Cloud Networking (VITMMA02)
Server Virtualization
Data Center Gear

Markosz Maliosz PhD

Department of Telecommunications and Media Informatics
Faculty of Electrical Engineering and Informatics
Budapest University of Technology and Economics

Spring 2017

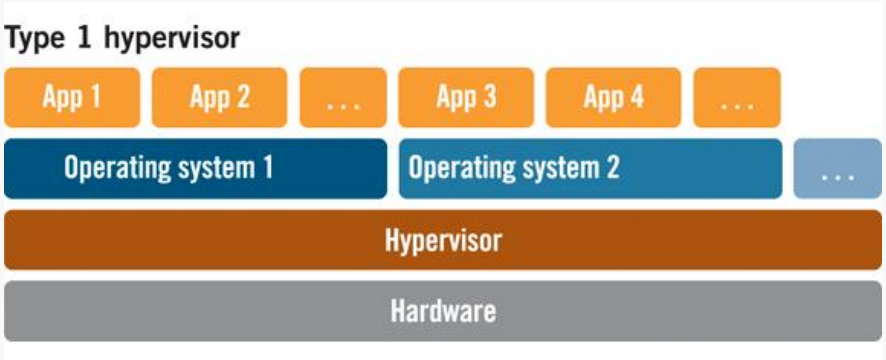
SERVER VIRTUALIZATION

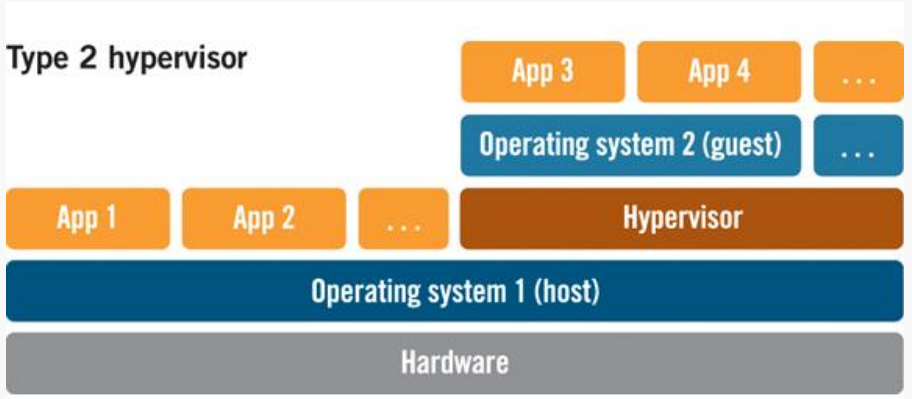
Server Virtualization

- Low server utilization ⇨ virtualization
  - PCs, servers: 10%
  - storage: 50%
- Server CPU and network bandwidth utilization is growing
  - Today: 2-4(-10) VM / physical server (Virtual/Physical Machine)
  - forecast for 2018: 16 VM/PM
  - for processes with low resource requirements even 100 VM/PM
- Hypervisor
  - virtual machine monitor/manager (VMM)
  - terminology
    - hardware: host
    - VM: guest
  - running VMs on the host
    - separated memory and disk management, CPU scheduling

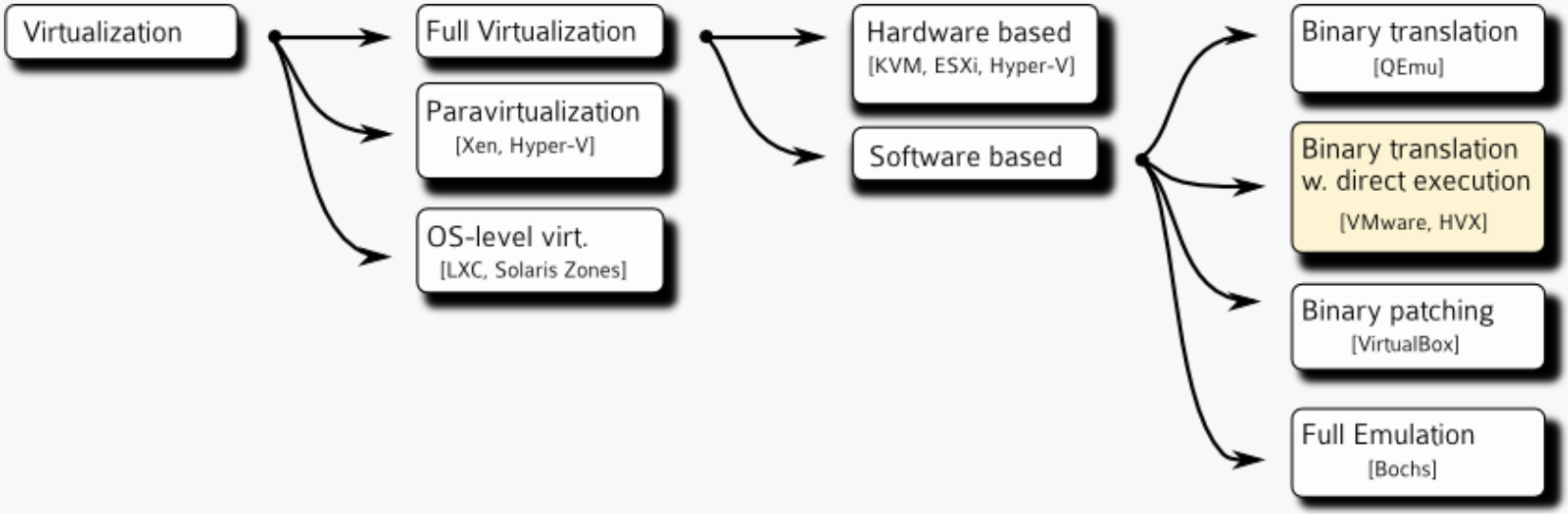
Server Virtualization

- Hypervisor
  - types, taxonomy: Gerald J. Popek and Robert P. Goldberg, „Formal Requirements for Virtualizable Third Generation Architectures”, 1974
    - Type 1: native (bare metal)
      - hypervisor is running directly on the hardware
      - e.g. Citrix XenServer, VMware ESX/ESXi, Microsoft Hyper-V
    - Type 2: hosted
      - hypervisor is running on the host OS (VM: guest)
      - e.g. VMware Workstation/Player, VirtualBox
    - other: Linux Kernel-based VM (KVM)
      - running as a kernel module, host OS is converted to Type 1
      - usually classified as Type 2





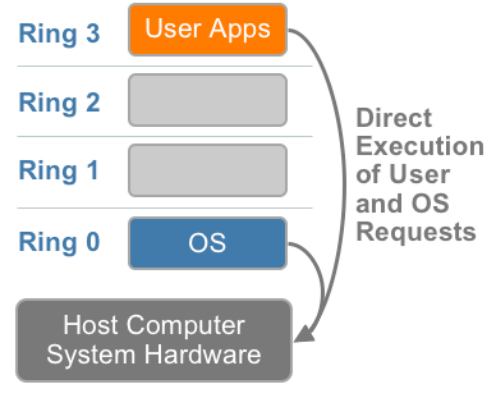
Types of Virtualization

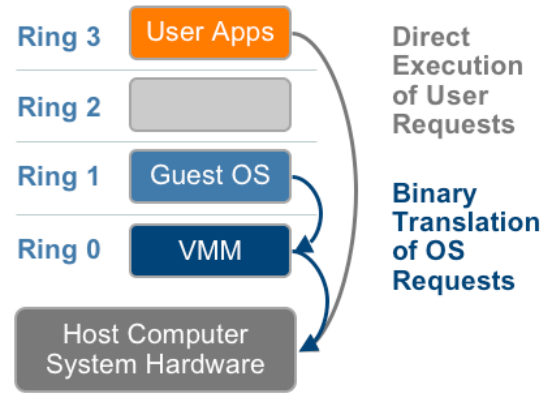


CPU Virtualization

- 2005-2006 hardware support for CPU virtualization: Intel VT-x and AMD-V
- Spread of virtualization software
- x86 CPU virtualization

