“LINE流”ネットワークの作り方とそれを支える光通信技術

次世代ネットワーク基盤セミナー - 2019/03/28

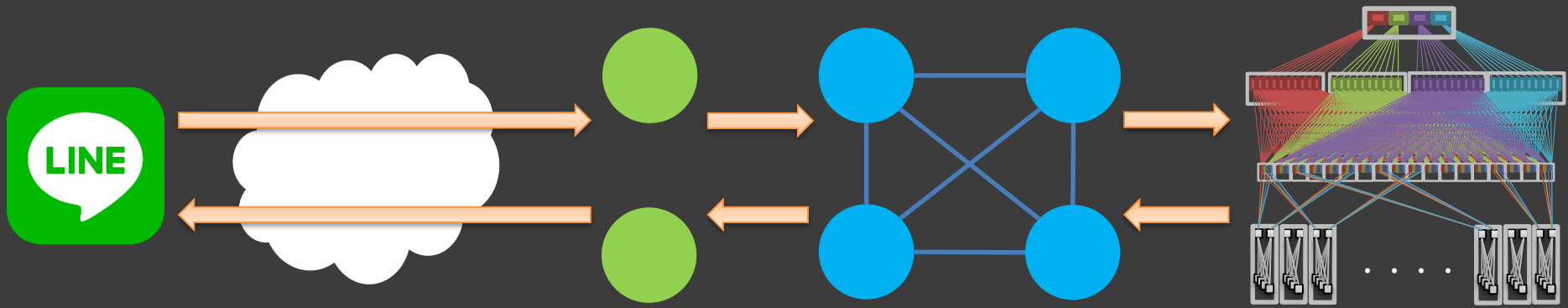
Masayuki Kobayashi
LINE Corporation

LINE

インフラのキャパシティと迅速なスケールが重要に

LINE Network Infrastructure

すべて自社で設計・構築・運用



User Internet Peering Edge DCI Data Center

164M+

