PORTABLE MACHINE LEARNING DEPLOYMENTS IN 

Salman Iqbal

SALMAN IQBAL

@soulmaniqbal


learnk8s

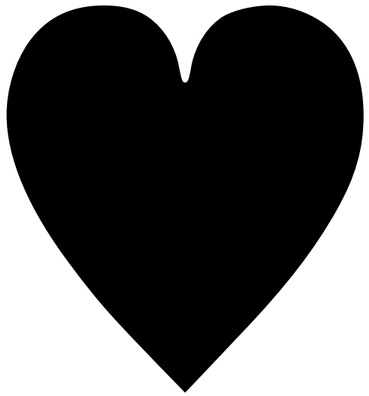

Office for
National Statistics

Cloud Native
Wales

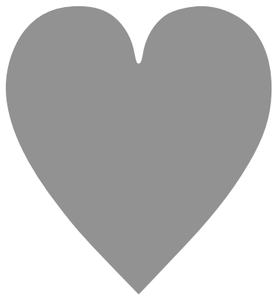
@CloudNativeWal

SOFTWARE DEVELOPER

SOFTWARE DEVELOPER

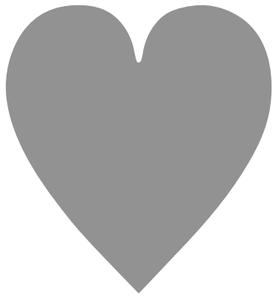
 BUILDING ON 

SOFTWARE DEVELOPER

 BUILDING THINGS ON 

DATA SCIENTIST

SOFTWARE DEVELOPER



BUILDING THINGS ON

DATA SCIENTIST

UK Mean Temperature for September 2019?

~ Met Office

13.1°C

UK Maximum Temperature for September 2019?

~ Met Office

27.7 °C

Weybourne (Norfolk)

AGENDA

Kubernetes

Containers

Kubeflow

Demo

THANKS DANIELE!

@danielepolencic

THANKS ECMWF!