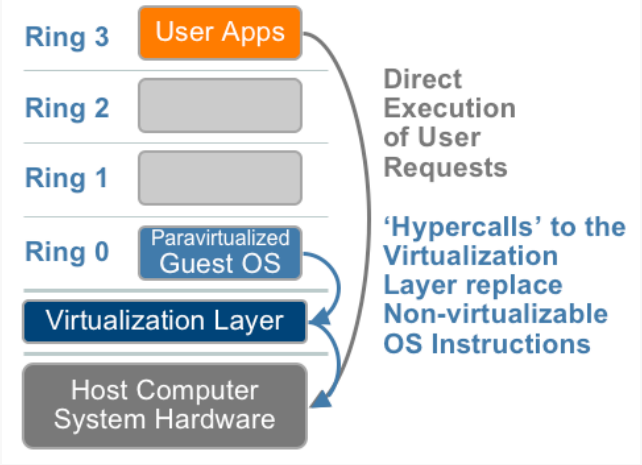
Privilege levels without virtualization

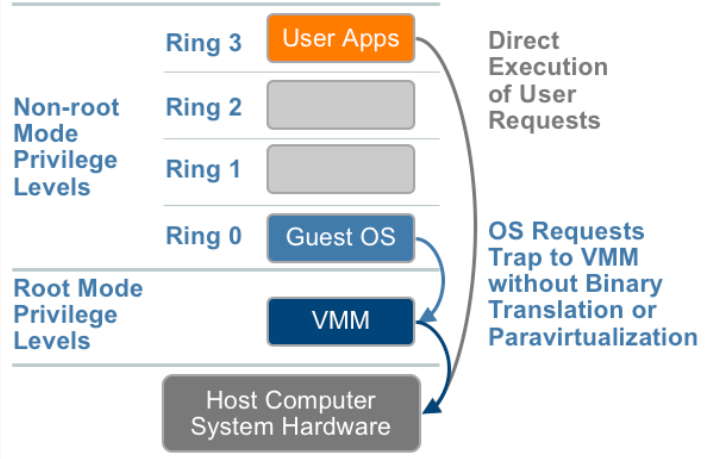




Full virtualization with dynamic binary instruction set translation

Paravirtualization (static translation)



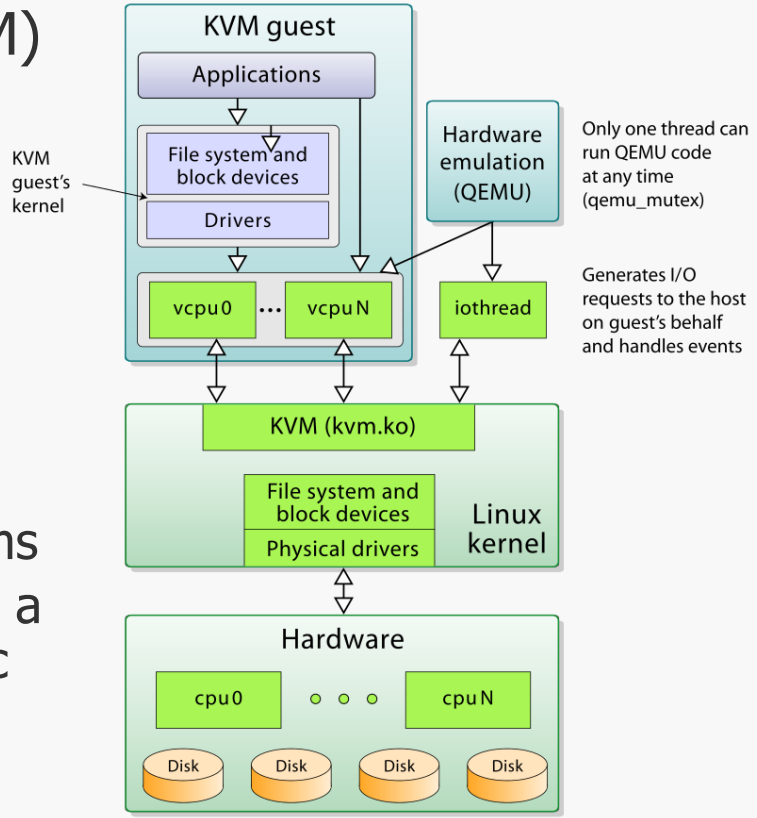


Hardware - assised virtualization

Source of figures: Vmware, Understanding Full Virtualization, Paravirtualization, and Hardware Assist, White Paper, 2007

