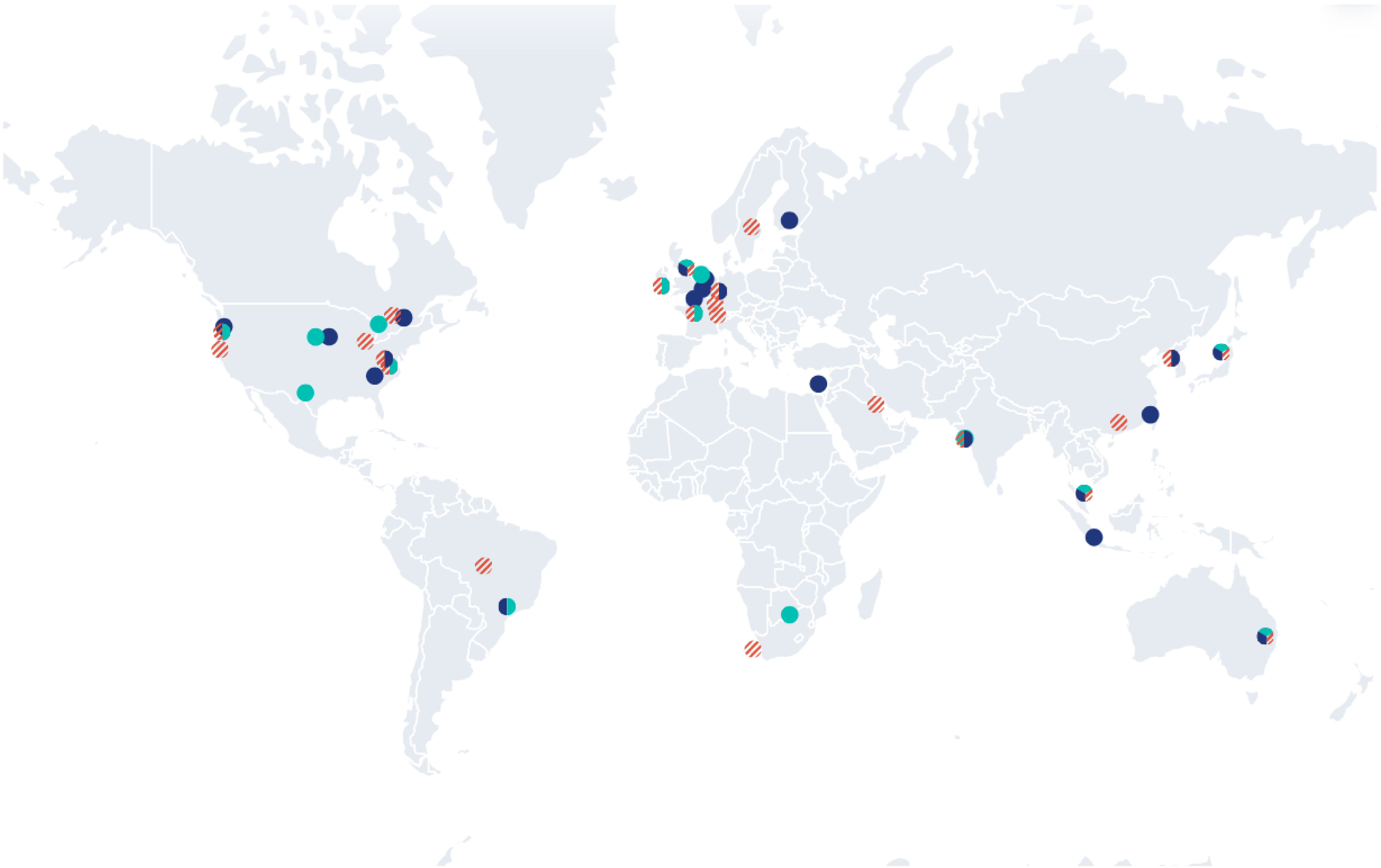
elastic | The Search AI Company

Building a global overlay network for cloud services

Christos Argyropoulos, Principal SRE

6 December 2024, GRNOG

Elastic cloud - A service with global coverage



- ● Google Cloud
- ● Microsoft Azure
- ▨ Amazon Web Services

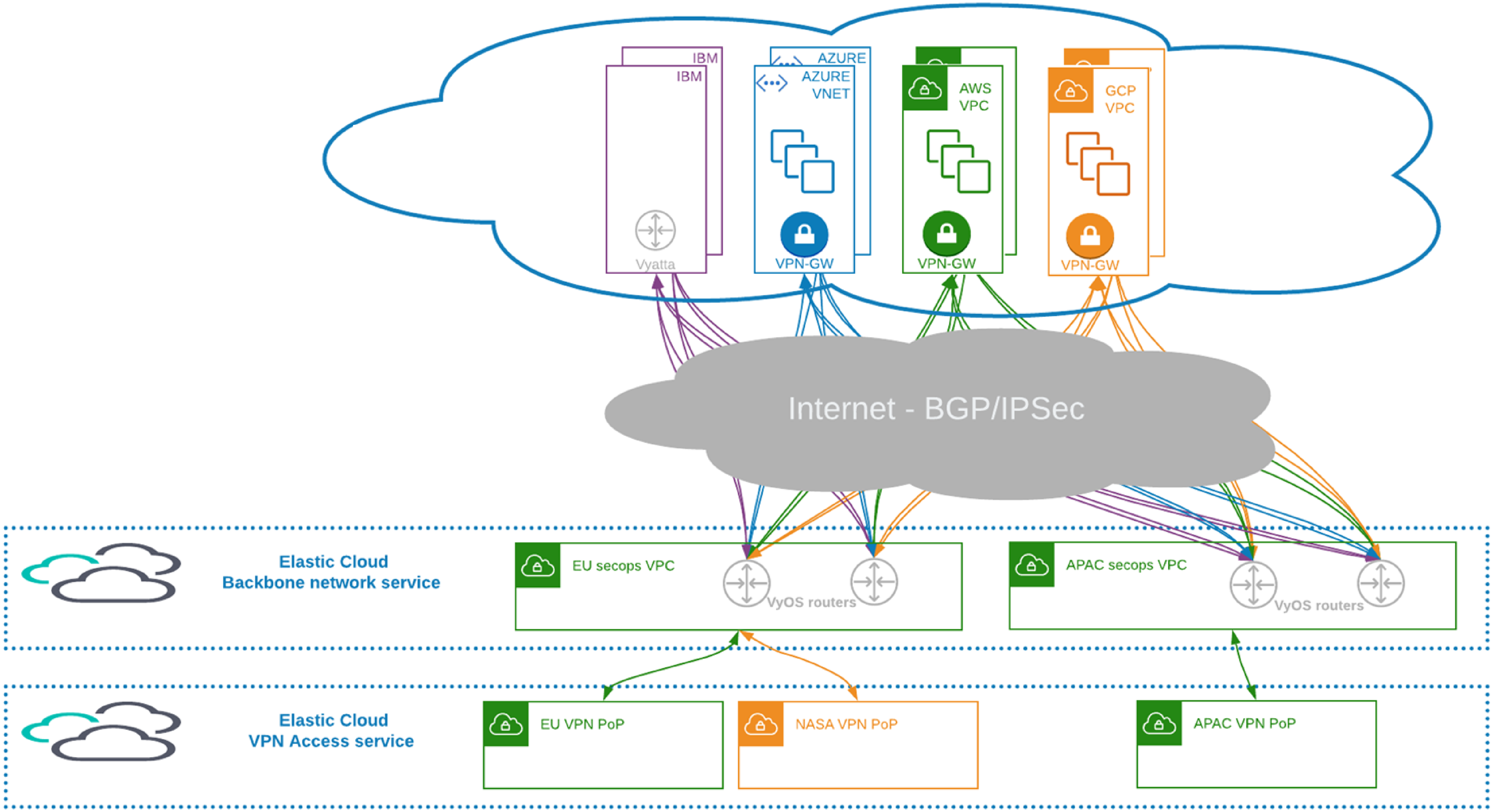
Vision for our overlay network

Build a global network fabric for a SaaS company

- Any-to-Any connectivity over a private network
    - Within same Cloud Service Provider (*scalability*)
    - Cross-CSP (*paradigm shift for a SaaS service*)
    - Simple design to reduce operational complexity
- Use-cases
    - Management-plane (host access, vaults, s/w releases)
    - Control-plane (internal platform APIs)
    - Future services (data-plane services)
        - Cross Cluster Search
        - Cross Cluster Replication

Overlay network

How we started



Vision

Requirements 1/2

- **Simplicity**
  - Operate 24/7/365 without dedicated network team
- **Scalability**
  - Connect > *n*x100 geo-regions > *k*x100 VPCs (clients)
  - Support 4 CSPs (AWS, Azure, GCP, IBM)
  - Possible further expansion
  - Multiple VPCs per region
- **Interoperability**
  - BGP for dynamic routing (the Internet cornerstone)
  - IPSec for tunneling encryption (CSP supported)
- **Reliability/Redundancy**
  - No single point of failure, high availability

