Arachne: Large Scale Data Center SDN Testing

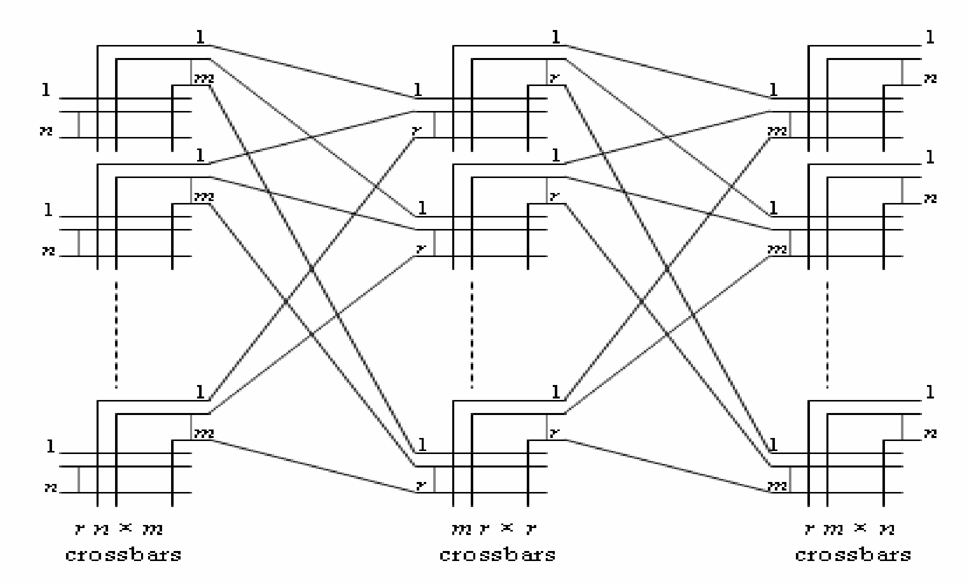
Alex Aring
Jamal Hadi Salim

Agenda

- Clos + SDN context and history
- Introduce Arachne
- Arachne Addressing + Naming
- L2 vs L3 mode
- Deployment Layout
- Workflow
- Challenges And Solutions
- Future

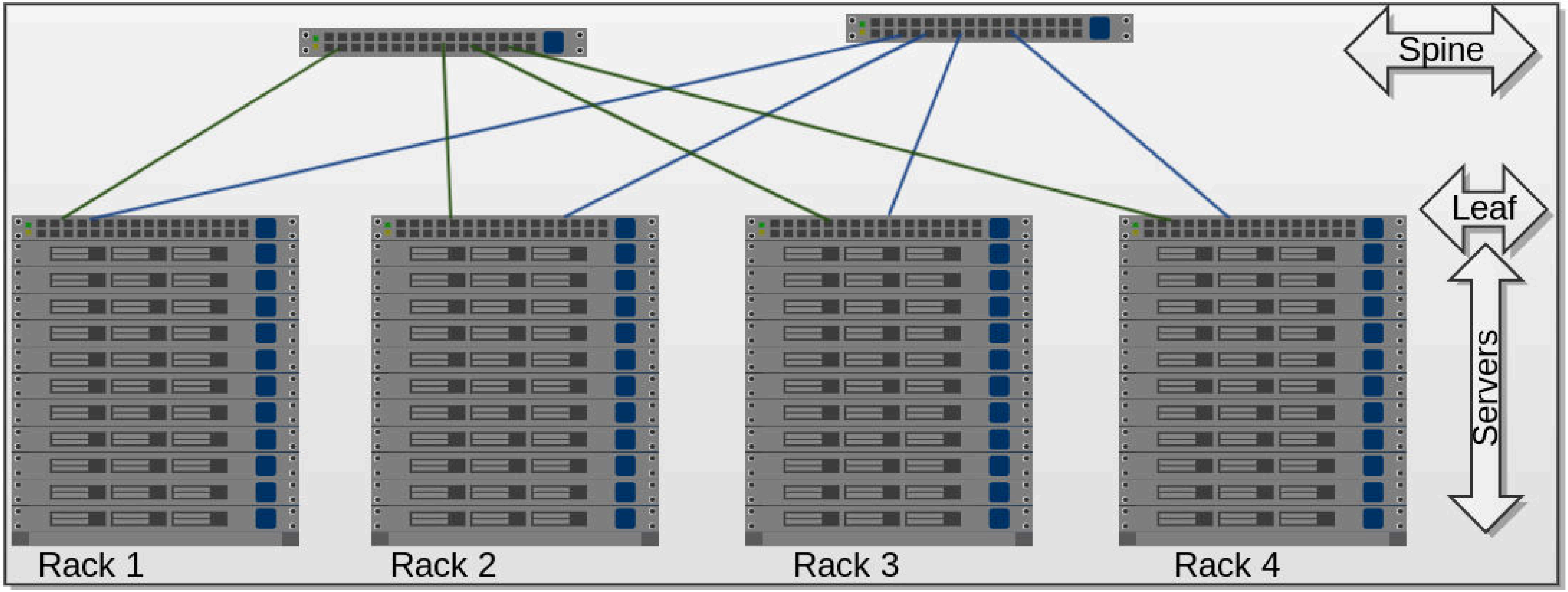
In the Beginning Was The Phone Network.....

