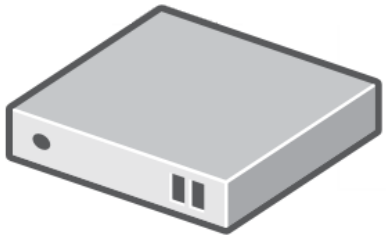
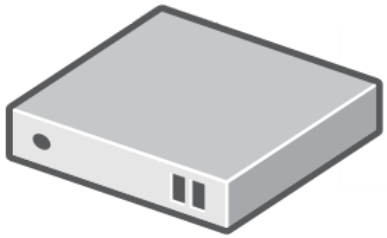
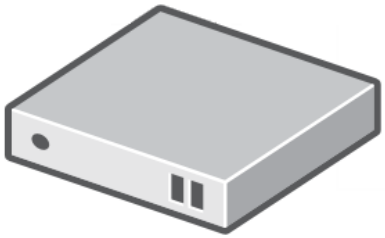
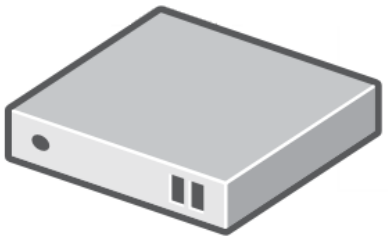
oVirt

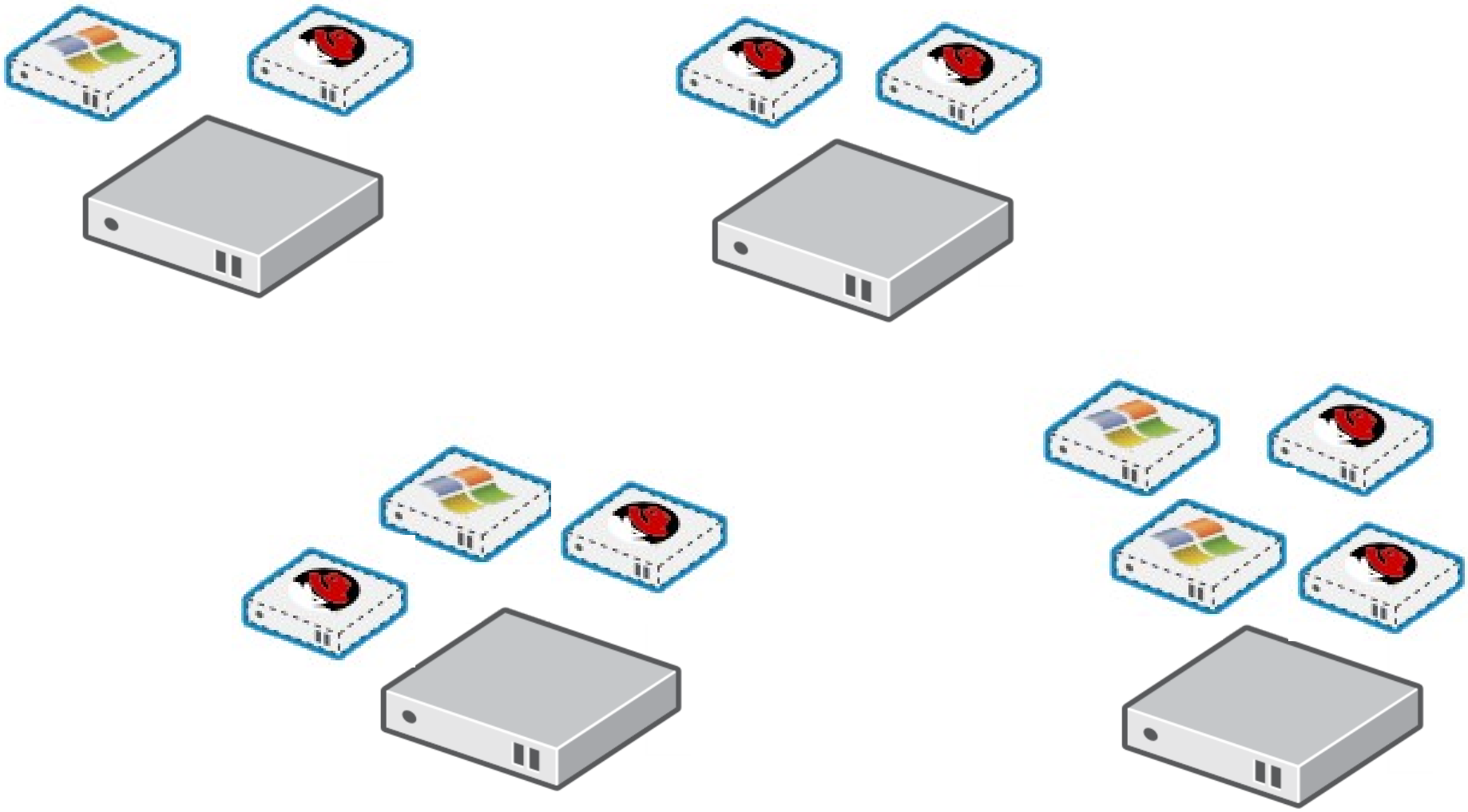
High Availability with No Split Brains!

Arik Hadas
Principal Software Engineer
Red Hat
27/01/2018

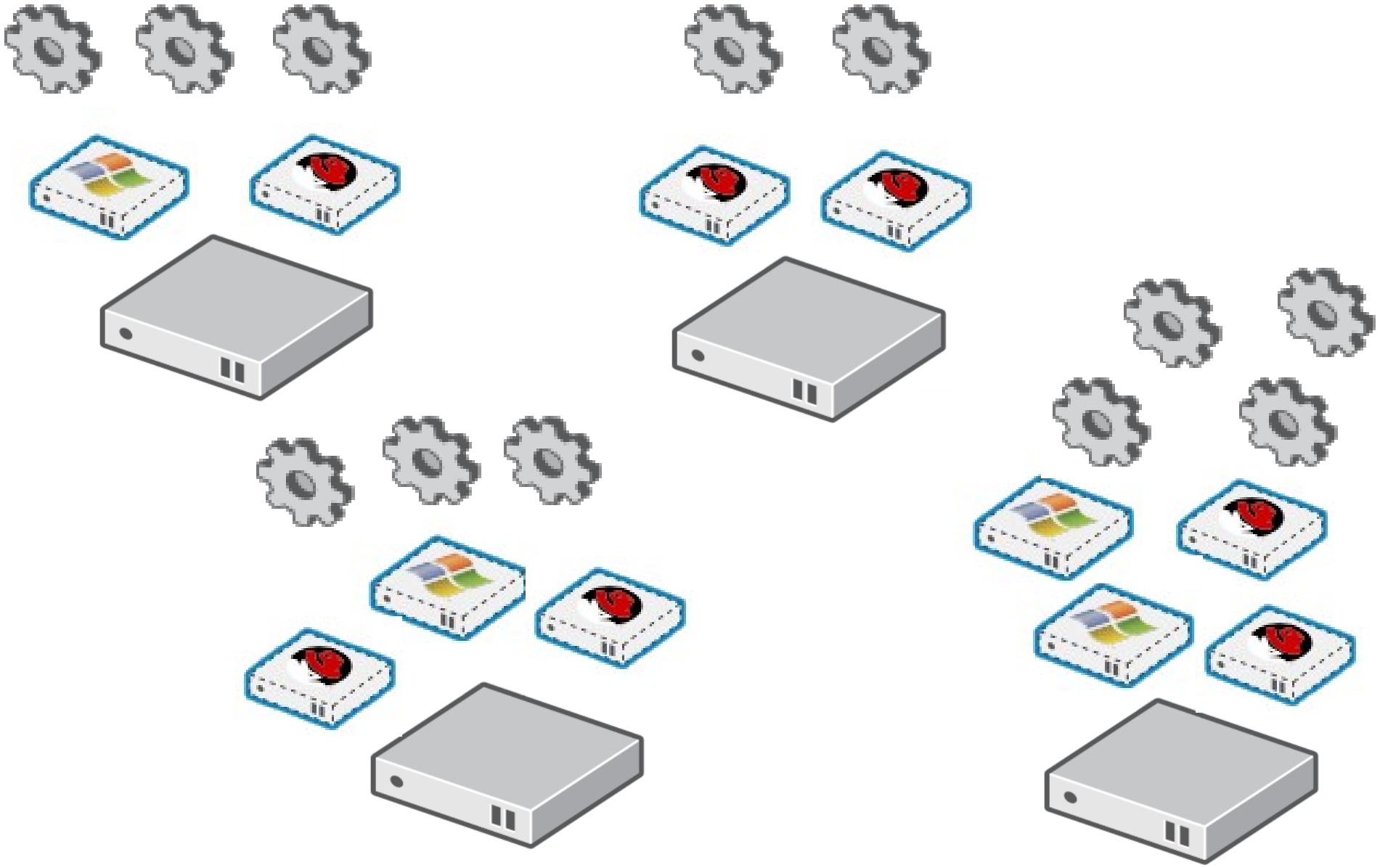
oVirt Virtual Data Center – Physical Servers

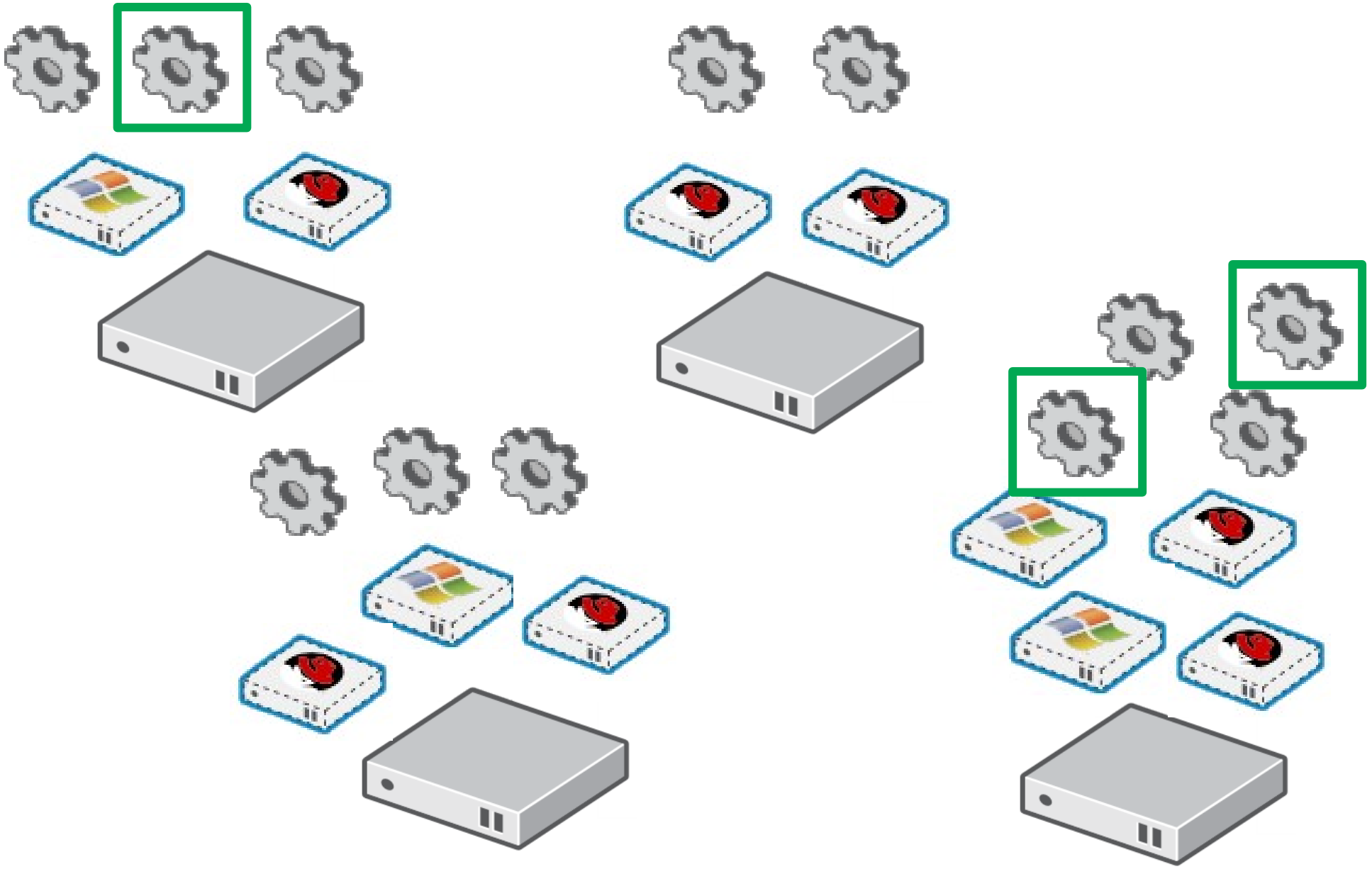
oVirt Virtual Data Center – Virtual Machines



