CDN事業者の Egress トラフィック エンジニアリング - SIGCOMM2017発表から -

小谷 大祐 (京都大学 学術情報メディアセンター)

{CDN, OTT, コンテンツ} 事業者, Hyper Giant

- コンテンツをインターネットの利用者に配信している事業者
  - Google, Facebook, Akamai, J-Stream, など
  - テキスト、画像、ビデオなど
- 巨大なトラフィックを生み出す
  - Facebook: 数Tbps レベル [Shuff 2015]
  - Akamai: 数十Tbps レベル (2017/09/15 のプレスリリース)

https://www.akamai.com/uk/en/about/news/press/2017-press/new-akamai-peak-traffic-record-demonstrates-increased-importance-of-live-online-events.jsp

コンテンツ配信の最適化

Webサイトの応答時間 (ページが表示されるまでの時間) が重要
[Nielsen 2010]

- 時間がかかるとユーザが離れてしまう

https://www.nngroup.com/articles/website-response-times/

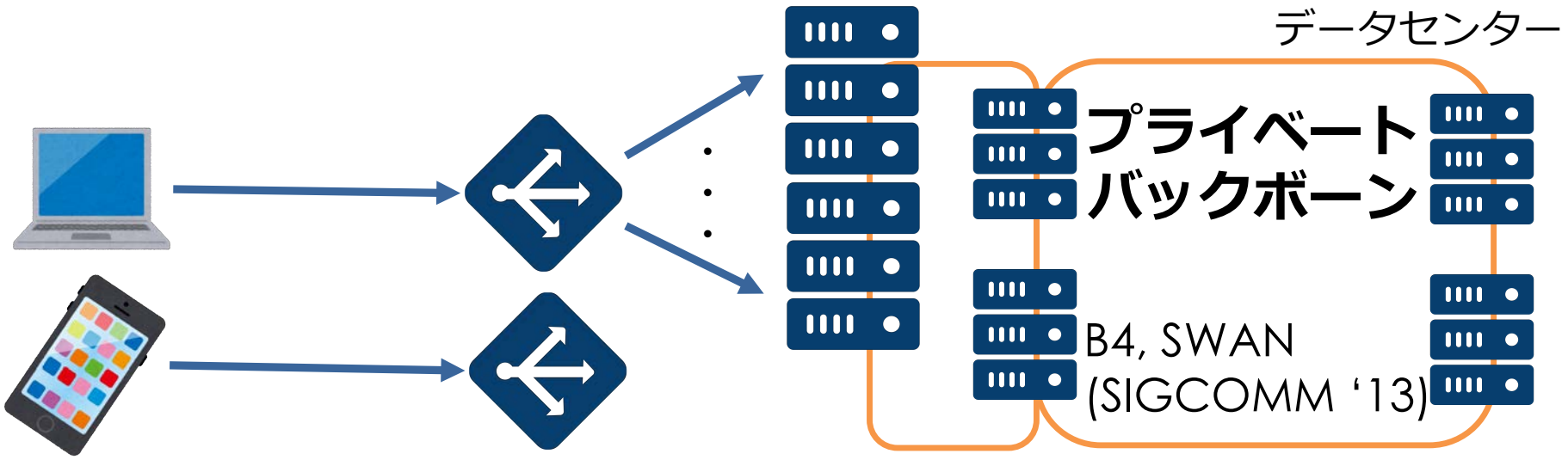
- 例: プロトコルの改良
  - 複数のTCPコネクション → 一つの TCP コネクション (SPDY, HTTP/2)
    - コネクション数の削減
    - TLS の handshake やヘッダの重複のオーバヘッドの削減
    - TCP の Slow Start の影響の軽減
    - プライオリティ
  - TCP → UDP (QUIC)
    - TCP の Head of Line Blocking 問題の影響の軽減

大規模コンテンツ事業者のコンテンツ配信基盤

- 膨大なトラフィックを、効率よく小さい応答時間で捌く

100Gbps/IP 以上までスケールアウトするロードバランサ

ECMP や Consistent Hashing 等を利用
Ananta (SIGCOMM '13), Maglev (NSDI '16)

