Global IP Routing for Kubernetes Clusters

Antonios Chariton

WARNING

This presentation contains IPv4, which is known to cause controversy and polarize communities.

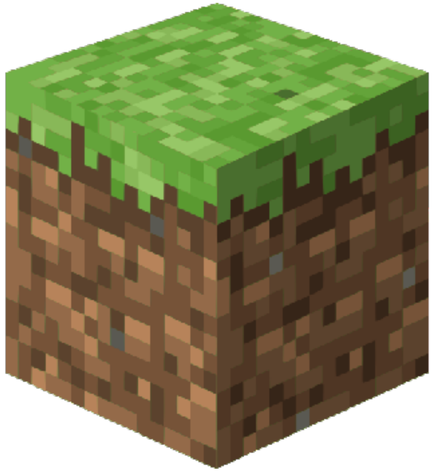
Please deploy IPv6!

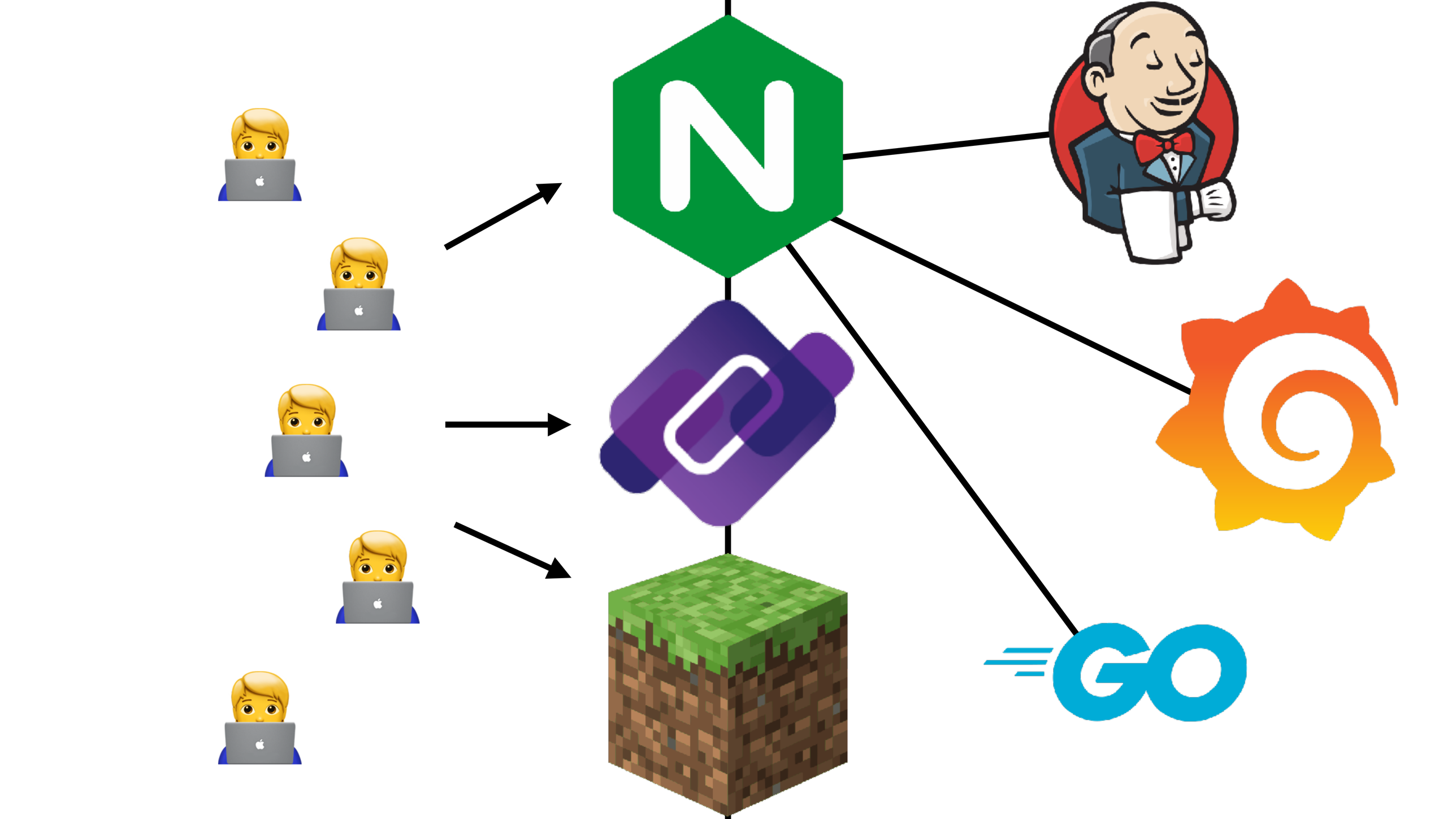
WARNING

This presentation contains IPv4, which is known to cause controversy and polarize communities.

Please deploy IPv6!

