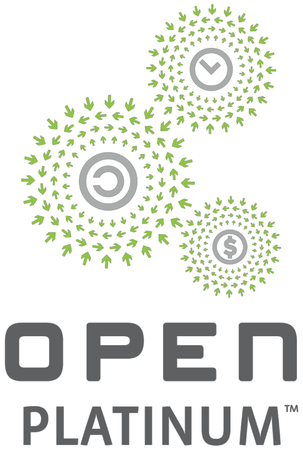
Open. Together.

OCP
REGIONAL
SUMMIT

CLOS design with Facebook Minipack

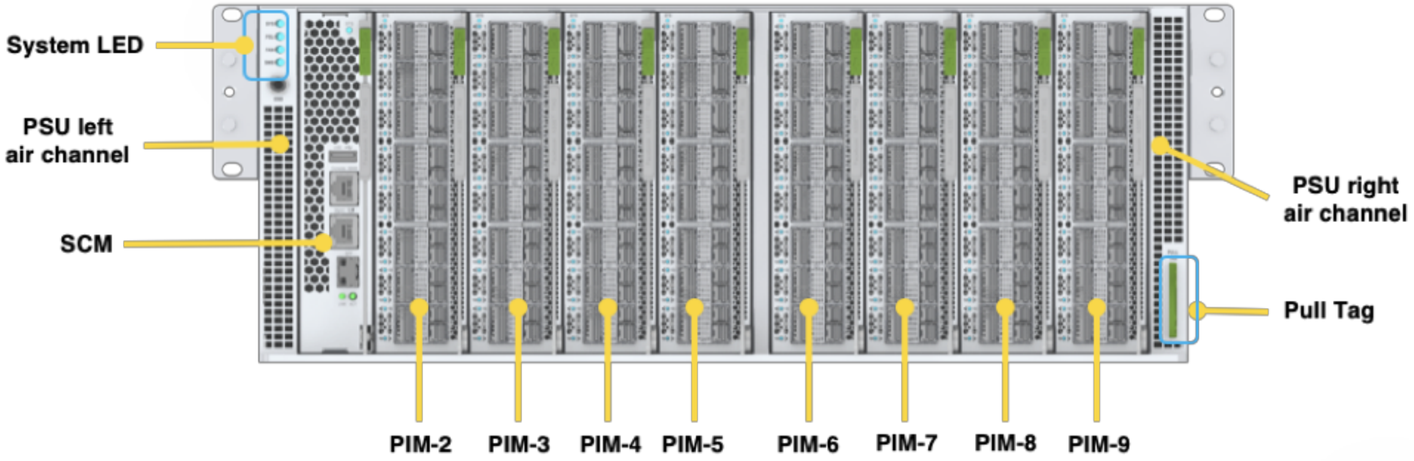
In Enterprise environments

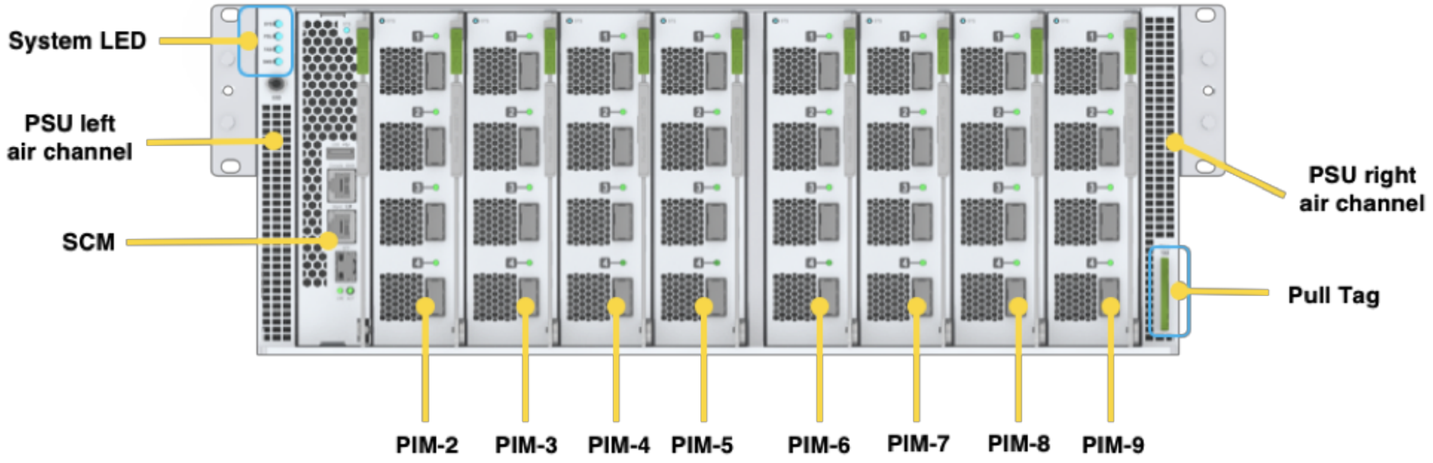


Cumulus Networks | Andreas la Quiante

