary IP Address).Sử dụng khác địa chỉ IP chính (Primary IP Address). Chắc chắn rằng VNI phải được ánh xạ (Mapping) đến Group. Đây là các yêu cầu để vPC đảm bảo tính cấu hình nhất quán của vPC cluster.
- Đối với lưu lượng Multicast, các nút VPC nhận được (S,G) khi tham gia vào RP (Rendezvous point) sẽ trở thành DF (Designated Forwarder).

- Đối với thiết bị sử dụng VPC, lưu lượng BUM (Broadcast, Unknow-unicast và Multicast) từ host sẽ được truyền trên kết nối Peer-Link của vPC.
- Đối với VPC, khi cổng Loopback có 2 địa chỉ IP bao gồm địa chỉ IP chính (Primary IP Address) và địa chỉ IP phụ (Secondary IP Address).-Địa chỉ IP chính là địa chỉ duy nhất và được sử dụng bởi giao thức lớp 3. Địa chỉ IP phụ trên Loopback là cần thiết vì cổng NVE sẽ sử dụng cho địa chỉ IP VTEP. Địa chỉ IP phụ phải giống nhau trên cả 2 vPC Peers.
- vPC Peer-gateway feature phải được bật (Enable) trên cả 2 Peers (Vì đây là Switch Nexus nên cần phải sử dụng Feature để bật các tính năng mở rộng trên Switch).
- Cần sử dụng Peer-switch, Peer Gateway, IP Arp Sync, cấu hình Ipv6 Nd Sync cho việc cải thiện tốc độ hội tụ trong mô hình VPC.
- Khi NVE hay Loopback shut trong VPC configurations: + Nếu NVE hay Loopback chỉ shut trên VPC switch đầu tiên (Primary VPC switch), thì toàn bộ VxLAN VPC khi kiểm tra tính nhất quán sẽ là **fails**. Sau đó NVE, Loopback và VPCs bị down trên VPC switch phụ.
- Nếu NVE hay Loopback chỉ Shut trên VPC switch phụ, thì toàn bộ VxLAN VPC kiểm tra tính nhất quán cũng sẽ là **Fails**. Sau đó NVE, Loopback, và Switch phụ sẽ được chuyển sang down trên secondary. Lưu lượng sẽ tiếp tục chuyển luồn thông qua VPC switch chính. Đối với cách đạt được kết quả tốt nhất, chúng ta cần giữ cả 2 NVE và Loopback phải **UP** trên cả VPC switch chính và phụ.
- Cấu hình Anycast RPs dự phòng trong Network cho Multicast Load-Balancing và RP dự phòng phải được hỗ trợ trong mô hình VPC VTEP.

2. Thực hiện Lab

Phần 1. Cấu hình Virtual Port Channel (vPC).

Sử dụng dây kết nối giữa e1/2 của NX01 và e1/2 của NX02 để gửi các gói keep alive duy trì kết nối VPC.

```
NX01(config)#int mgmt 0
NX01(config-if)#vrf member management
NX01(config-if)#ip add 172.16.12.1/24
```

```
NX02(config)#int mgmt 0
NX02(config-if)#vrf member management
NX02(config-if)#ip address 172.16.12.2/24
```

Tiến hành bật tính năng vpc và lacp trên các switch nexus.

```
NX01(config)#feature vpc
NX01(config)#feature lacp
```

```
NX02(config)#feature vpc
NX02(config)#feature lacp
```

Cấu hình để hai cổng e1/3 và e1/4 trên NX01 và NX02 tham gia vào channel 100 tạo thành một bó etherchannel 100 và cặp dây này sẽ là đường peer link (**peer-link** được sử dụng để đồng bộ hóa trạng thái giữa 2 thiết bị vPC thông qua các gói điều khiển vPC).

```
NX01(config)#int e1/3-4
NX01(config-if-range)#switchport mode trunk
NX01(config-if-range)#channel-group 100 mode active
NX01(config-if-range)#no shut
```

```
NX02(config)#int e1/3-4
NX02(config-if-range)#switchport mode trunk
NX02(config-if-range)#channel-group 100 mode active
NX02(config-if-range)#no shut
```

