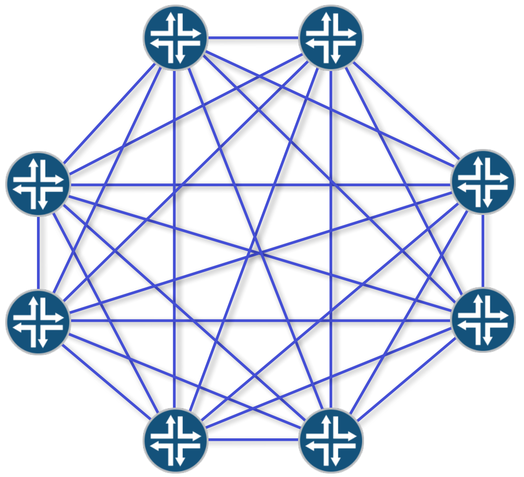
Boas práticas no uso de route reflectors

Luis Balbinot <lbalbinot@br.digital>

IX Fórum Regional Porto Alegre - 10 de março de 2023

O que são e por quê usamos route reflectors?

- São componentes de uma malha iBGP que replicam ou refletem rotas de seus clientes
- Clientes são agrupados em grupos, chamados de *clusters* e identificados por um *Cluster-ID*
- Para evitar loops um refletor descarta rotas que tenham seu *Cluster-ID* no caminho
- Reflexão de rotas segue algumas regras específicas
  - Rotas recebidas de clientes do RR são anunciadas para os clientes do RR e não clientes do RR
  - Rotas recebidas de não clientes do RR são anunciadas apenas para clientes do RR
  - Rotas recebidas de um vizinho eBGP são anunciadas para clientes do RR e não clientes do RR
  - Se existem múltiplas rotas para um mesmo destino, apenas a melhor (segundo o RR) é refletida
  - O RR não mode alterar atributos das mensagens, incluindo o endereço de next-hop

