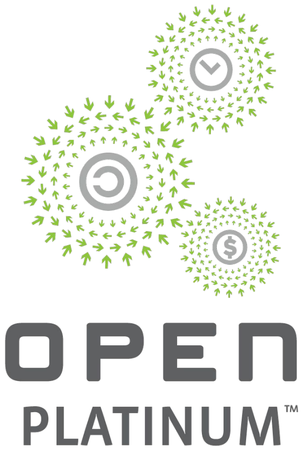
Open. Together.

OCP
SUMMIT

Networking
Software

The Linux Kernel, Ecosystem, and Community for Open Switch Hardware

Roopa Prabhu, Director Engineering, Cumulus Networks



This talk is about...

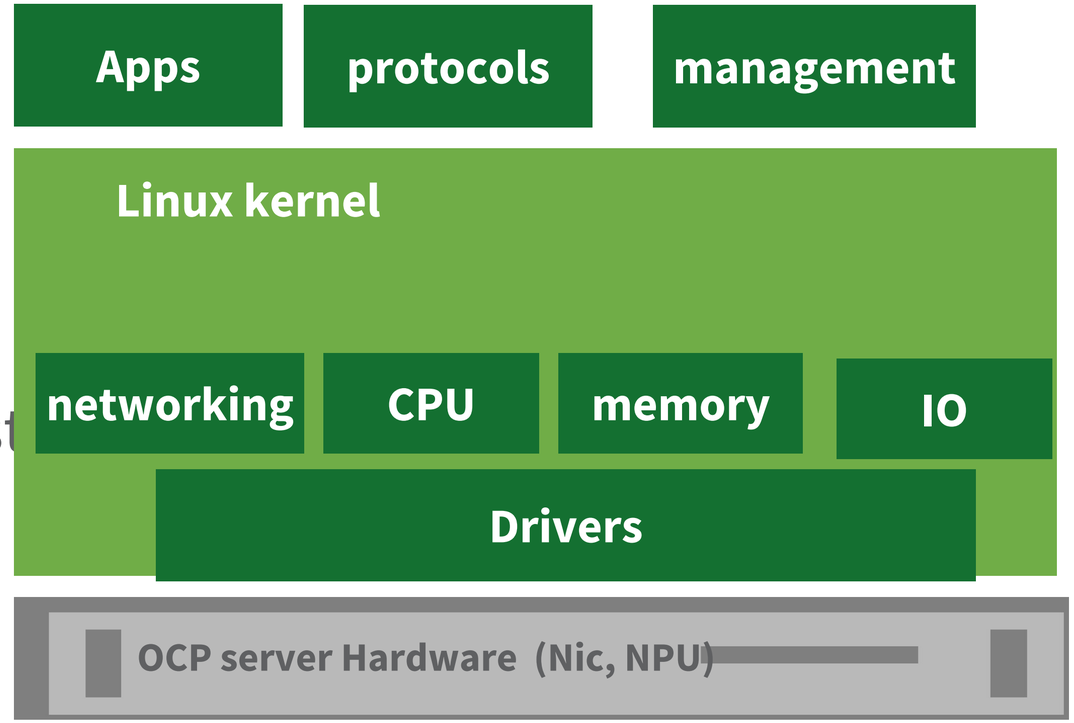
NETWORKING

- Open Switch Hardware and Linux Networking
- Disaggregated hardware and software stacks
- Journey of Open Switch Hardware in the Linux kernel and Community
  - Linux Switch Hardware Offload
  - Linux networking features for the Datacenter Fabric
- Leveraging Linux ecosystem for Open Switch Hardware
- Building Open Data center networking fabrics with Linux and Open Switch Hardware

Open Switch hardware and Linux: Revolution or Evolution ?

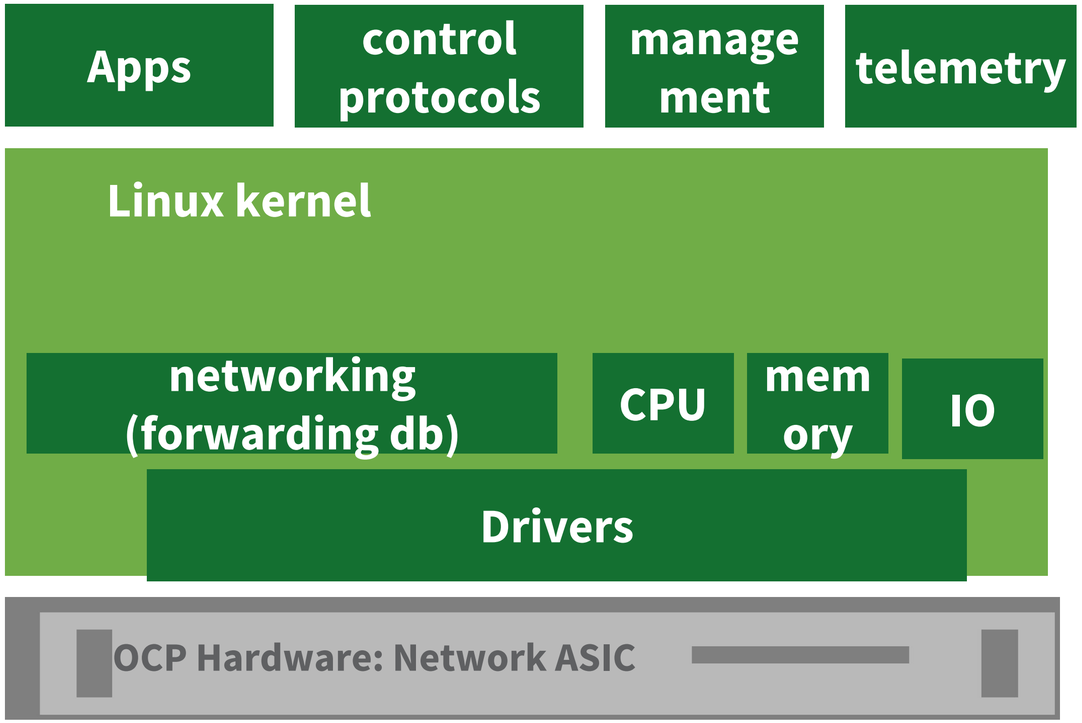
Open Compute Hardware and Linux networking

- Disaggregated Hardware and Software Stacks
- Linux hardware offload Model: Software accelerated by hardware (Network, memory, disk)
- Virtual hardware models: provide ability to test without HW