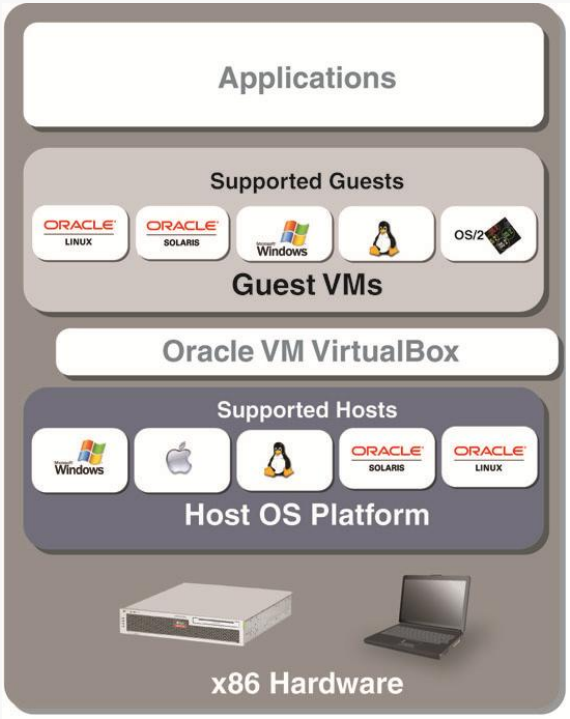
Platform Virtualization Software

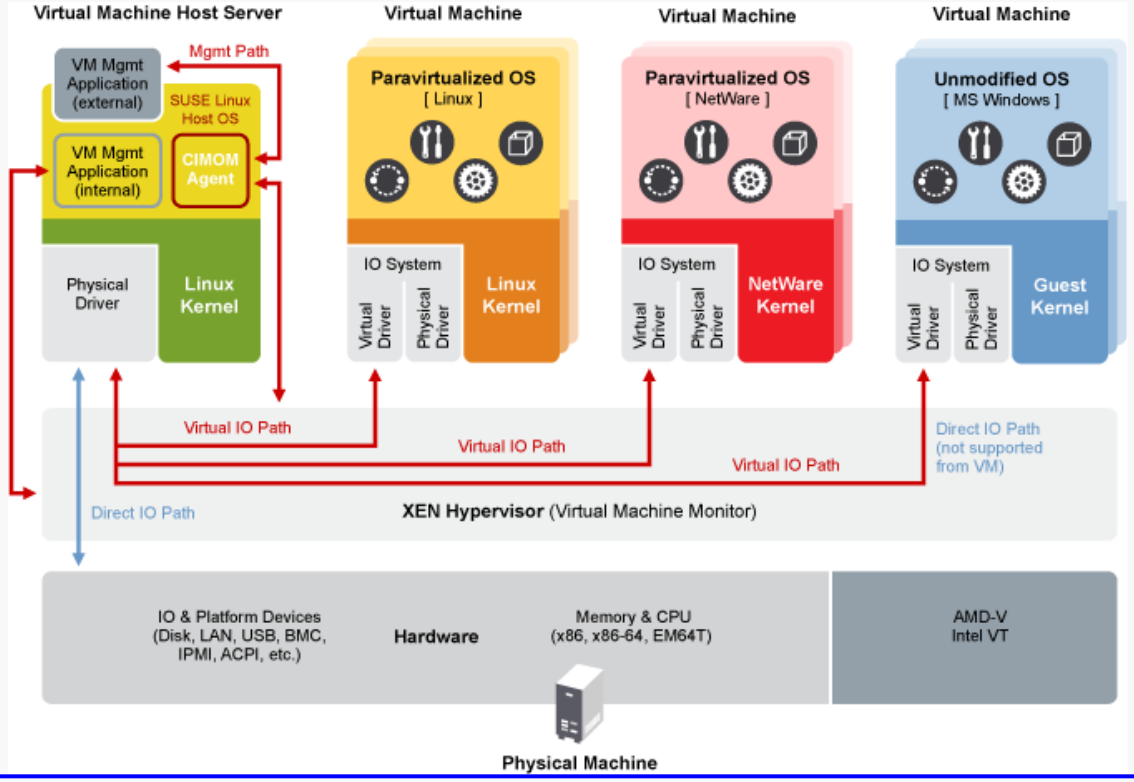
- Free and Open Source Software
  - Kernel-based Virtual Machine (KVM)
    - Type 2
    - part of Linux kernel
    - requirement: hardware-assisted virtualization
  - QEMU
    - generic and open source machine emulator and virtualizer
    - emulation: can run OSes or programs
    - programs made for one machine on a different machine, by using dynamic translation
    - virtualization: Xen or KVM
      - if host and guest is the same arch.
      - otherwise only software virtualization



Platform Virtualization Software

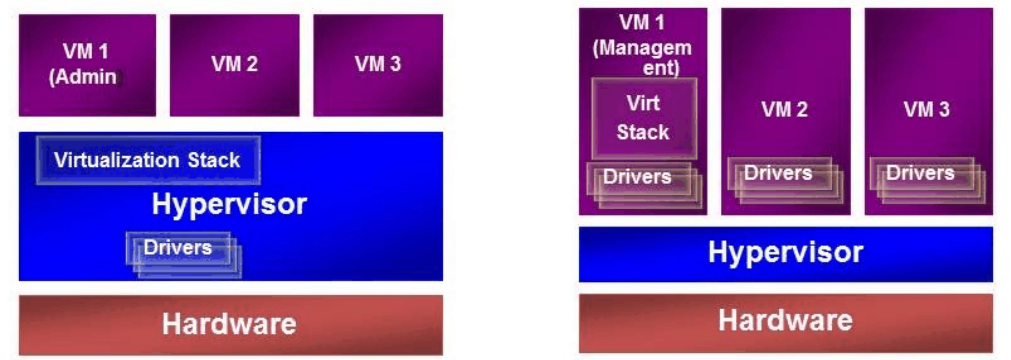
- Free and Open Source Software
  - Oracle VirtualBox
    - Type 2
    - software- or hardware-assisted virtualization
  - Xen
    - Type 1
    - paravirtualization or hardware-assisted virtualization





Platform Virtualization Software

- Closed Source / Commercial Products
  - VMware ESXi
    - Type 1
    - paravirtualization or hardware-assisted virtualization
    - small size: approx. 200 MB
    - monolithic VMkernel
      - hypervisor contains and manages all kind of device drivers
  - Microsoft Hyper-V
    - Type 1
    - partitions
      - parent partition (Admin, Management) : x86-64 Windows Server
      - child partitions : VMs
    - paravirtualization or hardware-assisted virtualization
    - larger size: approx. 5GB core, or 10GB full
    - mikorkernel
      - device drivers at VM level



NETWORKING IN DATA CENTERS: NETWORK DEVICES

Data Center Network

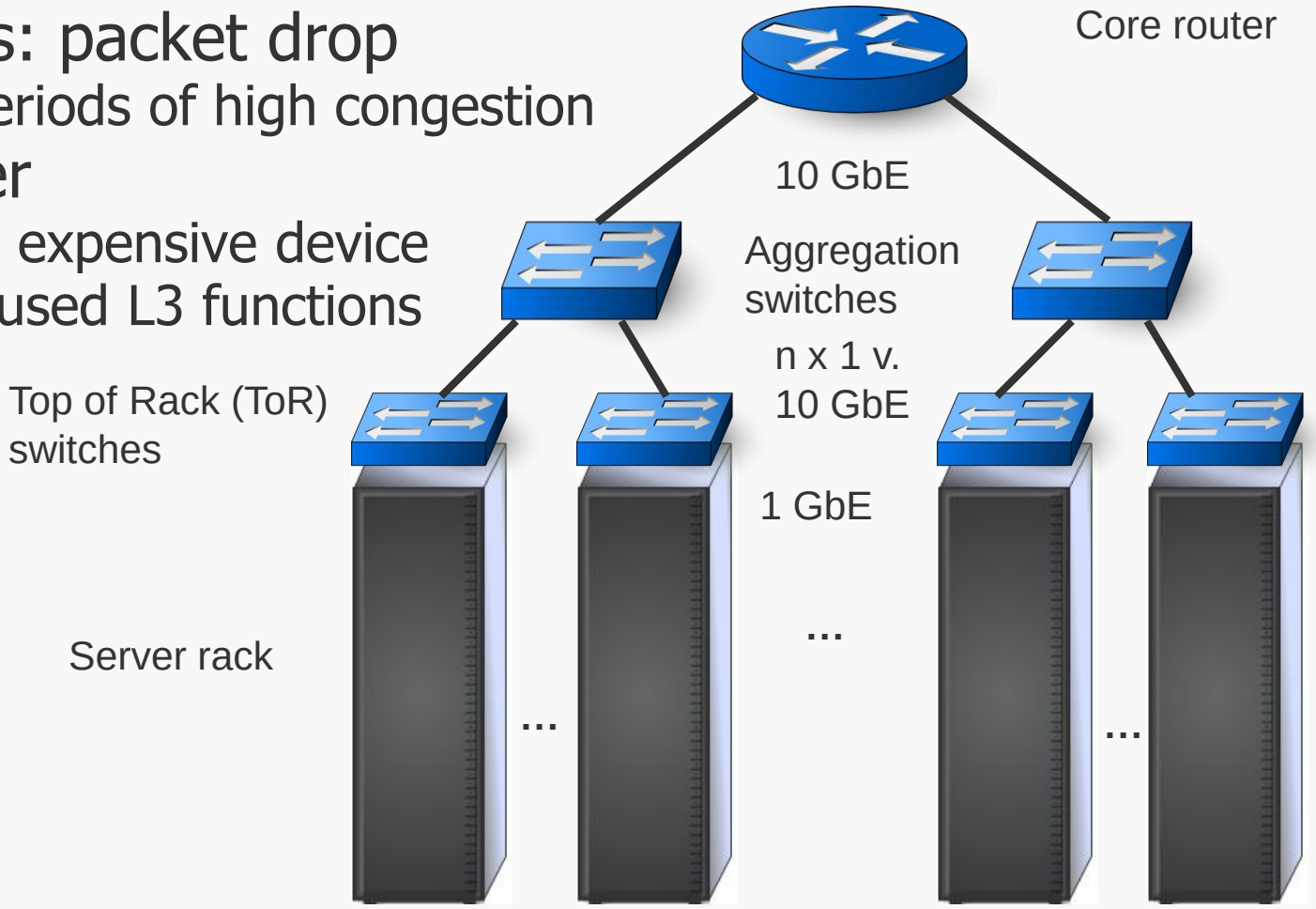
- Servers arranged in racks
  - several 10 or 100 thousand servers
- Reducing Capital and Operating Expenses (CapEx, OpEx)
- Network devices
  - Network Interface Card – NIC
  - switch, bridge
  - router
  - cabling (copper or fiber optic)
  - new trend
    - build network hardware according to custom specification and add own software (e.g. Google) ⇨ Software Defined Networking (SDN)

