AppDirectモード用のVMware ESXiでのDCPMMの設定

内容

概要

このドキュメントでは、ホストマネージドモードでIntel® Optine™ Persistent Memory(PMEM)を使用して、Unified Computing System(UCS)Bシリーズサーバ でESXiを設定するプロセスについて説明します。

前提条件

要件

次の項目に関する知識があることが推奨されます。

- UCS B シリーズ
- インテル® Optane™データ・センター・パーシステント・メモリー・モジュール(DCPMM)の概念
- VMware ESXiおよびvCenter Serverの管理

この設定を行う前に、次の要件が満たされていることを確認します。

- B200/B480 M5仕様ガイドのPMEMガイドラインを参照してください。
- CPUが第2世代のIntel® Xeon®スケーラブルプロセッサであることを確認してください。
- PMEM/Dynamic Random Access Memory(DRAM)比は、KB 67645に基づく要件を満たします。

- ESXiは6.7 U2 + Express Patch 10(ESXi670-201906002)以降です。以前の6.7リリースはサポートされていません。
- UCS Managerとサーバのバージョンは4.0(4)以降です。最新の推奨バージョンについては、

www.software.cisco.com/を参照してください。

使用するコンポーネント

このドキュメントの情報は、次のソフトウェアとハードウェアのバージョンに基づいています。

- UCS B480 M5
- UCS Manager 4.1(2b)

このドキュメントの情報は、特定のラボ環境にあるデバイスに基づいて作成されました。このドキュメントで使用するすべてのデバイスは、初期（デフォルト）設定の状態から起動しています。本稼働中のネットワークでは、各コマンドによって起こる可能性がある影響を十分確認してください。

背景説明

App Directモードに設定されたUCSサーバでは、VMware ESXi仮想マシンはOptine DCPMM永続的メモリ不揮発性デュアルインラインメモリモジュール(NVDIMM)にアクセスします。

Intel Optane DCPMMは、IPMCTL管理ユーティリティからUnified Extensible Firmware Interface (UEFI)シェルまたはOSユーティリティを使用して設定できます。このツールは、次のアクションのいくつかを実行するように設計されています。

- モジュールの検出と管理
- モジュールファームウェアのアップデートと設定
- 状態の監視
- 目標、地域、および名前空間のプロビジョニングと設定
- PMEMのデバッグとトラブルシューティング

UCSは、サービスプロファイルに接続された永続的なメモリポリシーを使用して設定できるため、使いやすくなります。

オープンソースの不揮発性デバイス制御(NDCTL)ユーティリティは、LIBNVDIMM Linuxカーネルサブシステムを管理するために使用されます。NDCTLユーティリティを使用すると、システムは設定をプロビジョニングし、OS用の領域および名前空間として実行できます。

ESXiホストに追加された永続メモリは、ホストによって検出され、フォーマットされ、ローカルPMemデータストアとしてマウントされます。ESXiはPMEMを使用するためにVirtual Machine Flying System(VMFS)-Lファイルシステム形式を使用し、ホストごとに1つのローカルPMEMデータストアだけがサポートされます。

PMEMデータストアは、他のデータストアとは異なり、従来のデータストアとしてのタスクをサポートしません。vmxおよびvmware.logファイルを含むVMホームディレクトリをPMEMデータストアに配置することはできません。

PMEMは、次の2種類のモードでVMに提示できます。ダイレクトアクセスモードと仮想ディスクモード。

- ダイレクトアクセスモード
  VMは、NVDIMMの形式でPMEM領域を表示することで、このモードに設定できます。このモ

ードを使用するには、VMオペレーティングシステムがPMem対応である必要があります。NVDIMMはバイトアドレス可能メモリとして動作するため、NVDIMMモジュールに保存されたデータは電源サイクルにわたって保持されます。NVDIMMは、PMEMのフォーマット時にESXiによって作成されたPMemデータストアに自動的に保存されます。

- 仮想ディスクモード
  ハードウェアバージョンをサポートするために、VM上に常駐する従来のOSおよびレガシーOSを対象としています。VM OSはPMEM対応である必要はありません。このモードでは、従来のSmall Computer System Interface(SCSI)仮想ディスクを作成し、VM OSで使用できます。

