

“Container Automate Scaling with HPA (K8S)”

By Praparn Luengphoonlap
Email: eva10409@gmail.com

“Swarm & K8S” We are one big community

Agenda

- The brief history of Docker and K8S
- Horizontal Pod Autoscaler Concept
- Demo Session
 - Enable kubernetest dashboard/metric server
 - Demo: Deploy simple python webpage
 - Demo: Generate load and HPA monitor

The brief history of Docker & K8S

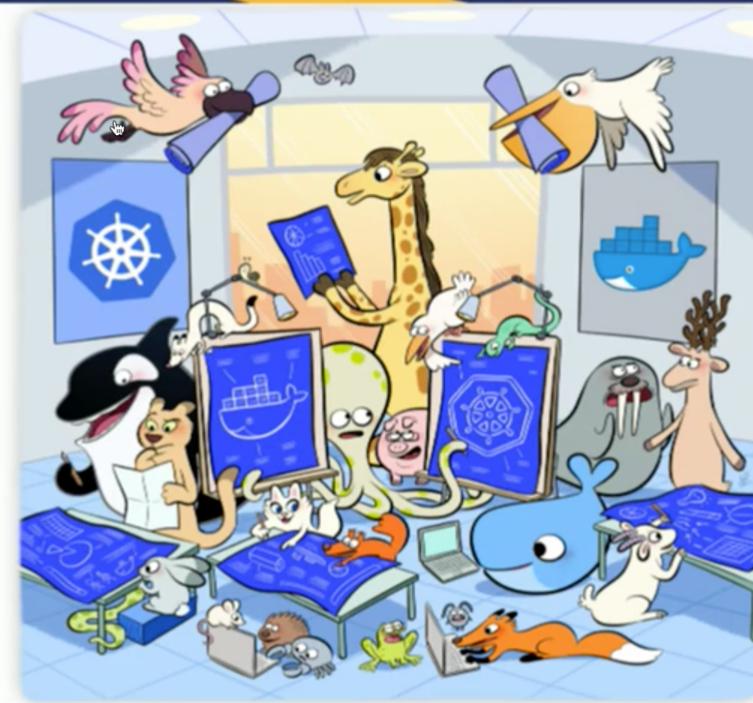


"Swarm & K8S" We are one big community

The brief history of Docker & K8S

- “We are one big community”

