

(19) 日本国特許庁 (JP)

(12) 特 許 公 報 (B2)

(11) 特許番号

特許第4807861号
(P4807861)

(45) 発行日 平成23年11月2日 (2011. 11. 2)

(24) 登録日 平成23年8月26日 (2011. 8. 26)

(51) Int. Cl.	F I
H04L 12/28 (2006.01)	H04L 12/28 200Z
H04L 29/10 (2006.01)	H04L 13/00 309C
G06F 13/00 (2006.01)	G06F 13/00 353N
G06F 9/54 (2006.01)	G06F 9/46 480Z

請求項の数 19 (全 22 頁)

(21) 出願番号	特願2008-503471 (P2008-503471)	(73) 特許権者	390009531
(86) (22) 出願日	平成18年3月15日 (2006. 3. 15)		インターナショナル・ビジネス・マシーンズ・コーポレーション
(65) 公表番号	特表2008-535343 (P2008-535343A)		INTERNATIONAL BUSINESS MACHINES CORPORATION
(43) 公表日	平成20年8月28日 (2008. 8. 28)		アメリカ合衆国10504 ニューヨーク州 アーモンク ニュー オーチャードロード
(86) 国際出願番号	PCT/EP2006/060734		
(87) 国際公開番号	W02006/103169		
(87) 国際公開日	平成18年10月5日 (2006. 10. 5)	(74) 代理人	100108501
審査請求日	平成20年12月17日 (2008. 12. 17)		弁理士 上野 剛史
(31) 優先権主張番号	11/097, 608	(74) 代理人	100112690
(32) 優先日	平成17年4月1日 (2005. 4. 1)		弁理士 太佐 種一
(33) 優先権主張国	米国 (US)	(74) 代理人	100091568
			弁理士 市位 嘉宏

最終頁に続く

(54) 【発明の名称】 サーバ環境においてオフロードをネットワーク化するためのホスト・イーサネット・アダプタ

(57) 【特許請求の範囲】

【請求項 1】

イーサネット・アダプタであって、

少なくとも1つのプロセッサからパケットを受信し、且つ少なくとも1つのプロセッサへパケットを送信することを可能にする複数の層を含んでおり、前記複数の層は、前記少なくとも1つのプロセッサのパーティション化を考慮に入れた多重分離メカニズムを含み、1つの論理スイッチが前記イーサネット・アダプタの各物理ポートのために利用され、各論理ポートが論理スイッチ上に個々のポートを有しており、1つの論理スイッチが複数の論理ポートを提供し、前記複数の論理ポートの各々は前記プロセッサのパーティションをサポートし、そして、前記パーティション間の通信が可能にされる、前記イーサネット・アダプタ。

【請求項 2】

前記複数の層は、

媒体アクセス・コントローラ (MAC) 及び並直列変換回路 (Serdes) 層 (以下、第1層) であって、ネットワーク上の他の装置からの及びネットワーク上の他の装置への複数のインターフェースを提供し、及び、当該提供のために同じチップの I/O が使用される、前記第1層と、パケット高速化及び仮想化層 (以下、第2層) であって、前記第1層からパケットを受信し、且つ前記第1層にパケットを提供する、前記第2層と、

10

20

ホスト・インターフェース層（以下、第3層）であって、前記第2層と通信し、且つ前記プロセッサのプライベート・バスとインターフェースする、前記第3層とを含む、請求項1に記載のイーサネット・アダプタ。

【請求項3】

前記第2層が、適切なときにはパケット・ヘッダをデータ・ペイロードから分離することによってパケット・ヘッダ分離を提供し、且つ、受信バッファを汚さずに処理されるように当該パケット・ヘッダをプロトコル・スタックに向け、それによって或るトランザクションのために待ち期間を短縮する、請求項2に記載のイーサネット・アダプタ。

【請求項4】

前記第3層がコンテキスト管理を提供する、請求項2に記載のイーサネット・アダプタ。

10

【請求項5】

前記第1層は、パケットを受信するための受信イーサネット機能と、パケットを送信するための送信イーサネット機能とを含む、請求項2に記載のイーサネット・アダプタ。

【請求項6】

前記複数の層は、
媒体アクセス・コントローラ及び並直列変換回路（MAC及びSerdes）層と、
前記MAC及びSerdes層からパケットを受信し、且つ前記MAC及びSerdes層にパケットを提供するパケット高速化及び仮想化層と、
前記パケット高速化及び仮想化層と通信し、前記プロセッサのプライベート・バスとインターフェースするホスト・インターフェース層と
を含む、請求項1に記載のイーサネット・アダプタ。

20

【請求項7】

前記MAC及びSerdes層は、パケットを受信するための受信イーサネット機能と、パケットを送信するための送信イーサネット機能とを含む、請求項6に記載のイーサネット・アダプタ。

【請求項8】

前記多重分離メカニズムは、前記プロセッサ内で利用される複数のオペレーティング・システムへの直接経路を考慮に入れている、前記プロセッサ内で利用される複数のアプリケーションへの直接経路を考慮に入れている、又は前記プロセッサ内で利用される複数のオペレーティング・システムへの直接経路と前記プロセッサ内で利用される複数のアプリケーションへの直接経路とを考慮に入れている、請求項1～7のいずれか一項に記載のイーサネット・アダプタ。

30

【請求項9】

或るトランザクションのための待ち期間を短縮するために、適切なときにはパケット・ヘッダがデータ・ペイロードから分離される、請求項6又は7に記載のイーサネット・アダプタ。

【請求項10】

ネットワーク・インターフェース・コントローラ（NIC）であって、
プロセッサのプライベート・バスに接続されるように成されているインターフェースと、
イーサネット・アダプタと
を含んでおり、
前記イーサネット・アダプタは、少なくとも1つのプロセッサからパケットを受信し、且つ少なくとも1つのプロセッサへパケットを送信することを可能にする複数の層を含んでおり、
前記複数の層は、前記少なくとも1つのプロセッサのパーティション化を考慮に入れた多重分離メカニズムを含み、
1つの論理スイッチが前記イーサネット・アダプタの各物理ポートのために利用され、各論理ポートが論理スイッチ上に個々のポートを有しており、1つの論理スイッチが複数

40

50

の論理ポートを提供し、前記複数の論理ポートの各々は前記プロセッサのパーティションをサポートし、そして、前記パーティション間の通信が可能にされる、前記NIC。

【請求項11】

前記複数の層は、

媒体アクセス・コントローラ（MAC）及び並直列変換回路（Serdes）層（以下、第1層）であって、ネットワーク上の他の装置からの及びネットワーク上の他の装置への複数のインターフェースを提供し、及び、当該提供のために同じチップのI/Oが使用される、前記第1層と、

パケット高速化及び仮想化層（以下、第2層）であって、前記第1層からパケットを受信し、且つ前記第1層にパケットを提供する、前記第2層と、

ホスト・インターフェース層（以下、第3層）であって、前記第2層と通信し、且つ前記プロセッサのプライベート・バスとインターフェースする、前記第3層と

を含む、請求項10に記載のNIC。

【請求項12】

前記第2層が、適切なときにはパケット・ヘッダをデータ・ペイロードから分離することによってパケット・ヘッダ分離を提供し、且つ、受信バッファを汚さずに処理されるように当該パケット・ヘッダをプロトコル・スタックに向け、それによって或るトランザクションのために待ち期間を短縮する、請求項10に記載のNIC。

【請求項13】

前記第3層がコンテキスト管理を提供する、請求項10に記載のNIC。

【請求項14】

前記第1層は、パケットを受信するための受信イーサネット機能と、パケットを送信するための送信イーサネット機能とを含む、請求項10に記載のNIC。

【請求項15】

前記複数の層は、

媒体アクセス・コントローラ及び並直列変換回路（MAC及びSerdes）層と、

前記MAC及びSerdes層からパケットを受信し、且つ前記MAC及びSerdes層にパケットを提供するパケット高速化及び仮想化層と、

前記パケット高速化及び仮想化層と通信し、前記プロセッサのプライベート・バスとインターフェースするホスト・インターフェース層と

を含む、請求項10に記載のNIC。

【請求項16】

前記MAC及びSerdes層は、パケットを受信するための受信イーサネット機能と、パケットを送信するための送信イーサネット機能とを含む、請求項15に記載のNIC。

【請求項17】

前記多重分離メカニズムは、前記プロセッサ内で同時に利用される複数のオペレーティング・システムへの直接経路を考慮に入れている、前記プロセッサ内で利用される複数のアプリケーションへの直接経路を考慮に入れている、前記プロセッサ内で利用される複数のオペレーティング・システムへの直接経路と前記プロセッサ内で利用される複数のアプリケーションへの直接経路とを考慮に入れている、請求項10～16のいずれか一項に記載のNIC。

【請求項18】

或るトランザクションのための待ち期間を短縮するために、適切なときにはパケット・ヘッダがデータ・ペイロードから分離される、請求項15又は16に記載のNIC。

【請求項19】

請求項10乃至18のいずれか1項に記載のNICを含むサーバ・システム。

【発明の詳細な説明】

【技術分野】

【0001】

10

20

30

40

50

本発明は、一般的にはサーバ環境に関し、特にその様な環境において利用されるアダプタに関する。

【0002】

関連出願との相互参照

本出願は下記の同時係属中の米国特許出願と関連している。

本願と同じ日に出願されて本発明の譲受人に譲渡された、“数個のイーサネット・ポートと、簡易化されたフレーム・バイ・フレーム上部構造により処理されるラップ伝送されるフローに配慮するための方法及びシステム (Method and System for Accommodating Several Ethernet Ports and a Wrap Transmitted Flow Handled by a Simplified Frame-By-Frame Upper Structure)”と題された米国特許出願第 11/096363 号 (代理人事件整理番号 RPS920050060US1/3486P)。

