IBM

Open by design™

Elastic Load-Balancing Using Octavia deep dive

Dean H. Lorenz, IBM Research – Haifa

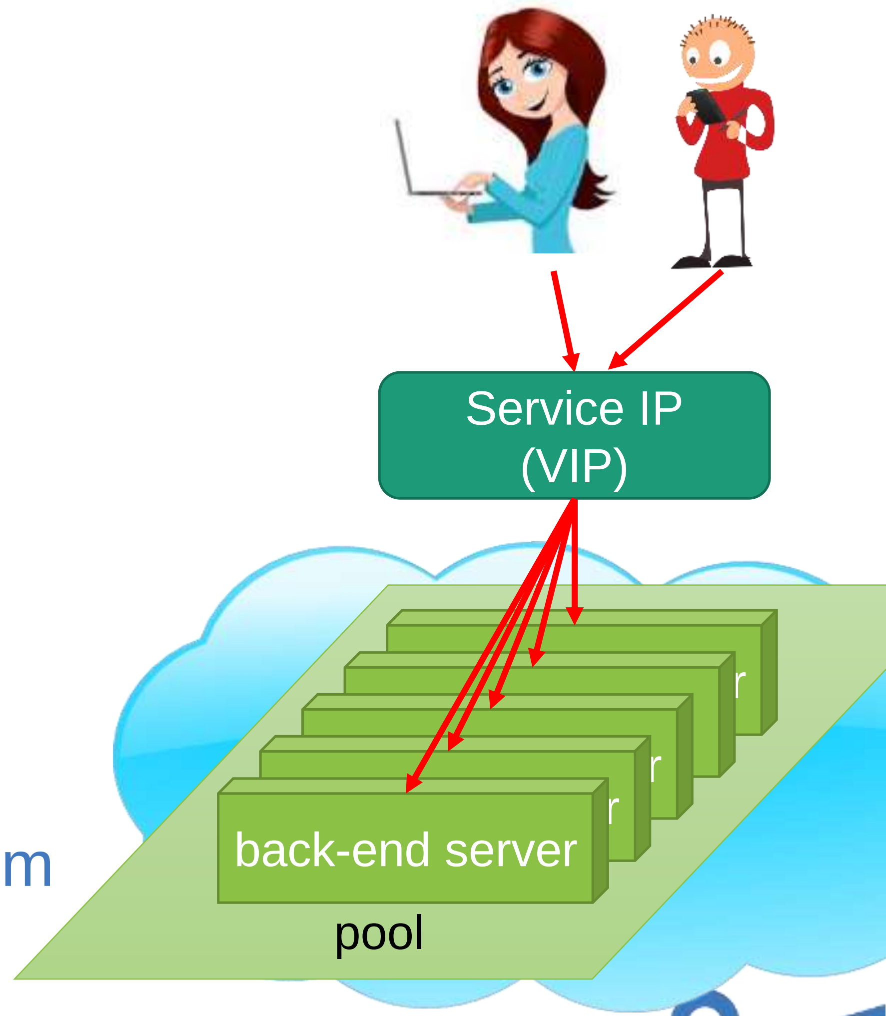
Allan Hu, Cloud Networking Services, IBM NSJ

Octavia
OpenStack Load Balancing

OpenStack Austin 2016

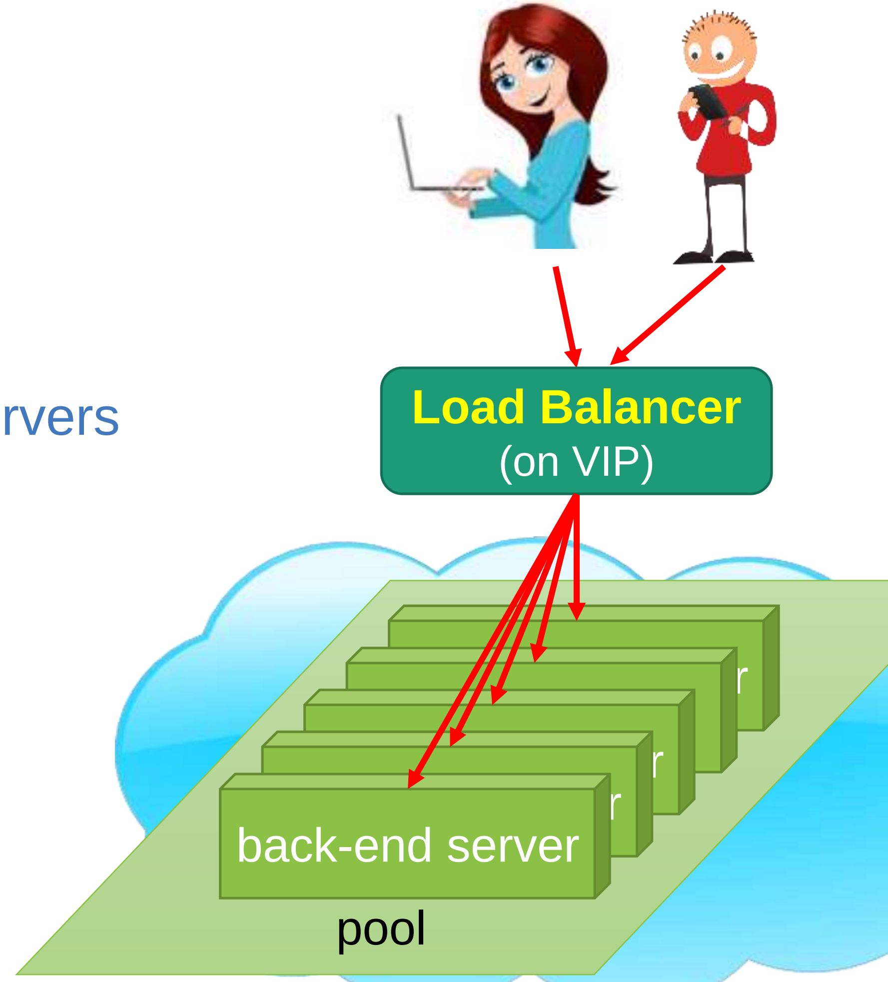
Load Balancing 101

- Users access a service
  - Service hosted on cloud
- **Pool** of back-end servers (aka **members**)
  - High availability:
    - server failure ≠ service failure
  - Performance:
    - add/remove servers to match load
- One service IP (aka **VIP**)
  - Clients do not know which back-end serves them
  - Need to split incoming VIP traffic



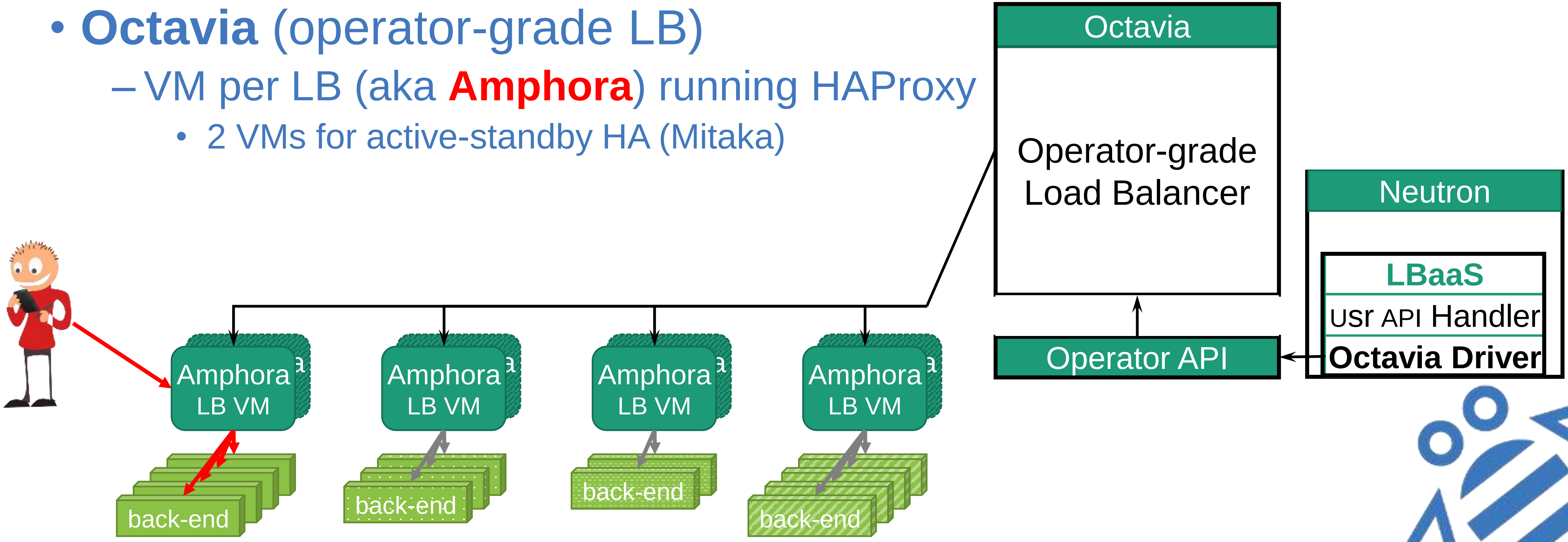
Load Balancing 101 (2)

- Load balancer
  - Distribute new VIP connections to members
  - High availability: avoid failed servers
  - Performance: avoid overloaded servers
    - LB is not the pool manager: does not add/remove servers
    - But uses all available servers, reports broken ones
  - **Health Monitor + Stats Collector**
- **LB Algorithm** / Policy
  - Balance something
    - # connections, CPU load…
  - **Affinity**: similar packets go to same back-end
    - All packets from same flow (minimum affinity)
    - All packets from same source (quicker TLS handshakes)
    - All packets from same HTTP user



