



# Linux での RoCEv2 を使用した NVMeoF の構成

- [Linux 上で RoCE v2 を使用するファブリック \(NVMeoF\) を介して NVMe を使用する際のガイドライン \(1 ページ\)](#)
- [Linux の要件 \(2 ページ\)](#)
- [Cisco Intersight での RoCE v2 for NVMeoF の構成 \(2 ページ\)](#)
- [ホストシステムでの NVMeoF の RoCE v2 の構成 \(7 ページ\)](#)
- [デバイス マッパー マルチパスの設定 \(11 ページ\)](#)
- [Cisco Intersight を使用した RoCE v2 インターフェイスの削除 \(12 ページ\)](#)

## Linux 上で RoCE v2 を使用するファブリック (NVMeoF) を介して NVMe を使用する際のガイドライン

### 一般的なガイドラインと制限事項

- Cisco では、[UCS ハードウェアとソフトウェアの互換性](#)をチェックして、NVMeoF のサポートを判断することを推奨します。NVMeoF は、Cisco UCS B シリーズ、C シリーズ、および X シリーズのサーバでサポートされています。
- RoCE v2 を使用した RDMA 上の NVMe は、Cisco UCS VIC 1400、VIC 14000、および VIC 15000 シリーズのアダプタでサポートされています。
- RoCE v2 インターフェイスを作成する際には、Cisco Intersight が提供する Linux-NVMe-RoCE アダプタ ポリシーを使用します。
- Ethernet Adapter ポリシーでは、キューペア、メモリ領域、リソースグループ、および優先度の設定値を、Cisco が提供するデフォルト値以外に変更しないでください。キューペア、メモリ領域、リソースグループ、および優先度の設定が異なると、NVMeoF の機能が保証されない可能性があります。
- RoCE v2 インターフェイスを構成する場合は、Cisco.com からダウンロードした `enic` と `enic_rdma` の両方のバイナリドライバを使用して、一致する `enic` と `enic_rdma` ドライバの

セットをインストールします。inbox enic ドライバを使用して Cisco.com からダウンロードしたバイナリ enic\_rdma ドライバを使用しようとしても、機能しません。

- RoCE v2 は、アダプタごとに最大 2 つの RoCE v2 対応インターフェイスをサポートしません。
- NVMeoF ネームスペースからのブートはサポートされていません。
- レイヤ 3 ルーティングはサポートされていません。
- RoCE v2 はボンディングをサポートしていません。
- システムクラッシュ時に crashdump を NVMeoF ネームスペースに保存することはサポートされていません。
- NVMeoF は、usNIC、VxLAN、VMQ、VMMQ、NVGRE、GENEVE オフロード、および DPDK 機能とともに使用することはできません。
- Cisco Intersight は、RoCE v2 対応の vNIC に対してファブリック フェールオーバーをサポートしません。
- Quality of Service (QoS) no drop クラス構成は、Cisco Nexus 9000 シリーズスイッチなどのアップストリームスイッチで適切に構成する必要があります。QoS の設定は、異なるアップストリームスイッチ間で異なります。
- スパニング ツリー プロトコル (STP) によって、フェールオーバーまたはフェールバック イベントが発生したときに、ネットワーク接続が一時的に失われる可能性があります。この問題が発生しないようにするには、アップリンクスイッチで STP を無効にします。

