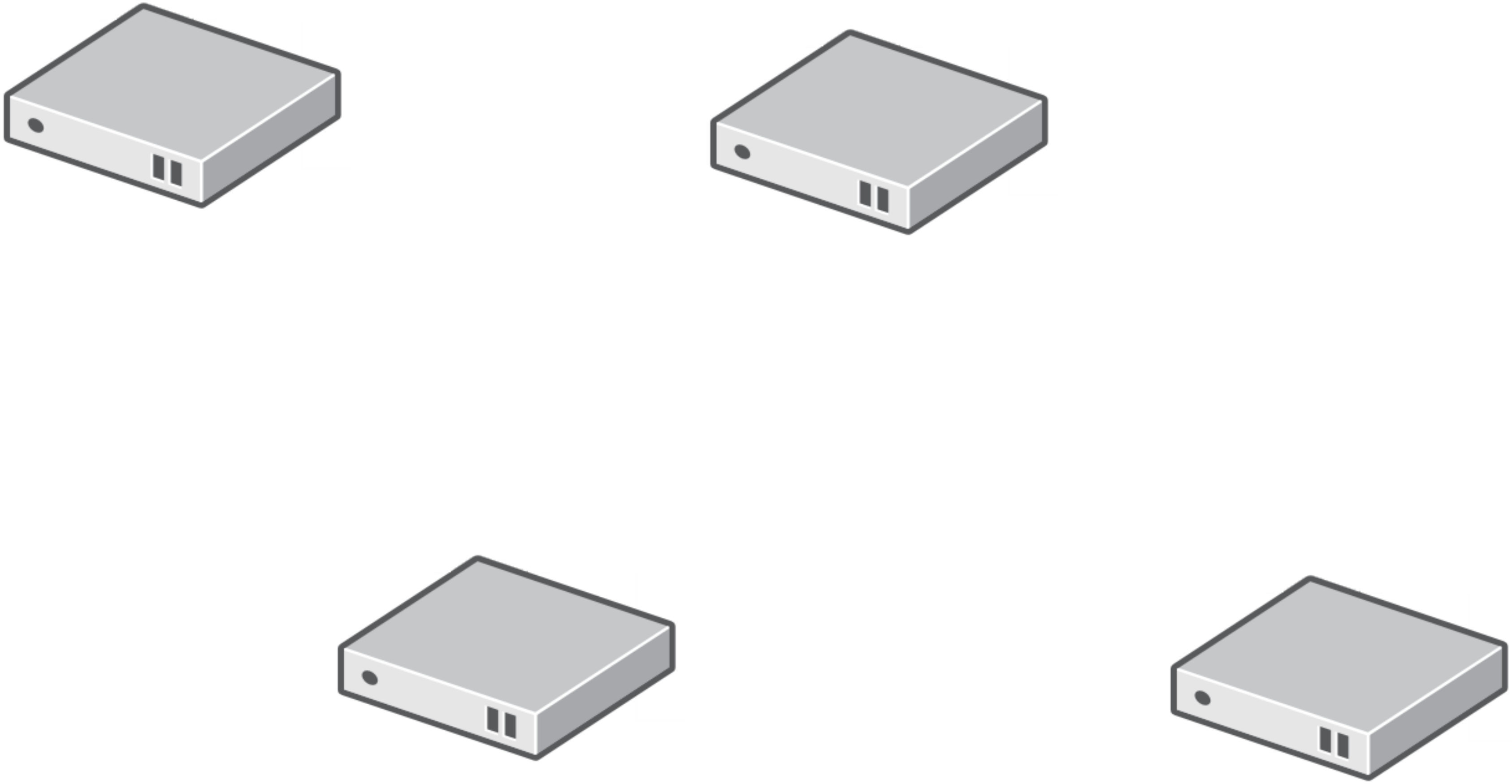
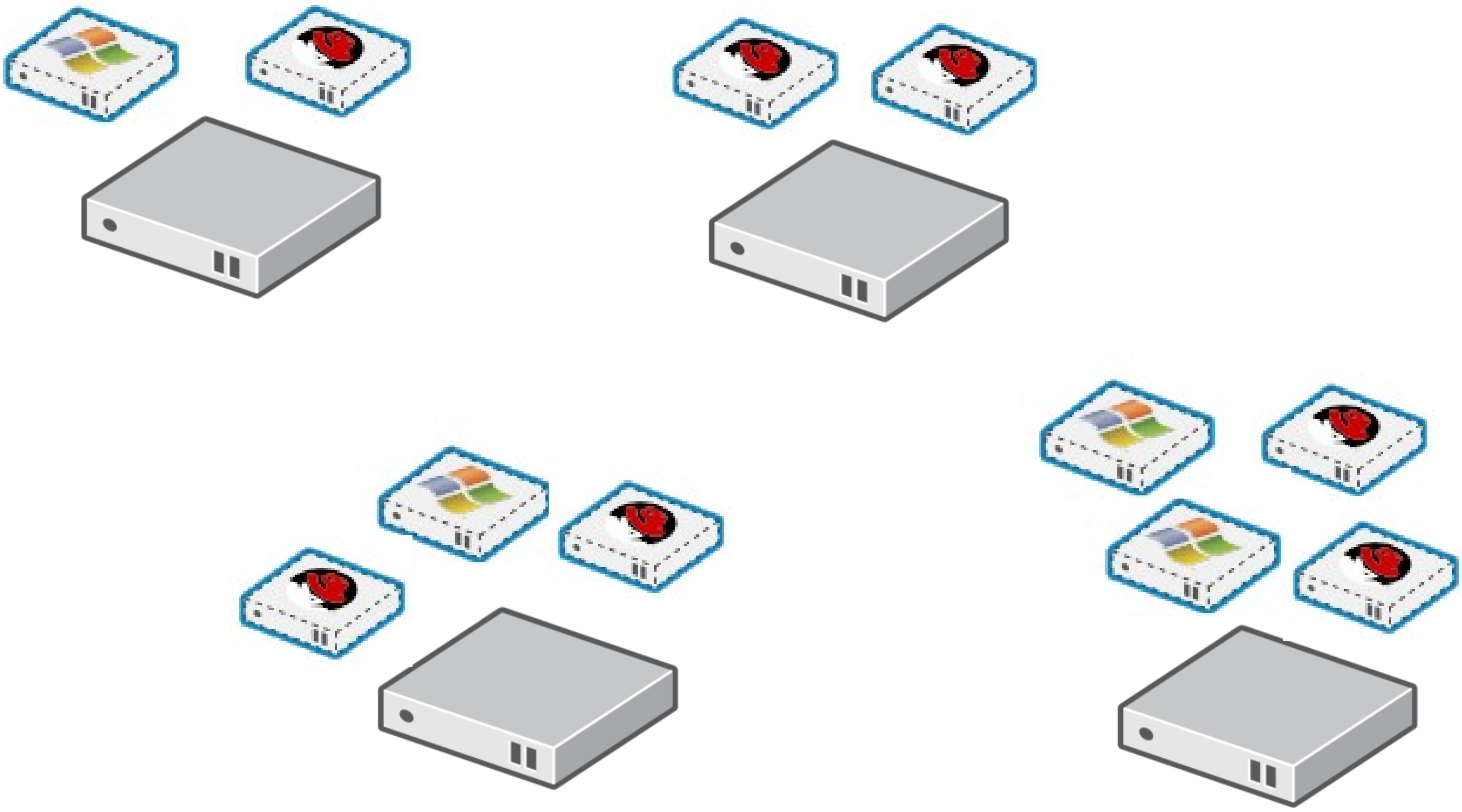
High Availability with No Split Brains!

