Server Guy's Guide to Virtual Networks

Ivan Pepelnjak (@ioshints)

CCIE#1354 Emeritus
NIL Data Communications

Server view of the network ... before we were “blessed” with virtualization

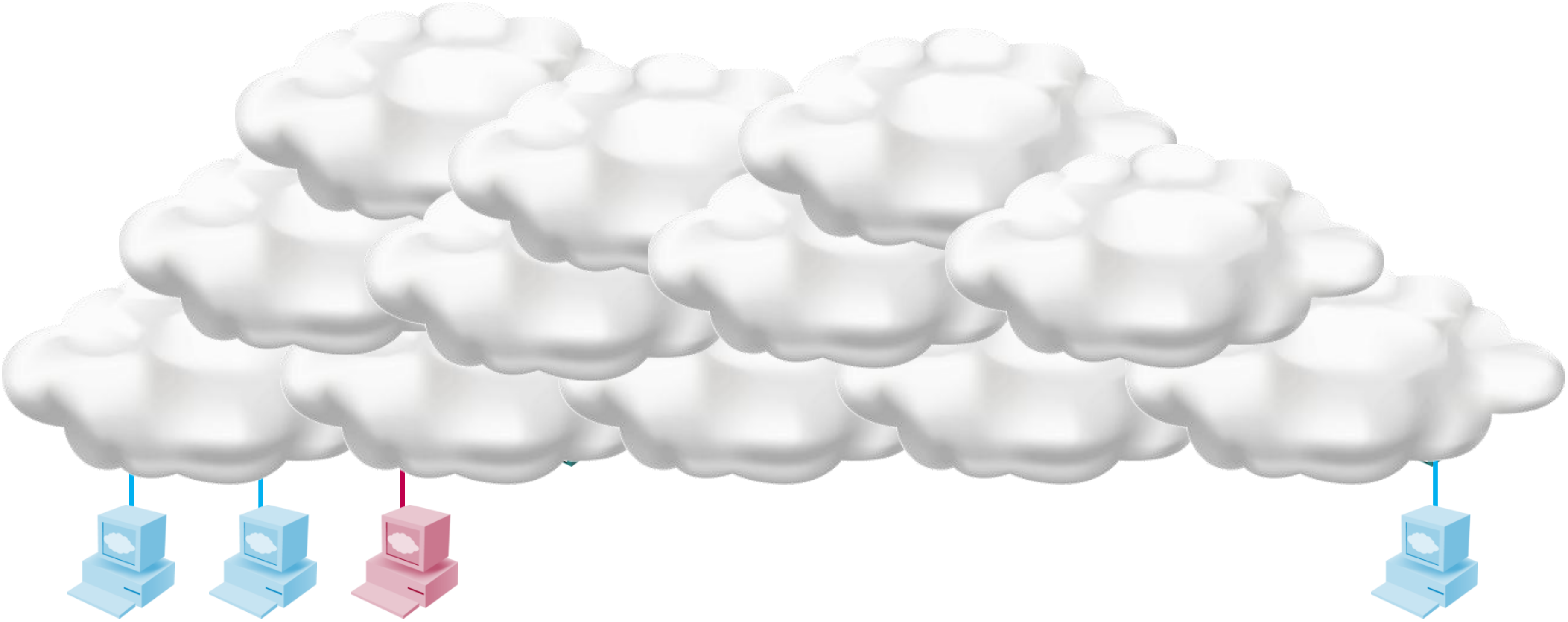
Who is Ivan Pepelnjak (@ioshints)

- Networking engineer since 1985
- Focus: real-life deployment of advanced technologies
- Chief Technology Advisor @ NIL Data Communications
- Consultant, blogger (blog.ipspace.net), book and webinar author (www.ipspace.net)
- Teaching “Scalable Web Application Design” at University of Ljubljana

Current interests:

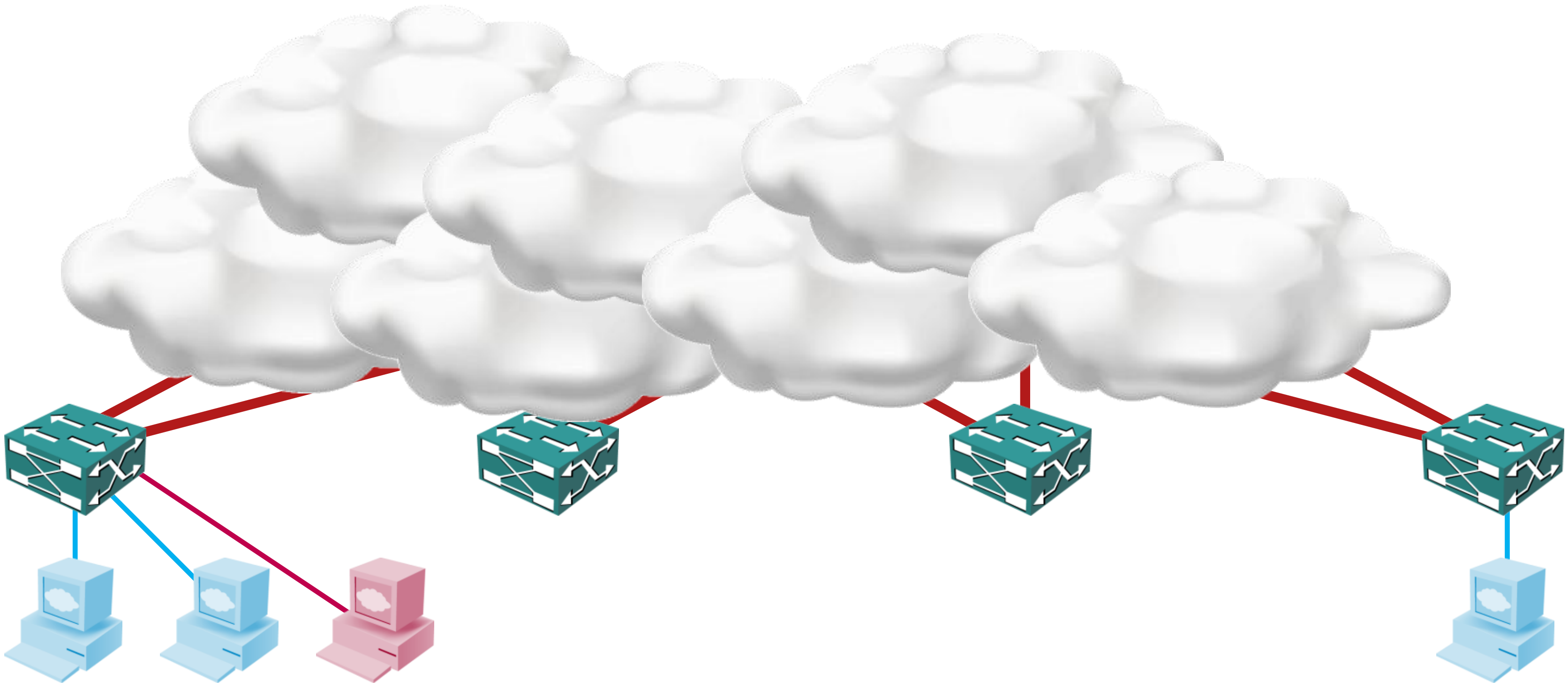
- Large-scale data centers and network virtualization
- Networking solutions for cloud computing
- Scalable application design
- Core IP routing/MPLS, IPv6, VPN

The Good Old World



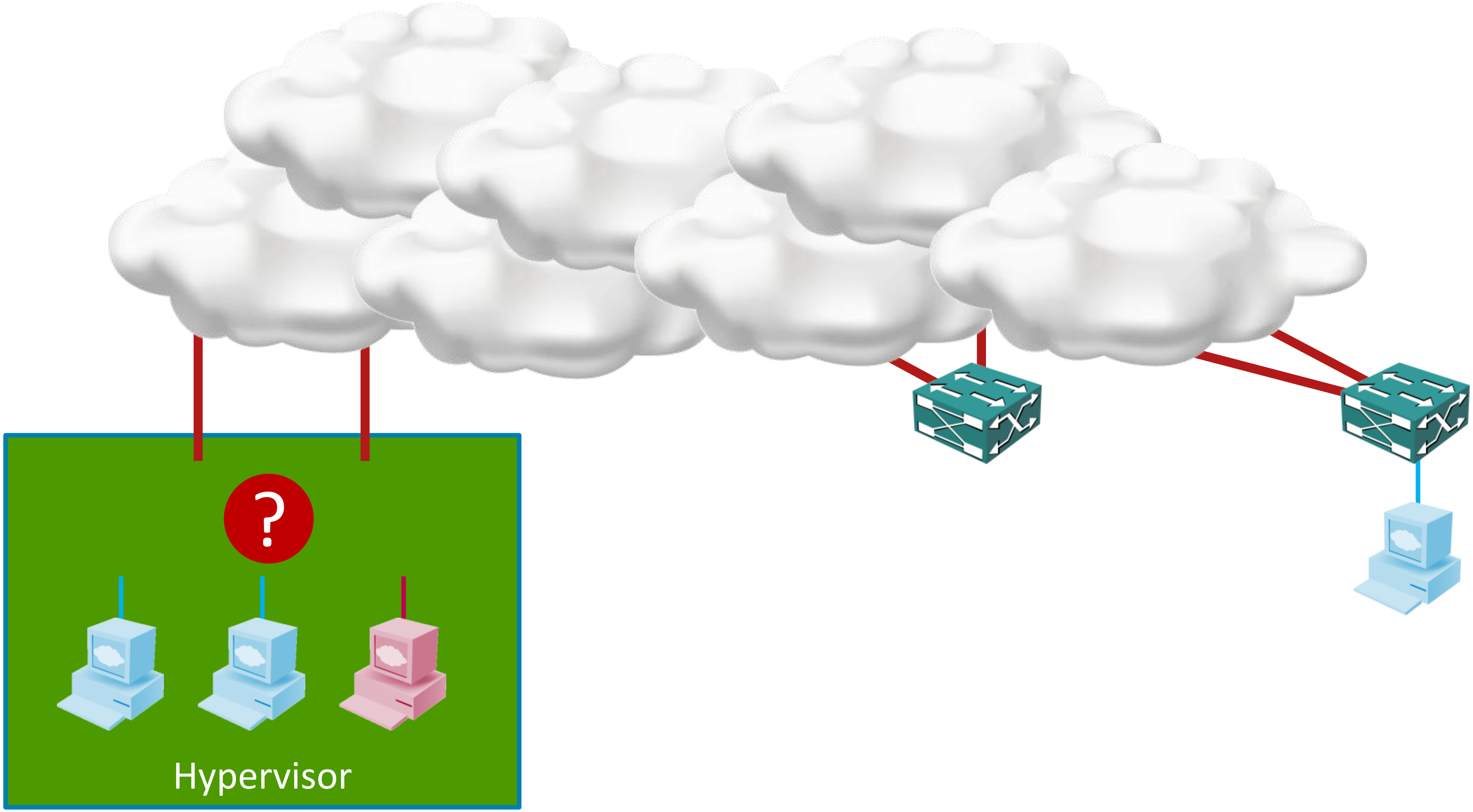
Who cares what’s inside the clouds … as long as it works.

