# Enhancing High Availability in the Context of OpenStack

Qiming Teng

tengqim@cn.ibm.com
IBM Research

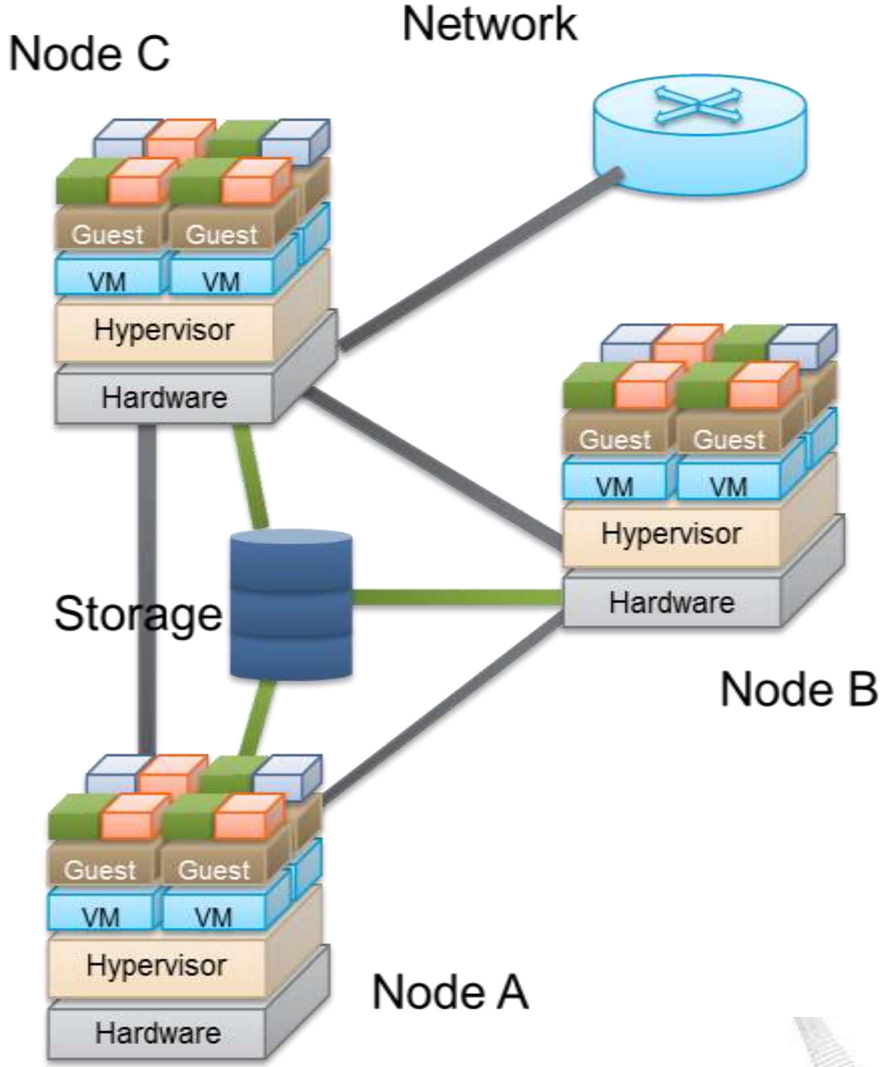
OpenStack Summit
**May 12-16, 2014**
Atlanta, Georgia

# Agenda

- High Availability (HA) Overview
- Four Types of HA in OpenStack
- OpenStack HA
- VM/Application HA Options
  - VM/App HA Orchestrated
  - Open Questions
- HA as a Service?

# High Availability Overview

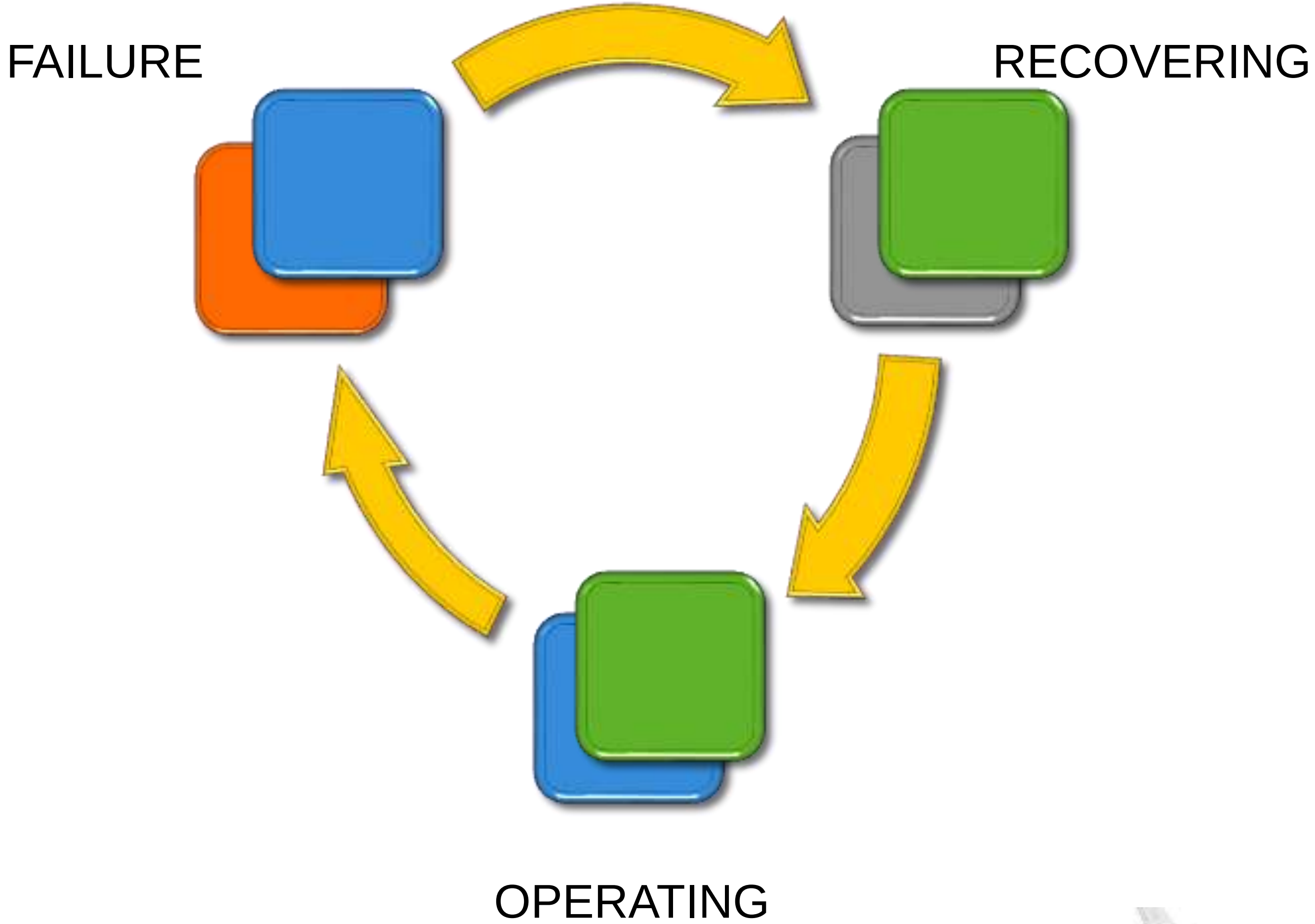
- Why HA?
  - Single system
    - Hardware failures
    - Hypervisor defects
    - OS (host/guest) crashes
    - Application bugs
  - In cloud
    - Shared, virtualized storage
    - Shared, virtualized network