Open Networking Switch Hardware and Linux networking

- Disaggregated Hardware and Software Stacks
- Linux hardware acceleration Model: Software networking accelerated by Hardware
- Linux network forwarding plane is the Model
- Virtual Linux forwarding model: provides ability to test networking without HW

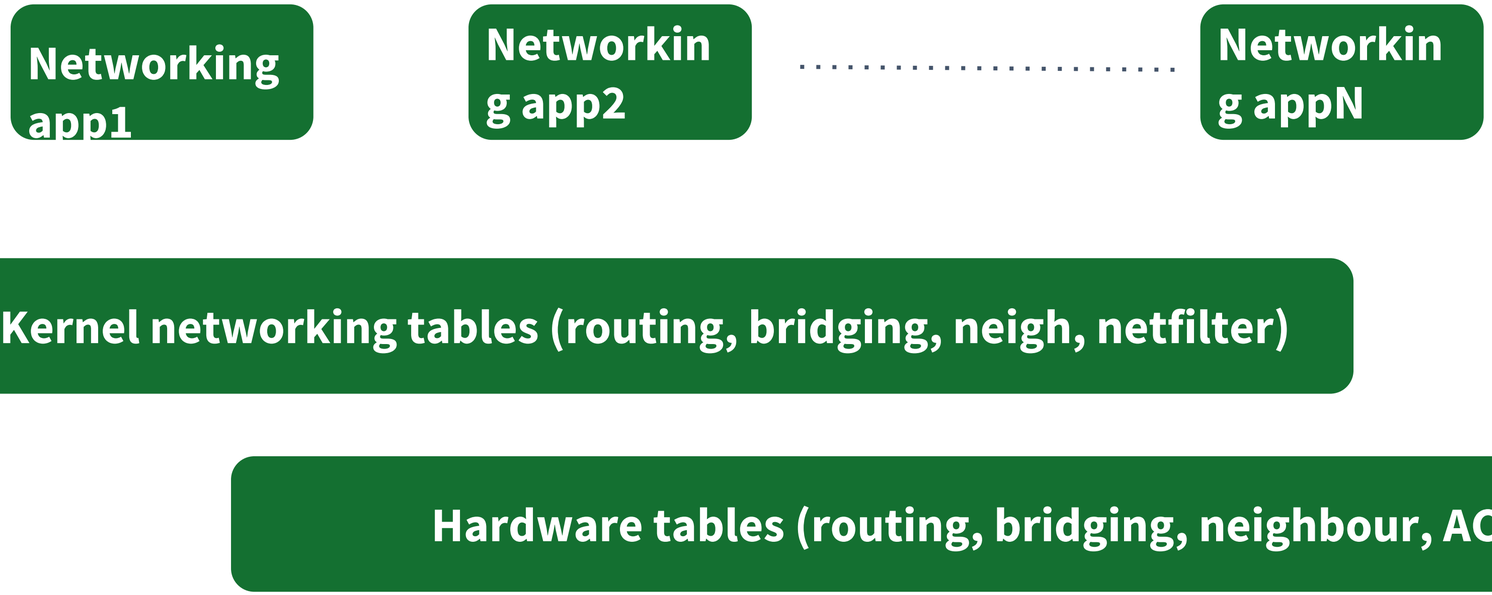


It’s a natural Evolution..

- Unified Architectures for all Open Hardware
- Unified deployment and operational models
  - Ability to simulate and test workflows with Linux software forwarding plane
- Vast Linux ecosystem to support all Open Hardware
- Cross technology pollination
  - Faster pace of innovation

No special appDB, configDB or appLib

Linux networking API (Netlink), Linux kernel tables and Linux kernel hardware offload API



OCP Hardware running Linux

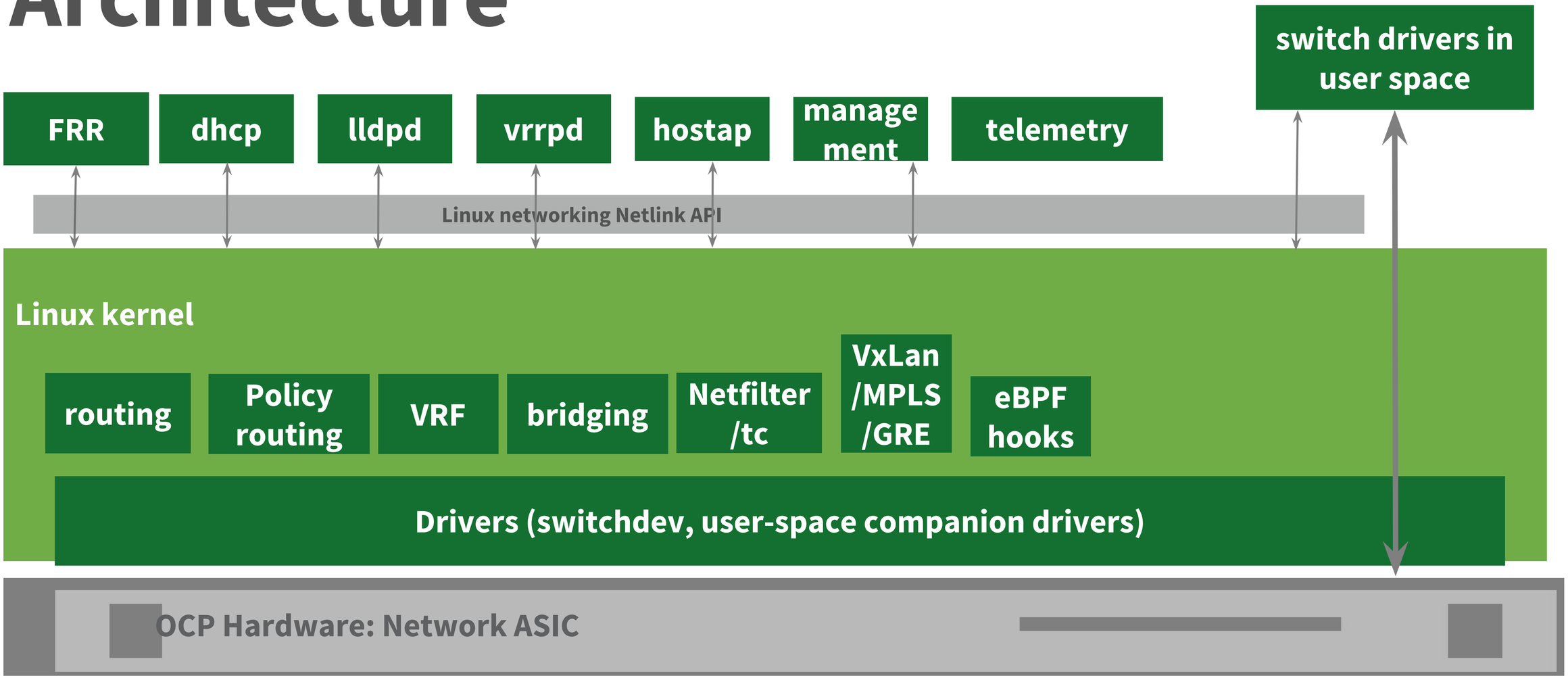
All Open hardware in this presentation is OCP hardware