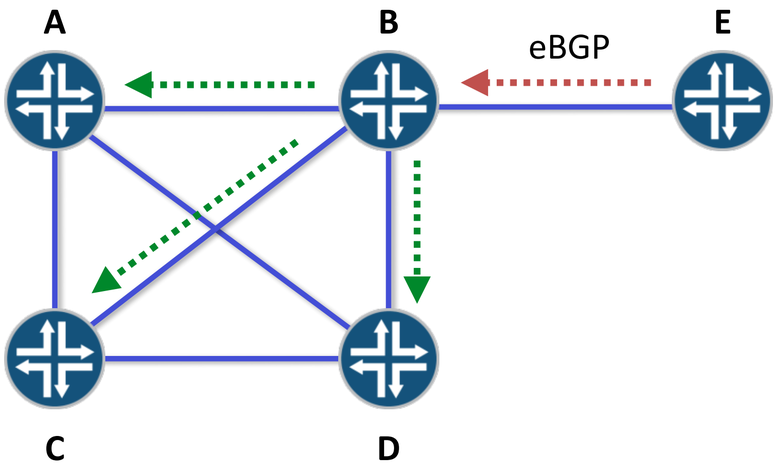


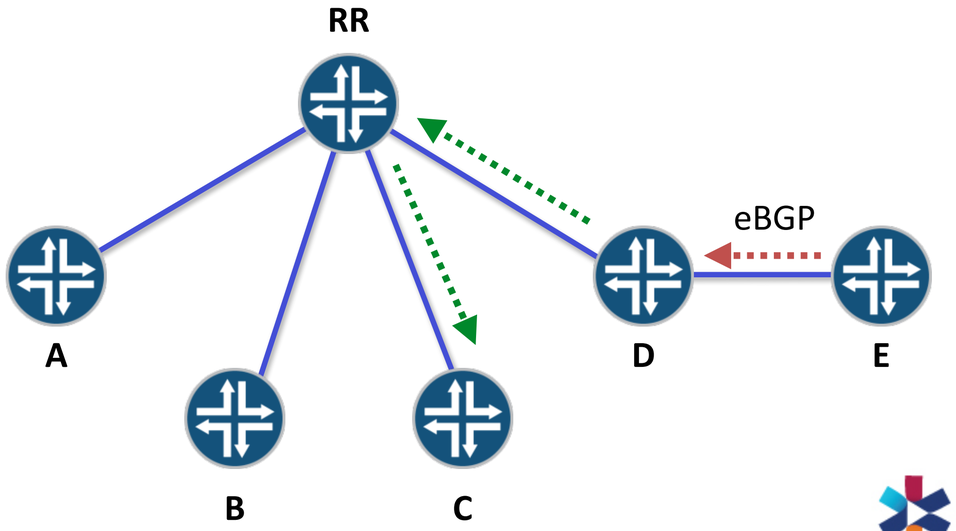
N(N-1)/2

- Objetivo é reduzir a quantidade de sessões iBGP e o tamanho da RIB de adjacências (Adj-RIB-In)
- Reduzir custo operacional da rede
- Reduzir a quantidade de UPDATES que são trocados (menos gasto de CPU e memória nos roteadores), deixando o processamento pesado aos RRs, que podem ser dedicados

Cuidados com o uso de RRs

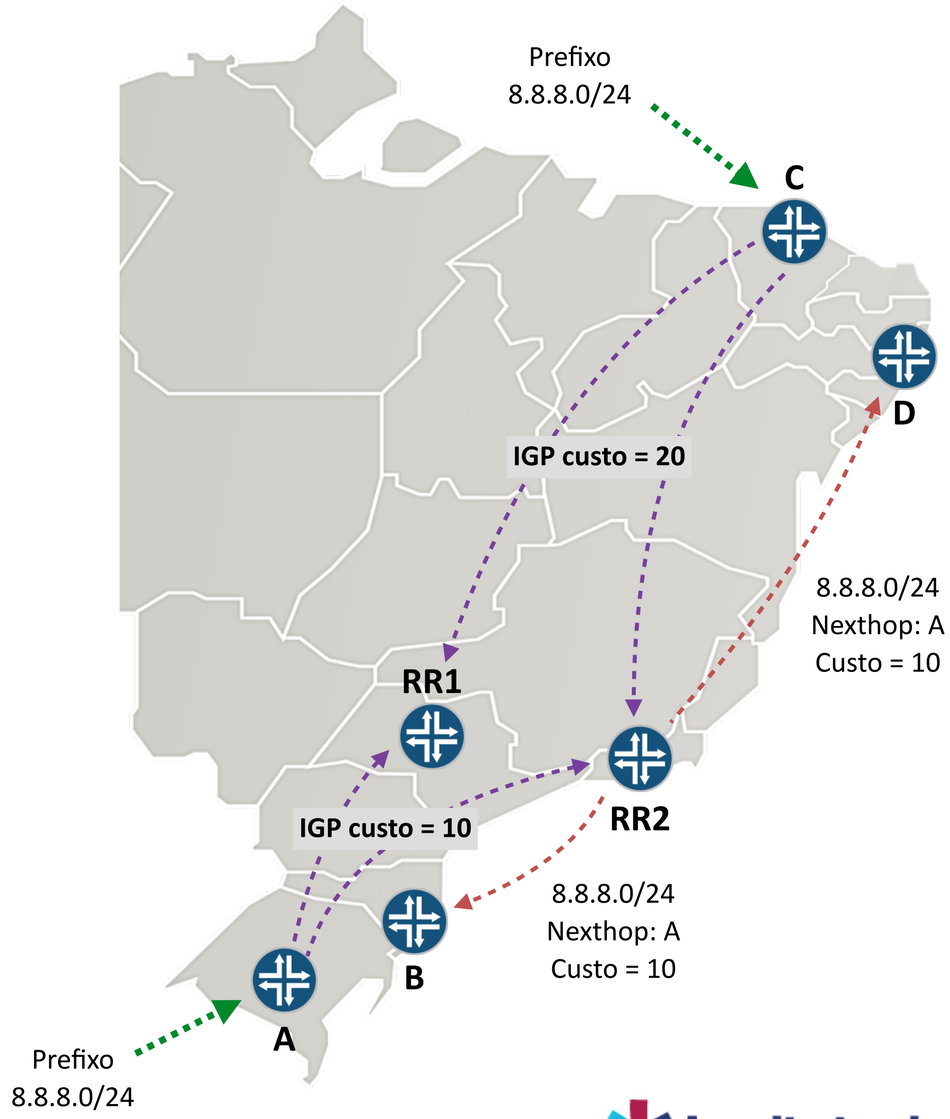
- **Aumento do tempo de convergência**
  - Em uma topologia full-mesh um UPDATE vai direto ao vizinho
  - Com RRs é preciso passar por um ou mais RRs
  - Há um aumento de tempo não só pelo fluxo de pacotes, mas também pelo tempo de processamento em cada estágio