Systems Engineer …, CNX$_{1999}$, … CCIE, … , CCONP 2019::9$_{2019}$, …

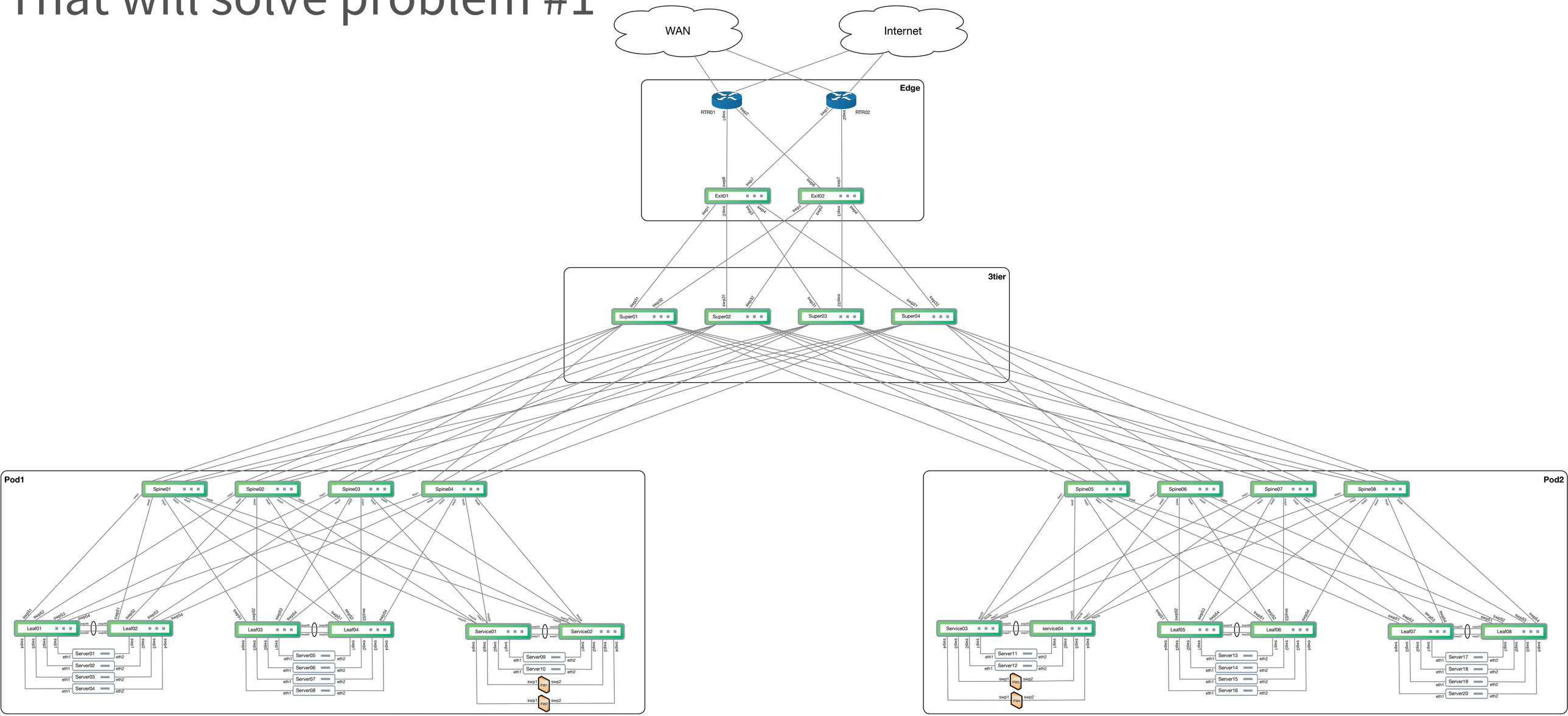
Once up on a time, there was a switch



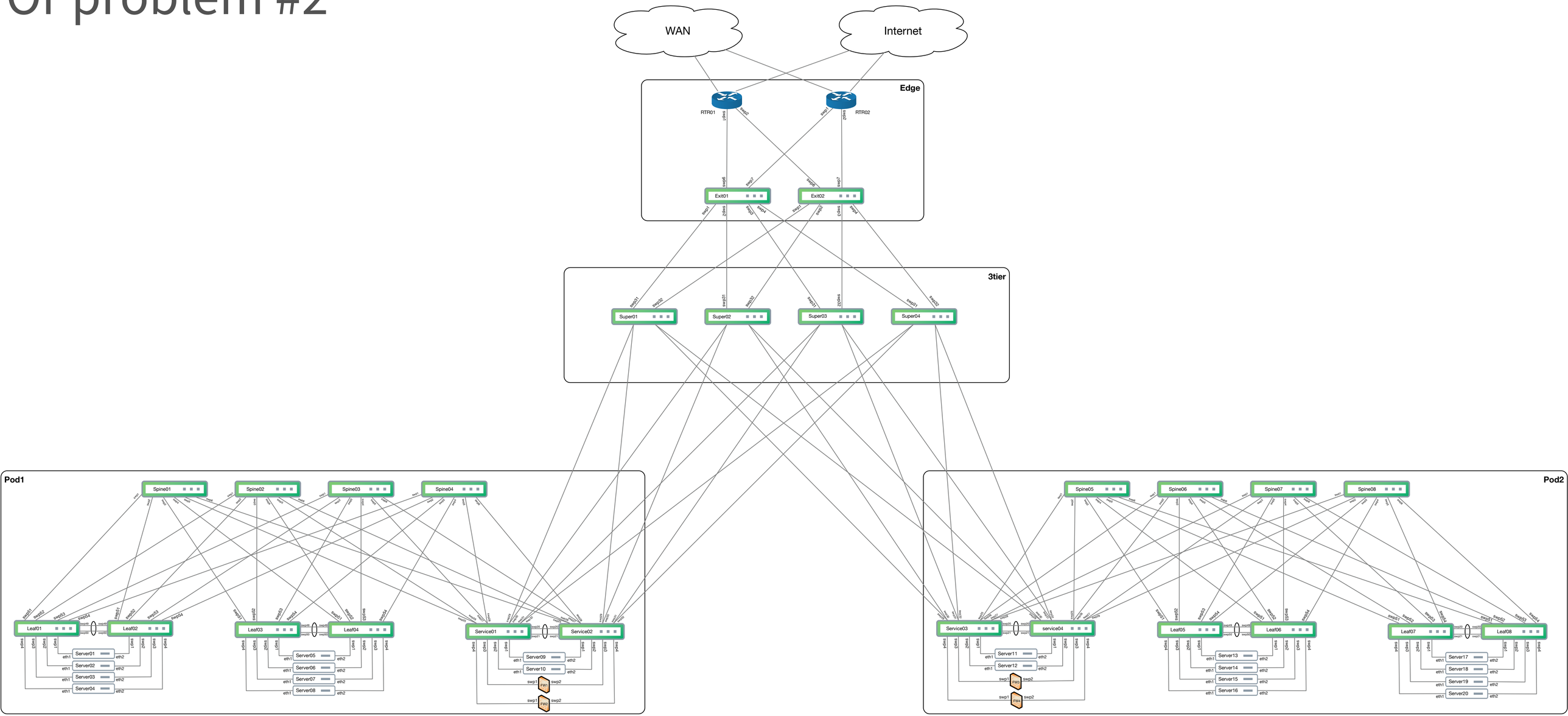


PIM == Port Interface Module

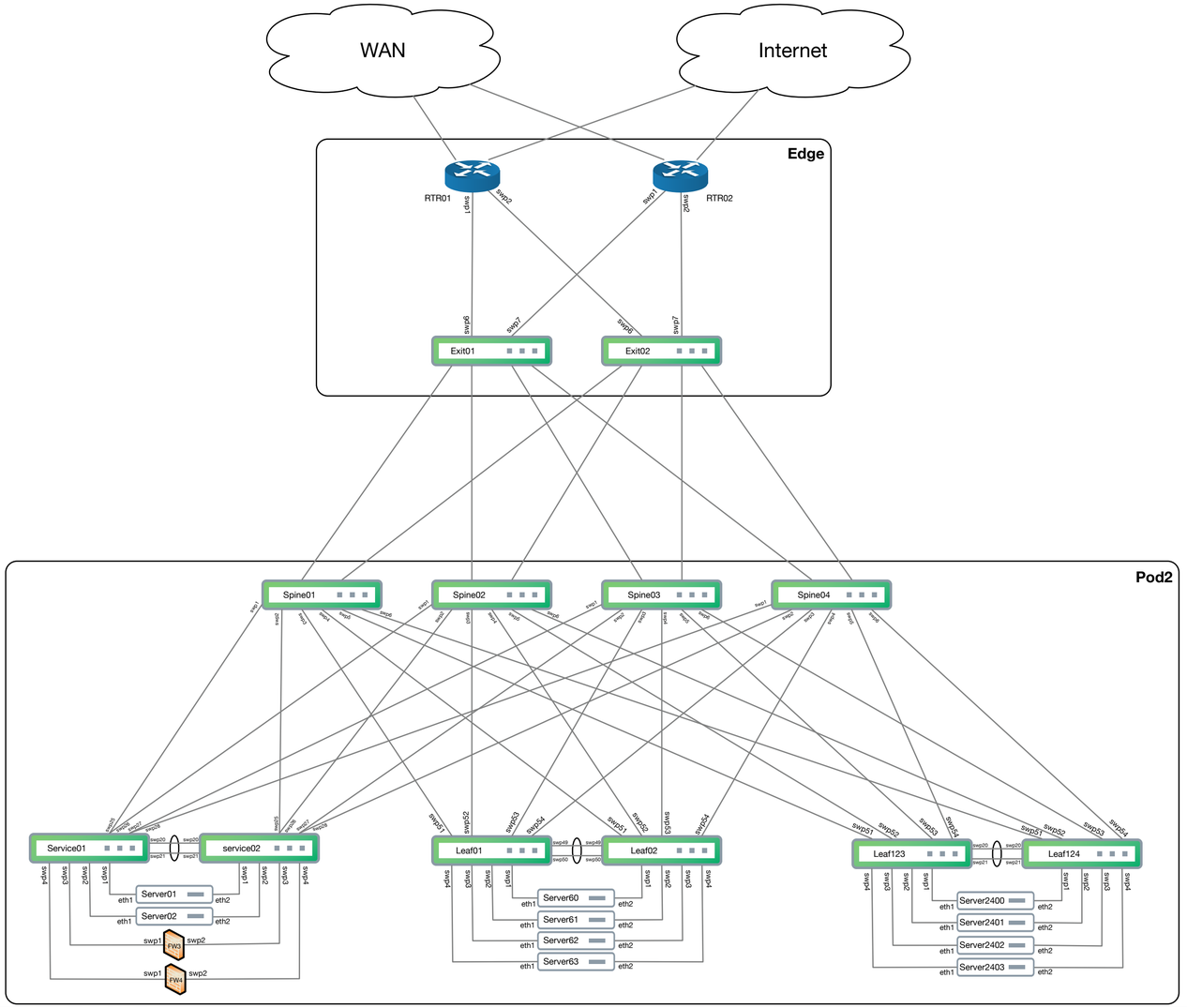
That will solve problem #1



Or problem #2


WAN
Internet
Edge
RTR01
RTR02
Exit01
Exit02
3tier
Super01
Super02
Super03
Super04
Pod1
Spine01
Spine02
Spine03
Spine04
Leaf01
Leaf02
Leaf03
Leaf04
Service01
Service02
Server01
Server02
Server03
Server04
Server05
Server06
Server07
Server08
Server09
Server10
FW1
FW2
Pod2
Spine05
Spine06
Spine07
Spine08
Service03
service04
Leaf05
Leaf06
Leaf07
Leaf08
Server11
Server12
Server13
Server14
Server15
Server16
Server17
Server18
Server19
Server20
FW3
FW4

