MTC WEB SERVICES (MWS)
NETWORK INCIDENT

Postmortem

Boris Khasanov

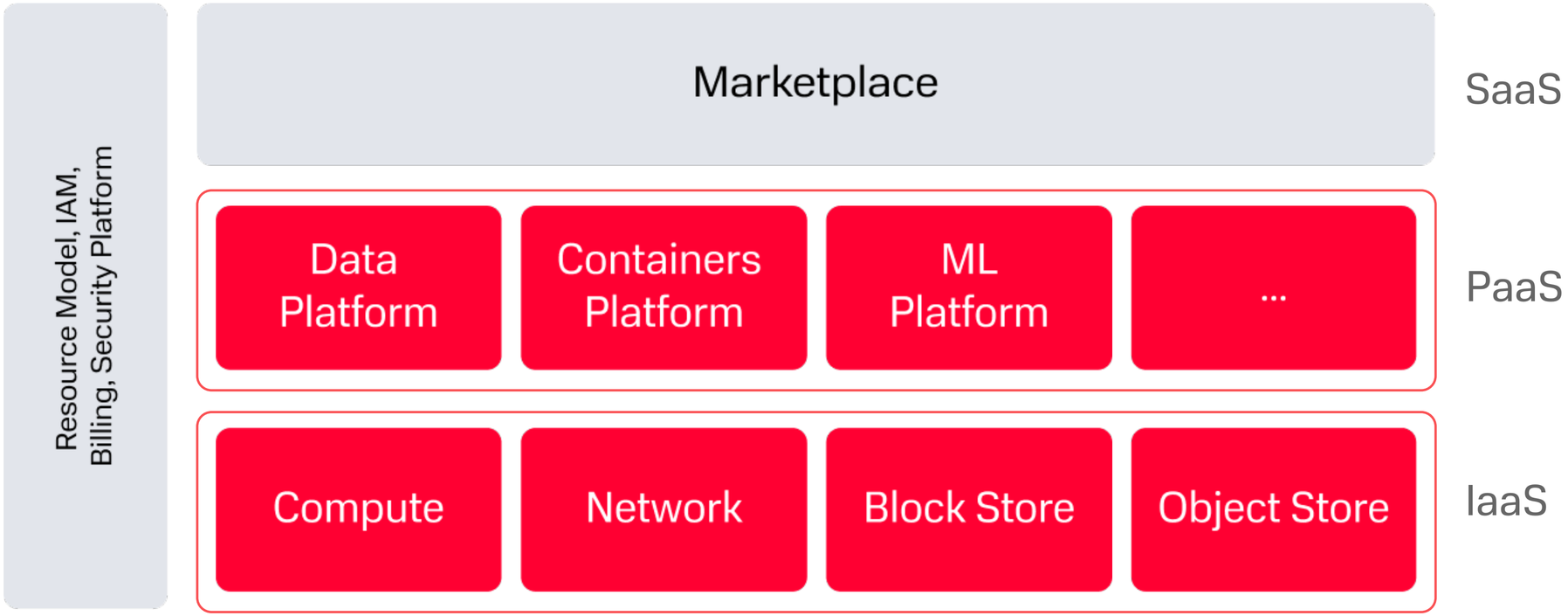
Senior network architect

Content

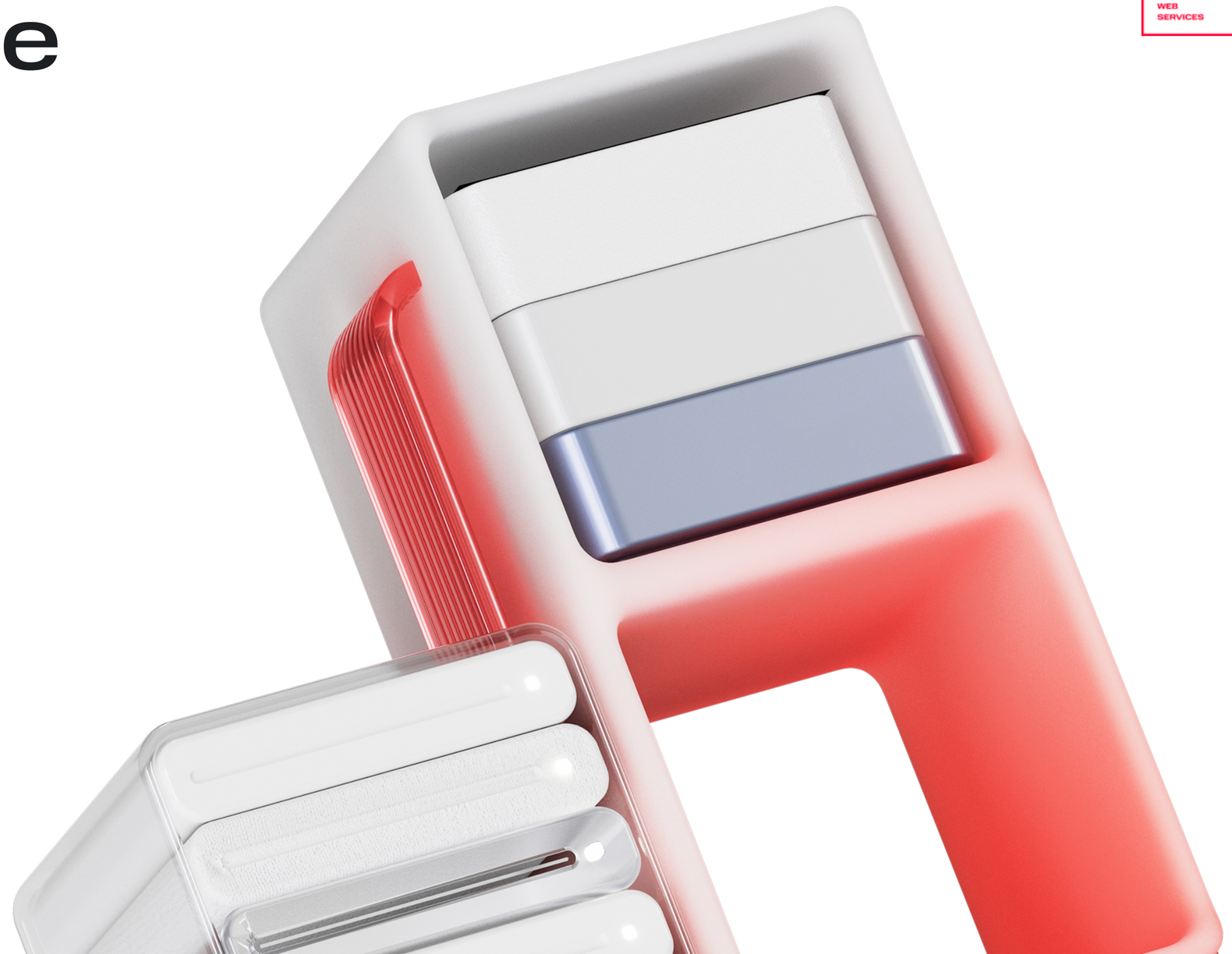
MWS Overview

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.

