Chesapeake NETCRAFTSMEN

CISCO PARTNER Gold Certified

Strategic Datacenter Design

February, 2014

Dr. Peter J. Welcher

About the Speaker

- Dr. Pete Welcher
  - Cisco CCIE #1773, CCSI #94014, CCIP, CCDP
  - Specialties: Large datacenter and network design and assessment, IP multicast, QoS, MPLS, Large-Scale Routing & Switching, High Availability, Management of Networks
  - Customers include large enterprises and hospitals, federal agencies, universities, large financial organizations, large App service provider
  - Taught many of the Cisco courses over the years, taught Nexus 5K/7K class once a month
  - Reviewer for many Cisco Press books, book proposals; designed and reviewed 2.0 revisions to the Cisco DESGN and ARCH courseware; tech reviewer for 2.1 version of ARCH book
  - Presented lab session on MPLS VPN Configuration at CPN 2003-2004, and Networkers 2005-2007; presented BGP lab session at Cisco Live 2008-2010; presented lab session on Nexus in 2011; may present in 2012
- Many blogs at http://www.netcraftsmen.net/welcher and in the archive

Chesapeake NETCRAFTSMEN

CISCO PARTNER Gold Certified

Abstract

This talk will cover a variety of Datacenter Topics, updating previous talks. The emphasis will be on strategic design: how do new technologies affect the datacenter, how do they solve business problems.

The presentation will contain a mix of slides and whiteboarding. We will briefly review FabricPath, OTV and Cisco 1000v. We will then discuss VXLAN technology with a segue into VMware NSX, as well as Cisco DFA (Dynamic Fabric Automation) and Cisco ACI (Application Centric Infrastructure).

In each case, the focus will be on what we can do with the technology, some idea of how it works, and where and when the technology will be appropriate, and pros and cons.

Since this talk will cover a lot of ground at a high level, time will be left at the end for questions and design or technology pro/con discussions, also sharing of actual experiences.

Chesapeake NETCRAFTSMEN 3 Copyright 2014

Agenda

- New Nexus Switches
- FabricPath
- OTV
- 1000v and Virtual Appliances
- VXLAN
- VMware NSX
- DFA
- ACI
- Automation and SDN
- Conclusions and Summary

4 Copyright 2014

Chesapeake NETCRAFTSMEN

CISCO PARTNER Gold Certified

New Nexus Switches – 1

- Nexus 9K series: (mostly) ACI ready
  - 9396PX: 48 x 1/10G + 12 x 40G non-blocking
  - 93128TX: 96 x 1/10G + 8 x 40G non-blocking
  - 9508: up to 288 x 40 G non-blocking
  - Some line cards mix 1/10 G and 40 G ports*
  - 40 G BiDi optics
- Nexus 7700 / 70xx series: 7706, 7710, 7718
  - F3 line card: 12x40G // 12 x 100G, 24 x 40G, 48 x 10G
  - FabricPath, OTV, VPLS, LISP, MPLS

Chesapeake NETCRAFTSMEN 5 Copyright 2014

New Nexus Switches – 2

- Nexus 5K series: new 5672UP, 56128P
  - 5672UP: 72 UP 10 G ports + 6 x 40 G uplinks
  - 56128: 128 x 10 G + 12 x 40 G, 1 microsec latency
- Nexus 3K series: new 3172TQ
  - L2/L3 ToR, 48 SFP+ for 100/1G/10G, 6 QSFP+ for 40G or 4x10G
- [**Gratuitous pictures showing "yes that's a Cisco chassis" omitted here]

Chesapeake NETCRAFTSMEN 6 Copyright 2014

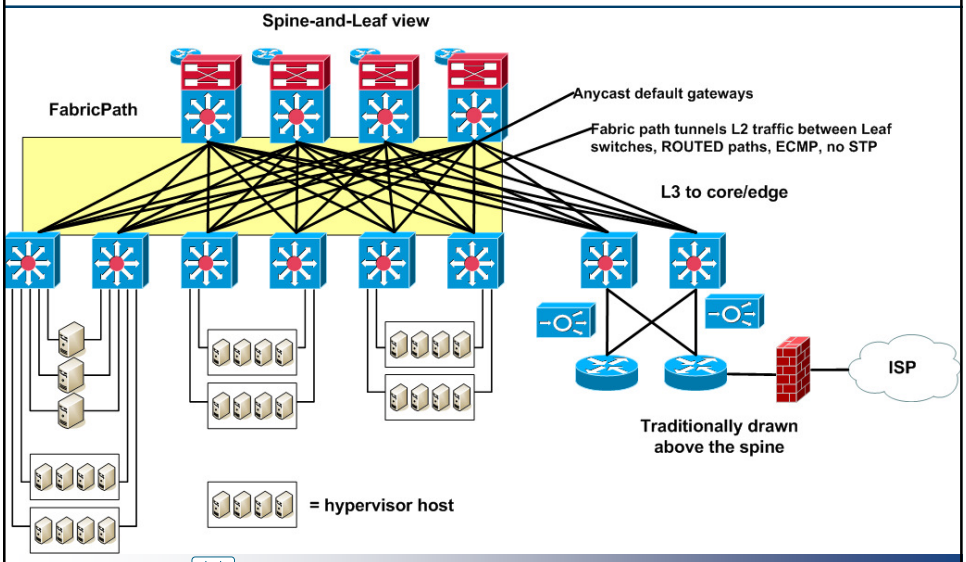
Planning Implications

- More ports, faster speeds
  - What do you need?
  - Will there be a UCS chassis with 40 G technology coming?
- Some design needed… also positioning matters: Spine/Leaf, ACI, DFA support
  - Get help when building a Bill of Materials!
- Check features and forward compatibility
- Consider hardware platform and code maturity

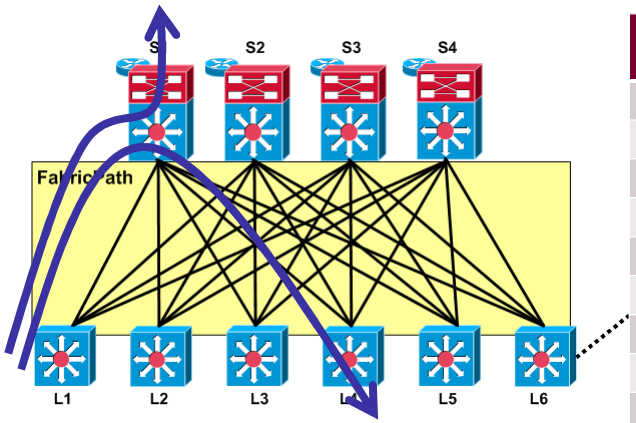
Agenda

- New Nexus Switches
- FabricPath
- OTV
- 1000v and Virtual Appliances
- VXLAN
- VMware NSX
- DFA
- ACI
- Automation and SDN
- Conclusions and Summary