WE ARE
ONE BIG
COMMUNITY



“Swarm & K8S” We are one big community

The brief history of Docker & K8S

- Official support orchestrator both K8S and Swarm

Docker with Swarm and Kubernetes

1 →

The best enterprise
container security and
management

Docker Enterprise Edition

Docker Community Edition



2 ←

The best container
development workflow

3 →

Native Kubernetes
integration provides full
ecosystem
compatibility

4 ←

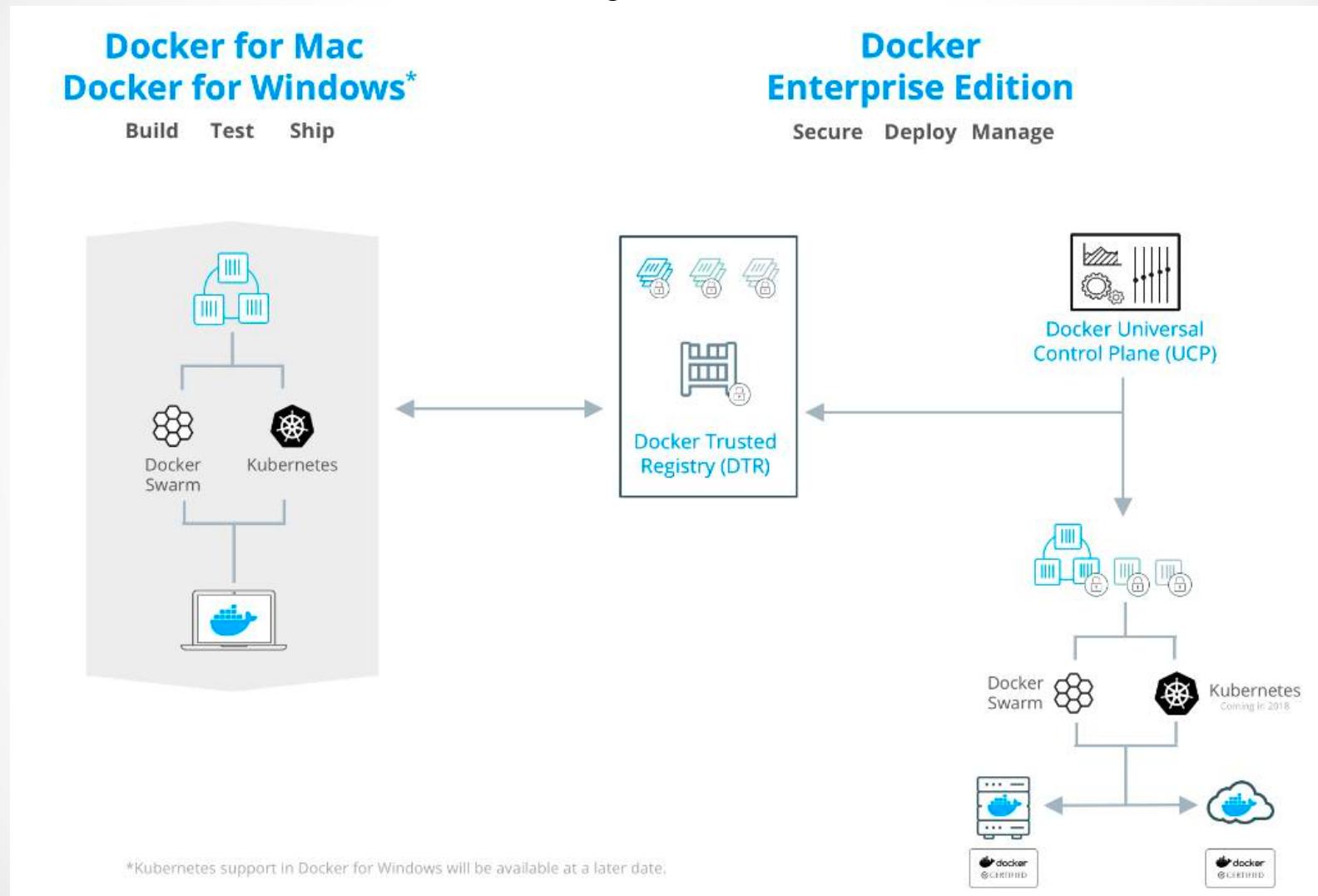
Industry-standard
container runtime

dockercon¹⁷EU

“Swarm & K8S” We are one big community

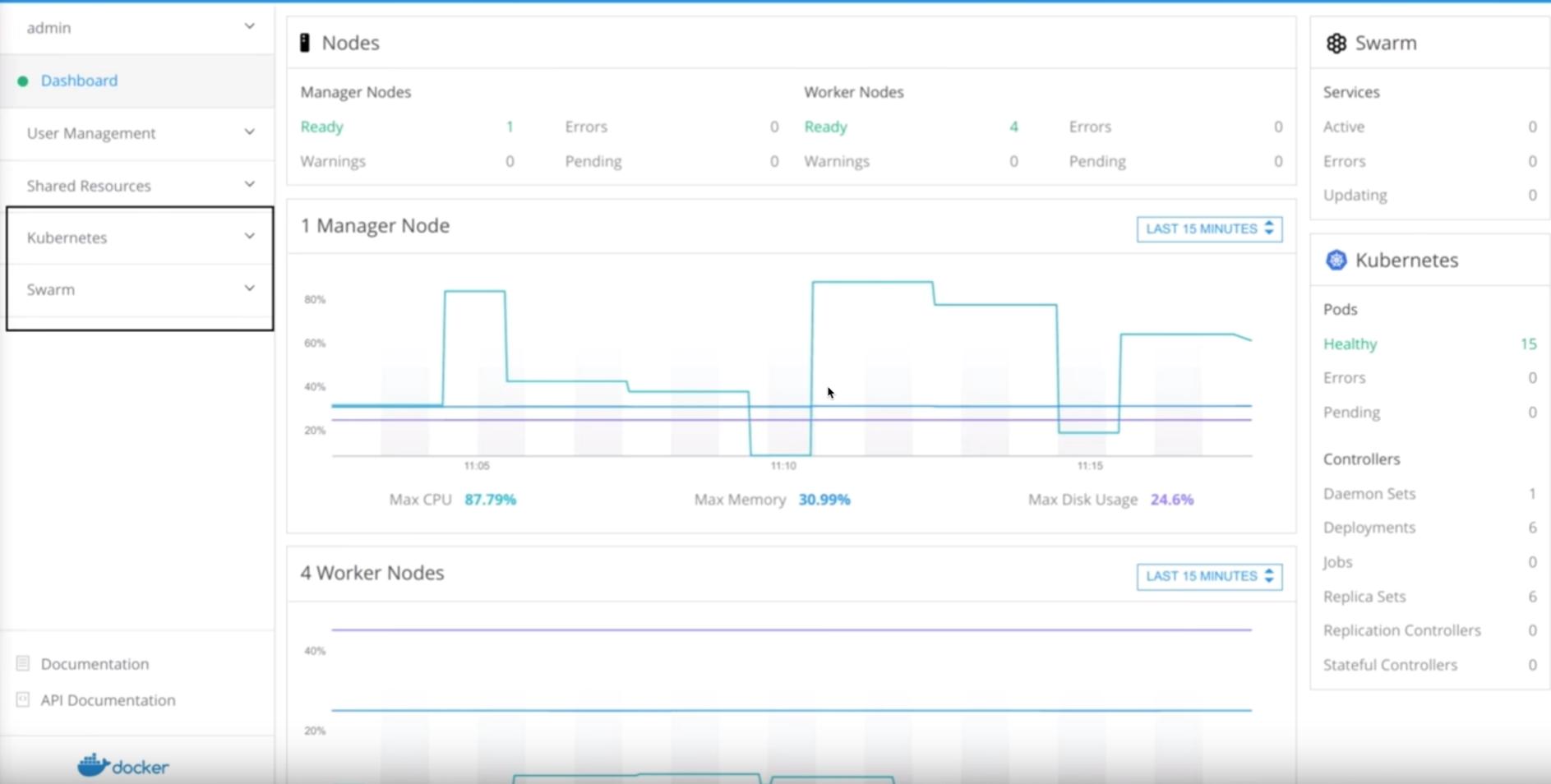
dockercon¹⁷EU

The brief history of Docker & K8S



“Swarm & K8S” We are one big community

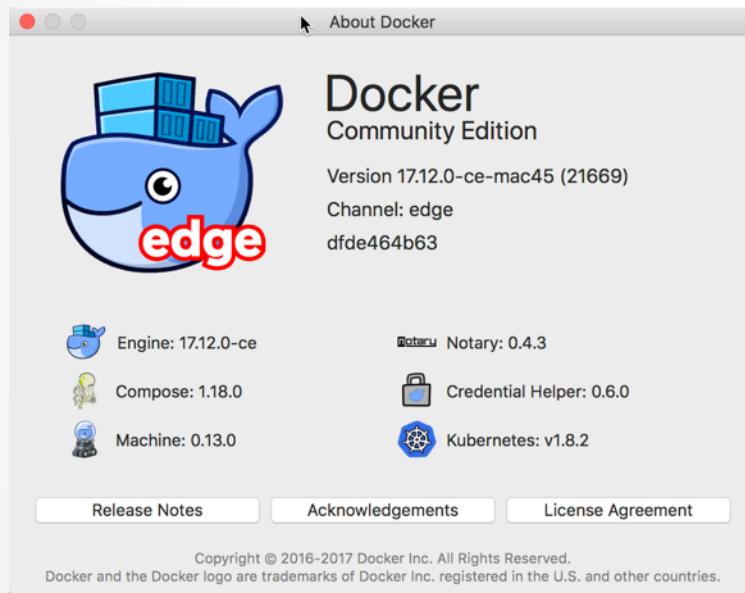
The brief history of Docker & K8S



“Swarm & K8S” We are one big community

The brief history of Docker & K8S

- Docker Community Edition (CE)
 - Docker for Windows (**Available Now with Edge Edition !!!**)
 - Docker for MAC (**Available Now with Edge Edition !!!**)



"Swarm & K8S" We are one big community

The brief history of Docker & K8S

What are the best Docker orchestration tools?

11

OPTIONS CONSIDERED

77

RECOMMENDATIONS

Mar 13, 2018

LAST UPDATED

THE BEST 1 OF 11 OPTIONS i WHY?

