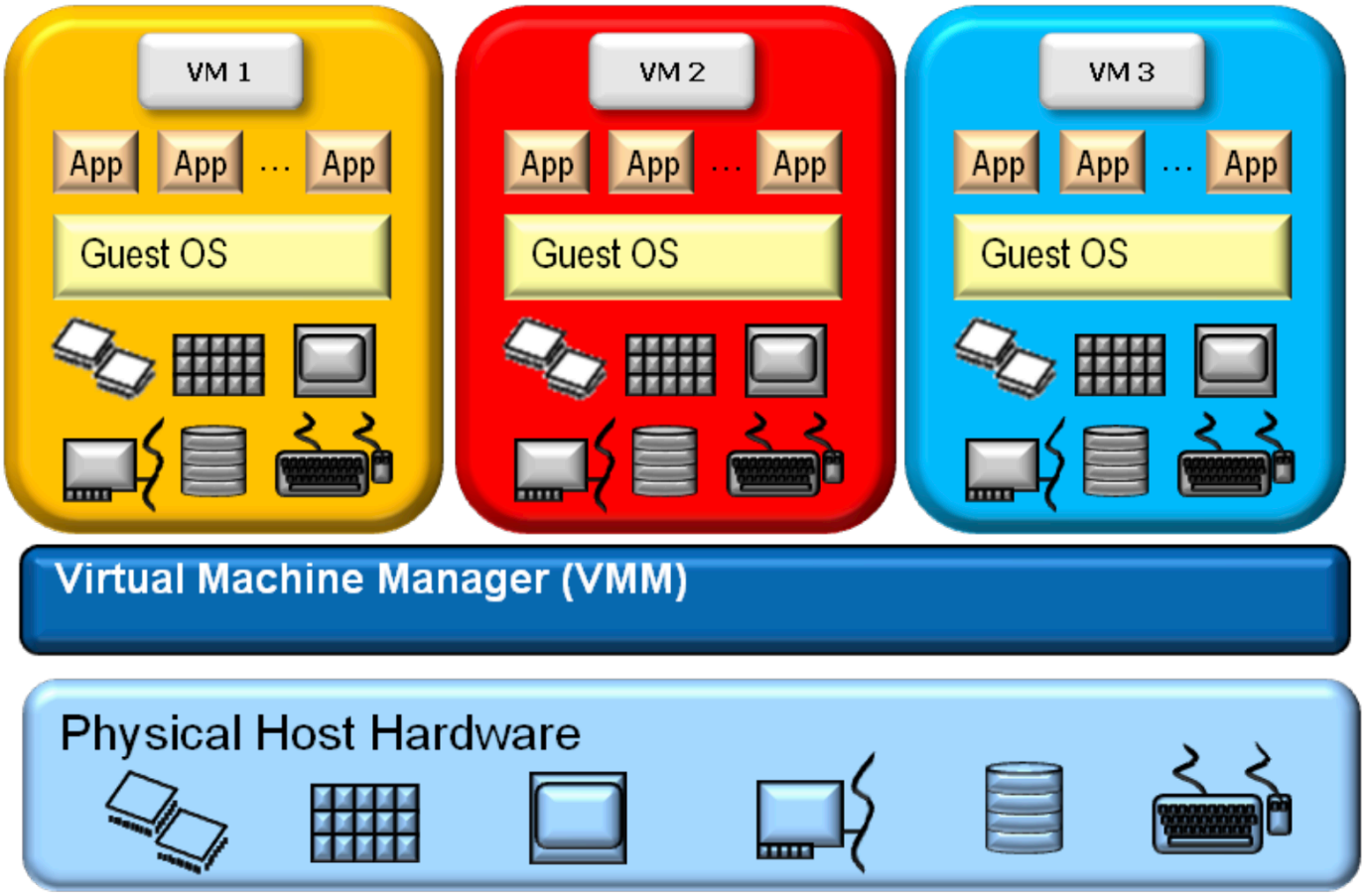
Network Virtualization and Host SDN

ECE/CS598HPN

Multi-tenant datacenters



Networking between VMs

- Conventional approach:
  - Physical network treats each VM as a host directly attached to it.
  - The vSwitch in the hypervisor extends physical network.

Issues with this approach

- Physical network is aware of tenant addresses.
  - Difficult to scale.

