(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5022412号
(P5022412)
(45)発行日　平成24年9月12日(2012.9.12)　(24)登録日　平成24年6月22日(2012.6.22)

(51)Int.Cl.　F I
H04L 12/56 (2006.01)　H04L 12/56 100Z

請求項の数 10 (全 25 頁)

(21)出願番号　特願2009-180328(P2009-180328)
(22)出願日　平成21年8月3日(2009.8.3)
(65)公開番号　特開2011-35686(P2011-35686A)
(43)公開日　平成23年2月17日(2011.2.17)
審査請求日　平成23年8月1日(2011.8.1)

(73)特許権者　000004226
日本電信電話株式会社
東京都千代田区大手町二丁目３番１号
(74)代理人　100064414
弁理士　磯野　道造
(74)代理人　100127720
弁理士　大石　恵
(74)代理人　100162374
弁理士　中村　新二
(72)発明者　クリステル　ペルサー
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内
(72)発明者　増田　暁生
東京都千代田区大手町二丁目３番１号　日本電信電話株式会社内

最終頁に続く

(54)【発明の名称】経路情報管理システム、経路情報管理方法、およびプログラム

(57)【特許請求の範囲】
【請求項１】
ＩＰ（Internet Protocol）ネットワークを構成するＡＳ（Autonomous System）内で、他のＡＳから該他のＡＳのＩＰアドレスの集合であるプリフィクスを含む外部経路情報を取得し、かつパケットを他のＡＳに転送する複数のルータと、同一のプリフィクスを含む外部経路情報を前記ルータから収集して経路計算を実行する２以上の経路情報管理装置とが相互に通信可能に接続される経路情報管理システムであって、
前記経路情報管理装置は、
前記ルータから収集した前記プリフィクスごとに、そのプリフィクスを含む外部経路情報を用いて、前記ＡＳ内の２以上のルータの一まとまりを単位として前記ＡＳ内に複数設定されるルータ集合に属する少なくともいずれか１つのルータにおける、パケットの転送先のプリフィクスに至る転送経路を計算する経路計算を実行する経路計算部と、
前記経路計算部によって経路計算されたルータ集合ごとの経路計算の結果を含む経路計算結果情報を送信する当該ルータ集合に属する所定数のルータを選択する格納先選択部と、
当該経路計算結果情報を、前記格納先選択部によって選択された当該所定数のルータに送信する入出力部と、
を備え、
前記ルータは、
転送すべきパケットを受信したときに、前記経路情報管理装置から受信した前記経路計

算結果情報を参照して、該パケットの転送先を決定する転送経路決定部と、
前記転送経路決定部によって決定された転送先に前記転送すべきパケットを送信する入出力部と、
を備える
ことを特徴とする経路情報管理システム。
【請求項２】
前記経路情報管理装置は、さらに、
前記ルータ集合ごとに、少なくとも１以上のルータのルータＩＤとして、当該ルータの属する前記ルータ集合のルータ集合ＩＤのビット列の後にハッシュ関数によって生成されたハッシュ値のビット列を連結したビット列を記憶する格納先ノード情報を格納する記憶部と、
前記ルータ集合ごとに、前記経路計算結果情報の格納先の検索に用いる検索キーとして、当該ルータ集合のルータ集合ＩＤのビット列、前記プリフィクスのビット列、前記プリフィクスに対するマスク値のビット列、およびパディングとしてすべて「０」のビット列を連結したビット列を生成する検索キー生成部と、
を備え、
前記格納先選択部は、
前記格納先ノード情報に記憶しているルータＩＤと前記検索キーとの距離を、当該ルータＩＤと前記検索キーとのビットごとの排他的論理和を示すビット列において、左から見て最初に「１」が出現したビット位置について右から数えたビットの桁数の数値として算出し、その算出した距離の小さい順に所定数のルータを当該ルータ集合から送信先として選択する
ことを特徴とする請求項１に記載の経路情報管理システム。
【請求項３】
前記経路計算結果情報は、前記検索キーと、その検索キーに含まれるプリフィクスに対応する前記ルータ集合ごとの経路計算の結果とを関連付けた情報である
ことを特徴とする請求項２に記載の経路情報管理システム。
【請求項４】
前記ルータは、さらに、
そのルータの属するルータ集合内の自身を除く少なくとも１以上のルータのルータＩＤを記憶する取得先ノード情報と前記経路計算結果情報とを記憶する記憶部と、
転送すべきパケットを受信したとき、当該ルータの属するルータ集合ＩＤのビット列、当該パケットの宛先ＩＰアドレスのビット列、前記宛先ＩＰアドレスに対するマスク値としてすべて「１」のビット列、およびパディングとしてすべて「０」のビット列を連結して形成される、取得キーを生成する取得キー生成部と、
前記取得先ノード情報に記憶しているルータＩＤと前記取得キーとの距離を、前記ルータＩＤと前記取得キーとのビットごとの排他的論理和を示すビット列において、左から見て最初に「１」が出現したビット位置について右から数えたビットの桁数の数値として算出し、その算出した距離の小さい順に所定数のルータを当該ルータ集合から取得先として選択する取得先選択部と、
を備え、
前記転送経路決定部は、
自身の前記記憶部に記憶する前記経路計算結果情報の検索キーと前記取得キーとが一致するか否かを判定し、一致するものがある場合には、その検索キーに関連付けられた前記ルータ集合ごとの経路計算の結果に基づいてパケットの転送先を決定し、一致するものがない場合には、前記取得先選択部によって取得先として選択されたルータに前記取得キーを送信して、当該ルータの記憶する前記経路計算結果情報の検索キーと前記取得キーとが一致する前記ルータ集合ごとの経路計算の結果を取得して、その取得した前記ルータ集合ごとの経路計算の結果に基づいてパケットの転送先を決定する
ことを特徴とする請求項３に記載の経路情報管理システム。

【請求項５】
前記転送経路決定部は、さらに、前記経路計算結果情報の検索キーと前記取得キーとが一致するものがない場合、
前記取得キーの宛先ＩＰアドレス部分の最後のゼロでないビットをゼロに変更するとともに、前記取得キーのマスク値部分の当該宛先ＩＰアドレス部分の変更位置を含む右側のビットをゼロに変更し、その変更したビット列を新たな取得キーに設定し、その新たな取得キーを用いて、自身の前記取得先選択部によって選択された取得先のルータから、さらにその取得先のルータの前記取得先選択部によって選択された取得先のルータと経路計算結果情報とを取得することを繰り返し、取得されたすべての前記経路計算結果情報の検索キーと前記新たな取得キーとが一致するか否かを判定して、前記経路計算結果情報の検索キーと前記取得キーとが最長一致となるときの前記ルータ集合ごとの経路計算の結果に基づいてパケットの転送先を決定することを、前記取得キーの宛先ＩＰアドレス部分のビット列がすべてゼロになるまで繰り返し実行する
ことを特徴とする請求項４に記載の経路情報管理システム。
【請求項６】
前記取得先ノード情報は、前記ルータＩＤを複数の範囲に区分し、自身のルータのルータＩＤから前記距離の近いルータＩＤほどその区分の範囲を狭く設定し、前記距離の遠いルータＩＤほどその区分の範囲を広く設定して、それぞれの前記範囲ごとにその範囲に属するルータのルータＩＤを記憶し、
前記取得先選択部は、前記取得キーと前記範囲の境界のルータＩＤとを比較して、前記取得キーを含む前記範囲を抽出し、前記取得キーと前記抽出した範囲に属するルータのルータＩＤとの前記距離を算出し、前記距離の近い順に所定数のルータを取得先として選択する
ことを特徴とする請求項４に記載の経路情報管理システム。
【請求項７】
請求項４に記載の経路情報管理システムにおいて用いられる経路情報管理装置の経路情報管理方法であって、
前記経路情報管理装置は、
前記ＡＳ内の２以上のルータの一まとまりを単位として前記ＡＳ内に複数設定されるルータ集合ごとに、少なくとも所定数以上のルータのルータＩＤとして、当該ルータの属する前記ルータ集合のルータ集合ＩＤのビット列の後にハッシュ関数によって生成されたハッシュ値のビット列を連結したビット列を記憶する格納先ノード情報を格納する記憶部を備え、
前記ルータから収集した前記プリフィクスごとに、そのプリフィクスを含む外部経路情報を用いて、前記ＡＳ内の２以上のルータの一まとまりを単位として前記ＡＳ内に複数設定されるルータ集合に属する少なくともいずれか１つのルータにおける、パケットの転送先のプリフィクスに至る転送経路を計算する経路計算を実行する経路計算ステップと、
前記ルータ集合ごとに、前記経路計算ステップによって経路計算されたルータ集合ごとの経路計算の結果を含む経路計算結果情報の格納先の検索に用いる検索キーとして、当該ルータ集合のルータ集合ＩＤのビット列、前記プリフィクスのビット列、前記プリフィクスに対するマスク値のビット列、およびパディングとしてすべて「０」のビット列を連結したビット列を生成する検索キー生成ステップと、
前記格納先ノード情報に記憶しているルータＩＤと前記検索キーとの距離を、当該ルータＩＤと前記検索キーとのビットごとの排他的論理和を示すビット列において、左から見て最初に「１」が出現したビット位置について右から数えたビットの桁数の数値として算出し、その算出した距離の小さい順に所定数のルータを当該ルータ集合から送信先として選択する格納先選択ステップと、
当該経路計算結果情報を、前記格納先選択ステップによって選択された当該所定数のルータに送信する入出力ステップと
を実行することを特徴とする経路情報管理装置の経路情報管理方法。

【請求項８】
請求項４に記載の経路情報管理システムにおいて用いられるルータの経路情報管理方法であって、
前記ルータは、
前記ＡＳ内の２以上のルータの一まとまりを単位として前記ＡＳ内に複数設定されるルータ集合ごとに、少なくとも所定数以上のルータのルータＩＤとして、当該ルータの属する前記ルータ集合のルータ集合ＩＤのビット列の後にハッシュ関数によって生成されたハッシュ値のビット列を連結したビット列を記憶する取得先ノード情報と、前記経路情報管理装置から送信される検索キーと前記ルータ集合ごとの経路計算の結果とを関連付けた情報を示す経路計算結果情報とを記憶する記憶部を備え、
転送すべきパケットを受信したとき、当該ルータの属するルータ集合ＩＤのビット列、当該パケットの宛先ＩＰアドレスのビット列、前記宛先ＩＰアドレスに対するマスク値としてすべて「１」のビット列、およびパディングとしてすべて「０」のビット列を連結して形成される、取得キーを生成する取得キー生成ステップと、
前記取得先ノード情報に記憶しているルータＩＤと前記取得キーとの距離を、前記ルータＩＤと前記取得キーとのビットごとの排他的論理和を示すビット列において、左から見て最初に「１」が出現したビット位置について右から数えたビットの桁数の数値として算出し、その算出した距離の小さい順に所定数のルータを当該ルータ集合から取得先として選択する取得先選択ステップと、
転送すべきパケットを受信したときに、自身の前記記憶部に記憶する前記経路計算結果情報の検索キーと前記取得キーとが一致するか否かを判定し、一致するものがある場合には、その検索キーに関連付けられた前記ルータ集合ごとの経路計算の結果に基づいてパケットの転送先を決定し、一致するものがない場合には、前記取得先選択部によって取得先として選択されたルータに前記取得キーを送信して、当該ルータの記憶する前記経路計算結果情報の検索キーと前記取得キーとが一致する前記ルータ集合ごとの経路計算の結果を取得して、その取得した前記ルータ集合ごとの経路計算の結果に基づいてパケットの転送先を決定する転送経路決定ステップと、
前記転送経路決定ステップにおいて決定された転送先に前記パケットを転送する入出力ステップと
を実行することを特徴とするルータの経路情報管理方法。
【請求項９】
請求項７に記載の経路情報管理方法を、コンピュータとしての経路情報管理装置に実行させるためのプログラム。
【請求項１０】
請求項８に記載の経路情報管理方法を、コンピュータとしてのルータに実行させるためのプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、ＩＰ（Internet Protocol）ネットワークにおいて、インタードメインの経路情報をＡＳ（自律システム；Autonomous System）内で分散保持する技術に関する。
【背景技術】
【０００２】
インターネット等のＩＰネットワークは、異なる組織によって運営される多数のＡＳが接続されて形成されている。ＡＳ間は、それぞれのＡＳの境界ルータによって接続されている。ＡＳは、そのＡＳの外部を宛先とするパケットを転送する必要があるため、インターネットの経路情報をＡＳ間でＢＧＰ（Border Gateway Protocol）を用いて交換している。そして、ＡＳ内のすべてのルータは、ＡＳ外部の外部経路情報をすべて保持する必要がある。具体的には、ＡＳの境界ルータ同士は、ｅＢＧＰ（external BGP）によって、ＡＳ外部の宛先プリフィクス(prefix)（ＩＰアドレスの集合）に対する外部経路情報（ＡＳ

