Janog52@Nagasaki

2023/07/06

5GC構築におけるClos IP Fabric活用

ソフトバンク株式会社
インターネットサービス部
Closデータネットワーク課

ヤダブ　リチュラジ

SoftBank

本日の内容

SoftBank

1. キャリアバックボーンとClos Fabric
   - 5GSAのアーキテクチャをIP Clos Fabricで表現してみよう＋

➡ 2. Clos Fabric
   - Fabricを作る際に考慮することをディスカッション

3. CGN/Firewall for 5GC Fabric
   - 5GC FabricのSecurityを支える技術

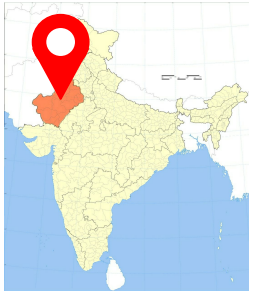
自己紹介

SoftBank

名前： ヤダブ　リチュラジ
Yadav　Rituraj

苗字ではないが、
リチュラジと呼んで
ください。

出身： インド　ラジャスタン州



仕事内容：
- インターネット中継網(Sgi/N6)の設計・構築
- 5GC 各NF間通信用のIP Fabric NW設計・構築
- インターネットNF（CGN/FW/DDoS等）の設計・構築

JANOG歴：
- 現地参加3回目
- 初登壇

Clos IP Fabricとは

SoftBank

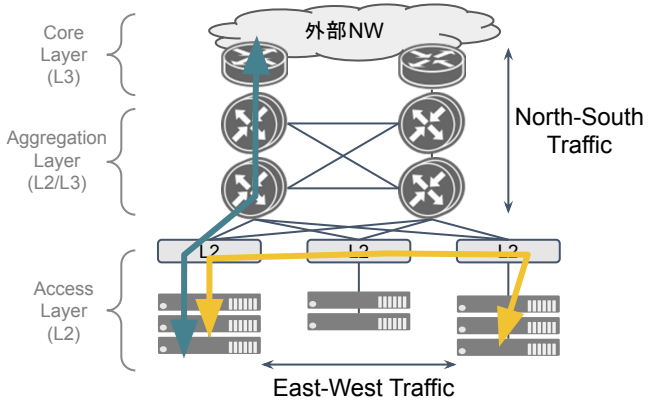
命名の由来

Clos IP Fabric

- 従来データセンターのL2 NWから進化した**IP(L3)Fabric**のこと
- 1952年 マルチステージネットワークトポロジーを考案した**Charles Clos**氏

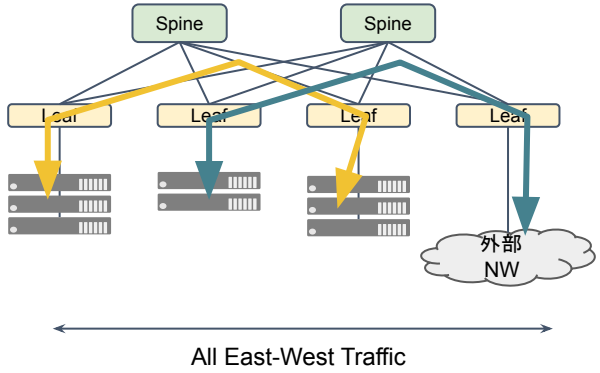
従来のDCネットワークトポロジー

■L2SW、L3SW、コアルータの3層型トポロジー
■DC内サーバ間通信がEast-West、ユーザ⇔サーバ間通信がNorth-South
■スケーラビリティに限界あり
■L2ループ対策が必要
■L2/L3共存のため設計の複雑化、トラブルシューティングが困難



