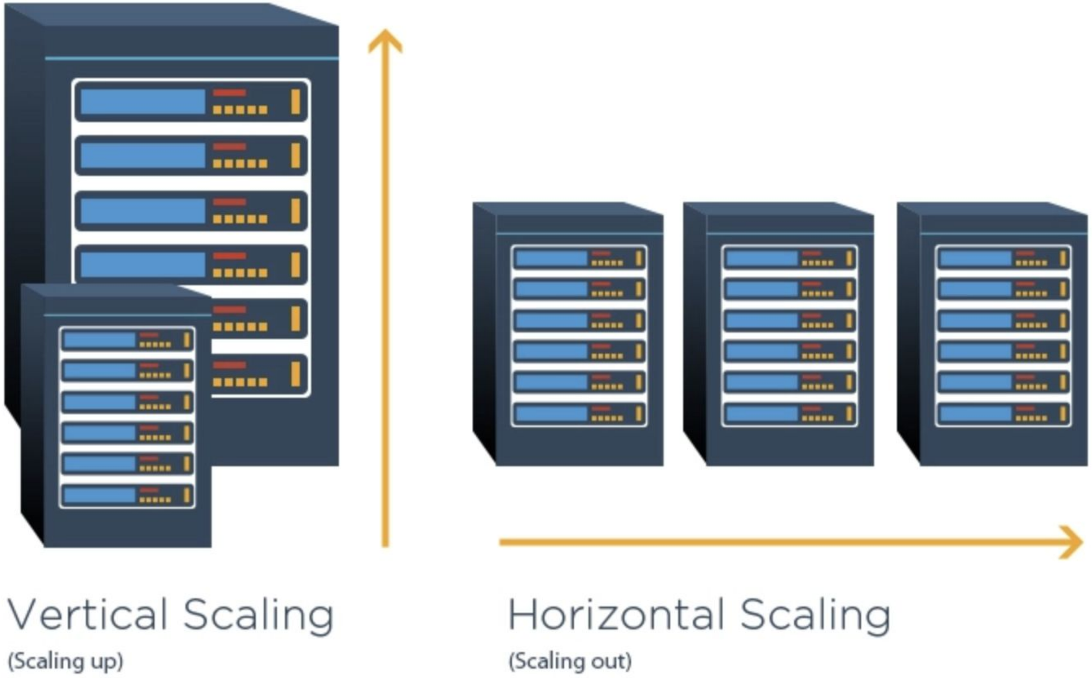
Scalability and Load balancing

OpsCI

What is it ?

- Respond to everyone

- Automatisation



The coffee shop problem

Clients are waiting in line to buy coffee

You have **baristas** and **coffee** machine

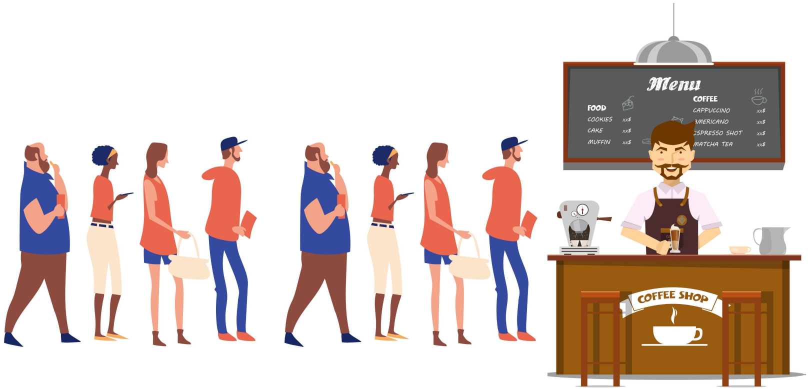
You must execute **actions** to **provide** them coffee before they leave **impatiently**