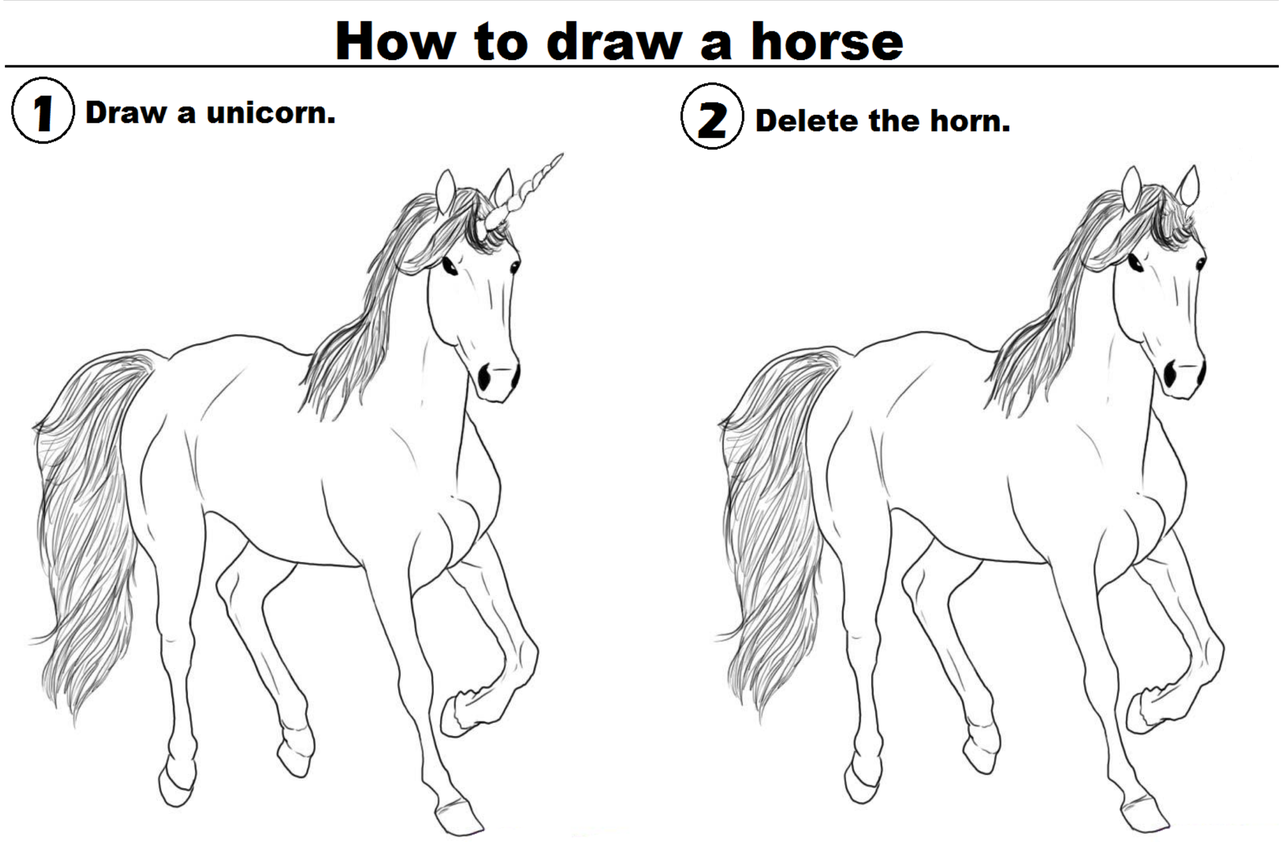
And simplified the datacenter network



Or not ?

Design caveats

- Physical network implementation
- Minipack use-case
- Tomahawk 3, speed over features
- Pod redundancy
- Traffic flows
- Optical costs
- 400G -> 100G



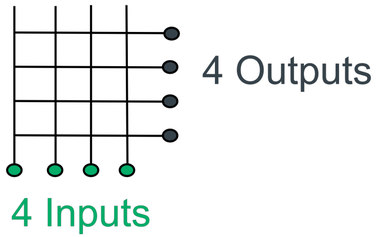
Clos concepts

In late 1930's, scale problem emerged as telephone networks grew.