AlphaGo

13 LAYER POLICY NETWORK

13 LAYER POLICY NETWORK

30 MILLION POSITIONS

13 LAYER POLICY NETWORK

30 MILLION POSITIONS

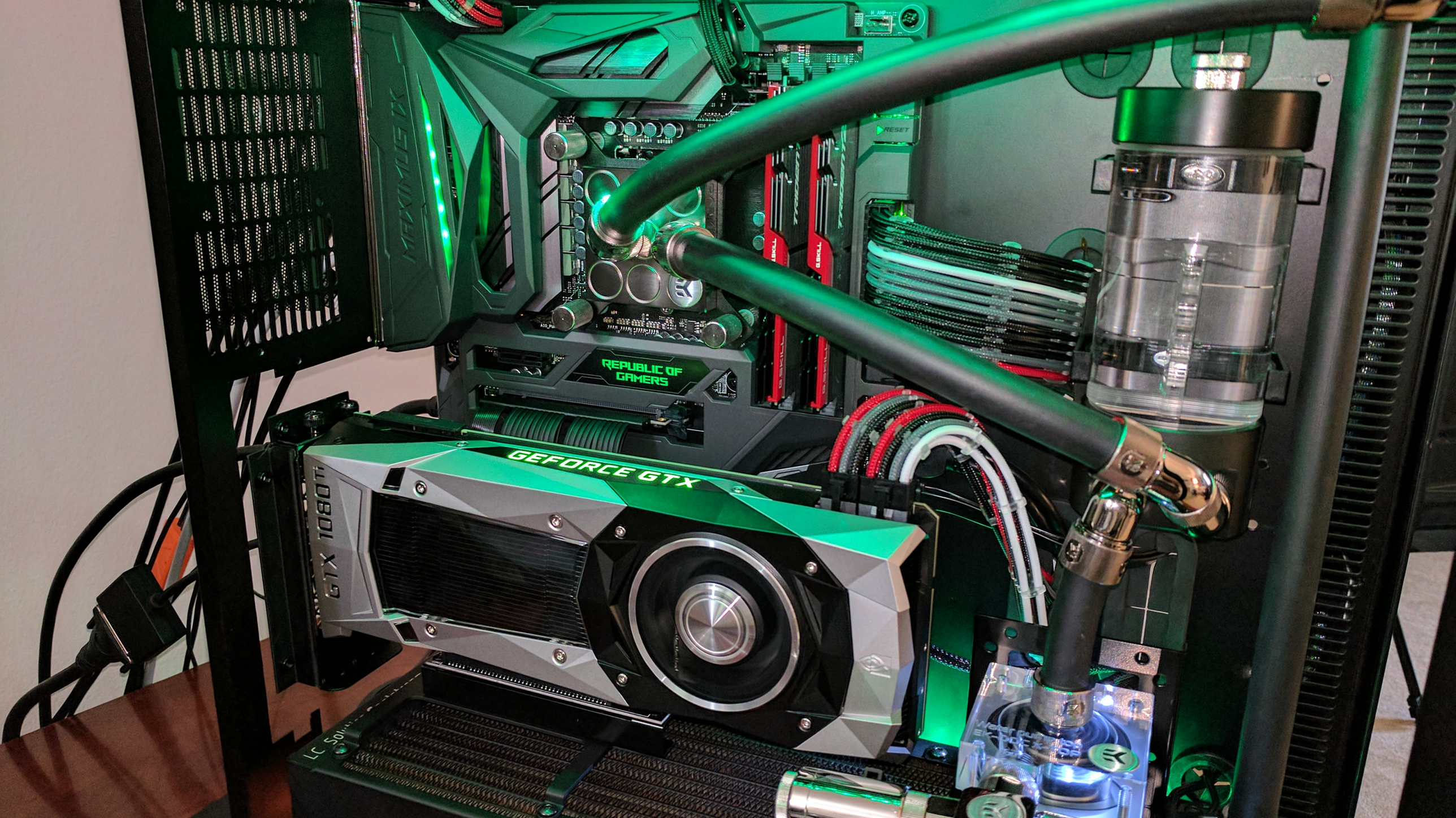
1,920 CPUS & 280 GPUS

CHALLENGES

1. COMPILE ONCE RUN ANYWHERE

~~1. COMPILE ONCE RUN ANYWHERE~~

1. WORKS ON MY MACHINE


REPUBLIC OF GAMERS
GEFORCE GTX
GTX 1080Ti

- [ ] BIG DATA
- [ ] SCALE WITH TEAM AND MODEL
- [ ] COST FOR MONEY
- [ ] QUICK FIX

BIG DATA

☐ SCALE WITH TEAM AND MODEL

☐ COST FOR MONEY

☐ QUICK FIX

BIG DATA

SCALE WITH TEAM AND MODEL

☐ COST FOR MONEY

☐ QUICK FIX

❌ BIG DATA

❌ SCALE WITH TEAM AND MODEL

❌ COST FOR MONEY

⬜ QUICK FIX

BIG DATA

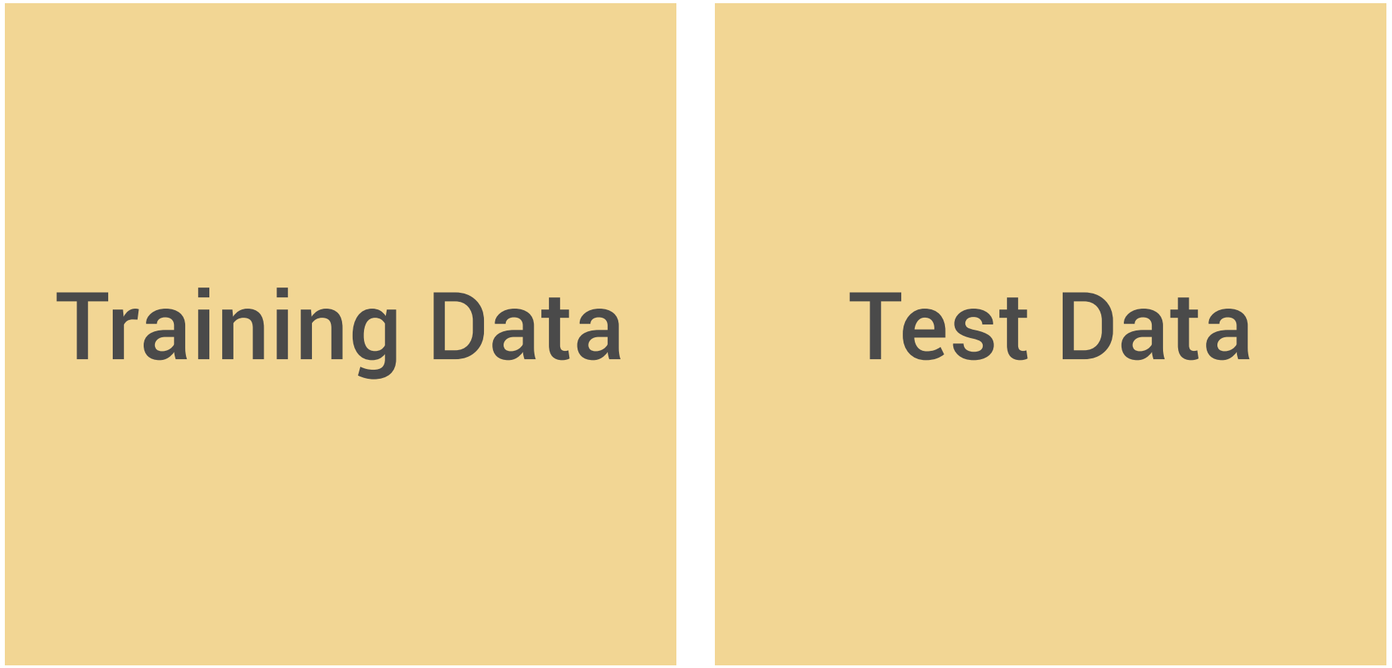
SCALE WITH TEAM AND MODEL

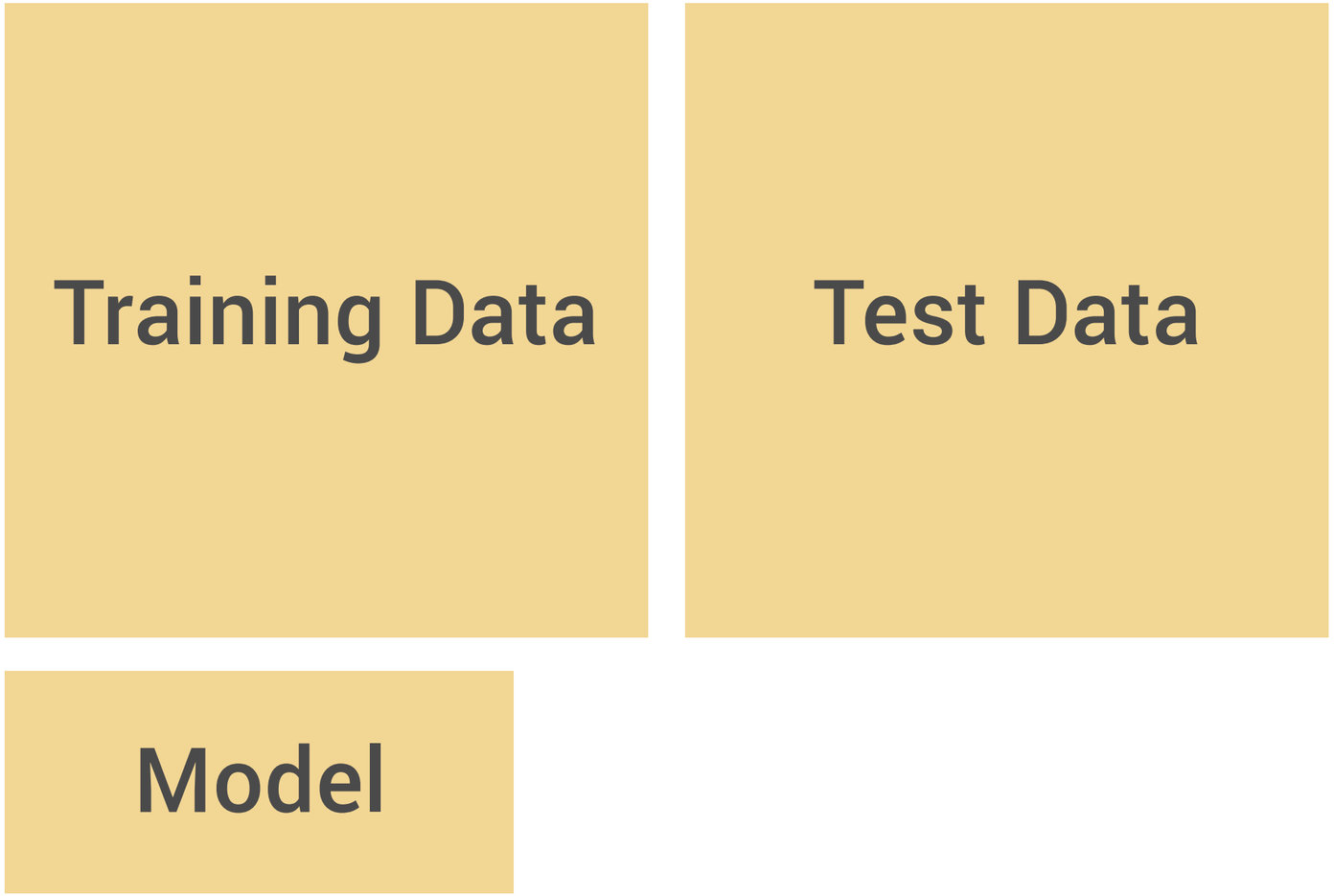
COST FOR MONEY

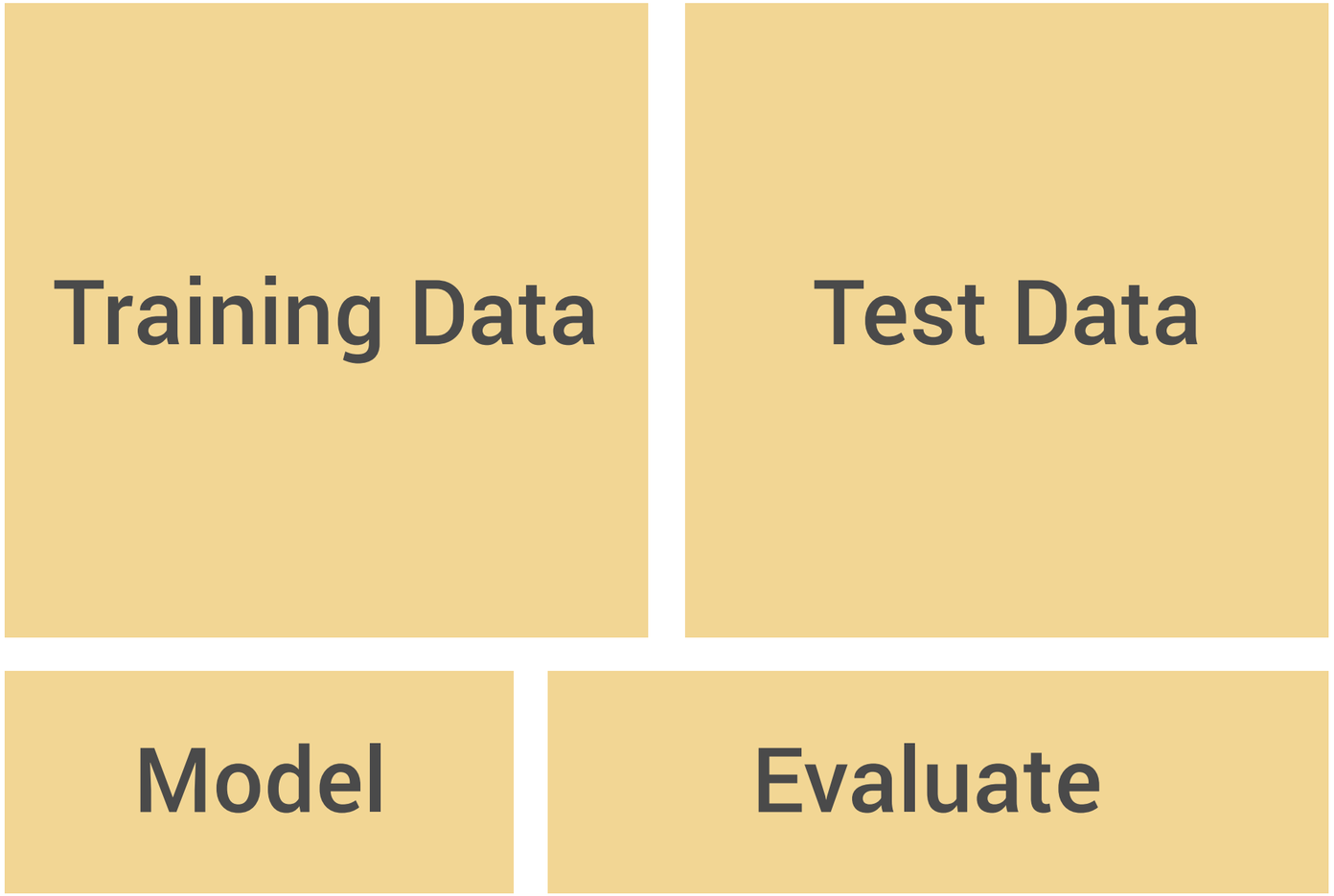
QUICK FIX

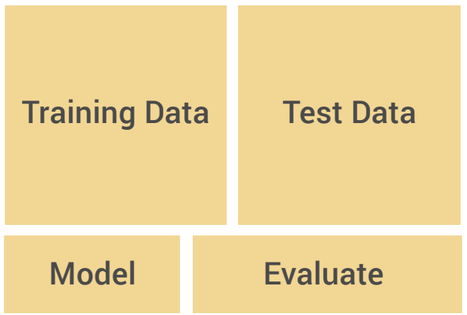
2. COMPLEXITY

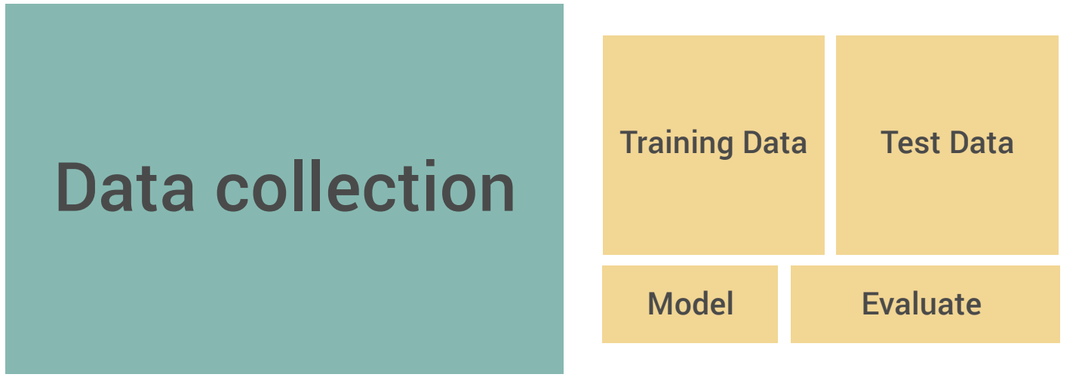
Training Data

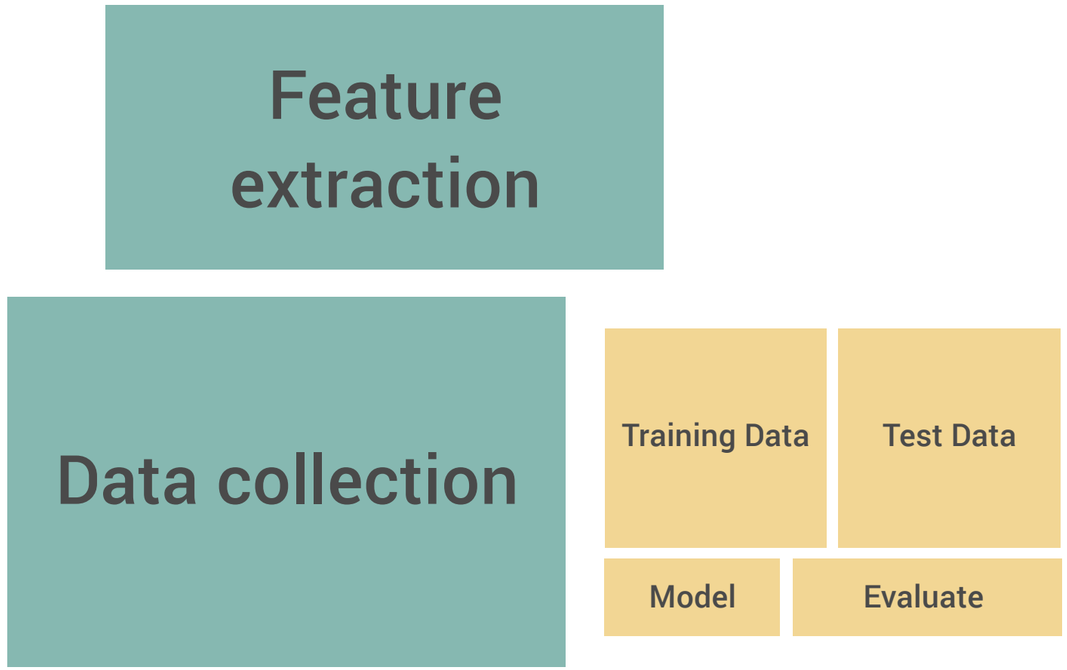

Training Data
Test Data

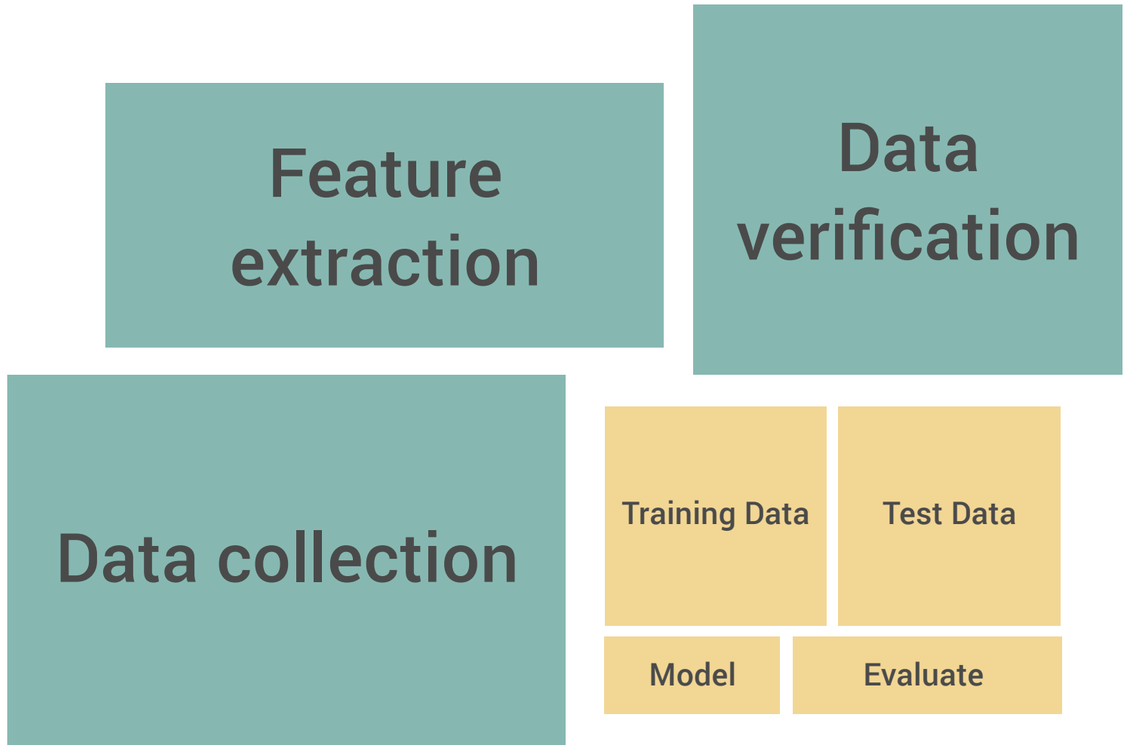

Training Data
Test Data
Model

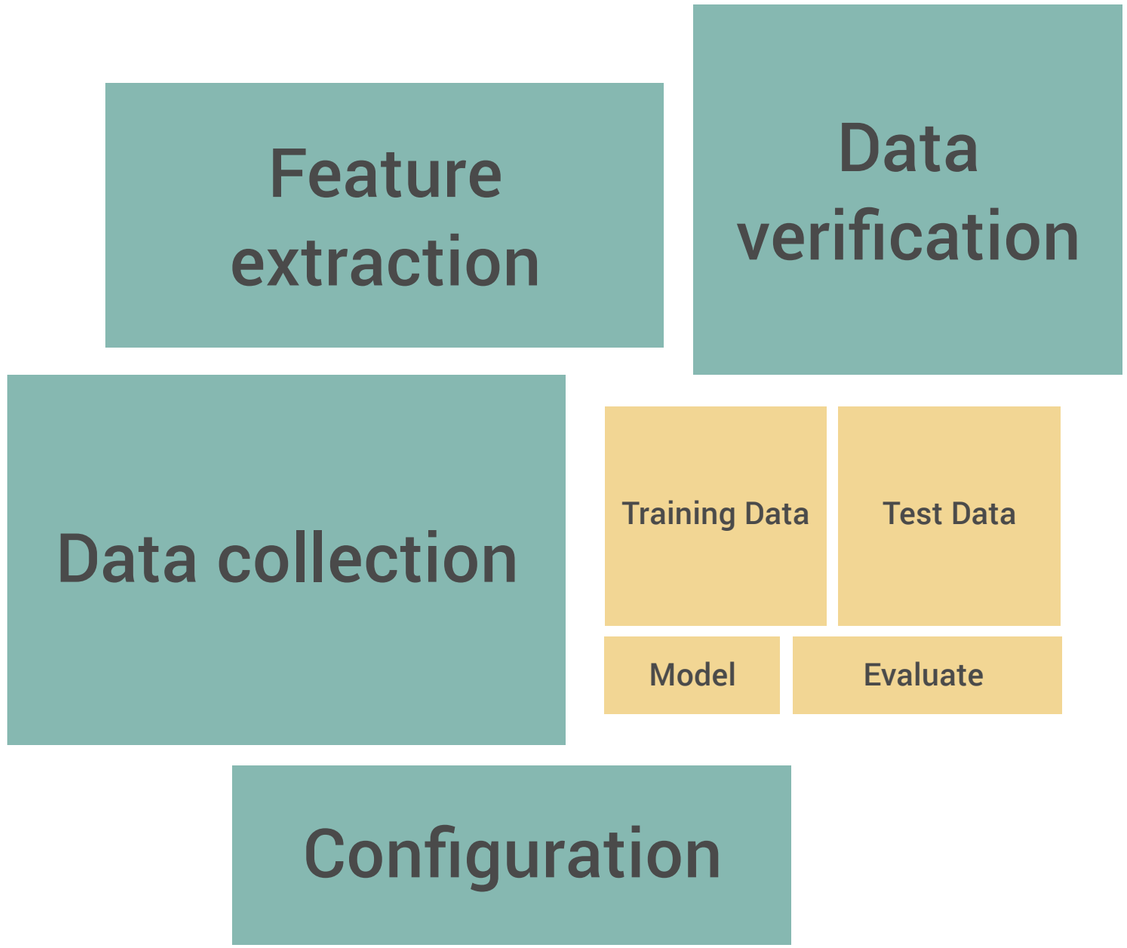

Training Data
Test Data
Model
Evaluate

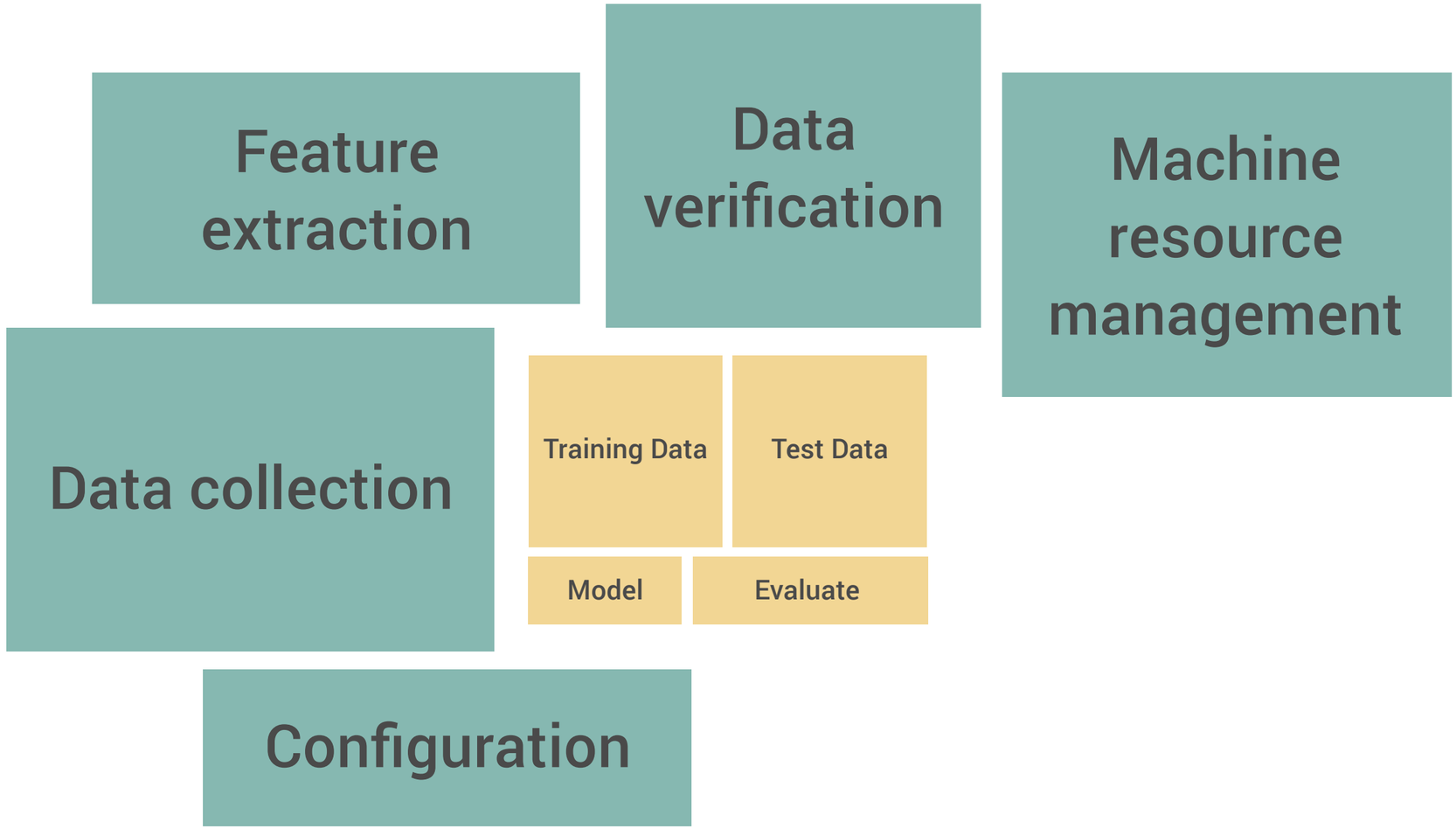

Training Data
Test Data
Model
Evaluate

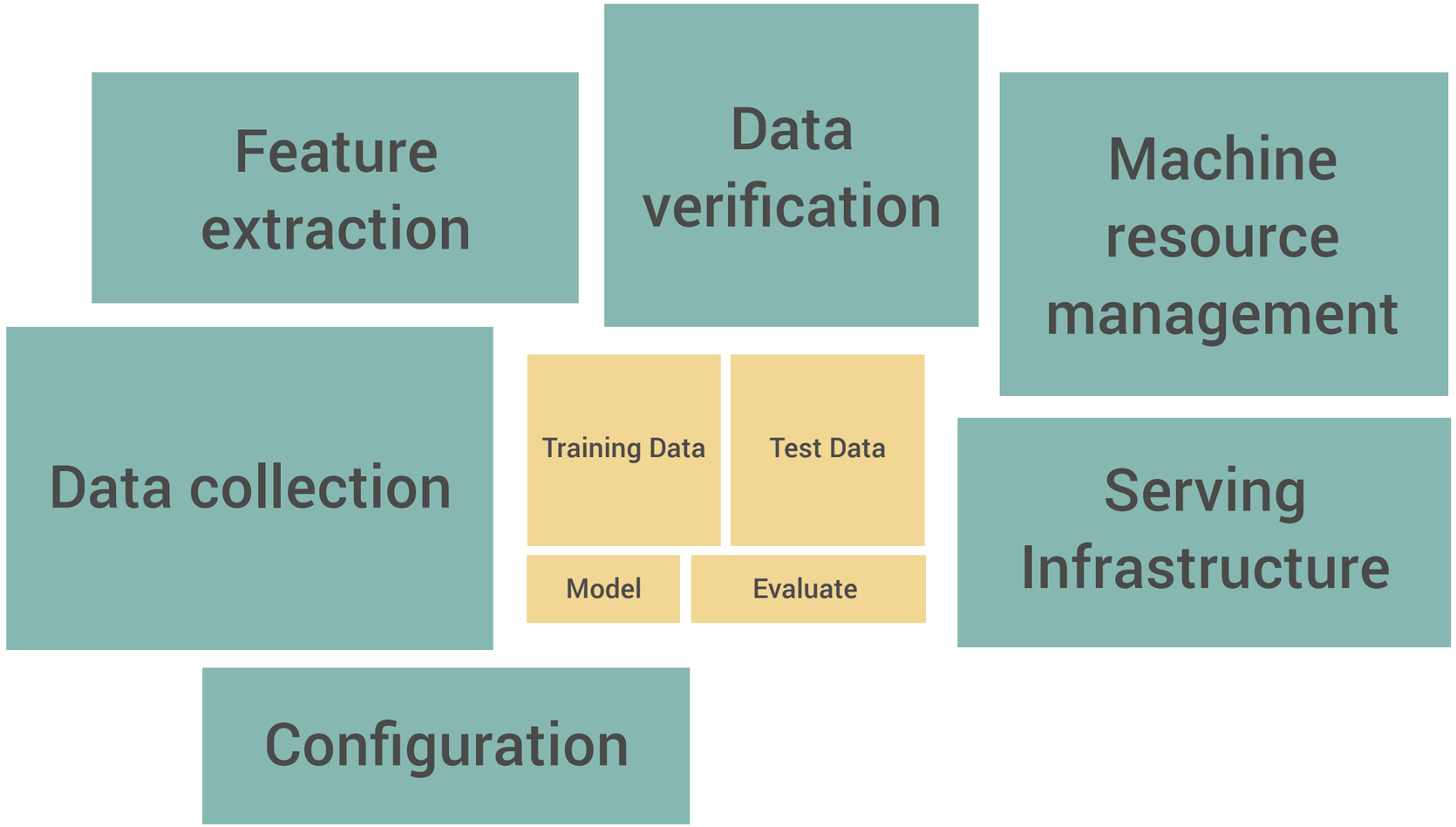

Data collection
Training Data
Test Data
Model
Evaluate

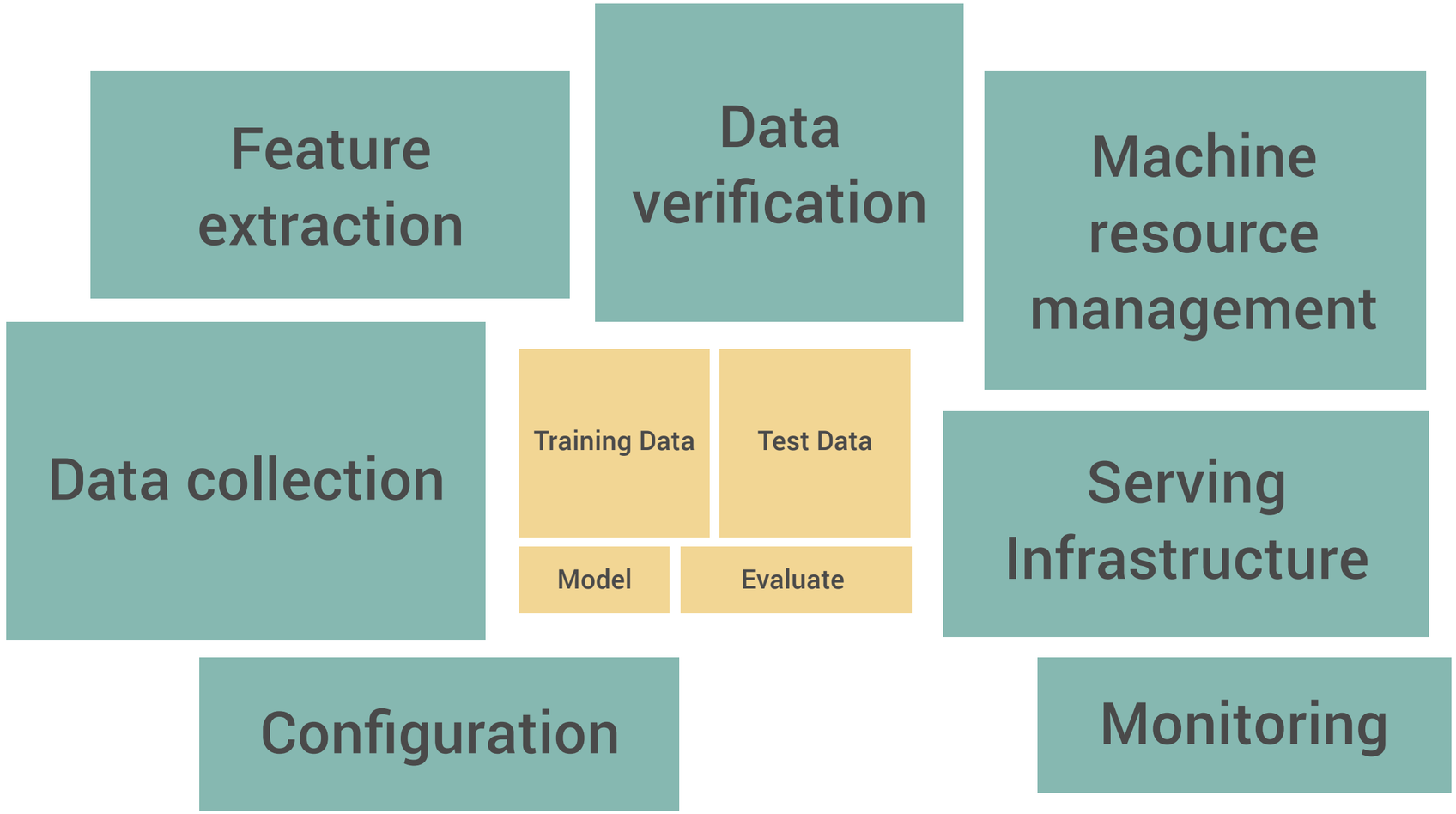

Feature extraction
Data collection
Training Data
Test Data
Model
Evaluate

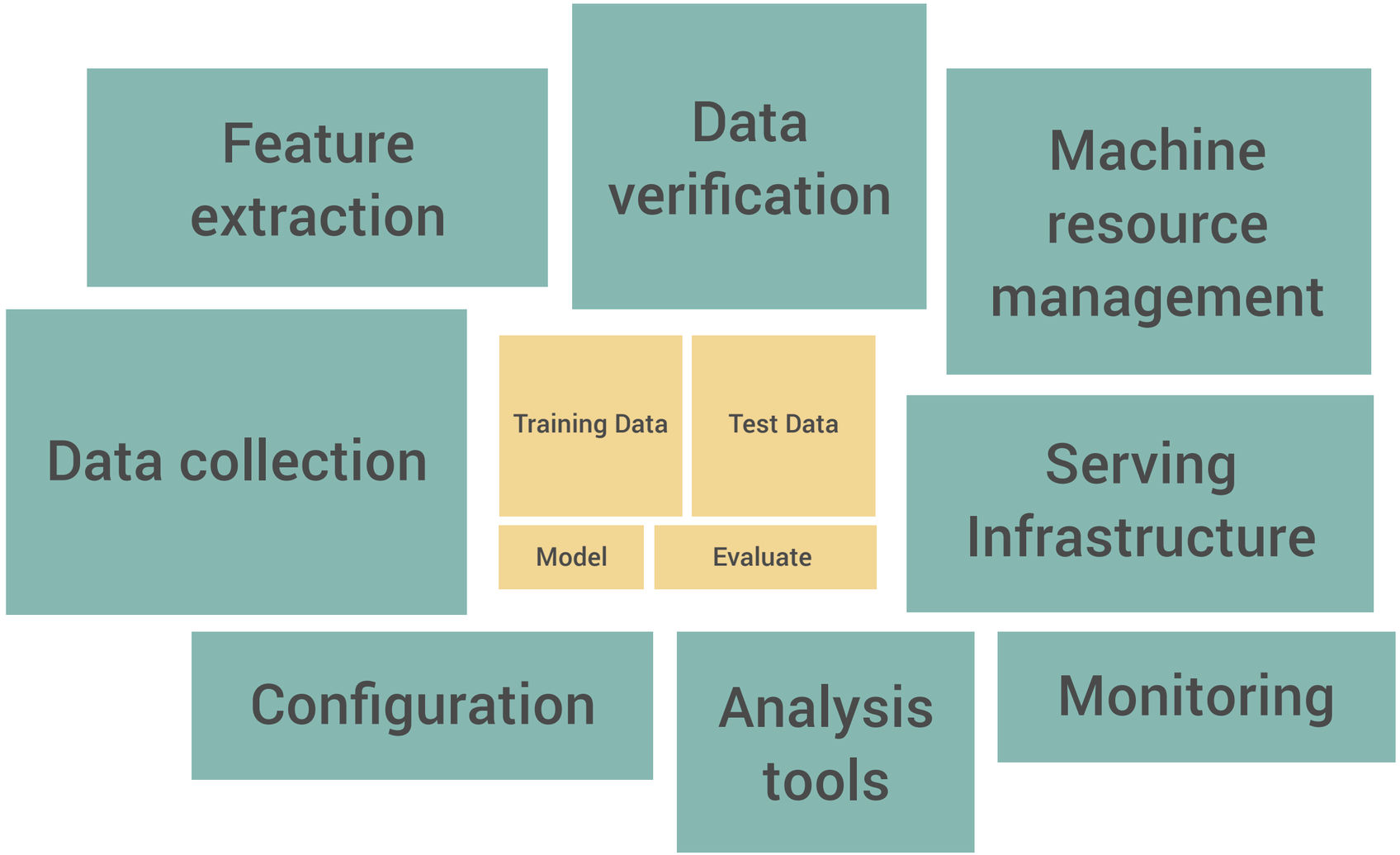

Feature extraction
Data verification
Data collection
Training Data
Test Data
Model
Evaluate


Feature extraction
Data verification
Data collection
Training Data
Test Data
Model
Evaluate
Configuration


Feature extraction
Data verification
Machine resource management
Data collection
Training Data
Test Data
Model
Evaluate
Configuration


Feature extraction
Data verification
Machine resource management
Data collection
Training Data
Test Data
Serving Infrastructure
Model
Evaluate
Configuration


Feature extraction
Data verification
Machine resource management
Data collection
Training Data
Test Data
Serving Infrastructure
Model
Evaluate
Configuration
Monitoring


Feature extraction
Data verification
Machine resource management
Data collection
Training Data
Test Data
Serving Infrastructure
Model
Evaluate
Configuration
Analysis tools
Monitoring

