

SR-MPLS Egress Peer Engineering を 使って測るインターネットレイテンシー

JANOG48 Meeting

中村 遼

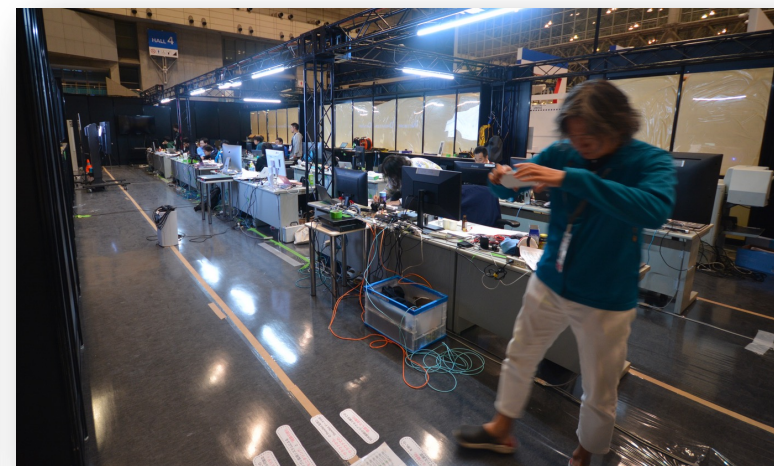
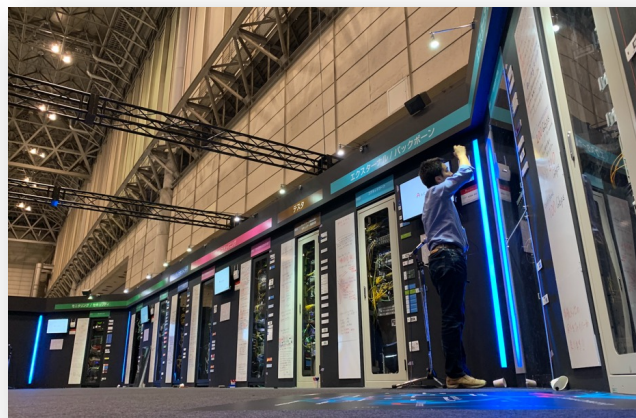
東京大学/Interop Tokyo ShowNet NOC Team

自己紹介

- <https://github.com/upa>
- 東京大学 情報基盤センター ネットワーク研究部門 助教
 - 業務: 東京大学キャンパスネットワークの運用
 - 研究: Linux kernelのNetwork stackやPCIeまわりのネットワーク+システムソフトウェア分野がメイン
 - コンテナを高速化しようとしたたり、PCIeのTransaction LayerをCで実装したり
 - 他にもルーティングや仮想ネットワークなどその時々によっていろいろ
- Interop Tokyo ShowNet
 - 2012年からNOC
 - 主にL2/L3、2016年くらいまではSDN/NFVも担当

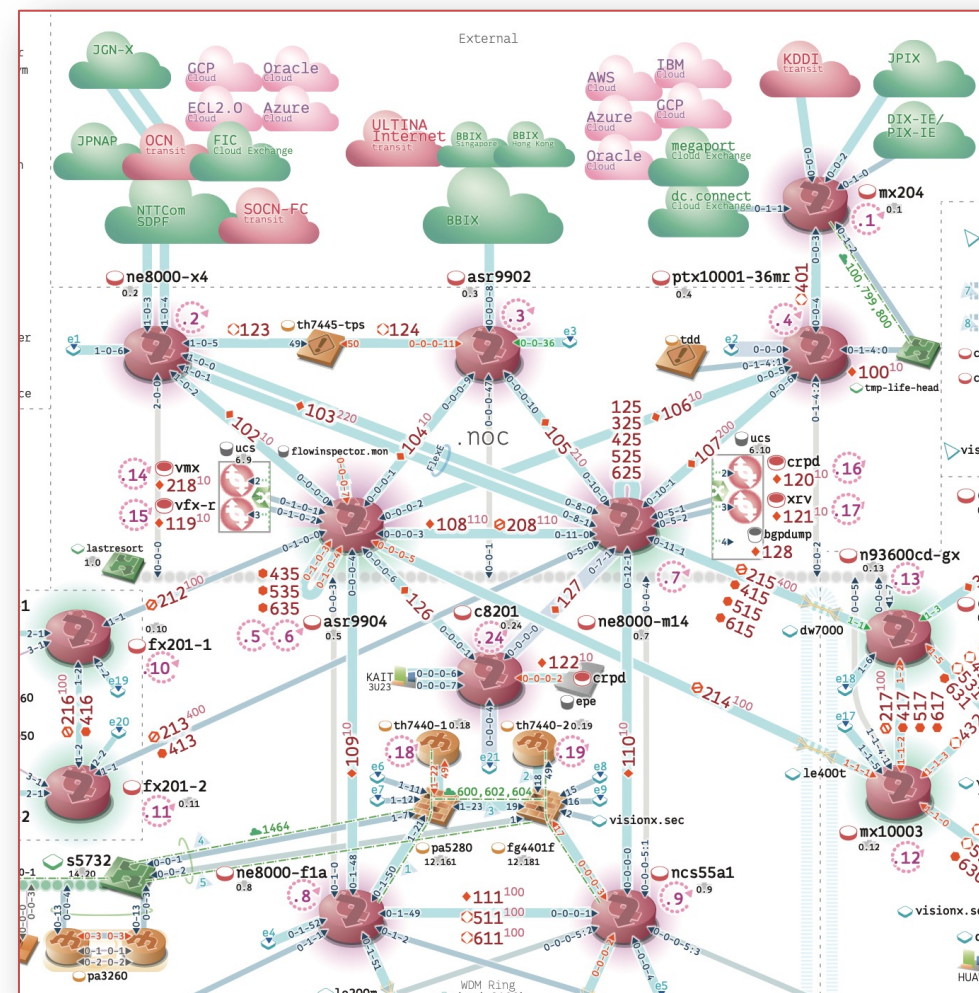
Interop Tokyo **SHOWNET** *Face the Future*

- **Interop:** 世界最大のネットワーク機器と技術の展示会
 - 毎年6月に幕張メッセで開催、来場者約14万人
- **ShowNet:** 世界最大のデモンストレーションネットワーク
 - その年々の最新技術/プロトコルの相互接続性検証やデモを実施
 - 出展社や来場者へのネットワーク接続性を提供するLive Network



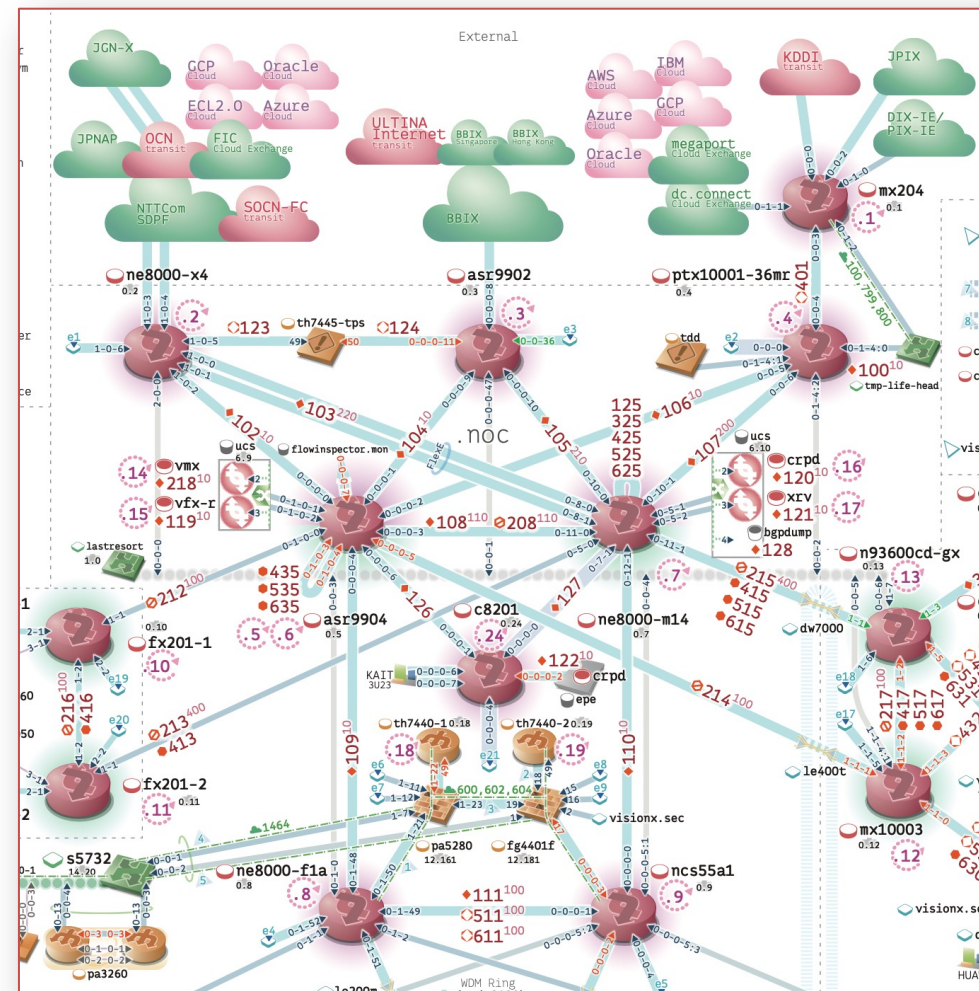
ShowNet 2021: SR-based Backbone

- SR-MPLSによる柔軟なトラフィック制御
 - Flex Algoによる優先制御
 - Egress Peer Engineeringによる計測実験
- 世界最先端・最大規模のSRv6相互接続性検証
- 400Gbps、Flex Ethernetによる200Gbpsなどの広帯域リンク



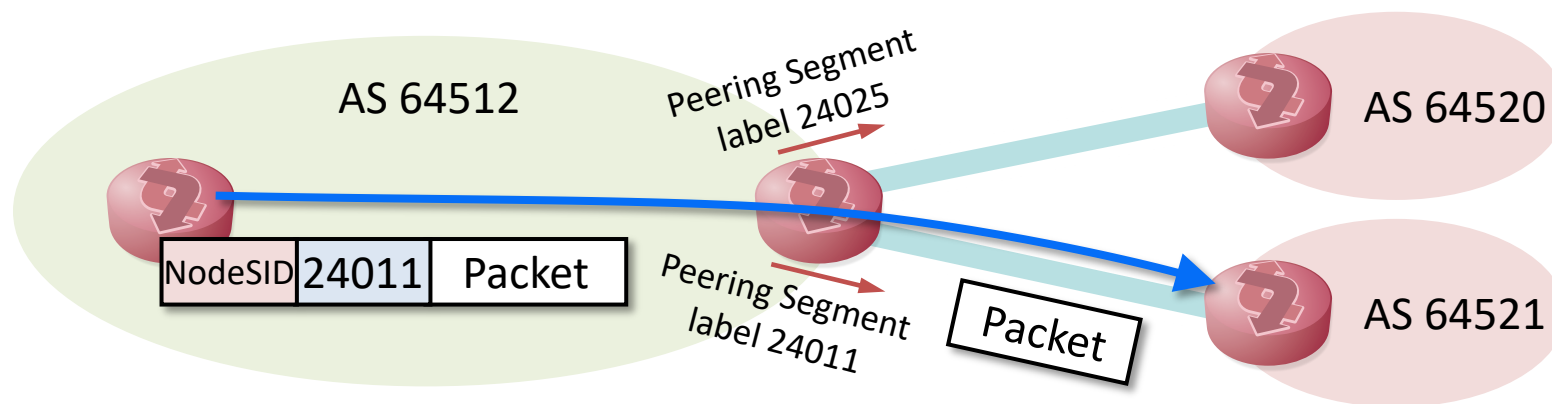
ShowNet 2021: SR-based Backbone

- SR-MPLSによる柔軟なトラフィック制御
 - Flex Algoによる優先制御
 - Egress Peer Engineeringによる計測実験
- 世界最先端・最大規模のSRv6相互接続性検証
- 400Gbps、Flex Ethernetによる200Gbpsなどの広帯域リンク



SR-MPLS Egress Peer Engineering

- パケットをLongest Prefix Matchによらずに狙ったピアへ
 - draft-ietf-spring-segment-routing-central-epe
 - ASBRがBGP PeerごとにSRのセグメント(ラベル)をアサイン
 - BGP Peering Segments
 - そのセグメント行きのパケットは、そのピアへ送信される
 - これを利用して送信方向のトラフィックを制御できる

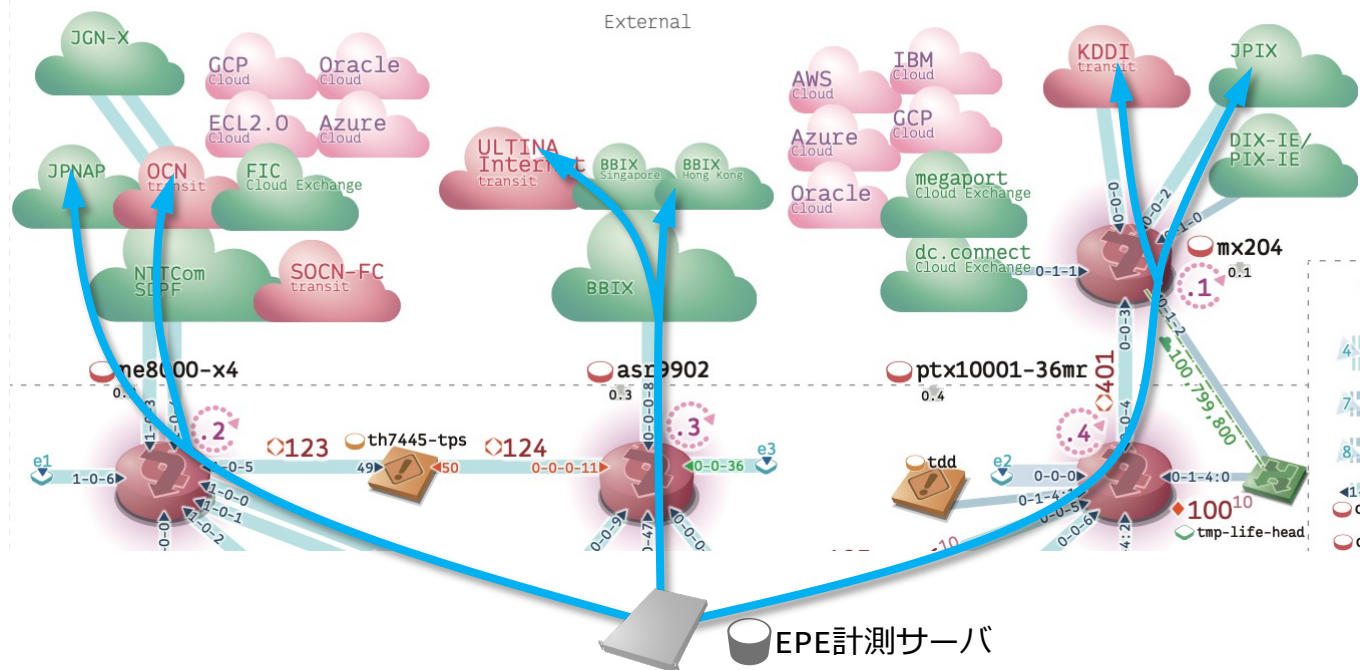


はたしてEPEを使う効果とは？

- レイテンシーという観点から実際に色々計測してみよう
 - Peer経由はTransit経由よりもRTTが短いはずだが、実際どれほど？
 - BestpathはどれくらいBestなのか？
 - レイテンシーという観点から見たEPEを使う効果を明らかにする
 - と同時に、単純なはなし、わたし気になります！
- もちろん実際はトラフィック量も大きな要素
 - 吐き出すトラフィックの傾向と対外接続に大きく依存
 - テレメトリとセッتناコントローラの賢さ次第

