2020/07/09@Online
Internet Weekショーケース

ISPにおける経路設計 (IGP・BGP)

KDDI総合研究所
宮坂拓也
ta-miyasaka@kddi-research.jp

はじめに (1/2)

- 本資料は**ISPにおけるIGP/BGPの基本的設計**について共有するものです
  - IGPについては、本資料では**OSPF**を例にして説明します

- OSPF, IS-IS, BGPの**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

はじめに (2/2)

- 本資料はInternet Week 2019で開催された以下プログラムを40分に濃縮してお送りするバージョンです
- 詳細情報を知りたい方は以下発表資料を参照ください

ISPにおける経路設計(IGP)	https://www.nic.ad.jp/ja/materials/iw/2019/proceedings/s07/s7-miyasaka-2.pdf
BGP設計:前半	https://www.nic.ad.jp/ja/materials/iw/2019/proceedings/s08/s8-matsuzaki.pdf
BGP設計:後半	https://www.nic.ad.jp/ja/materials/iw/2019/proceedings/s08/s8-hirai.pdf

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

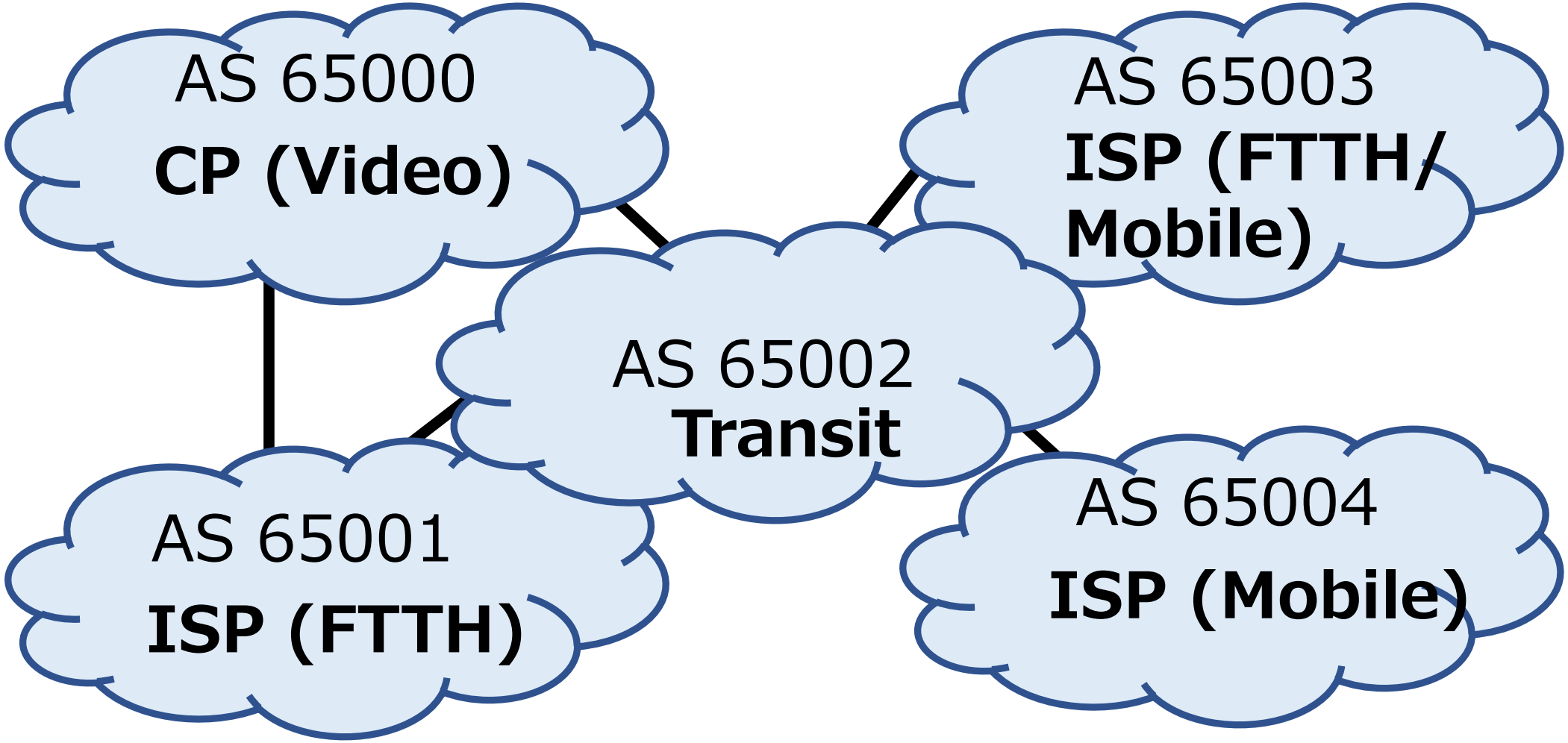
Agenda

1. インターネットとは？
2. ISPにおける経路設計デザイン概論
3. ISPにおけるプロトコル設計：BGP編
4. ISPにおけるプロトコル設計：IGP編

Agenda

1. インターネットとは？
2. ISPにおける経路設計デザイン概論
3. ISPにおけるプロトコル設計：BGP編
4. ISPにおけるプロトコル設計：IGP編

InternetとAS



InternetとAS

- Internet
  - 複数のネットワークを相互接続して構成される全世界規模の大規模ネットワーク

- Autonomous System (AS)
  - 統一のルーティングポリシのもとで運用されているIPプレフィックスの集まり[1]
    - Internetを構成する単一単位とも言える
  - ASの識別子としてAS番号が割り当てられる
    - 例：KDDI=2516

[1] https://www.nic.ad.jp/ja/materials/iw/2019/proceedings/s08/s8-matsuzaki.pdf

インターネットと経路広報

サーバーIPアドレス
10.0.1.0/24
2001:db8:1::/48

経路広報

AS 65000
CP (Video)

AS 65001
ISP (FTTH)

AS 65002
Transit

AS 65003
ISP (FTTH/Mobile)

AS 65004
ISP (Mobile)

インターネットと経路広報

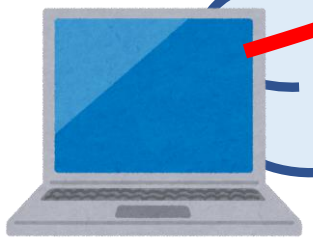
AS 65000
CP (Video)

AS 65003
ISP (FTTH/ Mobile)

AS 65002
Transit

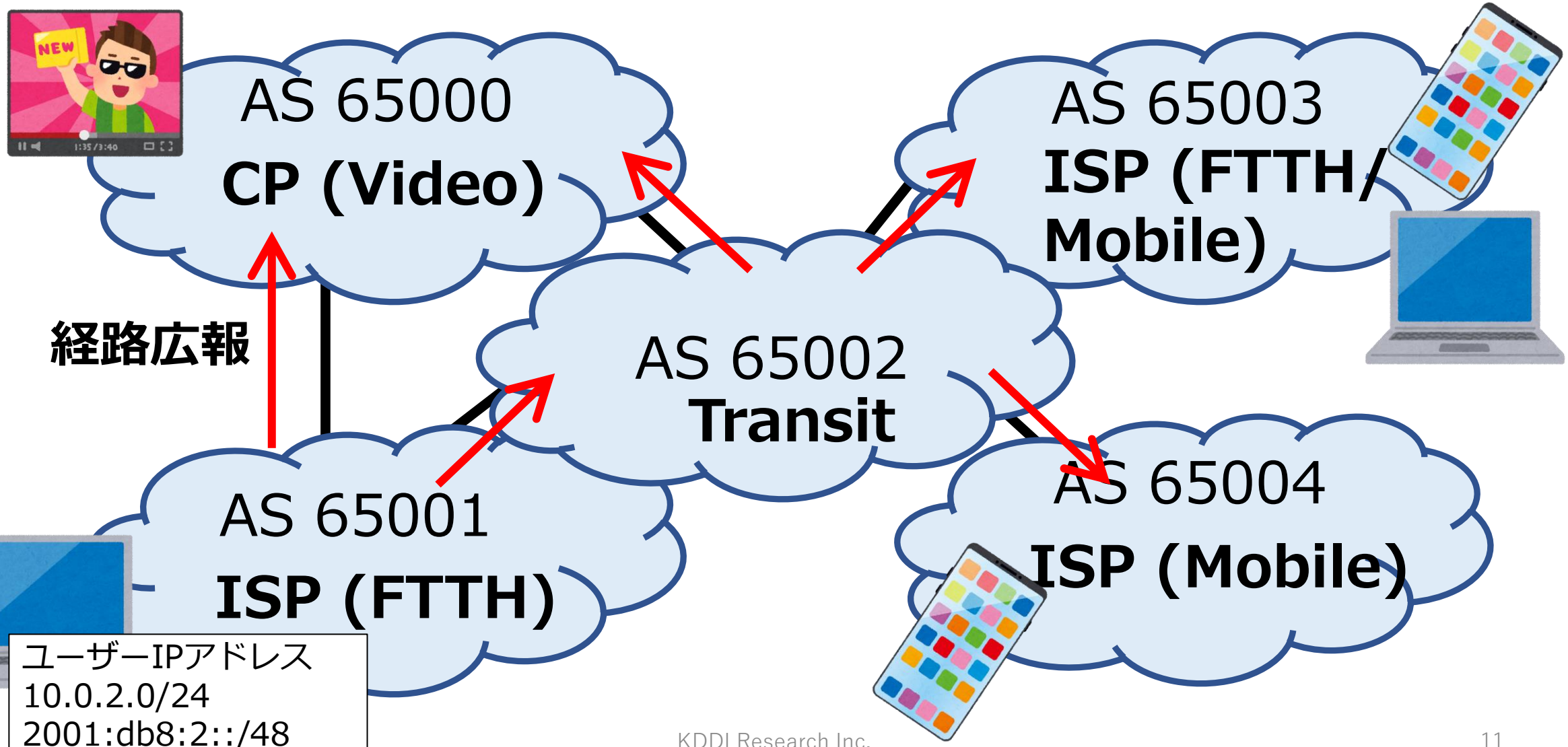
リクエスト
To 2001:db8:1::1
From 2001:db8:2::1

AS 65001
ISP (FTTH)

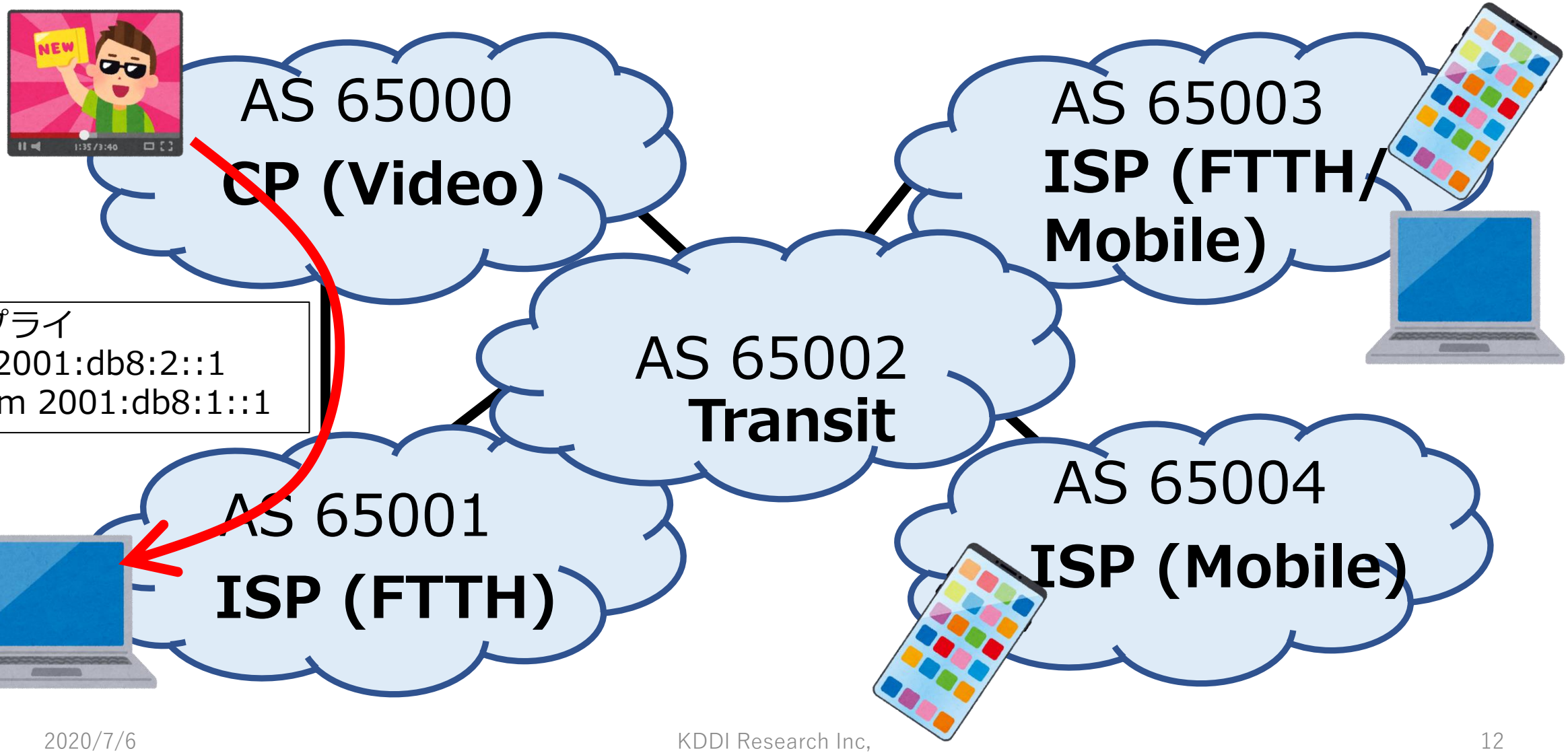


AS 65004
ISP (Mobile)