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

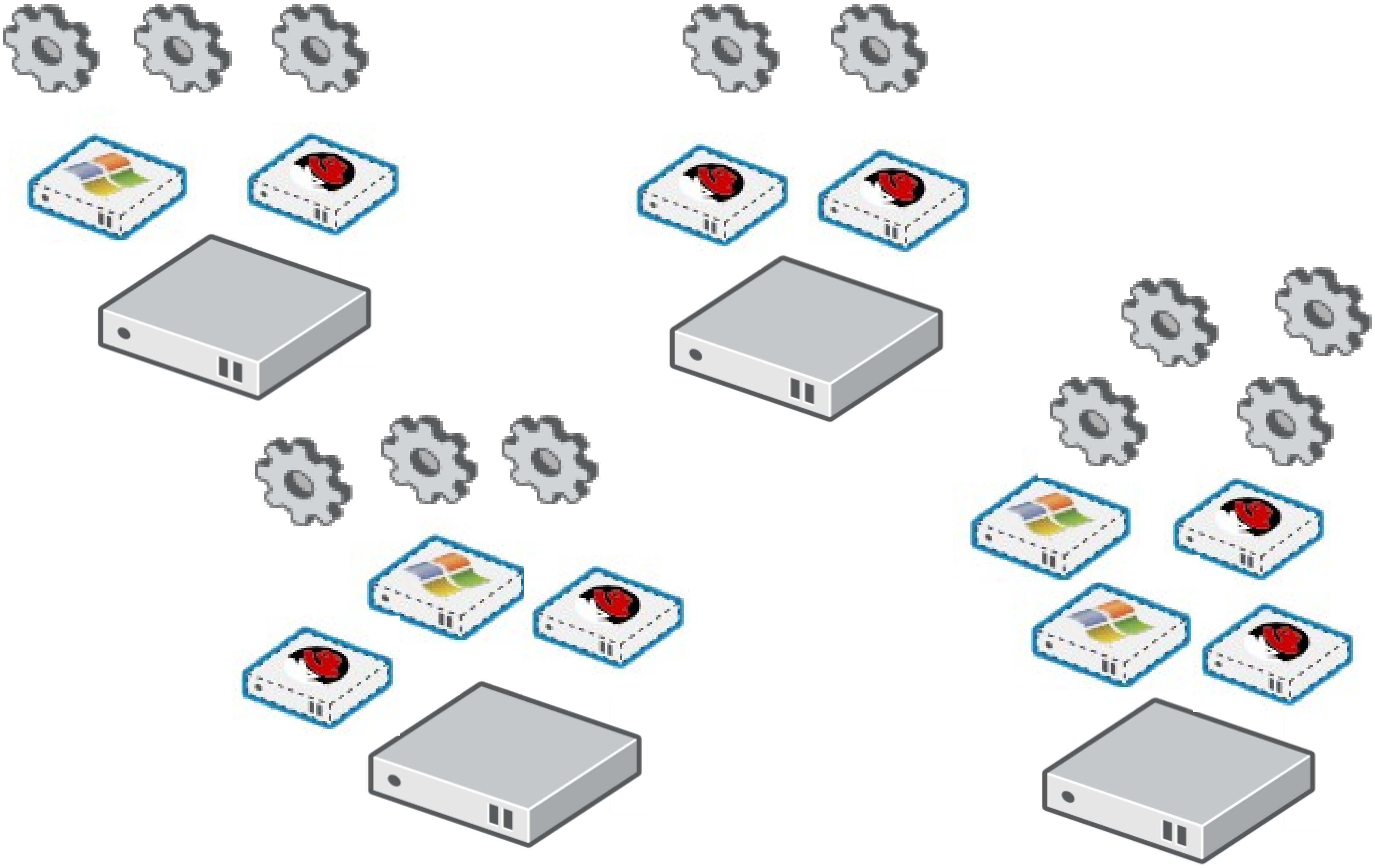
Virtual Data Center – Physical Servers



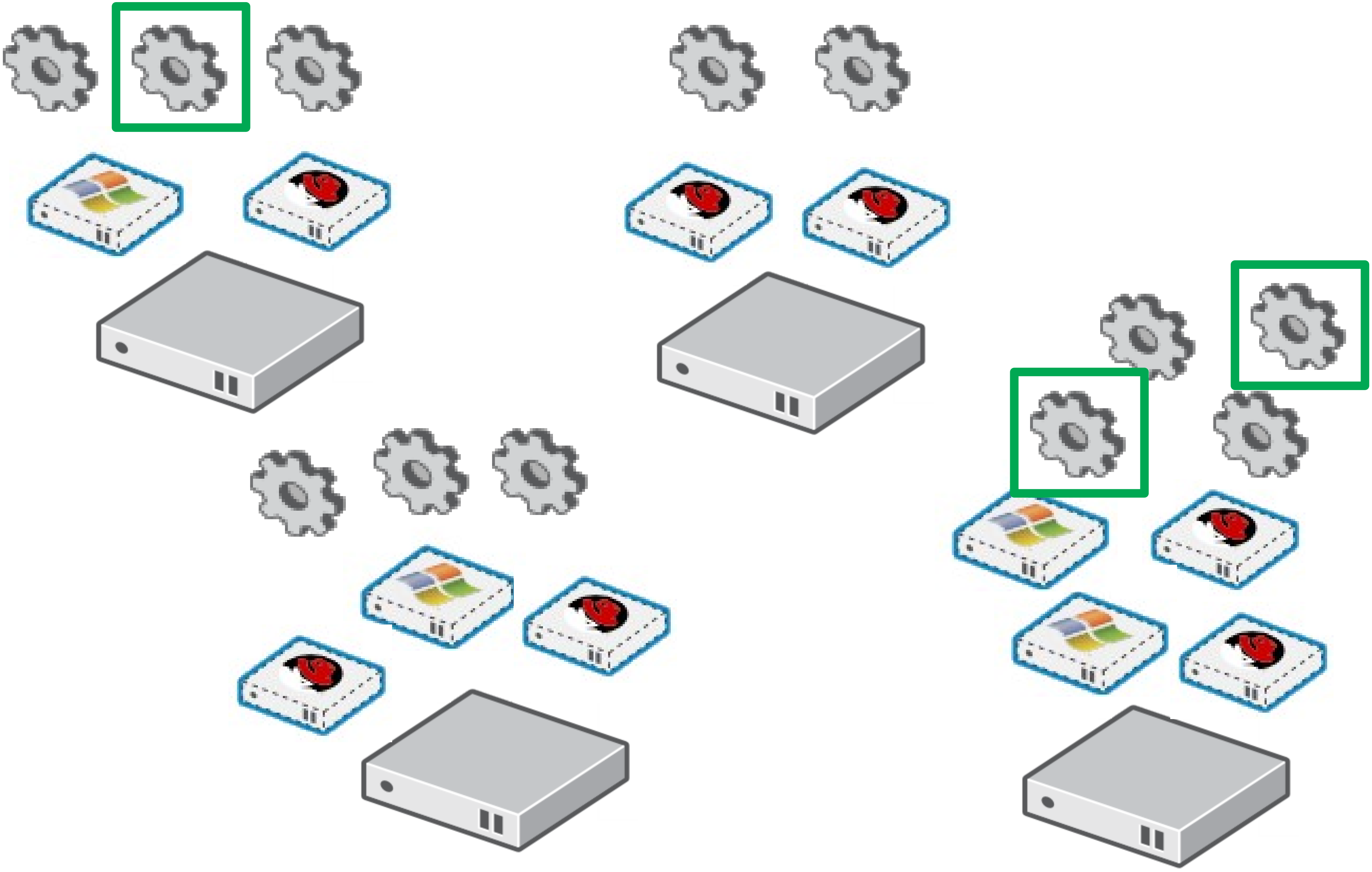
Virtual Data Center – Virtual Machines



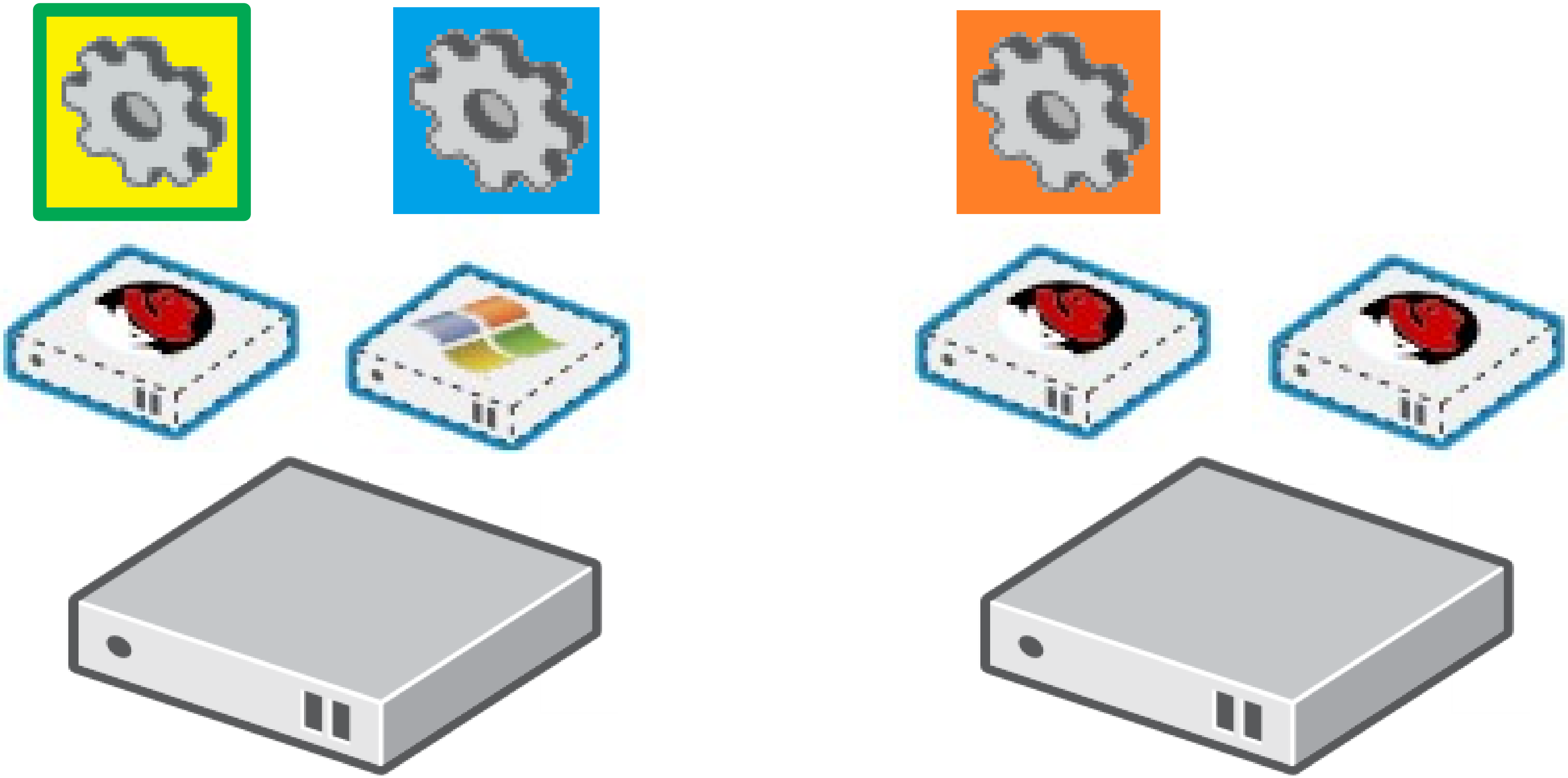
Virtual Data Center - Applications



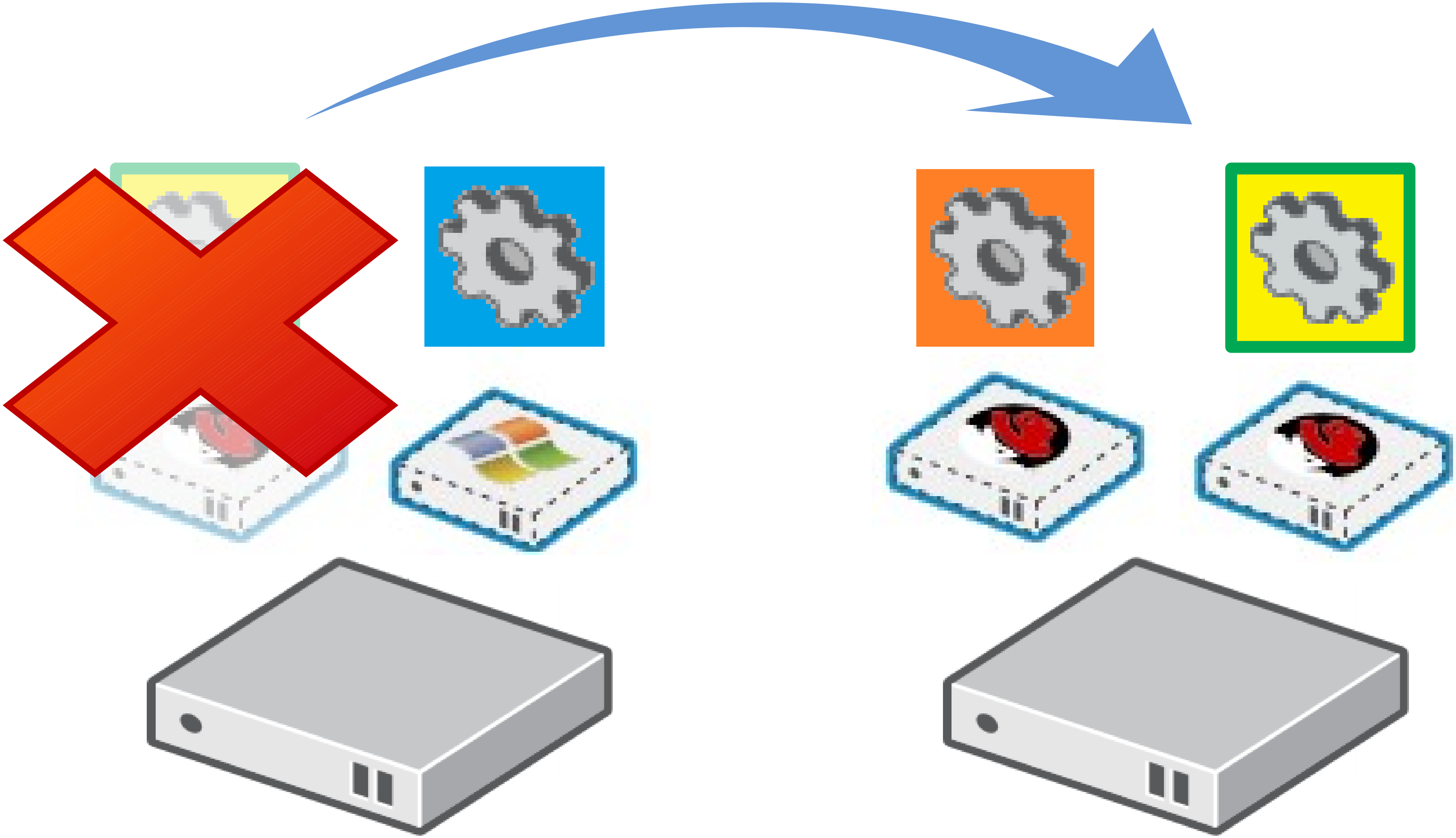
Some Applications are More Critial



High Availability - Application-Level



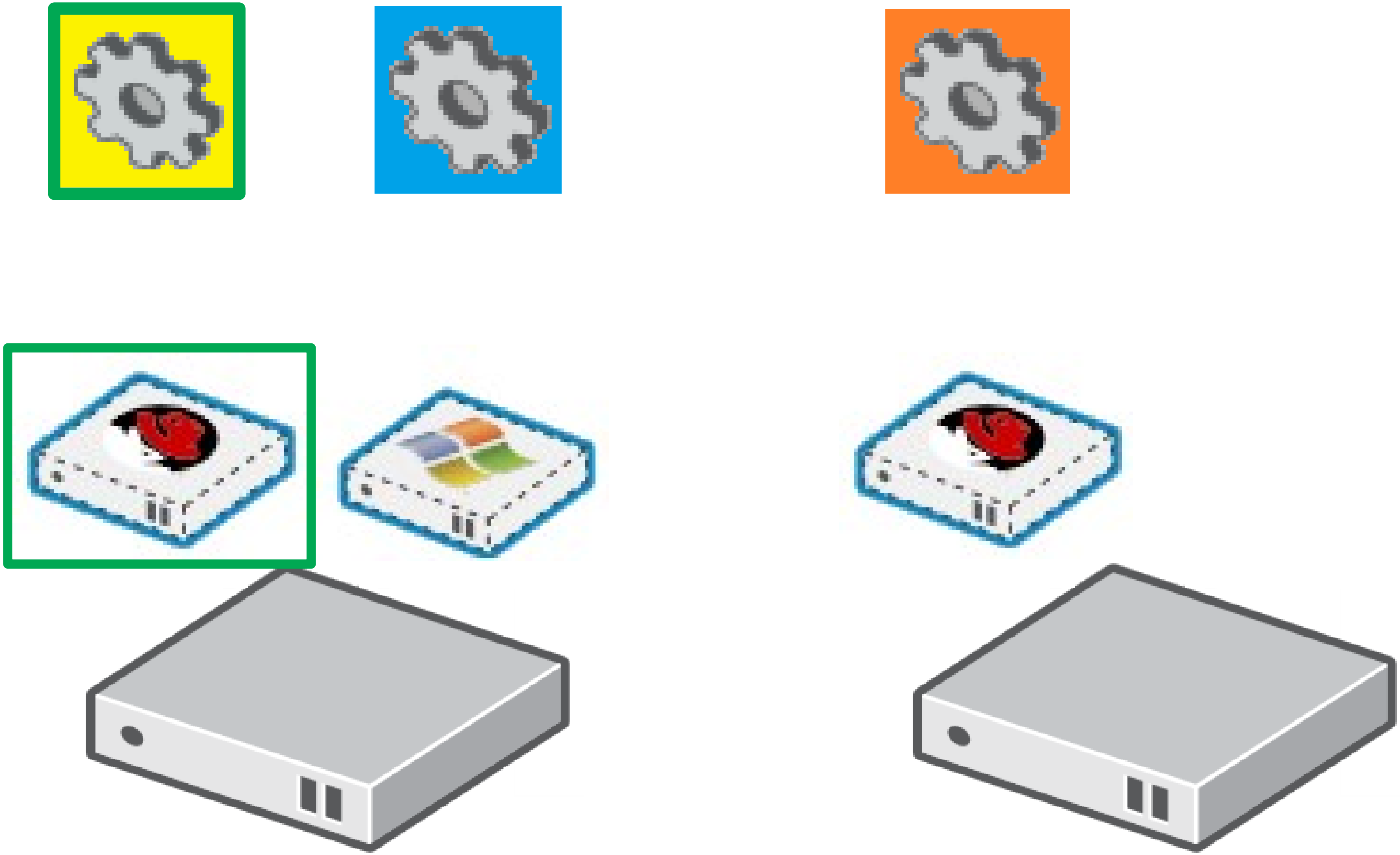
High Availability - Application-Level



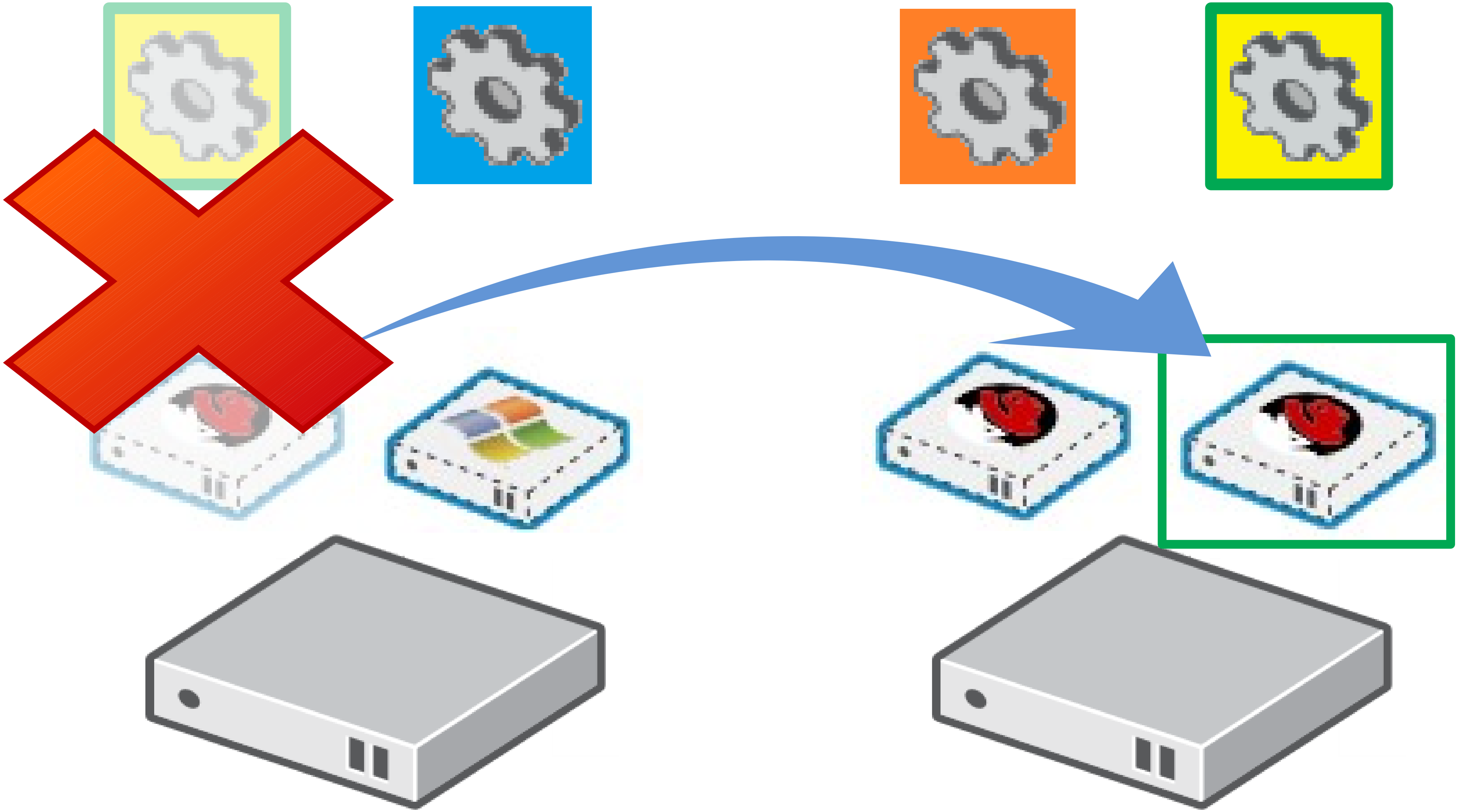
High Availability - Application-Level

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

High Availability - VM-Level



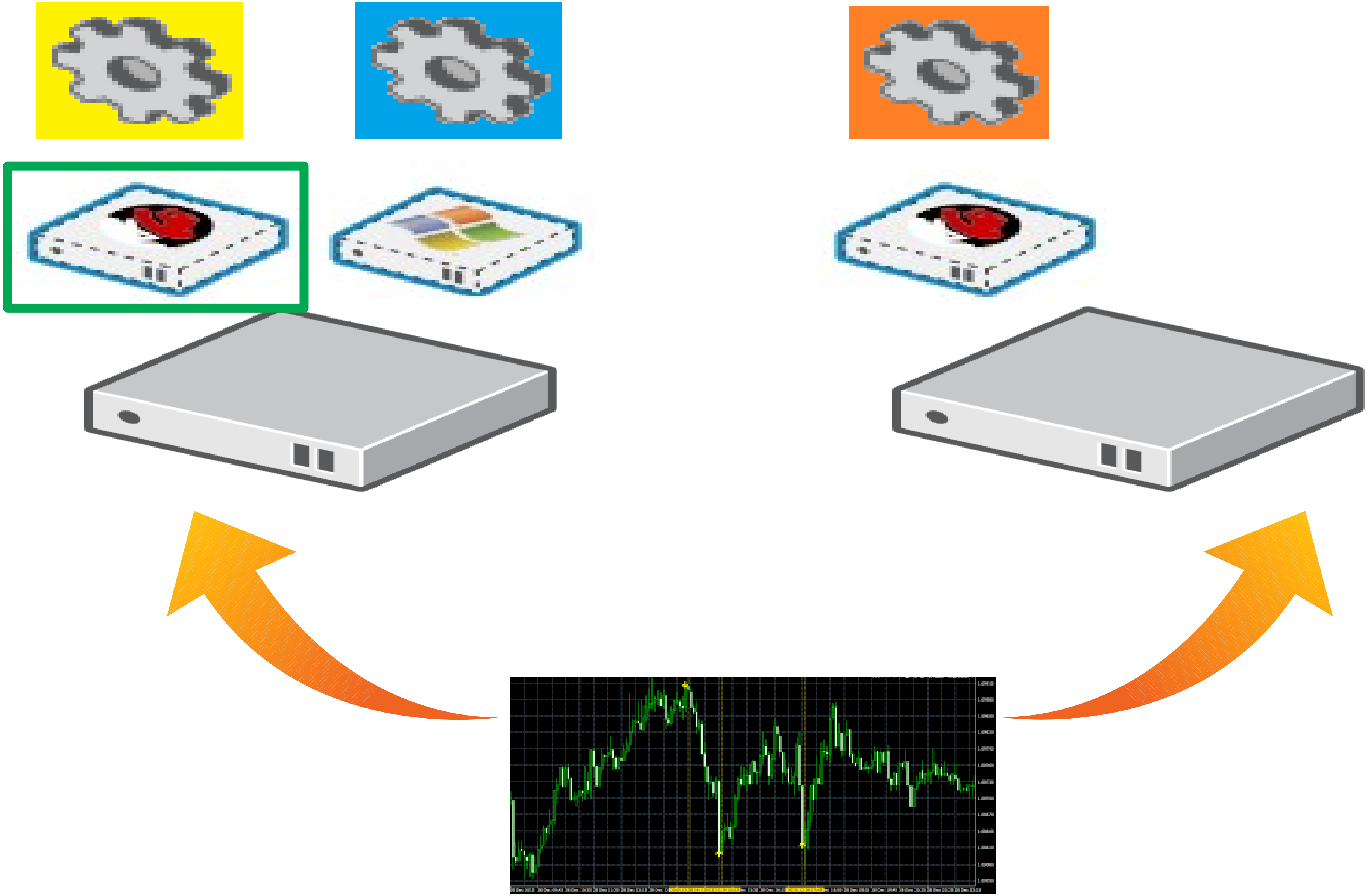
High Availability - VM-Level



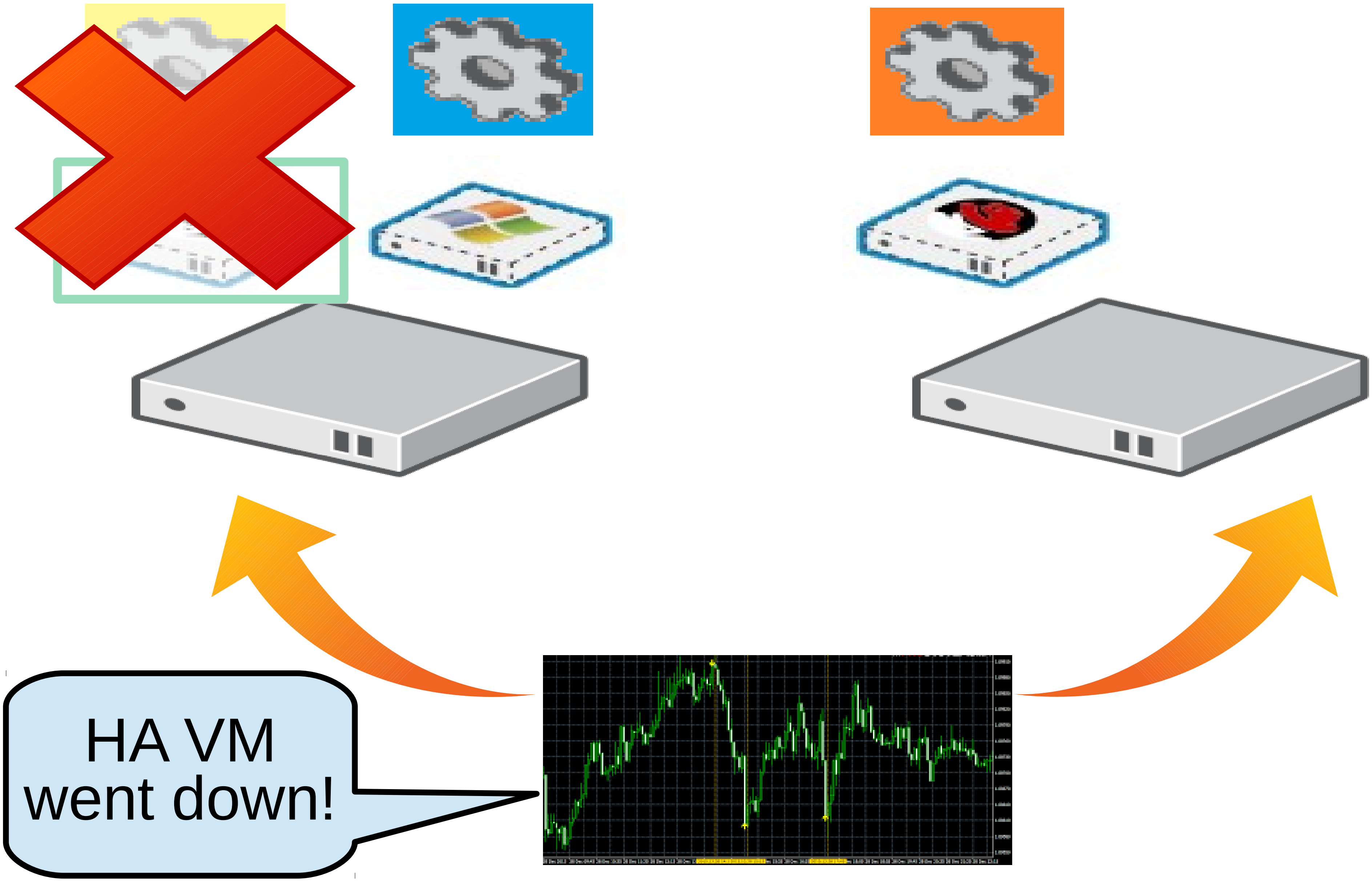
High Availability - VM-Level

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

Central Monitoring Unit

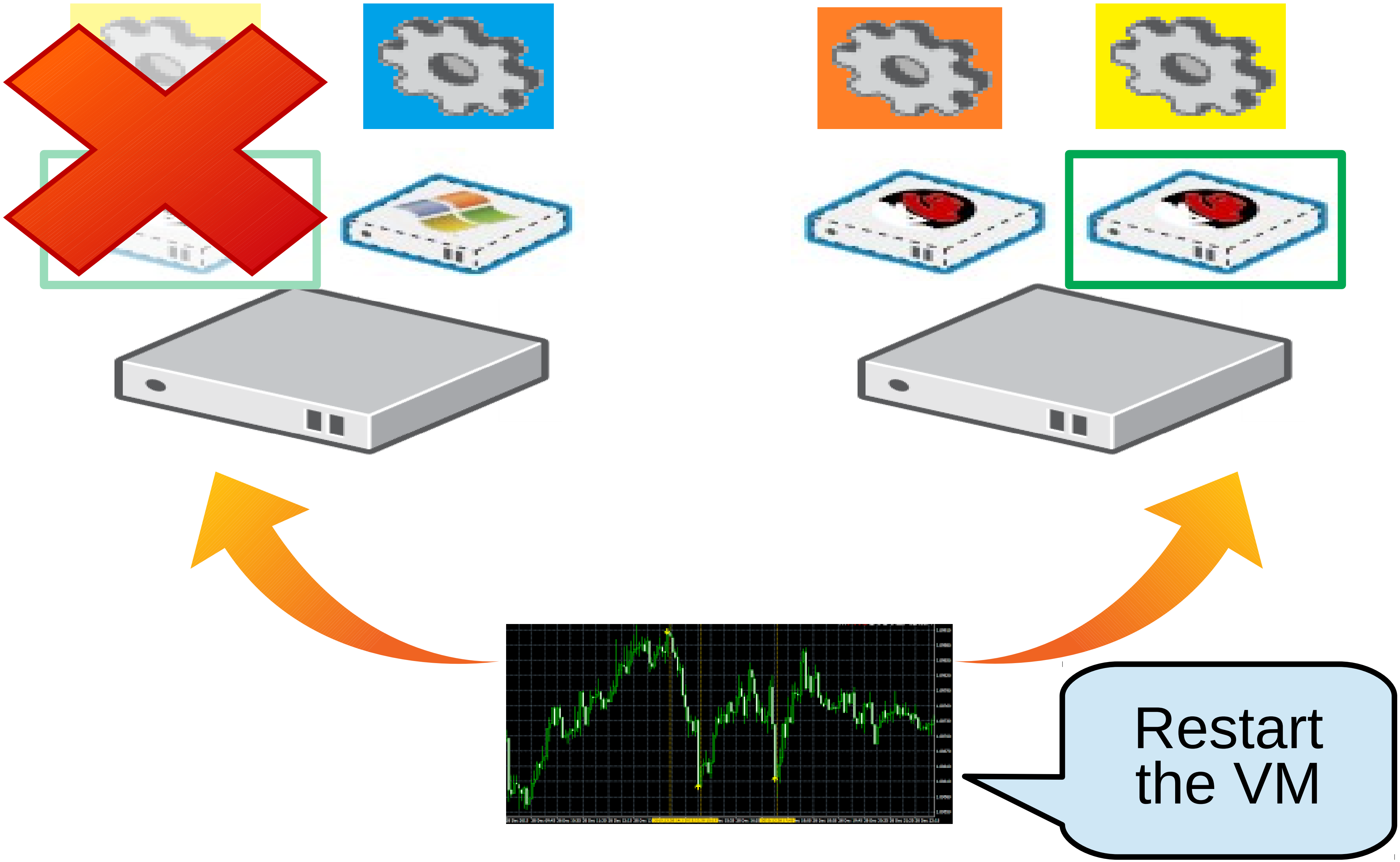


Fault Detection



HA VM went down!