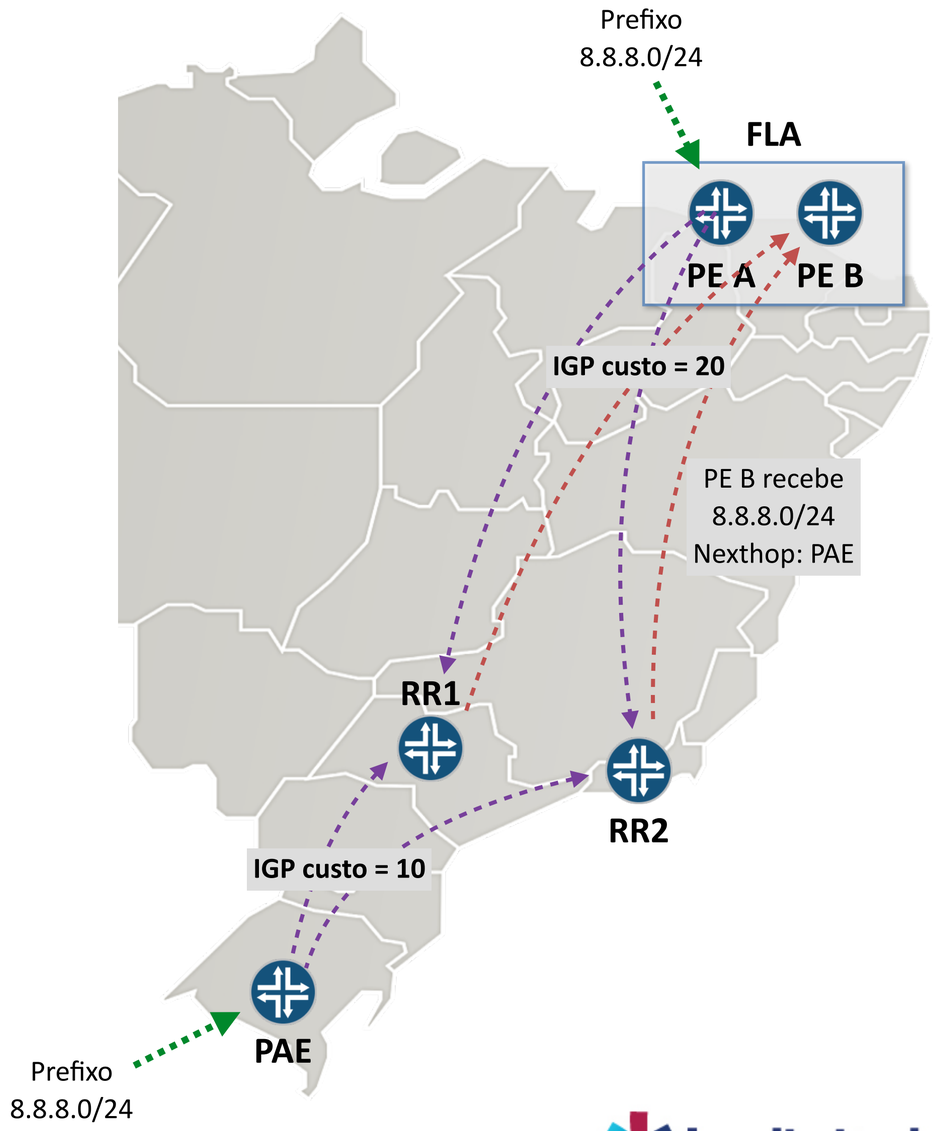
Cuidados com o uso de RRs

- **Roteamento não otimizado entre POPs**
  - O RR seleciona o melhor caminho baseado em suas informações locais de roteamento e o anuncia para seus clientes, o que pode resultar na escolha de um pior caminho em alguns casos
  - No exemplo, o prefixo 8.8.8.0/24 é aprendido pelos roteadores A (Porto Alegre) e C (Fortaleza), mas como os RRs estão em São Paulo e no Rio eles vão escolher a rota vinda de Porto Alegre por causa do custo menor do IGP e irão anunciá-la para todos os clientes
  - Por consequência, os clientes de Recife (D) ao invés de irem direto para Fortaleza (12ms), vão descer até por Porto Alegre (58ms)



