DigitalOcean

The Gateway to the Cloud

OvS in a Layer-3 Routed Datacenter

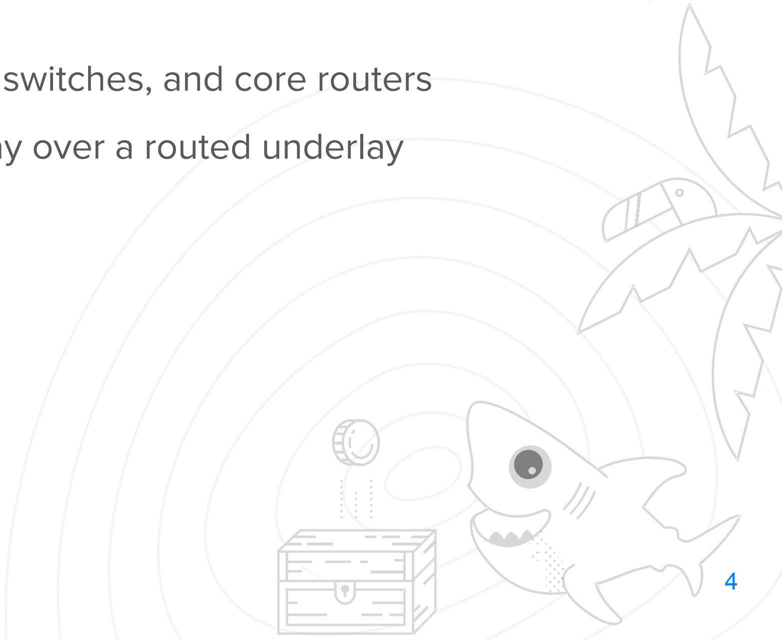
Carl Baldwin
Jacob Cooper

OVS+OVN ‘19

Layer 3 Public Network

Public Traffic to Droplets

- Context
  - droplets know they have public addresses (no NAT)
  - need better IP mobility to reclaim stranded IPs
  - avoid relying more on a global database
  - avoid a hardware refresh: use existing NICs, switches, and core routers
  - not related to VPC which uses its own overlay over a routed underlay



DO Started with Layer 2

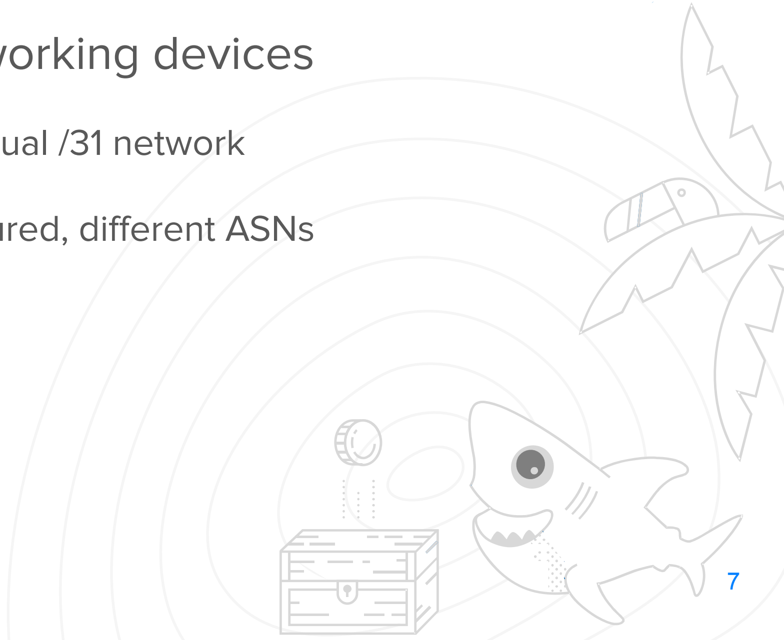
- Pros
  - easy to deploy
  - simple integration with servers (vlan tags only)
  - simple integration with droplets (VMs)
  - good for relatively small scale
- Cons
  - chatty
  - scalability issues (scale-up model)
  - huge blast radius
  - hard to troubleshoot
  - limited IP mobility

Layer 3 Advantages

- quiet
- scalable (scale-out approach)
- better IP mobility
- highly redundant
- minimal blast radius (single rack or even single server)
- easy to troubleshoot
- very easy to isolate faulty device (easy maintenance)