10

本願と同じ日に出願されて本発明の譲受人に譲渡された、“ネットワーク接続テーブルを提供するための方法及び装置 (Method and Apparatus for Providing a Network Connection Table)”と題された米国特許出願第 11/096571 号 (代理人事件整理番号 RPS920050061US1/3487P)。

本願と同じ日に出願されて本発明の譲受人に譲渡された、“オペレーティング・システム・パーティションのためのネットワーク通信 (Network Communications for Operating System Partitions)”と題された米国特許出願第 11/097051 号 (代理人事件整理番号 RPS920050062US1/3488P)。

20

本願と同じ日に出願されて本発明の譲受人に譲渡された、“ホスト・イーサネット・アダプタのための設定可能なポート (Configurable Ports for a Host Ethernet Adapter)”と題された米国特許出願第 11/097652 号 (代理人事件整理番号 RPS920050073US1/3502P)。

本願と同じ日に出願されて本発明の譲受人に譲渡された、“ホスト・イーサネット・アダプタ (HEA) においてチェックサムを構文解析し、フィルタリングし、計算するためのシステム及び方法 (System and Method for Parsing, Filtering, and Computing the Checksum in a Host Ethernet Adapter (HEA))”と題された米国特許出願第 11/096365 号 (代理人事件整理番号 RPS920050074US1/3503P)。

30

本願と同じ日に出願されて本発明の譲受人に譲渡された、“ホスト・イーサネット・アダプタ (HEA) において待ち時間を短縮する方法のためのシステム及び方法 (System and Method for a Method for Reducing Latency in a Host Ethernet Adapter (HEA))”と題された米国特許出願第 11/096353 号 (代理人事件整理番号 RPS920050075US1/3504P)。

40

本願と同じ日に出願されて本発明の譲受人に譲渡された、“ネットワーク伝送のためのブラインド・チェックサム及び訂正のための方法及び装置 (Method and Apparatus for Blind Checksum and Correction for Network Transmissions)”と題された米国特許出願第 11/097055 号 (代理人事件整理番号 RPS920050076US1/3505P)。

本願と同じ日に出願されて本発明の譲受人に譲渡された、“パケット・ヘッダ・ルックアップを実行する方法及びシステム (Method and System for Performing a Packet Header Lookup)”と題された

50

米国特許出願第 11/096362 号（代理人事件整理番号 RPS920050082US1/3512P）。

本願と同じ日に出願されて本発明の譲受人に譲渡された、“ホスト・イーサネット・アダプタ（HEA）においてブラインド・チェックサムを計算するシステム及び方法（System and Method for Computing a Blind Checksum in a Host Ethernet Adapter（HEA））”と題された米国特許出願第 11/097430 号（代理人事件整理番号 RPS920050089US1/3516P）。

【背景技術】

【0003】

10

図1は、在来のサーバ・システム10を示す。サーバ・システム10は、メイン・メモリ14に接続されたプロセッサ12を含む。プロセッサ12は、そのプライベート・バス（GX）16を介して、ネットワーク・インターフェース・システム18を含むシステムに接続される。ネットワーク・インターフェース・システム18は、PCIバス22等を介してアダプタ20に接続されている。周知されているように、PCI22バスは、それを通して流れることのできるトラフィックの量に影響を及ぼす限られた帯域幅を有する。

【0004】

インターネット及びそのアプリケーションは、サーバが満たさなければならないクライアントのリクエストの数を途方も無く増やした。各クライアントのリクエストは、ネットワークI/O及び記憶装置I/Oの両方を生じさせる。更に、10ギガビットのイーサネット及びIP記憶装置の出現はデータ・センタ通信を単一のバックボーン・インフラストラクチャ：イーサネット、TCP/IP、に整理統合することを可能にする。

20

【0005】

しかし、10ギガビットの速度でのTCP/IPは、主流サーバにおいて膨大な処理及びメモリ帯域幅を消費し、従ってアプリケーションを動作させるサーバの能力を厳しく制限する。

【発明の開示】

【発明が解決しようとする課題】

【0006】

今日のサーバのネットワーク・インターフェース・コントローラ（NIC）においては、TCP及びIPチェックサム、ラージ・センド（Large Send）（又はTCPセグメンテーション・オフロード）のような機能の限られたオフローディングがサポートされる。しかし、これらの機能は1Gまでは充分であるが、10G以上のようなより大きな速度については問題を解決しない。

30

【0007】

TCP/IPプロトコル・スタック全体をサーバから完全にオフロードするためにTCPオフロード・エンジンを用いることが知られている。しかし、TOEのインプリメンテーションは、一般に、ハードウェアで、或いは割合に複雑なピコ・プロセッサ・アーキテクチャにおいてはピココードで、実現される。デバッグ、問題判別及びスタック・メンテナビリティ問題もある。更に、ピコエンジンはメイン・プロセッサ・ロードマップに従わないのでピココードを用いるときにはスカビリティ（scalability）問題がある。最後に、オフロード・エンジンは、通例、新しいプロトコルとAPIとを導入し、従ってアプリケーション及び相互接続性問題における変化を必要とする。

40

【0008】

従って、必要とされるのは、上で特定された問題を克服する、イーサネット環境で大帯域幅データを考慮するシステム及び方法である。本発明は、このような必要に対処する。

【課題を解決するための手段】

【0009】

イーサネット・アダプタが開示される。このイーサネット・アダプタは、このアダプタがプロセッサからパケットを受け取り、またプロセッサにパケットを送ることを可能にす

50

る複数の層を含む。この複数の層は、プロセッサの分割を考慮する多重分離メカニズムを含む。

【0010】

ホスト・イーサネット・アダプタ（H E A）は、イーサネット及びT C P高速化への新しいアプローチを提供する統合イーサネット・アダプタである。T C P／I P高速化特徴のセットがツールキット・アプローチに導入されており、サーバT C P／I Pスタックは、必要とされるときに必要に応じてこれらのアクセラレータを使用する。サーバとネットワーク・インターフェース・コントローラとの間のインターフェースは、P C Iバスを回避することによって合理化されている。

【0011】

H E Aは、ネットワーク仮想化をサポートする。H E Aは、その性能に影響を及ぼすこと無く必須の絶縁及び保護を提供する複数のO Sにより共有され得る。

【発明を実施するための最良の形態】

【0012】

本発明は一般的にサーバ環境に、特にその様な環境で利用されるアダプタに関連する。以下の記述は、当業者が本発明を製造し使用することを可能にするために、特許出願及びその要件という関係において提示されている。本明細書に記載されている好ましい実施態様及び一般の原理及び特徴に対する種々の改変が当業者にとっては自明であろう。従って、本発明は、示されている実施態様に限定されるように意図されてはならず、本明細書に記載されている原理及び特徴と矛盾しない最も広い範囲を与えられるべきである。

【0013】

図2は、本発明によるサーバ・システム100のブロック図である。サーバ・システム100は、メモリ104とインターフェース・アダプタ・チップ106との間に接続されたプロセッサ102を含む。インターフェース・アダプタ・チップ106は、プロセッサ102のプライベート（G x）バスとのインターフェース108と、ホスト・イーサネット・アダプタ（H E A）110とを含む。H E A 110は、プロセッサ102から信号を受信し、またプロセッサ102に信号を送信する。

【0014】

H E A 110は統合イーサネット・アダプタである。サーバ内において、アクセラレータ特徴のセットがT C P／I Pスタックに設けられている。プロセッサ102とインターフェース・アダプタ・チップ106との間のインターフェース100は、P C Iバスを回避することによって合理化されて、多重分離及びマルチキューイング及びパケット・ヘッダ分離を可能にするインターフェース技術を提供している。

【0015】

H E A 110は、G X+バスに直接接続され、従って非常に大きな帯域幅（866 M h zで55.42 G b p s）を持っていて2つの10 G b p sポートの全40 G b p s帯域幅を実際にサポートすることにより、無比の性能レベルを達成する。64ビットP C I－X 133 M H zバスは8.51 M b p sに限定されており、2つの10 G b p sポートのスループットと調和するように少なくともP C I E x p r e s s x 8バスが要求されるということに注意しなければならない。G Xバス上に存在することは、中間の論理を除去するので、転送待ち時間を改善する。

【0016】

そのとき、レガシー・サーバ環境との両立性に配慮しながら高速システムでの改善された機能性に配慮するイーサネット・アダプタが設けられる。この改善された機能性の重要な特徴のうちの幾つかが以下に記載される。

【0017】

高速化機能

H E A 110は、高度な高速化特徴をサポートする。1つの重要な所見は、現在の高速化機能は送信側（すなわち、プロセッサからパケットを送信する）では十分に働くが受信側（すなわちアダプタを通してパケットを受信する）では充分ではないということである

10

20

30

40

50

。H E A 1 1 0 は、パケット多重分離及びマルチキューイング、及びヘッダ分離のような新しい特徴を導入することによってこのギャップに対処する。

【0018】

H E A 1 1 0 の新しい特徴の全てはオプションであり、必要なときにそれらを利用することはT C P / I P スタック次第である。例えば、T C P / I P スタックは、H E A 1 1 0 を使用し、またスループット、短待ち時間及び仮想化サポートのようなH E A の他の特徴を利用することができる。

【0019】

パケット多重分離及びマルチキューイング

マルチキューイング及び多重分離は、仮想化、パー・コネクション・キューイング、及びO S バイパスのような機能をサポートする重要な特徴である。H E A 多重分離は、キュー・ペア、完了キュー及びイベント・キューというコンセプトを用いる。O S プロトコル・スタック要件及び短パケット待ち時間短縮により良く対処するために拡張機能が付け加えられている。