Photo : Cassell & Co., Ltd



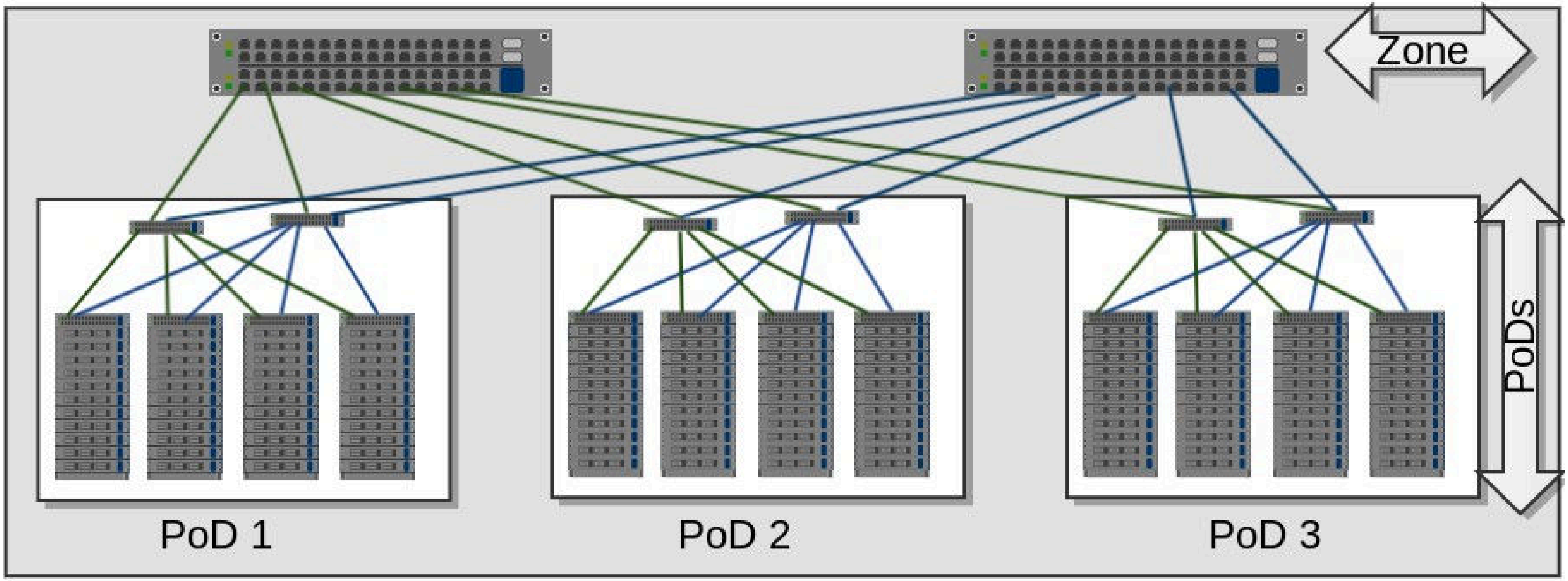
- Removing Humans From The Bridge
  - Almon B. Strowger, undertaker,
    - 1892
- Scaling and Modularity
  - Charles Clos, scholar,
    - 1952
- Separation of Control and Datapath
  - Phone phreaks and companies, SS7
    - 1975

Data Centre: 3-stage Clos Network



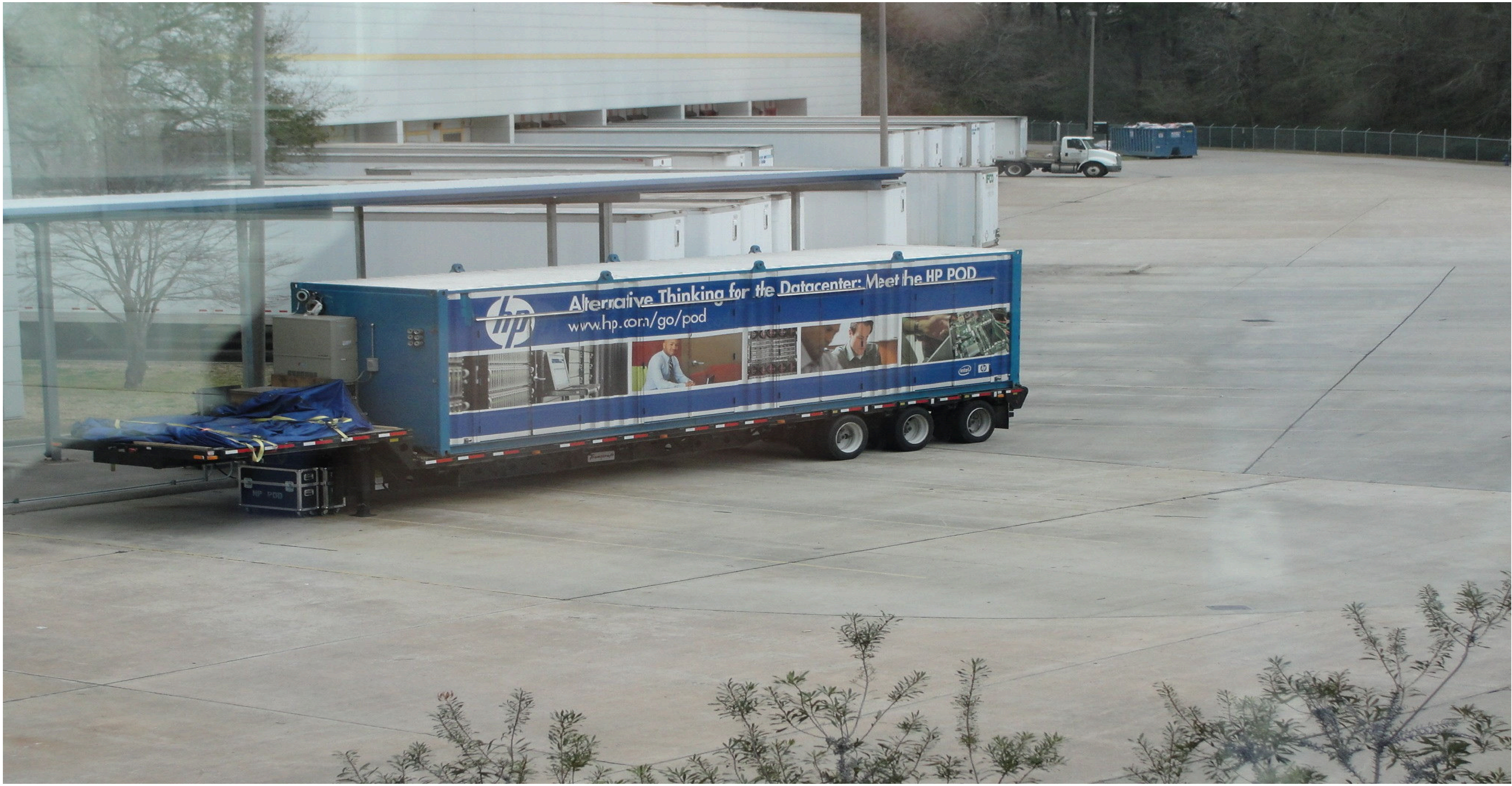
- Can wheel in a new Rack at Runtime
  - Connect cables to Spines and power up
- Arachne Design Goal