- Address space tied to physical network.
  - VMs get IP from subnet of first L3 router they are attached to.
  - Hinders VM mobility.
  - VMs can't run their on IP address management schemes.

Networking between VMs

- Different workloads (VM clusters) have different service requirements.
  - Some require L2 routing.
  - Some require L3 routing.
  - Some require special services (e.g. L4 load balancing).

Virtualization techniques

- VLANs: virtualized L2 domains.
- VRF: virtualized L3 forwarding tables.
- NAT: virtualized IP address space.
- MPLS: virtualized paths.

Point solutions that virtualize singular aspects.

Need for a more holistic and global approach.

Network Virtualization

Allows creating virtual networks

(each with independent service models, topologies, and addressing schemes)

over the same physical network.

These virtual networks are created, configured, and managed via global abstractions rather than individual box-by-box configuration.

Network Virtualization

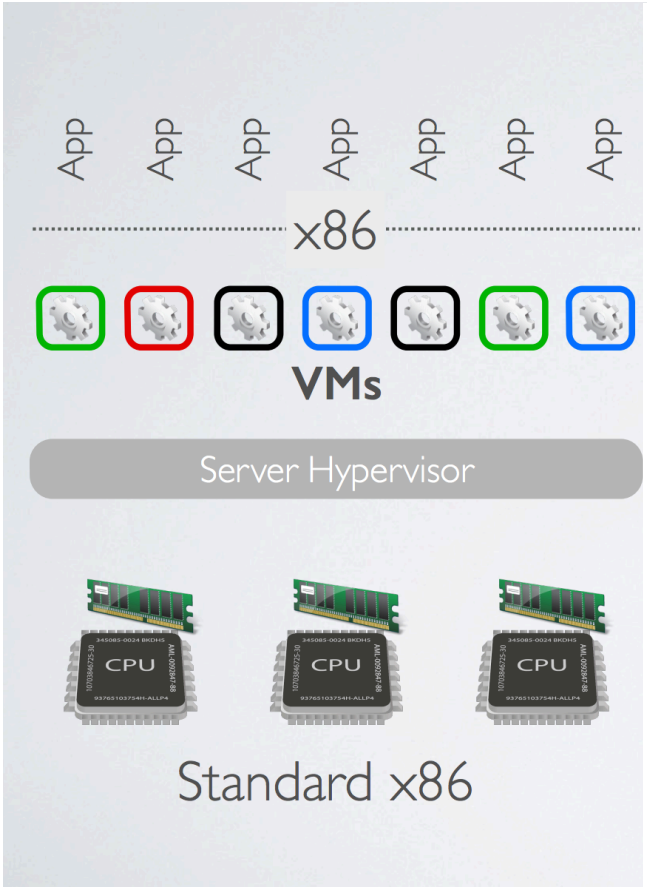
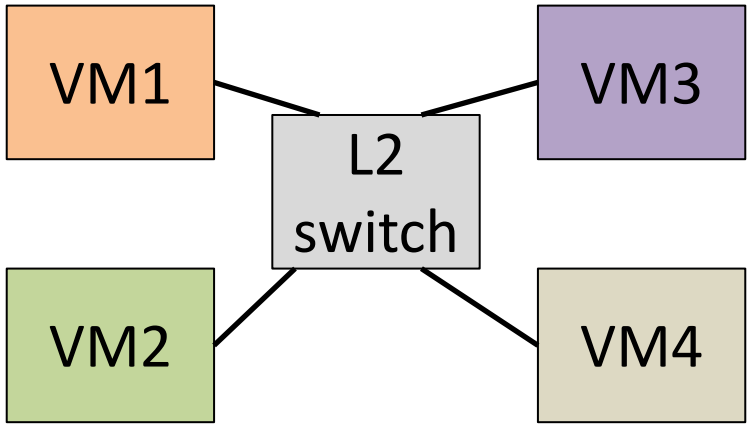
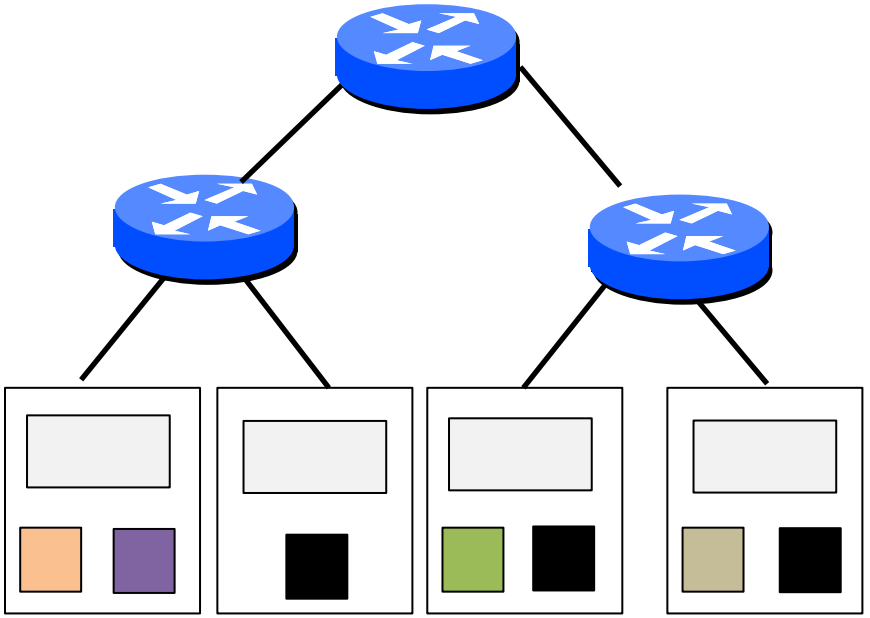