Layer 3 Fabric

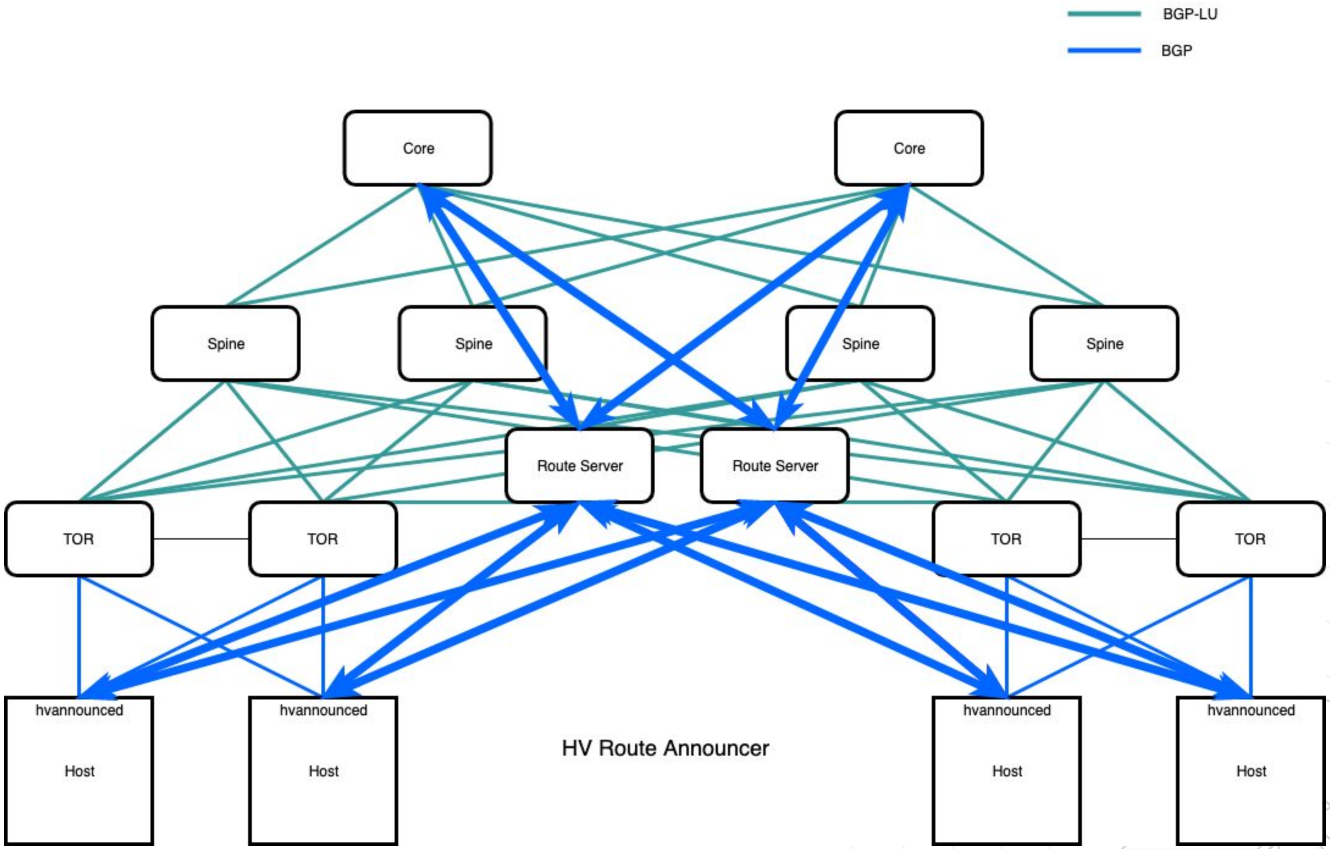
- IP fabric CLOS topology
- BGP is the only routing protocol in the datacenter
    - “Use of BGP for Routing in Large-Scale Data Centers” RFC7938
- More complicated configuration on networking devices
    - each fabric port has assigned IP from individual /31 network
    - each device has many BGP sessions configured, different ASNs
    - automation is a must have



Public Traffic Over IP Fabric

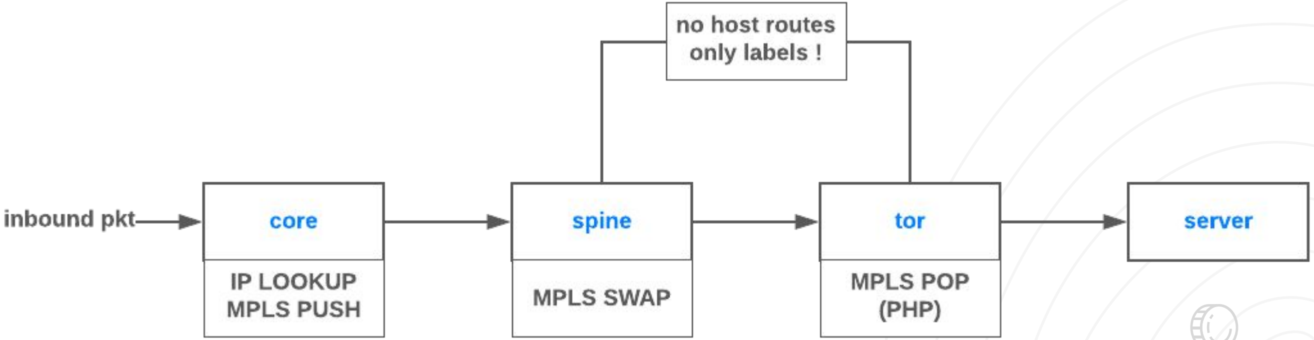
- host route per droplet
  - provides IP mobility (avoid constraining placement and migration)
  - no prefix aggregation up to the data center edge
- many host routes - challenges
  - leaf/spine switches need information about all host routes to perform IP forwarding
  - FIB (Forwarding Information Base) size limitation, especially in TORs and spines
- full encapsulation for public traffic is not yet feasible for us