Active Users in JP, TW, TH and ID

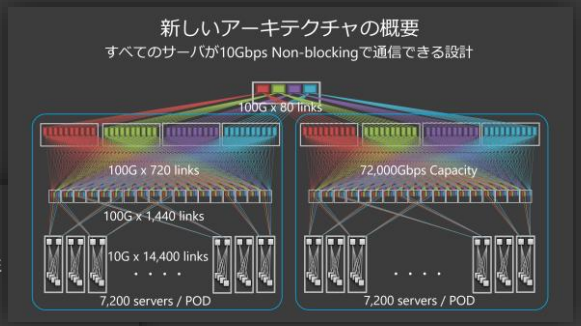
30,000+

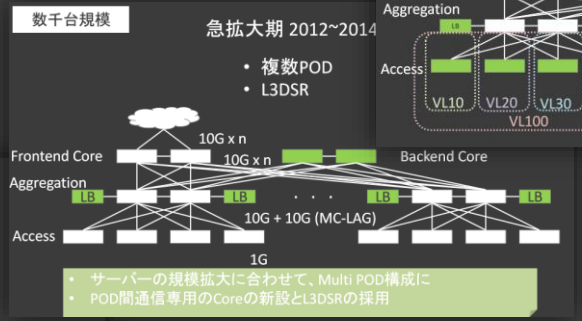
Physical Servers

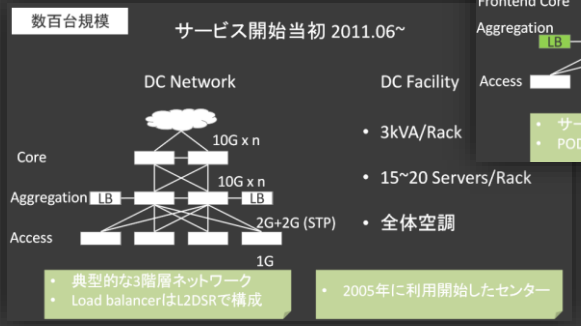
1Tbps+

User Traffic

ネットワークアーキテクチャの変遷







10G/40Gから100Gが主役に

データセンタネットワークへの要求

急増するトラフィックを効率的に捌くことが重要