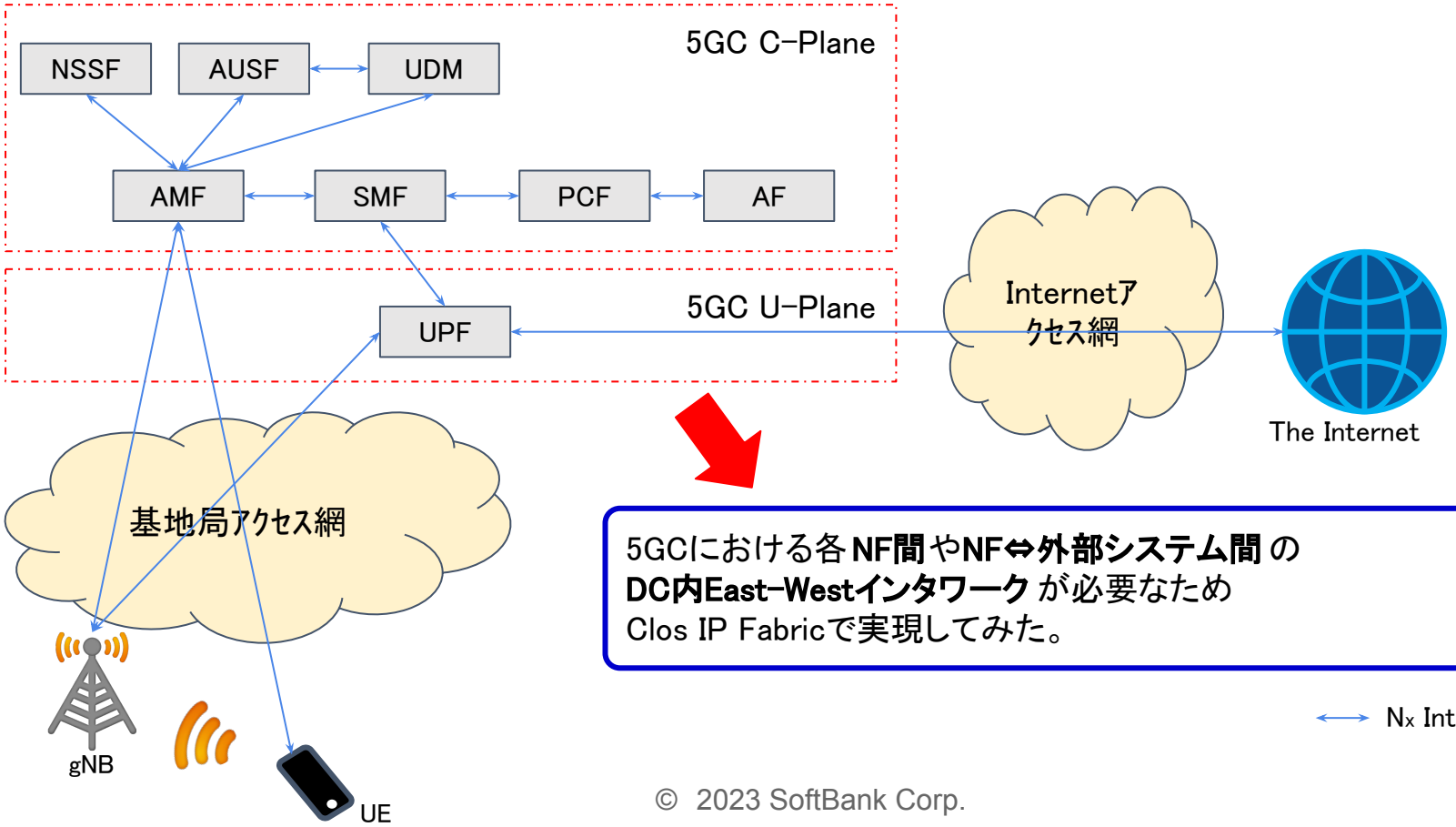
Clos IP Fabricポロジー

■Spine-Leafメッシュ型Topology（2層、3層タイプあり）
■Underlay IP Networkで構築のためL2ループ対策不要
■DC内すべての通信がEast-West
■スケールアウトが容易
■通信フローの設計が統一化可能なため、設計もメンテナンスも楽



