RPKI in DCの運用から見るBGP送信元検証の課題と未来

LINEヤフー株式会社
坂本 要

LINEヤフー

© LY Corporation

登壇者紹介

坂本 要

JANOG歴

初参加 JANOG38 若者支援プログラム

所属

ネットワーク本部 ネットワーク1部 サービスネットワーク1

業務内容

データセンターネットワークの構築運用
運用改善のためのソフトウェア開発運用
ネットワーク監視、ソフトウェア実行基盤の構築運用
新技術の調査検証

© LY Corporation

本プログラムの概要

本発表では、インターネットの経路送信元検証のために生まれたRPKIのテクノロジーをデータセンターネットワークに適用したLINEヤフーの事例を紹介します。

課題解決のためにRPKIを選択した背景とRPKIをデータセンターネットワークに適用するために開発したシステムの構成を、デモも交えて共有します。

最後に検証や開発の過程で感じた課題などを踏まえて、送信元検証のあり方について議論できればと思います。

© LY Corporation

目次

01 LINEヤフーのデータセンターネットワーク

02 送信元検証の課題と注目した解決策

03 データセンターネットワークでのRPKI適用までの道のり

04 デモ

05 まとめ・議論

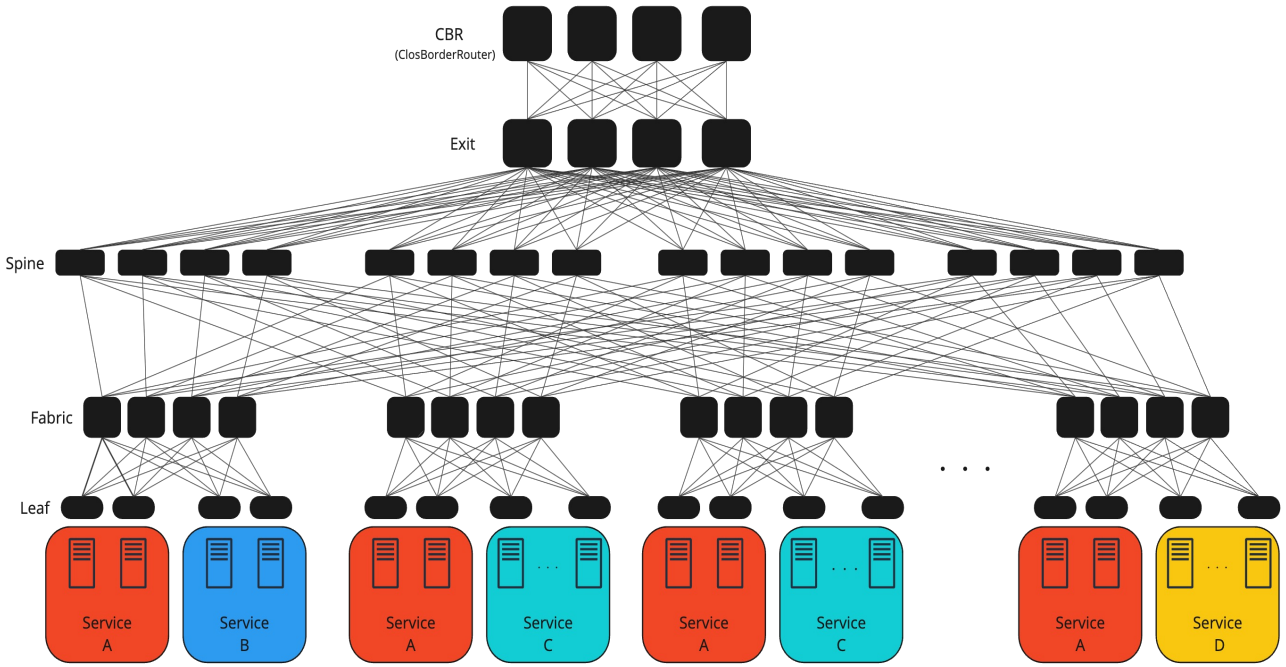
© LY Corporation

LINEヤフーの
データセンターネットワーク

© LY Corporation

全体概要

Clos Networkを利用



- Clos Network Topology
- BGP Routing on the Host
  - サーバやロードバランサまでBGP接続を延伸
- BGPの運用ナレッジを最大限活用可能な構成
- 6年以上の安定した商用運用実績

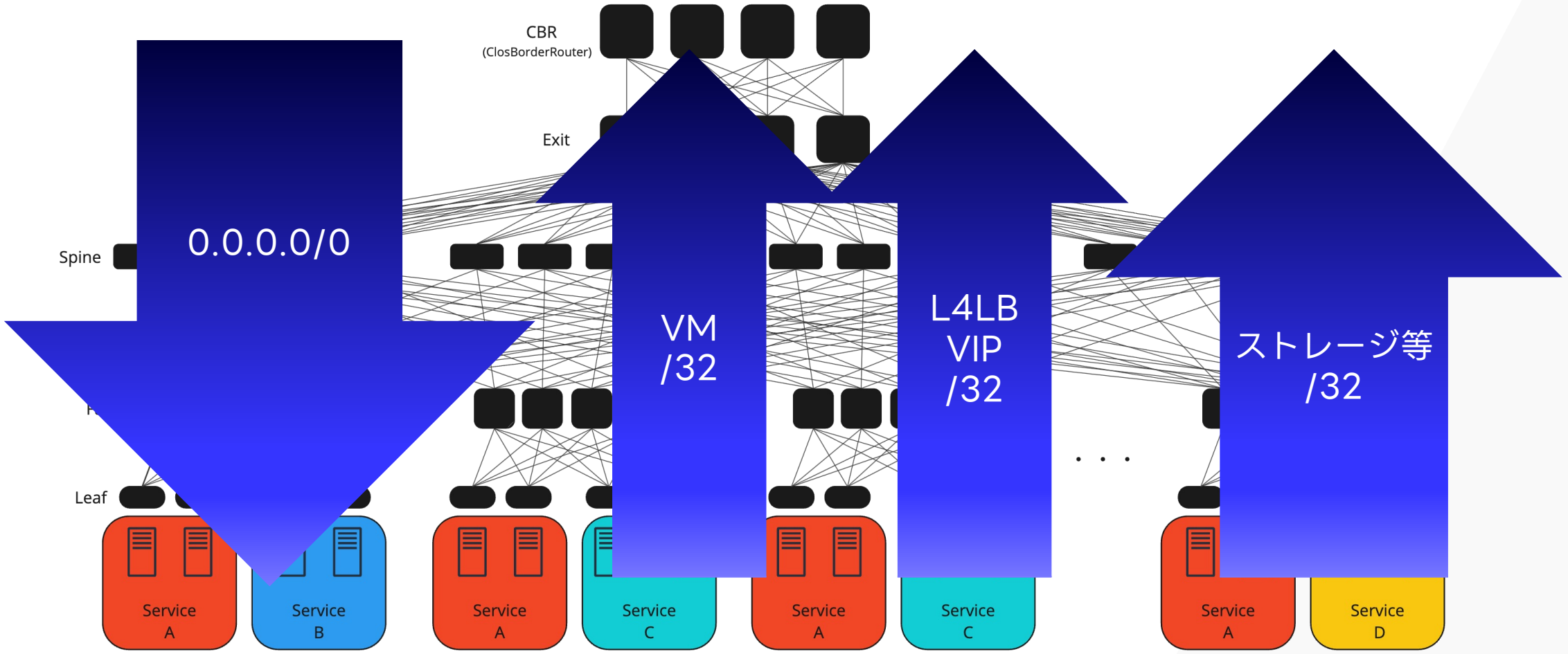
https://www.janog.gr.jp/meeting/janog43/application/files/7915/4823/1858/janog43-line-kobayashi.pdf

https://www.janog.gr.jp/meeting/janog46/wp-content/uploads/2020/06/janog46-yahoo-clos-takahashi.pdf

© LY Corporation

Clos Network ルーティングアーキテクチャ

設計



上流への経路はHVや物理サーバ、ストレージアプライアンス等から広報される

© LY Corporation

BGPで接続されるプラットフォームの多様化

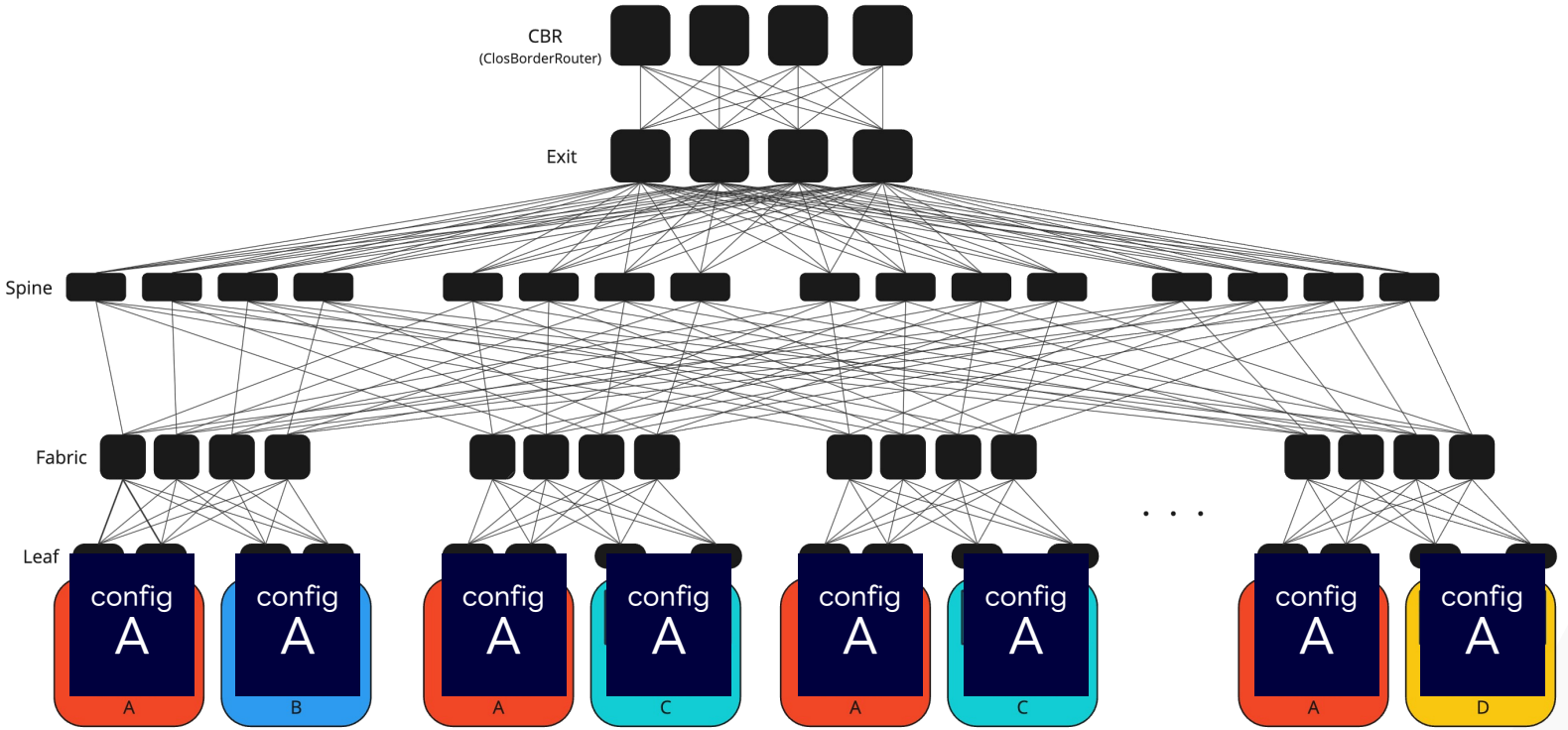
Clos Networkが社内の主な構成へと変化することで接続されるプラットフォームも多様化