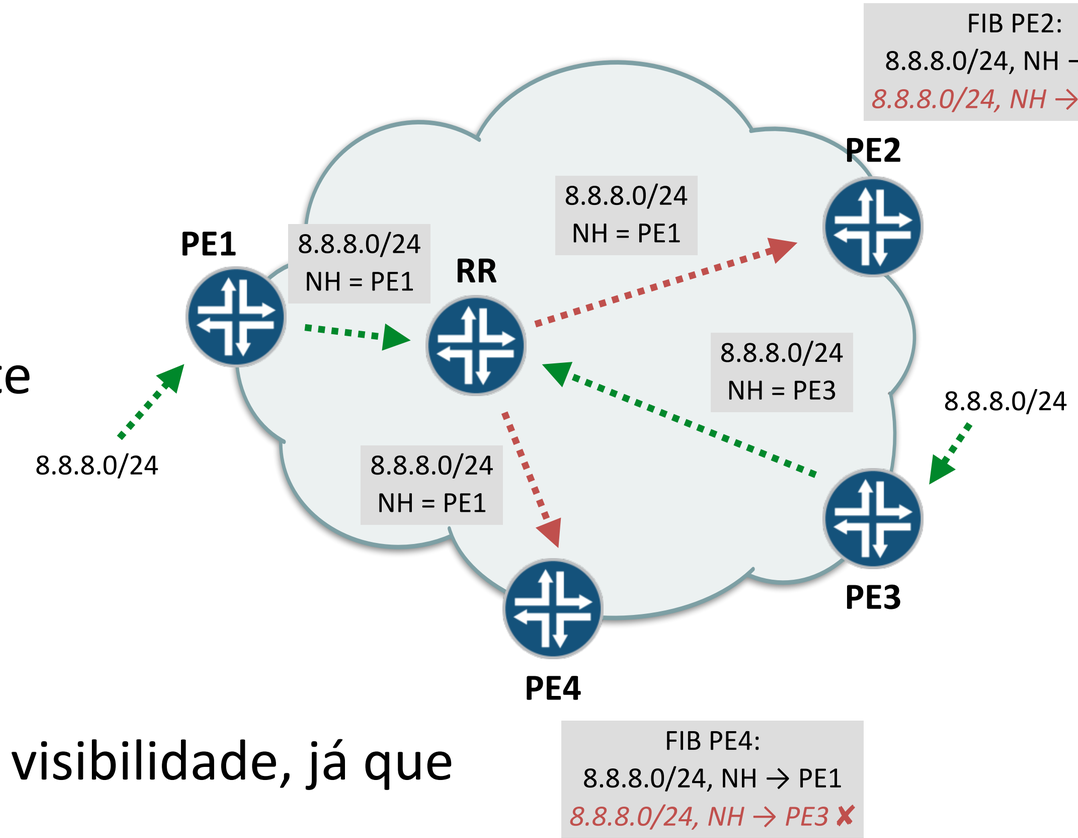
Cuidados com o uso de RRs

- **Roteamento não otimizado dentro do POP**
  - Ocorre quando dois roteadores dentro do mesmo POP acabam usando caminhos piores entre si
  - Em Fortaleza o PE A recebe o prefixo 8.8.8.0/24, mas dentro do mesmo POP o PE B que também é cliente dos RRs vai receber o prefixo escolhido por eles com o menor custo (10 via PAE), enquanto que poderia usar um next-hop local para o mesmo destino
  - Isso ressalta o quão importante é posicionar os RRs próximos de seus clientes de reflexão



Cuidados com o uso de RRs

- **Menor diversidade de caminhos**
  - Para mais rápida convergência costuma-se pré-calcular um caminho de backup
  - Assim com temos mecanismos de Fast ReRoute para MPLS e IP também é possível fazer o mesmo com prefixos BGP (BGP PIC Edge)
  - Isso é simples quando temos um full-mesh e visibilidade de todos os caminhos
  - Com a introdução dos RRs nós perdemos essa visibilidade, já que os RRs refletem apenas um caminho
  - Também se perde informações para usar BGP multipath, inviabilizando o balanceamento