11 Options Considered

Price

Last Updated



THE BEST
Kubernetes

GET IT HERE

Mar 13, 2018



Docker Swarm

GET IT HERE

Dec 11, 2017



Docker Compose

GET IT HERE

Sep 23, 2017



Nomad

GET IT HERE

Sep 23, 2017



OpenShift

GET IT HERE

Jan 22, 2018

See Full List

RELATED QUESTIONS

What are the best power user tools for macOS?

What are the best continuous integration tools?

What are the best developer tools for Mac OSX?

What are the best software tools for live streaming?

What are the best log aggregation & monitoring tools?

What are the best mockup and wireframing tools for websites?

What are the best diff tools for Git?

What are the best mind mapping tools?

What are the best 2D animation tools for game development?

What are the best power user tools for Windows?

Reference: <https://www.slant.co/topics/3929/~docker-orchestration-tools>

“Swarm & K8S” We are one big community



Horizontal Pods Autoscaler

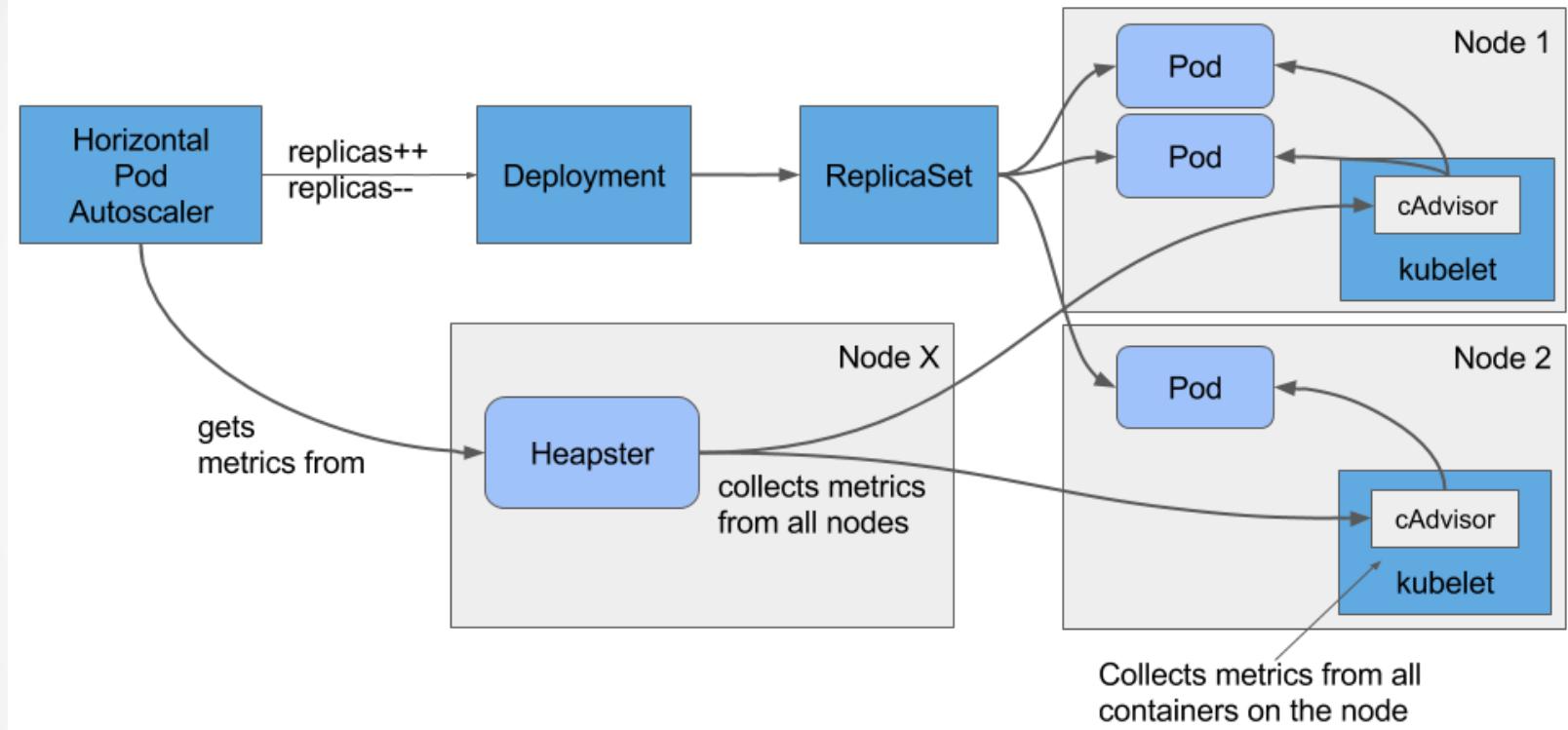


“Swarm & K8S” We are one big community

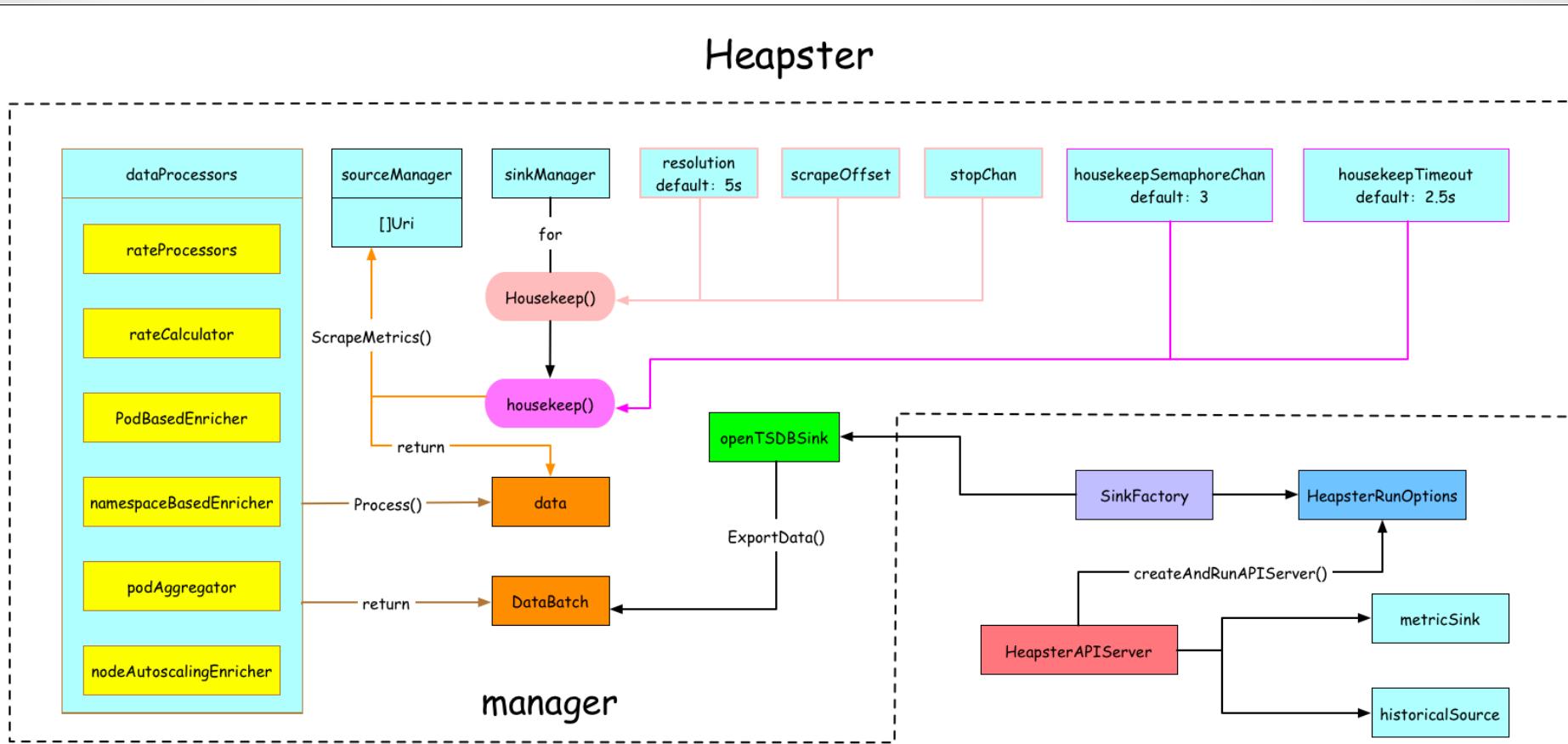
Horizontal Pods Autoscaler

- Normally application (Container inside) was designed base on requirement and performance test result
 - Ex:
 - Spec: 1 Container: 200 Concurrent API Call/Sec
 - Traffic Average: 5000 Concurrent API Call/Sec
 - Total: ~ 25-30 Containers (+ overcap 15%)
- Orchestrator was support scaling as request (Swarm/K8S/Mesos etc)
- But...
 - Why we need to scale bigger farm when empty workload ?
 - Are we confidence that all workload was handled ?
 - How can we handle workload spike ?