OK, Maybe We'll Peek In

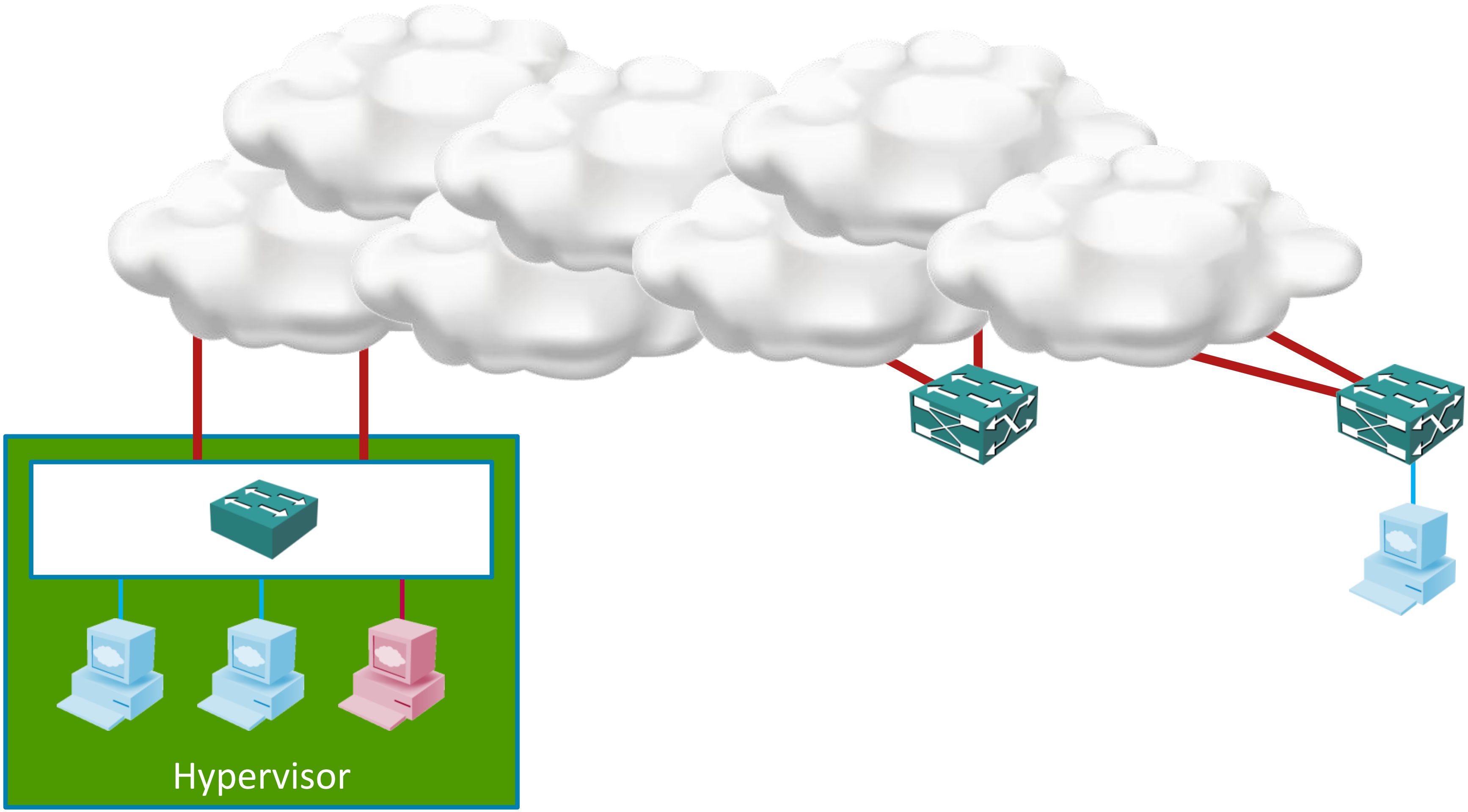


Ah, those things are called "switches". Nice to know

Welcome Server Virtualization



Welcome Server Virtualization



There’s a switch in my hypervisor!

The Usual Response

- Denial – I don't need to know about it
- Anger – Why do I have to deal with networking?
- Bargaining – Maybe I could figure things out with Google/Bing
- Depression – I don't get it. I don't want to know about networking.
- Acceptance – OK, let's talk with the networking team

But wait, there's more …

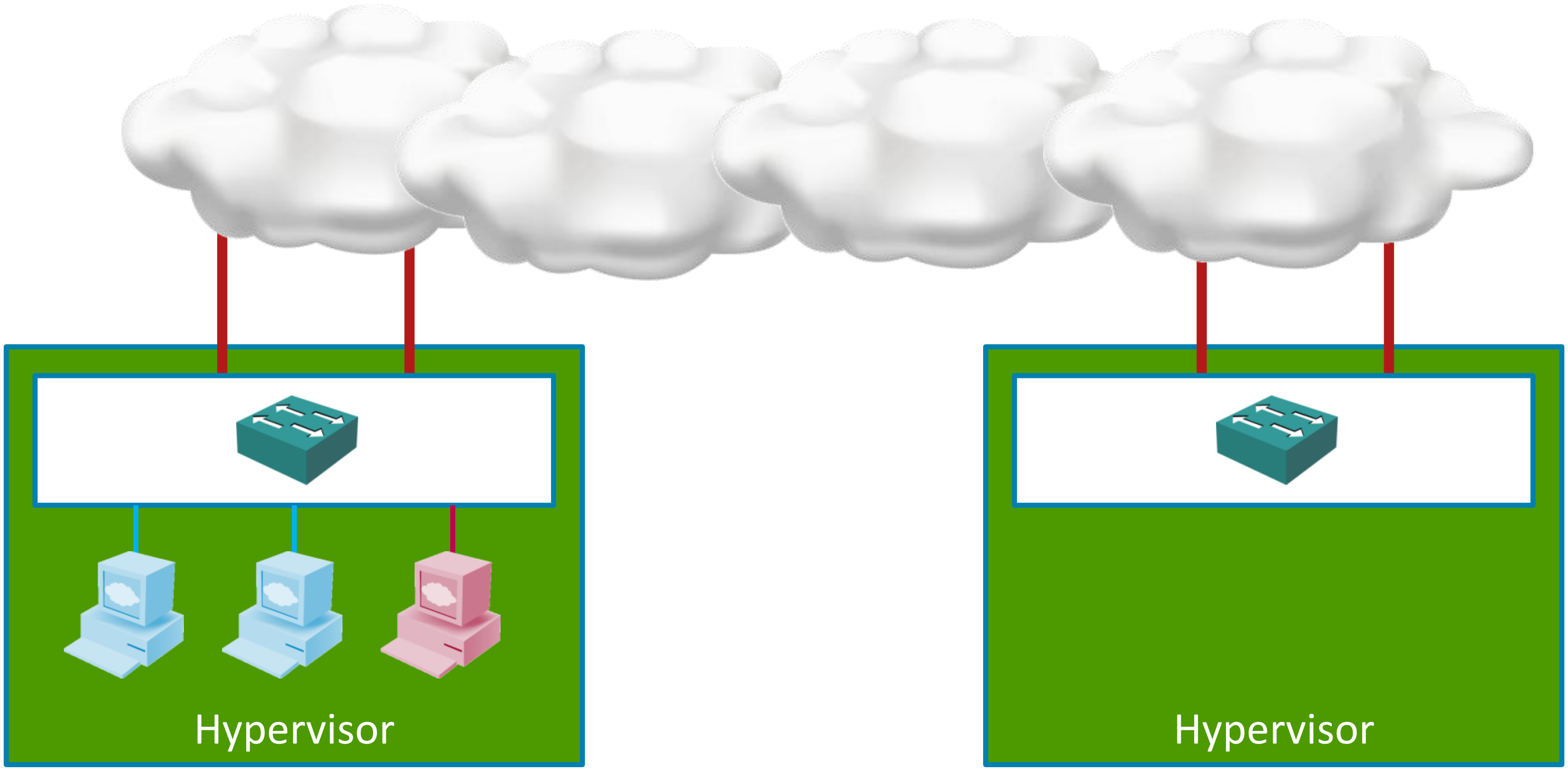
- Hypervisor switches are exceedingly simple
- They lack the basic features we need in secure & stable networks
- They use different terminology and configuration/management mechanisms than physical switches
- Who will manage the virtual switches?