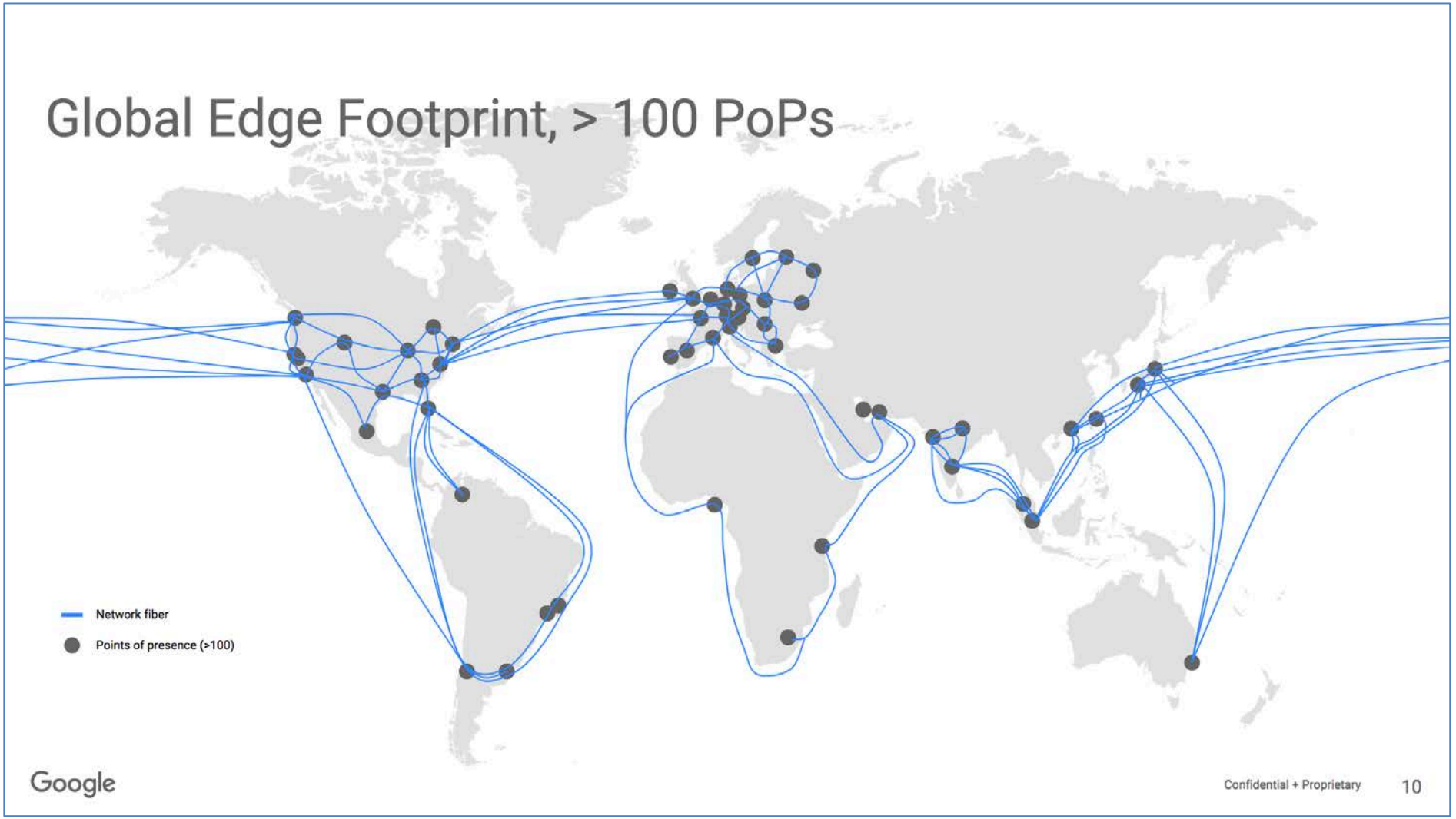


今回はこちら

世界中に PoP (Point of Presence) と Proxy を設置

TCP や TLS の Handshake 時間の短縮, 静的コンテンツのキャッシュ

例: Google の PoP



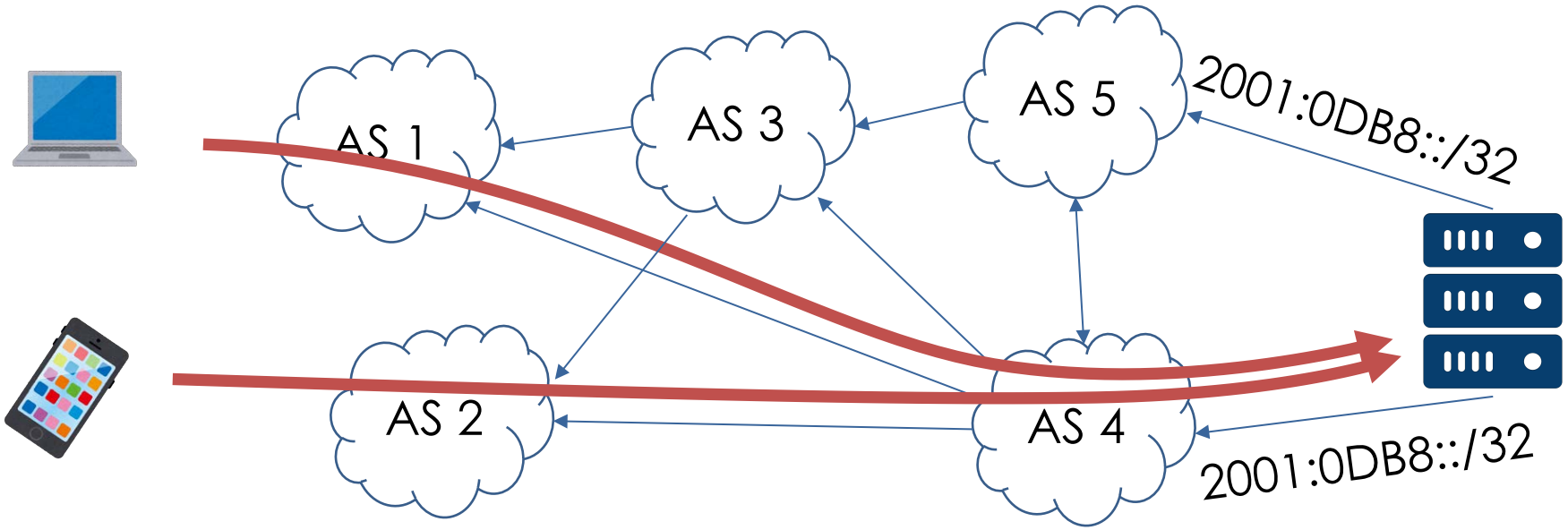
出典: ACM SIGCOMM '17 の KK Yap 氏の発表資料より

BGP の限界

各ルータは Best Path に従ってパケットを転送するので複数の経路があったときに有効活用できない

- 例: Best Path ではないが遅延が少ないパスを使う
  輻輳していないリンクに少しトラフィックを迂回させる

複数の PoP のサーバの負荷やトラフィックの状況に応じた PoP 間の負荷分散が困難



利用者のトラフィックの制御

Ingress (利用者 → コンテンツ事業者):
広域負荷分散技術による最適な PoP の選択

- Anycast や DNS を活用、PoP のサーバの負荷も考慮
- FastRoute [NSDI'15] by Microsoft
  End-User Mapping [SIGCOMM '15] by Akamai

Egress (コンテンツ事業者 → 利用者):
パケットが最適なパスを通るように制御

- Peering /Transit のリンクの選択
- Edge Fabric [SIGCOMM '17] by Facebook
  Espresso [SIGCOMM '17] by Google