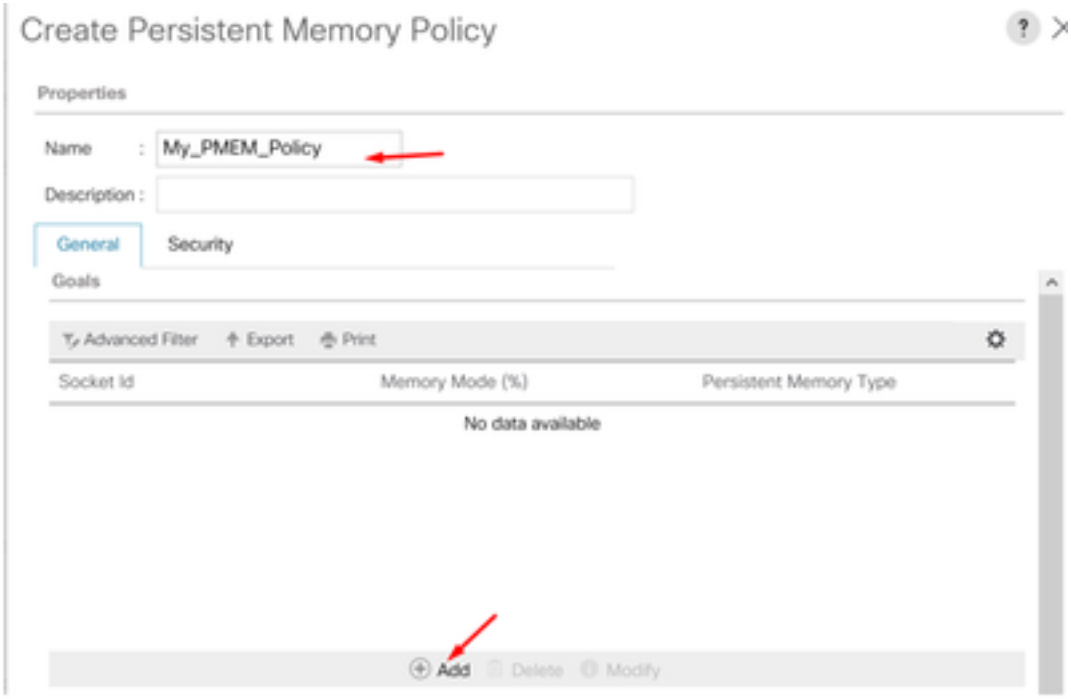
このドキュメントでは、ダイレクトアクセスモードで仮想マシン(VM)を使用するための設定について説明します。

設定

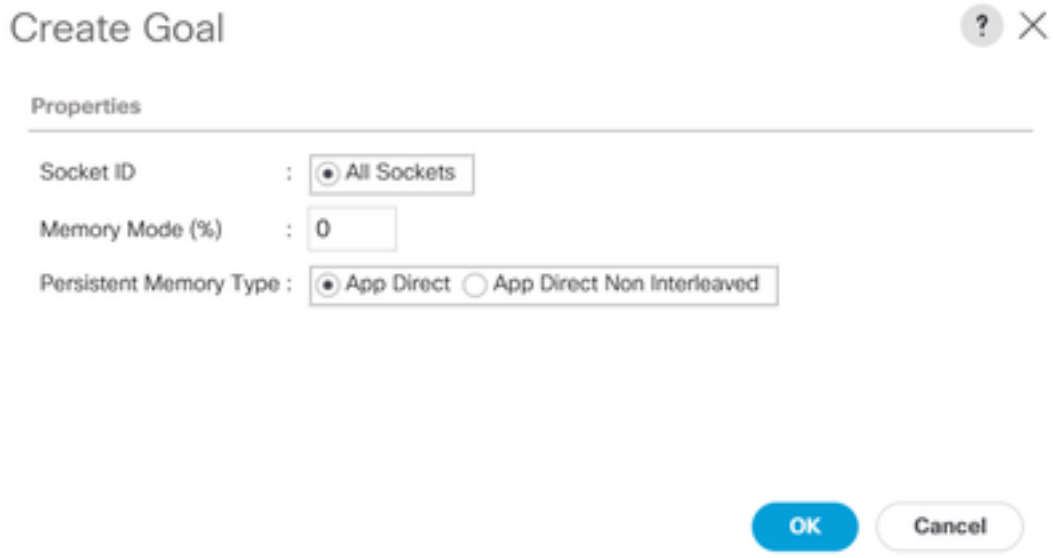
この手順では、Intel Optine DCPMMを使用してUCSブレードサーバ上でESXiを設定する方法について説明します。