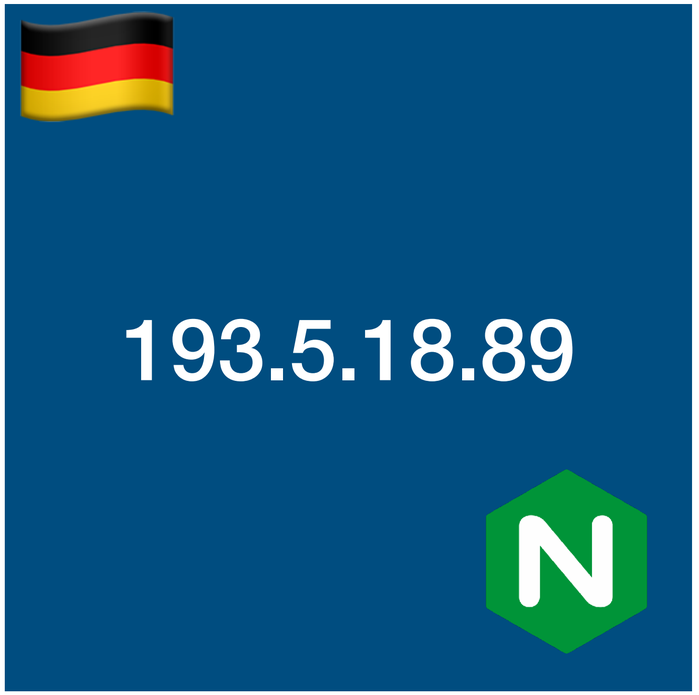

SAS
1.2TB10k
SAS
1.2TB10k
SAS
1.2TB10k
iDRAC Quick Sync
iDRAC Quick Sync






193.5.16.5


193.5.17.3


193.5.18.89

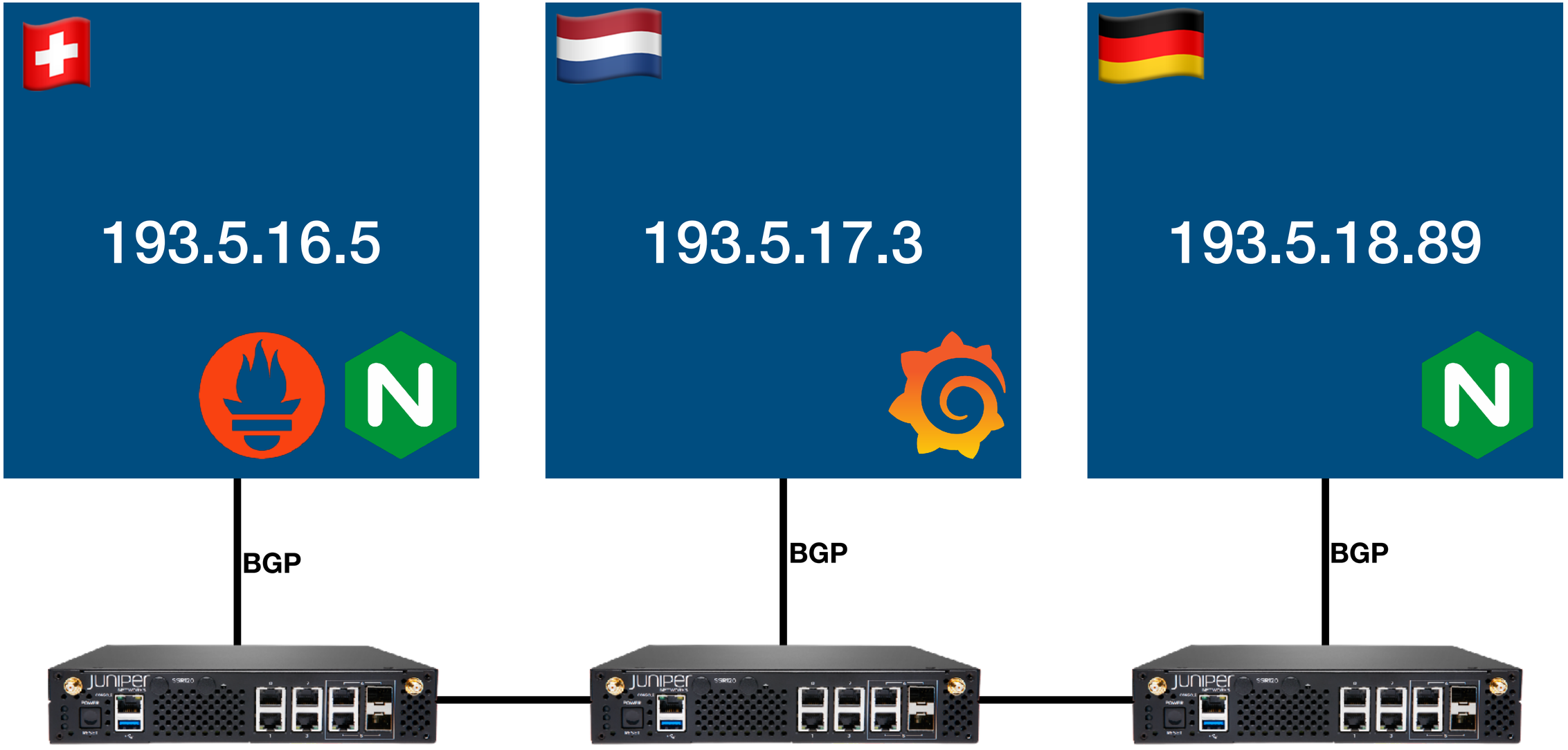

```
www IN A 193.5.16.5
www IN A 193.5.18.89
```


```
www IN A 193.5.16.5
www IN A 193.5.18.89
```

