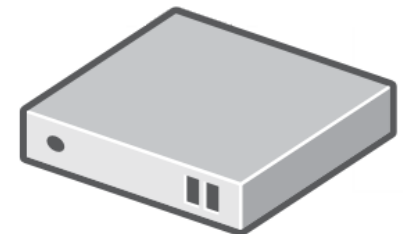
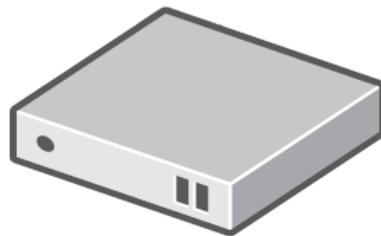
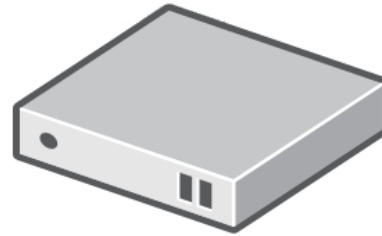
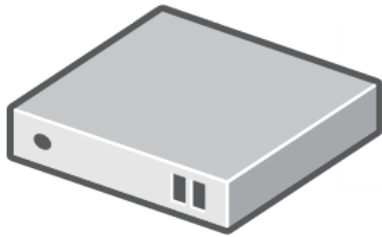
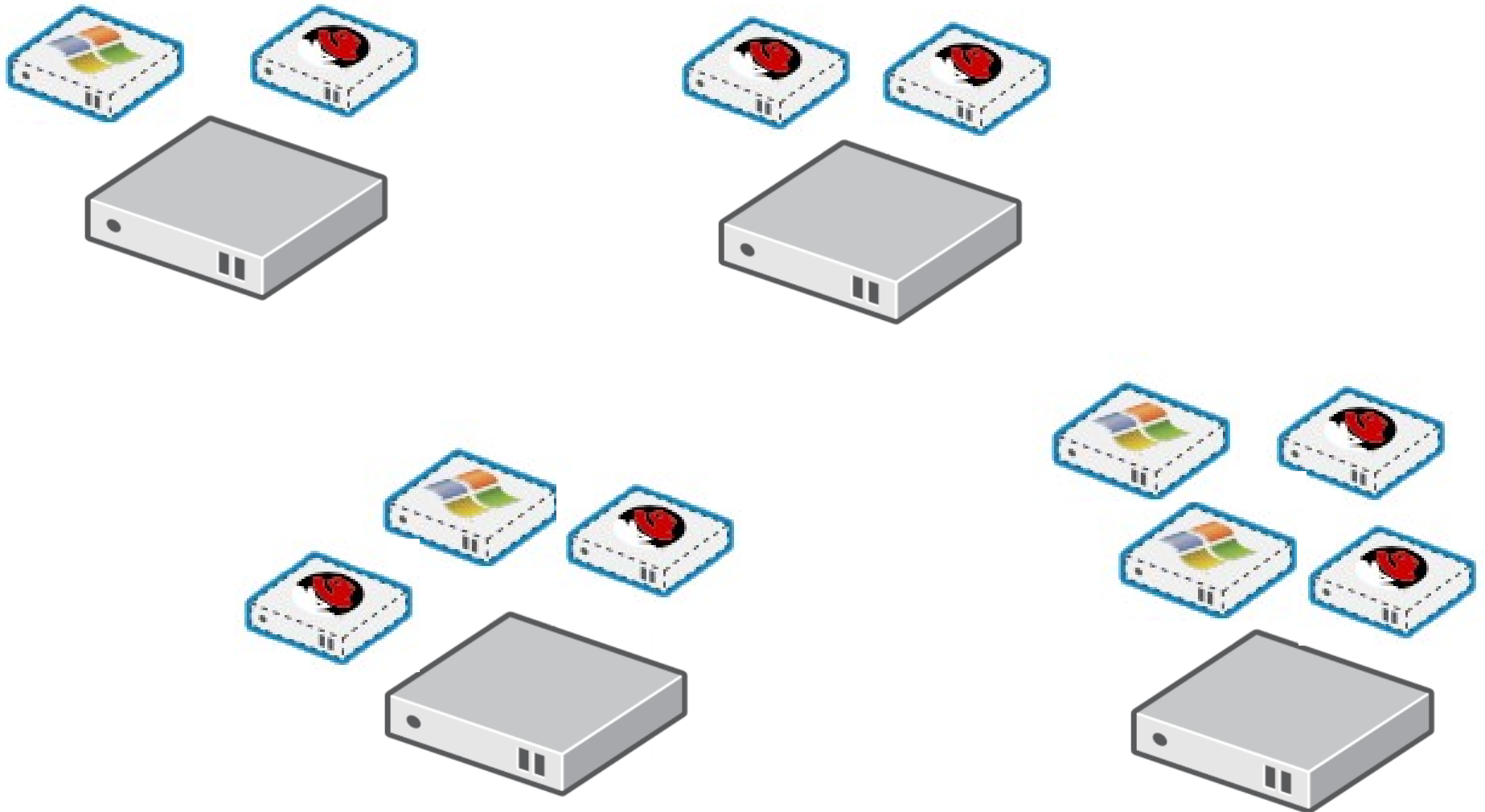


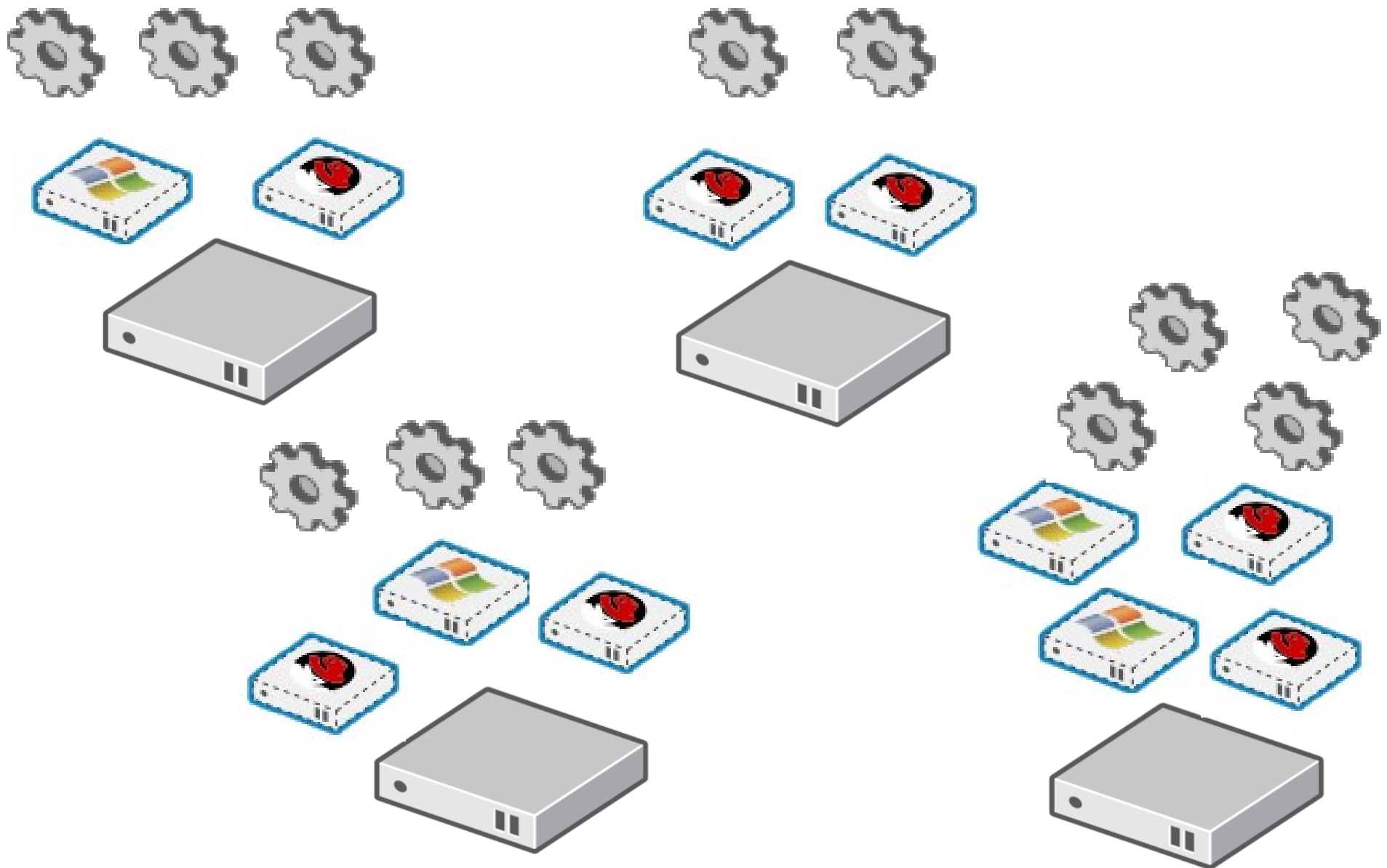
High Availability with No Split Brains!

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

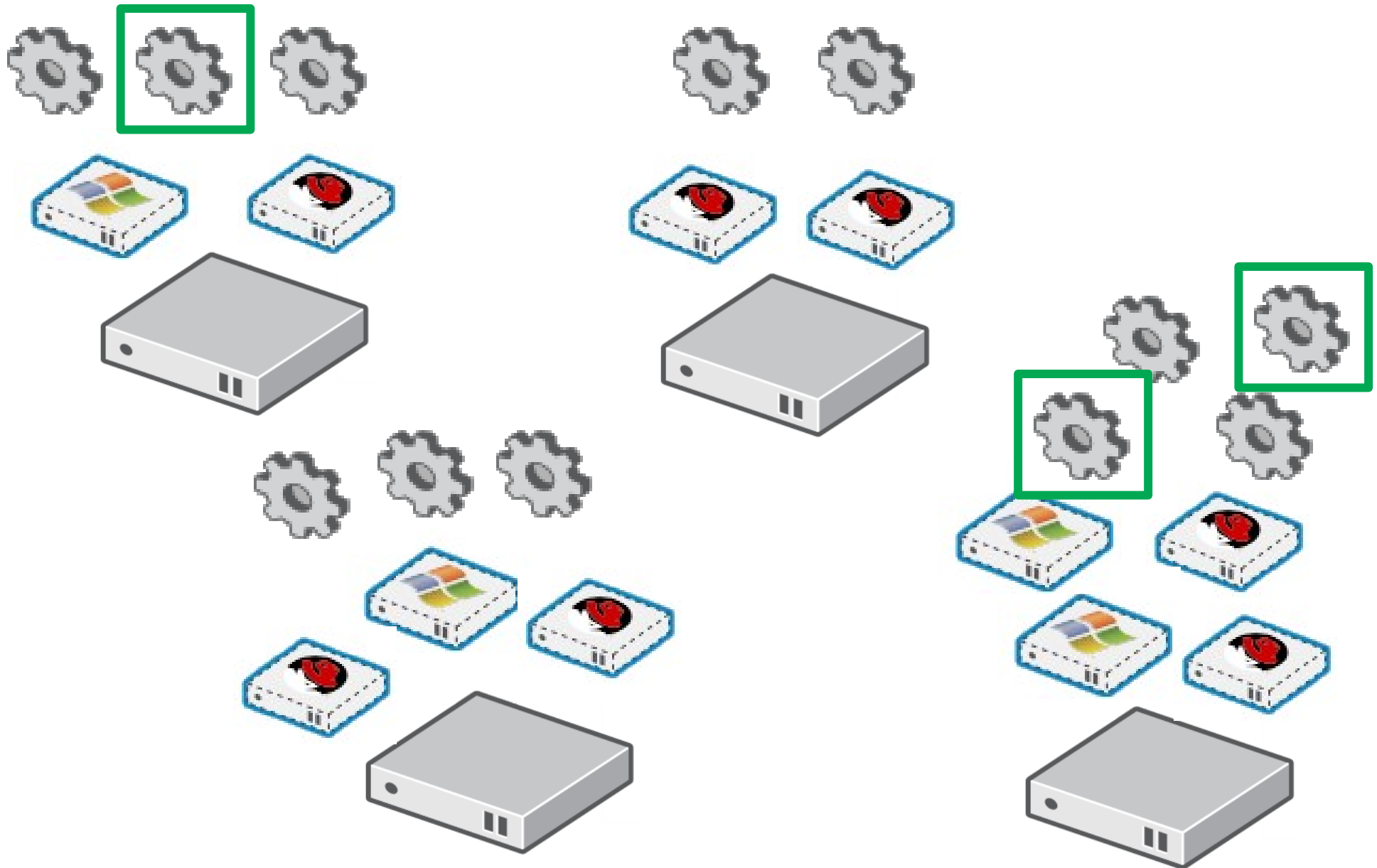




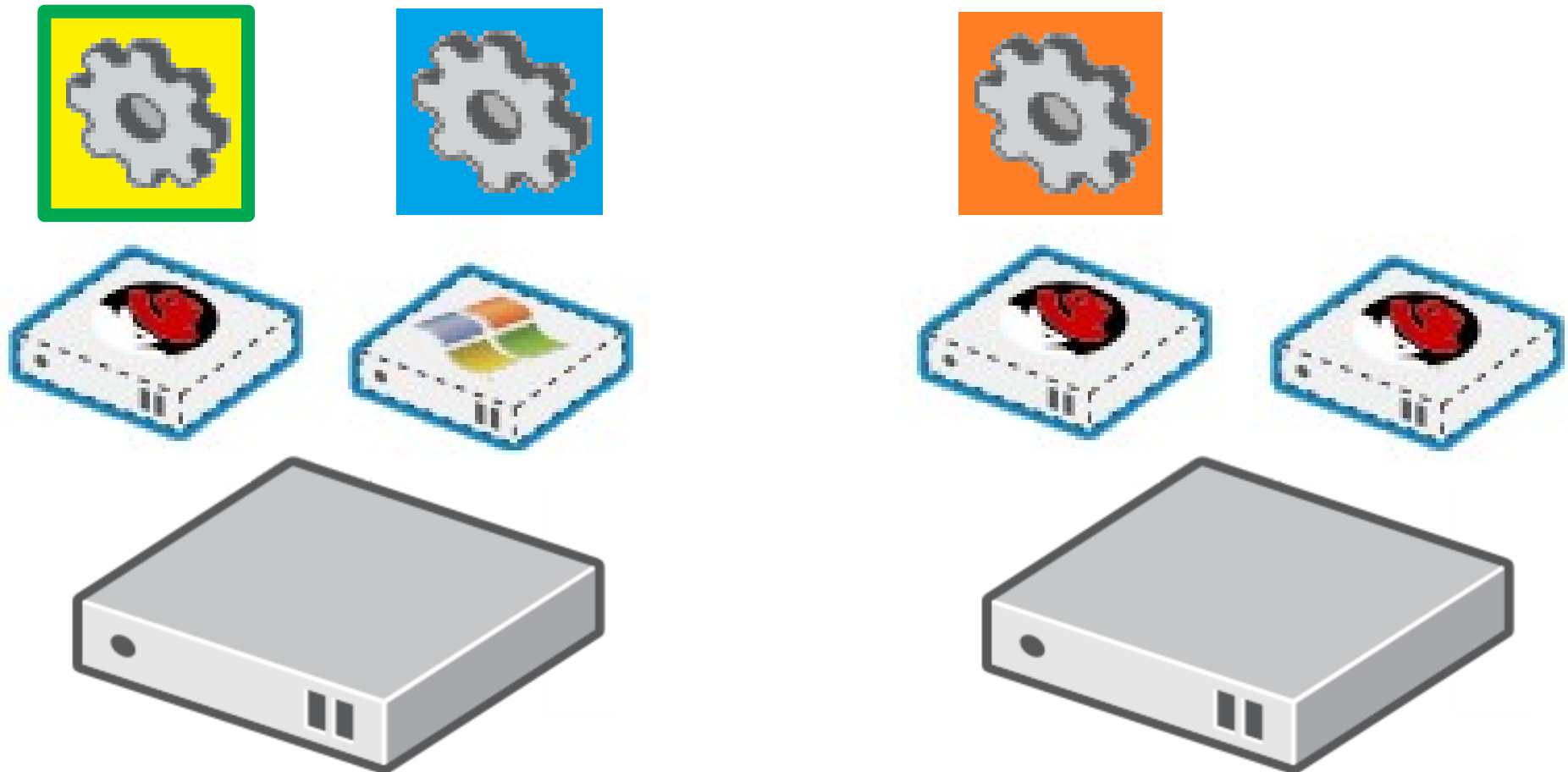
oVirt Virtual Data Center - Applications