Data Centre: 5-stage Clos Network



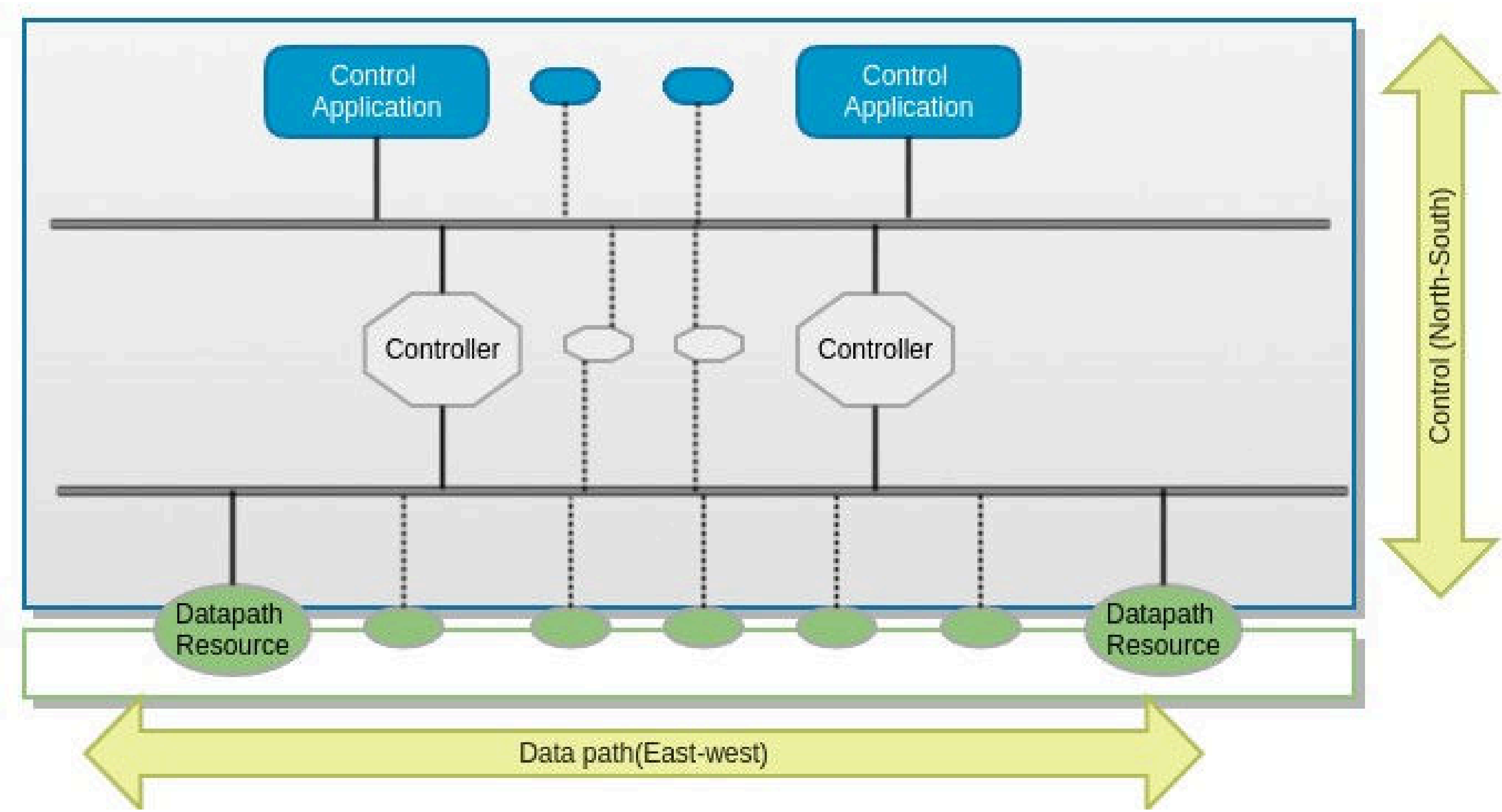
- Can Truck in a PoD at Runtime
  - connect cables to the Zones, power up
- Arachne Design Goal