サービス プロファイルの設定

1. UCS Manager GUIで、[Servers] > **[Persistent Memory Policy]に移動**し、図に示すように**[Add]**をクリックします。



2.目標を作**成**し、図に示すように**Memory Mode**が0%であることを確認します。

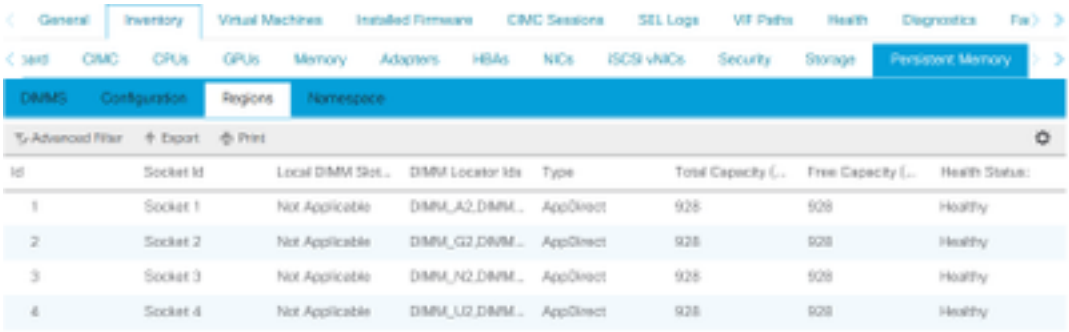


3. PMEMポリシーを目的のサービスプロファイルに追加します。

[Service Profile] > [Policies] > [Persistent Memory Policy]に移動し、作成したポリシーを適用します。

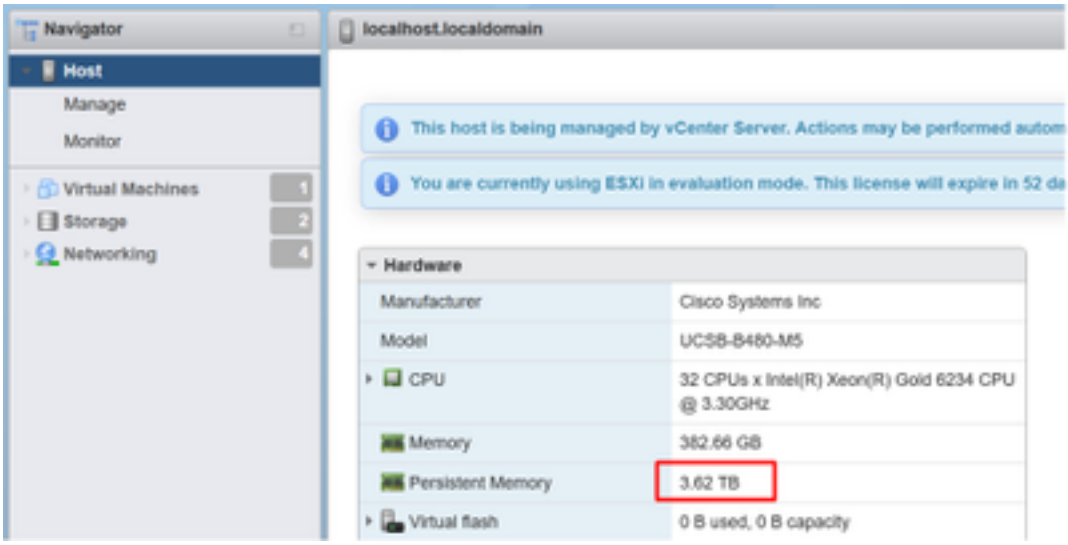
4. 地域の健全性を確認します。

選択した[サーバ] > [インベントリ] > **[永続メモリ] > [リージョン]に移動します**。タイプAppDirectが表示されます。このメソッドは、CPUソケットごとに1つの領域を作成します。