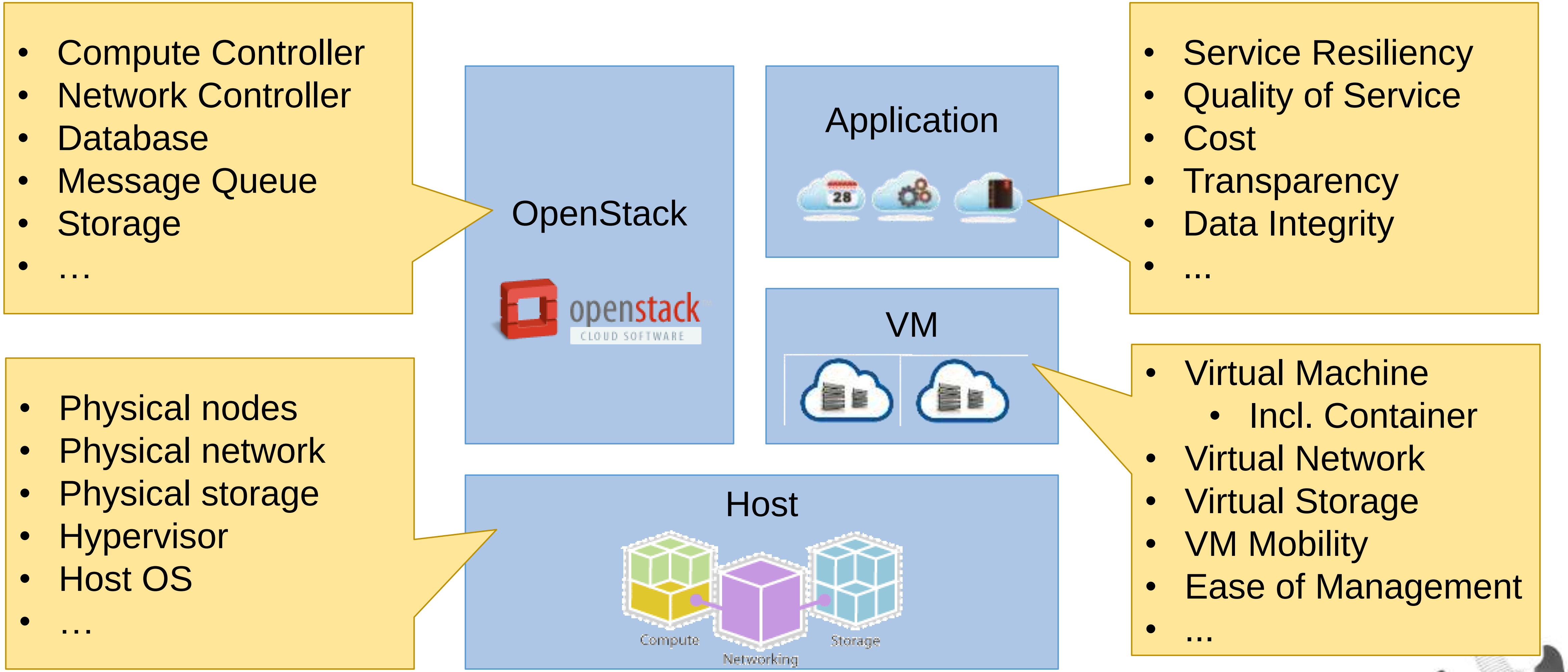
- Use cases
  - Server consolidation in private cloud
  - Selling point for public cloud
  - Ease of management
    - Planned/unplanned downtime
  - (potentially) a user consumable service

# How to Achieve HA?

- Three Technologies
  - Redundancy
    - Capacity Planning
    - Cost
  - Detection
    - Watchdog
    - Heartbeat messages
  - Recovery
    - Transparency
    - Data consistency
    - Interruption time
- Implications
  - Automatic
  - Autonomous



# Four Types of High Availability in an OpenStack Cloud



- Compute Controller
- Network Controller
- Database
- Message Queue
- Storage
- …

- Physical nodes
- Physical network
- Physical storage
- Hypervisor
- Host OS
- …

OpenStack

Application

VM

Host

- Service Resiliency
- Quality of Service
- Cost
- Transparency
- Data Integrity
- ...

- Virtual Machine
  - Incl. Container
- Virtual Network
- Virtual Storage
- VM Mobility
- Ease of Management
- ...

# OpenStack HA: Deployment Pattern

- Main Focus
  - Avoid SPOF (Single Point of Failure) in OpenStack services
    - Controller, Network, Compute, Swift, etc.
  - Stateful versus Stateless services
- Implementation
  - Primarily based on Pacemaker/Corosync Linux-HA stack, plus a load-balancer
  - Keepalived/haproxy

- A Deployment Pattern, not part of OpenStack core components
  - HA Guide documentation
  - Chef cookbooks
  - TripleO elements