Example hardware [1]

- FaceBook Wedge 100
- FaceBook Voyager
- Edgecore Networks AS7712-32X
- and more

Open Switch Hardware
and
Linux Kernel Networking

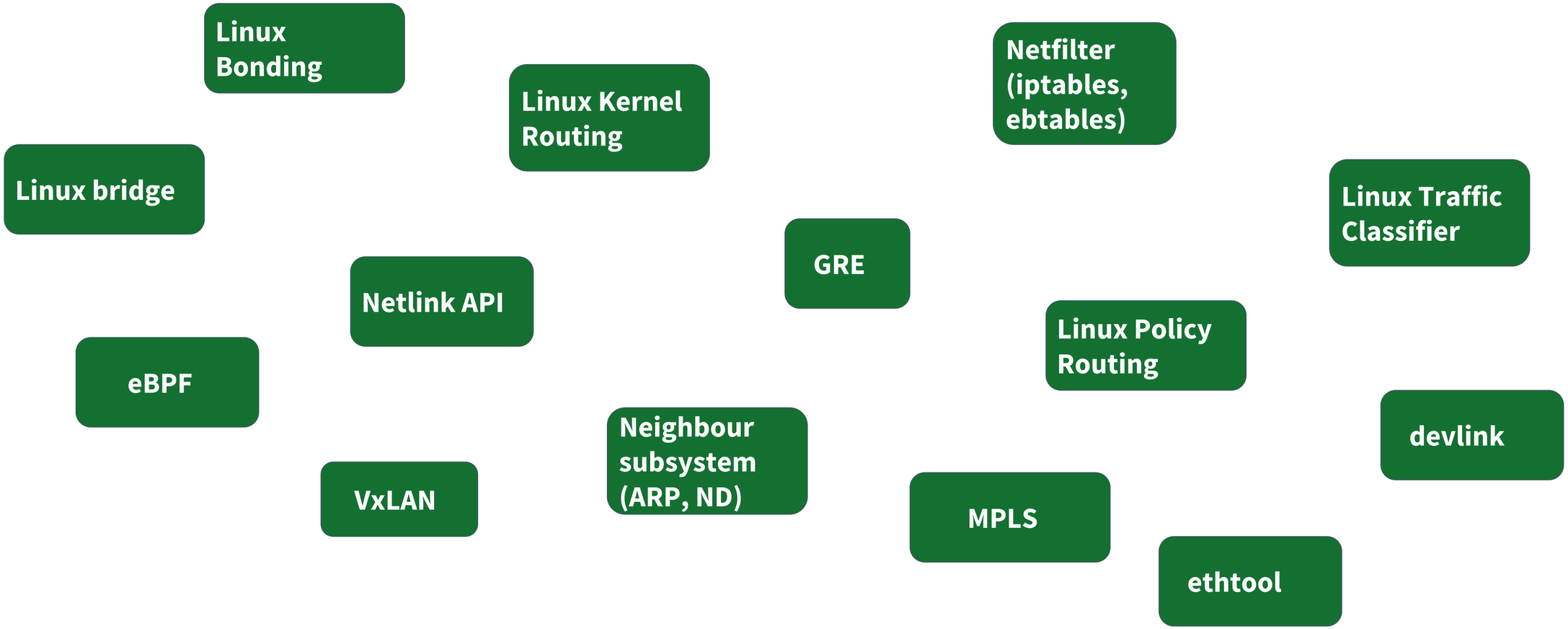
Architecture



Linux Kernel Switch ASIC offload support

- Networking community and maintainer
- New abstractions: switchdev-ops, notifiers-ops, netdev-ops
- Extensions to “Netlink API” [2] to support Switch ASIC deployments
- Linux kernel gets new features to support switch ASIC deployments: VRF [4], VxLAN [5], E-VPN dataplane [3,5], MPLS [10]
  - These features in turn have found uses in other software and host deployments

