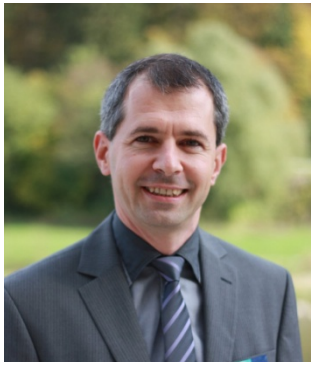
Cloud Networking – From Theory to Practice

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

Who is Ivan Pepelnjak ... in 30 Seconds



- 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 (blog.ioshints.info), book author

Focus:

- Core routing/MPLS, IPv6, VPN, Data centers, Virtualization

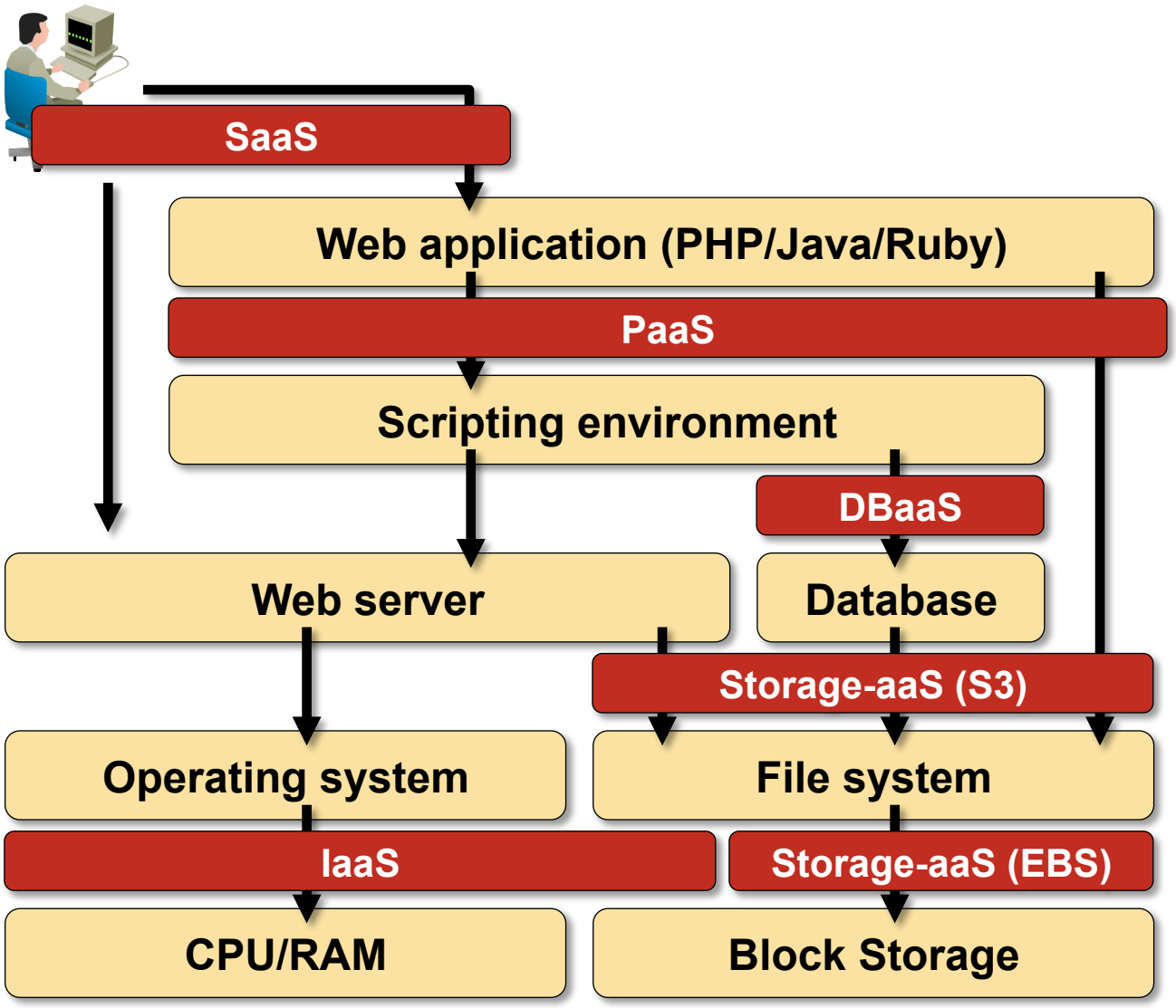
© ipSpace.net / NIL Data Communications 2012

Disclaimers

- This presentation is an analysis of currently available virtual networking architectures
- It's not an endorsement or bashing of companies, solutions or products mentioned on the following slides
- It describes *features* not *futures*
- The crucial question: Does It Scale?