Basic scenario

- one barista
- one coffee machine
- one client line

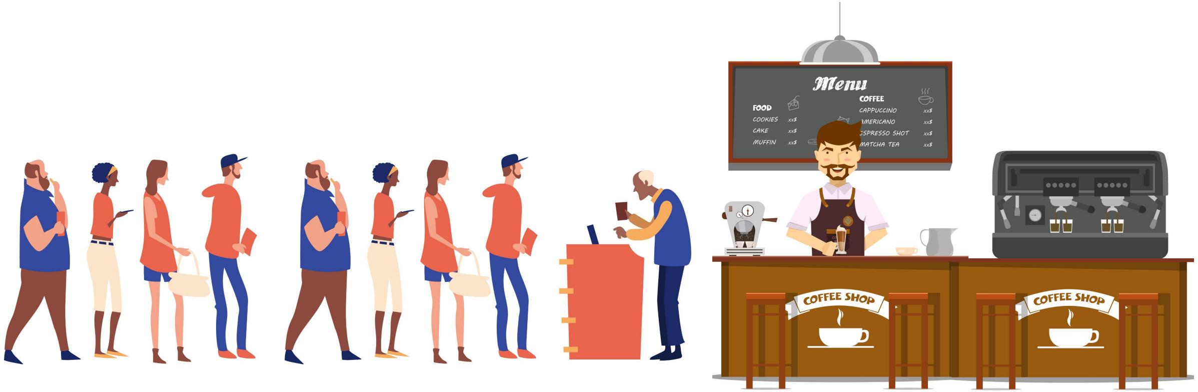
Borderline case and bottleneck

- too much people are waiting aka people arrive faster than you serve coffee
- the coffee machine is broken
- people are in the hurry
- on people ask for 24 coffees
- 50 people arrive at the same time
- etc

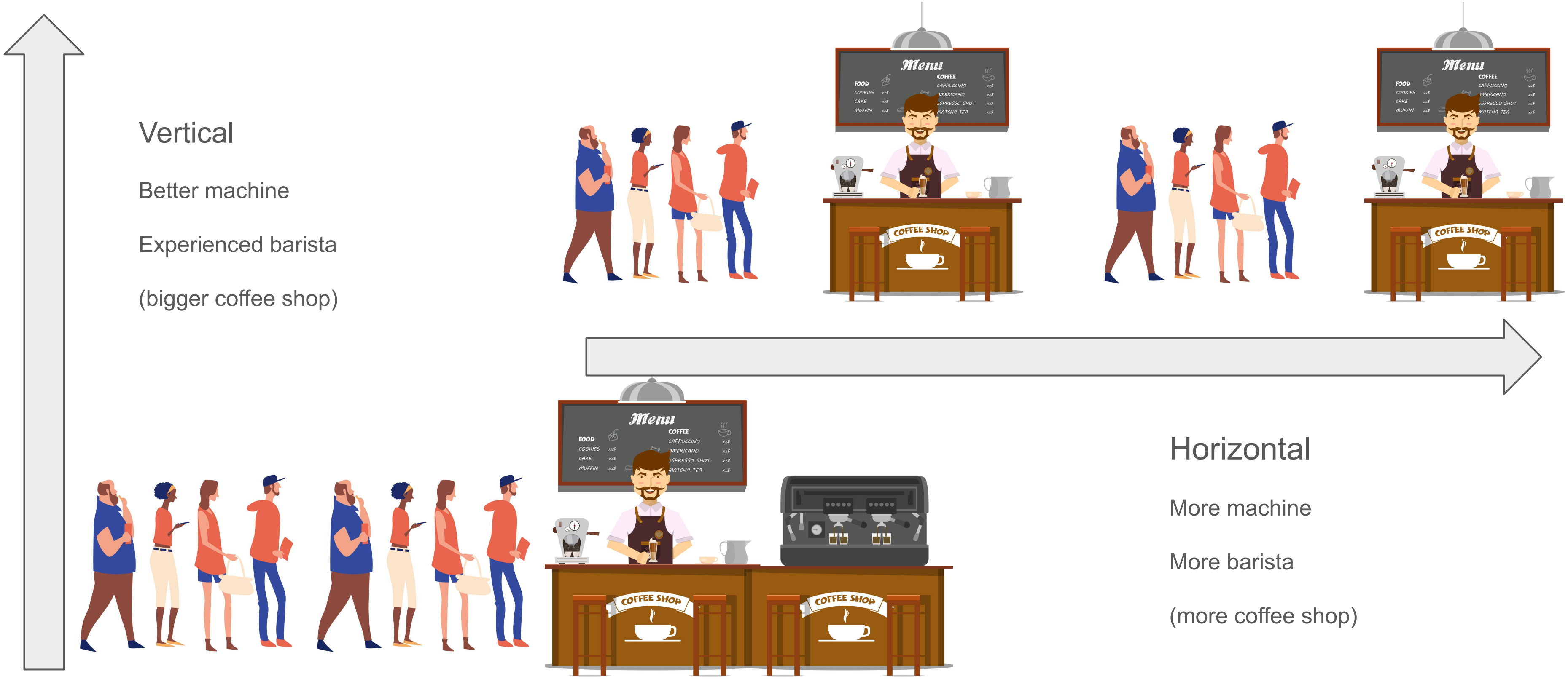


Upscaling your coffee shop

- use a bigger coffee machine
- put two people to serve coffee (split task or not)



