インターネットルーティング

Matsuzaki ‘maz’ Yoshinobu

<maz@iij.ad.jp>

インターネットがもたらすモノ

- コンテンツ、サービス
- 新たな視点、異文化

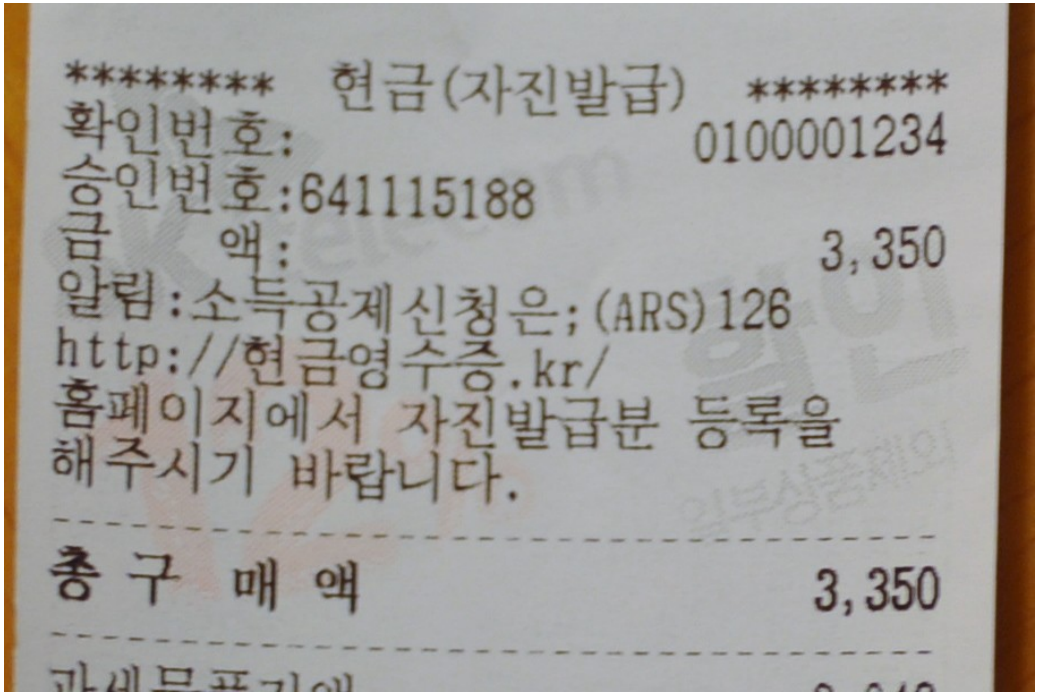
- 圧倒的な変革

守りたいものと、実装

- 破壊されうるもの
  - 伝統的価値、考え方
  - 社会体制や現地の産業
- 検閲と制御
  - パケットフィルタリング
  - DNSブロッキング
  - spoofing

ドメイン名がIDNな時代

- .বাংলা
- .ලංකා
- .இலங்கை
- .한국
- السعودية .
- .рф
- .ελ
- .日本



現地の方には親しみやすく見えるかもしれないが、
その言語を理解しない人にはアクセス不可能なURL

地域や国による制度

- ローミング料金
  - ローミング費用はEUとして規制
  - 通常の国内料金は事業者次第
- 音声通話
  - 音声通話の提供は許認可事業
  - VoIPも音声通話とみなす
- 関税
  - 精密機器に対する関税

