

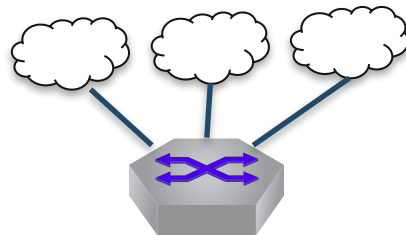
Spotifyから学んだこと -BGP SDN Internet RouterからAristaFlexRouteへ-

Shishio Tsuchiya
shitsuchi@arista.com

Spotify

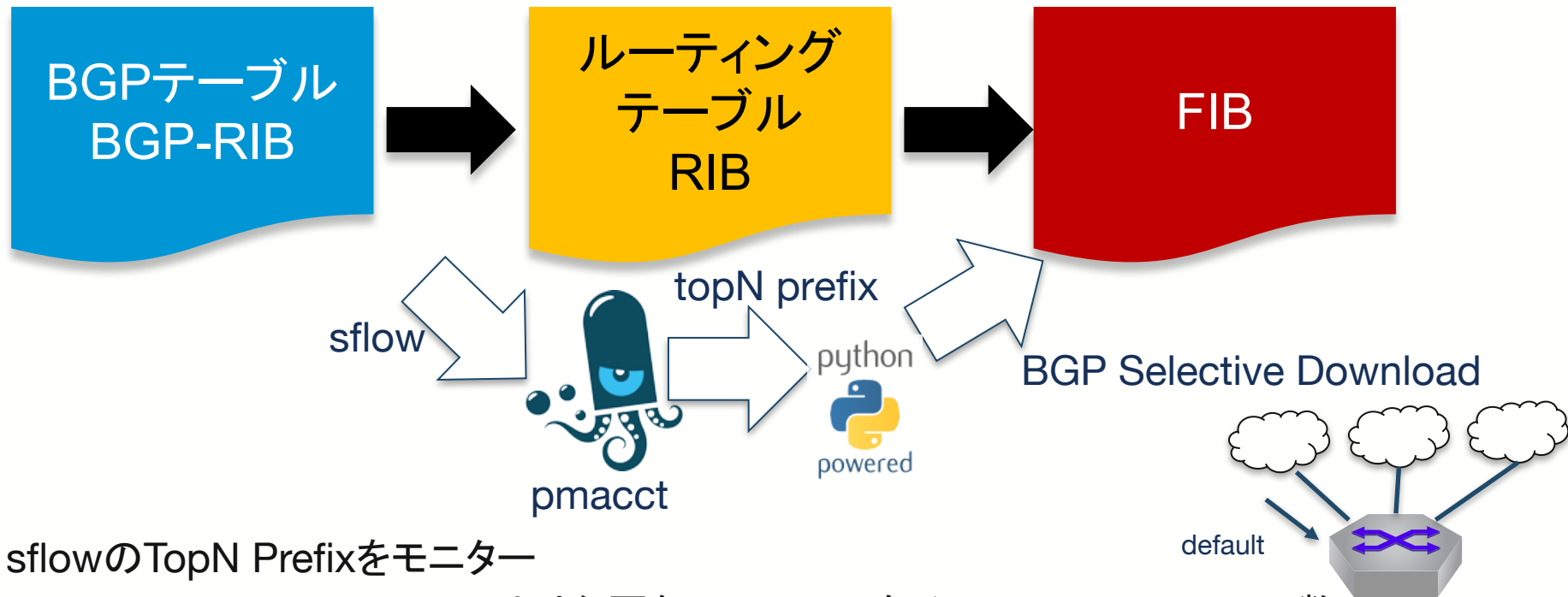


- スtockホルムの音楽定額配信プロバイダー
- 会員数は約2400万人(有料会員数600万人)
- 4つの主要データセンター
 - スtockホルム、ロンドン、アッシュバーン、サンノゼ
 - ユーザは最も近いデータセンターへ接続
- 低価格、省電力、低遅延のマーチャントシリコンを使用したいが、Trident+は32k、Aradは64k
- インターネットは+500K
- BGPフルルートは本当に必要か？



http://www.pmacct.net/dbarroso_plucente_waltzing_v0.6.pdf

BGP Selective Download by pmacct topN Prefix



- sflowのTopN Prefixをモニター
- BGP Selective downloadにより必要なPrefixのみをインストールしてルート数の制限があるマーチャントシリコンでBGPインターネットルーティングを実施

