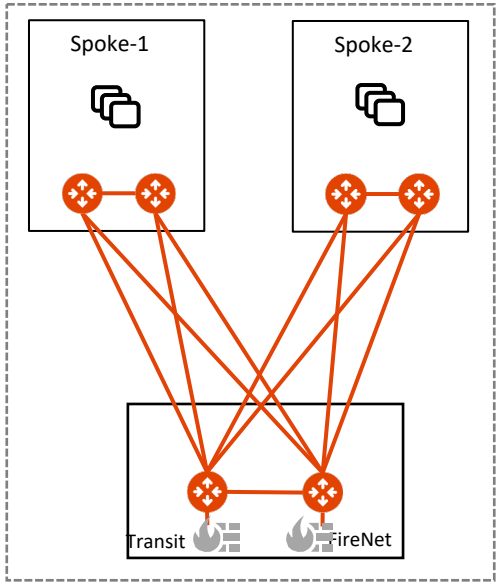
ActiveMesh

ACE Solutions Architecture Team

Overview

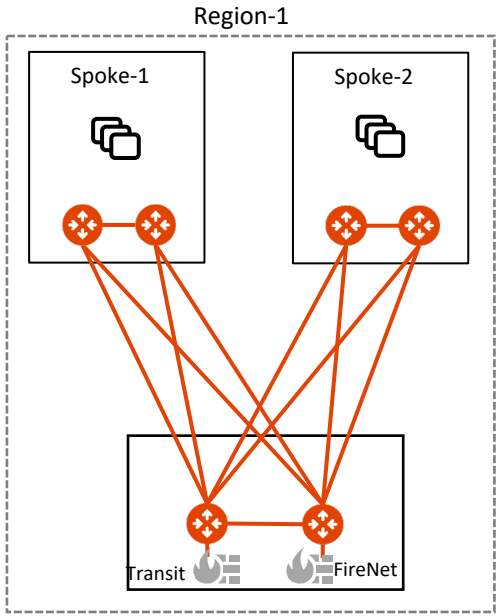
What is it ?

- Provides network **resiliency**, improved convergence time and high performance
- Two Aviatrix gateways in a VPC/VNet/VCN form a cluster
- Both gateways forward traffic simultaneously via ECMP
- Each gateway in a Spoke VPC/VNet/VCN builds IPsec tunnels to **both** Transit gateways
- Number of Transit and Spoke gateways as well as their **instance sizes** are independent of each other:
  - Maximum **2x** Transit Gateways can be deployed per Transit VPC/VNet/VCN
  - Maximum **15x** Spoke Gateways can be deployed per Spoke VPC/VNet/VCN