Data Centre: Clos Network



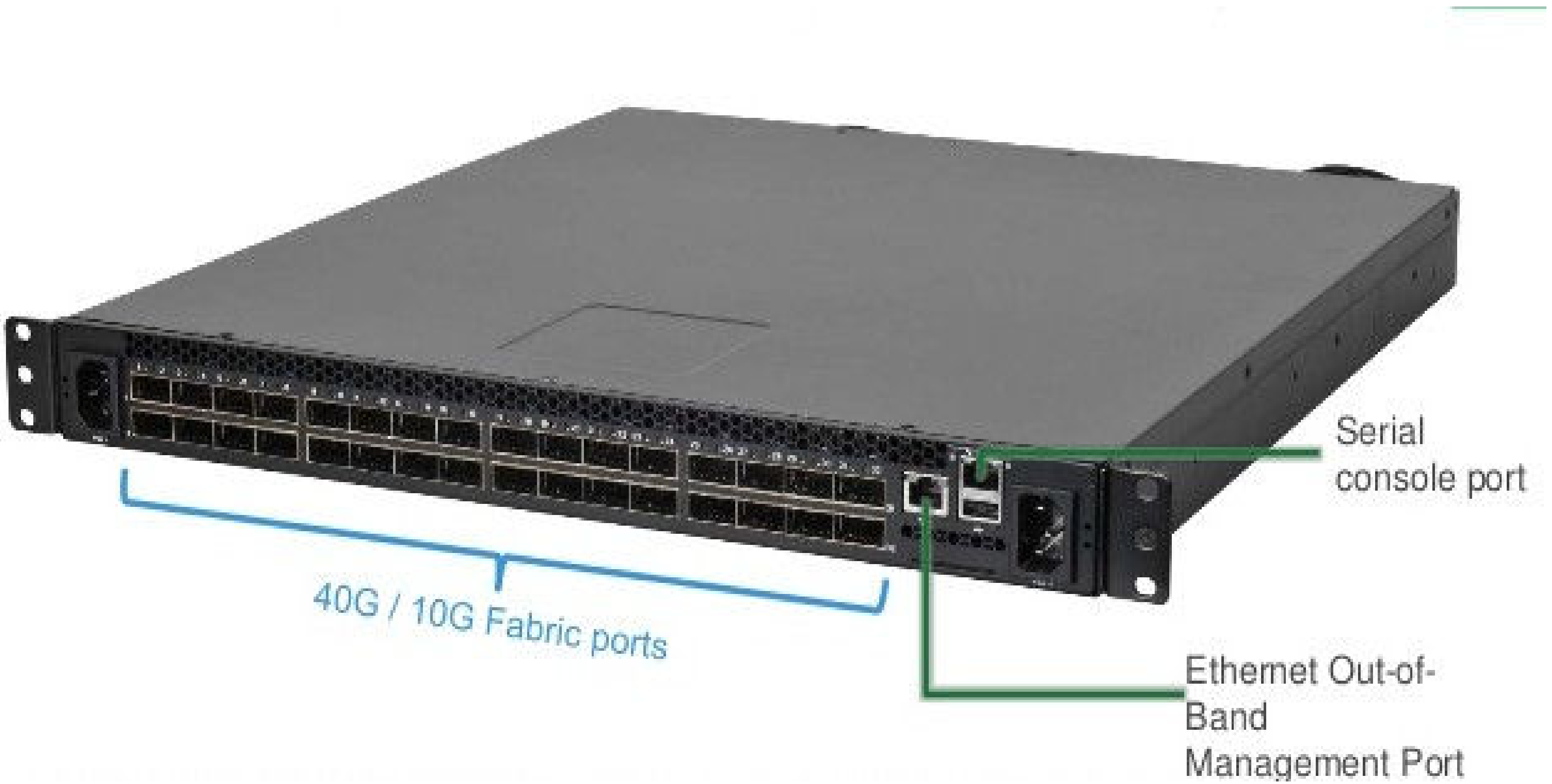
- Trucking-In a PoD

Software Define Networking



- Separate Control and Datapath Networks

Separating Control Path



- Use Management switch port

Introducing Arachne

- Control/Datapath testing
  - Small to very large scale testing of resources, controllers, applications
    - Cheaply: CPU, Memory
- Any SDN approach that uses Clos infrastructure
  - Plug in a rack or a PoD
- Reuse or create new open source components
  - <u>MUST be</u> Linux netdev based
    - Yes, we are known Linux bigots

Reuse Attempts

- VMs
  - Cumulus VX
  - Consumed too much memory and CPU
- Docker
  - Too much resources and complexity
- Mininet
  - Lightweight
  - Too specific to OF+OVS
  - Proprietary topology definitions
- Ansible
  - Static playbook inventories vs dynamic design
  - More dependencies with packaging

Arachne Components

- Patched Iproute2
- Patched Linux Kernel
  - Bridge, IP forwarding
- Python 3
- Dot file
- Qemu

Arachne Addressing Design: E.164

- Influence from E.164 in the telephony world
  - Country Code => ZoneID
  - Area Code => PoDID
  - Subscriber => Depends on type of node (host/leaf/spine/zone)
- Why geographical Addressing?
  - Simplifies automation (wheel/truck in a rack/PoD)
  - Simplifies debugging
  - Simplifies switching/routing
  - Simplifies policy management

Arachne Addressing Design

MAC Address

OUI	ZoneID	PODID	Role	Dir	PortID
(24b)	(3b)	(6b)	(2b)	(2b)	(11b)

Zone Switch IPv4 Address (NH selection)

"169.254"	Role	ZoneID	Zone SwitchID	"1"
(16b)	(2b)	(3b)	(6b)	(5b)

Spine Switch IPv4 Address (NH selection)

"169.254"	Role	PoDID	SpineID	"1"
(16b)	(2b)	(6b)	(4b)	(4b)

Leaf Switch IPv4 Address (NH selection)

"169.254"	Role	PoDID	RackID	LeafID	"1"
(16b)	(2b)	(6b)	(4b)	(2b)	(2b)

