学士論文 2020年度 (令和2年度)

ShadowLB: A Transparently Accelerated Load Balancer

慶應義塾大学 環境情報学部
橘 直雪

学士論文要旨 - 2020 年度 (令和 2 年度)

ShadowLB: A Transparently Accelerated Load Balancer

近年，ロードバランサーはハードウェアアプライアンスのものからソフトウェアのものへ変遷しつつある．これにより，データセンターは物理スペースや経済的なコスト問題を解決した．これらの高速ソフトウェアロードバランサにはカーネルバイパス系パケット処理フレームワーク等を利用した高速データプレーン技術を用いたパケット処理性能の向上が図られているが，これらの高速データプレーン技術は既存のコントロールプレーンの API を利用することができず，開発者はデータプレーンよりもコントロールプレーンの開発にコストを掛けているのが実情である．

そこで本研究では，既存のコントロールプレーン機能を流用しつつ，データプレーンのみ透過的に高速化させる機構,shadowLB を設計，実装した．shadowLB は既存のコントロールプレーンのデファクトスタンダードである ipvs の API を用いながら，テールレイテンシにおいて X 倍の性能を評価にて示した．

キーワード:
1. Load Balancer，2. XDP，3. 負荷分散

慶應義塾大学 環境情報学部
橘 直雪

Abstract of Bachelor's Thesis - Academic Year 2020

ShadowLB: A Transparently Accelerated Load Balancer

In recent years,load balancers have been shifting from hardware appliances to software. In this way, data centers have solved the physical space and economic cost problems. These high-speed software load balancers are designed to improve packet processing performance by using high-speed data plane technology that utilizes kernel bypass packet processing frameworks, etc.

However, these high-speed data plane technologies cannot utilize existing control plane APIs, and developers are spending more resources on control plane development than on data plane.

In this study, we designed and implemented shadowLB, a mechanism that transparently accelerates only the data plane while reusing the existing control plane functions. We have shown that shadowLB can achieve X times performance in tail latency while using the ipvs API, which is the de facto standard for existing control planes.

Keywords :
1. Data center network, 2. Network operation, 3. IPv6 transition mechanism

Keio University Faculty of Environment and Information Studies
Naoyuki Tachibana

目 次

図 目 次

表 目 次

第1章　序論

本章では本研究の背景と全体の構成について記述する．

1.1　背景

1.1.1　ロードバランサを取り巻く環境

大規模ネットワークサービスの台頭

スマートフォンやタブレットの普及，また Wi-Fi スポットの増加などに伴い，インターネットのトラフィックは近年急速に増加している [1]．それに伴い，インターネットサービスを展開する際には，大量のトラフィックを処理するために，ロードバランサーを設置して複数のサーバにトラフィックを分散するというアプローチが一般的に取られている．

ロードバランサーのソフトウェア化

従来，ロードバランサは F5 Networks[1], Cirtix[2]等を代表とするハードウェアアプライアンス製品が一般的であった．しかし，ハードウェアロードバランサは処理能力は優れているものの，以下のような課題を抱えていた．

- **物理スペースの圧迫**ルータ，スイッチ，サーバと同様，1U 以上の設置スペースを必要とする．
- **経済的コスト**データセンター向けネットワーク製品は高価であるとともに，メンテナンス費用，電力等様々なコストが発生する
- **冗長化が困難**物理アプライアンスであることから，ネットワークトポロジに組み込むまでに最短でも 1〜2 日は必要とする．そのため，急増するネットワークトラフィックに即座に対応できない．

このようなオペレーター，ユーザー双方にとってコストが高いハードウェアロードバランサーに代わり，Google[3]の Magrev[2] や Microsoft[4]の Ananta[3] といったように，大規模

[1] https://www.f5.com/ja_jp
[2] https://www.citrix.com/ja-jp/
[3] https://www.google.com/
[4] https://www.microsoft.com/ja-jp

なサービスを提供している企業はロードバランサを自社技術を用いてソフトウェア化し始めた．各企業の高速ロードバランサの詳細に関しては，2.1 章にて説明する．

1.2 本研究の目的

本研究では，Layer-4 ロードバランサにおける

1.3 本論文の構成

本論文の構成を以下に示す.

第 2 章では，IPv6 シングルスタックネットワークにおける IPv4 サービス提供手法に関してそれぞれの特徴や利点を紹介し比較する．

第 3 章では，IPv4/IPv6 プロトコル変換を利用した IPv4 サービス提供手法の一つである SIIT-DC のアーキテクチャと，解決すべき課題について明らかにする．

第 4 章では，SIIT-DC の課題を解決するために考えられる手法について論ずる．

第 5 章では，本研究において提案するダイナミックなアドレス変換テーブル広告手法の要件と構成について記述する．またメッセージングプロトコルとして採用した BGP の技術的利点について述べる．

第 6 章では，本提案手法の BGP メッセージのペイロード設計と第 7 章でも評価実験に用いる PoC(Proof of Concept) の具体的な実装について述べる．

第 7 章では，第 3 章で述べた課題に対して，本提案手法が有用であることを検証するための実証実験の概要及び具体的なシナリオについて述べ，結果を考察する．

第 8 章では，本研究のまとめと本研究の展望について検討する．

第2章　ソフトウェアロードバランサにまつわる技術

本章では既存のソフトウェアロードバランサを論ずる上で必要不可欠な高速データプレーンの技術について論ずる.

2.1　高速なソフトウェアロードバランサ

第3章　SIIT-DCのデザインと現状の課題

第??項で述べたPv4/IPv6トランスレーションを用いたIPv4サービス提供手法の一つとして，SIIT-DCがインターネット標準化されている．本章ではSIIT-DCのデザインとメリット及び考えられる運用，そして現状の課題について述べる．

3.1　SIIT-DC

3.1.1　概要

SIIT-DCとは，ステートレスIP/ICMP変換アルゴリズム[4]を利用して，IPv4インターネット・ネットワークからのアクセスをIPv6シングルスタックネットワーク上のホストに提供するためのネットワークデザインである．2016年にIETF IPv6 Operations WG[1]での議論を基にRFC7757として標準化された[5].

3.1.2　用語

SIIT-DCで利用される用語，及び特徴的な役割を有する機器・技術について述べる．

SIIT(Stateless IP/ICMP Translation Algorithm)

SIITとはIPv4/IPv6トランスレーションに用いられるプロトコル変換機能の略称である．RFC2765[6]で初めて標準化され，その後RFC6145[7]により一部の仕様が実運用のユースケースに合わせて変更された．現在はIPv6拡張ヘッダーを扱う機構などが追加されたRFC7915[4]が現行の標準仕様である．

BR(Border Relay)

BRとは，SIIT-DCネットワークにおいてIPv4インターネットとIPv6ネットワークとの間でSIITによるIPv4/IPv6トランスレーションを行う機器もしくは他の役割を有する

[1]IPv6 ネットワークの運用要件や関連する技術仕様の策定を行うワーキンググループ．https://datatracker.ietf.org/wg/v6ops/about/

機器の一機構である．IPv4 インターネットと IDC 内の IPv6 シングルスタックネットワークの各境界部に所在し，後述する EAMT を参照した 1:1 のアドレス変換を行う．IDC ネットワークに IPv4 インターネットとの接続点が複数ある場合，接続点ごとに最低一つの BR を配備する．