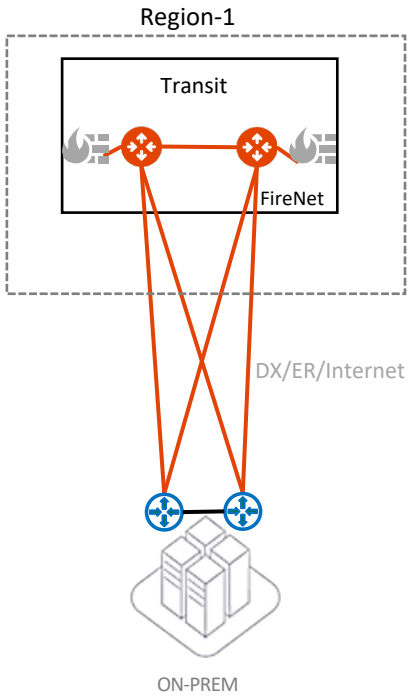


Use Cases

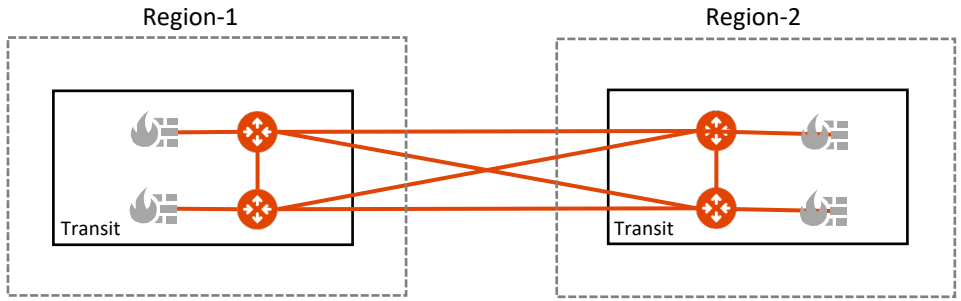
Intra-Region Spoke-Spoke



Cloud to On-Prem



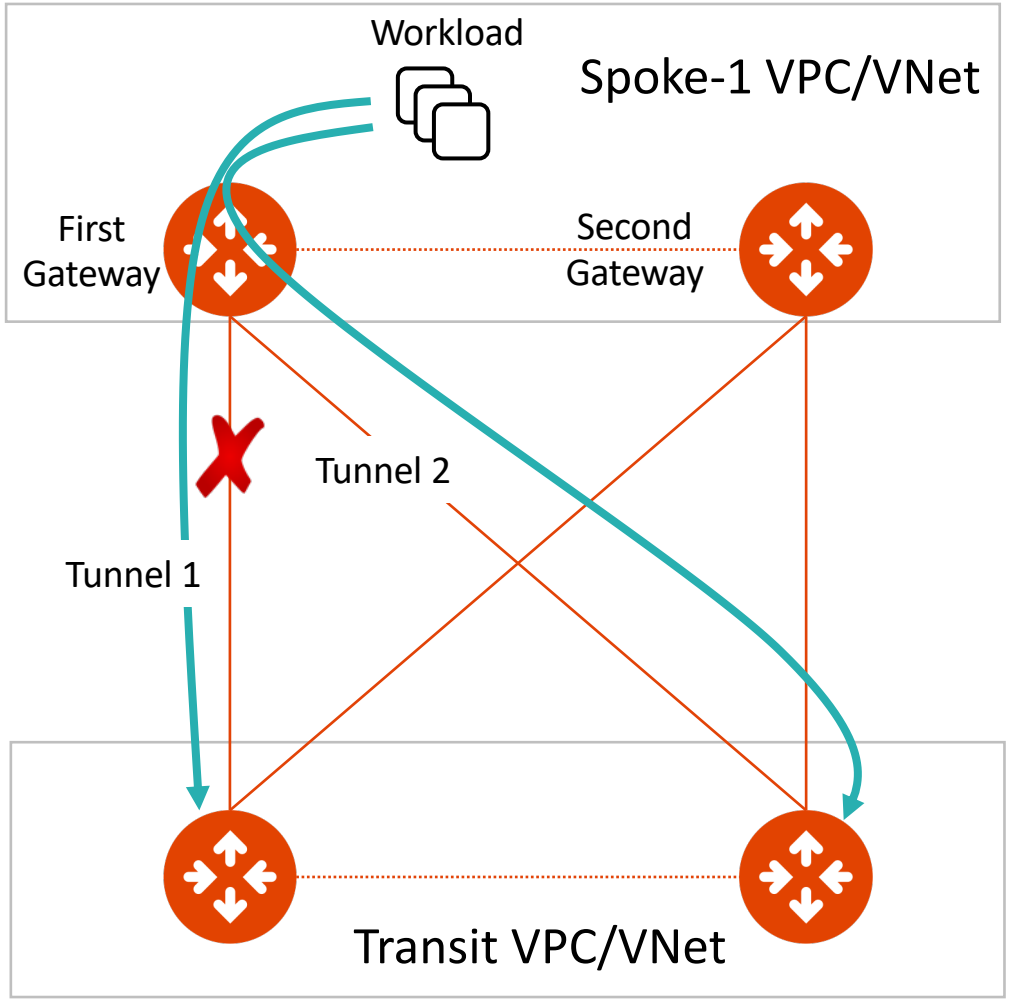
Inter-Region / Multi-Cloud



Resiliency

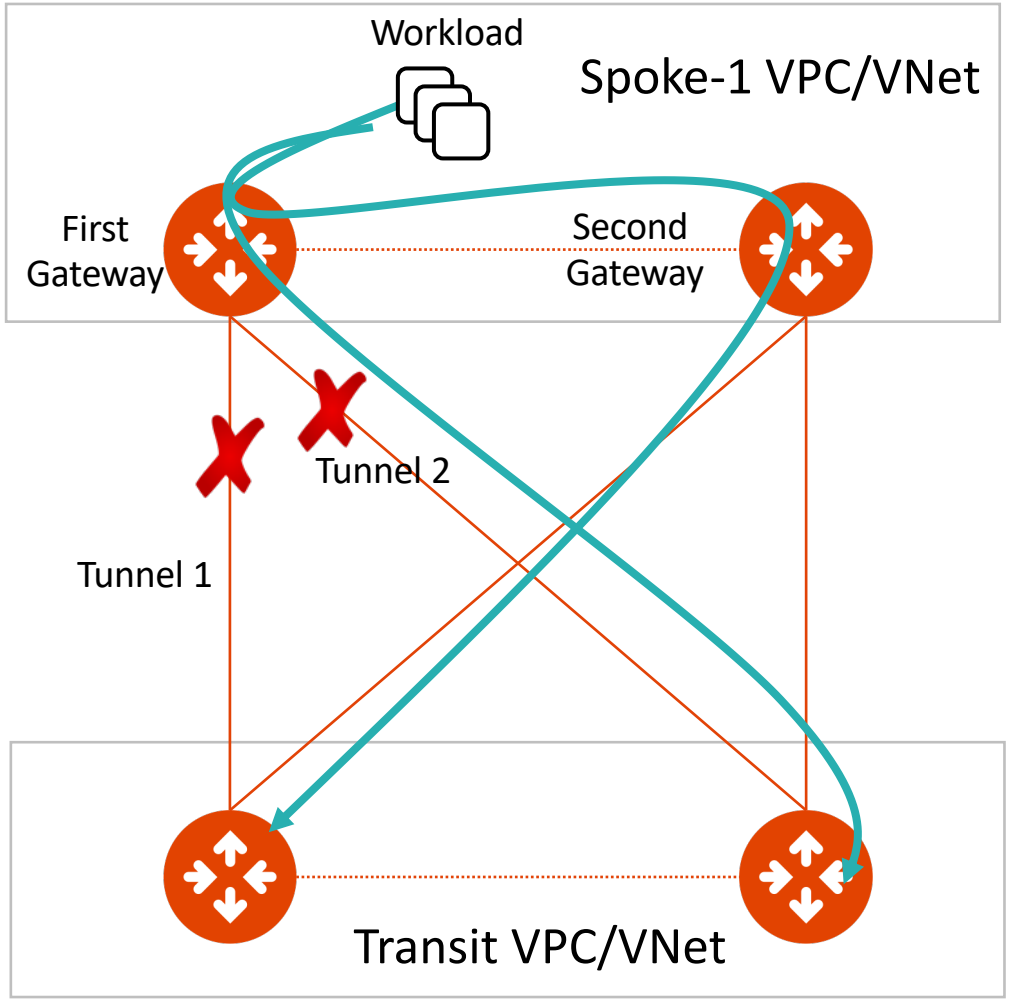
Failover Scenario 1

- Workload in Spoke-1 VPC/VNet traverses Primary gateway, Tunnel 1, onto Transit to Spoke-2 VPC/VNet (not shown)
- If Tunnel 1 at the Primary Spoke Gateway fails,
  - Then the traffic uses Tunnel 2 connected to the Secondary Transit Gateway
  - This tunnel was already active and was forwarding half of the traffic (same metric 100)
- **No re-convergence** of the routes in the VPC/VNet route table
- Gateway handles the change on its own
- Controller is aware of the tunnel going down event, but **it is not involved** in making the change