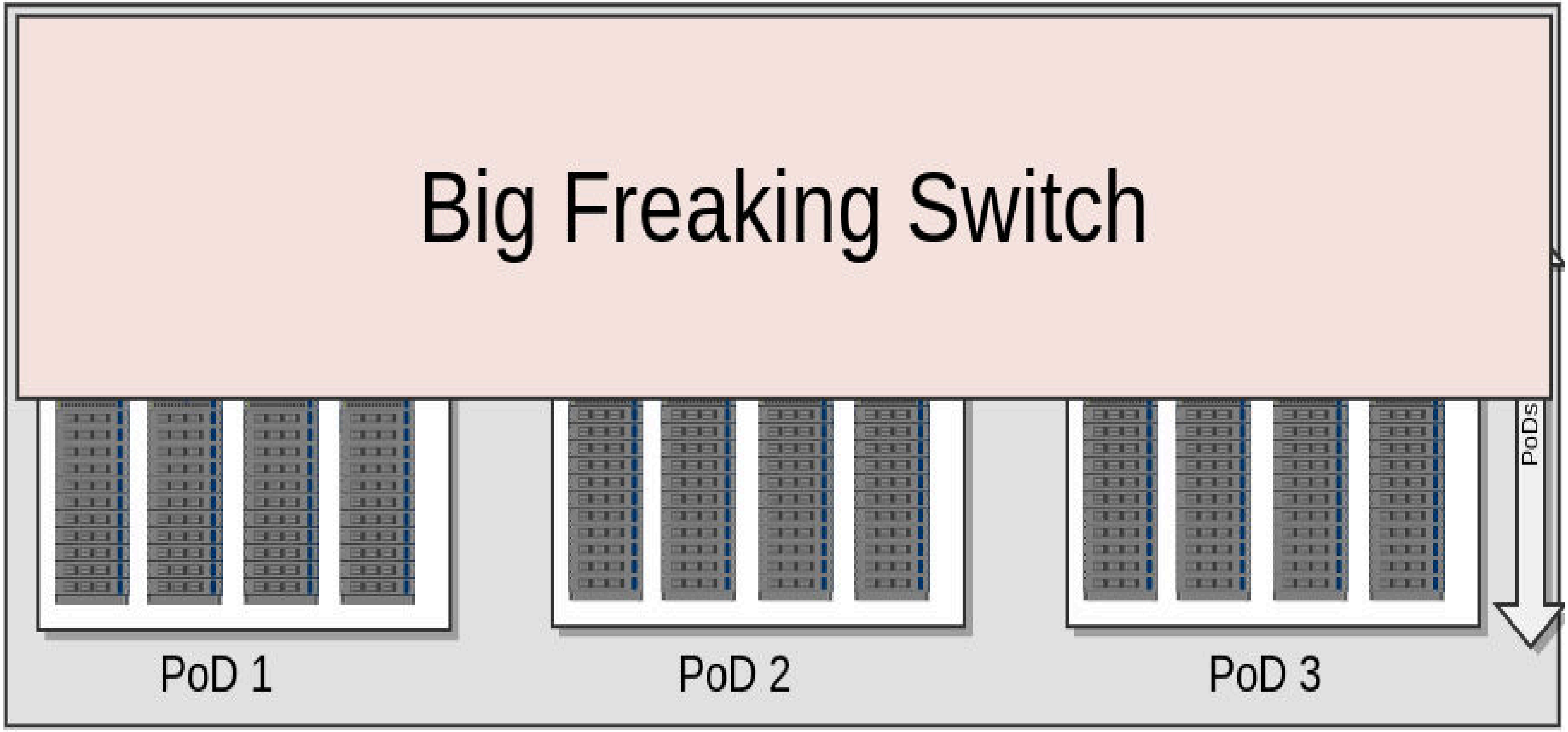
Host IPv4 Address

"10"	ZoneID	PODID	Rackid	NodeBits Available for use
(8b)	(3b)	(6b)	(4b)	(11b)

Arachne Node Name Design

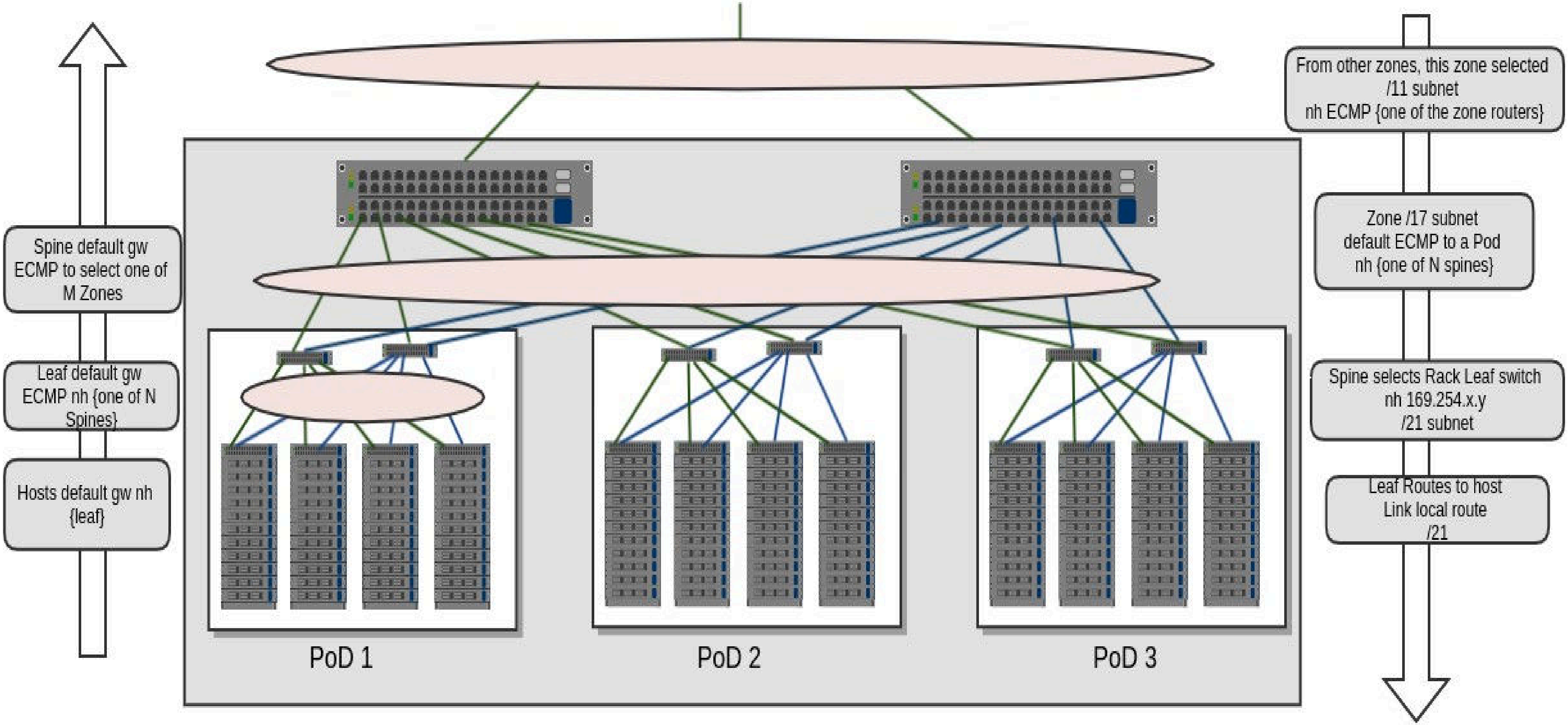
- Host
  - H<Hostid>_R<Rackid>_P<Podid>_Z<Zoneid>
- Leaf
  - L<Leafid>_R<Rackid>_P<Podid>_Z<Zoneid>
- Spine
  - S<Spineid>_P<Podid>_Z<Zoneid>
- Zone
  - ZS<Zone switch id>_Z<Zoneid>