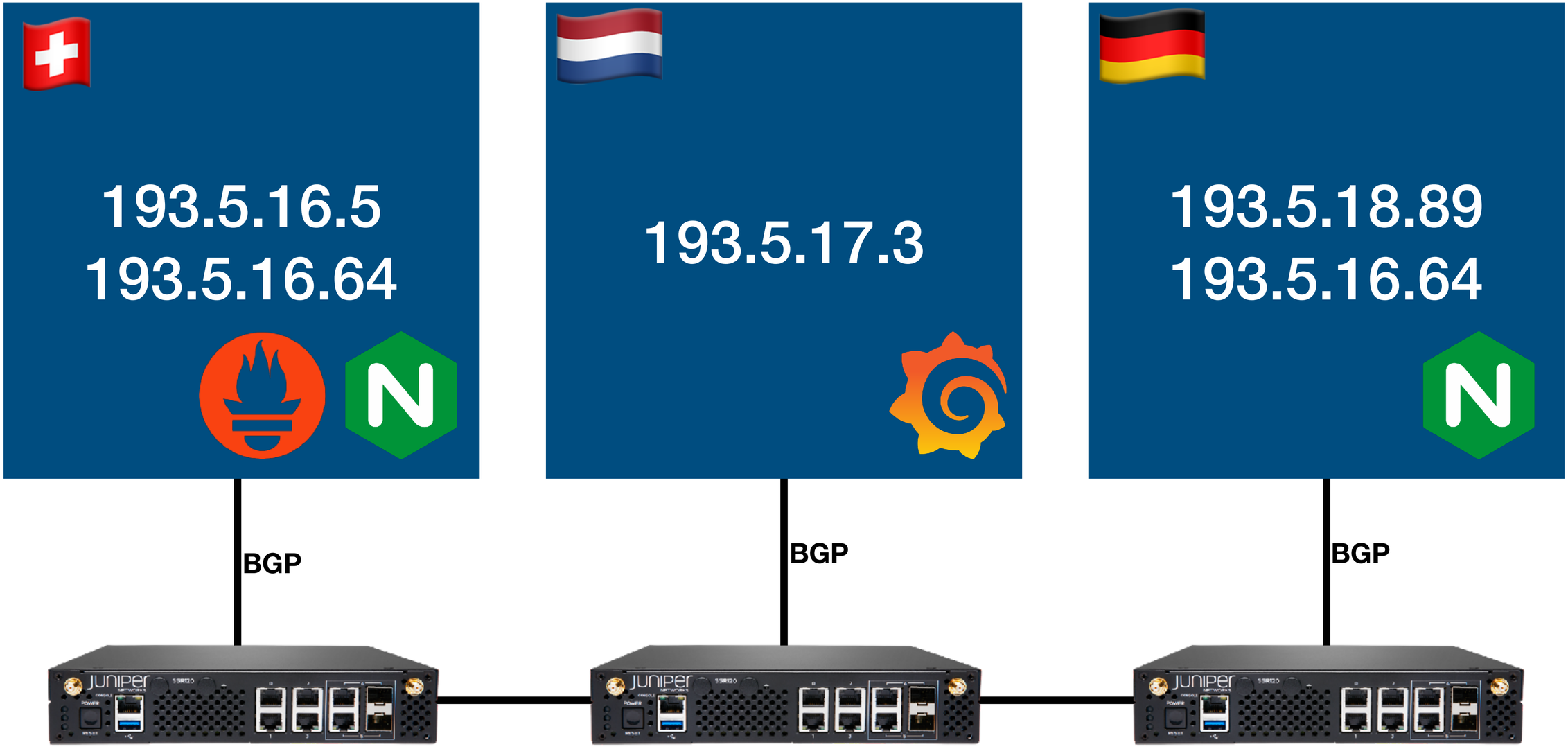

```
www IN A 193.5.16.5
www IN A 193.5.18.89
```


```
www IN A 193.5.16.5
www IN A 193.5.17.3
```


```
www IN A 193.5.16.5
www IN A 193.5.17.3
```