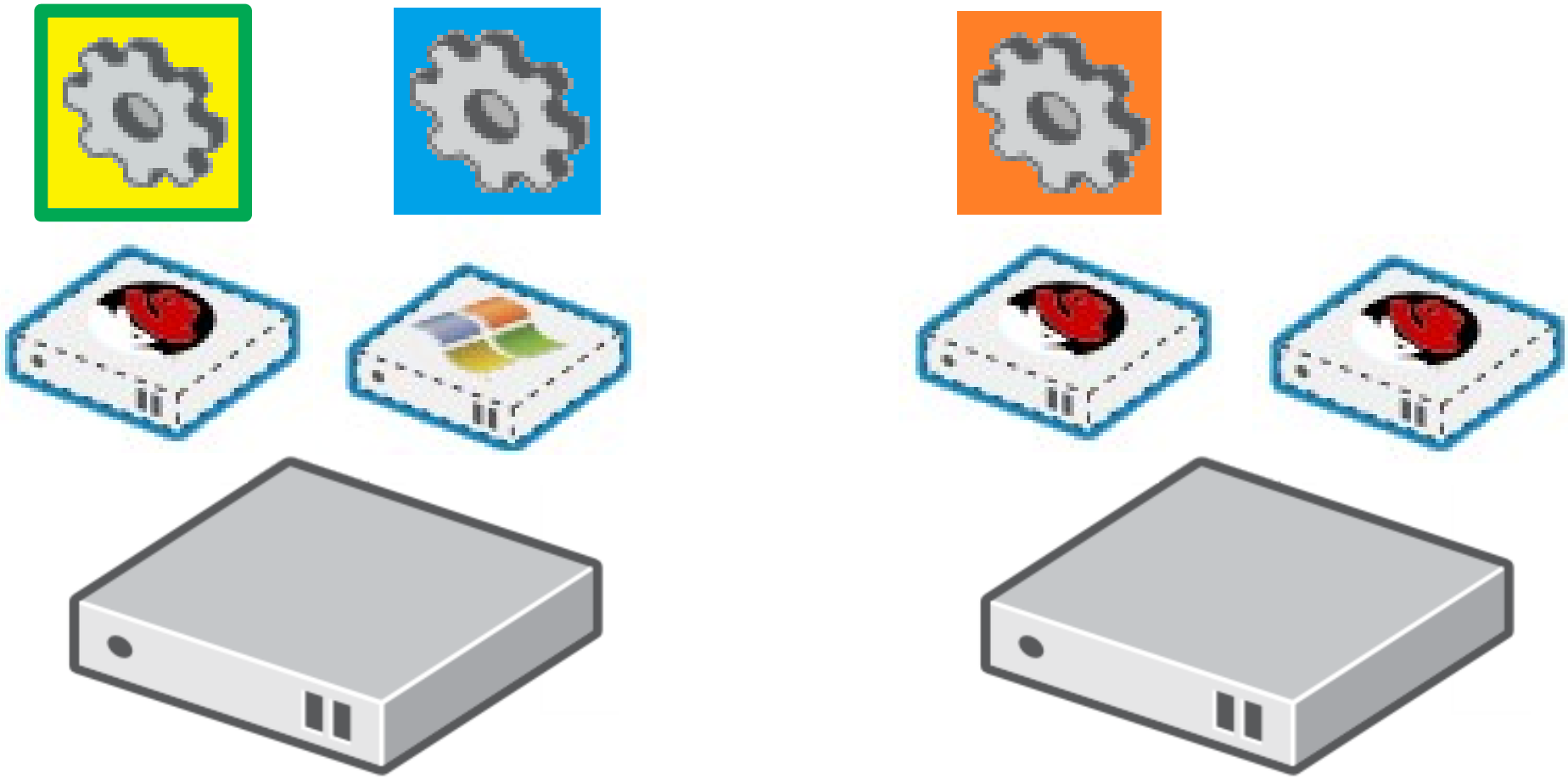
Virtual Data Center - Applications



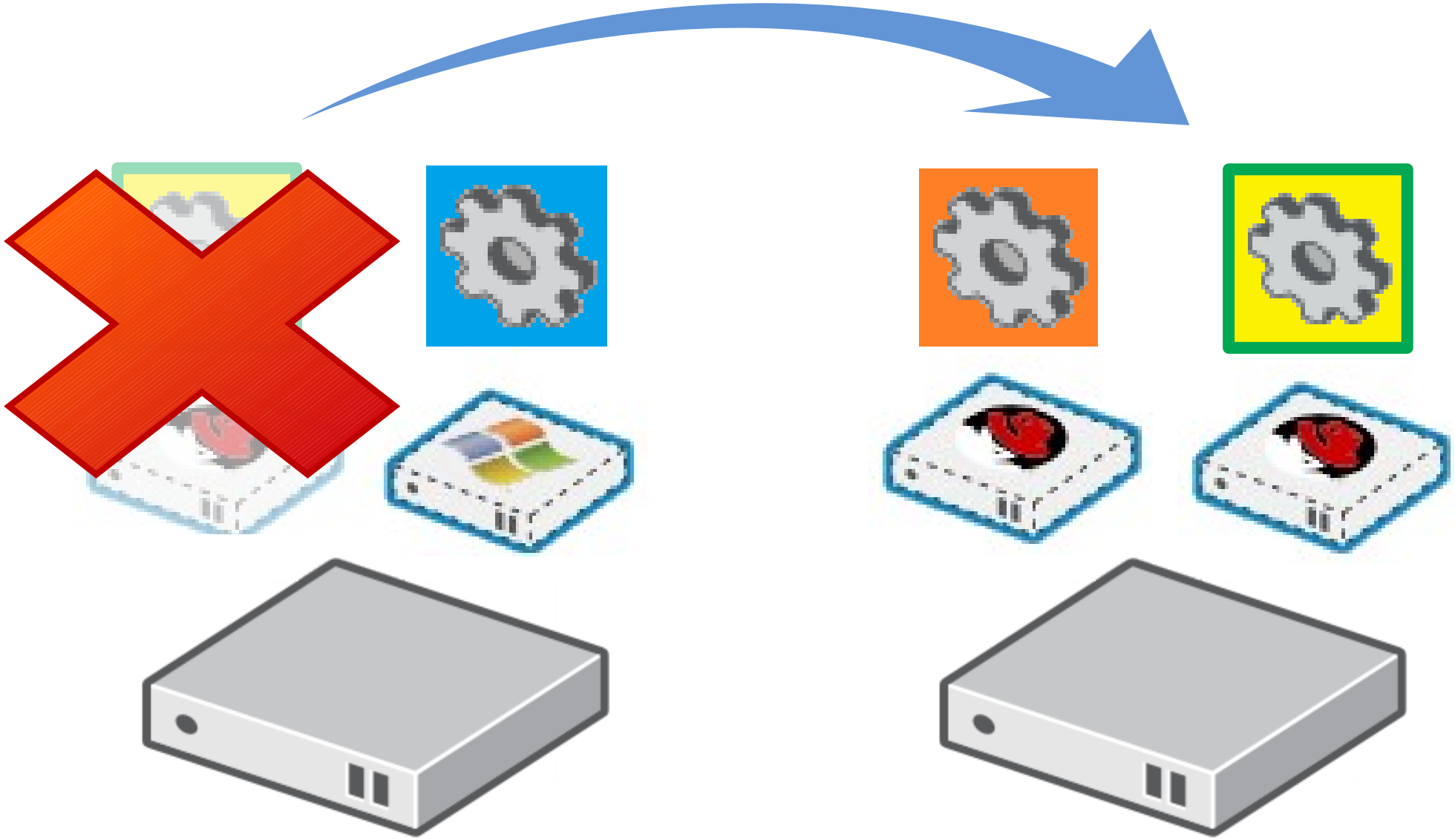
Some Applications are More Critical



High Availability - Application-Level



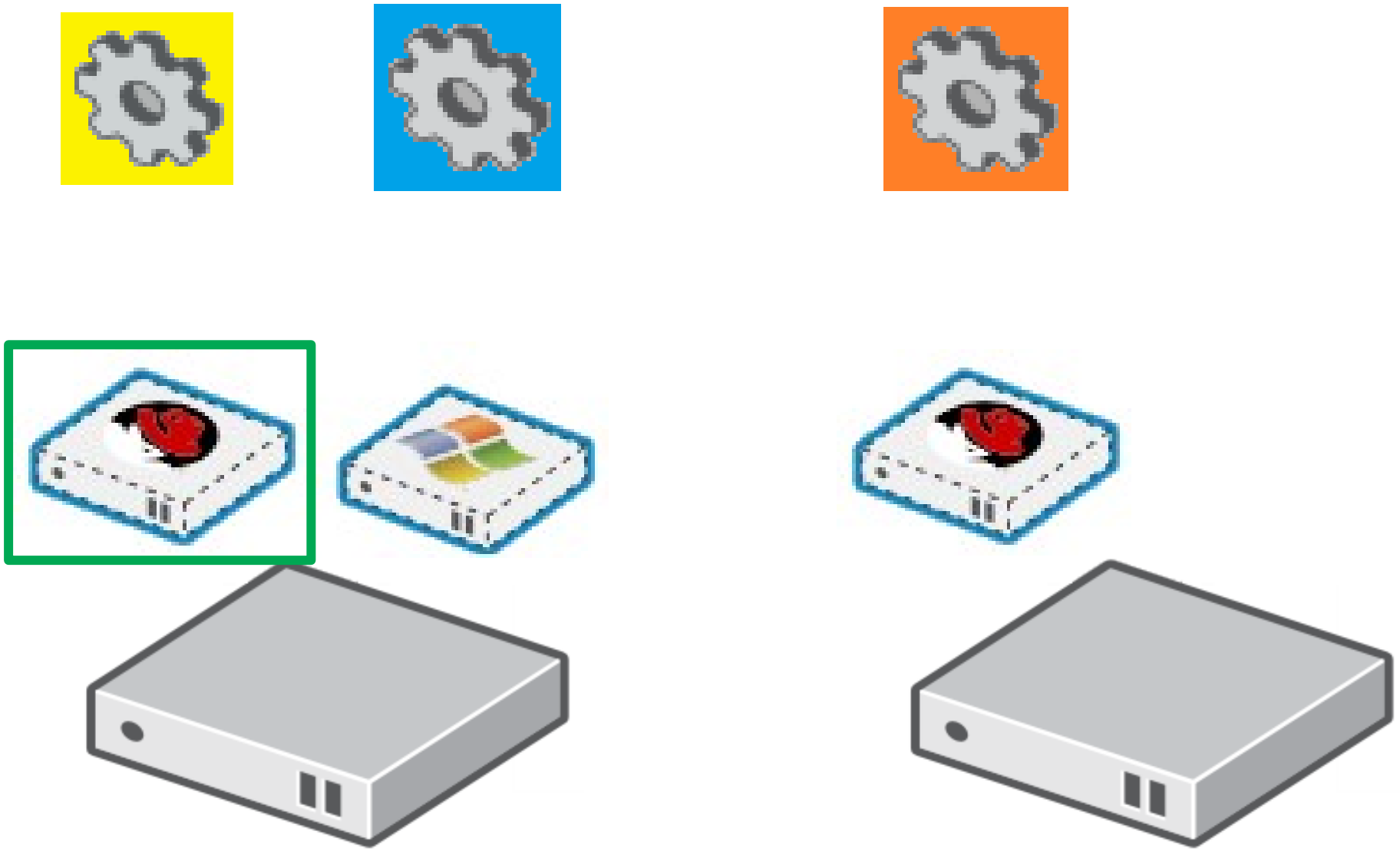
oVirt High Availability - Application-Level



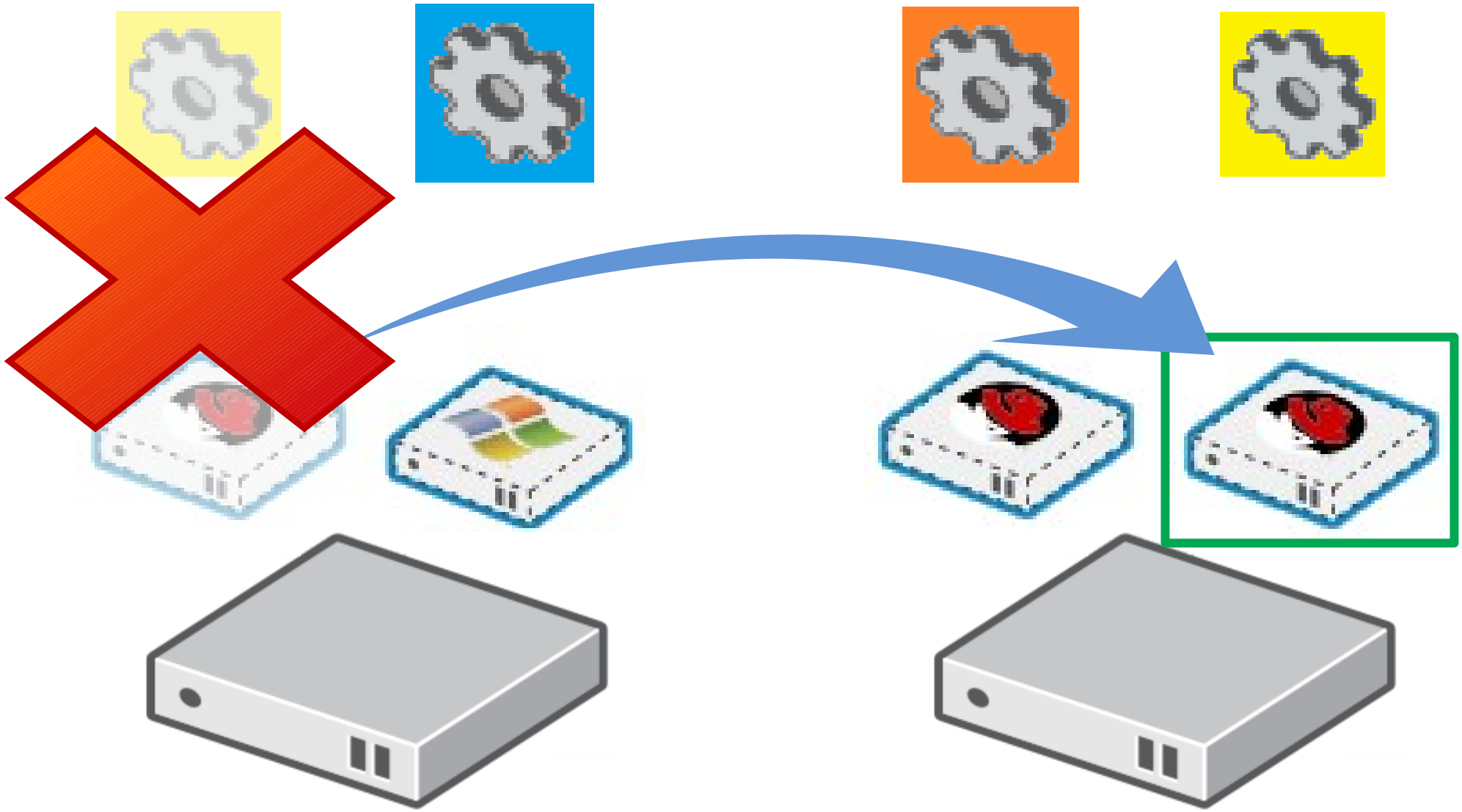
High Availability - Application-Level

- Higher resource consumption
- More responsibility on the application
- Backup starts in a different environment
  - Different IP address(es)
  - Different disk(s)

High Availability - VM-Level



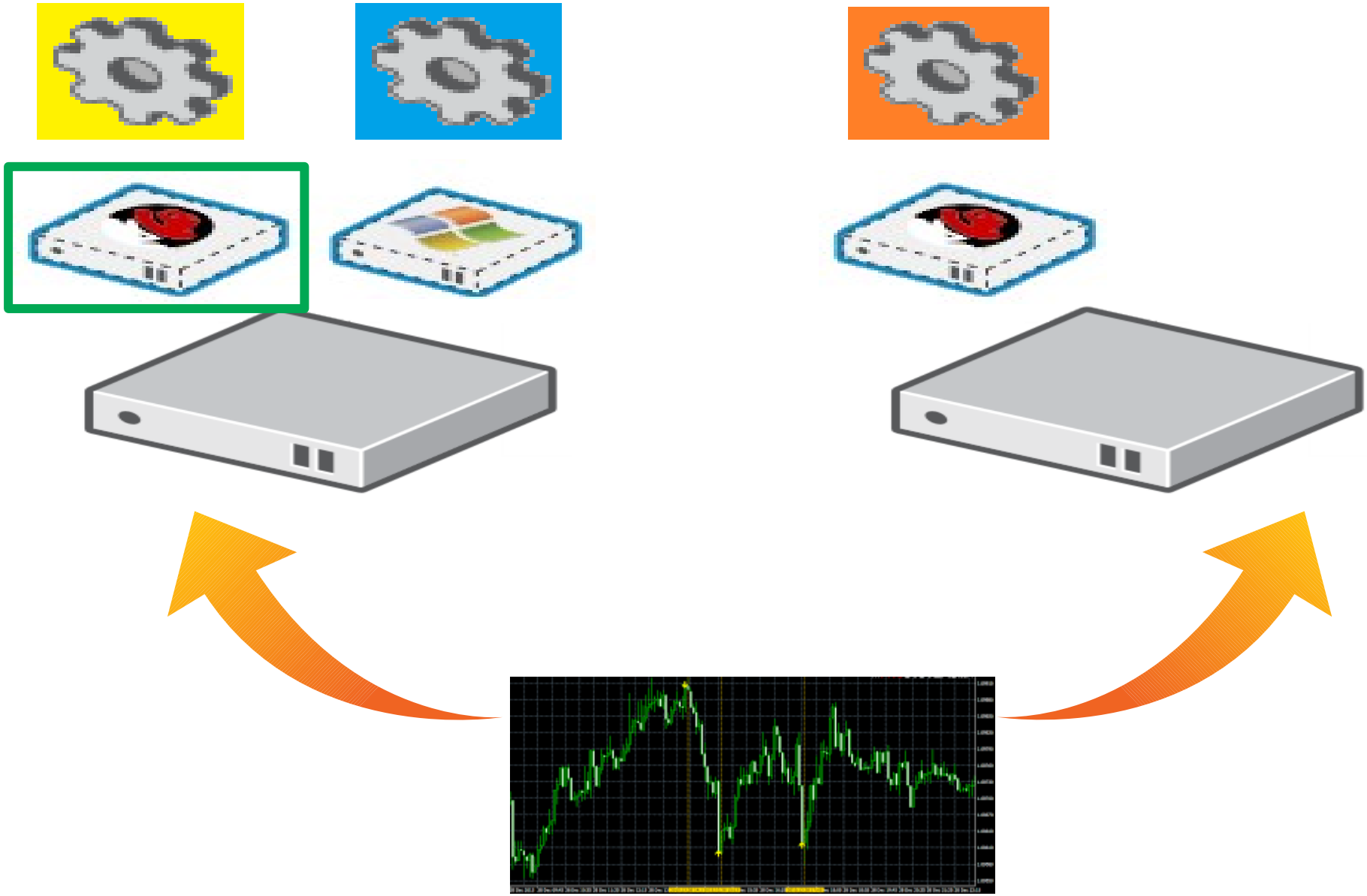
oVirt High Availability - VM-Level



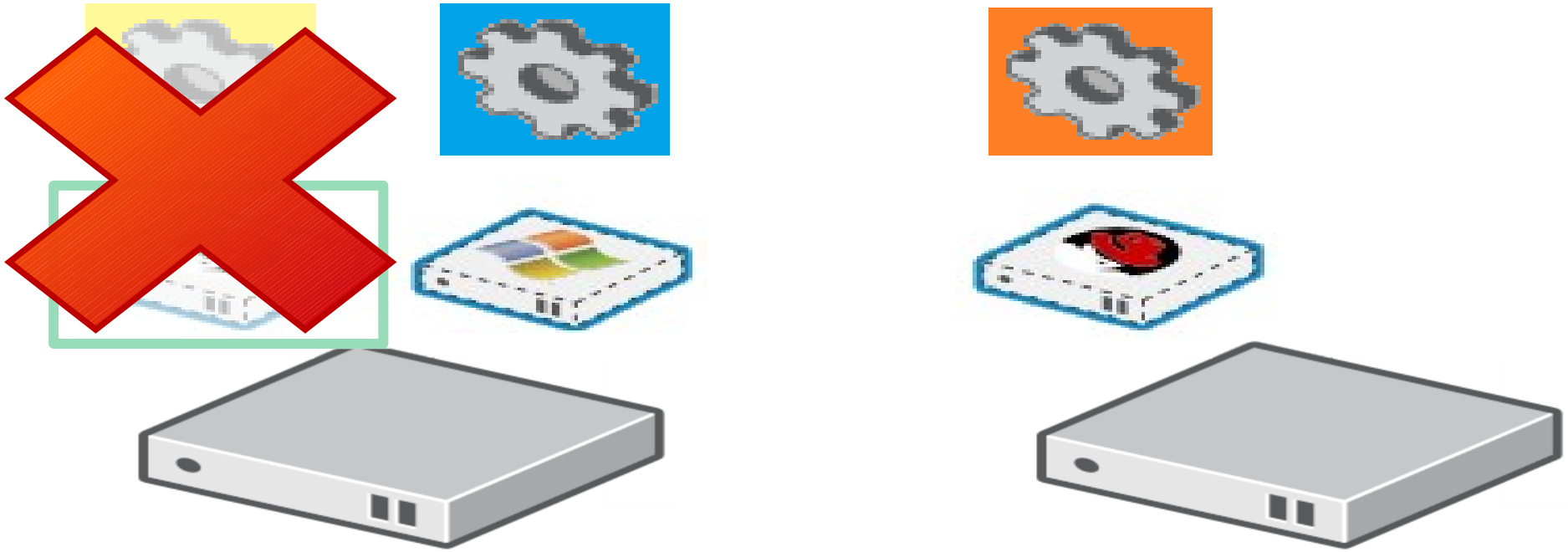
High Availability - VM-Level

- More efficient resource consumption
- Implemented at the infrastructure level
- VM always start in the same environment
  - Same IP address(es)
  - Same disk(s)

oVirt Central Monitoring Unit

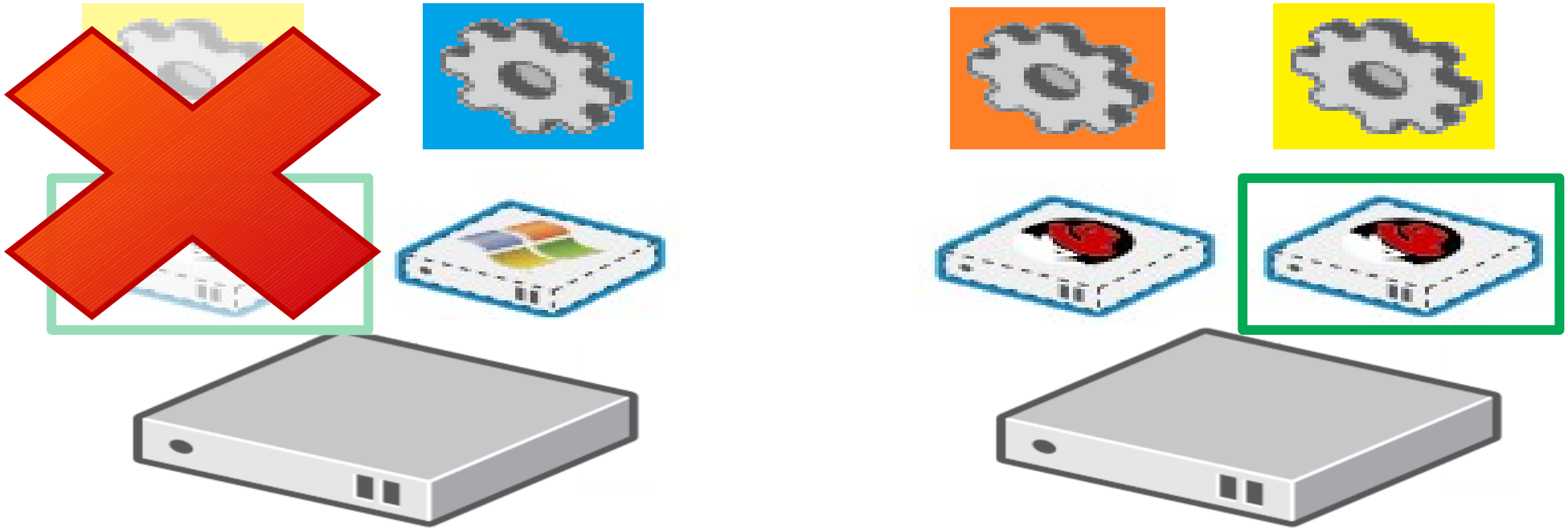


Fault Detection



HA VM went down!