広帯域

トラフィックは日々増加、特にシステム間の通信が膨大

高密度な100G opticsはデータセンタネットワークの構築に必須

スケーラビリティ

素早い事業展開を支える構成

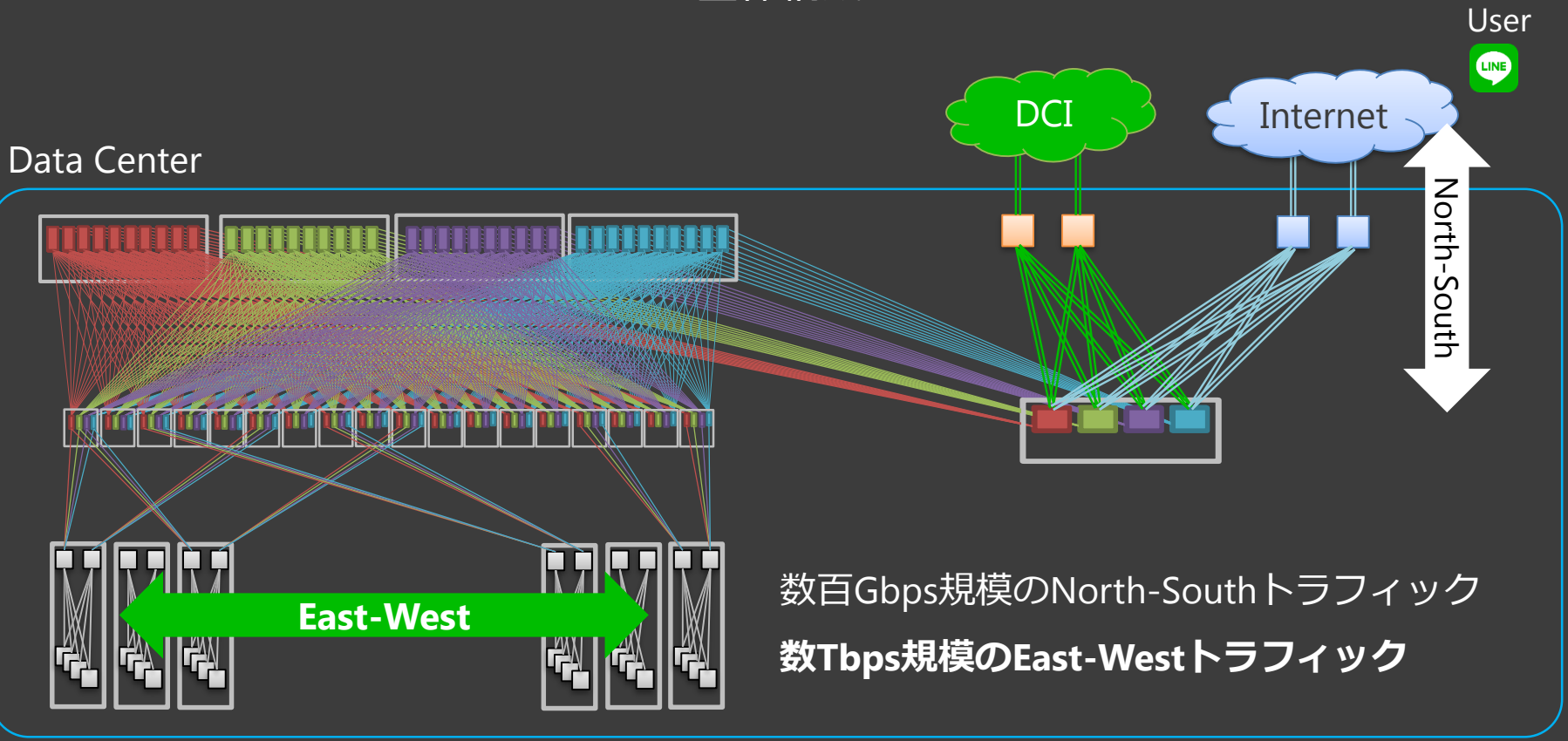
スケールするためにはシンプルで繰り返し可能であることが重要

安定運用

機器は必ずいつか故障する、壊れることを前提として設計

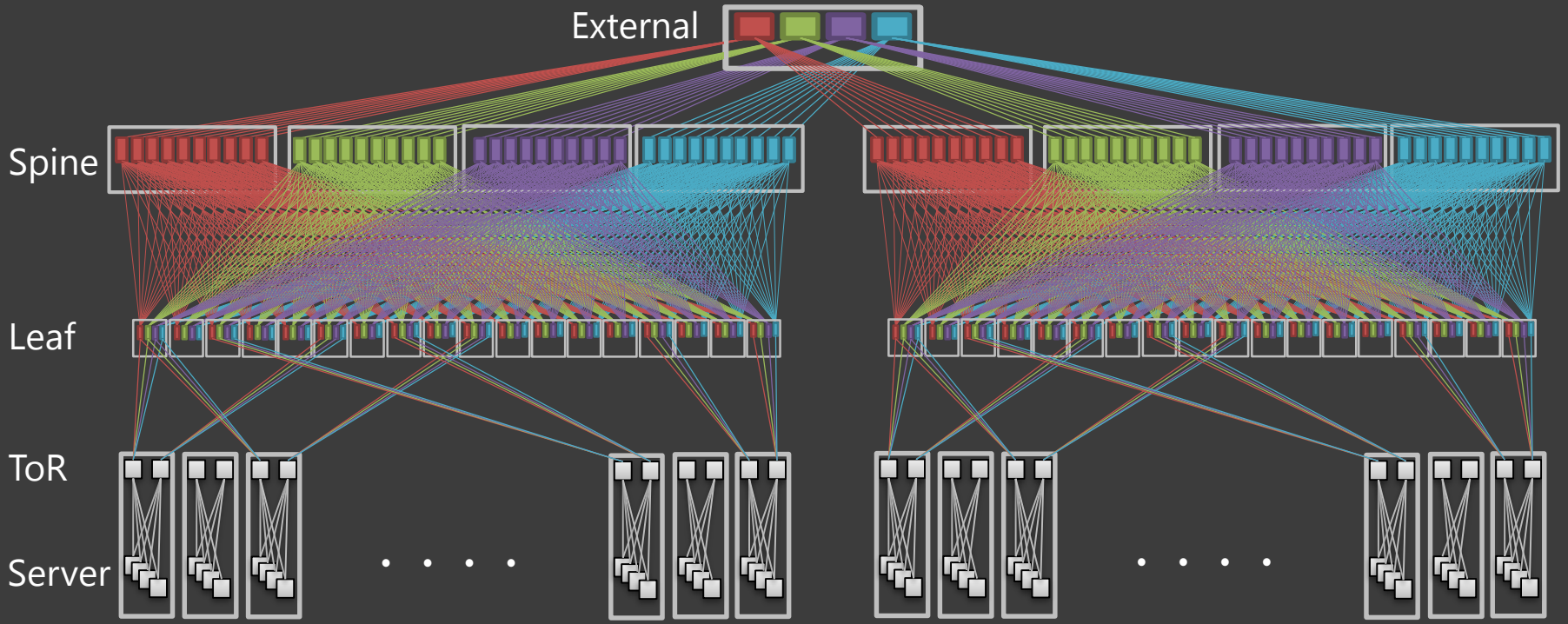
現在のデータセンタネットワーク

全体構成