外部のどの境界ルータにパケットを転送すれば宛先に到達するかの情報）を交換している。そして、その外部経路情報は、ｉＢＧＰ（internal BGP）によって、ＡＳ内のルータに広告される。ただし、ｉＢＧＰを用いてルータ間で広告するためには、ＩＧＰ（Interior Gateway Protocol）によってルータ間が通信可能に接続されている必要がある。

【０００３】

広告された外部経路情報を受信したルータは、受信した外部経路情報に基づいて、宛先プリフィクスごとに、どの境界ルータに転送すれば宛先に到達できるのかを示した転送経路、すなわち、最適な転送経路を計算して保持していた。したがって、ルータは、宛先プリフィクスの数に対応する外部経路情報および経路計算によって算出した転送経路情報を記憶する必要があった。また、ルータは、宛先プリフィクスの数に対応する外部経路情報をメッセージとして広告する必要があった。

【０００４】

近年は、インターネットの急速な拡大にともなって、インターネット上に存在する宛先プリフィクスの数が増大してきている。そのため、インターネットサービスプロバイダ等のネットワーク事業者は、ネットワーク設備であるルータ等における、保持すべき経路情報の記憶容量の拡張や経路計算の処理負荷量の増大に対処せざるを得なくなり、設備更改のコスト負担増を強いられている。また、ルータ一台当りの、保持すべき経路数の増大、送信する必要のあるメッセージ数の増大、および経路計算の処理負荷量の増大は、転送すべきパケットを受信してからＢＧＰ転送経路を決定するまでの時間の増大をもたらすため、ネットワーク制御の安定性に大きな影響を及ぼす虞がある。

【０００５】

従来技術では、ルータの処理負荷の低減のために、ルートリフレクタ（ＲＲ）（非特許文献１参照）またはＡＳコンフェデレーション（非特許文献２参照）を導入することによって、ｉＢＧＰを用いて広告するｉＢＧＰセッション数の削減が図られている。ルートリフレクタ（ＲＲ）は、ピアを確立したルートリフレクタクライアントとの間だけで外部経路情報を転送し、ピアを確立していないルートリフレクタクライアント同士では転送を行わない。そのため、ルートリフレクタ（ＲＲ）は、ＡＳ内に確立すべきピア数を、ｉＢＧＰを用いてＡＳ内の各ルータ間をフルメッシュで接続する場合より減少することが可能である。また、ＡＳコンフェデレーションは、ＡＳ内にいくつかのサブＡＳを構成することによって、ピア数をフルメッシュの場合より減少することが可能である。このように、ルートリフレクタ（ＲＲ）またはＡＳコンフェデレーションでは、ピア数を減少できるため、ルータの処理負荷の増大を抑制することが可能である。

【０００６】

また、ルータの記憶容量を低減するために、Compact Routing（非特許文献３，４参照）が提案されている。Compact Routingは、パケットを一旦経路の集約点となるルータに向けて転送し、その集約点となるルータから宛先アドレスへ転送する技術である。すなわち、Compact Routingは、経路を集約することによって、各ルータの経路情報の保持に要する記憶容量を低減している。

【先行技術文献】

【非特許文献】

【０００７】

【非特許文献１】T.Bates, E.Chen, and R.Chandra,"BGP Route Reflection : An Alternative to Full Mesh Internal BGP (IBGP)", RFC4456, IETF, 2006年4月

【非特許文献２】P.Traina, D.McPherson, and J.Scudder, "Autonomous System Confederations for BGP", RFC5065, IETF, 2007年8月

【非特許文献３】D.Krioukov, kc claffy, K.Fall, and A.Brady, "On Compact Routing for the Internet", SIGCOMM Comput. Commun. Rev., Vol.37, No.3, 2007年7月, p.41-52

【非特許文献４】P.Francis, X.Xu, and H.Ballani, "FIB Suppression with Virtual Aggregation", internet Draft, draft-ietf-grow-va-00.txt, 2009年5月

【発明の概要】
【発明が解決しようとする課題】
【０００８】
しかしながら、非特許文献１に記載のルートリフレクタ（ＲＲ）を用いる方法、および非特許文献２に記載のＡＳコンフェデレーションを用いる方法は、ｉＢＧＰトポロジがフルメッシュでなくなり、また外部経路情報を交換する際にｉＢＧＰセッションにおいて宛先プリフィクス（以降、プリフィクスともいう。）ごとに１つの経路情報しか広告しない。そのため、ｉＢＧＰセッション数が削減されて、ルータの処理負荷の低減や、経路情報の記憶容量の低減にある程度の効果は認められる。しかし、ルータが一部の経路情報しか記憶しないため、不完全な外部経路情報に基づいて経路計算が実行されてしまうので、経路計算の結果が必ずしも最適とならず、また計算された経路がループとなって収束しないことがあるという問題が発生する。
【０００９】
また、非特許文献３，４に記載のCompact Routingを用いる方法は、パケットを一旦経路の集約点となるルータに集めるため、パケットの転送経路が必ずしも最適（最短）とならず、ネットワークのリソースを無駄に使用するという問題が発生する。
【００１０】
そこで、本発明の課題は、前記した問題を解決しつつ、すなわち、すべての経路情報に基づいて経路計算を実行して最適な転送経路を算出することを満足しつつ、インターネットの今後の拡大に対応できる新しいルーティングアーキテクチャとして、ルータ一台当りに保持すべき経路数（記憶容量）の増大、送信するメッセージ数の増大、および処理負荷の増大を抑制する技術を提供することを目的とする。
【課題を解決するための手段】
【００１１】
本発明は、ＩＰ（Internet Protocol）ネットワークを構成するＡＳ（Autonomous System）内で、他のＡＳから該他のＡＳのＩＰアドレスの集合であるプリフィクスを含む外部経路情報を取得し、かつパケットを他のＡＳに転送する複数のルータと、同一のプリフィクスを含む外部経路情報を前記ルータから収集して経路計算を実行する２以上の経路情報管理装置とが相互に通信可能に接続される経路情報管理システムであって、前記経路情報管理装置が、前記ルータから収集した前記プリフィクスごとに、そのプリフィクスを含む外部経路情報を用いて、前記ＡＳ内の２以上のルータの一まとまりを単位として前記ＡＳ内に複数設定されるルータ集合に属する少なくともいずれか１つのルータにおける、パケットの転送先のプリフィクスに至る転送経路を計算する経路計算を実行する経路計算部と、前記経路計算部によって経路計算されたルータ集合ごとの経路計算の結果を含む経路計算結果情報を送信する当該ルータ集合に属する所定数のルータを選択する格納先選択部と、当該経路計算結果情報を、前記格納先選択部によって選択された当該所定数のルータに送信する入出力部と、を備え、前記ルータが、転送すべきパケットを受信したときに、前記経路情報管理装置から受信した前記経路計算結果情報を参照して、該パケットの転送先を決定する転送経路決定部と、前記転送経路決定部によって決定された転送先に前記転送すべきパケットを送信する入出力部と、を備えることを特徴とする。
【００１２】
このような構成によれば、経路情報管理装置がプリフィクスに係る外部経路情報を格納するため、ルータに保持すべき経路数（記憶容量）を削減することができる。また、経路情報管理装置は、同一のプリフィクスに係る外部経路情報についてすべてを収集して経路計算を実行するため、経路計算結果が最適とならないという問題は発生しない。また、ＡＳ内に複数のルータ集合を設けた場合、そのルータ集合内のＢＧＰ経路はプリフィクスに対して共通になるため、経路情報管理装置は、ルータ集合内の一つのルータに対してだけ経路計算を実行すれば良くなる。そのため、経路情報管理装置は、すべてのルータに対して経路計算を実行する必要はなく処理負荷を低減することができる。また、ルータは、経路計算を実行しなくて良いため、処理負荷を低減することができる。さらに、経路情報管

理装置によって算出されたルータ集合ごとの経路計算の結果は、ルータ集合内の所定数のルータにしか送信しないため、経路計算の結果をメッセージとして送信するときのメッセージ数を削減することができる。

【００１３】

本発明は、前記経路情報管理装置が、さらに、前記ルータ集合ごとに、少なくとも１以上のルータのルータＩＤとして、当該ルータの属する前記ルータ集合のルータ集合ＩＤのビット列の後にハッシュ関数によって生成されたハッシュ値のビット列を連結したビット列を記憶する格納先ノード情報を格納する記憶部と、前記ルータ集合ごとに、前記経路計算結果情報の格納先の検索に用いる検索キーとして、当該ルータ集合のルータ集合ＩＤのビット列、前記プリフィクスのビット列、前記プリフィクスに対するマスク値のビット列、およびパディングとしてすべて「０」のビット列を連結したビット列を生成する検索キー生成部と、を備え、前記格納先選択部が、前記格納先ノード情報に記憶している前記ルータＩＤと前記検索キーとの距離を、前記ルータＩＤと前記検索キーとのビットごとの排他的論理和を示すビット列において、左から見て最初に「１」が出現したビット位置について右から数えたビットの桁数の数値として算出し、その算出した距離の小さい順に所定数のルータを当該ルータ集合から送信先として選択することを特徴とする。

【００１４】

このような構成によれば、ルータＩＤのビット列にはルータ集合ＩＤとハッシュ値とを含み、検索キーのビット列にはルータ集合ＩＤとプリフィクスとを含み、ルータＩＤと検索キーとの距離の小さい順に所定数のルータが選択される。したがって、ルータ集合ＩＤが一致して、プリフィクスに距離の近いハッシュ値を有するルータが選択される。そのため、選択される経路計算結果情報の格納先のルータがほぼ均等化される。すなわち、格納先のルータがルータ集合内でほぼ均等に選択されるようになり、ルータの記憶容量の平均化が実現できる。言い換えると、一台当りのルータの記憶容量は、すべての経路計算結果情報を記憶していた従来の一台のルータの記憶容量に比べて、低減できる。

【００１５】

本発明は、前記経路計算結果情報が、前記検索キーと、その検索キーに含まれるプリフィクスに対応する前記ルータ集合ごとの経路計算の結果とを関連付けた情報であることを特徴とする。