BGP で得られる経路を活用しつつ、
従来の経路制御だけには頼らない (ルールを無視した)
トラフィックの制御

本セッションのご講演

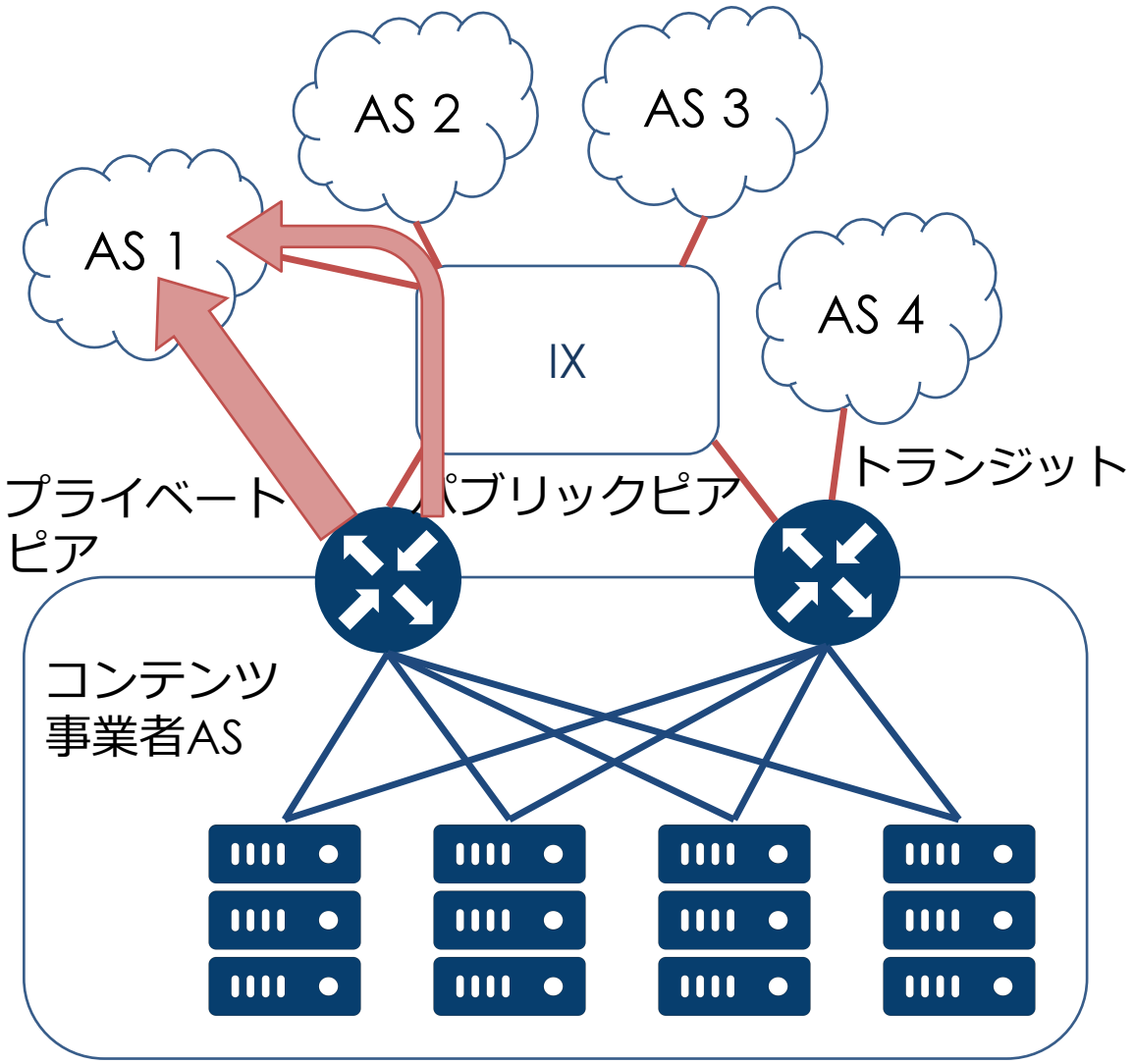
CDN のトラフィックエンジニアリング
鍋島 公章 様 (J-Stream)

CDN の価値観や今後の方向性

BGPによるドメイン間経路制御の現状と将来：障害事例と対策
西塚 要 様 (NTTコミュニケーションズ)

BGPで制御できないトラフィックがある状況で
今後の経路制御はどうしていくべきか

Egress トラフィックの制御の例



AS 1 への送信:
普段はプライベートピア経由

When:

- プライベートピアの輻輳
- プライベートピア経由の通信の品質の悪化
(遅延の増加や輻輳の発生)

