



# やっておこう？ あんな試験 こんな試験

JANOG37.5 Interim Meeting

2016/4/15

伊藤忠テクノソリューションズ株式会社  
平谷 創



伊藤忠テクノソリューションズ株式会社

## 自己紹介

- 今日は、CTC(SIer)としてお話しさせていただきますが、こんな過去を持っていたりなんかします。

- 某社にて、某金融機関のインターネットバンキング立上げに従事



- こんな会社でBackbone運用や、DC系NW運用  



- こんな会社でBackbone運用や、DC系NW運用  

- いまここ   
Challenging Tomorrow's Changes

- 将来 ???

## はじめに

昨今、クラウド、SDN、WB、オーバーレイ技術など華やかなキーワードが巷を賑わしておりますが・・・

それらを支えているネットワークの基本的な部分って、ちゃんと機能していますよね？

我々は、今なおその基本的な部分を検証という形で支えています。

そこで本日は、日々、キャリア様、コンテンツプロバイダ様、etc 向けの検証を実施している経験と、少々の個人的な実体験に基づくお話をさせて頂ければと思います。

華麗な姿ではなく、水中の  
縁の下的なお話をします



## 検証項目

- 我々が、一番シビアな環境での検証を実施する場合以下のような検証を実施しています。
  - 単体検証試験
  - 構成(結合)検証試験
  - 運用試験
  - etc
- 細かい情報はお見せできませんので、簡単に説明しますと

## 単体検証試験

- 試験対象機(HW別、OS別)あたり、約300項目近い試験を実施
- 試験対象次第では、省略する項目もある

| 大分類         | 中分類                                               | テスト内容                                         | 備考     |
|-------------|---------------------------------------------------|-----------------------------------------------|--------|
| HW          | Chassis,RP,Fabric,LC,Optics,Env,etc               | 各種OIR,Reload,shut,no shut,etc                 | 約30項目  |
| Mgmt        | Login,AAA,SNMP,Syslog,NTP,(T)FTP,etc              | 各機能における正常動作確認                                 | 約30項目  |
| Control     | ARP,RIB,FIB,Static,OSPF,BGP,LDP,TE,L2 VPN,LAG,etc | 機能動作確認,multipath,設定追加削除動作,Timer,BFD,GR,負荷,etc | 約190項目 |
| Aging       | w Operation,wo Operation,NDR                      | 24H,48H,コマンド連続投入,etc                          | 約10項目  |
| Performance | 1byte Sweep                                       | 異LC,同LC,Label有無,etc                           | 約20項目  |
| Specific    | Security,QoS,xFlow,script,etc                     | ACL,PortScan,Counter,QoS,etc                  | 約30項目  |

## 構成(結合)検証試験

- 試験構成(実環境を極力模擬)により、試験項目は大きく異なる。

| 大分類         | 中分類                   | テスト内容                                                   | 備考 |
|-------------|-----------------------|---------------------------------------------------------|----|
| Link(物理)    | 1本Link全断,LAG全断,LAG1本断 | 通信loss有無, 通信断時間計測<br><b>装置によっては、LAG本数によつては偏りが出る場合も！</b> |    |
| Link(サイレント) | 1本Link全断,LAG全断,LAG1本断 |                                                         |    |
| Node(筐体)    | 電源Off/On, Reload      | 通信loss有無, 通信断時間計測                                       |    |
| Node(LC,RP) | LC/RP抜去, SwitchOver   | 通信loss有無, 通信断時間計測                                       |    |

- 経路負荷

| Protocol | Type   | Prefix                     | 備考                                   |
|----------|--------|----------------------------|--------------------------------------|
| BGP      | eBGP   | IPv4 : 1M弱 / IPv6 100K弱    |                                      |
| OSPF     | Type 1 | OSPFは、状況に応じて実環境の倍程度        |                                      |
| OSPF     | Type 2 | 同上                         |                                      |
| OSPF     | Type 5 | 同上 + 実環境に応じてmetric 1, 2 混在 | <b>Metric-type 見落とすとLoopすることもある！</b> |

