CHAPTER 23

VXLAN

この章では、Virtual Extensible Local Area Network（VXLAN）の実装時に発生する可能性がある問題を識別して解決する方法について説明します。

この章は、次の内容で構成されています。

VXLAN に関する情報

概要

VXLAN は、IP カプセル化で MAC とオーバーレイ アプローチを使用することによって、LAN セグメントを作成します。カプセル化は、仮想イーサネット モジュール（VEM）内で、カプセル化される元の仮想マシン（VM）からオリジナル レイヤ 2（L2）フレームを伝送します。ネットワークで送信する MAC フレームをカプセル化する際に、送信元 IP アドレスとして使用する IP アドレスが各 VEM に割り当てられます。このカプセル化されたトラフィックの送信元として使用する VEM ごとに、複数の vmknics を設定できます。カプセル化は、ペイロード フレームの MAC アドレスのスコープに使用される VXLAN 識別子を伝送します。

接続された VXLAN は vNIC のポート プロファイル コンフィギュレーション内で指定され、VM の接続時に適用されます。各 VXLAN は、割り当てられた IP マルチキャスト グループを VXLAN セグメント内でのブロードキャスト トラフィックの伝送に使用します。

VM が VEM に接続する際、VEM 上の特定の VXLAN セグメントに参加する 1 番目の VM の場合、IGMP の参加は VXLAN の割り当てられたマルチキャスト グループに対して発行されます。VM がネットワーク セグメントのパケットを送信する際、ルックアップはフレームの宛先 MAC および

VXLAN 識別子を使用して L2 テーブルで行われます。結果がヒットの場合、L2 テーブル エントリには、フレームをカプセル化するようにリモート IP アドレスが含まれ、フレームは、リモート IP アドレス宛ての IP パケットに格納されて送信されます。結果が失敗（ブロードキャスト/マルチキャストまたは不明なユニキャストは、このバケットになります）の場合、フレームはセグメントの割り当てられた IP マルチキャスト グループに設定される宛先 IP アドレスでカプセル化されます。

カプセル化されたパケットをネットワークから受信すると、カプセル化が解除され、内部フレームおよび VXLAN ID の送信元 MAC アドレスがルックアップ キーとして L2 テーブルに追加され、カプセル化ヘッダーの送信元 IP アドレスは、テーブル エントリのリモート IP アドレスとして追加されます。

VXLAN の VEM L3 IP インターフェイス

VXLAN に接続している vEthernet インターフェイスが VEM にある場合、VEM は少なくとも 1 つの IP/MAC ペアで VXLAN パケットを終端させる必要があります。このため、VEM は IP ホストとして機能します。この目的で、VEM によりサポートされるのは IPv4 アドレッシングだけです。

VEM レイヤ 3（L3）コントロールの設定方法と同様に、VXLAN に使用する IP アドレスは、**capability vxlan** コマンドを持つ vmknic にポート プロファイルを割り当てることによって設定されます。

vPC-HM MAC-Pinning が必要なサーバ コンフィギュレーションで複数のアップリンク、またはサブグループ上の VXLAN トラフィックの伝送をサポートするには、最大 4 つの vmknics と **capability vxlan** を設定できます。同じ ESX/ESXi ホスト内のすべての VXLAN vmknics は、**capability vxlan** パラメータを持つ必要がある同じポート プロファイルに割り当てることを推奨します。

ローカル vEthernet インターフェイスによってソースされた VXLAN トラフィックは、フレームの送信元 MAC アドレスに基づいてこれらの vmknics 間で分散されます。VEM は自動的に複数の VXLAN vmknics を別々のアップリンクにピン接続します。アップリンクに障害が発生すると、VEM は自動的に vmknic を動作アップリンクに再ピン接続します。

カプセル化されたトラフィックが異なるサブネットに接続されている VEM に送られる際、VEM は VMware ホストのルーティング テーブルを使用しません。代わりに、vmknic はリモート VEM IP アドレスに対して ARP を開始します。アップストリーム ルータは、プロキシ ARP 機能を使用して応答するように設定する必要があります。

フラグメンテーション

VXLAN カプセル化のオーバーヘッドは 50 バイトです。フラグメンテーションによるパフォーマンスの低下を回避するには、VXLAN パケットを交換するすべての VEM 間のインターコネクト インフラストラクチャ全体を、VM VNIC が送信するように設定されているよりも 50 バイト多く伝送するように設定する必要があります。たとえば、デフォルト VNIC 設定の 1500 バイトを使用している場合、VEM アップリンク ポート プロファイル、アップストリーム物理スイッチ ポート、およびスイッチ間リンクとルータ（存在する場合）は、少なくとも 1550 バイトの MTU を伝送するように設定する必要があります。これが不可能な場合、ゲスト VM 内の MTU を 50 バイト小さく、たとえば、1450 バイトに設定すること推奨します。

