S7 サービスプロバイダ
バックボーン設計入門 前編

ISPにおける経路設計

KDDI総合研究所
宮坂拓也

はじめに

- 本資料は**ISPにおけるIGPの基本的設計**について共有するものです

- ISPにおけるIGPの基本的設計を説明するために、本資料ではIGPの一つである**OSPF**を例にして説明します

- OSPFやIS-ISといった**プロトコル自体の詳細説明は実施しません**
  - 多くの書籍・web解説があるのでそちらを参照してください

自己紹介

- 名前：宮坂拓也 (みやさかたくや)
- 経歴：
  - 2011/4：KDDI入社
  - 2011/4～2018/3：KDDIのバックボーンネットワーク(IP/MPLS)の設計開発
  - 2018/4～現在：KDDI総合研究所にて、ネットワーク関連の研究開発
- その他活動：
  - JANOG 運営委員
  - IETF 主にRouting areaでの標準化活動

Agenda

1. IGP基本事項おさらい
2. IGP経路広報ポリシー
3. IGP エリアデザインとBGPとの関係性について
4. 障害時に早く切り替えるために

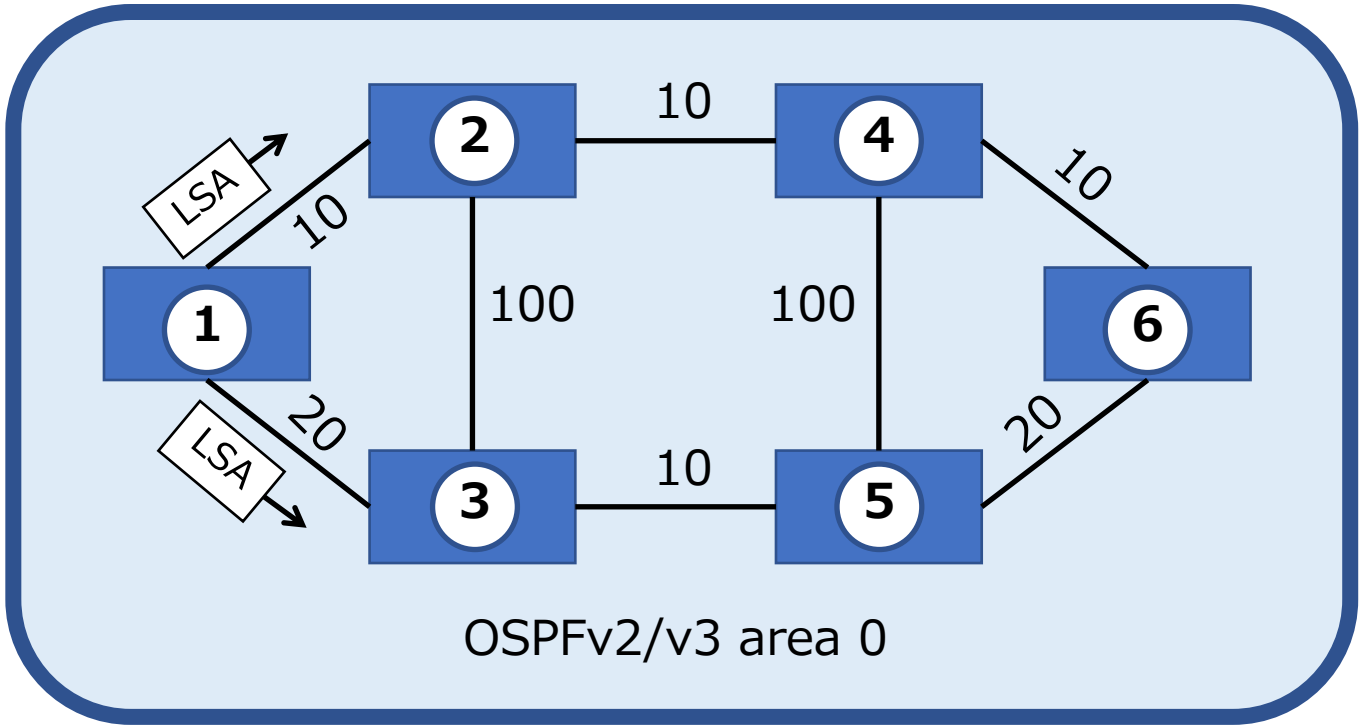
Agenda

1. IGP基本事項おさらい
2. IGP経路広報ポリシー
3. IGP エリアデザインとBGPとの関係性について
4. 障害時に早く切り替えるために

IGPの役割

- IGP
  - Interior Gateway Protocol
    - ネットワーク内の各ノードの経路情報を広報するプロトコル
    - リンクステート型ルーティングプロトコル
  - 現在主に利用されているプロトコル
    - OSPF
      - https://datatracker.ietf.org/doc/rfc2328/ (OSPFv2)
      - https://datatracker.ietf.org/doc/rfc5340/ (OSPFv3)
    - IS-IS
      - https://datatracker.ietf.org/doc/rfc1142/

IGP / リンクステート型ルーティング



- IGP
  - 自身のリンク情報と、他のノードから受信したリンク情報を隣接ノードに広報し、ネットワーク内でリンク情報を伝搬させる
  - 各ノードは現在のリンク情報(接続情報・コスト)を元に、各ノードへの最短経路を計算

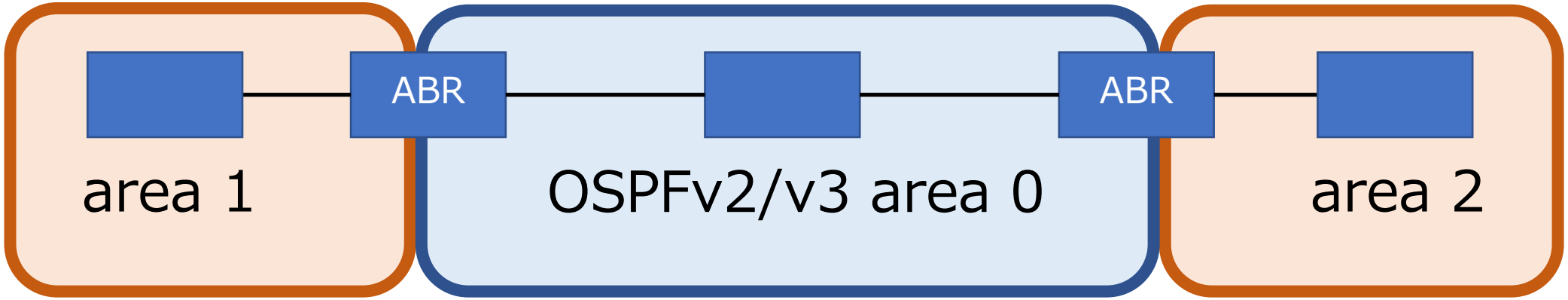
IGP：Neighbor

- Neighbor：リンクステート情報を交換する関係
- Neighbor Type
  - Point-to-point：2ルーターのみの接続
  - Broadcast：複数のルーターが同一セグメント上で接続
- Keep-alive
  - Helloパケットを定期的に送信して、死活監視を行う

IGP：Area

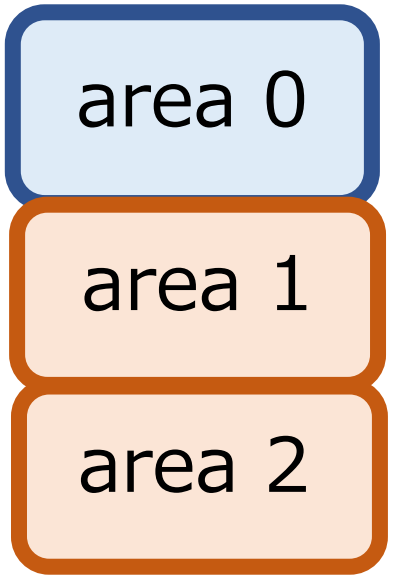
- エリア：ネットワークのグルーピング (RFC2328)
- バックボーンエリア (area 0)
  - すべてのエリアのIGP経路を交換・伝搬させる
- Non-バックボーンエリア (area 1,2…)
  - そのエリアの経路は詳細なリンクステート情報をもらう
  - Area 0から、サマライズされた他のエリアの経路をもらう
  - Area 0へは、そのエリアをサマライズした経路をわたす

