Software Defined Data Centers

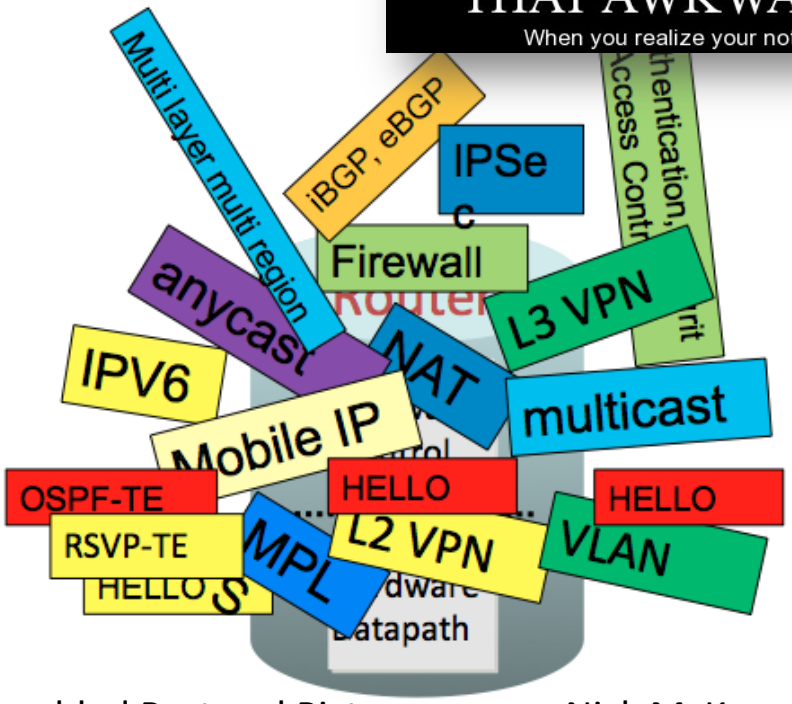
Brent Salisbury
Network Architect
University of Kentucky
brent@uky.edu

My Obligatory Rationalizing

Change is Bad

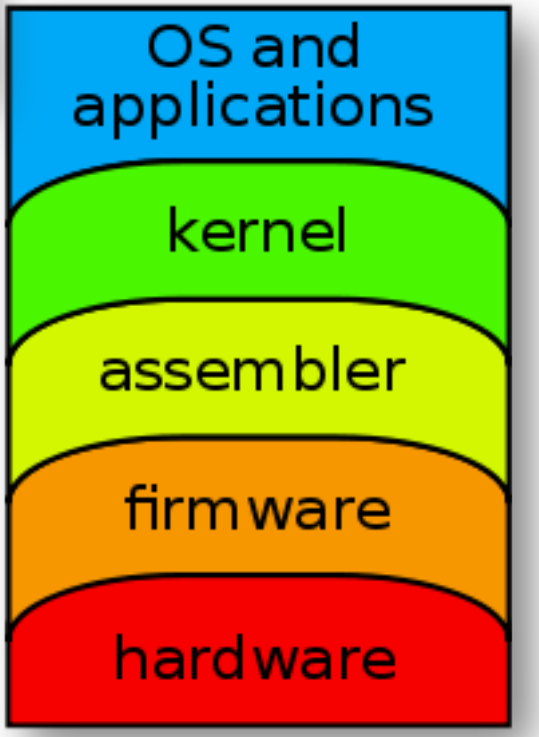
- We are operating far to close to the hardware.
  - Do systems administrators configure their services in x86 Bios? Guess what? We do.
- Generic components decomposed into resources to consume anywhere, anytime.
- Abstraction of Forwarding, State and Management.
  - Forwarding: Networking gear with flow tables and firmware.
  - State: Bag of protocols destruction.
  - Management: Orchestration, CMDB etc. Join the rest of the data center (and world)

A Quick Recap



<

Doh!



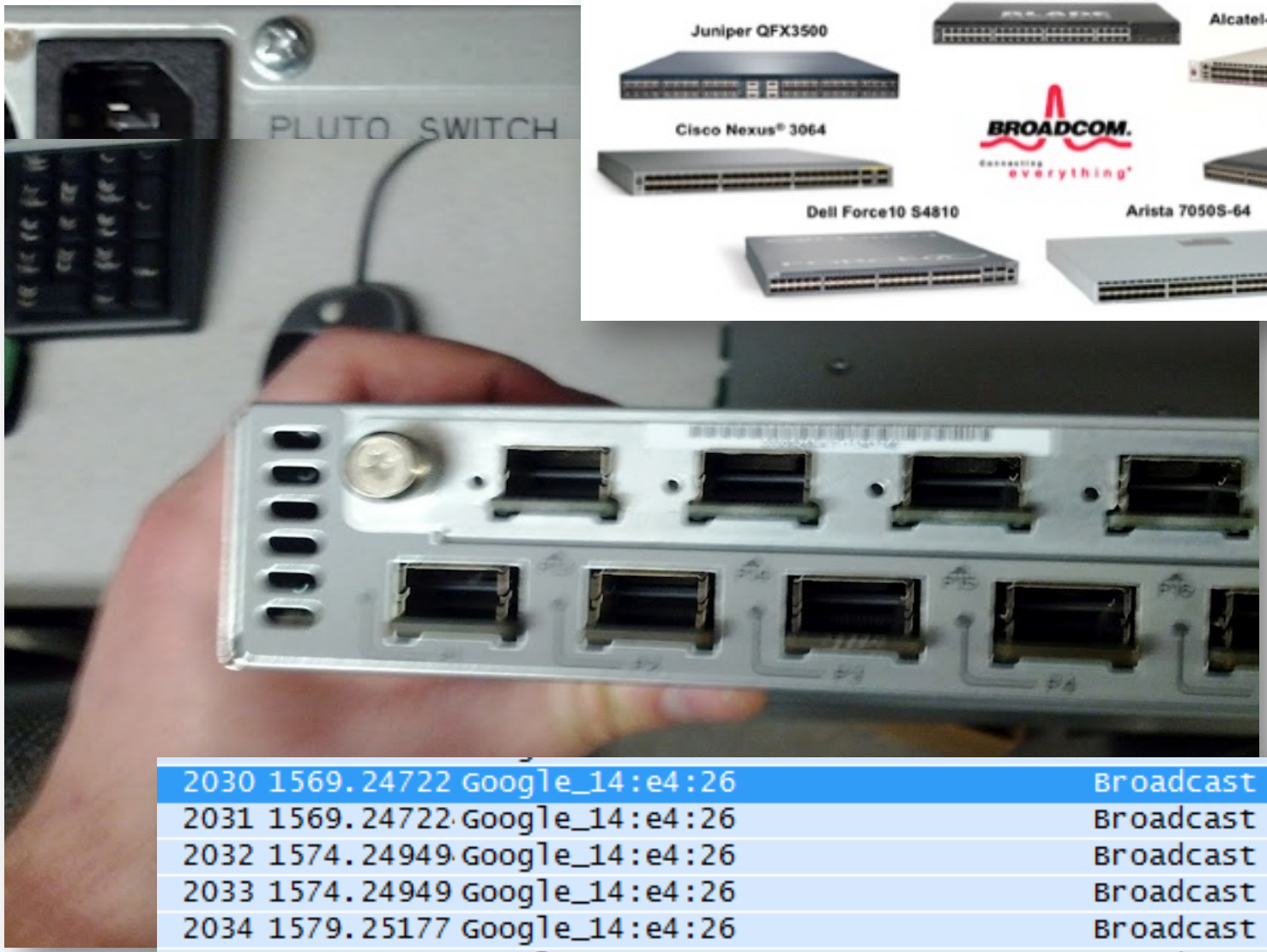
Jumbled Protocol Picture source: Nick McKeown -Stanford

The Problem Has Always Been the Edge

- Security Policy at the Edge.
- Multi-Tenancy at the Edge.
- QOS Policy at the Edge.
- Management at the Edge.
- Cost at the Edge.
- Complexity at the Edge.

Commoditization: A Collage of Disruption

Google's Pluto



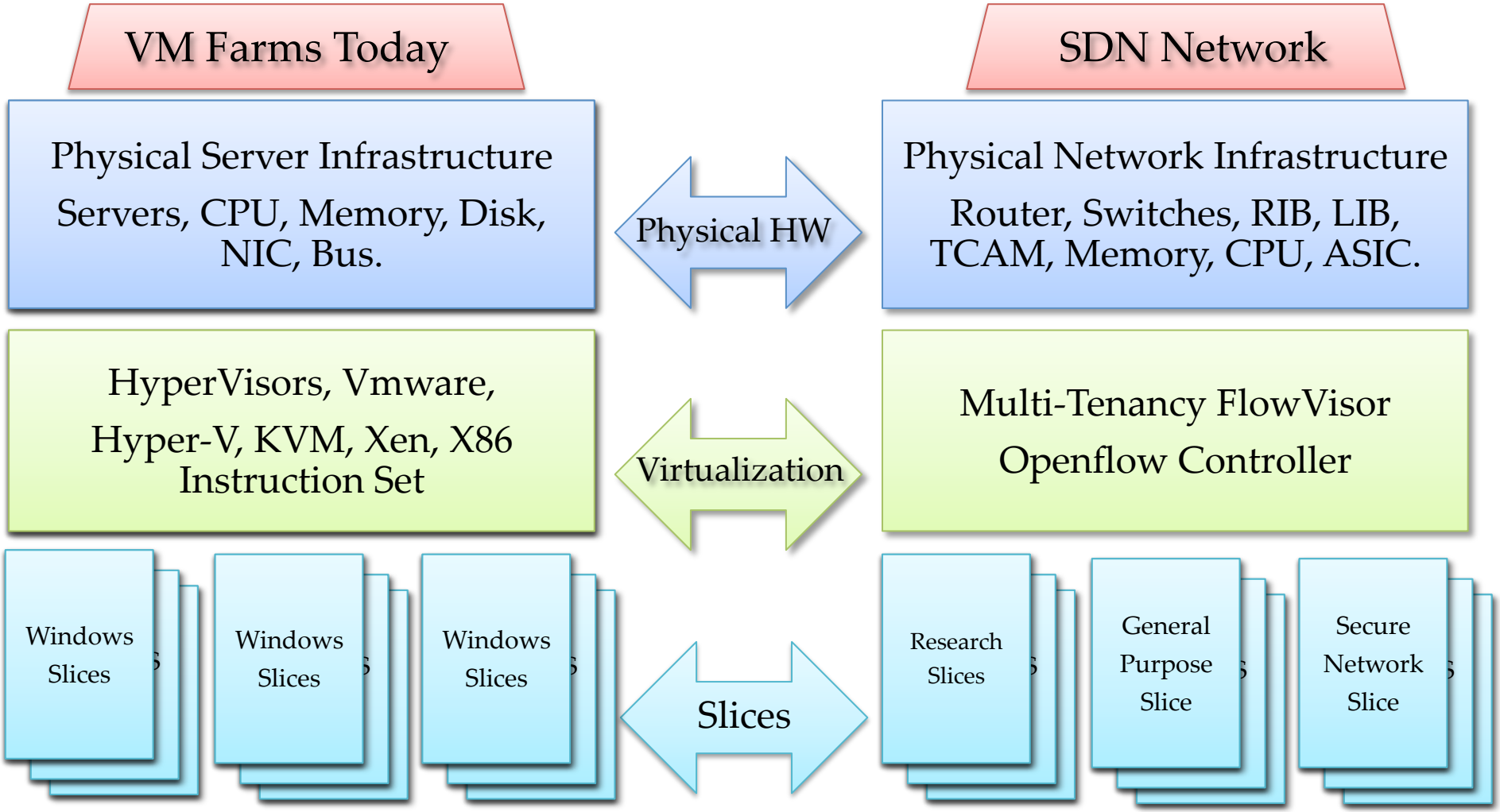
2030	1569.24722	Google_14:e4:26	Broadcast ARP
2031	1569.24722	Google_14:e4:26	Broadcast ARP
2032	1574.24949	Google_14:e4:26	Broadcast ARP
2033	1574.24949	Google_14:e4:26	Broadcast ARP
2034	1579.25177	Google_14:e4:26	Broadcast ARP
2035	1579.25177	Google_14:e4:26	Broadcast ARP

What Changed? Why Now? #1 HW Commoditization

1. *Commodity Hardware.* Off the shelf "Merchant Silicon". – If all vendors are using the same pieces and parts where is the value.

