# Understanding CEF Weight Distributions In Unequal-cost Load Sharing

## **Contents**

**Introduction** 

**Prerequisites** 

Requirements

**Components Used** 

**UCMP Overview** 

**Initial Configurations** 

Metric Weights/Load

**UCMP Variance Determination** 

**Understanding Weights** 

**Determining Weight Values** 

Weight

Normalized weight

Manipulating CEF Weight/Load Ratios

Example 1: Weight/Load Ratio of 26/5

Example 2: Weight/Load Ratio of 30/1

## Introduction

This document describes the aspects of Understanding, Configuring, and Verify Unequal-cost multipath in IOS-XR. Also we go through examples of weight manipulations to show how the path metric to a destination influences the load on a link.

# **Prerequisites**

This document does not have prerequisites.

# Requirements

Examples below are based on IOS-XR 6.4.1.

# **Components Used**

The information in this document was created from the devices in a specific lab environment. All of the devices used in this document started with a cleared (default) configuration. If your network is live, make sure that you understand the potential impact of any command.

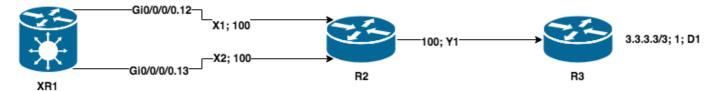
## **UCMP** Overview

The unequal cost multipath (UCMP) load-balancing provides the capability to load balance traffic proportionally across multiple paths, with different cost. Generally, higher bandwidth paths have lower Interior Gateway Protocol (IGP) metrics configured, so that they form the shortest IGP paths.

With the UCMP load-balancing enabled, protocols can use even lower bandwidth paths or higher cost paths for traffic, and can install these paths to the forwarding information base (FIB). These protocols still install multiple paths to the same destination in FIB, but each path will have a 'load metric/weight' associated with it. FIB uses this load metric/weight to decide the amount of traffic that needs to be sent on a higher bandwidth path and the amount of traffic that needs to be sent on a lower bandwidth path.

Traditionally, EIGRP has been the only IGP that supports UCMP feature, but in IOS-XR UCMP is supported for all IGPs, static routing, and BGP. In this document, we will explain the UCMP feature using OSPF as the basis of our examples, but the information here also applies to IS-IS and other UCMP-capable protocols.

#### **Topology Diagram**



# **Initial Configurations**

```
XR1
hostname XR1
interface GigabitEthernet0/0/0/0.12
description TO R2
ipv4 address 12.0.0.1 255.255.255.0
encapsulation dot1q 12
interface GigabitEthernet0/0/0/0.13
description TO R2
ipv4 address 13.0.0.1 255.255.255.0
encapsulation dot1q 13
! router ospf 1 address-family ipv4 area 0 ! interface GigabitEthernet0/0/00.12
                                                                                      cost 100
  interface GigabitEthernet0/0/0/0.13
   cost 100
  1
!
end
R2
hostname R2
1
interface Ethernet0/0.12
description TO XR1
encapsulation dot1Q 12
 ip address 12.0.0.2 255.255.255.0
```

```
interface Ethernet0/0.13
description TO XR1
encapsulation dot1Q 13
ip address 13.0.0.2 255.255.255.0
interface Ethernet0/1
description TO R3
ip address 172.16.23.2 255.255.255.0
ip ospf cost 100
router ospf 1
network 0.0.0.0 255.255.255.255 area 0
end
RЗ
hostname R3
!
interface Loopback0
description FINAL_DESTINATION
ip address 3.3.3.3 255.255.255.255
interface Ethernet0/0
description TO R2
ip address 172.16.23.3 255.255.255.0
router ospf 1
network 0.0.0.0 255.255.255.255 area 0
1
end
```

# **Metric Weights/Load**

In IOS-XR, when we install multiple paths to a destination, the destination is assigned a weight value that indicates the load distribution for a particular link. This value is inversely proportional to the path metric to the destination, the higher the cost, the lower the weight is assigned. This allows CEF to intelligently perform load-sharing of links when routing to destinations.

When ECMP paths are installed, weight values assigned are always set to 0 for all paths, this means that traffic is load-shared equally. If we check CEF we can confirm that weights of 0 have been assigned for each path.