Boas práticas

- **Topologia de referência**
  - Provedor de médio porte, com abrangência nacional, dividido em quatro regiões (sul, sudeste, nordeste e centro-oeste)
  - Serviços de Internet e VPN (L2VPN e L3VPN)
  - 20 POPs de roteamento, com quatro PEs em cada
  - Backbone IP/MPLS sem BGP no núcleo
  - IPv6 via 6PE
  - Flowspec

Distribuição geográfica dos RRs e clusters

- Cada POP com um par de RRs, implementados nos roteadores
- Em nossa topologia serão dois clusters em cada RR, um para serviços VPN e outro para Internet
  - Um único cluster facilita a operação, mas reduz a flexibilidade e aumenta o risco
  - Maior resiliência dos serviços em caso de falhas
  - Facilidade de crescimento de forma independente
    - PEs para VPN só participam dos clusters com as famílias de VPN, por exemplo
  - Impacto por modificações é reduzido

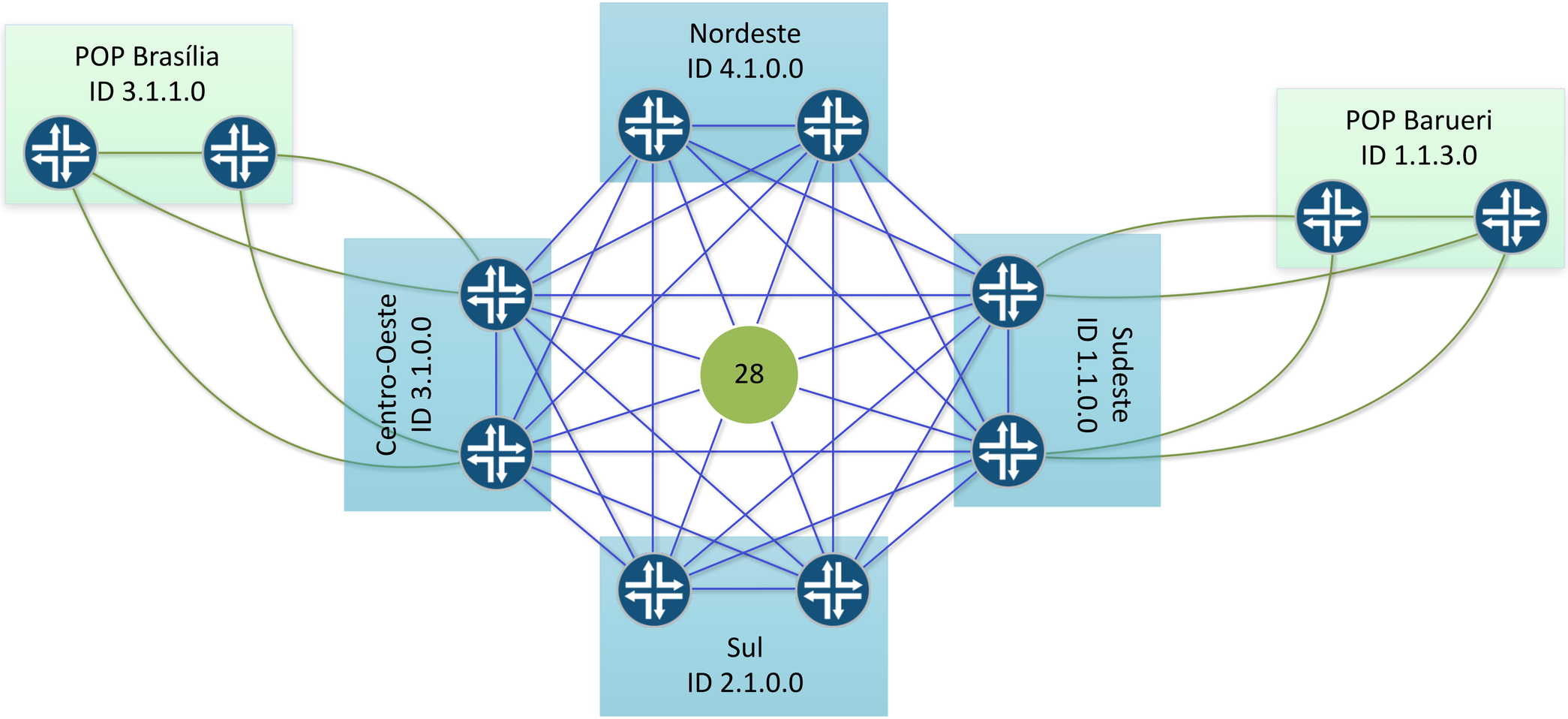
Pares de RRs com mesmo Cluster-ID ou não?