oVirt Automatic Restart

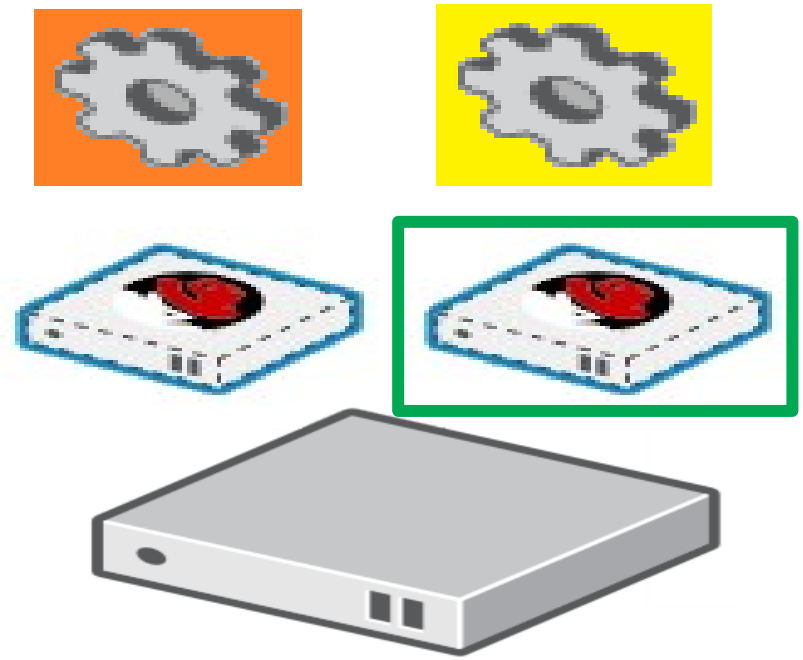


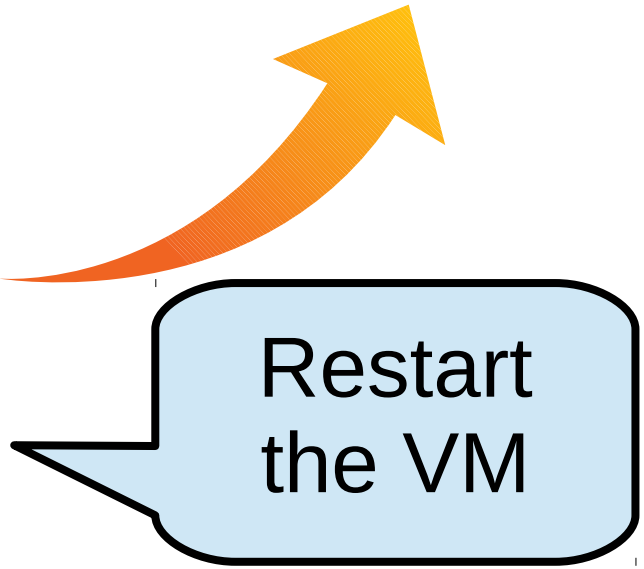
Restart the VM

Automatic Restart – Not That Simple

What if:

- Inaccessible resources
- VM is locked
- VM is being intentionally shut down

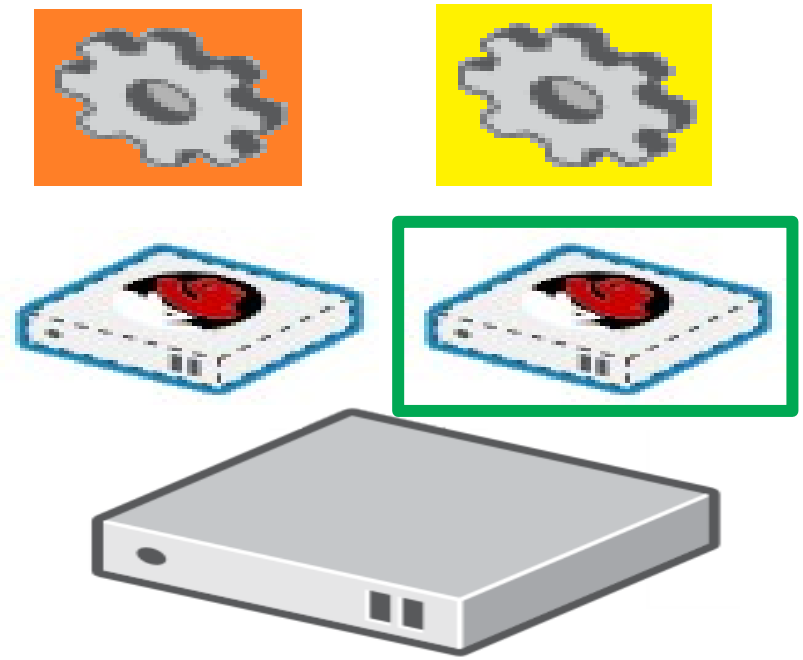


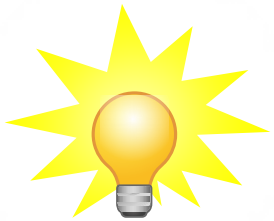


oVirt Automatic Restart – Not That Simple

What if:

- Inaccessible resources
- VM is locked
- VM is being intentionally shut down

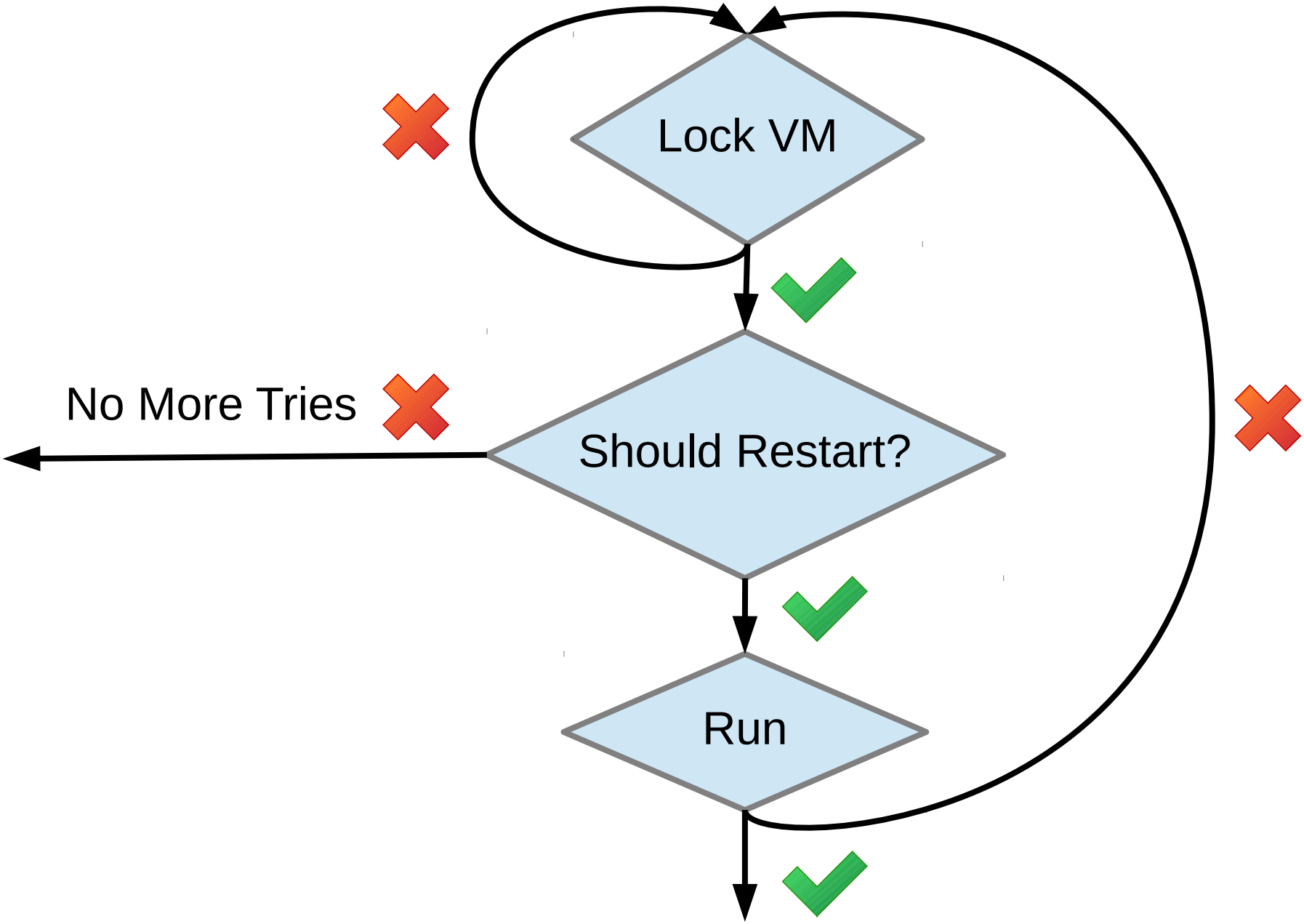




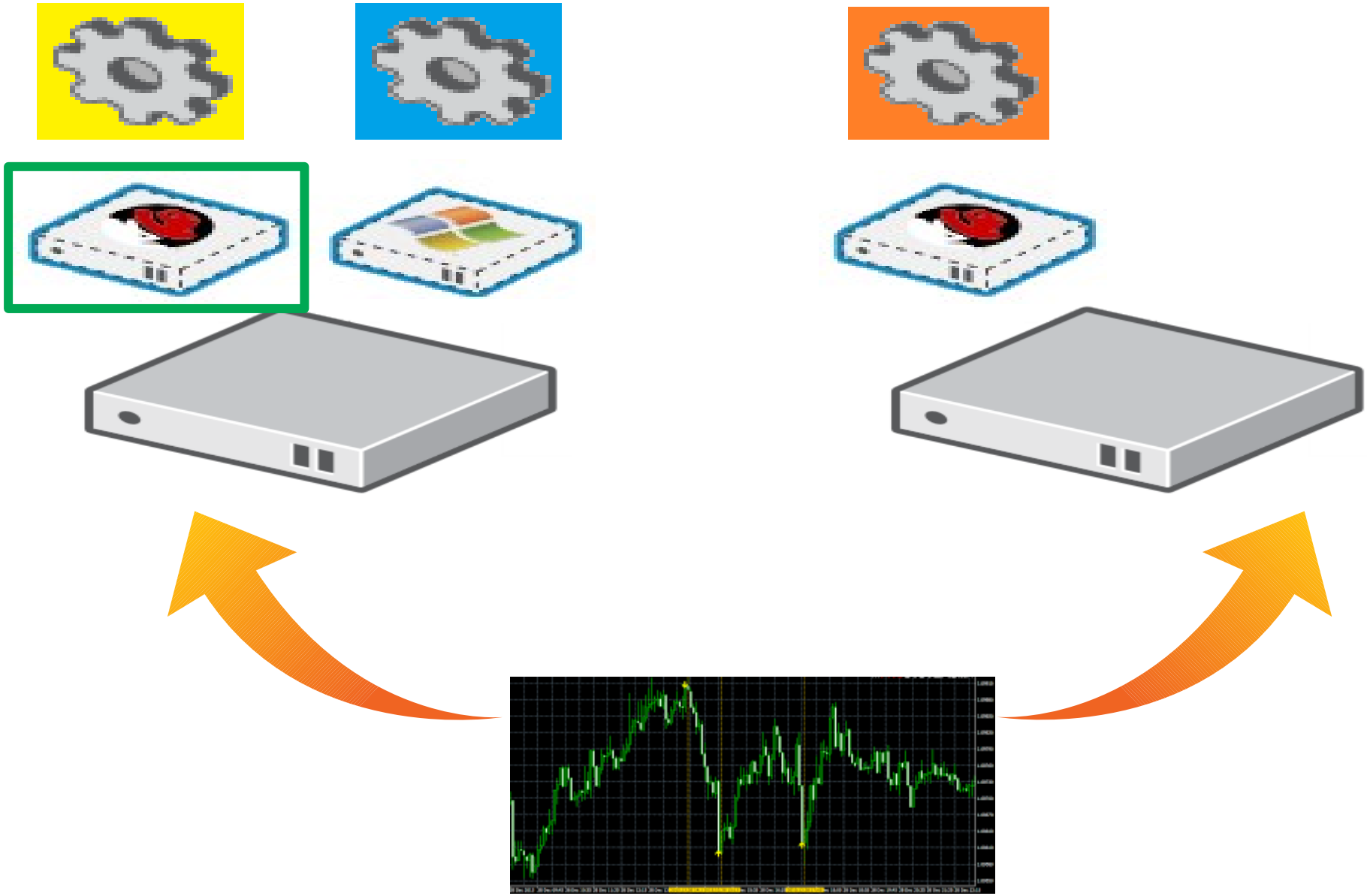
AutoStartVmsRunner

https://github.com/oVirt/ovirt-engine/blob/master/backend/manager/modules/bll/src/main/java/org/ovirt/engine/core/bll/AutoStartVmsRunner.java

