IPv6 PMTU Discovery Blackholeの盲点

株式会社ビーコンエヌシー

データセンター事業部

國武　功一

自己紹介

@kunitake

github.com/kunitake

２児の父親

明日長女の５歳の誕生日なのに、高松までうどん食べに来た父を許しておくれ……

この発表の動機

- いつまでたってもPTMU Discovery Blackholeがなくならない
- 起きてても意外と気づいてない
  - 偉そうに語ってる自分もミスったorz
  - ごく最近でも、IPv6に強そうな某社がやってた模様

Agenda

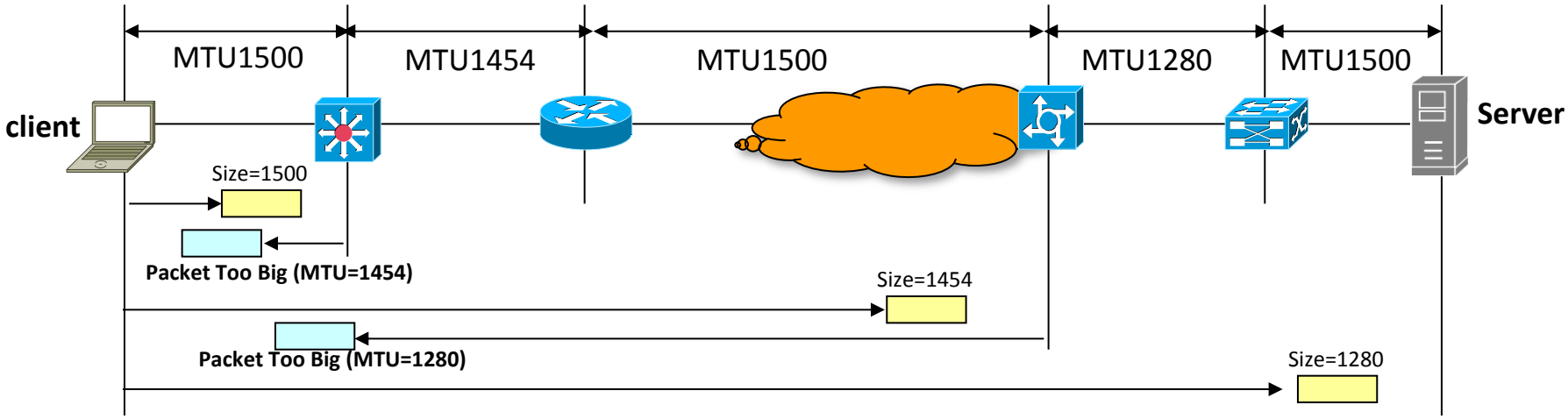
- PMTU Discoveryおさらい
- よく知られた事例
- Filteringの盲点
- そもそも Too bigを発生させない方法
- 最後に