© ipSpace.net / NIL Data Communications 2012

Cloud Services Taxonomy 101



- **IaaS** is most interesting for networking engineers
- All others are just TCP/IP applications - we know how to do that

What's different?

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

Key ingredients

- Scalability
- Orchestration
- Customer-driven deployment

© ipSpace.net / NIL Data Communications 2012

What Type of IaaS Service Do You Offer?

Business decisions:

- What is your added value?
- What is your differentiator from Amazon and Rackspace?
- Will you focus on enterprise apps or new-world (scale-out) apps?
- Will you be low-cost or feature-rich?

Technical questions:

- Simple compute capacity or full-blown virtual private networks?
- TCP or UDP cloud?
- IP Multicast support?

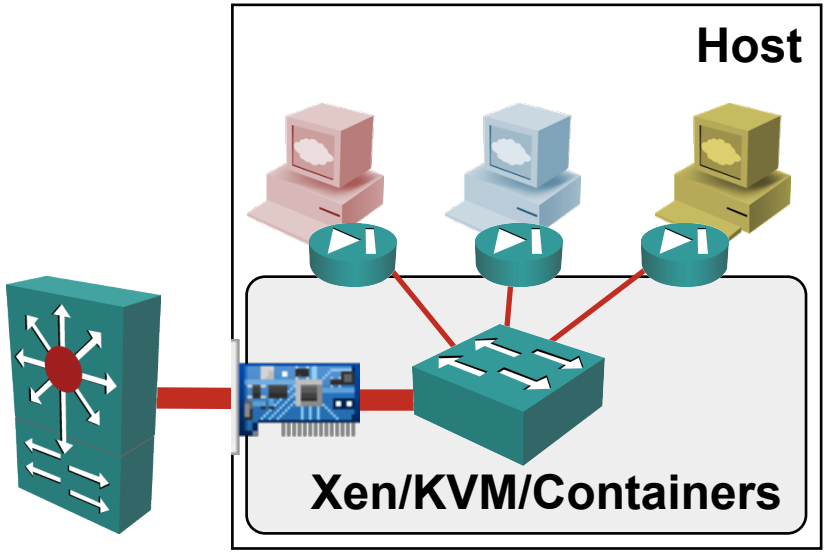
IaaS Lite: Multi-Tenant Isolation With Firewalls

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)
- Predefined configurations or user-controlled firewalls