ER(Edge Relay)

ER とは，IDC 内の IPv4 ネットワークと IPv6 ネットワーク間での多：多の IPv4/IPv6 トランスレーションを行う機器である．

SIIT-DC ではそのオプションとして，IPv4 ネットワーク内の IPv4 しか利用出来ないホストが，SIIT-DC を利用して IPv4 サービスを提供するユースケースをサポートする SIIT-DC Dual Translation Mode[8] が定義されており，ER はその中での利用が想定されている．

通常，ER が有する後述の EAMT は IDC ネットワーク内の IPv4 ネットワークアドレスと，その IPv4 ネットワークを示す IPv6 サービスアドレスにより構成される．

IPv4 サービスアドレス

IPv4 サービスを提供する IPv6 シングルスタックネットワークに属するホストに割り当てる IPv4 アドレス (群) を IPv4 サービスアドレスと呼称する．このアドレス宛に送信されたパケットは，BR/ER によって対応する IPv6 サービスアドレスに変換される．

なお，IPv4 サービスアドレスは IPv4 インターネットに経路広告されている必要がある．

IPv6 サービスアドレス

ER/BR を介してアプリケーションやホストに割り当てられた IPv6 アドレス (群) を IPv6 サービスアドレスと呼称する．IPv4 クライアントは SIIT-DC のアーキテクチャをを通じて，この IPv6 サービスアドレスが割り当てられたホストと通信することが出来る．

変換プレフィックス

変換プレフィックス (Translation Prefix) とは，全ての IPv4 アドレスをマッピングするために用いられる，プレフィックス長が 96bit の IPv6 ネットワークプレフィックスである [9]．IANA によって主に WKP(Well Known Prefix) として 64:ff9b::/48 が予約 [10, 11] されているが，運用者の裁量で ISP 自身に割り当てられた NSP(Network Specific Prefx)[2]を利用する事ができる．

IPv4 アドレスと IPv6 アドレスの間で変換を実行する際に，BR/ER は変換前の IP ヘッダーのアドレスフィールドを、変換プレフィックスが挿入・削除された状態に書き換える。

[2]主に RIR から割り当てられた IPv6 Global Unicas Address を指す.

なお SIIT-DC ネットワークにおいて，変換プレフィックス宛のパケットは各 BR/ER の IPv6 インターフェース宛に IGP(Interior Gateway Protocol) などで経路広告される必要がある．

EAM(Ecplicit Address Mapping)/EAMT(Ecplicit Address Mapping Table)

EAM とは，EAM アルゴリズム [12] によって結びつけられた IPv4 サービスアドレスと IPv6 サービスアドレスのペアを表す．

EAM において，それぞれ同数の IPv4 サービスアドレスと IPv6 サービスアドレスによって構成される．標準では結び付けられた IPv6 サービスアドレスが IPv4 サービスアドレスより多い状態が想定されているが，IPv6 サービスアドレスのホスト部が若いものから優先して変換するため，余剰分のアドレスは無視される．

また，BR 及び ER が変換を行う際に参照する EAM 群が記録されたテーブルを EAMT と定義している．以後 EAMT もしくは変換テーブルと呼称する．

3.1.3 ネットワーク設計

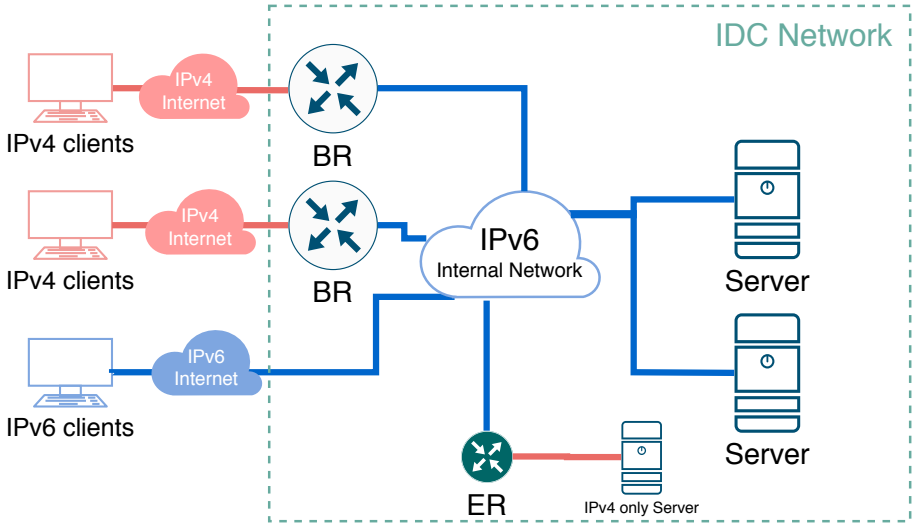

図 3.1: SIIT-DC ネットワーク

基本的な SIIT-DC ネットワークを図 3.1 に示す．

BR は IPv4 インターネットとの各接続点に配置される．各 IPv4 サービスアドレスは自組織のアドレスとして，IPv4 インターネットに経路広告される必要がある．

SIIT-DC ネットワークでは，変換プレフィックス宛のパケットは BR に対してルーティングされる．BR が複数ある場合，BR がネットワークプロトコルに利用する変換プレフィックスを別個に用意するか，同一の変換プレフィックスを各 BR にエニーキャスト [13] によっ

てルーティングさせる．エニーキャストを使用した場合，BR の障害時には別の BR へとトラフィックを迂回させることが可能である．

ER は IDC 内の IPv4 ネットワークとの接続点に配置され，IPv4 のみを持つホストが IDC 内の IPv6 ネットワークを介して IPv4 インターネットにサービス提供を行う場合に利用される．

3.1.4 SII-DC のメリット

SIIT-DC を用いた IPv4 サービスの提供によるメリットとして，以下の点が挙げられる，

デプロイメントが容易

SIIT-DC では，IDC の IPv6 ネットワークと IPv4 インターネットとの接続点に BR を設置を行うのみにより，基本的な IPv4 サービスの提供が可能である．そのため IDC のネットワークトポロジに限定されないシンプルな IPv4 サービス提供が期待できる．

アドレス単位での IPv4 アドレスの効率的な利用が可能

通常の IP ネットワークにおいて，サーバに対する IP アドレスアサインメントはサブネット単位での割り当てを行う必要がある．従来，事前に同一サブネットに属するホスト数を見積持った上で不足が生じないようにネットワークサイズを設定する必要があるめ，ネットワークサイズを超えるサービスの拡大が必要になった場合，サブネット全体の再設計が不可欠であった．また IP ネットワークには，ネットワークアドレスやブロードキャストアドレス，そしてデフォルトゲートウェイとなるルータのインターフェースのアドレスを確保する必要があり，ネットワークサイズが断片化されるほど，実質的に利用できないアドレスの割合が大きくなる問題が会った．

しかしながら SIIT-DC ではサーバごとにアドレスを割り当てることが出来るため，従来利用できなかった IPv4 アドレスを再利用することで，IPv4 アドレスの効率的な利用を実現できる．第??項で述べたように今後益々IPv4 アドレスの調達が困難になることが予想されるため，IPv4 アドレスの効率的な利用は事業者の負担軽減に繋がる．

