OTUS

Дизайн сетей ЦОД

otus.ru

Защита проекта
Тема: Организация георасределенной отказоустойчивой СПД ЦОД с использованием технологии VXLAN

Такменев Андрей

План защиты проекта

- Цели проекта
- Используемые технологии
- Схема решения
- Оборудование решения
- Блок технической информации
- Масштабирование решения
- Вопросы

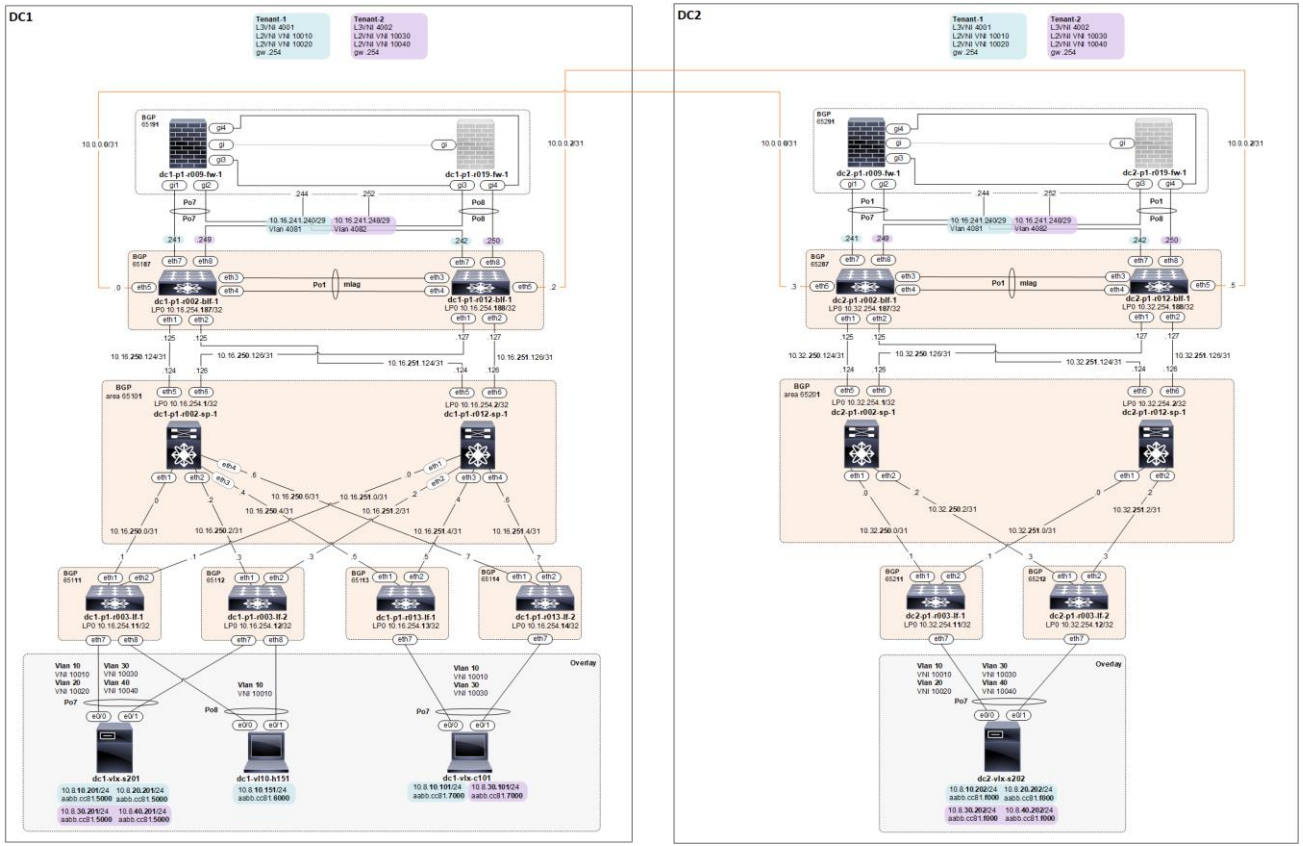
Цели проекта

1. Организация отказоустойчивой георасперределенной СПД ЦОД с использованием современной архитектуры
2. Обеспечение возможности взаимодействия между собой оконечного оборудования и сервисов площадок ЦОД на канальном и сетевом уровнях
3. Обеспечение возможности увеличения как количества площадок, так и количества подключаемого оконечного оборудования на каждой площадке