Tạo một vpc domain với ID=100, trong domain này chúng ta cần cấu hình các thông tin như peer device (trong lab này là **peer-switch**), đường link keep alive (chỉ định 2 ip cổng mgmt 0 của NX01 và NX02), thời gian reload nếu có 1 đường bị down (delay restore <1-65535>)…

```
NX01(config)#vpc domain 100
NX01(config-vpc-domain)#peer-switch
NX01(config-vpc-domain)#peer-keepalive destination 172.16.12.2 source
172.16.12.1 vrf management
NX01(config-vpc-domain)#peer-gateway
NX01(config-vpc-domain)#layer3 peer-router
NX01(config-vpc-domain)#ip arp synchronize
NX01(config-vpc-domain)#delay restore 5
```

```
NX02(config)#vpc domain 100
NX02(config-vpc-domain)#peer-switch
NX02(config-vpc-domain)#peer-keepalive destination 172.16.12.1 source
172.16.12.2 vrf management
NX02(config-vpc-domain)#peer-gateway
NX02(config-vpc-domain)#layer3 peer-router
NX02(config-vpc-domain)#ip arp synchronize
NX02(config-vpc-domain)#delay restore 5
```

Trên interface portchannel 100 chúng ta thực hiện cấu hình trunk để trao đổi thông tin giữa các vlan và chỉ định đây là cặp dây peer link.

```
NX01(config)#int port-channel 100
NX01(config-if)#switchport
NX01(config-if)#switchport mode trunk
NX01(config-if)#spanning-tree port type network  //Lệnh này để active tính
năng bridge assurance
NX01(config-if)#vpc peer-link
```

```
NX02(config)#int port-channel 100
NX02(config-if)#switchport
NX02(config-if)#switchport mode trunk
NX02(config-if)#spanning-tree port type network
```

```
NX02(config-if)#vpc peer-link
```

Tạo vlan 10 và access port e1/2 trên cả 2 switch Nexus vào vlan 10. Cũng như cấu hình gom 2 port e1/2 của NX01 và e1/2 của NX02 vào 1 channel group có id là 10 (đây chính là điểm đặc trưng của virtual port channel).

```
NX01(config)#vlan 10
NX01(config)#int e1/2
NX01(config-if)#switchport
NX01(config-if)#switchport access vlan 10
NX01(config-if)#channel-group 10 mode active
NX01(config-if)#no shut
```

```
NX02(config)#vlan 10
NX02(config)#int e1/2
NX02(config-if)#switchport
NX02(config-if)#switchport access vlan 10
NX02(config-if)#channel-group 10 mode active
NX02(config-if)#no shut
```

Trên interface port channel 10, chúng ta tiến hành gán port-channel 10 này vào vpc 10 và access vlan 10.

```
NX01(config)#int port-channel 10
NX01(config-if)#vpc 10
NX01(config-if)#switchport mode access
NX01(config-if)#switchport access vlan 10
NX01(config)#port-channel load-balance src ip
```

```
NX02(config)#int port-channel 10
NX02(config-if)#vpc 10
NX02(config-if)#switchport mode access
NX02(config-if)#switchport access vlan 10
NX02(config)#port-channel load-balance src ip
```

Trên switch access, chúng ta gán port e0/0 và e0/1 trên cả hai switch Nexus vào etherchannel. Etherchannel này có id là 10 và ở chế độ access vlan 10.

```
Switch_Access(config)#vlan 10
Switch_Access(config)#int range e0/0-1
Switch_Access(config-if-range)#switchport access vlan 10
Switch_Access(config-if-range)#channel-group 10  mode active
Switch_Access(config)#port-channel load-balance src-ip
```

Thực hiện kiểm tra cấu hình vPC qua câu lệnh **show vpc** hoặc **show vpc detail**.

```
NX01(config)# show vpc
Legend:
                (*) - local vPC is down, forwarding via vPC peer-link

vPC domain id                     : 100
```

```
Peer status                         : peer adjacency formed ok
vPC keep-alive status               : peer is alive
Configuration consistency status    : success
Per-vlan consistency status         : success
Type-2 consistency status           : success
vPC role                            : primary
Number of vPCs configured           : 1
Peer Gateway                        : Enabled
Dual-active excluded VLANs          : -
Graceful Consistency Check          : Enabled
Auto-recovery status                : Disabled
Delay-restore status                : Timer is off.(timeout = 30s)
Delay-restore SVI status            : Timer is off.(timeout = 10s)
Operational Layer3 Peer-router      : Enabled
Virtual-peerlink mode               : Disabled

vPC Peer-link status
--------------------------------------------------------------------
id     Port   Status Active vlans
--     ----   ------ --------------------------------------------------
1      Po100  up     1,10

vPC status
----------------------------------------------------------------------------
Id     Port          Status Consistency Reason                Active vlans
--     ------------  ------ ----------- ------                ---------------
10     Po10          up     success     success               10
```

Chú ý những thông tin được in đậm, nếu các trạng thái portchannel 10 và 100 là up đường peerlink alive thì cấu hình đã thành công.