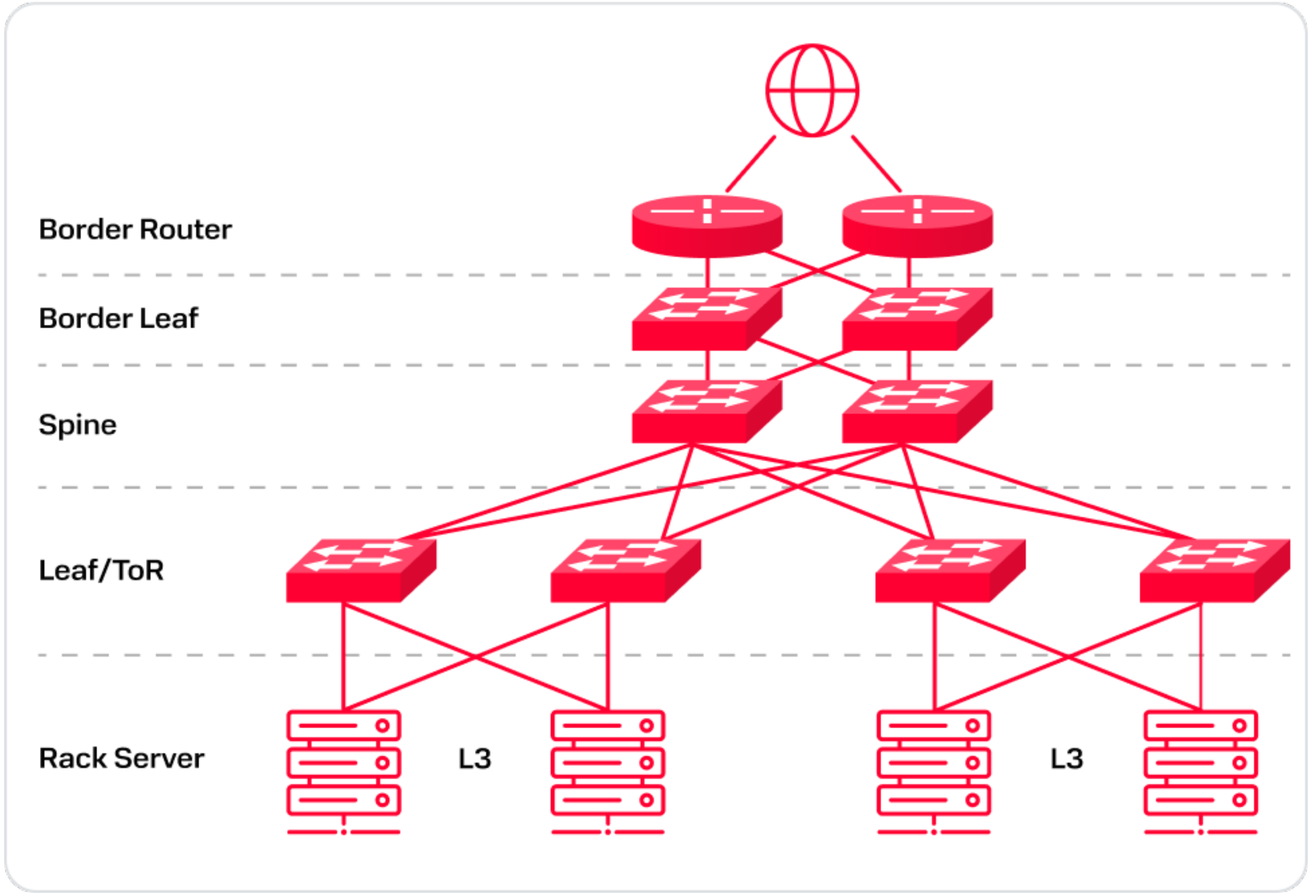


MWS Network Architecture

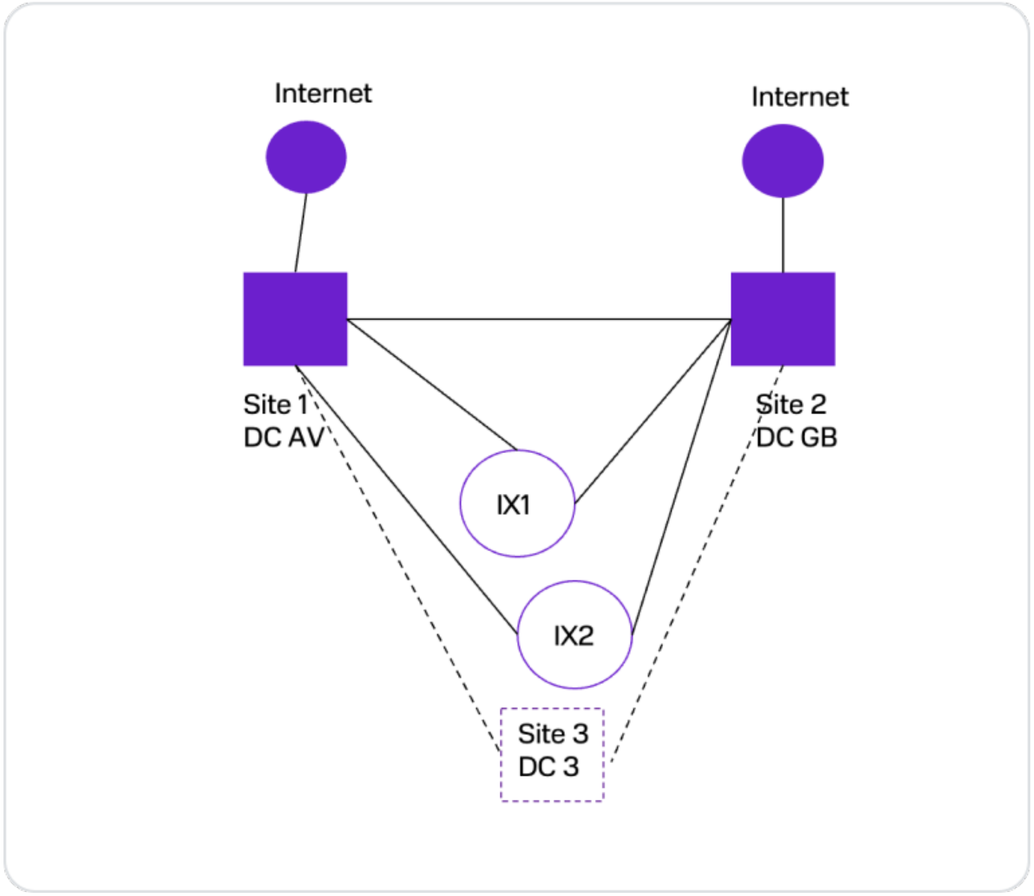


MWS Network baseline: Classic CLOS