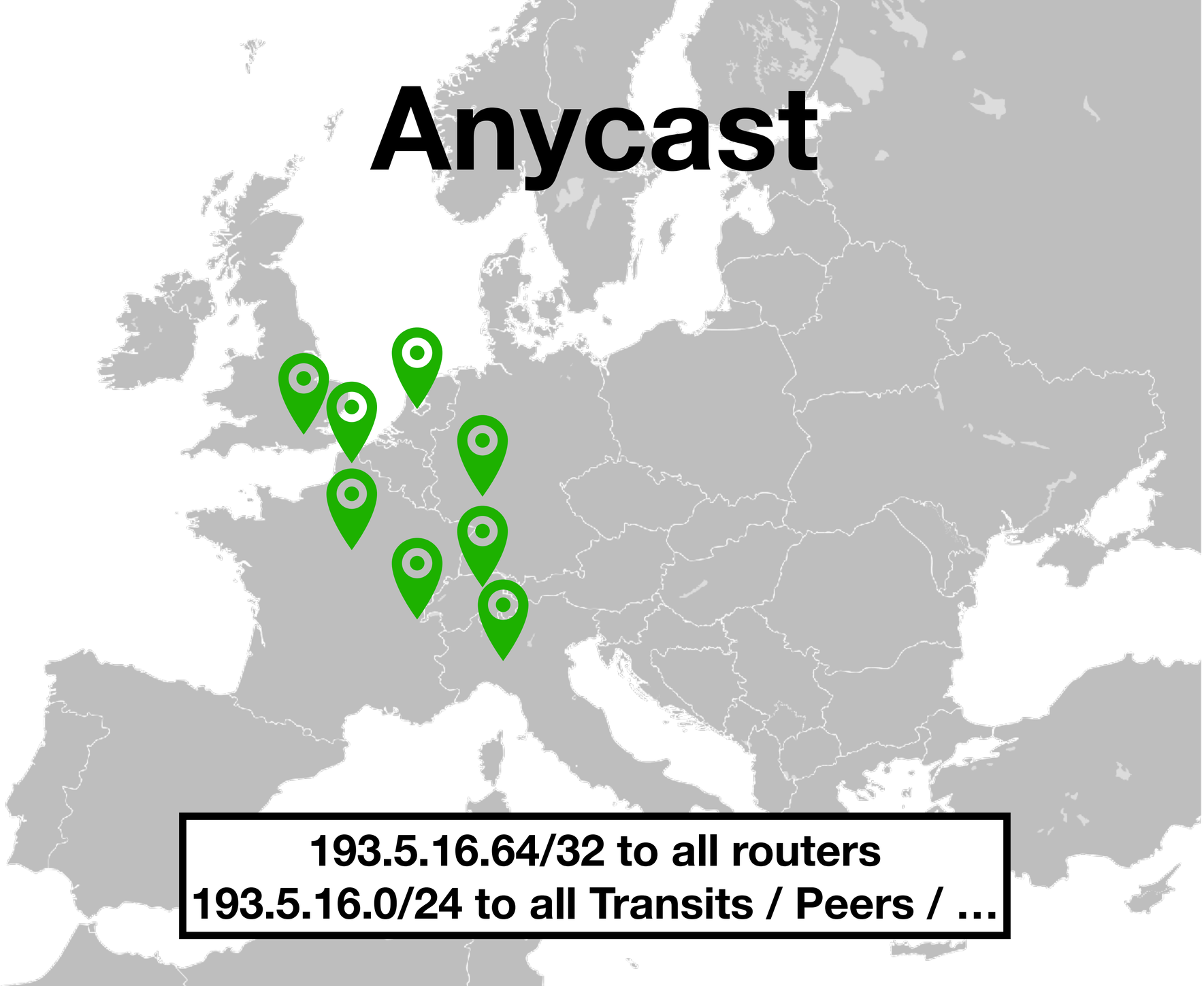

193.5.16.5
193.5.17.3
193.5.18.89
BGP
BGP
BGP

```
---
apiVersion: v1
kind: Service
metadata:
  name: exposed-service
  annotations:
   metallb.universe.tf/loadBalancerIPs: “193.5.16.64,2a0d:3dc0::16:64”
spec:
  ports:
  - port: 443
    targetPort: 443
  selector:
    run: my-nginx
  type: LoadBalancer
  ipFamilyPolicy: PreferDualStack
  externalTrafficPolicy: Local
```


