(19)日本国特許庁(JP)　(12)公表特許公報(A)　(11)特許出願公表番号
特表2017-538201
(P2017-538201A)
(43)公表日　平成29年12月21日(2017.12.21)

(51)Int.Cl.		F I		テーマコード（参考）
G06F 13/10	(2006.01)	G06F	13/10	330C
G06F 9/50	(2006.01)	G06F	9/46	465Z
G06F 9/46	(2006.01)	G06F	9/46	350

審査請求 未請求　予備審査請求 未請求　（全 25 頁）

(21)出願番号　特願2017-522887 (P2017-522887)
(86)(22)出願日　平成27年10月28日 (2015.10.28)
(85)翻訳文提出日　平成29年4月27日 (2017.4.27)
(86)国際出願番号　PCT/US2015/057860
(87)国際公開番号　WO2016/069773
(87)国際公開日　平成28年5月6日 (2016.5.6)
(31)優先権主張番号　62/072,847
(32)優先日　平成26年10月30日 (2014.10.30)
(33)優先権主張国　米国 (US)
(31)優先権主張番号　62/075,000
(32)優先日　平成26年11月4日 (2014.11.4)
(33)優先権主張国　米国 (US)
(31)優先権主張番号　62/076,336
(32)優先日　平成26年11月6日 (2014.11.6)
(33)優先権主張国　米国 (US)

(71)出願人　502303739
オラクル・インターナショナル・コーポレイション
アメリカ合衆国カリフォルニア州９４０６５レッドウッド・シティー，オラクル・パークウェイ５００
(74)代理人　110001195
特許業務法人深見特許事務所
(72)発明者　タソウラス，エバンジェロス
ノルウェー、エヌ－１３２５　リュサケール、ピィ・オゥ・ボックス・１３４
(72)発明者　ヨンセン，ビョルン・ダグ
ノルウェー、エヌ－０６８７　オスロ、ビルベルクグレンダ、９

最終頁に続く

(54)【発明の名称】ダイナミッククラウドにサブネット管理（ＳＡ）クエリキャッシングを提供するためのシステムおよび方法

(57)【要約】
システムおよび方法はクラウド環境においてサブネット管理をサポートすることができる。クラウド環境における仮想マシンマイグレーション中に、サブネットマネージャは効率的なサービスを遅延させる障害ポイントになる可能性がある。システムおよび方法は、マイグレーション後に仮想マシンに複数のアドレスを確実に保持させることによって、この障害ポイントを軽減することができる。当該システムおよび方法はさらに、マイグレートされた仮想マシンとの通信を回復させるときに、仮想マシンが利用することができるローカルキャッシュに、クラウド環境内の各々のホストノードを関連付けることを可能にし得る。

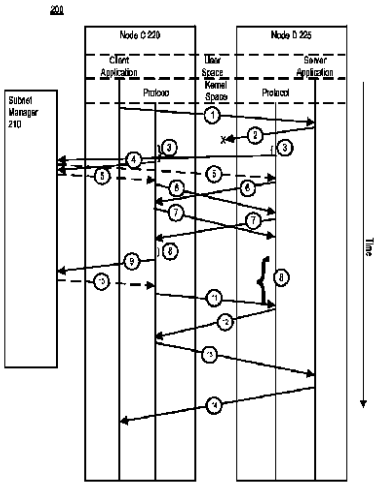

FIGURE 3

【特許請求の範囲】
【請求項１】
クラウド環境においてサブネット管理をサポートするための方法であって、
少なくとも第１のハイパーバイザおよび第１のホストチャネルアダプタに関連付けられた第１のホストノードを含む複数のホストノードを前記クラウド環境内に設けるステップと、
複数のアドレスに関連付けられた第１の仮想マシンを前記第１のホストノード上に設けるステップと、
前記第１の仮想マシンを、前記クラウド環境内の前記複数のホストノードのうち前記第１のホストノードから、設けられた第２のホストノードにまでマイグレートするステップとを含み、前記第２のホストノードは、少なくとも第２のハイパーバイザおよび第２のホストチャネルアダプタに関連付けられており、
前記複数のホストノードの各々はローカルキャッシュを含み、各々のローカルキャッシュは１つ以上のパス記録を含む、方法。
【請求項２】
前記第１の仮想マシンを、前記クラウド環境内の前記第１のホストノードから前記設けられた第２のホストノードにまでマイグレートする前記ステップは、
前記第１のハイパーバイザから前記第１の仮想マシンを分離するステップを含み、前記第１のハイパーバイザから前記第１の仮想マシンを分離する前記ステップは、前記第１の仮想マシンに関連付けられた第１の仮想機能を前記第１の仮想マシンから分離するステップを含み、マイグレートする前記ステップはさらに、
前記第１の仮想マシンに関連付けられた前記複数のアドレスを前記第２のホストノードに与えるステップと、
前記第２のハイパーバイザに関連付けられた第２の仮想機能に前記複数のアドレスを割当てるステップと、
前記第１の仮想マシンを、前記第１のホストノードから前記第２のホストノード上の第２の仮想マシンにまでマイグレートするステップと、
前記第１の仮想マシンに関連付けられた前記複数のアドレスに前記第２の仮想マシンをエクスポーズするステップとを含む、請求項１に記載の方法。
【請求項３】
前記クラウド環境内において、前記第１の仮想マシンを前記第１のホストノードから前記設けられた第２のホストノードにまでマイグレートするステップの後、前記第２の仮想マシンと第３の仮想マシンとの間に通信を確立するステップをさらに含み、前記第３の仮想マシンは、前記複数のホストノードのうち第３のホストノード上に設けられ、
前記クラウド環境内において前記第１の仮想マシンを前記第１のホストノードから前記設けられた第２のホストノードにまでマイグレートするステップの前に、前記第１の仮想マシンと前記第３の仮想マシンとが通信していた、請求項２に記載の方法。
【請求項４】
前記第２の仮想マシンと第３の仮想マシンとの間に通信を確立する前記ステップは、
前記第３のホストノードに関連付けられたローカルキャッシュ内に、少なくとも前記第１の仮想マシンに関連付けられた前記複数のアドレスを含む第１のパス記録を記憶するステップと、
前記第１の仮想マシンが前記第１のホストノードから前記設けられた第２のホストノードにまでマイグレートするとき、前記第３のホストノードによって、通信における途切れを検出するステップと、
前記第３のホストノードに関連付けられた前記ローカルキャッシュにおける前記第１のパス記録を前記第３のホストノードによって検索するステップと、
前記少なくとも第１のパス記録に基づいて、前記第２の仮想マシンと第３の仮想マシンとの間に通信を確立するステップとを含む、請求項３に記載の方法。
【請求項５】

前記第１のパス記録は、前記第１の仮想マシンに関連付けられるとともに前記第３のホストノードによってサブネットマネージャに送信される前記複数のアドレスに関するクエリに基づいて作成され、前記サブネットマネージャは前記クラウド環境に関連付けられている、請求項４に記載の方法。

【請求項６】

前記第１のパス記録が作成された後、前記第１の仮想マシンに関連付けられた前記複数のアドレスに関するクエリはいずれも、それ以上、前記第３のホストノードによって前記サブネットマネージャに送信されない、請求項５に記載の方法。

【請求項７】

前記第１の仮想マシンに関連付けられるとともに前記第３のホストノードによって前記サブネットマネージャに送信される前記複数のアドレスに関する前記クエリに応じて、前記サブネットマネージャはマーク付けされたパス記録を戻し、前記マーク付けされたパス記録はパス・キャッシング可能マークを含み、前記パス・キャッシング可能マークは、前記第１のパス記録が通信における前記途切れの間も持続することを示している、請求項５または６に記載の方法。

【請求項８】

前記複数のホストノードのうち別のホスト上に設けられた別の仮想マシンに関連付けられた別の複数のアドレスに関する、前記サブネットマネージャに対する別のクエリに応じて、前記サブネットマネージャはマーク付けされた別のパス記録を戻し、前記マーク付けされた別のパス記録はパス・キャッシング可能マークを含み、前記パス・キャッシング可能マークは、前記別のパス記録が通信における別の途切れの間も持続することを示している、請求項７に記載の方法。

【請求項９】

前記第２の仮想マシンと前記第３の仮想マシンとの間の前記通信はインフィニバンド・プロトコルに基づいている、請求項３から８のいずれかに記載の方法。

【請求項１０】

前記クラウド環境内において前記第１の仮想マシンを前記第１のホストノードから前記設けられた第２のホストノードにまでマイグレートするステップの後、前記第２の仮想マシンと第３のエンティティとの間に通信を確立するステップをさらに含み、前記第３のエンティティは、物理的なホスト、記憶装置、または、マイグレートされた第１の仮想マシンと前記インフィニバンド・プロトコルによって以前から通信している別のエンティティのうちの１つであり、さらに、

前記第３のエンティティに関連付けられたローカルキャッシュ内に第１のパス記録を記憶するステップを含み、前記第１のパス記録は、少なくとも、前記第１の仮想マシンに関連付けられた前記複数のアドレスを含み、さらに、

前記第１の仮想マシンが前記第１のホストノードから前記設けられた第２のホストノードにまでマイグレートするときに、前記第３のエンティティによって、通信における途切れを検出するステップと、

前記第３のエンティティに関連付けられた前記ローカルキャッシュにおける前記第１のパス記録を前記第３のエンティティによって検索するステップと、

少なくとも前記第１のパス記録に基づいて、前記第２の仮想マシンと前記第３のエンティティとの間に通信を確立するステップとを含み、

前記クラウド環境内において前記第１の仮想マシンを前記第１のホストノードから前記設けられた第２のホストノードにまでマイグレートするステップの前に、前記第１の仮想マシンと前記第３のエンティティとが通信していた、請求項２に記載の方法。