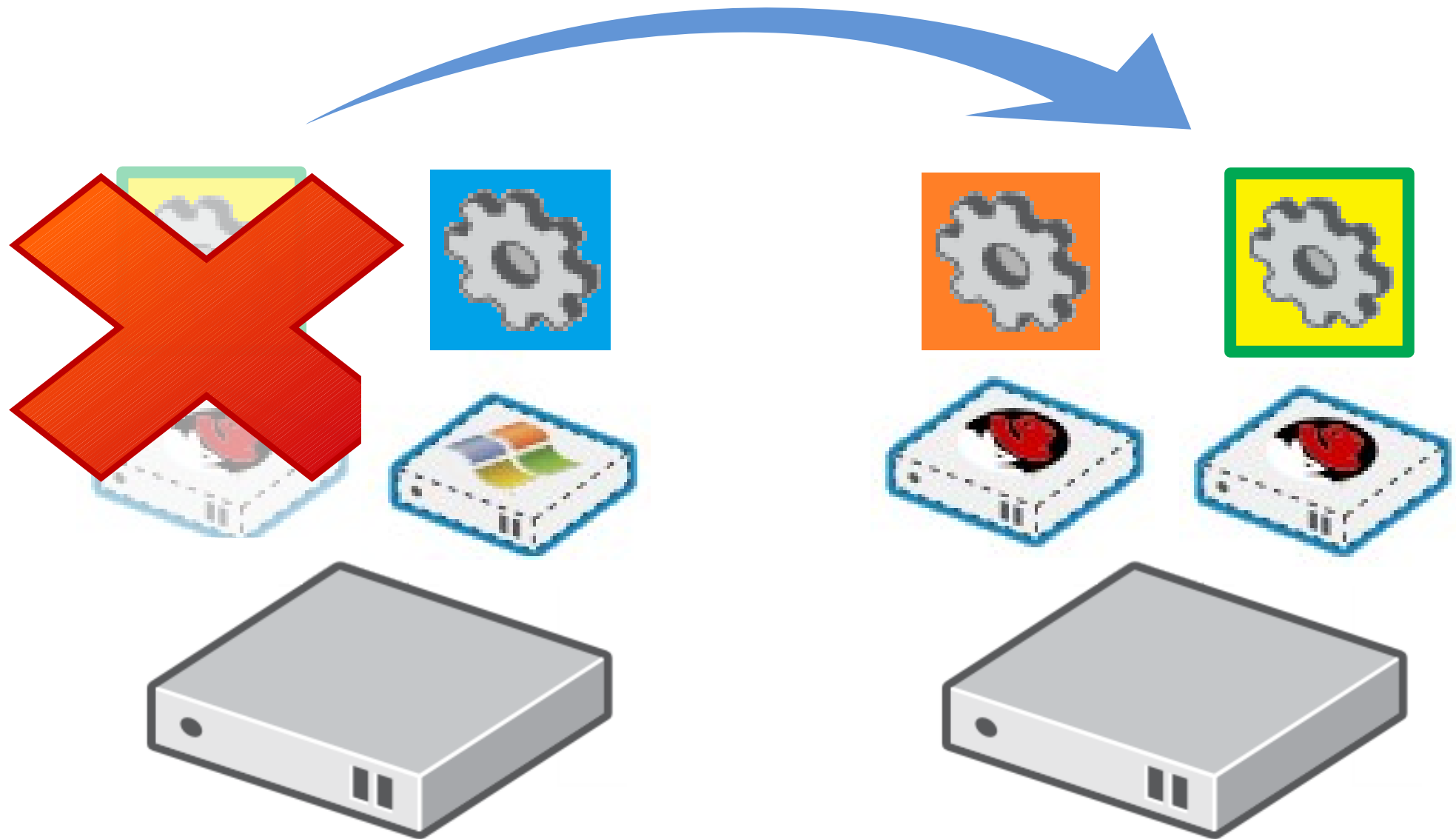
oVirt Some Applications are More Critical



High Availability - Application-Level

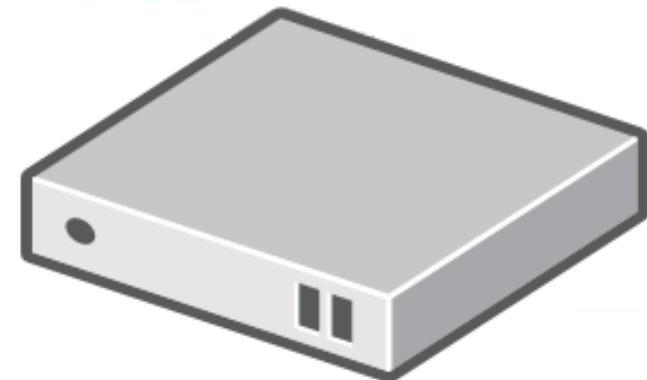
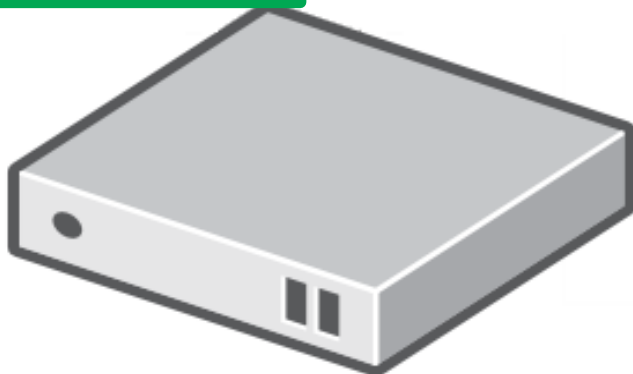
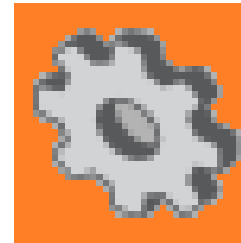
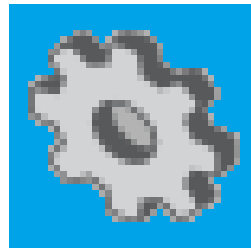
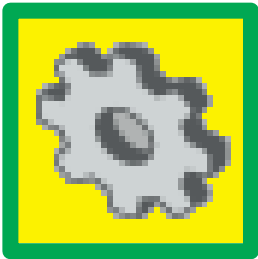


High Availability - Application-Level

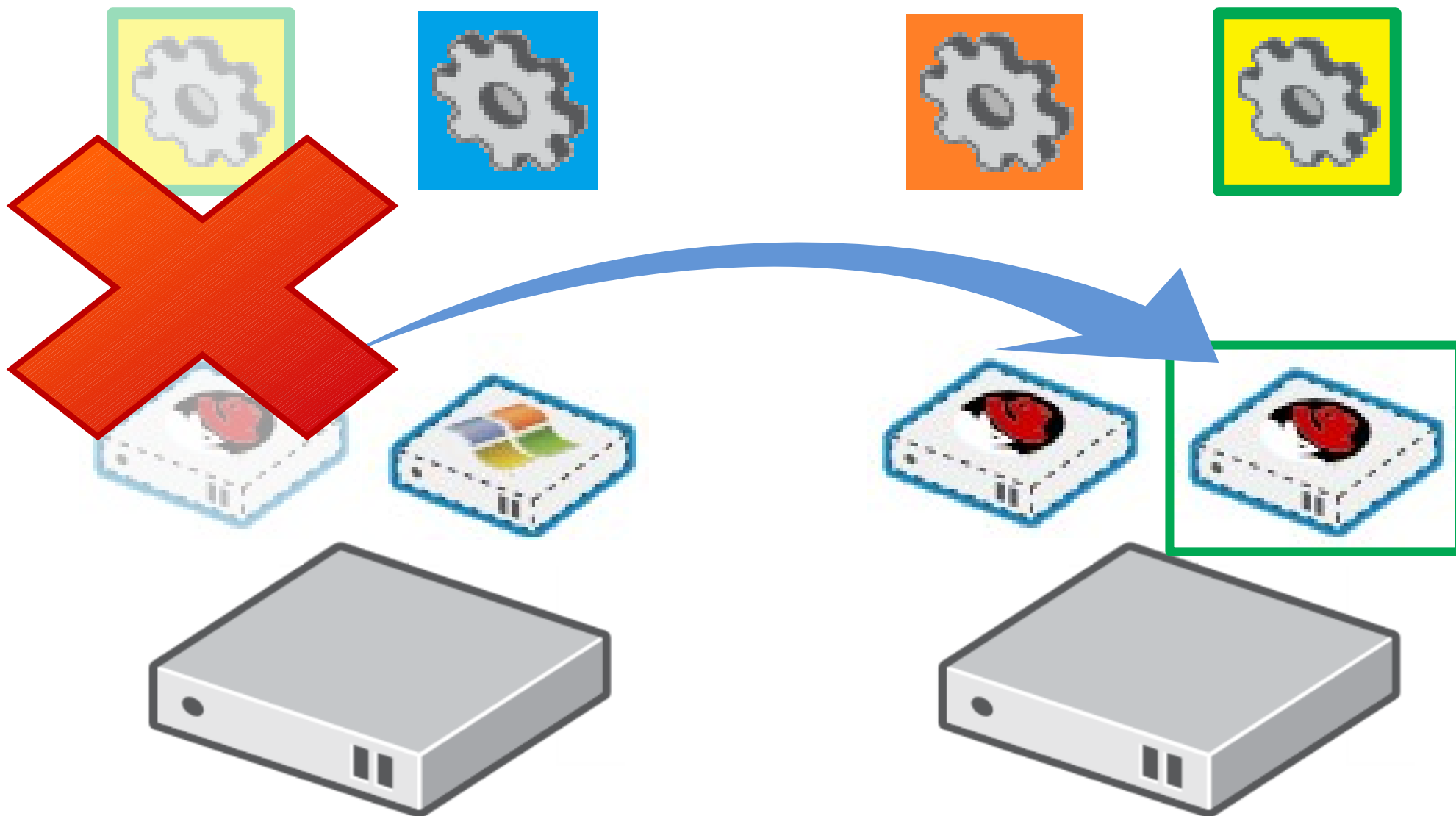


- Higher resource consumption
- More responsibility on the application
- Backup starts in a different environment
 - Different IP address(es)
 - Different disk(s)

