IPv6 トラブルシューティング ~ ISP編 ~

Matsuzaki ‘maz’ Yoshinobu

<maz@iij.ad.jp>

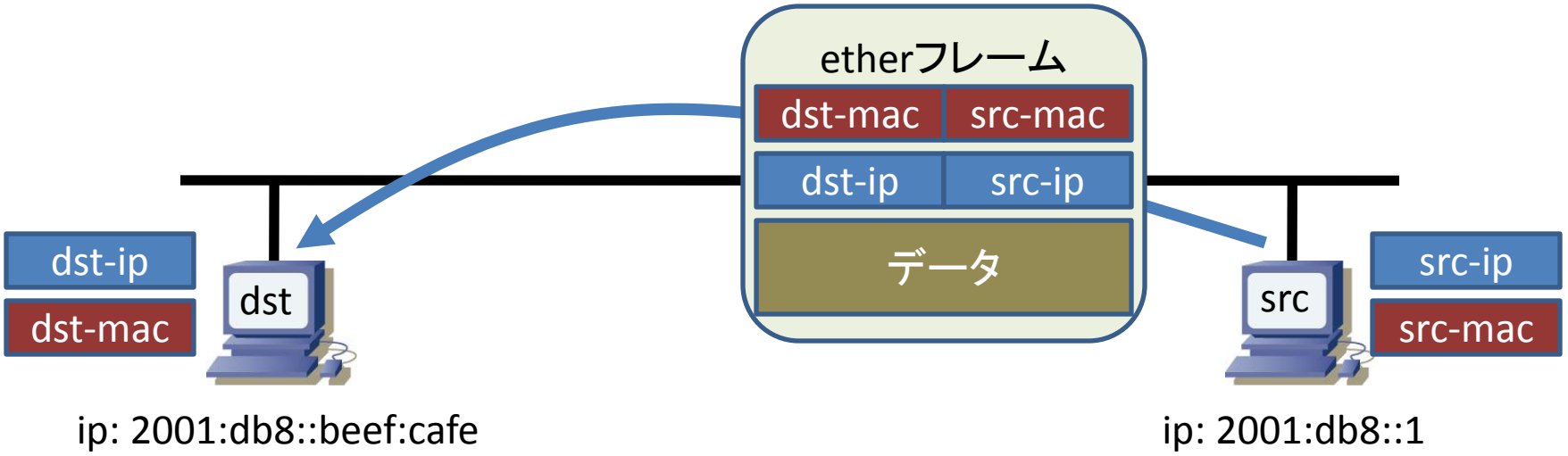
内容

- 主にISPのネットワークで起こりそうなトラブルと、その対策を紹介する
  - 疎通
  - 経路制御
- 正しい状態を知る
- トラブル事例を知る

IPv6パケット送信

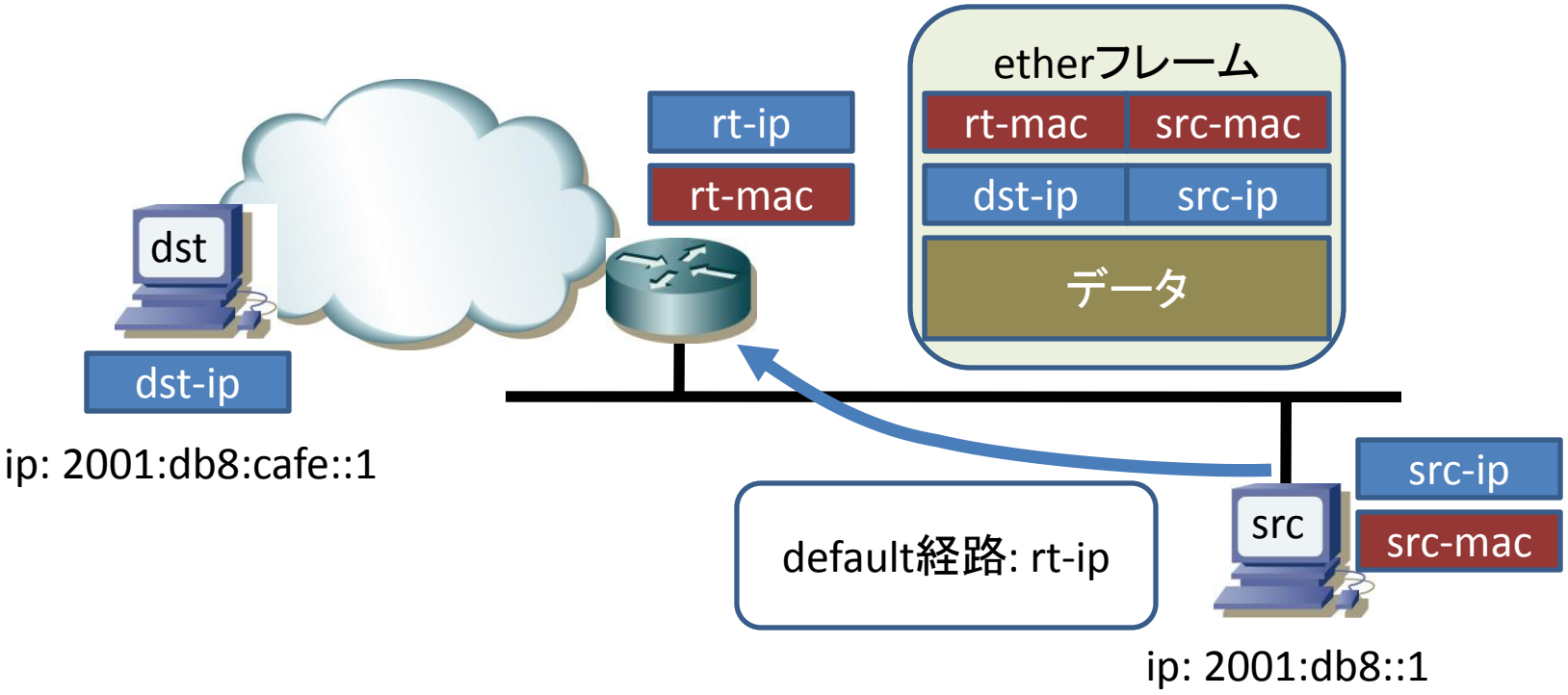
- on-link情報のあるネットワークには直接送信

inet6 2001:db8::1 prefixlen 64
↓
2001:db8::～2001:db8::ffff:ffff:ffff:ffffが同じセグメント上にある



IPv6パケット送信 2

- 遠くには経路情報に従ってルータに投げる



ndp (Neighbor Discovery Protocol)