FabricPath



FabricPath Routes Tunneled Traffic



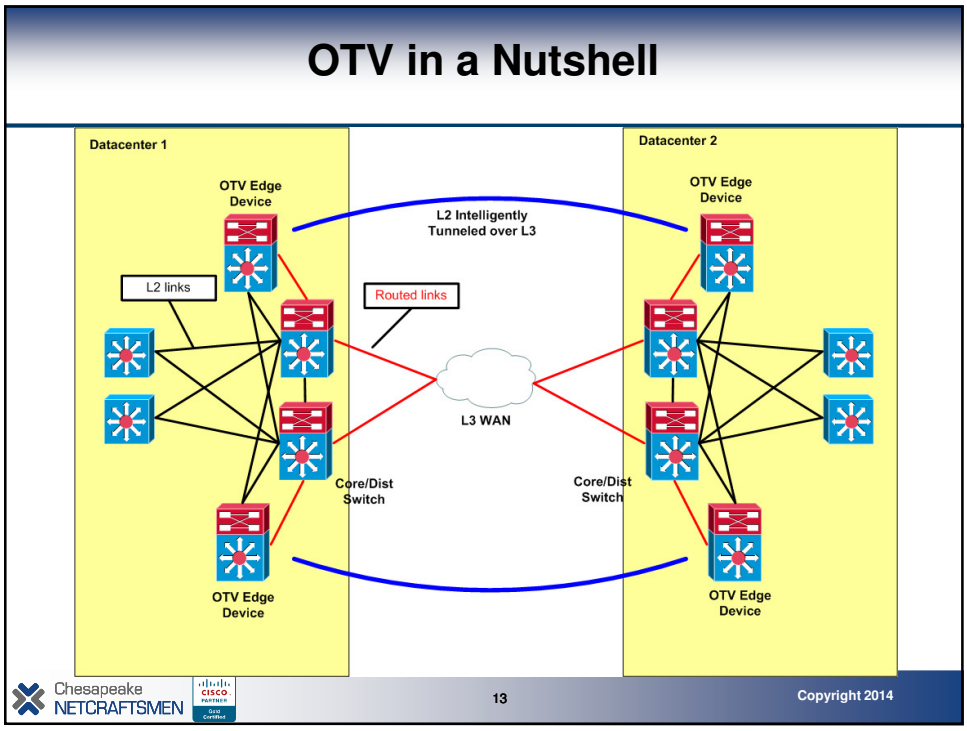
Switch	Forward on Interface
S1	1
S2	2
S3	3
S4	4
L1	1-4
L2	1-4
L3	1-4
L4	1-4
L5	1-4

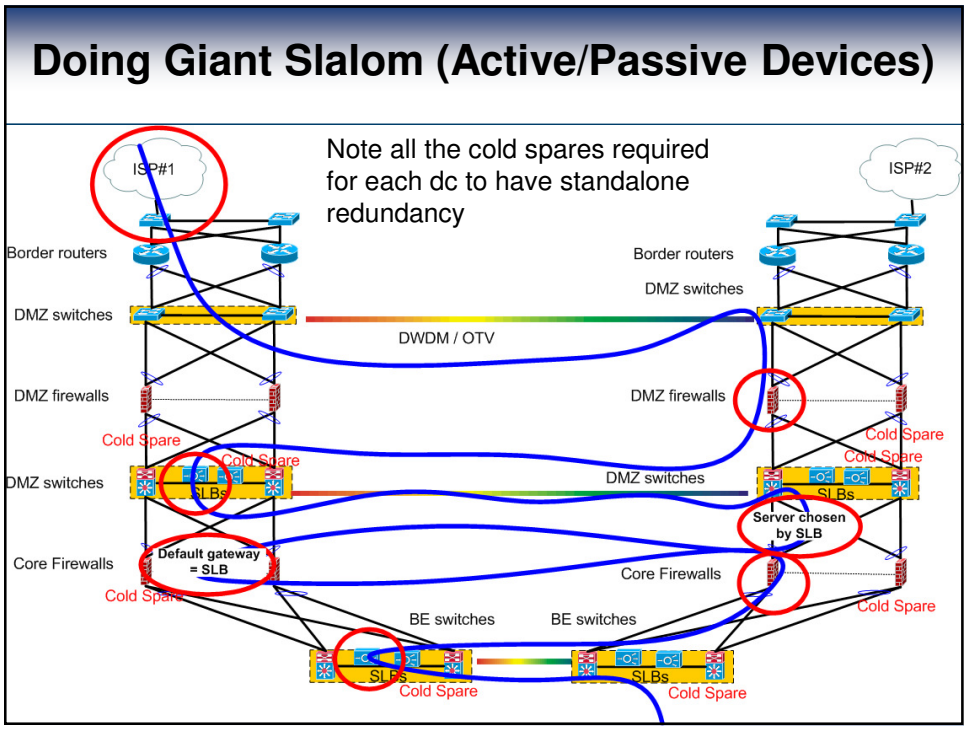
Some Implications

- FabricPath is now mature technology, rather easy to deploy, can deploy incrementally
- FabricPath gives you a way to do without Spanning Tree (and associated risk) on a Nexus datacenter core
  - Legacy STP can be preserved at edges
  - FP fabric looks like one giant switch to legacy edge devices
- Routing off FP at the spine / core is in principle a potential bottleneck
  - VLAN to VLAN traffic problem?
  - Usually don't need that much bandwidth to users / Internet
- Steps towards Cisco DFA and future technology readiness

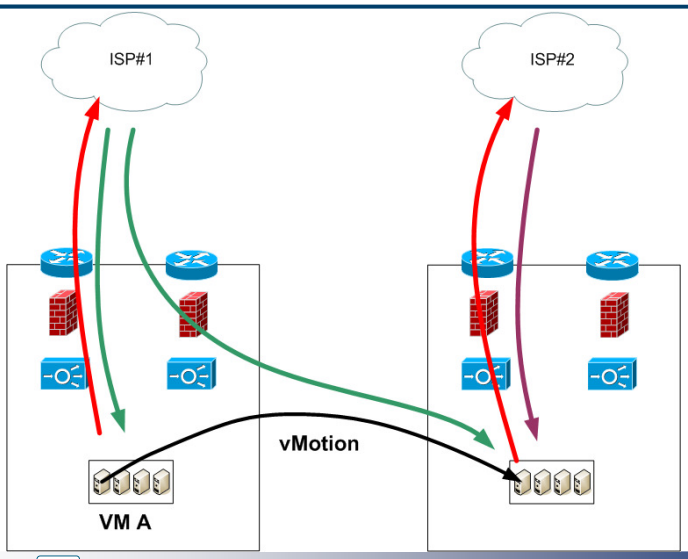
Agenda

- New Nexus Switches
- FabricPath
- OTV
- 1000v and Virtual Appliances
- VXLAN
- VMware NSX
- DFA
- ACI
- Automation and SDN
- Conclusions and Summary