- Only deployment, no runtime management service

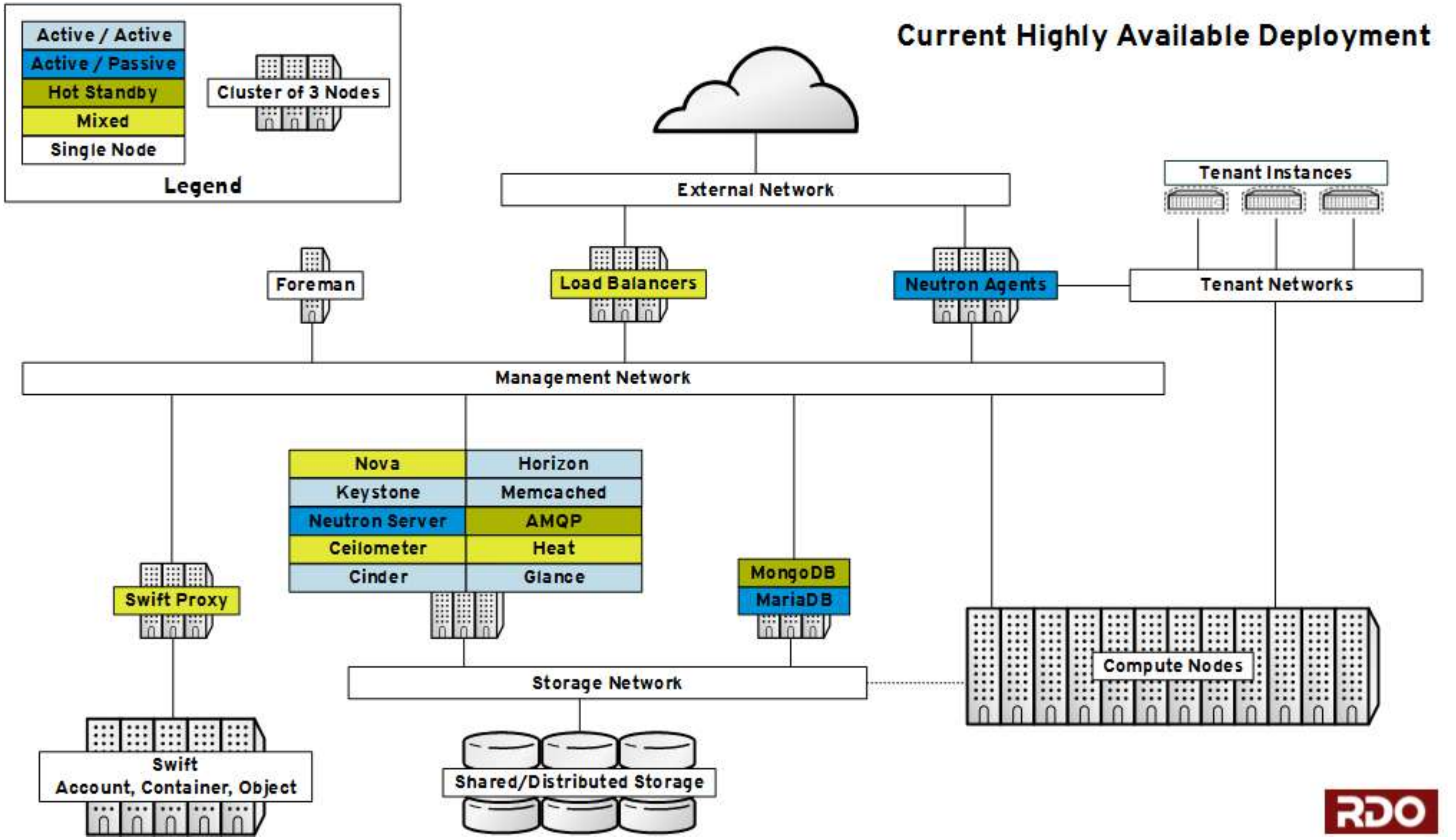
OpenStack

High Availability Guide

openstack

openstack

# An example setup (RDO)



# OpenStack HA: Intrinsic Supports

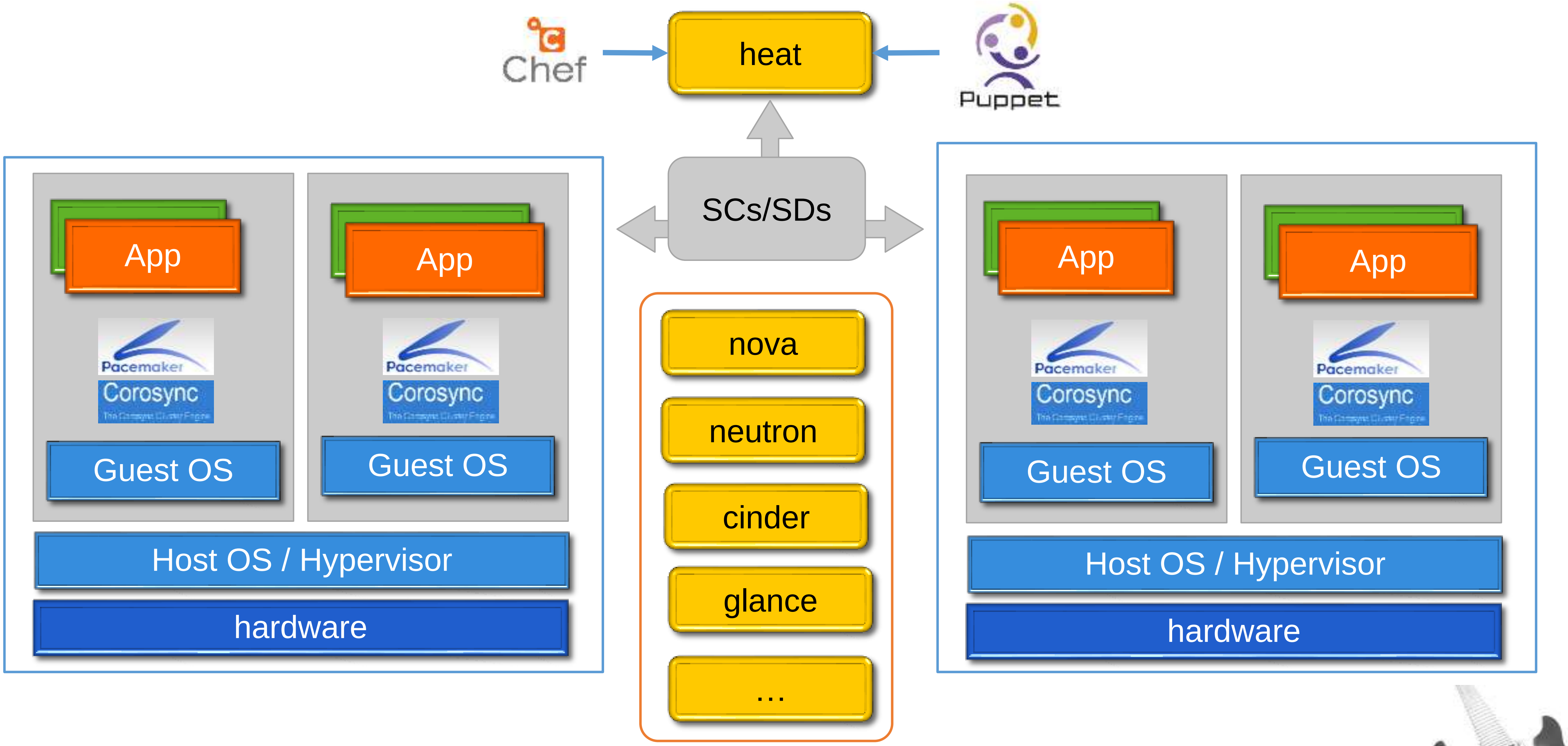
- Nova
  - Host Aggregates
  - Availability Zones
  - Service Groups
    - Internal heartbeat messages, zookeeper/memcached/matchmaker
  - …
- Message Queues
  - QPID heartbeats (60 seconds interval)
  - ZeroMQ w/ MatchMaker
- Cinder
  - Storwize driver (heartbeat: 10 seconds)
  - Contrib services
- Swift
- ...

# OpenStack HA: Internal Heartbeats

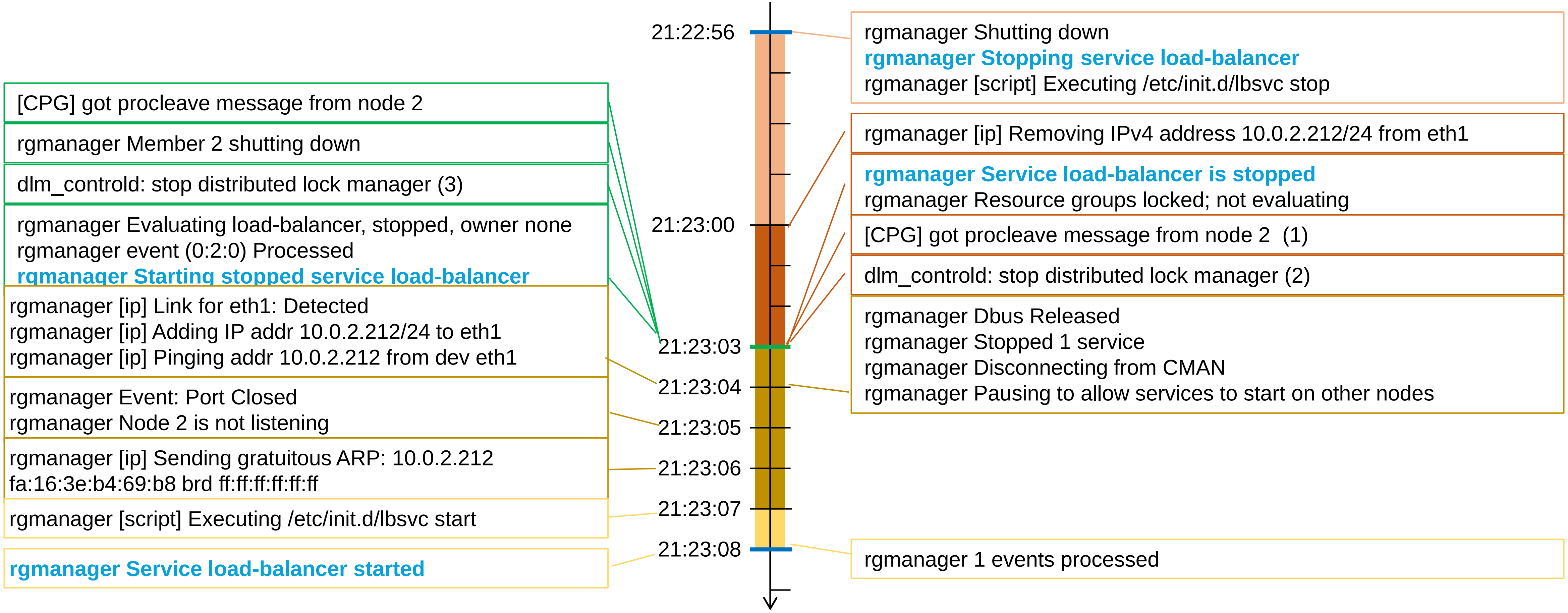
```
[tengqm@node1 ~]$ nova service-list
+----+------------------+-------+----------+---------+-------+----------------------------+-----------------+
| Id | Binary           | Host  | Zone     | Status  | State | Updated_at                 | Disabled Reason |
+----+------------------+-------+----------+---------+-------+----------------------------+-----------------+
| 1  | nova-conductor   | node1 | internal | enabled | up    | 2014-04-27T20:37:09.000000 | -               |
| 3  | nova-cert        | node1 | internal | enabled | up    | 2014-04-27T20:37:05.000000 | -               |
| 4  | nova-scheduler   | node1 | internal | enabled | up    | 2014-04-27T20:37:06.000000 | -               |
| 5  | nova-consoleauth | node1 | internal | enabled | up    | 2014-04-27T20:37:05.000000 | -               |
| 6  | nova-compute     | node1 | nova     | enabled | up    | 2014-04-27T20:37:06.000000 | -               |
+----+------------------+-------+----------+---------+-------+----------------------------+-----------------+
```

```
[tengqm@node1 ~]$ neutron agent-list
Starting new HTTP connection (1): 9.186.106.171
Starting new HTTP connection (1): 9.186.106.171
+--------------------------------------+--------------------+-------+-------+----------------+
| id                                   | agent_type         | host  | alive | admin_state_up |
+--------------------------------------+--------------------+-------+-------+----------------+
| 0f9b8470-577e-4439-84f1-36ce92eac77d | Metadata agent     | node1 | :-)   | True           |
| 7ac10787-9a62-4a96-868f-bd90bb46d52b | L3 agent           | node1 | :-)   | True           |
| c89d0bac-8a41-44ee-8df0-389a9c8db428 | Open vSwitch agent | node1 | :-)   | True           |
| e138db2d-bf3b-4ac2-89ab-50dbb8771a7b | DHCP agent         | node1 | :-)   | True           |
+--------------------------------------+--------------------+-------+-------+----------------+
```