- etherではパケット送信にMACアドレスが必要
  - 機器のIPv6アドレスからMACアドレスを知りたい
- ndpで解決
  - RFC4861, RFC5942
    - ICMPv6を利用してMACアドレスを問い合わせる
  - 送り先を未学習ならmulticastアドレス宛て
    - IP: ff02::1:ff00:0000 ～ ff02::1:ffff:ffff
      - 送信先IPアドレスの下位24bitを利用して生成
    - MAC: 33:33:00:00:00:00 ～ 33:33:ff:ff:ff:ff
      - 送信先IPアドレスの下位32bitを利用して生成

ndpでMACアドレス解決

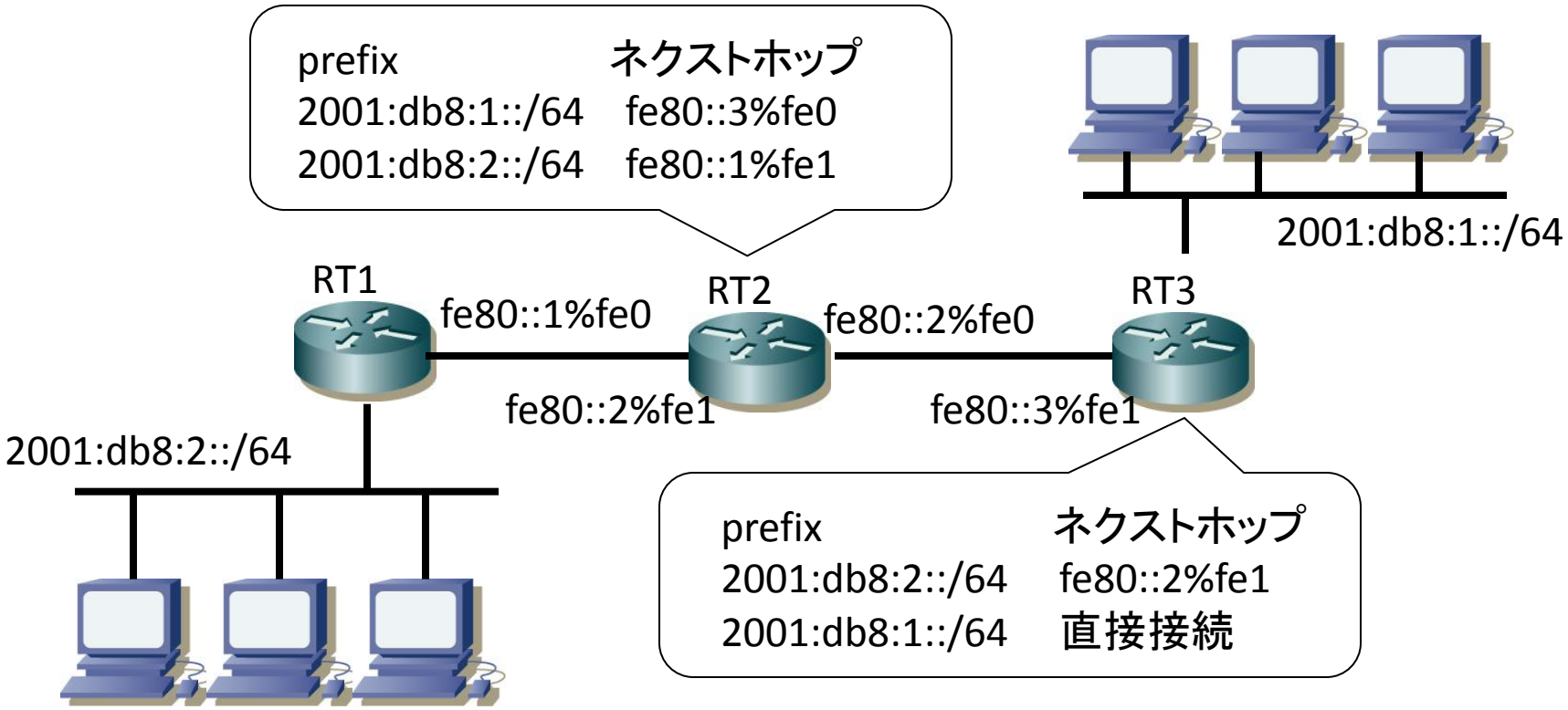
```
IP6 2001:db8::1 > ff02::1:ffef:cafe
 ICMP6, neighbor solicitation, who has 2001:db8::beef:cafe
 source link-address option: 00:19:bb:27:37:e0
        0x0000:  3333 ffef cafe 0019 bb27 37e0 86dd 6000
        0x0010:  0000 0020 3aff 2001 0db8 0000 0000 0000
        0x0020:  0000 0000 0001 ff02 0000 0000 0000 0000
        0x0030:  0001 ffef cafe 8700 9a90 0000 0000 2001
        0x0040:  0db8 0000 0000 0000 0000 beef cafe 0101
        0x0050:  0019 bb27 37e0
IP6 2001:db8::beef:cafe > 2001:db8::1
 ICMP6, neighbor advertisement, tgt is 2001:db8::beef:cafe
 destination link-address option: 00:16:17:61:64:86
        0x0000:  0019 bb27 37e0 0016 1761 6486 86dd 6000
        0x0010:  0000 0020 3aff 2001 0db8 0000 0000 0000
        0x0020:  0000 beef cafe 2001 0db8 0000 0000 0000
        0x0030:  0000 0000 0001 8800 c1fd 6000 0000 2001
        0x0040:  0db8 0000 0000 0000 0000 beef cafe 0201
        0x0050:  0016 1761 6486
```

IPv6ルーティング

- 基本的な考え方はIPv4と同じ
  - 経路情報＝宛先prefix+ネクストホップ
    - ネクストホップはリンクローカルアドレスが基本
  - 最長一致のルール
    - 細かい経路が優先される
    - 2001:db8::/32 と 2001:db8::/64の経路情報があると、2001:db8::1 宛ての経路はより細かい /64側の経路を選択する

経路情報

- 宛先prefixとネクストホップ



static route

- 宛先prefixとネクストホップを設定
- ネクストホップ
  - インタフェース名/ネクストホップアドレス
    - ネクストホップアドレスはリンクローカルが基本だが、グローバルを書いてもちゃんと動く