Multi-tenant isolation options

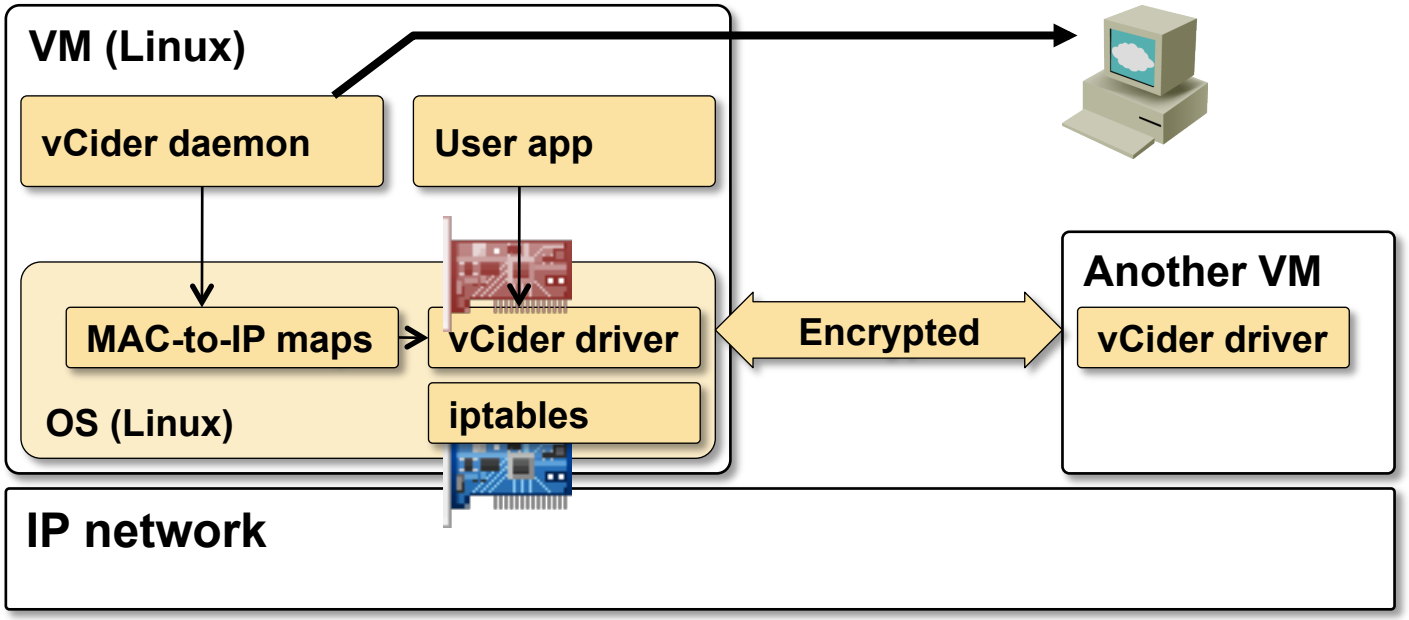
- Packet filters (example: iptables) applied to VM interfaces (XenServer/KVM)
- Private VLANs implemented in vSwitch (VMware VDS, Nexus 1000V)
- Virtual firewalls (VMware vShield App, Juniper VGW)
- Virtual firewalls with service insertion (Nexus 1000V + VSG)



Scalability: unlimited (see also: *Internet*)

© ipSpace.net / NIL Data Communications 2012

Sample Over-the-Cloud Virtual Network: vCider



- VM-based 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

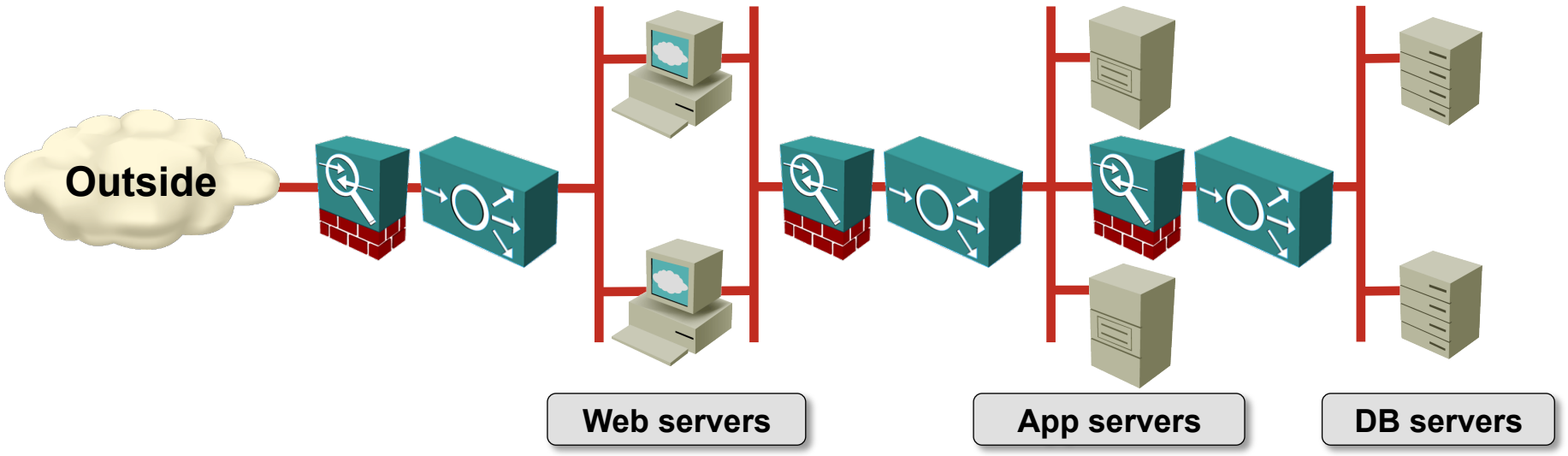
Drawbacks:

- Linux only
- Requires VM changes (device driver)

Alternative: CloudSwitch (nested hypervisor on Amazon EC2)

© ipSpace.net / NIL Data Communications 2012

Virtual Segments: Typical Customer Requirements



Requirements

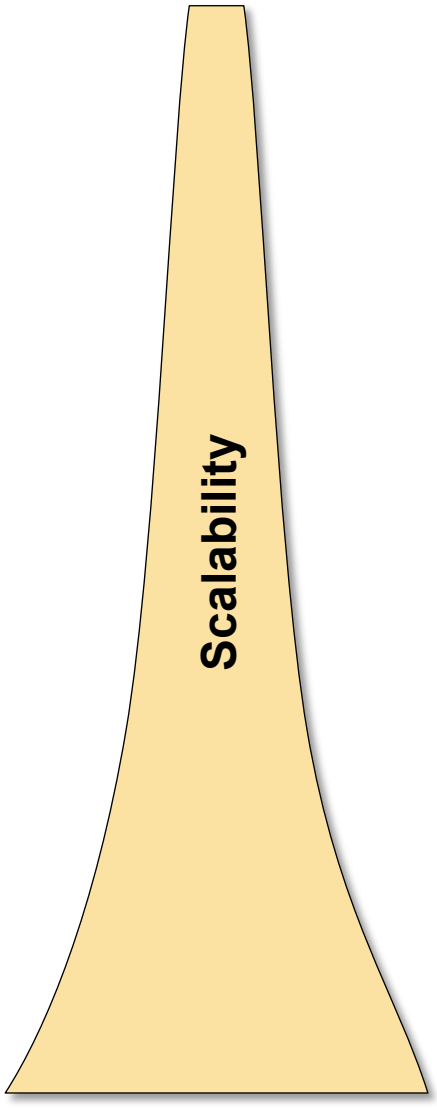
- Multiple logical segments
- Routing, load balancing or firewalling between segments
- Usually one NIC per VM
- Unlimited scalability and mobility

Implementation decisions

- VM mobility?
- L2 or L3 segments?
- Support for IP MC and L2 flooding?
- Virtual or physical appliances (LB, FW)?

© ipSpace.net / NIL Data Communications 2012

Solution Space and Scalability



VLANs — **4096 segments**

VM-aware Networking (Arista VM Tracer)

Edge Virtual Bridging (EVB, 802.1Qbg) — **Emerging**

vCDNI – VMware (L2 over L2)

EVB with PBB/SPB (L2 over L2) — **Theoretical**

VXLAN (Cisco) / NVGRE (Microsoft)
L2 over IP — **No control plane**

Nicira NVP (L2 over IP + Control Plane)

Amazon EC2 (IP over IP + Control Plane)

© ipSpace.net / NIL Data Communications 2012

Architectural Models

Stupid edge (VLAN-aware vSwitch) + Stupid core

- Traditional VLAN model

Stupid edge + Smart core

- VM-aware networking, EVB

With sufficient thrust, pigs fly just fine **RFC 1925**

Can we afford the fuel costs ... And who wants to fly pigs anyway?
Randy Bush

Smart edge + simple core

- vCDNI (L2 core), VXLAN, NVGRE, Nicira NVP, Amazon (L3 core)

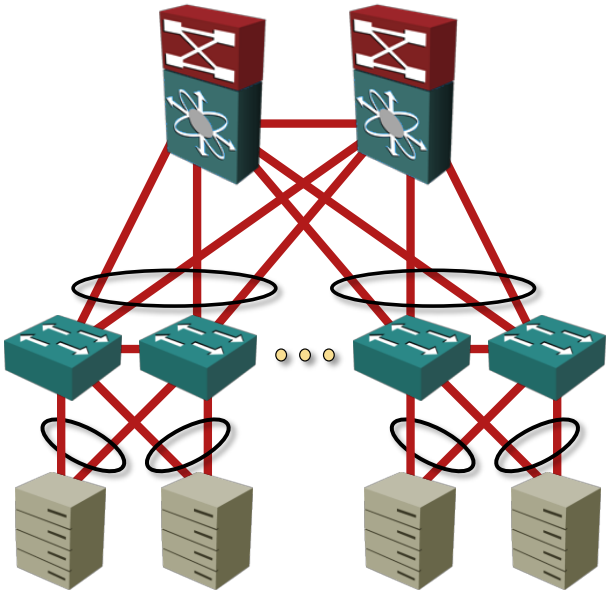
End-to-end protocol design should not rely on the maintenance of state inside the network **RFC 3439**