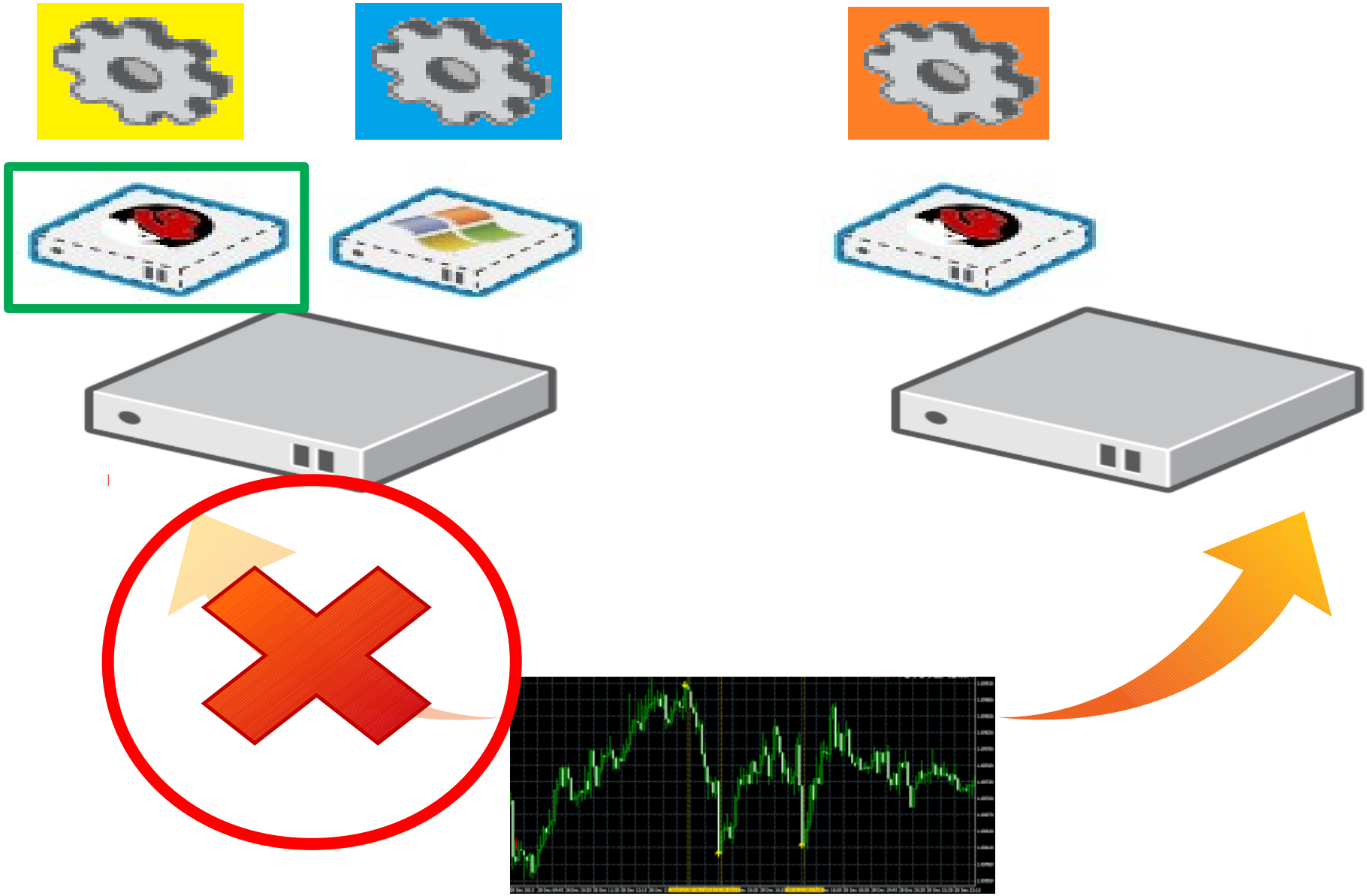
AutoStartVmsRunner



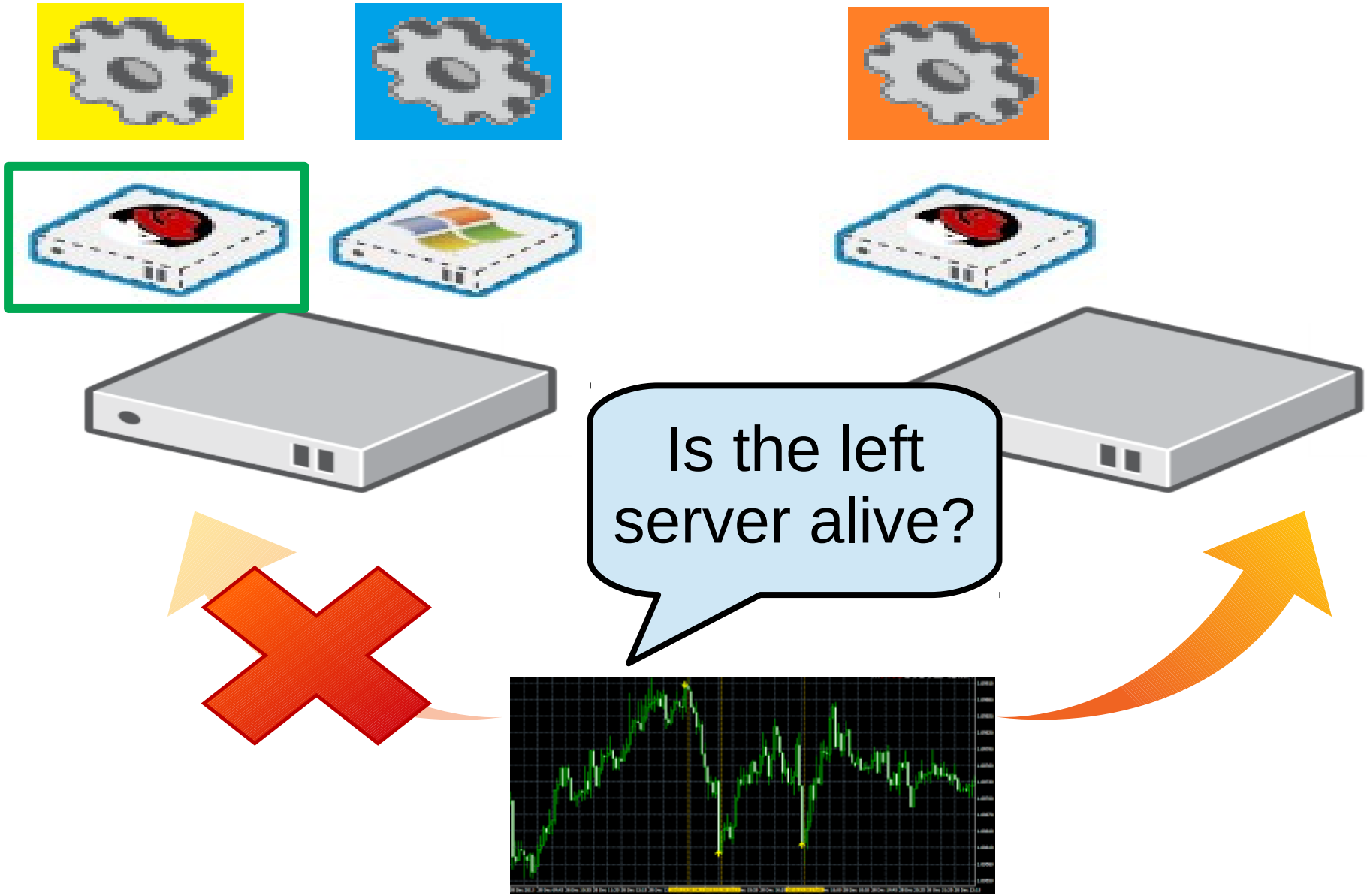
oVirt Fault Detection – Even More Complex



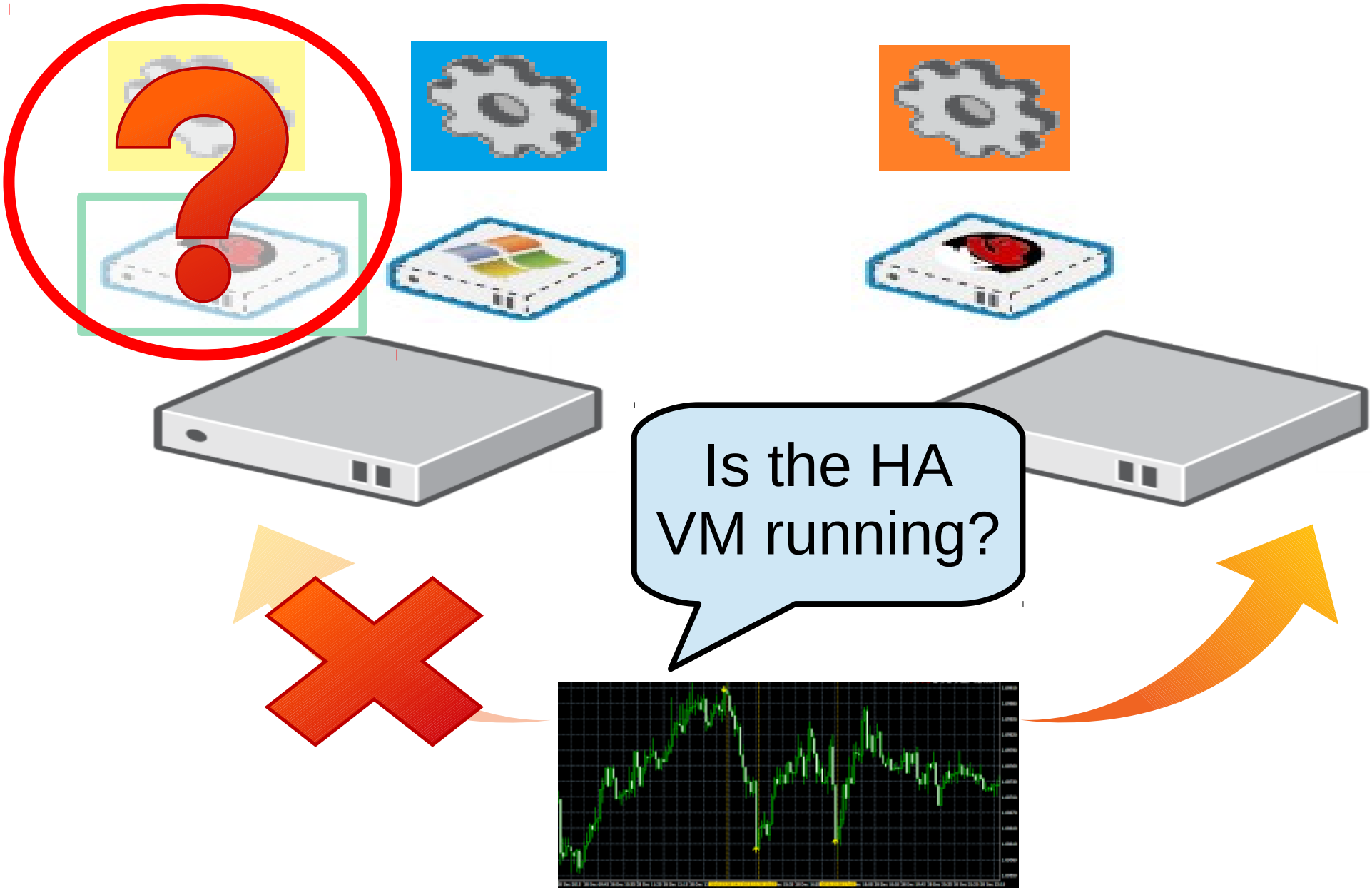
oVirt Fault Detection – Even More Complex



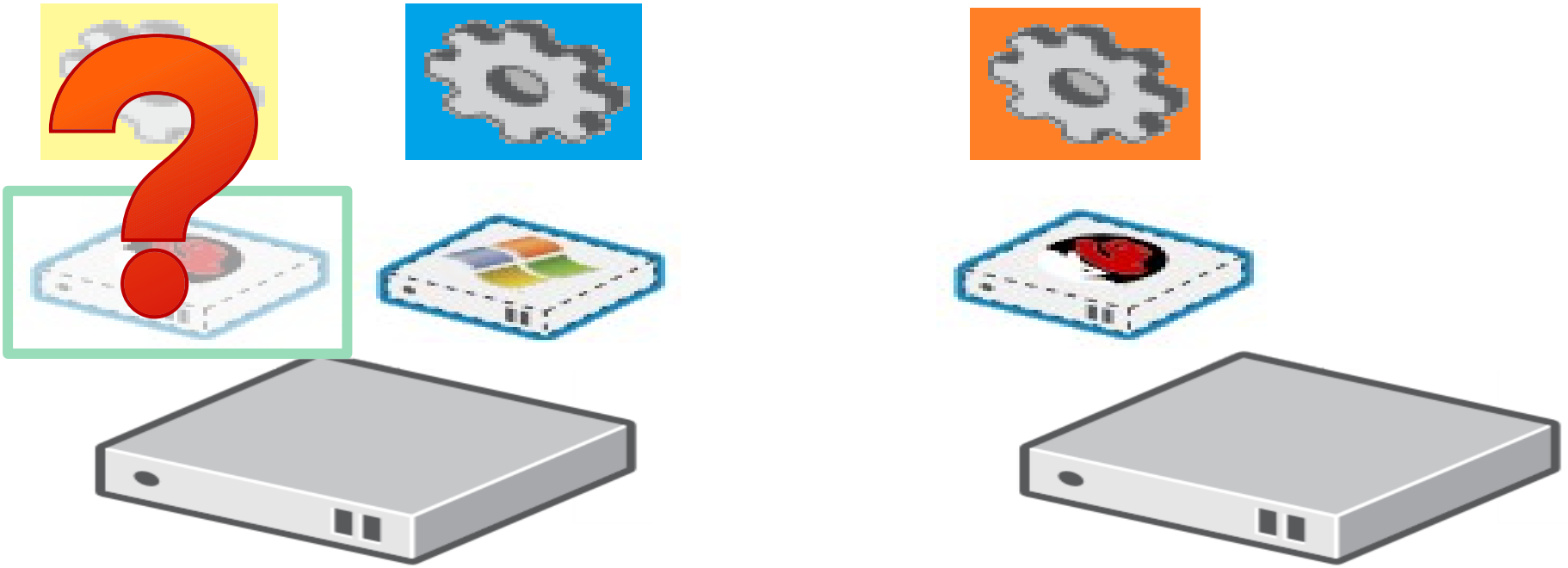
oVirt Fault Detection – Even More Complex

