ARISTA ホワイトペーパー

Arista FlexRoute™エンジン

アリスタネットワークスの Arista 7500 シリーズは、革新的なスイッチング・プラットフォームとして 2010 年 4 月に発表され、さまざまな賞を受けた製品です。データセンターのパフォーマンス、効率性、ネットワーク全体の信頼性を極限まで引き出すことができます。モジュール型データセンター・スイッチの他製品と比べて、スイッチングのパフォーマンスを 5 倍に高速化しながら、消費電力は 10 分の 1、占有スペースは 2 分の 1 へと、技術革新を続けてきました。

2013 年に発表された Arista 7500E シリーズは、機能を維持しながらポート密度とパフォーマンスを 3 倍に高めました。

それから 3 年後に登場した Arista 7500R ユニバーサル・スパイン・プラットフォームでは、パフォーマンスと密度が 3.8 倍以上になり、インターネットのフル・ルーティング・テーブルをサポートするなど、機能が大幅に向上しています。

FlexRoute™ は、Arista 7500R ユニバーサル・スパイン・プラットフォームと Arista 7280R ユニバーサル・リーフ・プラットフォームで、ハードウェア内に 100 万プレフィックスを超える IP フォワーディング容量を実現する技術です。このホワイトペーパーでは、FlexRoute エンジンについて詳しく説明します。

ARISTA FLEXROUTE エンジン

Arista FlexRoute エンジンは、インターネットのフル・ルーティング・テーブルをハードウェア内でサポートします。レイヤ 3 での IP フォワーディング機能を備え、将来、IPv4 と IPv6 の経路が 100 万経路以上に増加しても十分な余裕があります。革新的な FlexRoute エンジンは、Arista 7500R と 7280R のユニバーサル・スパイン/リーフ・プラットフォームにレイヤ 3 のフォワーディング・テーブルを作成するという、特許取得済みのアルゴリズム・アプローチを採用しています。これは、このプラットフォームのルーティング性能を最大化する、アリスタ独自のキーテクノロジーです。

ハードウェア面から見ると、FlexRoute は各ラインカード（図 1）またはシステム上の分散されたパケット・プロセッサで実行される入力パケット処理の一環として、IPv4 と IPv6 に対する LPM（プレフィックス最長一致）レイヤ 3 検索を実行します。

