LinuxCon Japan 2015

FUJITSU

shaping tomorrow with you

Live Migrate Guests w/PCI Pass-Through Devices

Jun 4th, 2015

Taku Izumi
FUJITSU LIMITED

Table of Contents

FUJITSU

FUJITSU

What is “Live migrate guests with PCI pass-through devices?

Background

FUJITSU

- Live Migration of KVM guests
  - Relocating running VMs from one physical host to another
  - Achieve high availability of guests
- PCI pass-through
  - Assign host physical PCIe devices to a guest, giving it a direct access to PCIe devices
  - Achieve high (almost native) I/O performance and low-latency of guests
- Live Migration and PCI pass-through don’t go together
  - Guests w/pass-through devices cannot be live-migrated
    - Internal information of PCIe devices (ex. registers) cannot be migrated
    - Target host may not have the same hardware

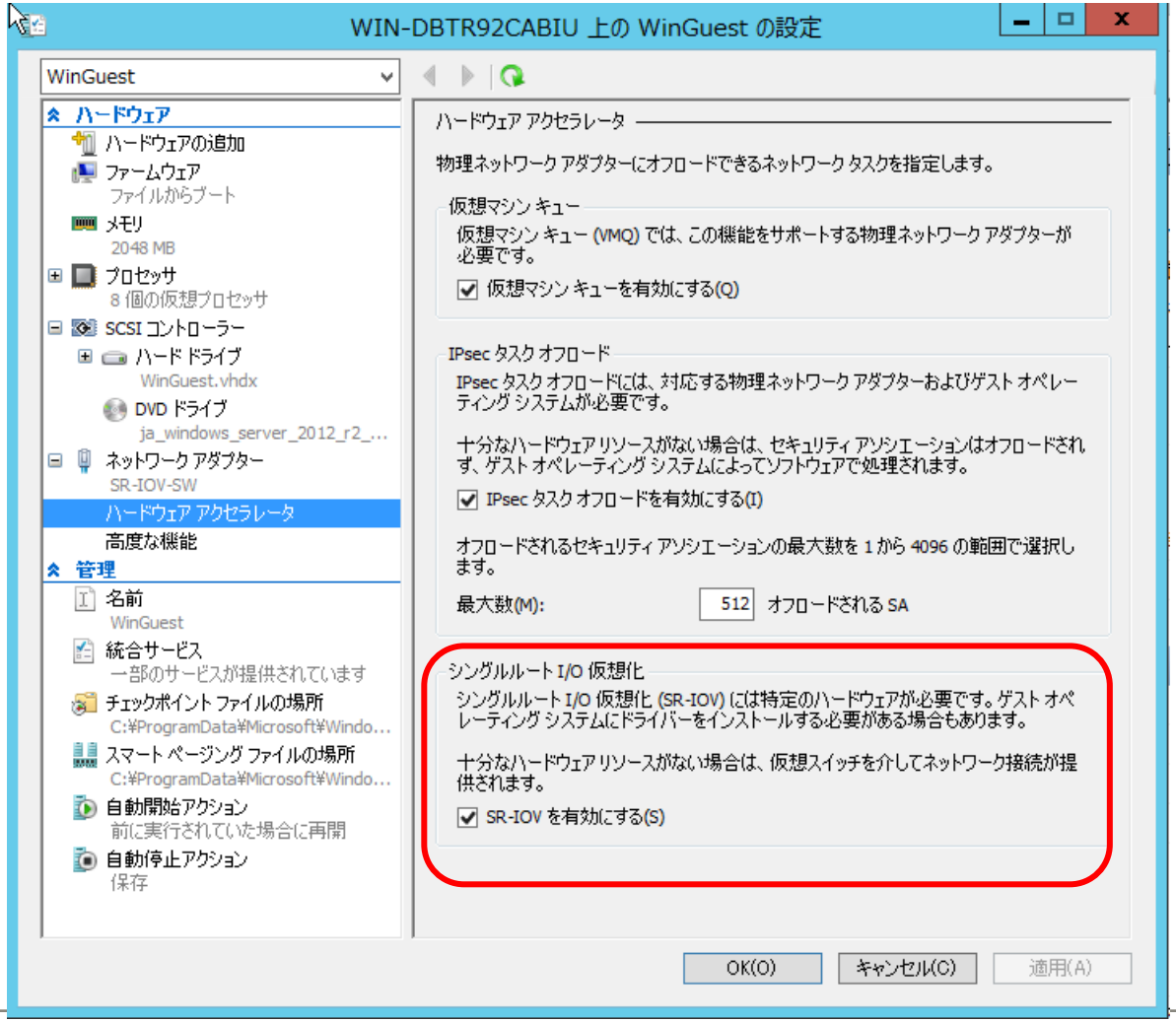
Trends of other hypervisors

FUJITSU

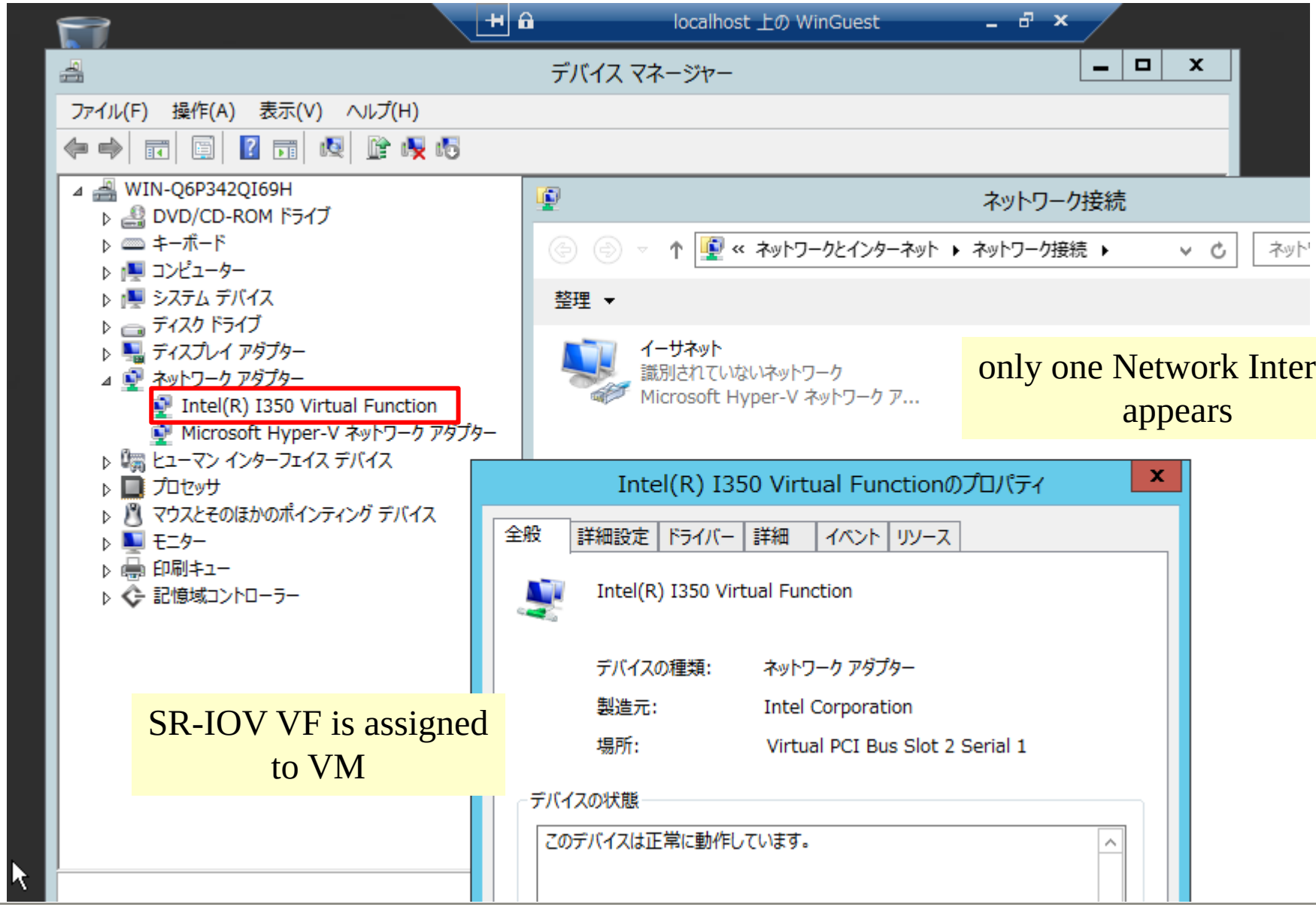
- Hyper-V
  - Hyper-V 1.0 doesn’t support PCI pass-through feature
  - Hyper-V 2.0 supports PCI pass-through of SR-IOV VF of NIC
    - Only Windows VMs are supported
    - Live migration is fully supported
- VMWare vSphere
  - VMDirectPath and vMotion don’t go together in most cases
  - vMotion with VMDirectPath I/O Gen2 is supported
    - Necessitate Cisco Virtual Interface Card
- Xen
  - Live Migration and PCI pass-through don’t go together

Approach of Hyper-V

- Unlike KVM, pass-throughed SR-IOV VF is used to accelerate para-virtualized NIC