Traffic Optimality



OTV / DCI IMPACT

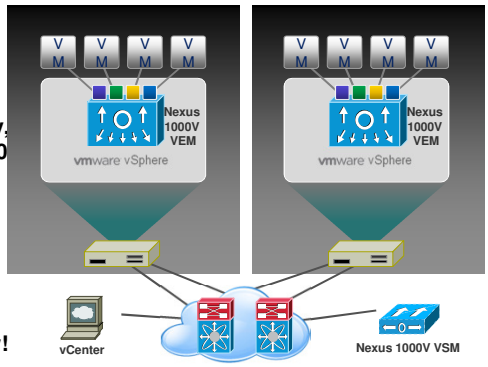
- **Pros:**
  - **Mature technology … still need STP defenses!**
  - **Supports vMotion, DRS, SRM – if close enough (distance limits!)**
  - **Less need for MS clustering @ L2 now…**
  - **FW, SLB clustering support (good/bad?)**
- **Cons:**
  - **DCI: creates shared fate**
  - **vMotion can abruptly discover an application latency sensitivity**
  - **Clustered FW, SLB = SPoF, can have other issues**
    - **Traffic tromboning**
  - **Gateway optimality issues: LISP and the Internet – when?**
    - **/32 scaling issue moved from BGP to LISP: good solution?**

Agenda

- New Nexus Switches
- FabricPath
- OTV
- 1000v and Virtual Appliances
- VXLAN
- VMware NSX
- DFA
- ACI
- Automation and SDN
- Conclusions and Summary

1000v Functionality

- 1000v and Family:
  - VSM + VEMs – distributed vSwitch
  - Virtual interfaces tied to VMs
  - VSG, ASA 1000v, vNAM, CSR 1000v, Imperva WAF 1000v, NetScaler 1000 other virtual appliances
  - vPath 2: service chaining
  - Secure multi-tenancy
  - 1000v InterCloud
  - Up to 128 VMware hosts / VSM now!
- 1000v and VXLAN:
  - Multicast VXLAN
  - Unicast VXLAN – using VSM at L3…
  - VXLAN gateway



Cisco content

1000v IMPACT

- **Pros:**
  - Network visibility into the dvSwitch
  - Port profiles for consistent provisioning
  - Cisco features, include SGT ACL enforcement (tags based on user ID or server groupings)
- **Cons:**
  - Have to talk to server / VWware admins ☹ (joking!)
  - VMware dvSwitch has increased functionality overlap
  - Market share? Mind share?
  - Adds some complexity to VMware admin, especially upgrades

Non-Cisco Virtual Appliances

- **There are many available, more coming:**
  - Brocade Vyatta, NetScaler VM, Palo Alto firewall / Panorama management software (for NSX for vSphere)
  - More functional than base NSX or hypervisor, or basic cloud functionality (or why buy them?)
  - Scale up versus scale out choice
- **vMotion and statefulness: with a v-appliance, you can take your state with you!**

vAppliances: IMPACT

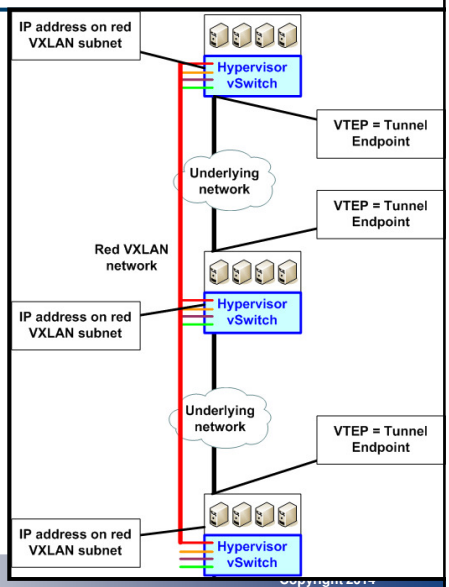
- Pros:
  - Puts the FW and SLB near the app
  - Localizes configuration, impact of mistakes
  - May be able to port current skills and configurations
- Cons:
  - Performance (1 Gbps? N x 10 Gbps?)
  - Sheer number of v-devices to manage
  - Lifecycle management, license costs
  - Learning curve
  - Who manages what? What about vRouters, etc. in the cloud?
- ACI partners: GUI manages hardware
- NSX partners: GUI or Plug-In manages v-appliance

Agenda

- New Nexus Switches
- FabricPath
- OTV
- 1000v and Virtual Appliances
- VXLAN
- VMware NSX
- DFA
- ACI
- Automation and SDN
- Conclusions and Summary

What is VXLAN?

- Vendors seem to vary usage
- A VXLAN is a logical switched network, a virtual VLAN: "VXLAN network"
- A VXLAN tunnel is a stateless IP tunnel
  - Allows devices attached to a VXLAN (logical switch) to communicate at L2 across different subnets

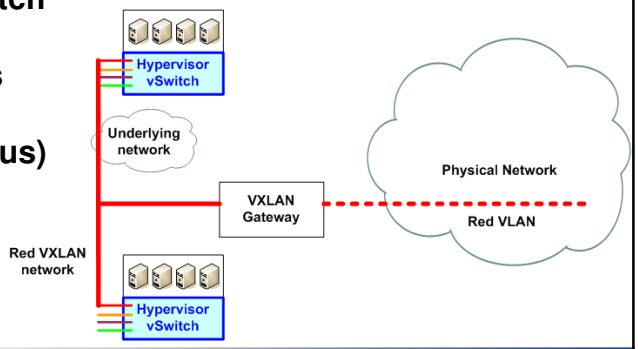


Why VXLAN?

- Cisco: it allows you to interconnect 1000v components separated at L3
- VMware:
  - Allows vMotion across hypervisor hosts in different VLANs, subnets – more scalable datacenter!
  - vMotion etc. can't change interface IP easily, need to stay in same subnet
  - Solution: VXLANs as virtual VLANs

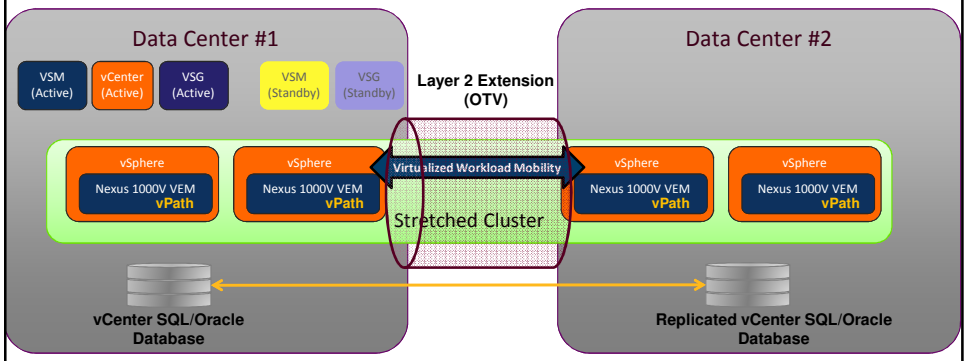
More Terminology

- **VXLAN Gateway: something that bridges a VXLAN (network) to a VLAN**
  - **Could be hypervisor**
  - **Could be a switch with hardware encaps/decaps support (e.g. some Nexus)**