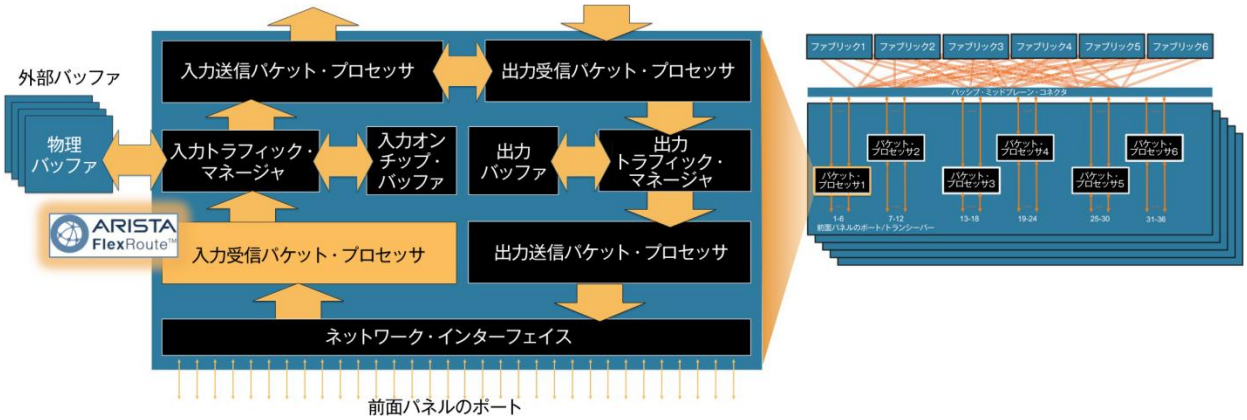

図 1: ラインカード上のパケット・プロセッサ内にある Arista FlexRoute エンジン

内部的には、アルゴリズム・アプローチを使用して検索を実行します。 従来の LPM アプローチと比較すると、FlexRoute は、アクティブではない（活動度が低い）シリコンと、使用効率に優れたトランジスタ（高密度ストレージ）を組み合わせて、LPM フォワーディング・テーブルを維持します。その結果、同じプロセス・ノードで別のアプローチを実行した場合と比べて、消費電力が劇的に低下し、処理ポート数とスループットが増加します。

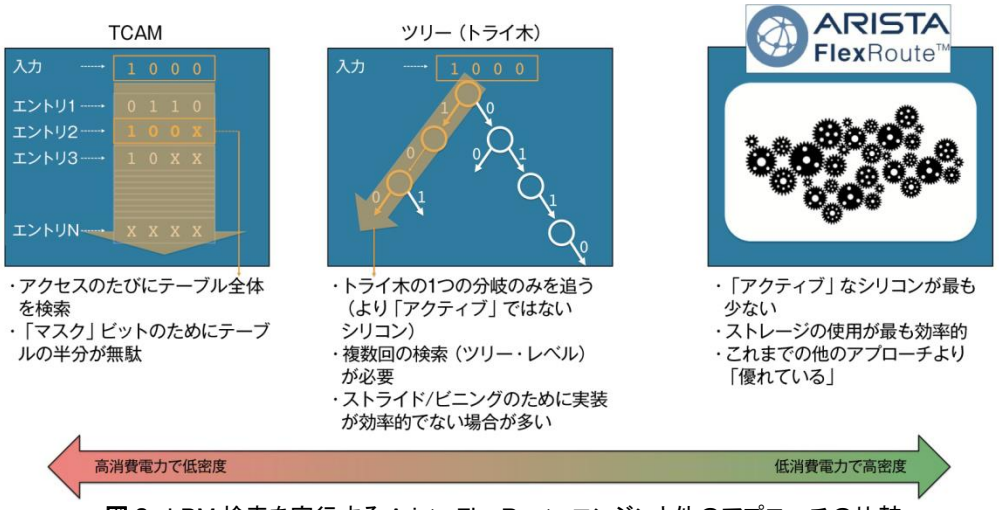

図 2: LPM 検索を実行する Arista FlexRoute エンジンと他のアプローチの比較

LPM 検索に使用するアルゴリズムは、インターネット・ルーティング・テーブルのこれまでの増大と、予測されるルーティング・テーブルの進化に関する既知のトレンドに基づいて最適化されます。たとえば、FlexRoute は、IPv4 プレフィックス空間のデアグリゲーション（/23 プレフィックスを 2 つの/24 プレフィックスに分割するなど）が引き続き増大するという予測に基づいて最適化されます。IPv6 アナウンスメントの急増（大部分のプレフィックス・アナウンスメントは/32 と/48 です）に基づいても最適化されます。LPM テーブルを拡大する従来の方法は、テーブルとメモリのサイズを増やす（つまりトランジスタを増やし、消費電力/発熱を増やし、ポート密度を低下させる）か、ツリー構造における検索深度を増やす（パフォーマンスを低下させる）というものでしたが、FlexRoute で使用するアルゴリズム・アプローチは、このようなトレンドとインターネット・ルーティング・テーブルの進化によって、さらに効率化されます。

パス、プレフィックス、インターネットの増大

FlexRoute で LPM を実現するために使用するアルゴリズムは、増大し続けるインターネット・ルーティング・テーブルに対して何年もの余裕をもたらすとアリスタは確信しています。100 万プレフィックスを超えてどこまで拡張できるのか、時間の経過とともに効率性がどのように高まるのかについてはここでは説明しませんが、インターネット・ルーティング・テーブルが現在のサイズ（2016 年 5 月：最大 649,000 プレフィックス [61 万の IPv4、最大 29,000 の IPv6]）までどのように増えてきたか、今後どのように増えると予測されるかをご覧ください。

