Routing w Sieci - Praktyczne aspekty implementacji protokołu BGP

Rafał Szarecki

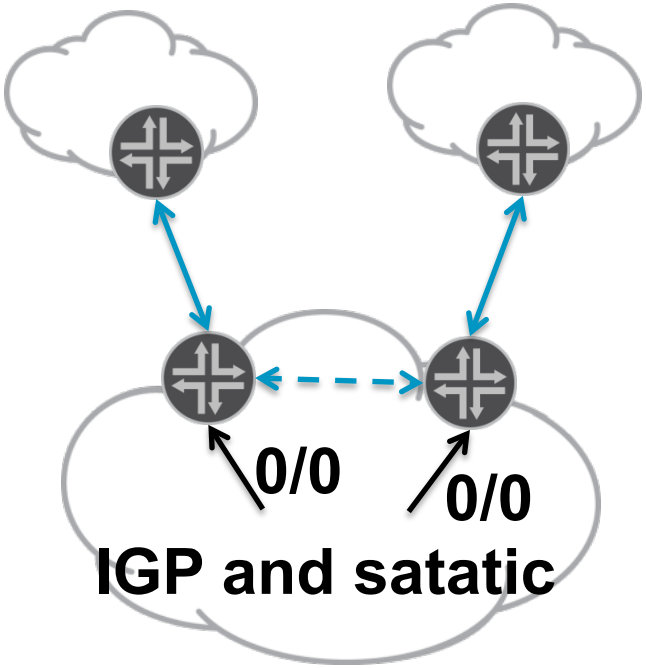
15/03/2011

BGP – Czym jest

- Protokół routingu miedzy operatorskiego

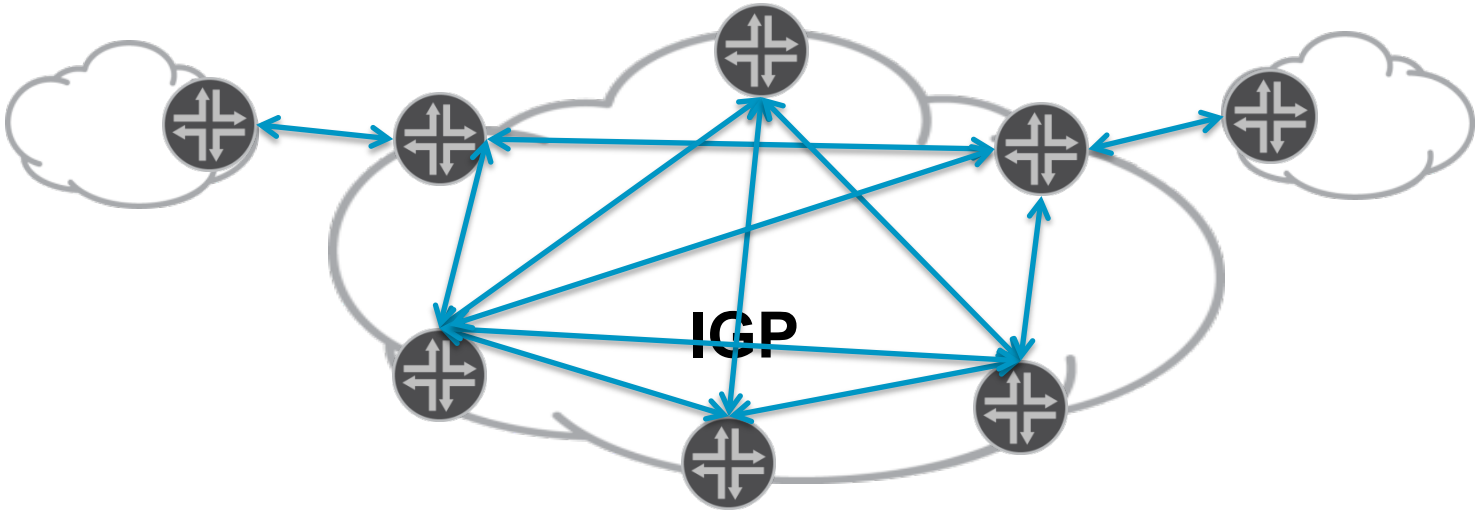
  Pomiędzy ASBR różnych AS

  90% sieci bez BGP



BGP – Nie tylko routing miedzyoperatorski

- Nie tylko – e.g. L3VPN – większość implementacji wewnątrz jednego AS.
  - Pomiędzy PE tego samego AS
  - 90+% sieci używa BGP (BGP free core)
- Protokół sygnalizacyjny osiągalności (reachability) usług.
- Protokół sygnalizacyjny auto-konfiguracji (auto-discovery) usług.
- Wszystkie routery (service-aware) muszą być w stanie wymieniać informacje pomiędzy sobą.



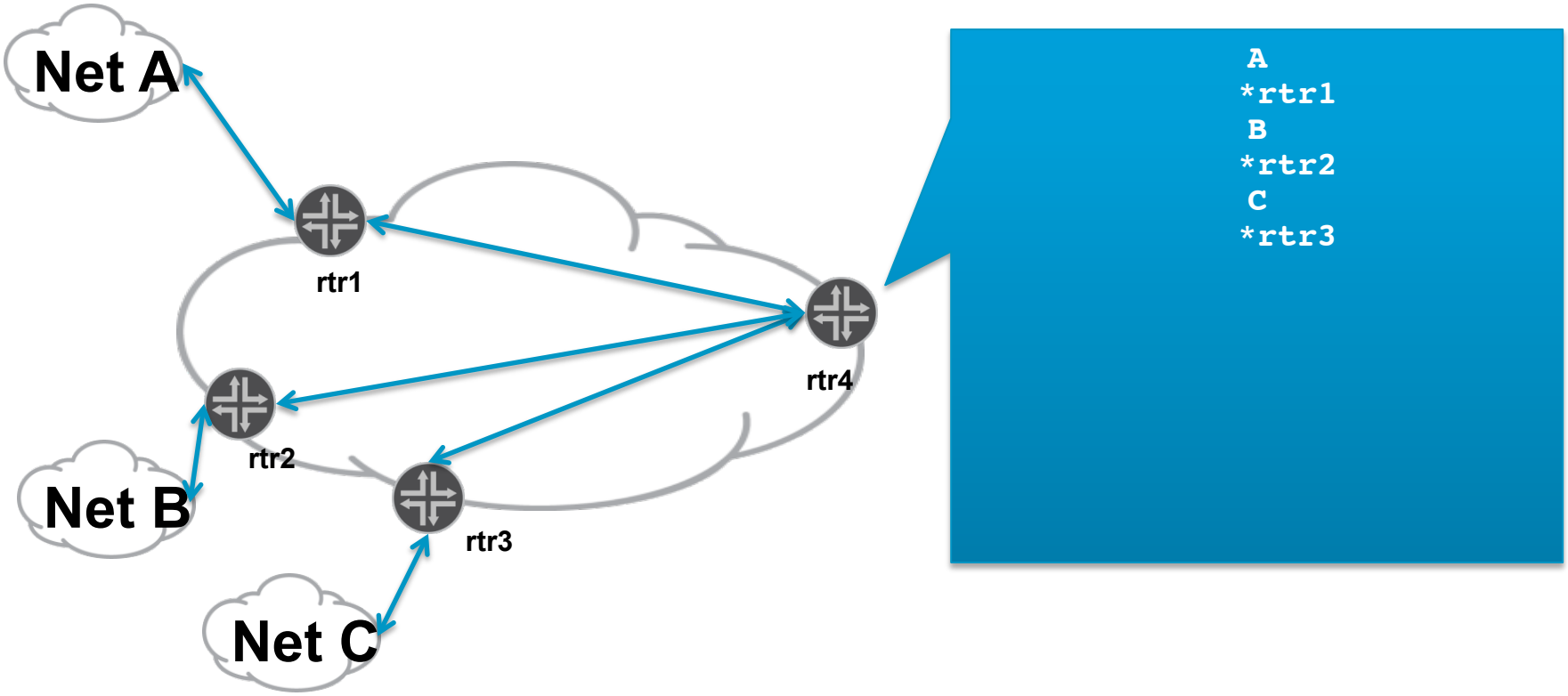
Internal BGP – Infrastruktura krytyczna

- IBGP nie odpowiada za wybór drogi wewnątrz AS. Odpowiada za wybór wyjścia z AS dla danej sieci docelowej.
- Infrastruktura krytyczna
  - Skalowalność
  - Elastyczność
  - Dostępność

