Cloud Networking – From Theory to Practice

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

Who is Ivan Pepelnjak (@ioshints)

- Networking engineer since 1985
- Consultant, blogger (blog.ioshints.info), book and webinar author
- Currently teaching “Scalable Web Application Design” at University of Ljubljana

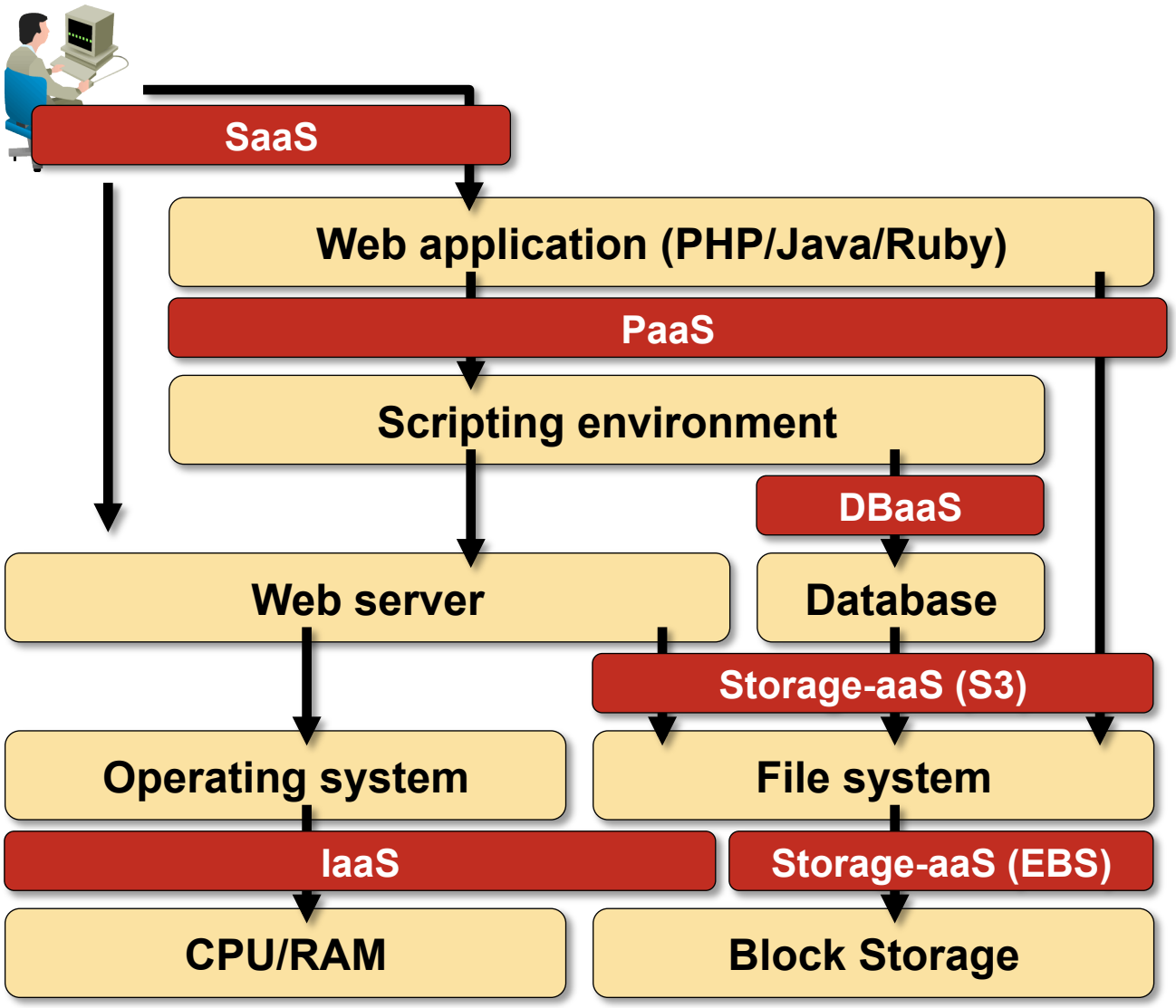
Focus:

- Large-scale data centers and network virtualization
- Networking solutions for cloud computing
- Scalable application design
- Core IP routing/MPLS, IPv6, VPN

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?

