

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号
特許第6287864号
(P6287864)

(45) 発行日 平成30年3月7日(2018.3.7)

(24) 登録日 平成30年2月16日(2018.2.16)

(51) Int.Cl.	F I
H O 4 L 12/28 (2006.01)	H O 4 L 12/28 2 O O D
H O 4 L 12/865 (2013.01)	H O 4 L 12/865
H O 4 L 12/70 (2013.01)	H O 4 L 12/70 D
H O 4 L 12/851 (2013.01)	H O 4 L 12/851
H O 4 L 12/46 (2006.01)	H O 4 L 12/46 V

請求項の数 10 (全 28 頁)

(21) 出願番号	特願2014-556279 (P2014-556279)	(73) 特許権者	000004237
(86) (22) 出願日	平成25年6月26日 (2013. 6. 26)		日本電気株式会社
(65) 公表番号	特表2015-527755 (P2015-527755A)		東京都港区芝五丁目7番1号
(43) 公表日	平成27年9月17日 (2015. 9. 17)	(74) 代理人	100109313
(86) 国際出願番号	PCT/JP2013/004014		弁理士 机 昌彦
(87) 国際公開番号	W02014/002500	(74) 代理人	100124154
(87) 国際公開日	平成26年1月3日 (2014. 1. 3)		弁理士 下坂 直樹
審査請求日	平成28年5月16日 (2016. 5. 16)	(72) 発明者	辻 聡
(31) 優先権主張番号	特願2012-147327 (P2012-147327)		東京都港区芝五丁目7番1号
(32) 優先日	平成24年6月29日 (2012. 6. 29)		日本電気株式会社内
(33) 優先権主張国	日本国 (JP)		
		審査官	衣鳩 文彦

最終頁に続く

(54) 【発明の名称】 Q o S 制御システム、Q o S 制御方法及びプログラム

(57) 【特許請求の範囲】

【請求項1】

トラフィックを識別する識別情報と、当該トラフィックに属するパケットに適用する優先度値と、前記パケット内に前記優先度値を格納する位置を示す情報と、を定めたQ o S (Quality of Service) パラメータを保持し、前記トラフィックに属するパケットに優先度値を格納してから送信するネットワークインタフェースと、

各トラフィックに適用するQ o S パラメータを含んだQ o S ポリシーを保持して、前記ネットワークインタフェースを管理するネットワークインタフェース管理部と、を含み、
前記ネットワークインタフェースは、さらに、

優先度が設定された1つ以上のキューから構成される優先度キューと、前記優先度キューを構成するキューの優先度に応じて、パケットの送信をスケジューリングするスケジューラと、を備える

Q o S 制御システム。

【請求項2】

前記ネットワークインタフェースは、外部からのパケット送信要求に応じてメモリからパケットを取得し、前記パケットに適合する識別情報を持つ前記Q o S パラメータを用いて前記パケットに優先度値を格納してから送信する請求項1のQ o S 制御システム。

【請求項3】

前記ネットワークインタフェースは、送信パケットに適合する識別情報を持つ前記Q o S パラメータを保持していない場合、前記ネットワークインタフェース管理部に対して前

記 Q o S パラメータの送信を要求する請求項 1 又は 2 の Q o S 制御システム。

【請求項 4】

前記ネットワークインタフェース管理部は、前記 Q o S ポリシーから、前記ネットワークインタフェースからの前記 Q o S パラメータの送信要求を受けたパケットと同じ識別情報を持つ前記 Q o S パラメータを取得し、前記ネットワークインタフェースに通知する請求項 3 の Q o S 制御システム。

【請求項 5】

前記ネットワークインタフェースは、前記優先度キューを指定するための情報に基づいて、前記 Q o S パラメータを適用したパケット、パケットヘッダ、またはパケット送信要求のいずれかを前記優先度キューのいずれかに格納する

10

請求項 1 乃至請求項 4 のいずれか一項に記載の Q o S 制御システム。

【請求項 6】

前記ネットワークインタフェースは、前記ネットワークインタフェース管理部に対して、前記優先度キューの使用状況を通知する機能を有する請求項 5 に記載の Q o S 制御システム。

【請求項 7】

前記ネットワークインタフェース管理部は、前記 Q o S ポリシーとして、さらに、前記優先度キューを構成するキューの識別情報を保持し、

前記ネットワークインタフェースから通知された前記優先度キューの使用状況に基づいて、前記キューの識別情報と一致するキューの識別情報を持つ Q o S ポリシーを選択し、

20

パケット送信要求の発行元に対して、前記選択した Q o S ポリシーを通知し、
前記パケット送信要求の発行元は、前記 Q o S ポリシーを含む通知に従って、パケット送信要求の抑制又はパケット送信要求の送信再開を行う請求項 6 の Q o S 制御システム。

【請求項 8】

前記ネットワークインタフェースは、前記ネットワークインタフェースにて優先度値の付与を行うか否かを設定可能であり、

前記ネットワークインタフェース管理部からの前記優先度値の付与の可否の設定を変更することにより、前記ネットワークインタフェースで優先度値の付与を行うか否かを制御する請求項 1 から 7 のいずれかの Q o S 制御システム。

【請求項 9】

30

各トラフィックに適用する Q o S (Quality of Service) パラメータを含んだ Q o S ポリシーを参照して、優先度が設定された 1 つ以上のキューから構成される優先度キューを有するネットワークインタフェースに対して、トラフィックを識別する識別情報と、当該トラフィックに属するパケットに適用する優先度値と、前記パケット内に前記優先度値を格納する位置を示す情報と、を含む前記 Q o S パラメータを通知するステップと、

前記ネットワークインタフェースに、前記優先度キューを構成するキューの優先度に応じて、パケットの送信をスケジューリングさせるステップと、

そのスケジューリングの結果に応じて、前記ネットワークインタフェースに、前記 Q o S パラメータを用いて、前記識別情報に適合するパケットに優先度値を格納してから送信させるステップと、を含む

40

Q o S 制御方法。

【請求項 10】

各トラフィックに適用する Q o S (Quality of Service) パラメータを含んだ Q o S ポリシーを参照して、優先度が設定された 1 つ以上のキューから構成される優先度キューを有するネットワークインタフェースに対して、トラフィックを識別する識別情報と、当該トラフィックに属するパケットに適用する優先度値と、前記パケット内に前記優先度値を格納する位置を示す情報と、を含む前記 Q o S パラメータを通知する処理と、

前記ネットワークインタフェースに、前記優先度キューを構成するキューの優先度に応

50

じて、パケットの送信をスケジューリングさせる処理と、

そのスケジューリングの結果に応じて、前記ネットワークインタフェースに、前記QoSパラメータを用いて、前記識別情報に適合するパケットに優先度値を格納してから送信させる処理と、をコンピュータに実行させる、
コンピュータ・プログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、日本国特許出願：特願2012-147327号（2012年6月29日出願）の優先権主張に基づくものであり、同出願の全記載内容は引用をもって本書に組み込み記載されているものとする。

【0002】

本発明は、QoS（Quality of Service）制御システム、QoS制御方法及びプログラムに関し、特に、I/O仮想化技術を使用可能なサーバ仮想化環境に好適なQoS制御システム、QoS制御方法及びプログラムに関する。

【背景技術】

【0003】

近年、サーバの性能向上、信頼性向上により、サーバ資源の仮想化（サーバ仮想化）の導入がデータセンタなどで進んでいる。サーバ仮想化技術により、これまで個々の物理的なサーバで稼動させていたサービスを仮想マシン上で稼動させ、この仮想マシンを1台の物理的なサーバに収容することで、物理的なサーバの台数を削減することができる。これによって、サーバの導入コストや運用コストを削減することが可能となる。

【0004】

図15は、サーバ仮想化技術を導入する前のデータセンタのネットワークアーキテクチャの一例を示したものである。図15によると、データセンタのネットワークは大きく分けて3層に分かれており、サーバ10が接続されるTop-of-Rackスイッチ（TORスイッチ）12、TORスイッチ12同士を接続するアグリゲーションスイッチ14、アグリゲーションスイッチ14とデータセンタ外部のネットワークを接続するコアスイッチ16から構成される。サーバ10はネットワークインタフェースカード（NIC）18を介してTORスイッチ12に接続される。

【0005】

一方、図16は、図15の環境にサーバ仮想化技術を導入したネットワークアーキテクチャを示したものである。図16によると、サーバ10内に1つ以上の仮想マシン20が稼動しており、仮想マシン20は、仮想スイッチ22を介して、サーバ10外部のネットワークと接続される。仮想スイッチ22はソフトウェアで実現されたスイッチであり、各仮想マシン20とNIC18とを接続し、仮想マシン20に対して、サーバ10内部の仮想マシン20同士の通信、およびサーバ10外部のネットワークへの接続を提供する。つまり、サーバ仮想化技術を導入したデータセンタでは、TORスイッチ12の先にさらにネットワークスイッチが接続されたアーキテクチャとなる。

【0006】

サーバ仮想化環境では、仮想化のオーバーヘッドが存在するため、仮想マシン20に高い通信性能を提供することが難しいという課題がある。これを解決するために、CPU（Central Processing Unit）やチップセット、NIC18がI/Oの仮想化をサポートするようになってきた。このような技術の例として、SR-IOV（Single Route I/O Virtualization）やIOMMU（Input Output Memory Management Unit）といったI/O仮想化技術が存在する。

【0007】

このような技術を用いると、仮想マシン20は仮想スイッチ22を介さずに直接NIC18にアクセスすることが可能となり、仮想化のオーバーヘッドをなくすることができる。結

10

20

30

40

50

果として、仮想マシン20に高い性能を提供することができる。

【0008】

図17は、I/O仮想化技術を用いたサーバ10内のアーキテクチャを示したものである。図17を参照すると、サーバ10で稼動する仮想マシン20のいくつかが仮想スイッチ22を介さずに、直接NIC18に接続されている。仮想スイッチ22は、サーバ仮想化環境では、一般的に仮想マシンモニタ24、あるいはハイパーバイザーと呼ばれる、サーバ仮想化環境を管理する領域で稼動する。NIC18はI/O仮想化をサポートするために、複数のインタフェースを提供する。仮想マシンモニタ24と接続するための物理インタフェース180と、仮想マシン20と接続するための仮想インタフェース181を備え、仮想インタフェース181は1枚のNIC18に対して、複数備えられる。図17において、仮想マシン20のうち、VM aとVM bが仮想インタフェース181を介して、直接NIC18に接続されている。