計測実験の概要@ShowNet 2021

- Egress ASを変えながら、インターネット全体にping/traceroute
 - ASBRの3台が(実験に参加する)Peer ASごとにラベルをアサイン
 - RTTと経路の情報から、Egress ASの選び方で遅延がどう変わるかを分析



About Measurement Experiment

In ShowNet 2021, we deploy a demonstration of Egress Peer Engineering (EPE) by Segment Routing. In a part of the demonstration, we conduct a measurement experiment to clarify the effectiveness derived from traffic engineering with EPE. The experiment sends ICMP or UDP probe packets that have arbitrary destination IP addresses to ASes peering ShowNet. The traffic rate is approximately 15Mbps, and the packets are sent for dozens of minutes to one hour periodically.

This experiment aims to demonstrate EPE and measure the efficiency of using such a traffic engineering technique. The measurement results are used only for sharing information with network communities and research purposes. We appreciate your understanding and cooperation.

☒ agree to join the measurement experiment of the demonstration of EPE

Peerする際に実験に参加するかどうかは選択可能

ASBRのConfig抜粋

NE8000-X4

```
bgp 290
 peer X.X.X.X as-number ASNUM
 peer X.X.X.X group ix-peer-v4
 peer X.X.X.X password cipher "Secret"
 peer X.X.X.X egress-engineering
```

ASR9902

```
router 290
 neighbor X.X.X.X
 remote-as ASNUM
 use neighbor-group ix-peer-v4
 egress-engineering
 password encrypted "Secret"
!
```

MX204

```
protocols {
  bgp {
    group ix-peer-v4 {
      type external;
      import import-from-ix-peer-v4;
      export export-to-ix-peer-v4;
      neighbor X.X.X.X {
        authentication-key "Secret"
        peer-as ASNUM;
        egress-te-node-segment {
          label 1000097;
        }
        egress-te;
      }
    }
  }
}
```

計測パケットをEPEで狙ったAS越しに送信

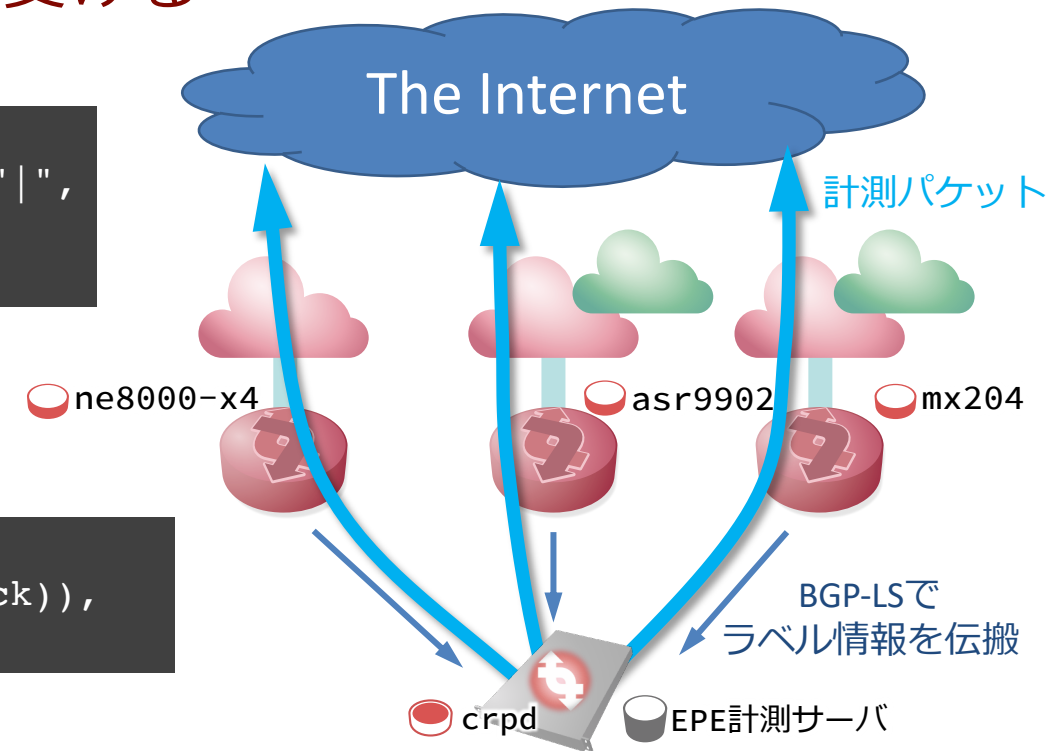
1. 計測サーバでJuniper cRPDが動作しBGP-LSを受ける
 - cRPDが学習したトポロジ情報を取り出す

```
c = [ "docker", "exec", "-it", "crpd", "cli",  
      "show", "route", "table", "lsdist.0", "extensive", "|",  
      "display", "json", "|", "no-more" ]  
jsonstr = subprocess.check_output(c).decode()
```

2. 計測パケットを送信したいEgress AS行きのLabel Stackを経路にセットして

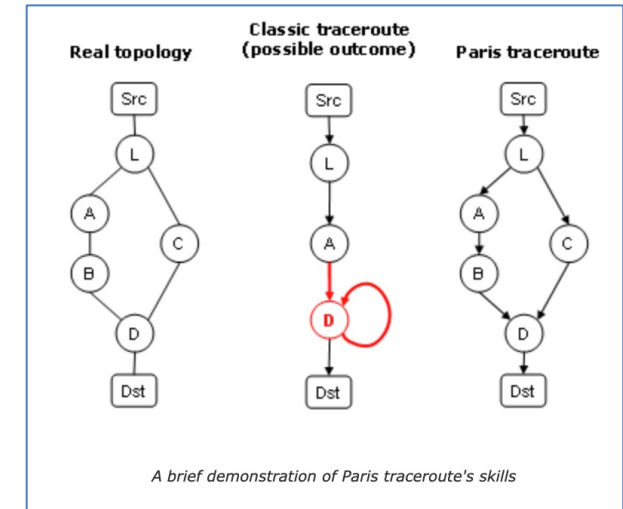
```
next_route = ["ip", "route", "add", "to", dest,  
              "encap", "mpls", "/".join(map(str, labelstack)),  
              "via", nexthop ]
```

3. あとはひたすらpingとtraceroute



pingとtraceroute

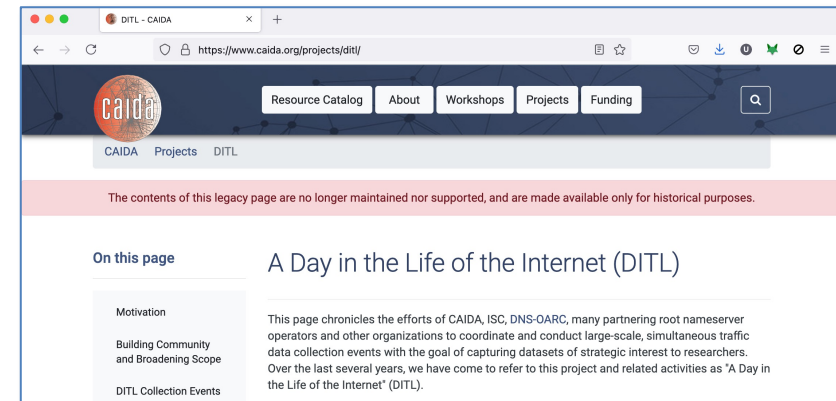
- Scamperを使用
 - <https://www.caida.org/catalog/software/scamper/>
 - ping, tracerouteを含む様々な計測メソッド
 - 複数のアドレスを並列計測
 - 計測ではUDP PingとParis Tracerouteを使用
- さらに並列化
 - LinuxのNetwork Namespaceを使って複数Scamperプロセスで複数Egress ASの計測を同時実行



<https://paris-traceroute.net/>

計測の宛先

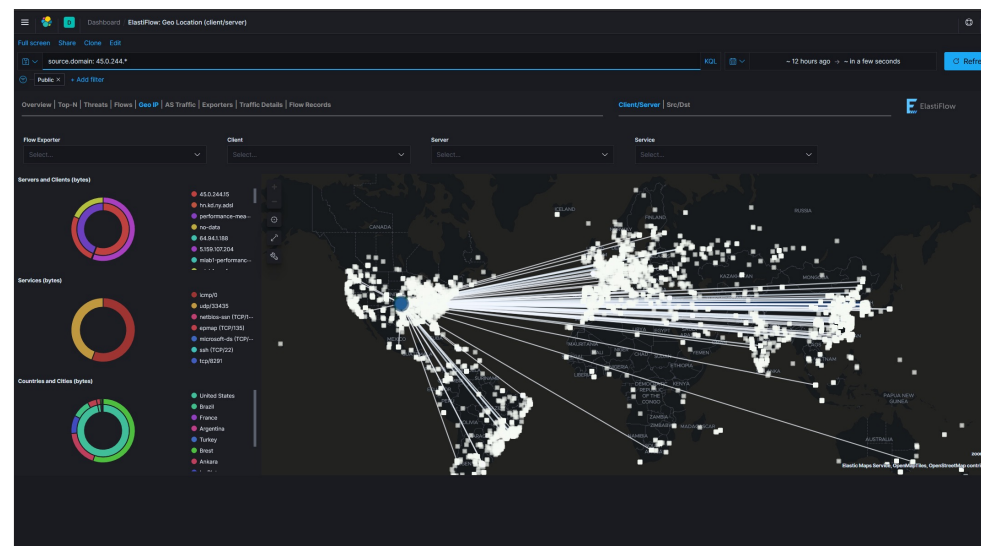
- 2,638,382 IPv4アドレス
 - 2020年のA Day in the Life of the Internetの某所のトレースからICMP Echo Replyを送信したIPアドレスを抽出
 - このトレースポイントの下にPinger Projectの計測箱がいた
 - そこから2021年3月31日にpingに応えたIPアドレスを抽出
 - さらに/24ごとに1 IPアドレスを抽出



<https://www.caida.org/projects/ditl/>

いざ計測

- 参加してくださったASとのeBGP Peer越しにひたすらping/traceroute
 - 101 Peers
 - 45 Egress ASes
 - 計測期間: 4月12日から4月16日
 - 4月12日にPeeringを開始
 - Interop展示会は4月14日から16日
 - 4月16日がJANOG48応募締め切り



とある実験参加ASで観測した計測パケット

解析結果の前に

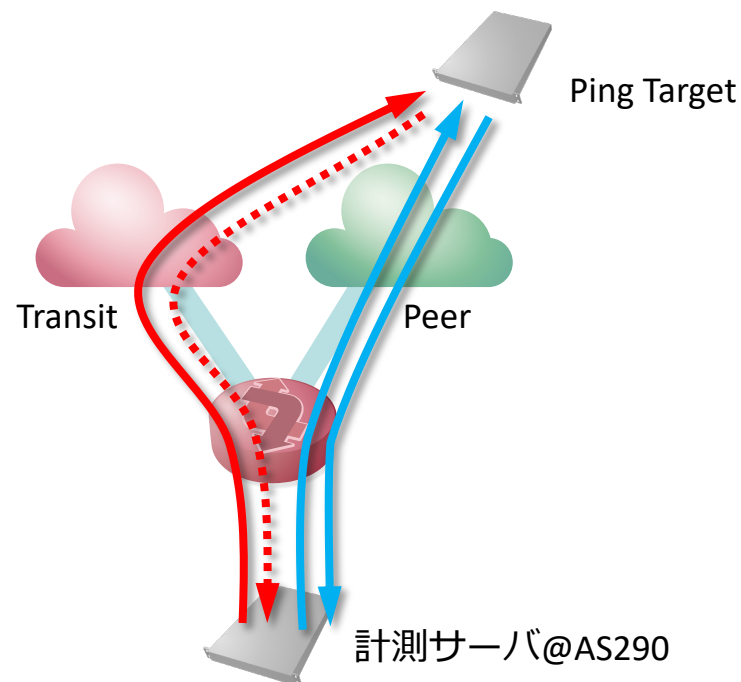
- 明らかにしたいのは「遅延という観点から見たEPEの効果」
 - どのASが良い悪いという話ではないです
 - 以後グラフ中のASを示す色は、グラフごとに異なります
 - RTTは長くなる方向に誤差が大きいです
 - 計測結果から複数回の試行のうち最小のRTTを値として使っています
 - AS290というとてもスペシフィックな環境からの計測結果です
 - どこでも必ずこうなる、というわけではありません

傾向を把握する一助/参考になれば幸いです

解析結果1:
PeerはTransitよりどれくらい近いのか？

解析の概要

- Peerから学習したPrefixに含まれる宛先に対して
 - ✓ そのPeer経由でping
 - ✓ Transit AS経由でping
 - 上記のRTTを比較
- EgressはTransit経由でも戻りは(おそらくほぼ)Peer経由
 - 戻りは同じ経路
 - RTTの差は行きの差のみ



