

グローバルインターネットにおける 大容量トラフィックコントロールとリソースマネジメント

2014年1月23日

NTTアメリカ 科学忍者隊IPENG
ショーン モリス
吉村 知夏

NTTコミュニケーションズ株式会社
清水 香里

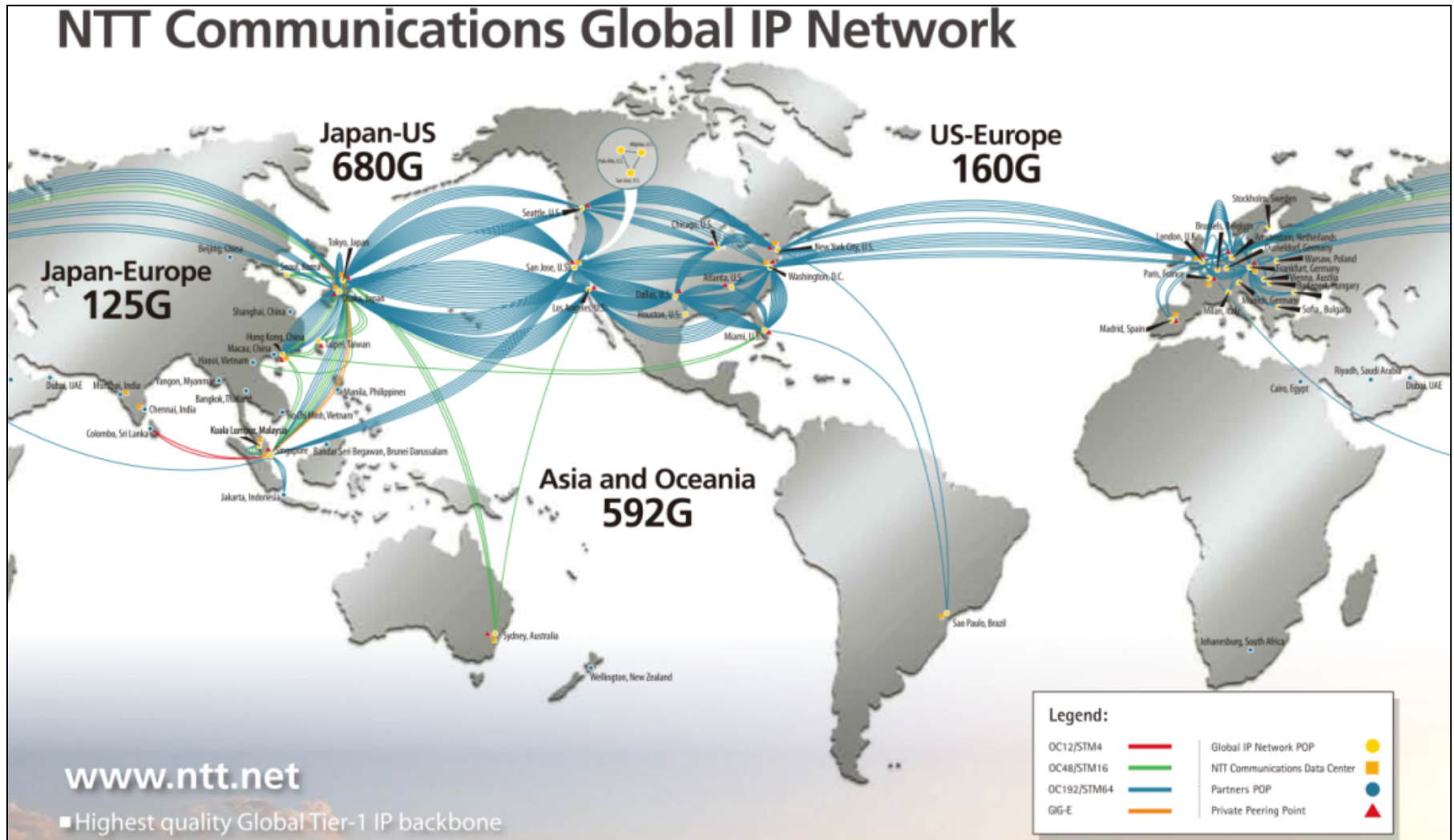
目次

1. グローバルIPネットワーク(GIN)の紹介
2. グローバルインターネットの“ベストデザイン”とは？
3. なぜ、グローバルIPネットワークにMPLS-TE？
4. 昨今のトラフィック状況サマリー
5. 予想できないトラフィックを如何にさばくか？ (How to)

In English

1. What is GIN?
2. “Best design” for Global Internet
3. Why to implement MPLS-TE for GIN
4. Current traffic summary
5. How to do for the current huge traffic

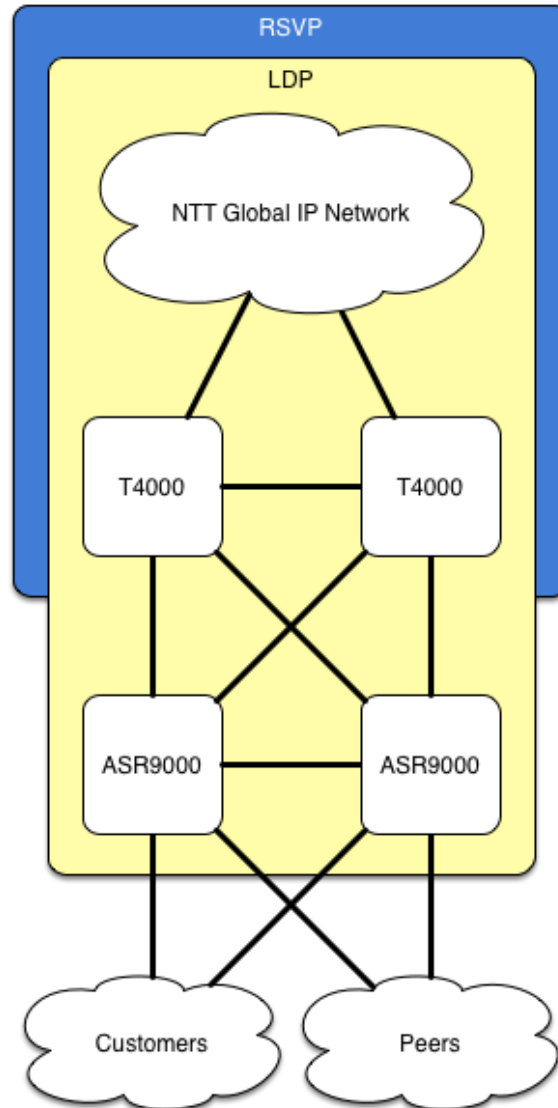
1. グローバルIPネットワーク(GIN)の紹介 (1/2)



1. グローバルIPネットワーク(GIN)の紹介 (2/2)