10

【0009】

このようなデータセンタの利用方法の1つとして、IaaS (Infrastructure as a Service) が存在する。IaaSでは、データセンタの物理資源をサーバ仮想化技術やVLAN (Virtual LAN) などの技術を用いて論理的に分割し、異なるユーザに計算機とネットワークの環境を提供する。例えば、計算機環境は仮想マシン20の形で、ネットワークはVLANの形で提供される。ユーザは提供された仮想マシン20にOS (Operating System) を導入し、ネットワークを介して、仮想マシン20同士を通信させ、何らかのサービスを実現することが可能である。

20

【0010】

IaaSのような環境では、複数のユーザが同一の環境を利用するため、計算資源やネットワーク資源の適切な割り当てと、アイソレーションが重要である。例えば、ネットワーク資源であれば、ユーザごとのQoS制御を行うことで、あるユーザの通信が、別のユーザの通信に影響を与えないことが求められる。I/O仮想化技術を用いない場合、仮想スイッチ22にて、QoS制御を行うことができる。QoS制御の一例として、MAC (Media Access Control) フレームに含まれるCoS (Class of Service) フィールドや、IP (Internet Protocol) パケットのToS (Type of Service) フィールド、DSCP (DiffServ Code Point) フィールドに設定された値に基づいて、フレームやパケットを優先度付けされた複数のキューに振り分ける方法が存在する。これらの値は、QoS制御を行うネットワーク区間の入り口となるスイッチにて、IPヘッダに含まれるIPアドレスなどの情報を元に設定される。この設定処理をマーキングと呼ぶ。スイッチはキューの優先度に基づいて、フレームやパケットの転送をスケジューリングしたり、リンク帯域を割り当てたりする。仮想スイッチ22でマーキングを行うと、ToRスイッチ12がマーキングの処理をサポートする必要がなくなるため、データセンタの構築コストを低減できるというメリットがある。

30

【0011】

ところが、I/O仮想化技術を用いた場合、ToRスイッチ12に、仮想スイッチ22にて本来行われるはずだったQoS制御の設定がなされていないフレームやパケットが届くことになる。また、I/O仮想化技術を利用する仮想マシン20 (図17のVM a、VM b) はNIC18を介してパケットを直接ネットワークに送出することができるため、ユーザに都合の良いようにマーキングしたパケットを送信することが可能である。例えば、図17のVM aのユーザが全てのパケットを最高の優先度を持つように設定することも可能である。パケットのフォーマット上は正常なパケットがサーバ10から送出されるため、ToRスイッチ12からすると、サーバ10から送出されたパケットが仮想スイッチ22を経由したものであるかどうかを判断できない。結果として、データセンタ管理者のQoSポリシーに沿わないQoS制御が行われる可能性がある。

40

【0012】

50

さらに、サーバ10とTORスイッチ12の間の物理リンクの帯域が、ユーザ間で適切に共有されない可能性がある。先に述べたように、I/O仮想化技術を用いると、仮想スイッチ22が提供するQoS制御がきかなくなるため、I/O仮想化技術を利用する仮想マシン20がリンクを占有することが可能となってしまう。

【0013】

このような問題を解決するための技術の一例が、特許文献1に記載されている。特許文献1では、ユーザの仮想マシンとI/O仮想化をサポートしたI/Oアダプタの間に、仮想マシンのI/Oを制御するI/O論理パーティションが存在し、システム管理者が所定のポリシーに従ってI/O論理パーティションを制御することで、仮想マシンに適切なI/O制御を提供することができる。I/O論理パーティションはI/Oデバイスの仮想インターフェースと接続される。

10

【0014】

また、前述の問題を解決するための技術の別の例が、特許文献2に記載されている。特許文献2では、NICを仮想化し、複数の優先度付けされた仮想ネットワークデバイスとする技術が記載されている。トラフィックフローを優先度に基づいて各仮想ネットワークデバイスに割り当てることで、優先度の高いトラフィックフローが、優先度の低いトラフィックフローに影響されることを低減することができる。

【先行技術文献】

【特許文献】

【0015】

【特許文献1】特開2012-073660号公報

【特許文献2】特開2010-283828号公報

【発明の概要】

【発明が解決しようとする課題】

【0016】

以下の分析は、本発明によって与えられたものである。

【0017】

上述した特許文献1に示される技術には、以下に示す問題点が存在する。すなわち、ユーザの仮想マシンと、I/Oデバイスとの間にI/Oを制御する管理領域が存在する、つまり、従来の仮想化環境で仮想マシンモニタを経由して通信を行うのと同様であるため、I/O仮想化をサポートすることのメリットである高い性能が得られない。

30

【0018】

さらに、上述した特許文献2に示される技術には、以下に示す問題点が存在する。すなわち、本技術をサーバ仮想化環境に適用すると、各仮想マシンに複数の仮想ネットワークデバイスを割り当て、トラフィック単位に仮想ネットワークデバイスを使い分けることと等しくなる。各仮想ネットワークデバイスがI/O仮想化技術により、仮想マシンから直接のアクセスが可能になるとすると、仮想ネットワークデバイスの選択は仮想マシンに委ねられることになり、ユーザの都合に合わせて仮想ネットワークデバイスが使用されることとなる。結果として、ネットワークがシステム管理者のQoSポリシーとは異なる形で利用されることになる。

40

【0019】

先に述べたI/O仮想化技術を用いることで発生する2つの問題は、I/O仮想化技術により、複数の仮想マシン20が、TORスイッチ12の各ポートに直接接続された状態となっているにもかかわらず、仮想マシン20とTORスイッチ12の間に、制御するポイントが存在しないことが原因である。これによって、仮想マシン20の間で通信を調停することができなくなっている。TORスイッチ12の1つのポートに複数のサーバが接続される状況は、通常は起こりえないため、特許文献1、2等に記載された技術ではこれらの問題を解決できない。

【0020】

本発明の目的は、I/O仮想化技術を使用可能なサーバ仮想化環境において、各仮想マ

50

シンが I / O 仮想化技術を利用して通信を行う場合においても、システム管理者が定める Q o S ポリシーに従って、トラフィックの Q o S 制御が可能な Q o S 制御システム、Q o S 制御方法及びプログラムを提供することにある。

【課題を解決するための手段】

【0021】

本発明の第1の視点によれば、トラフィックを識別する識別情報と、当該トラフィックに属するパケットに適用する優先度値と、前記パケット内に前記優先度値を格納する位置を示す情報と、を定めた Q o S (Q u a l i t y o f S e r v i c e) パラメータを保持し、前記トラフィックに属するパケットに優先度値を格納してから送信するネットワークインタフェースと、各トラフィックに適用する Q o S パラメータを含んだ Q o S ポリ

10

【0022】

本発明の第2の視点によれば、Q o S (Q u a l i t y o f S e r v i c e) ポリシーを参照して、ネットワークインタフェースに対して、トラフィックを識別する識別情報と、当該トラフィックに属するパケットに適用する優先度値と、前記パケット内に前記優先度値を格納する位置を示す情報と、を含む Q o S パラメータを通知するステップと、前記ネットワークインタフェースに、前記 Q o S パラメータを用いて、前記識別情報に適合するパケットに優先度値を格納してから送信させるステップと、を含む Q o S 制御方法が提供される。本方法は、送信パケットに優先度値を格納してから送信するコンピュータ

20

【0023】

本発明の第3の視点によれば、Q o S (Q u a l i t y o f S e r v i c e) ポリシーを参照して、ネットワークインタフェースに対して、トラフィックを識別する識別情報と、当該トラフィックに属するパケットに適用する優先度値と、前記パケット内に前記優先度値を格納する位置を示す情報と、を含む Q o S パラメータを通知する処理を Q o S 制御システムを構成するコンピュータに実行させ、前記ネットワークインタフェースに、前記 Q o S パラメータを用いて、前記識別情報に適合するパケットに優先度値を格納してから送信させるプログラムが提供される。なお、このプログラムは、コンピュータが読み取り可能な（非ランジエントな）記憶媒体に記録することができる。即ち、本発明は、

30

【発明の効果】

【0024】

本発明によれば、I / O 仮想化技術を使用可能なサーバ仮想化環境において、各仮想マシンが I / O 仮想化技術を利用して通信を行う場合における Q o S ポリシーに従ったトラフィックの Q o S 制御が可能となる。

【図面の簡単な説明】

【0025】

【図1】本発明の第1の実施形態に係る Q o S 制御システムの構成を示した図である。

【図2】本発明の第1の実施形態に係る Q o S 制御システムのパケット転送処理（N I C のフローテーブルに該当エントリが存在しない場合）を示したシーケンス図である。

40

【図3】本発明の第1の実施形態に係る Q o S 制御システムのパケット転送処理（N I C のフローテーブルに該当エントリが存在する場合）を示したシーケンス図である。

【図4】本発明の第1の実施形態に係る Q o S 制御システムの N I C の動作を表したフローチャートである。

【図5】本発明の第1の実施形態に係る Q o S 制御システムの仮想マシンモニタの動作を表したフローチャートである。

【図6】本発明の第2の実施形態に係る Q o S 制御システムの構成を示した図である。

【図7】キューの優先度に基づいて優先度キューに格納されたパケットを送出する処理を示したフローチャートである。

50

【図 8】本発明の第 2 の実施形態に係る Q o S 制御システムの変形構成（優先度キュー容量小）を示した図である。

【図 9】図 8 の構成における優先度キューを用いたパケット送出処理を示したフローチャートである。

【図 1 0】本発明の第 3 の実施形態に係る Q o S 制御システムの構成を示した図である。

【図 1 1】本発明の第 3 の実施形態に係る Q o S 制御システムの動作（スレーブポリシーテーブルへのエントリ追加）を示したシーケンス図である。

【図 1 2】本発明の第 3 の実施形態に係る Q o S 制御システムの別の動作（スレーブポリシーテーブルからのエントリ削除）を示したシーケンス図である。

【図 1 3】本発明の第 3 の実施形態に係る Q o S 制御システムの仮想マシンのパケット送信要求の発行処理を示したシーケンス図である。

【図 1 4】本発明の第 3 の実施形態に係る Q o S 制御システムの仮想マシン 2 0 のパケット送信要求の発行処理を示したフローチャートである。

【図 1 5】サーバ仮想化技術を導入する前のデータセンタのネットワークアーキテクチャの一例を示した図である。

【図 1 6】図 1 5 の環境にサーバ仮想化技術を導入したネットワークアーキテクチャを示した図である。

【図 1 7】I / O 仮想化サポート技術を用いたサーバ内のアーキテクチャを示した図である。

【発明を実施するための形態】

【0 0 2 6】

はじめに本発明の一実施形態の概要について図面を参照して説明する。なお、この概要に付記した図面参照符号は、理解を助けるための一例として各要素に便宜上付記したものであり、本発明を図示の態様に限定することを意図するものではない。