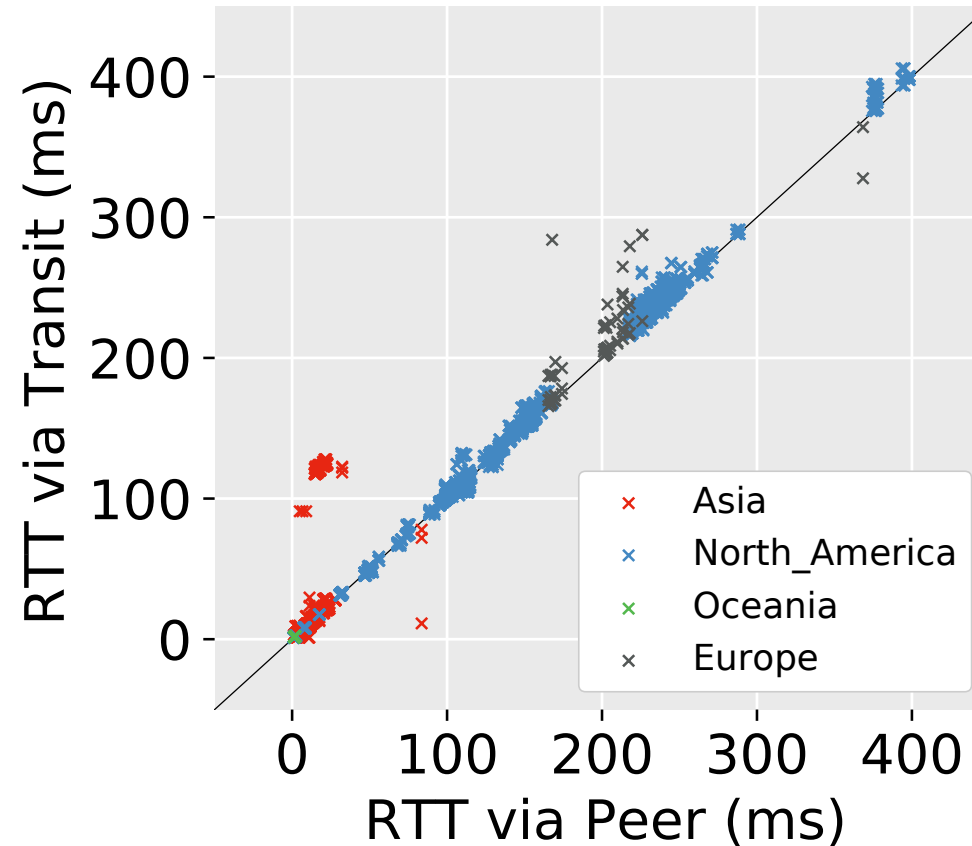
Peer vs Transit

- 各IPアドレスへのRTTをscatter plot

- X軸: Peer経路 のRTT
- Y軸: Transit経路のRTT

- 点が $x = y$ の直線より、

- 上にある ↑
Peer経路の方がRTTが短い
- 下にある ↓
Transit経路の方がRTTが短い

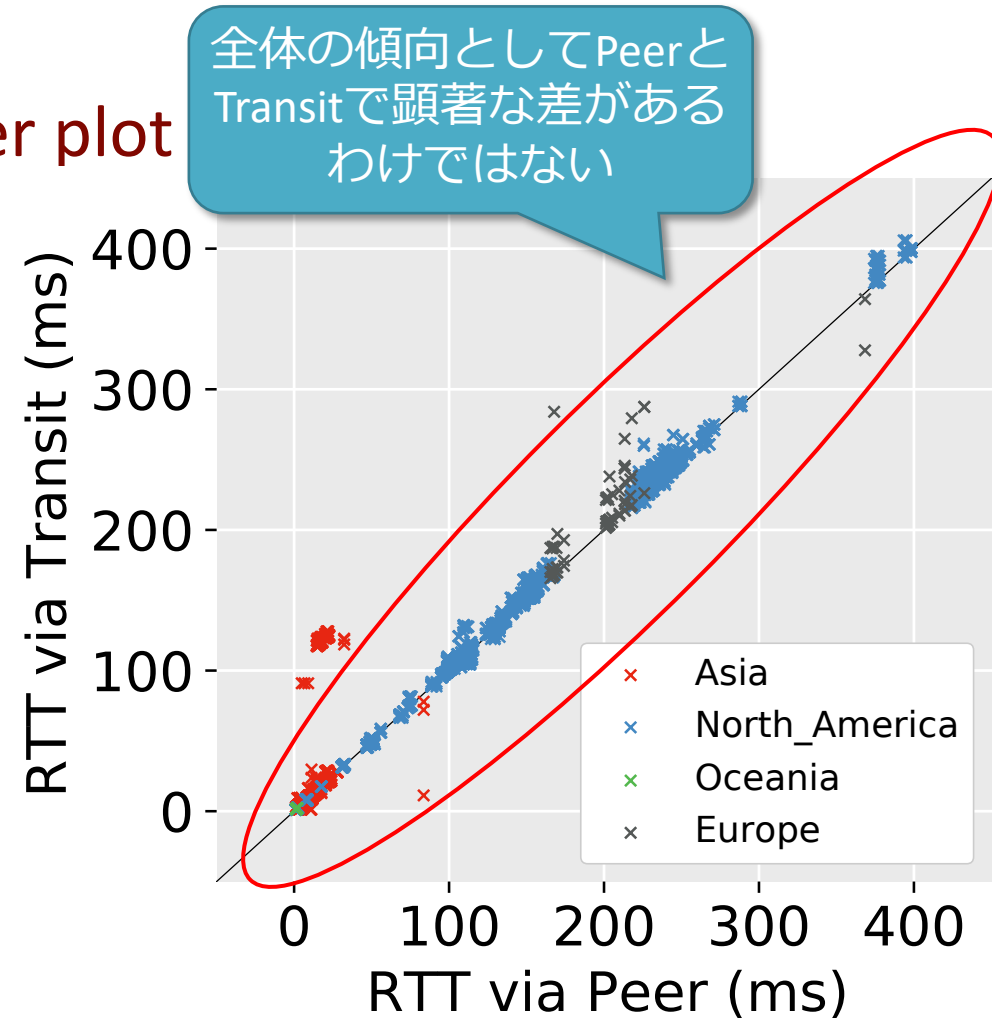


Peer vs Transit

- 各IPアドレスへのRTTをscatter plot

- X軸: Peer経由のRTT
- Y軸: Transit経由のRTT

- 点が $x = y$ の直線より、
 - 上にある ↑
Peer経由の方がRTTが短い
 - 下にある ↓
Transit経由の方がRTTが短い

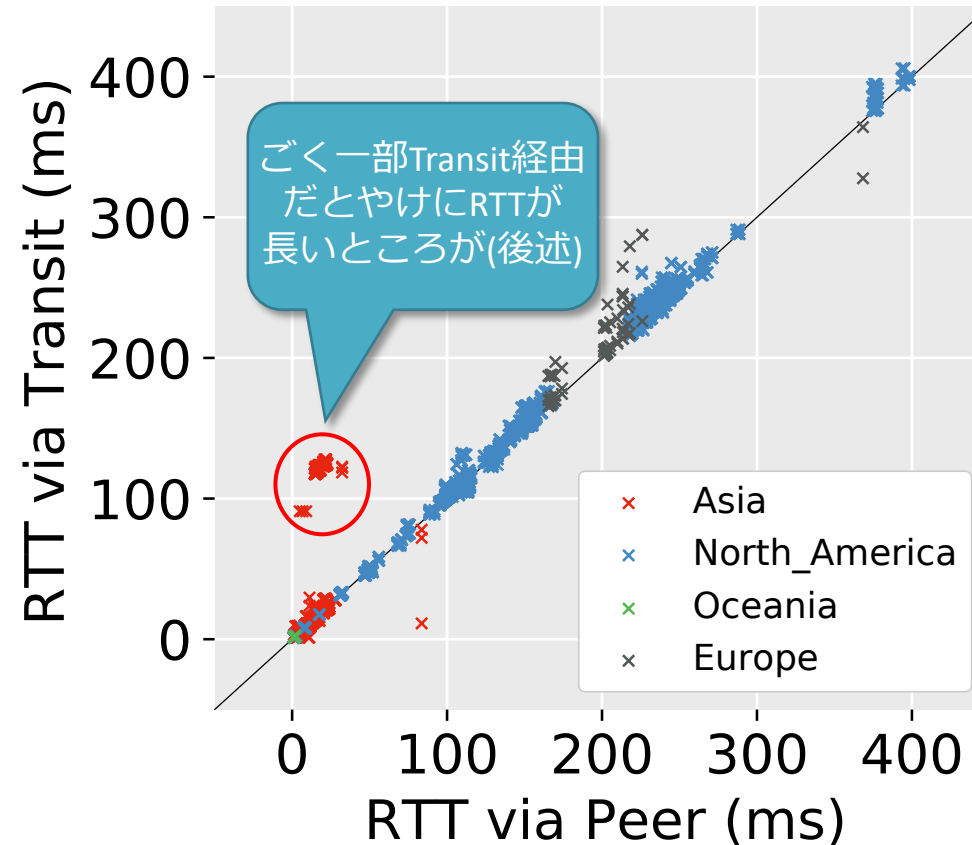


Peer vs Transit

- 各IPアドレスへのRTTをscatter plot

- X軸: Peer経路 のRTT
- Y軸: Transit経路のRTT

- 点が $x = y$ の直線より、
 - 上にある ↑
Peer経路の方がRTTが短い
 - 下にある ↓
Transit経路の方がRTTが短い

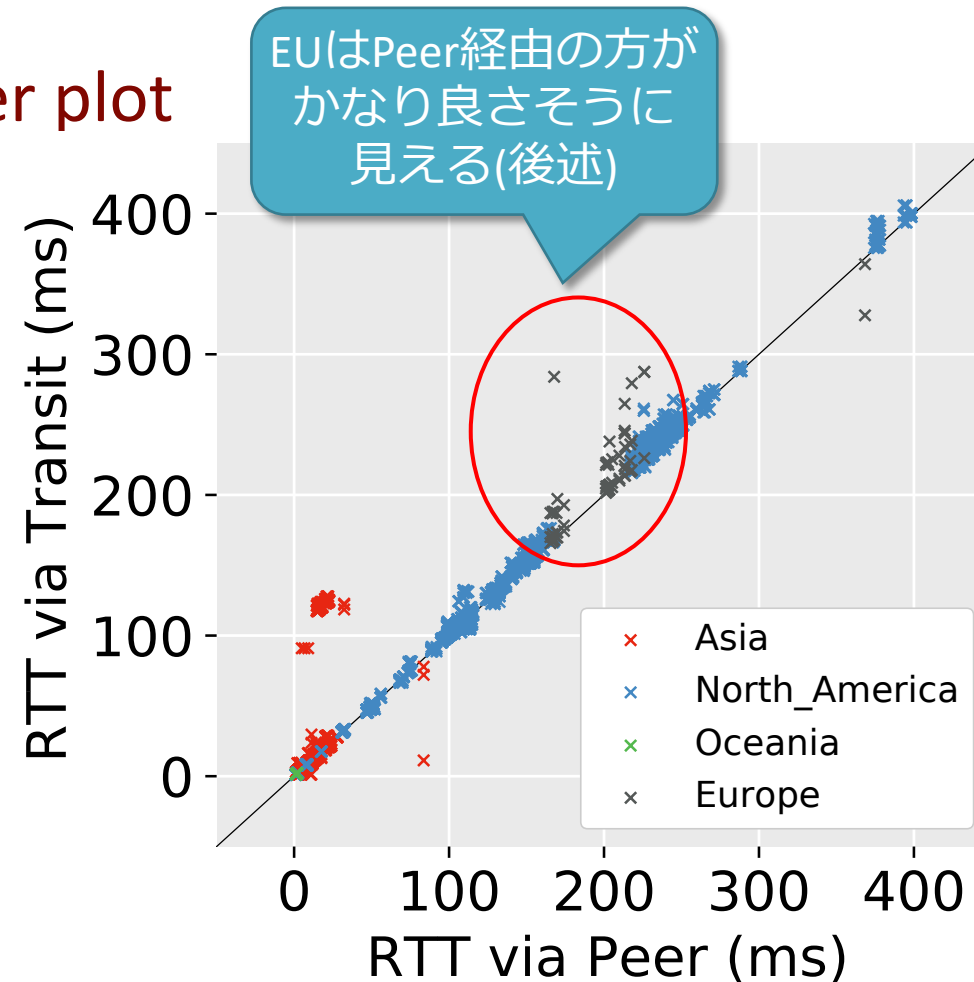


Peer vs Transit