【請求項１１】

クラウド環境においてサブネット管理をサポートするシステムであって、

１つ以上のマイクロプロセッサと、

前記１つ以上のマイクロプロセッサ上で実行されるプロセッサとを含み、前記プロセッサは、

少なくとも第１のハイパーバイザおよび第１のホストチャネルアダプタに関連付けられた第１のホストノードを含む複数のホストノードを前記クラウド環境内に設けるステップと、

複数のアドレスに関連付けられた第１の仮想マシンを前記第１のホストノード上に設けるステップと、

前記第１の仮想マシンを、前記クラウド環境内の前記複数のホストノードのうち前記第１のホストノードから、設けられた第２のホストノードにまでマイグレートするステップとを含む複数のステップを実行するように動作し、前記第２のホストノードは、少なくとも第２のハイパーバイザおよび第２のホストチャネルアダプタに関連付けられており、

前記複数のホストノードの各々はローカルキャッシュを含み、各々のローカルキャッシュは１つ以上のパス記録を含む、システム。

【請求項１２】

前記プロセッサはさらに、

前記第１のハイパーバイザから前記第１の仮想マシンを分離するステップを実行するように動作し、前記第１のハイパーバイザから前記第１の仮想マシンを分離する前記ステップは、前記第１の仮想マシンに関連付けられた第１の仮想機能を前記第１の仮想マシンから分離するステップを含み、前記プロセッサはさらに、

前記第１の仮想マシンに関連付けられた前記複数のアドレスを前記第２のホストノードに与えるステップと、

前記複数のアドレスを、前記第２のハイパーバイザに関連付けられた第２の仮想機能に割当てるステップと、

前記第１の仮想マシンを、前記第１のホストノードから前記第２のホストノード上の第２の仮想マシンにまでマイグレートするステップと、

前記第１の仮想マシンに関連付けられた前記複数のアドレスに前記第２の仮想マシンをエクスポーズするステップとを実行するように動作する、請求項１１に記載のシステム。

【請求項１３】

前記プロセッサはさらに、

前記クラウド環境内において、前記第１の仮想マシンを前記第１のホストノードから前記設けられた第２のホストノードにまでマイグレートするステップの後、前記第２の仮想マシンと前記複数のホストノードのうち第３のホストノード上に設けられた第３の仮想マシンとの間に通信を確立するステップを実行するように動作し、

前記クラウド環境内において前記第１の仮想マシンを前記第１のホストノードから前記設けられた第２のホストノードにまでマイグレートするステップの前に、前記第１の仮想マシンと前記第３の仮想マシンとが通信していた、請求項１２に記載のシステム。

【請求項１４】

前記プロセッサはさらに、

前記第３のホストノードに関連付けられたローカルキャッシュ内に第１のパス記録を記憶するステップを実行するように動作し、前記第１のパス記録は、少なくとも前記第１の仮想マシンに関連付けられた前記複数のアドレスを含み、前記プロセッサはさらに、

前記第１の仮想マシンが前記第１のホストノードから前記設けられた第２のホストノードにまでマイグレートするときに、前記第３のホストノードによって、通信における途切れを検出するステップと、

前記第３のホストノードに関連付けられた前記ローカルキャッシュにおける前記第１のパス記録を前記第３のホストノードによって検索するステップと、

少なくとも前記第１のパス記録に基づいて、前記第２の仮想マシンと第３の仮想マシンとの間に通信を確立するステップとを実行するように動作する、請求項１３に記載のシステム。

【請求項１５】

前記第１のパス記録は、前記第１の仮想マシンに関連付けられるとともに前記第３のホストノードによってサブネットマネージャに送信される前記複数のアドレスに関するクエ

リに基づいて作成され、前記サブネットマネージャは前記クラウド環境に関連付けられている、請求項１４に記載のシステム。
【請求項１６】
前記第１のパス記録が作成された後、前記第１の仮想マシンに関連付けられた前記複数のアドレスに関するクエリはいずれも、それ以上、前記第３のホストノードによって前記サブネットマネージャに送信されない、請求項１５に記載のシステム。
【請求項１７】
前記第２の仮想マシンと前記第３の仮想マシンとの間の前記通信はインフィニバンド・プロトコルに基づいている、請求項１３から１６のいずれかに記載のシステム。
【請求項１８】
クラウド環境においてサブネット管理をサポートするための命令が格納されている非一時的な機械読取可能な記憶媒体であって、前記命令は、実行されると、システムに、
少なくとも第１のハイパーバイザおよび第１のホストチャネルアダプタに関連付けられた第１のホストノードを含む複数のホストノードを前記クラウド環境内に設けるステップと、
複数のアドレスに関連付けられた第１の仮想マシンを前記第１のホストノード上に設けるステップと、
前記クラウド環境内において、前記第１の仮想マシンを、前記複数のホストノードのうちの前記第１のホストノードから、設けられた第２のホストノードにまでマイグレートするステップとを含む複数のステップを実行させ、前記第２のホストノードは、少なくとも第２のハイパーバイザおよび第２のホストチャネルアダプタに関連付けられており、
前記複数のホストノードの各々はローカルキャッシュを含み、各々のローカルキャッシュは１つ以上のパス記録を含む、非一時的な機械読取可能な記憶媒体。
【請求項１９】
前記複数のステップはさらに、
前記第１のハイパーバイザから前記第１の仮想マシンを分離するステップを含み、前記第１のハイパーバイザから前記第１の仮想マシンを分離する前記ステップは、前記第１の仮想マシンに関連付けられた第１の仮想機能を前記第１の仮想マシンから分離するステップを含み、前記複数のステップはさらに、
前記第１の仮想マシンに関連付けられた前記複数のアドレスを前記第２のホストノードに与えるステップと、
前記第２のハイパーバイザに関連付けられた第２の仮想機能に前記複数のアドレスを割当てるステップと、
前記第１の仮想マシンを、前記第１のホストノードから前記第２のホストノード上の第２の仮想マシンにまでマイグレートするステップと、
前記第２の仮想マシンを、前記第１の仮想マシンに関連付けられた前記複数のアドレスにエクスポーズするステップとを含む、請求項１８に記載の非一時的な機械読取可能な記憶媒体。
【請求項２０】
前記複数のステップはさらに、
前記クラウド環境内において、前記第１の仮想マシンを、前記第１のホストノードから前記設けられた第２のホストノードにまでマイグレートするステップの後、前記第２の仮想マシンと第３の仮想マシンとの間に通信を確立するステップを含み、前記第３の仮想マシンは、前記複数のホストノードのうち第３のホストノード上に設けられ、
前記クラウド環境内において前記第１の仮想マシンを前記第１のホストノードから前記設けられた第２のホストノードにまでマイグレートするステップの前に、前記第１の仮想マシンと前記第３の仮想マシンとが通信していた、請求項１９に記載の非一時的な機械読取可能な記憶媒体。
【請求項２１】
前記複数のステップはさらに、

前記第３のホストノードに関連付けられたローカルキャッシュ内に第１のパス記録を記憶するステップを含み、前記第１のパス記録は、少なくとも前記第１の仮想マシンに関連付けられた前記複数のアドレスを含み、前記複数のステップはさらに、

前記第１の仮想マシンが前記第１のホストノードから前記設けられた第２のホストノードにまでマイグレートするときに、前記第３のホストノードによって、通信における途切れを検出するステップと、

前記第３のホストノードに関連付けられた前記ローカルキャッシュにおける前記第１のパス記録を前記第３のホストノードによって検索するステップと、

少なくとも前記第１のパス記録に基づいて、前記第２の仮想マシンと第３の仮想マシンとの間に通信を確立するステップとを含む、請求項２０に記載の非一時的な機械読取可能な記憶媒体。

【請求項２２】

前記第１のパス記録は、前記第１の仮想マシンに関連付けられるとともに前記第３のホストノードによってサブネットマネージャによって送信される前記複数のアドレスに関するクエリに基づいて作成され、前記サブネットマネージャは前記クラウド環境に関連付けられており、

前記第１のパス記録が作成された後、前記第１の仮想マシンに関連付けられた前記複数のアドレスに関するクエリはいずれも、それ以上、前記第３のホストノードによって前記サブネットマネージャに送信されない、請求項２１に記載の非一時的な機械読取可能な記憶媒体。

【発明の詳細な説明】

【技術分野】

【０００１】

著作権表示：

この特許文献の開示の一部は、著作権保護の対象となる題材を含んでいる。著作権の所有者は、特許商標庁の包袋または記録に掲載されるように特許文献または特許情報開示を誰でも複製できることに対して異議はないが、その他の点ではすべての如何なる著作権をも保有する。

【０００２】

発明の分野：

本発明は概してコンピュータシステムに関し、特にクラウド環境に関する。

【背景技術】

【０００３】

背景：

インフィニバンド（InfiniBand）サブネットにおいては、サブネットマネージャ（Subnet Manager：ＳＭ）が潜在的な障害となる。インフィニバンド・サブネットのサイズが大きくなると、ホスト間のパスの数が多項式的に増加し、多くの同時発生的なパス決定の要求を受取ったときにＳＭがタイムリーにネットワークのために機能することができなくなるかもしれない。このスケーラビリティの問題は、ダイナミック仮想化クラウド環境においてさらに増幅される。インフィニバンド相互接続された仮想マシン（Virtual Machine：ＶＭ）がライブマイグレートする（live migrate）と、ＶＭアドレスが変更される。これらのアドレス変更により、結果として、通信ピアが新しいパス特徴を解釈するためにサブネット管理（Subnet Administration：ＳＡ）パス記録クエリをＳＭに送信したときに、ＳＭに負荷が追加されることとなる。