インターネットと経路広報



インターネットと経路広報



インターネットと経路広報

- インターネットにおける経路広報
  - **経路生成**：各ASが自身が所有するIPアドレス情報を他のASに通知する (ISP、CP)
  - **経路転送**：また、必要があれば、他のASから受信したIPアドレス情報を、さらに別のASに転送する (トランジット)
    - この経路広報に用いられる通信プロトコルが**BGP**
    - インターネット上の全ASで生成された経路全体を**インターネットフルルート**と呼ぶ
      - 現在(2020年7月) IPv4が約80万経路、IPv6が約9万経路である

本発表対象

AS 65000

CP (Vid

サーバーIPアドレス
10.0.1.0/24
2001:db8:1::/48

経路広報

AS 65003

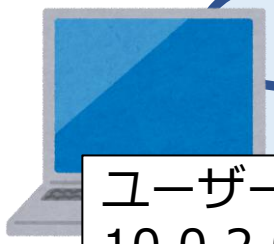
本発表では、ISPにおける経路設計についてBGP/IGPそれぞれについて紹介する

AS 65001

ISP (FTTH)

AS 65004

ISP (Mobile)

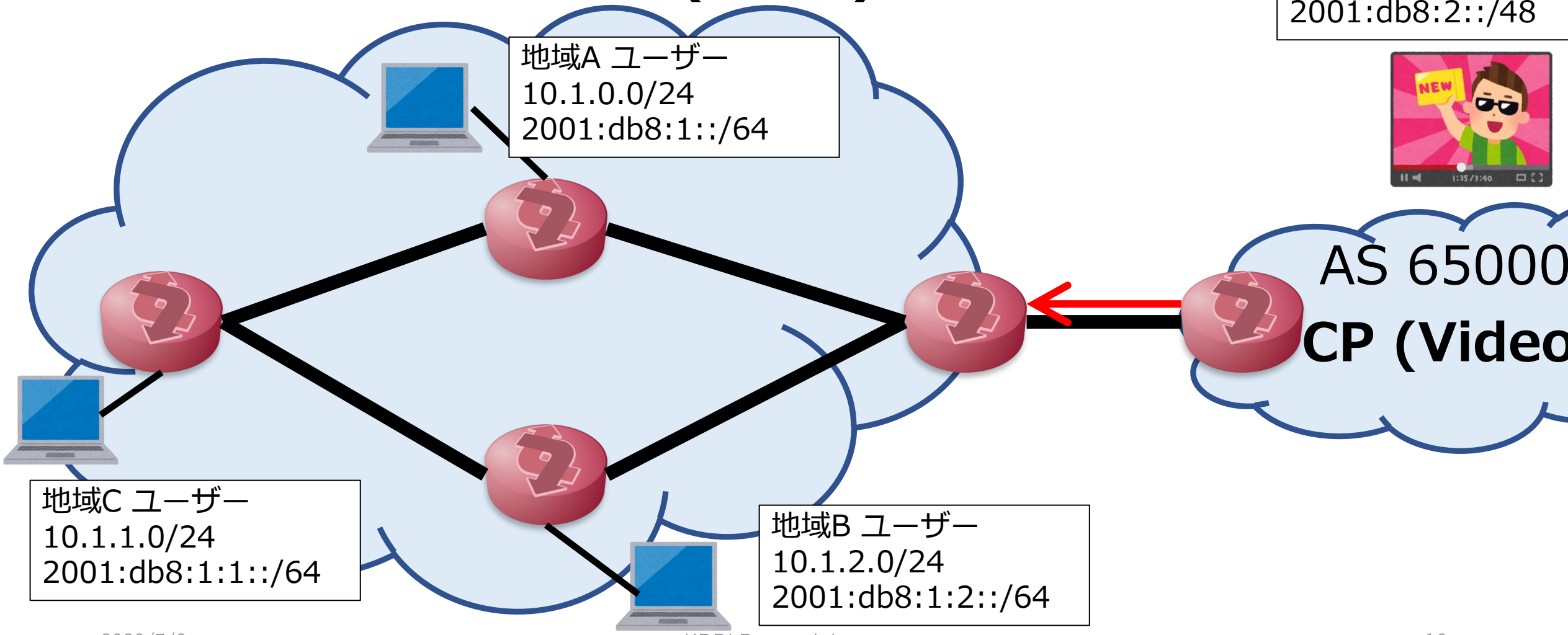


ユーザーIPアドレス
10.0.2.0/24
2001:db8:2::/48

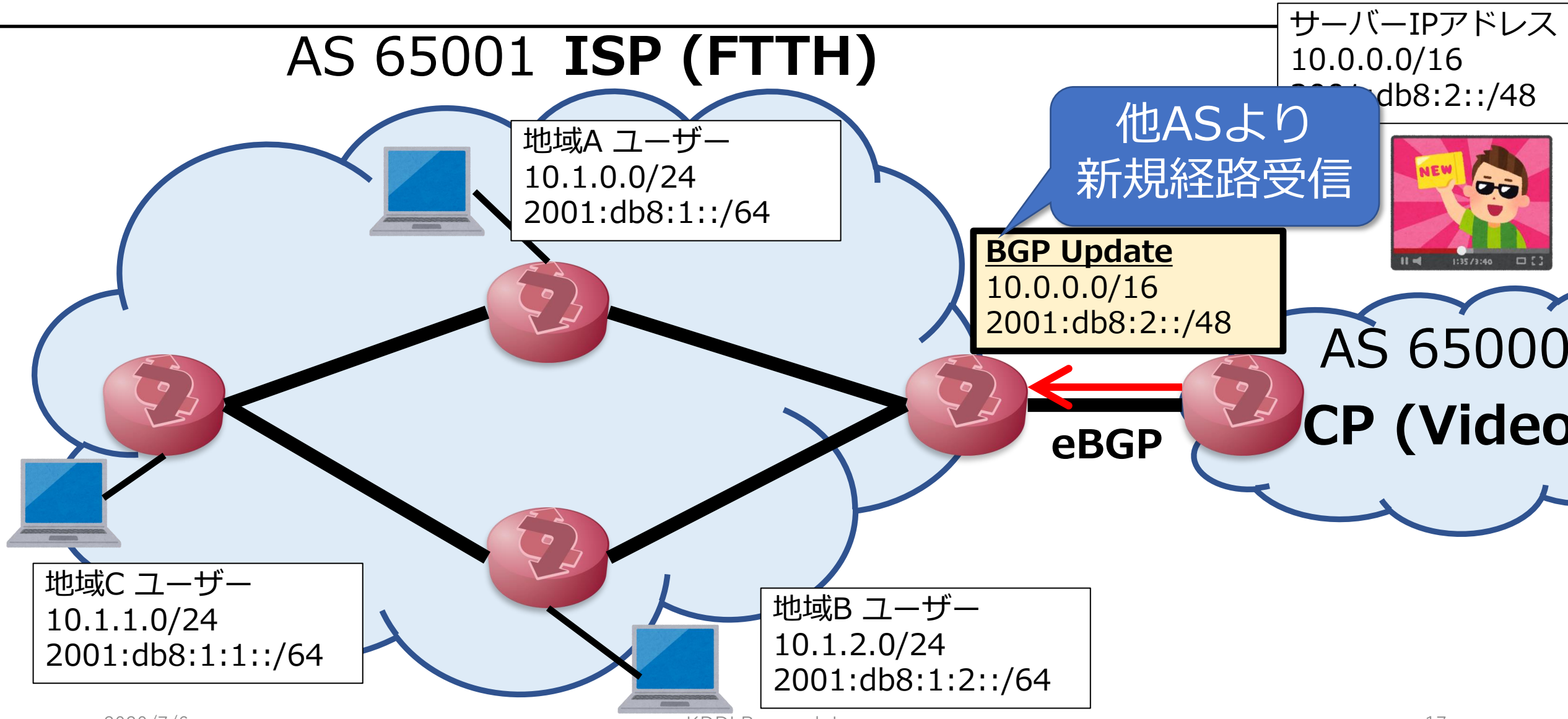
Agenda

1. インターネットとは？
2. ISPにおける経路設計デザイン概論
3. ISPにおけるプロトコル設計：BGP編
4. ISPにおけるプロトコル設計：IGP編

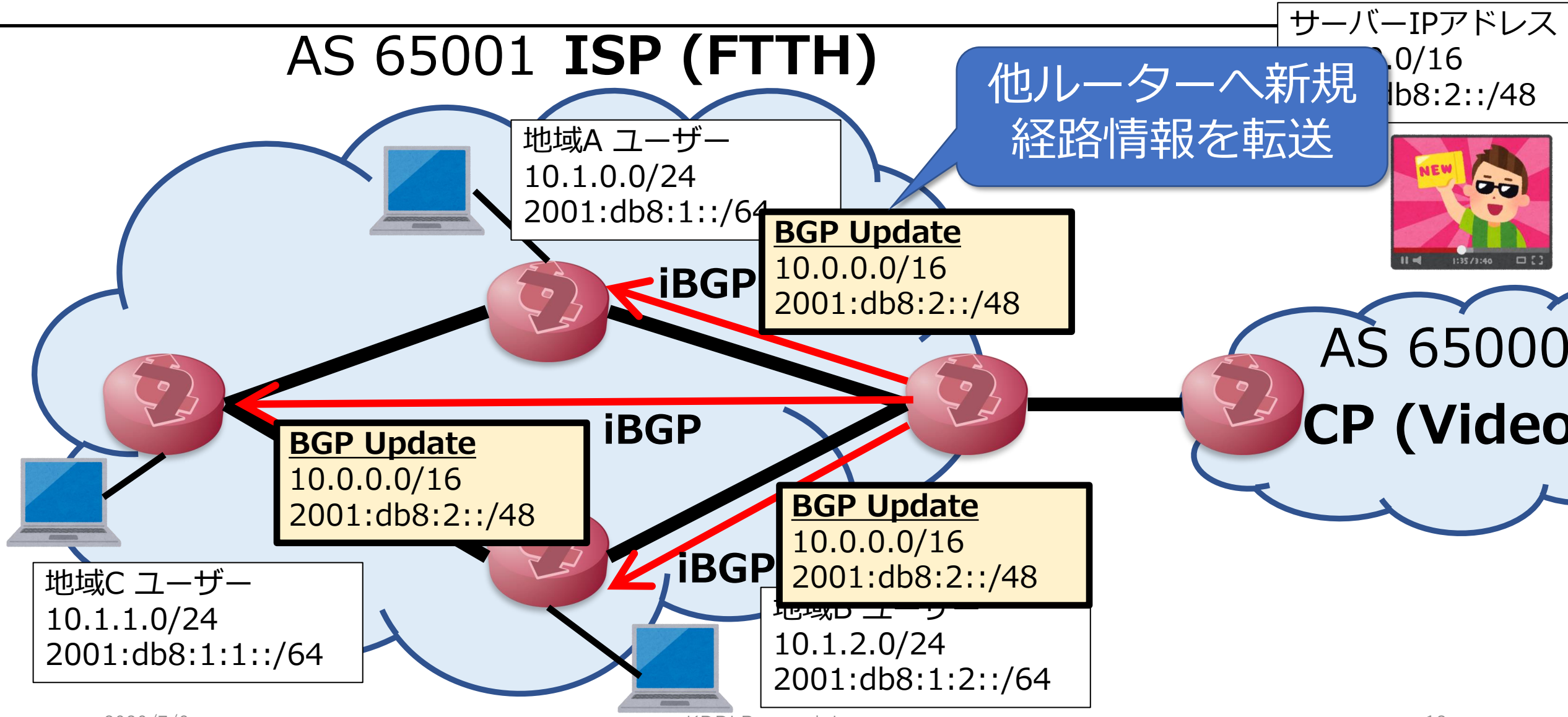
経路設計


AS 65001 ISP (FTTH)
サーバーIPアドレス
10.0.0.0/16
2001:db8:2::/48
地域A ユーザー
10.1.0.0/24
2001:db8:1::/64
NEW
AS 65000
CP (Video
地域C ユーザー
10.1.1.0/24
2001:db8:1:1::/64
地域B ユーザー
10.1.2.0/24
2001:db8:1:2::/64