Maintain Network / Security Policies across Datacenters

Nexus 1000V VSM Pair & VSG Pair (or VSG/VSG hosted on Nexus 1010s)



Migrate virtual workloads seamlessly across Data Centers
Maintain transparency to network & security policies (via N1KV & VSG)

VXLAN and DCI

- VXLAN can be used between datacenters
  - But is doing so a good idea?
- CAUTION: Latency is still YOUR problem – think long distance vMotion
- Is VXLAN gatewaying and OTV for DCI a better answer?
- VMware (and likely 1000v InterCloud) supports encrypted VXLAN tunnels

VXLAN and BUM Traffic

- A VXLAN network acts like a VLAN
- Floods BUM (broadcast, unknown unicast, multicast) traffic
- VXLAN started out using multicast groups for this (one IPmc group per VXLAN network)
  - Various ways to optimize and scale this
- VMware and 1000v now support stateless unicast tunnels
  - MAC, IP to VTEP IP mapping needed

VXLAN IMPACT

- **Pros:**
  - Provides L2-like adjacency / mobility over L3
  - There can be more VXLANs than VLANs (24 bit or 16 million versus 12 bits or approximately 4096)
  - Unless you need that much segmentation (big datacenter), why not just use FabricPath, do the flooding etc. in hardware?
- **Cons:**
  - BUM flooding still is … flooding
  - Manageability, robustness of multicast?
  - VMware admins can create VXLAN "spaghetti" whether you like it or not
  - Troubleshooting requires understanding fundamentals, ARP

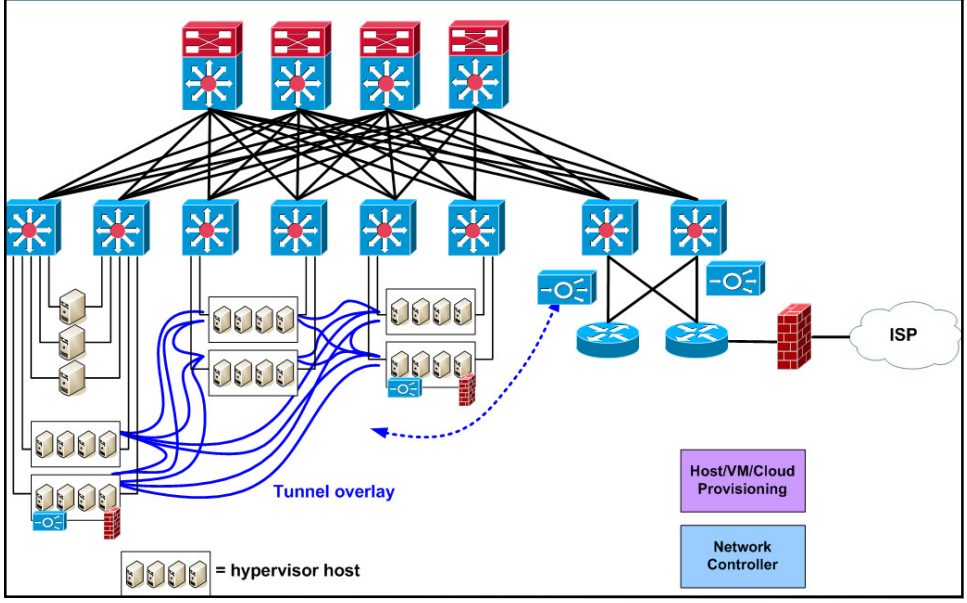
Agenda

- New Nexus Switches
- FabricPath
- OTV
- 1000v and Virtual Appliances
- VXLAN
- VMware NSX
- DFA
- ACI
- Automation and SDN
- Conclusions and Summary

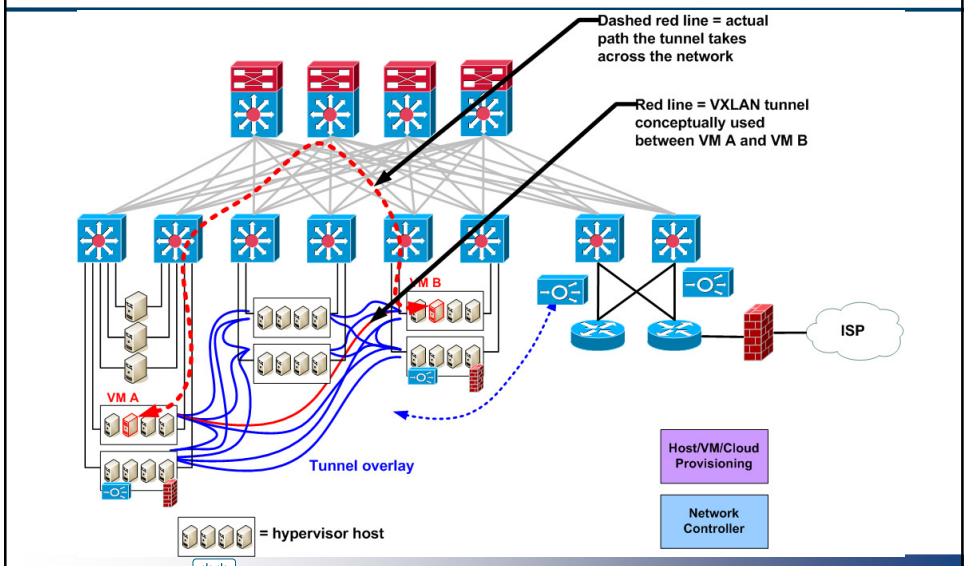
VMware NSX

- Evolution of VMware vShield Edge (vSE), etc.
- Leverages overlay tunnels: VXLAN or STT or NVGRE encaps
- Based on ovswitch and OVSDB
  - Information distributed by controller using OVSDB- protocol
  - Separate tunnel interface created on each host for each new host added to vSphere
  - Controller cluster distributes MAC to IP (tunnel interface) and segment ID mapping info
- Intended to be controlled via cloud provisioning tools, including OpenStack via Neutron plug-in (can only have one Neutron plug-in right now)
- Physical appliances can tie in by running ovswitch internally