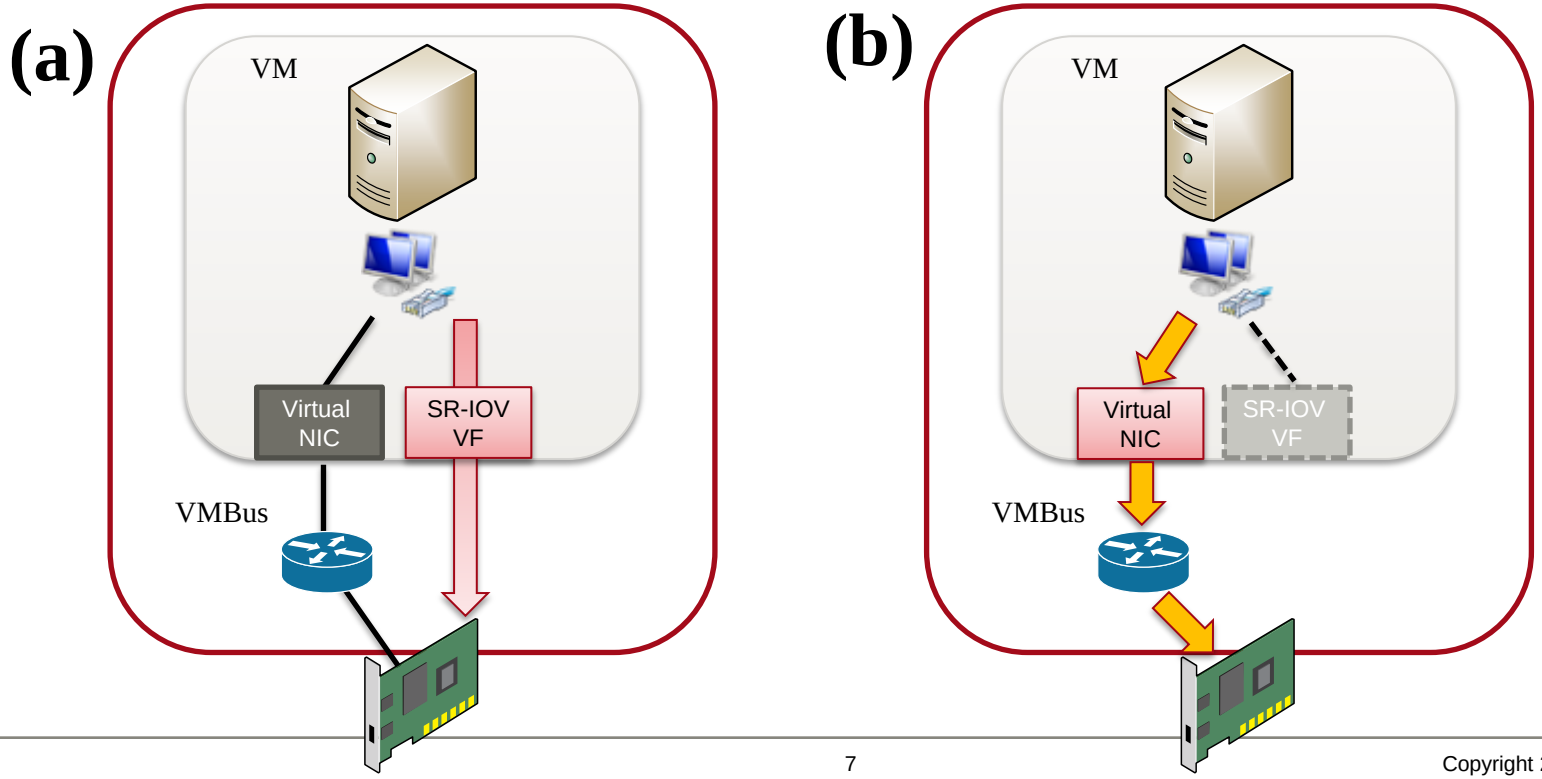


Approach of Hyper-V cont.



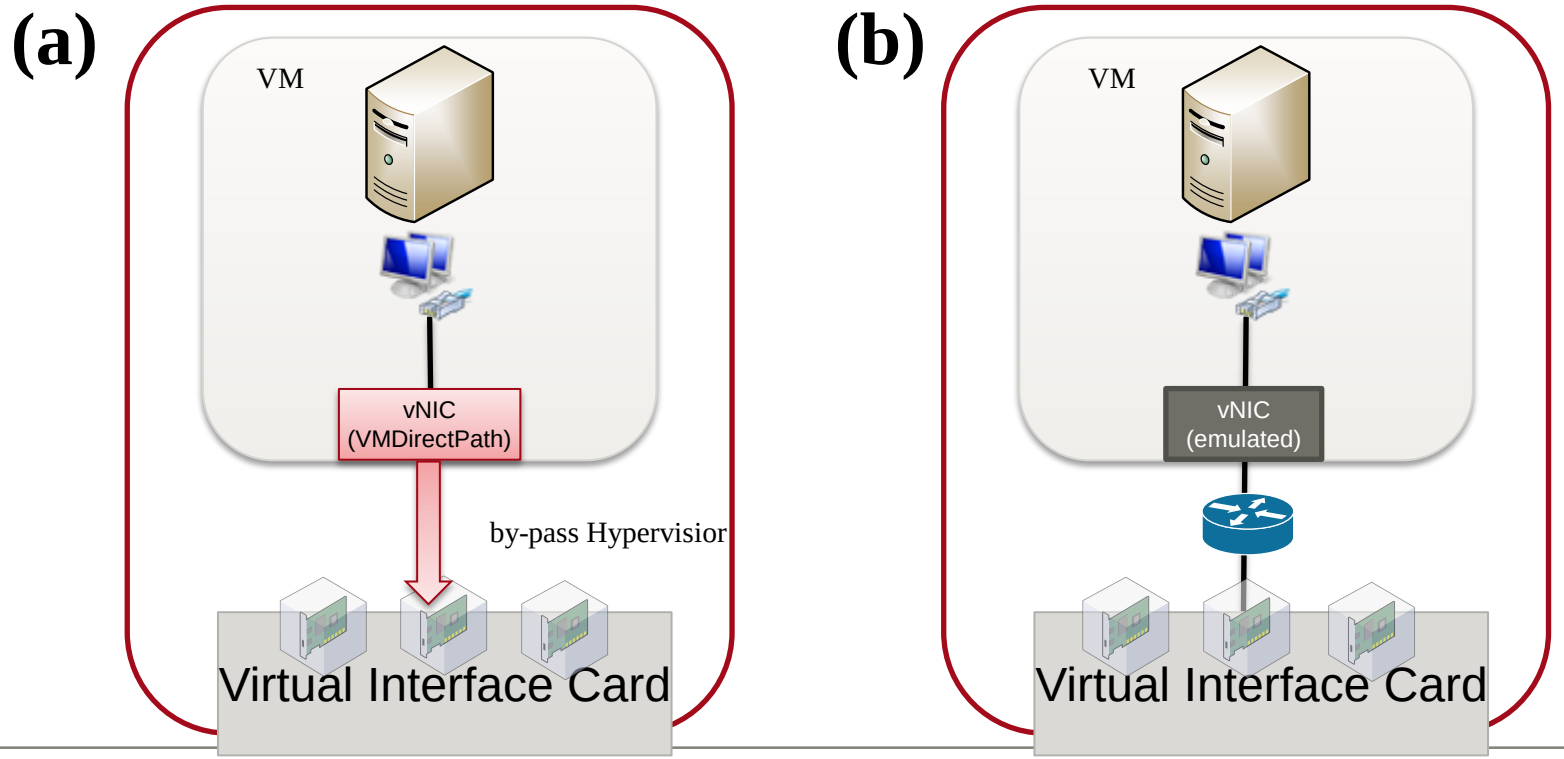
Approach of Hyper-V cont.