- 各IPアドレスへのRTTをscatter plot

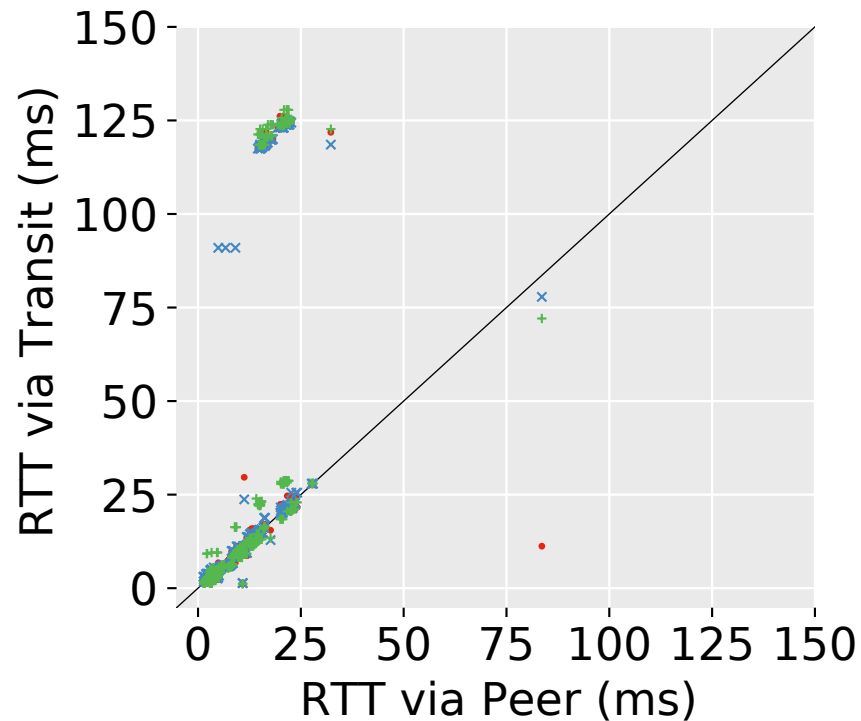
- X軸: Peer経路 のRTT
- Y軸: Transit経路のRTT

- 点が $x = y$ の直線より、
 - 上にある ↑
Peer経路の方がRTTが短い
 - 下にある ↓
Transit経路の方がRTTが短い

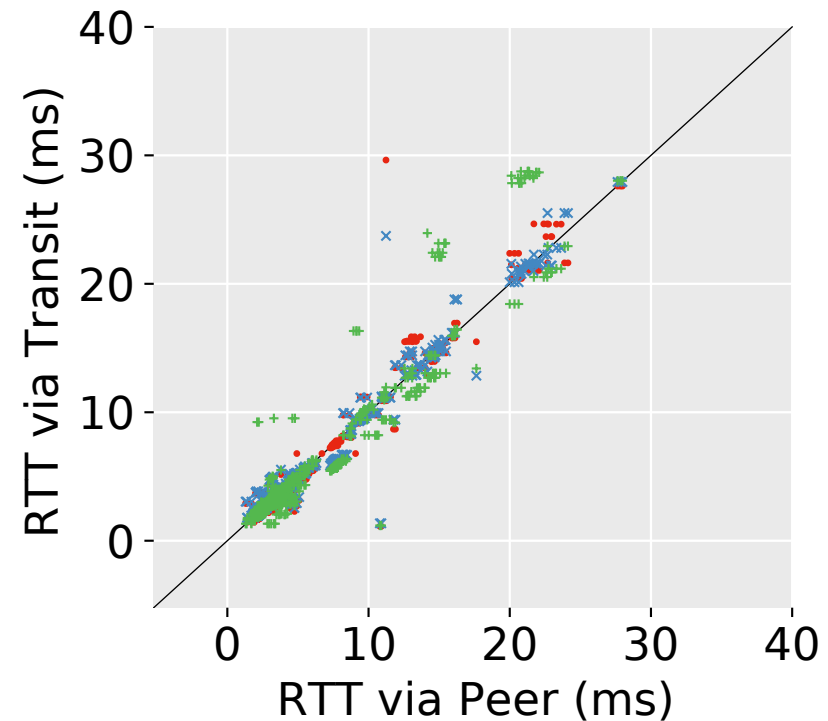


Peer vs Transit: Asia

- AsiaなIPアドレスだけを抜き出してプロット(実質JP)
 - 色分けは経由したTransit AS (Egress AS)

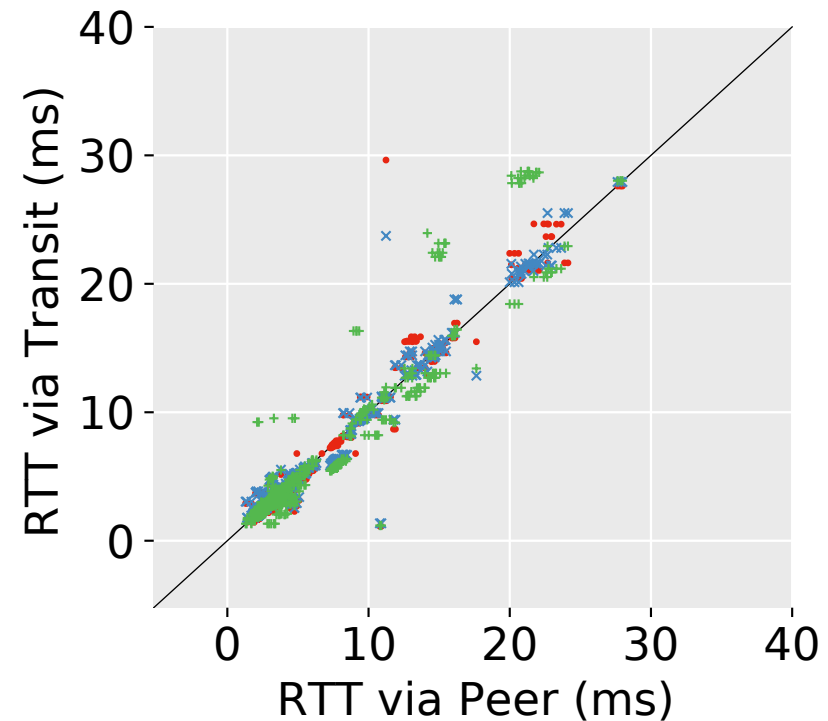
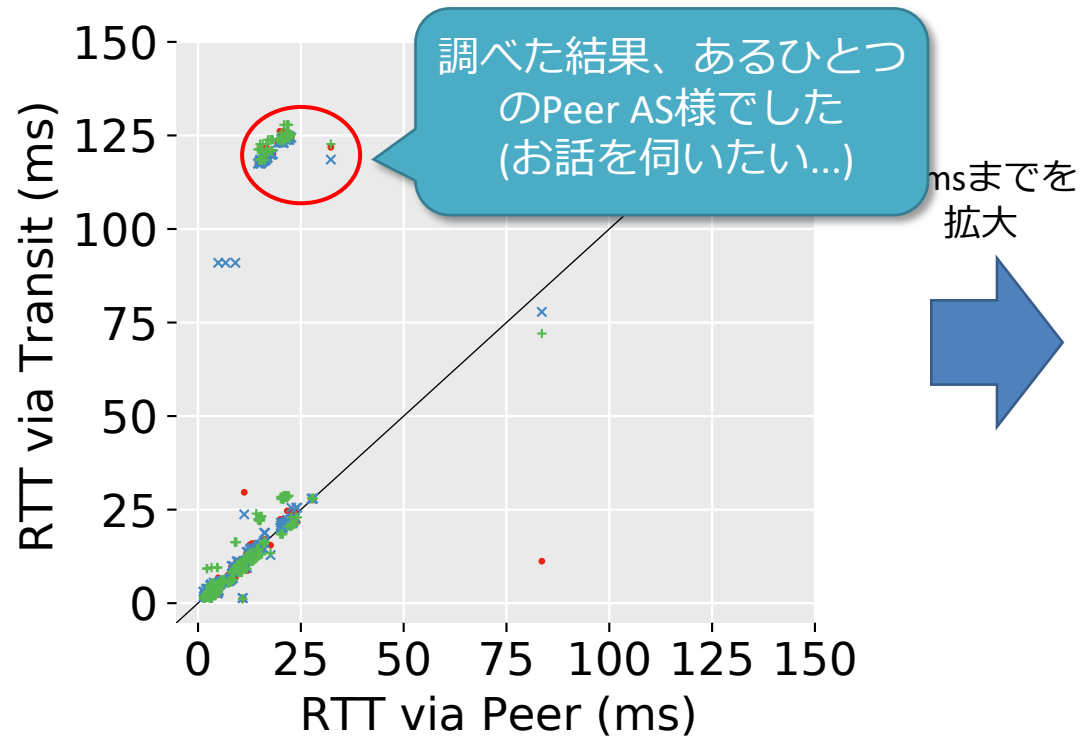


40msまでを
拡大



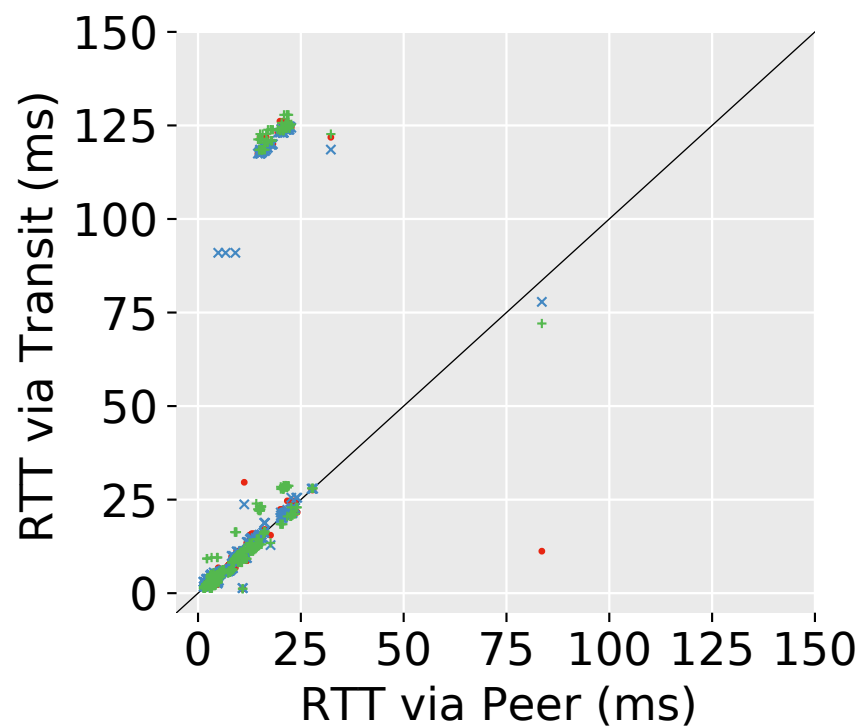
Peer vs Transit: Asia

- AsiaなIPアドレスだけを抜き出してプロット(実質JP)
 - 色分けは経由したTransit AS (Egress AS)

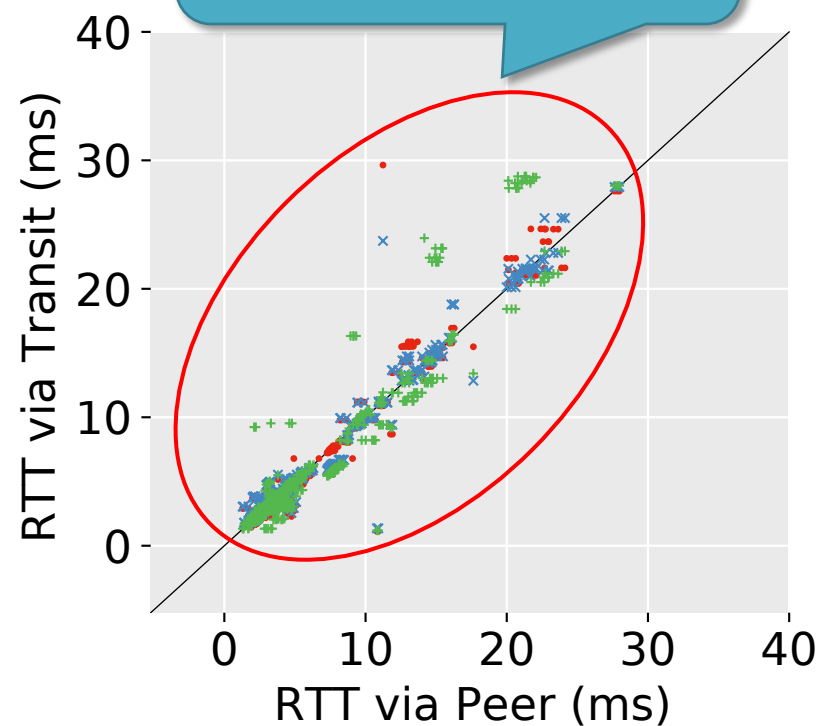


Peer vs Transit: Asia

- AsiaなIPアドレスだけを抜き出してプロット(第1図)
 - 色分けは経由したTransit AS (Egress AS)