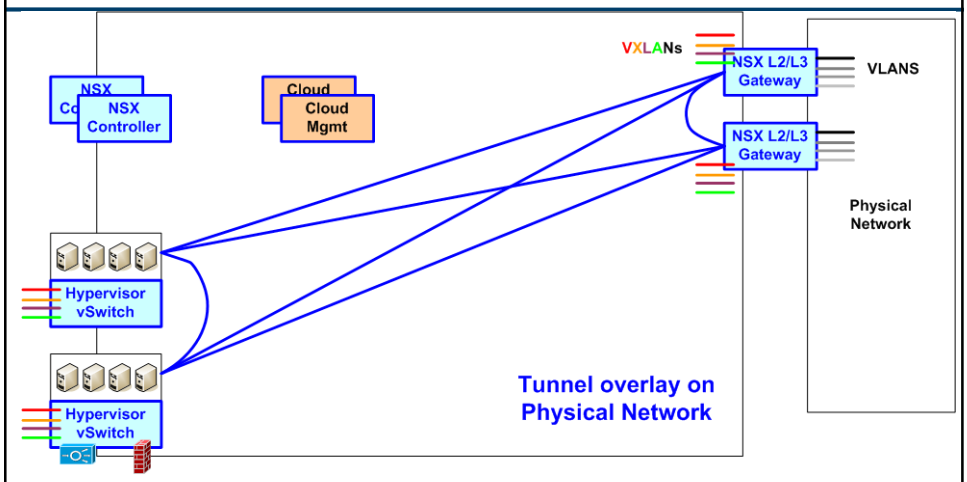
NSX VXLAN Tunneling



Tunnels and the Physical World



A Different View of the World



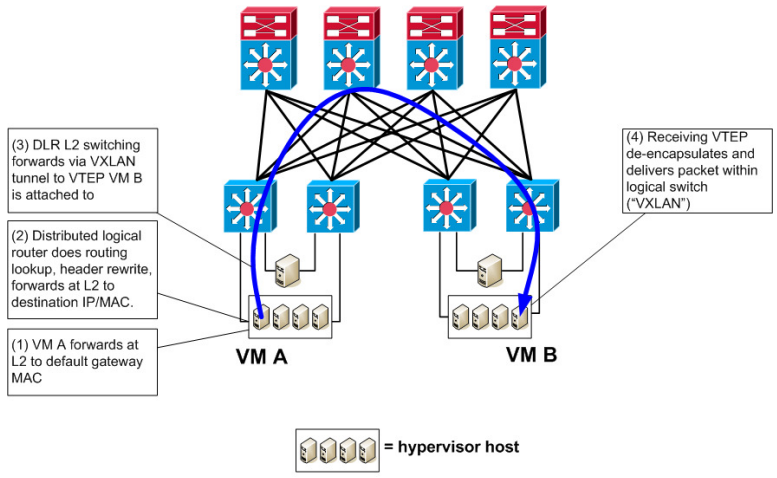
- **Copyrighted image: Steinberg_New_Yorker_Cover.png**

NSX – 2

- **Assumes robust L2 or L3 network**
  - **CLOS tree / spine-and-leaf for max throughput recommended**
  - **Traffic patterns may become "interesting"**
- **Lets you stand up multiple VXLANs**
  - **Act like VLANs, but don't depend on network VLANs**
  - **Can gateway (bridge) to VLANs (NSX or hardware)**
  - **Tunnels can extend to public cloud, w/ encryption for security**
- **Can have distributed routing and firewalling between VXLANs**
  - **Up To 10 Gbps, done in VMware kernel**

NSX Distributed Routing between VXLANs

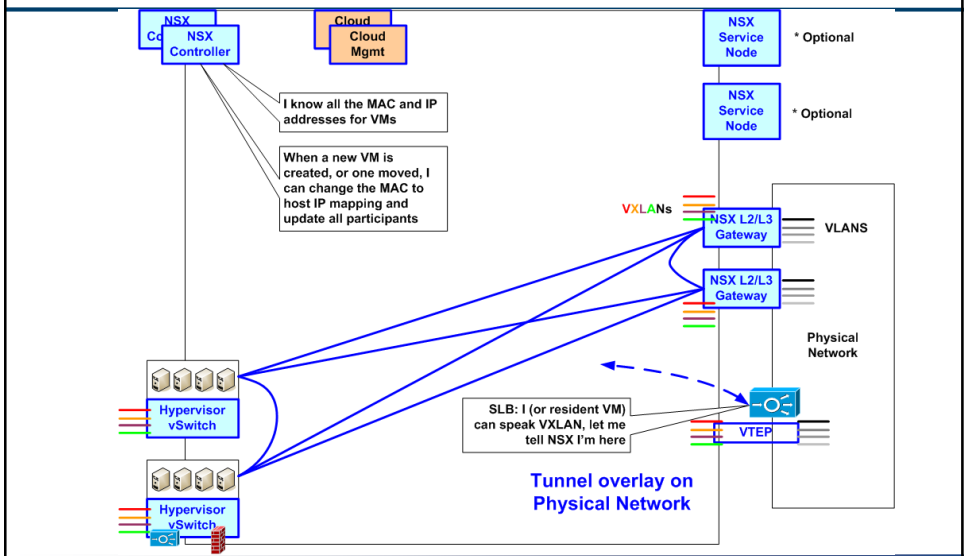
NSX: Routing from A to B, in different VXLANs



NSX – 3

- Can have edge routing to the physical world
  - Think virtual edge router that can do NAT and firewall functions as well
  - Stateful firewall for return traffic
  - OSPF and BGP
  - IPsec, DNS, DHCP
- Service nodes can be added to handle BUM traffic
  - Acts somewhat like an ATM LANE BUS server

NSX and ARP



NSX – 4

- Business case: agility, speed of deployment
- VMware admins can use GUI to stand up edge NAT + FW + SLB + VMs with internal FW zoning and segment routing
  - Like vSE, can "clone" a service pod (App, middleware, DB, FW, SLB VMs) and just adjust the public VIP for NAT
  - If you don't think the NSX FW and SLB are good enough, can use virtual or physical ones
- NSX appears to have hit critical technical mass

VMware NSX IMPACT

- Pros:
  - GUI-based automation with "good enough" networking
  - Leverages open source elements so good chance to dominate the market, cross-hypervisor
  - Still will depend on high-speed physical switched network
  - VMware part of datacenter may become opaque to network staff – still need switches if have big iron, physical servers, etc. in datacenter
- Cons:
  - Doesn't configure physical switches
  - Bottleneck potential between physical and virtual worlds?
  - GUI for setting up BGP and some features takes a bunch of clicking – but don't have to buy and rack a router, either. Can create templates.
  - Early commentary: per-VM licensing, costly if actually implemented

Agenda

- New Nexus Switches
- FabricPath
- OTV
- 1000v and Virtual Appliances
- VXLAN
- VMware NSX
- DFA
- ACI
- Automation and SDN
- Conclusions and Summary

What is DFA?