【0 0 2 7】

本発明は、その一実施形態において、ネットワークインタフェースカード上に Q o S パラメータを保持する手段（図 1 の 1 8 5）と、パケットをホストメモリから取得する手段（図 1 の 1 8 3）と、取得したパケットのパケットヘッダに対して、優先度値を格納してから送信する手段（図 1 の 1 8 2）と、前記ネットワークインタフェースカードを管理する手段（図 1 の 2 4）とを備える構成にて実現できる。

【0 0 2 8】

より具体的には、Q o S パラメータは、トラフィックを識別する識別情報と、当該トラフィックに属するパケットに適用する優先度値と、前記パケット内に前記優先度値を格納する位置を示す情報と、を含んでいる。前記ネットワークインタフェースカードを管理する手段は、このような Q o S パラメータを含んだ Q o S ポリシー（図 1 の 2 4 0）を保持している。前記ネットワークインタフェースカードを管理する手段は、Q o S ポリシーに基づいて、Q o S パラメータを、ネットワークインタフェースカードに通知する。

【0 0 2 9】

ネットワークインタフェースカードは、仮想マシンから発行されるパケット送信要求を受けると、ホストメモリからパケットを取得する。次に、ネットワークインタフェースカードは、前記ネットワークインタフェースカードを管理する手段から通知された Q o S パラメータを、前記取得したパケットのパケットヘッダの情報を用いて、前記 Q o S パラメータを保持する手段を参照することで取得し、Q o S パラメータに含まれる優先度値を取得する。ネットワークインタフェースカードは、該パケットのパケットヘッダに前記取得した優先度値を格納してから送信する。なおここで、Q o S パラメータを保持する手段が、ホストメモリから取得したパケットに適合する Q o S パラメータを保持していない場合、ネットワークインタフェースカードは、前記ネットワークインタフェースカードを管理する手段に対し、適切な前記 Q o S パラメータを問い合わせ、適切な前記 Q o S パラメータを取得する。このようにして、ネットワークインタフェースカードに設定された優先度値が送信パケットに適用される。

【0030】

以上のように、I/O仮想化技術を用いて、仮想マシンがネットワークインタフェースカードに対して直接、パケットの送信要求を発行し、パケットを送信できるような環境においても、システム管理者のQoSポリシーが適用されて、パケットが送信されるようになる。その理由は、ネットワークインタフェースカードが、パケット送信要求に従って取得したパケットのヘッダ情報を元にして、QoSパラメータが格納されたテーブルを参照し、システム管理者のポリシーに従ったQoSパラメータをパケットヘッダに適用するよう構成したためである。

[第1の実施形態]

続いて、本発明の第1の実施形態について図面を参照して詳細に説明する。なお、以下の説明では、ネットワークに転送するデータの単位を「パケット」に統一して表記するが、このことは、本技術が適用されるネットワークの種類等を制限するものではない。また、以降、「フロー」とも呼ばれる一連のパケット列を「トラフィック」と表記する。例えば、FTP (File Transfer Protocol) によって、あるファイルをダウンロードする際に、ファイルのダウンロード開始から完了までに流れたパケットをまとめてトラフィックとする。

【0031】

図1は、本発明の第1の実施形態に係るQoS制御システムの構成を示した図である。図1を参照すると、NIC18T-1と、仮想マシン(VM)20と、仮想マシンモニタ(VMM)24と、を含んだQoS制御システム10aが示されている。仮想マシン20と、仮想マシンモニタ24とは、サーバ10の環境下で動作するコンピュータプログラムである。いずれもQoS制御システム10aを構成するサーバ等のCPU、メモリ、ディスク、I/Oといった各種資源を利用して実現される。以降、QoS制御システム10aを構成するサーバ等のメモリを「ホストメモリ」と表記する。

【0032】

NIC18T-1は、NIC18の送信側の構成を示したものであり、物理インタフェース180と、仮想インタフェース181と、TXコントローラ(TX Ctrl)182と、DMAコントローラ(DMAC)183と、ホストバスインタフェース(Host Bus I/F)184と、フローテーブル185と、TXインタフェース(TX I/F)188と、を含んでいる。

【0033】

物理インタフェース180は、I/O仮想化技術における、仮想マシンモニタ24に対するインタフェースであり、NIC18T-1の動作を制御するためのインタフェースを備える。例えば、フローテーブル185に設定するエントリを通知する、仮想インタフェース181のステータスの読み出し、などが挙げられる。物理インタフェース180の実装例としては、SR-IOVのフィジカルファンクションが挙げられる。SR-IOVにおいて、フィジカルファンクションはPCI-Expressのインタフェースが提供する一般的な機能を全て備える。

【0034】

仮想インタフェース181は、仮想マシン20がI/O仮想化技術を用いてNIC18T-1に直接アクセスするためのインタフェースである。仮想インタフェース181の実装例としては、SR-IOVのバーチャルファンクションが挙げられる。SR-IOVにおいて、バーチャルファンクションはフィジカルファンクションの一部のリソースのみを利用して実現される。

【0035】

仮想マシン20は、仮想インタフェース181を通じてパケット送信要求をNIC18T-1に発行する。パケット送信要求には、少なくとも送信パケットサイズと、送信パケットが格納されたホストメモリのアドレスと、が含まれる。場合によっては、送信パケットがホストメモリ上で連続な領域に格納されていない場合がある。この場合は、パケット送信要求に、サイズとアドレスをリストとして、複数個含めるが、本発明に本質的な差異

を与えないため、ホストメモリ上の連続した領域にパケットが格納されているとして説明する。

【0036】

TXコントローラ182は、NIC18T-1のパケット送信処理を制御するコントローラである。TXコントローラ182は、内部にパケットバッファを備え、一定数のパケットを保持できることが望ましい。TXコントローラ182は、仮想インタフェース181から発行されるパケット送信要求から、送信パケットサイズ、送信パケットが格納されたホストメモリのアドレスを取得し、DMAコントローラ183に対して送信パケットの読み出し要求を発行する。続いて、読み出されたパケットに対して、QoSにおける優先度を示す値（以降、優先度値と表記する）をセット（マーキング）し、TXインタフェース188にパケットを転送する。

10

【0037】

フローテーブル185は、送信パケットに対して設定するQoSのパラメータを保持するテーブルである。QoSのパラメータとして、少なくとも、優先度値をパケットヘッダのどのフィールドにセットするかを示す情報と、セットする優先度値が含まれる。例えば、CoSフィールド／ToSフィールド／DSCPフィールドのいずれかを指定する値と、指定されたフィールドにセットする優先度値と、がQoSのパラメータに少なくとも含まれる。また、Cisco Systems社が実装するISL（Inter Switch Link）方式のように、ベンダ独自の方式でトラフィックの優先度付けを可能にしている場合もあるため、例えば、パケットヘッダの先頭からのバイト数／ビット数と、セットする優先度値、セットする優先度値のサイズ（バイト数／ビット数）をフローテーブルに保持することも考えられる。このことは、NIC18T-1がサポートする機能に依存する話であるが、本実施の形態に実質的な差異を与えない。

20

【0038】

フローテーブル185の各エントリは、QoSのパラメータを検索するキーとして、パケットのヘッダ情報に含まれる、トラフィックの識別情報を、QoSのパラメータと対応付けて保持する。トラフィックの識別情報として、パケットヘッダに含まれるIPアドレスやTCP（Transmission Control Protocol）やUDP（User Datagram Protocol）のポート番号、などが挙げられる。

【0039】

フローテーブル185は、例えば、TCAM（Ternary Content Addressable Memory）のような形で実現することができる。

30

【0040】

TXコントローラ182は、取得したパケットのヘッダの情報を用いて、フローテーブル185を参照し、パケットにセットする優先度値を取得する。TXコントローラ182は、取得した優先度値を用いて、パケットに対してマーキングを行い、TXインタフェース188にパケットを転送する。

【0041】

ネットワークによっては、パケットにエラー検出用のデータが付与されている場合がある。例えば、イーサネット（登録商標）においては、パケットの末尾にFCS（Frame Check Sequence）が付与されており、パケットの宛先MACアドレスから、ペイロード部の末尾までのデータを用いてCRC（Cyclic Redundancy Check）値を計算した結果が付与される。マーキングを行うと、パケットの内容が書き換わるため、TXコントローラ182にて、エラー検出用のデータの再計算を行い、パケットの適切なフィールドに適用することが望ましい。

40

【0042】

フローテーブル185に、該当するエントリが存在しない場合、TXコントローラ182は、物理インタフェース180を通じて、仮想マシンモニタ24に対して、このパケットに適用するQoSパラメータの問い合わせを行う。この問い合わせには、パケットヘッダの情報のうち、フローテーブル185に登録可能な情報（上記トラフィックの識別情報

50

として用いる情報)が含まれる。例えば、TXコントローラ182は、ホスト側のCPUに対する割り込みといった形で、仮想マシンモニタ24に対して、問い合わせを行ったことを示す通知を行う。

【0043】

DMAコントローラ183は、TXコントローラ182からの要求に従って、ホストバスデータインタフェース184を通じて、メモリに格納されたデータを読み出す処理を行う。

【0044】

ホストバスインタフェース184は、サーバ10と、NIC18T-1とを接続するデータバスとのインタフェースであり、例えば、PCI-Expressにおいて、DMA転送用のコマンド発行のためのインタフェースや、ホスト側から転送されてきたデータを格納するレジスタなどが該当する。

【0045】

TXインタフェース188は、NIC18T-1のネットワークへの物理的なインタフェースであり、TXコントローラ182から受け取ったパケットをネットワークに送出する。その際に、TOSスイッチ12がパケットの始まりを検出するための信号をパケットの前に付加するなどの処理を行う。例えば、ネットワークがイーサネットであれば、プリアンブル部がそれに当たる。

【0046】

仮想マシン20は、システムの管理者からアサインされたサーバ10の資源(CPUやメモリ、ディスク、I/Oインタフェースなど)を用いて、ユーザがアプリケーションを動かすことのできる、仮想的な計算機である。一般に、IaaS環境では、ユーザは自身でOSを用意し、仮想マシン20に導入することで、アプリケーションの実行環境を用意する。