- "We want to create a dynamic where we have a very good base set of vendor-agnostic instructions. On the other hand, we need to give room for switch/chip vendors to differentiate." -Nick McKeown
- "You don't have to have an MBA to realize there is a problem. We are still ok but not for very long." *-Stuart Selby,Verizon*
- When you run a large data center it is cheaper per unit to run a large thing rather than a small thing, unfortunately in networking that's not really true. *-Urs Hoezle, Google*
- "Work with existing silicon today; tomorrow may bring dedicated OpenFlow silicon." -David Erickson
- "The path to OpenFlow is not a four lane highway of joy and freedom with a six pack and a girl in the seat next to you, it's a bit more complex and a little hard to say how it will work out, but I'd be backing OpenFlow in my view" – *Greg Ferro Etherrealmind.com*

Not New Ideas



Conceptually Simple, Yet Powerful

-Flow Switching

Switch Port	MAC src	MAC dst	Eth type	VLAN ID	IP Src	IP Dst	IP Prot	TCP sport	TCP dport	Action
port3	00:20..	00:1f..	0800	vlan1	1.2.3.4	5.6.7.8	4	17264	80	port6

VLAN Switching

Switch Port	MAC src	MAC dst	Eth type	VLAN ID	IP Src	IP Dst	IP Prot	TCP sport	TCP dport	Action
*	*	00:1f..	*	vlan1	*	*	*	*	*	port6, port7, port9

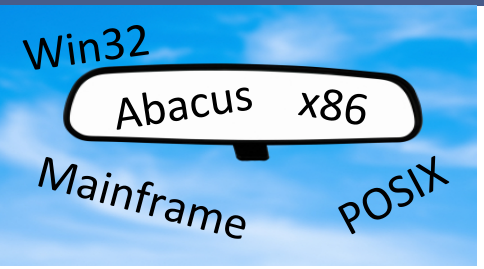
Routing

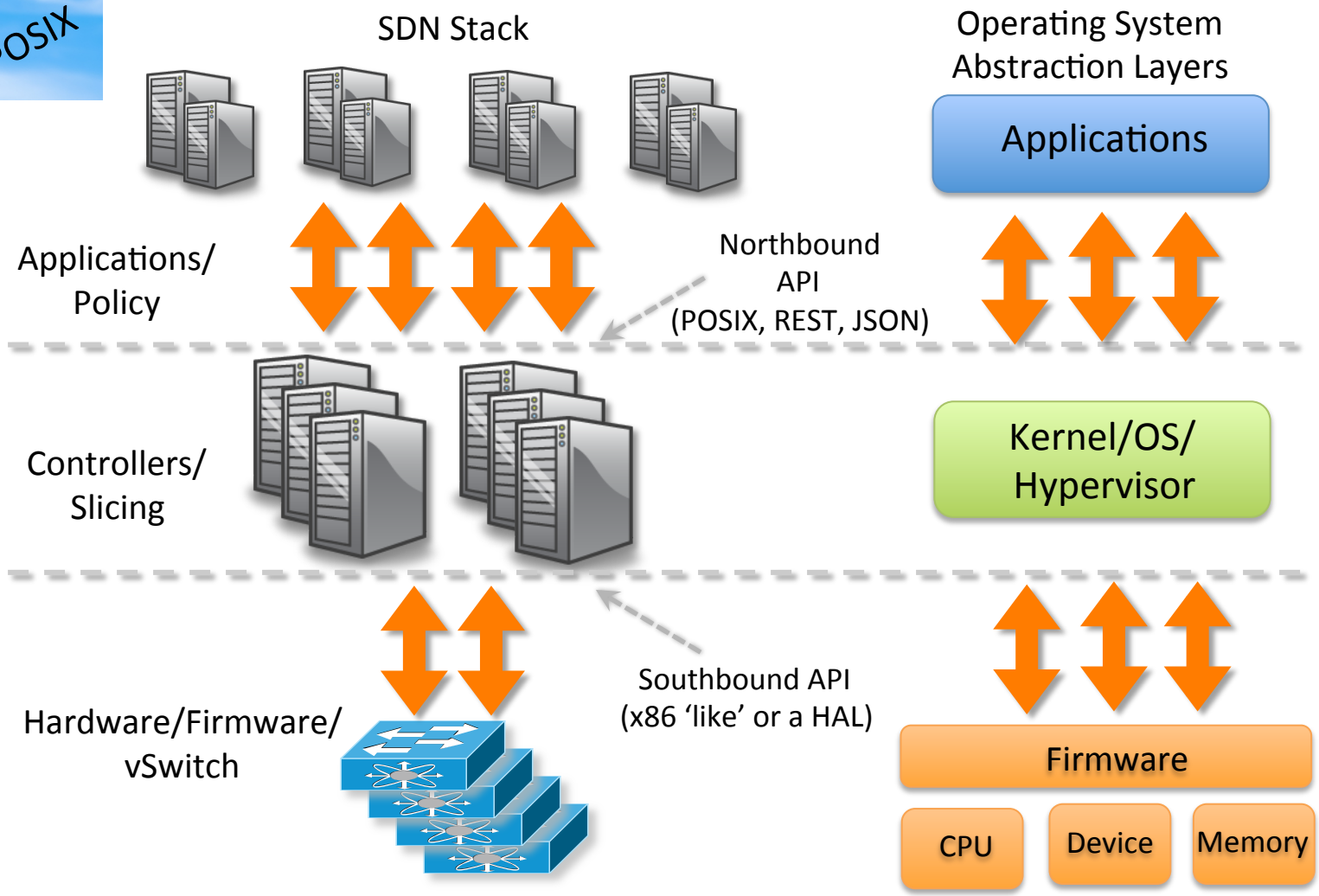
Switch Port	MAC src	MAC dst	Eth type	VLAN ID	IP Src	IP Dst	IP Prot	TCP sport	TCP dport	Action
*	*	*	*	*	*	5.6.7.8	*	*	*	port6

-Firewall

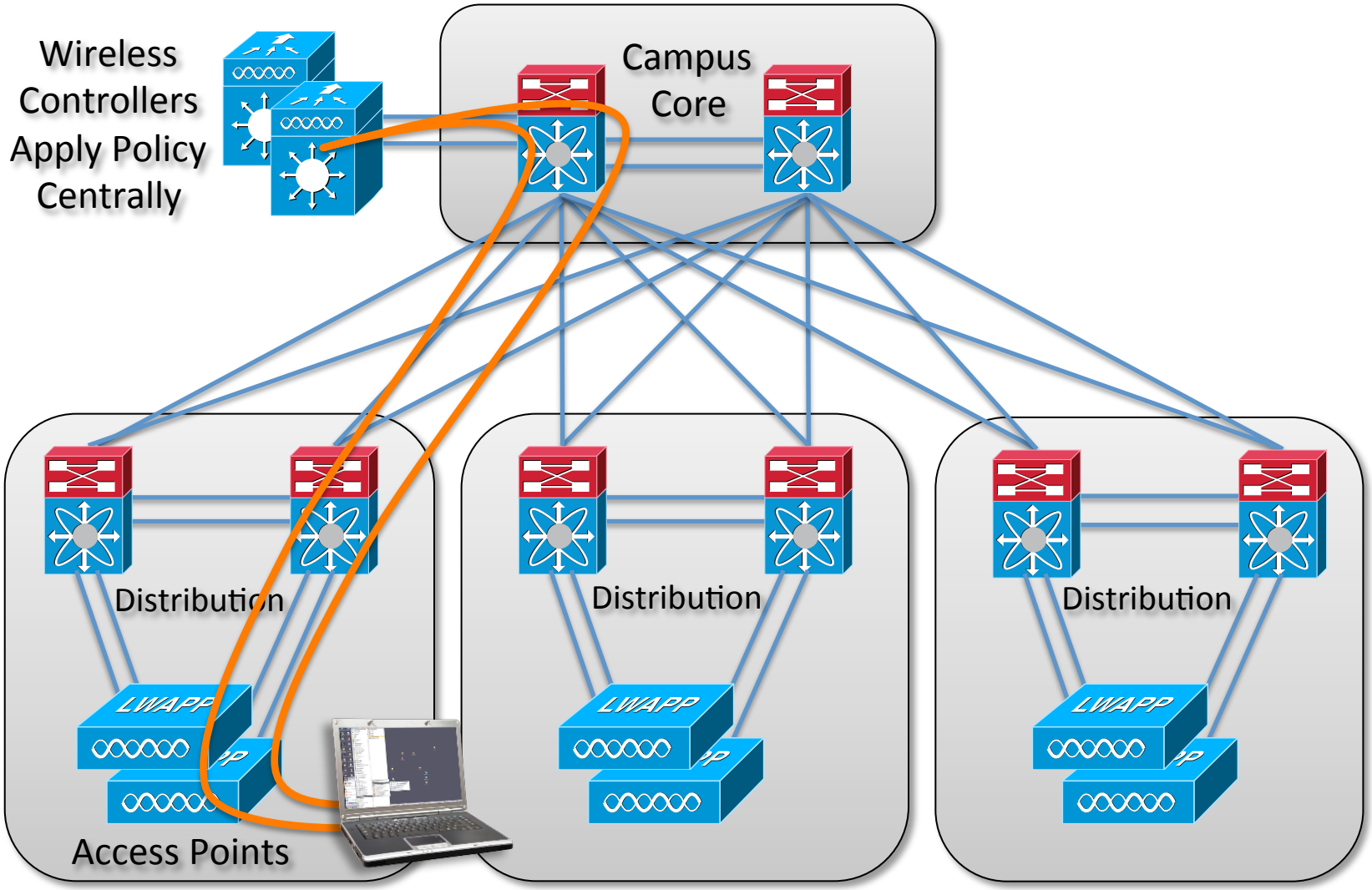
Switch Port	MAC src	MAC dst	Eth type	VLAN ID	IP Src	IP Dst	IP Prot	TCP sport	TCP dport	Action
*	*	*	*	*	*	*	*	*	22	drop