Linux Kernel building blocks



Linux Networking Ecosystem

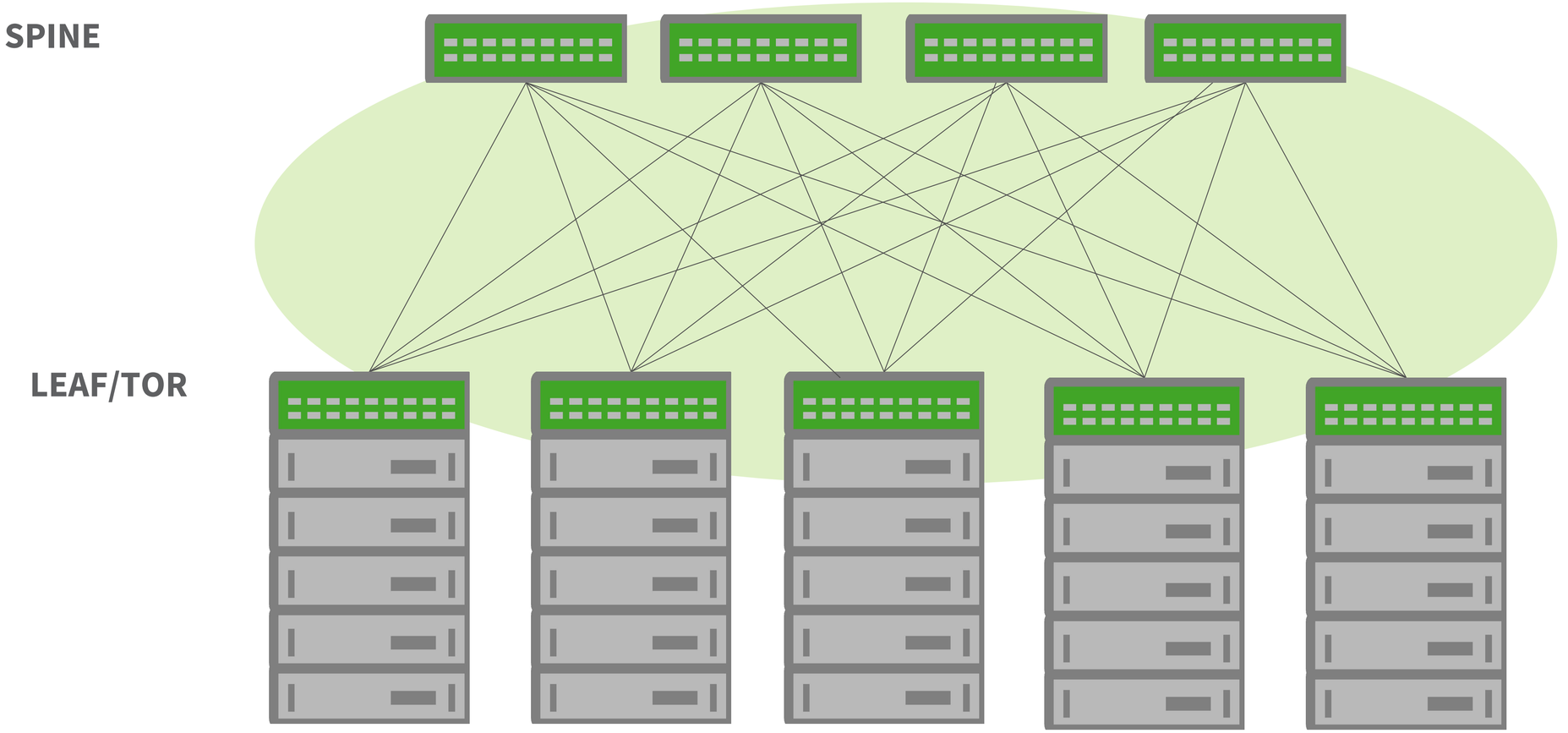
- Free Range Routing suite [3]
- Linux Dhcp [6] , vrrpd [7], lldpd [8], wpa [9], networking tools
- iproute2 [17] and ifupdown2 [18] for network configuration
- Systemd for service monitoring [16]
- Linux traffic classifier
- Linux netfilter: iptables, ip6tables, ebtables

Linux Networking: Latest cool things

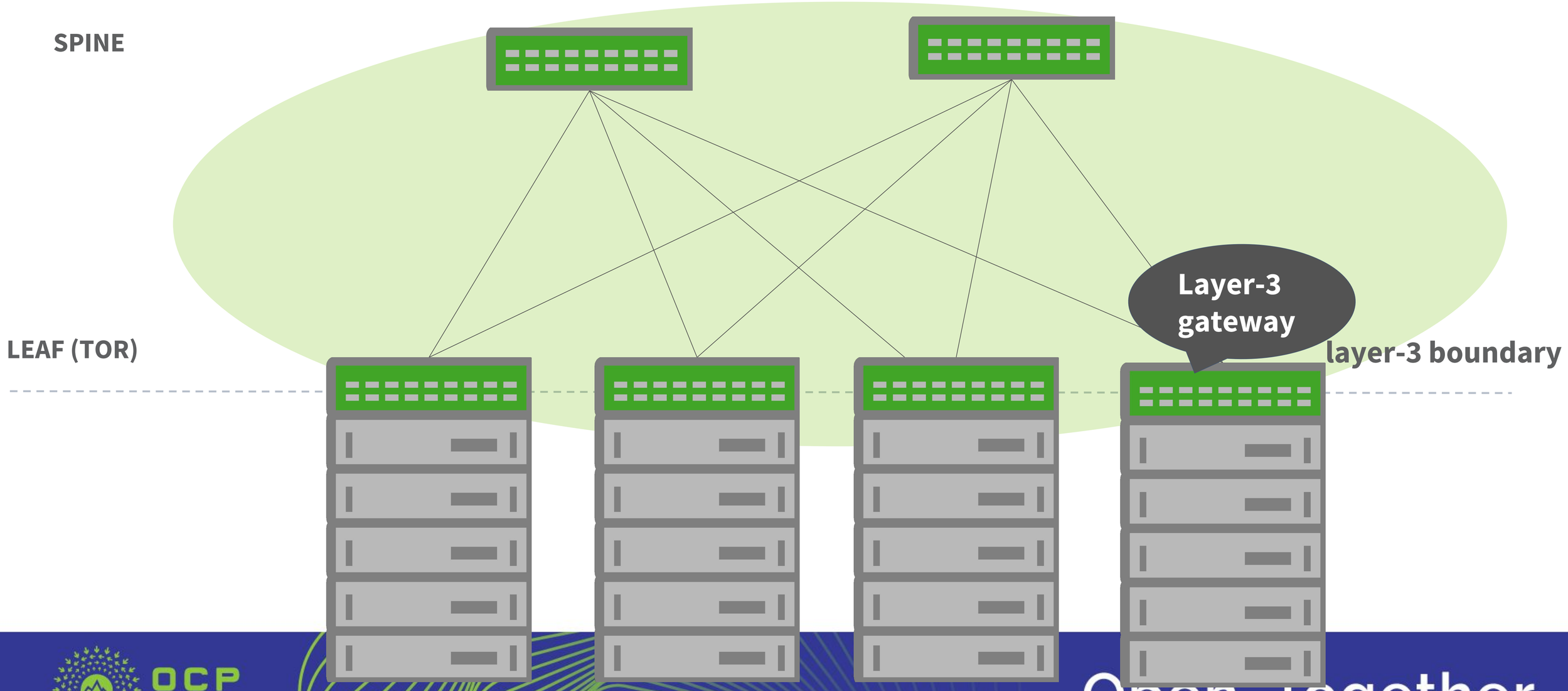
- Network programmability with eBPF hooks in
  - Network cgroups
  - Network tracing
  - TCP analytics and congestion algorithms
  - Filtering: Linux traffic classifier and Netfilter (bpfFilter)
  - Accelerated Datapath with XDP
  - Socket API’s

