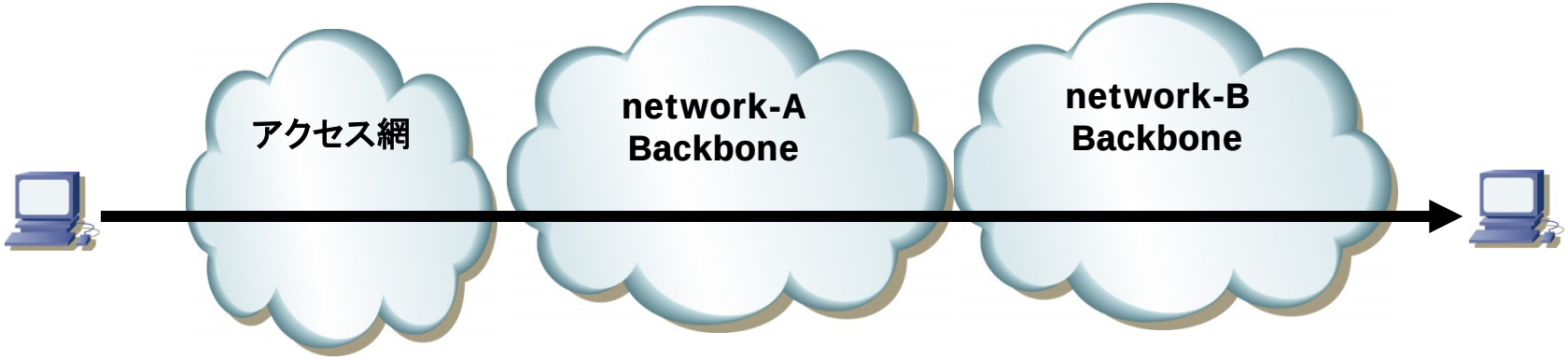
IIJのバックボーンネットワーク運用

Matsuzaki ‘maz’ Yoshinobu
<maz@iij.ad.jp>

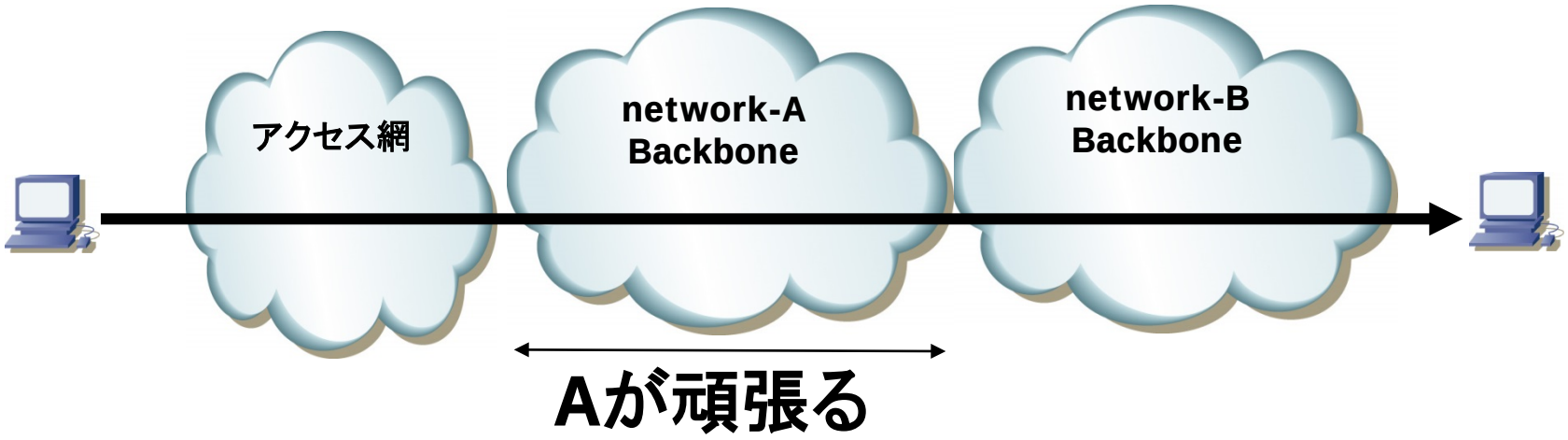
インターネットの接続

- ユーザはどこを経由するか気にしない
- 途中の皆が頑張れば、品質良く通信できる

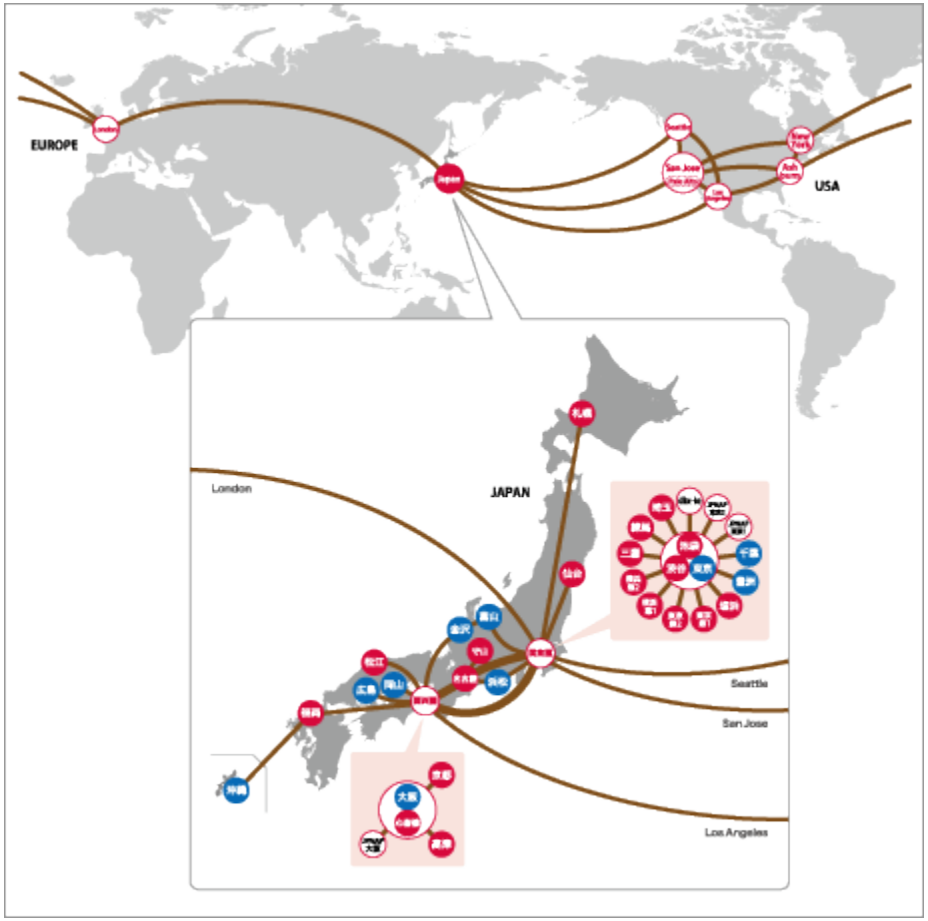


各バックボーン

- それぞれの範囲で管理を頑張る



IIJのバックボーン



東京-ロンドンのRTT

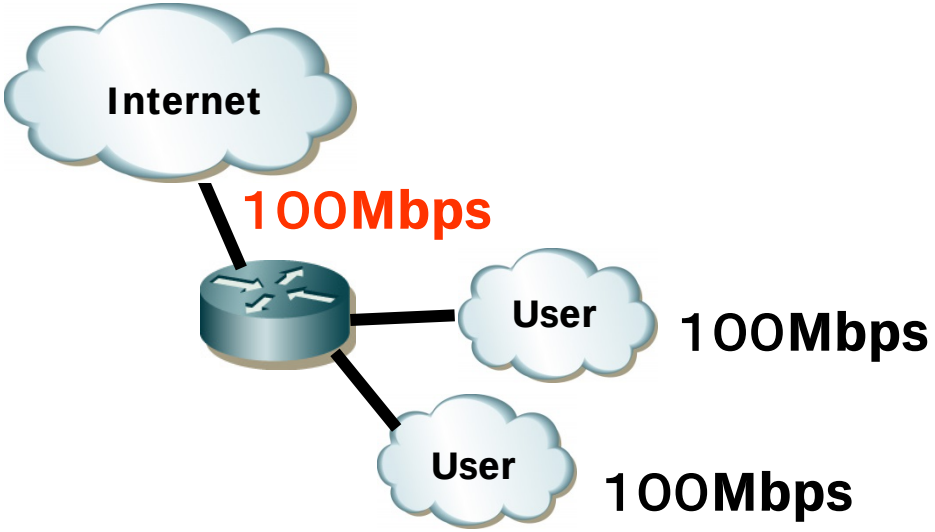
```
% ping -c10 ldn001bb10.iij.net
PING ldn001bb10.iij.net (58.138.97.197): 56 data bytes
64 bytes from 58.138.97.197: icmp_seq=0 ttl=58 time=168.767 ms
64 bytes from 58.138.97.197: icmp_seq=1 ttl=58 time=168.642 ms
64 bytes from 58.138.97.197: icmp_seq=2 ttl=58 time=168.532 ms
64 bytes from 58.138.97.197: icmp_seq=3 ttl=58 time=168.629 ms
64 bytes from 58.138.97.197: icmp_seq=4 ttl=58 time=168.627 ms
64 bytes from 58.138.97.197: icmp_seq=5 ttl=58 time=168.644 ms
64 bytes from 58.138.97.197: icmp_seq=6 ttl=58 time=168.604 ms
64 bytes from 58.138.97.197: icmp_seq=7 ttl=58 time=168.585 ms
64 bytes from 58.138.97.197: icmp_seq=8 ttl=58 time=168.647 ms
64 bytes from 58.138.97.197: icmp_seq=9 ttl=58 time=168.607 ms

--- ldn001bb10.iij.net ping statistics ---
10 packets transmitted, 10 packets received, 0.0% packet loss
round-trip min/avg/max/stddev = 168.532/168.628/168.767/0.057 ms
```

