2007－OS－104 (12)
2007／1／31

フリースケールネットワーク方式

村上 健一郎[1] 菅原 俊治[2] 明石 修[3] 福田 健介[4] 廣津 登志夫[5]
[1]法政大学ビジネススクール [2]NTTコミュニケーション科学基礎研究所 [3]NTT未来ねっと研究所
[4]国立情報学研究所 [5]豊橋技術科学大学

本論文では、ネットワークを再帰的に拡張できるフリースケールネットワークFSN(Free Scale Network)方式を提案する。FSNではネットワークの構成単位をレルムと呼ぶ。例えばインターネットは一つのレルムである。FSNのレルムには固有のアドレスがあり、ネットワークアドレスの拡張部として使用する。しかし、従来のホストやルータなどの装置を変更することなく、アドレス空間を越えて透過的にネットワークを拡張できる。これは、FSNの多重仮想空間方式によるものである。この利点に加え、プロトコルを変更する必要がないためにネットワークの運用も互換性である。従って、これまでの拡張方式で必要であった膨大な設備投資や運用コストを回避しながら、ネットワークをオンデマンドに拡張することが可能となる。

Free Scale Network Architecture

Ken Murakami[1] Toshiharu Sugawara[2] Osamu Akashi[3] Kensuke Fukuda[4] Toshio Hirotsu[5]
[1]Hosei Business School [2]NTT Communication Science Labs [3]NTT Network Innovation Labs
[4]National Institute of Informatics [5]Toyohashi University of Technology

The Free Scale Network (FSN) architecture provides conventional networks with an unlimited extension capability. FSN consists of hierarchical or meshed realms. Each realm has its unique realm address. The address works as a prefix of the conventional addresses in the realm. FSN gateway interconnects them transparently by a multiple virtual space (MVS) mechanism. In MVS, a local address space is reserved for virtual addresses and an address in the space is allocated dynamically every time a host requests the gateway to resolve address of an FQDN. Thus, it enables the conventional hosts and routers to access whole the address space with no modification.

1. はじめに

本論文では、ネットワークを自由に拡張できるフリースケールネットワークFSN(Free Scale Network)方式を提案する。FSN ではネットワークの構成単位をレルム(realm)と呼ぶ。例えば、現在のインターネットはレルムである。レルム内のネットワークやホストのアドレスは、それぞれのレルムで独立して割り当てられる。各レルムは、FSN 内で一意のレルムアドレス RA(Realm Address)を持つ。レルムには新たなレルムを接続することができ、新たなレルムは元のレルムからは単一のホストに見える。このため、FSN は階層的に何層にも拡張できる能力を持つ。これが本方式の第一の特徴である。

第二の特徴は、従来のホストやルータなどの装置、および、ネットワーク運用との互換性である。これは、各レルム内では、従来のネットワークプロトコルが使用されるからである。新たに追加する装置は、レルム間を接続するレルムゲートウェイ RG (Realm Gateway)だけである。

第三の特徴は、階層的な構造を基本としながらも、信頼性向上のために任意のショートカットを持つことができる点である。階層構造では、親のレルム配下に子レルムが隠れ、子レルムの RA はトップレベルの経路情報に載らない。親レルムが障害に陥れば子レルムは孤立する。そこで、必要なレルムには、トップレベルのレルムと同じく、グローバルアドレスを RA の別名として付与し、経路情報にそれを載せる。このため、障害が発生しても別経路があれば、そこを通じて通信が継続できる。

第四の特徴は、多重仮想アドレス空間方式である。この方式では、従来のホストがアクセスできるアドレス空間内に、拡張によって広がった実アドレスをマッピングするための仮想アドレス空間を設ける。マッピングや仮想空間の管理は RG が行う。このため、各ホストは拡張されたネットワークの全空間を透過的にアクセスできる。また、ホストごとに仮想空間を用意し、特定のホストの異常な振舞いが他に波及することを防止する。