Automatic Restart

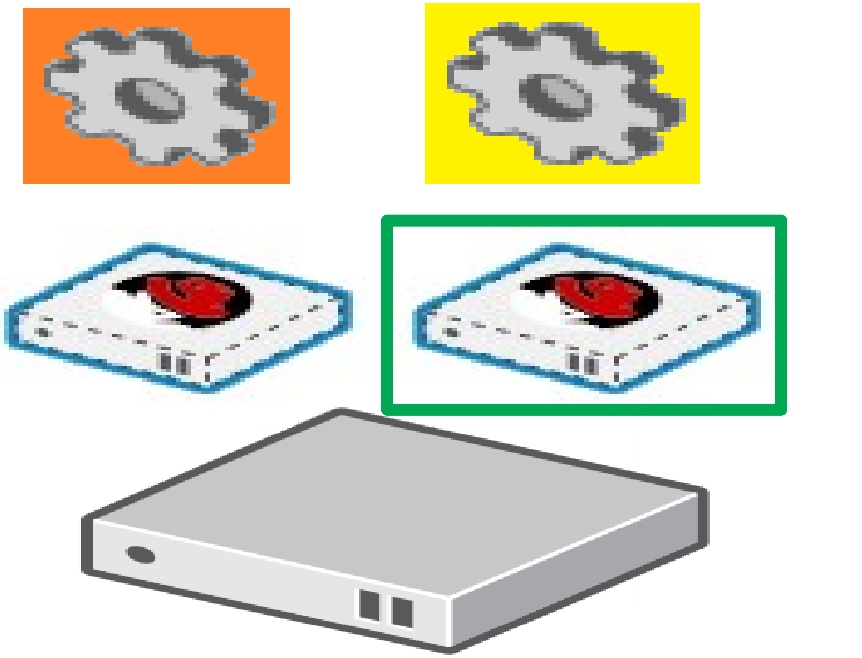


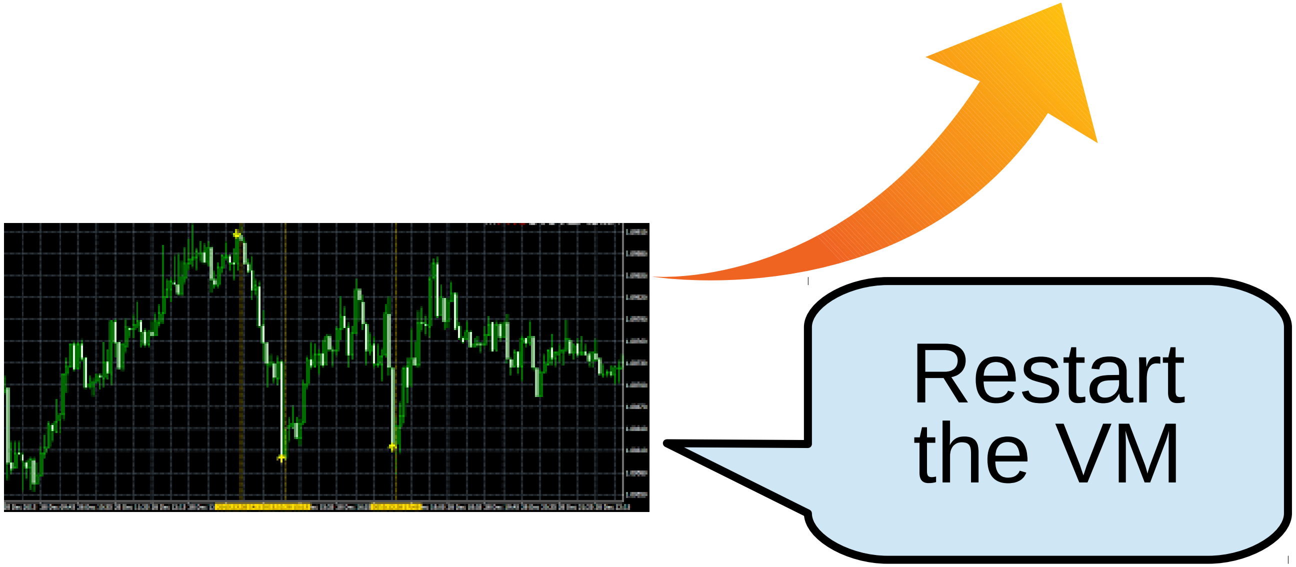
Restart the VM

Automatic Restart – Not That Simple

What if:

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

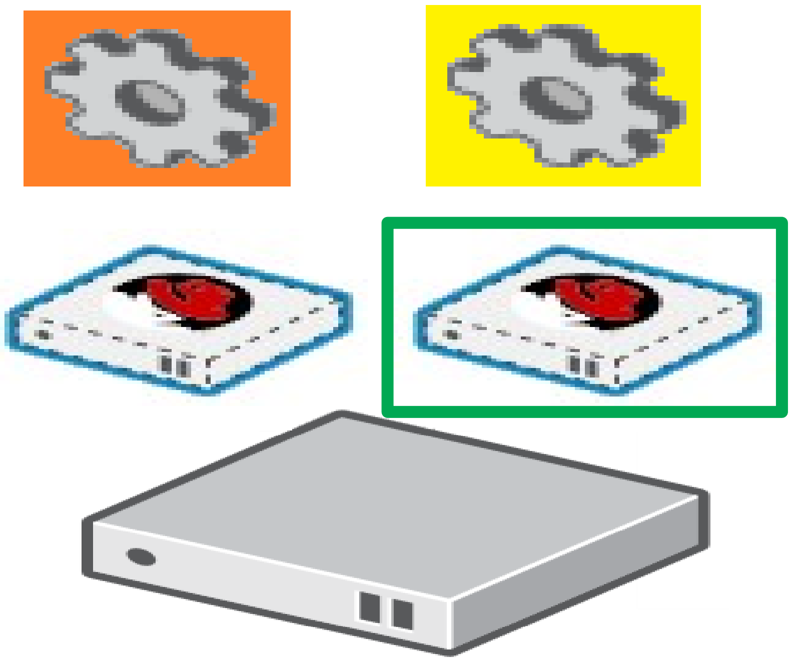



Restart the VM

Automatic Restart – Not That Simple

What if:

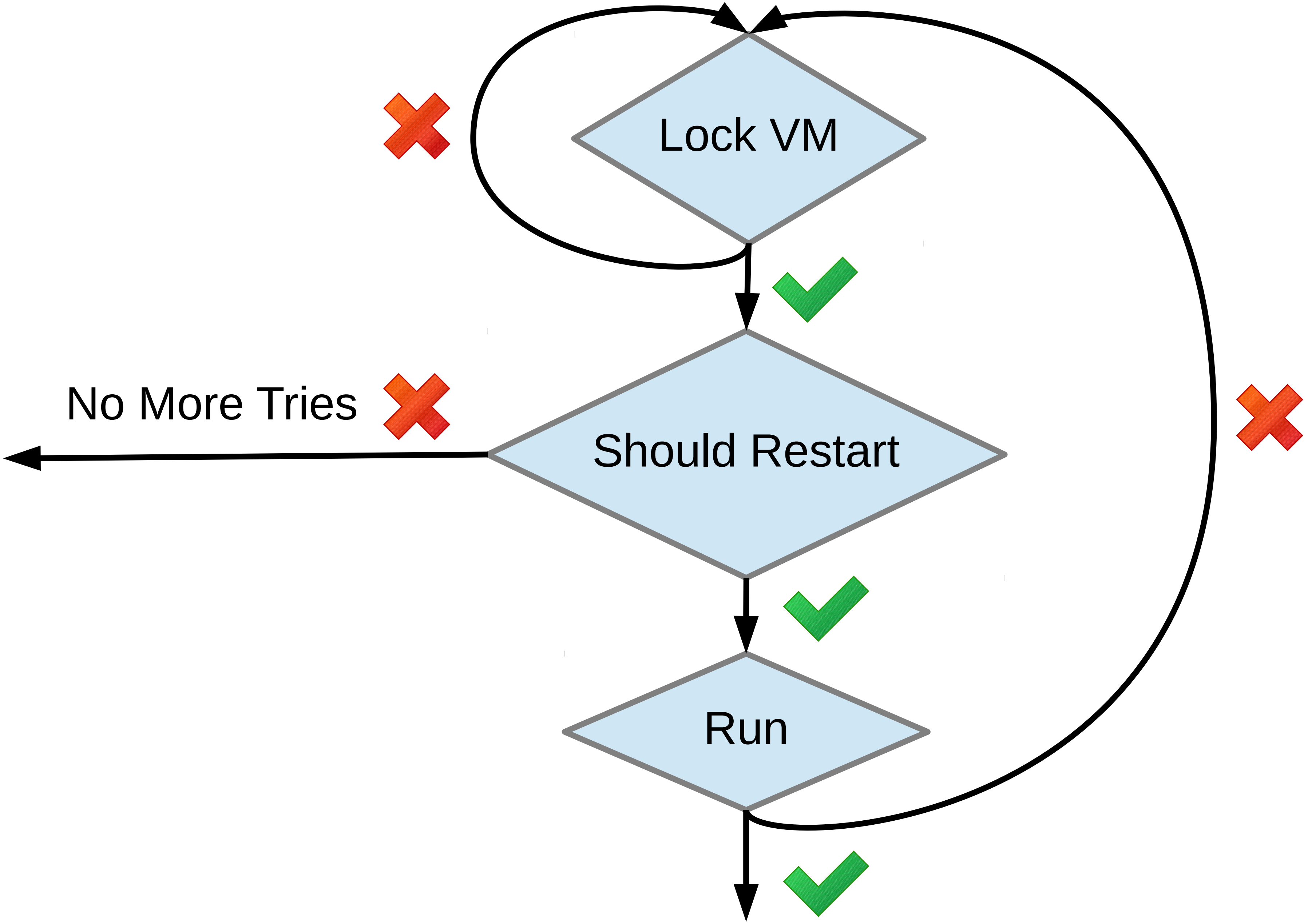
- Inaccessible resources
- VM is locked
- VM is being intentionally shut down