バックボーンは共用設備

- ネットワークの基幹部分
  - サービスや利用者で共用されるネットワーク
- 様々な要件の通信が通る
  - 品質や遅延
  - それぞれの要件に耐えうる設計

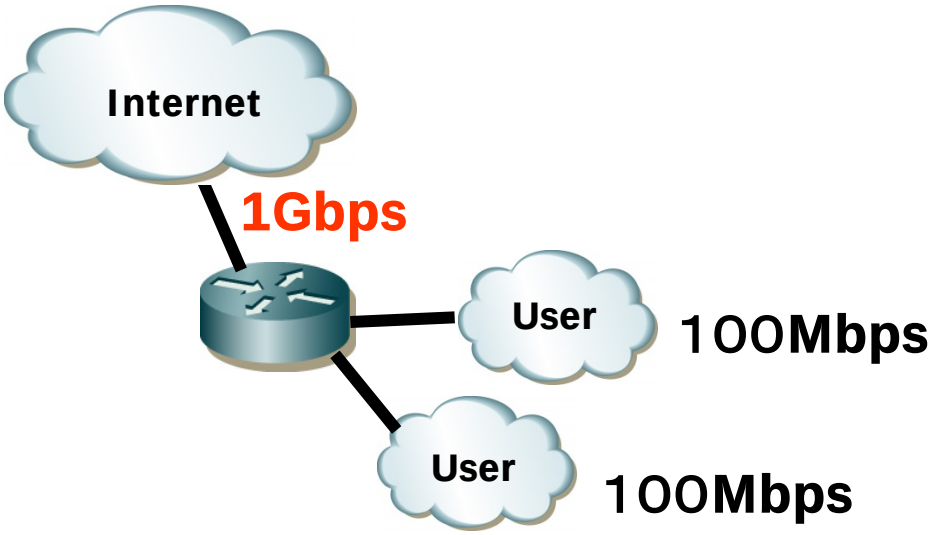
設備設計

- over-subscription
  - 統計多重の効果を期待
  - 同時に100%使うわけではない
  - 利用率の見込みが必要



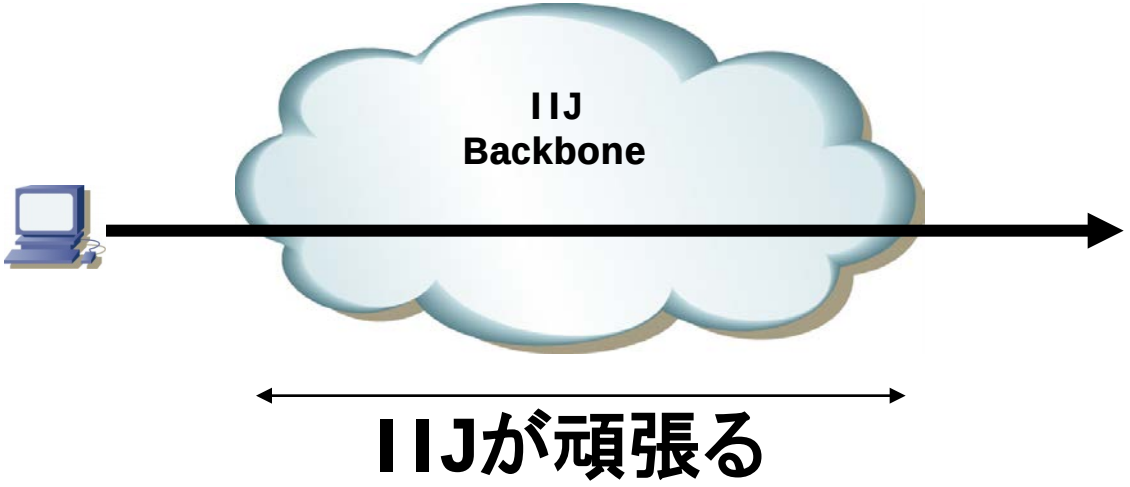
設備設計2

- over-provisioning
  - –設備を過剰に供給する

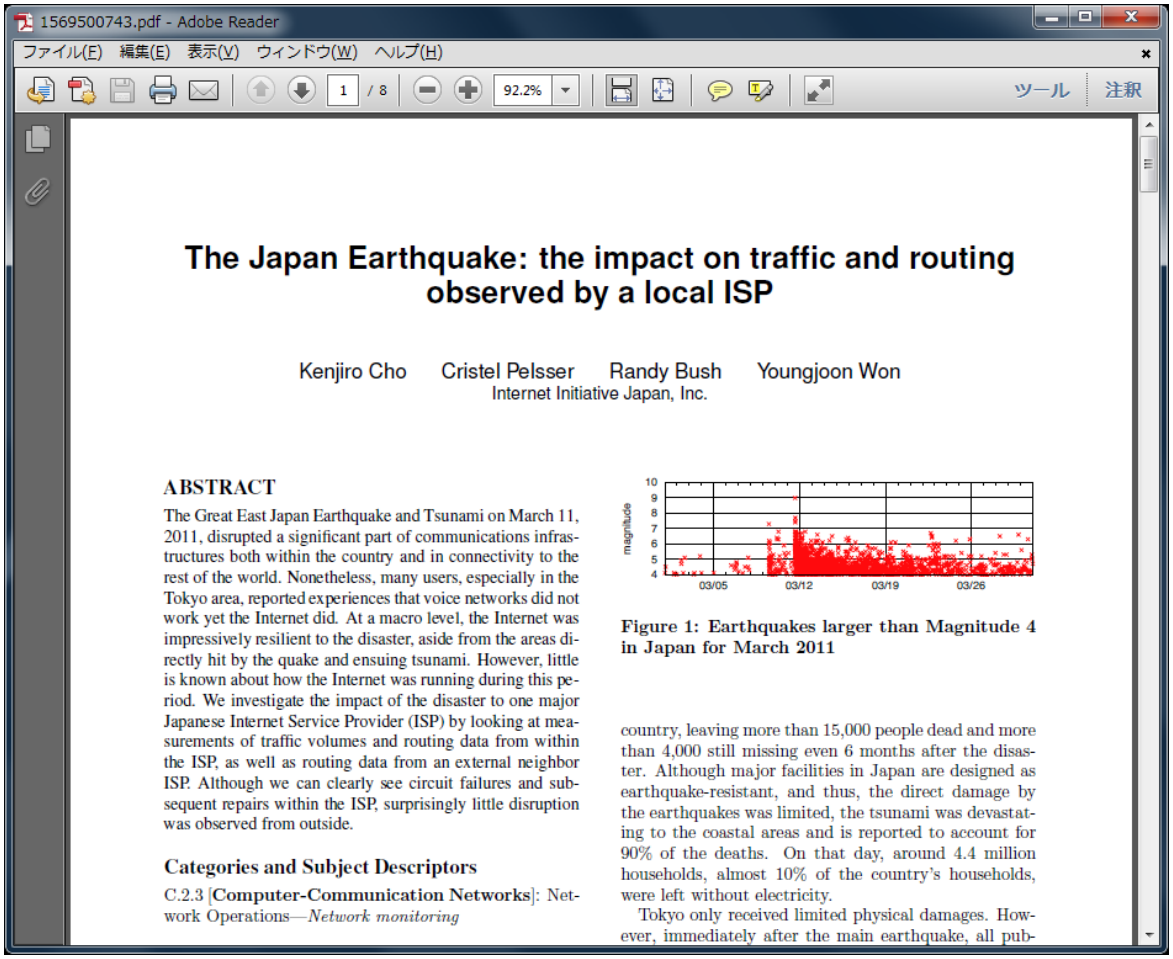


バックボーンの帯域設計

- over-subscriptionだが統計多重効果がばっちり
  - 設備投資の軽減
- 需要＝利用帯域に対してはover-provisioning
  - 迂回帯域の確保
  - 低遅延などの効果



iijlabによる研究事例


1569500743.pdf - Adobe Reader

ファイル(F) 編集(E) 表示(V) ウィンドウ(W) ヘルプ(H)

1 / 8 92.2% ツール 注釈

The Japan Earthquake: the impact on traffic and routing observed by a local ISP

Kenjiro Cho Cristel Pelsser Randy Bush Youngjoon Won
Internet Initiative Japan, Inc.