Cloud Services Taxonomy 101



Interesting: **IaaS**

Others run over TCP

Key ingredients

- Scalability
- Orchestration
- On-demand

What IaaS Service Will You Offer?

What is your added value?

- Differentiator from Amazon and Rackspace?
- Enterprise apps or new-world (scale-out) apps?
- Low-cost or feature-rich?

Technical questions:

- Simple compute capacity or app stack support?
- TCP or UDP cloud?
- IP Multicast support?

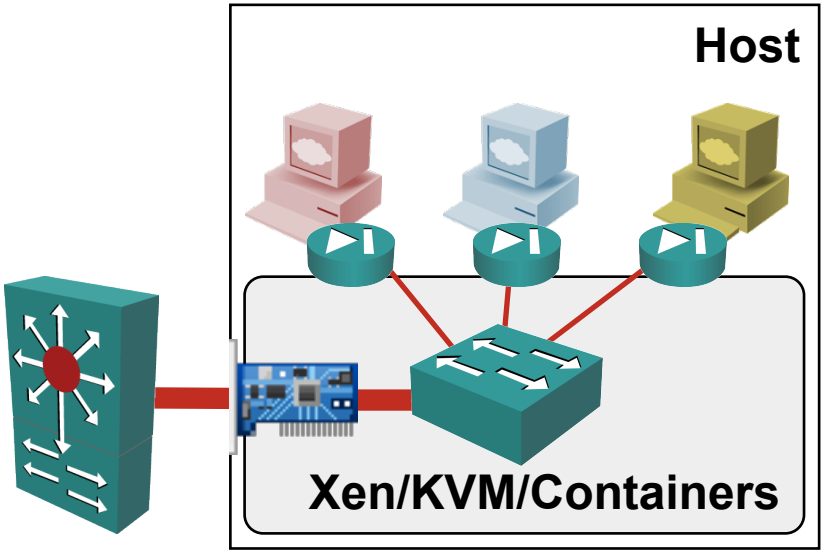
IaaS Lite: Multi-Tenant Isolation

Making life easier for the cloud provider

- Customer VMs attached to “random” L3 subnets
- VM IP addresses allocated by the IaaS provider
- Predefined configurations or user-controlled firewalls