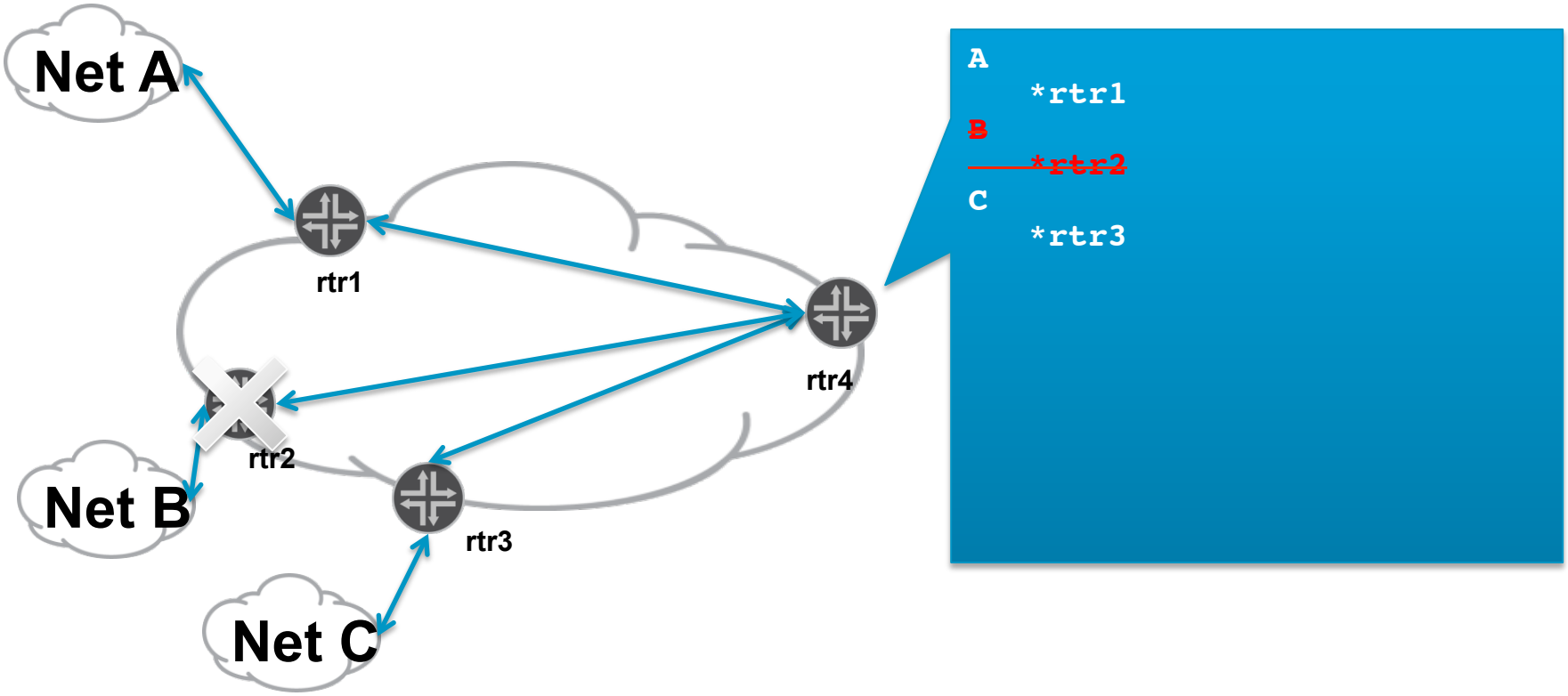
Architectura IBGP – Full MESH (1)

- IBGP full mesh
  - BGP pracuje w oparciu on TCP.
  - iBGP nie rozgłasza ścieżek pomiędzy sesjami wewnętrznymi.
- IBGP full mesh to GOOD THING
  - Rozproszone przetwarzanie – wysoka dostępność, minimalne zmiany.
  - Pełny obraz o dostępnych ścieżkach na każdym routerze – zawsze optymalny routing z punktu widzenie każdego routera z osobma.
- IBGP full mesh limits
  - Czynnnik Ludzki – nie lubimy pełnej kraty. Ludzie kochają hierarhię.
  - Ilość sesji TCP per router
    - Wyzwanie dla operatora
    - Wyzwanie da OS
  - Wiele ścieżek do tej samej sieci docelowej – tylko jeśli wiele wyjść.

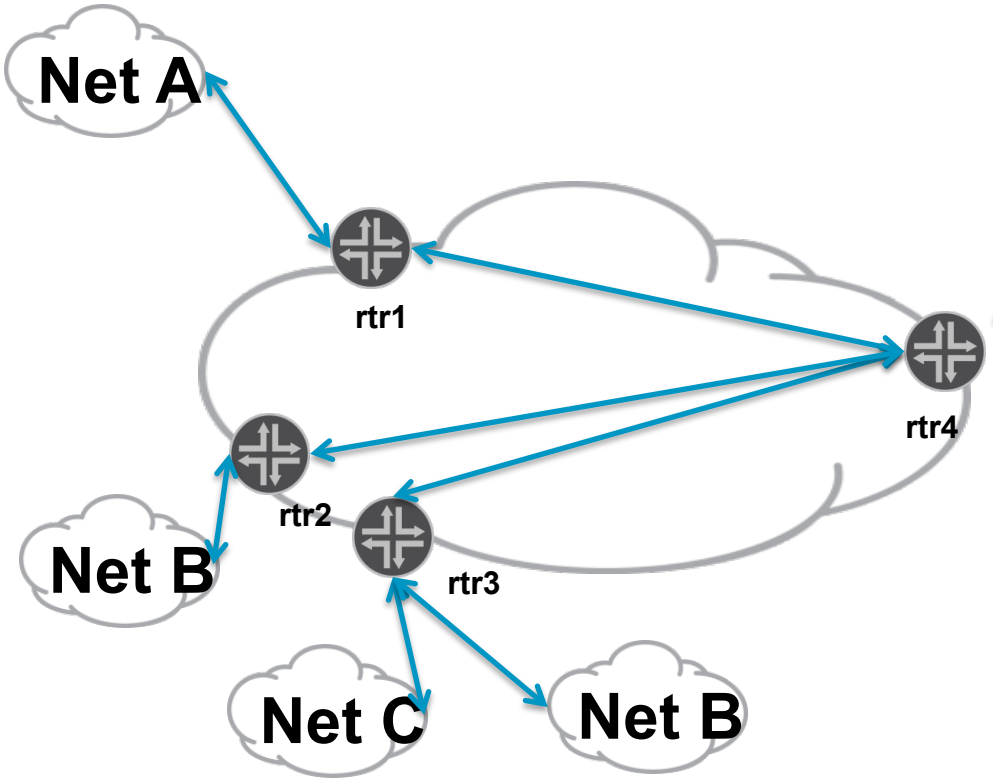
ArchiteKtura IBGP – Full MESH (2)



ArchiteKtura IBGP – Full MESH (2)



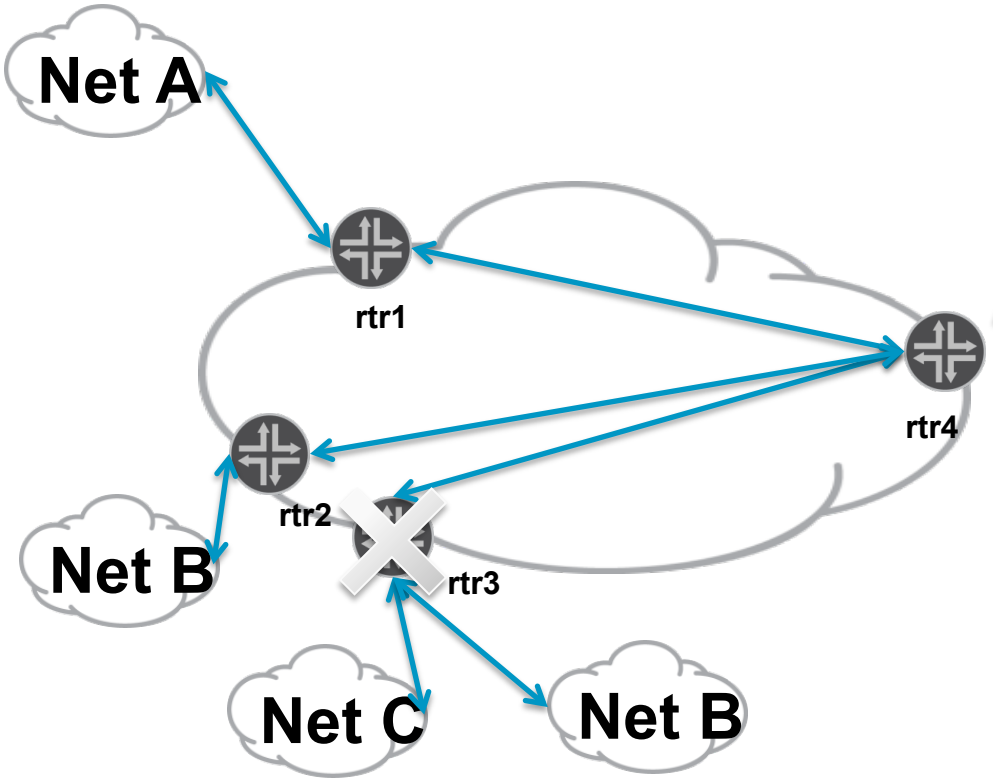
ArchiteKtura IBGP – Full MESH (2)




```
A
    *rtr1
B
    *rtr2
     rtr3
C
    *rtr3
```

