OPTIMAL ROUTING VS. ROUTE REFLECTOR VNF - RECONCILE THE FIRE WITH WATER

Rafal Jan Szarecki #JNCIE136

Solution Architect, Juniper Networks.

AGENDA

- Route Reflector VNF - goals
- Route Reflector challenges and the traditional answer
- Route Reflector VNF – the solution

RR VNF

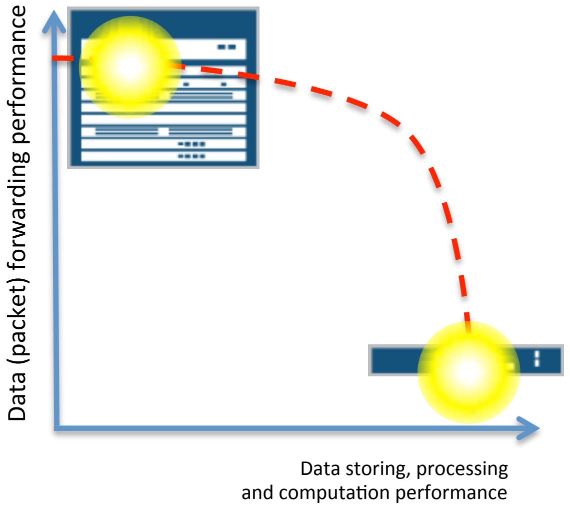
- Have RR’s (multiple/all) running on VM in Data Center, wherever DC is. So RR as a VNF.
- Ensure that RR Client (RRC) receives paths that are at least as optimal as in the traditional design.
- Keep session and path scale under control.

VNF – Virtual Network Function – and instantiation of network function based on NFV concept.

Route Reflector vs. Router

ROUTERS

- Forward packets/data
- Data storage and processing is overhead (cost of doing business)
- RR is all about data storage and processing
  - cost of doing business
  - side affect of BGP support on routers.

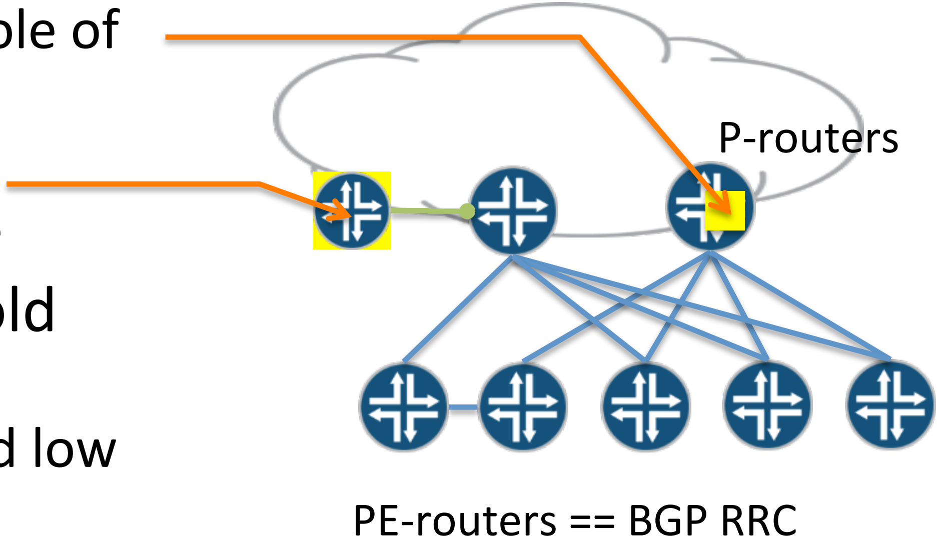

Data (packet) forwarding performance
Data storing, processing and computation performance

SERVERS

- Computes and stores data
- x86 appliance is much cheaper then router, but…
- VM is even cheaper in DC, more scalable, and easier to instantiate