IGPネットワーク内のリンクステートデータを減らすことで、計算量を削減することができる



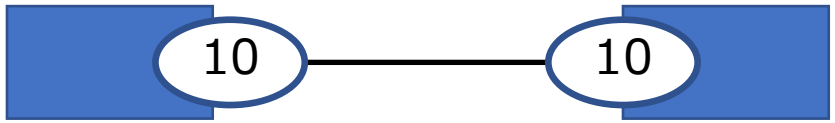
IGP：Area注意点

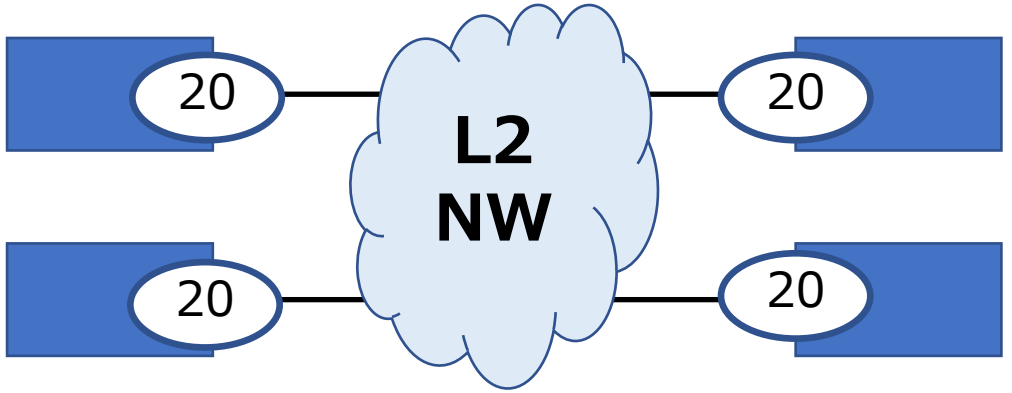
- バックボーンエリアは分断できない
- エリアの階層構造はできない
  - Virtual Linkという解決策もあるが、ネットワークが複雑になるため、基本的に利用しない方が良い



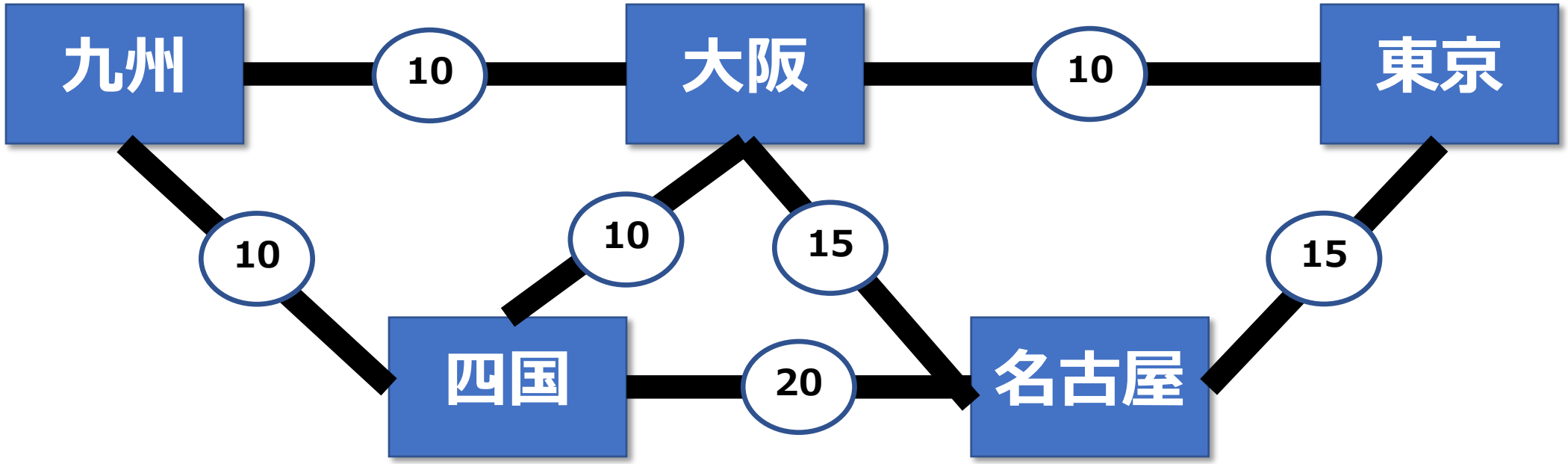
IGP：コスト

- 対象リンクの”重み”を指定し、IGPによって広報する
- SPF計算では、各ネットワークのコストを元に、ネットワーク内の各ノードへの最短経路をダイクストラ法により計算を行う
- 帯域幅を元に自動で計算することもできるが、ネットワークデザインに合わせて手動で設定するのが望まれる