## Linux の要件

Linux での RoCE v2 の構成と使用には、次のものがが必要です。

- InfiniBand カーネル API モジュール ib\_core
- NVMeoF 接続をサポートするストレージアレイ

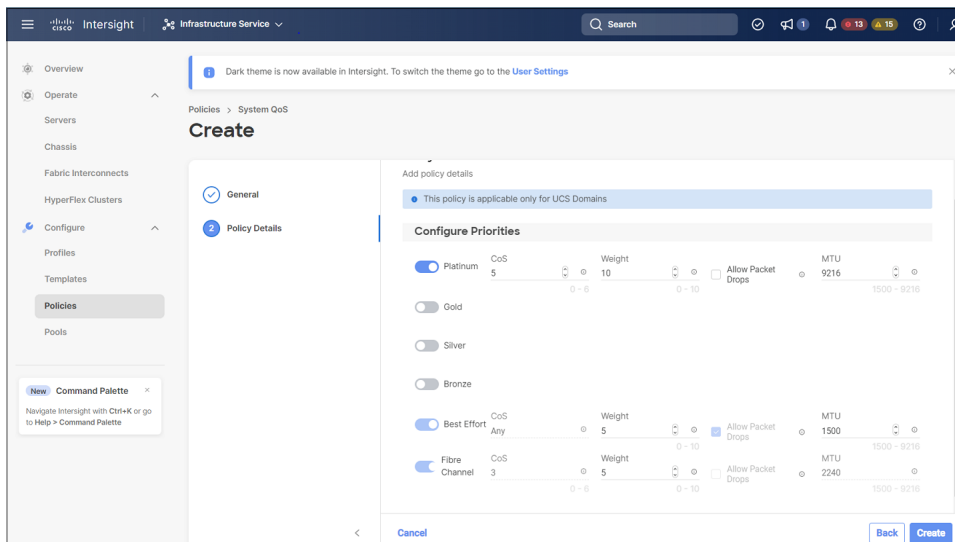
## Cisco Intersight での RoCE v2 for NVMeoF の構成

Cisco Intersight で RoCE v2 インターフェイスを構成するには、次の手順に従います。

RDMA パケット ドロップの可能性を回避するには、ネットワーク全体で同じ非ドロップ COS が構成されていることを確認してください。次の手順に従えば、システム QoS ポリシーで非ドロップクラスを構成して、RDMA でサポートされているインターフェイス用に使用できます。

## 手順

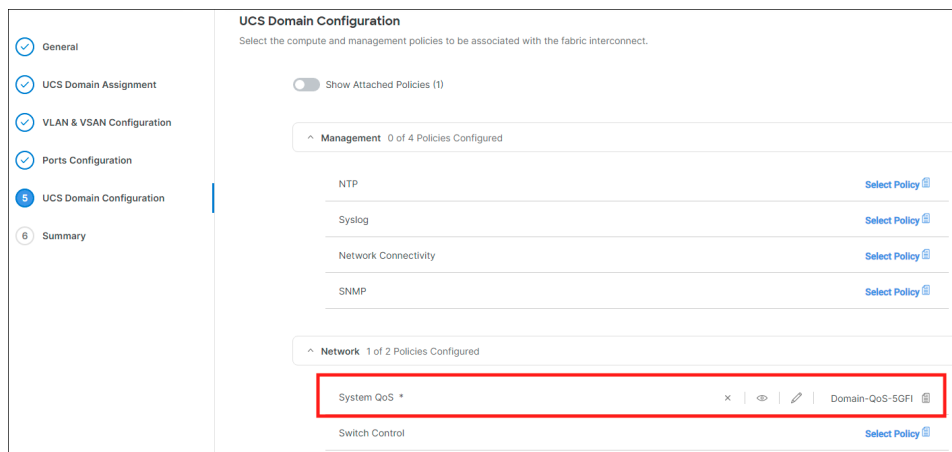
- ステップ 1 [構成 (CONFIGURE)] > [ポリシー (Policies)] に移動します。[ポリシーの作成 (Create Policy)] をクリックし、[UCS ドメイン (UCS Domain)] プラットフォームタイプを選択し、[システム QoS (System QoS)] を検索または選択して、[Start (開始)] をクリックします。
- ステップ 2 [全般 (General)] ページでポリシー名を入力し、[次へ (Next)] をクリックします。次に、[ポリシーの詳細 (Policy Details)] ページで、次のようにシステム QoS ポリシーのプロパティ設定を構成します。
- [優先順位 (Priority)] で、[プラチナ (Platinum)] を選択します。
  - [パケットドロップを許可 (Allow Packet Drops)] チェックボックスをオフにします。
  - [MTU] については、値を 9216 に設定します。



