Cloud Computing Networking

Under the Hood

Ivan Pepelnjak (ip@ioshints.info)
NIL Data Communications

Podatkovne komunikacije
Data Communications

Who is @ioshints?

- Networking engineer since 1985 (DECnet, Netware, X.25, OSI, IP ...)
- Technical director, later Chief Technology Advisor @ NIL Data Communications
- Started the first commercial ISP in Slovenia (1992)
- Developed BGP, OSPF, IS-IS, EIGRP, MPLS courses for Cisco Europe
- Architect of Cisco’s Service Provider (later CCIP) curriculum
- Consultant, blogger, book author

Focus:

- Core routing/MPLS, IPv6, VPN, Data centers, Virtualization
- Rock climbing, mountain biking ;)

The troubles usually start here ...

Someone Got the Next Great Idea

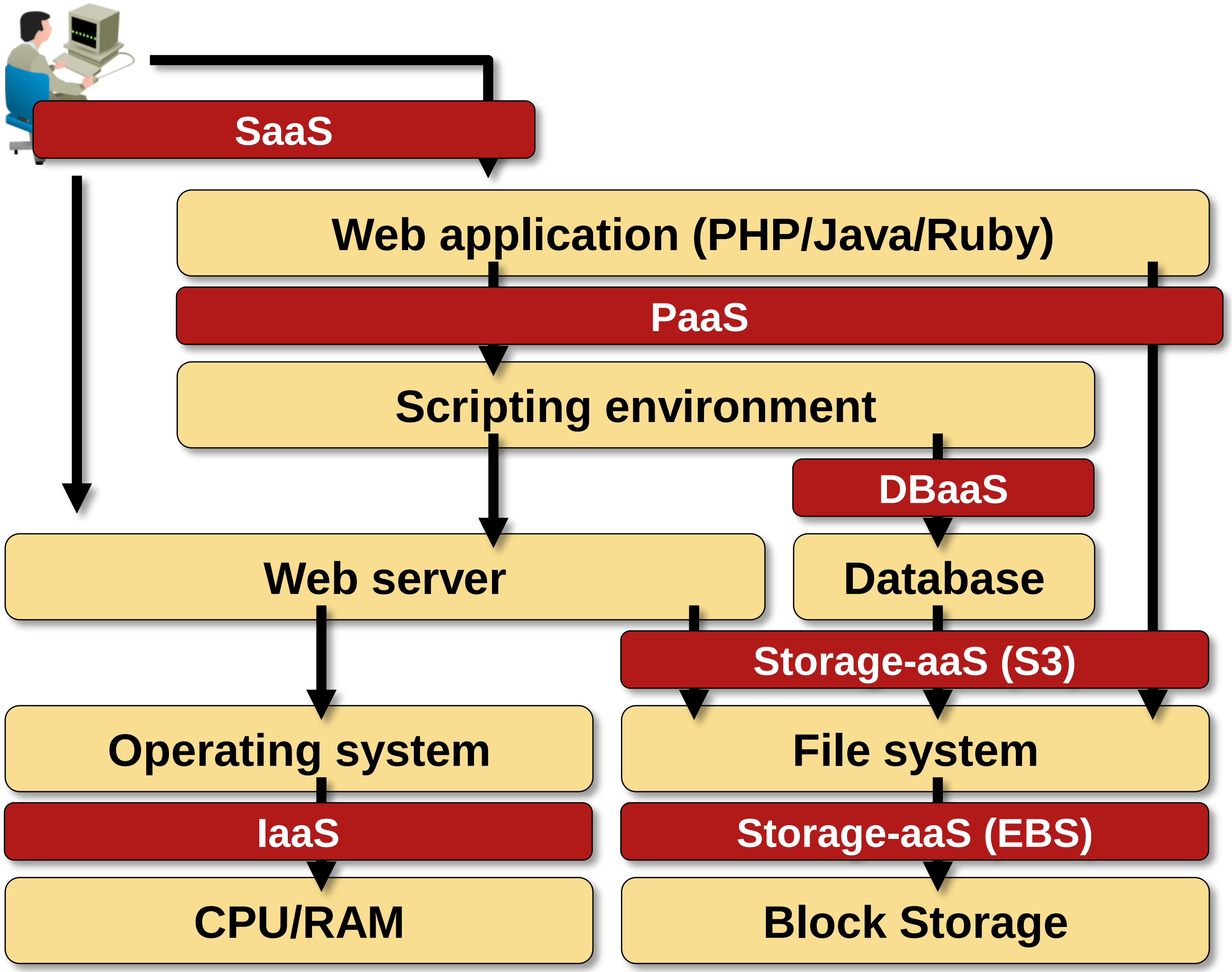
We’re launching the
cloud services next week

Cloud? What cloud? What services?

I don’t care. Get it done!

Image: Ambro / FreeDigitalPhotos.net

Cloud Services Taxonomy 101



What's different?

- Scalable
- Elastic
- Location-independent
- On-demand

Key ingredients

- Scalability
- Orchestration
- Customer-driven deployment

Do We Care?