cisco ios

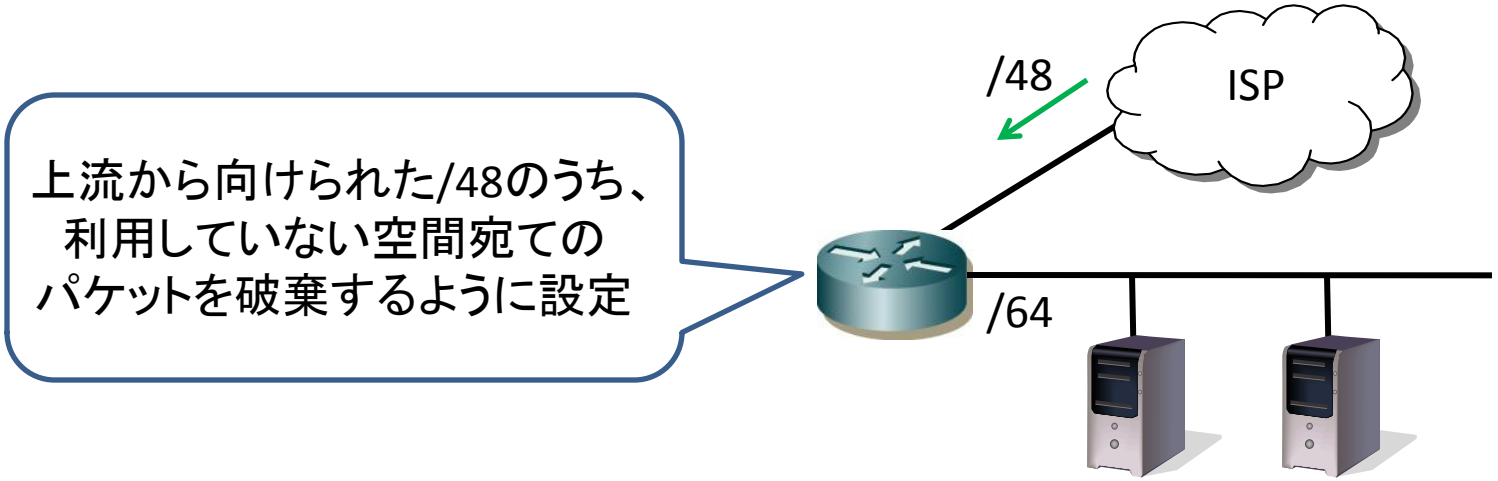
```
ipv6  route  2001:db8:110::/48  2001:db8:ff10::2
```

juniper junos

```
routing-options { rib inet6.0 { static {
2001:db8:110::/48 {
next-hop 2001:db8:ff10::2;}}}}
```

static routeの使い所

- PA (Provider Aggregatable)経路の生成
- single homeな顧客向けの経路設定
  - IPv6ではアドレス空間がいっぱいあるので、必ず顧客側でも利用していない空間宛てのパケットを破棄するルールを設定しておく



OSPFv3 (OSPF for IPv6)

- 基本的な考え方はOSPFv2と同じ
- IPv6用に一部拡張
  - しかしルータID等は32bit長のまま
  - トポロジ情報はIPv6非依存で構築できる
- リンク(インタフェース)単位で実行される
- LSAを整理
  - トポロジと経路の分離
  - IPアドレス+netmask → prefix+prefix長

OSPFv3関連 RFC

- [RFC2328] OSPFv2
  - 基本的な考え方はOSPFv2で理解しておく
- [RFC2740] OSPFv3
  - OSPFv2からの差分で記述されている部分が多い
- [RFC4552] Authentication/Confidentiality for OSPFv3
  - IPSECで守りましょう

OSPFv3で利用されるID

- ルータID
  - ルータを識別する32bit長のID
  - OSPFのAS内で一意であれば良い
- インタフェースID
  - ルータ内でインタフェースを識別する32bit長のID
  - MIB-II IfIndexとかでも良い
- インスタンスID
  - 同じセグメント上で異なるOSPFを同時に稼働させたい場合に利用するID
  - 通常は0を利用

OSPFv3の設定

- コスト設計はOSPFv2と同様
  - トポロジの違いは考慮しなければならないものの、OSPFv2と同様に運用できる
- 各インタフェース単位でOSPFv3が動く
  - OSPFv2では基本的にサブネット単位で制御
  - あるインタフェースでOSPFv3を有効にすると、設定されているIPv6 prefixは自動的に経路広告されてしまう
    - 不用意なアドレス設定に要注意

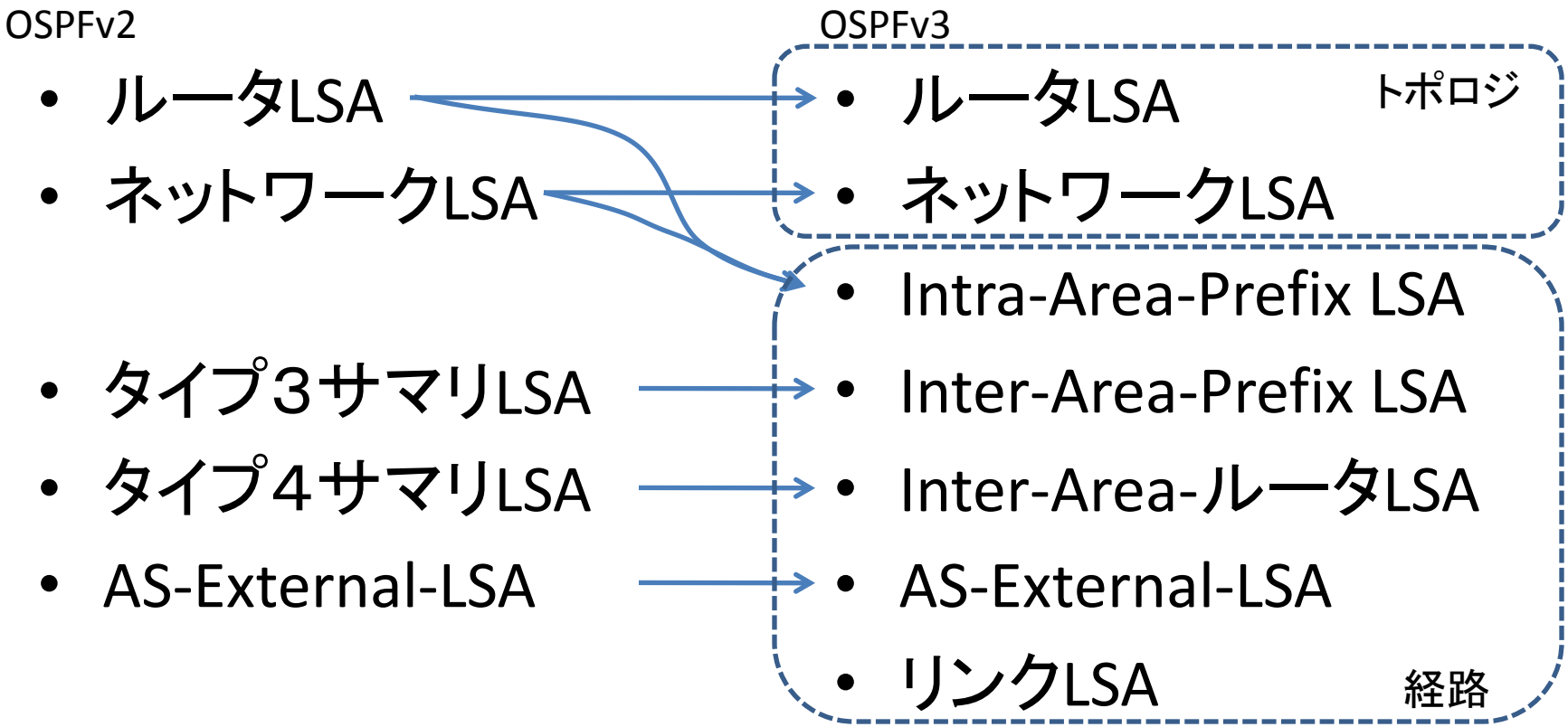
OSPFv3パケットの送信先

- ALLSPFRouter - [ff02::5] (OSPFv2 [224.0.0.5])
  - 全OSPFルータが受信する
- AllDRouters - [ff02::6] (OSPFv2 [224.0.0.6])
  - DRとBDRのみが受信する
  - point-to-point接続ではALLSPFRouter宛
  - ブロードキャストネットワークでは
    - HelloはALL SPF Router宛
    - DR,BDRからのLS update, LS AckはALLSPFRouter宛
    - DROtherからのLS update, LS AckはAllDRRouter宛
  - それ以外はネイバへのユニキャスト宛
    - virtualリンク以外はリンクローカルアドレスを利用