【発明の概要】

【課題を解決するための手段】

【０００４】

概要：

システムおよび方法は、クラウド環境においてサブネット管理をサポートすることができる。クラウド環境における仮想マシンマイグレーション（migration）中に、サブネッ

トマネージャは効率的なサービスを遅延させる障害ポイントになる可能性がある。システムおよび方法は、マイグレーション後に仮想マシンに確実に複数のアドレスを保持させることによって、この障害ポイントを軽減させることができる。当該システムおよび方法は、マイグレートされた仮想マシンとの通信を回復させるときに、クラウド環境内の各々のホストノードを、仮想マシンが利用できるローカルキャッシュに関連付けることをさらに可能にし得る。

【図面の簡単な説明】

【０００５】

【図１】一実施形態に従った、クラウド環境におけるＶＭライブマイグレーションをサポートする例を示す図である。

【図２】一実施形態に従った、２つのホスト間の接続を確立するプロトコルの例を示す図である。

【図３】一実施形態に従った、接続の途絶時に２つのノード間で継続中の通信の例を示す図である。

【図４】一実施形態に従った、クラウド環境においてＳＡパス・キャッシングをサポートする例を示す図である。

【図５】本発明の一実施形態に従った、クラウド環境においてＳＡパス・キャッシングをサポートする例を示す図である。

【図６】一実施形態に従った、クラウド環境においてサブネット管理をサポートするための方法の例を示す図である。

【発明を実施するための形態】

【０００６】

詳細な説明：

本発明は、同様の参照符号で同様の要素を示している添付の図面において、限定によってではなく例示によって説明されている。この開示において「ある」または「１つの」または「いくつかの」実施形態を言及している場合、このような言及は必ずしも同じ実施形態に対するものではなく、「少なくとも１つ」という意味である。

【０００７】

本発明の以下の記載は、高性能ネットワークプロトコルについての一例としてインフィニバンド（ＩＢ）ネットワークプロトコルを用いている。他のタイプの高性能ネットワークプロトコルが制限なしに使用可能であることは当業者にとって明らかであるだろう。

【０００８】

クラウド環境において仮想マシン（virtual machine：ＶＭ）マイグレーション・サブネット管理（subnet administration：ＳＡ）パス・キャッシングをサポートすることができるシステムおよび方法がこの明細書中に記載される。

【０００９】

一実施形態に従うと、高性能コンピューティング（High Performance Computing：ＨＰＣ）を提供することができるクラウドコンピューティングを提供することができる。ＨＰＣ・アズ・ア・サービス（HPC-as-a-service）をコンピューティングクラウド内に提供することができ、このＨＰＣ・アズ・ア・サービスは仮想ＨＰＣ（virtual HPC：ｖＨＰＣ）クラスタを考慮に入れたものであって、高性能相互接続ソリューションを用いてこの仮想ＨＰＣクラスタに対応し得る。

【００１０】

一実施形態に従うと、各々のＩＢサブネットはサブネットマネージャ（ＳＭ）を利用することができる。各々のＳＭは、ネットワーク初期化、トポロジー発見、パス演算、ならびに、スイッチおよびホストチャネルアダプタ（Host Channel Adapter：ＨＣＡ）上のＩＢポートの構成のために機能し得る。大規模なサブネットでは、ノード間で利用可能なパスが多項式的に大きくなる可能性があり、パス決定のための多くの同時発生的な要求を受取ったときに、ＳＭが潜在的な障害になる可能性がある。このスケーラビリティの問題は、ダイナミックな仮想化クラウド環境においては、ＩＢ相互接続されている仮想マシンが

ライブマイグレートしたときにさらに増幅される。

【００１１】

効率的な仮想化をサポートするために、高帯域および低レイテンシを維持しながらも、ＩＢホストチャネルアダプタ（ＨＣＡ）はシングルルートＩ／Ｏ仮想化（Single Root I/O Virtualization：ＳＲ－ＩＯＶ）をサポートすることができる。ＩＢ接続されたノードは各々、３つのさまざまなアドレスを有する。ライブマイグレーションが発生すると、パススルーされたインターフェイス（the passed through interface）を取外すことでダウンタイムが発生するにもかかわらず、ＩＢアドレスのうち１つ以上を変更することができる。マイグレーション中のＶＭと通信する他のノードは接続性を失い、サブネット管理（ＳＡ）パス記録クエリをＳＭに送信することによって再接続するための新しいアドレスを発見しようと試みる。基盤ネットワークにおいて結果として生じるＳＭ宛ての通信は有効になり得る。大規模なネットワークにおいては、ＳＭ宛てのメッセージがこのようにＶＭのマイグレートによって溢れかえると、ＳＭに対する負荷が増えるのに応じて、全体的なネットワークレイテンシが大きくなる可能性がある。

【００１２】

一実施形態に従うと、さらに、ＶＭマイグレーションによって引起こされてＳＭによって受取られるＳＡ要求の量を減らすことによって、ＳＭに対する負荷を減らすことが望ましい。方法およびシステムは、ＶＭがマイグレーション後にその同じアドレスを保持することができるシステムを実現することによって、これを達成することができる。加えて、２つのノード間の初期接続が確立された後に、ＳＡパス・キャッシングメカニズムを用いてＳＡクエリの数を激減させることができる。

【００１３】

一実施形態に従うと、インフィニバンドは一般に３つの異なるタイプのアドレスを用いている。まず、１６ビットのローカル識別子（Local Identifier：ＬＩＤ）が挙げられる。少なくとも１つのＬＩＤは、ＳＭによって各々のＨＣＡポートおよび各々のスイッチに割当てられている。ＬＩＤを用いて、サブネット内のトラフィックをルーティングすることができる。ＬＩＤが１６ビット長であるので、６５５３６個の固有のアドレスの組合せを構成することができ、そのうち４９１５１（０×０００１－０×ＢＦＦＦ）個だけをユニキャストアドレスとして用いることができる。結果として、利用可能なユニキャストアドレスの数は、ＩＢサブネットの最大サイズを定義することとなる。

【００１４】

第２のタイプのアドレスは、一般に、製造業者によって各々の装置（たとえば、ＨＣＡおよびスイッチ）ならびに各々のＨＣＡポートに割当てられる６４ビットのグローバル固有識別子（Global Unique Identifier：ＧＵＩＤ）である。ＳＭは、ＳＲ－ＩＯＶ　ＶＦが使用可能にされたときに有用になり得る追加のサブネット固有のＧＵＩＤをＨＣＡポートに割当て得る。

【００１５】

第３のタイプのアドレスは１２８ビットのグローバル識別子（Global Identifier：ＧＩＤ）である。ＧＩＤは、一般に、有効なＩＰｖ６ユニキャストアドレスであり、少なくとも１つが各々のＨＣＡポートおよび各々のスイッチに割当てられている。ＧＩＤは、ファブリックアドミニストレータによって割当てられたグローバルに固有の６４ビットプレフィックスと各々のＨＣＡポートのＧＵＩＤアドレスとを組合わせることによって形成される。

【００１６】

本発明の以下の記載は、高性能ネットワークについての一例としてインフィニバンドネットワークを用いる。他のタイプの高性能ネットワークを制限なしで使用できることは当業者にとって明らかであるだろう。また、本発明の以下の記載は、仮想化モデルについての一例としてＫＶＭ仮想化モデルを用いている。他のタイプの仮想化モデル（たとえばＸｅｎ）を制限なしに使用できることは当業者にとっては明らかであるだろう。

【００１７】

本発明の以下の記載は、加えて、ＯｐｅｎＳｔａｃｋ、ＯｐｅｎＳＭおよびＲＤＳ　Ｌｉｎｕｘ（登録商標）カーネルモジュールを利用する。ＯｐｅｎＳｔａｃｋは、データセンタを通じて処理、記憶およびネットワークリソースのプールを制御する相互に関係付けられたプロジェクトのグループを含むクラウド・コンピューティング・ソフトウェアプラットホームである。ＯｐｅｎＳＭは、ＯｐｅｎＩＢの上で実行させることができるインフィニバンド対応のサブネットマネージャおよびアドミニストレーションである。リライアブル・データグラム・ソケット（Reliable Datagram Socket：ＲＤＳ）は、データグラムを提供するための高性能で低レイテンシかつ信頼性のある非接続プロトコルである。他の同様のプラットフォームを制限なしで利用できることは当業者にとって明らかであるだろう。

【００１８】

本発明の一実施形態に従うと、仮想化は、クラウドコンピューティングにおける効率的なリソース利用および融通性のあるリソース割当てに有益であり得る。ライブマイグレーションは、アプリケーションにトランスペアレントな態様で物理サーバ間で仮想マシン（ＶＭ）を移動させることによってリソース使用を最適化することを可能にする。これにより、シングルルートＩ／Ｏ仮想化（ＳＲ－ＩＯＶ）法を利用する仮想化は、ライブマイグレーションによって統合、リソースのオン・デマンド・プロビジョニングおよび融通性を可能にし得る。

【００１９】

ＩＢアーキテクチャは、直列なポイント・ツー・ポイントの全二重技術である。ＩＢネットワークはサブネットとも称することができ、この場合、サブネットは、スイッチおよびポイント・ツー・ポイントリンクを用いて相互接続された１セットのホストで構成されている。ＩＢサブネットは、サブネットにすべてのスイッチ、ルータおよびホストチャネルアダプタ（host channel adaptor：ＨＣＡ）の構成を含むネットワークを初期化して提供する役割を果たし得る少なくとも１つのサブネット・マネージャ（ＳＭ）を含み得る。