設計

- 技術的障壁
  - 技術動向、業界動向
  - 製品の特長、制限
  - 検証
- ベンダへの外注
  - サービス毎、地域毎に違うベンダが担う場合も

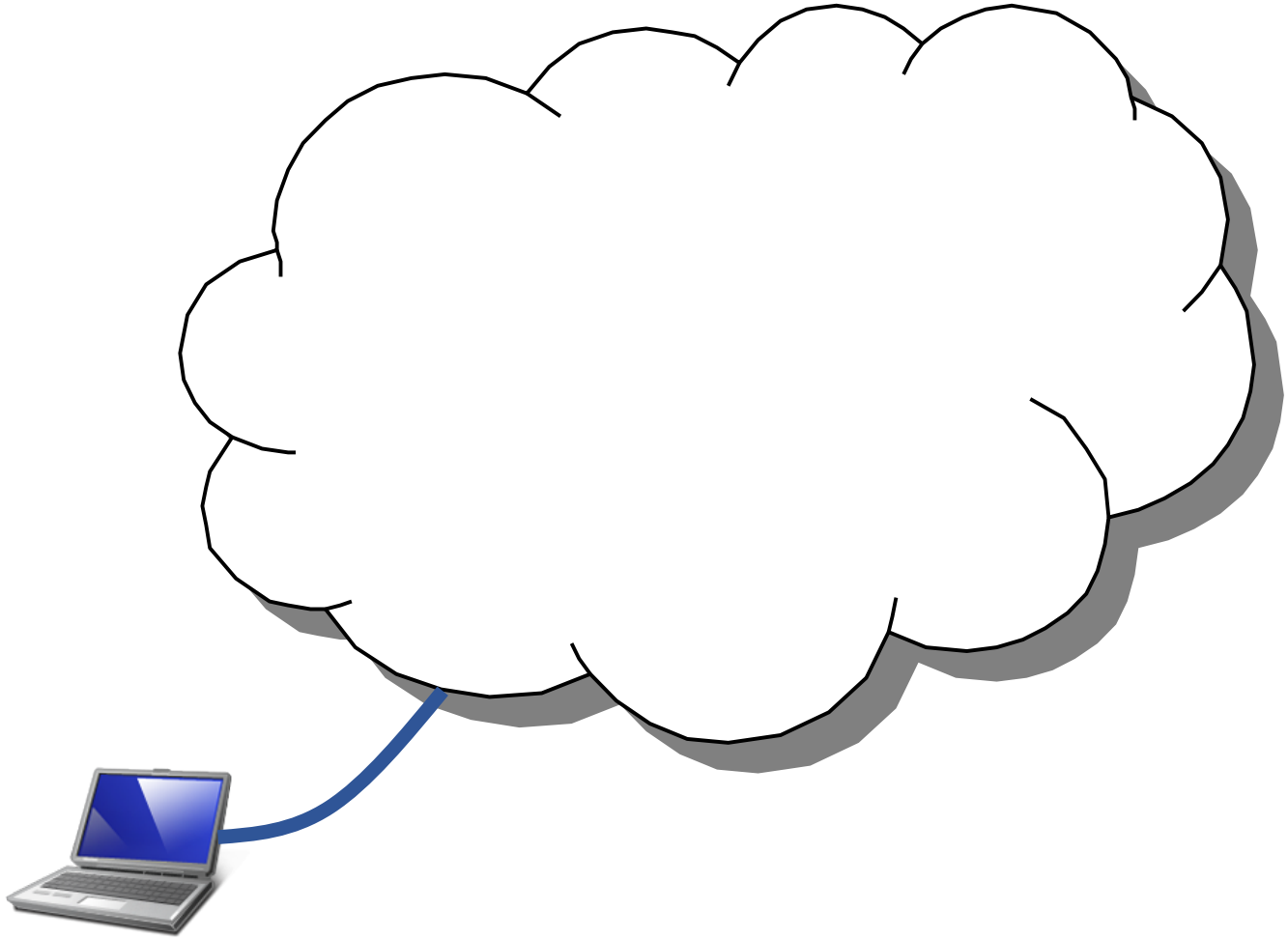
英語＋技術力

- 発展途上国からの人材流出
  - より良い生活環境
  - 安定した社会基盤
- 頻繁な転職
  - より良い給与・待遇

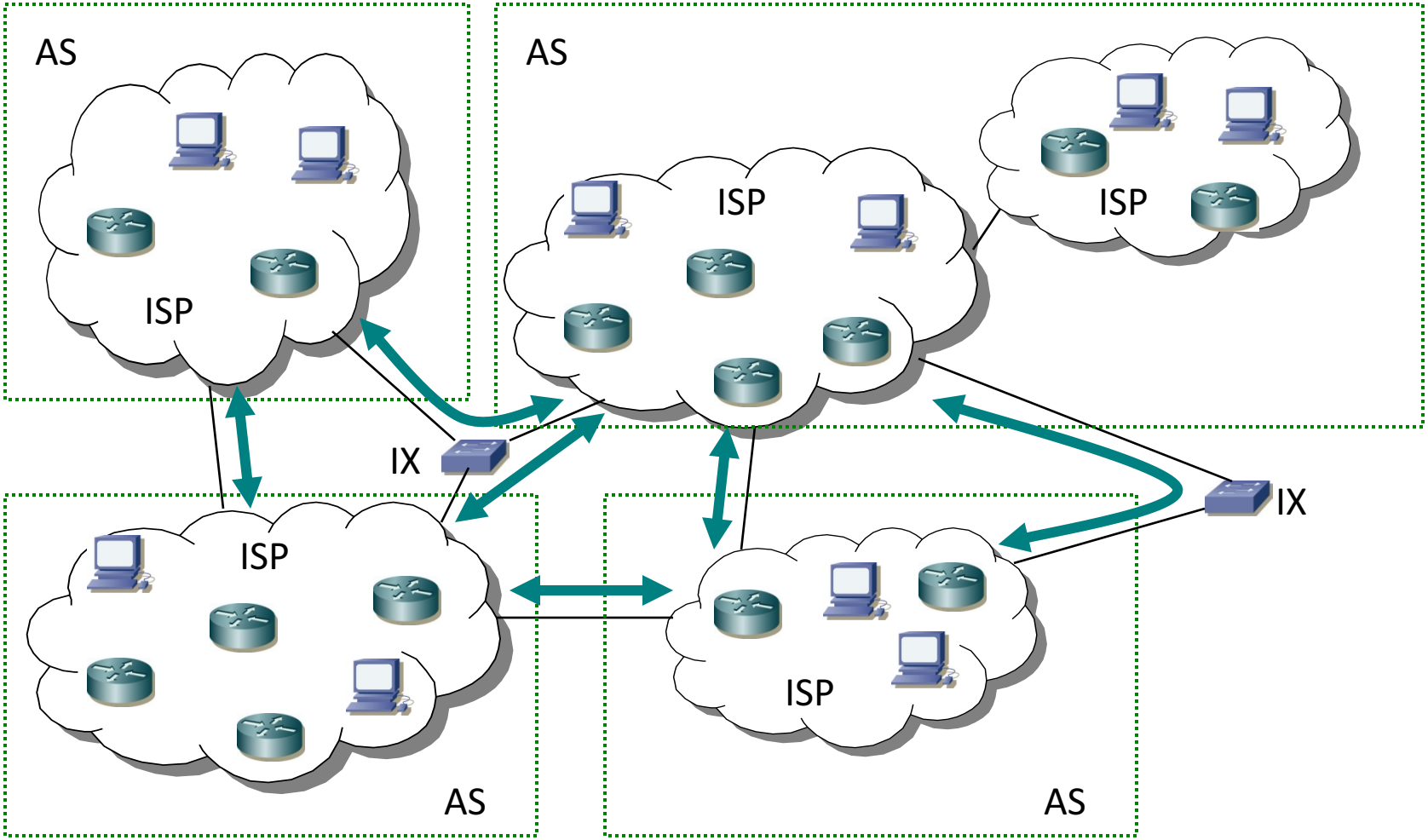
- 頻発する初歩的な障害の原因にも繋がっている
  - ずさんな管理と独断による施工
  - 自動化への期待がある一方、自動化を管理する人材が課題