【００１６】

このような構成によれば、経路計算結果情報が、検索キーと、その検索キーに含まれるプリフィクスに対応するルータ集合ごとの経路計算の結果とを関連付けた情報となっている。そのため、検索キーを媒介として、検索キーのプリフィクスに距離の近いノードＩＤのルータに、当該プリフィクスに対応するルータ集合ごとの経路計算の結果を記憶させることができる。したがって、あるプリフィクスに係る経路計算結果情報を探索する場合、当該プリフィクスを含む検索キーを用いれば、その検索キーに距離の近いノードＩＤのルータおよびそのルータに記憶されている当該プリフィクスに対応するルータ集合ごとの経路計算の結果を取得することができる。すなわち、一台当りのルータの記憶容量を抑制するために分散して記憶されている経路計算結果情報の中から、検索キーによって、その検索キーに含まれるプリフィクスに対応するルータ集合ごとの経路計算の結果を検索することが可能となる。

【００１７】

本発明は、前記ルータが、さらに、そのルータの属するルータ集合内の自身を除く少なくとも１以上のルータのルータＩＤを記憶する取得先ノード情報と前記経路計算結果情報とを記憶する記憶部と、転送すべきパケットを受信したとき、当該ルータの属するルータ集合ＩＤのビット列、当該パケットの宛先ＩＰアドレスのビット列、前記宛先ＩＰアドレスに対するマスク値としてすべて「１」のビット列、およびパディングとしてすべて「０」のビット列を連結して形成される、取得キーを生成する取得キー生成部と、前記取得先ノード情報に記憶しているルータＩＤと前記取得キーとの距離を、前記ルータＩＤと前記取得キーとのビットごとの排他的論理和を示すビット列において、左から見て最初に「１

」が出現したビット位置について右から数えたビットの桁数の数値として算出し、その算出した距離の小さい順に所定数のルータを当該ルータ集合から取得先として選択する取得先選択部と、を備え、前記転送経路決定部が、自身の前記記憶部に記憶する前記経路計算結果情報の検索キーと前記取得キーとが一致するか否かを判定し、一致するものがある場合には、その検索キーに関連付けられた前記ルータ集合ごとの経路計算の結果に基づいてパケットの転送先を決定し、一致するものがない場合には、前記取得先選択部によって取得先として選択されたルータに前記取得キーを送信して、当該ルータの記憶する前記経路計算結果情報の検索キーと前記取得キーとが一致する前記ルータ集合ごとの経路計算の結果を取得して、その取得した前記ルータ集合ごとの経路計算の結果に基づいてパケットの転送先を決定することを特徴とする。

【００１８】

また、本発明は、前記転送経路決定部が、さらに、前記経路計算結果情報の検索キーと前記取得キーとが一致するものがない場合、前記取得キーの宛先ＩＰアドレス部分の最後のゼロでないビットをゼロに変更するとともに、前記取得キーのマスク値部分の当該宛先ＩＰアドレス部分の変更位置を含む右側のビットをゼロに変更し、その変更したビット列を新たな取得キーに設定し、その新たな取得キーを用いて、自身の前記取得先選択部によって選択された取得先のルータから、さらにその取得先のルータの前記取得先選択部によって選択された取得先のルータと経路計算結果情報とを取得することを繰り返し、取得されたすべての前記経路計算結果情報の検索キーと前記新たな取得キーとが一致するか否かを判定して、前記経路計算結果情報の検索キーと前記取得キーとが最長一致となるときの前記ルータ集合ごとの経路計算の結果に基づいてパケットの転送先を決定することを、前記取得キーの宛先ＩＰアドレス部分のビット列がすべてゼロになるまで繰り返し実行することを特徴とする。

【００１９】

このような構成によれば、ルータは、転送すべきパケットを受信したとき、まず、自身の記憶部に記憶されている経路計算結果情報の検索キーと取得キーとが一致する経路計算結果情報を抽出する。そして、一致するものがない場合には、他のルータに記憶されている経路計算結果情報を探索して、経路計算結果情報の検索キーと取得キーとが一致する経路計算結果情報を抽出することができる。すなわち、一台当りのルータの記憶容量を抑制するために分散管理されている経路計算結果情報を、取得キーと検索キーとを照合して、検索することが可能となる。また、検索キーと取得キーとが一致する経路計算結果情報が見つからない場合の処理として、取得キーの宛先ＩＰアドレス部分のビットの有効な桁を短くしていくことによって、検索キーと取得キーとが最長一致となる経路計算結果情報を自身および他のルータから抽出することができる。したがって、転送すべきパケットを受信したルータ自身が必要な経路計算結果情報を記憶していなくても、他のルータから転送により適した経路計算結果情報を取得することができる。すなわち、ルータ集合内のルータによって経路計算結果情報が分散管理されるため、一台当りのルータの記憶容量は、すべての経路計算結果情報を記憶していた従来の一台のルータの記憶容量に比べて、低減できる。

【００２０】

本発明は、前記取得先ノード情報が、前記ルータＩＤを複数の範囲に区分し、自身のルータのルータＩＤから前記距離の近いルータＩＤほどその区分の範囲を狭く設定し、前記距離の遠いルータＩＤほどその区分の範囲を広く設定して、それぞれの前記範囲ごとにその範囲に属するルータのルータＩＤを記憶し、前記取得先選択部が、前記取得キーと前記範囲の境界のルータＩＤとを比較して、前記取得キーを含む前記範囲を抽出し、前記取得キーと前記抽出した範囲に属するルータのルータＩＤとの前記距離を算出し、前記距離の近い順に所定数のルータを取得先として選択することを特徴とする。

【００２１】

このような構成によれば、ルータＩＤを複数の範囲に区分して記憶しているため、取得キーを含む範囲に属するルータＩＤと取得キーとの距離計算を行えば良くなる。したがっ

て、ルータに記憶されているすべてのルータＩＤと取得キーとの距離計算を行う必要がなく、計算処理量（処理負荷）を低減することが可能となる。

【００２２】

本発明は、前記経路情報管理システムにおいて用いられる経路情報管理装置の経路情報管理方法であって、前記経路情報管理装置が、前記ＡＳ内の２以上のルータの一まとまりを単位として前記ＡＳ内に複数設定されるルータ集合ごとに、少なくとも所定数以上のルータのルータＩＤとして、当該ルータの属する前記ルータ集合のルータ集合ＩＤのビット列の後にハッシュ関数によって生成されたハッシュ値のビット列を連結したビット列を記憶する格納先ノード情報を格納する記憶部を備え、前記ルータから収集した前記プリフィクスごとに、そのプリフィクスを含む外部経路情報を用いて、前記ＡＳ内の２以上のルータの一まとまりを単位として前記ＡＳ内に複数設定されるルータ集合に属する少なくともいずれか１つのルータにおける、パケットの転送先のプリフィクスに至る転送経路を計算する経路計算を実行する経路計算ステップと、前記ルータ集合ごとに、前記経路計算ステップによって経路計算されたルータ集合ごとの経路計算の結果を含む経路計算結果情報の格納先の検索に用いる検索キーとして、当該ルータ集合のルータ集合ＩＤのビット列、前記プリフィクスのビット列、前記プリフィクスに対するマスク値のビット列、およびパディングとしてすべて「０」のビット列を連結したビット列を生成する検索キー生成ステップと、前記格納先ノード情報に記憶しているルータＩＤと前記検索キーとの距離を、当該ルータＩＤと前記検索キーとのビットごとの排他的論理和を示すビット列において、左から見て最初に「１」が出現したビット位置について右から数えたビットの桁数の数値として算出し、その算出した距離の小さい順に所定数のルータを当該ルータ集合から送信先として選択する格納先選択ステップと、当該経路計算結果情報を、前記格納先選択ステップによって選択された当該所定数のルータに送信する入出力ステップとを実行することを特徴とする。

【００２３】

また、本発明は、前記経路情報管理システムにおいて用いられるルータの経路情報管理方法であって、前記ルータが、前記ＡＳ内の２以上のルータの一まとまりを単位として前記ＡＳ内に複数設定されるルータ集合ごとに、少なくとも所定数以上のルータのルータＩＤとして、当該ルータの属する前記ルータ集合のルータ集合ＩＤのビット列の後にハッシュ関数によって生成されたハッシュ値のビット列を連結したビット列を記憶する取得先ノード情報と、前記経路情報管理装置から送信される検索キーと前記ルータ集合ごとの経路計算の結果とを関連付けた情報を示す経路計算結果情報とを記憶する記憶部を備え、転送すべきパケットを受信したとき、当該ルータの属するルータ集合ＩＤのビット列、当該パケットの宛先ＩＰアドレスのビット列、前記宛先ＩＰアドレスに対するマスク値としてすべて「１」のビット列、およびパディングとしてすべて「０」のビット列を連結して形成される、取得キーを生成する取得キー生成ステップと、前記取得先ノード情報に記憶しているルータＩＤと前記取得キーとの距離を、前記ルータＩＤと前記取得キーとのビットごとの排他的論理和を示すビット列において、左から見て最初に「１」が出現したビット位置について右から数えたビットの桁数の数値として算出し、その算出した距離の小さい順に所定数のルータを当該ルータ集合から取得先として選択する取得先選択ステップと、転送すべきパケットを受信したときに、自身の前記記憶部に記憶する前記経路計算結果情報の検索キーと前記取得キーとが一致するか否かを判定し、一致するものがある場合には、その検索キーに関連付けられた前記ルータ集合ごとの経路計算の結果に基づいてパケットの転送先を決定し、一致するものがない場合には、前記取得先選択部によって取得先として選択されたルータに前記取得キーを送信して、当該ルータの記憶する前記経路計算結果情報の検索キーと前記取得キーとが一致する前記ルータ集合ごとの経路計算の結果を取得して、その取得した前記ルータ集合ごとの経路計算の結果に基づいてパケットの転送先を決定する転送経路決定ステップと、前記転送経路決定ステップにおいて決定された転送先に前記パケットを転送する入出力ステップとを実行することを特徴とする。

【００２４】

このような構成によれば、経路情報管理装置がプリフィクスに係る外部経路情報を格納するため、ルータに保持すべき経路数（記憶容量）を削減することができる。また、経路情報管理装置は、同一のプリフィクスに係る外部経路情報についてすべてを収集して経路計算を実行するため、経路計算結果が最適とならないという問題は発生しない。また、ＡＳ内に複数のルータ集合を設けた場合、そのルータ集合内のＢＧＰ経路はプリフィクスに対して共通に定まるため、経路情報管理装置は、ルータ集合内の１つのルータに対してだけ経路計算を実行すれば良くなる。そのため、経路情報管理装置の処理負荷を低減することができるとともに、ルータでの経路計算が不要となる。また、経路情報管理装置によって算出されたルータ集合ごとの経路計算の結果は、ルータ集合内で分散管理されるため、ルータ集合ごとに所定数のルータに格納（送信）されれば良い。そのため、経路計算の結果をメッセージとして送信するときのメッセージ数を削減することができる。また、ルータは、転送すべきパケットを受信したとき、まず、自身の記憶部に記憶されている経路計算結果情報の検索キーと取得キーとが一致する経路計算結果情報を抽出する。そして、一致するものがない場合には、他のルータに記憶されている経路計算結果情報を探索して、経路計算結果情報の検索キーと取得キーとが一致する経路計算結果情報を抽出することができる。すなわち、一台当りのルータの記憶容量を抑制するために分散管理されている経路計算結果情報を、取得キーと検索キーとを照合して、検索することが可能となる。したがって、転送すべきパケットを受信したルータ自身が必要な経路計算結果情報を記憶していなくても、他のルータからより転送に適した経路計算結果情報を取得することができる。すなわち、ルータ集合内のルータによって経路計算結果情報が分散管理されるため、一台当りのルータの記憶容量は、すべての経路計算結果情報を記憶していた従来の一台のルータの記憶容量に比べて、低減できる。