- IGP is IS-IS
- Roughly 150 routers in backbone iBGP mesh
- BGP confederated sub-ASNs are used for affiliated companies/groups
- No route reflectors, flat BGP topology
- Current FIB is ~484k routes
- Current RIB is ~4.9mm routes
- Juniper T-Series Core
- Cisco ASR9000 Edge
- Junipers run a full mesh of RSVP-TE LSPs
 - A limited number of ASR9000s as well
- Pseudowire Ethernet service uses LDP tunneled within the RSVP infrastructure

MPLS/POP Architecture



2.1 グローバルインターネットの“ベストデザイン”とは？

- In networking, there is no such thing as a “Best Design” that remains true for all time.
- Every decision in networking is a result of **cost benefit analysis** given the available technology and the **business problem** at that moment.
- NTT GIN is built primarily as a wholesale IP backbone, multi-service (L2VPN) came later
- Others MPLS networks are purpose built for other services (L1 transport, L2 transport, L3 VPN)
- Other MPLS networks are built to support multiple services or to be aggregators of other networks

- 「ベストデザイン」の正解は無い。何がベストかは、時によって変わる
- コスト分析とビジネスニーズに基づくデザイン
- GINは元々はトランジットを提供するIPバックボーンだったが、その後L2VPNを提供

BGP-Free Core vs. full iBGP

- GIN Backbone runs a full iBGP mesh
- Network is designed to be flat (customers and peers can land on core routers)
- No desire to extend MPLS-TE mesh to pure edge devices
- We feel we would be unable to rely on platforms that would deliver promised capex savings
- IP traffic will still forward in case of MPLS failure
- IPv4/v6 dual-stack support in RSVP-TE was lacking until recently

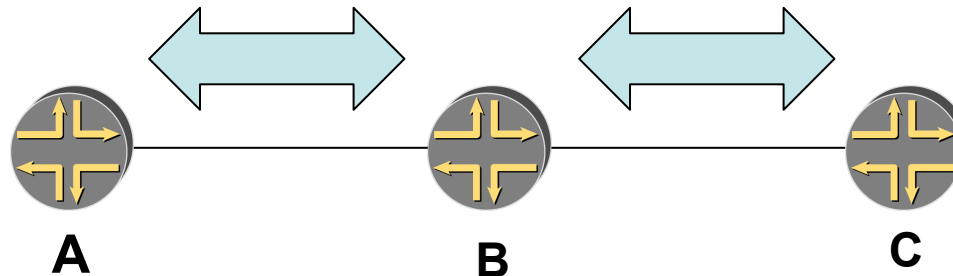
- **バックボーンはiBGPのフルメッシュ構成（ルートリフレクタ無し）**
 - **MPLSダウン時にIPフォワーディングが可能**
 - **v6はIPフォワーディングが必要だった（RSVP-TEのサポートが後発だった）**
- **エッジルータでMPLS-TEは導入していない**

3.1 なぜ、グローバルIPネットワークにMPLS-TE?

- Originally, GIN network did not use MPLS and ran BGP and ISIS directly over the point to point links.
- MPLS/TE was introduced in 2002 for operational improvement reasons primarily due to three factors:
 - Better **visibility into router to router pair flow** traffic data
 - Ability to do pseudo **admission control** on traffic
 - **Finer grain traffic splitting** along parallel or equivalent paths
- **2002年にMPLS-TEを導入**
- **導入理由**
 1. ルータ間のトラフィックフローを見やすく可視化
 2. トラフィックのアドミッションコントロール
 3. よりきめ細やかなトラフィック分散制御

Also Traffic and Topology Growth!

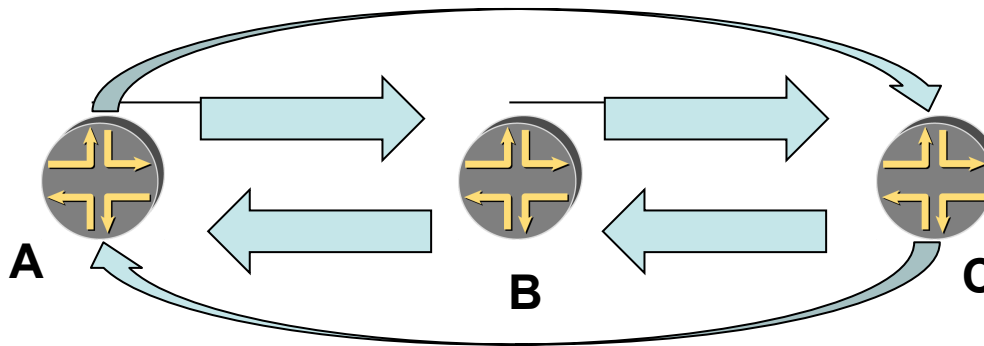
3.2 なぜ、グローバルIPネットワークにMPLS-TE?



- In traditional IP Networks, traffic statistics can only be collected on per interface basis
 - A \leftrightarrow B, B \leftrightarrow C traffic can be collected, but not A \leftrightarrow C.
 - A \leftrightarrow C can be interpolated based on traffic information from A, B and C, but as the network scales to more routers and links, the more **computationally expensive the interpolation, and lower the accuracy** of the result.
 - While other traffic collection technologies such as sFlow and Netflow can give similar data, in router vendor implementations, measuring LSP traffic is the most scalable and reliable way to measure router to router traffic in a complex topology network.
- 伝統的なIPネットワークでは、ルータをまたぐトラフィックデータの取得は難しい
- A \leftrightarrow Cのトラフィックフローを見たい場合、
 - A \sim B、B \sim Cのデータからがんばって計算する
 - flowデータ (Netflow, sFlow)
 - いずれも完璧ではない

3.3 なぜ、グローバルIPネットワークにMPLS-TE?

- ▶ With more accurate statistics, one can do better capacity planning and also day to day traffic management
- ▶ With MPLS, one can create a full mesh of LSPs between routers, limited only by manageability



- Once full mesh of LSPs are created,
 - A traffic matrix can be built between all nodes in the network
 - Traffic statistics can be immediately used to reconfigure the network (no topology changes needed)
- **A⇔CにLSPを設定することで、A⇔Cのトラフィックデータを取得できる**
- **フルメッシュのLSPを設定すれば、バックボーン内の全ルータ間のフローが把握可能**

3.4 ASN-based stats vs. node-based stats

- In GIN backbone ASN-to-ASN statistics are useful for traffic engineering on the edge.
- For the core, router-to-router statistics are more useful because traffic sources cannot be easily moved between POPs
- Additionally, in a very large IP Network ASN-to-ASN traffic matrix is a very computationally difficult problem to solve and analyze
- **バックボーンのトラフィックエンジニアリングのためには、ルータ間に流れるトラフィックフローや帯域を把握することが重要**
 - BGP ASフロー（Src/Dst ASのデータ）の把握も重要だが、どちらかという
とエッジ向け
 - ネットワークが複雑になるに伴い、ASフローデータを正確に把握することは
難しくなる
 - GINではフルメッシュLSPのフローデータを見ている

3.5 なぜ、グローバルIPネットワークにMPLS-TE?