Most cloud services are TCP-based applications

- SaaS, PaaS, DBaaS, Storage-aaS (iSCSI or HTTP interface)

Requirements

- Scalable & robust L3 network
- Lots of east-west traffic (replication, distributed file systems, multi-tier architectures)
- Scalable local & global load balancing is a major requirement
- Choose load balancers that offer high-level APIs (*expect* scripts or *ssh 'copy scp:file running-config'* don't count)
- Some cloud services might be implemented on top of IaaS service (server virtualization)

IaaS is a really tough nut to crack

IaaS 101

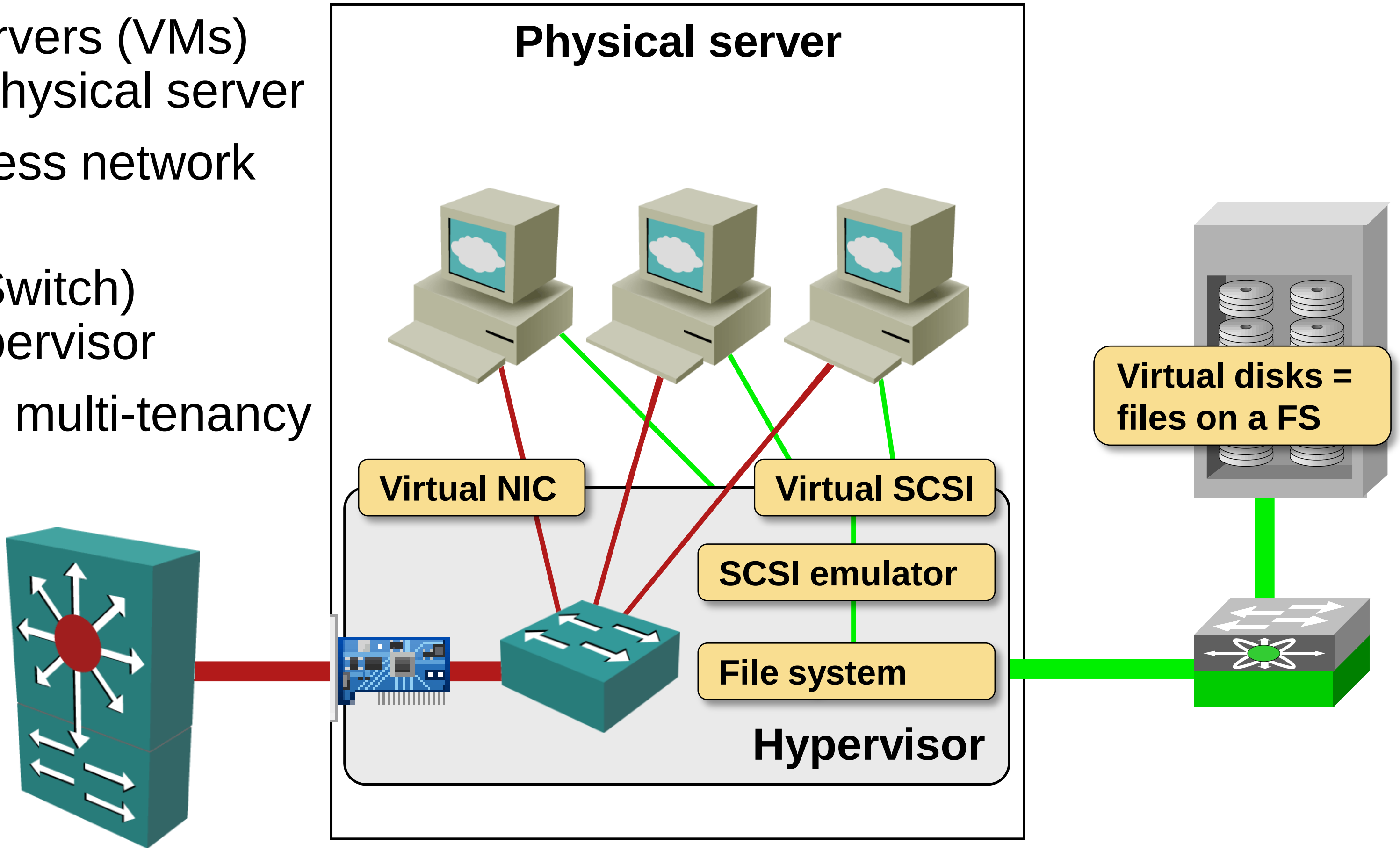
Hypervisor-based solutions:

- Multiple virtual servers (VMs) running inside a physical server

Q: how do VMs access network and storage

A: Virtual switch (vSwitch) embedded in hypervisor

Q: How do we solve multi-tenancy issues?



This-is-What-You-Get Approach

Making life easier for the cloud provider (early Amazon EC2)

- Customer VMs attached to “random” L3 subnets
- VM IP addresses allocated by the IaaS provider (example: DHCP)

Multi-tenant isolation

- Packet filters (example: iptables) applied to VM interfaces
- Predefined configurations or user-controlled firewalls

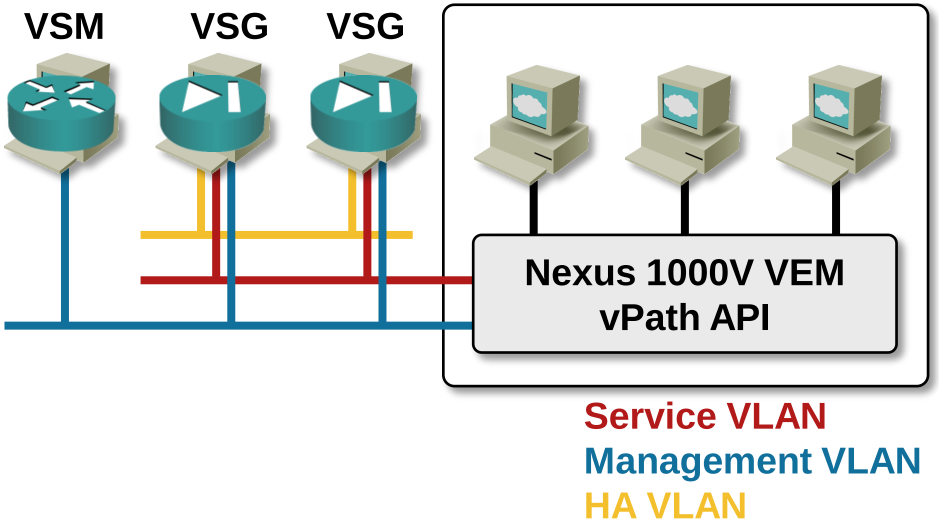
Implementation options

- XenServer/KVM with iptables
- vSphere with Cisco’s Virtual Security Gateway
- External firewalls (caveat: doesn’t address inter-VM attacks)