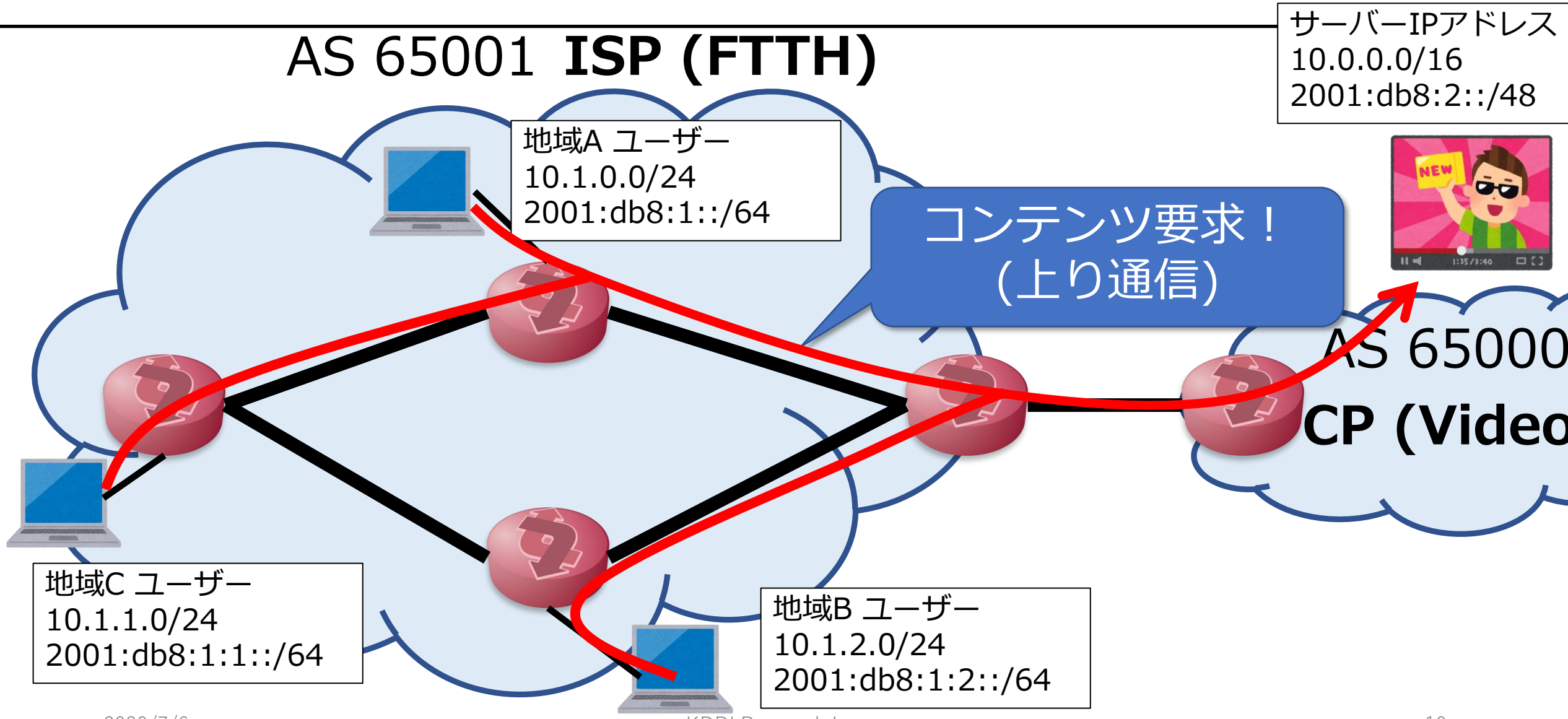
経路設計：インターネット経路受信編



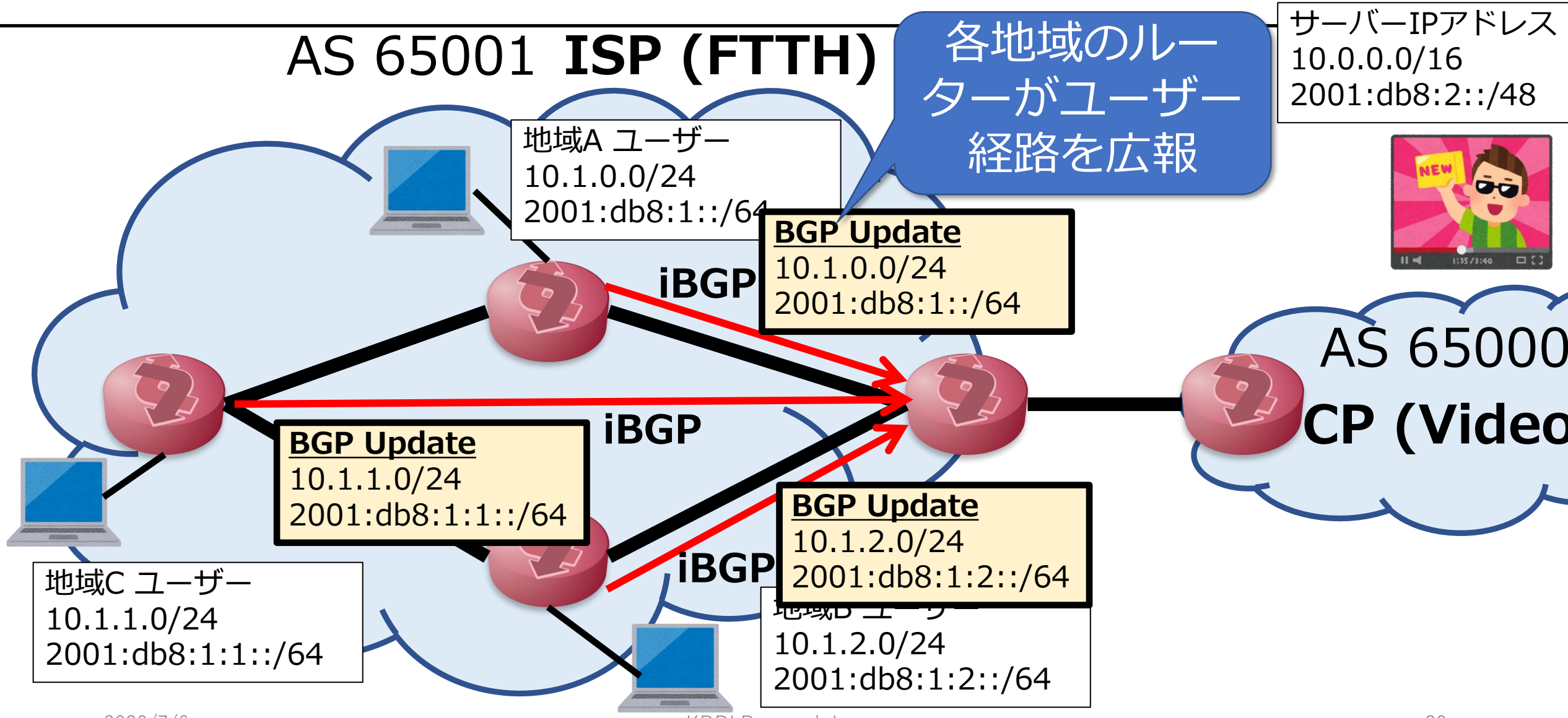
経路設計：インターネット経路受信編



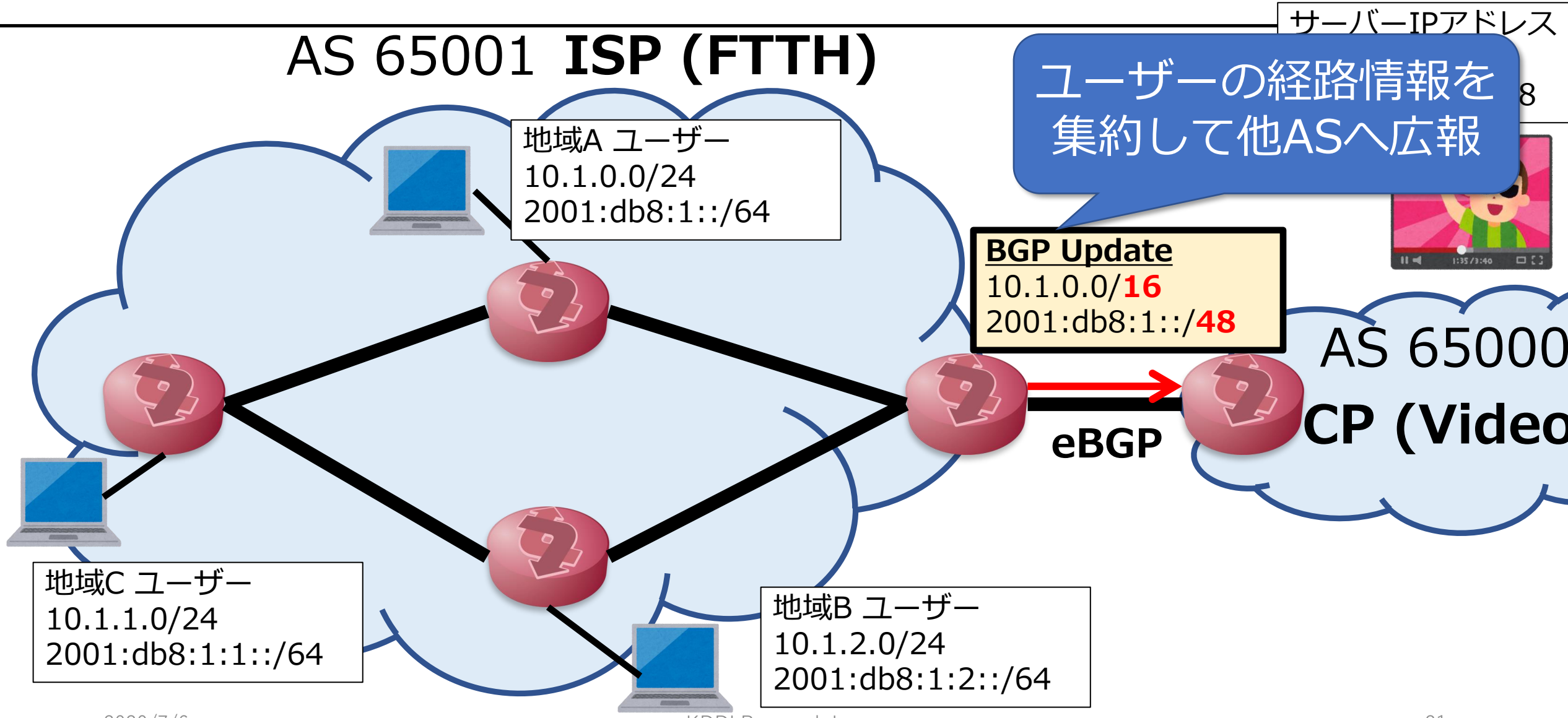
経路設計：インターネット経路受信編



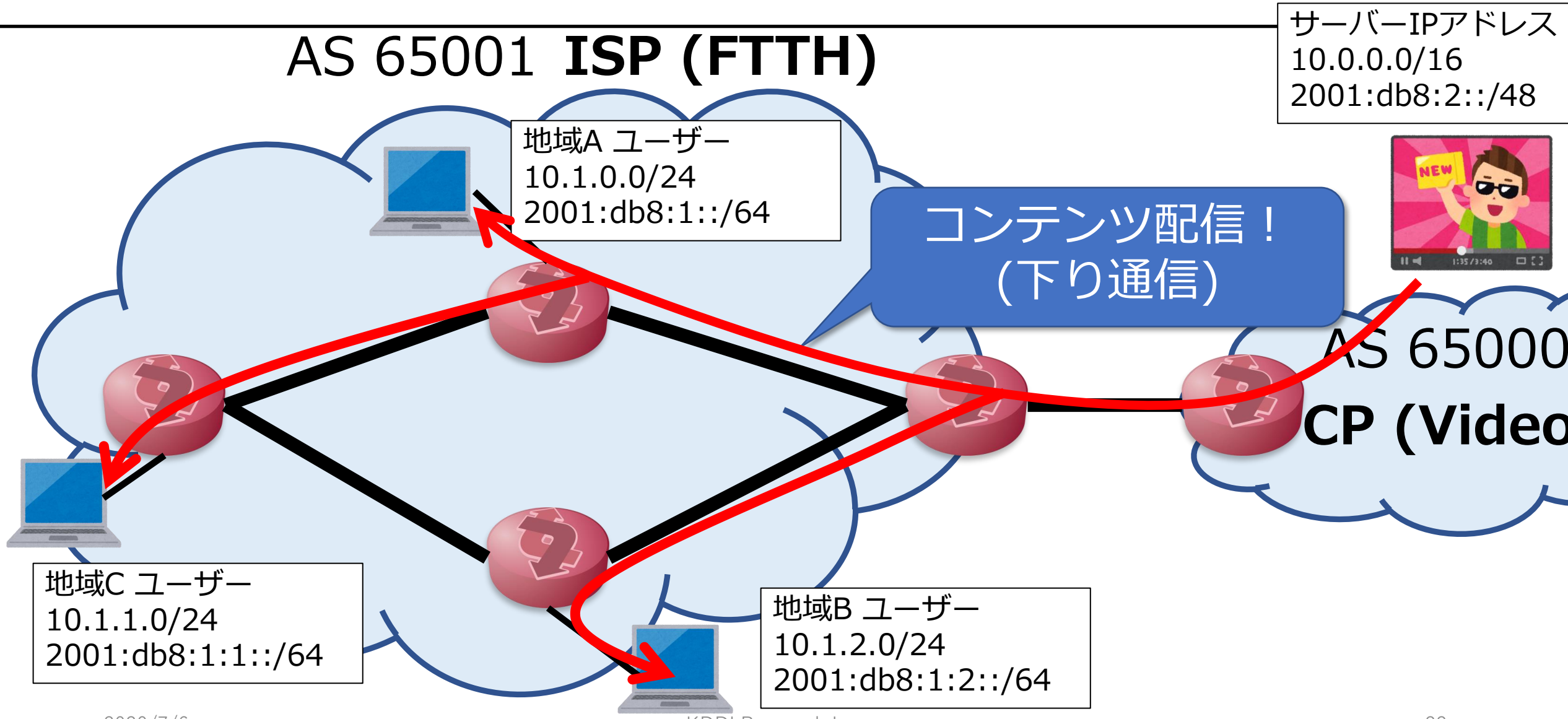
経路設計：ユーザー経路広報編



経路設計：インターネット経路受信編



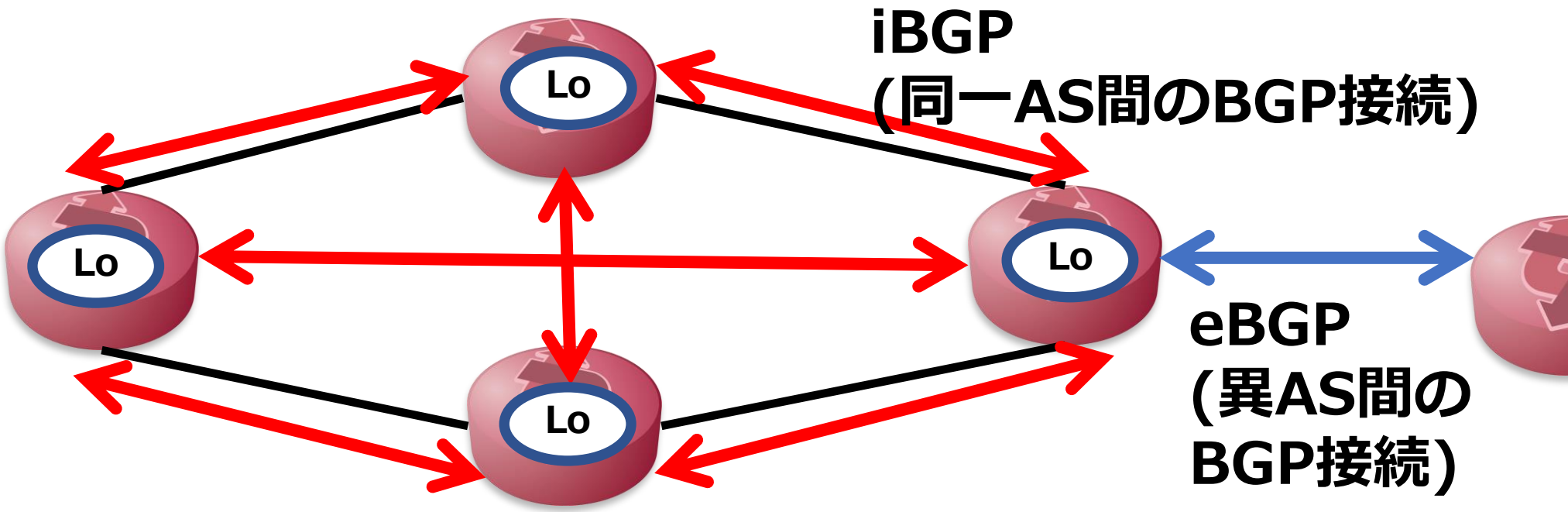
経路設計：インターネット経路受信編



Agenda

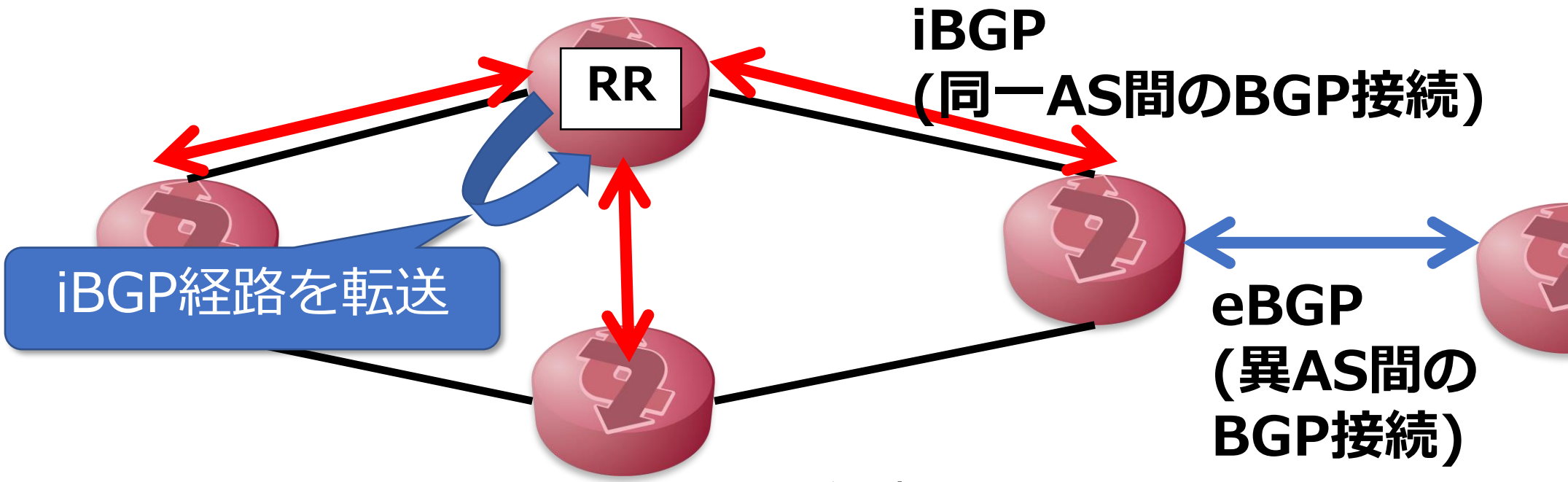
1. インターネットとは？
2. ISPにおける経路設計デザイン概論
3. ISPにおけるプロトコル設計：BGP編
4. ISPにおけるプロトコル設計：IGP編