【００２０】

ＩＢは、リモートダイレクトメモリアクセス（remote direct memory access：ＲＤＭＡ）および従来の送信／受信セマンティックをともに提供するために、１組のリッチな伝送サービスをサポートする。用いられる伝送サービスに依存せずに、ＩＢ　ＨＣＡはキュー対（queue pair：ＱＰ）を用いて通信を行う。ＱＰは、通信設定中に作成され、提供されるＱＰ数、ＨＣＡポート、宛先ＬＩＤ、キューサイズおよび伝送サービスなどの１セットの初期属性を有し得る。ＨＣＡは多くのＱＰを処理することができ、各々のＱＰは、送信キュー（send queue：ＳＱ）および受信キュー（receive queue：ＲＱ）などの１対のキューからなる。通信に関与している各々のエンドノードにはこのような１つの対が存在している。送信キューは、リモートノードに転送されるべき作業要求を保持し、受信キューは、リモートノードから受取ったデータで何を行なうべきかについての情報を保持している。ＱＰに加えて、各々のＨＣＡは、１セットの送信キューおよび受信キューに関連付けられている１つ以上の完了キュー（completion queue：ＣＱ）を有している。ＣＱは、送信キューおよび受信キューに掲示された作業要求についての完了通知を保持する。たとえ通信の複雑さがユーザからは見えないようになっていたとしても、ＱＰ状態情報はＨＣＡに保持されている。

【００２１】

ネットワークＩ／Ｏ仮想化：

一実施形態に従うと、Ｉ／Ｏ仮想化（I/O virtualization：ＩＯＶ）を用いることにより、Ｉ／Ｏリソースを共有して、さまざまな仮想マシンからリソースへのアクセスの保護を提供することができる。ＩＯＶは、仮想マシンにエクスポーズされ得る論理装置をその物理的な実装から分離することができる。このような１タイプのＩＯＶとして、直接的な装置の割当てがある。

【００２２】

一実施形態に従うと、直接的な装置の割当ては、ＶＭ間で装置を共有せずにＩ／Ｏ装置

をＶＭに連結することを必要とし得る。直接的な割当て（または装置のパススルー）は最小限のオーバーヘッドで固有の性能に近いものを提供することができる。物理装置は、ハイパーバイザをバイパスしてＶＭに直接取付けられており、ゲストＯＳは未改良のドライバを用いることができる。不利な面としては、共有がなされないせいでスケーラビリティが制限されてしまう点であり、１枚の物理ネットワークカードが１つのＶＭと連結されている。

【００２３】

一実施形態に従うと、シングルルートＩＯＶ（Single Root IOV：ＳＲ－ＩＯＶ）は、ハードウェア仮想化によって、物理装置がその同じ装置の複数の独立した軽量のインスタンスとして現われることを可能にし得る。これらのインスタンスは、パススルー装置としてＶＭに割当てることができ、仮想機能（Virtual Function：ＶＦ）としてアクセスすることができる。ＳＲ－ＩＯＶは、純粋に直接的に割当てする際のスケーラビリティの問題を軽減する。

【００２４】

残念ながら、ＳＲ－ＩＯＶなどの直接的な装置の割当て技術によれば、実現されたシステムがデータセンタ最適化のためにトランスペアレントなライブマイグレーション（ＶＭマイグレーション）を使用する場合、クラウドプロバイダに対して問題が提起される可能性がある。ライブマイグレーションの本質は、仮想マシンのメモリ内容が遠隔のハイパーバイザにまでコピーされるという点にある。さらに、仮想マシンをソース・ハイパーバイザにおいて一時停止させると、その動作は、それがコピーされた宛先において再開される。基礎をなすシステムが（ＳＲ－ＩＯＶなどの）直接的な装置の割当てを利用する場合、ネットワークインターフェイスがハードウェアに接続されている場合にはそのネットワークインターフェイスの完全な内部状態をコピーすることができない。ＶＭに割当てられたＳＲ－ＩＯＶ　ＶＦが分離されると、ライブマイグレーションが実行されることとなり、その宛先において新しいＶＦが取付けられることとなる。

【００２５】

ＩＢ　ＶＦを用いるＶＭがライブマイグレートされている状況においては、ＶＭの３つのアドレスをすべて変更することにより、基礎をなすネットワークファブリックおよびＳＭに対する明らかな影響を取り入れることができる。ＶＭが異なるＬＩＤを有する異なる物理ホストに移されるので、ＬＩＤが変化する。宛先において異なるＶＦが取付けられると、ＳＭによってソースＶＦに割当てられている仮想ＧＵＩＤ（virtual GUID：ｖＧＵＩＤ）が同様に変化し得る。次に、ｖＧＵＩＤを用いてＧＩＤが形成されるので、ＧＩＤも変化するだろう。結果として、マイグレートされたＶＭが、突然、新しい１セットのアドレスに関連付けられ、マイグレートされたＶＭの通信ピアが、マイグレートされたＶＭとの失われた接続を回復しようとして、同時発生したＳＡパス記録クエリバーストをＳＭに送信し始めることができる。これらのクエリは、ＳＭに対する余分なオーバーヘッドと、副次的作用としての追加のダウンタイムとをもたらす可能性がある。マイグレートされたノードがネットワークにおける他の多くのノードと通信する場合、ＳＭが障害となり、全体的なネットワーク性能を阻害する可能性がある。

【００２６】

一実施形態に従うと、この明細書中に記載される方法およびシステムは、クラウドプロバイダに提示されているＳＲ－ＩＯＶなどの直接的な装置の割当て技術を用いて、仮想マシンのライブマイグレーションに関連付けられた問題を低減および／または排除することができる。当該方法およびシステムは、ＩＢ　ＶＦを用いるＶＭがライブマイグレートされている状況において提示される問題を克服することができる。

【００２７】

仮想マシン（ＶＭ）ライブマイグレーション

図１は、一実施形態に従った、クラウド環境におけるＶＭのライブマイグレーションをサポートする例を示す。図１に示されるように、インフィニバンド（ＩＢ）サブネット１００は、異なるハイパーバイザ１１１～１１３をサポートする複数のホストノードＡ１０

１～Ｃ１０３を含み得る。
【００２８】
加えて、各々のハイパーバイザ１１１～１１３は、さまざまな仮想マシン（ＶＭ）がその上で実行されることを可能にする。たとえば、ホストノードＡ１０１上のハイパーバイザ１１１はＶＭ　Ａ１０４をサポートすることができ、ホストノードＢ上のハイパーバイザ１１２はＶＭ　Ｂ１０５をサポートすることができる。ＶＭ　ＡおよびＶＭ　Ｂが実行されているノード同士は通信することができる。
【００２９】
さらに、ホストノードＡ１０１～Ｃ１０３の各々は１つ以上のホストチャネルアダプタ（ＨＣＡ）１１７～１１９に関連付けることができる。図１に示されるように、ホストノードＡ１０１上のＨＣＡ１１７は、ＶＭ　Ａ１０４によって使用され得るＱＰａ１０８などのキュー対（ＱＰ）を利用することができ、ホストノードＢ１０２上のＨＣＡ１１８は、ＶＭ　Ｂ１０５によって使用され得るＱＰｂ１０７を利用することができる。
【００３０】
本発明の一実施形態に従うと、入出力仮想化（input/output virtualization：ＩＯＶ）を用いて、ＶＭにＩ／Ｏリソースを提供し、複数のＶＭから共有のＩ／Ｏリソースへのアクセスの保護を提供することができる。ＩＯＶは論理装置を分離することができ、この論理装置は、その物理的実装からＶＭにエクスポーズされている（exposed）。たとえば、シングルルートＩ／Ｏ仮想化（ＳＲ－ＩＯＶ）は、ＩＢネットワークにわたる仮想化において高性能を達成するためのＩ／Ｏ仮想化アプローチである。
【００３１】
また、ＩＢサブネット１００は、ネットワーク初期化、スイッチおよびＨＣＡ上のＩＢポートの構成、トポロジー発見およびパス演算のために機能し得るサブネットマネージャ１１０を含み得る。
【００３２】
図１に示されるように、ＶＭ　Ｂ１０５は、（たとえば、ハイパーバイザ１１１上のＶＭ　Ａ１０５と通信しながら）ハイパーバイザ１１２からハイパーバイザ１１３へとマイグレートさせることができる。
【００３３】
マイグレーションの後、新しいＶＭ　Ｂ′１０６は、宛先ホストノードＣ１０３における新しい１セットのアドレスに対して突然エクスポーズされる可能性がある。さらに、ピアＶＭ（たとえば、ＶＭ　Ａ１０４）は、失われた接続を回復させようとしている間に、ＳＭ１１０にサブネット管理（ＳＡ）パス記録クエリを送信し始めることができる（ＶＭ　Ｂ′はまた、新しいホストノード上で実行されると、ＳＡパス要求をＳＭに送信することもできる）。これは以下の事実に起因している。すなわち、一般に、ＶＭ　Ｂがホストノードホストノードにマイグレートする場合などのようにＶＭがマイグレートすると、これに応じて、ＶＭのアドレス（ＬＩＤ、ＧＵＩＤ、ＧＩＤ）が変化する。というのも、これらＶＭのアドレス（ＬＩＤ、ＧＵＩＤ、ＧＩＤ）は、概して、ＳＲ－ＩＯＶを用いているときにはハードウェアに接続されているからである。サブネットマネージャへのこれらのＳＡパスクエリは、著しいダウンタイムと、インフィニバンドＳＭ１１０に対する余分なオーバーヘッドとをもたらす可能性がある。多くのマイグレーションが大規模なデータセンタにおいてかなり短い時間フレーム内で行なわれる場合、または、マイグレートされたノードがネットワークにおける他の多くのノードと通信している場合、ＳＭ１１０が障害になる可能性がある。なぜなら、このＳＭ１１０がタイムリーに応答することができないかもしれないからである。
【００３４】
本発明の一実施形態に従うと、当該システムは、ＶＭ　Ｂ１０４がマイグレートしてＩＢアドレス情報が変化した場合に、関与しているホストノードＡ１０１～Ｃ１０３によって生成されるＳＡクエリの量を減らすことができる。
【００３５】