- [ ] LEARN DEVELOPMENT
- [ ] LEARN OPS & INFRA
- [ ] LEARN DATA SCIENCE

LEARN DEVELOPMENT

LEARN OPS & INFRA

LEARN DATA SCIENCE


LEARN DEVELOPMENT


LEARN OPS & INFRA

LEARN DATA SCIENCE

LEARN DEVELOPMENT

LEARN OPS & INFRA

LEARN DATA SCIENCE

DATA SCIENTIST – EXPECTATION

3. TRAINING HUNDREDS OF MODELS

Recommendation

Antispam

Churn rate

Recommendation

Antispam

Churn rate

Recommendation

v1

Recommendation

Antispam

Churn rate

Recommendation
v1

Recommendation
v2

Recommendation

Antispam

Churn rate

Recommendation
v1

Recommendation
v2

Recommendation
v3

Recommendation

Antispam

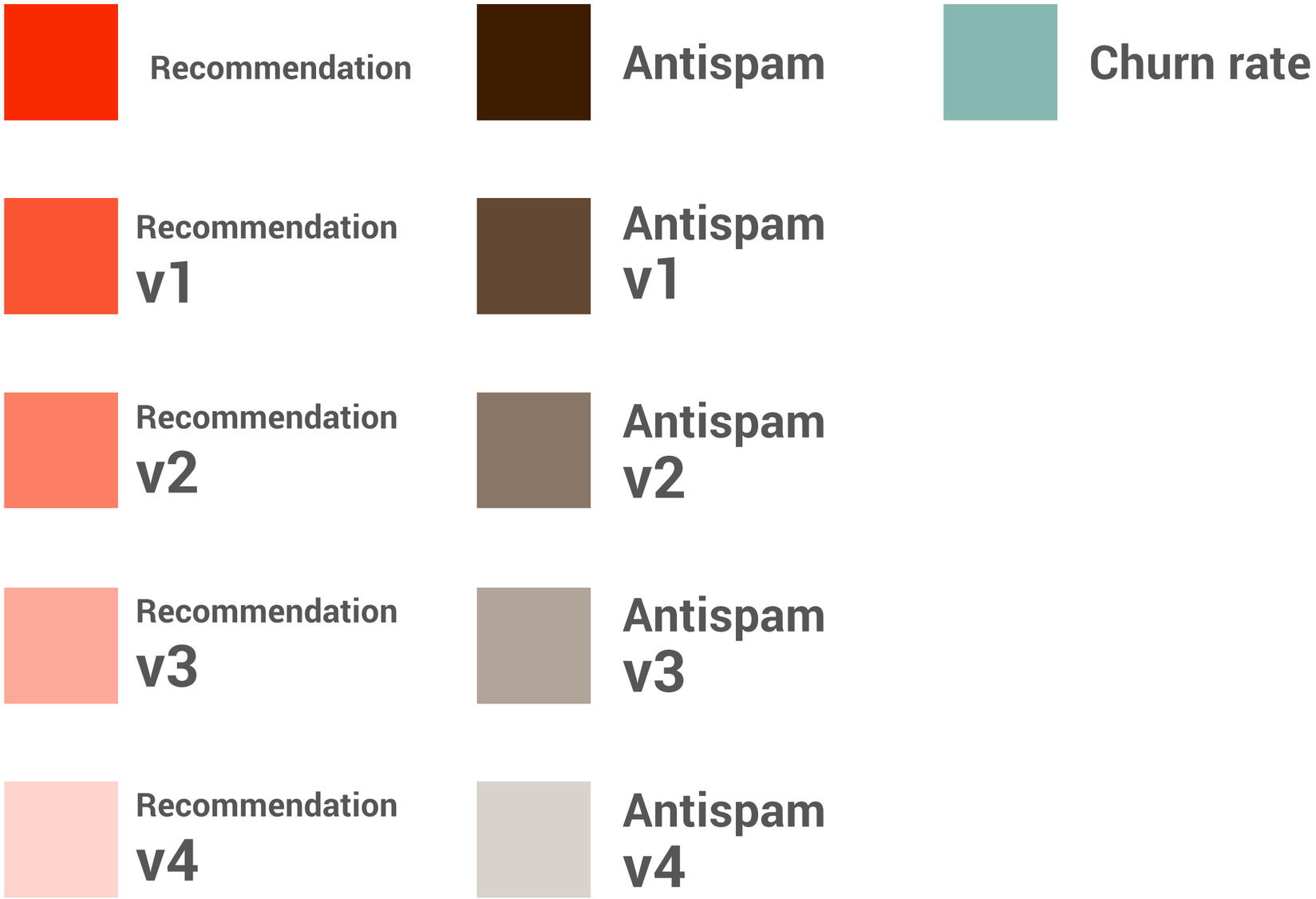
Churn rate

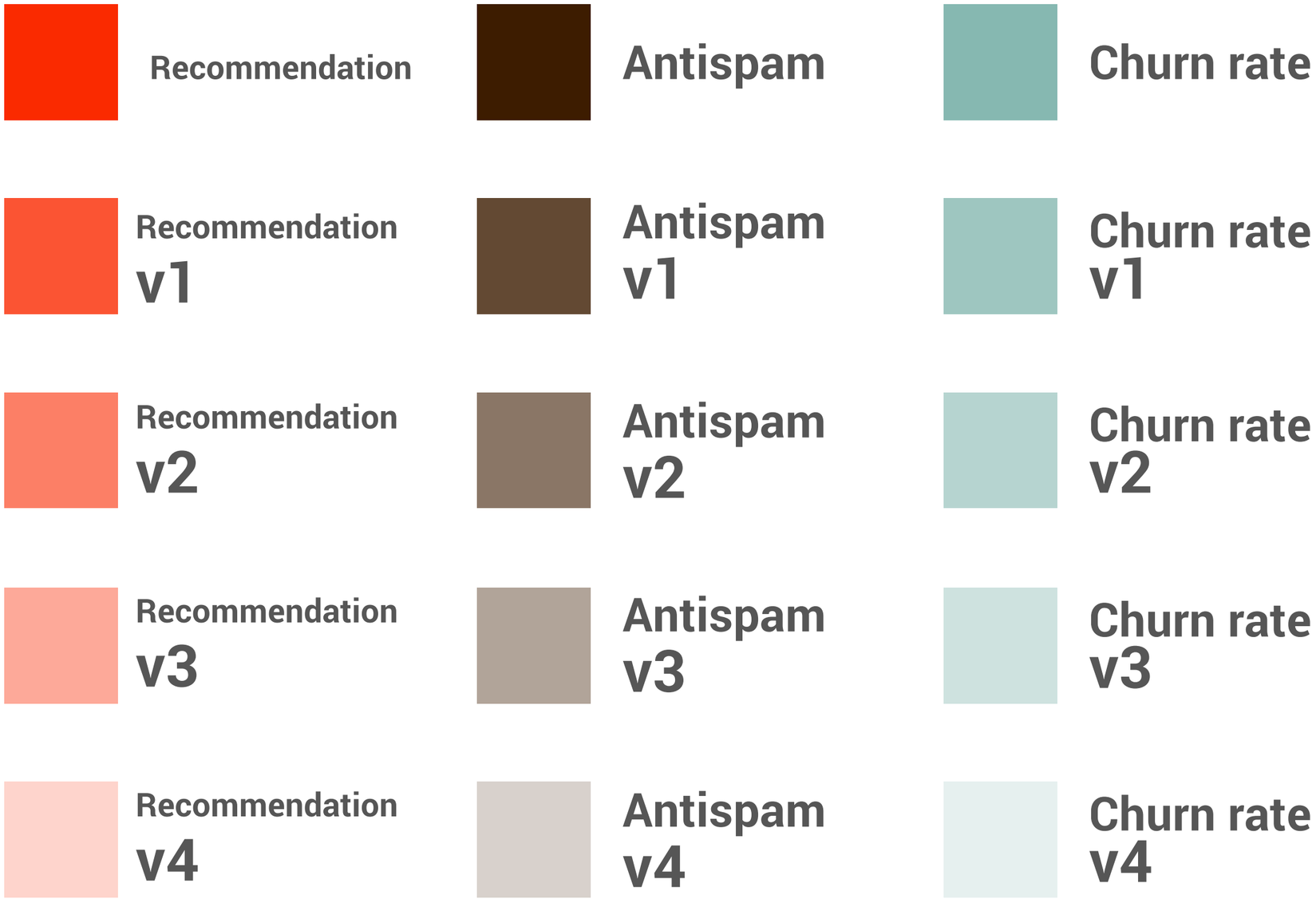
Recommendation
v1

Recommendation
v2

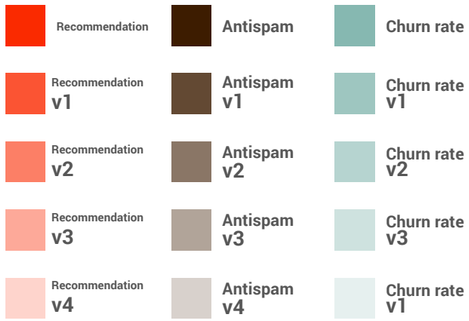
Recommendation
v3

Recommendation
v4

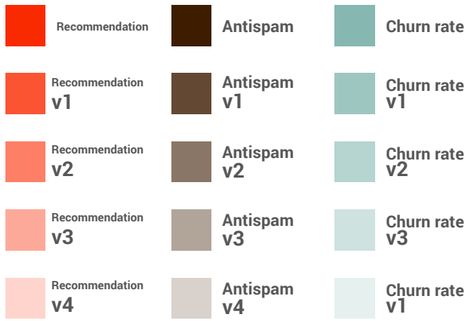

Recommendation
Antispam
Churn rate
Recommendation
v1
Antispam
v1
Recommendation
v2
Antispam
v2
Recommendation
v3
Antispam
v3
Recommendation
v4
Antispam
v4


Recommendation
Antispam
Churn rate
Recommendation
v1
Antispam
v1
Churn rate
v1
Recommendation
v2
Antispam
v2
Churn rate
v2
Recommendation
v3
Antispam
v3
Churn rate
v3
Recommendation
v4
Antispam
v4
Churn rate
v4

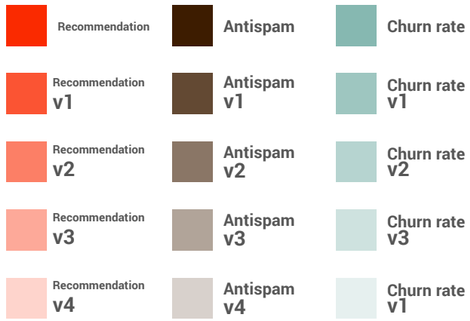
Monday


Recommendation
Antispam
Churn rate
Recommendation v1
Antispam v1
Churn rate v1
Recommendation v2
Antispam v2
Churn rate v2
Recommendation v3
Antispam v3
Churn rate v3
Recommendation v4
Antispam v4
Churn rate v1

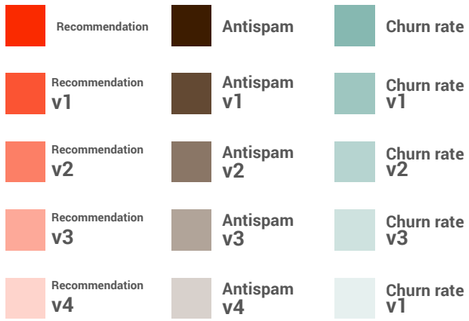
Monday



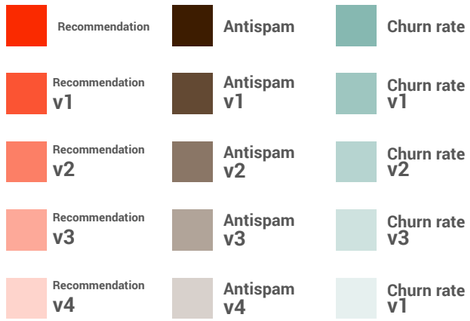
Tuesday



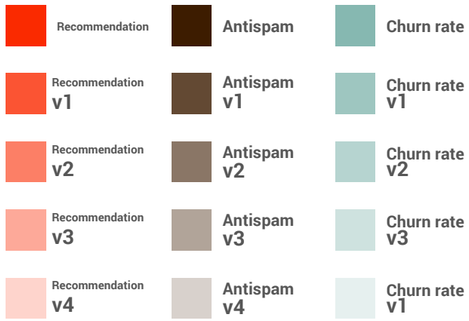
Monday



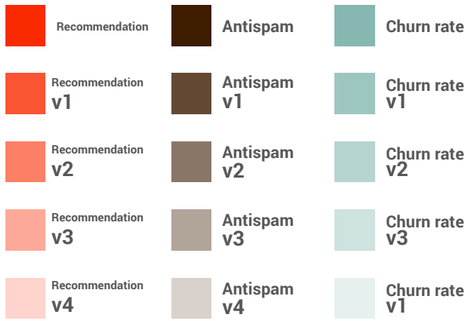
Tuesday



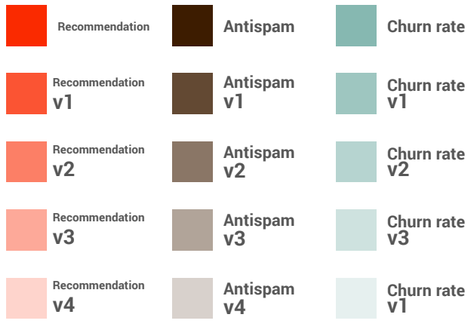
Wednesday



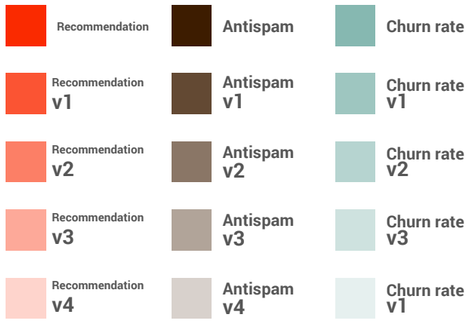
Monday



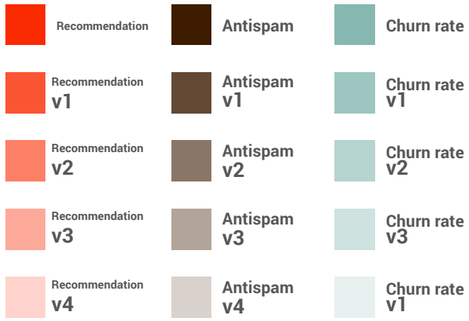
Tuesday



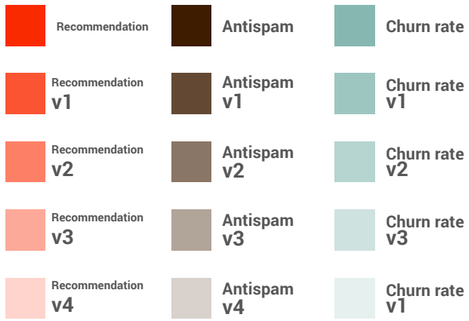
Wednesday



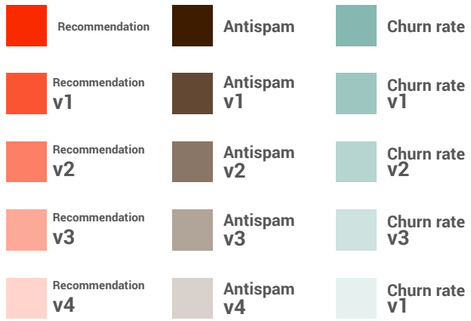
Thursday



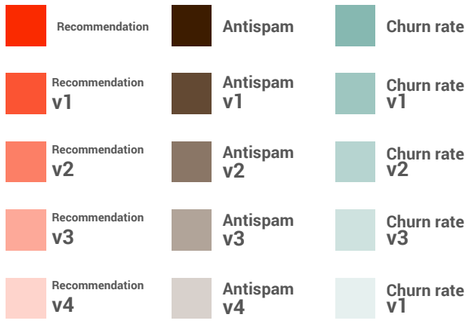
Monday



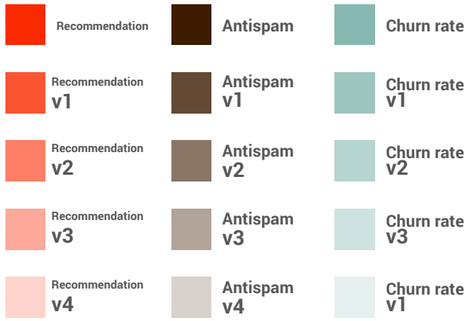
Tuesday



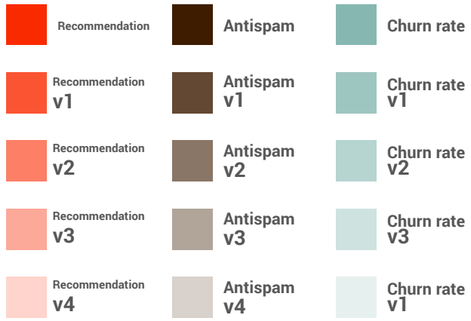
Wednesday



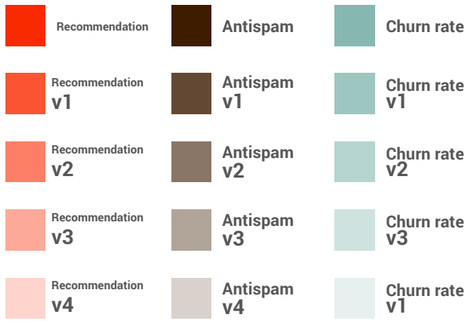
Thursday



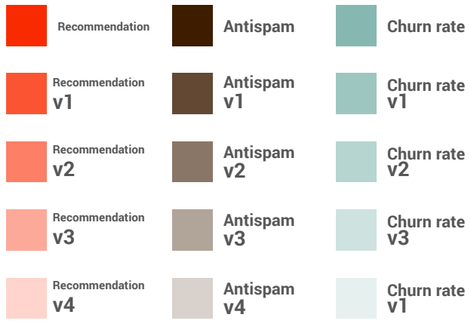
Friday



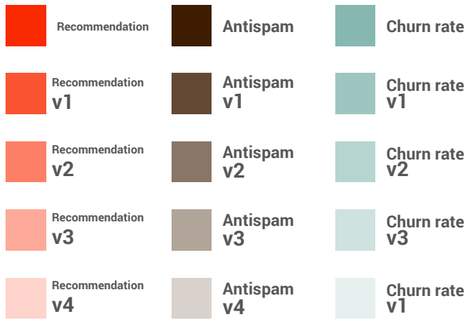
Monday



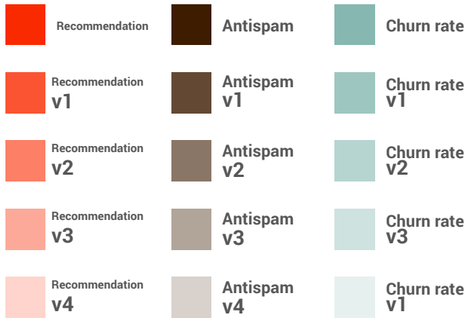
Tuesday



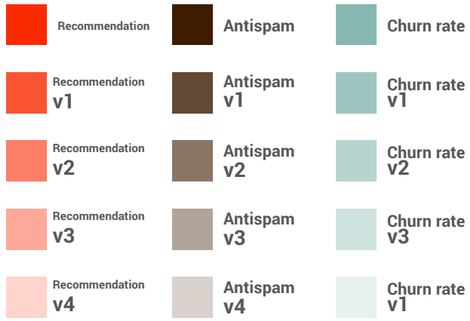
Wednesday



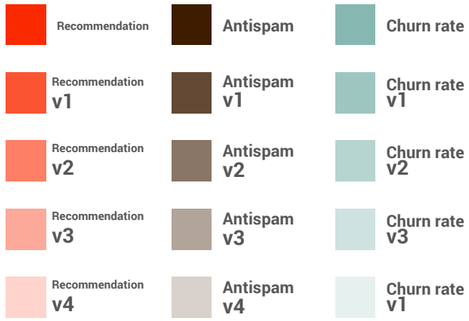
Thursday



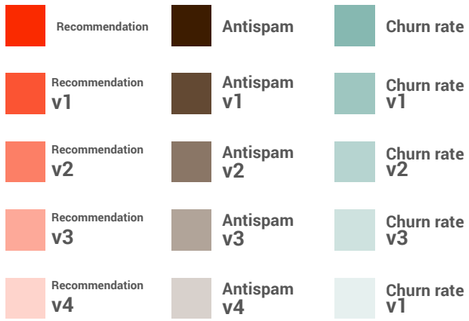
Friday



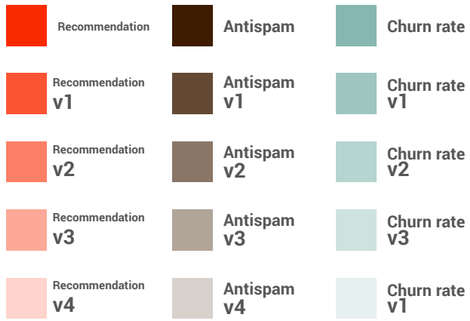
Saturday



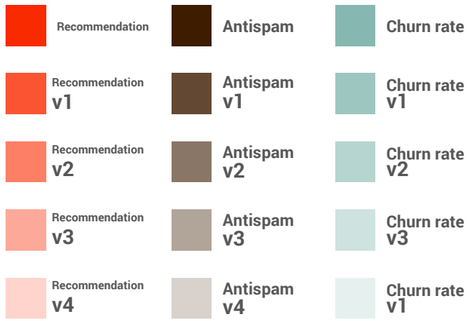
Monday



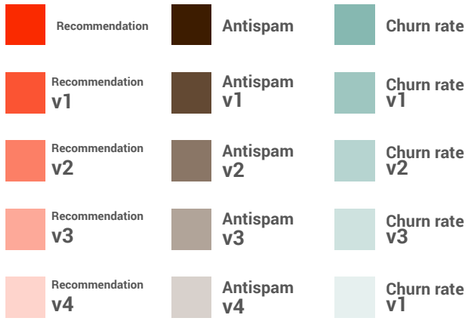
Tuesday



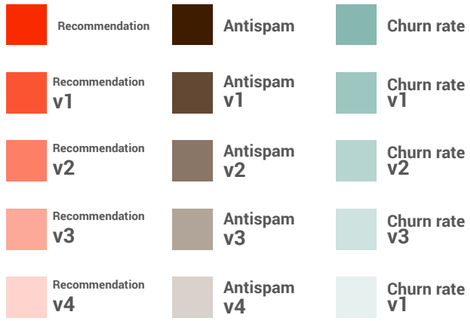
Wednesday



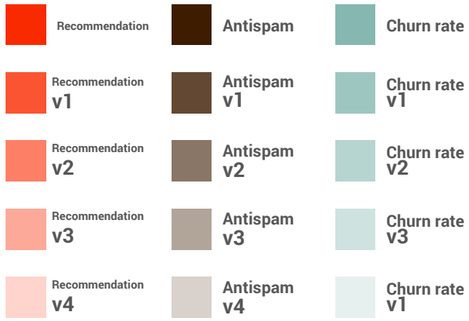
Thursday



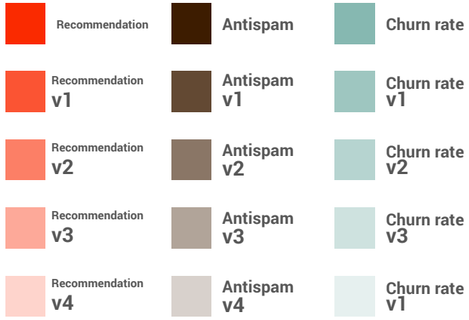
Friday



Saturday



Sunday



MANUAL AUTOMATION

SLOW!

SLOW!

ERROR PRONE!!

SLOW!

ERROR PRONE!!

BORING!!!

WHAT CAN YOU DO?

WHAT IF...

YOU COULD USE  ?

RENT KIT BY THE HOUR

RENT KIT BY THE HOUR

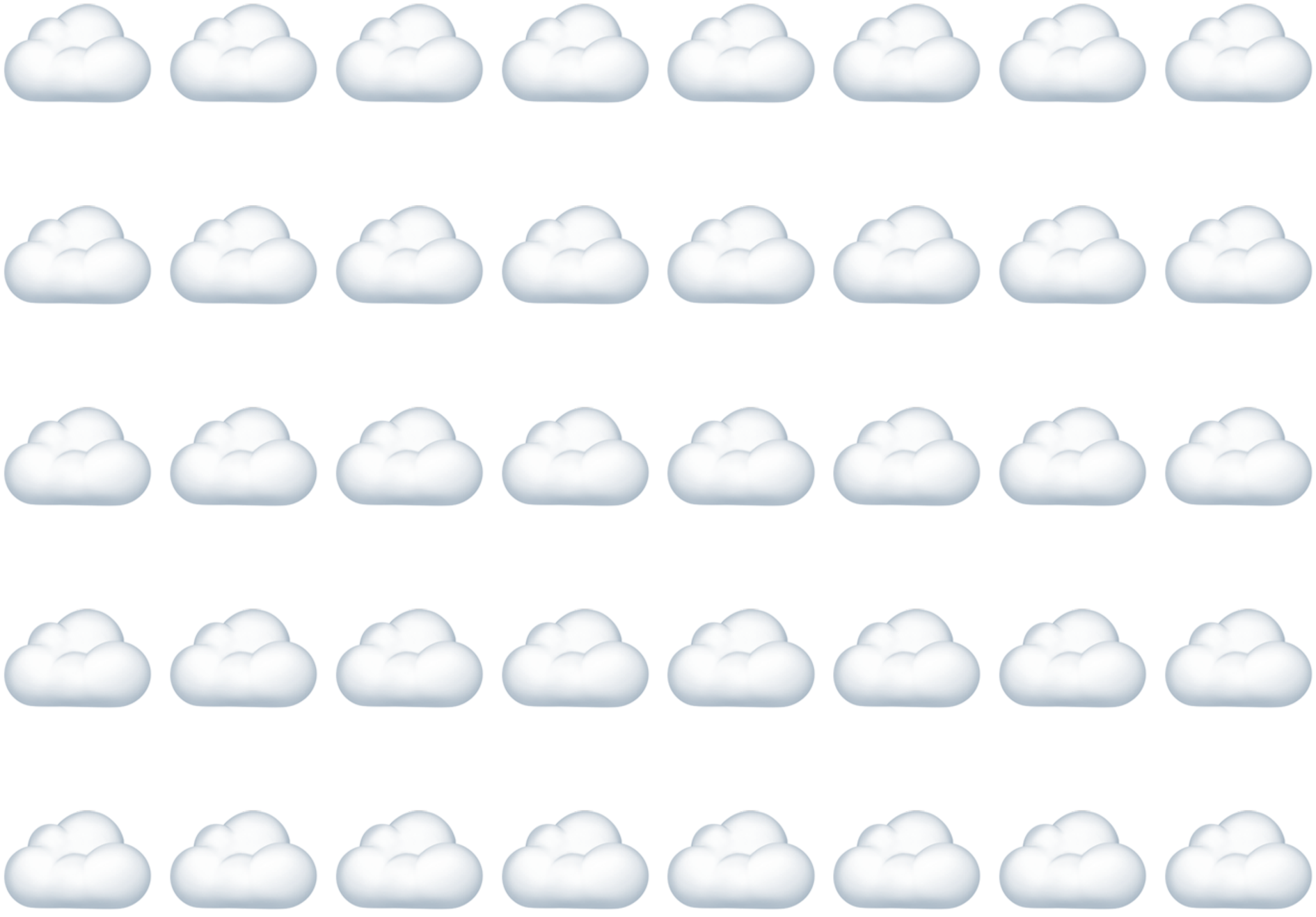
BIGGER MACHINES

RENT KIT BY THE HOUR

BIGGER MACHINES

SCALE WITH THE TEAM

WHAT IF...