CLOS with dual-homing servers

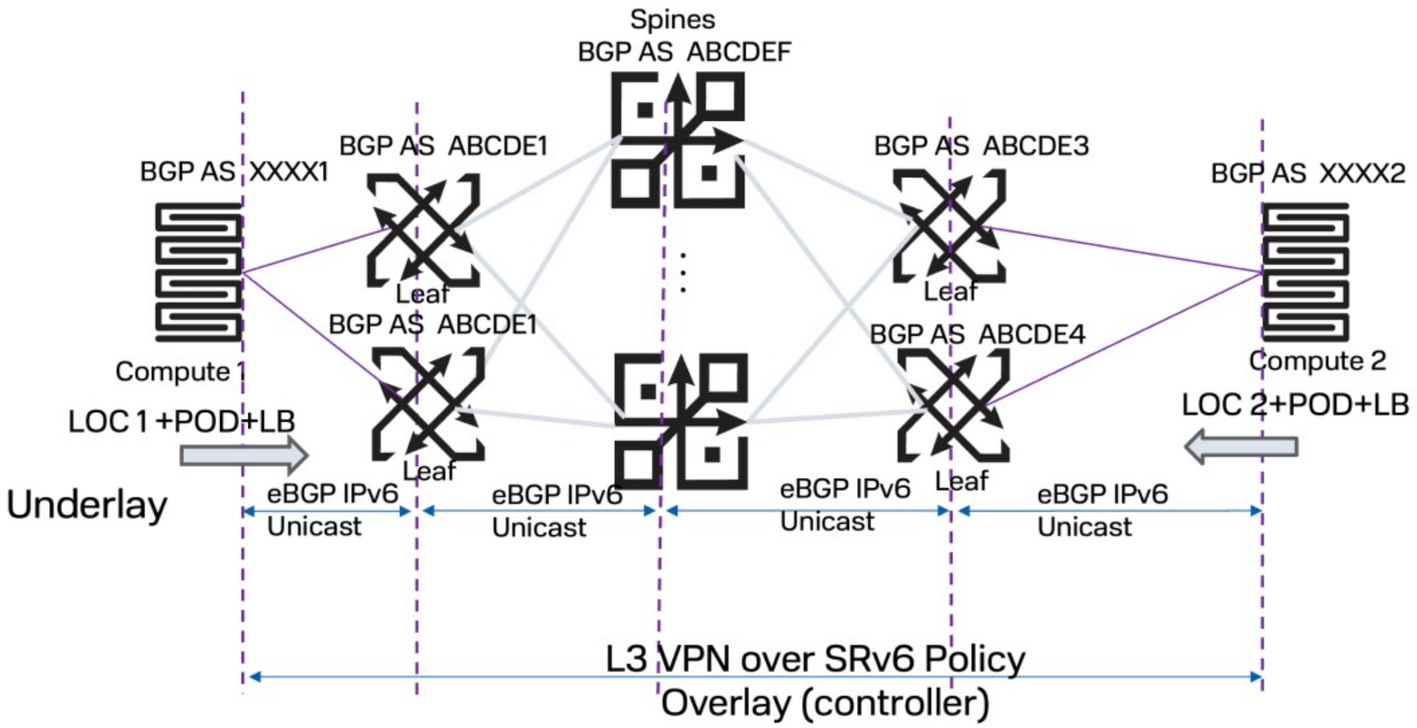


Two DCs and AZs now, 3rd in the roadmap



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.