ステップ 3 [作成 (Create)] をクリックします。

ステップ 4 システム QoS ポリシーをドメインプロファイルに関連付けます。

## LAN 接続ポリシーで RoCE 設定を有効化する



(注)

詳細については、「ドメインポリシーの構成」の「システム QoS ポリシーの作成」および「ドメインプロファイルの構成」を参照してください。

システム QoS ポリシーが正常に作成され、ドメインプロファイルに展開されます。

#### 次のタスク

LAN 接続ポリシーで RoCE v2 vNIC 設定を使用してサーバプロファイルを構成します。

## LAN 接続ポリシーで RoCE 設定を有効化する

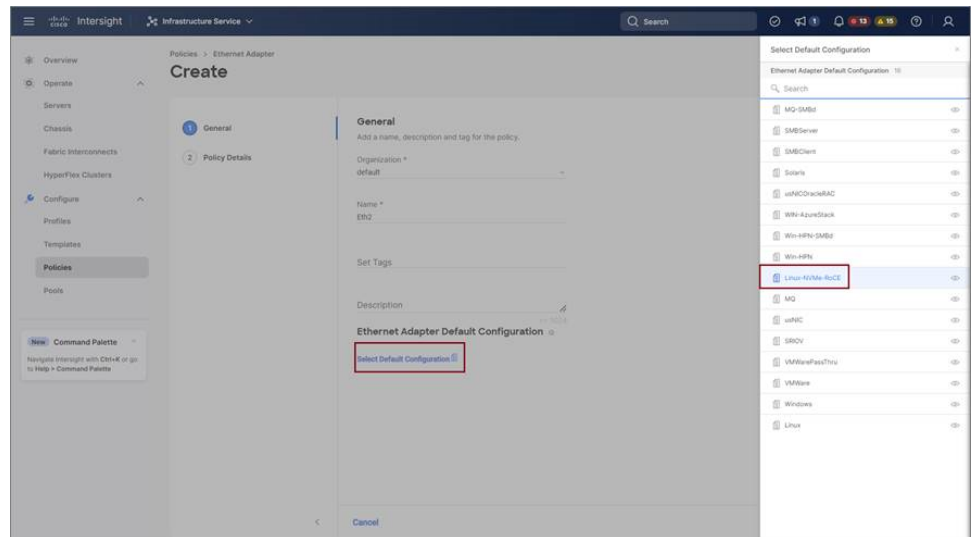
RoCE v2 vNIC を構成するには、次の手順に従います。Cisco Intersight LAN 接続ポリシーでは、次のように Linux 構成向けのイーサネットアダプタポリシーの RoCE 設定を有効にできます。

#### 手順

- ステップ 1 [構成 (CONFIGURE)] > [ポリシー (Policies)] に移動します。[ポリシーの作成 (Create Policy)] をクリックし、[UCS サーバ (UCS Server)] プラットフォームタイプを選択し、[LAN 接続ポリシー (LAN Connectivity policy)] を検索または選択して、[Start (開始)] をクリックします。
- ステップ 2 ポリシーの [全般 (General)] ページで、ポリシー名を入力し、[ターゲットプラットフォーム (Target Platform)] として [UCS サーバ (スタンドアロン) (UCS Server (Standalone))] または [UCS サーバ (FI アタッチ) (UCS Server (FI-Attached))] を選択し、[次へ (Next)] をクリックします。
- ステップ 3 [ポリシーの詳細 (Policy Details)] ページで、[vNIC の追加 (Add vNIC)] をクリックして新しい vNIC を作成します。