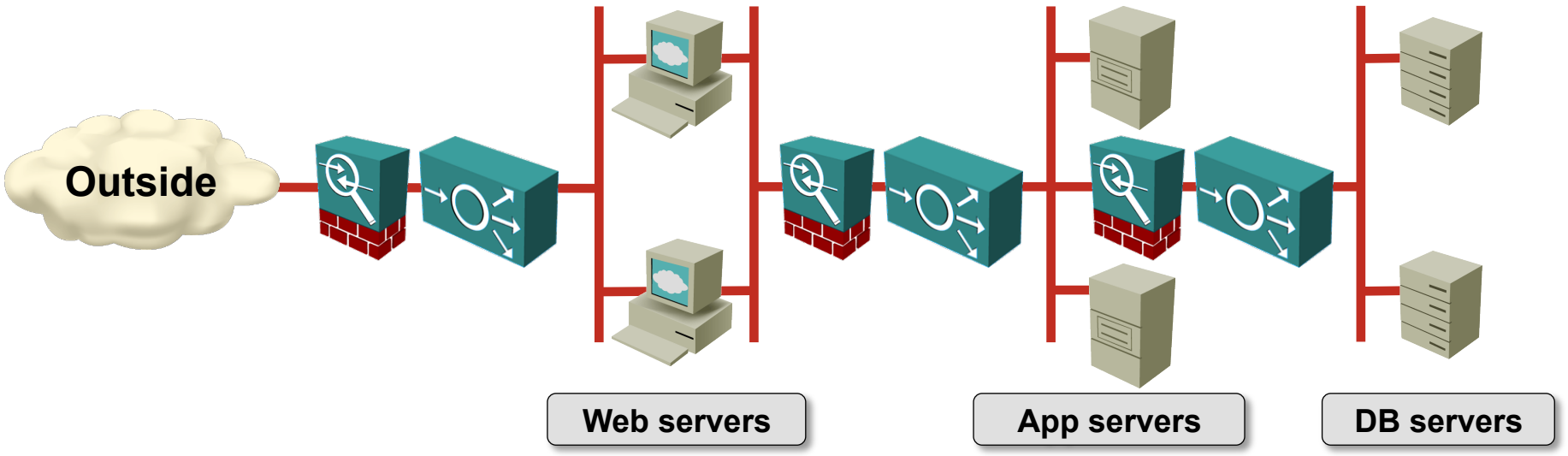
Multi-tenant isolation options

- Packet filters (ex: iptables)
- Private VLANs in vSwitch
- Virtual firewalls



Scalability: unlimited (see also: *Internet*)

What Customers Want



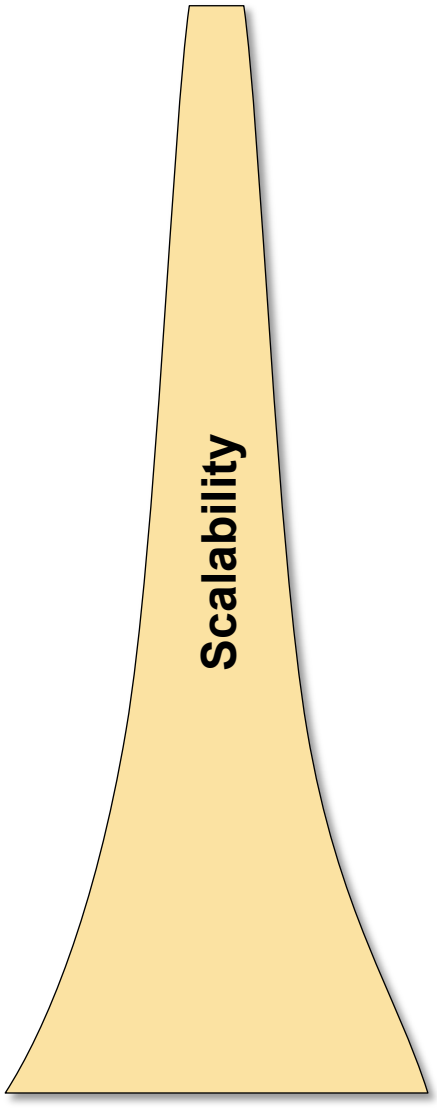
Requirements

- Multiple logical segments
- Load balancing and firewalling
- 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?

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)

Architectural Models

VLANs: Stupid edge + Stupid core

Stupid edge + Smart core

- VM-aware networking, EVB (802.1Qbg)

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

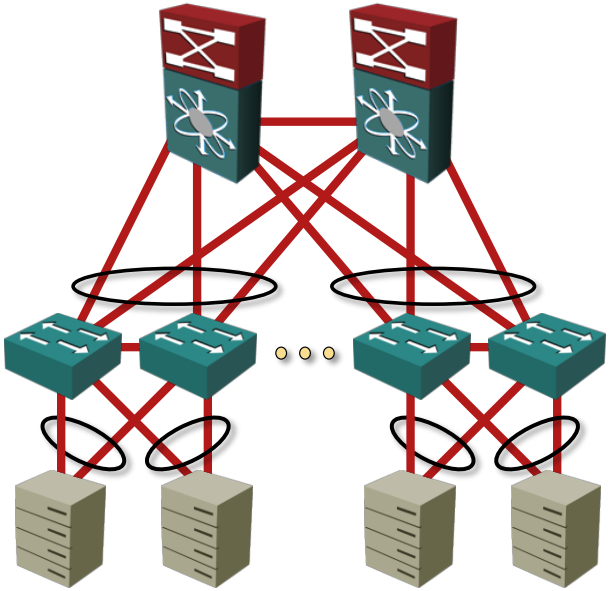
- vCDNI (L2 core), VXLAN, NVGRE, Nicira NVP, Amazon

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

VLANs: Bridging Has Failed Before

Your vendor has a solution:

- Two core switches and MLAG aggregation: ~ 1900 ports (Arista)
- QFabric – Juniper (~ 6000 ports)
- FabricPath – Cisco (over 10K ports)