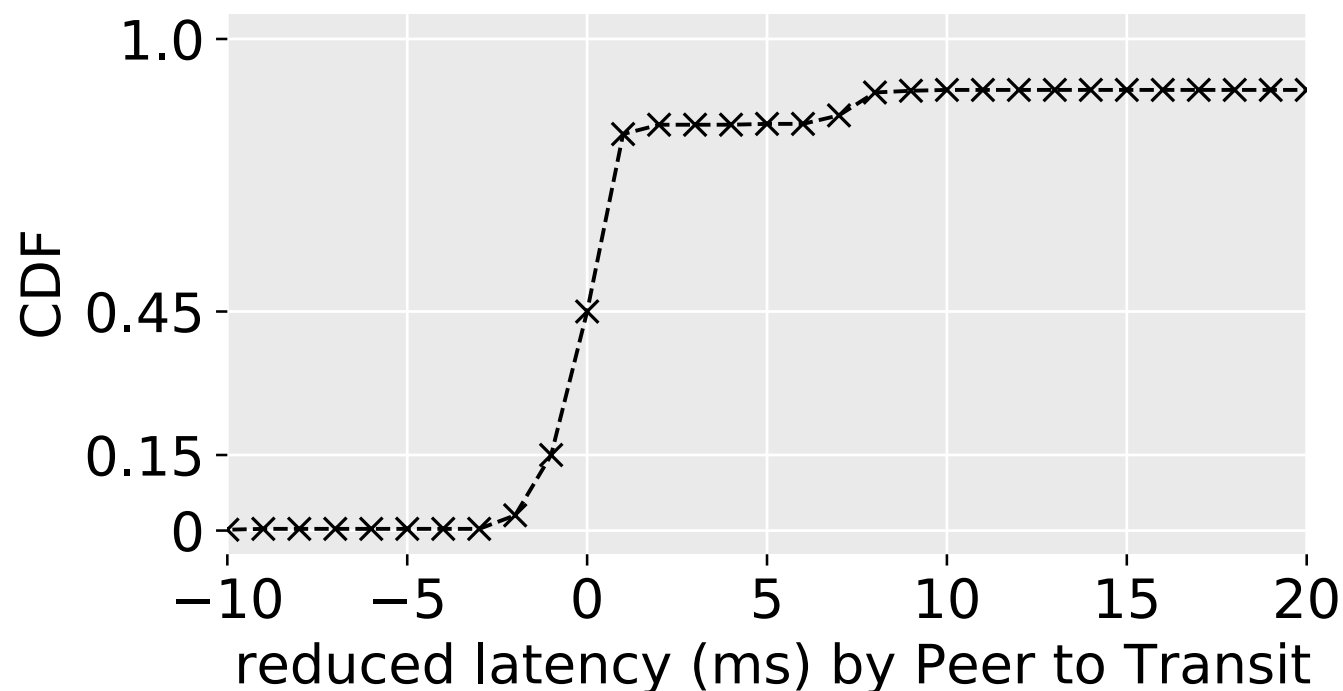


40msまでを
拡大



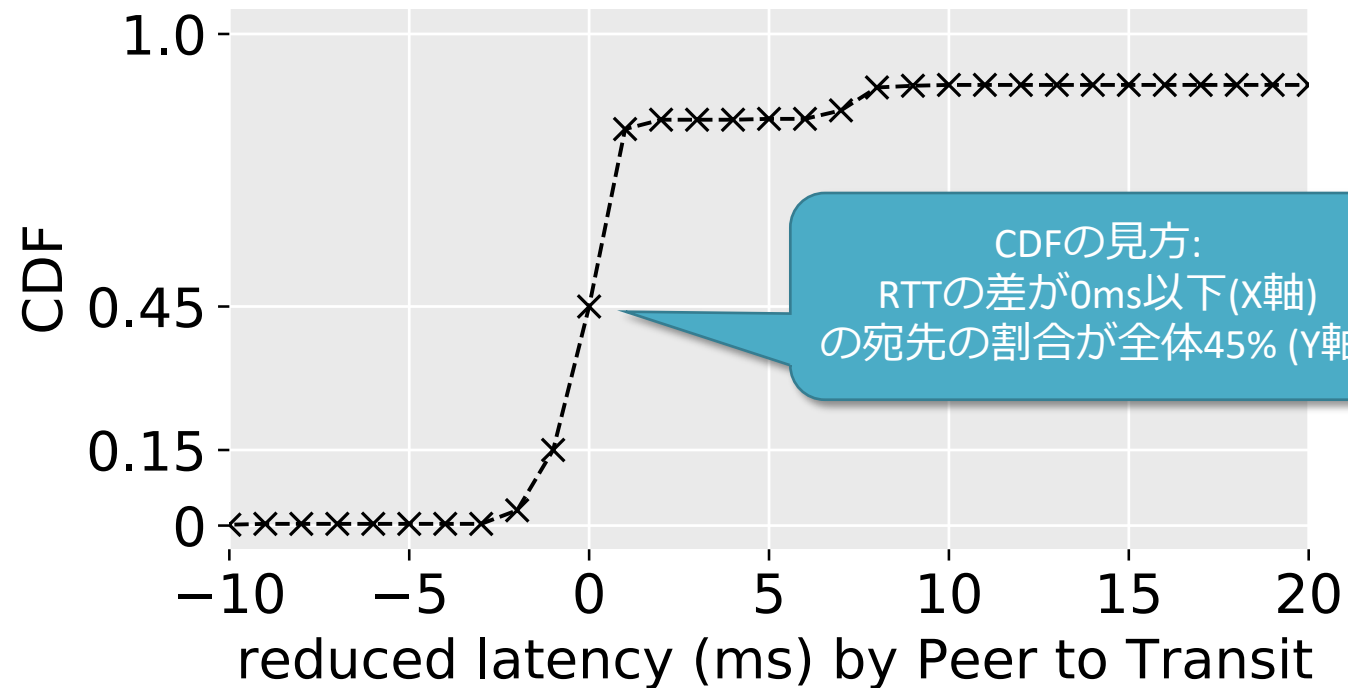
Peer vs Transit: Asia

- 実際どれくらい差がありそうか？
 - $RTT(\text{transit}) - RTT(\text{peer}) = \text{Peer経由で短くなるlatency}$
 - この値(reduced latency)を宛先アドレスごとに出して累積密度分布(CDF)でプロット



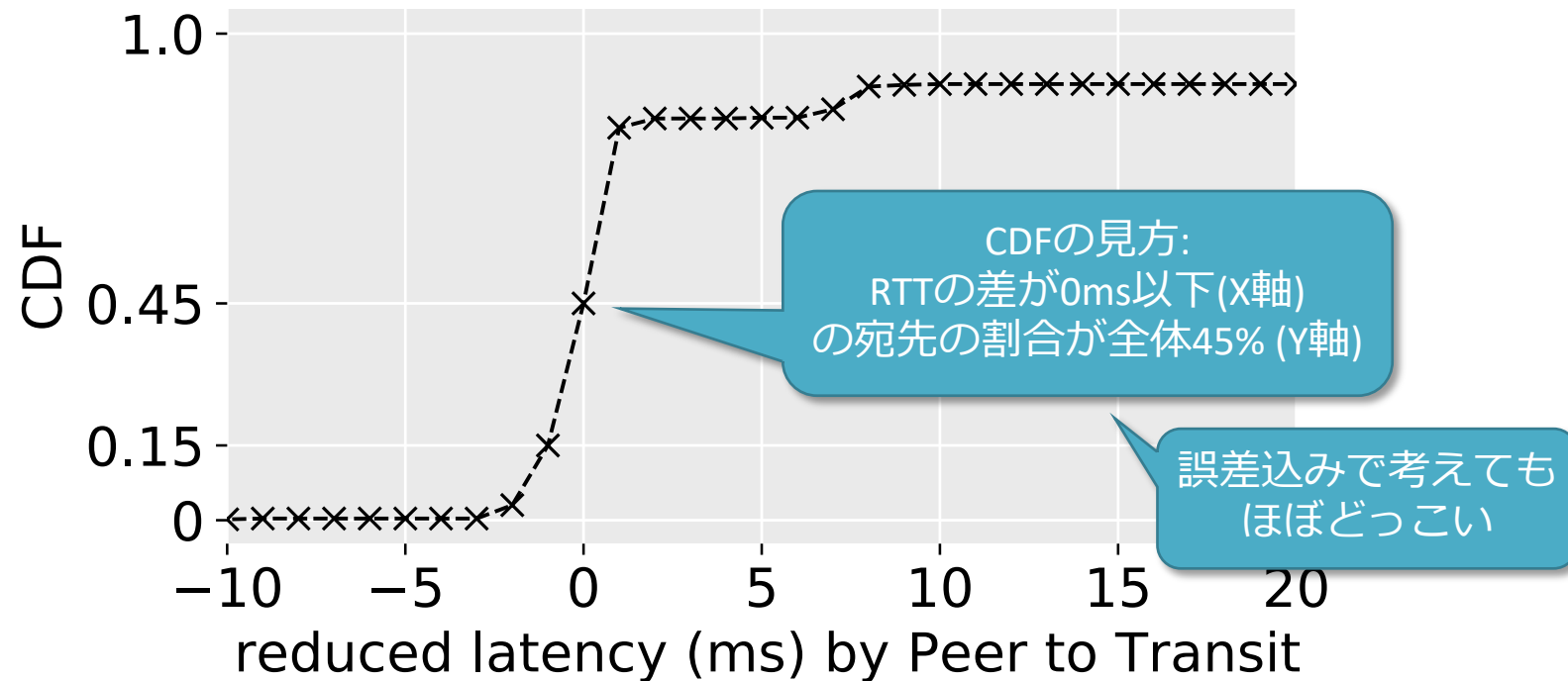
Peer vs Transit: Asia

- 実際どれくらい差がありそうか？
 - $RTT(\text{transit}) - RTT(\text{peer}) = \text{Peer経由で短くなるlatency}$
 - この値(reduced latency)を宛先アドレスごとに出して累積密度分布(CDF)でプロット



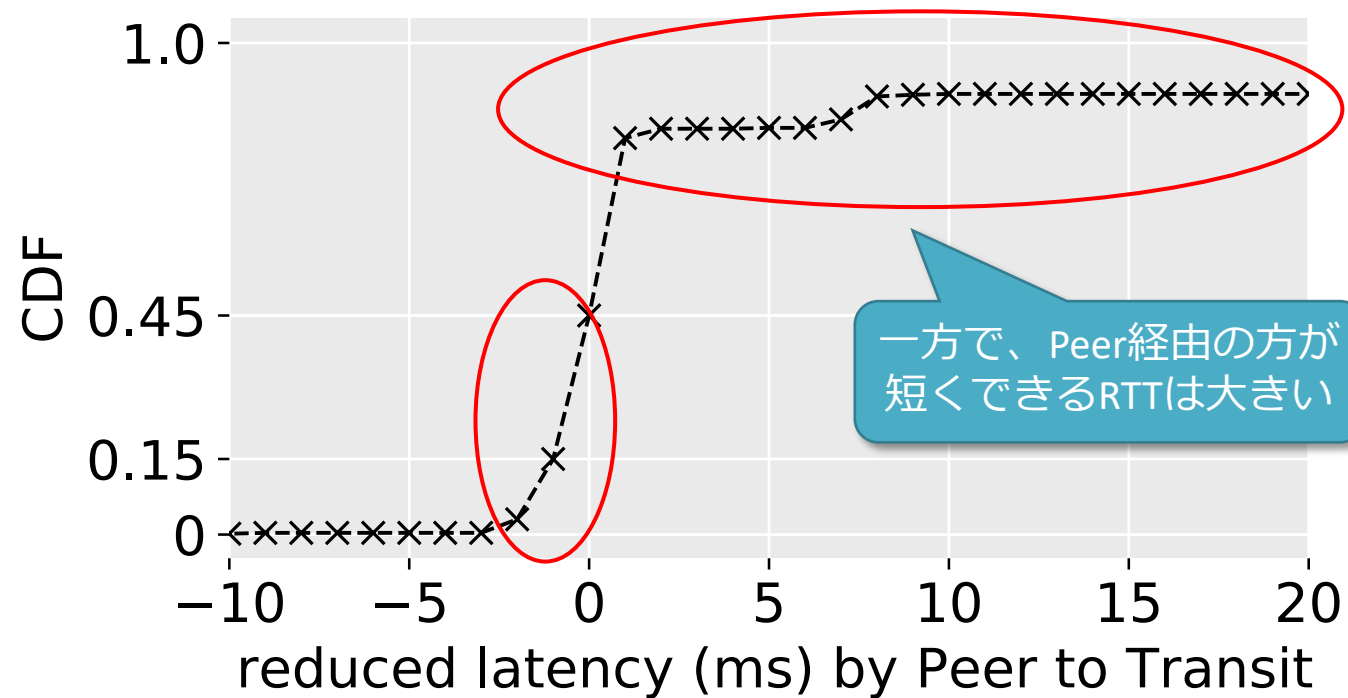
Peer vs Transit: Asia

- 実際どれくらい差がありそうか？
 - $RTT(\text{transit}) - RTT(\text{peer}) = \text{Peer経由で短くなるlatency}$
 - この値(reduced latency)を宛先アドレスごとに出して累積密度分布(CDF)でプロット



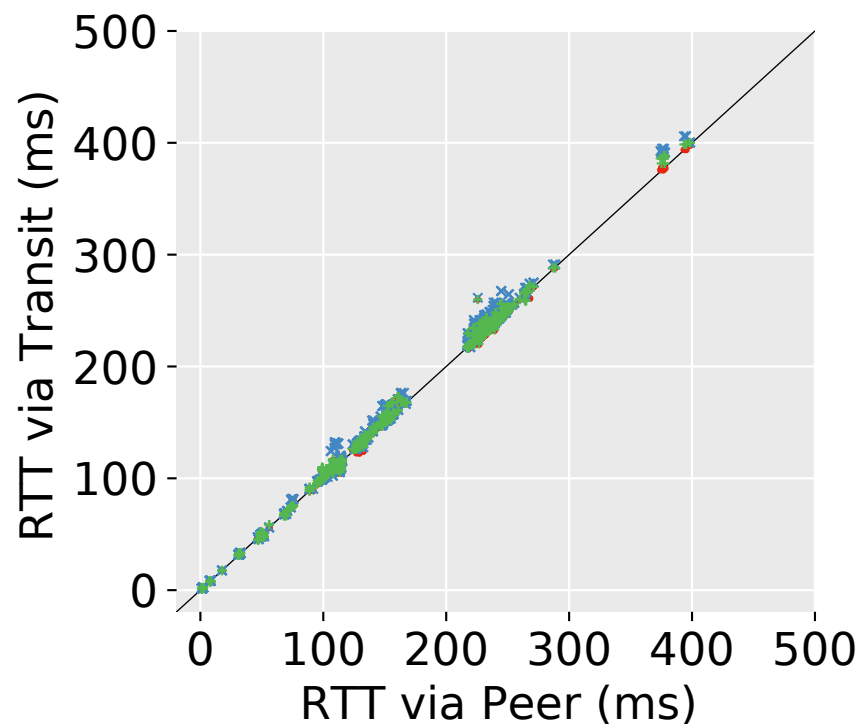
Peer vs Transit: Asia

- 実際どれくらい差がありそうか？
 - $RTT(\text{transit}) - RTT(\text{peer}) = \text{Peer経由で短くなるlatency}$
 - この値(reduced latency)を宛先アドレスごとに出して累積密度分布(CDF)でプロット