Phần 2. Cấu hình PIM và vxlan

Trong bài Lab về *VxLan overlay network Multicast,* chúng ta đã cấu hình vxlan kết hợp với pim (là giao thức định tuyến multicast). Địa chỉ ip của multicast group tạo bởi 3 router R1-R2-R3 là 230.1.1.1 và Rendezvous Point (RP) có địa chỉ ip là 10.1.1.1 là R1 loopback.

Một lưu ý ở hai interface loopback 0 trên NX01 và NX02 (như đã nói ở phần I), Địa chỉ IP chính 200.200.200.1 là địa chỉ duy nhất và được sử dụng bởi giao thức lớp 3.

Địa chỉ IP phụ 100.100.100.1 trên Loopback là cần thiết vì cổng NVE sẽ sử dụng cho địa chỉ IP VTEP. Địa chỉ IP phụ phải giống nhau trên cả 2 vPC Peers.

```
NX01(config)#feature nv overlay
NX01(config)#feature vn-segment-vlan-based
NX01(config)#feature ospf
NX01(config)#feature pim
NX01(config)#router ospf 1
NX01(config-router)#router-id 200.200.200.1
NX01(config)#ip pim rp-address 10.1.1.1 group-list 224.0.0.0/4
NX01(config)#int lo0
```

```
NX01(config-if)#ip add 200.200.200.1/32
NX01(config-if)#ip address 100.100.100.1/32 secondary
NX01(config-if)#ip router ospf 1 area 0
NX01(config-if)#ip pim sparse-mode
NX01(config)#int e1/1
NX01(config-if)#no switchport
NX01(config-if)#ip add 20.1.1.1/30
NX01(config-if)#ip router ospf 1 area 0
NX01(config-if)#ip pim sparse-mode
NX01(config)#int nve 1
NX01(config-if-nve)#no shut
NX01(config-if-nve)#source-interface loopback 0
NX01(config-if-nve)#member vni 10000 mcast-group 230.1.1.1
NX01(config-if-nve)#vlan 10
NX01(config-vlan)#vn-segment 10000
```

```
NX02(config)#feature nv overlay
NX02(config)#feature vn-segment-vlan-based
NX02(config)#feature ospf
NX02(config)#feature pim
NX02(config)#router ospf 1
NX02(config-router)#router-id 200.200.200.2
NX02(config)#ip pim rp-address 10.1.1.1 group-list 224.0.0.0/4
NX02(config)#int lo0
NX02(config-if)#ip add 200.200.200.2/32
NX02(config-if)#ip address 100.100.100.1/32 secondary
NX02(config-if)#ip router ospf 1 area 0
NX02(config-if)#ip pim sparse-mode
NX02(config)#int e1/1
NX02(config-if)#no switchport
NX02(config-if)#ip add 20.1.1.5/30
NX02(config-if)#ip router ospf 1 area 0
NX02(config-if)#ip pim sparse-mode
NX02(config)#int nve 1
NX02(config-if-nve)#no shut
NX02(config-if-nve)#source-interface loopback 0
NX02(config-if-nve)#member vni 10000 mcast-group 230.1.1.1
NX02(config-if-nve)#vlan 10
NX02(config-vlan)#vn-segment 10000
```

```
NX03(config)#feature nv overlay
NX03(config)#feature vn-segment-vlan-based
NX03(config)#feature ospf
NX03(config)#feature pim
NX03(config)#router ospf 1
NX03(config-router)#router-id 100.100.100.2
NX03(config)#ip pim rp-address 10.1.1.1 group-list 224.0.0.0/4
NX03(config)#int lo0
NX03(config-if)#ip add 100.100.100.2/32
NX03(config-if)#ip router ospf 1 area 0
NX03(config-if)#ip pim sparse-mode
```

```
NX03(config-if)#int e1/1
NX03(config-if)#no switchport
NX03(config-if)#ip add 30.1.1.1/30
NX03(config-if)#ip router ospf 1 area 0
NX03(config-if)#ip pim sparse-mode
NX03(config)#int e1/2
NX03(config-if)#switchport
NX03(config-if)#switchport access vlan 10
NX03(config-if)#no shut
NX03(config)#int nve 1
NX03(config-if-nve)#no shut
NX03(config-if-nve)#source-interface loopback 0
NX03(config-if-nve)#member vni 10000 mcast-group 230.1.1.1
NX03(config-if-nve)#exit
NX03(config)#vlan 10
NX03(config-vlan)#vn-segment 10000
```