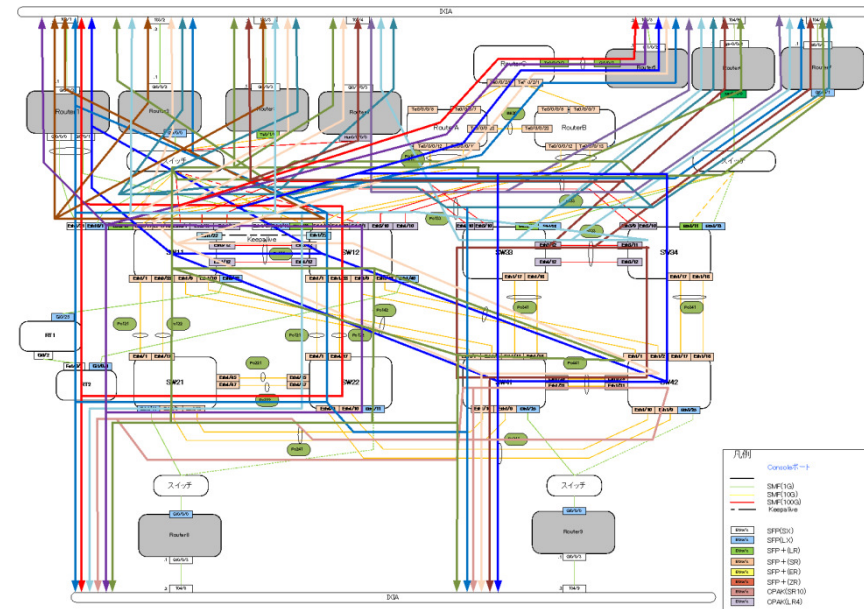
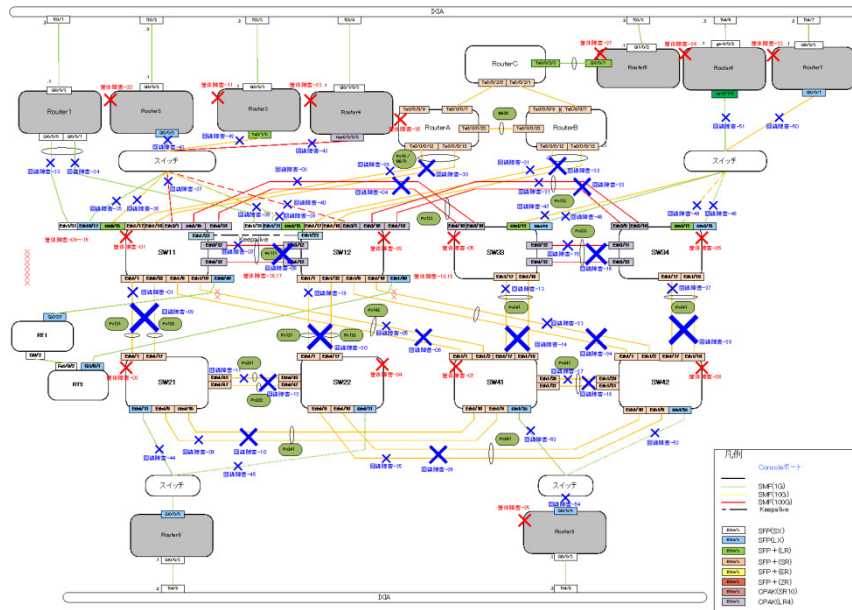
## 構成(結合)検証試験