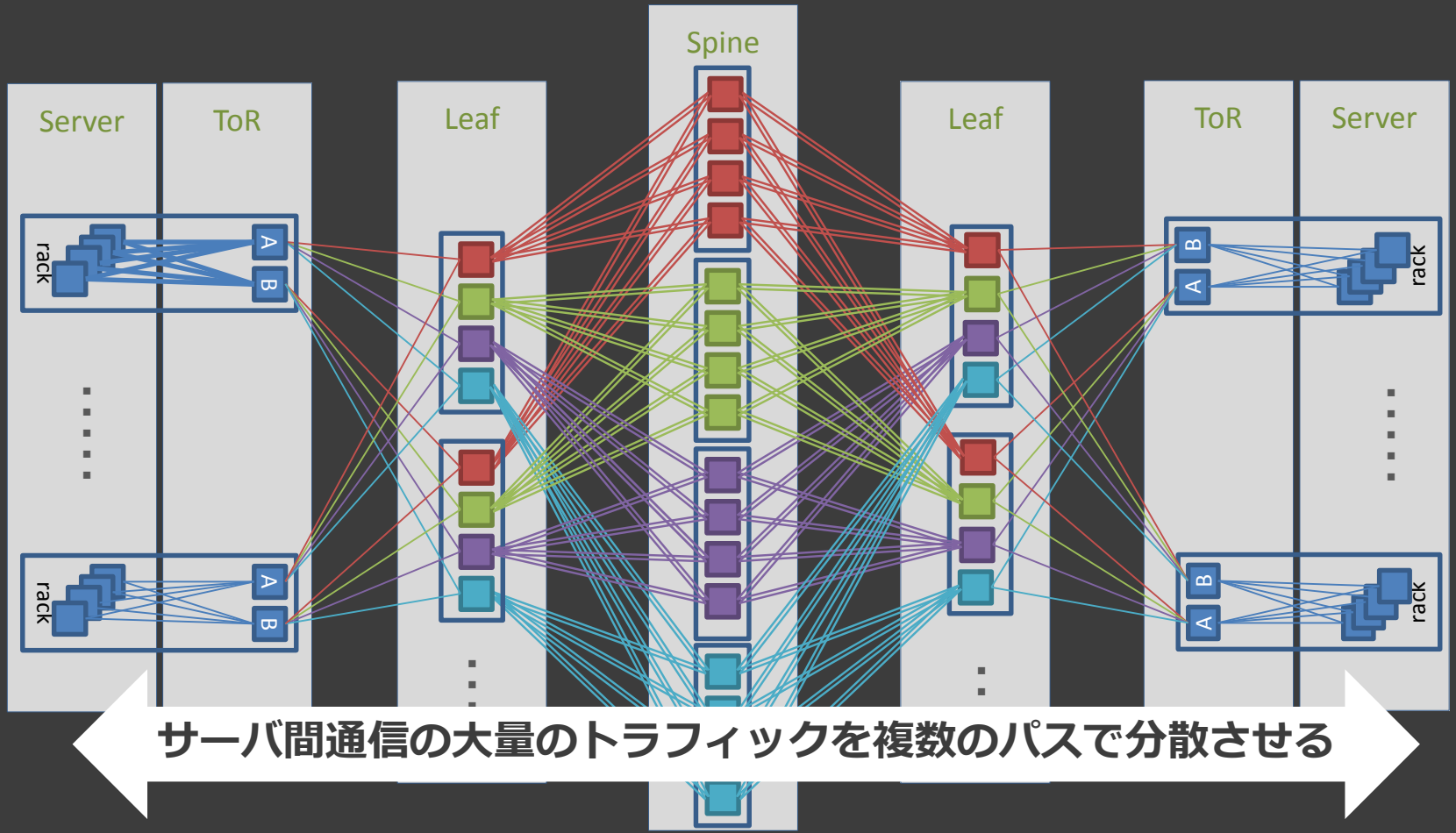


現在のデータセンタネットワーク

水平スケール型アーキテクチャ(CLOS Network)を採用

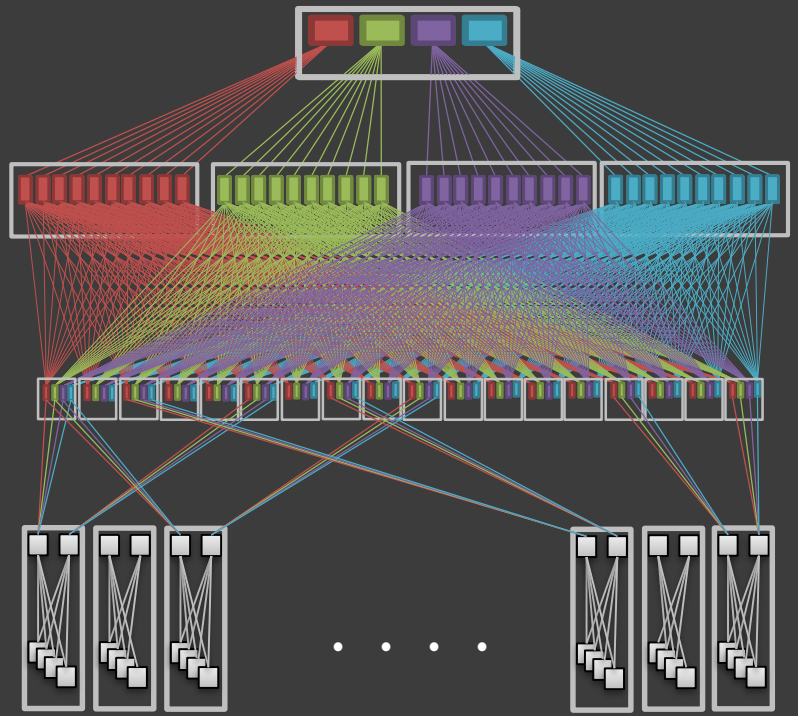


CLOS network (展開図)



現在のデータセンタネットワーク

設計の目的



スケーラビリティの最大化

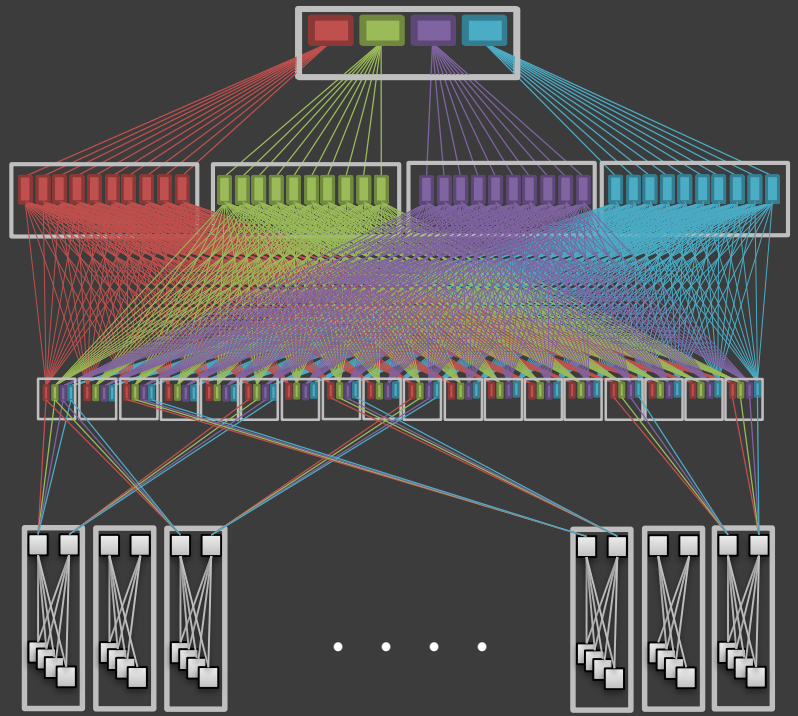
- 同種のコンポーネント(スイッチ)で構成
- 同一階層のスイッチを横展開
- 多数のLinkでトラフィックを分散

