Switch as Border Router

Small ISP without FIRT

ITNOG on the WEB 05/2020

nicola modena
@nmodena

Agenda

- Motivation
- RIB vs FIB
- IXP and Transit
- Selective routing installation and propagation
- Route Selection for optimized routing
- Self assessment
- Switch/Asics selection

Software or Hardware based forwarding

+ Flexibility, Cost
- Perf, Scale

+ Performance, Cost
- Flexibility, FIB Scale(*)

+ Perf, Flexibility, Scale
- Cost

Border router requirement

Border, Core & Aggregation routers for modern ISP must have:

- High Forwarding capacity
- High speed port density
- High RIB / FIB capacity
- Enhanced Load Balancing
- HW access lists (on the edge)
- Deep buffers (on link speed changes)

RIB vs FIB

Routing Information Base hold all the destination from the best path route selection

One or more entry for each destination are combined and installed in the forwarding table

FIB on ASICs uses fixed size TCAM for efficient lookup.

example: Broadcom Trident III: 128K FIB Entry

Routing Information Base

from Best Path Selection

```
SHOW IP ROUTE                C
SHOW ROUTE                   J
```

DRAM

Forwarding Information Base

```
SHOW IP CEF                  C
SHOW ROUTE FORWARDING-TABLE  J
```

TCAM

ISP Backbone without FIRT

use of just default-route to reach external destination (*)

PROS

- **Lower Resources (FIB)**
- **Faster Convergence**
- **Easier Management**
- **Cheap devices**

CONS

- **sub-optimal routing**
- **No Transit Customers**

Possible solution : Route selection & Selective FIB Download

(*) backbone with multiple and distributed exit points

1

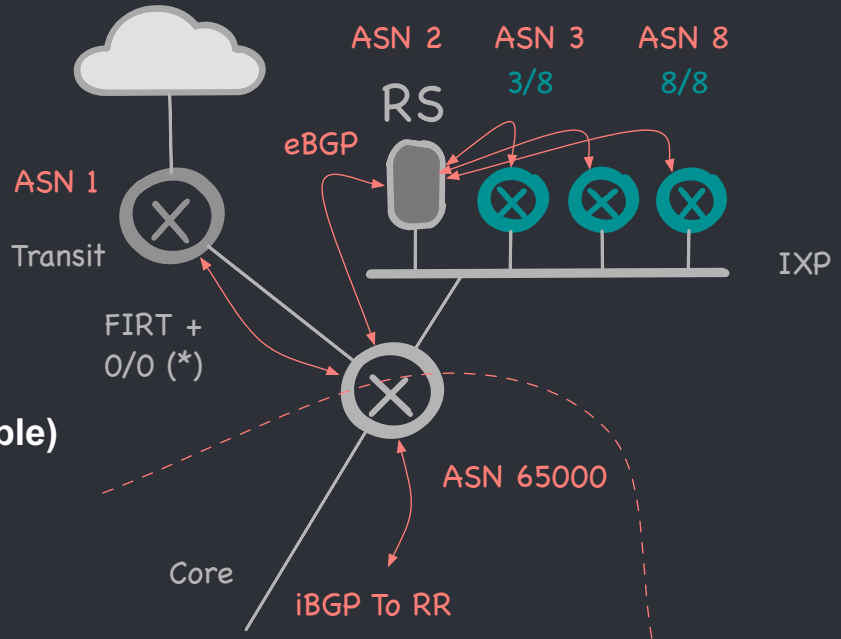
A pratical example

Border router with peers and transit

IXP Routing

IXP Peering
- **eBGP peering with Route Server**
- **no RS ASN in path**
- **next-hop unchaged**
- **BGP update of all members (customizable)**