Vision

Requirements 2/2

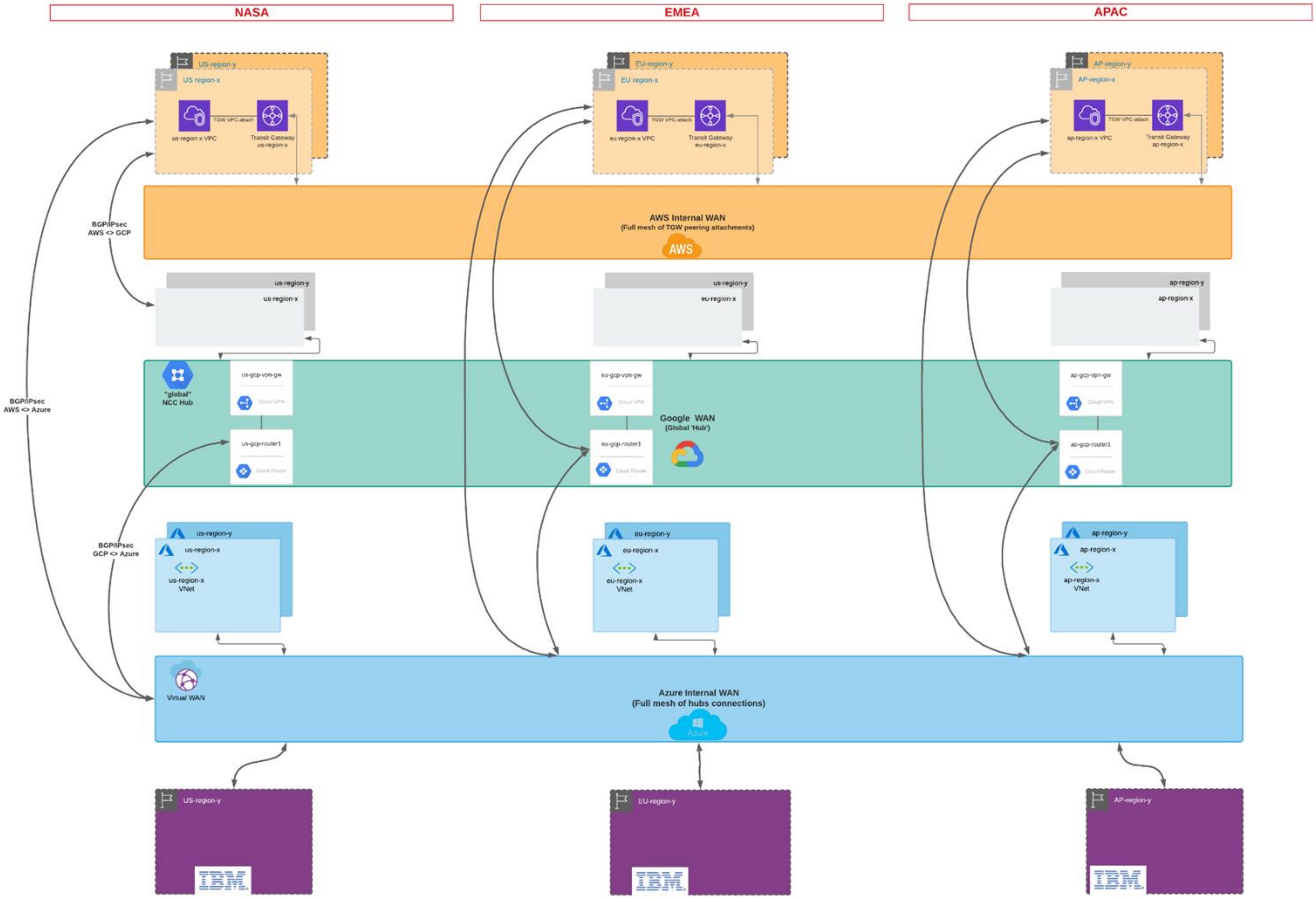
- **Routing**
  - Any-to-any connectivity
  - Traffic geo-localization (avoid extra-costs, high latencies)
  - No static routes, just

BGP

- **Automation** (e.g Terraform, Ansible)
- **Monitoring/Alerting**
- **IPv6 path**
- **Implement Identity and Access Management for the** networking equipment

Solution #1

Cloud Native



Solution #1 - Cloud native service

- **Pros:**
  - Service not Devices/Appliances
    - Managed network fabric
    - Infrastructure abstraction
  - Network fabric resiliency/scalability
    - CSPs take care of some managements tasks
    - Less pressure on the SRE team
    - Automation
  - Native integration on the provider's network infra
  - Assured future integrations with peering services

Solution #1 - Cloud native service

- **Cons (as captured in 2022):**
  - **Immaturity**
    - GCP WAN (NCC) in private-GA with critical features not supported
    - AWS TGW basic feature (dynamic routing among TGWs)
    - Azure routing policies not yet supported
    - *CSPs planned roadmap did not solve shortest path problem (for cross-CSP traffic)*
    - *Scaling Caps (# of routes)*