- Hypervisor
- ストレージ
- 物理サーバ
- ロードバランサ

これらはネットワークチーム以外が運用主体であり広報される経路の**信頼**と**検証**の難しさが浮き彫りに

© LY Corporation

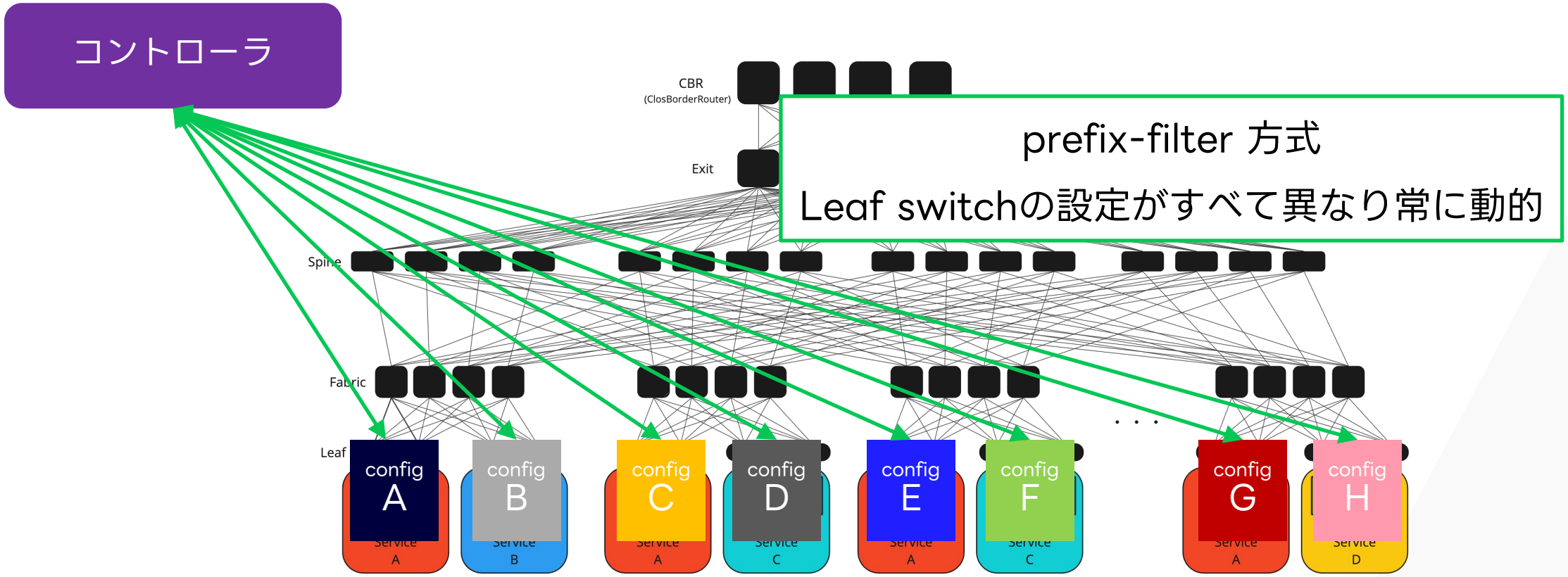
旧ヤフーの経路制御ポリシー



- 下流から0.0.0.0/0の受信を拒否、その他の経路は許可
- 構築後に機器の設定が変化しない
- 外部のコントロールプレーンを必要とせず、シンプルで楽だった

© LY Corporation

旧LINEの経路制御ポリシー



- SDNベースのコントローラでフィルタ制御
- 構築作業+経路フィルタ+運用を一括管理するオーケストレータを自社開発
- 過去4年間の運用で非常にうまく機能した

© LY Corporation

送信元検証の課題と注目した解決策

© LY Corporation

現在の送信元検証の課題

旧各社で分かれている経路制御ポリシー

会社の統合でネットワークの経路制御ポリシーを合わせたい

Clos NWで採用する機種の多様化

新しいNOSを採用するたびにオーケストレータの改修が必要

動的に変更される設定

テンプレート化できない

反映までにラグがある

© LY Corporation

過去に発生した送信元検証に関わる事故の事例

規模の拡大・多様化につれて表面化

Kubernetes内部経路を意図せず広報

- CNIプラグインによって作成された仮想NICにIPアドレスが付与
- 不適切な設定によりそれらが意図せず広報された

L3DSRのVIPをリアルサーバが広報

- リアルサーバのloに設定したVIPがそのまま広報された
- VIPに対する通信がLBを通さずにリアルサーバに到達する事態に発展
- LBとリアルサーバでECMP
  - クライアントの位置次第ではリアルサーバ側に吸い込まれる

いずれも送信元検証が適切に行われていれば防げたケース

© LY Corporation

BGP送信元検証

課題の解決策

解決すべき課題の本質

- BGP Mis-Origination の
  - 発生防止
  - 発生検知
  - 影響最小化

が可能

課題解決にむけた方針

- 標準化されたプロトコル
- 実装が広く普及し安定
- 運用負荷が低い

採用する技術

- RPKI

© LY Corporation

データセンターネットワークでのRPKI適用までの道のり

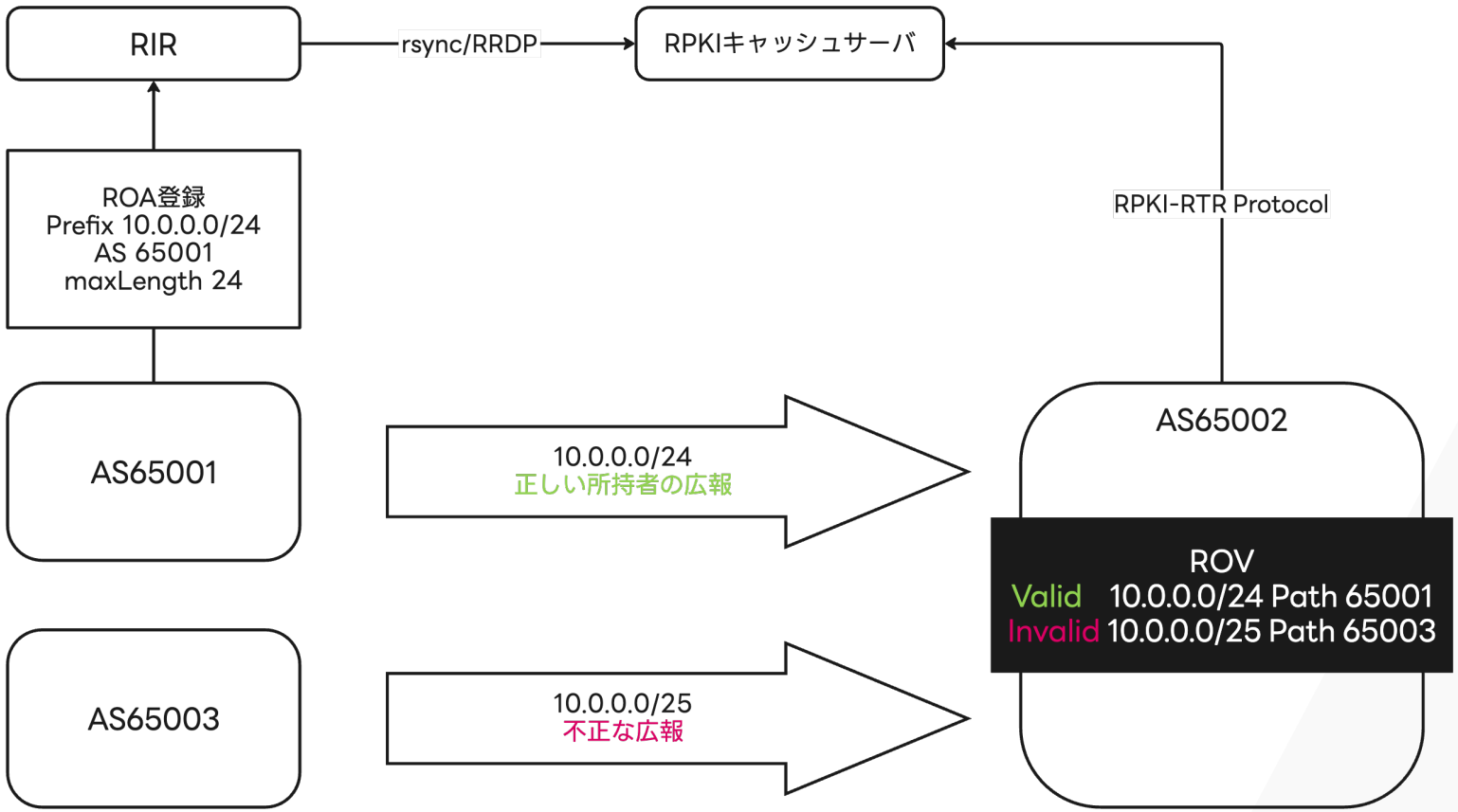
© LY Corporation

RPKIの基本

RPKI（Resource Public Key Infrastructure）

IPアドレスやAS番号と、その正当な保持者の関連を証明するための仕組み。

ROA（Route Origin Authorization）の作成と、ROV（Route Origin Validation）の実施で効果を発揮する。



© LY Corporation

データセンターネットワークのRPKI

インターネットRPKIとの違い

インターネット

- 異なる組織の集合
  （Inter-domain routing）
- 複雑な仕組みが必要
  - TAL、RP、証明書検証、署名

データセンターネットワーク

- 自組織で完結
- RIRのような正しさを担保する第三者が存在しない
- 自分たちが作成した仕組み、情報を信頼
- 実装はシンプルにできる
  - AS、Prefix、Maxlenの組み合わせが照合できれば良い
  - 認証局や証明書の検証等は不要