Transit
- **FIRT Full Internet Routing Table**
- **and/or default-route**

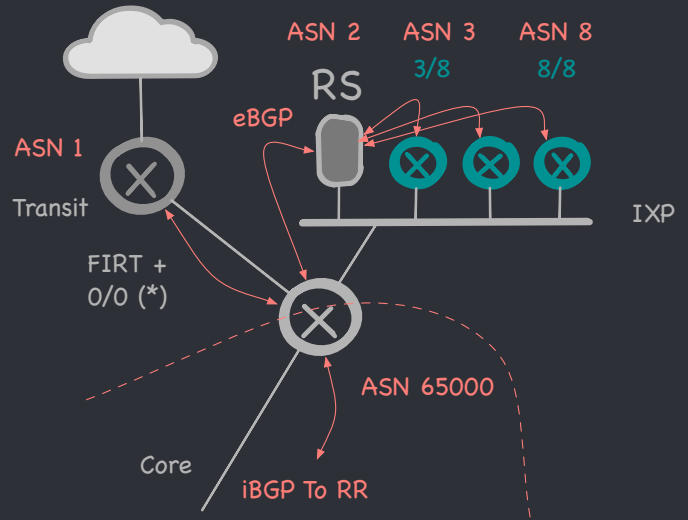


Our goal is to optimize router resources for this scenario

Border router with selective FIB install

IXP bgp configuration

```
protocols {
    bgp {
        group IXP {
            neighbor 172.16.0.1 {
            import from-IXP;
            export only-my-AS;
            peer-as 2;
}}}}

policy-options {
    policy-statement from-IXP {
        term accept-all {
            then {
                community add FIB;
            }
        }
    }
    community FIB members 65000:800;
}
```



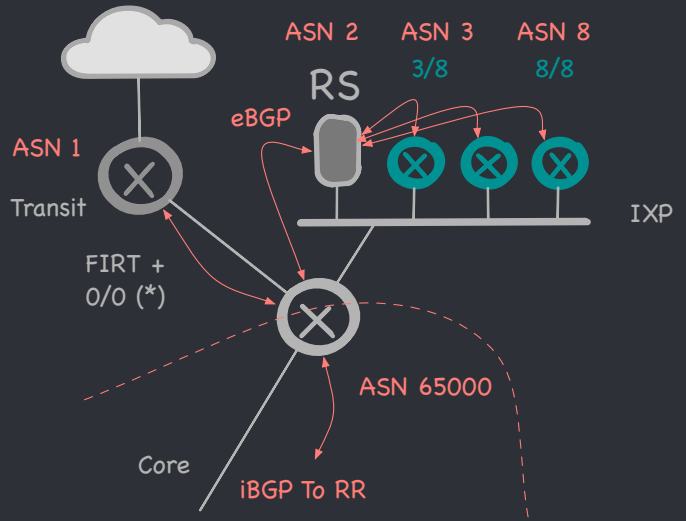
Tag all the relevant prefixes with a community

OPT: filter-out irrilevant destinations reached in other peering points or trough transits

Border router with selective FIB install

Transit bgp import policy

```
policy-options {
    policy-statement from-TRANSIT {
        term default-route {
            from {
                route-filter 0.0.0.0/0 exact;
            }
            then {
                community add FIB;
                next policy;
            }
        }
        term default {
            then {
                community add RIB-ONLY;
            }
        }
    }
    community FIB members 65000:800;
    community RIB-ONLY members 65000:801;
}
```



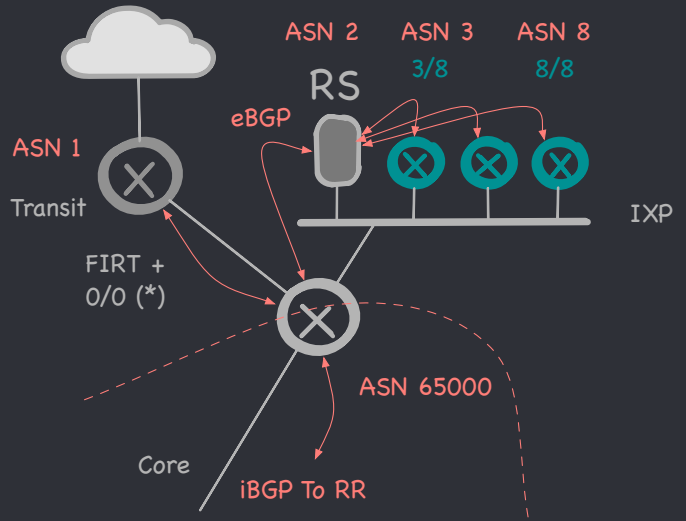
Tag all transit routes with a different community except the default-route.