193.5.16.5
193.5.16.64
193.5.17.3
193.5.18.89
193.5.16.64
BGP
BGP
BGP

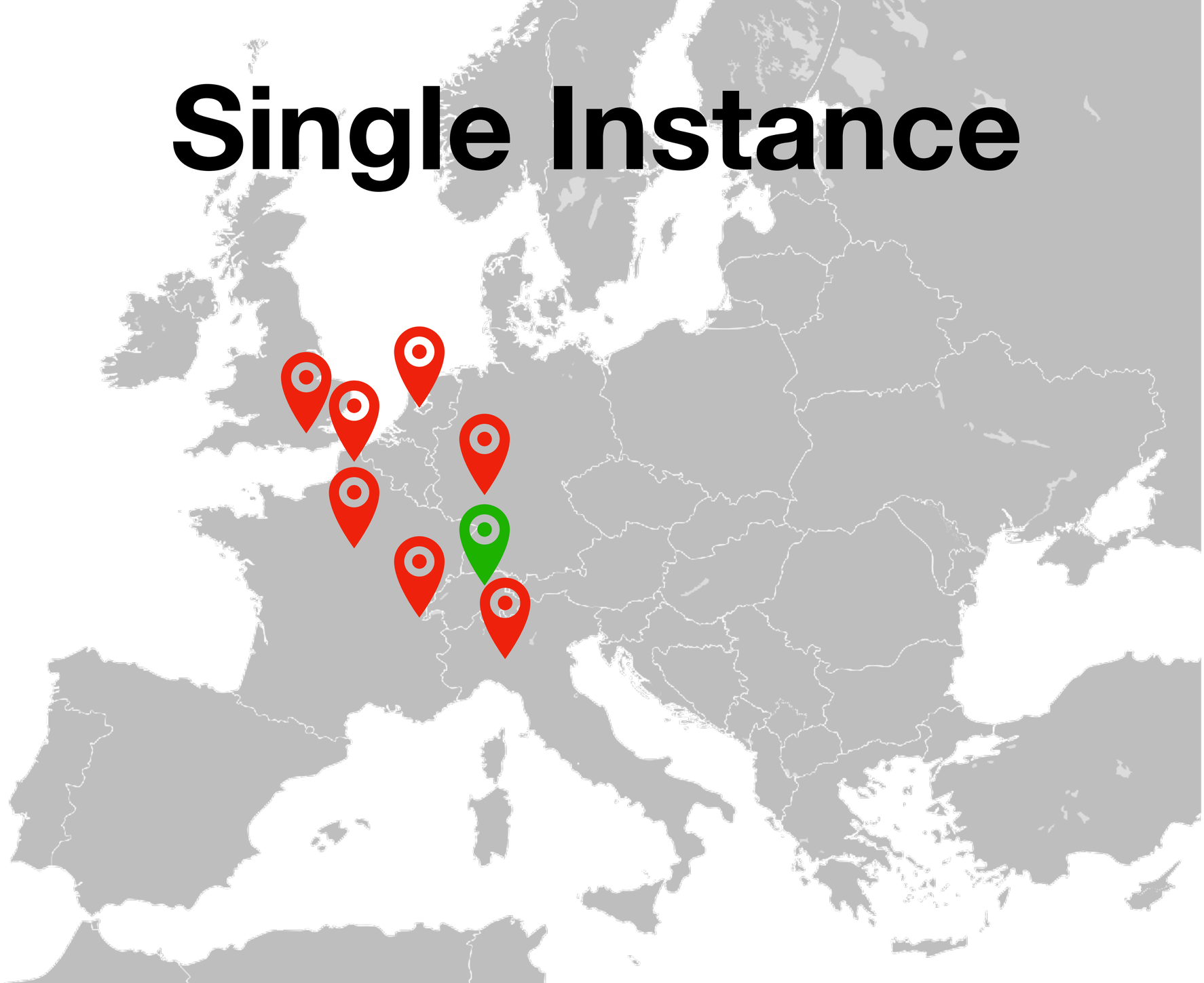
```
www IN A 193.5.16.64
www IN A 193.5.17.64
www IN A 193.5.18.64
www IN A 193.5.19.64
```

Anycast



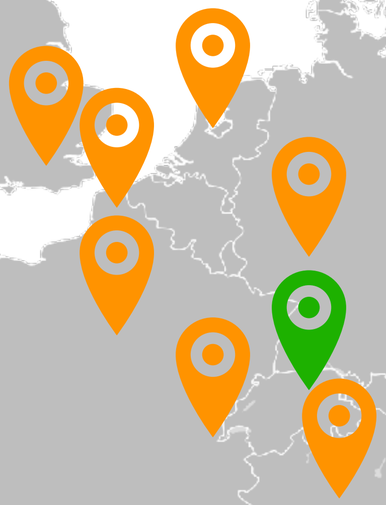
193.5.16.64/32 to all routers
193.5.16.0/24 to all Transits / Peers / …

Single Instance



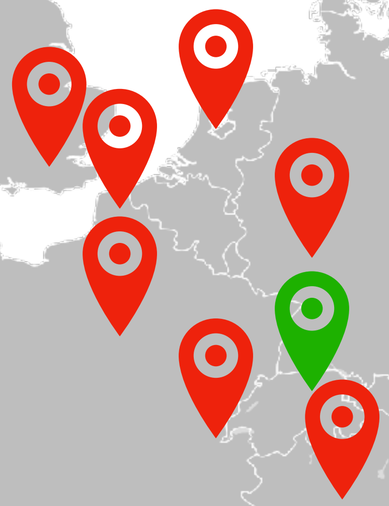
Few Instances

Important
Single Instance



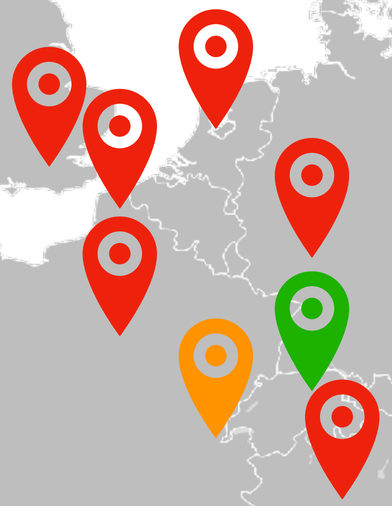
193.5.16.65/32 to ZRH routers
193.5.16.0/24 to all Transits / Peers / …

Lazy
Single Instance



193.5.17.64/32 to ZRH routers
193.5.17.0/24 to ZRH Transits / Peers / …

Group Lazy
Single Instance



193.5.17.64/32 to ZRH routers
193.5.17.0/24 to CH Transits / Peers / …

IPAM Accounting

- Anycast
  - 1 Prefix, advertised anywhere
- Important Unicast
  - 1 Prefix, advertised anywhere (can be same as Anycast)
- Lazy Unicast
  - N Prefixes, one per group (PoP, Metro, Country, …)

But IPv4 is expensive!