図１に示されるように、当該システムは、最初に、たとえば、ＶＭ　Ｂ１０４から仮想機能（virtual function：ＶＦ）１１５を分離することによって、ハイパーバイザ１１２からＶＭ　Ｂ１０４を分離することができる。次いで、システムは、たとえば、これらのアドレスを、ホストノードＣ１０３上のハイパーバイザ１１３上において次に利用可能な仮想機能（すなわちＶＦ′１１６）に割当てることによって、ＶＭ　Ｂ１０４に関連付けられているアドレス情報１２０を宛先ホストノードＣ１０３に与えることができる。最後に、ＶＭ　Ｂ１０４がＶＭ　Ｂ′１０６としてハイパーバイザ１１３にマイグレートされた後、当該システムは、（たとえば、ＱＰｂ′１０９を介する）ピアＶＭとの通信を回復させるために、ＶＭ　Ｂ′１０６をアドレス情報１２０にエクスポーズすることができる。

【００３６】

これにより、宛先ホストノードＣ１０３へのマイグレーションの後、新しいＶＭ　Ｂ′１０６を元の１セットのアドレスにエクスポーズすることができ、ピアＶＭ　Ａ１０４が、ＳＡパス記録クエリをＳＭ１１０に送信する必要がなくなる。

【００３７】

一実施形態に従うと、システムは、ＩＢ　ＳＲ－ＩＯＶ　ＶＦが取付けられているＶＭのＶＭライブマイグレーションをサポートすることができる。リモートダイレクトメモリアクセス（ＲＤＭＡ）は、ＶＭのマイグレーション後に通信を回復させるために、リライアブル・データグラム・ソケット（ＲＤＳ）プロトコルなどのプロトコルによって利用することができる。

【００３８】

一実施形態に従うと、システムは、ＯｐｅｎＳｔａｃｋ、ＯｐｅｎＳＭおよびＲＤＳ　Ｌｉｎｕｘカーネルモジュールを利用することができる。加えて、ＬＩＤｔｒａｃｋｅｒと称され得るプログラムは、各々のＶＭに関連付けられたＩＢアドレスを常に把握しておくために用いることができ、マイグレーションプロセスを編成することができる。

【００３９】

一実施形態においては、プログラムは、ＯｐｅｎＳＭのオプションhonor_guid2lid_fileを使用可能にし得る。次いで、ＯｐｅｎＳＭによって生成されたファイルguid2lidは、プログラムによってパーズされ、昇順などの順序でＧＵＩＤによってソートされ得る。ＬＩＤは、１から開始してＧＵＩＤに割当てられる。ＧＵＩＤに割当てられた各々のＬＩＤは、物理ホストのためのベースＬＩＤとも称することができる。

【００４０】

一実施形態においては、ベースＬＩＤが割当てられると、ＩＢ対応のＯｐｅｎＳｔａｃｋ演算ノードの各々は、ＶＭを実行させるためにスキャンすることができる。実行されていることが判明した各々のＶＭには、４９１５１（最上位のユニキャストＬＩＤ）から開始して降順でＬＩＤが割当てられ得る。ＶＭに割当てられたこれらのＬＩＤは浮動ＬＩＤとも称され得る。

【００４１】

一実施形態においては、浮動ＬＩＤは、ＶＭが実行されているＯｐｅｎＳｔａｃｋ演算ノードにおけるベースＬＩＤと置換えることができる。ハイパーバイザはＬＩＤをＶＭと共有する。いくつかの実施形態においては、１つのＶＭをハイパーバイザごとに実行させることができ、他のＶＭがその時点で実行されていないハイパーバイザに対して、或るＶＭがマイグレートされ得る。他の実施形態においては、複数のＶＭをハイパーバイザ上で実行させることができ、他のＶＭが宛先ハイパーバイザ上でその時点で実行されているかどうかにかかわらず、或るＶＭを別のハイパーバイザにマイグレートすることができる。

【００４２】

一実施形態においては、VM_xのためのマイグレーションがＯｐｅｎＳｔａｃｋ　ＡＰＩなどのＡＰＩから指示されると、ＳＲ－ＩＯＶ　ＶＦをＶＭから分離することができる。装置の取外しが完了し、マイグレーションが進行中である場合、ＯｐｅｎＳｔａｃｋは、プログラムに対し、$Hypervisor_y$などの１つのハイパーバイザからＨｙｐｅ

ｒｖｉｓｏｒ$_{z}$などの宛先ハイパーバイザにまでＶＭ$_{x}$が移動していることを通知することができる。次いで、プログラムは、Ｈｙｐｅｒｖｉｓｏｒ$_{y}$のＬＩＤをそのベースＬＩＤに戻すことができ、Ｈｙｐｅｒｖｉｓｏｒ$_{z}$はＶＭ$_{x}$に関連付けられた浮動ＬＩＤを得ることができる。プログラムはまた、宛先ハイパーバイザであるＨｙｐｅｒｖｉｓｏｒ$_{z}$において、ＶＭ$_{x}$に関連付けられたｖＧＵＩＤを次に利用可能なＳＲ－ＩＯＶ　ＶＦに割当てることができる。マイグレーション中、ＶＭにはネットワーク接続性がない。

【００４３】

一実施形態に従うと、これらの変更は再始動によって加えることができる。さらに、マイグレーションが完了すると、ＯｐｅｎＳｔａｃｋは次に利用可能なＳＲ－ＩＯＶ　ＶＦをＨｙｐｅｒｖｉｓｏｒ$_{z}$上のＶＭ$_{x}$に追加することができ、ＶＭはそのネットワーク接続性を復帰させることができる。ＶＭは、マイグレーションの前に有していた同じＩＢアドレス（ＬＩＤ、ｖＧＵＩＤおよびＧＩＤ）にエクスポーズすることができる。ＶＭの視点からは、マイグレートするのに必要な時間だけＩＢアダプタが分離されており、かつ、アドレスが変化しなかったので同じＩＢアダプタが再び取付けられたように、見えている。

【００４４】

サブネット管理（ＳＡ）パス・キャッシング

一実施形態に従うと、エンドノードにおけるローカルのＳＡパス・キャッシングメカニズムは、初期接続が２つのノード間で確立された後にＳＡクエリを減らすかまたは排除することができる。キャッシングスキームは一般的なものであり得るとともに、使用可能にされると、ライブマイグレーションの実施の有無にかかわらず、ＳＭに対する負荷を軽減させることができる。

【００４５】

図２は、一実施形態に従った、２つのホスト間の接続を確立するプロトコルの例を示す。より特定的には、図２は、２つのホスト間の接続を確立するためにＲＤＳなどのプロトコルを用いることを例示している。

【００４６】

一実施形態に従うと、接続を確立する前に、ＩＰ・オーバー・ＩＢ（IP over IB：ＩＰｏＩＢ）をすべての通信ピアにおいて設定することができる。ＲＤＳなどのプロトコルは、特有のＩＢポートのＩＰｏＩＢアドレスを用いてポートのＧＩＤアドレスを決定することができる。ＧＩＤアドレスが決定された後には、プロトコルは、パス記録照合を実行してＩＢ通信を確立するのに十分な情報を有することができる。

【００４７】

図２に示されるように、インフィニバンド・サブネット２００内では、サブネットマネージャ２１０は、ノードＣ２２０とノードＤ２２５との間、より特定的には、ノードＣ上のクライアント側アプリケーションとノードＤ上のサーバ側アプリケーションとの間、にパス通信を提供することができる。図２においては、上位層アプリケーションのクライアント側がノードＣにおいて実行され、アプリケーションのサーバ側はノードＤにおいて実行される。アプリケーションのクライアント側は、ＲＤＳソケットなどのソケットを作成して、アプリケーションのサーバ側と通信するよう試みる（ステップ１）ことができる。ＲＤＳなどのプロトコルは、ＳＡパス記録要求をノードＣからＳＭに送信する（ステップ２）ことができる。サブネットマネージャは、プロトコルに応答を提供する（ステップ３）ことができる。この応答は、クライアント側アプリケーションのターゲットについてのアドレス情報を含み得る。サブネットマネージャから応答を受取った後、プロトコルは、接続要求を送信することによってノードＤとの接続を開始するよう試みる（ステップ４）ことができる。接続が成功すれば、プロトコルは、たとえば、両側でのＲＤＭＡ＿ＣＭ＿ＥＶＥＮＴ＿ＥＳＴＡＢＬＩＳＨＥＤイベントによって通信チャネルを確立する（ステップ５）ことができる。この時点において、上位層アプリケーションが通信可能となる（ステップ６）。

【００４８】

初期接続において何かエラーがあれば、クライアント側（ノードＣ）のプロトコルは、ランダムなバックオフメカニズムとの接続を確立するために再試行しようとすることができる。サーバは、クライアントが通信しようと意図していることにまだ気づいていない。接続が確率された後に何か不具合があれば、ＲＤＳ側（アプリケーション視点からはクライアントおよびサーバ）がともに積極的にピアと再接続しようとするだろう。接続プロセスにおけるランダムなバックオフメカニズムは、両側が接続に関与している場合には、競合状態を回避するのに有用である。