Abstraction

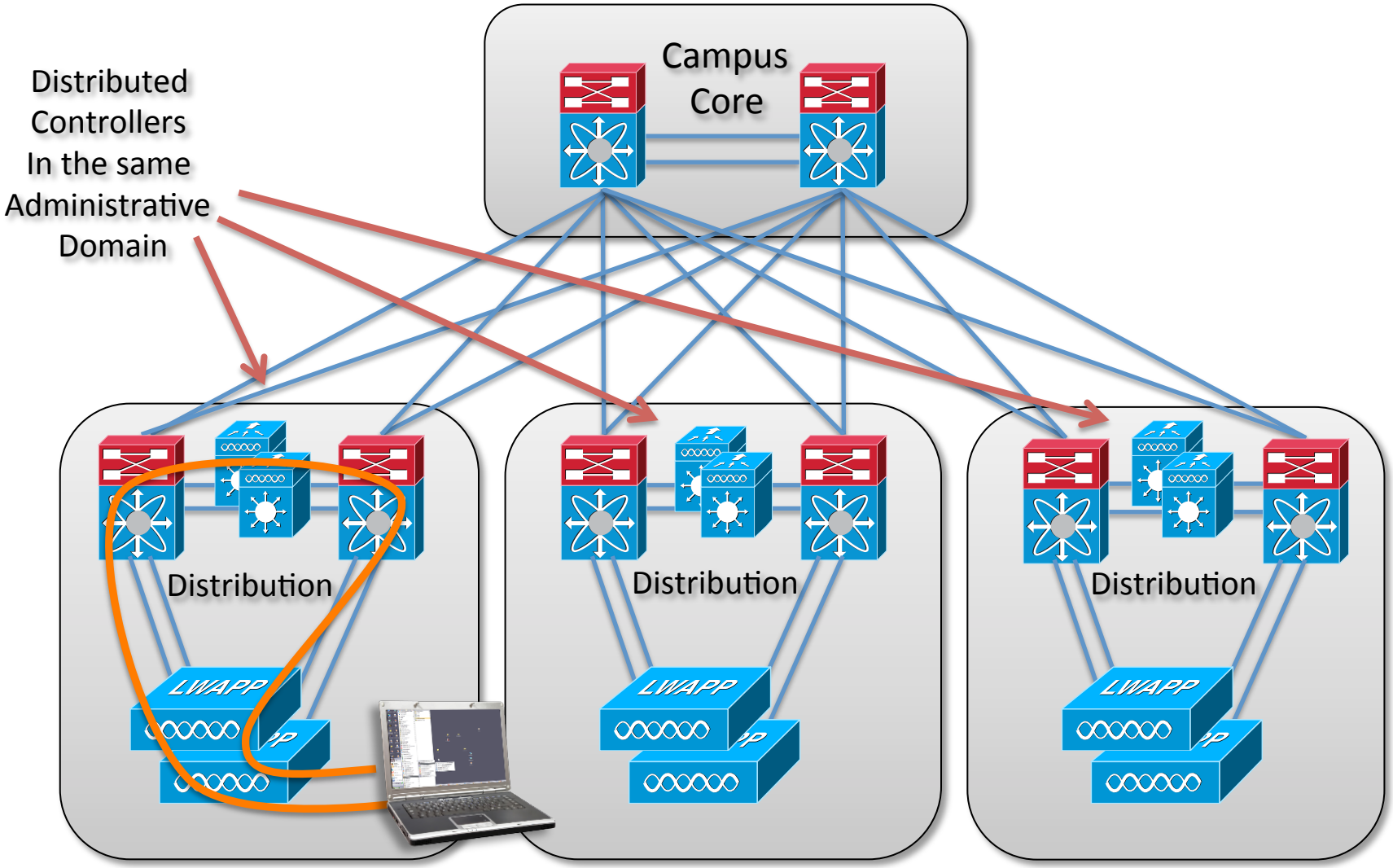




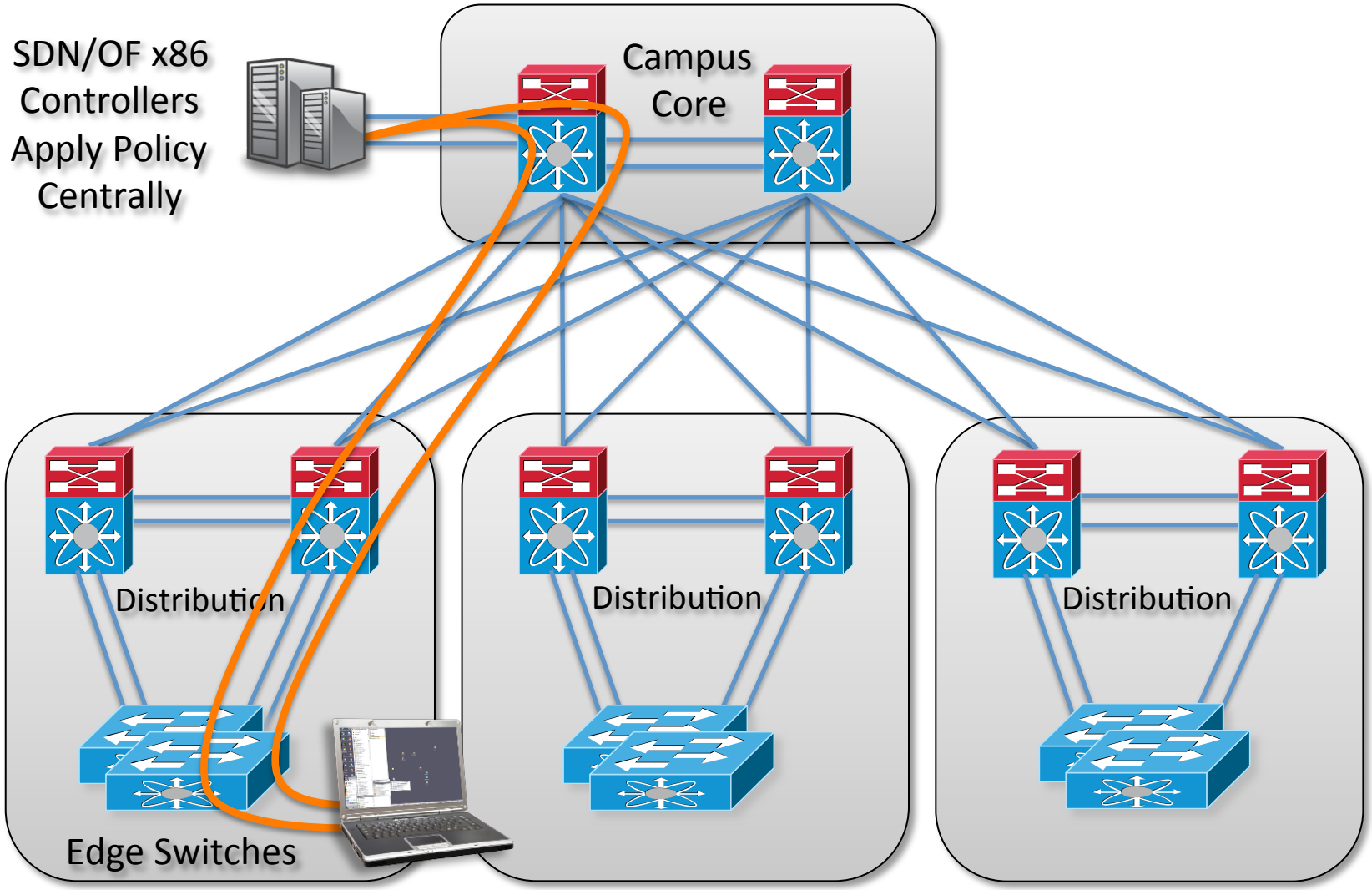
Enterprise Wireless Today



Enterprise Wireless at Larger Scale Today

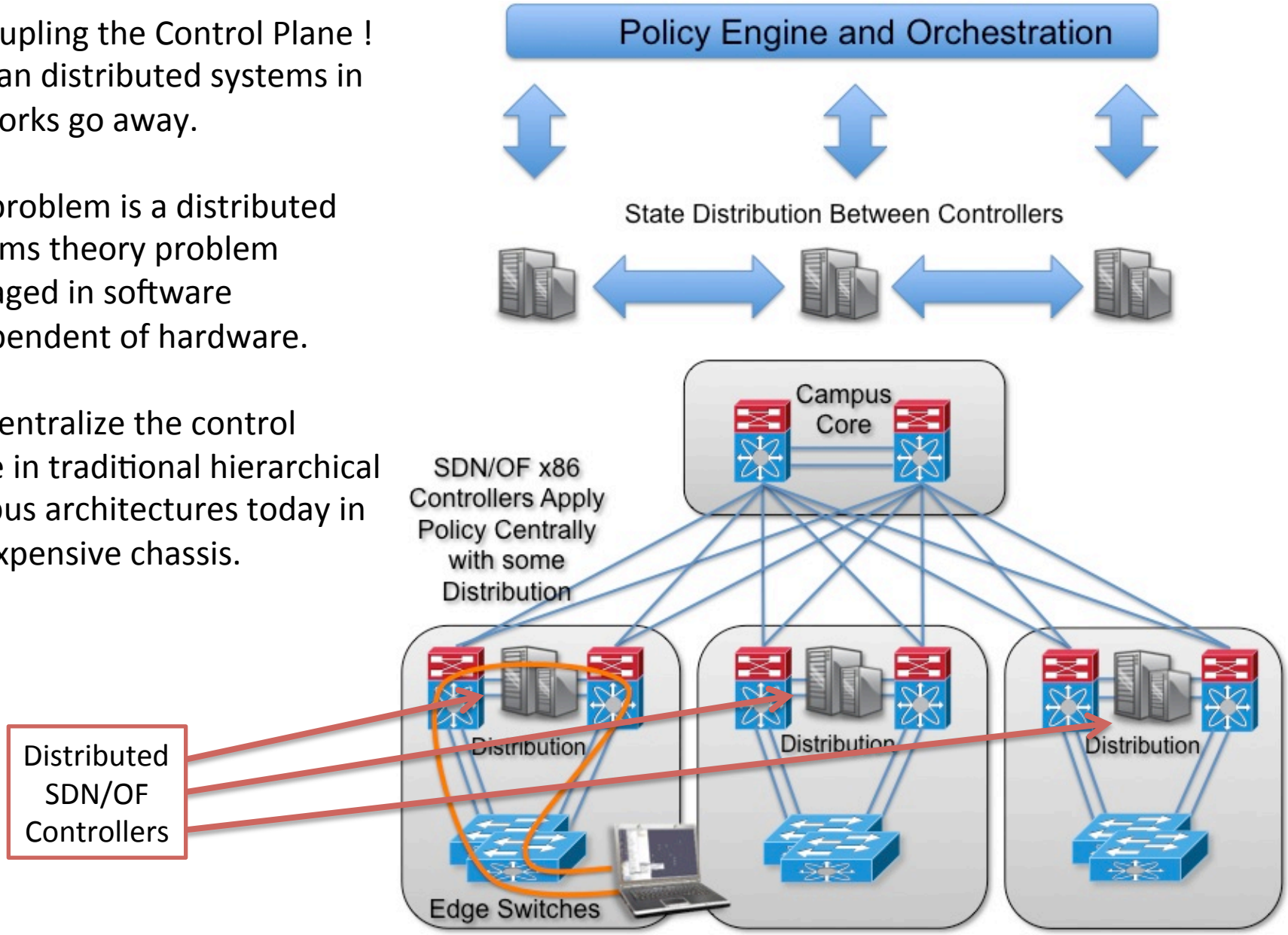


Decoupled Control Plane (NOS)



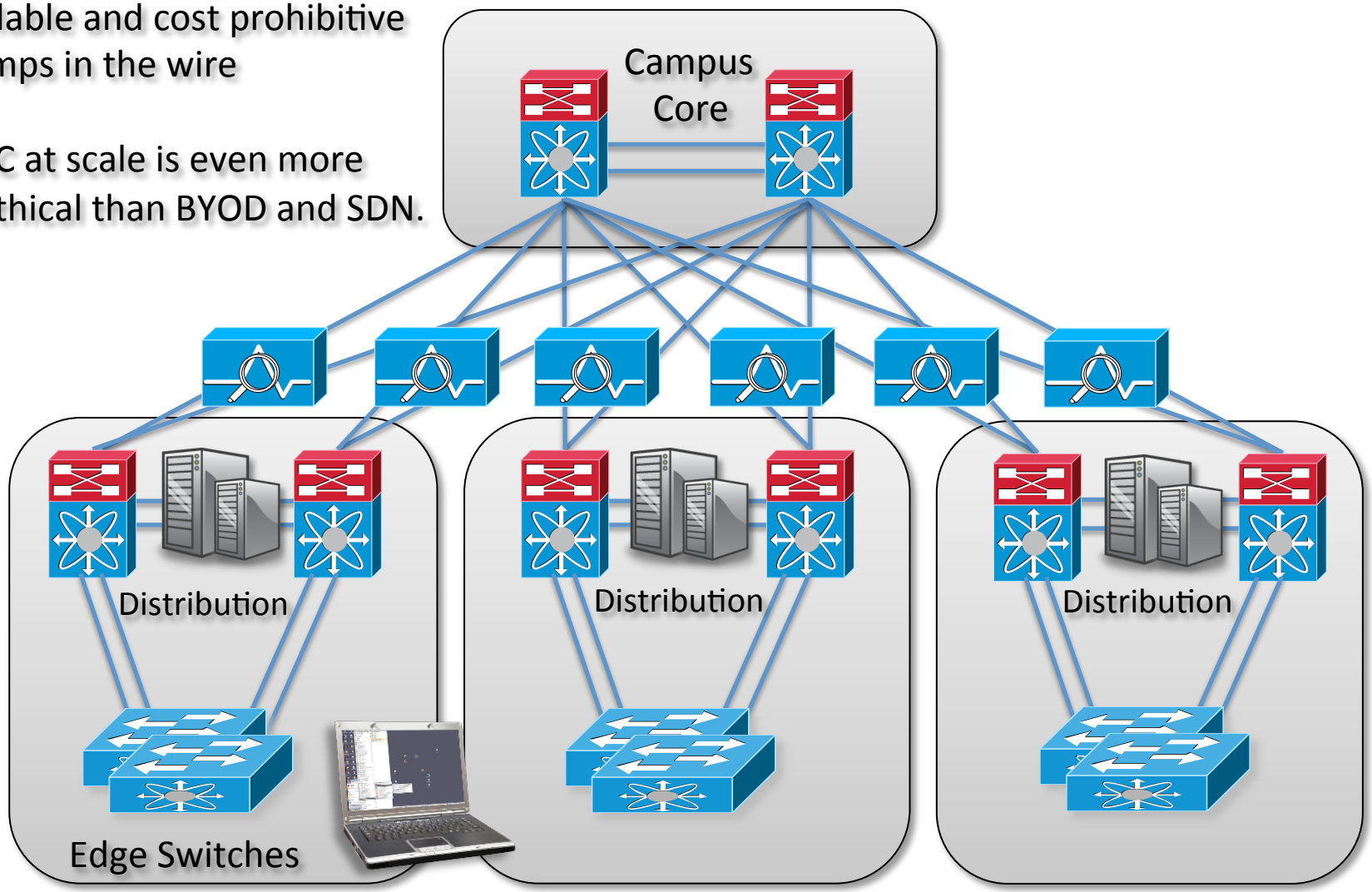
Policy Application in Wired Networks

- Decoupling the Control Plane ! = mean distributed systems in networks go away.
- The problem is a distributed systems theory problem managed in software independent of hardware.
- We centralize the control plane in traditional hierarchical campus architectures today in big expensive chassis.



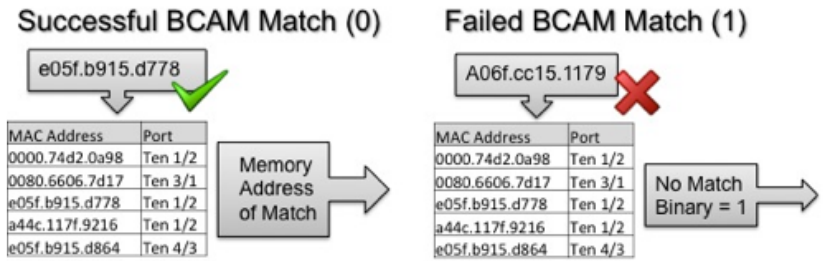
The Alternative is More of the Same

- The Alternative to apply policy is Business as usual. Un-scalable and cost prohibitive bumps in the wire
- NAC at scale is even more mythical than BYOD and SDN.



Open vSwitch – Scale HW vs. SW

- VM rack density East-West traffic could be problematic for general purpose top of rack.
- 100K+ entries in a rack is unrealistic in HW today.



Packet-in with match in TCAM – Action is forward to port 0/2

Port 0/3	*	*	*	*	192.168.1.1/32	*	*	*	*

TCAM Lookup

Ingress Port	Ethec Src	Ether Dst	Ether Type	Vlan ID	IP Dst	IP Src	TCP Dst	TCP Src	IP Proto
Port 0/1	*	*	*	*	*	*	80	*	*
*	*	*	*	*	192.168.1.0/20	*	*	*	*
*	*	*	*	*	192.168.1.0/24	*	25	*	*
Port 0/3	*	*	*	*	192.168.1.1/32	*	*	*	*