【0020】

システム要件及び構成に依存して、H E A は、入ってくるパケットを、

- － 宛先M A C アドレス（通例、1つのM A C アドレスと、パーティションあたりに1つのデフォルト・キューと）、
 - － 確立された接続についての接続識別子（プロトコル、ソースI P アドレス、宛先I P アドレス、ソース・ポート、宛先ポート）、
 - － 宛先ポート及びオプションでT C P 接続セットアップ・パケット（S Y N）についての宛先I P アドレス、
- に基づいて多重分離することができる。

【0021】

パケット・ヘッダ分離

H E A 1 1 0 は、T C P / I P ヘッダをデータ・ペイロードから分離することができる。この特徴は、アプリケーションにより掲示された受信バッファを汚さずに処理されるようにヘッダをプロトコル・スタックに向けることを可能にする。

【0022】

拡張された特徴

サーバ環境においてH E A 1 1 0 により多くの拡張された特徴が提供される。それらの特徴のうちの幾つかが以下に列挙される。

【0023】

1. 複数の受信キュー。キュー・ペアのコンセプトが、ペアあたりに2つ以上の受信キューをサポートするように拡張される。これは、スタックがそのバッファ・プール・メモリをより良く管理することを可能にする。例えば、1つのキューを小さなパケットに、1つを中型のパケットに、そして1つを大きなパケットに割り当てることができる。H E A 1 1 0 は、受信されたパケットのサイズに応じて、そのためのキューを選択する。

【0024】

2. 短待ち時間キュー。送信側においてディスクリプタ（W Q E）は即値データを含むことができ、その様な場合にはデータを送信するためにインダイレクションすなわちシステム・メモリからの付加的なD M A は不要である。受信側において、短待ち時間キューはバッファを供給せずに、即値パケット・データを受信する。H E A 1 1 0 は受信キューに直接書き込む。短いパケットはこの特徴を利用し、D M A 動作の劇的減少という結果をもたらす。すなわち、パケットあたりに1つのD M A リード及び1つのD M A ライトに対してパケットあたりに唯一のD M A ライト。

【0025】

受信短待ち時間キューは、パケット・ヘッダ分離をサポートするためにも使用される。ヘッダは短待ち時間キューに書き込まれ、ペイロードはその場合の受信キューにおいて示されているバッファにD M A される。

【0026】

要約すると、多重分離及びマルチキューイング、及びパケット・ヘッダ分離は、仮想化の基本的構成要素であり、短待ち時間操作を提供する。更に、これらの特徴が在来のOSプロトコル・スタック性能を改善するためにも使用され得ることに留意するべきであり、例えばパー・コネクション・キューイングは、コードの除去に配慮しており、そしてもっとも重要なことに、メモリ内の適切な情報の位置を突き止めるために消費されるメモリ・アクセスと、関連した機能停止／キャッシュ汚染とを減少させる。

【0027】

HEA110の特徴をより詳しく記述するために、以下の記述を添付図面と関連させて参照されたい。

10

【0028】

図3は、本発明によるHEA110の略ブロック図である。見たとおりに、HEA110は3層アーキテクチャを有する。第1層は、イーサネット・ネットワーク上の他の装置からの及びそれらの装置への複数のインターフェースを提供する媒体アクセス・コントローラ(MAC)及び並直列変換回路(Serdes)層202を含む。層202において、複数のインターフェースを提供するために同じチップI/Oが使用される。例えば、好ましい実施態様では、10ギガビット・インターフェース又は1ギガビット・インターフェースを提供するために同じチップI/Oが利用される。

【0029】

第2層は、パケット高速化及び仮想化層204を含む。層204は、パケットを受信して、仮想化を可能にするためにパケットのフローを多重分離することに配慮している。層204は、パケットに基づくサーバのオペレーティング・システムの仮想化又は分割を可能にする。層204は、ゼロ・コピー操作を可能にし、従って改善された待ち時間を提供するためにパケット・ヘッダ分離も提供する。また、層204はホスト・インターフェース層206を通してプライベート・バス(Gx)と直接相互に作用するので、短待ち時間、大帯域幅接続が提供される。

20

【0030】

第3層はホスト・インターフェース層206を含む。ホスト・インターフェース層206は、プロセッサのGx又はプライベート・バスへのインターフェースを提供し、層204と通信する。層206は、TCPスタックのために効率的バッファ管理を可能にするためにキュー・ペア(QP)あたりに複数の受信サブ・キューに配慮している。ホスト層206は、データ・パケットの所与のフローのためにコンテキスト管理を提供する。

30

【0031】

HEA100の層202、204及び206の各々の特徴をより詳しく記述するために、以下の検討を添付図面と関連させて参照されたい。

【0032】

MAC及びSerdes層202

図4は、MAC及びSerdes層202をより詳しく示すHEA110のブロック図である。見たとおりに、この実施態様には1つの10ギガビットMAC302と4つの1ギガビットMAC304a及び304bとがある。MAC302、304a及び304bは、受信されたパケットを整列させて符号化するためのアナログ符号化ユニット308a、308b及び308cを含む。MAC302、304a及び304bは高速並直列変換回路(HSS)306に接続されている。高速Serdes306は、1つの10ギガビット・ソース又は4つの1ギガビット・ソースからデータを受信することができる。

40

【0033】

受信イーサネット機能(RxNet)概観

このセクションは、層202内の受信イーサネット機能の高レベル構造及びフローを示す。以下でより詳しく説明されるRxアクセラレータ・ユニット400は、パケット高速化及び仮想化層204の一部である。

【0034】

50

図5は、R x N e t の1つについてコンポーネント及びデータフローを示す。データは、インターフェース302に到着して、この実施態様ではパケット・データを64ビット(10G)又は32ビット(1G)の平行・データ・バスにおいてアセンブルし整列させ高速S e r d e s 304と、アナログ符号化ユニット308a及び308bと、M A Cとによって処理される。フレームの開始及び終了及び他のパケット情報を示す制御信号も生成される。データ及び制御は、層206(図2)の受信パケット・プロセッサ(R P P)による処理に備えて構文解析、フィルタリング、チェックサム及びルックアップ機能を実行するR x A c c e l ユニット400を通過する。この実施態様では、R x A c c e l ユニット400に入るとクロックは4.6nsクロックに変換され、データ幅は128bに変換される。

10

【0035】

データがR x A c c e l ユニット400を通過してホスト層206内のデータ・バッファに流れるとき、R x A c c e l ユニット400は制御及びデータを調査してその処理を開始する。R x A c c e l ユニット400の結果がパケットの末端と同期化されるようにデータ・フローはR x A c c e l ユニット400内で遅延させられる。このとき、R x A c c e l ユニット400の結果は、M A C 302からの何らかのオリジナル制御情報と共にコマンド・キューに渡される。この制御情報は、データと共にバッファに格納される。

【0036】

R x A c c e l ユニット400は、ルックアップ・エントリがキャッシングされていないければ、G Xバス・インターフェースを通してメイン・メモリに行かなければならないかもしれない。G Xバスは4.6nsで動作する。ホスト層206は、R x A c c e l ユニット400からのキュー・ペア解像度(queue pair resolution)情報を非同期的に読むことができる。

20

【0037】

送信イーサネット機能(T x E n e t)概観

このセクションは、イーサネット及び高速化機能の送信構造及びフローの概観を提供する。以下でより詳しく説明されるT x アクセラレータ・ユニット500は、パケット高速化及び仮想化層204の一部である。

【0038】

図6は、1つのT x E n e t についてのコンポーネント及びデータフローを示す。パケット・データ及び制御はH E A 110のT x A c c e l 500コンポーネントから到着する。T x アクセラレータ(T x A c c e l)ユニット500は、制御情報を解釈して、ユニット500を通過して流れるパケットのヘッダのフィールドを改変する。それは、制御情報又はパケット・ヘッダにおいて見出された情報に基づいてラップ対ポート決定(w r a p v e r s u s p o r t d e c i s i o n)を行う。それは、また、T x M A C 302及び304のための適切な制御を生成する。T x A c c e l ユニット500がパケット・ヘッダを、M A C 302及び304へ流れる前に、更新できるように、データ・フローはT x A c c e l ユニット500において遅延させられる。出口で、データ幅は128ビットから64ビット(10G)又は32ビット(1G)に変換される。データ及び制御は、M A C 302及び304の異なるクロック・ドメインに入るためにT x A c c e l ユニット500のクロック変換機能を通過する。M A C 302及び304、アナログ変換器508a及び508b、及び高速S e r d e s 306は、イーサネット・インターフェースのためにパケットをフォーマットする。

30

40

【0039】

パケット高速化及び仮想化層204

図7は、パケット高速化及び仮想化層204をより詳しく示すH E A 110のブロック図である。H E A 層204は、前記の受信(R x A c c e l)高速化ユニット400と送信高速化(T x A c c e l)ユニット500とを含む。R x A c c e l ユニット400は、受信バックボーン(R B B)402と、パーサ・フィルタ・チェックサム・ユニット(P F C)404と、ルックアップ・エンジン(L U E)406と、M I Bデータベース4

50

08とを含む。TxAccelユニット500は、送信バックボーン502と、ルックアップ・チェック504と、MIBエンジン506とを含む。Rx高速化ユニット400及びTx高速化ユニット500の動作が以下で詳しく説明される。