ステップ 4 [vNIC の追加 (Add vNIC)] ページで、構成パラメータに従って RoCE v2 vNIC を有効にします。

- a) [全般 (General)] セクションで、仮想イーサネット インターフェイスの名前を入力します。
- b) スタンドアロン サーバの [Consistent Device Naming (CDN)] セクションまたは FI アタッチ サーバの [フェールオーバー (Failover)] セクションで、次の手順を実行します。
  - [イーサネットアダプタ (Ethernet Adapter)] の下で、[ポリシーの選択 (Select Policy)] をクリックします。
  - [ポリシーの選択 (Select Policy)] ウィンドウで、[新規作成 (Create New)] をクリックして、イーサネットアダプタ ポリシーを作成します。
  - [全般 (General)] ページで、ポリシーの名前を入力し、[デフォルトの構成を選択 (Select Default Configuration)] をクリックします。[デフォルトの構成 (Default Configuration)] ウィンドウで [Linux-NVMe-RoCE] を検索して選択し、[次へ (Next)] をクリックします。
  - [ポリシーの詳細 (Policy Details)] で、RoCE のデフォルト構成パラメータを確認し、[作成 (Create)] をクリックします。



- [追加 (Add)] をクリックして設定を保存し、新しい vNIC を追加します。

(注)

\* が付いているすべてのフィールドは必須です。適切なポリシーに従って入力または選択されていることを確認してください。

ステップ 5 [作成 (Create)] をクリックし、RoCE v2 設定によって LAN 接続ポリシーを完成させます。

ステップ 6 LAN 接続ポリシーをサーバプロファイルに関連付けます。

(注)

詳細については、「[UCS サーバポリシーの構成](#)」の「[LAN接続ポリシーの作成](#)」および「[イーサネットアダプタポリシーの作成](#)」および「[UCS サーバプロファイルの構成](#)」を参照してください。

---

イーサネットアダプタポリシーの vNIC 設定を含む LAN 接続ポリシーが正常に作成および展開され、RoCE v2 設定が有効になります。

### 次のタスク

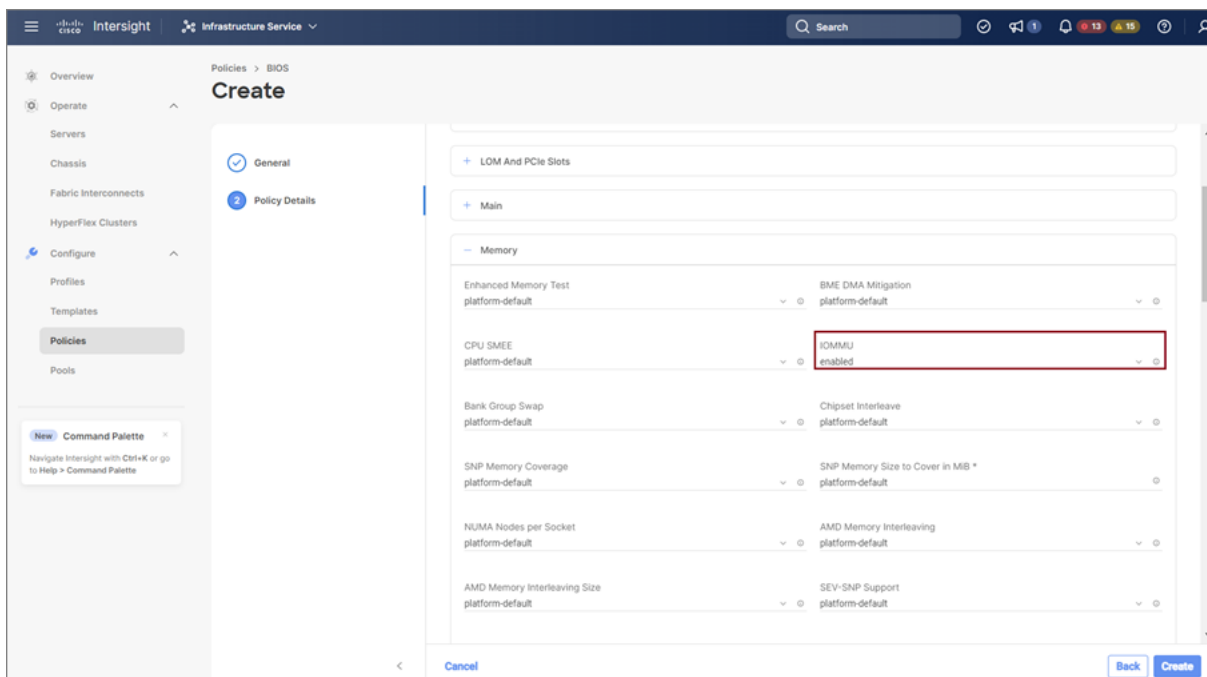
RoCE v2 のポリシー構成が完了したら、続いて、BIOS ポリシーで IOMMU を有効にします。