Используемые технологии

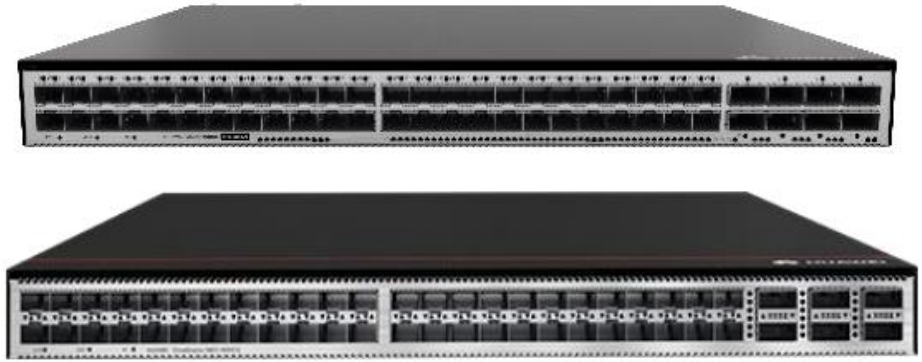
1. Протокол VXLAN (data plane) + BGP EVPN (control plane)
2. Протокол eBGP (реализация underlay, overlay, DCI)
3. Статические (leaf) и динамические (spine) peer group в BGP
4. Балансировка трафика (ECMP в BGP)
5. Уменьшение времени сходимости (bfd, таймеры bgp)
6. Технологии EVPN Multihoming (leaf), MLAG (border leaf)
7. Изоляция сегментов СПД ЦОД (tenant)
8. МЭ+VRF Lite + BGP AS-path prepend (взаимодействие между tenant)
9. Архитектура Multipod (взаимодействие между площадками)

Схема решения



Оборудование решения

Тип	Модель	Порты	Юниты
spine	Huawei CloudEngine 9860-4C-EI-A	128 x 100 GE QSFP28 or 32 x 400 GE QSFP-DD	4U
leaf	Huawei CloudEngine 6885-48YS8CQ	48 x 10/25 GE SFP28 or 48 x 50 GE SFP56 8 x 40/100 GE QSFP28 or 8 x 200 GE QSFP56	1U
leaf	Huawei CloudEngine 6863H-48S6CQ	48 x 10/25 GE SFP28 6 x 40/100 GE QSFP28	1U



Блок технической информации

Распределение ASN и наименование

Для случая 2xDC, 2xPOD или 4xDC, 1xPOD, leaf < 70 шт.

Тип	Номер AS	X,DC/POD	Y
sspine	65x00	1-4	-
spine	65x0y	1-4	1-8
leaf	65xyy	1-4	11-84
boleaf	65xyy	1-4	85-90
fw	65xyy	1-4	91-95
br	65xyy	1-4	96-99
host	646yy	1 DC/POD	0-99
host	647yy	2 DC/POD	0-99
host	648yy	3 DC/POD	0-99
host	649yy	4 DC/POD	0-99

Наимнование ACO определяется следующим образом

dcX-pX-rXXX-XX-X

Для остальных вариантов DC/POD или leaf > 70 шт.

		DC	POD	ТШ	Тип	Номер
AS	42	X	X	XXX	XX	X

Соответствие типа оборудования и его номера

Тип	Оборудование
0	host
1	leaf
2	spine
3	sspine
4	fw
5	-
6	-
7	-
8	-
9	br

Размещение оборудования в DC1

ТШ	Имя для ASN	Оборудование	Сокращение	ASN 4 байта	ASN 2 байта
2	dc1-p1-r002-02-1	dc1-p1-r002-sp-1	spine-1	4211002021	65101
12	dc1-p1-r012-02-1	dc1-p1-r012-sp-1	spine-2	4211012021	65101
3	dc1-p1-r003-01-1	dc1-p1-r003-lf-1	leaf-11	4211003011	65111
3	dc1-p1-r003-01-2	dc1-p1-r003-lf-2	leaf-12	4211003012	65112
13	dc1-p1-r013-01-1	dc1-p1-r013-lf-1	leaf-13	4211013011	65113
13	dc1-p1-r013-01-2	dc1-p1-r013-lf-2	leaf-14	4211013012	65114
2	dc1-p1-r002-01-1	dc1-p1-r002-blf-1	boleaf-187	4211002011	65187
12	dc1-p1-r012-01-1	dc1-p1-r012-blf-1	boleaf-188	4211012011	65188
9	dc1-p1-r009-04-1	dc1-p1-r009-fw-1	fw-1	4211009041	65191
19	dc1-p1-r019-04-1	dc1-p1-r019-fw-1	fw-2	4211019041	65191