OSPFv3パケットの送信元

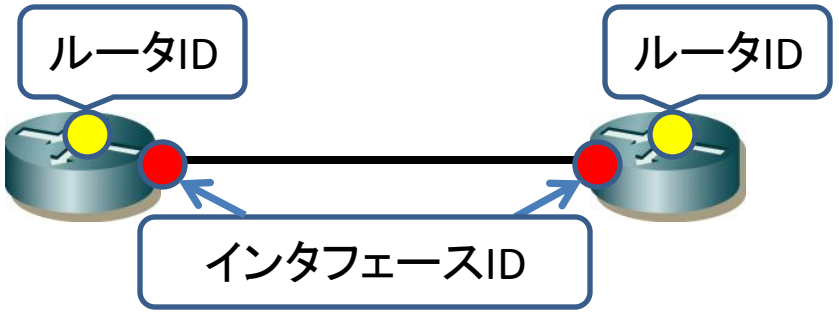
- 基本的にリンクローカルアドレスを利用
  - OSPFパケットを送出するインタフェースのアドレス
  - virtualリンクではグローバルアドレス
- つまりインタフェースにリンクローカルアドレスさえ付いていれば、OSPFv3は動く

LSAの変更

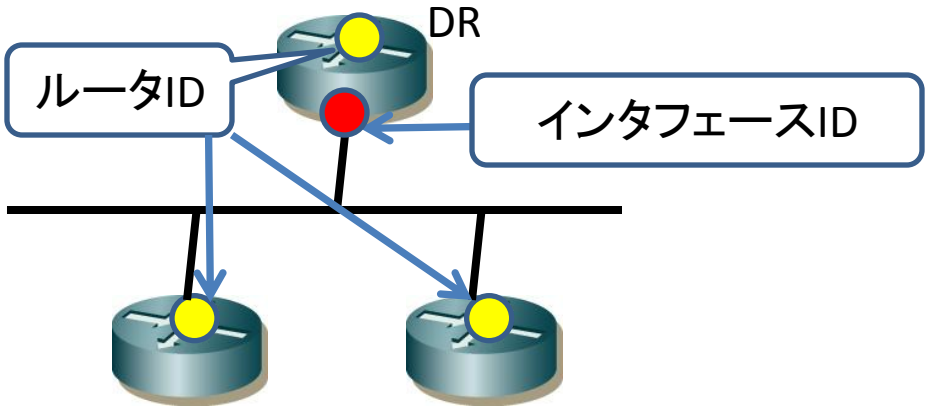


OSPFv3のルータLSAとネットワークLSA

- IPv6非依存でトポロジ情報だけを運ぶ
- ルータLSA
  - ルータの接続情報

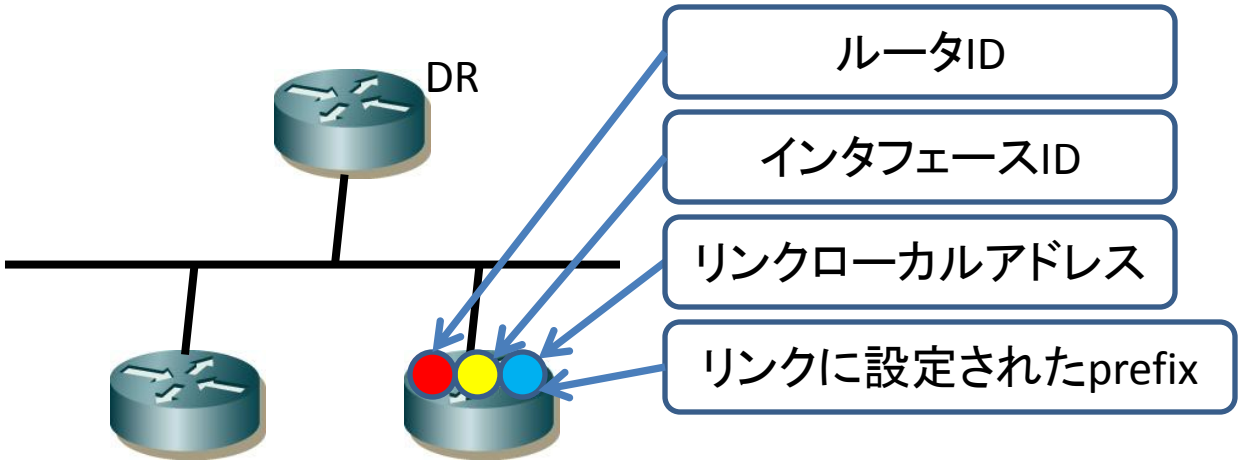


- ネットワークLSA
  - リンクに接続しているルータのリスト



リンクLSA

- リンク内のみで交換されるLSA
  - リンクローカルアドレスの通知
  - リンク上で有効なprefixの通知



Intra-Area-Prefix LSA

- OSPFv2の時にルータLSAやネットワークLSAが運んでいた経路情報を運ぶ
  - Stubネットワークやtransitネットワークの経路
  - loopbackの経路もこのLSAで運ばれる
- リンクLSAをDRが収集して、経路情報を代表してIntra-Area-Prefix LSAとして広報する
  - リンク上の一部ルータにだけ設定されているprefixでもリンクのDRが代表して広報する

構築される経路情報

- ネクストホップアドレスにリンクLSAで通知されたリンクローカルアドレスを採用
  - リンクローカルアドレスがとても重要
- LSA別の経路優先度はOSPFv2とほぼ同じ
  - エリア内経路
    - Intra-Area-Prefix LSA
  - エリア間経路
    - Inter-Area-Prefix LSA
  - 外部経路
    - AS-External-LSA

BGP4+

- BGPのマルチプロトコル対応版
  - 昨今だと、ほとんどの商用実装で対応済み
  - もちろんIPv6の経路もこれで運べる
- 基本的な動きはBGP4に準拠
  - 経路優先度など
- 付随する機能もマルチプロトコル対応済み
  - Route-Refresh等