- MPLS/TE does not restrict the number of LSPs between two origin and destination routers.
 - This means that one could create multiple parallel LSPs between a single pair of routers
 - This is useful if the traffic between two routers is such that the LSP sizes are too big to fit into a single link, or too big for that LSP and other LSPs to fit together on a single link.
 - There is no restriction on how many diverse paths those parallel LSPs can take any between the two points.
 - Different fill strategies can be used to optimize parallel paths between devices (least fill, most fill, random)
 - **ルータ間に設定できるLSP数には制限が無い**
 - 1ルータペアに複数のLSPを設定可能
 - 1回線にトラフィックが乗り切らなかった場合は、複数のLSPを設定することで、複数回線にトラフィックを分散させることができる
 - **LSPがどの回線を通るかは、様々な設定が可能で柔軟性が高い**
 - よりきめ細やかなトラフィック分散が可能
- least fill:** 使用帯域のより少ないパスに優先的にLSPを乗せる
most fill: 使用帯域のより多いパスに優先的にLSPを乗せる

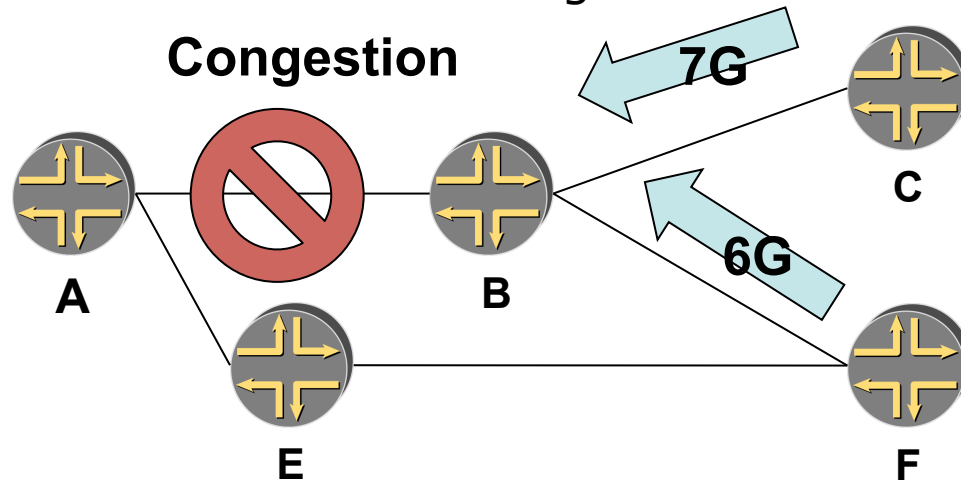
3.6 なぜ、グローバルIPネットワークにMPLS-TE?

- Many of the shortcomings of IP routing compared to MPLS/TE have since the early 2000s been addressed, e.g.
 - “wide” metric for ISIS/OSPF that allows for computation of link metric that includes other information such as load
 - Sophisticated metric calculation programs that allows for traffic adjustments in face of multiple diverse paths between nodes
 - IP forwarding performance has caught up to MPLS forwarding
- However, to date there is no equivalent way to obtain **detailed traffic statistics** that MPLS full meshed LSPs provides in pure IP routing.
- Should the network requirement or the available technology change, GIN network will change along with it and in the future GIN may not be using MPLS/TE.
- pure IPルーティングには、2000年代始めから指摘されてきたように、様々な欠点があった（ただし、その多くは現在すでに改善されている）
 - 帯域情報などを載せられる、拡張IGPが無かった
 - 効率的なトラフィック分散を実現するため、IGPメトリックの計算プログラムが必要になる場合があった
 - IPフォワーディングのパフォーマンスは、MPLSフォワーディングに比べて劣っていた
- トラフィックフロー分析の観点に立つと、今のところMPLS-TE(LSP)が最適
- ネットワーク要件や技術の変化に応じて、今後MPLS-TEを使わなくなるかもしれない



3.7 なぜ、グローバルIPネットワークにMPLS-TE?

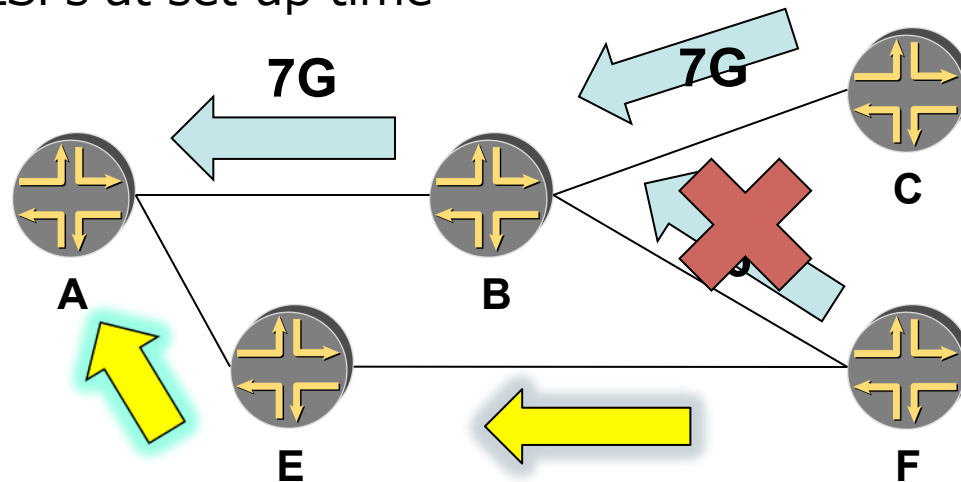
- In a pure IP network without MPLS, there was no good way to do admission control into traffic routing.



- ▶ Assuming that all links are 10G above, if C->B->A and F->B->A are the best paths for routers C and F, there will be congestion between A and B.
- ▶ In traditional IP routing, there's no concept of load, just simply a number that represented the "cost" of that link.
- ピュアなIPネットワークでは、IGPコストに従ってトラフィックが流れる。
- 流れる先の回線が輻輳する可能性がある。
- アドミッションコントロール（流れる先の回線帯域を把握して、流してよいトラフィック量を判断する）の概念は無い。

3.8 なぜ、グローバルIPネットワークにMPLS-TE?