Figure from NSDI'14 talk on "Network Virtualization in Multi-tenant Datacenters"

Network Virtualization

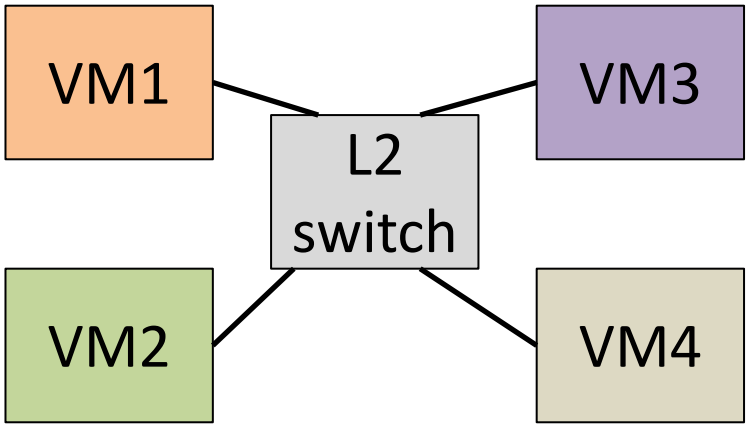


Abstraction



Physical Topology

Network Virtualization

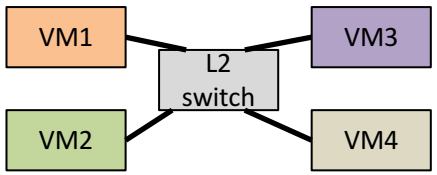


Abstraction

Control abstraction: Tenants configure their policy using a centralized controller, agnostic of physical layout, as they would have configured such a physical network.

Packet abstraction: Packets get the same switching, routing and filtering service as they would in the tenant's home network.