Path MTU Discoveryおさらい

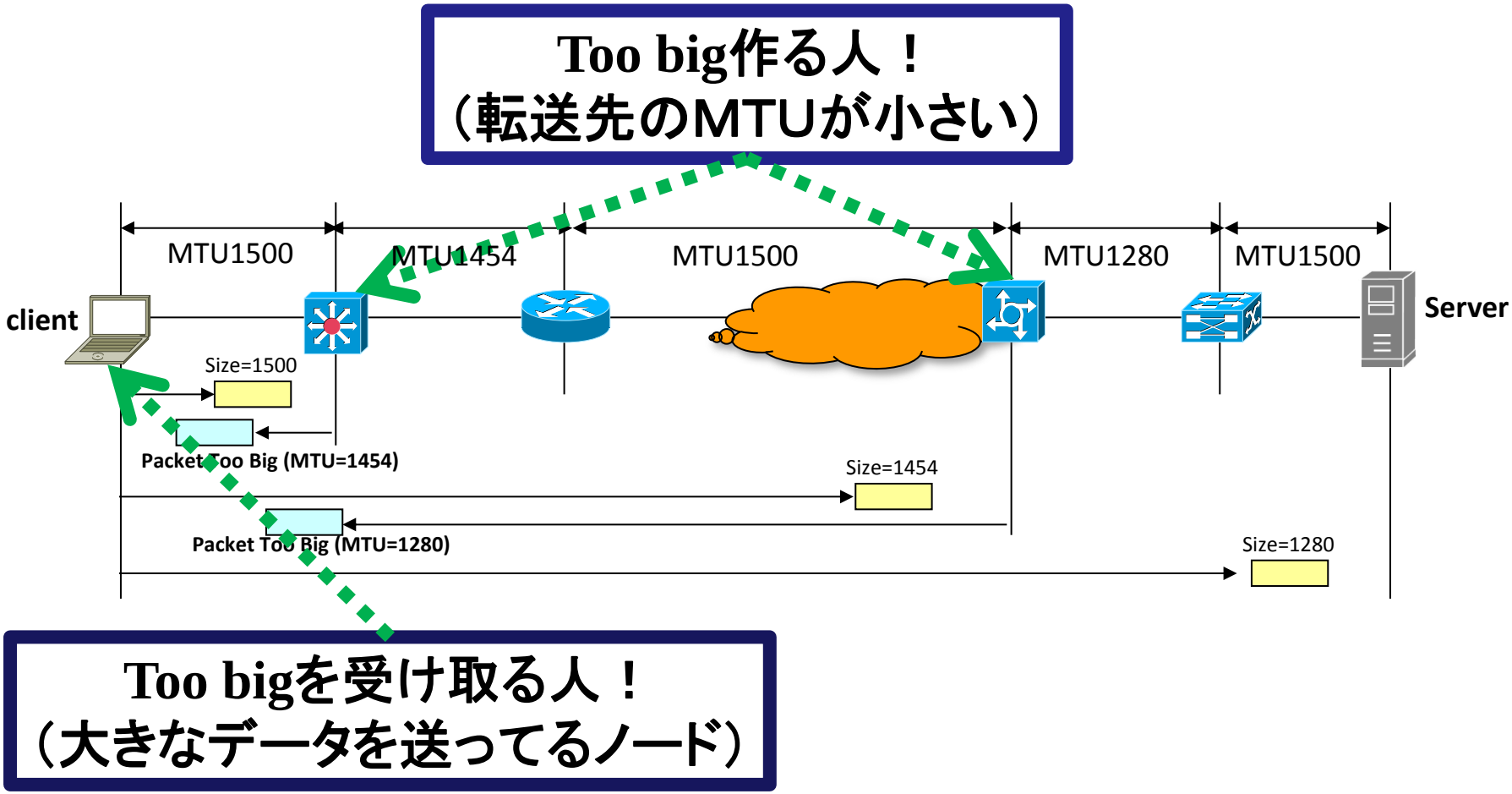
Path MTU Discoveryおさらい

- IPv6 では中継ノードでフラグメントしない（始点ノードが実施）
  - IPv4 ではルータ等の中継ノードがフラグメントを実施
  - 送信パケットに対する ICMPv6 Error Message を受信時、MTU を変更
    - 最初のリンクのMTU が初期値
    - ICMPv6 Packet Too Big Message 受信時、始点ノードでフラグメントして再送
  - IPv6最小MTU は、1280byte
    - L2 SWのMTUにひっかかった場合は破棄される
    - Path MTU Discovery の実装が難しいノードは 1280byte 固定

Path MTU Discoveryおさらい



Path MTU Discoveryおさらい



Blackholeの主な原因

1. Too big パケットが作れない
2. Too bigパケットが受信・転送できない

Too big 受け取るのはだれ？

- コンテンツを送信する側
  - ウェブサーバ
  - メール送信者(大きな添付ファイルとか)
  - Dropbox的ななにか

PMTUD Blackholeなぜ困る？

- IPv4と違って、IPv6では途中ルータがパケットをフラグメントすることは禁止されており、PMTU Dicoveryが動作することによる再送を期待しているため。

Too bigが届かないと、通信ができない！

実際の切り分けデモ

- 一発でわかるワンパターン体験
- 今日日のエンジニアは、telnetすら使わないとか人から聞きましたが…

僕らには
Ｈａｐｐｙ Ｅｙｅｂａｌｌ
があるじゃないか！！

↑TCPのセッションは張れてる
のでだめです…

つまり…

Too bigを必ず届ける、受け取る！

or/and

Too bigを発生させない！

がIPv6通信には必須

よく知られた事例

ここを疑え！

- L3’s icmp rate limit
- Firewall Policy
- LB構成でToo bigが

1. Too bigパケットが作れない

別の意味で、ここを疑え！（後述）

2. Too bigパケットが転送できない

ここを疑え！

- 頑張っちゃって link local addressのみでネットワーク作っちゃった、かつ異なるMTUサイズを混ぜちゃった

1. Too bigパケットが作れない

RFC4291: Routers must not forward any packets with link-local source or destination addresses to other link.