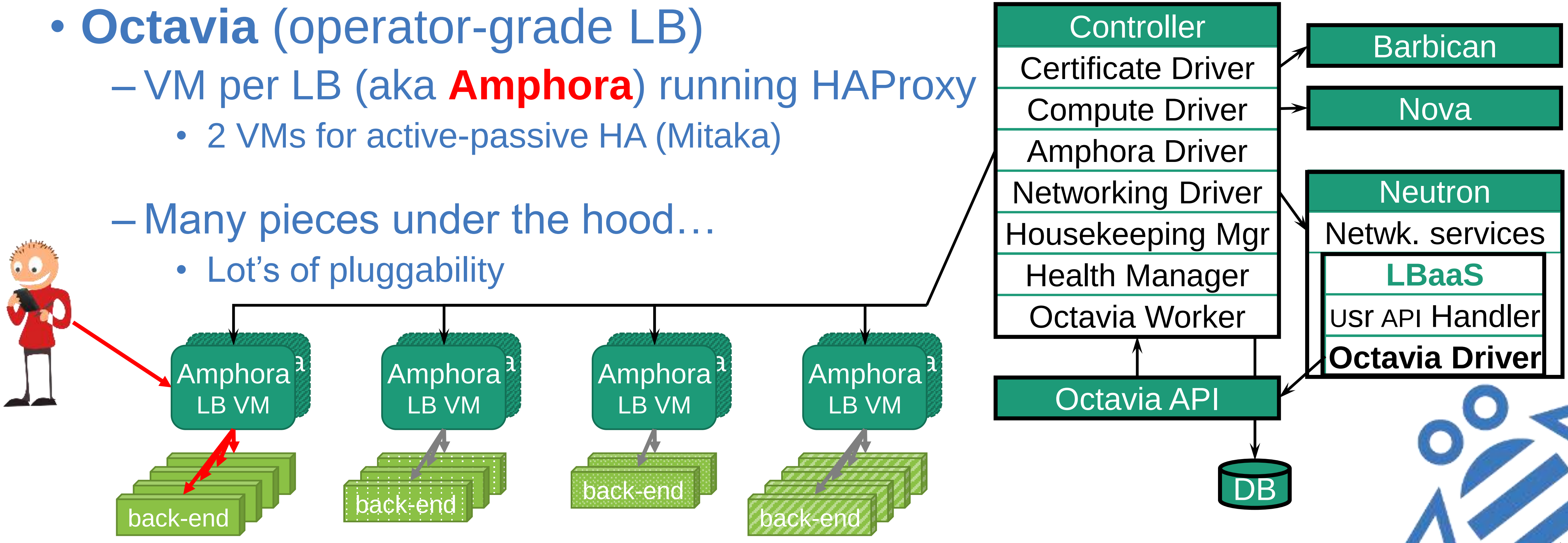
Load-Balancing as a Service (LBaaS)

- Neutron LBaaSv2 API
  - LB (VIP) → Listeners (protocol) → Pool → Members, Health monitor
    - `neutron lbaas-{loadbalancer,listener,pool,member,healthmonitor}-CRUD,`
      `CRUD: {create,delete,list,show,update}`
- **Octavia** (operator-grade LB)
  - VM per LB (aka **Amphora**) running HAProxy
    - 2 VMs for active-standby HA (Mitaka)



Load-Balancing as a Service (LBaaS)

- Neutron LBaaSv2 API
  - LB (VIP) → Listeners (protocol) → Pool → Members, Health monitor
    - `neutron lbaas-{loadbalancer,listener,pool,member,healthmonitor}-CRUD,`
      `CRUD: {create,delete,list,show,update}`
- **Octavia** (operator-grade LB)
  - VM per LB (aka **Amphora**) running HAProxy
    - 2 VMs for active-passive HA (Mitaka)
  - Many pieces under the hood…
    - Lot's of pluggability

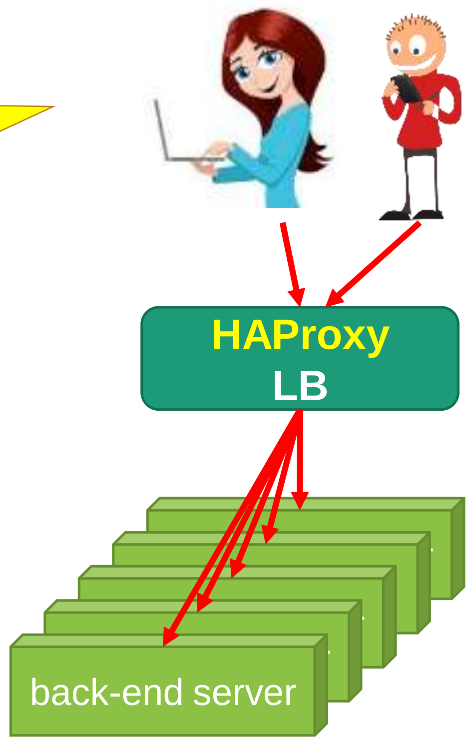


Amphora can do even more

- HAProxy is great
  - L7 Content Switching
  - Monitor back-end health
  - Cookie insertion (session stickiness)
  - SSL termination
  - Authentication
  - Compression
  - …

New in Octavia (Mitaka)

Not supported in Octavia (yet)

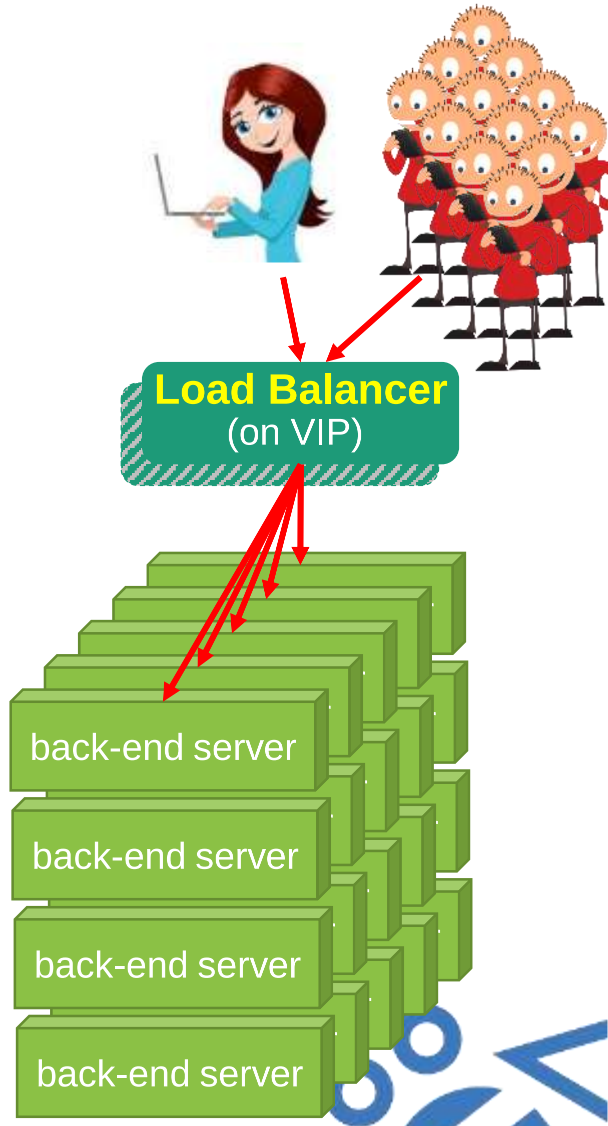


- Would be nice to include other functions
  - E.g., cache, FW, rewrite, …

⚠ The more it does, the more resources it needs

Remind me again; why did I need a LB?

- High availability
  - Amphora is single point of failure
  - *But active-standby just added in Mitaka*
- Performance:
  - Huge, successful service…
  - Amphora might not be able to handle load



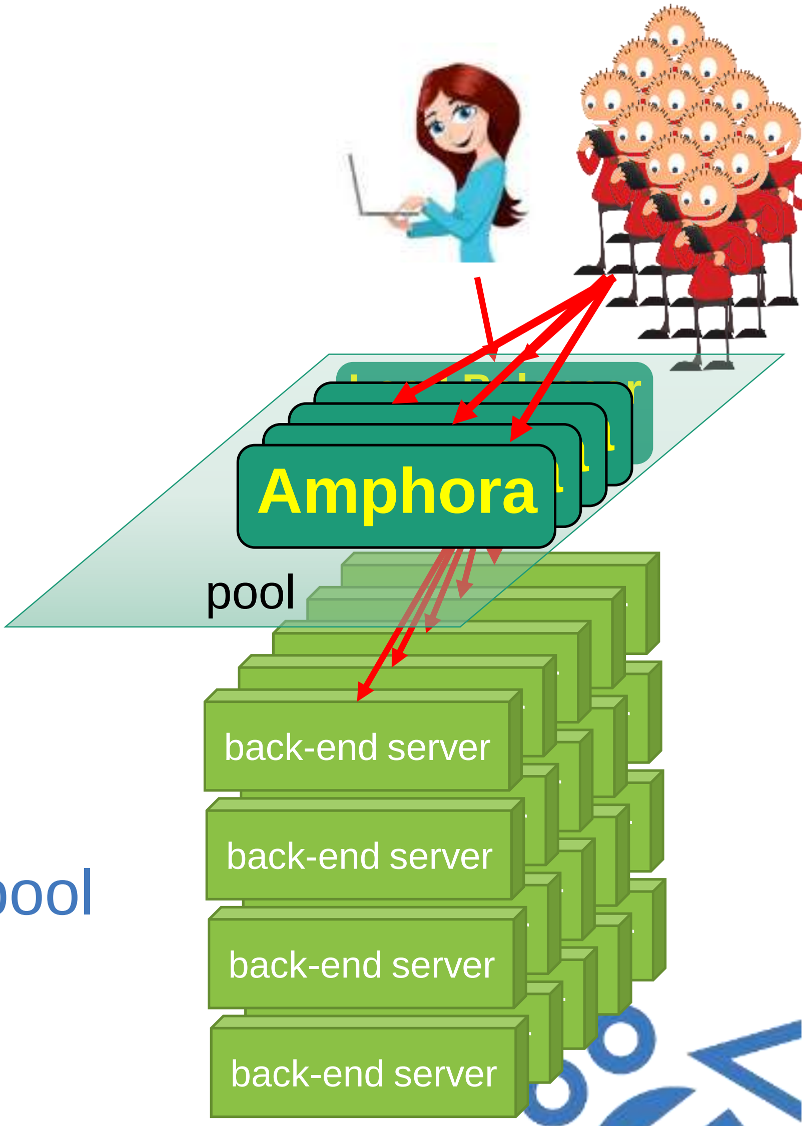
Elastic Load Balancing (ELB)

Remind me again; why did I need a LB?