### ● 通信フロー

| 種別      | 内容                         | 備考                     |
|---------|----------------------------|------------------------|
| 通信経路    | 基本的に想定される全end-to-end通信経路模擬 |                        |
| IPレンジ   | UDP:若番IP～老番IP vs 若番～老番IP   | <b>Convergenceに影響！</b> |
| パケットサイズ | IMIXか、MIN/MAX              |                        |
| PPS     | 1フローあたり、10000PPS程度が多い      | 1パケloss = 0.1msec      |

- 各通信経路毎に以下の設定を行うため、フロー数は膨大に！
  - Src/Dst IPレンジの設定
  - パケットサイズの設定

# 構成試験における障害ポイント、フローを図に起こすと！





## 運用試験

- 基本的には、運用で利用するコマンドをすべて実行し、通信断の有無を確認する。
- メンテナンス(迂回)を想定したシナリオを実施し、通信断の有無を確認する。
- SNMP等のPollingによる負荷や、syslogによる負荷を確認する。

# 本題

## お題 その1

- ARPの学習時間ってきにしていますか？
- DC系コアスイッチにてMaster/Slave切り替えをしたときに想定以上に通信断が起きたことはありませんか？※FHRP等
- Uplink側(上位ルータ),Downlink側(ToR)どちらから作業しますか？
- Uplinkからやると、どうなるか！

⇒ Slaveは、配下のNodeのARPを所持していないことが多いです。  
この状態で、Uplinkを切り替えるとNode向けのトラフィックが一気に流れ込み、ARP解決処理が走ります。

数百台規模のNodeなら良いかもしれませんが、仮に数千？数万？だったら！

※機器/構成によっては、mac tableでも同様の問題があります。

※mac tableの場合は未学習ならflooding

## お題 その1 に対する試験の重要性と、解決方法

- 単体検証試験にて、ARP検証があります。
- 試験機を使い、ARP Request/ReplyによるARP学習時間を計測。
- 我々は、通常1万IPの学習試験を実施しています。
- とある、ハイエンドなルータでも15sec程度かかります！
- お使いの装置の性能(学習時間、CPU利用率)を把握しておきましょう！

## 解決方法

- 100%の解決策ではありませんが、極力DownLink側からの切り替えをお勧めいたします。
- セグメントが複数ある場合は、セグメントをわけて実施することをお勧めいたします。
- コアスイッチではGARPを出すことで、配下からの通信は切り替わり、その後順次ARP/mac tableが学習され、インパクトが減ります。

## お題 その2

- シャーシ型の装置においてLCのメモリって気にしてますか？
- 例えば、回線増強においてスクエア構成だった物を、たすき掛け構成に変えたので、iBGPのMultipathしなきゃ！とか

⇒とある、分散型フォワーディングに対応した装置にてMultipathにしたことにより、LCのメモリが枯渇しこんな事象が起きました。

- I. Linkは上がったままだ
- II. どうもフォワーディングは止まっているようだ
- III. ん、connectされた経路情報は再配送されているようだ

装置冗長取ってるから、Linkも落ちてくれれば、切り替わったのに・・・

ちなみにメモリ問題は、メモリリークでもおきますよ！

## お題 その2 に対する試験の重要性と、解決方法

- 構成検証試験にて、実環境を極力模擬した上で、実環境の倍近い経路負荷をかけて検証を実施しています。
- 可能なら、将来を見据えた拡張構成まで模擬できると嬉しいですね。
- 単体/構成検証試験にAgingがあります、最大でも48H程度ですが高負荷状態にてパケロス、メモリ利用率等を試験しています。加えてOperationコマンドをマクロ等で投入し続けることで、特定コマンドによるメモリリークを発見できる場合もあります。

