SRv6 USAGE IN MTC WEB SERVICES (MWS)

The Cloud Use Case

Boris Khasanov

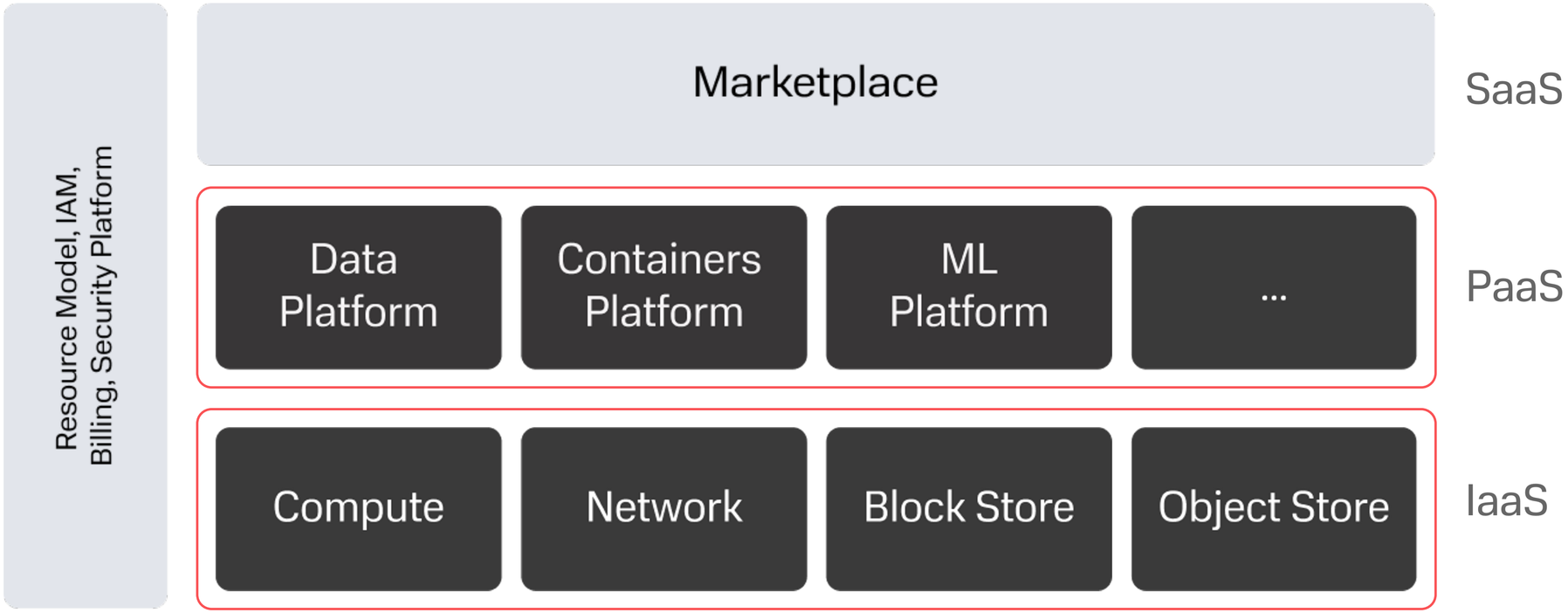
Senior network architect

Content

MWS Overview

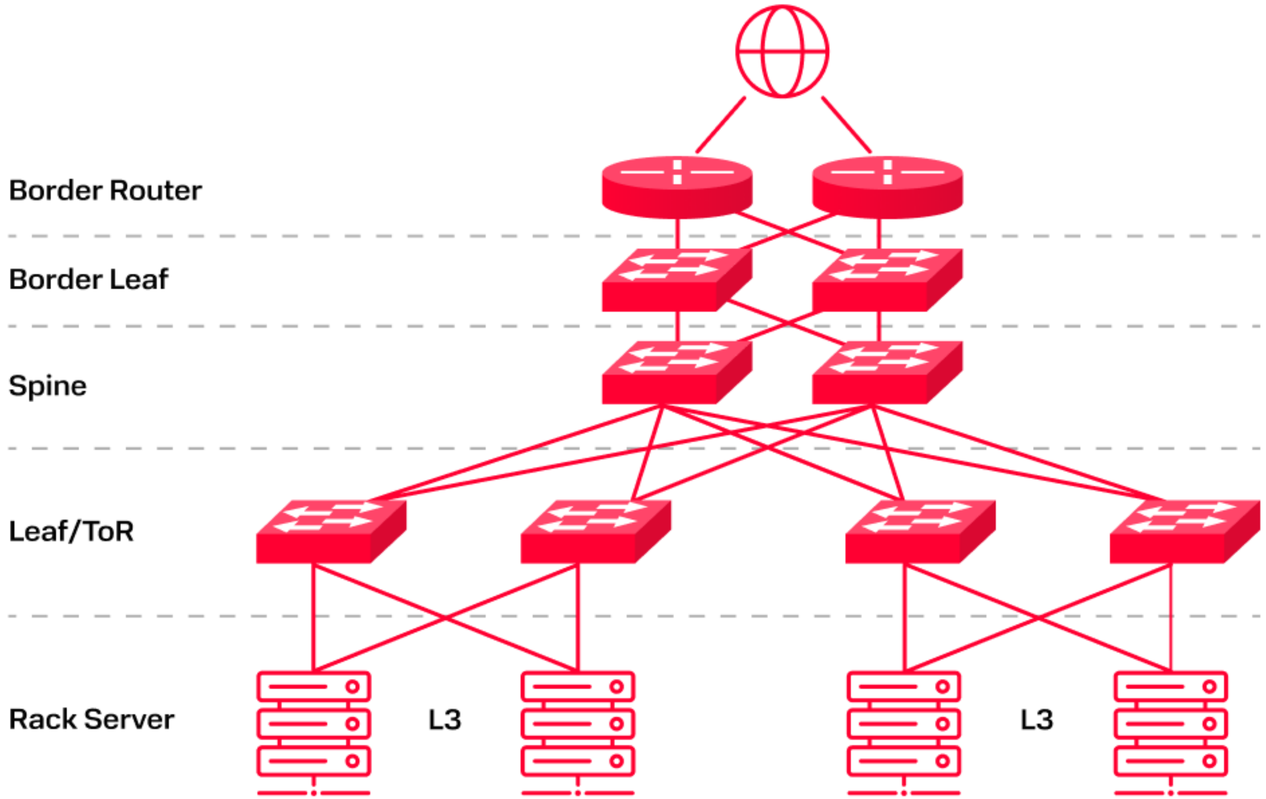
MTC Web Services (MWS) Overview

- MWS is the subsidiary company of MTC (mobile and wireline SP).
- MWS focuses on building the new (greenfield) vendor-independent Public Cloud.
- Milestones&Roadmap:
  - Q2 2025 - Object Storage in GA, IaaS in Preview.
  - Q3 2025 - IaaS in GA, PaaS in Preview.
  - Q4 2025 - PaaS in GA.