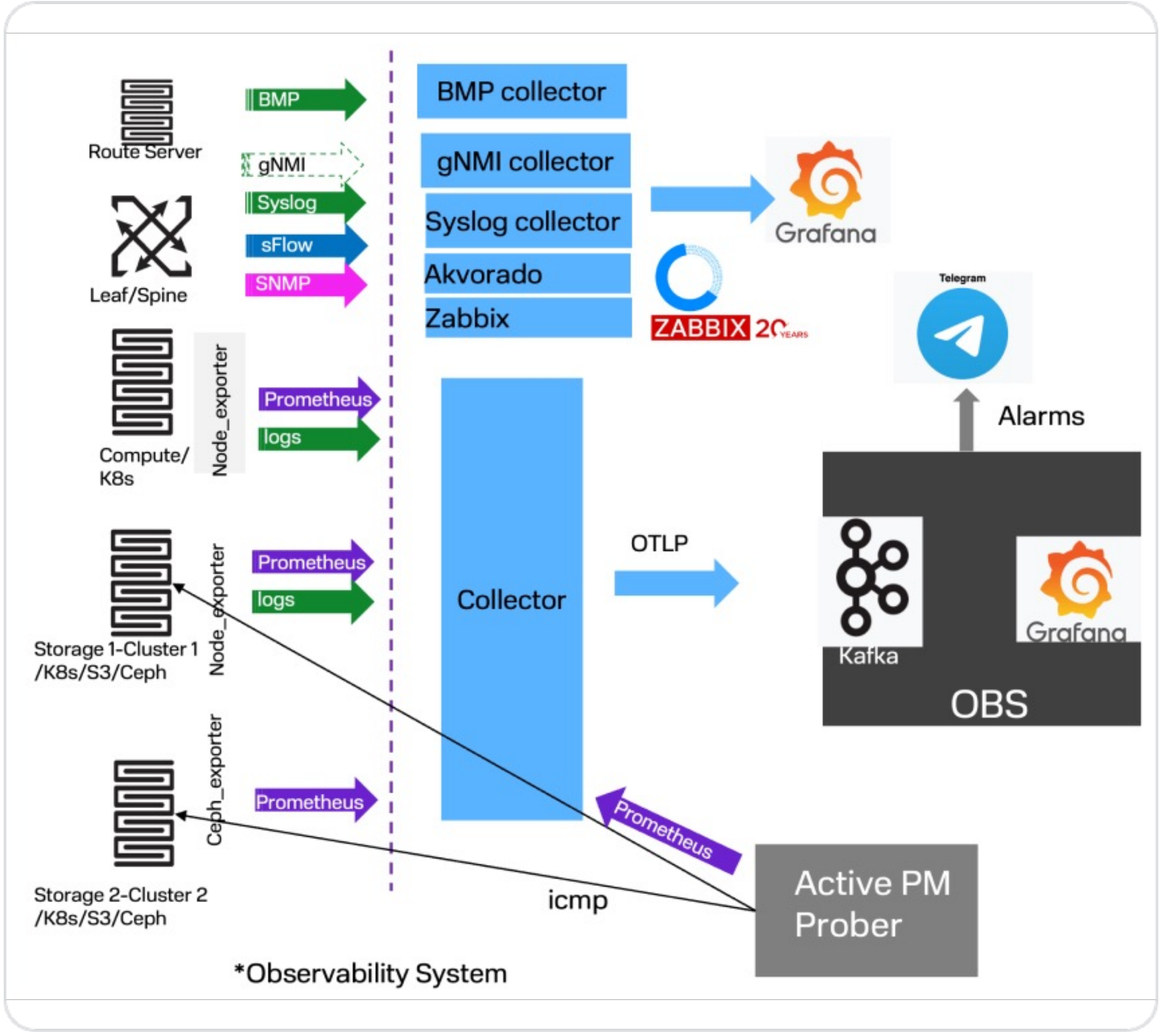


Observability overview

Data collection and OBS* architecture

At the moment we have **three** monitoring loops:

- OBS for host/K8s clusters
- Number of collectors for the network monitoring.