Virtual Security Gateway (Cisco)

Virtual Security Gateway = stateful FW

- NX-OS in a VM
- Interacts with Nexus 1000V VEM
- Redundant architecture
- VSG can serve many hosts/tenants

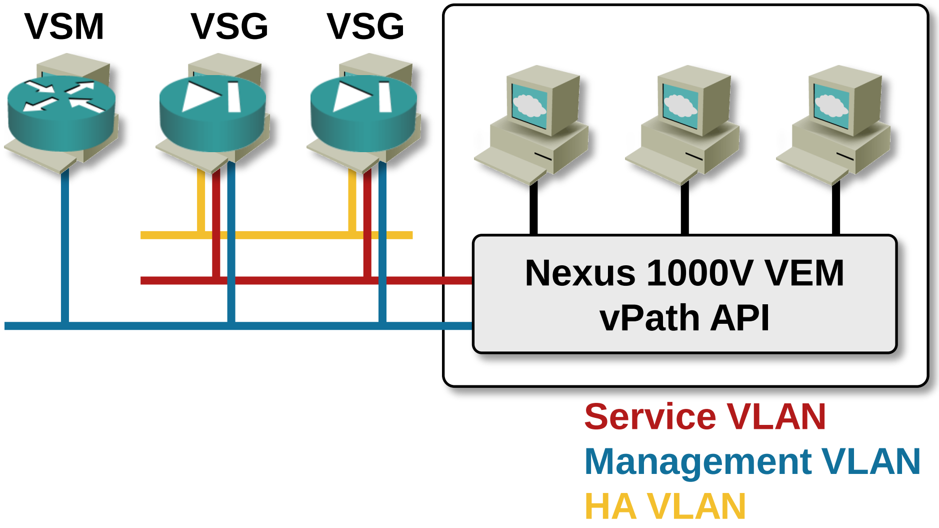


Principles of operation

- VN-service defined on port profile in Nexus 1000V
- Nexus 1000V forwards VM traffic to VSG on service VLAN
- VSG inspects and returns the traffic
- VSG can download 6-tuple (+VLAN) to VEM (fast-path offload)
- All subsequent packets in the same session are switched by VEM

VSG Multi-Tenant Deployment

VSG IP address and security profile name configured on port profile



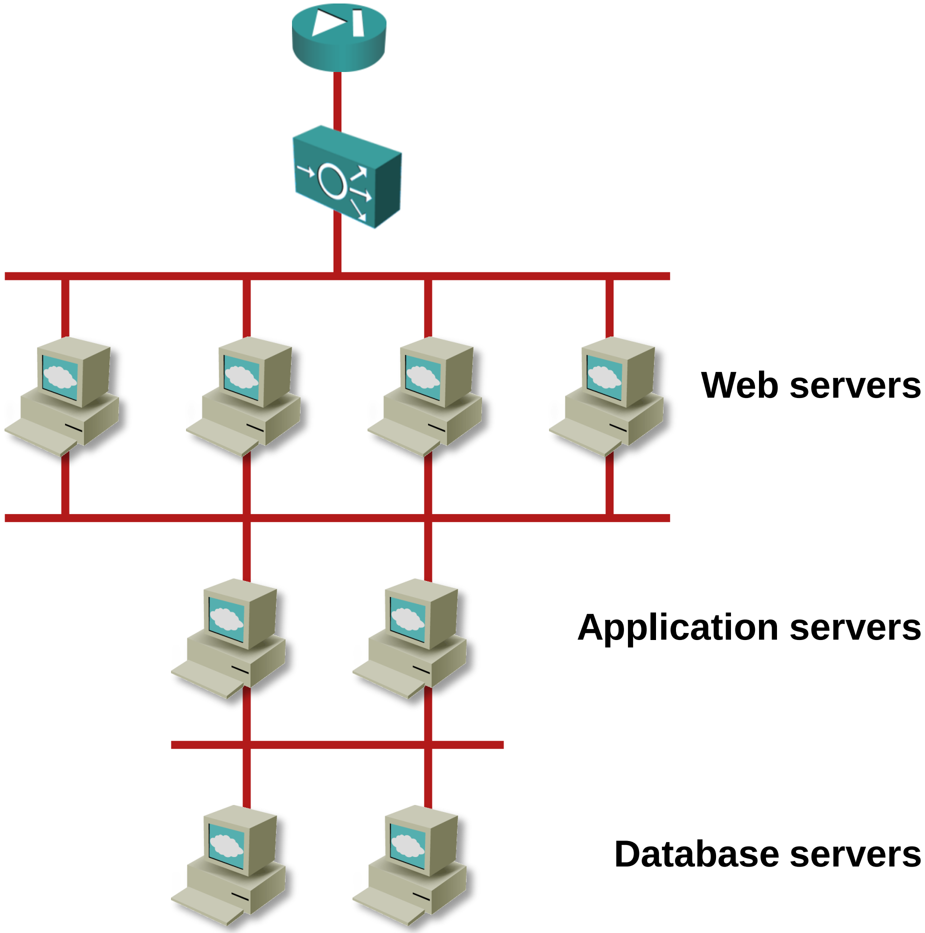
Option#1 – central VSG

- Each tenant has a different security profile
- Centralized firewall management

Option#2 – per-tenant VSG

- Each tenant gets a separate service VLAN
- Tenant's VSG configured in VSM port profile
- Tenant manages its own VSG