【００２５】

本発明は、前記経路情報管理方法を、コンピュータとしての経路情報管理装置またはルータに実行させるためのプログラムとした。

【００２６】

このようなプログラムをインストールされたコンピュータは、このプログラムに基づいた機能を実現することができる。

【発明の効果】

【００２７】

本発明によれば、すべての経路情報に基づいて経路計算を実行して最適な転送経路を算出することを満足しつつ、ルータ一台当りに保持すべき経路数（記憶容量）の増大、送信するメッセージ数の増大、および処理負荷の増大を抑制する技術を提供することができる。

【図面の簡単な説明】

【００２８】

【図１】本実施形態の概要を表す図である。

【図２】本実施形態におけるルートサーバの機能の一例を示す図である。

【図３】本実施形態におけるルータの機能の一例を示す図である。

【図４】ルートサーバの決定方法の例を示す図である。

【図５】本実施形態における距離の定義を示す図である。

【図６】ルートサーバの格納先ノードＤＢの一例を示す図である。

【図７】ルータの取得先ノードＤＢの一例を示す図である。

【図８】ルータの経路計算結果ＤＢの一例を示す図である。

【図９】本実施形態における経路計算結果の格納処理の流れを示す図である。

【図１０】本実施形態における他のルータから転送経路情報を取得する処理の流れを示す図である。

【図１１】取得キーｋｄのビット変更時の操作を示す図である。

【発明を実施するための形態】

【００２９】

次に、本発明を実施するための形態（以降「本実施形態」と称す）について、適宜図面を参照しながら詳細に説明する。

【００３０】

≪本実施形態の概要≫

初めに、本実施形態の概要について、図１を用いて説明する。図１に示すように、経路情報管理システム３０は、ＡＳ１内に、外部経路情報の収集およびルーティングの管理を行う２以上のルートサーバ（経路情報管理装置）１０Ａ，１０Ｂ（１０）および境界ルータを含む複数のルータ２０（２０Ａ，２０Ｂ，・・，２０Ｌ）を備えている。そして、各ルートサーバ１０は、ルータ２０と通信可能に接続される。各ルータ２０間は、相互に通信可能に接続される。また、各ルータ２０は、一点鎖線で示される、いずれかのＰｏＰ（Point of Presence）に属している。ＰｏＰは、例えば、同じ市や同じ建物等の同一のロケーションに存在するルータの集合（ルータ集合）を表している。なお、図１では、ＡＳ１は、ＰｏＰ１，ＰｏＰ２，ＰｏＰ３，ＰｏＰ４によって構成されているが、ＰｏＰの数は２以上であれば、４つに限られない。また、ルートサーバ１０の台数、ルータ２０の台数は図１に示した台数に限られない。また、ルータ２０間の接続は、図１に示した実直線に限られない。

【００３１】

ここで、経路情報管理システム３０において、ルータ２０の記憶容量、送信する必要のあるメッセージ数、および処理負荷を抑制するために設けた３つの特徴について説明する。

（特徴１）プリフィクスに係る外部経路情報を、ルートサーバ１０に集約する。

（特徴２）ルートサーバ１０は、ＰｏＰごとにプリフィクスに係る経路計算を実行し、その経路計算の結果を、ＰｏＰごとにそのＰｏＰに属する一部のルータ２０に格納（送信）する。

（特徴３）ＰｏＰ内では、そのＰｏＰに属するルータ２０間で、経路計算結果を分散管理し、パケットの転送に必要な経路計算結果を相互に取得する。

【００３２】

（特徴１について）

図１に示すように、ＡＳ１内のルータ２０Ａは、ＡＳ２からプリフィクスＰ１に係る外部経路情報をｅＢＧＰによって受信したものとする。そして、受信したプリフィクスＰ１に係る外部経路情報は、ルートサーバ１０Ａに収集される。また、ルータ２０Ｋが、ＡＳ３からプリフィクスＰ７に係る外部経路情報をｅＢＧＰによって受信したものとする。そして、受信したプリフィクスＰ７に係る外部経路情報は、ルートサーバ１０Ｂに収集される。すなわち、各ルートサーバ１０Ａ，１０Ｂは、それぞれ、異なるプリフィクスに係る外部経路情報を収集する。なお、どのルートサーバ１０が、どのプリフィクスに係る外部経路情報を収集するかは、予め決められている。そのプリフィクスを収集するルートサーバの決定方法については、後記する。

【００３３】

このようにして、ルートサーバ１０は、いずれかのプリフィクスについてはすべての外部経路情報を収集できるので、すべての外部経路情報に基づいて経路計算を行うことが可能となる。つまり、背景技術の欄で説明したルートリフレクタ（ＲＲ）を用いる方法やＡＳコンフェデレーションを用いる方法に見られるような、一部の外部経路情報に基づいて経路計算を実行してしまうことにともなう、経路計算結果が最適とならないというケースや、経路計算が収束しないというケースは起こらない。また、各ルートサーバ１０がプリフィクスごとに外部経路情報を分散して担当するので、ＡＳ１内のルータ２０の台数の増加に対応してルートサーバ１０の台数を増加することによって、ルートサーバ１０一台当りの記憶容量の増加や処理負荷の増加を抑制することが可能となる。

【００３４】

また、従来の境界ルータは、外部経路情報の受信リストおよび送信リストを記憶するＲＩＢ（Routing Information Base）と、パケットの転送先を記憶するＦＩＢ（Forwarding

Information Base）とを備えていた。しかし、本実施形態の境界ルータ２０では、外部経路情報の受信リストおよび送信リストは、ルートサーバ１０に集約されるため、ＲＩＢが不要となる。すなわち、ルータ２０に保持すべき経路数（記憶容量）を削減することができる。

【００３５】

（特徴２について）

ルートサーバ１０は、プリフィクスごとにＰｏＰ内の１つのルータについて、経路計算を行えば良い。この理由は、以下の通りである。すなわち、通常は、ＡＳ１内のＰｏＰ間に張られるリンクのＩＧＰリンクコストは、ＰｏＰ内のリンクコストより大きい値に設定されている。そして、経路計算においては、経路を構成するリンクごとに割り当てられたリンクコストを合計した値が最も小さい経路が優先して選択される。このようにして、同じＰｏＰ内を宛先とするパケットは、そのＰｏＰから外に出ないようにして、ＰｏＰ間を行ったり来たりするような無駄な転送を防いでいる。そのため、同じＰｏＰ内のルータは、同じプリフィクスを宛先とするＢＧＰ経路については同一の経路（ネクストホップ）をとることになる。つまり、ＰｏＰ内のすべてのルータにとって、ＡＳ外部向けの転送経路は共通になる。そこで、同じプリフィクスを宛先とする経路計算は、ＰｏＰ内の１つのルータに対してだけ行えば良いことになる。以降、これを、"ＰｏＰごとの経路計算"と呼ぶことにする。なお、ＰｏＰごとの経路計算は、ルートサーバ１０が、ルータ２０における経路計算の代わりに行うもので、公知のリンクコストを用いた経路計算手法を用いることができる。ただし、このＰｏＰごとの経路計算は、ルートサーバ１０が、自身を起点として経路計算を実行するのではなく、ＰｏＰ内の１つのルータ２０を起点として実行する点において、公知の経路計算手法と異なるだけである。したがって、ＰｏＰごとの経路計算は、ＡＳ１内のすべてのルータ２０について経路計算が行われていた従来の方法に比べると、処理負荷の低減になっていると言える。

【００３６】

また、ルートサーバ１０は、経路計算結果を各ＰｏＰの一部のルータに格納（送信）する（図１の点線）。これは、後記する特徴３が備わっているために可能となっている。そして、すべてのルータ２０に対して経路計算結果をメッセージとして配信していた従来に比べて、本実施形態では、メッセージ数を削減することができる。

【００３７】

（特徴３について）

境界ルータ２０は、転送すべきパケットを受信して、そのパケットの転送に必要な経路計算結果を自身の記憶部２３（図３参照）に記憶していないとき、自身の属するＰｏＰ内の他のルータ２０から、必要な経路計算結果を取得する（図１の実線矢印）。そして、境界ルータ２０は、取得した経路計算結果に基づいて、受信したパケットを転送することができる。言い換えると、ＰｏＰ内のルータ２０は、ルートサーバ１０から格納（送信）された経路計算結果を分散管理している。すなわち、各ルータ２０は、一部の経路計算結果しか記憶しなくてもよいので、従来より記憶容量を低減することができる。この分散管理の方法の詳細については、後記する。

【００３８】

≪ルートサーバ１０の機能≫

次に、本実施形態の経路情報管理システム３０に用いられるルートサーバ１０の機能について、図２を用いて説明する。なお、図２は、本実施形態のルートサーバ１０の一部の機能のみを示し、ＢＧＰやＩＧＰ等に係る機能については記載を省略している。

【００３９】

図２に示すように、ルートサーバ１０は、外部経路情報や経路計算結果等の各種情報の入出力を司る入出力部１１と、外部経路情報に基づいて経路計算する処理部１２と、記憶部１３とを備える。入出力部１１は、他のルータ２０等との情報を入出力する。入出力部１１における情報の入出力は、処理部１２からの指示に基づいて行われる。

【００４０】

処理部１２は、ルートサーバ１０が備える図示しないＣＰＵ（Central Processing Unit）と図示しないメインメモリとによって構成されている。そして、このＣＰＵがメインメモリにアプリケーションプログラムを展開して、それを実行することにより種々の演算処理を具現化する。なお、アプリケーションプログラムは、記憶部１３に格納されている。また、処理部１２は、外部経路情報収集部１２１、経路計算部１２２、格納先選択部１２３、および検索キー生成部１２４を備えている。

【００４１】

外部経路情報収集部１２１は、境界ルータ２０から広告されてくる外部経路情報を入出力部１１を介して受信する。なお、同一のプリフィクスに係る外部経路情報は、すべて、同じルートサーバ１０に収集される。そして、外部経路情報収集部１２１は、受信した外部経路情報を記憶部１３の外部経路情報ＤＢ１３１に記憶する。経路計算部１２２は、外部経路情報ＤＢ１３１に新たに記憶された外部経路情報を読み出して、プリフィクスごとに前記した"ＰｏＰごとの経路計算"を実行し、ＰｏＰごとの経路計算の結果、すなわち最適な転送経路および次善の転送経路を算出する。予め次善の転送経路を算出しておく理由は、最適な転送経路が輻輳や故障等によって利用できない場合に、直ちにこの次善の転送経路に切り替えて用いられるようにするためである。

【００４２】

格納先選択部１２３は、格納先ノードＤＢ１３２に記憶されているルータ２０のルータＩＤと、後記するプリフィクスを含む検索キーとを用いて、経路計算部１２２によって算出されたＰｏＰごとの経路計算の結果の格納（送信）先であるＰｏＰ内のルータ２０を選択する。そして、格納先選択部１２３は、入出力部１１を介し、ＰｏＰごとの経路計算の結果を、ＰｏＰごとに選択したルータ２０に格納（送信）する。また、検索キー生成部１２４は、ＰｏＰごとの経路計算の結果を格納するときに用いる検索キーを生成する。なお、格納方法および検索キーの生成については、後記する。

【００４３】

記憶部１３は、外部経路情報ＤＢ１３１および格納先ノードＤＢ１３２を格納している。外部経路情報ＤＢ１３１は、境界ルータ２０から広告されたプリフィクスに係る外部経路情報を記憶している。例えば、外部経路情報は、プリフィクスと、そのプリフィクスに関連付けられたパス属性情報とを含んでいる。パス属性情報は、AS-pathやネクストホップに至るリンクに割り当てられたリンクコスト等である。なお、AS-pathは、経由してきたＡＳを記録しておくためのものであり、ＡＳを経由するごとにＡＳ番号が追加される。そして、AS-pathに記録されたＡＳ番号の数が少ない方の経路が優先して用いられる。外部経路情報ＤＢ１３１に記憶されている外部経路情報は、経路計算部１２２によって、経路計算を実行するときに用いられる。また、格納先ノードＤＢ１３２は、ルータ２０を識別可能なルータＩＤを記憶している。このルータＩＤは、ＡＳの管理者によって予め設定される。そして、格納先ノードＤＢ１３２は、格納先選択部１２３が格納先のルータ２０を選択するときに参照される。格納先ノードＤＢ１３２の具体例については、後記する。