- Active PM Prober (pinger) for storage (S3/Ceph) end-points liveness detection. The results are sent as metrics into the OBS.

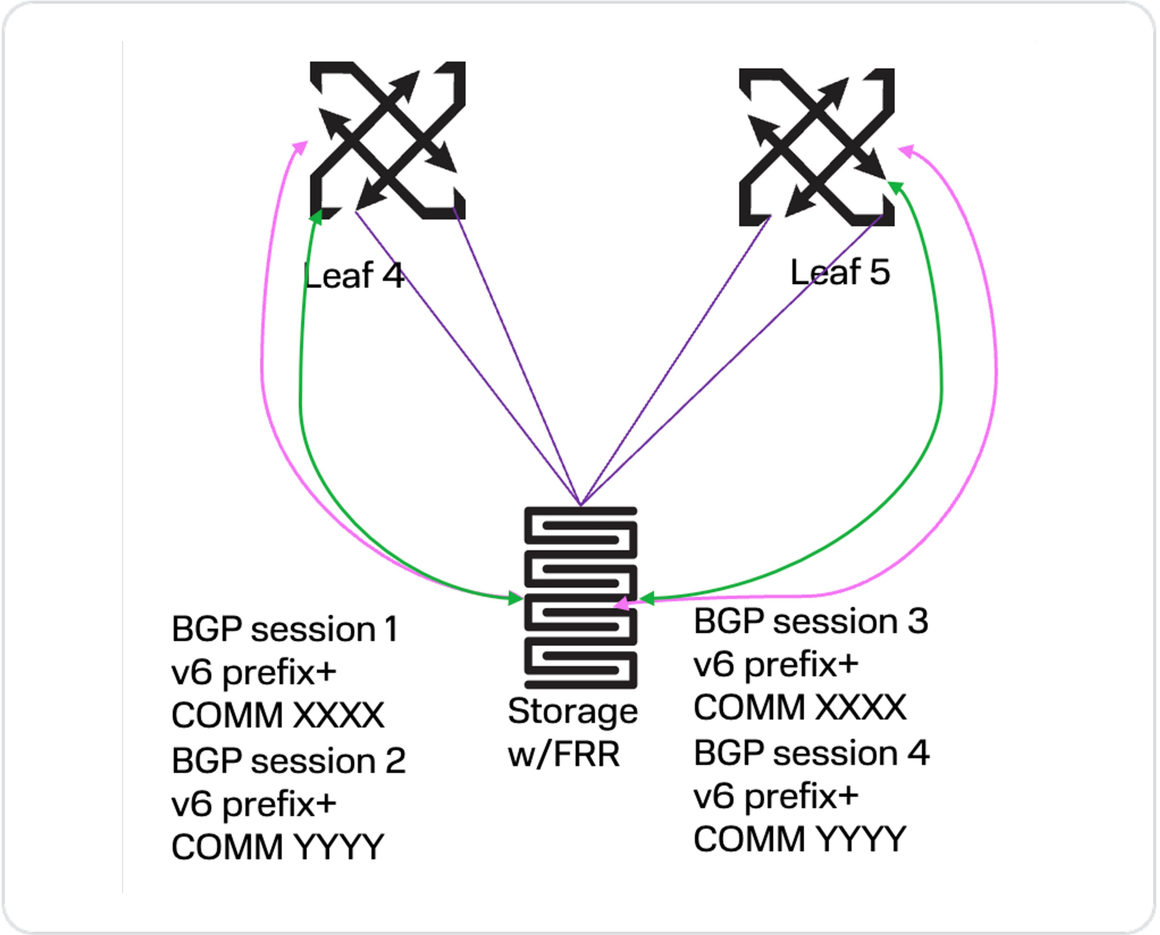


Network incident description

Does dual-homed server connection provide 100% redundancy?