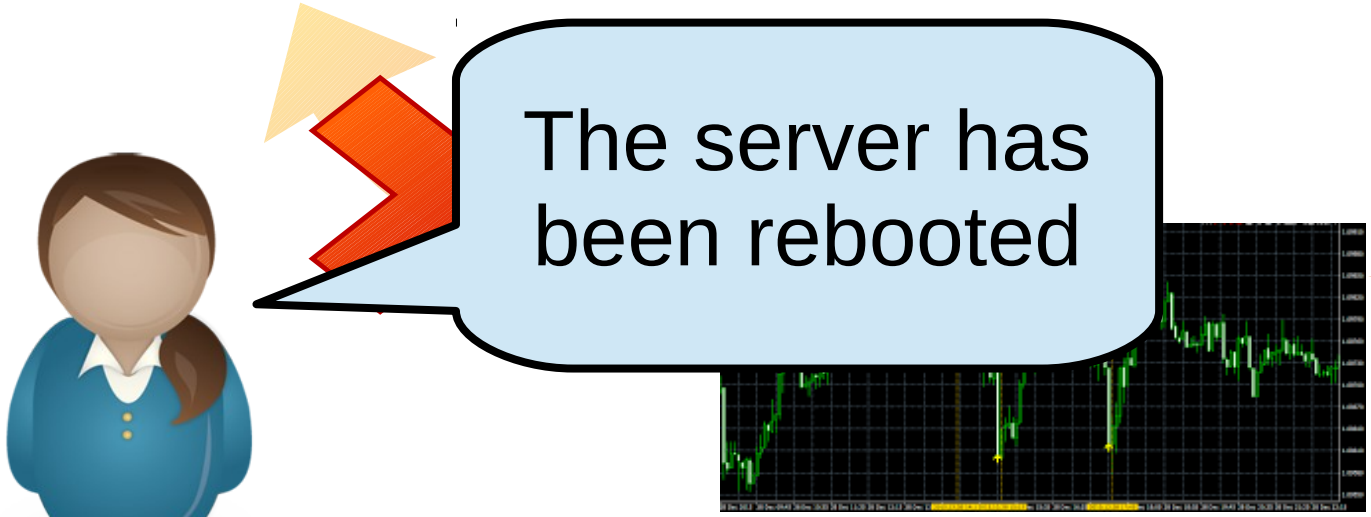


oVirt Fault Detection – Even More Complex

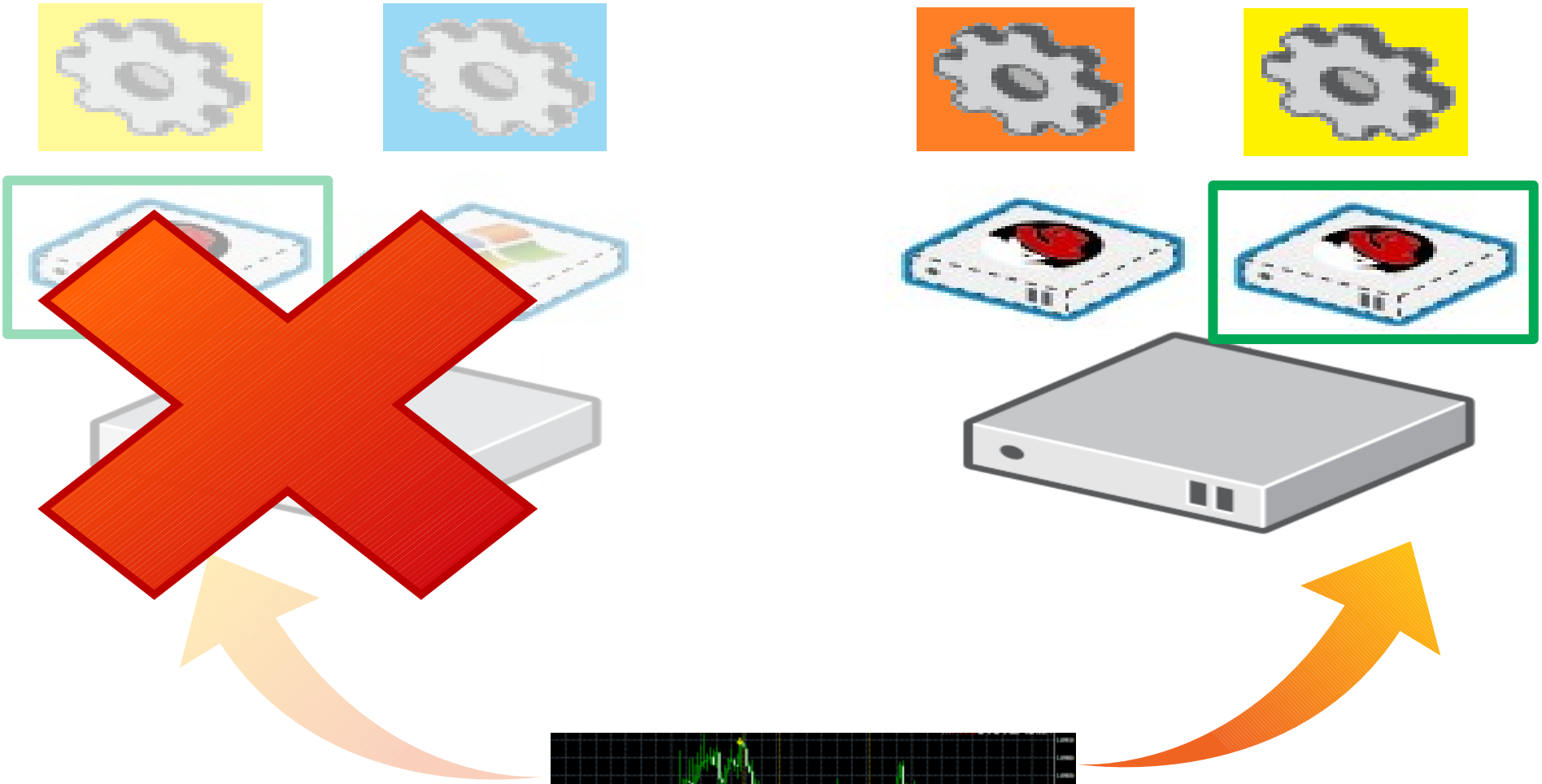


oVirt Fault Detection – Manual Confirmation





oVirt Fault Detection – Manual Confirmation



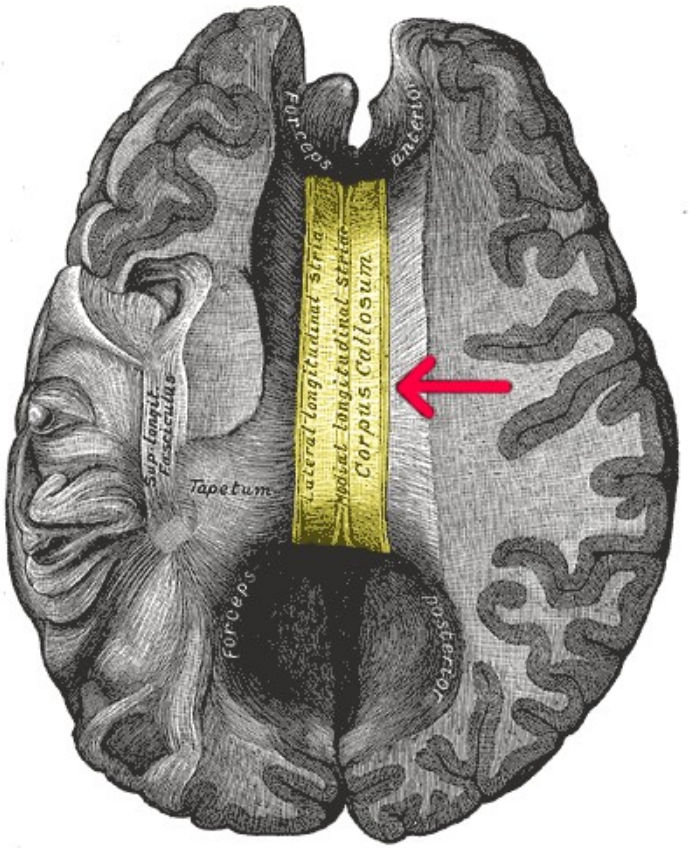
Restart the VM

Fault Detection – Manual Confirmation

- Slow
- Error-prone
  - Mistakes may lead to a split-brain

Split Brain of Virtual Machines

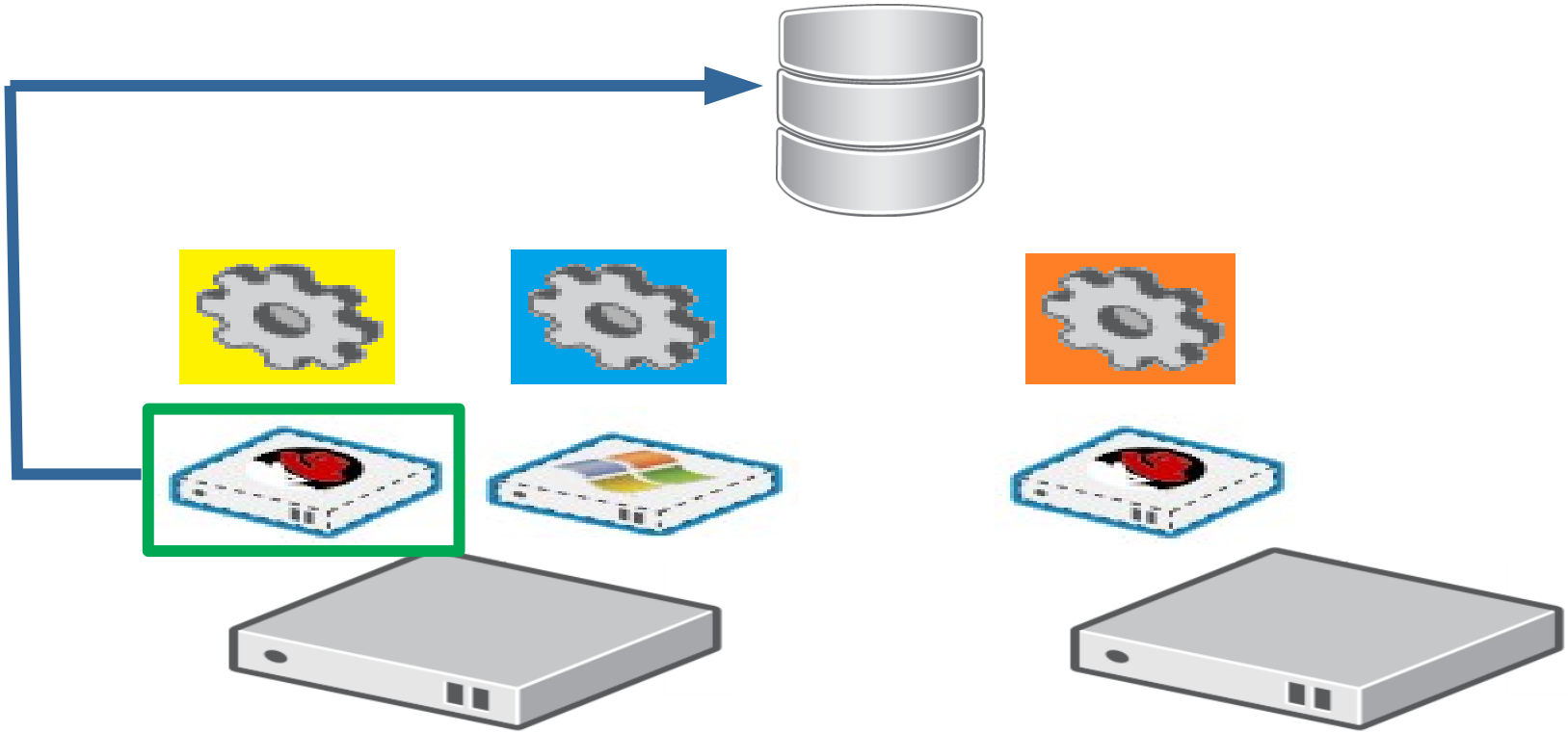
A scenario in which several instances of the same VM run simultaneously



oVirt Split Brain Due to a False Confirmation

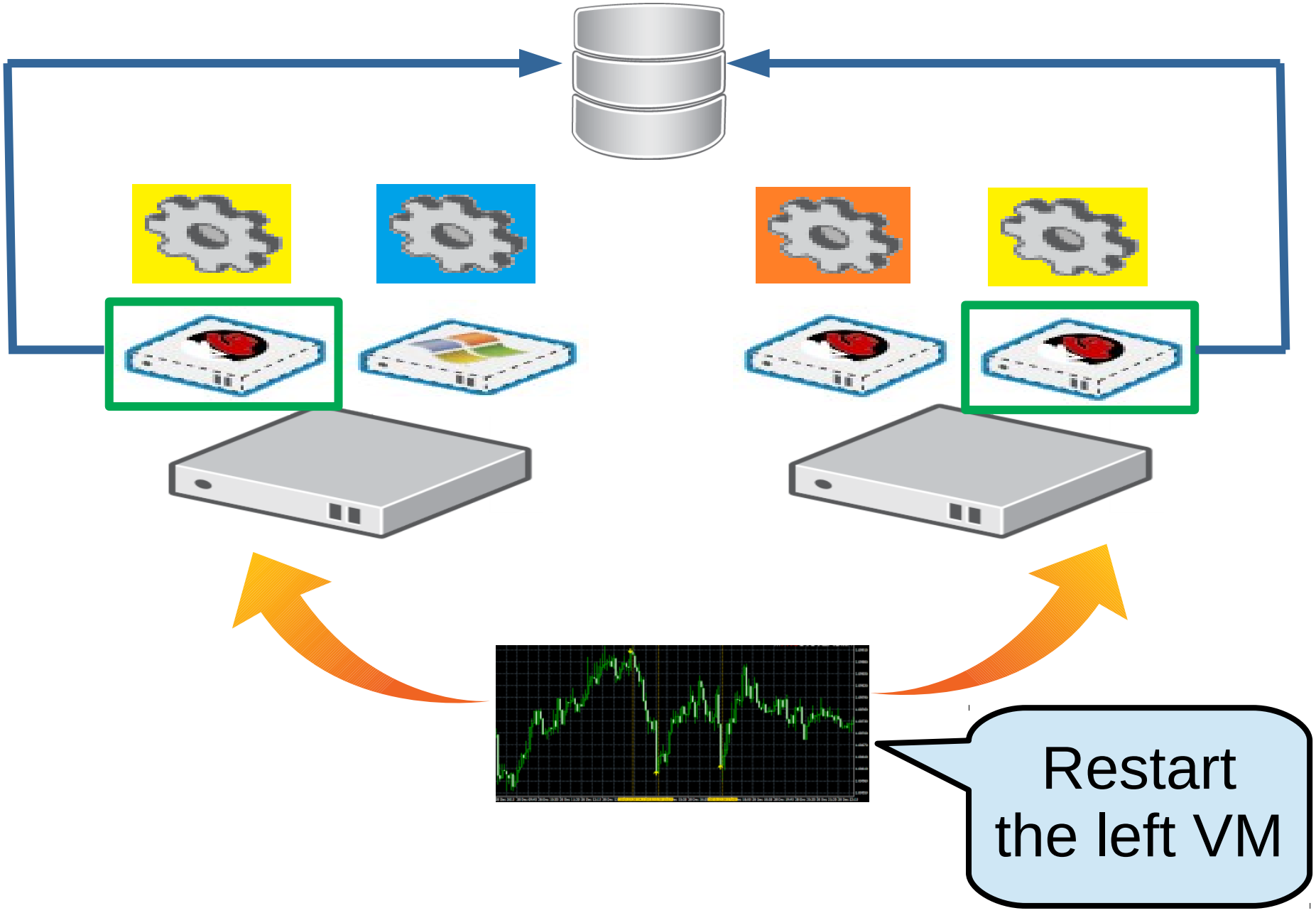
May lead to data corruption!

oVirt Split Brains May Happen Due to Bugs

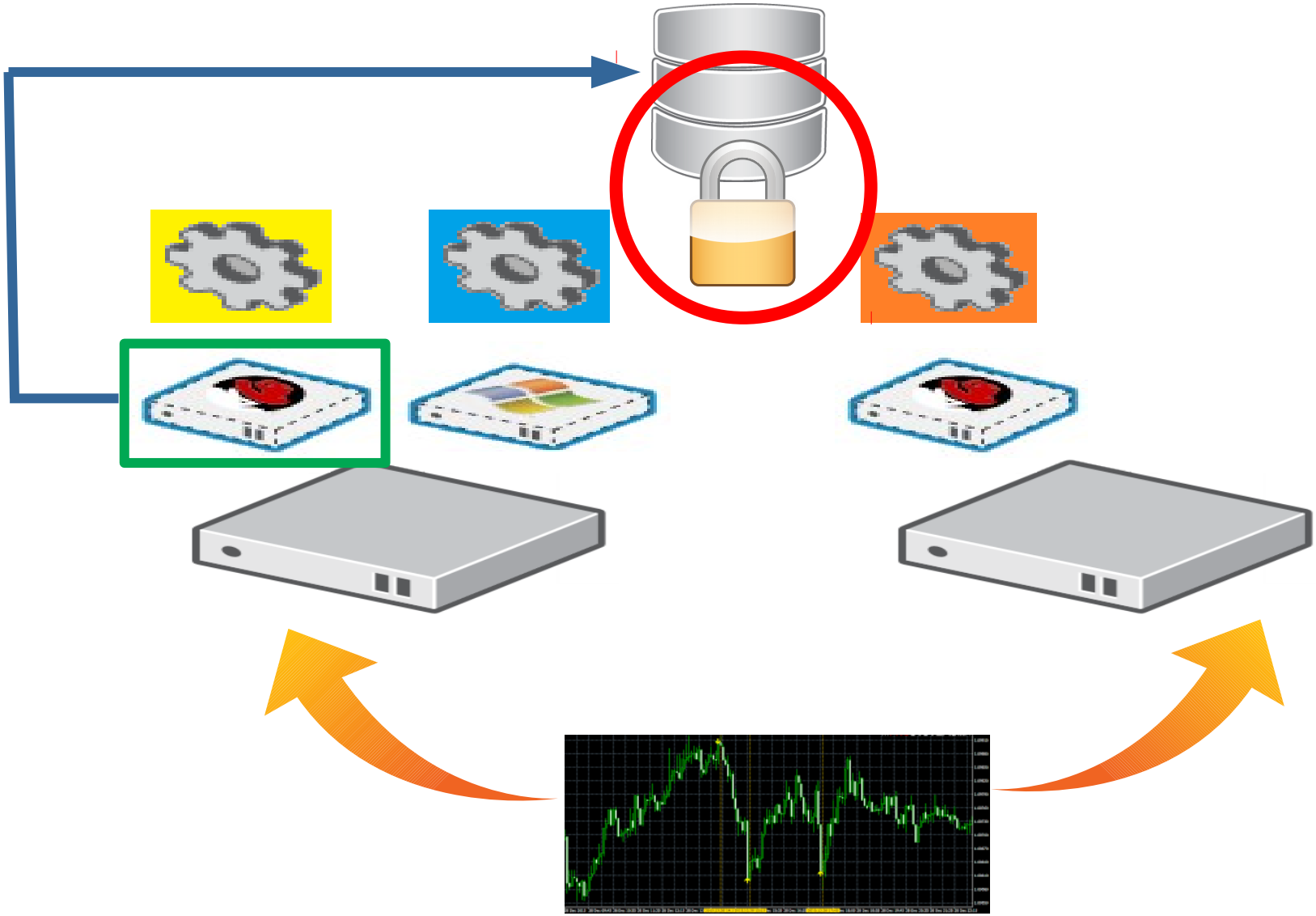


Only the right VM is reported

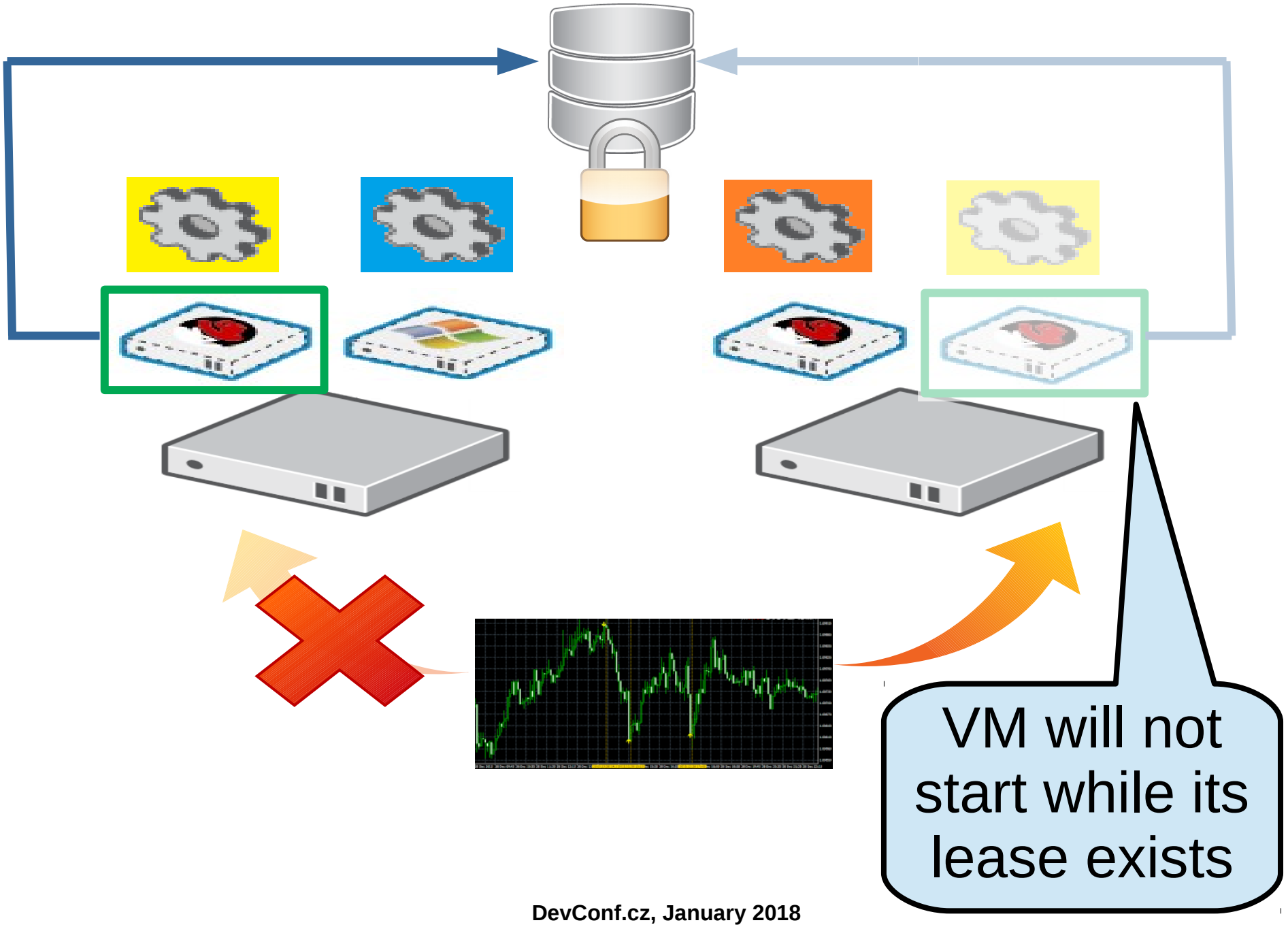
oVirt Split Brains May Happen Due to Bugs