Arachne L2 Mode



- Simple
- One big broadcast domain
  - STP to avoid loops

Arachne L3 Mode



Spine default gw
ECMP to select one of
M Zones

Leaf default gw
ECMP nh {one of N
Spines}

Hosts default gw nh
{leaf}

From other zones, this zone selected
/11 subnet
nh ECMP {one of the zone routers}

Zone /17 subnet
default ECMP to a Pod
nh {one of N spines}

Spine selects Rack Leaf switch
nh 169.254.x.y
/21 subnet

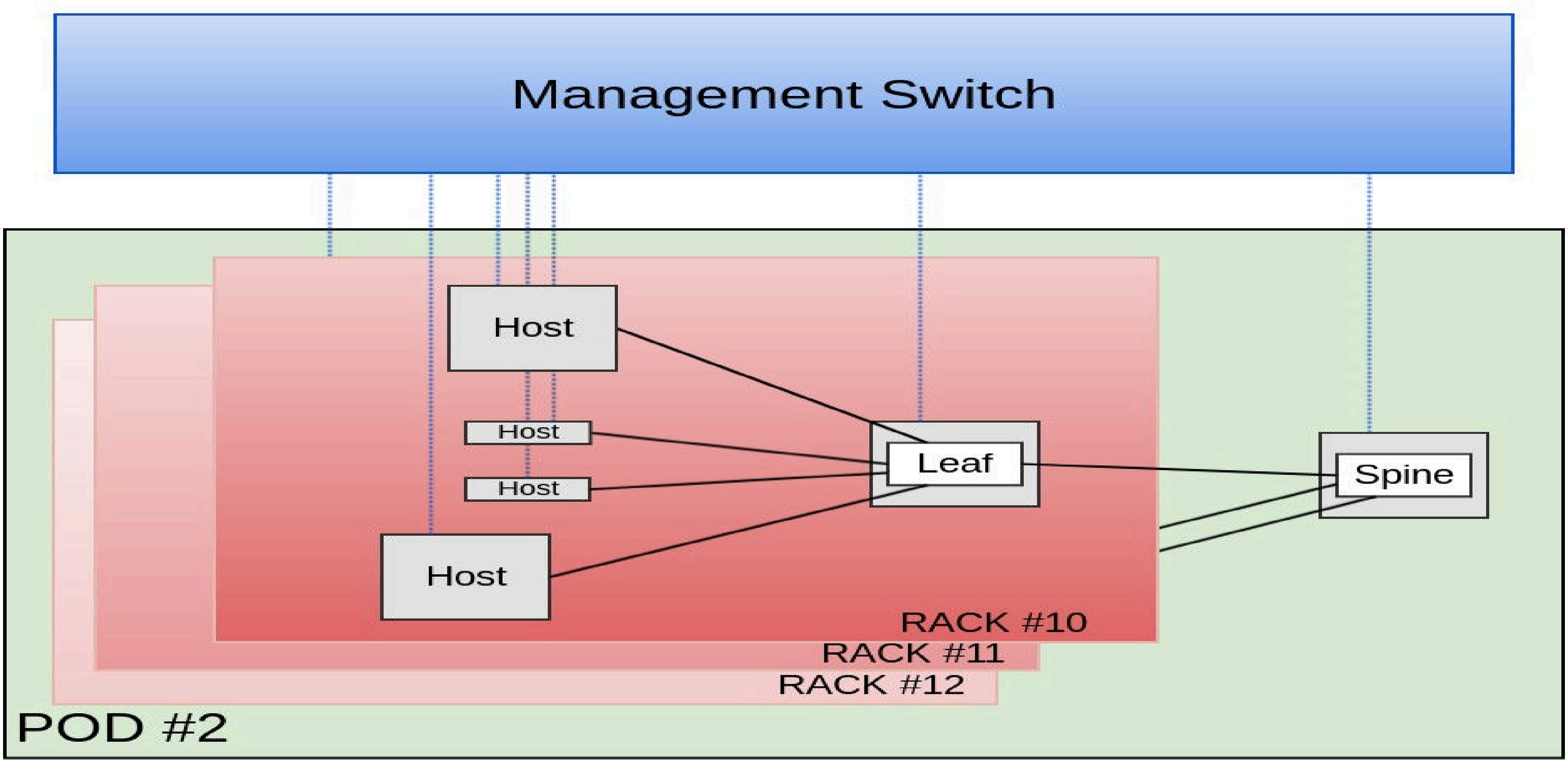
Leaf Routes to host
Link local route
/21