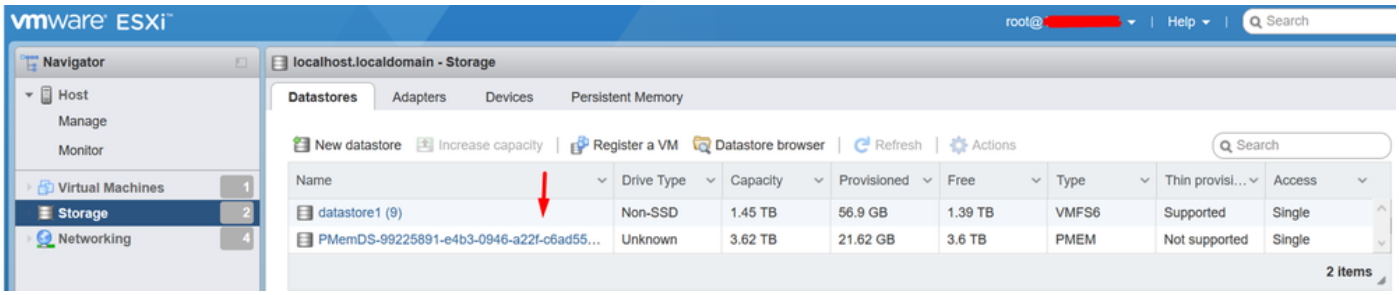


ESXiの確認

1. Webコンソールで、ホストに使用可能な合計PMEMが表示されます。



2. ESXiは、図に示すように、PMEMの総量で構成される特別なデータストアを表示します。



仮想マシンNVDIMMの設定

1. ESXiでは、仮想マシンはNVDIMMとしてOptine DCPMM PMEMにアクセスします。NVMDIMMを仮想マシンに割り当てるには、vCenterから仮想マシンにアクセスし、**[Actions] > [Edit Settings]に移動**して、**[ADD NEW DEVICE]をクリック**して、図に示すように**[NVDIMM]**を選択します。

> **注：** 仮想マシンを作成する場合は、OSの互換性がインテル® Optine™永続メモリーをサポートする最小要件バージョンを満たしていることを確認してください。それ以外の場合は、**NVDIMMオプションが選択可能な項目に表示されません。**

2.図に示すようにNVDIMMのサイズを設定します。

仮想マシンでの名前空間の構成

1. NDCTLユーティリティは、PMEMまたはNVDIMMの管理と設定に使用されます。

この例では、設定にRed Hat 8が使用されています。Microsoftには、永続的なメモリ名前空間の管理用のPowerShellコマンドレットがあります。

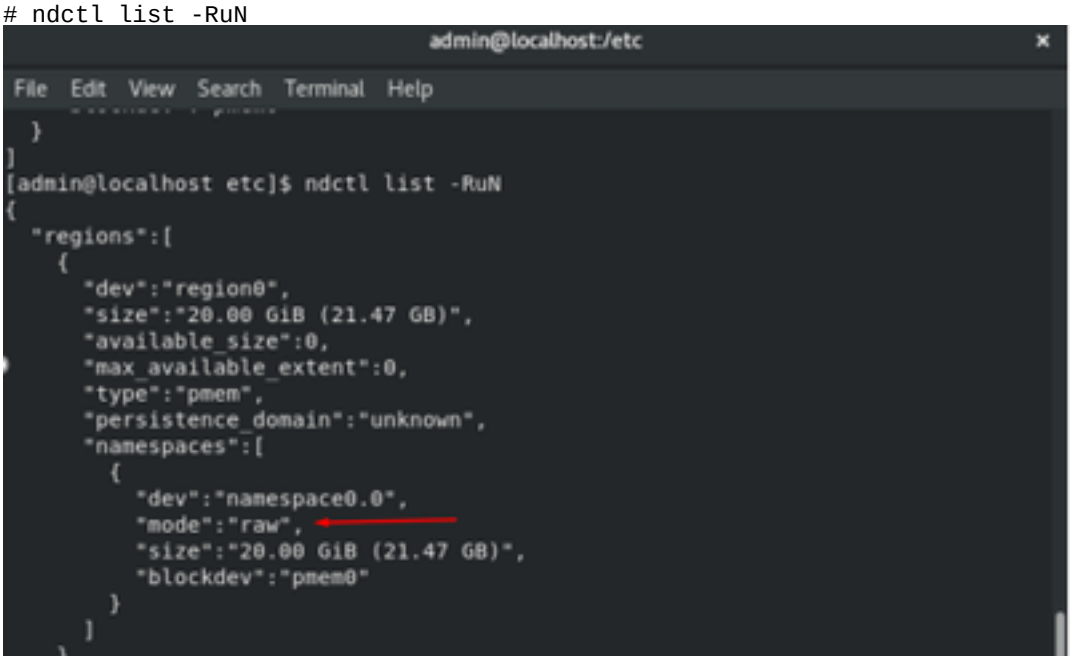
Linuxディストリビューションに従って、利用可能なツールを使用してNDCTLユーティリティをダウンロードします