## **UCMP Variance Determination**

If we want to enable UCMP, let's begin by setting cost differently on XR1, for this, we will set cost as below:

```
router ospf 1
address-family ipv4
area 0
 interface Loopback0
 interface GigabitEthernet0/0/0/0.12
  cost 50
 interface GigabitEthernet0/0/0/0.13
  cost 100
  !
 !
end
RP/0/RP0/CPU0:XR1#show route 3.3.3.3/32
Routing entry for 3.3.3.3/32
 Known via "ospf 1", distance 110, metric 151, type intra area
 Installed Nov 11 22:32:48.289 for 00:00:32
 Routing Descriptor Blocks
   12.0.0.2, from 3.3.3.3, via GigabitEthernet0/0/0/0.12
     Route metric is 151
 No advertising protos.
```

To consider other paths for UCMP we need to determine if these are eligible. IOS-XR uses a percentage criteria for IS-IS and OSPF, this is based on the **ucmp variance <value>** router process command. The two paths we have are:

```
path metric 1 (pm1) = 151
path metric 2 (pm2) = 201
```

Loop free next-hops will be installed based on UCMP <= (Variance \* Primary path metric) / 100.

How much primary path has to grow to reach the worst path metric (pm2) in this case is 134 percent of 151, which results in 202. This is the exact variance value we need to configure to make the path eligible.

```
!
router ospf 1
 ucmp variance 134
```

```
RP/0/RP0/CPU0:XR1#show route 3.3.3.3/32
Routing entry for 3.3.3.3/32
Known via "ospf 1", distance 110, metric 151, type intra area Installed Nov 11 22:36:45.720 for 00:00:09
Routing Descriptor Blocks
12.0.0.2, from 3.3.3.3, via GigabitEthernet0/0/0/0.12
Route metric is 151, Wt is 4294967295
13.0.0.2, from 3.3.3.3, via GigabitEthernet0/0/0/0.13
Route metric is 151, Wt is 3226567396
```

#### **Spoiler**

No advertising protos.

**Note**: The variance value does not have any impact on the weight results. In this case a minimum variance of 134 or a variance of 10000 (max value) would have lead to the same weight results, instead, the cost values are the ones that influence the resulting weights, as these are inversely proportional to each other.

Note: The variance value does not have any impact on the weight results. In this case a minimum variance of 134 or a variance of 10000 (max value) would have lead to the same weight results, instead, the cost values are the ones that influence the resulting weights, as these are inversely proportional to each other.