TCAM Lookup

0/5	*	*	*	*	172.24.16.5/32	*	80	*	*

Packet-in with NO match in TCAM – Action is Punt to Controller

Action Bucket

Send Packet to Port 0/2 In (n)RAM

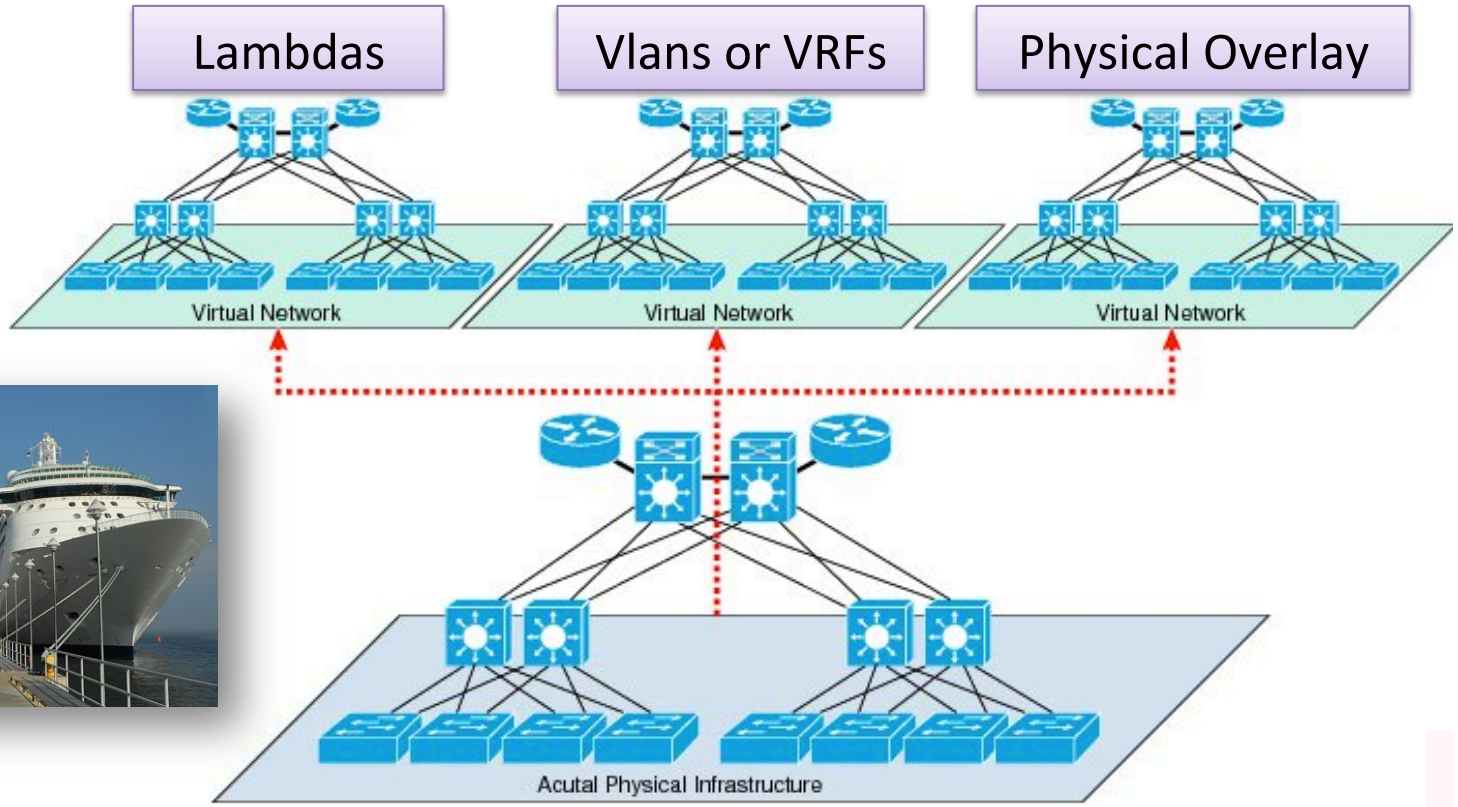
Action Bucket

Send Packet to Controller

Where Do You Start? (Physical Devices)

Ships in the night..

- Whatever your method of virtualization or an overlay network on hardware lifecycle. The idea is to create test bed and migrate as warranted.
- Virtualization today is done in segments or layers rather than a complete network solution.
- Most vendor early OpenFlow enabled code supports hybrid adoptions.



Ez Deployment Scenario

New Flow Processing -- struct ofp_packet_in (POX L2 LearningAlgorthym)

1. Update Source Address in (T)CAM or SW tables.

2. Is destination address a Ethertpe LLDP or Bridge Filtered MAC, or is?

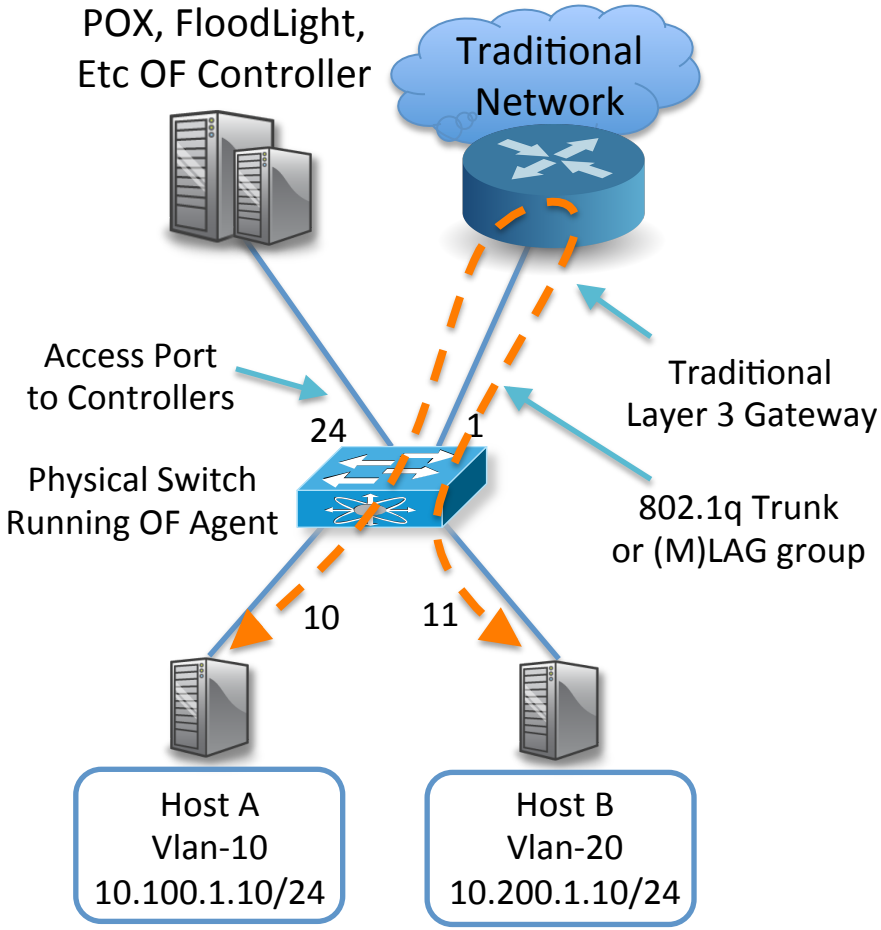
Drop or FWD to Controller or even hand off to STP. LLDP may be used to build a topology (important for future).

3. Is Multicast? Yes Flood.

4. Is the destination address in port mac address table. If no Flood.

5. Is output port the same as input port? Drop to prevent loops.

6. Install flow and forward buffered and subsequent packets.



What Changed? Why Now? #2 The Data Center

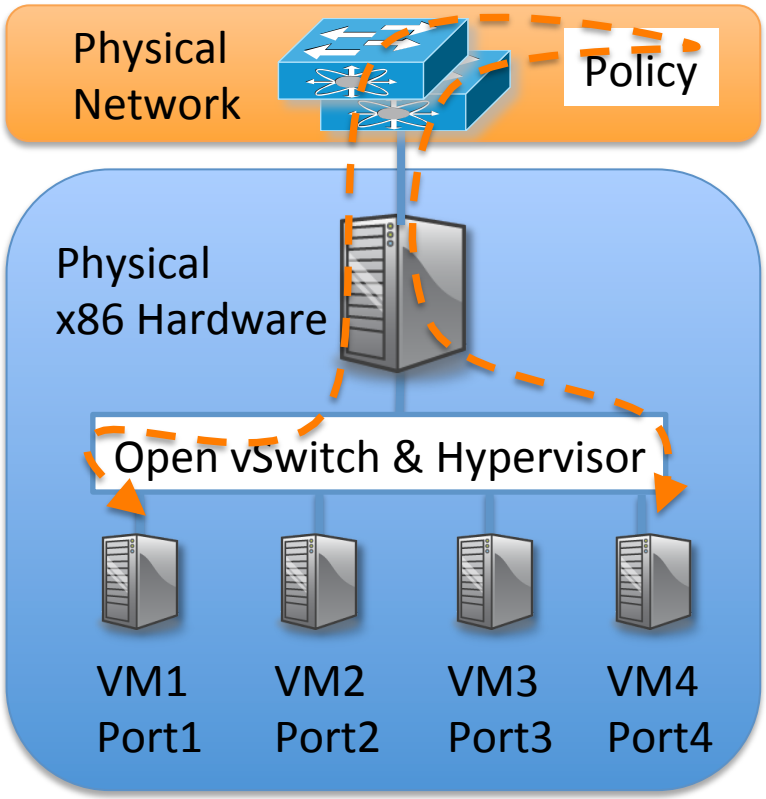
- "The network is in my way" *-James Hamilton, Amazon*
- Networking is complex because the appropriate abstractions have not yet been defined." *–A Case for Expanding OpenFlow/SDN Deployments On University Campuses*
- "If you look at the way things are done today, it makes it impossible to build an efficient cloud. If you think about the physical network because of things like VLAN placements, you are limited on where you can place workloads. So even without thinking about the application at all, there are limits on where you can place a VM because of capacity issues or because of VLAN placement issues." – Martin Casado

- The tools we have today for automation: snmp, netconf, subprocess.Popen(Python), Net::Telnet(Perl),#!/bin/bash, autoexpect, etc.

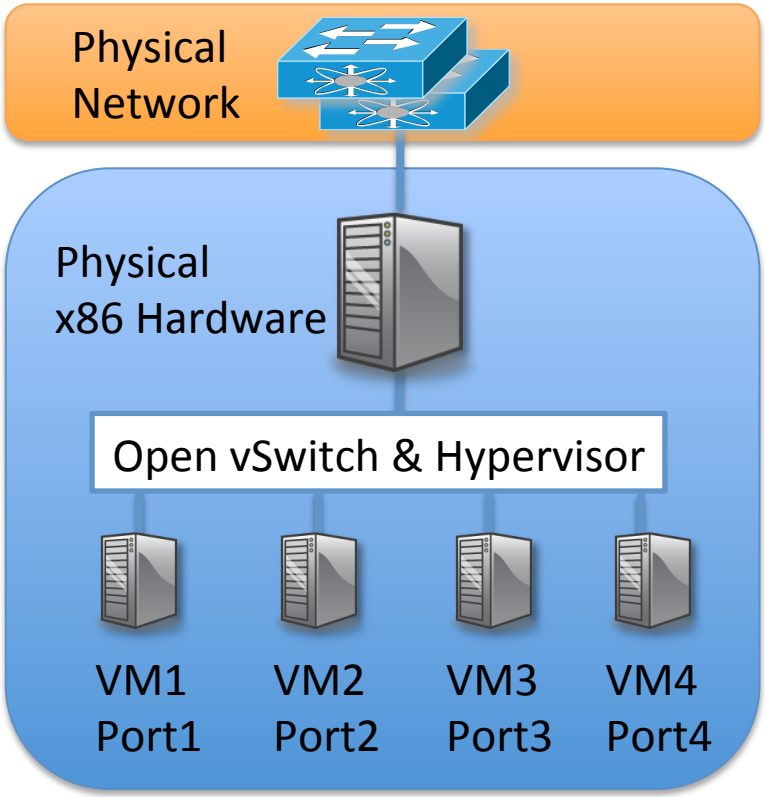
What Changed? #2 The Data Center

- Public Cloud Scale
- VID Limitations - ~4094 Tags
- ¼ of Servers are Virtualized
- Customers want flat networks but they do not scale.
- Complexity in the network substrate to support bad application design.
- Required- Flexible & Open APIs to consume Network Resources.
- East-West policy application.
- East-West BW Consumption.
- L2 Multi-Tenancy.
- Hypervisor Agnostic.
- VM port characteristic mobility.
- Traffic Trombone for Policy.

The Edge Needs to Be Smarter but also manageable: Below is neither



Virtual Switching (Open vSwitch)

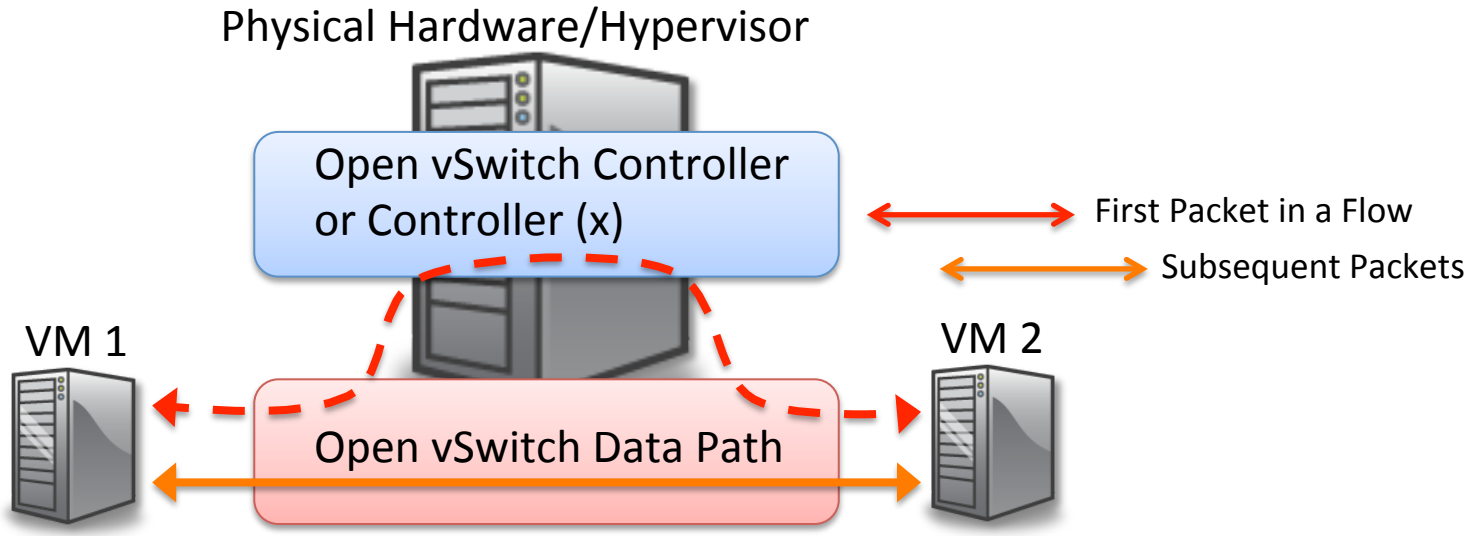


% ovs-appctl fdb/show br0

port	VLAN	MAC	Age
0	0	00:0f:cc:e3:0d:d8	6
1	0	00:50:56:25:21:68	1
2	0	10:40:f3:94:e0:82	1
3	10	10:40:f3:94:e0:82	1
4	10	00:0f:cc:e3:0d:d8	1