以下に、いくつかの例を示します。

```
# yum install ndctl # zypper install ndctl # apt-get install ndctl
```

2. ESXiがデフォルトで作成したNVDIMM領域と名前空間を確認します。NVDIMMが仮想マシンに割り当てられている場合、スペースが設定と一致していることを確認します。名前空間のモードがrawに設定されていることを確認します。これは、ESXiが名前空間を作成したことを意味します。確認するには、次のコマンドを使用します。

```
# ndctl list -RuN
```



を選択します。（オプション）ネームスペースがまだ作成されていない場合は、次のコマンドを

使用してネームスペースを作成できます。

```
# ndctl create-namespace
```

ndctl create-namespaceコマンドは、デフォルトでfsdaxモードで新しいネームスペースを作成し、新しい/dev/pmem([x].[y])デバイスを作成します。名前空間がすでに作成されている場合は、この手順をスキップできます。

4. PMEMアクセスモードを選択します。設定に使用できるモードは次のとおりです。

- セクタモード：

  ストレージを高速ブロックデバイスとして表示します。これは、永続的なメモリを使用できないレガシーアプリケーションに役立ちます。

- Fsdaxモード：

  永続メモリデバイスがNVDIMMへの直接アクセスをサポートできるようにします。ファイルシステムの直接アクセスでは、直接アクセスプログラミングモデルを使用できるようにするために、 fsdaxモードを使用する必要があります。このモードでは、NVDIMM上にファイルシステムを作成できます。