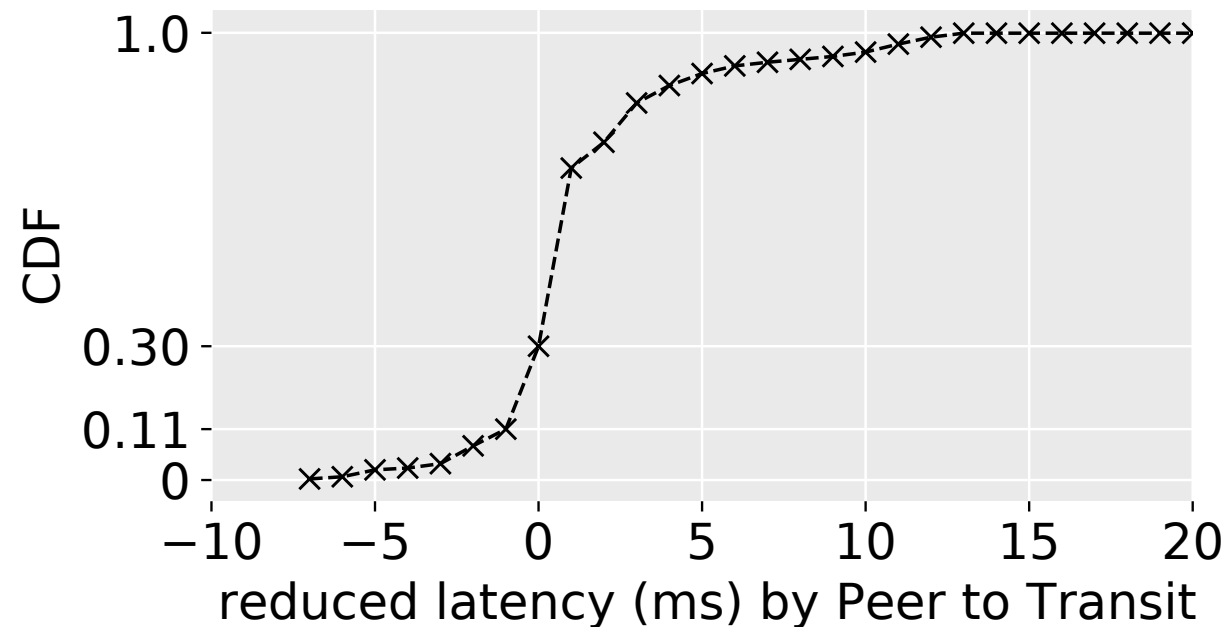
Peer vs Transit: North America

- 北米までのRTTが最もTransitとPeerによる違いがなさそうな傾向
 - 若干上方向に膨らみが見て取れる



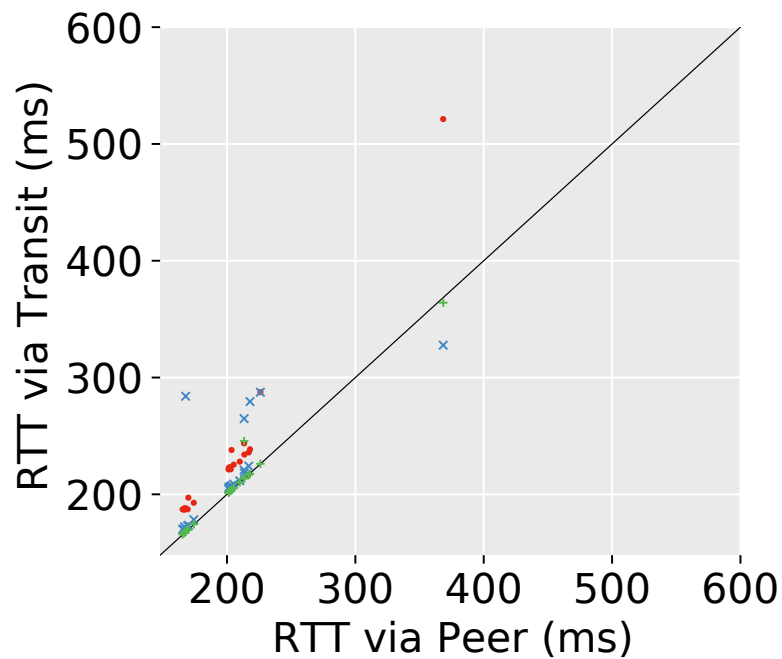
Peer vs Transit: North America

- CDFでみるとRTTはやはりPeer優位
 - Transitの方が短かったのは30%
 - 誤差(RTTは長くなりがち)があることに注意

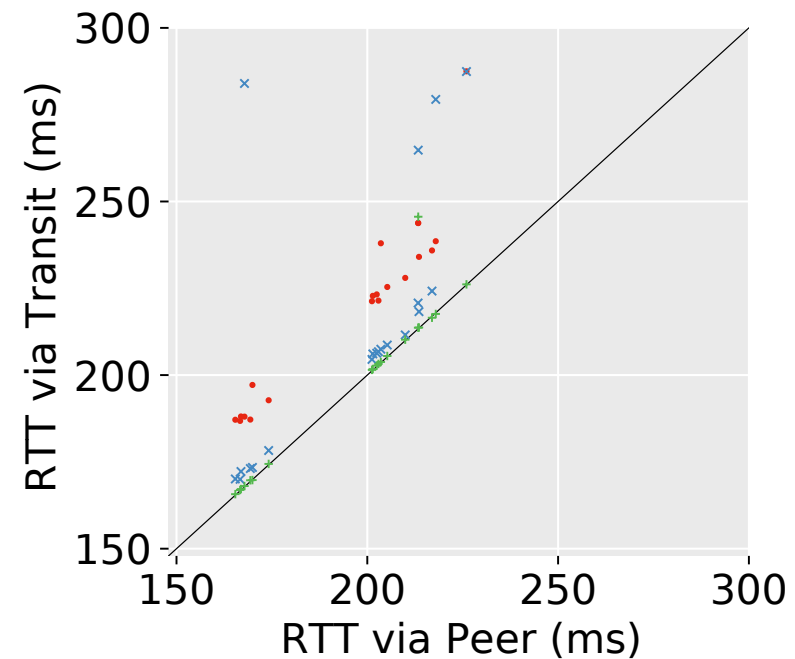


Peer vs Transit: Europe

- Peer経路に対して、Transit ASによるRTTの差が大きい
 - おそらくどのファイバルートを通るかに依存?



300msまでを
拡大



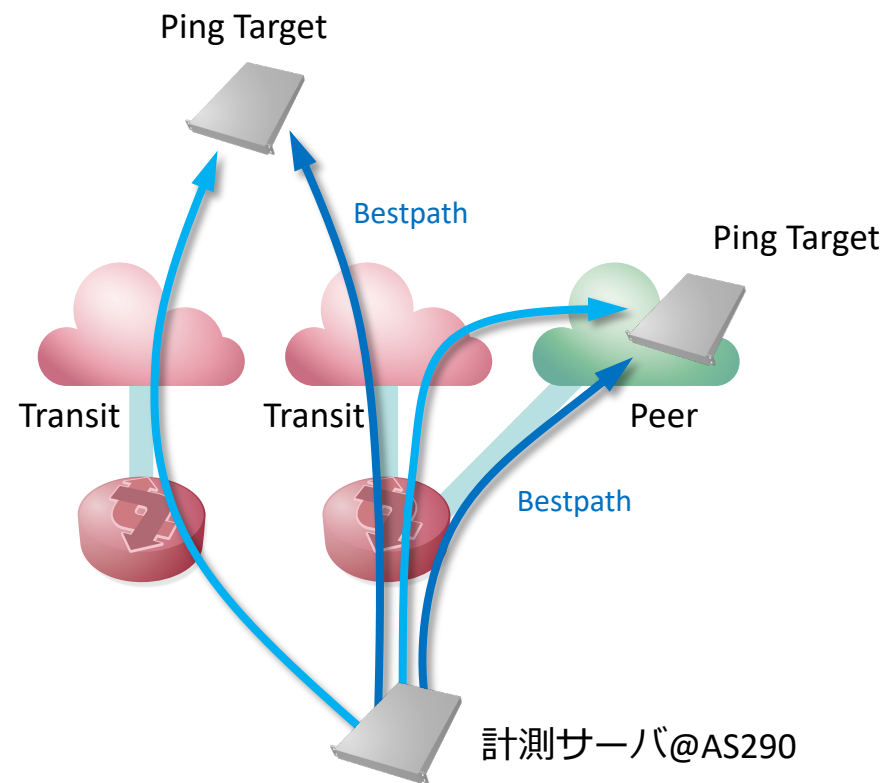
Peer vs Transitまとめ

- PeerはTransitよりどれくらい近いのか？
 - 全体の傾向としては、あまりかわらない
 - それでもやはりPeer経由の方がRTTは短くなる
 - 地域によって異なる
 - 日本国内は、PeerもTransitもどっこいどっこい
 - 北米もどっこいだが、よりPeerの方がRTTが短い傾向があった
 - 70%の宛先について、Peerの方がTransitよりgt 0ms 短かった
 - ヨーロッパは明らかにPeerの方がRTTが短く、おそらくファイバレートによるものと思われる

**計測結果2:
BestpathはどれくらいBestなのか？**

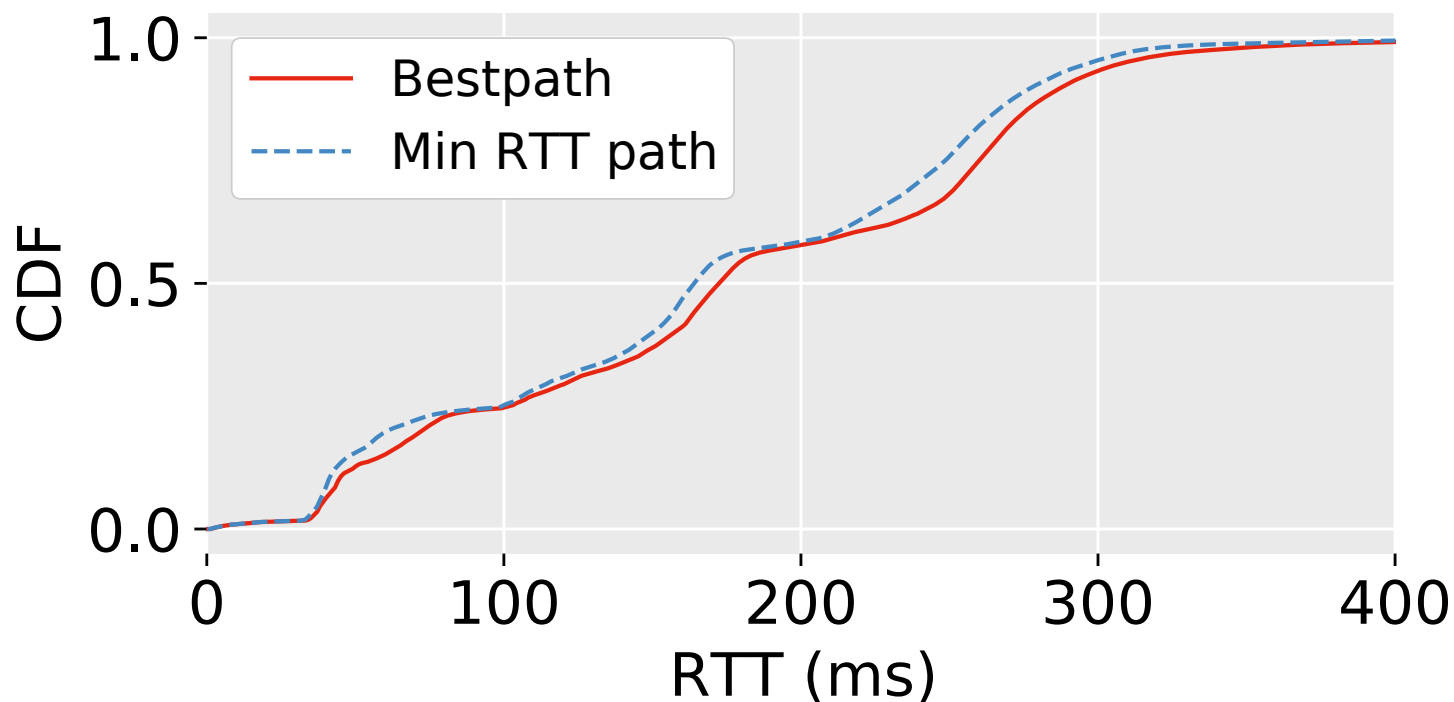
解析の概要

- 全ての宛先IPアドレスについて
 - ✓ BestpathなEgress AS経路のRTT
 - ✓ Bestpathを含む全Egress AS経路のRTT
 - 上記それぞれの最小値を比較
 - つまり、BestPathと、EPE含めた最短RTTのPathの比較
- ShowNet側の制御
 - 国内Peerを優先