Solution #1 - Cloud native service

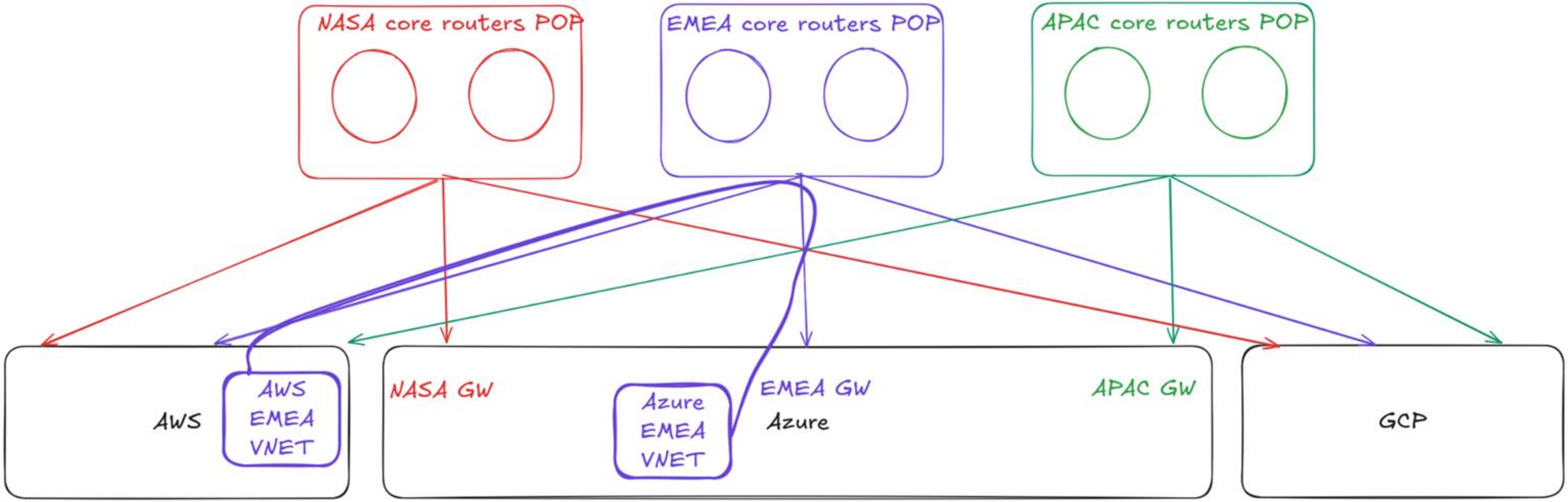
- **Challenge: cross-CSP shortest path selection**
  - *“Choose the shortest cross-CSP path in terms of latency, but choose an alternative path in case of failure to the primary path”*
- **Demand: A common ground to the BGP attributes used for CSP routing announcements**
- **Fall-back: Use S/W routers between the CSPs to implement the shortest path routing logic using BGP policies**

Solution #1 - Cloud native service

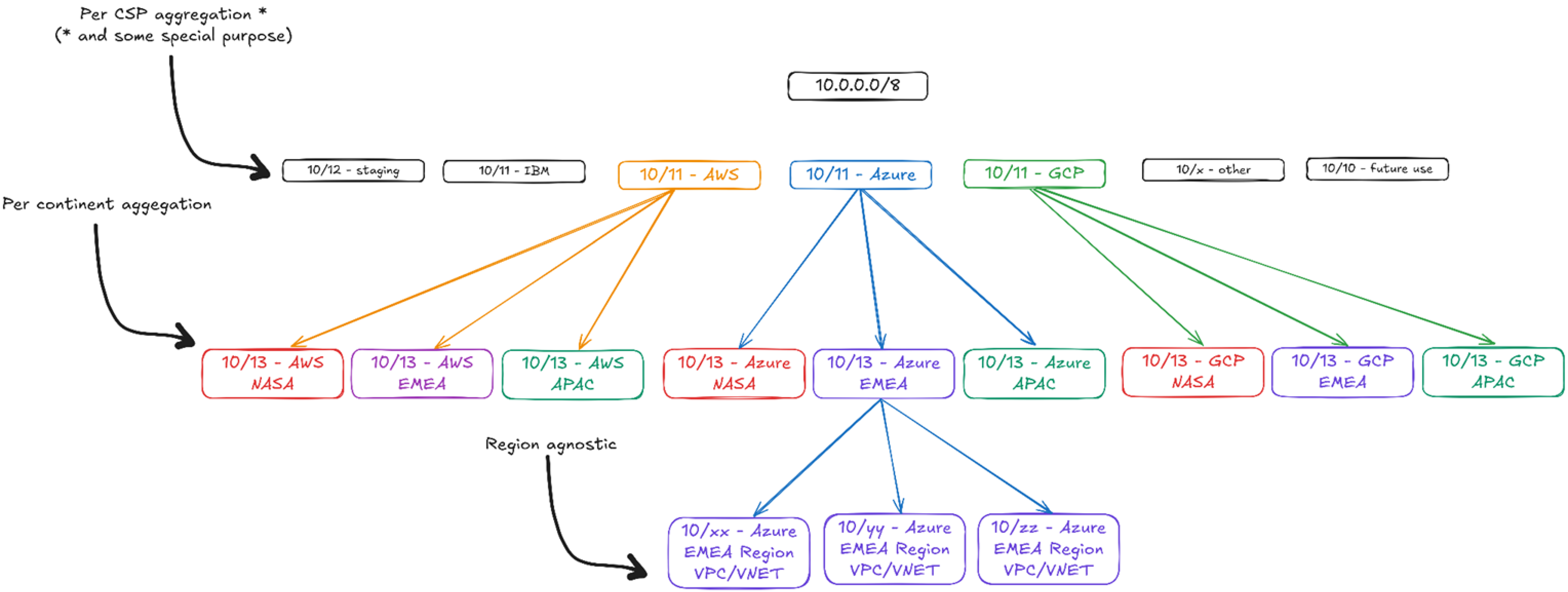
Enhancements Requests

- **Infuse CSPs with the cross-CSP SaaS concept**
  - **GCP**
    - Working with the GCP Network Product Management
    - Explain what is needed to the Dev Leads
  - **Azure**
    - Provide input to their Dev team for their routing policies
  - **AWS**
    - Working with their Network SAs and Product Team