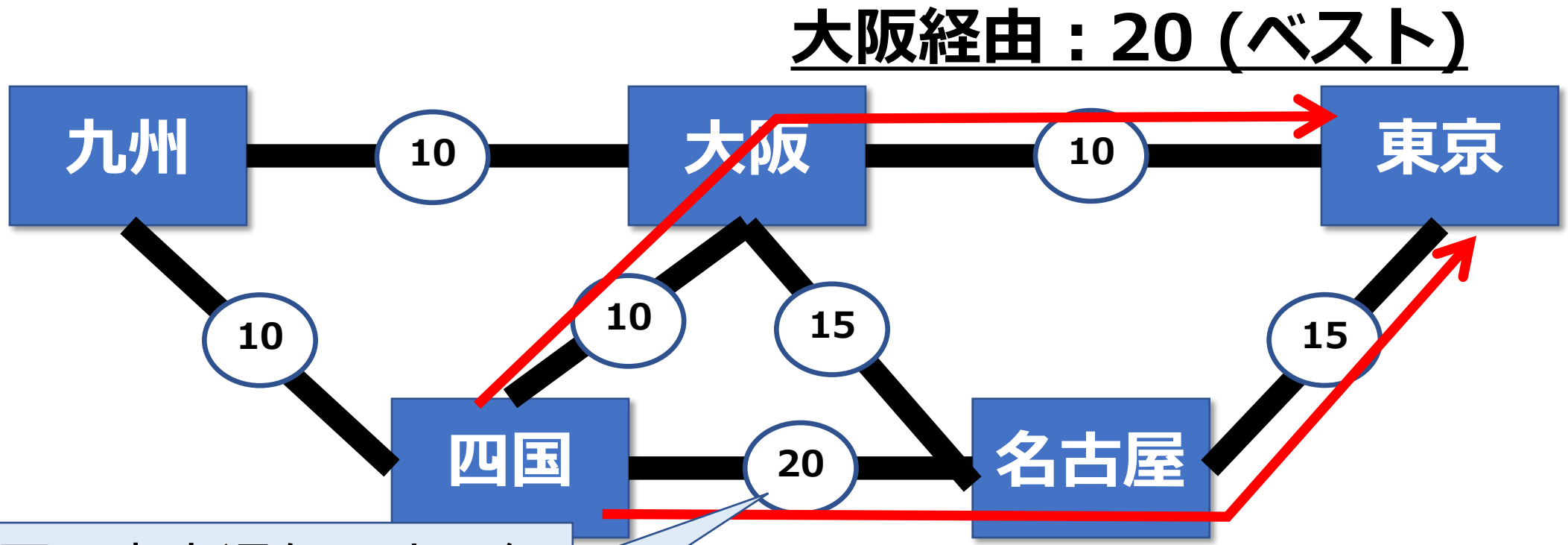




IGP：コスト設計の例

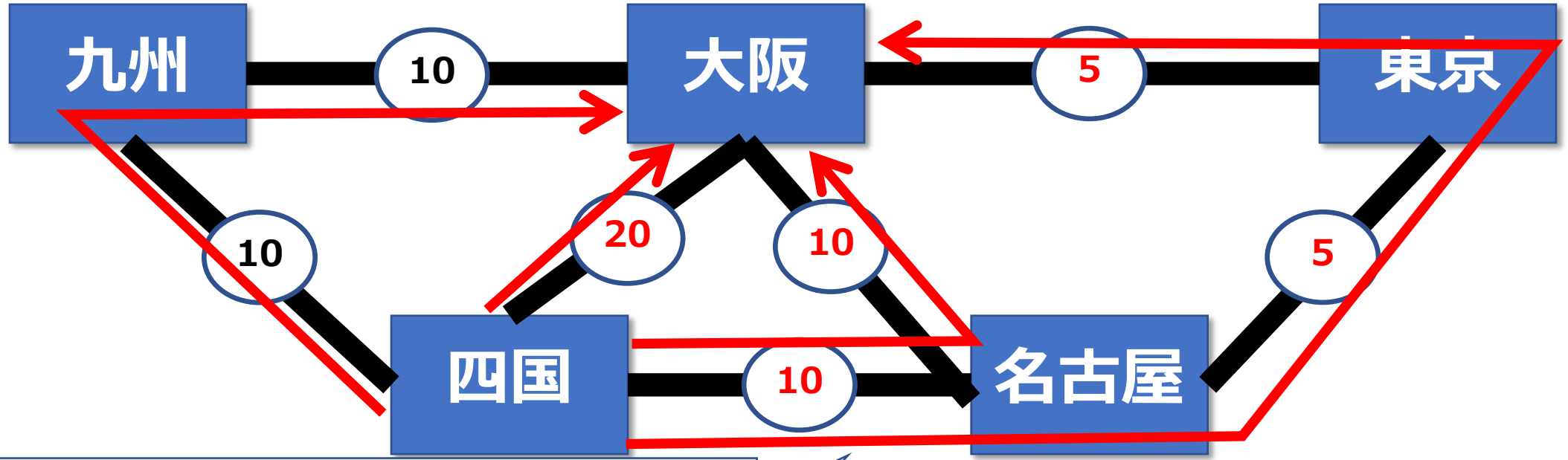


IGP：コスト設計の例



四国→東京通信は大阪経由にしたいので、名古屋向け経路を少し高める

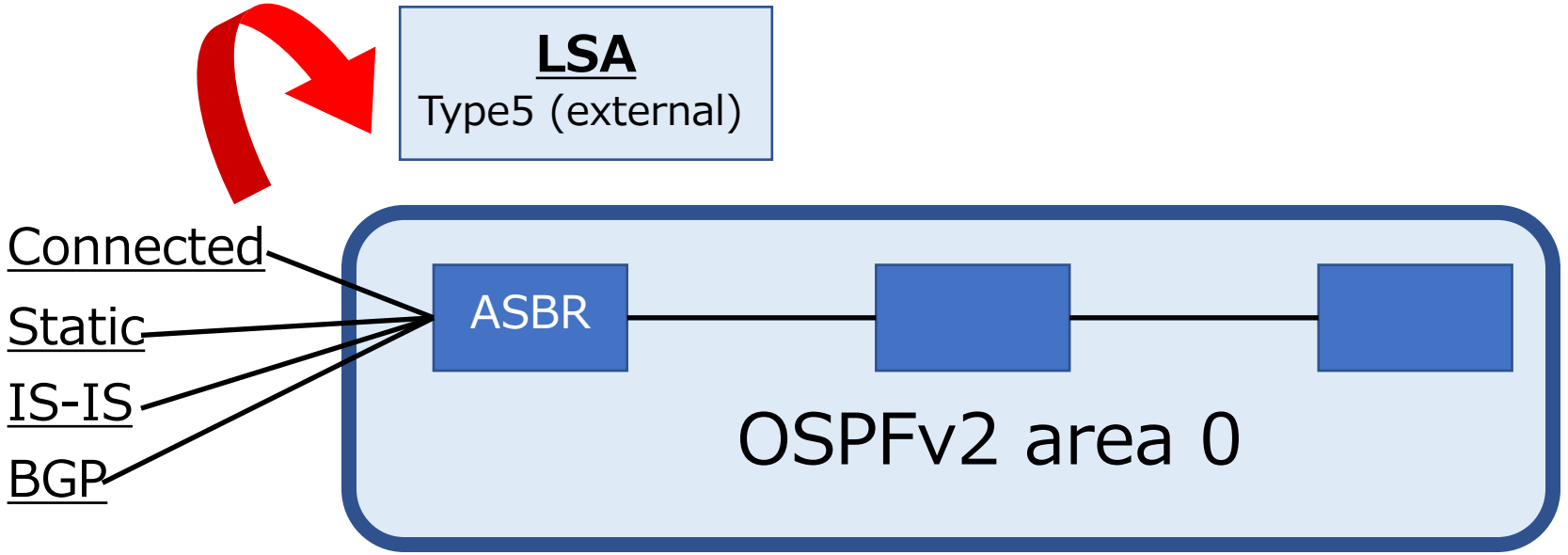
IGP：ECMP (Equal Cost Multi-Path)



四国→大阪通信はすべての経路で等しいコストになるので、トラフィックが分散される

IGP：外部経路(Redistribution)

- 異なるプロトコルを外部経路として、IGPに経路を注入することができる
- 後述するが、IGPは計算コストが高いプロトコルであるため、**IGPへの外部経路の注入はできるだけ避け**、BGPへのRedistributionを基本にすることが良い

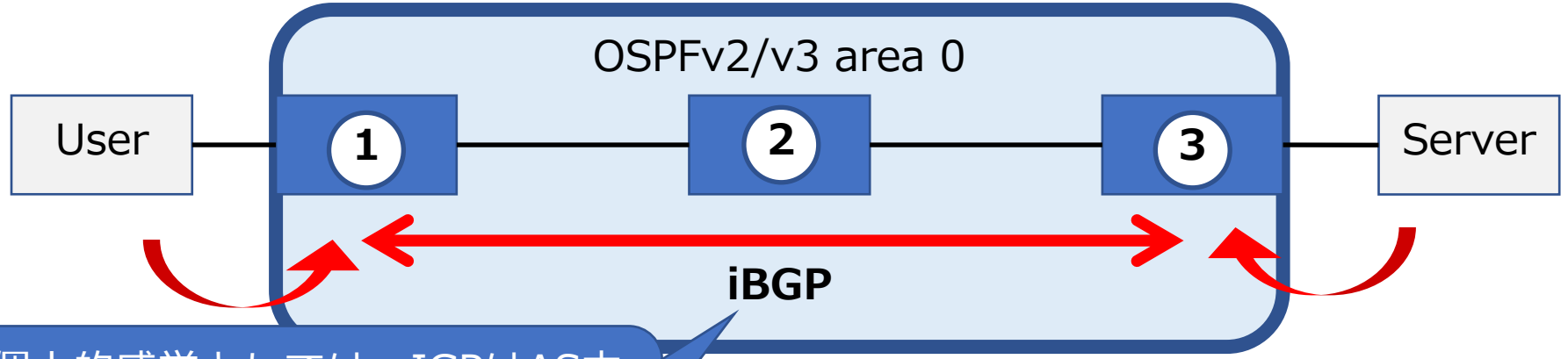


Agenda

1. IGP基本事項おさらい
2. IGP経路広報ポリシー
3. IGP エリアデザインとBGPとの関係性について
4. 障害時に早く切り替えるために

IGPにどんな経路をのせるのか？

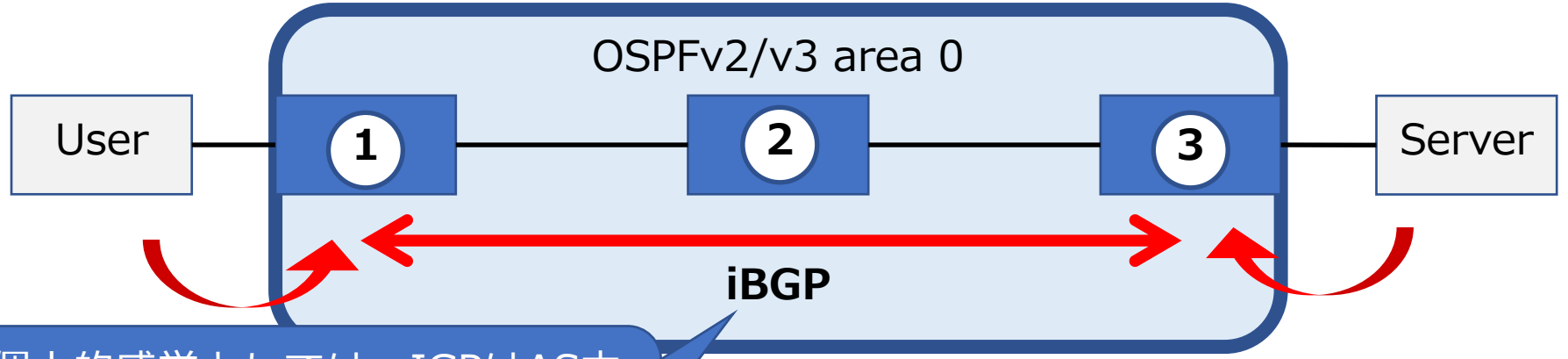
- IGPの基本的役割
  - 各ルーターのインターフェース経路を広報すること
    - Loopbackアドレス
      - iBGPを貼るためのアドレス
    - リンクアドレス
      - iBGPの宛先であるLoopbackアドレスへ導くアドレス



個人的感覚としては、IGPはAS内で**iBGPを接続するために利用する**プロトコルというイメージ

IGPにどんな経路をのせるのか？

- IGPの基本的役割
  - ユーザー経路やサーバー経路といった、実際にトラフィックがのる経路はIGPにのせない (BGPにのせる)
    - IGPはBGPに比べて計算量が大きく、なるべくIGP経路を最小にし、それ以外のものはBGPで経路広報するデザインが好まれる