The FIRT is accepted and may be propagated internally for further usage

Border router with selective FIB install

Controlled FIB install

```
routing-options {
    forwarding-table {
        export selective-FIB-install;
    }
}

policy-options {
   policy-statement selective-FIB-install {
        term RIB-ONLY {
            from {
                protocol bgp;
                community RIB-ONLY;
            }
            then reject;
        }
    }
    community RIB-ONLY members 65000:801;
}
```



deploy a policy to control RIB to FIB download and reject RIB-ONLY tagged route

HINT: Almost every platform has this feature deployed for out-of-band route reflectors

Border router with selective FIB install

Checking for the RIB : all entries are present

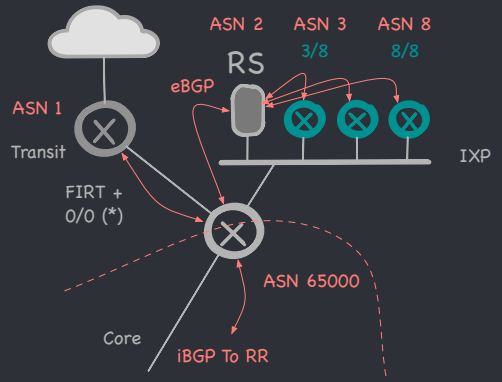
```
nmodena@MX-05> show bgp summary
Table          Tot Paths  Act Paths Suppressed    History Damp State    Pending
inet.0
Peer                     AS      InPkt     OutPkt    OutQ   Flaps Last Up/Dwn
172.16.0.1                2       1294       1304       0       0     9:39:54 Establ
  inet.0: 6/6/6/0
172.16.1.106              1    2073299          9       0       0       53:42 Establ
  inet.0: 42793/42793/42793/0

nmodena@MX-05> show route
inet.0: 42822 destinations, 42823 routes (42822 active, 0 holddown, 0 hidden)
+ = Active Route, - = Last Active, * = Both

0.0.0.0/0          *[BGP/170] 01:10:09, localpref 100
                      AS path: 1 I, validation-state: unverified
                    >  to 172.16.1.106 via et-0/0/4.0
1.0.0.0/24         *[BGP/170] 01:10:09, localpref 100
                      AS path: 1 49673 24811 13335 I, validation-state: unverified
                    >  to 172.16.1.106 via et-0/0/4.0
1.0.4.0/24         *[BGP/170] 01:10:09, localpref 100
                      AS path: 1 49673 6939 4826 38803 56203 I, validation-state: unverified
                    >  to 172.16.1.106 via et-0/0/4.0
1.0.5.0/24         *[BGP/170] 01:10:09, localpref 100
                      AS path: 1 49673 6939 4826 38803 56203 I, validation-state: unverified
                    >  to 172.16.1.106 via et-0/0/4.0
```

IXP

TRANSIT

FIRT

Border router with selective FIB install



Checking for the FIB : only relevant entries are installed

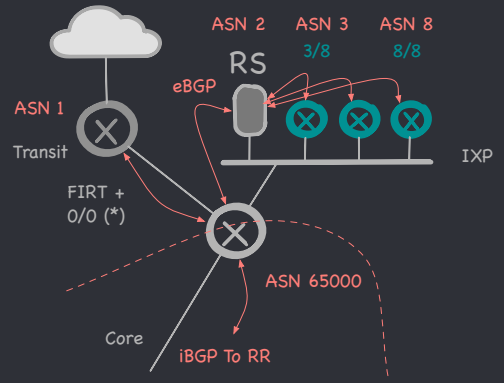
```
nmodena@MX-05> show route forwarding-table
Routing table: default.inet
Internet:
Enabled protocols: Bridging,
Destination        Type RtRef Next hop           Type Index    NhRef Netif
default            user     0 50:0:0:b:0:0       ucst      596     4 et-0/0/4.0
default            perm     0                    rjct       36     1
0.0.0.0/32         perm     0                    dscd       34     3
3.0.0.0/8          user     0 172.16.0.3         ucst      599     5 et-0/0/5.0
3.0.1.0/24         user     0 172.16.0.3         ucst      599     5 et-0/0/5.0
3.0.2.0/24         user     0 172.16.0.3         ucst      599     5 et-0/0/5.0
8.0.0.0/8          user     0 172.16.0.8         ucst      598     5 et-0/0/5.0
8.0.1.0/24         user     0 172.16.0.8         ucst      598     5 et-0/0/5.0
8.0.2.0/24         user     0 172.16.0.8         ucst      598     5 et-0/0/5.0
10.0.0.0/8         user     0                    dscd       34     3
10.0.35.0/24       intf     0                    rslv      584     1 et-0/0/3.0
10.0.35.0/32       dest     0 10.0.35.0          recv      582     1 et-0/0/3.0
[…]
```