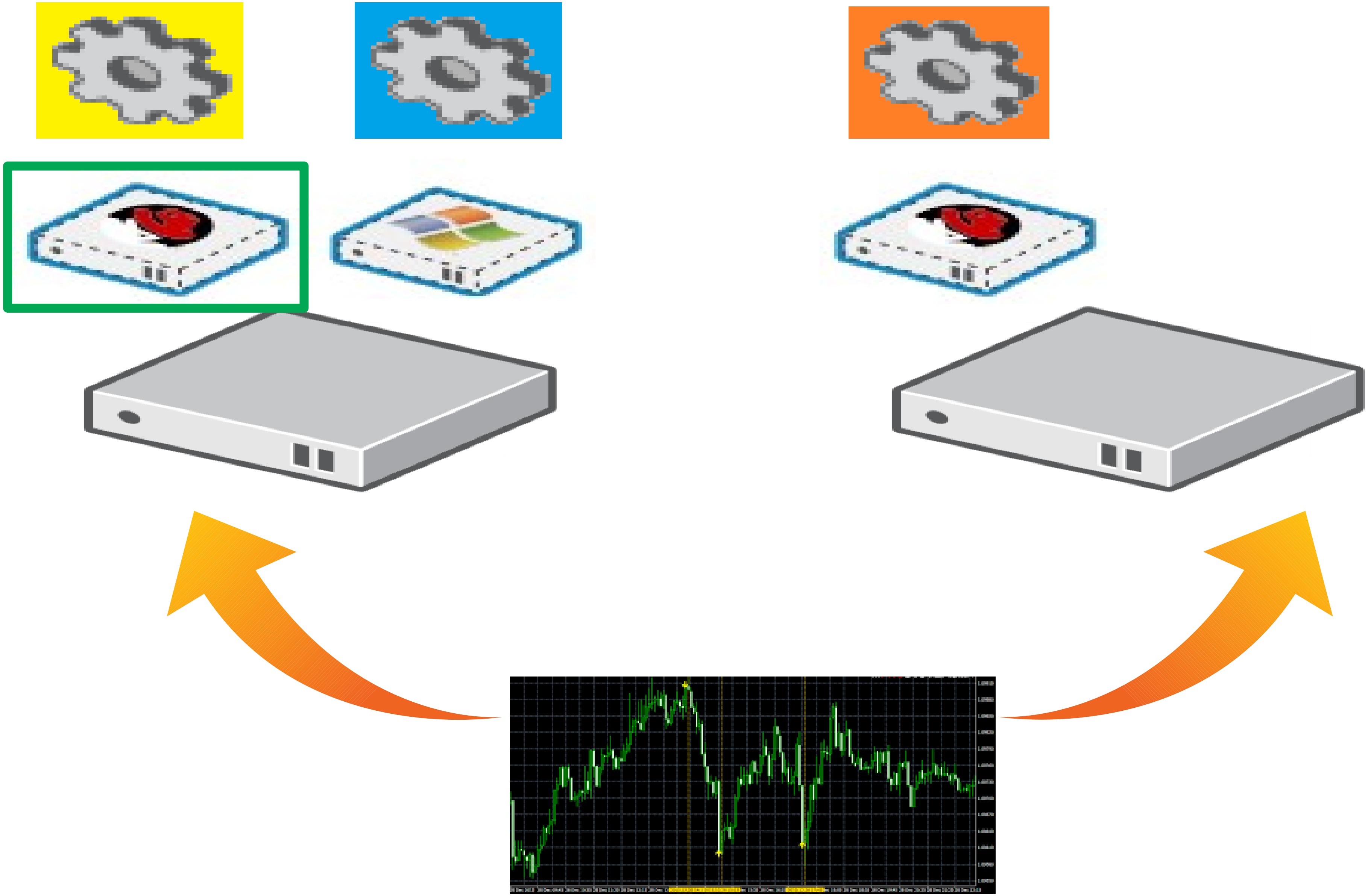
AutoStartVmsRunner

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

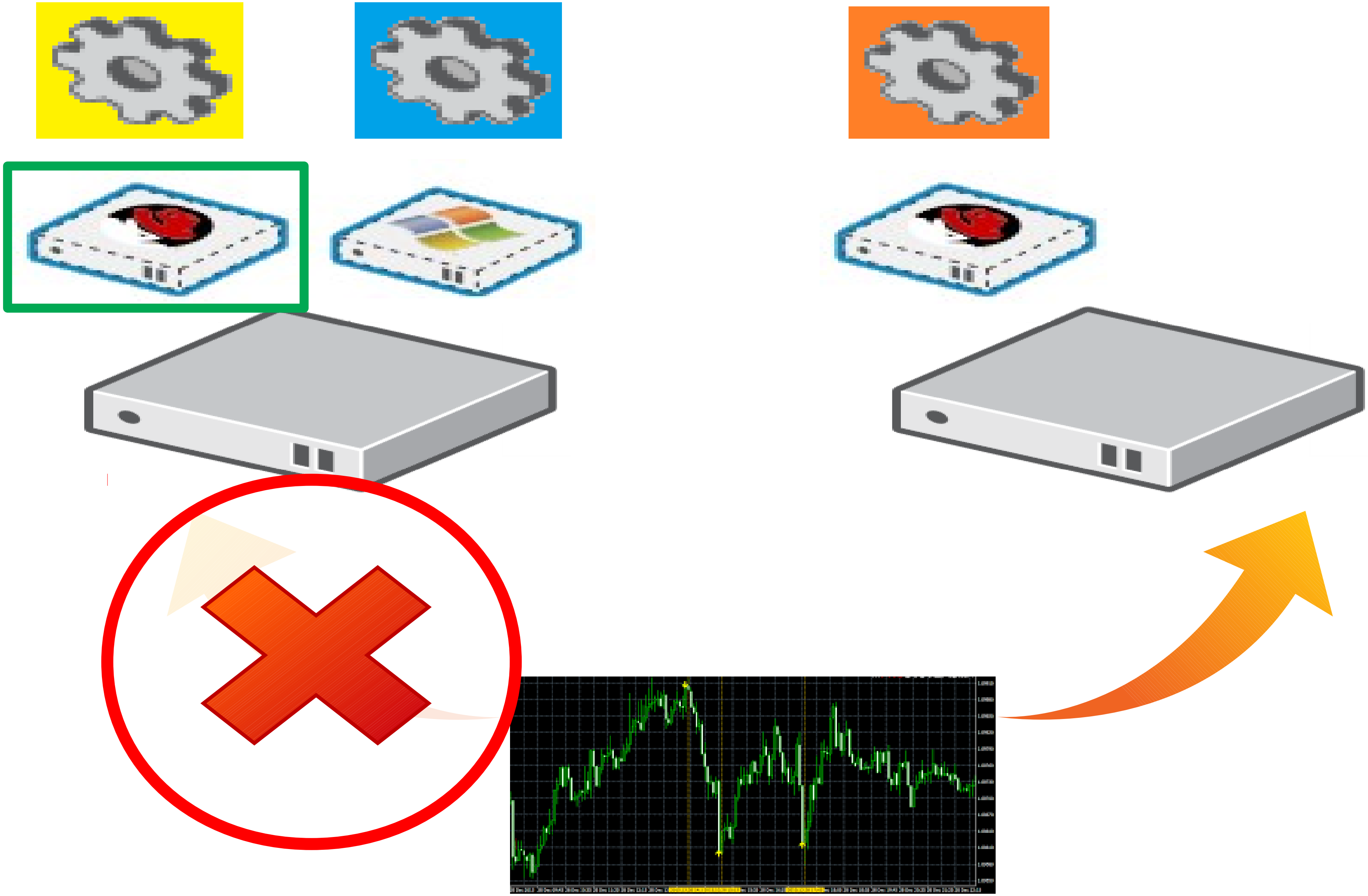
AutoStartVmsRunner



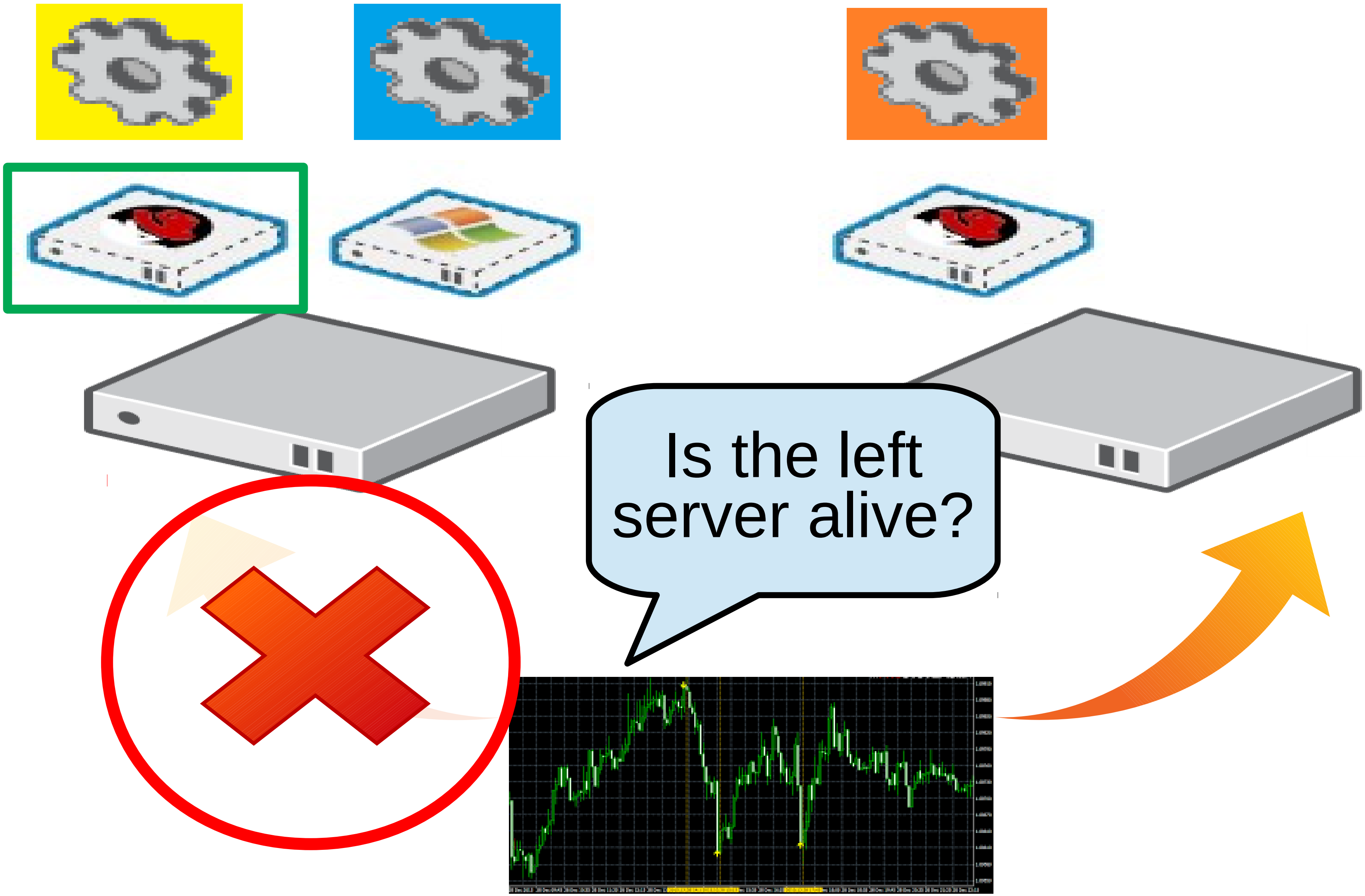
Fault Detection – Even More Complex



Fault Detection – Even More Complex

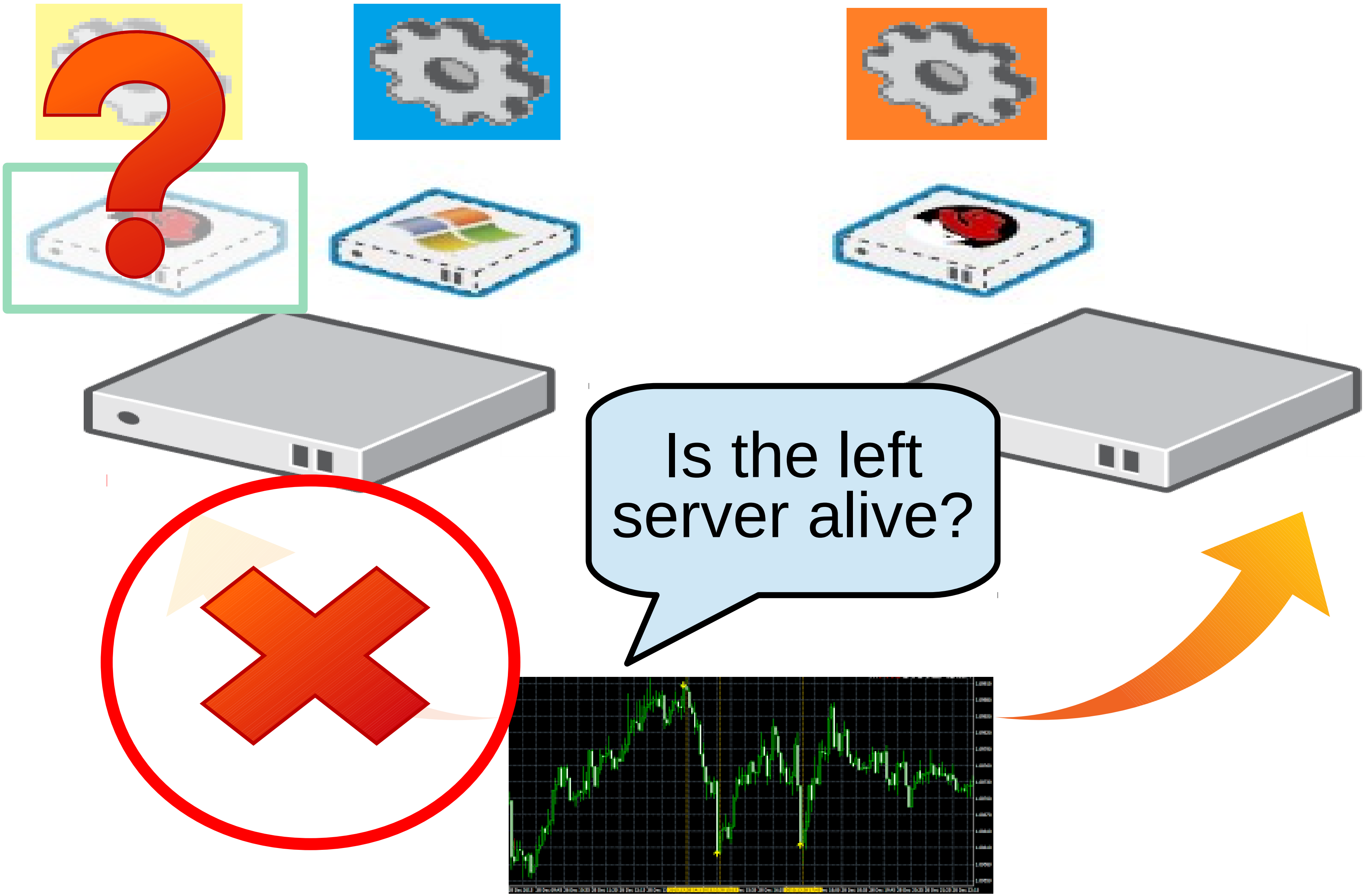


Fault Detection – Even More Complex



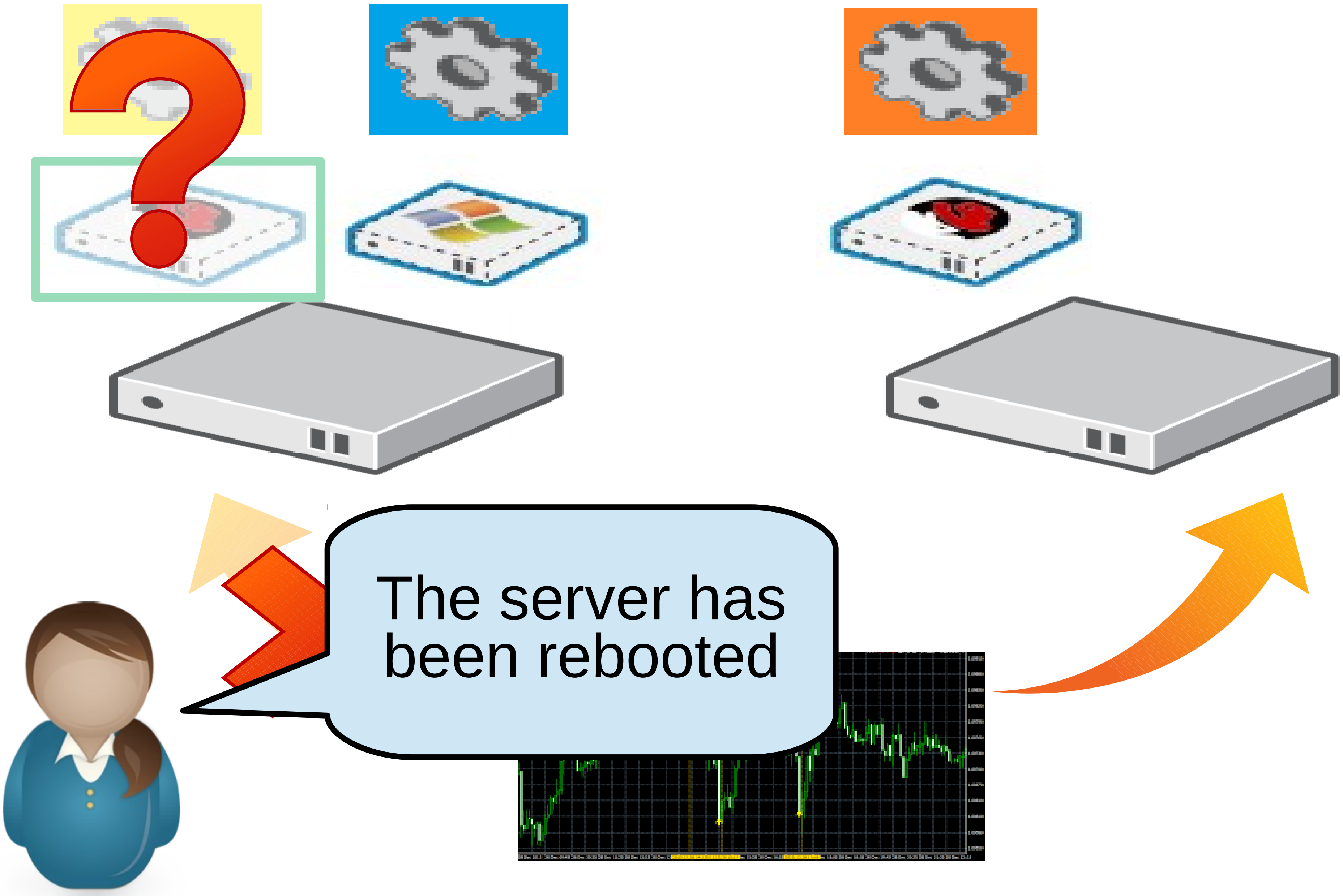
Is the left server alive?

Fault Detection – Even More Complex

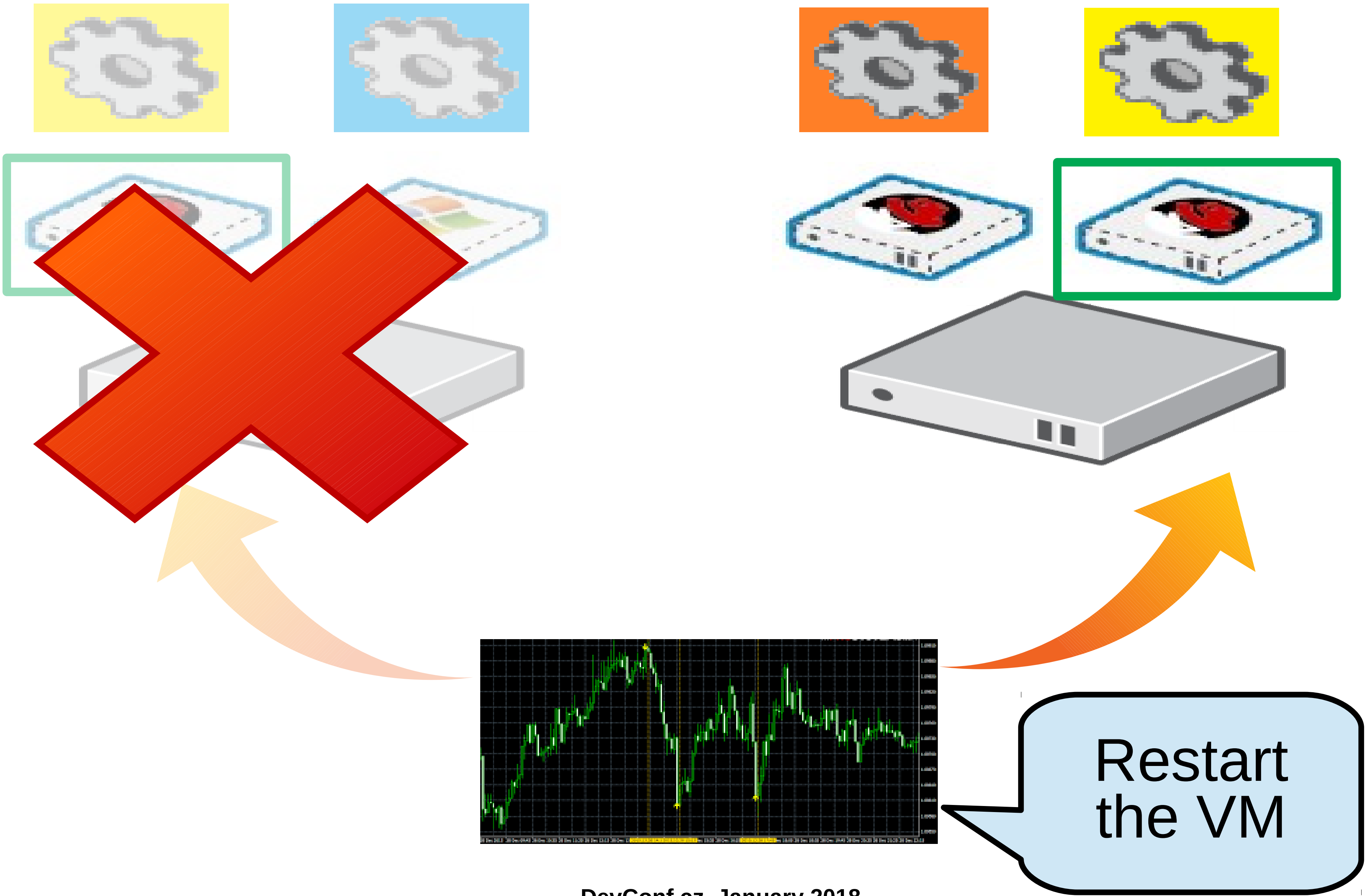


Is the left server alive?

Fault Detection – Manual Confirmation



Fault Detection – Manual Confirmation



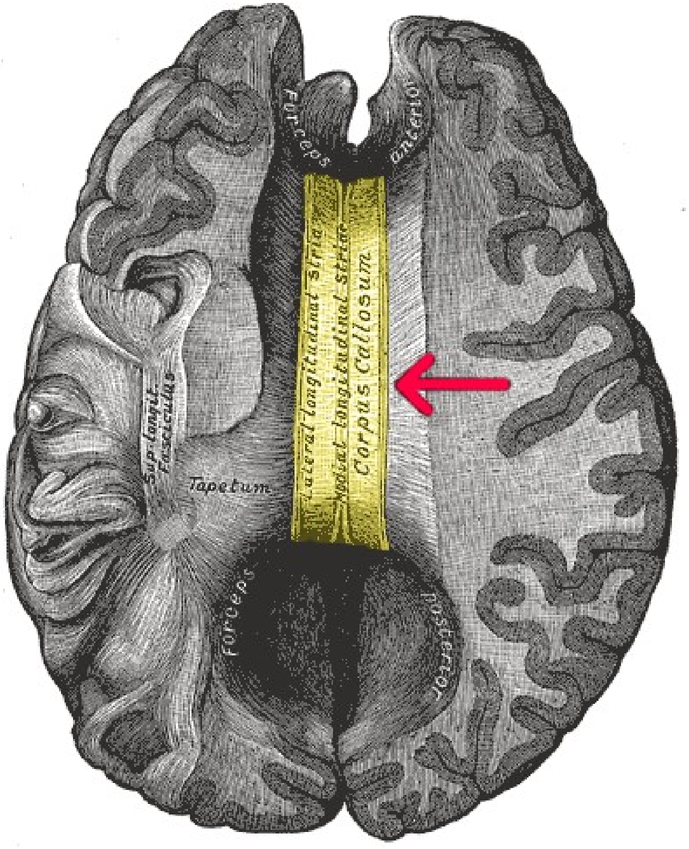
Restart the VM

Fault Detection – Manual Confirmation

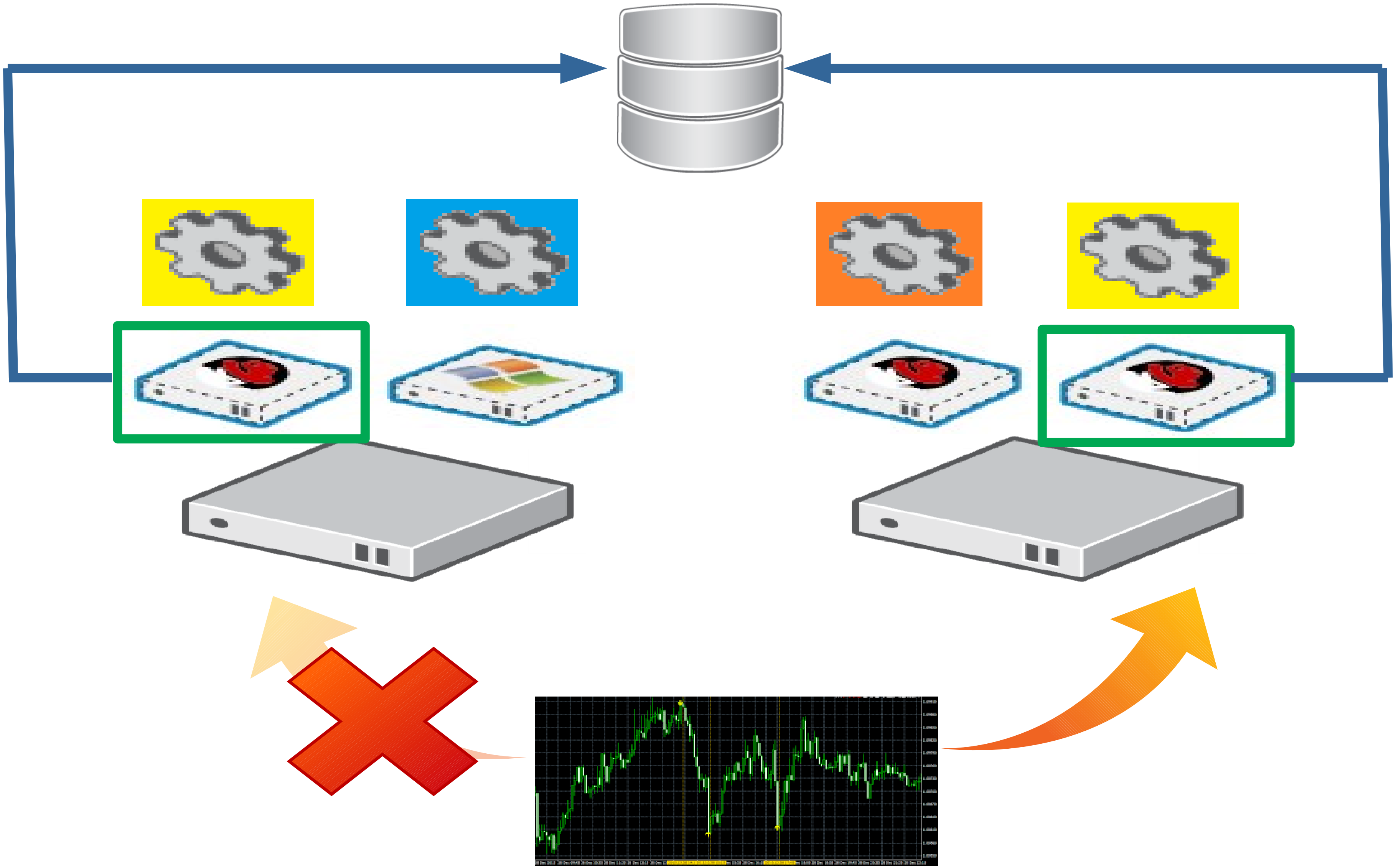
- Slow
- Error-prone
  - Mistakes may lead to a split-brain