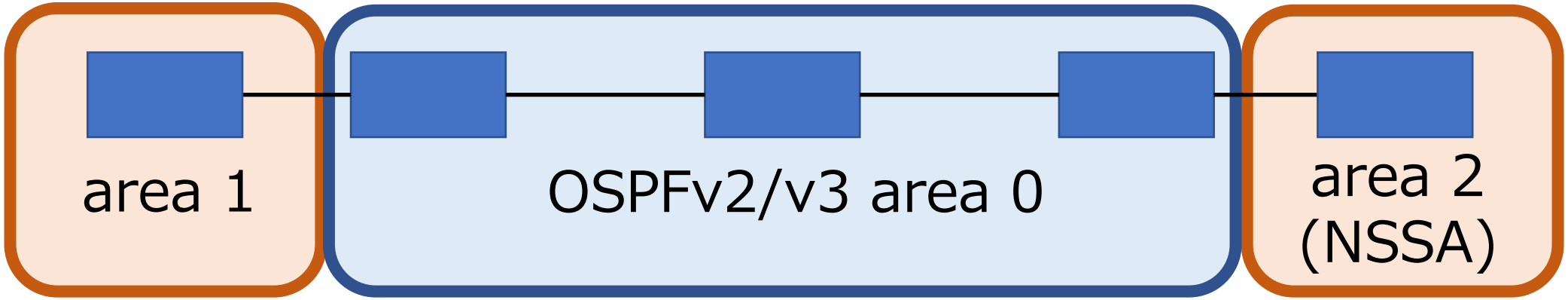


Agenda

1. IGP基本事項おさらい
2. IGP経路広報ポリシー
3. IGP エリアデザインとBGPとの関係性について
4. 障害時に早く切り替えるために

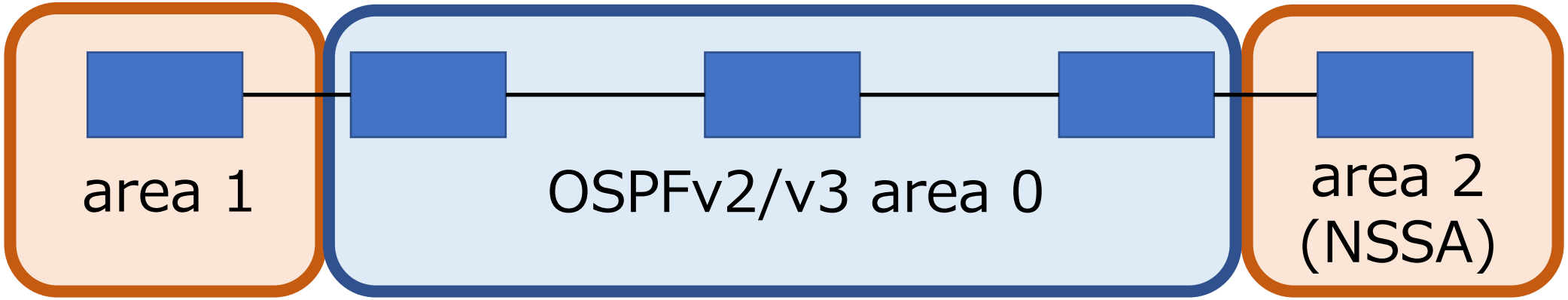
IGP Areaデザイン (1/2)

- IGP Areaデザイン
  - もちろん、バックボーンエリア(area 0)を中心に考えて設計
  - エリアを分ける理由・モチベーション
    - 非力なルーター/サーバーをエリアに所属させたい
    - 運用ポリシー上、エリアを分けたい
      - バックボーンネットワークをarea 0に、各地域のIGP areaをsub areaにする



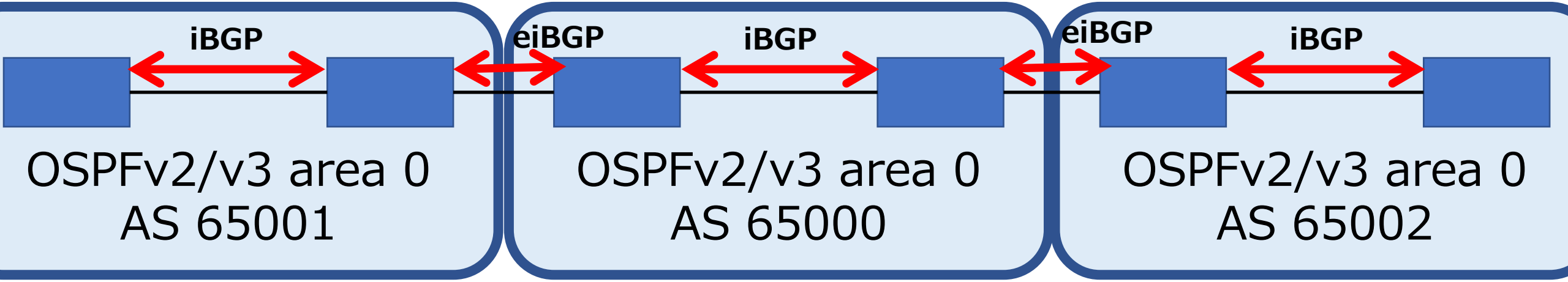
IGP Areaデザイン (1/2)

- IGP Areaデザイン
  - エリアを分ける場合、基本的には標準エリアにする、更に非力なものがいればtotally-stubやNSSAなどを検討
  - RSVP-TE, Segment Routingなどが、IGP Inter-areaの動作をサポートしない事や、あまりこなれていないということもある



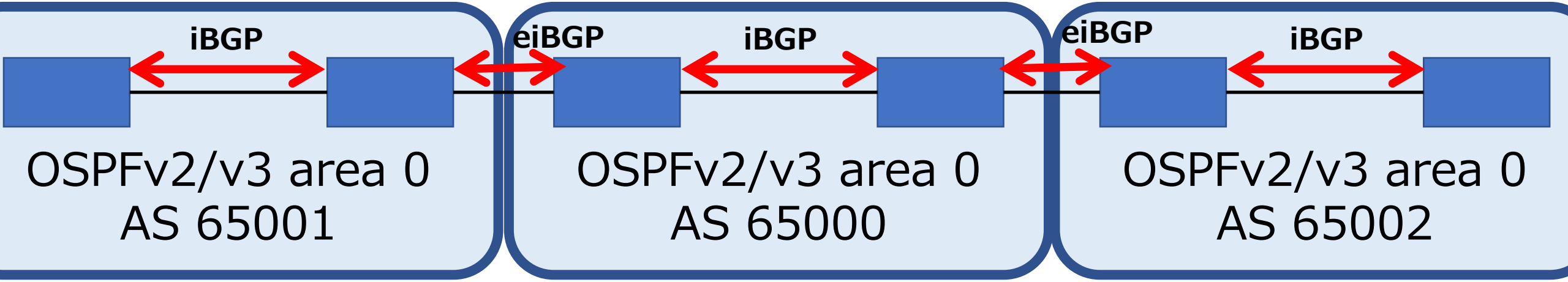
IGP Areaデザイン (2/2)

- ASとエリアの関係性
  - **「1つのASに、1つのバックボーンエリア」**がシンプルで楽
    - IGP＝AS内詳細経路をRouting, BGP=AS間で必要な経路をRouting
    - グローバルASをプライベートASで切っている場合も、同様が良い



IGP Areaデザイン (2/2)

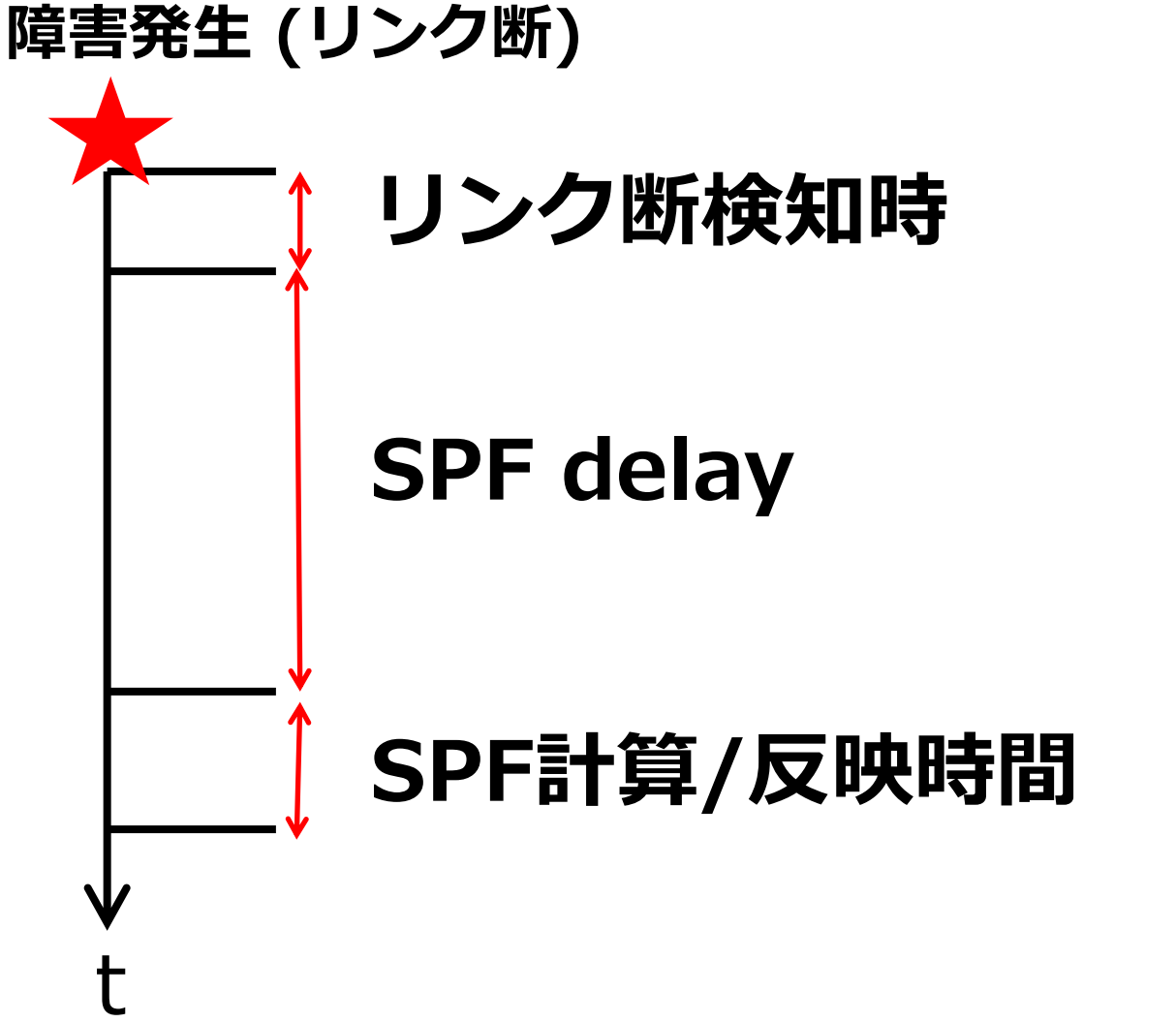
- ASとエリアの関係性
  - **(基本的に) BGPにIGPの経路をのせない！**
    - 先に書いた通り、BGPはサービス経路のみのせる
  - **(基本的に) IGPにBGPの経路をのせない！**
    - 誤ってインターネットフルルートがIGPに流れたら死亡します