- High availability
  - Amphora is single point of failure
  - *But active-standby just added in Mitaka*
- Performance:
  - Huge, successful service…
  - Amphora might not be able to handle load

Elastic Load-Balancing (ELB)

- Pool of Amphorae
- Need to split incoming VIP traffic over Amphorae pool
- Déjà vu…



Cost-effectively provide LBaaS for cloud workloads

- Customers expect the cloud to support their **elastic** workloads
  - Cheap for small workloads (free tier)
  - Acceptable performance for large workloads
    - No matter how large
- LbaaS should
  - Use as little resources as possible for small workloads
  - Have the resources to handle huge workloads
- Existing Octavia topologies have per LB
  - **One** active VM
    - Too small for large workloads? Too much for free tier? Maybe use containers?
  - (optionally) One **idle** standby VM
    - 50% utilization

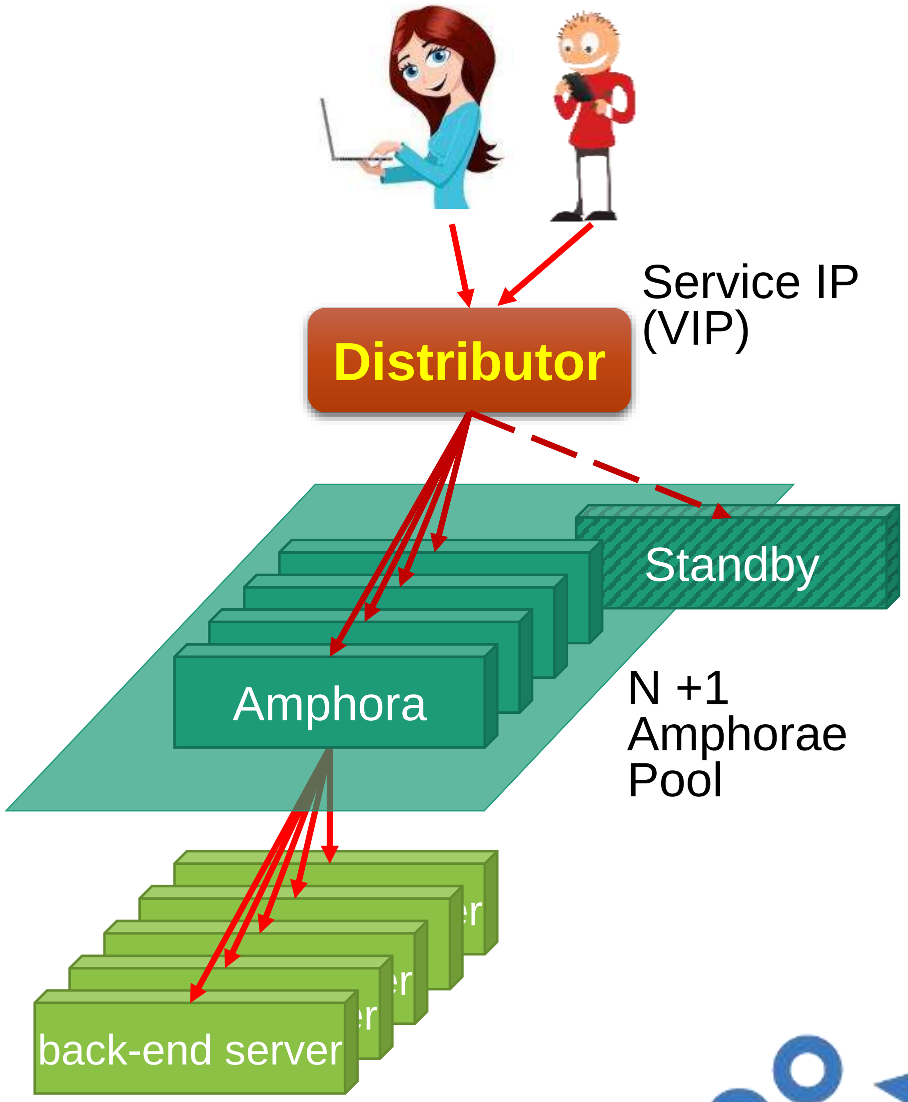
IBM

Introducing:

Active-Active, N+1 Topology

- *N* Amphorae, all active
  - Can handle large load
- 2-stage VIP traffic splitting
  1) **Distributor** to Amphorae
  2) Amphora to Back-end servers
- Standby Amphora
  - Ready to replace a failed Amphora
    - Takes over the load
  - Failed Amphora recreated as standby
  - Can generalize to more than one standby
    - $N + k$

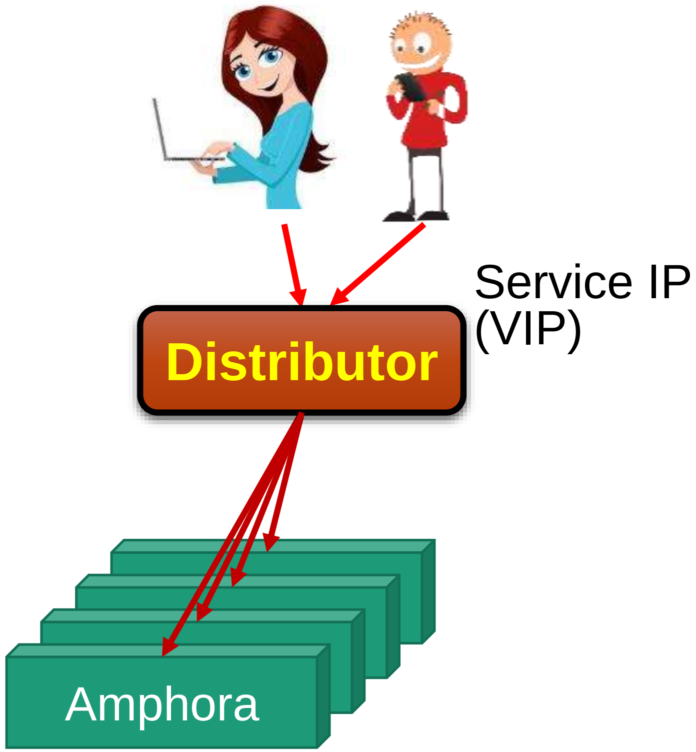
Open by design™



Disclaimer: Active-Active topology is still a draft blue-print ☹ (+ demo code ☺)

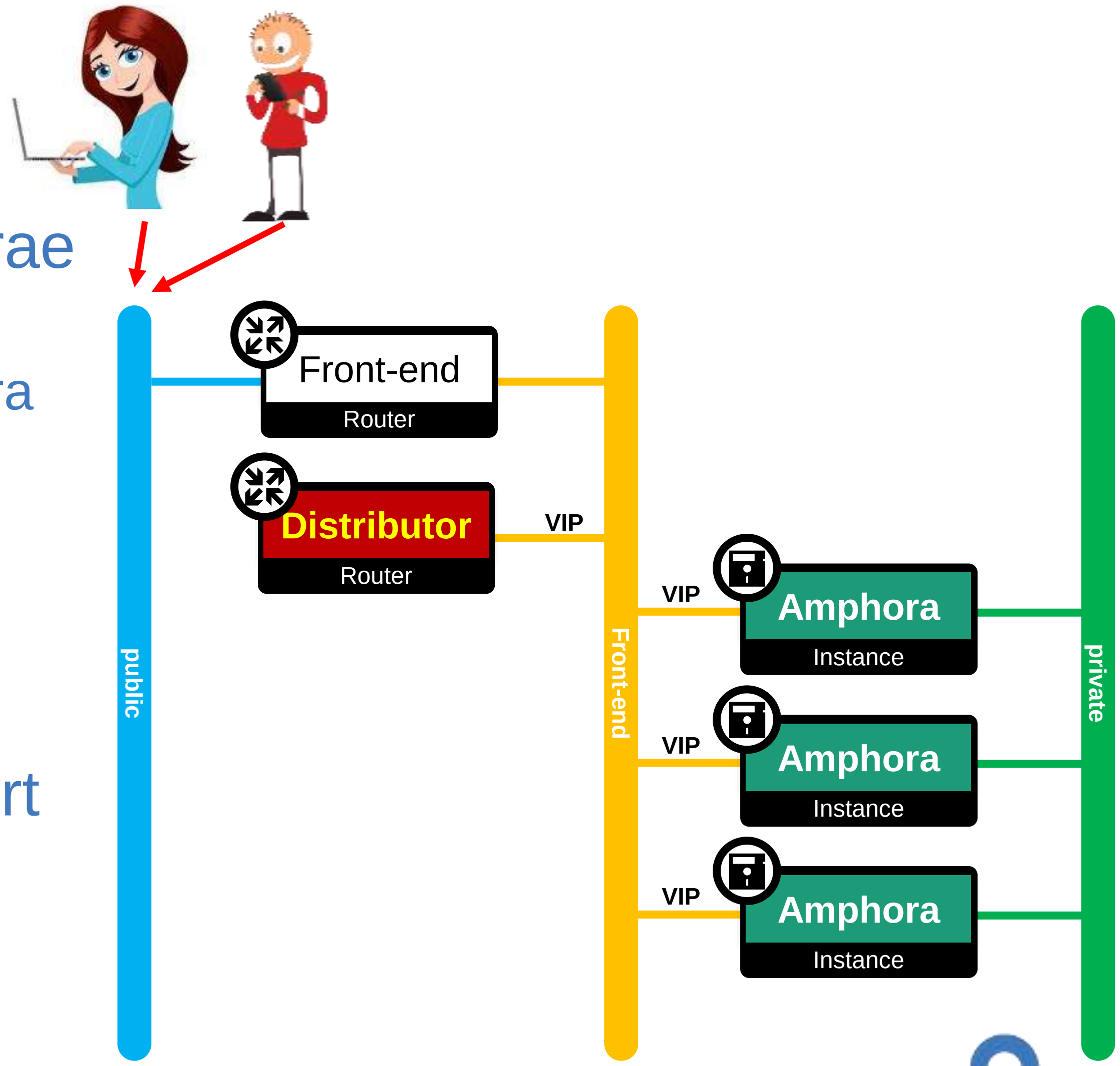
The Distributor

- Equivalent to a GW router
  - Should have similar high availability attributes
  - Needs to handle entire VIP load
  - HW is a good match
- "Not so smart" LB
  - More like ECMP
  - L3 only, but **must have per-flow affinity**
    - Cannot break TCP
- Could be shared (multi-tenant)
  - SSL termination is only at Amphora
- Could be DNS
  - If you have enough (public) IPs



