Red Hat Virtualization 4.0

SR-IOV 実装に関するハードウェアの考慮事項

SR-IOV を Red Hat Virtualization に実装するにあたってのハードウェアの考慮事項

Last Updated: 2018-08-02

Red Hat Virtualization 4.0 SR-IOV 実装に関するハードウェアの考慮事項

SR-IOV を Red Hat Virtualization に実装するにあたってのハードウェアの考慮事項

Red Hat Virtualization Documentation Team
Red Hat Customer Content Services
rhev-docs@redhat.com

法律上の通知

概要

本ガイドは、Red Hat Enterprise Linux での SR-IOV 実装と、Red Hat Virtualization でのデバイス割り当てに関するハードウェアの考慮事項について説明します。

目次

1. はじめに

Single Root I/O Virtualization (SR-IOV) は、単一の PCI Express (PCIe) エンドポイントを複数の別個のデバイスとして使用できるようにするハードウェアリファレンスです。これは、Physical Function (PF) と Virtual Function (VF) の 2 つの PCIe 機能を導入することによって実現します。

Physical Function とは、SR-IOV 機能を含む従来の PCIe 機能で、PCIe デバイスを完全に設定/制御することができます。各 PCIe デバイスには、独立した PF を 1 つから 8 つまで使用することができます。

Virtual Function とは、データの移動に必要なリソースと、最小限の設定リソースセットを含む軽量の PCIe 機能のことです。複数の VF を各 PF 上に作成して、各 PF が異なる数の VF をサポートすることが可能です。許容される VF の合計数は、PCIe デバイスベンダーによって異なり、またデバイスによっても違います。

PCIe の仕様は、PCIe ヘッダーのデバイス数フィールドを再換算して 9 個以上の機能を使用可能にする Alternative Routing ID Interpretation (ARI) の実装によって、より多くの VF をサポートします。この換算は、PCIe デバイスおよびそのデバイスのすぐ上位のポート (ARI をサポートするルートポートまたはスイッチ) の両方に依存します。

システムファームウェア (BIOS または UEFI) は、メモリー、I/O ポートアパーチャー、および PCIe バス番号の範囲などのリソースを割り当てます。このため、十分なリソースを割り当てるためにはファームウェアが SR-IOV をサポートおよび有効化している必要があります。

1.1. SR-IOV に関するハードウェア考慮事項の概要

- ファームウェア (BIOS または UEFI) が SR-IOV をサポートしていること。拡張機能がデフォルトで有効化されているかどうかを確認してください。有効化されていない場合には、手動で有効にします。これは、仮想化の拡張機能 (VT-d または AMD-Vi) を有効にするのと同様です。具体的な詳細は、ベンダーのマニュアルを参照してください。
- ルートポートまたは PCIe デバイスのすぐ上位のポート (例: PCIe スイッチ) が ARI をサポートしていること。
- PCIe デバイスが SR-IOV をサポートしていること。

ハードウェアがこれらの要件を満たしているかどうかを確認するには、ベンダーの仕様とデータシートを参照してください。

`lspci -v` コマンドを使用して、システムにインストール済みの PCI デバイスの情報を表示することができます。

2. デバイス割り当てを使用するための追加のハードウェア考慮事項

デバイス割り当て機能により、PCIe デバイスに仮想ゲストを直接割り当てて、ゲストへの完全なアクセスとネイティブに近いパフォーマンスを提供することができます。この機能を SR-IOV と併せて実装すると、VF が仮想ゲストに直接アタッチされます。この方法で、複数の仮想ゲストを単一の PCIe デバイスの VF に直接割り当てることができます。

仮想マシンを PCIe デバイスに直接割り当てるために SR-IOV を有効にする必要はありません。また、VF 作成の用途はデバイスの割り当てだけではありません。2 つの機能は補完関係にあり、これらを併用するには、追加のハードウェア考慮事項があります。