過去、現在、未来のインターネットの増大

APNIC（Asia Pacific Regional Internet Registry）のチーフ・サイエンティストである Geoff Huston は、ほぼ 10 年間にわたり、グローバル・インターネット・ルーティング・テーブルに関する調査、分析、解説を行ってきました。2016 年 1 月、Geoff は APNIC Labs の一員として、このトピックについてのこれまでの分析と解説をもとに、2015 年のインターネット・ルーティング・テーブルに関する分析[1] を発表しました。

インターネットを構成する IPv4 およびと IPv6 プレフィックスの正確な数は、地域や、地域ごとのサマライズ によって異なりますが、プレフィックスの大まかな数はかなり明確になり、その結果としてトレンドも明確になっています。APNIC 地域内にあるオーストラリアと日本の観点から、受動測定ポイントとして AS131072 のグローバル・ルーティング・テーブルとそのデータを使用すると、収集されたデータは IPv4 および IPv6 プレフィックス空間が以下のように拡大していることを示しています。

表 1: IPv4 および IPv6 アナウンスメントのこれまでの増加（出典：Geoff Huston/APNIC Labs、表 1 と 2 は[1]から）

指標	2013 年 1 月	2014 年 1 月	2015 年 1 月	2016 年 1 月
IPv4 プレフィックス	441,000	488,000（+10%）	530,000（+9%）	587,000（+11%）
IPv6 プレフィックス	11,900	16,700（+40%）	21,000（+26%）	27,200（+30%）
合計（IPv4 + IPv6）	**452,000**	**504,700（+11%）**	**551,000（+9%）**	**614,200（+11%）**

Regional Internet Registry のプレフィックスの割り当てと実際のプレフィックス経路のアナウンスメント（より限定的なプレフィックスがアドバタイズされるなど）、およびそのトレンドが時間の経過とともにどのように増加していくかを考慮し、これまでのトレンドに基づくと今後のプレフィックスのアナウンスメント、アップデート、デアグリゲーションはどのようになるかという前提で、同じレポートで将来の増加についても予測しています。IPv6 の予測はやや困難なので、レポートでは線形増加（L）と指数関数的増加（E）の両方に基づいて予測を行っています。実際には、両者の間に収まる可能性が最も高くなります。

表 2: IPv4 および IPv6 アナウンスメントの将来予測される増加（出典：Geoff Huston/APNIC Labs、表 3 と 4 は[1]から）

指標	2016 年 1 月（実際）	2017 年 1 月（予測）	2018 年 1 月（予測）	2019 年 1 月（予測）	2020 年 1 月（予測）	2021 年 1 月（予測）
IPv4 プレフィックス	586,879	629,000（+7%）	675,000（+7%）	722,000（+7%）	769,000（+7%）	816,000（+6%）
IPv6 プレフィックス（L）	27,241	30,421（+12%）	35,113（+15%）	39,806（+13%）	44,498（+12%）	49,203（+11%）
IPv6 プレフィックス（E）		37,968（+39%）	51,303（+35%）	69,322（+35%）	93,669（+35%）	126,671（+35%）
合計（線形増加の IPv6）	**614,120**	**659,421（+7%）**	**710,113（+8%）**	**761,806（+7%）**	**813,498（+7%）**	**865,203（+6%）**
合計（指数関数的増加の IPv6）		**666,968（+9%）**	**726,303（+9%）**	**791,322（+9%）**	**862,669（+9%）**	**942,671（+9%）**

表 2 に要約されている[1]の予測はあくまでも予測ですが、基礎となるデータは、たとえ増加速度が急激でも、IPv4 および IPv6 プレフィックスのアナウンスメントの合計が累積 100 万エントリを超えるには 5 年以上の余裕があることを明らかに示しています。