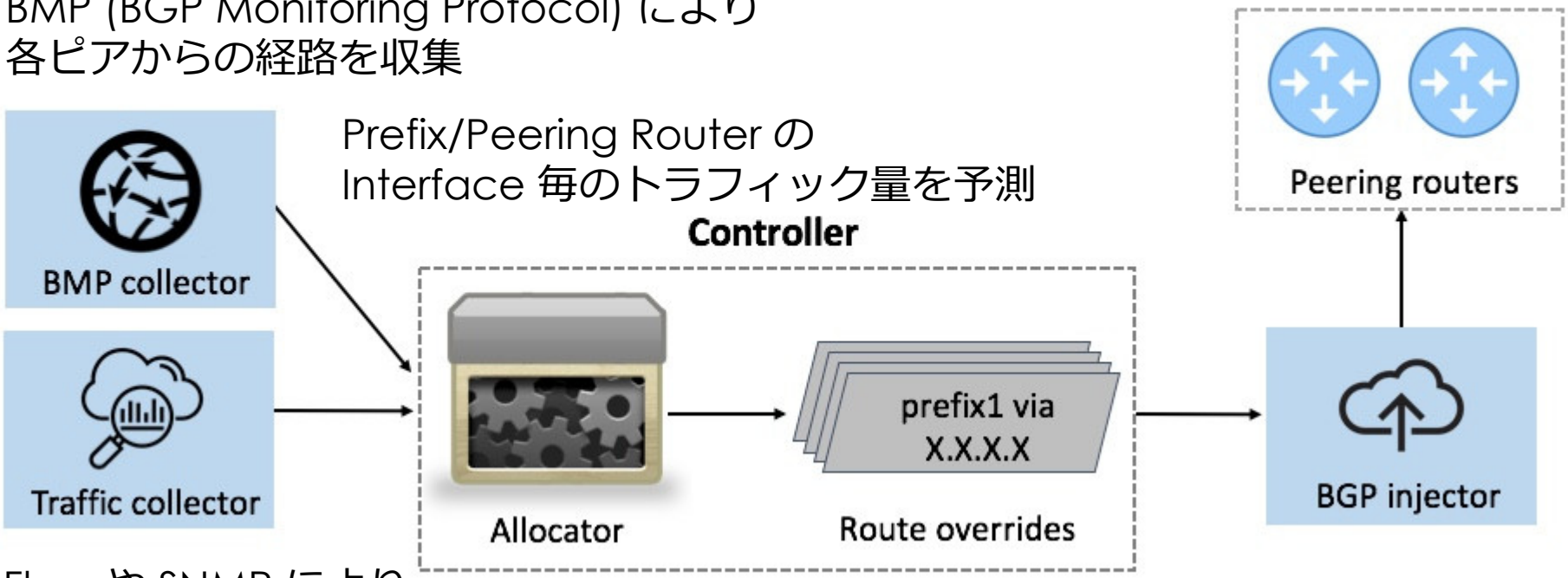
Then:

- IX/トランジット経由へ
- ビデオはプライベートピア、テキストや画像はIX/トランジット経由へ

Edge Fabric (Facebook)