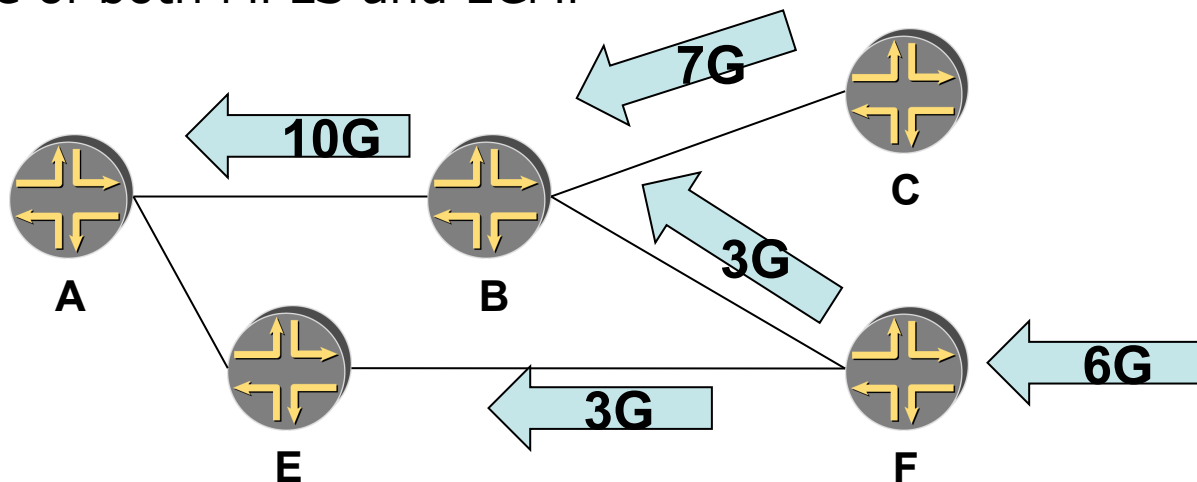
- With MPLS TE, if traffic information is known in advance that can be used to do admission control by assigning bandwidth to the individual LSPs at set up time



- ▶ Since $7G+5G > 10G$ of A-B link, either the LSP between C->A or F->A will fail to establish.
- ▶ The LSP that fails to establish will be routed around the “longer” path, which in above will most likely be the F->A LSP via F->E->A.
- **MPLS-TEでは、アドミッションコントロールが可能。**
- **トラフィック量を予め把握しておき、輻輳しそうな通り道を回避する**
 - **帯域が溢れそうな通り道には、そもそもLSPを張ることができない**

3.9 MPLS plus ECMP

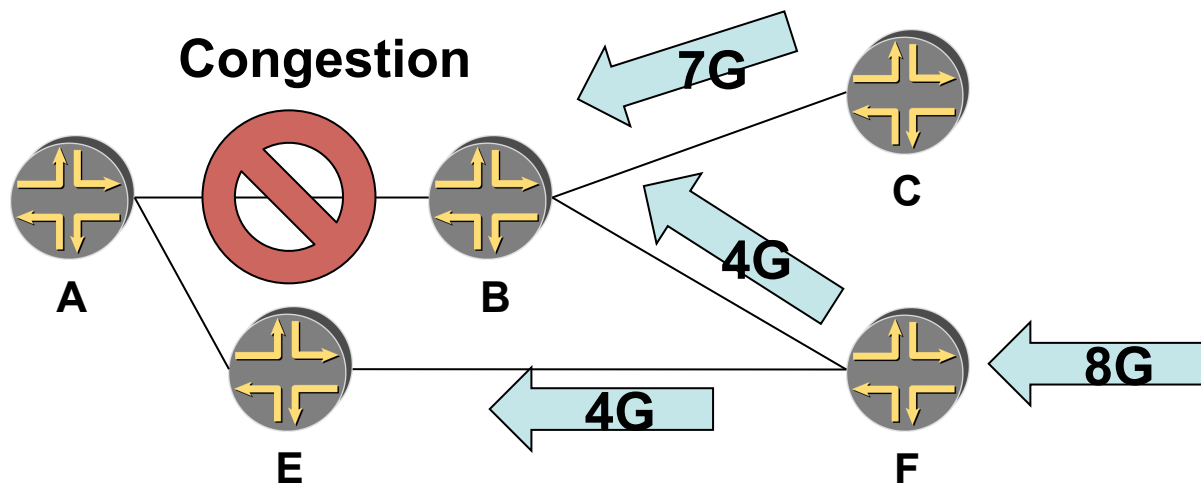
- If instead we created 2 LSPs between F and A we could take advantage of both MPLS and ECMP



- ▶ High value traffic could also be assigned it's own LSP between F and A and given priority access to the link between A and B.
- **2LSP(ここではF-B-A, F-E-A)を張れば、MPLS-TEでもECMPが利用できる**
- **F->A宛トラフィックを、2つの通り道に分散可能**

3.10 MPLS vs. ECMP

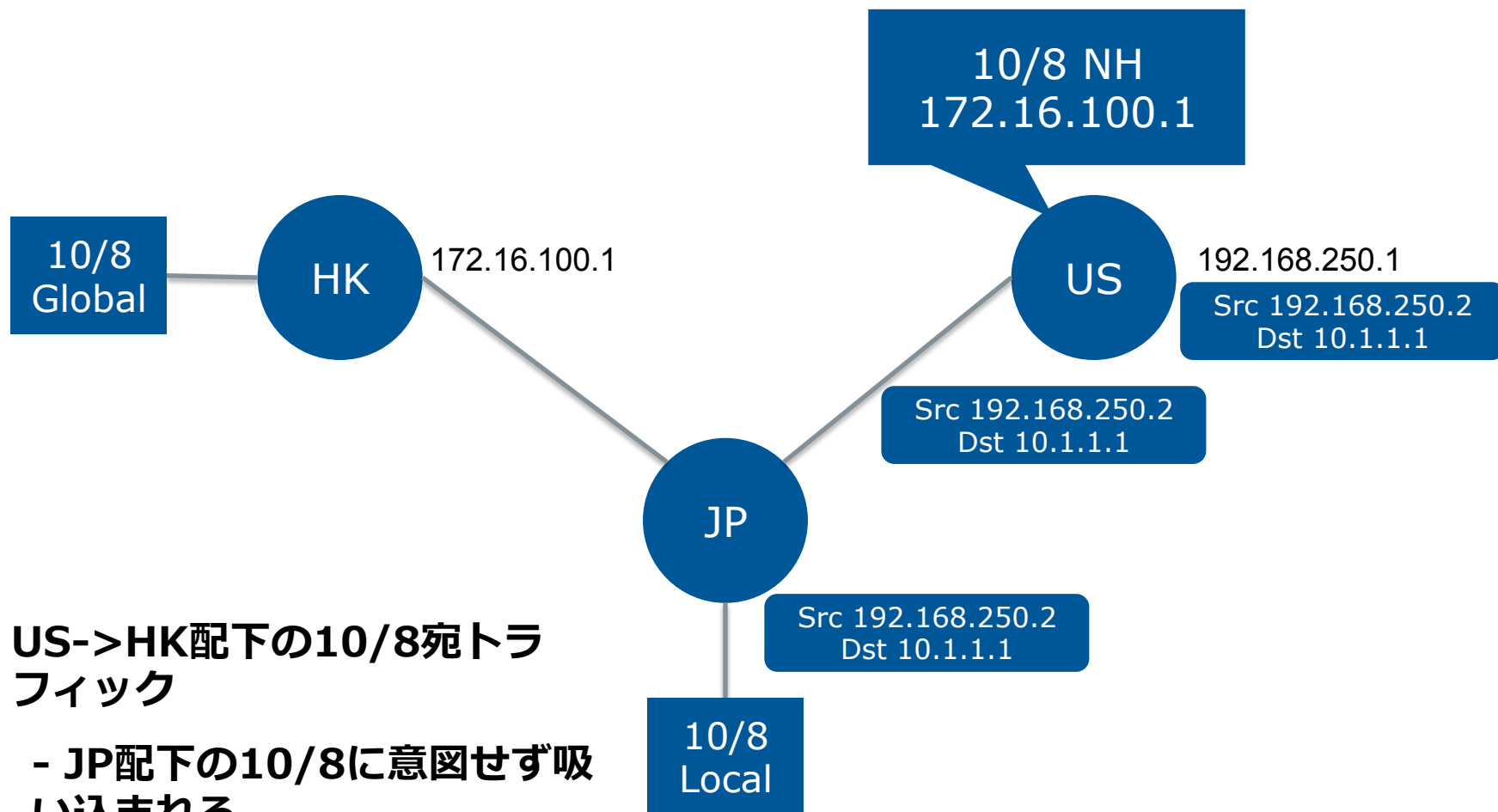
- With ECMP, traffic is balanced without regard to downstream capacity