【0047】

仮想マシンモニタ24は、仮想スイッチ22のほか、ポリシーテーブル240が備えられる。仮想マシンモニタ24は、NIC18T-1からの問い合わせに応じて、ポリシーテーブル240の内容を参照し、適切なQoSパラメータを選択し、NIC18T-1に通知する。

【0048】

ポリシーテーブル240には、ネットワーク管理者が定めた、トラフィックに対するプライオリティのポリシーが格納される。ポリシーの例として、パケットのIPアドレスや、TCP/UDPのポート番号のような、トラフィックを識別する情報と、そのトラフィックの優先度値と、その優先度値をセットするパケットの位置を示す情報と、から構成する方式が考えられる。この場合、ポリシーテーブル240は、フローテーブル185のスーパーセットとなる。別のポリシーの例として、優先度値に、COSやTOSなどの特定の優先度付けを行うための方式と非依存な値を保持するということも考えられる。この場合、ポリシーテーブル240を参照して得られた優先度値を、実際に使用するQoS制御の方式に合った設定に変換する処理が必要である。例えば、ポリシーテーブル240に、優先度を0~63の64段階(=6ビット)で保持するとし、ネットワーク側で使用するQoS制御の仕組みがTOSである場合、TOSの優先度は8段階であるため、{ポリシーテーブル240の値、TOSの優先度} = {0~7、0} {8~15、1} {16~23、2} {24~31、3} {32~39、4} {40~47、5} {48~55、6} {56~63、7} のように対応させる。

【0049】

TXコントローラ182は、フローテーブル185に該当するエントリが存在しないパケットを扱う場合に、仮想マシンモニタ24に対し、パケットに適用するQoSパラメータを問い合わせる。仮想マシンモニタ24が、ポリシーテーブル240を参照し、設定するQoSパラメータをTXコントローラ182に通知すると、TXコントローラ182は、通知されたQoSパラメータをフローテーブル185に登録する。また、以後、TXコ

10

20

30

40

50

ントローラ182は、通知されたQoSパラメータに含まれる優先度値を、対象のパケットに適用する。この通知は、物理インタフェース180のレジスタを介して行うことが例として考えられる。つまり、仮想マシンモニタ24が、物理インタフェース180の特定のレジスタにQoSパラメータを書き込むことで、TXコントローラ182にトリガがかかり、QoSパラメータのフローテーブル185への登録と、パケットへの適用が行われるように実現することが考えられる。このような仕組みを用いると、仮想マシンモニタ24が、事前にフローテーブル185にフローエントリを登録しておくことも可能である。例えば、通信頻度の高いトラフィックに対するポリシーをあらかじめフローテーブル185に登録しておくことで、仮想マシンモニタ24への問い合わせを削減することができる。

10

【0050】

次に、図2、図3のシーケンス図、図4、図5のフローチャートを参照して、本実施形態の全体の動作について詳細に説明する。

【0051】

図2は、本発明の第1の実施形態に係るQoS制御システムのパケット転送処理（NICのフローテーブルに該当エントリが存在しない場合）を示したシーケンス図である。具体的には、図2は、仮想マシン20がパケットの送信要求を発行したものの、NIC18T-1のフローテーブル185に、送信するパケットに該当するエントリが存在しない場合のパケット転送処理を示している。

【0052】

20

まず、仮想マシン20が、仮想インタフェース181を通じて、パケットの送信要求をNIC18T-1に発行する（図2のステップS200）。送信要求には、少なくとも転送するパケットのサイズと、そのパケットが格納されたホストメモリのアドレスが含まれている。発行されたパケット送信要求はNIC18T-1に備えられるTXコントローラ182で解釈される。

【0053】

次に、TXコントローラ182は、仮想マシン20からパケット送信要求を受け取ると、パケット送信要求に含まれるパケットのサイズと、メモリアドレスと、の情報を用いて、DMAコントローラ183に対し、パケットの読み出し要求を発行し、パケットを取得する（図2のステップS182）。取得したパケットは、TXコントローラ182が備えるパケットバッファに格納される。

30

【0054】

次に、TXコントローラ182は、フローテーブル185から、取得したパケットのヘッダ情報に適合するトラフィック識別情報を持つエントリを検索する。この検索の結果、該当するエントリが存在しないことが分かると、TXコントローラ182は、仮想マシンモニタ24に対し、物理インタフェース180を通じて、このパケットに適用するQoSパラメータの問い合わせを行う（図2のステップS185）。この問い合わせには、少なくともパケットヘッダのフィールドのうち、フローテーブル185を検索する際に用いることができるキーと同じフィールドの情報、即ち、上記トラフィックの識別情報として用いる情報が含まれる。

40

【0055】

次に、仮想マシンモニタ24は、NIC18T-1からQoSパラメータの問い合わせがあると、問い合わせに含まれるパケットヘッダの情報を用いて、ポリシーテーブル240を参照し、このパケットに適用するQoSパラメータを取得する（図2のステップS240）。先に述べたように、ポリシーテーブル240に保持されているポリシーが、特定の優先度付けを行うための方式と非依存である場合、ポリシーテーブル240から取得したポリシーを、ネットワーク側で使用する優先度付けの方式に対応させる処理を実行する。

【0056】

次に、仮想マシンモニタ24は、物理インタフェース180を通じてNIC18-T1

50

に対して、取得したQoSパラメータを通知する（図2のステップS242）。

【0057】

次に、TXコントローラ182は、仮想マシンモニタ24から通知されたQoSパラメータをフローテーブル185に登録する（図2のステップS186）。

【0058】

次に、TXコントローラ182は、取得したQoSパラメータに含まれる優先度値をパケットヘッダに適用する（図2のステップS187）。このときに、TXコントローラ182にて、パケットのエラー検出用のデータの再計算を行い、パケットの該当するフィールドに適用する。

【0059】

次に、TXコントローラ182は、優先度値をセットしたパケットをTXインタフェース188に転送し、ネットワークにパケットを送出する（図2のステップS188）。

【0060】

なお、図2のステップS186と、図2のステップS187～ステップS188の処理に依存は無いため、処理手順が入れ替わっても良い。つまり、図2のステップS187～ステップS188の処理を実行した後に、図2のステップS186の処理時を実行することも、同時に実行することも可能である。

【0061】

ここで、TXコントローラ182が、パケット全体を保持するバッファを備えない場合について、動作を説明する。この場合、図2のステップS182でパケット全体を読み出したとしても、図2のステップS185～ステップS186のQoSパラメータを問い合わせている間、パケットのパケットヘッダ以降の領域（ペイロード部）を保持できないため、破棄することになる。結果として、QoS制御システム10aを構成するサーバのPCI-Expressなどの内部バス帯域を無駄に使用してしまうことになる。

【0062】

この場合、まず、TXコントローラ182は、図2のステップS182において、DMAコントローラ183に対してパケットヘッダのみを読み出す要求を発行し、送信するパケットのパケットヘッダを取得する。この場合、パケットヘッダのサイズが事前に明らか場合は、そのサイズを利用することが可能であるが、パケットヘッダのサイズが可変である場合には、パケットの先頭から、固定値でデータサイズを指定することが考えられる。

【0063】

以降の図2のステップS185～ステップS187の処理は、TXコントローラ182が、パケット全体を保持するバッファを備える場合と同一である。その後、ステップS188の処理を実行する際に、TXコントローラ182は、再度DMAコントローラ183に、パケットのペイロード部の読み出し要求を発行し、図2のステップS187でQoSパラメータを適用したパケットヘッダの後続にペイロード部を付加して、パケットを送信する。この場合、パケットのエラー検出のためのデータを計算し、パケットに適用するのは、パケットのペイロード部を読み出したあととなる。

【0064】

以上の動作により、仮想マシン20から発行されたパケット送信要求に従って、NIC18T-1にてパケットに対して、優先度付けを行い、ネットワークにパケットが送出された。

【0065】

図3は、本発明の第1の実施形態に係るQoS制御システムのパケット転送処理（NICのフローテーブルに該当エントリが存在する場合）を示したシーケンス図である。より具体的には、図3は、仮想マシン20がパケットの送信要求を発行し、NIC18T-1のフローテーブル185に、送信するパケットに該当するエントリが存在する場合のパケット転送処理を示している。なお、図2と同様の処理については同一の符号を付し、詳細な説明を省略する。

10

20

30

40

50

【0066】

TXコントローラ182は、仮想マシン20からパケットの送信要求を受けると、DMAコントローラ183経由で送信するパケットのパケットヘッダを取得する（図3のステップS200～ステップS182）。そして、TXコントローラ182は、パケットヘッダの情報を用いてフローテーブル185を参照し、パケットに適用する優先度値を取得する（図3のステップS184）。

【0067】

以下図2と同様に図3のステップS187において優先度値を設定されたパケットは、ネットワークに送出される（図3のステップS188）。

【0068】

図4は、図2、図3のシーケンス図における、TXコントローラ182の処理を詳細に示したフローチャートである。

【0069】

まず、TXコントローラ182は、仮想マシン20からパケット送信要求が発行されると（図4のステップF180のYes）、パケット送信要求に含まれる、送信パケットが格納されたホストメモリのアドレスと、送信するパケットのサイズの情報を用いて、DMAコントローラ183に対し、パケットの読み出し要求を発行する（図4のステップF182）。TXコントローラ182は、仮想マシン20からパケット送信要求が発行されていない場合（図4のステップF180のNo）、処理を開始しない。

【0070】

次に、TXコントローラ182は、フローテーブル185から、前記取得したパケットのヘッダ情報に適合するトラフィック識別情報を持つエントリを検索する（図4のステップF183）。

【0071】

フローテーブル185に該当エントリが存在する場合（図4のステップF183の「エントリが存在する」）、TXコントローラ182は、フローテーブル185の該当エントリから、パケットに適用する優先度値を取得する（図4のステップF184）。

【0072】

一方、フローテーブル185にエントリが存在しない場合（図4のステップF183の「エントリが存在しない」）、TXコントローラ182は、物理インタフェース180を通じて、仮想マシンモニタ24に対して、QoSパラメータを問い合わせる（図4のステップF185）。

【0073】

次に、TXコントローラ182は、仮想マシンモニタ24から、QoSパラメータの通知がなされると（図4のステップF186のYes）、取得したQoSパラメータをフローテーブル185に登録する（図4のステップF187）。一方、仮想マシンモニタ24から、QoSパラメータの通知がなされていない場合（図4のステップF186のNo）には、TXコントローラ182は、仮想マシンモニタ24からの通知を待つ。

【0074】

次に、TXコントローラ182は、取得した優先度値をパケットヘッダに適用する（図4のステップF188）。このときに、TXコントローラ182は、パケットのエラー検出のためのデータを計算し、パケットに適用する。