© LY Corporation

RPKI in DCで達成したいこと

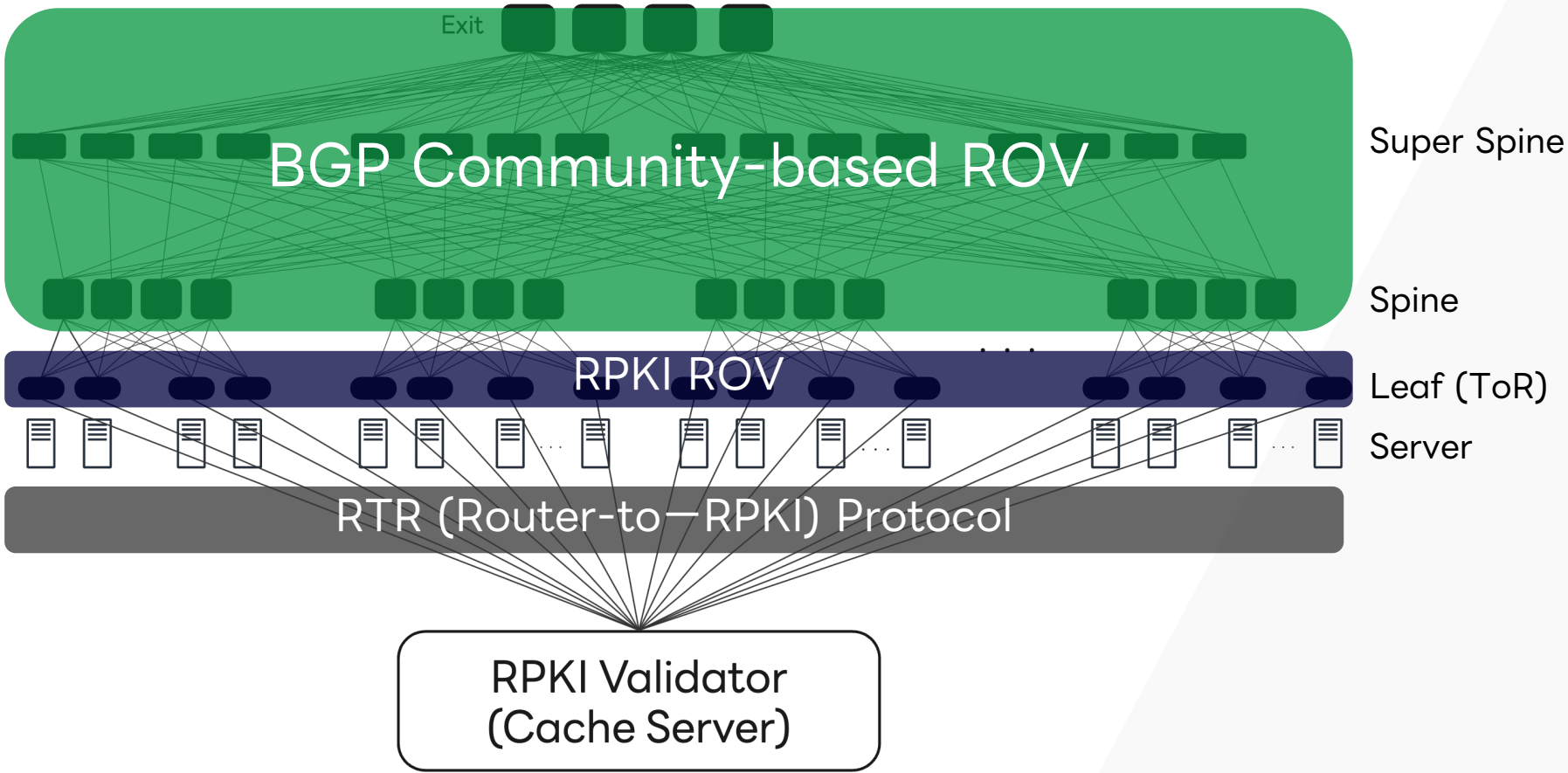
経路ハイジャックの防止と発生検知

- 経路ハイジャックが発生してもネットワークに影響が出ないようにしたい
  - 適切な制御とハイジャックが発生したことを検知できるようにする必要がある
- 従来の解決策である経路フィルタでは発生の検知はできない
  - 単にdropするのではなく、優先度を下げるような柔軟な処理もできない
- 経路ハイジャックの発生が即座に検知できる仕組み

© LY Corporation

RPKI in DCのためにNOSに必要な実装

- LeafはRPKIキャッシュサーバと同期したデータを元にROV
- ROVステータスをRFC8097で定義されたBGPコミュニティを使用して上流に通知する
- Leafより上流の機器はBGPコミュニティをベースにROV



© LY Corporation

RFC 8097

BGP Prefix Origin Validation State Extended Community

BGP Community-based ROVのためにRFC 8097で定義された拡張コミュニティを使用

検証状態を下記の拡張コミュニティを使って下流から上流に送信する

Meaning	Value
Valid	0x4300:0.0.0.0:0
Unknown (Not Found)	0x4300:0.0.0.0:1
Invalid	0x4300:0.0.0.0:2

https://datatracker.ietf.org/doc/html/rfc8097

© LY Corporation

LeafのRPKI ROV

サーバから受信した経路を検証

```
Leaf#show ip bgp
 Network
 * >   V 192.0.2.1/32          # valid
 * >   I  198.51.100.0/32      # max length invalid
 * >   I  203.0.113.0/32       # origin asn invalid
 * >   U 10.0.0.0/24           # roa not found - unknown
 * >      10.0.0.1/32          # Except ROV – local originated
 * >   V 10.100.0.0/24         # valid
```

```
> cat /etc/stayrtr/rpki.json
{
        "roas": [
                {"prefix": "10.100.0.0/24","maxLength": 24,"asn": 65010},
                {"prefix": "192.0.2.0/24","maxLength": 32,"asn": 65010},
                {"prefix": "198.51.100.0/24","maxLength": 24,"asn": 65010},
                {"prefix": "203.0.113.0/32","maxLength": 32,"asn": 65200}
        ]
}
```

ROA取得

RPKI Validator
(Cache Server)

update

RTR (Router-to-RPKI) Protocol

Leaf
ROV Device

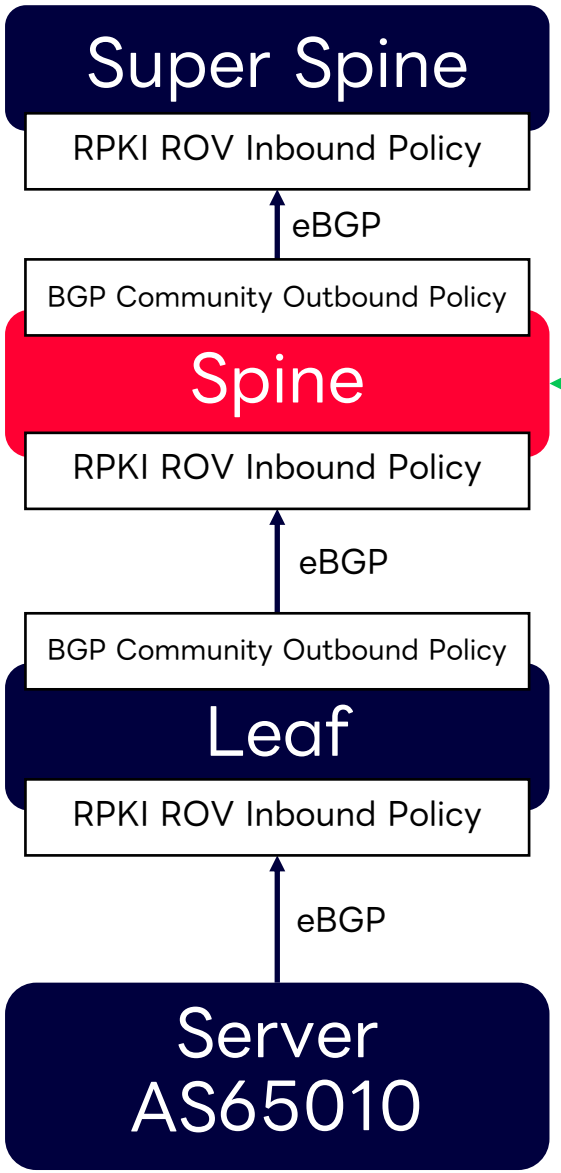
192.0.2.1/32
198.51.100.0/32
203.0.113.0/32
10.0.0.0/24
10.0.0.1/32
10.100.0.0/24

eBGP

Server
AS65010

© LY Corporation

SpineのRPKI ROV



RPKI BGP Community (RFC8097 non-transitive) を透過的に扱いたい

- コミュニティに基づいてローカルでROVを実行する
- ROV実行後、Community値を再付与し上流のeBGPネイバーに通知する

```
Spine#show ip bgp 198.51.100.0/32 detail
BGP routing table information for VRF default
Router identifier 10.0.0.3, local AS number 65100
BGP routing table entry for 198.51.100.0/32
 Paths: 1 available
  65000 65010
    10.0.3.1 from 10.0.3.1 (10.0.0.1)
      Origin INCOMPLETE, metric 0, localpref 100, IGP metric 0, weight 0, tag 0
      Received 00:00:10 ago, valid, external, best
      Extended Community: Origin-Validation-State:invalid
      Rx SAFI: Unicast

Spine#show ip bgp 192.0.2.0/32 detail
BGP routing table information for VRF default
Router identifier 10.0.0.3, local AS number 65100
BGP routing table entry for 192.0.2.0/32
 Paths: 1 available
  65000 65010
    10.0.3.1 from 10.0.3.1 (10.0.0.1)
      Origin INCOMPLETE, metric 0, localpref 100, IGP metric 0, weight 0, tag 0
      Received 00:00:29 ago, valid, external, best
      Extended Community: Origin-Validation-State:valid
      Rx SAFI: Unicast
  Not advertised to any peer.

Spine#show ip bgp 10.0.0.0/24 detail
BGP routing table information for VRF default
Router identifier 10.0.0.3, local AS number 65100
BGP routing table entry for 10.0.0.0/24
 Paths: 1 available
  65000 65010
    10.0.3.1 from 10.0.3.1 (10.0.0.1)
      Origin INCOMPLETE, metric 0, localpref 100, IGP metric 0, weight 0, tag 0
      Received 00:01:30 ago, valid, external, best
      Extended Community: Origin-Validation-State:not-found
      Rx SAFI: Unicast
  Not advertised to any peer.
```

© LY Corporation

RPKI in DC

RPKI 設定例

```
rpki cache stayrtr
     host ******* vrf management
     refresh-interval 5
     retry-interval 30
     expire-interval 600
     !
     transport tcp
!
rpki origin-validation
     ebgp local
     ebgp send
```

Leaf設定イメージ

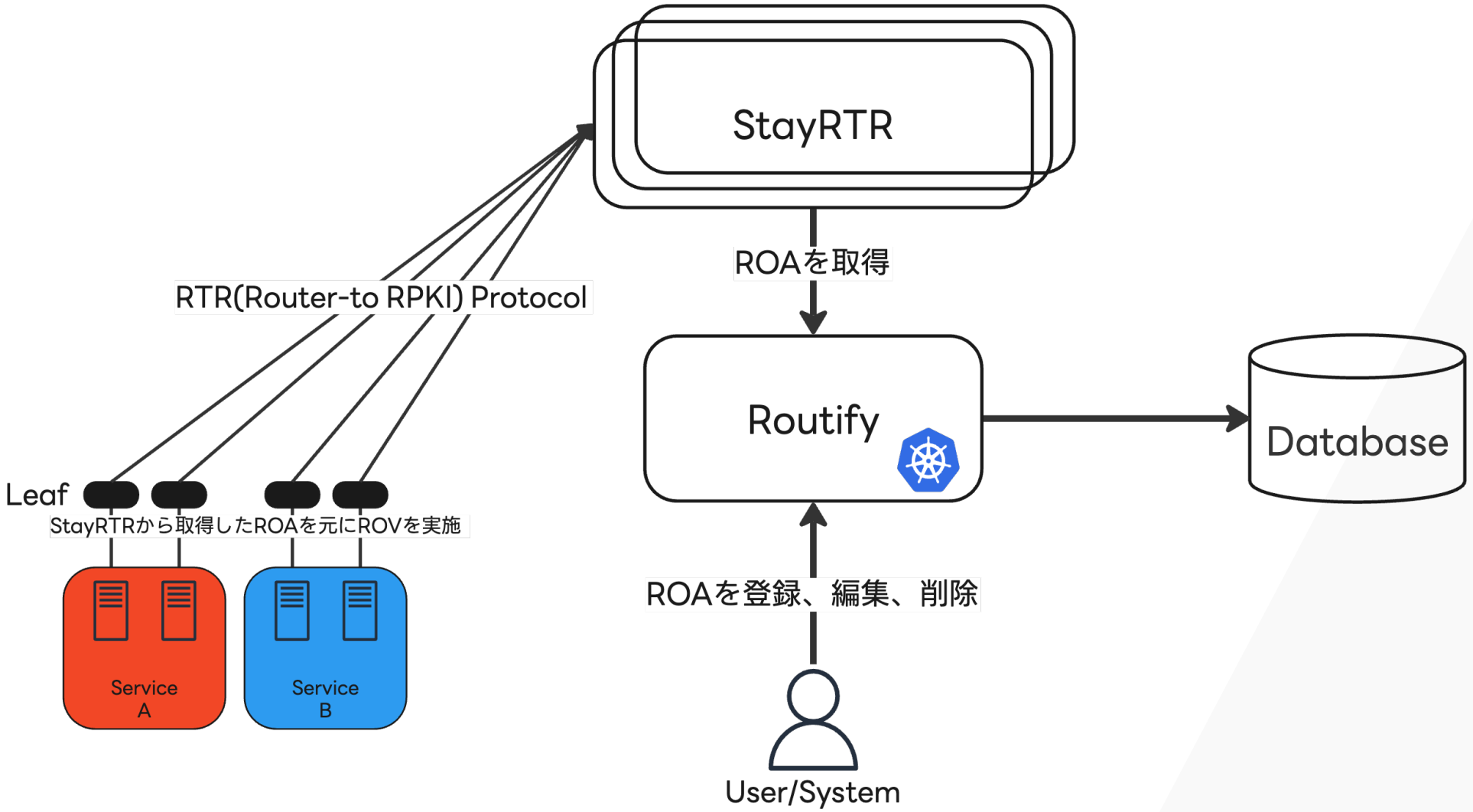
```
rpki origin-validation
     ebgp community
     ebgp send
```