### 解決方法

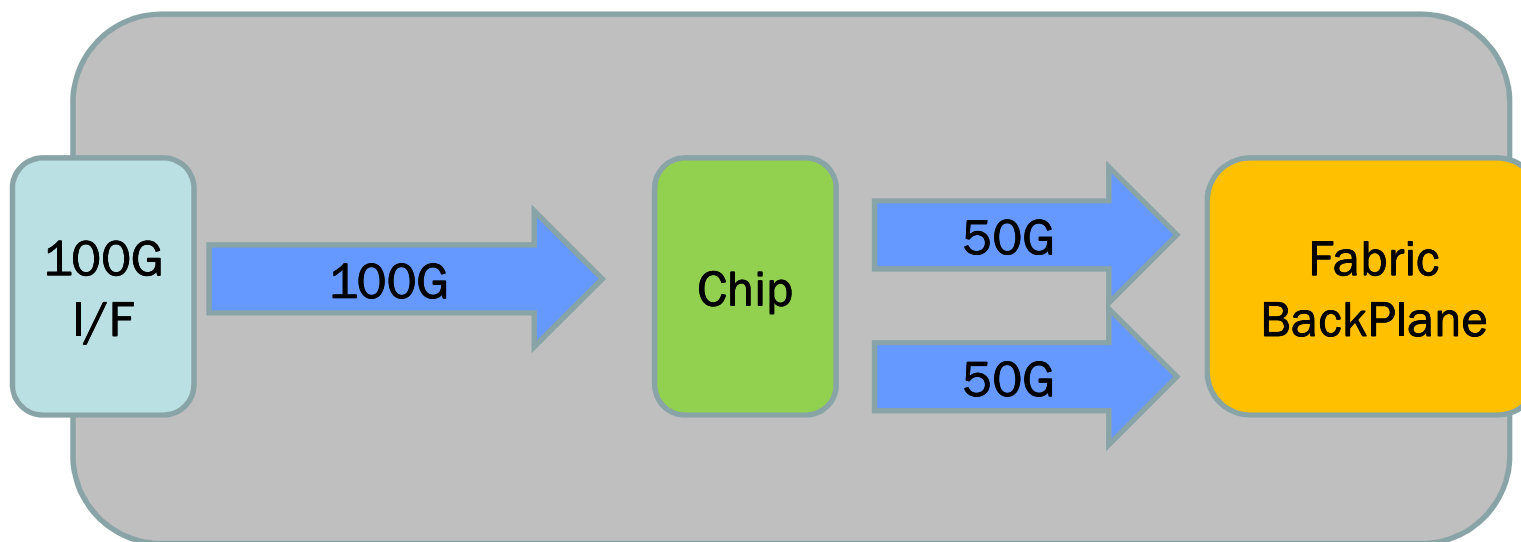
- 通常運用時には、LCのメモリも監視してください。
- 各種LCにて搭載しているメモリを把握しておきましょう。また1経路あたり消費するメモリサイズも把握しておくの良いかと思います。

### その他、解決方法？

- CiscoさんのVIRLとか仮想環境でこの辺って試験できるんでしょうか？
- 他、どのような試験方法があるかご意見頂きたいです！

## お題 その3

- ルータ/LC等のアーキテクチャを理解して検証ってしていますか？
- 例えば、100GbpsのI/Fを持つLCでも内部的にFabric/BackplaneとつながるまでにどのようなASICがあり、どのようなフローでパケットが流れていくのか！
- また、物理パスと論理パスも把握しておきましょう！
- 例えば、こんなアーキテクチャのLCがあった場合、何が問題か！



## お題 その3 に対する試験の重要性と、把握しておきたい事

- 一見すると100Gbpsをさばけそうに思えますが、注意しなければいけないことがあります
- 1. 50Gbps×2本のパスに分かれた際のロードバランスはどうなっているか？
- 2. 50Gbps×2本のパスの中の論理パスはどうなっているか？
- 3. Fabric/BackplaneまでTotal100Gbpsしかないってことは？

### 1, 2で起きる問題

- 1フロー(1対のsrc/dst IP)で、100Gbpsを出せない。  
論理パスにおいて例えば、10Gbps×16本などに分かれている場合もある。その場合は1フローではMaxで10Gbpsしか出ない事になる。