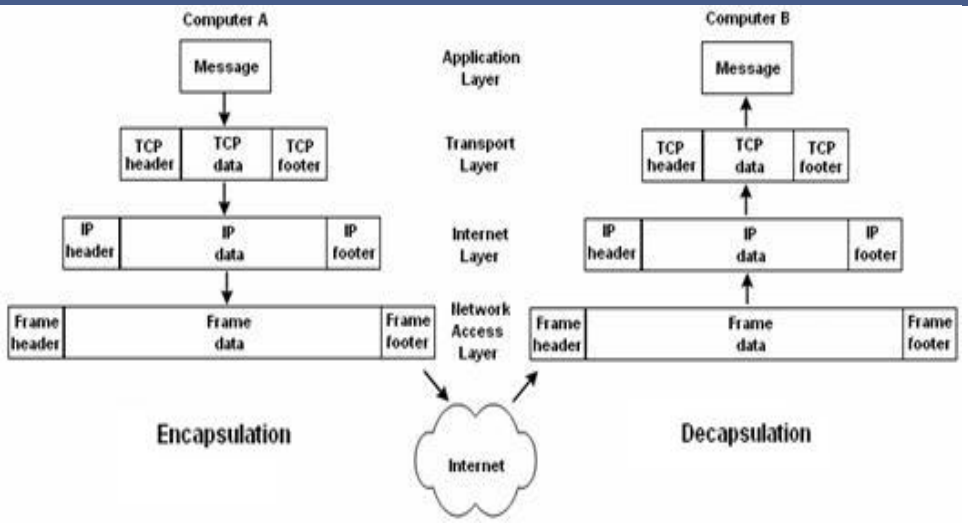
- **Security-**Vlan Layer2 isolation, Traffic filtering
- **QOS-**Traffic queuing and shaping
- **Monitoring-** Netflow, sFlow, SPAN, RSPAN
- **Control**- OpenFlow or NextGen
- **Features**: Bonding, GRE, VXLan, Capwap, STT, Ipsec.
- **Hypervisor Support:** KVM, XEN, Xenserv, Vbox.
- **Orchestrator Support:** OpenStack, CloudStack.
- **License**: Apache2 and GPL (upstream kernel mod)

Open vSwitch Forwarding



- First Packet in the flow goes to the OVS controller (slowpath) and subsequent are forwarded by the OVS data path (fastpath)
- Underlying Open vSwitch is a flow-table forwarding model similar to that used by OpenFlow.
- When attached to a controller datapaths are determined by the OpenFlow Controller.
- Multiple tenants can share the same tunnel.
- Actions: Forward, Drop, encapsulate and send to controller.

Refresh of Encapsulation



Layer 5+ | Application Header | YouTube

Layer 4 | TCP Header | Application Header | YouTube

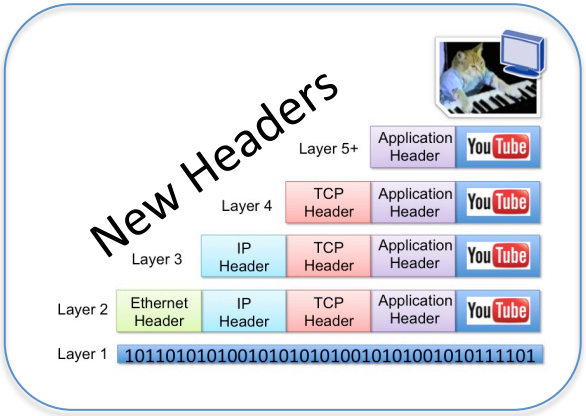
Layer 3 | IP Header | TCP Header | Application Header | YouTube

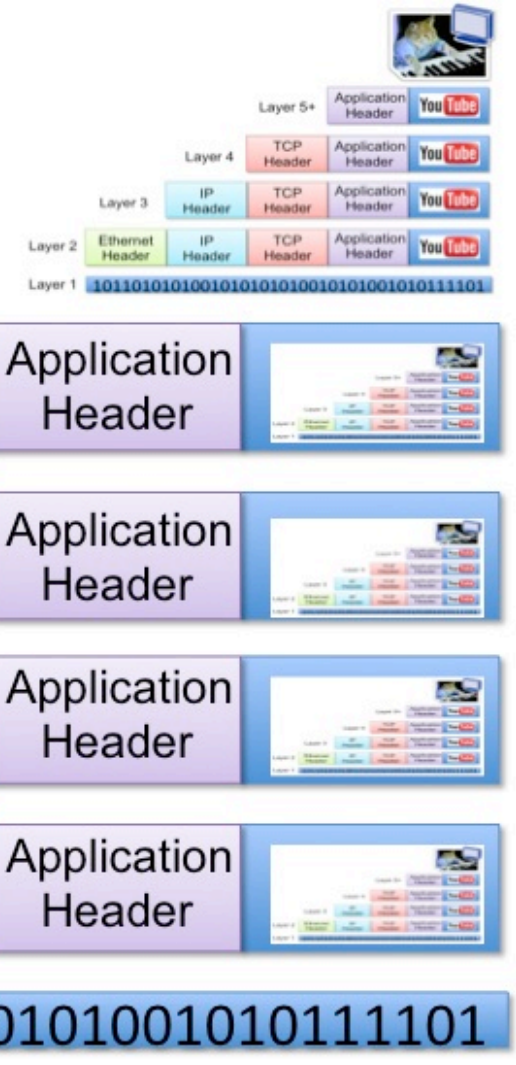
Layer 2 | Ethernet Header | IP Header | TCP Header | Application Header | YouTube

Layer 1 | 10110101010010101010100101010010101111101

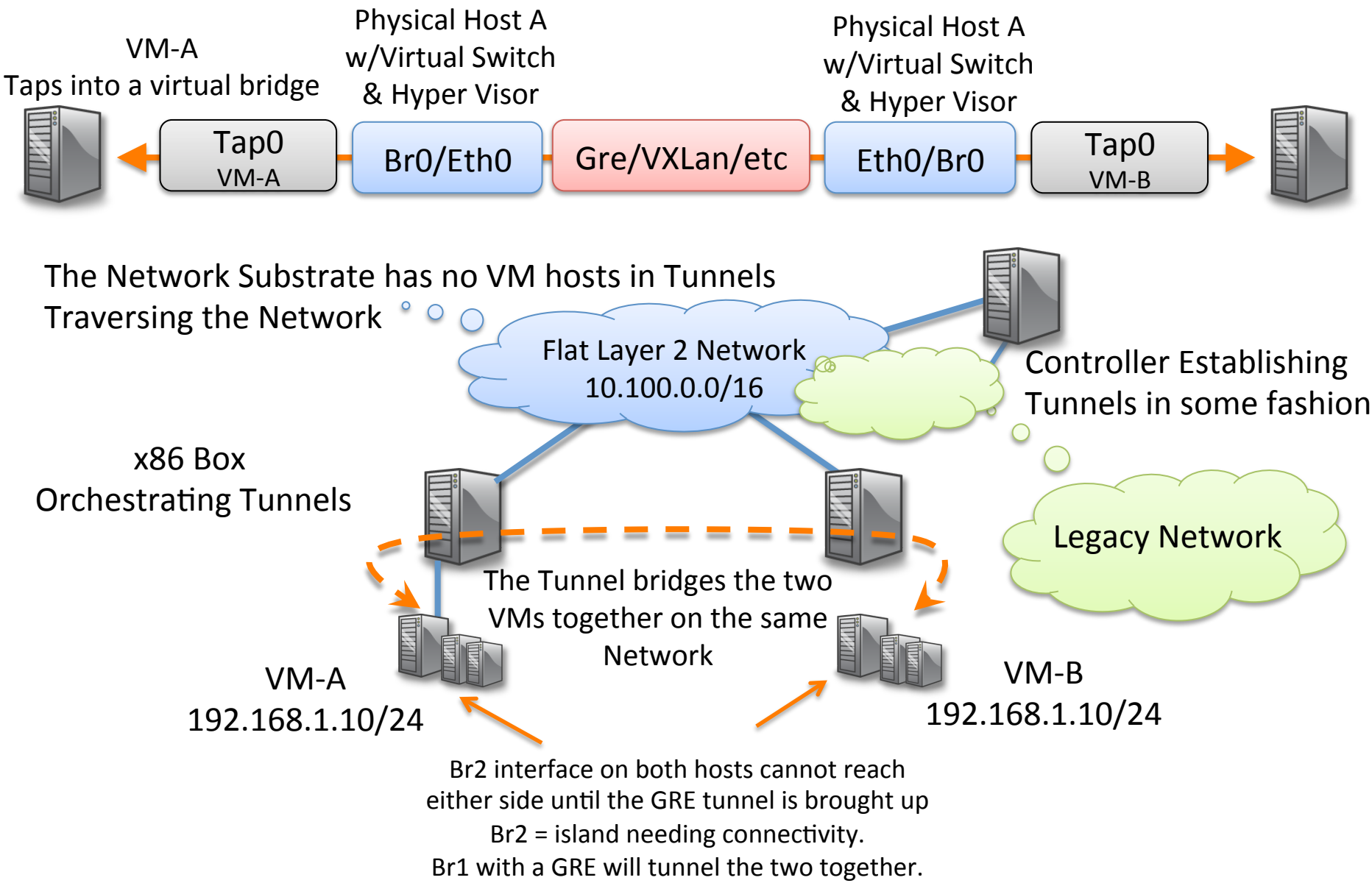
New Encapsulation to Traverse Physical Net

Encapsulated In New Headers the Original Packet + Headers is now Payload.



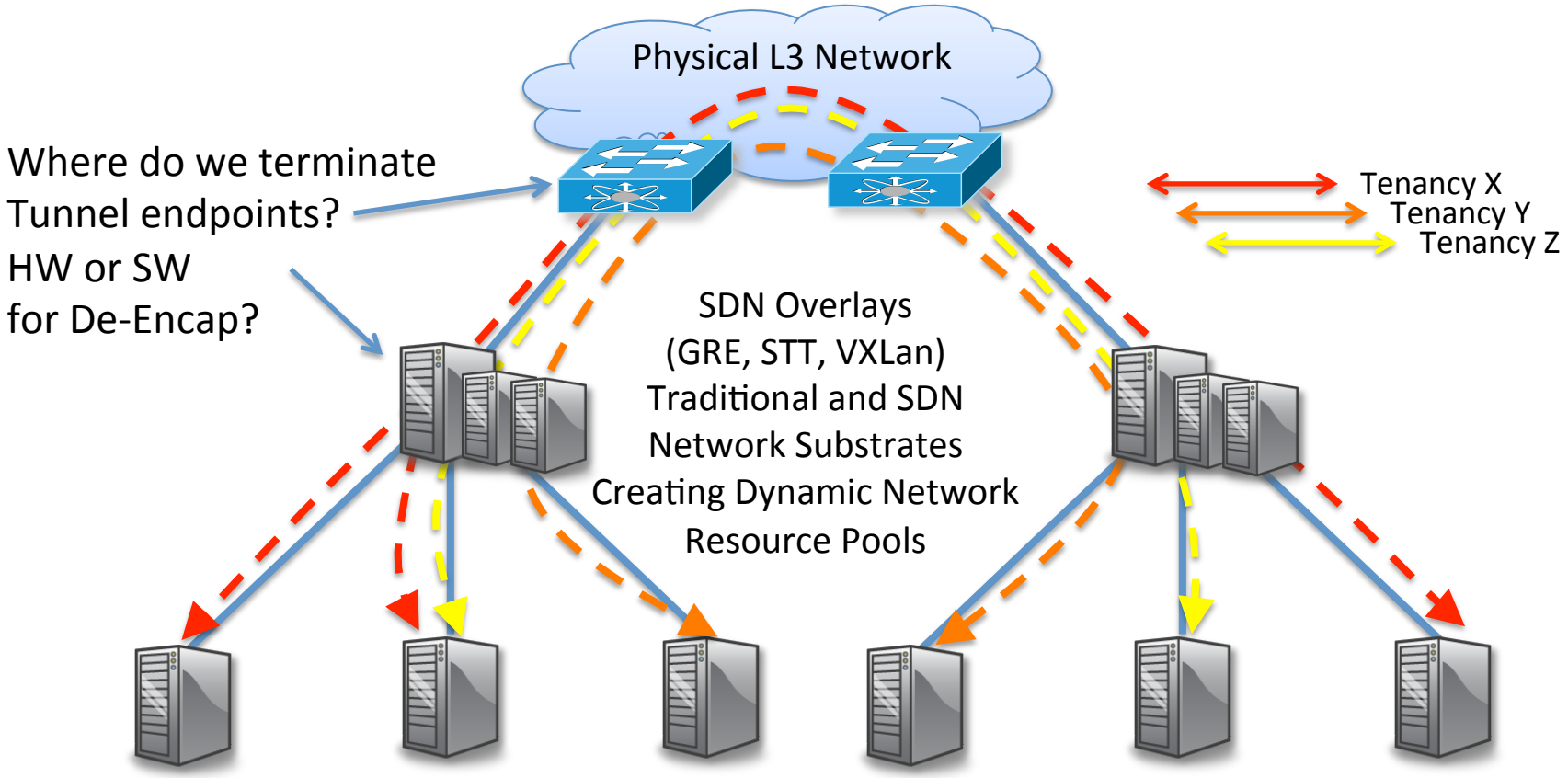


Tunneling/Overlays



Data Center Overlays

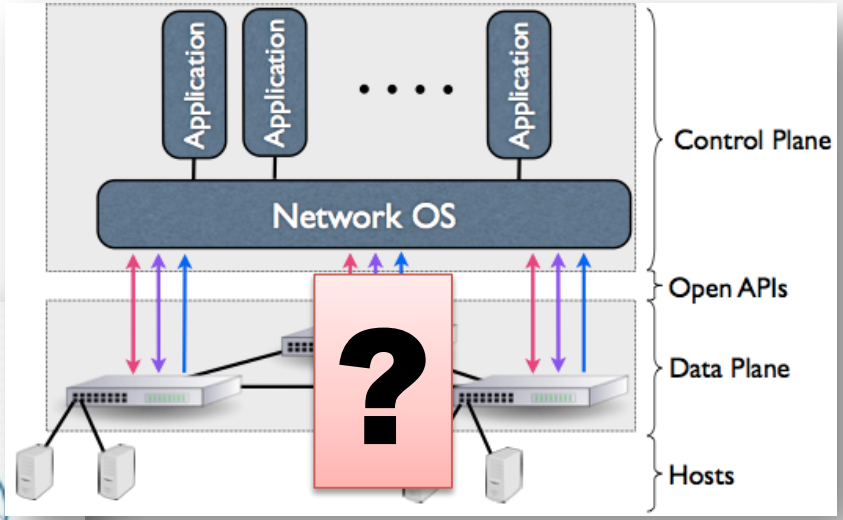
- Early SDN adoptions are happening today in Data Centers. decouple the virtual from the physical discreetly.



