Towards optimizing Network Utilization and Deployment in Virtualized environments

LinuxCon 2012, San Diego

Shyam Iyer

Dell| OS Advanced Engineering

- Network topology
- Virtualized ?
  - How many Virtual Machines do you run
    - On a single server ?
- How many NIC ports do you run on your server ?
  - Onboard NIC ports
  - Any additional NIC cards ?
    - 10G ? 40G?
  - CNA adapters ?
  - NPAR
    - Common NPAR devices support 4/8/16 NPAR functions
  - SR-IOV
    - Some commodity cards support upto 64 VFs

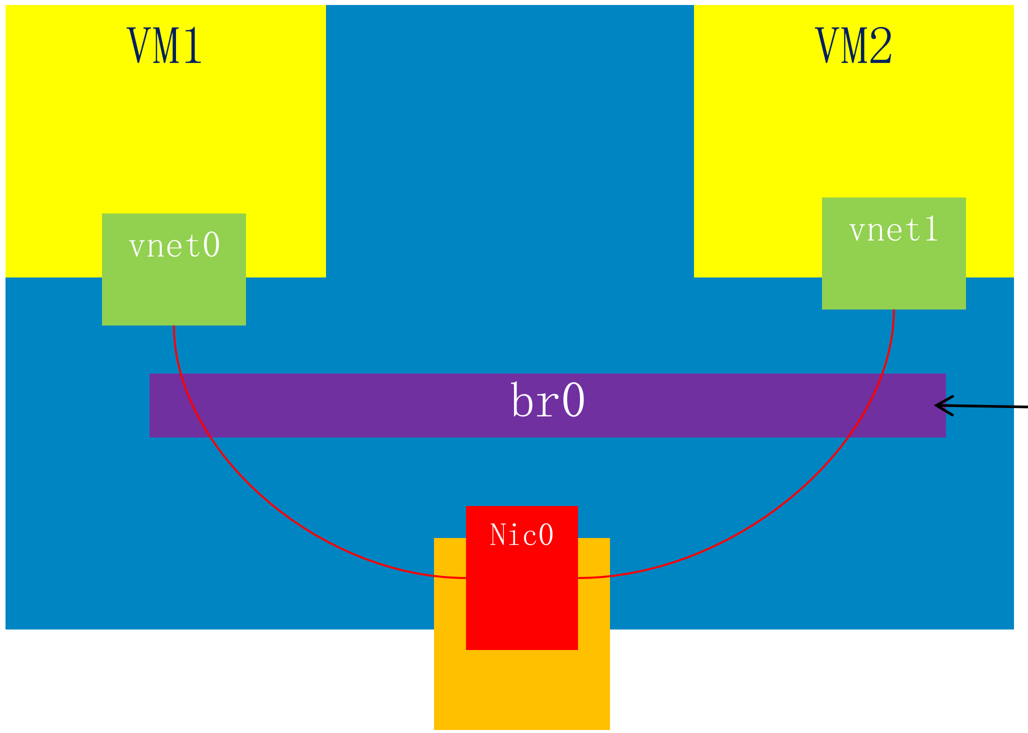
Management

- Are you running different workloads on the same fabric
  - Traditional LAN ethernet
  - iSCSI
  - FCoE
  - Infiniband
  - RDMA
- What is your management touchpoint ?
  - Server
    - Host operating system
    - Hypervisor management software
    - Tool provided by NIC/CNA vendor
    - Out of Band server management
  - Switches
    - Top of the rack switch
      - Access switch
      - Gateway switch
  - Storage