# VM/Application HA: Guest Clusters



# VM/Application HA Timeline (reboot VM-2)



[CPG] got procleave message from node 2

rgmanager Member 2 shutting down

dlm_controld: stop distributed lock manager (3)

rgmanager Evaluating load-balancer, stopped, owner none
rgmanager event (0:2:0) Processed
**rgmanager Starting stopped service load-balancer**

rgmanager [ip] Link for eth1: Detected
rgmanager [ip] Adding IP addr 10.0.2.212/24 to eth1
rgmanager [ip] Pinging addr 10.0.2.212 from dev eth1

rgmanager Event: Port Closed
rgmanager Node 2 is not listening

rgmanager [ip] Sending gratuitous ARP: 10.0.2.212
fa:16:3e:b4:69:b8 brd ff:ff:ff:ff:ff:ff

rgmanager [script] Executing /etc/init.d/lbsvc start

**rgmanager Service load-balancer started**

rgmanager Shutting down
**rgmanager Stopping service load-balancer**
rgmanager [script] Executing /etc/init.d/lbsvc stop

rgmanager [ip] Removing IPv4 address 10.0.2.212/24 from eth1

**rgmanager Service load-balancer is stopped**
rgmanager Resource groups locked; not evaluating

[CPG] got procleave message from node 2 (1)

dlm_controld: stop distributed lock manager (2)

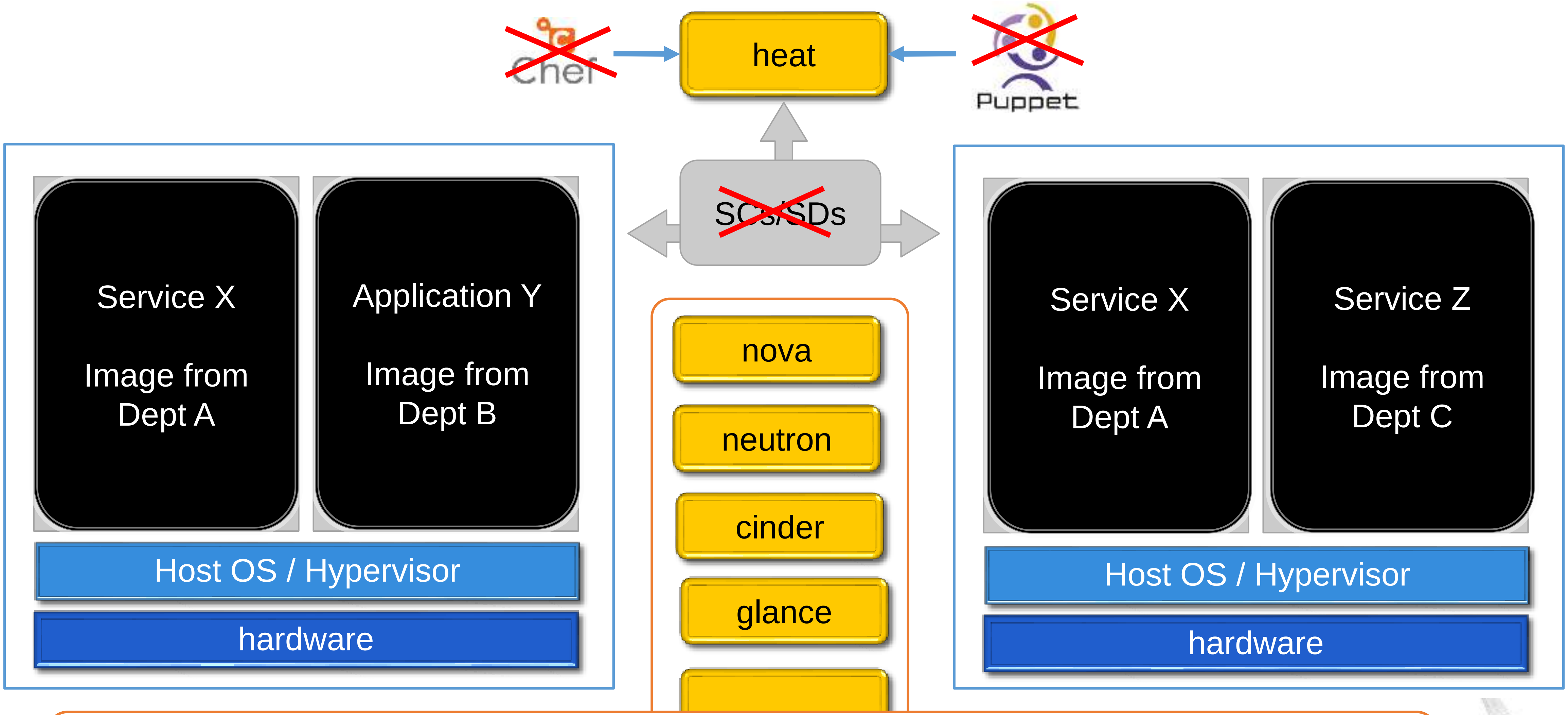
rgmanager Dbus Released
rgmanager Stopped 1 service
rgmanager Disconnecting from CMAN
rgmanager Pausing to allow services to start on other nodes

rgmanager 1 events processed

VM-1

VM-2(rebooted)