# **Understanding Weights**

We have two different types of weights in IOS-XR, weight and normalized weights. The usage of these is based on how many hash buckets are supported on a particular platform, XRv9000 support 32 hash buckets, ASR 9000 and CRS-X support 64 hash buckets respectively. This means that, when the router programs the weight values, the weighting cannot exceed the hash bucket limit of the particular platform. We can observe what normalized weights are programmed by issuing the **show cef refix> detail location <location> command. Based on the cost values set, we have a 18, 13 load distribution, which means 31 hash buckets have been assigned (18+13).** 

```
RP/0/RP0/CPU0:XR1#show cef 3.3.3.3/32 detail
3.3.3/32, version 23, internal 0x1000001 0x0 (ptr 0xd3ecb50) [1], 0x0 (0xd583610), 0x0 (0x0)
Updated Nov 11 22:36:45.723
remote adjacency to GigabitEthernet0/0/0/0.12
Prefix Len 32, traffic index 0, precedence n/a, priority 1
 gateway array (0xd4163d8) reference count 1, flags 0x0, source rib (7), 0 backups
               [2 type 3 flags 0x8401 (0xd4bc7b8) ext 0x0 (0x0)]
 LW-LDI[type=3, refc=1, ptr=0xd583610, sh-ldi=0xd4bc7b8]
 gateway array update type-time 1 Nov 11 22:36:45.723
LDI Update time Nov 11 22:36:45.729
LW-LDI-TS Nov 11 22:36:45.729
  via 12.0.0.2/32, GigabitEthernet0/0/0/0.12, 6 dependencies, weight 4294967295, class 0 [flags
0 \times 01
   path-idx 0 NHID 0x0 [0xe14b1b0 0x0]
   next hop 12.0.0.2/32
   remote adjacency
  via 13.0.0.2/32, GigabitEthernet0/0/0/0.13, 6 dependencies, weight 3226567396, class 0 [flags
0 \times 01
   path-idx 1 NHID 0x0 [0xe14b128 0x0]
   next hop 13.0.0.2/32
```

```
Hash OK Interface
                                   Address
     Y GigabitEthernet0/0/0/0.12 remote
0
1
     Y GigabitEthernet0/0/0/0.12 remote
     Y GigabitEthernet0/0/0/0.12 remote
2
        GigabitEthernet0/0/0/0.12 remote
3
     Y
     Y
         GigabitEthernet0/0/0/0.12 remote
5
     Y
        GigabitEthernet0/0/0/0.12 remote
6
     Y GigabitEthernet0/0/0/0.12 remote
7
     Y GigabitEthernet0/0/0/0.12 remote
     Y GigabitEthernet0/0/0/0.12 remote
9
     Y GigabitEthernet0/0/0/0.12 remote
     Y
        GigabitEthernet0/0/0/0.12 remote
10
         GigabitEthernet0/0/0/0.12 remote
11
12
     Y
         GigabitEthernet0/0/0/0.12 remote
13
     Y
         GigabitEthernet0/0/0/0.12 remote
14
     Y GigabitEthernet0/0/0/0.12 remote
     Y GigabitEthernet0/0/0/0.12 remote
     Y GigabitEthernet0/0/0/0.12 remote
16
     Y GigabitEthernet0/0/0/0.12 remote
17
18
     Y
         GigabitEthernet0/0/0/0.13 remote
     Y
         GigabitEthernet0/0/0/0.13 remote
20
        GigabitEthernet0/0/0/0.13 remote
     Y
21
     Y GigabitEthernet0/0/0/0.13 remote
     Y GigabitEthernet0/0/0/0.13 remote
22
23
     Y GigabitEthernet0/0/0/0.13 remote
24
     Y GigabitEthernet0/0/0/0.13 remote
        GigabitEthernet0/0/0/0.13 remote
25
     Y
26
         GigabitEthernet0/0/0/0.13 remote
        GigabitEthernet0/0/0/0.13 remote
27
     Y
28
     Y GigabitEthernet0/0/0/0.13 remote
29
     Y GigabitEthernet0/0/0/0.13 remote
         GigabitEthernet0/0/0/0.13 remote
```

remote adjacency

As we can observe, the sum of the normalized weight represents the amount of hash buckets assigned by the platform, in this case, we can never exceed 32 hash buckets, as per the limit of this particular platform. The weight of the primary path (pm1) is always set to 4294967295, which is the maximum weight (2^32) - 1.

# **Determining Weight Values**

## Weight

We can easily compute the weights with the formula **weight = best cost / worst cost \* 4294967295**. For example, weights for path 1 and path 2 are calculated below:

```
Weight_path_1 = always set to 4294967295
```

Weight\_path\_2 = 151 / 201 \* 4294967295 = 3226567470

#### **Spoiler**

Note: Loss of precision can happen when computing the values, as we are doing floating point calculations, and we must install integers in RIB and FIB.

Note: Loss of precision can happen when computing the values, as we are doing floating point calculations, and we must install integers in RIB and FIB.

## Normalized weight

As we mentioned, we cannot install in the CEF table weight values exceeding the amount of hash buckets by a platform, due to this we need to normalize the weights before programming them into hardware. Platform computes the normalize weights according to the formula **Normalized Weight** = (Path weight/Total weight) \* Maximum bucket size. Based on our example, we can compute this as follows:

```
normalized_weight_1 = (4294967295 * 32) / (3226567396 + 4294967295) = 18
normalized_weight_2 = (3226567396 * 32) / (3226567396 + 4294967295) = 13
```

#### **Spoiler**

Note: When the G.C.D is equal to 1, then above method is used, otherwise if G.C.D =! 1, then normalize weight will be division of the resulting G.C.D by the weight values.

Note: When the G.C.D is equal to 1, then above method is used, otherwise if G.C.D =! 1, then normalize weight will be division of the resulting G.C.D by the weight values.

## **Manipulating CEF Weight/Load Ratios**

In some scenarios we might want to determine what particular path metric value we need to configure to have a resulting weight/load distribution. We could determine the appropriate path metric by changing the cost of the links and based on until we reach or approximate the required value. Note that not all weights we might require are exactly possible, but we can approximate the distribution required.