- Devdaxモード：

  DAX文字デバイスを使用して永続メモリへのrawアクセスを提供します。 devdaxモードを使用するデバイス上でファイルシステムを作成することはできません。

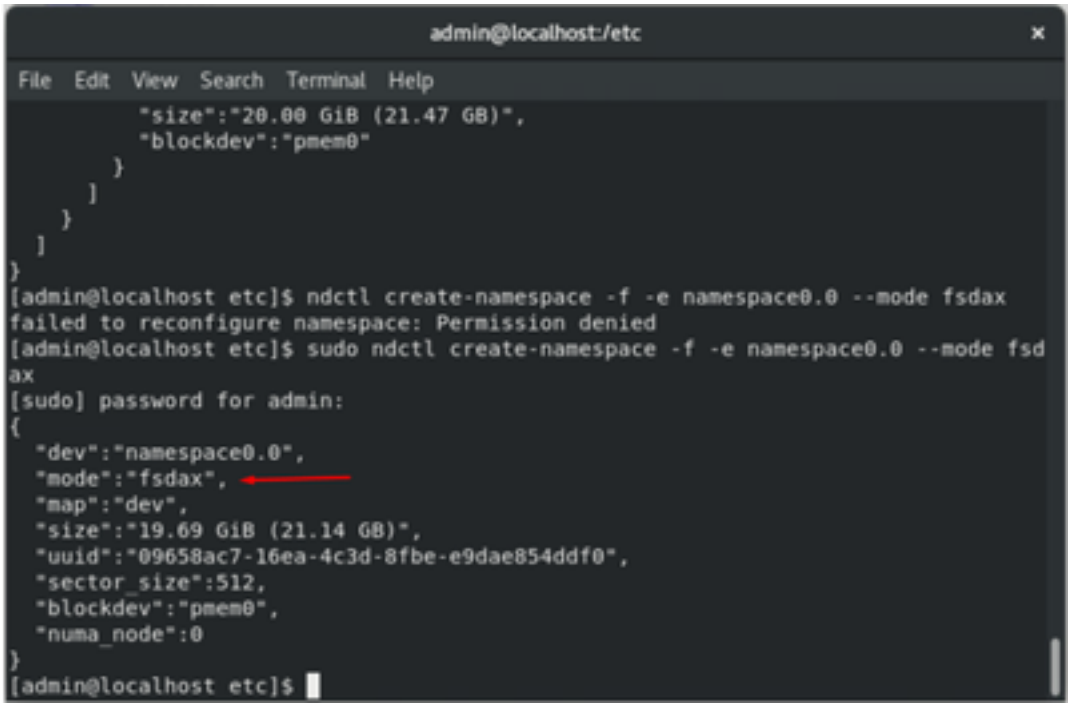
- Rawモード：

  このモードには複数の制限があるため、永続メモリの使用は推奨されません。

モードをfsdaxモードに変更するには、コマンドを使用します。

```
ndctl create-namespace -f -e
```

devがすでに作成されている場合は、dev名前空間を使用してモードをフォーマットし、 fsdaxに変更します。



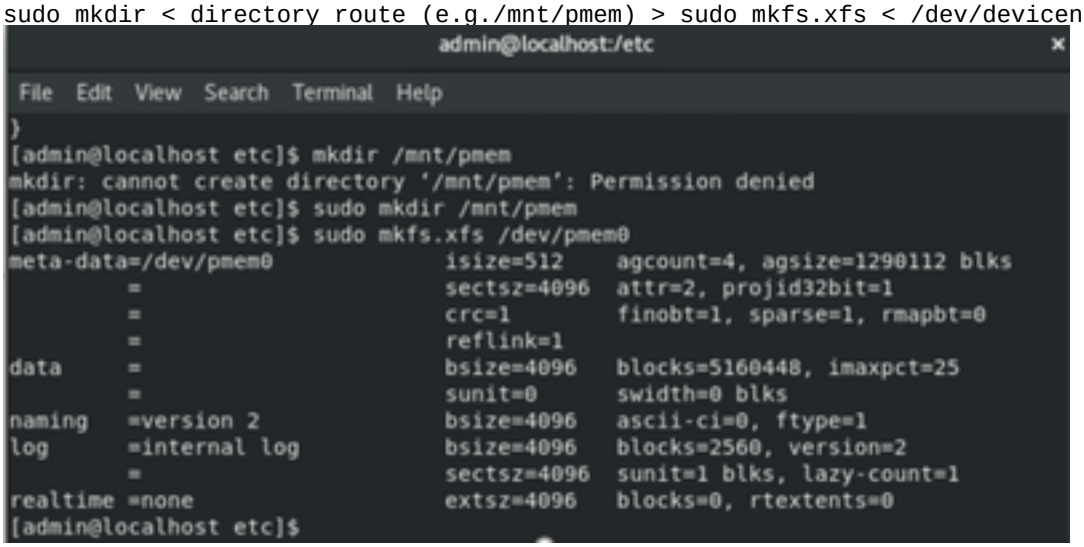
注：これらのコマンドでは、アカウントにルート権限が必要です。**sudoコマンド**が必要になる場合があります。

5.ディレクトリとファイルシステムを作成します。

ダイレクトアクセスまたはDAXは、アプリケーションがCPUから（ロードとストアを介して）永続的なメディアに直接アクセスし、従来のI/Oスタックをバイパスできるようにするメカニズムです。DAX対応の永続的メモリファイルシステムには、ext4、XFS、およびWindows NTFSがあります。

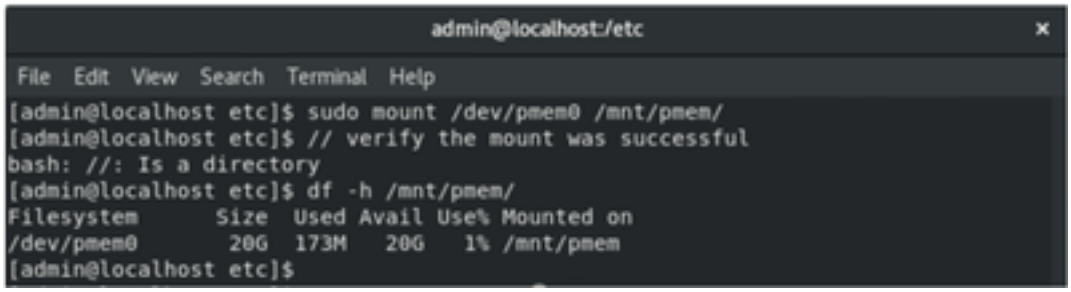
作成およびマウントされたXFSファイルシステムの例：

```
sudo mkdir < directory route (e.g./mnt/pmem) > sudo mkfs.xfs < /dev/devicename (e.g. pmem0) >
```