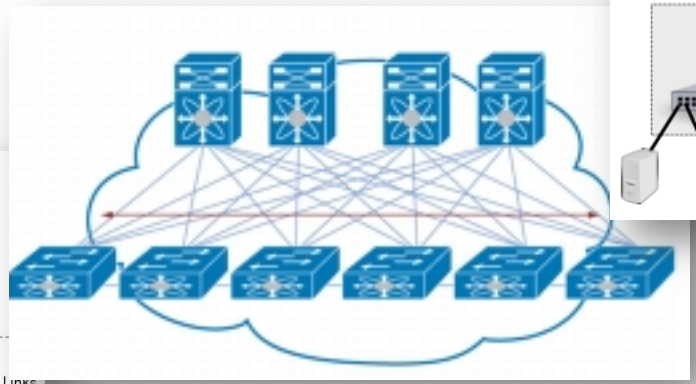
Resources Consumption (Storage, Network, Compute) Either Local or Hybrid Private/ Public Cloud. Visibility, OAM, Dynamic Provisioning, Brokerage and Analytics.

Evolution or Re-Invention?

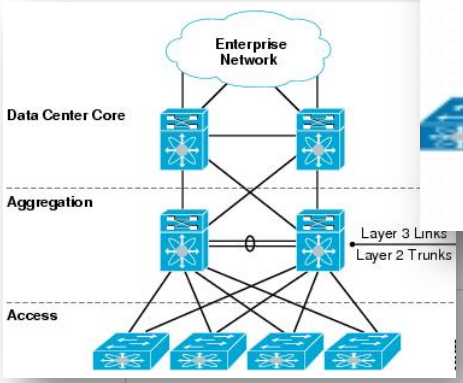
Software Defined

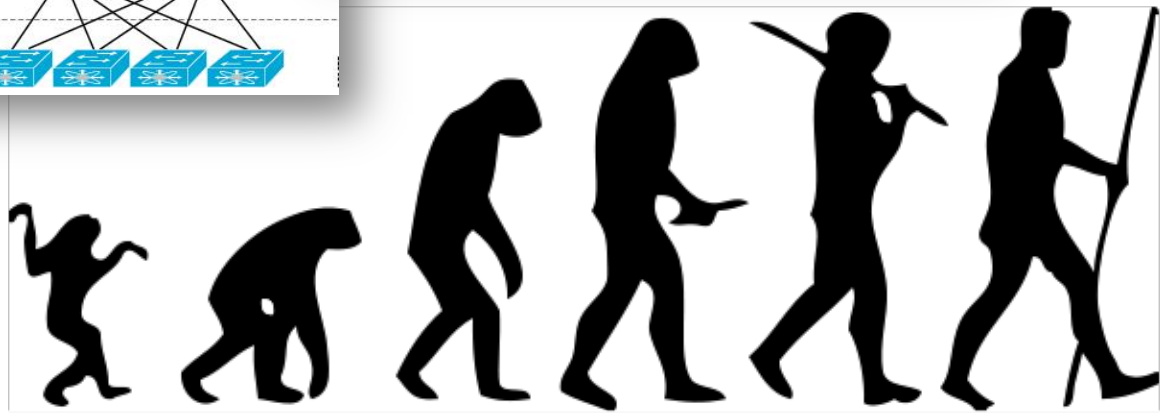


2-Tier Flat TRILL/SPB/MPLS
North-South 75/25

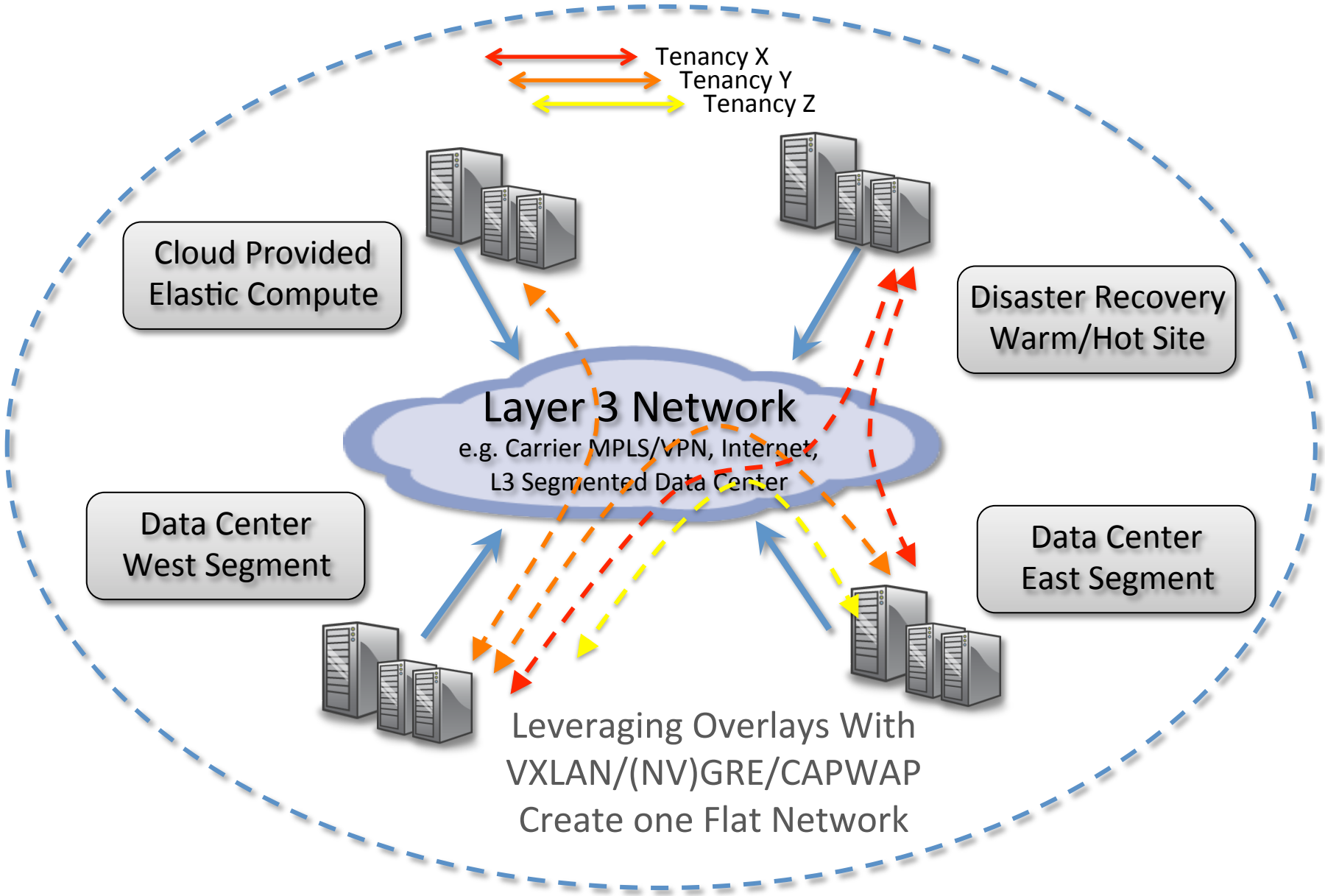


3-Tier
North-South 90/10

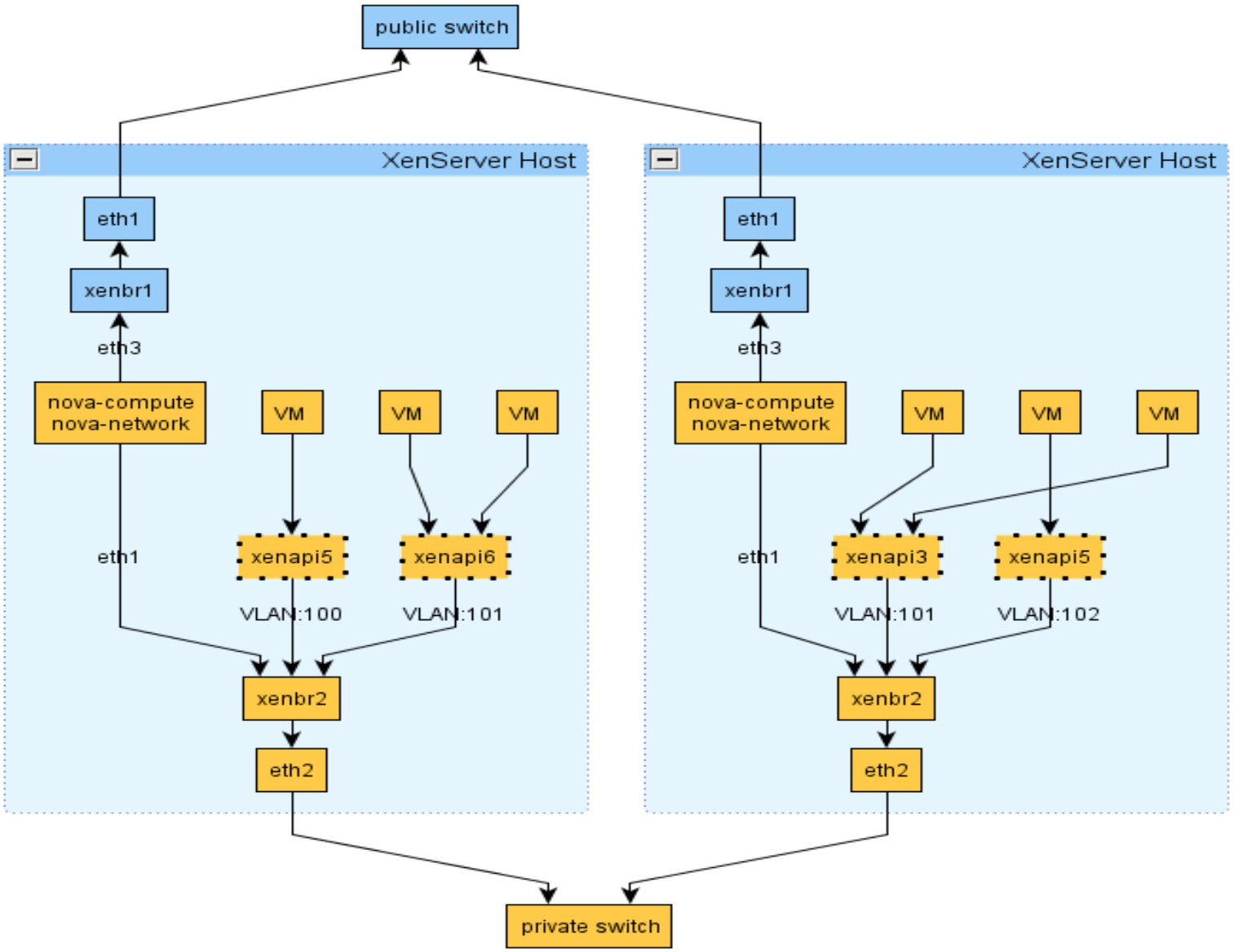




Does This Make Sense?