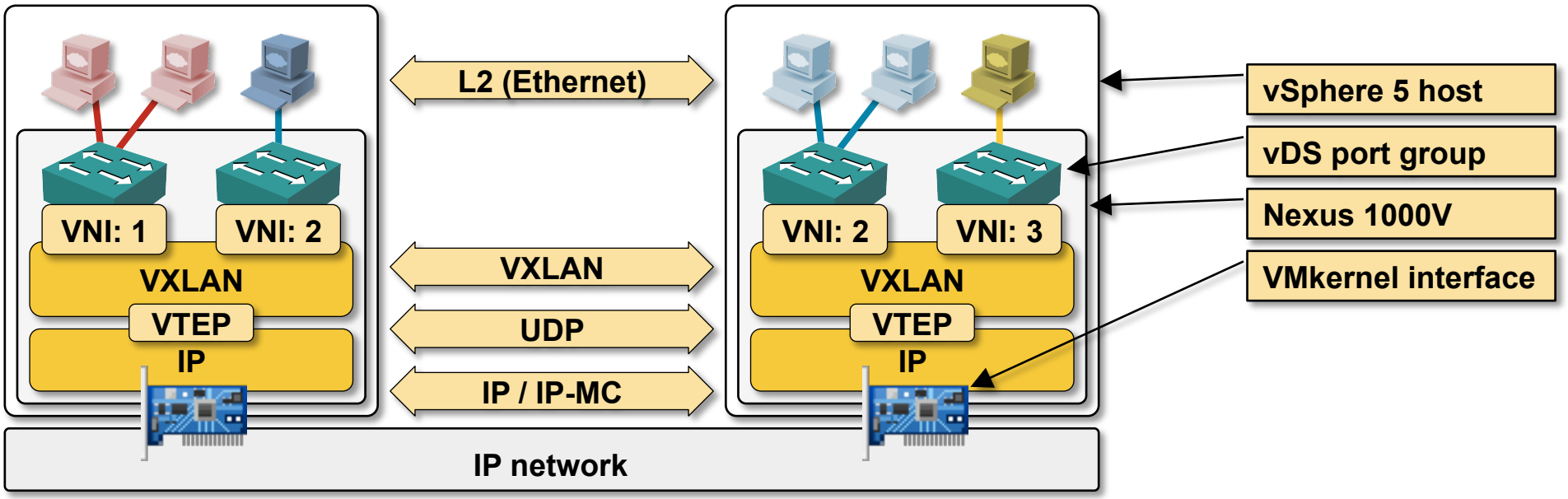


Reality checks:

- VMware vDS supports 350 hosts (Nexus 1000V: 64)
- We still have only 4K VLANs
- L2 network = single failure domain

You can run away from Spanning Tree, but broadcasts will eventually kill you

VXLAN/NVGRE: Where Is Control Plane?



- Virtual L2 segments over L3 transport
- UDP/LISP- or GRE-based encapsulation
- Dynamic MAC learning with L2 flooding over IPMC