What the Customers Think They Want



Requirements

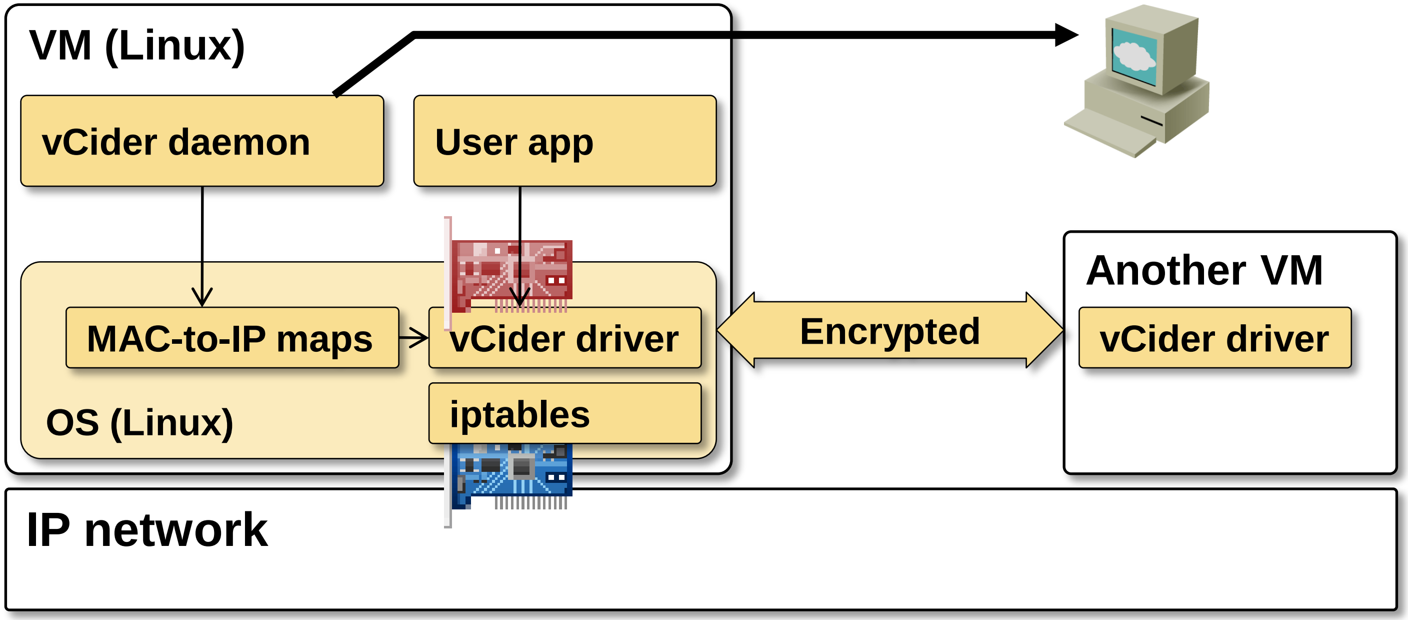
- Multiple logical segments
- Multiple NICs per VM
- Some segments isolated, other shielded with firewalls
- Unlimited scalability and mobility

Implementation options

- Userspace (vCider)
- Nested hypervisors (CloudSwitch)
- VLANs
- MAC-in-MAC (PBB, vCDNI)
- MAC-over-IP (VXLAN)
- IP-over-IP (Amazon EC2)

Remote access and software upgrades: the oops moment

vCider – Userspace MAC-over-IP



- Userland (VM) MAC-over-IP solution
- Each VM registers its node ID and IP address with vCider web-based service
- Customers can build on-demand networks
- All inter-VM traffic is encrypted

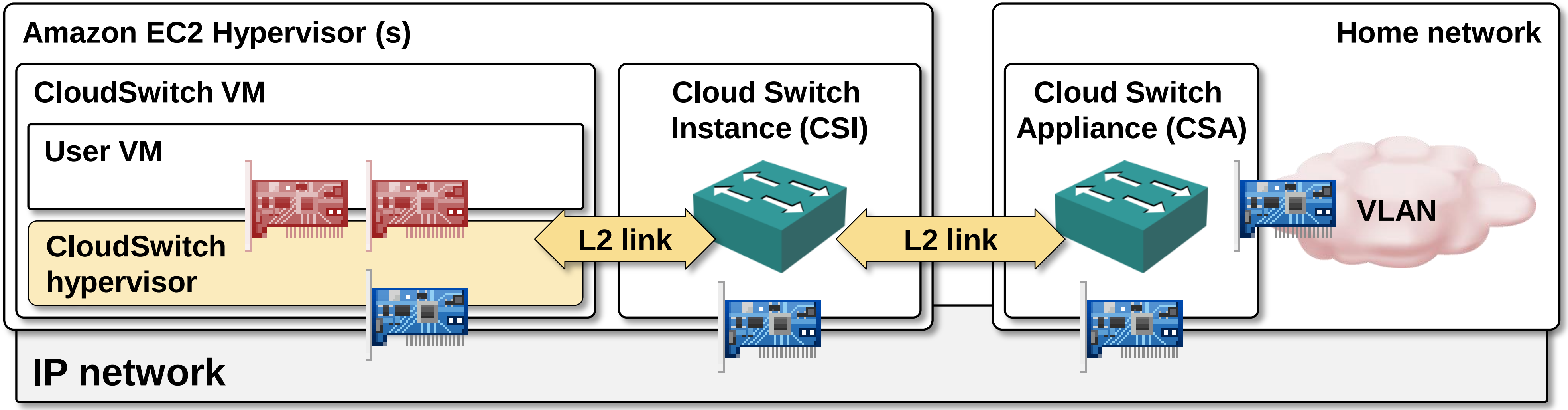
Benefits:

- Works with any virtualization system

Drawbacks:

- Linux only
- Requires VM changes (device driver)

CloudSwitch – Nested Hypervisors



- User VM runs within CloudSwitch VM within IaaS hypervisor (Amazon / Terramark)
- CloudSwitch VM provides multi-NIC support and MAC-over-IP services
- Per-cloud CSI: soft switch, VPN MAC-over-IP link to home network
- CSA: Control & VPN termination

Features:

- Works with any VMware VM (no VM modifications needed)
- Network and storage encryption
- Automatic cloud-side provisioning

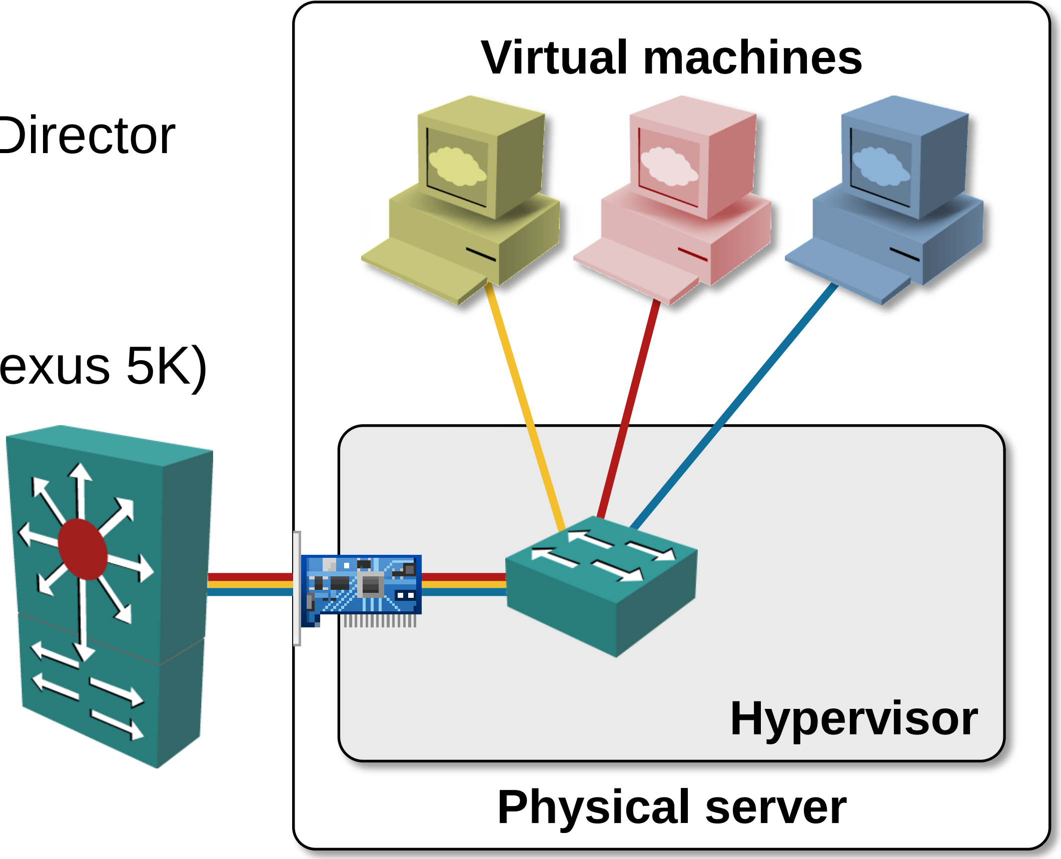
Primary use cases: Migration to cloud, cloudbursting

VLANs – The Ultimate Scalability Failure

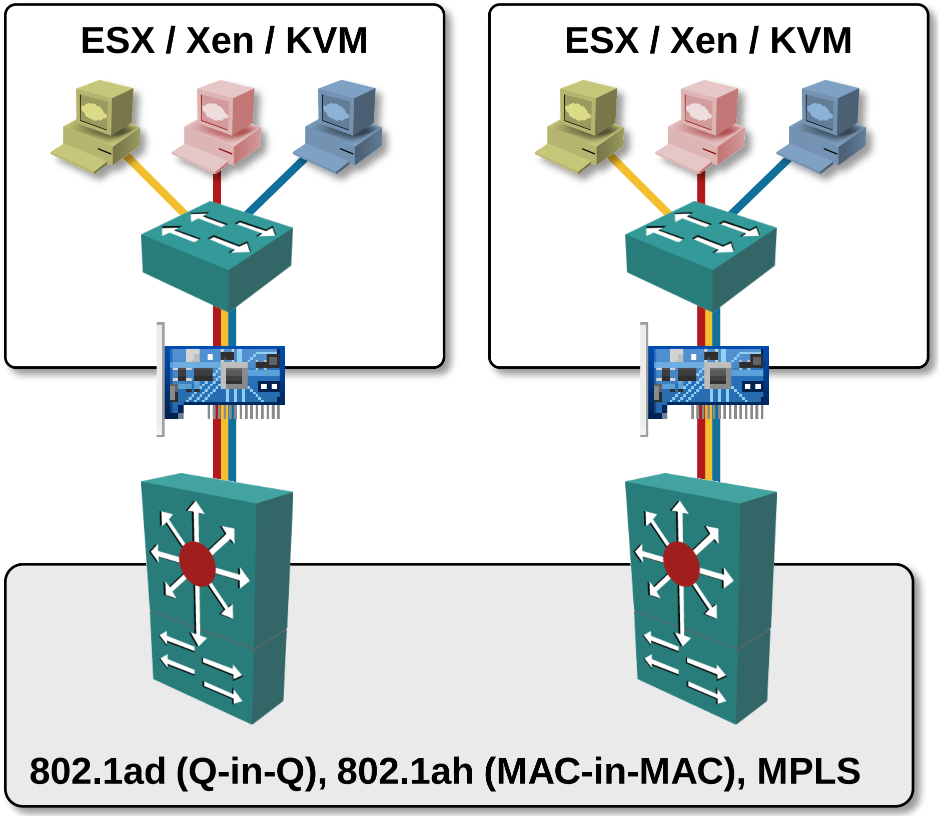
- VLAN-capable layer-2 vSwitch
- Each tenant gets one or more VLANs
- VM NICs are connected to VLANs
- Orchestration with tools like vCloud Director