BGP パス、経路、フォワーディング・エントリ

BGP のプレフィクスとパスと、フォワーディング・テーブルに保存されているエントリの関係について、しばしば誤解があります。

たとえば、3 つのアップストリーム・プロバイダー（BGP ネイバー）からトランジット・サービスを受けていて、各ネイバーが 60 万のプレフィックスを BGP で送信する場合、パスは 180 万（60 万 x 3 ネイバー）になりますが、プレフィックスは 180 万にはならず、一意の 60 万プレフィックスのままです。プレフックスがどのネイバーを使用して選択されるかは、BGP レベルで解決されます。1 つのプレフィックスに複数のイコール・コスト・パスがある場合は、経路のプレフィックスに ECMP（イコール・コスト・マルチ・パス）が使用されます。しかし、プレフィックスはいずれかのネクスト・ホップを指したり、ECMP の場合はネクスト・ホップ・エントリのグループを指したりするだけなので、プレフィックスの数はやはり 60 万しかありません。

BGP で受信するプレフィックスと、ルーティング・テーブル（RIB）およびフォワーディング・テーブル（FIB）にどのように保存されるかという関係を図 3 に示します。

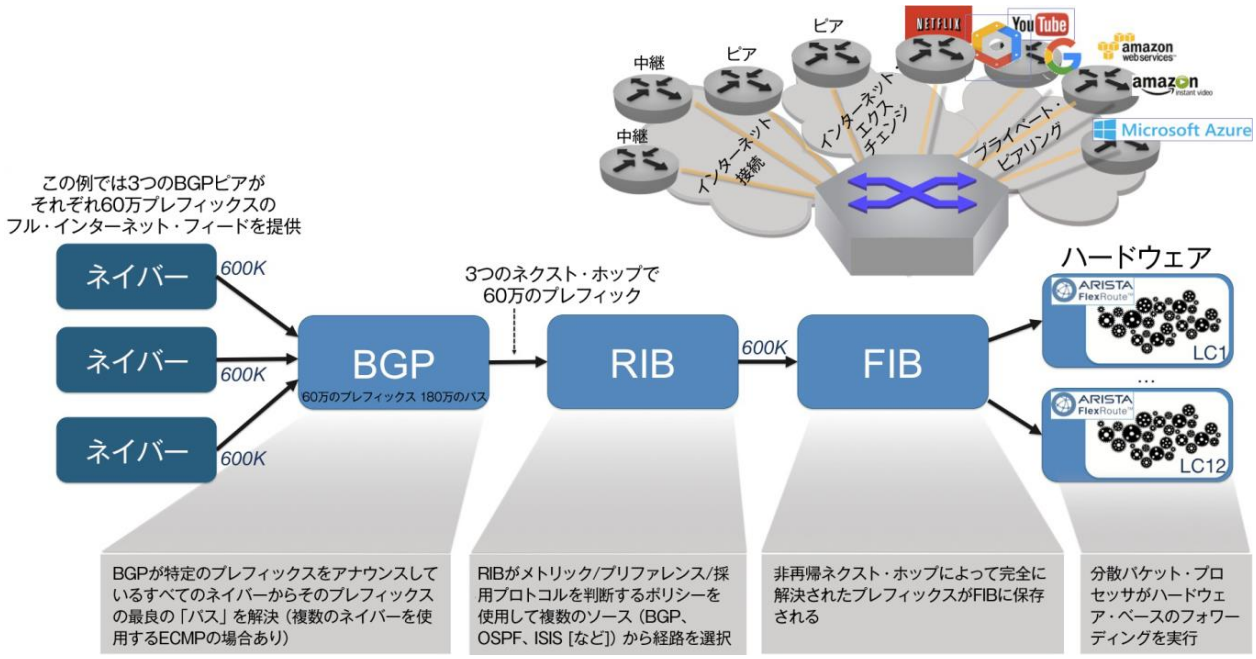

図 3: BGP で受信するプレフィックスと BGP から RIB、FIB への解決

トランジット・プロバイダーから受信するテーブル全体の数や、ピアの数にかかわらず、あるいは意図しないプレフィックスが間違ってアナウンスされても、中継プロバイダーやピアリング・プロバイダーが複数存在するためにプレフィックスの数が増えることはありません。