Spine設定イメージ

© LY Corporation

RPKI in DC

システム概略図

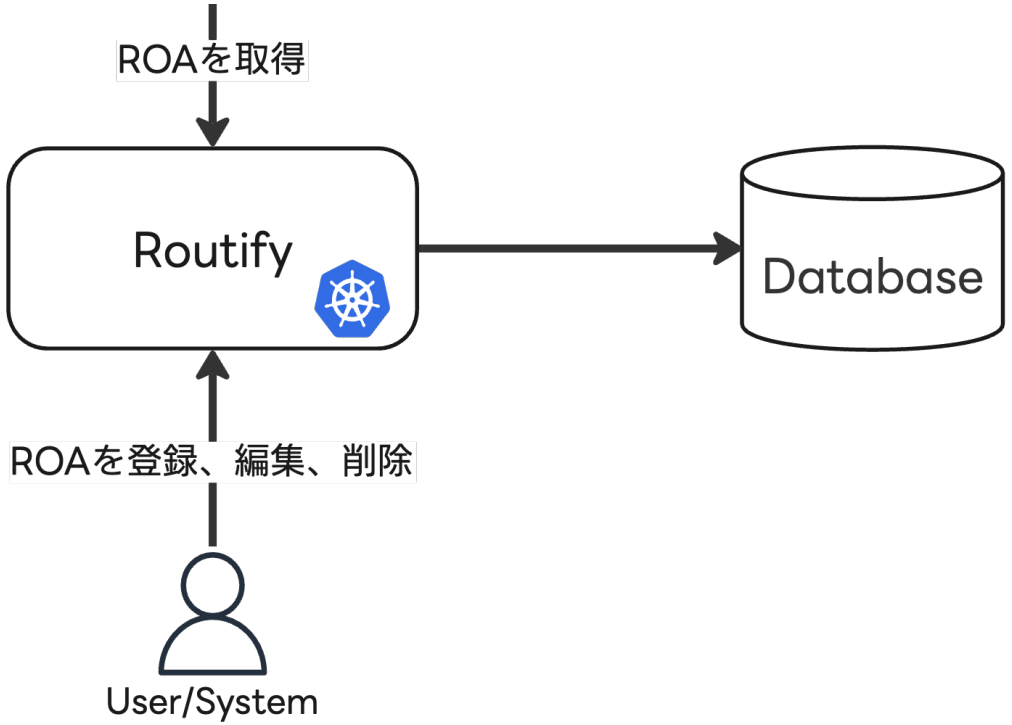


© LY Corporation

RPKI in DC

RPKIキャッシュサーバ

Routify



Go言語で実装した内製ソフトウェア

Kubernetesクラスタ上にデプロイ

ユーザや連携システムはRoutifyにROAを登録する

　CLIかREST APIで登録、削除が可能

登録されたROAはDBに書き込まれる

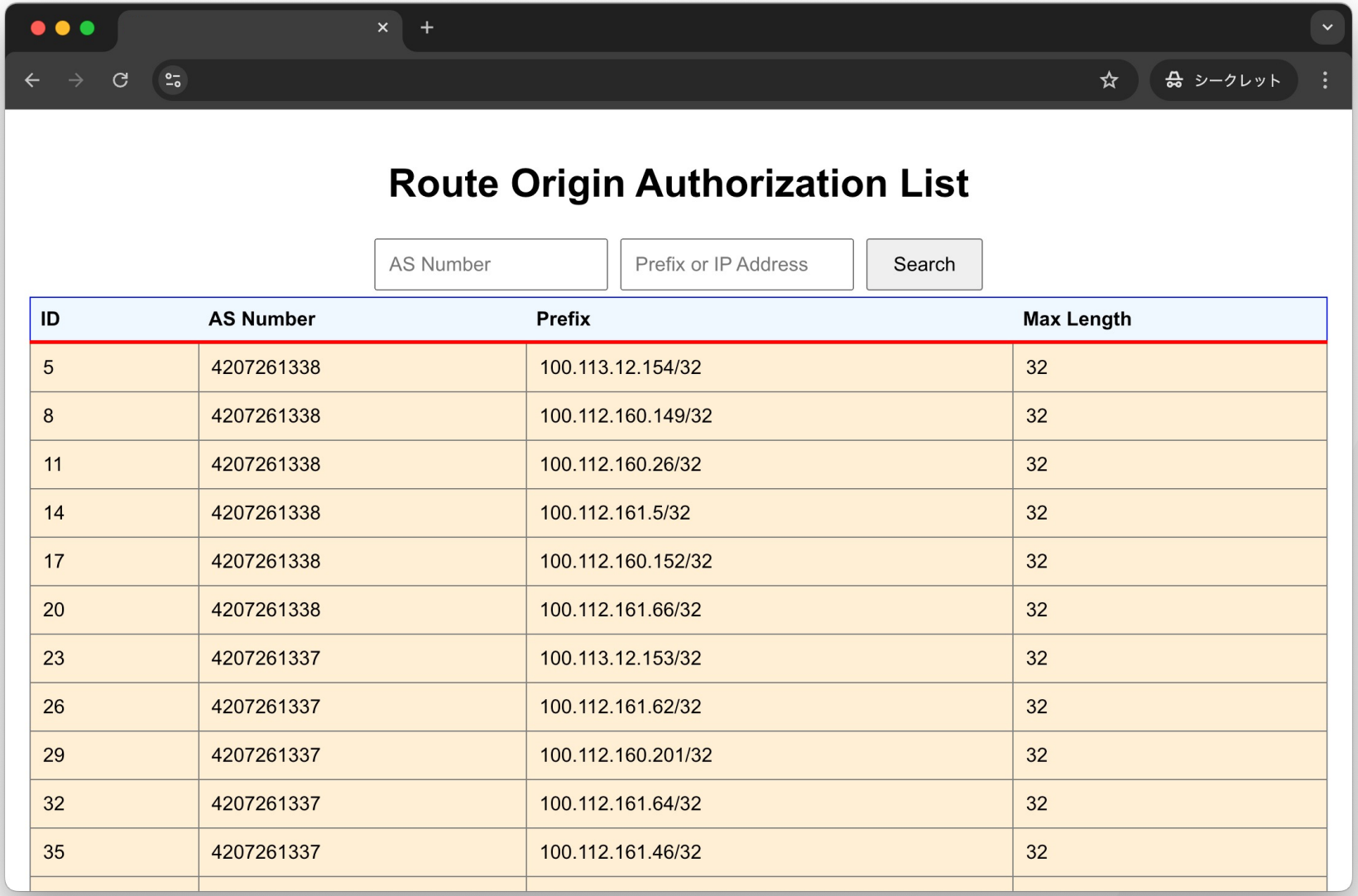
既存システムの置き換え、社内の各種ツールとの連携を念頭に内製を選択

　例: HypervisorはVM作成時にRoutifyエンドポイント経由でROAを登録する

© LY Corporation

RPKI in DC

Routify

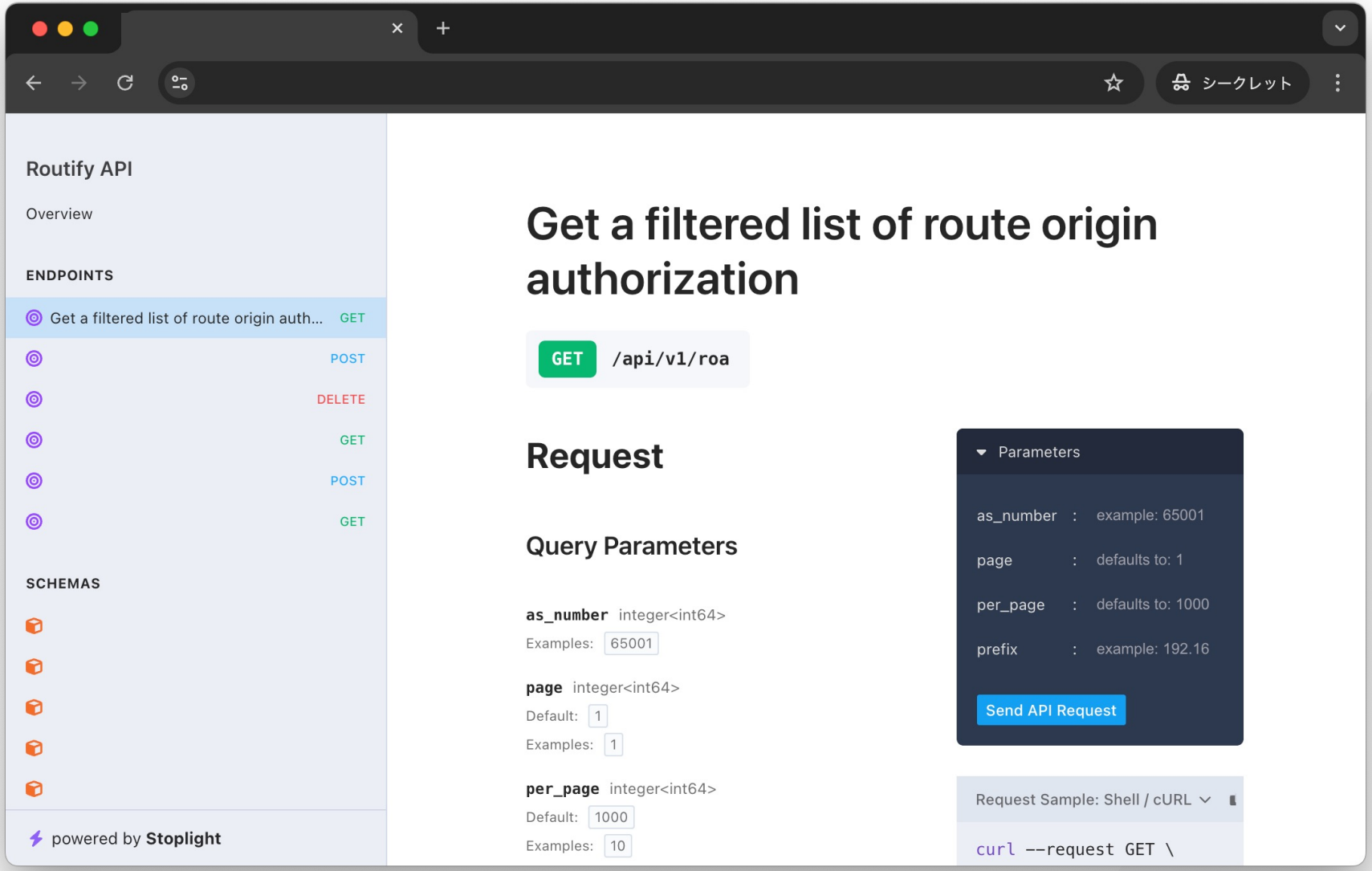


ID	AS Number	Prefix	Max Length
5	4207261338	100.113.12.154/32	32
8	4207261338	100.112.160.149/32	32
11	4207261338	100.112.160.26/32	32
14	4207261338	100.112.161.5/32	32
17	4207261338	100.112.160.152/32	32
20	4207261338	100.112.161.66/32	32
23	4207261337	100.113.12.153/32	32
26	4207261337	100.112.161.62/32	32
29	4207261337	100.112.160.201/32	32
32	4207261337	100.112.161.64/32	32
35	4207261337	100.112.161.46/32	32

