Virtual Switching in an Era of Advanced Edges

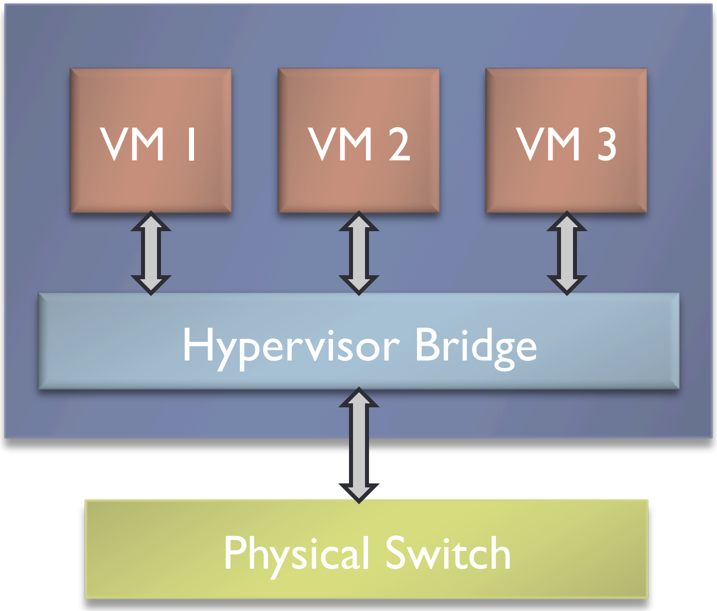
Justin Pettit
Ben Pfaff
Jesse Gross
Martin Casado
Nicira Networks

Simon Crosby
Citrix Systems

DC CAVES 2010

What is Virtualization?

- Multiple virtual machines on the same physical host
- Lowest layer is the hypervisor, which provides the illusion
- Built by OS people
- Historically, simple bridge

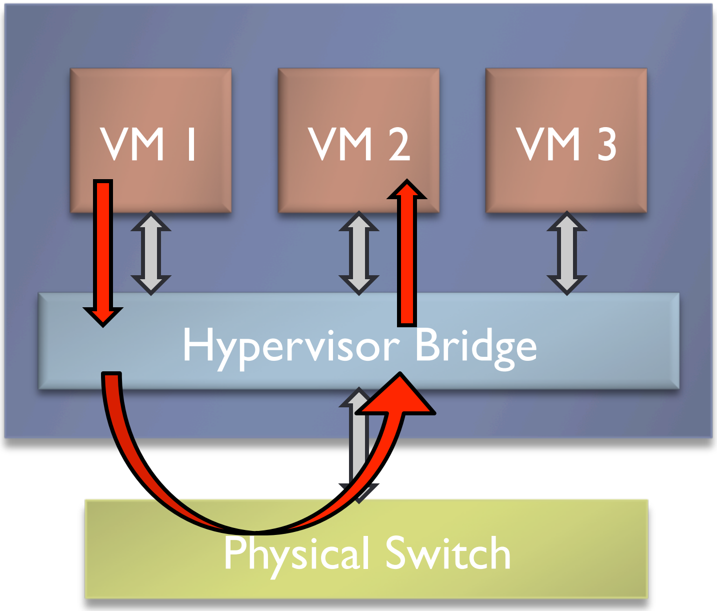


Impact of Virtualization on Networking

- IP doesn’t support mobility in a scalable manner
  - Flat networks and VLANs don’t scale
  - Policies don’t follow host movement
- Network infrastructure needs to change
  - Know logical context (directly or tags)
  - Adapt to changes in the virtualization layer (signals or inference)