5GCアーキテクチャにおけるClos IP Fabricの必要性

SoftBank



5GCにおける各**NF間**や**NF⇔外部システム間**の
DC内East-Westインタワークが必要なため
Clos IP Fabricで実現してみた。

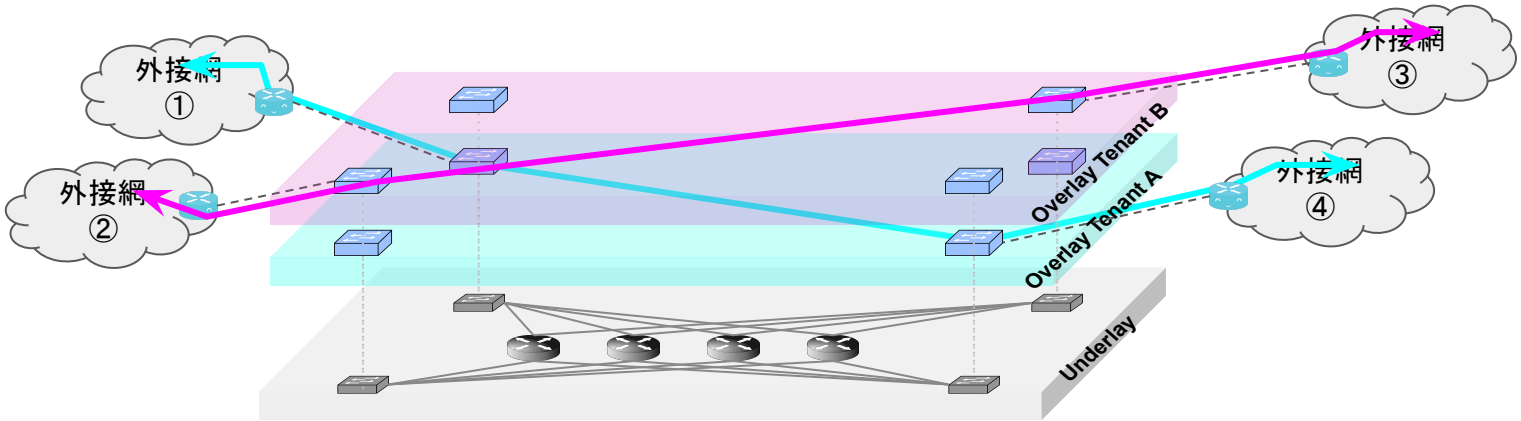
⟷ Nx Interface

【構成検討①】

・Fabricに収容される外部NWとは様々な通信要件があるため、**マルチテナント収容** 前提で設計

・マルチテナントを収容するため、Fabric内の通信レイアを2つで考える
　　■**Underlay[infra]**：各Spine-Leaf間の到達性を持たすレイア
　　■**Overlay[service]**　　：各外接NW/NF/ｻｰﾊﾞ間の通信を可能にするレイア

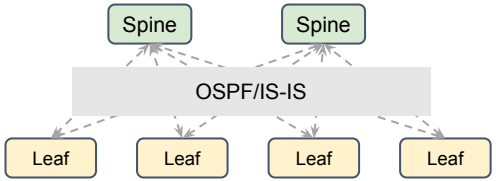
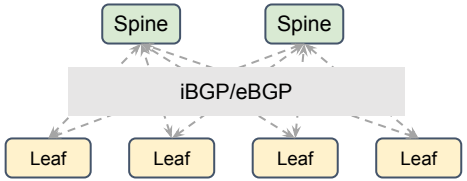
・各外接間の不要な通信を起こさないため、Overlayの面分けを考える　→　仮想的Routing等



SoftBank

【構成検討②】

Underlayプロトコル検討　　：各Spine/Leaf間の**リンクやLoopbackアドレス**を互いに知らせる。

元々ﾊﾞｯｸﾎﾞｰﾝ設計担当者として、2 **パターン比較→**		
Fabric内情報交換方式	Link State情報交換し、LSDB構築	NLRI交換し、ﾙｰﾃｨﾝｸﾞﾃｰﾌﾞﾙ構築
NW状態把握の範囲	全NWトポロジー	隣接ノードのみ
ECMPやTraffic制御方法	リンクコストベース、ECMP可	BGP属性ベース、ECMP可
スケーラビリティ	Leaf数増→LSDB増＝機器負荷増	交換する経路数が数百～数千程度のため大きな影響無し
障害時の影響	NW全体で再計算→断時間長め **LFA-FRR**によるバックアップパス有効化検討	BGP Multipathにより高速収束 さらに**BGP-PIC**利用しFIBへバックアップパス投入