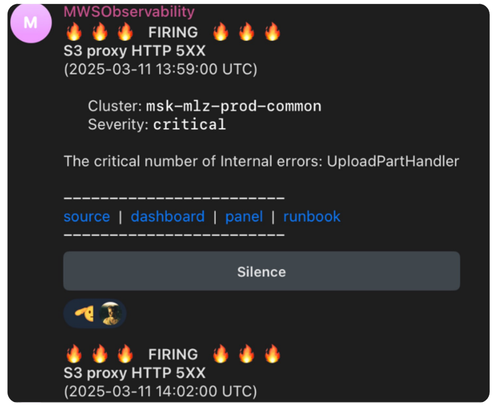
- Storage host have 4x100G Ethernet ports (2 service ports plus 2 sync ports).
- Dual homing into 2 Leaf switches (2x2 ports).
- **Two** different BGP sessions per each Leaf (for service prefix and for Ceph sync).
- Leaf 5 was scheduled for a planned
- swap during a maintenance window.

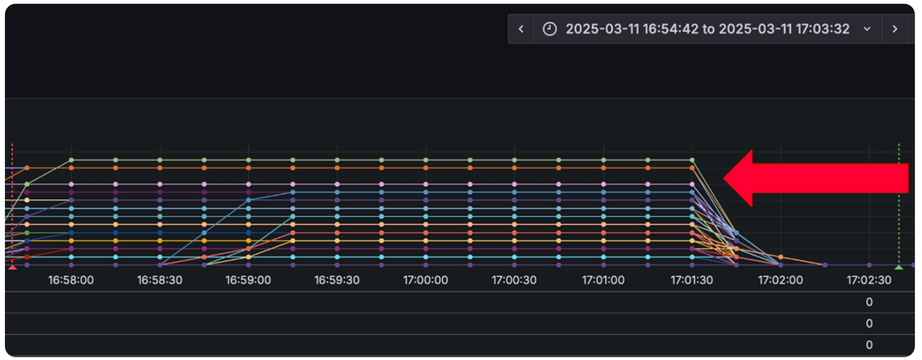
What can go wrong?



Timelines, the 1st signal

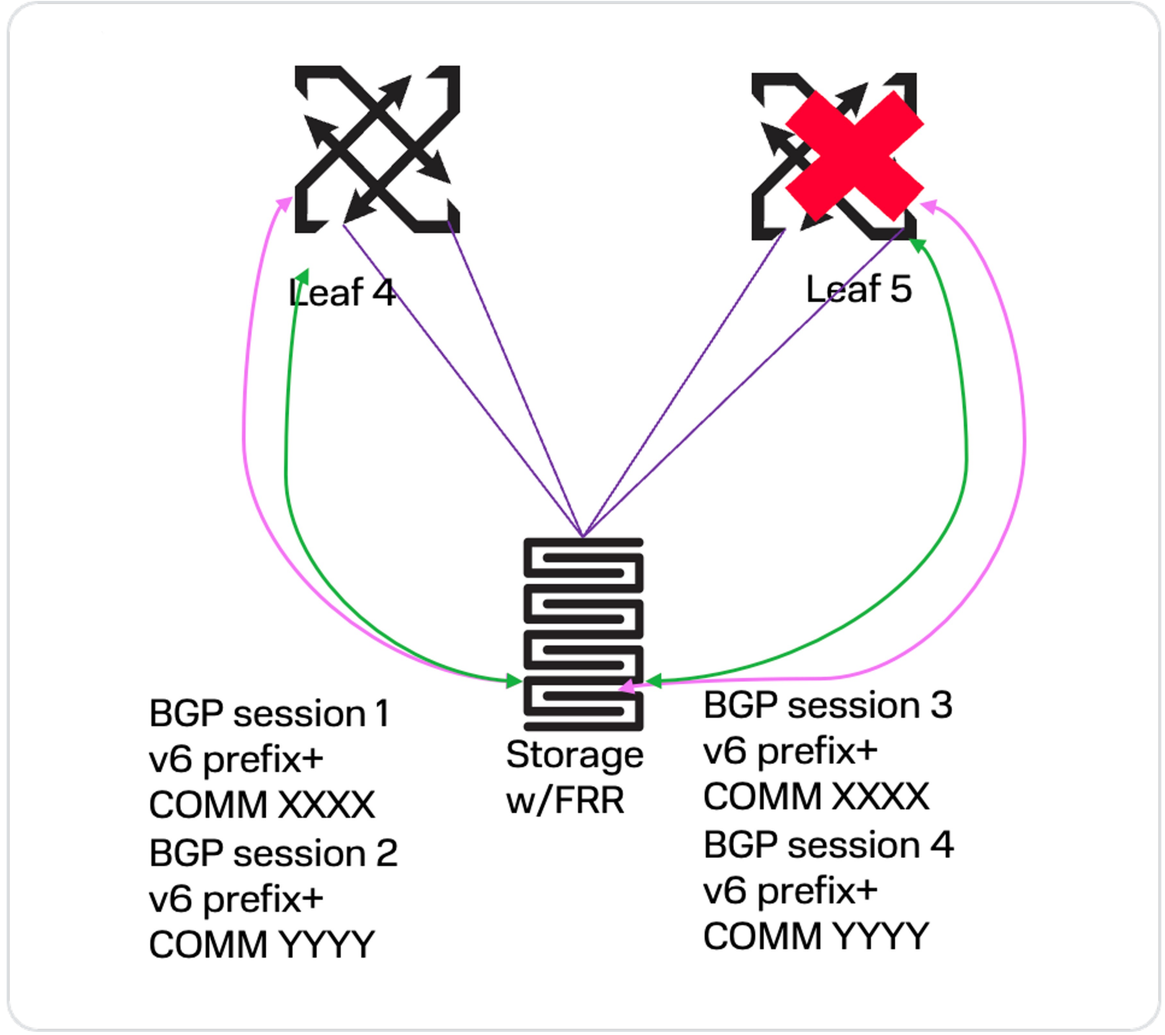
- About 4.59pm local time we've got the alarm signal from Active PM Prober that S3 End-Point is not available.
- The colleague on duty asked the networking team for help.
- That was the maintenance window time when Leaf 5 was replaced.
- Hm-m...
- How the host(s) became inaccessible if it is dual homed?