【0075】

次に、TXコントローラ182は、優先度値を設定したパケットを、TXインタフェースを通じてネットワークに送出する（図4のステップF189）。

【0076】

このようにして、TXコントローラ182は、パケットに優先度付けをして、ネットワークにパケットを送出する。

【0077】

図5は、図2のシーケンス図における、仮想マシンモニタ24の動作を詳細に示したフ

10

20

30

40

50

ローチャートである。

【0078】

まず、仮想マシンモニタ24は、NIC18T-1からQoSパラメータの問い合わせを受けると(図5のステップF238のYes)、NIC18T-1からの問い合わせに応じて、ポリシーテーブル240を参照し、該当するQoSパラメータを取得する(図5のステップF240)。なお、NIC18T-1からQoSパラメータの問い合わせがない場合、仮想マシンモニタ24は、処理を開始しない(図5のステップF238のNo)。

【0079】

次に、仮想マシンモニタ24は、取得したQoSパラメータを、物理インタフェース180を通じて、NIC18T-1に通知する(図5のステップF242)。

【0080】

このようにして、仮想マシンモニタ24は、TXコントローラ182からの問い合わせに応じて、QoSパラメータを通知する。

【0081】

ここで、仮想マシンモニタ24に内蔵される仮想スイッチ22に接続された仮想マシン20(例えば、図17のVMc、及びVMd)が、QoS制御システム10aに配置されている場合の、NIC18T-1の処理について説明する。

【0082】

仮想スイッチ22に接続された仮想マシン20が送信するパケットについては、仮想スイッチ22において、ポリシーテーブル240を参照することで、優先度値を設定することが可能である。

【0083】

また、一般に、I/O仮想化に対応したNIC18は、物理インタフェース180と、各仮想インタフェース181とを、固有のIDを用いて識別可能である。

【0084】

仮想スイッチ22から、NIC18T-1に対して、パケットを送信する際には、仮想スイッチ22が、NIC18T-1に対してパケット送信要求を発行する。

【0085】

そのため、仮想スイッチ22から物理インタフェース180を通じて発行された送信要求については、送信対象のパケットに、システム管理者のポリシーに基づいた優先度値の設定がなされているとし、TXコントローラ182は、図4に示したような処理を行わずに、パケット送信要求に含まれる転送サイズと、アドレスに基づいて、DMAコントローラ183がホストメモリから読み出したパケットを、そのままネットワークに送出することが可能である。

【0086】

一方で、仮想マシンモニタ24の負荷が上昇すると、優先度値の設定を行うオーバーヘッドにより、仮想スイッチ22に接続された仮想マシン20に係る通信の性能が劣化する可能性がある。

【0087】

このような場合に、優先度値の付加を、仮想スイッチ22で行わず、NIC18T-1で行うように、NIC18T-1の動作を切り替えることが考えられる。

【0088】

例えば、NIC18T-1に、物理インタフェース180から送信されたパケットに対して、優先度値の付加を行うかどうかを切り替えることができる設定レジスタを設ける。この設定レジスタに対する設定は、物理インタフェース180を介して、仮想マシンモニタ24から行うことができる。

【0089】

CPU使用率などの情報から、仮想マシンモニタ24の負荷が上昇したと仮想マシンモニタ24が判断した場合に、仮想マシンモニタ24から、この設定レジスタの設定を行う

10

20

30

40

50

ことで、NIC18T-1側でパケットに対する優先度値の付加を行うようにすることが可能となる。この場合、仮想スイッチ22からのパケット送信要求に対して、NIC18T-1が図4で示したような処理を実行することで、優先度が付与されたパケットがネットワークに送出されることになる。

【0090】

また、あるいは、仮想マシンモニタ24と、NIC18-T1の間のインタフェースを複数設ける、あるいは物理インタフェース180に、パケット送信要求を書き込むレジスタを複数設けることで、優先度値の付加を仮想スイッチ22で行うか、NIC18T-1で行うかを使い分けることができる。

【0091】

10

例えば、仮想マシンモニタ24とNIC18-T1の間のインタフェースとして、物理インタフェース180に加えて、仮想インタフェース181を割り当て、優先度値の付加を仮想スイッチ22で行う場合は、パケット送信要求を物理インタフェース180に対して、優先度値の付加をNIC18-T1で行う場合は、パケット送信要求を仮想インタフェース181に対して発行する。あるいは、物理インタフェース180に、パケット送信要求を書き込むレジスタが複数設けられている場合も同様に、パケット送信要求を発行するレジスタを、優先度値の付加を仮想スイッチ22で行うかどうかに応じて使い分ける。

【0092】

このようにすることで、仮想マシンモニタ24の負荷が上昇した際に、その上昇の原因となったトラフィックに対してのみ、このトラフィックに属するパケットに対する優先度値の付加をNIC18-T1に行わせるように制御することも可能となる。この場合、単位時間当たりに仮想スイッチ22を通過したパケット数や、データサイズをモニタすることで、負荷が上昇した原因のトラフィックを特定することが可能である。

20

【0093】

以上説明したように、本実施形態では、NIC18T-1は、フローテーブル185に登録されたQoSパラメータに従って、I/O仮想化技術を用いて送信されるパケットに対して優先度値をセットする。このQoSパラメータは、システム管理者のポリシーに従って、仮想マシンモニタ24からの指示によって、フローテーブル185にセット、あるいは送信パケットに適用されるものである。このため、I/O仮想化技術を用いて高い性能を実現でき、さらに、仮想マシン20のユーザが設定した優先度値がネットワークでの優先度制御に用いられることを防ぐことができる。

30

[第2の実施形態]

次に、本発明の第2の実施形態について図面を参照して詳細に説明する。図6は、本発明の第2の実施の形態に係るQoS制御システム10bの構成を示した図である。図6を参照すると、本発明の第2の実施形態は、図1に示した本発明の第1の実施形態の構成に、優先度キュー186と、スケジューラ187とが追加され、フローテーブル185-2と、ポリシーテーブル240-2とに変更が加えられている。その他本発明の第1の実施形態と同様の構成要素については同一の符号を付し、詳細な説明を省略する。

【0094】

優先度キュー186は、互いに異なる優先度を持った複数のキューから構成される。各キューは1段以上の深さを持ち、キューの各段は、TXコントローラ182が優先度値をセットしたパケットを格納できる領域を持つ。優先度キュー186に含まれるキューの個数は、一般的なデータセンタ向けネットワークスイッチがサポートする、ポートごとのQoS用の優先度キューの個数と同等が望ましいが、NIC18T-2の実装コストなどの要因により、それよりも少なくなることも考えられる。少なくとも、優先度キュー186を構成するキューの個数や、各キューのNIC18T-2内での優先度は、物理インタフェース180を通じて、仮想マシンモニタ24が知ることができる。

40

【0095】

フローテーブル185-2は、送信パケットに対して設定するパラメータを保持するテーブルである。各エントリには、QoSのパラメータとして、第1の実施形態のフローテ

50

ーブル185が保持するパラメータに加えて、優先度キュー186のどのキューに格納するかを示す、キューの識別情報が含まれる。

【0096】

スケジューラ187は、優先度キュー186に格納されたパケットを、パケットが格納されているキューの優先度に従って、ネットワークに送出する。パケットを送出するスケジューリング方式として、優先度の高いキューが空になるまで、優先度の低いキューのパケットを送出しない方式や、キューの優先度に基づいた重み付けをし、重み付けの比率に基づいてパケットを各キューから送出する方式、などが考えられる。どのようなスケジューリング方式を用いているか、複数のスケジューリング方式をサポートする場合、どれを使用するかは、物理インタフェース180を通じて、仮想マシンモニタが検知・指定することができる。

10

【0097】

ポリシーテーブル240-2には、ネットワーク管理者が定めた、トラフィックに対するQoSポリシーが格納される。QoSポリシーを構成する情報として、第1の実施形態のポリシーテーブル240に登録する情報に加えて、優先度キュー186のどのキューに格納するかを示す、キューの識別情報が含まれる。

【0098】

次に、図7のフローチャートを参照して、本実施形態の動作について詳細に説明する。

【0099】

図7は、NIC18T-2が、優先度キュー186に格納されたパケットを、キューの優先度に基づいてパケットを送出する処理を示したフローチャートである。本実施形態において、NIC18T-2は、図4のステップF188までの処理を実行した上で、図7に示す処理を行う。図7のステップF188までの処理は、本実施の第1の実施形態と同様（図4参照）であるため、詳細な説明を省略する。

20

【0100】

まず、TXコントローラ182は、優先度値をセットしたパケットを、その優先度値に基づいて、優先度キュー186の対応するキューに格納する（図7のステップF1880）。パケットにセットされる優先度値が表現することのできる優先度の高低が、優先度キュー186を構成するキューの個数で表現することのできる優先度の高低よりも広い場合、複数の優先度値をまとめて、1つのキューに対応させることが考えられる。例えば、パケットヘッダの優先度値を示すフィールドの上位数ビットの値を元に、各キューに割り振ることができる。

30

【0101】

次に、スケジューラ187は、キューの優先度に基づいて、キューに格納されたパケットを、TXインタフェース188に転送する。TXインタフェース188は、ネットワークに前記転送されたパケットを送出する（図7のステップF1882）。

【0102】

このようにして、NIC18T-2は、パケットの優先度に基づいて、パケットをネットワークに送出する。

【0103】

40

本実施形態では、パケットのエラー検出のためのデータを計算し、パケットに適用する処理は種々のタイミングで実行することができる。例えば、TXコントローラ182が図4のステップF188の時点で、パケットのエラー検出のためのデータの計算と適用を行うこととしてもよいし、スケジューラ188が、図7のステップF1882の時点で、前記パケットのエラー検出のためのデータの計算と適用を行うこととしてもよい。

【0104】

ここで、優先度キュー186に、パケット全体を保持するだけの容量が備えられていない場合について図面を用いて説明する。本実施形態では、パケットがいつ送信されるかは、パケットが格納された優先度キュー186を構成するキューの優先度依存する。キューの各段にパケット全体を保持するだけの容量が備えられていない場合、図4のステップ

50

F 1 8 2でパケット全体を読み出したとしても、パケットのペイロード部はN I C 1 8 T-2で破棄されることになる。

【0 1 0 5】

図8は、優先度キュー1 8 6に、パケット全体を保持するだけの容量が備えられていない場合のQ o S制御システム1 0 b-2の構成を示したものである。

【0 1 0 6】

図8を参照すると、Q o S制御システム1 0 b-2では、N I C 1 8 T-3が、Q o S制御システム1 0 bと異なる。

【0 1 0 7】

N I C 1 8 T-3では、図6に示した構成と比較して、スケジューラ1 8 7-2が、DMAコントローラ1 8 3に対して、ホストメモリからデータを読み出す要求を出すことができるようになっている。