- Normally VM issues I/O via SR-IOV VF device (a)
- When live-migration is triggered, SR-IOV VF is removed
- During live-migration, or VM can't find SR-IOV VF, VM issues I/O via traditional virtual NIC (b)



Approach of VMWare vSphere

- vNIC supports 2 modes
  - (a) VMDirectPath mode (High performance mode)
  - (b) emulated mode (Standard mode)
- When vMotion is triggered, vmkernel switch from VMDirectPath mode to emulated mode (vMotion supported)



Clue to the solution

FUJITSU

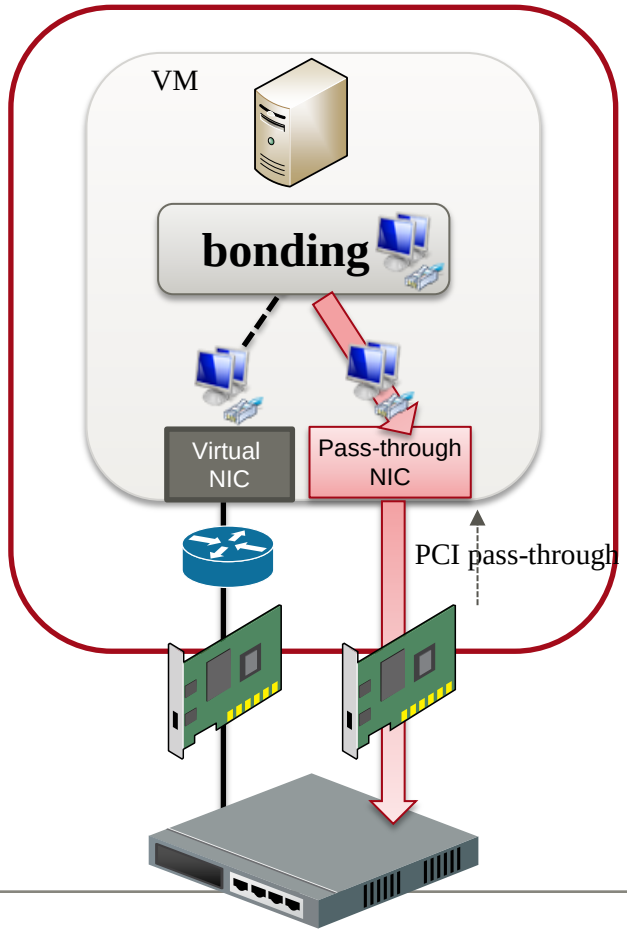
- Making a pair of para-virtualized device and PCI pass-through device
  - Normally use PCI pass-through device
  - When live-migration, switch from PCI-pass through device to para-virtualized device
- Redundant configuration to maintain connectivity when switching devices
- Removing PCI pass-through device before live-migration to avert migration failure and restoring it after live-migration