© ipSpace.net / NIL Data Communications 2012

Layer-2 Transport Doesn't Scale

Large-scale Layer-2 Switching Solutions:

- Clos fabric with two core switches and multi-chassis link aggregation – Arista (~ 1900 ports)
- QFabric – Juniper (~ 6000 ports)
- FabricPath – Cisco (~ 18000 ports)

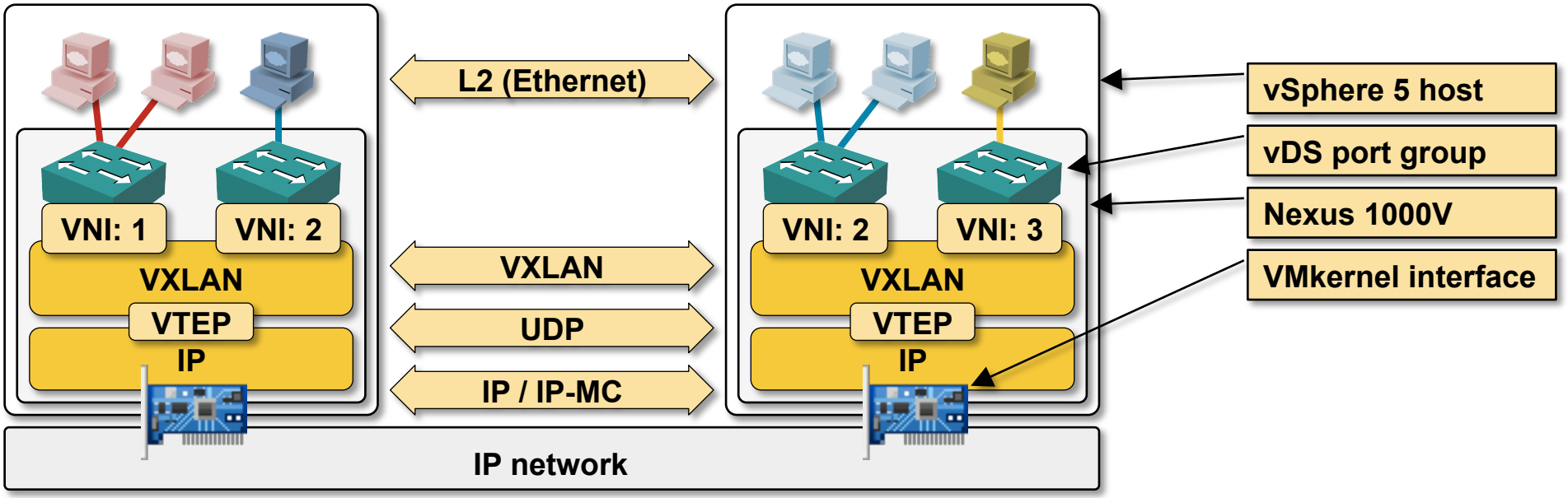


Reality checks:

- VMware vDS supports 300 servers
- Cisco's Nexus 1000V supports 64 servers

You can run away from Spanning Tree, but broadcasts will eventually kill you ... Not to mention that L2 network is a single failure domain

VXLAN/NVGRE: You Can't Scale w/o Control Plane



- Virtual layer-2 segments (VXLAN segments) over L3 transport infrastructure
- UDP-based encapsulation similar to OTV/LISP with 24-bit segment ID (VNI)
- IP multicast used for L2 flooding (dynamic MAC learning)