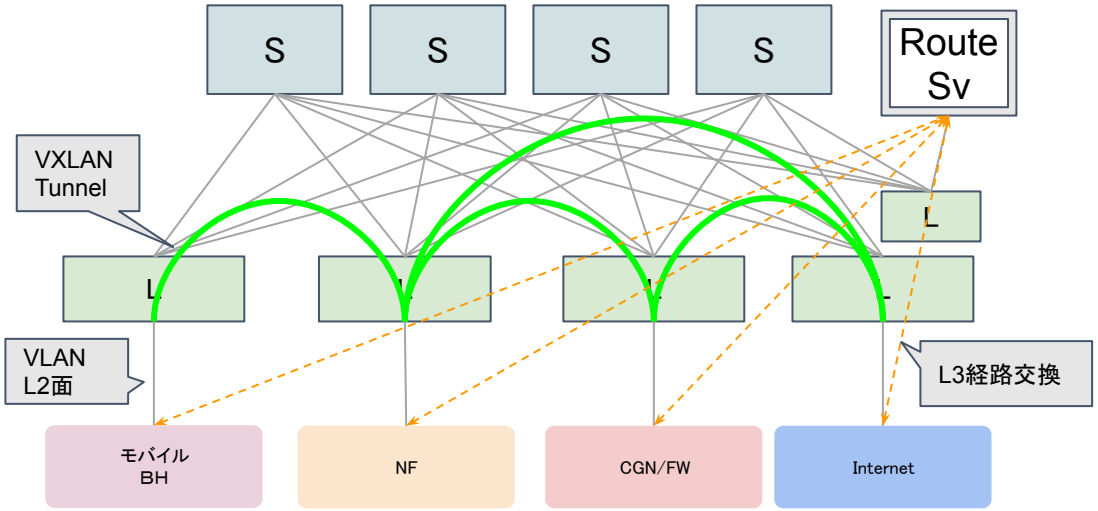
マルチパスが有効であれば障害時は残パスでサービス継続可能な、シンプル（使い慣れた）なBGP実装でも良いのでは？

会場/視聴者の皆様で Underlayプロトコル選定で悩んだことある方いますか？

【構成検討③-1】

Overlayプロトコル検討 ：MultiTenancyを前提で検討

Approach①　L2 Overlay (VXLAN)