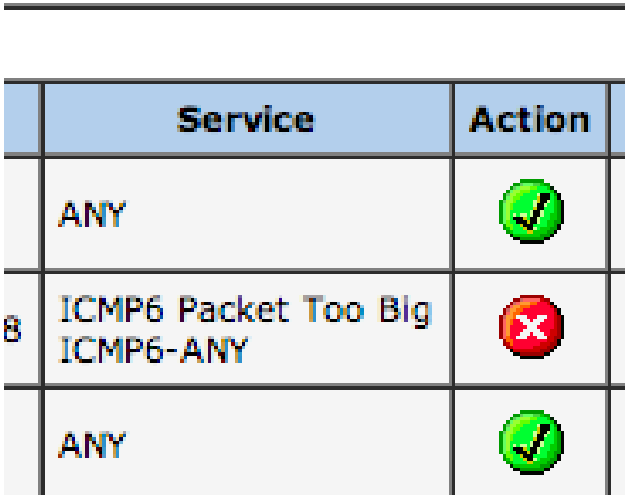
- UPS/UTMのおせっかい <- TODAY

2. Too bigパケットが転送できない

Filteringの盲点

Firewall見直し、これで安心？

- 実はFirewallのポリシーでは、事実上、Too big は落とせない（フロー上、自動的に許可される）



	Service	Action
	ANY	✓
8	ICMP6 Packet Too Big ICMP6-ANY	✗
	ANY	✓

なんと、こんな設定書いてもToo bigは落とせない…

iptables/ip6tablesでも

-A INPUT -m state ¥

--state ESTABLISHED,RELATED ¥

-j ACCEPT

RELATED

meaning that the packet is starting a new connection, but is associated with an existing connection, such as an FTP data transfer, or an ICMP error.

では、誰が落としてるの？

- こんなメッセージに見覚えはありませんか？

[00001] 2014-07-17 19:00:00 [Root]system-critical-00436: Large ICMP packet! From 2001:db8:ffff::222 to 2001:db8::80, proto 58 (zone Untrust, int ethernet0/1). Occurred 6 times.

たとえばこんな機能

- [ScreenOS] Large Size ICMP Packet (size > 1024) in IPv6 environment.

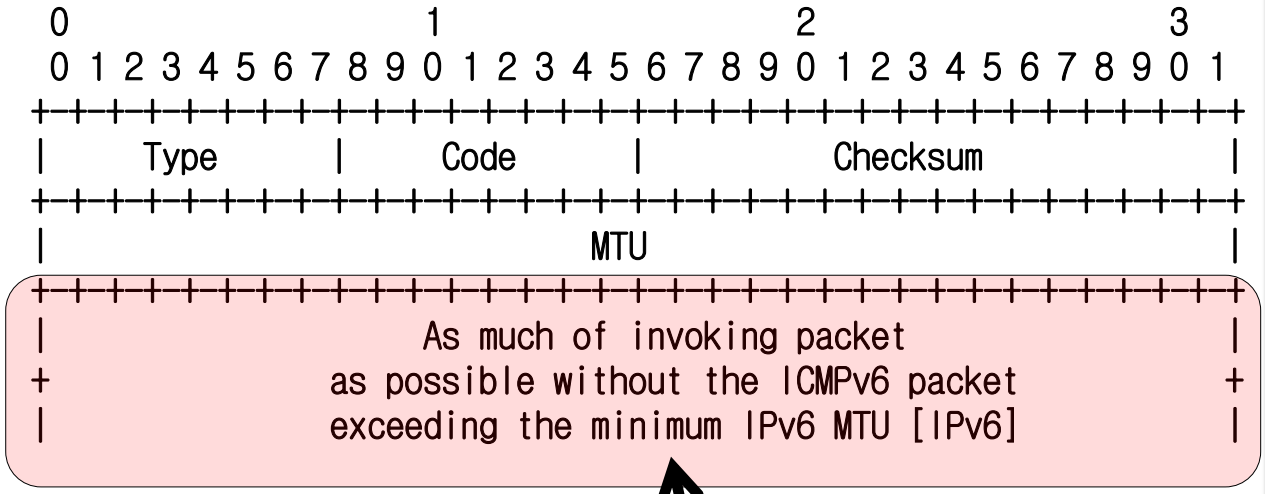
  http://kb.juniper.net/InfoCenter/index?page=content&id=KB26473&actp=RSS