- ▶ Since $7G + 4G > 10G$ of A-B link, congestion will occur on A-B
- ▶ The topology could be changed to allow for more efficient use by ECMP.
 - ▶ This may not be possible when long-haul or under-sea cables are involved
 - ▶ MPLS can solve this problem without a topology change
- **MPLS-TE無しのECMPでは、アドミッションコントロールができない**
- **せっかくトラフィックを分散しても、どこかで輻輳する可能性がある。**

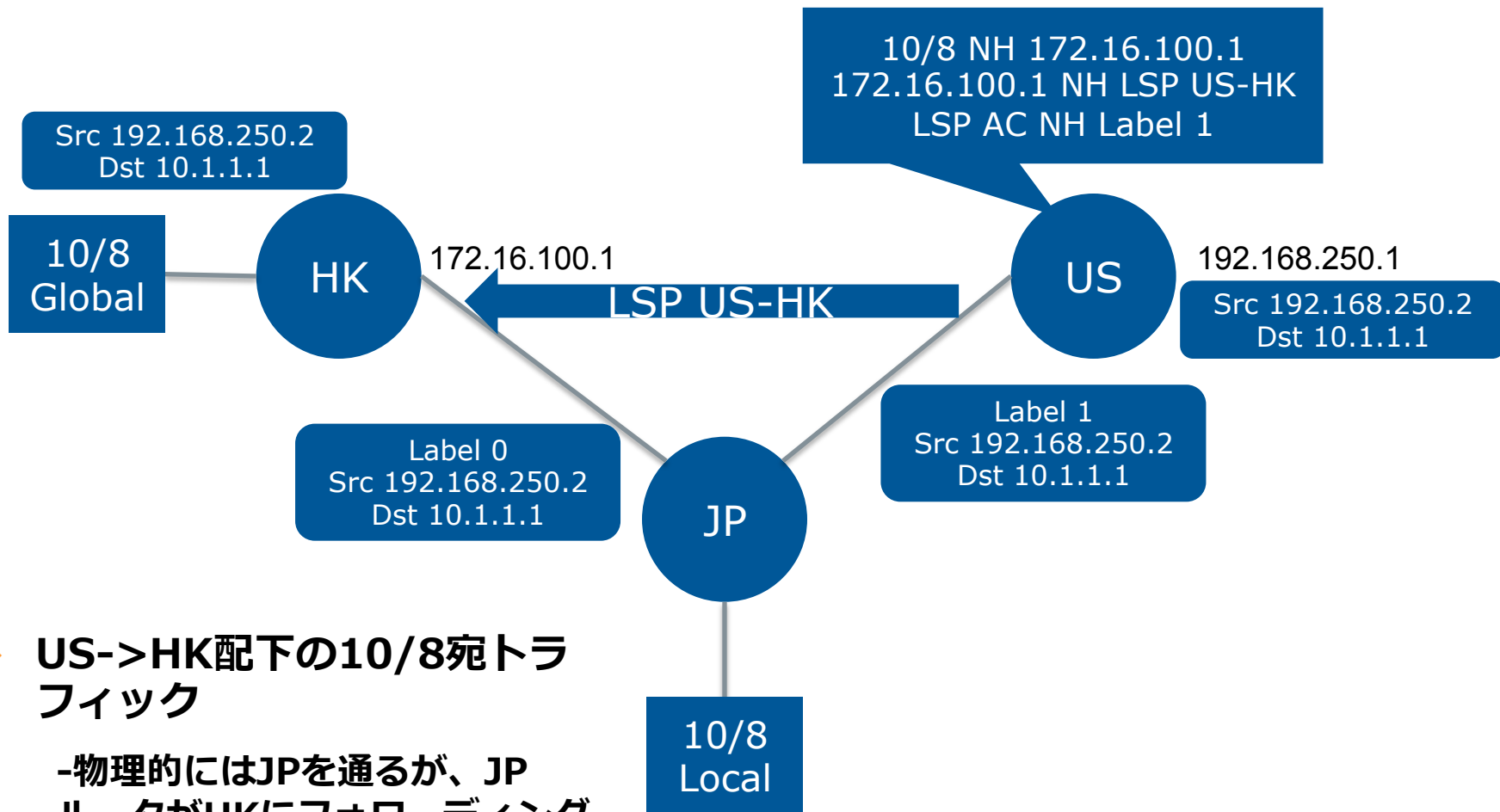
You're running an MPLS network, not an IP network