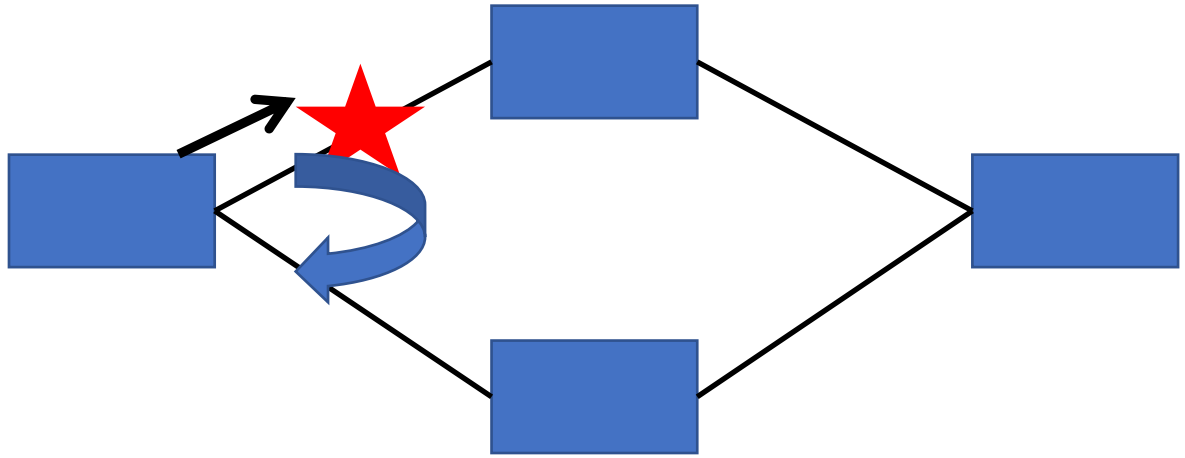


Agenda

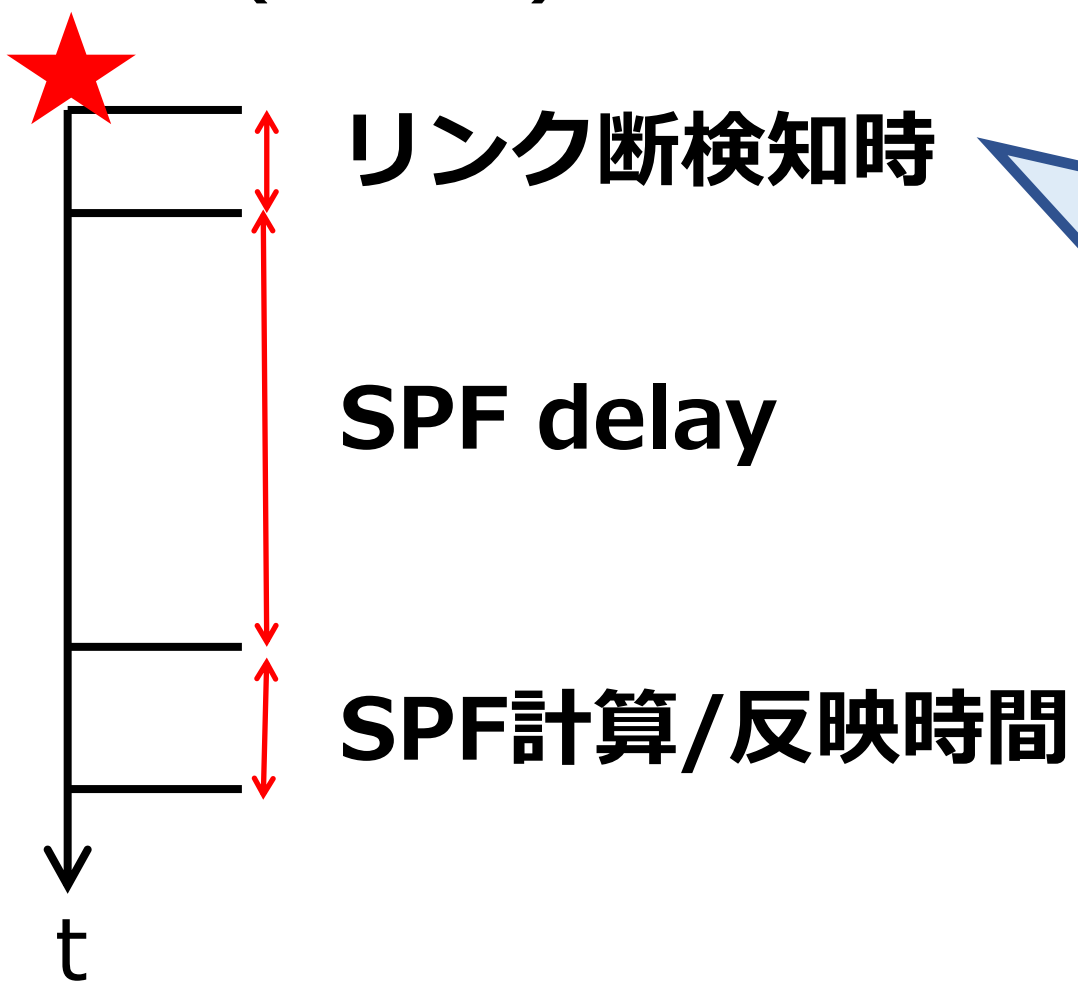
1. IGP基本事項おさらい
2. IGP経路広報ポリシー
3. IGP エリアデザインとBGPとの関係性について
4. 障害時に早く切り替えるために