### 3で起きる問題

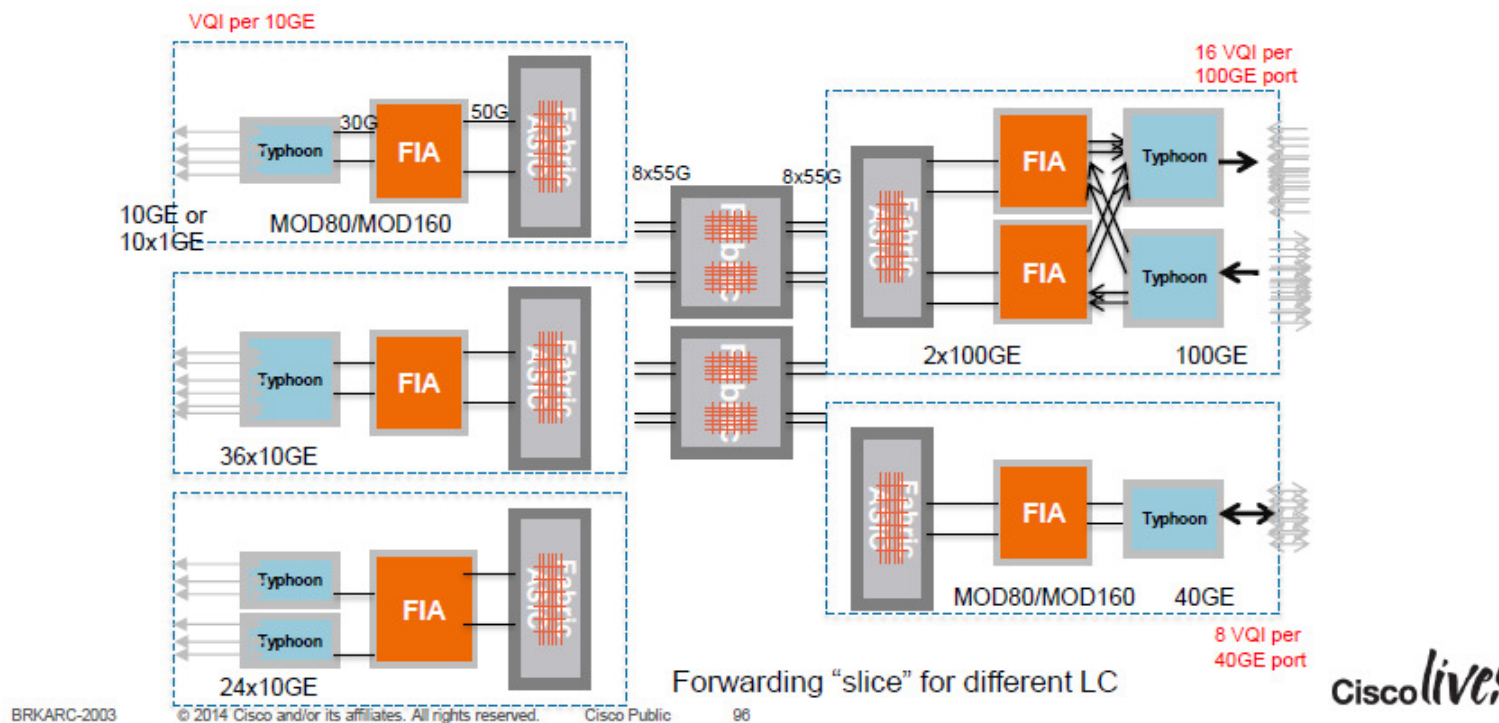
- 仮に別ポートへmirroringを設定した場合、Fabric経由でポート間通信がされる仕組みであれば、通常のトラフィックとあわせて100Gbpsとなる為、ポート間通信のアーキテクチャは理解しておくべき。