3.10 IP Forwarding



- ▶ **US->HK配下の10/8宛トラフィック**
 - JP配下の10/8に意図せず吸い込まれる

3.11 MPLS Forwarding

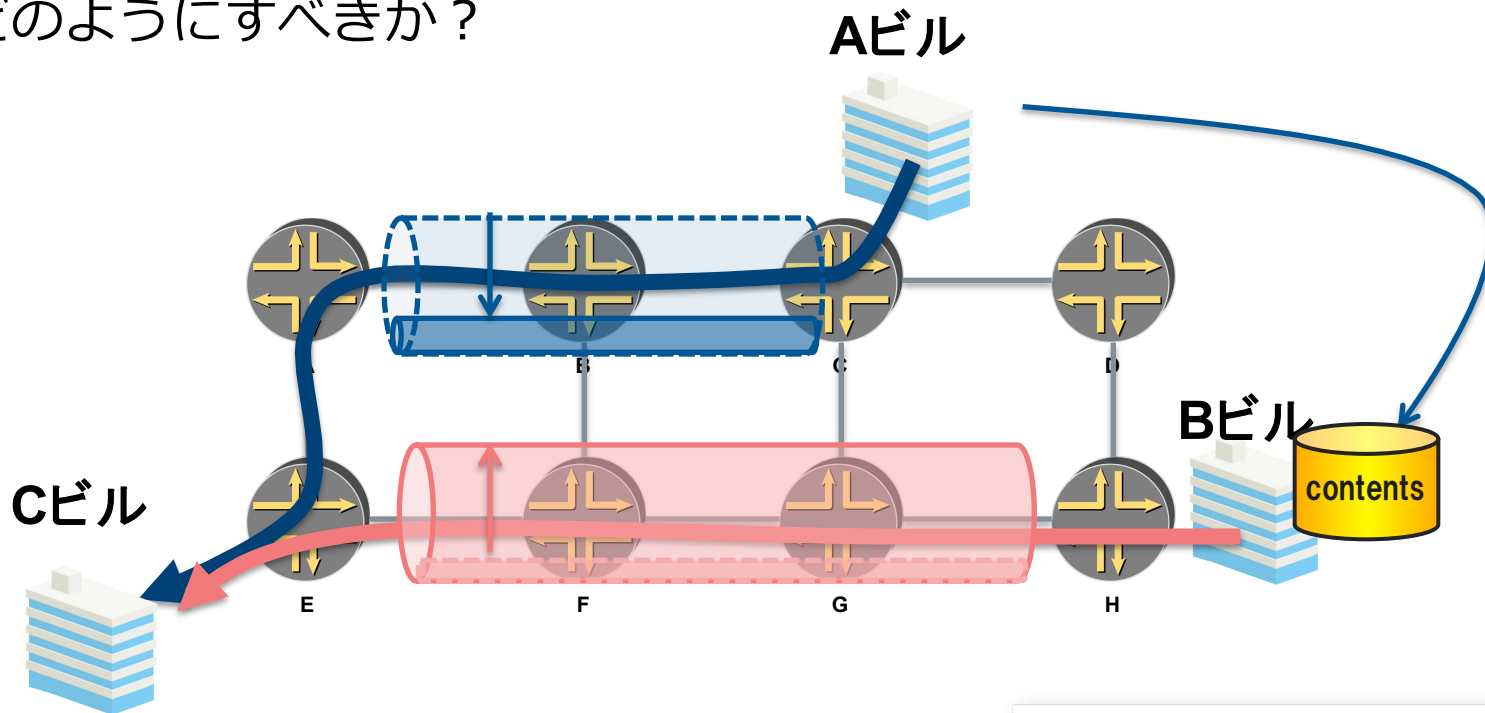


- ▶ **US->HK配下の10/8宛トラフィック**
 - 物理的にはJPを通るが、JPルータがHKにフォワーディングする

4. 昨今のトラフィック状況サマリー

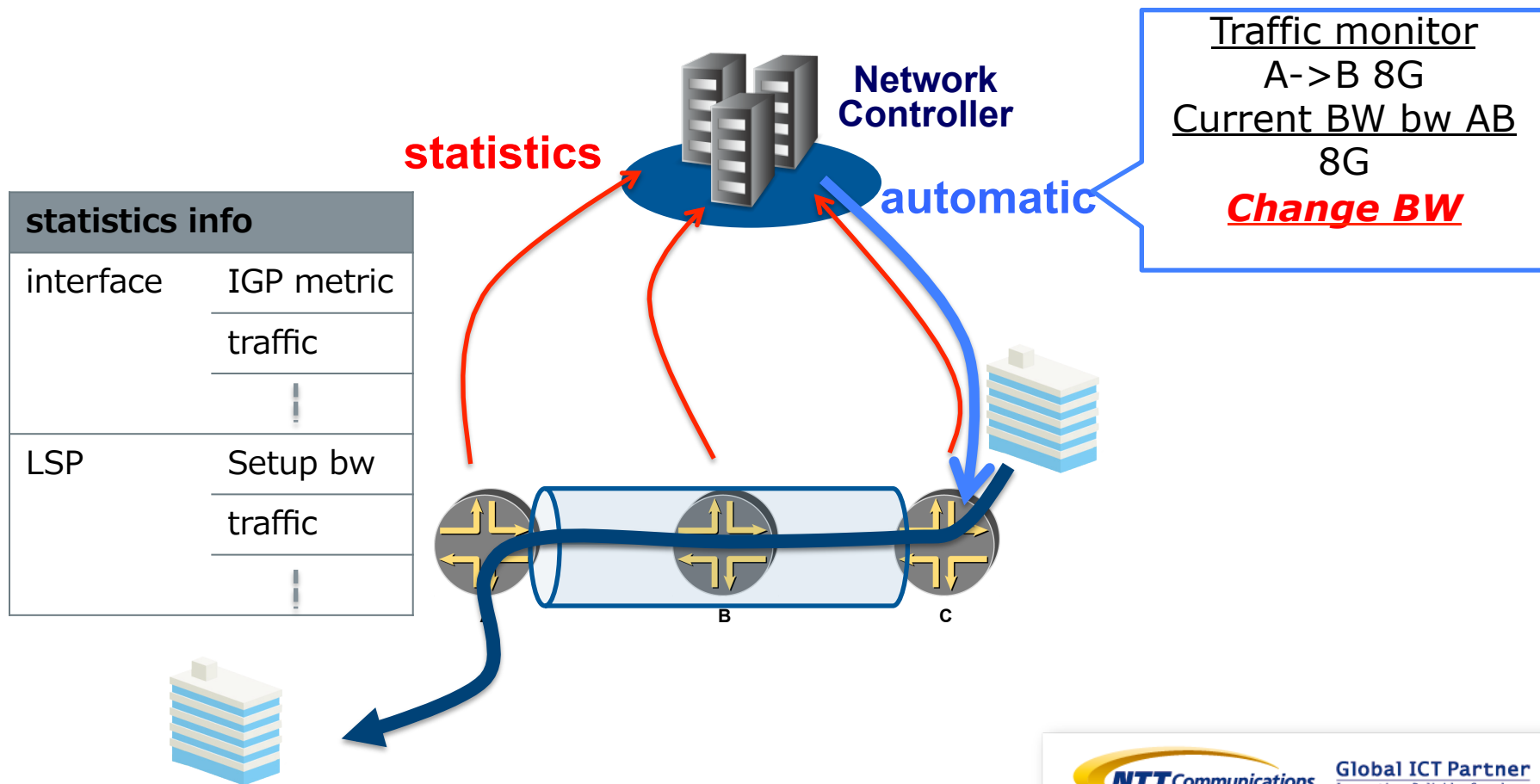
5.1.1 Auto bandwidth

- ◆ トラフィック交流が予想できない昨今、急にA地点からC地点へのトラフィックが減って、B地点からC地点へのトラフィックが増えるなんてザラに起きる。
- ◆ その変動にあわせて土管の太さをかえていかないと、パケロスするし、リソースの無駄使いにもなる。
- ◆ どのようにすべきか？