FSN は汎用の方式であるが、説明をわかりやすくするために、以下ではインターネットへの適用を想定して説明を行う。

2. 従来の研究

これまでに、従来のネットワークと互換性を保ちながら透過的に拡張できるネットワークアーキテクチャに関する試みがいくつか行われてきた。それは、互換性のないネットワークでは、次のような問題が発生するからである。(1)従来のネットワーク機器の更新、入れ替え、追加等で膨大な設備投資が必要となる。(2)新ネットワーク方式と旧ネットワーク方式との混在が運用を複雑化し、ネットワーク運用の低信頼化と高コスト化を招く。(3)新たな方式のネットワーク運用のためにオペレータの教育訓練コストが増大する。しかし、従来の研究は企業レベルよりも小さなネットワークを対象とし、国や多国籍企業の大規模ネットワークまで考慮したスケーラビリティを持つものがない。

最初に従来との互換性を念頭に方式を提案したのは RFC1385 の EIP (Extended Internet Protocol) [1]である。この方式では、パケットの形式を現在のインターネット、即ち、IPv4 と同じ形式とした。これは、IPv4 方式のインターネットを二階層に拡張したもので、主眼は、プライベートアドレスを使用したプライベートネットワークを、パブリックアドレスを使用したインターネットへ透過的に接続することである。ネットワークレイヤヘッダにはプライベートアドレスを、そのヘッダオプション部にパブリックアドレスを入れることで形式の互換性を保った。

EIP では、プライベートネットワーク内のホスト間通信に関しては従来との互換性があるが、すべてのルータおよびホストで EIP を解釈できるような改造をしない限り、透過的な通信はできないという問題がある。また、DNS の互換性、二階層以上の拡張性、経路制御等に関して十分な検討が行われていない。更に、我々の調査では、ヘッダオプション部を処理できるルータはほとんどなく、パケットが破棄されるかルータに異常をきたすかのどちらかである。

最近では、ネットワークレイヤのヘッダ部とトランスポートレイヤのヘッダ部との間に拡張情報を置くシムヘッダ方式の研究が行われている。しかし、これらも、ホストやルータで従来プロトコルとの互換性が確保できていない。例えば、IPNL (IP Next Layer)[2]は、各レルムを識別するための 2

バイトの番号 RN (Realm Number) を導入した。これに加え、各レルムを独自の DNS (Domain Name System) ゾーンに対応させるようにした。これは FQDN (Full Qualified Domain Name) でパケットを中継するためである。各ホストは、従来のグローバルアドレス、RN、拡張されたネットワークのプライベートアドレスとの３つの組み合わせで識別される。但し、これらは動的に変更される可能性がある。通信の開始時には FQDN でホストは指定され、レルム間を結ぶ中継装置 NL-router が中継時にアドレスの解決を行う。ホストが一度パケットを交換した後は、ホストはアドレスで識別されるようになる。このような複雑さは、経路障害等が発生した時に別の NL-router を選択できるようにする仕組みから発生する。

これに対し、IP4+4[3]は、EIP と同じくインターネットを二階層に拡張したものである。IP4+4 ではレルムゲートウェイがレルム間の接続を行う。パケットは IPv4 と互換性を持たせ、シムヘッダにパブリックとプライベートのアドレスを置く。レルムゲートウェイは中継の際に、次のレルムに応じてシムヘッダ内からネットワークレイヤ用ヘッダ内のアドレス部へアドレスのコピーを行う。即ち、IP4+4 は転送先までの経路を指定するソースルーティングを行っている。これでは動的経路制御ができず、大規模なネットワークの障害では経路の切り替えに対応できないという大きな問題を持つ。

以上の従来研究が持つ互換性、拡張性、信頼性などの問題に対し、FSN 方式は多重仮想アドレス空間方式、再帰的拡張方式、レルム間経路制御方式によってそれらを解決する。

3. 方式の概要

本章では、FSN の方式概要を説明する。

3.1 FSN のアーキテクチャ

FSN はレルムの集合である。図 1 は最大第四階層までレルムが繰り返し構造をもつ例であるが、その階層の数に制限はない。各レルム内はホストと、それらが接続されたネットワークセグメントを相互接続するルータとから構成される。一方、レルム間は RG で接続する。

各レルムは固有の RA を持つ。インターネットに直接接続されているトップレベルのレルムでは RA はグローバルアドレスである。R0 から R3 がその例である。なお、R0 は現在のインターネットである。ここでは、トップレベルのアドレスをレベル 0 (L0)と呼ぶ。その配下の子レルムは、それぞれの親レルム内で一意の RA を持つ。但し、R4 のように子レルムや孫レルムも L0 のアドレスを RA の別名として持つことができる。