スケーラビリティ

SIIT-DC の標準仕様 [5] では明示的に述べられていないが，本論文では BR を並行して複数配置することで，スケールアウトが可能なネットワークデザインを立案する．本ネットワークデザインでは，ECMP 及びエニーキャスト [13] を利用することにより，BR の数を水平に増加させることで，IPv4 サービスの提供容量をリニアに増加させることが可能である．

図 3.2 に本ネットワークデザインに則って BR 配置を行ったスケーラブルな SIIT-DC ネットワークの例を示す．IPv4 クライアントからのアクセスはいずれかの BR にフォワー

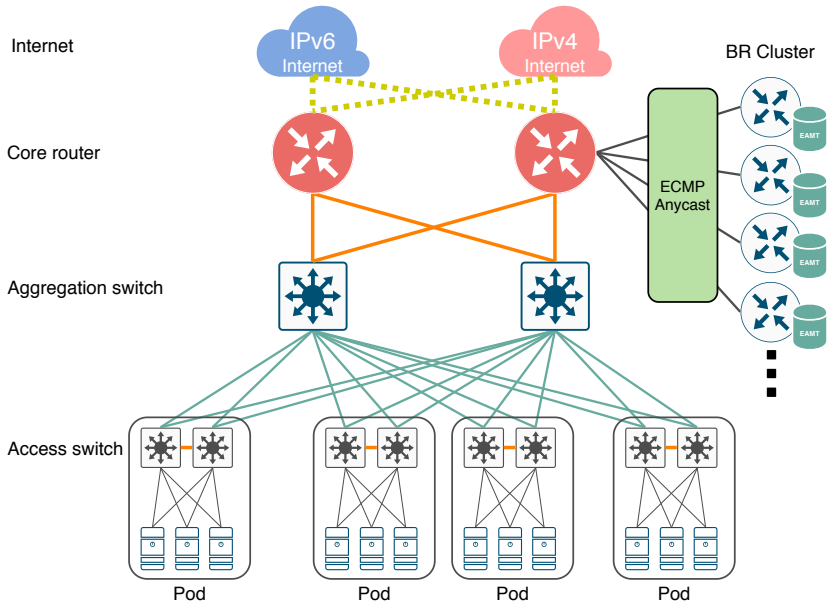

図 3.2: BR を水平スケールすることが出来る SIIT-DC ネットワーク

ディングされた後，IPv6 プロトコルに変換された上で再度コアルータを介して IDC ネットワーク内の IPv4 サービス提供サーバに到達する．

3.1.5 基本的なパケットの流れ

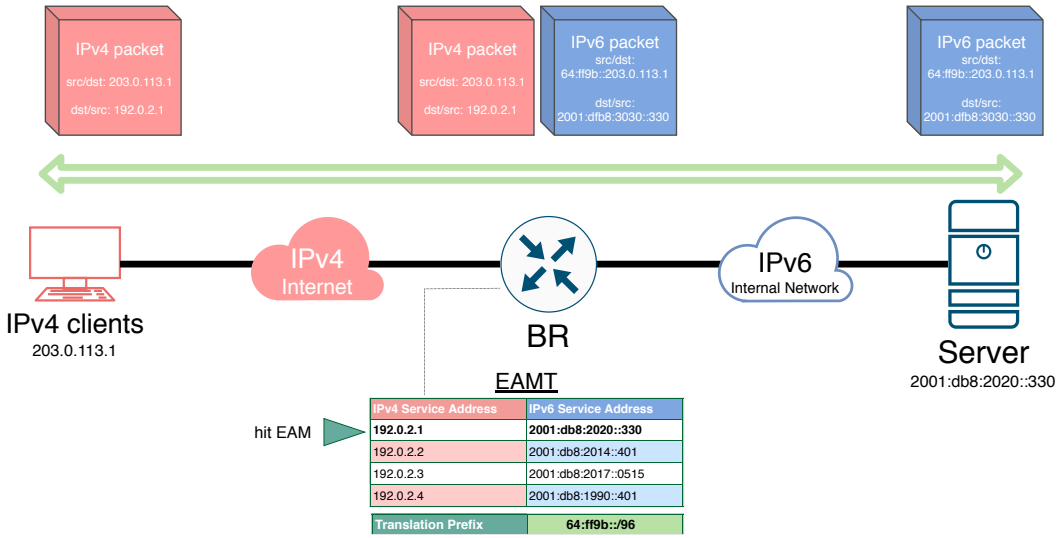

図 3.3: SIIT-DC パケットの流れ

SIIT-DC における基本的な IPv4 クライアントからのトラフィックの流れは以下の様になる．一連のパケットの送信元・送信先のアドレスの遷移を図 3.3 に示す．

IPv4 クライアントの IPv4 サービスアドレス宛のパケットは IPv4 インターネットに接続する BR に到達後，当該 BR が有する EAMT に従って IPv6 サービスアドレス宛の IPv6 パケットに変換される．このパケットの送信元アドレスは変換プレフィックスに埋め込まれた IPv6 アドレスとして表現される．IDC 内の IPv6 ネットワークを介して IPv6 サーバに到達した後，IPv6 サーバは送信元アドレスへの応答パケットを送信する．3.1.2 項で述べたように，変換プレフィックス宛のパケットは IPv6 ネットワークを経由して BR にルーティングされる．IPv6 サーバからの応答を受け取った BR は EAMT を参照し，送信元アドレス (IPv6 サービスアドレス) を IPv4 サービスアドレスに書き換え，送信先アドレス (IPv4 クライアントの IPv4 アドレス) から変換プレフィックスを除去書き換えたのち，IPv4 インターネットを介して IPv4 クライアントに返送される．

3.2 SIIT-DC の課題

本節では SIIT-DC の現状の課題及びそれに起因して起こる事象に関して述べる．

3.2.1 一貫した EAMT の必要性

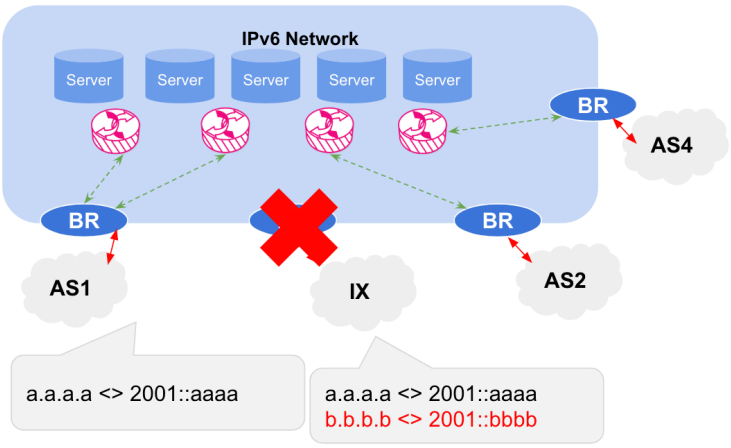

図 3.4: BR に障害が発生した場合に適切にフェイルオーバーが出来ないケース

3.1.2 項で述べたように，SIIT-DC では対外接続点ごとに BR を配置するネットワークデザインを採用することで，IPv6 シングルスタックネットワークに最小限の IPv4 ネットワークを追加するだけで IPv4 サービスの提供を可能にしている．また 3.1.3 項や 3.1.4 項で触れたように，複数の BR で共通した変換プレフィックスをエニーキャストで IDC ネットワーク内に広告する運用を行うことにより，BR 及び対外接続点の障害時に他の BR を