デバイス割り当て機能には、CPU とファームウェアで I/O Memory Management Unit (IOMMU) がサポートされている必要があります。IOMMU は I/O Virtual Addresses (IOVA) と物理メモリーアドレス間の変換を行います。これにより、仮想ゲストがデバイスをゲストの物理アドレスでプログラムすることが可能となります。それらのアドレスは、IOMMU がホストの物理アドレスに変換することができます。

IOMMU グループとは、システム内の他すべてのデバイスから分離することができるデバイスのセットです。IOMMU グループは、IOMMU の粒度と、システム内の他の IOMMU グループからの分離の両方を備えたデバイスの最小のセットです。これにより IOMMU は、IOMMU グループ外にあり IOMMU の制御が及ばないデバイス間の直接メモリーアクセス (DMA) を制限しつつ、IOMMU グループとのトランザクションを区別することができます。

仮想ゲストと PCIe デバイスの Virtual Function の間のトランザクションの分離は、デバイスの割り当てに必須です。PCIe およびサーバーの仕様で定義されている Access Control Service (ACS) 機能は、IOMMU グループ内で分離を維持するためのハードウェア標準です。ネイティブの ACS がない場合や、この機能が実装されていることをハードウェアベンダーが明言していない場合には、IOMMU グループのいずれかのマルチファンクションデバイスによって、IOMMU で保護されていない状態で発生する機能間のピアツーピア DMA が公開され、IOMMU グループが拡張されて、適切に分離されていない機能が含まれてしまうリスクを伴うことになります。

ネイティブの ACS サポートは、サーバーのルートポートにも推奨されます。このサポートがない場合には、それらのポートにインストールされたデバイスは、一緒にグループ化されます。ルートポートには、プロセッサーベース (northbridge) とコントローラーハブベース (southbridge) の 2 種類があります。前述したように、デバイス割り当て機能を SR-IOV と併用する場合には、仮想ゲストが VF に割り当てられるので、それらのポートが ACS と ARI の両方をサポートする必要があります。

Intel の Xeon プロセッサー E5 ファミリー、Xeon プロセッサー E7 ファミリー、および High End Desktop Processors には、プロセッサーベースのルートポートにおける ACS のネイティブサポートが含まれています。

Intel デバイスには、通常、コントローラーハブベースのルートポートでの ACS のネイティブサポートは含まれませんが、Red Hat Enterprise Linux 7.2 カーネルにはある程度の柔軟性があり、X99、X79、および 5 シリーズから 9 シリーズまでのチップセットのルートポートで ACS と同等の分離を有効にすることができます。

PCIe デバイスをインストールする際には、ベンダーの仕様を参照してルートポートが ACS をサポートしていることを確認した上で、プロセッサーおよびコントローラーハブベースのルートポートを決定してください。

加えて、I/O トポロジー内の PCIe スイッチまたはブリッジも ACS をサポートしている必要があります。そうでない場合には、IOMMU グループが拡張される可能性があります。

2.1. デバイス割り当て機能に関するハードウェア考慮事項の概要

- CPU が IOMMU (例: VT-d または AMD-Vi)をサポートしていること。IBM POWER8 はデフォルトで IOMMU をサポートしています。
- ファームウェアが IOMMU をサポートしていること。
- 使用する CPU ルートポートは、ACS または ACS と同等の機能をサポートしていること。
- PCIe デバイスが ACS または ACS と同等の機能をサポートしていること。
- PCIe デバイスとルートポート間の PCIe スイッチとブリッジはすべて ACS をサポートしていることを推奨します。たとえば、スイッチが ACS をサポートしていない場合には、そのスイッチの背後にあるデバイスはすべて同じ IOMMU グループを共有し、同じ仮想マシンにしか割り当てることができません。

ハードウェアがこれらの要件を満たしているかどうかを確認するには、ベンダーの仕様とデータシートを参照してください。

lspci -v コマンドを使用して、システムにインストール済みの PCI デバイスの情報を表示することができます。