これが設定されていない場合、VEM は Path MTU（PMTU）Discovery を実行するかどうかを VM に通知しようとします。VM が小さい MTU でパケットを送信しない場合、VM は IP パケットをフラグメント化します。フラグメンテーションは、IP レイヤだけで行われます。伝送に大きすぎるフレームを VM が送信する場合、VXLAN カプセル化を追加した後、およびフレームに IP パケットを含まない場合、フレームはドロップされます。

スケーラビリティ

VXLAN の最大数

Cisco Nexus 1000V は、合計 2048 の VLAN または VXLAN または 2048 を超えない任意の組み合わせをサポートします。この番号は、Cisco Nexus 1000V のポートの最大数と一致します。これにより、各ポートが異なる VLAN または VXLAN に接続できるようにします。

サポートされる機能

ここでは、次の項目について説明します。

ジャンボ フレーム

ジャンボ フレームは、VXLAN カプセル化のオーバーヘッドに対応するための空き容量が少なくとも 50 バイトあり、物理スイッチ/ルータ インフラストラクチャがこれらのジャンボ サイズの IP パケットを転送できる限り、でサポートされます。

VXLAN 機能のグローバルなディセーブル化

安全策として、**no feature segmentation** コマンドは、VXLAN のポート プロファイルに関連付けられたポートがある場合は許可されません。この機能をディセーブルにする前に、すべてのアソシエーションを削除する必要があります。**no feature segmentation** コマンドは、上のすべての VXLAN ブリッジドメイン設定をクリーンアップします。

VXLAN のトラブルシューティング コマンド

VXLAN 属性を表示するには、次のコマンドを使用します。

ここでは、次の項目について説明します。

VSM コマンド

特定のセグメントに属しているポートを表示する方法

```
switch(config)# show system internal seg_bd info segment 10000
Bridge-domain: A
Port Count: 11
Veth1
Veth2
Veth3
```

vEthernet ブリッジ ドメインの設定を表示する方法

```
switch(config)# show system internal seg_bd info port vethernet 1
Bridge-domain: A
segment_id = 10000
Group IP: 225.1.1.1
```

vEthernet ブリッジ設定と ifindex を引数として表示する方法

```
switch(config)# show system internal seg_bd info port ifindex 0x1c000050
Bridge-domain: A
segment_id = 10000
Group IP: 225.1.1.1
```

ブリッジ ドメイン ポートの合計数を表示する方法

```
switch(config)# show system internal seg_bd info port_count
Number of ports: 11
```

ブリッジ ドメイン内部コンフィギュレーションを表示する方法

```
switch(config)# show system internal seg_bd info bd vxlan-home

Bridge-domain vxlan-home (2 ports in all)
Segment ID: 5555 (Manual/Active)
Group IP: 235.5.5.5
State: UP               Mac learning: Enabled
is_bd_created: Yes
current state: SEG_BD_FSM_ST_READY
pending_delete: 0
port_count: 2
action: 4
hwbd: 28
pa_count: 0
Veth2, Veth5
switch(config)#
```

VXLAN vEthernet インターフェイスの情報を表示する方法

```
switch# show system internal seg_bd info port
if_index = <0x1c000010>
Bridge-domain vxlan-pepsi
rid = 216172786878513168
swbd = 4098

if_index = <0x1c000040>
Bridge-domain vxlan-pepsi
rid = 216172786878513216
swbd = 4098

switch#
```

追加の **show** コマンド：

show system internal seg_bd info {pss | sdb | global | all}

show system internal seg_bd {event-history | errors | mem-stats | msgs}

VEM コマンド

VXLAN vEthernet プログラミングの確認手順

```
~ # vemcmd show port segments
                           Native  Seg
  LTL   VSM Port  Mode     SegID   State
   50       Veth5   A        5555   FWD
   51       Veth9   A        8888   FWD
~ #
```

VXLAN vmknic プログラミングの確認手順

```
~ # vemcmd show vxlan interfaces
LTL            IP       Seconds since Last
                        IGMP Query Received
(* Interface on which IGMP Joins are sent)
----------------------------------------
 49         10.3.3.3         50          *
 52         10.3.3.6         50
~ #
```

vmknics が適切なトランスポート VLAN にあるかどうか確認するには、“vemcmd show port vlans” を使用してください。

VEM ブリッジ ドメインの作成の確認手順

```
~ # vemcmd show bd  bd-name vxlan-home
BD 31, vdc 1, segment id 5555, segment group IP 235.5.5.5, swbd 4098, 1 ports,
"vxlan-home"
Portlist:
     50  RedHat_VM1.eth0

~ #
```