Recommendations

- Talk with the networking team
- Figure out a way to get what you need while keeping the network stable
- Option: Use third-party enterprise-grade virtual switches (Cisco Nexus 1000V, IBM Distributed Virtual Switch 5000V)

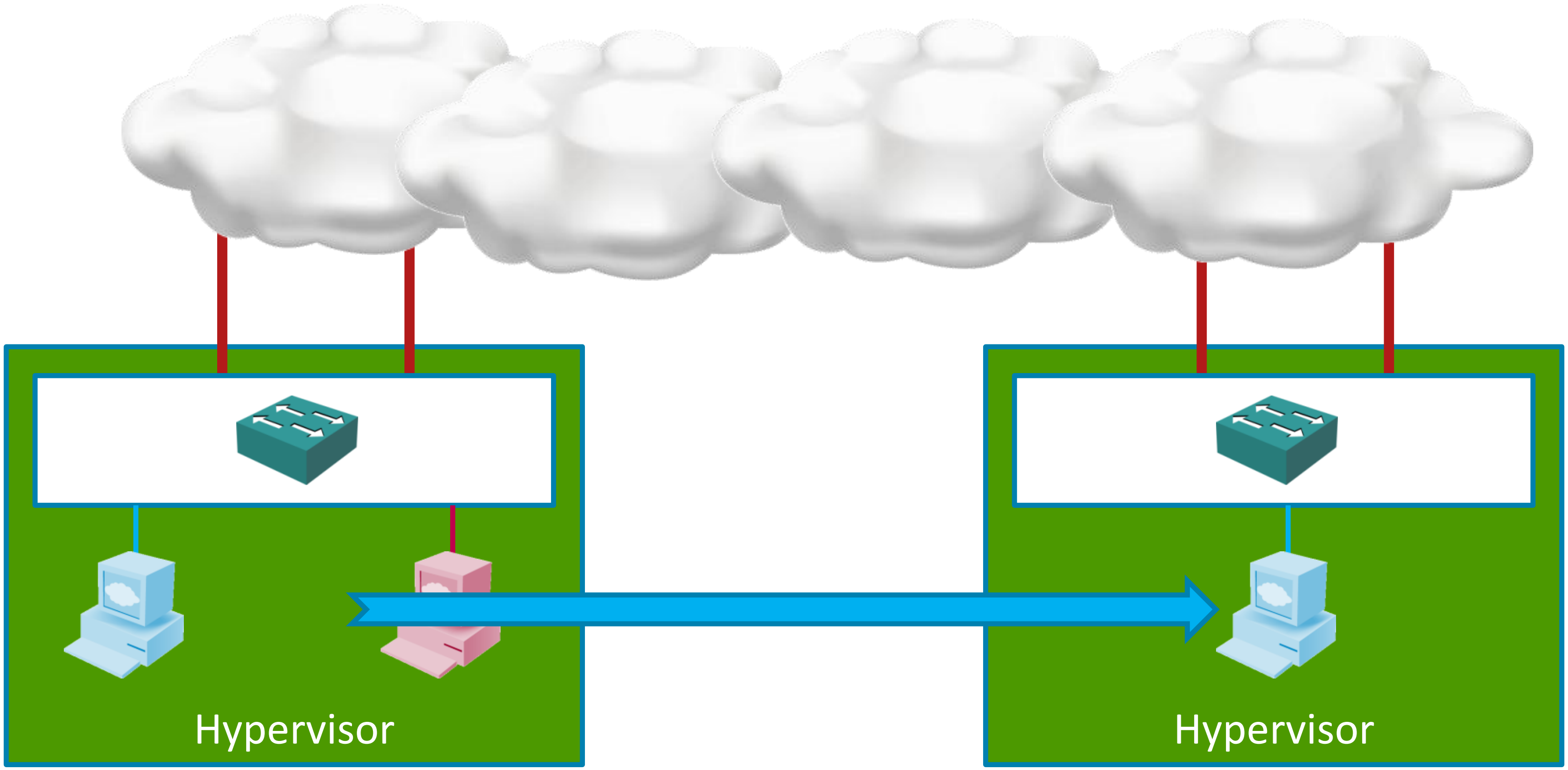
Don’t trust biased whitepapers and consultants ;)

Life Gets Better with Live Migration



- Running VM is moved to another hypervisor
- Application sessions must not be disrupted
- It actually works, but you need a layer-2 (bridged) domain

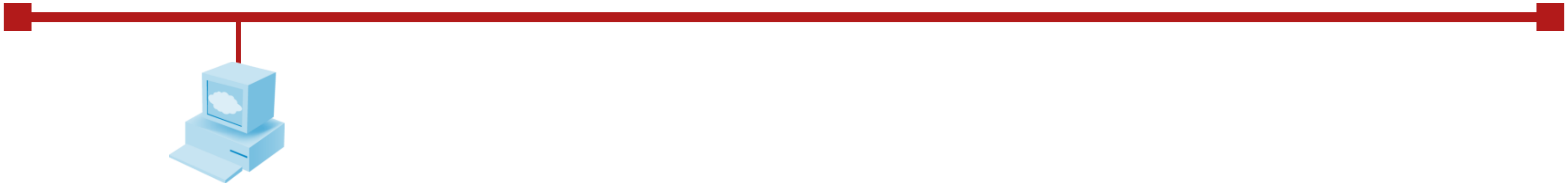
Life Gets Better with Live Migration



- Running VM is moved to another hypervisor
- Application sessions must not be disrupted
- It actually works, but you need a layer-2 (bridged) domain

What the **** is a layer-2 domain?

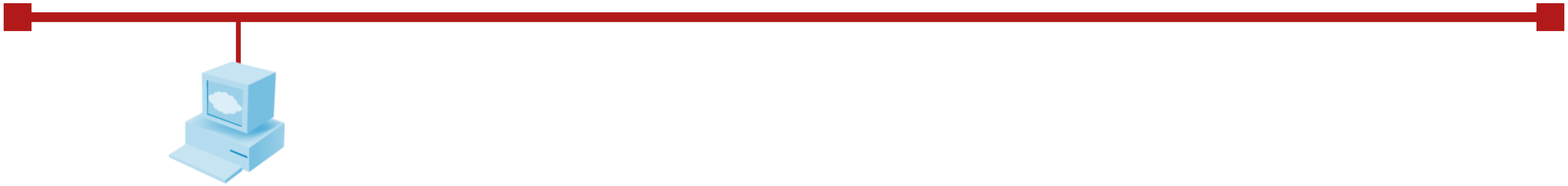
Remember Coaxial Cable Ethernet?



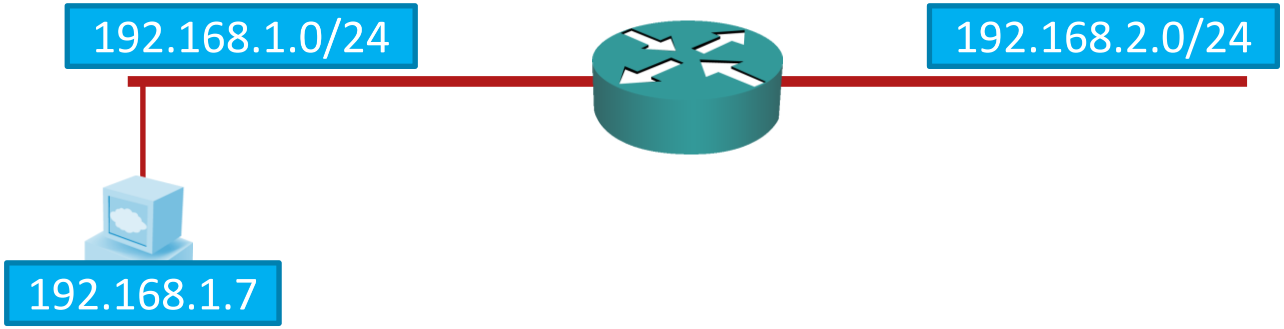
- Moving a server along a cable is a no-brainer – of course you won’t lose the user sessions unless you disconnect the server

Source: Wikipedia

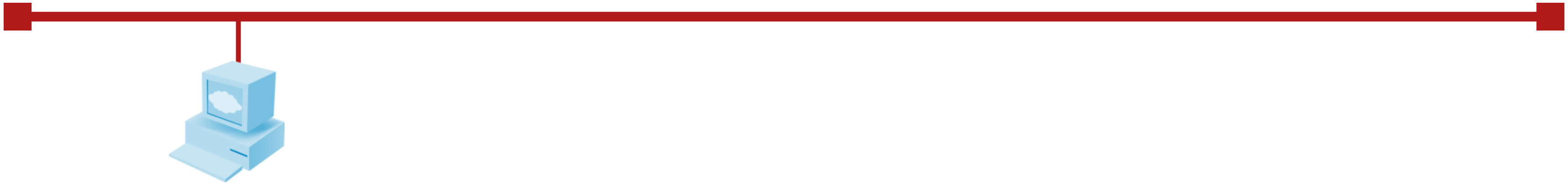
Remember Coaxial Cable Ethernet?



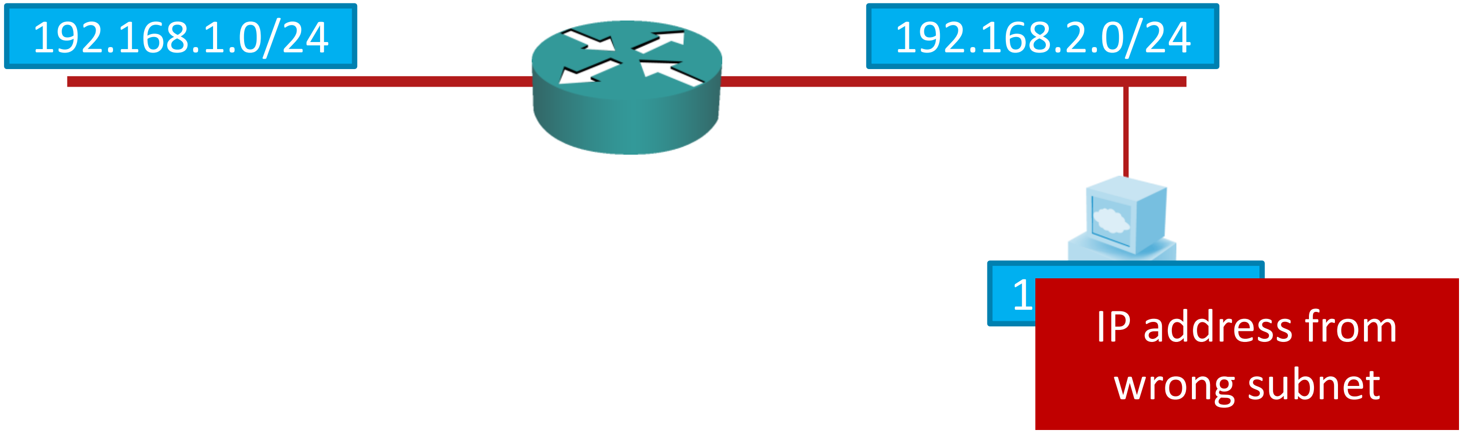
- Moving a server along a cable is a no-brainer – of course you won't lose the user sessions unless you disconnect the server
- The same trick doesn't work across routers (layer-3 switches) or between data centers (don't even think about that)



Remember Coaxial Cable Ethernet?