Our SDN SW Distributor

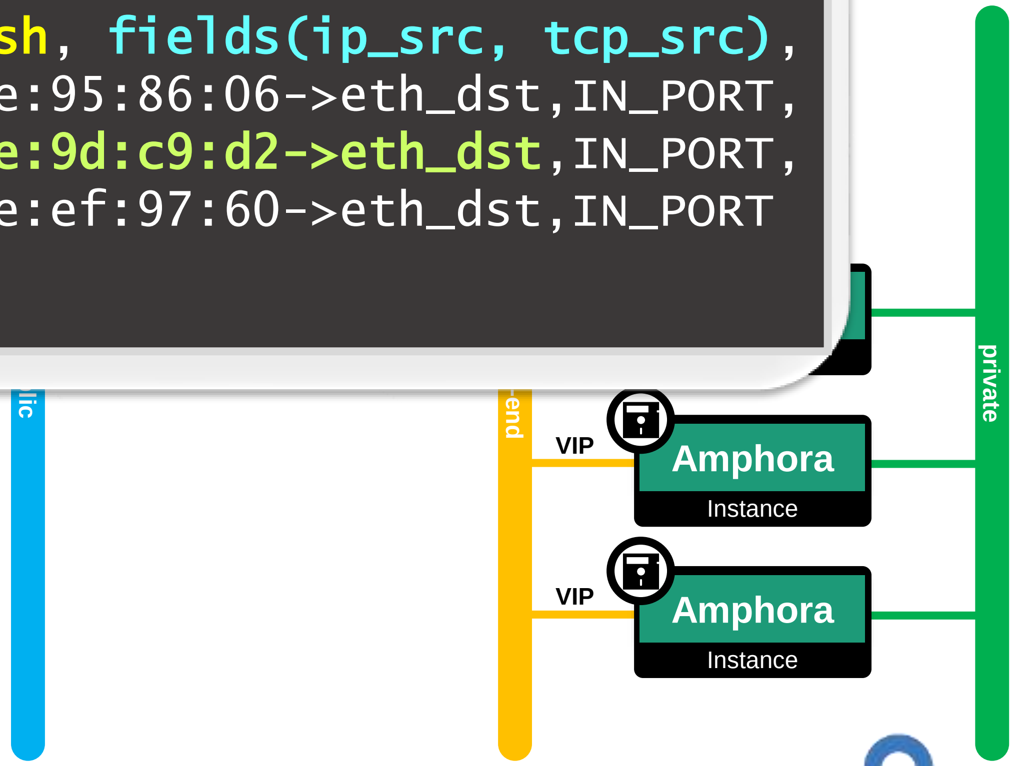
- 1-arm Direct Routing
  - Co-located on same LAN as Amphorae
  - L2 forwarding
    - Replace own MAC with MAC of Amphora
  - Direct Server Return
    - Return traffic goes directly to GW
  - Amphorae do not advertise VIP
- OpenFlow rules (using groups)
  - Select Amphora by hash of SrcIP:Port
- OVS VM
  - Can be any OpenFlow switch
  - Multi-tenant
  - No HA for now ☹



Our SDN SW Distributor

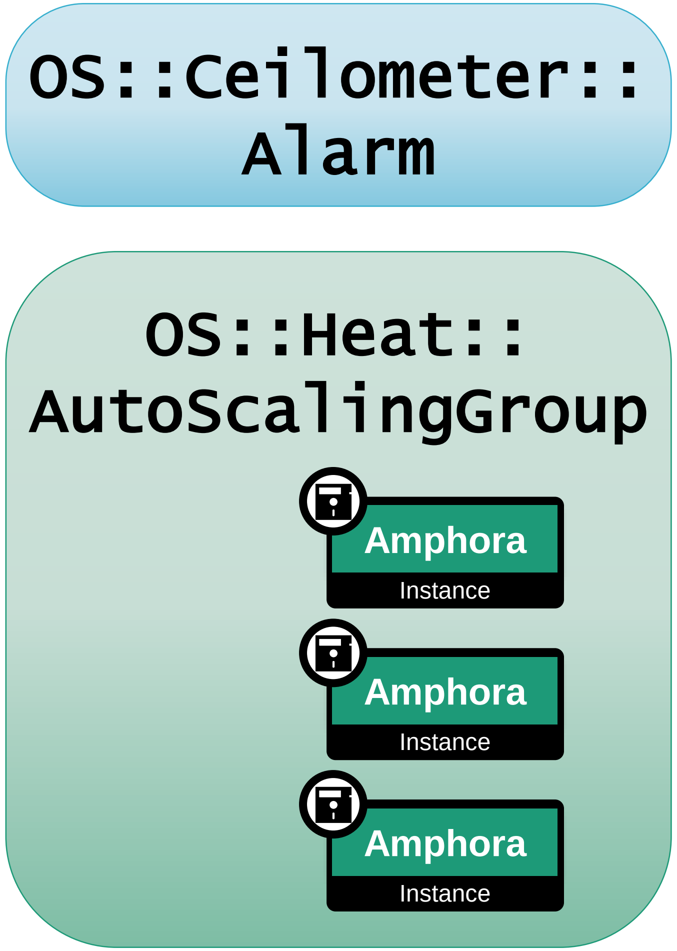
```
$ sudo ovs-ofctl -O OpenFlow 15 dump-groups br-data
OFPST_GROUP_DESC reply (OF1.5) (xid=0x2):
group_id=1, type=select, selection_method=hash, fields(ip_src, tcp_src),
bucket=bucket_id:0,actions=set_field:fa:16:3e:95:86:06->eth_dst,IN_PORT,
bucket=bucket_id:1,actions=set_field:fa:16:3e:9d:c9:d2->eth_dst,IN_PORT,
bucket=bucket_id:2,actions=set_field:fa:16:3e:ef:97:60->eth_dst,IN_PORT
$
```

- OpenFlow rules (using groups)
  - Select Amphora by hash of SrcIP:Port
- OVS VM
  - Can be any OpenFlow switch
  - Multi-tenant
  - No HA for now ☹



Elastic LB – Auto Scaling

- Amphorae pool is an auto-scale group
  - Use **Heat** to manage Amphora stack
    - Octavia Compute Driver (similar to Nova Driver)
    - Being replaced with a **Cluster Manager Driver**
  - Manage cluster of *N* Amphorae
    - Detect & replace failed Amphorae
    - Add/remove Amphorae when overloaded/underloaded
- Use **Ceilometer** to monitor Amphorae
- Octavia controller still does all the work….
  - Configure each Amphora
  - Monitor Amphorae at the application level
    - ***Do we need Ceilometer?***
  - Add/remove forwarding rules to **Distributor**
    - **Need to handle Affinity!**



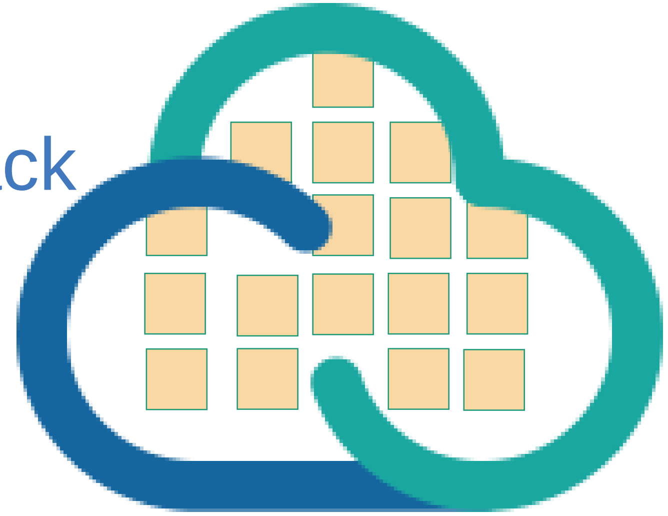
<u>Disclaimer</u>:

Not even a blue-print yet ☹

(but demo code ☺)

IBM Cloud

- Based on open standards
- Several cloud offerings running OpenStack operating system
- A large scale of workloads
- Benefit of load-balancer
  - High-availability
  - Performance
- Benefit of ELB
  - Load-balancer HA
  - Accommodate more workloads
  - Allow pay-per-use (cost efficient)

