ExpEther(エクスプレスイーサ)による 単一ホスト仮想化対応 I/O のマルチホスト同時共有

鈴木 順 飛鷹 洋一 吉川 隆士 岩田 淳
NEC システムプラットフォーム研究所

1. はじめに

コンピューティングリソースを、サービスの要求に応じて高速かつ柔軟に再構成する技術は今後益々重要になると考えられる [1]。また、これらの再構成された論理的なリソースは、プラットフォームを構成する物理的なリソースに対し仮想化されていなければならない。これらの要求を I/O に関して満たすのが I/O 仮想化技術である。

我々が提案している ExpEther は[2]、イーサネットを用いて複数のホストと I/O を接続し、接続するイーサネット上に論理的な PCI Express (PCIe)スイッチを形成する。これにより、ホストと I/O が 1:N のツリー接続であった PCIe を、M:N のメッシュ接続に拡張する。

ExpEther では、VLAN / MAC を用いて I/O の接続先ホストを切り替えることにより、ホストに対する I/O の割り当てを柔軟に切り替えることができる。

一方、最近 PCI-SIG では、I/O が単一ホスト内で複数の仮想マシン(VM)からのアクセスを認識する SR-IOV (Single-Root I/O Virtualization) を標準化した[3]。本稿では、ExpEther に SR-IOV 対応 I/O を接続することにより、複数ホストから I/O を同時に共有する方式を提案する。

2. ホストに対する I/O の割り当て変更

本節ではまず、ExpEther を用いてホストに対する I/O の割り当てを柔軟に切り替えられることを実験で示す。図 1 に ExpEther の構成を示す(用いる I/O は(a)従来の I/O)。

ExpEther ブリッジは、ホストと I/O をイーサネットに接続し、ホストと I/O 間で PCIe のパケット(TLP, Transaction Layer Packet)をトンネリングする。ホストには、ExpEther ブリッジ対

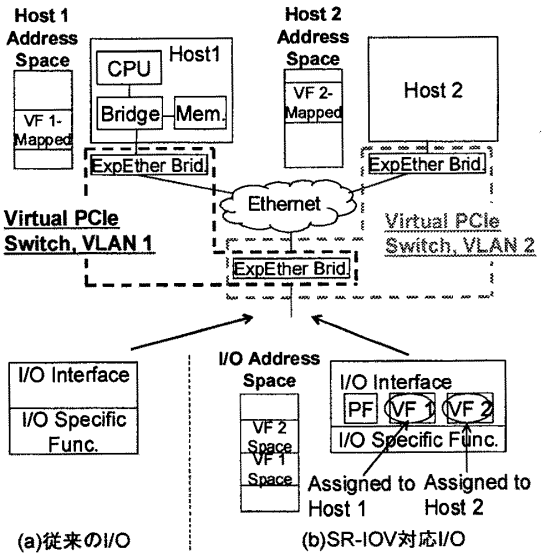

図 1. ExpEther の構成

が PCIe スイッチと等価となる。つまり、イーサネットの End-End で一段の論理的な PCIe スイッチを構成する。論理的な PCIe スイッチには、1 つの VLAN が割り当てられる。

ホストに対する I/O の割り当て変更は、I/O 側の ExpEther ブリッジの VLAN 変更と、ホストで動作する OS の I/O ホットプラグおよびホットリムーブの機能を用いて実現される。

本実験では、図 1(a)の従来の I/O として、一般的な GbE の Network Interface Card (NIC)を用いた。ホスト 1 および 2 では、ネットワーク帯域を測定する Iperf サーバを動作させた [4]。NIC の先は、GbE を用いて Iperf クライアントを動作させるクライアントサーバと接続した。クライアントサーバでは、5s 毎にホスト 1 と 2 に対するネットワーク帯域を測定した。

図 2 に NIC の割り当てをホスト 1 からホスト 2 に切り替えた際に、クライアント側で測定した各ホストの帯域変化を示す。ホットリムーブおよびホットプラグ処理に約 17s を要し、NIC の割り当てがホスト間で切り替えられることがわかる。

Simultaneous Multi-Host Sharing of I/O with ExpEther Interconnect
Jun Suzuki, Yoichi Hidaka, Takashi Yoshikawa, and Atsushi Iwata
System Platforms Research Laboratories
NEC Corporation