Failover Scenario 2

- Workload in Spoke-1 VPC/VNet traverses Primary gateway, Tunnel 2, onto Transit to Spoke-2 VPC/VNet (not shown)
- If both Spoke ←→ Transit tunnels fail on Primary Spoke gateway:
  - The traffic gets forwarded from the Primary Spoke gateway through the interconnected link to the Secondary Spoke Gateway
  - Secondary Spoke Gateway forwards the traffic to any of the Transit Gateways via ECMP (usual behavior – metric 100 on both downstream links)
- No re-convergence of the routes in the VPC/VNet route table
- Gateway handles the change on its own
- Controller is aware of the tunnel going down event, but it is not involved in making the change

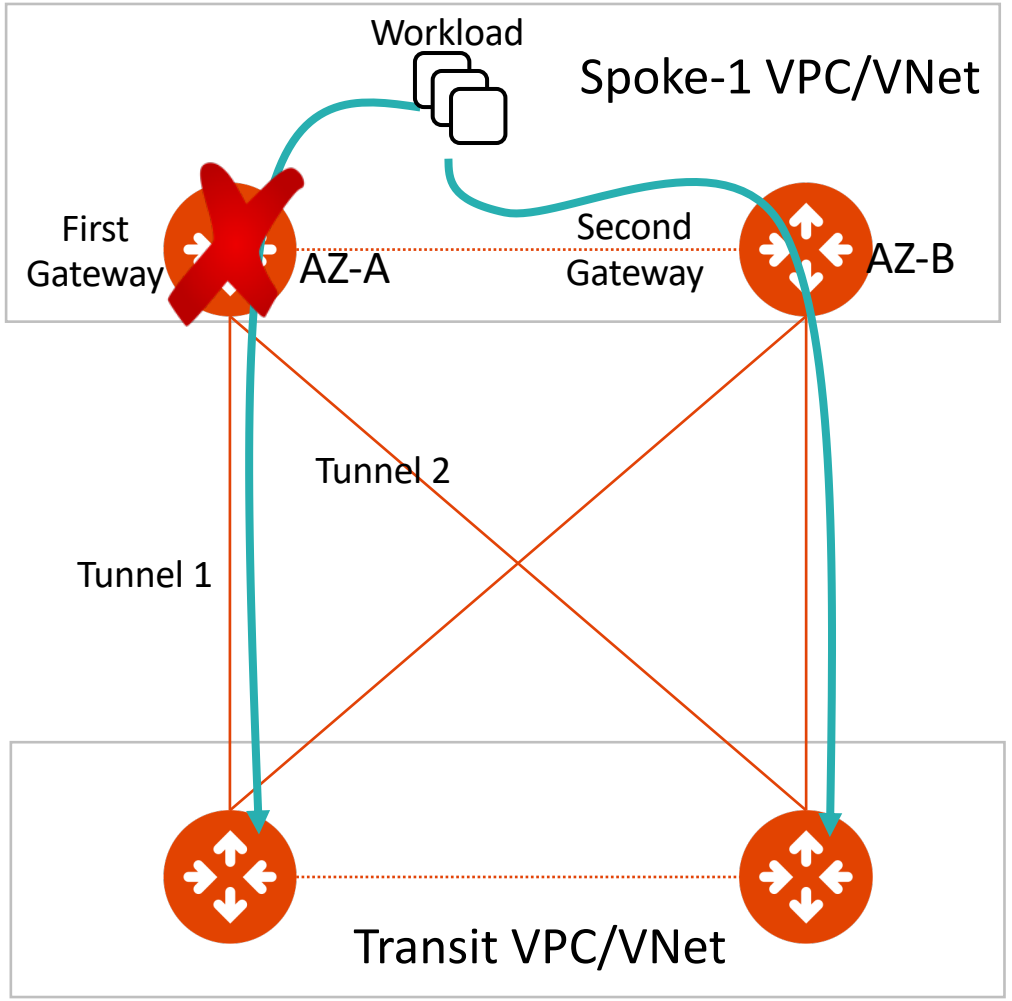


Failover Scenario 3

AZ-A
route table RT-A
Initially: 10.0.0.0/8 → Spoke-AZ-A-GW NIC
After failover: 10.0.0.0/8 → Spoke-AZ-B-GW NIC

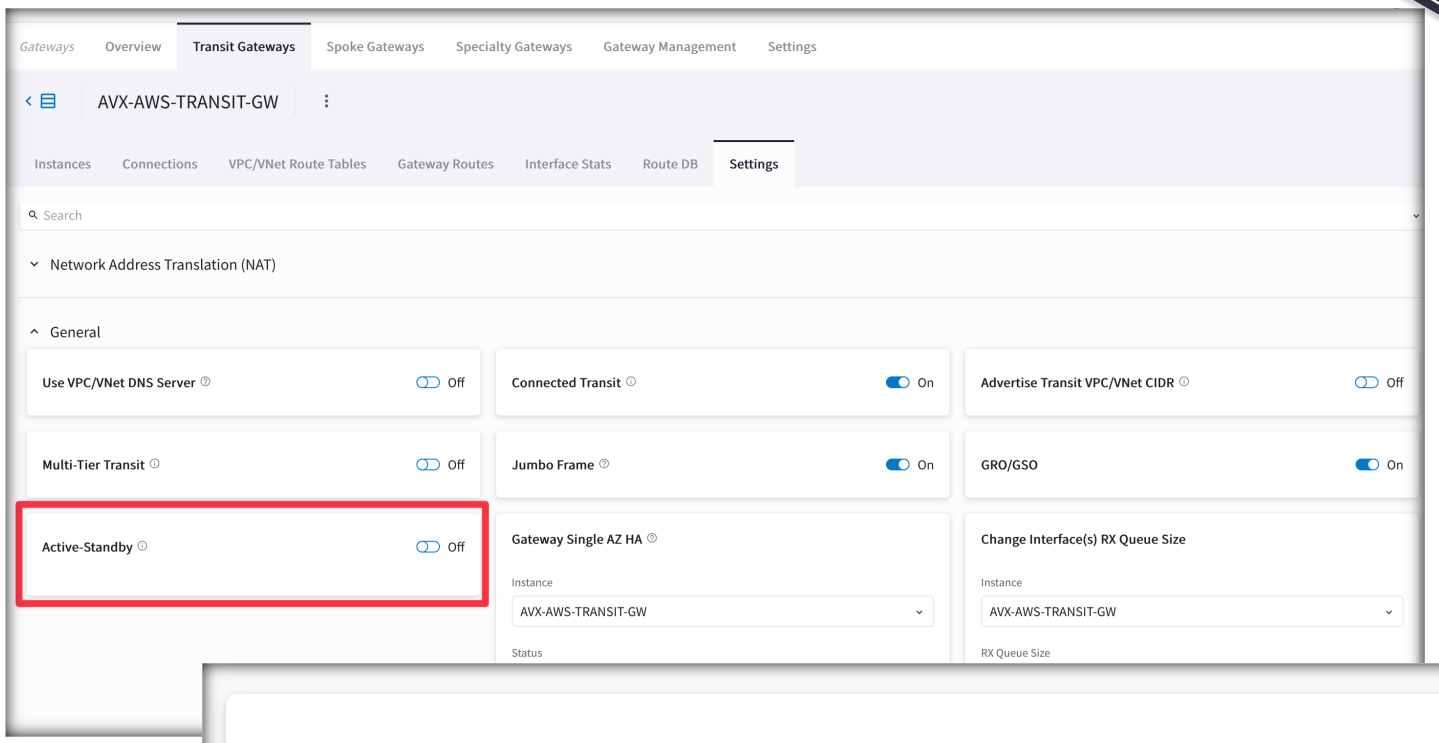
AZ-B
route table RT-B
10.0.0.0/8 → Spoke-AZ-B-GW NIC

