draft-agt-rtgwg-dragonfly-routingを試してみた

ネットワンシステムズ株式会社
平河内 竜樹

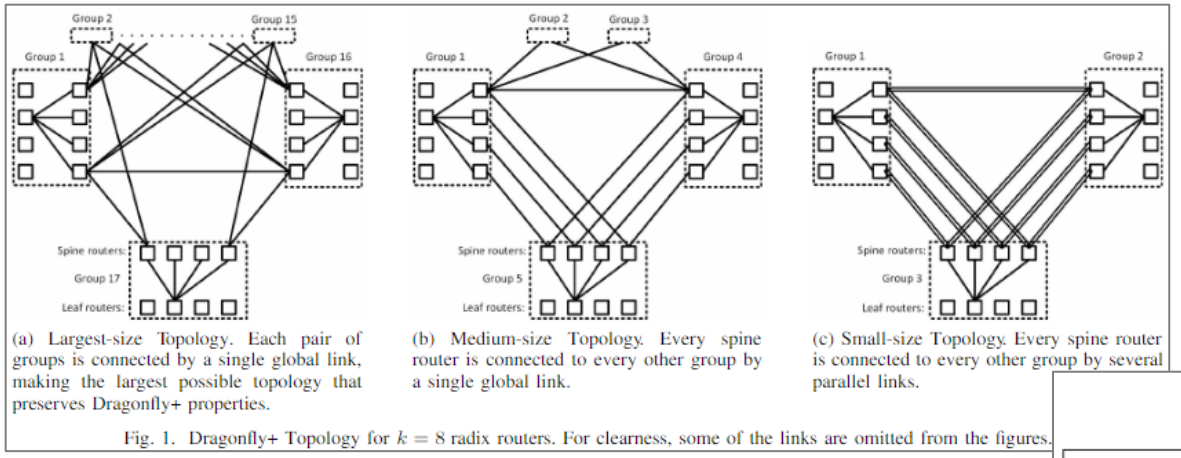
Dragonfly+ってなぁに？

■leaf-spine networkの各グループをspine間接続で直結するトポロジ

□主に大規模分散システムで選択肢となる

➢Fat Treeと比較すると「ホスト収容効率」や「遅延」などの点で優位

JANOG52『AI/ML基盤の400G DCネットワークを構築した話』でも言及あり



(a) Largest-size Topology. Each pair of groups is connected by a single global link, making the largest possible topology that preserves Dragonfly+ properties.

(b) Medium-size Topology. Every spine router is connected to every other group by a single global link.

(c) Small-size Topology. Every spine router is connected to every other group by several parallel links.

Fig. 1. Dragonfly+ Topology for $k = 8$ radix routers. For clearness, some of the links are omitted from the figures.

最大ホスト数　　：多

ホップ数　　：少

TABLE I
TOPOLOGIES COMPARISON FOR ROUTER RADIX OF 36.

	Dragonfly+	Dragonfly	3-level Fat Tree with 2:1 blocking ratio	3-level Fat Tree non-blocking	SlimFly
Maximal number of hosts	105,300	26,406	15,552	11,664	6,144
Number of hosts per router	9	9	9	7.2	12
Group size	324	162	432	324	108
Uniform random network throughput	≈ 100%	≈ 100%	≈ 50%	100%	≈ 100%
Worst case permutation network throughput	≈ 50%	≈ 8.33% (≈ 42% by [6])	≈ 50%	100%	≈ 50%
Number of VLs	2	3 (4 by [6])	1	1	4
Diameter	3	3	4	4	2
Maximal Assured Route Length	6	5 (6 by [6])	4	4	4

Source: *Dragonfly+: Low Cost Topology for Scaling Datacenters*

Dragonfly+ってなぁに？

■当該トポロジに対応したパス選出とロードバランシングが必要

□IP Networkにおけるルーティング ⇒ draft-agt-rtgwg-dragonfly-routing

➢VRF/BGPを活用することで、 ホップ数が異なるパスを等コストとして扱う

1. High priority : Minimal route (LGL) [Minimal path]
2. Medium priority : Intermediate spine route (LGGL) [Non-minimal path]
3. Low priority : Intermediate leaf route (LGLLGL) [Non-minimal path]