できるだけ BGP Best Path に従いつつ、
それに従わない経路を BGP の経路を注入し上書き

BMP (BGP Monitoring Protocol) により
各ピアからの経路を収集



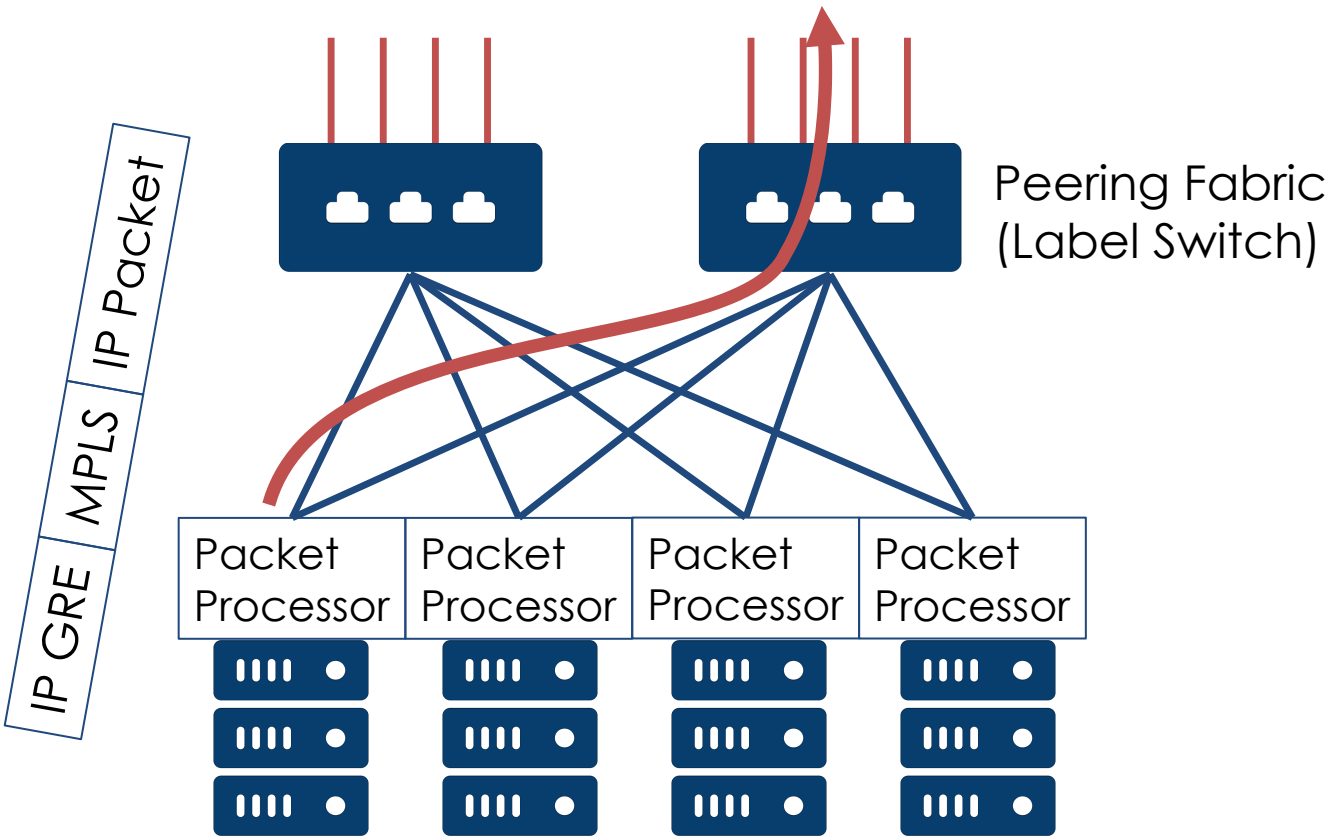
sFlow や SNMP により
Prefix/ Peering Router の
Interface 毎のトラフィック量を取得

輻輳が起きないように
上書きする経路を生成
(Prefix の分割を含む: 1 x /20 → 2 x /21)