## IOMMU BIOS 設定の有効化

Linux カーネルで IOMMU を有効にする前に、次の手順を実行して、RoCE v2 vNIC を使用するようサーバのサービスプロファイルを構成し、IOMMU BIOS ポリシーを有効にします。

### 手順

- 
- ステップ 1 [構成 (CONFIGURE)] > [ポリシー (Policies)] に移動します。[ポリシーの作成 (Create Policy)] をクリックし、[UCS サーバ (UCS Server)] プラットフォームタイプを選択し、[BIOS] を検索または選択して、[Start (開始)] をクリックします。
  - ステップ 2 [全般 (General)] ページで、ポリシーの名前を入力し、[次へ (Next)] をクリックします。
  - ステップ 3 [ポリシーの詳細 (Policy Details)] ページで、次の BIOS を構成します。
    - a) [すべてのプラットフォーム (All Platforms)] を選択します。
    - b) [メモリ (Memory)] グループを展開します。
    - c) [IOMMU] ドロップダウンリストで、IOMMU 構成の設定を有効にする BIOS 値を選択します。



ステップ 4 [作成 (Create)] をクリックします。

ステップ 5 BIOS ポリシーをサーバプロファイルに関連付け、サーバを再起動します。

(注)

詳細については、「[サーバポリシーの構成](#)」の「[BIOS ポリシーの作成](#)」および「[サーバプロファイルの構成](#)」を参照してください。

BIOS ポリシーが正常に作成され、サーバプロファイルに展開されます。

#### 次のタスク

ホストシステムで RoCE v2 for NVMeoF を構成します。

## ホストシステムでの NVMeoF の RoCE v2 の構成

### 始める前に

IOMMU 対応 BIOS ポリシーを使用して、RoCE v2 vNIC を使用するサーバのサービスプロファイルを設定します。

## 手順

**ステップ 1** 編集のために /etc/default/grub ファイルを開きます。

**ステップ 2** GRUB\_CMDLINE\_LINUX の末尾に intel\_iommu=on を追加します。

```
sample /etc/default/grub configuration file after adding intel_iommu=on:
# cat /etc/default/grub
GRUB_TIMEOUT=5
GRUB_DISTRIBUTOR="$(sed 's, release .*$,,g' /etc/system-release)"
GRUB_DEFAULT=saved
GRUB_DISABLE_SUBMENU=true
GRUB_TERMINAL_OUTPUT="console"
GRUB_CMDLINE_LINUX="crashkernel=auto rd.lvm.lv=rhel/root rd.lvm.lv=rhel/swap biosdevname=1
rhgb quiet intel_iommu=on
GRUB_DISABLE_RECOVERY="true"
```

**ステップ 3** ファイルを保存した後、新しい grub.cfg ファイルを生成します。

レガシー ブートの場合 :

```
# grub2-mkconfig -o /boot/grub2/grub.cfg
```

UEFI ブートの場合 :

```
# grub2-mkconfig -o /boot/grub2/efi/EFI/redhat/grub.cfg
```

**ステップ 4** サーバをリブートします。IOMMU を有効にした後で、変更を反映するためにサーバを再起動します。

**ステップ 5** サーバが intel\_iommu=on オプションを使用して起動されていることを確認します。

```
cat /proc/cmdline | grep iommu
```

出力の最後に含まれることに注意してください。

```
[root@localhost basic-setup]# cat /proc/cmdline | grep iommu
BOOT_IMAGE=vmlinux-3.10.0-957.27.2.el7.x86_64 root=/dev/mapper/rhel-root ro crashkernel=auto
rd.lvm.lv=rhel/root rd.lvm.lv=rhel/swap rhgb quiet intel_iommu=on LANG=en US.UTF-8
```