安定運用≒故障時の縮退率最小化

- N+1の冗長構成
- 数本のLinkが故障しても影響は少ない
- 機器故障の縮退率をできるだけ小さく

現在のデータセンタネットワーク

ディスアグリゲーションによるユーザの選択肢の増加



ハードウェアの選択

要件を満たすハードウェア(≒ASIC)を選択

ホワイトボックススイッチの採用

ソフトウェアの選択

運用の効率化と自動化が目標

自分たちに最適な実装(Network OS)の選択

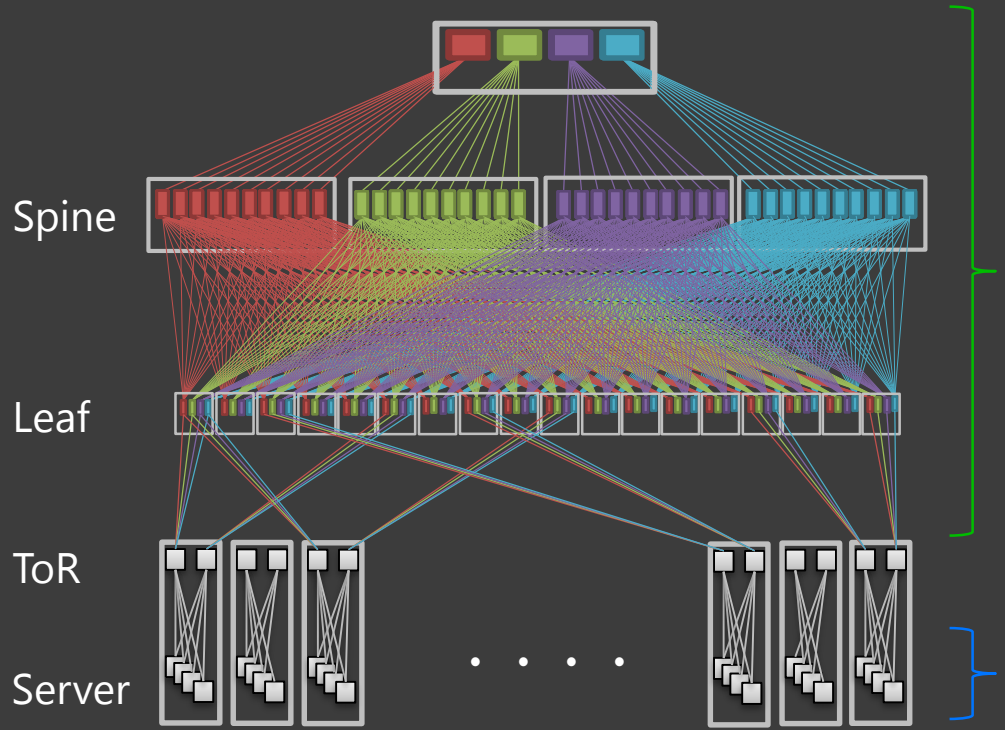
ファイバシステムの選択

数多くの伝送規格

サードパーティ製品を含めた選択肢

現在のデータセンタネットワーク

ネットワーク全体を100G化



100G switch 472台 / POD

ToR: 360

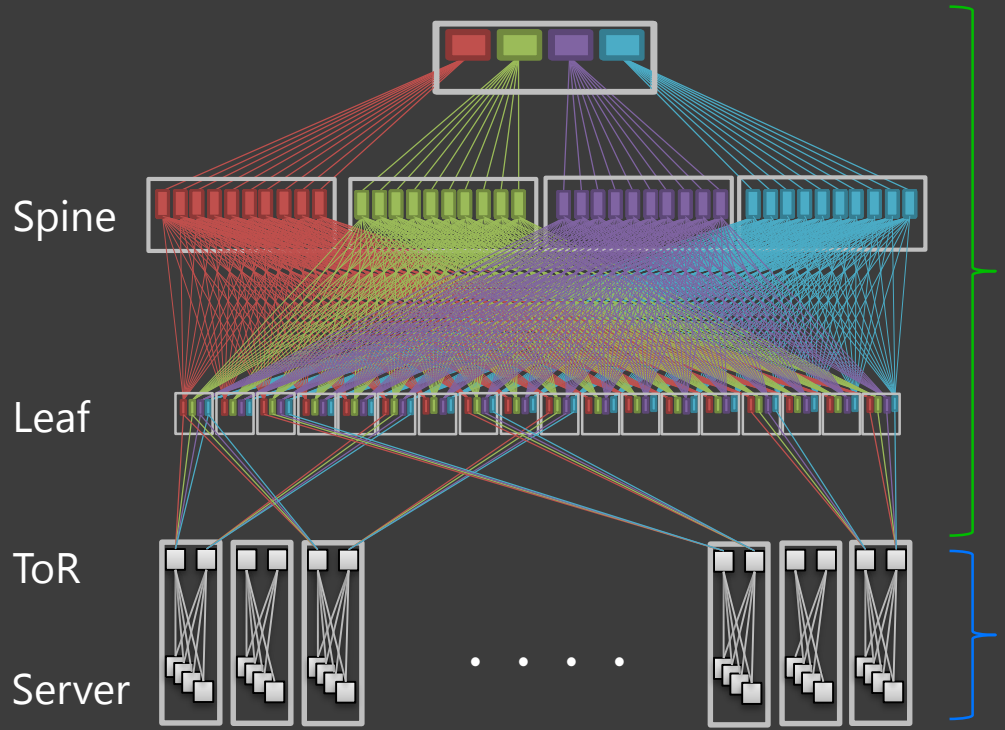
Leaf: 72

Spine: 40

7,200 servers / POD

現在のデータセンタネットワーク

ネットワーク全体を100G化



100G x 2,240 links / POD

ToR Uplink: 100G x 1,440

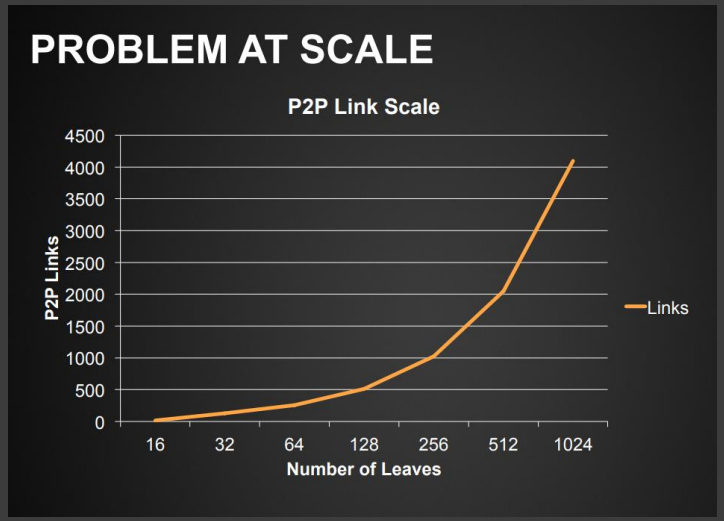
Leaf Uplink: 100G x 720

Spine Uplink: 100G x 80

10G x 14,400 links / POD

アーキテクチャとファイバシステムの関係

コスト構造の変化



MULTI-STAGE CLOS ARCHITECTURES
https://www.nanog.org/sites/default/files/monday.general.hanks.multistage.10.pdf

P2P Linkの数が大幅に増加

光トランシーバの数も増加 → Link数 x2

水平スケール型アーキテクチャの宿命

光トランシーバの導入コストに占める割合が増加

1,000+
100G Switches

5,000+
100G links

10,000+
100G Transceivers

LINEの光トランシーバ利用例

一般に普及している規格を採用し特殊実装は回避

1. **100G-SR4**

   同一ルーム内の機器(ラック)間配線に最も多く利用、CWDM4より安価

   SWDM4(DLC)と比較対象になるが、対応機種が少ない

2. **100G-CWDM4**

   DCのフロアやビルを跨ぐ機器接続に利用(既設のSMFを流用可能)