Параметры VXLAN в Overlay-сети

В решении используется два tenant

VRF	Тип VNI	Номер VNI	Номер VLAN	Значение RT	Значение RD
tenant-1	L3VNI	4001	4001	4001:4001	RID:4001
tenant-1	L2VNI	10010	10	10010:10	RID:10
tenant-1	L2VNI	10020	20	10020:20	RID:20
tenant-2	L3VNI	4002	4002	4002:4002	RID:4002
tenant-2	L2VNI	10030	30	10010:30	RID:30
tenant-2	L2VNI	10040	40	10020:40	RID:40

Параметры VXLAN в Overlay-сети

В решении используется следующие параметры EVPN Multihoming

DC	Оборудование	Порт	ESI	ES-Import RT	LACP system-id
1	dc1-p1-r003-lf-1	Po7	0000:0101:0011:0007:0000	01:01:00:11:00:07	0101.0011.0007
1	dc1-p1-r003-lf-1	Po8	0000:0101:0011:0008:0000	01:01:00:11:00:08	0101.0011.0008
1	dc1-p1-r003-lf-2	Po7	0000:0101:0011:0007:0000	01:01:00:11:00:07	0101.0011.0007
1	dc1-p1-r003-lf-2	Po8	0000:0101:0011:0008:0000	01:01:00:11:00:08	0101.0011.0008
1	dc1-p1-r013-lf-1	Po7	0000:0101:0013:0007:0000	01:01:00:13:00:07	0101.0013.0007
1	dc1-p1-r013-lf-2	Po7	0000:0101:0013:0007:0000	01:01:00:13:00:07	0101.0013.0007
2	dc2-p1-r003-lf-1	Po7	0000:0201:0011:0007:0000	02:01:00:11:00:07	0201.0011.0007
2	dc2-p1-r003-lf-2	Po7	0000:0201:0011:0007:0000	02:01:00:11:00:07	0201.0011.0007

Выводы команд underlay

```
dc1-p1-r002-sp-1#show ip bgp summary
BGP summary information for VRF default
Router identifier 10.16.254.1, local AS number 65101
Neighbor Status Codes: m - Under maintenance
  Neighbor      V AS           MsgRcvd   MsgSent  InQ OutQ  Up/Down State   PfxRcd PfxAcc
  10.16.250.1   4 65111            699       677    0    0 00:24:29 Estab   1      1
  10.16.250.3   4 65112            712       675    0    0 00:24:30 Estab   1      1
  10.16.250.5   4 65113            697       684    0    0 00:24:29 Estab   1      1
  10.16.250.7   4 65114            699       689    0    0 00:24:29 Estab   1      1
  10.16.250.125 4 65187            702       685    0    0 00:24:29 Estab   8      8
  10.16.250.127 4 65187            684       651    0   19 00:24:29 Estab   8      8
```