## 次のタスク

enic および enic\_rdma ドライバをダウンロードします。

## Cisco enic および enic\_rdma ドライバのインストール

enic\_rdma ドライバには enic ドライバが必要です。enic および enic\_rdma ドライバをインストールする場合は、Cisco.com で一致する enic および enic\_rdma ドライバのセットをダウンロードして使用してください。inbox enic ドライバを使用して Cisco.com からダウンロードしたバイナリ enic\_rdma ドライバを使用しようとしても、機能しません。



## 手順

**ステップ 1** enic および enic\_rdma rpm パッケージをインストールします。

```
# rpm -ivh kmod-enic-<version>.x86_64.rpm kmod-enic_rdma-<version>.x86_64.rpm
```

(注)

enic\_rdma のインストール中に、enic\_rdmalibnvdimm モジュールは、RHEL 7.7 へのインストールに失敗することがあります。nvdimm-security.conf dracut モジュールは add\_drivers 値にスペースを必要とするためです。回避策については、次のリンクの指示に従ってください。

<https://access.redhat.com/solutions/4386041>

[https://bugzilla.redhat.com/show\\_bug.cgi?id=1740383](https://bugzilla.redhat.com/show_bug.cgi?id=1740383)

**ステップ 2** enic\_edma ドライバはインストールされていますが、動作中のカーネルでロードされません。サーバを再起動して、実行中のカーネルに enic\_rdma ドライバをロードします。

**ステップ 3** enic\_rdma ドライバと RoCE v2 インターフェイスのインストールを確認します。

```
[root@localhost ~]# dmesg | grep enic_rdma
[  3.137083] enic_rdma: Cisco VIC Ethernet NIC RDMA Driver, ver 1.2.0.28-877.2
2 init
[  3.242663] enic 0000:1b:00.1 eno6: enic_rdma: FW v3 RoCEv2 enabled
[  3.284856] enic 0000:1b:00.4 eno9: enic_rdma: FW v3 RoCEv2 enabled
[ 16.441662] enic 0000:1b:00.1 eno6: enic_rdma: Link UP on enic_rdma_0
[ 16.458754] enic 0000:1b:00.4 eno9: enic_rdma: Link UP on enic_rdma_1
```

**ステップ 4** vme-rdma カーネル モジュールをロードします。

```
# modprobe nvme-rdma
```

サーバの再起動後に、nvme-rdma カーネルモジュールがアンロードされます。サーバの再起動ごとに nvme-rdma カーネルモジュールをロードするには、次を使用して nvme\_rdma.conf ファイルを作成します。

```
# echo nvme_rdma > /etc/modules-load.d/nvme_rdma.conf
```

(注)

インストール後の enic\_rdma の詳細については、`rpm -q -l kmod-enic_rdma` コマンドを使用して README ファイルを抽出します。

### 次のタスク

ターゲットを検出し、NVMe 名前スペースに接続します。システムでストレージへのマルチパス アクセスが必要な場合は、[デバイス マッパー マルチパスの設定 \(11 ページ\)](#) についてのセクションを参照してください。

## NVMe ターゲットの検出

NVMe のターゲットを検出し、NVMe ネームスペースを接続するには、次の手順を使用します。

### 始める前に

まだインストールされていない場合は、**nvme cli** バージョン 1.6 以降をインストールします。



(注) **nvme-cli** バージョン 1.7 以降がインストールされている場合は、下のステップ 2 はスキップします。

RoCEv2 インターフェイスで IP アドレスを設定し、インターフェイスがターゲット IP に対して ping を実行できることを確認します。

### 手順

**ステップ 1** /etc で nvme フォルダを作成し、ホスト nqn を手動で生成します。

```
# mkdir /etc/nvme
# nvme gen-hostnqn > /etc/nvme/hostnqn
```

**ステップ 2** settos.sh ファイルを作成し、IB フレームでプライオリティフロー制御 (PFC) を設定するスクリプトを実行します。

(注)