UseCases

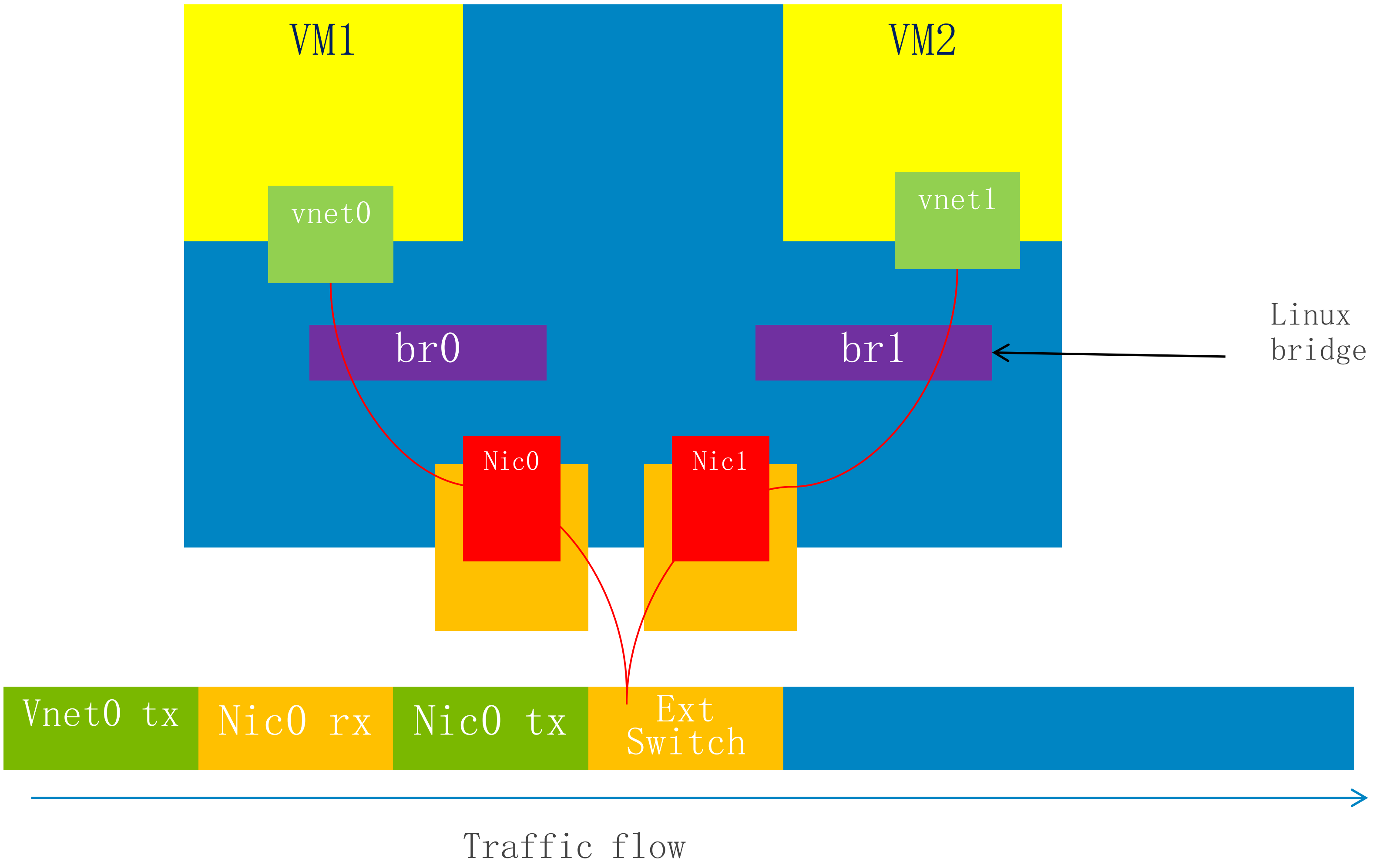
- They are not exhaustive
- Feedback

VM to VM communication

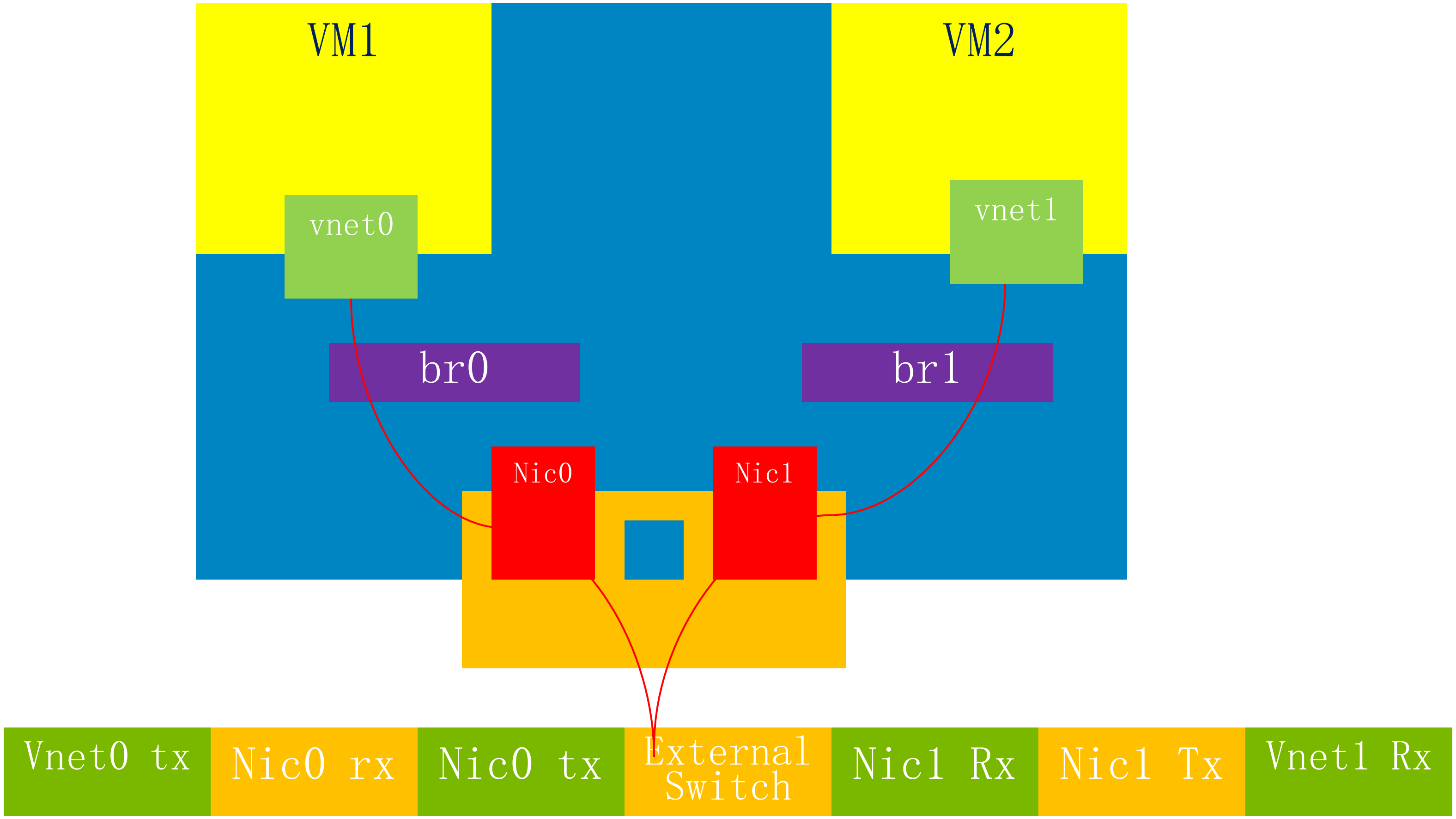


Linux bridge

Traffic going in/going out of the VM



VM to VM



VM1

VM2

vnet0

vnet1