Open Switch Hardware
and
Linux Networking in the Data Center

Modern Data Center Network



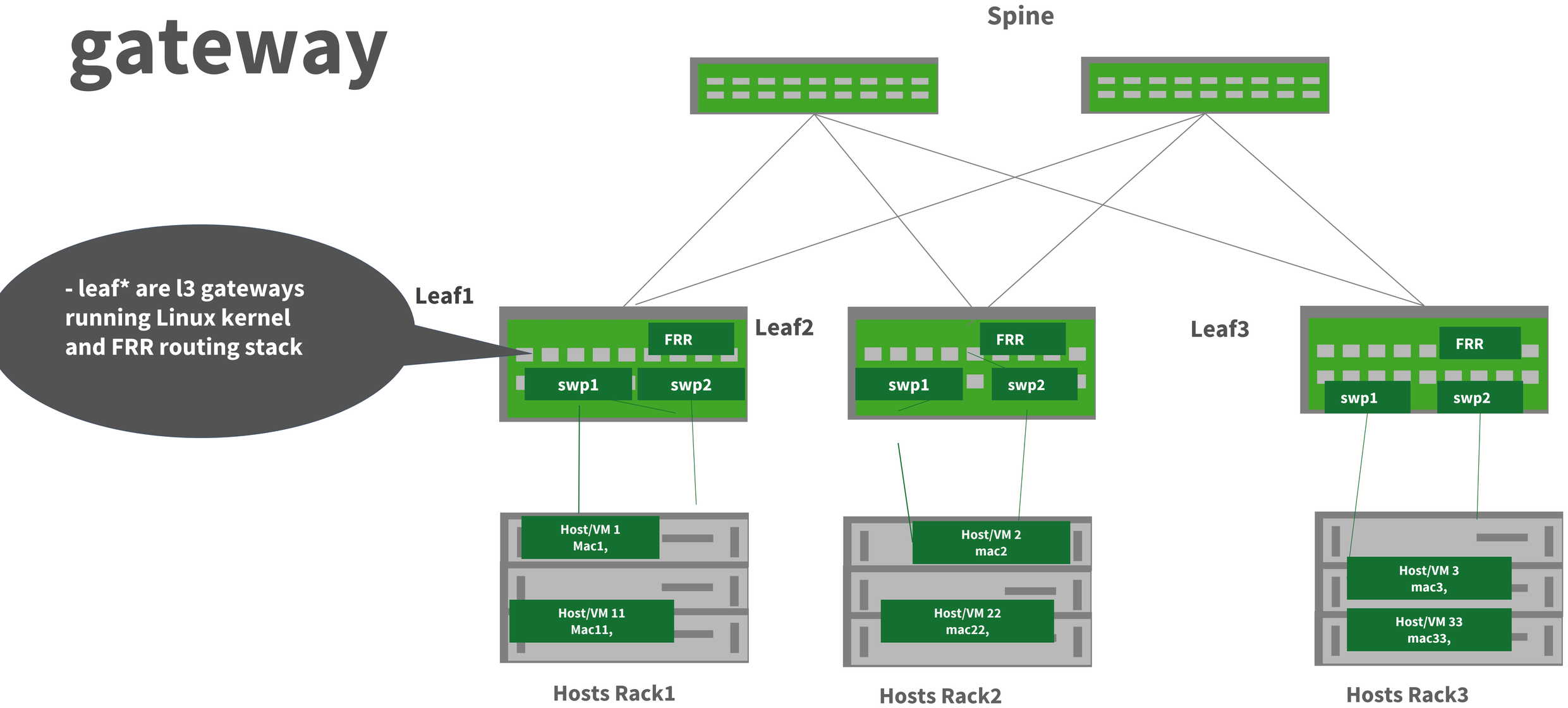
Data center Layer-3 gateway



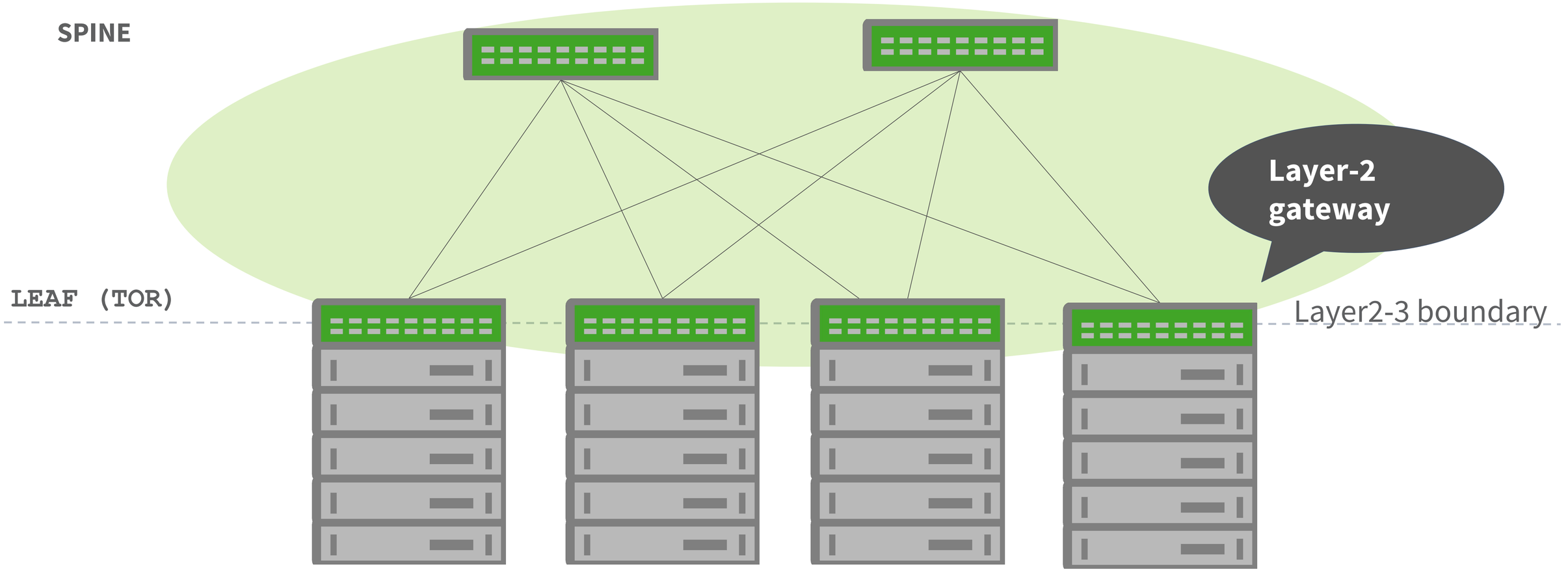
Open Layer3 gateway

- Linux kernel: routing FIB, VRF and neighbour subsystem
- Open Linux routing protocol stack: FRR (Free Range Routing)
- Open switch ASIC hardware with layer3 support