NVMeoF トラフィックの送信に失敗しないようにするには、サーバを再起動するごとにこのスクリプトを作成して実行する必要があります。

```
# cat settos.sh
#!/bin/bash
for f in `ls /sys/class/infiniband`;
do
    echo "setting TOS for IB interface:" $f
    mkdir -p /sys/kernel/config/rdma_cm/$f/ports/1
    echo 186 > /sys/kernel/config/rdma_cm/$f/ports/1/default_roce_tos
done
```

**ステップ 3** 次のコマンドを入力して、NVMe ターゲットを検出します。

```
nvme discover --transport=rdma --traddr=<IP address of transport target>
```

例えば、50.2.85.200 でターゲットを検出するには、次のようにします。

```
# nvme discover --transport=rdma --traddr=50.2.85.200

Discovery Log Number of Records 1, Generation counter 2
=====Discovery Log Entry 0=====
trtype: rdma
adrfam: ipv4
subtype: nvme subsystem
treq: not required
```

```
portid: 3
trsvcid: 4420
subnqn: nqn.2010-06.com.purestorage:flasharray.9a703295ee2954e
traddr: 50.2.85.200
rdma_prtype: roce-v2
rdma_qptype: connected
rdma_cms: rdma-cm
rdma_pkey: 0x0000
```

(注)

IPv6 を使用して NVMe ターゲットを検出するには、traddr オプションの次に IPv6 ターゲットアドレスを指定します。

**ステップ 4** 次のコマンドを入力して、検出された NVMe ターゲットに接続します。

```
nvme connect --transport=rdma --traddr=<IP address of transport target port>> -n <subnqn value from nvme discover>
```

例えば、50.2.85.200 のターゲットと上記の subnqn 値を検出するには、次の手順を実行します。

```
# nvme connect --transport=rdma --traddr=50.2.85.200 -n
nqn.2010-06.com.purestorage:flasharray.9a703295ee2954e
```

(注)

IPv6 を使用して検出した NVMe ターゲットに接続するには、traddr オプションの次に IPv6 ターゲットアドレスを指定します。

**ステップ 5** nvme list コマンドを使用して、マッピングされたネームスペースを確認します。

```
# nvme list
Node              SN                      Model                      Namespace
Usage              Format                    FW Rev
-----
/dev/nvme0n1      09A703295EE2954E      Pure Storage FlashArray   72656
  4.29 GB /      4.29 GB   512   B + 0 B   99.9.9
/dev/nvme0n2      09A703295EE2954E      Pure Storage FlashArray   72657
  5.37 GB /      5.37 GB   512   B + 0 B   99.9.9
```

## デバイス マッパー マルチパスの設定

システムがデバイス マッパー マルチパス (DM マルチパス) を使用して構成されている場合は、次の手順に従ってデバイス マッパー マルチパスをセットアップします。

### 手順

**ステップ 1** まだインストールされていない場合は、device-mapper-multipath パッケージをインストールします。

**ステップ 2** Multipathd を有効にして開始します。

```
# mpathconf --enable --with_multipathd y
```

ステップ 3 etc/multipath.conf ファイルを編集して、次の値を使用します。

```
defaults {
    polling_interval      10
    path_selector         "queue-length 0"
    path_grouping_policy  multibus
    fast_io_fail_tmo     10
    no_path_retry         0
    features               0
    dev_loss_tmo          60
    user_friendly_names   yes
}
```