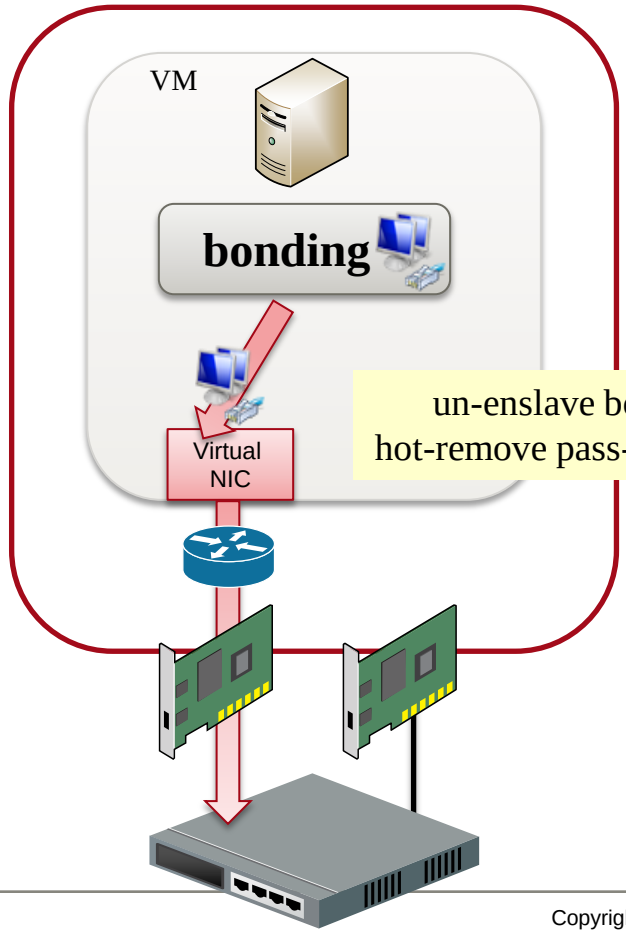
FUJITSU

Proposal Solution

Basic idea

- Configure bonding device (active-backup mode) on VM
  - Enslaving a para-virtualized NIC and pass-through NIC
- Hot-remove pass-through NIC before live-migration
  - Hot-add new one after live-migration





Basic idea cont.

FUJITSU

- Pros. and Cons.
  - Pros.
    - Teaming feature is available on Linux and Windows guest
    - Hot-plug of pass-through device works well on KVM guests
    - The last piece of puzzle is automation method
  - Cons.
    - The kind of pass-through devices is limited to Network Card
    - The configuration of bonding driver may conflict guest user's network configuration
- Design concept
  - Extend domain schema (XML) for bonding configuration
  - Enhance qemu-guest-agent to do guest side operation

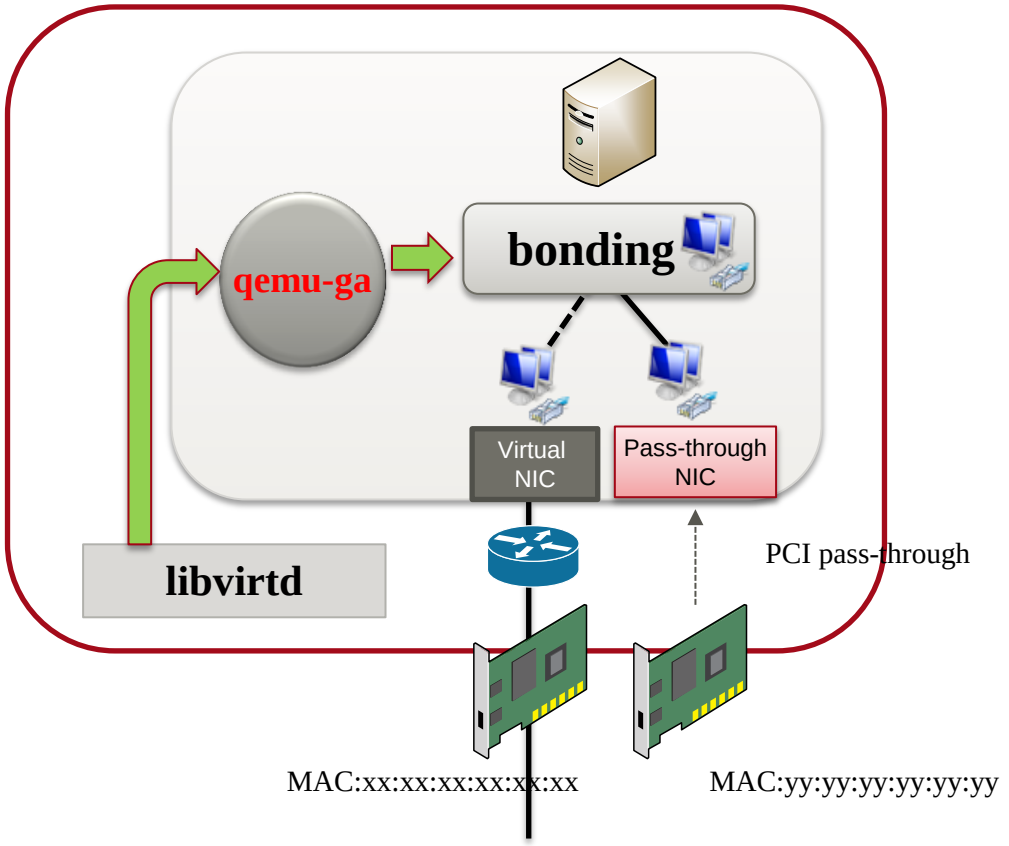
Implementation: On VM startup

FUJITSU

- qemu-guest-agent creates bonding device
  - enslaving a para-virtualized NIC and pass-through NIC