実際の FLEXROUTE のリソース使用率

アリスタの革新的な FlexRoute エンジンは、IPv4 と IPv6 を合わせて 100 万プレフィックス以上の容量を持つインターネット・ルーティング・テーブルとプレフィックスの分散を核として設計、構築されています。FlexRoute は、FlexRoute ライセンスと以下の CLI コマンドで有効になります。

```
arista(config)# ip hardware fib optimize prefixes profile internet
arista(config)# ipv6 hardware fib optimize prefixes profile internet
```

複数の導入事例で使用される実際のハードウェア容量とリソースの例を以下に示します。

実際の例 1：Internet2 エッジ・ルーター（IPv4 のみ）

IPv4 のこの導入例（Internet2 エッジ・ルーター）では、2 つの BGP ネイバーから最大 595,000 プレフィクスを受信しています。ルーティング・テーブル（RIB）の一意のプレフィックスは最大 579,000 になります。この例で最も使用されているハードウェア・リソースの使用率は 62%です。リソース使用率を表示するには、EOS コマンド「show hardware capacity」を入力します。

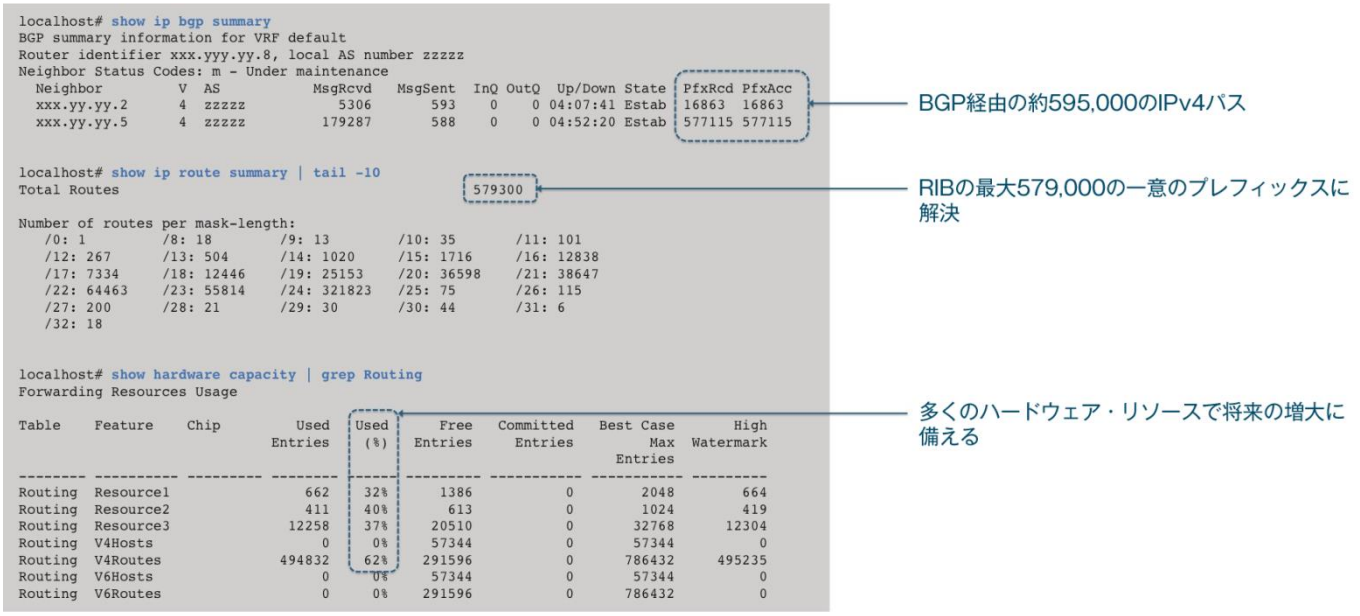

図 4：Internet2 に接続されているルーター

実際の例 2：フル IPv4/IPv6 を提供している大規模ホスティング・プロバイダー

この導入例では、大規模ホスティング・プロバイダーがデバイスをエッジ・ルーターとして使用し、多数の内部プレフィックスに加え、フル・インターネット IPv4 と IPv6 も提供しています。この例では、BGP 経由で最大 290 万の IPv4 パスと最大 204,000 の IPv6 パスを受信しています。ルーティング・テーブル（RIB）には、最大 854,000 の IPv4 プレフィックスと最大 45,000 の IPv6 プレフィックスが保存されます。合わせて最大 90 万プレフィックスなので、最も使用されているハードウェア・ルーティング・リソースの使用率は 83%になります。