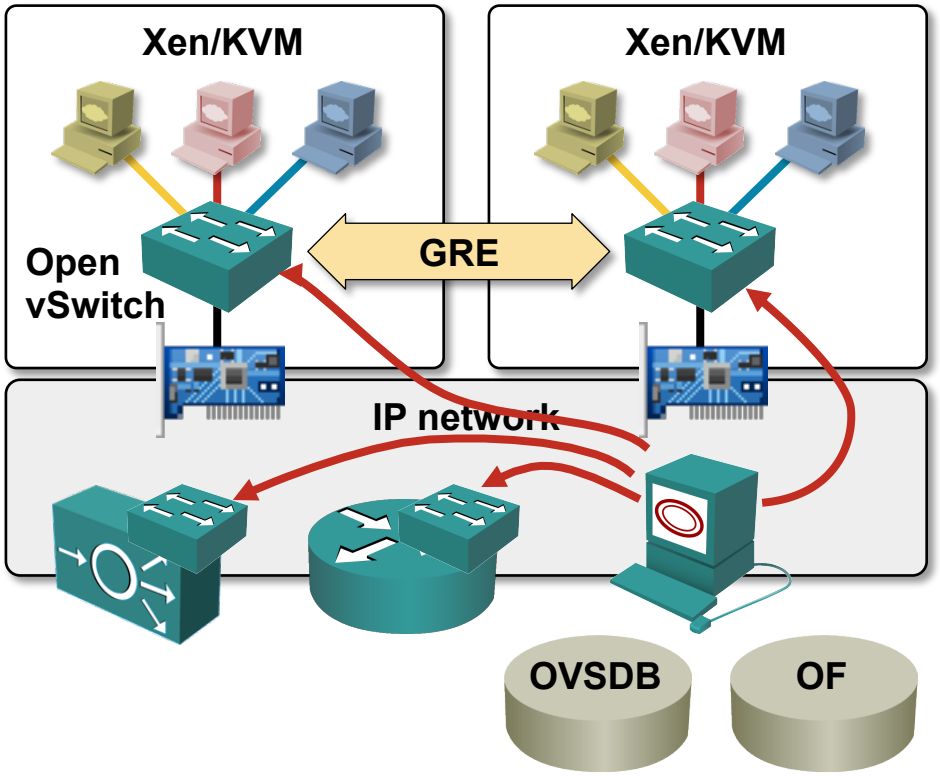
Large "broadcast domains" or enormous amount of (*,G) and (S,G) state
Dynamic MAC learning through flooding *does not scale*

© ipSpace.net / NIL Data Communications 2012

Open vSwitch With Nicira NVP (OpenFlow)

MAC-over-IP with control plane

- OpenFlow-capable vSwitches (OVS)
- P2P GRE tunnels provisioned with OVSDB
- MAC-to-IP mapping downloaded to OVS with OpenFlow
- Third-party physical devices with OVS



Benefits

- Proper control plane
- No reliance on flooding
- No IP multicast in the core

Drawbacks

- L2 flooding within the virtual subnets (ARP proxy?)