■Fabricでは外接ノードと経路交換しない

■Routeサーバ(RS)を外接先として立てて、
L3接続必要な各外接NWがRSとPeering

■各LeafでVLAN⇔VNIﾏｯﾋﾟﾝｸﾞしVXLANでTraffic転送

■考えられる**メリット**
- ✔ とてもシンプルなFabric実装
- ✔ 経路保持不要（CPU/メモリー節約）

■考えられる**デメリット**
- ？ PBRや経路リーク等がとても困難
- ？ サービスチェイニング対応困難

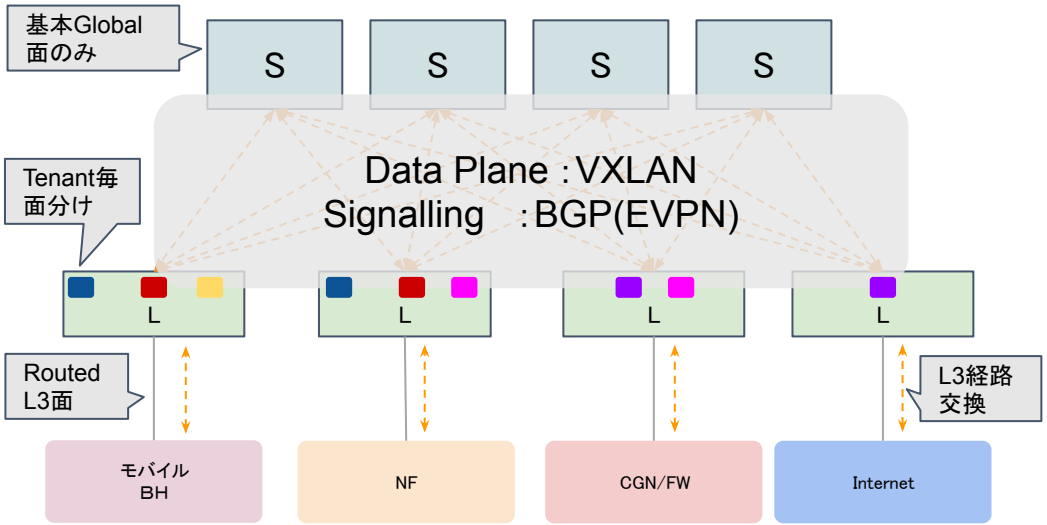
**ｲﾝﾀｰﾈｯﾄ世界のDC内は
このような設計が普通？**

SoftBank

【構成検討③-2】

Overlayプロトコル検討 ：MultiTenancyを前提で検討

Approach②　L3 Overlay (VXLAN+BGP/EVPN)



■Fabricでは外接ノードとL3で接続し経路交換

■各Tenant通信のためにL3面分け（VRF等）

■RR/RSはFabric網内で立ち上げ

■考えられるメリット

- ✔ PBRや経路リーク等複雑な通信制御に対応
- ✔ Fabric網内で経路管理、経路制御可能
- ✔ 各通信フローの理解、切り分けが容易

■考えられるデメリット

- ？ 高性能な機器が必要（コスト＋＋）
- ？ 各外部NWごとに外接設計が必要

EVPN/VXLANが主流？
他のプロトコルでL3 Fabric
作ってる方いますか？

EVPN/VXLANとは

SoftBank

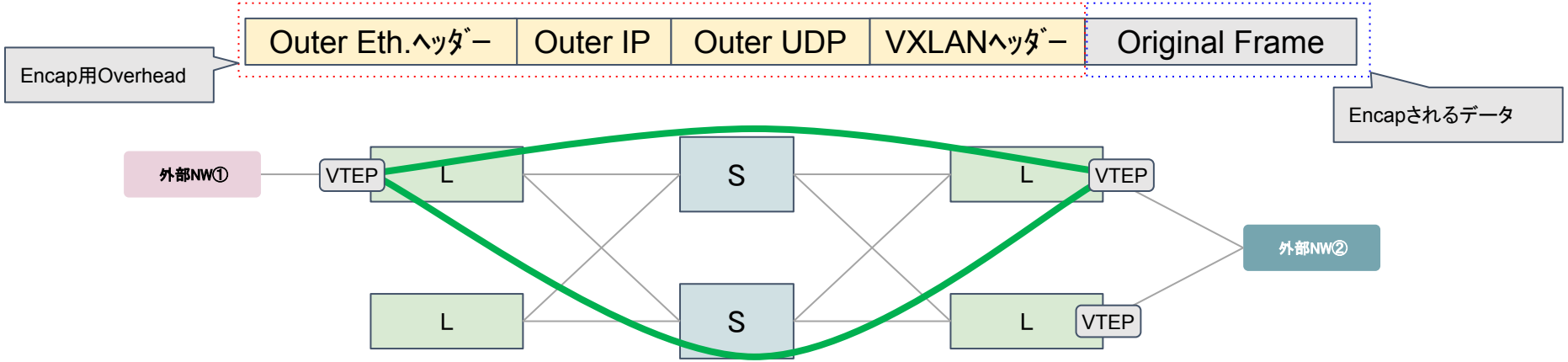
VXLAN

■RFC7348にて標準化されているプロトコル

■簡単にはVLANの数制限(1-4094)の拡張版といえるが、VLANでできない機能が追加されている
- ・L3 NWを経由し、分離されている L2 NW間の、トンネリングが可能
- ・ユニークなID数が24-bitのため1600万程度が使える
- ・トンネル作るのに必要なのは各 End-pointのみ（VTEP）