【００４４】

≪ルータ２０の機能≫

次に、本実施形態のルータ２０の機能について、図３を用いて説明する。なお、図３は、本実施形態のルータ２０の一部の機能のみを示し、ＢＧＰやＩＧＰ等に係る機能については記載を省略している。

【００４５】

図３に示すように、ルータ２０は、ルートサーバ１０や他のルータ２０との間で送受信される情報の入出力を制御する入出力部２１と、処理部２２と、記憶部２３とを備える。入出力部２１における情報の入出力は、処理部２２からの指示に基づいて行われる。

【００４６】

処理部２２は、ルータ２０が備える図示しないＣＰＵ（Central Processing Unit）と図示しないメインメモリによって構成されている。そして、このＣＰＵがメインメモリにアプリケーションプログラムを展開して、それを実行することにより種々の演算処理を具

現化する。なお、アプリケーションプログラムは、記憶部２３に格納されている。また、処理部２２は、ルートサーバ選択部２２１、転送経路決定部２２２、取得先選択部２２３、カプセル化処理部２２４、および取得キー生成部２２５を備えている。

【００４７】

ルートサーバ選択部２２１は、記憶部２３のルートサーバＩＤＤＢ２３１に記憶されているルートサーバＩＤ情報を参照して、後記するルートサーバ決定方法に基づいて、他のＡＳから受信した外部経路情報を広告するルートサーバ１０を選択する。そして、ルートサーバ選択部２２１は、入出力部２１を介して、他のＡＳから受信した外部経路情報を、選択したルートサーバ１０へ広告する。

【００４８】

転送経路決定部２２２は、転送すべきパケットを受信したときに、記憶部２３の経路計算結果ＤＢ２３２に記憶している経路計算結果情報を探索し、そのパケットの転送先を決定する。しかし、経路計算結果ＤＢ２３２は、経路計算に必要な経路計算結果情報の一部しか記憶していない場合がある。その場合には、転送経路決定部２２２は、取得先選択部２２３によって選択された他のルータ２０から不足している経路計算結果情報を取得する。そして、転送経路決定部２２２は、その取得した経路計算結果情報に基づいて、パケットの転送経路（転送先）を決定する。

【００４９】

取得先選択部２２３は、転送すべきパケットを受信したときに、取得先ノードＤＢ２３３に記憶しているルータ２０に係る情報を参照して、不足する経路計算結果情報を保持するルータ２０を選択する。また、カプセル化処理部２２４は、転送すべきパケットに対して、出口となる境界ルータ２０のアドレスのヘッダでカプセル化する処理を実行する。また、出口の境界ルータ２０では、カプセル化処理部２２４は、カプセル化されたパケットから、その出口の境界ルータ２０のアドレスのヘッダを取り外す処理を実行する。また、取得キー生成部２２５は、経路計算結果情報の取得先のルータ２０を選択するときに用いる取得キーを生成する。なお、取得先の選択方法および取得キーの生成方法については、後記する。

【００５０】

記憶部２３は、ルートサーバ１０から配信される経路計算結果情報を格納する経路計算結果ＤＢ２３２、ルートサーバ１０のＩＤ番号を格納するルートサーバＩＤＤＢ２３１、および経路計算結果情報を分散管理するルータ２０に係る情報を格納する取得先ノードＤＢ２３３を格納する。経路計算結果ＤＢ２３２は、転送経路決定部２２２の経路計算において参照される。ルートサーバＩＤＤＢ２３１は、ルートサーバ選択部２２１が、外部経路情報を広告するルートサーバ１０を選択するときに参照される。取得先ノードＤＢ２３３は、取得先選択部２２３が必要な経路計算結果情報の取得先のルータ２０を選択するときに参照される。経路計算結果ＤＢ２３２、ルートサーバＩＤＤＢ２３１、および取得先ノードＤＢ２３３の具体例については、後記する。

【００５１】

≪ルートサーバ決定方法≫

ここで、ルータ２０におけるルートサーバ決定方法について、図４を用いて説明する（適宜図３参照）。図４は、○印で示すルートサーバ１０とそれをカバーするハッシュ空間を表している。図４中の○印の中の数字は、ルートサーバ１０のＩＤ番号を示している。そして、図４は、ハッシュ空間に、３つのルートサーバ１０Ａ，１０Ｂ，１０Ｃが存在する場合を表している。例えば、ルートサーバ１０ＡのＩＤ番号が５、ルートサーバ１０ＢのＩＤ番号が１０、ルートサーバ１０ＣのＩＤ番号が１５、であるものとして説明する。なお、これらのルートサーバ１０Ａ，１０Ｂ，１０ＣのＩＤ番号に係る情報が、図３の記憶部２３のルートサーバＩＤＤＢ２３１に記憶されることになる。

【００５２】

まず、next(m)という関数を導入する。next(m)はハッシュ空間においてハッシュ空間を時計回り（つまりＩＤを大きくする方向）で次のノードのＩＤを返す。例えば、next(2

)=5、next(6)=10、next(15)=15となる。すなわち、next(m)は、ハッシュ空間のある値m以上の範囲において最小の値となるＩＤを返す。ここで、mは、プリフィクスの値をハッシュ関数に代入して算出されたハッシュ値である。ただし、図４に示すハッシュ空間では、０＜m≦１５である。

【００５３】

そして、図４では、ハッシュ空間におけるルートサーバ１０Ａの分担範囲は０超え５以下、ルートサーバ１０Ｂの分担範囲は５超え１０以下、ルートサーバ１０Ｃの分担範囲は１０超え１５以下となる。このようにして、ルートサーバ１０のＩＤ番号を予め決めておけば、プリフィクスに関する経路情報を送信するルートサーバ１０を特定することが可能となる。

【００５４】

≪経路計算結果情報の分散管理≫

前記したように、ＰｏＰ内のルータ２０がＢＧＰ経路については同一の経路をとるので、経路計算結果情報をＰｏＰ内のルータ２０によって分散管理することが可能となる。そして、経路計算結果情報を分散管理することによって、ルータ２０が保持すべき経路数（記憶容量）を抑制することが可能となる。以下に、分散管理の方法について説明する。

【００５５】

（フレームワーク）

まず、分散管理のためのフレームワークについて説明する。ＡＳ内のルータ２０には、そのルータを識別するために、ＡＳの管理者によってユニークなルータＩＤが割り当てられる。また、経路計算結果情報は、分散管理が可能なように、検索キーと経路計算結果の転送経路情報（最適な転送経路および次善の転送経路）とが紐付けられて形成される。ただし、ルータＩＤと検索キーとは、同じビット長で表現される。

【００５６】

まず、検索キーは、ルートサーバ１０において、経路計算結果情報を、どのルータ２０に格納（送信）するかの選択に用いられる。そして、ルートサーバ１０は、後記する検索キーとルータＩＤとの距離に基づいて、検索キーに距離の近いルータＩＤのルータを、距離の近い順にｋ個選択する。ｋ個選択する意味は、冗長性の確保のためである。したがって、冗長性を考えなければ、ｋ＝１であっても良い。そして、ルートサーバ１０は、選択したｋ個のルータ２０に、経路計算結果情報を格納（送信）する。

【００５７】

ここで、距離の定義について、図５を用いて説明する。図５では、説明を簡単にするために、ルータＩＤおよび検索キーのビット長をそれぞれ５ビットで表している。各ルータａ，ｂ，ｃ，ｄのルータＩＤがそれぞれ10011，10101，10110，11000であったとする。そして、検索キーが、10111であったとする。この場合、ルータＩＤと検索キーとのExclusive ORを算出する。そして、Exclusive ORの演算結果のビット列において、左から見て最初に「１」が出現した位置について右から数えた桁数の数値を、「距離」と定義する。すなわち、各ルータａ，ｂ，ｃ，ｄに対して、ルータＩＤと検索キーとの距離は、それぞれ３，２，１，４となる。したがって、検索キーはルータｃに最も近いと判定される。そして、冗長性を考慮してｋ＝２とした場合には、ルータｃの次に距離が近いのは、ルータｂと判定される。このようにして、検索キーおよびその検索キーに紐付けられた転送経路情報によって形成される経路計算結果情報は、ルータｂ，ｃに格納されることになる。この距離を用いることによって、距離の近いものほど、左から数えたビット列の一致する長さが長くなり、距離が０の場合は、完全に一致することを表す効果がある。

【００５８】

次に、ルータ２０が、他のルータ２０から必要な経路計算結果情報を取得するときに、取得先のルータ２０を選択する方法について説明する。この場合、経路計算結果情報を格納するときと同じ距離を用いて、必要な経路計算結果情報を格納したルータ２０を特定することができる。以下に、ルータＩＤ、検索キー、および転送経路情報の構成と、格納先ノードＤＢ１３２、取得先ノードＤＢ２３３、および経路計算結果ＤＢ２３２の一例につ

いて、図６，７，８を用いて説明する。なお、本実施形態では、例えば、ビット長は１６０ビットとして説明する。

【００５９】

ルータＩＤの構成を式（１）に示す。

＜ルータＩＤ＞：：＝＜ＰｏＰのＩＤ＞＜ハッシュ値＞　　・・式（１）

ただし、ＰｏＰのＩＤは、そのルータ２０が属するＰｏＰを識別するＩＤであり、３２ビットとする。ＰｏＰのＩＤは、ＡＳ管理者によって設定されるユニークな識別情報である。また、ハッシュ値は、例えば、ハッシュ関数の一つである、ランダム値を基にしたＭＤ５（Message Digest 5）により算出したＭＤ５ハッシュ値であって、１２８ビットとする。

【００６０】

経路計算結果情報の格納先のルータ２０を選択するときに用いる検索キーｋｒの構成を式（２）に示す。

＜ｋｒ＞：：＝＜ＰｏＰのＩＤ＞＜プリフィクス値＞＜マスク値＞＜パディング＞

・・式（２）

ただし、ＰｏＰのＩＤは、前記と同様である。つまり、ＰｏＰのＩＤによって、どのＰｏＰのための検索キーであるかが特定され、同じＰｏＰのＩＤを含むノードＩＤを見つけやすくすることができる。プリフィクス値は、外部経路情報に含まれるプリフィクスであり、３２ビットのＩＰｖ４（Internet Protocol version 4）である。マスク値は、プリフィクス値に対するマスク値であり、３２ビットである。このマスク値は、ＩＰｖ４で用いられるサブネットマスクに相当する。パディングは、その値をすべて「０」とし、６４ビットである。なお、「：：＝」は、左辺のビット列と右辺のビット列とが同じになることを表し、右辺は＜＞のビット列をこの順で連結することを表している。

【００６１】

経路計算結果情報を形成する、検索キーｋｒに紐付けられる転送経路情報ｖａｌｕｅの構成を式（３）に示す。

＜ｖａｌｕｅ＞：：＝＜バージョン番号＞＜ｎ＞＜ＮＨ_１＞＜優先度１＞・・

・・＜ＮＨ_ｎ＞＜優先度ｎ＞　　・・式（３）

ここで、バージョン番号は、ルートサーバ１０が経路計算を実行するごとに付与される。ｎは、ｖａｌｕｅに格納する経路計算の結果である転送経路（ＮＨ＿＊）の数である。ＮＨ＿＊は、プリフィクスに対応する転送経路（ネクストホップ情報）を表す。優先度は、転送経路を用いる優先順位を指定するもので、優先度の高いものから順に用いられる。なお、転送経路情報ｖａｌｕｅは、複数の転送経路を格納しているので、選択された転送経路の障害時には、次に優先度の高い転送経路を直ちに用いることができる。