【0040】

受信高速化 (RxAccel) ユニット400

このセクションはRxAccelユニット400の高レベル構造を説明する。図8は、RxAccelユニット400の詳細図である。RxAccelユニット400は、受信バックボーン(RBB)402と、パーサ、フィルタ及びチェックサム・ユニット(PFC)404と、ローカル・ルックアップ・ユニット(LLU)406と、リモート・ルックアップ・ユニット(RLU)408と、MIBデータベース410とを含む。

10

【0041】

データは、受信MACからRxAccelユニット400を通して変更されずに流れる。RBB402はデータのフローを管理し、クロック及びデータ・バス幅変換機能を担う。受信MACから受信された制御及びデータは、高速化機能を実行し廃棄決定を行うためにPFC404により使用される。PFC404は、RBB402のためにキュー・ペア番号(QPN)を解明するためにフレームから抽出された5タプル・キーを含む制御及びデータをLLU406に渡す。LLU406は、QPNを直ちに見出すか、又はスロットを確保するためにキャッシュ・エントリを割り当てる。もし現在のキーがキャッシュ内に無ければ、LLU406はメイン・メモリにおいてキーを捜す。PFC404は、パケット統計を格納するためにMIBデータベース410へインターフェースする。

20

【0042】

Tx高速化500

このセクションは、送信高速化ユニット500(TxAccel)の高レベル構造とフローを説明する。

【0043】

図9は、TxAccelユニット500が2つの送信バックボーン(XBB)502a及び502b、2つの送信チェックサム・ユニット(XCS)504a及び504b、2つの送信MIB506a及び506b、1つのラップ・ユニット(WRP)508、及び1つのポーズ・ユニット(PAU)論理510から構成されていることを示す。データは、ENOpからTxAccelを通して流れて、IP及びTCPチェックサム・フィールドを調整するために改変される。XBB502a及び502bは、データのフローを管理し、クロック及びデータ・バス幅変換機能を担う。ENOpから受信された制御及びデータは、チェックサム機能を実行するためにXCS504a及び504bにより使用される。パケットがMAC層202によって送信(又は廃棄)された後、送信ステータスはアカウントिंगのためにTxAccelに戻る。XBB502は、情報をTxAccelユニット500のクロック・ドメインに変換する。ステータス情報は、XCS504によりパケットから得られたオリジナル情報と合併されてMIBカウンタ論理506a及び506bに渡される。MIB論理506a及び506bは、MIBアレイ内の適切なカウンタを更新する。ラップ・ユニット(WRP)508は、XCS504a及び504bがラッピングするように決定したパケットを受信側に転送する役割を担う。ポーズ・ユニット(PAU)510は、受信バッファの占有率に基づいてポーズ・フレームを送信するようにMACに命令する。

30

40

【0044】

ホスト・インターフェース層206

図10は、ホスト・インターフェース層206をより詳しく示すHEA110のブロック図である。ホスト・インターフェース層206は、層204からパケットを受信するための入力バッファ602とパケットを層204に提供するための出力バッファ604とを含む。層206は、入力バッファ内のパケットを適切に処理するための受信パケット・プロセッサ(RPP)606を含む。コンテキスト管理メカニズム908は、TCPスタックのための有効なバッファ管理を可能にするために、前のキューあたりに複数のサブ・キ

50

ユーを提供する。

【0045】

多重分離機能

ホスト・インターフェース層206のコンポーネントと関連する層204のRxユニット400は、パケットがプロセッサの適切な部分に提供されることを確実にするためにパケットを多重分離する。従って、受信されたパケットは、それらがサーバの適切な部分へ流れることを確実にするために、多重分離されなければならない。

【0046】

この多重分離機能の詳細を説明するために、図8及び9と関連させて下記を参照されたい。

【0047】

HEAアダプタにおける多重分離インプリメンテーション

受信パケット・プロセッサ(RPP)606が受信されたパケットに対して作用し得るようになる前に、キュー・ペア・コンテキストが検索されなければならない。QP接続マネージャは、QP番号を用いてこれを実行する。QP番号はTCP/IPパケットで運ばれないので、該番号は他の手段で判定されなければならない。2つの一般的種類のQP、すなわちパー・コネクションQP及びデフォルトQP、がある。

【0048】

パー・コネクション・キュー・ペア(QP)

パー・コネクションQPは、IPパケットのフラグメンテーションが予期されなくて短待ち時間が予期される長命の接続のために使用されるように意図されている。それは、HEA110により提供されるユーザ・スペーシング・キューイング・メカニズムをアプリケーションがサポートすることを必要とする。この実施態様では、宛先MACアドレスを用いて初めに論理ポートが見出されなければならない。パー・コネクションQPのために3つのタイプのルックアップが存在する。

【0049】

1. 特定の宛先IPアドレス及び宛先TCPポートのための新しいTCP接続。もしパケットがTCP SYNパケットであったならば、TCP/IP(DA, DP, 論理ポート)に基づいてルックアップが実行される。

【0050】

2. 特定の宛先TCPポートだけのための新しいTCP接続(DAを無視する)。もしパケットがTCP SYNパケットであったならば、TCP/IP(DP, 論理ポート)に基づいてルックアップが実行される。

【0051】

3. 現存するTCP/UDP接続。もしパケットが断片化されていないユニキャストTCP又はUDPパケットであったならば、TCP/IP5タプル及び論理ポートに基づいてルックアップが実行される。

【0052】

デフォルト・キュー・ペア

パケットについてパー・コネクションQPを見出し得ないか、又はMACアドレスについてパー・コネクション・ルックアップが可能にされないか、又はパケットが再循環させられたマルチキャスト/ブロードキャスト・パケットであれば、デフォルトQPが使用される。一般にデフォルトQPはカーネル・ネットワーキング・スタックによって処理される。これらのタイプのデフォルトQPがHEA110に存在する。

【0053】

1. 論理ポートあたりのデフォルトOSキュー。論理ポートは、それ自身のデフォルト・キューを有する論理イーサネット・インターフェースに対応する。各論理ポートは論理スイッチ上に別のポートを有する。

MACアドレスに基づいてルックアップが実行される。

デフォルトOSキューへのダイレクト索引(論理ポート番号)に、再循環させられた(

10

20

30

40

50

ラッピングされた) マルチキャスト／ブロードキャスト・パケットが供給される。

【0054】

2. マルチキャスト (MC) 又はブロードキャスト (BC) キュー。

パケットがMACルックアップ・テーブル内のMACアドレスのうちの1つと一致しないマルチキャスト又はブロードキャスト・パケットである場合の、設定された値。

【0055】

3. スーパー・デフォルト・ユニキャスト (UC) キュー。

UCパケットが設定されたMACアドレスのうちの1つと一致しなければ、デフォルトUC QPNが使用され得る。

【0056】

このメカニズムは、接続ごとのキューイングと論理ポートごとのキューイング (OSキュー) という両極端の間の柔軟性に配慮している。両方のモデルが、それ自身のキューイングを有する幾つかの接続、及びデフォルト論理ポート・キューで待ち合わせさせられる幾つかの接続と共に働くことができる。

【0057】

接続ルックアップはRxAccelユニット400によって実行される。1つのそのようなユニットが各ポート・グループのために存在する。RxAccelユニット400内で、各コンポーネントはプロセスの一部を実行する。PFC404は、パケット・ヘッダから必要とされるフィールドを抽出し、宛先MACアドレスに基づいて論理ポート番号を判定する。ローカル・ルックアップ・ユニット (LLU) 406とリモート・ルックアップ・ユニット (RLU) 408とは、QP番号を解明する責任を有する。LLU406は、ローカル・リソースだけ (キャッシュ及びレジスタ) を用いてQPNを見出そうと試みる。

【0058】

LLU406の目的は、受信されたパケットに関連付けられているQP番号を判定しようとする試みることである。QP番号は、VLIM及びRPP606により必要とされる。それは、可能ならばこのタスクをローカルに (すなわち、システム・メモリに行かずに) 実行する。

【0059】

QP番号は、数個の方法、すなわち

- TSキャッシュにおけるルックアップ
- デフォルト・パーティションQP
- デフォルトUC QP

のうちの1つでローカルに見出され得る。

【0060】

もし一致がローカルに見出されなければ、エントリがシステム・メモリ内に現在あるか否か調べるために予備チェックが行われる。もしあるならば、検索を実行するためにRLU408が呼び出される。もしRLU408が使用中ならば、リクエストのキューを形成することができ、それはRLU408に、それが自由になると、提供される。

【0061】

LLU406は、一時的キューイングに用いるQP番号又はキュー・インデックス或いはその両方を提供するRBB402と通信する。もし適切なエントリがキャッシュにおいて利用できなければ、LLU406はRBB402に検索がビジーであることを示す。この場合にはそのパケットは落とされなければならない。

【0062】

キュー・インデックス解明が要求されて解明されたときにLLU406はQPNをホスト層406に提供する。RLU408は、システム・メモリ・テーブルを用いてQPNを見出そうと試みる。

【0063】

LLU406は、TCP／UDPパケットのためのQPNを見出すためにローカル64

10

20

30

40

50

エントリ・キャッシュを利用する。もしそのエントリが該キャッシュにおいて見出されたならば、RLU408は呼び出されなくてもよい。もしそのエントリが該キャッシュにおいて見出されなければ、そのエントリが接続テーブル内にあるか否か調べるために予備チェックが該キャッシュにおいて行われる。該キャッシュは、数個の設定されたキューがあるときにメイン・メモリへの不要なアクセスを無くするのに役立つ。