用いて IPv4 サービスの提供を継続することが出来る．この機構を有効に作用させるためには，SIIT-DC ネットワーク内の全ての BR で一貫した EAMT の保持が求められる．

しかしながら現状の SIIT-DC 及び EAMT の仕様 [5, 8, 12] では，BR は他の BR との間で EAMT を共有するためのメッセージング機構を有さない，これは BR 間で EAMT の不一致が発生した場合に，差異となった EAM に該当する IPv4 サービス宛のトラフィックを別の BR へ迂回出来なくなるケースが発生することを意味する．

3.2.2 変更追従性の欠如

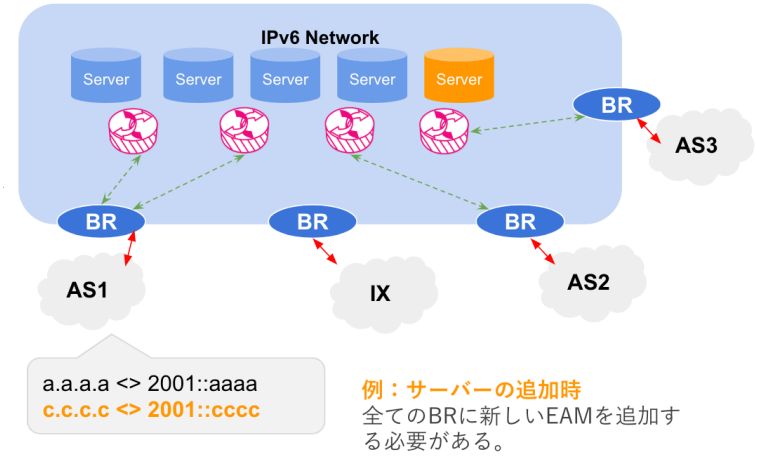

図 3.5: サーバを追加した際，全ての BR への設定追加が必要になる．

プライベートクラウド環境が一般的に利用される IDC ネットワークでは，日々多くのサーバやアプリケーションが追加・廃止・変更される．一方で 3.2.1 項で触れたように，SIIT-DC で IPv4 提供サービスを冗長に運用するためには，IPv4 提供サービスに該当する EAM が BR の EAMT に保持されることが要求される．IPv4 提供サービスの構成に変更があった場合，全ての BR の EAMT を更新する必要がある．

しかしながら現状 SIIT-DC 及び EAMT の仕様 [5, 8, 12] において，IPv4 サービスを行うサーバの存在や状態によってダイナミックに EAMT を更新する機構は存在しない．そのため，IDC ネットワークにおける IGP などによって IPv6 サービスへの到達性が検証されていたとしても，IPv4 サービスの場合はリアルタイムな構成変更に追従することが出来ない．

3.3 本章のまとめ

第 3.2 項で述べたように，現状の SIIT-DC 及び EAMT の仕様は EAMT の一貫性を担保する手法の検討がなされておらず，それに起因した障害時の適切なフェイルオーバーの

実行や IPv4 サービスの増減時の変更追従に関しての課題がある．NPO 日本ネットワークセキュリティ協会 (JNSA) らの調査によれば IT システムの障害の原因の約半数は人為ミスに分類されるものにあり [14]，サービスの安定的な稼働を実現するためには単調な繰り返し動作を含む運用をシステムによって減らす必要がある．

第4章　手法の検討

4.1　概要

本研究ではSIIT-DCにおける動的なEAMTの管理・制御手法の実現を目指す．以後このような機構をダイナミックEAMT機構と呼称する．

本章では考えられる手法を大別した上でその特徴と利点及び欠点を挙げ，最も適した手法を検討する．

4.2　求められる要件

前で述べたIPv4サービス提供手法の機能要件とSIIT-DCの現状の課題を総括し，EAMTを動的に制御する手法に求められる要件を下記のように定義した．

1. **BR間のEAMTの一貫性**
   障害時の適切なフェイルオーバを実現するためには，ネットワーク内の各BRが有するEAMTの一貫性が保証される必要がある．

2. **変更追従性**
   近年のIDCでは多数の物理サーバーを統合的に管理するプライベートクラウド環境やコンテナオーケストレーション環境[1]が普及しており，アプリケーション・サービスの追加及び削除が頻繁に行われている．サービスの障害を検知し，適切に冗長系に移行するための手法として，SLB(Servver Load Balancer)が広く利用されている．SIIT-DCのIPv4サービス提供の場でも，サービスの状態の変動にリニアに対応しフェイルオーバーできるような働きが求められる．

3. **スケーラビリティ**
   IPv6シングルスタックネットワークにおけるIPv4サービスの提供では水平スケールが容易に行える仕組みを備える必要がある．IPv4サービスを行うサーバの増設や，対外接続点が増えた場合のBRの拡大に十分に適用するスケーラビリティを有することが望ましい．

4. **デプロイメントの容易さ**
   SIIT-DCの最も特筆すべきメリットの一つにデプロイメントの容易さが挙げられる．これを損なうことくダイナミックEAMTを実現する必要がある．

[1]Container Orchestration. コンテナ型仮想化統合管理環境

4.3 アプローチの分類と比較

はダイナミック EAMT を実現するアプローチとして，二つのアプローチを考察する．それぞれのアプローチで考えられる実装と実際の構成，及び第 5.5 節で述べた各要件への適合性を定性的に評価する．

本節ではスケーラビリティの評価のために，制御に必要な通信コネクション数による比較を行う．以後 BR の数を M，IPv4 サービスを提供するサーバの数を N とし，総通信コネクション数を C として表現する．

4.3.1 中央管理型アプローチ

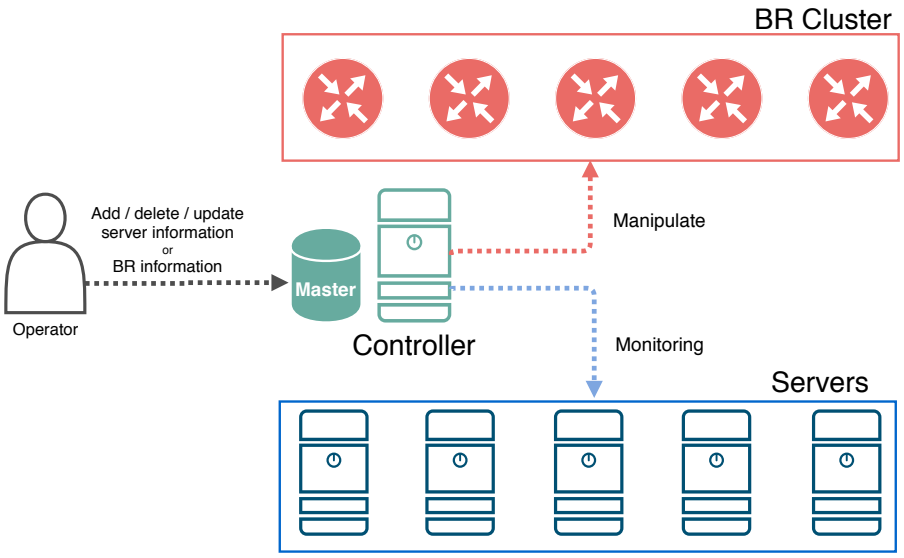

図 4.1: 中央管理型アプローチによるダイナミック EAMT

中央管理型アプローチとは，複数の BR の EAMT を統合的に管理する「コントローラ」を IDC ネットワーク上に配置し，各 BR がネットワークを介してこれを参照する機構である．図 4.1 に中央管理型アプローチによってダイナミック EAMT を実現した SIIT-DC の各コンポーネントの関係図を示す．