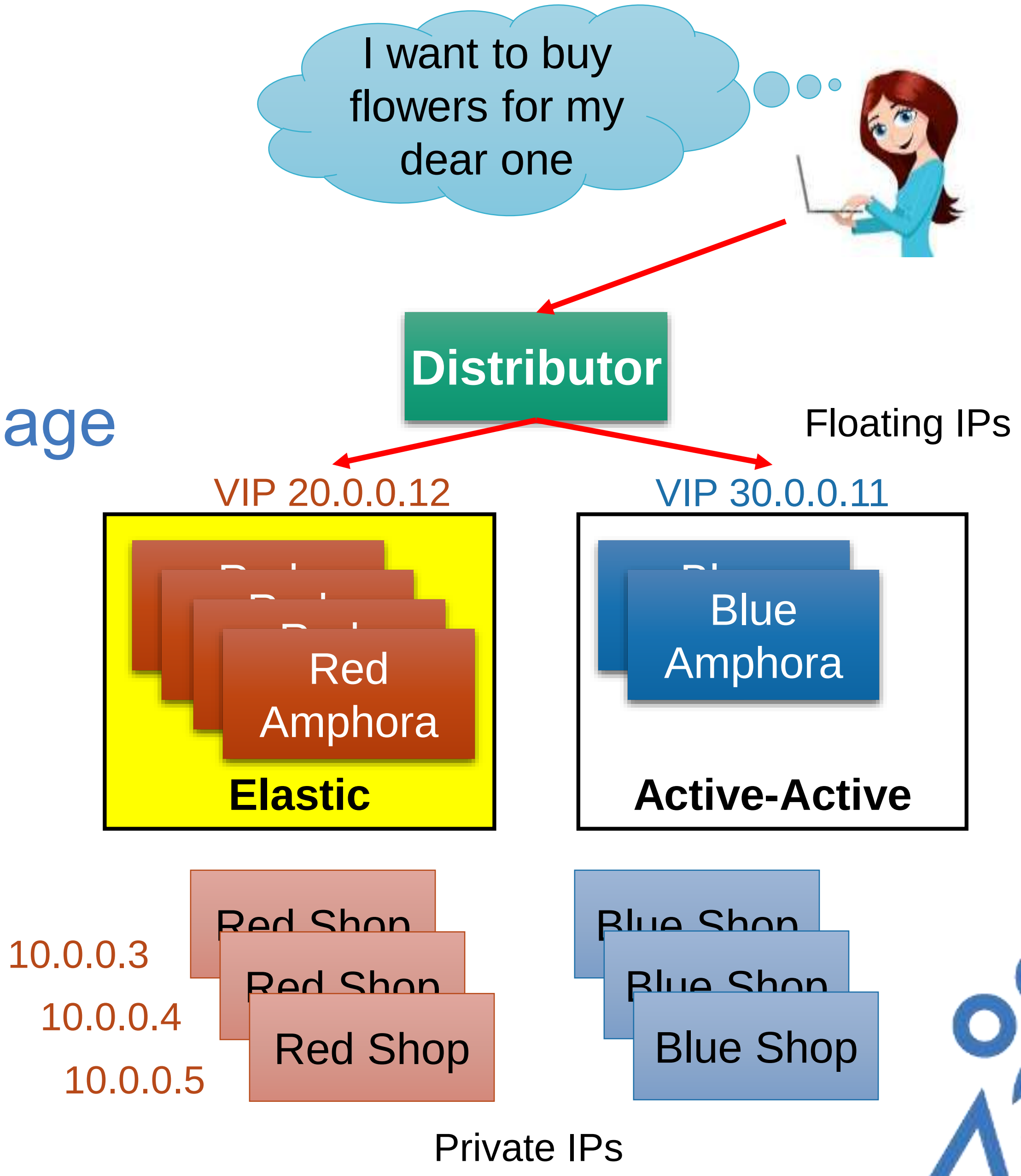


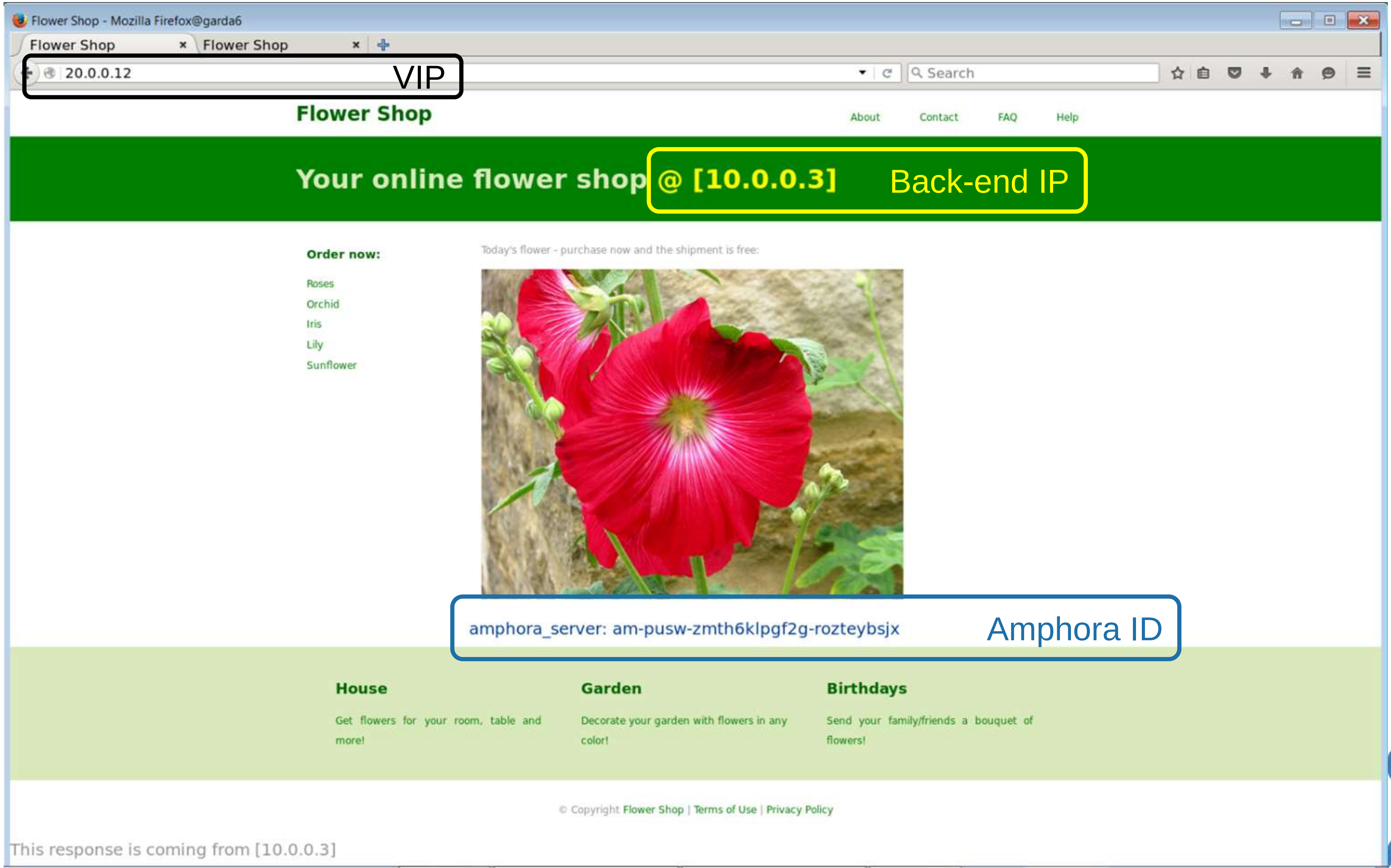
Demo (screenshots)

https://www.youtube.com/watch?v=l302AURPViI

Demo Story

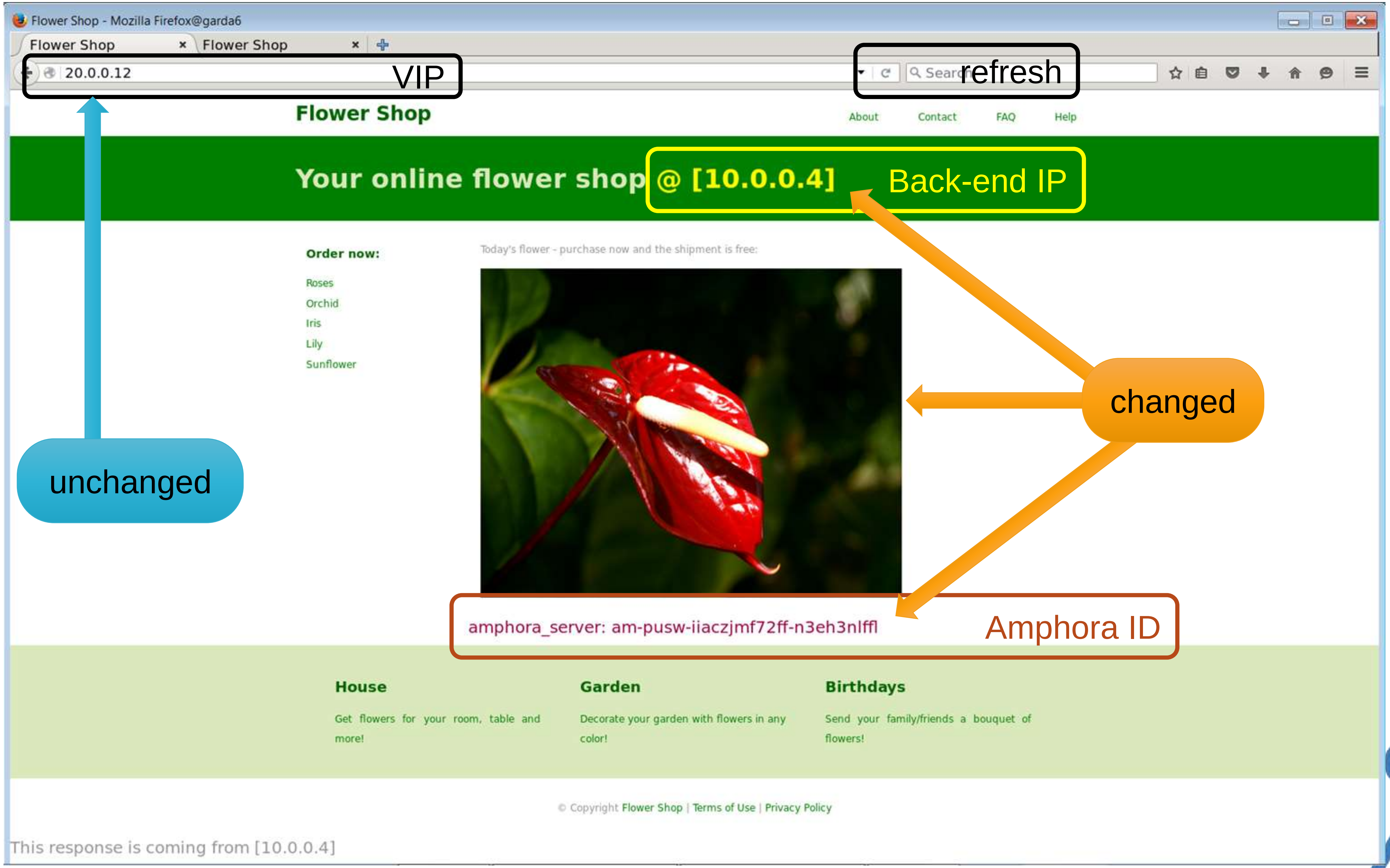
- Two web flower shops:
  - Red shop
  - Blue shop
- Each "shop" returns a flower page
  - Red or Blue flower
  - Different flower per back-end
  - Back-end IP inserted into page
- # of Amphorae doing LB for the red shop is auto-scaled by Heat (Ceilometer alarms)
- HAProxy injects Amphora ID
  - For demo purposes only