ABSTRACT

The Great East Japan Earthquake and Tsunami on March 11, 2011, disrupted a significant part of communications infrastructures both within the country and in connectivity to the rest of the world. Nonetheless, many users, especially in the Tokyo area, reported experiences that voice networks did not work yet the Internet did. At a macro level, the Internet was impressively resilient to the disaster, aside from the areas directly hit by the quake and ensuing tsunami. However, little is known about how the Internet was running during this period. We investigate the impact of the disaster to one major Japanese Internet Service Provider (ISP) by looking at measurements of traffic volumes and routing data from within the ISP, as well as routing data from an external neighbor ISP. Although we can clearly see circuit failures and subsequent repairs within the ISP, surprisingly little disruption was observed from outside.

Categories and Subject Descriptors

C.2.3 [**Computer-Communication Networks**]: Network Operations—*Network monitoring*

Figure 1: Earthquakes larger than Magnitude 4 in Japan for March 2011

country, leaving more than 15,000 people dead and more than 4,000 still missing even 6 months after the disaster. Although major facilities in Japan are designed as earthquake-resistant, and thus, the direct damage by the earthquakes was limited, the tsunami was devastating to the coastal areas and is reported to account for 90% of the deaths. On that day, around 4.4 million households, almost 10% of the country's households, were left without electricity.

Tokyo only received limited physical damages. However, immediately after the main earthquake, all pub-

http://conferences.sigcomm.org/co-next/2011/workshops/SpecialWorkshop/papers/1569500743.pdf

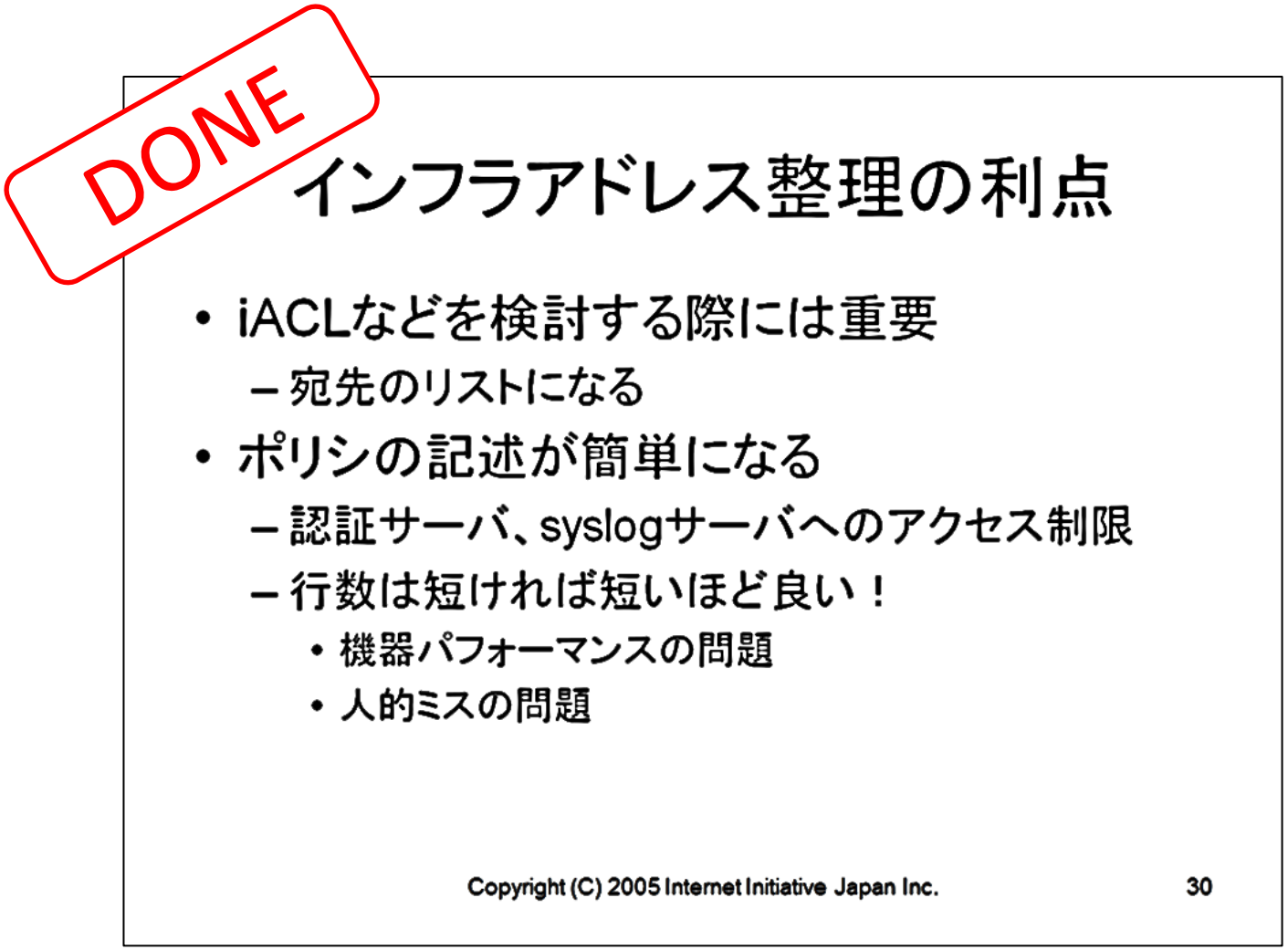
ペーパーのサマリ

- IIJ網内では様々な変動を検出していた
  - 地域的停電によるトラフィック減
  - 回線の切断
    - 国内、対米回線
- 外部のBGP記録でみるとほとんど変化無し
- 冗長化とover-provisioningは非常に良く機能しており、回線切断の影響はほとんど無し
  - ただし、”今回は”耐えられた点に注意

ネットワークを守る

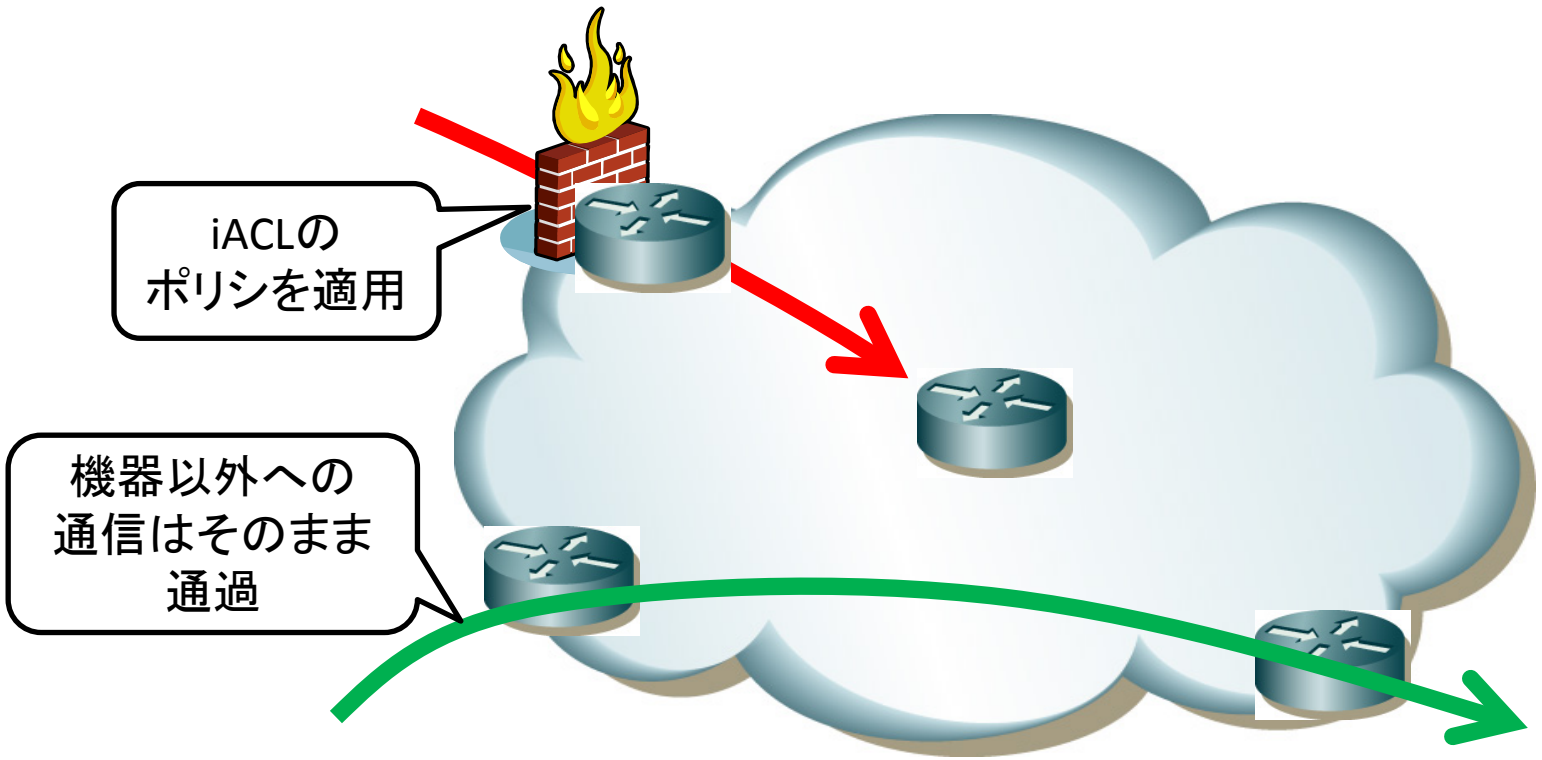
- 機器へのアクセス認証
  - アカウント管理
  - コマンド履歴
- 機器へのアクセスコントロール
  - vtyアクセス
  - OSPFやiBGP
  - snmp, syslog, ntp