- 3 prefixes
- 4 paths

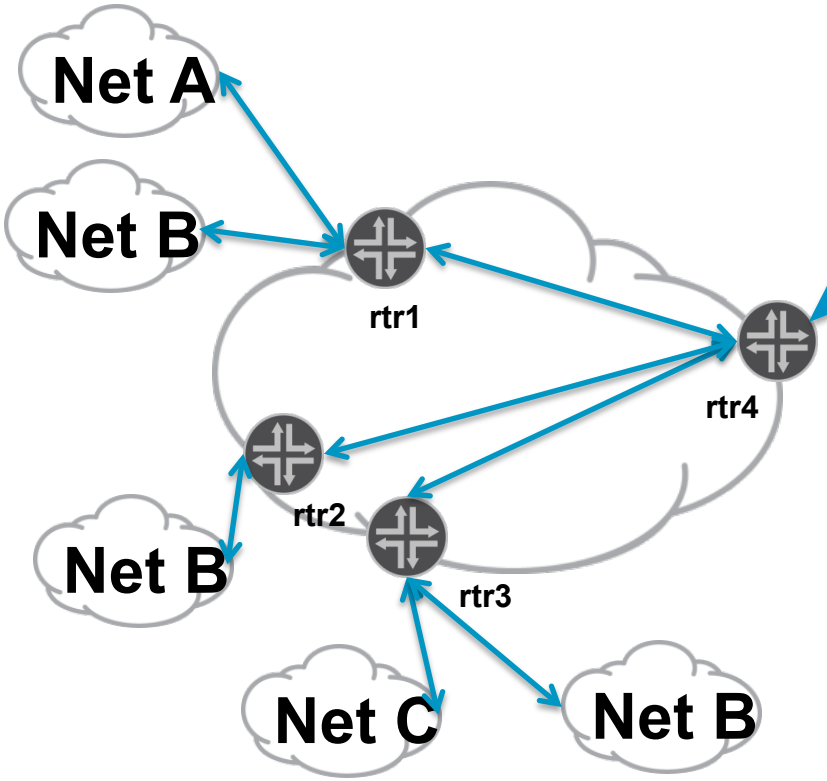
ArchiteKtura IBGP – Full MESH (2)




```
A
    *rtr1
B
    *rtr2
    *rtr3
C
    *rtr3
```

- 3 prefixes
- 4 paths

ArchiteKtura IBGP – Full MESH (2)




```
A
    *rtr1
B
    *rtr1 – lowest IGP cost to rtr1
     rtr2
     rtr3
C
    *rtr3
```

- 3 prefixes
- 5 paths

Architectura IBGP – Route Reflection

- IBGP RR

  iBGP RR rozgłasza ścieżeki pomiędzy 2 sesjami wewnętrznymi, o ile conajmniej jedna z nich należy do klastra.

  Dopisuje cluster ID do listy klastrów w NLRI – zapobieganie pentlom i rozszeżenie

  Originator ID – zapobieganie pentlom.

- Czynnnik ludzki jest zadowolony
- Wiele ścieżek do tej samej sieci docelowej.
- IBGP RR caveats

  Nieoptymalne trasy.

  Ilość sesji TCP na RR taka sama jak w Full Mesh

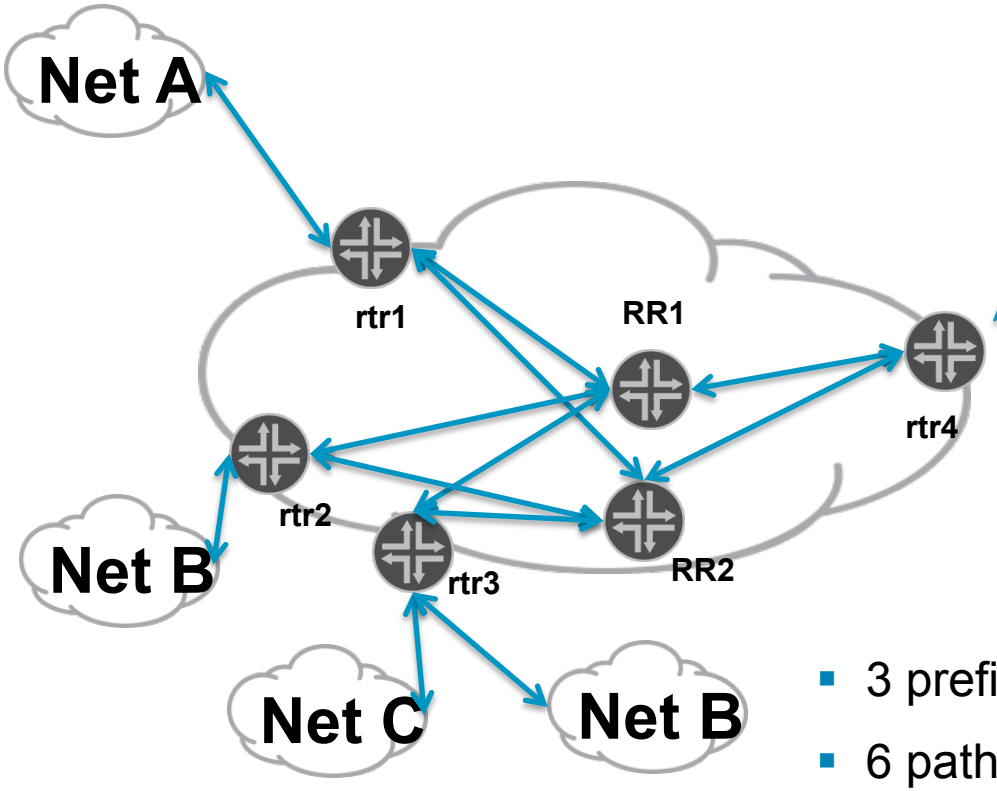
    Wyzwanie dla OS RR

    Wyzwania dla operatora

    Dużo więcej danych to wysłania !

  Opóźnienie – centralne przetwarzanie, dodatkowy element.

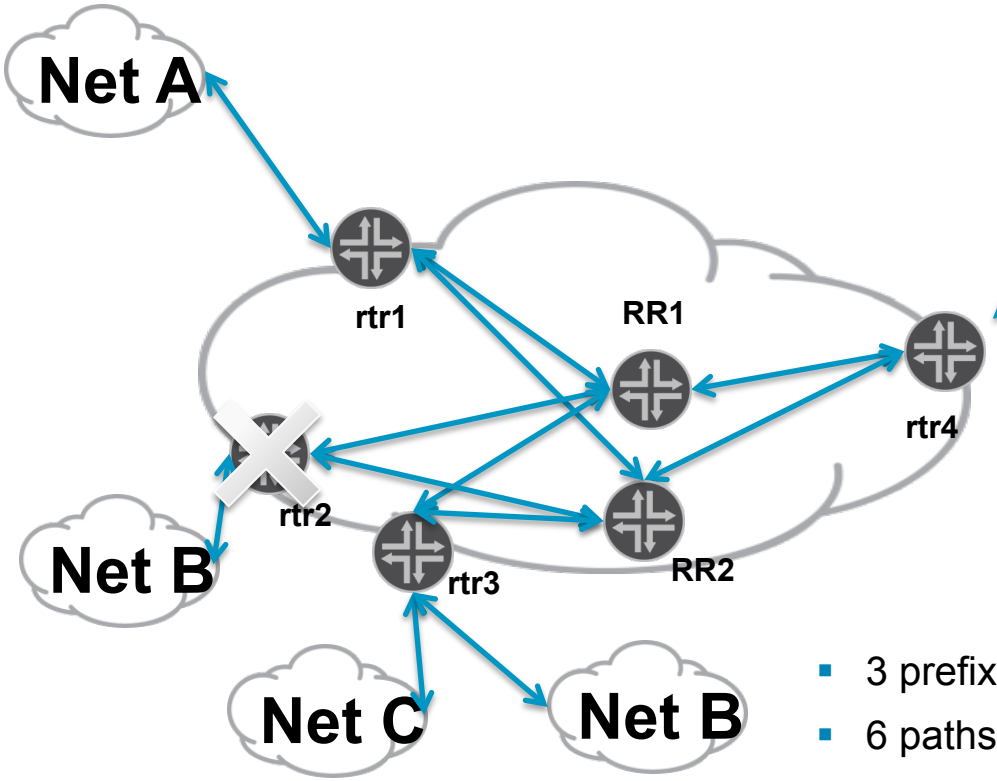
ArchiteKtura IBGP – RR




```
A
    *rtr1 – via RR1
     rtr1 – via RR2
B
    *rtr2 – via RR1
     rtr2 – via RR2
C
    *rtr3 – via RR1
     rtr3 – via RR2
```

- 3 prefixes
- 6 paths
- Path via RR1 preferred – lower RID.
- Lost of redundancy for prefix B in case of rtr2 fail.