© LY Corporation

RPKI in DC

Routify



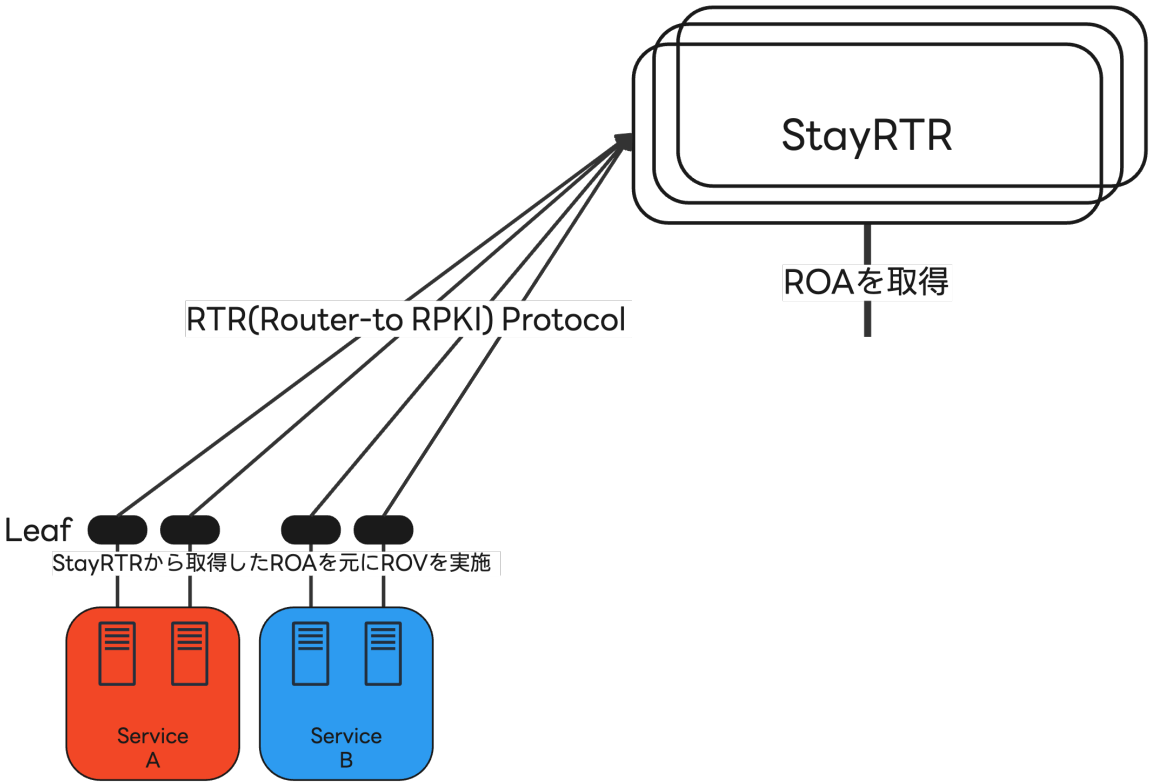
© LY Corporation

```
% routifyctl roas -h
ROAを管理するためのコマンド

Usage:
  routifyctl roas [AS OR PREFIX OR IPADDRESS]... [flags]
  routifyctl roas [command]

Aliases:
  roas, r, ro, roa

Available Commands:
  add         ROAを追加
  delete      roaのDELETEフラグをONにする
  update      ROAのMaxLengthを更新

Flags:
  -h, --help          help for roas
  -l, --length int    Page Size (default 1000)
  -p, --page int      Page Number (default 1)

Global Flags:
  -c, --config string    設定ファイルのパスを指定。指定しない場合以下の順序で探します。[./routify.yml,
/etc/routify/routify.yml]

Use "routifyctl roas [command] --help" for more information about a command.
```

© LY Corporation

RPKI in DC

RPKI-RTR



StayRTR

GoRTRをベースにしたOSSのRPKI-RTR実装

複数台のVMにデプロイ、L3DSRでVIPを構成

LeafはRPKIキャッシュとしてVIPを参照

ROAの情報はjsonフォーマットのRPKIデータをRoutifyから取得している

機能がシンプルで馴染みある言語で実装されていることが決め手となり採用

https://github.com/bgp/stayrtr

© LY Corporation

RPKI in DC

Validationの方針

```
route-map rpki-rov permit 10
   description RPKI Valid Prefix
   match origin-as validity valid
   set local-preference 100
!
route-map rpki-rov permit 20
   description RPKI Unknown Prefix
   match origin-as validity not-found
   set local-preference 80
!
route-map rpki-rov permit 30
   description RPKI Invalid Prefix
   match origin-as validity invalid
   set local-preference 1
!
```

設定イメージ

- いきなりDropはしない
  - 最終的にはinvalidな経路はDropしたい
- LPの制御から始める
- 既に動いているネットワークに安全に影響を抑えながら導入していきたい
- 下流から受信した経路のみROVする
- 上流から受信したデフォルトルートへのROVは行わない
  - 事故防止

© LY Corporation

RPKI in DC

動作確認できたNOSの一覧

	RPKI-RTRプロトコル	RFC8097
EOS 4.30.3M	✅	✅
Cumulus Linux 5.5.1	❌	❌
NX-OS 10.3(4a)	✅	🔄 *1
ArcOS 7.1.1.Alpha1	✅	❌

今後はこれらの実装に対応しているNOSを優先的に採用する方針で検討中

*1
RFC8097で定義されたExtended CommunityをeBGPで送受信できない
Cisco社にご協力頂き当該機能については対応実装中のステータス

© LY Corporation

RPKI in DC

その他

- SLURM (Simplified Local Internet Number Resource Management with the RPKI)
  - 今現在は利用していない
  - 運用が複雑化するので今後も利用することは避けたい
- Anycast経路の取り扱い
  - L4LBのVIPなど、異なるOriginから同一経路が広報されるケース
  - 特定のOriginだけが先にValidになると吸われてしまう
  - 今回作成したシステム上で複数のOriginから同一経路が広報されても問題ないことを確認

© LY Corporation

RPKI in DC

その他

- ASPA (Autonomous System Provider Authorization)
  - 今回のシステムではサーバがOriginのAS-PATHが1個の経路をLeafで検証する
  - 検証可能なPATHの長さがない
- 防げないケース
  - Origin ASごと不正経路を広報されるようなパターン
  - 物理サーバにAS番号自体を間違って設定