Typical design w/ RR

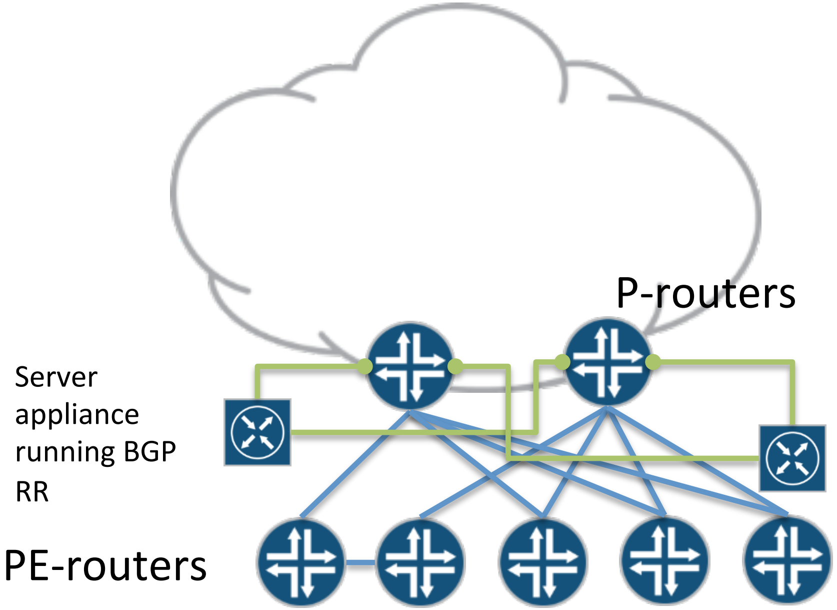
- Router performs RR function
  - In addition to its primary role of forwarder, OR
  - As a dedicated platform
- RR cluster – regional scope
- RR as retirement plan for old routers ?
  - Old router == slow CPU and low Memory
  - Short time "optimization"



Let’s optimize platform

- Replace Router based RR, with server appliance
  - Carrier-grade SW now available (commercial and/or public domain)
  - Better price to compute performance

- What we do **not** gain?
  - Interfaces to connect RR appliance into network. Also on network router.
  - Maintenance
    - trips still needed,
    - new skills needed to manage.
  - Servers HW has shorter live cycles.



Let’s optimize quantity – central RR

- That isn’t a new idea
  - Quite common and works well for VPN in many providers – 2 central RR’s
  - Not that popular for Internet – why?
- Two fundamental issues w/ Central RR’s:
  - Session and Path scaling
  - Suboptimal routing.

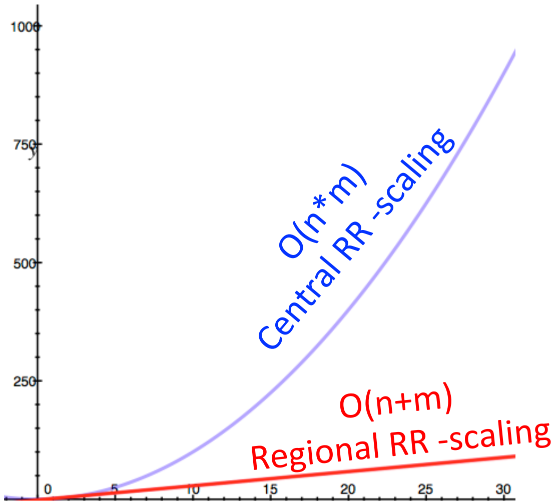
Session scaling problem

- Central RR would have to keep 1000’s of sessions (say N).

- Pressure on:
    - Memory – not big deal nowadays
    - CPU – session maintenance (keep-a-lives, BFD, counters)
    - Communication path - Any update needs to be sent N times.

Session scaling - traditional answer

- Divide and conquer
- Regional RRs - mitigates the problem.
- O(n+m) instead of O(n*m) sessions
  - n number of regions
  - m number of RRCs in each region
  - N=n*m

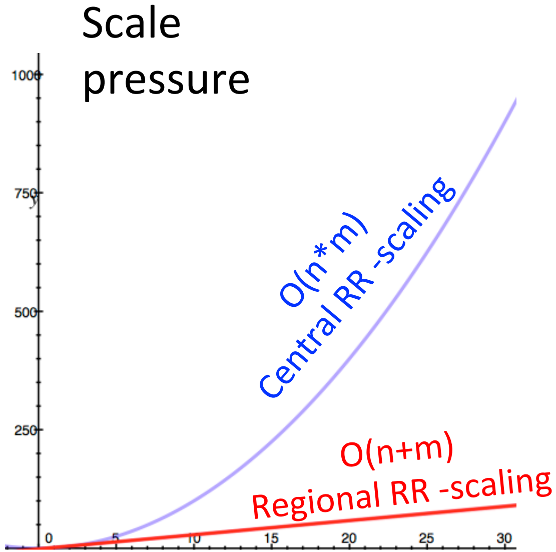


Path scaling problem

- Central RR learns all prefixes from all exit nodes
  - VPN
    - A given customer site is (usually) single or dual-homed – avg. is <2 paths per prefix
    - Typically there are less VPN prefixes than on Internet (and can always be partitioned)
  - Internet
    - Large providers have 10’s peers from who all internet prefixes are accepted.
    - Provided 0.5M prefixes - 10’s of millions of paths.
- Pressure on:
  - CPU
    - best path selection among more paths takes more time.
    - More paths – higher churn
    - There is no such thing as “too-fast convergence”