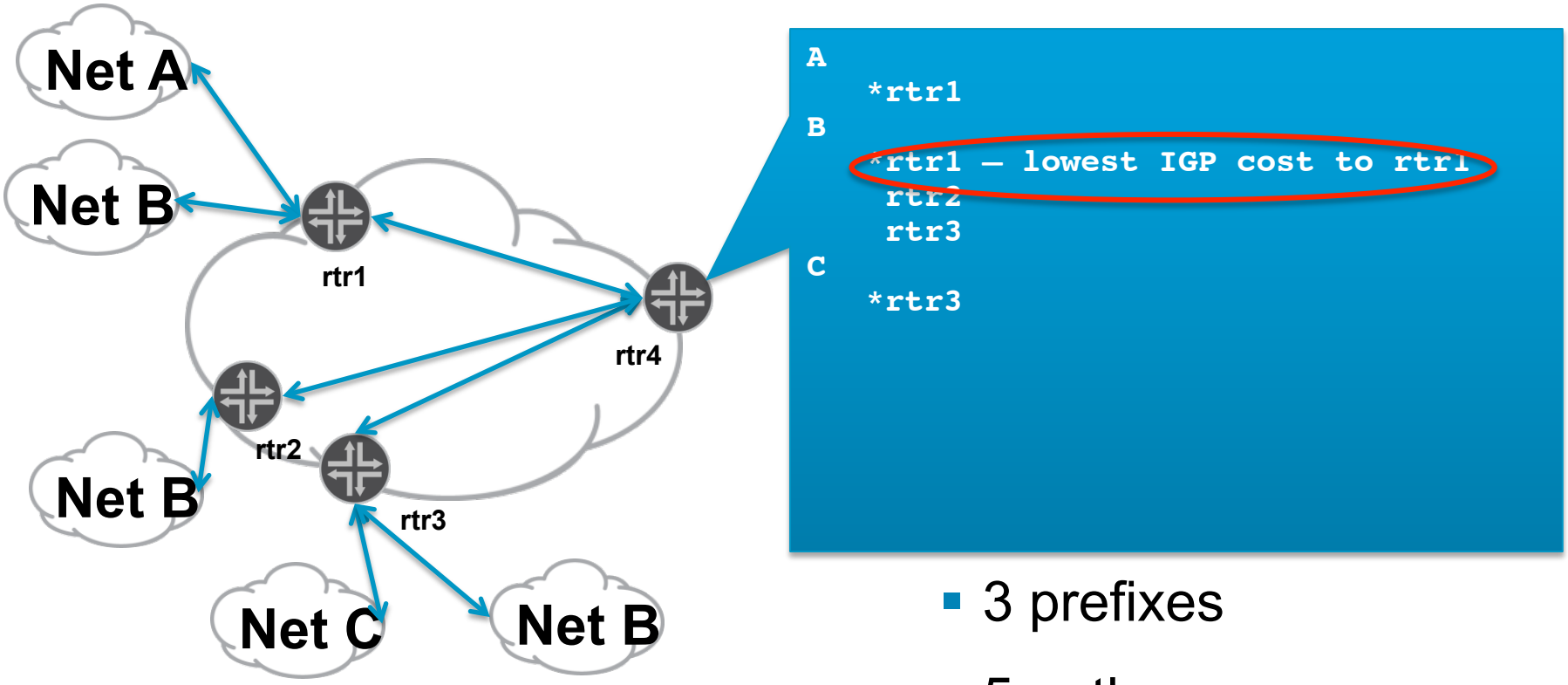
ArchiteKtura IBGP – RR




```
A
    *rtr1 – via RR1
     rtr1 – via RR2
B
    *rtr2 – via RR1
     rtr2 – via RR2
    *rtr3 – via RR1
     rtr3 – via RR2
C
    *rtr3 – via RR1
     rtr3 – via RR2
```

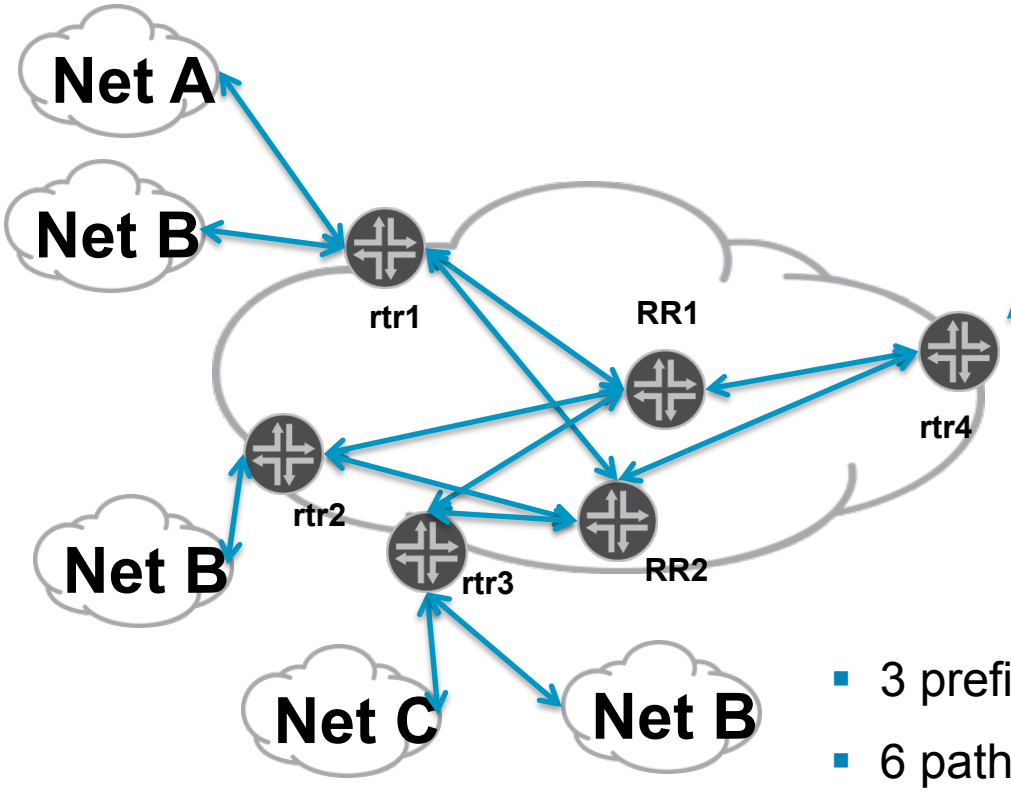
- 3 prefixes
- 6 paths
- Path via RR1 preferred – lower RID.
- Lost of redundancy for prefix B in case of rtr2 fail.
- WITHRAW and UPDATE from RR needed.

ArchiteKtura IBGP – Full MESH



- 3 prefixes
- 5 paths

ArchiteKtura IBGP – RR

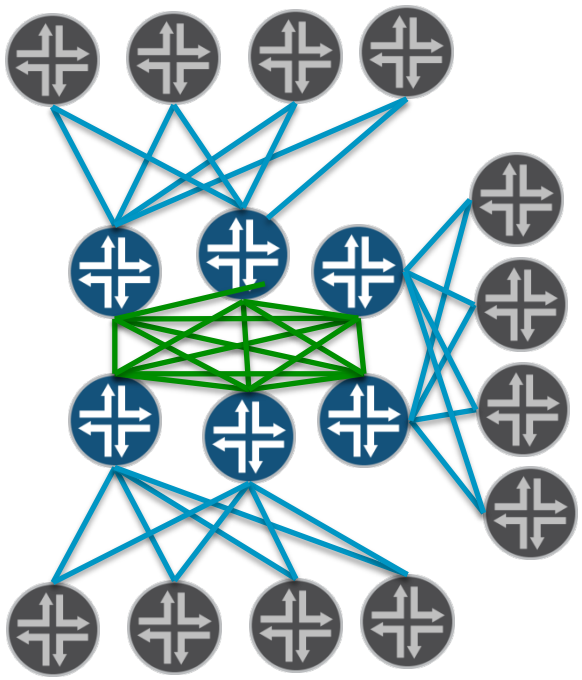



```
A
    *rtr1 – via RR1
     rtr1 – via RR2
B
    *rtr2 – via RR1
     rtr2 – via RR2
C
    *rtr3 – via RR1
     rtr3 – via RR2
```

- 3 prefixes
- 6 paths
- No path to Net B via closest exit – rtr1.
- rtr2 is closest exit form RRs PoV.