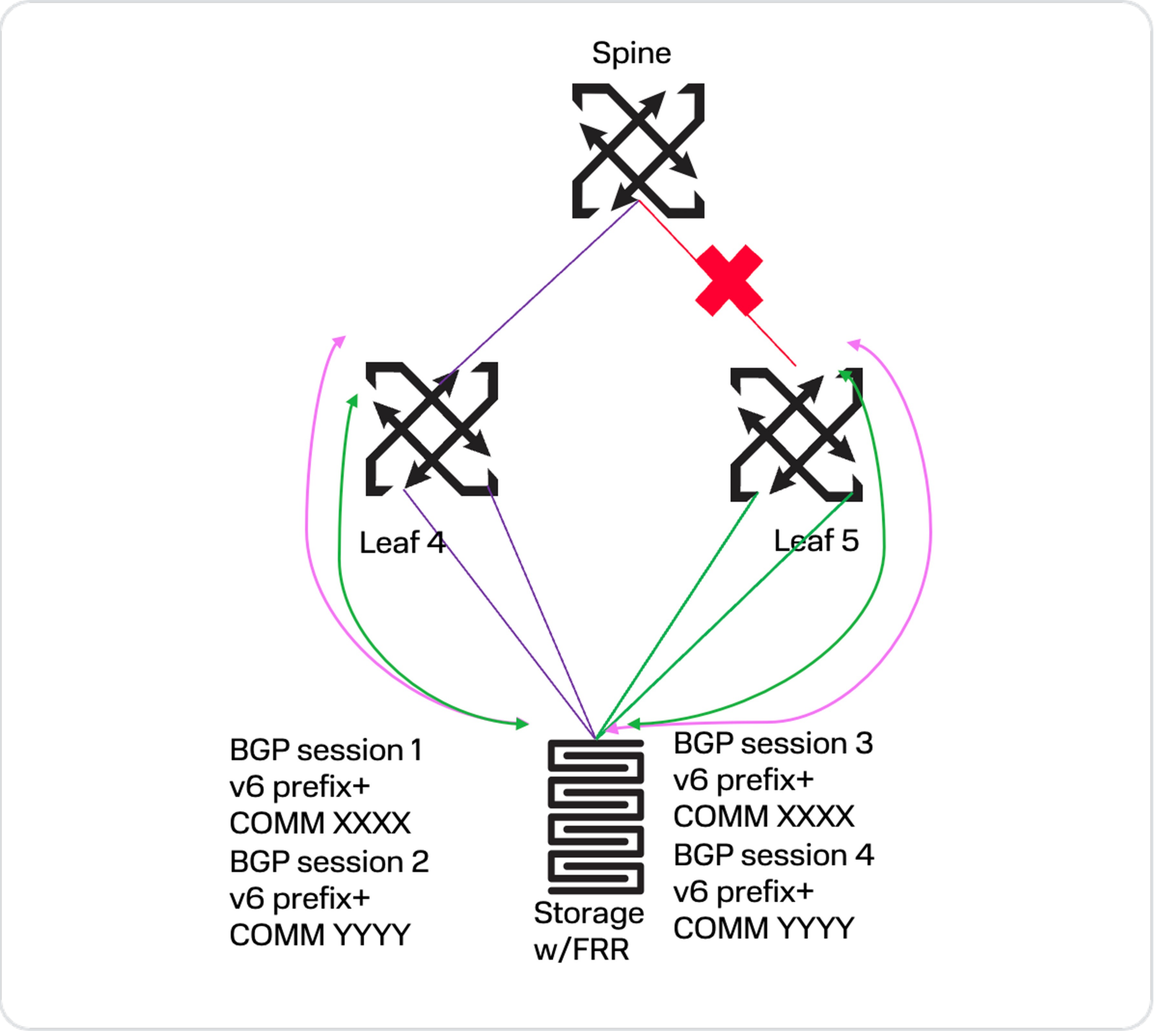
Correlation with the switch swap

- Obviously, we immediately found
  that at that time we had the switch swap work.
- But why traffic didn't flow thru the Leaf 4?



Switch swap details

- After the new Leaf 5 switch came up, firstly downlink ports towards the hosts became UP.
- As soon as downlink ports were UP the host started to send traffic.
- Then Uplink ports became UP.
- So — traffic was blackholed during that time (~3 mins).
- Why?