Data Center Network

- Why Ethernet?
  - high performance-to-cost ratio
  - ease of configuration
  - high speed: 1, 10, 40 (, 100) GbE
  - 10 GbE
    - since 2006 devices appear on the marker
    - since 2012 wide spread use
  - storage network traffic
    - Fibre Channel over Ethernet (FCoE)
- Challenges
  - different requirements compared to the LAN: scalability, reliability, bisection bandwidth, automated address allocation
- Convergence
  - not only data communication, but also storage network traffic is on the same network
    - data loss not tolerated
    - minimum bandwidth guarantees

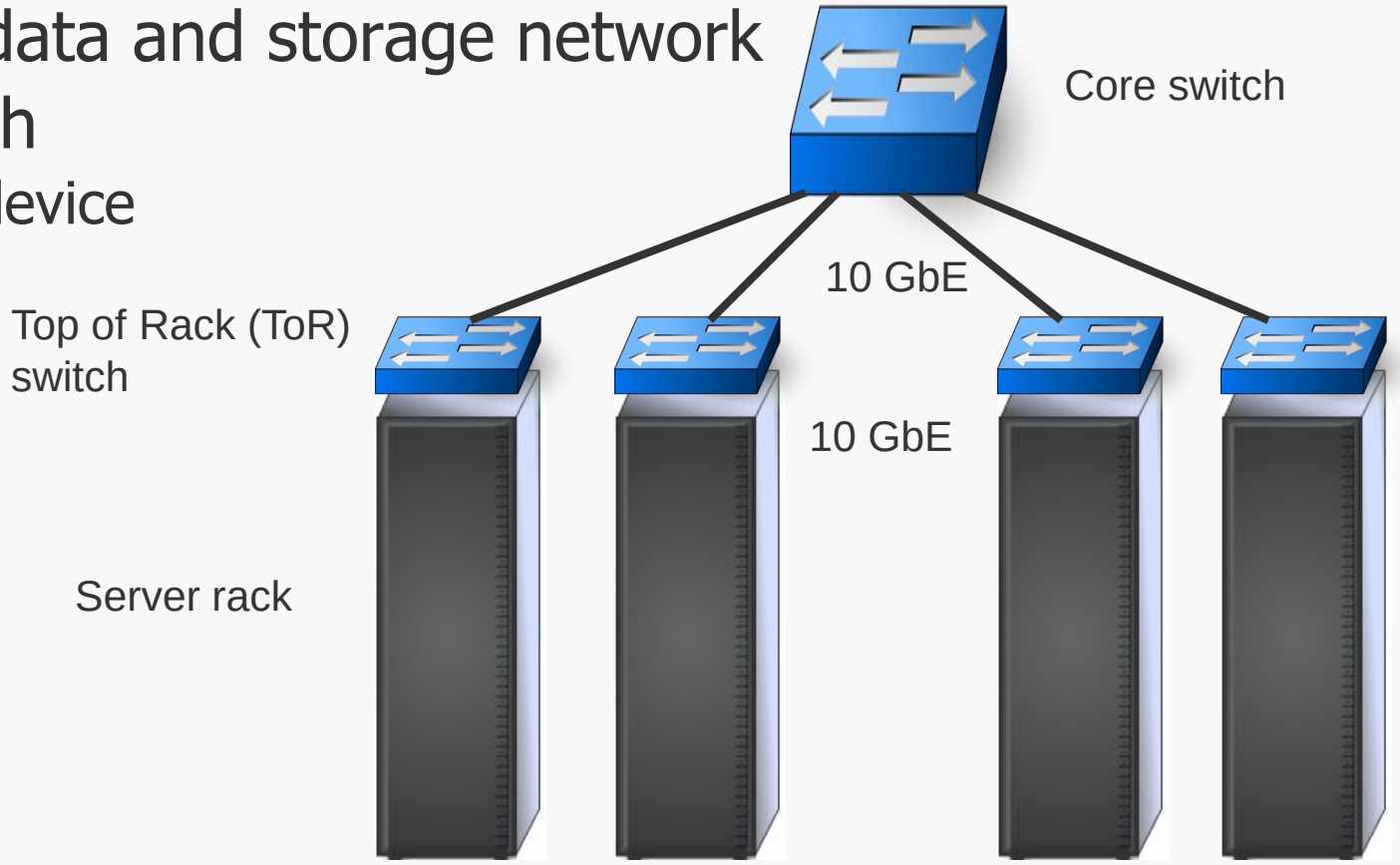
Enterprise Data Center

- Traffic between servers over multiple devices (hops)
  - latency, latency variation
- Traffic loss: packet drop
  - during periods of high congestion
- Core router
  - complex, expensive device
  - many unused L3 functions



Cloud Data Center

- Traffic between servers over few hops
  - flat(ter) network topology
  - lower latency and latency variation
- Common data and storage network
- Core switch
  - simpler device

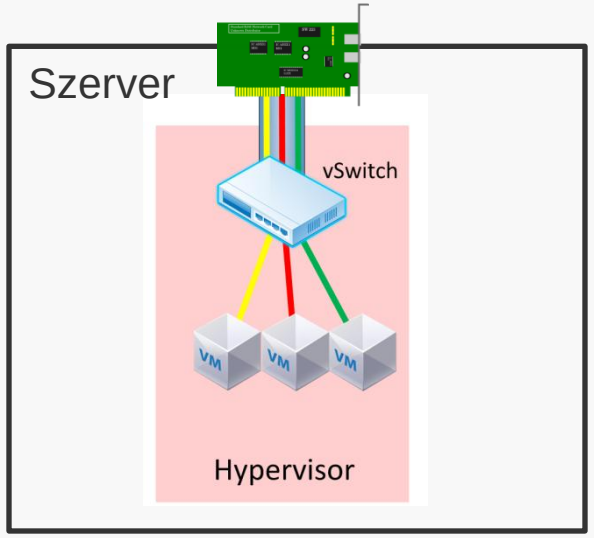


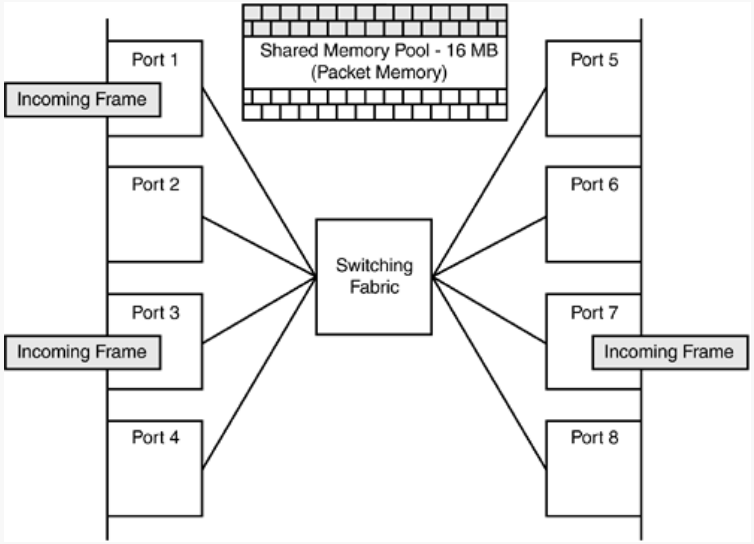
Data Center Switch Types

- Virtual Switch, vSwitch
  - between VMs on the same physical server
- Top of Rack (ToR) switch
- End of Row (EoR) switch
- Aggregation switch
- Core switch/router