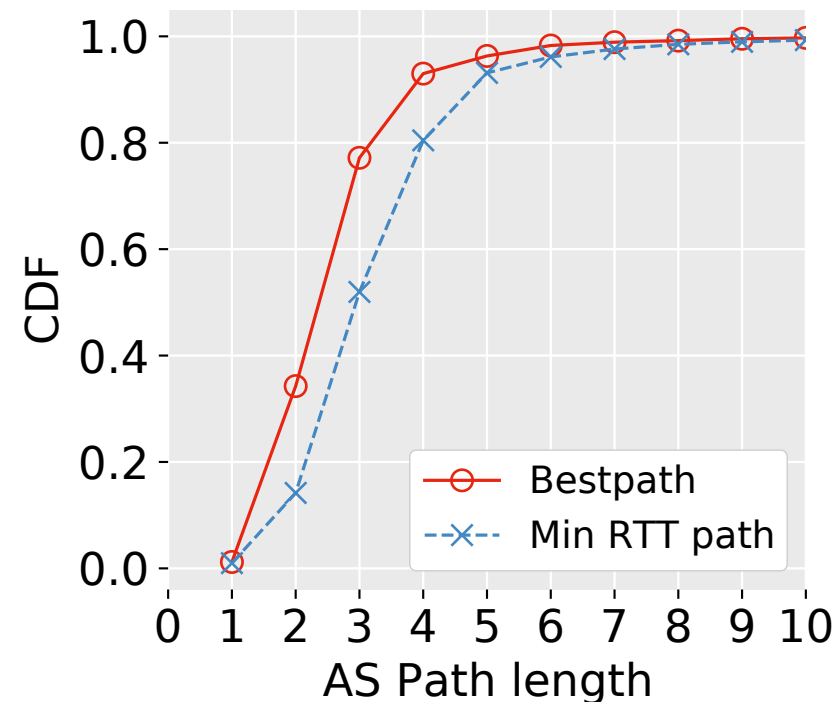


BestpathとMinimum RTT Pathを比較

宛先アドレスごとにBestpath経路のRTTと最も短かったRTTの分布

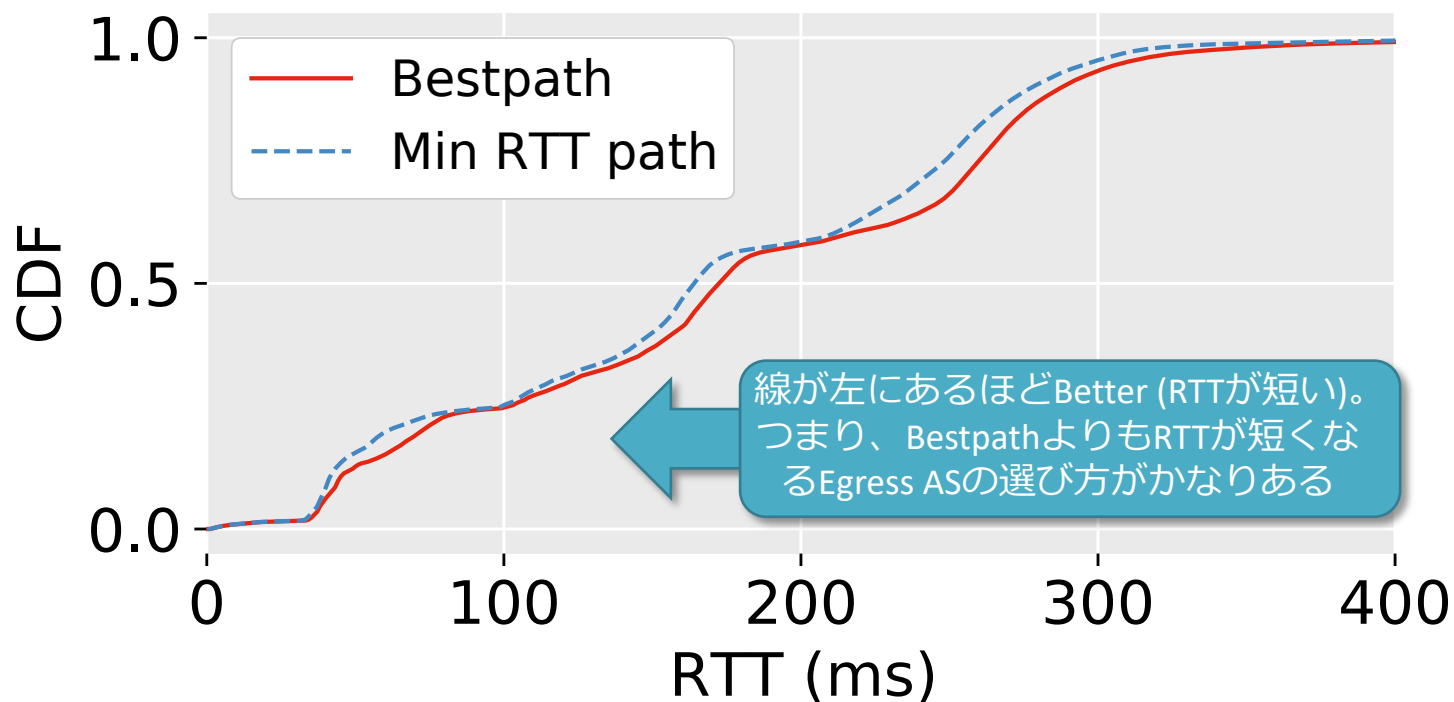


宛先ごとのAS Path長比較

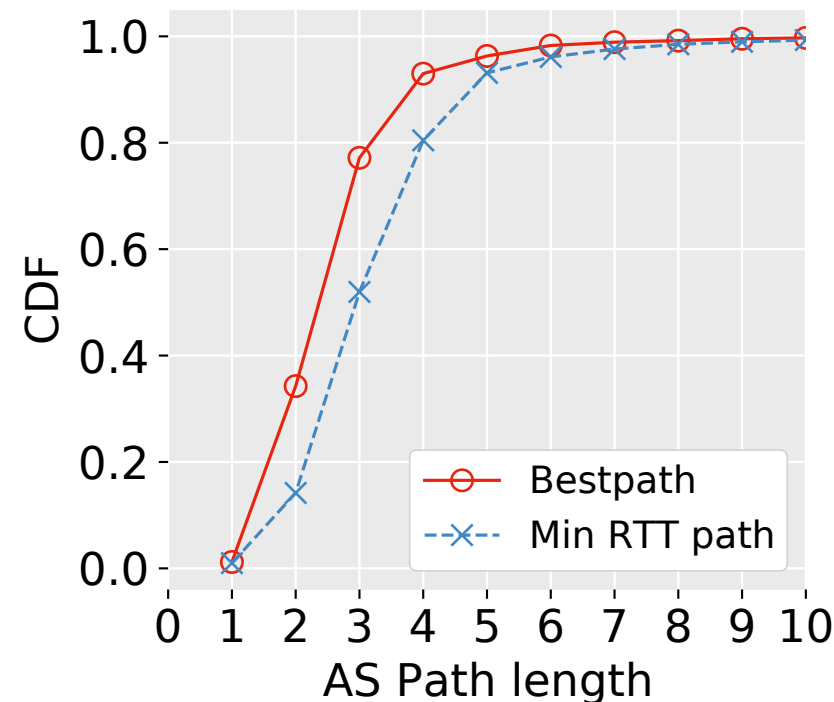


BestpathとMinimum RTT Pathを比較

宛先アドレスごとにBestpath経路のRTTと最も短かったRTTの分布

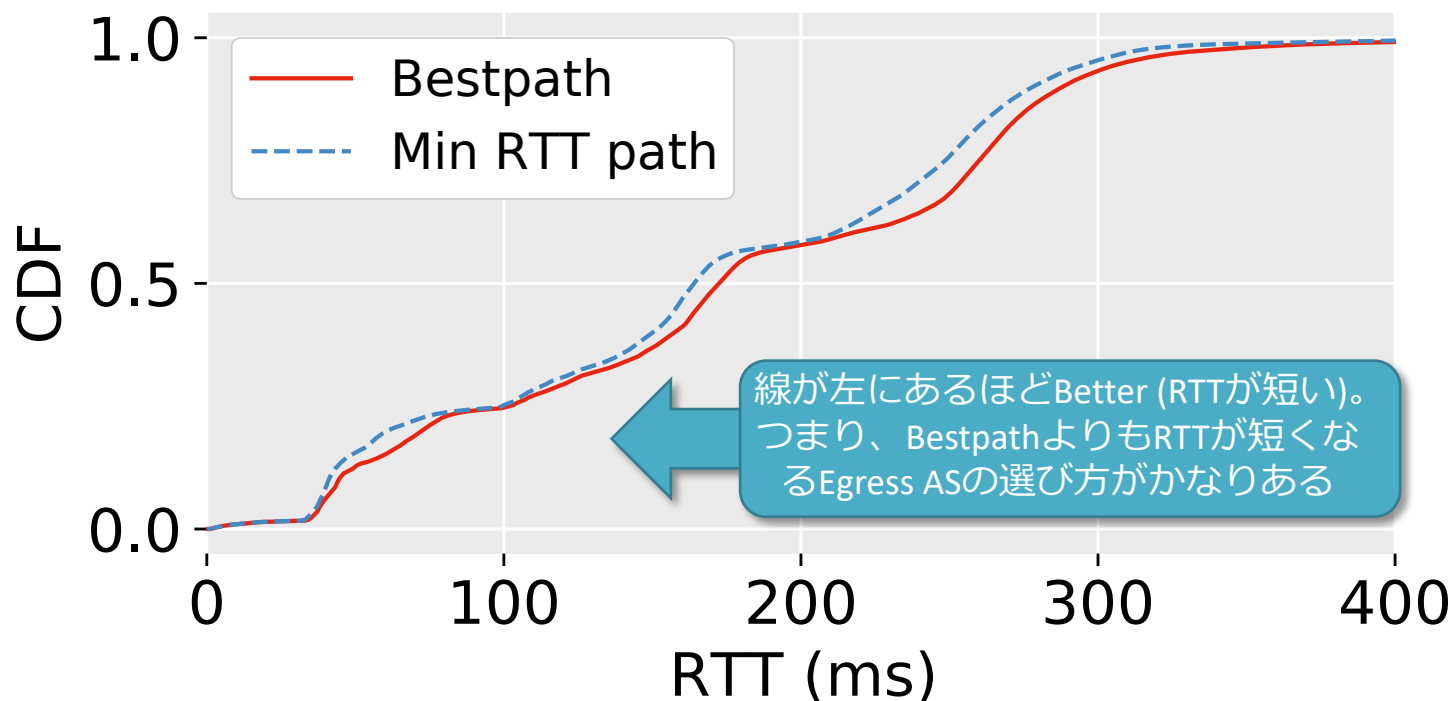


宛先ごとのAS Path長比較

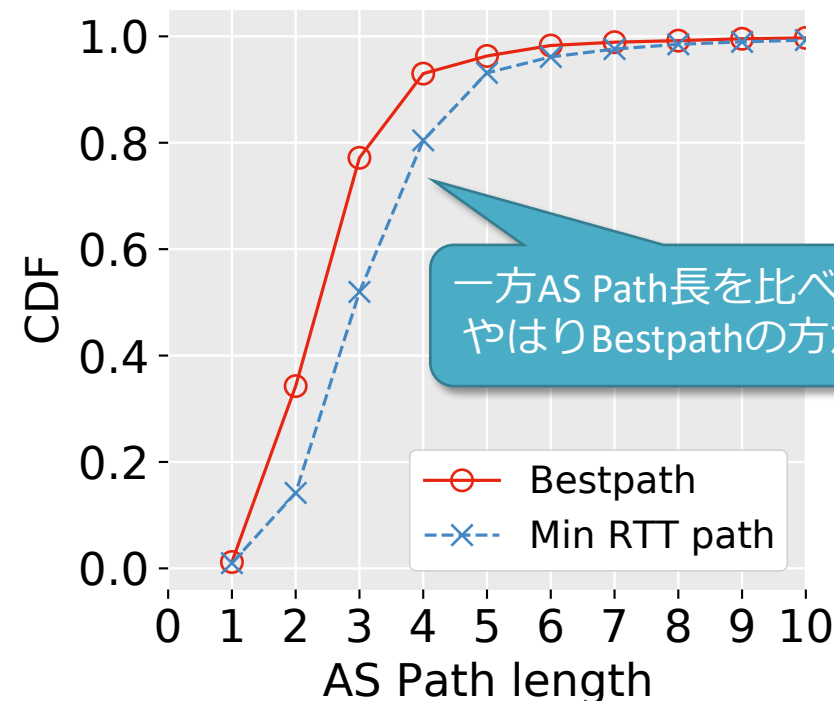


BestpathとMinimum RTT Pathを比較

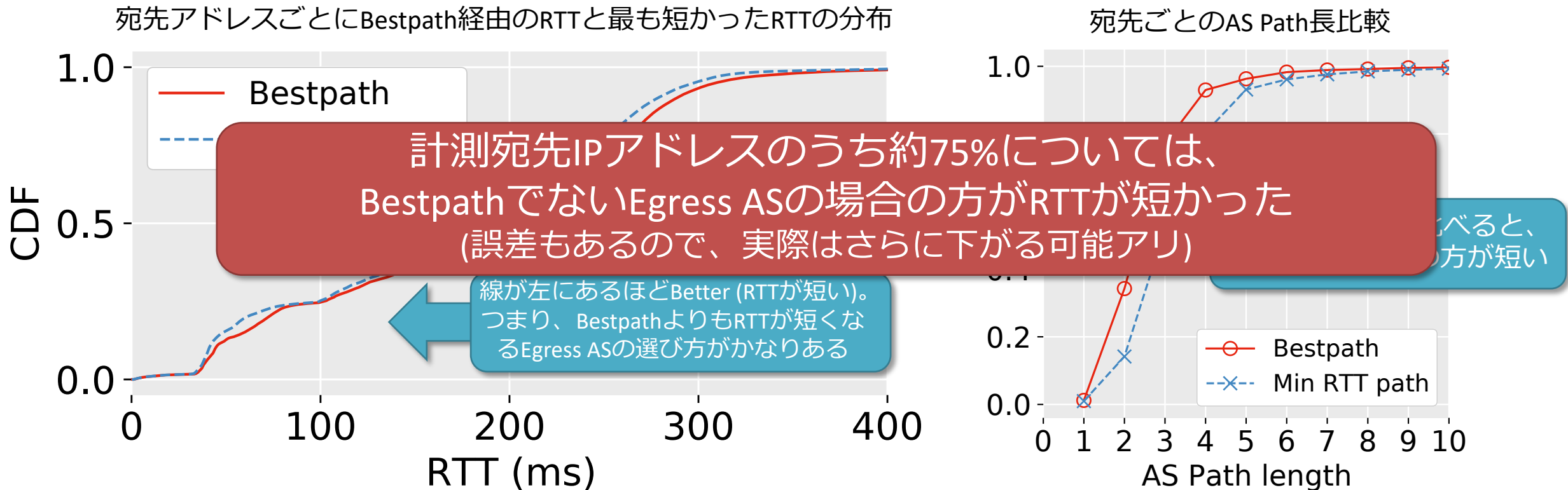
宛先アドレスごとにBestpath経路のRTTと最も短かったRTTの分布



宛先ごとのAS Path長比較

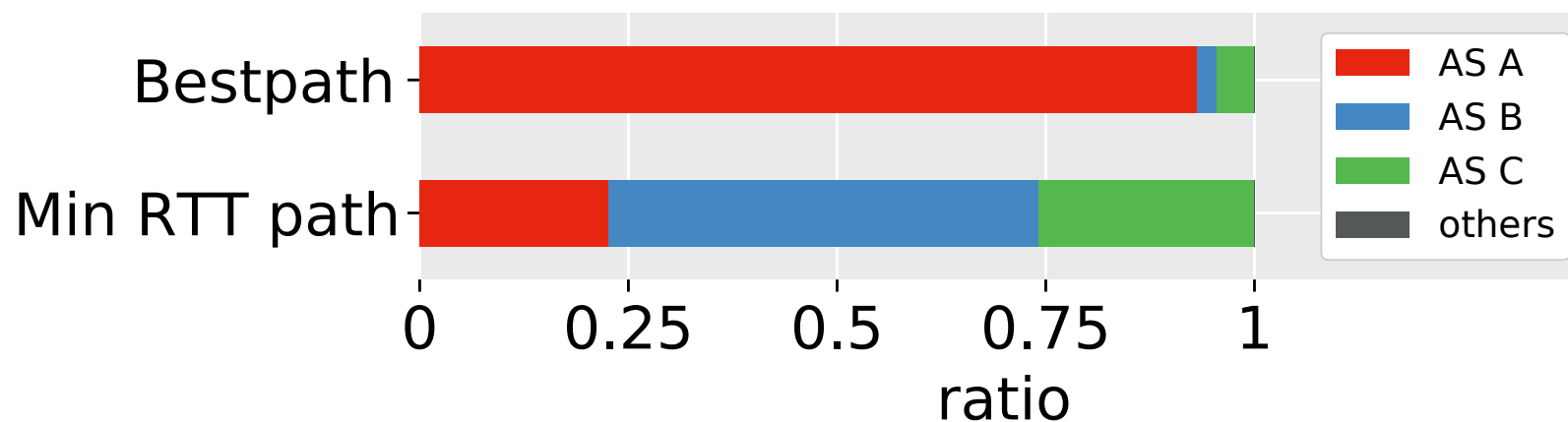


BestpathとMinimum RTT Pathを比較



BestpathとMinimum RTT Pathを比較

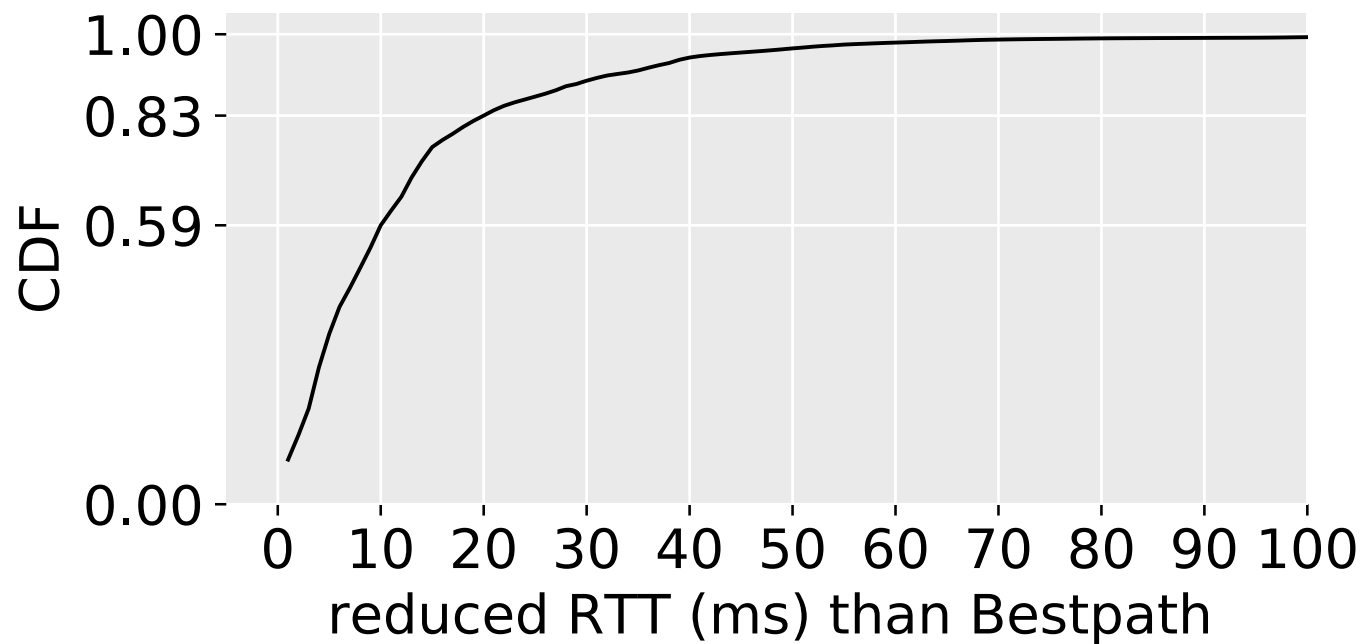
- Egress ASの割合
 - BestpathではあるASがEgress ASの約93%を占めていた
 - Minimum RTTになったEgress ASの割合では22%