BGP Control Plane



IP/MPLS fabric

- MPLS loosens constraints related to limited resources on fabric switches
- added MPLS with one additional NLRI to existing BGP sessions (BGP-LU)
    - “Using BGP to Bind MPLS Labels to Address Prefixes” RFC8277



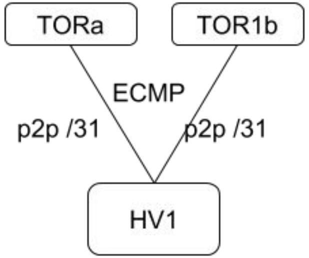
Hypervisors

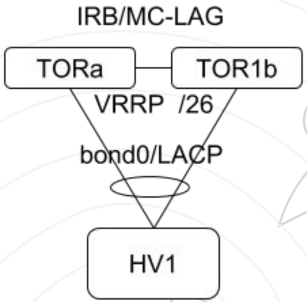
Open vSwitch

- heavy investment as our HV datapath
  - firewall, floating IP, VPC, public addresses, etc.
- ultimate control over packet forwarding
- parallel data paths for layer 2 and layer 3
  - key to a seamless pivot

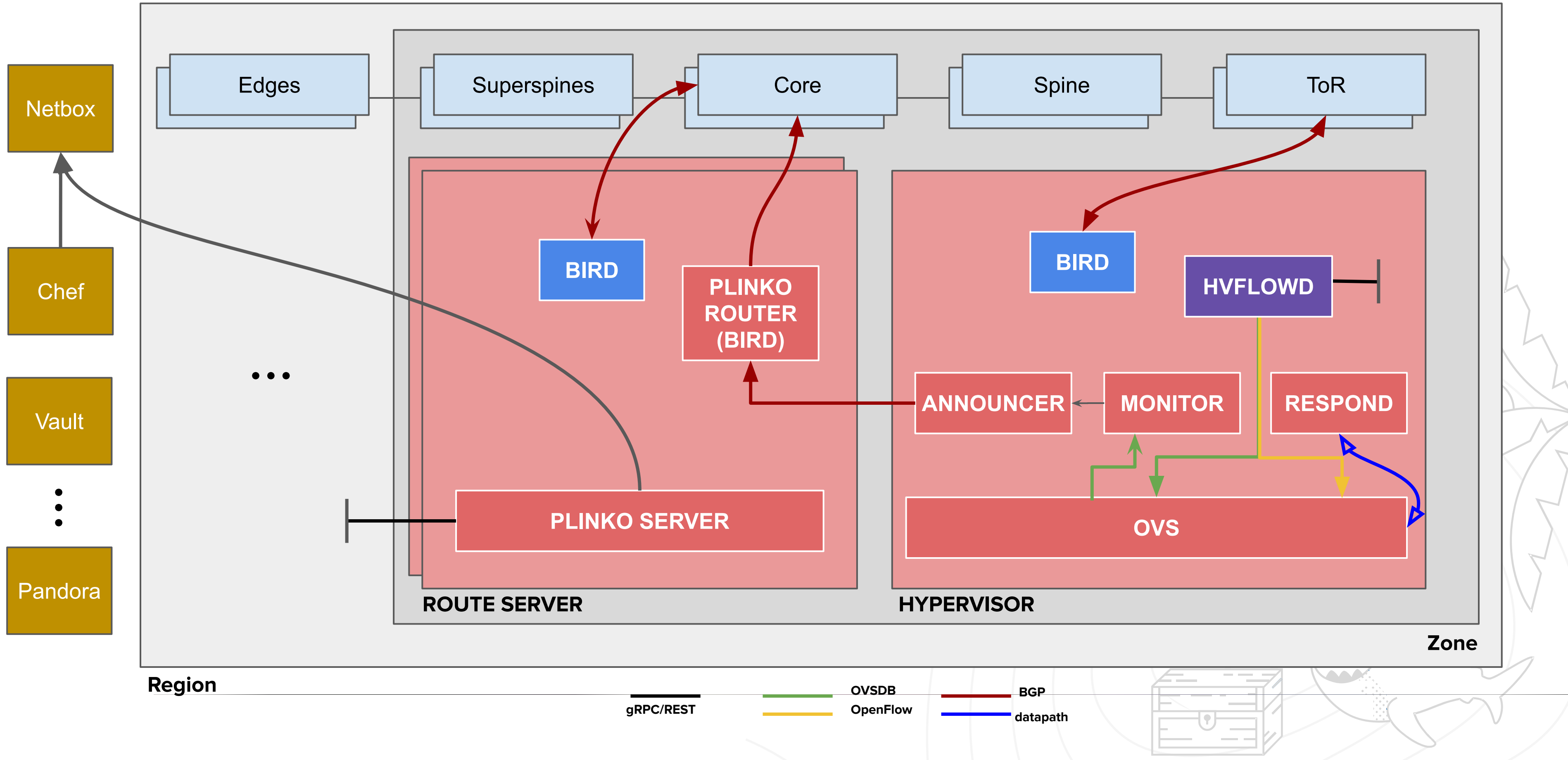
Open vSwitch Challenges

- usually associated with Layer 2
  - built-in bond implementation
- no integration with routing protocols
  - made L3 to the host tricky (for now?)
  - still running L2 to the TOR (impedance mismatch) ------>
  - group with *type=select*, *selection_method=dp_hash*
- not labeling MPLS on the HV
  - older version performed poorly on MPLS labeling
  - BGP-LU implementations limited for HVs anyway





L3/MPLS Deployment Architecture



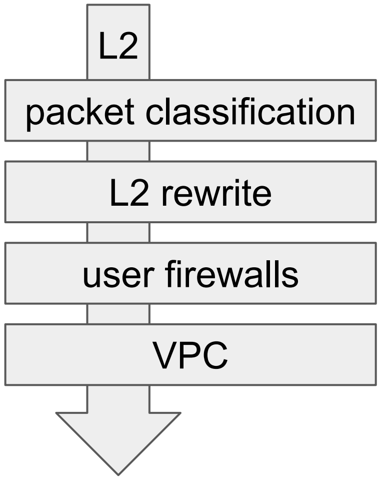
Bespoke HV Services

- hvflow
  - writes metadata about droplet interfaces to ovsdb
  - programs OvS with flow rules that handle droplet networking
- announced
  - watches ovsdb to learn droplet addresses
  - announces host routes to route servers to draw ***ingress*** traffic over L3 network
- respond
  - droplets ***think*** they are still on an L2 network