- O recomendado é utilizar o mesmo Cluster-ID em RRs redundantes
- Economiza memória e CPU nos RRs pois as rotas enviadas de um RR para o outro são descartadas
- Utilizar Cluster-IDs diferentes possui alguns benefícios em tempo de convergência, mas em topologias muito específicas fora do nosso escopo
- Com o mesmo Cluster-ID é obrigatório que todos os clientes sempre mantenham sessões com ambos RRs

Hierarquia de RRs

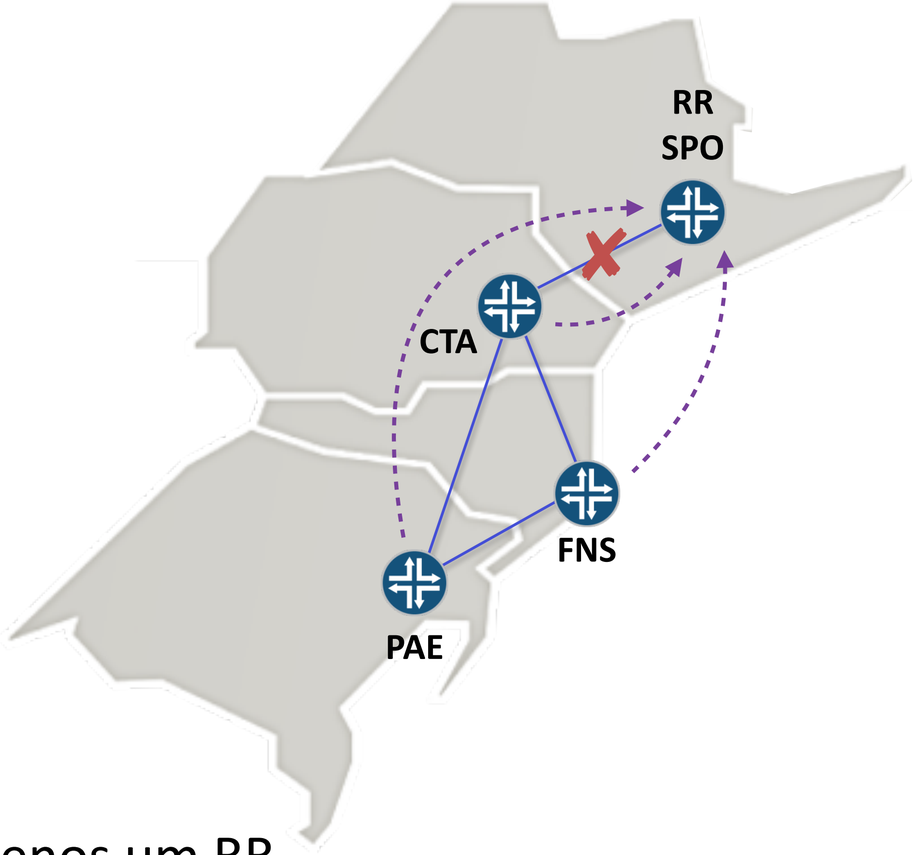
- Com 20 POPs e com dois RRs por POP o full-mesh iBGP fica grande
  - 40*39 / 2 = 780 sessões
  - 1560 sessões com dois clusters em cada RR
- Por isso criamos um segundo nível de hierarquia entre os RRs, com um novo par de RRs em cada uma das nossas quatro regiões
  - Full-mesh entre os 8 RRs das regiões fica com apenas 28 sessões iBGP
  - Simples de automatizar e gerenciar
- Cada RR dos POPs se liga com os dois RRs da sua região
- Em POPs muito pequenos pode-se fazer um full-mesh e ligar direto aos RRs regionais
- Dentro do POP dos RR regionais eles podem operar como RRs do POP também
- Em certos casos um POP pode se ligar em duas regiões ao mesmo tempo