バックアップ

インターネット

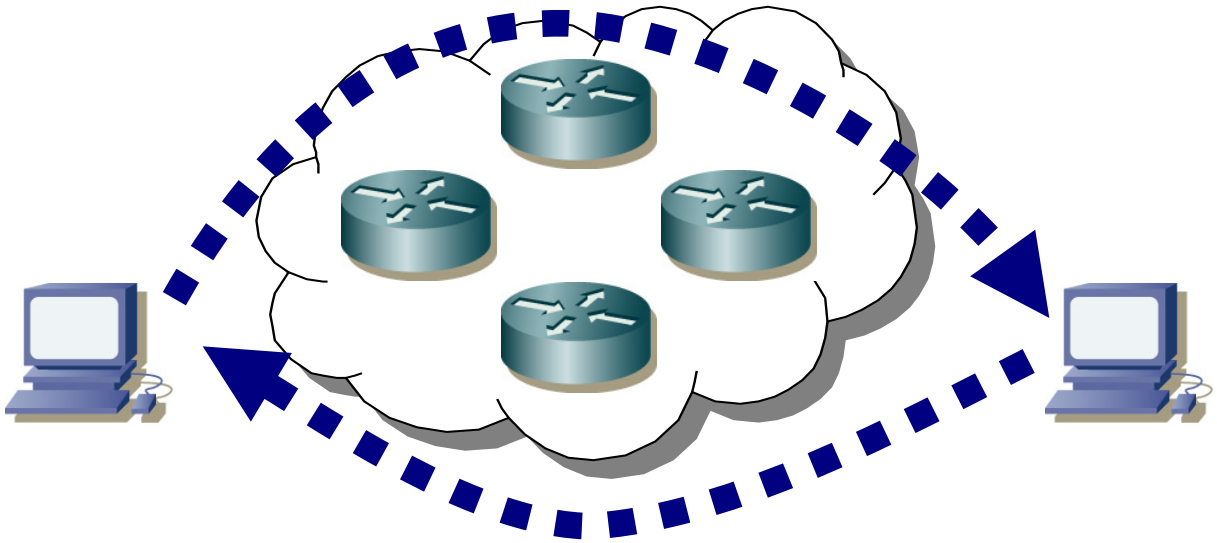


ネットワーク



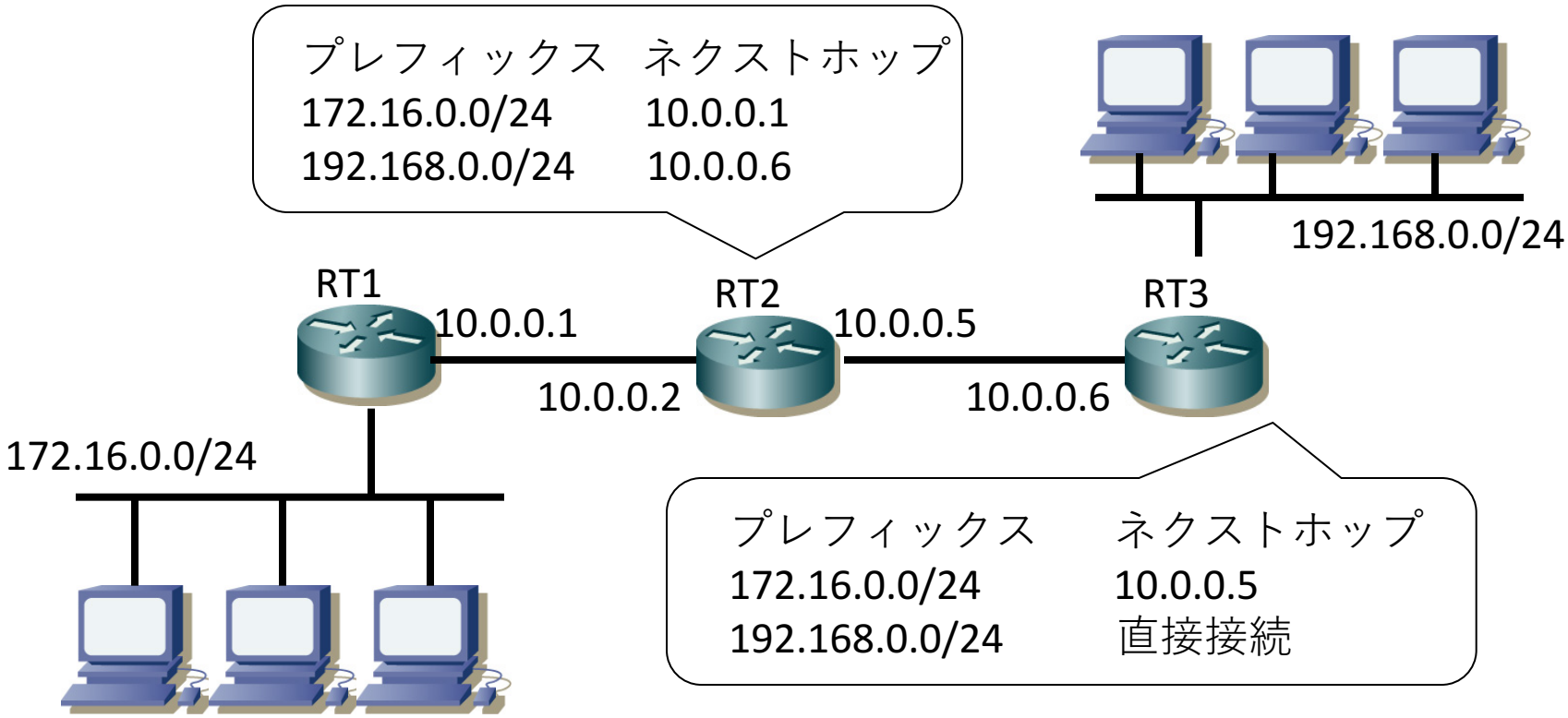
パケットと経路

- 送信元から宛先まで経路に矛盾が無ければ、パケットが届く
- 双方向で問題が無ければ、相互に通信できる
  - 行きと帰りの経路は違うかもしれない



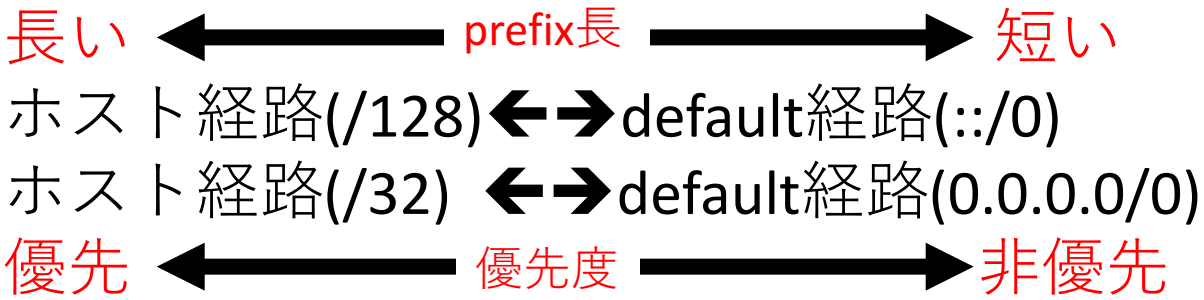
経路情報

- “宛先プレフィックス”＋“ネクストホップ”の集合



経路の優先順位

1. prefix長が長い(経路が細かい)ほど優先



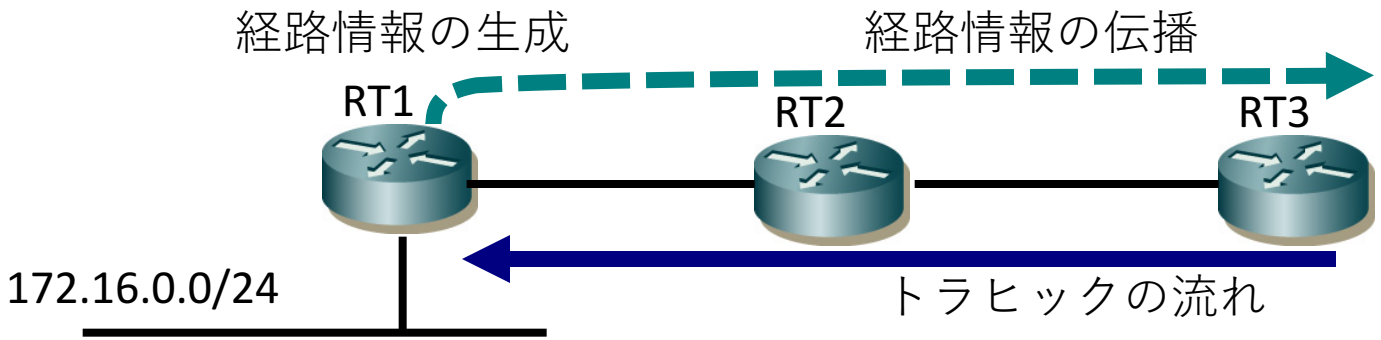
2. 経路種別で優先
   - ① connected経路
   - ② static経路
   - ③ 動的経路(ospf, bgp, etc...)
     - 内訳はベンダ依存