High Availability - VM-Level

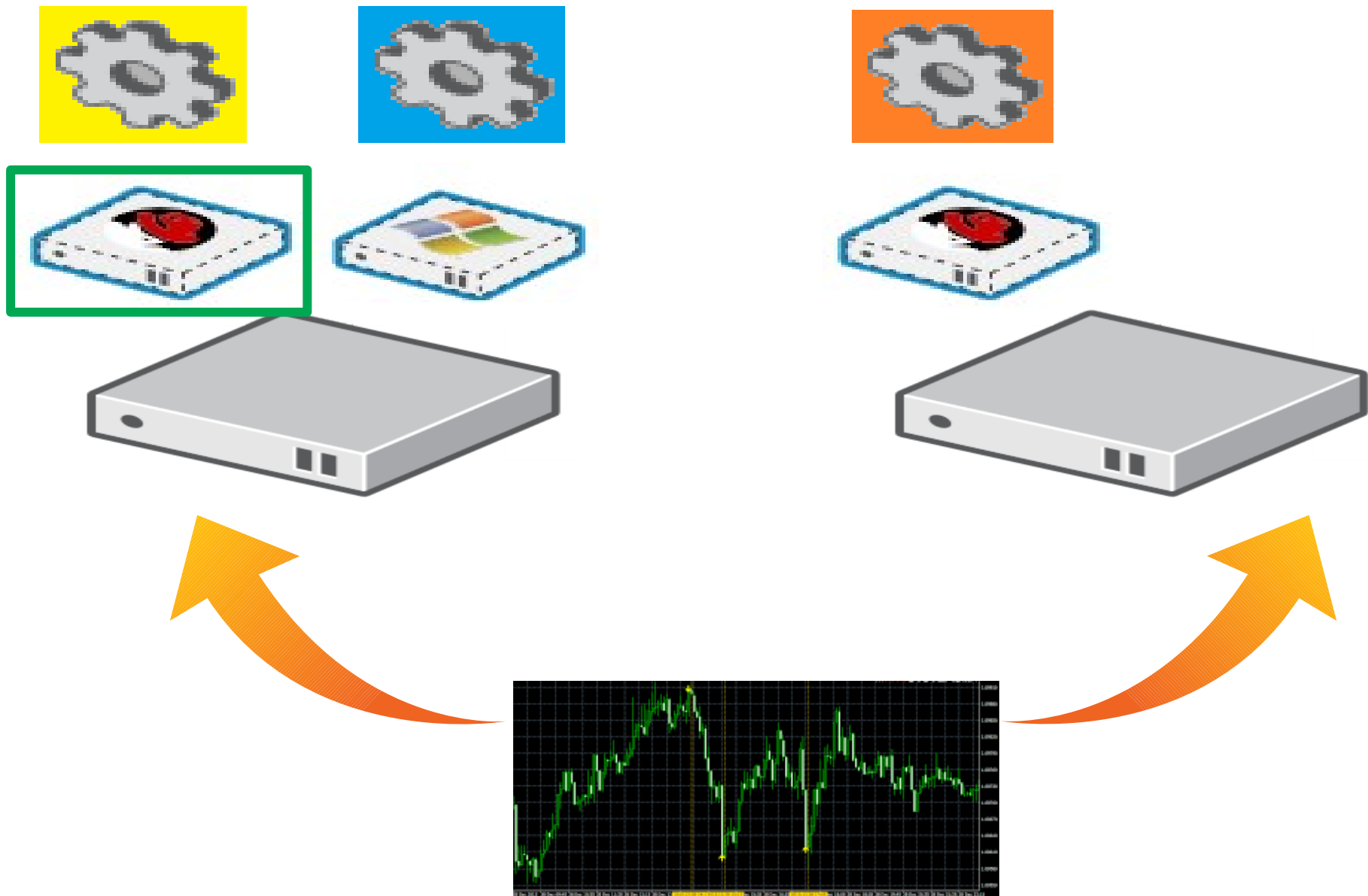


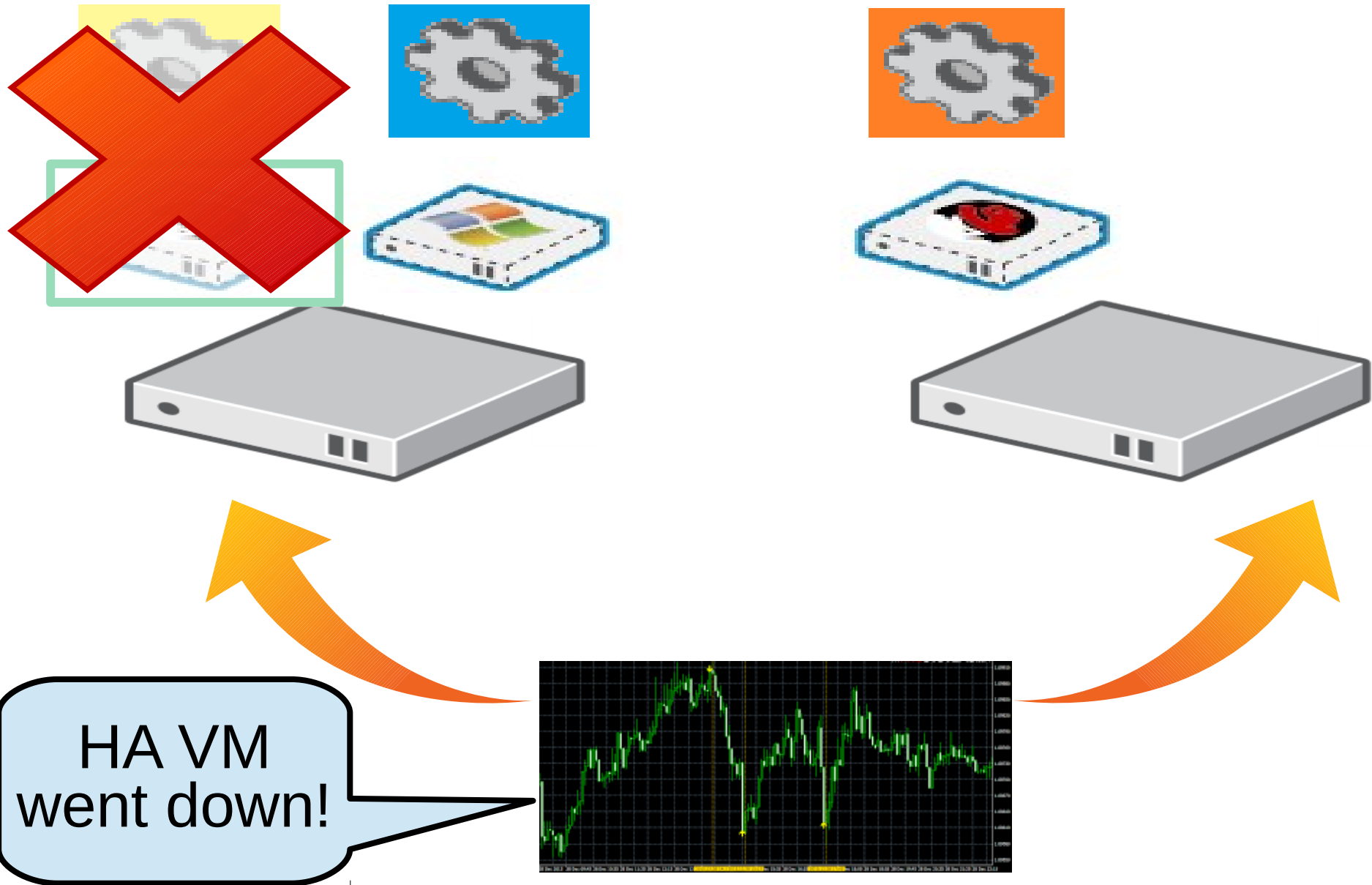
High Availability - VM-Level



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

Central Monitoring Unit



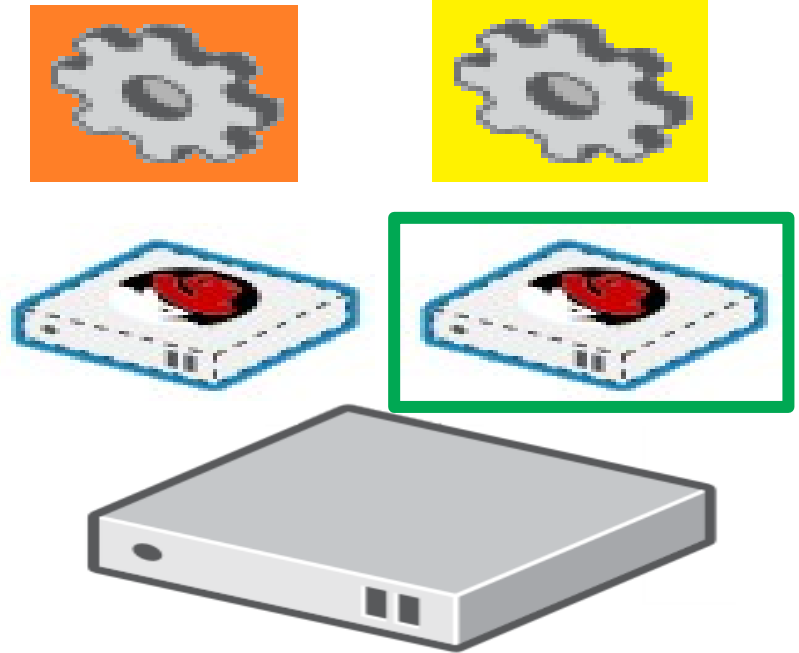




Automatic Restart – Not That Simple

What if:

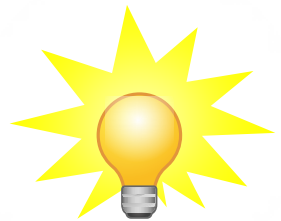
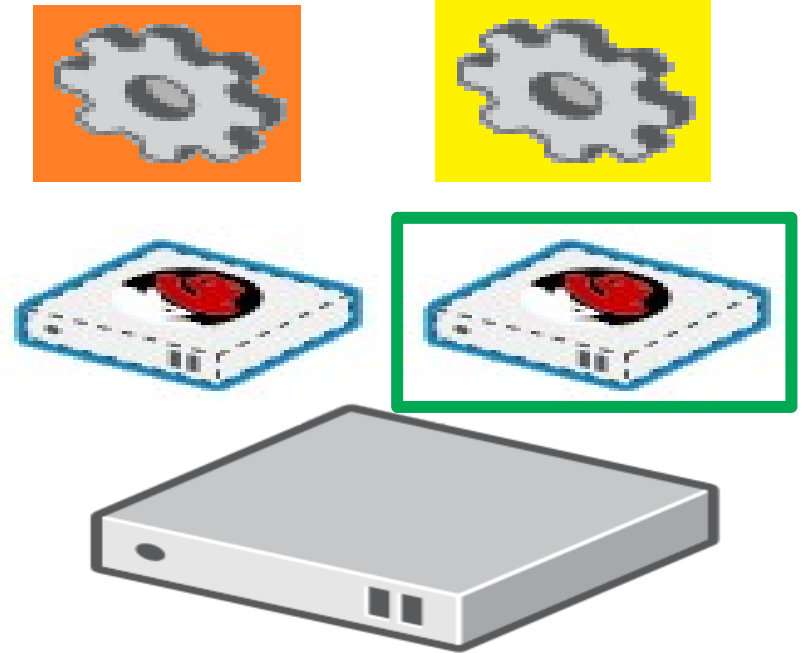
- Inaccessible resources
- VM is locked
- VM is being intentionally shut down



Restart
the VM

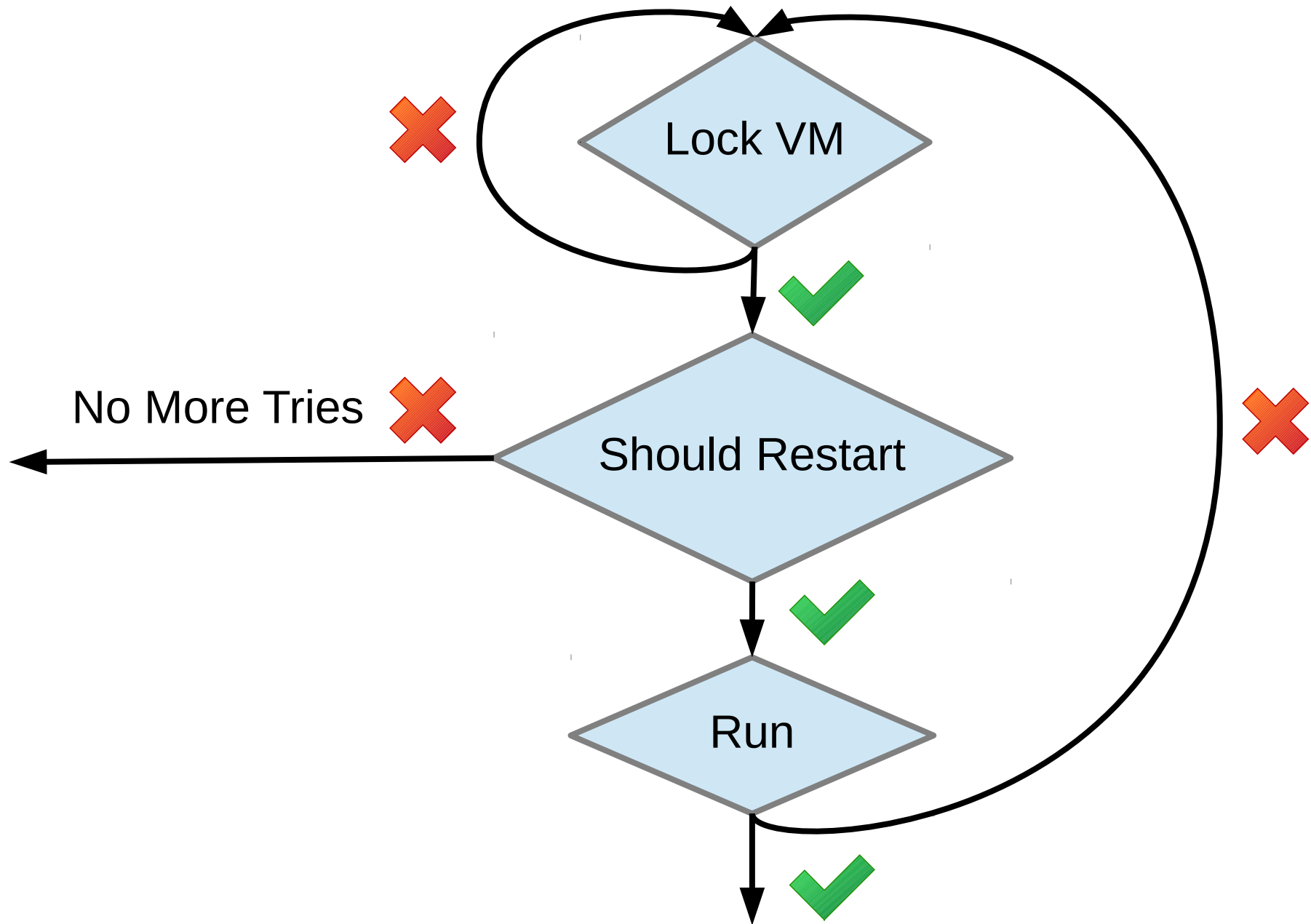
What if:

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

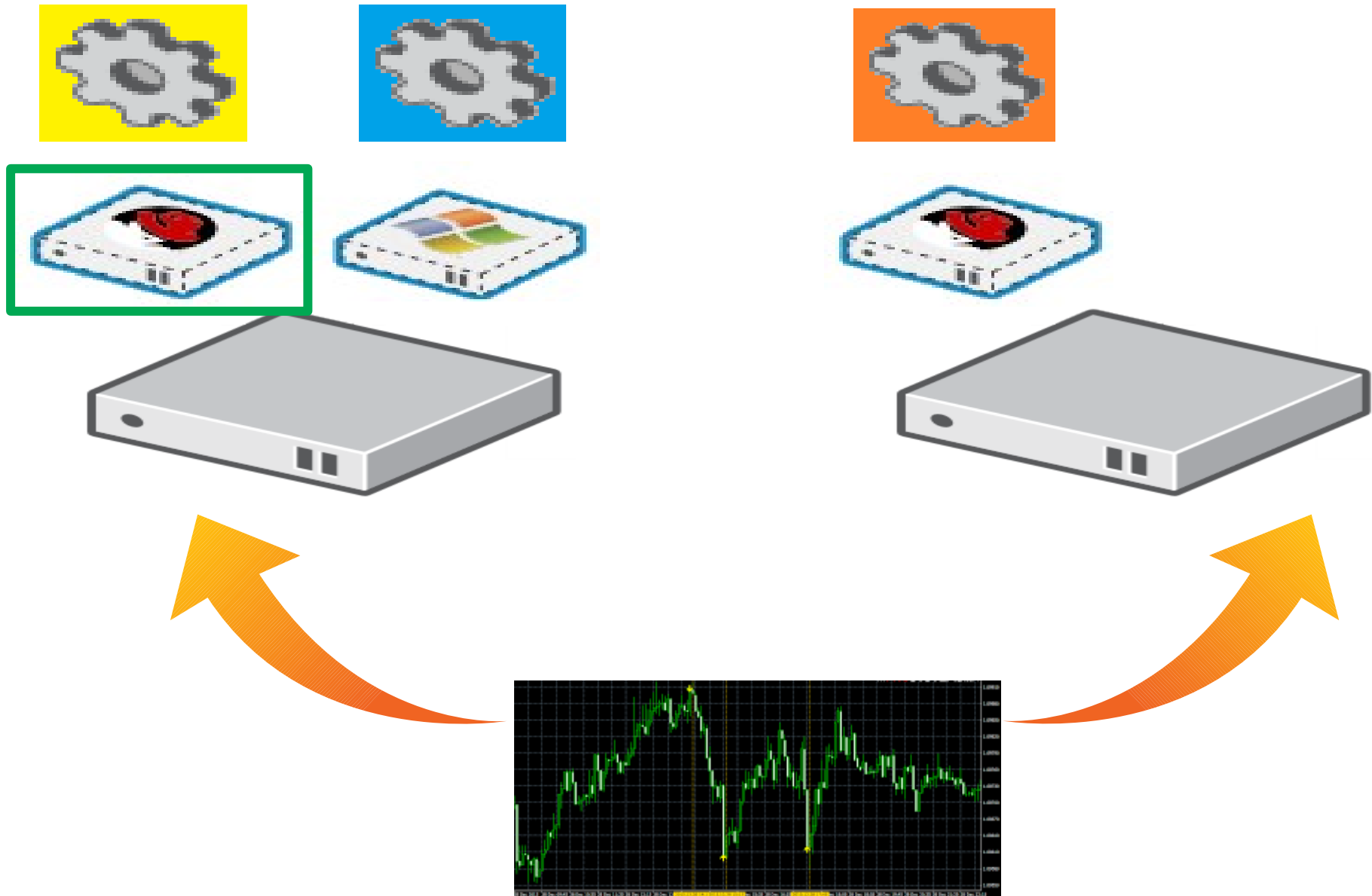


AutoStartVmsRunner

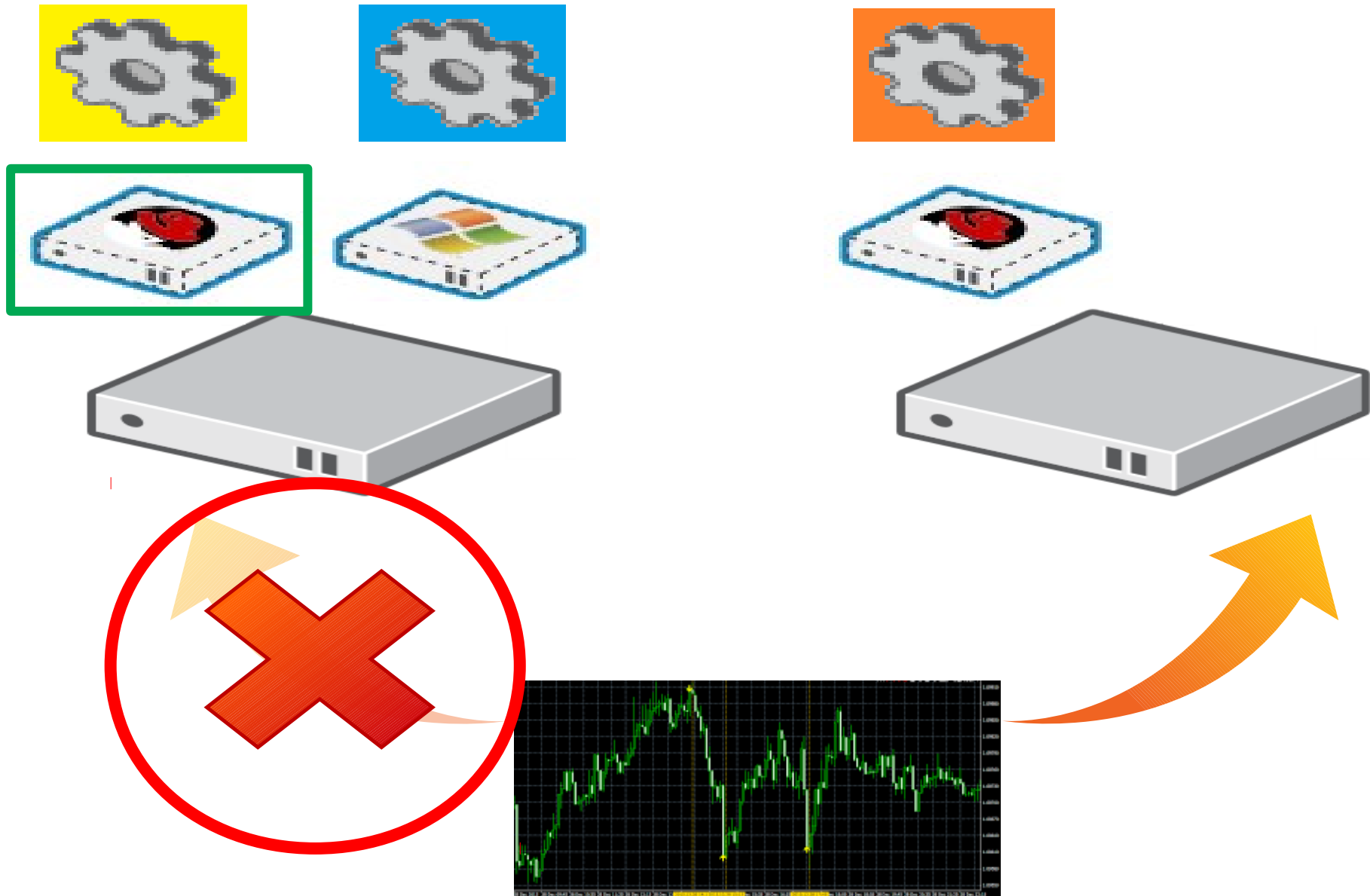
<https://github.com/oVirt/ovirt-engine/blob/master/backend/manager/modules/bll/src/main/java/org/ovirt/engine/core/bll/AutoStartVmsRunner.java>