- Static Routing
- ECMP in presence of multiple next hops

Getting Intimate With Arachne

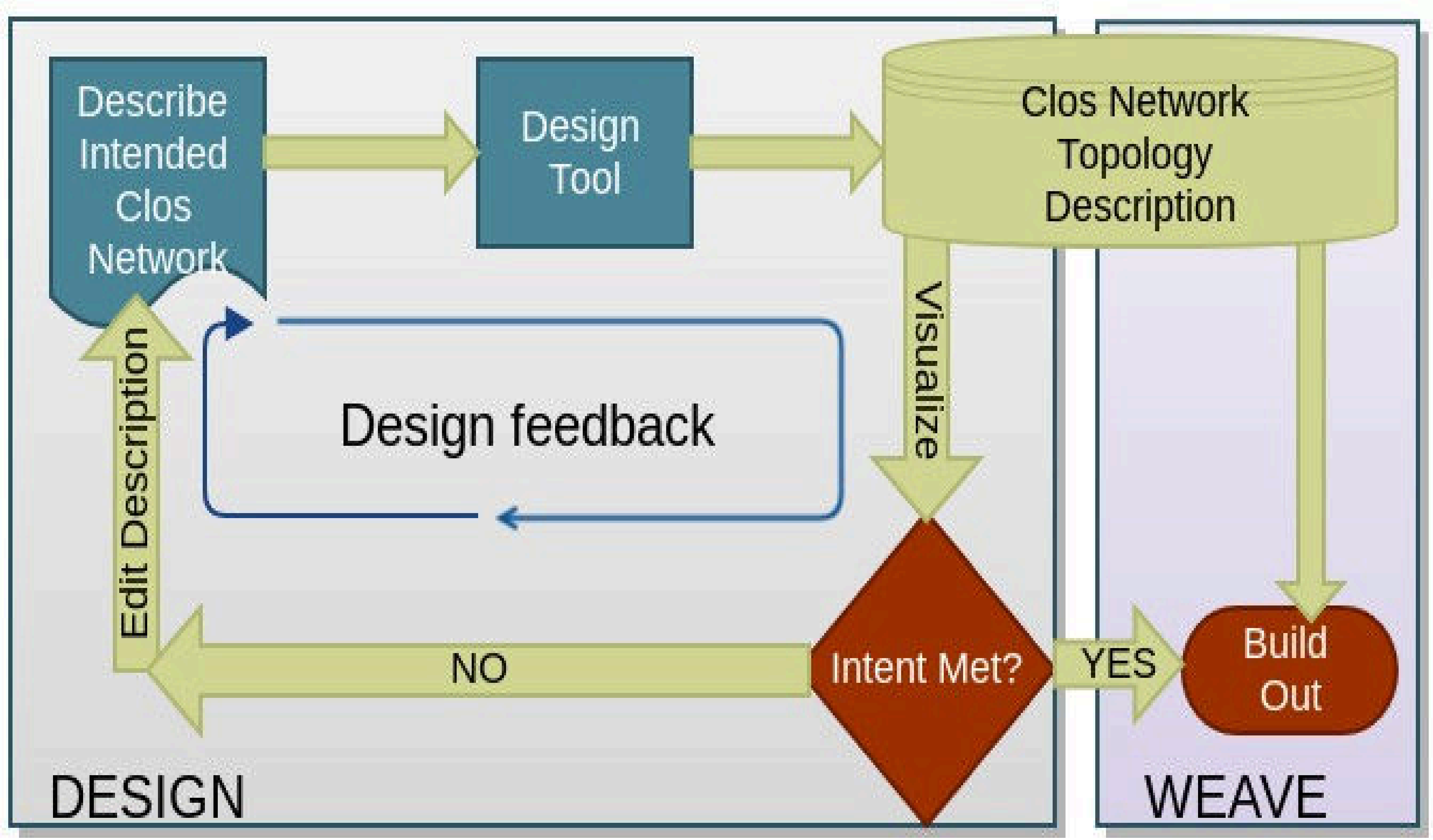
- Constitutes two parts
  - A fabric design component
  - A fabric weaving component

Arachne Container Deployment



- Each node (host/leaf/spine/zone) is a container
- Switches are Linux Bridges inside containers
- Veth as a port/cable