Technical WEEK 2005



iACL(infrastructure ACL)

- パケット流入口で機器へのアクセスを制限



iACLとポリシ

- 許可したい通信
  - トラブルシュートに利用する程度の通信
    - traceroute, pingなどなど
    - 主に他のネットワーク運用者向け

→ 完全フィルタではなく、帯域制限を実装
  - 様子を見ながら制限値を煮詰める予定

現状のiACLポリシ

送信元IPアドレス	宛先IPアドレス	制御
IIJインフラアドレス	IIJインフラアドレス	破棄
any	IIJインフラアドレス	帯域制限

IPv6/IPv4で同様のポリシを実装

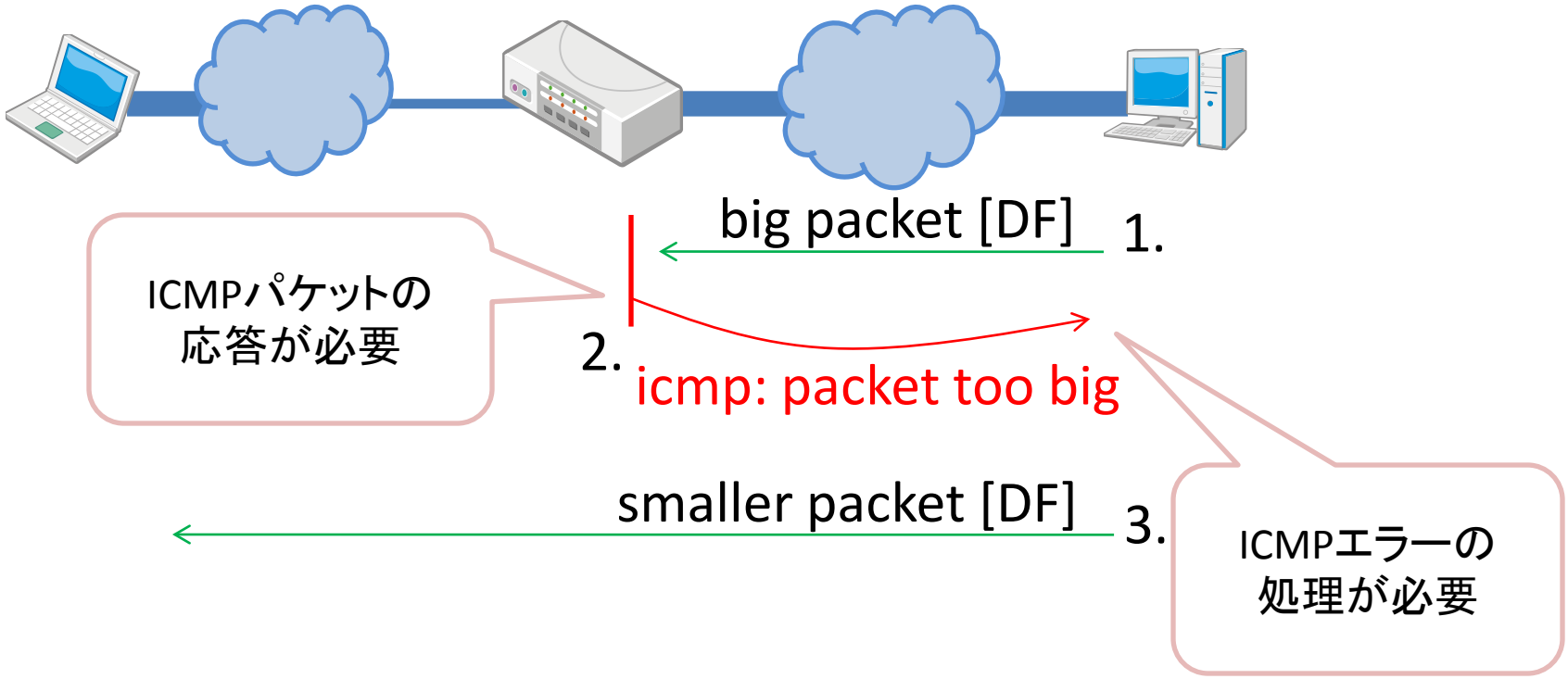
ところでICMP

- Path MTU discoveryなどで必須です
  - 安易なフィルタをしてはいけない
  - IPv4であろうと、IPv6であろうと、です

誤解: × セキュリティのためにICMPをフィルタ

正解: ○ 適切にICMPを処理

Path MTU Discovery



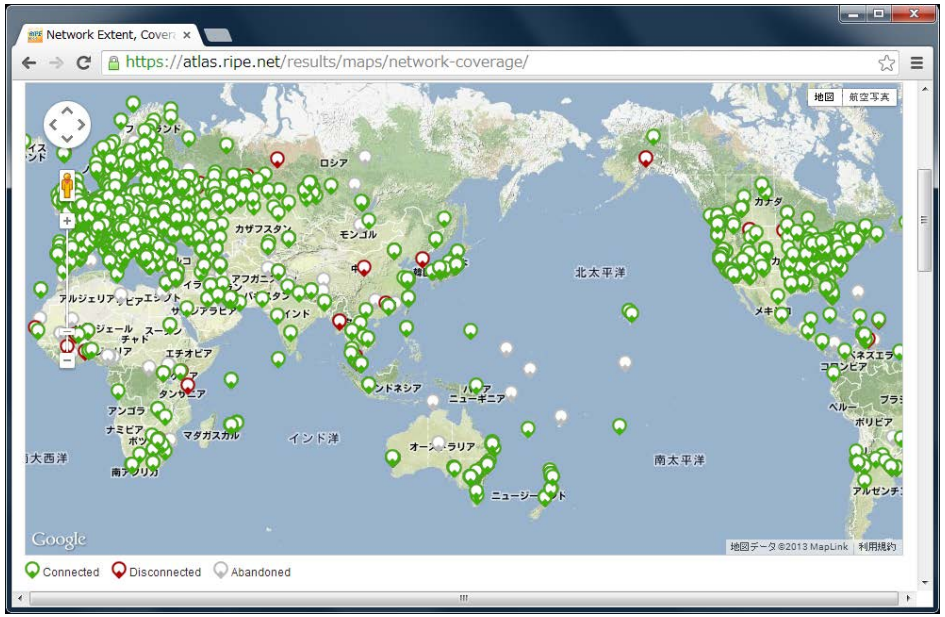
ICMP生成の制限

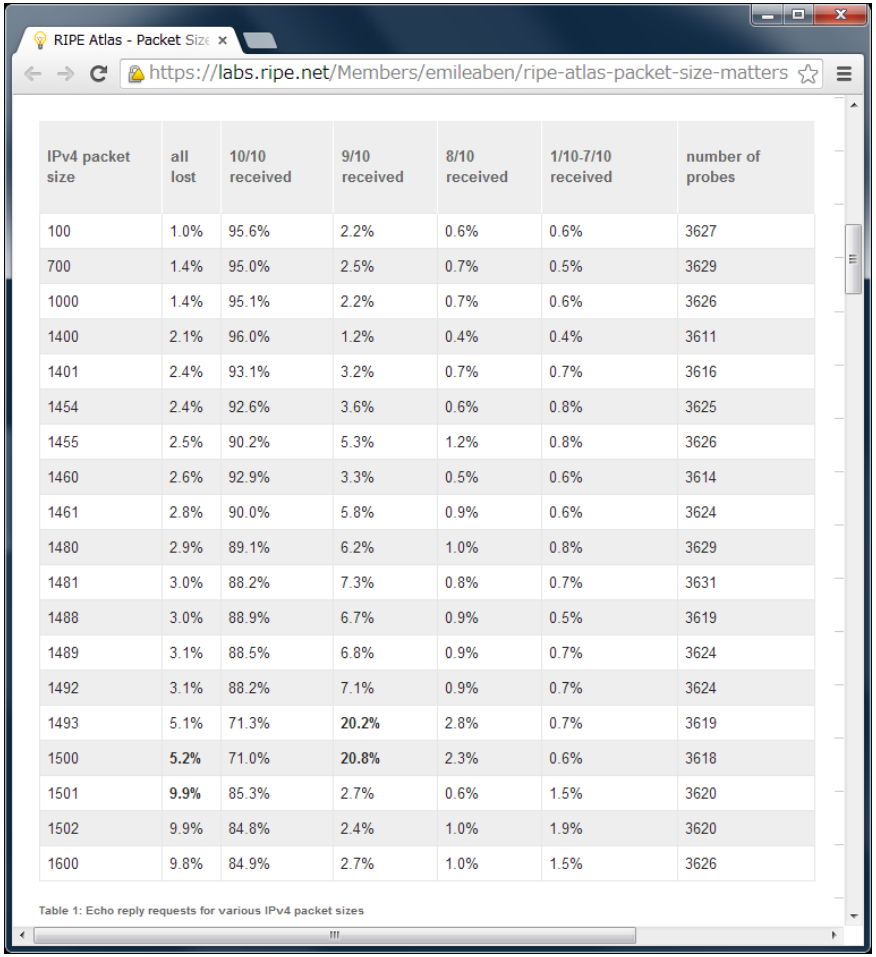
- cisco ios
  - ip icmp rate-limit unreachable 500
    - means icmp errors are limited to one every 500msec
  - ipv6 icmp error-interval 100
    - means icmp errors are limited to one every 100msec
- juniper junos
  - icmpv4-rate-limit {packet-rate 1000;};
    - means max 1000pps for icmp to/from RE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