Path scaling - traditional answer

- Regional RRs - mitigates the problem.
- Less sessions -> less paths per prefix.
- O(n+m) instead of O(n*m) sessions
  - n number of regions
  - m number of RRCs in each region



Routing Optimality problem

- RR advertises only one path per prefix
  - It chooses one using BGP path-selection. BGP NextHop (NH) closer to RR - wins.
  - RRC is not aware about alternative paths – can't select it, even if it would be optimal.
- L3VPN overcomes it by having PE-unique Route Distinguisher.
- Internet doesn't have this option (no RD). Add-path instead.
  - Add all path -> scaling explosion
  - Add 2nd-best path -> still could be very un-optimal exit from RRC perspective.

Routing Optimality – traditional answer

- Regional RRs - mitigates the problem.
- Regional RR select closest - in-region – exit ASBR.
  - Semi-optimal from RRC perspective
  - Good enough.

topology independent RR VNF

- Goals
  - Have RR's (multiple/all) running on VM in Data Center, wherever DC is. So RR as a VNF.
  - Ensure that RR Client (RRC) receives paths that are at least as optimal as in the traditional design.
  - Keeps session and path scale under control.

- Assumption: DC infrastructure may have L3 elements that do not participate in any WAN routing protocol (IGP, BGP)

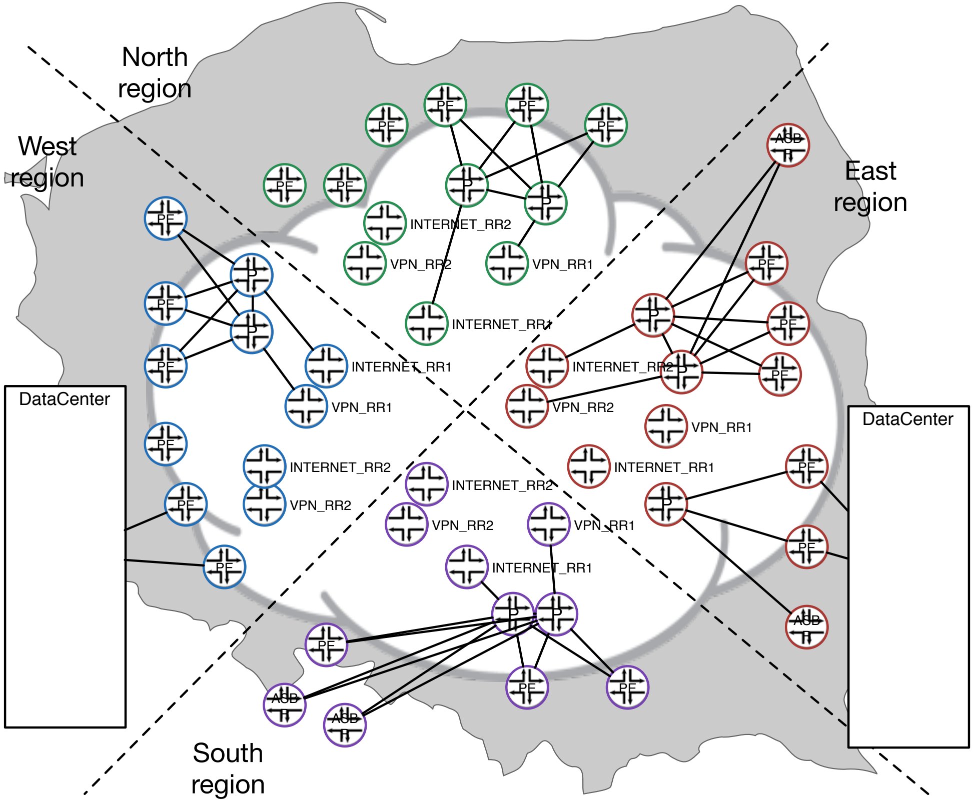
BGP path selection