Transit

Peers at IXP

Internal Network

Border router with selective FIB install



Compare RIB and FIB size

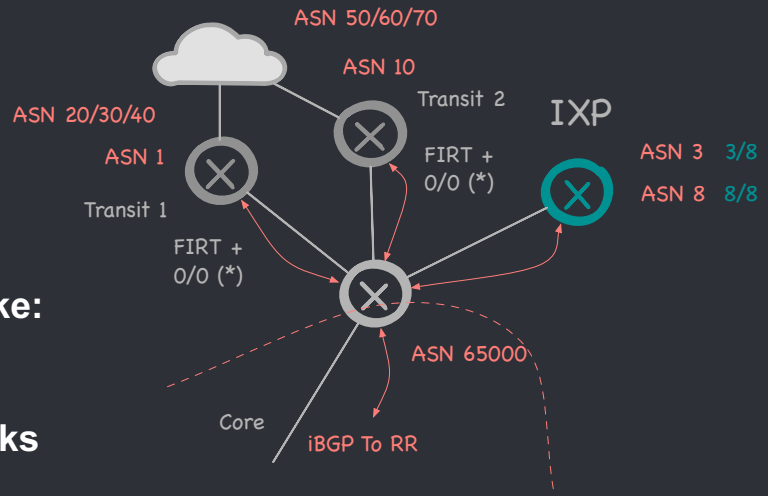
```
nmodena@MX-05> show route summary
inet.0: 42822 destinations, 42823 routes (42822 active, 0 holddown, 0 hidden)
              Direct:      5 routes,      5 active
               Local:      4 routes,      4 active
                OSPF:      9 routes,      9 active
                 BGP:  42802 routes,  42801 active
           Aggregate:      2 routes,      2 active
                 LDP:      1 routes,      1 active

inet.3: 8 destinations, 8 routes (8 active, 0 holddown, 0 hidden)
                 LDP:      8 routes,      8 active

nmodena@MX-05> show route forwarding-table summary
Routing table: default.inet
         user:         61 routes
         perm:          5 routes
         intf:          9 routes
         dest:         19 routes
```

Is it really THAT simple ?

yes, but only if you have a SINGLE TRANSIT provider and all the remaining prefix fits into the FIB

…but this is not true in most cases…

Border router with multiple Transit



Multiple transit are used to get optimal paths like:

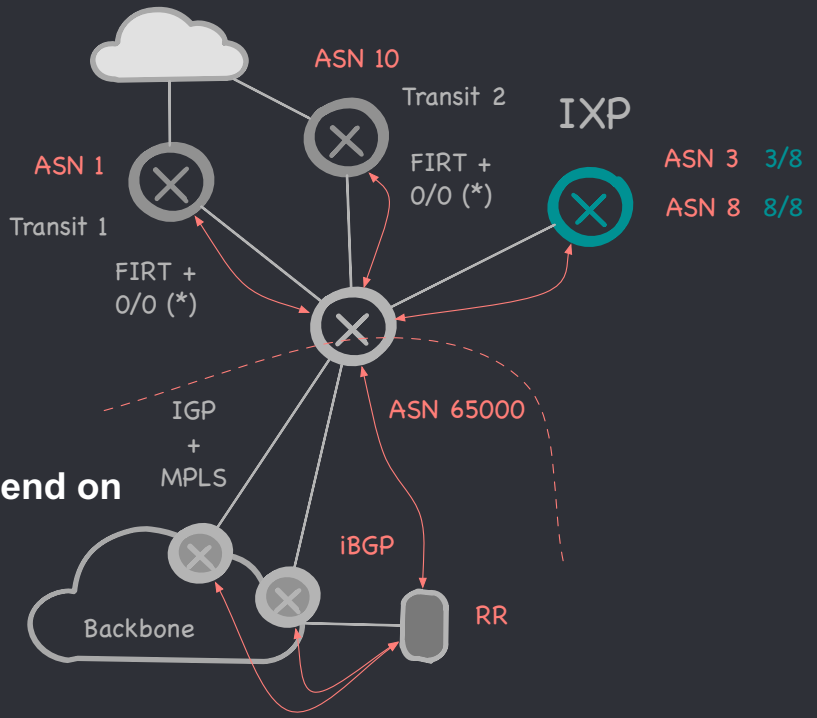
- **for ASN 20/30/40 use Transit 1**
- **for ASN 50/60/90 use Transit 2**
- **load balance the remainig traffic on both links**

this requires to

- **create policy to identify relevant prefixes**
- **check if the expected FIB size fit's in the TCAM**
- **mark this new prefixes for FIB download**

it is usually a procedure performed manually and if necessary

Border router and internal reachability



Also internal destination may be considered

and also can be reduced / summarized but depend on

- **addressing plan**
- **summarization strategy**
- **number of interfaces**
- **load balancing**
- **use of MPLS**