iBGP Neighbor Design



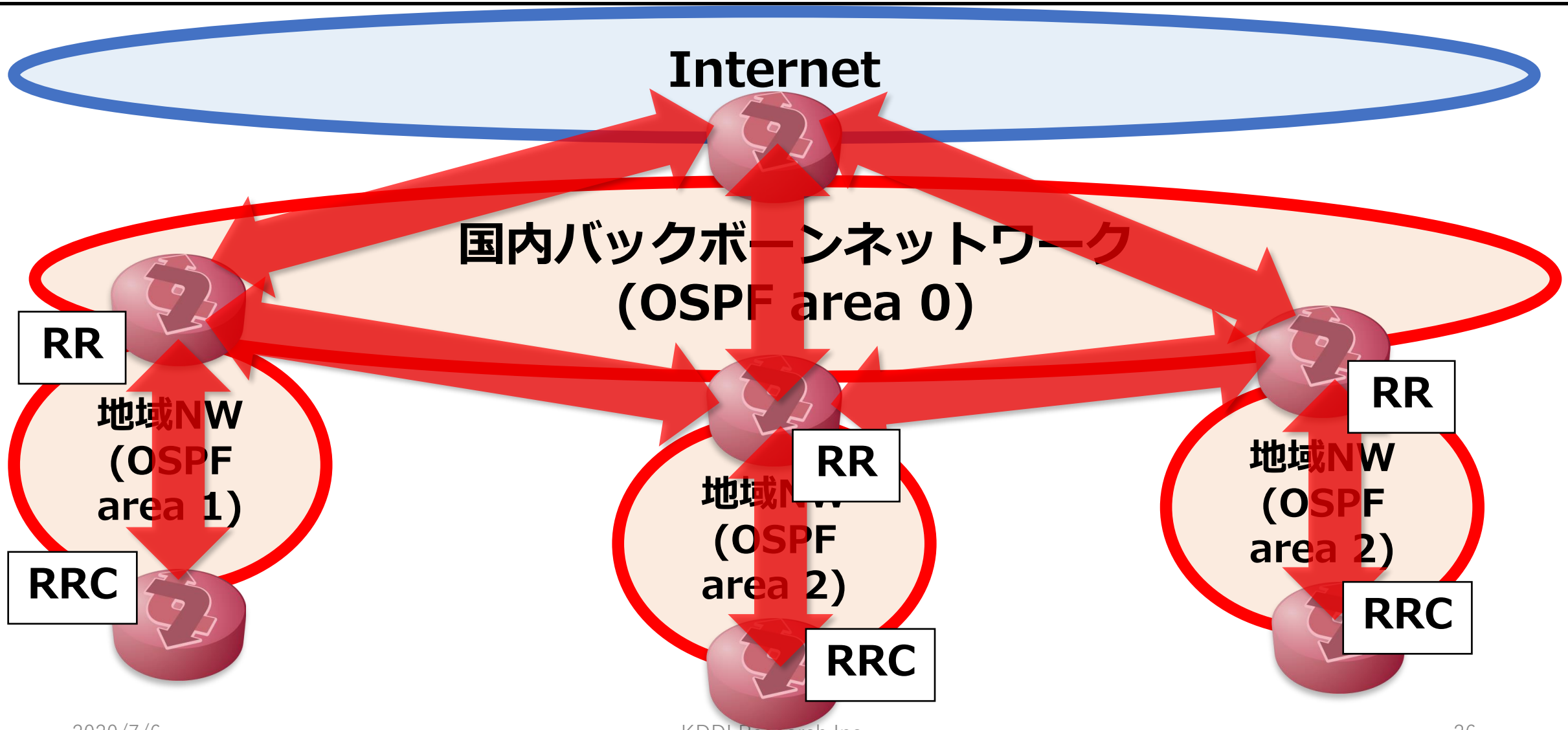
- iBGPは**全ルーター間でフルメッシュに接続**する必要あり
  - 各ルーターのLoopbackアドレスで接続する
  - iBGPには、「あるiBGP peerから受信した経路を、他のiBGP peerに転送してはいけない」というルールがあるため
  - ルーターの数が増えた時にiBGPネイバー数が肥大化する ☹

iBGP Neighbor Design

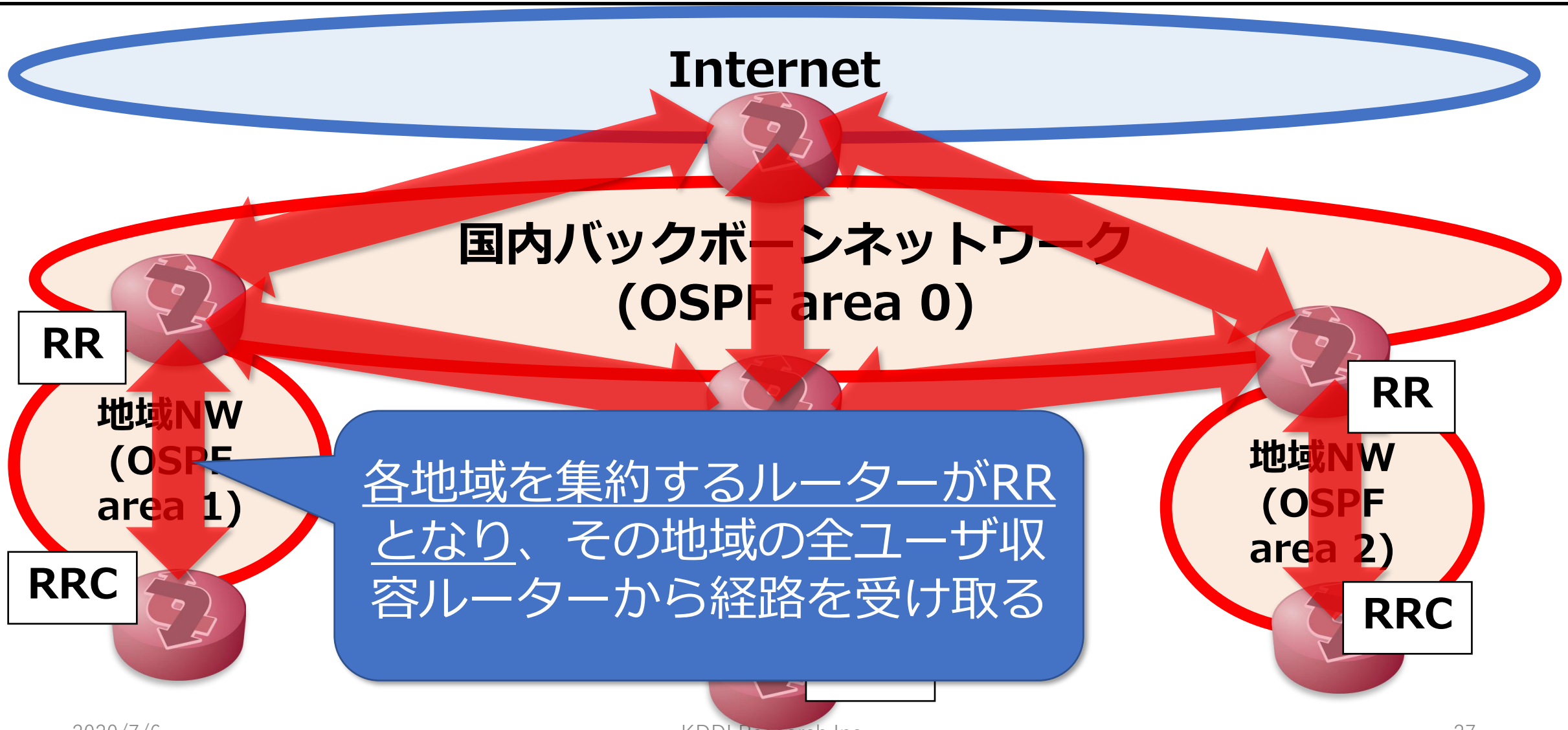


- **Route Reflector(RR)**を用いることで、網内のルーターはRRのみとiBGP peerを接続すればよくなるが、以下注意も必要
  - RRを冗長化していないと、RR障害時に網全体が通信断
  - 設定によっては、最適なRoutingにならないこともあるので要注意
- その他として、**Confederation**も使える