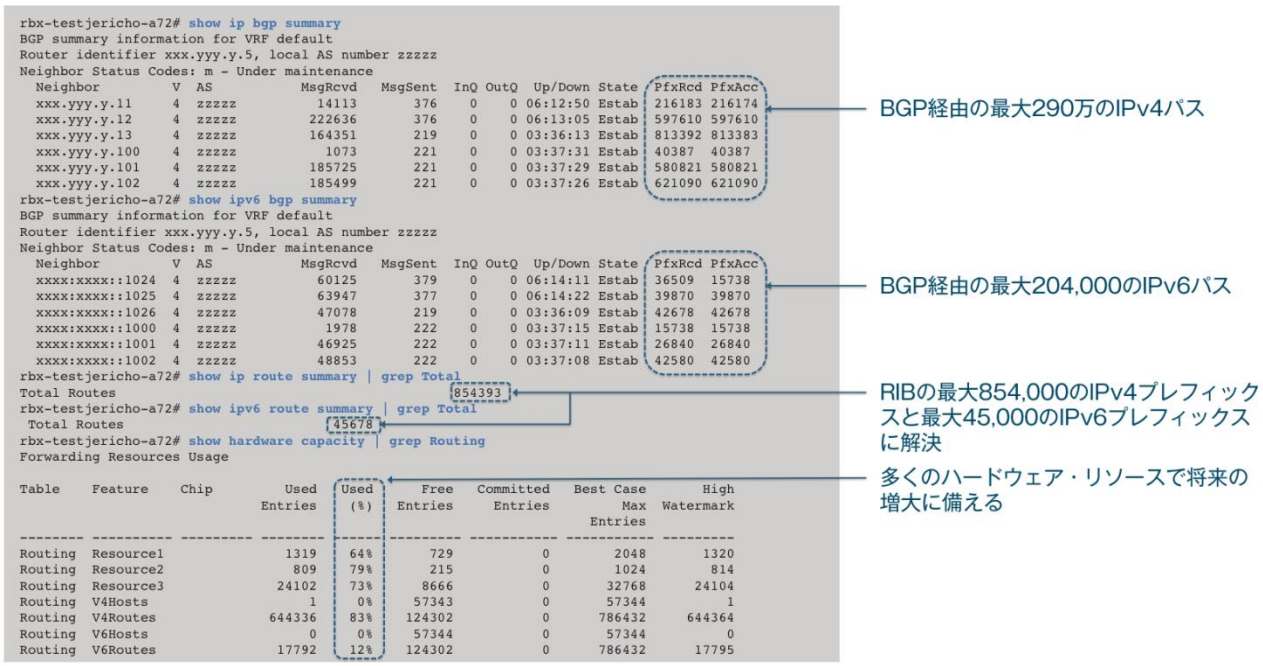

図 5：IPv4 と IPv6 の両方を提供する大規模ホスティング・プロバイダー

実際の例 3: 大手クラウド企業のフル IPv4/IPv6 インターネット・エッジ・ルーター

この導入事例では、大手クラウド企業がデバイスをエッジ・ルーターとして使用し、複数のトランジット・プロバイダー経由で IPv4 と IPv6 の両方を利用しています。この例では、最大 230 万の IPv4 パスと最大 14 万の IPv6 パスで IPv4 と IPv6 の両方を提供するフル・フィードが 4 つあります。ルーティング・テーブル(RIB)には、最大 575,000 の IPv4 プレフィックスと最大 35,000 の IPv6 プレフィックスが保存されます。