- Dynamic Fabric Automation
  - Available now / 1Q 2014
  - See online documents for h/w, s/w required, what's supported
- DFA is incremental innovation using a mix of current components and some adapted technology (e.g. MP-BGP from MPLS VPN)
  - Marketed as evolutionary
  - Can pick and choose

Components of DFA – 1

- Uses free DCNM Essential Edition as "CPoM", single centralized point of management
- Leverages POAP to build Spine-Leaf fabric, device auto-configuration
  - Power-On Auto Provisioning
  - Template creation tool
  - Cabling consistency check

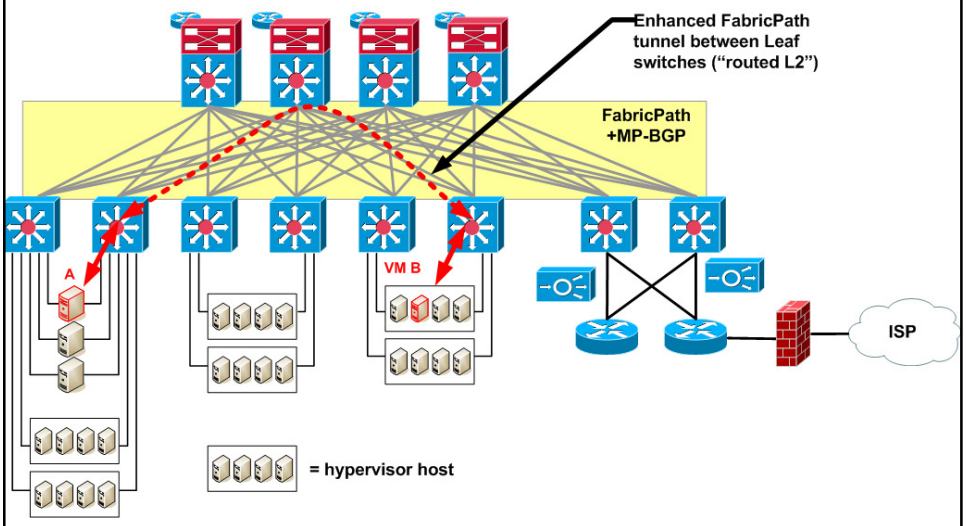
Components of DFA – 2

- Uses manual, VDP, DHCP, traffic to learn MAC / VLAN / VRF / Leaf association (VRF = tenant)
  - Tied to auto-config profiles
  - VLAN managed by vCenter, OpenStack, 1000v, etc.
  - Fully automatic with UCS Director, vCloud Director, OpenStack

Components of DFA – 3

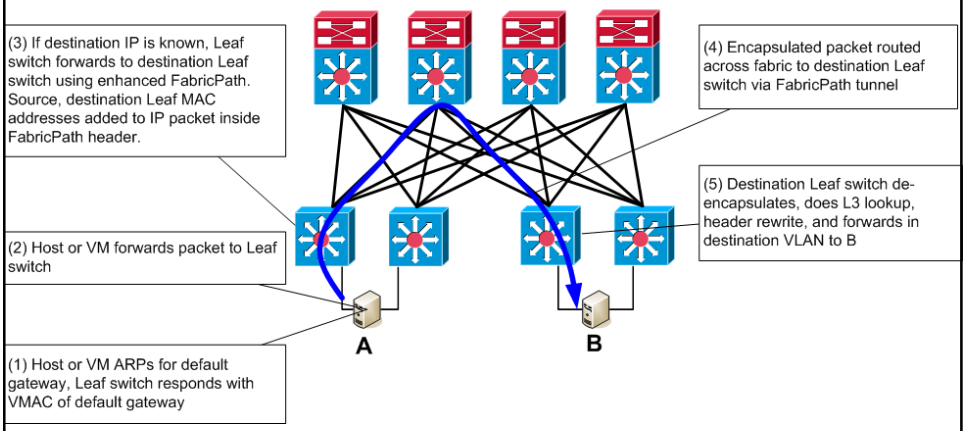
- Leaf switches act as distributed VLAN gateway
  - "Proxy gateway", proxy-ARP
  - N5K uses FabricPath behavior, anycast gateway
- VRF or "segment ID" isolates groups of L2, L3 networks
- MP-BGP tracks host IP location (Leaf switch) and VRF
- FabricPath+ tunnels traffic between Leaf switches (DFA-A = FabricPath, DFA-B adds VXLAN and other features)

DFA Uses FabricPath



DFA Routing

DFA: Routing from A to B



Impressions / Impact

- **DFA is incremental, less risky**
  - **Uses recent hardware: N6K, N7K F2, N5K with some limitations, N2K**
- **Moderate learning curve figuring how the moving parts fit together**
- **Some initial planning required**
  - **Getting the right hardware and software**
  - **Building templates etc.**
  - **Identifying gaps in present device feature support**
- **DFA: some work in progress but can start soon and fill in gaps / add functionality as you learn**

Agenda

- New Nexus Switches
- FabricPath
- OTV
- 1000v and Virtual Appliances
- VXLAN
- VMware NSX
- DFA
- ACI
- Automation and SDN
- Conclusions and Summary

What Is ACI?

- Application Centric Infrastructure
- New N9K hardware with merchant + custom silicon
  - Developed by Insieme (Cisco spin-in startup)
  - NX-OS, reduced to specific functionality
  - VXLAN, NVGRE, VLAN termination at Leaf nodes
  - VXLAN tunnel transport over fabric

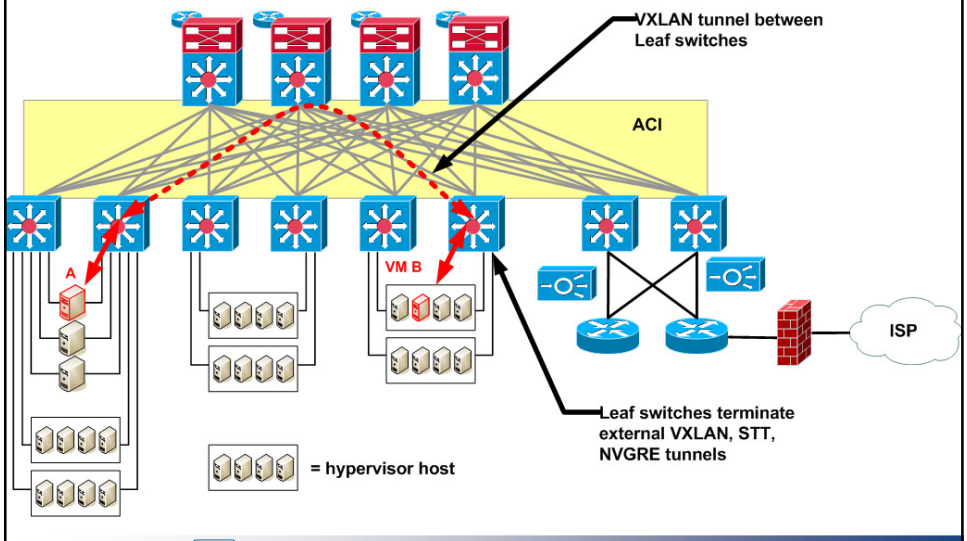
ACI – 2

- ACI fabric controller (APIC)
  - Stateless policy model
  - Telemetry, Health reporting
  - Service chaining
  - Hypervisor agnostic
  - Application agility
- GUI / code base similar to UCS Manager approach
- Northbound API for integration with OpenStack
- Southbound API