- HPA will response for monitor workload on Pods (Now base on CPU) and automatic trigger deployment to scale-up application

Horizontal Pods Autoscaler



Horizontal Pods Autoscaler



© Jimmy Song <https://github.com/rootsongjc/kubernetes-handbook>

Horizontal Pods Autoscaler

[GitHub, Inc. \[US\] | https://github.com/juju-solutions/bundle-canonical-kubernetes/issues/484](https://github.com/juju-solutions/bundle-canonical-kubernetes/issues/484)

This repository Search Pull requests Issues Marketplace Explore Watch 16 Star 57 Fork 23

juju-solutions / bundle-canonical-kubernetes Issues 124 Pull requests 1 Projects 1 Wiki Insights New issue

Autoscaler with CDK v1.9 #484

Open ktsakalozos opened this issue on Jan 31 · 1 comment

ktsakalozos commented on Jan 31 Member +
Starting with v1.9 HPA uses new resource metrics API that is not available in CDK out of the box. If you want to use autoscaler you should do a:

\$ juju config kubernetes-master controller-manager-extra-args="--horizontal-pod-autoscaler-us

Based on the discussion here [kubernetes/kubernetes#57673](#) we can either set the --horizontal-pod-autoscaler-use-rest-clients=false or deploy this Metrics server: <https://kubernetes.io/docs/tasks/debug-application-cluster/core-metrics-pipeline/>

Issue reported here as well: [kubernetes/kubernetes#57996](#)

hyperbolic2346 commented on Jan 31 Member +
The metrics server is deployed by default in kube-up. I would think we should deploy it as well.

wwwtyro added this to Feature Request in CDK on Feb 1

hyperbolic2346 referenced this issue in [kubernetes/kubernetes](#) on Feb 22 Adding metrics server #60174 Open