Geo-aware routing



IPAM Policy



Per CSP aggregation *
(* and some special purpose)

Per continent aggegation

Region agnostic

10.0.0.0/8

10/12 - staging

10/11 - IBM

10/11 - AWS

10/11 - Azure

10/11 - GCP

10/x - other

10/10 - future use

10/13 - AWS NASA

10/13 - AWS EMEA

10/13 - AWS APAC

10/13 - Azure NASA

10/13 - Azure EMEA

10/13 - Azure APAC

10/13 - GCP NASA

10/13 - GCP EMEA

10/13 - GCP APAC

10/xx - Azure EMEA Region VPC/VNET

10/yy - Azure EMEA Region VPC/VNET

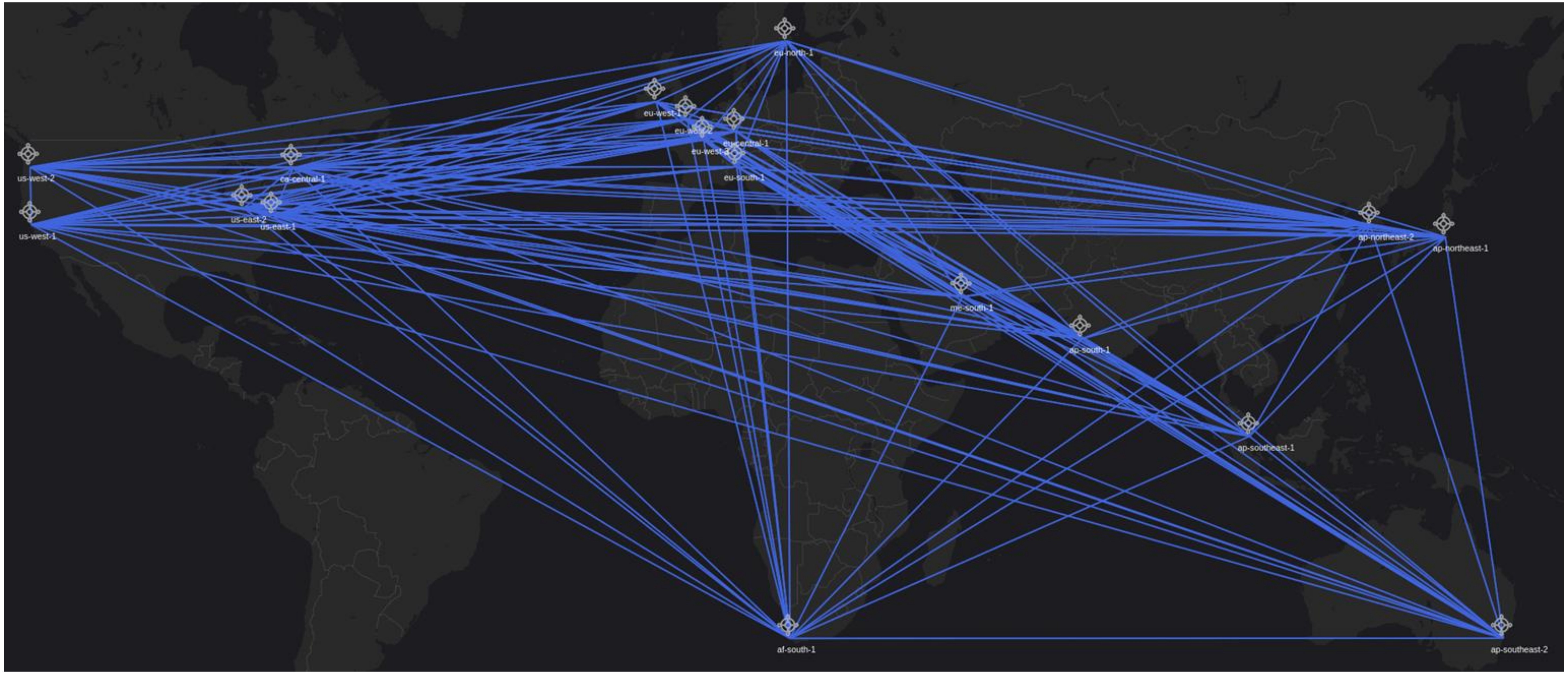
10/zz - Azure EMEA Region VPC/VNET

IPAM subnet allocation - Terraform

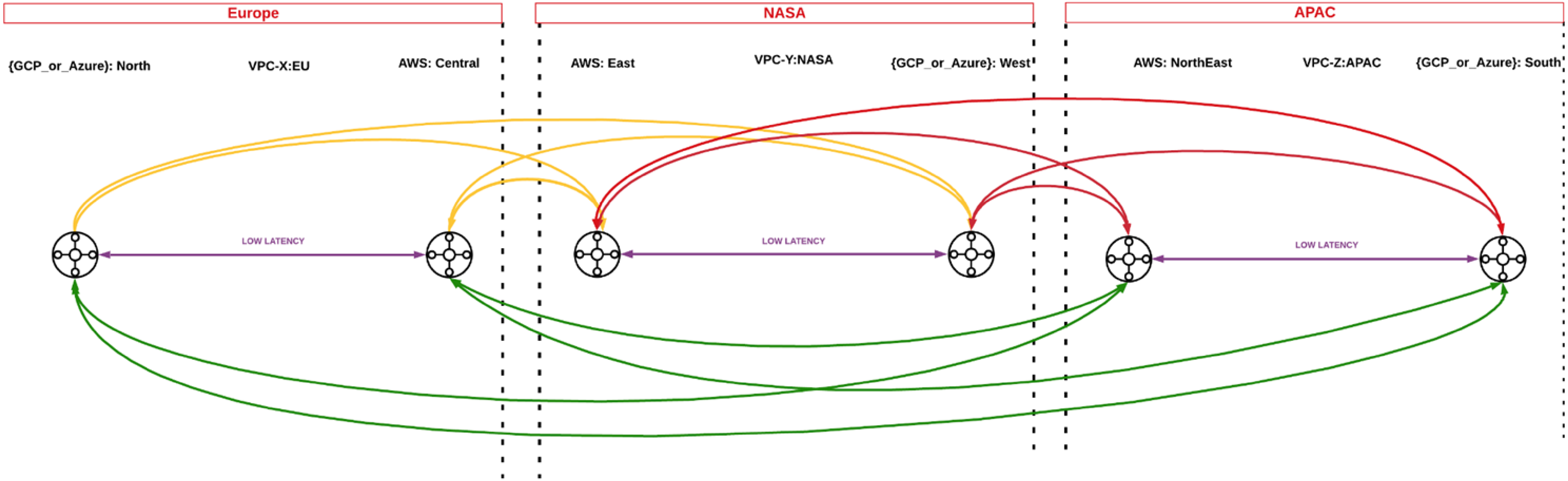
```
# This module returns the "parent" prefix that the new prefix will be allocated under
# Based on a combination of the CSP, Environment, and Region.
module "parent_prefix" {
  source = "../../modules/terraform-netbox-parent-prefix"
  csp              = "gcp"
  environment      = "qa"
  geographic_region = "nasa"
}

# This module returns the "shared" Pod and Service prefixes used for all k8s clusters
module "k8s_prefixes" {
  source = "../../modules/terraform-netbox-k8s-prefixes"
}

# This module allocates the "next_available" prefixes under the "parent" prefix defined above
module "next_available_prefix" {
  source       = "../../modules/terraform-netbox-next-available-prefix"
  new_prefixes  = local.new_prefixes
  parent_prefix = module.parent_prefix.prefix.prefix
}
```

