sipgate

~~Bye Bye VLANs.~~ Hello Layer 3 Connectivity.

Rudolph Bott

sipgate GmbH

2026

Hi, I'm Rudi - and I've spent years fighting layer 2 technologies.

Network & infrastructure at **sipgate**, Düsseldorf

We run business telephony at scale.
Both VoIP and mobile.

99% of our servers run Debian 💙

→ **What I do all day**

Operate AS15594, build and maintain Debian based infrastructure, hack on Ganeti.

→ **What I also do**

DadOps, hiking in the mountains and taking photos.

→ **Disclaimer**

The next 30 minutes are about datacenter networking, not so much home labs :-)
Also: this targets single-tenant networks.

01 **The Layer 2 status quo**
A quick recap.

02 **Why overlays didn't fix it**
VXLAN promised L2 anywhere - we got two networks to debug.

03 **The Layer 3 alternative**
Every server a router. Every service a /32.

04 **The building blocks**
BGP unnumbered, dummy interfaces, ECMP, host firewalling.

05 **In the wild**
Generic apps, Kubernetes CNI, Ganeti VM clusters.

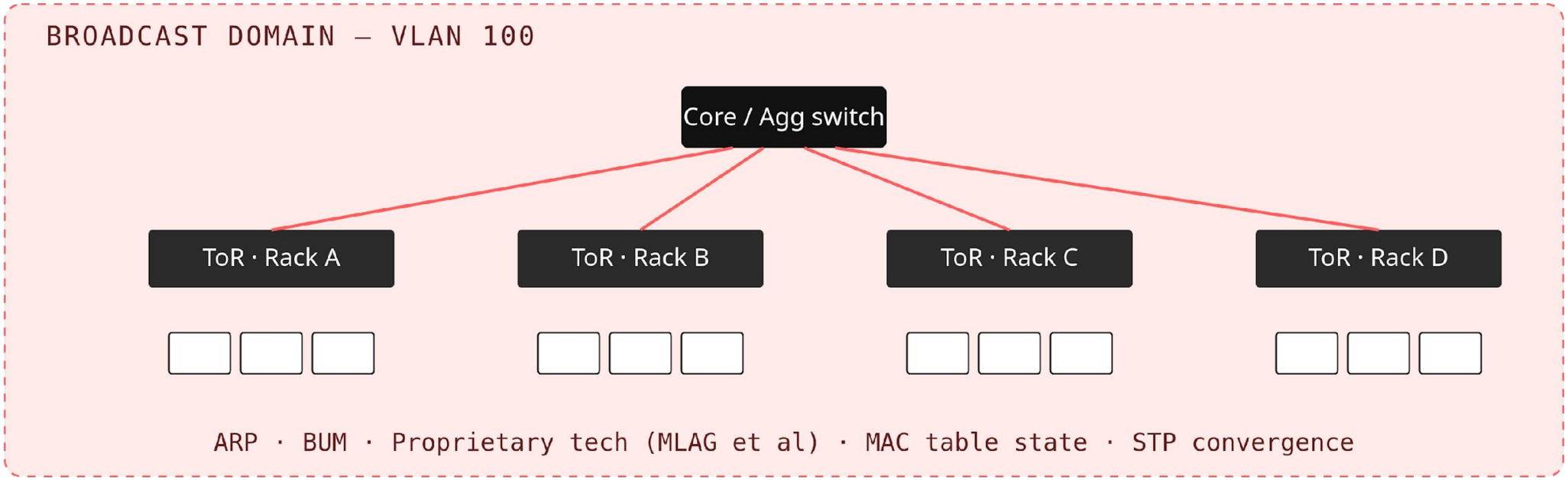
06 **Trade-offs & caveats**
Where this is the wrong answer.

01

The Layer 2 status quo

VLANs, MACs, broadcast domains.

VLANs scale to one, maybe few racks. ~~Not to a datacenter. Not beyond.~~



~~Broadcast domain stretched across racks~~ ~~Trunk links (every VLAN on every link)~~

The Layer 2 tax: four things you debug at 3am.

01 SPANNING TREE

Loop prevention by blocking links.

One topology change → convergence storm. The redundancy you built sits idle until something breaks - and then you hit a bad SFP which drops packets under load.

02 PROPRIETARY SOLUTIONS — MLAG & CO.

Vendor-specific magic to fake redundancy.