5.1.2 Auto bandwidth

- ◆ 自動化のやり方：
 - ◆ 集中制御：最近流行っぽくやってみる。



5.1.3 Auto bandwidth

- ◆ 自動化のやり方：
 - ◆ 分散制御：ノードが自立的にやってみる。
 - ◆ 例えば某社ルータだと;

```
[edit protocols mpls statistics]  
auto-bandwidth;
```

```
[edit protocols mpls label-switched-path label-switched-path-name]  
auto-bandwidth
```

```
    adjust-interval (in seconds); (default 24 hours)  
    adjust-threshold (percent); (default 5)  
    adjust-threshold-overflow-limit (number); (if the max avg bw is  
exceeded this many times adjust immediately)  
    minimum-bandwidth (bps);  
    maximum-bandwidth (bps);  
    monitor-bandwidth; (calculate changes but don't apply)
```

5.1.4 Auto bandwidth

◆ 某社ルータだと;

mpls traffic-eng
auto-bw collect frequency (minutes) (how frequently output information is collected)

interface tunnel-te tunnel-id

auto-bw

application (in minutes) (time between bw adjustments)

bw-limit min (kbps) max (kbps)

adjustment-threshold (percentage) min (minimum-bandwidth)

overflow-threshold (percentage) min (bandwidth kps) limit (if the max avg bw is exceeded this many times adjust immediately)

5.2.1 Re-optimization

- ◆ トラフィック量の変動が激しく、トポロジがドラスティックに変わる（例えば、新規ケーブル調達（数十Gベース）やケーブル障害）。その都度、フルメッシュのパスルートを人手を介して考えていたら、人件費がばかにならないし、障害時は、復旧の長期化によりトラフィックを守る事ができない。
- ◆ どのようにすべきか？

5.2.2 Re-optimization

- ◆ 自動化すべき。ルータによる分散処理を使う場合、某社ルータだとこんな感じで設定いれるだけ;

```
[edit protocols mpls]
```

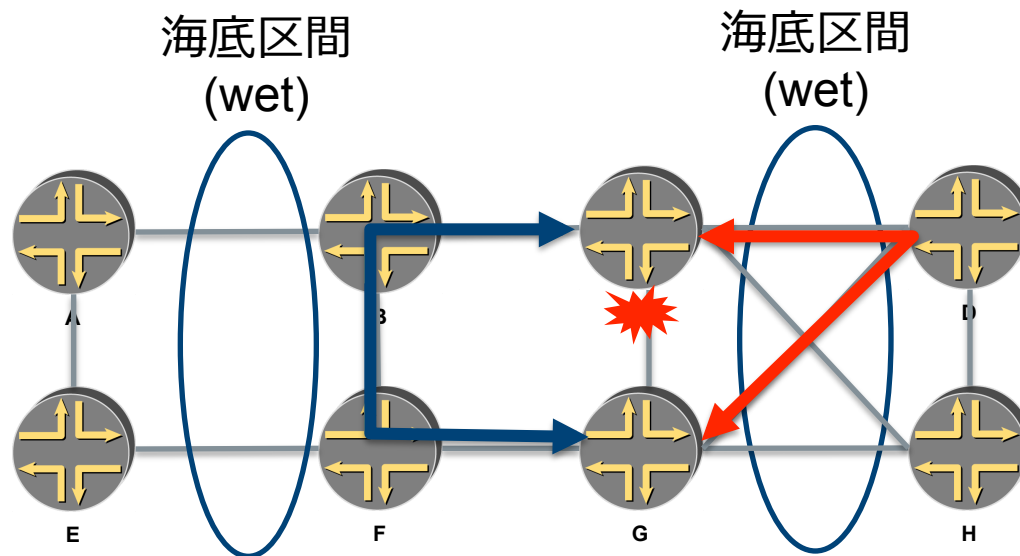
```
optimize-aggressive;
```

```
optimize-timer (second); (default 1800sec)
```

- ◆ 東日本大震災時、エンジニアは回線復旧のみを考えれば、あとはノードやシステムが勝手にベストルートにのせてくれた。

5.3.1 MPLS Link Coloring

- ◆ Intra-Regionのパスが太平洋/大西洋の海を行って来いするようなことは障害時であっても避けたい。
- ◆ グローバルにおいて、海底ケーブルコストがコストに閉める割合が非常に高い。
- ◆ どのようにすべきか？

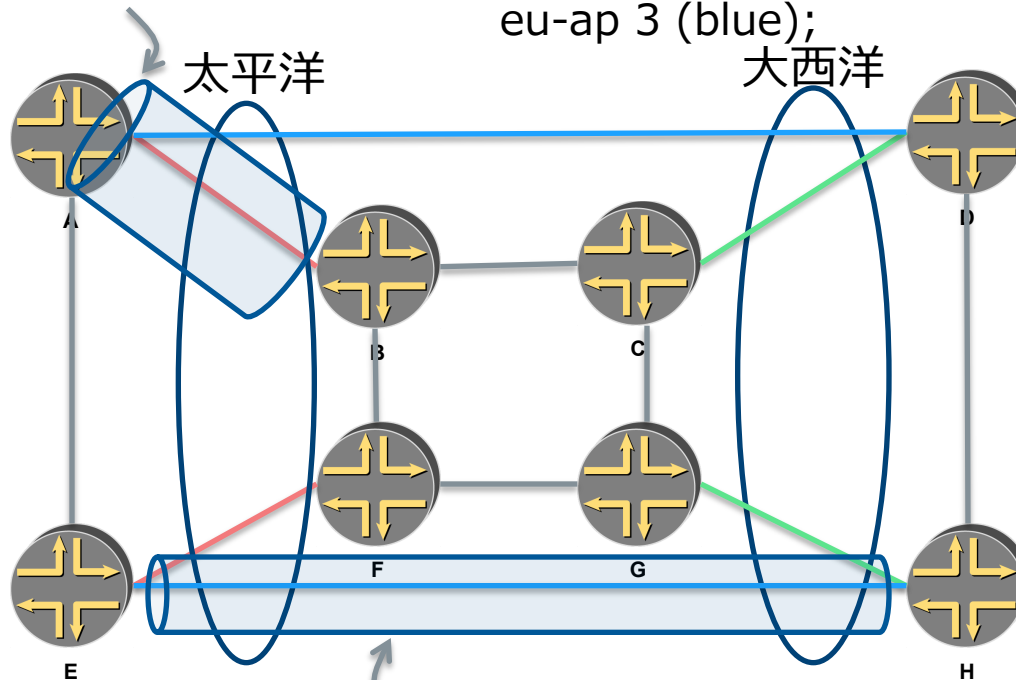


5.3.2 MPLS Link Coloring

- ◆ リンクに色をつけて、Intra-regionを意識した経路計算を自動化する。

全ルータに色づけの定義を実施
[edit protocols mpls admin-groups]
na-ap 1 (red);
na-eu 2 (green);
eu-ap 3 (blue);