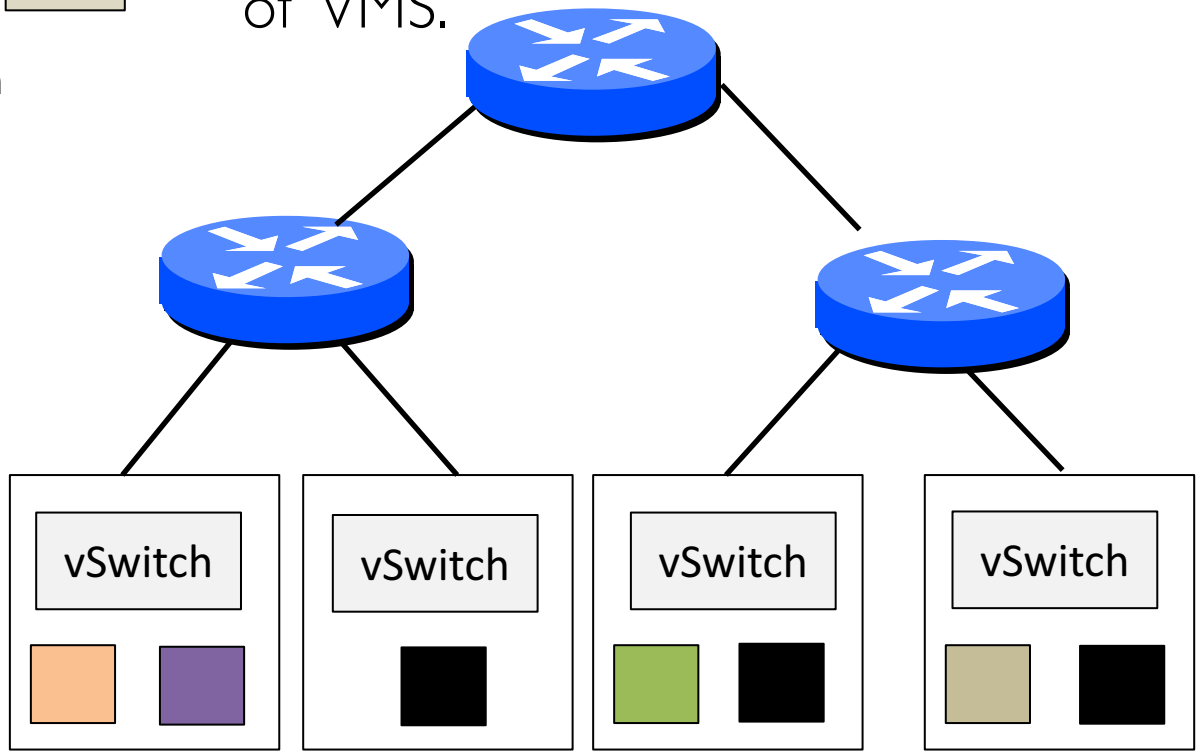
vSwitches provide the abstraction



Abstraction

All the smartness is in the vSwitch:

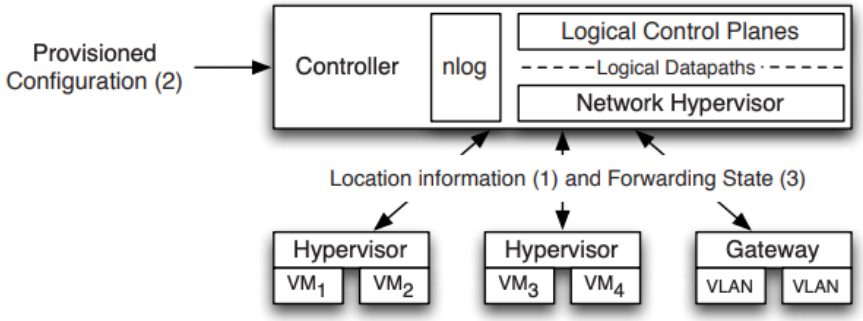
- VMs behave as if they are on their own network.
- Physical network routes between servers, agnostic of VMS.