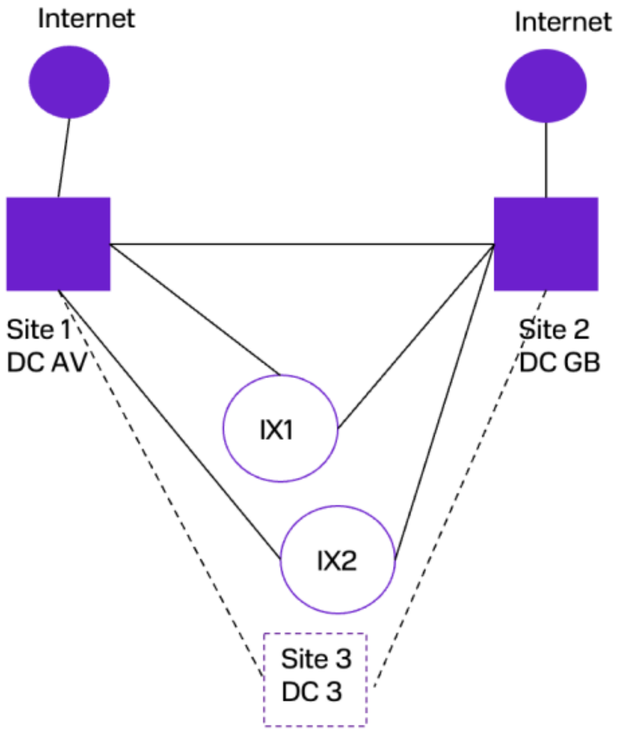


MWS Network baseline: Classic CLOS

CLOS with dual-homing servers



Two DCs and AZs now, 3rd in the roadmap

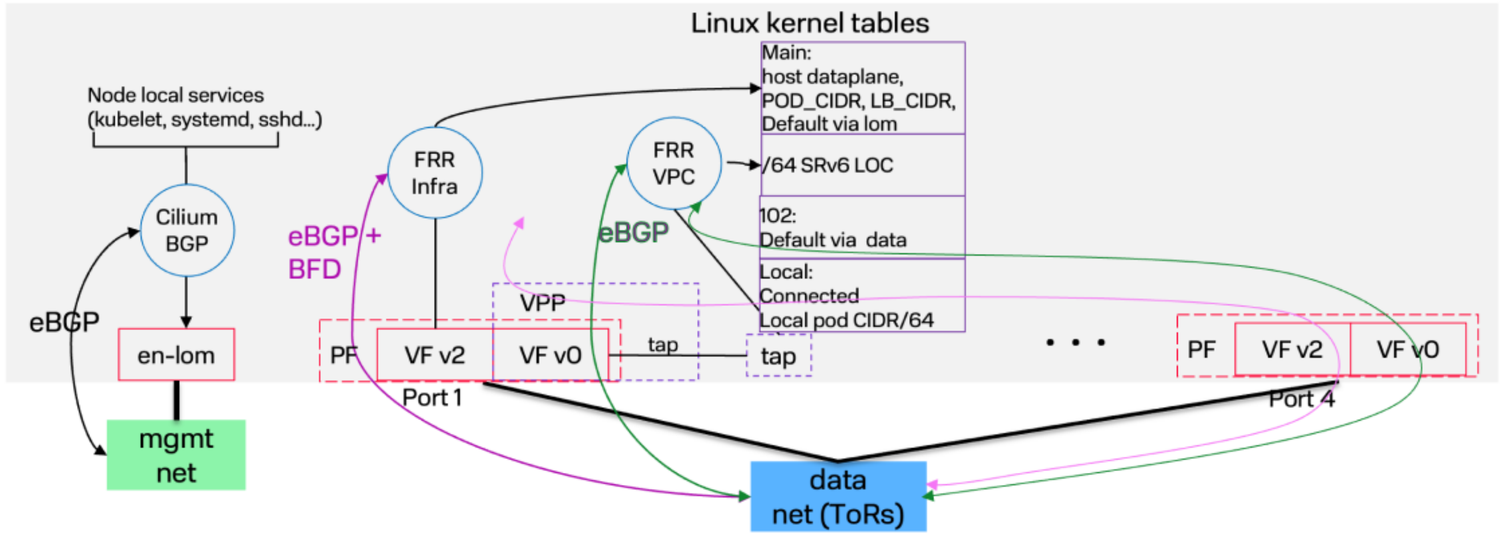


MWS Network Architecture

MWS Network Underlay: Compute Host connectivity and routing

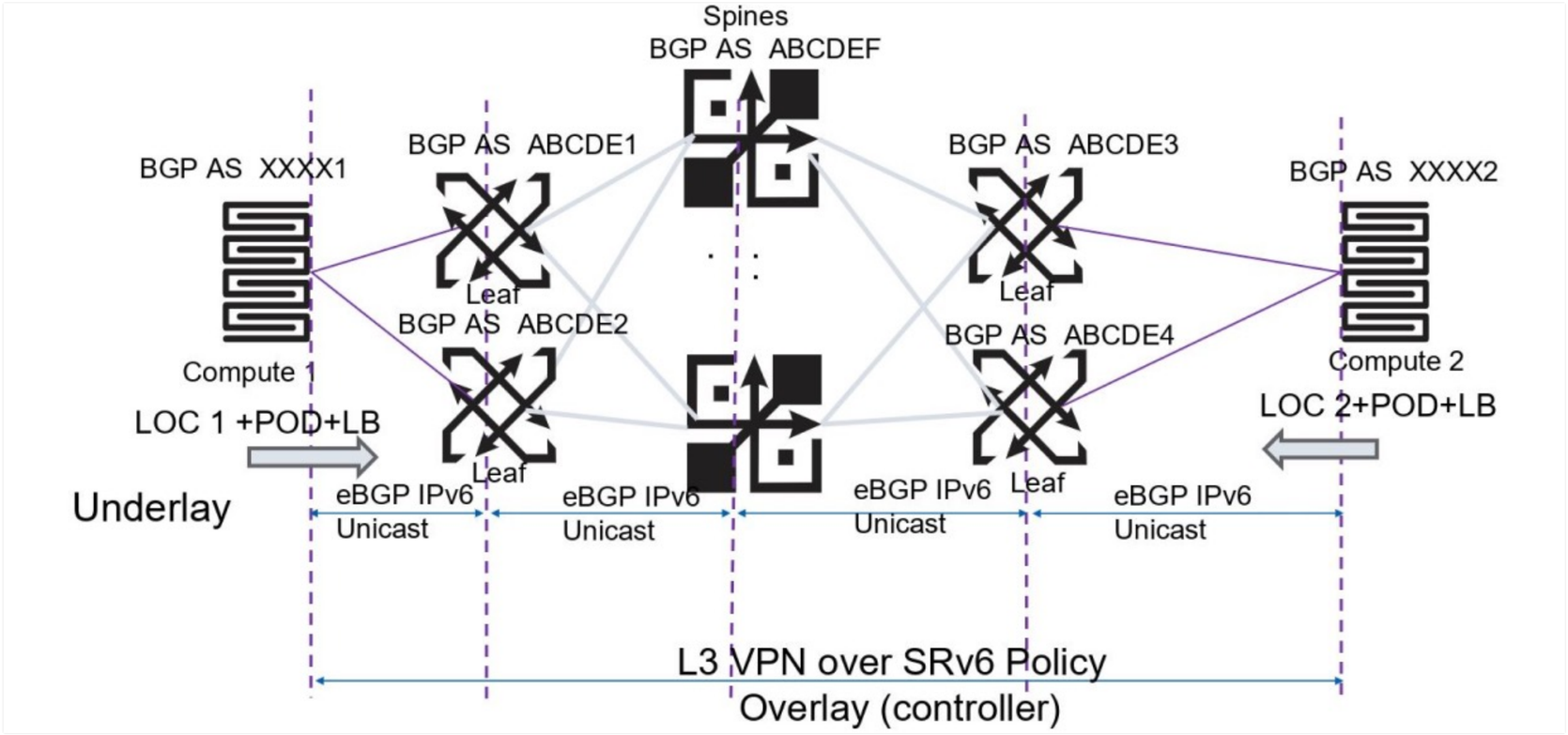
- Use K8s PODs (linux netns) with Cilium CNI.
- Each POD have unique IPv6 addresses.
- Service = VIP with LB for PODs (Cilium eBPF)
- Examples: Node Port, LB
- Physical separation of data and management
  lom intf used for a host and services management
- Dataplane forwarding on each port:
- SR-IOV divides physical intf into virtual functions (VFs)
- V0 VF used for VPP (dataplane forwarding also tap intf)
  V2 VF used for infrastructure traffic (stays in the kernel)

- Host control plane for K8s local services announced by Cilium BGP into the management network.
- Two different FRRs are used for the routing in the data network, FRR Infra for infrastructure prefixes
- (POD_CIDR, LB_CIDR, LOOPBACK) and FRR VPC for SRv6 LOC (IPv6 Unicast) and some other cloud internal services.
- Thus there are 2 BGP sessions per each data port plus 1 management.
- Host uses linux kernel tables for different prefixes.



MWS Underlay IP-fabric

- eBGP only based IP-fabric.
- Routing on hosts (SRv6 locator advertisement by FRR).
- IPv6 forwarding in the dataplane (ECMP).
- VRFs configs and routes per host for an overlay are provisioned from internal SDN-controller.
- Initially Route Servers were used for VPN BGP full mesh, now SDN-controller is used for VPN prefixes distribution.



MWS SRv6 Cloud
Use case

Why SRv6 for a Cloud?

Previous Cloud network architecture

- Proprietary vendor X with HW-based overlay and vendor Y virtualization and SW-overlay
- BP/VXLAN vendor Z with HW-based overlay and vendor Y virtualization and SW-overlay