Trên các router multicast, chúng ta cấu hình ip và định tuyến ospf kết hợp định tuyến multicast với giao thức PIM.

```
R2(config)#ip multicast-routing
R2(config)#router ospf 1
R2(config-router)#router-id 2.2.2.2
R2(config)#int e0/1
R2(config-if)#ip add 20.1.1.2 255.255.255.252
R2(config-if)#ip ospf 1 area 0
R2(config-if)#no shut
R2(config)#int e0/0
R2(config-if)#ip add 40.1.1.2 255.255.255.252
R2(config-if)#ip ospf 1 area 0
R2(config-if)#no shut
R2(config)#int range e0/0-2
R2(config-if-range)#ip pim sparse-mode
R2(config)#ip pim rp-address 10.1.1.1
R2(config)#int e0/0
R2(config-if)#ip igmp join-group 230.1.1.1
```

```
R3(config)#ip multicast-routing
R3(config)#router ospf 1
R3(config-router)#router-id 3.3.3.3
R3(config)#int e0/0
R3(config-if)#ip add 30.1.1.2 255.255.255.252
R3(config-if)#ip ospf 1 area 0
R3(config-if)#no shut
R3(config)#int e0/1
R3(config-if)#ip add 50.1.1.2 255.255.255.252
R3(config-if)#ip ospf 1 area 0
R3(config-if)#no shut
R3(config)#int range e0/0-1
R3(config-if-range)#ip pim sparse-mode
R3(config)#ip pim rp-address 10.1.1.1
R3(config)#int e0/1
```

```
R3(config-if)#ip igmp join-group 230.1.1.1
```

```
R1(config)#ip multicast-routing
R1(config)#router ospf 1
R1(config-router)#router-id 1.1.1.1
R1(config)#int lo0
R1(config-if)#ip add 10.1.1.1 255.255.255.255
R1(config-if)#ip pim sparse-mode
R1(config-if)#ip ospf 1 area 0
R1(config)#int e0/0
R1(config-if)#ip add 50.1.1.1 255.255.255.252
R1(config-if)#ip ospf 1 area 0
R1(config-if)#ip pim sparse-mode
R1(config)#int e0/1
R1(config-if)#ip add 40.1.1.1 255.255.255.252
R1(config-if)#ip ospf 1 area 0
R1(config-if)#ip pim sparse-mode
R1(config)#ip pim rp-address 10.1.1.1
```

3. KIỂM TRA CẤU HÌNH

Dùng lệnh debug nve < > với các mode hỗ trợ để có thể xem log các gói vxlan đóng/mở.

```
NX01#debug nve packet
NX02#debug nve packet
NX03#debug nve packet
```

Dưới đây là một log entry mẫu để chúng ta có thể thấy 1 gói tin nào đó đang đi qua NX01 đang được decapsulate gói vxlan có layer 4 port là 4789, VNI là 10000, có source ip là 100.100.100.2 là ip VTEP NX03 gửi đến dest là multicast 230.1.1.1.

```
NX01# 2019 Dec 20 04:14:56.942057 nve: nve_ip_udp_decapsulate: VxLAN
Packet - UDP Port: 4789 VNI: 10000 VTEP: 0x49000001 Source: 100.100.100.2
Dest: 230.1.1.1
```

Ngoài ra, chúng ta có thể thực hiện các lệnh **show nve vni** – để theo dõi trạng thái interface nve với các thông số VNI, **show nve peers** – để quan sát peer của thiết bị chúng ta gõ lệnh, **show nve interface** – qua lệnh này chúng ta có thể quan sát interface nve trên 1 VTEP dùng đóng gói VxLan với sự kết hợp VPC.

```
NX03# show nve vni
Codes: CP - Control Plane        DP - Data Plane
       UC - Unconfigured         SA - Suppress ARP
       SU - Suppress Unknown Unicast
       Xconn - Crossconnect
       MS-IR - Multisite Ingress Replication

Interface VNI      Multicast-group   State Mode Type [BD/VRF]      Flags
--------- -------- ----------------- ----- ---- ------------------ -----
nve1      10000    230.1.1.1         Up    DP   L2 [10]

NX02# show nve vni
…
Interface VNI      Multicast-group   State Mode Type [BD/VRF]      Flags
--------- -------- ----------------- ----- ---- ------------------ -----
nve1      10000    230.1.1.1         Up    DP   L2 [10]
```

```
NX01(config)# show nve peers
Interface Peer-IP                 State LearnType Uptime   Router-Mac
--------- ----------------------  ----- --------- -------- ------------
nve1      100.100.100.2           Up    DP        01:40:15 n/a
```

```
NX01(config)# show nve interface
Interface: nve1, State: Up, encapsulation: VXLAN
 VPC Capability: VPC-VIP-Only [notified]
 Local Router MAC: 5001.0001.0007
 Host Learning Mode: Data-Plane
 Source-Interface: loopback0 (primary: 200.200.200.1, secondary:
100.100.100.1
```

```
NX02(config)# show nve interface
Interface: nve1, State: Up, encapsulation: VXLAN
 VPC Capability: VPC-VIP-Only [notified]
 Local Router MAC: 5001.0002.0007
 Host Learning Mode: Data-Plane
 Source-Interface: loopback0 (primary: 200.200.200.2, secondary:
100.100.100.1)
```

```
NX03# show nve interface
Interface: nve1, State: Up, encapsulation: VXLAN
 VPC Capability: VPC-VIP-Only [not-notified]
 Local Router MAC: 5001.0003.0007
 Host Learning Mode: Data-Plane
 Source-Interface: loopback0 (primary: 100.100.100.2, secondary: 0.0.0.0)
```

```
NX01# show nve vni data-plane
Codes: CP - Control Plane        DP - Data Plane
       UC - Unconfigured         SA - Suppress ARP
       SU - Suppress Unknown Unicast
       Xconn - Crossconnect
```

```
        MS-IR - Multisite Ingress Replication

Interface VNI      Multicast-group   State Mode Type [BD/VRF]      Flags
--------- -------- ----------------- ----- ---- ------------------ -----
nve1      10000    230.1.1.1         Up    DP   L2 [10]
```