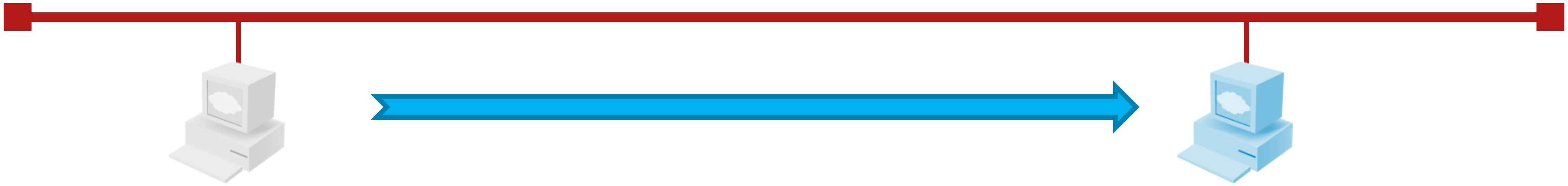


- Moving a server along a cable is a no-brainer – of course you won’t lose the user sessions unless you disconnect the server
- The same trick doesn’t work across routers (layer-3 switches) or between data centers (don’t even think about that)



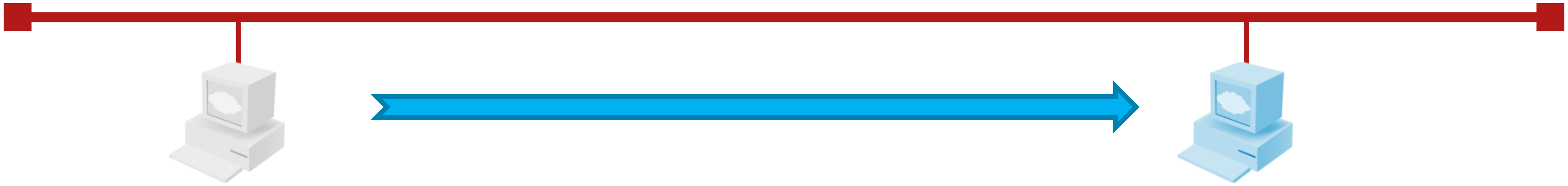
IP address from wrong subnet

We Need Virtual Coaxial Cables

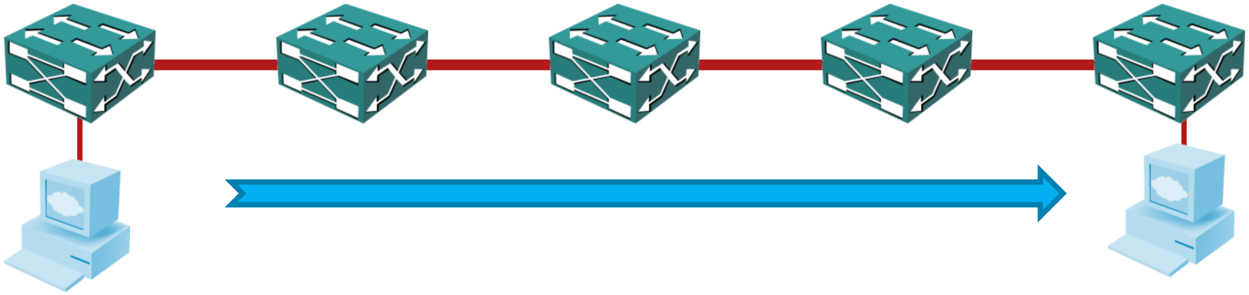


- What we need is a cable … but it should be virtual
- Actually, we need one single IP subnet
- Single IP subnet = single LAN (that’s how IP works)
- Network devices should be transparent
  ➔ bridges or layer-2 switches

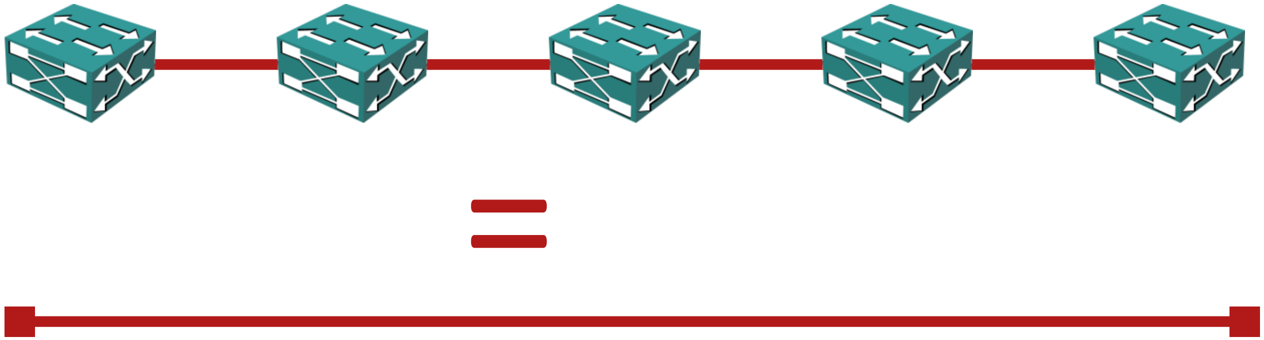
We Need Virtual Coaxial Cables



- What we need is a cable … but it should be virtual
- Actually, we need one single IP subnet
- Single IP subnet = single LAN (that’s how IP works)
- Network devices should be transparent
  ➔ bridges or layer-2 switches



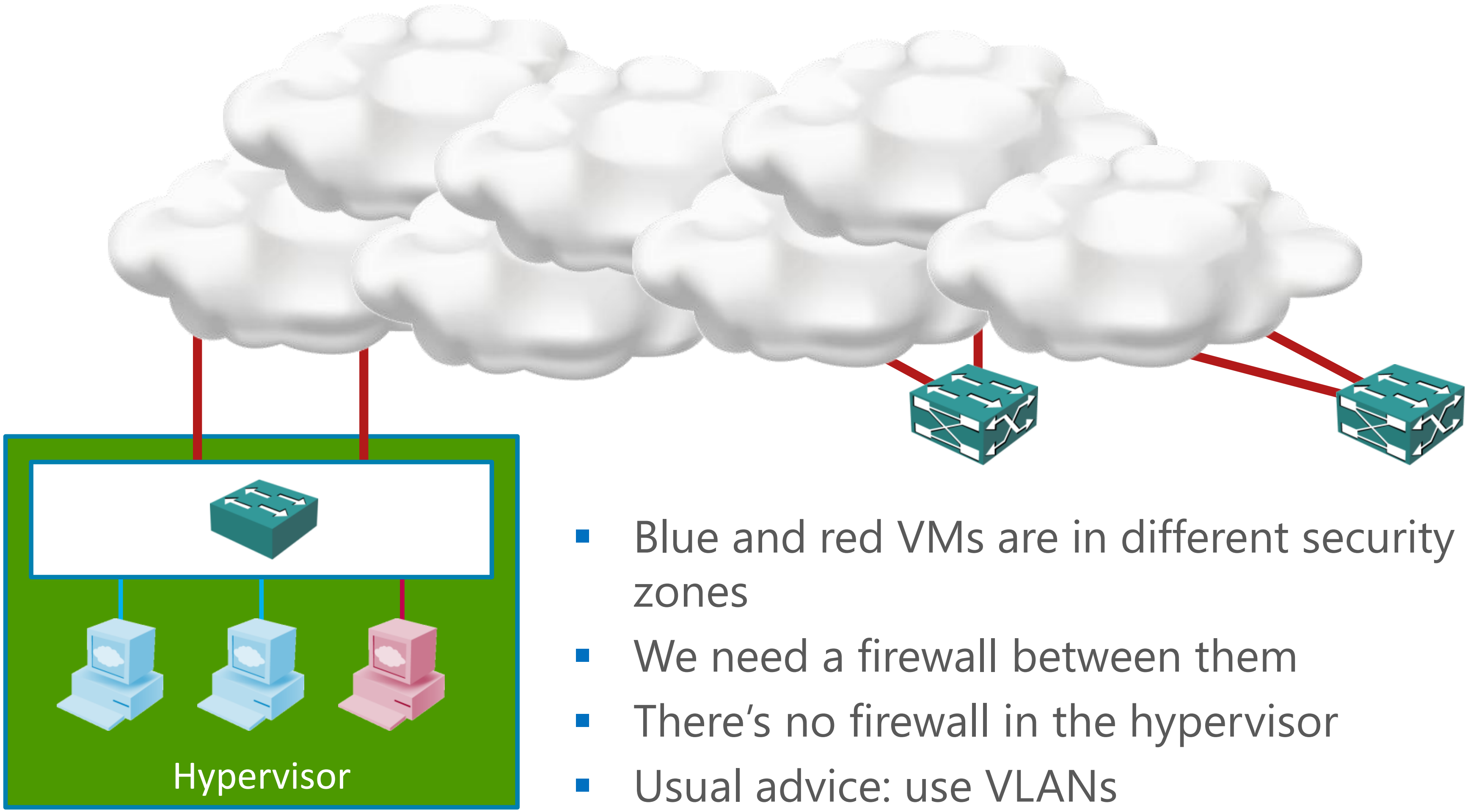
Layer-2 Domain = Virtual Cable



- Single cut in coaxial cable ➔ you lose the cable
- Layer-2 domain = cable
- Single problem ➔ you lose layer-2 domain (whole data center?)
- Got it?