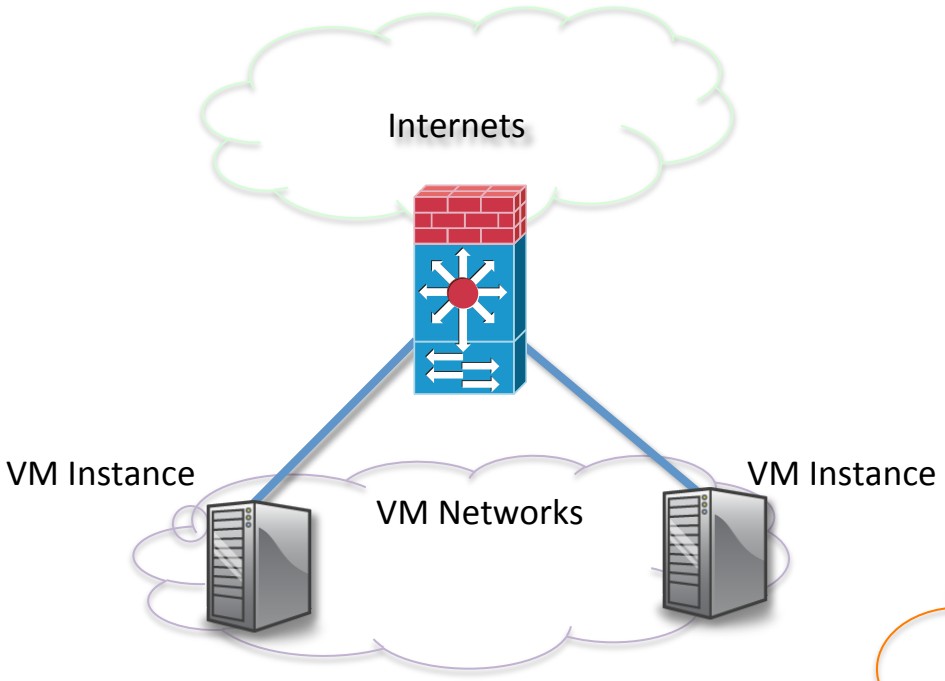


OpenStack Vlan Networking

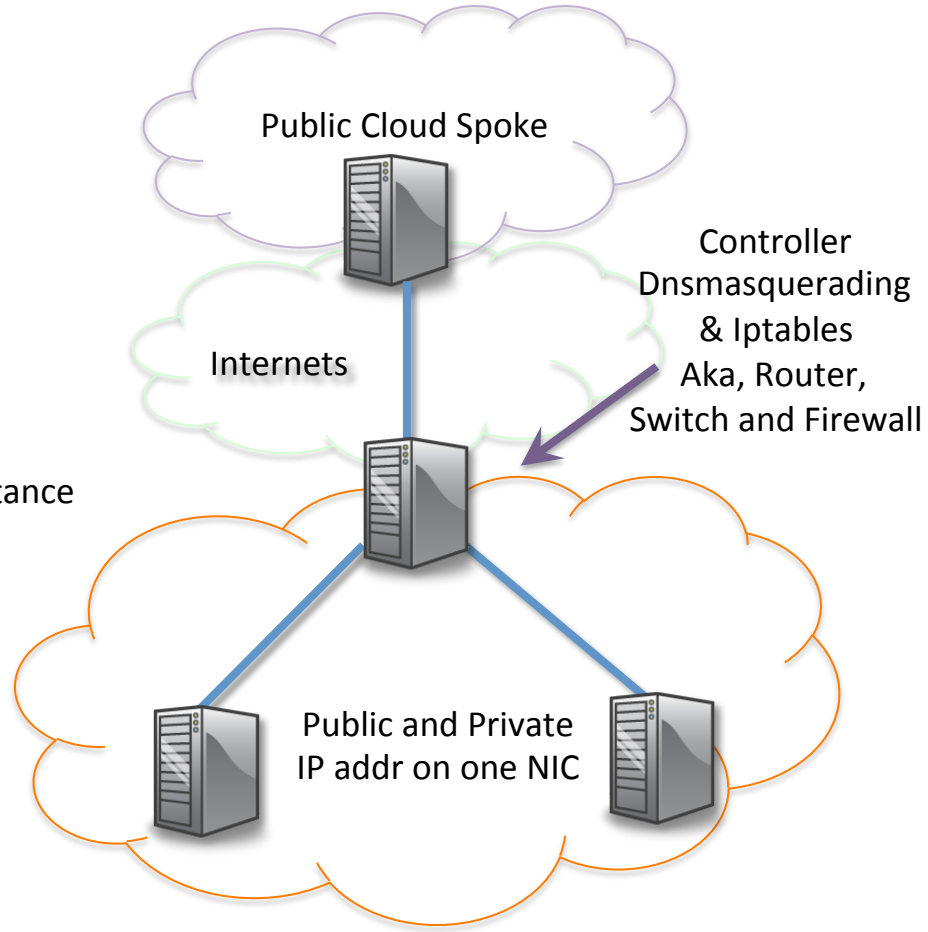


Hybrid Cloud Look

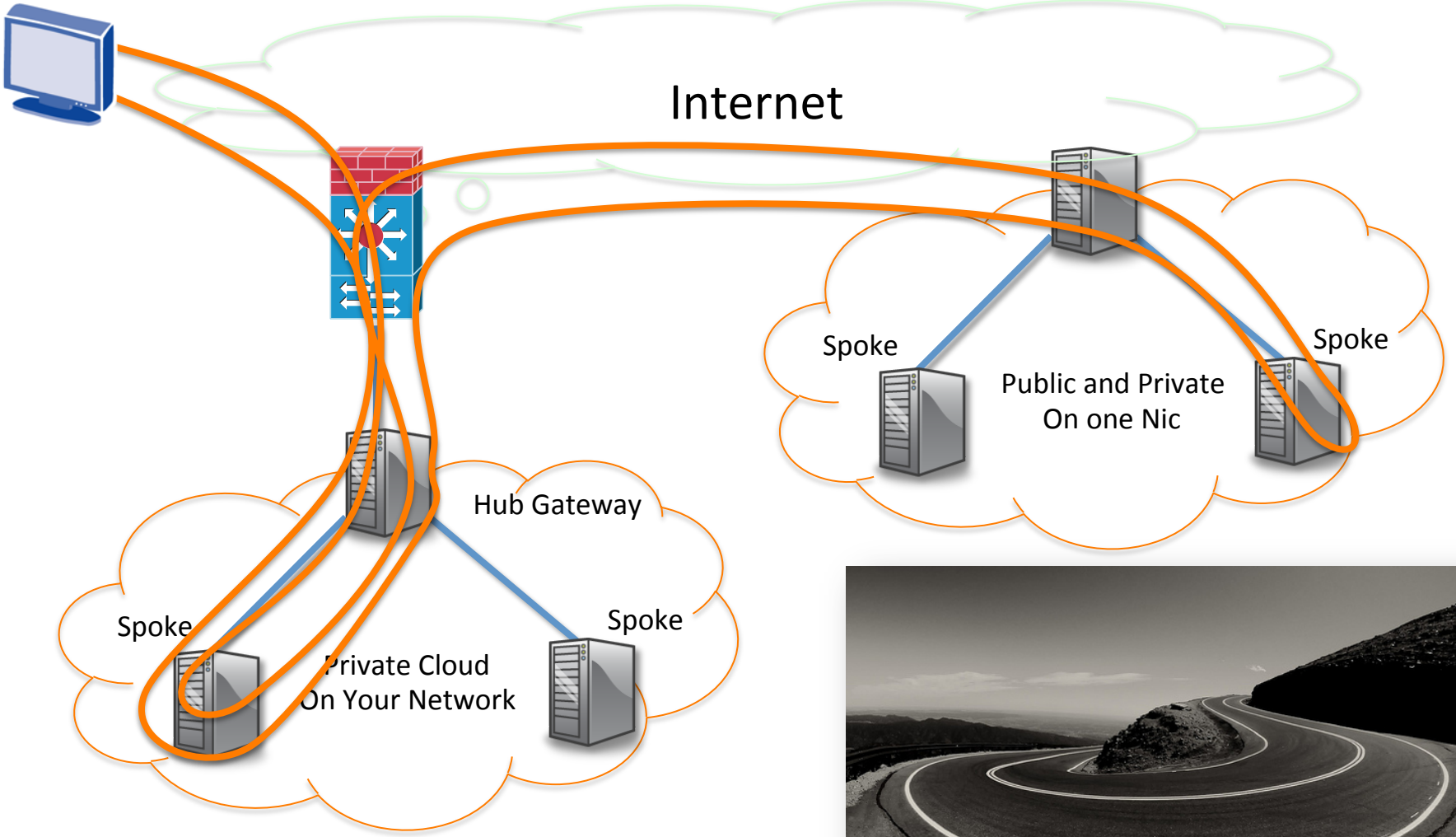
How Public Cloud Feels



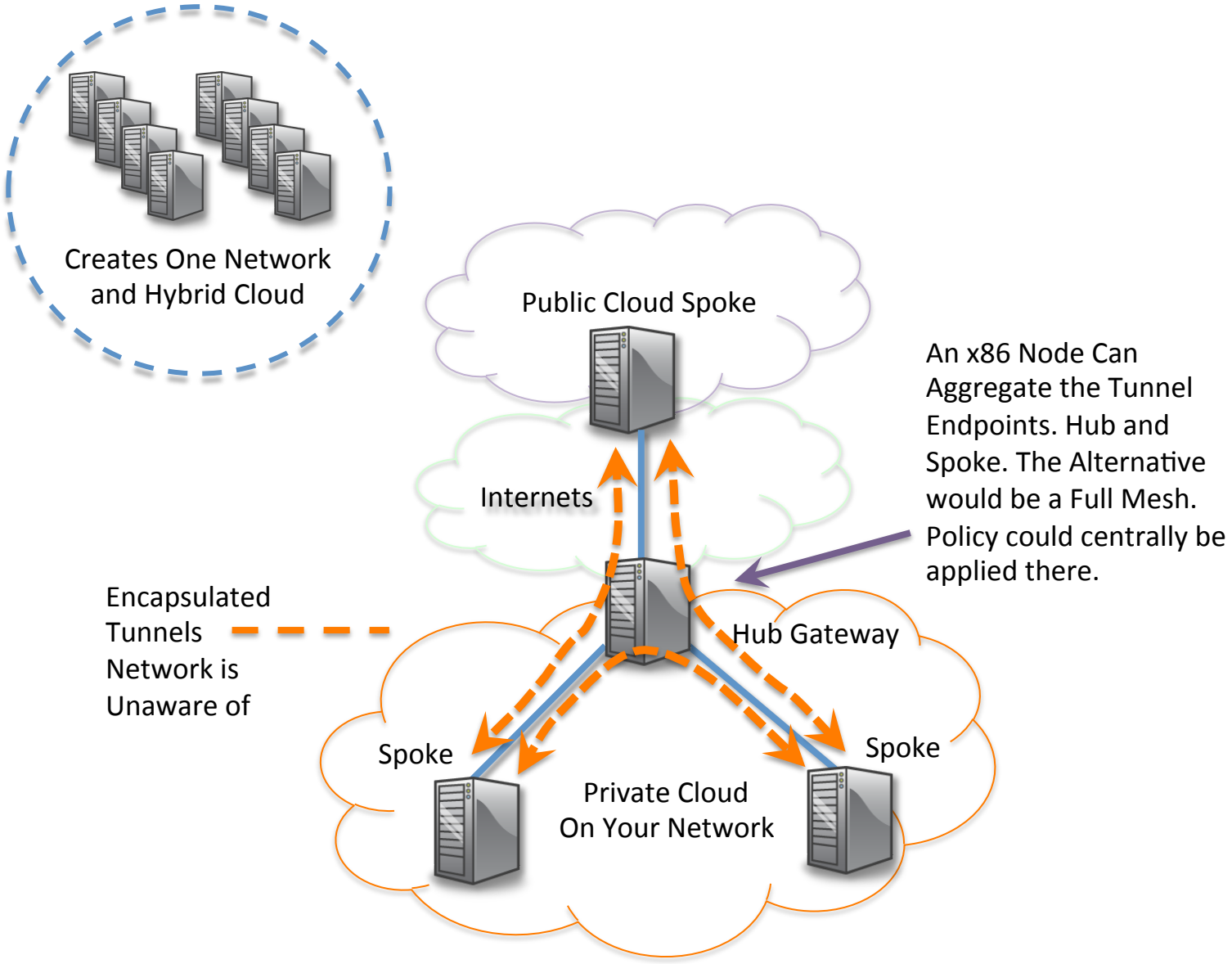
How it Really is:



Hybrid Cloud - IMO Not as Bad as It Looks, this exists today in most DCs

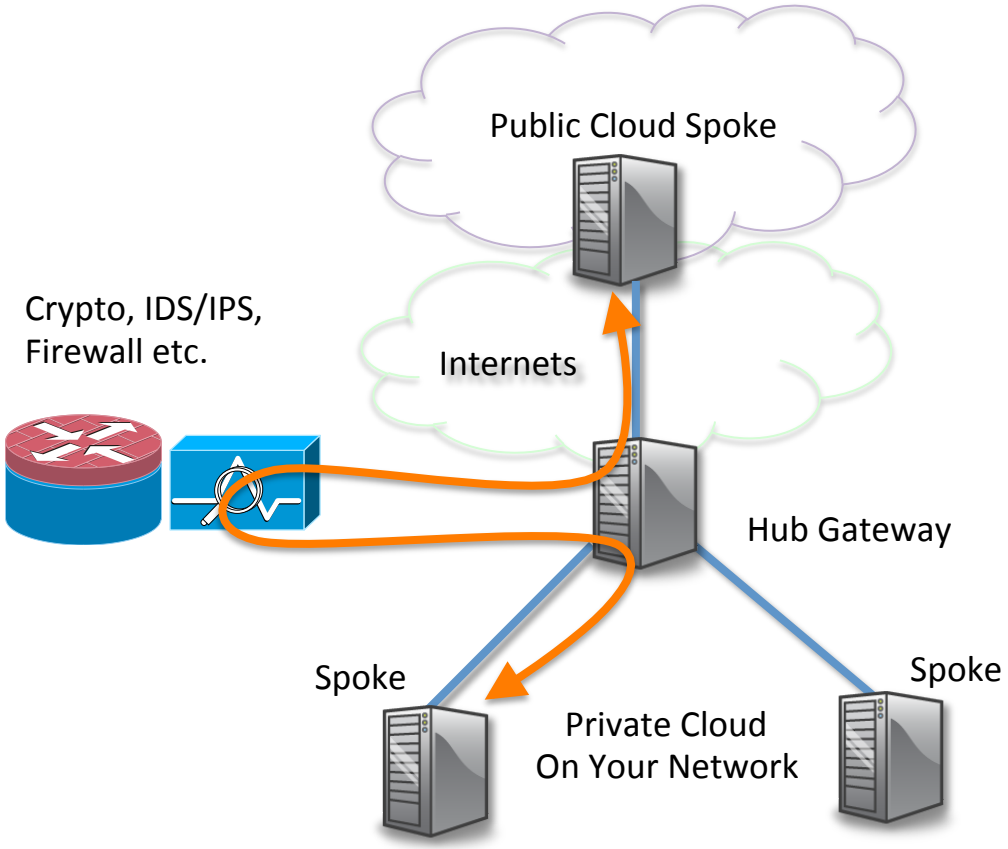


Tunneling & Hybrid Cloud



De-Dup Policy is the best Reason for Tunnels

- Leverage existing centralized policy Application and Orchestration.
- That all said sending the client directly to a cloud provider outside of a tunnel via the Internet is by far the easiest and most scalable solution.