© LY Corporation

デモ

© LY Corporation

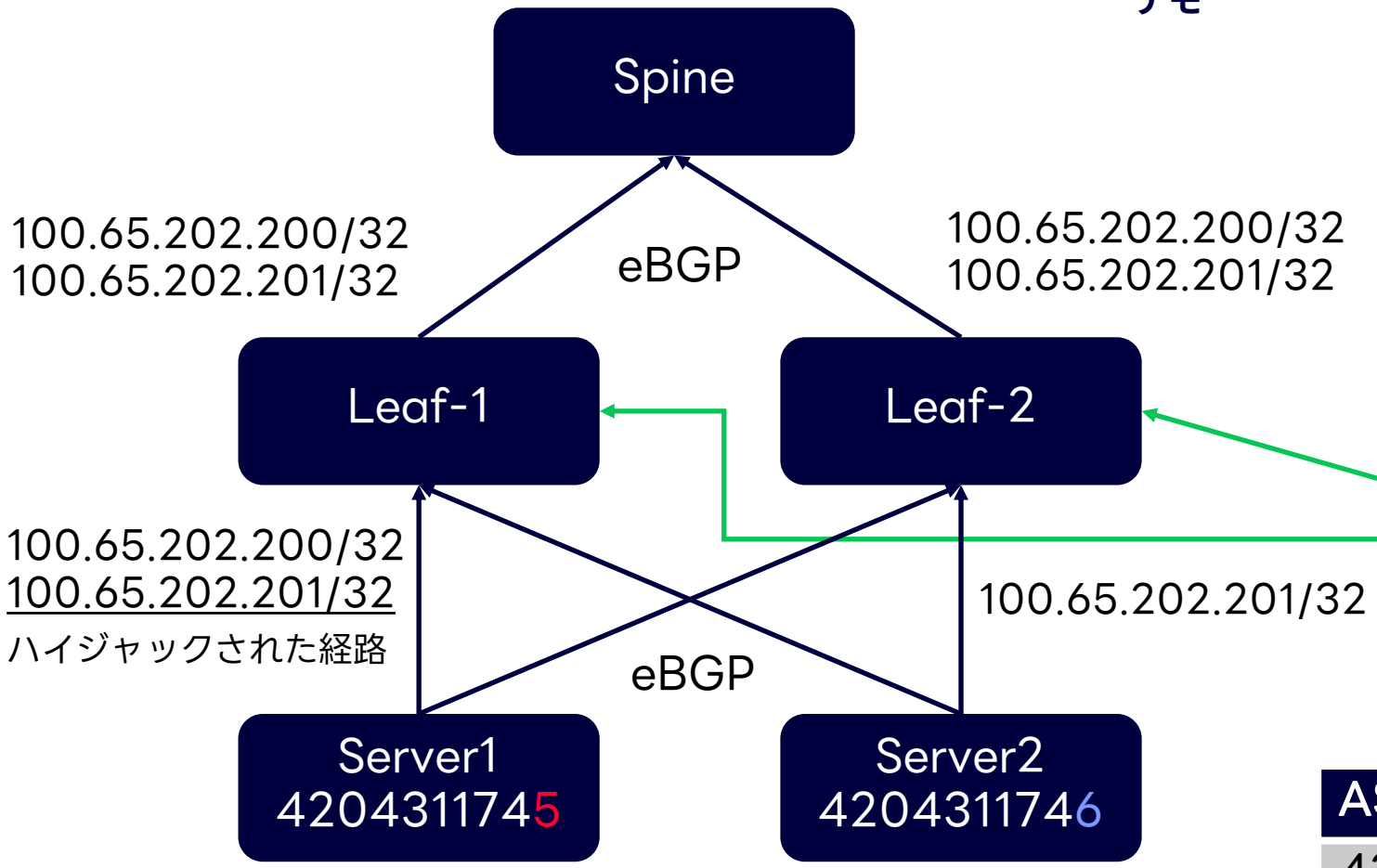
RPKI in DC

デモ

ステップ1 Server1が正しい経路を広報

ステップ2 Server1が誤った経路を広報

ステップ3 Server2が正しい経路を広報



ASN	Prefix
4204311745	100.65.202.200/32
4204311746	100.65.202.201/32

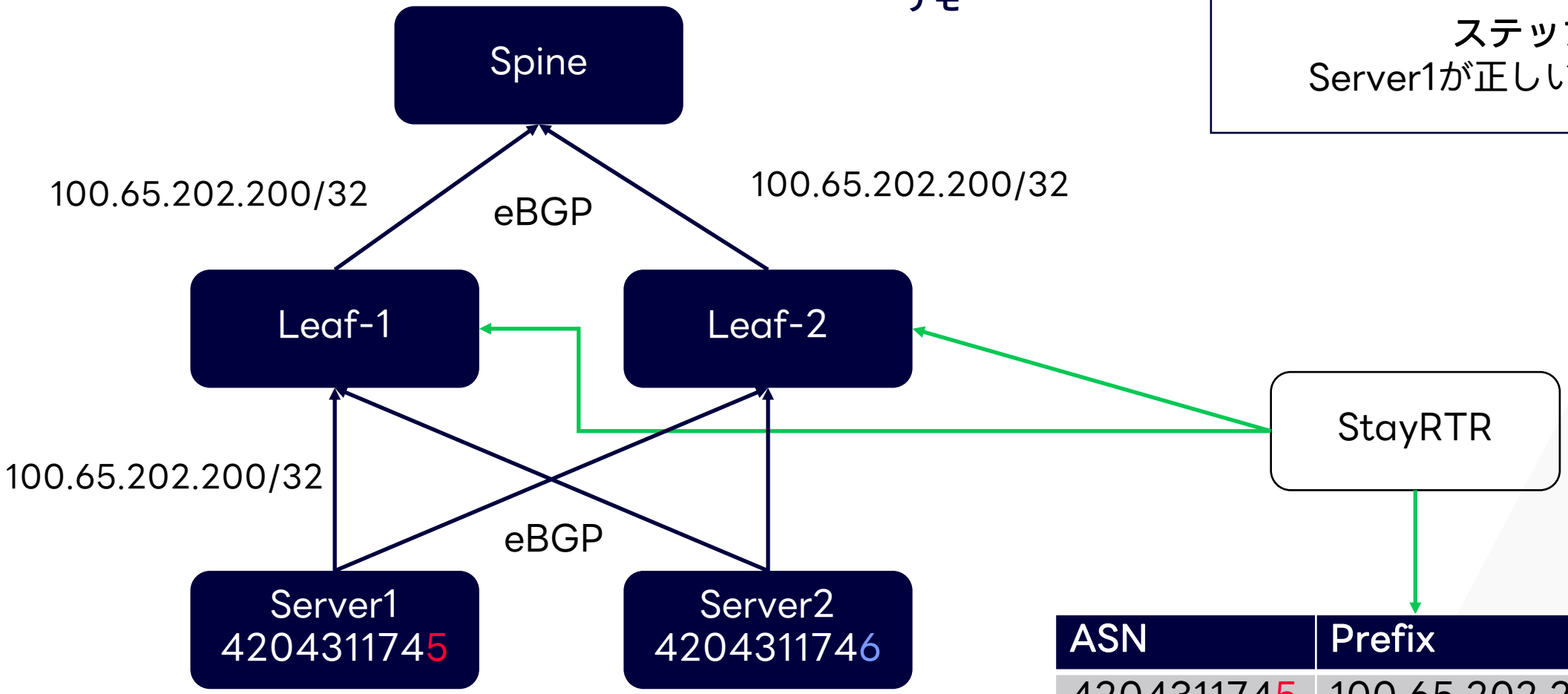
Routify

© LY Corporation

RPKI in DC

デモ

ステップ1
Server1が正しい経路を広報



ASN	Prefix
4204311745	100.65.202.200/32
4204311746	100.65.202.201/32

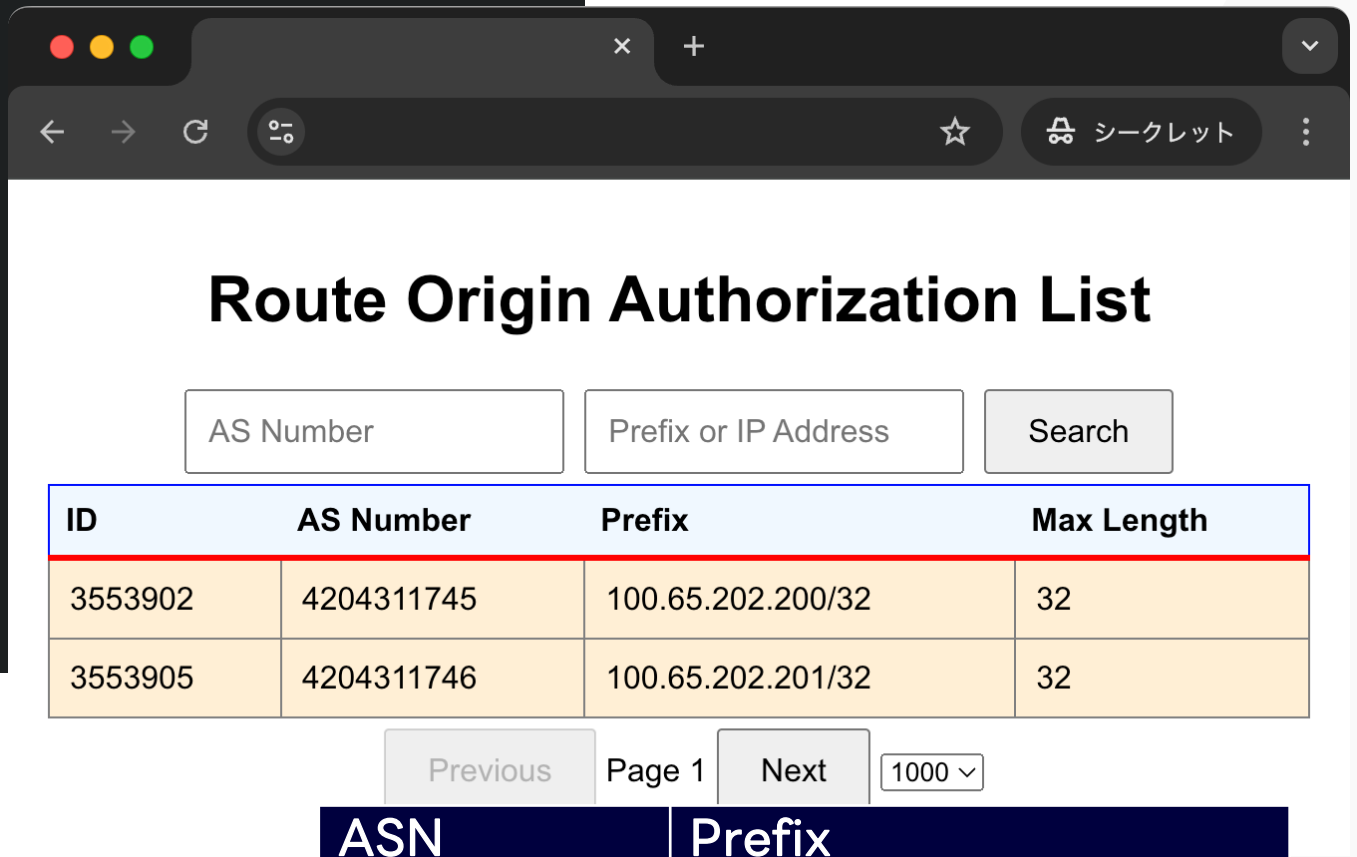
Routify

© LY Corporation

ステップ1
Server1が正しい経路を広報

```
[kanasaka@stayrtr03]~% curl -s http://                                  /api/v1/stayrtr | jq .
{
  "$schema":
  "metadata": {
    "generated":                ,
    "generated_time":
  },
  "roas": [
    {
      "asn": 4204311745,
      "prefix": "100.65.202.200/32",
      "maxLength": 32
    },
    {
      "asn": 4204311746,
      "prefix": "100.65.202.201/32",
      "maxLength": 32
    }
  ]
}
```

StayRTRはRoutifyのエンドポイントからROAを取得する



ASN	Prefix
4204311745	100.65.202.200/32
4204311746	100.65.202.201/32

© LY Corporation

ステップ1
Server1が正しい経路を広報

Leaf-1

```
#sh bgp rpki roa ipv4
Synced Caches: stayrtr03
Unsynced Caches:

                  Max
Prefix            Len  ASN         Caches
----------------- ---- ----------- ----------------------------------------------
100.65.202.200/32  32   4204311745  stayrtr03
100.65.202.201/32  32   4204311746  stayrtr03
```

Leaf-1

```
#sh ip bgp 100.65.202.200/32
BGP routing table information for VRF default
Router identifier 10.147.48.7, local AS number 4243200519
BGP routing table entry for 100.65.202.200/32
 Paths: 1 available
 4204311745
   fe80::8581:cc86:de50:33f3%Vl1001 from fe80::8581:cc86:de50:33f3%Vl1001 (100.65.202.193)
     Origin INCOMPLETE, metric 0, localpref 100, IGP metric 1, weight 0, tag 0
     Received 03:13:42 ago, valid, external, best, AS Origin valid
     Rx SAFI: Unicast
```

Leaf-2

```
#sh bgp rpki roa ipv4
Synced Caches: stayrtr03
Unsynced Caches:

                  Max
Prefix            Len  ASN         Caches
----------------- ---- ----------- ----------------------------------------------
100.65.202.200/32  32   4204311745  stayrtr03
100.65.202.201/32  32   4204311746  stayrtr03
```

Leaf-2

```
#sh ip bgp 100.65.202.200/32
BGP routing table information for VRF default
Router identifier 10.147.48.9, local AS number 4243200521
BGP routing table entry for 100.65.202.200/32
 Paths: 1 available
 4204311745
   fe80::6880:3276:34e6:a01f%Vl1001 from fe80::6880:3276:34e6:a01f%Vl1001 (100.65.202.193)
     Origin INCOMPLETE, metric 0, localpref 100, IGP metric 1, weight 0, tag 0
     Received 03:12:23 ago, valid, external, best, AS Origin valid
     Rx SAFI: Unicast
```

ASN	Prefix
4204311745	100.65.202.200/32
4204311746	100.65.202.201/32

© LY Corporation

ステップ1
Server1が正しい経路を広報

RPKI in DC

デモ

Spine

```
sh ip bgp 100.65.202.200/32
BGP routing table information for VRF default
Router identifier 10.147.48.22, local AS number 4243200534
BGP routing table entry for 100.65.202.200/32
 Paths: 2 available
  4243200521 4204311745
    fe80::626b:5bff:fe9c:fad5%Et2/1 from fe80::626b:5bff:fe9c:fad5%Et2/1 (10.147.48.9)
      Origin INCOMPLETE, metric 0, localpref 100, IGP metric 1, weight 0, tag 0
      Received 18:43:38 ago, valid, external, ECMP head, ECMP, best, ECMP contributor, AS Origin valid
      Extended Community: Origin-Validation-State:valid
      Rx SAFI: Unicast
  4243200519 4204311745
    fe80::626b:5bff:fe9d:df37%Et1/1 from fe80::626b:5bff:fe9d:df37%Et1/1 (10.147.48.7)
      Origin INCOMPLETE, metric 0, localpref 100, IGP metric 1, weight 0, tag 0
      Received 18:43:38 ago, valid, external, ECMP, ECMP contributor, AS Origin valid
      Extended Community: Origin-Validation-State:valid
      Rx SAFI: Unicast
```