br0

br1

Nic0

Nic1

Vnet0 tx

Nic0 rx

Nic0 tx

External Switch

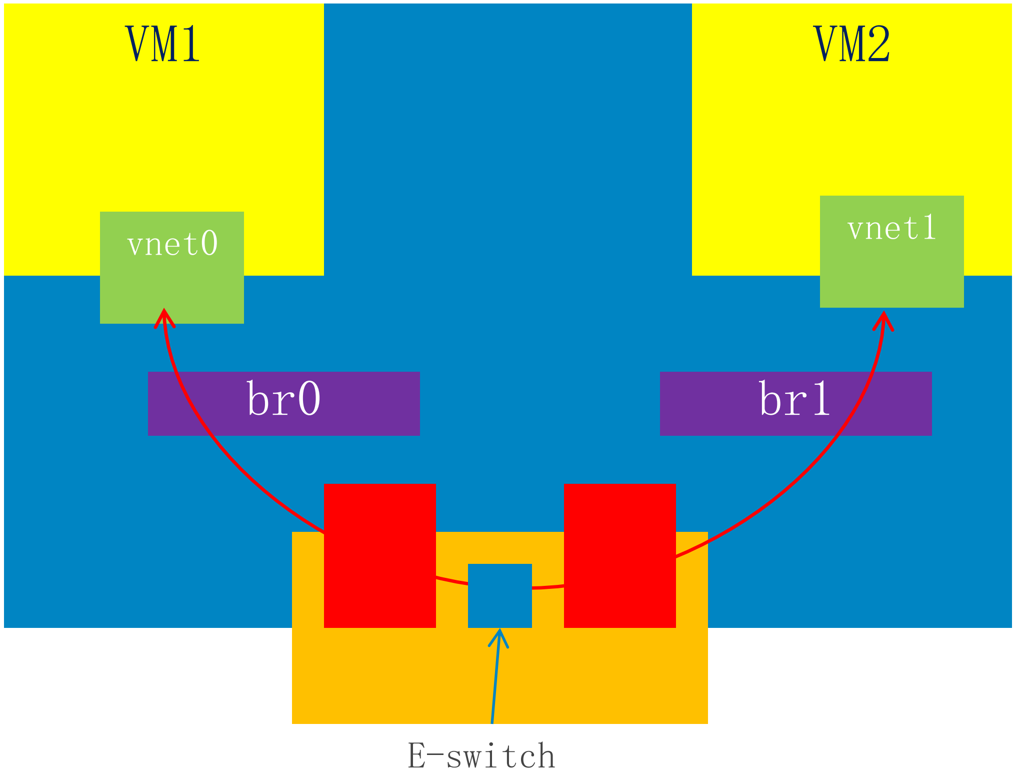
Nic1 Rx

Nic1 Tx

Vnet1 Rx

VM to VM communication

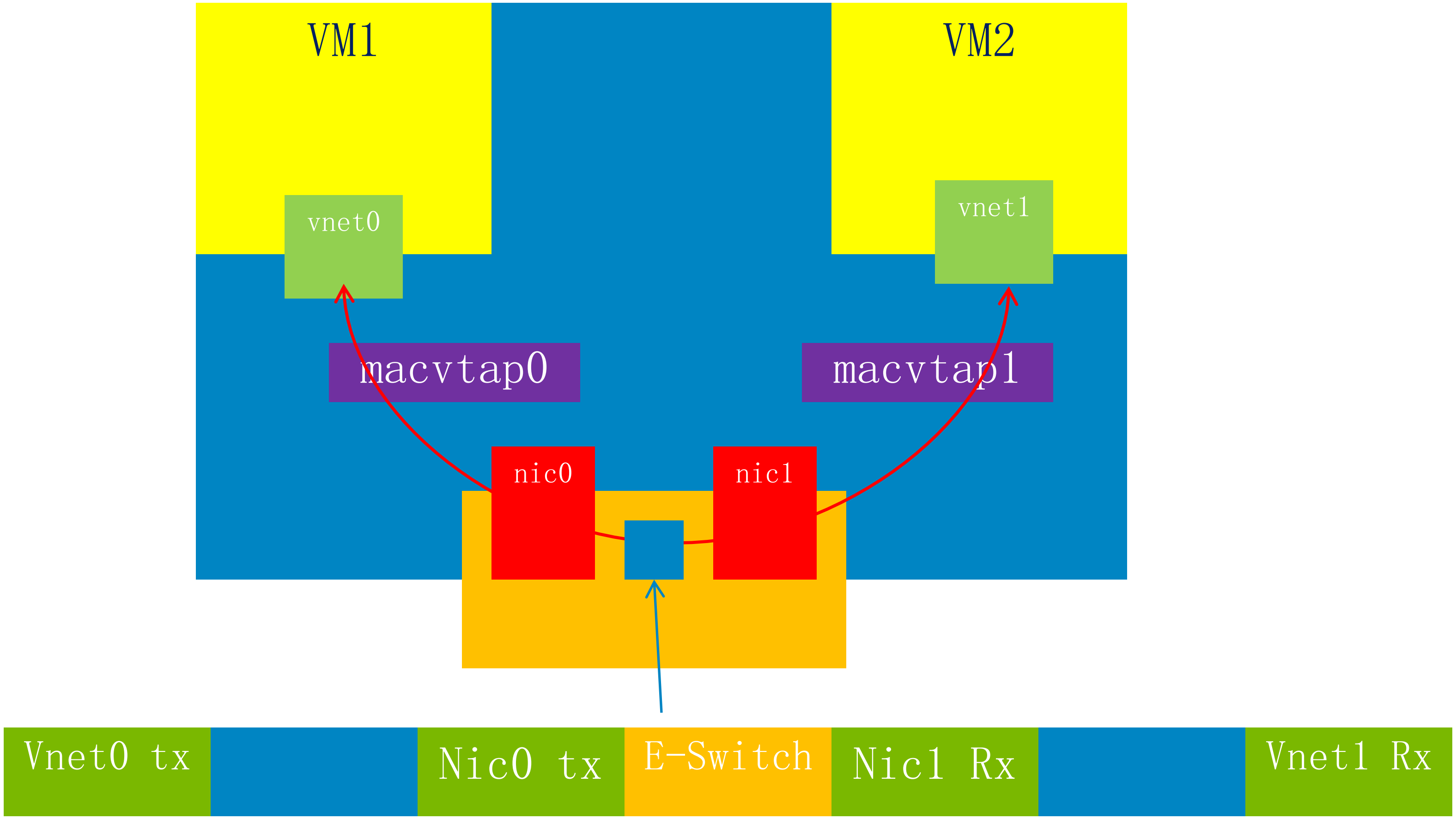
- Optimized for traffic flow via e-switch