Before continuing, take the following restrictions into account:

- a.) Not all weight/load distributions are exactly possible, but we can do an approximation.
- b.) Never exceed the hash bucket limits. This means that the sum of all path weights cannot exceed the hash buckets, if this happens, then the weight must be normalized. Meaning that, when adding up all weights, we do not exceed the hash bucket limit.
- c.) ASR 9000 and CRS-X have a 64 hash bucket limit, XRv9000 have a 32 hash bucket limit.
- d.) When using pre-6.4.1, weight distribution is different, and the path with the least weight is always set to a weight of 1 while other paths are multiples of this path which means that it can be higher than 1.

## Example 1: Weight/Load Ratio of 26/5

Following the same topology before, we want to have a 26/5 weight distribution between the two links.

i.) Initially, the costs are equally set on all paths (100 + 100 + 1) = 201.

- ii.) If we will set the UCMP variance to the maximum value, to consider all next-hops.
- iii.) If we check the RIB, we can see the default state where XR1 is doing ECMP.

```
RP/0/RP0/CPU0:XR1#show cef 3.3.3.3/32 detail
3.3.3/32, version 27, internal 0x1000001 0x0 (ptr 0xd3ecb50) [1], 0x0 (0xd583610), 0x0 (0x0)
Updated Nov 11 23:08:25.290
remote adjacency to GigabitEthernet0/0/0/0.12
Prefix Len 32, traffic index 0, precedence n/a, priority 1
 gateway array (0xd416218) reference count 2, flags 0x0, source rib (7), 0 backups
               [3 type 3 flags 0x8401 (0xd4bc6f8) ext 0x0 (0x0)]
 LW-LDI[type=3, refc=1, ptr=0xd583610, sh-ldi=0xd4bc6f8]
 gateway array update type-time 1 Nov 11 23:08:25.290
LDI Update time Nov 11 23:08:25.297
LW-LDI-TS Nov 11 23:08:25.297
  via 12.0.0.2/32, GigabitEthernet0/0/0/0.12, 4 dependencies, weight 4294967295, class 0 [flags
0 \times 01
   path-idx 0 NHID 0x0 [0xe14b1b0 0x0]
   next hop 12.0.0.2/32
   remote adjacency
  via 13.0.0.2/32, GigabitEthernet0/0/0/0.13, 4 dependencies, weight 4294967295, class 0 [flags
0x0]
   path-idx 1 NHID 0x0 [0xe14b128 0x0]
   next hop 13.0.0.2/32
   remote adjacency
   Weight distribution:
   slot 0, weight 4294967295, normalized_weight 1, class 0
   slot 1, weight 4294967295, normalized_weight 1, class 0
   Load distribution: 0 1 (refcount 3)
   Hash OK Interface
         Y
            GigabitEthernet0/0/0/0.12 remote
             GigabitEthernet0/0/0/0.13 remote
```

For this example, we will use a case where you want the following weights:

```
W1 = 26 (primary best cost)
```

W2 = 5 (secondary best cost)

We need to take a leg path, for this path, the cost should known already, in this case reference path will be the path via Gi0/0/0/0.12. The leg path will be pre-computed with cost from end to end, the path metric and weight required for this path are:

- i.) X1+Y1+D1 = 100 + 100 + 1 = 201. (Note the variables attached to each link in the topology).
- ii.) Weight 1 = 26
- iii.) Weight 2 = 5
- iv.) pm1 = 201 (primary leg path); Weight = 26
- v.) pm2 = unknown yet (secondary path); Weight = 5

Computing the weights.

Path metric of pm2: pm2 = (26/5) \* 201 = 1045

Determining cost of link X2 on XR1.

```
X2 = pm2-(x2+y1+d1)
1045-(100+100+1) = 844
```

3

4

Y

Y GigabitEthernet0/0/0/0.12 remote

Y GigabitEthernet0/0/0/0.12 remote

GigabitEthernet0/0/0/0.12 remote

Configuring OSPF cost on X2 link.

```
router ospf 1
ucmp variance 10000
area 0
!
  interface GigabitEthernet0/0/0/0.13
  cost 844
```