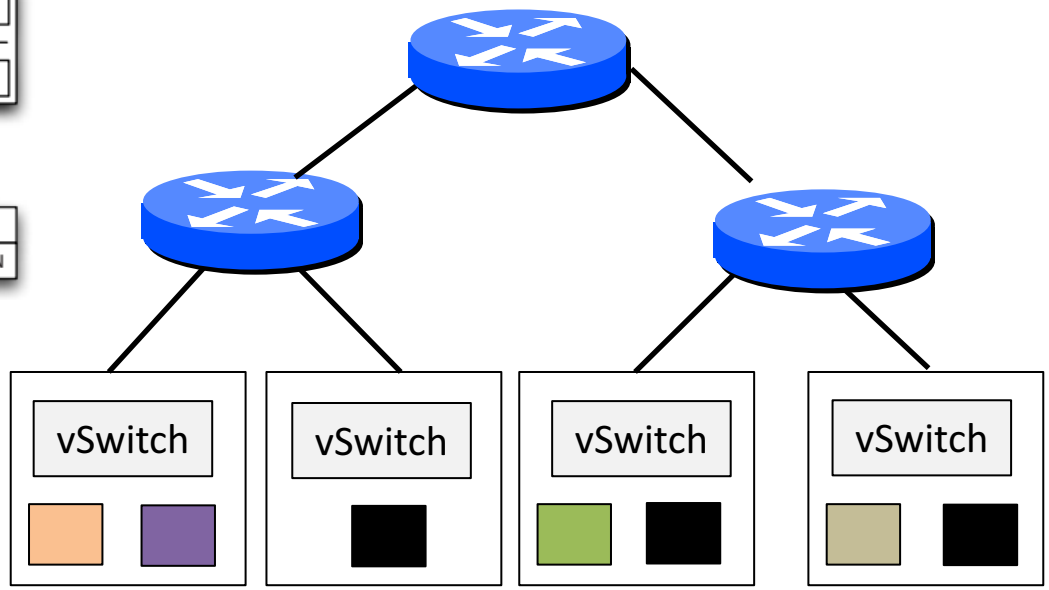


Physical Topology

vSwitches provide the abstraction

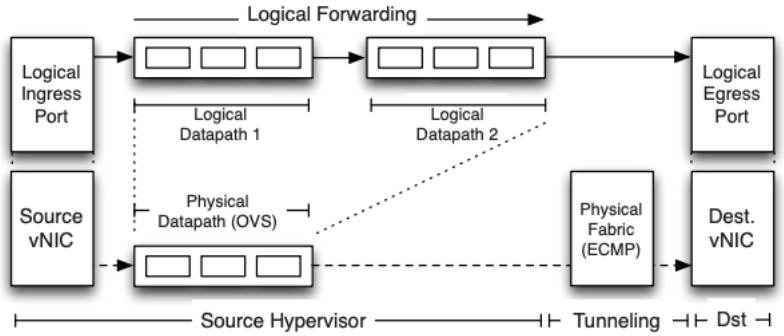
A centralized controller cluster translates the logical (abstract) pipeline into match-action tables executed in the vSwitch.