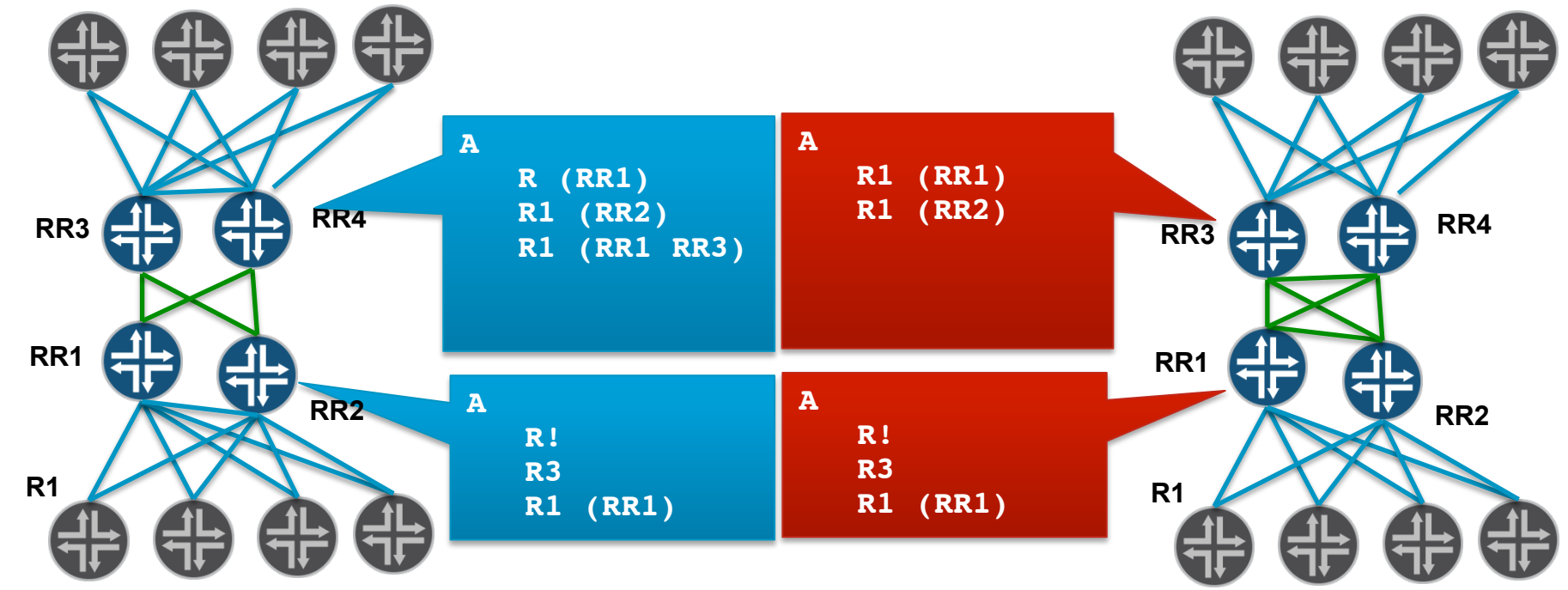
Projektowanie infrastruktury z RR

- Zapobieganie sub-optymalnemu routingowi.
  - Dla VPN – RT powinny być oparte na RID nie na ASN.
    - Więcej unikalnych prefixów VPNv4 – wymagania na pamięć i pasmo.
  - Regionalizacja
    - RR nadal wybierają najlepsze wyjscie z własnego punktu widzenia. Ten punkt jest bliżko punktu widzenia klientów.
    - iBGP full mesh między RR.
- Rozsądna ilość sesji na RR
  - Regionalizacja
  - ~ 50 per RR.
- Ile RR w Full Mesh ?
- Wzrost ilości scieżek na RR



Full mesh – 66 sesji, 11 per router
3*4*2 + 15 = 39 sesji, 9 per RR (-18%)
Full mesh – 276 sesji, 23 per router
6*4*2 + 66 = 114 sesji, 15 per RR (-35%)