Vnet0 tx	Nic0 rx	Nic0 tx	E-Switch	Nic1 Rx	Nic1 Tx	Vnet1 Rx

VM to VM communication

- Macvtap



VM1

VM2

vnet0

vnet1

macvtap0

macvtap1

nic0

nic1

Vnet0 tx

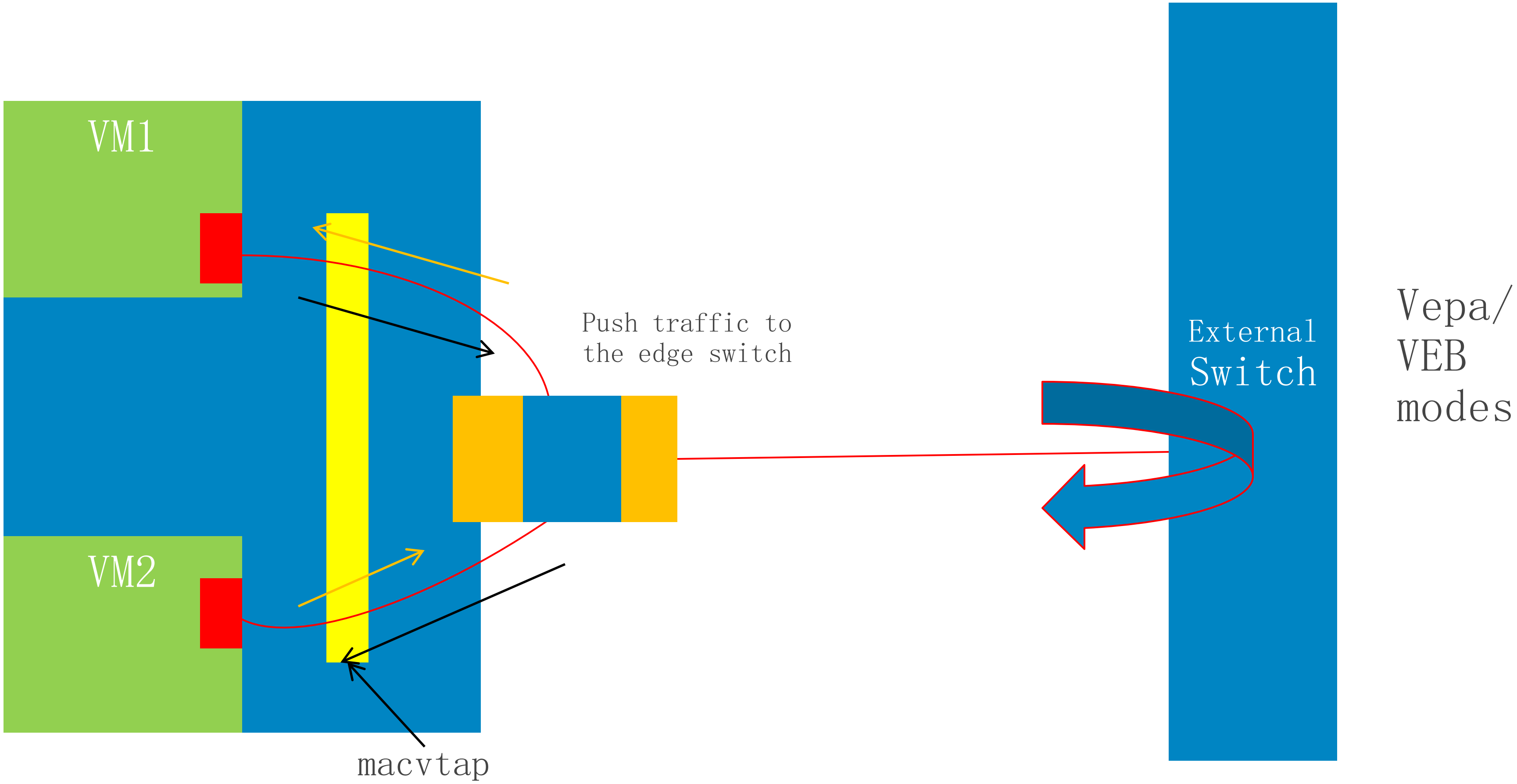
Nic0 tx

E-Switch

Nic1 Rx

Vnet1 Rx

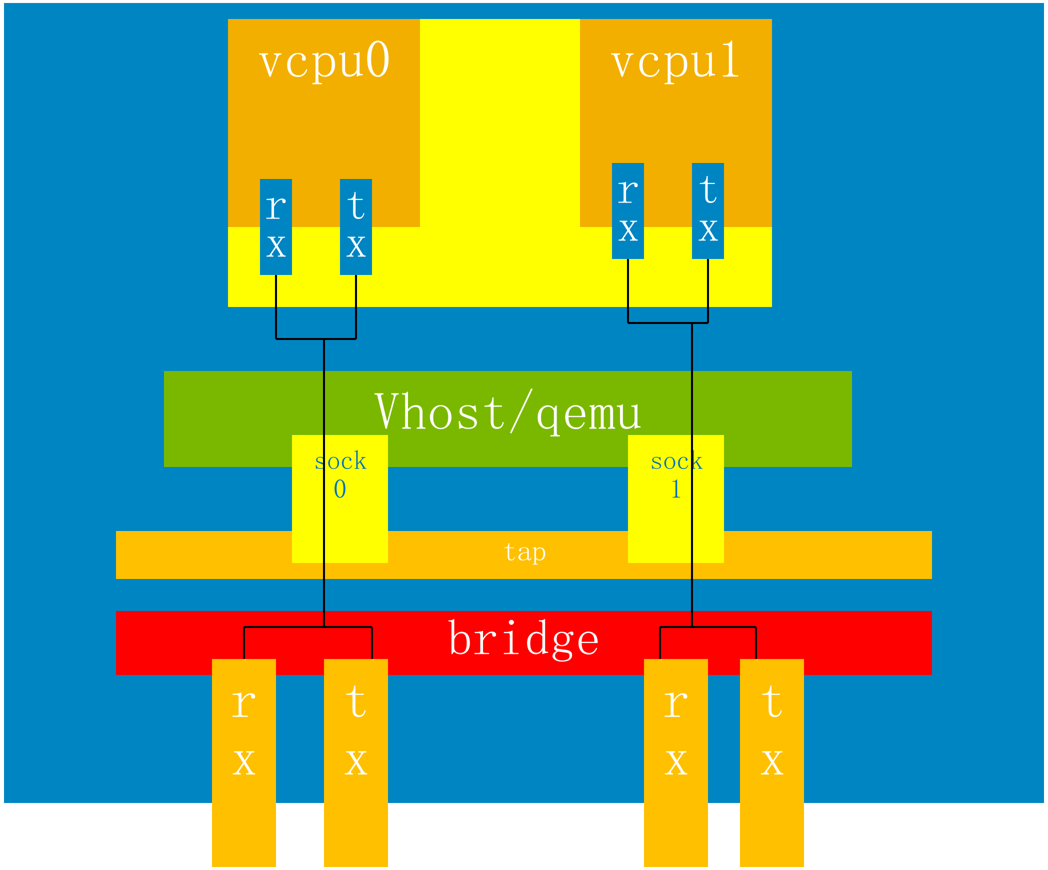
A typical macvtap use-case



QoS Requirements

- Example requirement
- Virtual Machine 1
  - Queue1, priority band 1, 100Mbps
    - Data queue
  - Queue2, priority band 2, 10Mbps
    - Backup queue
  - Some motivations
    - It is the same IP address
    - Possibly in the same vlan
    - Queue based prioritization and classification of workload