Virtual Switch

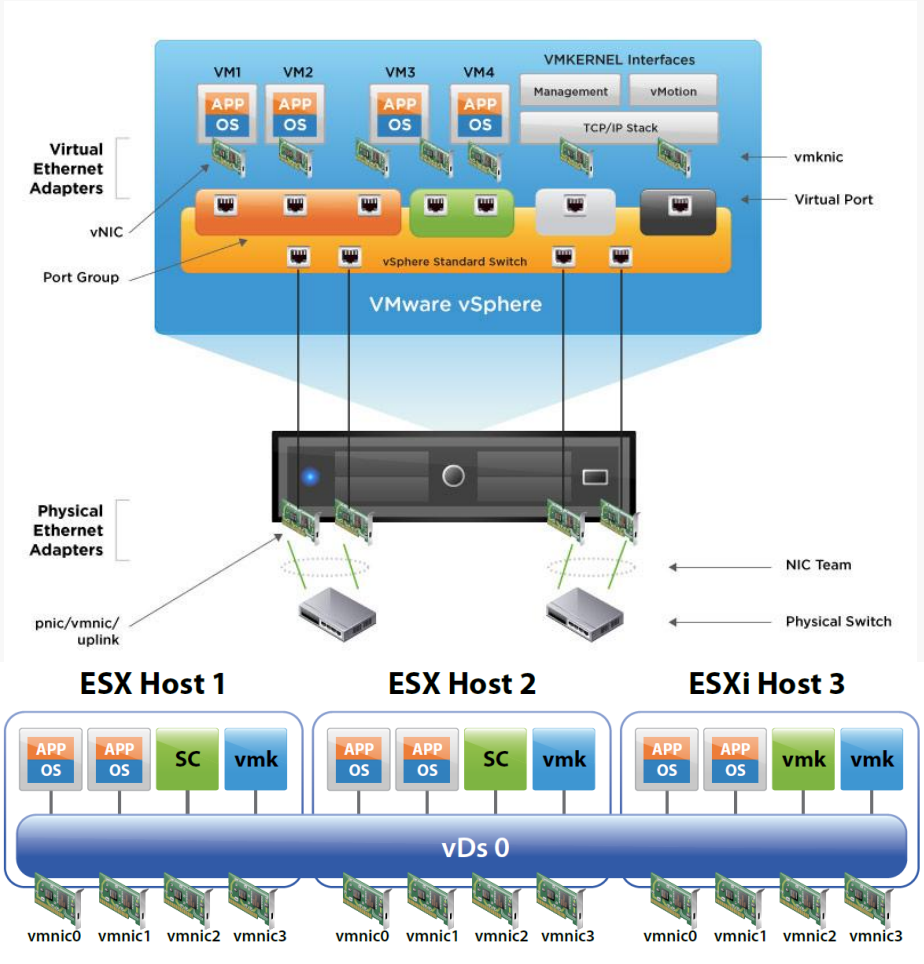
- Hypervisor
  - configuration of VMs and virtual switches
- vSwitch
  - attached to physical NIC, typically all VMs are connected to the same vSwitch
    - limiting the bandwidth of the VMs input/output traffic
  - server CPU is used for switching
  - in practice the vSwitch is implemented with shared memory
    - data (frames, packets) is stored in memory of the servera
    - VMs exchange pointers to this data
    - high bitrate!
  - also part of the network
    - uniform configuration and management would be ideal with the physical switches





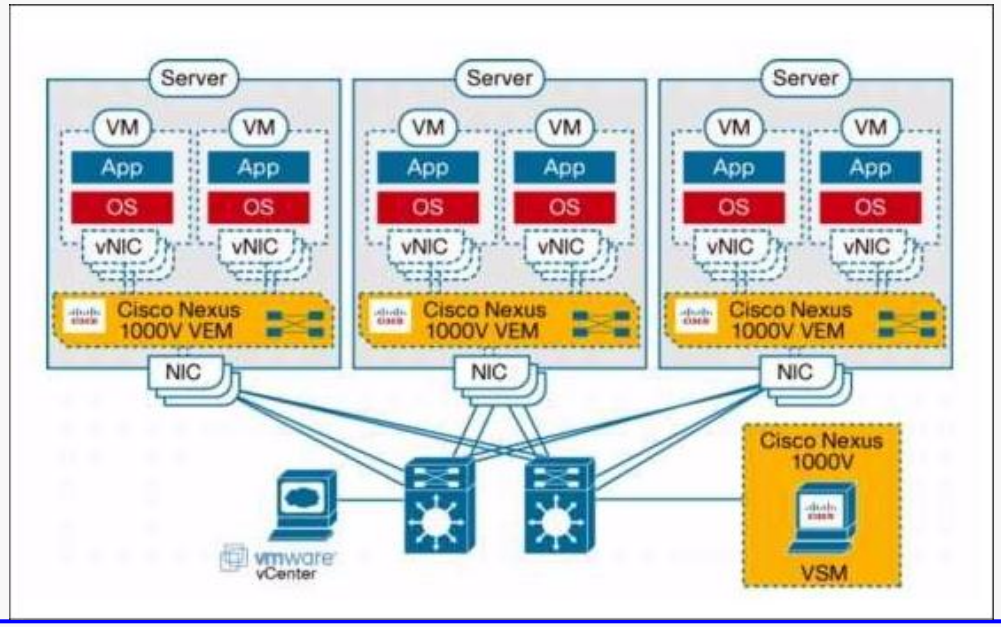
Virtual Switching: VMware vSwitch

- multiple VM – one (or multiple) NIC
- software vSwitch – hypervisor
  - VMware vSphere (ESXi)
    - vNetwork Standard Switch (VSS)
    - abstract, distributed switch: vSphere Distributed Switch (VDS)
      - unifying multiple physical servers
    - New features (with significant CPU usage)
      - traffic monitoring
      - VLAN isolation
      - traffic shaping (max. bandwidth)
      - vMotion support