BGP Selective Download

```
!  
ip prefix-list ALLOWDEFAULT  
  seq 10 permit 0.0.0.0/0  
!  
ip prefix-list PLIST1  
  seq 1 permit 112.76.0.0/15  
  seq 2 permit 108.161.189.0/27  
  seq 3 permit 108.161.188.0/28  
  seq 4 permit 117.110.0.0/15  
  seq 5 permit 108.161.187.0/27  
!...  
! ###Truncated for brevity ###  
!  
route-map FIBIN permit 5  
  match ip address prefix-list ALLOWDEFAULT  
!  
route-map FIBIN permit 10  
  match ip address prefix-list PLIST1  
!
```

```
router bgp 65000  
  router-id 172.254.254.1  
  bgp route install-map FIBIN  
  neighbor 127.0.0.9 remote-as 65000  
  neighbor 127.0.0.9 description "pmaacct"  
  neighbor 127.0.0.9 transport remote-port 1179  
!  
sflow sample 16384  
sflow destination 127.0.0.9 9999  
sflow source-interface Loopback0  
sflow run  
!
```

BGP Selective Download

TopNをPermit

<https://eos.arista.com/arista-eos-bgp-selective-route-download/>

srdtool.py

URL Based Prefix-list

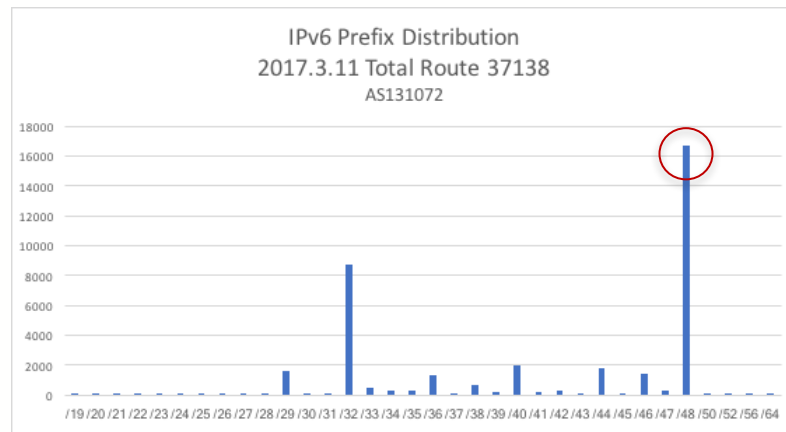
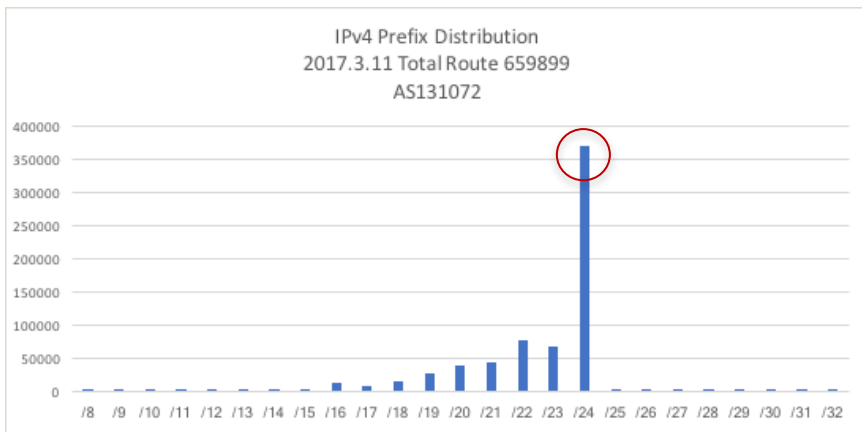
```
vEOS2(config)#ip prefix-list martiaon source flash:martin.txt
vEOS2(config)#show ip prefix-list martiaon
ip prefix-list martiaon source flash:/martin.txt
vEOS2(config)#refresh ip prefix-list martiaon
refreshed ip prefix-list martiaon source flash:/martin.txt
Num entries: 6
vEOS2(config)#show ip prefix-list martiaon detail
ip prefix-list martiaon source flash:/martin.txt
! Num entries: 6. Update time: 2017-03-12 12:40:33
    seq 5 permit 0.0.0.0/8 le 32
    seq 10 permit 127.0.0.0/8 le 32
    seq 15 permit 169.254.0.0/16 le 32
    seq 20 permit 192.0.2.0/24 le 32
    seq 25 permit 224.0.0.0/4 le 32
    seq 30 permit 240.0.0.0/4 le 32
vEOS2(config)#ip prefix-list martiaon source ?
file:    URL to load prefix-list entries
flash:   URL to load prefix-list entries
http:    URL to load prefix-list entries
https:   URL to load prefix-list entries
scp:     URL to load prefix-list entries
```