Virtio-net Multiqueuing

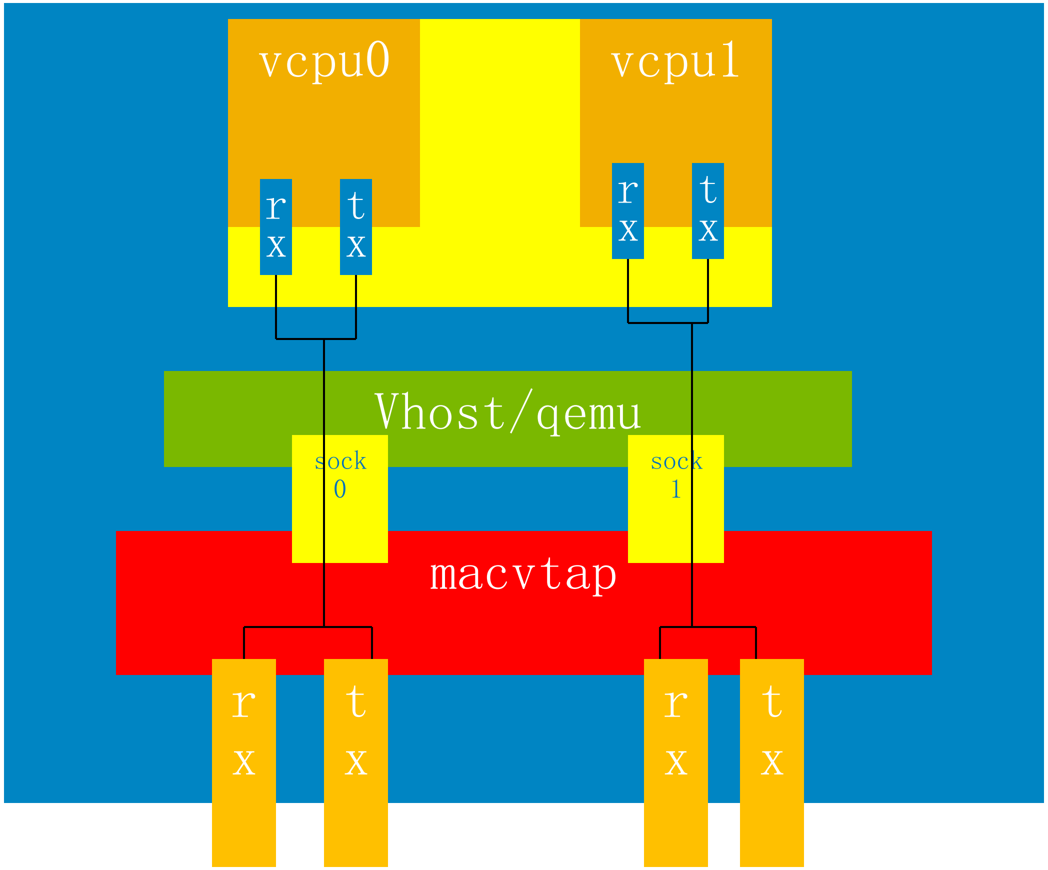


Linux bridge

Original architecture design from

http://www.linux-kvm.org/page/Multiqueue

Virtio-net with macvtap

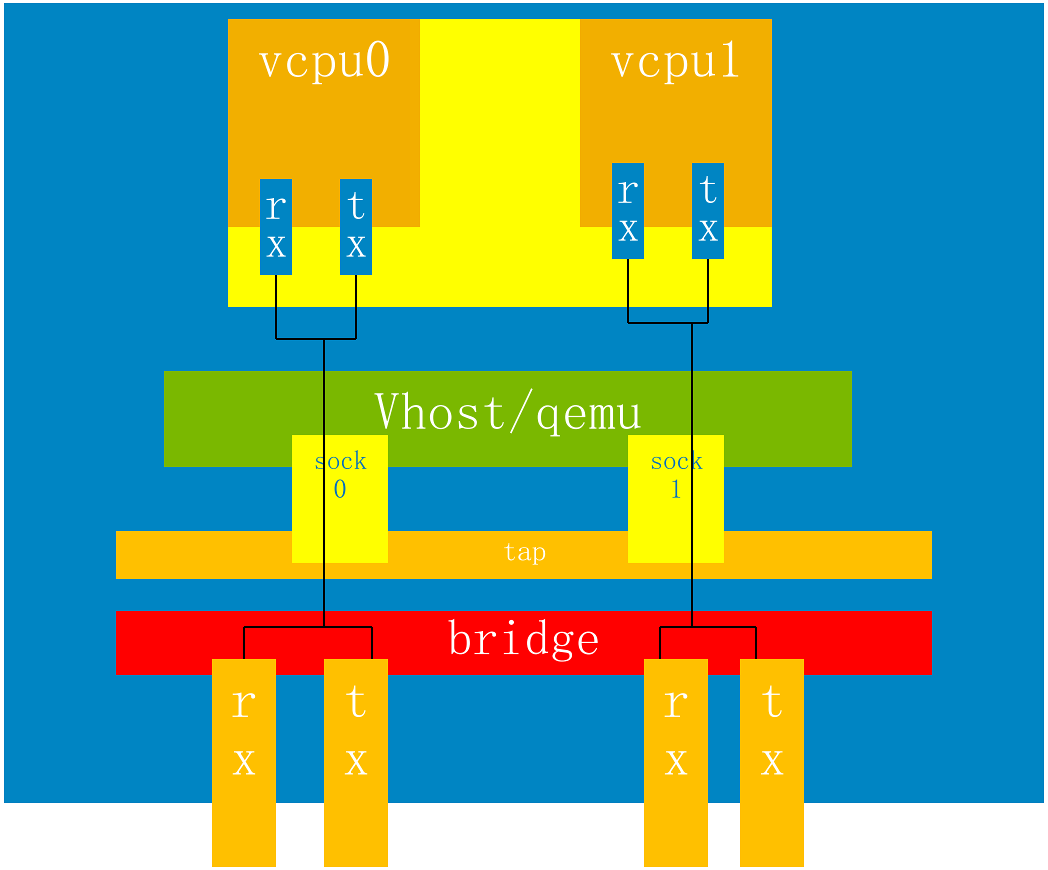


macvtap

Original architecture design from

http://www.linux-kvm.org/page/Multiqueue

With Openvswitch



Openvswitch ??