Assignees
No one assigned

Labels
None yet

Projects
Feature Request in CDK

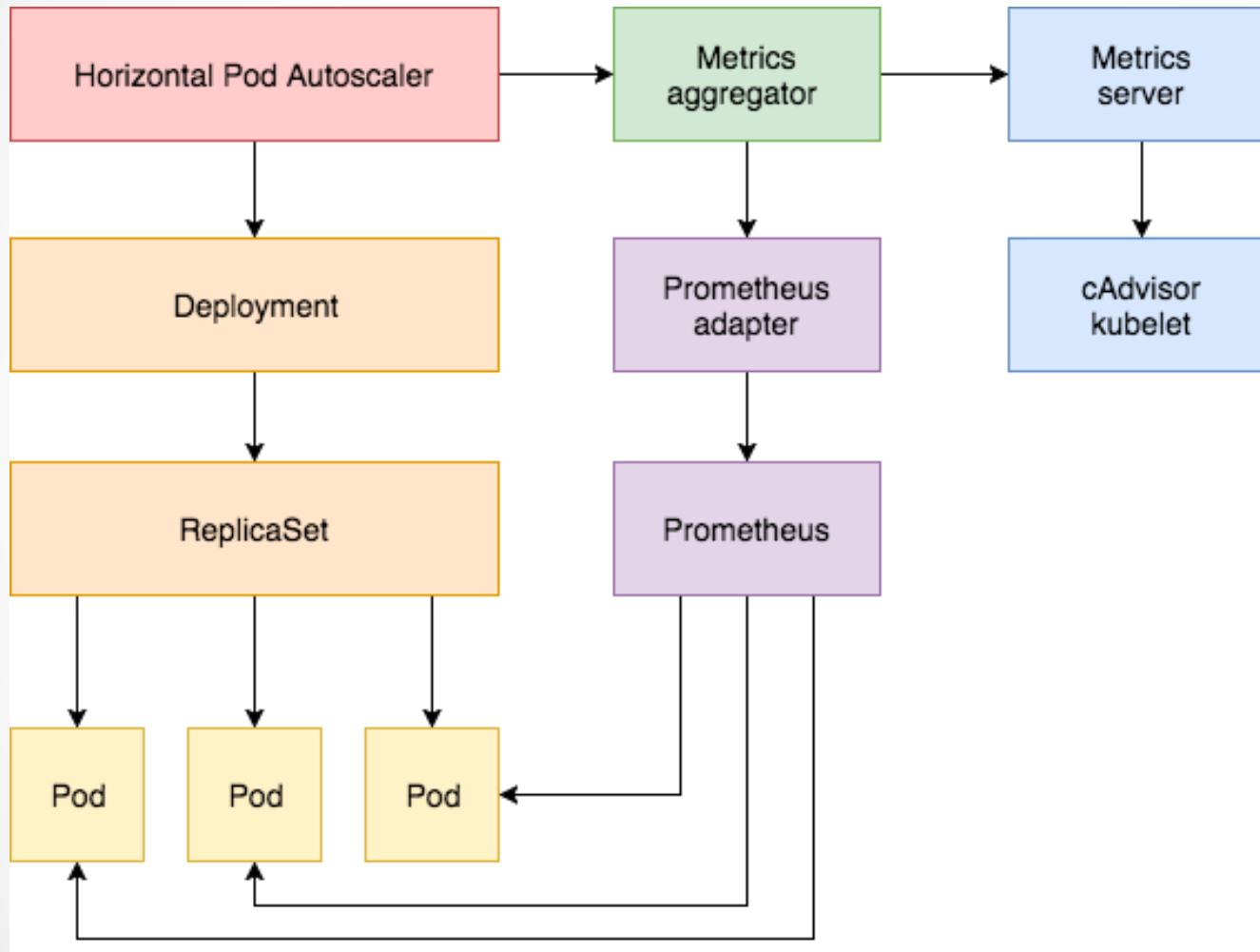
Milestone
No milestone

Notifications
Subscribe
You're not receiving notifications from this thread.

2 participants

"Swarm & K8S" We are one big community

Horizontal Pods Autoscaler

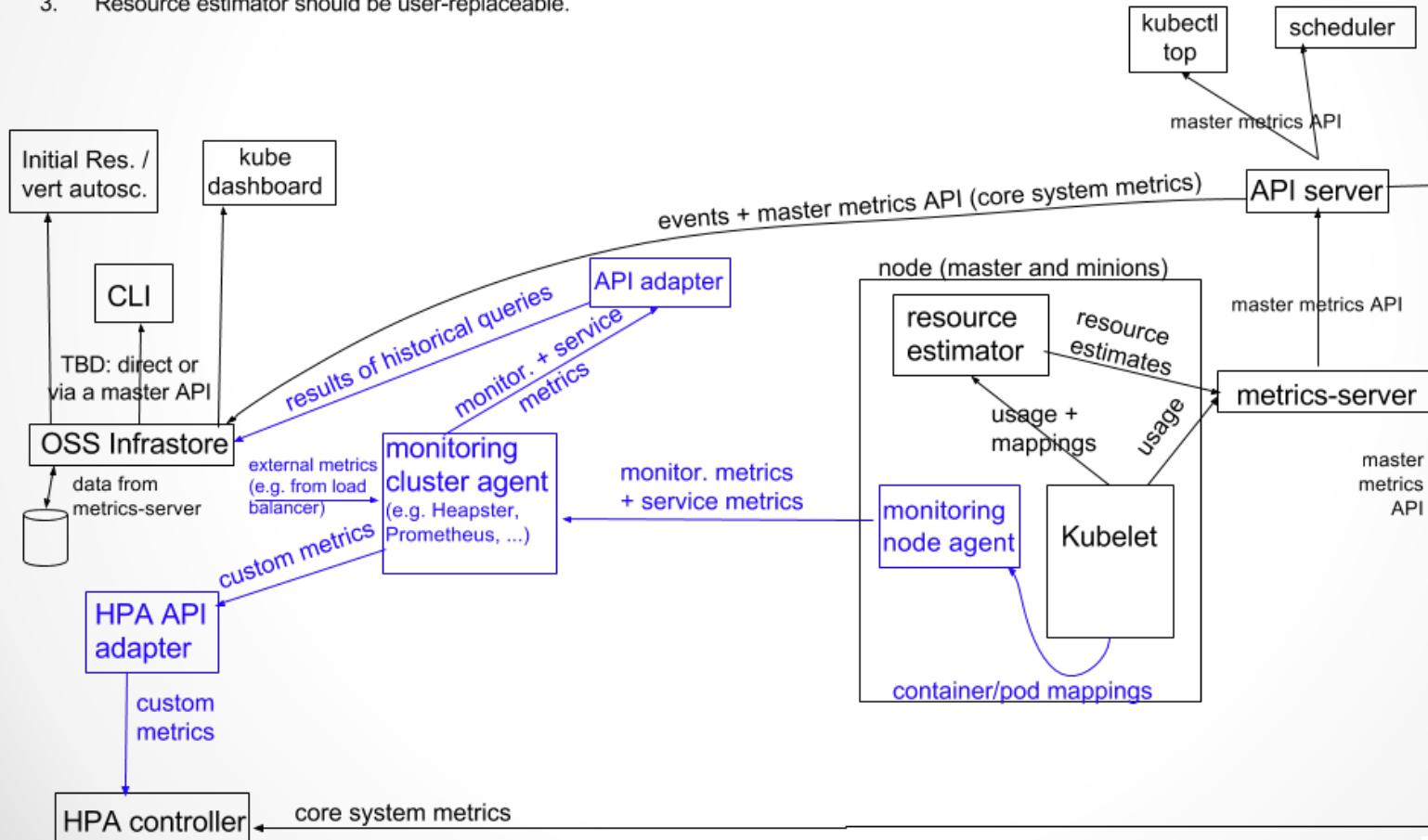


Horizontal Pods Autoscaler

Monitoring architecture proposal: OSS
(arrows show direction of metrics flow)

Notes

1. Arrows show direction of metrics flow.
2. **Monitoring pipeline is in blue**. It is user-supplied and optional.
3. Resource estimator should be user-replaceable.