Mystery clarification...

- We have static default routes pointing towards Leaf 5 for different routing tables.
- Why? Because we have default route in GRT towards LoM interface.
- It is used for automated host provisioning etc.
- Thus after host data ports and BFD went UP it started to send a traffic towards Leaf 5 even his UL ports were Down.
- The mystery of the incident is clarified!

```
ipv6 route ::/0 fd00:99:a:8::1 210 table 101 bfd
ipv6 route ::/0 fd00:99:b:7::1 210 table 101 bfd
ipv6 route ::/0 fd00:99:a:7::1 210 table 102 bfd
ipv6 route ::/0 fd00:99:b:8::1 210 table 102 bfd
```

How to fix the problem?

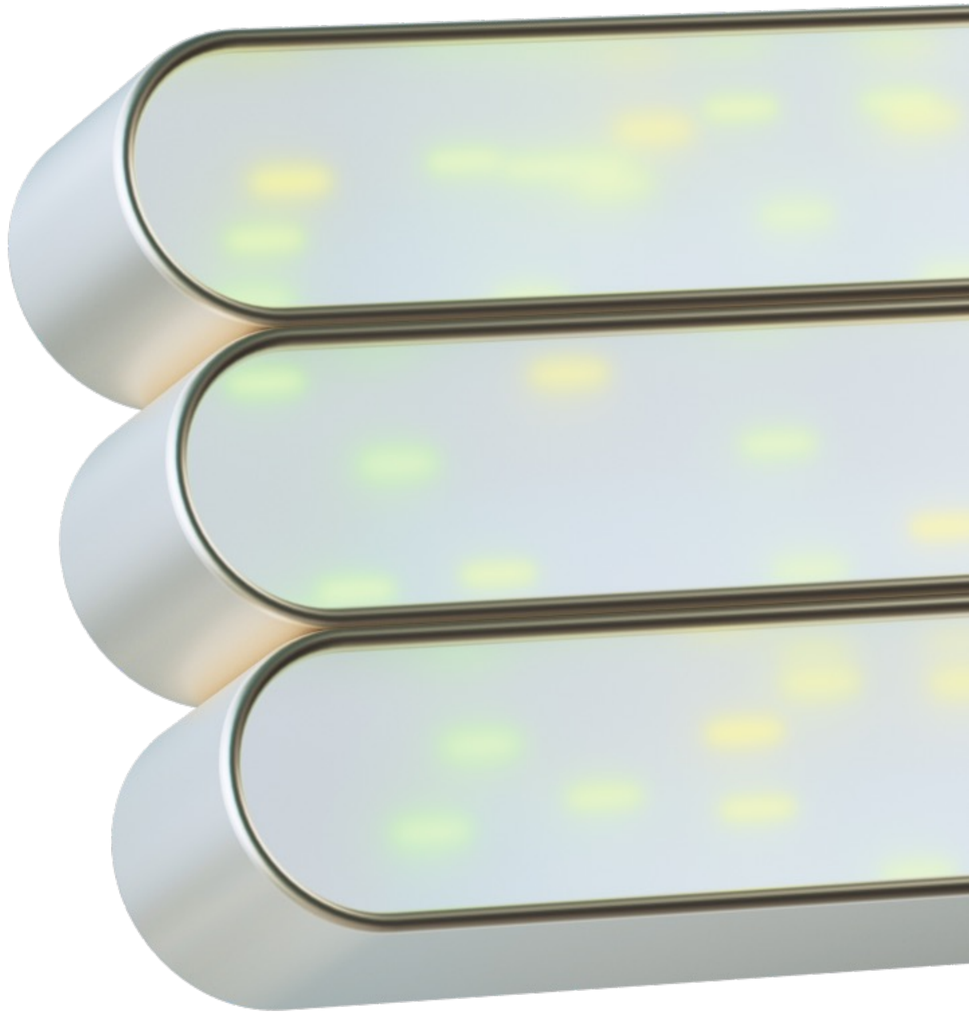
What's next ?

Fixing the problem

- Unfortunately the switches vendor does not have yet a feature which prevents bringing UP DL ports while UL ports are not UP.
- Thus we move away from the static default routes scheme towards BGP originated default route towards the hosts.
- But here we met another "the chicken or the egg" problem: BGP vs. BFD, who's the first to bring up the session?
- That problem and solution are discussed in draft-ietf-idr-bgp-bfd-strict-mode .

Meanwhile the workaround:

- During the automated switch provision phase, just after an installation, we will generate BGP config towards the hosts with BGP sessions turned off (neghbor x..x.x shutdown).
- Will turn them on using special hook after HW installation team will confirm that all ports are UP.
- So we will have BGP originated default on storage hosts.



Next steps

01

Bringing UP BMP on the switches

Getting this feature from vendor, testing and running.

Choosing the exact collector for the production.

02

gNMI telemetry on the switches

Moving towards the NOS version which supports gNMI at scale. Moving gNMI from PoC to prod.

03

Traffic "coloring" on the hosts for E2E visibility

Bringing some tools like alternative marking, PoC then prod.

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.

Thank you!

Boris Khasanov

Senior network architect