Dùng PC9 với ip 192.168.1.11 ping PC8 192.168.1.10 ta có thể thấy gói ICMP đã chạy thông suốt. Sau đó chúng ta thực hiện lệnh trace để xem các thiết bị mà gói ICMP đã đi qua khi PC9 gửi đi. Có thể thấy rằng, mặc dù giữa 2 PC phải đi qua rất nhiều hop là các switch, các router có địa chỉ ip thuộc các lớp mạng khác nhau nhưng khi thực hiện trace thì ta chỉ có thể thấy gói ICMP đi thẳng tới hop là destination của gói tin ICMP đó, giống với cách mà ta ping nội bộ trong 1 lớp mạng LAN mà không cần đến sự dẫn đường của router, và đó cùng là đặc trưng mà VxLan tạo ra.

```
84 bytes from 192.168.1.10 icmp_seq=73 ttl=64 time=25.978 ms
84 bytes from 192.168.1.10 icmp_seq=74 ttl=64 time=29.079 ms
84 bytes from 192.168.1.10 icmp_seq=75 ttl=64 time=38.823 ms
84 bytes from 192.168.1.10 icmp_seq=76 ttl=64 time=25.102 ms
84 bytes from 192.168.1.10 icmp_seq=77 ttl=64 time=27.508 ms
84 bytes from 192.168.1.10 icmp_seq=78 ttl=64 time=47.915 ms
84 bytes from 192.168.1.10 icmp_seq=79 ttl=64 time=31.212 ms
84 bytes from 192.168.1.10 icmp_seq=80 ttl=64 time=26.929 ms
84 bytes from 192.168.1.10 icmp_seq=81 ttl=64 time=32.308 ms
84 bytes from 192.168.1.10 icmp_seq=82 ttl=64 time=26.471 ms
84 bytes from 192.168.1.10 icmp_seq=83 ttl=64 time=36.621 ms
84 bytes from 192.168.1.10 icmp_seq=84 ttl=64 time=24.943 ms
84 bytes from 192.168.1.10 icmp_seq=85 ttl=64 time=34.156 ms
84 bytes from 192.168.1.10 icmp_seq=86 ttl=64 time=33.253 ms
84 bytes from 192.168.1.10 icmp_seq=87 ttl=64 time=28.839 ms
84 bytes from 192.168.1.10 icmp_seq=88 ttl=64 time=25.181 ms
84 bytes from 192.168.1.10 icmp_seq=89 ttl=64 time=29.152 ms
84 bytes from 192.168.1.10 icmp_seq=90 ttl=64 time=25.962 ms
84 bytes from 192.168.1.10 icmp_seq=91 ttl=64 time=64.069 ms
84 bytes from 192.168.1.10 icmp_seq=92 ttl=64 time=25.364 ms
84 bytes from 192.168.1.10 icmp_seq=93 ttl=64 time=28.891 ms
84 bytes from 192.168.1.10 icmp_seq=94 ttl=64 time=28.903 ms
84 bytes from 192.168.1.10 icmp_seq=95 ttl=64 time=33.800 ms
```

```
VPCS> trace 192.168.1.11
traceroute to 192.168.1.11, 8 hops max
 1 192.168.1.11     0.001 ms
```