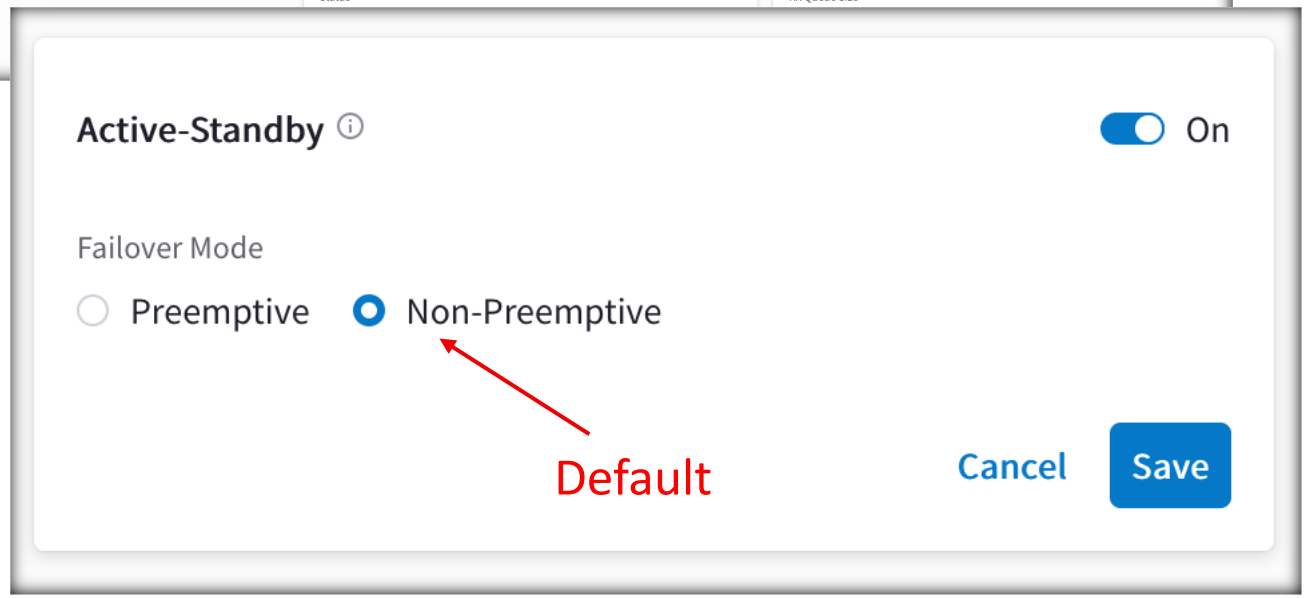
- Workload in Spoke-1 VPC/VNet needs to reach Spoke-2 VPC/VNet (not shown), but the Gateway is down
- If the Primary Gateways fails, the Controller will detect this event through the periodic keepalive messages exchanged between itself and all the gateways
- In this scenario, the Controller will **reprogram the routing table** in the AZ-A, updating the next-hop of the three RFC1918 routes with the ENI of the Second Spoke Gateway, in AZ-B



Active-Standby Mode (introduced in Controller version 6.6)