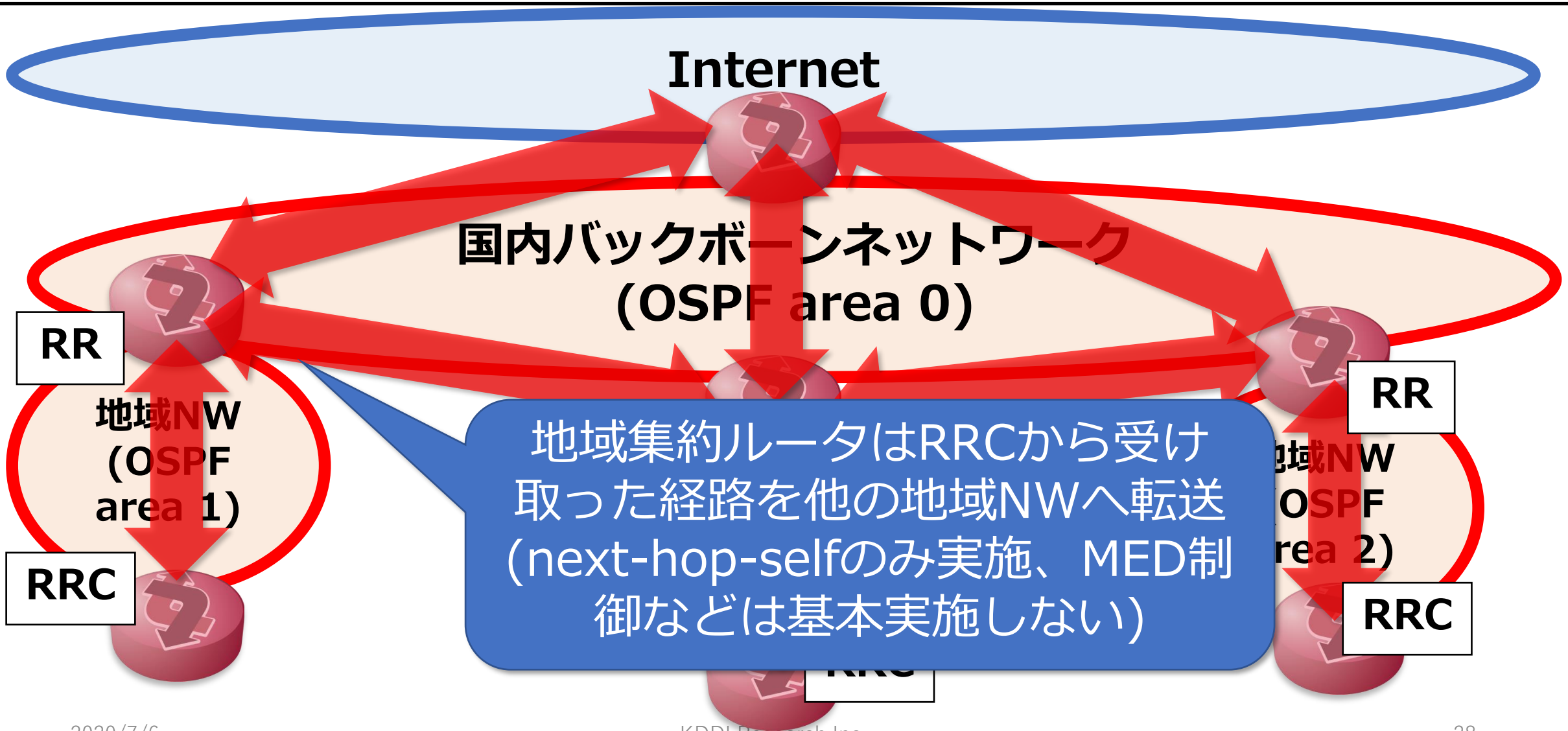
とあるISPにおける設計例



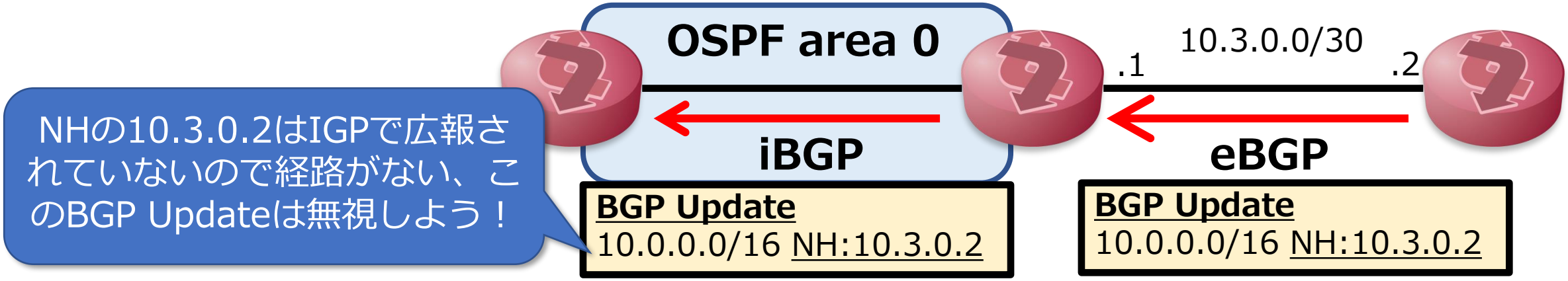
とあるISPにおける設計例



とあるISPにおける設計例



Next-hopに気をつけろ！



- BGPの経路広報時には、送信すべきNext-hopのIPアドレスが付与されている(NEXT_HOP attribute)
- BGPの特性として、受信側ルーターでその**Next-hopのIPアドレスの経路情報がないと経路を無視してしまう**
- 上例の解決策？
  - Next-hop-self
  - eBGPで利用しているリンク情報をIGPで広報

Next-hop-selfする？しない？

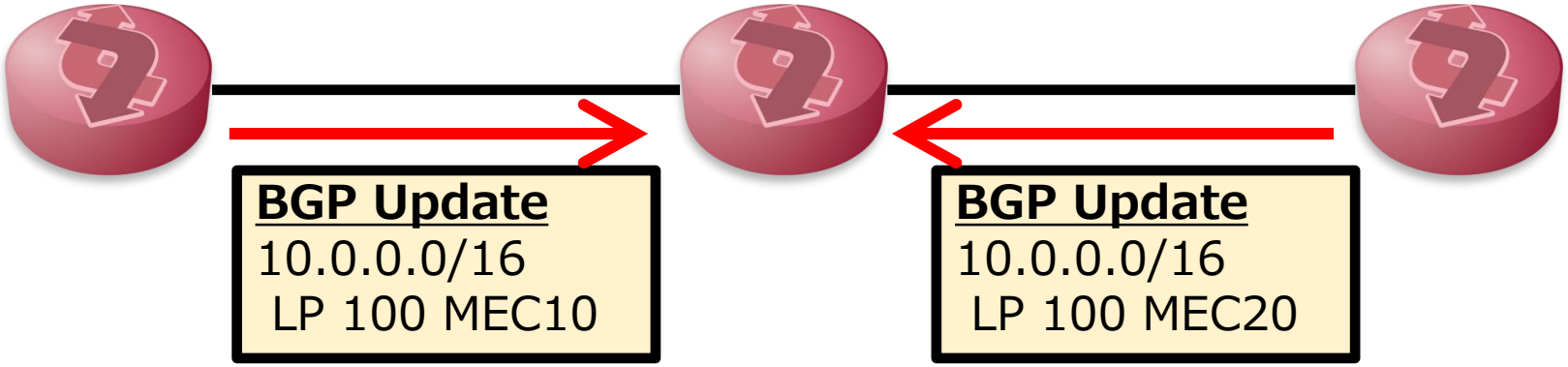
- それだけで、1時間は議論できる内容なので、設計時にはよく注意していきましょう
  - 参考：http://irs.ietf.to/wiki.cgi?page=IRS28

時間	プログラム	発表者	発表資料
	開場諸注意	司会 調整中	
18:00-	あやしい経路の、お話(仮)	IIJ 松崎さん	{{}}
18:30-	iBGPでのBGPネクストホップの話(next-hop selfするしない話)	(株)FORNEXT 篠宮さん	irs28_shino_bgp-nexthop_01.pdf(1988)
19:10-	JC1005のnext-hop-self	NTTCom 川上さん	https://www.slideshare.net/yuyarin/nexthopself-in-jc1005-irs28
19:20-	告知タイム JPIRRの経路ハイジャック通知、内容増えるってよ	JPNIC 岡田さん	
	告知タイム OsakaPeeringFestival開催のお知らせ	Jstream 佐藤 太一さん	
	告知タイム JANOG42開催のお知らせ	JANOG42実行委員長 松下さん、鈴木さん	
19:30-	クロージング、次回いつごろか相談		
19:50-21:00	懇親会 会場にてビアバッシュ形式	会費：1500円	集金済みの方にはシールをお渡ししますので、胸につけてください。終了時のゴミ収集、ゴミ捨てにご協力ください。

より詳細は・・・

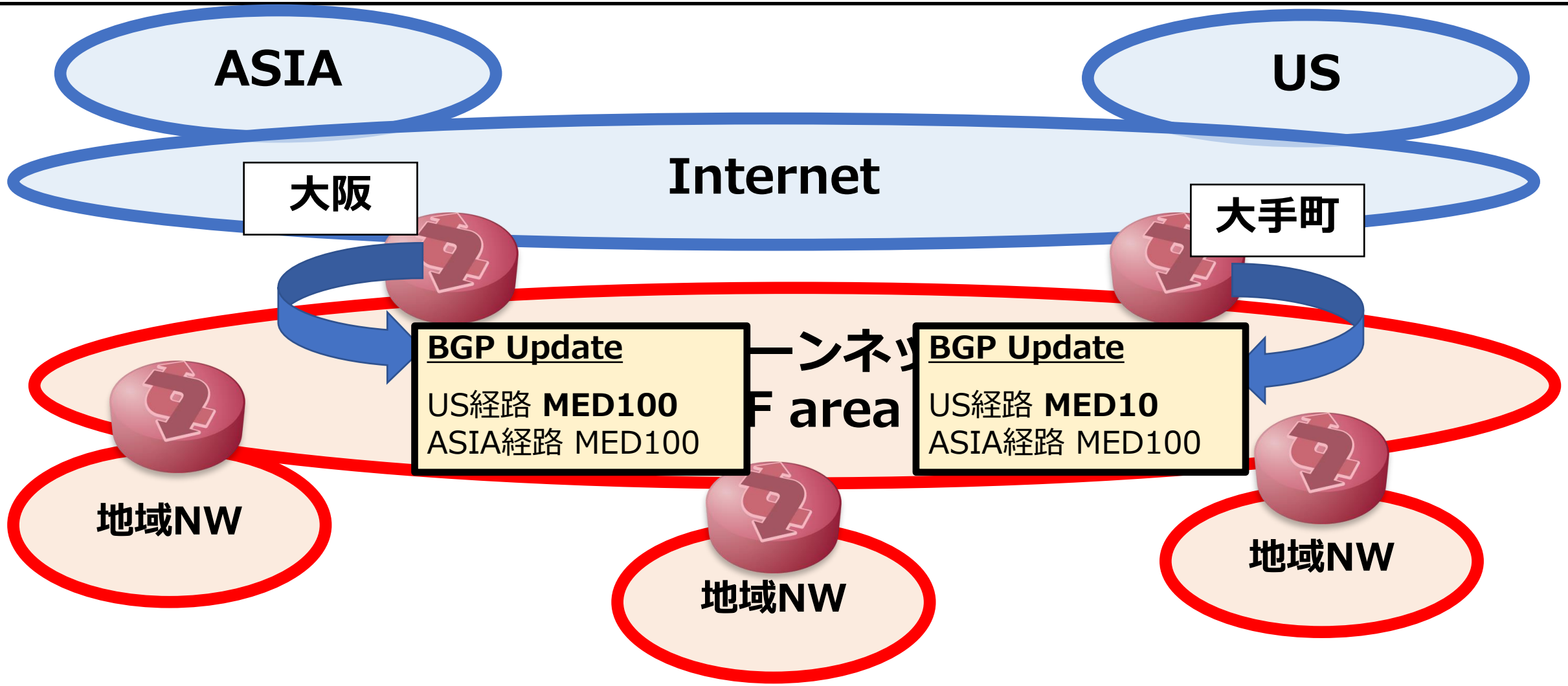
- https://www.nic.ad.jp/ja/materials/iw/2019/proceedings/s08/s8-matsuzaki.pdf