Virtual Switching: VMware vSwitch

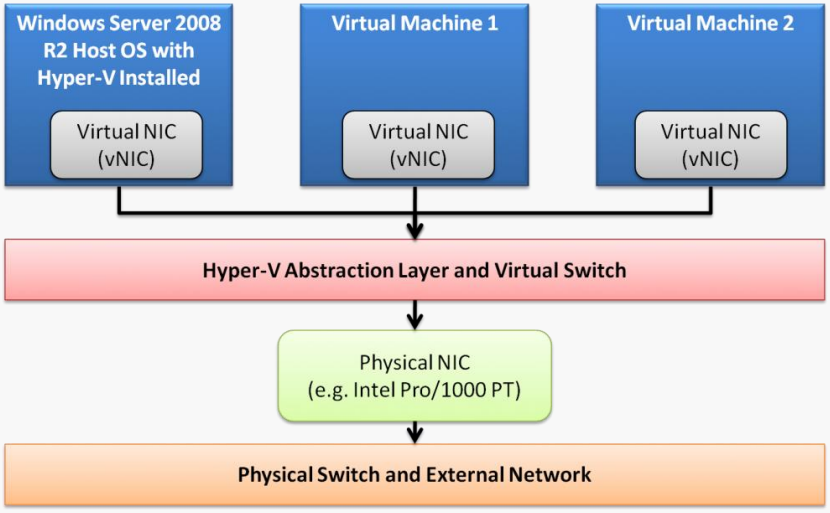
- VMware vSphere (ESXi)
  - Cisco Nexus 1000V
    - Cisco/VMware collaboration
    - feature set is close to a physical switch, but with limitations
    - components
      - Virtual Ethernet Module: runs inside the hypervisor
      - Virtual Supervisor Module: manages VEMs
    - integrated Cisco Command Line Interface (CLI) and VDS API
    - VXLAN support

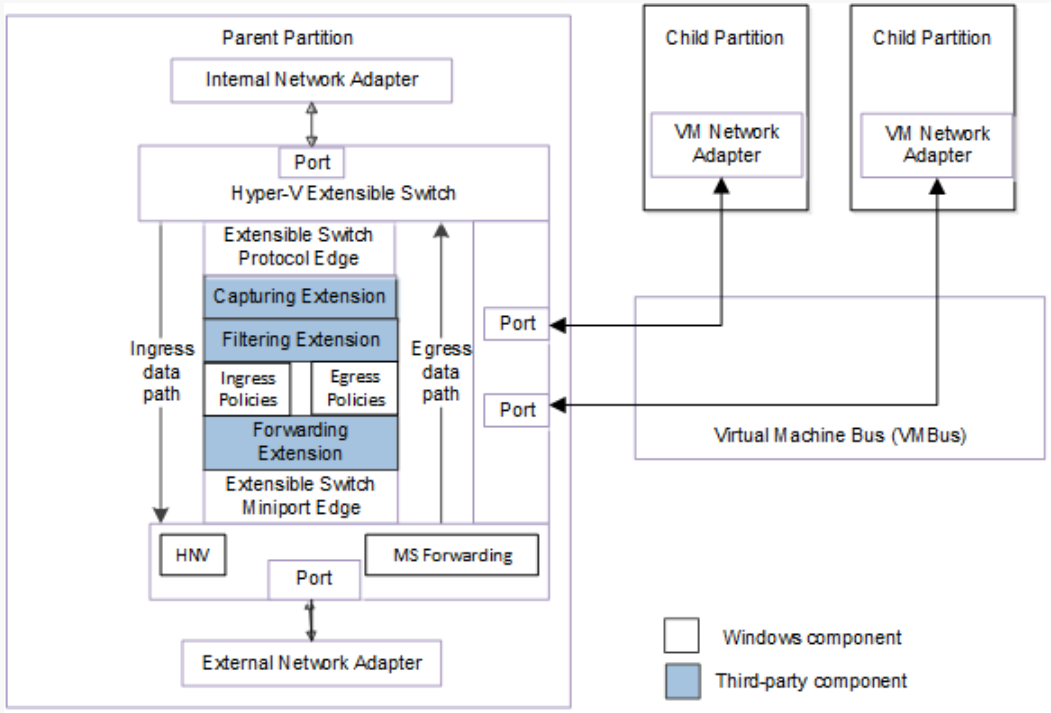


Virtual Switching: Microsoft Hyper-V

- Private/Internal/External modes
- Hyper-V 3.0: Windows Server 2012
  - Hyper-V Extensible Switch
  - traffic classification, filtering and monitoring
  - guarantee a minimum and/or limit the outbound speed
  - congestion control
  - VM queues
  - live migration
  - extensibility
  - Cisco Nexus 1000V can be integrated