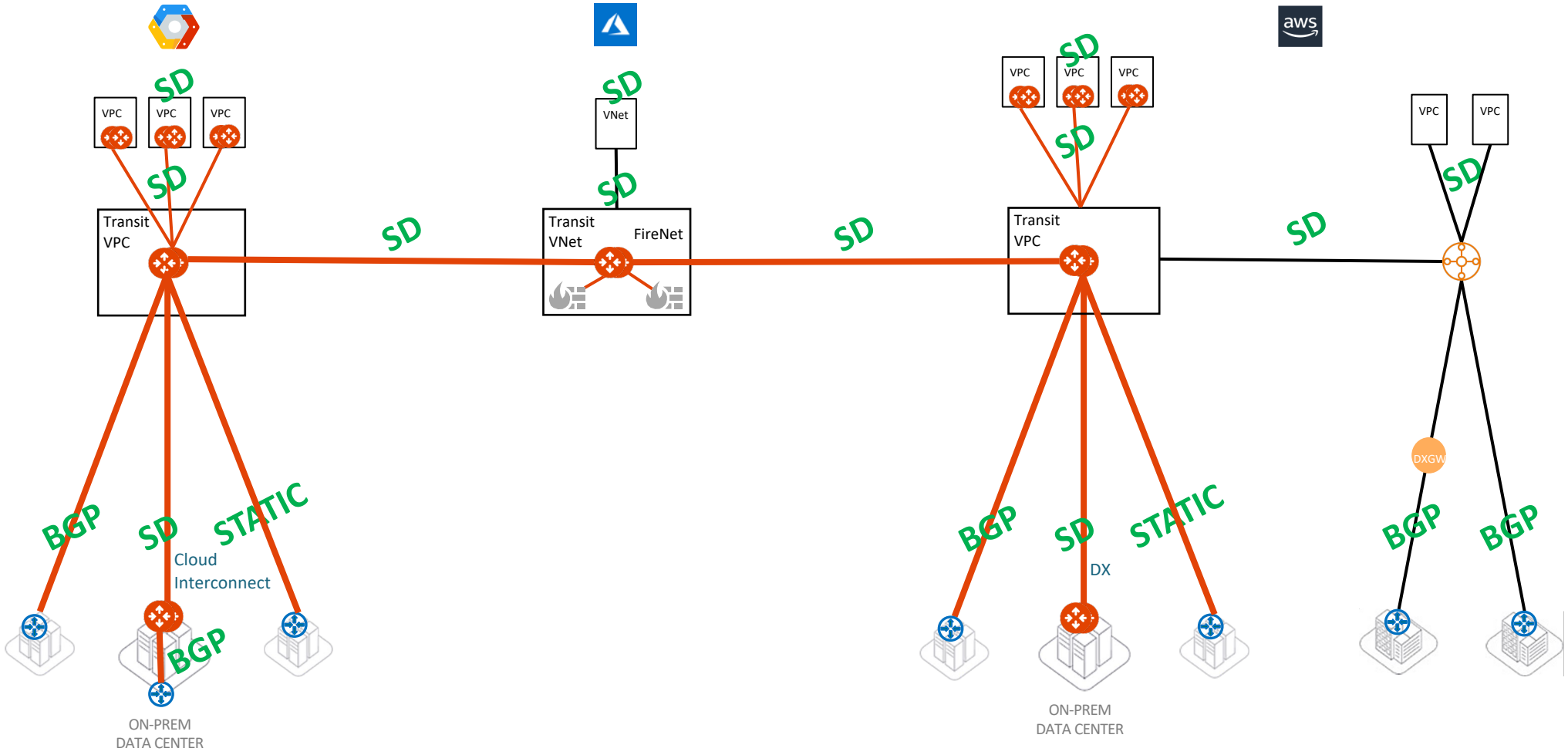
- **Use case:** Deployment scenario where on-prem device such as firewall does not support asymmetric routing on two tunnels
- Upon failure, Secondary gateway takes over from Primary
- Primary does not become active unless there is a manual switchover or Secondary failure
- UI provides option for customer to choose Preemptive or Non-preemptive behavior.





Aviatrix Control Plane

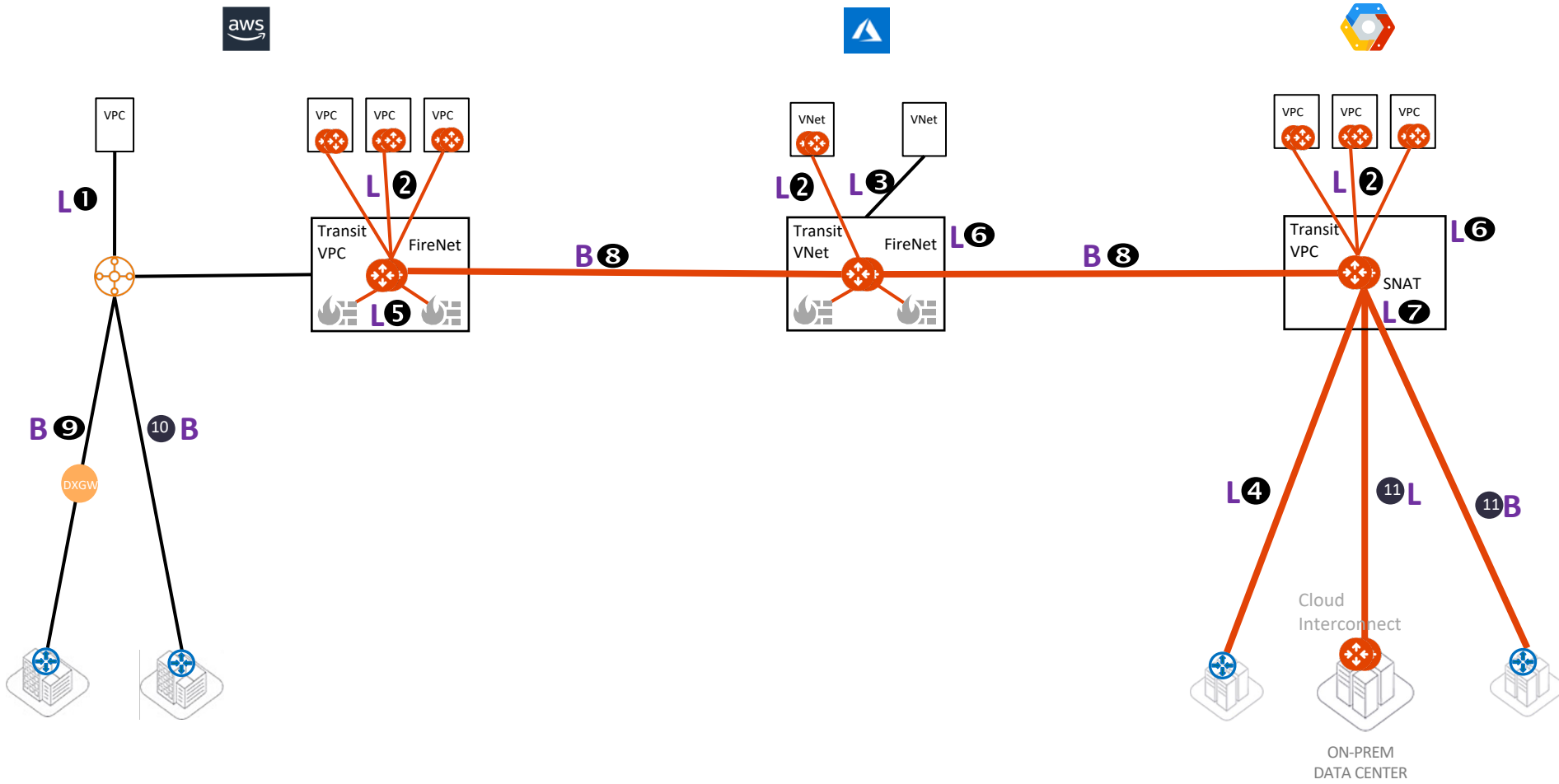
Route Programming: Software-Defined (SD) / Static / Dynamic