Guest configuration XML

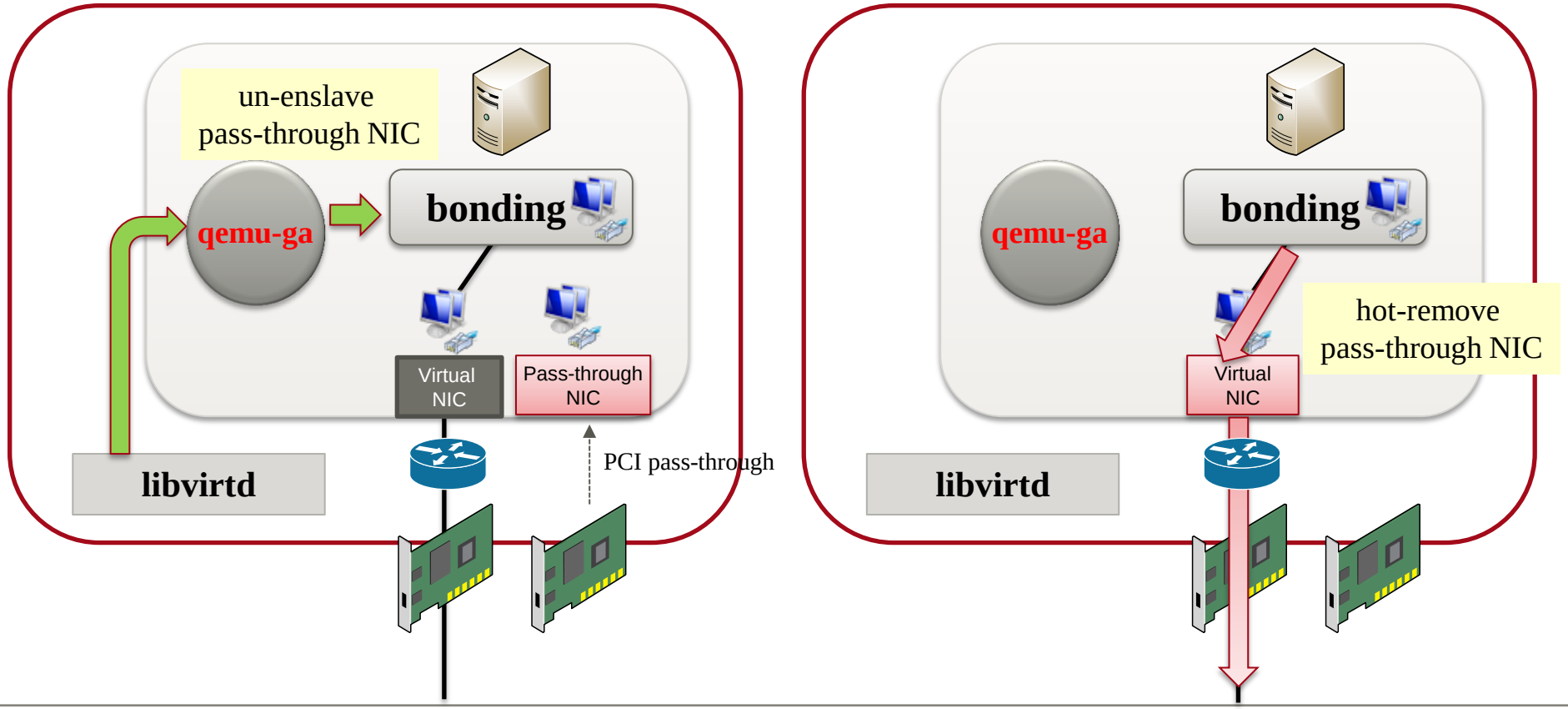
```
<domain type='kvm'>
  ....
  <devices>
    <hostdev mode='subsystem' type='pci'
        managed='yes'>
      <driver name='vfio' type='bond'/>
      <bond>
        <interface address='xx:xx:xx:xx:xx:xx'/>
        <interface address='yy:yy:yy:yy:yy:yy'/>
      </bond>
      <source>
        <address domain='0x0000' bus='0x49'
            slot='0x00' function='0x0'/>
      </source>
    </hostdev>
  ....
  </devices>
....
</domain>
```