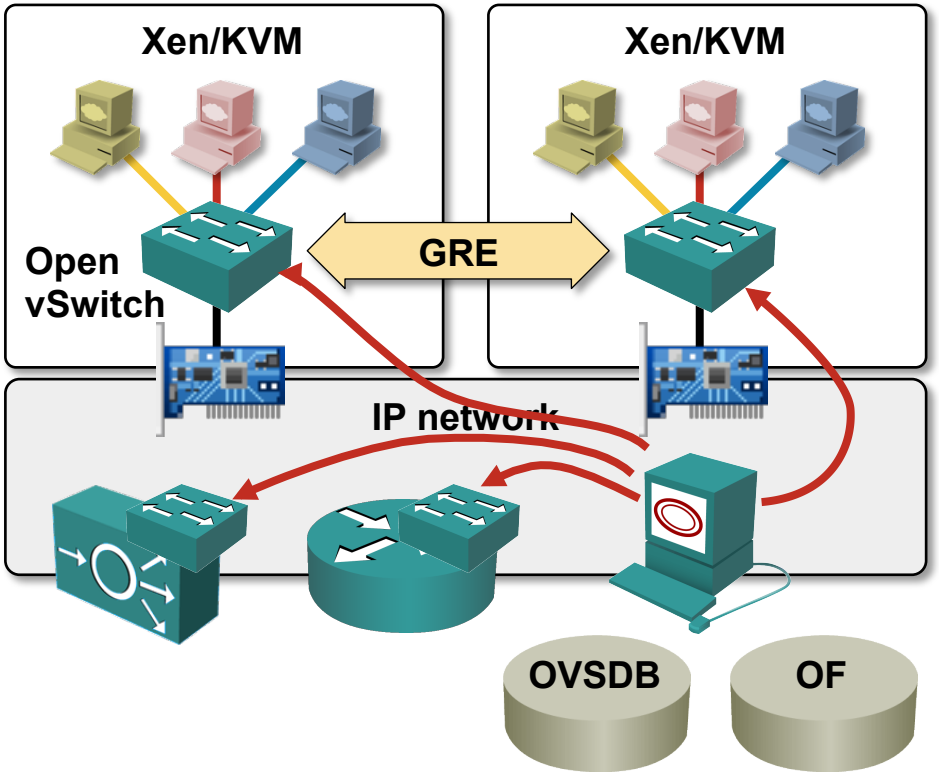
Large "broadcast domains" or enormous amount of (*,G) and (S,G) state

Dynamic MAC learning through flooding *does not scale*

Open vSwitch With Nicira NVP (OpenFlow)

L2-over-IP with control plane

- OpenFlow-capable vSwitches
- IP tunnels (GRE, STT ...)
- MAC-to-IP mappings downloaded with OpenFlow
- Third-party physical devices



Benefits

- No reliance on flooding
- No IP multicast in the core

Open questions

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

Rule-of-Thumb Guidelines

100s tenants, 100s servers → VLANs

1000s tenants, 100s servers → vCDNI or Q-in-Q

Few tenants per server → VM-aware networking

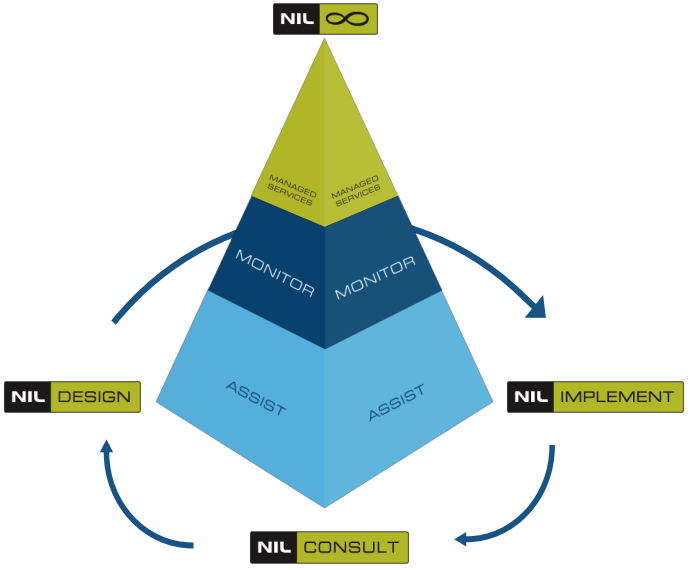
Few 1000s servers, many tenants → VXLAN / NVGRE