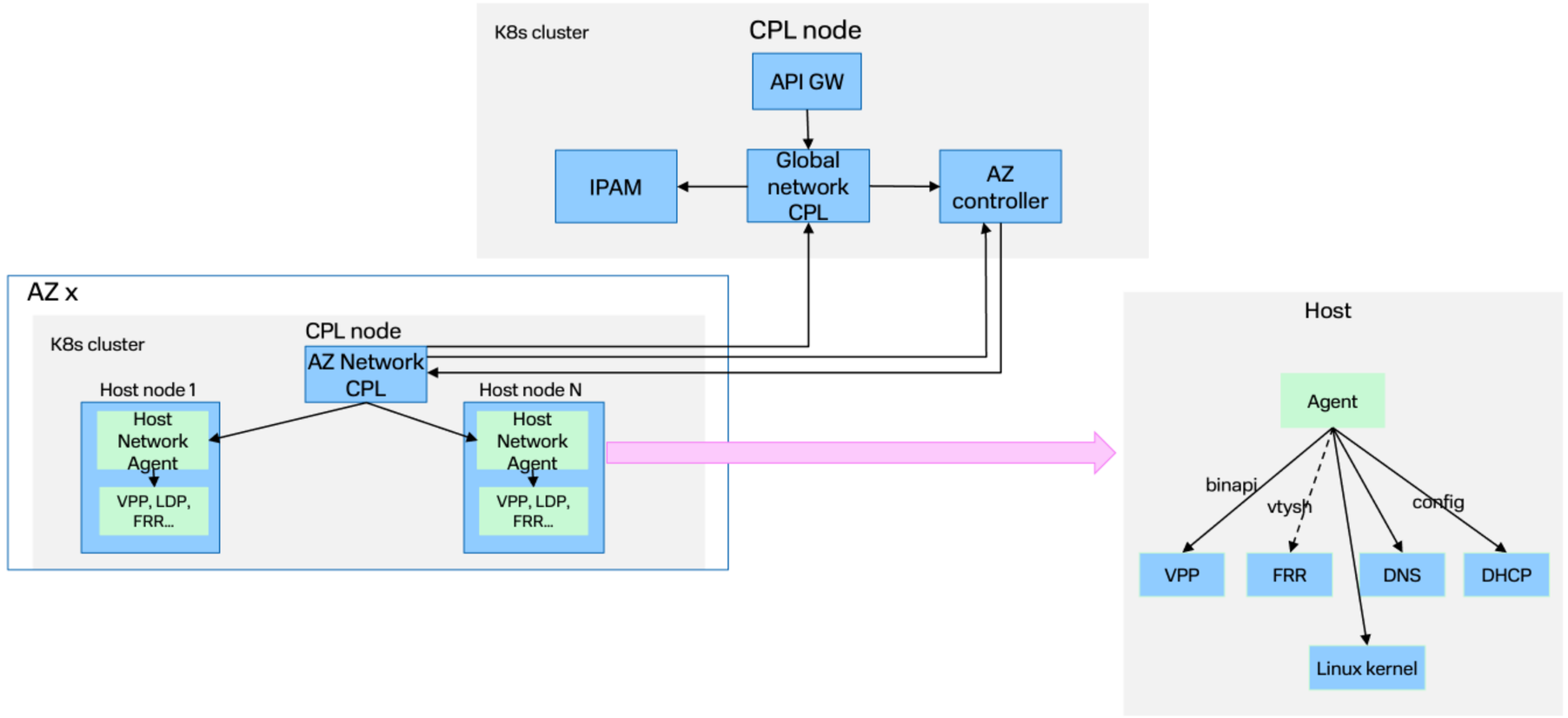
Vendor lock-in, incompatibility, issues with observability and operations...

SRv6-based

Allows to make SIMPLE separation: pure IPv6 underlay and SRv6-based overlay on hosts.

- No more switches vendors lock-in and proprietary architectures...
- SRv6 programmability to use those Network Features which we need at the moment
- Simplicity in the network, complexity on the hosts and in virtualization control-plane.
- IP-friendly (IPv6).
- SRv6 can be used as inside DC as outside for DCI up to Edge Computing.

MWS Centralized Virtualization and SRv6 Control Plane: CPL



The CPL (aka SDN-controller) is responsible for provisioning and tracing all VM/VPC related parameters including SRv6.

Simplified SRv6 Overlay work procedure

1. Host agent reads the assigned host prefix (i.e. 2a02:5501:0:aaa1/64) from the config file and sends it towards CPL together with SRV6 LOC (i.e. 2a02:5501:0:aaa1:100::/80).