この例で最も使用されているハードウェア・リソースの使用率は 88%です。

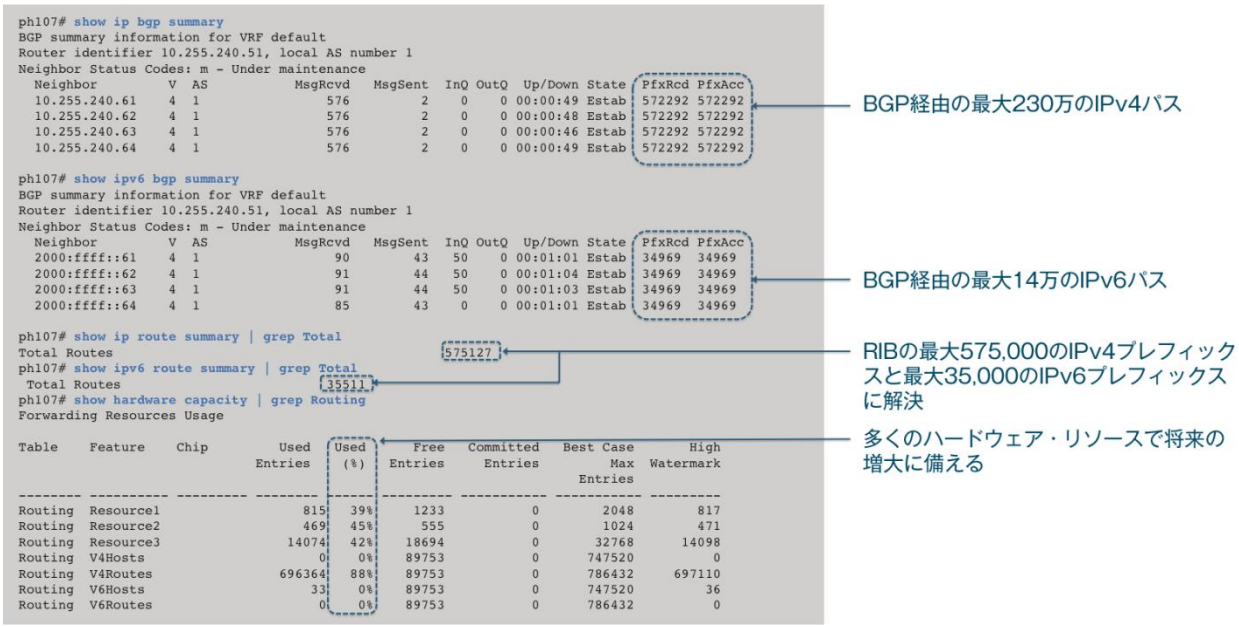

図 6: IPv4 と IPv6 の両方を提供する 4 つのフル・フィード(4 つの中継プロバイダー)を持つ大手クラウド・プロバイダー

ハードウェア・リソースのまとめ

アルゴリズム・アプローチを採用しているため、使用する正確なリソースは導入事例によって異なります。前述の例では、インターネットのフル・ルーティング・テーブルに対するフォワーディングに十分すぎる容量がありました。フォワーディング・リソースには、今後何年も増大する経路数をサポートするだけの余裕があります。