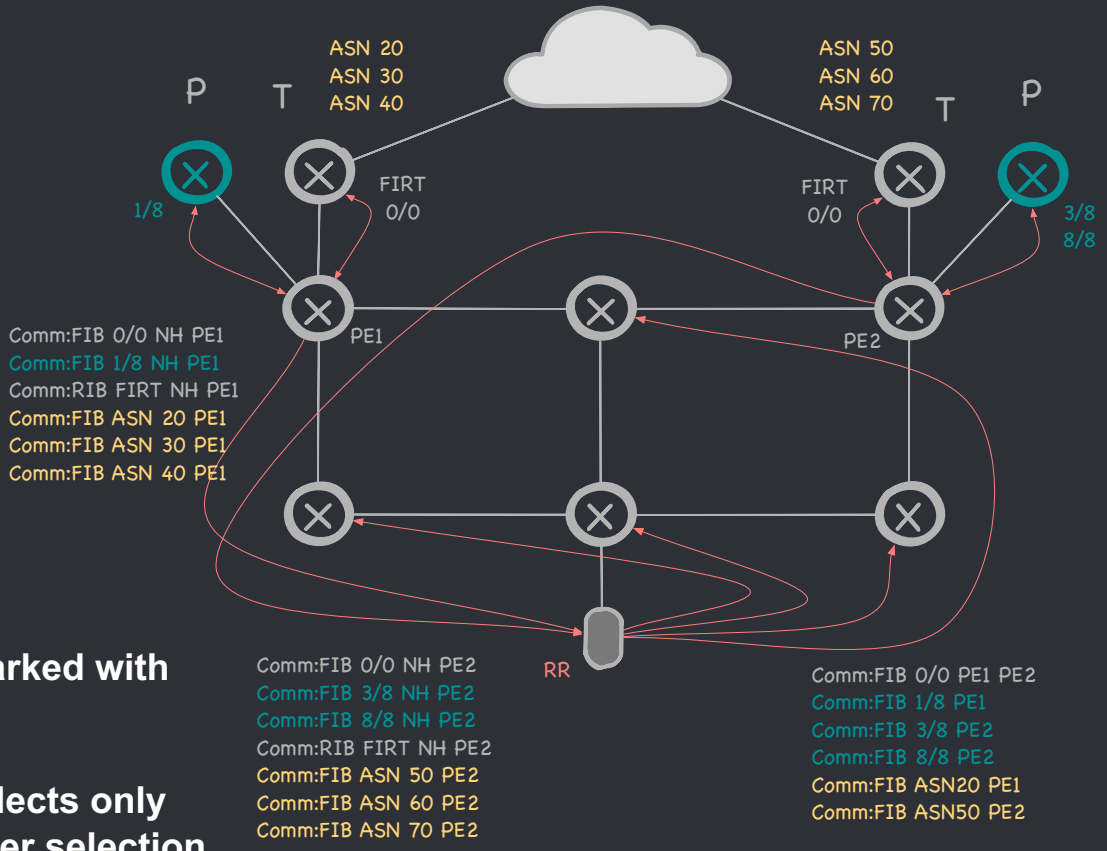
Hint: some modular devices has also FIB localization capability