【0064】

もしRLU408が呼び出されたならば、それは、メモリ???から128バイトのダイレクト・テーブル(DT)エントリを取り出すために6タプル(論理ポート番号を含む)のハッシュを使用する。このDTエントリは8個に及ぶ6タプル・パターンと関連するQPNとを含む。もし一致が見出されれば、それ以上の動作は不要である。

10

【0065】

RLU408が呼び出されなければならないときには、パケットが入力バッファに入れられるときにQPNを急いで判定することはできない。実際には、QPNは数パケット後に判定され得る。この理由から、RxAccelユニット400は、パケット・キューイングのためにQPN又はキュー・インデックスをホスト層206に提供することができる。QPNが提供されるのであれば、ホスト層206(アンローダ)は、RPPにより処理されるようにパケットを直接キューに入れることができる。キュー・インデックスが提供されるのであれば、ホスト層206は、QPNの解明を待つためにこのパケットを保持しなければならない。QPNは、常に、RPPがディスパッチされる時まで判定される。

【0066】

20

仮想化

高速データ経路は複数のパーティションにより共有されそうであり、また高速イーサネット処理能力がサーバにおいては重要なので、HEAが

- 複数のパーティションによるアダプタ・シェアリングを提供し、
- ネイティブの処理能力、すなわち、“専用アダプタの場合のように(as with a dedicated adapter)”、を考慮に入れ、
- ネイティブの価値付加特徴、すなわち、“専用アダプタの場合のように”(ラージ・センド、パー・コネクション・キューイング、...)を考慮に入れ、
- パーティション間の分離を考慮に入れ、
- パーティション接続性を提供する

30

ことが非常に重要である。

【0067】

パーティションは、トランスペアレントに、すなわちそれらが同じ物理的サーバに配置されているか、それとも本物のイーサネットによって接続された別々の物理的サーバに配置されているかにかかわらず同じように、通信することができなければならない。

【0068】

今日、イーサネット仮想化は、アダプタを持っているサーバ・パーティションにおけるスイッチング又はルーティングによってサポートされ、この余分のホップは性能ボトルネック(データ・コピー、3ドライブ・ドライブ、...)を生じさせる。HEA110は、使用パーティションとアダプタとの間の直接データ及び制御経路(余分のホップではない)を提供するように設計される。換言すれば、HEAは、各パーティションに、それ自身の“仮想”アダプタ及び“論理”ポートを供給する。HCAの場合のように、全てのHEAリソース及び機能はパーティションごとに割り当て/イネーブルされ得、パーティション相互の保護及び分離を提供するために正確に同じメカニズムが使用される。

40

【0069】

データ経路

データ経路に関して、図11に示されているように、HEA110は、マルチキャスト処理及びパーティション910a-910c間通信を提供するために、物理的ポート902及び904ごとに論理層2スイッチ906及び908を提供する。HEA内でのこのサポートの実現は、全体としてのシステムのソリューションを簡単に保ち(特に、ソフトウ

50

エアについての透明性)、高性能を提供する。パーティション間の通信のために全てのH E Aハードウェア高速化及び保護が利用可能である。

【0070】

上記のフローをサポートするために、考えやすい方法は、所与の物理的ポートに関連付けられた全ての論理ポートとその物理的ポート自体とが取り付けられている論理層2スイッチ902及び904を想像することである。問題は、この論理スイッチをどのようにしてどこに実現するかであり、選択肢は、ファームウェア/ソフトウェアにおける完全なエミュレーションからH E Aハードウェアにおける完全なインプリメンテーションまでにわたる。物理的ポートごとに1つの論理層2スイッチがあり、これらの論理スイッチは互いに接続されない。

10

【0071】

システム構成

レガシーOS TCP/IPスタックを伴う仮想化されたH E A

図12は、レガシーOS TCP/IPスタック1102, 1104及び1106と共に使用されるH E Aを示す。

- アプリケーション1108a-1108cは変化していない
- TCP/IPスタック1102, 1104及び1106は変化していない
- H E A110をサポートするデバイス・ドライバ1107a, 1107b及び1107cが必要とされる。

【0072】

TCP/IPスタック(OS)は、TCP接続ごとの短パケット(単数又は複数)多重分離のための短待ち時間キューのような特徴を利用するためにオプションで拡張され得る。見たとおり、パケットの多重分離はMACアドレスとパーティションごとのQPNとに基づいて実行される。

20

【0073】

レガシーOSスタック及びユーザ・スペースTCP/IPを伴う仮想化されたH E A

図13は、幾つかのパーティションがユーザ・スペースTCPスタック1220及びレガシーOSスタック108a及び1208bをサポートするシステムにおいて使用されたH E A110を示す。

- ユーザ・スペースTCPをサポートするアプリケーションはソケット拡張APIを使用するように要求され得る
- 他のパーティションは通常のOS TCP/IPスタックを使用することができる
- ユーザ・スペースTCP(User Space TCP)をサポートするパーティション内の或るアプリケーションも通常のTCP/IPスタックを使用することができる(デフォルト経路)。ユーザ・スペースTCP1220は、カスタマー識別(Cid)情報とカスタマーのためのQPNとに基づいてH E Aにより多重分離される。

30

【0074】

論理スイッチは、アダプタにおいて完全にサポートされる。H E Aハードウェアの複雑さを最小にするために、H E Aは、マルチキャスト/ブロードキャスト・パケット複製のためにソフトウェア・エンティティ、すなわちマルチキャスト・マネージャ、に依拠する。H E Aは、パケット・コピーを宛先パーティションに渡すためにマルチキャスト・マネージャにアシストを提供する。

40

【0075】

外部ユニキャスト・トラフィック

送信ユニキャスト・トラフィックは、パーティションに割り当てられたQPを通して処理される。それは接続ごとの専用キュー・ペア又は論理ポートごとの単一キュー・ペア或いはその両方であり得る。送信キューの間での公平なスケジューリングがH E Aにより提供される。システム構成に依存して、QPアクセスはアプリケーション(ユーザ・スペース)に又はOSスタック(特権)にだけ与えられ得る。

【0076】

50

受信されたユニキャスト・トラフィックは次のように多重分離される。

- ー 宛先論理ポートを見出すために宛先MACアドレスでルックアップが行われる
- ー もし宛先論理ポートのためにパー・コネクション・キューイングが可能にされるならば、その接続に関連するQPを見出すために接続IDで第2ルックアップが行われる
- ー もしパー・コネクション・キューイングが可能にされないか、又は接続QPが特定の接続のためにセットアップされていなければ、入ってくるメッセージは、宛先論理ポートに関連する“デフォルト”QPにキューイングされる。

【0077】

パーティション間ユニキャスト・トラフィック

図14は、サポートされるパー・コネクション・キューイングを含む全てのHEA高速化特徴を示すブロック図である。完全な透明性がパーティションのデバイス・ドライバに提供される。

【0078】

パーティション・スタックは、パケットを送信するためにパー・コネクションQP又はデフォルトQPを使用する。パケットがHEA送信側によって処理されるとき、HEAは、宛先MACアドレスが同じ物理的ポート上に定義された論理ポートに関連するMACアドレスであることを検出する（換言すれば、宛先MACアドレスは、送信論理リンクとは同じ層2論理スイッチに属する受信論理リンクを識別する）。従って、HEAはパケットをラップする。HEA受信側は、パケットを、あたかもそれが物理的リンクから受信されたかのように処理し、従ってまさに同じ高速化特徴が使用される。

【0079】

IPの場合には、IPスタックは、同じIPサブネット上に位置する宛先パーティションの宛先MACアドレスを発見するために通常のメカニズムを使用することができる。このパーティションは同じサーバ上に配置されてもされなくても良く、これはスタック及びデバイス・ドライバの両方のためにトランスペアレントである。

【0080】

外部及びパーティション間マルチキャスト／ブロードキャスト・トラフィック

HEAは、関係のあるパーティションにマルチキャスト・パケット及びブロードキャスト・パケットを複製するための備えを持っていない。代わりに、それは、受信した全てのMC/BCパケットを、マルチキャスト・マネージャ機能により所有されるQPに転送する。この機能は、必要に応じてパケットを複製し、HEAトランスポート能力を使用してコピーを関係のあるパーティションに配布する。

【0081】

受信

図15は、インバウンド・マルチキャスト送信を示す。受信されたマルチキャスト・パケット及びブロードキャスト・パケットは初めにHEAフィルタリング機能を通る。もし廃棄されなければ、パケットは、マルチキャスト・マネージャ1500により所有されるQPに向けられる。パケットはシステム・メモリに転送され、マルチキャスト・マネージャ1500は起動される。マルチキャスト・マネージャ1500は、どの論理ポートがパケットのコピーを受信するべきかを決定し（マルチキャスト・フィルタリング）、パケット複製を処理する。マルチキャスト・マネージャは、パケットを受信側パーティション1502a-1502cに再分配するためにHEA110機能を使用することができる。

【0082】

そうするためにマルチキャスト・マネージャは、受信されたパケットを参照するn個の受け側ディスクリプタ(WQE)をその送信キューにエンキューする。パケットはその受け側にそっくりそのまま送られなければならない、特にマルチキャスト宛先MACアドレスの代わりにその種々の受け側のユニキャスト・アドレスを用いることは容認できないことに注意しなければならない。従って、パケット・ディスクリプタは、HEAがパケットをその適切な宛先に向けることができるように情報を含んでいなければならない。この情報は、受け側のデフォルトQP又はその論理ポートID又はそのMACアドレスで

10

20

30

40

50

あり得る。パケットが送信されるべく選択されると、H E A 送信側は、Q P 及びW Q E の両方に含まれている情報のおかげで、パケットがラップを介して送られる必要があると判定する。データと共に、受け側Q Pを判定する情報がH E A 受信側に転送される。H E A 受信側は、パケットを受け側Q Pにエンキューするためにこの情報を使用する。