Open switch hardware and Linux L3 gateway



Hybrid layer2 - layer3 data center network

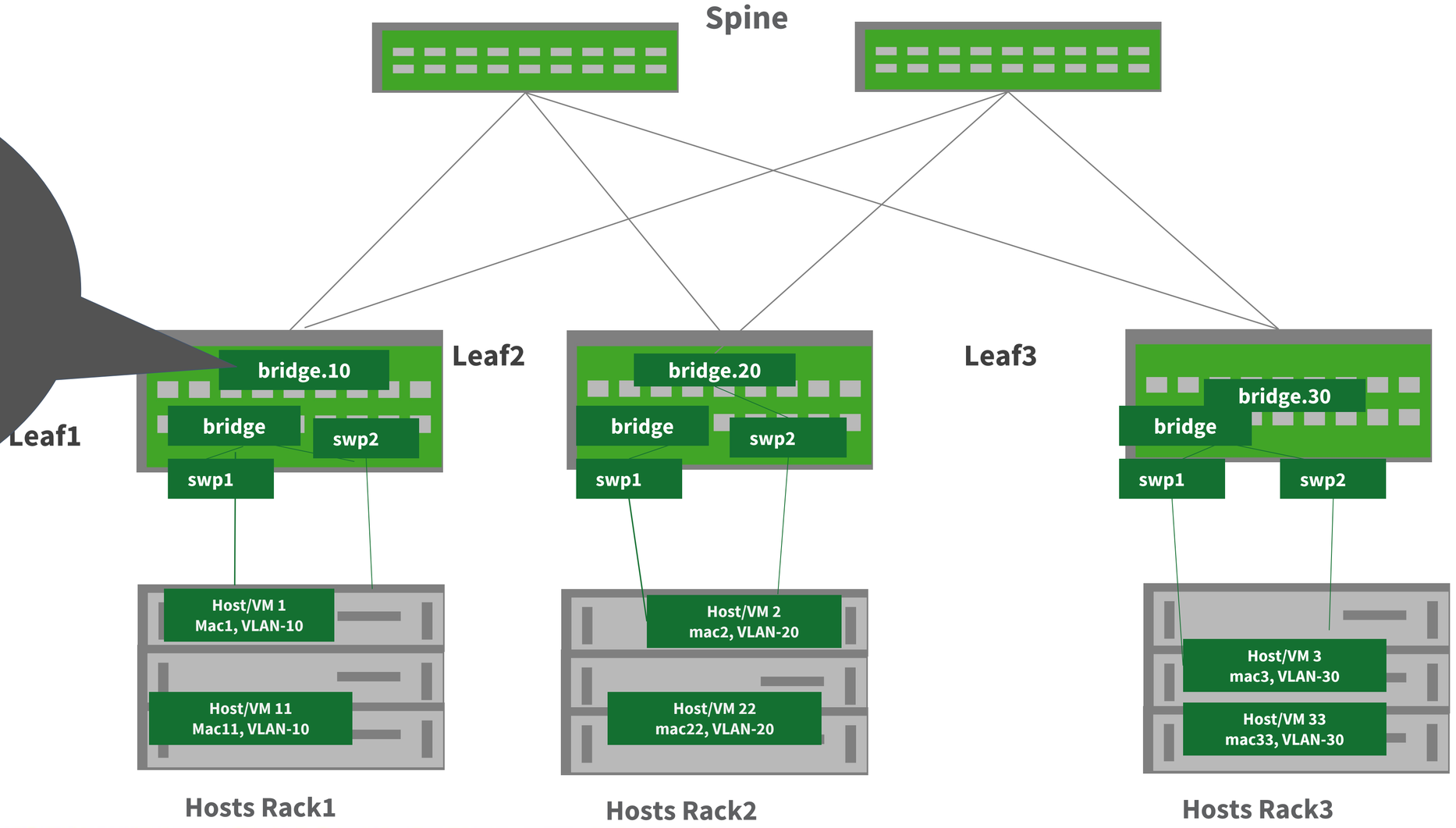


Open Layer2 gateway

- Linux kernel bridge driver and forwarding database:
  - STP, IGMP snooping
- Open Linux protocol implementations
- Open switch ASIC hardware with Layer2 support

Open switch hardware and Linux L2 gateway

- leaf* are l2 gateways running Linux bridge
- Bridge within the same vlan and rack and route between vlans
- bridge.* Linux vlan interfaces act as SVIs for routing