BGP4+関連RFC

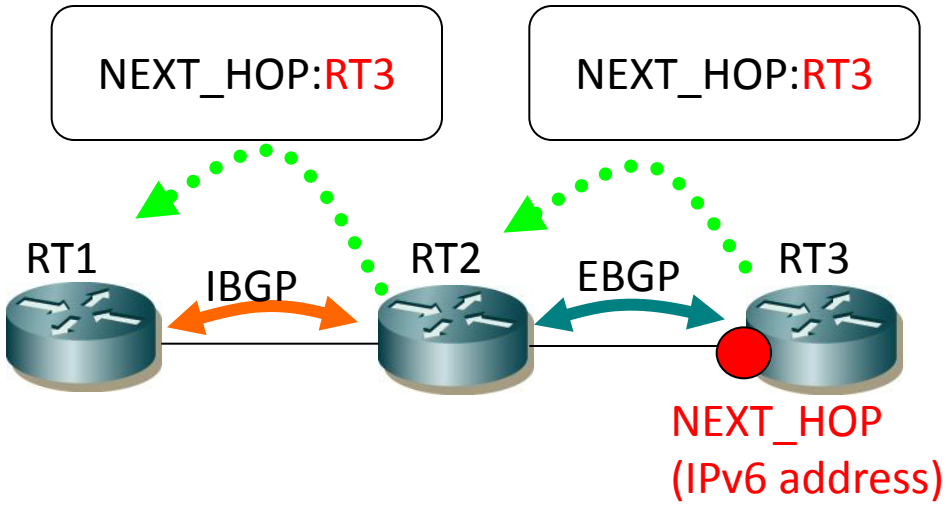
- [RFC4271] A Border Gateway Protocol 4
  - 基本中の基本
- [RFC4760] Multiprotocol Extensions for BGP-4
  - マルチプロトコル対応
- [RFC2545] Use of BGP-4 Multiprotocol Extensions for IPv6 Inter-Domain Routing
  - IPv6利用時のネクストホップなどの検討
- 後はこの辺も
  - [RFC1997] BGP Communities Attribute
  - [RFC4451] BGP MED Considerations
  - [RFC4456] BGP Route Reflection
  - [RFC4893] BGP Support for Four-octet AS Number Space
  - [RFC5004] Avoid BGP Best Path Transitions from One External to Another
  - [RFC5065] AS Confederations for BGP

IPv6とBGP4+

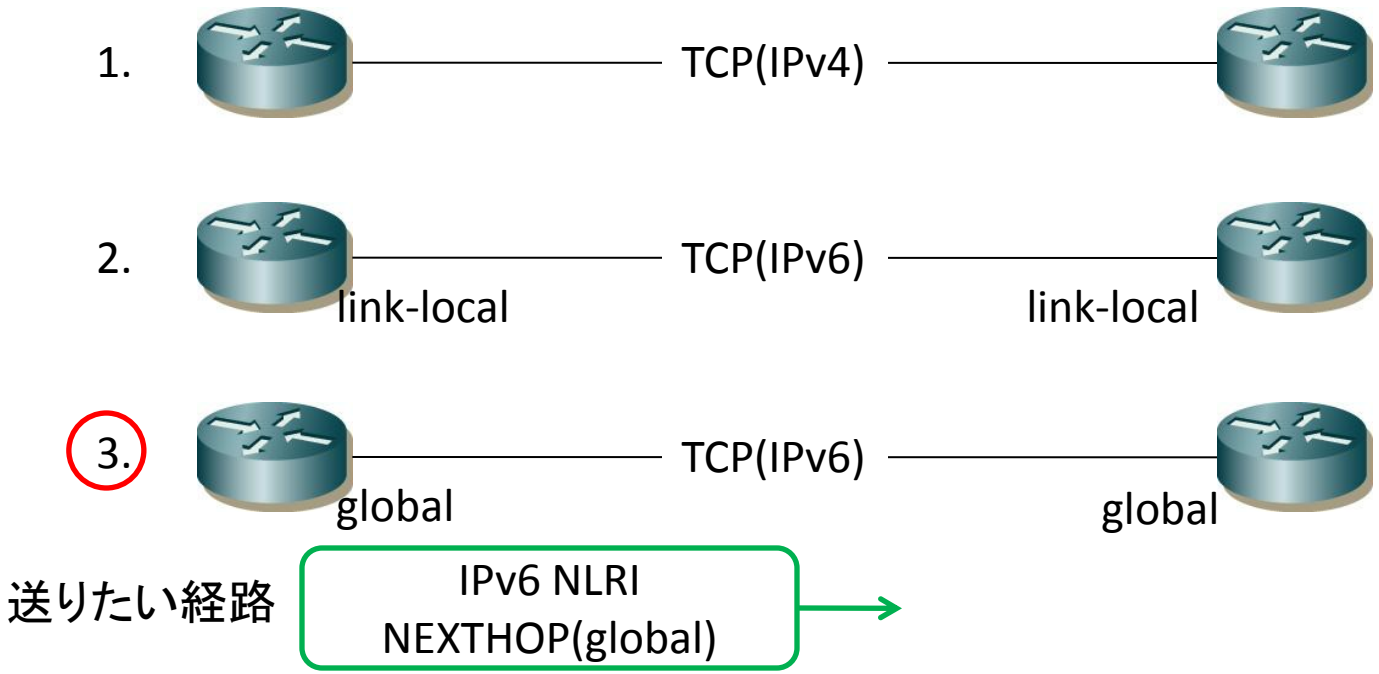
- IPv6 prefixはパス属性として交換されてる
  - MP_REACH_NLRI (Type14)
  - MP_UNREACH_NLRI (Type15)
- BGP OPENで相互にmultiprotocol対応を通知
  - OPENメッセージのCapability広告を利用
  - 接続するには双方が対応している必要がある

BGPのネクストホップ属性

- 基本はグローバルアドレスを指定
  - リンクローカルアドレスが含まれる場合もある
  - iBGPではネクストホップをそのまま伝搬するので、リンクローカルのみでは動作しない



BGP接続と交換する経路



- IPv6グローバルアドレスでBGP接続しておくのが簡単

iBGP、eBGPと接続

- eBGP
  - インタフェースのグローバルアドレスでBGP接続
- iBGP
  - loopbackのグローバルアドレスでBGP接続
  - PathMTUDが有効なので注意
- 静的に設定したアドレスを利用する方が便利
  - EUI64形式だと機器の変更時に設定変更が発生

BGP4+設定

- TCPの接続に関する部分
  - 接続するIPアドレス
  - MD5認証の利用など
- 交換する経路に関する部分
  - 交換する経路のアドレスファミリ
  - 経路フィルタ等のポリシー

※一つのTCP接続上で複数のアドレスファミリの経路を扱うことはできるが、お勧めしない

経路フィルタ

- IPv4と同様にprefixでのフィルタも記述できる

cisco ios

```
ipv6 prefix-list  LIST-NAME seq 5 permit 2001:db8::/32
```

juniper junos

```
policy-options  { policy-statement {  POLICY-NAME {
            term TERM-NAME {
                    from { route-filter 2001:db8::/32 exact;}
                    then next policy;
            }
            then reject;
}}
```