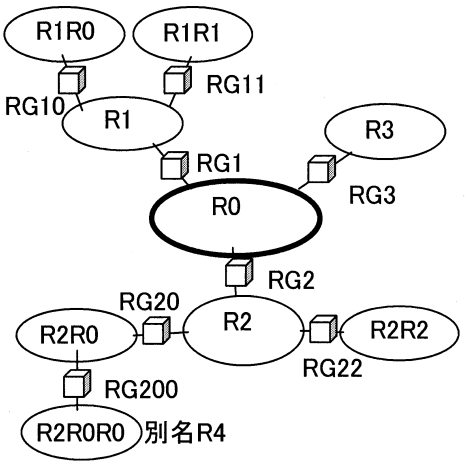

図 1. FSN のアーキテクチャ

3.2 実アドレスと仮想アドレス

RA は階層的に標記する。例えば、R1 配

下のレルムは、それぞれのR1内での一意のRAであるR0、R1を親レルムのRAであるR1と連接し、R1R0、R1R1と標記する。実際のアドレスもそれぞれの階層のレルムにおけるRAを連接したものとなる。これを図2に示す。ここでL0からLn-1まではレルムアドレス、HAがレルムRn-1の中のホストアドレスである。

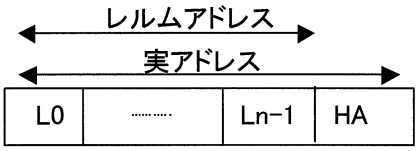

図2. FSN実アドレス形式

FSN内のホストは図2の実アドレスを持つが、ローカルレルム内では、HAの部分だけを利用する。ローカルレルム内では従来のインターネットと互換の経路制御が行われており、転送プロトコルに変更はない。ホストは転送先ホストのアドレスをネットワークレイヤヘッダに入れてパケットを送り出す。ルータは、この転送先アドレスを参照してパケットを中継し、最終的に転送先のホストまで当該パケットを転送する。

一方、レルム間の通信では、実アドレス全体が使用される。しかし、ホストは実アドレスが扱えない。そこで、従来のプライベートアドレス領域の一部を仮想空間として使用し、実アドレスをこの空間にマッピングする。このマッピングの様子を図3に示す。

レルムが小規模で、内部のホスト数がプライベートアドレス空間内に収まる場合には、プライベート空間以外の外部のL0空間を従来通り透過的に使用することもできる。一方、国や多国籍企業レベルの大規模なレルムの場合はローカルアドレスとして全空間を使用することになるので、外部のL0空間も仮想空間にマッピングして使用することになる。いずれの場合も、外部との通信の透過性は確保される。

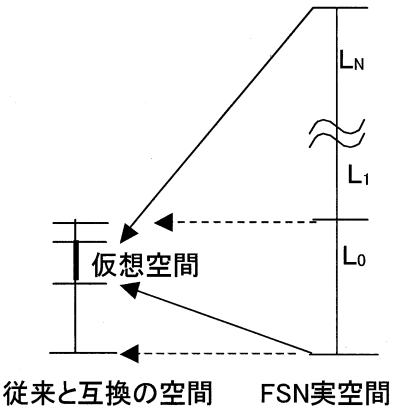

図3. アドレスマッピング

3.3 アドレスマッピング

マッピングは、RGによって以下のように行われる。転送にあたって、転送元のホストは、まず、転送先ホストのFQDNをアドレスに変換する必要がある。このアドレス解決要求はRGへ送られる。RGはDNSリゾルバーの機能を持っており、DNSサーバをアクセスして、当該FQDNの実アドレスを得る。DNSではアドレスを示すAレコードに加え、HINFOレコードに当該レルムを示すFQDNが入っている。RGは、HINFOレコードが空でなければ、このFQDNのアドレス解決も行い、最後に、それらのアドレスを連結して実アドレスを得る。この様子を図4に示す。