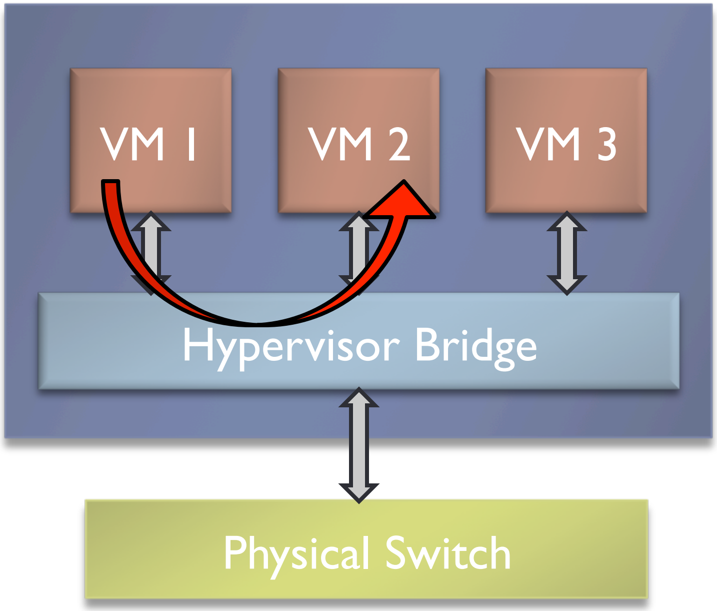
Hairpin Switching

- Use hardware that's already in the network
- Bridge already dumb, make it dumber (and simpler)
- All traffic bounces off the adjacent switch



Switching at the Edge

- Strengths
  - Greater context
  - Enforce policies early
  - Inter-VM traffic has less overhead
- Weaknesses
  - CPU overhead
  - Additional switches to configure and monitor
  - Historically, feature-weak