Route Classification

1. AWS TGW Attachment [L]
2. Aviatrix Spokes (VPC/VNet) [L]
3. Azure Native Spokes [L]
4. Aviatrix Transit GW – on-prem (*static*) [L]
5. Firewall Egress 0/0 [L]
6. Transit VPC/VNet associated prefixes [L]
7. Transit GW SNAT IP [L]
8. Remote Transit GW (Transit Peering) [B]
9. TGW DXGW [B]
10. TGW VPN [B]
11. Site2Cloud BGP on Transit GW (including Edge routes) and Site2Cloud BGP on Spoke GW [B]

L = Routes considered local by Controller
B= BGP learned routes

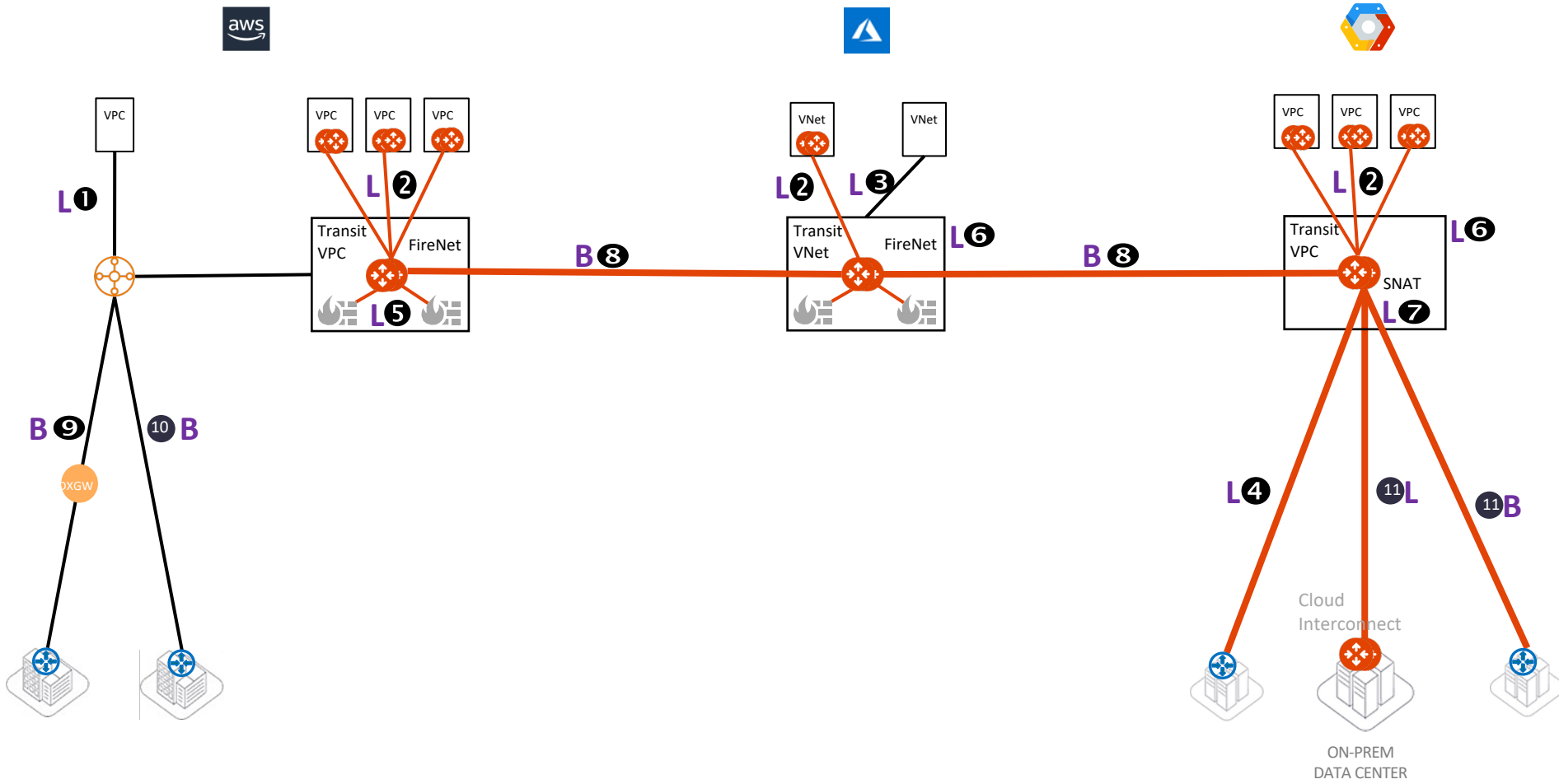


Path Selection Algorithm for Deterministic Next-Hop Selection

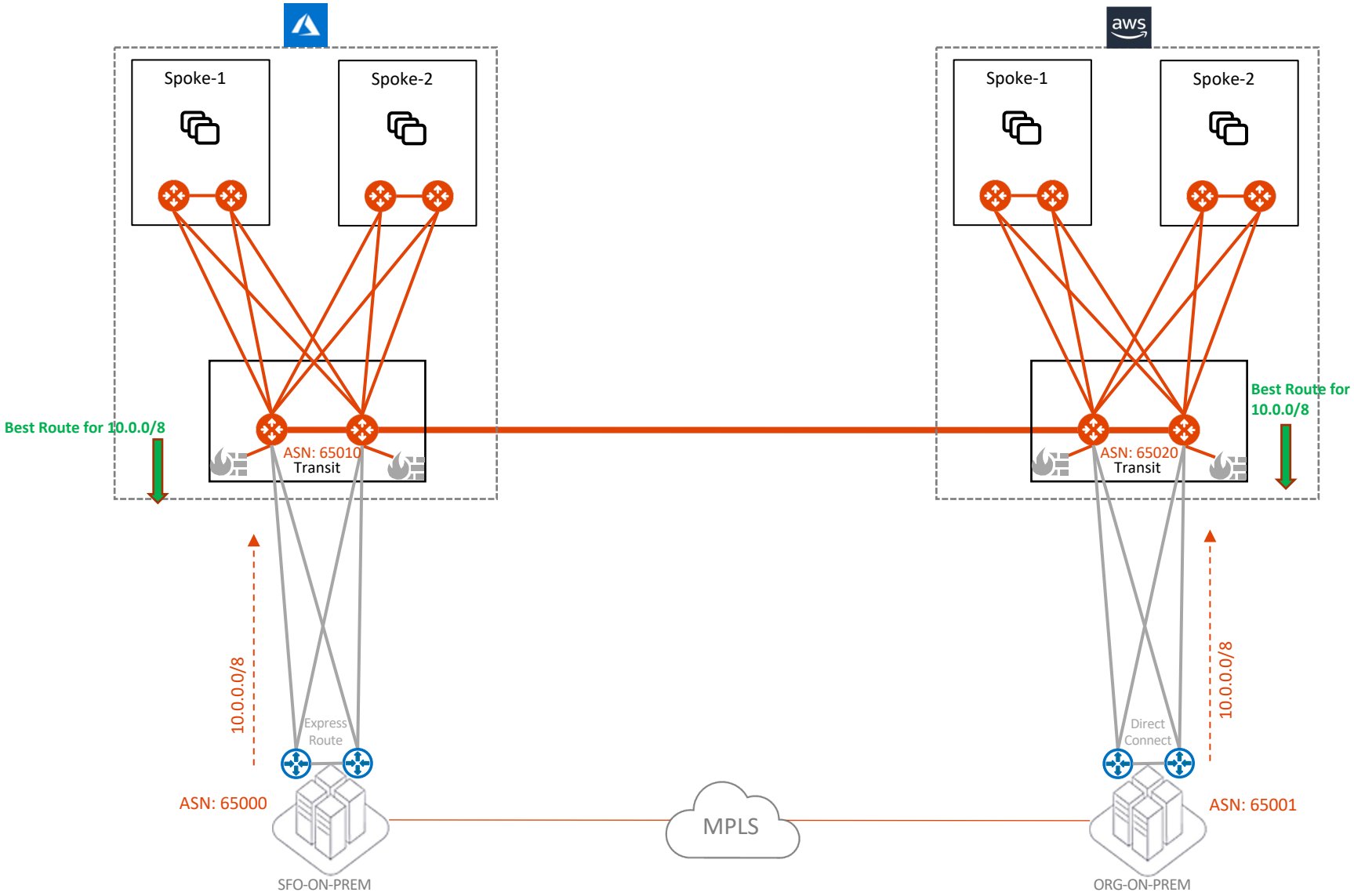
L = Routes considered local by Controller
B= BGP learned routes

Route Selection Algorithm

1. Longest prefix match
2. If equal length, then local route is chosen
3. If routes are of the same type, then shortest AS-path length is chosen
4. If AS-path length is the same, then lowest metric is chosen
5. If metric is the same, then
   - If ECMP is enabled, then traffic is distributed to available routes
   - If ECMP disabled, then the route first programmed in the table is chosen
     - If programmed at the same time, then lower integer IP next hop is chosen