ネットワーク構成上の注意点

- MTUが小さくなる区間
  - トンネル区間やVPNに要注意
  - 適切にICMPを応答できているか
- 他のネットワークでのMTU
  - 適切にICMPを受信できているか
- TCP MSSの書き換えはうまく動いている
  - PPPoEとともに流行
  - ほとんどの通信がTCPなので、大部分救える

RIPE Atlas - Packet Size Matters





IPv4 packet size	all lost	10/10 received	9/10 received	8/10 received	1/10-7/10 received	number of probes
100	1.0%	95.6%	2.2%	0.6%	0.6%	3627
700	1.4%	95.0%	2.5%	0.7%	0.5%	3629
1000	1.4%	95.1%	2.2%	0.7%	0.6%	3626
1400	2.1%	96.0%	1.2%	0.4%	0.4%	3611
1401	2.4%	93.1%	3.2%	0.7%	0.7%	3616
1454	2.4%	92.6%	3.6%	0.6%	0.8%	3625
1455	2.5%	90.2%	5.3%	1.2%	0.8%	3626
1460	2.6%	92.9%	3.3%	0.5%	0.6%	3614
1461	2.8%	90.0%	5.8%	0.9%	0.6%	3624
1480	2.9%	89.1%	6.2%	1.0%	0.8%	3629
1481	3.0%	88.2%	7.3%	0.8%	0.7%	3631
1488	3.0%	88.9%	6.7%	0.9%	0.5%	3619
1489	3.1%	88.5%	6.8%	0.9%	0.7%	3624
1492	3.1%	88.2%	7.1%	0.9%	0.7%	3624
1493	5.1%	71.3%	**20.2%**	2.8%	0.7%	3619
1500	**5.2%**	71.0%	**20.8%**	2.3%	0.6%	3618
1501	**9.9%**	85.3%	2.7%	0.6%	1.5%	3620
1502	9.9%	84.8%	2.4%	1.0%	1.9%	3620
1600	9.8%	84.9%	2.7%	1.0%	1.5%	3626

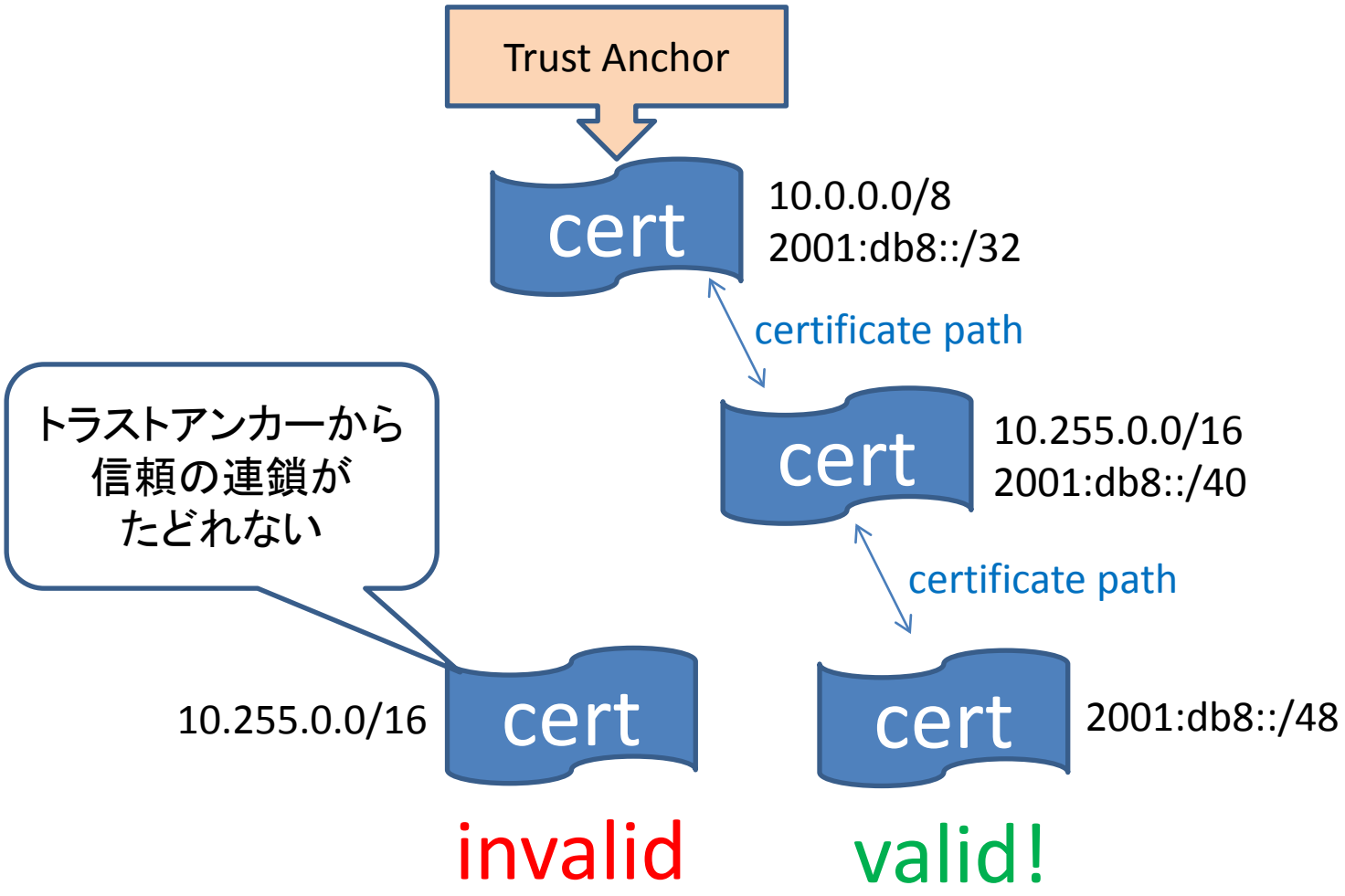
Table 1: Echo reply requests for various IPv4 packet sizes