Horizontal Pods Autoscaler

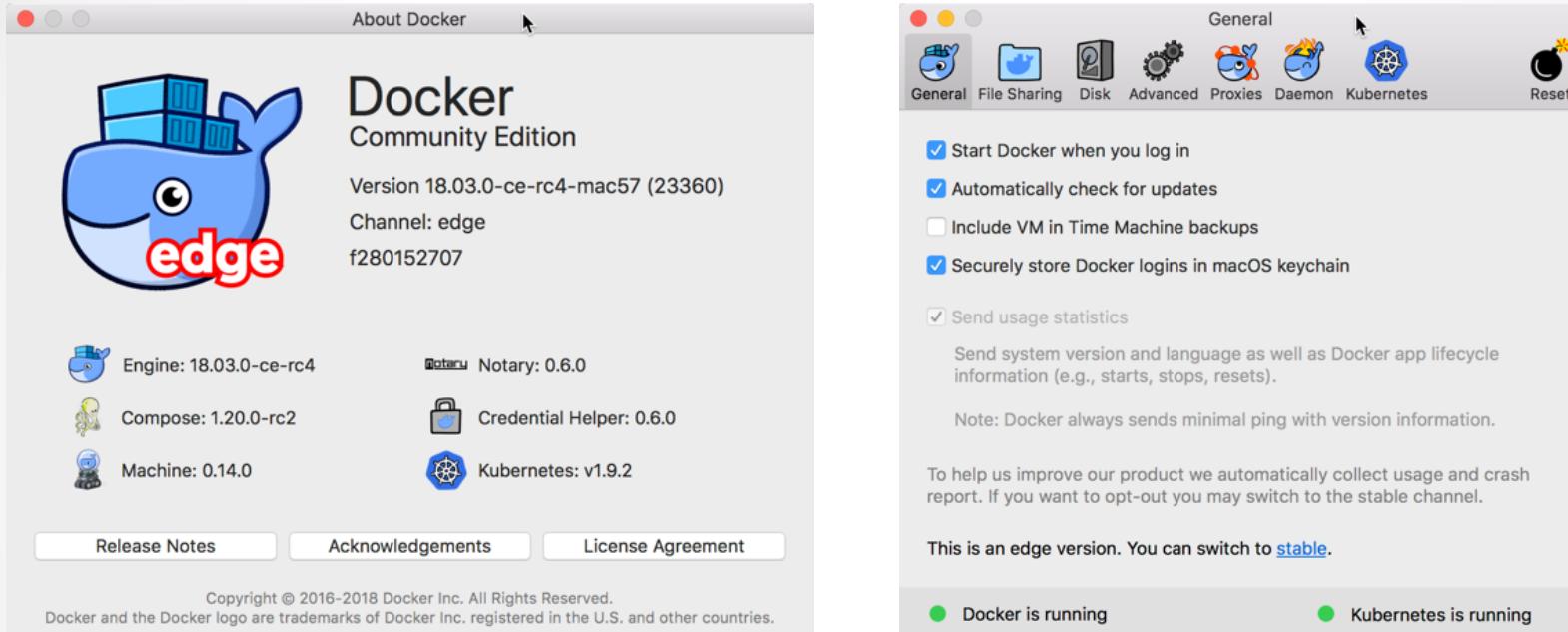
- Example: Deployment for python

```
1  apiVersion: v1
2  kind: Service
3  metadata:
4    name: webtest
5    labels:
6      name: web
7      owner: Praparn_L
8      version: "1.0"
9      module: WebServer
10     environment: development
11 spec:
12   selector:
13     name: web
14     owner: Praparn_L
15     version: "1.0"
16     module: WebServer
17     environment: development
18
19   type: NodePort
20   ports:
21     - port: 5000
22       name: http
23       targetPort: 5000
24       protocol: TCP
25       nodePort: 32500
```

```
27  apiVersion: apps/v1
28  kind: Deployment
29  metadata:
30    name: webtest
31    labels:
32      name: web
33      owner: Praparn_L
34      version: "1.0"
35      module: WebServer
36      environment: development
37 spec:
38   replicas: 1
39   selector:
40     matchLabels:
41       name: web
42       owner: Praparn_L
43       version: "1.0"
44       module: WebServer
45       environment: development
46 template:
47   metadata:
48     labels:
49       name: web
50       owner: Praparn_L
51       version: "1.0"
52       module: WebServer
53       environment: development
54 spec:
55   containers:
56     - name: webtest
57       image: labdocker/cluster:webservicelite_v1
58       resources:
59         requests:
60           cpu: "200m"
61       ports:
62         - containerPort: 5000
63           protocol: TCP
```

Horizontal Pods Autoscaler

- Example: Deployment for python



```
[praparns-MBP% kubectl config use-context docker-for-desktop
Switched to context "docker-for-desktop".
[praparns-MBP% kubectl config get-contexts
CURRENT      NAME                  CLUSTER          AUTHINFO        NAMESPACES
*   minikube     minikube           minikube        minikube
*   docker-for-desktop   docker-for-desktop-cluster  docker-for-desktop
      first       first            default-cluster
      local
praparns-MBP%
```

"Swarm & K8S" We are one big community

Horizontal Pods Autoscaler

- Example: Deployment for python

```
praparns-MacBook-Pro% kubectl create -f https://raw.githubusercontent.com/praparn/docker_meetup_20180329/master/webtest_deploy_hpa.yml
service "webtest" created
deployment "webtest" created
praparns-MacBook-Pro% kubectl get deployment/webtest -o wide
[    kubectl get svc/webtest -o wide
NAME      DESIRED   CURRENT   UP-TO-DATE   AVAILABLE   AGE
webtest   1          1          1           1           <invalid>
          environment=WebServer, name=web, owner=Praparn_L, version=1.0
NAME      TYPE        CLUSTER-IP   EXTERNAL-IP   PORT(S)      AGE
webtest   NodePort   10.108.136.22 <none>        5000:32500/TCP <invalid>
          environment=development, module=WebServer, name=web, owner=Praparn_L, version=1.0
praparns-MacBook-Pro%
```

The screenshot shows a web browser window with the following details:

- Address Bar:** 127.0.0.1:32500
- Toolbar:** Includes icons for back, forward, search, and refresh.
- Bookmark Bar:** Shows various bookmarks including "Apps", "NMac Ked - Mac OS...", "Medium", "jenkins", "vagrant", "Mesos", "Vue", "docker", "NGINX", "Taiwan", "Kubernetes", and "Other Bookmarks".
- Content Area:** Displays the text "Welcome Page from Container Python Lab Web Version 1.00".
- Status Bar:** Checkpoint Date/Time: Mon Mar 26 16:43:13 2018