- Telephone network backbone/core should be nonblocking
- Circuit switching was still electromechanical

Inputs & Outputs connected at ***crosspoints***

Every input to reach every output needs a ***full mesh (n^2)***



4 * 4 = 16 Crosspoints

Clos concepts

Scaling relied on building switches with greater numbers of ports

Existing design strategy not sustainable/scalable

- As # lines grow in a single switch, number of crosspoints explodes
- 100 inputs = 10,000 | 200 inputs = 40,000 | 500 inputs = **250,000**
- Number of crosspoints approximated cost of the switch.
- **Crosspoints in backplane was limiting factor (cost)**

Scalability can’t come from single switches with more ports

Sound familiar?

Physical network & Cost

Physical considerations

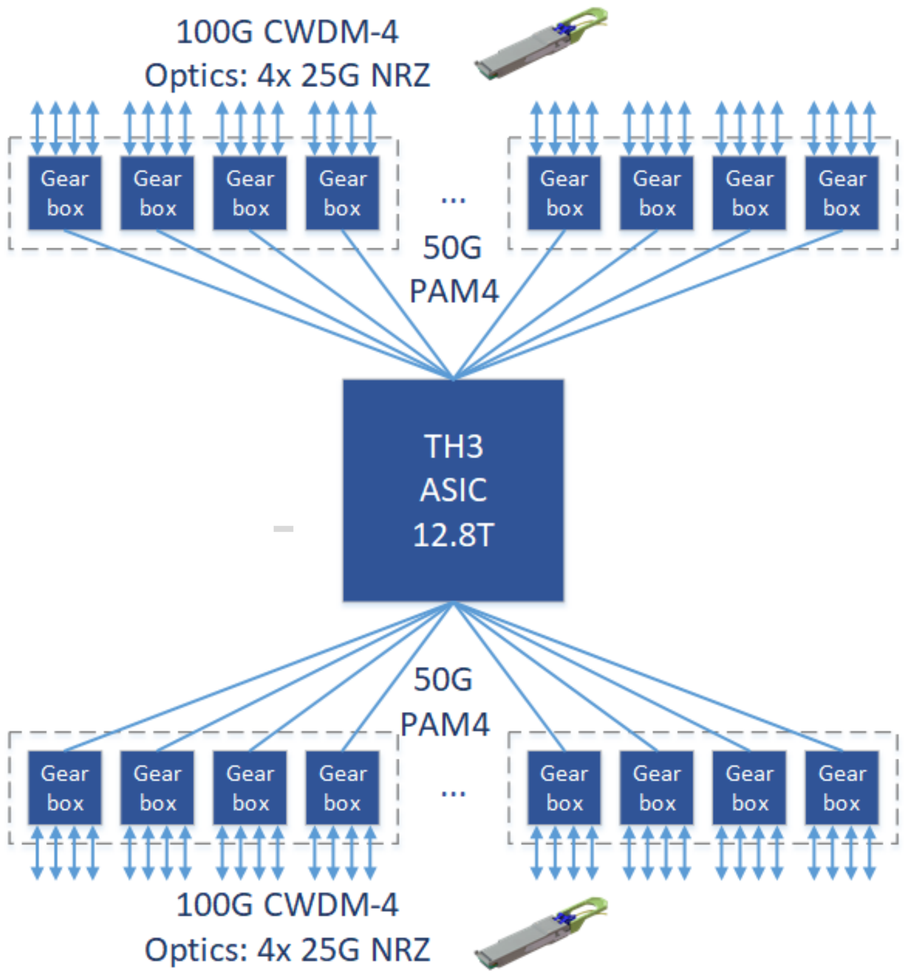
- Physical datacenter floors
- Locations of minipack
- Cabling routes
- Fiber types
- Optical costs