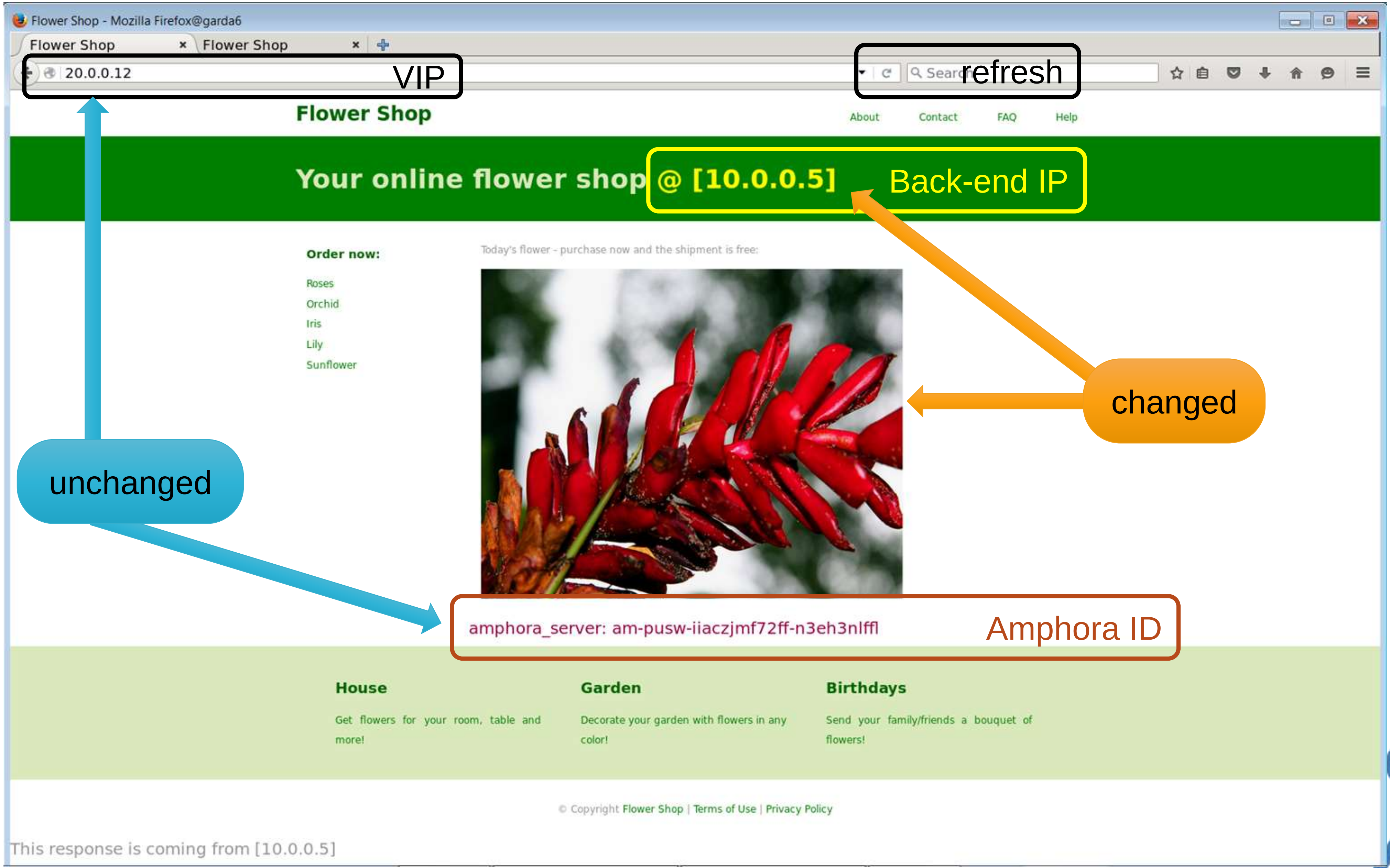



Flower Shop - Mozilla Firefox@garda6
Flower Shop
Flower Shop
20.0.0.12
VIP
Search
Flower Shop
About
Contact
FAQ
Help
Your online flower shop @ [10.0.0.3]
Back-end IP
Order now:
Roses
Orchid
Iris
Lily
Sunflower
Today's flower - purchase now and the shipment is free:
amphora_server: am-pusw-zmth6klpgf2g-rozteybsjx
Amphora ID
House
Get flowers for your room, table and more!
Garden
Decorate your garden with flowers in any color!
Birthdays
Send your family/friends a bouquet of flowers!
© Copyright Flower Shop | Terms of Use | Privacy Policy
This response is coming from [10.0.0.3]