ASN	Prefix
4204311745	100.65.202.200/32
4204311746	100.65.202.201/32

© LY Corporation

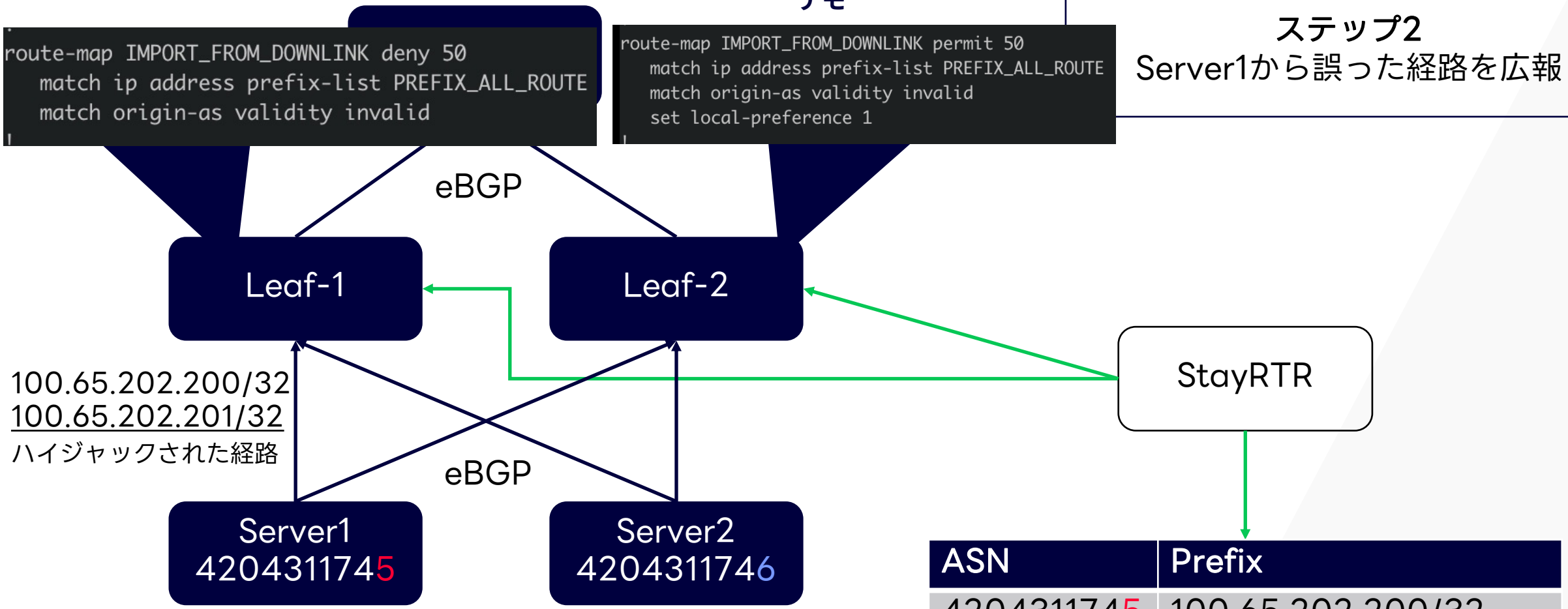
RPKI in DC

デモ

```
route-map IMPORT_FROM_DOWNLINK deny 50
   match ip address prefix-list PREFIX_ALL_ROUTE
   match origin-as validity invalid
!
```

```
route-map IMPORT_FROM_DOWNLINK permit 50
   match ip address prefix-list PREFIX_ALL_ROUTE
   match origin-as validity invalid
   set local-preference 1
!
```

ステップ2
Server1から誤った経路を広報



ASN	Prefix
4204311745	100.65.202.200/32
4204311746	100.65.202.201/32

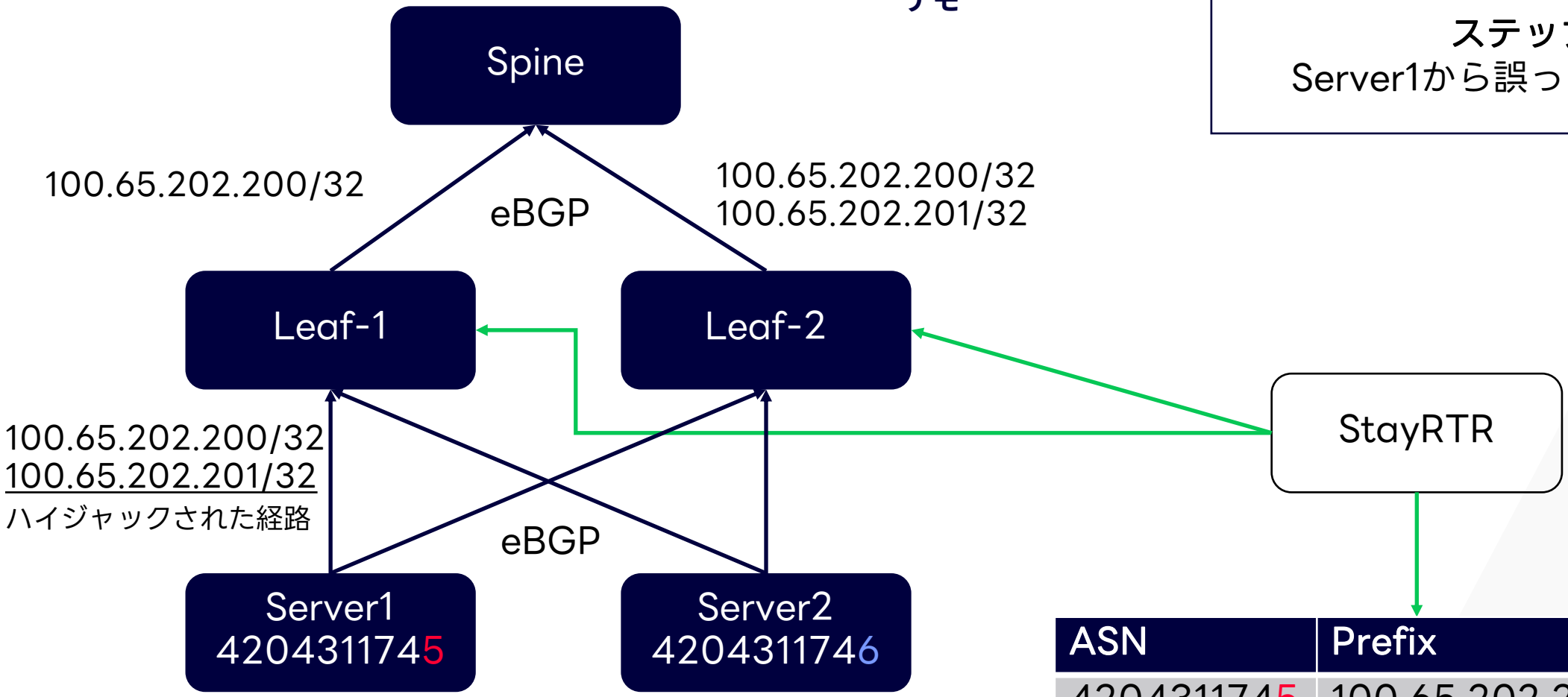
Routify

© LY Corporation

RPKI in DC

デモ

ステップ2
Server1から誤った経路を広報



ASN	Prefix
4204311745	100.65.202.200/32
4204311746	100.65.202.201/32

Routify

© LY Corporation

ステップ2
Server1から誤った経路を広報

Leaf-1

```
sh ip bgp 100.65.202.201/32
BGP routing table information for VRF default
Router identifier 10.147.48.7, local AS number 4243200519
```

Leaf-1

```
sh ip route bgp

VRF: default
Codes: C - connected, S - static, K - kernel,
       O - OSPF, IA - OSPF inter area, E1 - OSPF external type 1,
       E2 - OSPF external type 2, N1 - OSPF NSSA external type 1,
       N2 - OSPF NSSA external type2, B - Other BGP Routes,
       B I - iBGP, B E - eBGP, R - RIP, I L1 - IS-IS level 1,
       I L2 - IS-IS level 2, O3 - OSPFv3, A B - BGP Aggregate,
       A O - OSPF Summary, NG - Nexthop Group Static Route,
       V - VXLAN Control Service, M - Martian,
       DH - DHCP client installed default route,
       DP - Dynamic Policy Route, L - VRF Leaked,
       G  - gRIBI, RC - Route Cache Route,
       CL - CBF Leaked Route

 B E      100.65.202.200/32 [20/0]
           via fe80::8581:cc86:de50:33f3, Vlan1001
```

Leaf-2

```
>sh ip bgp 100.65.202.201/32
BGP routing table information for VRF default
Router identifier 10.147.48.9, local AS number 4243200521
BGP routing table entry for 100.65.202.201/32
 Paths: 1 available
 4204311745
    fe80::6880:3276:34e6:a01f%Vl1001 from fe80::6880:3276:34e6:a01f%Vl1001 (100.65.202.193)
      Origin INCOMPLETE, metric 0, localpref 1, IGP metric 1, weight 0, tag 0
      Received 18:00:40 ago, valid, external, best, AS Origin invalid
      Rx SAFI: Unicast
```

Leaf-2

```
>sh ip route bgp

VRF: default
Codes: C - connected, S - static, K - kernel,
       O - OSPF, IA - OSPF inter area, E1 - OSPF external type 1,
       E2 - OSPF external type 2, N1 - OSPF NSSA external type 1,
       N2 - OSPF NSSA external type2, B - Other BGP Routes,
       B I - iBGP, B E - eBGP, R - RIP, I L1 - IS-IS level 1,
       I L2 - IS-IS level 2, O3 - OSPFv3, A B - BGP Aggregate,
       A O - OSPF Summary, NG - Nexthop Group Static Route,
       V - VXLAN Control Service, M - Martian,
       DH - DHCP client installed default route,
       DP - Dynamic Policy Route, L - VRF Leaked,
       G  - gRIBI, RC - Route Cache Route,
       CL - CBF Leaked Route

 B E      100.65.202.200/32 [20/0]
           via fe80::6880:3276:34e6:a01f, Vlan1001
 B E      100.65.202.201/32 [20/0]
           via fe80::6880:3276:34e6:a01f, Vlan1001
```

ASN	Prefix
4204311745	100.65.202.200/32
4204311746	100.65.202.201/32

© LY Corporation

ステップ2
Server1から誤った経路を広報

RPKI in DC

デモ

Spine

```
>sh ip bgp 100.65.202.201/32
BGP routing table information for VRF default
Router identifier 10.147.48.22, local AS number 4243200534
BGP routing table entry for 100.65.202.201/32
 Paths: 1 available
  4243200521 4204311745
    fe80::626b:5bff:fe9c:fad5%Et2/1 from fe80::626b:5bff:fe9c:fad5%Et2/1 (10.147.48.9)
      Origin INCOMPLETE, metric 0, localpref 1, IGP metric 1, weight 0, tag 0
      Received 00:08:35 ago, valid, external, best, AS Origin invalid
      Extended Community: Origin-Validation-State:invalid
      Rx SAFI: Unicast
```