© ipSpace.net / NIL Data Communications 2012

Rule-of-Thumb Guidelines

Few hundred tenants, few hundred servers → VLANs

Thousands of tenants, few hundred servers → vCDNI or Q-in-Q

Hundreds of tenants, few thousand servers → VM-aware networking

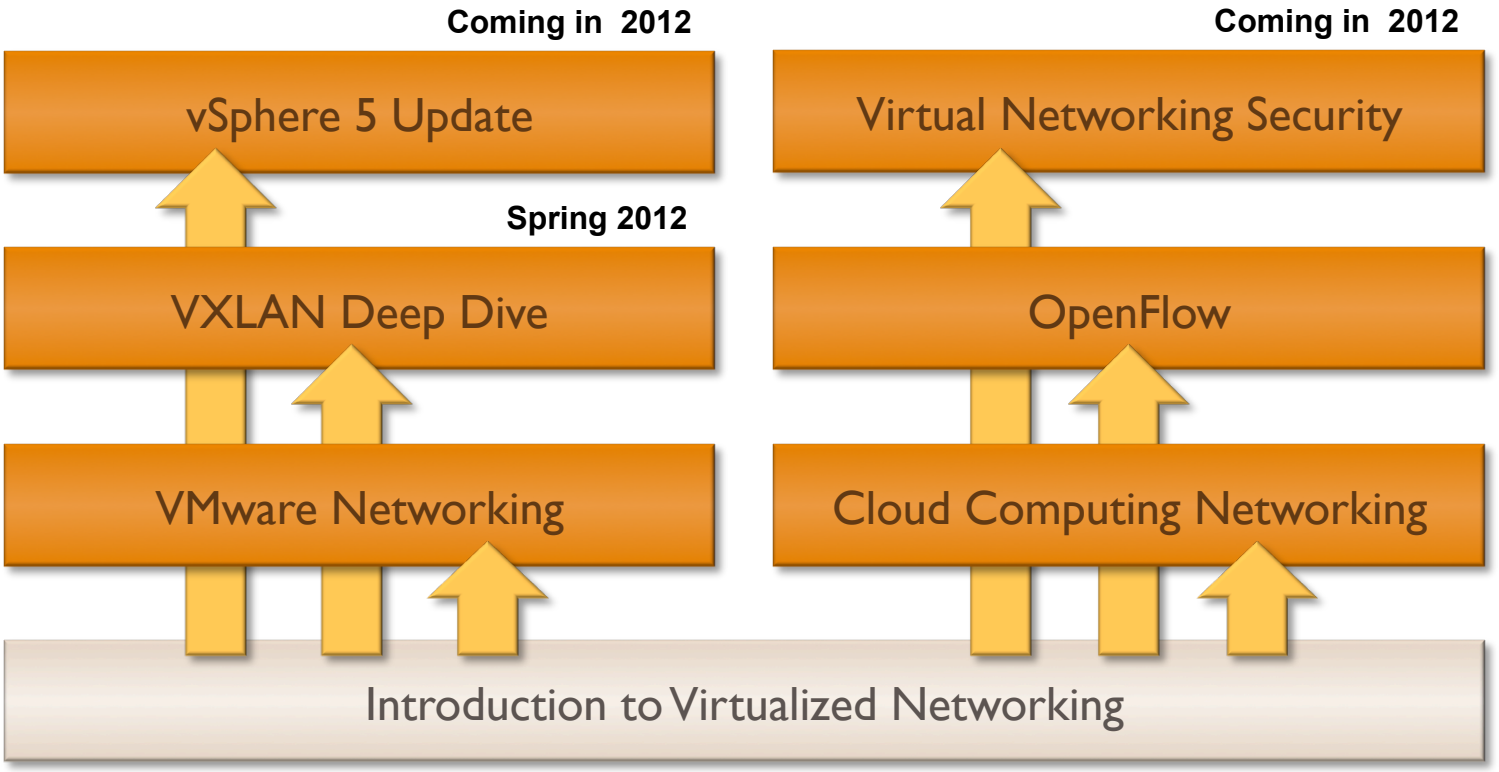
Few thousand servers, thousands of tenants → VXLAN / NVGRE

More than that → L2 over IP with control plane

You can scale low-end solutions by splitting your DC in availability zones

© ipSpace.net / NIL Data Communications 2012

More information: Virtualization Webinars



Availability

- Live sessions
- Recordings of individual webinars
- Yearly subscription

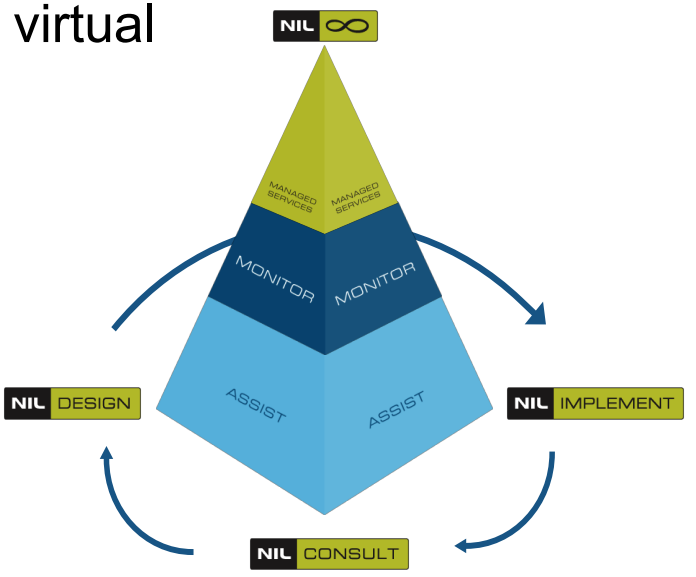
Other options

- Customized webinars
- ExpertExpress
- On-site workshops

More information @ http://www.ioshints.info/Webinars

First Steps

- Start with the business requirements and service definitions
- Build-or-buy decision
- Select the automation/orchestration tools
- Orchestration tool might dictate hypervisors and/or virtual networking technologies
- Design the network



Need help?

- ExpertExpress for quick discussions, reviews or second opinions

NIL's Professional/Learning Services

- In-depth design/deployment projects
- Cloud-related training
- Details: www.nil.com, flipit.nil.com

© ipSpace.net / NIL Data Communications 2012

Questions?