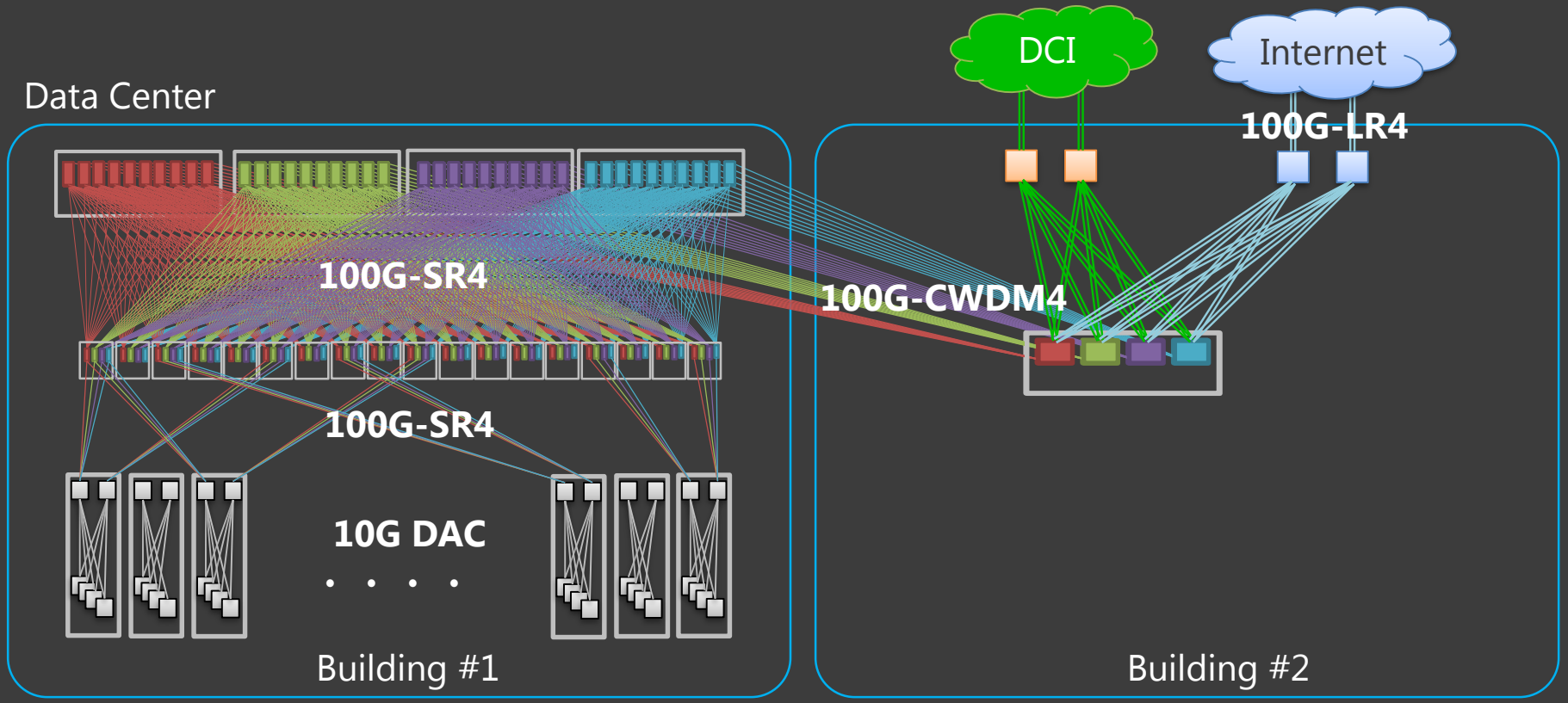
   組み合わせによってはFECの動作に注意が必要(な場合がある)