Vertical vs horizontal scaling



Vertical

Better machine

Experienced barista

(bigger coffee shop)

Horizontal

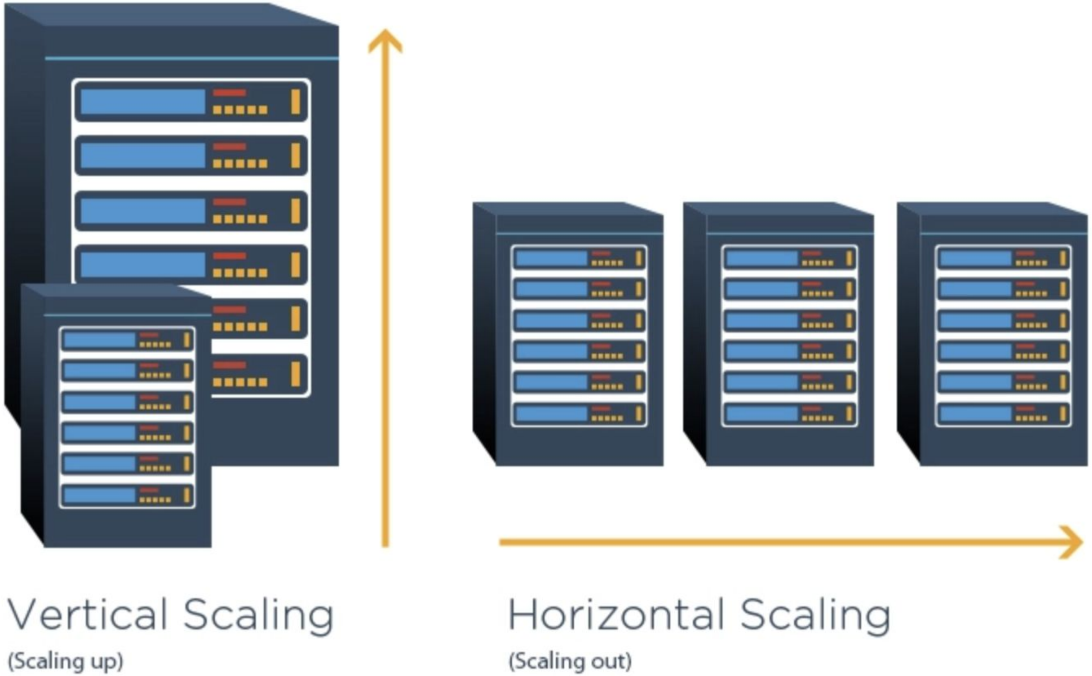
More machine

More barista

(more coffee shop)

Vertical vs horizontal

- resources limit
- entry costs
- organization

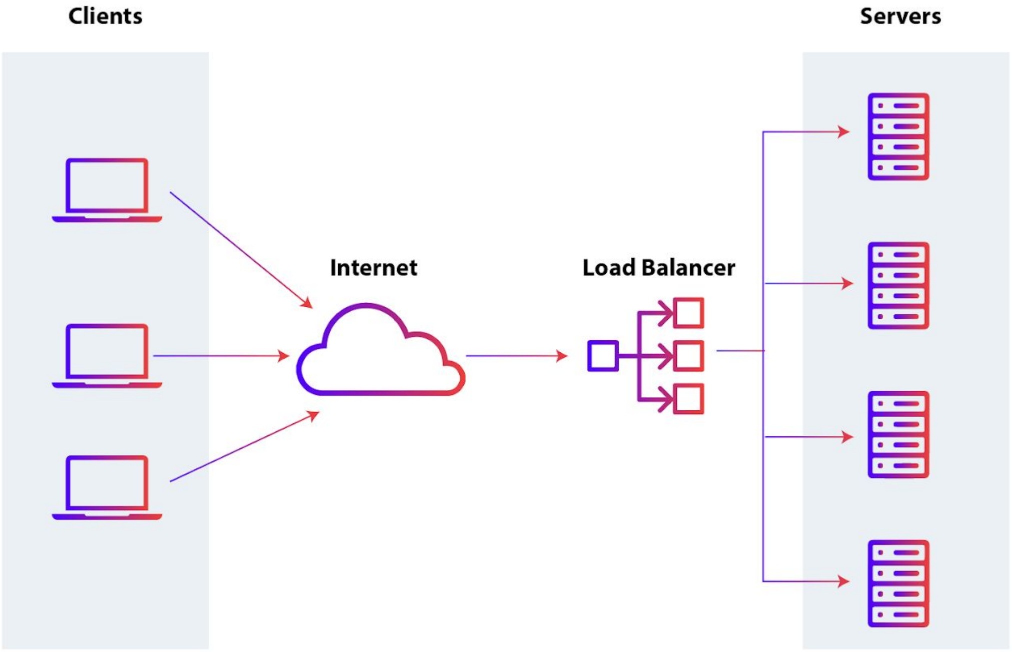


Load balancing

- split load

- single entry point

- load :
  - number of requests
  - CPU usage
  - memory usage
  - version management