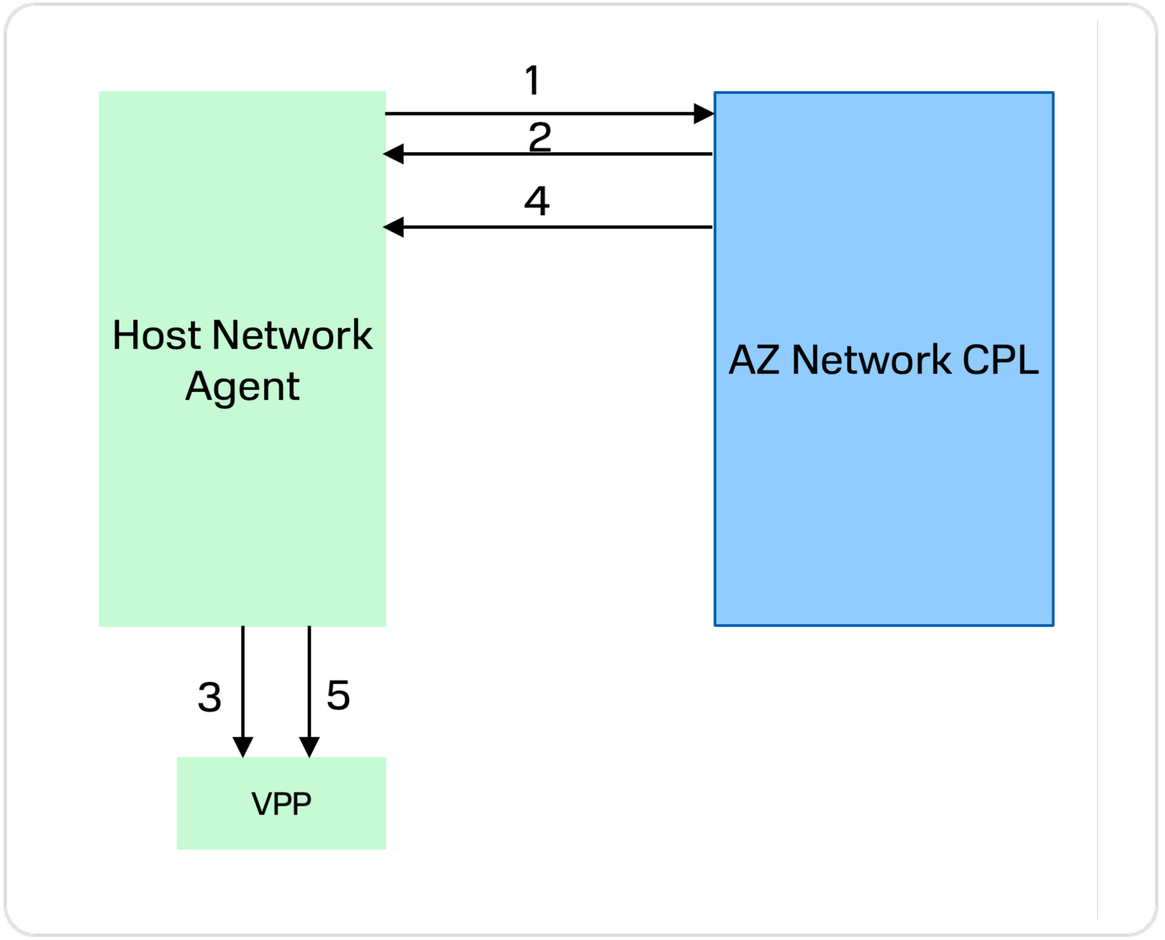
2. A CPL replies with:

- 2.1. Network attachment ID (or sidClass, some number, i.e. 1000) that is an identity of a network for a VM.

1. Host Agent adds Network attachment ID/sidClass to LOC for a Local SID generation for VPP (i.e. END.DT4, 2a02:5501:0:aaa1:100:1000::/128).

2. CPL also sends VPN Prefixes List (i.e. pairs of IPv4 with SRv6 SID as NH) back to the Host Agent.

3. Network Agent creates SR Policy per each remote SID with crafted BSID (Remote SID+sidClass+Policy ID). SR Policy have one SL with one SID.