■Clos IP Fabricでは、L3であるUnderlayのうえで、Overlay(Data Plane)で複数L2/P2P面の実現が可能

■Overheadがかなり大きいが、どうでしょうか？**VXLANより良い事例あれば聞きたい**！



EVPN/VXLANとは

SoftBank

EVPN

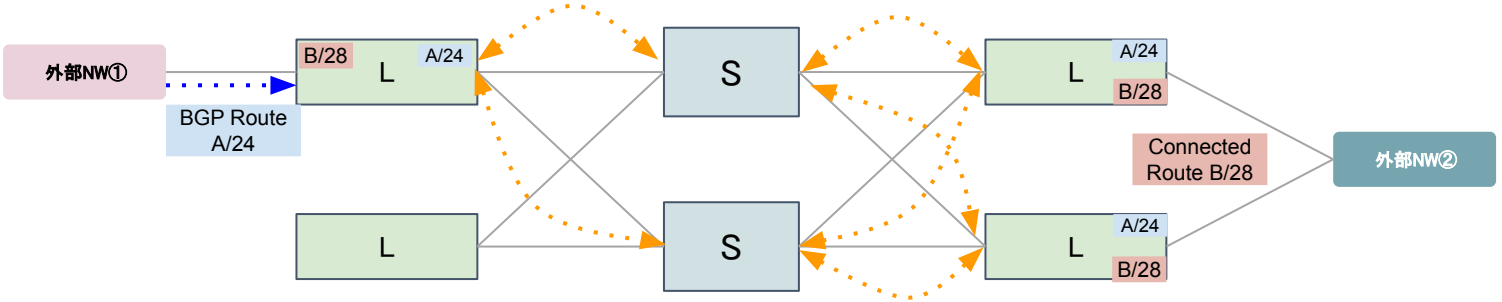
■BGPの拡張機能として、複数VPNの経路を交換するプロトコル

■L3経路(Prefix)だけでなく、L2経路(MAC情報)も交換可能なため、L3 FabricでL2面適用可能

■EVPNで交換可能なRoute-type

Type 1 ：Ethernet Auto Discovery
Type 2 ：MAC/IP Advertisement
Type 3 ：Multicast Route
Type 4 ：Ethernet Segment
Type 5 ：IP Prefix（v4/v6）

■EVPNでVXLANトンネル自動作成が可能なため、Overlay ControlとData Plane間のインタワークが良い



SoftBank

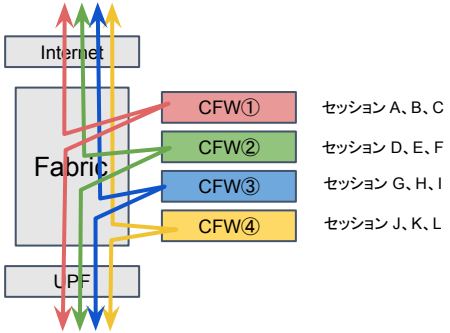
【サービスチェイニングについて】

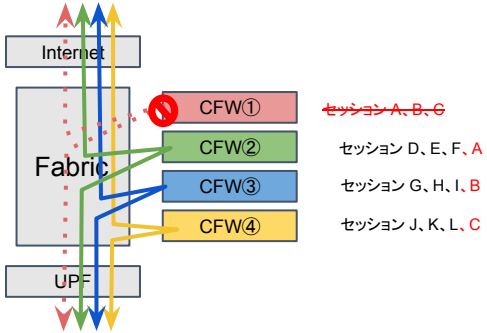
アプライアンスのロードバランス等を行う際に FabricでTraffic分散等が必要