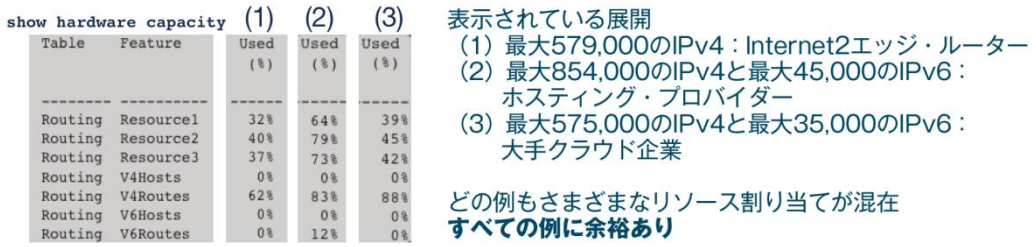

図 7: 例のハードウェア・リソース使用率のまとめ

アリスタは、FlexRoute で使用するアルゴリズムと手法を引き続き改良し、さらなる容量の強化を目指しています。

ARISTA EOS、SYSDB、NETDB

Arista 7500R と 7280R のユニバーサル・スパイン/リーフ・プラットフォームの中心となるのは、Arista EOS®(Extensible Operating System)です。EOS は、コアとなるソフトウェアの構成要素として、モジュール性、プログラマビリティ、フォールト・コンテインメント(障害の封じ込め)、対障害性を備えた、マルチプロセス状態共有アーキテクチャを強力な基盤としています。

システム状態を効率性に優れた一元的なシステム・データベース(SysDB)に保存し、自動パブリッシュ/サブスクライブ/通知モデルを使用してアクセスします。また、NetDB を使用してルーティング・スタックを拡張でき、従来のルーターや、ルーターのコントロール・プレーンを使用する旧来型アプローチよりも高速なコンバージェンスで何百万もの経路と何百ものネイバーをサポートします。

多くのネットワーク・ベンダーは、自社のコントロール・プレーンは高速で、拡張性があり、堅牢だと主張しますが、障害に反応するのに何秒もかかったり、経路をハードウェア内にプログラムするのに何分もかかる可能性があるというただし書きが付いています。Arista EOS には業界屈指のコンバージェンスと経路のプログラミング能力があり、中断に対して 1 秒以内（一般的にはミリ秒単位）の対応時間で対応します。旧来型アプローチと異なり、FlexRoute が考慮すべき重要事項は、データプレーンでの高速なプレフィックス・プログラミングと、ハードウェア内のフォワーディング・テーブルの Make-Before-Break プログラミングをサポートして、隣接エントリを中断させないようにすることです。

まとめ

アリスタの FlexRoute エンジンは、インターネットのルーティング・テーブル全体をハードウェア内でサポートします。レイヤ 3 の IP フォワーディング機能を備え、将来、IPv4 と IPv6 の経路がどちらも 100 万経路以上に増えても十分な余裕があります。革新的な FlexRoute エンジンは、Arista 7500R と 7280R のユニバーサル・スパイン/リーフ・プラットフォームにレイヤ 3 のフォワーディング・テーブルを作成するという特許取得済みのアルゴリズム・アプローチを採用しています。これは、このプラットフォームのルーティング性能を最大化する、アリスタ独自のキーテクノロジーです。

リファレンス

[1] 2015 年のインターネット・ルーティング・テーブルに関する分析、Geoff Huston（APNIC）。
https://labs.apnic.net/?p=767

アリスタネットワークスジャパン合同会社

〒170-6045　東京都豊島区東池袋 3-1-1　サンシャイン 60　45F
Tel:03-5979-2012(代表）

西日本営業本部
〒530-0001　大阪市北区梅田 2-2-2　ヒルトンプラザウエストオフィスタワー19 階
Tel: 06-6133-5681

お問い合わせ先
japan-sales@arista.com

www.arista.com/jp

2016 年 6 月