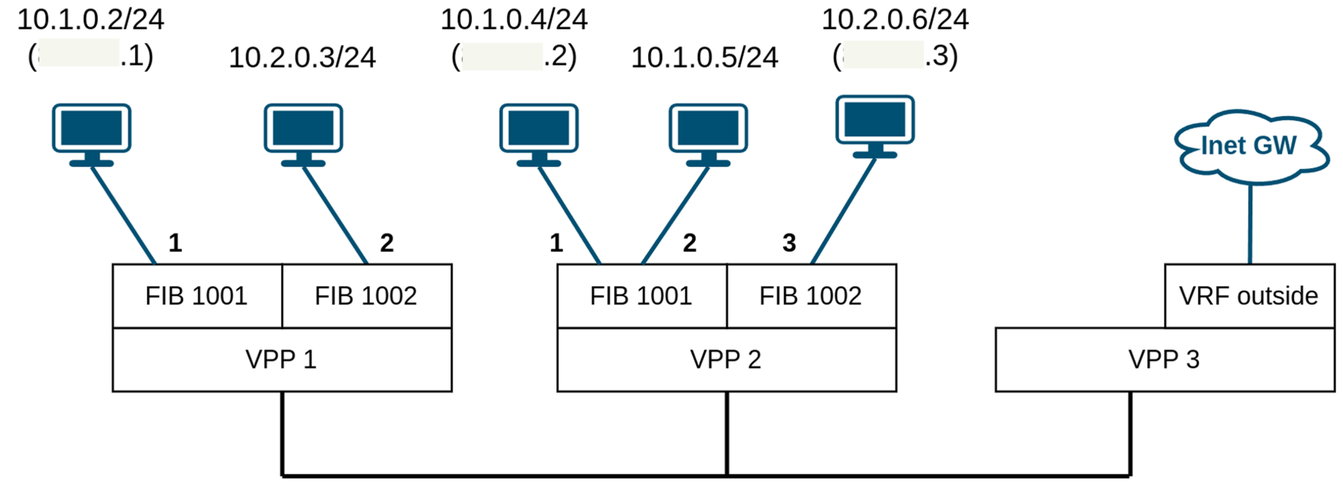


SRv6 Configuration Template Example*

```
  "spec": {
    "networks": [
      {
        "name": "projects/glko-project/networkAttachments/glko-network",
        "version": 1,
        "vrfId": 320917,
        "srv6Options": {
          "sid": 4954,
          "routeDistinguisher": "192.168.4.181:4954",
          "routeTargetImport": [
            "320917:1"
          ],
          "routeTargetExport": [
            "320917:1"
          ]
        },
        "hostNetworkNumber": 859,
        "mtu": 1500,
        "subnets": [
          {
            "name": "projects/glko-project/networkAttachments/glko-network/subnetAttachments/glko-subnet",
            "version": 1,
            "cidr": "192.168.11.0/24",
            "gatewayAddress": "192.168.11.1/24",
            "dhcpOptions": {
```

*This is just short extraction, other parameters skipped for the sake of simplicity.

Simplified SRv6 Overlay details



VPP 1

IPv4 FIB table 1001
```
0.0.0.0/0    via NAT(outside)
10.1.0.2/32 via vhost1
10.1.0.4/32 via SR 2a:22::11:0
10.1.0.5/32 via SR 2a:22::11:0
```

IPv4 FIB table 1002
```
10.2.0.3/32 via vhost2
10.2.0.6/32 via SR 2a:22::12:0
```

IPv4 FIB table 555 (ouside)
```
0.0.0.0/0    via SR 2a:33::55:0
 .1/32 via NAT(VRF 1001)
```

FIB table (underlay)
```
2a:22::0/64 via VPP2 (via switch)
2a:33::0/64 via VPP3 (via switch)
```

Localsids
```
2a:11::11:0/128 => DT4 table=1001
2a:11::12:0/128 => DT4 table=1002
2a:11::55:0/128 => DT4 table=555
```