MLAG, Virtual Chassis, stacking, fabric-paths - two (or more) switches pretending to be one. Works until a firmware bump, peer-link flap, or split-brain reminds you they're not. And every vendor's flavour is different.

03 BROADCAST & ARP

Every host hears every neighbour's noise.

Misbehaving NIC, gratuitous ARP loop, IPv6 RA storm - the whole VLAN feels it. Blast radius = broadcast domain.

04 MAC TABLE PRESSURE

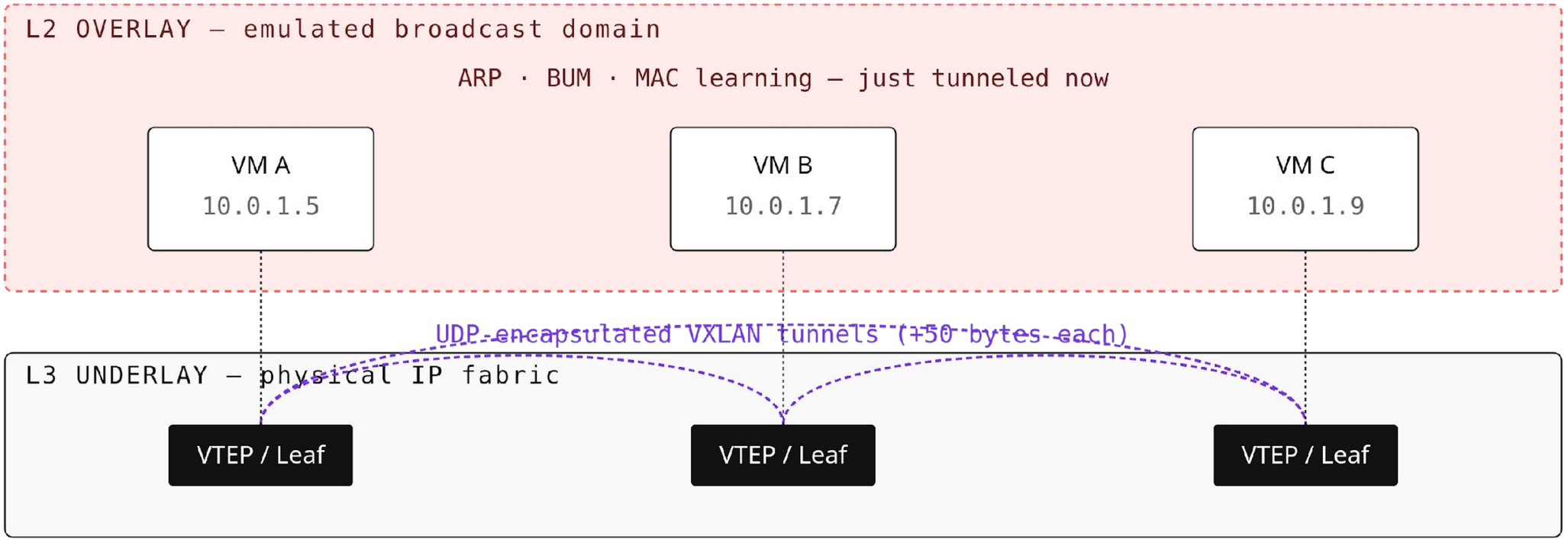
Hardware tables are finite.

Add VMs, containers, virtual MACs or a bad actor - ToR hits its ceiling, starts flooding unknown unicast. Performance dies silently. Granted: not exactly an everyday/everyperson problem :-)

02

Welcome our lord and savior: Overlays

Add more network to your network


L2 OVERLAY — emulated broadcast domain
ARP · BUM · MAC learning — just tunneled now
VM A
10.0.1.5
VM B
10.0.1.7
VM C
10.0.1.9
UDP-encapsulated VXLAN tunnels (+50 bytes each)
L3 UNDERLAY — physical IP fabric
VTEP / Leaf
VTEP / Leaf
VTEP / Leaf

The L2 problems didn't go away. They got encapsulated.