その他のパス属性

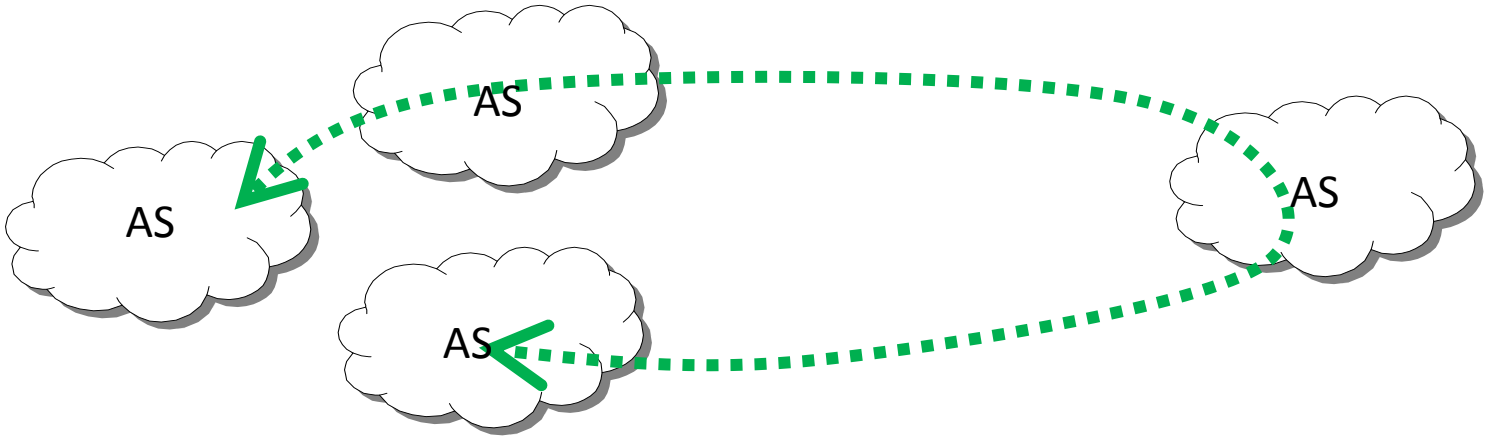
- IPv4での場合と同じ
  - Local Preference
  - AS Path
  - MED
  - NEXT_HOP METRIC
  - BGP Community

トラブル事例

- IPv6ネットワークトポロジ
- NDP
- routing loop
- Path MTU Discovery
- BGP接続
- ルータのバグ/仕様

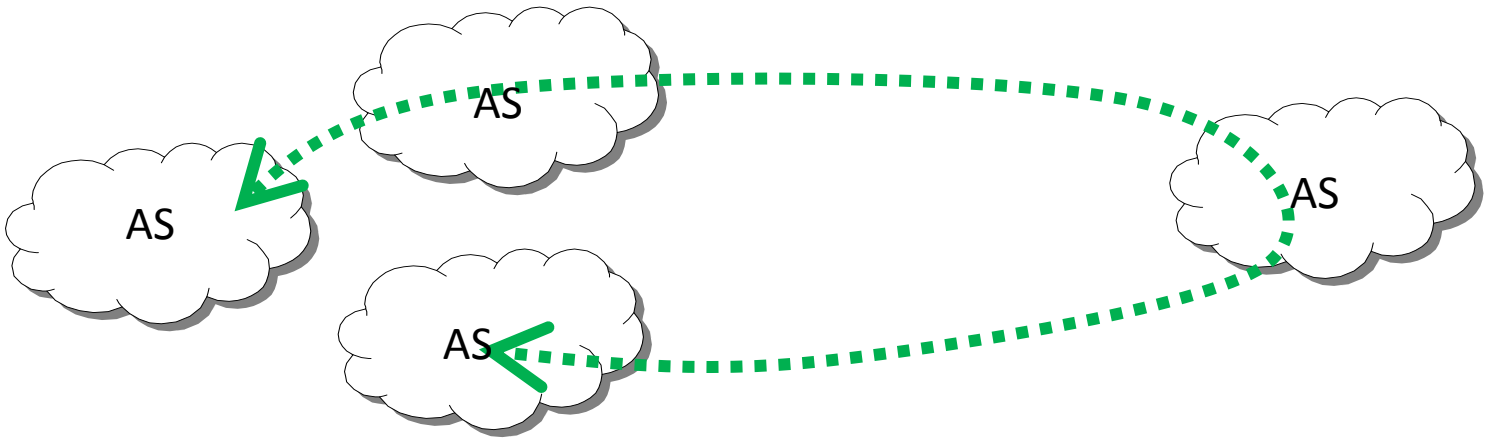
IPv6ネットワークトポロジ – AS間

- 問題
  - 近傍にあるASへの通信が遠回り
- 特徴
  - ヨーロッパなど遠方にあるサイトではほぼ誤差
  - 国内等、近傍にあるネットワークで遅延が大きい