Challenges

- Scalability (250 VLANs verified on Nexus 5K)
- Tight integration with network infrastructure
- VLAN sprawl
- Large-scale bridging required to support VM mobility



Carrier Ethernet-Based Creative Solutions



Authors:

Kurt Bales – Cisco Metro Ethernet (Q-in-Q)
Derick Winkworth – Juniper MX (Q-in-Q, VPLS)

Basic idea

- Use VLANs on vSwitch as before
- Use Carrier Ethernet (not DC) switches
- Map VLANs into bridging instances
- Transport bridging instances with Q-in-Q, VPLS or MAC-in-MAC

Benefits

- Breaks through the VLAN barriers
- No need for large-scale bridging in the core (VPLS case)
- More controlled bridging with PBB

Drawbacks

- Complex
- Extensive coordination/orchestration

Edge Virtual Bridging (802.1Qbg) Might Actually Help

Technologies

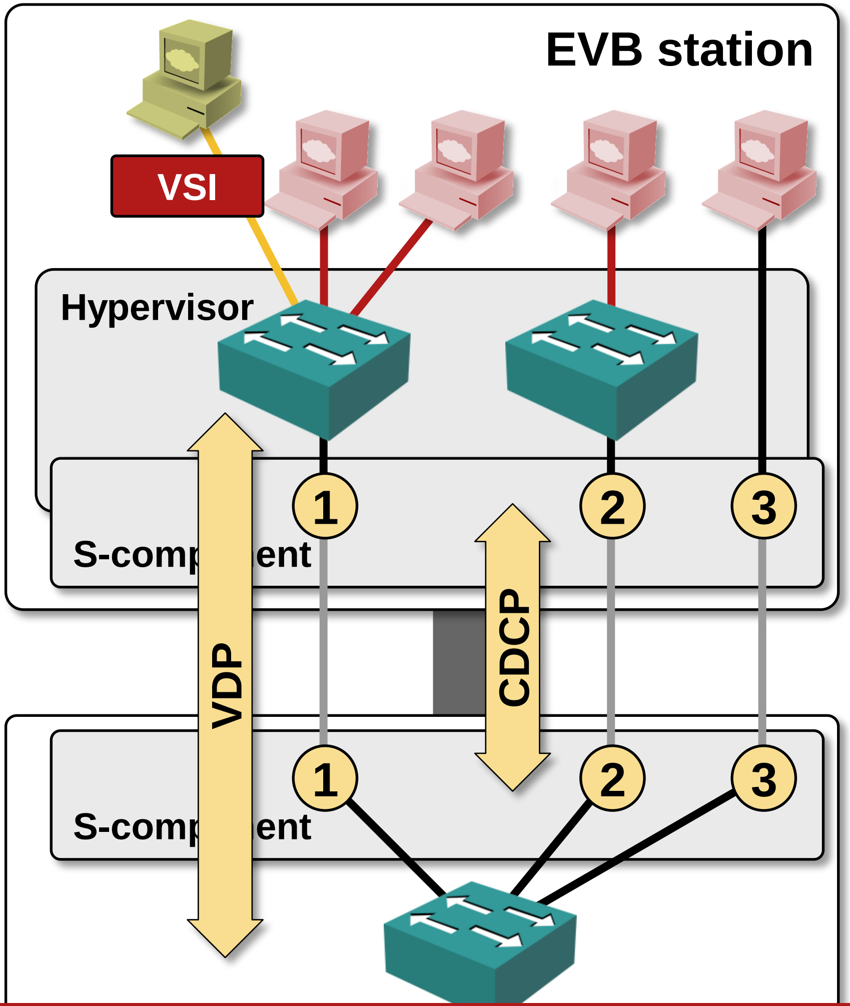
- 802.1Q or 802.1ad (Q-in-Q) tagging
- Outer tag used for virtual links

VSI Discovery Protocol (VDP)

- Hypervisor (EVB station) requests EVB bridge support for new/moved VMs
- Specifies VLAN/GroupID/MAC of the VM
- Breaks through the 12-bit VLAN barrier
- Automatic VLAN provisioning on the access switches (EVB bridge)

S-Channel Discovery and Configuration Protocol

- Creates multiple logical links (S-channels) through the same physical adapter



No vSwitch supports it yet, works in PowerPoint only

vCloud Director Networking Infrastructure (vCDNI)