10

【0 1 0 8】

このような構成において、N I C 1 8 T-3が、パケットに優先度値を付与し、優先度に基づいてパケットを送信する処理について、図4、図9を用いて、その差異を説明する。

【0 1 0 9】

図9は、N I C 1 8 T-3が、パケットに優先度値を付与し、優先度に基づいてパケットを送信する処理の一部を示したフローチャートである。

【0 1 1 0】

20

まず、図4のステップF 1 8 2において、T Xコントローラ1 8 2は、DMAコントローラ1 8 3を通じて、パケットヘッダのみを読み出す。

【0 1 1 1】

次に、T Xコントローラ1 8 2は、図4のステップF 1 8 4、または図4のステップF 1 8 5～ステップF 1 8 7の手順に従ってパケットに適用する優先度値を取得する。

【0 1 1 2】

次に、T Xコントローラ1 8 2は、図4のステップF 1 8 8で、優先度値をパケットヘッダに適用する。

【0 1 1 3】

次に、T Xコントローラ1 8 2は、パケットヘッダに適用した優先度値に基づいて、パケットヘッダと、パケットの転送要求を優先度キュー1 8 6に格納する（図9のステップF 1 8 8 0-2）。

30

【0 1 1 4】

次に、スケジューラ1 8 7-2は、キューの優先度に基づいて、キューに格納されたパケット転送要求とパケットヘッダを読み出し（図9のステップF 1 8 8 2-2）、パケット転送要求に基づいて、DMAコントローラ1 8 3に対し、パケットのペイロード部の読み出しを要求する（図9のステップF 1 8 8 4-2）。

【0 1 1 5】

次に、スケジューラ1 8 7-2は、読み出されたパケットのペイロード部と、図9のステップF 1 8 8 2-2で読み出したパケットヘッダからパケットを生成し、T Xインタフェース1 8 8を通じて、ネットワークにパケットを送出する（図9のステップF 1 8 8 6-2）。

40

【0 1 1 6】

このようにして、優先度キュー1 8 6の各キューのサイズが小さい構成においても、パケットに優先度値を付与し、優先度に基づいてパケットを送信することが可能となる。

【0 1 1 7】

ここで、図9のステップF 1 8 8 0-2で格納されたパケット転送要求について、既に読み出したパケットヘッダの領域の分だけ、転送サイズと、パケットが格納されたホストメモリのアドレスを変更しておくことも考えられる。また、このような変更をせずに、スケジューラ1 8 7-2が、図9のステップF 1 8 8 4-2でDMAコントローラ1 8 3に

50

対し、パケットのペイロード部の読み出しを要求する際に、転送サイズと、読み出し元のアドレスを変更することも考えられる。さらに、転送要求には変更を加えずに、図9のステップF1884-2でパケット全体を読み出し、パケットヘッダ部を破棄するというとも考えられる。

【0118】

また、図8の構成の場合、パケットのエラー検出のためのデータを計算し、パケットに適用する処理は、スケジューラ187-2が、図9のステップF1886-2の時点で行うことになる。

【0119】

ここで、仮想マシンモニタ24に内蔵される仮想スイッチ22に接続された仮想マシン20が、QoS制御システム10bに配置されている場合の、NIC18T-2の処理について説明する。

【0120】

仮想スイッチ22に接続された仮想マシン20が送信するパケットについては、仮想スイッチ22において、ポリシーテーブル240-2を参照することで、優先度値を設定することが可能である。

【0121】

そのため、仮想スイッチから物理インタフェース180を通じて発行された送信要求については、送信対象のパケットに、システム管理者のポリシーに基づいた優先度値の設定がなされているとし、TXコントローラ182は、図4、図7で示したような処理を行わずに、パケット送信要求に含まれる転送サイズと、アドレスに基づいて、DMAコントローラ183がホストメモリから読み出したパケットに設定された優先度値に基づいて、優先度キュー186の適切なキューにパケットを格納する。

【0122】

一方で、仮想マシンモニタ24の負荷が上昇すると、優先度値の設定を行うオーバーヘッドにより、仮想スイッチ22に接続された仮想マシン20に係る通信の性能が劣化する可能性がある。

【0123】

このような場合に、優先度値の付加を、仮想スイッチ22で行わず、NIC18T-2で行うように、NIC18T-2の動作を切り替えることが考えられる。この方法については、本発明の第1の実施形態に記載したものと同様であるため、詳細な説明を省略する。

【0124】

以上のように、本実施形態では、NIC18T-2、NIC18T-3が備える優先度キュー186に、フローテーブル185に登録されたQoSパラメータに従って付与された優先度値に基づいて、パケットが格納される。優先度キュー186に格納されたパケットは、キューの優先度に基づいて送信時刻がスケジューリングされる。これによって、優先度に基づいて、サーバ10とTORスイッチ12の間のリンクを使用することが可能となる。

[第3の実施形態]

次に、本発明の第3の実施形態について図面を参照して詳細に説明する。図10は、本発明の第3の実施形態に係るQoS制御システム10cの構成を示した図である。図10を参照すると、第2の実施形態の構成の仮想マシン20にスレーブポリシーテーブル242が追加された構成が示されている。その他本発明の第2の実施の形態と同様の構成要素については同一の符号を付し、詳細な説明を省略する。

【0125】

スレーブポリシーテーブル242には、ポリシーテーブル240-2のエントリの一部が登録される。具体的には、優先度キュー186を構成する各キューに、あふれそうなキューが存在する場合に、そのキューの優先度に対応するポリシーが登録される。例えば、優先度キュー186を構成するキューが0～7番の8つあり、この番号がそのままキュー

10

20

30

40

50

の優先度を示すとする。このときに、TCPパケットにおいて、宛先ポート番号として80番または、443番を持つトラフィックに対して、キュー番号が5番のキューを使用するようにポリシーが設定されているとする。このキューがあふれそうな場合に、宛先ポート番号が80番のトラフィックを示すポリシーと、宛先ポート番号443番のトラフィックを示すポリシーがスレーブポリシーテーブルに登録される。

【0126】

スレーブポリシーテーブル242に登録されているポリシーは、そのポリシーにマッチするトラフィックが利用する優先度キュー186のキューがあふれそうなことを示す。そのため、各仮想マシン20は、パケットを構築した際に、パケット送信要求をNIC18T-2に発行する前に、スレーブポリシーテーブル242に、送信要求で送信しようとするパケットのパケットヘッダに適合するトラフィック識別情報を持つエントリが存在するか否かをチェックする。ここでもし、エントリが存在した場合には、優先度キュー186の該当するキュー（5番キュー）があふれそうであるため、仮想マシン20は、パケットの送信要求の発行を延期するように動作することができる。

10

【0127】

続いて、図11のシーケンス図を参照して、本実施形態の動作について詳細に説明する。

【0128】

図11は、優先度キュー186を構成するキューのいずれかがあふれそうになった場合に、スレーブポリシーテーブルにエントリを追加する処理を示したシーケンス図である。

20

【0129】

まず、NIC18T-2は、優先度キュー186を構成するキューのいずれかがあふれそうになると、そのキューの識別情報を物理インタフェース180を通じて、仮想マシンモニタ24に通知する（図11のステップS182-3）。この通知には、あふれそうなキューの識別情報が含まれる。また、キューのいずれかがあふれそうになるか否かの判定は、各キューの空き領域（収容可能パケット数）が所定のしきい値TH1未満となったか否かにより判定できる。このしきい値TH1は、仮想マシンモニタ24が、物理インタフェース180から設定可能なように、NIC18-T2を構成することが考えられる。この場合、NIC18-T2は、設定されたしきい値TH1と、キューの空き領域を比較し、キューがあふれそうになるか否かの判定を行う。

30

【0130】

次に、仮想マシンモニタ24は、NIC18T-2からの通知に含まれる、あふれそうなキューの識別情報をもとに、ポリシーテーブル240-2を参照し、該当するポリシーを抽出する（図11のステップS242-3）。ポリシーテーブル240に登録されているポリシーの構成要素には、第2の実施形態にて述べたように、優先度キュー186のどのキューに格納するかを示す情報が含まれているため、これと比較することで抽出可能である。

【0131】

次に、仮想マシンモニタ24は、抽出したポリシーを、このポリシーが適用される可能性のある仮想マシン20に通知する（図11のステップS244-3）。

40

【0132】

ここで、「ポリシーが適用される可能性がある」ことについて、例を用いて説明する。例えば、「TCPパケットの宛先ポート番号が80番のトラフィック」のような、トラフィックの送信元、宛先を限定しないポリシーの場合は、全ての仮想マシンに通知される。一方で、「VLAN IDが1024番のトラフィック」のようなポリシーの場合、ポリシーが適用される仮想マシン20が、VLAN IDが1024の仮想マシン20に限定されるため、VLAN IDが1024のVLANに属する仮想マシン20にのみ通知される。また、「送信元MACアドレスが00:11:22:33:44:55で、VLAN IDが1024のトラフィック」のようなポリシーの場合、仮想マシン20が一意に定まるため、その仮想マシン20に対してのみ通知される。

50

【0133】

次に、ポリシーの通知を受けた仮想マシン20は、スレーブポリシーテーブル242に、通知されたポリシーを登録する（図11のステップS202-3）。

【0134】

なお、仮想マシン20に通知されるポリシーには、設定する優先度値の情報を含まないように通知することも可能である。

【0135】

次に、図12のシーケンス図を参照して、本実施形態の動作に係るスレーブポリシーテーブルからのエントリの削除の処理について詳細に説明する。

【0136】

図12は、優先度キュー186を構成するキューのうち、あふれそうになっていたキューのいずれかについて、空き領域に余裕ができた場合に、スレーブポリシーテーブル242からエントリを削除する処理を示したシーケンス図である。また、キューの空き領域に余裕ができたか否かの判定は、各キューの空き領域（収容可能パケット数）が所定のしきい値TH2以上となったか否かにより判定できる。このしきい値TH2は、前述のこのしきい値TH1と同一の値でもよいし、異なる値でもよい。このしきい値TH2の設定方法の例は、しきい値TH1の場合と同様であるため、詳細な説明を省略する。