# VM/Application HA: Guest Clusters

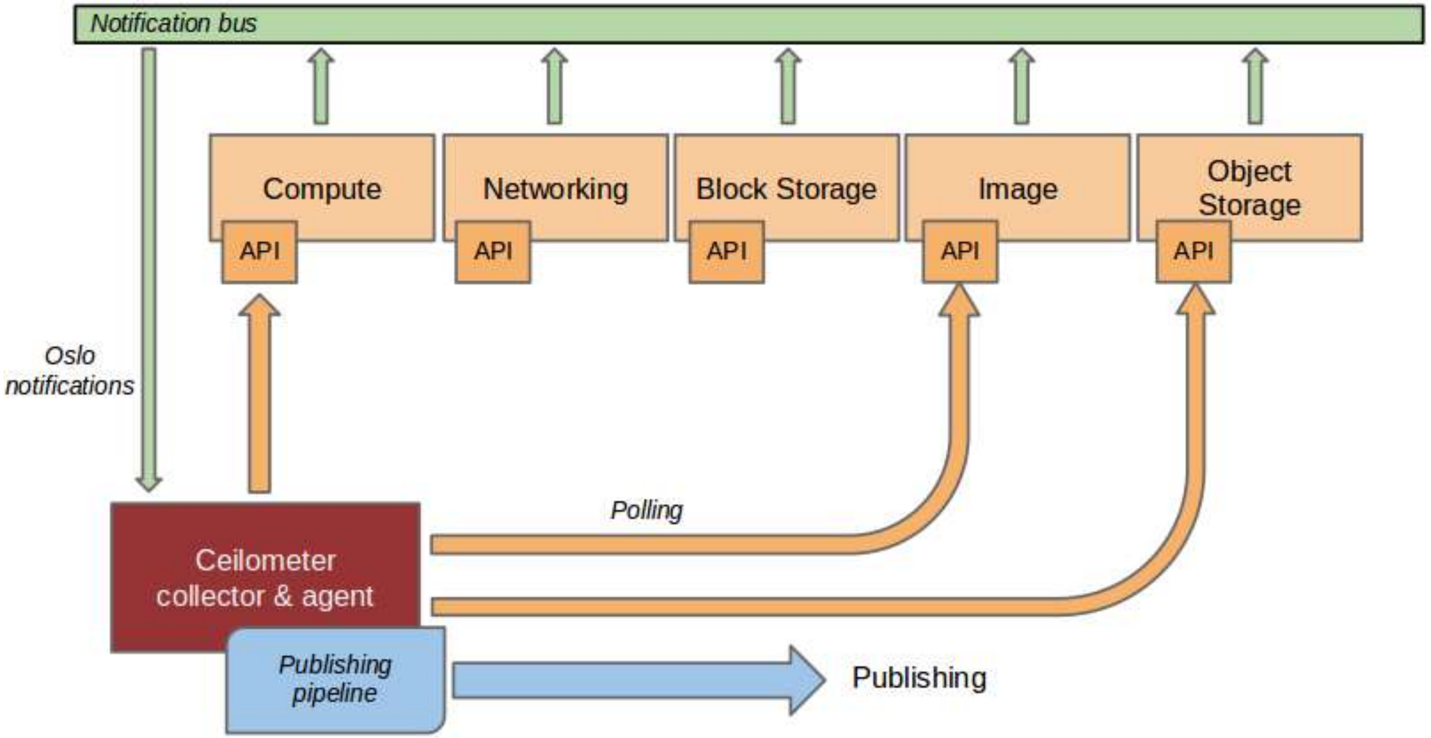


**LIMITATIONS**

- Ease of management - Application Specific - Intrusive

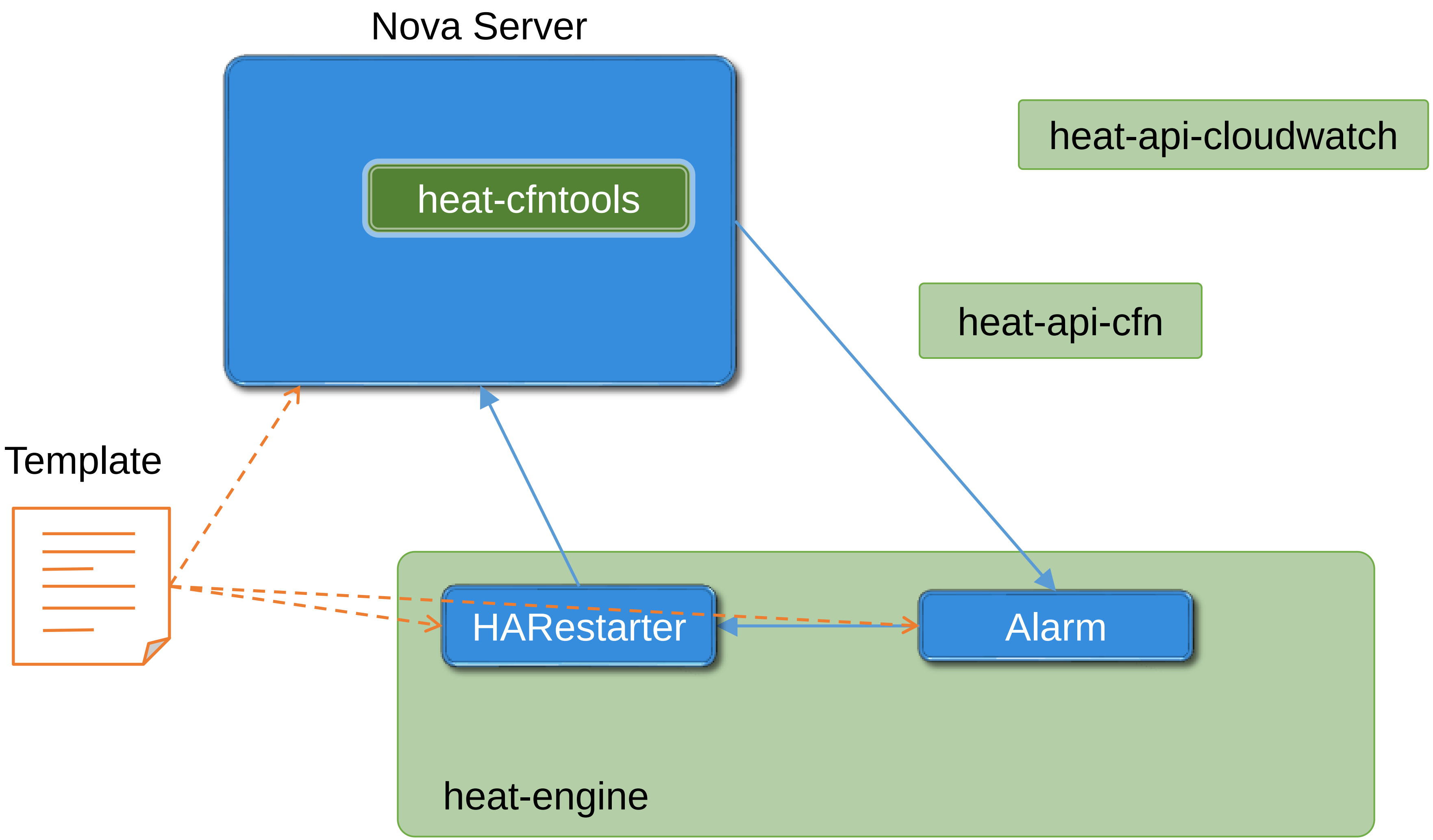
# VM/Application HA: Intrinsic Supports

- Redundancy
  - Nova
    - Server Groups
    - Virtual Ensembles ?
    - Virtual Clusters ?
  - Heat
    - `InstanceGroup` resource
    - `ResourceGroup` resource
- Detection
  - RPC notification, `oslo.messaging`
  - Ceilometer
- Recovery
  - Fencing support in nova, cinder, neutron [undergoing]
  - VM reboot, rebuild, evacuation …
  - `OS::Heat::HARestarter` resource in Heat (deprecating)
  - …