WHAT THE PITCH SAID

- L2 anywhere, L3 wherever you want
- Decouple tenants from physical topology
- Move workloads without renumbering
- Hardware-accelerated on modern silicon

WHAT YOU ACTUALLY MAINTAIN

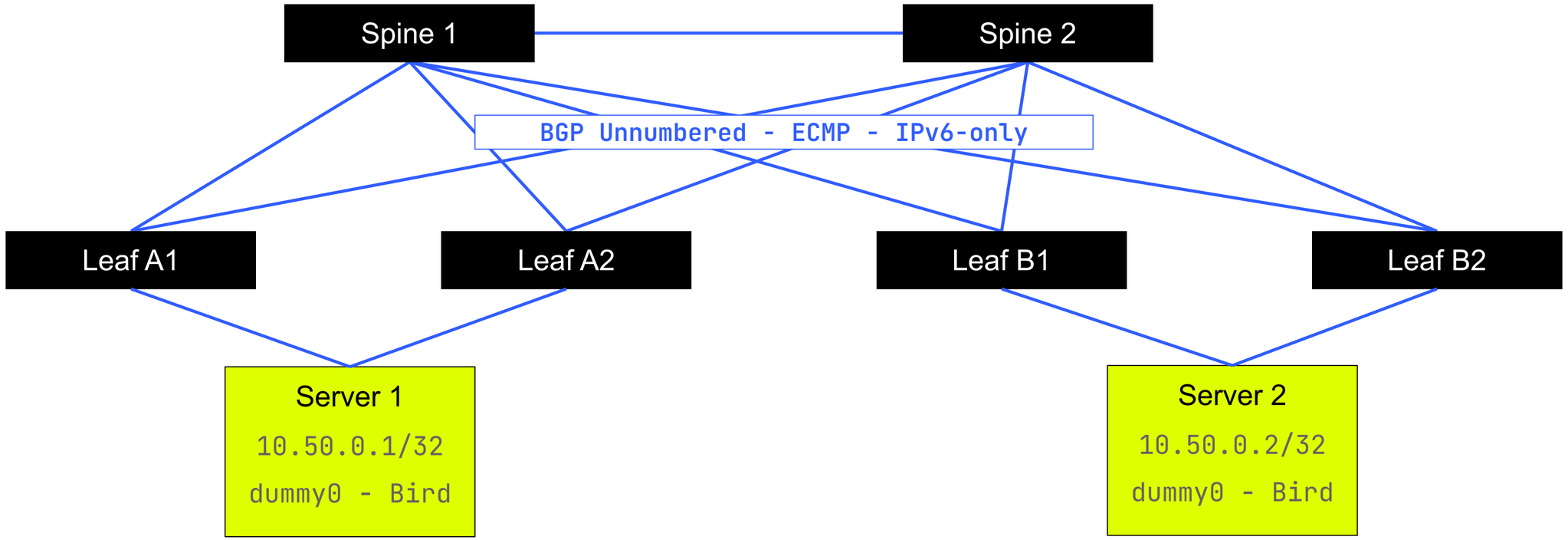
- Two networks: an underlay *and* an overlay
- MTU bookkeeping - every tunnel is +50 bytes
- EVPN control plane to learn and operate
- More technologies, more failure modes

03

What if servers just spoke L3?

Every host a router. Every service a /32.

Spine-leaf, BGP everywhere, ECMP between every pair of nodes.