RENT KIT BY THE HOUR

~~RENT KIT BY THE HOUR~~

RENT FLEET OF MACHINE BY THE HOUR

RENT FLEET OF MACHINE BY THE HOUR

BIGGER MACHINES

RENT FLEET OF MACHINE BY THE HOUR

~~BIGGER MACHINES~~

MIXED NODES W/ CPUs, GPUs, TPUs

RENT FLEET OF MACHINE BY THE HOUR

MIXED NODES W/ CPUS, GPUS, TPUS

SCALE WITH THE TEAM

RENT FLEET OF MACHINE BY THE HOUR

MIXED NODES W/ CPUs, GPUs, TPUs

~~SCALE WITH THE TEAM~~

RENT FLEET OF MACHINE BY THE HOUR

MIXED NODES W/ CPUS, GPUS, TPUS

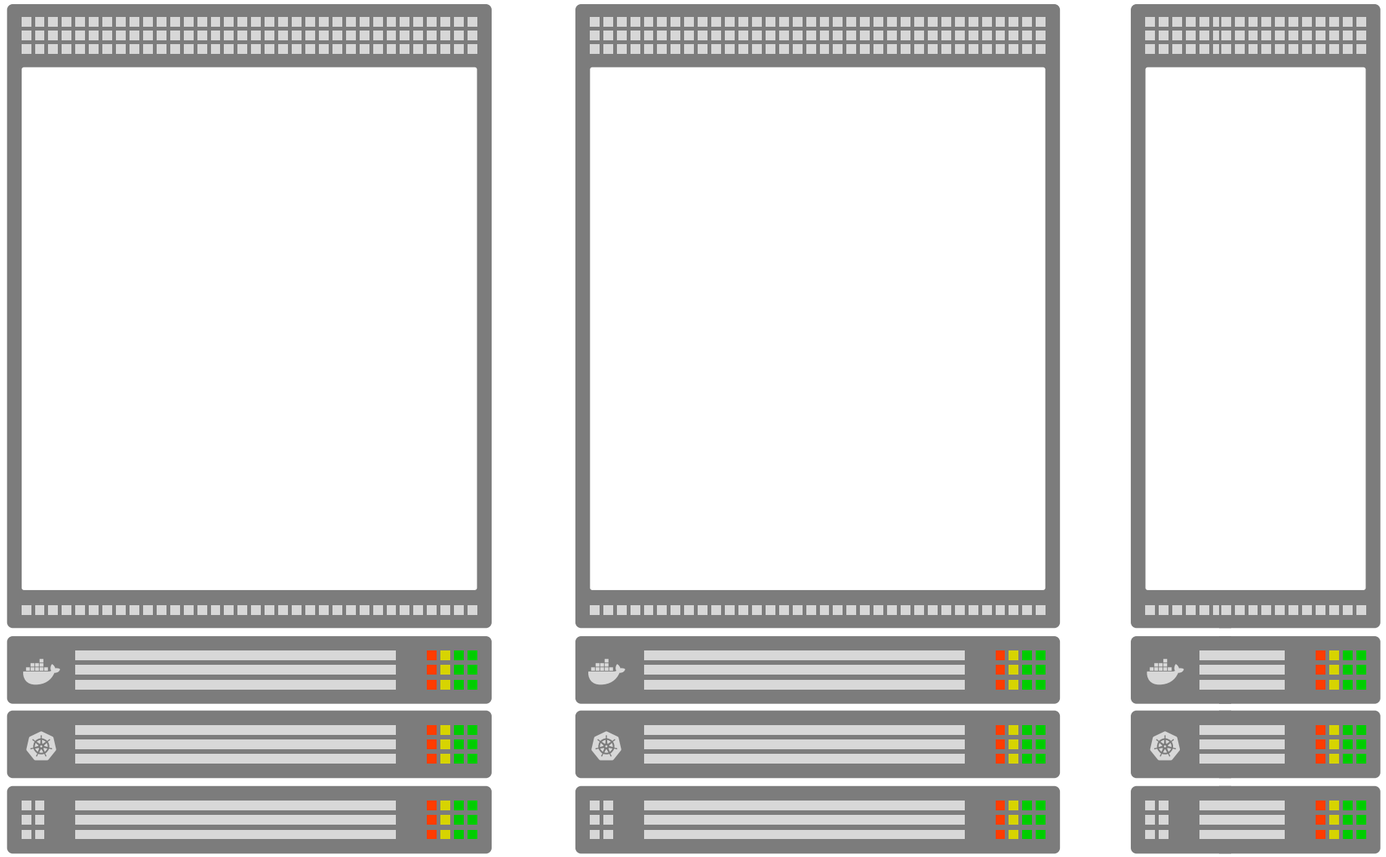
SCALE WITH THE COMPANY

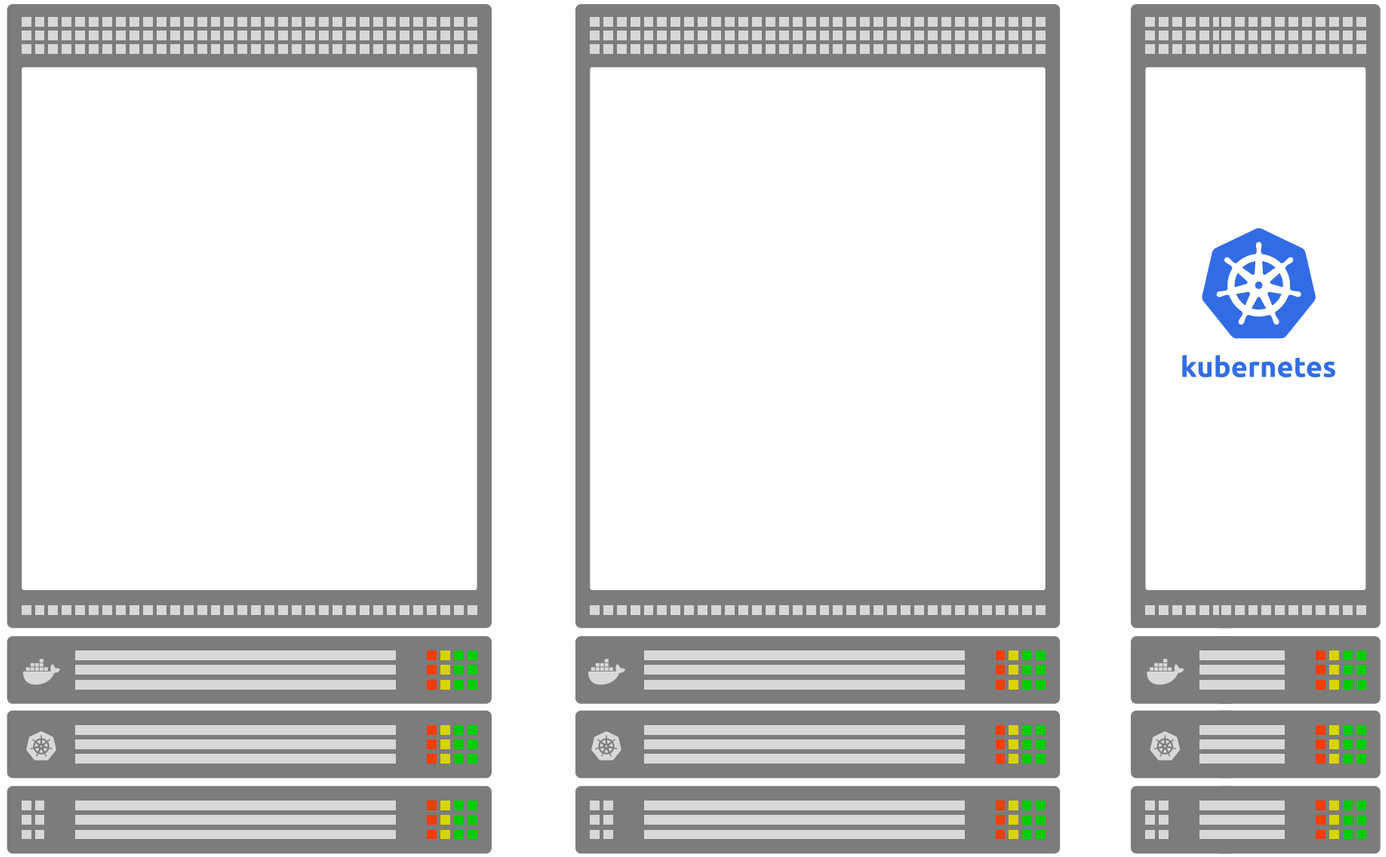
WHAT IF...

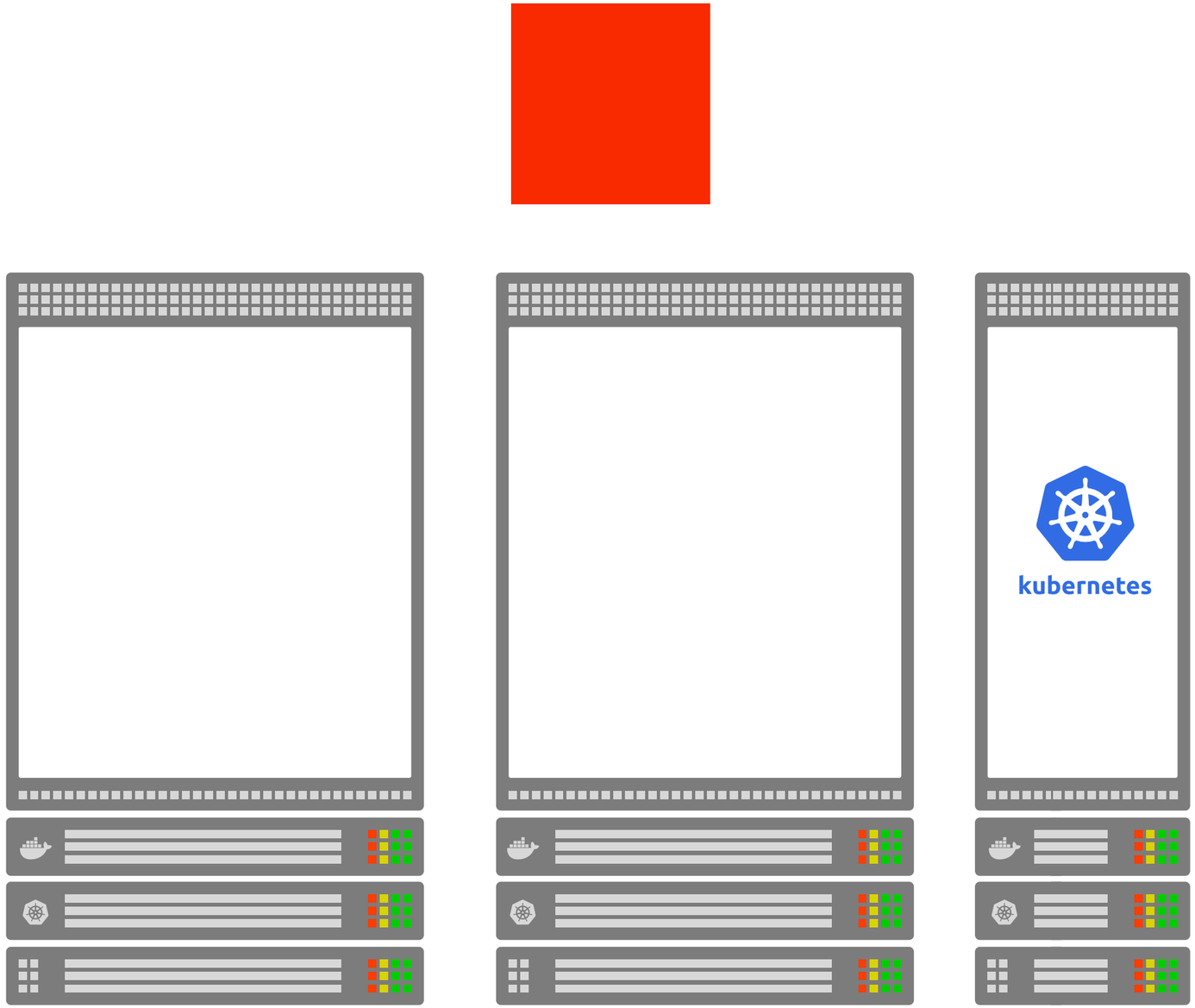
THIS IS READY TO BE USED?

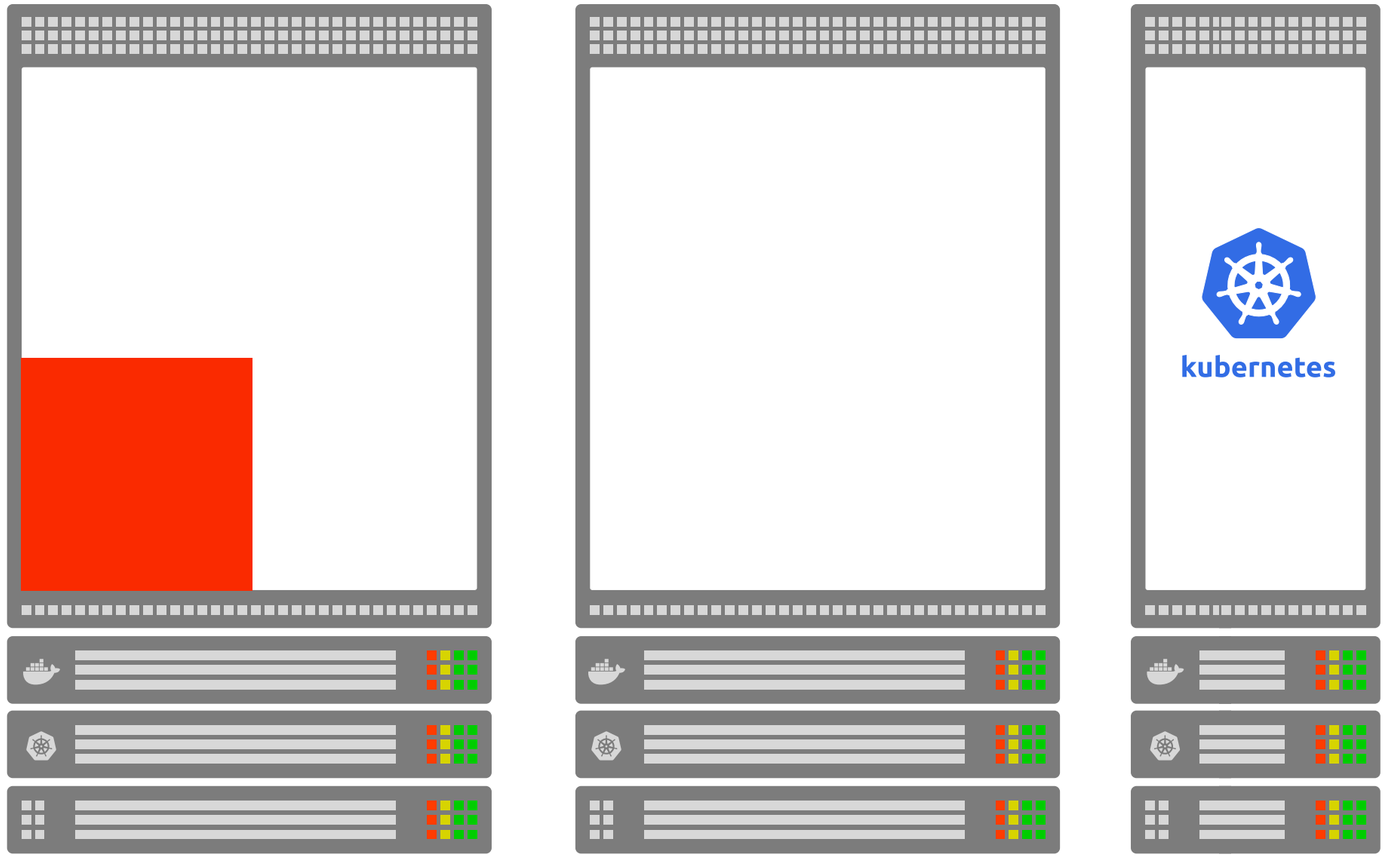
kubernetes

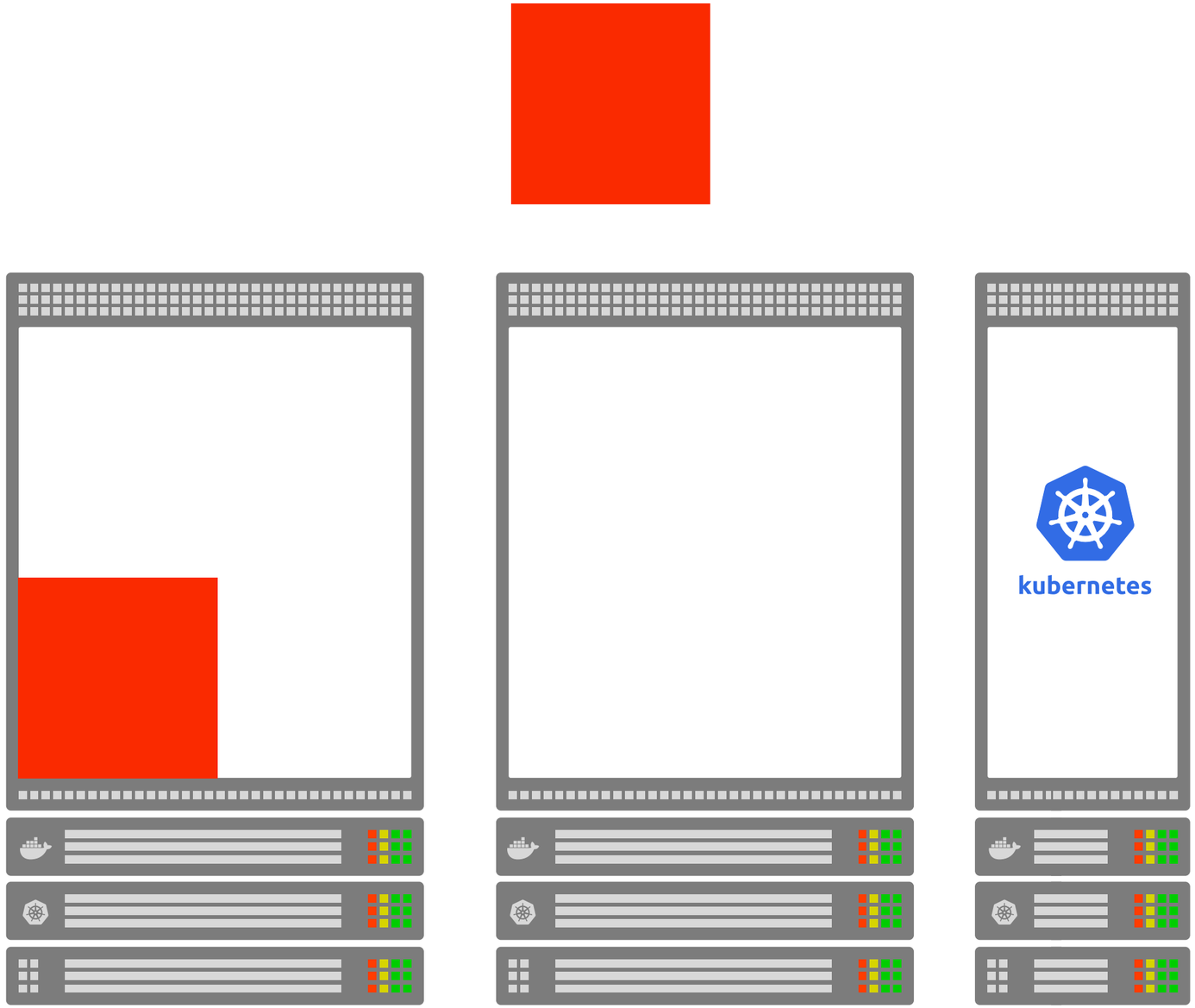
WHAT IS IT?