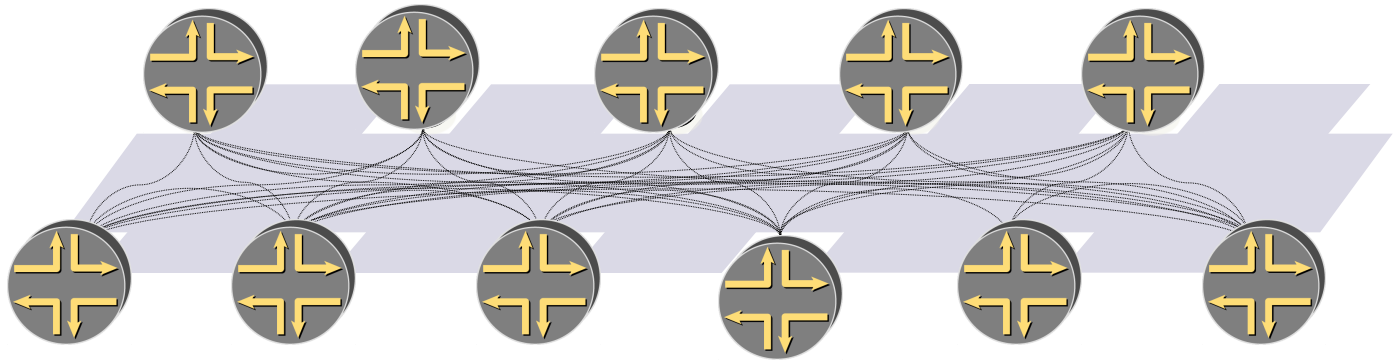
Path E-H
(exclude [eu-ap na-eu])



Path E-H
(exclude [na-ap na-eu])


5.4.1 Priority for Large LSP

- ◆ フルメッシュでノード間を結ぶと数千本/serviceの論理パスが必要。ネットワークに存在するパスは数万本になる。
- ◆ この論理パスをベストに配置していくことが、物理リソースの有効利用強いてはコスト削減に繋がる。
- ◆ どのようにすべきか？



5.4.2 Priority for Large LSP

- ◆ 論理パスを配置する方法として、
 - ◆ 大きなトラフィックさばくためのパス（setup bwが大きいパス）からルートを決めていく。
 - ◆ 障害時も、大きなトラフィックを運ぶパスを最適化できると、パケロスが最小化できる。

Priority	setup	hold	BW値（例）
high	0	0	-
	1	1	30G to 40G
	2	2	20G to 30G
	3	3	10G to 20G
	4	4	5G to 10G
	5	5	1G to 5G
	6	6	1G
low	7	7	-

5.5.1 Spread LSPs for resource allocation

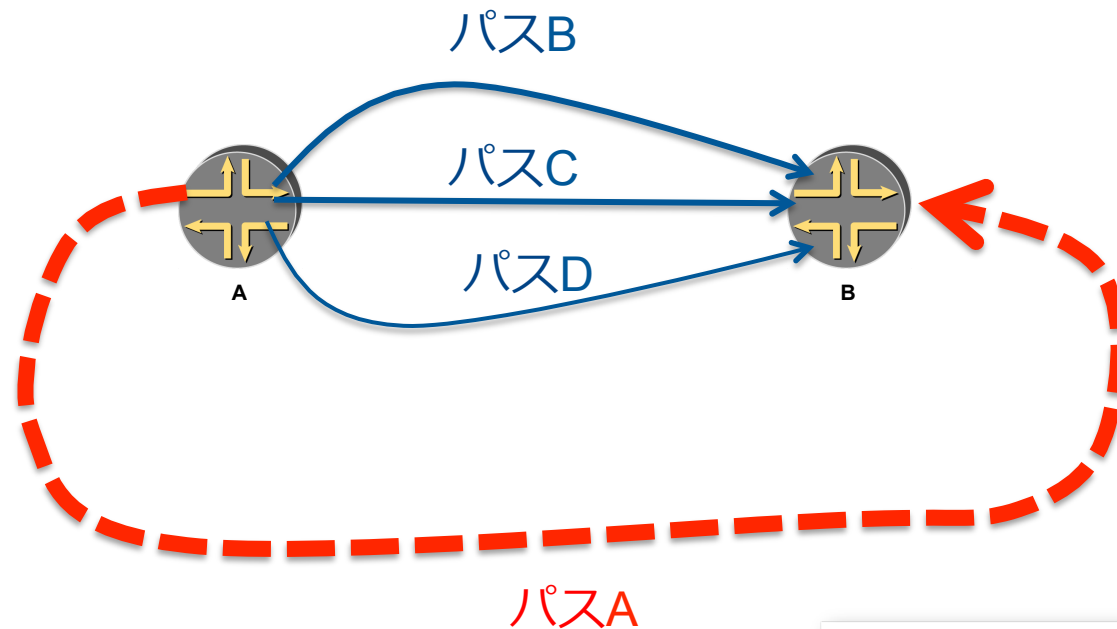
- ◆ ルータ間トラフィックは1年で約10G以上増加している。
- ◆ 全自動でパス帯域を増やしていくと最適化できない場合も存在してくる。
- ◆ どのようにすべきか？

	Nam	Sour	Dest	Setup BW
1	r...	r...	r...	23283.30
2	r...	r...	r...	15685.10
3	r...	r...	r...	15450.10
4	r...	r...	r...	15382.10
5	r...	r...	r...	15318.70

	Nam	Sour	Dest	Setup BW
1	r...	r...	r...	28102.60
2	r...	r...	r...	27900.60
3	r...	r...	r...	27821.60
4	t...	r...	r...	26516.30
5	r...	r...	r...	26509.10

5.5.2 Spread LSPs for resource allocation

- ◆ パス帯域を増やすのではなく、複数にセパレートしてしまう。
- ◆ パスAを分割し、パスB+パスC+パスDにする。
- ◆ Latencyが悪いネットワークにならないように、ルートが異なっても可能な限りのshortestをキープする。



Questions?