# Ciscolive!とか見てみよう！

- どなたでもアカウント作成できると思いますので！
- こんな情報まで載ってます※ciscolive! BRKARC-2003より

## Understand VQI and Internal Link Bandwidth

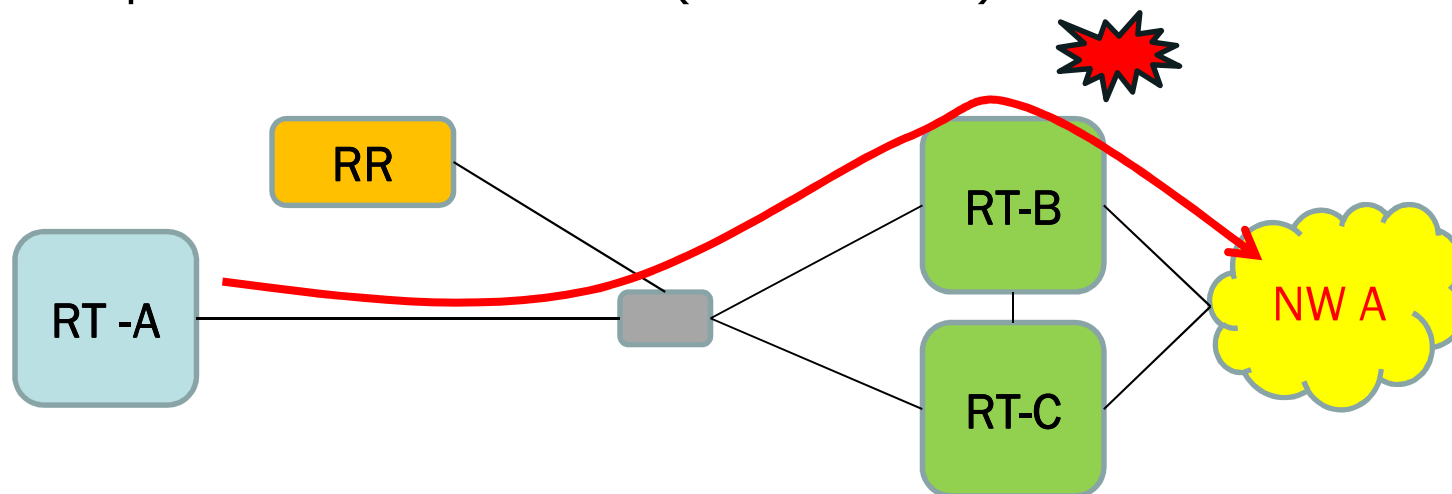


## アーキテクチャを理解することでこんな便利なことも？

- 通常パケロスはshow interface等で確認することが多いかと思いますが、先ほどのLCアーキテクチャを知っていると、こんなコマンドがあることに気づきます！※ciscoさんを例に挙げています
  - ◆ **show controllers fabric fia stats**
  - ◆ **show controllers fabric fia drops**
  - ◆ **show controllers fabric fia errors**
- show interface コマンドではわからないパケットのDropポイント(ボトルネックポイント)等の発見に役立つこともあるかもしれません！
- 中には、internal系のコマンドもあり、開発者用のコマンドもあったりしますので、そこはご了承ください