ACI – 3

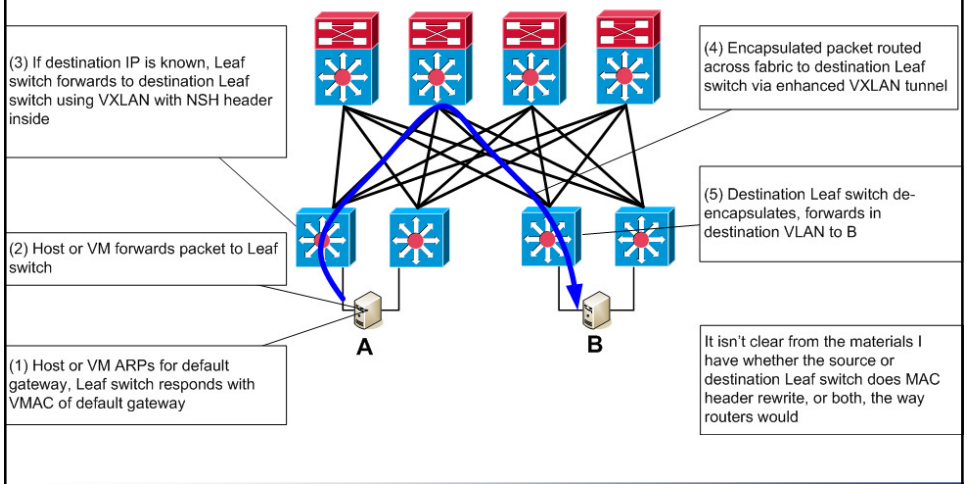
- ACI Fabric
  - Routed, runs IS-IS
  - (Seems like FabricPath re-purposed for VXLAN)
  - 1M IPv4 + IPv6 endpoints at edge
  - 64,000+ tenants
  - 1/10 Gbps edge
  - De-coupling of endpoint ID, location, policy from topology
  - Spine switches DB of MAC, IP, tenant, Leaf location
  - Fabric can do smart Load Balancing

VXLAN Tunnel Termination, Internal VXLAN



Routing in ACI

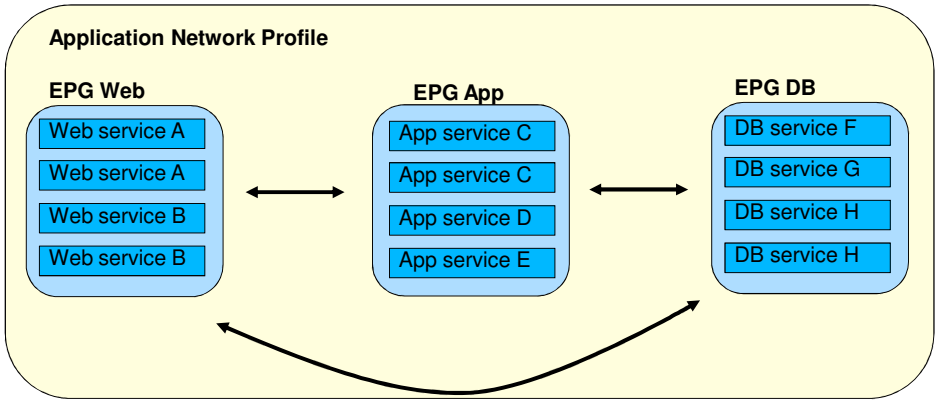
ACI: Routing from A to B



Policy in ACI

- (Cisco slides show this with lots of graphics)
- End Point Group (EPG) = Groups of servers or services
  - VLAN or VXLAN ID, perhaps VM port group
  - IP address or subnet
  - Future: perhaps DNS name, L4 ports
- Application Network Profile (ANP) = group of EPGs and in/out policies for allowed traffic between pairs of EPGs
  - Get out of the IP address security business?

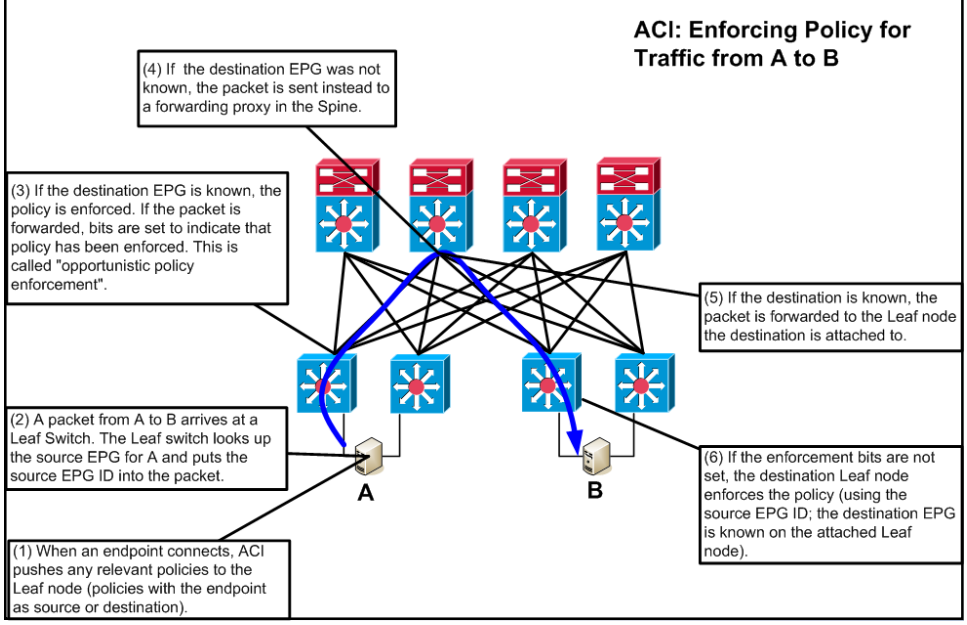
Application Network Profile



Policy in ACI – 2

- Contract: what allowed between EPGs and how it behaves:
  - Filter (TCP port 80)
  - Action (permit)
  - Label
- Action could be: permit, deny, redirect, log, copy, mark (DSCP)
- Security rules, QoS C&M, etc.
- Support for service chaining, templates

Policy Enforcement in ACI



Chesapeake NETCRAFTSMEN

CISCO PARTNER Gold Certified

Agenda

- New Nexus Switches
- FabricPath
- OTV
- 1000v and Virtual Appliances
- VXLAN
- VMware NSX
- DFA
- ACI
- Automation and SDN
- Conclusions and Summary

What Problem Are We Trying to Solve?