BGP経路制御

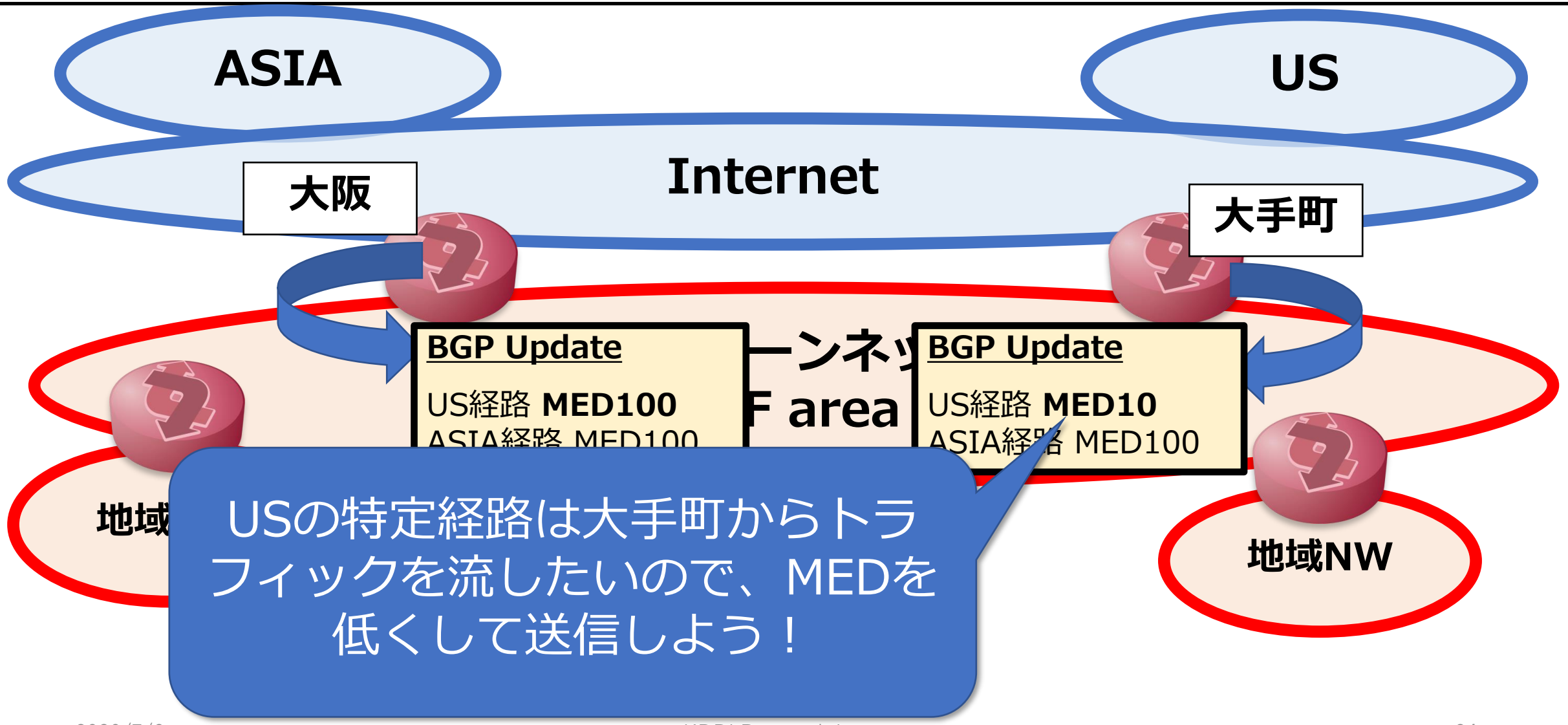


- BGPの最適経路選択アルゴリズムはIGPと比べ複雑、かつベンダー毎に使用が異なることもあり注意が必要
- 複数拠点間から同一prefixが流れて来るケースにおいて、**オペレータが意識して経路制御設計する必要がある**
  - 例：大手町と大阪の２拠点でインターネット経路を受信している
- また、日々の網内トラフィック状況や、ユーザー申告(例：特定のサイト遅いんですけど…)に応じて**日々変更が必要**なことも

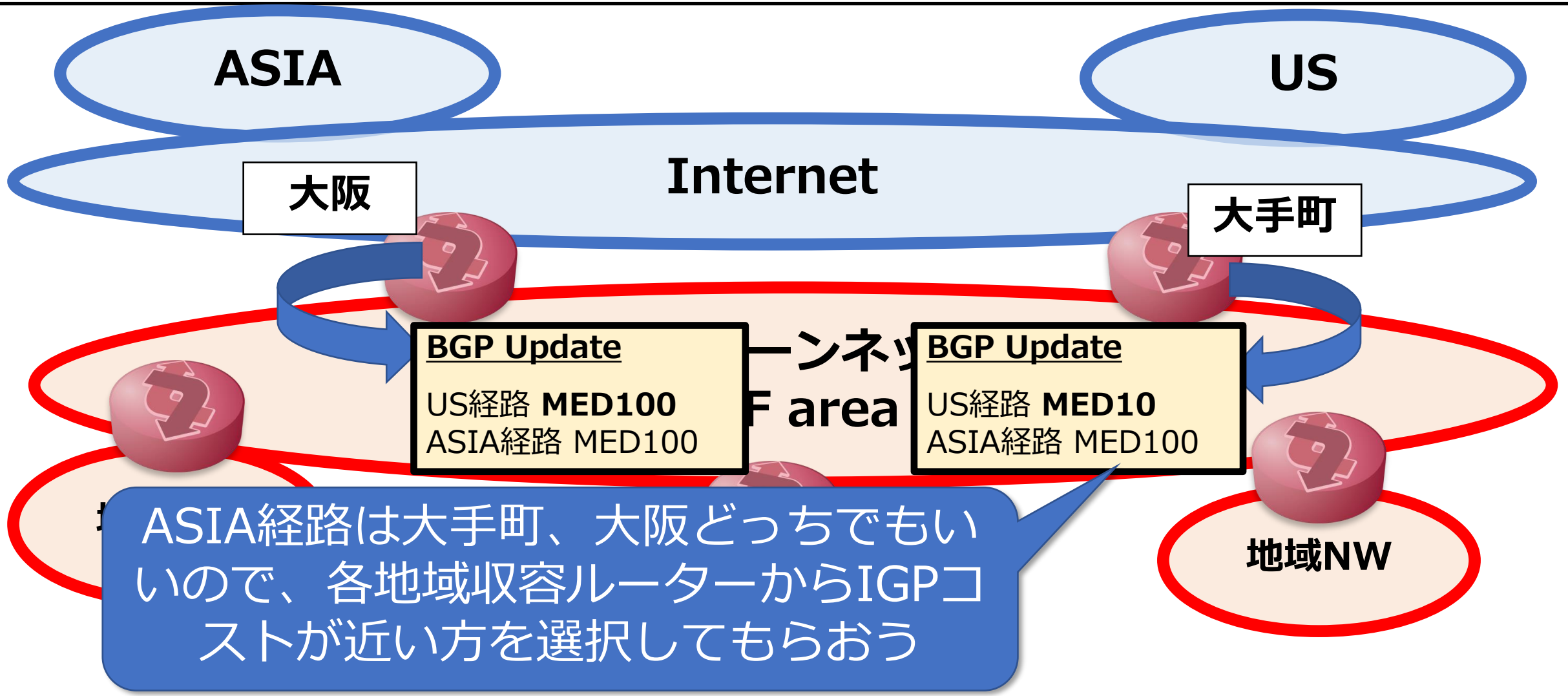
とあるISPにおける設計例



とあるISPにおける設計例



とあるISPにおける設計例



より詳細は・・・

- https://www.nic.ad.jp/ja/materials/iw/2019/proceedings/s08/s8-hirai.pdf

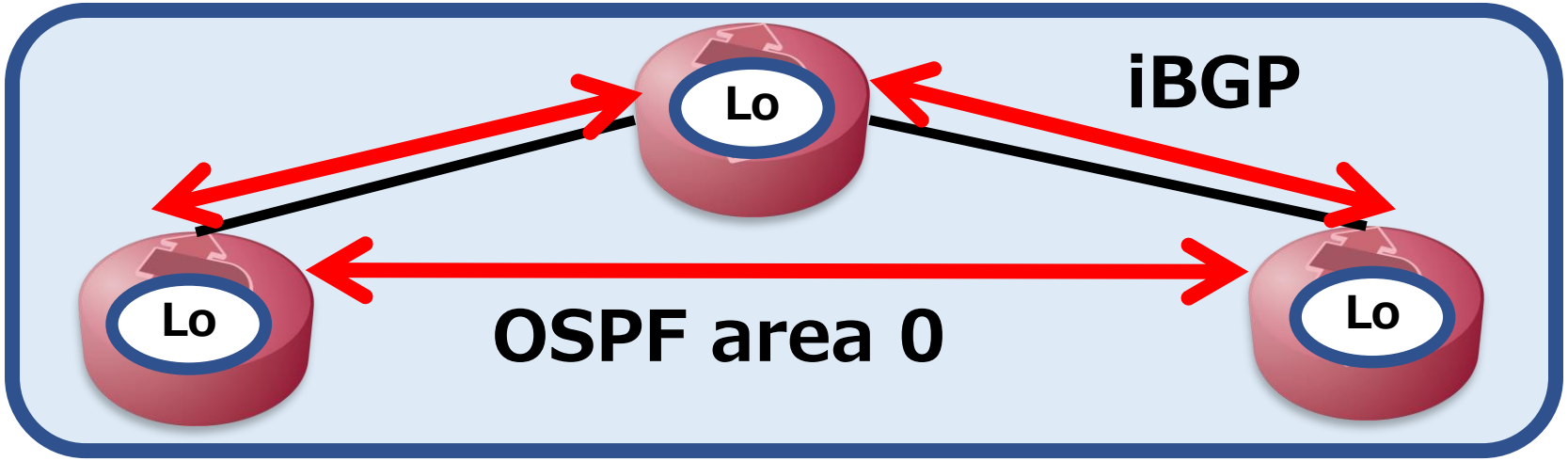
S8 サービスプロバイダ
バックボーン設計入門 後編
BGP設計 後半

Norisuke Hirai
SoftBank Corp./BBIX, Inc.

Agenda

1. インターネットとは？
2. ISPにおける経路設計デザイン概論
3. ISPにおけるプロトコル設計：BGP編
4. ISPにおけるプロトコル設計：IGP編

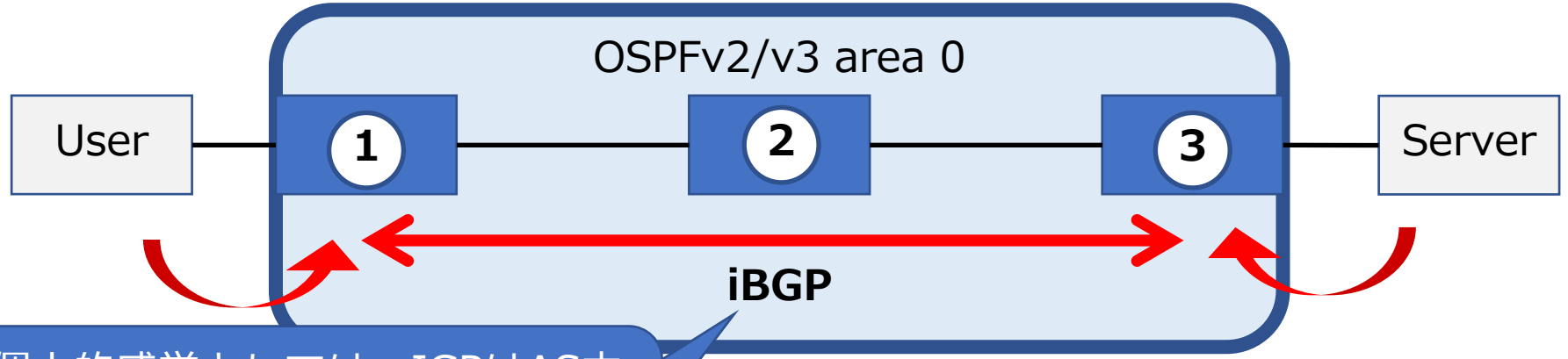
ISPにおいてIGPは何のためにある？



- 基本的には、IGPは「AS内ルーターのiBGP接続をするための経路情報(Loopbackアドレス+リンク情報)を網内に広報するため」にある
- IGP/BGPの基本的な使い分けとしては：
  - BGP：実際にユーザートラフィックが流れる経路情報を交換するもの
  - IGP：BGPの接続のためのルーター経路情報を交換するもの
    - 逆に、IGPにユーザアドレス・サーバアドレスは出来る限りのせないこと！

IGPにどんな経路をのせるのか？

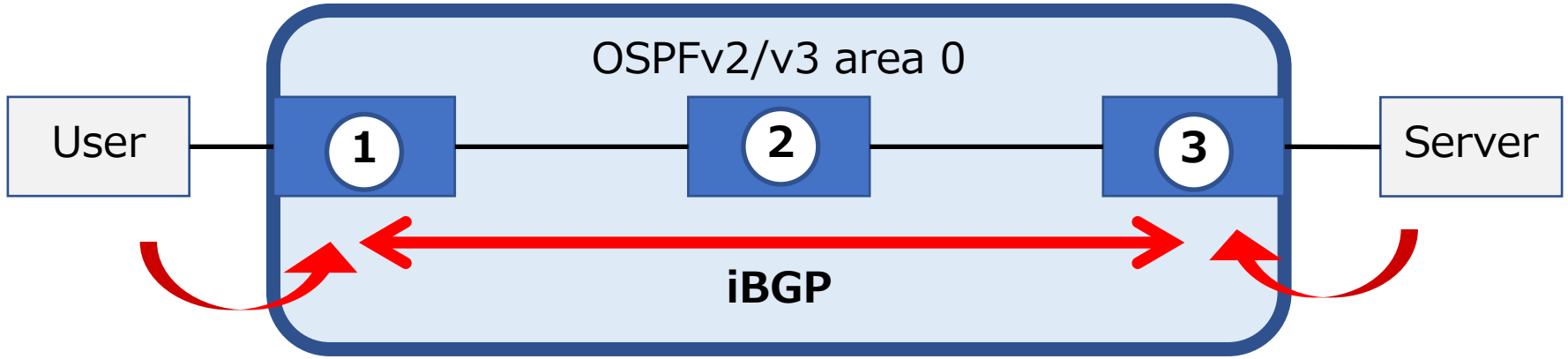
- IGPの基本的役割
  - 各ルーターのインターフェース経路を広報すること
    - Loopbackアドレス
      - iBGPを貼るためのアドレス
    - リンクアドレス
      - iBGPの宛先であるLoopbackアドレスへ導くアドレス