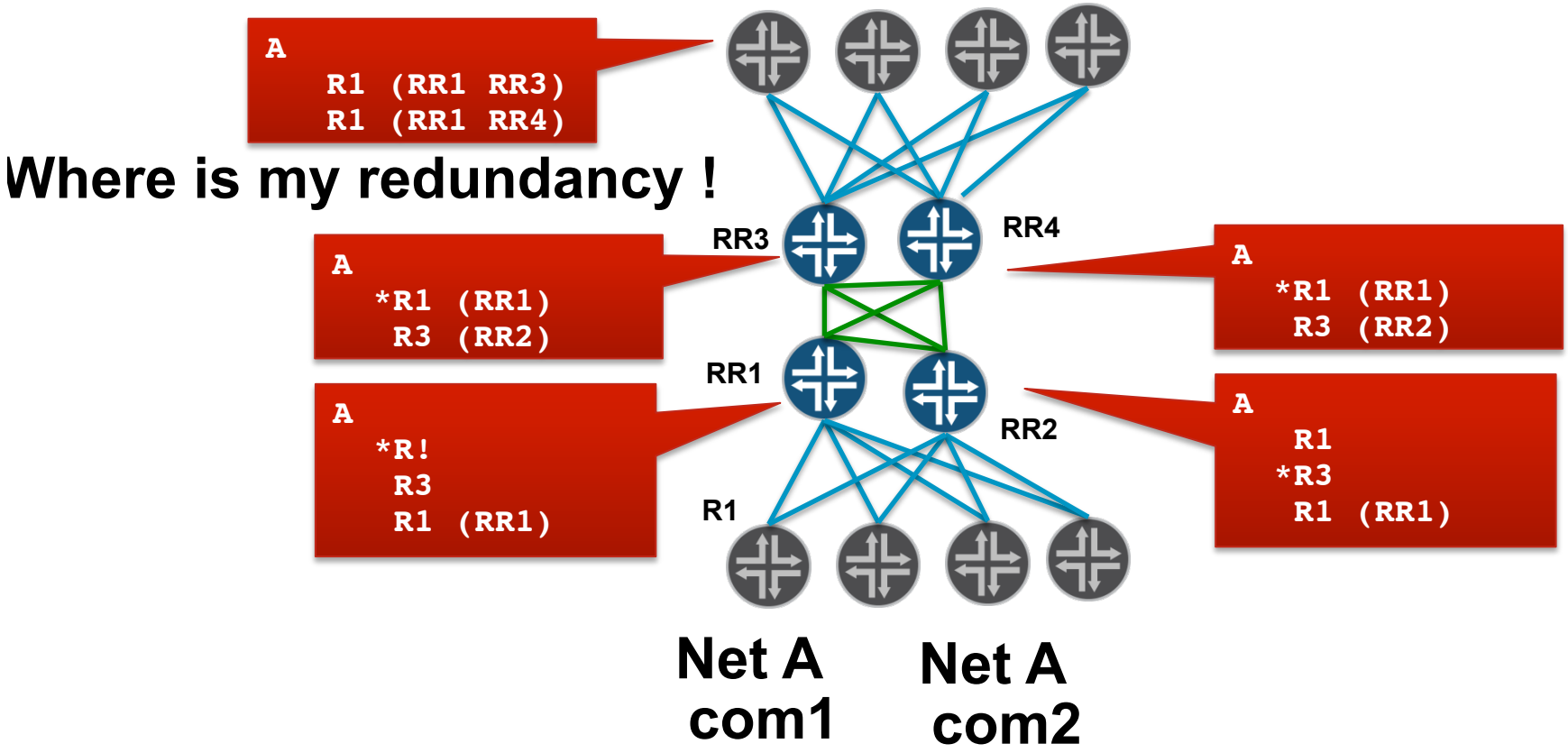
Ilość ścieżek na RR



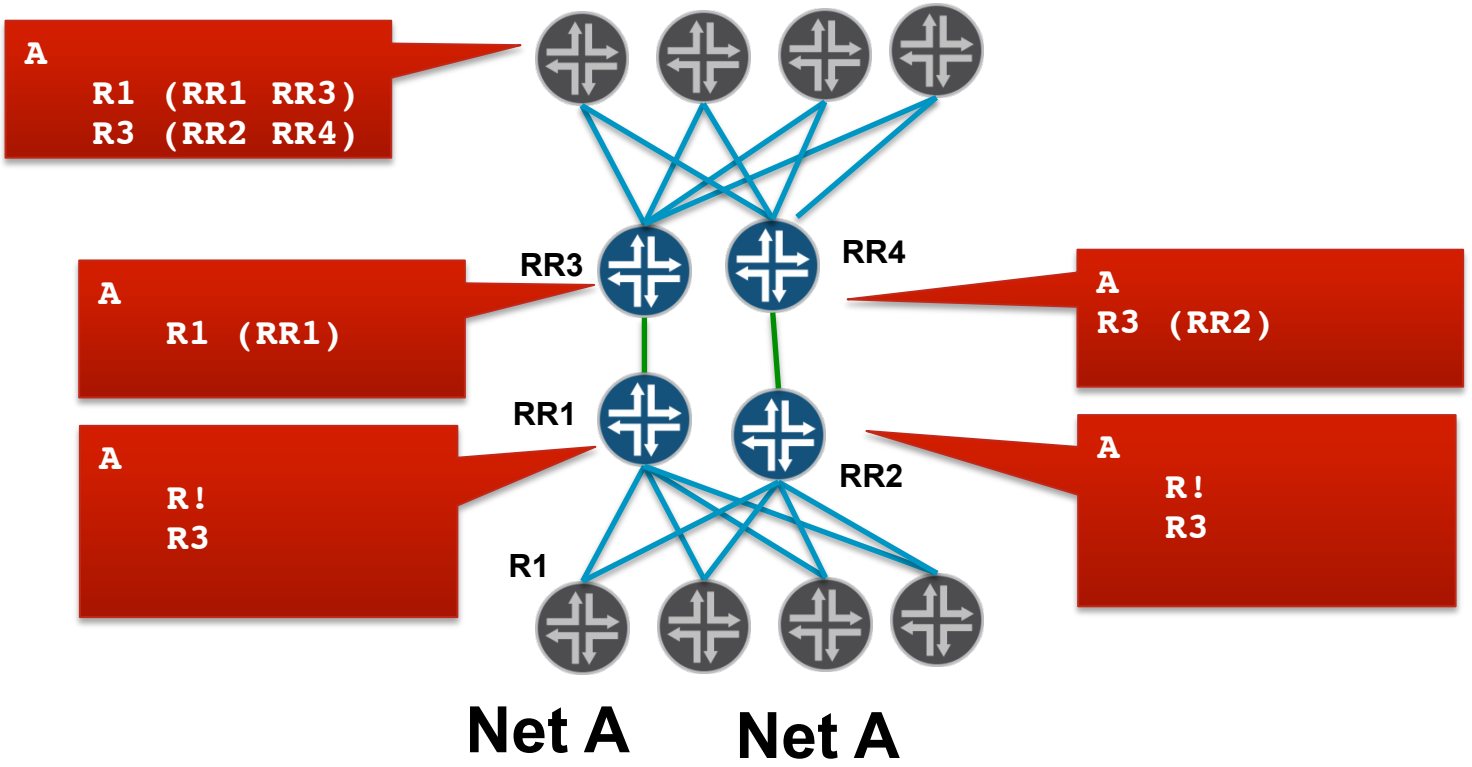
Net A **Net A**

- Shared Cluster-ID is another option

Exit redundancy w/ RR



Exit redundancy w/ RR

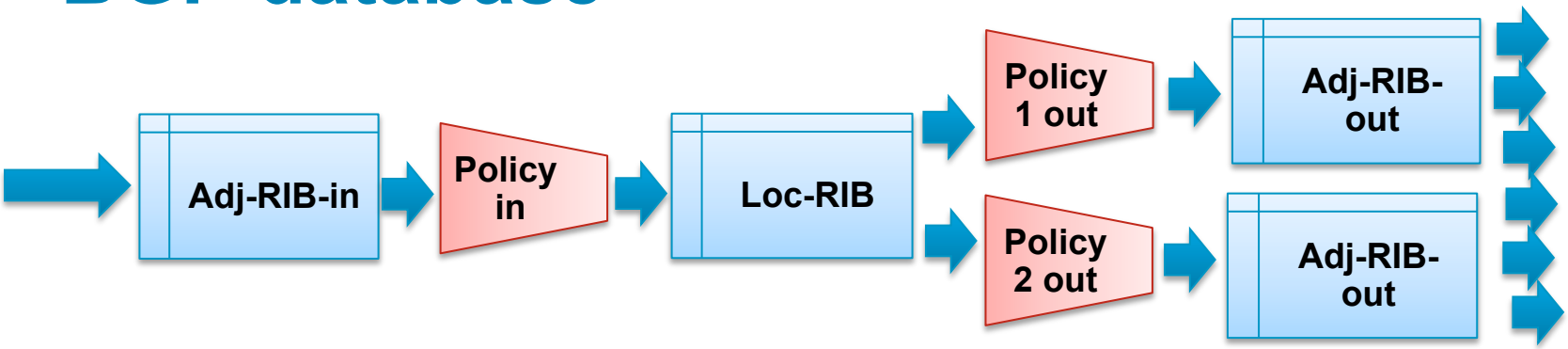


- Shared Cluster-Dual RR Plane

Dedicated Route Reflectors

- Dylemat
  - Dedykowany RR – nie przełącza pakietów, zajmuje się wyłącznie routingiem.
  - RR jako funkcja routera przełączającego ruch.
- RR jako funkcja routera COREowego (Internet-Free Core)
  - Brak prefixów internetowych na routerach COREowych – **bezpieczeństwo** (i wydajność).
  - Bezpieczeństwo
    - Zablokowanie zapisu dróg sieci nauczonych z BGP do FIBa – router nie wyśle pakietu poza AS.
    - Wykożystanie wirtualizacji.
  - Ekonomiczne rozwiązanie, ograniczona wydajność, konkuruje o zasoby control plane z innymi zadaniami.
- RR jako funkcja routera brzegowego
  - UWAGA na NEXT-HOP-SELF
  - Nie można odciąć od internetu – router brzegowy.
  - Ryzykowne – skuteczny atak może zainfekować WSZYSTKIE routery BGP.
  - Ekonomiczne rozwiązanie, ograniczona wydajność, konkuruje o zasoby control plane z innymi zadaniami.
- Dedykowany RR
  - Tylko jedna wada – Koszt
  - Dedykowany != scentralizowany.

BGP database



- RIBs may (and should) be implemented as single table w/ pointers and flags – memory consumption, less transfers.
- Adj-RIB-in - Keeps all recived PATHs, which pass sanity check. (e.g. AS loop)
- Changes in “Policy in” – no need for session clear or re-advertisment of NLRIs

  JUNOS – no-need: `set protocols bgp {group g-name {neighbour n-addr}} keep none`

  IOS need to enabe: `neighbour x.x.x.x soft-reconfiguration inbound`
- Loc-RIB paths elligable to be active, after policy actions.
- Adj-RIB-out – routes after export policy

  Per neighbour poliicy = Multiple policies – multiple RIBs – more Memory needed
- Group as many as possible peers with same egress policy – single Adj-RIB-out for all members.

  Less memory

  Less CPU

BGP refresh capability,

- If memory is constrain, free-up Adj-RIB-in

  JUNOS: `set protocols bgp {group g-name {neighbour n-addr}} keep none`

  IOS – no need – default.

- If “policy in” changes, BGP peer need to re-send all prefixes

  Slow process.

  Need to request peer for retransmission.

    BGP refresh capability.

    Negotiated on session start.

- Affects all address families

Securing infrastricture

- Cryptographic-based session authentication.
  - SHA is stronger then MD5, replace periodically
  - BGP over IPSec
- Prefix black-holing
  - Disables whole network/host.
  - Requires pre-configuration on all BGP routers (community, policy/route-map, route to null0)
- Flow Spec routes
  - Use BGP to program FF on remote routers.
    - Matches source, destination, protocol, ports, packet size, TCP flag, etc
    - Actions: discard, count, rate-limit, redirect to other routing instance, etc
  - No pre-configuration needed except enabling family in BGP.
  - Works across AS borders.

```
[protocol bgp ]
family inet {
    flow;
}
export flow-routes;

[routing-options flow]
route attack1 {
    match {
        destination 1.1.1.1/32;
        source 2.2.0.0/20;
        protocol TCP;
        destination-port 22;
        fragment is-fragment;
    }
    then {
        discard;
        sample;
    }
}
```

Ku Szybkiej Zbieżności

Zbieżność BGP

- BGP jest wolne
  - Tak bo jest na TCP
  - Tak bo ma złożone algorytmy i polityki
  - Tak bo zwykle przenosi wiele prefixów i ścieżek.
  - Tak bo RR wprowadzają dodatkowe opóźnienie
  - Jest dużo szybsze dziś
    - Nie stosuje się dampeninigu
    - IOS nie opiera się na periodycznym BGP scanner job – event driven (JunOS – event driven day one)
- Wiele można osiągnąć przez design.

BGP path selecton

- Higher local preference.
- If the path includes an AS path: Prefer the route with a shorter AS path.
- Prefer the route with the lower origin code.
- Depending on whether nondeterministic routing table path selection behavior is configured, there are two possible cases:
- ‚Prefer the path with the lowest multiple exit discriminator (MED) metric*.
- Prefer strictly external (EBGP) paths over external paths learned through interior sessions (IBGP).
- For BGP, prefer the path whose next hop is resolved through the IGP route with the lowest metric.
- 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.
- For BGP, prefer the path with the shortest cluster list length; length is 0 for no list.
- For BGP, prefer the path from the peer with the lowest peer IP address.

BGP PATH SELECTION

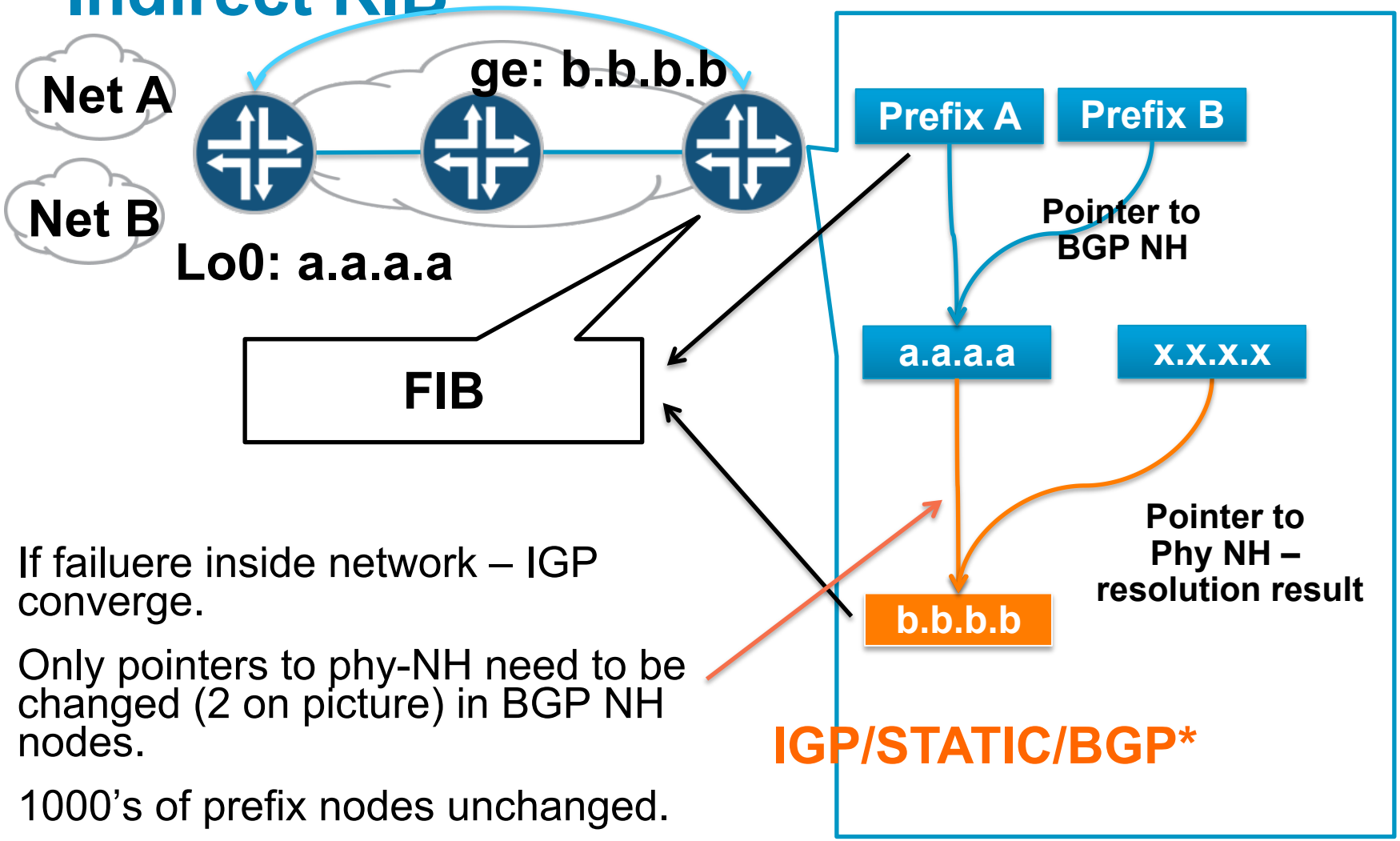
- The BGP Next-HOP has to be reachable
- Higher local preference.
- If the path includes an AS path: Prefer the route with a shorter AS path.
- Prefer the route with the lower origin code.
- Depending on whether nondeterministic routing table path selection behavior is configured, there are two possible cases:
- ,Prefer the path with the lowest multiple exit discriminator (MED) metric*.
- Prefer strictly external (EBGP) paths over external paths learned through interior sessions (IBGP).
- For BGP, prefer the path whose next hop is resolved through the IGP route with the lowest metric.
- 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.
- For BGP, prefer the path with the shortest cluster list length; length is 0 for no list.
- For BGP, prefer the path from the peer with the lowest peer IP address