- Speed of deployment / agility?
- Staff productivity?
- Infrastructure as code?
- Visibility (esp. L2 and security appliances?)
  - Virtual relationship to Physical
  - Application problems to sets of servers and VMs to physical network
  - Resource constraints reaching max capacity

Vendor-Centric Automation

- Big vendor canned solutions:
  - Dynamic Fabric Automation (DFA)
  - Application Centric Infrastructure (ACI)
  - VMware NSX for vSphere
- APIs allow extending these tools
  - Or integrating them

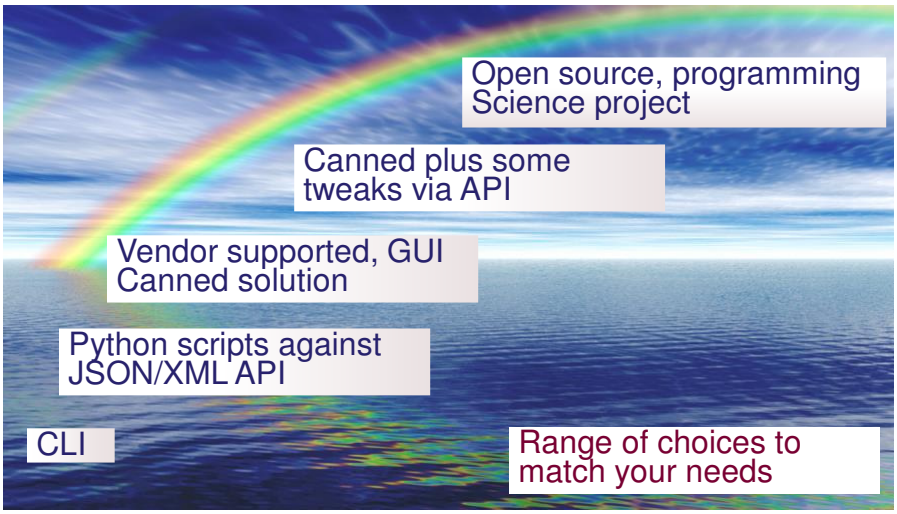
Open Source

- OpenDaylight
- OpenStack
- OpenVSwitch
  - A lot of energy going into the above
- OpenFlow **
  - Some very innovative ideas
  - Market passing it by? Or not?
  - My take: OF focus on flows is TMI, as a configuration tool it lacks any abstraction
  - Might be useful for control of hypervisors later?

Some SDN Thoughts

- SDN has many flavors, now means automation
- OpenFlow type SDN: what vendor is the market going to coalesce around? HP? Arista? NEC?
  - Issues with pie in the sky versus hardware realities?
  - Are tools really going to be vendor-neutral?
- SDN / automation will be needed for IoE
  - Cloud-managed, cloud data collection and analysis?
  - How many firms will be establishing a footprint in your home, sensor networks, etc.?

Spectrum of SDN Choices



Thought: Your Size Affects Your Choices

- If you have up to 400 VM's running on a small 4-8 CPU chassis with integrated flash and storage ("ultraconvergence"), do you need anything more than VMware / hypervisor? Do you care about cloud?
- If you're big enough, open source and coding aren't scary (risk, cost), dev team worth it (speed, savings)
- DFA hits a mid-ground with "current" Cisco switches, way to get started, do ACI on next core/access refresh?
- ACI is more visionary, may provide more gain, requires appropriate N9K hardware, chips
- Conclusion: start small, try something out, learn

Key SDN Questions

- What do you wish to automate? Physical, virtual, cloud?
- What do you hope to gain by doing so?
- Platform support
  - Will your HP server software play well with Cisco automation and vice versa?
- What are the limitations of a given tool or set of tools?
  - Vendor lock in?
  - What does the tool not do?
  - Dependencies, degree of complexity
  - Skills to deploy and maintain the tool?
  - What support model fits your needs? Is such support available?
  - GUI versus Puppet, Chef (text / code-like tools)

Agenda

- New Nexus Switches
- FabricPath
- OTV
- 1000v and Virtual Appliances
- VXLAN
- VMware NSX
- DFA
- ACI
- Automation and SDN
- Conclusions and Summary

Design Conclusions

- Spine / leaf topology is the new datacenter topology
  - L2 or L3: depends on the automation approach
  - Spine and Leaf may require different switches/cards
- Set and forget infrastructure has its advantages
- Items to understand:
  - How do you add services?
  - How do you do multi-tenant, security zones, etc. ?
  - How does failover work?
  - How to balance complexity versus robustness?

Operations Conclusions

- Most automation products will use VXLAN tunnels
  - Tunnel reporting must tie the overlay to the underlay, help troubleshoot
- If it isn't visible/manageable, it shouldn't be in the datacenter
- Knowing which apps map to which VMs and where those VMs are at a given time will be important
- Automation needs to make things simpler, not just wrap a GUI around the same complexity

Business Processes

- Business Process Changes coming!
  - Need to end stovepipes, get groups talking across former boundaries
  - When is a physical device appropriate, when is a virtual appliance better?
  - Who designs, deploys, manages virtual appliances?
  - Cloud? External connectivity, virtual network components for applications?
  - Need integrated tools, or at least change management and alerting log / dashboard with easy filtering

In Case You Wanted to Know

- Some other technologies:
  - TRILL: Designed by a committee. FabricPath is better in some ways, also Cisco-proprietary.
  - LISP: I'm not holding my breath waiting for ISP's to deploy lots of costly PITR routers. Until they do, LISP for the Internet seems like a non-starter.

Summary

- SDN/automation:
  - "May you live in interesting times!"
  - Change is happening fast, but datacenters change slowly
  - Many startups, few will survive
  - Startups will have to find a niche with the big ecosystems: Cisco and VMware? HP's mind-share?
  - I'm not convinced we all have to learn to program
  - There are lots of choices, pick what's best for your organization
  - Opportunity: gain efficiency, agility in datacenter
- <u>References: See links in my recent blogs</u>

Chesapeake NETCRAFTSMEN

CISCO PARTNER Gold Certified

Thanks

- Thanks to Cisco for the Cancun and Milan CiscoLive slideware, which helped immensely!
- Thanks to Ivan Pepelnjak, Brad Hedlund, Scott Lowe, for the NSX Architecture podcast plus slide deck
  - See Ivan and Brad's blogs for more NSX info and other good information
  - See Scott's blog for OpenSwitch, NSX, and other topics

Any Questions?

- For a copy of the presentation, email me at pjw@netcraftsmen.net
- About Chesapeake Netcraftsmen:
  - Cisco Gold Partner, 2nd in U.S. to meet 2012 broad certification requirements:
    - Data Center Architecture
    - Borderless Networks Architecture
    - Collaboration Architecture
  - Cisco Customer Satisfaction Excellence rating
  - We've done some large and very large data center assessments, designs, and deployments, large UC deployments, WLAN, etc.
  - Designed and assessed networks for several federal agencies, several well-known hospitals, large mortgage firms, stock firms, App commerce datacenters, law firms…

CISCO PARTNER Gold Certified

Chesapeake NETCRAFTSMEN

CISCO PARTNER Gold Certified

Chesapeake NETCRAFTSMEN

My e-mail: pjw@netcraftsmen.net

Company telephone: 888-804-1717

Company e-mail: info@netcraftsmen.net

76

Copyright 2014