IPv6ネットワークトポロジ – AS間

- 想定される原因
  - 国内で相互接続が十分ではない
- 解決策
  - IX等を利用して、IPv6の相互接続を進める



IPv6ネットワークトポロジ – 網内

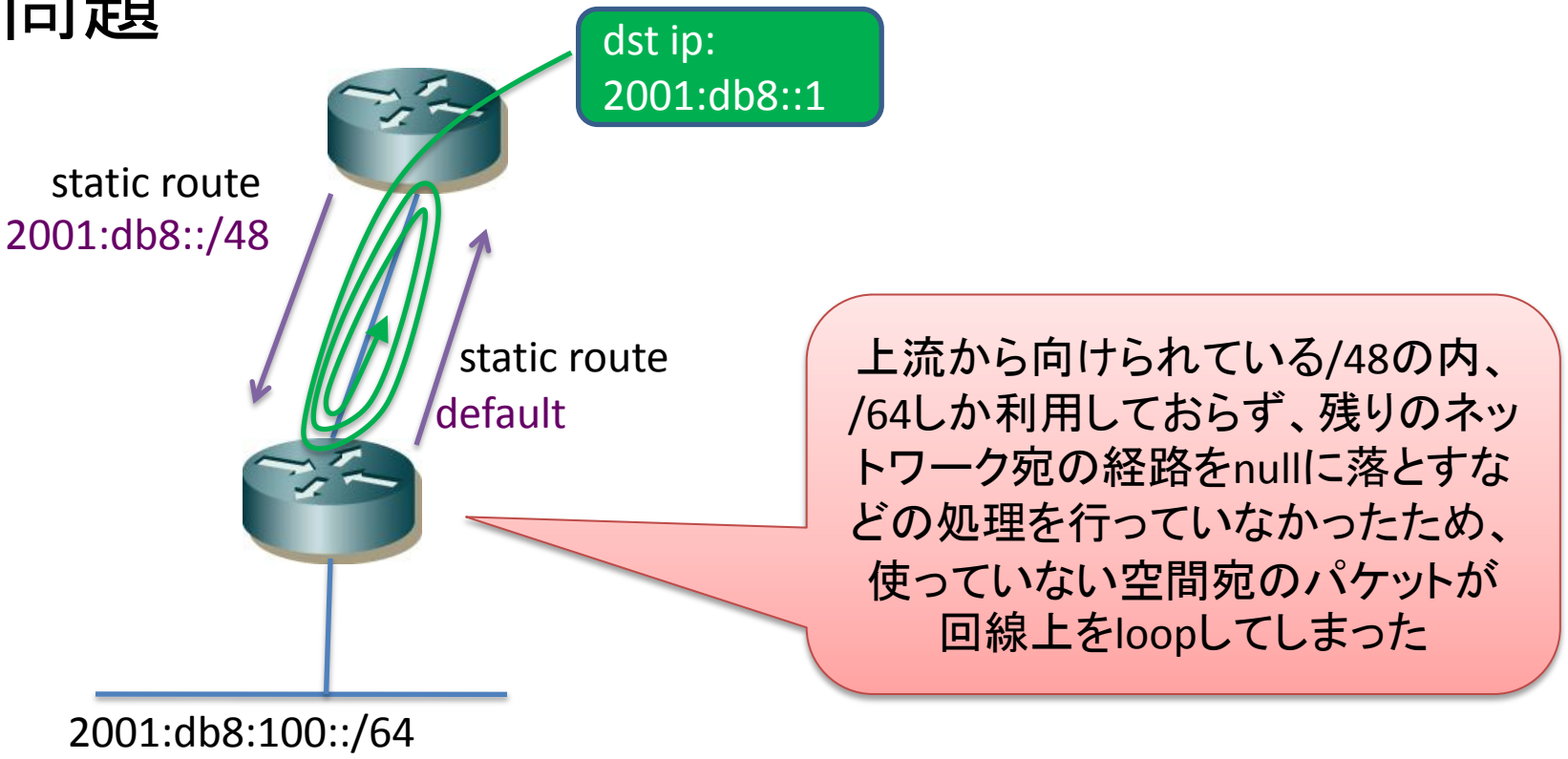
- 問題
  - トラヒックが思わぬ所を経由する
- 想定される原因
  - IPv4とIPv6でネットワーク構成が異なる
    - IGPはIPv4とIPv6で異なるトポロジ情報を維持
- 解決策
  - できるだけIPv4/v6で同じトポロジを心がける, or
  - IPv4とIPv6で異なるトラヒック制御を考える

ndpテーブル

- 問題
  - ルータのndpのテーブルサイズがあふれる
- 想定される原因
  - 標準で/64 = 2^64アドレス空間であり、膨大
  - 外部からアクセスがあればndpでL2アドレス解決
- 解決策
  - 対策済みのルータ実装を利用する, or
  - 適切なprefix長に変えてしまう

loop – 終端されていない経路

- 問題



loop – 終端されていない経路

- 解決策
  - 静的にむけられている経路を必ず終端する
    - 全体を終端しておいて、各ネットワークへの到達性は細かな経路で確保する

cisco ios

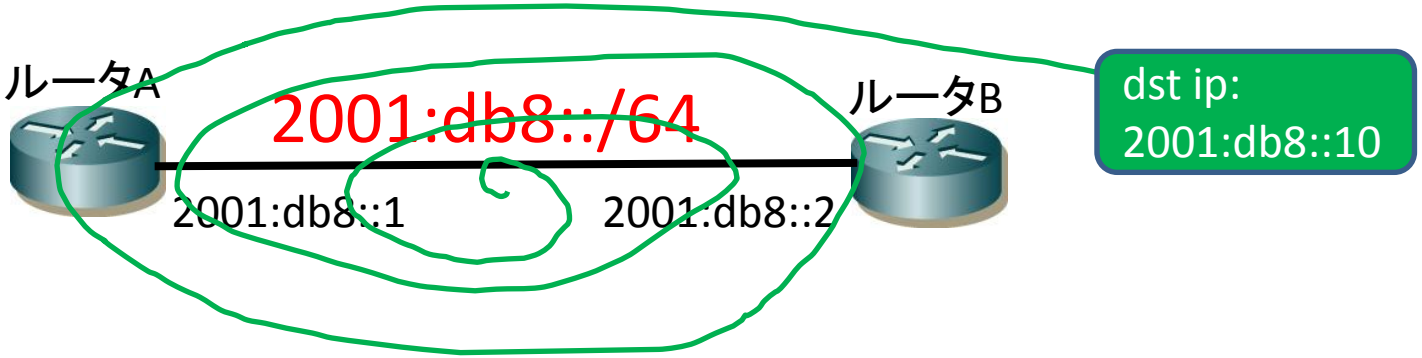
```
ipv6 route 2001:db8::/48 Null0
```

juniper junos

```
routing-options  { rib inet6.0 {  static {
                route 2001:db8::/48 discard;
}}}
```

loop – point-to-pointリンク

- 問題

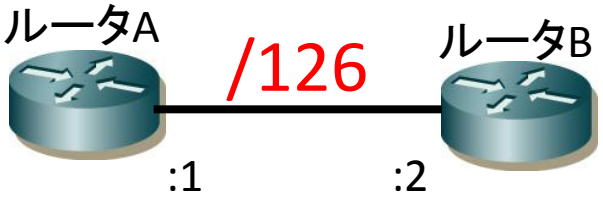


loop – point-to-pointリンク

- 主にルータ間で多用されている回線
  - POS、Serialなどなど
  - 実はtunnelもpoint-to-pointリンク
- パケットを回線に投げれば対向に届く
  - Layer2アドレス解決のためのarpとかいらない
- リンク上に/64等のネットワークがあるように設定している
- RFC4443では、ループを起こさない様に規定