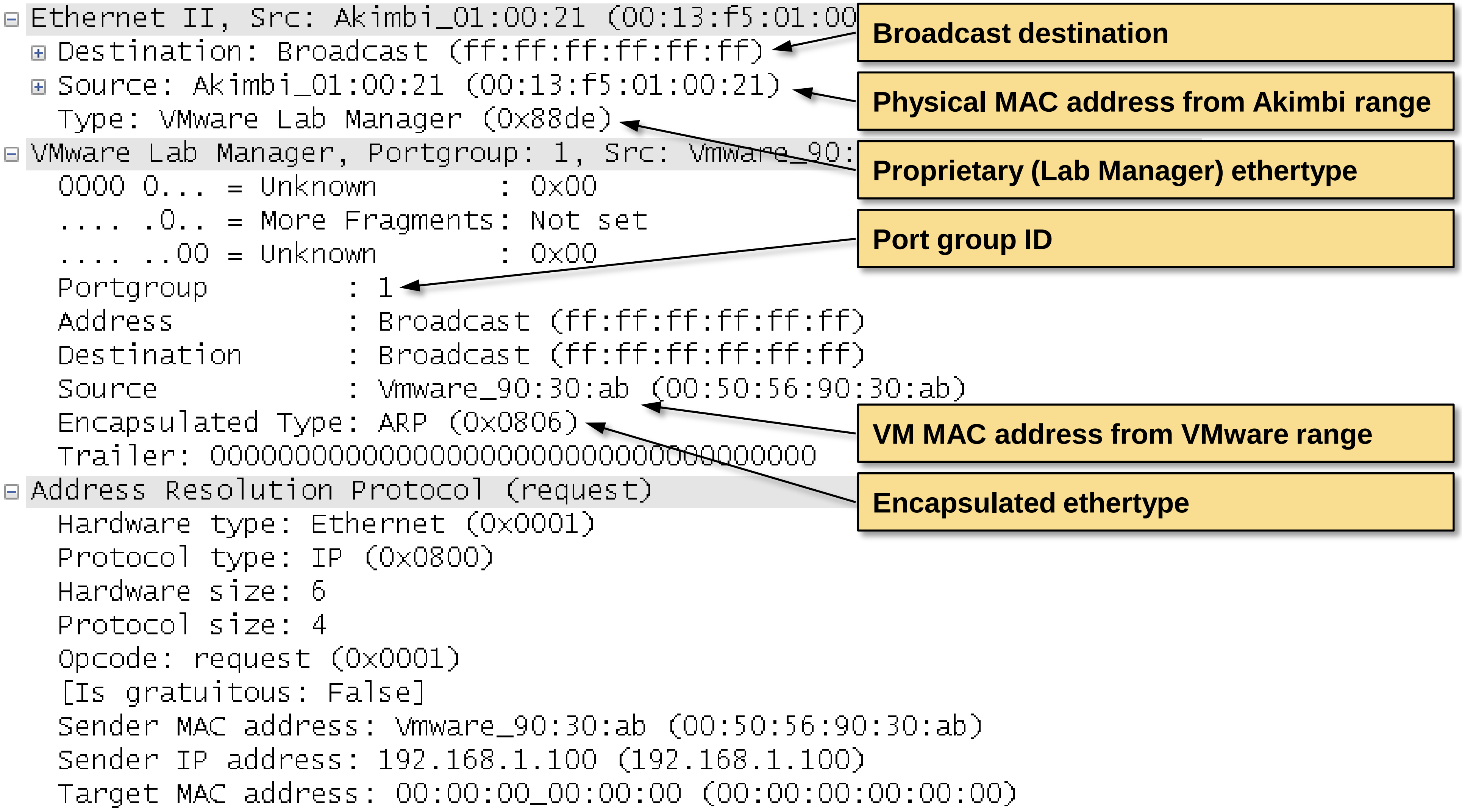
Principle of operations

- Proprietary MAC-in-MAC encapsulation (Port Group Isolation – PGI)
- Port Group ID in the PGI (VMware Lab Manager) header
- MAC frames exchanged between MAC addresses of vSphere hosts
- VM broadcasts/multicasts mapped to physical broadcasts/multicasts
- Dynamic MAC address learning and VM-MAC-to-ESX-MAC discovery

Availability

- vShield Edge (protected port groups only)
- vCloud Director (vCDNI)
- Lab Manager

Sample Wireshark Trace – Broadcast Packets



```
⊟ Ethernet II, Src: Akimbi_01:00:21 (00:13:f5:01:00
  ⊞ Destination: Broadcast (ff:ff:ff:ff:ff:ff)
  ⊞ Source: Akimbi_01:00:21 (00:13:f5:01:00:21)
    Type: VMware Lab Manager (0x88de)
⊟ VMware Lab Manager, Portgroup: 1, Src: Vmware_90:
    0000 0... = Unknown        : 0x00
    .... .0.. = More Fragments: Not set
    .... ..00 = Unknown        : 0x00
    Portgroup           : 1
    Address             : Broadcast (ff:ff:ff:ff:ff:ff)
    Destination         : Broadcast (ff:ff:ff:ff:ff:ff)
    Source              : Vmware_90:30:ab (00:50:56:90:30:ab)
    Encapsulated Type: ARP (0x0806)
    Trailer: 000000000000000000000000000000000000
⊟ Address Resolution Protocol (request)
    Hardware type: Ethernet (0x0001)
    Protocol type: IP (0x0800)
    Hardware size: 6
    Protocol size: 4
    Opcode: request (0x0001)
    [Is gratuitous: False]
    Sender MAC address: Vmware_90:30:ab (00:50:56:90:30:ab)
    Sender IP address: 192.168.1.100 (192.168.1.100)
    Target MAC address: 00:00:00_00:00:00 (00:00:00:00:00:00)
```