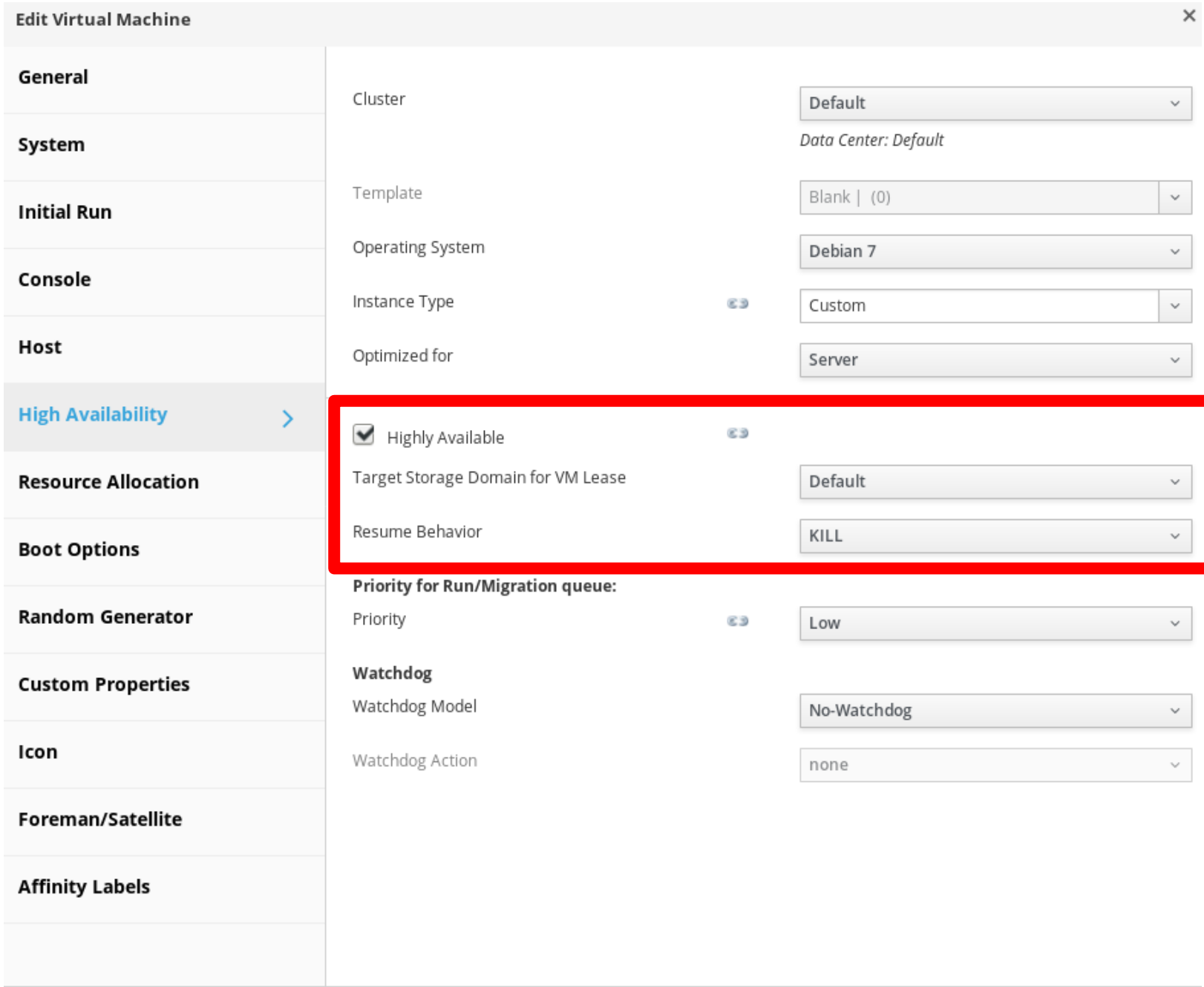
VM Leases: Our Solution to Split Brains



VM Leases: Our Solution to Split Brains

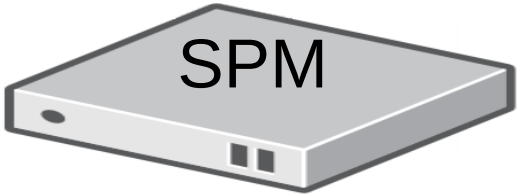


VM Lease Creation



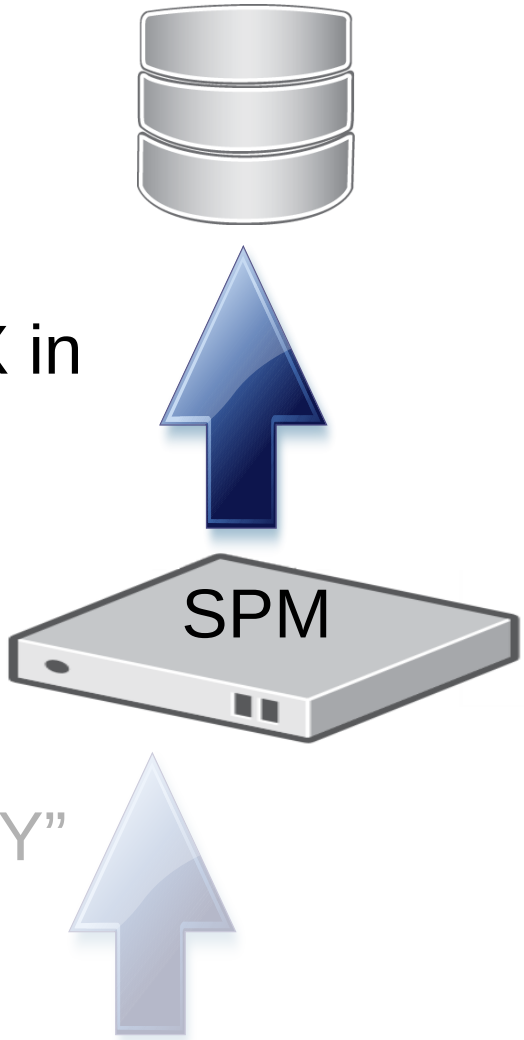
VM Lease Creation

VM Lease Creation

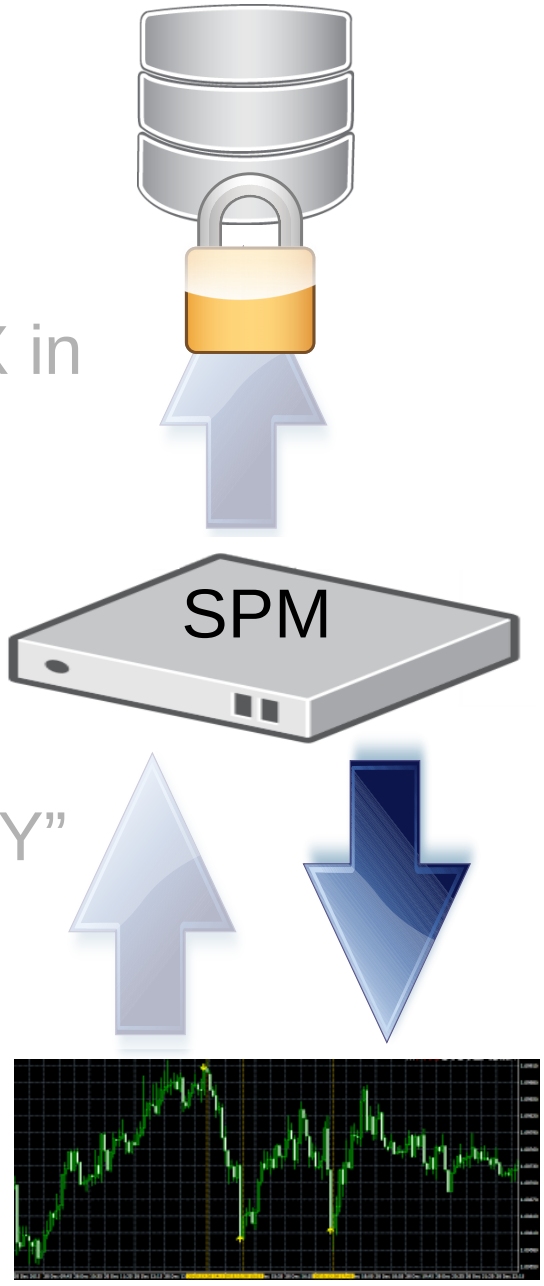


“Create a VM Lease for
VM X in storage domain Y”

VM Lease Creation



VM Lease Creation



xleases volume

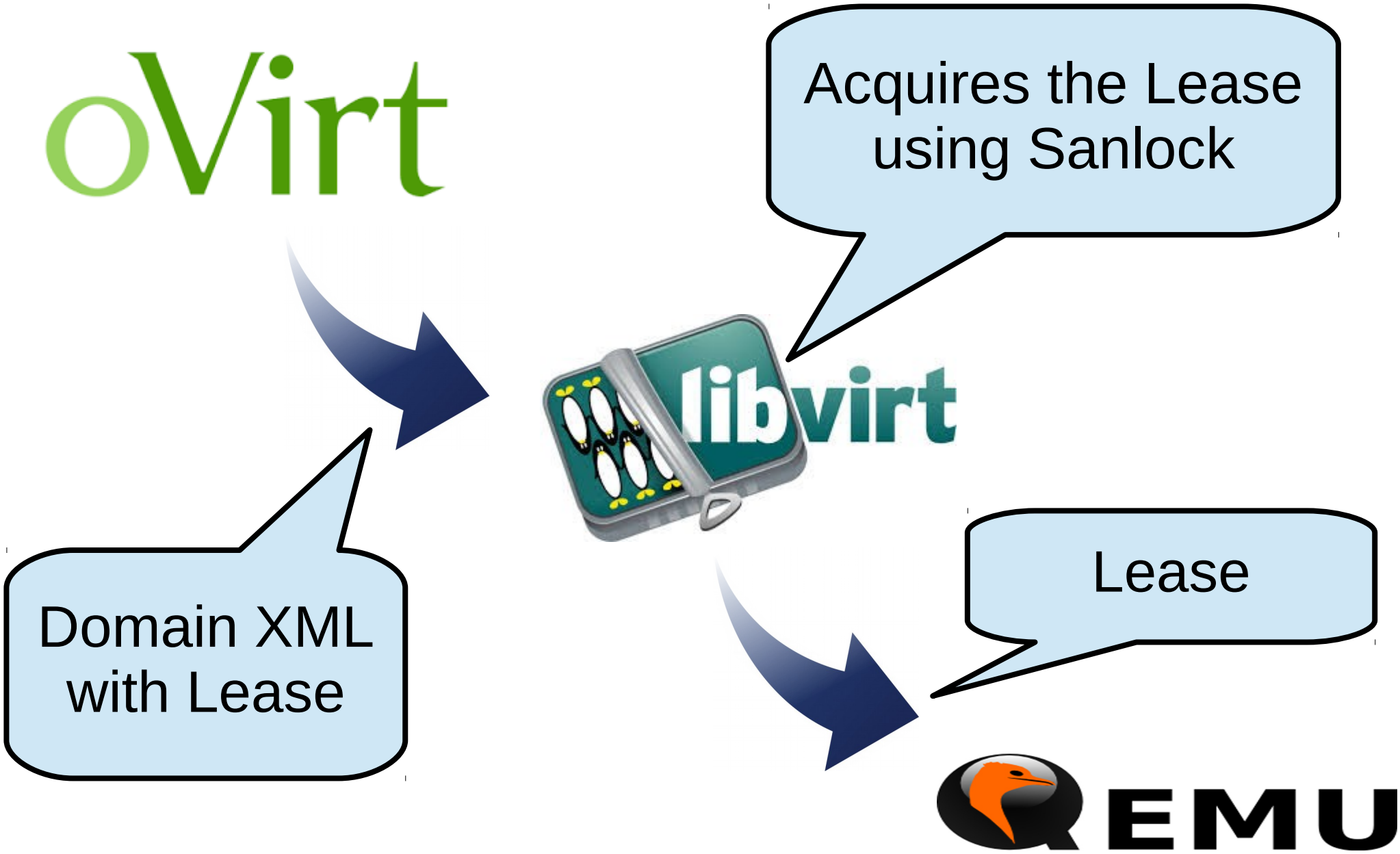
- Sanlock does not manage leases allocation
- Volume layout:

- Same format in block and file storage
- Deep Dive - VM leases (youtube)

Running a VM with a Lease

```
<domain type='kvm' id='6'>
  <name>fedora8</name>
   ... skipped ...
  <devices>
   ... skipped ...
   <lease>
    <lockspace>571184ae-79da-41fb-a3fb-c3117991abae</lockspace>
    <key>cbd783e4-45f8-4b51-93ca-4460d4dad772</key>
    <target path='/rhev/data-center/mnt/10.35.1.90:_srv_Default/571184ae-
      79da-41fb-a3fb-c3117991abae/dom_md/xleases' offset='3145728'/>
   </lease>
   ... skipped ...
 </domain>
```