BGP-routed L3 link

Host running a routing daemon

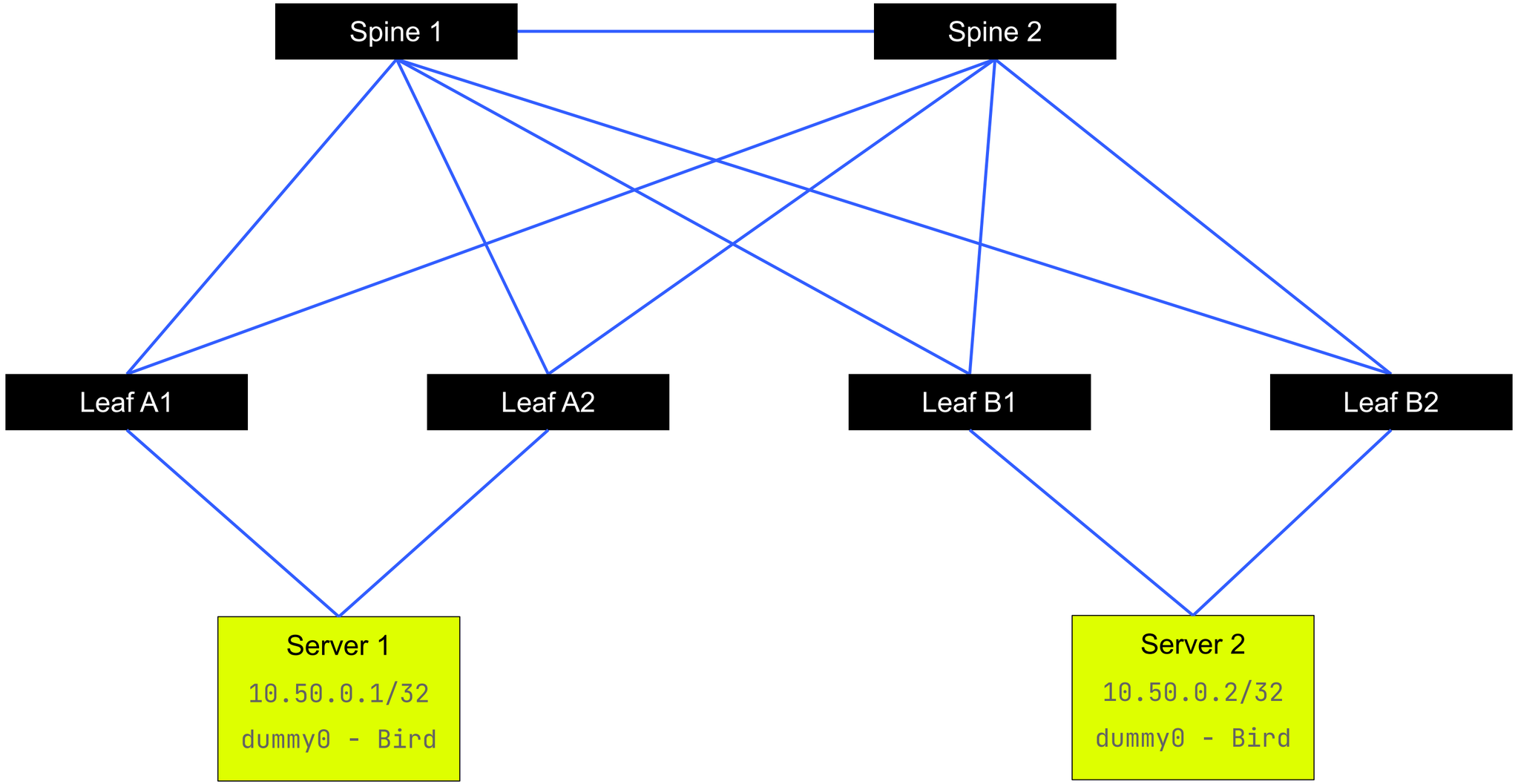
No bridges. No VLANs. No large broadcast domains.