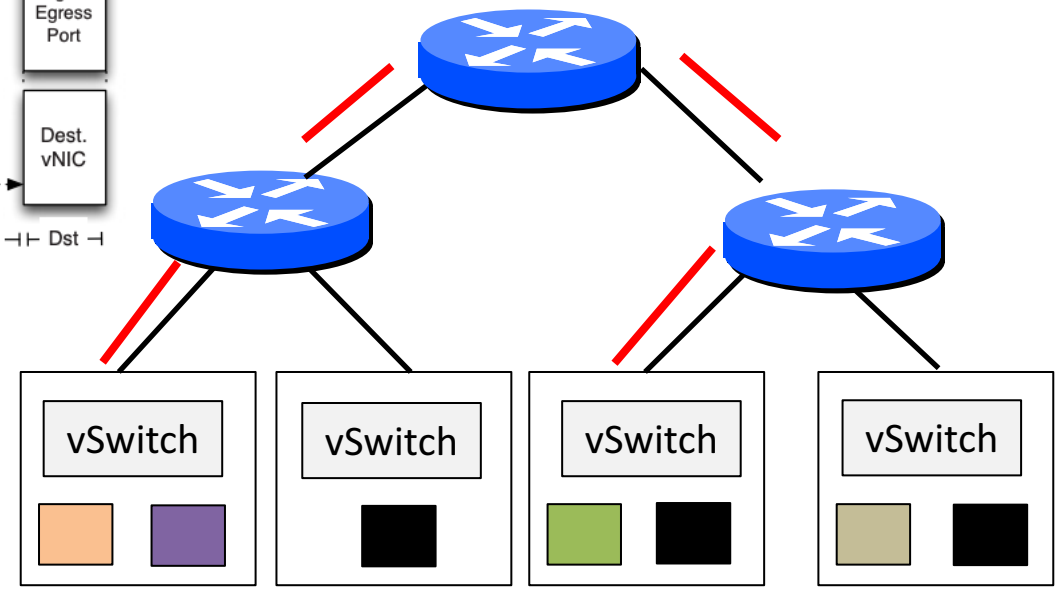




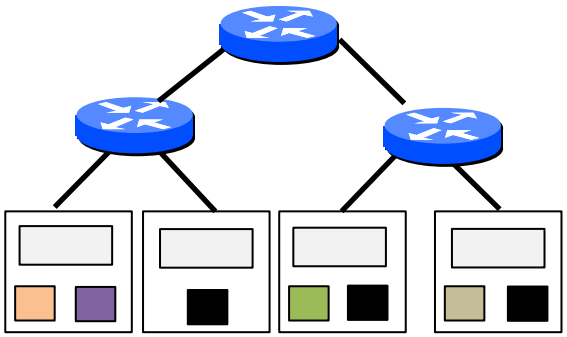
vSwitches provide the abstraction

- vSwitches transparently tunnel packets between VMs on different servers.
- Physical switches simply route packets from one server to another using IP-routing.

