
SciLifeLab

SCALING MADE SIMPLE: HOW KUBERNETES HANDLES GROWING APPS

ÁLVARO REVUELTA

SciLifeLab

ABOUT ME

Alvaro Revuelta.

- Systems Developer
- Laboratory for Life Sciences - SciLifeLab
- Follow the presentation

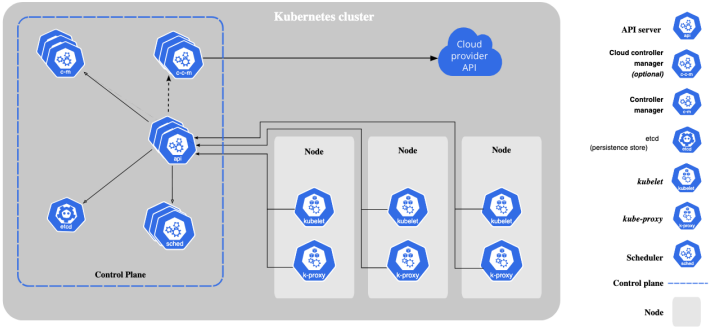
SciLifeLab

PRESENTATION STRUCTURE

1. K8s intro.
2. What is Scaling?
3. How Kubernetes Solves Scaling (Including live demo).
4. Final

SciLifeLab

1. K8S INTRO

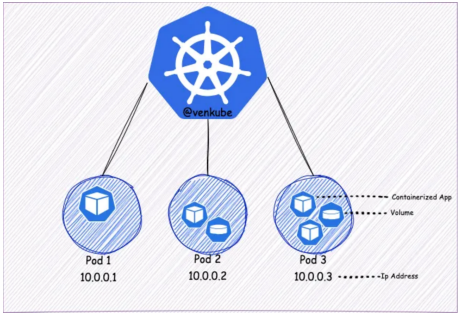


SOURCE: KUBERNETES.IO

SciLifeLab

1. K8S INTRO

- Open-source platform widely used in cloud infrastructures
- Managing of containers efficiently
- Key components
  - **Master Nodes**.
  - **Worker Nodes**.
  - **Pods**: Smallest deployable unit
  - **Deployments**: Kubernetes resource where we specify the number and characteristics of our pods.
  - **Controllers**: Ensure the desired state of the applications.



Source: VenKube

2. WHAT IS SCALING?

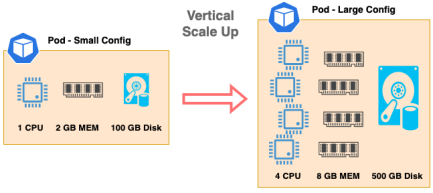
What happens when your app gets popular?

- **Scaling up** means adding more infrastructure, changing computing power or databases as needed.
- Ex: Larger hard drive, more CPU, etc.

- **Scaling out** is more associated with distributed architectures
- You would, for example, add more nodes to divide the workload among them.

SciLifeLab

3. HOW KUBERNETES SOLVES SCALING.



Source: portworx

Source: portworx

SciLifeLab

3. HOW KUBERNETES SOLVES SCALING.

The official documentation already documents it pretty well → link).

Scaling out

- We can manually resize the number of replicas (pods) for a deployment.
- Since v 1.23 (2022) there exists an automatic autoscaling (HPA) which monitors usage and automatically scales up or down.

Scaling up

- We can manually resize the resources of a pod. New pods will be created before the olds ones are deleted to ensure availability (Don't worry I will show it now).
- There is also an Autoscaler, but it is not enabled by default.
  - However, a good tutorial can be found in AWS.

SciLifeLab

3. HOW KUBERNETES SOLVES SCALING.

DEMO TIME!

SciLifeLab

3. HOW KUBERNETES SOLVES SCALING.

Other interesting resources/concepts.

1. Cluster Autoscaler: Also not enabled by default.
2. Autoscaling based on events:
    - For example, based on the number of queries or messages in a queue.
    - Can be combined with an autoscaling based on time, specifying to add resources during peak hours.
    - Check KEDA project.

SciLifeLab

4. FINAL

CONTACT:
LINKEDIN - ALVARO REVUELTA MARTINEZ
ALVARO.REVUELTA@SCILIFELAB.UU.SE

SciLifeLab

4. FINAL

Q & A

SciLifeLab