Load balancing

Everybody want the coffee :

- network load balancing

Menu
COFFEE SHOP
Menu
COFFEE SHOP
Menu
COFFEE SHOP

Load balancing

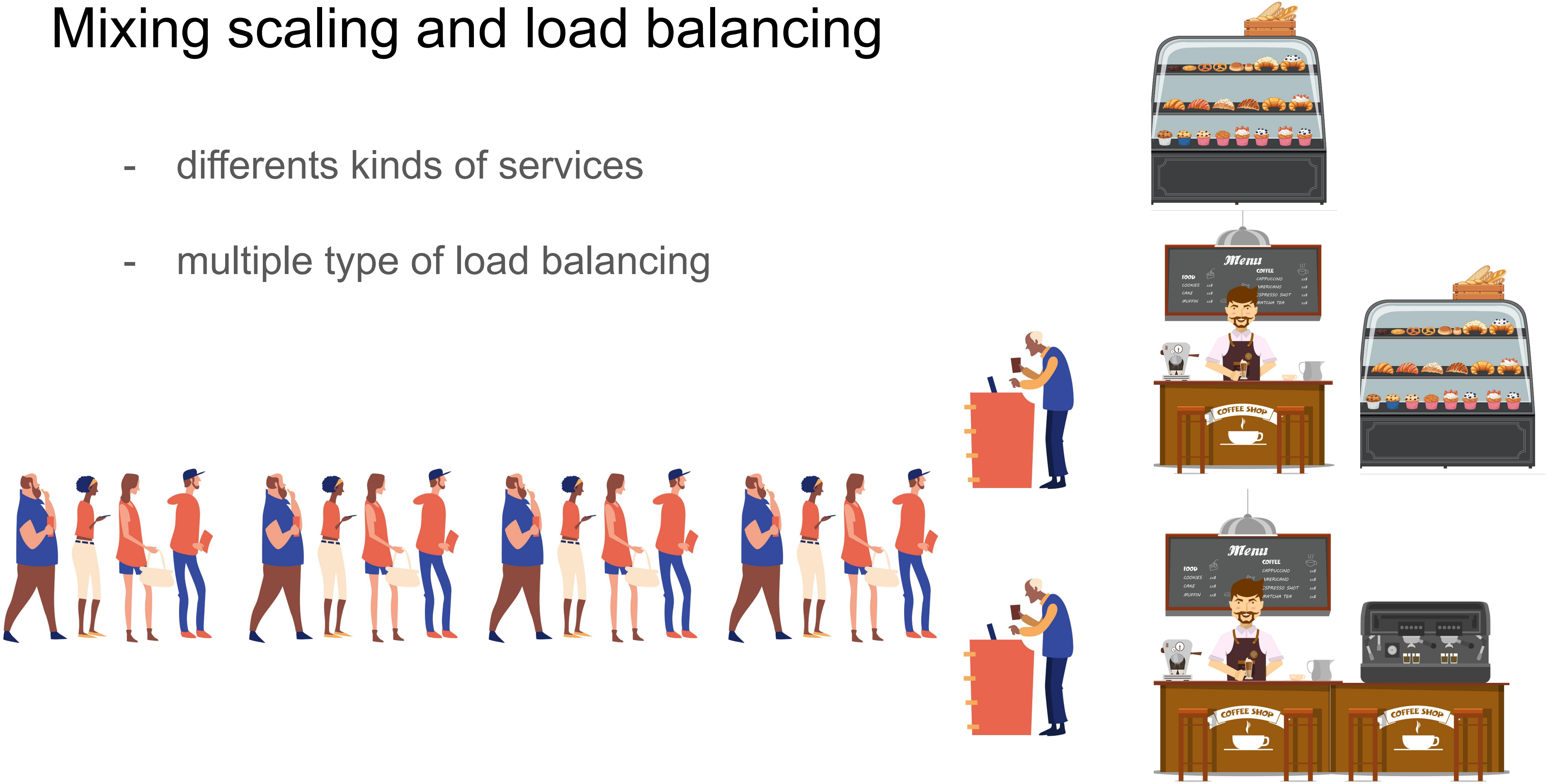
Clients want differents things :