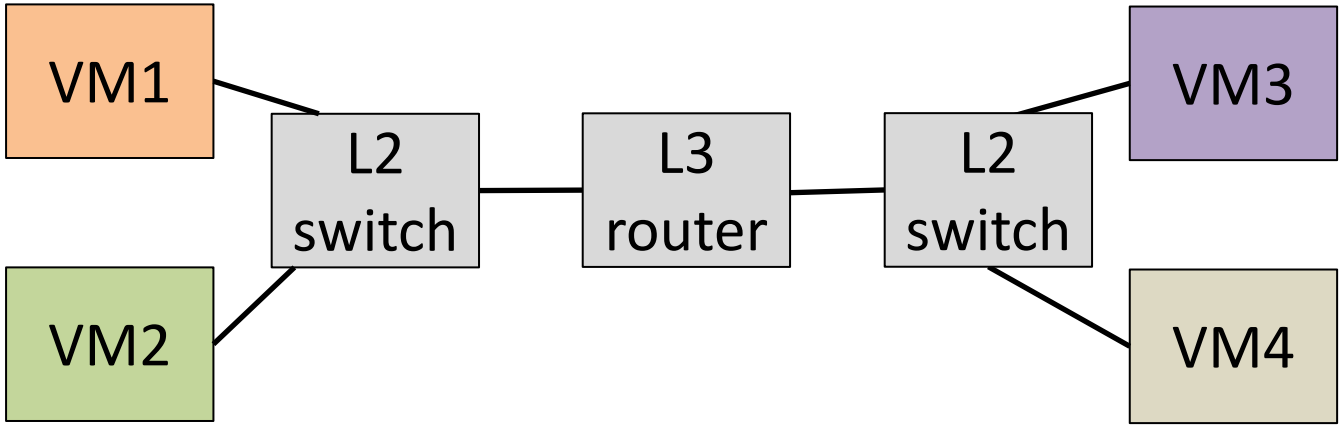




More complex datapaths



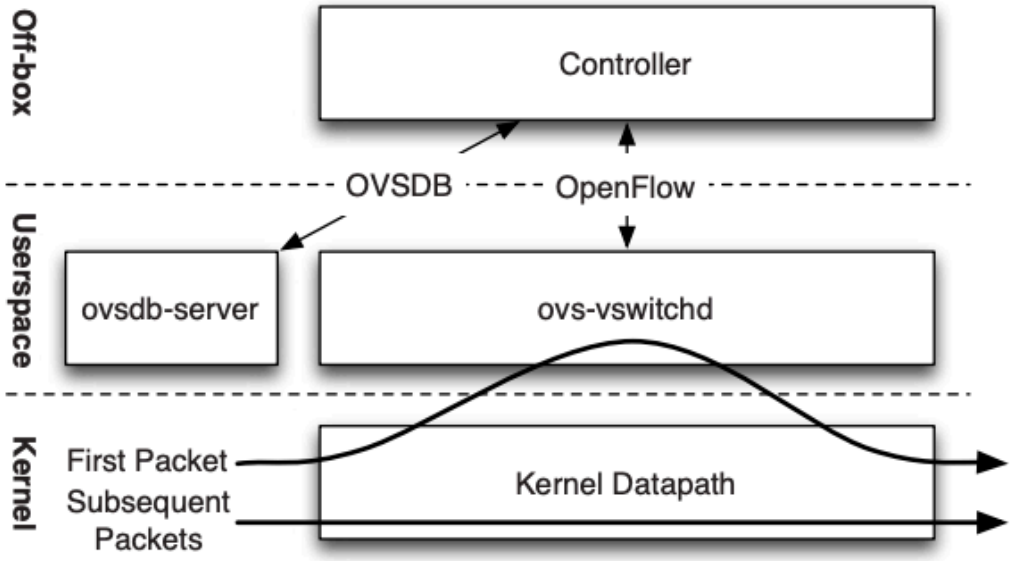
Physical Topology



Abstraction

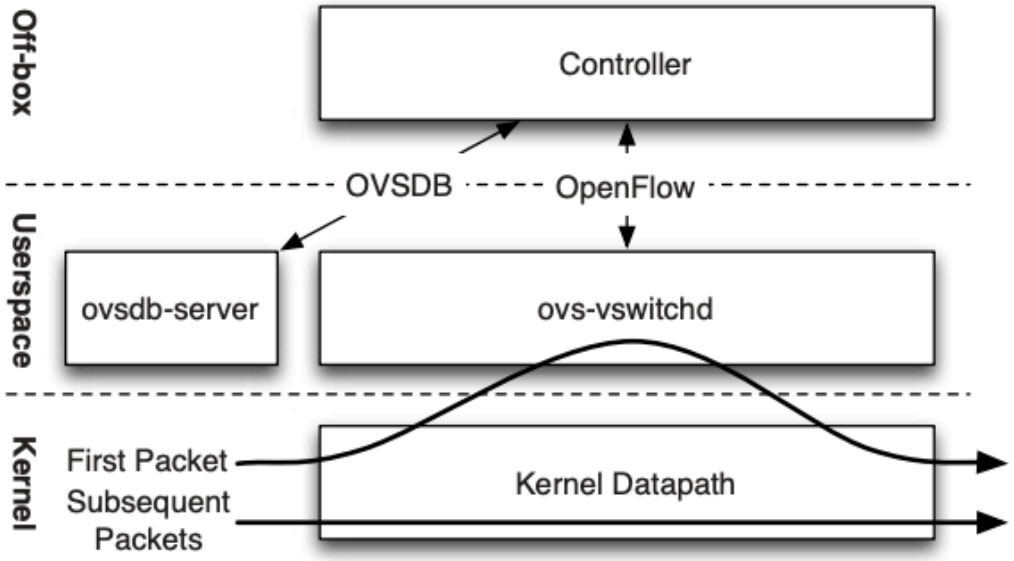
OpenvSwitch (OVS)

- An open-source virtual switch developed at Nicira (acquired by VMWare which in turn got acquired by Broadcom).
- One of the most successful SDN product.



OpenvSwitch (OVS)

- Controller-vSwitch interface: familiar OpenFlow + OVS config. protocol.
- First packet sent to userspace: walks through the entire logical datapath.
- Subsequent packets handled by an exact match flow table in the kernel.

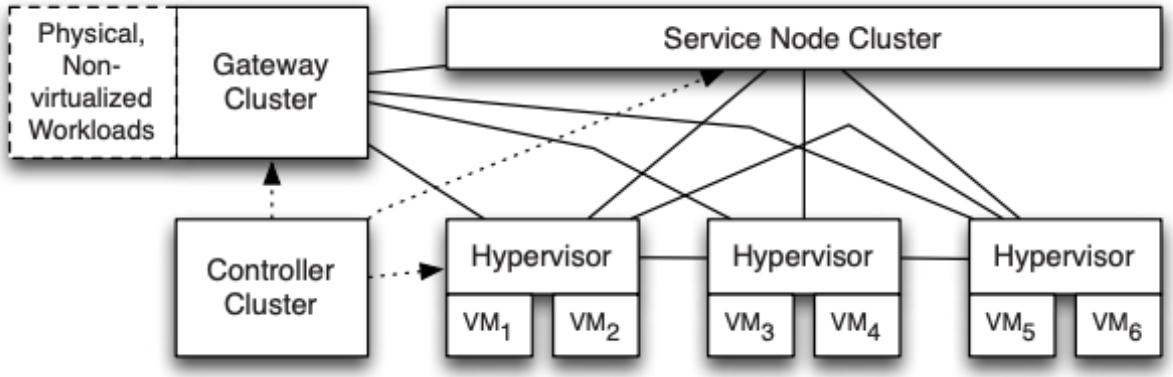


OpenvSwitch 10 years later

- SIGCOMM'21
- Tight kernel user-space coupling hinders innovation and impacts performance.
- Bring most packet processing into userspace using express datapath (XDP).
- An alternative was to use OVS over DPDK, but that was incompatible with commonly used tools and systems.