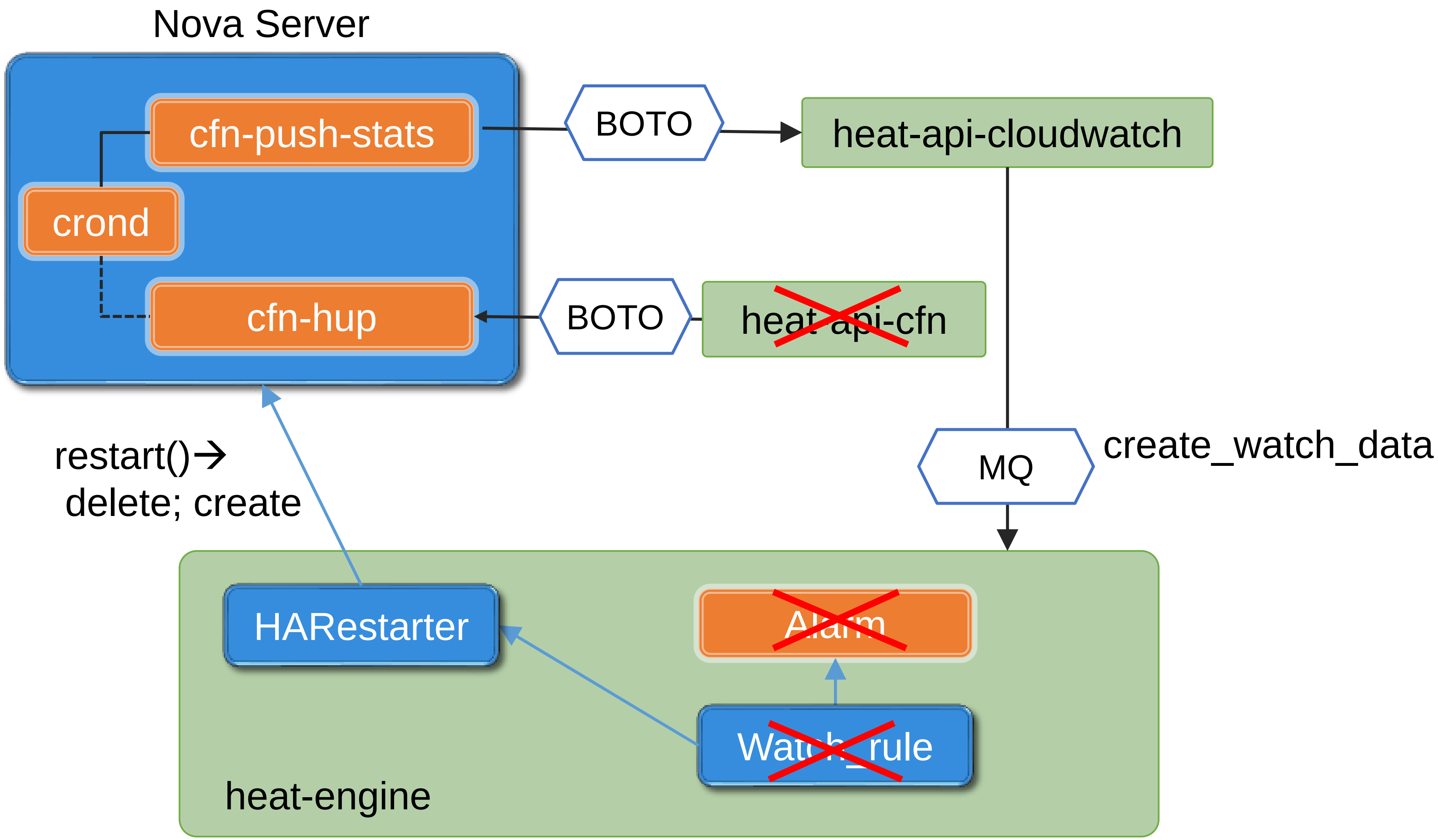


Ceilometer Notification

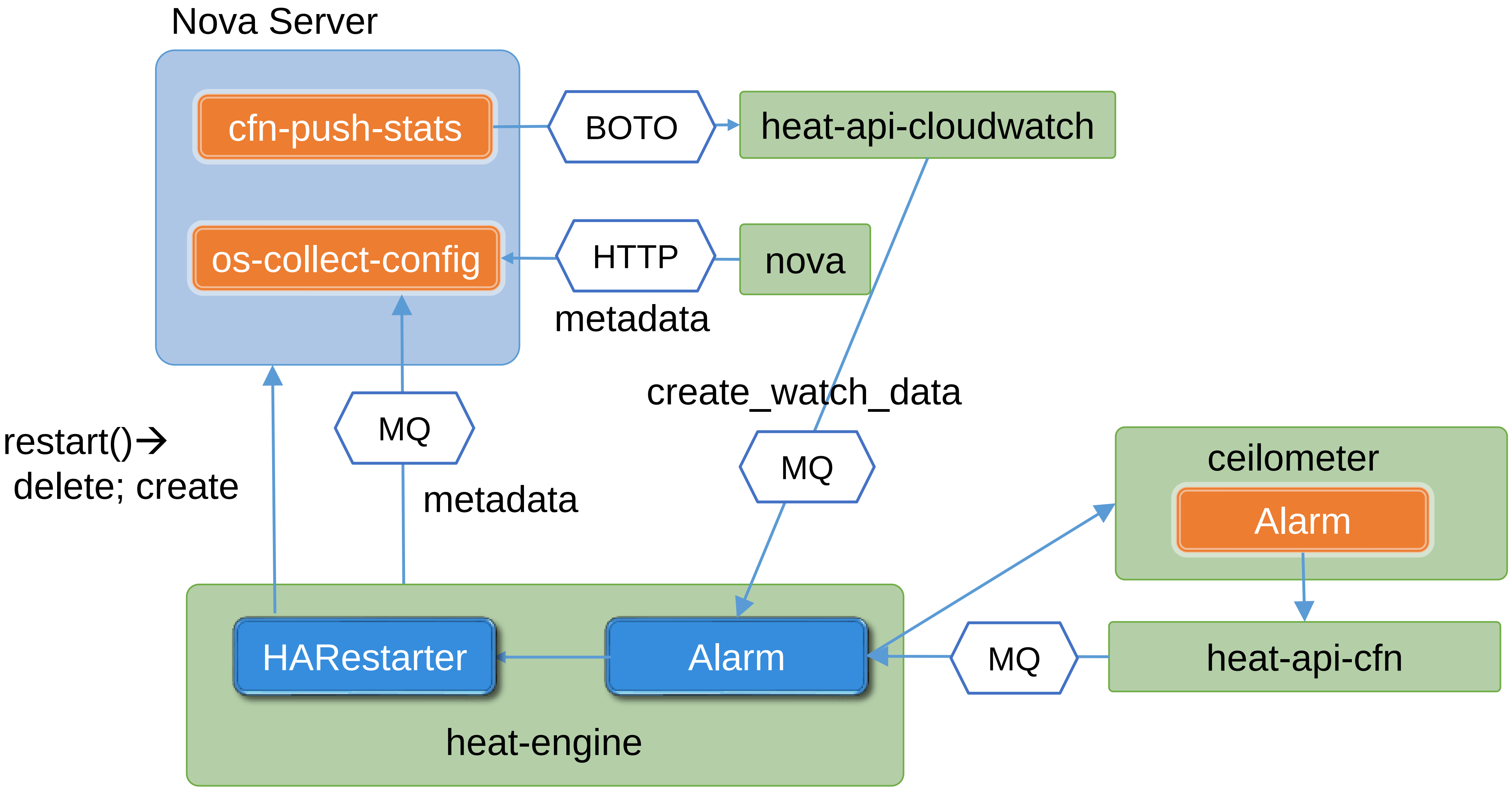
# VM/Application HA: Heat Orchestrated – yesterday



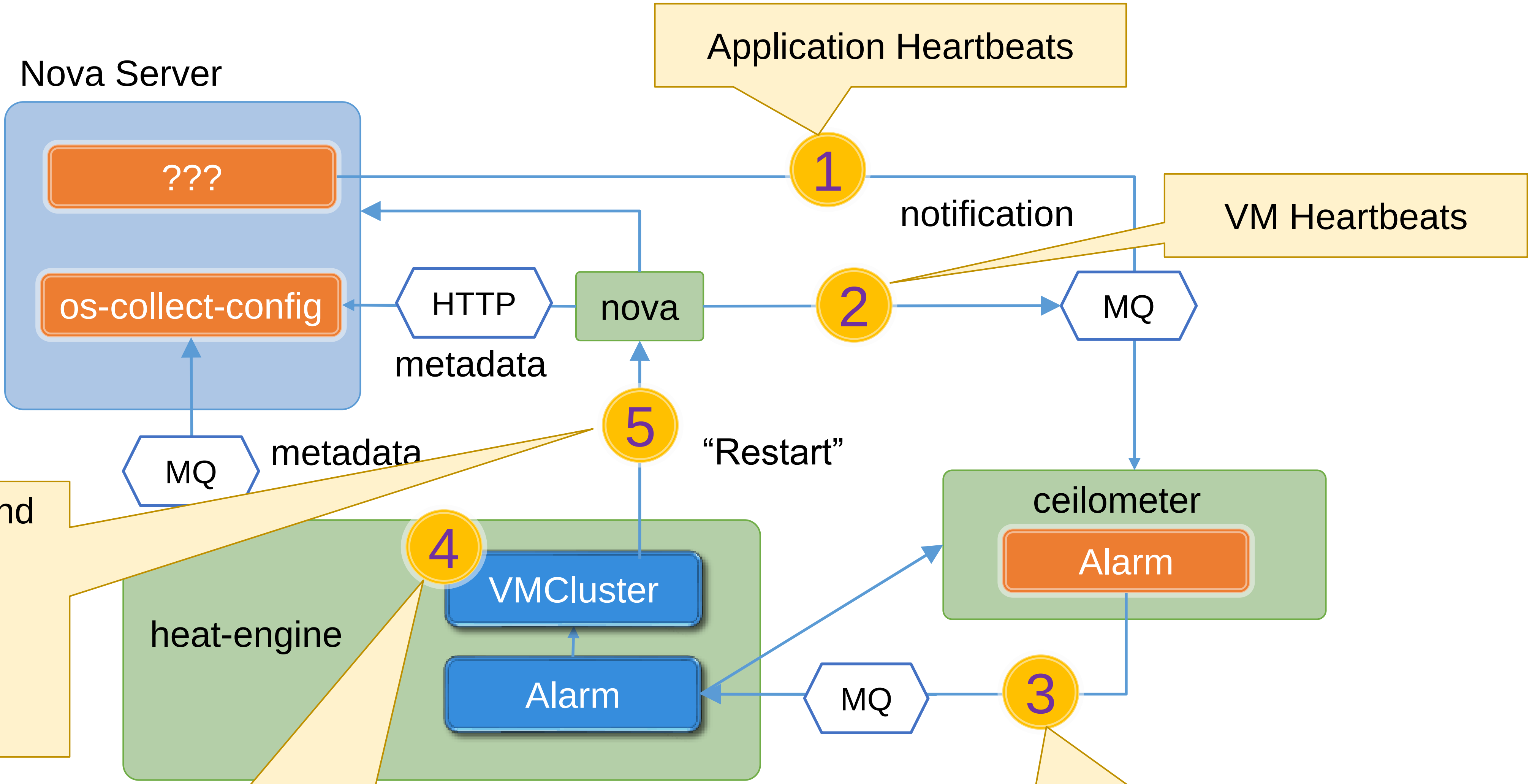
# VM/Application HA: Heat Orchestrated – yesterday



# VM/Application HA: Heat Orchestrated – today



# VM/Application HA: Heat Orchestrated – tomorrow?



Native support VM and application recovery:

- reboot
- rebuild
- migrate
- remote-restart

A notion of VM groups for

- VM/Application redundancy
- HA policies

A native signal / alarm

# VM/Application HA: Open Questions

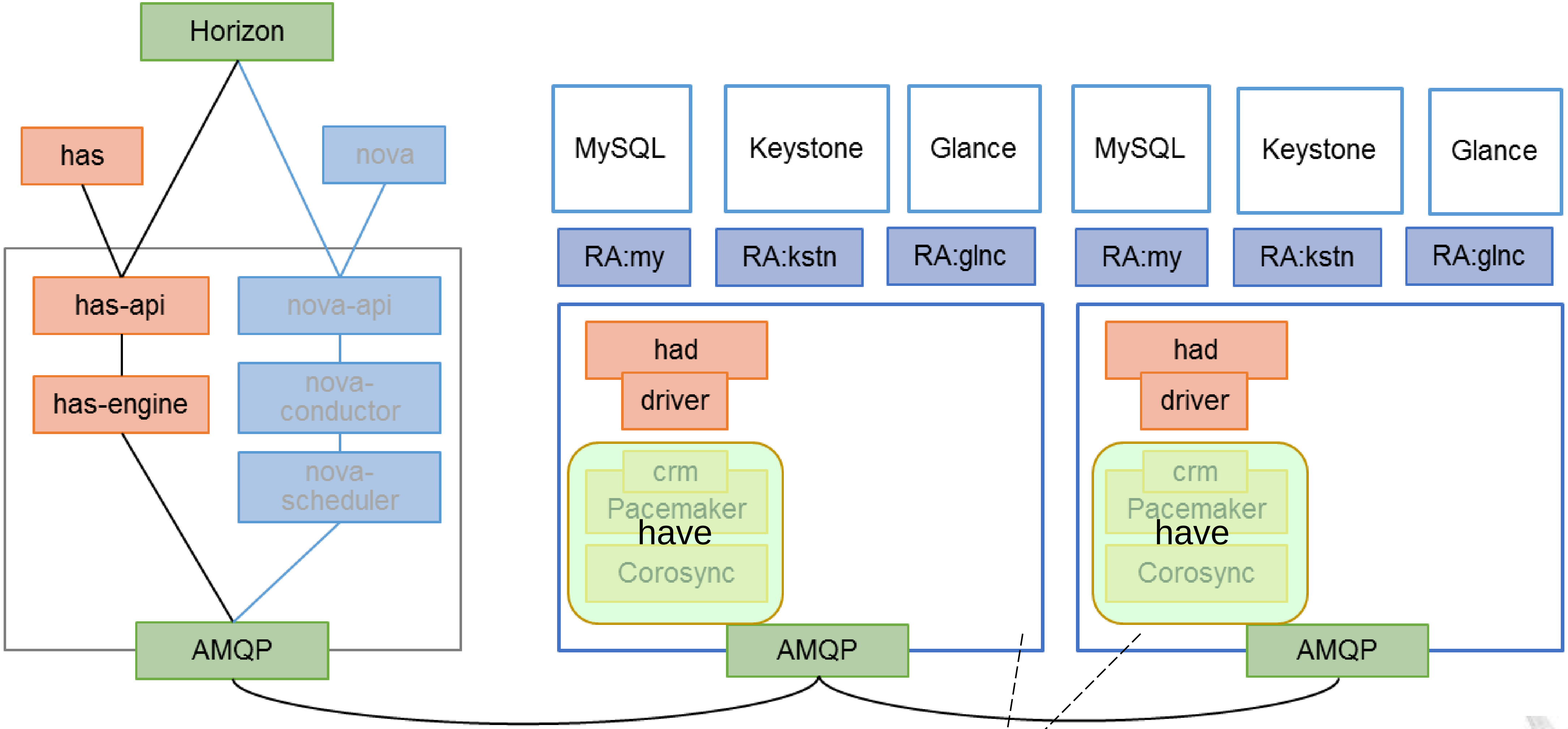
- Physical placement of VMs
  - No shared PDU/rack, no shared network switch
  - HA-aware scheduling, e.g. server priority
- Detection of failures
  - High availability and QoS, e.g. desired latency/throughput versus reality
  - Reliable detection, application involvement, …
- Reasoning of failures
  - Root cause, trend analysis
  - Log collection and analysis
- HA management / orchestration
  - As a cross-cutting concern, involving not only compute, but also storage and network
    - Stack availability?
  - Capacity planning / reservation
- Leverage existing HA software
  - Can we leverage supports from hypervisors?
  - Can/should we generalize this into a service?

# High Availability as a Service (HAaaS)

- Generic HA management service
  - Applicable to different levels of HA
    - Host, VM, App, OpenStack
  - Applicable to different hypervisors
    - vSphere, KVM, Xen, HyperV, PowerVM, …)
  - Functionality determined via user authentication

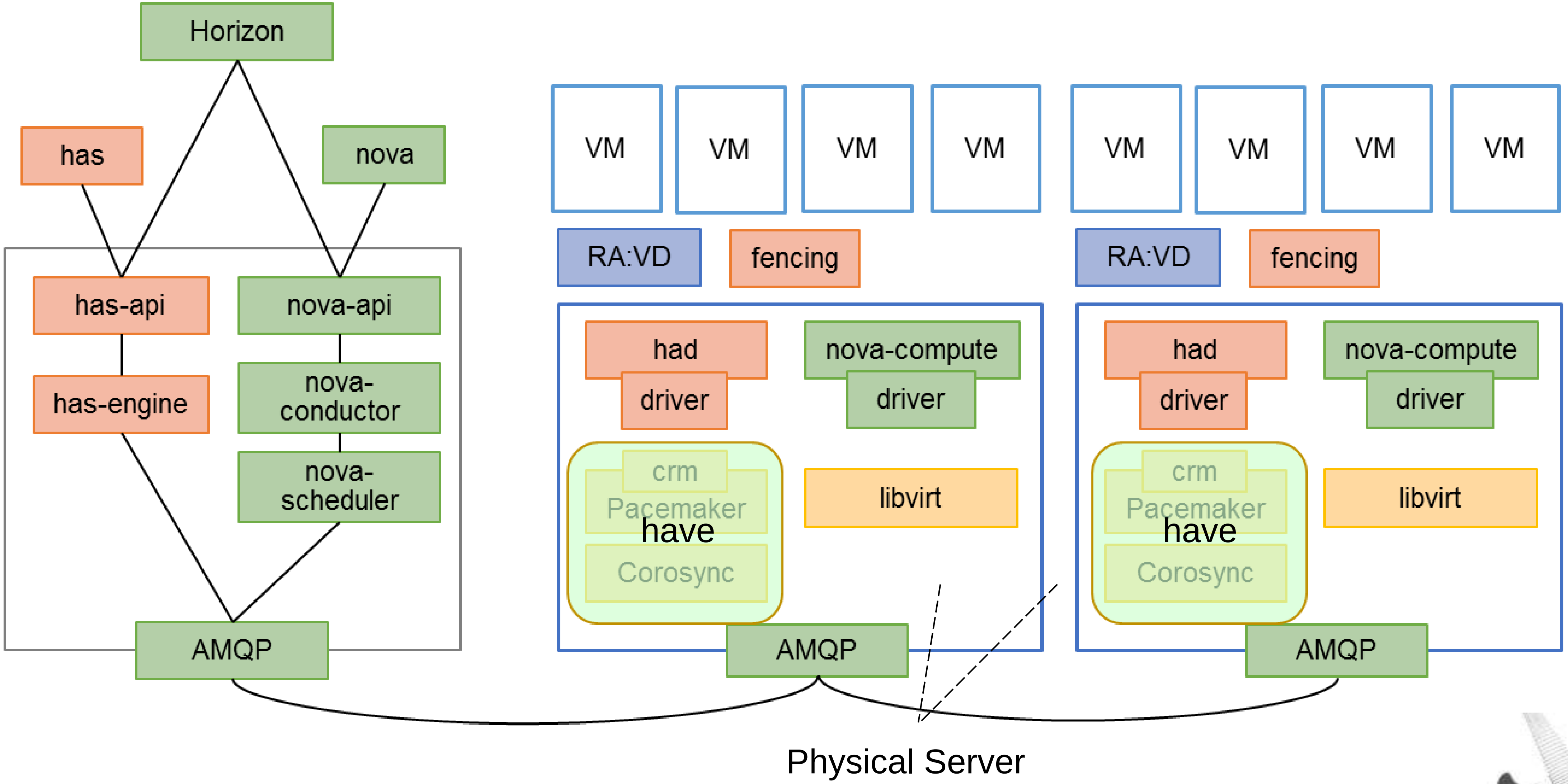
- Well-defined service APIs
  - Clusters management
  - Application/Service resource definition
  - HA policies
    - Fail-over domain
    - Fail-over priority, operation, timeout, retries, …