Remember: layer-2 (bridged) domain = single failure domain

Next: Security



- Blue and red VMs are in different security zones
- We need a firewall between them
- There's no firewall in the hypervisor
- Usual advice: use VLANs

What the **** is a VLAN?

Remember the Virtual Cables?

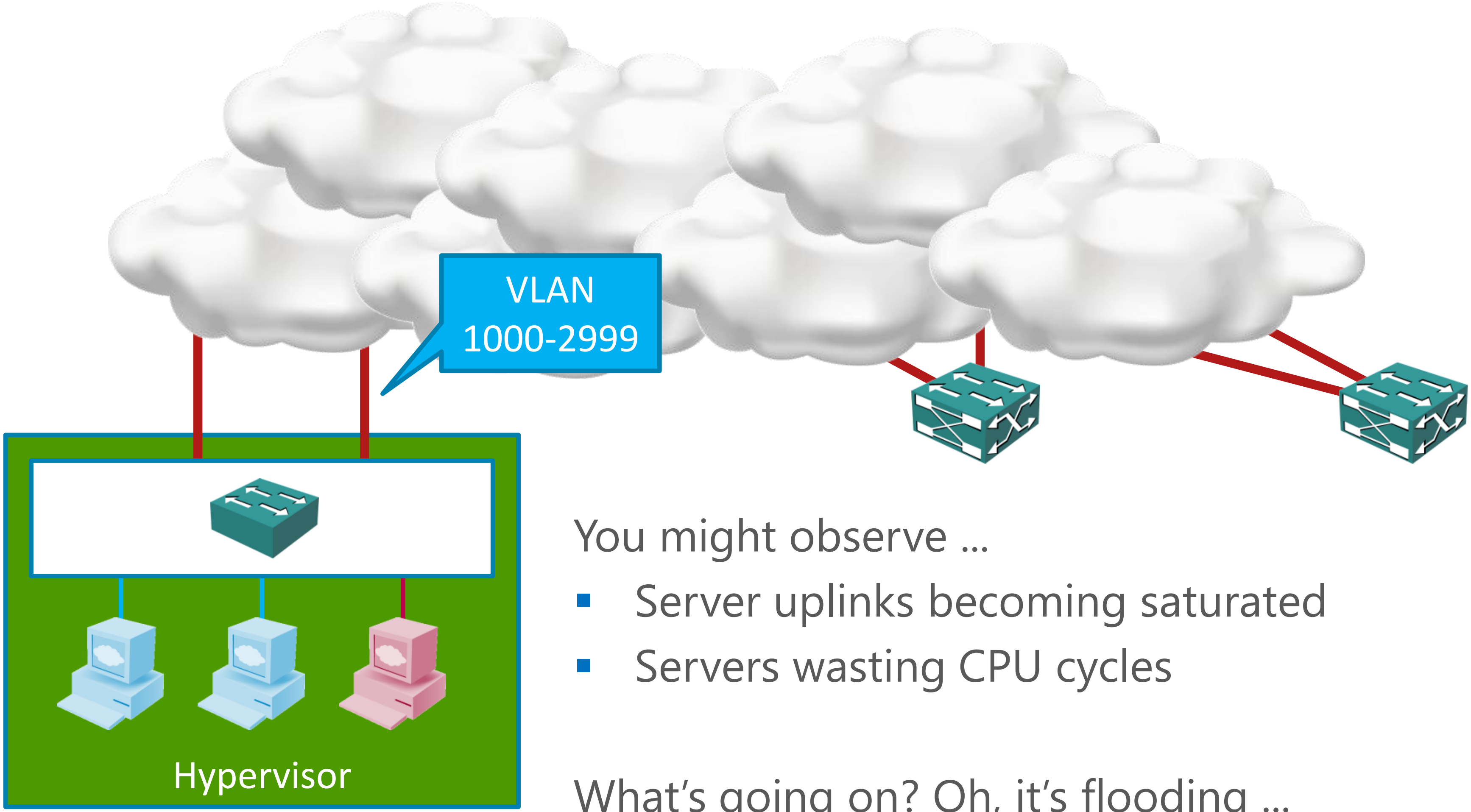
VLAN = virtual cable (only a bit more virtual than before)

- Cable number (VLAN tag) inserted in every packet (802.1Q)
- ~4000 VLANs
- VLAN = single failure domain
- VLAN numbers have to be synchronized across data center

Solutions

- Every VLAN provisioned on every server uplink
- Virtualization engineer talking with networking engineer ;)
- Network-Hypervisor integration (e.g. VM-FEX, EVB/VEPA)
- Overlay Virtual Networking

Every VLAN on Every Server Uplink



You might observe …

- Server uplinks becoming saturated
- Servers wasting CPU cycles

What’s going on? Oh, it’s flooding …

What the **** is flooding?

Flooding 101

- Every device can “hear” every other device on a coax cable
- Cable behavior is emulated with *flooding* in bridged LANs
  - Multicast and broadcast packets (reasonable)
  - Unknown unicast packets (why???)
- Some server solutions rely on cable-like behavior (Microsoft NLB)

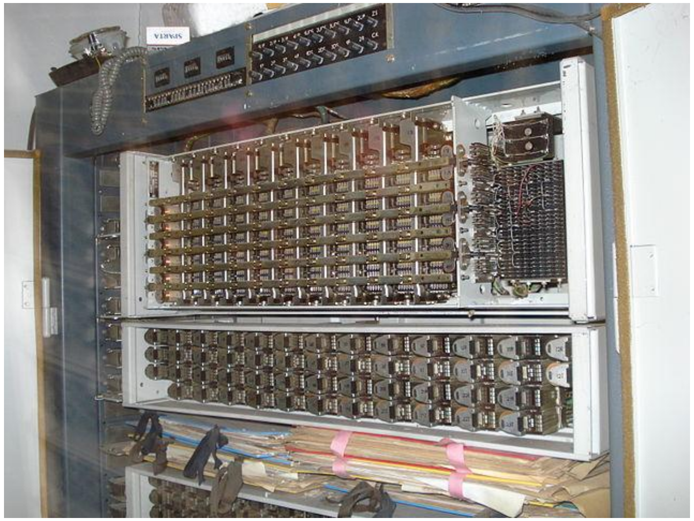
The ugly consequences

- Every server gets every flooded packet through every uplink ➔ wasted bandwidth
- Every server has to *process* every flooded packet ➔ wasted CPU

OK. I get it. What can we do?

More of the Same?

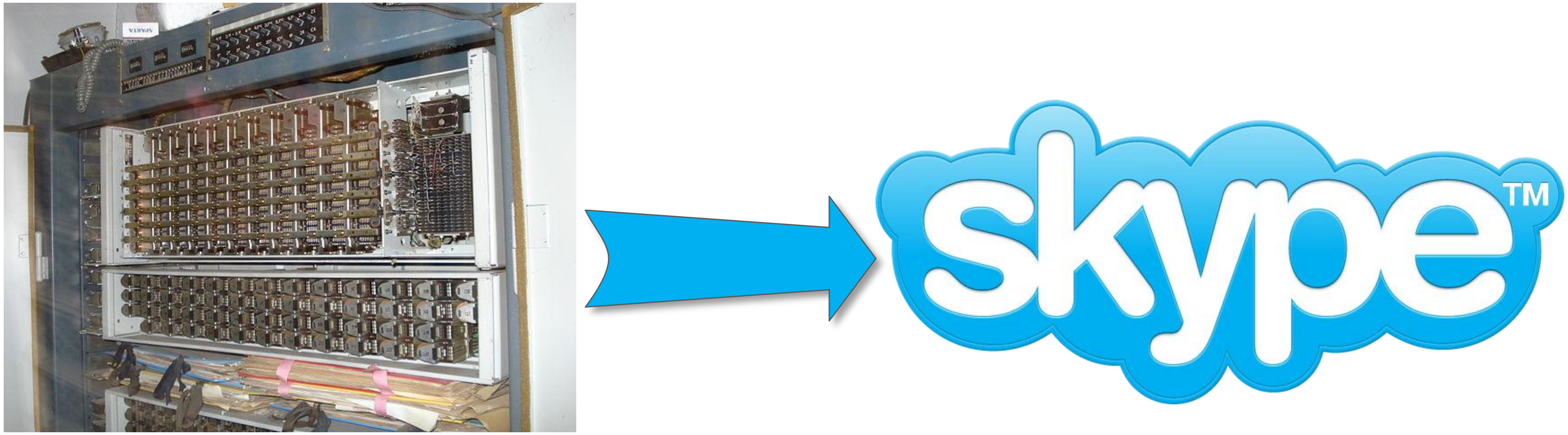
What the networking industry is proposing:



- EVB (802.1Qbg) or equivalent (VM tracer, HyperLink, VM-FEX …)
- TRILL, SPB (802.1Qaq) or equivalent (FabricPath, VCS Fabric, QFabric)
- 802.1ad (Q-in-Q) or 802.1ah (PBB)
- 802.1ak (MVRP) or equivalent (VTP)
- Numerous other features (e.g. BPDU guard, storm control)