⇒Traffic分散のHashはどのような方式が良いのか？

✔通信断が出るフロー以外のフローは影響無しにしたい

✔様々なアプライアンスを折り返しする通信フローは管理すべき？





■PBR等によりSrc IPでTraffic分散
■上下通信は同じCFWを経由する

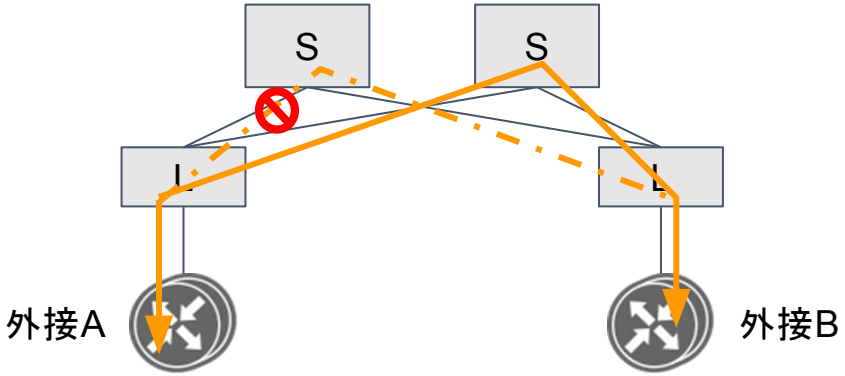
■フロー障害時は、他フローへ影響無し
■障害フローは他フロー再分散

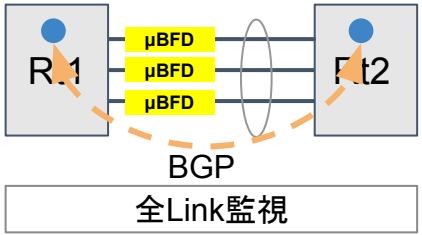
【耐障害性の考え】

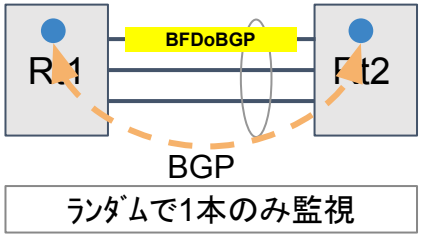
◇各通信フローにおいて **常にECMP**でトラフィックをロードバランス
◇単一や二重障害時は、障害を受けたパスだけ排除され、サービス通信は継続

サイレント障害検知手法

—LACP　　:L2レイヤーで高速検知（ポピュラー、実装も幅広い）
—BFD：L3レイヤーで高速検知（ポピュラー、ただ**μBFD**実装は少ない）







他に耐障害性の考察ありますか？

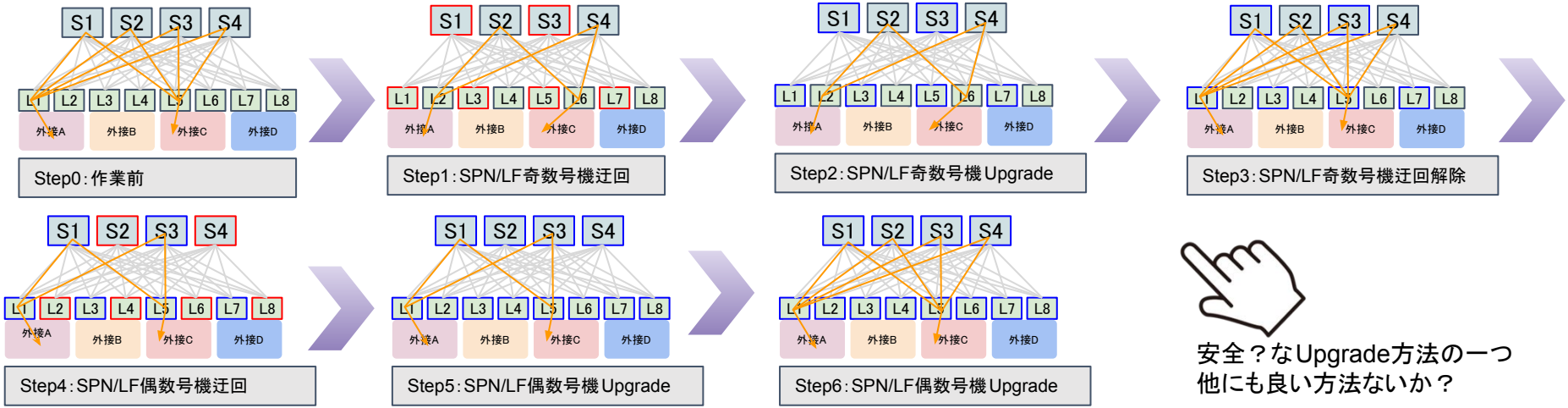
SoftBank

【メンテナンスやOSアップグレードの考え】

■通信断無し（or 許容範囲）でメンテ（迂回等）を実施する

- ✔ ヒューマンミスの削減
- ✔ 迂回作業工数削減

} 作業自動化（SDNコントローラ経由等）



安全？なUpgrade方法の一つ
他にも良い方法ないか？

SoftBank

【SDNコントローラによる構築自動化】

■Fabricという大規模なインフラを構築するのに多大な構築作業と期間が発生する。

■迅速にフィールド展開するための考察:
- ーHLD/LLDのDB化
- ーZTP有効化
- ーSDNコントローラよりインフラへの設定 (Config)の一斉配信