Hyper-V Networking Basic Diagram

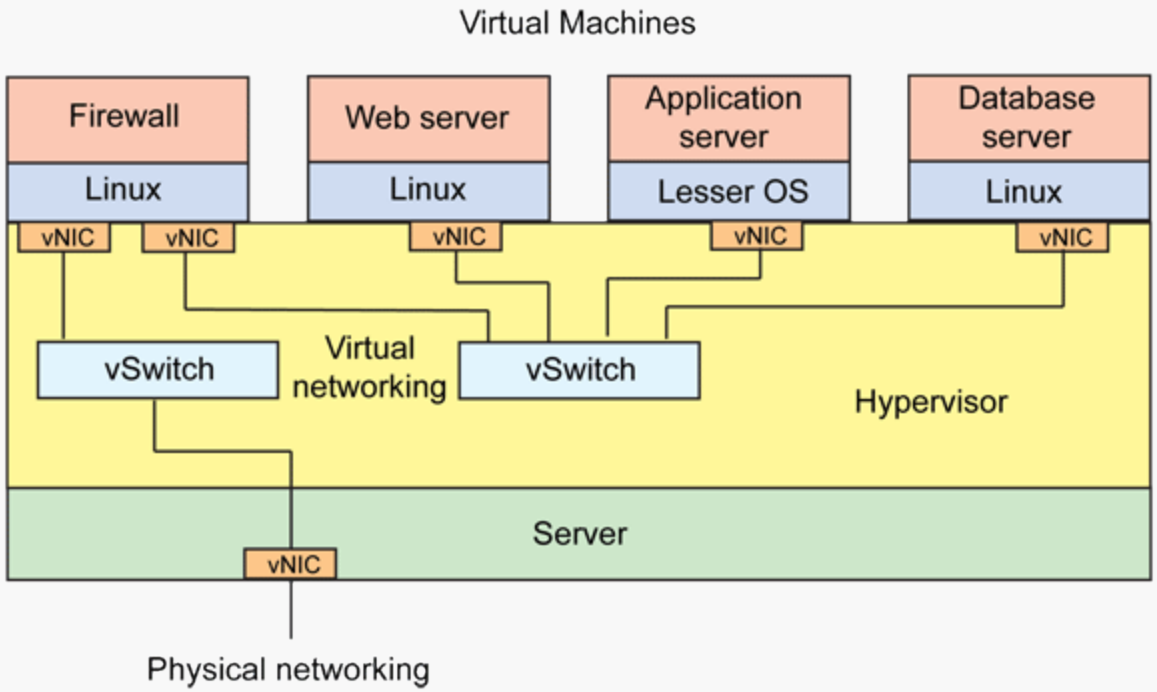




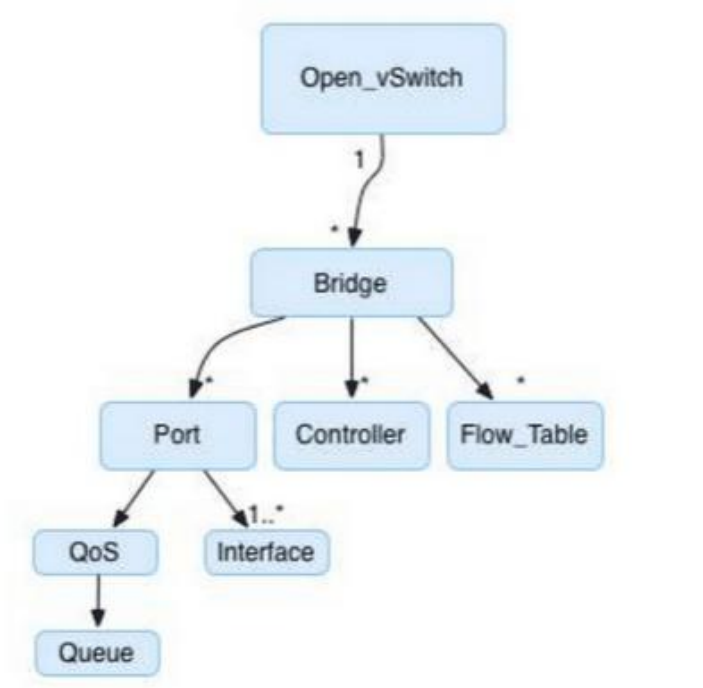
Virtual Switching: Open vSwitch

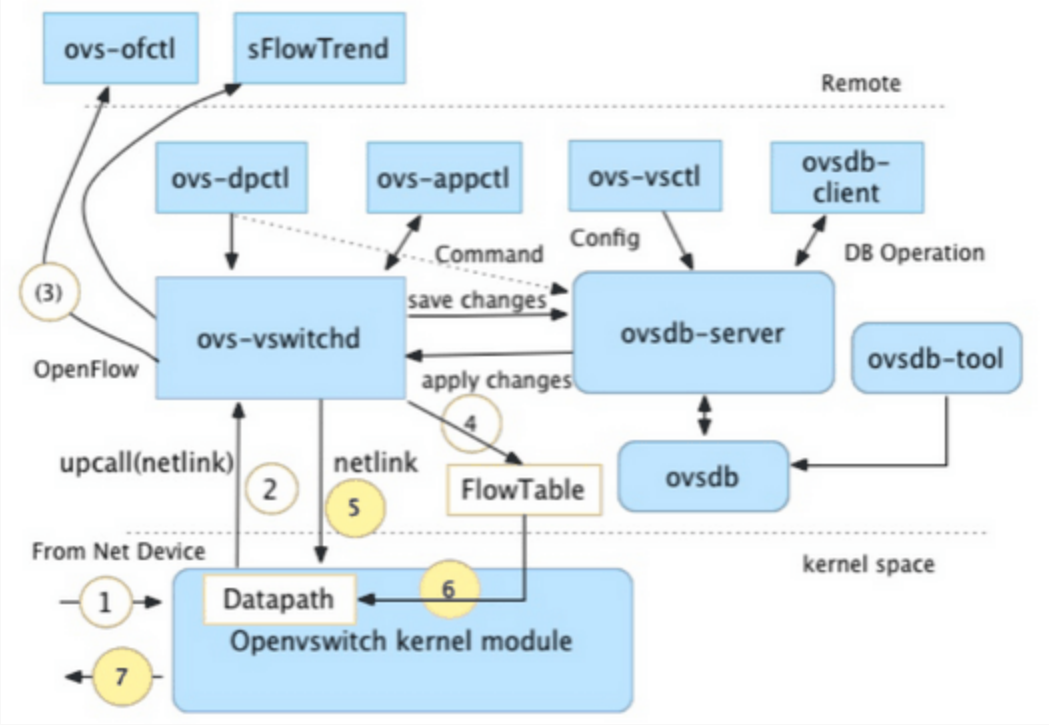
- open source implementation
- standard management protocols
- Features
  - Visibility into inter-VM communication via e.g. NetFlow
  - 802.1Q VLAN
  - STP (IEEE 802.1D-1998)
  - QoS control
  - Per VM interface traffic policing
  - NIC bonding
  - OpenFlow protocol support (including many extensions for virtualization)
  - IPv6 support
  - Multiple tunneling protocols (GRE, VXLAN, STT, and Geneve, with IPsec support)
  - Kernel and user-space forwarding engine options
  - user-space control
- Characteristics
  - Mobility of state: all network state (e.g. an entry in an L2 learning table, ACLs, QoS policy, etc.) associated with a network entity (say a virtual machine) should be easily identifiable and migratable between different hosts
  - Responding to network dynamics: VM startup, shutdown, migration
  - Maintenance of logical tags: tunneling, VM identification
  - Hardware integration: can be the control plane of a hardware switch
  - distributed vSwitch: with OpenFlow
- Platforms
  - XenServer, Xen, KVM, VirtualBox, OpenStack, OpenNebula, Linux (kernel), FreeBSD

Virtual Switching: Open vSwitch



Open vSwitch data structures and architecture



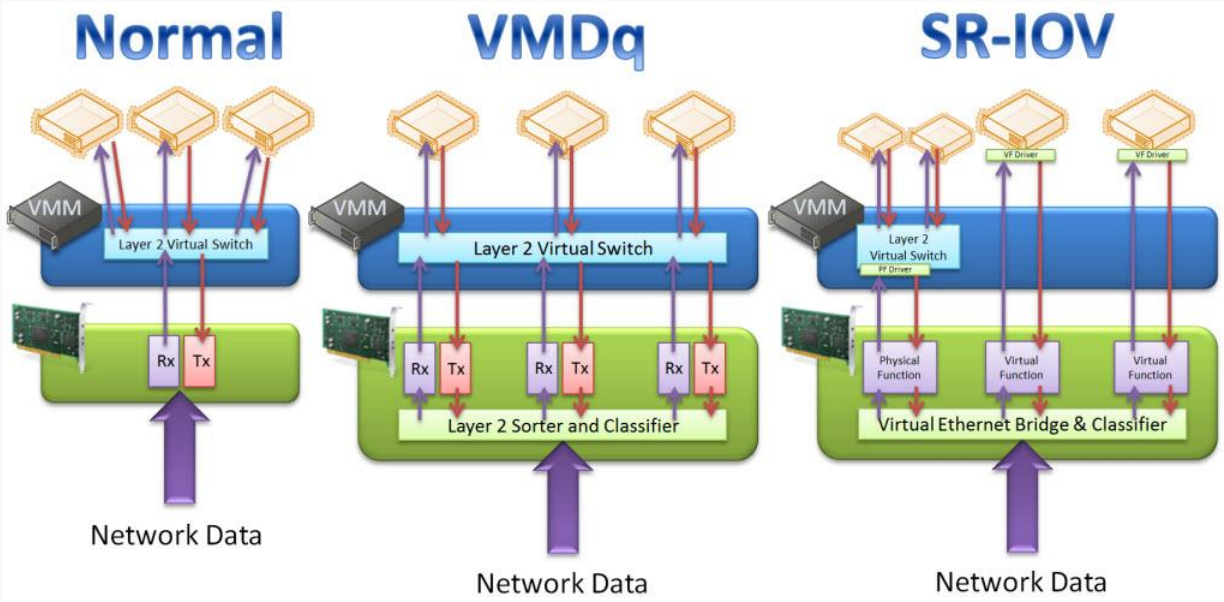


Linux Bridge

- alternative switch technology
- simpler than OVS
  - simple model, not flow-based
  - forwarding layer in the Linux kernel
  - less code-base
  - easier to troubleshoot
  - set-up: brctl, ip route
- Tunneling
  - supported GRE Tunnels
  - VXLAN support: from Linux kernel 3.7 (2012)