https://labs.ripe.net/Members/emileaben/ripe-atlas-packet-size-matters

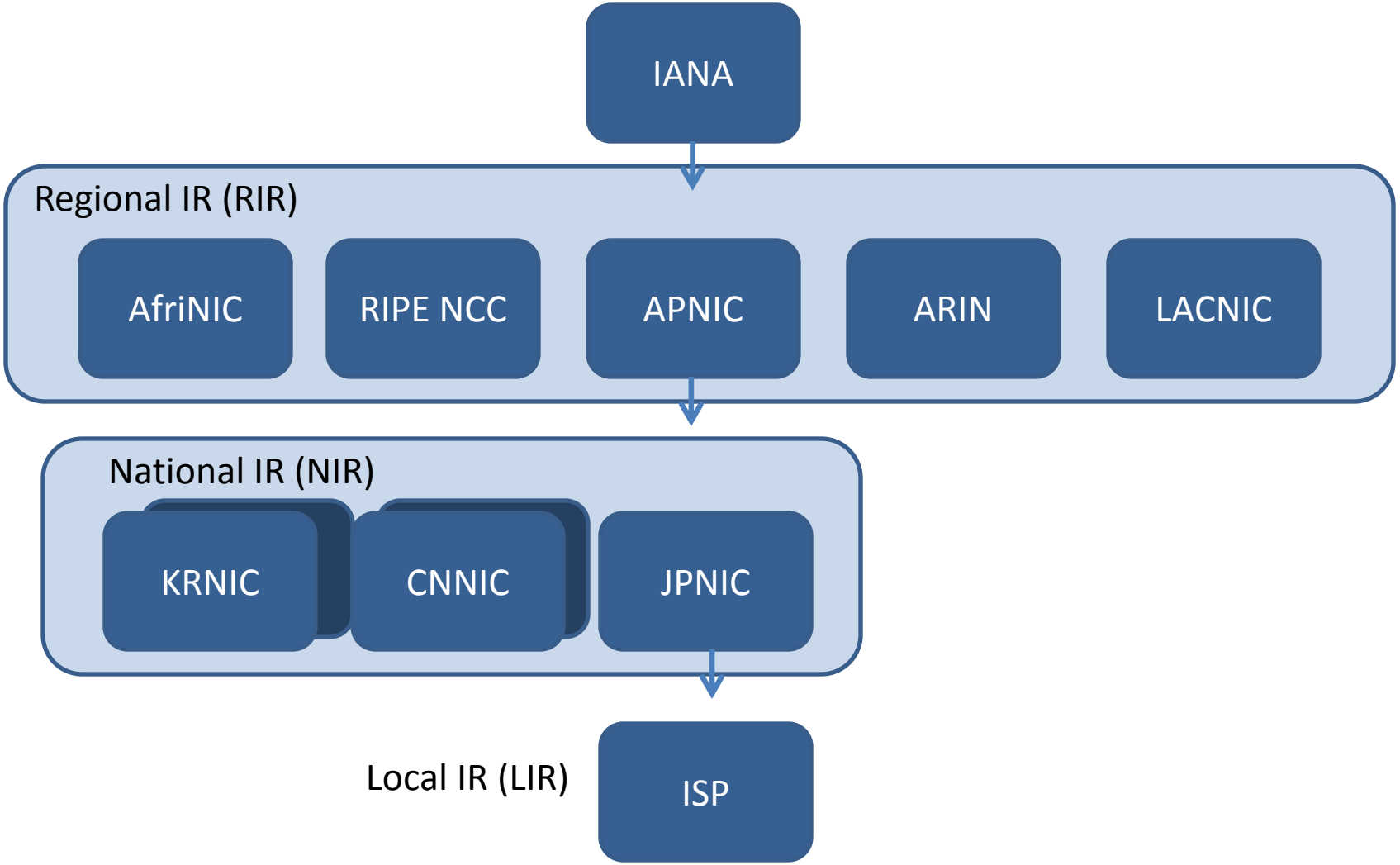
Resource Public Key Infrastructure

- RPKIと呼ばれてます
- インターネット資源のための電子証明書
  - 公開鍵暗号技術
  - インターネット資源の利用権を検証できる

RPKI 構成



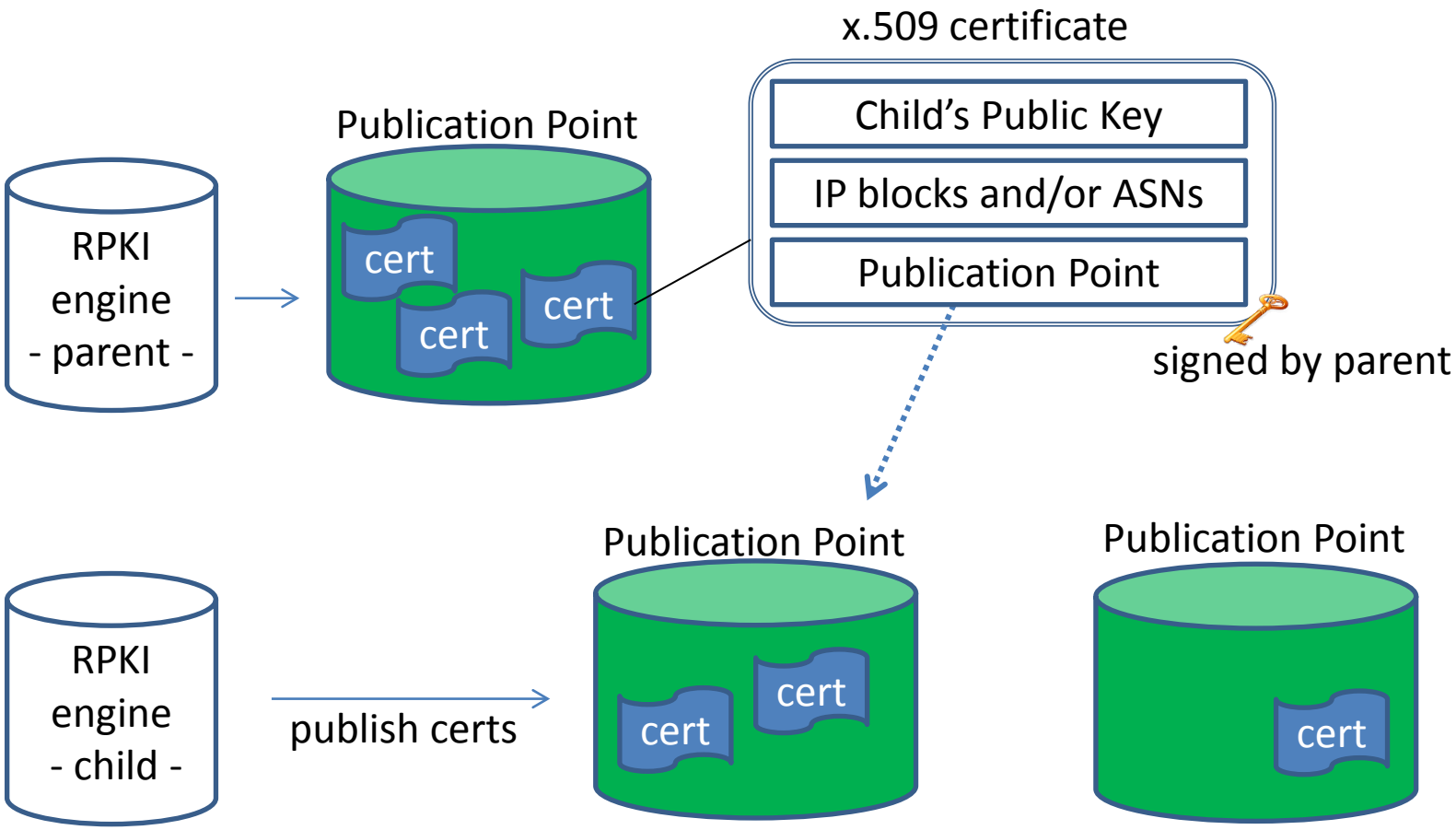
証明書とリソース分配



Trust Anchor Locations (TALs)

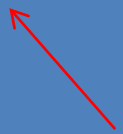
- rsyncのURLと公開鍵情報
  - RFC6490
- 全てのRIRはRPKIに対応済み
  - それぞれのRIRがTALを公開
  - https://www.ripe.net/lir-services/resource-management/certification/rir-trust-anchor-statistics

RPKI publicationサーバ



実際の電子証明書

```
$ openssl x509 -inform DER -text -in nUoKQJmirKA2dIS40zY34cs7tKc.cer
 :
Subject Information Access:
        CA Repository - URI:rsync://rpki.apnic.net/member_repository/XXX/XX/
 :
sbgp-autonomousSysNum: critical
                Autonomous System Numbers:
                  2497-2528
                  2554
 :
sbgp-ipAddrBlock: critical
        IPv4:
                  1.0.16.0/20
                  1.0.64.0/18
:
```



publication point

Route Origin Attestations (ROAs)

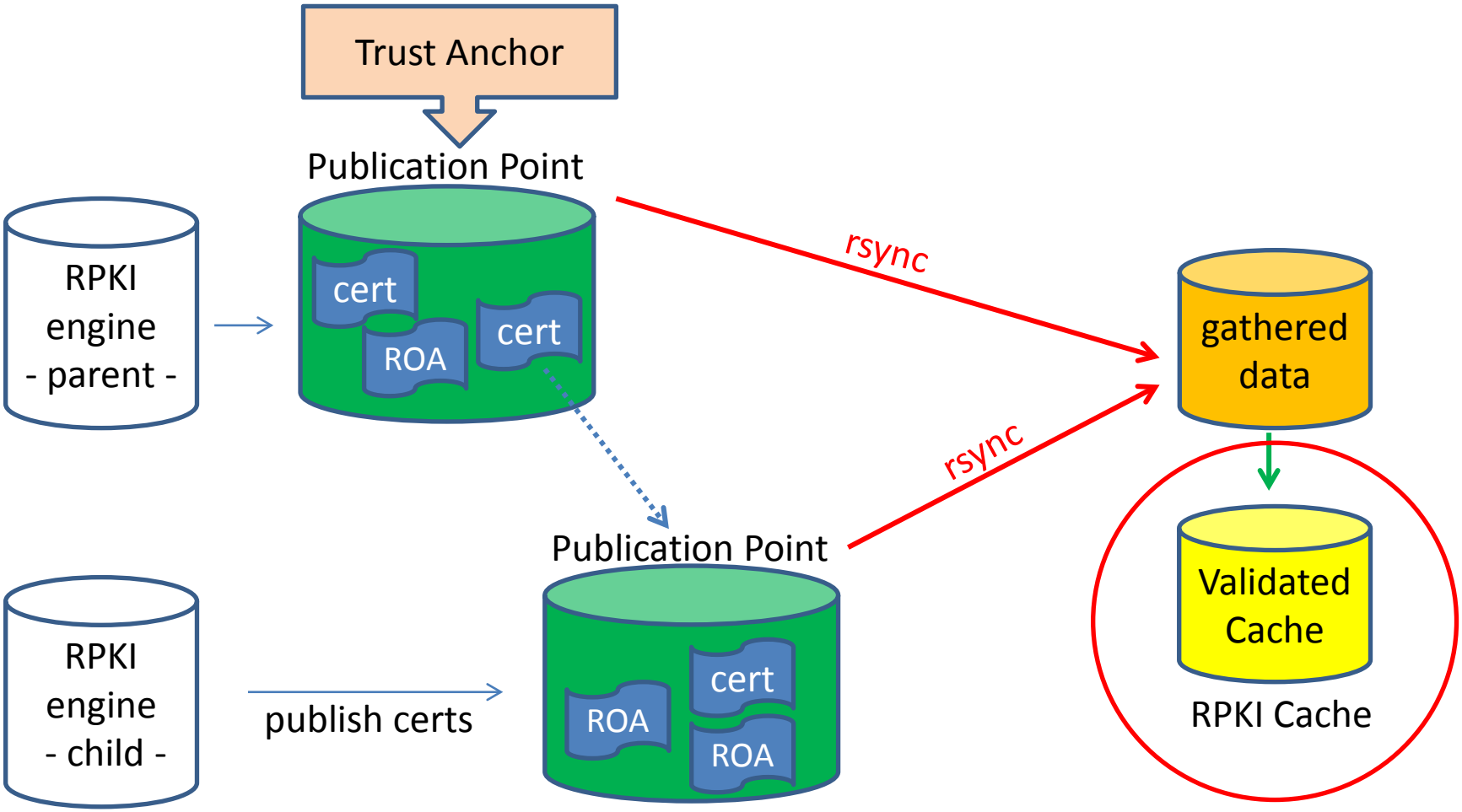
- AS番号とprefixを含む電子証明書
  - そのASから該当のprefixを広報することを宣言
    - IRRにおけるrouteやroute6 objectと同等
  - IPアドレスブロックの保持者がROAを生成できる
- ‘maximum length’オプション
  - 広報するprefixの最大prefix長を宣言
  - 細かい経路の広報を行う時に利用できる