HLD/LLD
DB情報

パラメータ連携

SDNコントローラ

ZTP、プロビジョニング

最低限コントローラ機能
- ◆Config履歴、変化管理
- ◆トポロジー管理
- ◆マルチベンダー・マルチOS対応
- ◆他に？

課題と感じたこと

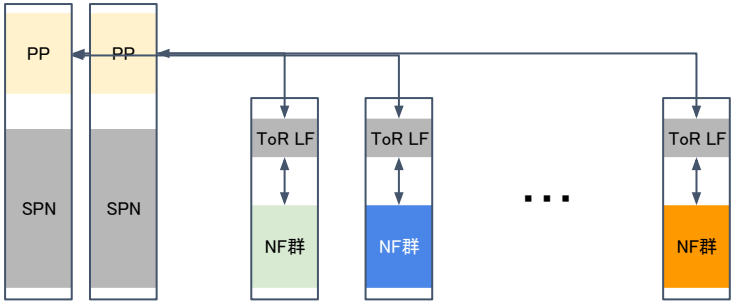
SoftBank

【Leaf＝ToRスイッチの場合の課題】

・C-PlaneのNFは通信容量が少ない　→Leafのポート有効活用が図りにくい
・NF基盤のﾗｯｸが増えばToR Leafも増える　→SPNのポート消費もあり、Fabric収容限界が到来する

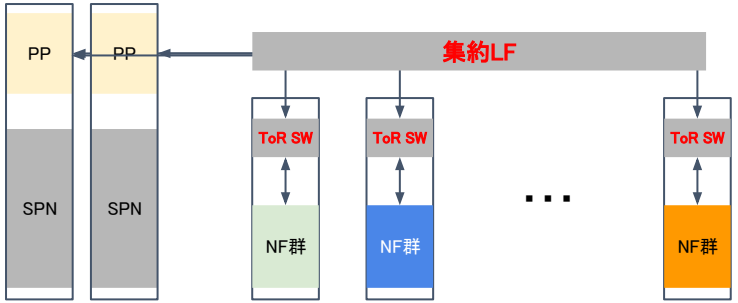
👉**ToRとして汎用L3SWに置き換え**　→Fabricポート節約、NF毎の通信要件異なってもFabric設計が共通

従来のDCアプローチ



・ﾗｯｸ毎の接続方式が違う
・帯域ほぼ余ってるのにLFが増え続ける
・Fab全体の管理台数が増え管理が大変
・経路やﾌﾛｰの切り分けが大変

こんなアプローチありなのか？



・Leaf外接設計が共通！
・Leaf台数も大幅に削減
・ToR SWはNF毎の好きな設計で良し
・Fabric全体管理し易さUP！

今後の展望

SoftBank

5GSAのさらなる進化のため Clos IP FabricHW/OS/デザインの向上を目指す
今後は以下の Topicで情報収集や情報共有活動をしていきます。

WBS/MultiVendor化

■WBSでのSpine/Leaf実装　:コンパクトな構成で早期立ち上げができるとうれしい
■マルチベンダーサポート　:機器HW/OSに依存せず一元管理

Fabric機能の進化

■Leaf SWにNFの機能を追加することによりサービスチェイニングの最適化
→UPF、NAT、FW、LB、DDoS防御

■NW Slicingのさらなる浸透とオンデマンド対応
→（例）セキュリティ対策を充実したスライス（異常や感染診断等）
→（例）低遅延スライス
→（例）IoT用スライス

将来JANOGにて上記の議論をさせてください。

Fin