6.ファイルシステムをマウントし、正常に実行されたことを確認します。

```
sudo mount
```



VMはPMEMを使用する準備ができました。

トラブルシュート

エラーが見つかった場合は、-o daxマウントオプションを使用してこのDAX対応ファイルシステムをマウントすることをお勧めします。

```
[admin@localhost etc]$ sudo mount -o dax /dev/pmem0 /mnt/pmem/
mount: /mnt/pmem: wrong fs type, bad option, bad superblock on /dev/pmem0, missi
ng codepage or helper program, or other error.
```

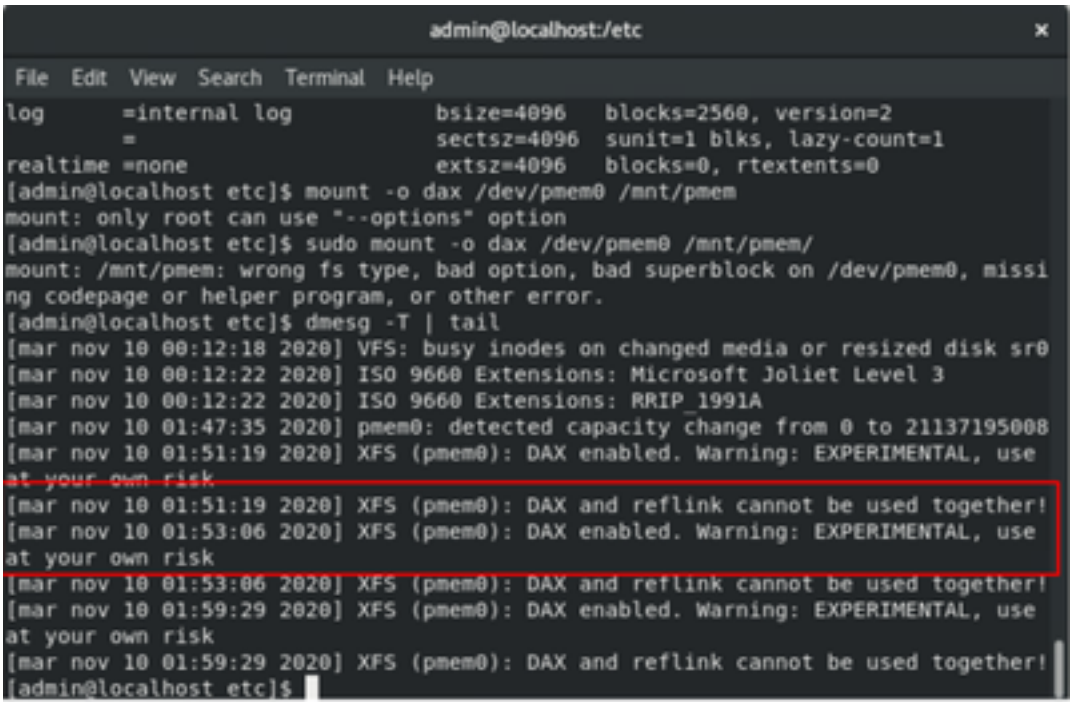
整合性を確保するために、ファイルシステムの修復が実行されます。

```
[admin@localhost etc]$ sudo xfs_repair /dev/pmem0
[sudo] password for admin:
Phase 1 - find and verify superblock...
Phase 2 - using internal log
        - zero log...
        - scan filesystem freespace and inode maps...
        - found root inode chunk
Phase 3 - for each AG...
        - scan and clear agi unlinked lists...
        - process known inodes and perform inode discovery...
        - agno = 0
        - agno = 1
        - agno = 2
        - agno = 3
        - process newly discovered inodes...
Phase 4 - check for duplicate blocks...
        - setting up duplicate extent list...
        - check for inodes claiming duplicate blocks...
        - agno = 0
        - agno = 1
        - agno = 2
        - agno = 3
Phase 5 - rebuild AG headers and trees...
        - reset superblock...
Phase 6 - check inode connectivity...
        - resetting contents of realtime bitmap and summary inodes
        - traversing filesystem ...
        - traversal finished ...
        - moving disconnected inodes to lost+found ...
Phase 7 - verify and correct link counts...
done
[admin@localhost etc]$
```