Cross-region connectivity

AWS CloudWAN case - Today

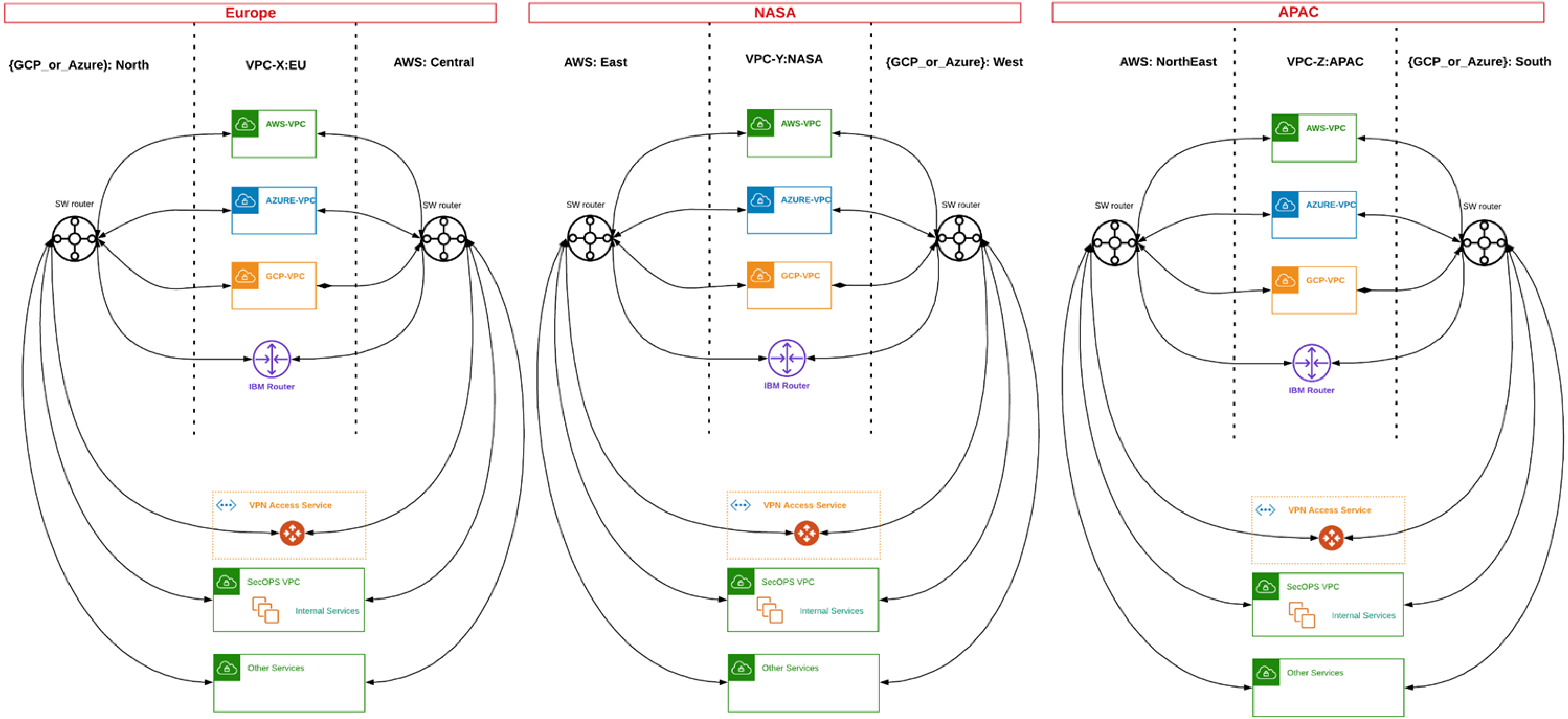


Solution #2 - Software routers



Solution #2 - Software routers

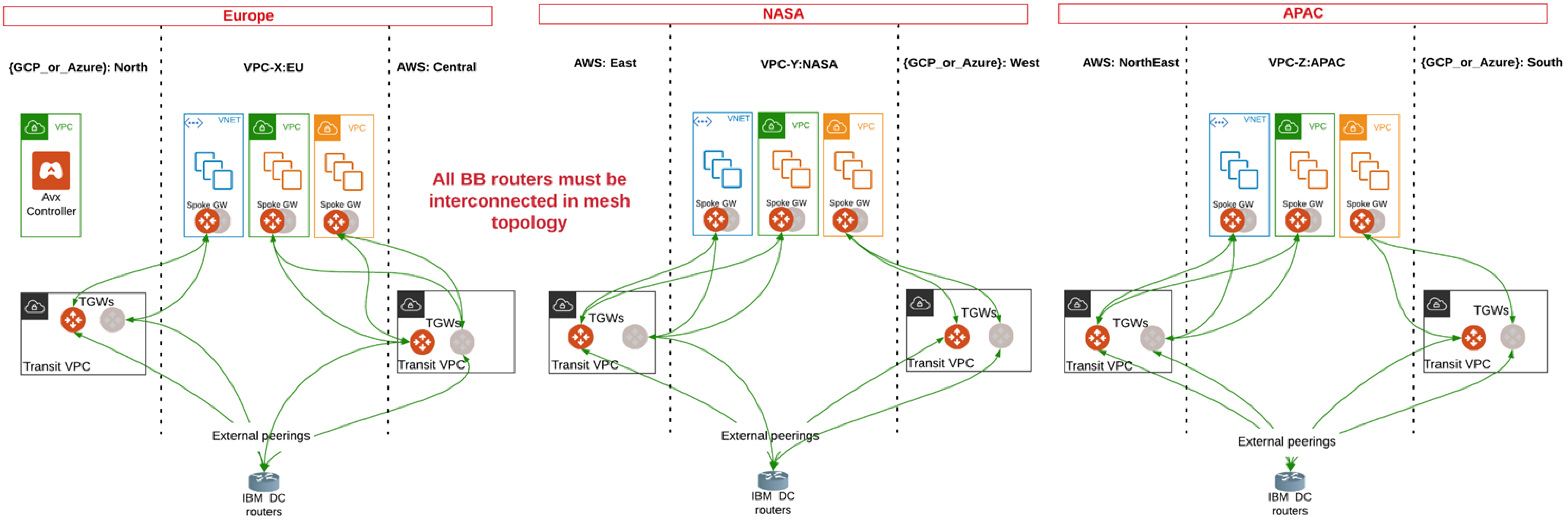
Client peerings



Solution #2 - Software routers

- **Pros:**
  - Full control of the network layer/protocols
  - Cross-vendor compatibility if vanilla network protocols are used
  - Easier migration from the previous topology
  - No vendor lock-in as the routers can be replaced gracefully
- **Cons:**
  - Steep learning curve for SREs with no network background (low-level network protocols details exposed)
  - Network protocol know-how building/investment
  - Indirect costs
    - Management costs (e.g. OS upgrades)
    - Security incidents handling

Solution #3 - SDN Vendor



All BB routers must be interconnected in mesh topology

Solution #3 - SDN Vendor

- **Pros:**
  - Centralized control/management plane (Controller)
  - Single pane of glass for monitoring and alerting
  - Abstracts the multi-CSP management/control plane
  - Established channel & partially tested solution
- **Cons:**
  - No high availability to the controller
  - Security concerns (immaturity)
    - Security incident handling and immature IAM narrative
  - Indirect costs (e.g. OS upgrades, security incidents)
  - TCO (~ 150% of native CSP) for licensing & resources
  - Scalability (no running deployment at our scale)