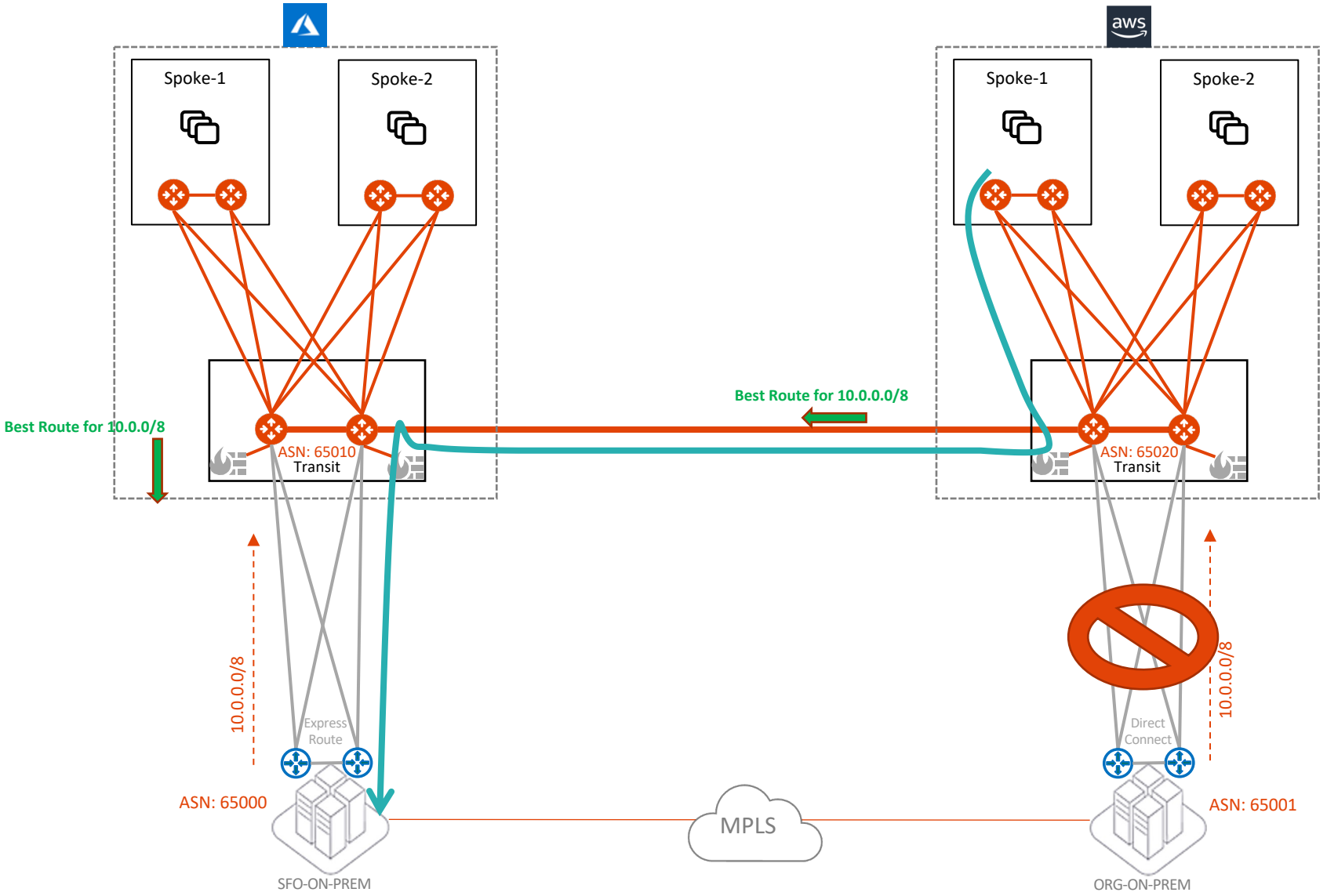


Example of using Transit as an alternate path 1/3



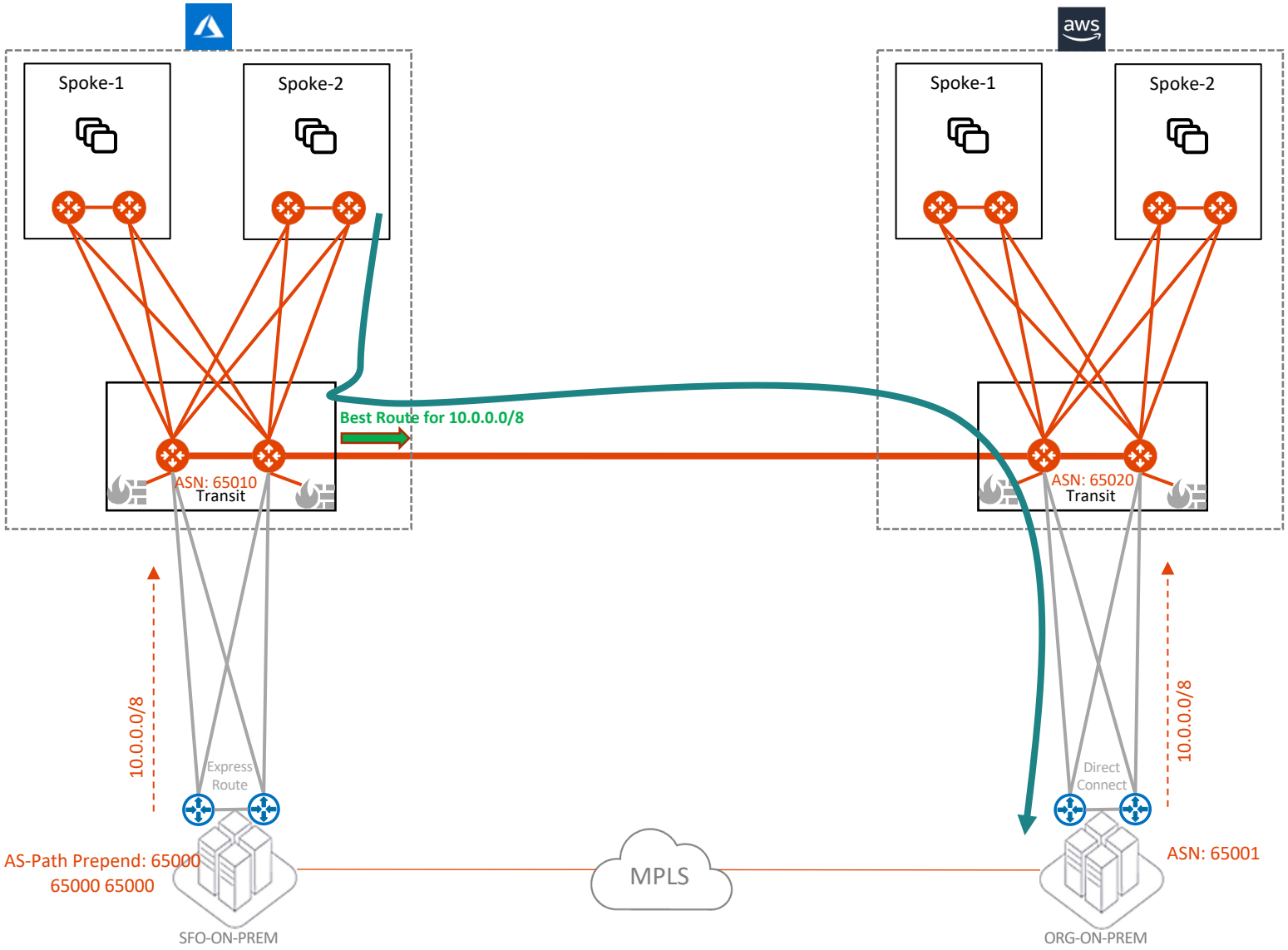
- At steady state
  - Each transit is learning 10/8 locally from on-prem
  - For each transit, Controller DB will have 10/8 via local and peer transit
  - Route via peer will have as-path-len 2
  - Each transit and its spokes will get to on-prem via local private path (DX/ER) as best path

Example of using Transit as an alternate path 2/3



- When on-prem connection goes down
  - For e.g., DX is down
  - Only route to 10/8 now is via Azure Transit

Example of using Transit as an alternate path 3/3



Use AS-PATH Prepend

- E.g, SFO on-prem ER is going under planned maintenance
- You want to avoid sending any traffic through SFO on-prem ER
- You can send AS-paths from SFO on-prem so that AWS Transit becomes the preferred path

Next: Lab 5 – HPE with ActiveMesh