IBM

Open by design™

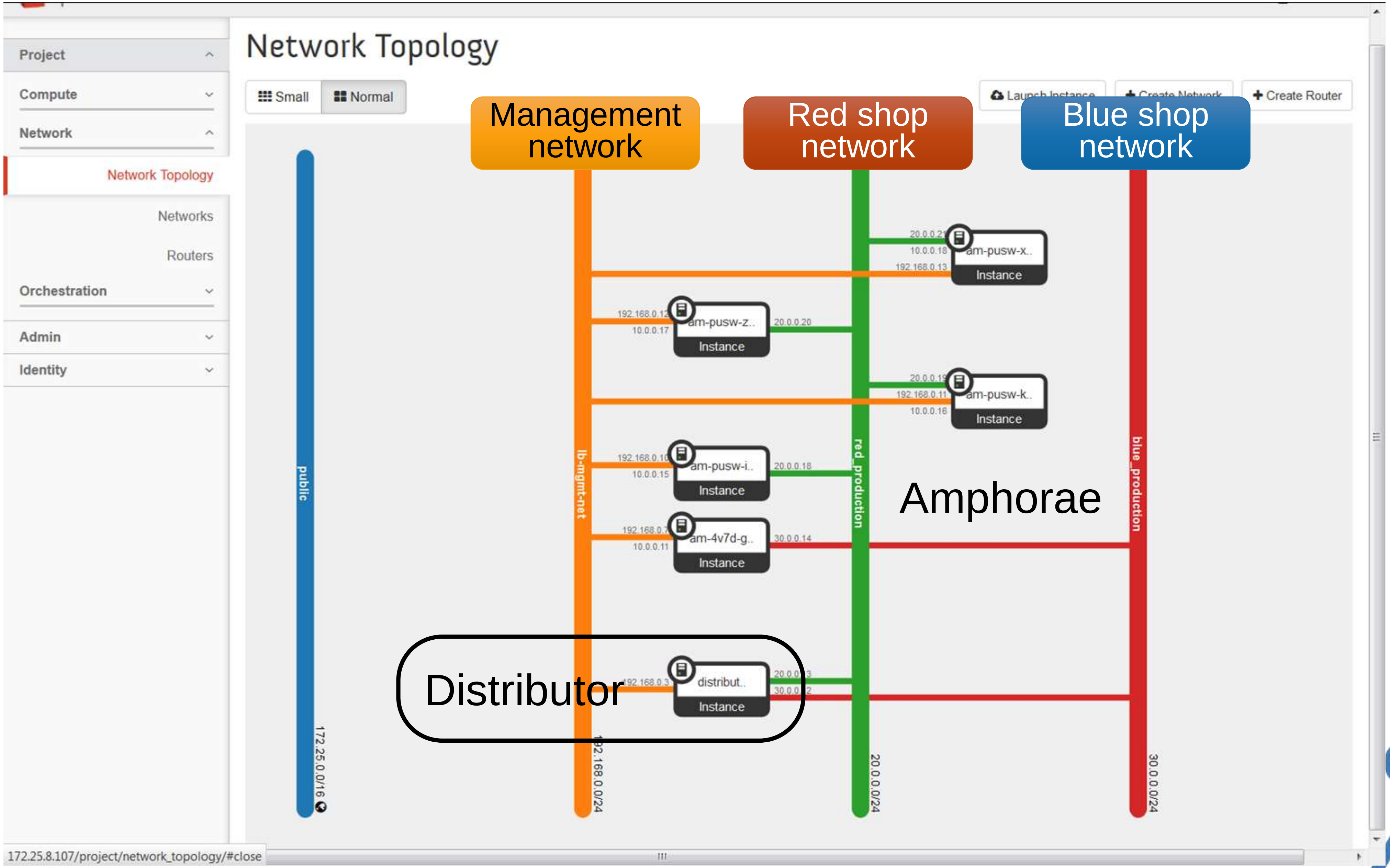


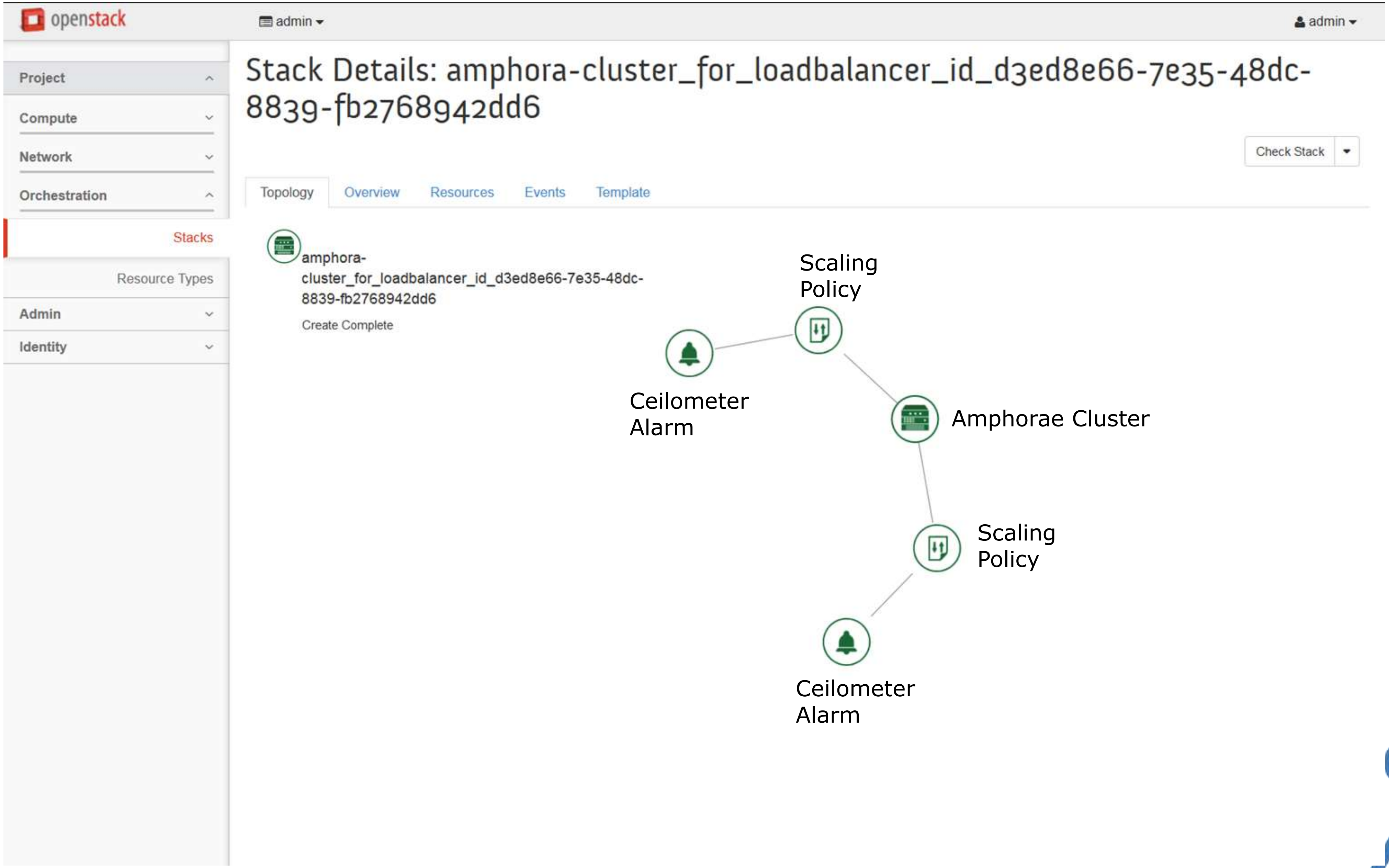


Flower Shop - Mozilla Firefox@garda6
Flower Shop
Flower Shop
20.0.0.12
VIP
refresh
Flower Shop
About
Contact
FAQ
Help
Your online flower shop @ [10.0.0.5]
Back-end IP
changed
unchanged
Order now:
Roses
Orchid
Iris
Lily
Sunflower
Today's flower - purchase now and the shipment is free:
amphora_server: am-pusw-iiaczjmf72ff-n3eh3nlffl
Amphora ID
House
Get flowers for your room, table and more!
Garden
Decorate your garden with flowers in any color!
Birthdays
Send your family/friends a bouquet of flowers!
© Copyright Flower Shop | Terms of Use | Privacy Policy
This response is coming from [10.0.0.5]

2 Elastic Load-Balancers

```
stack@garda6 [1937] ~/devstack/SharedRepository/CIL/tools  (master *)
$ neutron lbaas-loadbalancer-list
+--------------------------------------+---------+-------------+---------------------+----------+
| id                                   | name    | vip_address | provisioning_status | provider |
+--------------------------------------+---------+-------------+---------------------+----------+
| 6379f6f7-9c8b-459a-8469-30e5f08e7da5 | red_lb  | 20.0.0.12   | ACTIVE              | octavia  |
| d3ed8e66-7e35-48dc-8839-fb2768942dd6 | blue_lb | 30.0.0.11   | ACTIVE              | octavia  |
+--------------------------------------+---------+-------------+---------------------+----------+
```


Network Topology
Project
Compute
Network
Network Topology
Networks
Routers
Orchestration
Admin
Identity
Small
Normal
Launch Instance
Create Network
Create Router
Management network
Red shop network
Blue shop network
public
lb-mgmt-net
red_production
blue_production
172.25.0.0/16
192.168.0.0/24
20.0.0.0/24
30.0.0.0/24
am-pusw-x..
am-pusw-z..
am-pusw-k..
am-pusw-i..
am-4v7d-g..
distribut..
Instance
Amphorae
Distributor
172.25.8.107/project/network_topology/#close


openstack
admin
admin
Project
Compute
Network
Orchestration
Stacks
Resource Types
Admin
Identity
Stack Details: amphora-cluster_for_loadbalancer_id_d3ed8e66-7e35-48dc-8839-fb2768942dd6
Check Stack
Topology
Overview
Resources
Events
Template
amphora-cluster_for_loadbalancer_id_d3ed8e66-7e35-48dc-8839-fb2768942dd6
Create Complete
Scaling Policy
Ceilometer Alarm
Amphorae Cluster
Scaling Policy
Ceilometer Alarm

openstack | admin | admin

Project
Compute
Network
Orchestration
Stacks
Resource Types
Admin
Identity

Stack Details: amphora-cluster_for_loadbalancer_id_d3ed8e66-7e35-48dc-8839-fb2768942dd6

Check Stack

Topology | Overview | Resources | Events | Template

Stack Resource	Resource	Stack Resource Type	Date Updated	Status	Status Reason	
scaleup_policy	df754645f9de48499c641898adafb28d	OS::Heat::ScalingPolicy	2 hours, 10 minutes	Create Complete	state changed	**Scale-up Policy**
cpu_alarm_low	9cb13922-8860-42e8-b304-528ab3a6d1e7	OS::Ceilometer::Alarm	2 hours, 10 minutes	Create Complete	state changed	**Scale-down Alarm**
scaledown_policy	c79987f999d3494a9bc30cfacf928dde	OS::Heat::ScalingPolicy	2 hours, 10 minutes	Create Complete	state changed	**Scale-down Policy**
asg	8efcdcc4-7e5f-4c21-9a6a-fac6a8f50938	OS::Heat::AutoScalingGroup	2 hours, 10 minutes	Create Complete	state changed	**Amphorae Cluster**
cpu_alarm_high	b9e41970-f8ea-4e9c-9496-35aff64d84e0	OS::Ceilometer::Alarm	2 hours, 10 minutes	Create Complete	state changed	**Scale-up Alarm**

Displaying 5 items

```
stack@garda6 [1900] ~/devstack/SharedRepository/CIL/tools  (master *)
$ ceilometer alarm-show d18df35b-646d-456f-8d32-8d5aeccc51d0
+---------------------------+------------------------------------------------------------------------+
| Property                  | Value                                                                  |
+---------------------------+------------------------------------------------------------------------+
| alarm_actions             | ["http://172.25.8.77:8000/v1/signal/arn%3Aopenstack%3Aheat%3A%3Ac355a7a0 |
|                           | 21614562bb74b555a54445ab%3Astacks%2Famphora-                            |
|                           | cluster_for_loadbalancer_id_6379f6f7-9c8b-459a-8469-30e5f08e7da5        |
|                           | %2F96b90c9e-40b6-469a-859f-bbba989a76d4%2Fresources%2Fscaleup_policy?Tim |
|                           | estamp=2016-02-11T10%3A23%3A18Z&SignatureMethod=HmacSHA256&AWSAccessKeyI |
|                           | d=8f7b2c5a84fc4ff18e73b4c990d0c982&SignatureVersion=2&Signature=xHwGUQGh |
|                           | iOfnh5gkq8jWq%2BbmVmjsQeBjrG40w7T2bpU%3D"]                              |
| alarm_id                  | d18df35b-646d-456f-8d32-8d5aeccc51d0                                    |
| comparison_operator       | gt                                                                      |
| description               | Alarm when cpu_util is gt a avg of 40.0 over 120 seconds                |
| enabled                   | True                                                                    |
| evaluation_periods        | 1                                                                       |
| exclude_outliers          | False                                                                   |
| insufficient_data_actions | None                                                                    |
| meter_name                | cpu_util                                                                |
| name                      | amphora-cluster_for_loadbalancer_id_6379f6f7-9c8b-459a-8469-30e5f08e7da5 |
|                           | -cpu_alarm_high-hvxnvmyv2q2l                                            |
| ok_actions                | None                                                                    |
| period                    | 120                                                                     |
| project_id                | c355a7a021614562bb74b555a54445ab                                        |
| query                     | metadata.user_metadata.stack == amphora-                                |
|                           | cluster_for_loadbalancer_id_6379f6f7-9c8b-459a-8469-30e5f08e7da5        |
| repeat_actions            | True                                                                    |
| severity                  | low                                                                     |
| state                     | ok                                                                      |
| statistic                 | avg                                                                     |
| threshold                 | 40.0                                                                    |
| type                      | threshold                                                               |
| user_id                   | 73ed098273b24c73a23224f613219256                                        |
+---------------------------+------------------------------------------------------------------------+

stack@garda6 [1901] ~/devstack/SharedRepository/CIL/tools  (master *)
$
```

Alarm fires when avg of cpu_util > 40% over 2 minutes

Scale-up Ceilometer Alarm:

- statistic: avg
- comparison_operator: gt
- type: threshold
- threshold: 40.0
- period: 120
- state: unknown/ok/alarm
- alarm_actions: Scale-up URL

```
stack@garda6 [1902] ~/devstack/SharedRepository/CIL/tools  (master *)
$ ceilometer alarm-show c5ac9295-8835-4aa3-9706-82be0f3a1785
+---------------------------+----------------------------------------------------------------------------+
| Property                  | Value                                                                      |
+---------------------------+----------------------------------------------------------------------------+
| alarm_actions             | ["http://172.25.8.77:8000/v1/signal/arn%3Aopenstack%3Aheat%3A%3Ac355a7a0 |
|                           | 21614562bb74b555a54445ab%3Astacks%2Famphora-                               |
|                           | cluster_for_loadbalancer_id_6379f6f7-9c8b-459a-8469-30e5f08e7da5           |
|                           | %2F96b90c9e-40b6-469a-859f-bbba989a76d4%2Fresources%2Fscaledown_policy?T   |
|                           | imestamp=2016-02-11T10%3A23%3A18Z&SignatureMethod=HmacSHA256&AWSAccessKe   |
|                           | yId=0b05dafa7d9a4b01bea766e4ceb5346b&SignatureVersion=2&Signature=cyWpHs   |
|                           | 8SPlSxe7Ea1z5y9DEOjE3uQDqvYfnkEHb%2FwXI%3D"]                               |
| alarm_id                  | c5ac9295-8835-4aa3-9706-82be0f3a1785                                       |
| comparison_operator       | lt                                                                         |
| description               | Alarm when cpu_util is lt a avg of 10.0 over 120 seconds                   |
| enabled                   | True                                                                       |
| evaluation_periods        | 1                                                                          |
| exclude_outliers          | False                                                                      |
| insufficient_data_actions | None                                                                       |
| meter_name                | cpu_util                                                                   |
| name                      | amphora-cluster_for_loadbalancer_id_6379f6f7-9c8b-459a-8469-30e5f08e7da5   |
|                           | -cpu_alarm_low-cn4b3y6t4kgk                                                |
| ok_actions                | None                                                                       |
| period                    | 120                                                                        |
| project_id                | c355a7a021614562bb74b555a54445ab                                           |
| query                     | metadata.user_metadata.stack == amphora-                                   |
|                           | cluster_for_loadbalancer_id_6379f6f7-9c8b-459a-8469-30e5f08e7da5           |
| repeat_actions            | True                                                                       |
| severity                  | low                                                                        |
| state                     | alarm                                                                      |
| statistic                 | avg                                                                        |
| threshold                 | 10.0                                                                       |
| type                      | threshold                                                                  |
| user_id                   | 73ed098273b24c73a23224f613219256                                           |
+---------------------------+----------------------------------------------------------------------------+
stack@garda6 [1903] ~/devstack/SharedRepository/CIL/tools  (master *)
$
```

Alarm fires when avg of cpu_util < 10% over 2 minutes

Scale-down Ceilometer Alarm:

- statistic: avg
- comparison_operator: lt
- type: threshold
- threshold: 10.0
- period: 120
- state: unknown/ok/alarm
- alarm_actions: Scale-dn URL

IBM

Open by design™

Start the Stress…

```
Creating ping stress for 600 seconds against 20.0.0.12
ARPING to 20.0.0.12 from 20.0.0.11 via eth0
Unicast reply from 20.0.0.12 [fa:16:3e:cf:67:6f] 6.831ms
Sent 1 probe(s) (1 broadcast(s))
Received 1 replies (0 request(s), 0 broadcast(s))
Creating high stress for [1] more seconds (using port 13980)
Waiting (sleeping) for 600 seconds
Waiting (no stress) for [1] more seconds
```

Elastic Load-Balancers Under Stress

```
Info from Ceilometer

+--------------------------------------+----------+-------+---------------+------+---------------------
| Resource ID                          | Name     | Type  | Volume        | Unit | Timestamp
+--------------------------------------+----------+-------+---------------+------+---------------------
| 2f13d7da-c8b6-404e-b969-2caa8f580d0e | cpu_util | gauge | 45.2290716999 | %    | 2016-02-11T12:50:21.
| 11901a40-0fb8-4c1a-b6d8-e347623a15e3 | cpu_util | gauge | 32.6827604832 | %    | 2016-02-11T12:50:21.
| 95de6adc-f3ab-4d57-9b8c-2e3b7c238063 | cpu_util | gauge | 53.2962528019 | %    | 2016-02-11T12:50:21.
```

cpu_util > 40% (as specified in the alarm) – scale-up alarm triggered

A new Amphora VM will be added to the cluster (by Heat Engine)

Elastic Load-Balancers Stress Free

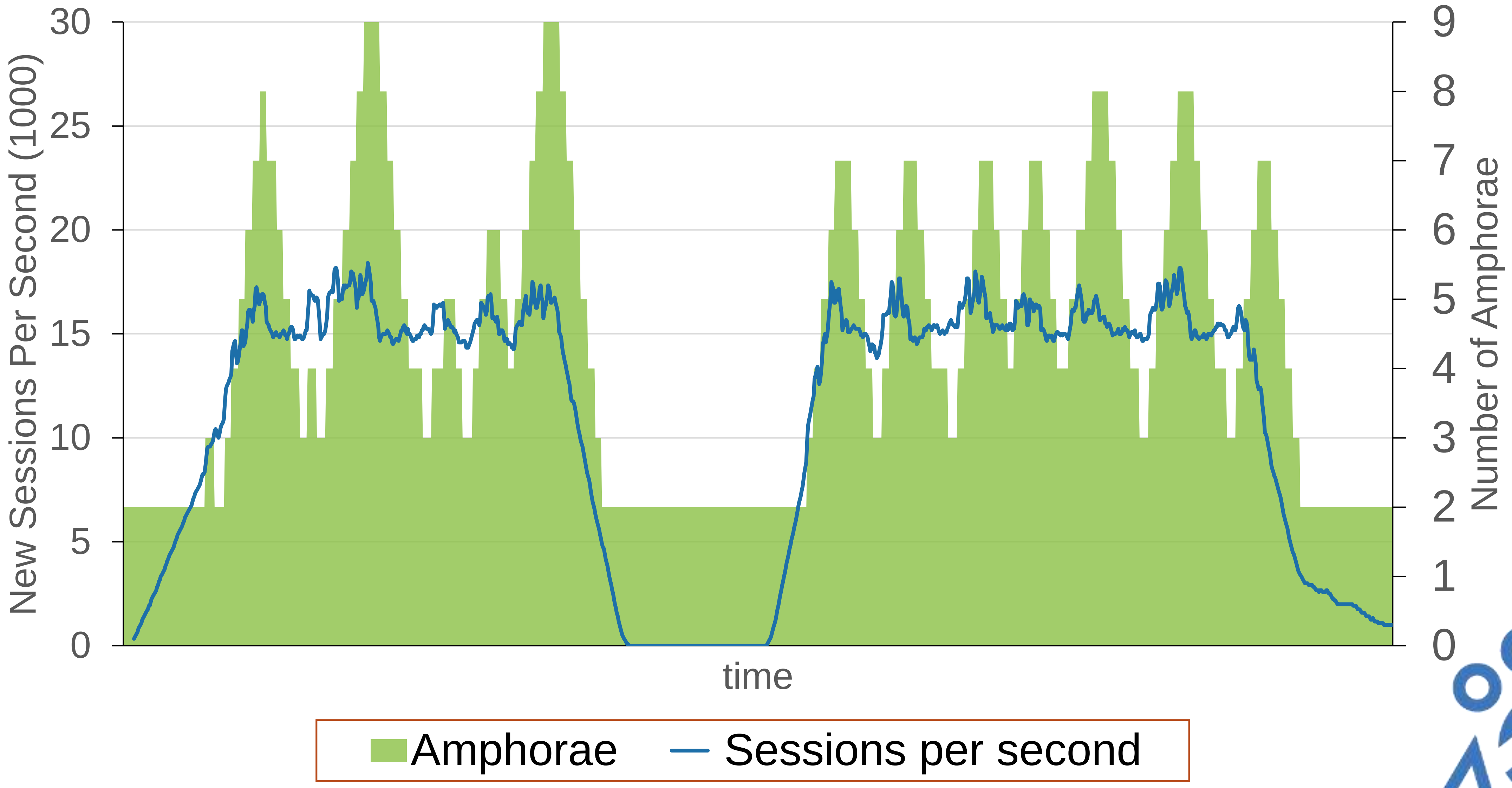
```
Info from Ceilometer

+--------------------------------------+----------+-------+----------------+------+----------------------------+
| Resource ID                          | Name     | Type  | Volume         | Unit | Timestamp                  |
+--------------------------------------+----------+-------+----------------+------+----------------------------+
| 2f13d7da-c8b6-404e-b969-2caa8f580d0e | cpu_util | gauge | 6.63408730771  | %    | 2016-02-11T12:13:21.719179 |
| 79232dcc-a9f1-4b56-b901-02c23cd6f4b8 | cpu_util | gauge | 7.01812390202  | %    | 2016-02-11T12:13:21.699397 |
```

cpu_util < 10% (as specified in the alarm) – scale-down alarm triggered

An existing Amphora VM will be removed from the cluster (by Heat Engine)

Sample Run (simulated HTTPS load)



Equal Balancing at Each Level

```
stack@garda6 [1894] ~/devstack/SharedRepository/CIL/tools  (master *)
$ for i in {1..100}; do curl -I 20.0.0.12 2>/dev/null | grep '^backend-server'; done | sort  | uniq -c
     34 backend-server: 10.0.0.3
     33 backend-server: 10.0.0.4
     33 backend-server: 10.0.0.5

stack@garda6 [1895] ~/devstack/SharedRepository/CIL/tools  (master *)
$ for i in {1..100}; do curl -I 20.0.0.12 2>/dev/null | grep '^amphora'; done | sort  | uniq -c
     20 amphora_server: am-pusw-2hpxzzgsnomu-b5ewlqrj2f
     22 amphora_server: am-pusw-mikvap7iockv-x6zxxpdpb7
     31 amphora_server: am-pusw-xlzy4aesdwxj-6qrbagv2dz
     27 amphora_server: am-pusw-zmth6klpgf2g-rozteybsjx

stack@garda6 [1896] ~/devstack/SharedRepository/CIL/tools  (master *)
$ _
```

End of Demo

https://www.youtube.com/watch?v=l302AURPViI

Amphora Containers

- Lower cost per LB instance
  - Containers use less resources
  - Can be packed tighter
- Container less powerful
  - Horizontal scaling allows large workloads
- Faster creation
  - No need for +1 ?
- Better availability
  - Larger N → better spread
  - Container migration

Thank you.

Questions?

Blueprints: (active-active-topology, active-active-distributor)
https://review.openstack.org/#/c/234639