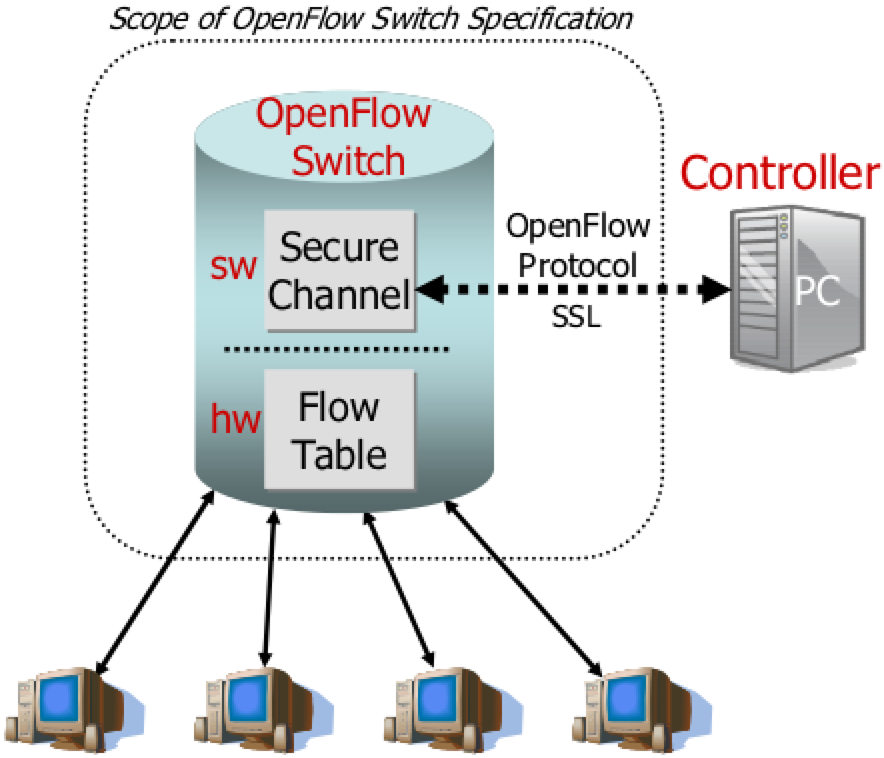
Original architecture design from

http://www.linux-kvm.org/page/Multiqueue

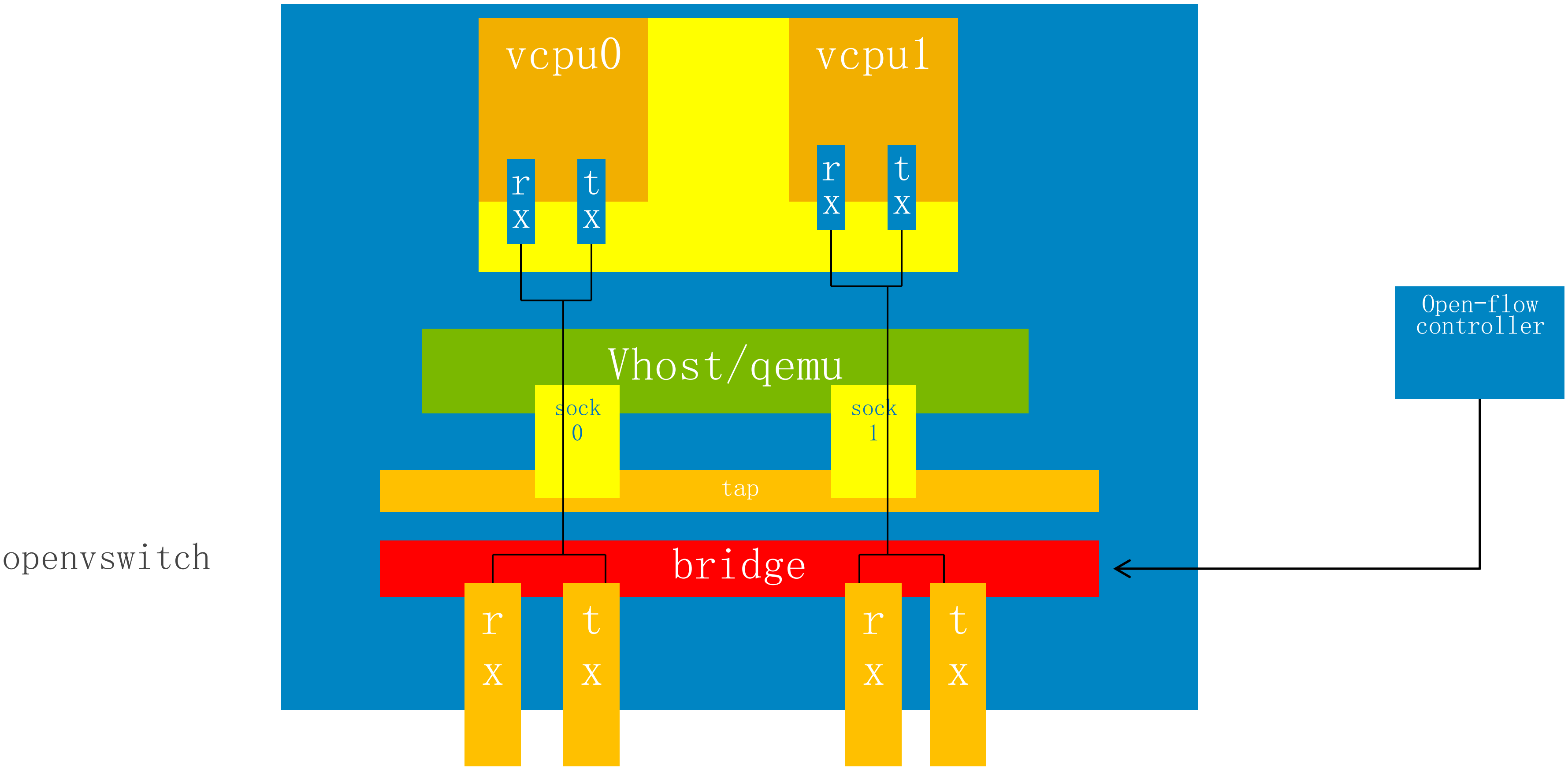
Software Defined networking

- Switching
  - Control Plane
  - Data Plane
- Leading protocols
  - Open Flow
    - Maintained by ONF(open networking foundation)
    - Implementations
      - Open vswitch
- Open Source Network: Software-Defined Network (SDN) and OpenFlow – Insop Song, Ericsson
  - http://lcna2012.sched.org/event/68f58321a544a862253caa8503c8a831?iframe=no&w=900&sidebar=yes&bg=no#.UD6soLnNV38
- www.openflow.org
- www.openvswitch.org