# HAaaS: OpenStack HA

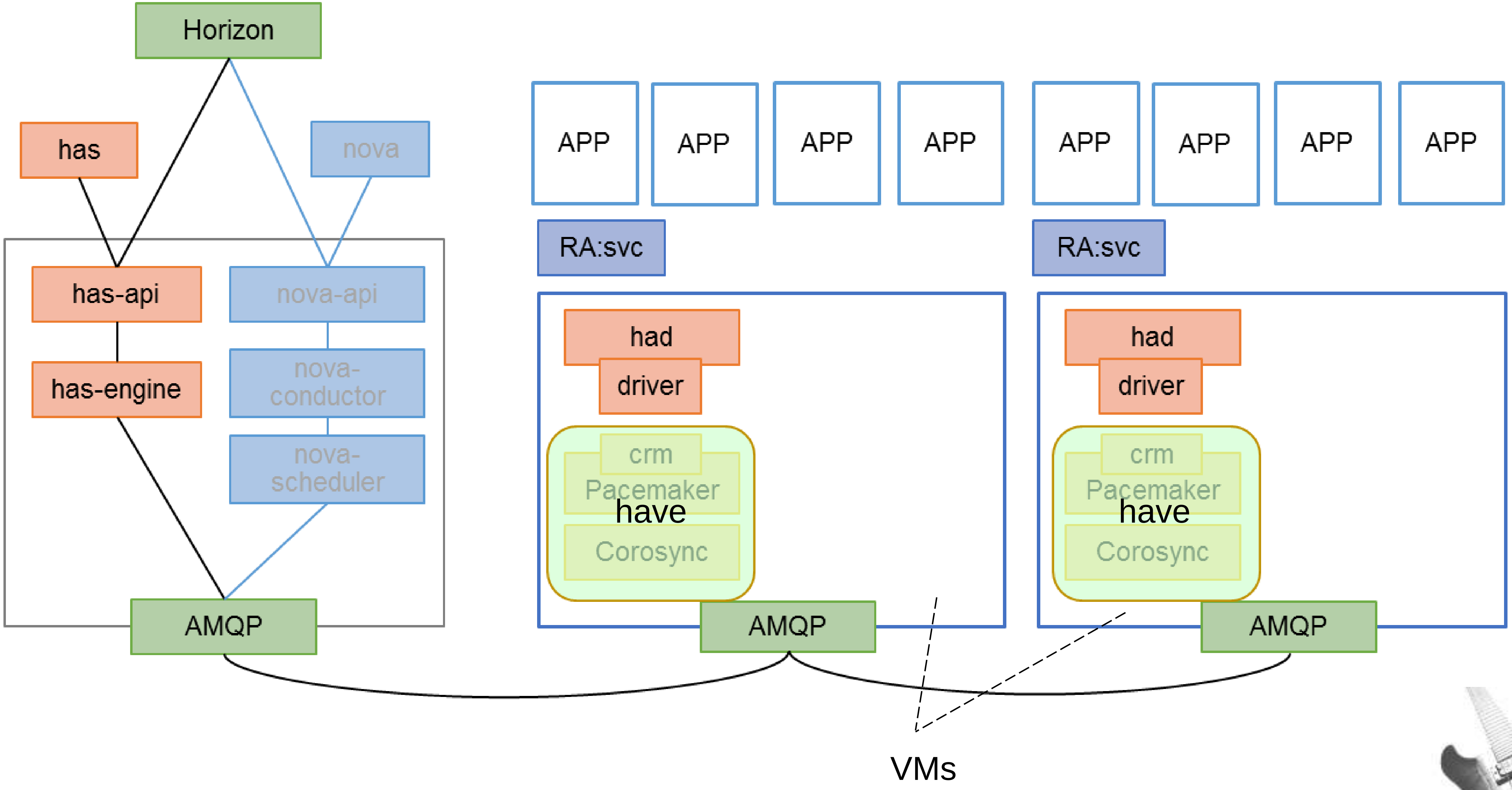


VM or Physical Servers

# HAaaS: VM HA


Horizon
has
nova
has-api
nova-api
has-engine
nova-conductor
nova-scheduler
AMQP
VM
VM
VM
VM
RA:VD
fencing
had
driver
nova-compute
driver
crm
Pacemaker
have
Corosync
libvirt
AMQP
VM
VM
VM
VM
RA:VD
fencing
had
driver
nova-compute
driver
crm
Pacemaker
have
Corosync
libvirt
AMQP
Physical Server

# HAaaS: Application HA



# Service Interfaces (1)

- Cluster Management
  - cluster_create,
  - cluster_destroy,
  - cluster_start,
  - cluster_stop,
  - cluster_suspend,
  - cluster_resume,
  - cluster_set_attr,
  - cluster_get_attr,
  - cluster_by_host,
  - cluster_get_status,
  - cluster_get_log,
  - ...

- Node Management (physical/virtual)
  - node_join_cluster
  - node_leave_cluster
  - node_get_attr
  - node_set_attr
  - node_startup
  - node_shutdown
  - node_reboot
  - node_evacuate
  - node_get_status
  - ...

# Service Interfaces (2)

- Resource Management
  - resource_create
  - resource_destroy
  - resource_get_attr
  - resource_set_attr
  - ...
- Fencing Management
  - Fencing_dev_add
  - Fencing_dev_del
  - Fencing_dev_associate
  - Fencing_dev_deassociate
  - Fencing_dev_set_opts
  - Fencing_dev_get_opts
  - ...
- Service Management (aka. resource groups)
  - service_create
  - service_destroy
  - service_add_resource
  - service_del_resource
  - service_list
  - service_get_attr
  - service_set_attr
  - service_start
  - service_stop
  - service_restart
  - service_relocate
  - ...

# IBM Sponsored Sessions

## Monday, May 12 – Room B314

**12:05-12:45**

OpenStack is Rockin' the OpenCloud Movement! Who's Next to Join the Band ?
Angel Diaz, VP Open Technology and Cloud Labs
David Lindquist, IBM Fellow, VP, CTO Cloud & Smarter Infrastructure

## Wednesday, May 14 - Room B312

**9:00-9:40**

Getting from enterprise ready to enterprise bliss - why OpenStack and IBM is a match made in Cloud heaven.
Todd Moore - Director, Open Technologies and Partnerships

**9:50-10:30**

IBM and OpenStack: Enabling Enterprise Cloud Solutions Now.
Tammy Van Hove -Distinguished Engineer, Software Defined Systems

**11:00-11:40**

Taking OpenStack beyond Infrastructure with IBM SmartCloud Orchestrator.
Andrew Trossman - Distinguished Engineer, IBM Common Cloud Stack and SmartCloud Orchestrator

**11:50-12:30**

IBM, SoftLayer and OpenStack - present and future
Michael Fork - Cloud Architect

# IBM Technical Sessions

**Monday, May 12**

| | |
|---|---|
| 3:40 - 4:20 | An Overview of Cloud Auditing Support for OpenStack |
| 3:40 - 4:20 | Hosting hybrid (bare-metal + virtualized) applications on OpenStack |

**Tuesday, May 13**

| | |
|---|---|
| 11:15 - 11:55 | Enhancing High Availability in Context of OpenStack |
| 2:00 - 2:40 | Training your cluster to take care of itself and let you eat dinner in peace |
| 5:30 - 6:10 | Optimizing OpenStack for large scale Cloud Foundry deployments |
| 5:30 - 6:10 | Turning the Heat up on DevOps: Providing a web-based editing experience around Heat templates |

**Wednesday, May14**

| | |
|---|---|
| 9:50 - 10:30 | Linux Containers - NextGen Virtualization for Cloud |
| 2:40 - 3:20 | A practical approach to deploying a highly available and optimally performing OpenStack |

**Thursday, May 15**

| | |
|---|---|
| 9:50 - 10:30 | Federated Identity & Federated Service Provider Support for OpenStack Clouds |
| 1:30 - 2:10 | Network Policy Abstractions in Neutron |
| 2:20 - 3:00 | Hybrid Cloud with OpenStack: Bridging Two Worlds |

- ● Apps on OpenStack
- ● Compute
- ● Networking
- ● Operations
- ● Public & Hybrid Clouds
- ● Security
- ● Ecosystem

Be sure to stop by the IBM booth to see some demos and get your rockin' OpenStack T-shirt while they last.

Thank you !