Split Brain of Virtual Machines

A scenario in which several VMs that may write to the same disk(s) are running simultaneously

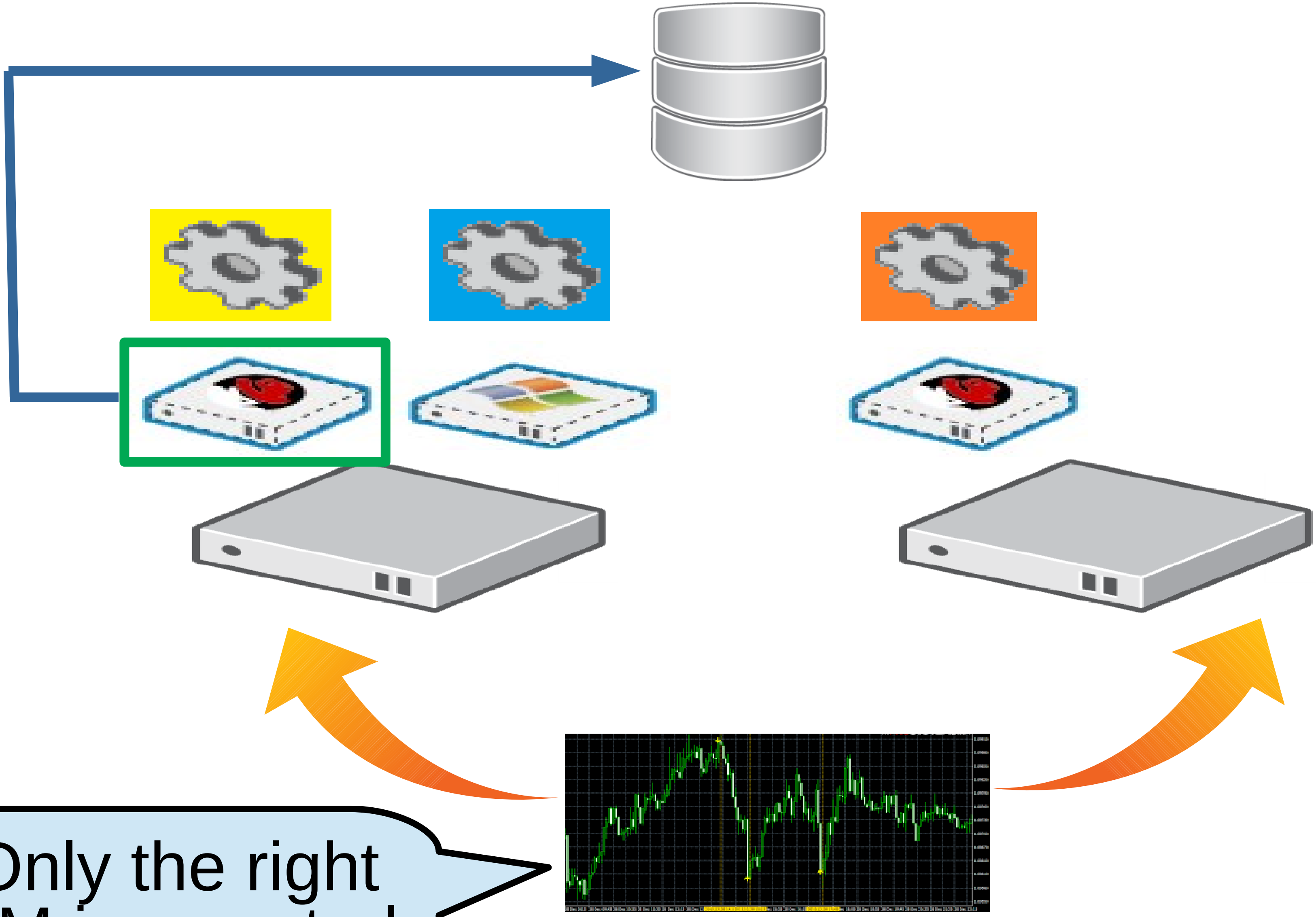


Split Brain - Example



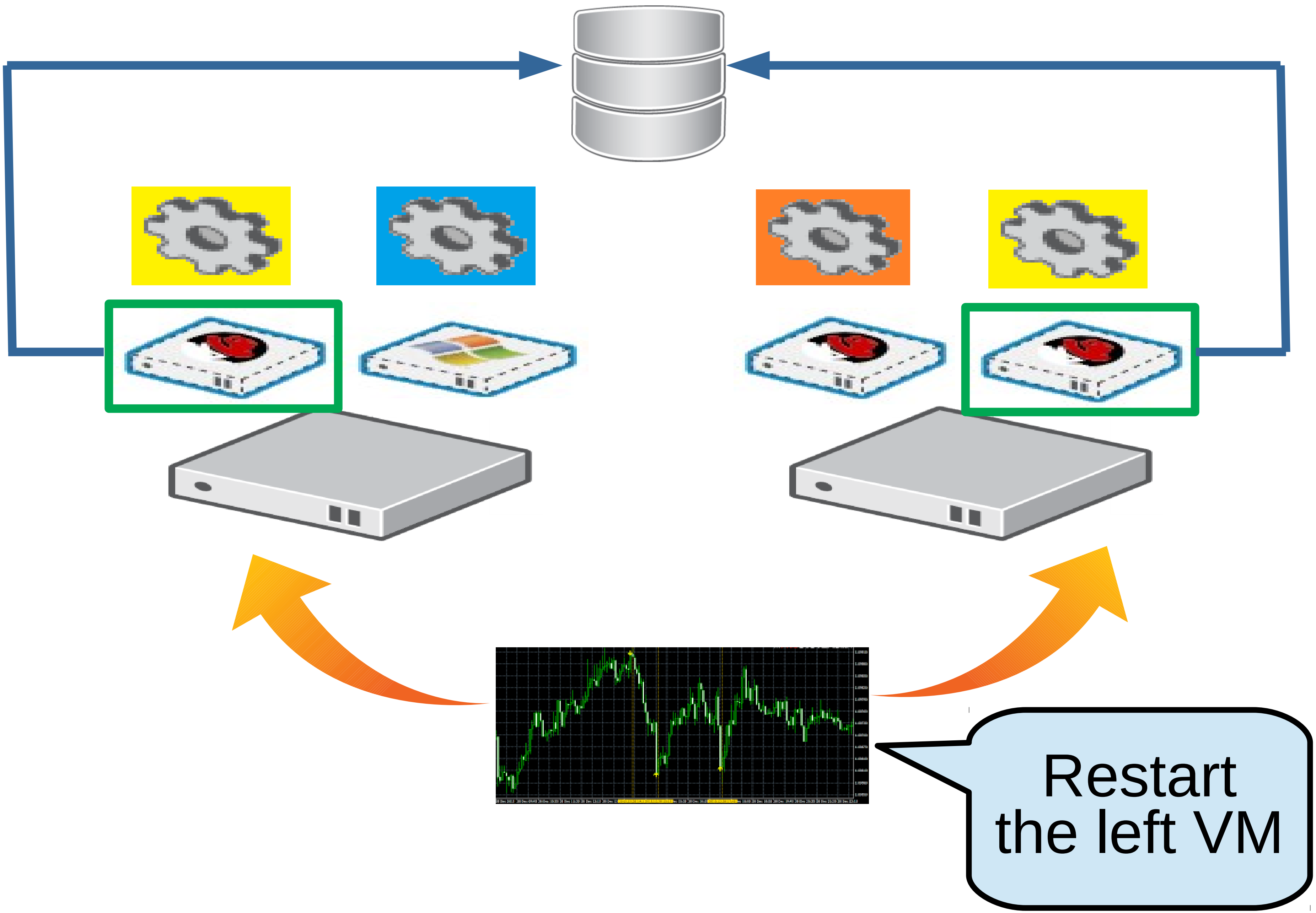
May lead to data corruption!

Split Brains May Happen Due to Bugs



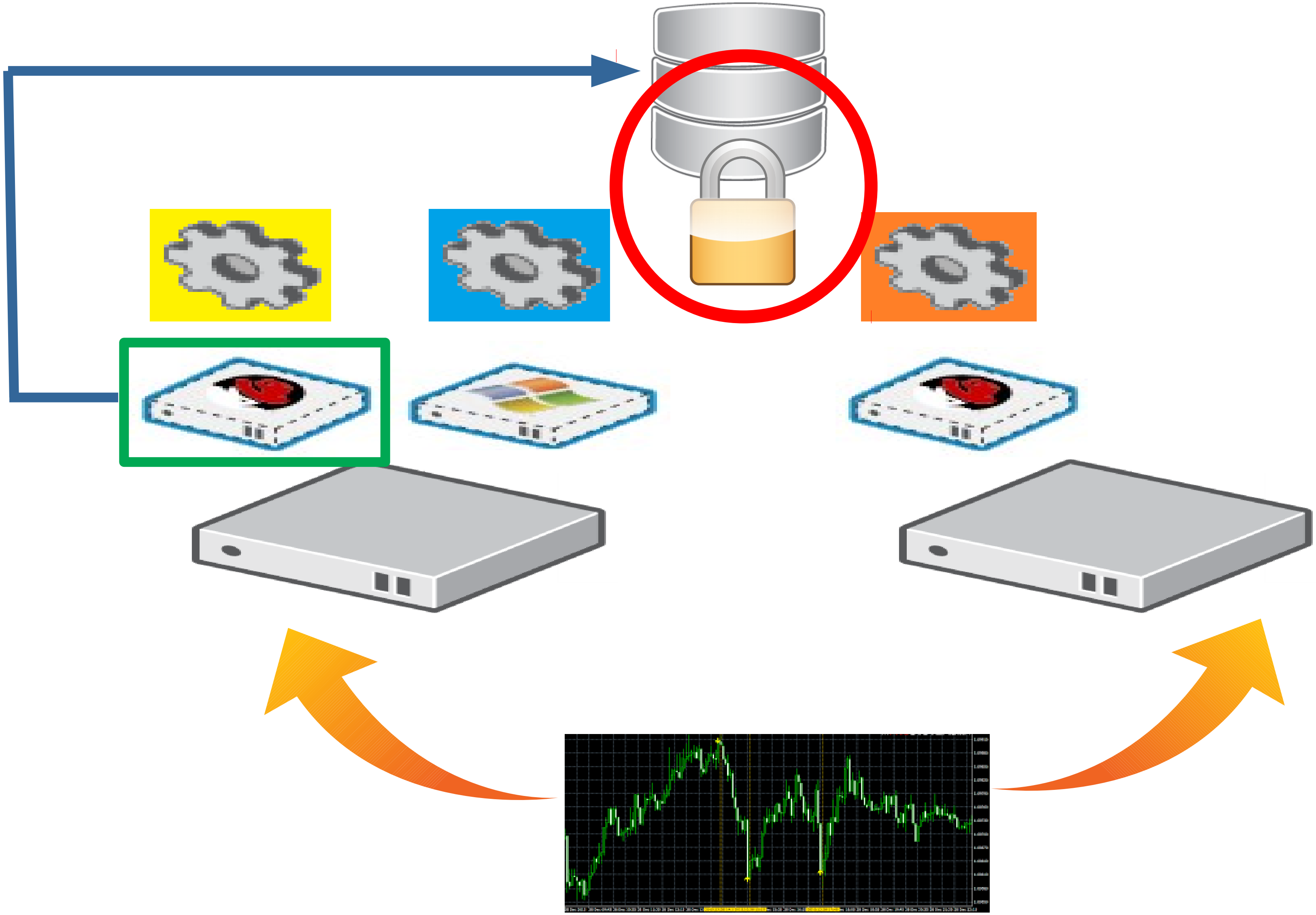
Only the right VM is reported (on startup)

Split Brains May Happen Due to Bugs

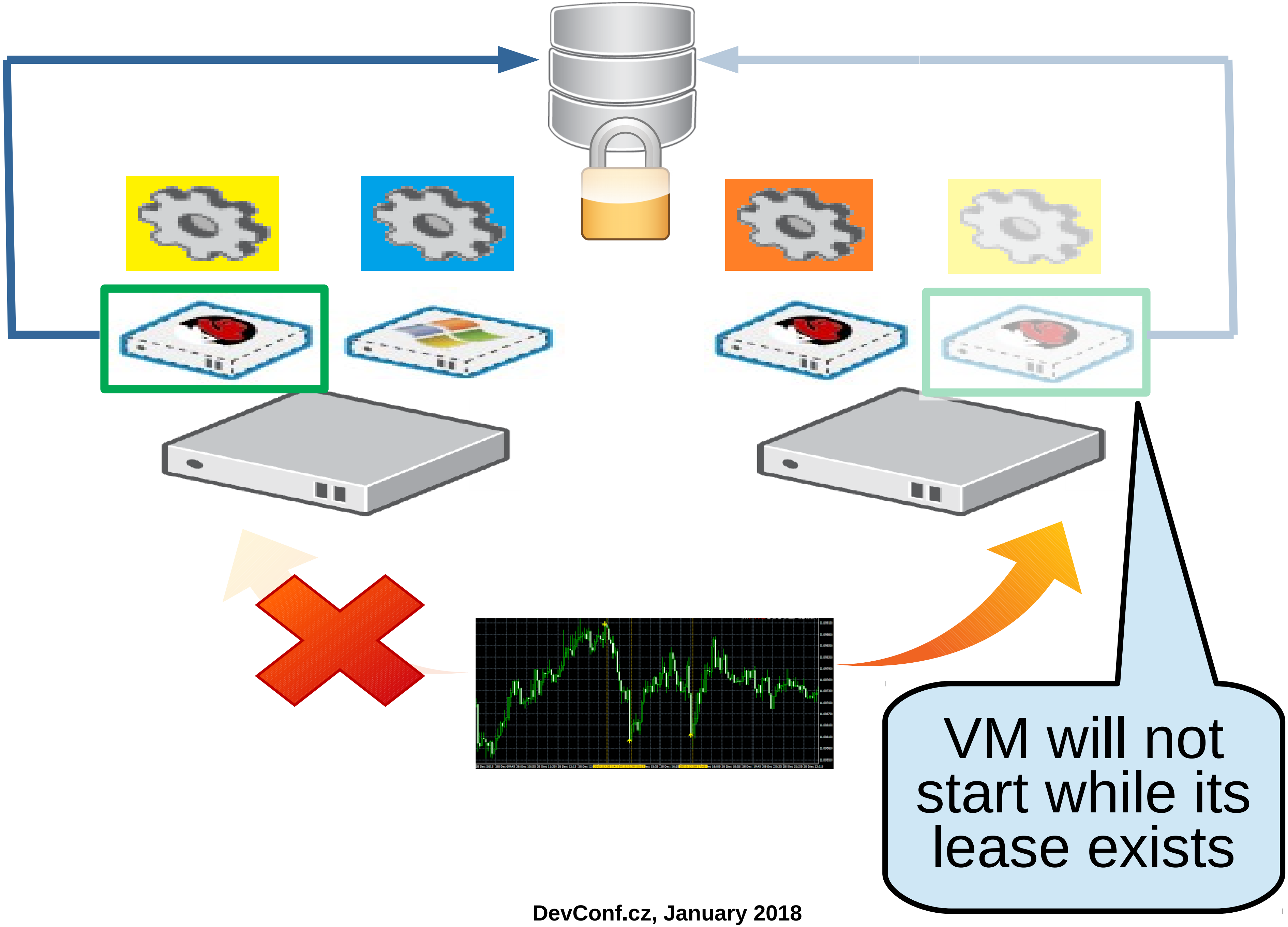


Restart the left VM

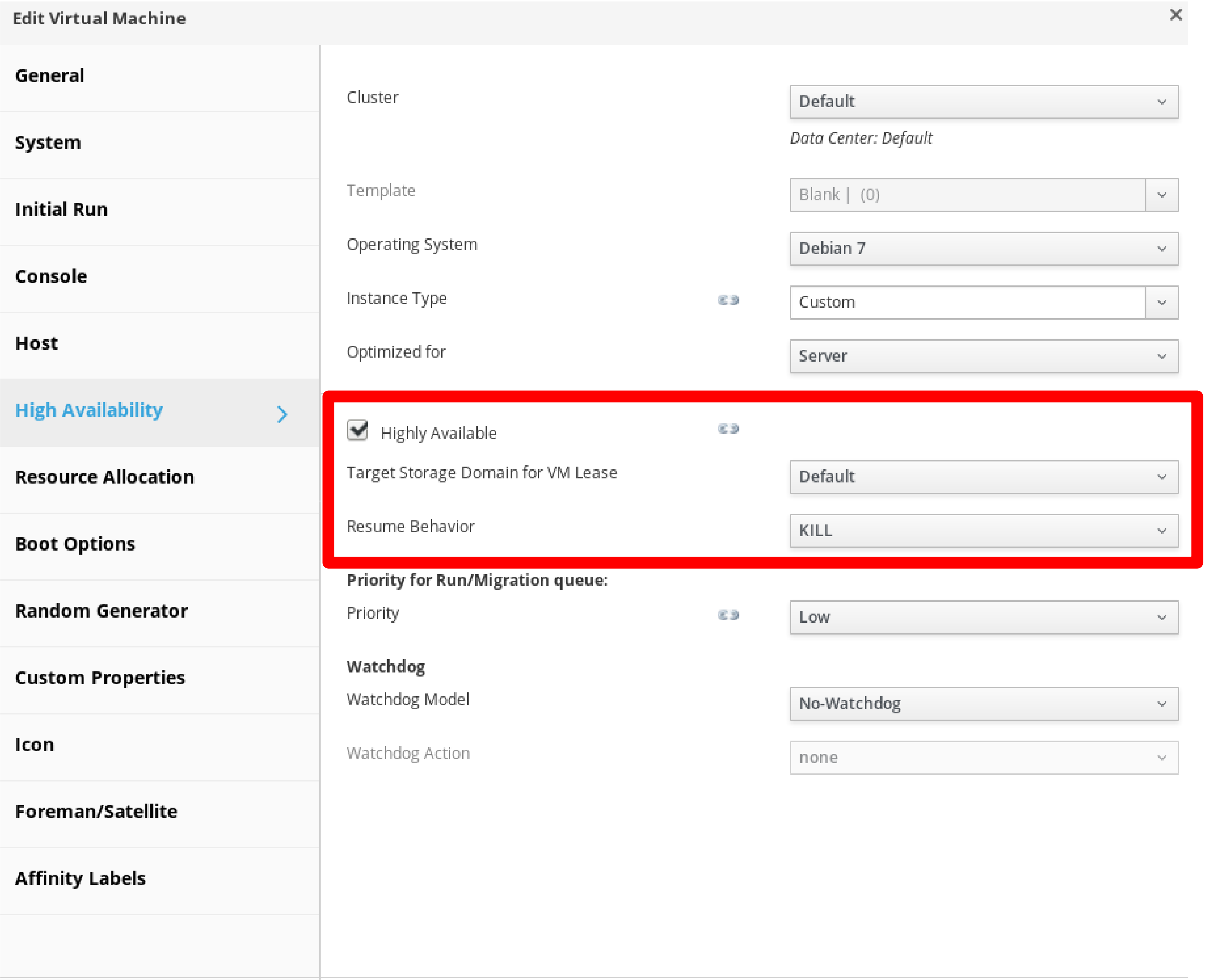
Our Solution to Split Brains: VM Leases



Our Solution to Split Brains: VM Leases


VM will not start while its lease exists

VM Lease Creation



Edit Virtual Machine

Cluster	Default *Data Center: Default*
Template	Blank | (0)
Operating System	Debian 7
Instance Type	Custom
Optimized for	Server