- Prefix-listの管理はISP毎にルールも異なり、非常に大変
- URLベースで読み込みが可能 / detailやbashで確認可能
- refreshで読み込みやファイル更新時のアップデートが可能
- 4.18.0F

FIB Optimization Technology

- FIBにはLPM(Longest Prefix Match)とLEM(Longest Exact Match)の2つのテーブルが存在
- LPMは様々なPrefix長のテーブルで、LEMは固定長のテーブル
- 一般的にはLPMはルーティングテーブルに使用され、LEMはホストテーブルに使用される
- インターネットルーティングテーブルを見てみると、偏りが見られる

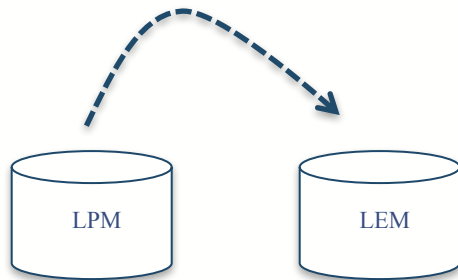
インターネットルーティングテーブル



- 一方、インターネットルーティングテーブルはIPv4は+660K/IPv6+37K
- プレフィックス毎の分布を見ると、IPv4では/24が55.91% IPv6は/48 44.96%(APNIC Lab AS131072のケース)
- ルーティングテーブルはLPMを利用するので、多いエントリーはLEMに移す

何をLPMからLEMに移すのか？

- EOSではshow ip route sumでPrefix Distributionを表示
- maskcount.pyではこれをカウントする



```
bgp-edge2>sh ip route sum
```

Route Source	Number Of Routes
connected	14
static (persistent)	0
static (non-persistent)	0
VXLAN Control Service	0
static nexthop-group	0
ospf	0
Intra-area: 0 Inter-area: 0 External-1: 0 External-2: 0	
NSSA External-1: 0 NSSA External-2: 0	
ospfv3	0
bgp	349512
External: 1781 Internal: 347731	
isis	0
Level-1: 0 Level-2: 0	
rip	0
internal	27
attached	61
aggregate	0
Total Routes	349614

Number of routes per mask-length:

/8: 9	/9: 5	/10: 20	/11: 62	/12: 153
/13: 271	/14: 564	/15: 1026	/16: 6962	/17: 4524
/18: 7277	/19: 15269	/20: 22422	/21: 23962	/22: 40007
/23: 32956	/24: 193252	/25: 195	/26: 100	/27: 57
/28: 76	/29: 75	/30: 65	/31: 19	/32: 225