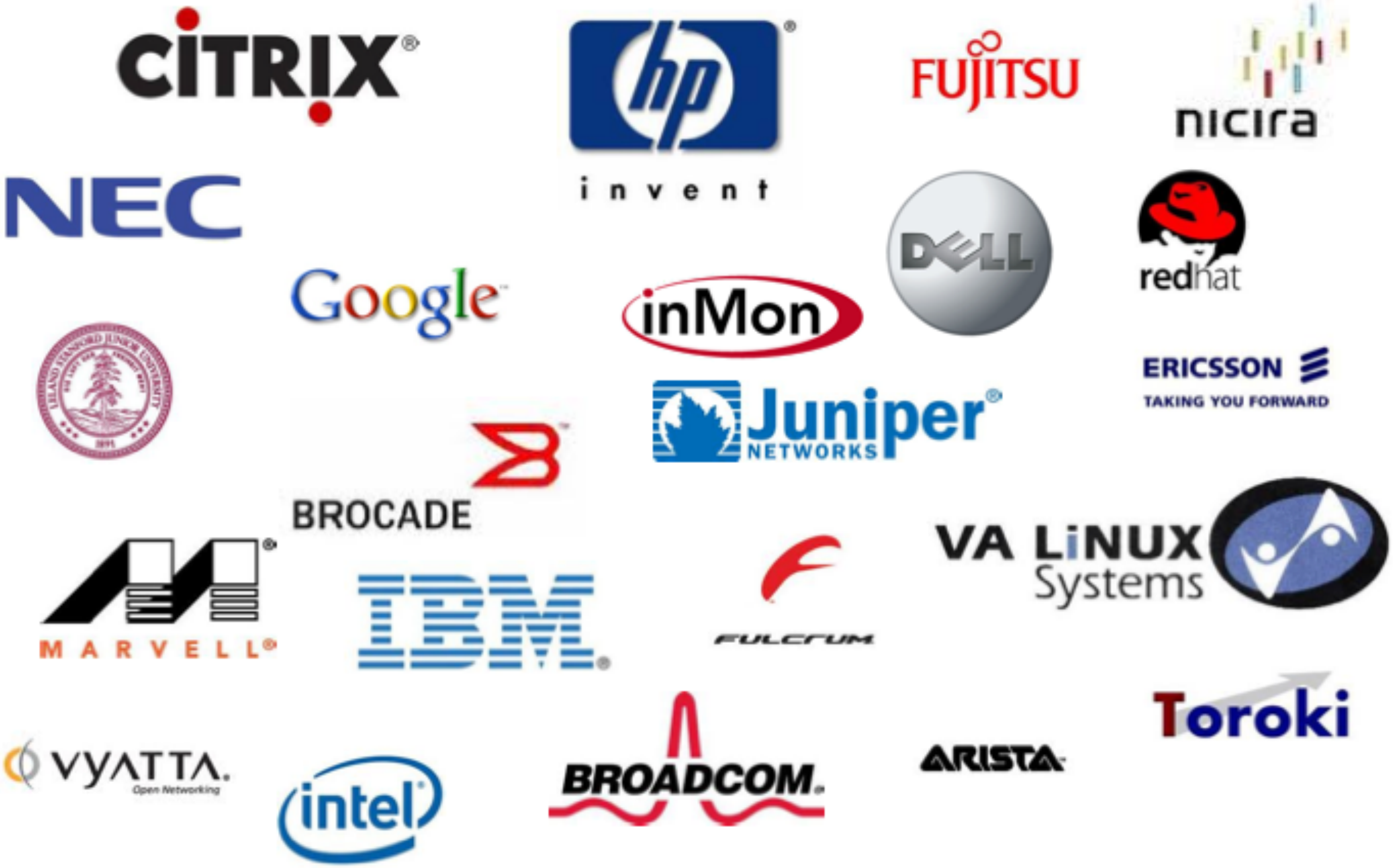
Advanced Edge Switches

- Hardware-offloading
- Centralized management
- Approaching feature-parity with hardware switches
  - Visibility
  - ACLs
  - Quality of Service
- Examples: VMware vSwitch, Cisco Nexus 1000V, Open vSwitch

Open vSwitch

- Visibility (NetFlow, sFlow, SPAN/RSPAN)
- Fine-grained ACLs and QoS policies
- Centralized control through OpenFlow
- Port bonding, GRE, and IPsec
- Works on Linux-based hypervisors: Xen, XenServer, KVM, VirtualBox
- In the process of being upstreamed to Linux
- Open source, commercial-friendly Apache 2 license
- Multiple ports to physical switches