1. Verify that the next hop can be resolved.
2. Choose the path with the lowest **preference value** (routing protocol process preference).
3. For BGP, prefer the path with higher **local preference**.
4. For BGP, prefer the path with the **shortest autonomous system** (AS) **path** value.
5. For BGP, prefer the route with the **lower origin code**.
6. For BGP, prefer the path with the lowest multiple exit discriminator (**MED**) metric.
7. Prefer strictly **external BGP** (EBGP) paths over external paths learned through **internal BGP** (IBGP) sessions.
8. For BGP, prefer the path whose next hop is resolved through the IGP route with the lowest metric.
9. For BGP, if both paths are external, prefer the currently active path to minimize route-flapping.
10. For BGP, prefer the path from the peer with the lowest router ID. For any path with an originator ID attribute, substitute the originator ID for the router ID during router ID comparison.
11. For BGP, prefer the path with the **shortest cluster list length**. The length is 0 for no list.
12. For BGP, prefer the path from the peer with the **lowest peer IP address**.

Example

BEFORE

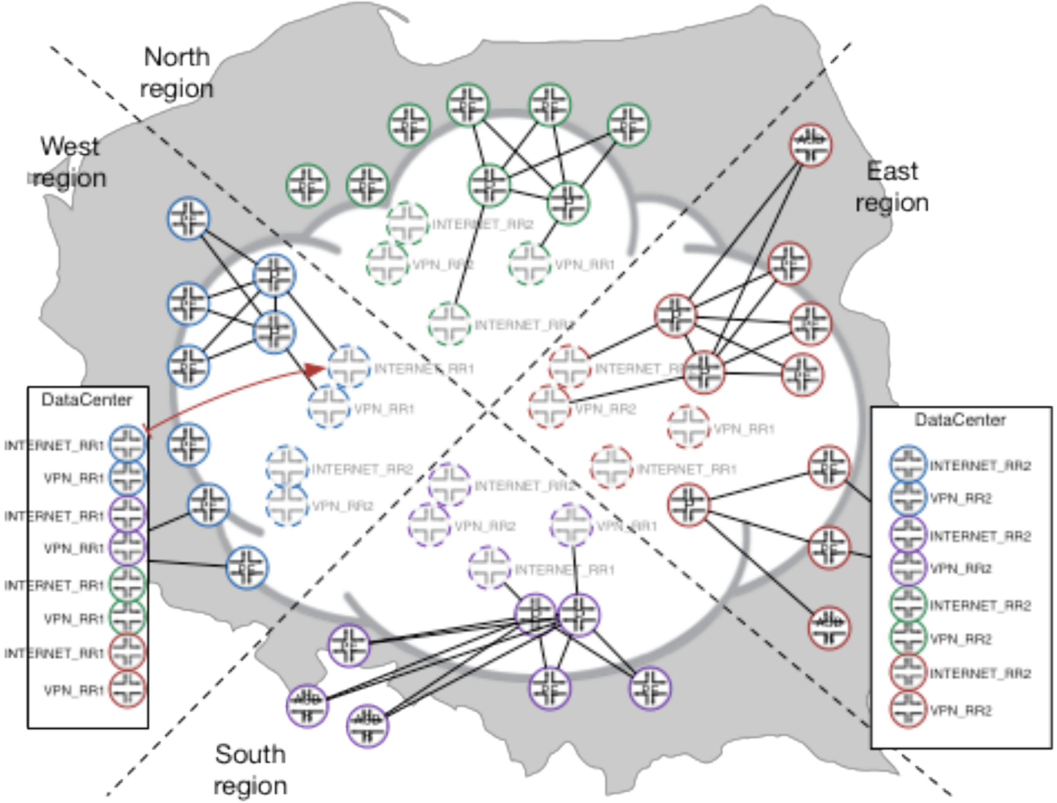
- Separate RR's for VPN and Internet
- 4 regions, 16 RR's all together.
- spread amongst 8 locations for geo-redundancy
- DC are in low-real-estate-cost parts of country.