【００６２】

次に、ルートサーバ１０の記憶部１３の格納先ノードＤＢ１３２（図２参照）の一例を、図６を用いて説明する。格納先ノードＤＢ１３２は、ＰｏＰごとに、そのＰｏＰに属するルータ２０の、ルータ名、ＩＰアドレス、およびルータＩＤを記憶する。なお、ＰｏＰ内では経路計算結果情報が分散管理されるため、格納先ノードＤＢ１３２は、ＡＳ内のすべてのルータ２０に係る情報について記憶しなくても良い。

【００６３】

また、ルータ２０の記憶部２３の取得先ノードＤＢ２３３（図３参照）の一例を、図７を用いて説明する。なお、図７に示す取得先ノードＤＢ２３３は、ある一つのルータ２０に記憶されているものを表している。ｒａｎｇｅは、ルータＩＤをいくつかに区分した範囲を示している。例えば、図４に示すハッシュ空間がルータＩＤとして用いられるＩＤ全体を示すものとすると、ルートサーバ１０の分担範囲がｒａｎｇｅに相当する。そして、その範囲は、自身のルータＩＤに距離の近い範囲は狭く、距離の遠い範囲は広く設定される。図７では、冗長性をもたせて、取得先のルータの台数ｋ＝２とした場合を表しており、ｒａｎｇｅごとに、２つの取得先のルータに係る情報を記憶した状態を表している。

【００６４】

また、ルータ２０の記憶部２３の経路計算結果ＤＢ２３２（図３参照）の一例を、図８を用いて説明する。経路計算結果ＤＢ２３２は、ＰｏＰ内の一部の経路計算結果情報を記憶している。なお、経路計算結果情報は、検索キーｋｒと転送経路情報ｖａｌｕｅとの組（ｋｒ，ｖａｌｕｅ）によって表される。

【００６５】

（ルートサーバ１０の格納処理の流れ）

次に、本実施形態における、ルートサーバ１０による格納処理の流れについて、図９を用いて説明する（適宜図２参照）。ステップＳ９０１では、経路計算部１２２が、外部経路情報ＤＢ１３１を参照して、各プリフィクスに対してＰｏＰごとの経路計算を実行し、転送経路情報ｖａｌｕｅを生成する。ステップＳ９０２では、検索キー生成部１２４が、ルータ２０から受信したプリフィクスについて、そのプリフィクスを含む検索キーＫｒを生成する。ステップＳ９０３では、格納先選択部１２３が、ＰｏＰごとに、検索キーｋｒとルータＩＤとの距離の近い（小さい）順にｋ個（予め決められている所定数）のルータＩＤを抽出する。そして、ステップＳ９０４では、格納先選択部１２３が、抽出したルータＩＤのルータ２０に、そのルータ２０が記憶している取得先のルータに係る情報をｋ個取得する要求ｆｉｎｄ＿ｎｏｄｅ（ｋｒ）を送信する。なお、ｆｉｎｄ＿ｎｏｄｅ（ｋｒ）を受信したルータ２０の取得先選択部２２３は、検索キーｋｒとｒａｎｇｅ（図７参照）の境界のルータＩＤとを比較して、検索キーｋｒを含むｒａｎｇｅを抽出し、検索キーｋｒと抽出したｒａｎｇｅに属するルータ２０それぞれのルータＩＤとの距離を算出し、距離の近い順にｋ個の取得先のルータに係る情報を、ルートサーバ１０に返信する。

【００６６】

ステップＳ９０５では、格納先選択部１２３は、ｆｉｎｄ＿ｎｏｄｅ（ｋｒ）の送信先のルータ２０から、取得先のルータに係る情報をｋ個取得する。ステップＳ９０６では、格納先選択部１２３は、ステップＳ９０５で取得したルータの中で、ｆｉｎｄ＿ｎｏｄｅ（ｋｒ）を未送信のルータがあるか否かを判定する。そして、未送信のルータがある場合（ステップＳ９０６でＹｅｓ）、ステップＳ９０７で、格納先選択部１２３が、未送信のルータに、ｆｉｎｄ＿ｎｏｄｅ（ｋｒ）を送信する。そして、処理は、ステップＳ９０５へ戻る。また、未送信のルータがない場合（ステップＳ９０６でＮｏ）、ステップＳ９０８では、格納先選択部１２３は、ＰｏＰごとに、検索キーｋｒに距離の近い順にｋ個のルータＩＤを選択し、選択したｋ個のルータＩＤのルータに、経路計算結果情報（ｋｒ，ｖａｌｕｅ）を格納（送信）する。なお、ステップＳ９０８で格納された経路計算結果情報（ｋｒ，ｖａｌｕｅ）は、ルートサーバ１０が経路計算を再実行した場合には、削除される。また、ルートサーバ１０が、ｆｉｎｄ＿ｎｏｄｅ（ｋｒ）の送信先から新しく取得したルータＩＤは、格納先ノードＤＢ１３２に格納し、更新していく。

【００６７】

（他のルータ２０から転送経路情報を取得する処理）

次に、他のルータ２０から転送経路情報を取得する処理について、その概要を説明する（適宜図３参照）。まず、ルータ２０の取得キー生成部２２５は、転送すべきパケットが到着すると、転送先の決定のために、転送すべきパケットの宛先ＩＰアドレスを含む取得キーｋｄを生成する。なお、この取得キーｋｄの詳細については、後記する。そして、転送経路決定部２２２は、自身の記憶する経路計算結果ＤＢ２３２を参照して、取得キーｋｄと一致する、経路計算結果情報に含まれる検索キーｋｒを探索する。取得キーｋｄに一致する検索キーｋｒが見つかった場合には、その検索キーｋｒに紐付けられた転送経路情報ｖａｌｕｅが読み出され、パケットの転送に用いられる。

【００６８】

また、取得先選択部２２３は、転送すべきパケットが到着すると、取得キーｋｄに距離の近いｋ個のルータＩＤのルータ２０を選択する。なお、この距離は、前記した距離と同様の定義である。そして、取得先選択部２２３は、その選択したルータ２０に、そのルータ２０が保持する取得キーｋｄに距離の近いｋ個のルータに係る情報と、取得キーｋｄに一致する検索キーｋｒの経路計算結果情報（ｋｒ，ｖａｌｕｅ）がある場合には経路計算

結果情報（ｋｒ，ｖａｌｕｅ）とを取得するための要求ｆｉｎｄ＿ｖａｌｕｅ（ｋｄ）を送信する。そして、転送経路決定部２２２は、取得した複数の経路計算結果情報（ｋｒ，ｖａｌｕｅ）の中から、取得キーｋｄに最長一致（例えば、複数の検索キーｋｒと取得キーｋｄとの距離を算出したとき、最も距離の近いビット列を見つけること）となる検索キーｋｒを含む経路計算結果情報（ｋｒ，ｖａｌｕｅ）を選択する。次に、転送経路決定部２２２は、選択した該経路計算結果（ｋｒ，ｖａｌｕｅ）から転送経路情報ｖａｌｕｅを読み出して、パケットの転送に用いる。

【００６９】

（取得キーｋｄの構成）

ここで、取得キーｋｄの構成を式（４）に示す。

＜ｋｄ＞：：＝＜ＰｏＰのＩＤ＞＜ＩＰアドレス＞＜マスク値＞＜パディング＞

・・式（４）

ただし、ＰｏＰのＩＤは、転送すべきパケットを受信したルータの属するＰｏＰのＩＤである。つまり、ＰｏＰのＩＤによって、どのＰｏＰのための取得キーであるかが特定される。ＩＰアドレスは、転送すべきパケットの宛先ＩＰアドレスであり、３２ビットである。マスク値は、プリフィクス値に対するマスク値であり、３２ビットである。このマスク値は、ＩＰｖ４で用いられるサブネットマスクに相当し、すべて「１」に設定する。パディングは、その値をすべて「０」とし、６４ビットである。

【００７０】

このように、境界ルータ２０では、転送すべきパケットが到着すると、転送先の決定のために、分散管理されている転送経路情報ｖａｌｕｅを取得する必要がある。そのため、境界ルータ２０は、転送経路情報ｖａｌｕｅとの組を形成している検索キーｋｒを生成できればよいが、転送すべきパケットからは、宛先ＩＰアドレスしか取得できない。そして、その宛先ＩＰアドレスを示すビット列と、宛先のプリフィクスを示すビット列とは完全に一致するとは限らない。すなわち、境界ルータ２０は、取得キーｋｄを生成する場合、経路計算結果情報（ｋｒ，ｖａｌｕｅ）を格納するときに用いた検索キーｋｒと同じビット列を生成できるとは限らない。そこで、ルータ２０が、取得キーｋｄを用いて行う転送経路情報の取得処理の流れについて、図１０を用いて説明する（適宜図３参照）。

【００７１】

（転送経路情報の取得処理の流れ）

図１０に示すように、ステップＳ１００１では、取得キー生成部２２５が、転送すべきパケットの宛先ＩＰアドレスを含む取得キーｋｄを生成する。ステップＳ１００２では、取得先選択部２２３が、取得キーｋｄと取得先ノードＤＢ２３３に記憶されているルータＩＤとを比較して、取得キーｋｄに距離の近い順にｋ個のルータを選択する。ステップＳ１００３では、取得先選択部２２３が、選択したルータに、取得キーｋｄに距離の近いｋ個のルータに係る情報と、取得キーｋｄに一致する検索キーｋｒが存在すれば、その検索キーｋｒを含む経路計算結果情報（ｋｒ，ｖａｌｕｅ）の転送経路情報ｖａｌｕｅとを取得するための要求ｆｉｎｄ＿ｖａｌｕｅ（ｋｄ）を送信する。

【００７２】

ステップＳ１００４では、転送経路決定部２２２は、ｆｉｎｄ＿ｖａｌｕｅ（ｋｄ）の送信先のルータ２０から、取得キーｋｄに距離の近いｋ個のルータ２０に係る情報と、取得キーｋｄに一致する検索キーｋｒが存在すれば、その検索キーｋｒを含む経路計算結果情報（ｋｒ，ｖａｌｕｅ）の転送経路情報ｖａｌｕｅとを取得する。なお、ｆｉｎｄ＿ｖａｌｕｅ（ｋｄ）を受信したルータ２０の取得先選択部２２３は、取得キーｋｄとｒａｎｇｅ（図７参照）の境界のルータＩＤとを比較して、取得キーｋｄを含むｒａｎｇｅを抽出し、取得キーｋｄと抽出したｒａｎｇｅに属するルータ２０それぞれのルータＩＤとの距離を算出し、距離の近い順にｋ個の取得先のルータに係る情報を、送信元のルータ２０に返信する。

【００７３】

ステップＳ１００５では、転送経路決定部２２２が、転送経路情報ｖａｌｕｅが取得さ

れたか否かを判定する。そして、転送経路情報ｖａｌｕｅが取得された場合（ステップＳ１００５でＹｅｓ）、ステップＳ１００６では、転送経路決定部２２２が、転送経路情報ｖａｌｕｅに含まれる転送経路（式（３）におけるＮＨ＿＊）と、現在使われている転送経路とのどちらが、転送すべきパケットのＩＰアドレスと最長一致かを比較し、最長一致の方の転送経路を現在使われている転送経路に設定する。ただし、ステップＳ１００５で取得された転送経路情報ｖａｌｕｅに含まれる転送経路および現在使われている転送経路のいずれもが、転送すべきパケットのＩＰアドレスとの距離が同じ場合には、転送経路情報ｖａｌｕｅのバージョン番号が最新の方に、現在使われている転送経路を設定する。そして、ステップＳ１００７では、転送経路決定部２２２が、現在使われている転送経路を不図示のＦＩＢに設定する。そして、処理は、ステップＳ１００８へ進む。なお、ステップＳ１００５で、転送経路情報ｖａｌｕｅが取得されなかった場合（ステップＳ１００５でＮｏ）、処理は、ステップＳ１００８へスキップする。