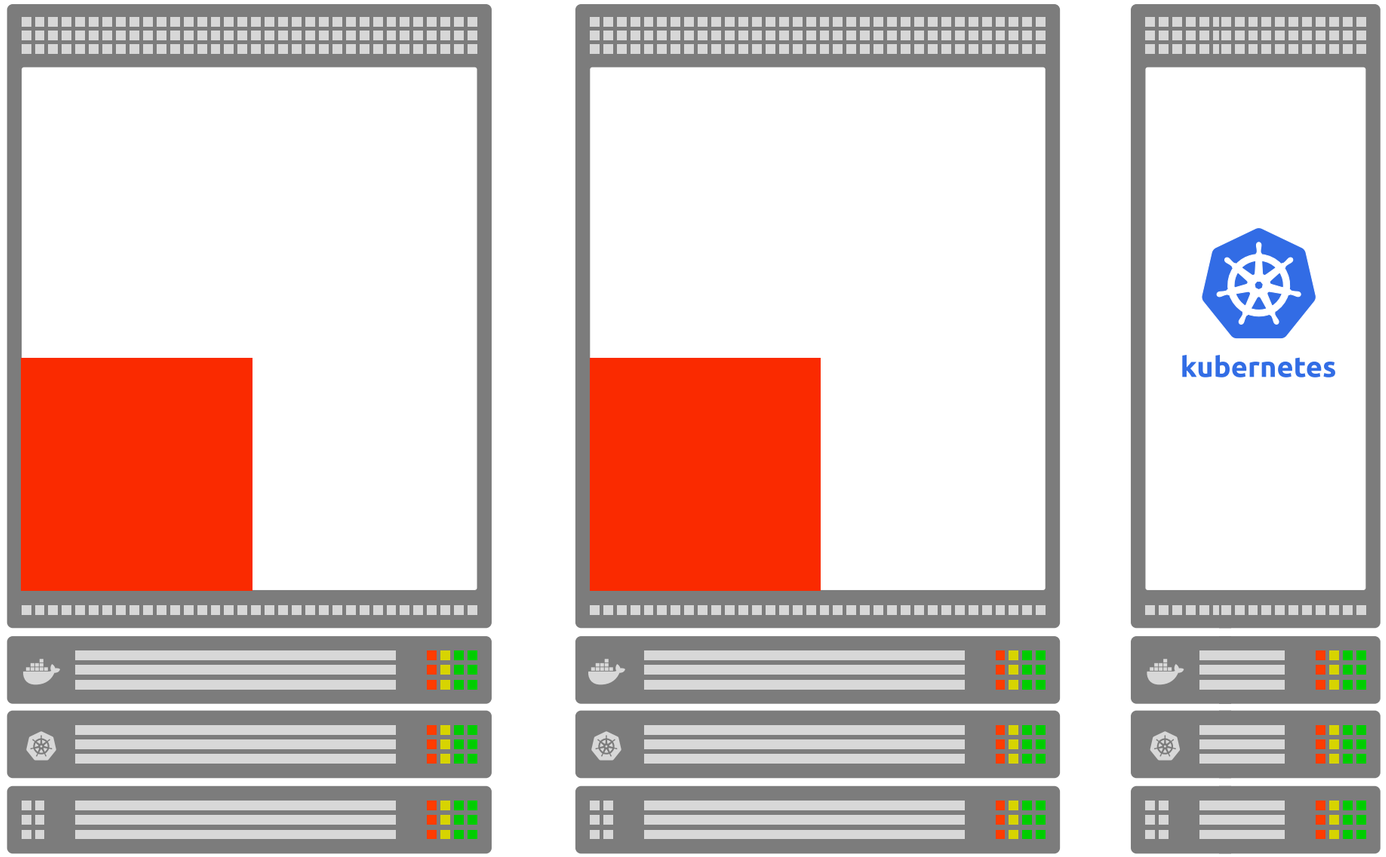


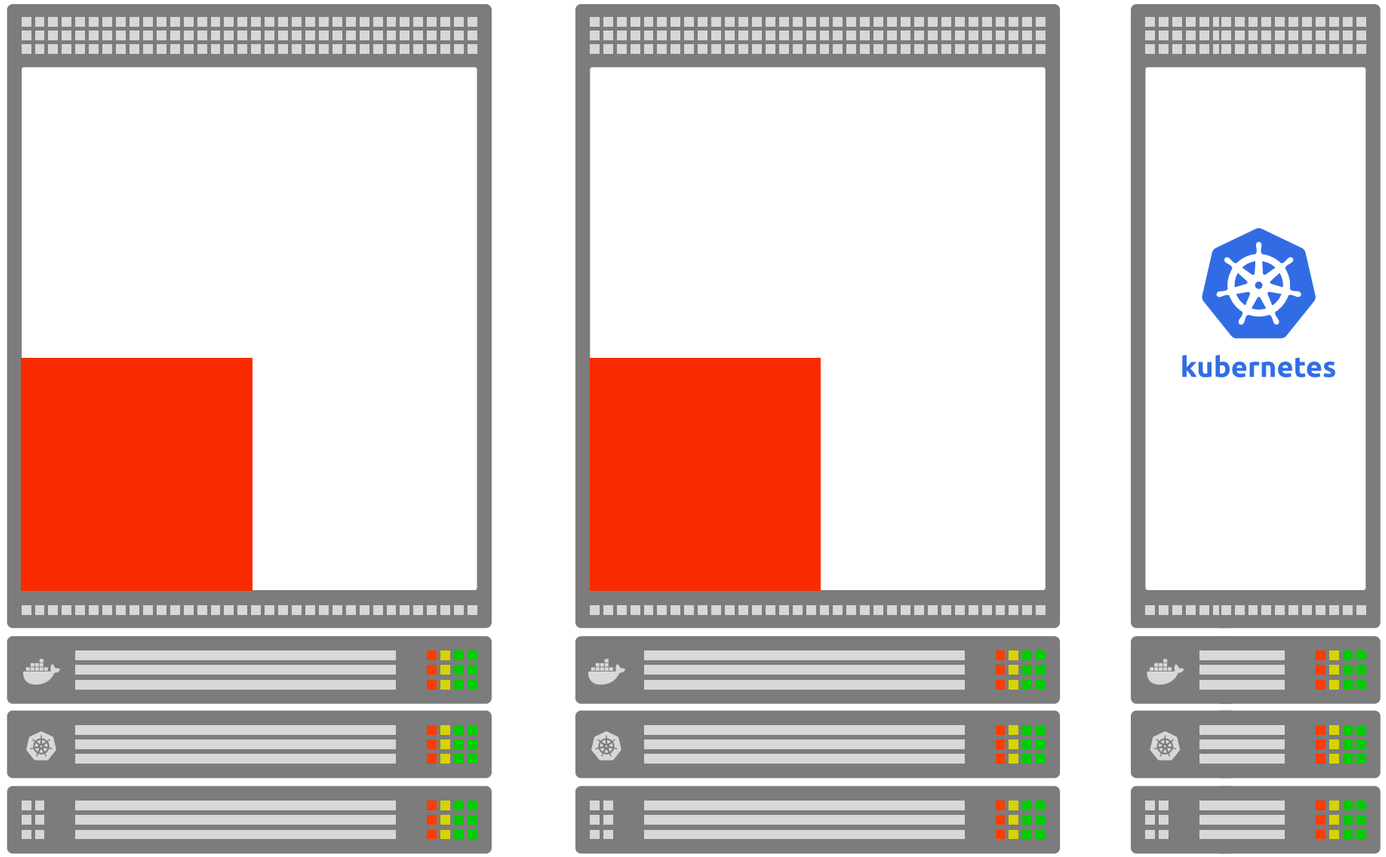

kubernetes

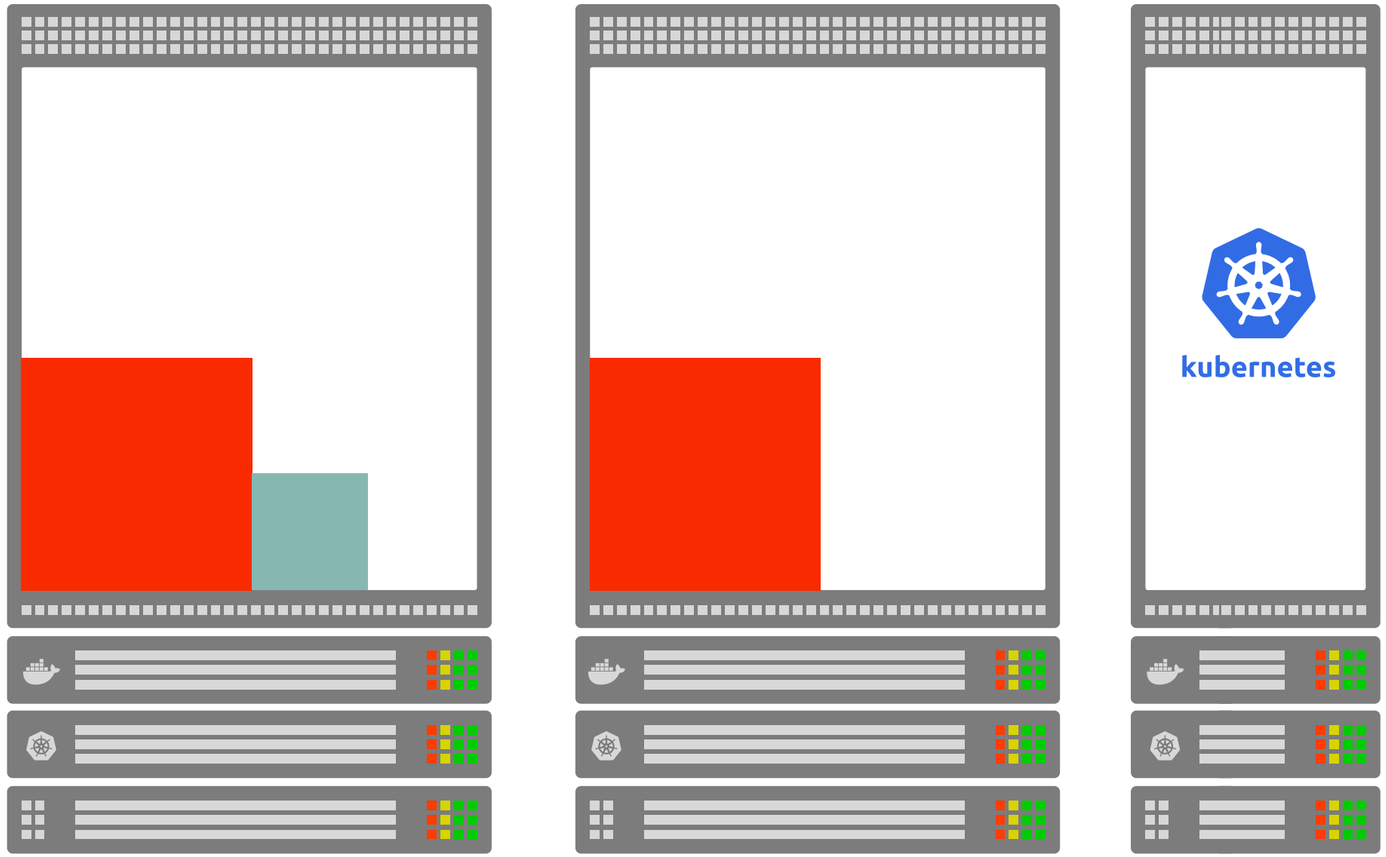

kubernetes

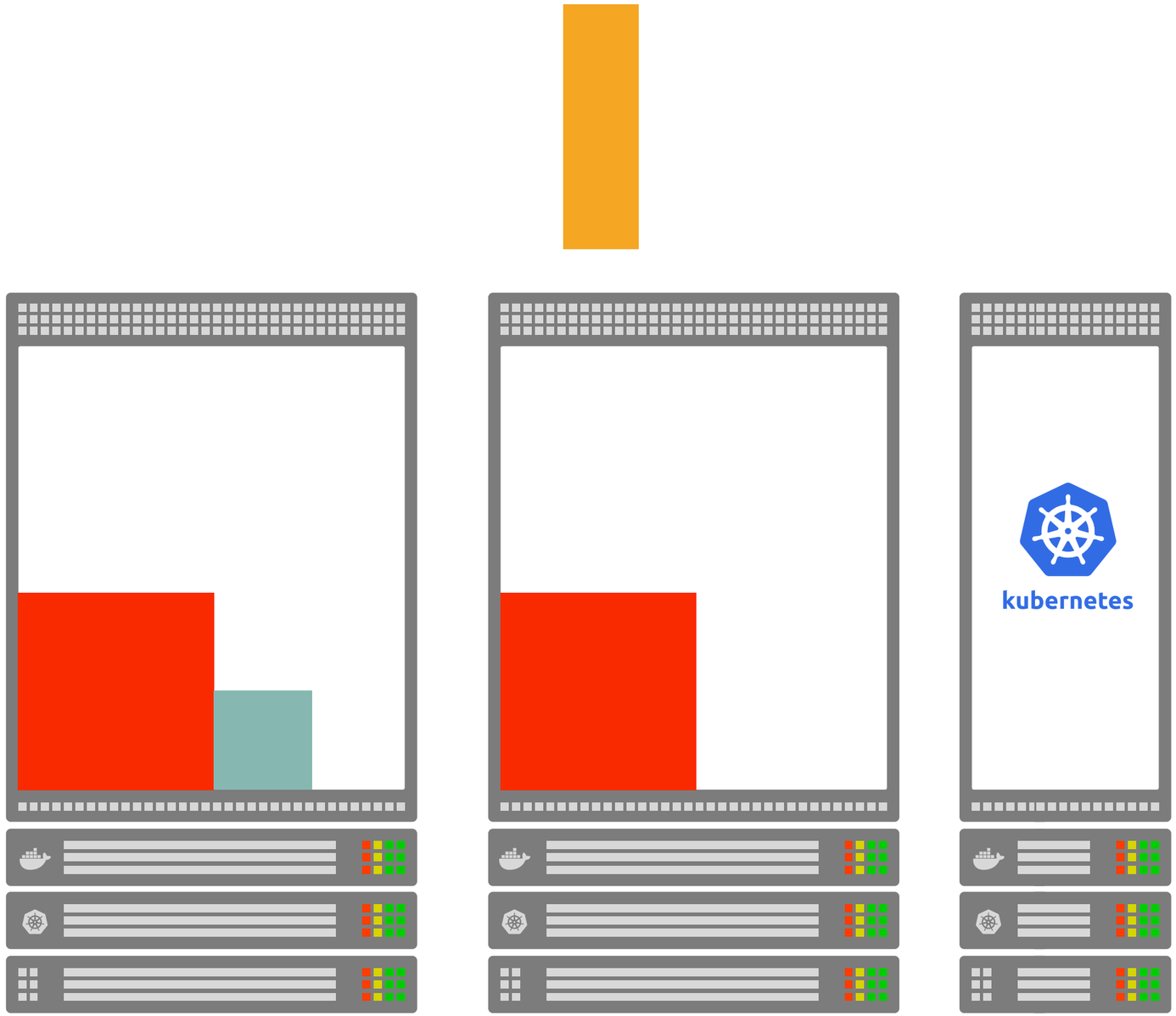

kubernetes

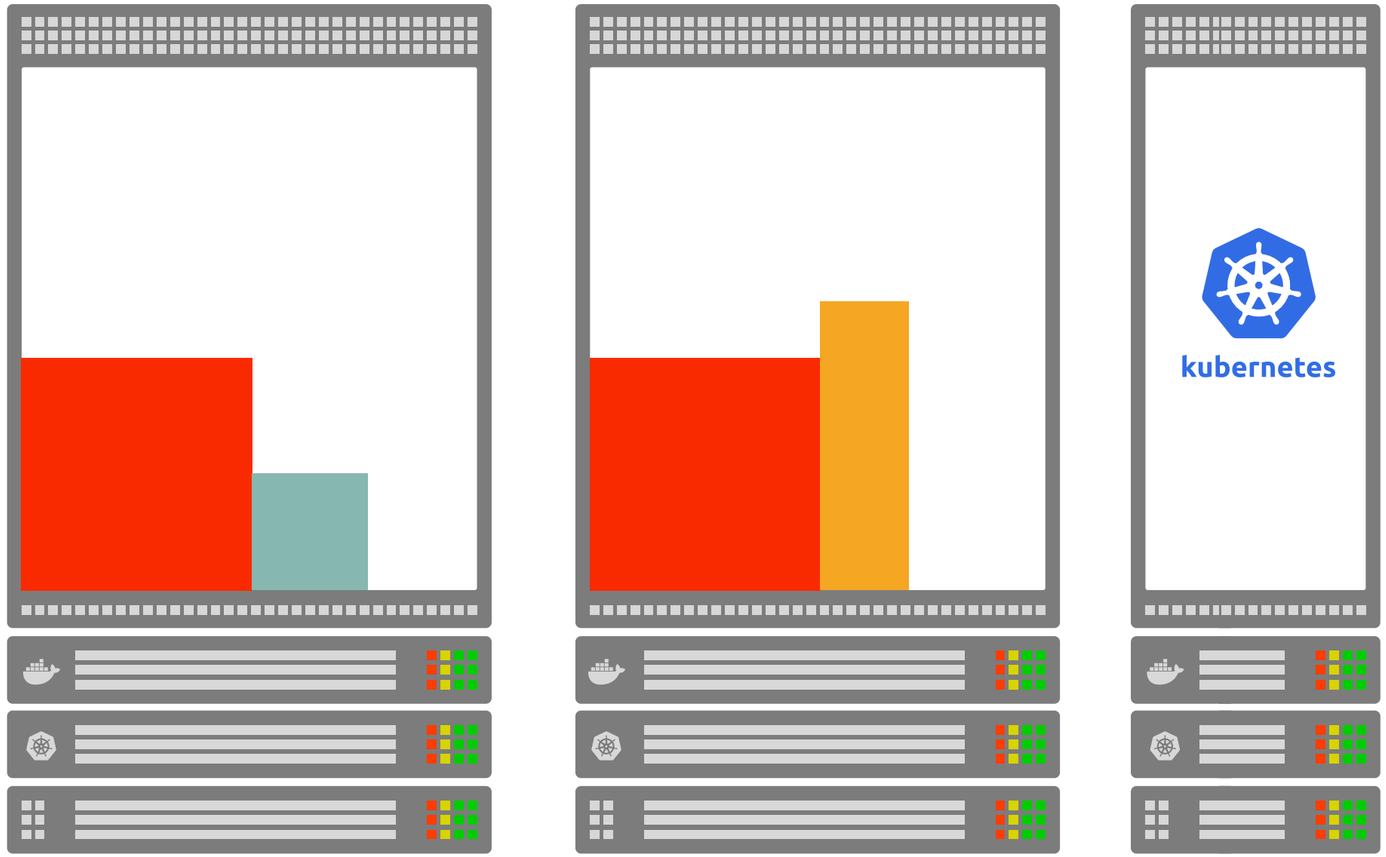

kubernetes

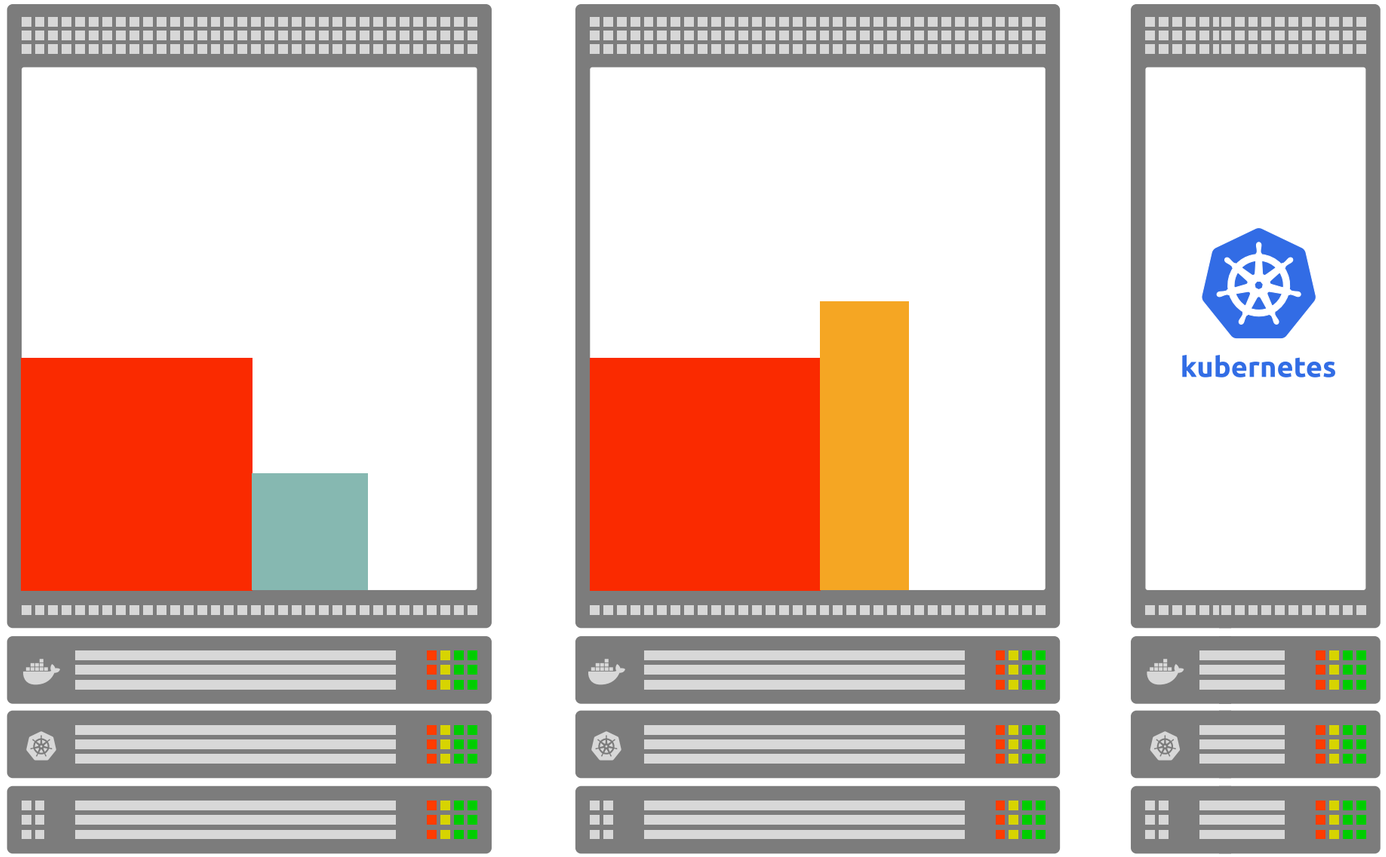

kubernetes

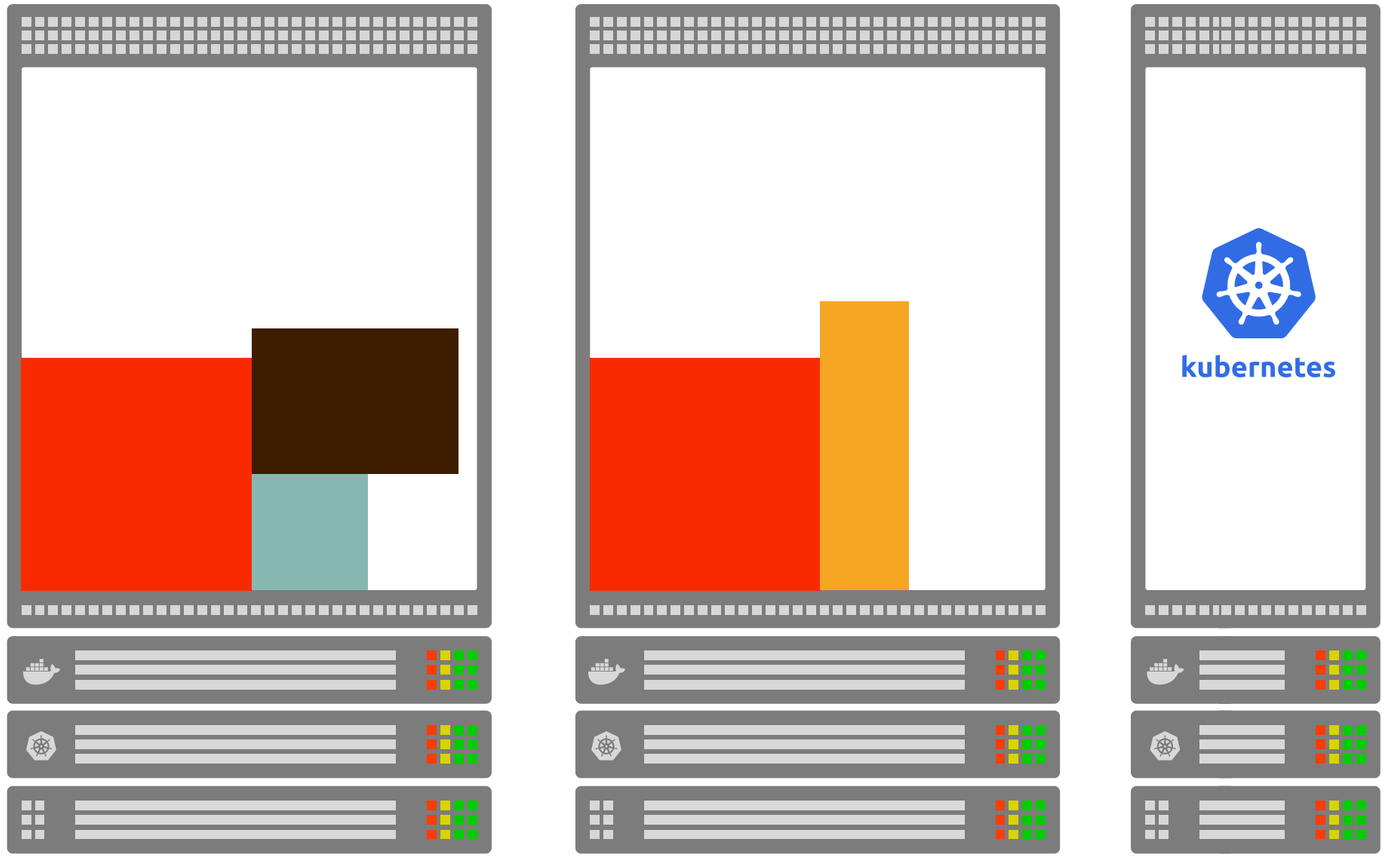

kubernetes

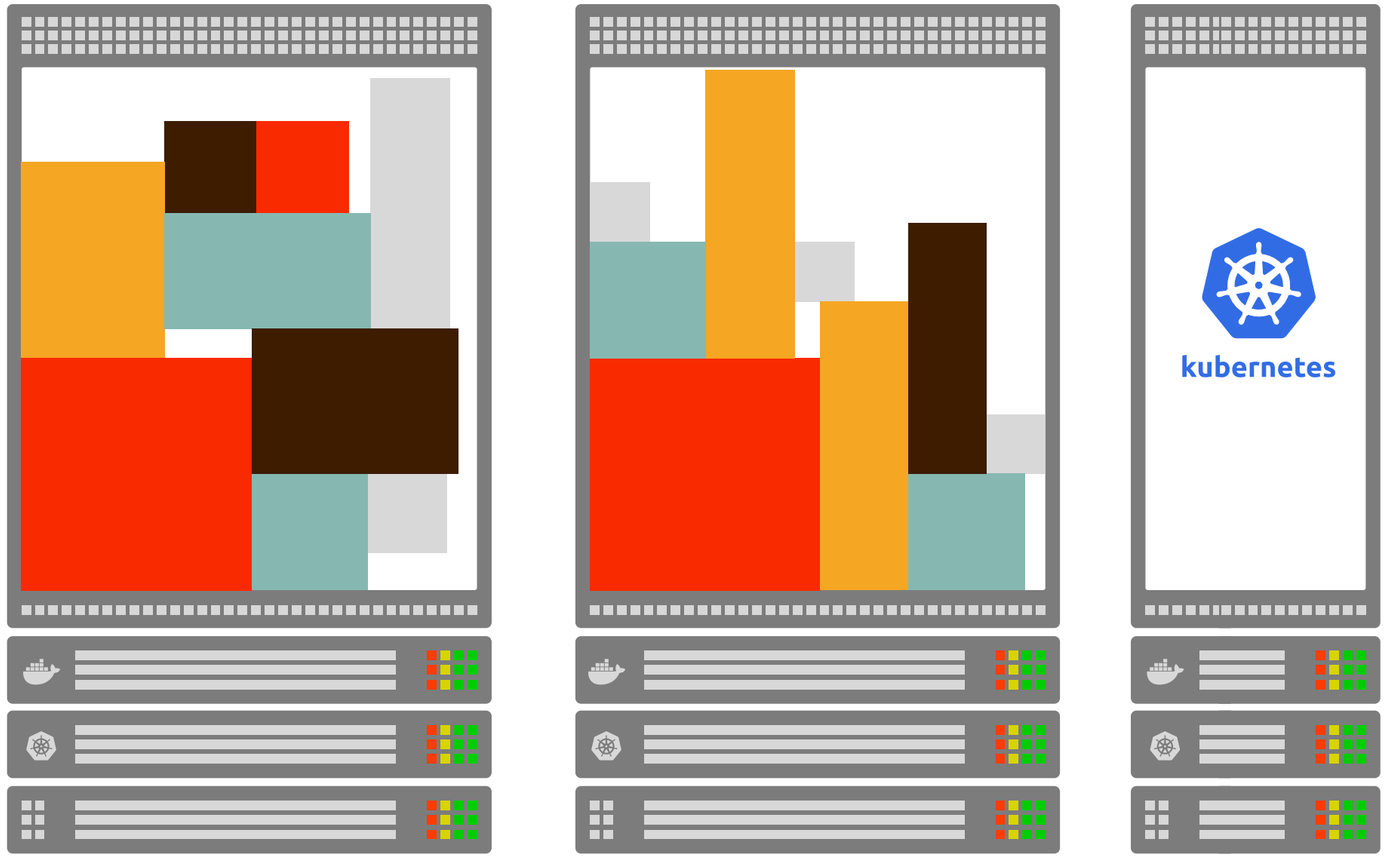

kubernetes


kubernetes


kubernetes