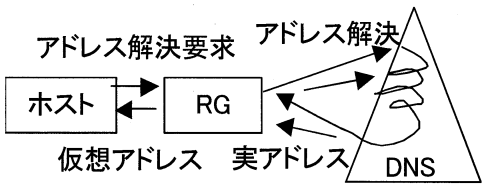

図4. アドレス解決

もしも、実アドレスが透過的にレルム外にパケットを中継すべきアドレス空間内にある場合には、そのままホストに返す。それ以外の場合には、実アドレスと仮想アドレスとのマッピングを保持するTLB(Translation Look-ahead Buffer)内の転送元ホストに関するテーブルを調査する。多重仮想アドレス方式のために、テーブルは転送元ホストごとに設けられている。すでに実アドレスに対して仮想アドレスが割り当てられていれば、それをアドレス解決要求への応答として返す。割り当てられていなければ、未使用の仮想アドレスVAを選択し、それを返す。同時に当該仮想アドレス用のエントリをTLBに作成する。

3.4 パケットの転送

VAを受け取った転送元ホストはそれを転送先としてパケットを送信し、通信を開始する。RGは仮想アドレス空間や到達可能なL0アドレス空間の経路情報をレルム内部にアドバタイズしているので、当該パケットはRGへ届く。

RGは、転送先アドレスが透過的に中継しなければならない空間であれば、そのままパケットを転送する。一方、それが仮想アドレス空間内であれば、実アドレスへの変換を行う。まず、TLBを参照して実アドレスを得る。そして、図5に示すシムヘッダ部を付加し、転送先および転送元の実アドレスをシムヘッダに入れる。

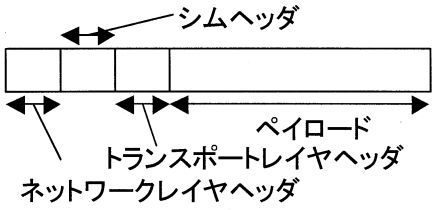

図5. FSNのパケット形式

RGは上記の変換処理を行った後、あるいは、FSNパケットを外部から受信した時にレルムの経路情報を入れたテーブルRRT(Realm Routing Table)を参照してネットワークレイヤヘッダ内の転送先アドレスを書き換える。RRTの例を図6に示す。

転送先RA	次RG	属性等
R2R0R0	RG20	
R2R0	RG20	
R2	-	IF1
R1	RG1	
R0	-	IF0
R3	RG3	
R4	RG20	

図6. レルム経路テーブル例 (RG2)

このテーブルには各RAについて、次に渡すべきRGのアドレスが入っている。RGはシムヘッダ内の転送先実アドレスをキーとしてテーブルを検索し、一致するエントリがあれば、それに対応した次RGのアドレスをネットワークレイヤヘッダの転送先アドレスにコピーする。また、転送元アドレスにはその RG 自身のインタフェースのアドレスを入れる。これは、レルム内を中継中に何らかのエラーが発生した場合、それに対するICMP(Internet Control Message Protocol)のパケットを受信して処理するためである。

アドレスのマッチングにはlongest match first アルゴリズムを採用する。これは、最短距離でパケットを中継するためである。例えば、R2R0R0 あてのパケットの場合、R2や R2R0 のエントリがあっても、R2R0R0のエントリがあれば、それが優先される。なお、4.3でも説明するが、RRTでは子レルムのアドレスを親レルムのアドレスにアグリ

ゲートして消去する場合がある。

受信したパケットが自レルム内のホストあてである場合には、FSN パケットを通常のパケットに逆変換する。その際、転送元アドレスには仮想アドレスを入れなければならない。そこで、当該 RG の TLB を検索し、シムヘッダ内の転送元実アドレスに対応する仮想アドレスがあれば、それを使用する。該当するエントリがなければ、新たにエントリを作成し、未使用の仮想アドレスを割り当てる。そして、当該アドレスをネットワークレイヤヘッダの転送元アドレスとする。転送先アドレスには、シムヘッダ内の転送先実アドレスのホスト部分をコピーする。

3.5 シムヘッダの形式

シムヘッダの形式を図 7 に示す。シムヘッダは転送中に長さが変わることのない固定長のヘッダである。但し、その長さは、転送元と転送先のホストがあるレルムの階層によって変わる。

シムヘッダには、転送先ホストの実アドレス DRA(Destination Real Address)および転送元ホストの実アドレス SRA(Source Real Address)以外に、転送元ホストの仮想アドレス SVA も入れられている。これは、転送先ホスト側で応答パケットのヘッダチェックサムを作成する場合に、最終的にホストへ到着した時のヘッダチェックサムを予想して計算しておくために使う。この場合、中継中の再計算が不要となる。なお、誤った仮想アドレスへの変換が行われた時にエラーを検出するため、転送先ホストの FQDN もオプション領域に入れられている場合がある。