04

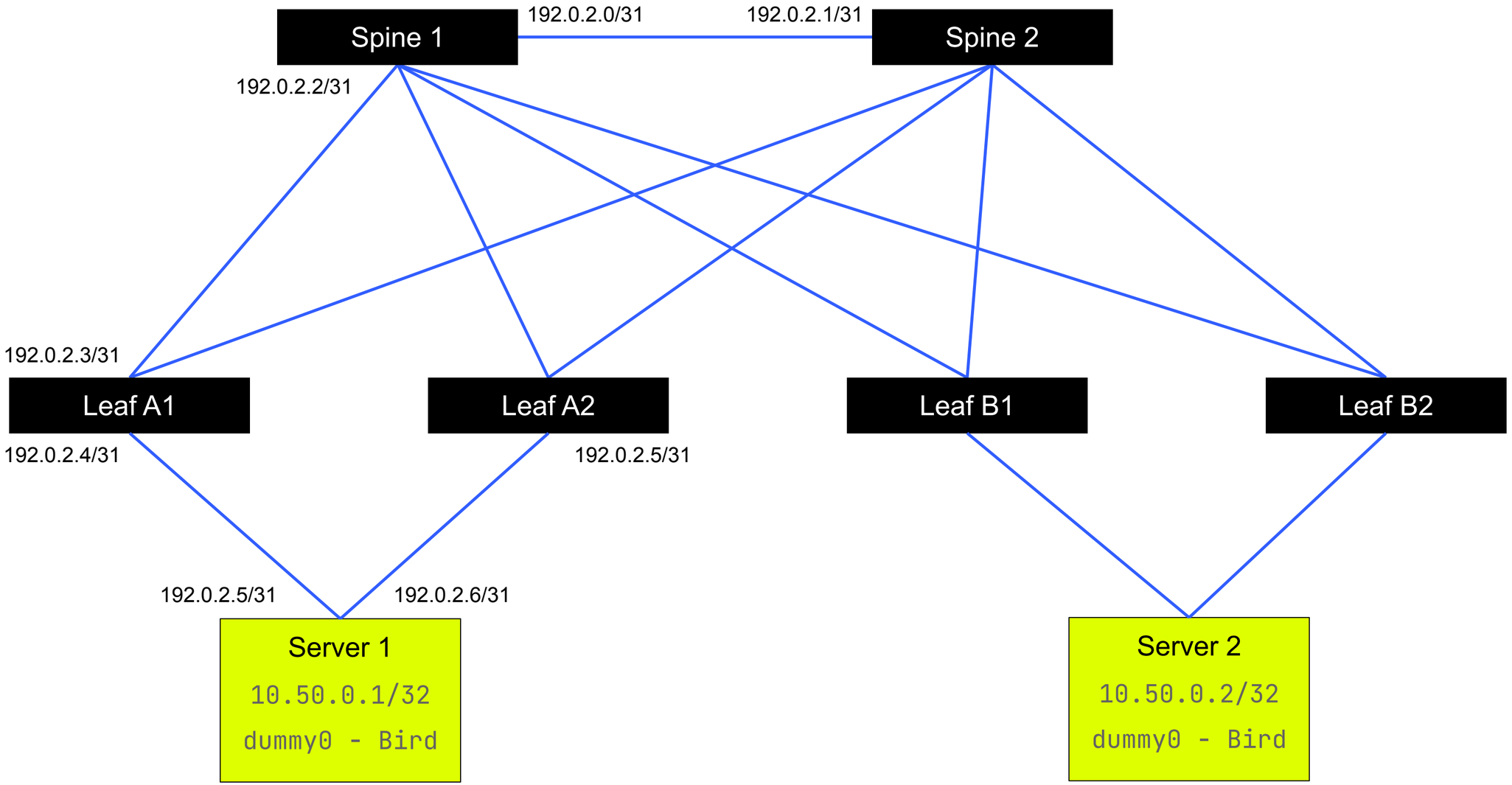
Building Blocks

BGP unnumbered · dummy0 · ECMP

Easy, right?


Spine 1
Spine 2
Leaf A1
Leaf A2
Leaf B1
Leaf B2
Server 1
10.50.0.1/32
dummy0 - Bird
Server 2
10.50.0.2/32
dummy0 - Bird

Easy, right?


Spine 1
192.0.2.0/31
192.0.2.1/31
Spine 2
192.0.2.2/31
192.0.2.3/31
Leaf A1
Leaf A2
Leaf B1
Leaf B2
192.0.2.4/31
192.0.2.5/31
192.0.2.5/31
192.0.2.6/31
Server 1
10.50.0.1/32
dummy0 - Bird
Server 2
10.50.0.2/32
dummy0 - Bird

Easy, right?


192.0.2.2
192.0.2.3/31
Leaf A1
192.0.2.4/31
192.0.2.5/31
Server 1
10.50.0.1/32
dummy0 - Bird
MOTHER OF GOD
Leaf B2
Server 2
10.50.0.2/32
dummy0 - Bird

BGP unnumbered removes the addressing headaches. But what is it?

- **IPv6 link-local addressing**
  No IPAM, no IP configuration required. fe80::/10 to the rescue.

- **IPv6 Router Advertisements & Dynamic BGP Neighbors**
  Let and L3 devices auto-discover their neighbors

- **IPv4 over IPv6 next-hop**
  You don’t need an IPv4 gateway to route IPv4 traffic.

```
203.0.113.0/24 proto bird metric 32
    nexthop via inet6 fe80::1 dev eth0 weight 1
    nexthop via inet6 fe80::1 dev eth1 weight 1
```

BGP unnumbered removes the addressing headaches. But what is it?