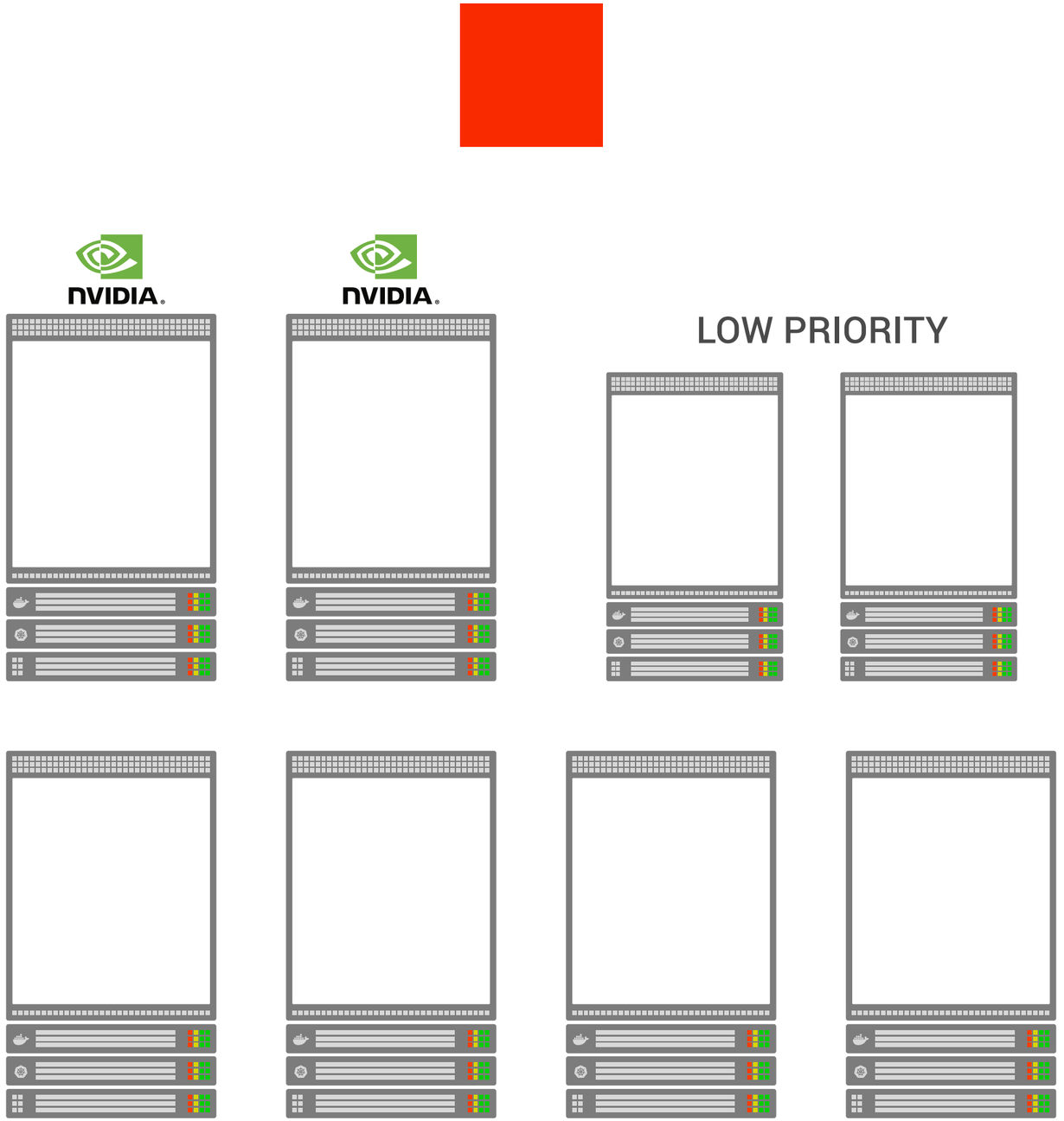

kubernetes


kubernetes

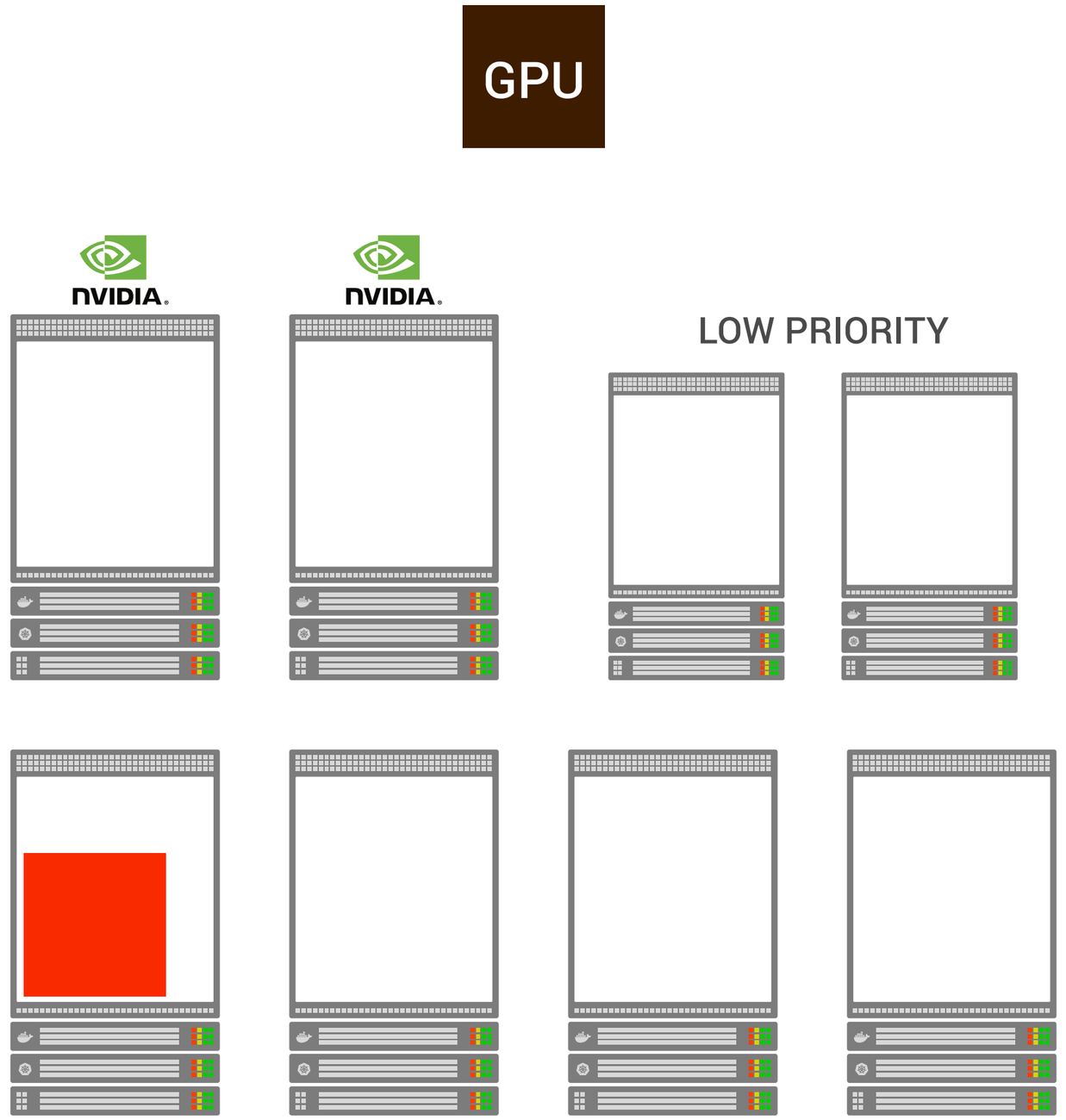

kubernetes

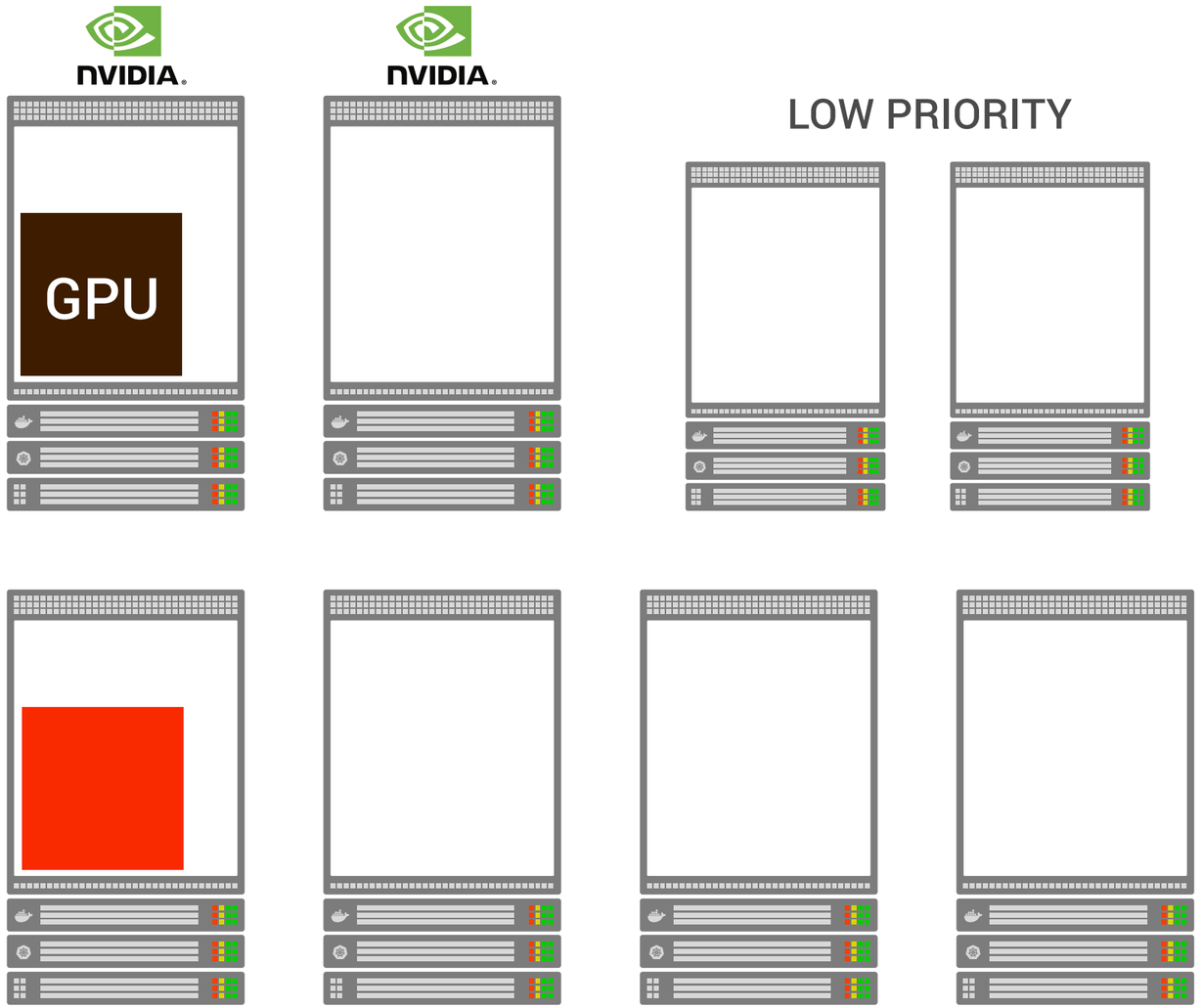
ALSO


NVIDIA
NVIDIA
LOW PRIORITY

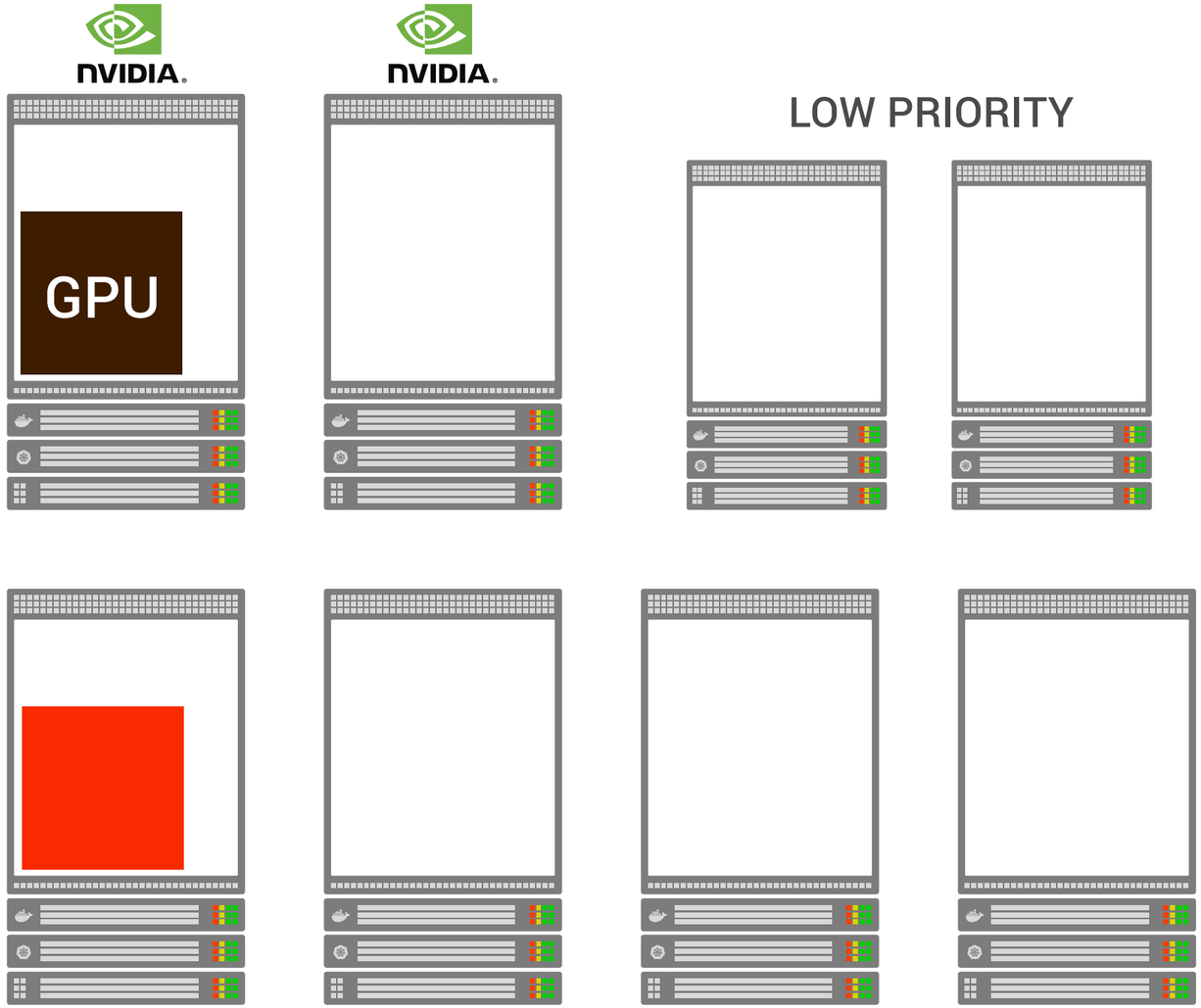

LOW PRIORITY

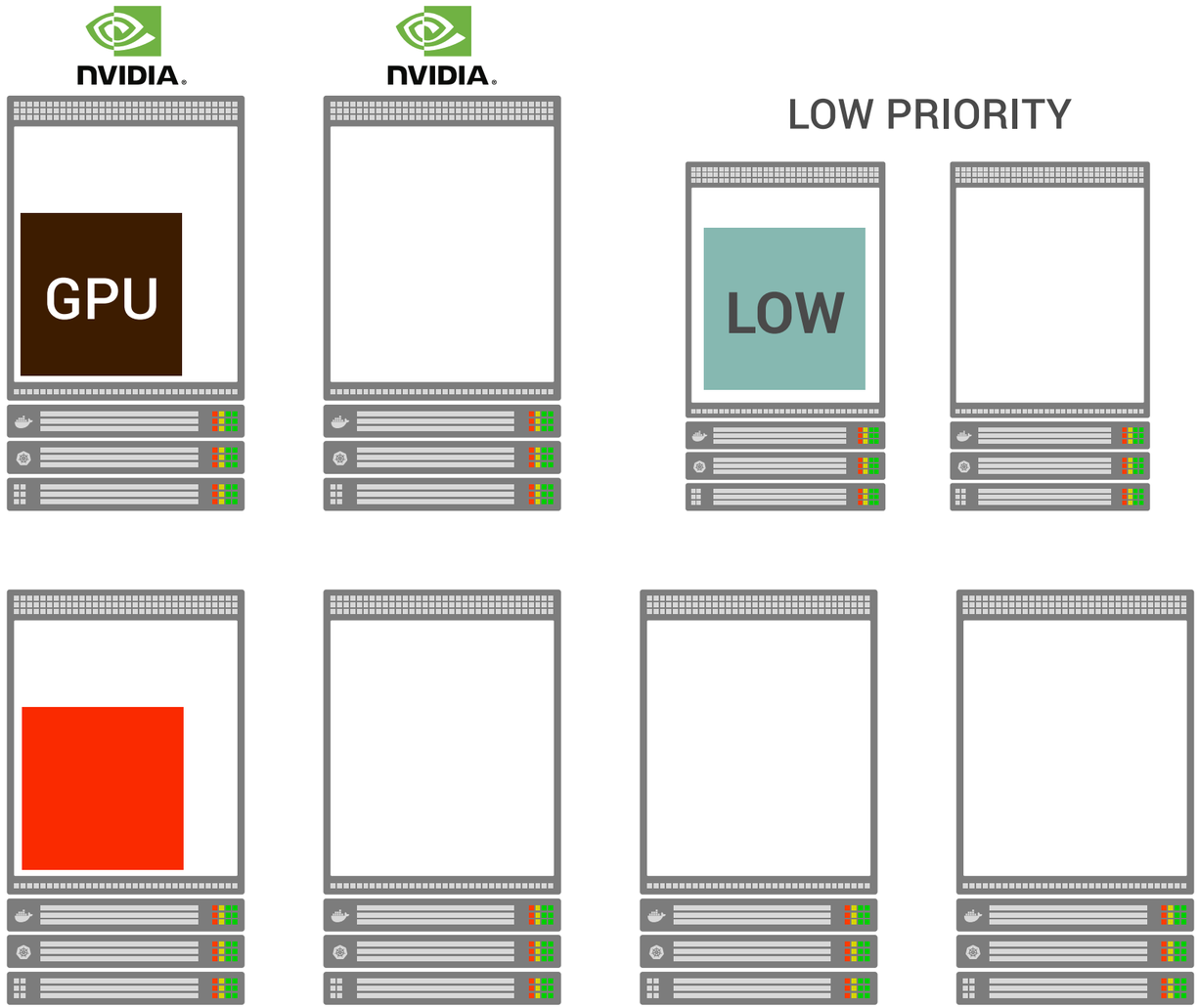

LOW PRIORITY


GPU
NVIDIA
NVIDIA
LOW PRIORITY

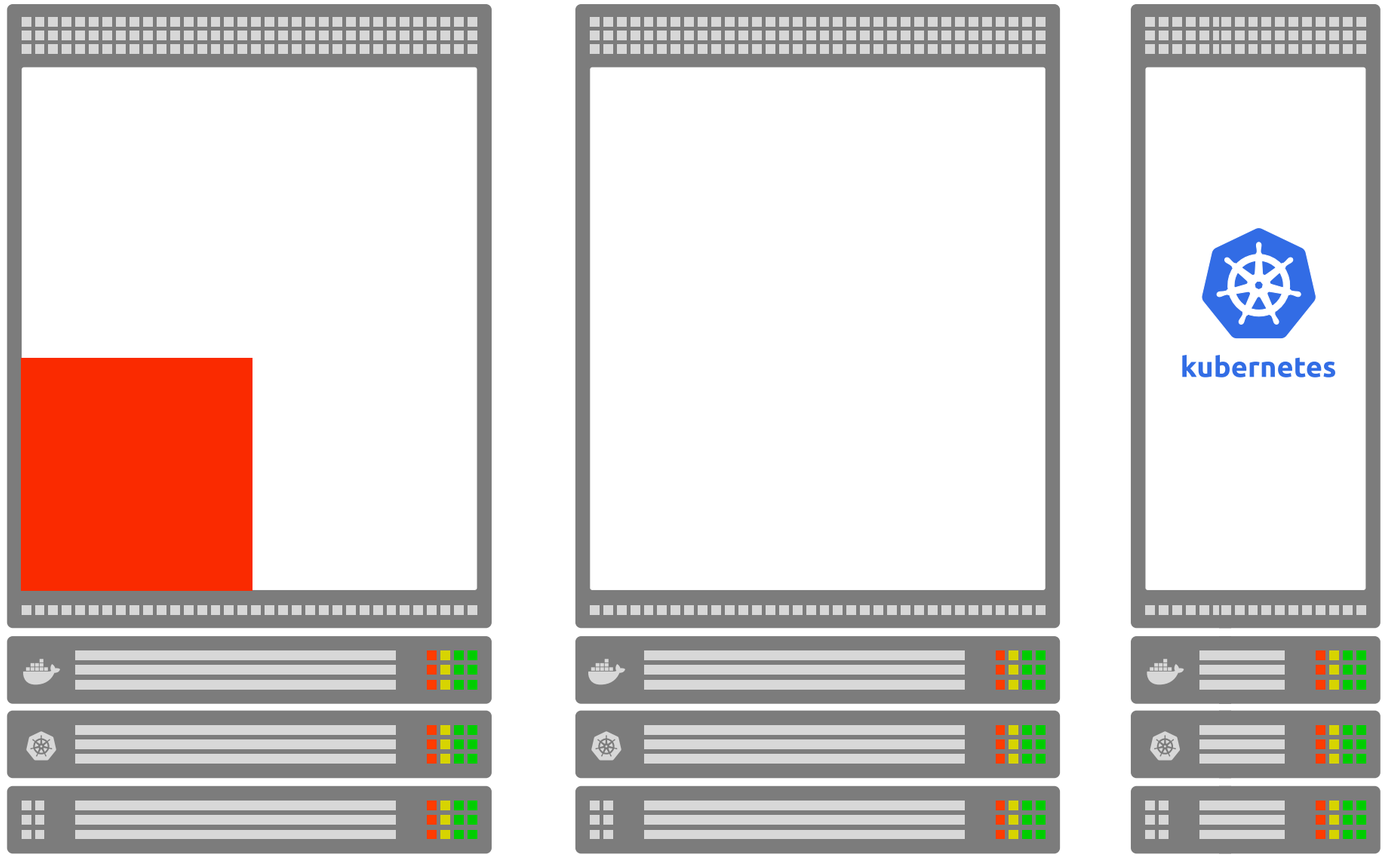

NVIDIA
NVIDIA
LOW PRIORITY
GPU