… and you still have a single failure domain

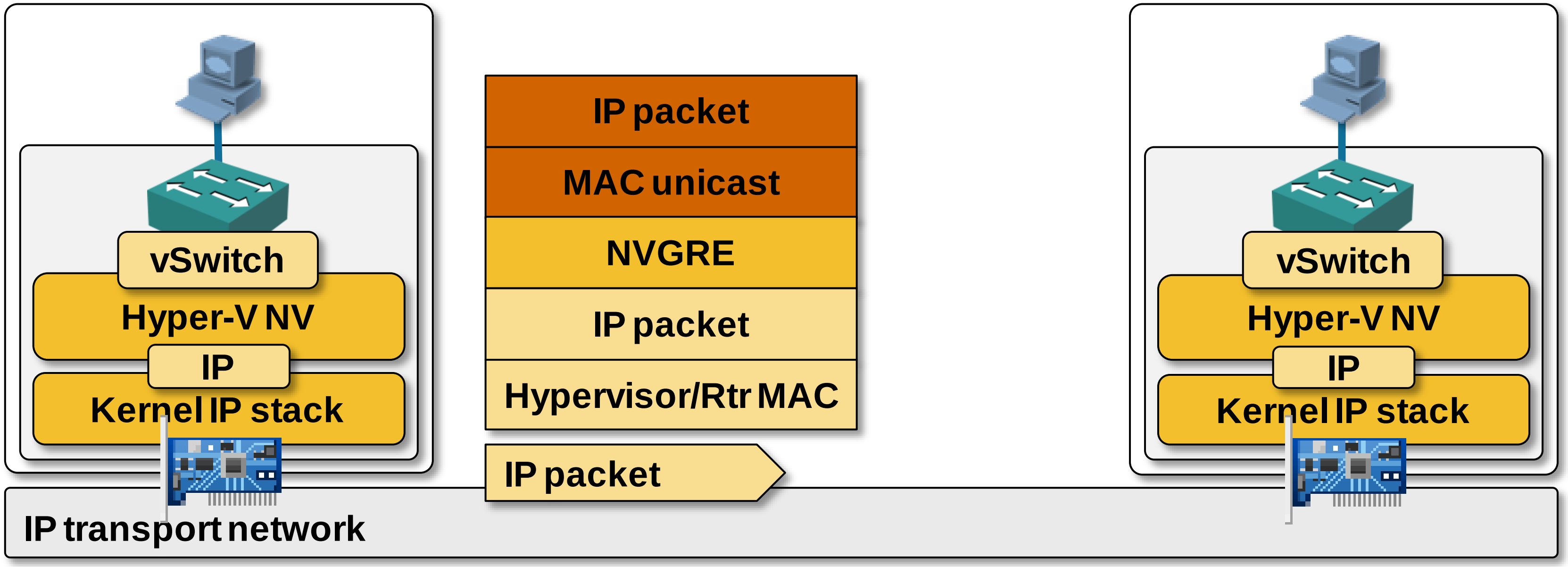
Decoupling Makes Things Simpler



- Data Center network provides fast IP transport
- Hypervisors implement virtual networks
- Virtual-to-physical interface through firewall and load balancer appliances (virtual or physical)

Keep Layer-2 domains small

- Limit live migration diameter (e.g. single cluster)
- Decouple virtual networks from physical world (VXLAN, Hyper-V Network Virtualization – NVGRE)



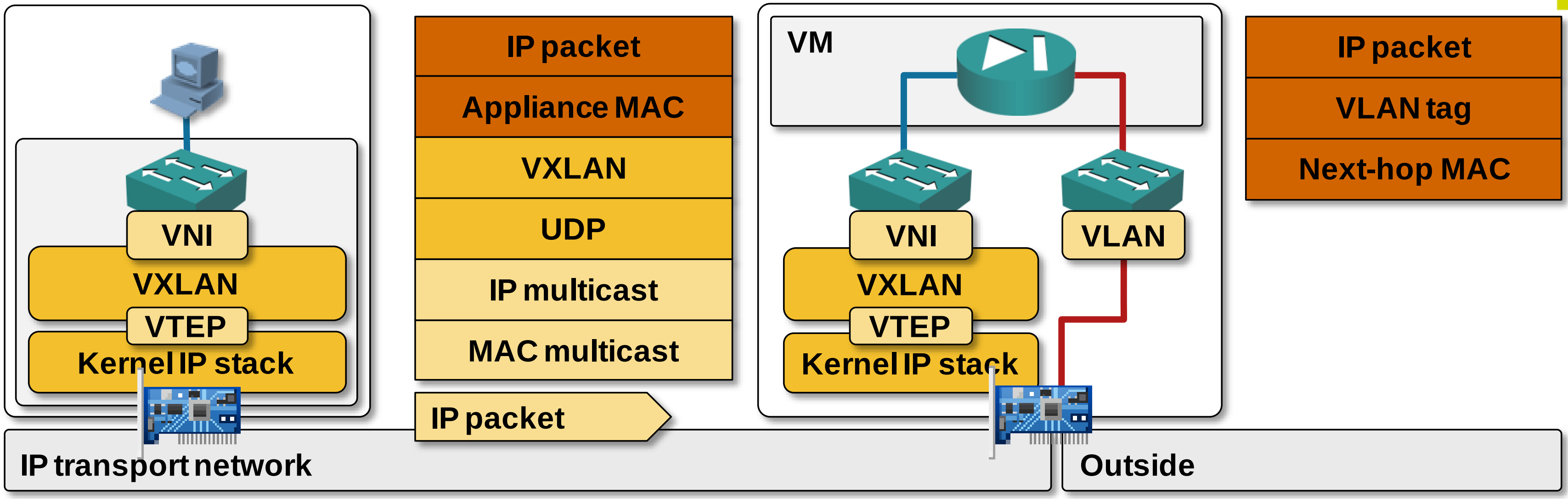
Overlay Virtual Networks 101

What we got so far:

- Network provides pure IP transport
- Hypervisors implement virtual networks
- Everything is configured through System Center (or vShield Manager or vCloud Director)

Excuse me – my clients still live in real world!

VM-Based Network Service Appliances



- Firewall (or load balancer) = x86-based device with 2+ interfaces
- Package the software in virtual disk format
- Deploy a VM with 2+ interfaces (one in VLAN, one in NVGRE segment)
- Most vendors offer VM-based solutions (Cisco vASA, F5 LTM VE, VMware vShield Edge, CloudStack, OpenStack Network Node …)

Overlay Virtual Networks – Bigger Picture

The basics:

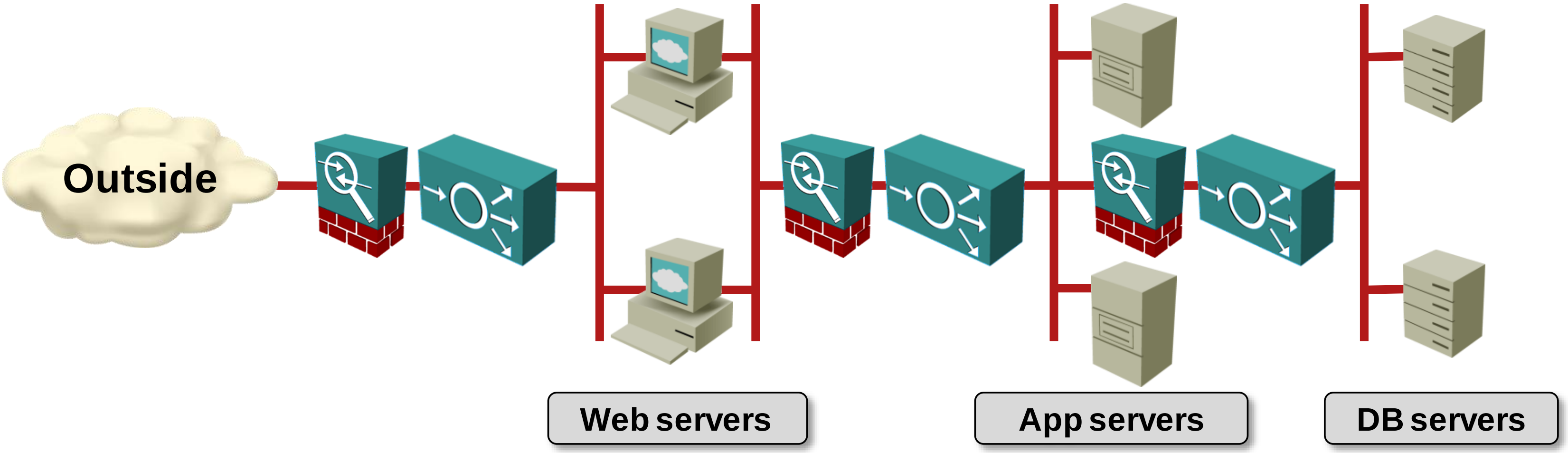
- Network provides pure IP transport
- Hypervisors implement virtual networks
- Everything is configured through System Center (or vShield Manager or vCloud Director)

Connecting virtual and physical:

- Overlay networking-aware physical appliances (F5)
- Overlay networking-aware L2 and L3 switches (Arista)
- VM-based network services (firewalls/load balancers)

Does that mean I can configure my own firewall?