Выводы команд underlay

dc1-p1-r002-sp-1#show ip bgp vrf all

	Network	Next Hop	Metric	AIGP	LocPref	Weight	Path
* >	**10.16**.254.1/32	-	-	-	-	0	i
* >	10.16.254.11/32	10.16.250.1	0	-	100	0	65111 i
* >	10.16.254.12/32	10.16.250.3	0	-	100	0	65112 i
* >	10.16.254.13/32	10.16.250.5	0	-	100	0	65113 i
* >	10.16.254.14/32	10.16.250.7	0	-	100	0	65114 i
* >Ec	10.16.254.187/32	10.16.250.127	0	-	100	0	65187 i
* ec	10.16.254.187/32	10.16.250.125	0	-	100	0	65187 i
* >Ec	10.16.254.188/32	10.16.250.127	0	-	100	0	65187 i
* ec	10.16.254.188/32	10.16.250.125	0	-	100	0	65187 i
* >Ec	10.32.254.1/32	10.16.250.127	0	-	100	0	65187 65287 65201 i
* ec	**10.32**.254.1/32	10.16.250.125	0	-	100	0	65187 65287 65201 i
* >Ec	10.32.254.2/32	10.16.250.127	0	-	100	0	65187 65287 65201 i
* ec	10.32.254.2/32	10.16.250.125	0	-	100	0	65187 65287 65201 i
* >Ec	10.32.254.11/32	10.16.250.127	0	-	100	0	65187 65287 65201 65211 i
* ec	10.32.254.11/32	10.16.250.125	0	-	100	0	65187 65287 65201 65211 i
* >Ec	10.32.254.12/32	10.16.250.127	0	-	100	0	65187 65287 65201 65212 i
* ec	10.32.254.12/32	10.16.250.125	0	-	100	0	65187 65287 65201 65212 i
* >Ec	10.32.254.187/32	10.16.250.127	0	-	100	0	65187 65287 i
* ec	10.32.254.187/32	10.16.250.125	0	-	100	0	65187 65287 i
* >Ec	10.32.254.188/32	10.16.250.127	0	-	100	0	65187 65287 i
* ec	10.32.254.188/32	10.16.250.125	0	-	100	0	65187 65287 i

Выводы команд overlay

```
dc1-p1-r003-lf-1#show vxlan vtep
Remote VTEPS for Vxlan1:

VTEP                 Tunnel Type(s)
-------------------- --------------
10.16.254.12         unicast, flood
10.16.254.13         unicast, flood
10.16.254.14         unicast, flood
10.16.254.187        unicast
10.16.254.188        unicast
10.32.254.11         unicast, flood
10.32.254.12         unicast, flood
10.32.254.187        unicast
10.32.254.188        unicast

Total number of remote VTEPS:  9
```

```
dc1-p1-r003-lf-1#show vxlan address-table
          Vxlan Mac Address Table
----------------------------------------------------------------------
VLAN  Mac Address     Type      Prt  VTEP             Moves   Last Move
----  -----------     ----      ---  ----             -----   ---------
  10  aabb.cc81.7000  EVPN      Vx1  10.16.254.13     2       0:28:38 ago
                                     10.16.254.14
  10  aabb.cc81.f000  EVPN      Vx1  10.32.254.11     1       0:28:37 ago
                                     10.32.254.12
  20  aabb.cc81.f000  EVPN      Vx1  10.32.254.11     1       0:28:37 ago
                                     10.32.254.12
  30  aabb.cc81.7000  EVPN      Vx1  10.16.254.13     2       0:28:38 ago
                                     10.16.254.14
  30  aabb.cc81.f000  EVPN      Vx1  10.32.254.11     1       0:28:37 ago
                                     10.32.254.12
  40  aabb.cc81.f000  EVPN      Vx1  10.32.254.11     1       0:28:37 ago
                                     10.32.254.12
4093  5000.0003.3766  EVPN      Vx1  10.16.254.13     1       2 days, 6:55:36 ago
4093  5000.0015.f4e8  EVPN      Vx1  10.16.254.14     1       2 days, 6:55:32 ago
4093  5000.0045.abdf  EVPN      Vx1  10.16.254.188    1       1 day, 2:34:21 ago
4093  5000.0068.a17f  EVPN      Vx1  10.32.254.188    1       21:43:45 ago
4093  5000.0088.fe27  EVPN      Vx1  10.16.254.187    1       1 day, 2:34:22 ago
4093  5000.00ba.c6f8  EVPN      Vx1  10.32.254.11     1       21:43:57 ago
4093  5000.00d5.5dc0  EVPN      Vx1  10.16.254.12     1       2 days, 6:55:36 ago
4093  5000.00d5.e2ad  EVPN      Vx1  10.32.254.187    1       21:44:00 ago
4093  5000.00d8.ac19  EVPN      Vx1  10.32.254.12     1       21:43:50 ago
...
Total Remote Mac Addresses for this criterion: 24
```