Example

AFTER

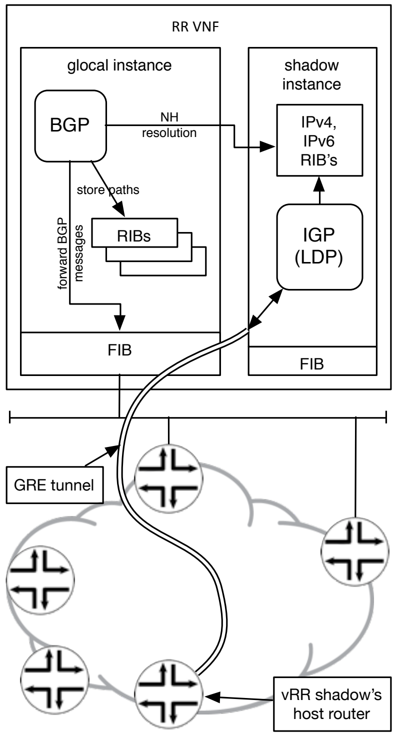
- Separate RR's for VPN and Internet
- 4 regions, 16 RR's all together.
- 16 spread amongst **2 locations** for geo-redundancy
- DC are on low-real-estate-cost parts of country.
- **RR's path selection works as they would be in "BEFORE" state**.



The Solution

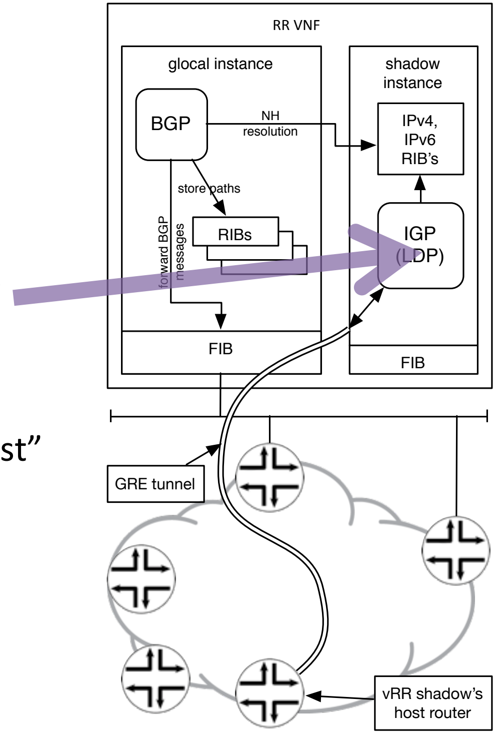
- Static route used to reach peer's loopbacks
- BGP sessions with the same set of RRCs and peer RR as usual/before.
  - Global/default instance
- "shadow" routing context/instance
  - Used only to resolve (global instance) BGP path NH
  - Let IGP of the network build routing tables of the "shadow" instance
  - IGP metrics/topology in this instance are as RR would be installed in given region.

- RESULT : BGP best PATH selection is as the RR would be directly connected to the "shadow's host".



Shadow instance

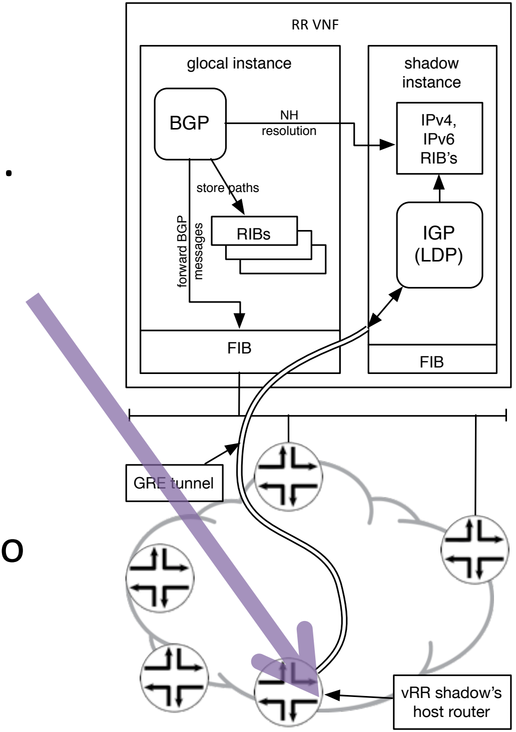
- IGP derives reachability of RRC - fast BGP convergence via NextHop tracking.
- Terminates GRE tunnel (tunnel payload).
  - IGP and BFD run inside tunnel. Metric is MAX-1
  - For security reasons, no transit traffic inside tunnel (filters or other means)
- IGP adjacency with the “shadow’s host” router(s)
  - Metrics/Topology view as RR would be connected to “shadow’s host”
  - Over GRE tunnel
- Policies
  - IGP set to overload state.
  - IGP routes are not inserted into FIB of “shadow” instance.
  - Only host routes (/32 and /128; loopbacks) are accepted from IGP’s LSDB into routing table. Only these routes are visible for BGP NH resolution.