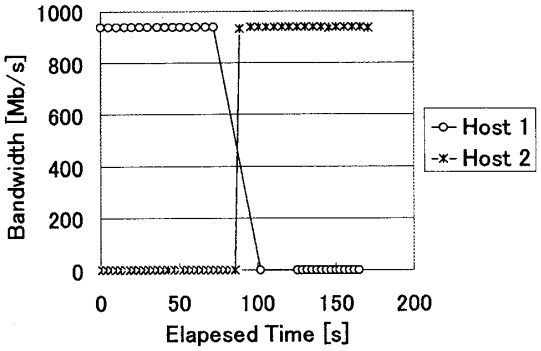

図 2. NIC 割り当て変更時の各ホストの帯域変化

3. SR-IOV 対応 I/O

本稿の提案方式では、SR-IOV に対応する I/O を ExpEther に接続し、I/O を複数ホスト間で同時に共有する。SR-IOV 対応 I/O の構成を図 1(b)に示す。SR-IOV 対応 I/O は、通常単一ホストに占有して用いられ、ホスト内の個々の VM に対し、VM が直接アクセスできるインタフェースを提供する。これにより、VM はソフトウェアの中間層を介さず I/O に直接アクセスすることができ、I/O アクセスのレイテンシおよびスループットが向上する [5]。

SR-IOV 対応 I/O 内の PF (Physical Function)は I/O リソースを制御するためのインタフェースであり、VF (Virtual Function)の数などを設定する。VF は VM に個別に割り当てられるインタフェースであり、VM は VF に対し、従来の I/O と同様に I/O 命令を発行する。

4. 提案方式

提案方式では、図 1(b)に示す SR-IOV 対応 I/O を接続し、I/O の VF を各ホストに個別に割り当てることで I/O の同時共有を実現する。

SR-IOV 対応 I/O は、通常単一ホストの物理メモリ空間に連続してマップされる。I/O と接続する ExpEther ブリッジは、図 1 に示すように I/O を予めコンフィグレーションし、コンフィグレーションしたメモリ領域から個々の VF 成分を識別し、各 VF をそれぞれのホストにリマップする。

I/O と接続する ExpEther ブリッジは、TLP を転送する際にパケットに記載されているアドレスをスワップする。図 3 にホストから I/O にパケットを転送する際のアドレススワップを示す。ホストが送信するパケットのあて先アドレスは、ExpEther ブリッジが VF をマップしたアドレス(Mem_H)である。ExpEther ブリッジは、あて先アドレスを予め I/O をコンフィグレーションした

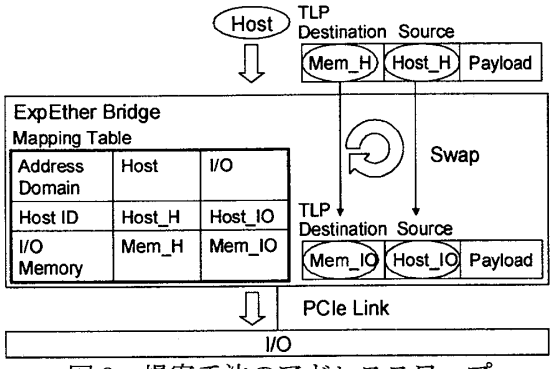

図 3. 提案手法のアドレススワップ

アドレス(Mem_IO)にスワップする。また、送信元 ID をホストが定めた値(Host_H)から、ホストを一意に識別するために ExpEther ブリッジが定める値(Host_IO)にスワップする。

5. まとめ

本稿では、SR-IOV 対応 I/O を、ExpEther を用いて複数ホスト間で同時に共有する方式を提案した。ExpEther は、イーサネットを用いて複数のホストと I/O を接続し、接続するイーサネット上に論理的な PCIe スイッチを形成する。これにより、ExpEther は従来の I/O に対して、I/O のホストに対する割り当てを柔軟に変更可能とする。また、本稿で提案した I/O の同時共有方式により、現在 PCI-SIG で標準化が行われている MR-IOV (Multi-Root I/O Virtualization)に対応する I/O に加え[6]、SR-IOV に対応する I/O をホスト間で同時に共有することが可能となる。

謝辞

本研究の一部は、総務省の委託研究「次世代バックボーンに関する研究開発」プロジェクトの成果である。

参考文献

[1]Cisco White Paper, "Cisco VFrame and VMware Integration," 2007.

[2]J. Suzuki *et al.*, 14th IEEE Symposium on High-Performance Interconnects (HotI), pp. 45-51, 2006.

[3]PCI-SIG, "Single Root I/O Virtualization and Sharing Specification Revision 1.0."

[4]http://dast.nlanr.net/projects/Iperf/

[5]H. Raj *et al.*, 16th international symposium on High performance distributed computing (HPDC), pp. 179-188, 2007.

[6]PCI-SIG, "Multi-Root I/O Virtualization and Sharing Specification Revision 0.9."