## お題 その4

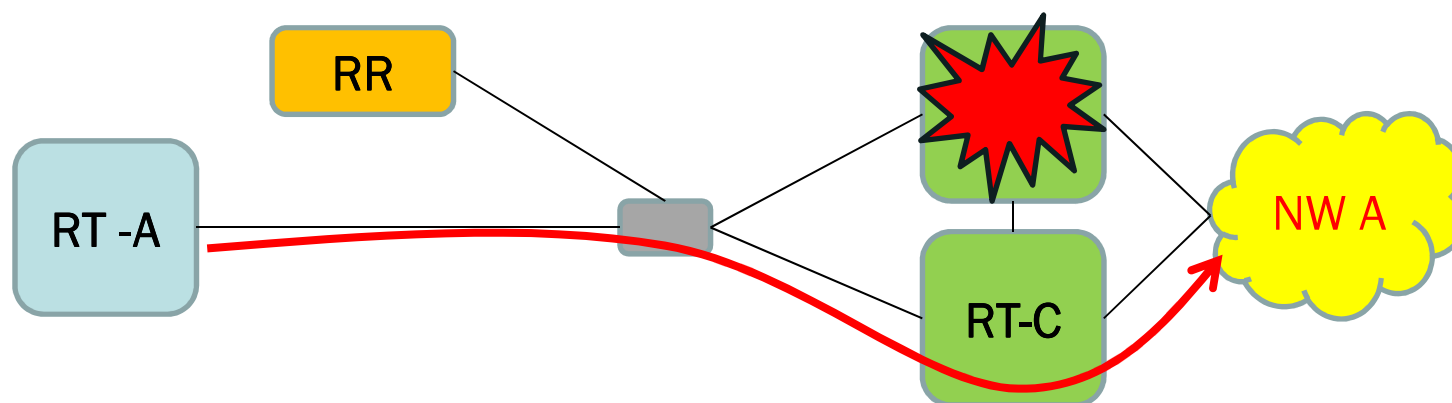
- 通信断時間の予測ってしていますか？
- 例えば、以下の構成で、以下の障害がおきると？ ※Cisco想定ですが
- ※multipathはない環境を想定(IGPはOSPF)



- 正常時 RT-A ⇒ RT-B ⇒ NW A
- RT-Bで障害が発生したら？

## お題 その4

- RT-A ⇒ RT-C ⇒ NW A に切り替わるけど、何が起きているのか！



- RRがSPF delay 5sec後に RT-B Down検知
- RRにてNHTで1sec後にwithdrawn送信
- RT-CよりRRへbest経路のupdate 1sec程度
- RRにて advertise-delay 5sec後に RT-C(best)をupdate
- RT-AにてRIB/FIB更新後に経路の切り替わり 1sec程度

**1 2 ～ 1 3 秒の通信断が発生します！**

## お題 その4 に対する試験の重要性と、把握しておきたい事

- 通信断時間を想定しておくことで、拠点毎(end-to-end)での通信断時間を把握することができます。
- フロー毎(IPの組み合わせ)に通信断時間が違う場合があります。
- タイマーに対する理解度があがり、タイマーチューニングに役立ちます。
- 想定断時間より収束が早い場合、検証環境のミスconfigや、構築ミスの発見に繋がるほか、新OS/HWになったことによる仕様変更等を発見できる場合があります。

## RIB/FIB問題

- タイマーはルーティングプロトコルに設定されており、想定断を求める重要な要素ですが、RIB/FIB同期にて以下のような事象が起きる場合があります。例えばとあるハイエンドルータにて50万経路を学習させる
  - BGP Tableにすべてインストールされるまでに43秒
  - RIBにすべてインストールされるまでに48秒
  - FIBにすべてインストールされるまでに、**95秒**

**想定断時間を把握しているからこそ、その他要因における通信断時間の発見に繋がります！**

## 最後に、

- 何はともあれ、ベースとなるところが正常に動き、上位層側では何も気にすることなく、全力で利用してもらおうインフラを作ろう！
- その為にも、まだまだNWとしての基本機能検証は必要ではないか！

## 将来！

- 例えば、データプレーン、コントロールプレーンの分離や、ソフトウェアルータとの使い分け等で、コンバージェンスの改善や、NW機器のリソースを気にする必要は減ってくるのか？
- NW機器として、CPU性能の向上によりそこは改善していくのか？

## 課題！

- すべてを人力でやるのは、限界。早く自動化しよう！

つぎは、縁の下(水中)ではなく、華やかな？水面の試験のお話を  
J u n i p e r 中嶋様お願いします！