LDP/RSVP may be needed as well, depending on RR VNF implementation for BGP labeled NLRI resolution.

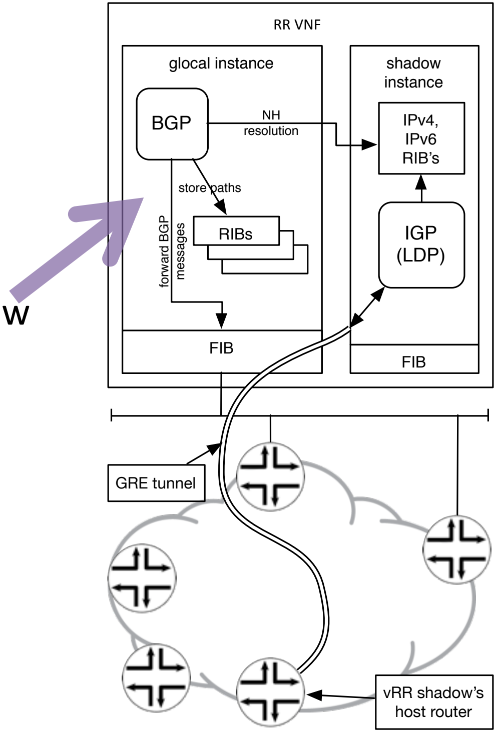
Shadow's host

- Regular router deep in the network. The one that would connect dedicated RR in a traditional design.
- It holds GRE tunnel to RR VNF.
  - Tunnels destination is the RR's loopback (global context)
  - Tunnels payload is terminated into shadow instance of RR.
  - IGP and BFD runs inside tunnel. Metric is MAX-1
  - For security reasons, tunnel configuration ensures no transit traffic (filters or other means)



RR VNF global instance

- BGP protocol is configured in the global context as usual on Route Reflectors.
- Only static routes toward client's loopback's exist in the default instance FIB (other routes filtered out).
  - 2 gateways
  - BFD
- BGP NH resolver configured to use routing tables (RIB) of the shadow instance.
  - NH tracking
  - IGP Topology/metrics form remote regon point of view
- Routing and forwarding of GRE packets (after encapsulation) to "shadow's host" router.
- Forwarding of BGP (and BFD) packets to BGP peers (RRC and other RR's).
  - Not tunneled.
  - Follow shortest IGP path



RR VNF - network resiliency

- 2 GRE tunnels are recommended. Each to a separate “shadow’s host” router.

- Ideally both “shadow’s host” should have same IGP path cost to members of given RR cluster.

Routing recursion avoidance

- Running protocol over tunnel (IGP over GRE) – It is asking for a problem.
  - Recursion → packet losses → IGP session flaps → BGP NH invalid → BGP withdraws
  - Total network failure !
- Use of separate routing instances **solves the problem**:
  - Default routing instance is used to send GRE (tunneling) packets (and it is based on static routing only)
  - Shadow routing instance participate in IGP routing. It may learn the tunnel destination address but:
    - this route is never used to forward GRE (tunneling) packets.
    - this route could be used to forward **tunneled** packets – reach "shadow's host" for IGP, BFD, OAM, etc.

RR VNF – prevent transit traffic

- RR VNF has limited forwarding capacity
- But it is part of the IGP w/ 2 interfaces (GRE redundancy) – could end up on the shortest path for some traffic.
- To prevent:
  - High tunnel metric
  - Overload bit/state
  - Filters on entry into GRE at “shadow’s host”

Extending concept

- Multiple "shadow" instances – different topology and different egress node selection per address family. Example:
  - Internet and L3VPN BGP paths resolved in "shadow" but,
  - VPLS paths resolved in "shadow2" and
  - Inet.2 paths resolved in "shadow3", etc
- Each of the "shadowX" have GRE tunnels to other "shadow's hosts" → different topology view → different IGP cost to BGP NHs.

Summary

- Instantiate RR as VNF could provide optimal routing
  - Regardless of physical VNF (Data Centre) location
  - Evolutionary alternative solution to ORR

- NH resolution and resulting Path selection could be differentiated on per Address Family basis.

- Protection of RR VNF against unwanted transit traffic needs to be in place.

THANK YOU