Hierarquia de RRs



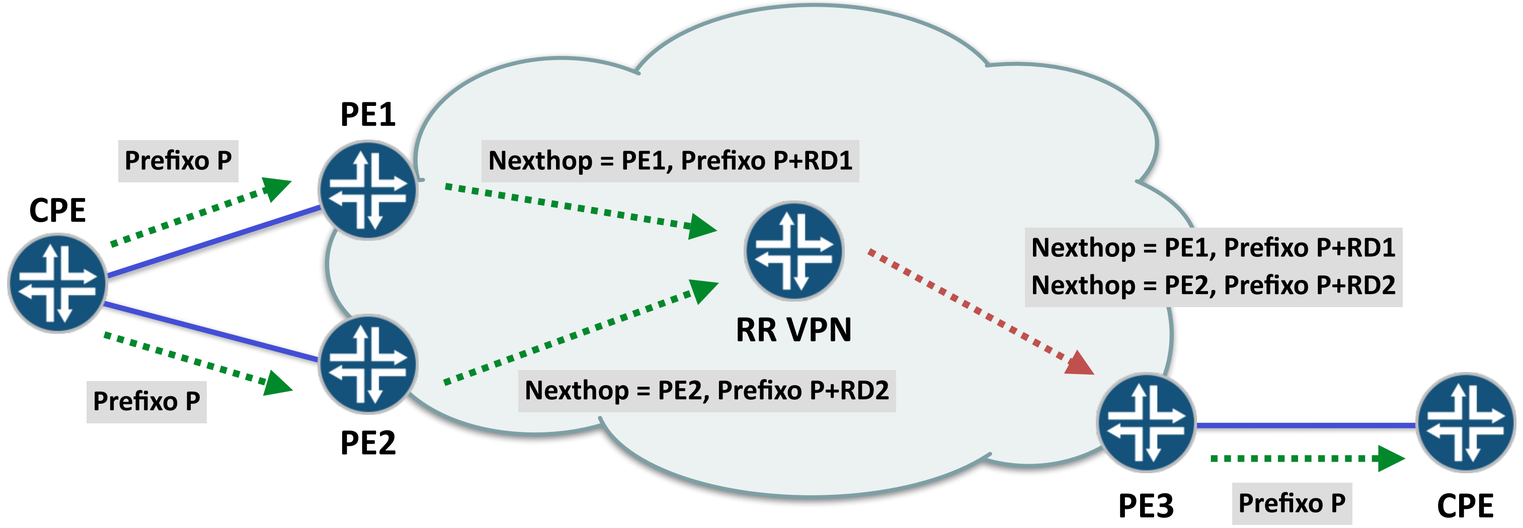
Posicionando o RR regional

- Não existe regra genérica, exceto a de que eles devem ficar próximos aos clientes e no caminho do tráfego
- É preciso ter amplo conhecimento do backbone para decidir
  - Saber quais enlaces são confiáveis
  - Quais são os caminhos principais e a direção do tráfego
  - Mapear enlaces que compartilham infraestrutura (*shared risk*)
- Distribuir de forma que eventos de falha no backbone não deixem regiões isoladas
- Em caso de isolamento deve-se garantir que os roteadores daquela região continuem trocando rotas entre si por pelo menos um RR
- Eventualmente pode-se mover ou até mesmo adicionar um terceiro RR em uma região problemática



Resolvendo a diversidade de caminhos

- No caso das VPNs é simples, basta usar um RD único por serviço
- Se o cliente está ligado em dois PEs e envia a mesma rota para ambos, cada uma terá um RD diferente e isso garante que sempre existirão duas rotas distintas na tabela, mesmo passando pelos RRs



Resolvendo a diversidade de caminhos

- Para tráfego Internet é um pouco mais complicado
- Uma opção seria colocar tráfego Internet em cima de VRFs, mas não é o nosso caso
- Uma solução é utilizar Add-Path (RFC 7911)
  - Conceito similar ao do RD das VPNs
  - Adiciona um Path ID único aos prefixos para diferenciá-los
  - Habilita o uso de mecanismos de convergência rápida
- Existem vários modos de operação que precisam ser considerados, dependendo de cada caso (Add-All, Add-N, etc)
- No nosso caso o modo Add-N é o mais adequado e resolve o problema da diversidade

Perguntas?

Luis Balbinot <lbalbinot@br.digital>