経路の種類

- 静的経路
  - connected経路
    - ルータが直接接続して知っている経路
  - static経路
    - ルータに静的に設定された経路
- 動的経路
  - ルーティングプロトコルで動的に学習した経路
    - OSPFやIS-IS、BGPなどで学習した経路

動的経路制御の基本アイディア

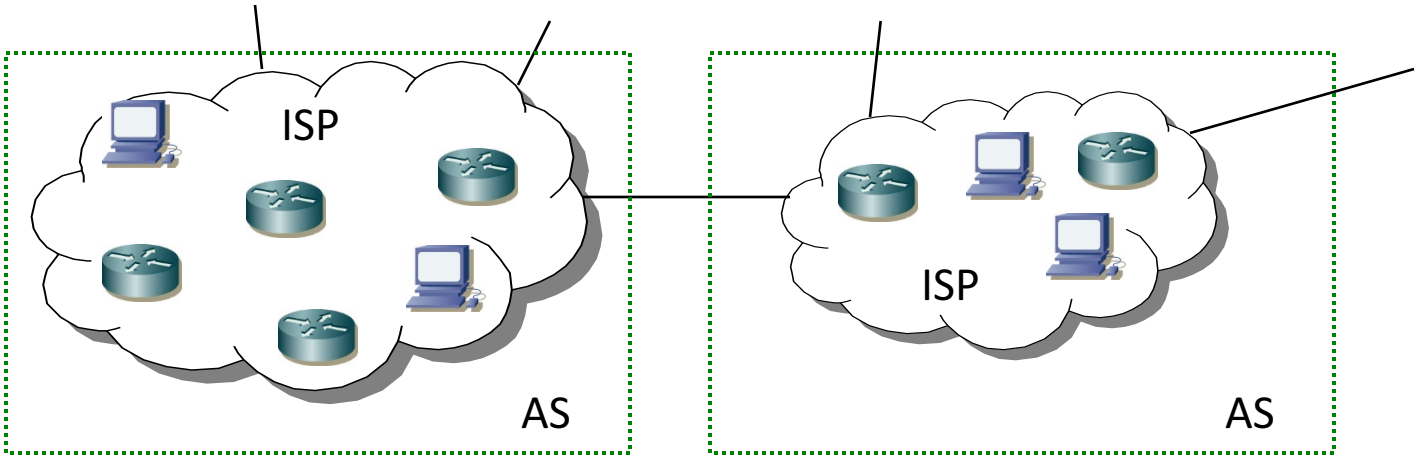
- 検知 – ルータがネットワークの変化を検知
- 通知 – 情報を生成し他のルータに伝達
- 構成 – 最適経路で経路テーブルを構成



経路情報の伝搬の方向とトラヒックの流れは逆になる

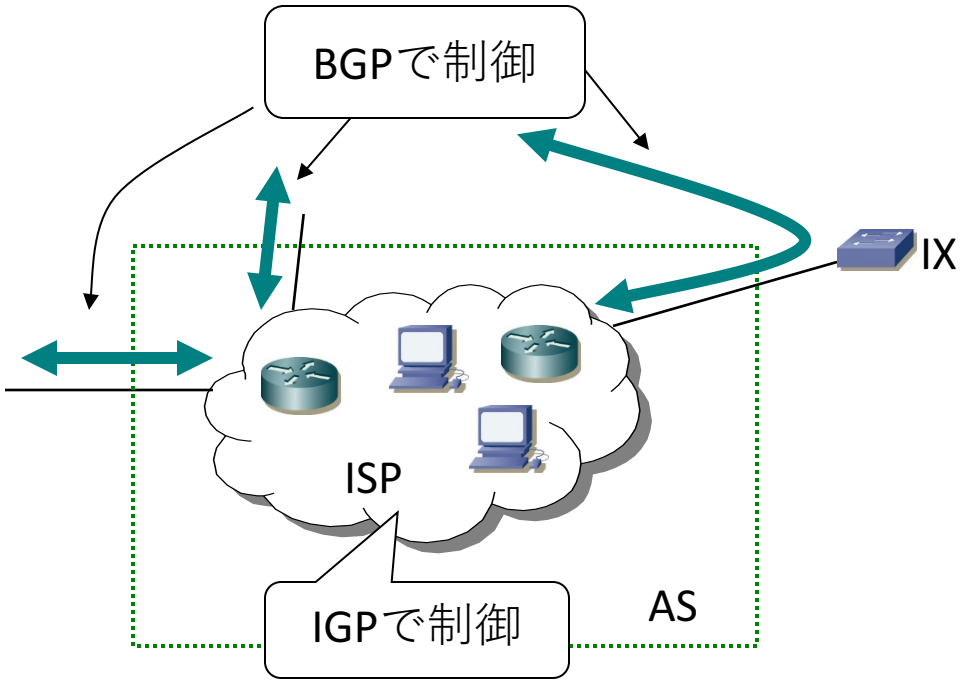
AS

- Autonomous System
- 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
- インターネットではASの識別子として、IRから一意に割り当てられたAS番号を利用する