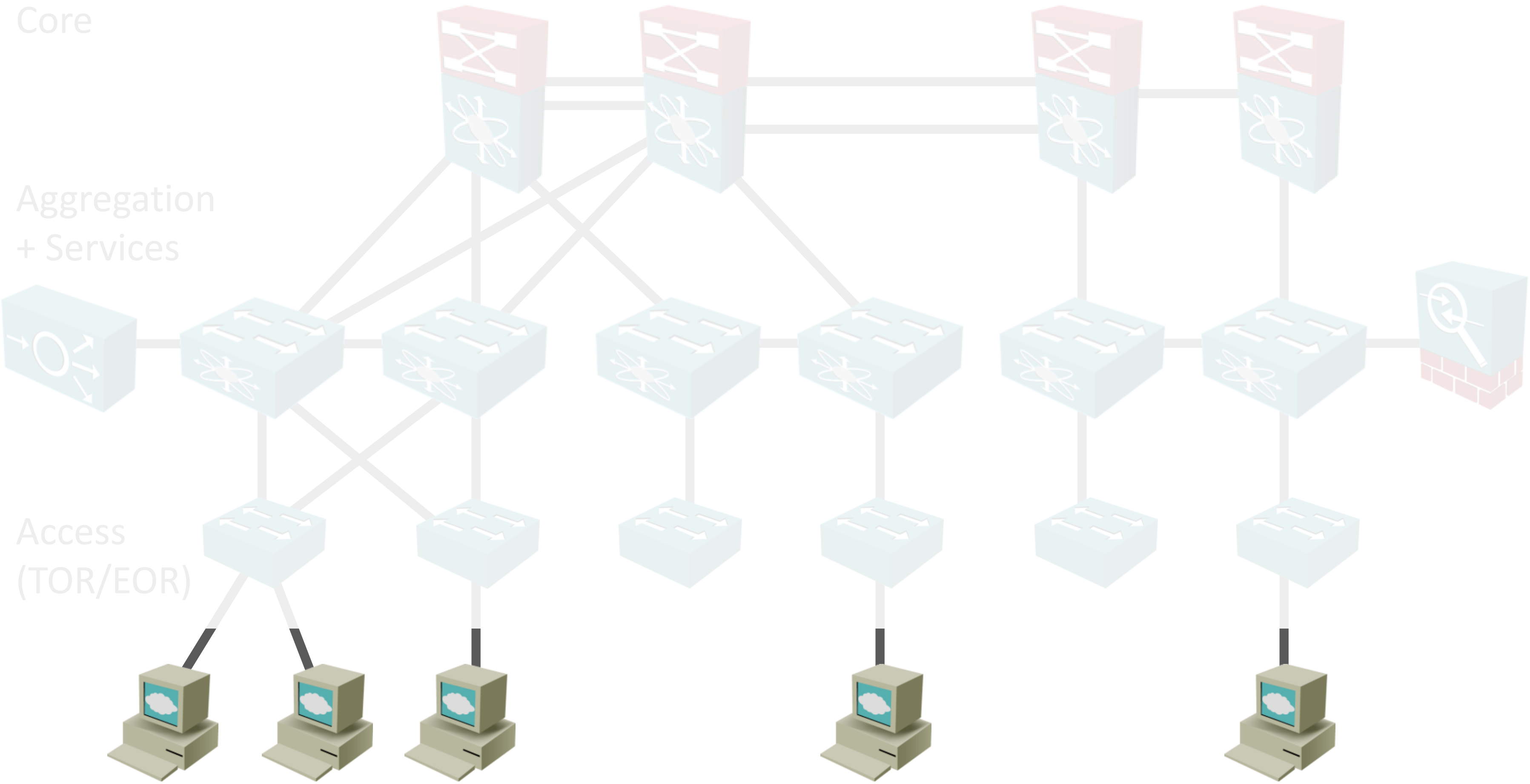
Per-Application-Stack Network Services



- Most application stacks need network services (firewalls, load balancers)
- Typical solution: large all-in-one physical appliances
  - Complex (1000s of rules), hard to operate/change
- Alternative: per-application/tenant VM appliances
  - Offered by most cloud orchestration solutions
  - Hint: easy disaster recovery ;)

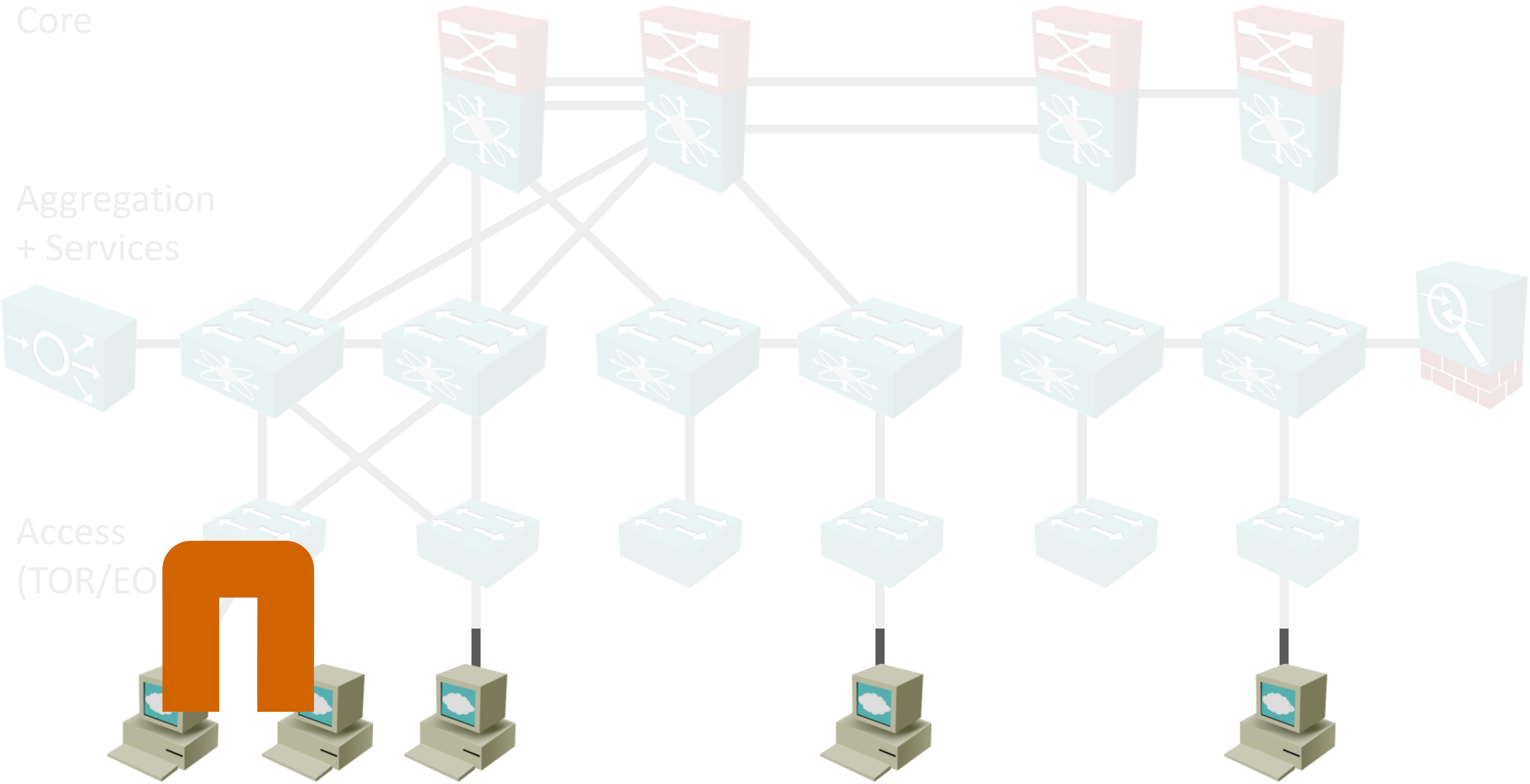
Remember: With great power comes great responsibility

Finally: Bandwidth



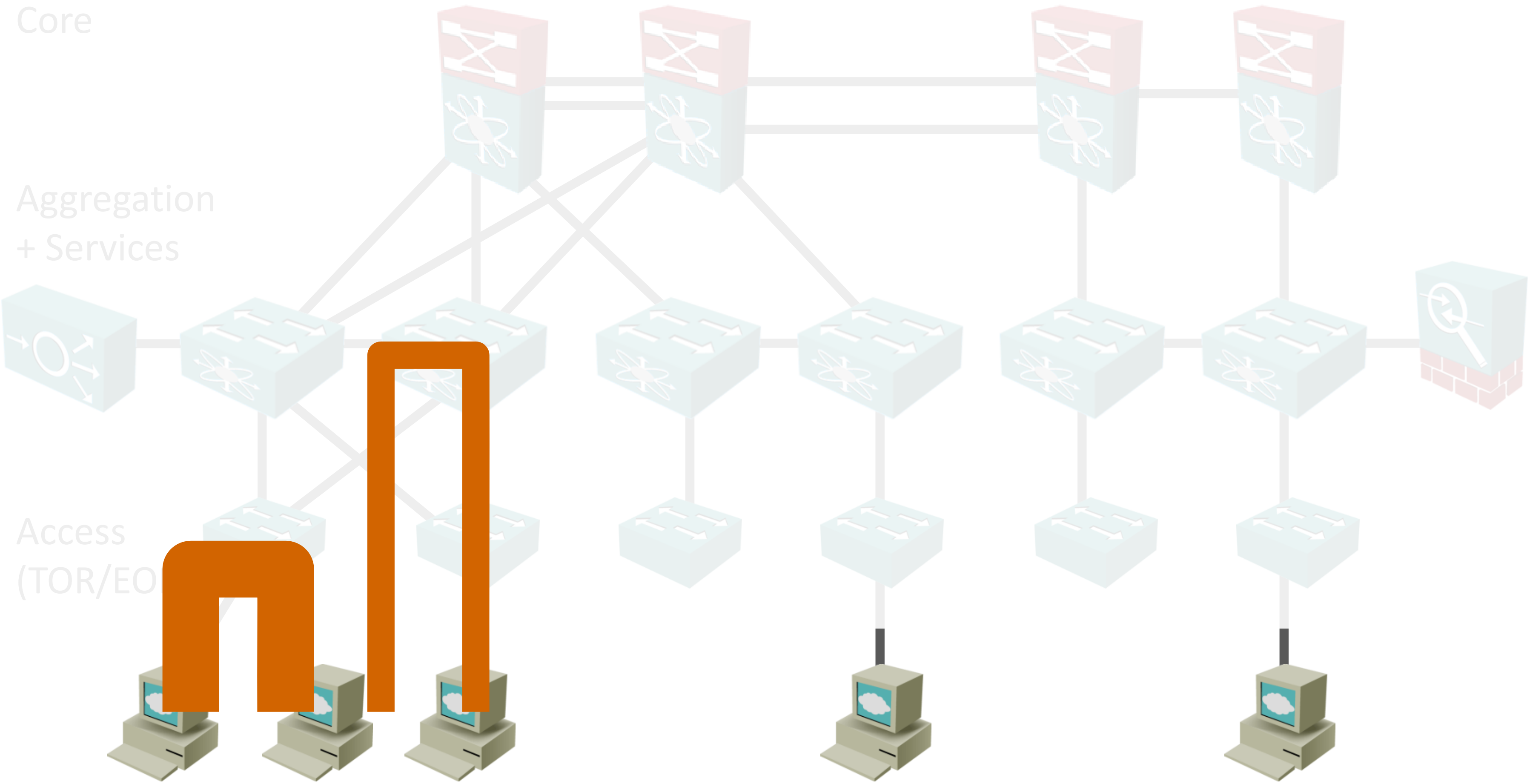
We need *equidistant endpoints* to simplify workload placement