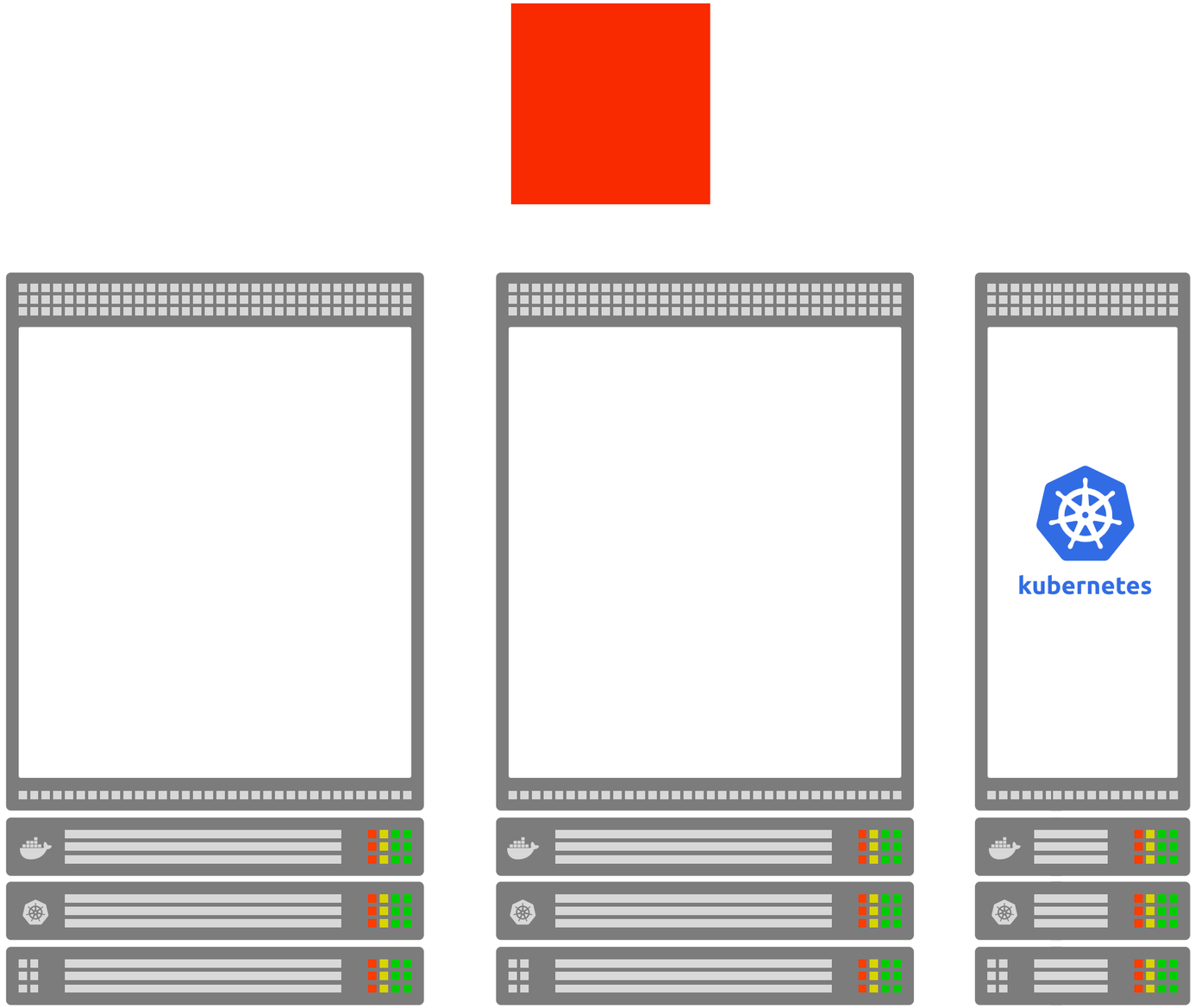
LOW

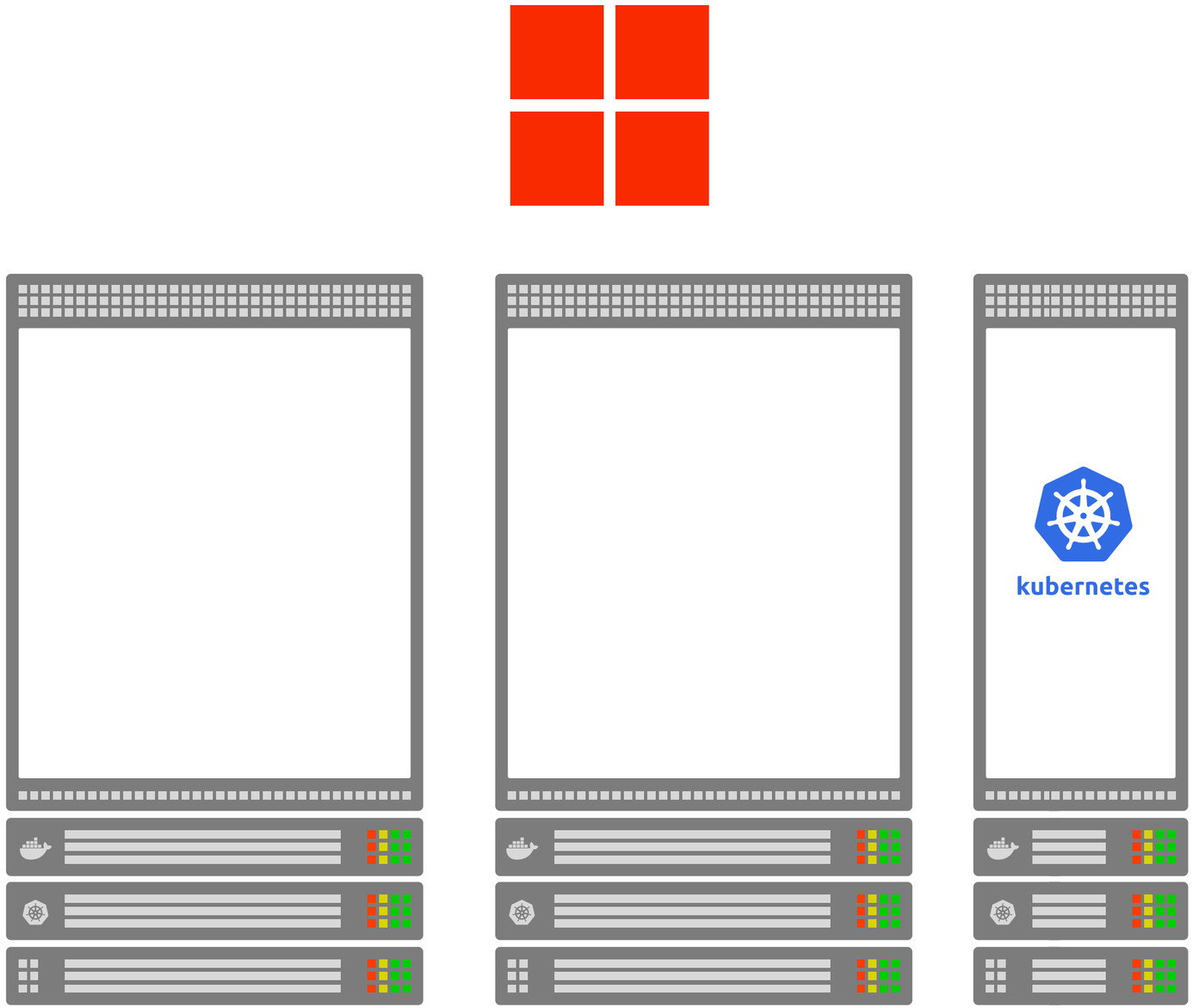


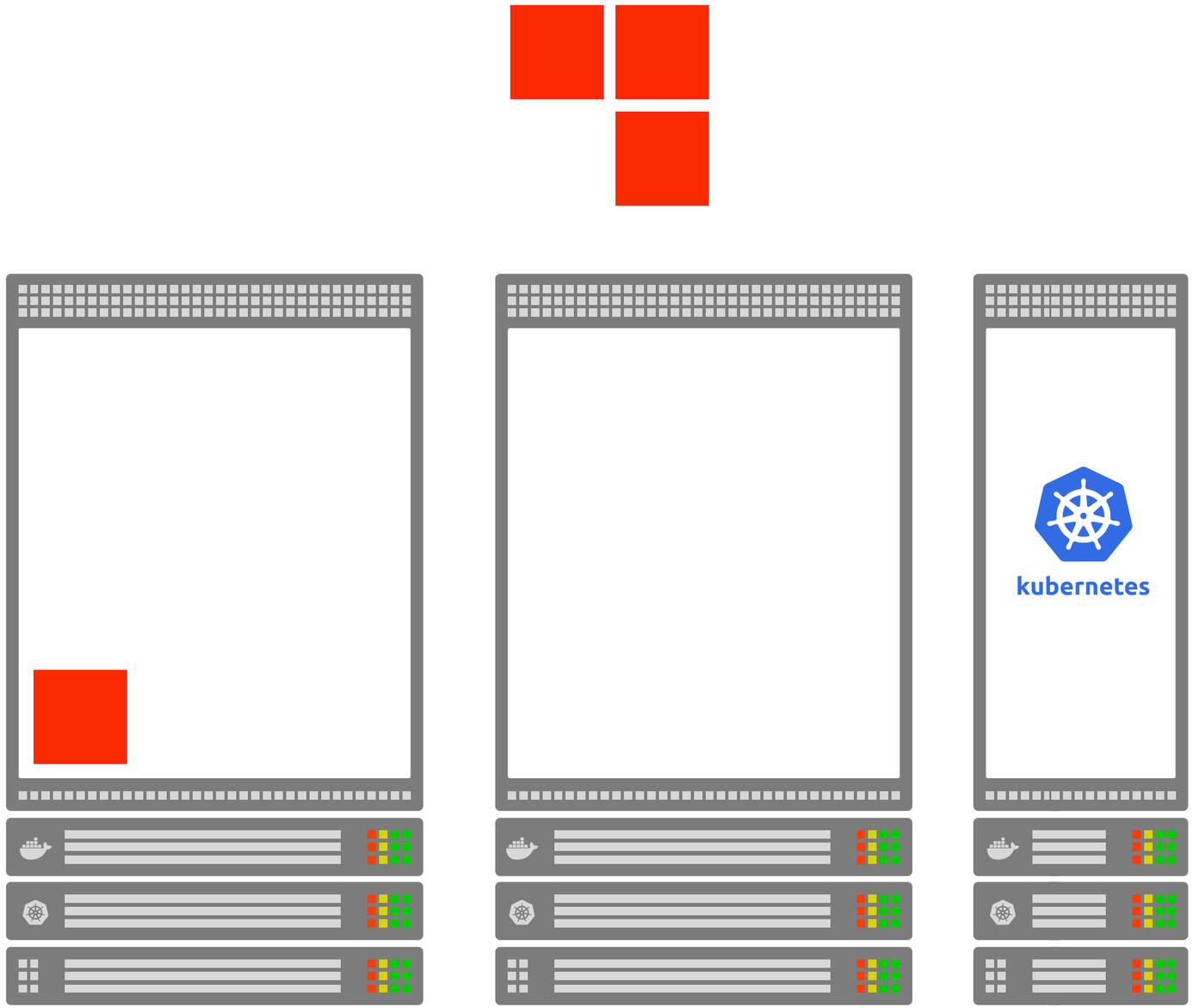

NVIDIA
NVIDIA
LOW PRIORITY
GPU
LOW

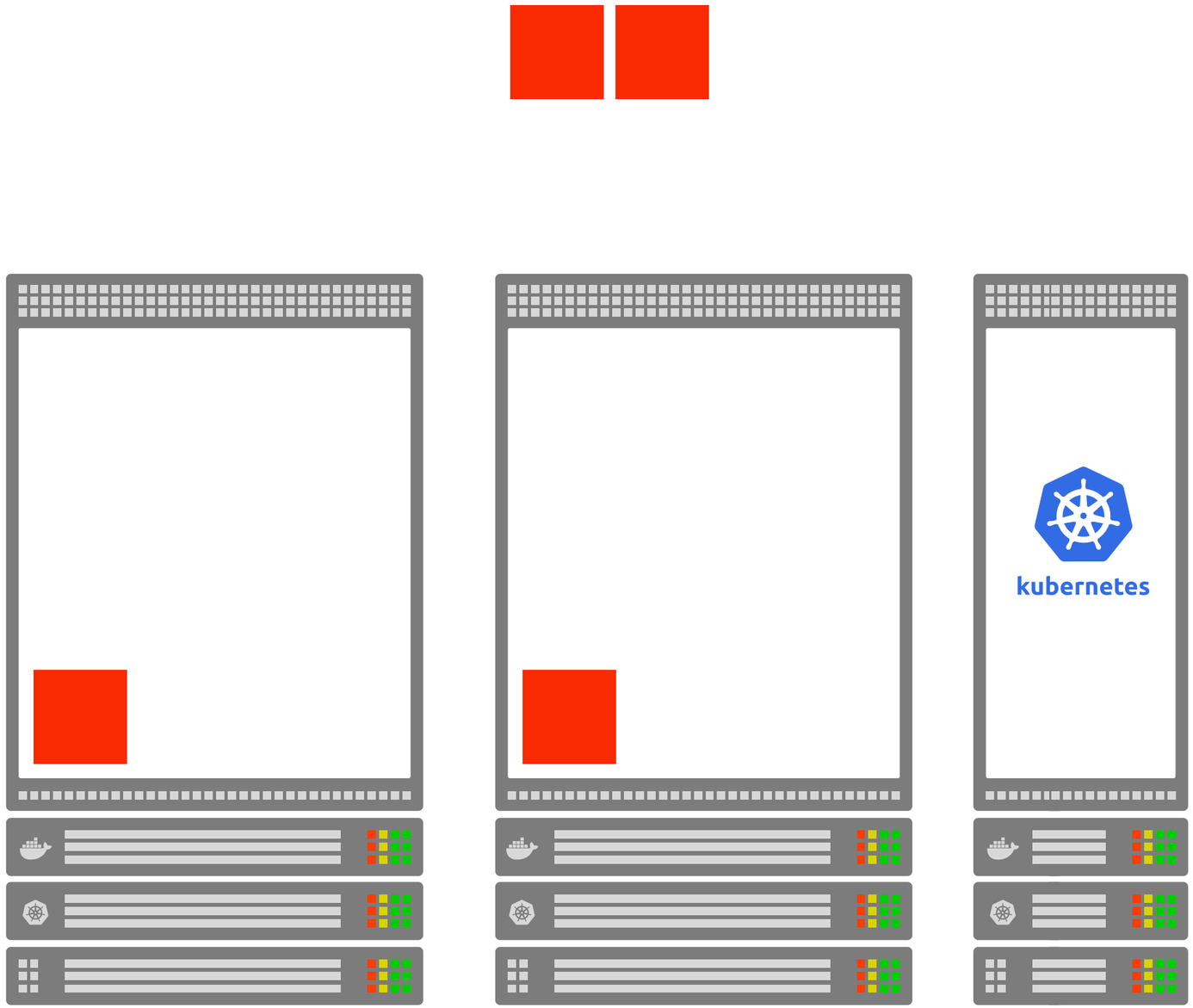
STILL SINGLE JOB...

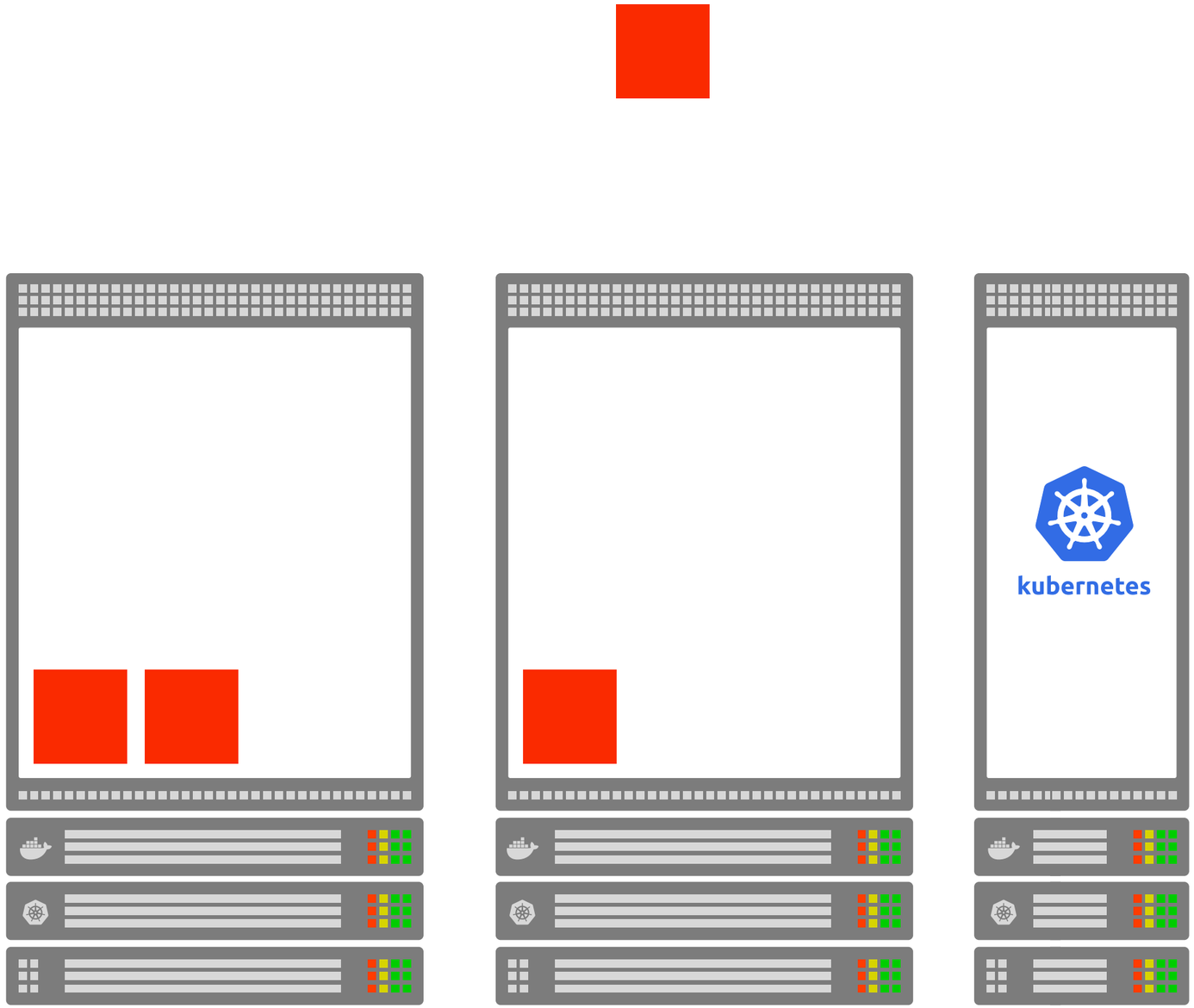

kubernetes

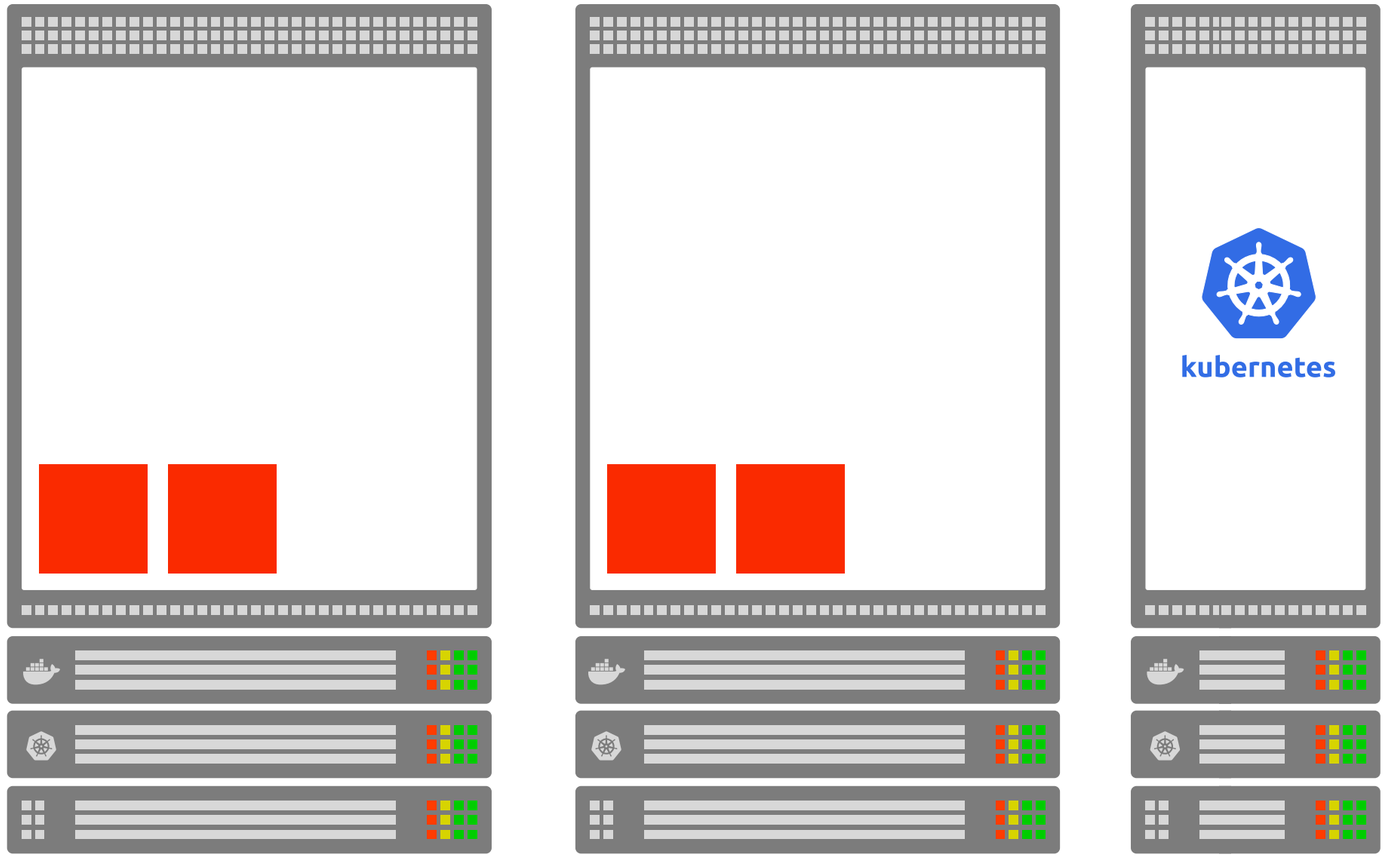

kubernetes


kubernetes


kubernetes


kubernetes

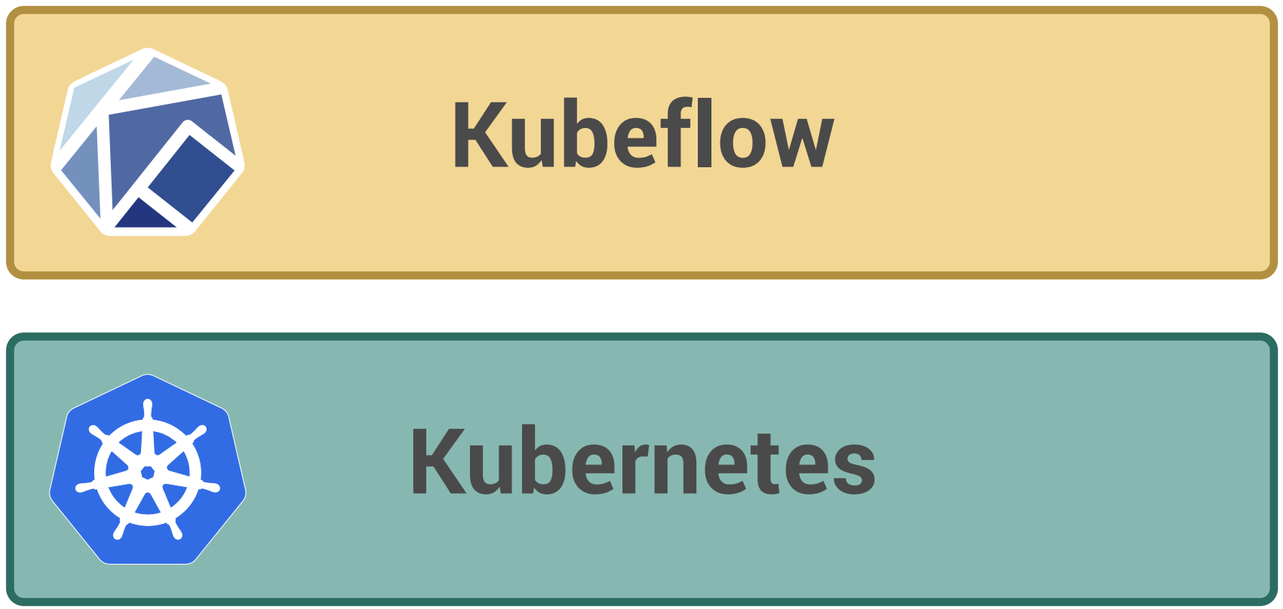

kubernetes


kubernetes

HOW CAN I  JOBS?