Two special transport nodes

- Broadcast/multicast carried out by "service node" overlay.
- Gateway nodes handle traffic to/from tenant's physical network.
- Replication for fault-tolerance.
  - flow-hashing based load balancing across service nodes.
  - leader election among gateway nodes.



Tunneling packets between servers

- Challenge: how to support TSO/LRO?
- Solution:
  - STT: Fake (stateless) TCP header after outer IP header.
  - Issues?

Controller Scalability

- Multiple controller instances.
  - multiple logical controller instances and multiple physical controller instances.
  - also enables fault-tolerance.

- Enable incremental changes to flow tables.
  - use of a Datalog inspired domain-specific language (nlog).

Proactive Model

- Controllers "push" all necessary forwarding state to the hypervisors.
- Controllers do not process any packets.

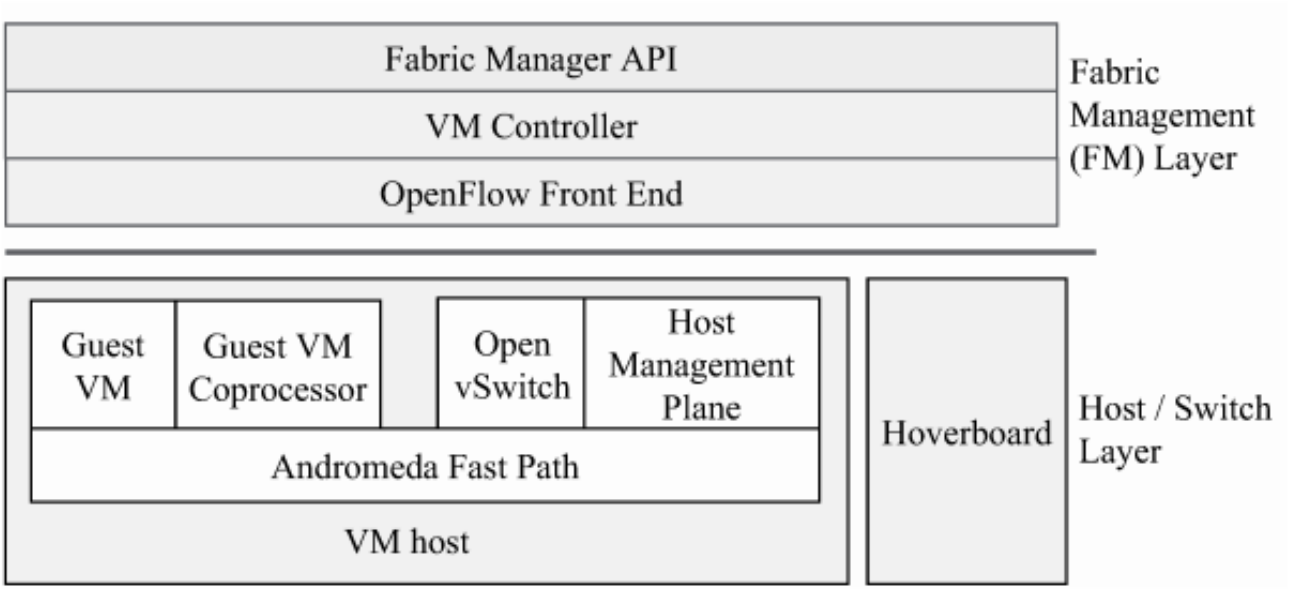
Your thoughts?

- What did you like about the paper?

- What are its limitations?

Network Virtualization at Google

- Andromeda (NSDI'18)



- SNAP and PonyExpress (SOSP'19): Google's host-networking stack

Network Virtualization at Microsoft

Virtual Filtering Platform (VFP), NSDI'17

- Overcomes certain limitations of OVS
  - Support for stateful actions.
  - Customized encapsulation/decapsulation
  - User-defined actions.

AccelNet (NSDI'18): offload on customized hardware to make use of SR-IOV.