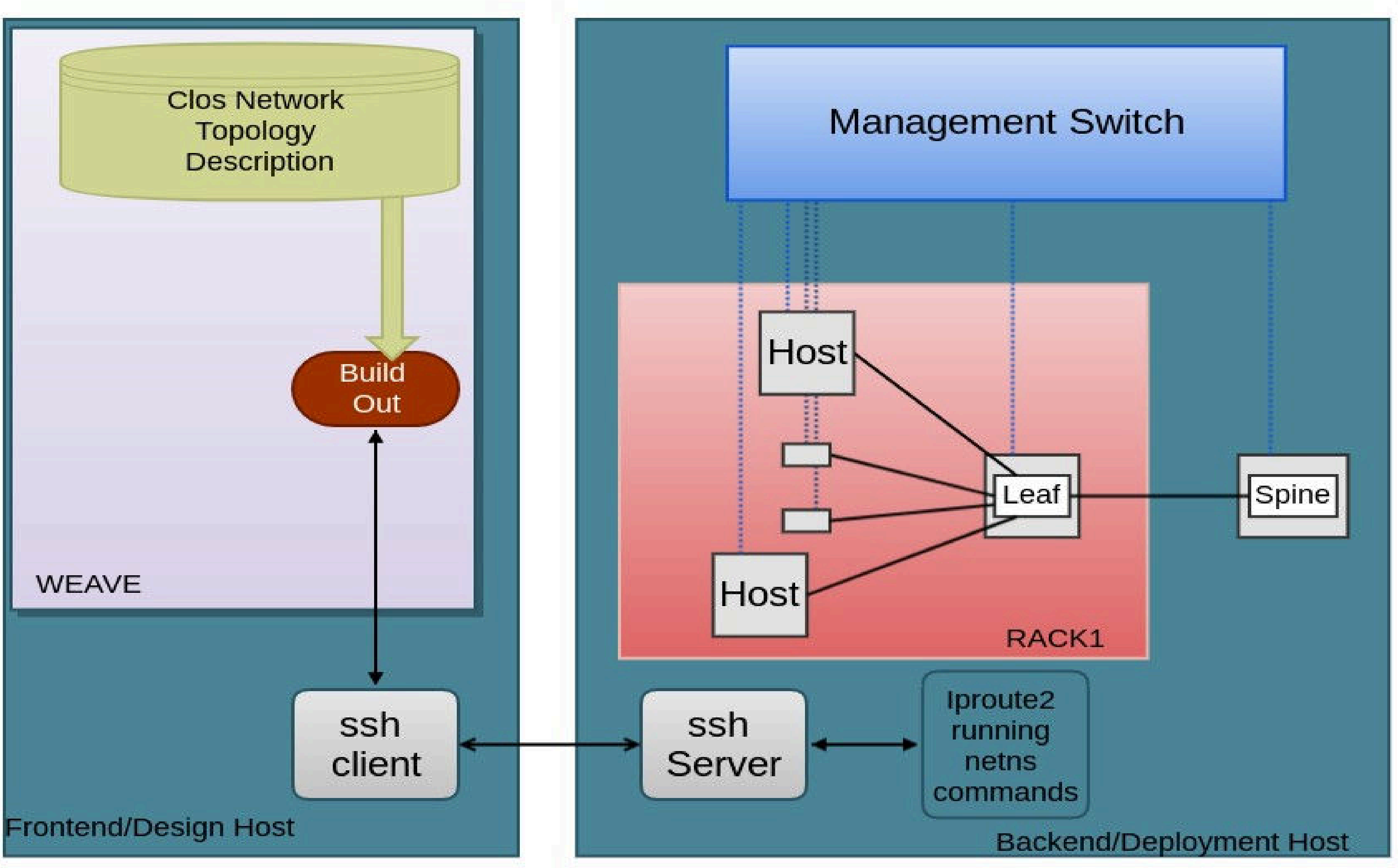
Arachne Workflow



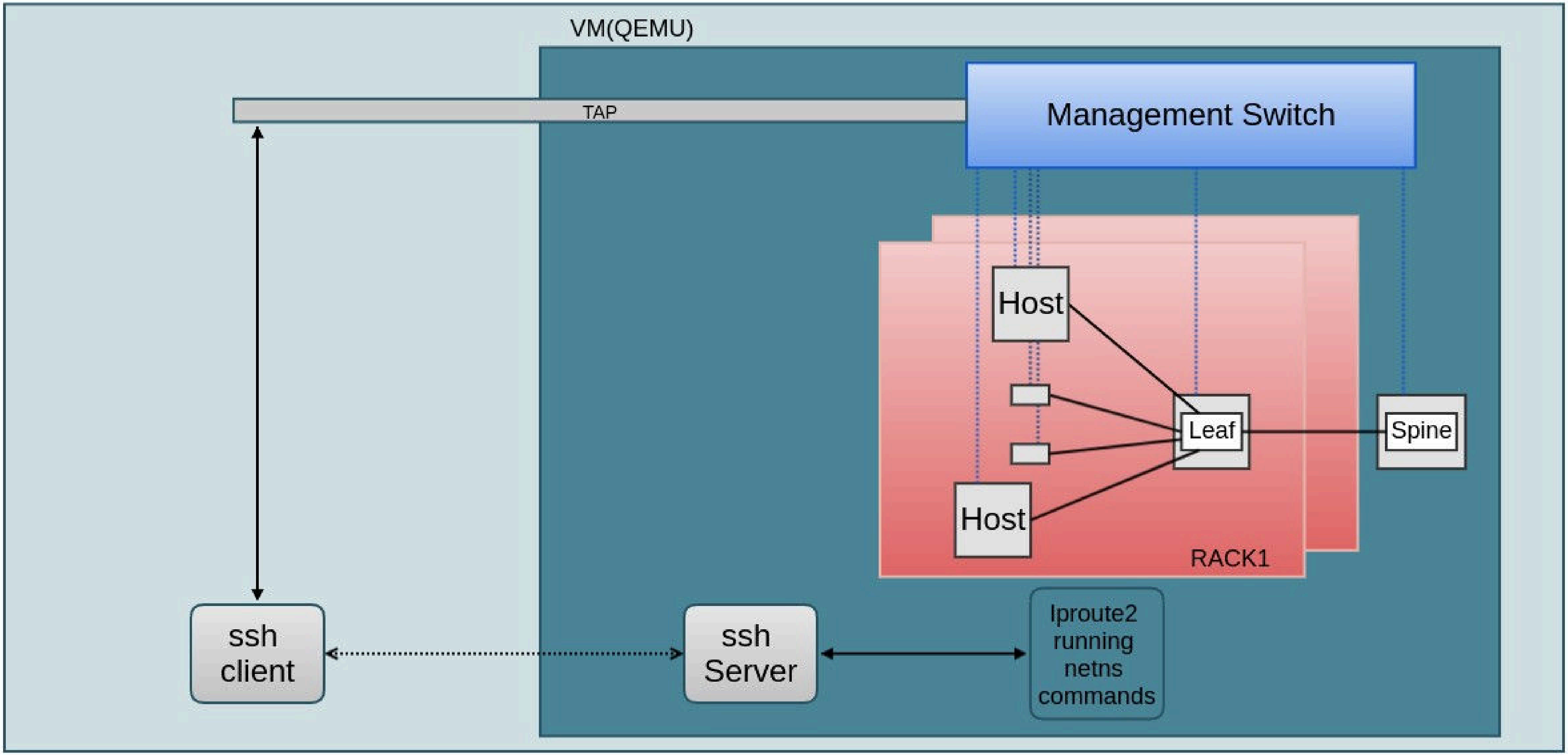
Network Intent Description

- Number of zones Needed
  - Only one zone is supported for now
- Number of PoDs in a Zone
- Number of spines in a PoD
- Number of racks per PoD
- Number of hosts per rack
  - Arachne supports a single leaf switch per rack for now

Demo: Designing With Arachne



Demo: Weaving with Arachne



Trials And Tribulations: Tooling Issues

- Iproute2 hostname
  - Iproute2 patch
- Veth names
  - Fix our naming conventions
- DHCP and IP binding
  - Use dhcp client hooks for binding
- IPv6 stateless autoconfig
  - Disable IPv6
- Python2/3 mess
  - Static binaries; use pyinstaller

Trials And Tribulations: Bridge Issues

- LLC not respected by veth
  - Patch kernel
- Bridge favoring lowest MAC address as source
  - Management ARP confusion
  - Fix MAC address on bridge

Trials And Tribulations: Scaling Issues

- Management DHCP slowing us down
  - Use static IP addresses
  - 192./8 saga
  - 192.168/16 insufficient
    - Use 25./8
- Bridge port limit of 1024
  - Patch Kernel
- ARP table overflow
  - Tweak ARP GC params
- Shared fs resulting in running out of fds
  - Increase fd limits

Future Work

- Runtime Addition of Racks and PoDs
- Publish numbers on large size networks
- Use embedded NIC switches and physical switches
- IPv6
- 7-stage Clos
- Constrained Design Templates
- Chaos Monkey
- Open Source

Attribution: Images

- “Women operators working at McGill Montreal, Quebec, Canada”
  - https://commons.wikimedia.org/wiki/File:Telephone_exchange_Montreal_QE3_33.jpg
- Original Clos Network
  - https://commons.wikimedia.org/wiki/File:Closnetwork.png
- HP PoD
  - http://storagenerve.com/wp-content/uploads/2010/03/DSC00086.jpg