Conclusions (cloud native solution)

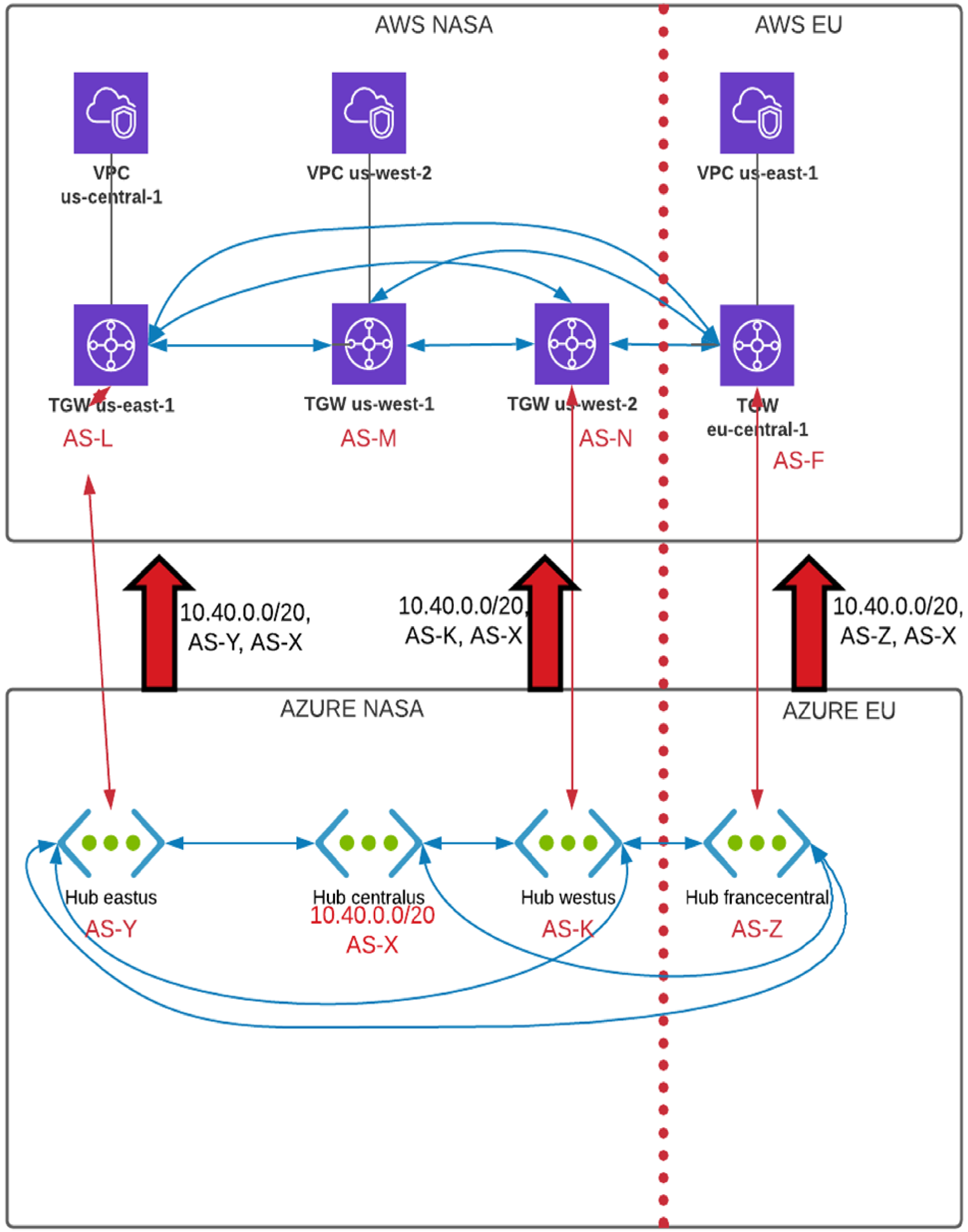
- **Simplified and automated operations - Lifesaver**
  - (Most) SREs lack deep networking expertise, intentional focus on other skills as doesn't match our core business
- **Segmentation - Lifesaver**
  - Not included to our initial list of requirements

- **Provisioning speed, Scalability - Lifesaver**
  - Able to build and wire multiple VPCs in multiple regions in less than 10 minutes
    - in AWS, yes
    - in Azure under certain circumstances
    - Enabler for Kubernetes cluster roll-out in new VPCs
- **Costs - Headache of 0.02$/GB**

Thank you!

elastic | The Search AI Company

Challenge: cross-CSP shortest path selection



TGW us-west-1 wants to reach centralus

10.40.0.0/20, AS-L, AS-Y, AS-X
10.40.0.0/20, AS-N, AS-K, AS-X
10.40.0.0/20, AS-F, AS-Z, AS-X

1. Longest-prefix match criteria tie
2. AS-PATH criteria tie
3. MED criteria not applicable
4. ECMP or older wins

==> Non-determenistic traffic tromobone
==> Route-maps required