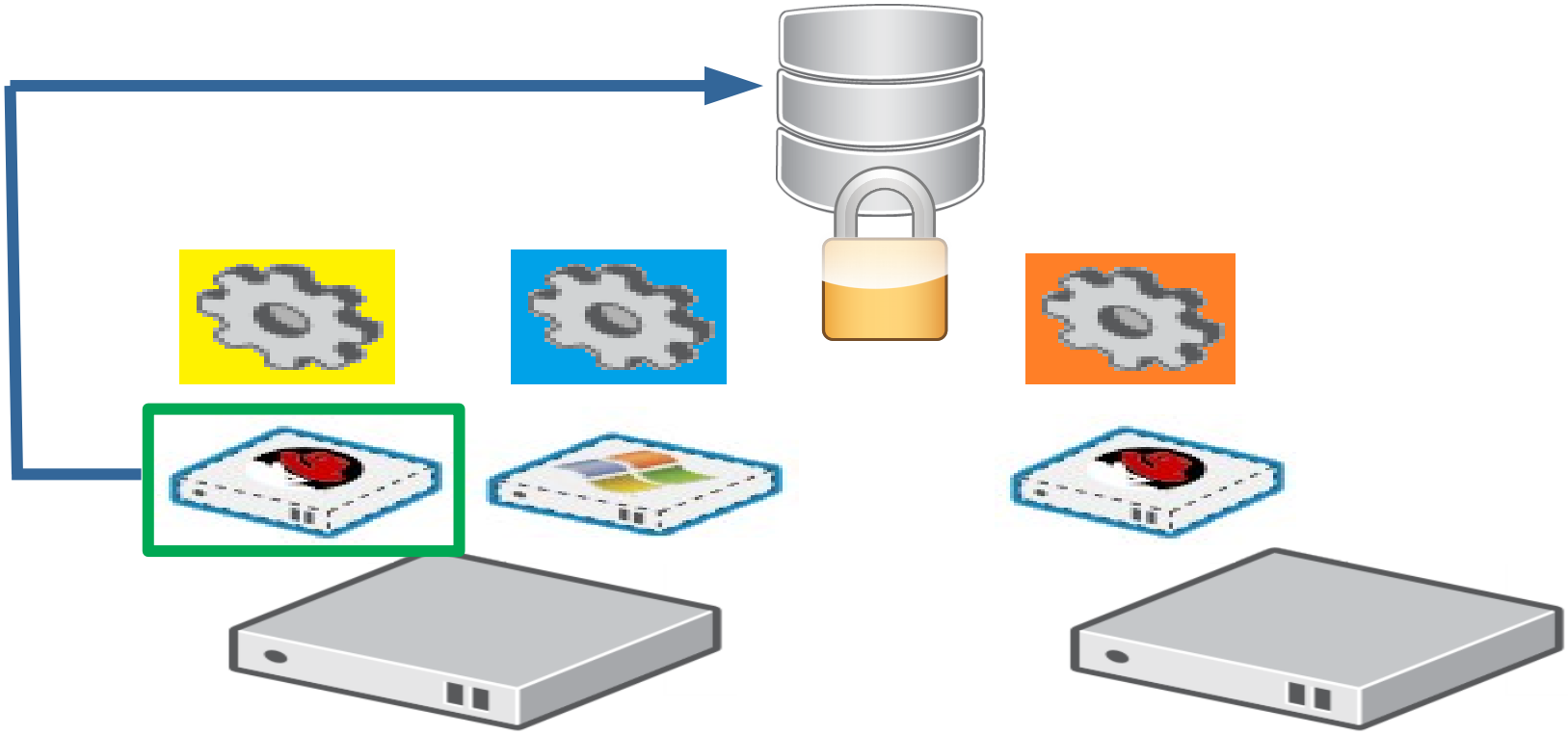
Running a VM with a Lease



oVirt Non-Responsive Host Treatment

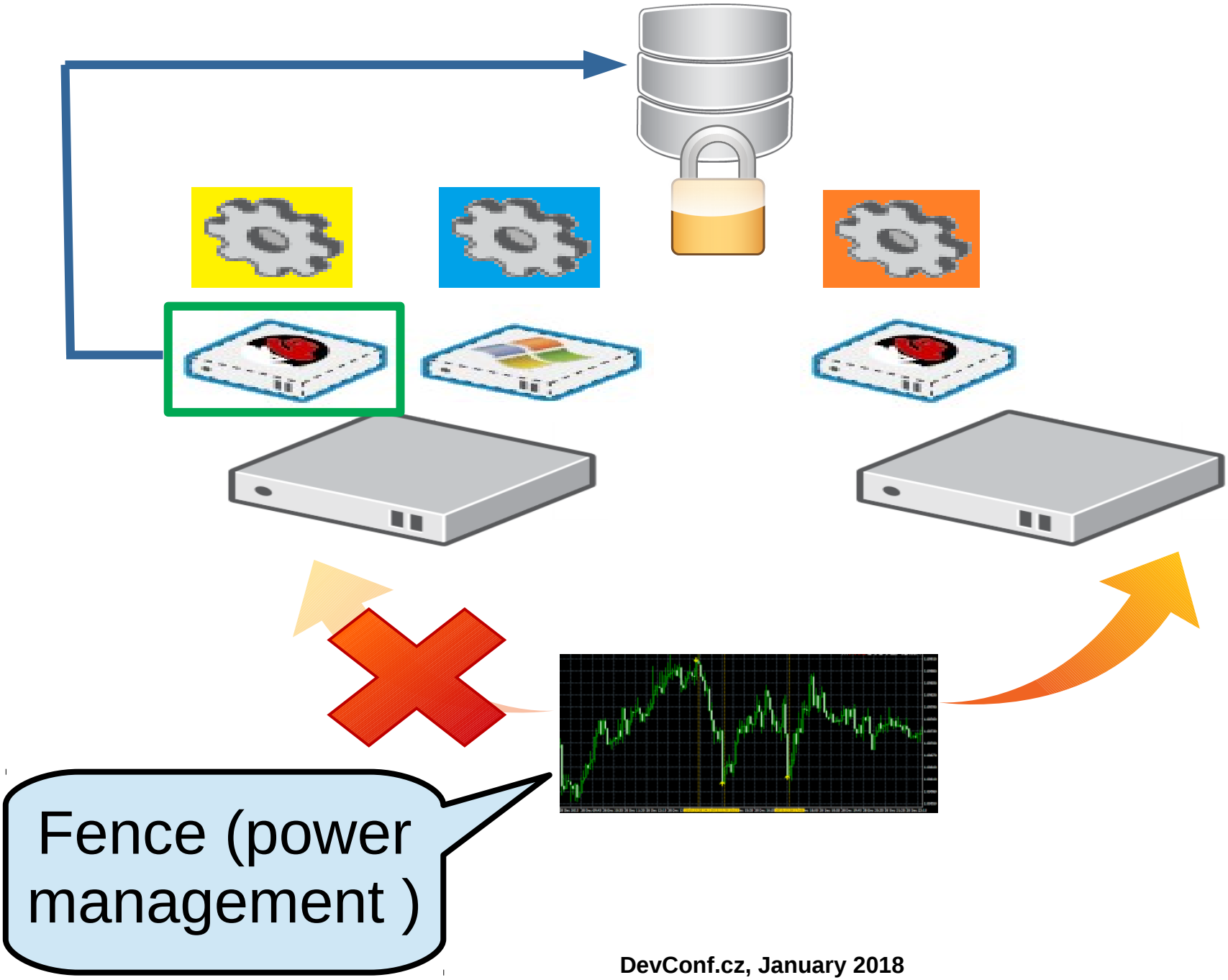
Non-Responsive Host Treatment

Non-Responsive Host Treatment

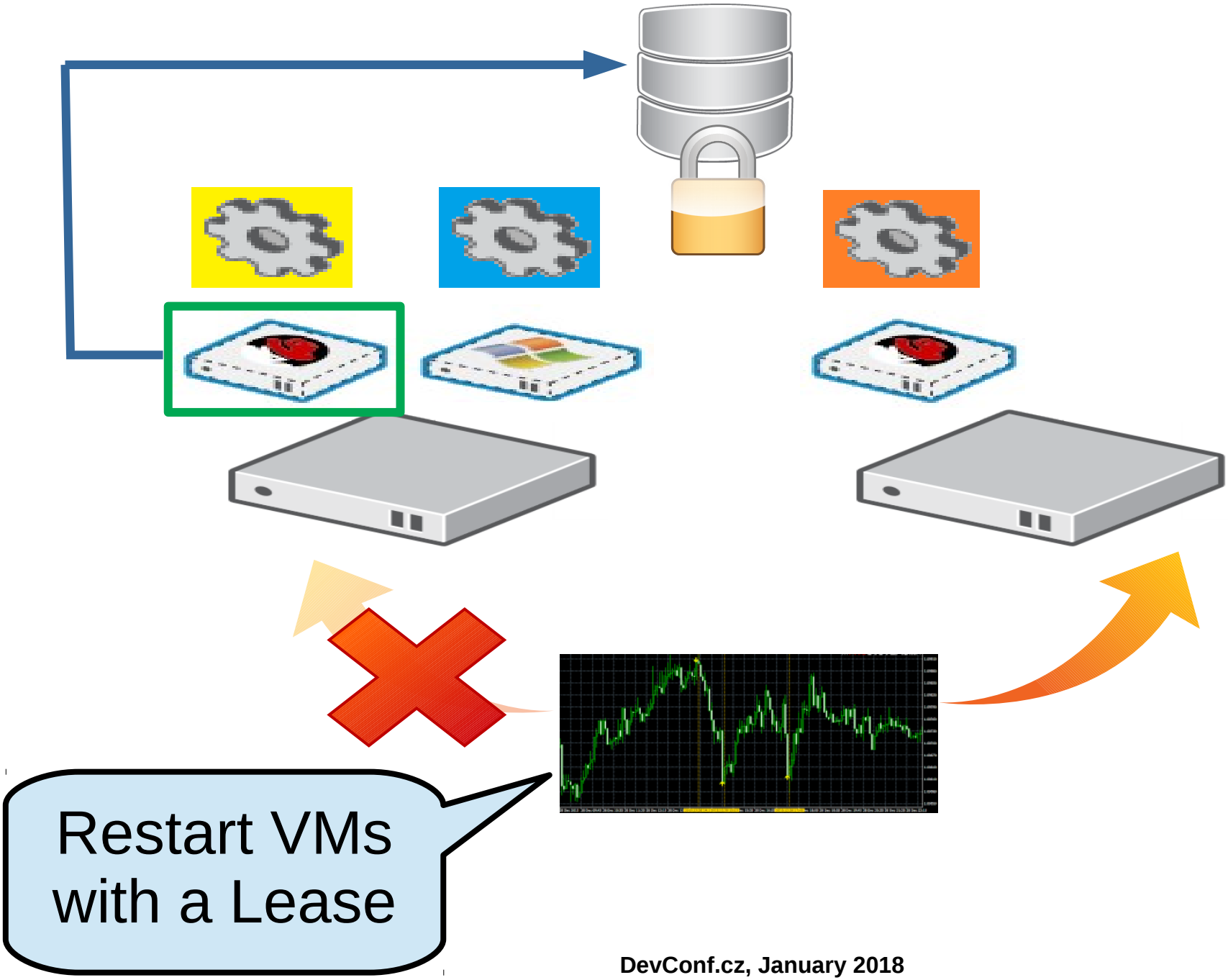


60+ sec of grace period

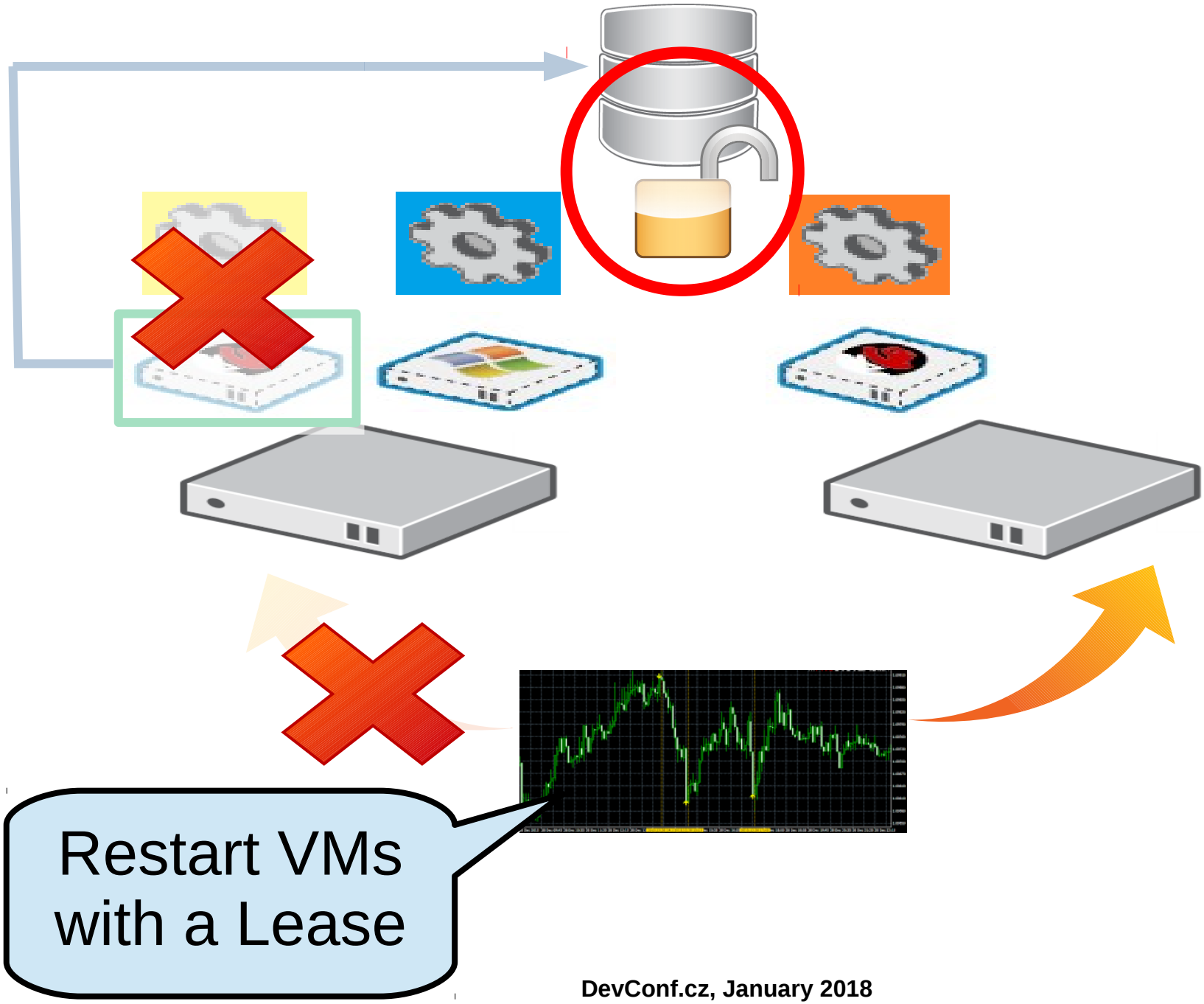
Non-Responsive Host Treatment



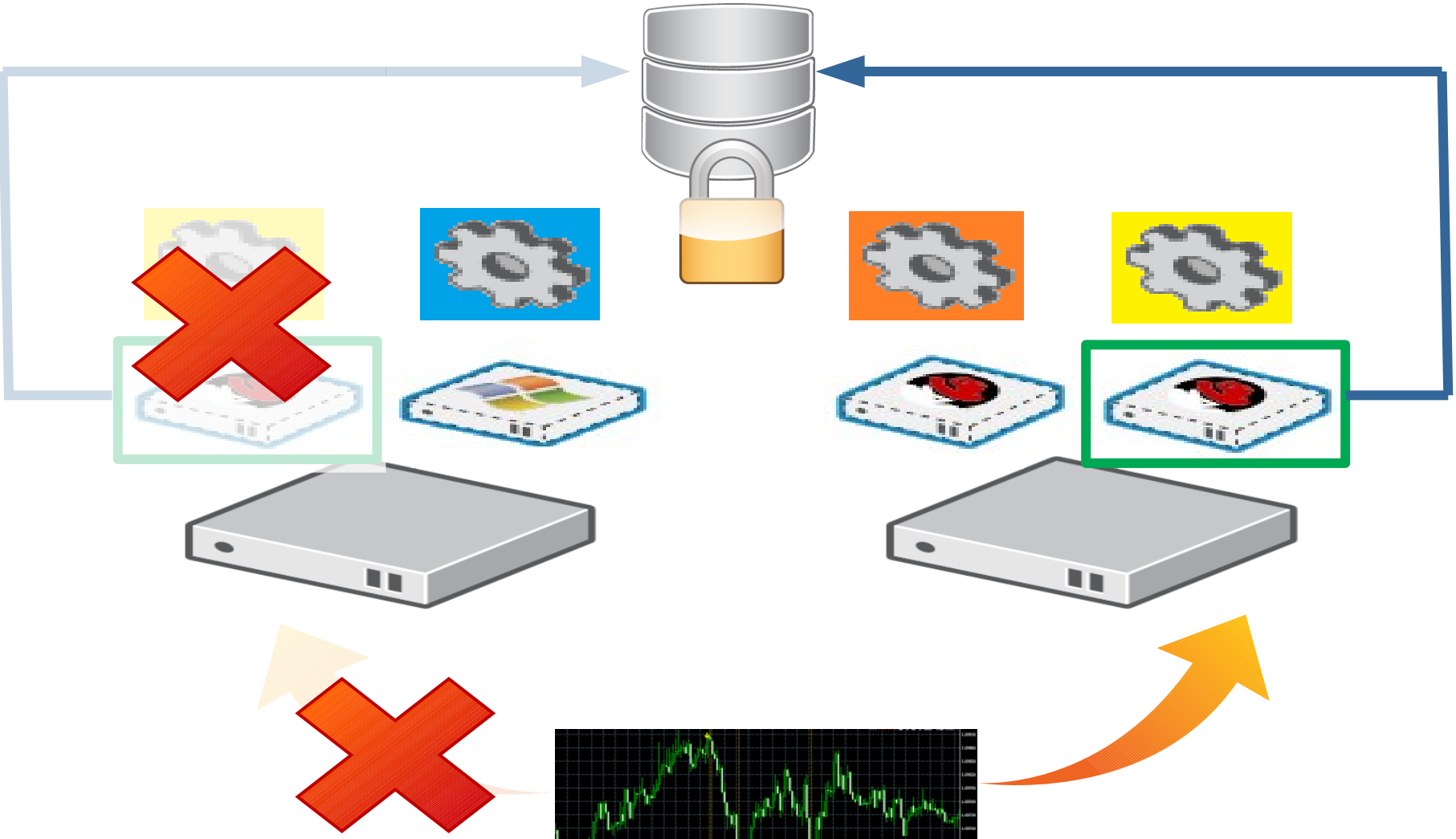
Non-Responsive Host Treatment



(1) Non-Responsive Host + VM is Down

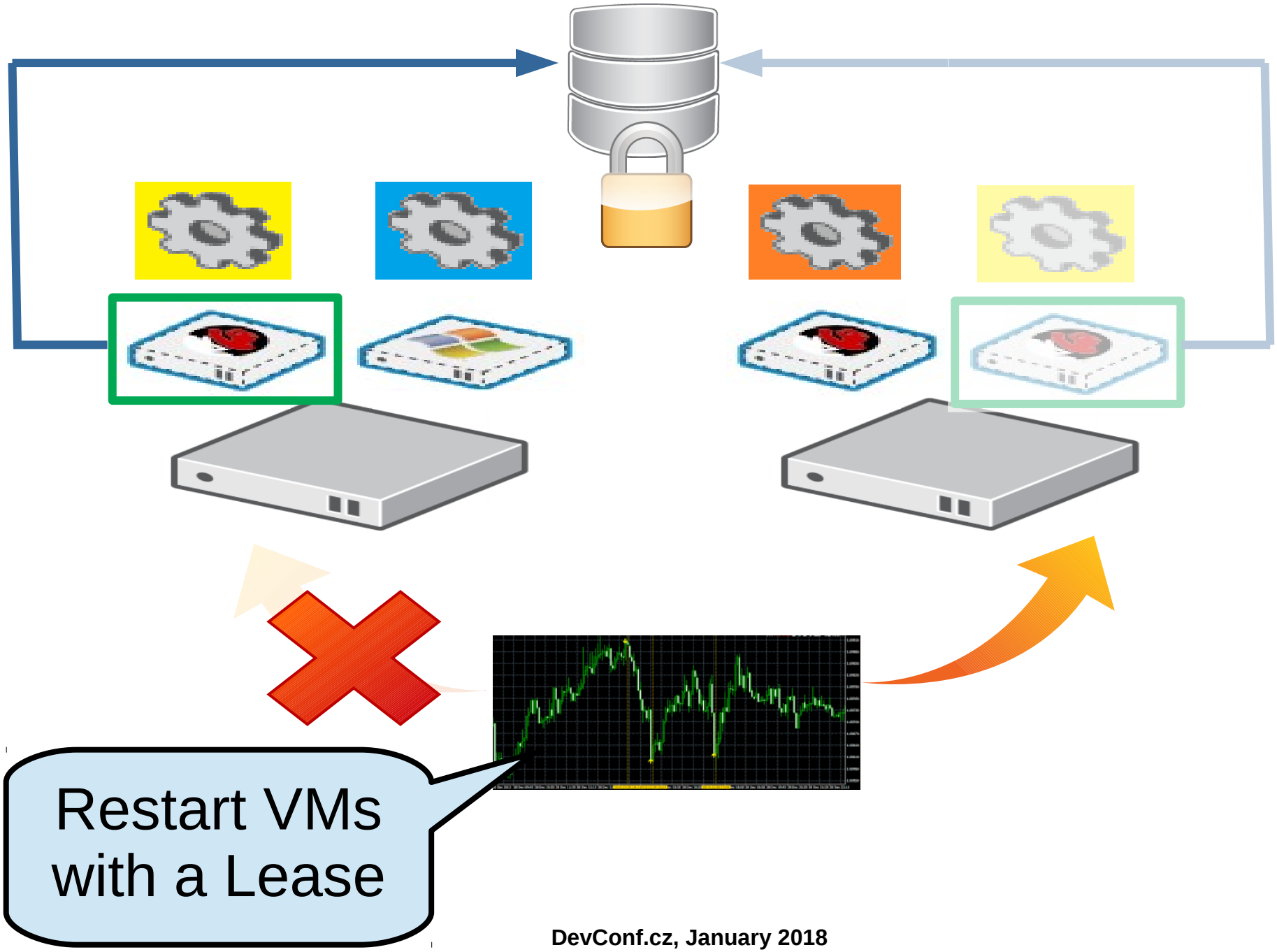


(1) Non-Responsive Host + VM is Down

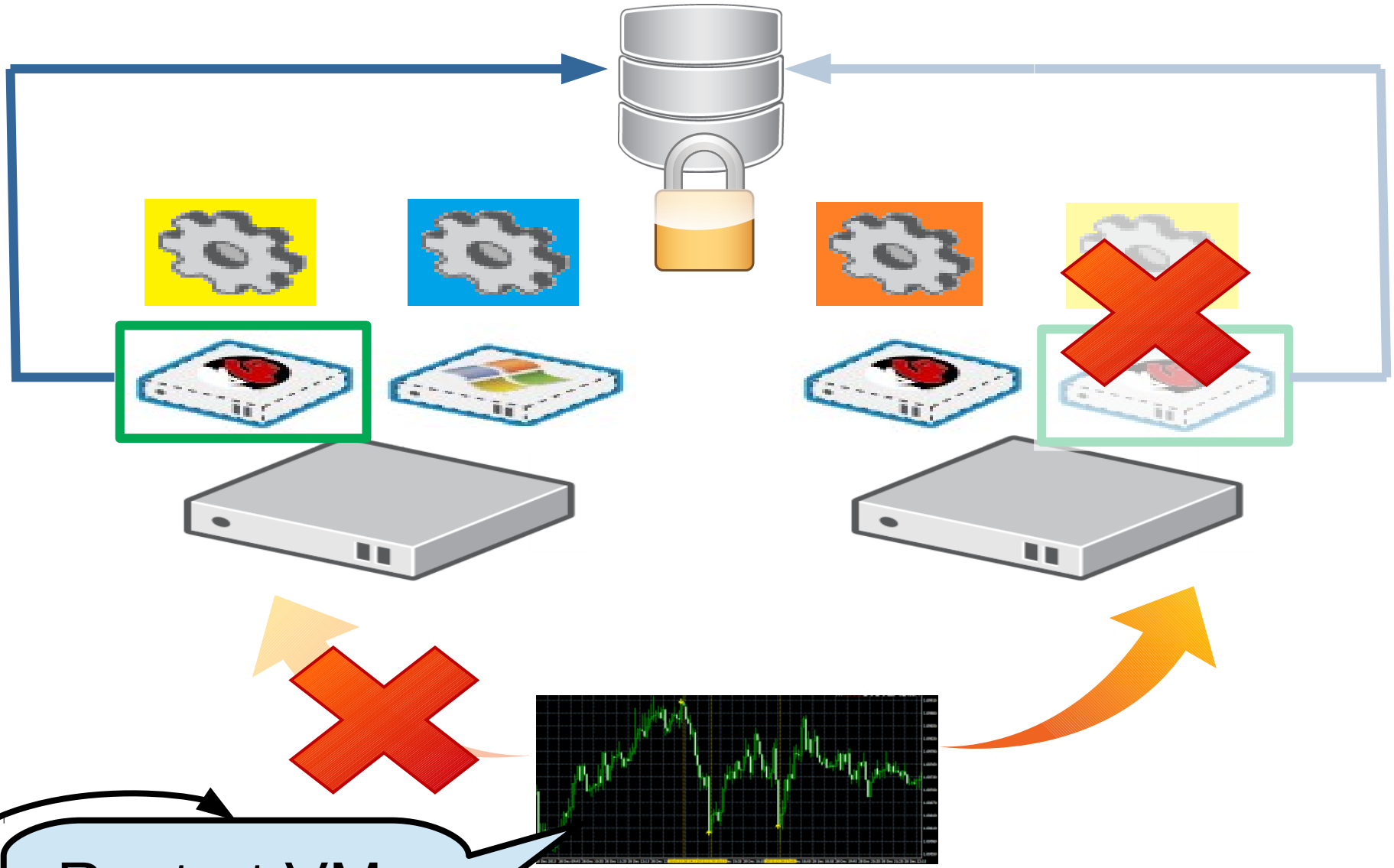


VM starts on another host

(2) Non-Responsive Host + VM is UP

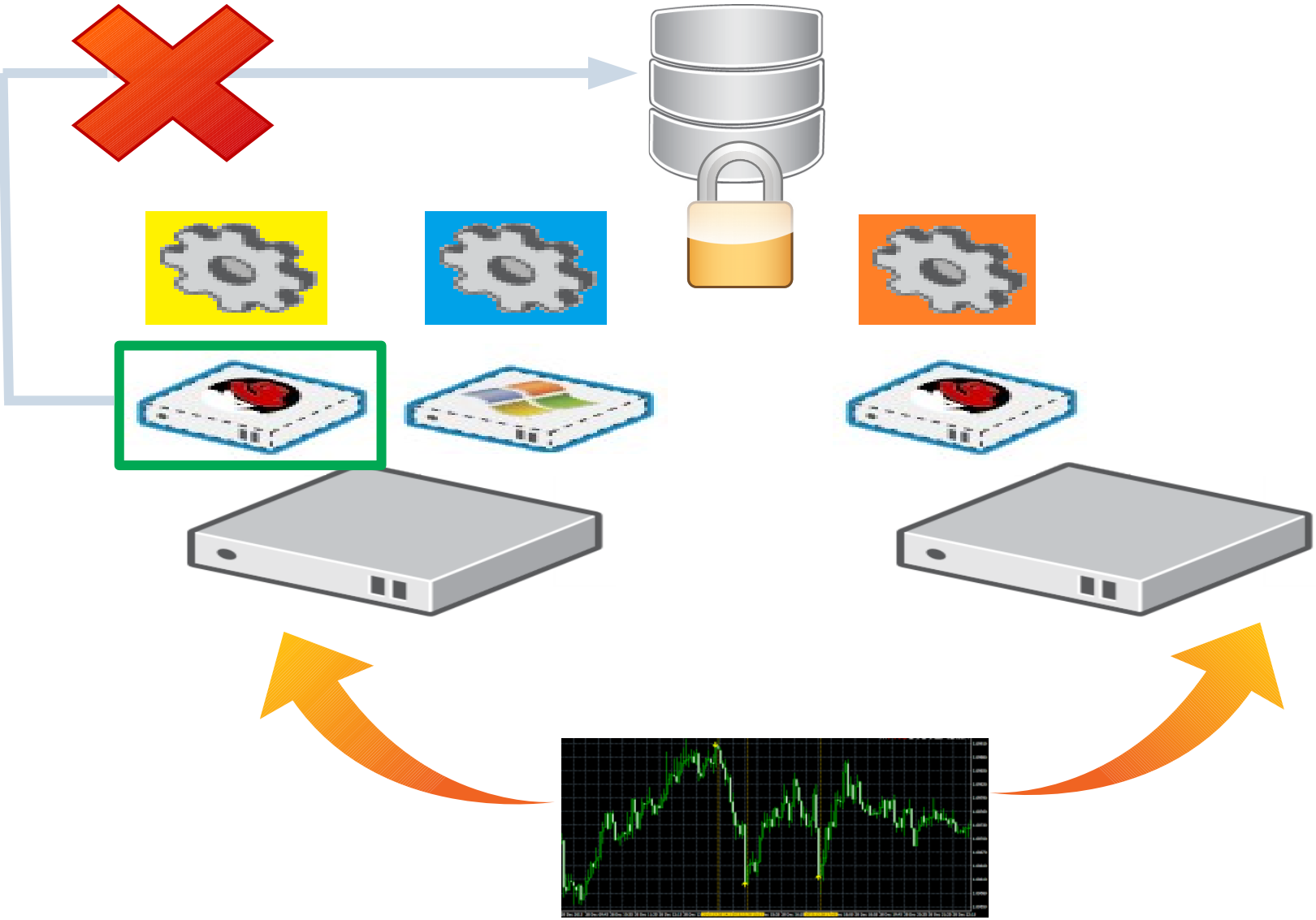


(2) Non-Responsive Host + VM is UP