【0083】

送信

図16は、アウトバウンド・マルチキャスト送信を示す。ブロードキャスト／マルチキャスト・パケットは、発信パーティションにより通常の手続きを用いて送信される。パケットがH E A 送信側によって処理されるとき、H E A は、宛先M A C アドレスがブロードキャスト又はマルチキャストされることを検出し、またW Q E において“ フォースアウト (F o r c e _ O u t) ” オプションがセットされていないことを検出する。従って、H E A はパケットをラップする。H E A 受信側は、パケットを上記のように処理する。マルチキャスト・マネージャは、下記の追加事項を伴って上記のようにパケットを処理する。

－ それは、センダが受け側のリストから除去されることを確実にしなければならず、それはパケットのソースM A C アドレスをフィルタとして用いて行う。

－ パケット複製プロセス中にV L A N フィルタリングが実行され得る。パケットはV L A N の構成要素に送られるだけである。

－ 内部複製が行われると、パケットは物理的ポートから送出されなければならない。それは、そのQ P のフォースアウト機能をイネーブルしてW Q E の“ フォースアウト ” ビットをセットすることによって、その様に行う。このビットがセットされるたとき、H E A は直ぐにパケットを物理的リンク上に送り出す。

【0084】

マルチパーティション環境におけるマルチキャスト・フィルタリング

受信側において、H E A はマルチキャスト・フィルタリングを提供する。H E A は、他の“ 直ぐに入手可能な (o f f t h e s h e l f) ” アダプタと同様に、宛先M A C アドレスのハッシュ値と物理的ポートあたりに1つのフィルタリング・テーブルへのルックアップとに基づいてベスト・エフォート・フィルタリングを提供する。この機能の意図は、マルチキャスト・トラフィックを制限することであるが、“ 最終の ” フィルタリングはスタックに任される。マルチ・パーティションの場合、関与する全てのパーティションからのフィルタリング要求はマルチキャスト・マネージャにより併合され、その後にH E A で設定されるべきである。

【0085】

マルチキャスト・マネージャは、関係するパーティションへのパケット配布を処理するときにパーティションごとのマルチキャスト・フィルタリングを行うことができる。

【0086】

パケット・ヘッダ分離

H E A 1 1 0 は、データ・ペイロードからT C P / I P ヘッダを分離することができる。この特徴は、ゼロ・コピー操作を可能にし、従って待ち時間を改善する。

【0087】

パケット・ヘッダ分離は、Q P コンテキストで設定されたときH E A 1 1 0 により実行される。設定されたとき、イーサネット / I P / T C P 又はイーサネット / I P / U D P ヘッダは、パケットのボディから分離されて異なるメモリに置かれる。通常、T C P / I P スタックはヘッダを処理し、アプリケーションはボディを処理する。ハードウェアでの分離は、ユーザ・データをユーザ・バッファに整列させ、従ってコピーを回避することを可能にする。

【0088】

層204内のP F C 4 0 4 (図8) は、ヘッダ全長 (8 ビット) をホスト・インターフェース206のR P P 6 0 6 (図10) に渡して、イーサネット、I P 及びT C P / U D P ヘッダのバイト数を示す。ヘッダ分割が行われなるときにはヘッダ長さは0にセットされる。

【0089】

Q Pは2つ以上の受信キュー（R Q）のために設定されなければならない。

【0090】

パケットがT C P又はU D P（ヘッダ長さがゼロでない）ならば、R P P 6 0 6はヘッダをR Q 1 W Q Eの中に置く。R P P 6 0 6は、その後、パケットのデータ部分のために適切なR Qを選択する（R Q 2又はR Q 3）を選択する。R Q 2又はR Q 3 W Q Eのディスクリプタは、残りのデータを配置するために使用される。R P P 6 0 6は、完全な情報でC Q Eが生成されるべきことを示す。ヘッダ分割フラグがセットされる。C Q Eのコリレータ（c o r r e l a t o r）フィールドのコリレータが、使用されたR Q 2又はR Q 3 W Q Eからコピーされる。第1 W Q Eに置かれたヘッダ・バイトの総数もC Q Eに置かれる。

10

【0091】

もしヘッダがR Q 1 W Q E内の使用可能なスペースより大きければ、W Q Eはなるべく多くのデータで満たされ、C Q Eにおいてヘッダ長すぎフラグ（H e a d e r T o o L o n g f l a g）がセットされる。ヘッダの残りの部分はデータと共にR Q 2 / R Q 3 W Q Eに置かれる。

【0092】

ヘッダ分割モードがA L Lにセットされてヘッダ分割が実行される（ヘッダ長さがゼロでない）時には、パケットのボディはR Q 1 W Q Eには決して置かれていない。Q Pは、オプションで、短いパケットを完全にR Q 1 W Q Eの中に置くように設定され得る（ヘッダ分割モード＝M L）。もしその様に設定されれば、パケット長さがR Q 2 スレシールドより小さければ、R Q 1 W Q Eだけが使用され、ヘッダ分離は行われない。ボディがR Q 1 及びR Q 2 / R Q 3 に決して分割されないことに注意しなければならない。

20

【0093】

もしパケットがI Pフラグメントであるか又はT C PでもU D Pでもなく（ヘッダ長さがゼロ）、またパケットが大きすぎてR Q 1 W Q Eに入らなければ、パケット全体がR Q 2 又はR Q 3 W Q Eを用いて配置される。ヘッダ総数はゼロにセットされる。ヘッダ分割フラグはオフである。（競争情報がR Q 1 W Q Eに置かれていないのであれば）R Q 1 W Q Eは消費されない。

【0094】

従ってH E A 1 1 0はT C P / I Pヘッダをデータ・ペイロードから分離することができる。この特徴は、アプリケーションにより揭示された受信バッファを汚さずに処理されるようにヘッダをプロトコル・スタックに向けることを可能にし、従って或るトランザクションのために待ち期間を短縮する。

30

【0095】

概要

従って、本発明のホスト・イーサネット・アダプタ（H E A）は、プロセッサのプライベート・バスに直接接続され、従って2つの1 0 G b p sポートの全4 0 G b p s帯域幅をサポートする充分な帯域幅（例えば8 6 6 M h zで5 5 . 4 2 G b p s）を有することによって無比の性能レベルを達成する。プロセッサのプライベート・バス上に該アダプタを有することで中間の論理を除去し、従って転送待ち時間を改善する。H E A 1 1 0を利用して、在来のN I Cの場合より高速で、短待ち時間で簡単な論理を考慮したネットワーク・インターフェース・コントローラ（N I C）を提供することができる。

40

【0096】

本発明は、示された実施態様に従って記述されたけれども、当業者は、それらの実施態様に変化形があり得ること、それらの変化形が本発明の範囲内にあることを直ぐに認めるであろう。従って、添付されている請求項の範囲から逸脱せずに当業者により多くの改変がなされ得る。

【図面の簡単な説明】

【0097】

50

【図 1】 在来のサーバ・システムを示す。

【図 2】 本発明によるサーバ・システムのブロック図である。

【図 3】 本発明による H E A の簡単なブロック図である。

【図 4】 M A C 及び S e r d e s 層をより詳しく示す、H E A のブロック図である。

【図 5】 R x N e t のうちの 1 つのためのコンポーネント及びデータフローを示す。

【図 6】 1 つの T x E n e t のためのコンポーネント及びデータフローを示す。

【図 7】 パケット高速化及び可視化層をより詳しく示す H E A のブロック図である。

【図 8】 R x A c c e l ユニットのより詳しい図である。

【図 9】 R x A c c e l ユニットが 2 つの送信バックボーン (X B B) と、2 つの送信チェックサム・ユニットと、2 つの送信 M I B と、1 つのラップ／ユニットと 1 つのポーズ・ユニットとから構成されることを示す。

【図 10】 ホスト・インターフェース層をより詳しく示す H E A 1 1 0 のブロック図である。

【図 11】 1 つの物理的ポートにつき 1 つの論理層 2 スイッチを提供する H E A を示す。

【図 12】 レガシー O S T C P / I P スタックと共に使用される H E A を示す。

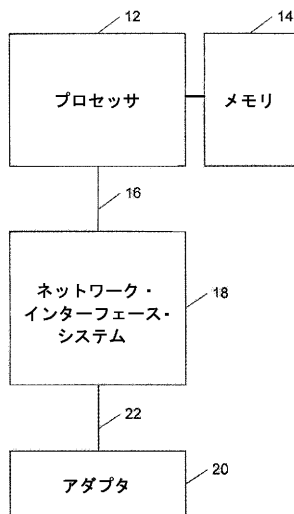
【図 13】 幾つかのパーティションがユーザ・スペース T C P スタックをサポートしているシステムに使用された H E A を示す。