Length フィールドは、このシムヘッダの長さを示す。また、SRA offset フィールドは SRA の開始位置を示す。SRA length フィールドは実アドレスに続く領域の位置を示す。current DRA は、すでに中継中に一致した DRA 部分を示すもので、経路テーブルの検索を高速化するためのものである。Protocol フィールドは TCP (Transmission Control Protocol)や UDP (User Datagram Protocol)などのトランスポートレイヤのプロトコル種別を示す。

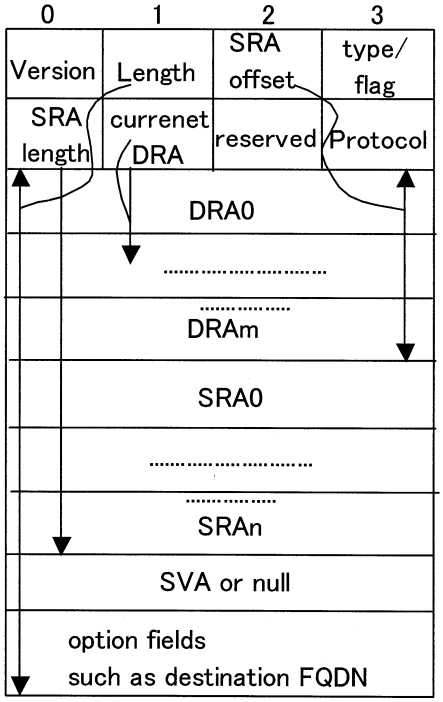

図 7. シムヘッダの形式概要

FSN では中継途中にパケットの細分化が発生すると最初のフラグメント以外はシムヘッダが存在しないという問題が発生する。そこで、ネットワークヘッダにはフラグメント禁止(Don't Fragment)のフラグを立てる。また、MTU 検出[4]によって、細分化がホスト内だけでしか発生しないようにする。その再組み立ては転送先のホストで行われる。

4. レルムゲートウェイ RG

この章では、RG の機能について説明する。

4.1 RG の概要

RG は、レルム間を結ぶパケット中継装置である。これは以下の 6 つの機能を備える。

(1) DNS リゾルバーとして働く。
(2) 仮想アドレスの割り当てや回収など、TLB の管理を行う。
(3) 同一レルムに属する RG 間でオンデマンドあるいは定期的に TLB を交換する。
(4) 同一レルム内の RG 間で RA の交換をすることによって RA 経路の動的生成を行う。
(5) 同一レルム内のルータと内部経路の交換をすることにより、レルム内部あてのパケットの中継に必要な経路情報を得る。
(6) レルム間でパケットを中継する。

すでに(1), (2)は説明をしたので、以下では、(3)から(6)の機能について説明する。図 8 に RG の内部構造を示しておく。

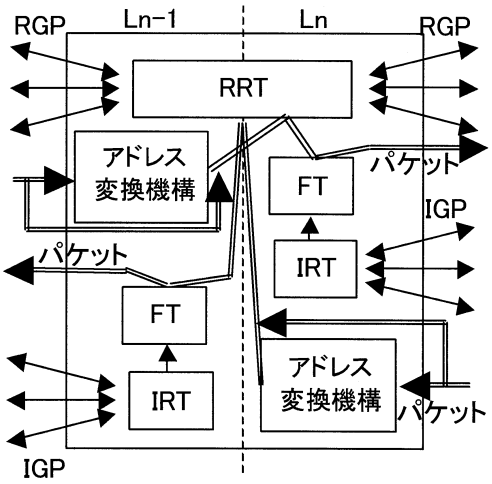

図 8　RG の内部構造

4.2 仮想空間の分割と VARP

同一レルムに複数の RG が接続されている場合、障害が発生した時に RG を変更して障害からの回復を行う。このために、各 RG は仮想空間を分割して担当する。即ち、RG はホストからのアドレス解決要求の際にレルムアドレス経路情報を参照し、最も目的のレルムに近い RG に VA の割り当てを要求する。そして、その VA をホストへ返す。この RG 間の要求と応答のプロトコルを VARP (Virtual Address Resolution Protocol) と呼ぶ。また、RG は、オンデマンドあるいは定期的に TLB を交換し、コピーを保持しておく。ある RG がダウンした場合には、他の RG も TLB のコピーを保持しているので、仮想アドレスを継続して使用できるようになり、通信の継続が可能となる。