Open flow architecture



Source: www.openflow.org



openvswitch

vcpu0

vcpu1

rx

tx

rx

tx

Vhost/qemu

sock 0

sock 1

tap

bridge

rx

tx

rx

tx

Open-flow controller

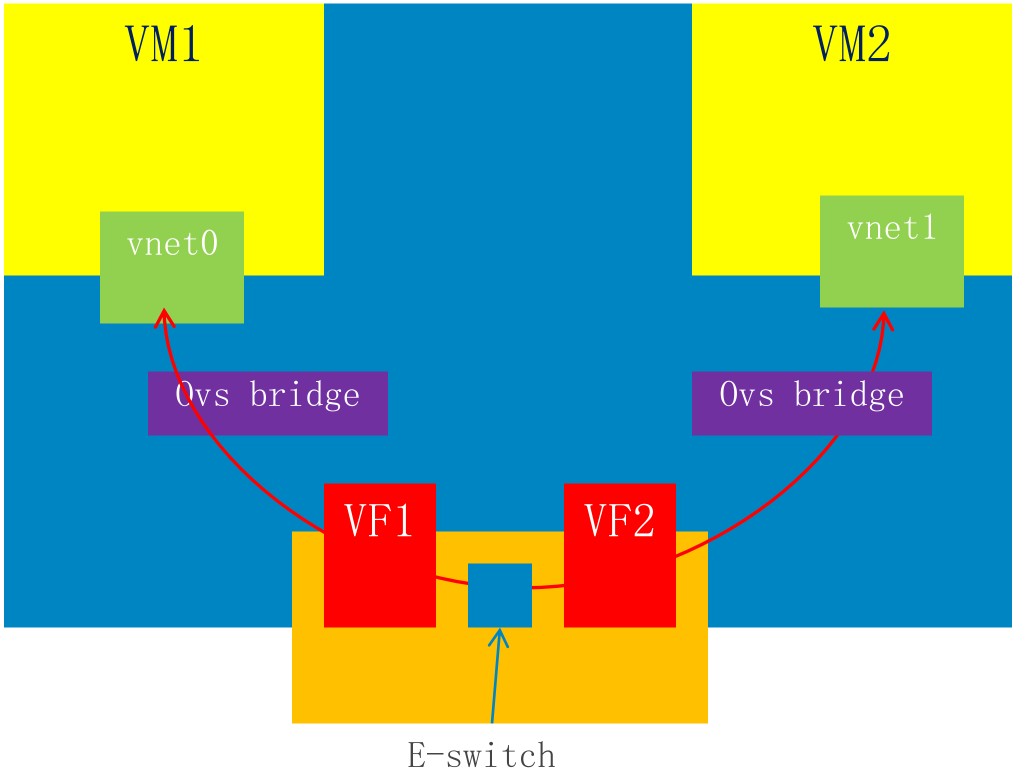
Original architecture design from

http://www.linux-kvm.org/page/Multiqueue

What about SR-IOV ?

- Doesn’t SR-IOV bypass all types of host-bridging when assigned to the VM ?
- Bridging with SR-IOV
  - Don’t Assign VFs to the virtual machines
  - Create vnic ports for the virtual functions/physical functions
  - Bridge the vnic ports with the virtual function/physical function
- Pros
  - Manageability
  - Even better manageability with open flow
- Cons
  - Some possible performance drop.
    - Guest vf driver only vs guest virtio-net + host vf driver
    - Room for improvement

VM to VM communication



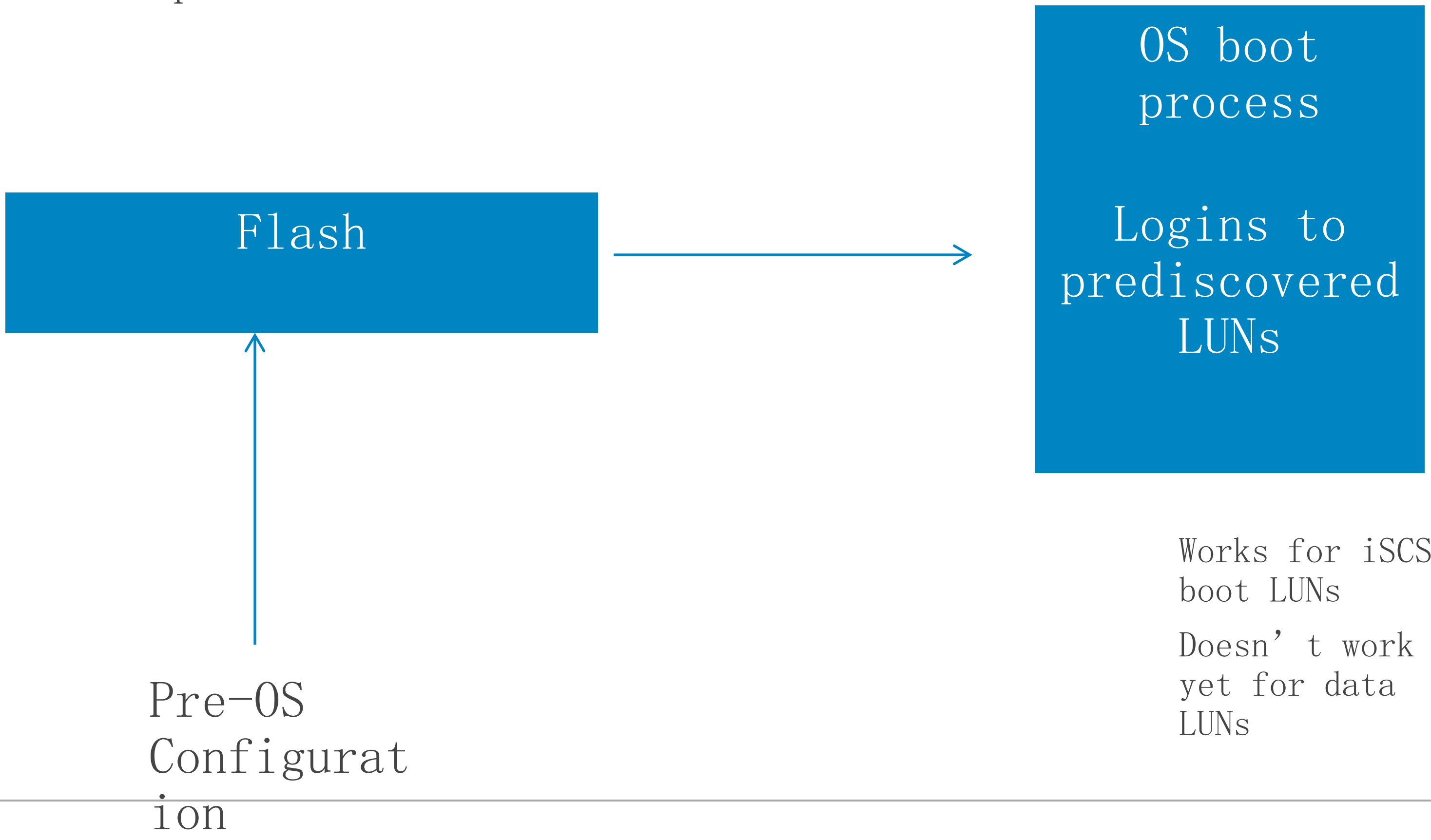
Vnet0 tx	Nic0 rx	Nic0 tx	E-Switch	Nic1 Rx	Nic1 Tx	Vnet1 Rx