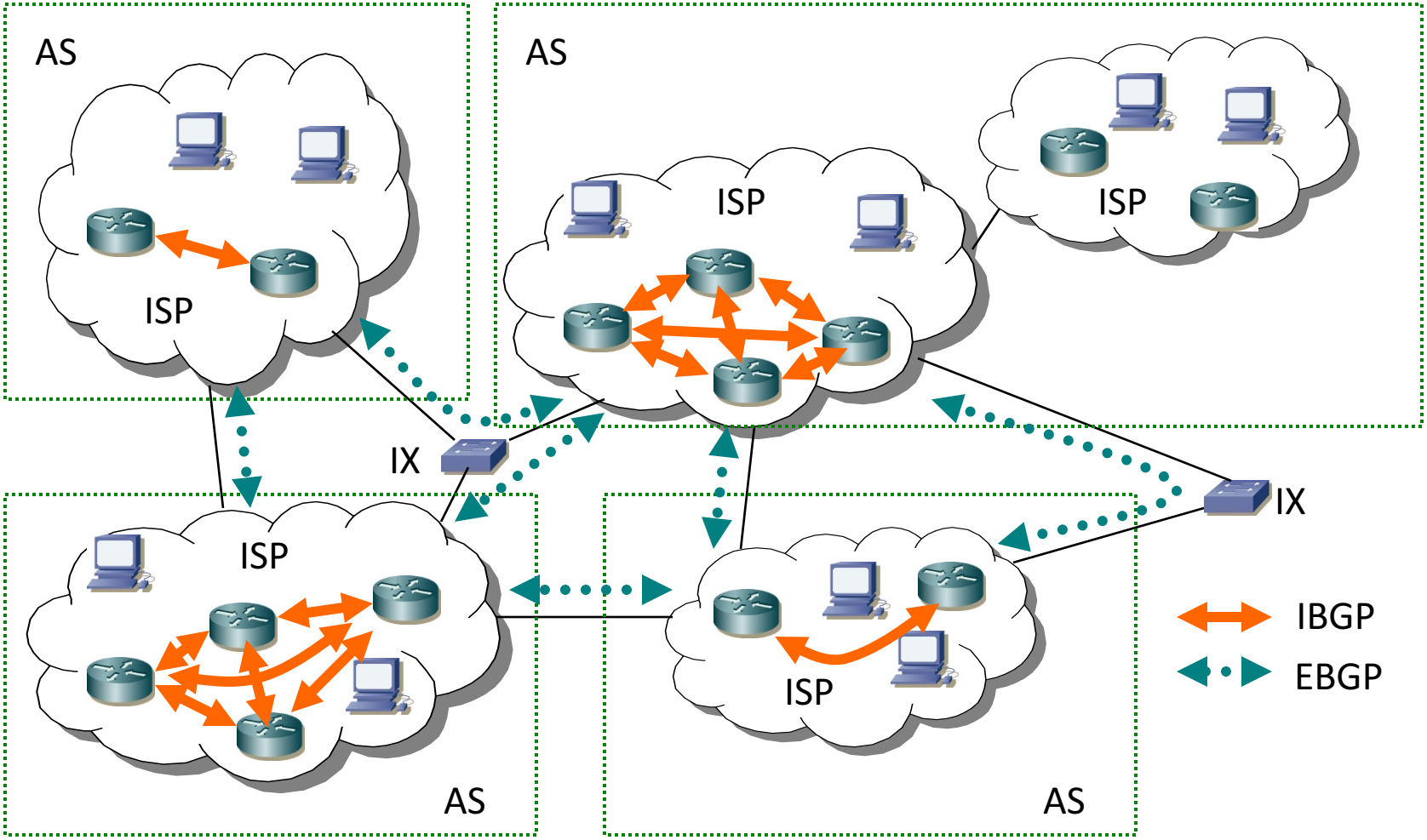


IGPとEGP

- IGP
  - OSPF、 IS-IS、 BGP等
  - AS内
- EGP
  - 事実上BGPのみ
  - AS間



BGPの世界



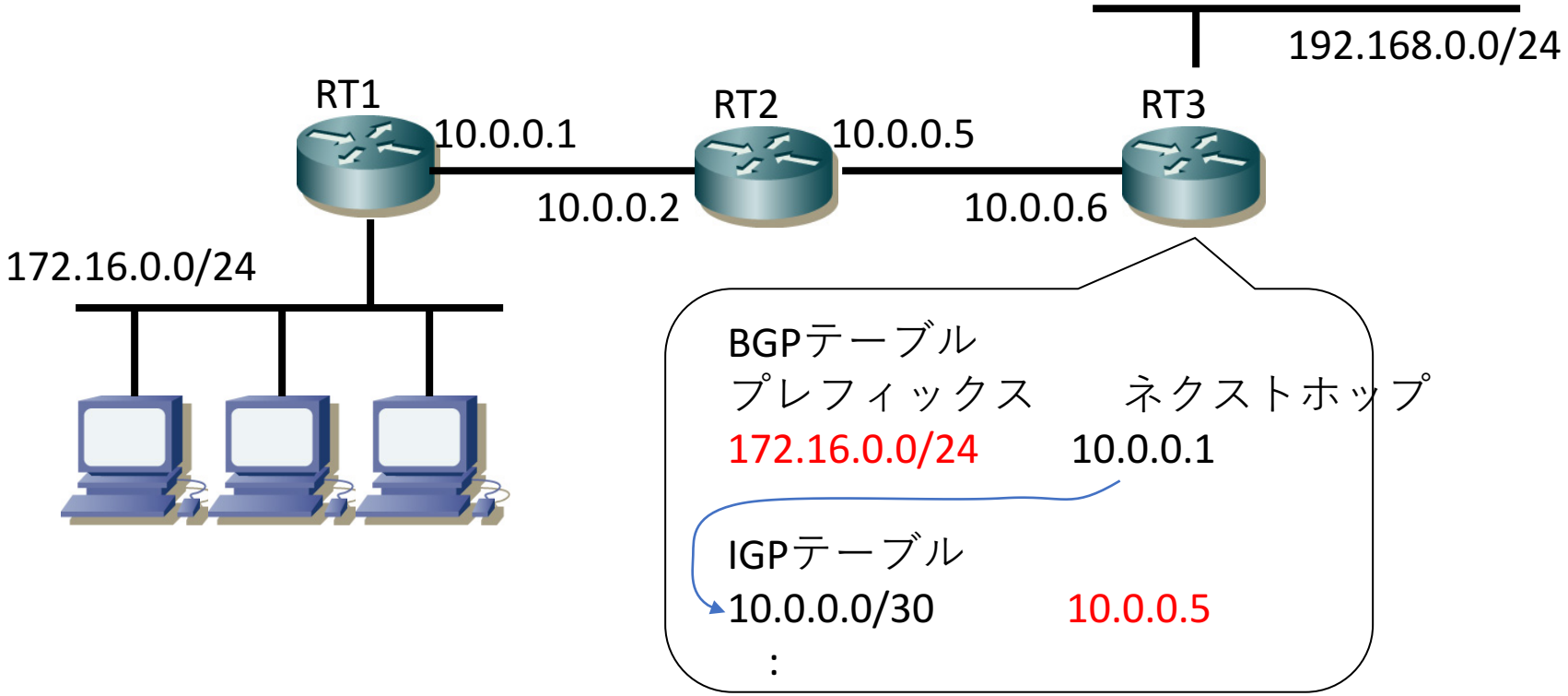
ISPでのプロトコルの利用法

- OSPF or IS-IS
  - ネットワークのトポロジ情報
  - 必要最小限の経路で動かす
  - 切断などの障害をいち早く通知、迂回
- BGP
  - その他全ての経路
    - 顧客の経路や他ASからの経路
  - 大規模になっても安心
  - ポリシに基づいて組織間の経路制御が可能

BGPの基本アイディア

- 準備
  - 経路交換したいBGPルータとTCPでネイバを構築
  - (ネイバ|ピア|BGPセッション)を張るとも言う
- 通知
  - ベスト経路に変更があればUPDATEとしてネイバに広報
  - 受信した経路は幾つかの条件を経て、他のネイバに広報
- 構成
  - 各ルータが受信経路にポリシを適用し、パス情報を元にベスト経路を計算

BGPと再帰経路



BGPで学習したネクストホップアドレスをさらに
経路情報で再帰的に探して、ルータが実際に
パケットを送出する宛先を見つけ出す
「172.16.0.0/24宛は10.0.0.5(RT2)にフォワード」