中央管理型アプローチではコントローラが各 BR に投入する EAM が記録された「マスターテーブル」を保持し，それを元に各 BR のデータプレーンにルールを書き込む手法を取る．マスターテーブルに記載される EAM はオペレーターがネットワークの構成変更に合わせて追加・削除・更新を行い，それぞれの IPv4 サービスを提供するサーバ群に対してはコントローラからプル型[2]の外部監視[3]によりサーバの状態変化を検知しマスターテーブルを更新する．

[2]pull-based monitoring. コントローラから各サーバに能動的に情報を取得する

[3]External monitoring. 監視対象でエージェントを稼働することなく，外部から得られる情報を利用して監視を行うこと．

本アプローチの実装手法としては，OpenFlow[4]などを用いた集中コントローラ型 SDN フレームワークを利用する方法が考えられる [15]．類似事例として，Sheng らによって Open Flow を利用して各アクセススイッチに IPv4/IPv6 トランスレーション機構をデータプレーンとして導入するデータセンターネットワークデザインの提案がなされている [16]．

要件評価

- BR 間の EAMT の一貫性
  本アプローチでは各 BR の EAMT が一つのマスターテーブルからレプリケーションされるために，十分な一貫性が保証される．
- 変更追従性
  基本的には EAM 情報の更新はオペレーターのマスターテーブルへの記入までの時間はコントローラのサーバ監視性能に依存する．
- スケーラビリティ
  コントローラの数を L とすると，EAMT の制御に必要とする総通信コネクション数 C は以下の通りになる．
  $$C = L(M + N) \tag{4.1}$$
  一方，変更追従性と同じく，管理対象のサーバの収容台数に関しては，コントローラの実装・性能がボトルネックとなる設計である．
- デプロイメントの容易さ
  コントローラに求められる機器の性能・機能要件が大きいため，標準的な SIIT-DC よりデプロイメントのコストは高い．

4.3.2 分散管理型アプローチ

分散管理型アプローチとは，IPv4 サービスを提供するサーバがエージェントプロセスを介して自身の IPv4 サービスアドレスと IPv6 サービスアドレスを広告し，その広告情報を受け取った BR が自身の EAMT に反映させる機構である．図 4.2 に中央管理型アプローチによってダイナミック EAMT を実現した SIIT-DC の各コンポーネントの関係図を表す．

サーバ群は各 BR と EAM を広告するためのコネクションを確立する．IPv4 サービスを提供するサーバと BR の間の IP ネットワークが何らかの原因により疎通不能になると，当該サーバの広告も同時に停止されるため，該当 BR の EAMT から該当する EAM のレコードが削除される．

[4]Open Networking Foundation により標準化されているデータプレーン制御用通信プロトコル．`https://www.opennetworking.org/`

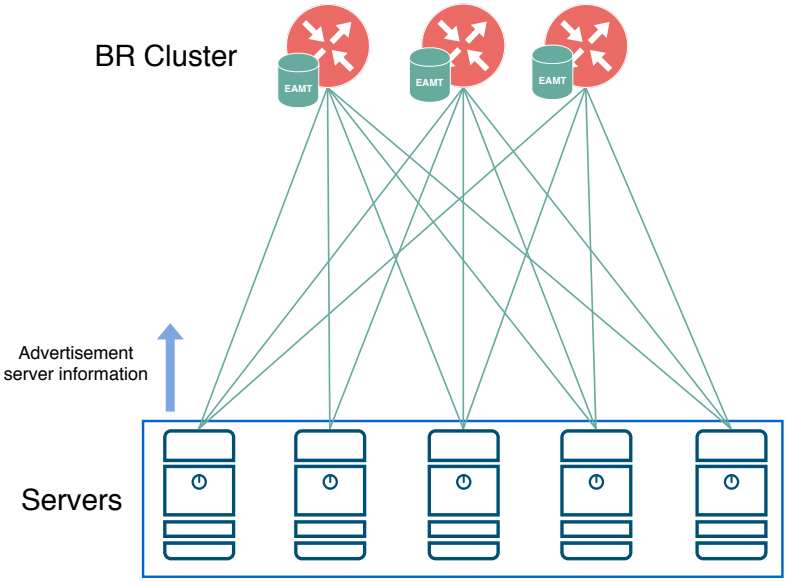

図 4.2: 分散管理型アプローチによるダイナミック EAMT

要件評価

- BR 間の EAMT の一貫性
  各 BR 間で EAMT の一貫性を保証する機構は無いが，当該 BR と疎通できないサーバは障害時に自身の IPv4 サービスアドレス宛のトラフィックを当該 BR に経由させることが出来ないため，問題にならない．

- 変更追従性
  サーバ自身のエージェントプロセスが直接 BR に広告を行うため，実際の変更にリニアに対応出来る．

- スケーラビリティ
  EAMT の制御に必要とする通信コネクション数 C は以下の通りになる．

$$C = M \cdot N \tag{4.2}$$

  サーバ群・各 BR 間でフルメッシュでのコネクションが必要なため，SIIT-DC ネットワーク自体が小規模の場合のみ採用可能である．

- デプロイメントの容易さ
  各サーバ・BR にエージェントを導入する必要があるが，システム自体の機能は軽量である．

表 4.1: 各アプローチの比較

手法	EAMT の一貫性	変更追従性	コネクション数	デプロイメントの容易さ
オペレーターによる手動設定	無し	無し	——	——
中央管理型アプローチ	有り	(監視機構の実装依存)	$\frac{L(2M+2N+L-1)}{2}$	困難 (コントローラーの実装依存)
分散管理型アプローチ	無し	有り	$M \cdot N$	有り

4.4 アプローチの検討

表 4.1 に 5.5 で述べたダイナミック EAMT に求められる要件に関する両アプローチの比較を示す．中央管理型アプローチが各 BR 間での EAMT の一貫性，スケーラビリティの二要素で優位であるが，コントローラの役割が非常に大きくなり機能要件が高くなるため，変更追従性とデプロイメントの容易さの面での障壁が高いという問題を抱えている．一方で分散管理型アプローチはシンプルな構成であるためデプロイメントが比較的容易であり変更への追従がリニアであるが，各サーバが通信コネクションを多量に貼らなくてはならない点でスケーラビリティに難がある．

第5章　ダイナミックEAMT実現手法の設計

第4章では，ダイナミックを設計する上で考えられる二種類のアプローチについて，求められる要件に照らし合わせて評価・検討を行った．本章では検討結果の得られた内容を基に，本研究において提案するダイナミック EAMT の実現手法の設計に関して論じる．

5.1　概要

第4.4で述べたように，分散管理型アプローチと中央管理型アプローチの双方に優位点があり，ネットワークやサービスの規模に合わせて柔軟に選択可能であると望ましい．

本研究では，動的経路制御プロトコルである BGP を利用したサーバ・BR 間のメッセージングにより，SIIT-DC ネットワークにおけるダイナミック EAMT 機構を提案する．本提案手法は，IBGP(Internal BGP)・RR(Route Reflector) 構成を採用することで，ネットワークやサービスの規模に合わせてスケールアウトすることが可能であり，両アプローチの優位点を備えていると言える．

5.2　BGP

5.2.1　概要

BGP とはインターネットにおいて自律システム[1]間の経路情報交換に用いられるパスベクタ型の動的経路制御プロトコルである．現在有効なバージョンは BGP4 であり，RFC4271 で定義されている [17].