Host Physical Network

- Network planning issues with storage in the same converged network on the host server
- How do we separate storage infrastructure as we boot the host server ?
  - VLANs ?
  - But I need that LUN as soon as I boot up

HBA discovery already programmed

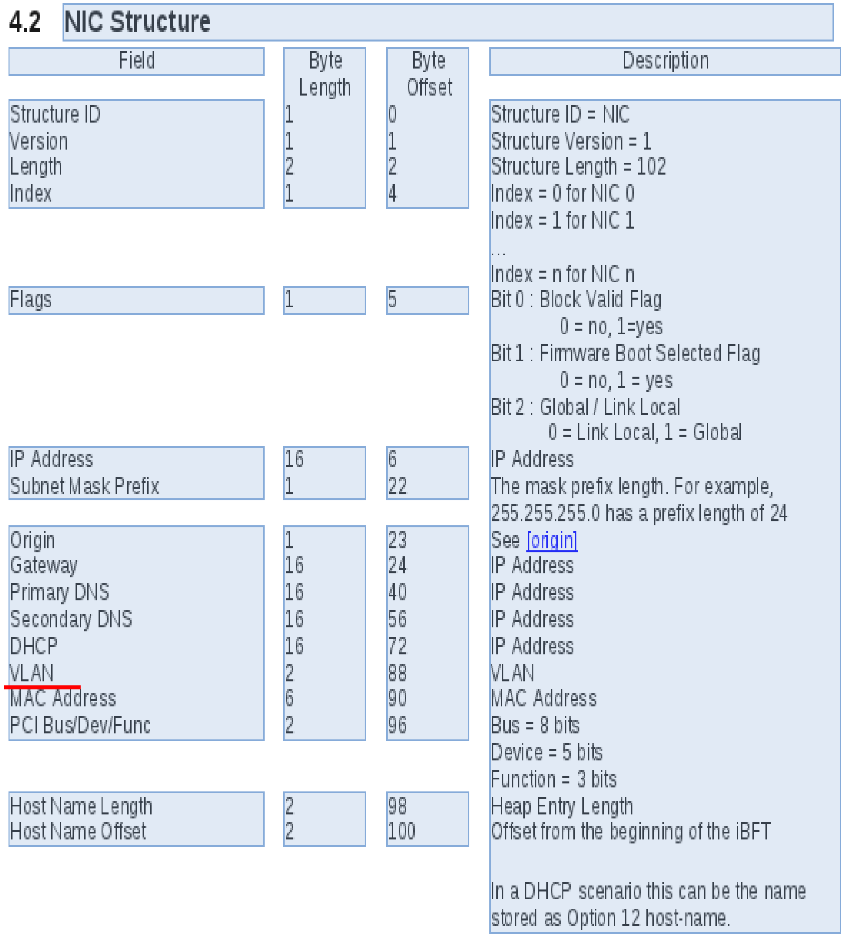
- Flash contains information about discovered HBAs
- Pre configured iSCSI targets
- OS boots up with data LUNs



Works for iSCSI boot LUNs

Doesn’ t work yet for data LUNs

Separate VLANs for iSCSI Network



4.2 NIC Structure

Field	Byte Length	Byte Offset	Description
Structure ID	1	0	Structure ID = NIC
Version	1	1	Structure Version = 1
Length	2	2	Structure Length = 102

Index	1	4	Index = 0 for NIC 0
			Index = 1 for NIC 1
			...
			Index = n for NIC n

Flags	1	5	Bit 0 : Block Valid Flag 0 = no, 1=yes Bit 1 : Firmware Boot Selected Flag 0 = no, 1 = yes Bit 2 : Global / Link Local 0 = Link Local, 1 = Global
IP Address	16	6	IP Address
Subnet Mask Prefix	1	22	The mask prefix length. For example, 255.255.255.0 has a prefix length of 24
Origin	1	23	See [origin]
Gateway	16	24	IP Address
Primary DNS	16	40	IP Address
Secondary DNS	16	56	IP Address
DHCP	16	72	IP Address
VLAN	2	88	VLAN
MAC Address	6	90	MAC Address

PCI Bus/Dev/Func	2	96	Bus = 8 bits
			Device = 5 bits
			Function = 3 bits

Host Name Length	2	98	Heap Entry Length
Host Name Offset	2	100	Offset from the beginning of the iBFT In a DHCP scenario this can be the name stored as Option 12 host-name.

- PreAssign VLANs for the storage network
- Find OS assigning the same vlan id for the storage network
  - Eg: IBFT table(See picture) contains vlan id field that OS can use to recreate the vlan.

Summarizing

- Key Optimization drivers
  - Workload driven vs traditional
  - Orchestration
  - Automation
  - Compute
  - Storage
  - Network

Thanks..

- Feedback
  - shyam_iyer<at>dell<dot>com
- Virtio-net multiqueue work
  - http://www.linux-kvm.org/page/Multiqueue
- Forwarding FDB table to E-switch work
  - http://lwn.net/Articles/491521/
- Virtio-net integration with openvswitch
  - https://blueprints.launchpad.net/lpc/+spec/lpc2012-net-openswitch-harmonizing
- Discovering iSCSI HBA information from storage adapter’s flash
  - https://groups.google.com/forum/?fromgroups=#!topic/open-iscsi/5sbB76c0BZg
- http://linux.dell.com/files/presentations/LinuxCon2012