- AS Path長は長いがRTTが短いパス、というのは結構ある

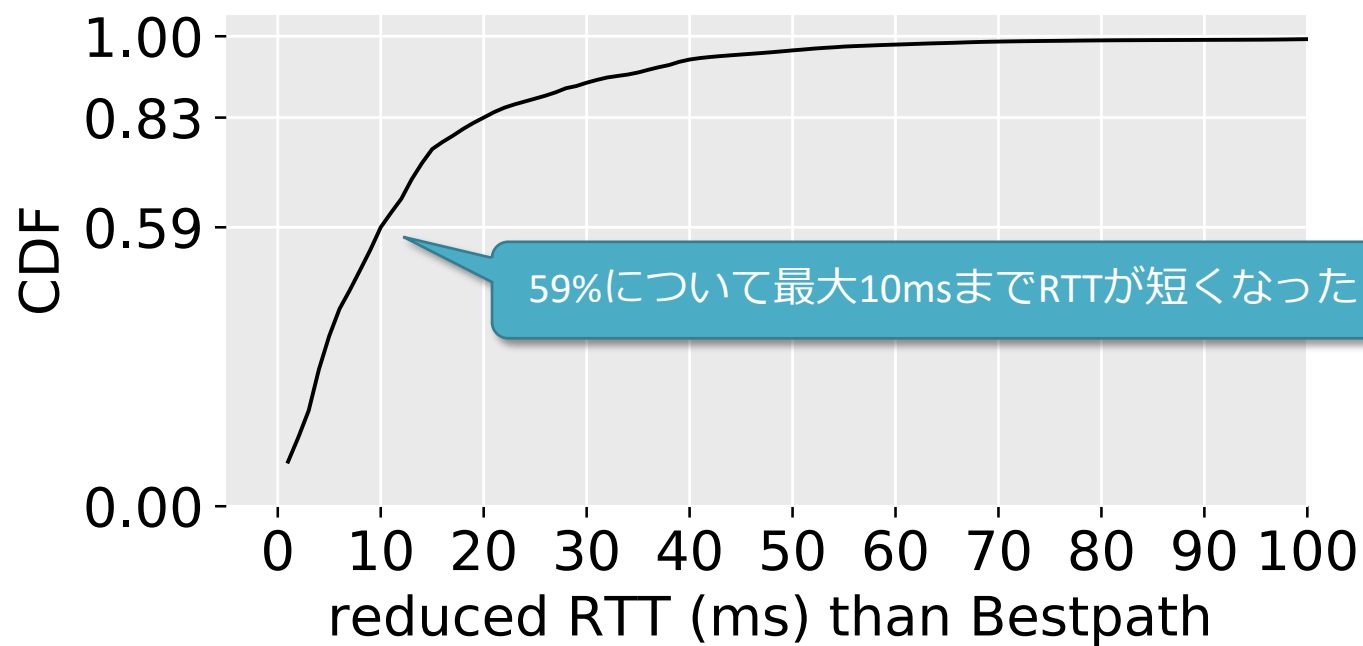
BestpathとMinimum RTT Pathを比較

- EPEによって短くなったRTTの分布
 - $RTT(\text{bestpath}) - RTT(\text{min}) = \text{EPE}$ で短くなるレイテンシー
 - この値が0より大きかったもの(全宛先のうち約75%)を累積密度分布(CDF)でプロット



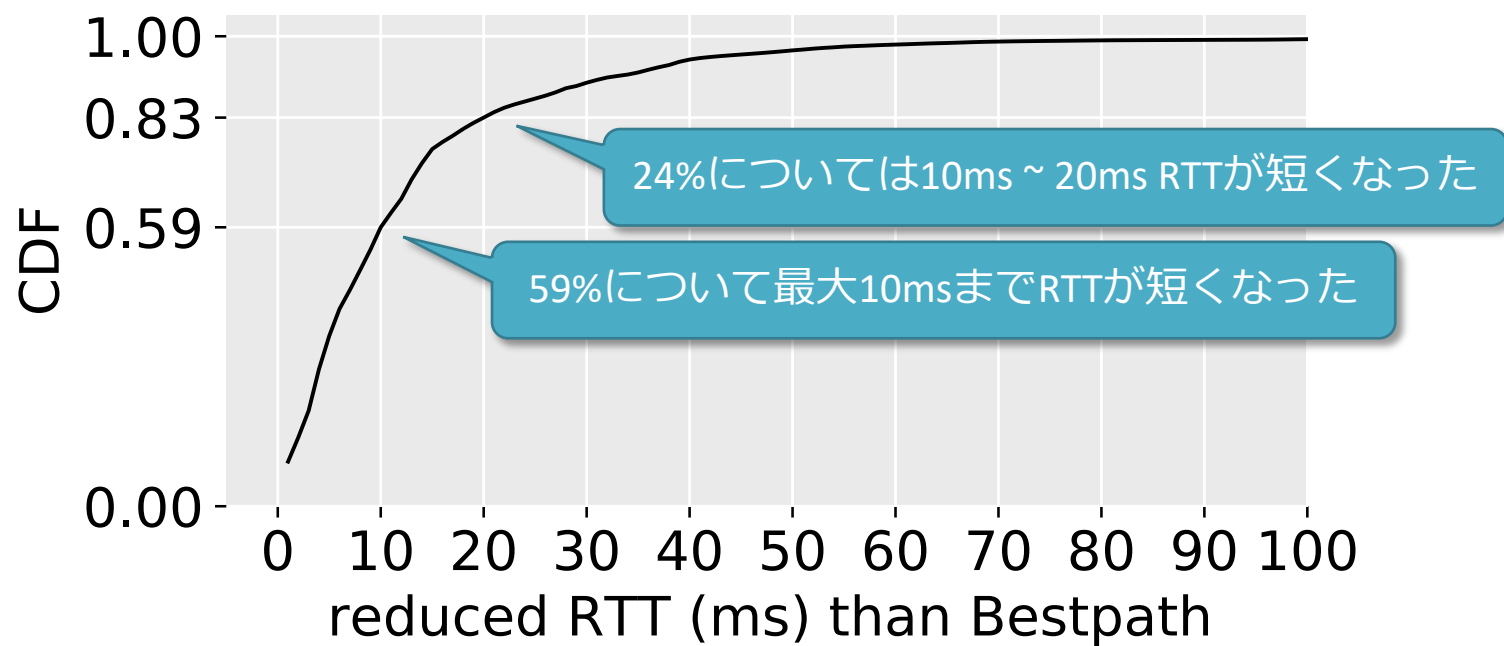
BestpathとMinimum RTT Pathを比較

- EPEによって短くなったRTTの分布
 - $RTT(\text{bestpath}) - RTT(\text{min}) = \text{EPE}$ で短くなるレイテンシー
 - この値が0より大きかったもの(全宛先のうち約75%)を累積密度分布(CDF)でプロット



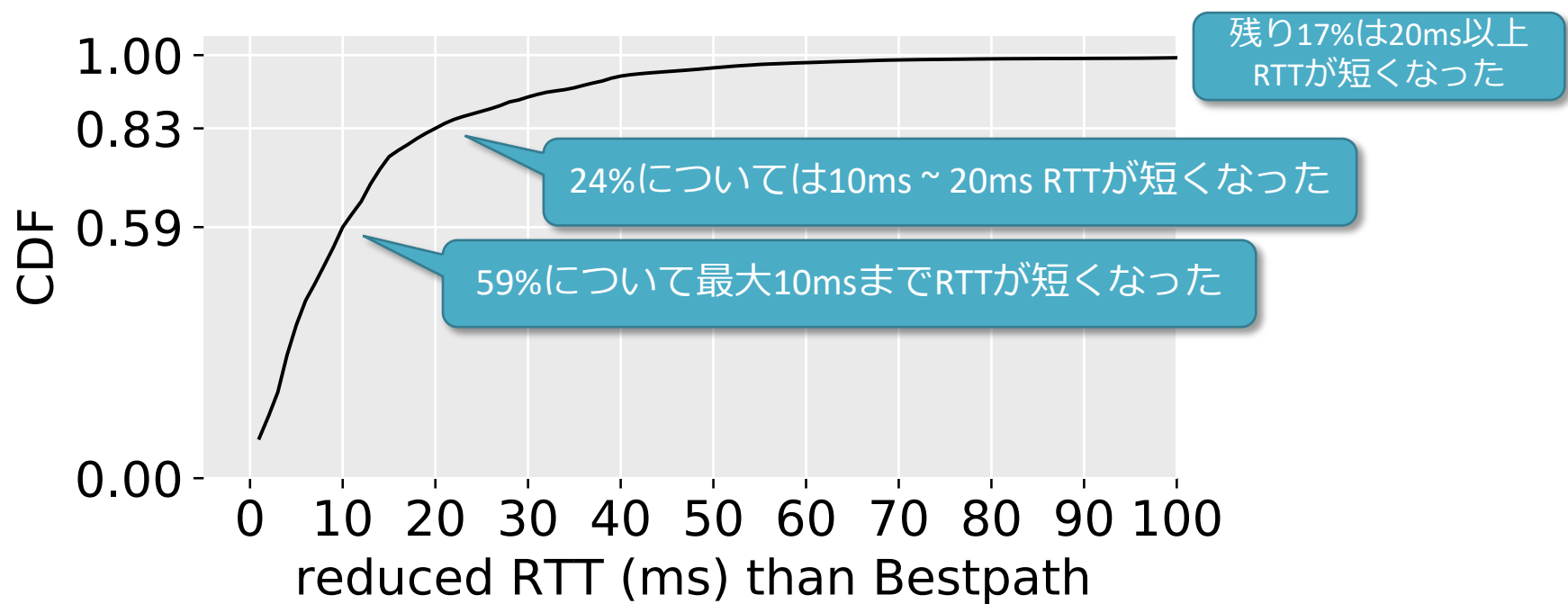
BestpathとMinimum RTT Pathを比較

- EPEによって短くなったRTTの分布
 - $RTT(\text{bestpath}) - RTT(\text{min}) = \text{EPE}$ で短くなるレイテンシー
 - この値が0より大きかったもの(全宛先のうち約75%)を累積密度分布(CDF)でプロット



BestpathとMinimum RTT Pathを比較

- EPEによって短くなったRTTの分布
 - $RTT(\text{bestpath}) - RTT(\text{min}) = \text{EPE}$ で短くなるレイテンシー
 - この値が0より大きかったもの(全宛先のうち約75%)を累積密度分布(CDF)でプロット



BestpathとMinimum RTT Pathまとめ

- レイテンシーという観点から見たEPEを使う効果
 - BestpathよりもRTTが短いパスは結構ある
 - 今回の計測では約263万宛先IPアドレスのうち約75%
 - EPEでBestpathよりもRTTを短くできたのは全宛先のうち75%
 - その75%のうち、59%について最大10msまでRTTが短くなった
 - その75%のうち、24%については10ms ~ 20ms RTTが短くなった
 - その75%のうち、17%については20ms以上RTTが短くなった
 - (くだいですが)誤差があります
 - RTTは長い方に誤差がでるので、Bestpath経由のRTTがたまたま長く計測されていたアドレスもあり得ます

SR-MPLS Egress Peer Engineeringを 使って測るインターネットレイテンシー

- AS290 ShowNetからEgress ASを変えながらping/traceroute
 - Peer vs Transit
 - 全体の傾向はあまりかわらないが、それでもやはりPeer経由の方がRTTは短い
 - 地域差も大きい。海外についてはおそらくファイバルートに依存
 - Bestpath vs Minimum RTT Path
 - BestpathよりもRTTが短くなるPathはかなりあった(宛先のうち約75%)
 - その75%のうちの、59%は最大10ms、41%は10ms以上RTTが短くなった
- 誤差(以下略)
- ShowNetという単一ASでの計測結果です
 - 対外接続のあり方や、戻りの経路など、結果が変わる要素はたくさん

おわりに

- EPE、使ってみたくありませんでしたか？
 - RTTだけみても、確かに短くできる
 - 一方EPEを使わなくてもBGPまわりを見直す意味はあるかも
- ここまでの解析はpingの結果のみ
 - tracerouteのデータも...
 - 続きは(きっと)論文で
- Interop Tokyo ShowNet
 - 相互接続性検証とデモンストレーション
 - そしてコミュニティへの還元



Special Thanks



and, all ASes peering ShowNet!

来年も、どうぞよろしく願いいたします