SID format

Host \ / VRF

2a:22::12:0

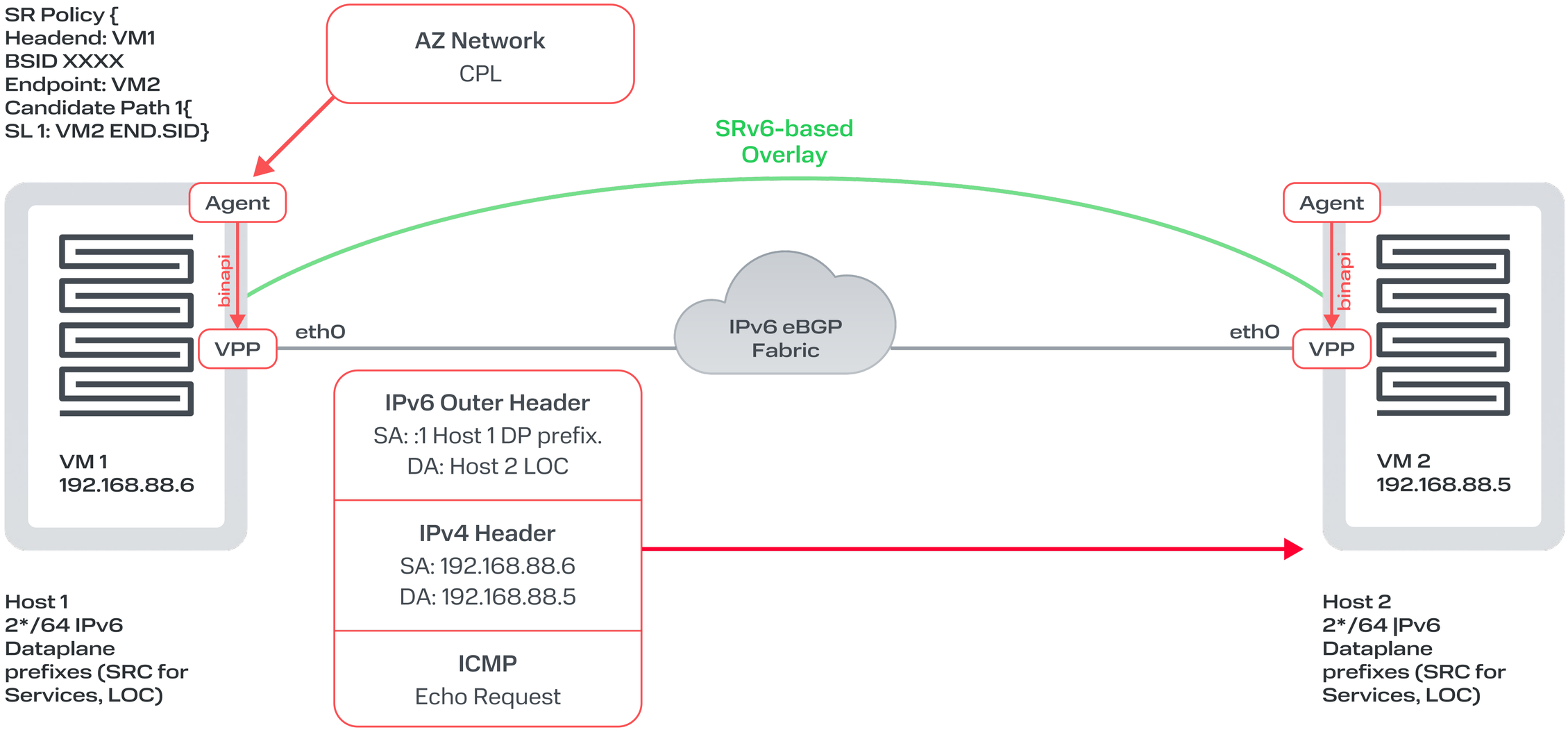
VPP 3

IPv4 FIB table 555 (outside)
```
0.0.0.0/0    via lookup-in OUTSIDE
 .1/32 via SR 2a:11::55:0
 .2/32 via SR 2a:22::55:0
 .3/32 via SR 2a:22::55:0
```

FIB table (underlay)
```
2a:11::0/64 via VPP1 (via switch)
2a:22::0/64 via VPP2 (via switch)
```

Localsids
```
2a:33::55:0/128 => DT4 table=555
```

SRv6 Forwarding Example



Next Steps

Next steps

01

Readiness for GA

Testing and fixing all remaining issues.

02

Border HA

Implementing a solution for borders HA with centralized CP.

03

SR Policy extension

Research for more advanced SR Policy implementation (more SLs, more SIDs, wECMP)

04

Active PM & Incident Management

Research and Implementation of active SRv6 PM.

Developing Labeling system for the network metrics during incidents (using Yang-model as reference) for further post-mortem analysis. Using LLMs for network anomaly detection.

References

My recent article on Linkedin : https://www.linkedin.com/pulse/vpp-binapi-magic-new-way-provision-sr-policy-boris-hassanov-aofef

Thank you!

Boris Khasanov

Senior network architect

Backup

HOST IPV6-ADDRESSES

Host control plane address	Assigned for lom intf
Two host dataplane prefixes	/64 used as SRC for services and /64 for SRv6 LOC
POD_CIDR	/64, Cilium IPAM assigns IPv6 addresses for the PODs from that prefix
LB_CIDR	/64 for cluster