NEXT HOP-TRACKING

- Jesli na ma drogi do BGP next-hop danj scieżki, to ta ścieżka jest ‘uszkodzona’ i nie można jej używać.
  - Proste jak dostajemy nowy update BGP – sprawdamy BGP NH
  - A co jeśli ścieżka BGP była OK, ale ostatni wypadł routing to BGP NH? – musimy ją unieważnić.
    - Okresowe przeglądanie
    - NH-tracking – wyzwalane przez zbieżność IGP.
- Droga do BGP NH musi spełniać pewne warunki:
  - Jeśli jest to droga BGP, to maska > od maski prefixu rozwiązywanego.
  - IPv4 – może być plane IP (AFI=1 SAFI=1), lub labeled IPv4 (AFI=1 SAFI=4)
  - L3VPN, 6PE, VPLS – musi być labeled IPv4
- Problem – routing domyślny. BGP NH zawsze dostępny!
- Potityka które drogi mogą być wykożystywane do rozwiązania BGP NH.

```
policy-statement IGP-host-routes {
    term 1 {
        from {
            protocol isis;
            route-filter 0.0.0.0/0
prefix-length-range /32-/32;
        }
        then accept;
    }
    term 2 {
        from rib inet.3;
        then accept;
    }
    term last {
        then reject;
    }
}
[routing-options resolution ]
rib inet.0 {
    resolution-ribs [ inet.3 inet.
0 ];
    import IGP-host-routes;
}
```

Indirect RIB

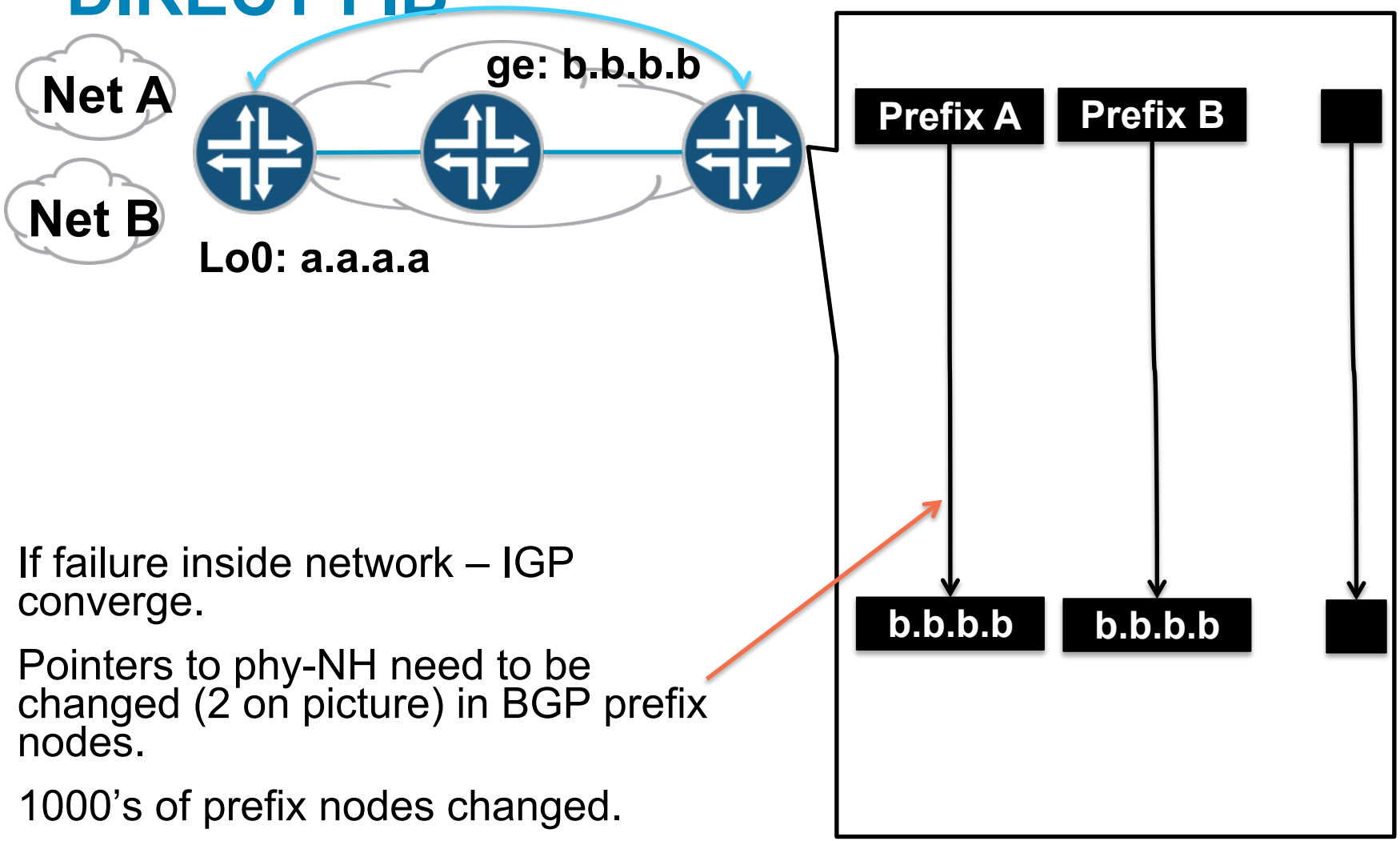


If failuere inside network – IGP converge.

Only pointers to phy-NH need to be changed (2 on picture) in BGP NH nodes.

1000's of prefix nodes unchanged.