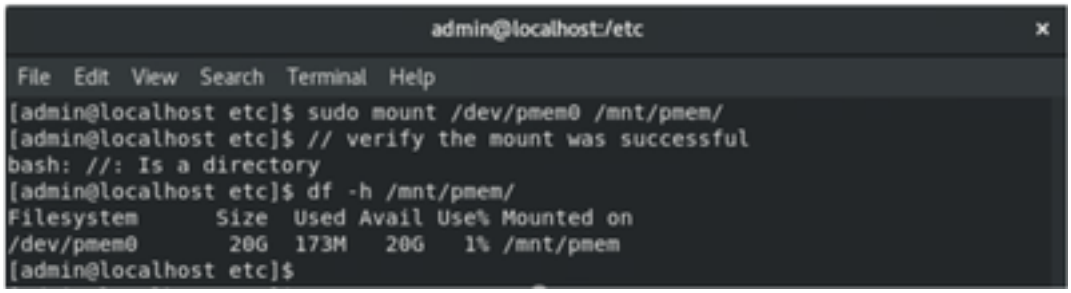
回避策として、-o daxオプションを使用せずにマウントを**マウントできます**。

> **注：**xfsprogs**バージョン**5.1では、デフォルトではreflinkオプションを有効にしてXFSファイルシステムを作**成すること**になります。以前は、デフォルトで無効になっていました。**reflink**と**dax**のオプション**は相互に排他的**で、マウントが失敗します。

「DAXとreflinkを一緒に使用することはできません！」 dmesgでは、mountコマンドが失敗するとエラーが表示されます。

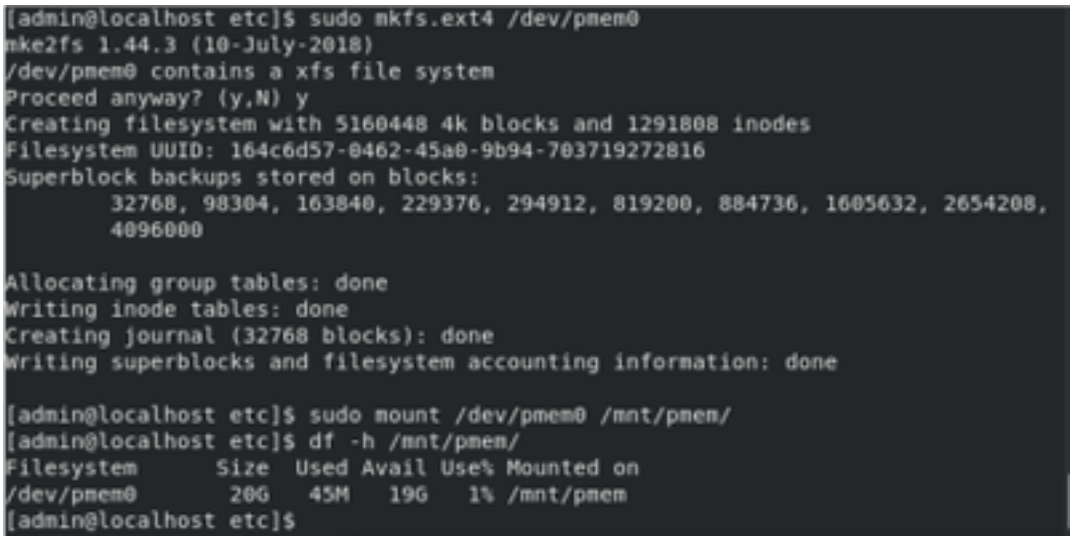


回避策として、-o daxオプションを削除します。



ext4 FSでマウントします。

EXT4ファイルシステムはreflink機能を実装していないがDAXをサポートしているため、代替として使用できます。



関連情報

- クイック スタート ガイドインテル®オプタン™ DC永続メモリーのプロビジョニング
- 永続メモリの設定
- インテル® Optine™永続メモリーの管理ユーティリティーipmctlおよびndctl
- テクニカル サポートとドキュメント – Cisco Systems