リモート IP の学習の確認手順

```
~ # vemcmd show l2 bd-name vxlan-home
Bridge domain   31 brtmax 4096, brtcnt 2, timeout 300
Segment ID 5555, swbd 4098, "vxlan-home"
Flags:  P - PVLAN  S - Secure  D - Drop
       Type          MAC Address   LTL   timeout   Flags    PVLAN    Remote IP
    Dynamic   00:50:56:ad:71:4e   305         2                      10.3.3.100
     Static   00:50:56:85:01:5b    50         0                       0.0.0.0

~ #
```

統計情報を表示する方法

```
~ # vemcmd show vxlan-stats
  LTL  Ucast   Mcast   Ucast   Mcast    Total
       Encaps  Encaps  Decaps  Decaps   Drops
   49       5   14265       4      15       0
   50       6   14261       4      15     213
   51       1      15       0       0      10
   52       0      11       0       0      15

~ #
```

VXLAN vEthernet/vmknic のポート単位の詳細な統計情報を表示する方法

```
~ # vemcmd show vxlan-stats ltl 51
```

すべてのブリッジ ドメインの VXLAN vmknic のポート単位ブリッジ単位の詳細なドメイン統計情報を表示する方法

```
~ # vemcmd show vxlan-stats ltl <vxlan_vmknic_ltl> bd-all
```

指定したブリッジ ドメインの VXLAN vmknic のポート単位ブリッジ単位の詳細なドメイン統計情報を表示する方法

```
~ # vemcmd show vxlan-stats ltl vxlan_vmknic_ltl bd-name bd-name
```

VEM パケット パスのデバッグ

VEM1 の VM から VEM2 の VM への VXLAN トラフィックをデバッグするには、次のコマンドを使用します。

- VEM1：パケットがセグメント vEthernet からスイッチに着信していることを確認します。

  ```
  vempkt capture ingress ltl vxlan_veth
  ```

- VEM1：VXLAN のカプセル化を確認します。

  ```
  vemlog debug sflisp all
  vemlog debug sfvnsegment all
  ```

- VEM1：リモート IP が学習されたことを確認します。

  ```
  vemcmd show l2 bd-name segbdname
  ```

  リモート IP が学習されていない場合、パケットはマルチキャストにカプセル化されて送信されます。たとえば、VM からの最初の ARP 要求はこの方法で送信されます。

- VEM1：カプセル化されたパケットがアップリンク送信されることを確認します。

  どのアップリンクが使用されるかを調べるには、**vemcmd show vxlan-encap ltl** *ltl* コマンドまたは **vemcmd show l2lisp-encap mac** *mac* を使用します。

  ```
  vempkt capture egress ltl uplink
  ```

- VEM1：任意の障害の統計情報を確認します。

  ```
  vemcmd show vxlan-stats all
  vemcmd show vxlan-stats ltl veth/vxlanvmknic
  ```

- VEM2：カプセル化されたパケットがアップリンクに到達していることを確認します。

  ```
  vempkt capture ingress ltl uplink
  ```

- VEM2：VXLAN カプセル開放を確認します。

  ```
  "vemlog debug sflisp all"
  "vemlog debug sfvnsegment all"
  ```

- VEM2：VXLAN vEthernet で送信されるカプセル解放されたパケットを確認します。

  ```
  vempkt capture egress ltl vxlan_veth
  ```

- VEM2：任意の障害の統計情報を確認します。

  ```
  vemcmd show vxlan-stats all
  vemcmd show vxlan-stats ltl veth/vxlanvmknic
  ```

VEM マルチキャストのデバッグ

VEM マルチキャストをデバッグするには、次のコマンドを使用します。

- VEM の IGMP ステート：

  vemcmd show igmp *vxlan_transport_vlan* **detail**

（注）　このコマンドは、セグメント マルチキャスト グループの出力を表示しません。マルチキャスト テーブル スペースを節約するため、セグメントのグループは、VEM 上の IGMP スヌーピングによって追跡されません。

- IGMP クエリー：

vemcmd show vxlan interfaces コマンドを使用して、IGMP クエリーが受信されていることを確認します。

- vmknic からの IGMP Join:

VMware スタックが Join を送信しているかどうかを確認するには、**vempkt capture ingress ltl** *first_vxlan_vmknic_ltl* コマンドを使用します。