障害対策：リンク障害発生時の断時間





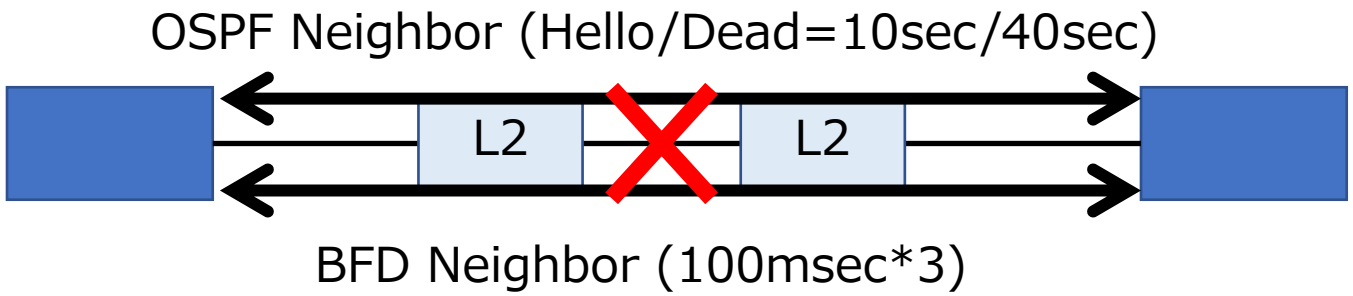
障害対策：リンク障害発生時の断時間



ルーターのリンクが落ちたことを検知するまでの時間。

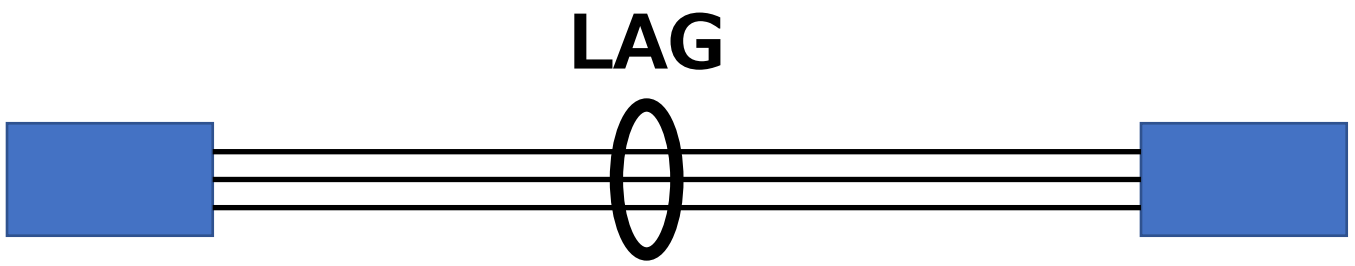
サイレント障害・途中にL2SWなどがある場合は要注意
→次のスライド

障害対策：リンク障害発生時の断時間



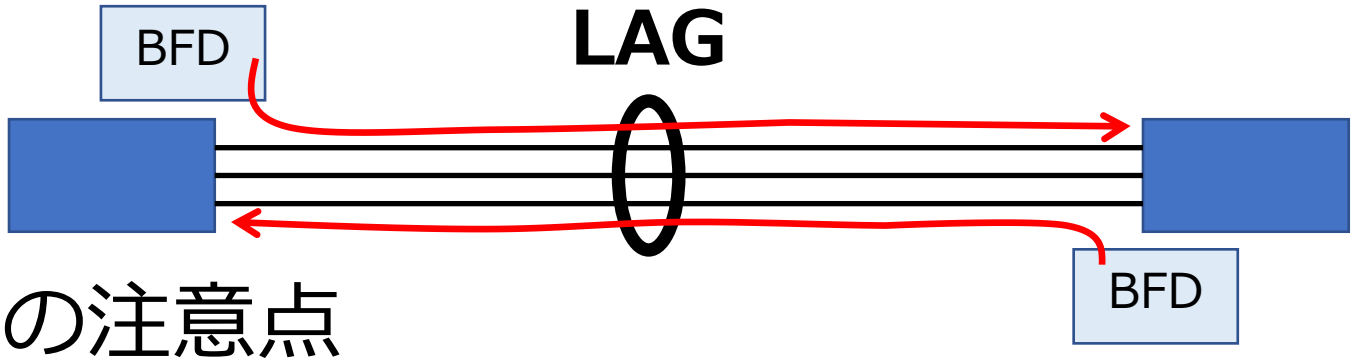
- リンクサイレント障害時の検知方法
  - Hello PacketのDead Timerで検知するのではなく、BFDなどのサイレント障害検知用プロトコルを利用しましょう
  - 最近の機器では少ないが、もしBFDなどをサポートしない場合はHello/Dead timerを短くするしかないが、短くしすぎるとルーターのCPU利用率に要注意

障害対策：LAGとリンク障害 (1/2)



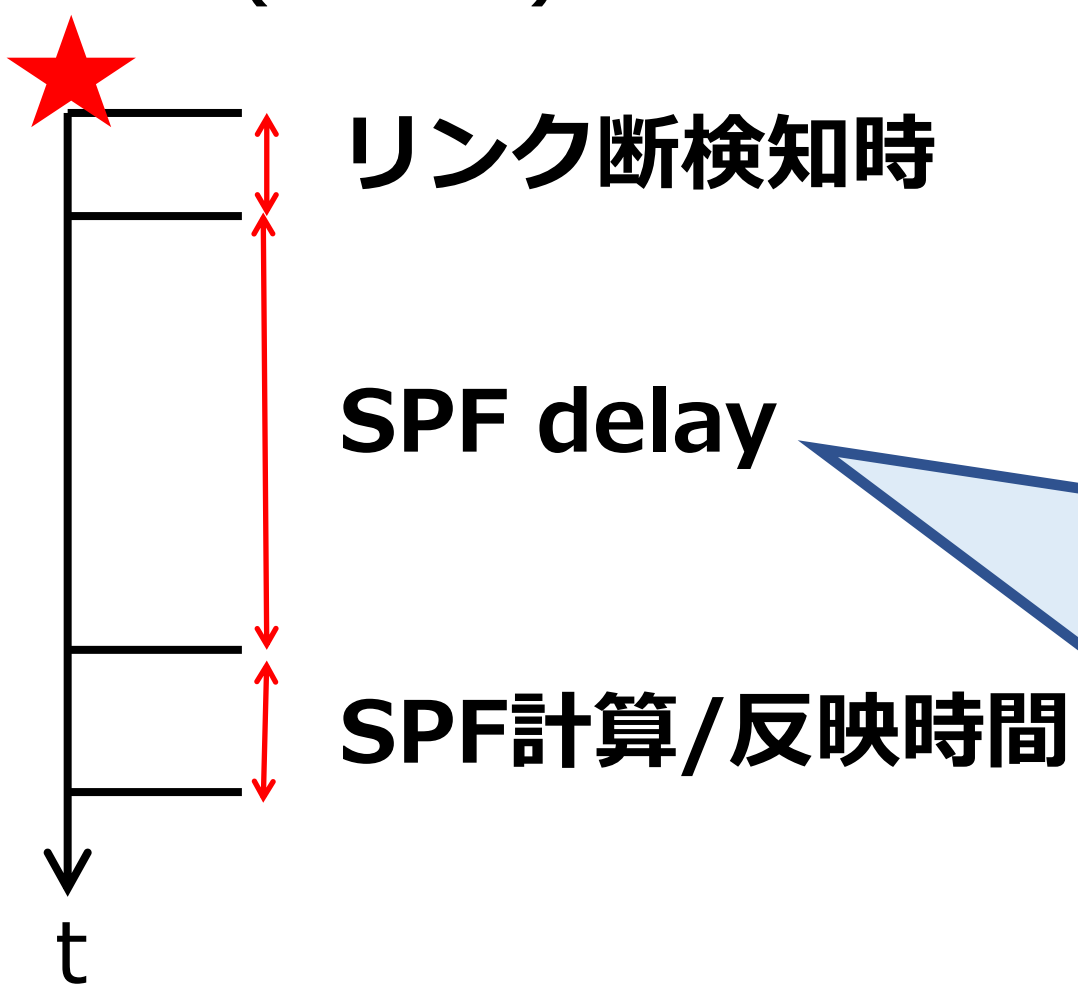
- Link Aggregation (LAG) / Bonding
  - 複数のインターフェースを束ねて論理的に１つのインターフェースにする技術
    - IGPネイバーの数やリンクステート情報を少なくすることができるので、安定性が増す
    - 論理ネットワークトポロジーもわかりやすくなる
    - LACPによるメンバー管理・死活監視が可能

障害対策：LAGとリンク障害 (2/2)



- LAGとBFDの注意点
  - LAGインターフェース上でBFDを有効にした場合、**通常ではBFDパケットがLAGメンバーすべてに通ってくれない**ことがある。つまり、たまたまBFDパケットが通っていないリンクでサイレント障害が起きた場合はブラックホールされてしまう ☹
  - BFD on LAG (RFC 7130)
    - すべてのLAGメンバー上でmicro-BFDセッションを確立
    - 比較的新しい技術であるため、実装されているか？異ベンダー間で相互接続できるか？などは要確認