☑ Highly Available

Target Storage Domain for VM Lease	Default
Resume Behavior	KILL

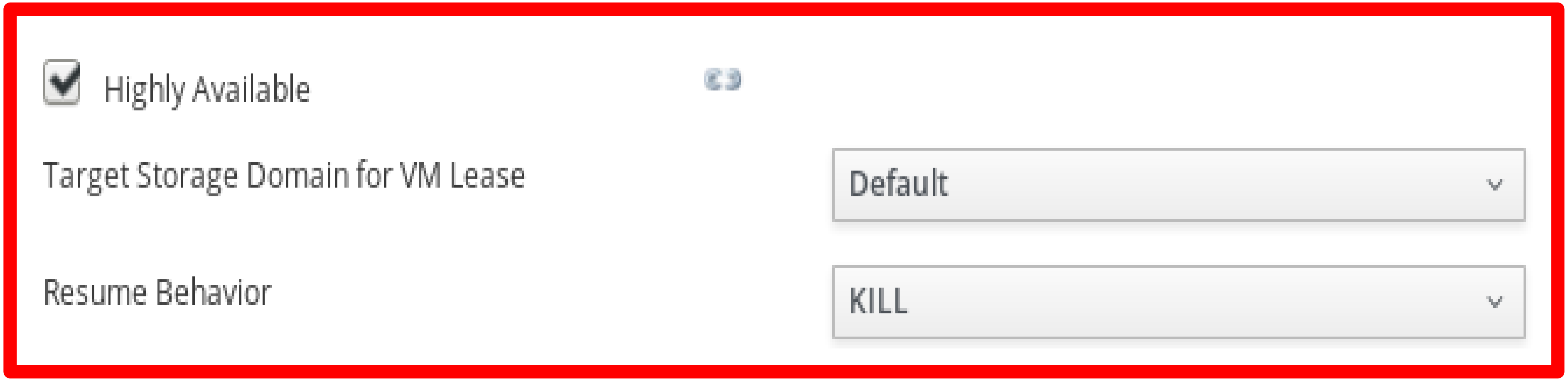
Priority for Run/Migration queue:

Priority	Low

Watchdog

Watchdog Model	No-Watchdog
Watchdog Action	none

VM Lease Creation

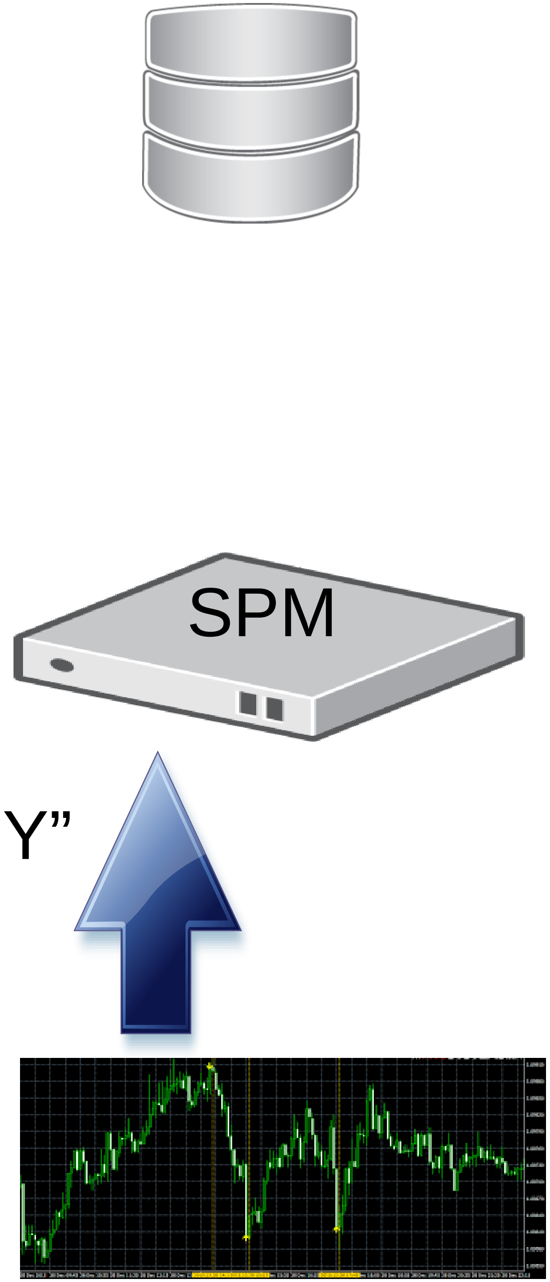


☑ Highly Available

Target Storage Domain for VM Lease: Default

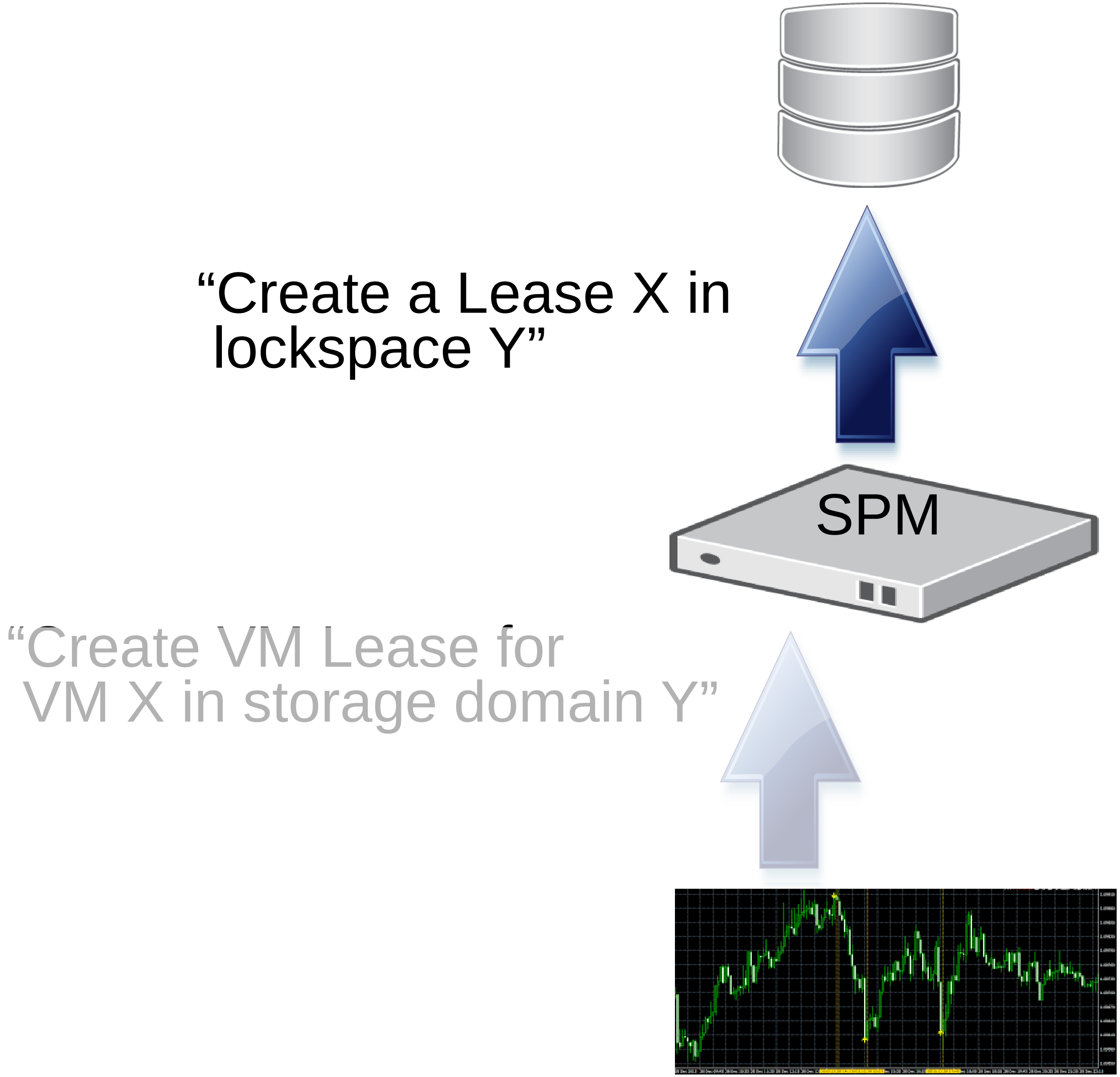
Resume Behavior: KILL

VM Lease Creation



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

VM Lease Creation

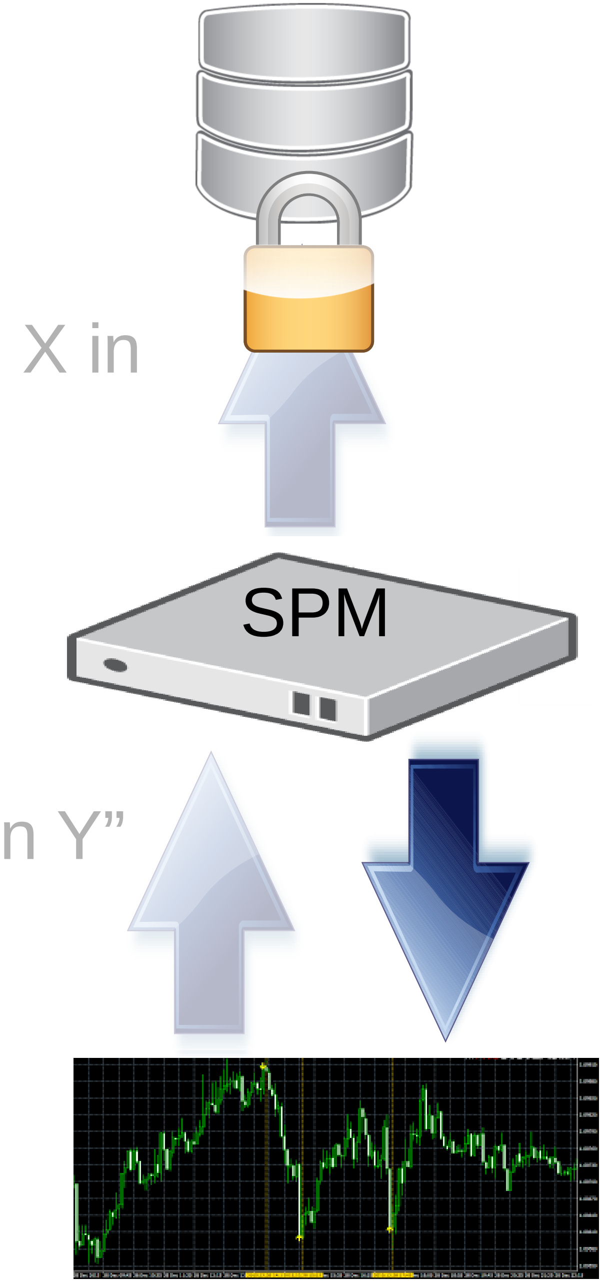


“Create a Lease X in lockspace Y”

SPM

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

VM Lease Creation



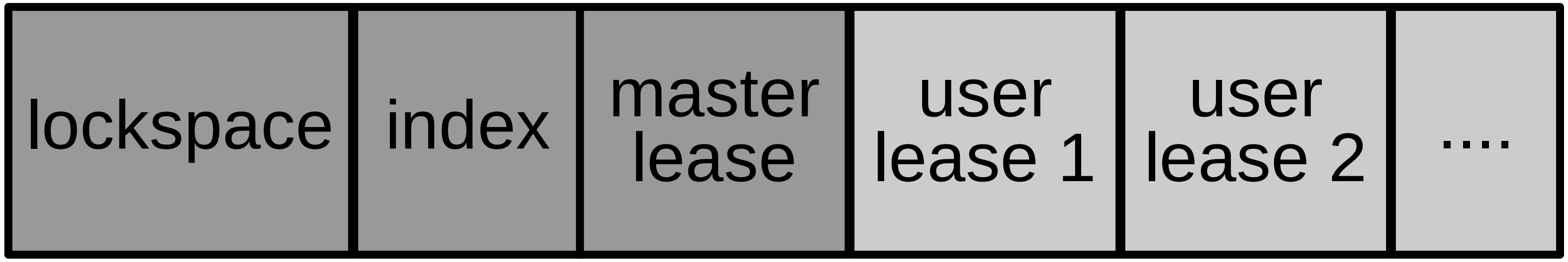
“Create a Lease X in lockspace Y”

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

“Path P to xleases volume and Lease offset O”

xleases volume

- Sanlock does not manage leases allocation
- Volume layout:



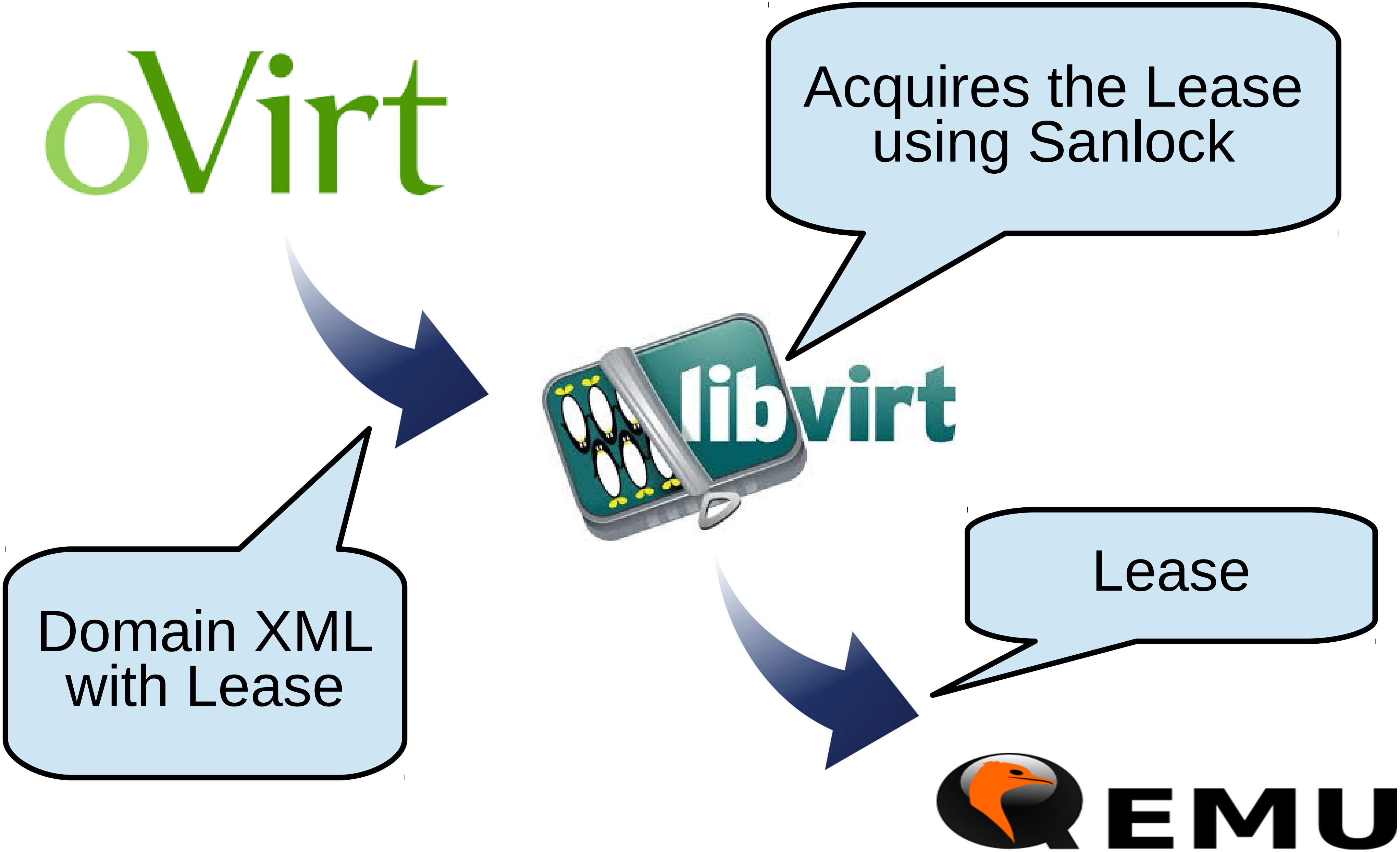
lockspace | index | master lease | user lease 1 | user lease 2 | ....