【0137】

まず、NIC18T-2は、優先度キュー186を構成するキューのうち、あふれそうになっていたキューについて、そのキューの空きに余裕ができると、そのキューの識別情報を物理インタフェース180を通じて、仮想マシンモニタ24に通知する（図12のステップS184-3）。

【0138】

次に、仮想マシンモニタ24は、NIC18T-2からの通知に含まれる、空きに余裕ができたキューの識別情報をもとに、ポリシーテーブル240を参照し、該当するポリシーを抽出する（図12のステップS246-3）。ポリシーテーブル240に登録されているポリシーの構成要素には、第2の実施形態にて述べたように、優先度キュー186のどのキューに格納するかを示す情報が含まれているため、これと比較することで抽出可能である。

【0139】

次に、仮想マシンモニタ24は、抽出したポリシーを、このポリシーが適用される可能性のある仮想マシン20に通知する（図12のステップS248-3）。

【0140】

次に、ポリシーの通知を受けた仮想マシン20は、通知されたポリシーを、スレーブポリシーテーブル242から削除する。削除する際に、スレーブポリシーテーブル242を検索するために、通知されたポリシーに含まれるトラフィックを識別する情報を用いることができる。

【0141】

次に、図13のシーケンス図、図14のフローチャートを参照して、上記スレーブポリシーテーブル242を利用した仮想マシン20のパケット送信要求の抑制処理について詳細に説明する。

【0142】

図13は、スレーブポリシーテーブル242を利用して、仮想マシン20がパケット送信要求を発行する処理を示したシーケンス図である。

【0143】

まず、仮想マシン20は、送信するパケットを生成すると、パケット送信要求をNIC18T-2に発行する前に、生成したパケットのパケットヘッダの情報を用いて、スレーブポリシーテーブル242を参照する（図13のステップS206-3）。

【0144】

次に、スレーブポリシーテーブル242にエントリが存在しない場合、仮想マシン20

10

20

30

40

50

は、パケット送信要求を、仮想インタフェース181を通じて、NIC18T-2に発行する（図13のステップS208-3）。

【0145】

以降、NIC18T-2と仮想マシンモニタ24が、図4のステップF180～ステップF188、図5のステップF238～ステップF242、図7の処理を実行することで、パケットがネットワークに送出される。

【0146】

図14は、スレーブポリシーテーブル242を利用して、仮想マシン20がパケット送信要求を発行する処理を示したフローチャートである。

【0147】

まず、仮想マシン20は、送信するパケットを生成すると、パケット送信要求をNIC18T-2に発行する前に、生成したパケットのパケットヘッダの情報を用いて、スレーブポリシーテーブル242を参照する（図14のステップF206-3）。

【0148】

次に、仮想マシン20は、スレーブポリシーテーブル242にエントリが存在しない場合（図14のステップF206-3の「エントリが存在しない」）、パケット送信要求を、仮想インタフェース181を通して、NIC18T-2に発行する（図14のステップF208-3）。

【0149】

一方、仮想マシン20は、スレーブポリシーテーブル242にエントリが存在した場合（図14のステップF206-3の「エントリが存在する」）、パケット送信要求の発行を、該当するエントリを削除する通知が仮想マシンモニタ24からなされるまで遅延させる（図14のステップF210-3）。

【0150】

なお、スレーブポリシーテーブル242にエントリが存在するにもかかわらず、このエントリに該当するパケットのパケット送信要求を、仮想マシン20が発行した場合、NIC18T-2にて送信要求を破棄することで、キューがあふれることを防ぐことができる。

【0151】

以上説明したように、本実施形態では、優先度キュー186を構成するキューのいずれかがあふれそうになると、NIC18T-2から仮想マシンモニタ24に対して、通知がなされ、仮想マシンモニタ24を通じて、影響のある仮想マシン20に通知される。これにより、仮想マシン20が、キューのあふれそうな通信要求の発行を行うことを抑制することができる。

【0152】

なお、本発明の第1から第3実施形態では、ポリシーテーブル240、あるいはポリシーテーブル240-2が、仮想マシンモニタ24に備えられ、ネットワークインタフェースカード18T-1、ネットワークインタフェースカード18T-2、ネットワークインタフェースカード18T-3が、フローテーブル185、フローテーブル185-2に登録されていないQoSパラメータについて問い合わせる際に、仮想マシンモニタ24に問い合わせるような構成をとっているが、サーバ10の外部に備えられた、システム管理者が管理するコントローラにポリシーテーブル240を備え、このコントローラに問い合わせるという構成をとることも可能である。外部のコントローラに問い合わせる方法として、TXインタフェース188を通じて問い合わせる方法や、サーバ10が他の管理用外部インタフェースを備える場合に、この管理用外部インタフェースを用いることが考えられる。この場合、仮想マシンモニタ24に対して、QoSパラメータの問い合わせの要求を、外部コントローラに発行してもらうように依頼するという構成が考えられる。

【0153】

また、本発明の第1から第3実施形態では、サーバ仮想化環境を前提とした構成をとっているが、本発明の適用先はサーバ仮想化環境に限定されない。例えば、サーバ仮想化環

10

20

30

40

50

境を用いずに、QoS制御システム10a、10b、10b-2、10cを構成するサーバ上にベアメタルでOSを導入した場合でも、システム管理者が運用する外部コントローラから、NIC18を制御し、本発明のQoS制御を適用することも可能である。

【0154】

最後に、本発明の好ましい形態を要約する。

[第1の形態]

(上記第1の視点によるQoS制御システム参照)

[第2の形態]

第1の形態のQoS制御システムにおいて、

前記ネットワークインタフェースは、外部からのパケット送信要求に応じてメモリからパケットを取得し、前記パケットに適合する識別情報を持つQoSパラメータを用いて前記パケットに優先度値を格納してから送信することができる。

10

[第3の形態]

第1又は第2の形態のQoS制御システムにおいて、

前記ネットワークインタフェースは、送信パケットに適合する識別情報を持つQoSパラメータを保持していない場合、前記ネットワークインタフェース管理部に対してQoSパラメータの送信を要求することができる。

[第4の形態]

第3の形態のQoS制御システムにおいて、

前記ネットワークインタフェース管理部は、前記QoSポリシーから、前記ネットワークインタフェースからの前記QoSパラメータの送信要求を受けたパケットと同じ識別情報を持つQoSパラメータを取得し、前記ネットワークインタフェースに通知することができる。

20

[第5の形態]

第1から第4いずれか一の形態のQoS制御システムにおいて、

前記ネットワークインタフェースは、さらに、

優先度が設定された1つ以上のキューから構成される優先度キューと、前記優先度キューを構成するキューの優先度に応じて、パケットの送信をスケジューリングするスケジューラと、を備える構成とすることができる。

[第6の形態]

30

第5の形態のQoS制御システムにおいて、

前記ネットワークインタフェースは、前記QoSパラメータとして、トラフィック毎に前記優先度キューを指定するための情報を保持することができる。

[第7の形態]

第6の形態のQoS制御システムにおいて、

前記ネットワークインタフェースは、前記優先度キューを指定するための情報に基づいて、前記QoSパラメータを適用したパケット、パケットヘッダ、またはパケット送信要求のいずれかを前記優先度キューのいずれかに格納することができる。

[第8の形態]

第5から第7いずれか一の形態のQoS制御システムにおいて、

40

前記ネットワークインタフェースは、前記ネットワークインタフェース管理部に対して、前記優先度キューの使用状況を通知する機能を有することができる。

[第9の形態]

第8の形態のQoS制御システムにおいて、

前記ネットワークインタフェース管理部は、前記QoSポリシーとして、さらに、前記優先度キューを構成するキューの識別情報を保持し、

前記ネットワークインタフェースから通知された前記優先度キューの使用状況に基づいて、前記キューの識別情報と一致するキューの識別情報を持つQoSポリシーを選択し、パケット送信要求の発行元に対して、前記選択したQoSポリシーを通知し、

前記パケット送信要求の発行元は、前記QoSポリシーを含む通知に従って、パケット

50

送信要求の抑制又はパケット送信要求の送信再開を行うことができる。

[第10の形態]

第1から第9いずれか一の形態のQoS制御システムにおいて、

前記ネットワークインタフェースは、前記ネットワークインタフェースにて優先度値の付与を行うか否かを設定可能であり、

前記ネットワークインタフェース管理部からの前記優先度値の付与の要否の設定を変更することにより、前記ネットワークインタフェースで優先度値の付与を行うか否かを制御することができる。

[第11の形態]

(上記第2の視点によるQoS制御方法参照)

10

[第12の形態]

(上記第3の視点によるプログラム参照)

なお、第11の形態、第12の形態は、第1の形態と同様に、第2～第10の形態に展開することが可能である。

【0155】

なお、上記の特許文献および非特許文献の各開示を、本書に引用をもって繰り込むものとする。本発明の全開示（請求の範囲を含む）の枠内において、さらにその基本的技術思想に基づいて、実施形態ないし実施例の変更・調整が可能である。また、本発明の請求の範囲の枠内において種々の開示要素（各請求項の各要素、各実施形態ないし実施例の各要素、各図面の各要素等を含む）の多様な組み合わせ、ないし選択が可能である。すなわち、本発明は、請求の範囲を含む全開示、技術的思想にしたがって当業者であればなし得るであろう各種変形、修正を含むことは勿論である。