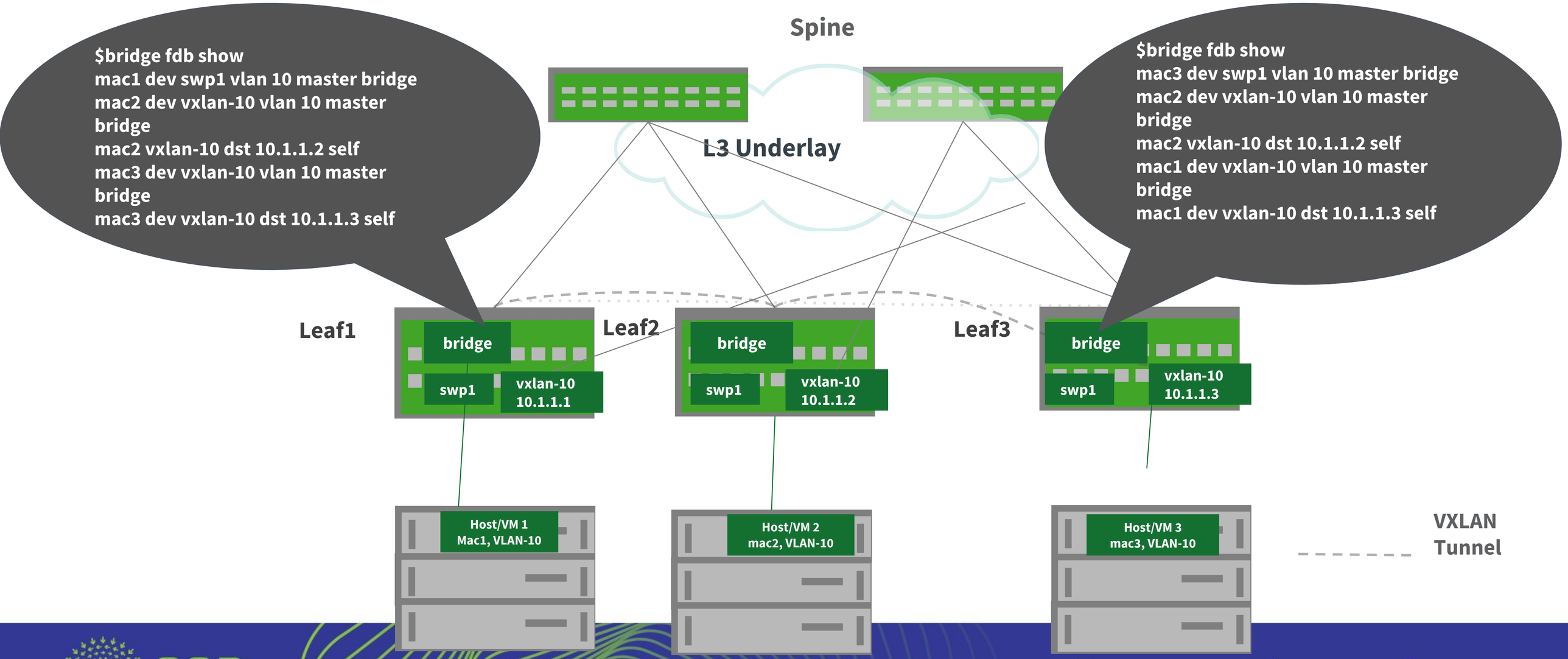


Network Virtualization and Overlay gateways

Open VxLan overlay gateway

- Linux kernel vxlan data and forwarding plane
- Linux bridge driver
- Open switch ASIC hardware with vxlan support

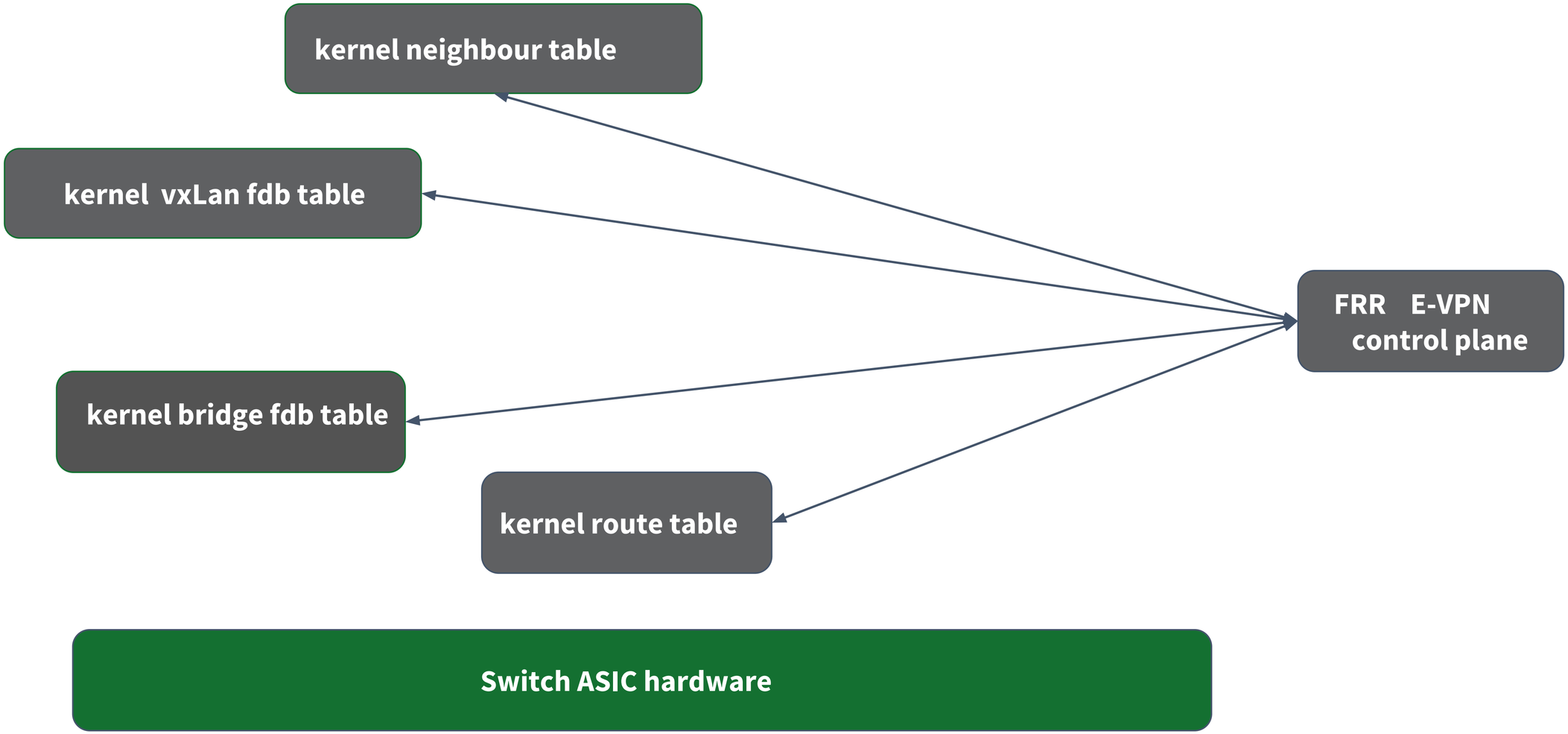
Open switch hardware and Linux overlay gateway



Open Linux E-VPN Data center Fabric

- Linux kernel VxLan data and forwarding plane
- Linux kernel routing, bridge and neighbour subsystem
- Open switch ASIC hardware with VxLan and routing support
- Open E-VPN control plane: FRR (Free range routing)