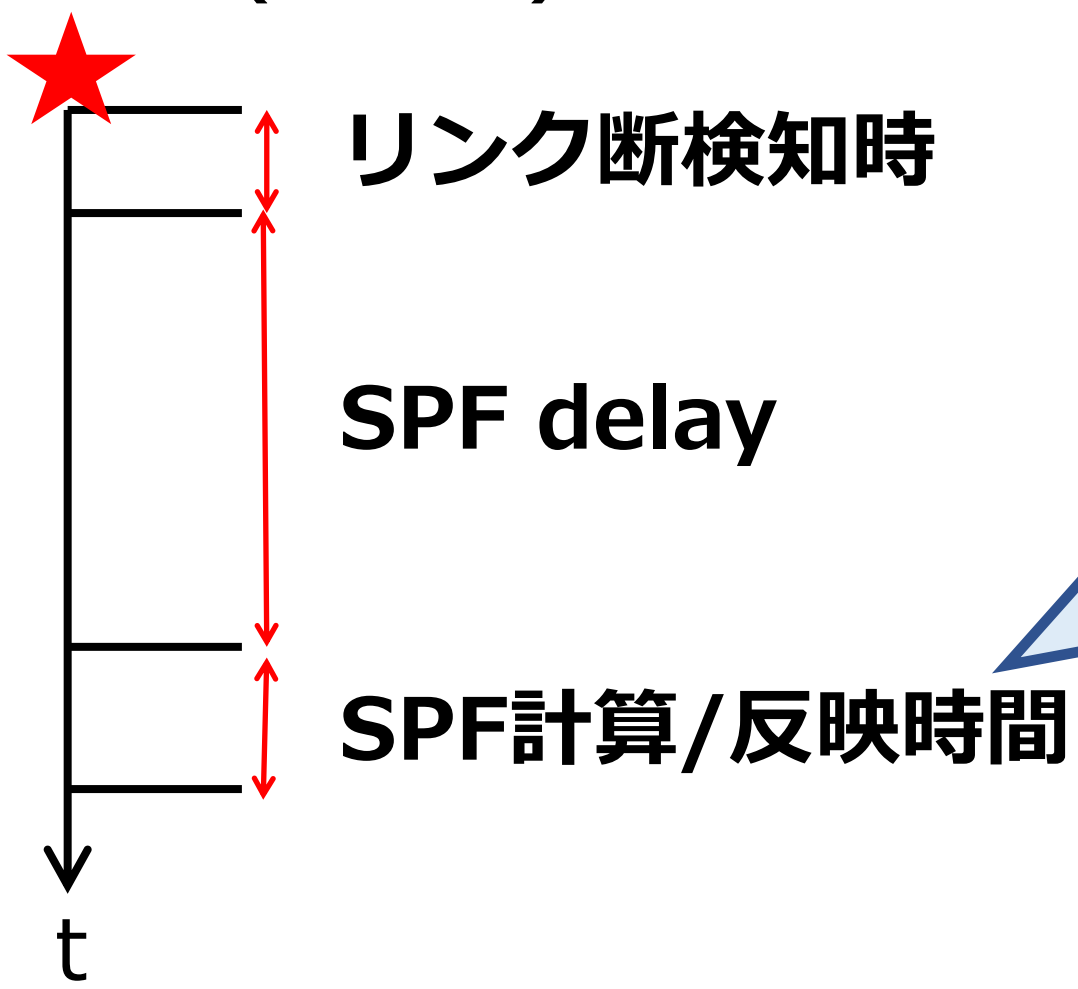
障害対策：リンク障害発生時の断時間



リンクステート情報にアップデートがあったときに、フラップ防止のために、一定時間待つ時間。

大体ここが支配的要因&デフォルトだと大きすぎることもあるので、ネットワークに合わせてチューニングする必要がある。

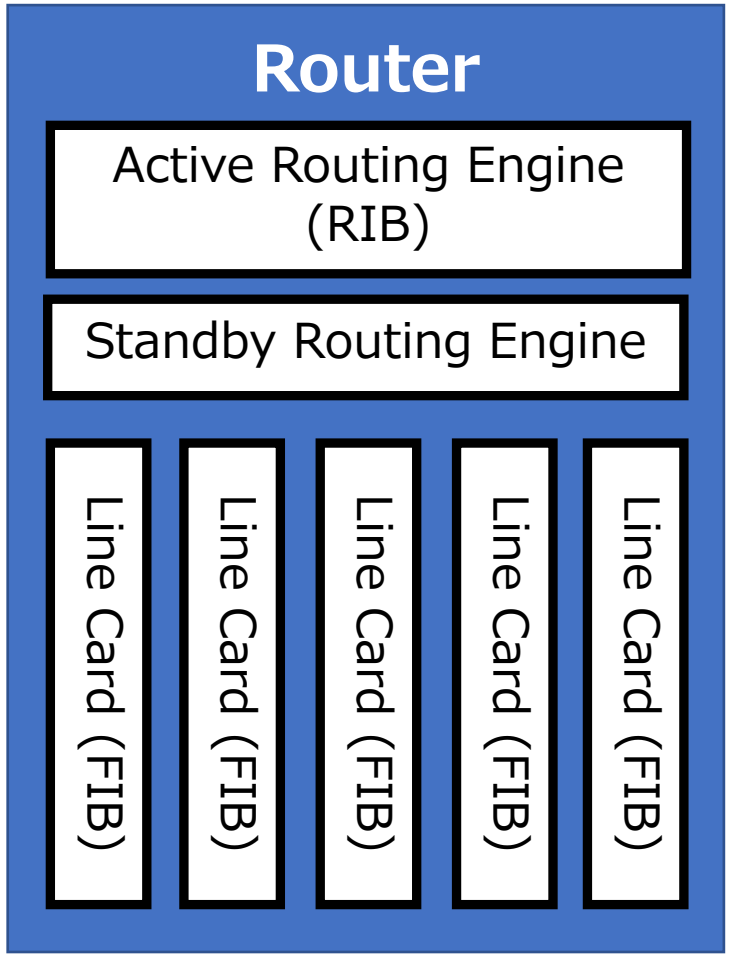
障害対策：リンク障害発生時の断時間



更新されたネットワークにおける、SPF計算を実施。

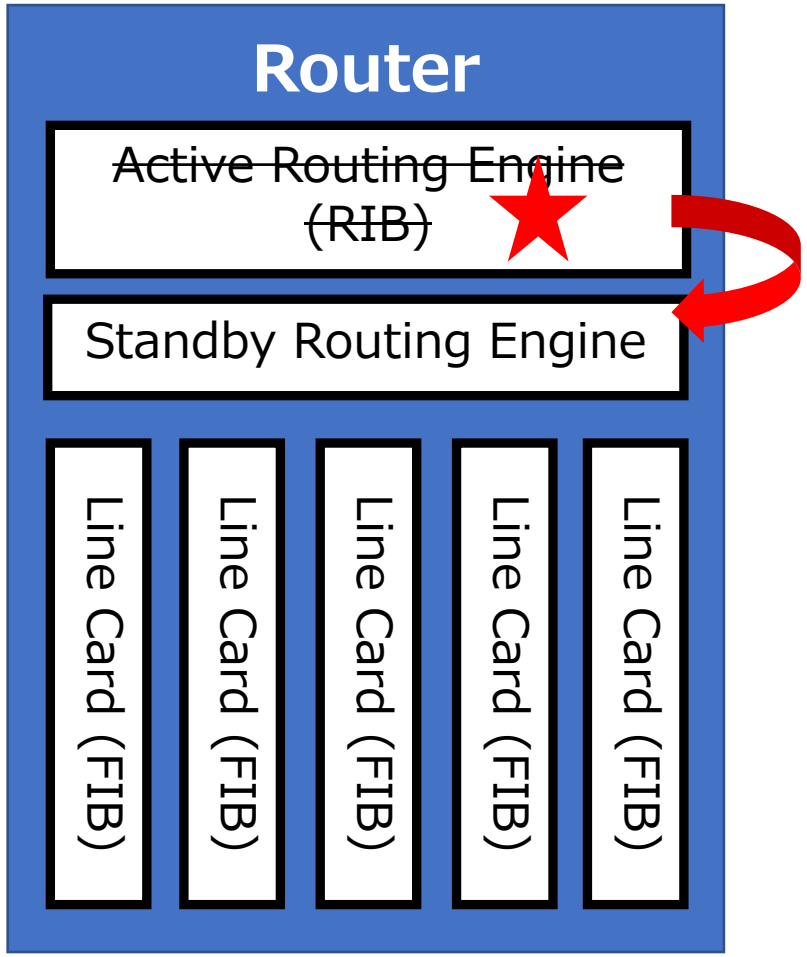
結果をRIB/FIBに反映し、ネットワークの切替を行う。

障害対策：NSF+GR / NSR for IGP

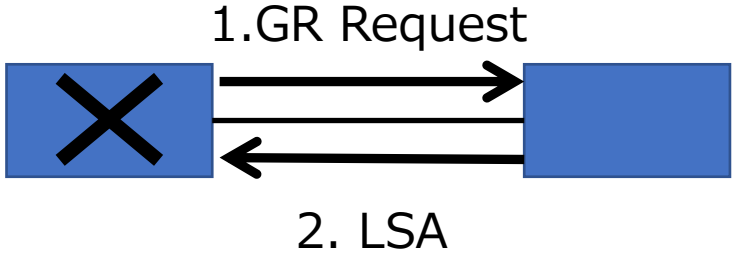


- シャーシ型ルーターだと、経路計算などを実施する頭脳の部分である、Routing Engine(RE)が冗長化されている場合が多い
- ActiveなRouting Engineが何らかの原因で落ちた時に、トラフィック断を防ぐ検討を行う必要がある
- 候補技術
  - Non Stop Forwarding + Graceful Restart
  - Non Stop Routing