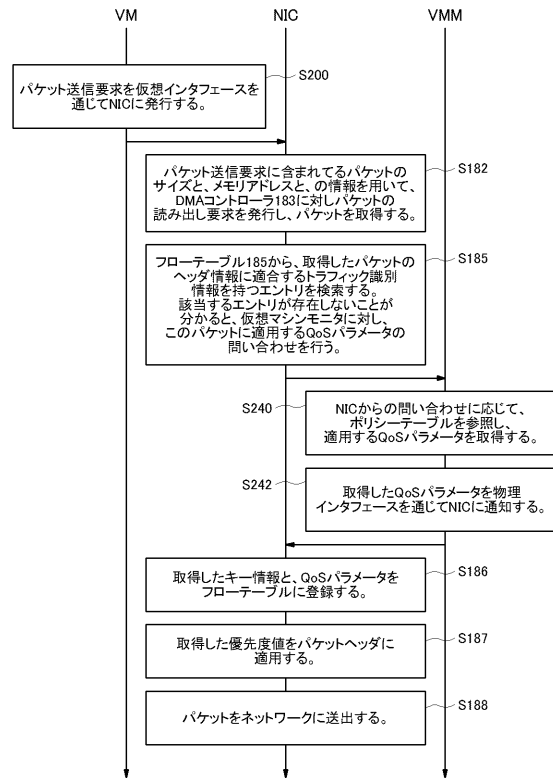
20

【符号の説明】

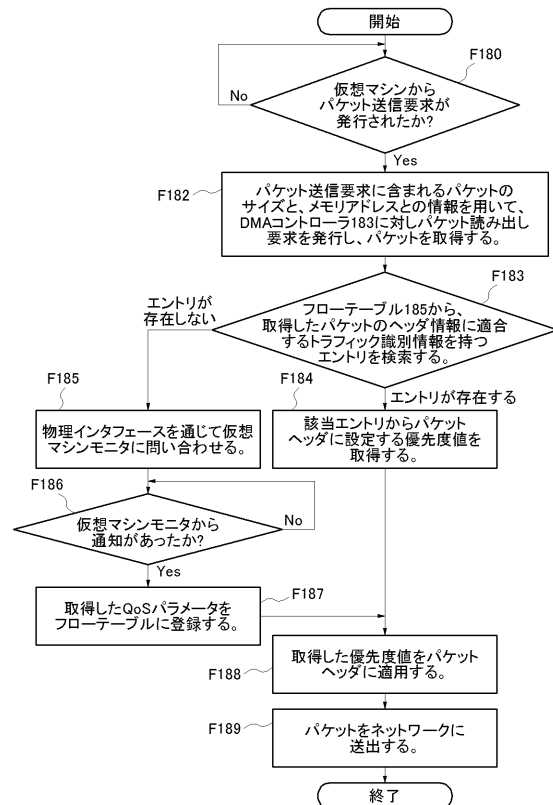
【0156】

1	データセンタネットワークアーキテクチャ	
2	データセンタネットワークアーキテクチャ	
10	サーバ	
10a	QoS制御システム	
10b	QoS制御システム	
10b-2	QoS制御システム	30
10c	QoS制御システム	
12	Top-of-Rackスイッチ	
14	アグリゲーションスイッチ	
16	コアスイッチ	
18	ネットワークインタフェースカード	
18T-1～18T-3	ネットワークインタフェースカード送信部	
20	仮想マシン	
22	仮想スイッチ	
24	仮想マシンモニタ	
180	物理インタフェース	40
181	仮想インタフェース	
182	TXコントローラ	
183、183-2	DMAコントローラ	
184	ホストバスインタフェース	
185、185-2	フローテーブル	
186	優先度キュー	
187、187-2	スケジューラ	
188	TXインタフェース	
240	ポリシーテーブル	
242、240-2	スレーブポリシーテーブル	50

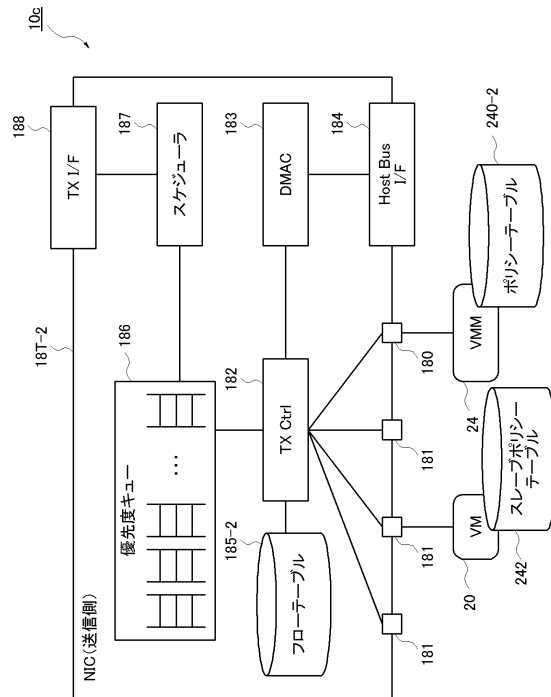
【図 2】



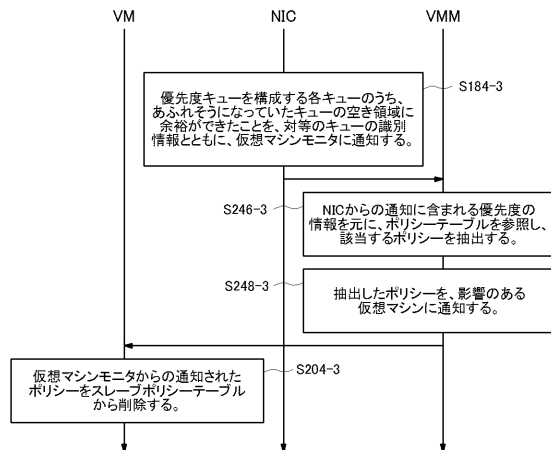
【図 4】



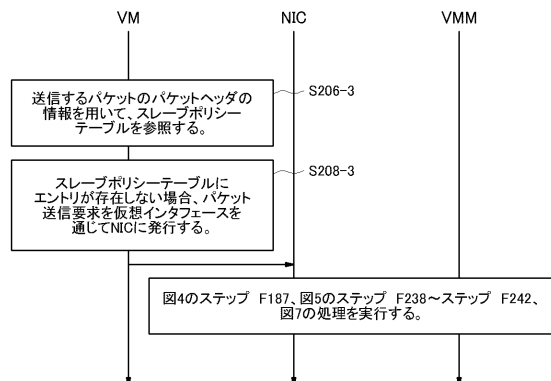
【図 10】



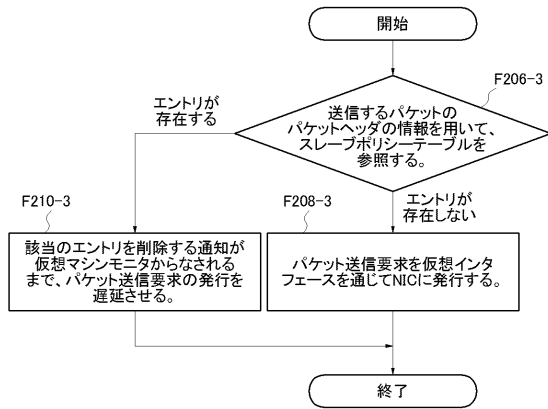
【图 1 2】



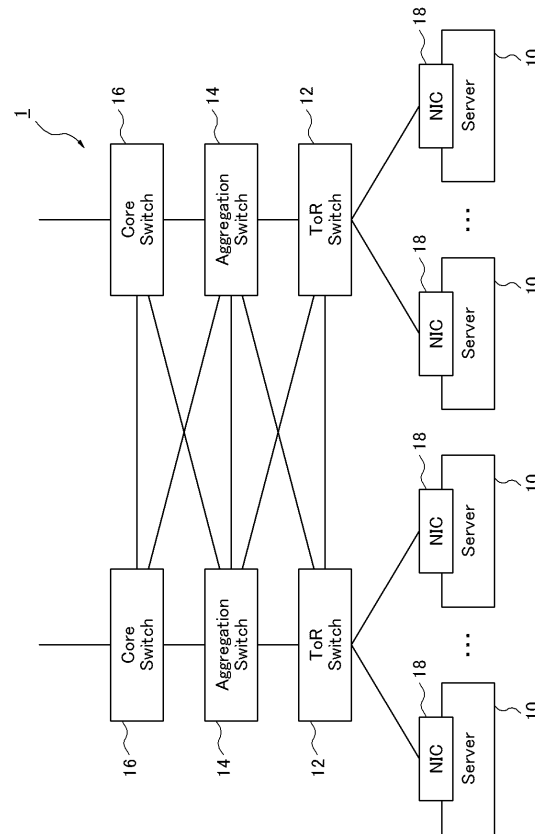
【图 13】



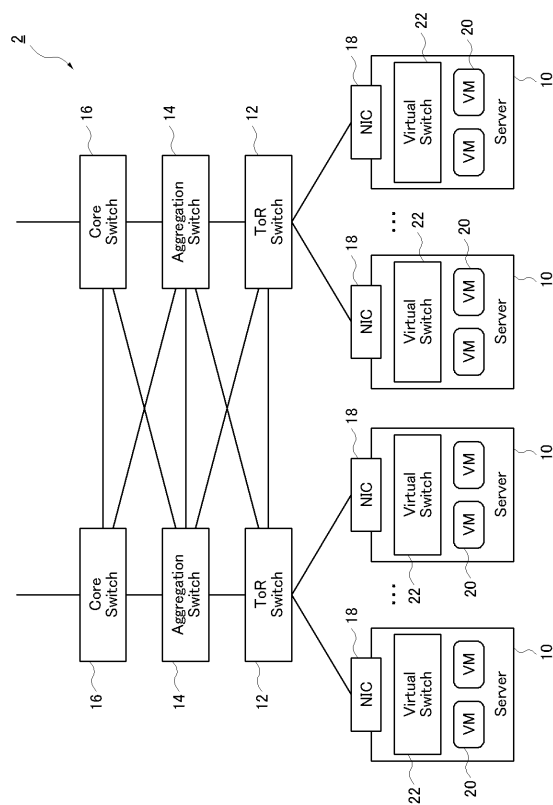
【図 14】



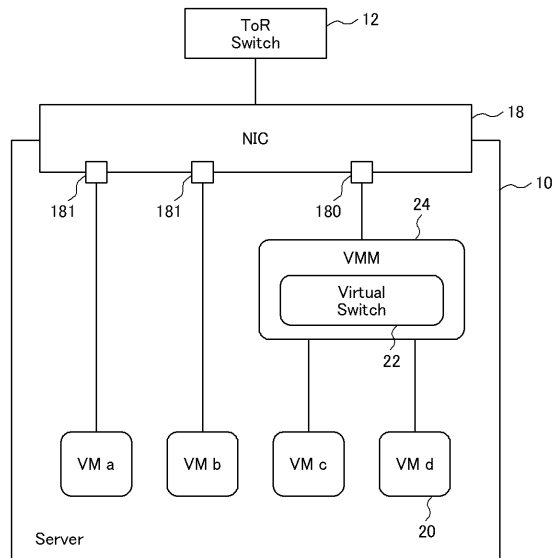
【図 15】



【図 16】



【図 17】



フロントページの続き

(出願人による申告)平成23年度、独立行政法人情報通信研究機構「高度通信・放送研究開発委託研究／新世代ネットワークを支えるネットワーク仮想化基盤技術の研究開発」、産業技術力強化法第19条の適用を受ける特許出願

(56)参考文献 特表2007-531353(JP, A)
特開2003-092595(JP, A)
特開2003-069631(JP, A)

(58)調査した分野(Int.Cl., DB名)
H04L 12/00~12/955