【図 14】 パー・コネクション・キューイングを含む全ての H E A サポート高速化特徴を示す。

【図 15】 インバウンド・マルチキャスト送信を示す。

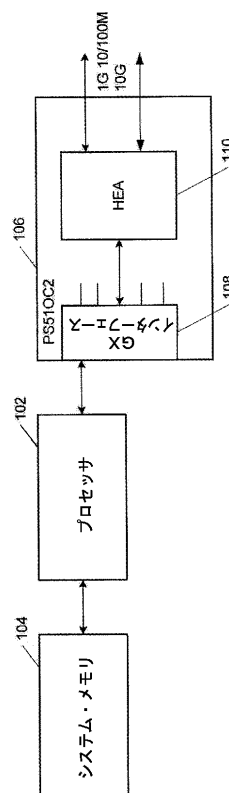
【図 16】 アウトバウンド・マルチキャスト送信を示す。

【図 1】



10

【図 2】

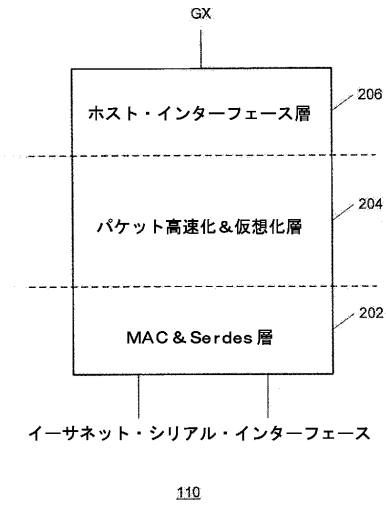


100

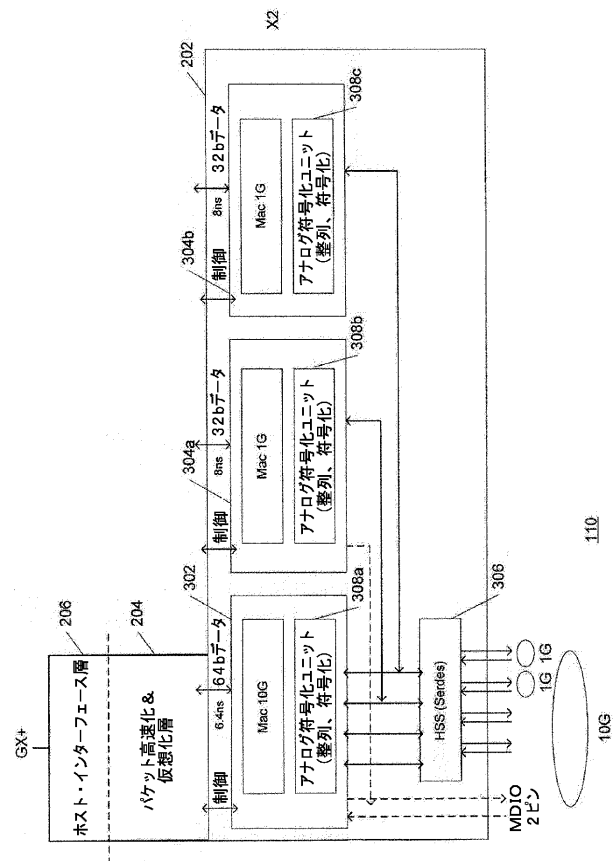
10

20

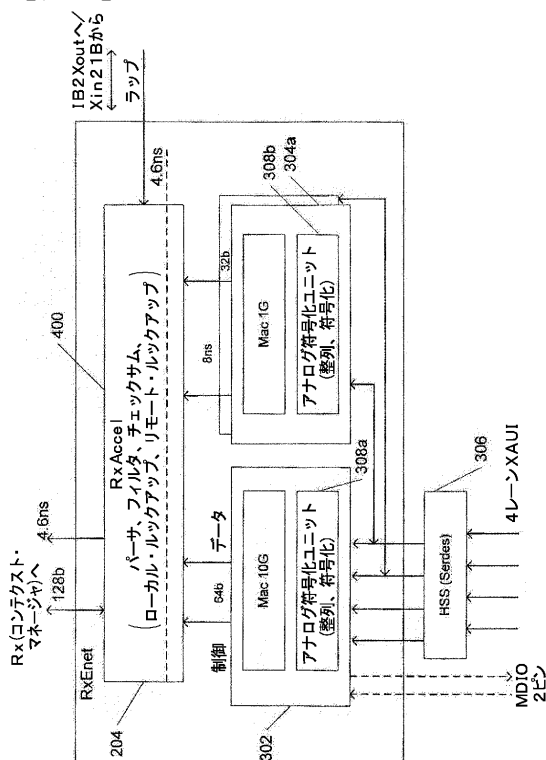
【図 3】



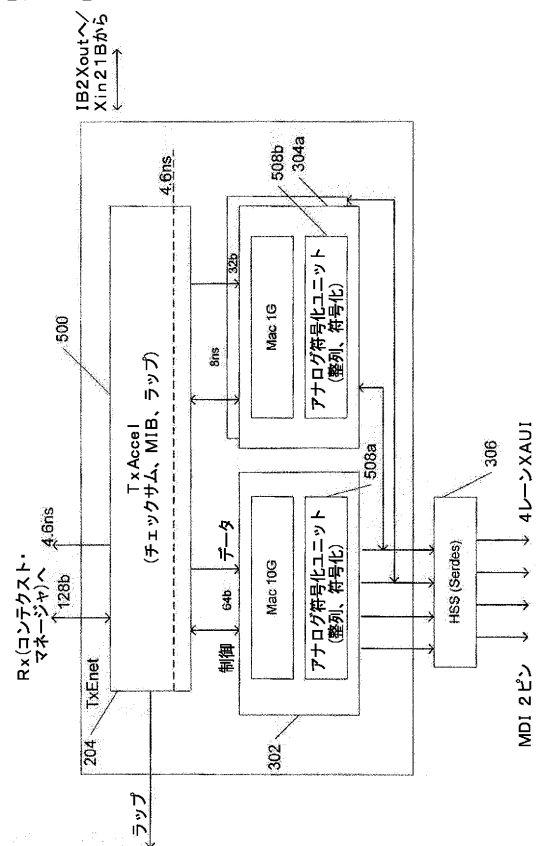
【図 4】



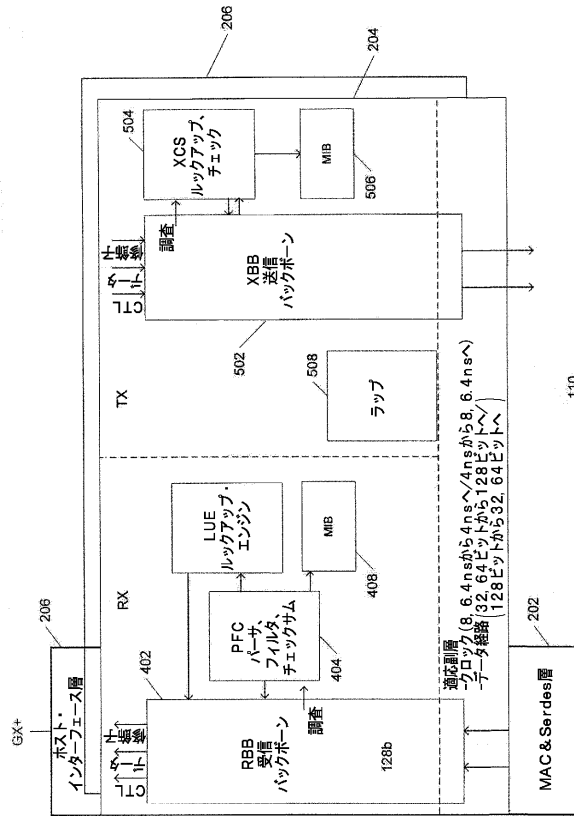
【図 5】



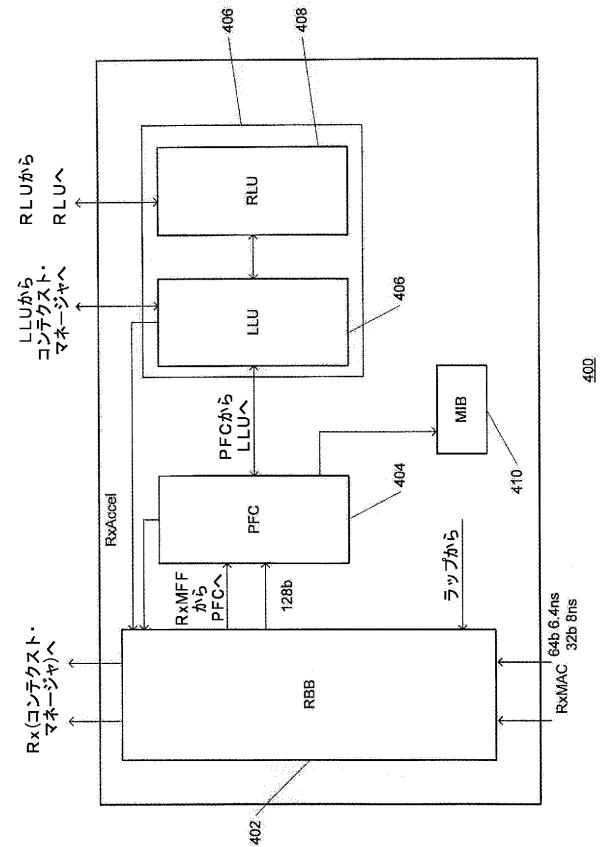
【図 6】



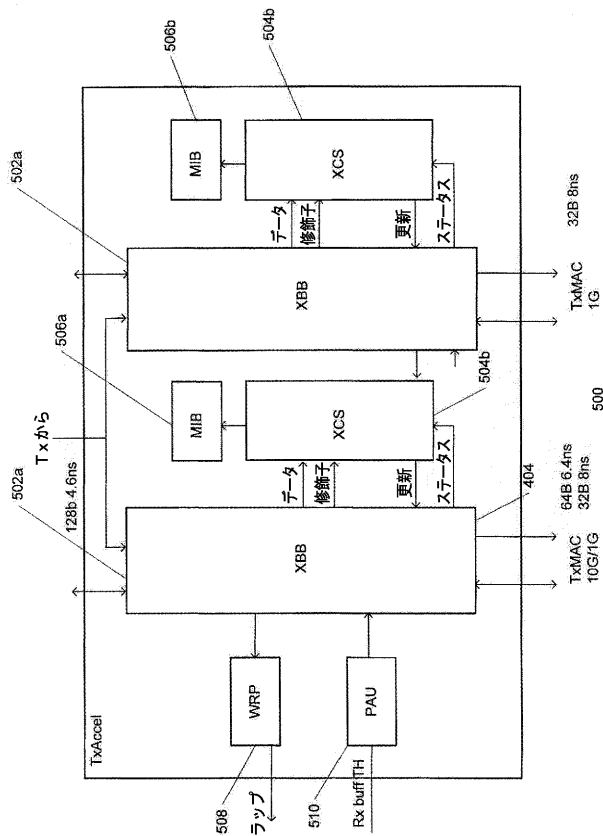
【図 7】



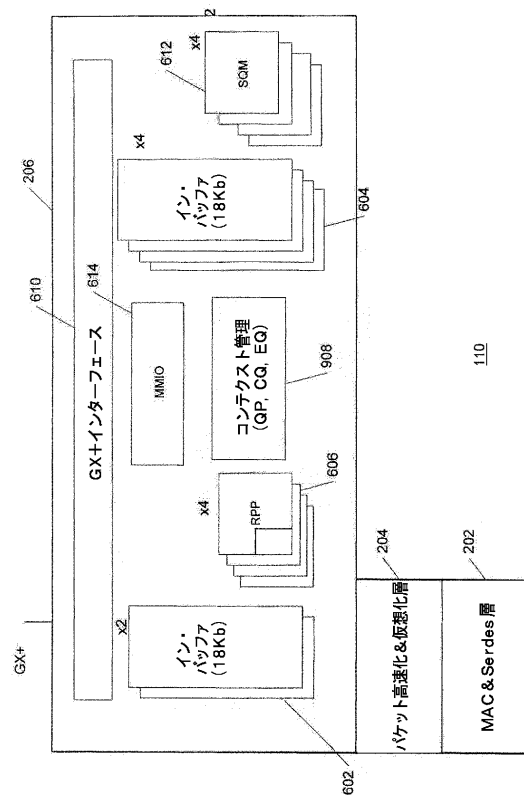
【図 8】



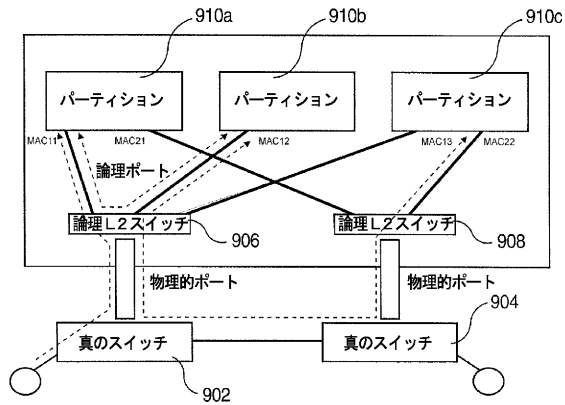
【図 9】



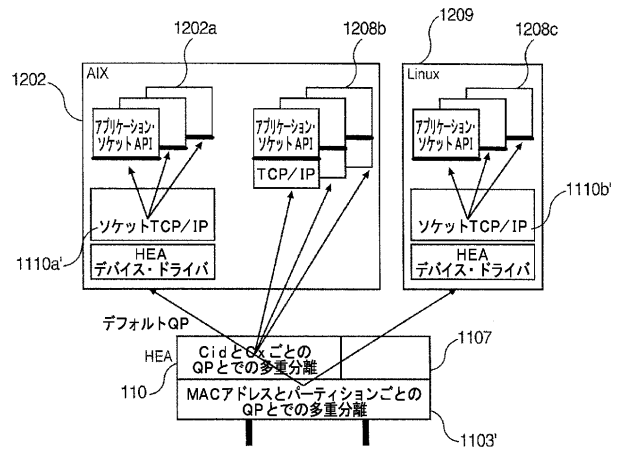
【図 10】



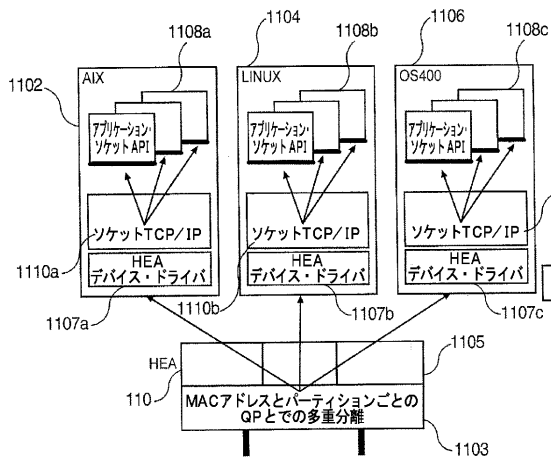
【図 1 1】



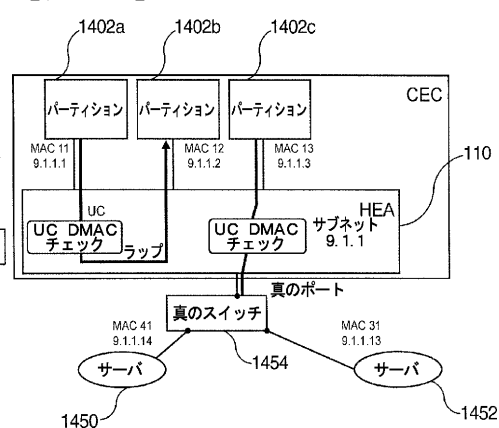
【図 1 3】



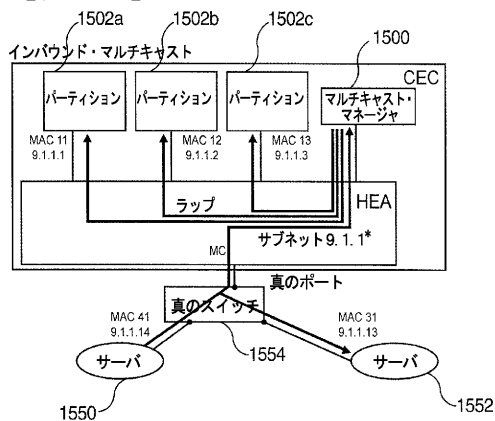
【図 1 2】



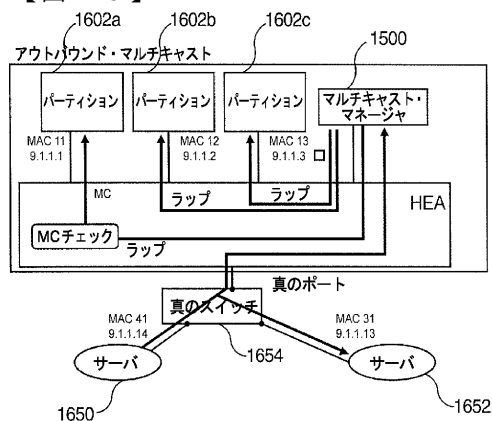
【図 1 4】



【図 1 5】



【図 1 6】



フロントページの続き

- (74)代理人 100086243
弁理士 坂口 博
- (72)発明者 アリミリ、ラビ、クマール