3. **100G-LR4**

   主にルータでPeeringなどの対外接続用に利用

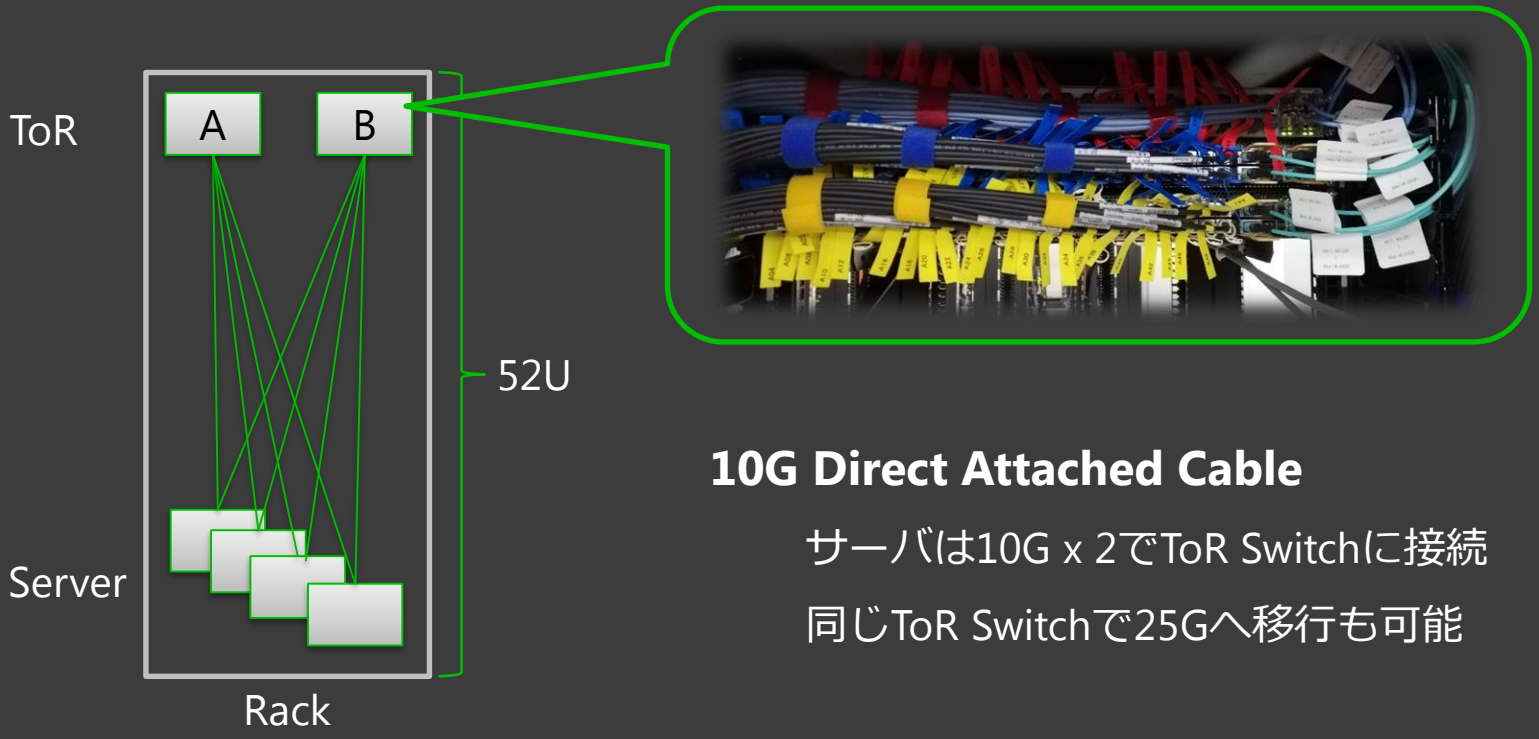
LINEの光トランシーバ利用例

各規格の伝送特性ごとに適材適所で利用



ラック内配線

サーバ接続



10G Direct Attached Cable

サーバは10G x 2でToR Switchに接続

同じToR Switchで25Gへ移行も可能

ラック間配線

MPOケーブルの配線

100G-SR4

MPO12(MMF)でスイッチ間を接続

MPOの取り扱いに慣れる必要がある

MPO 128port Patch Panel

各ラックのToR SwitchをLeaf Switchに接続

構成上、多ポートのパッチパネルが必要

ファイバシステム

配線の整合性を自動で確認

```
$ sudo  ptmctl -d
--------------------------------------------------------------------------------------------------
port    cbl     exp              act              sysname    portID  portDescr  match    last
        status  nbr              nbr                                            on       upd
--------------------------------------------------------------------------------------------------
swp1    pass    S1-01R01A:swp54  S1-01R01A:swp54  S1-01R01A  swp54   swp54      IfName   3d:1h: 2m:56s
swp2    pass    S1-01R02A:swp54  S1-01R02A:swp54  S1-01R02A  swp54   swp54      IfName   3d:20h:47m:21s
swp3    pass    S1-01R03A:swp54  S1-01R03A:swp54  S1-01R03A  swp54   swp54      IfName   3d:1h: 6m:51s
swp4    pass    S1-01R04A:swp54  S1-01R04A:swp54  S1-01R04A  swp54   swp54      IfName   3d:21h:11m:43s
swp5    pass    S1-01R05A:swp54  S1-01R05A:swp54  S1-01R05A  swp54   swp54      IfName   3d:21h:11m: 9s
swp6    pass    S1-01R06A:swp54  S1-01R06A:swp54  S1-01R06A  swp54   swp54      IfName   3d:21h:11m:42s
swp7    pass    S1-01R07A:swp54  S1-01R07A:swp54  S1-01R07A  swp54   swp54      IfName   3d:21h:11m:43s
swp8    pass    S1-01R08A:swp54  S1-01R08A:swp54  S1-01R08A  swp54   swp54      IfName   3d:21h:11m:42s
(snip)
```

大量のケーブルを手作業で確認するのは困難

予め定義された接続情報とLLDPで取得した情報を比較、配線ミスを自動で検出

配線作業後の確認にかかる時間を大幅に短縮

ファイバシステムの運用

基本的に扱いは消耗品

故障時の対応

障害Linkのトラフィックを迂回して、新品に交換するのみ

多くの場合、翌営業日対応で問題ない

→ 機器が壊れることを前提に設計されたネットワークの利点

導入時に直面した問題

100G FEC(誤り訂正)機能

異なるベンダの機器間のデフォルト設定でCWDM4がLink Upせず

機器本体とトランシーバ、設定の組み合わせを総当たりで検証し切り分け

LINEでは一部を明示的にOFFで運用している

→ 推奨されるものでないため事前の検証は必須 & 最大伝送距離に注意！

```
set interfaces et-4/0/5 gigether-options fec none
```

```
interface Ethernet1/33
  fec off
```

サードパーティ品の場合、自前で調査・解決することが必要

まとめ

データセンタネットワークでも100G opticsはすでに当たり前の存在に

ディスアグリゲーションの流れで、ユーザの選択肢が増えつつある

各コンポーネントについて、より深い理解が欠かせなくなっている

アーキテクチャの変更に伴いファイバシステムのコストが支配的になった

伝送特性や規格の詳細を理解し、自社にとって最適な構成で利用する

導入にあたっての各種検証は必須、不明点はコミュニティなどで相談する

最後に

最近の光通信技術のトレンドへの期待と現状

400G

現時点で今すぐに400Gの機器をデータセンタに導入する予定は無い

伝送路のオープン化

こちらに大きく期待している

既存のスイッチで利用可能な100G QSFP28で伝送が実現できると嬉しい

数年以内をターゲットに80km未満のDCIの実現に期待


LINE