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？

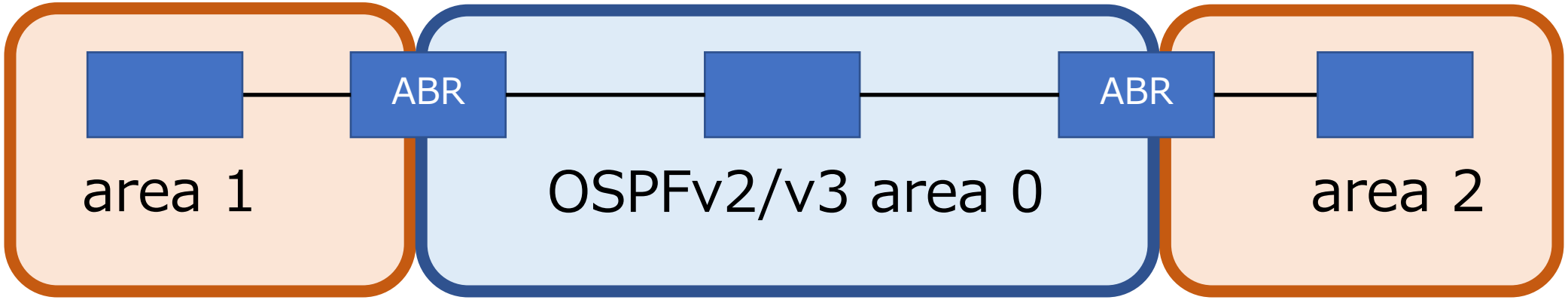
- IGPの基本的役割
  - ユーザー経路やサーバー経路といった、実際にトラフィックがのる経路はIGPにのせない (BGPにのせる)
    - IGPはBGPに比べて計算量が大きく、なるべくIGP経路を最小にし、それ以外のものはBGPで経路広報するデザインが好まれる
    - IGPが分断されている場合、経路再広報(Redistribution)が必要となってしまう



IGP：Area

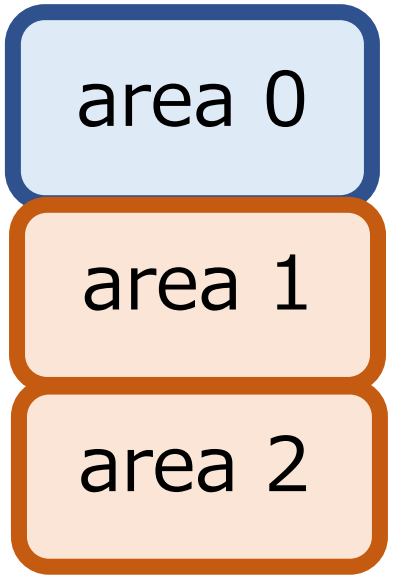
- エリア：ネットワークのグルーピング (RFC2328)
- バックボーンエリア (area 0)
  - すべてのエリアのIGP経路を交換・伝搬させる
- Non-バックボーンエリア (area 1,2…)
  - そのエリアの経路は詳細なリンクステート情報をもらう
  - Area 0から、サマライズされた他のエリアの経路をもらう
  - Area 0へは、そのエリアをサマライズした経路をわたす

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



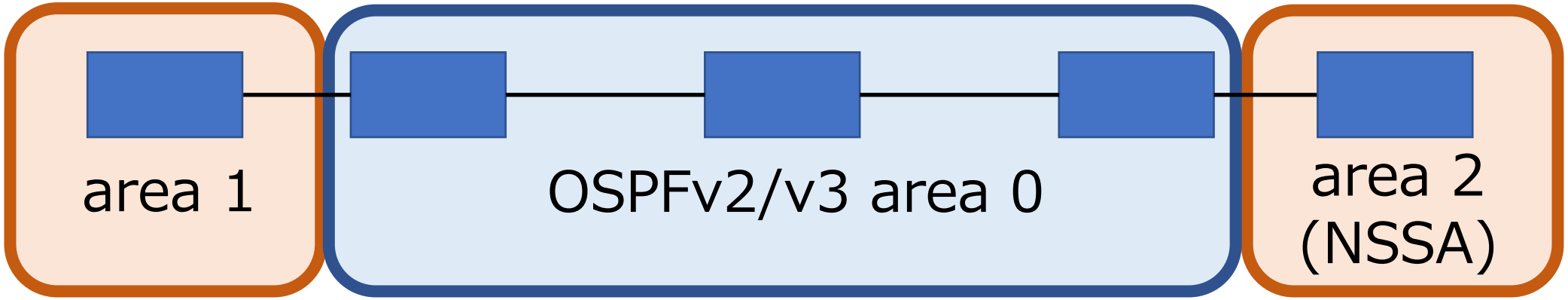
IGP：Area注意点

- 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



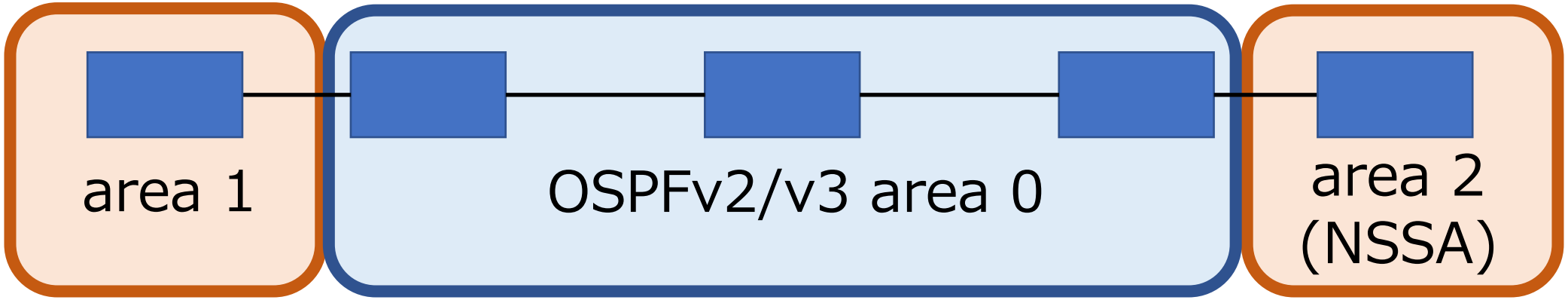
IGP Areaデザイン

- 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



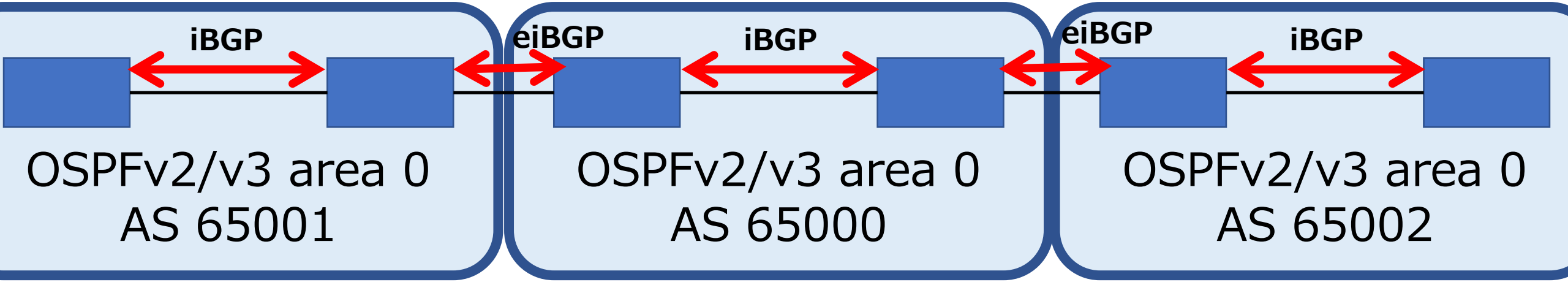
IGP Areaデザイン

- 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



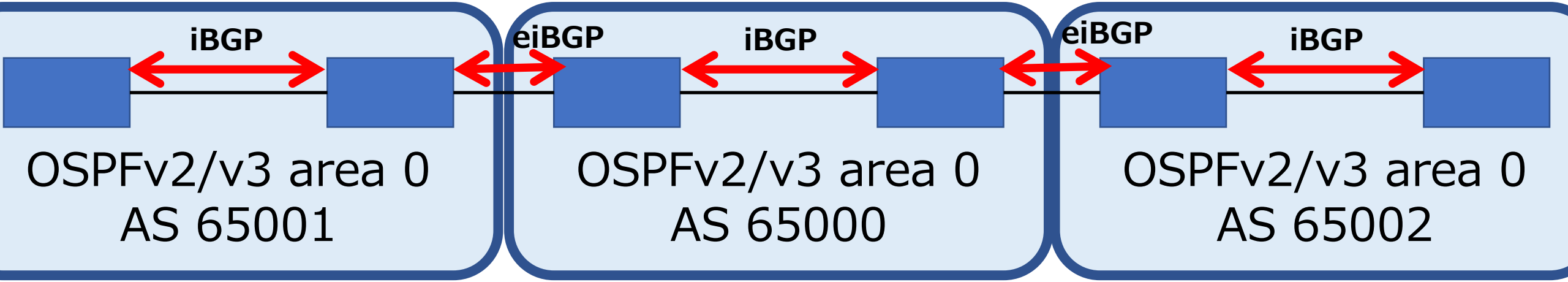
IGP Areaデザイン

- 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
    - Confederation機能を利用して、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