Verifying weight/load distribution we can see that the weights required have been assigned appropriately in CEF as we predicted in the calculations.

```
RP/0/RP0/CPU0:XR1#show cef 3.3.3.3/32 detail
3.3.3/32, version 37, internal 0x1000001 0x0 (ptr 0xd3ecce0) [1], 0x0 (0xd5835d8), 0x0 (0x0)
Updated Nov 11 23:17:47.945
remote adjacency to GigabitEthernet0/0/0/0.12
Prefix Len 32, traffic index 0, precedence n/a, priority 1
 gateway array (0xd4163d8) reference count 1, flags 0x0, source rib (7), 0 backups
              [2 type 3 flags 0x8401 (0xd4bc7b8) ext 0x0 (0x0)]
 LW-LDI[type=3, refc=1, ptr=0xd5835d8, sh-ldi=0xd4bc7b8]
 gateway array update type-time 1 Nov 11 23:17:47.945
LDI Update time Nov 11 23:17:47.956
LW-LDI-TS Nov 11 23:17:47.956
  via 12.0.0.2/32, GigabitEthernet0/0/0/0.12, 6 dependencies, weight 4294967295, class 0 [flags
0 \times 01
   path-idx 0 NHID 0x0 [0xe14b1b0 0x0]
   next hop 12.0.0.2/32
   remote adjacency
  via 13.0.0.2/32, GigabitEthernet0/0/0/0.13, 6 dependencies, weight 913532538, class 0 [flags
0x0]
   path-idx 1 NHID 0x0 [0xe14b128 0x0]
   next hop 13.0.0.2/32
   remote adjacency
   Weight distribution:
   slot 0, weight 4294967295, normalized_weight 26, class 0
   slot 1, weight 913532538, normalized_weight 5, class 0
   2)
   Hash OK Interface
                                     Address
   0
         Y
           GigabitEthernet0/0/0/0.12 remote
   1
         Y GigabitEthernet0/0/0/0.12 remote
   2
         Y GigabitEthernet0/0/0/0.12 remote
```

```
6
     Y
        GigabitEthernet0/0/0/0.12 remote
7
     Y GigabitEthernet0/0/0/0.12 remote
8
     Y GigabitEthernet0/0/0/0.12 remote
9
     Y GigabitEthernet0/0/0/0.12 remote
10
     Y GigabitEthernet0/0/0/0.12 remote
     Y GigabitEthernet0/0/0/0.12 remote
11
12
     Y
        GigabitEthernet0/0/0/0.12 remote
     Y
         GigabitEthernet0/0/0/0.12 remote
14
     Y
         GigabitEthernet0/0/0/0.12 remote
     Y GigabitEthernet0/0/0/0.12 remote
15
16
     Y GigabitEthernet0/0/0/0.12 remote
17
     Y GigabitEthernet0/0/0/0.12 remote
18
     Y GigabitEthernet0/0/0/0.12 remote
19
     Y
        GigabitEthernet0/0/0/0.12 remote
20
        GigabitEthernet0/0/0/0.12 remote
     Y
        GigabitEthernet0/0/0/0.12 remote
     Y GigabitEthernet0/0/0/0.12 remote
22
23
    Y GigabitEthernet0/0/0/0.12 remote
    Y GigabitEthernet0/0/0/0.12 remote
25
    Y GigabitEthernet0/0/0/0.12 remote
     Y GigabitEthernet0/0/0/0.13 remote
26
27
        GigabitEthernet0/0/0/0.13 remote
     Y
        GigabitEthernet0/0/0/0.13 remote
29
     Y GigabitEthernet0/0/0/0.13 remote
         GigabitEthernet0/0/0/0.13 remote
```

## Example 2: Weight/Load Ratio of 30/1

Same as before, we default cost to 100 on both XR1 interfaces.

```
W1 = 30 (primary best cost)
```

W2 = 1 (secondary best cost)

- i.) X1+Y1+D1 = 100 + 100 + 1 = 201. (Note the variables attached to each link in the topology).
- ii.) Weight 1 = 30
- iii.) Weight 2 = 1
- iv.) pm1 = 201 (primary leg path); Weight = 30
- v.) pm2 = unknown yet (secondary path); Weight = 1

Computing the weights.