Выводы команд EVPN Mutihomig

```
dc1-p1-r003-lf-1#show bgp evpn instance
EVPN instance: VLAN 10
  Route distinguisher: 0:0
  Route target import: Route-Target-AS:10010:10
  Route target export: Route-Target-AS:10010:10
  Service interface: VLAN-based
  Local VXLAN IP address: 10.16.254.11
  VXLAN: enabled
  MPLS: disabled
  Local ethernet segment:
    ESI: 0000:0101:0011:0008:0000
      Interface: Port-Channel8
      Mode: all-active
      State: up
      ES-Import RT: 01:01:00:11:00:08
      DF election algorithm: modulus
      Designated forwarder: 10.16.254.11
      Non-Designated forwarder: 10.16.254.12
    ESI: 0000:0101:0011:0007:0000
      Interface: Port-Channel7
      Mode: all-active
      State: up
      ES-Import RT: 01:01:00:11:00:07
      DF election algorithm: modulus
      Designated forwarder: 10.16.254.11
      Non-Designated forwarder: 10.16.254.12
```

```
dc1-p1-r003-lf-1#show port-channel dense
…
Number of channels in use: 2
Number of aggregators: 2

   Port-Channel       Protocol    Ports
------------------ -------------- ---------
   Po7(U)             LACP(a)     Et7(PG+)
   Po8(U)             LACP(a)     Et8(PG+)
```

Выводы команд VRF

```
dc2-p1-r002-blf-1#show ip bgp summary vrf all
BGP summary information for VRF default
Router identifier 10.32.254.187, local AS number 65287
Neighbor Status Codes: m - Under maintenance
  Description              Neighbor      V AS           MsgRcvd   MsgSent  InQ OutQ  Up/Down State   PfxRcd PfxAcc
  ### dc1-p1-r002-blf-1 ## 10.0.0.0      4 65187          30580     30476    0    0 21:40:34 Estab   8      8
  ### dc2-p1-r012-blf-1 ## 10.32.241.1   4 65287          30431     30458    0   19 00:24:58 Estab   13     13
  ### dc2-p1-r002-sp-1 ### 10.32.250.124 4 65201          30640     30629    0    0 21:40:33 Estab   3      3
  ### dc2-p1-r012-sp-1 ### 10.32.251.124 4 65201          30625     30649    0    0 21:40:33 Estab   3      3

BGP summary information for VRF tenant-1
Router identifier 10.32.241.241, local AS number 65287
Neighbor Status Codes: m - Under maintenance
  Description              Neighbor      V AS           MsgRcvd   MsgSent  InQ OutQ  Up/Down State   PfxRcd PfxAcc
  ### dc2-p1-r009-fw-1 ### 10.32.241.244 4 65291          24585     29619    0   19 00:24:57 Estab   1      1

BGP summary information for VRF tenant-2
Router identifier 10.32.241.249, local AS number 65287
Neighbor Status Codes: m - Under maintenance
  Description              Neighbor      V AS           MsgRcvd   MsgSent  InQ OutQ  Up/Down State   PfxRcd PfxAcc
  ### dc2-p1-r009-fw-1 ### 10.32.241.252 4 65291          24584     29618    0   38 00:24:57 Estab   1      1
```

Выводы команд VRF

```
dc2-p1-r002-blf-1#show ip bgp vrf all
BGP routing table information for VRF tenant-1
Router identifier 10.32.241.241, local AS number 65287
          Network                Next Hop              Metric  AIGP       LocPref Weight  Path
 * >      10.8.0.0/16            10.16.254.187         0       -          100     0       65187 65191 i
 *        10.8.0.0/16            10.32.241.244         0       -          100     0       65291 65291 65291 65291 i
 * >Ec    10.8.10.0/24           10.32.254.11          0       -          100     0       65201 65211 i
 *  ec    10.8.10.0/24           10.32.254.12          0       -          100     0       65201 65212 i
 *  ec    10.8.10.0/24           10.32.254.12          0       -          100     0       65201 65212 i
 *  ec    10.8.10.0/24           10.32.254.11          0       -          100     0       65201 65211 i
 *  Ec    10.8.10.0/24           10.16.254.11          0       -          100     0       65187 65101 65111 i
 *  ec    10.8.10.0/24           10.16.254.12          0       -          100     0       65187 65101 65112 i
 *  ec    10.8.10.0/24           10.16.254.14          0       -          100     0       65187 65101 65114 i
 *  ec    10.8.10.0/24           10.16.254.13          0       -          100     0       65187 65101 65113 i
 * >Ec    10.8.10.101/32         10.16.254.13          0       -          100     0       65187 65101 65113 i
 *  ec    10.8.10.101/32         10.16.254.14          0       -          100     0       65187 65101 65114 i
 * >Ec    10.8.10.151/32         10.16.254.11          0       -          100     0       65187 65101 65111 i
 *  ec    10.8.10.151/32         10.16.254.12          0       -          100     0       65187 65101 65112 i
 * >Ec    10.8.10.201/32         10.16.254.11          0       -          100     0       65187 65101 65111 i
 *  ec    10.8.10.201/32         10.16.254.12          0       -          100     0       65187 65101 65112 i
 * >Ec    10.8.10.202/32         10.32.254.11          0       -          100     0       65201 65211 i
 *  ec    10.8.10.202/32         10.32.254.11          0       -          100     0       65201 65211 i
 *  ec    10.8.10.202/32         10.32.254.12          0       -          100     0       65201 65212 i
 *  ec    10.8.10.202/32         10.32.254.12          0       -          100     0       65201 65212 I
          …
 * >      10.16.241.240/29       10.16.254.187         0       -          100     0       65187 i
 * >      10.32.241.240/29       -                     -       -          -       0       i
```