- 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

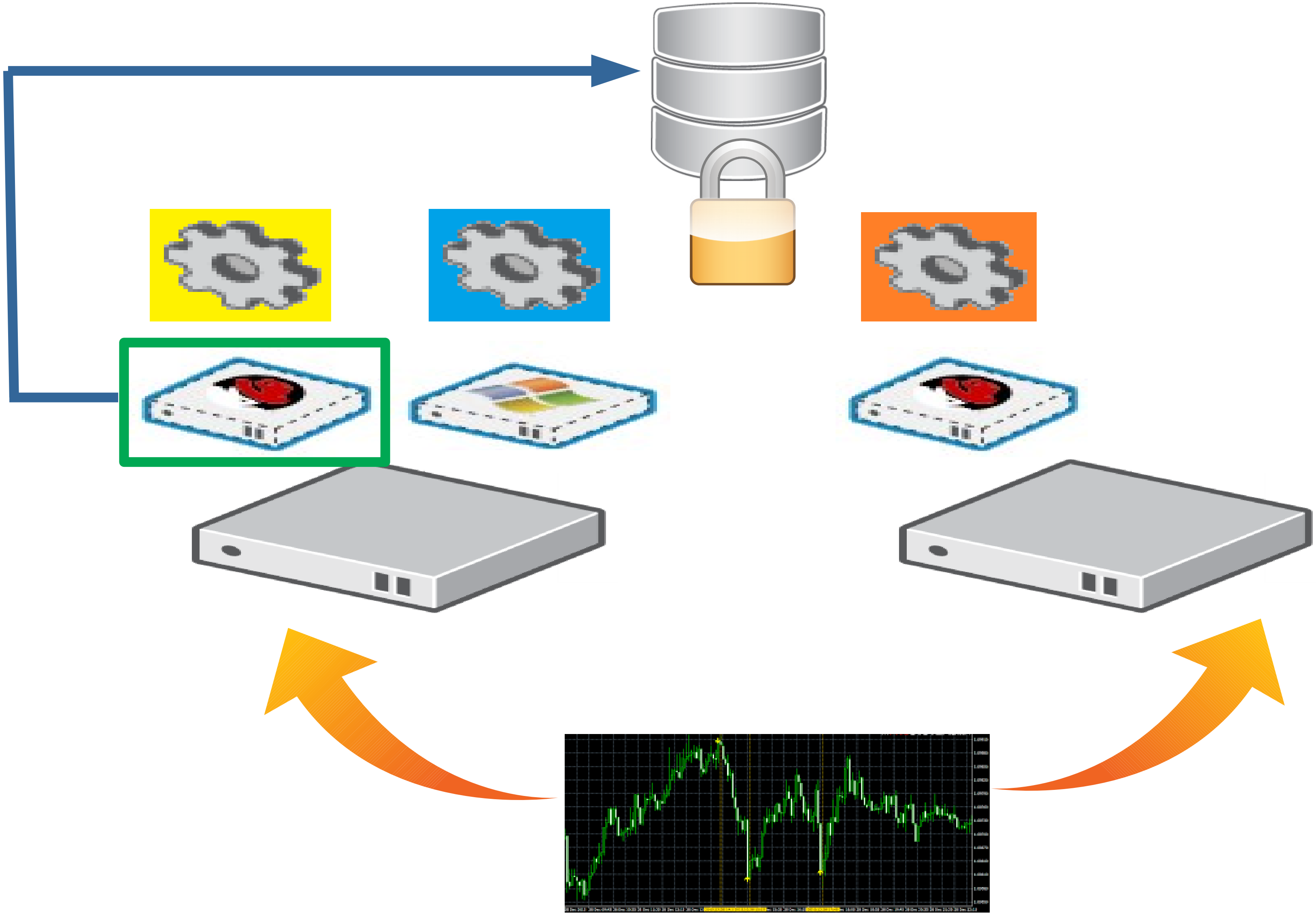


Acquires the Lease using Sanlock

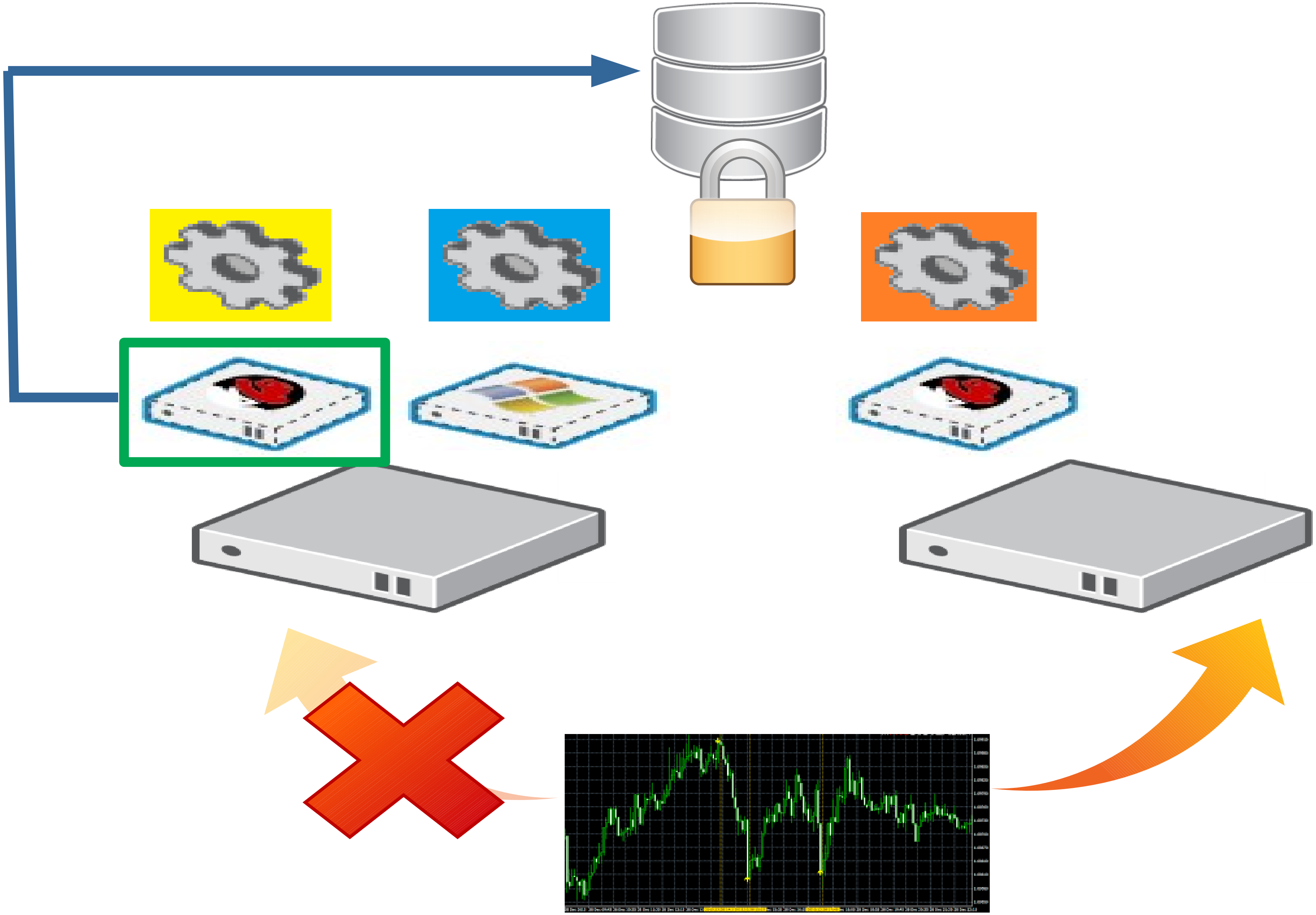
Domain XML with Lease

Lease

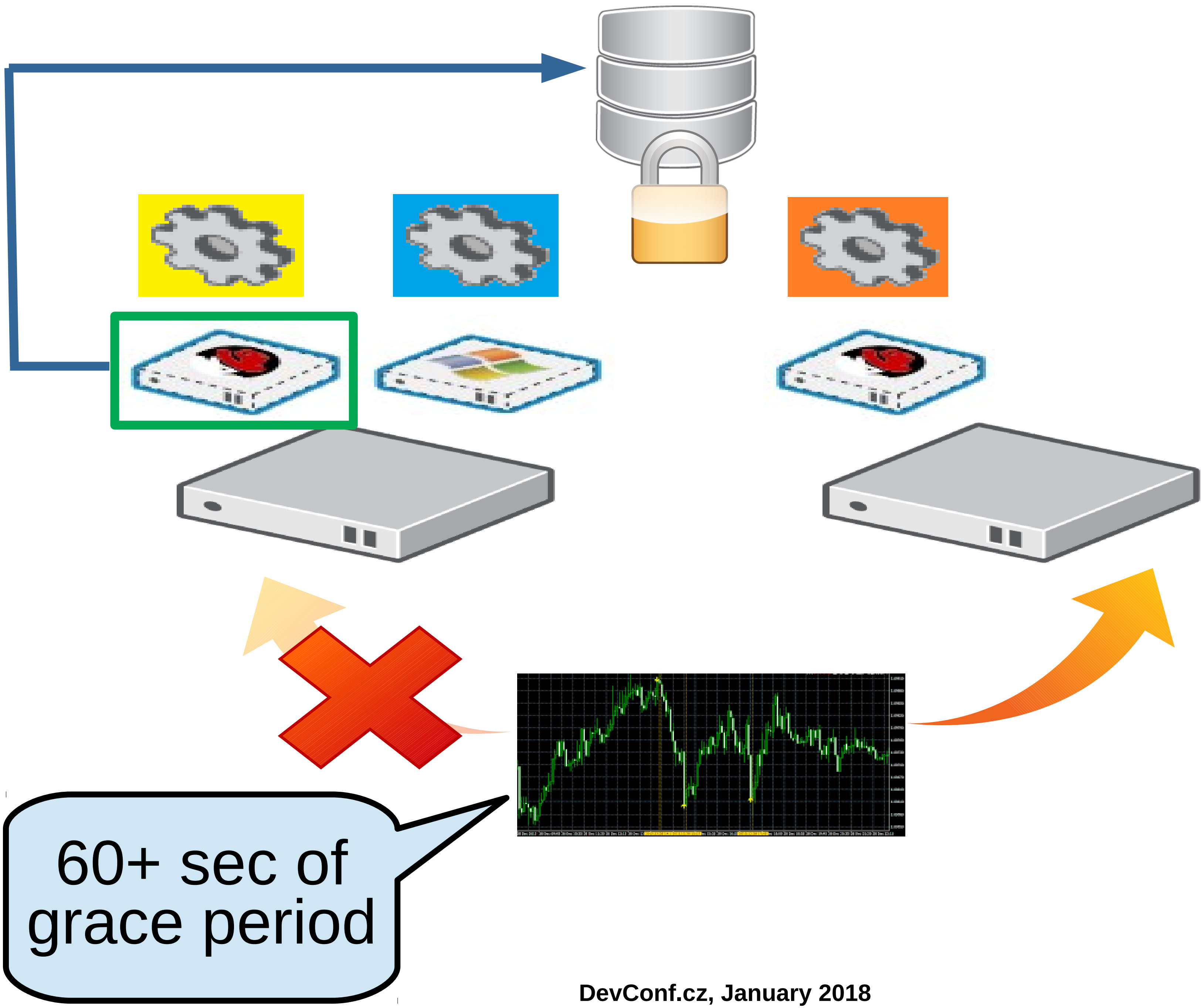
Non-Responsive Host Treatment



Non-Responsive Host Treatment

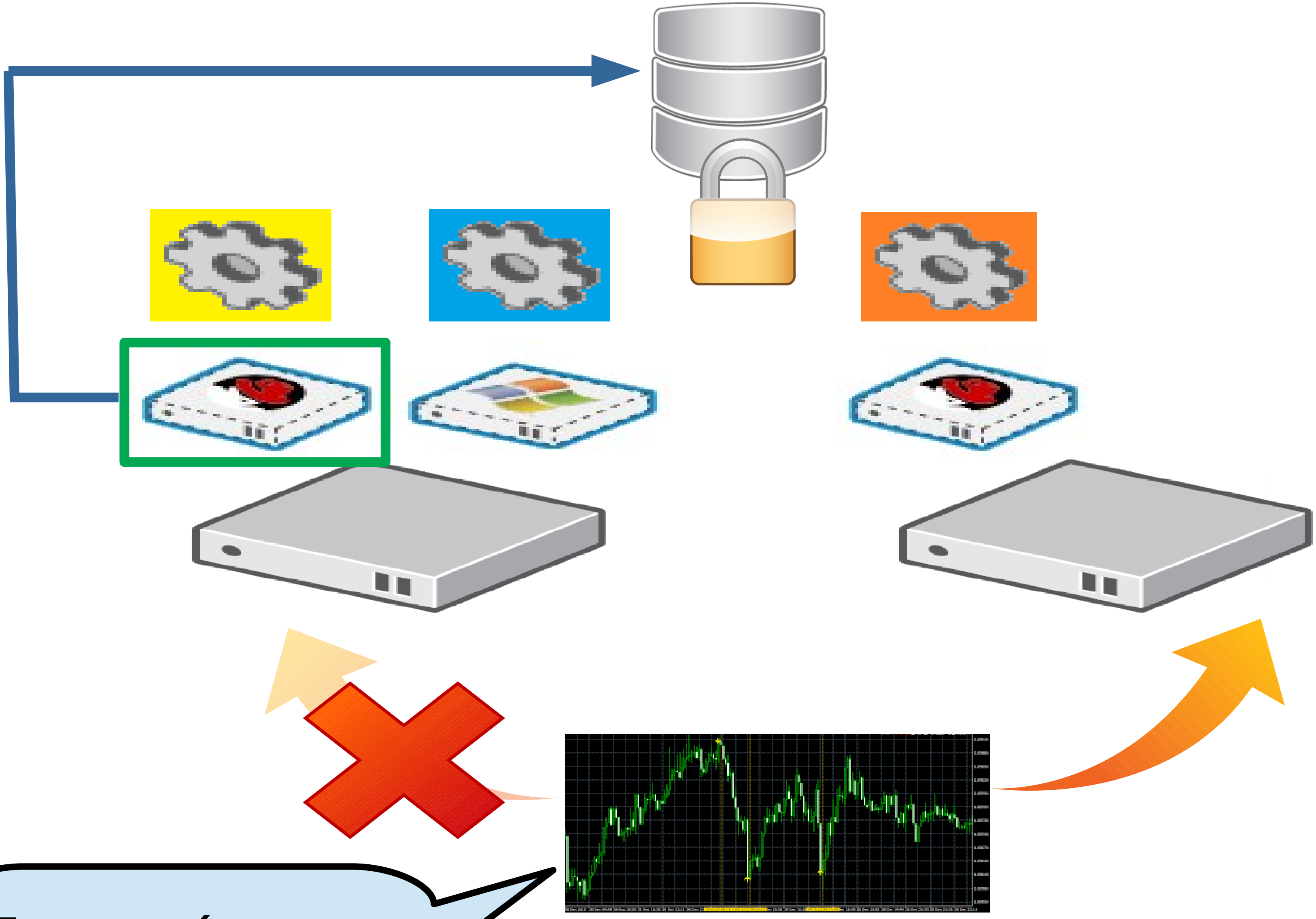


Non-Responsive Host Treatment



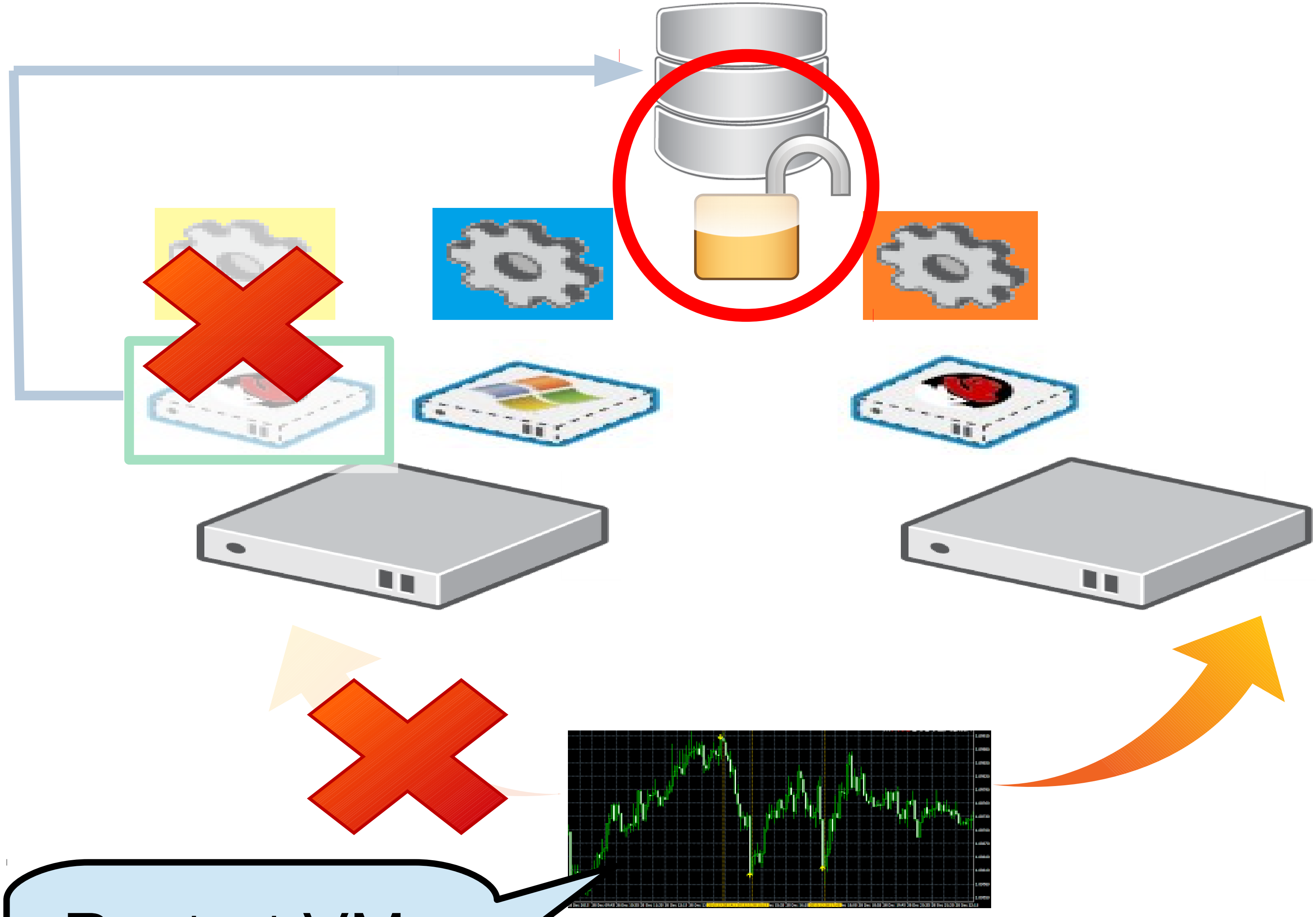
60+ sec of grace period

Non-Responsive Host Treatment



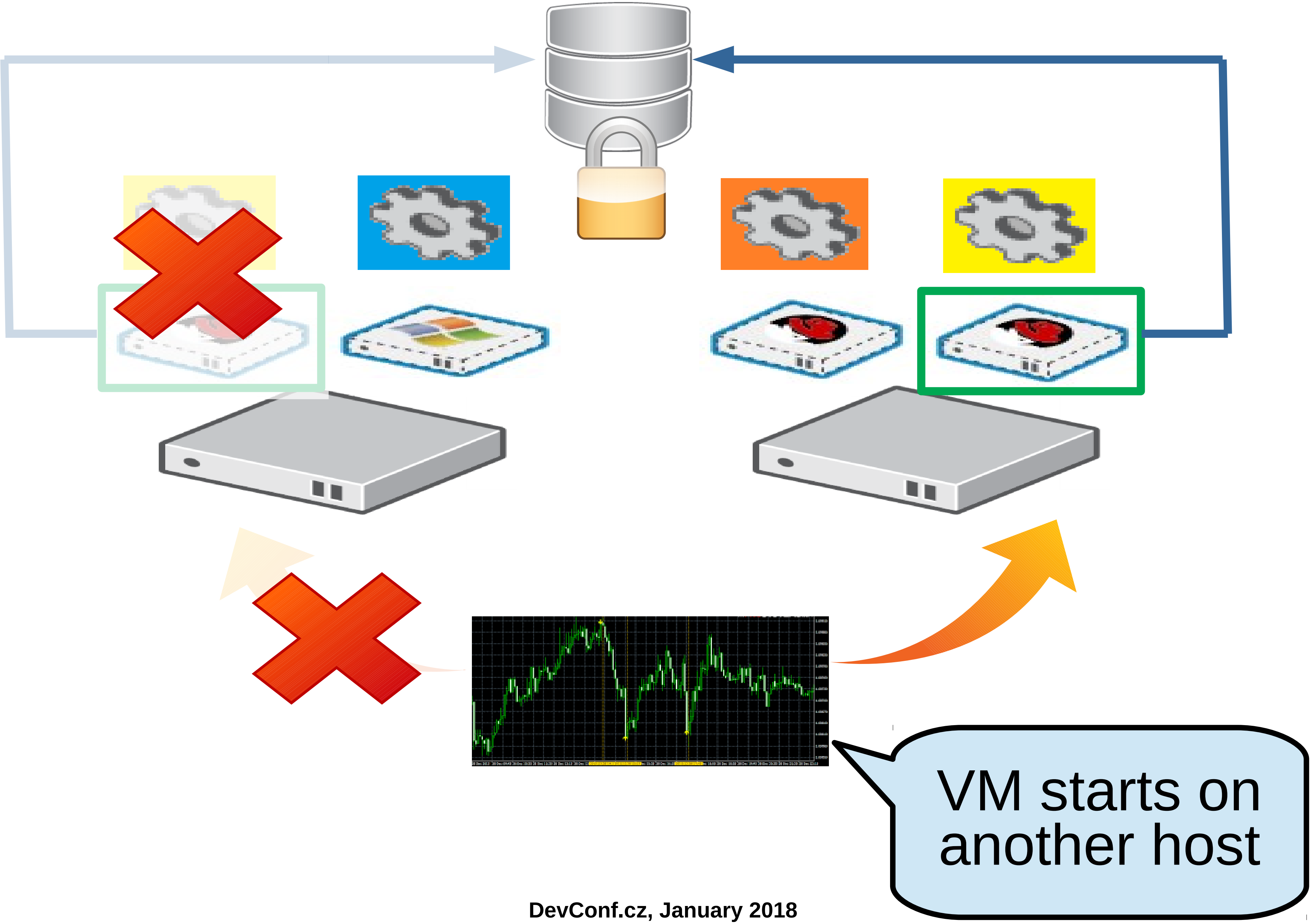
Fence (power Management)