5.2.2　用語

BGP において利用される用語のうち，本提案手法において重要なものを以下に列記する．

[1]Autonomous System. インターネットを構成するネットワークをそれぞれ独立的に運用する組織群を指す.

BGP スピーカ

BGP を実装された機器を BGP スピーカと呼ぶ.

BGP ピア

BGP で経路交換を行う関係にある機器をぞれぞれ BGP ピア (BGP Peer) と呼称する.

そのうち，自律システム間での接続関係にあるピアを EBGP(External BGP) ピア，同一自律システム内の BGP スピーカ同士の経路交換に用いられるピアを IBGP(Internal BGP) ピアと呼ぶ.

BGP コネクション

BGP コネクションとは BGP で経路交換に用いられる接続関係を指す．各機器は 1 対 1 の関係で BGP コネクションを確立する．BGP コネクションにはトランスポート層のプロトコルとして TCP[18] のポート番号 179 が利用され，フラグメンテーションや再送制御，応答確認，誤り制御等，TCP による高信頼なメッセージングが可能である．

また，BGP コネクションを維持・管理するために，BGP では以下のような 4 つのメッセージが定義されている．

BGP コネクションは BGP ピア間で TCP コネクションを確立したのちに OPEN メッセージにより各機能の対応関係を確認することにより確立され，KEEPALIVE メッセージによりセッションが維持される．UPDATE メッセージにより，BGP ピアへ広告する経路 (Adj-RIB-Out) に変更が生じたことを通知する．何らかの理由により BGP コネクションが確立出来なくった場合，NOTIFICATION メッセージを利用して切断を通知する．

Adj-RIB-In/Adj-RIB-Out/Loc-RIB

図 5.1 に BGP における経路受信・保持・送信の流れを示す．BGP ピアから受信した経路は Adj-RIB-In と呼ばれ，BGP スピーカの任意のフィルターやポリシーを適用した上で Loc-RIB と呼ばれるテーブルに保存される．BGP スピーカは Loc-RIB から任意のフィルターを適用した経路を BGP ピアに広告する．この広告する経路を Adj-RIB-Out と呼ぶ.

5.2.3 特徴

本提案手法においてダイナミック EAMT を実現するためのメッセージングプロトコルとして BGP を選択するに至った要素について述べる.

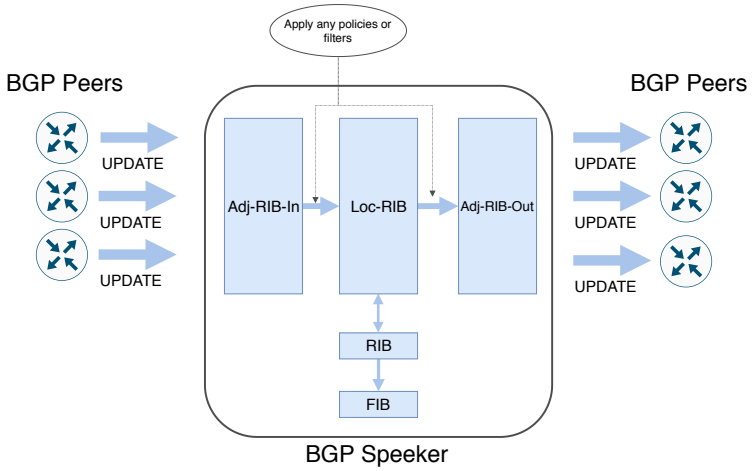

図 5.1: BGP スピーカの経路の扱い

マルチプロトコル

現行版である BGP4 では，OPEN メッセージにオプション値 (Capabilities Optional Parameter) を挿入することで，IANA によって定められた任意のネットワークプロトコル [19, 20] の経路を交換することが想定されている [21]．本提案手法で利用している IPv6 ユニキャスト経路もこの機構を用いる．

実装が一般的

BGP は自律システム間の経路交換プロトコルとしてインターネットで利用されているデフファクトスタンダードなプロトコルであり，OSS(Open Source Software)[2]にも多くの実装が存在する．広く普及したプロトコルを利用することにより，特別な実装を最小限にして本提案手法を実現することが出来る．

中継ネットワークに非依存

本提案手法で採用している IBGP では，TTL(Time to Live)[3]が 255 に設定されており，BGP ピア間で IPv4/IPv6 による到達性があればメッセージングを行うことが可能である．すなわち本提案手法は既存の SIIT-DC ネットワークに非依存であり，これは第 5.5 項で述べた要件の一つである，デプロイメントの容易さを充足する．

[2]ソースコードが公開されており，定められたライセンス規約に基づく範囲で自由に使用・改造が可能なソフトウェア．

[3]そのパケットが宛先ホストに到達するまでに許容される中継ルータ数．IPv6 プロトコルでは Hop Limit として同一の機能が実装されている [22]．

5.3 基本的なネットワーク設計

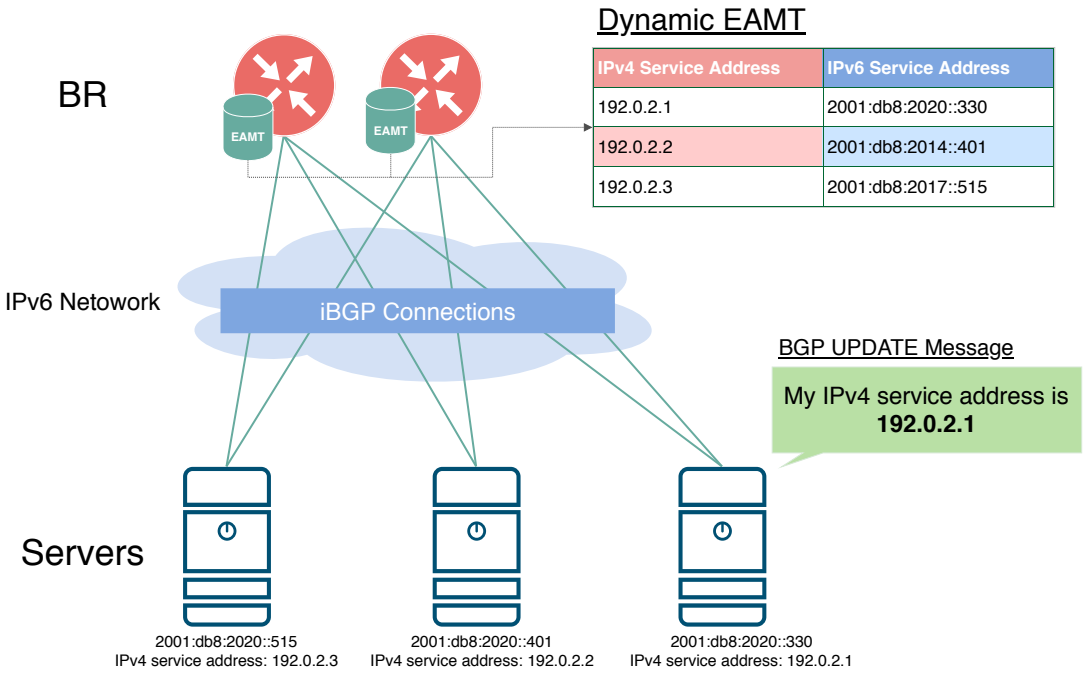

図 5.2: 本提案手法の基本機能を実装した SIIT-DC ネットワークの例

図 5.2 本提案手法の各要素の関係を示す.