- Room to grow

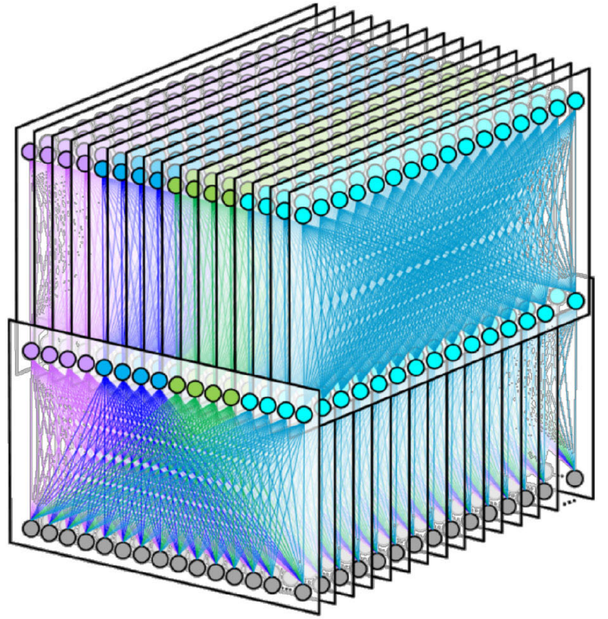
Use-cases & Tomahawk3

Minipack use-case

- Designed as fabric switch (spine)
- Servers are not directly connected
- No VxLAN encap/decap support

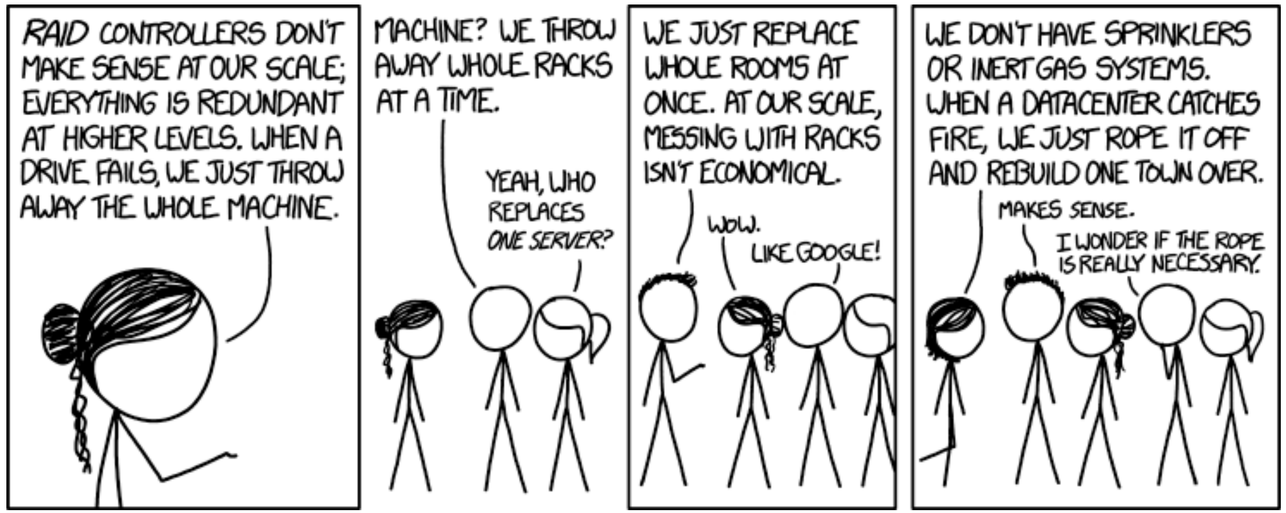


Redundancy



Redundancy with Minipack

- One pod design
- Number of Minipacks
- Edge “pods”