Dragonfly+では異なるホップ数となる3種類のパスが存在

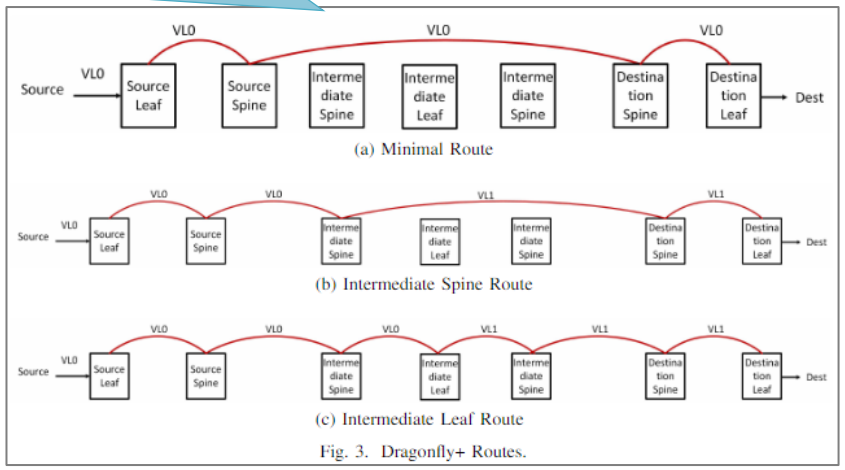
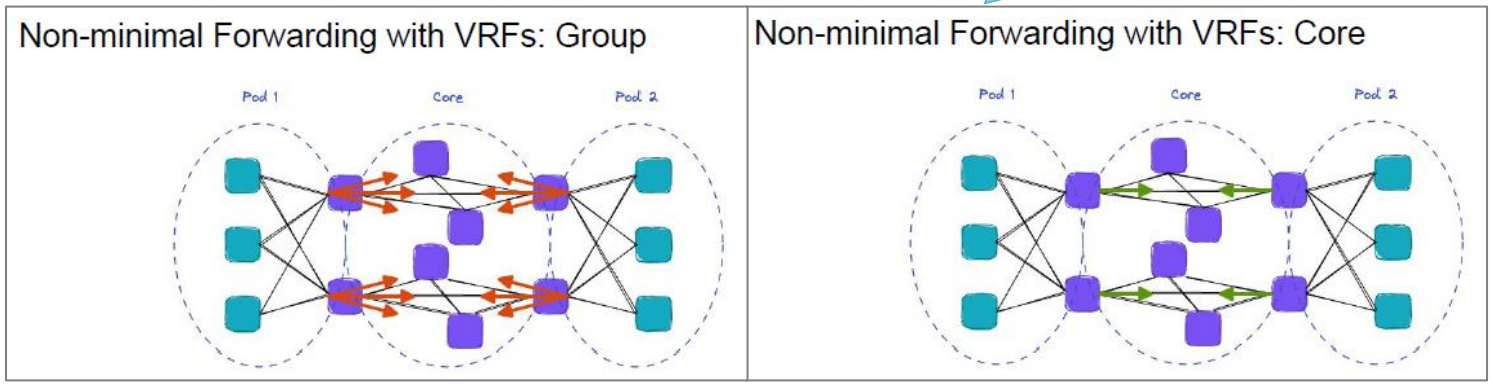


Fig. 3. Dragonfly+ Routes.