BR 数を N，サーバ数を M とした，ルートリフレクタを利用した本提案手法での必要な BGP コネクション数 C_a は式 5.1 のように表現できる.

$$C_a = M \cdot N \tag{5.1}$$

5.3.1 各ノードの役割と機能要件

BR

BR では下記のような 3 つの機能が必要となる.

- BGP デーモン
  各サーバと IBGP コネクションを確立し，Loc-RIB を生成する.
- SIIT 機構
  EAMT を保持し，それを参照して IPv4/IPv6 プロトコル変換を行う.
- EAMT 制御機構
  BGP デーモンが有するの Loc-RIB を参照し，EAMT を更新する.

IPv4 サービス提供サーバ

IPv4 サービス提供サーバでは以下の 2 つの機構が求められる．

- IPv4 サービス
  IPv4 によりインターネットに提供したいサービスを稼働させる．
- BGP デーモン
  自身が提供する IPv4 サービスアドレスを含んだ情報を広告する．

5.4 ルートリフレクタを活用したネットワーク設計

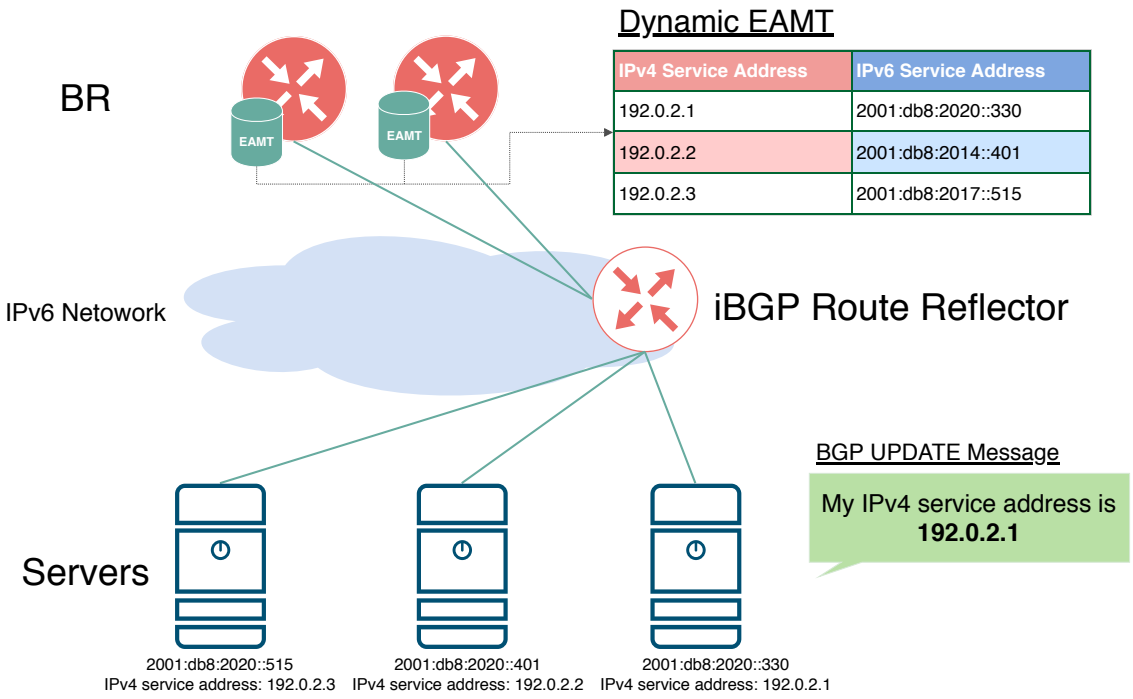

図 5.3: ルートリフレクタを採用した SIIT-DC ネットワークの例

通常，IBGP ではルートループを防ぐために異なる BGP ピアから受信した経路は他の BGP ピアに広告されない．そのため一つの IBGP スピーカが広告する経路を他の IBGP スピーカが受信するためには，BGP コネクションをフルメッシュで確立する必要がある [23]．

ルートリフレクタとは，Originator-ID と呼ばれる特殊な属性を Adj-RIB-Out に付与することでルートループを防ぎながら，IBGP ピアから受信した経路を他の IBGP に対して広告する特殊な BGP スピーカである [24]．ルートリフレクタは IBGP のコネクション数を削減するために広く利用されている．

ルートリフレクタを複数設置することで，負荷分散・冗長化構成を容易に実現することが出来る．一般的にはルートリフレクタ間はフルメッシュでの BGP コネクションを確立する設計を行うが，Gutiérrez らによればツリー型の BGP コネクション関係を一部で選択することにより，よりルートリフレクタに掛かる負荷を軽減出来ることが明らかになっている [25].

BR 数を N，サーバ数を M，ルートリフレクタの数を L とした，ルートリフレクタを利用した本提案手法での必要な BGP コネクション数 C_b は式 5.2 のように表現できる．なおルートリフレクタ間の BGP コネクションはフルメッシュを想定している．第 5.3 節で述べた基本的なネットワーク設計を行う場合と比較して，SIIT-DC ネットワークが大きくなった場合に BGP コネクションが大幅に削減できることがわかる．

$$C_b = \frac{L(2M + 2N + L - 1)}{2} \tag{5.2}$$

5.4.1 各ノードの役割と機能要件

BR 及び IPv4 提供サーバ

BR 及び IPv4 サービス提供サーバは，ルートリフレクタとのみ BGP コネクションを確立する．複数ルートリフレクタを配備する場合，それぞれとコネクションを確立することで冗長性を高めることが出来る．その他の機能は第 5.3 で述べたものと同様に配備する．

ルートリフレクタ

ルートリフレクタでは，ルートリフレクタ機能が有効となった BGP デーモンを配備する必要がある．各サーバ・BR と BGP コネクションを確立する．

5.5 各アプローチとの比較

第 4.3 節で検討した各アプローチと本提案手法を，第節で挙げた各性能要件に関して比較する．

表 5.1: 各手法の比較

手法	EAMT の一貫性	変更追従性	コネクション数	デプロイメントの容易さ
参考: オペレーターによる手動設定	無し	無し	——	——
中央管理型アプローチ	有り	(監視機構の実装依存)	$\frac{L(2M+2N+L-1)}{2}$	困難 (コントローラーの実装依存)
分散管理型アプローチ	無し	有り	$M \cdot N$	有り
提案手法 1: IBGP	有り	有り	$M \cdot N$	容易
提案手法 2: IBGP + ルートリフレクタ	有り	有り	$\frac{L(2M+2N+L-1)}{2}$	容易 (RR は容易にスケールアウト可能)