ASの運用

- 到達性の確保
  - 何はともあれ、到達性が重要
  - 大抵、どこかからtransitを購入して保険をかける
- トラヒックの制御
  - BGPは回線の空き具合を気にしない
  - 回線や設備はそんなに柔軟に変えられない
    - ホントは需要に応じて増強するのが一番きれい
  - それでも対処しなきゃいけない事案は出てくる

基本的なお作法

- PAブロックは割り振られたサイズで広報
  - 細かい経路やprivate AS&アドレス等を漏らさない
  - 広報する経路に責任をもつ
- 全ての接続点で一貫した経路広報
  - 相互接続しているASには、どの接続点でも同一の経路を広報
- 何らかトラヒック制御しようとする場合には、事前に相互接続先と相談

経路制御ポリシ

- あった方が運用に一貫性が出て良い
  - 意図しない経路制御を防止できる

- ポリシを考えるもと
  - 提供したい通信、自由度
  - トラヒック制御
  - 自身の経路制御の防御

対外接続

- EBGPで接続
- 他のASと経路交換
  - トランジットしてもらって到達性の確保
  - ピア（相互接続）で独自の接続性の向上
- 接続方法
  - 相互接続に合意
  - 専用回線やIXで接続

専用回線でEBGP（プライベートピア）

- インタフェースの合意
  - 速度や種別
- 必要に応じて回線手配と費用分担の調整
  - 構内回線や回線サービスなど
- その回線で利用するIPアドレス手配
  - どちらかの組織から持ち出しになる場合が多い
  - /30or/31, /64or/127
- ネイバの設定

IXでEBGP（パブリックピア）

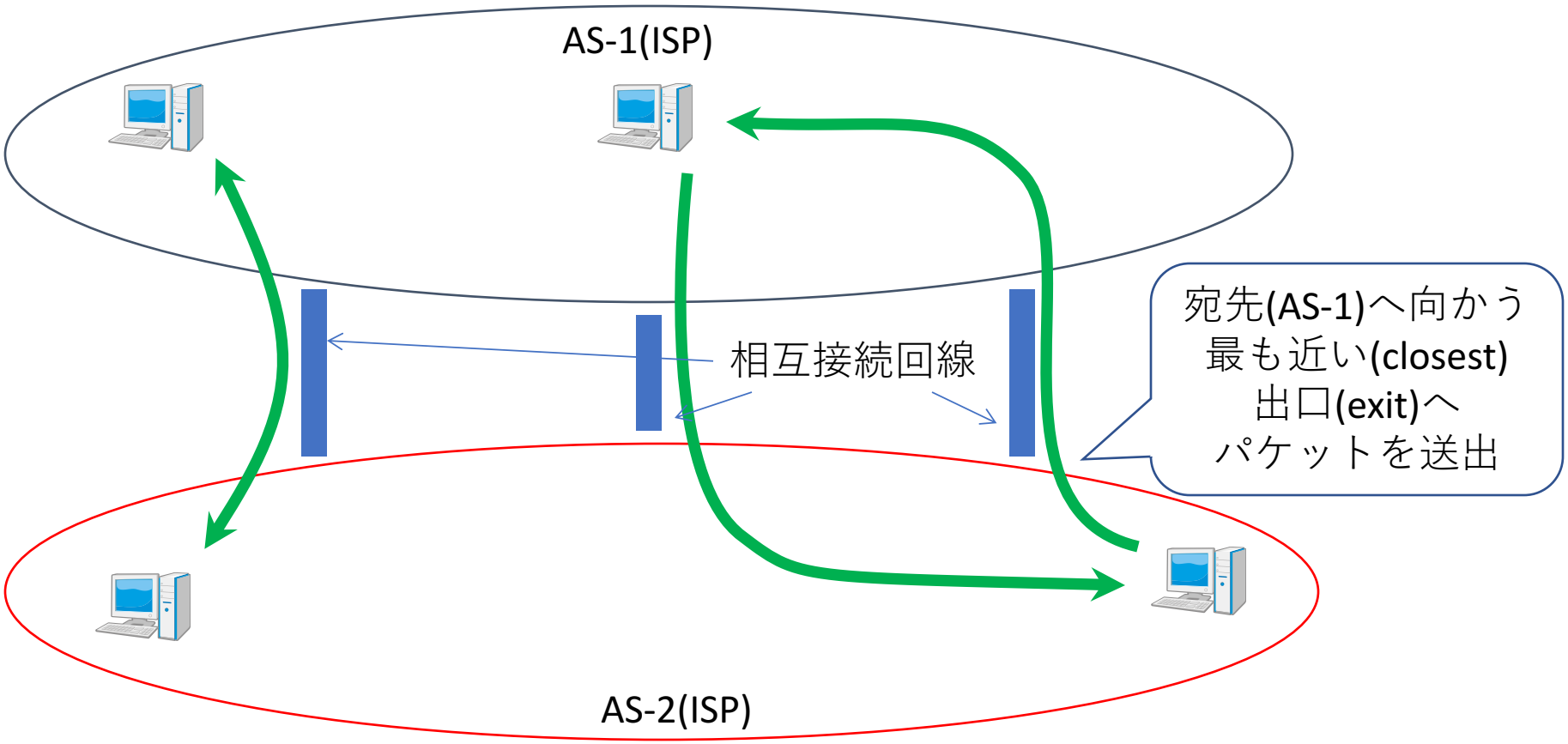
- お互いに同じIXに居る事の確認
- お互いのIPアドレスの通知
  - IXで提供される個別セッションサービスやVLANサービス等を利用する場合、IPアドレスの手配が必要な場合もある
- ネイバの設定