Enhancing VM Network Performance

- Virtual machine device queues (VMDq) – Intel
  - intensive CPU usage by vSwitch affects performance of the VMs
  - VMDq implemented on NIC
  - separate receive and transmit queues for VMs
    - based on MAC address and VLAN tag information
  - advantages
    - parallelization
    - filtering and sorting of packets is done by hardware
- PCIe single-root IO virtualization (SRIOV)
  - moves the vNIC functionality into the NIC
  - one physical function (PF) with multiple virtual functions (VFs)
  - vSwitch is by-passed, therefore it is used only in special applications
  - DMA between VF and VM
  - requires support from
    - VM network driver
    - hypervisor
  - approx. 10-15% CPU load reduction

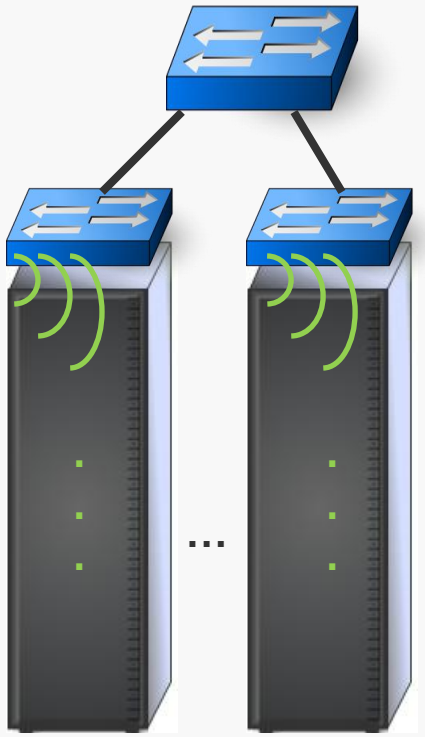


ToR Switch

- Typical configuration
  - servers connected using star topology
  - 48 x 10 GbE port towards servers
  - 4 x 40 GbE towards aggregation switche(s)
  - 480 Gbps ⇔ 160 Gbps
    - 3:1 oversubscription
      - typical values: 2,5:1 – 8:1
  - 1 rack – 1 switch
    - rack level redundancy
- ToR switch
  - low latency
  - large address space tables
  - also for storage traffic
- Possible extra functions
  - tunneling
  - filtering
  - metering
  - load balancing

Advantages:
- shorter and simpler cabling
- rack level management/redundancy

Disadvantages:
- more switches has to be managed
- scalability limits (STP, ports)
- control plane for each switch

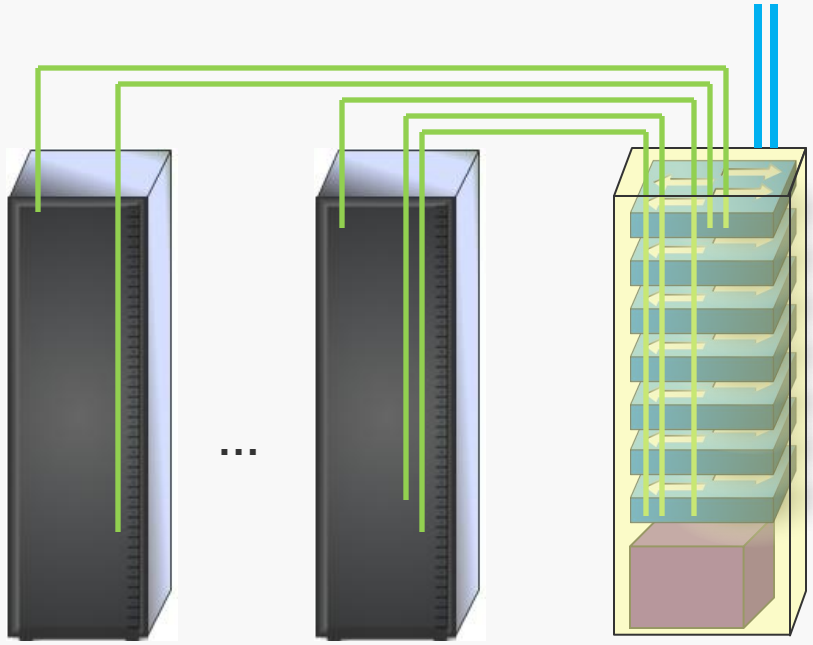


EoR Switch

- Cost reduction
  - large number of switch components together connecting servers and core switch
  - like switch cards plugged into a modular chassis
  - sharing common power, cooling and mgmnt. infrastructure
- Cabling
  - max. 100 m towards servers
  - cable distance ⇨ cost

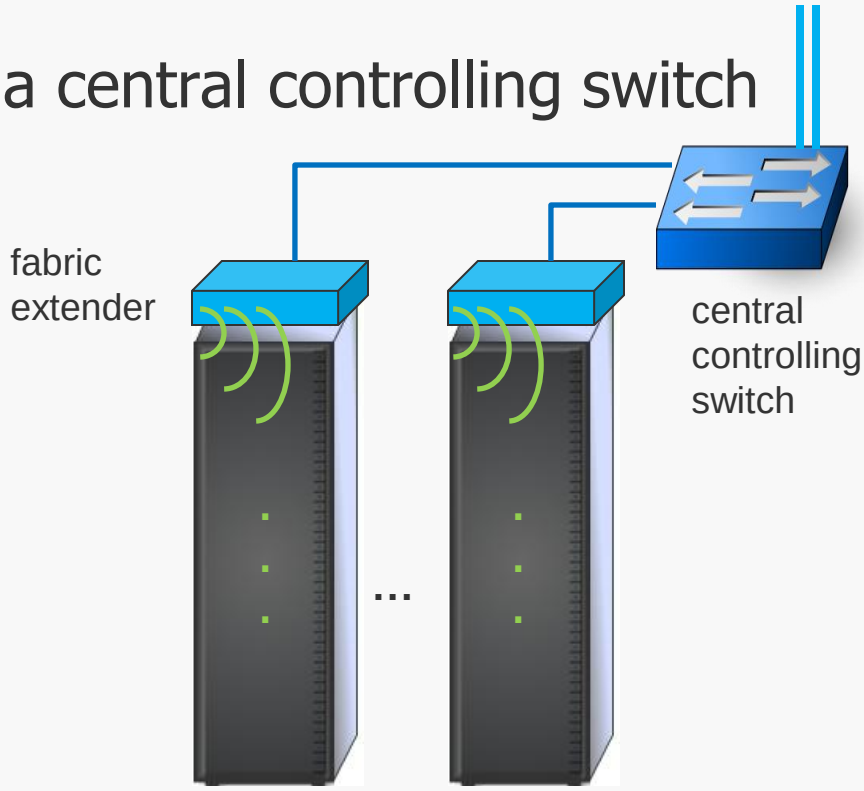
Advantages:
- central management processor
- less aggregation ports
- less STP instance
- one control plane

Disadvantages:
- expensive, inflexible cabling
- longer and hard to handle cables
- row level management/redundancy



Fabric Extenders (FEX)

- ToR / EoR hybrid solution
  - like a logically a distributed EoR switch
  - fabric extender on ToR: physical switch of limited functionality
  - controlled and monitored by a central controlling switch
  - cabling, like ToR switches



Advantages:
- central management processor
- less aggregation ports
- less STP instance
- one control plane
- cost effective cabling
- rack level management/redundancy

Disadvantages:
- not commonly used

Aggregation and Core Switches

- Aggregation: connecting ToR switches and Core switch
- Core Switch
  - connection to the outside network
  - high-bandwidth links and high port counts
  - modular design: cards attached to a common backplane
    - line cards (first stage)
    - switch cards (second stage)
    - processing cards (CPU, memory) – e.g. firewall, load balancer
    - management cards
  - complexity can be reduced by simple packet forwarding rules using various types of forwarding tags