  （**http://juni.pr/QJCruH**）

  多くのICMPパケットは大きくないため、1024バイト以上のICMPパケットを攻撃パケットやLoki (ICMP Tunnel)の通信などとみなして、たたき落としてくれる素敵機能。

なぜひっかかるのか？

3.2. Packet Too Big Message



こいつだ！

Too bigのパケット長

- 1280byte以下であることは仕様で決まっているが、どこまでパケットを頑張って詰め込むかは、実装依存

  今回問題になったFirewallでは、1000byte

  1000 byte < 1024byte....

１０秒でできるチェック

- Rancidリポジトリをexport
- Find一発。ね？簡単でしょ？

```
# find . –print0 |xargs -0 grep icmp-large -l
```

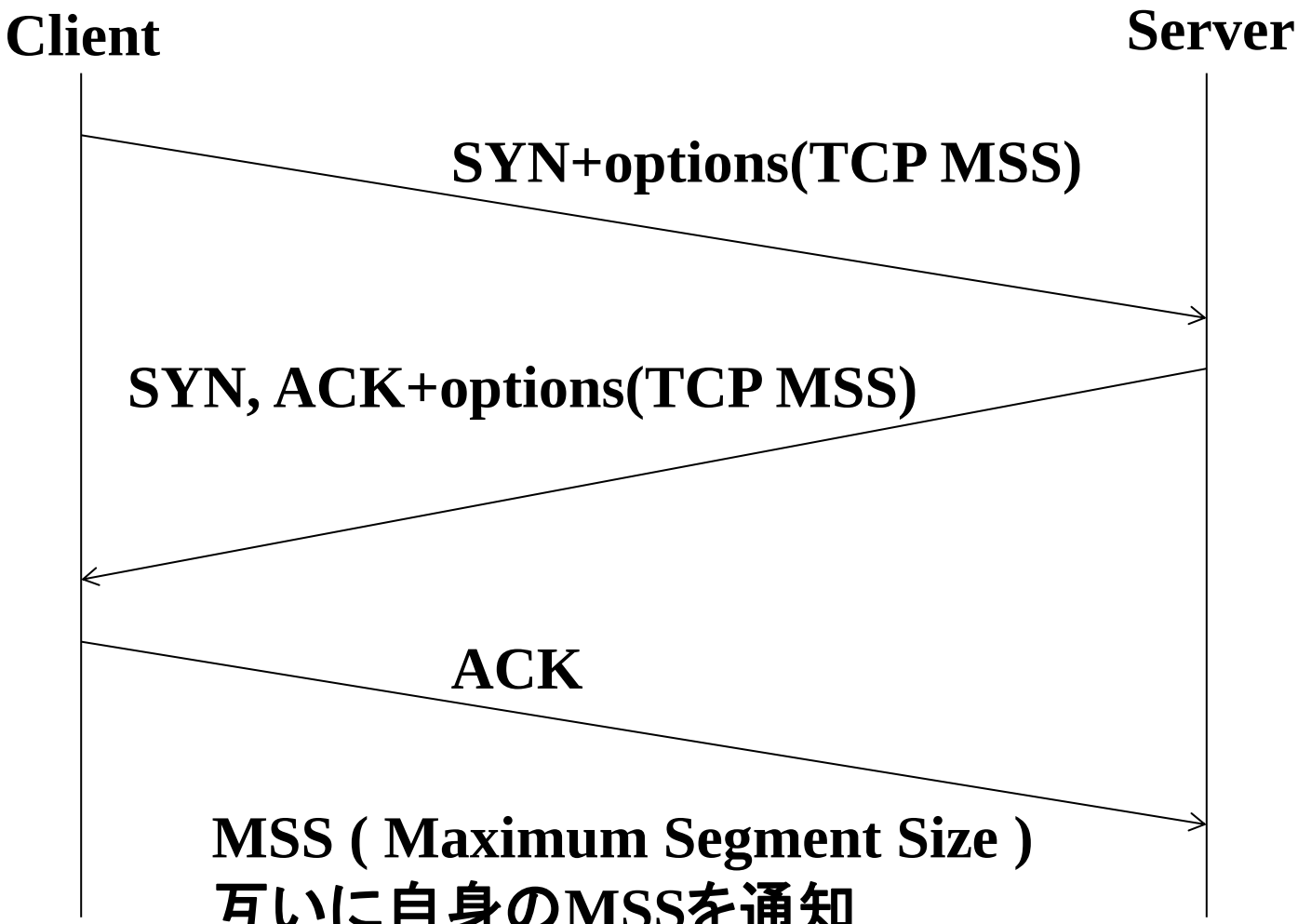
氷山の一角か

- 他の製品でも、UTM/IPSでこんな感じのおせっかい機能がありそう
- エンタープライズ向けのFirewallは、単純にポリシーだけを見ていると、足元をすくわれることも

今回の例では、IPv4/IPv6で設定が共通なので、最初は問題なくても、IPv4由来で途中で有効化すると、気付かないうちにIPv6で問題が発生する。またフィルタリングポリシーよりも優先される。

そもそも...
Too bigを
発生させない！

TCP 3way handshake復習



ということは…

- TCP MSS optionが動作する前提で、サーバ側のMTUを1280 octetsに設定しておけば、Packet Too bigは発生しない
- MTUサイズ変更の境界にいることの多いブロードバンドルータが TCP MSS Clampingに対応した場合も、Packet Too big は発生しない(ただし、Path MTUを知っているわけではないことに注意)

TCP MSS Clamping 嬉しさいろいろ

- Path MTU Discovery blackholeの回避
- サーバにおけるMTUキャッシュの削減
- TCP再送の抑制
- 輻輳回線での Too bigのドロップ時の影響回避
- IPv4で上手くやれてきた実績

TCP MSSで解決？

- 救えるのはTCPだけ。UDPは、もともとフラグメントを嫌って short packetでやり取りしていることが多いが、留意する必要がある
- サーバ側で MTU＝1280に設定することで、スループットが落ちる可能性
- 結局 Path MTU Discovery Blackhole問題の隠ぺいでしかない。