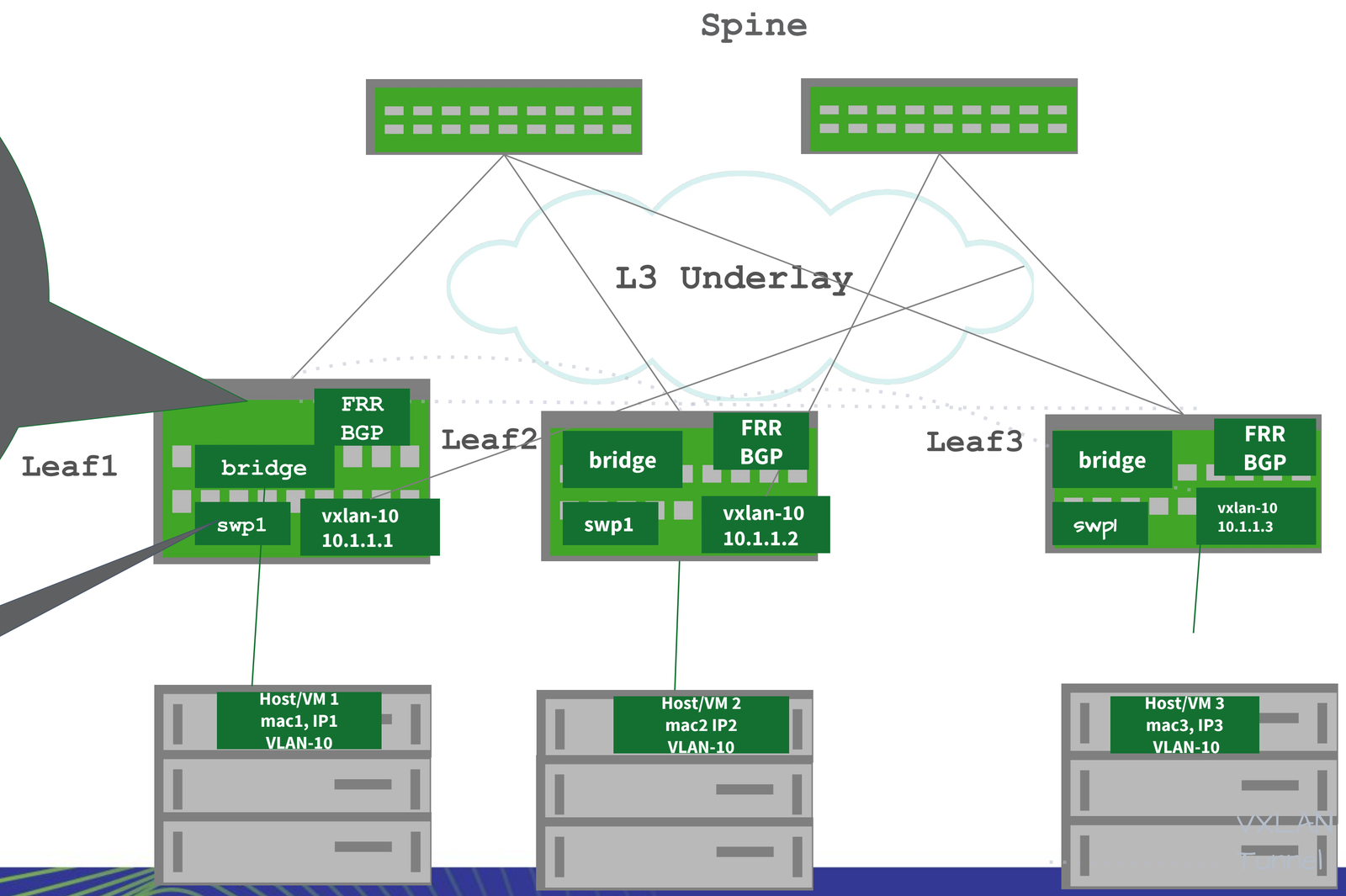
FRR E-VPN and Linux kernel



E-VPN with Open Hardware and Linux

(a) FRR BGP discovers local vlan-vni mapping via netlink
(b) BGP reads local bridge <mac, vlan> entries and distributes them to bgp E-vpn peers
(c) BGP learns remote <mac, vni> entries from E-VPN peers and installs them in the kernel bridge fdb table
(d) Kernel bridge fdb table has all local and remote mac's for forwarding

(a) Bridge learns local <mac, vlan> in its fdb



New and Ongoing work

- Scaling Linux routing API [13]
- Devlink hardware management API for Switch ASICS [14]
    - Extends beyond Switch Hardware: NICs, SRIOV, NPUs
    - Firmware management
- E-VPN updates for multihoming and multicast
- Debuggability: perf tracing/probes in networking subsystems
- Kernel networking selftests [11] and syzbot [12]

References

[1] Cumulus Linux hardware compatibility list: https://cumulusnetworks.com/products/hardware-compatibility-list/

[2] Netlink API: http://man7.org/linux/man-pages/man7/netlink.7.html

[3] FRR routing stack: https://frrouting.org/

[4] VRF https://cumulusnetworks.com/blog/vrf-for-linux/

[5] Linux bridge, VxLan and E-VPN https://www.netdevconf.org/2.2/slides/prabhu-linuxbridge-tutorial.pdf

[6] Linux Dhcp server: https://packages.debian.org/isc-dhcp-server

[7] VRRP: https://packages.debian.org/vrrpd

[8] LLDPD https://packages.debian.org/lldpd

[9] WPA (802.1x) https://packages.debian.org/wpa

References (Contd)

[10] MPLS in the Linux kernel:

https://netdevconf.org/1.1/tutorial-deploying-mpls-linux-roopa-prabhu.html

[11] Linux kernel selftests: testing hardware switch forwarding with VRFs :
https://marc.info/?l=linux-netdev&m=151981456405307&w=2

[12] syzbot: Tests Linux kernel branches: https://github.com/google/syzkaller/blob/master/docs/syzbot.md

[13] Scaling routing API https://lwn.net/Articles/763950/

[14] devlink api for switch ASICS: https://lwn.net/Articles/674867/

[15] E-VPN: Arp-ND suppression support: https://patchwork.ozlabs.org/cover/822906/

[16] systemd: https://wiki.debian.org/systemd

[17] iproute2: https://mirrors.edge.kernel.org/pub/linux/utils/net/iproute2/

[18] ifupdown2: https://packages.debian.org/ifupdown2

Call to Action

Linux networking Community:
mailing list: netdev@vger.kernel.org , http://vger.kernel.org/vger-lists.html#netdev

Free Range Routing Community: https://frrouting.org/ , https://frrouting.org/#participate

Linux debian ecosystem for packages/apps: eg https://packages.debian.org/jessie/

Cumulus Networks Hardware compatibility list: for native Linux network operating system support on OCP hardware: https://cumulusnetworks.com/products/hardware-compatibility-list/

Linux networking Conference to discuss new hardware and software support for switch ASICs: https://netdevconf.org/

Linux networking hardware offload workshop at the upcoming conference in Prague: https://www.netdevconf.org/0x13/session.html?workshop-hardware-offload

Open Network Install Environment:
https://www.opencompute.org/wiki/Networking/SpecsAndDesigns#Open_Network_Install_Environment


Open. Together.
OCP Global Summit | March 14–15, 2019
OCP
SUMMIT