【００７４】

ステップＳ１００８では、転送経路決定部２２２が、ｆｉｎｄ＿ｖａｌｕｅ（ｋｄ）を、未送信のステップＳ１００４で取得したルータがあるか否かを判定する。未送信のステップＳ１００４で取得したルータがある場合（ステップＳ１００８でＹｅｓ）、ステップＳ１００９では、取得先選択部２２３が、未送信のルータ２０へ、ｆｉｎｄ＿ｖａｌｕｅ（ｋｄ）を送信する。そして、処理は、ステップＳ１００４へ戻る。また、未送信のステップＳ１００４で取得したルータがない場合（ステップＳ１００８でＮｏ）、処理は、ステップＳ１０１０へ進む。

【００７５】

ステップＳ１０１０では、転送経路決定部２２２が、取得キーｋｄのＩＰアドレス部の最後（一番右）のゼロでないビットをゼロに変更し、同時にマスク値部分の当該ＩＰアドレスの変更位置のビットを含む右側のビットをゼロに変更し、その変更したビット列を新たな取得キーｋｄに設定する。この取得キーｋｄのビット変更時の操作の具体例について、図１１を用いて説明する。図１１に示すように、ビット変更前のＩＰアドレス部のビット列およびマスク値部分のビット列がそれぞれ＜１０１０１１００＞＜１１１１１１１１＞であったとする。ＩＰアドレス部の最後（一番右）のゼロでないビットは、右から３番目である。この右から３番目のビットをゼロに変更する。そして、マスク値部分については、右から３番目のビットを含む右側の３つのビットをゼロに変更する。

【００７６】

ステップＳ１０１１では、転送経路決定部２２２が、取得キーｋｄのＩＰアドレス部のホスト部分がすべてゼロかそれ以外かを判定する。取得キーｋｄのＩＰアドレス部のホスト部分の一部がゼロでない場合（ステップＳ１０１１でＮｏ）、処理は、ステップＳ１００２へ戻る。また、取得キーｋｄのＩＰアドレス部のホスト部分がすべてゼロの場合（ステップＳ１０１１でＹｅｓ）、ステップＳ１０１２では、転送経路決定部２２２が、現在使われている転送経路を含む経路計算結果情報のレプリケーションを、当初の取得キーｋｄに（ｋ＋１）番目に距離の近いルータＩＤのルータに対して実行する。そして、転送経路情報の取得処理を終了する。

【００７７】

なお、転送すべきパケットは、ステップＳ１００７でＦＩＢに設定された転送経路に基づいて転送される。この際、カプセル化処理部２２４は、転送すべきパケットを、出口の境界ルータのＩＰアドレスを宛先とするヘッダによってカプセル化して、ＩＧＰを用いて、そのパケットを転送する。このカプセル化の効果は、ルータ２０が他のＰｏＰのＢＧＰ経路情報を保持する必要がなくなること、およびＡＳ内で経由するルータ２０が意図しない方路へパケットを転送することを防止できること、である。

【００７８】

また、ステップＳ１０１０において取得キーｋｄのＩＰアドレス部のビット列を、最後（一番右）のゼロでないビットをゼロに変更し、再びステップＳ１００２へ戻って検索を続ける理由は、経路計算結果情報（ｋｒ，ｖａｌｕｅ）は、宛先ＩＰアドレスの集合であ

るプリフィクスに基づいて管理されているのに対して、取得キーｋｄは宛先ＩＰアドレスを用いて生成されているからである。すなわち、取得キーｋｄにそのまま一致する検索キーｋｒが見つかる可能性は低いと考えられる。そこで、ＩＰアドレス部のビット列を、最後（一番右）のゼロでないビットをゼロに変更しつつ、最長一致となるビット列を探索していくためである。

【００７９】

また、ステップＳ１００６において、最長一致となる転送経路が複数見つかった場合には、転送経路情報ｖａｌｕｅに含まれるバージョン番号が最新の方が選ばれる。また、ステップＳ１０１２において、レプリケーションする経路計算結果情報は、ｋ＋１番目に距離の近いルータ２０に格納されるが、ｋ＋１番目のルータ２０にすでに格納済みであれば、ｋ＋２番目のルータにレプリケーションされる。そして、レプリケーションされた経路計算結果情報は、所定時間が経過した後（タイムアウト後）、消去される。なお、その所定時間は、取得キーｋｄとルータＩＤとの距離の大きさに反比例して設定される。すなわち、タイムアウトの所定時間は、距離が遠いほど短く設定される。

【００８０】

（記憶容量の削減効果）

前記したように、本実施形態では、ルータ２０が記憶していた外部経路情報がルートサーバ１０に記憶されるため、従来のルータに比べて、記憶容量を低減することが可能である。また、ルートサーバ１０も、プリフィクスを分担して記憶する台数を増加することによって、一台当りの記憶容量を低減することができる。さらに、従来の各ルータでは、パケット転送のために必要な転送経路の情報（ＦＩＢに記憶される情報）については、すべてのプリフィクスに対して記憶しなければならなかったが、本実施形態では、ＰｏＰ内で分散管理することによって、ルータ２０の記憶容量を低減することが可能である。定量的には、ルータ２０一台当りの保持する経路数は、全経路数をＰｏＰ内のルータ数で除算した値に、レプリケーションされる情報の経路数を加算して算出される。なお、インターネットにおいては、トラヒックのほとんどは少数の宛先に集中する傾向があるため、レプリケーションされる経路数は、全経路数をＰｏＰ内のルータ数で除算した値に比較して小さい。

【００８１】

（メッセージ数および計算処理量の削減効果）

ルートサーバ１０が経路計算を実行するため、各ルータ２０は、経路計算を実行しないで良い。そのため、本実施形態のルータ２０では、従来のルータに比較して、計算処理量が低減できる。さらに、本実施形態では、ＡＳ内のルータ２０は、複数のＰｏＰのいずれかに属している。そして、ＰｏＰ内ではＢＧＰ経路については同一の転送経路となるので、ルートサーバ１０は、ＰｏＰごとに１つのルータ２０のための経路計算を実行すれば良く、また、その経路計算結果をメッセージとしてＰｏＰ内のルータ数より少ないｋ個のルータ２０に送信すれば良い。したがって、ルートサーバ１０が、ルータ２０ごとにすべての経路計算結果を送信する場合に比べて、定性的には、メッセージ数および計算処理量（処理負荷）を低減することができる。

【００８２】

定量的には、ＰｏＰ内のルータ数の平均をＲｐｏｐ、ルートサーバ１０からすべてのルータ２０に経路計算結果（メッセージ）を送信する場合のメッセージ数をＭとすると、経路計算結果の配信のために、ルートサーバ１０が送信するメッセージ数Ｒｍは、Ｍ／Ｒｐｏｐ＋Ｎｆとなる。ここで、Ｎｆは、ｆｉｎｄ＿ｎｏｄｅメッセージ数である。このＮｆは、初期状態では、最大でＰｏＰ内のルータ数分になることもあるが、ｆｉｎｄ＿ｎｏｄｅが繰り替えされることによって、ルートサーバ１０の格納先ノードＤＢ１３２が更新され、ルートサーバ１０が知り得ないルータ２０が無くなっていくため、最終的には、ｋ（冗長性）に収束する。そのため、メッセージ数は、削減される。

【００８３】

また、ｆｉｎｄ＿ｖａｌｕｅメッセージ数は、転送すべきパケットの宛先の分布に依存

する。インターネットでは、パケットの宛先のほとんどが、ごく一部のプリフィクスに集中することから、レプリケーションによって、他のルータに取得要求を行う必要がなくなっていく。したがって、ｆｉｎｄ＿ｖａｌｕｅメッセージ数は、ほとんどゼロに収束する。したがって、本実施形態では、ＰｏＰおよびレプリケーションの導入にともなって、メッセージ数を低減することができる。

【００８４】

以上、本実施形態の経路情報管理システム３０によれば、ルートサーバ１０がプリフィクスに係る外部経路情報を格納するため、ルータ２０に保持すべき経路数（記憶容量）を削減することができる。また、ルートサーバ１０は、同一のプリフィクスに係る外部経路情報についてすべてを収集して経路計算を実行するため、経路計算結果が最適とならないという問題は発生しない。また、ＡＳ内に複数のＰｏＰを設けた場合、そのＰｏＰ内のＢＧＰ経路はプリフィクスに対して共通になるため、ルートサーバ１０は、ＰｏＰ内の１つのルータ２０に対してだけ経路計算を実行すれば良くなる。そのため、ルートサーバ１０の処理負荷を低減することができるとともに、ルータ２０での経路計算が不要となる。また、ルートサーバ１０によって算出されたＰｏＰごとの経路計算の結果は、ＰｏＰ内で分散管理されるため、ＰｏＰごとにｋ個（所定数）のルータ２０に格納（送信）されれば良い。そのため、経路計算の結果をメッセージとして送信するときのメッセージ数を削減することができる。また、ＰｏＰ内のルータ２０によって経路計算結果情報が分散管理されるため、一台当りのルータ２０の記憶容量は、すべての経路計算結果情報を記憶していた従来の一台のルータ２０の記憶容量に比べて、低減できる。また、必要な経路計算結果の取得のための探索は、ＰｏＰ内のルータ２０だけに限られるため、その取得処理が早くなる。

【００８５】

なお、本実施形態は、これらに限定されるものではなく、その趣旨を変えない範囲で実施することができる。例えば、取得先選択部２２３が検索キーｋｒや取得キーｋｄに距離の近いｋ個のルータ２０を選択するときに、本実施形態では検索キーｋｒや取得キーｋｄを含むｒａｎｇｅで選択されているｋ個のルータを取得するものとして説明した。しかし、ルータ２０の記憶部２３に記憶されているルータＩＤと検索キーｋｒや取得キーｋｄとの距離を計算して、最も距離の近いｋ個のルータ２０を選択するようにしても構わない。また、図７では、一例として、記憶部２３の取得先ノードＤＢ２３３にｒａｎｇｅごとにｋ個のルータに係る情報を格納する場合を示した。しかし、ｒａｎｇｅごとにｋ個より多くのルータ２０に係る情報が記憶されている場合には、その中から最も距離の近いｋ個のルータ２０を選択するようにしても構わない。

【００８６】

また、転送経路決定部２２２が取得キーｋｄに一致する検索キーｋｒを探索するときに、経路計算結果情報（ｋｒ，ｖａｌｕｅ）をｒａｎｇｅごとに紐付けて記憶されていれば、探索範囲を絞ることができ、探索時間を小さくすることができる。また、ルータ２０は、ｆｉｎｄ＿ｖａｌｕｅ（ｋｄ）によって取得した転送経路情報ｖａｌｕｅを、自身の記憶部２３の図示しないキャッシュメモリに格納しておくと、探索を早めることができる。なお、キャッシュメモリに格納された情報は、所定時間後（タイムアウト後）に消去される。なお、レプリケーションやキャッシュは、転送経路情報を格納する１つのＲａｄｉｘＴｒｅｅによって管理される。