DIRECT FIB

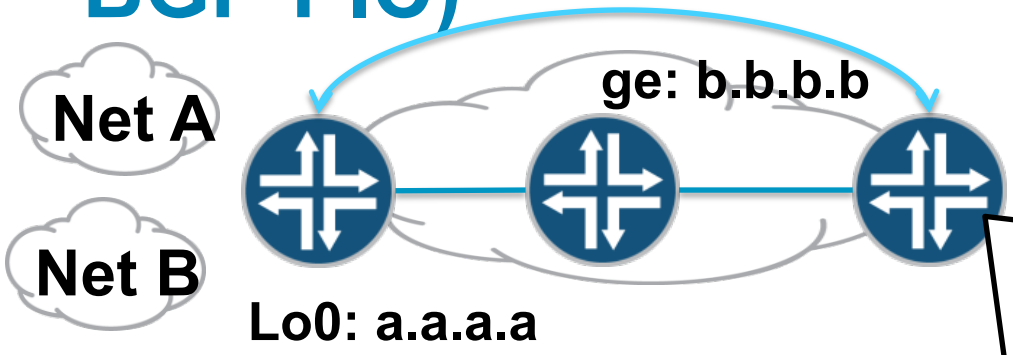


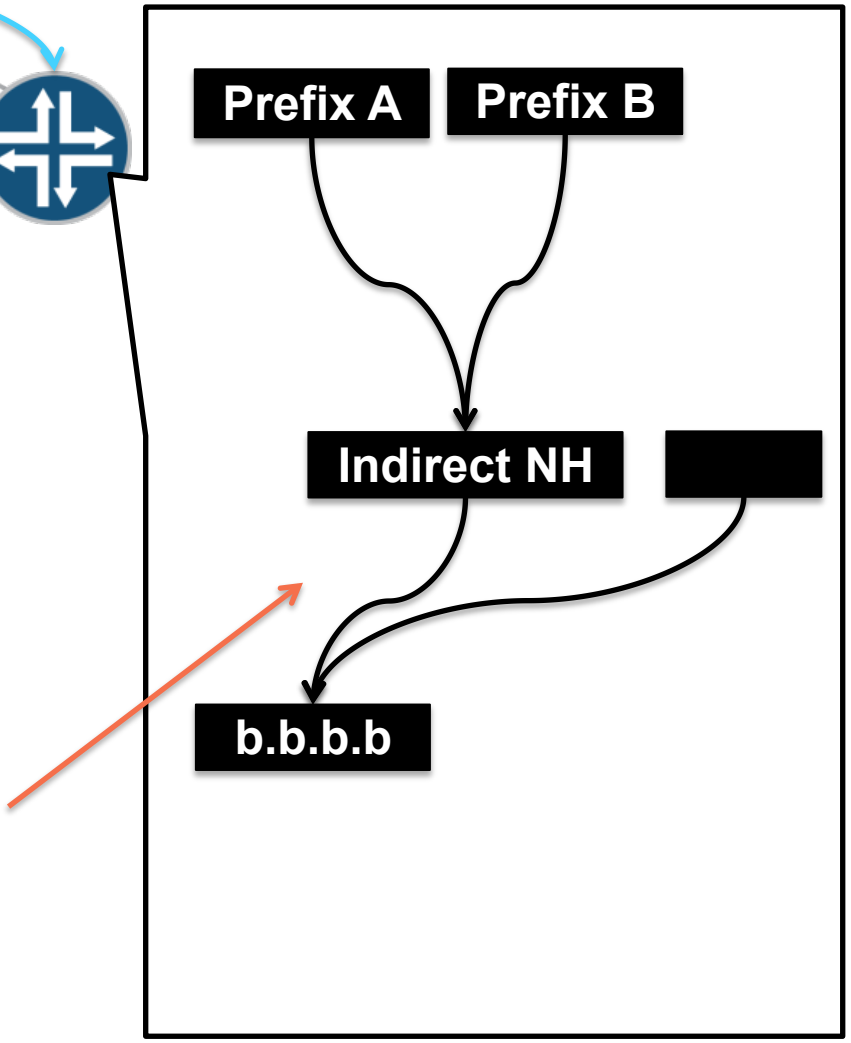
If failure inside network – IGP converge.

Pointers to phy-NH need to be changed (2 on picture) in BGP prefix nodes.

1000's of prefix nodes changed.

INDIRECT FIB (a.k.a Hierarchical FIB, BGP PIC)





If failure inside network – IGP converge.

Pointers to phy-NH need to be changed (2 on picture) in indirect nodes.

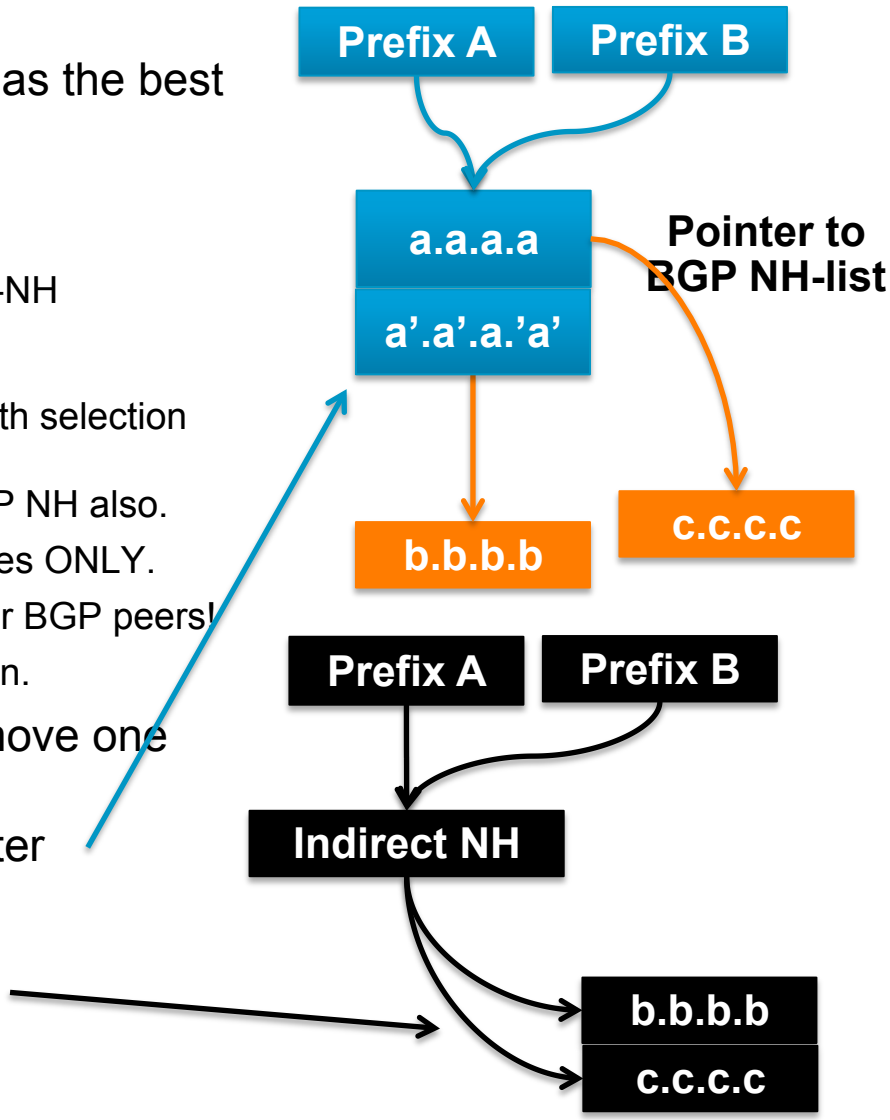
1000's of prefix nodes unchanged.

BGP PATH SELECTION

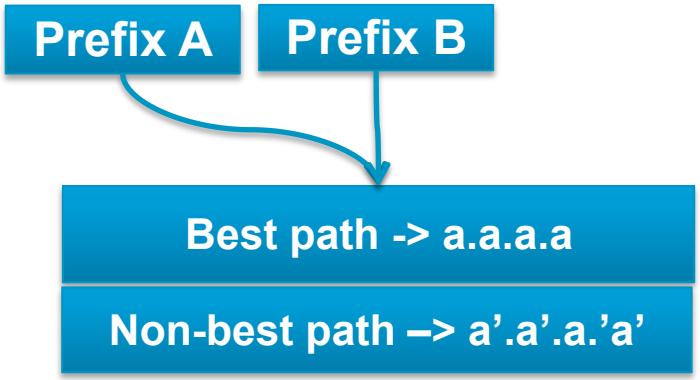
- The BGP Next-HOP has to be reachable
- Higher local preference.
- If the path includes an AS path: Prefer the route with a shorter AS path.
- Prefer the route with the lower origin code.
- Depending on whether nondeterministic routing table path selection behavior is configured, there are two possible cases:
- ,Prefer the path with the lowest multiple exit discriminator (MED) metric*.
- Prefer strictly external (EBGP) paths over external paths learned through interior sessions (IBGP).
- For BGP, prefer the path whose next hop is resolved through the IGP route with the lowest metric.
- 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.
- For BGP, prefer the path with the shortest cluster list length; length is 0 for no list.
- For BGP, prefer the path from the peer with the lowest peer IP address

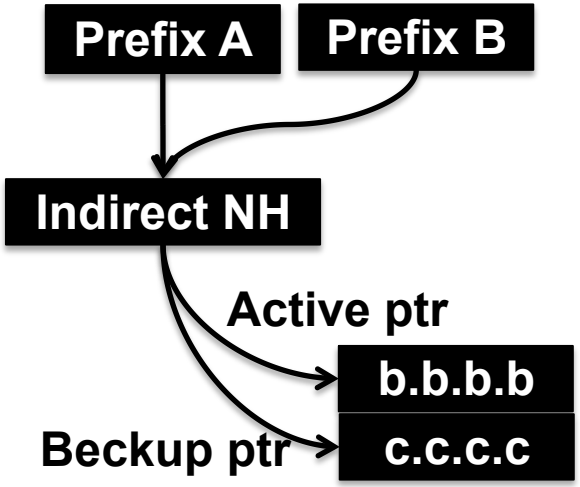
BGP Multipath

- BGP selects one and only one path as the best one.
  - Only one BGP NH for given prefix.
  - No BGP Load Balancing (per flow)
  - However resolution may result in 2 phy-NH
- BGP Multipath
  - removes RID and peer ID steps rom path selection process
  - Optionally may remove IGP cost to BGP NH also.
  - Multiple BGP NH for forwarding purposes ONLY.
  - Still only one path is announced to other BGP peers!
    - AD-PATH attribute is/will be solution.
- In case of iBGP peer failure, we remove one enty on BGP NH-list (BGP track)
- In case link failure remove one pointer



BGP FRR





- Usabel in E-BGP enviroment. – both BGP NH need to be directly connected.
- In IBGP Non-Best path may lead to loops.

Pytania

Dziękuję