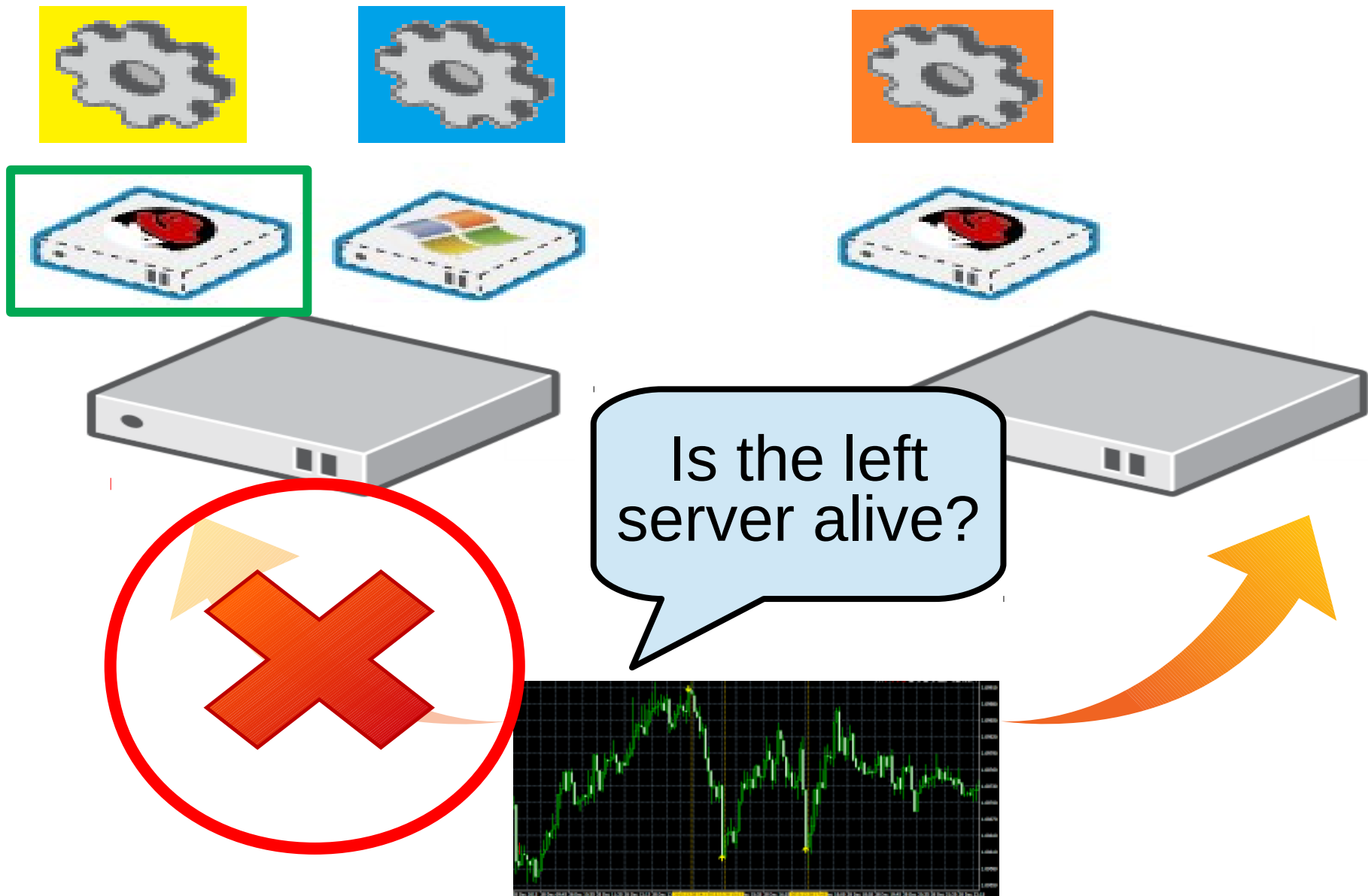
Fault Detection – Even More Complex



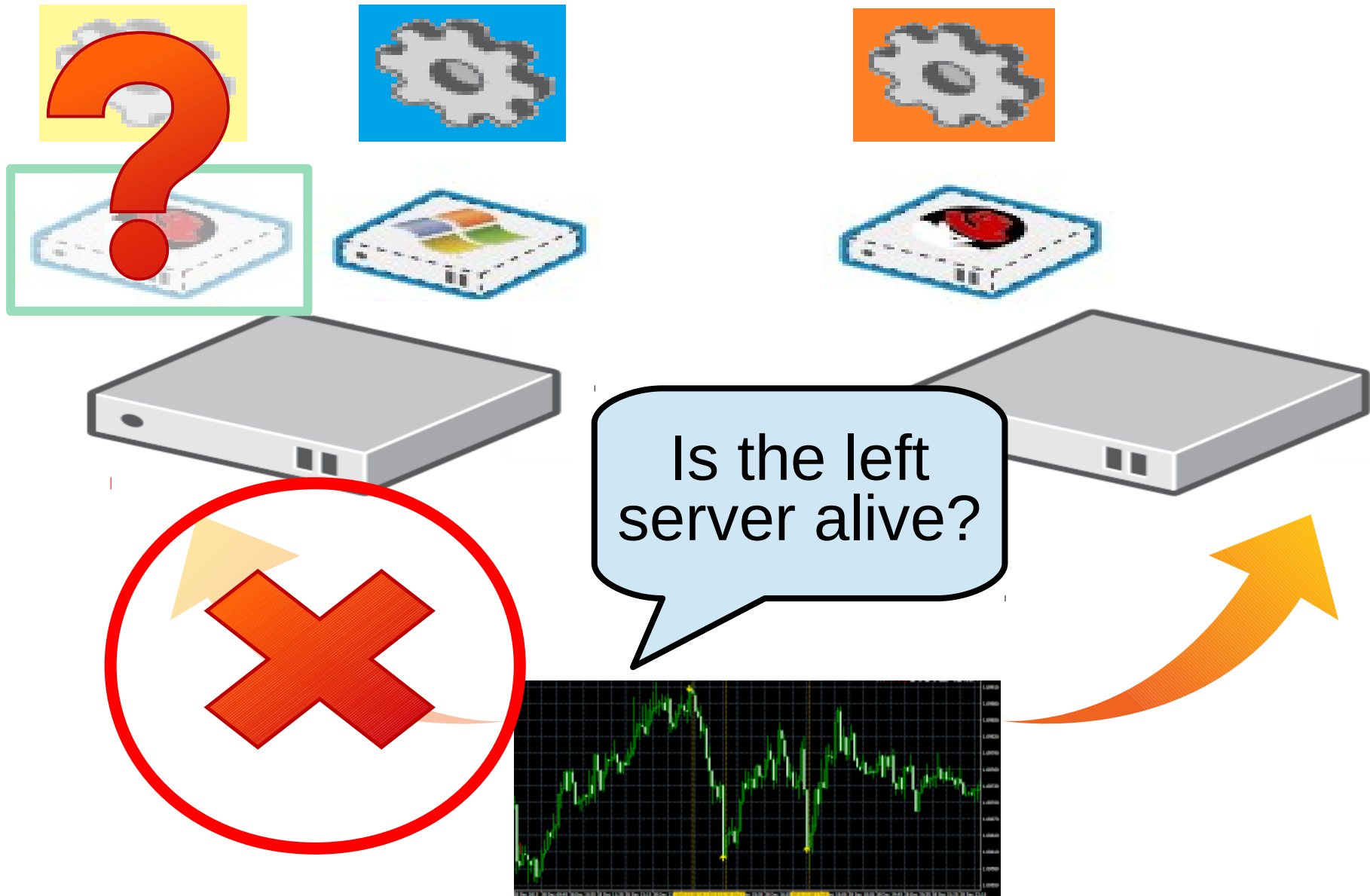
Fault Detection – Even More Complex

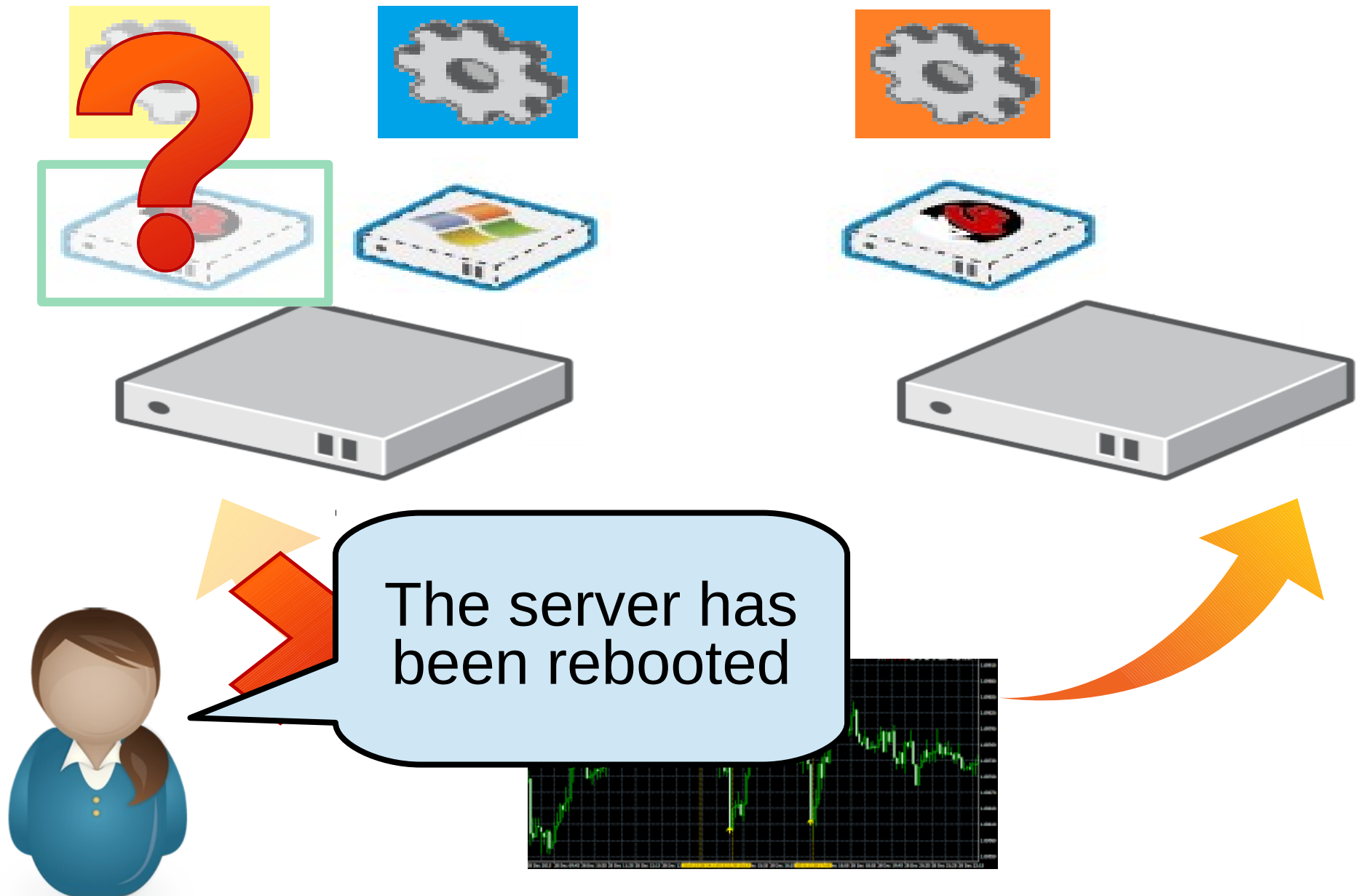


Fault Detection – Even More Complex

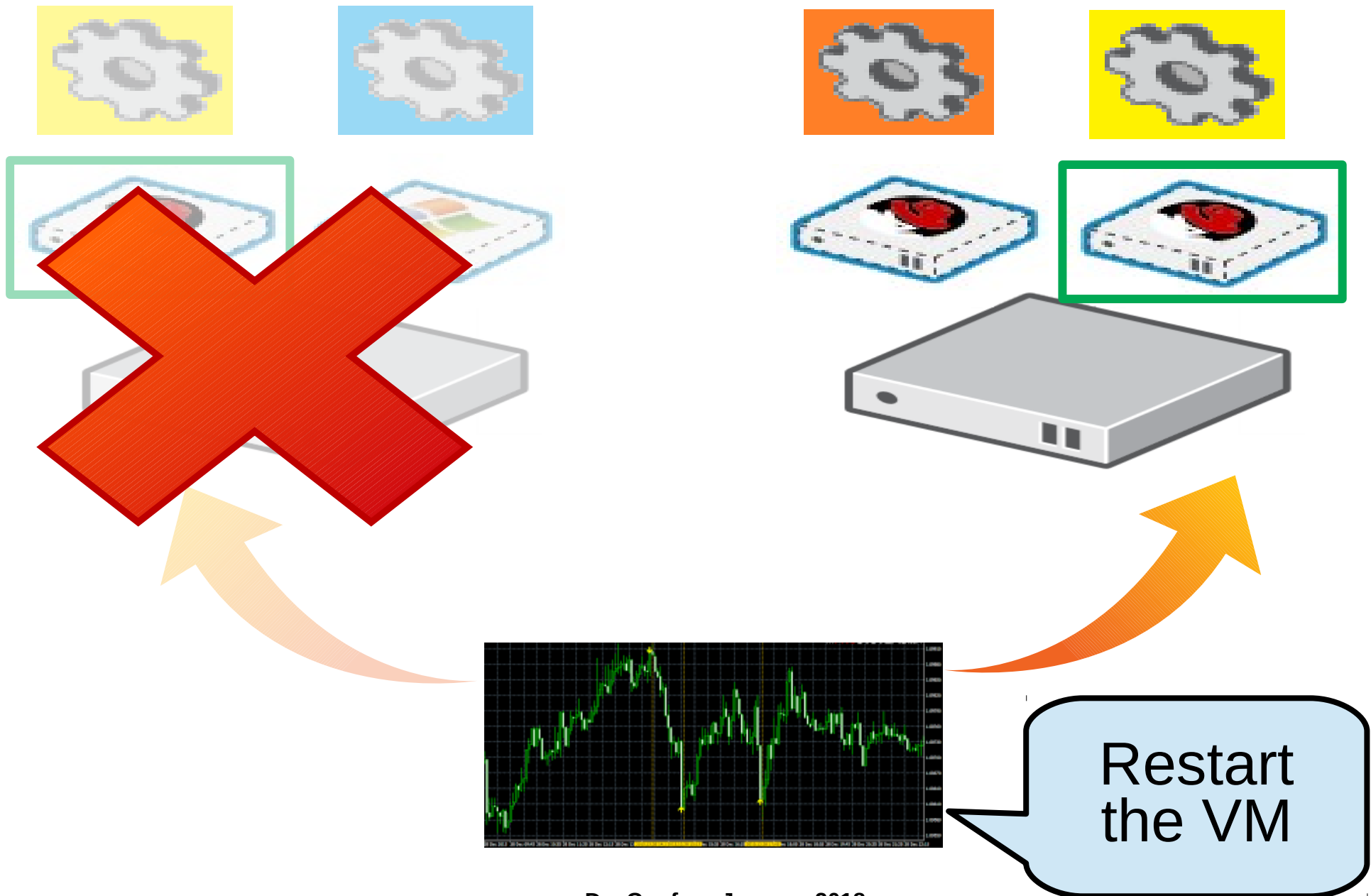


Fault Detection – Even More Complex





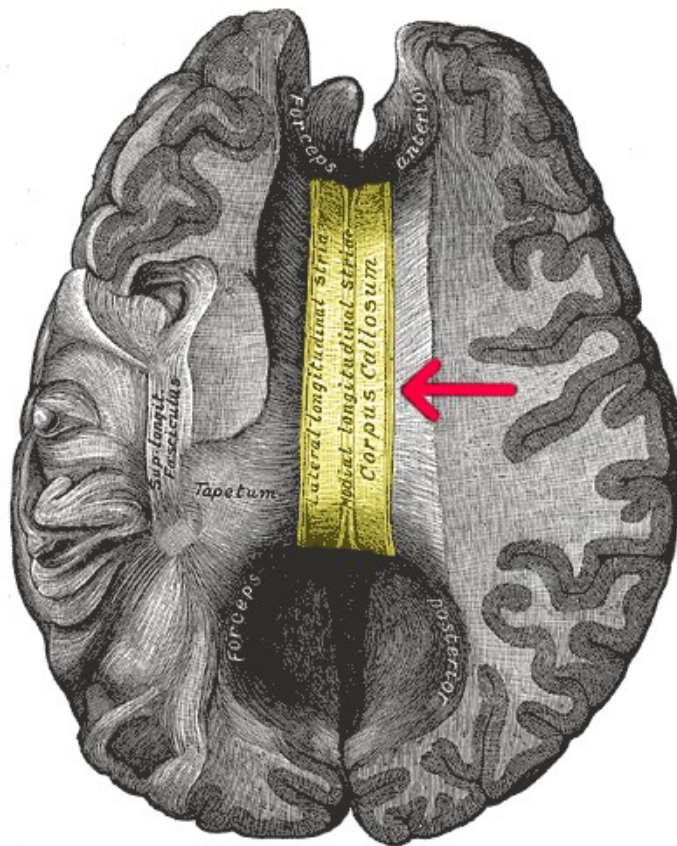
Fault Detection – Manual Confirmation



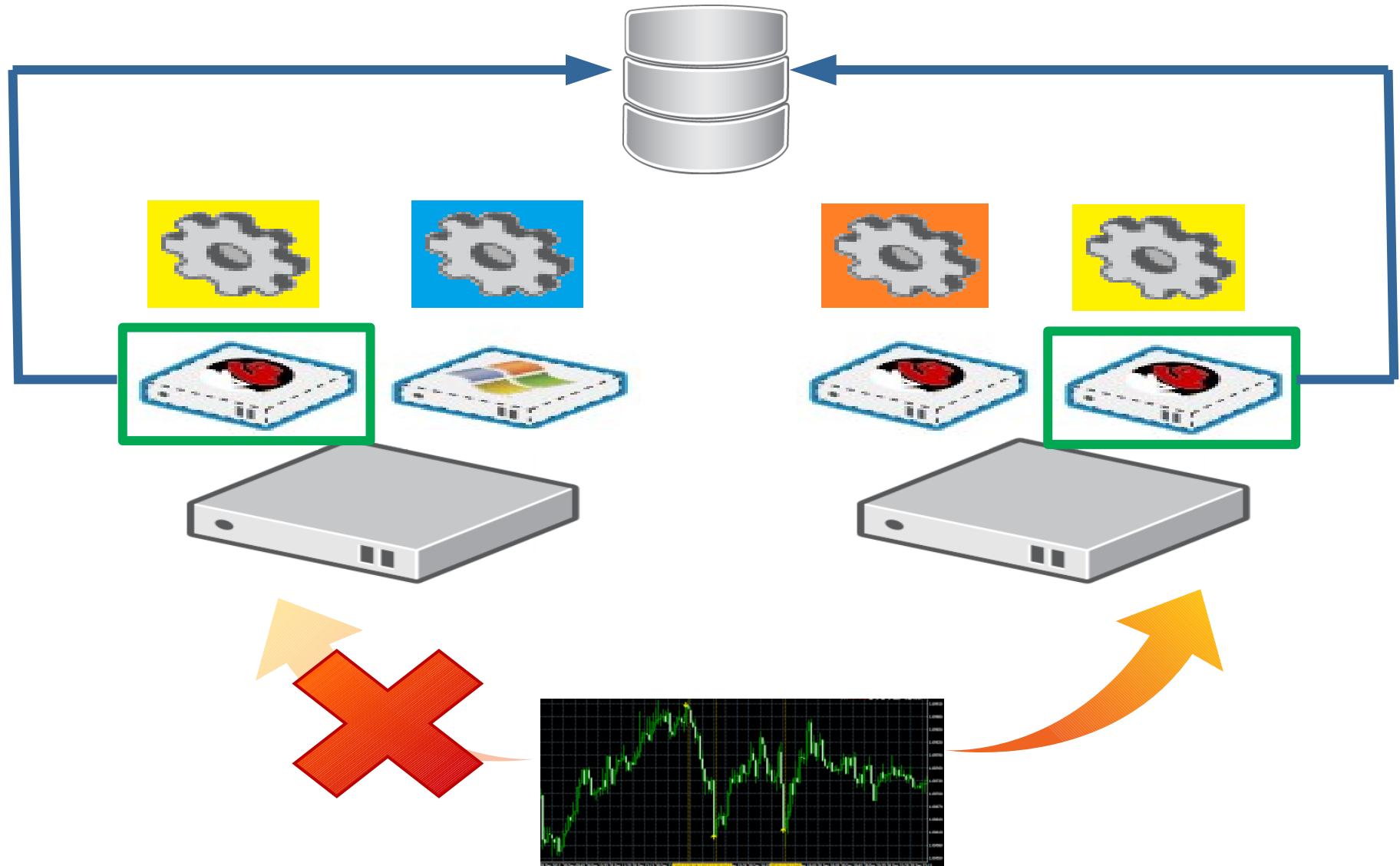
- Slow
- Error-prone
 - Mistakes may lead to a split-brain

Split Brain of Virtual Machines

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

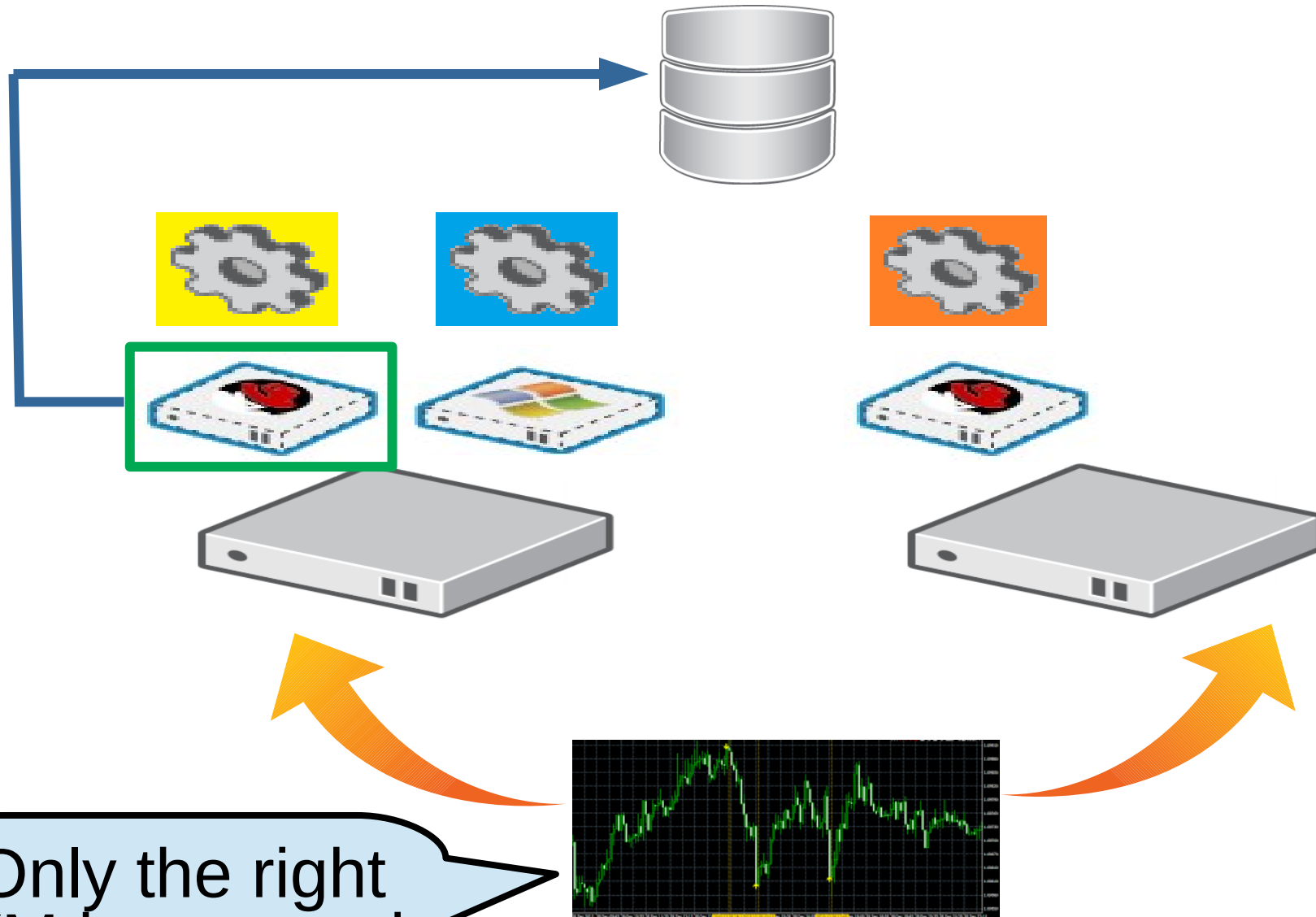


Split Brain Due to a False Confirmation



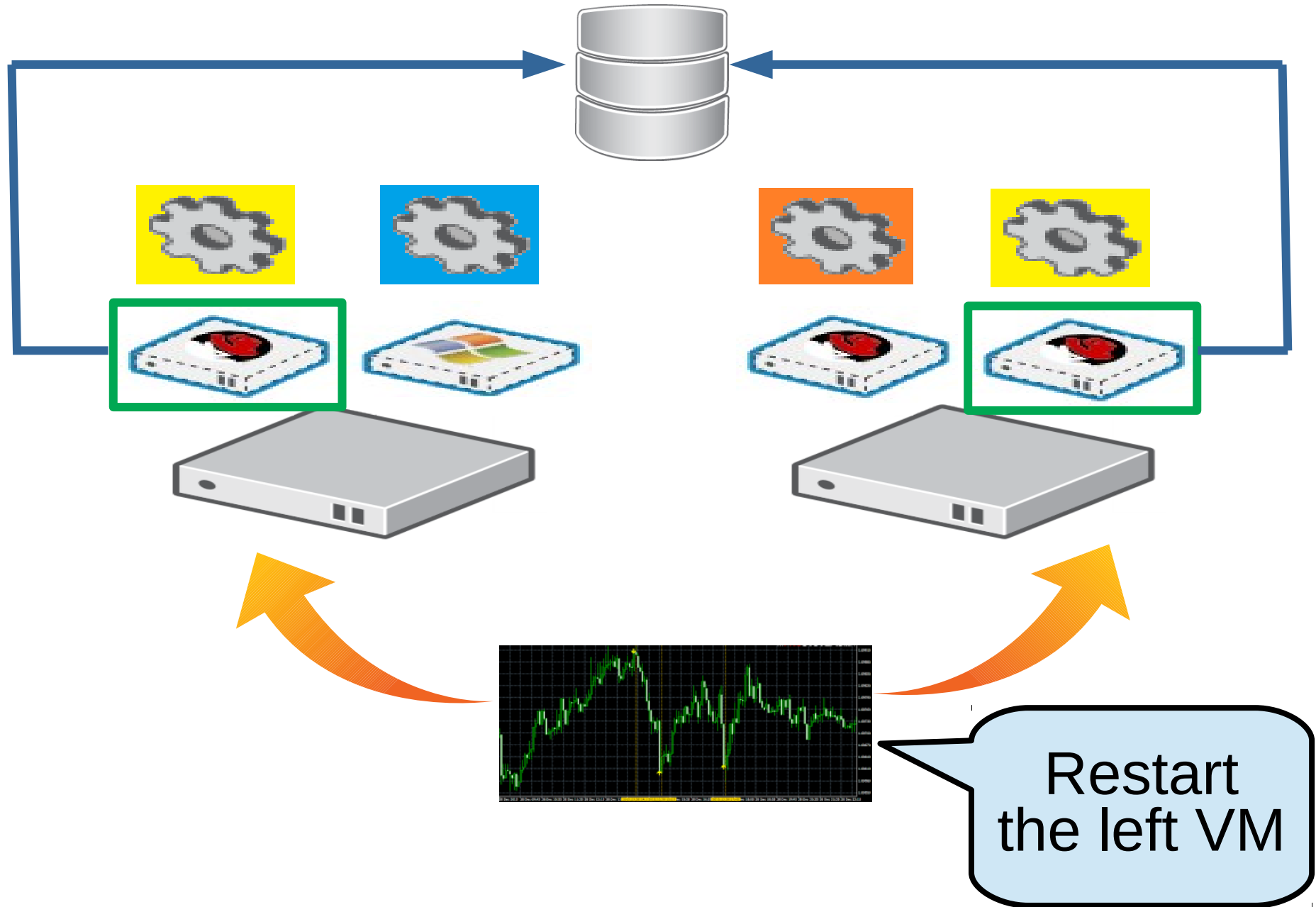
May lead to data corruption!