図は https://research.fb.com/steering-oceans-of-content-to-the-world/ より引用、一部追記

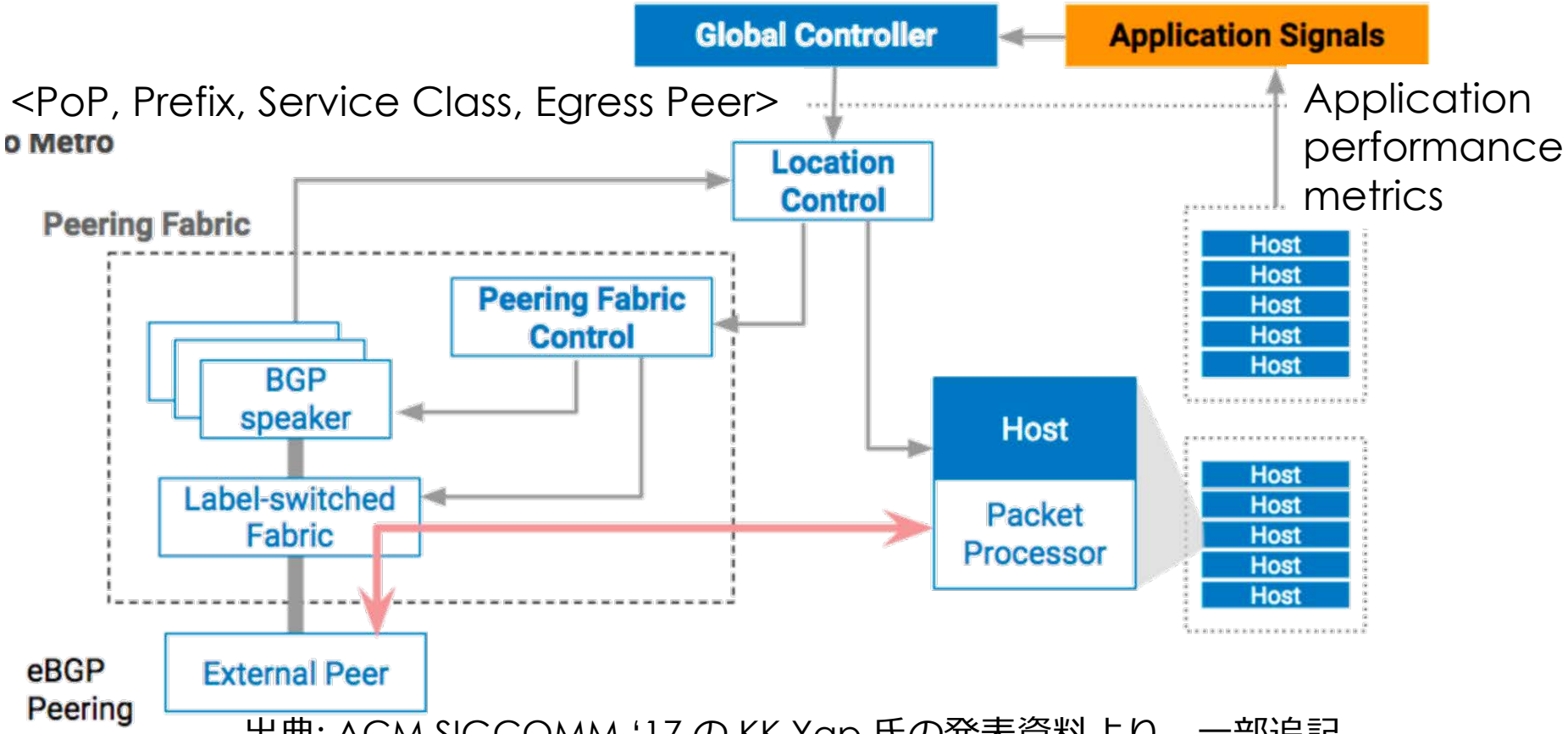
Espresso (Google)

Host の Packet Processor で Egress の Peer を指定 (ラベルづけ) → Edge Fabric より細かな単位で制御可能



Espresso (Google)

トラフィック予測や BGP で得られる経路情報を元に
遅延やスループットが同程度の Prefix 単位で
どの Service Class がどの Peer を利用するか選択



出典: ACM SIGCOMM '17 の KK Yap 氏の発表資料より、一部追記

従来の経路制御との違い

Application Performance のためには
従来の経路制御の仕組みに従わないことがある

- Edge Fabric の例:
  - Peer によっては Longest Prefix Match に従わないことがある
  - Peer から広報されてきた経路を細分化して扱う場合がある
- Espresso の例:
  - Peer 先から広報されてきた経路を、Prefix 内のホストとの遅延が同程度になるまで Prefix を細分化して扱う

トラフィックを吐き出すサーバと連携して制御できる

- 通信の両端のうち片端が意図的に変更される
- BGP によるトラフィックエンジニアリングが効かない

まとめ

- Hyper Giant は世界各地に PoP を置いている
- Hyper Giant にとっては Application Performance が重要
- Performance を優先した Egress の選択を行い、結果として BGP の Best Path に従わない場合がある
  - BGP が performance aware ではない

- このような制御がインターネットの経路制御に及ぼす影響を懸念
  - ミスするとインターネット全体への影響が大きい制御をやる事業者が増えると、経路障害の頻度が増えそう
  - 従来のトラフィックエンジニアリングの手法が効きづらくなる