4.3 RGP (Realm Gateway Protocol)

同一レルムに属する RG は、RA の経路情報および TLB の交換を行う。このプロトコルを RGP と呼ぶ。RGP は BGP(Border Gateway Protocol) [5]と同じくパスベクトル型の経路制御プロトコルである。但し、パスはレルムのパスである。また、ポリシールーティングも行わない。

RGP では以下の情報を交換する。

(1) レルムアドレス
(2) レルムまでのパス
(3) 経路情報の生成元(IGP, RGP, static などの経路がどこから生成されたのかを示す)
(4)経路の属性（優先度など）
(5) TLB の更新情報

各 RG は、すべての RG から受信した経路情報から、レルムアドレスごとに最適なパスを計算する。この場合、短いパスや優先度の高いパスが選定される。また、レルム経路のアグリゲーションも行う。例えば、R0 へのパスと R0R1 への同一のパスがあれば R0R1 のパスは削除される。この場合、デバッグの

ため、アグリゲーションを行ったRGのアドレスを経路の属性に入れておく。このようにして計算したベストパスは、3.3.4で説明したRRTに保存する。また、それを送ってきたRG以外のRGにアドバタイズする。

4.4 IGP (Interior Gateway Protocol)

IGPでは、レルム内部のルータから経路情報を受信し、ベストの経路を計算する。この結果はIRT(Internal Routing Table)に保存される。注意しなければならないのは、RGが複数のレルムにまたがっているため、経路情報も別のIRTに分離して保持しなければならないことである。但し、L0の経路情報に関してはすべての情報の中からベスト経路を選択し、他方のレルム側のルータにもアドバタイズする。また、設定によって、特定の経路にだけをIRTからRRTへインジェクトしたり、その逆を行うことができる。

パケットの処理の際には、まず、RRTによってパケットが自レルム内のホストあてか別レルム内のホストあてかを判別する。他レルムあてであれば、最適パス上の次RGに中継される。自レルムあてであれば、アドレスとパケットの逆変換を行い、従来のパケット形式に戻す。次に、IRTをもとにベストパスを入れたFT(Forwarding Table)を検索し、当該パケットを渡すルータを決定して転送する。

5. おわりに

本論文では、従来のホストやルータなどの装置も、ネットワーク運用も変更することなく、ネットワークを制限なく拡張できるFSN方式について論じた。本方式の第一の特徴は、ネットワークが階層的なレルムの繰り返し構造を持つことである。第二の特徴は、従来のホストやルータ等の装置の変更が必要ないことである。第三の特徴は、階層的な構造を持ちながらもショートカットによる経路の最適化や信頼性向上が図れることである。第四の特徴は、多重仮想アドレス空間方式を用いることによって、従来のホスト側のソフトウェアを改造することなく全空間をアクセスできる点である。第五の特徴は、FQDNの実アドレスを従来のネームサーバDNSで解決できることである。

FSNでは、装置や運用の互換性によって、これまで必要であった膨大な設備投資や運用コストを回避しながらネットワークを拡張することができる。従って、新世代の安価なネットワークシステムやサービスの提供が可能となる。また、これからネットワークが普及する発展途上国でもインターネットと互換性および相互接続性のある安価な新世代のネットワークの構築が可能となる。

参考文献

[1] Z. Wang, “ EIP: The Extended Internet Protocol”, RFC1385, Nov. 1992

[2] P. Francis, R. Gummadi, “ IPNL: A NAT-Extended Internet Architecture,” SIGCOMM' 01, August 2001.

[3] Zoltán Turányi, András Valkó, Andrew Campbell, ” 4+4: an architecture for evolving the Internet address space back toward transparency”, ACM SIGCOMM Computer Communication Review, Vol. 33, No 5, pp 43-54, October 2003.

[4] J. Mogul, S. Deering, “ Path MTU discovery”, RFC1191, November 1990.

[5] Y. Rekhter, T. Li, “ A Border Gateway Protocol 4 (BGP-4)”, RFC1771, March 1995.