"Swarm & K8S" We are one big community



Horizontal Pods Autoscaler

- Example: Deployment for python

```
[praparns-MacBook-Pro% kubectl autoscale deployment/webtest --min=1 --max=10 --cpu-percent=10
deployment "webtest" autoscaled
[praparns-MacBook-Pro% kubectl get hpa
NAME      REFERENCE      TARGETS      MINPODS      MAXPODS      REPLICAS      AGE
webtest   Deployment/webtest <unknown> / 10%    1            10           0            <invalid>
[praparns-MacBook-Pro% kubectl get hpa
NAME      REFERENCE      TARGETS      MINPODS      MAXPODS      REPLICAS      AGE
webtest   Deployment/webtest 1% / 10%    1            10           1            23s
praparns-MacBook-Pro%
```

The screenshot shows the Kubernetes UI interface. On the left, there's a sidebar with navigation links: Cluster, Namespaces, Nodes, Persistent Volumes, Roles, and Storage Classes. The main content area has a blue header bar with the text "Workloads > Deployments > webtest". Below the header, there's a message: "This Deployment does not have any old replica sets". Underneath this, there's a section titled "Horizontal Pod AutoScalers". A table lists the HPA configuration for the "webtest" deployment:

Name	Target CPU Utilization	Current CPU Utilization	Min Replicas	Max Replicas	Age	⋮
webtest	10%	1%	1	10	7 minutes	⋮

On the right side of the screen, there are buttons for "+ CREATE", "SCALE", "EDIT", and "DELETE".

Horizontal Pods Autoscaler

- Example: Deployment for python

```
praparns-MacBook-Pro% kubectl run -i --tty load-generator --image=busybox /bin/sh
If you don't see a command prompt, try pressing enter.
/ # wget -q -O- http://webtest.default.svc.cluster.local:5000
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 16:54:02 2018
/ # while true; sleep 0.01; do wget -q -O- http://webtest.default.svc.cluster.local:5000; done
```

```
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 16:55:30 2018
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 16:55:30 2018
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 16:55:30 2018
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 16:55:30 2018
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 16:55:30 2018
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 16:55:30 2018
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 16:55:30 2018
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 16:55:30 2018
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 16:55:30 2018
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 16:55:30 2018
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 16:55:30 2018
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 16:55:30 2018
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 16:55:30 2018
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 16:55:30 2018
```

Horizontal Pods Autoscaler

- Example: Deployment for python

```
praparns-MBP:~ praparn$ kubectl get hpa
NAME      REFERENCE      TARGETS      MINPODS      MAXPODS      REPLICAS      AGE
webtest   Deployment/webtest  1% / 10%    1            10           1            10m

praparns-MBP:~ praparn$ kubectl get hpa
NAME      REFERENCE      TARGETS      MINPODS      MAXPODS      REPLICAS      AGE
webtest   Deployment/webtest  27% / 10%   1            10           1            11m

praparns-MBP:~ praparn$ kubectl top pods
NAME                  CPU(cores)      MEMORY(bytes)
load-generator-5c4d59d5dd-m4tsb  43m          1Mi
webtest-7d89786977-tmg76       33m          28Mi

praparns-MBP:~ praparn$ kubectl top nodes
NAME                  CPU(cores)      CPU%      MEMORY(bytes)      MEMORY%
docker-for-desktop   468m          11%      1433Mi          75%

praparns-MBP:~ praparn$
```

```
praparns-MBP:~ praparn$ kubectl top nodes
NAME                  CPU(cores)      CPU%      MEMORY(bytes)      MEMORY%
docker-for-desktop   749m          18%      1508Mi          79%

praparns-MBP:~ praparn$ kubectl get hpa
NAME      REFERENCE      TARGETS      MINPODS      MAXPODS      REPLICAS      AGE
webtest   Deployment/webtest  15% / 10%   1            10           3            14m

praparns-MBP:~ praparn$ kubectl get hpa
NAME      REFERENCE      TARGETS      MINPODS      MAXPODS      REPLICAS      AGE
webtest   Deployment/webtest  15% / 10%   1            10           5            15m

praparns-MBP:~ praparn$ kubectl get hpa
NAME      REFERENCE      TARGETS      MINPODS      MAXPODS      REPLICAS      AGE
webtest   Deployment/webtest  8% / 10%    1            10           5            17m

praparns-MBP:~ praparn$
```

Horizontal Pods Autoscaler

- Example: Deployment for python

```
praparns-MBP:~ praparn$ kubectl describe hpa/webtest
Name:          webtest
Namespace:     default
Labels:        <none>
Annotations:   <none>
CreationTimestamp: Mon, 26 Mar 2018 23:45:03 +0700
Reference:    Deployment/webtest
Metrics:      resource cpu on pods  (as a percentage of request): 8% (17m) / 10%
  Min replicas: 1
  Max replicas: 10
Conditions:
  Type        Status  Reason
  AbleToScale False   BackoffBoth
  ScalingActive True    ValidMetricFound
  ScalingLimited False  DesiredWithinRange
Events:
  Type      Reason
  Normal   SuccessfulRescale
  6m       horizontal-pod-autoscaler
  New size: 3; reason: cpu resource utilization (percentage of request) above target
  Normal   SuccessfulRescale
  2m       horizontal-pod-autoscaler
  New size: 5; reason: cpu resource utilization (percentage of request) above target
praparns-MBP:~ praparn$
```

Horizontal Pods Autoscaler

- Example: Deployment for python

The screenshot shows the Kubernetes web interface with the following details:

Cluster	Action	Type	Status	Replicas	Last Scale Time	Created
Namespaces	New size: 3; reason: cpu resource utilization (percentage of request) above target	horizontal-pod-autoscaler	-	1	2018-03-26T16:56 UTC	2018-03-26T16:56 UTC
Nodes	Scaled up replica set webtest-7d8978 6977 to 3	deployment-controller	-	1	2018-03-26T16:56 UTC	2018-03-26T16:56 UTC
Roles	New size: 5; reason: cpu resource utilization (percentage of request) above target	horizontal-pod-autoscaler	-	1	2018-03-26T17:00 UTC	2018-03-26T17:00 UTC
Storage Classes	Scaled up replica set webtest-7d8978 6977 to 5	deployment-controller	-	1	2018-03-26T17:00 UTC	2018-03-26T17:00 UTC
Namespace	New size: 4; reason: All metrics below target	horizontal-pod-autoscaler	-	1	2018-03-26T17:06 UTC	2018-03-26T17:06 UTC

Navigation and search bars are visible at the top, along with a sidebar for cluster management.