あるASと複数拠点で相互接続

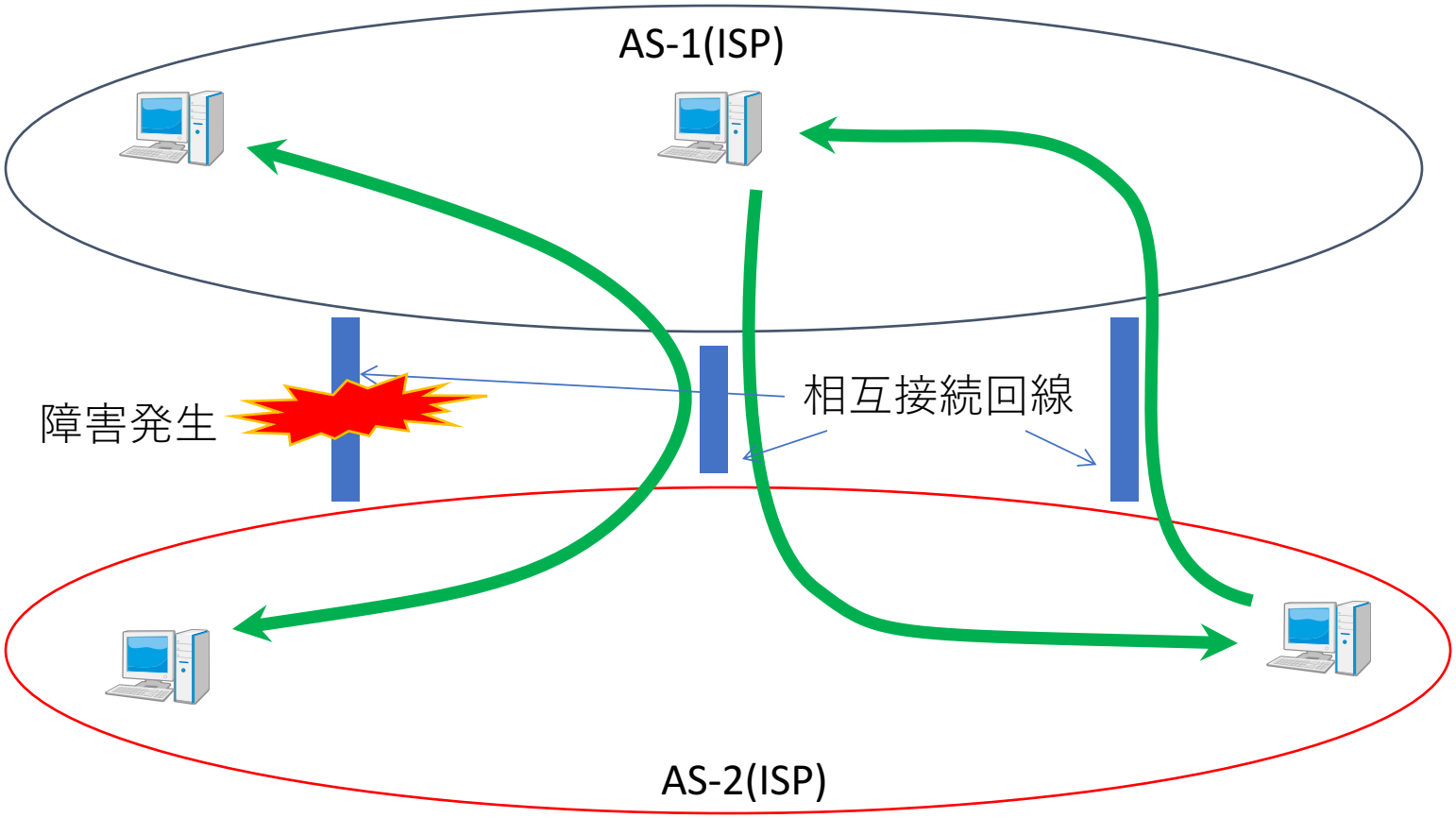
- トラヒック制御が課題になる
- **お互いに相手ネットワークの事は分からない**
- **最適な経路を選ぶには、宛先に近いネットワークに素早くパケットを渡せば良い**