【００４９】

図３は、一実施形態に従った、接続が途絶えた場合における２つのノード間の継続中の通信の例を示す。

【００５０】

図３においては、インフィニバンド・サブネット２００内では、サブネットマネージャ２１０は、ノードＣ２２０とノードＤ２２５との間にパス通信を提供することができ、接続が途絶えた（ステップ２）場合にノードＣとノードＤとの間に通信を継続させている（ステップ１）。接続の途絶えは、たとえば、ノード上で実行されているアプリケーションのうち１つのアプリケーションのライブマイグレーションに付随して起こる可能性がある。両方のプロトコルエンドは、接続がダウンしていると判断することができ、再接続しようと試みる前にいくらかのランダム時間（すなわちバックオフ時間）だけ待機する（ステップ３）ことができる。再接続を試みる前に両側が待機する時間は、図３に示されるように、同じであってもよく、または異なっていてもよい。ノードは、ＳＭにＳＡパス記録要求を送信する（ステップ４）ことによって再接続しようと試みることができる。ＳＡパス記録応答が受取られた（ステップ５）後、接続要求を送信する（ステップ６）ことができる。

【００５１】

図３に示される場合においては、ステップ３において２つのノードによって選択されたバックオフ時間はほとんど同じであった。このため、ノードＤがノードＣよりもわずかに速くＳＡパス記録応答を得て、ステップ６において最初に接続を開始しようとしていたとしても、ノードＣが接続要求自体を送信するより以前には接続要求はノードＣに到達していなかった。この場合、両方のプロトコルエンドは未処理の接続要求を有している。次いで、ノードがそれらのピアから接続要求を受取ると、これらのノードは接続要求を拒否する（ステップ７）だろう。ステップ８において、２つのノードは、再接続しようと再試行する前に、ランダムなバックオフ時間をもう一度選択した。このとき、ノードＤによって選択されたランダムなバックオフ時間はノードＣによって選択されたものより著しく長い。結果として、ノードＣが優先されることとなり、このノードＣが、接続確立プロセスを繰返し、ＳＡパス記録要求を送信し（ステップ８）、サブネットマネージャから応答を受取り（ステップ１０）、ノードＤに接続要求を送信し（ステップ１１）、そして、ノードＤがノードＣとの接続自体を開始しようと試みる前に接続要求がノードＤに到達する。図３に示される状況においては、ノードＤが入接続を受付ける（ステップ１２）。次いで、ステップ１３および１４において、上位層アプリケーションのために通信を再開することができる。

【００５２】

図３から推測すると、サブネットマネージャがＶＭマイグレーションの場合にＳＡパス要求を大量に受取る可能性がある（通信の遮断）ことが明らかになる。何千ものノードを有する大規模なサブネットにおいては、各々のノードから追加のＳＡクエリが１つだけ送信された場合でも、ＳＭには、最終的に何千ものメッセージが殺到する可能性がある。ライブマイグレーションがダイナミックＩＢベースのクラウドにおいて行なわれる場合、多くの過剰なＳＡクエリが送信される可能性がある。ＳＡクエリの量は、ネットワークにおけるノードの数が増加するのに応じて多項式的に増加する。これらの開示された方法およびシステムは、サブネットにおけるノードによってサブネットマネージャに送信されるＳＡクエリの数を減らすことができるキャッシングメカニズムを提供する。

【００５３】
図４は、本発明の一実施形態に従った、クラウド環境におけるＳＡパス・キャッシングをサポートする例を示す。図４に示されるように、インフィニバンド（ＩＢ）サブネット４００は、サブネットマネージャ（ＳＭ）４１０および複数のホストノードＡ４０１およびＢ４０２を含み得る。
【００５４】
送信元ホストノードＡ４０１（たとえば、ＶＭ　Ａ４１１）が宛先ホストノードＢ４０２（たとえば、ＶＭ　Ｂ４１２）と通信しようと初めて試みたとき、送信元ホストノードＡ４０１はＳＭ４１０にＳＡパス記録要求を送信することができる。次いで、送信元ホストノードは、ローカルキャッシュ４２１を用いてパス情報（たとえば、パス記録４２２）を記憶することができる。
【００５５】
さらに、送信元ホストノードＡ４０１が同じ宛先ホストノードＢ４０２に再接続しようと試みるとき、送信元ホストノードＡ４０１は、サブネットマネージャに要求を送信するのではなく、ローカルキャッシュ４２１におけるキャッシングテーブル内の宛先ホストノードのアドレスを検索することができる。
【００５６】
パス情報が発見された場合、送信元ホストノードＡ４０１は、パス記録４２２によって示されるように、パス４２０を用いて宛先ホストノードＢ４０２に接続することができる。この場合、ＳＡクエリはＳＭ４１０に送信されない。それ以外の場合、送信元ホストノードＡ４０１は、必要なパス情報を取得するためにＳＭ４１０にＳＡパス記録要求を送信することができる。
【００５７】
図５は、本発明の一実施形態に従った、クラウド環境におけるＳＡパス・キャッシングをサポートする例を示す。より特定的には、図５は、インフィニバンド環境のサブネット内におけるＳＡパス・キャッシングをサポートする例を示す。
【００５８】
図５に示されるように、インフィニバンド（ＩＢ）サブネット５００は、異なるハイパーバイザ５１１および５１２をサポートする複数のホストノードＡ５０１およびＢ５０２を含み得る。加えて、各々のハイパーバイザ５１２および５１３は、さまざまな仮想マシン（ＶＭ）がその上で実行されることを可能にする。たとえば、ホストノードＡ１０１上のハイパーバイザ５１１は、ＶＭ　Ａ５０４をサポートすることができ、ホストノードＢ上のハイパーバイザ５１２はＶＭ　Ｂ５０５をサポートすることができる。
【００５９】
さらに、ホストノードＡ５０１およびＢ５０２の各々は１つ以上のホストチャネルアダプタ（ＨＣＡ）５１７および５１８に関連付けることができる。図５に示されるように、ホストノードＡ５０１上のＨＣＡ５１７は、ＶＭ　Ａ５０４が使用可能なＱＰａ５０８などのキュー対（ＱＰ）を利用することができ、ホストノードＢ５０２上のＨＣＡ５１８は、ＶＭ　Ｂ５０５が使用可能なＱＰｂ５０７を利用することができる。
【００６０】
一実施形態に従うと、各々のホストノードはまたメモリ５３０および５４０をサポートすることができ、メモリ５３０および５４０の各々は（ローカルキャッシュなどの）キャッシュ５３５および５４５を含み得るとともに、各々のキャッシュは、キャッシュテーブルに記憶することができる１つ以上のパス記録５３７および５４７を含み得る。
【００６１】
また、ＩＢサブネット５００は、サブネットマネージャ５１０を含み得る。サブネットマネージャ５１０は、ネットワーク初期化、スイッチおよびＨＣＡ上のＩＢポートの構成、トポロジー発見およびパス演算のために機能し得る。
【００６２】
一実施形態に従うと、送信元ホストノードＡ５０１（たとえば、ＶＭ　Ａ５０４）が宛

先ホストノードＢ５０２（たとえば、ＶＭ　Ｂ５０５）と最初に通信しようと試みたとき、送信元ホストノードＡ５０１はＳＭ５１０にＳＡパス記録要求を送信することができる。次いで、送信元ホストノードは、ローカルキャッシュ５３５を用いてパス情報（たとえば、パス記録５３７）を記憶することができる。

【００６３】

さらに、送信元ホストノードＡ５０１が同じ宛先ホストノードＢ５０２に再接続しようと試みるとき、送信元ホストノードＡ５０１は、サブネットマネージャに要求を送信するのではなく、キャッシュ５３５における宛先ホストノードのアドレスを検索することができる。

【００６４】

一実施形態に従うと、パス情報が発見されれば、送信元ホストノードＡ５０１は、パス記録５３７に示されたパスを用いることによって宛先ホストノードＢ５０２に接続することができるが、この場合、ＳＡクエリはＳＭ５１０に送信されない。それ以外の場合、送信元ホストノードＡ５０１は、ＳＭ４１０にＳＡパス記録要求を送信して、必要なパス情報を取得することができる。

【００６５】

一実施形態に従うと、ホストノードＡ５０１がサブネットマネージャ５１０にＳＡパス記録要求を送信する状況においては、受取られた応答は、キャッシングフラグを含み得る。このキャッシングフラグは、宛先ホストノードＢ５０２（ＤＧＩＤ）の所与のＧＩＤアドレスに関連付けられたパス特徴を記憶するために（キャッシュ５３５内における）ローカルキャッシングテーブルを用いるようにホストノードＡ５０１に指示し得る。

【００６６】

図６は、一実施形態に従った、クラウド環境におけるサブネット管理をサポートするための方法の例を示す。例示的な方法６００はステップ６０１から開始され得る。ステップ６０１において、少なくとも第１のハイパーバイザおよび第１のホストチャネルアダプタに関連付けられた第１のホストノードを含む複数のホストノードをクラウド環境内に設ける。ステップ６０２において、当該方法は、引続き、複数のアドレスに関連付けられた第１の仮想マシンを第１のホストノード上に設け得る。ステップ６０３において、当該方法は、引続き、クラウド環境内において、第１の仮想マシンを、複数のホストノードのうちの第１のホストノードから、設けられた第２のホストノードにまでマイグレートする。第２のホストノードは、少なくとも第２のハイパーバイザおよび第２のホストチャネルアダプタに関連付けられている。複数のホストノードの各々はローカルキャッシュを含む。各々のローカルキャッシュは１つ以上のパス記録を含む。