ステップ 4 更新されたマルチパス デバイス マップを使用してフラッシュします。

```
# multipath -F
```

ステップ 5 マルチパス サービスを再起動します。

```
# systemctl restart multipathd.service
```

ステップ 6 マルチパス デバイスを再スキャンします。

```
# multipath -v2
```

ステップ 7 マルチパス ステータスを確認します。

```
# multipath -ll
```

## Cisco Intersight を使用した RoCE v2 インターフェイスの削除

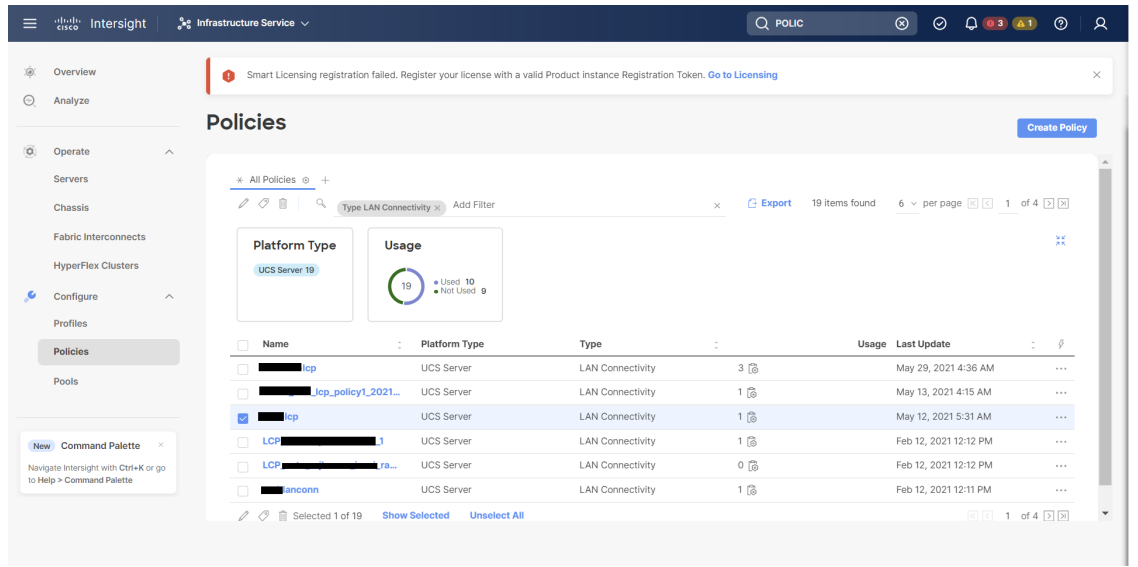
RoCE v2 インターフェイスを削除するには、次の手順を実行します。

### 手順

ステップ 1 [構成 (CONFIGURE) ]>[ポリシー (Policies) ]に移動します。[フィルタの追加 (Add Filter) ] フィールドで、[タイプ: LAN 接続 (Type: LAN Connectivity) ]を選択します。

ステップ 2 RoCE V2 構成用に作成された適切な LAN 接続ポリシーを選択し、ポリシー リストの上部または下部にある削除アイコンを使用します。

ステップ 3 ポリシーを削除するには、[削除 (Delete) ]をクリックします。



The screenshot displays the Cisco Intersight interface for managing policies. A notification at the top indicates a failed Smart Licensing registration. The main section is titled 'Policies' and shows a list of 19 items found, filtered by 'Type LAN Connectivity'. A 'Usage' chart shows 19 items used and 0 not used. The table below lists the policies:

Name	Platform Type	Type	Usage	Last Update
[redacted]_lcp	UCS Server	LAN Connectivity	3	May 29, 2021 4:36 AM
[redacted]_lcp_policy1_2021...	UCS Server	LAN Connectivity	1	May 13, 2021 4:15 AM
[redacted]_lcp	UCS Server	LAN Connectivity	1	May 12, 2021 5:31 AM
[redacted]_lcp	UCS Server	LAN Connectivity	1	Feb 12, 2021 12:12 PM
[redacted]_lcp	UCS Server	LAN Connectivity	0	Feb 12, 2021 12:12 PM
[redacted]_lanconn	UCS Server	LAN Connectivity	1	Feb 12, 2021 12:11 PM

**ステップ 4** RoCE v2 構成を削除したら、サーバプロファイルを再展開し、サーバを再起動します。



## 翻訳について

このドキュメントは、米国シスコ発行ドキュメントの参考和訳です。リンク情報につきましては、日本語版掲載時点で、英語版にアップデートがあり、リンク先のページが移動/変更されている場合がありますことをご了承ください。あくまでも参考和訳となりますので、正式な内容については米国サイトのドキュメントを参照ください。