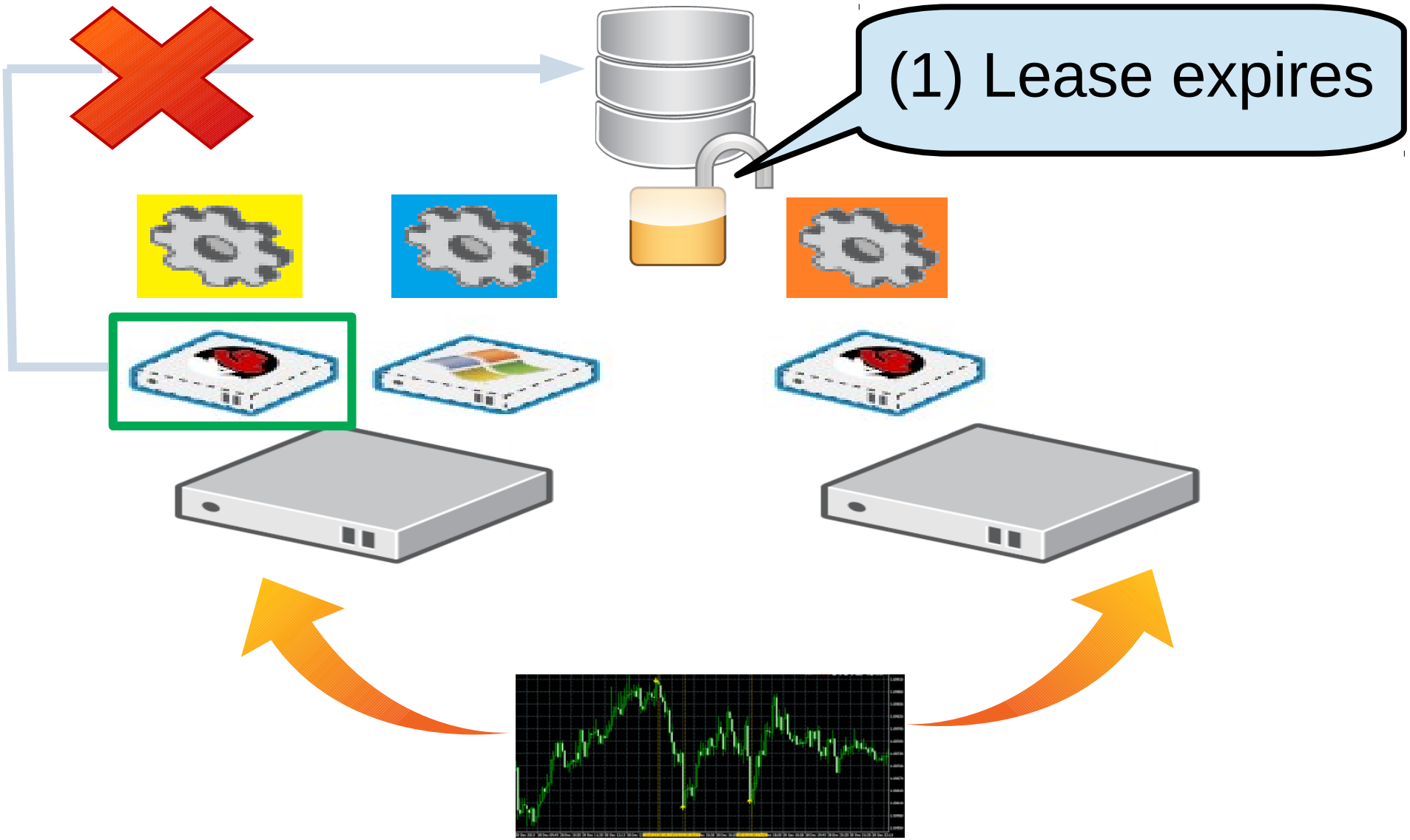


Restart VMs with a Lease

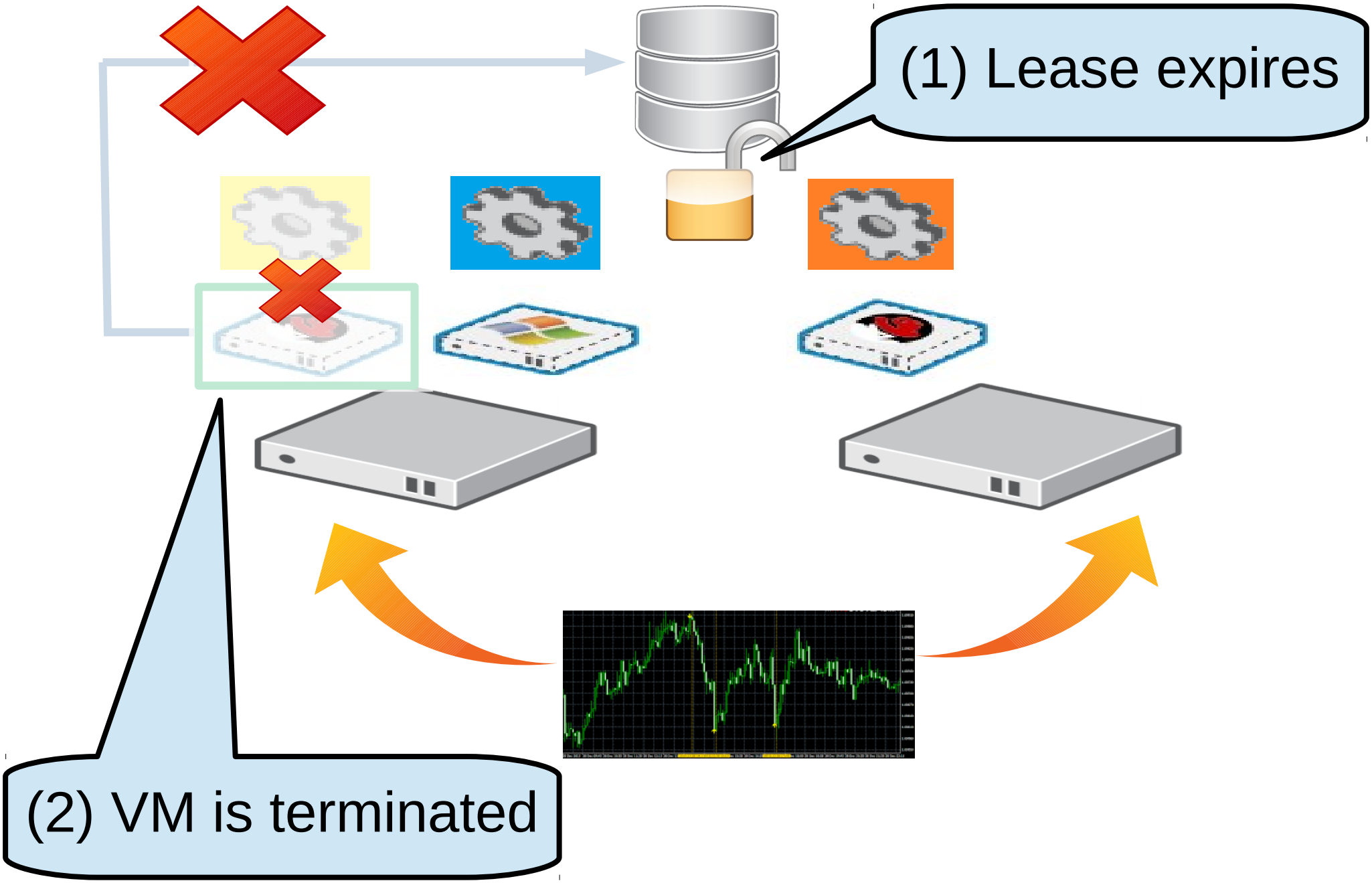
oVirt Disconnection From Storage Device



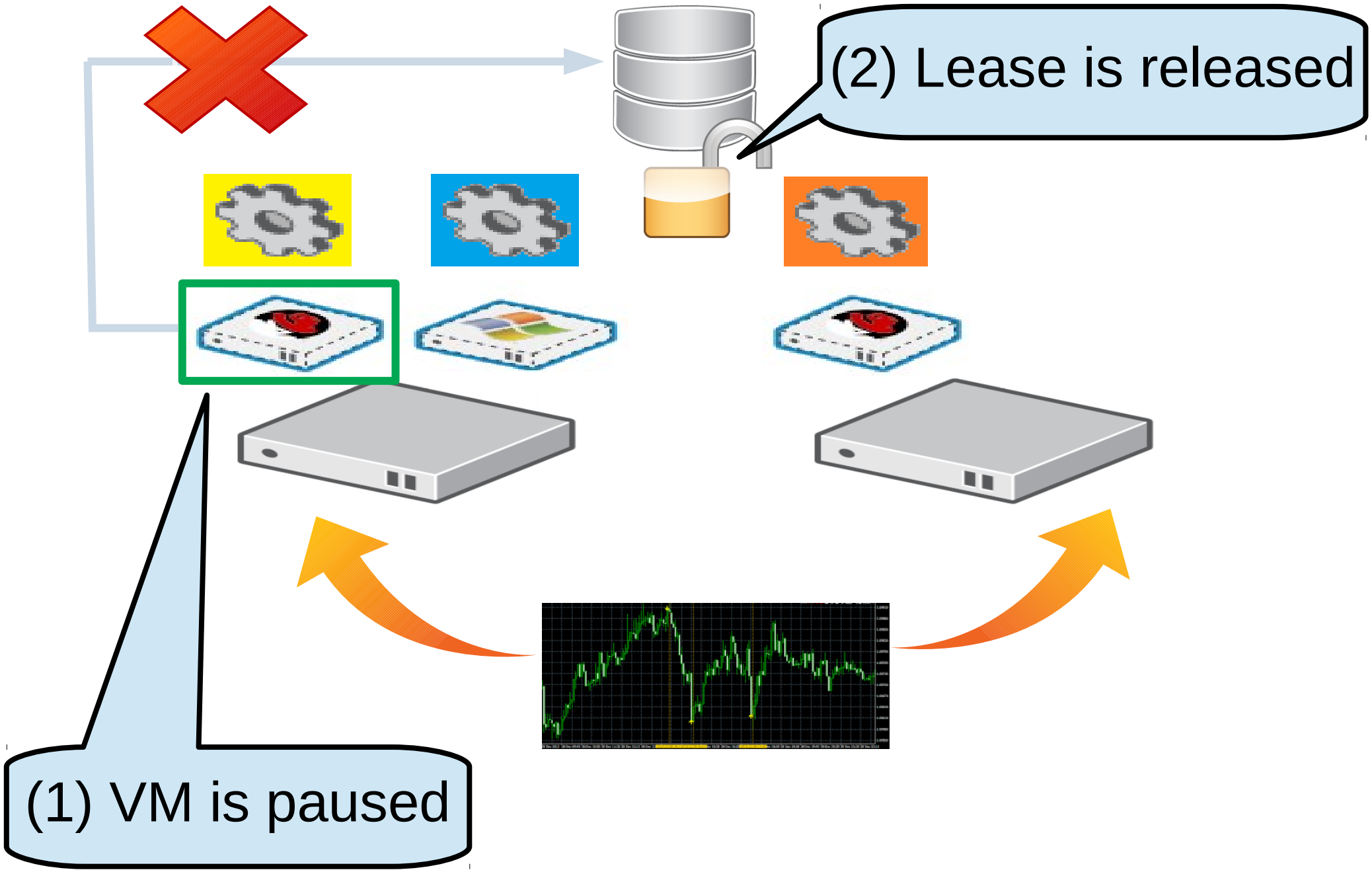
oVirt Disconnection From Storage Device (1)



oVirt Disconnection From Storage Device (1)



oVirt Disconnection From Storage Device (2)



Summary

- VM Lease – an important new element
  - Prevents split-brains
  - Enables automatic restart of unreported VMs
- Available since oVirt 4.1
  - Polished in oVirt 4.2
- Possible future enahncements:
  - May be used to restart paused VMs
  - Move together with the bootable disk

THANK YOU!

http://www.ovirt.org
ahadas@redhat.com
ahadas@irc.oftc.net#ovirt