Выводы команд DCI

```
dc1-p1-r002-blf-1#show ip bgp summary
BGP summary information for VRF default
Router identifier 10.16.254.187, local AS number 65187
Neighbor Status Codes: m - Under maintenance
  Description              Neighbor      V AS      MsgRcvd   MsgSent  InQ OutQ  Up/Down State   PfxRcd PfxAcc
  ### dc2-p1-r002-blf-1 ## 10.0.0.1      4 65287     33123     33259    0    0 21:39:50 Estab   6      6
  ### dc1-p1-r012-blf-1 ## 10.16.241.1   4 65187     37311     37299    0   38 22:40:51 Estab   13     13
  ### dc1-p1-r002-sp-1 ### 10.16.250.124 4 65101     37610     37753    0    0 00:24:29 Estab   5      5
  ### dc1-p1-r012-sp-1 ### 10.16.251.124 4 65101     37593     37713    0    0 00:23:32 Estab   5      5

dc2-p1-r012-blf-1#show ip bgp summary
BGP summary information for VRF default
Router identifier 10.32.254.188, local AS number 65287
Neighbor Status Codes: m - Under maintenance
  Description              Neighbor      V AS       MsgRcvd   MsgSent  InQ OutQ  Up/Down State   PfxRcd PfxAcc
  ### dc1-p1-r012-blf-1 ## 10.0.0.2      4 65187      30531     30446    0   38 21:39:42 Estab   8      8
  ### dc2-p1-r002-blf-1 ## 10.32.241.0   4 65287      30417     30415    0   19 00:24:13 Estab   13     13
  ### dc2-p1-r002-sp-1 ### 10.32.250.126 4 65201      30591     30561    0   19 21:39:43 Estab   3      3
  ### dc2-p1-r012-sp-1 ### 10.32.251.126 4 65201      30731     30727    0   19 00:27:45 Estab   3      3
```