【００６７】

一実施形態に従うと、第１の仮想マシンをマイグレートするステップは、ステップ６０４において、第１のハイパーバイザから第１の仮想マシンを分離するステップを含み得る。第１のハイパーバイザから第１の仮想マシンを分離するステップは、第１の仮想マシンに関連付けられた第１の仮想機能を第１の仮想マシンから分離するステップを含む。ステップ６０５において、当該方法は、次に、第１の仮想マシンに関連付けられた複数のアドレスを第２のホストノードに与える。ステップ６０６において、当該方法は、第２のハイパーバイザに関連付けられた第２の仮想機能に複数のアドレスを割当てることができる。ステップ６０７において、当該方法は、第１の仮想マシンを第１のホストノードから第２のホストノード上の第２の仮想マシンにまでマイグレートすることができる。ステップ６０８において、当該方法は、第２の仮想マシンを、第１の仮想マシンに関連付けられた複数のアドレスにエクスポーズするステップで終了し得る。

【００６８】

一実施形態に従うと、ＳＡパス記録キャッシングメカニズムは、ＲＤＳプロトコルなどのプロトコルにおいて実現することができ、キャッシングテーブルは、各々のノードのメモリに記憶することができる。以下の擬似コードに示されるプログラムを用いることができる：

【００６９】
【表１－１】

```
1: private bool SA PathCachingEnabled
2: private list SA PathRecordCacheTable
3:
4: procedure RDSMODULEINITIALIZATION
5: // The Caching table is initialized
6: SA PathRecordCacheTable = empty
7:
8: // The system does not know yet if SA Path Caching is
9: // enabled by the SM, so we assume not.
10: SA PathCachingEnabled = False
11: end procedure
12:
13: procedure (RE-)CONNECTIONESTABLISHMENT(DGID)
14: struct PathRecord DST Path = NULL
15:
16: // Use the cache only if the SA Path Caching is
17: // enabled by the SM
18: if SA PathCachingEnabled then
19: if DGID in SA PathRecordCacheTable.
20: DGIDs then
21: DST Path = Cached PathRecord
22: end if
23: end if
24:
25: // If DST Path is NULL at this point, either the
26: // cache is disabled by the SM, or the path
27: // characteristics for the host with the given DGID
28: // have never been retrieved. In any case, a
29: // send a PathRecord Query can be sent to the SM.
30: if DST Path == NULL then
31: SendAnewSA PathRecordQueryToTheSM
32: WaitForTheReply
```

【００７０】

【表１－２】

```
33: DST Path = PathRecordResponse
34:
35: // If caching is enabled by the SM the reply will
36: // have the reserved field in the PathRecord set
37: // to 1. If not, the reserved field is 0
38: if DST Path ! Reserved Field != 0 then
39: SA PathCachingEnabled = True
40:
41: // Insert the DST Path in the caching table
42: SA PathRecordCacheTable.append(
43: DST Path)
44: end if
45: end if
46: connect to(DST Path)
47: end procedure
```

【００７１】

一実施形態に従うと、送信元ホスト（source host：ＳＨｏｓｔ）が宛先ホスト（destination host：ＤＨｏｓｔ）と通信しようと最初に試みるとき、ＳＨｏｓｔはサブネットマネージャにＳＡパス記録要求を送信することができる。応答によってキャッシングフラグが立てられた場合、ＳＨｏｓｔは、ローカルキャッシングテーブルを用いて、ＤＨｏｓｔ（ＤＧＩＤ）の所与のＧＩＤアドレスに関連付けられたパス特徴を記憶することができる。さらに、ＳＨｏｓｔは、キャッシングがサブネットマネージャによってサポートされていることをこのとき認識し、このため、次にＳＨｏｓｔがいずれかのＤＨｏｓｔに接続または再接続しようと試みるときに、ＳＨｏｓｔはキャッシングテーブルを最初に検索することとなる。所与のＤＨｏｓｔについてのパス情報が発見された場合、ＳＨｏｓｔは、サブネットマネージャに送信されるＳＡクエリの送信が妨害される可能性があり、代わりに、ＳＨｏｓｔは、そのキャッシングテーブル内の情報を用いてＤＨｏｓｔに接続しようと試みることができる。

【００７２】

遮断された接続（ステップ２）に関して、図３を再び参照すると、上述のキャッシングメカニズムが使用可能にされるシステムにおいては、サブネットマネージャにＳＡクエリが送信される必要はない。図３に記載される場合においては、ステップ４、５、９および１０が省かれており、このため、接続がより高速に回復され、サブネットマネージャに対する負荷（たとえばＳＡパスの要求および応答）がより低下する。

【００７３】

この発明の多くの特徴は、ハードウェア、ソフトウェア、ファームウェア、またはそれらの組合せにおいて、それを使用して、またはその助けを借りて実行され得る。従って、この発明の特徴は、（たとえば、１つ以上のプロセッサを含む）処理システムを用いて実現され得る。

【００７４】

この発明の特徴は、ここに提示された特徴のうちのいずれかを実行するように処理システムをプログラミングするために使用可能な命令を格納した記憶媒体またはコンピュータ読取可能媒体であるコンピュータプログラム製品において、それを使用して、またはその助けを借りて実現され得る。記憶媒体は、フロッピー（登録商標）ディスク、光ディスク、ＤＶＤ、ＣＤ－ＲＯＭ、マイクロドライブ、および光磁気ディスクを含む任意のタイプ

のディスク、ＲＯＭ、ＲＡＭ、ＥＰＲＯＭ、ＥＥＰＲＯＭ、ＤＲＡＭ、ＶＲＡＭ、フラッシュメモリ装置、磁気カードもしくは光カード、ナノシステム（分子メモリＩＣを含む）、または、命令および／もしくはデータを格納するのに適した任意のタイプの媒体もしくは装置を含み得るものの、それらに限定されない。

【００７５】

この発明の特徴は、機械読取可能な媒体のうちのいずれかに格納された状態で、処理システムのハードウェアを制御するために、および処理システムがこの発明の結果を利用する他のメカニズムとやりとりすることを可能にするために、ソフトウェアおよび／またはファームウェアに組込まれ得る。そのようなソフトウェアまたはファームウェアは、アプリケーションコード、装置ドライバ、オペレーティングシステム、および実行環境／コンテナを含み得るものの、それらに限定されない。

【００７６】

この発明の特徴はまた、たとえば、特定用途向け集積回路（application specific integrated circuit：ＡＳＩＣ）などのハードウェアコンポーネントを使用して、ハードウェアにおいて実現されてもよい。ここに説明された機能を実行するようにハードウェアステートマシンを実現することは、関連技術の当業者には明らかであるだろう。

【００７７】

加えて、この発明は、この開示の教示に従ってプログラミングされた１つ以上のプロセッサ、メモリおよび／またはコンピュータ読取可能記憶媒体を含む、１つ以上の従来の汎用または専用のデジタルコンピュータ、コンピューティング装置、マシン、またはマイクロプロセッサを使用して都合よく実現され得る。ソフトウェア技術の当業者には明らかであるように、適切なソフトウェアコーディングが、この開示の教示に基づいて、熟練したプログラマによって容易に準備され得る。

【００７８】

この発明のさまざまな実施形態が上述されてきたが、それらは限定のためではなく例示のために提示されたことが理解されるべきである。この発明の精神および範囲から逸脱することなく、形状および詳細のさまざまな変更を行なうことができることが関連技術の当業者にとって明らかになるであろう。

【００７９】

この発明は、特定された機能およびそれら関係の実行を示す機能的構築ブロックの助けを借りて上述されてきた。説明の便宜上、これらの機能的構築ブロックの境界はしばしば、この明細書中では任意に規定されてきた。特定された機能およびそれらの関係が適切に実行される限り、代替的な境界を規定することができる。このため、そのようないかなる代替的な境界も、この発明の範囲および精神内に含まれる。

【００８０】

この発明の上述の説明は、例示および説明のために提供されてきた。それは、網羅的であるよう、またはこの発明を開示された形状そのものに限定するよう意図されてはいない。この発明の幅および範囲は、上述の例示的な実施形態のいずれによっても限定されるべきでない。多くの変更および変形が当業者には明らかであるだろう。これらの変更および変形は、開示された特徴の関連するあらゆる組合せを含む。実施形態は、この発明の原理およびその実用的応用を最良に説明するために選択され説明されたものであり、それにより、企図される特定の用途に適したさまざまな実施形態についての、およびさまざまな変更例を有するこの発明を、当業者が理解できるようにする。この発明の範囲は、添付の特許請求の範囲およびそれらの均等物によって定義されるよう意図されている。