  - special MAC (fe:00:00:00:01:01) steers ***egress*** traffic to the L3 network
  - handles all ARP, NDP, and ICMP echo (gateway only) requests

L3 Pivot

L3 Pivot

- Keeping the lights on was a large consideration in designing our network architecture
  - The pivot occurs without evacuating droplets from the HV



Simplified OvS table structure

- matches on per-droplet fields; IP, MAC, port, drops otherwise

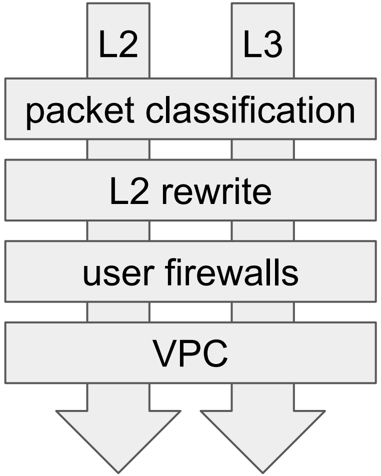
- matches on IP, MAC, vlan; tag/strip vlan

- implements user defined firewall rules

- performs VPC encap/decap

L3 Pivot

- With parallel data paths, flows were *backfilled* for existing droplets
  - Parallel data paths for legacy L2 and L3 traffic exist in network fabric and OvS tables
  - Allowed for L3 flows to be added without disrupting L2, and prior to traffic pivot
  - L3 flows are flagged for easy removal by setting a bit in the cookie field



Simplified OvS table structure - L3 data path

- fwd ARP/NDP/ICMP to respond; otherwise matches on port, MAC, etc

- L3 matches on fe:00:00:00:01:01 or L3 vlan

- implements user defined firewall rules

- performs VPC encap/decap

L3 Pivot

Install:
- Dormant L3 flows
- L3 services
- L3 config in fabric

Enable:
- Droplet reachability announcements on core switches

- Install Flows to fwd ARP/NDP/ICMP to respond proxy
- Trigger GARP from respond to droplet

Legacy L2 → L3 Backfill → L3 Ingress → L3 Egress

Thank you!