ROA

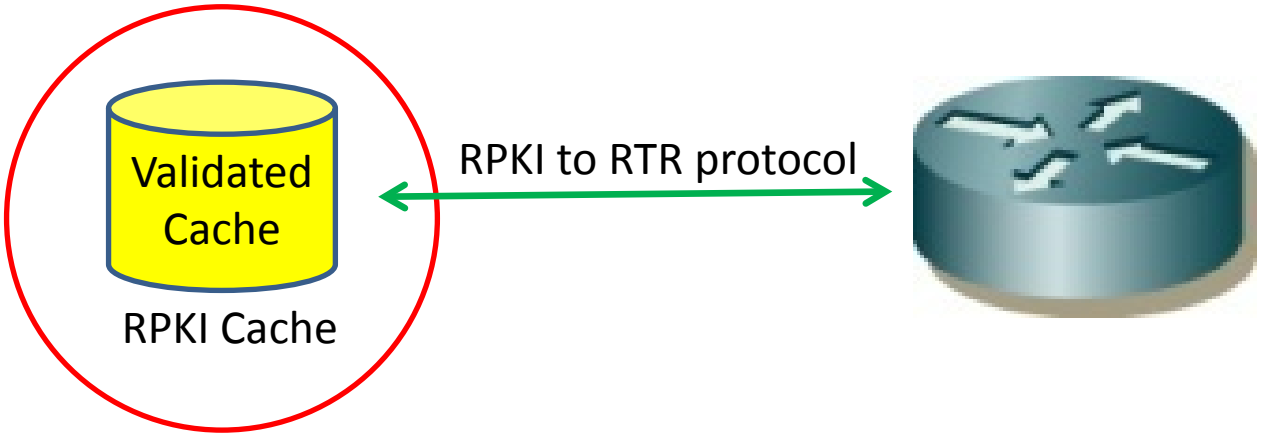
```
$ print_roa FksMMjbAOUZnFeuDv2yZmcAXJeY.roa
 :
asID:          2497
 addressFamily: 2
    IPaddress: 2001:240::/32
```

複数のASから広報するために、複数のROAを発行することもできる

RPKI cache



経路生成元の検証

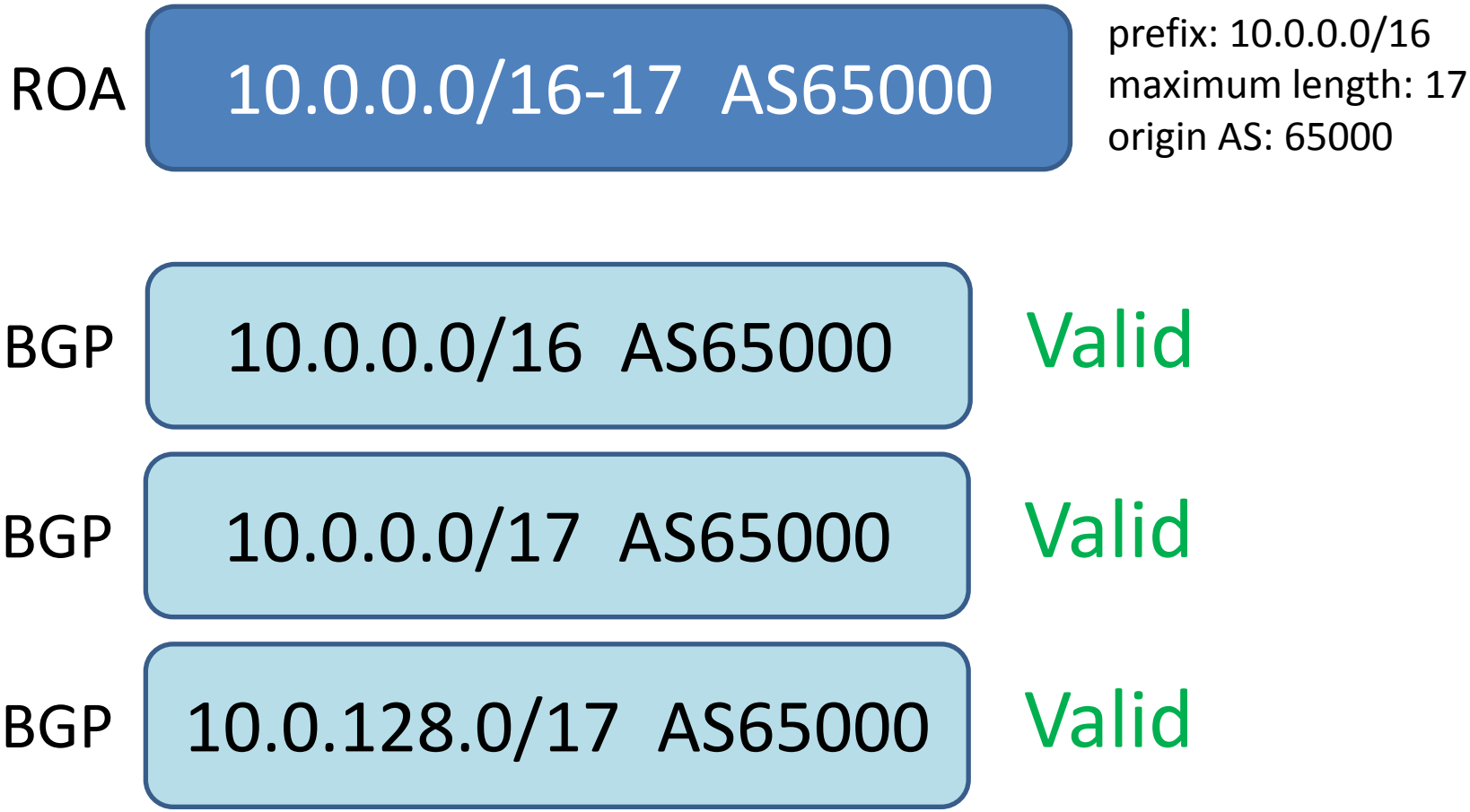


- ルータはRPKI CacheからROAの情報を取得
  - RPKIの電子証明書はRPKI Cacheで検証済み
- ルータのBGPで、経路情報とROAを突き合わて比較を行う