Split Brains May Happen Due to Bugs



Only the right
VM is reported
(on startup)

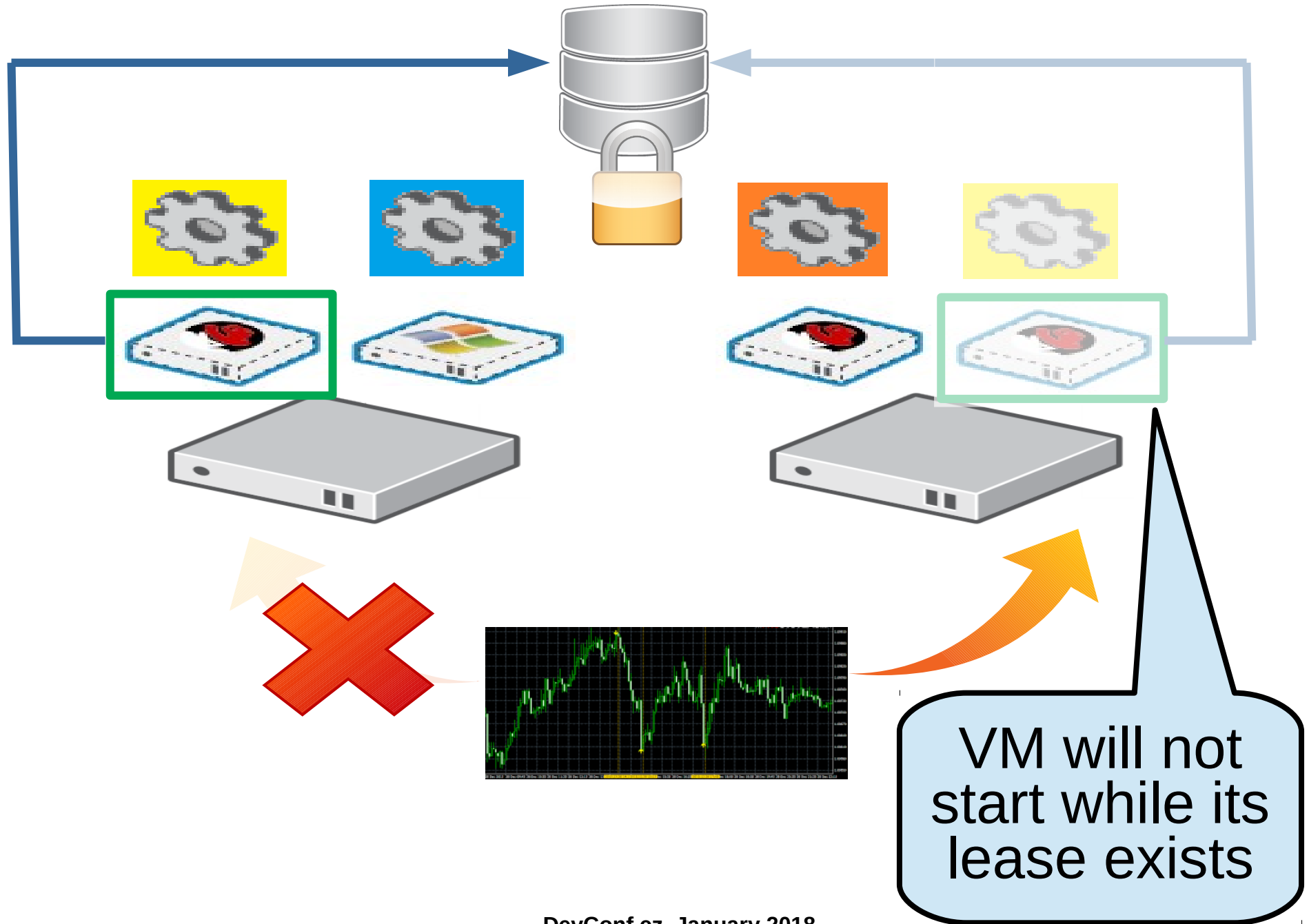
Split Brains May Happen Due to Bugs



VM Leases: Our Solution to Split Brains



VM Leases: Our Solution to Split Brains



Edit Virtual Machine

General

System

Initial Run

Console

Host

High Availability

Resource Allocation

Boot Options

Random Generator

Custom Properties

Icon

Foreman/Satellite

Affinity Labels

Cluster

Template

Operating System

Instance Type

Optimized for

☒ Highly Available

Target Storage Domain for VM Lease

Resume Behavior

Priority for Run/Migration queue:

Priority

Watchdog

Watchdog Model

Watchdog Action

Default

Data Center: Default

Blank | (0)

Debian 7

Custom

Server

Default

KILL

Low

No-Watchdog

none

☒ Highly Available



Target Storage Domain for VM Lease

Default

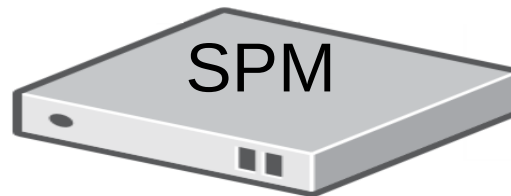


Resume Behavior

KILL



VM Lease Creation

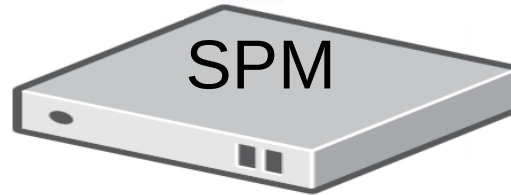


“Create VM Lease for
VM X in storage domain Y”



VM Lease Creation

“Create a Lease X in
lockspace Y”

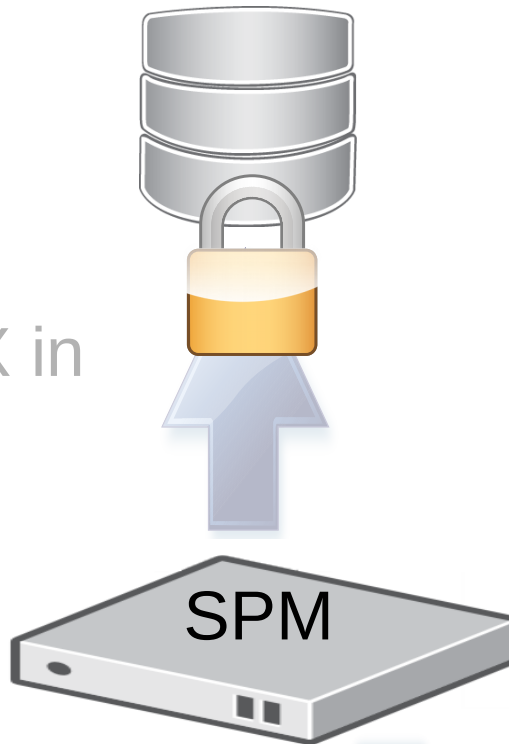


“Create VM Lease for
VM X in storage domain Y”



VM Lease Creation

“Create a Lease X in lockspace Y”

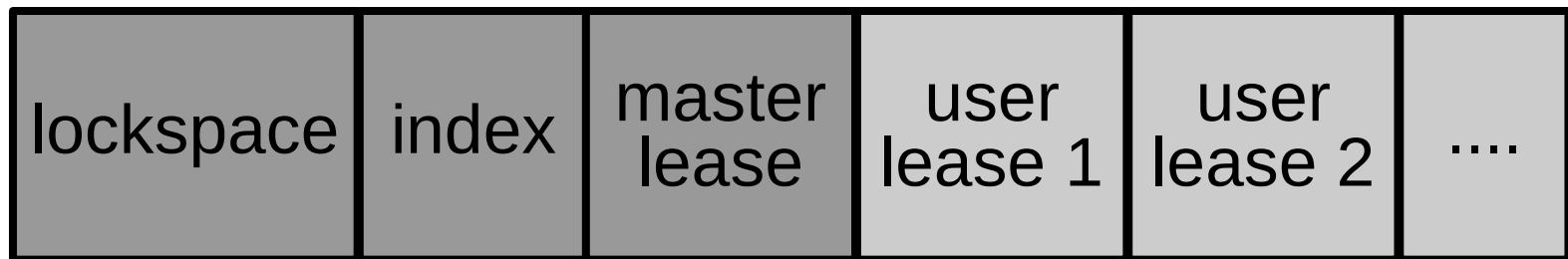


“Create VM Lease for VM X in storage domain Y”

“Path P to xleases volume and Lease offset O”