Join がアップストリーム ポートに送信されているかどうかを確認するには、**vempkt capture egress ltl** *uplink_ltl* コマンドを使用します。

VXLAN Datapath のデバッグ

ここに挙げるコマンドは、VXLAN 関連の問題のトラブルシューティングに使用できるものです。

ここでは、次の項目について説明します。

Vemlog デバッグ

ブリッジ ドメインのセットアップまたは設定をデバッグするには、次のコマンドを使用します。

```
vemlog debug sfbd all
```

ポート設定/CBL/vEthernet LTL ピン接続をデバッグするには、次のコマンドを使用します。

```
vemlog debug sfporttable all
```

（encap/decap の設定と決定）

```
vemlog debug sfvnsegment all
```

実際のパケット編集、VXLAN のインターフェイス処理、およびマルチキャストの処理をデバッグするには、次のコマンドを使用します。

```
vemlog debug sflisp all
```

DPA ソケットのマルチキャストの出入りをデバッグするには、次のコマンドを使用します。

echo "debug dpa_allplatform all" > /tmp/dpafifo

ブリッジ ドメイン設定をデバッグするには、次のコマンドを使用します。

echo "debug sfl2agent all" > /tmp/dpafifo

ポート設定をデバッグするには、次のコマンドを使用します。

echo "debug sfportagent all" > /tmp/dpafifo

capability l2-lisp のヒットレス再接続（HR）をデバッグするには、次のコマンドを使用します。

echo "debug sfportl2lisp_cache all" > /tmp/dpafifo

CBL のプログラミングをデバッグします。

echo "debug sfpixmagent all" > /tmp/dpafifo

HR

HR のセグメント情報をデバッグするには、次のコマンドを使用します。

echo "debug sfsegment_cache all" > /tmp/dpafifo (to debug segment info HR)

（キャッシュされたおよび一時セグメント情報リストの詳細がある場合）

echo "show vsm cache *vsm control mac*" **> /tmp/dpafifo**

Vempkt

Vempkt で VLAN/SegmentID が表示されるようになりました。VEM を通過するパケット パスを追跡するには、vempkt を使用します。

- Encap : Seg-VEth LTL の入力 – アップリンクの出力をキャプチャします。
- Decap : アップリンクの入力 – Seg-VEth LTL の出力をキャプチャします。

統計情報

ポート単位の統計情報の要約を表示するには、次のコマンドを使用します。

vemcmd show vxlan-stats

VXLAN vmknic のポート単位の詳細な統計情報を表示するには、次のコマンドを使用します。

vemcmd show vxlan-stats ltl *vxlan_vmknic_ltl*

VXLAN の vEthernet のポート単位の詳細な統計情報を表示するには、次のコマンドを使用します。

vemcmd show vxlan-stats ltl *vxlan_veth_ltl*

すべてのブリッジ ドメインの VXLAN vmknic のポート単位ブリッジ単位の詳細なドメイン統計情報を表示するには、次のコマンドを使用します。

vemcmd show vxlan-stats ltl *vxlan_vmknic_ltl* **bd-all**

指定したブリッジ ドメインの VXLAN vmknic のポート単位ブリッジ単位の詳細なドメイン統計情報を表示するには、次のコマンドを使用します。

vemcmd show vxlan-stats ltl *vxlan_vmknic_ltl* **bd-name** *bd-name*

ポート上で学習したスタティック MAC について、カプセル化およびその後のアップリンク PC へのピン接続に使用する VXLAN vmknic を表示するには、次のコマンドを使用します。

vemcmd show vxlan-encap ltl *vxlan_veth_ltl*

カプセル化およびその後のアップリンク PC へのピン接続に使用する VXLAN vmknic を表示するには、次のコマンドを使用します。

vemcmd show vxlan-encap mac *vxlan_vm_mac*

show コマンド

表 23-1 は、使用可能な **vemcmd show** コマンドのリストです。

表 23-1　vemcmd Show コマンド

コマンド	結果
vemcmd show vxlan interfaces	VXLAN カプセル化インターフェイスを表示します。
vemcmd show port vlans	ブリッジ ドメインのポート プログラミングと CBL 状態を確認します。
vemcmd show bd	ポートの SegmentID/ グループ / リストを表示します。
vemcmd show bd bd-name *bd-name-string*	1 セグメント ブリッジ ドメインを表示します。
vemcmd show l2 all	学習されているリモート IP を表示します。
vemcmd show l2 bd-name *bd-name-string*	1 セグメント ブリッジ ドメインのレイヤ 2 テーブルを表示します。
vemcmd show arp all	外部でカプセル化されたヘッダーの IP-MAC マッピングを表示します。