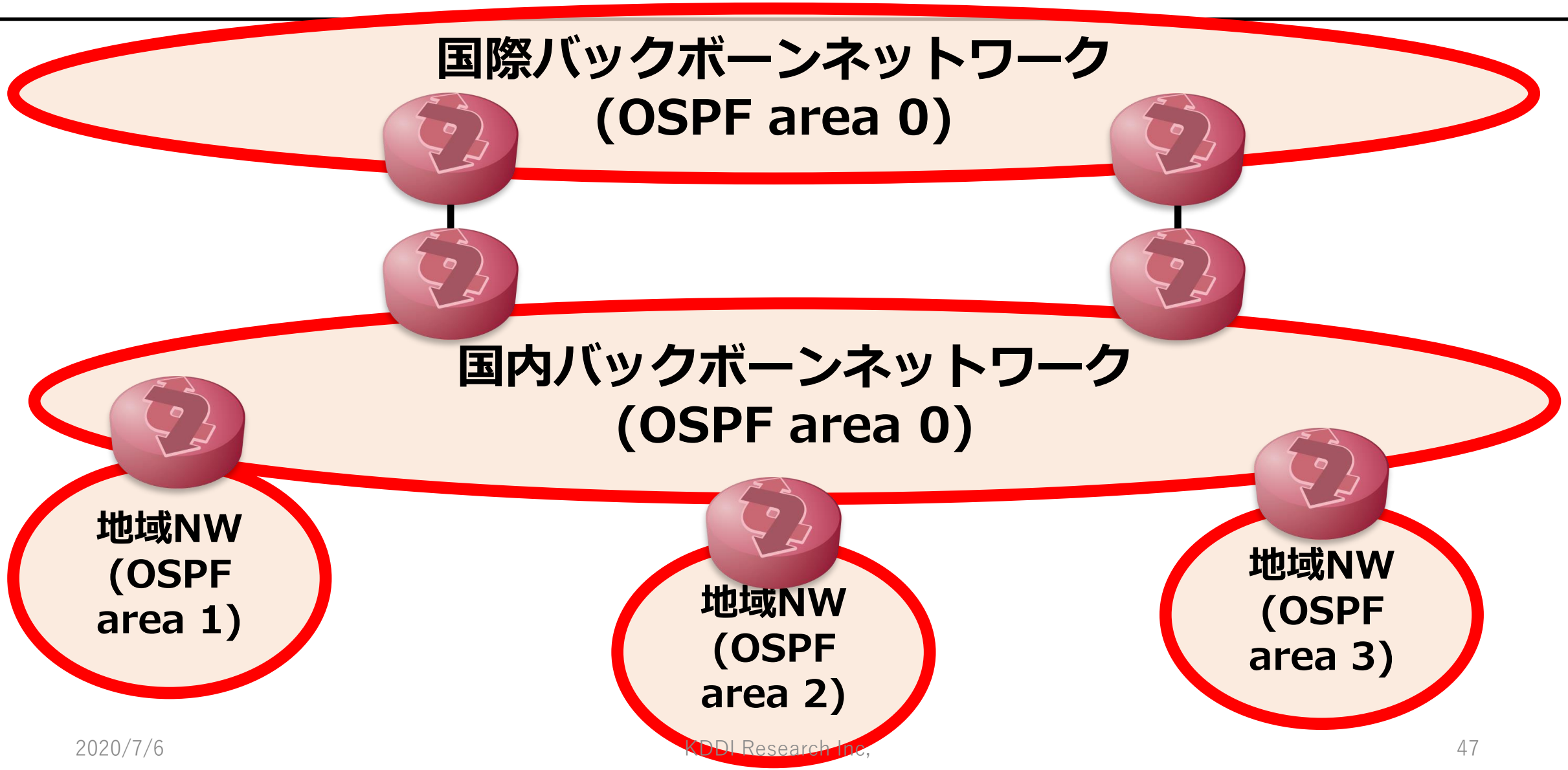


IGP Areaデザイン

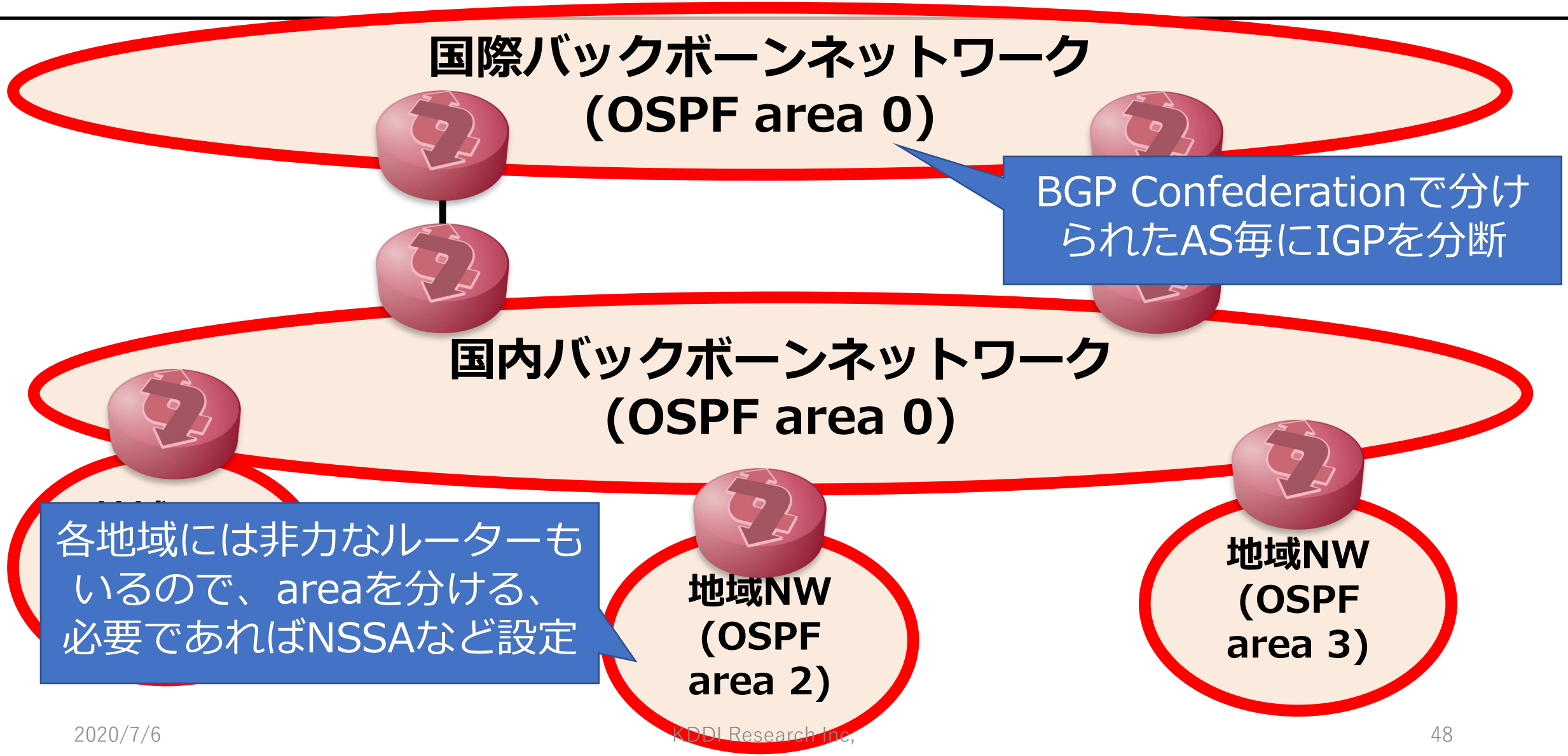
- ASとエリアの関係性
  - **(基本的に) BGPにIGPの経路をのせない！**
    - 先に書いた通り、BGPはサービス経路のみのせる
  - **(基本的に) IGPにBGPの経路をのせない！**
    - 誤ってインターネットフルルートがIGPに流れたら死亡します



とあるISPにおける設計例

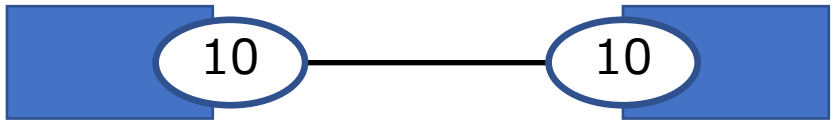


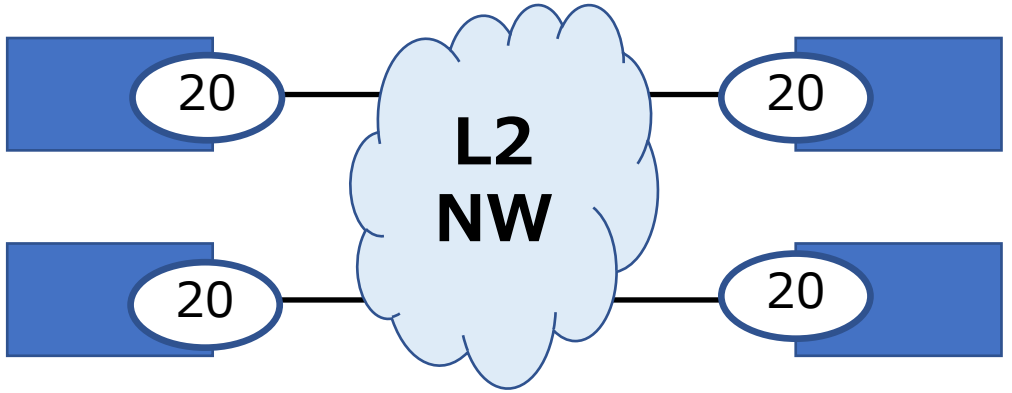
とあるISPにおける設計例



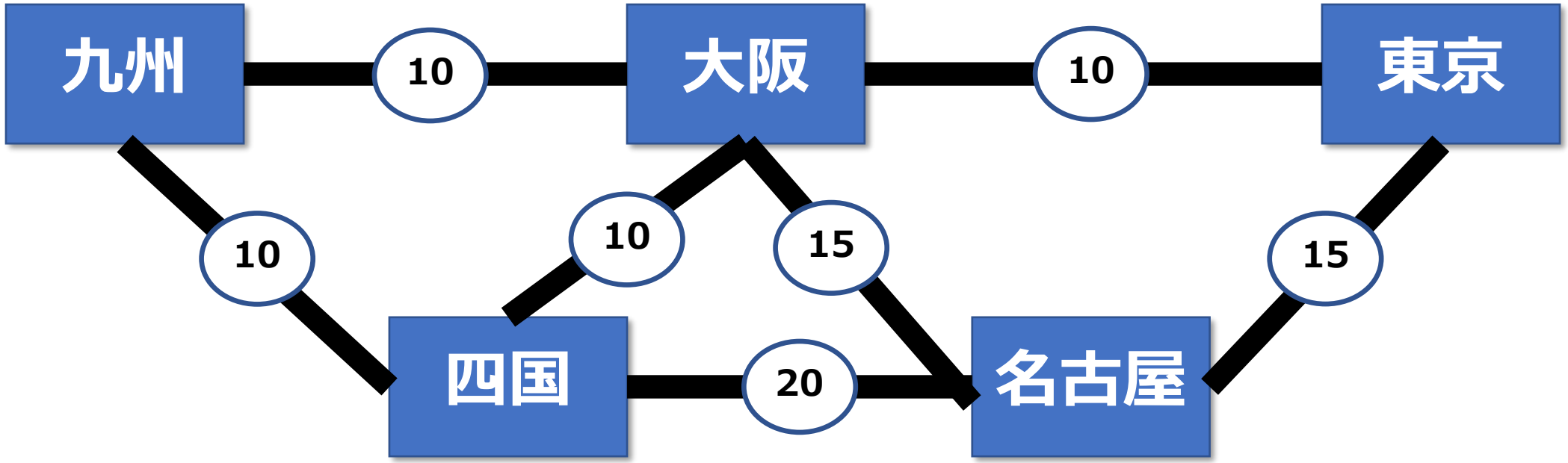
IGP：コスト

- 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

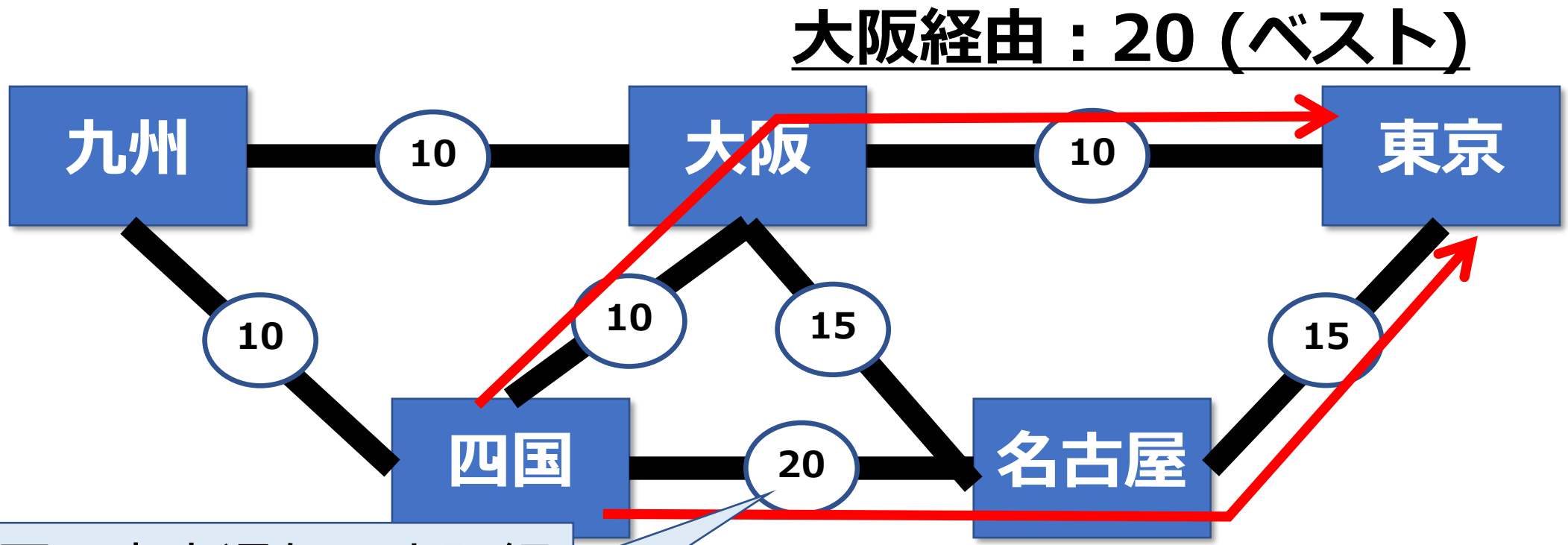




IGP：コスト設計の例

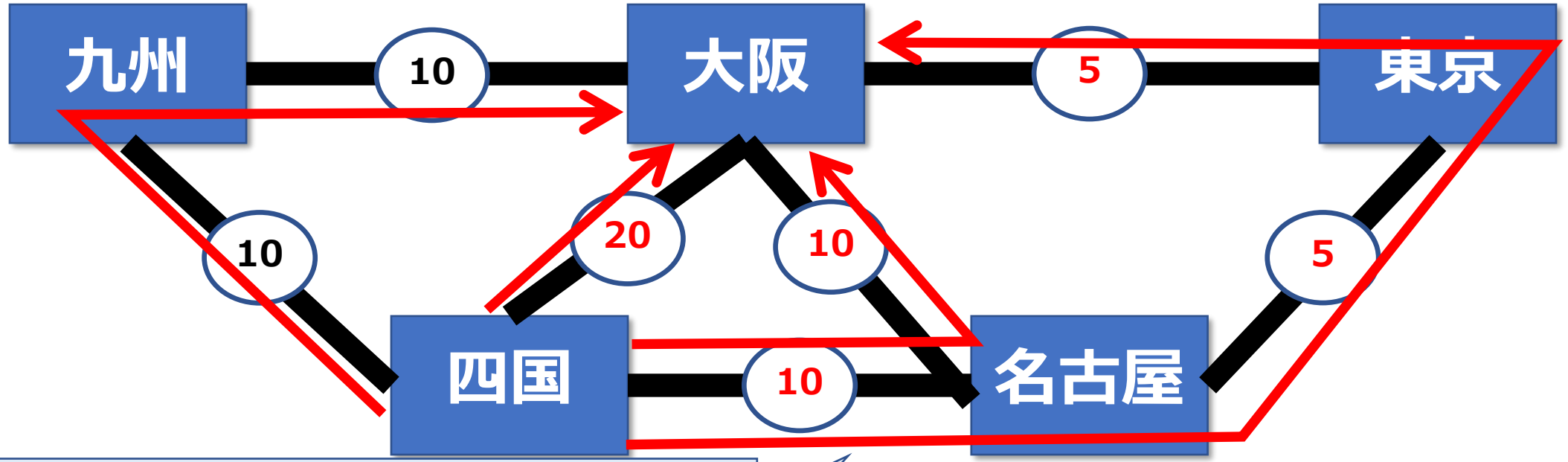


IGP：コスト設計の例



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：対障害対策

- リンク障害・ノード障害時に早期復旧するために、SLAと機器性能を考慮して以下をチューニング・導入検討していく (詳細は次ページ参考資料参照)
- リンク障害
  - SPF delay時間の調整
  - サイレント障害検知用プロトコルの導入
- ノード障害
  - Nos-stop-forwarding + Graceful Restart
  - Non-stop-routing
  - Max-metric on start-up

より詳細は・・・

- https://www.nic.ad.jp/ja/materials/iw/2019/proceedings/s07/s7-miyasaka-2.pdf

今日お話したこと：

1. インターネットとは？
2. ISPにおける経路設計デザイン概論
3. ISPにおけるプロトコル設計：BGP編
4. ISPにおけるプロトコル設計：IGP編

- インターネットにおけるISPの経路設計についてお話しましたが、**(1)まずはBGPの経路設計デザインを固め**、**(2)そのBGPデザインを実現するためにIGPの経路設計デザインをする** という流れが個人的には良いと思います