- 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>
```

oVirt

Acquires the Lease
using Sanlock

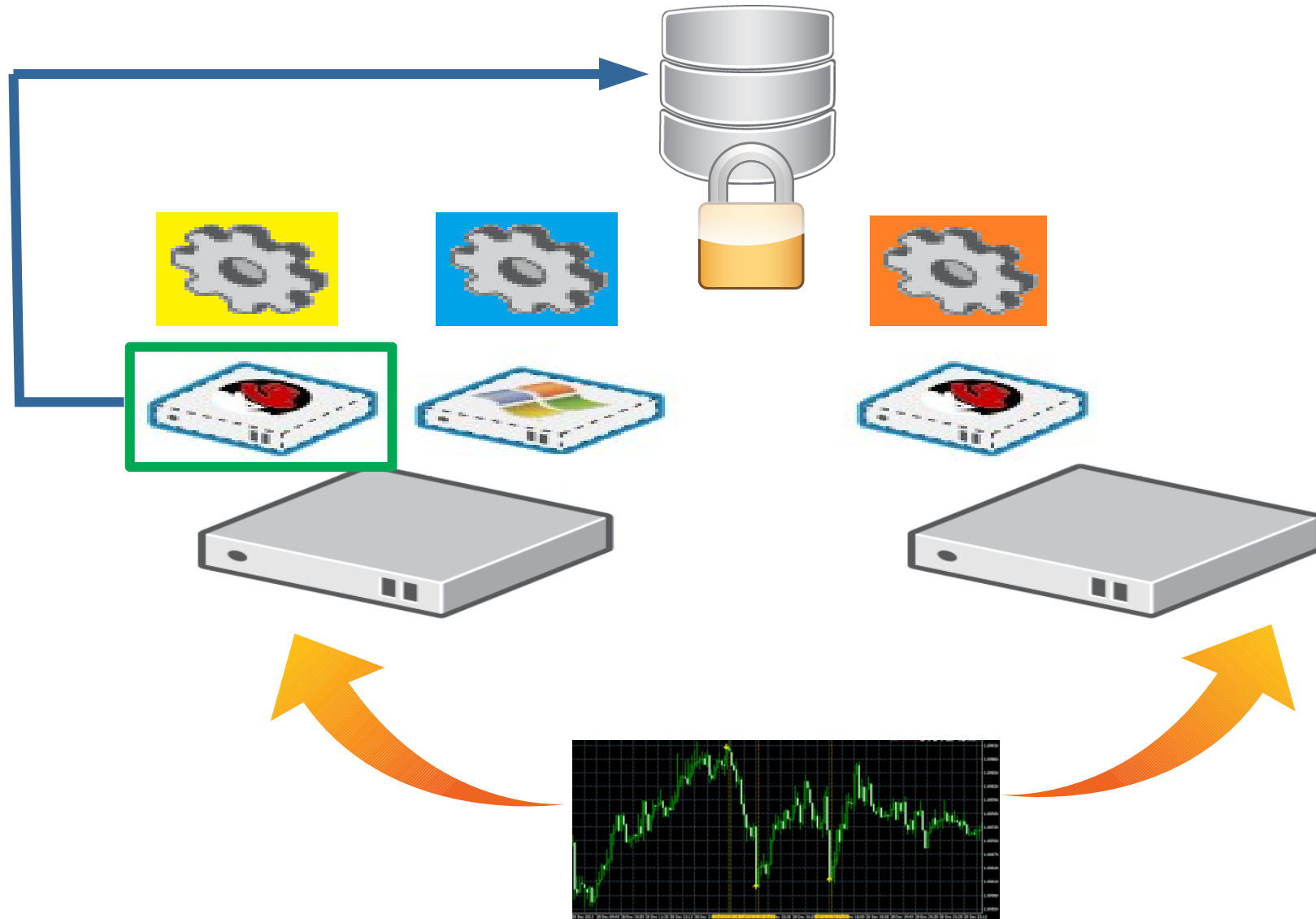


Domain XML
with Lease

Lease



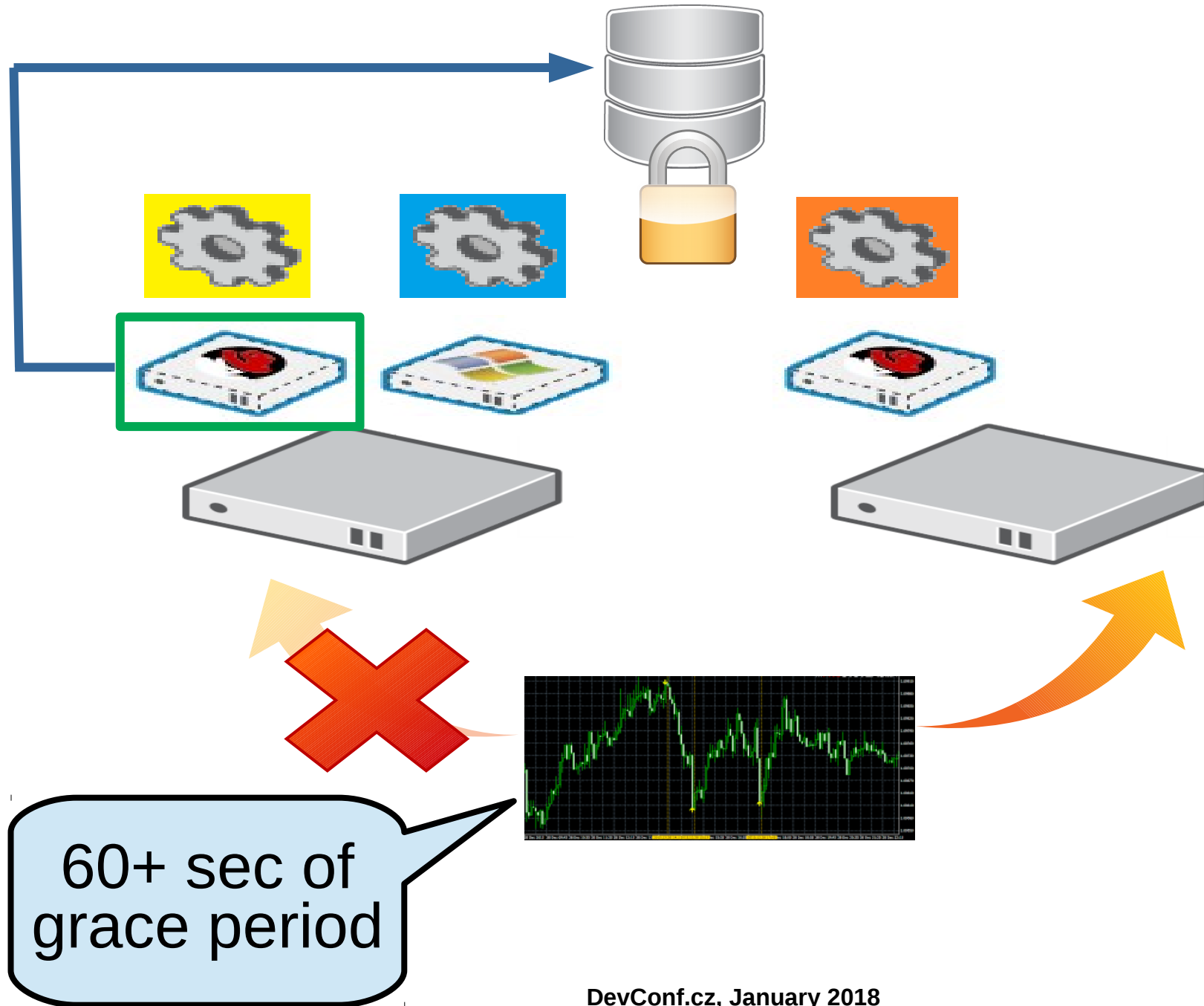
oVirt Non-Responsive Host Treatment



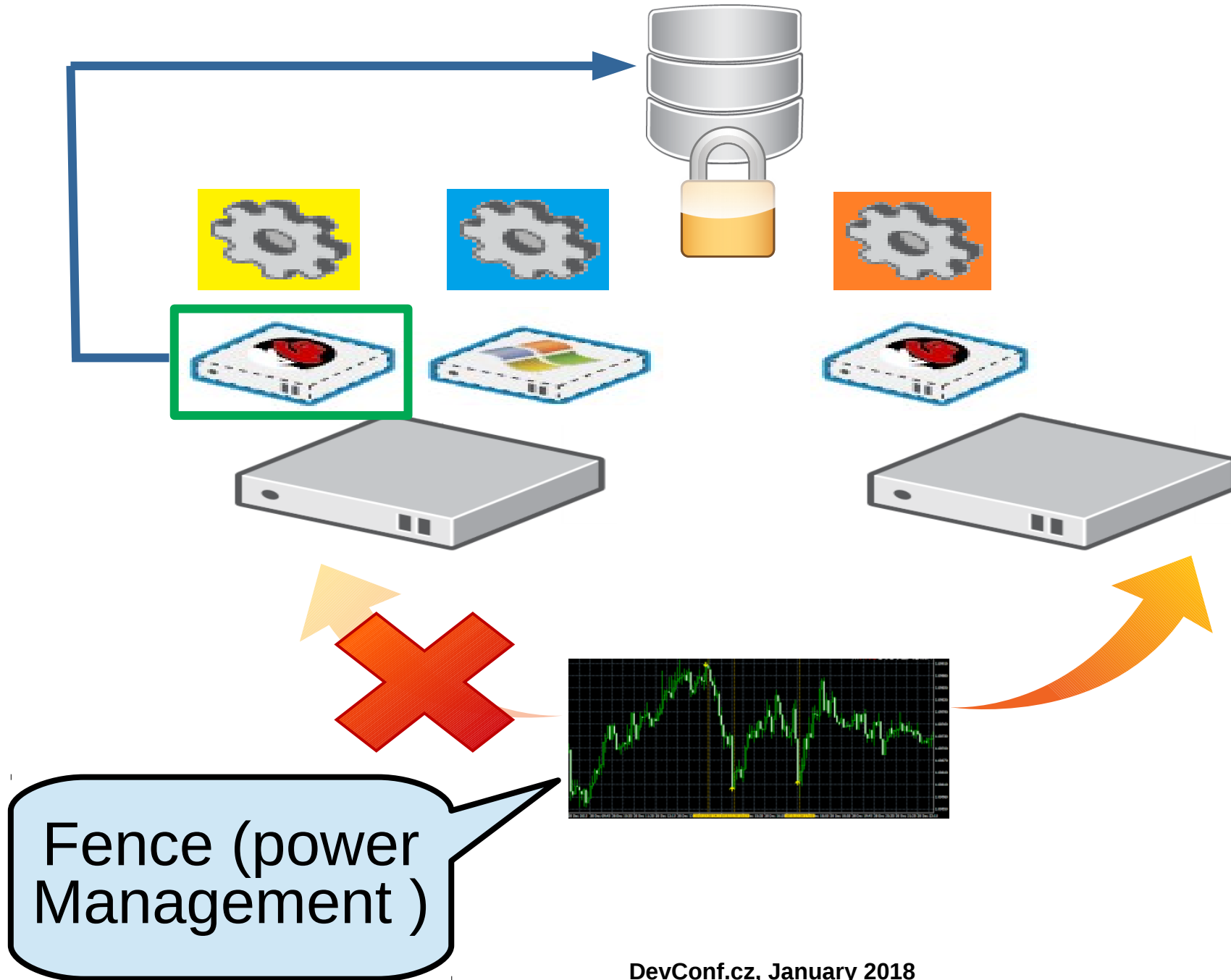
oVirt Non-Responsive Host Treatment



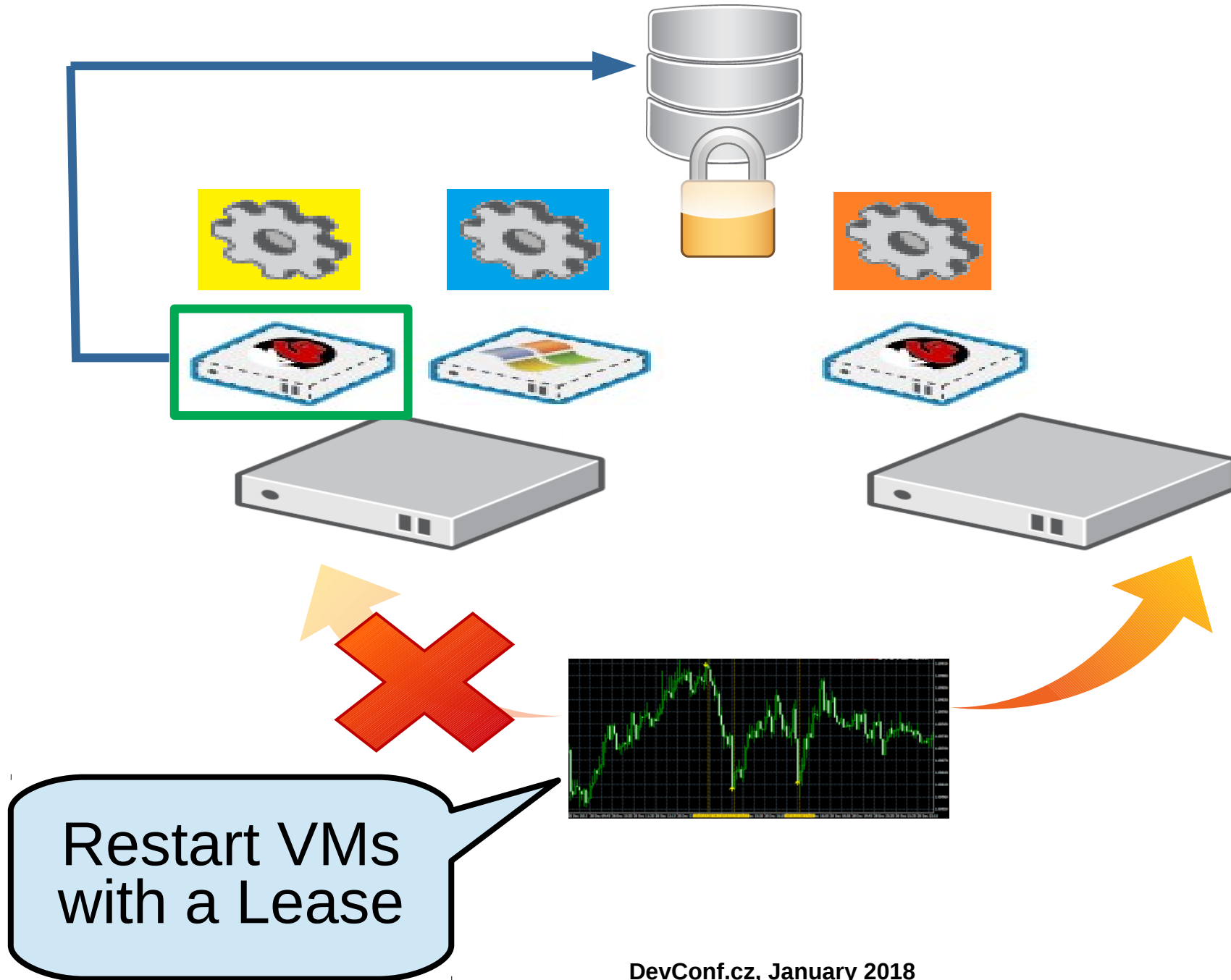
oVirt Non-Responsive Host Treatment



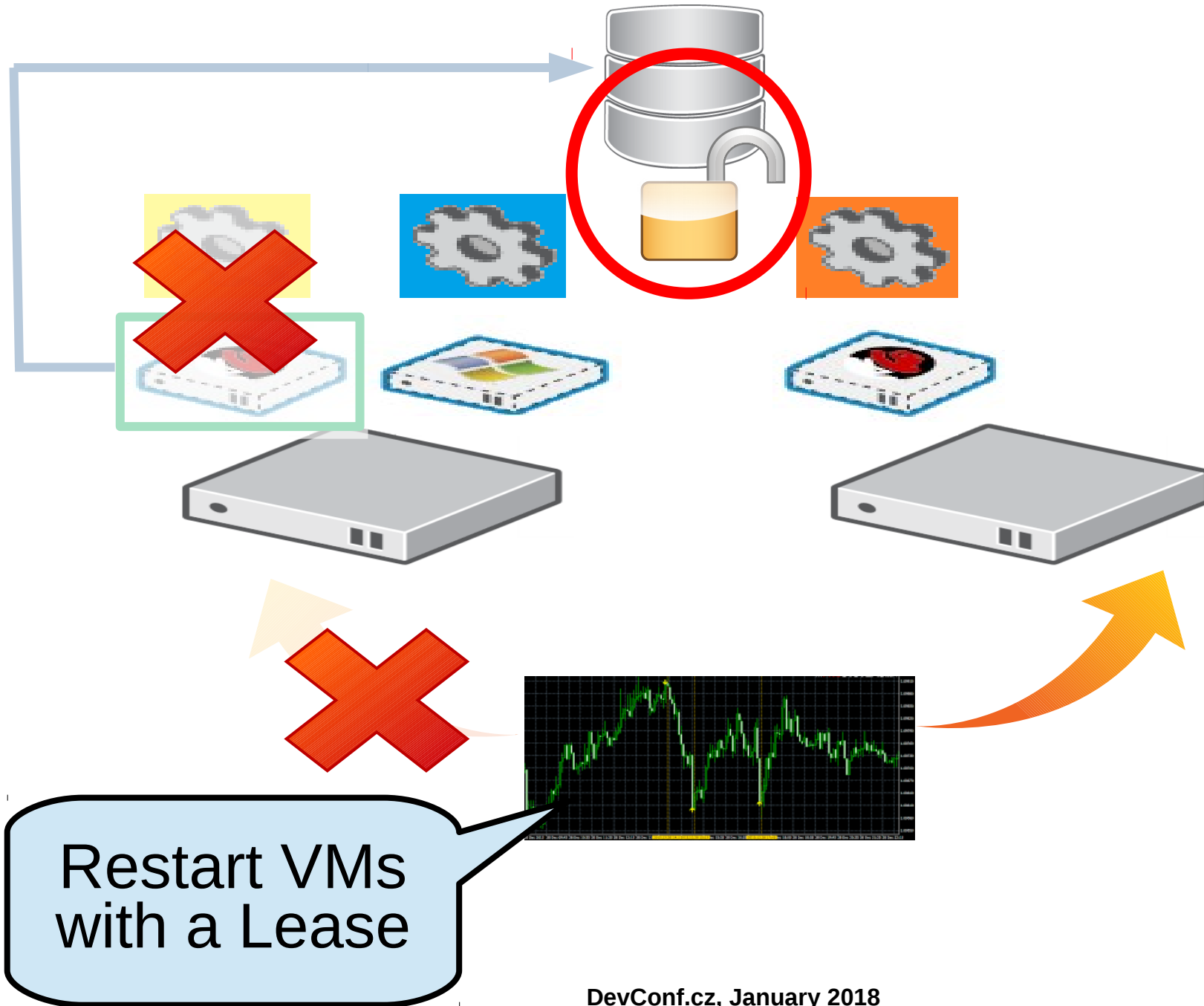
oVirt Non-Responsive Host Treatment



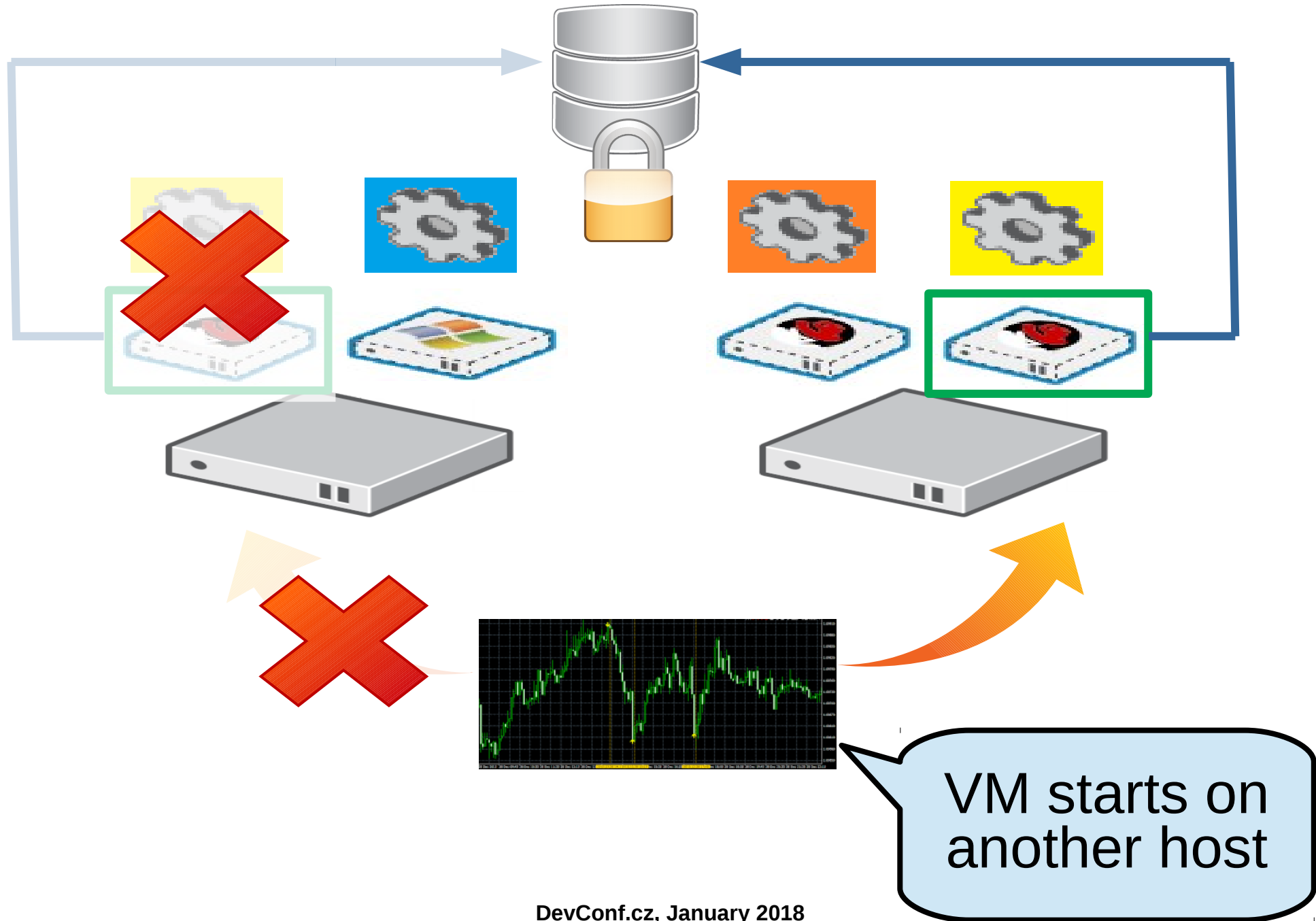
oVirt Non-Responsive Host Treatment