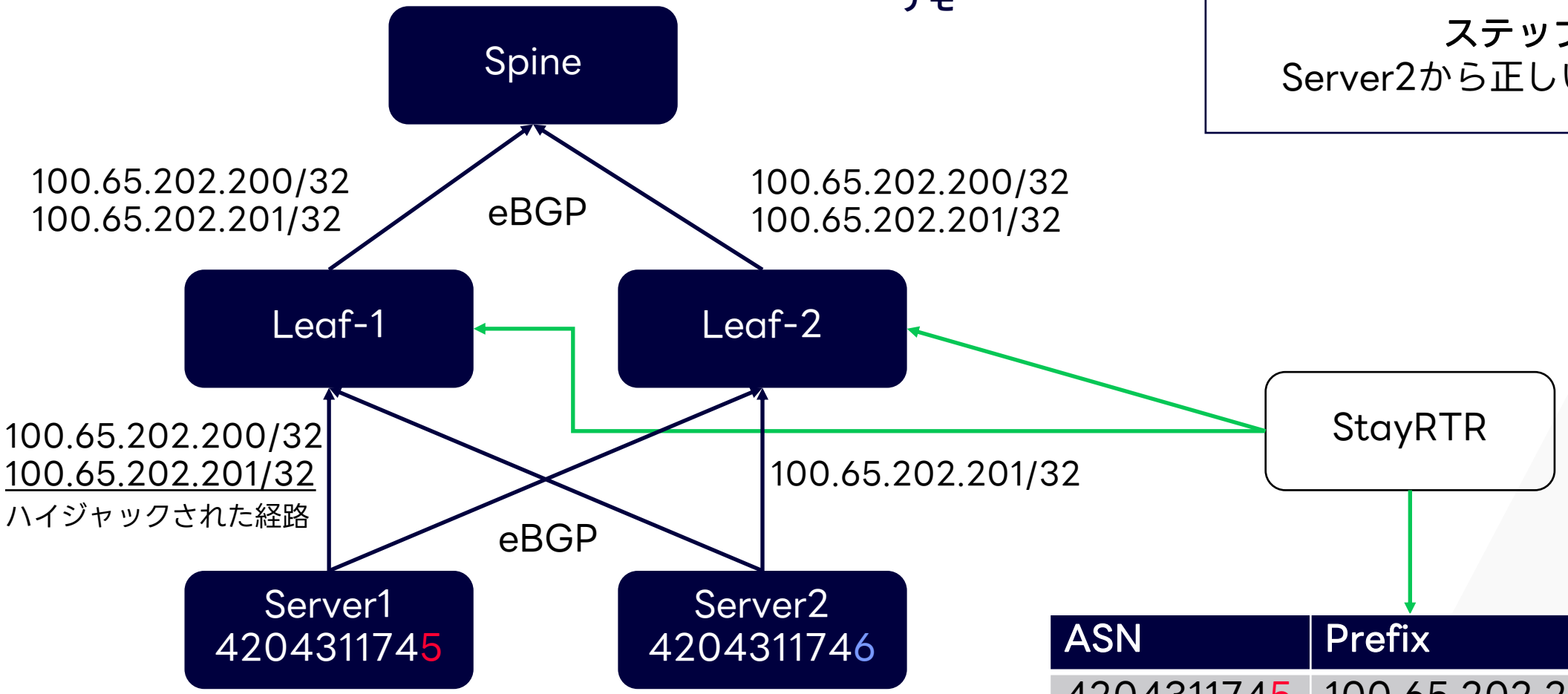
ASN	Prefix
4204311745	100.65.202.200/32
4204311746	100.65.202.201/32

© LY Corporation

RPKI in DC

デモ

ステップ3
Server2から正しい経路を広報



ASN	Prefix
4204311745	100.65.202.200/32
4204311746	100.65.202.201/32

Routify

© LY Corporation

ステップ3
Server2から正しい経路を広報

Leaf-1

```
sh ip bgp 100.65.202.201/32
BGP routing table information for VRF default
Router identifier 10.147.48.7, local AS number 4243200519
BGP routing table entry for 100.65.202.201/32
 Paths: 1 available
 4204311746
   fe80::37df:1e23:a553:3af9%Vl1002 from fe80::37df:1e23:a553:3af9%Vl1002 (100.65.202.194)
     Origin INCOMPLETE, metric 0, localpref 100, IGP metric 1, weight 0, tag 0
     Received 00:00:17 ago, valid, external, best, AS Origin valid
     Rx SAFI: Unicast
```

Leaf-1

```
sh ip route bgp

VRF: default
Codes: C - connected, S - static, K - kernel,
       O - OSPF, IA - OSPF inter area, E1 - OSPF external type 1,
       E2 - OSPF external type 2, N1 - OSPF NSSA external type 1,
       N2 - OSPF NSSA external type2, B - Other BGP Routes,
       B I - iBGP, B E - eBGP, R - RIP, I L1 - IS-IS level 1,
       I L2 - IS-IS level 2, O3 - OSPFv3, A B - BGP Aggregate,
       A O - OSPF Summary, NG - Nexthop Group Static Route,
       V - VXLAN Control Service, M - Martian,
       DH - DHCP client installed default route,
       DP - Dynamic Policy Route, L - VRF Leaked,
       G  - gRIBI, RC - Route Cache Route,
       CL - CBF Leaked Route

 B E      100.65.202.200/32 [20/0]
          via fe80::8581:cc86:de50:33f3, Vlan1001
 B E      100.65.202.201/32 [20/0]
          via fe80::37df:1e23:a553:3af9, Vlan1002
```

Leaf-2

```
sh ip bgp 100.65.202.201/32
BGP routing table information for VRF default
Router identifier 10.147.48.9, local AS number 4243200521
BGP routing table entry for 100.65.202.201/32
 Paths: 2 available
 4204311746
   fe80::1eb4:9a38:ad94:e256%Vl1002 from fe80::1eb4:9a38:ad94:e256%Vl1002 (100.65.202.194)
     Origin INCOMPLETE, metric 0, localpref 100, IGP metric 1, weight 0, tag 0
     Received 00:01:13 ago, valid, external, best, AS Origin valid
     Rx SAFI: Unicast
 4204311745
   fe80::6880:3276:34e6:a01f%Vl1001 from fe80::6880:3276:34e6:a01f%Vl1001 (100.65.202.193)
     Origin INCOMPLETE, metric 0, localpref 1, IGP metric 1, weight 0, tag 0
     Received 18:12:13 ago, valid, external, AS Origin invalid
```

Leaf-2

```
sh ip route bgp

VRF: default
Codes: C - connected, S - static, K - kernel,
       O - OSPF, IA - OSPF inter area, E1 - OSPF external type 1,
       E2 - OSPF external type 2, N1 - OSPF NSSA external type 1,
       N2 - OSPF NSSA external type2, B - Other BGP Routes,
       B I - iBGP, B E - eBGP, R - RIP, I L1 - IS-IS level 1,
       I L2 - IS-IS level 2, O3 - OSPFv3, A B - BGP Aggregate,
       A O - OSPF Summary, NG - Nexthop Group Static Route,
       V - VXLAN Control Service, M - Martian,
       DH - DHCP client installed default route,
       DP - Dynamic Policy Route, L - VRF Leaked,
       G  - gRIBI, RC - Route Cache Route,
       CL - CBF Leaked Route

 B E      100.65.202.200/32 [20/0]
          via fe80::6880:3276:34e6:a01f, Vlan1001
 B E      100.65.202.201/32 [20/0]
          via fe80::1eb4:9a38:ad94:e256, Vlan1002
```

ASN	Prefix
4204311745	100.65.202.200/32
4204311746	100.65.202.201/32

© LY Corporation

ステップ3
Server2から正しい経路を広報

RPKI in DC

デモ

Spine

```
>sh ip bgp 100.65.202.201/32
BGP routing table information for VRF default
Router identifier 10.147.48.22, local AS number 4243200534
BGP routing table entry for 100.65.202.201/32
 Paths: 2 available
  4243200521 4204311746
    fe80::626b:5bff:fe9c:fad5%Et2/1 from fe80::626b:5bff:fe9c:fad5%Et2/1 (10.147.48.9)
      Origin INCOMPLETE, metric 0, localpref 100, IGP metric 1, weight 0, tag 0
      Received 00:28:38 ago, valid, external, ECMP head, ECMP, best, ECMP contributor, AS Origin valid
      Extended Community: Origin-Validation-State:valid
      Rx SAFI: Unicast
  4243200519 4204311746
    fe80::626b:5bff:fe9d:df37%Et1/1 from fe80::626b:5bff:fe9d:df37%Et1/1 (10.147.48.7)
      Origin INCOMPLETE, metric 0, localpref 100, IGP metric 1, weight 0, tag 0
      Received 00:28:38 ago, valid, external, ECMP, ECMP contributor, AS Origin valid
      Extended Community: Origin-Validation-State:valid
      Rx SAFI: Unicast
```

ASN	Prefix
4204311745	100.65.202.200/32
4204311746	100.65.202.201/32

© LY Corporation

構築したシステムで実現・確認できたこと

- 経路ハイジャックの防止
  - 登録したROAを元にしたROVを行うことで、経路ハイジャックを防止
- ROVの結果に基づいた柔軟な制御
  - Dropに加えてLPを使った優先度制御が可能であることを確認
- 複数の手段でのROAの登録、編集、削除
  - CLIに加えてHTTP API経由で即時に反映できるシステムを実装
  - これにより、周辺のシステムと連携した動作が可能
- Extended Communityに基づいたROV
- プライベートIP、4byteASを使った環境におけるRPKI
  - データセンターネットワークでも問題なく動作

© LY Corporation

システム構築にあたって苦労した点

- データセンターでRPKIを利用する観点で情報が少ない
  - インターネットでの利用が前提となっている
- データセンターネットワーク内で利用しているNOSが要件を満たしていないことがある
  - 需要はある
  - 今後の対応に期待

© LY Corporation

まとめ・議論

© LY Corporation

RPKI in DCによる変化

旧各社で分かれている経路制御ポリシー

会社の統合でネットワークの経路制御ポリシーを合わせたい

検証とシステム実装を通して
RPKIを使った送信元検証
の導入に一定の目処

Clos NWで採用する機種の多様化

新しいNOSを採用するたびにオーケストレータの改修が必要

標準化された技術である
RPKIを採用することで
送信元検証に関する
オーケストレータの改修が不要に

動的に変更される設定

テンプレート化できない

反映までにラグがある

RPKIの導入で差分が解消
均一かつ恒久化された設定を実現

© LY Corporation

手段としてのRPKI

- 必ずしも「RPKI」である必要はない
- 信頼できるASとIPアドレスの組み合わせをスイッチが参照してROVする実装であればなんでもいい
  - スイッチ個別に異なるフィルタ設定がconfigとして入らないことが重要
- 標準化された実績のある仕組みに乗ることを考えるとRPKIになる

© LY Corporation

何を信頼するのか

- 経路のOriginの正当性を証明するレジストリは存在しない
- 正しさを保証する第三者が存在しない
- 自分たちが管理するシステムを自分たちで信頼している
- ROAとして登録したデータが誤ったものだと全てが崩れる
- 信頼性をどう担保していくか？
  - 誰がどう情報を管理できるようにするべきか

© LY Corporation

RPKIについて感じていること

- 課題の解決策として仕組みと実装が重いと感じた
- インターネットRPKIは更に複雑
  - なかなか普及しない一因なのでは

© LY Corporation

今後の取り組み

- RPKI in DCの普及拡大
  - 商用環境への適用をより拡大していく
- 監視の実装開発
  - ハイジャックの検知と通知
  - Validation Stateのメトリクス化

© LY Corporation

まとめ

本発表では、LINEヤフーのデータセンターネットワークの構成と直面している課題、その解決策として注目した技術であるRPKIの適用事例を紹介しました。

RPKIはインターネットの経路送信元検証のために生まれた技術です。

そのため、データセンターでの利用を前提とした時に機器やソフトウェアの対応状況は充実しているとは言い難い状況ではありますが、今回の取り組みを通してデータセンターネットワークでも十分活用可能な状態にあると感じました。

意図しない経路の広報と受信の適切な制御という普遍的な課題について、皆さんがどう向き合っているのか議論できれば幸いです。

© LY Corporation

議論ポイント

- BGP in DCでの経路フィルタの必要性
- 何を信頼するのか – 送信の元となるデータ
- RPKI導入の障壁
- RPKIの実装差異や運用で困ったこと
- 代替になりそうな技術について
- RPKIへの期待と課題

© LY Corporation

LINEヤフー

© LY Corporation