Finally: Bandwidth



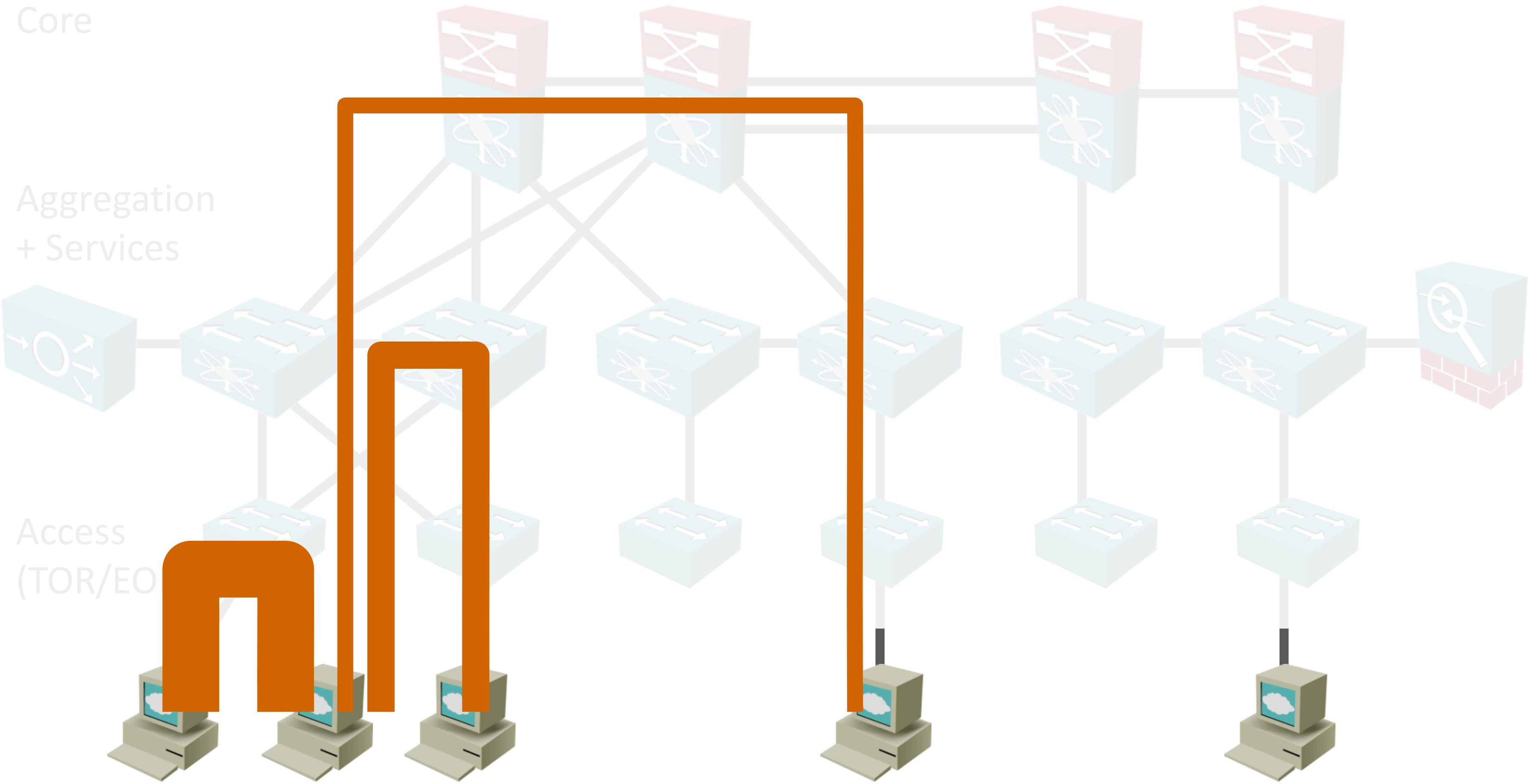
We need *equidistant endpoints* to simplify workload placement

Finally: Bandwidth



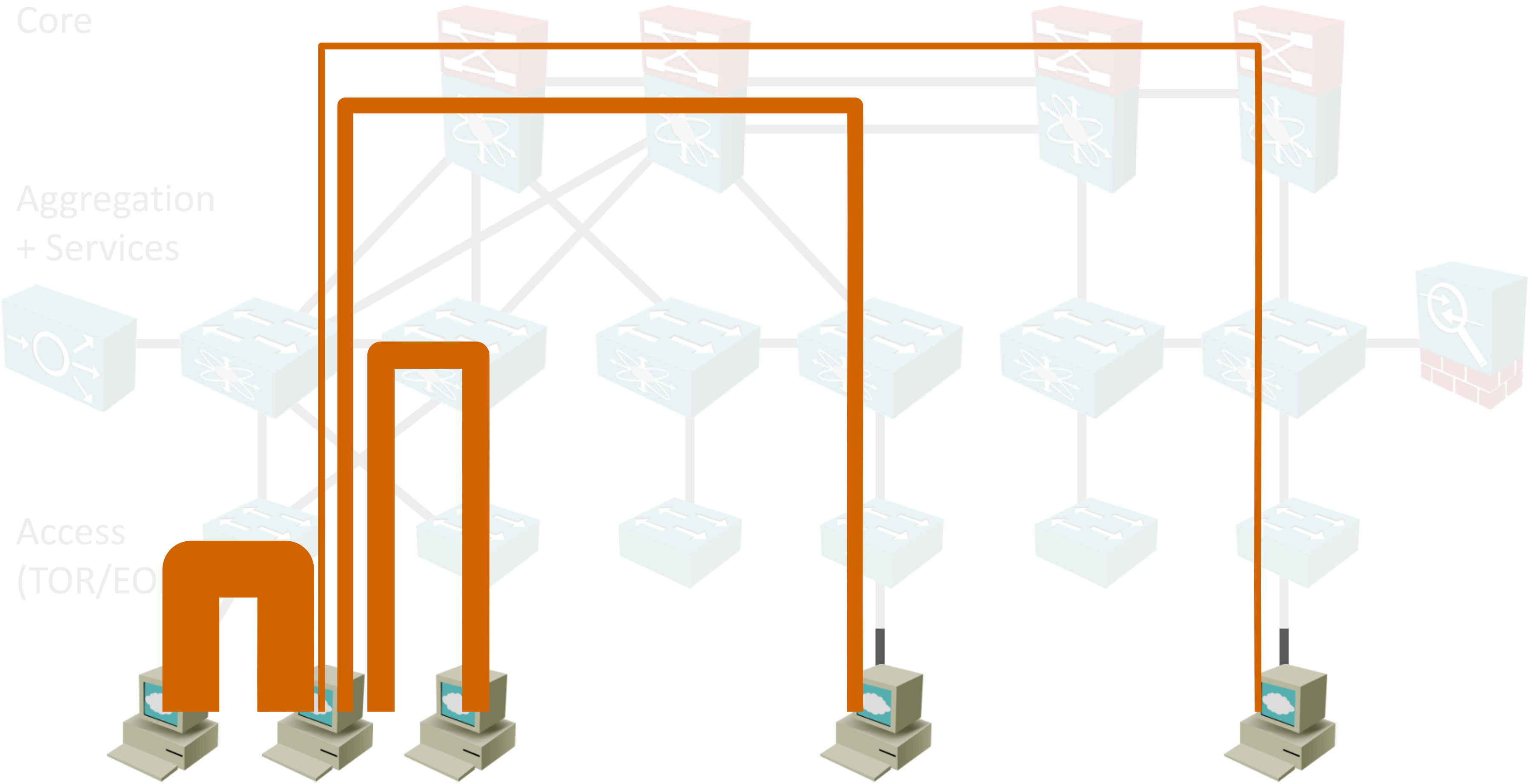
We need *equidistant endpoints* to simplify workload placement

Finally: Bandwidth



We need *equidistant endpoints* to simplify workload placement

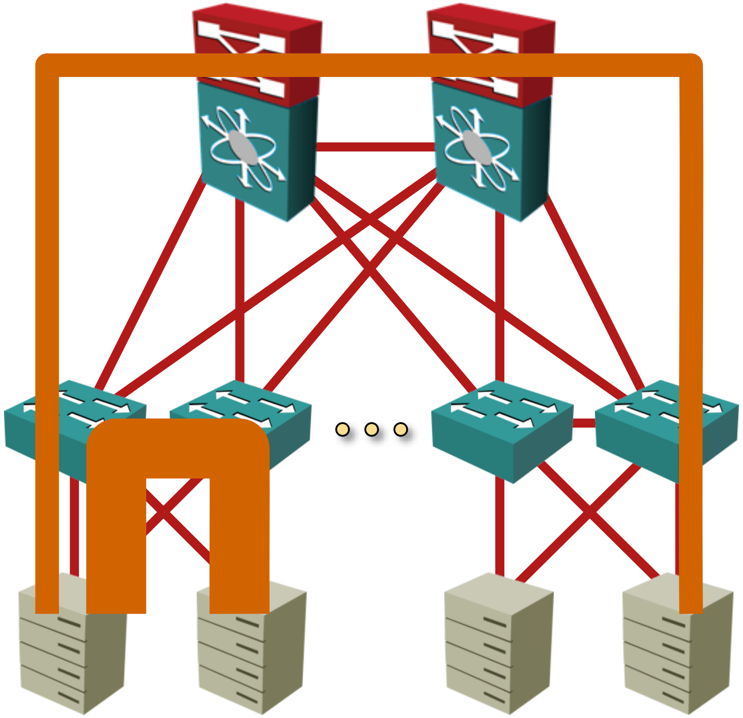
Finally: Bandwidth



We need *equidistant endpoints* to simplify workload placement

Welcome to Leaf-and-Spine World

- Modern data center network architectures give you equidistant endpoints
- Buzzword: Leaf-and-Spine
- There is no good reason not to use them



What can you do to make everyone's life easier?

- **Know your traffic!**
- High-end servers (high virtualization ratio)
- 10GE uplinks, 2 uplinks per server
- SR-IOV or similar NIC virtualization

Conclusions

Conclusions

- Compute, Storage and Network are merged in virtualized world ➔ there's no way out
- Start talking with the networking team: explain your challenges, listen to theirs (most of them are *not* excuses)
- Engage the networking team early in the planning/design process
- Consider overlay networks and virtual appliances in your 3-5 year planning

Questions?

Send them to ip@ipSpace.net or @ioshints