【００８７】

また、ルータ２０が、ルートサーバ１０の機能を兼ねていても構わない。また、本実施形態は、ＡＳ間の通信プロトコルであるｅＢＧＰを改変するものではないので、ＡＳ内に閉じて適用することができる。そして、本実施形態を適用したＡＳは、従来通り、他のＡＳと接続することが可能である。

【００８８】

また、本実施形態において、ルートサーバ１０（図２参照）の各部１１，１２の処理は、ルートサーバ１０をコンピュータで実現したときに搭載されるプログラムによって実現

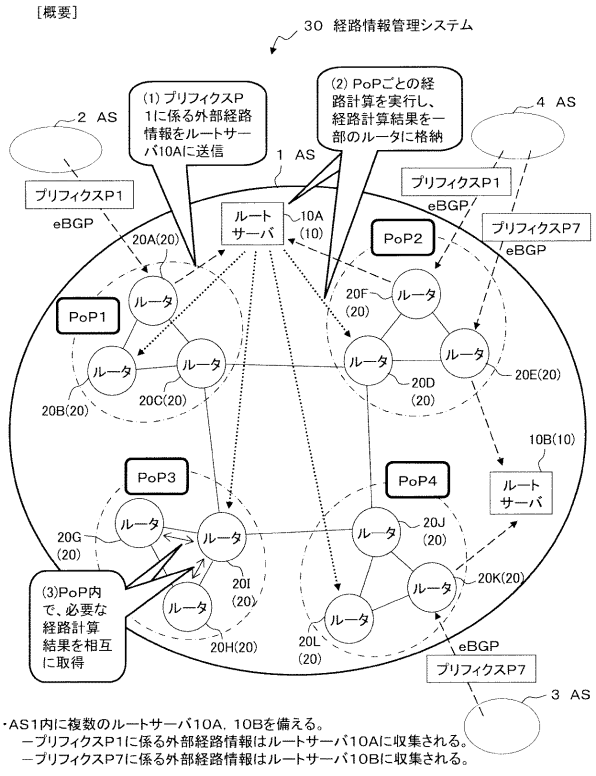
されてもよい。また、本実施形態において、ルータ（図３参照）の各部２１，２２の処理は、ルータ２０をコンピュータで実現したときに搭載されるプログラムによって実現されてもよい。このプログラムは、通信回線を介して提供することもできるし、ＣＤ－ＲＯＭ等のコンピュータ読み取り可能な記録媒体に書き込んで配布することも可能である。

【符号の説明】

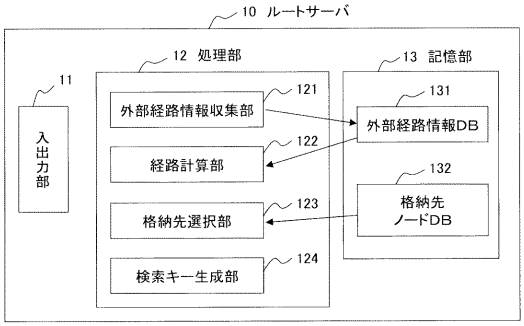
【００８９】

１　ＡＳ
１０　ルートサーバ（経路情報管理装置）
１１　入出力部
１２　処理部
１３　記憶部
２０　ルータ（境界ルータ、ルータ）
２２　処理部
２３　記憶部
３０　経路情報管理システム
１２１　外部経路情報収集部
１２２　経路計算部
１２３　格納先選択部
１２４　検索キー生成部
１３１　外部経路情報ＤＢ
１３２　格納先ノードＤＢ
２２１　ルートサーバ選択部
２２２　転送経路決定部
２２３　取得先選択部
２２４　カプセル化処理部
２２５　取得キー生成部
２３１　ルートサーバＩＤＤＢ
２３２　経路計算結果ＤＢ
２３３　取得先ノードＤＢ

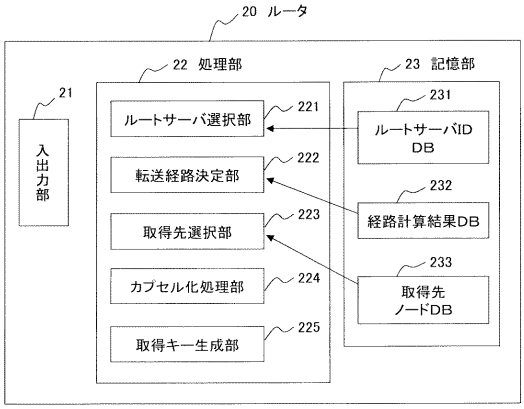
【図１】



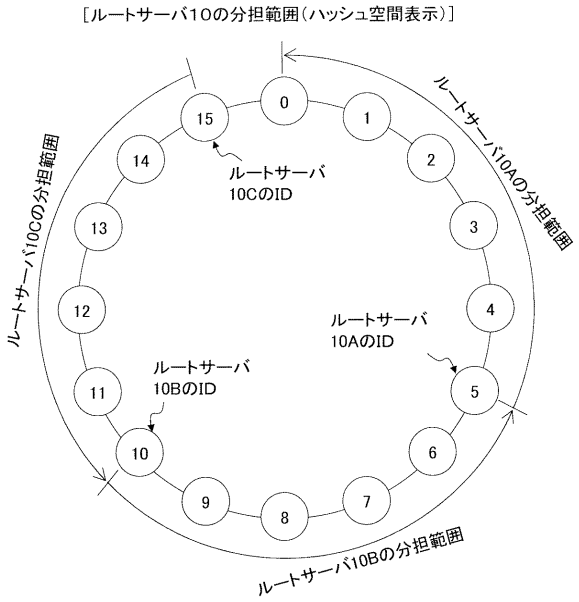
【図２】



【図３】



【図４】



・○内の数字はルートサーバのID番号を表す。
・ルートサーバ10Aの分担範囲：0超え～5以下
・ルートサーバ10Bの分担範囲：5超え～10以下
・ルートサーバ10Cの分担範囲：10超え～15以下

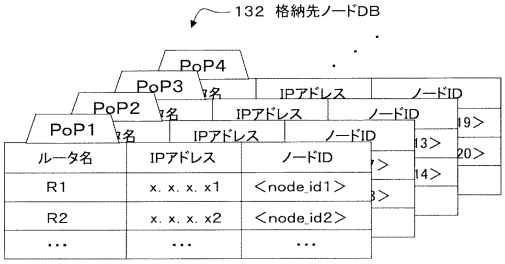
【図５】

[距離の定義]ノードIDと検索キーとのビット列のExclusive ORを演算し、Exclusive ORの演算結果のビット列において、左から見て最初に「1」が出現した位置を右からの数えた桁数の数値

（例）

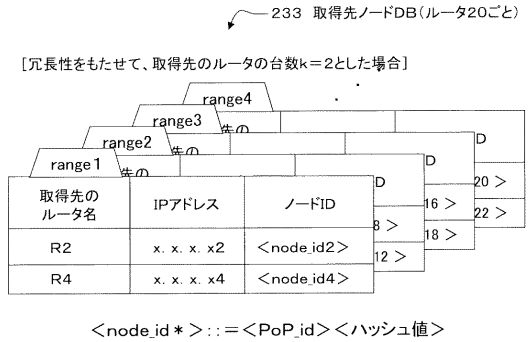
ルータ	ノードID	検索キー	Exclusive OR	距離	近さの順位
a	10011	10111	00100	3	3
b	10101		00010	2	2
c	10110		00001	1	1
d	11000		01111	4	4

【図６】

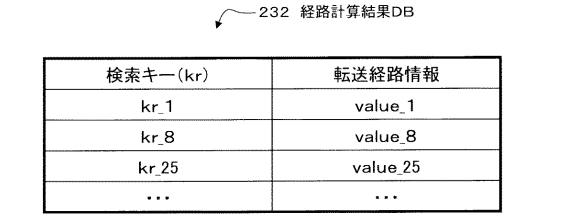


<node_id＊>:::=<PoP_id><ハッシュ値>

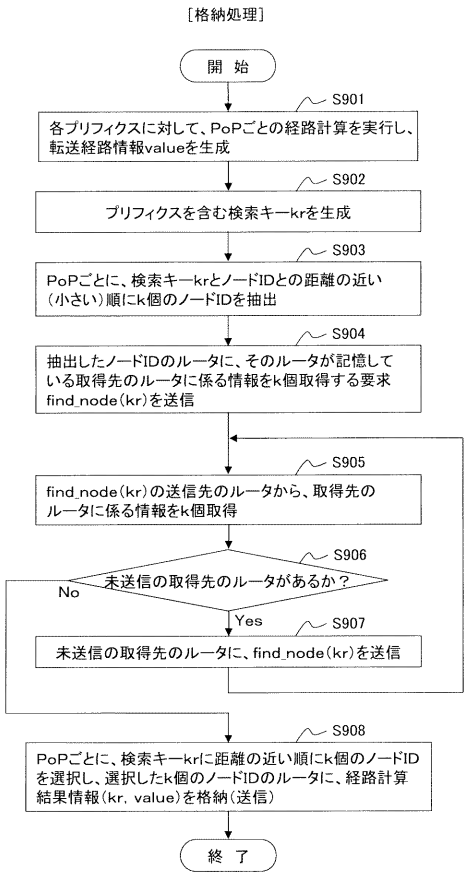
【図７】



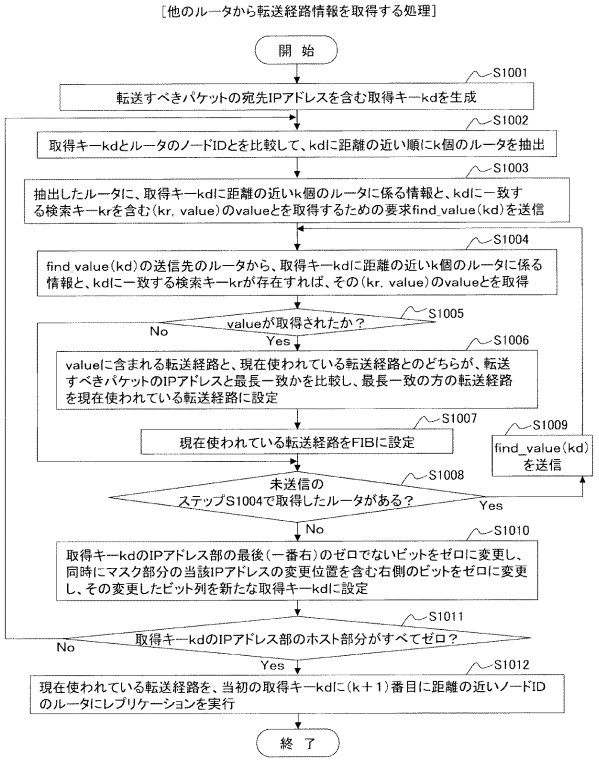
【図８】



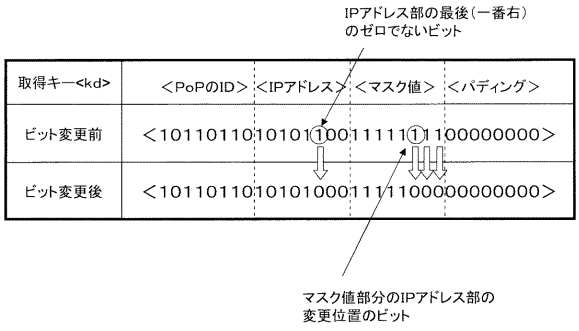
【図９】



【図１０】



【図１１】



フロントページの続き

審査官　安藤　一道

(56)参考文献　特開２００６－２７０７８１（ＪＰ，Ａ）
特開２００４－２６０６５５（ＪＰ，Ａ）
特開２００３－１４３１４３（ＪＰ，Ａ）
特開２０１０－１９９８００（ＪＰ，Ａ）

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｈ０４Ｌ　　１２／５６