Non-Responsive Host + VM is Down



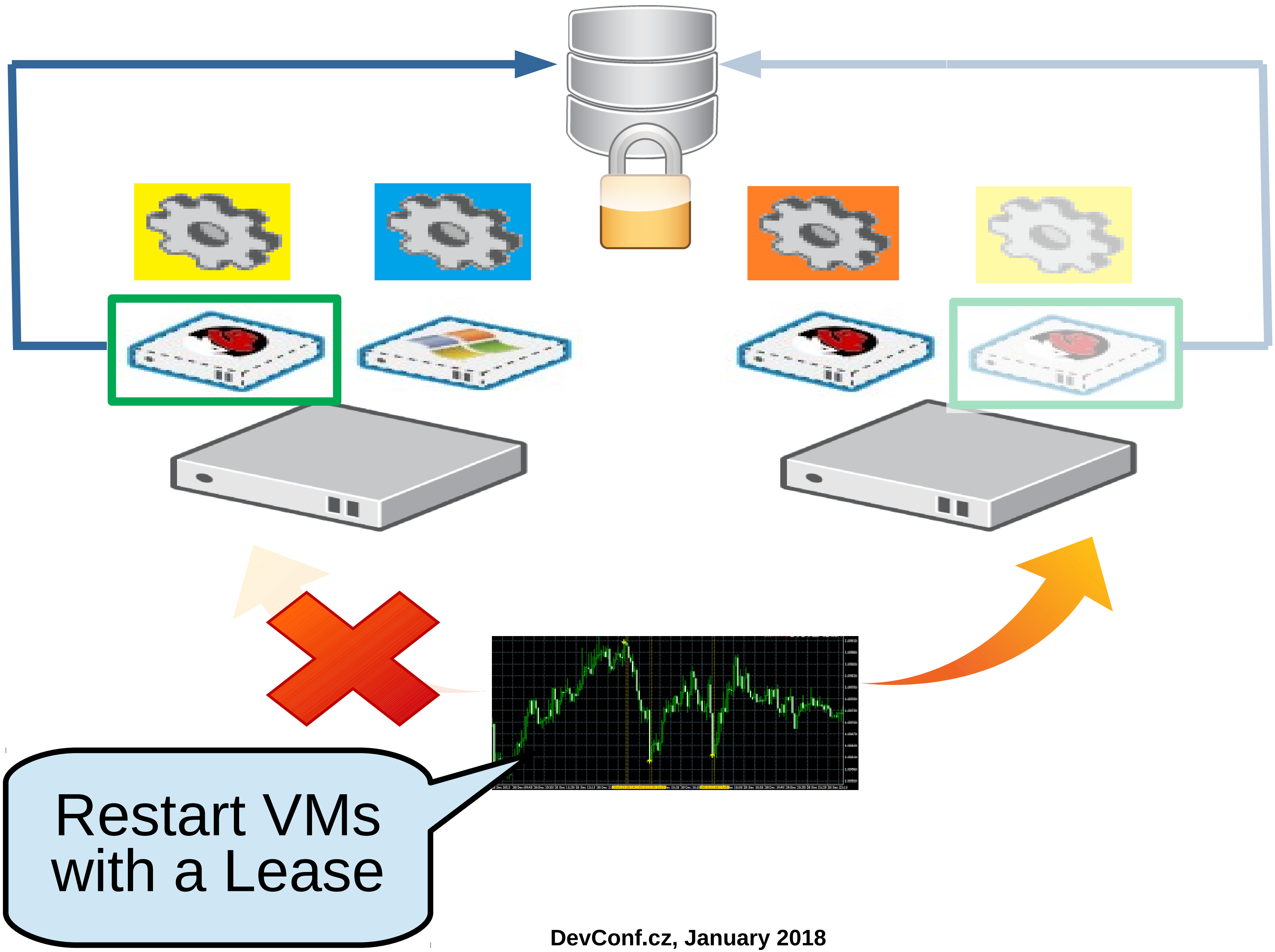
Restart VMs with a Lease

Non-Responsive Host + VM is Down

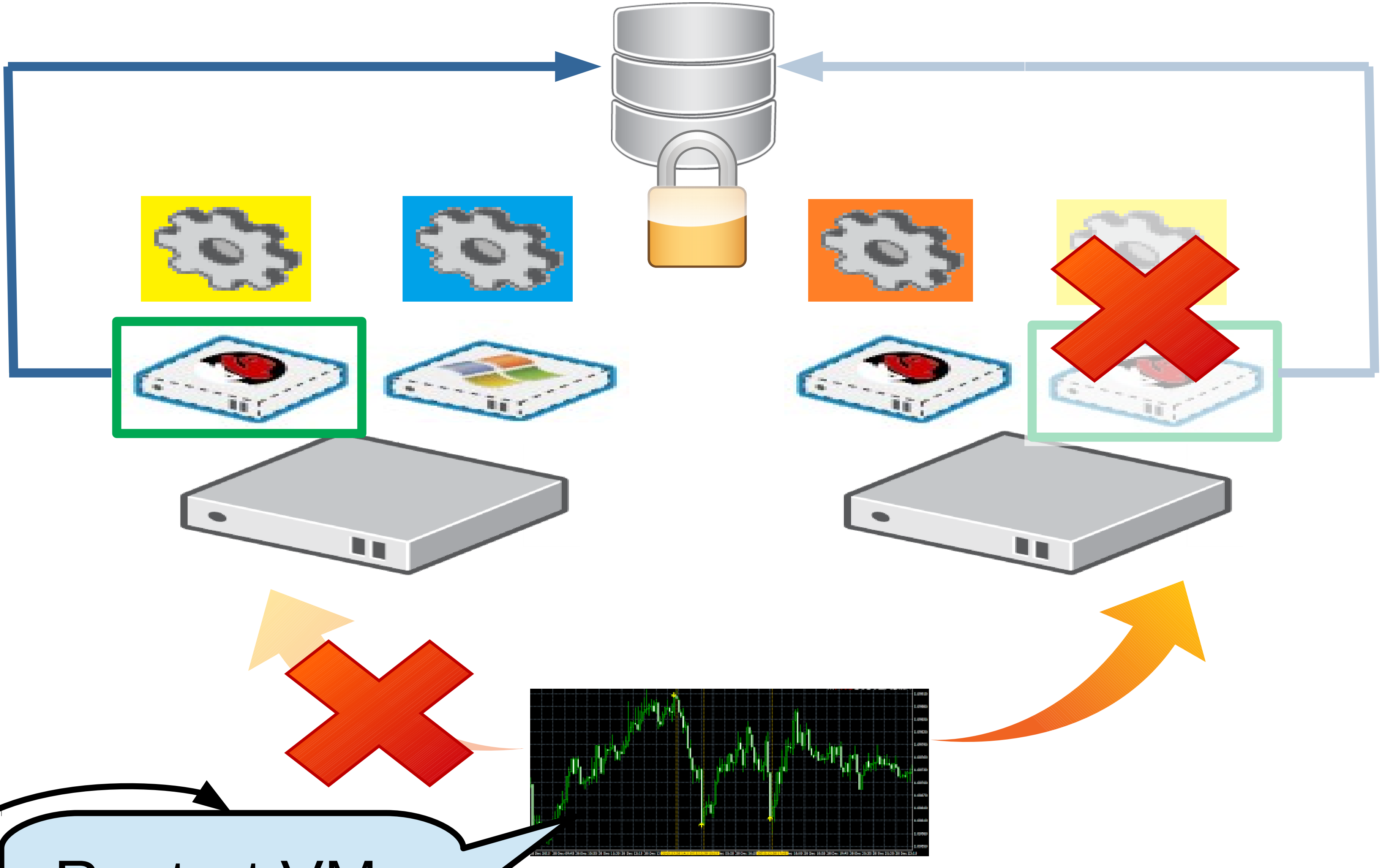


VM starts on another host

Non-Responsive Host + VM is UP

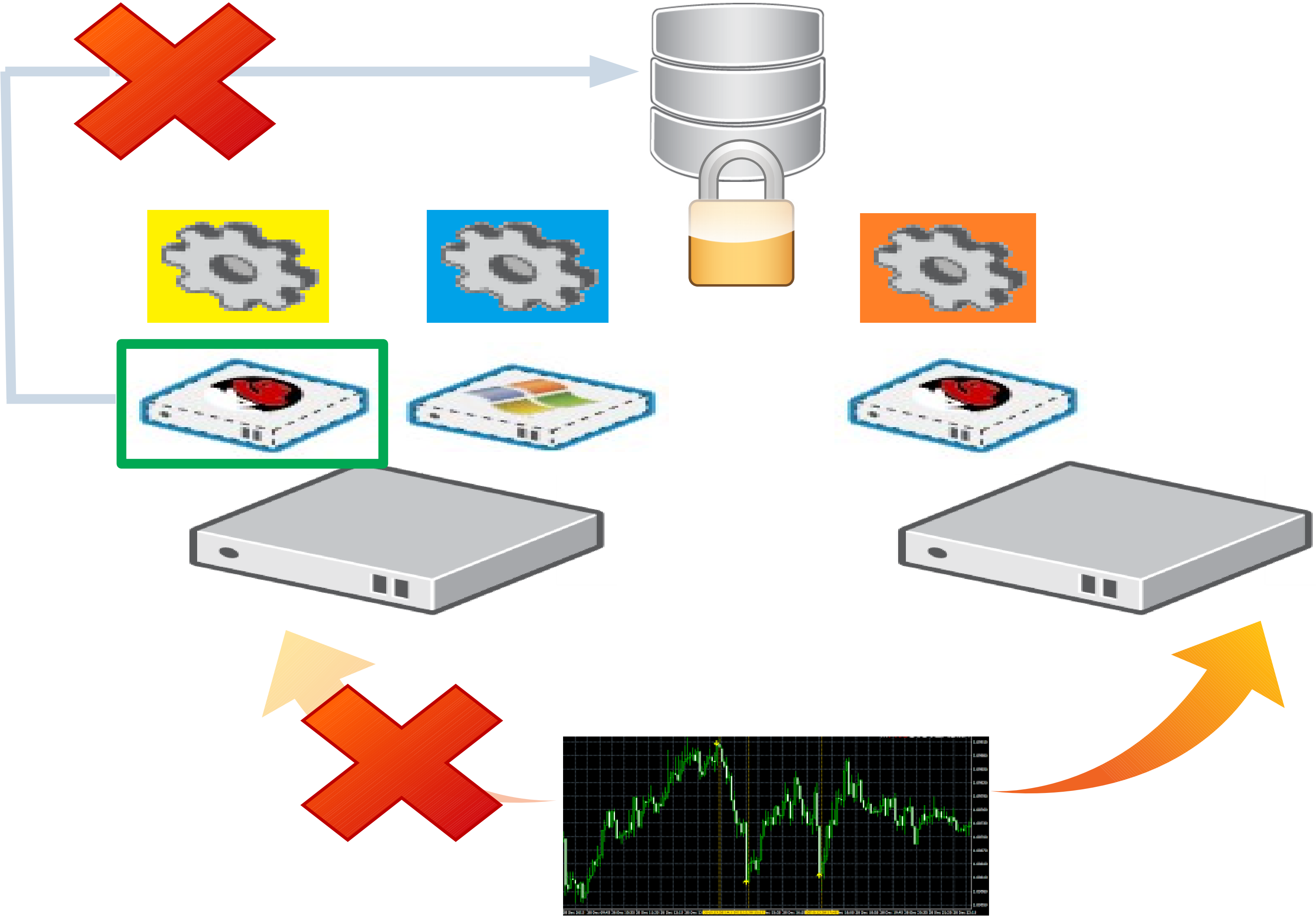


Non-Responsive Host + VM is UP

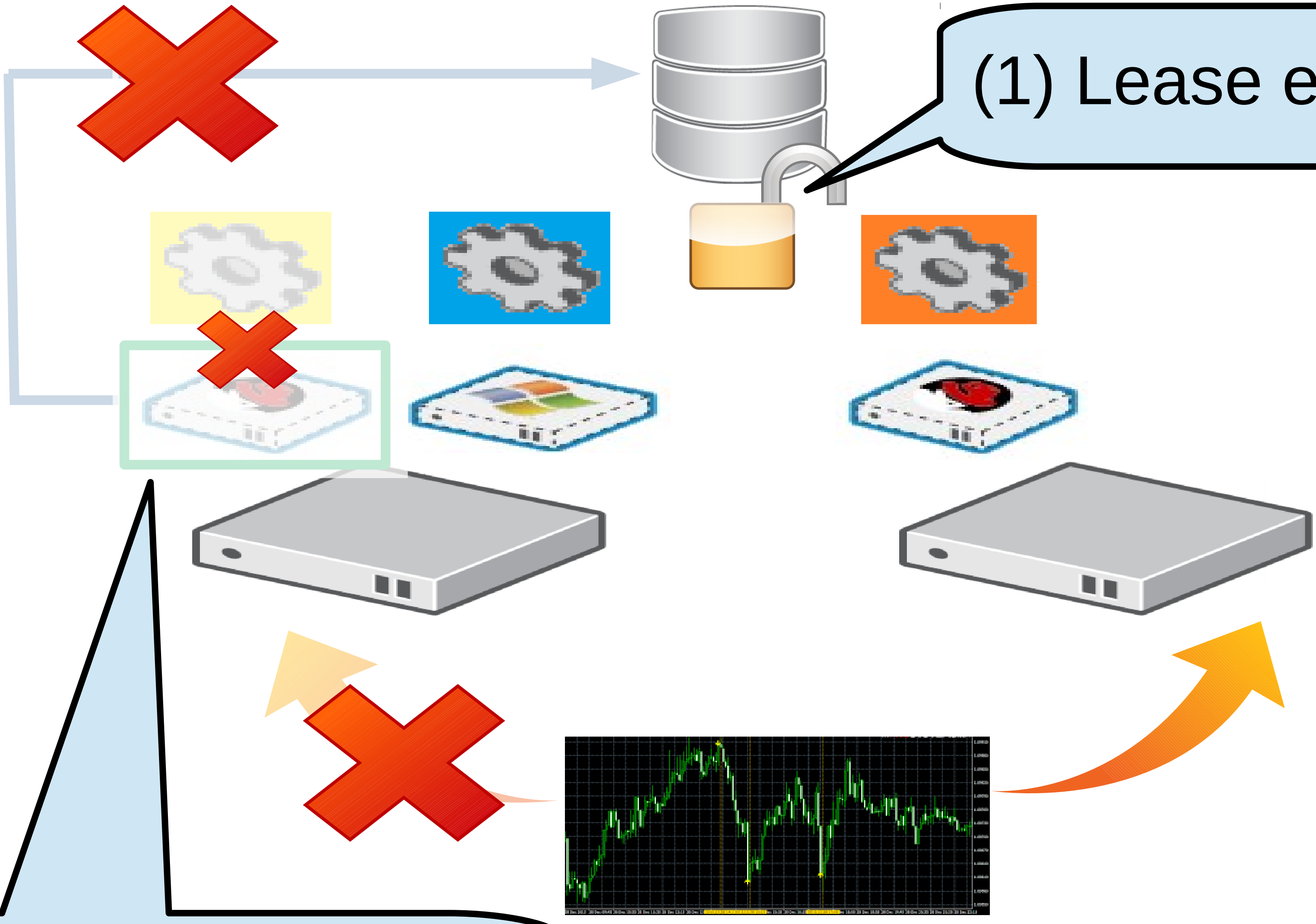


Restart VMs with a Lease

Disconnected Host + VM is UP



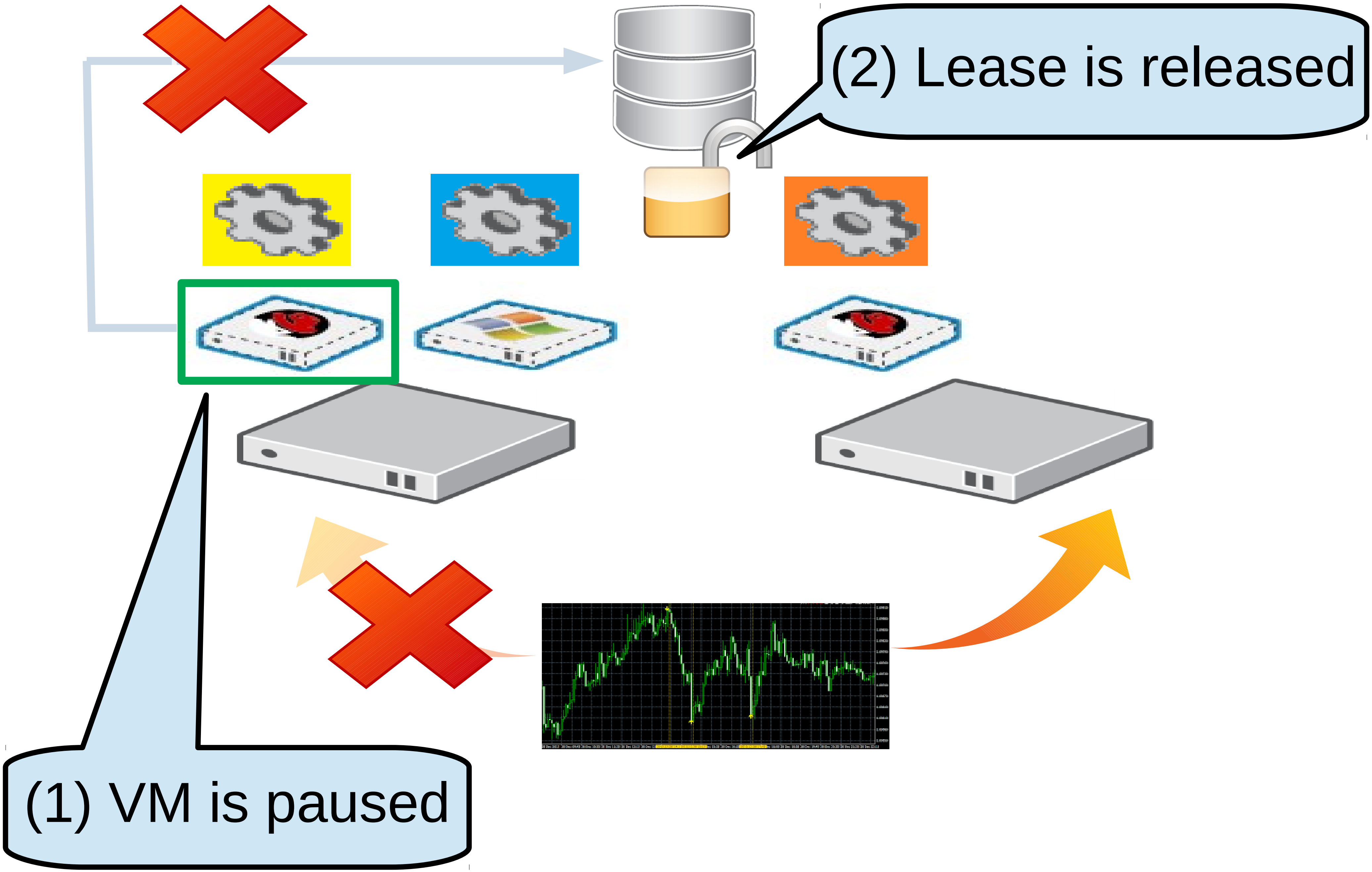
Disconnected Host + VM is UP (1)



(1) Lease expires

(2) VM is terminated

Disconnected Host + VM is UP (2)



(2) Lease is released

(1) VM is paused

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
- 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