Then just offer the better service over IPv6 only ;)

Can DNS Help?

- Move hosts per domain to different group
  - All nginx servers can serve all traffic
  - Anycast VIP -> Europe VIP
  - Zurich VIP -> Switzerland VIP

Fine-tuned Traffic Engineering

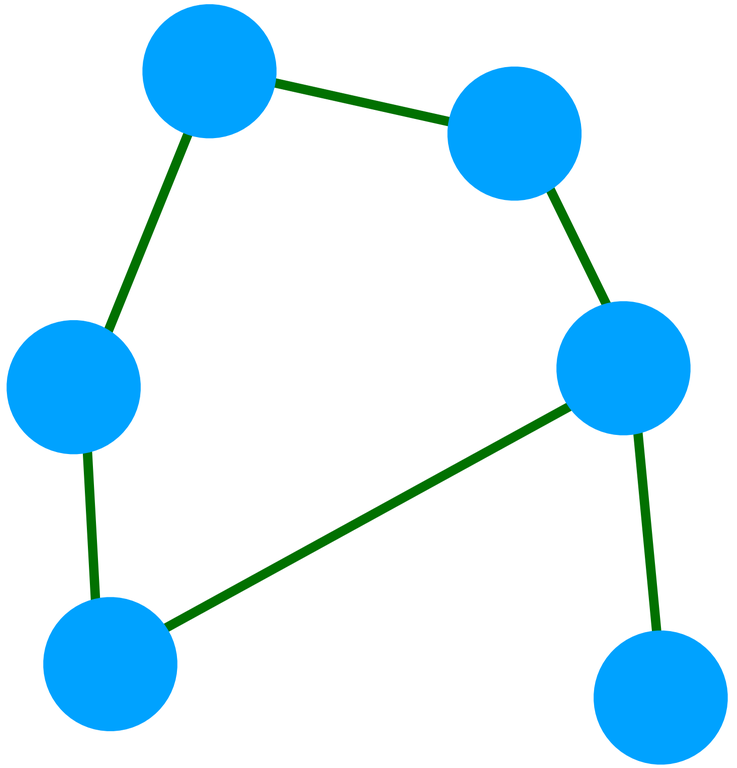
- Network Automation (BGP) can change group policy in < 1’
- DNS Automation can change group in ~ `$TTL` seconds
- Network data can drive decisions (link utilization, etc.)
- Careful of: RPKI, `route{6,}`, BGP Prefix Limits, Flap Dampening, Stuck Routes (in the IGP)

Service Traffic Class Priorities

www-nginx	1000
grafana	800
debian-mirrors	300
sysadmin-minecraft	15

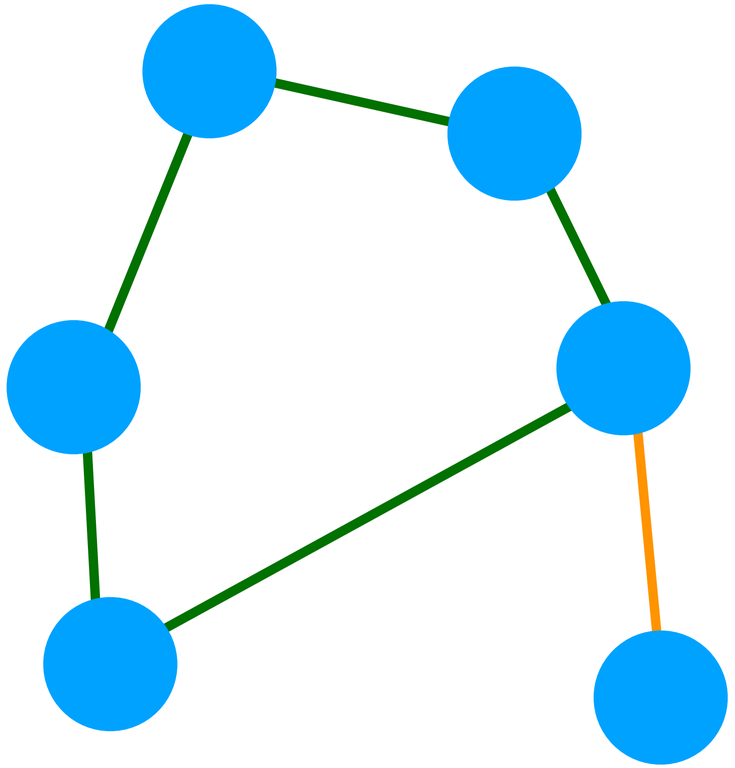
Service Traffic Class Priorities

auth-dns	1000	Global
www-nginx	800	Global
debian-mirrors	300	Global
sysadmin-minecraft	15	Global