Open vSwitch Contributors



Approaches Compared

- Cost
- Performance
- Tagging

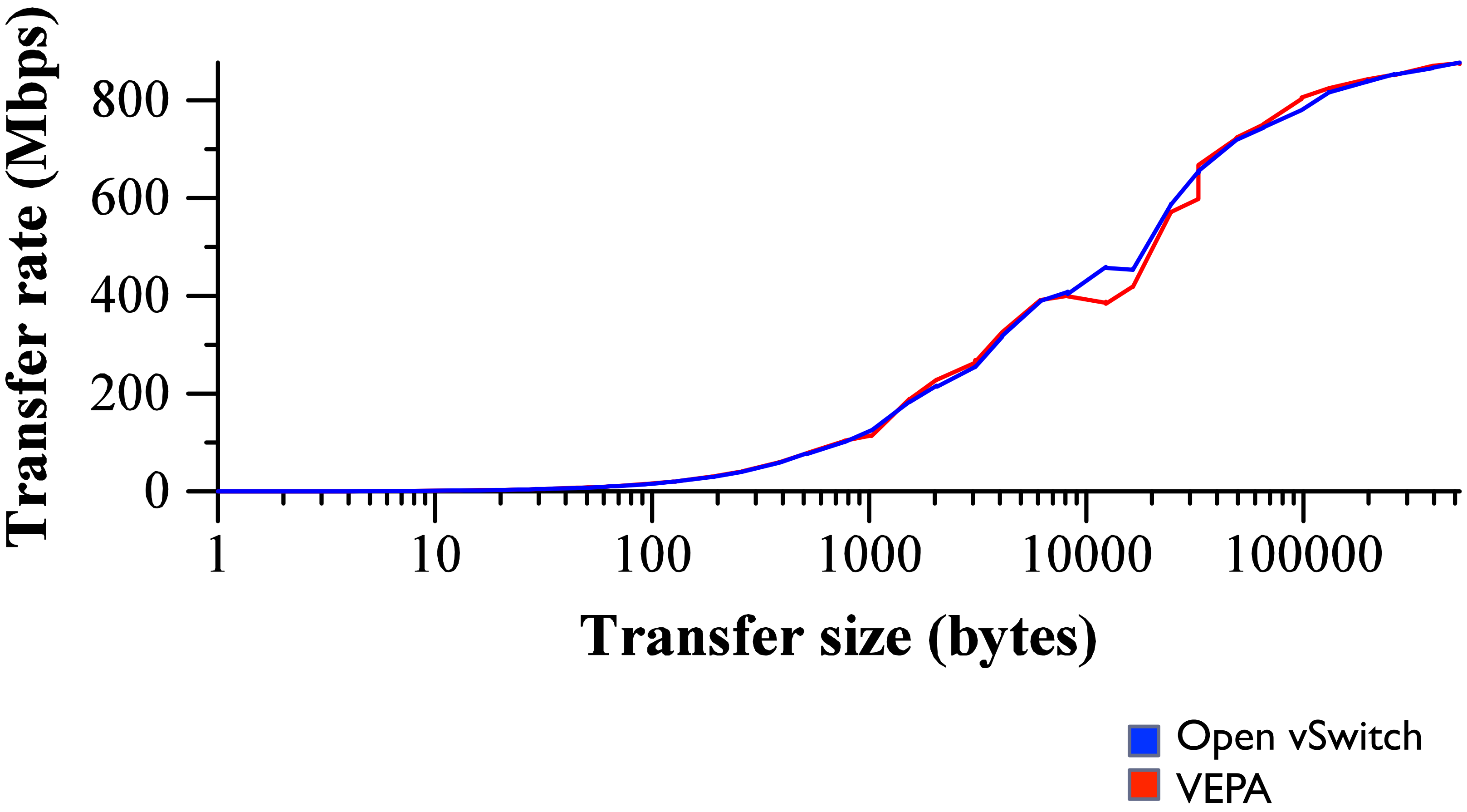
Cost

- Hairpin switching may be able to use existing equipment, but becomes aggregation device that must scale to a much larger number of virtual interfaces
- Edge can support larger number of policy rules
- Edge switch is just software, which makes it easy to add new features
- Without hardware acceleration, both approaches consume hypervisor CPU cycles
- Edge can always fall-back to software when hardware not available