Implementation: Just before live-migration

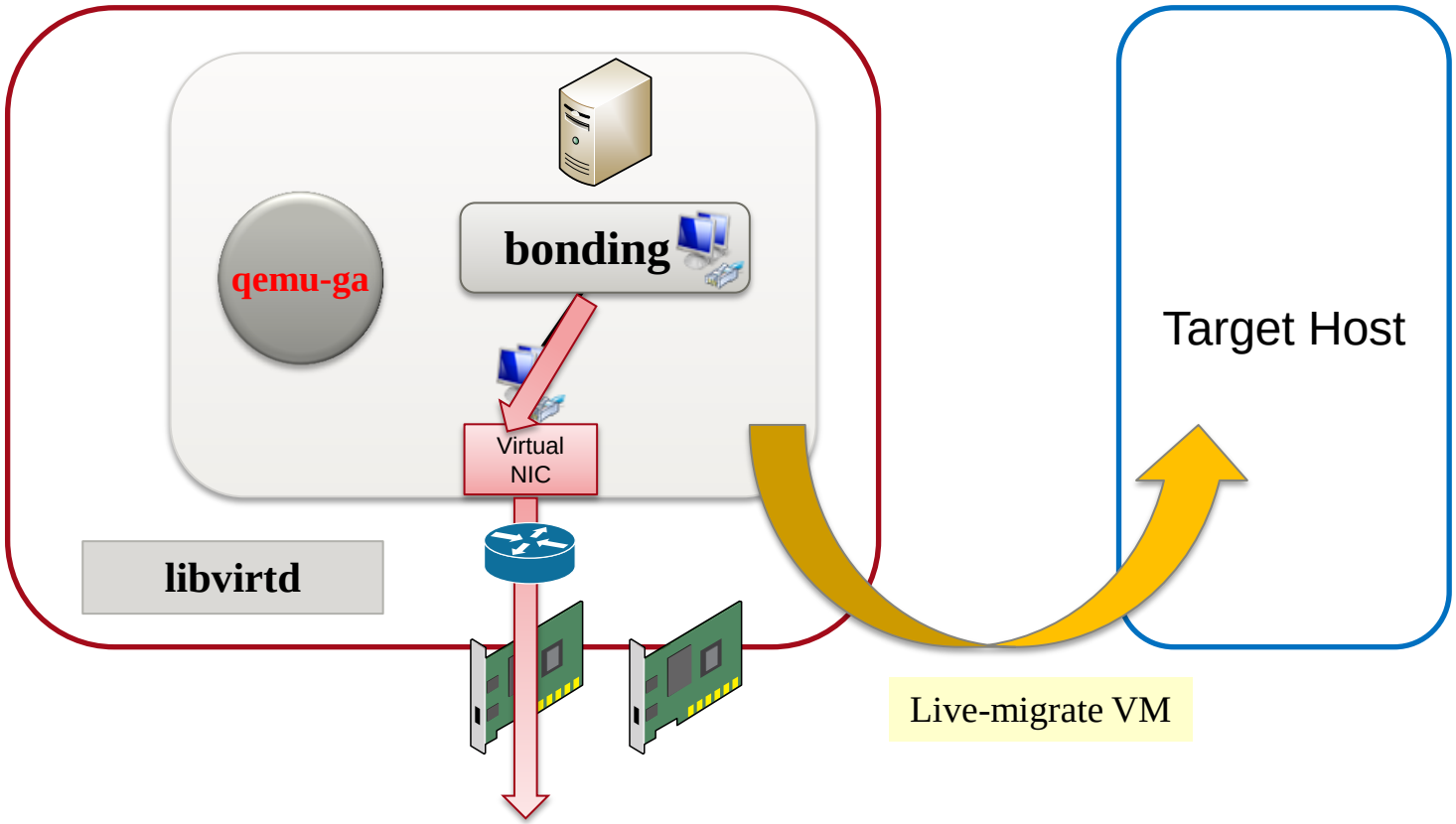
FUJITSU

- qemu-guest-agent un-enslaves pass-through NIC
- libvirtd hot-removes pass-through NIC