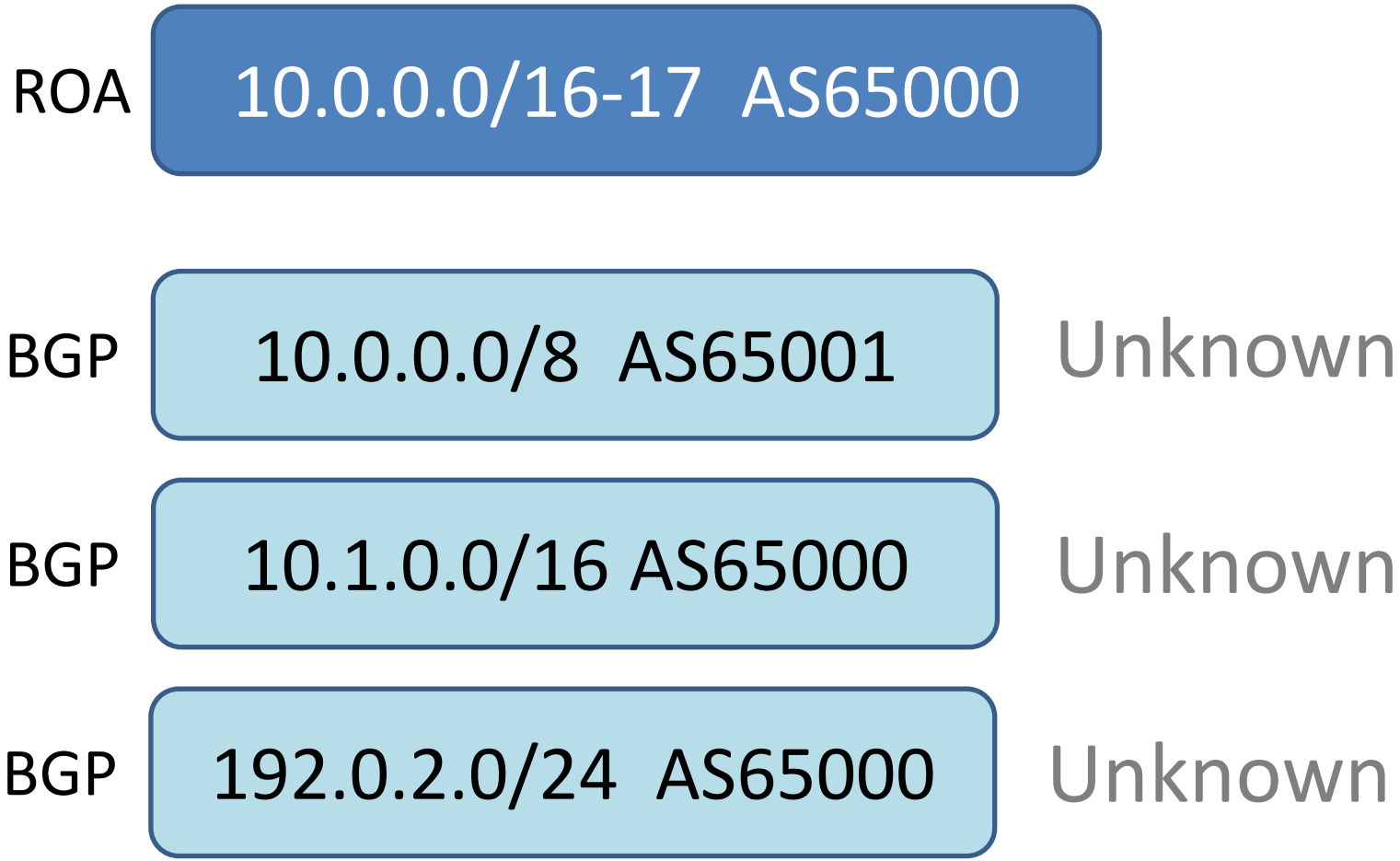
検証結果

- Valid
  - prefixとASに合致するROAが見つかった
- Unknown (Not found)
  - prefixに該当するROAが無かった
- Invalid
  - prefixに合致するROAが見つかったが、AS番号あるいはprefix長がROAと合致しない

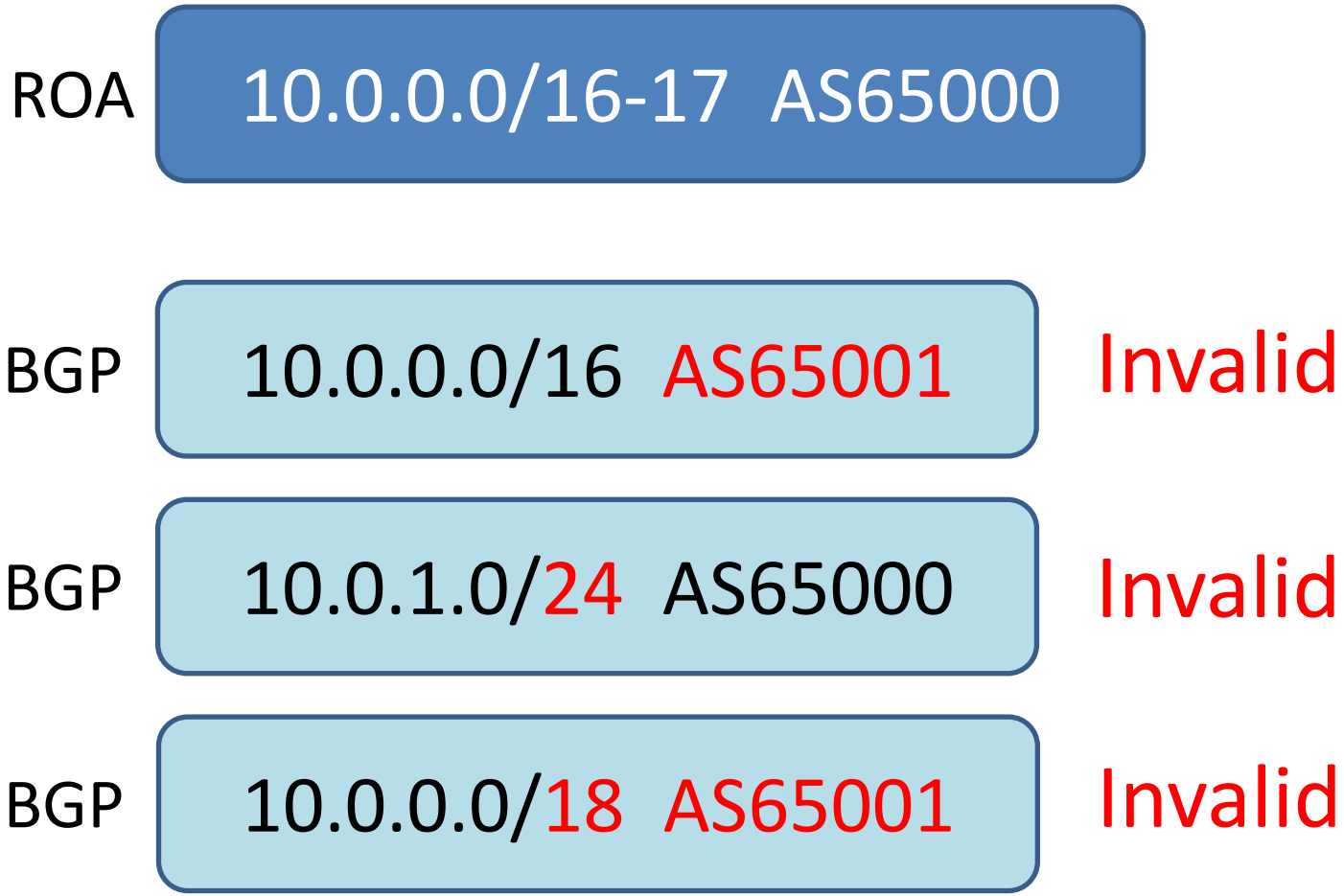
example - valid



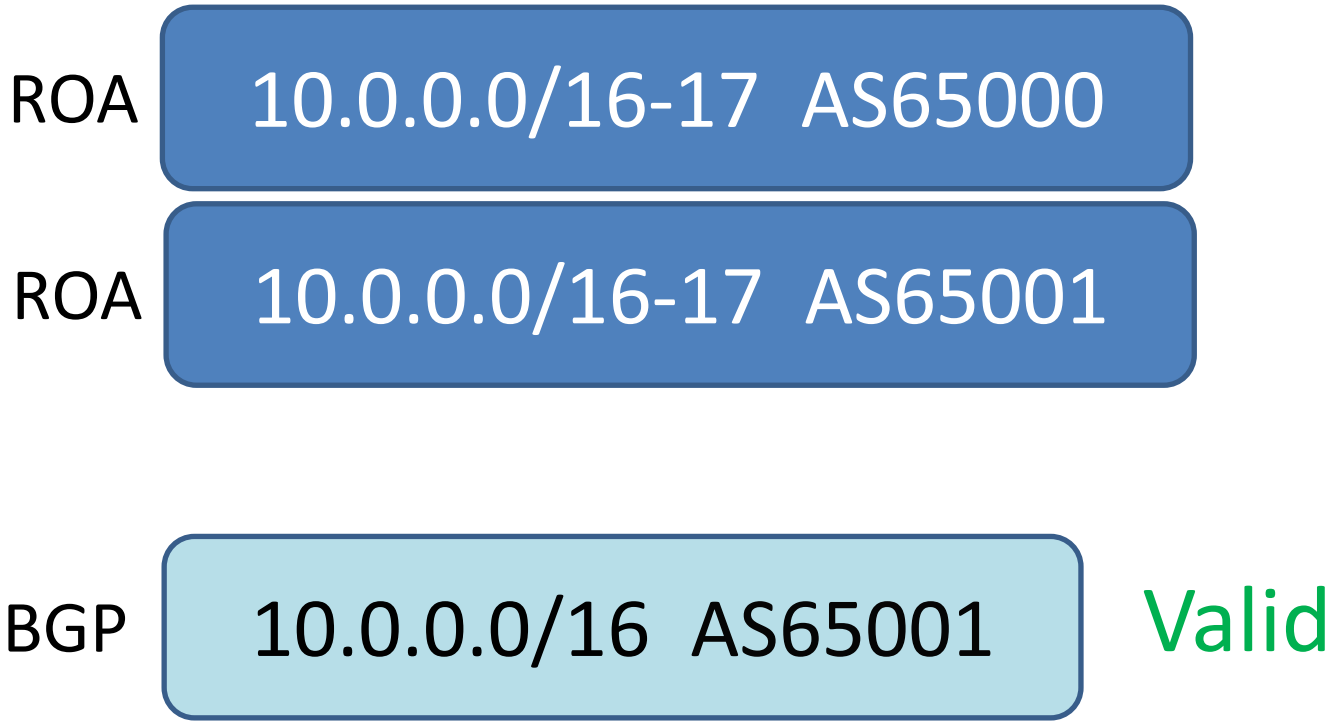
example - unknown



example - invalid



example - multiple origin ROA



RPKI現状

- 不正な経路広報の伝搬を予防しうる

- ソフトウェアは充実してきている
  - ルータのサポート
  - 証明書の検証環境
- 日本ではJPNICの対応待ち
  - テストベッドなどに参加して検証中

まとめ

- シンプルなポリシ
  - 冗長設計とover-provisioning
- 攻撃への備え
  - iACLの導入、ただし運用の利便性は確保
- 経路認証
  - RPKIの検証等に協力中