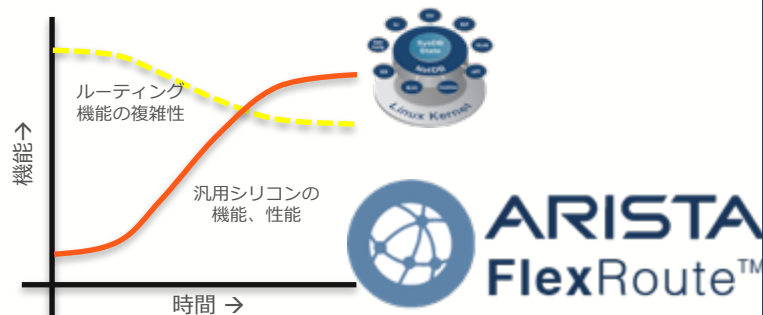
FIB Optimization

- /24と/32をLPMからLEMに移動

```
7280(config)#ip hardware fib optimize prefix-length 32 24
7280(config)#show platform arad ip route summary
Total number of VRFs: 1
Total number of routes: 30017
Total number of route-paths: 12335
Total number of lem-routes: 17682
Total number of /24 routes in lem: 17680
Total number of /32 routes in lem: 2
```

```
bgp-edge(config)#ip hardware fib optimize prefix-length 24 ?
0      Prefix length 0
1      Prefix length 1
10     Prefix length 10
11     Prefix length 11
12     Prefix length 12
13     Prefix length 13
14     Prefix length 14
15     Prefix length 15
16     Prefix length 16
17     Prefix length 17
18     Prefix length 18
19     Prefix length 19
2      Prefix length 2
20     Prefix length 20
21     Prefix length 21
22     Prefix length 22
23     Prefix length 23
25     Prefix length 25
26     Prefix length 26
27     Prefix length 27
28     Prefix length 28
29     Prefix length 29
3      Prefix length 3
30     Prefix length 30
31     Prefix length 31
32     Prefix length 32
4      Prefix length 4
5      Prefix length 5
6      Prefix length 6
7      Prefix length 7
8      Prefix length 8
9      Prefix length 9
<cr>
```

Arista FlexRoute



- FlexRouteではこれらのチューニングを自動的に実施
- 1Mを超えるルーティングをマーチャントシリコンのみで実施する為、省電力/低遅延を単体でサポート

```
bgp-edge2(config)#ip hardware fib optimize prefixes profile internet
Bgp-edge2(config)#ipv6 hardware fib optimize prefixes profile internet
```

```
bgp-edge2(config)#show ip bgp sum
```

```
BGP summary information for VRF default
```

```
Router identifier 109.105.96.55, local AS number 2603
```

```
Neighbor Status Codes: m - Under maintenance
```

```
Neighbor      V  AS      MsgRcvd   MsgSent   InQ  OutQ   Up/Down   State   PfxRcd  PfxAcc
109.105.97.24  4  2603      286314      96      0      0 00:19:38  Estab   348325 348325
(...)
```

```
bgp-edge2>sh hardware capacity
```

```
Forwarding Resources Usage
```

Table	Feature	Chip	Used Entries	Used (%)	Free Entries	Committed Entries	Best Case Max Entries	High Watermark
ECMP			13	0%	4082	0	4095	13
ECMP	Mpls		0	0%	4082	0	4095	0
ECMP	Routing		13	0%	4082	0	4095	13
EEDB		Jericho0	72	0%	81848	0	81920	81
EEDB	ARP	Jericho0	72	0%	81848	4096	81920	81
EEDB	MplsTunnel	Jericho0	0	0%	77824	0	81920	0
EEDB	Tunnel	Jericho0	0	0%	77824	0	81920	0
FEC			365	0%	126611	0	126976	440
FEC	Mpls		0	0%	126611	0	126976	0
FEC	Routing		365	0%	126611	0	126976	440
LEM			308259	39%	478173	0	786432	309157
LEM	MAC		15	0%	478173	0	786432	15
LEM	Mpls		0	0%	478173	0	786432	0
Routing	Resource1		550	26%	1498	0	2048	553
Routing	Resource2		331	32%	693	0	1024	335
Routing	Resource3		9944	30%	22824	0	32768	9964
Routing	V4Hosts		0	0%	81920	0	81920	0
Routing	V4Routes		297264	38%	478173	0	786432	298162
Routing	V6Hosts		0	0%	81920	0	81920	0
Routing	V6Routes		10980	2%	478173	0	786432	11000
TCAM	ACL	Jericho0	0	0%	22528	0	24576	0
TCAM	PBR	Jericho0	0	0%	22528	0	24576	0
TCAM	QOS	Jericho0	0	0%	22528	0	24576	0
TCAM	Subnets		0	0%	512	0	512	0

まとめ

- SpotifyはマーチャントシリコンでBGP Selectiveダウンロードを実施
 - これにより小容量のルータにて実際に使用してるユーザのみのインターネットルーティングを可能に
 - Prefix-listのURLロードなどの機能を追加
- インターネットルーティングにおけるFIBの最適化を実施
- これらのスクリプトは下記にて公開
 - <https://eos.arista.com/arista-eos-bgp-selective-route-download/>
 - **SDN Internet Router (sir)**
 - <https://github.com/dbarrosop/sir#sdn-internet-router-sir>
 - **fib_optimizer**
 - https://github.com/dbarrosop/sir_apps/tree/master/fib_optimizer
- FIBの最適化を自動的に柔軟に行えるようArista FlexRouteを開発
 - TCAMなどは使わずに省電力でのインターネットルーティングを実現
 - https://www.arista.com/assets/data/pdf/Whitepapers/FlexRoute-WP_jp.pdf



Thank You

www.arista.com