アメリカ合衆国 7 8 7 5 9 テキサス州 オースティン スパイスブラッシュ・ドライブ 9 2 2
1
- (72)発明者 カルビニャック、ジャン
アメリカ合衆国 2 7 6 0 6 ノースキャロライナ州 ラーレイ アトキンス・ファーム・コート
6 0 0 1
- (72)発明者 バッソ、クロード
アメリカ合衆国 2 7 6 1 3 ノースキャロライナ州 ラーレイ パーシー・コート 7 6 0 4
- (72)発明者 チャン、チージェン
アメリカ合衆国 2 7 5 2 3 ノースキャロライナ州 エイベックス ウォルドン・メドウ・ドライ
ブ 1 8 0 7
- (72)発明者 デイモン、フィリップ
アメリカ合衆国 2 7 5 1 6 ノースキャロライナ州 チャペル・ヒル マウンテン・クリーク・ロ
ード 8 1 0
- (72)発明者 フース、ロナルド、エドワード
アメリカ合衆国 5 5 9 0 1 ミネソタ州 ロチェスター シャレー・ドライブ エヌ・ダブリュー
9 6 0
- (72)発明者 シャルマ、サティア、プラカシュ
アメリカ合衆国 7 8 7 5 9 テキサス州 オースティン スワン・バレー・レーン 1 0 1 0 9
- (72)発明者 バイディアナサン、ナタラヤン
アメリカ合衆国 2 7 5 1 0 ノースキャロライナ州 カッロボロ メイナー・リッジ・ドライブ
2 0 9
- (72)発明者 フェルブランケン、ファブリス
フランス共和国 0 6 6 1 0 ラ・ゴード ルート・デ・カーニュ 9 1 5 2
- (72)発明者 フェリーリ、コリン、ビートン
アメリカ合衆国 2 7 5 0 2 ノースキャロライナ州 エイベックス ハゼルハースト・サークル
7 5 5 3

審査官 田畑 利幸

- (56)参考文献 米国特許出願公開第 2 0 0 5 / 0 0 5 3 0 6 0 (US, A1)
特表 2 0 0 4 - 5 2 3 0 3 5 (JP, A)
特開平 0 8 - 0 0 6 8 8 2 (JP, A)
特表 2 0 0 5 - 5 0 6 7 2 6 (JP, A)

(58)調査した分野(Int.Cl., DB名)

H04L 12/28
H04L 29/10
G06F 9/54
G06F 13/00