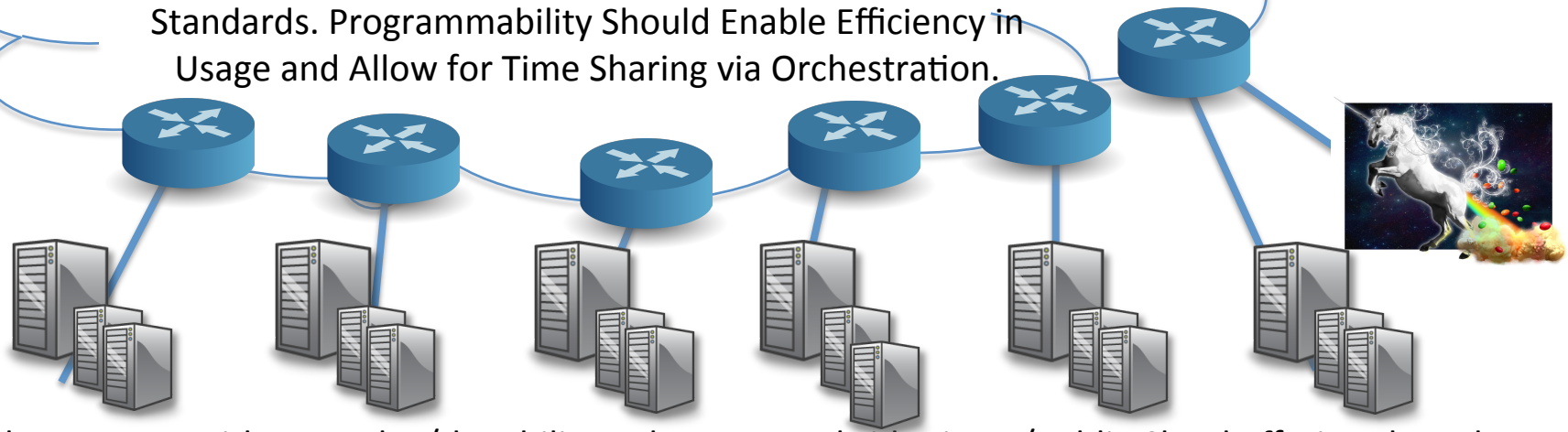
Public Cloud: The Internet will be the new LAN

Option 1:General Internet1 best effort connectivity through commodity I1 drains like a Cogent for example at ~ ¢25-¢50 cents per/mb. Capture that as service level as a lower tier SLA but priced significantly lower. Primary option.

Option 2: Dedicated peerings to any node from tenant to colo into the super-regional anyone selling resource pools with open APIs. Rackspace, HP, Dell, Piston Cloud. Companies whose end game is not 100% cloud.

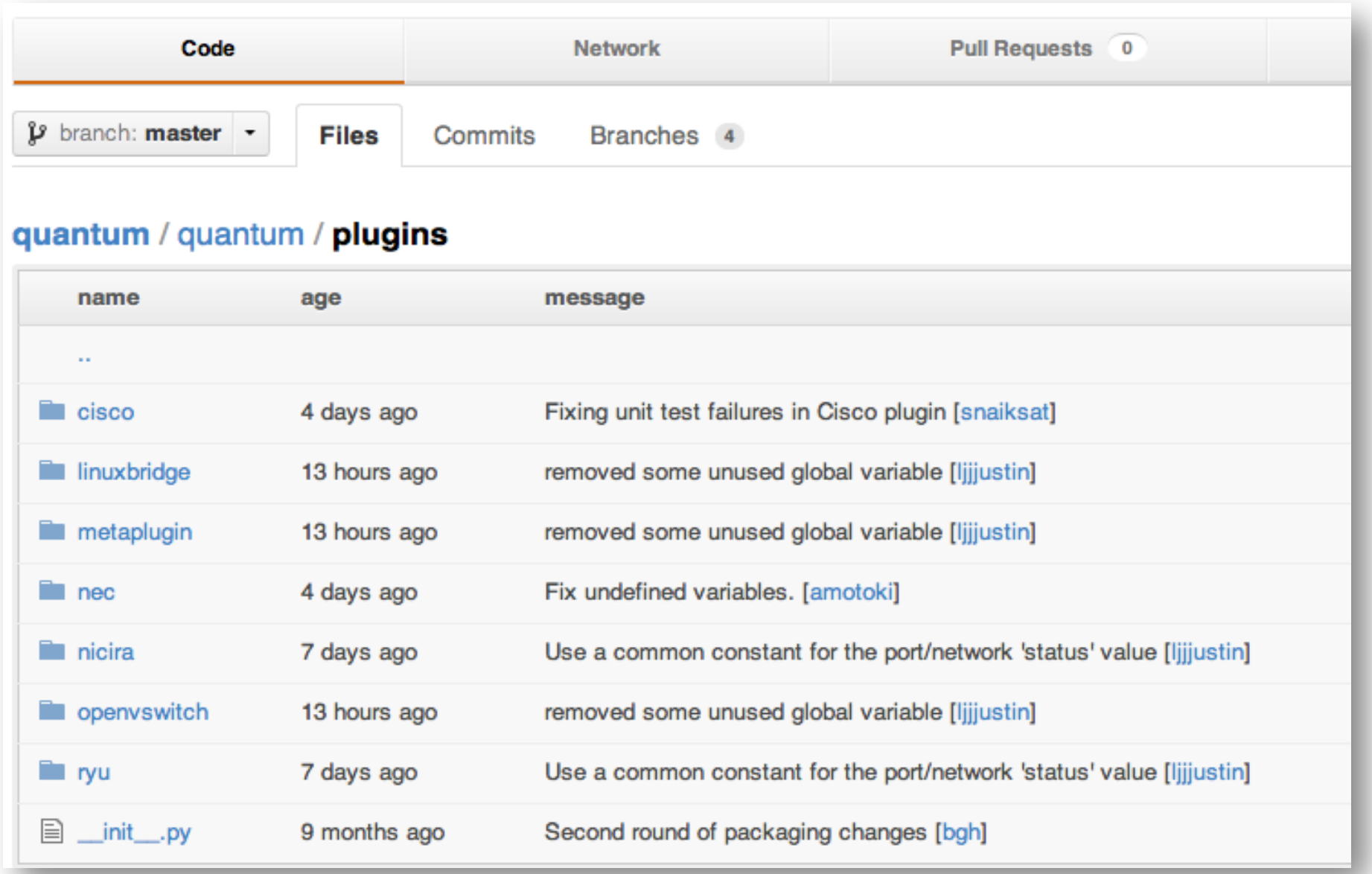
Option 3: Internet2, Ideally begin leveraging their peering and Colos globally for a broader net to capture competitively priced resources

Leverage Regional & Super-Regional Statewide Networks and Open Peerings to Cloud Providers. xAAS driven as a commodities market through Emerging Open API Standards. Programmability Should Enable Efficiency in Usage and Allow for Time Sharing via Orchestration.



OpenStack Resources Either Local w/the ability to leverage Hybrid Private/Public Cloud offerings based on the best market price that year, month maybe even day depending on the elasticity and flexibility to move workloads. Also balancing workloads amongst each other through scheduling and predictive analysis and magic. Tenants would be any community anchor, state, city, education non-profit etc.

Quantum: The OpenStack Network Plugin



Code | Network | Pull Requests 0

branch: master | Files | Commits | Branches 4

quantum / quantum / plugins

name	age	message
..		
cisco	4 days ago	Fixing unit test failures in Cisco plugin [snaiksat]
linuxbridge	13 hours ago	removed some unused global variable [ljjjustin]
metaplugin	13 hours ago	removed some unused global variable [ljjjustin]
nec	4 days ago	Fix undefined variables. [amotoki]
nicira	7 days ago	Use a common constant for the port/network 'status' value [ljjjustin]
openvswitch	13 hours ago	removed some unused global variable [ljjjustin]
ryu	7 days ago	Use a common constant for the port/network 'status' value [ljjjustin]
__init__.py	9 months ago	Second round of packaging changes [bgh]

Cisco Nexus Plugin Snippet

```
# @author: Edgar Magana, Cisco Systems, Inc.

"""
Nexus-OS XML-based configuration snippets
"""

import logging


LOG = logging.getLogger(__name__)


# The following are standard strings, messages used to communicate with Nexus,
EXEC_CONF_SNIPPET = """
      <config xmlns:xc="urn:ietf:params:xml:ns:netconf:base:1.0">
        <configure>
          <__XML__MODE__exec_configure>%s
          </__XML__MODE__exec_configure>
        </configure>
      </config>
"""


CMD_VLAN_CONF_SNIPPET = """
            <vlan>
              <vlan-id-create-delete>
                <__XML__PARAM_value>%s</__XML__PARAM_value>
                <__XML__MODE_vlan>
                  <name>
                    <vlan-name>%s</vlan-name>
```

Nicira NVP Plugin

```
class NVPCluster(object):
    """Encapsulates controller connection and api_client for a cluster.

    Accessed within the NvpPluginV2 class.

    Each element in the self.controllers list is a dictionary that
    contains the following keys:
        ip, port, user, password, default_tz_uuid, uuid, zone

    There may be some redundancy here, but that has been done to provide
    future flexibility.
    """
    def __init__(self, name):
        self._name = name
        self.controllers = []
        self.api_client = None

    def __repr__(self):
        ss = ['{ "NVPCluster": [']
        ss.append('{ "name" : "%s" }' % self.name)
        ss.append(',')
        for c in self.controllers:
            ss.append(str(c))
            ss.append(',')
        ss.append('] }')
        return ''.join(ss)

    def add_controller(self, ip, port, user, password, request_timeout,
                       http_timeout, retries, redirects,
                       default_tz_uuid, uuid=None, zone=None):
        """Add a new set of controller parameters.
```

FloodLight Rest Proxy

```
    def get_networks(self):
        ret = []
        try:
            vnetq = self.nova_vnets.select()
            vnets = vnetq.execute()
            for i in range(vnets.rowcount):
                (_, net_id, tenant_id, _) = vnets.fetchone()
                if tenant_id is None:
                    tenant_id = 'default'
                ret.append((tenant_id, net_id))
        except Exception, excp:
            LOG.error(str(excp))
            return None
        return ret


class ServerProxy(object):
    """
    REST server proxy to a network controller
    """

    def __init__(self, server, port, ssl, auth, timeout, base_uri, name):
        self.server = server
        self.port = port
        self.ssl = ssl
        self.base_uri = base_uri
        self.timeout = timeout
        self.name = name
        self.success_codes = range(200, 207)
        self.auth = None
        if auth:
            self.auth = 'Basic ' + base64.encodestring(auth).strip()
```

Challenges

- The idea that all networks should be built like the Internet is getting in the way. I have to build Campus and Healthcare networks in the same manner as Tier1 Service Providers build their networks and we build our Regional SP network in the state to find reliability and scale. Time and Space should be taken into account.

- The Internet will be the LAN in the next decade. The Carriers, LECs and Cable companies are not investing the capital necessary to scale.

- Distribution of state and mapping of elements in a decoupled Control Plane. This has been solved in systems today eg. Hadoop / HGFS. Tracking millions of elements over some distributed controllers and maintaining state is well within the realm of reality.

- “Big old tech companies are often incapable of investing in new ideas because they're addicted to the revenue streams from their current businesses and don't want to disrupt those businesses.” - Marc Andreessen Silicon Valley VC Firm

- If providing IaaS understand Self-Provisioning and Customer Experience.

How To Get Involved

- Get Involved with test beds with the community and vendors.
- Thought leadership and knowledge transfer.
- Dip your toes in the public and private Cloud.
- Installers for local OpenStack instances. @ http://www.rackspace.com/ & https://airframe.pistoncloud.com/
- Participate in the Internet2 working group. Co-Chairs Dan Schmeidt WILLYS@clemson.edu & Deniz Gurkan dgurkan@Central.UH.EDU
- http://incntre.iu.edu/openflow IU OpenFlow in a Day Class.
- Networking vendors need to realize one of the role reversal in Networking that has occurred (thanks in large part to R&E and DIY's). Just as in the x86 market it is consumer driven. The legacy echo chambers product managers live in are no longer acceptable.
- Mimic what the software industry perfected. Open and communities of practice.

Brent's Bookmarks – Comments, Questions, Nerd Rage?

- http://ioshints.info
- http://etherealmind.com
- http://nerdtwilight.wordpress.com/
- http://networkheresy.com/
- http://www.openflowhub.org/ (Floodlight)
- http://www.noxrepo.org/ (POX)
- First 10 minutes of McKeown's presentation for anyone with manager in title not to mention brings tears to my eyes.
- http://www.youtube.com/watch?v=W734gLC9-dw (McKeown)
- http://packetpushers.net
- http://www.codybum.com
- http://www.rackspace.com (Rackspace OpenStack Private Cloud build)
- http://www.networkworld.com/community/fewell
- http://networkstatic.net/ My Ramblings
- irc.freenode.net #openflow #openvswitch #openstack

Closing- Comments Questions Nerd Rage?