Traffic flows & Applications

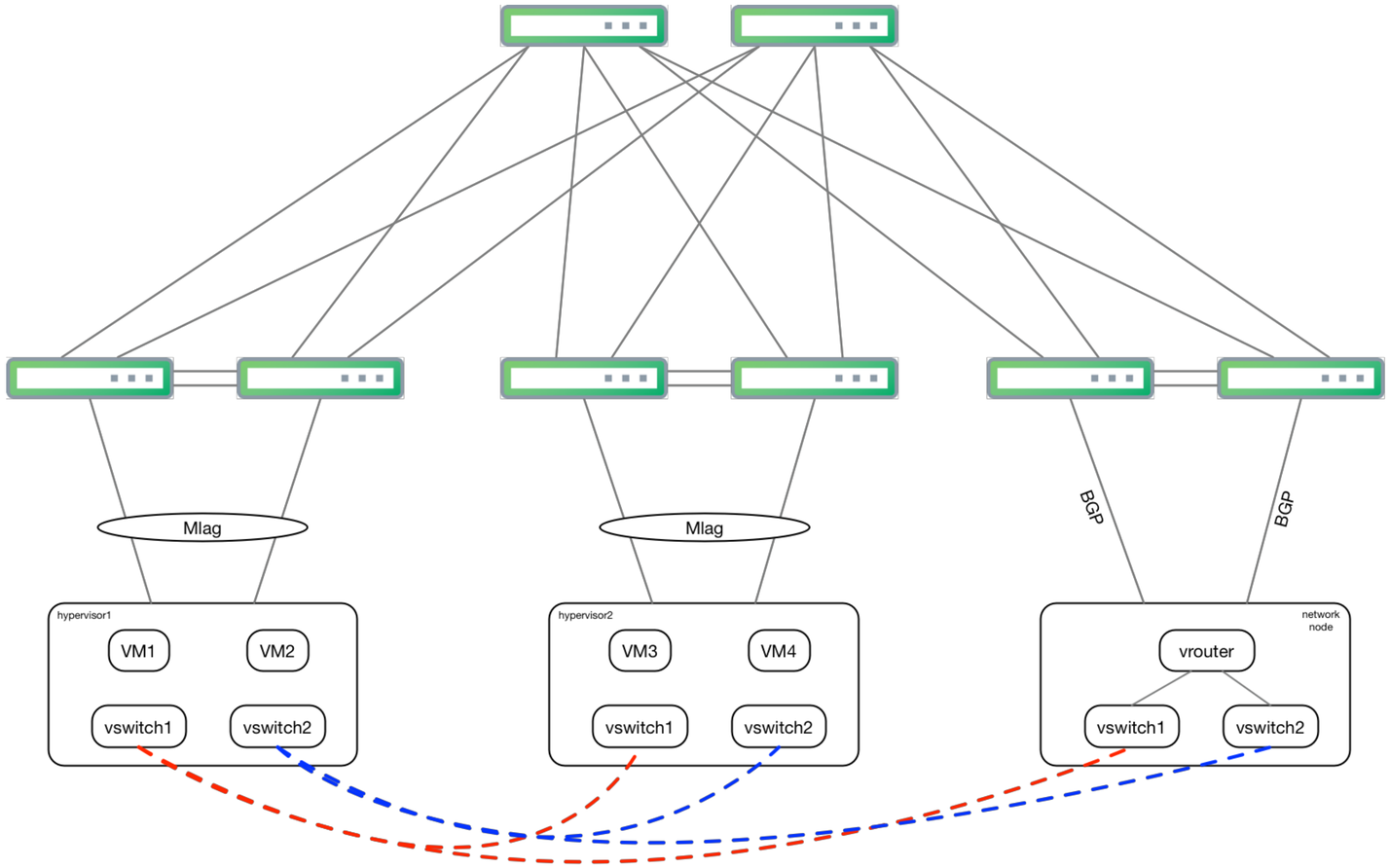
Defining traffic flows

- East-west vs. North South
- Storage
- Application traffic
- EVPN-VxLAN…… (BUM traffic (head-end-replication vs. EVPN-PIM))
- Kubernetes…….... (CNI (routing vs. overlay))
- Openstack
- Vmware NSX…… (Exit points)
- Bare metal hosts

Traffic flows & Applications

Overlay networks

- Prevent traffic tromboning
- Where to exit the overlay
- Overlay in Overlay
- Security



Limitations

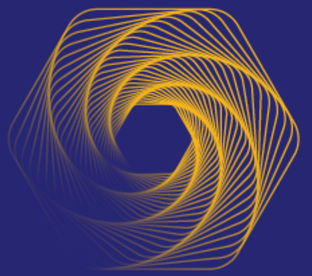
Scaling limitations

- Route entries / TCAM size
- MAC entries
- ACLs
- Route propagation
- Convergence time
- Route aggregation

Summary

Closs & Minipack

- Valid design option to simplify
- Very capable layer 3 spine
- Easy to operate/automate for L3 and EVPN-VxLAN



OCP REGIONAL SUMMIT

Open. Together.

OCP Regional Summit

26–27, September, 2019