- application load balancing

Mixing scaling and load balancing

- differents kinds of services

- multiple type of load balancing



Now using real servers

- coffee machine => database (postgreSQL, redis…)

- barista/cashier => server (nodeJS, React)

- menu => Frontend client (React in browser, Android, IOs)

- client => client

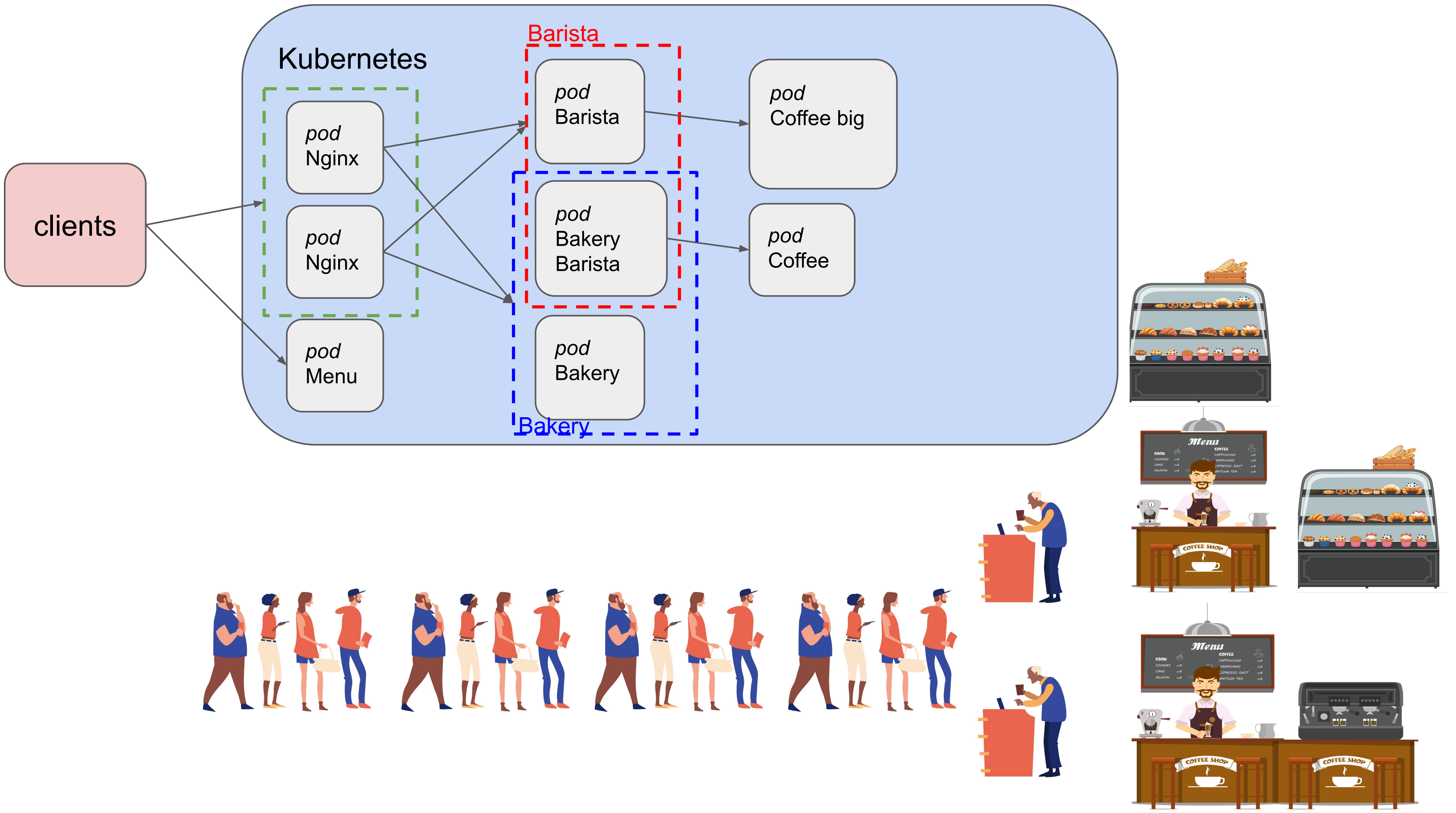
Managing everything

Scaling vertically : docker, kubernetes => giving more resources

Scaling horizontally : docker, kubernetes => create more containers/pods

Load balancing network : docker, kubernetes => create services and target groups

Load balancing application : reverse proxy => create routes



Kubernetes

clients

pod Nginx

pod Nginx

pod Menu

Barista

pod Barista

pod Bakery Barista

pod Bakery

Bakery

pod Coffee big

pod Coffee