Horizontal Pods Autoscaler

- Example: Deployment for python

```
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 17:12:49 2018
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 17:12:49 2018
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 17:12:49 2018
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 17:12:49 2018
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 17:12:49 2018
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 17:12:49 2018
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 17:12:49 2018
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 17:12:49 2018
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 17:12:49 2018
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 17:12:49 2018
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 17:12:49 2018
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 17:12:49 2018
<H1> Welcome Page from Container Python Lab Web Version 1.00 </H1>Checkpoint Date/Time: Mon Mar 26 17:12:49 2018
^C
/ # exit
Session ended, resume using 'kubectl attach load-generator-5c4d59d5dd-m4tsb -c load-generator -i -t' command when the pod is running
praparns-MacBook-Pro% 
```

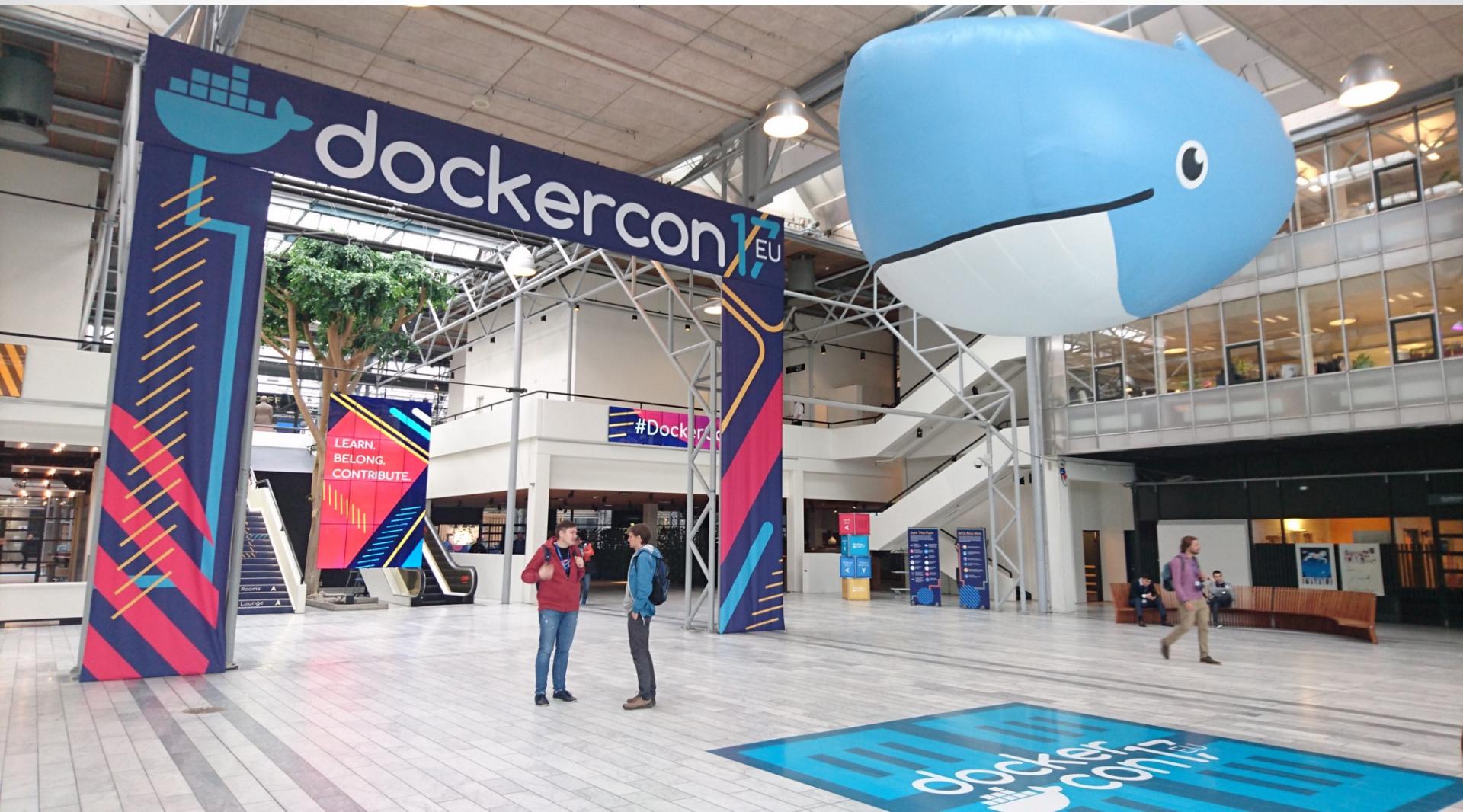
The screenshot shows the Kubernetes UI for managing workloads. On the left, a sidebar navigation includes 'Namespaces' (set to 'default'), 'Overview', 'Workloads' (selected), 'Cron Jobs', 'Daemon Sets', 'Deployments' (selected), 'Jobs', 'Pods', and 'Replica Sets'. The main content area has two tabs: 'Horizontal Pod AutoScalers' and 'Events'. The 'Horizontal Pod AutoScalers' tab displays a table with the following data:

Name	Target CPU Utilization	Current CPU Utilization	Min Replicas	Max Replicas	Age
webtest	10%	1%	1	10	31 minutes

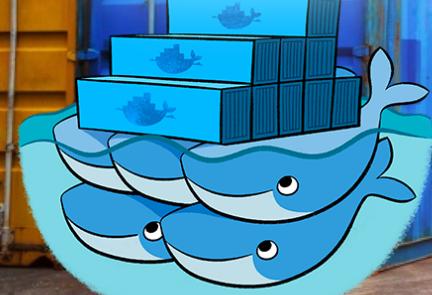
The 'Events' tab shows a single event:

Message	Source	Sub-object	Count	First seen	Last seen
Scaled up replica set webtest-7d89786977 to 1	deployment-controller	-	1	2018-03-26T16:33 UTC	2018-03-26T16:33 UTC

Demo Session



"Swarm & K8S" We are one big community



By Praparn Luengphoonlap
Email: eva10409@gmail.com

"Swarm & K8S" We are one big community

Q&A