loop – point-to-pointリンク

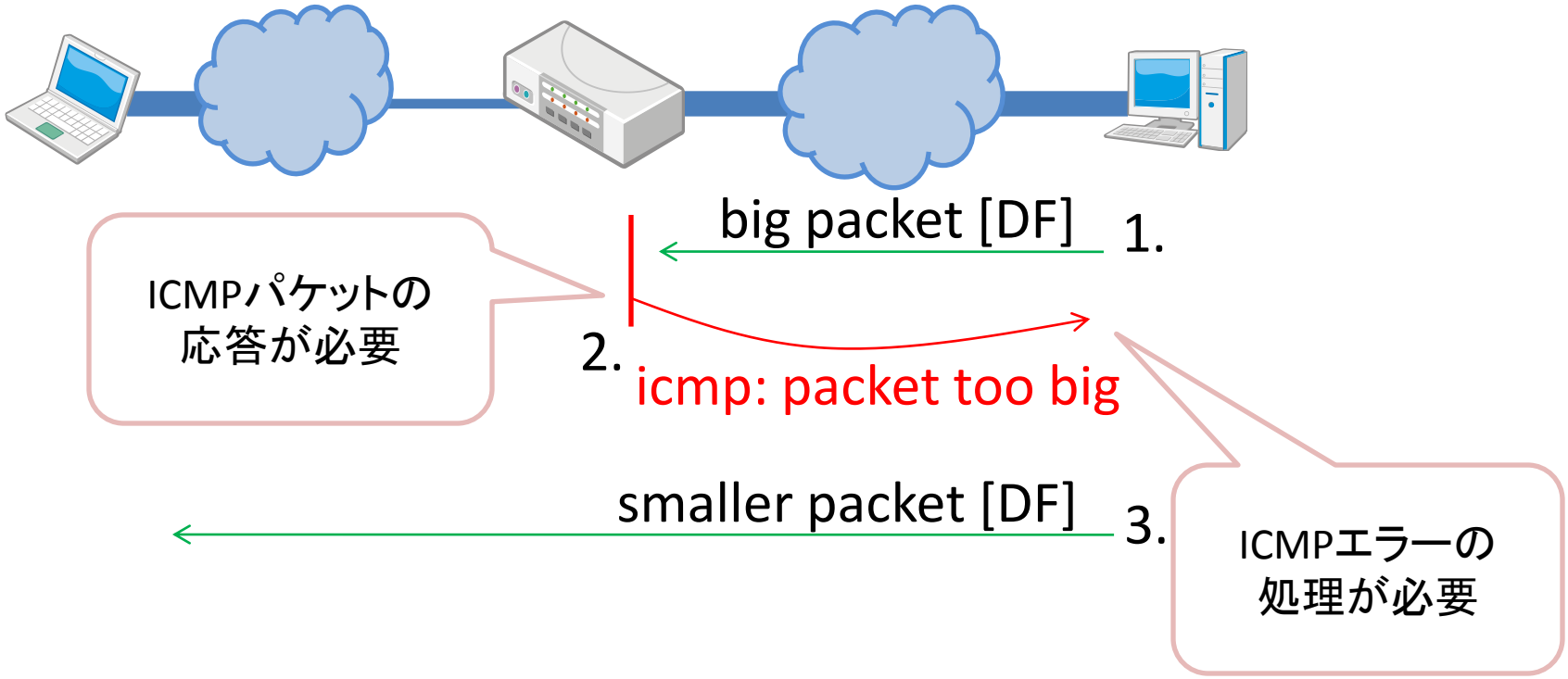
- 2001:db8::/126でもダメ
  - 2001:db8::0 ← Subnet Router-anycast address
  - 2001:db8::1 ← ルータA
  - 2001:db8::2 ← ルータB
  - 2001:db8::3 ← 空き



loop – point-to-pointリンク

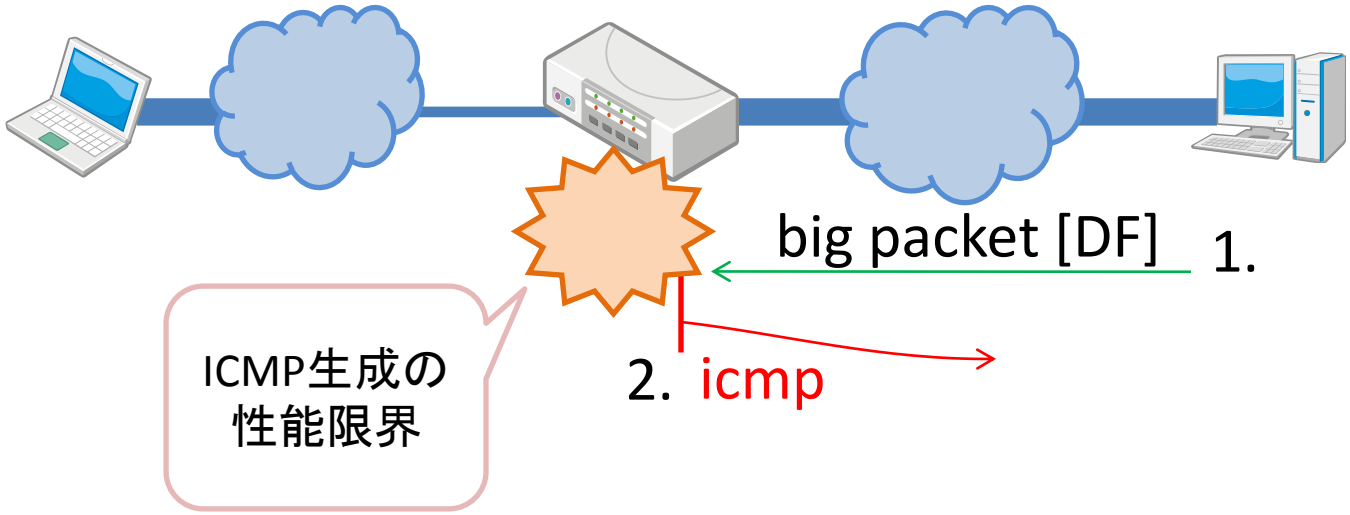
- 解決策
  - 対策済みのルータ実装を使う, or
  - /127アドレッシングを利用する
    - http://datatracker.ietf.org/doc/draft-ietf-6man-prefixlen-p2p/

Path MTU discovery



Path MTU discovery

- ICMPメッセージを生成するルータの性能



Path MTU discovery

- cisco ios
  - ip icmp rate-limit unreachable 500
    - means icmp errors are limited to one every 500msec
  - ipv6 icmp error-interval 100
    - means icmp errors are limited to one every 100msec
- juniper junos
  - icmpv4-rate-limit {packet-rate 1000;};
    - means max 1000pps for icmp to/from RE
  - icmpv6-rate-limit {packet-rate 1000;};
    - means max 1000pps for icmp to/from RE

Path MTU discovery

- 解決策
  - TCP MSSハックを実装してもらう, and
  - 頑張ってICMPエラーを返せるようにする

BGP接続 – nexthop

- 問題
  - 経路は交換できているけどトラヒックが流れない
- 想定される原因
  - トラヒックがそもそも少ない ☹
  - nexthopのアドレスに到達性がない
- 解決策
  - nexthopがglobalアドレスである事を確認, and
  - nexthopのアドレスを網内に広報する

BGP接続 – 到達性

- 問題
  - 経路広報しても、到達できないASがある
- 想定される原因
  - 他のネットワークでの経路フィルタ
- 解決策
  - 割り振られたサイズで経路広報してみる ex /32
  - 該当ASにコンタクトしてみる

ルータの制限事項に注意する

- テーブルサイズ
  - RIBテーブル
  - FIBテーブル
  - ndpテーブル
- 扱えるprefix長
  - /64より細かい経路で発生しやすい
  - この問題に関して、internet-draftを作成予定

まとめ

- トラブル傾向
  - 設計ミス
  - 設定ミス
  - 機器の仕様やバグ
- IPv6の特性を把握しつつ、今風な運用を心掛けることが必要