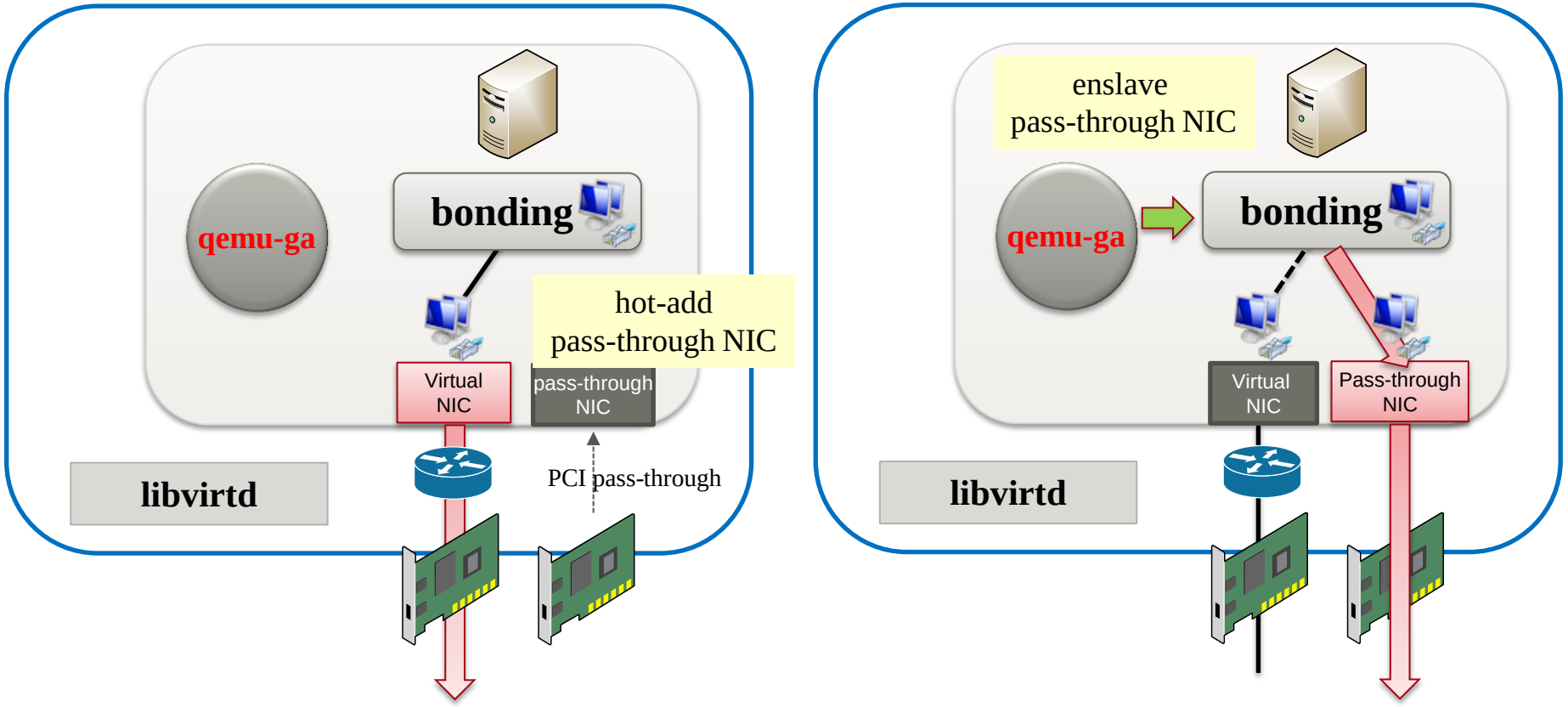
Implementation: live-migrate

- Now that VM has no pass-through device, VM can be live-migrated to other host



Implementation: Just after live-migration

- At target host, libvirtd hot-adds new pass-through NIC
- qemu-guest-agent enslaves it



FUJITSU

Current status

Feedback of community

FUJITSU

- We post first patch-set in April 2015
- No one disapprove of this feature itself, however we got many objection against our proposal solution
  - libvirtd should not be involved in this operation
- The discussion points are
  - Difficult error handling
  - Difficult selection of new NIC on target host
  - Confliction with guest-side network configuration

Difficult error handling

FUJITSU

- Hot-remove may fail on source host
  - Hot-remove operation may hung forever due to guest heavy traffic
  - This cause domain in undefined state during migration
- Hot-add may fail on target host
  - Hot-add failure would cause the migration bisect
  - How to recover the migration need to discuss
    - Migrate back to source ?
    - VM stop ?
- In case multiple NICs need to be migrated, some successful and some failed
  - How to deal with this case ?

Difficult selection of new NIC on target host

FUJITSU

- The target host almost doesn't has the same NIC devices
  - Unlikely has the same PCI address (Bus/Device/Function) that is on the source
- Need to prepare a set of free NICs for migration as candidates
- Need policy to decide which NIC to be hot-attach

Confliction with guest-side network conf.

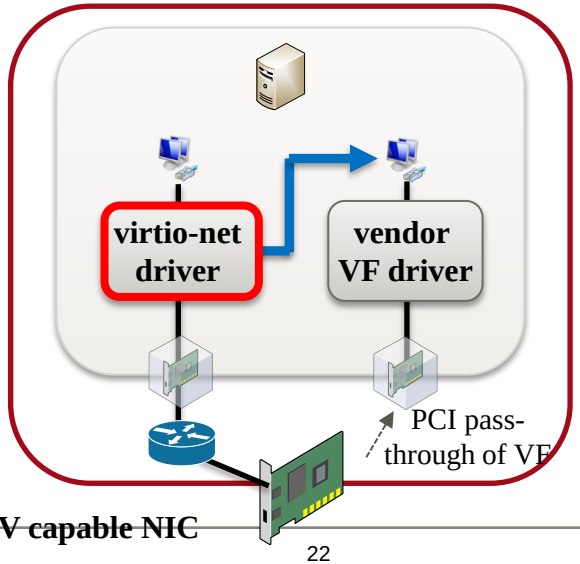
FUJITSU

- Creating bonding device is too intrusive into the guest
- Likely conflict with the existing configuration by NetworkManager or systemd-network in guest
- Libvirt dosn’t know configuration manner decided by guest admin, more likely create bonding device fail

Future Plan

FUJITSU

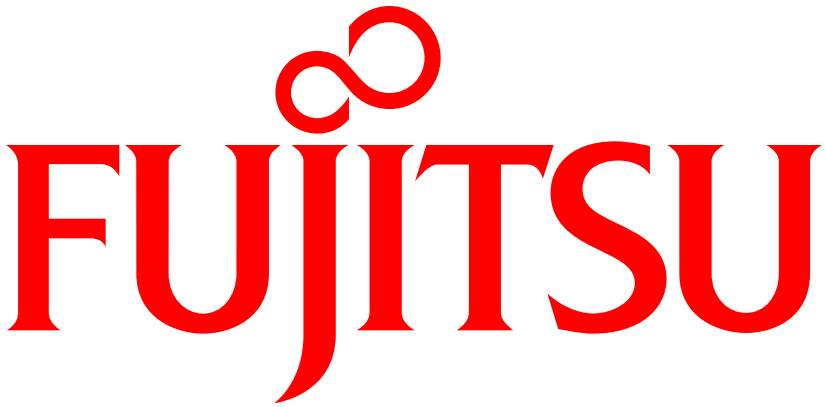
- We are now considering better approach
  1. Enhance management application (oVirt or OpenStack)
     - App have the policy to know the hot plug/unplug is complete. And then decide if to do migration
     - App can decide to migrate the appropriate hosts from the cluster
     - if due to some odd reasons migration failed, App can migrate to another appropriate host
  2. Add new type virtual NIC device and enhance para-virtualized driver to have bonding-like functionality



Summary

FUJITSU

- What is “Live migrate guests with PCI pass-through devices?
- Proposal Solution
- Current status



shaping tomorrow with you