Source: *Dragonfly+: Low Cost Topology for Scaling Datacenters*

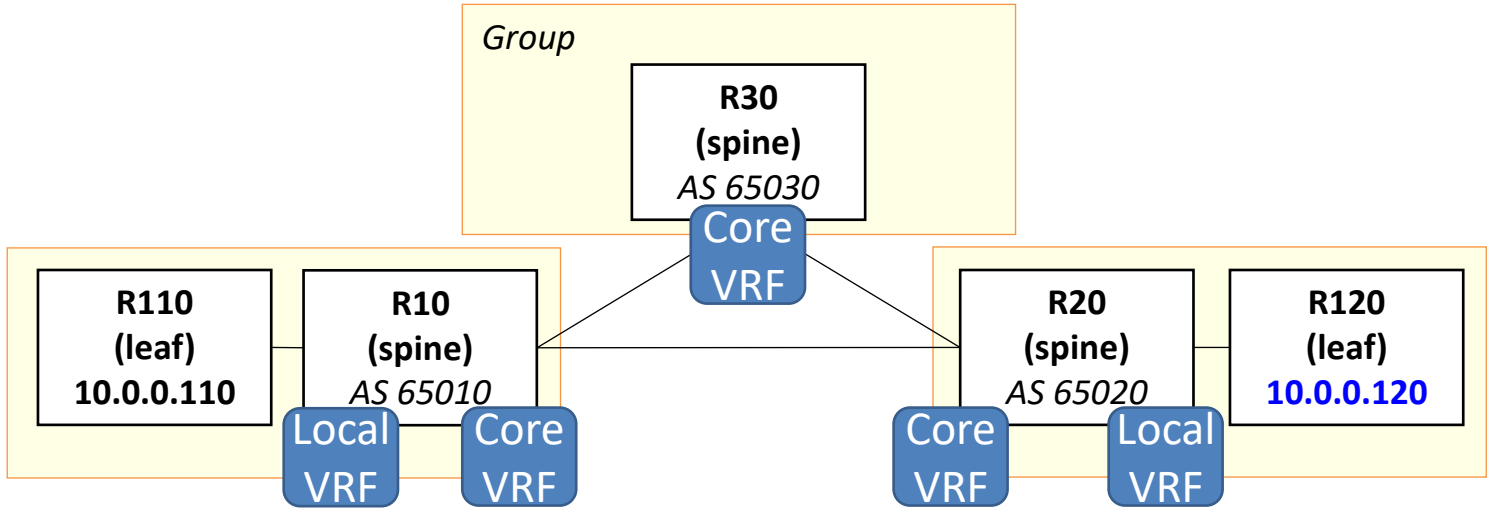
VRFを用い参照テーブルを使い分けることによってDragonfly+でマルチパスとループ回避の両立を実現



Source: *https://datatracker.ietf.org/meeting/117/materials/slides-117-rift-dragonfly-routing-00*

やってみた

■LGGL




```
R10#show bgp vpnv4 unicast all | b Net
     Network          Next Hop            Metric LocPrf Weight Path
Route Distinguisher: 65010:1 (default for vrf LOCAL) VRF Router ID 10.0.0.10
 *>i  10.0.0.110/32    10.10.110.110            0    100      0 i
 *m   10.0.0.120/32    10.10.30.30                           0 65030 65020 i
 *>                    10.10.20.20                           0 65020 65020 i
Route Distinguisher: 65010:2 (default for vrf CORE) VRF Router ID 10.0.0.10
 *>i  10.0.0.110/32    10.10.110.110            0    100      0 i
 *    10.0.0.120/32    10.10.30.30                           0 65030 65020 i
 *>                    10.10.20.20                           0 65020 i
R10#
```

Core VRF間で交換された経路をLocal VRFへインポートする際、交換時に付与されたCommunity属性に基づき、ASパスを調整（同じ長さとなる様に）

Non-minimal Routing with BGP - LGGL

BGP policies allow to implement additional logic taking into account network topology:
Simple counting scheme to limit number of hops announce will travel in the core and to prevent path hunting
- Add C1 when sending announce to the core (if neither C1 nor C2 are present)
- Add C2 when propagating announce with C1
- Don't propagate announces with C2

Make min (C1) and min+1 (C1 & C2) routes eligible for ECMP or WCMP on import into the pod VRF:
- prepend AS-PATH for routes with C1 only
- or rewrite AS-PATH

✓ **Local VRFではMinimal pathとNon-minimal pathでマルチパス**

```
R10#show ip route vrf LOCAL bgp | b Gate
Gateway of last resort is not set

      10.0.0.0/8 is variably subnetted, 4 subnets, 2 masks
B        10.0.0.110/32 [200/0] via 10.10.110.110, 00:33:26
B        10.0.0.120/32 [20/0] via 10.10.30.30 (CORE), 00:00:36
                       [20/0] via 10.10.20.20 (CORE), 00:00:36
R10#
```

✓ **Core VRFではMinimal pathだけが存在**

```
R10#show ip route vrf CORE bgp | b Gate
Gateway of last resort is not set

      10.0.0.0/8 is variably subnetted, 7 subnets, 2 masks
B        10.0.0.110/32 [200/0] via 10.10.110.110 (LOCAL), 00:33:17
B        10.0.0.120/32 [20/0] via 10.10.20.20, 00:04:27
R10#
```

Source: https://datatracker.ietf.org/meeting/117/materials/slides-117-rift-dragonfly-routing-00

■（続く）