【図１】

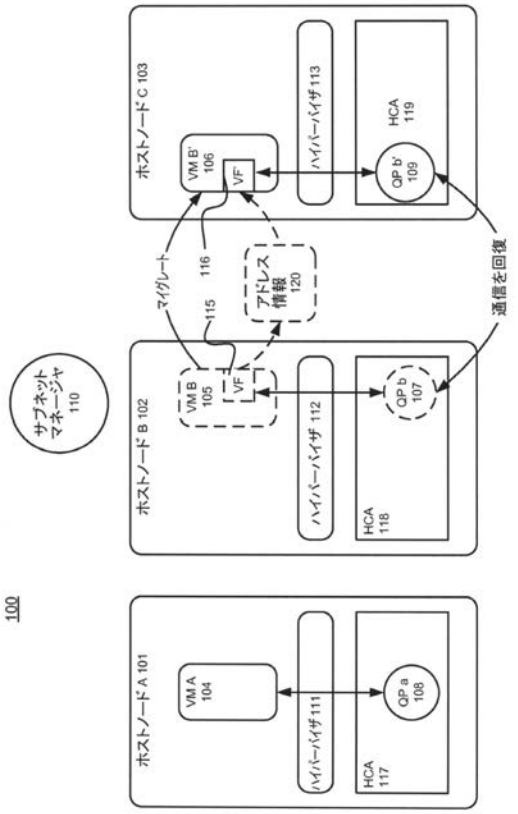

FIGURE 1

【図２】

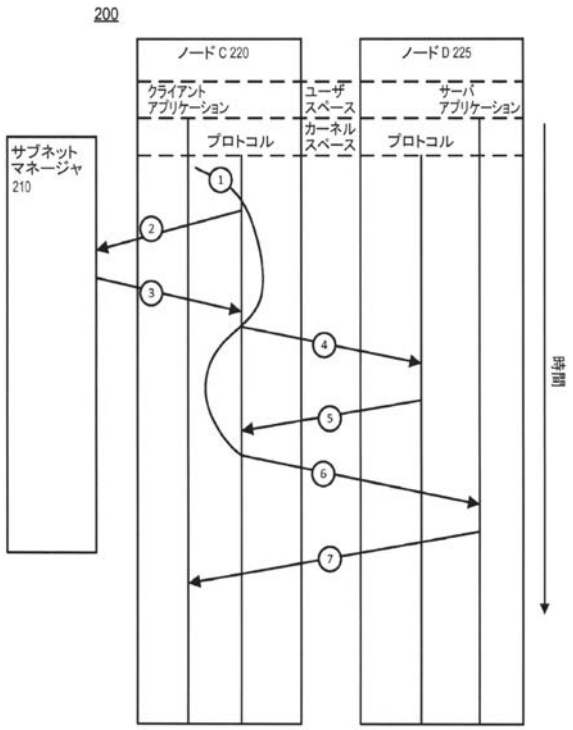

FIGURE 2

【図３】

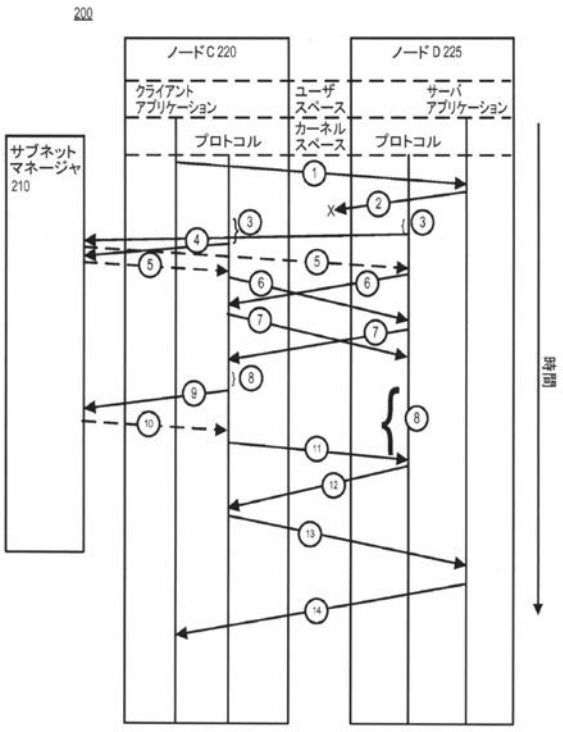

FIGURE 3

【図４】

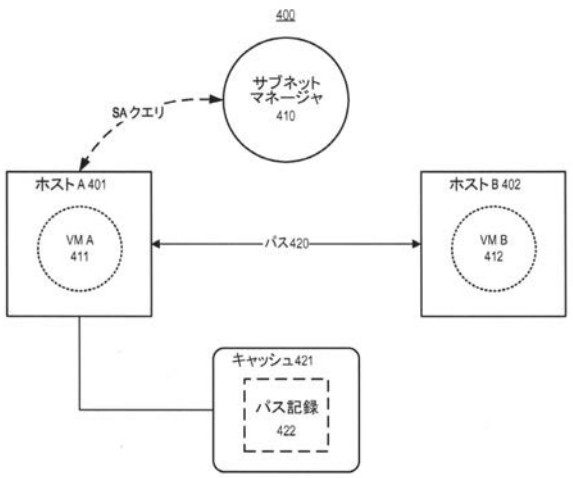

FIGURE 4

【図５】

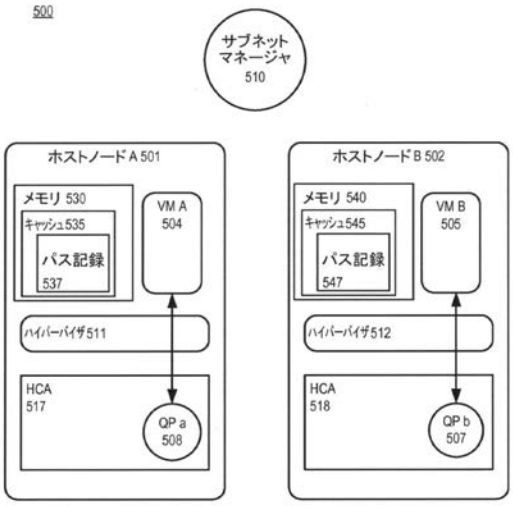

FIGURE 5

【図６】

FIGURE 6

【国際調査報告】

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2015/057860

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F9/48
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CLARK C ET AL: "Live migration of virtual machines", PROCEEDINGS OF THE SYMPOSIUM ON NETWORKED SYSTEMS DESIGN ANDIMPLEMENTATION, USENIX ASSOCIATION, BERKELEY, CA, US, 2 May 2005 (2005-05-02), - 4 May 2005 (2005-05-04), pages 273-286, XP002443245, figure 1 page 274, left-hand column, line 1 - line 13 page 275, left-hand column, line 10 - line 17 page 276, left-hand column, line 1 - line 27 page 276, left-hand column, line 47 - right-hand column, line 33 page 282, left-hand column, line 21 - line 30 page 285, left-hand column, line 16 - line -/--	1-22

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
28 January 2016	04/02/2016

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Wirtz, Hanno

2

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2015/057860

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
	30	
X	US 2008/189432 A1 (ABALI BULENT [US] ET AL) 7 August 2008 (2008-08-07)	1-3,5-9, 11-13, 15-20,22
A	claims 1,2 figure 3 paragraph [0009] - paragraph [0010] paragraph [0015] paragraph [0045] - paragraph [0055]	4,10,14, 21
A	US 2008/186990 A1 (ABALI BULENT [US] ET AL) 7 August 2008 (2008-08-07) claims 1-5 figure 2 paragraph [0012] - paragraph [0016] paragraph [0021] - paragraph [0022]	1-22
X,P	EVANGELOS TASOULAS ET AL: "A Novel Query Caching Scheme for Dynamic InfiniBand Subnets", 2015 15TH IEEE/ACM INTERNATIONAL SYMPOSIUM ON CLUSTER, CLOUD AND GRID COMPUTING, 4 May 2015 (2015-05-04), - 7 May 2015 (2015-05-07), pages 199-210, XP055244069, DOI: 10.1109/CCGrid.2015.10 ISBN: 978-1-4799-8006-2 the whole document	1-22

2

Form PCT/ISA/210 (continuation of second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2015/057860

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2008189432 A1	07-08-2008	NONE	
US 2008186990 A1	07-08-2008	NONE	

フロントページの続き

(31)優先権主張番号 62/121,294
(32)優先日 平成27年2月26日(2015.2.26)
(33)優先権主張国 米国(US)
(31)優先権主張番号 62/133,179
(32)優先日 平成27年3月13日(2015.3.13)
(33)優先権主張国 米国(US)
(31)優先権主張番号 14/924,281
(32)優先日 平成27年10月27日(2015.10.27)
(33)優先権主張国 米国(US)

(81)指定国 AP(BW,GH,GM,KE,LR,LS,MW,MZ,NA,RW,SD,SL,ST,SZ,TZ,UG,ZM,ZW),EA(AM,AZ,BY,KG,KZ,RU,TJ,TM),EP(AL,AT,BE,BG,CH,CY,CZ,DE,DK,EE,ES,FI,FR,GB,GR,HR,HU,IE,IS,IT,LT,LU,LV,MC,MK,MT,NL,NO,PL,PT,RO,RS,SE,SI,SK,SM,TR),OA(BF,BJ,CF,CG,CI,CM,GA,GN,GQ,GW,KM,ML,MR,NE,SN,TD,TG),AE,AG,AL,AM,AO,AT,AU,AZ,BA,BB,BG,BH,BN,BR,BW,BY,BZ,CA,CH,CL,CN,CO,CR,CU,CZ,DE,DK,DM,DO,DZ,EC,EE,EG,ES,FI,GB,GD,GE,GH,GM,GT,HN,HR,HU,ID,IL,IN,IR,IS,JP,KE,KG,KN,KP,KR,KZ,LA,LC,LK,LR,LS,LU,LY,MA,MD,ME,MG,MK,MN,MW,MX,MY,MZ,NA,NG,NI,NO,NZ,OM,PA,PE,PG,PH,PL,PT,QA,RO,RS,RU,RW,SA,SC,SD,SE,SG,SK,SL,SM,ST,SV,SY,TH,TJ,TM,TN,TR,TT,TZ,UA,UG,US

(72)発明者 グラン，アーンスト・ガンナー
ノルウェー、エヌ－１３２５ リュサケール、ピィ・オゥ・ボックス・１３４