表 5.1 にそれぞれの項目における比較結果を示す．なお，コントローラーもしくはルートリフレクタの導入数を L，サーバ数を M，BR の数を N としている．

5.5.1 EAMT の一貫性

BGP はインターネットの経路広告手法として広く利用されており，多数のルータ間で一定の一貫性を保つことがプロトコルレベルで保証されている．本提案手法では BGP を利用したメッセージングにより，SIIT-DC ネットワークにおけるダイナミック EAMT を実現しているため，BGP と同水準の一貫性の担保が可能である．

5.5.2 変更追従性

本提案手法では各 IPv4 サービス提供サーバが自身の EAM を BGP の経路情報として広告するため，サーバが広告するモデルを採用する分散管理型アプローチと同様に，経路広告の有無によりサーバのネットワーク健全性を保証する事ができる．

一方で 4.3.1 項で述べたように，中央管理型アプローチにおいても変更追従性を実現可能であるが，コントローラの監視機構の実装に依存するほか，何らかの手段によってマスターテーブルに対して変更を適用する手段を考慮する必要がある．

5.5.3 コネクション数

各サーバが直接 BR に対してコネクションを確立する分散管理型アプローチ及び提案手法１では，サーバ・BR の数が大きくなった場合に，EAMT の維持・管理に必要なコネクション数が増加する．

一方でルートリフレクタを採用した提案手法２と中央管理型アプローチでは，サーバ・BR の数が増加した場合にも，上記２手法と比較して少ないコネクションで EAMT を維持・管理することが可能である．

5.5.4 デプロイメントの容易さ

本提案手法では動的経路制御プロトコルとして一般的な BGP をメッセージングに利用しているため，OSS を含む多種多様な実装をそのまま利用することが可能であり，他の手法と比較して実装・デプロイメントが容易である．

また 5.4 節で述べたように，ルートリフレクタ間のトポロジや管理方法を工夫することで，更なるコネクション数とルートリフレクタの負荷軽減を実現する余地がある．

一方で，中央管理型アプローチでは，特別な監視・管理機構を持ったコントローラを実装する必要があり，他のデプロイメントの障壁が高くなる．

第6章　プロトコル設計と実装

本章では，第5章で述べた提案システムのメッセージ設計と実装について述べる．

第7章　評価

本章では，第 5 章及び第 6 章で設計・実装に関して述べた本提案手法に関して，第 3.2 節で指摘した SIIT-DC の課題に対して有効性があることを評価する．

第8章　結論

本章では，本研究の総括と今後の課題を示す.

8.1　本研究のまとめ

謝辞

参考文献

[1] Cisco Systems. Cisco annual internet report（2018〜2023 年）ホワイトペーパー. https://www.cisco.com/c/ja_jp/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html, 2020.

[2] Daniel E. Eisenbud, Cheng Yi, Carlo Contavalli, Cody Smith, Roman Kononov, Eric Mann-Hielscher, Ardas Cilingiroglu, Bin Cheyney, Wentao Shang, and Jinnah Dylan Hosein. Maglev: A fast and reliable software network load balancer. In *13th USENIX Symposium on Networked Systems Design and Implementation (NSDI 16)*, pages 523–535, Santa Clara, CA, 2016.

[3] Parveen Patel, Deepak Bansal, Lihua Yuan, Ashwin Murthy, Albert Greenberg, David A Maltz, Randy Kern, Hemant Kumar, Marios Zikos, Hongyu Wu, et al. Ananta: Cloud scale load balancing. *ACM SIGCOMM Computer Communication Review*, 43(4):207–218, 2013.

[4] Congxiao Bao, Xing Li, Fred Baker, Tore Anderson, and Fernando Gont. IP/ICMP Translation Algorithm. RFC 7915, June 2016.

[5] Tore Anderson. SIIT-DC: Stateless IP/ICMP Translation for IPv6 Data Center Environments. RFC 7755, February 2016.

[6] Erik Nordmark. Stateless IP/ICMP Translation Algorithm (SIIT). RFC 2765, February 2000.

[7] Xing Li, Fred Baker, and Congxiao Bao. IP/ICMP Translation Algorithm. RFC 6145, April 2011.

[8] Tore Anderson and S.J.M. Steffann. Stateless IP/ICMP Translation for IPv6 Internet Data Center Environments (SIIT-DC): Dual Translation Mode. RFC 7756, February 2016.

[9] Xing Li, Mohamed Boucadair, Christian Huitema, Marcelo Bagnulo, and Congxiao Bao. IPv6 Addressing of IPv4/IPv6 Translators. RFC 6052, October 2010.

[10] Tore Anderson. Local-Use IPv4/IPv6 Translation Prefix. RFC 8215, August 2017.

[11] IANA. Internet protocol version 6 address space. https://www.iana.org/assignments/ipv6-address-space/ipv6-address-space.xhtml, 2019. 最終閲覧: 2019-12-17.

[12] Tore Anderson and Alberto Leiva. Explicit Address Mappings for Stateless IP/ICMP Translation. RFC 7757, February 2016.

[13] Kurt Erik Lindqvist and Joe Abley. Operation of Anycast Services. RFC 4786, December 2006.

[14] NPO 日本ネットワークセキュリティ協会 (JNSA). 情報セキュリティインシデントに関する調査報告書. https://www.jnsa.org/result/incident/2018.html, 2018. 最終閲覧: 2019-12-21.

[15] Evangelos Haleplidis, Kostas Pentikousis, Spyros Denazis, Jamal Hadi Salim, David Meyer, and Odysseas Koufopavlou. Software-Defined Networking (SDN): Layers and Architecture Terminology. RFC 7426, January 2015.

[16] S. Maojia, B. Congxiao, and L. Xing. A sdn for multi-tenant data center based on ipv6 transition method. In *2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference*, pages 190–195, May 2016.

[17] Yakov Rekhter, Susan Hares, and Tony Li. A Border Gateway Protocol 4 (BGP-4). RFC 4271, January 2006.

[18] Transmission Control Protocol. RFC 793, September 1981.

[19] IANA. Address family numbers. https://www.iana.org/assignments/address-family-numbers/address-family-numbers.xml, 2019. 最終閲覧: 2019-12-21.

[20] IANA. Subsequent address family identifiers (safi) parameters. https://www.iana.org/assignments/safi-namespace/safi-namespace.xhtml, 2019. 最終閲覧: 2019-12-21.

[21] Tony J. Bates, Ravi Chandra, Yakov Rekhter, and Dave Katz. Multiprotocol Extensions for BGP-4. RFC 4760, January 2007.

[22] Dr. Steve E. Deering and Bob Hinden. Internet Protocol, Version 6 (IPv6) Specification. RFC 8200, July 2017.

[23] Mythili Vutukuru, Paul Valiant, Swastik Kopparty, and Hari Balakrishnan. How to construct a correct and scalable ibgp configuration. 2005.

[24] Enke Chen, Tony J. Bates, and Ravi Chandra. BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP). RFC 4456, April 2006.

[25] E. Gutiérrez, D. Agriel, E. Saenz, and E. Grampín. Rrloc: A tool for ibgp route reflector topology planning and experimentation. In *2014 IEEE Network Operations and Management Symposium (NOMS)*, pages 1–4, May 2014.