Выводы команд DCI

```
dc1-p1-r013-lf-2#show ip route vrf tenant-1
VRF: tenant-1
...
Gateway of last resort is not set
 B E      10.8.10.151/32 [20/0] via VTEP 10.16.254.12 VNI 4001 router-mac 50:00:00:d5:5d:c0 local-interface Vxlan1
                                via VTEP 10.16.254.11 VNI 4001 router-mac 50:00:00:72:8b:31 local-interface Vxlan1
 B E      10.8.10.201/32 [20/0] via VTEP 10.16.254.12 VNI 4001 router-mac 50:00:00:d5:5d:c0 local-interface Vxlan1
                                via VTEP 10.16.254.11 VNI 4001 router-mac 50:00:00:72:8b:31 local-interface Vxlan1
 B E      10.8.10.202/32 [20/0] via VTEP 10.32.254.11 VNI 4001 router-mac 50:00:00:ba:c6:f8 local-interface Vxlan1
                                via VTEP 10.32.254.12 VNI 4001 router-mac 50:00:00:d8:ac:19 local-interface Vxlan1
 C        10.8.10.0/24 is directly connected, Vlan10
 B E      10.8.20.201/32 [20/0] via VTEP 10.16.254.12 VNI 4001 router-mac 50:00:00:d5:5d:c0 local-interface Vxlan1
                                via VTEP 10.16.254.11 VNI 4001 router-mac 50:00:00:72:8b:31 local-interface Vxlan1
 B E      10.8.20.202/32 [20/0] via VTEP 10.32.254.11 VNI 4001 router-mac 50:00:00:ba:c6:f8 local-interface Vxlan1
                                via VTEP 10.32.254.12 VNI 4001 router-mac 50:00:00:d8:ac:19 local-interface Vxlan1
 B E      10.8.20.0/24 [20/0] via VTEP 10.16.254.12 VNI 4001 router-mac 50:00:00:d5:5d:c0 local-interface Vxlan1
                              via VTEP 10.16.254.11 VNI 4001 router-mac 50:00:00:72:8b:31 local-interface Vxlan1
 B E      10.8.0.0/16 [20/0] via VTEP 10.16.254.187 VNI 4001 router-mac 50:00:00:88:fe:27 local-interface Vxlan1
                             via VTEP 10.16.254.188 VNI 4001 router-mac 50:00:00:45:ab:df local-interface Vxlan1
 B E      10.16.241.240/29 [20/0] via VTEP 10.16.254.187 VNI 4001 router-mac 50:00:00:88:fe:27 local-interface Vxlan1
                                  via VTEP 10.16.254.188 VNI 4001 router-mac 50:00:00:45:ab:df local-interface Vxlan1
 B E      10.32.241.240/29 [20/0] via VTEP 10.32.254.187 VNI 4001 router-mac 50:00:00:d5:e2:ad local-interface Vxlan1
                                  via VTEP 10.32.254.188 VNI 4001 router-mac 50:00:00:68:a1:7f local-interface Vxlan1
```

Выводы команд трассировки

- tracert из из **DC-1** с dc1-vl10-h151 в другой tenant-2 (vlan 40) идет через МЭ DC1 (**10.16**)
- tracert из **DC-2** с dc2-vl30-s202 в другой tenant-1 (vlan 10) идет через МЭ DC1 (**10.16**)

```
dc1-vl10-h151#traceroute 10.8.40.202
Type escape sequence to abort.
Tracing the route to 10.8.40.202
VRF info: (vrf in name/id, vrf out name/id)
  1 10.8.10.254 28 msec 6 msec 7 msec
  2 10.16.241.242 134 msec 55 msec 36 msec
  3 10.16.241.244 171 msec 158 msec 126 msec
  4 10.16.241.249 431 msec 599 msec 252 msec
  5 10.8.30.254 268 msec 450 msec 265 msec
  6 10.8.40.202 265 msec *  392 msec
```

```
dc2-vlx-s202#traceroute vrf vlan30 10.8.10.101
Type escape sequence to abort.
Tracing the route to 10.8.10.101
VRF info: (vrf in name/id, vrf out name/id)
  1 10.8.30.254 34 msec 9 msec 9 msec
  2 10.16.241.249 70 msec 97 msec 72 msec
  3 10.16.241.252 136 msec 138 msec 290 msec
  4 10.16.241.241 501 msec 228 msec 373 msec
  5 10.8.10.254 362 msec 175 msec 401 msec
  6 10.8.10.101 267 msec *  271 msec
```

Масштабирование решения

1. Увеличение числа DC до 4 шт. или POD до 2 шт. в каждом DC (2 байтные ASN)
2. Увеличение числа DC до 8 шт., POD до 4 шт. в кажом DC (4 байтные ASN)
3. Увеличение числа spine в каждом POD до 6-8 шт. (количество uplink-портов на leaf)
4. Увеличение числа leaf в каждом POD до 70 шт. с использованием 2 байтных ASN
5. Увеличение числа leaf в каждом POD до 127 шт. с использованием 4 байтных ASN с учетом: :
   - портовой емкости spine (128 портов)
   - размера транспортного сегмента (сеть с маской /25)
   - физических ограничений по размещению leaf (1U) и spine (4U)
6. Увеличение числа tenant до 90 шт
7. Увеличение числа разделяемых сетевых сегментов до исчерпания блока 10.8.0.0/14

Спасибо за внимание!

Готов ответить на ваши вопросы