Broadcast destination

Physical MAC address from Akimbi range

Proprietary (Lab Manager) ethertype

Port group ID

VM MAC address from VMware range

Encapsulated ethertype

vCDNI Benefits and Drawbacks

Benefits

- Virtual L2 segments created between vSphere hosts without changes in network configuration

Drawbacks

- Relies on bridging (not scalable)
- VM broadcasts mapped into physical broadcasts (not scalable)
- Proprietary: vSphere-to-vSphere only, no network device implementation
- Network must support jumbo frames

Security implications

- Inter-host backbone must be secured (vCDNI does not provide security)
- VM in promiscuous port group can monitor all vCDNI traffic

VXLAN – Yet Another MAC-over-IP Solution

Somewhat more scalable vCDNI

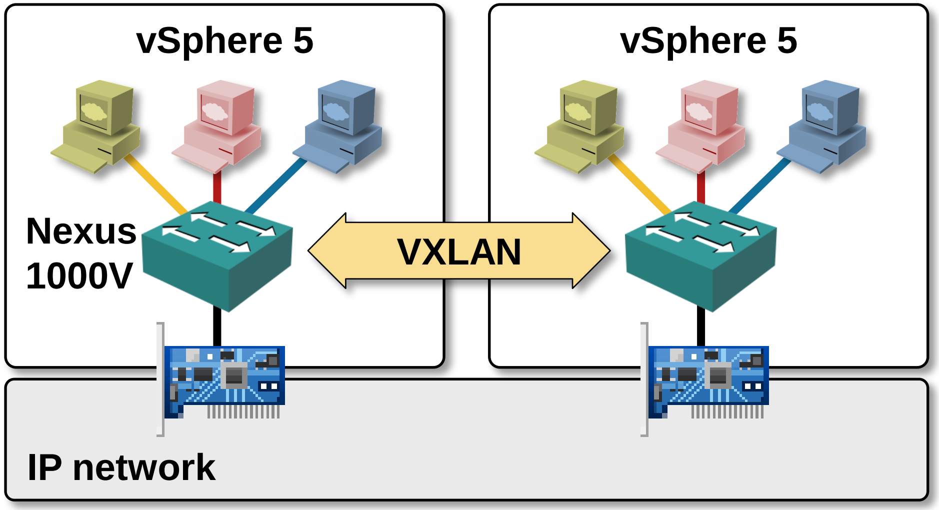
- MAC-in-UDP with 24-bit segment ID
- IP multicast used for L2 flooding
- Very simple encapsulation
- No control plane (dynamic MAC learning)
- No security

Benefits

- Standard. Definitely better than vCDNI
- Somewhat scalable intra-DC L2 solution

Drawbacks

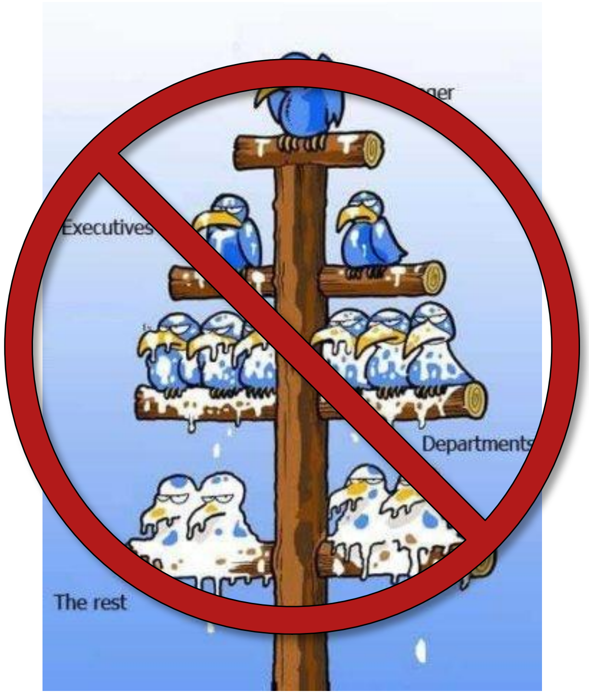
- Requires IP multicast
- No termination on physical devices (yet)
- Inter-subnet traffic goes through layer-3 VM appliances (vShield Edge, Vyatta, BIG-IP)
- No L2 traffic reduction (broadcast limitations, ARP proxy ...)



Parting Thoughts

We are no longer in VoiceLand

- You will either offer cloud services or compete on razor-thin bandwidth margins
- You need to move fast
- Start with the business requirements, not fancy technologies
- Apps/Server/Storage/Network engineers have to work together (DevOps ;)
- Iterative designs involving everyone yield best results
- No place for silos, pyramids and blame-shifting
- Don’t try to be everything to everyone (forget MS NLB)



However

- It’s a new and exciting opportunity
- Truly scalable products are not readily available
- Plenty of room for innovation/creativity

More information

Blogs & Podcasts

- Packet Pushers Podcast & blog (packetpushers.net)
- The Cloudcast (.net)
- Network Heresy (Martin Casado, Nicira)
- BradHedlund.com (Brad Hedlund, Cisco)
- RationalSurvivability.com (Christopher Hoff, Juniper)
- it20.info (Massimo Re Ferre, VMware)
- NetworkJanitor.net (Kurt Bales)
- Yellow bricks (Duncan Epping, VMware)
- ioshints.info (yours truly)

Webinars (@ www.ioshints.info/webinars)

- Introduction to Virtualized Networking & VMware Networking Deep Dive
- Data Center Fabric Architectures (upcoming)
- Data Center 3.0 for Networking Engineers

Questions?