More than that → L2 over IP with control plane

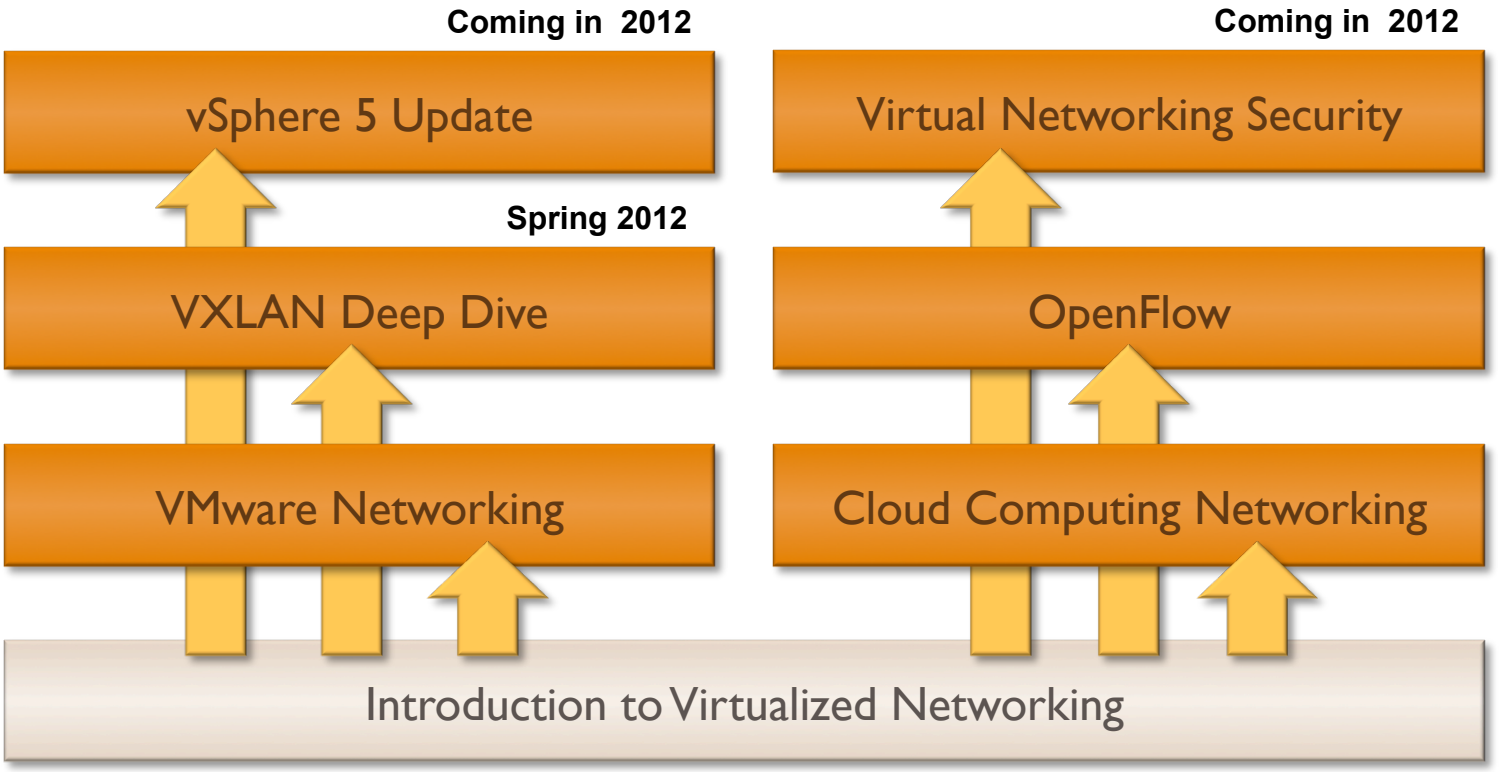
Scale low-end solutions by splitting DC in availability zones

First Steps

- Start: Business requirements and service definitions
- Build-or-buy decision
- Select the orchestration tools ➔ might dictate hypervisors and networking technologies
- Finally: Design the network
- First time: Get help



Reference: Virtualization Webinars



Availability

- Live sessions
- Recordings of individual webinars
- Yearly subscription

Other options

- Customized webinars
- ExpertExpress
- On-site workshops

More information @ http://www.ipspace.net/Webinars

Reference: Blogs and Podcasts

- Packet Pushers Podcast & blog (packetpushers.net)
- The Cloudcast (.net)
- Network Heresy (Martin Casado, Nicira)
- RationalSurvivability.com (Christopher Hoff, Juniper)
- High Scalability Blog
- it20.info (Massimo Re Ferre, VMware)
- NetworkJanitor.net (Kurt Bales)
- BradHedlund.com (Brad Hedlund, Dell Force 10)
- Yellow bricks (Duncan Epping, VMware)
- Twilight in the Valley of the Nerds (Brad Casemore)
- blog.ioshints.info & ipspace.net (yours truly)

Questions?