3

Evolution with optimized edge selection

distribution of selected external destinations

Layer-3 solution with optimized routing

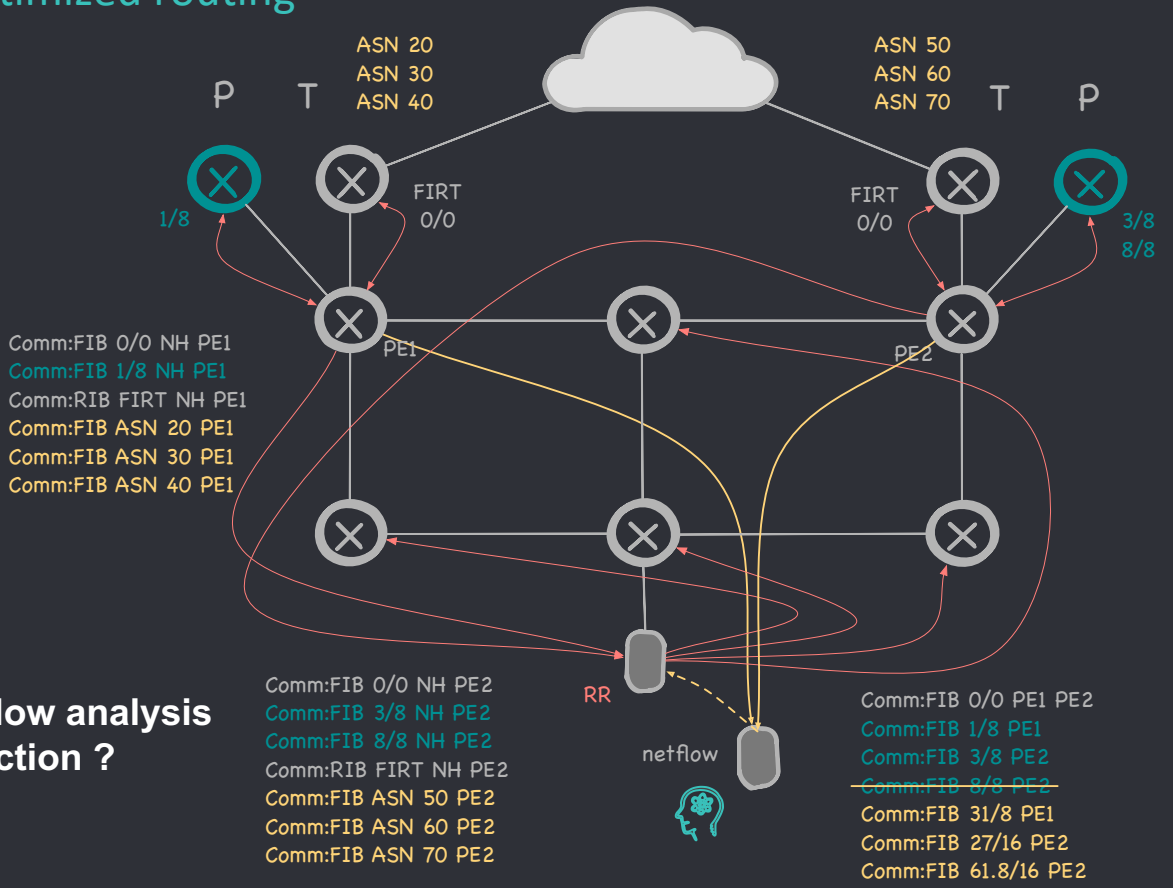


We can put all togheter:

Border routers advertise all external prefixes to RR marked with FIB or RIB-ONLY community

RR can hold the FIRT and reflects only relevant prefix to optimize peer selection

Layer-3 solution with optimized routing



why do not combine a netflow analysis for an automatic edge selection ?

What can go wrong ? ☺

How to identify relevant destinations

Manually : become a regex ninja

Automatic: create an SDN Solution like David Barroso with Paolo Lucente

https://labs.spotify.com/2016/01/26/sdn-internet-router-part-1/

https://blog.ipspace.net/2015/01/sdn-router-spotify-on-software-gone-wild.html

https://github.com/dbarrosop/sir

4 Summary

Summary - Take Away

- on ASICS resources are fixed/limited
- We can operate with unsynchronized RIB and FIB
- Easier on border router with hot-potato and default-route
- it’s possible to deploy on backbone but it’s not simple
- Partial Routing distribution and/or different hasing can lead to traffic loops
- keep an uniform forwarding information base in the backbone
- Troubleshooting requires new skills and updated procedure

- you can save a lot of money on high end device

5

Self Assessment

Is this solution good for me ?

Self Assesment

- how many prefixes I currently have in my rib ?
- how many transit or peering-points I am connected to ?
- what distance (RTT,$) there is between them ?
- Sub-optimal routing or double transit it’s an option ?
- I’m able to identify how many prefix I need to cover 90% of my destination ?
- How many time I spend troubleshooting transit issue ?
- how often I need to optimize transit traffic ?
- I have the knowledge to manage this system ?

-> balance between: TIME - KNOWLEDGE – COSTS - PERFORMANCE

Switch selection

Select Asic/Vendor based on your needs:

- FIB Size
- Filtering capability
- QOS with Filtering
- Load Balancing
- Buffers
- Netflow
- Telemetry
- Programmability
- CPU / RAM / Control Plane

0

THANK YOU

Questions ?

Nicola Modena - CCIE #19119 JNCIE-SP #986
linkedin.com/in/nmodena - @nmodena
nicola@modena.to

Acknowledge:

“Ammiraglio” Tiziano Tofoni friendship and in-depth discussions
Massimo Magnani (Arista) e Roberto Ricci (Cisco) for invaluable inputs
ITNOG community