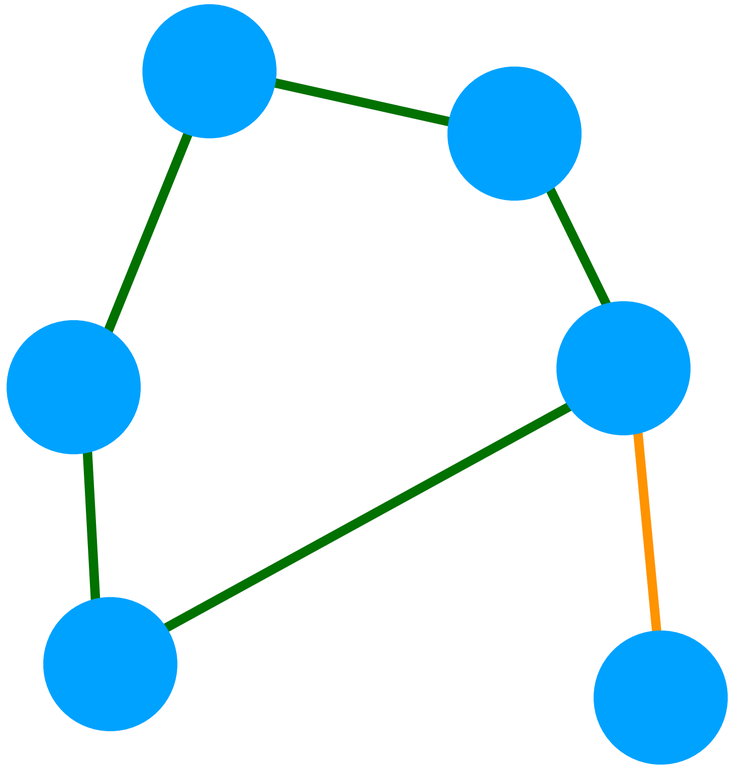
Service Traffic Class Priorities

auth-dns	1000	Global
www-nginx	800	Global
debian-mirrors	300	Global
sysadmin-minecraft	15	Global



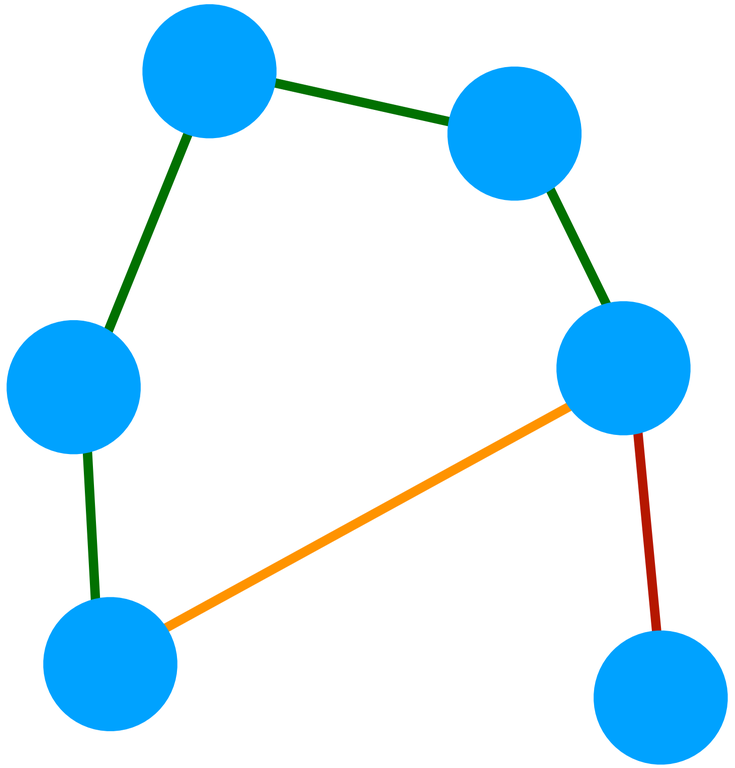
Service Traffic Class Priorities

auth-dns	1000	Global
www-nginx	800	Global
debian-mirrors	300	Global
sysadmin-minecraft	15	Ring-Only