  ＝ **closest exit(クローゼスト イグジット)**
  - BGPの素直な利用方法
  - 世界のISPが標準的に採用しているポリシ

closest exit



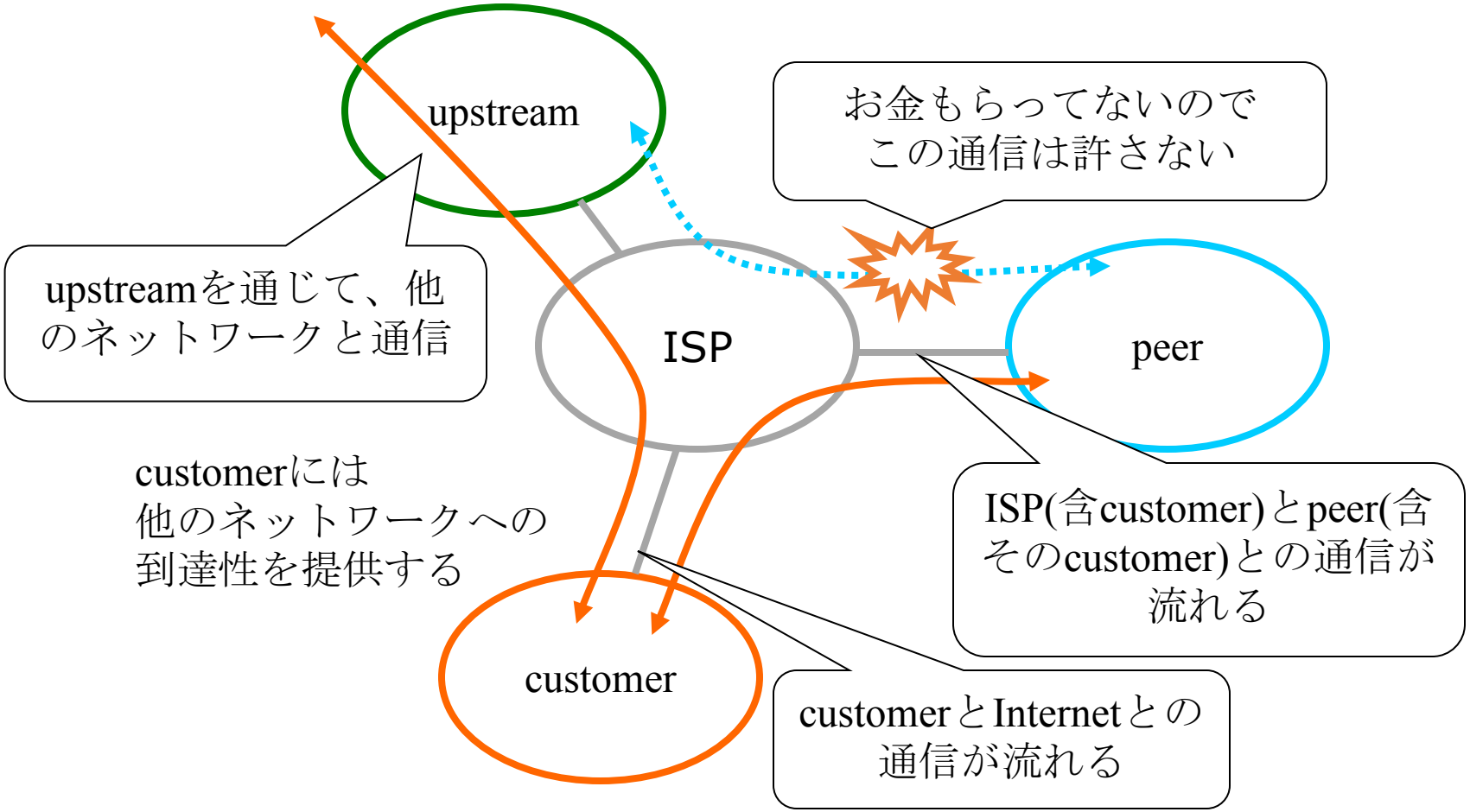
障害発生時のclosest exit



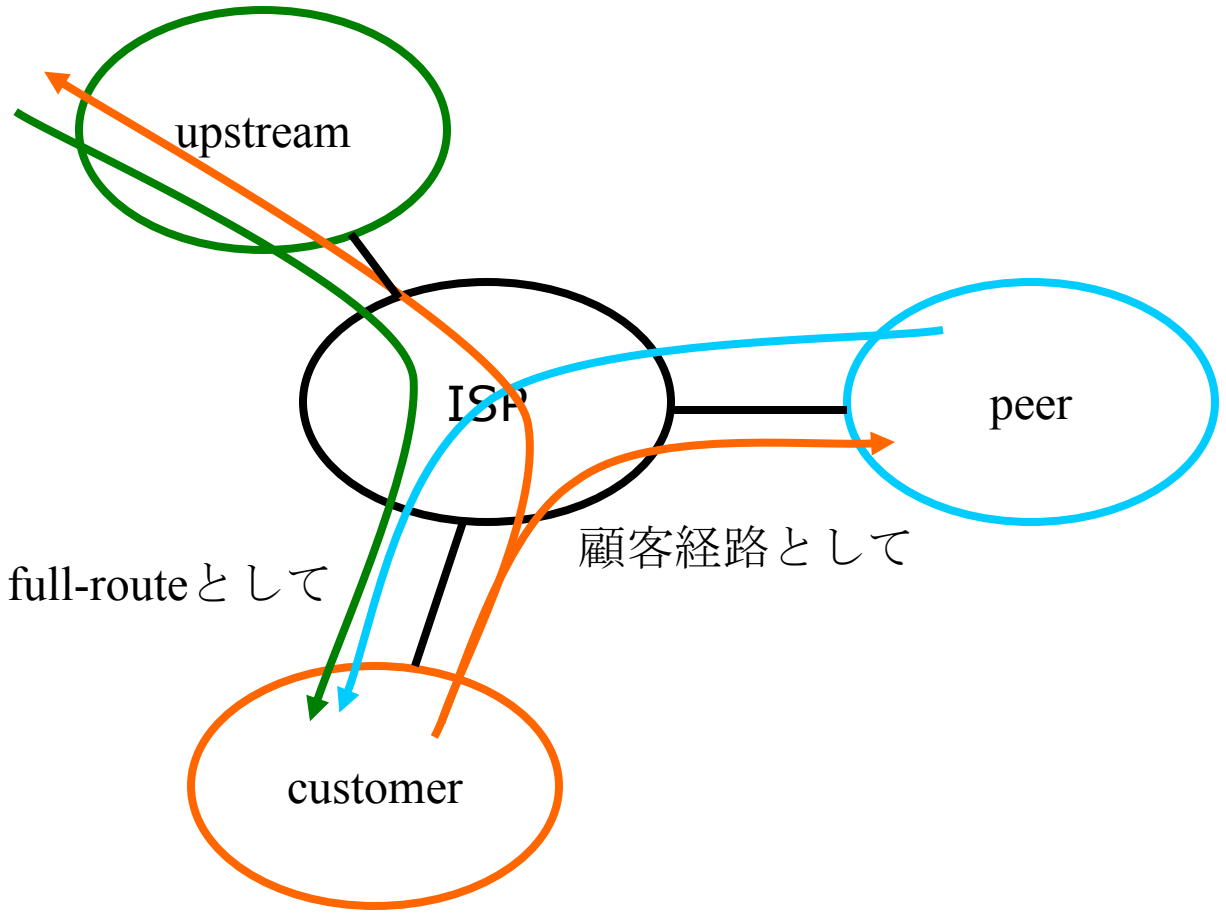
closest exitの特長

- 簡単なポリシで最適な経路を選びうる
  - BGPはclosest exitを前提として設計されている
- 相互接続ポイントが増えても、それまでと同じ経路制御ポリシのままで運用できる
  - 拡張性に優れる
  - 特別な設計が必要ない

顧客に提供したい通信


upstream
お金もらってないので
この通信は許さない
upstreamを通じて、他
のネットワークと通信
ISP
peer
customerには
他のネットワークへの
到達性を提供する
ISP(含customer)とpeer(含
そのcustomer)との通信が
流れる
customer
customerとInternetとの
通信が流れる

対応する経路広報の流れ



トランジットの実装方法

- 普通はBGP community
  - 顧客経路の受信時にtransit用のTAG付け
    - 顧客からの経路受信時に経路フィルタの併用が必須
  - 外部にはtransit用TAGがついた経路のみを広報
- 小規模なら経路フィルタでも実現可能
  - トランジットする経路をprefixフィルタで管理
  - 外部に広報するときに、このフィルタを適用
    - 顧客から広報されなくてもtransitしてしまうかも

受信経路の基本的な優先制御

- 経路優先度
  - customer ＞ peer　≧ transit
  - ほとんどのASが、LOCAL_PREFを使って実装
- customer経路は優先
  - 顧客にtransitを提供するために優先
    - BGPはベスト経路しか広報しないよね
    - 他から広報された経路が優先されちゃうとtransitできない
- peerとtransitから受信した経路の優先度は低め
  - 少なくともcustomerからの経路よりも低め

LOCAL_PREF

- AS内での経路優先度を示す優先度
- 経路受信時に明示的に設定しておくのが吉

接続相手	設定するLOCAL_PREF例
customer	200
peer	100
upstream	90

- LOCAL_PREFは強すぎるので、これ以外の制御に使わない方が良い

MED

- 隣接ASとの距離を示す値
  - あるASと複数接続がある場合に、それぞれの優先度を設定
  - eBGPで経路の広報元が値を設定しても良いし、受信側で適当な値を設定しても良い
  - バックアップ経路の指定や、拠点やIXなど狭い範囲での経路選択に利用される場合が多い
- 機器によって実装が違う場合があるので注意
  - 設定してなければ0として扱う (RFC4271)
    - MEDを利用した制御を行うなら、何らの値を明示的に設定するべし

MEDの評価

- non-deterministic-med (cisco default)
  - 受信経路の到着順序に従って最適経路を選択する
  - MEDの値が思い通りに評価されないことがあるため、普通使わない
- deterministic-med (juniper default)
  - 同一ASから受信した経路同士を先に比較して、その後再度最適経路を選択する
  - みんな使ってる
- always-compare-med
  - 異なるASから受信した経路でもMEDの値を評価する

受信経路のMED

- 受信時に上書き
  - 制御を提供しない場合
  - upstreamやpeerからの経路等
- 受信したMEDをそのまま利用
  - 制御を提供する場合
  - customerやpeerからの経路等

PA経路の生成

- 内部のルータでnull向けstatic経路から生成
  - 障害に備え、複数のルータで生成しておく
- 外部に広報する直前にsummary経路として生成
  - 障害に備え、summary経路の生成条件を明示しておく