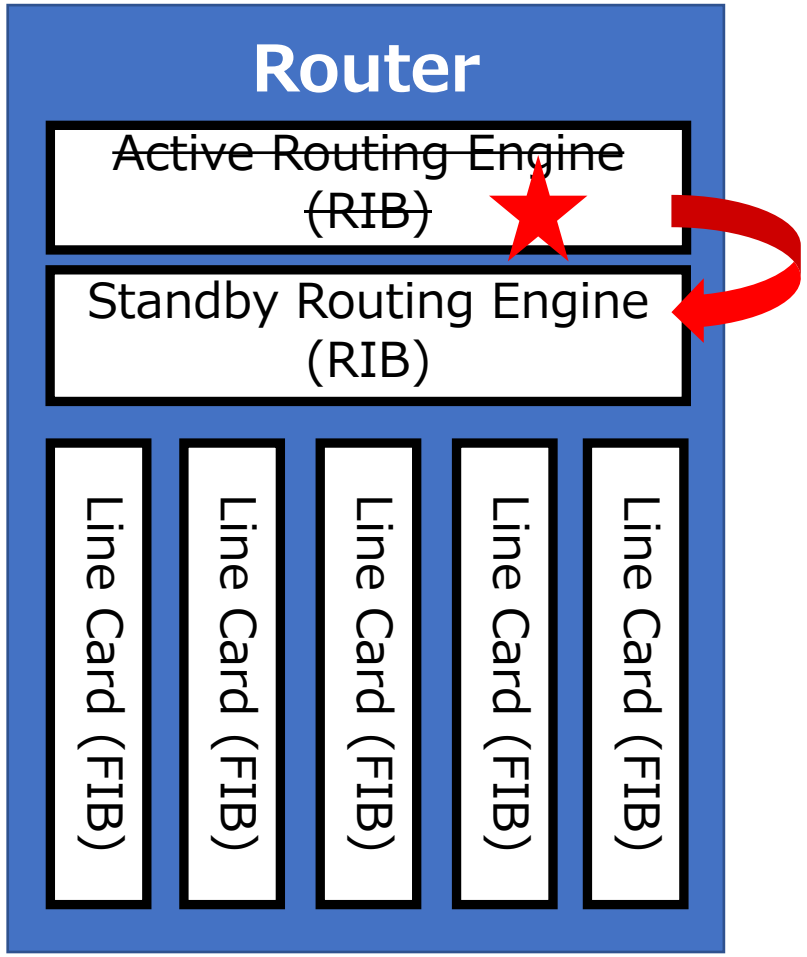
障害対策：**NSF+GR** / NSR for IGP



- Non Stop Forwarding + Graceful Restart
  - ActiveなREが死んだ時に・・・
    - RIBはREからなくなる
    - **FIBはそのままでパケット転送を維持 (NSF)**
  - IGPのRIBをどのように再構成するか？
    - **隣接ルータへ、REが死んだことを通知し、リンクステート情報を再送信してもらう (GR)**
    - RFCで手順は定義されているが、某ベンダーの独自実装などもあるので、インオペ試験など要確認



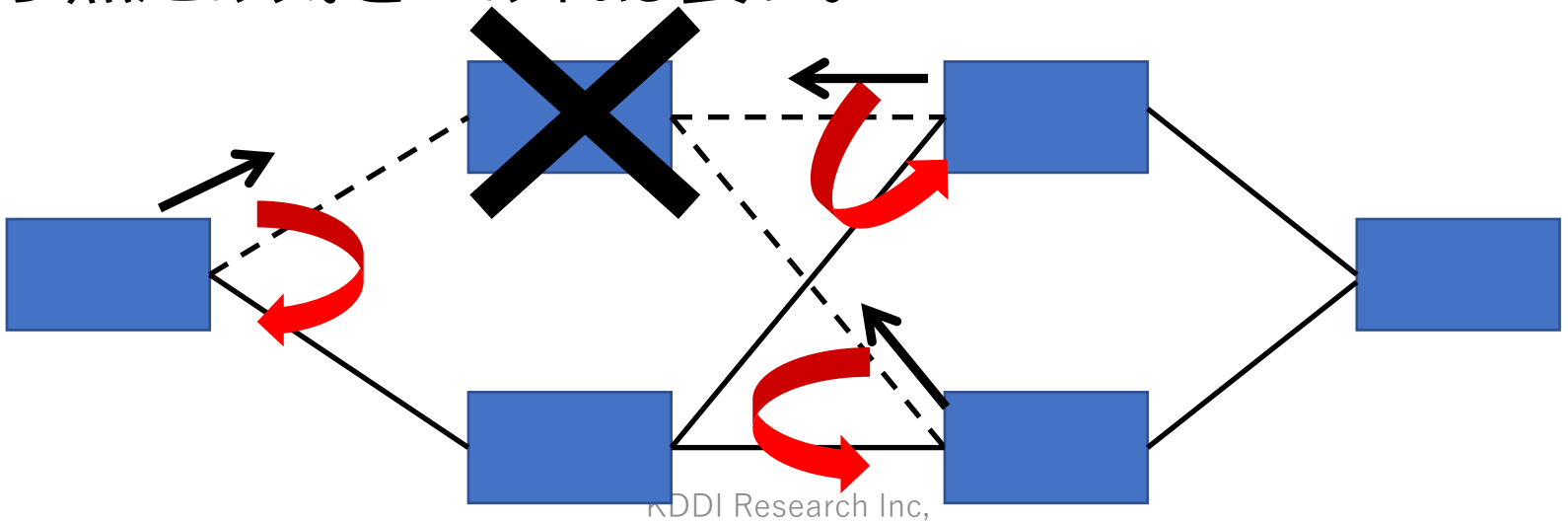
障害対策：NSF+GR / **NSR** for IGP



- Non Stop Routing
  - 通常時動作
    - Standby REにIGPのRIBやNeighbor情報を常に同期していく
  - ActiveなREが死んだ時に・・・
    - 同期していたRIB情報・Neighbor情報を利用して動作を継続
    - GRと比較して、隣接ルータとの協調動作は不要なので、インターオペラビリティを気にしなくていい！
    - ただ、実際のネットワークの経路数/Neighbor数を想定してNSRが動作するかは要検証。保険のためにNSF/GRを有効にするのもよい

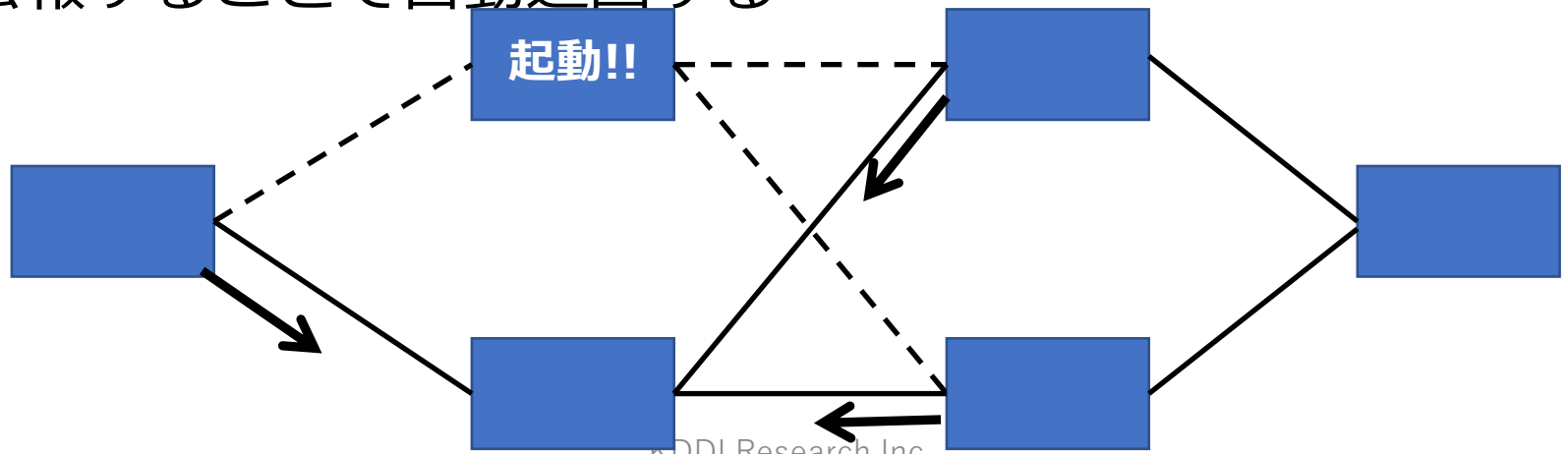
障害対策：ノード障害 (1/2)

- ルーター自体が何らかの原因でダウン/再起動した場合のIGPについて気をつけること
- ノードダウン時
  - IGPの観点では、基本的には先のリンク障害と変わらない。複数のリンク障害が同時に発生するという状況であるため。
  - ネットワークの設計として、Single Point of Failureを作らないという点だけ気をつければ良い。



障害対策：ノード障害 (2/2)

- ノード起動時
  - ノード起動時はルーター自身が不安定な状態になっているため、一定期間トラフィックを自動迂回したい
    - インターネットフルルートをBGPで受信し、FIBに落とすのに数分オーダーでかかることも・・・
  - **Max-metric on start-up**
    - ノード起動時、自身がIGPで広報するMetric(Cost)値を最大にして広報することで自動迂回する



まとめ

1. IGP基本事項おさらい

2. IGP経路広報ポリシー

3. IGP エリアデザインとBGPとの関係性について

4. 障害時に早く切り替えるために