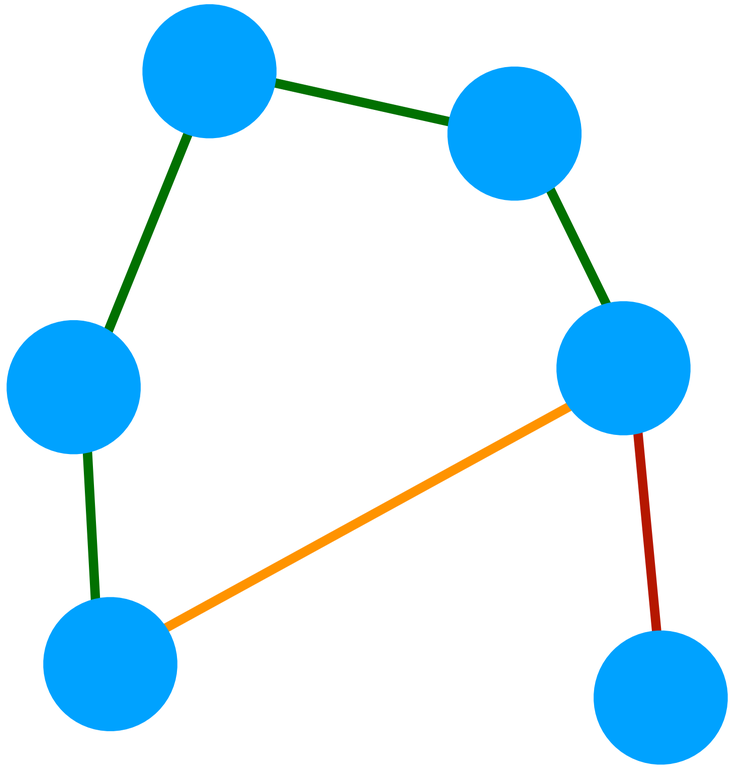
Service Traffic Class Priorities

auth-dns	1000	Global
www-nginx	800	Global
debian-mirrors	300	Global
sysadmin-minecraft	15	Ring-Only



Service Traffic Class Priorities

auth-dns	1000	Global
www-nginx	800	Global
debian-mirrors	300	Local
sysadmin-minecraft	15	Ring-Only



Group Creation

- Static
  - Groups are static, permanently configured on all IP Routers
  - IP Routers advertise the /48 - /24
- Dynamic
  - No pre-existing groups exist
  - Servers advertise the /48 - /24 when the /128 - /32 is advertised

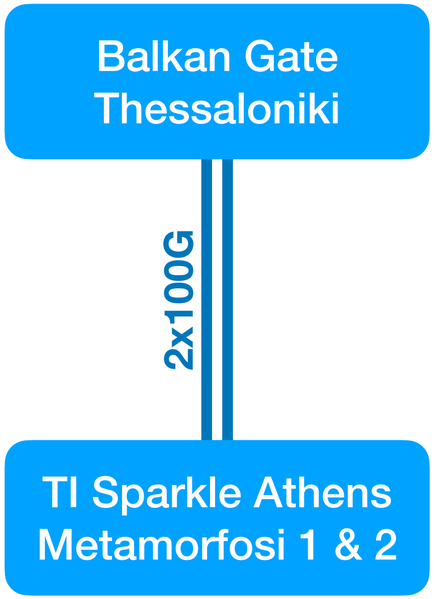
Incoming Traffic Volume

- Balance your L3 connectivity
  - A router in Amsterdam / Frankfurt likely won’t receive the same traffic as one in Athens (hotspots)
  - AMS failure can turn 90% peering in ATH to 5%
- Plan around cascading failures

One more thing…

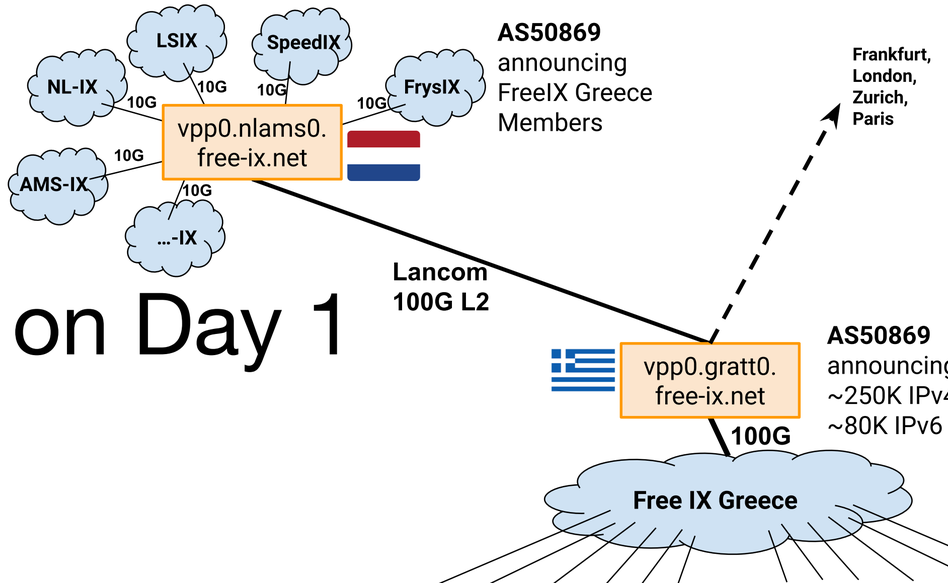
Free IX Greece

- New IXP in Greece (Thessaloniki & Athens)
  - One Common Peering LAN
- 1/10/25 GbE Ports Free
- info@free-ix.gr // free-ix.gr to connect!
  - Express interest for other DCs too — it helps :)



Free IX NET

- 100 GbE from Thessaloniki to Amsterdam
- Connection to several AMS IXPs as a Member
- 1 Gbps Free “Partial Transit” per Member
- Optional Service / Configurable
- ~333k Prefixes on Free IX Greece on Day 1



daknob@daknob.gov
@antonis@mastodon.social
linkedin.com/in/daknob/

Please replace “.gov” with “.net” in the e-mail address above so I can receive your message. :)