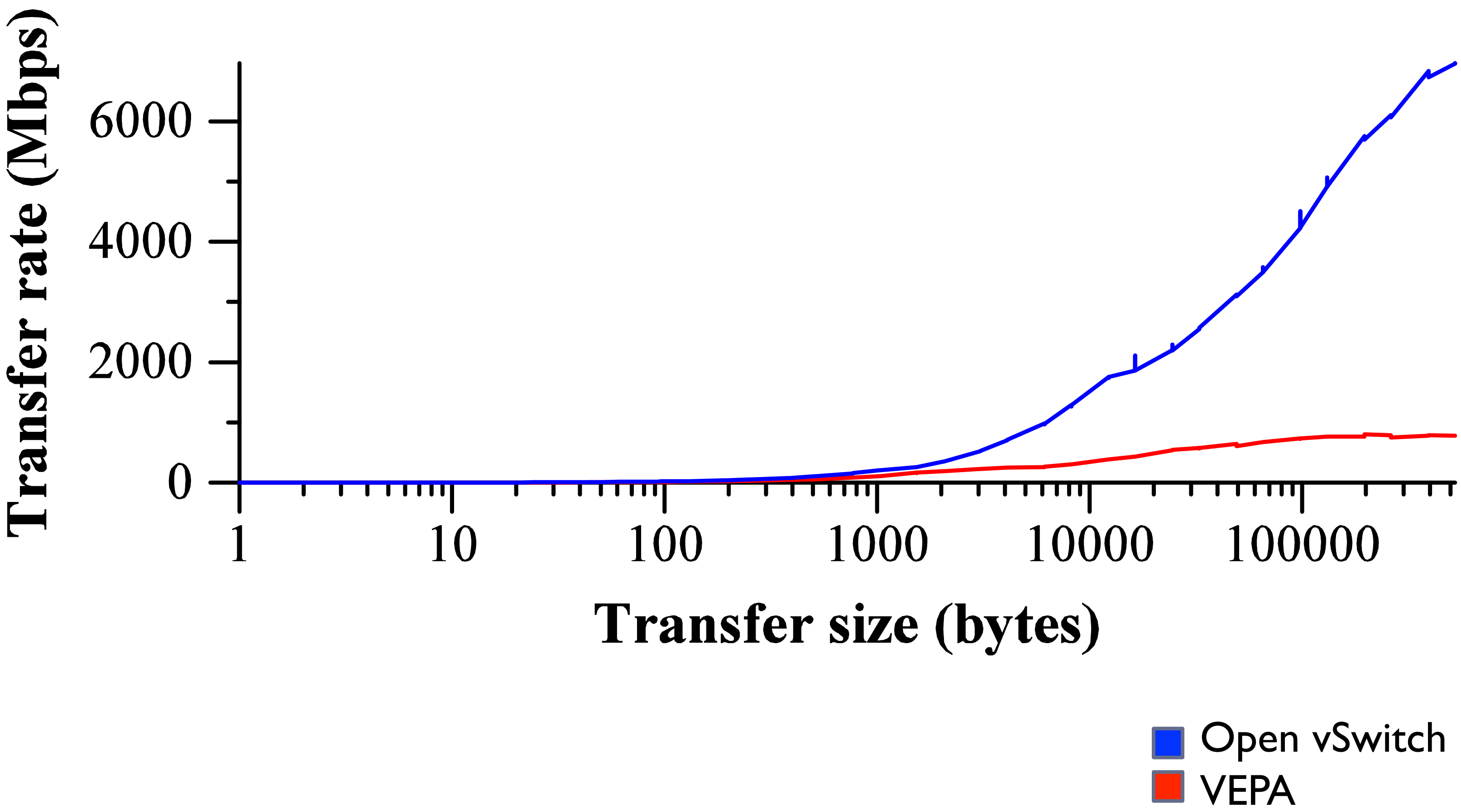
Performance

- Edge switches have been demonstrated at 40Gbps—at significant CPU overhead
- Traffic can be dropped closer to the source with edge switch—important in clouds with over-subscribed links and untrusted sources
- Both need offloading to not take CPU hit
- Checksum and TSO offloading provide big wins; SR-IOV even bigger
- Edge will be faster for local VM-to-VM traffic

Off-box Performance



On-box Performance



Tagging

- Without tags, hairpin switch must rely on fields that are easily spoofed
- Distinguish context, but don’t say anything about the contexts—need port profiles
- Tag space limited and may cause issues with multicasting and mobility
- On the plus side, may provide context throughout the network

Future

- NICs will do the heavy-lifting
  - New types of offloading
  - Bypass the hypervisor in the common case (e.g., SR-IOV)
  - Push the datapath into the NIC
- Edge is approaching feature-parity with high-end switches
- Physical switches adding same control interfaces as edge, for a unified control interface throughout the network

Conclusion

- Hairpin switches attractive when applying similar policies over all nodes or in aggregate with little local VM-to-VM traffic
- Edge switches provide more flexibility and fine-grained control at cost of hypervisor CPU cycles
- Best approach likely uses both
- Need common standardized control interface