- **No transfer-net hell, no dual stack**
  No /31 + /127 per cable, no assignment logic required.
- **No ARP, no neighbour learning, no DHCP**
  Link-local discovery is the kernel's job.
- **Uniform switchport and host configs**
  Plug a cable in. BGP comes up. Connectivity established.

blog.bott.im/bgp-unnumbered-in-2025-same-idea-different-implementations

Dummy interfaces are how a server owns an IP.

A virtual interface that owns an IP without needing a physical link.

- **Service IPs live here**
  HAProxy binds to the dummy IP. So does Apache, Postgres, anything.
- **Always-up by definition**
  Independent of which physical NIC is currently carrying traffic.
- **Announced via BGP**
  The dummy IP is the prefix the host advertises to its leafs.

```
ip link add dummy0 type dummy
ip link set dummy0 up
ip addr add 10.50.0.1/32 dev dummy0

# BIRD picks it up via 'protocol direct'
# Leafs learn it via BGP
# Reachable from every other server
```

ECMP allows you to use all available links

- **No Proprietary Protocols**
  ECMP, strictly speaking, isn't a protocol.
  It is a forwarding behavior.
- **Flow-Based Route Selection**
  Traffic is distributed by source/destination IPs & ports.
  A single flow is limited to a single link’s bandwidth.
- **Broadly Available**
  Equal-Cost-Multipath is built into all “enterprise” L3 network devices.
  Introduced in Linux 3.x, improved in 4.x.

05

In the wild.

Generic apps, Kubernetes, Ganeti

4 slides

Most apps don't notice the network is different.

- **Services expect an IP address**
  Whether it sits in eth0 or dummy0 does not matter.

- **Standard Application will just work™**
  HAProxy, Apache, nginx, Postgres…you name it.

Local BGP endpoint enables downstream clients.

Configure bird on hosts to accept local BGP connections

- **Anycast / HA services**
  Use other BGP software like exabgp to announce additional loopback IPs based on service health.
- **Virtual Machines**
  Run virtual machines with their own bird which connects to the host's bird via BGP.
- **Kubernetes CNI**
  Have cilium & co announce ingress IPs and pod networks through BGP.

blog.bott.im/look-ma-no-bridge-layer-3-vm-clusters-with-ganeti

Kubernetes: ditch the CNI overlay, route pods natively.

DEFAULT CNI BEHAVIOUR

- Pods get private IPs the underlay doesn't know about
- Tunnel encap (VXLAN, IP-in-IP, WireGuard) between every node
- SNAT pod traffic to the node IP for egress
- MTU pain. Pcap pain. Performance loss.

NATIVE ROUTING (CALICO BGP, CILIUM NATIVE)

- Pods get **real, routable** IPs
- Each node announces its pod prefix via BGP to the leafs
- No NAT for egress — pod IP is the source
- Same kernel data path as the rest of the host

The Kubernetes overlay was a workaround for not being able to ask the underlay for routes. With BGP, you can.

06

Trade-offs & caveats.

What this approach asks of you.

01 HARDWARE

L3-capable access switches.

Most modern silicon does this fine. If you're still on pure L2 hardware, it is probably old :-)

02 SOFTWARE

A routing daemon on every host.

FRR or BIRD. Both rock-solid, both packaged everywhere. Adds one daemon and a config file to your host fleet.

03 OPERATIONS

A team comfortable reading routing tables.

`ip route`, `birdc show route`, `vtysh` . Debugging connectivity problems changes, but OTOH you always head to deal with L2 **and** L3.

04 MINDSET

Trust the routing protocol.

BGP has run the internet for decades. It will run your datacenter. If you already operate your own AS, you have the knowledge anway.

When this is the wrong answer.

A LEGACY APPS

L2-bound clustering protocols.

Old Pacemaker/Corosync, some HA appliances, certain storage replication. They expect a shared broadcast domain.

B CUSTOMER / TENANT SEPARATION

Where Overlays DO make sense.

If you provide infrastructure/hosting services to customers, you probably want them strongly separated.

C DHCP

Provisioning with DHCP won’t work.

If you depend on a process including DHCP, you won’t get very far. (PXE & friends excluded, there are solutions for that).

D VM LIVE MIGRATION

Possible - but with larger network downtime.

Live VM migration will now be possible even across large distances / different datacenters. But the network failover time will be in the range of seconds, not sub-second.

Bye bye VLANs. Hello boring, scalable networks.