なぜ気づけない？

- テストテストテスト
  - Too bigを通すにはセオリーなだけに、設定を見直すだけで安心している？
  - 運用経験がまだ浅い

    ワンパターンな問題だと気づけるのは、場数を踏んでこそ
  - 監視がイケてない
    - ポートチェックだけでは、検知できない

再発防止策

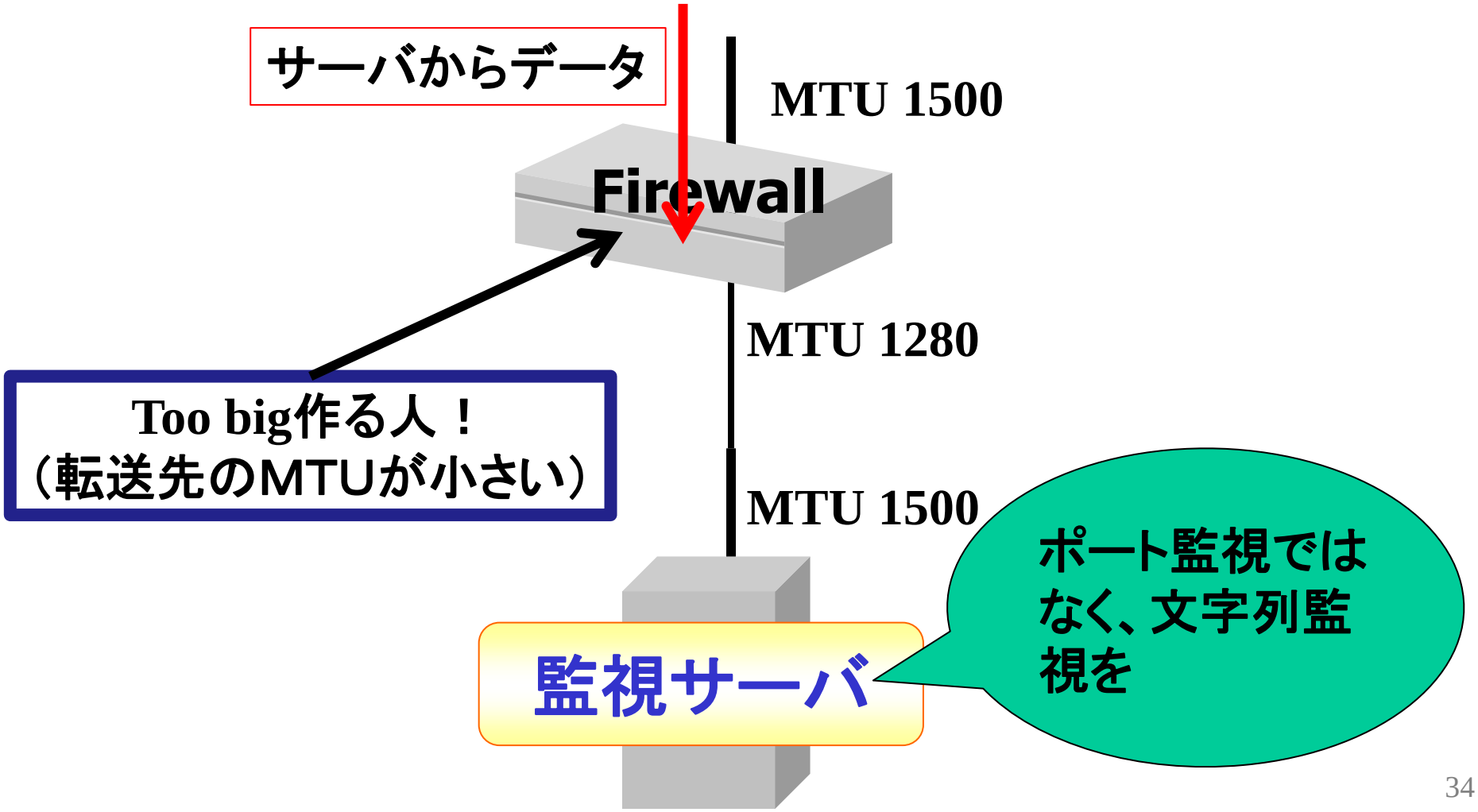
- 設定変更時

  いま入れようとしている設定が、IPv6にも影響を与える設定なのか、与えるなら、悪い影響を与えないかを検討する必要がある。

- ＭＴＵが小さくなるクライアント環境を用意し、継続的に監視・導入前テスト

  – ポート監視だけではだめ。１２８０バイト以上となるコンテンツが取得できることのチェックが必要

たとえばこんな監視環境



こんなことしちゃだめよ？

- MTUが小さくなる環境構築
  - クライアント環境で以下を実行するだけ

```
# ifconfig en0 mtu 1280
```

これは、Too bigを受け取れるかのテストにはならず、
TCP MSSによる調整が発生する。
逆にいえば、これで問題解消する場合は、
Path MTU Discovery Blackholeが原因

最後に

- 実はPMTU Discovery Blackholeを引き起こす環境を作るのは難しい。
- 起きてしまった時の、PMTU Discovery Blackholeの罪深さ

  TCPのセッションは普通に張れるので…

  - happy eyeballs では救えない
  - IPv4にフォールバックもしない

Special Thanks to

- ネットワークチーム(弊社事例解析)
- 資料作れと背中を押した人 by @mikiT_T
- ネタの提供 by @ taji_314159265

参考

- JANOG33.5 Interim Meeting
  - フラグメンテーションの今後を考えよう

    http://www.janog.gr.jp/meeting/janog33.5/doc/janog33.5_fragmentation.pdf
- 第6回IPv6オペレーションズフォーラム
  - IPv6 Path MTU の傾向と対策 - MTU が小さくて何が悪い!?

    https://speakerdeck.com/tsahara/ipv6-path-mtu-in-the-world-and-japan