Path metric of pm2: pm2 = (30/1) \* 201 = 6030

Determining cost of link X2 on XR1.

$$X2 = pm2-(x2+y1+d1)$$

$$6030 - (100 + 100 + 1) = 5829$$

Configuring OSPF cost on X2 link.

```
router ospf 1
  ucmp variance 10000
area 0
!
  interface GigabitEthernet0/0/0/0.13
  cost 5829
```

Verifying weight/load distribution we can see that the weights required have been assigned appropriately in CEF as we predicted in the calculations.

```
RP/0/RP0/CPU0:XR1#show cef 3.3.3.3/32 detail
3.3.3/32, version 40, internal 0x1000001 0x0 (ptr 0xd3ecce0) [1], 0x0 (0xd5835d8), 0x0 (0x0)
Updated Nov 11 23:31:58.207
remote adjacency to GigabitEthernet0/0/0/0.12
Prefix Len 32, traffic index 0, precedence n/a, priority 1
 gateway array (0xd416218) reference count 1, flags 0x0, source rib (7), 0 backups
               [2 type 3 flags 0x8401 (0xd4bc6f8) ext 0x0 (0x0)]
 LW-LDI[type=3, refc=1, ptr=0xd5835d8, sh-ldi=0xd4bc6f8]
 gateway array update type-time 1 Nov 11 23:31:58.207
LDI Update time Nov 11 23:31:58.208
LW-LDI-TS Nov 11 23:31:58.208
  via 12.0.0.2/32, GigabitEthernet0/0/0/0.12, 6 dependencies, weight 4294967295, class 0 [flags
0x0]
   path-idx 0 NHID 0x0 [0xe14b1b0 0x0]
   next hop 12.0.0.2/32
   remote adjacency
  via 13.0.0.2/32, GigabitEthernet0/0/0/0.13, 6 dependencies, weight 140784018, class 0 [flags
0x0]
   path-idx 1 NHID 0x0 [0xe14b128 0x0]
   next hop 13.0.0.2/32
   remote adjacency
   Weight distribution:
   slot 0, weight 4294967295, normalized_weight 30, class 0
   slot 1, weight 140784018, normalized_weight 1, class 0
   2)
   Hash OK Interface
                                      Address
         Y
             GigabitEthernet0/0/0/0.12 remote
    0
   1
         Y
             GigabitEthernet0/0/0/0.12 remote
    2
         Y
             GigabitEthernet0/0/0/0.12 remote
    3
             GigabitEthernet0/0/0/0.12 remote
         Y
             GigabitEthernet0/0/0/0.12 remote
    4
         Y
    5
         Y
             GigabitEthernet0/0/0/0.12 remote
             GigabitEthernet0/0/0/0.12 remote
    6
         Y
    7
         Y
             GigabitEthernet0/0/0/0.12 remote
    8
         Y
             GigabitEthernet0/0/0/0.12 remote
    9
         Y
             GigabitEthernet0/0/0/0.12 remote
   10
             GigabitEthernet0/0/0/0.12 remote
         Y
   11
         Y
             GigabitEthernet0/0/0/0.12 remote
             GigabitEthernet0/0/0/0.12 remote
    12
         Y
   13
         Y
             GigabitEthernet0/0/0/0.12 remote
   14
             GigabitEthernet0/0/0/0.12 remote
         Y
   15
         Y
             GigabitEthernet0/0/0/0.12 remote
    16
         Y
             GigabitEthernet0/0/0/0.12 remote
   17
         Y
             GigabitEthernet0/0/0/0.12 remote
   18
         Y
             GigabitEthernet0/0/0/0.12 remote
    19
         Y
             GigabitEthernet0/0/0/0.12 remote
```

GigabitEthernet0/0/0/0.12 remote

21	Y	<pre>GigabitEthernet0/0/0/0.12</pre>	remote
22	Y	<pre>GigabitEthernet0/0/0/0.12</pre>	remote
23	Y	<pre>GigabitEthernet0/0/0/0.12</pre>	remote
24	Y	<pre>GigabitEthernet0/0/0/0.12</pre>	remote
25	Y	<pre>GigabitEthernet0/0/0/0.12</pre>	remote
26	Y	<pre>GigabitEthernet0/0/0/0.12</pre>	remote
27	Y	<pre>GigabitEthernet0/0/0/0.12</pre>	remote
28	Y	<pre>GigabitEthernet0/0/0/0.12</pre>	remote
29	Y	<pre>GigabitEthernet0/0/0/0.12</pre>	remote
30	Y	<pre>GigabitEthernet0/0/0/0.13</pre>	remote