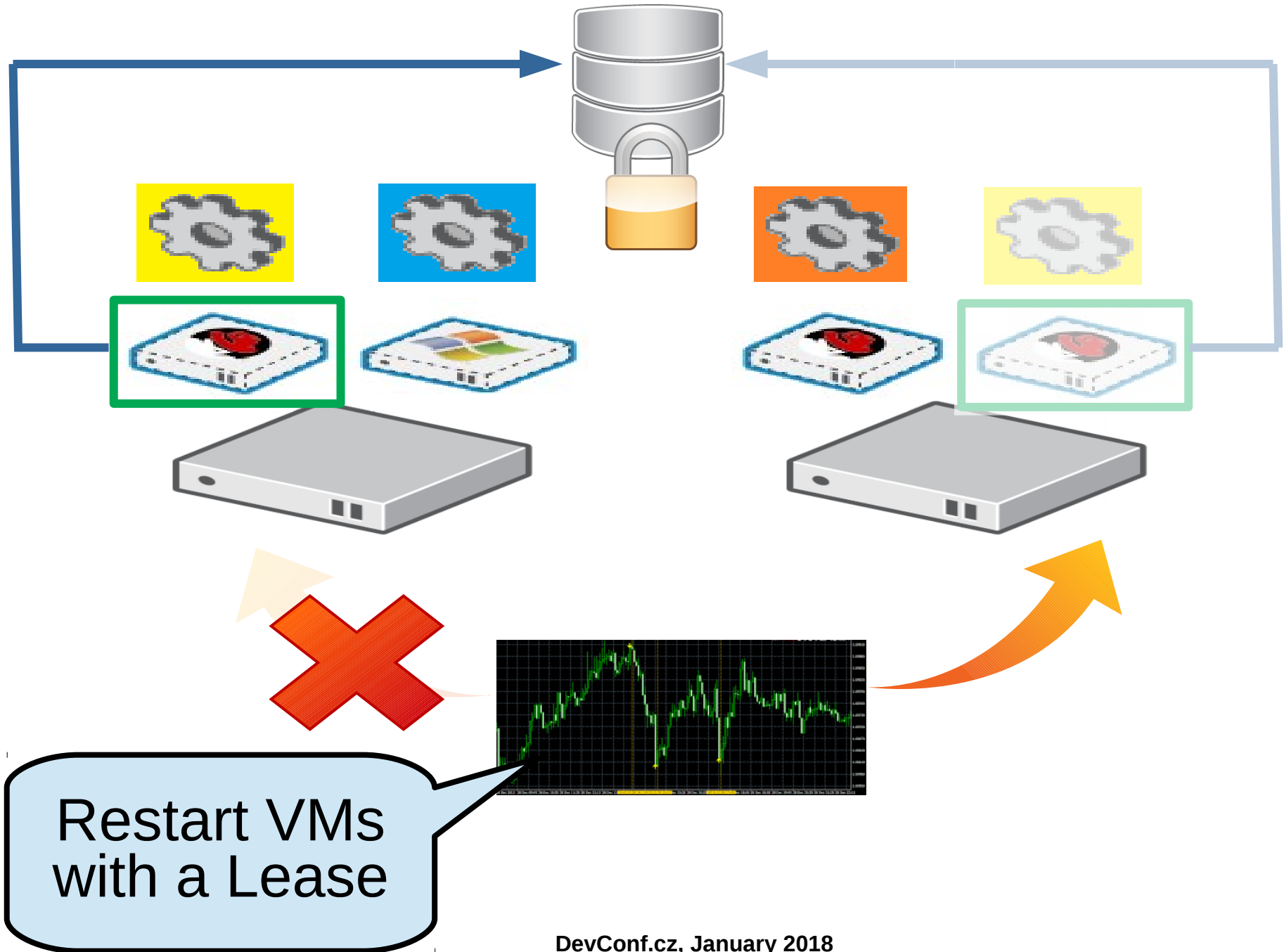
(1) Non-Responsive Host + VM is Down



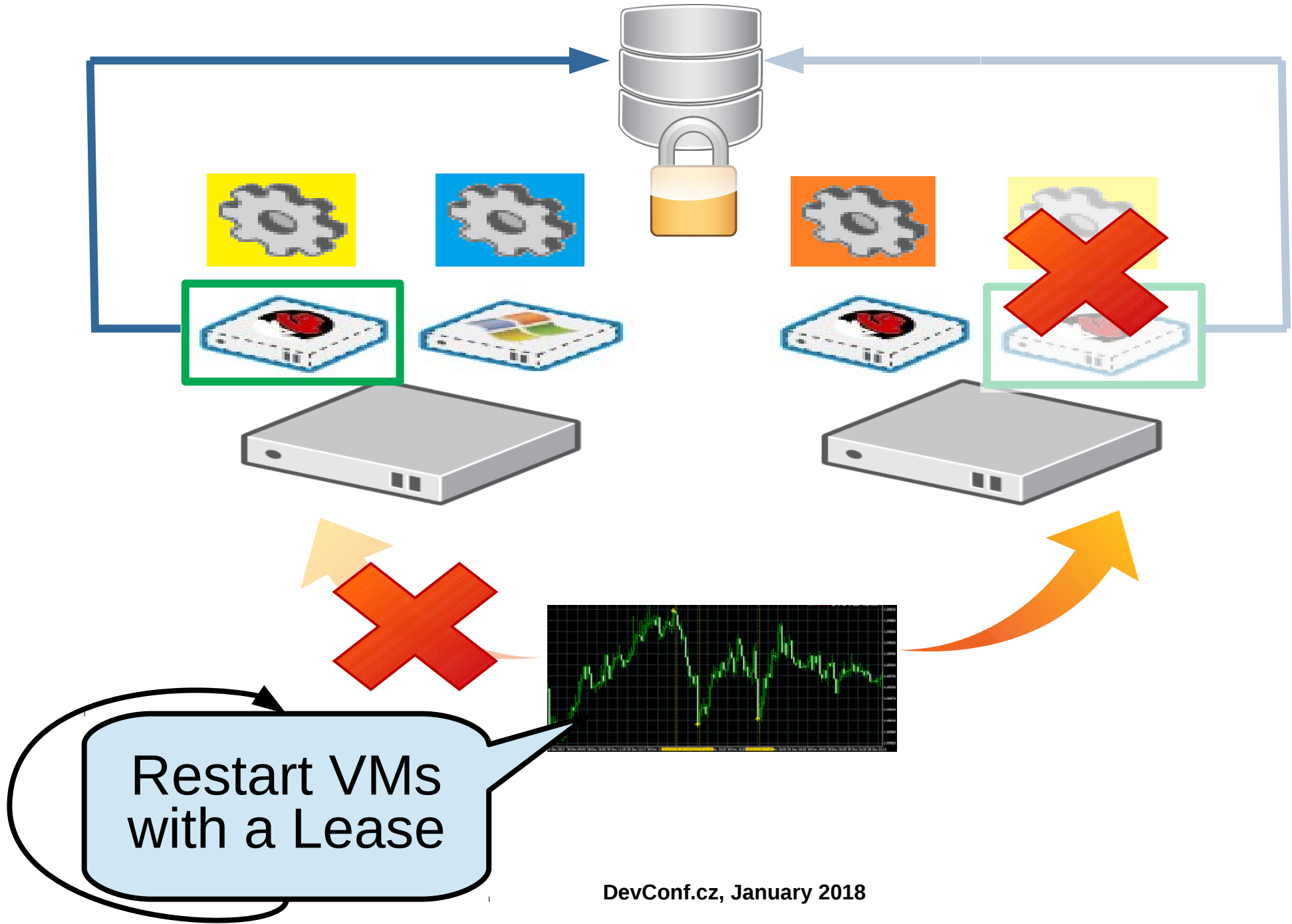
(1) Non-Responsive Host + VM is Down



(2) Non-Responsive Host + VM is UP



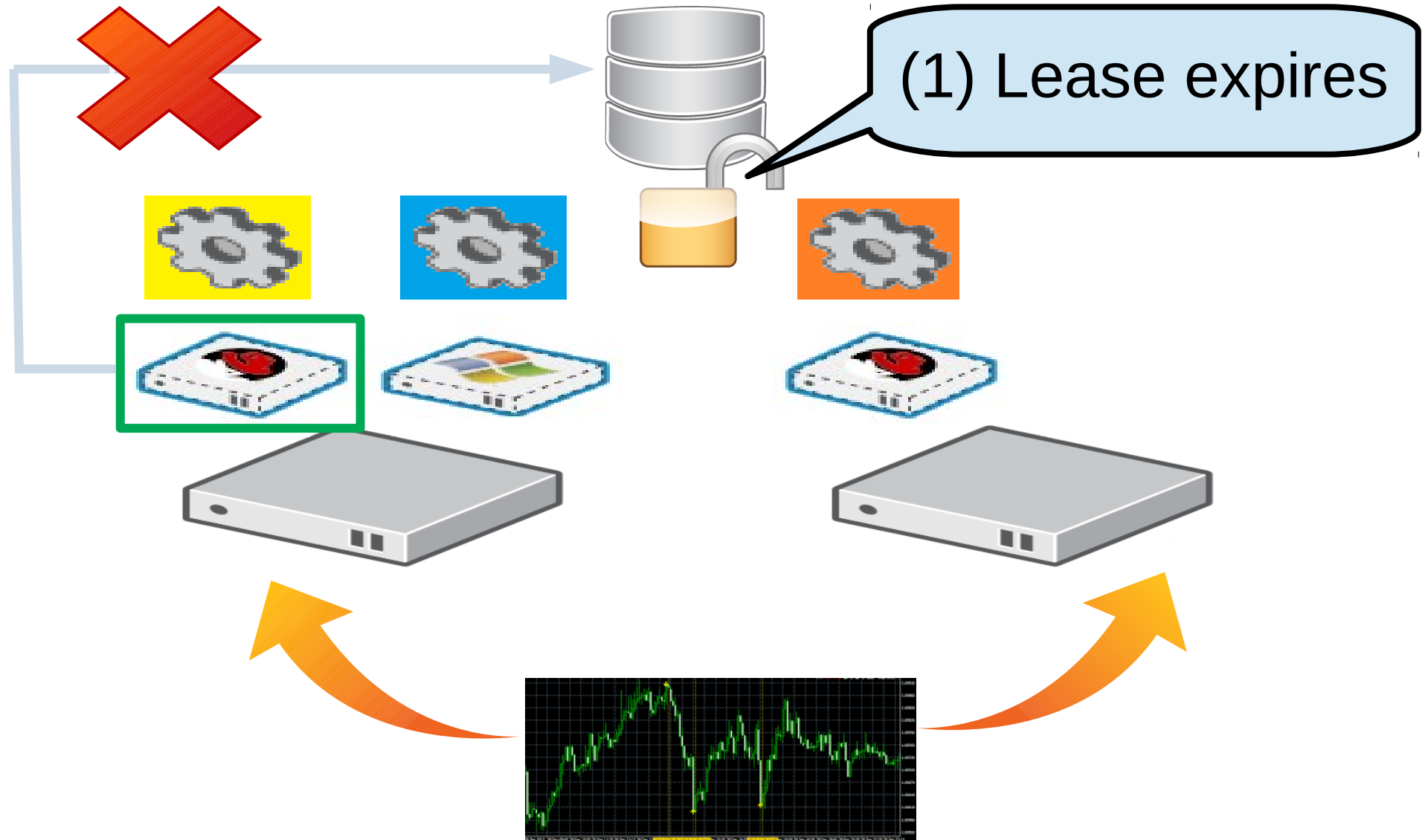
(2) Non-Responsive Host + VM is UP



Disconnection From Storage Device

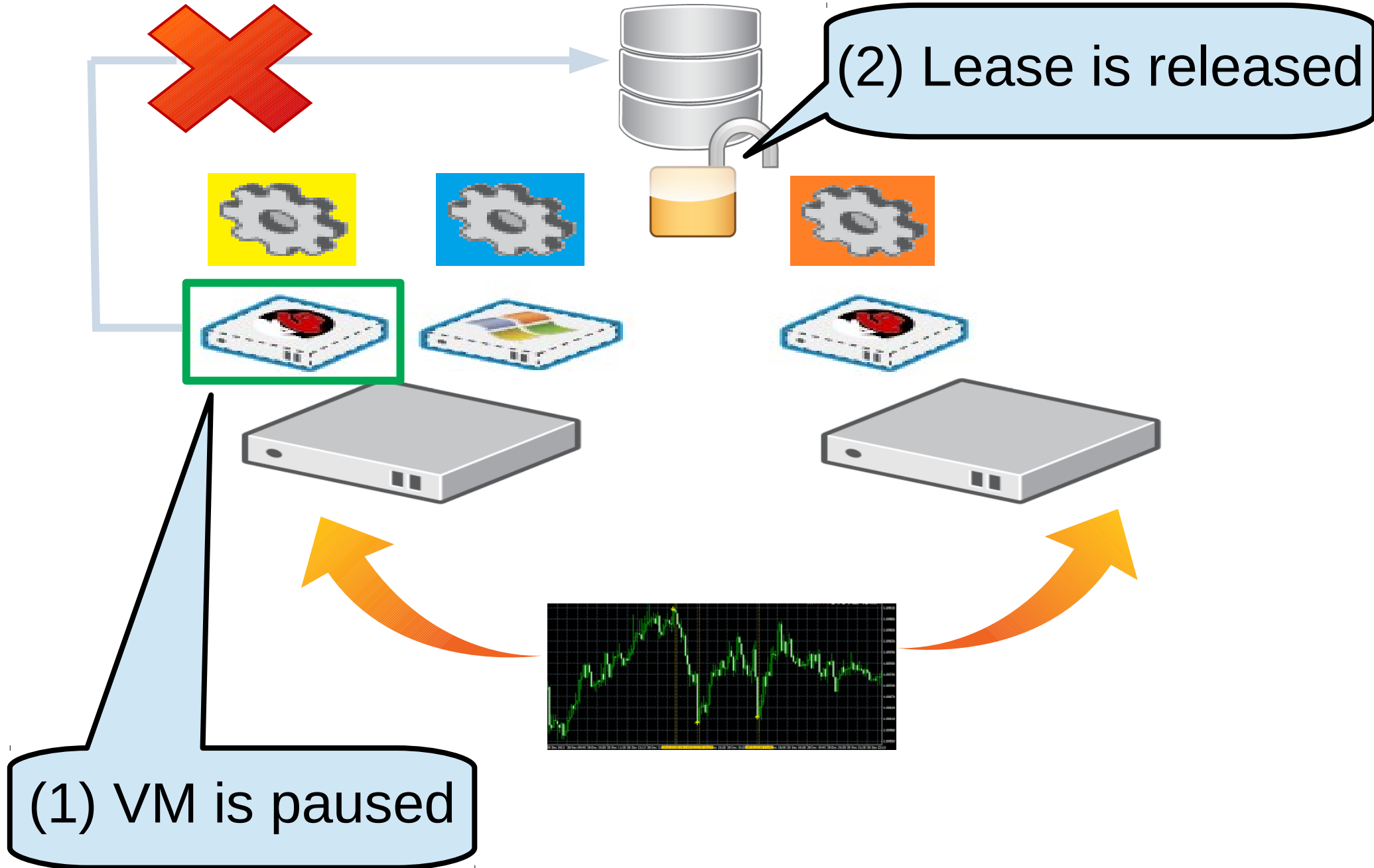


Disconnection From Storage Device (1)





Disconnection From Storage Device (2)



- 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 enhancements:
 - 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](irc://irc.oftc.net/#ovirt)