Kubeflow

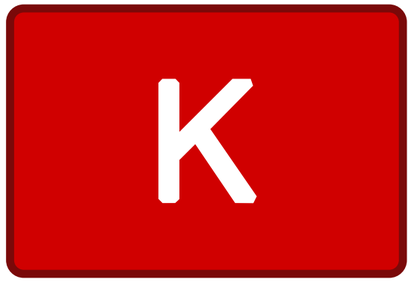
Kubernetes


Kubeflow
Kubernetes

K

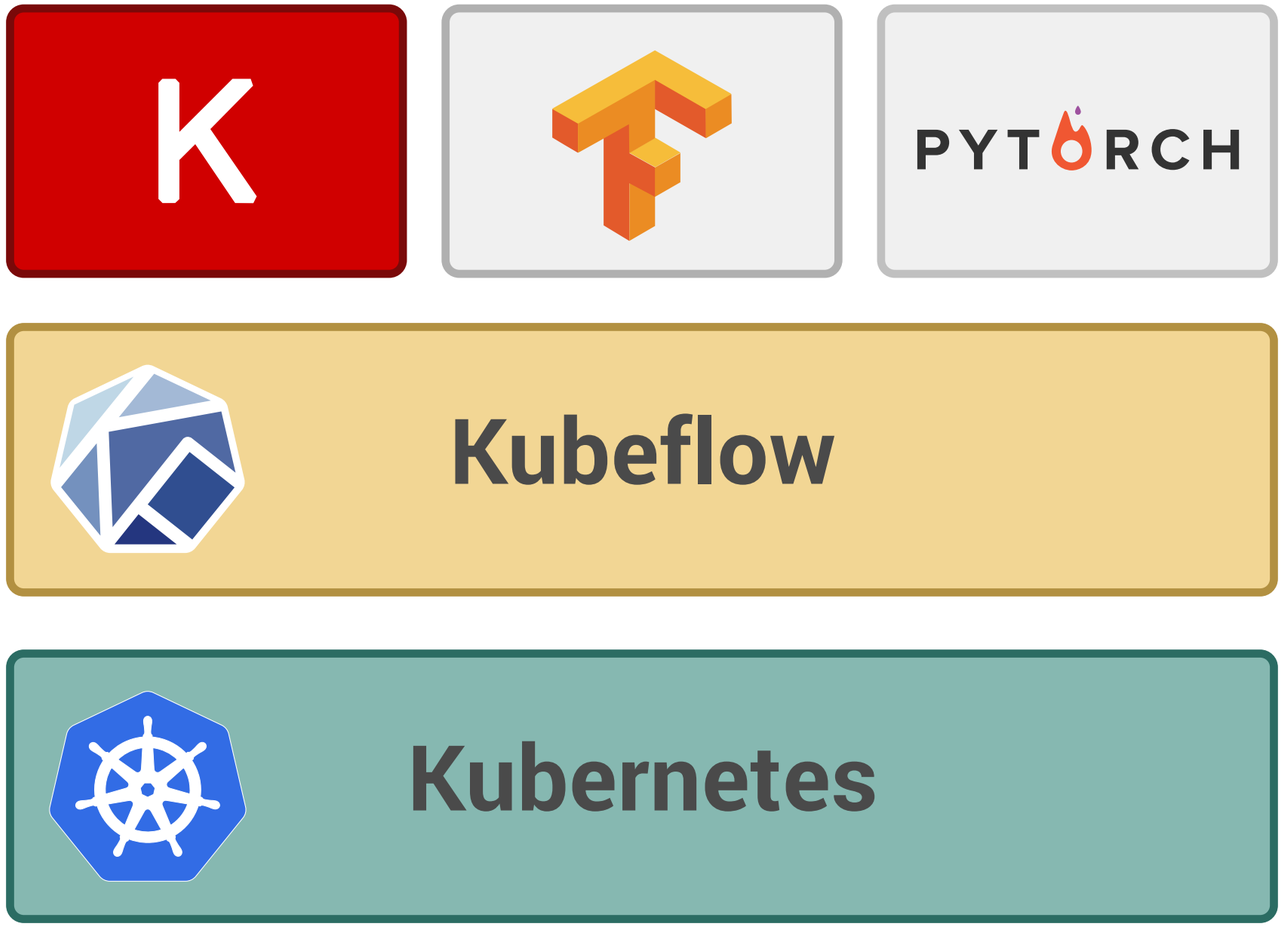
Kubeflow

Kubernetes



Kubeflow

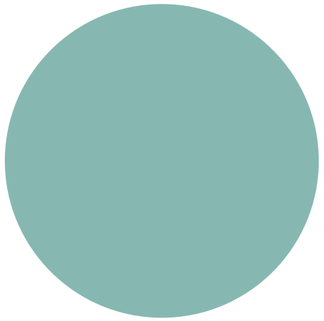
Kubernetes

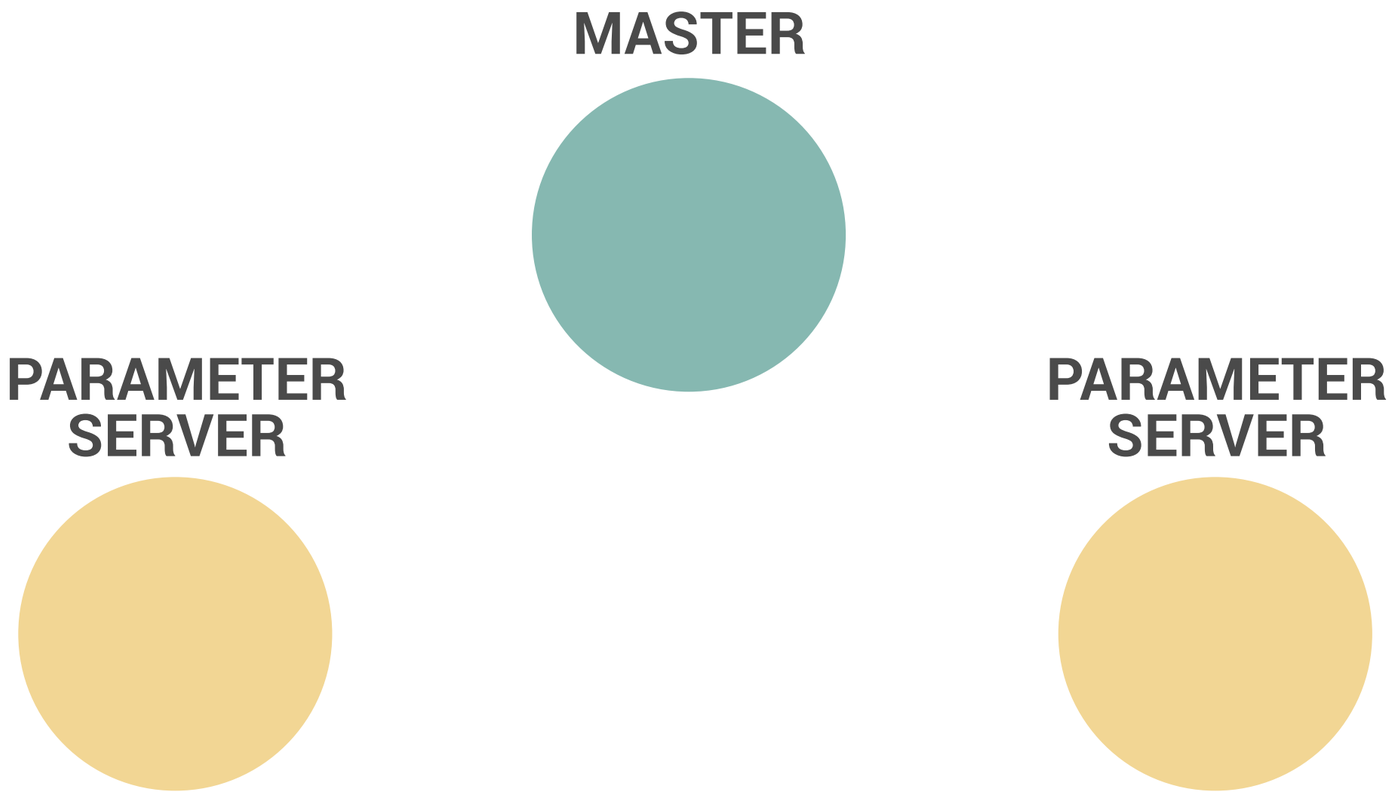

K
PYTORCH
Kubeflow
Kubernetes

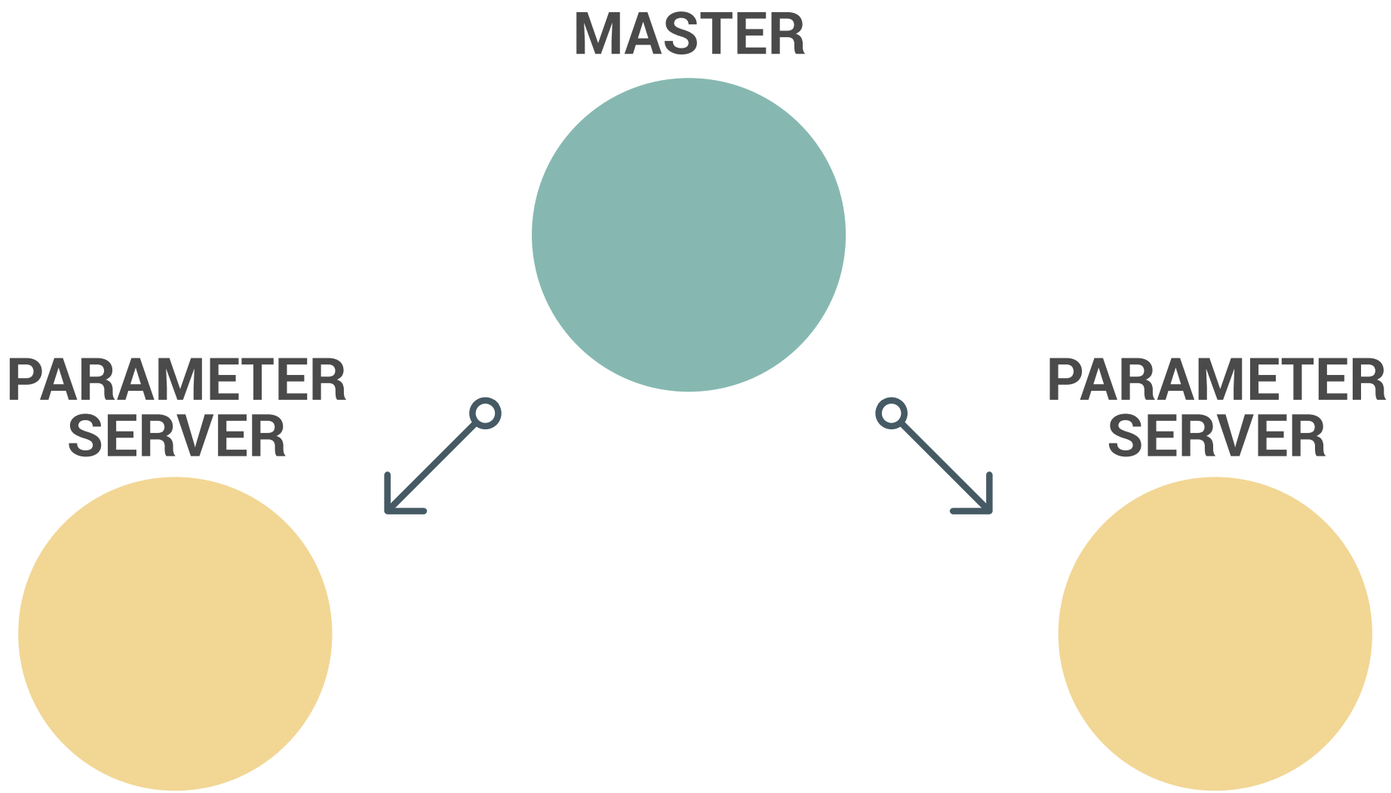

TensorFlow

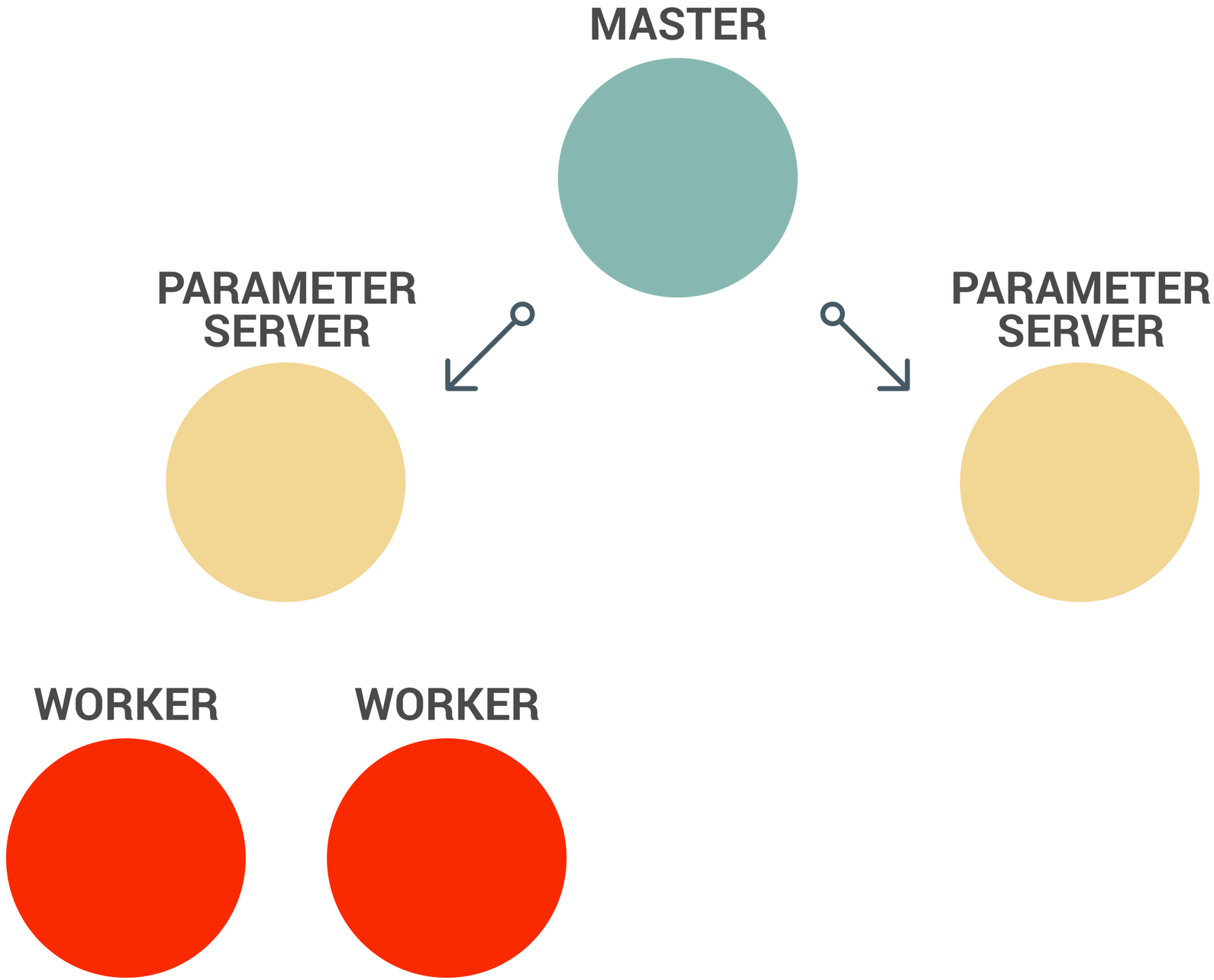
ARCHITECTURE

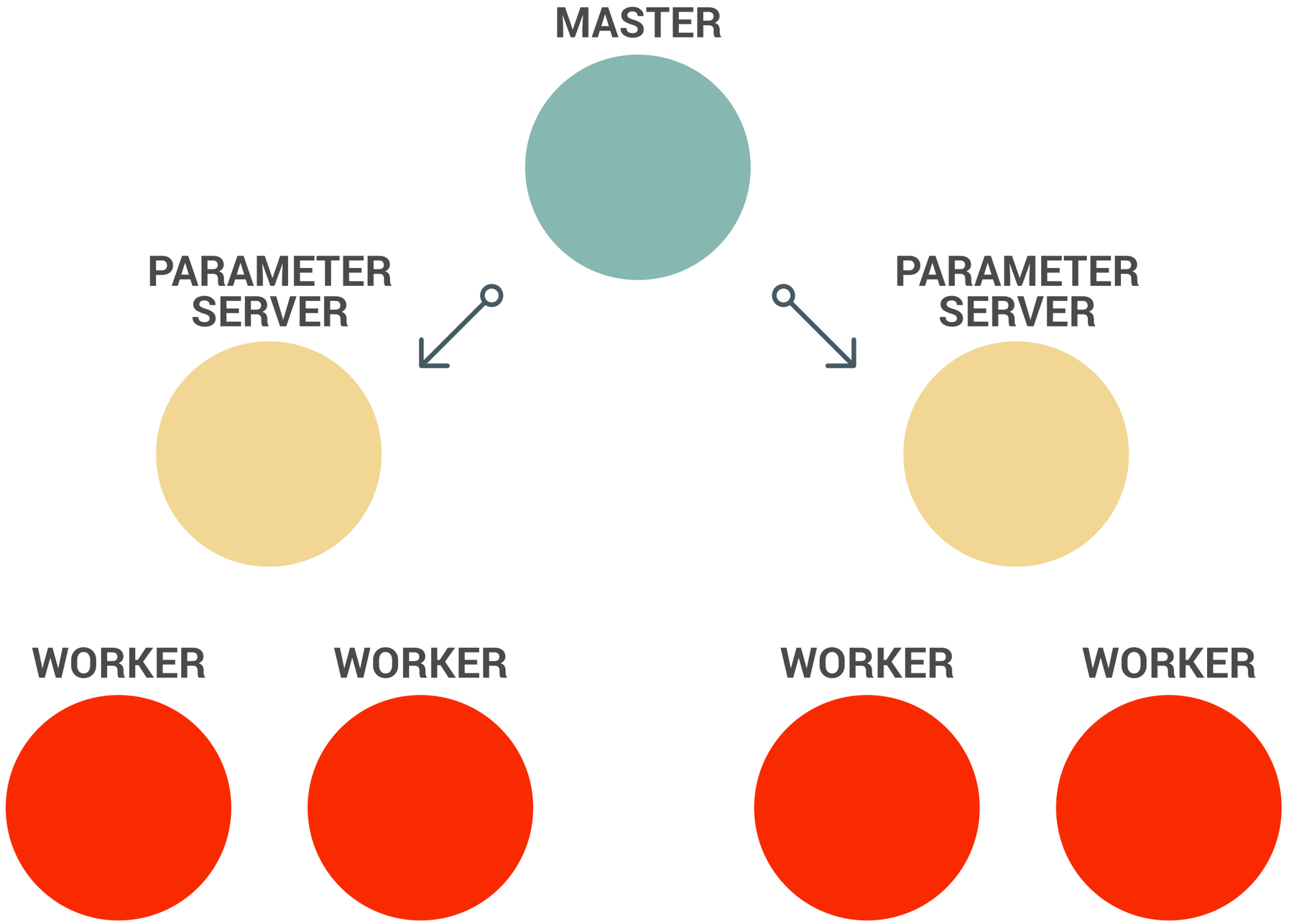
MASTER

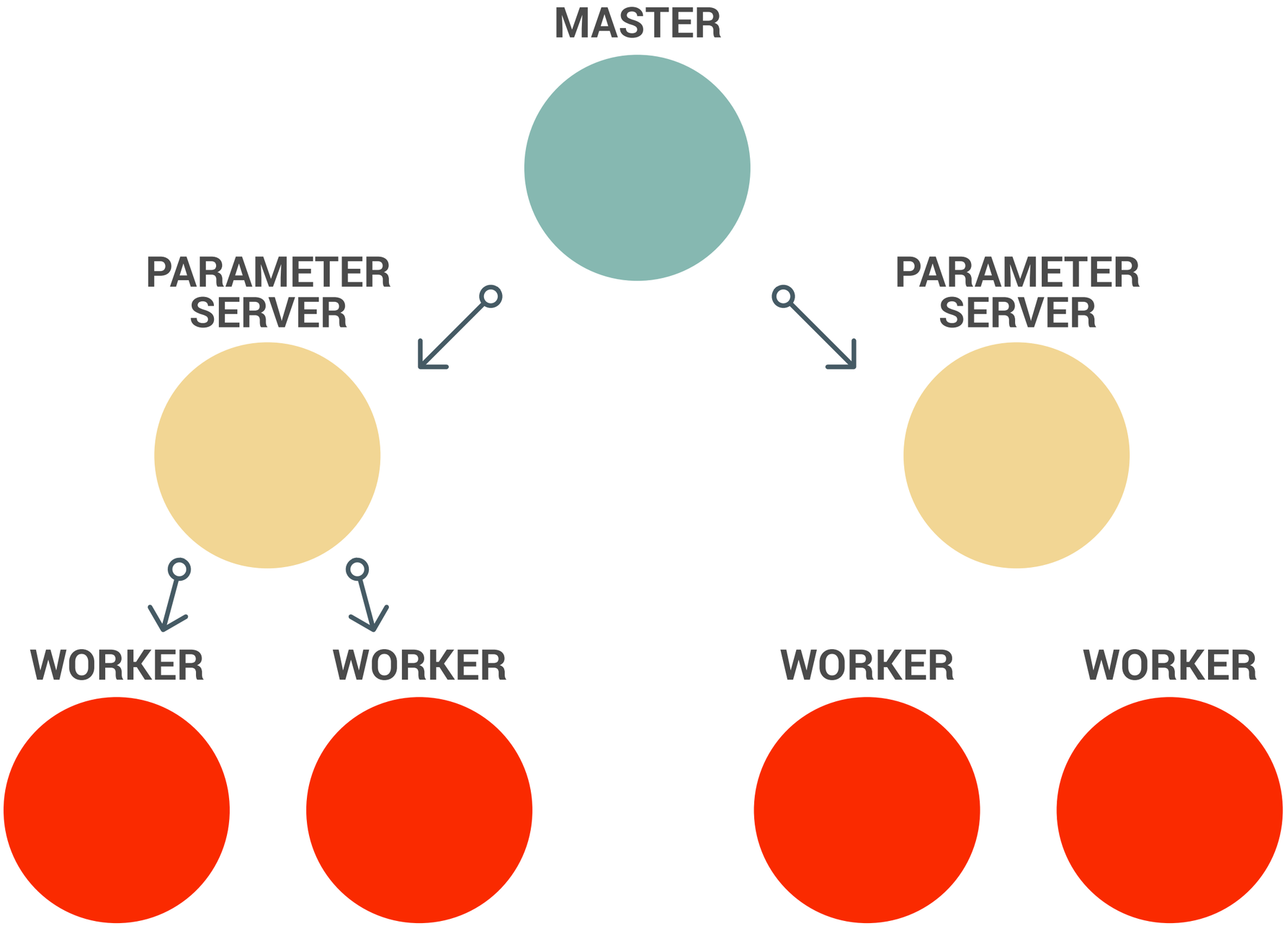


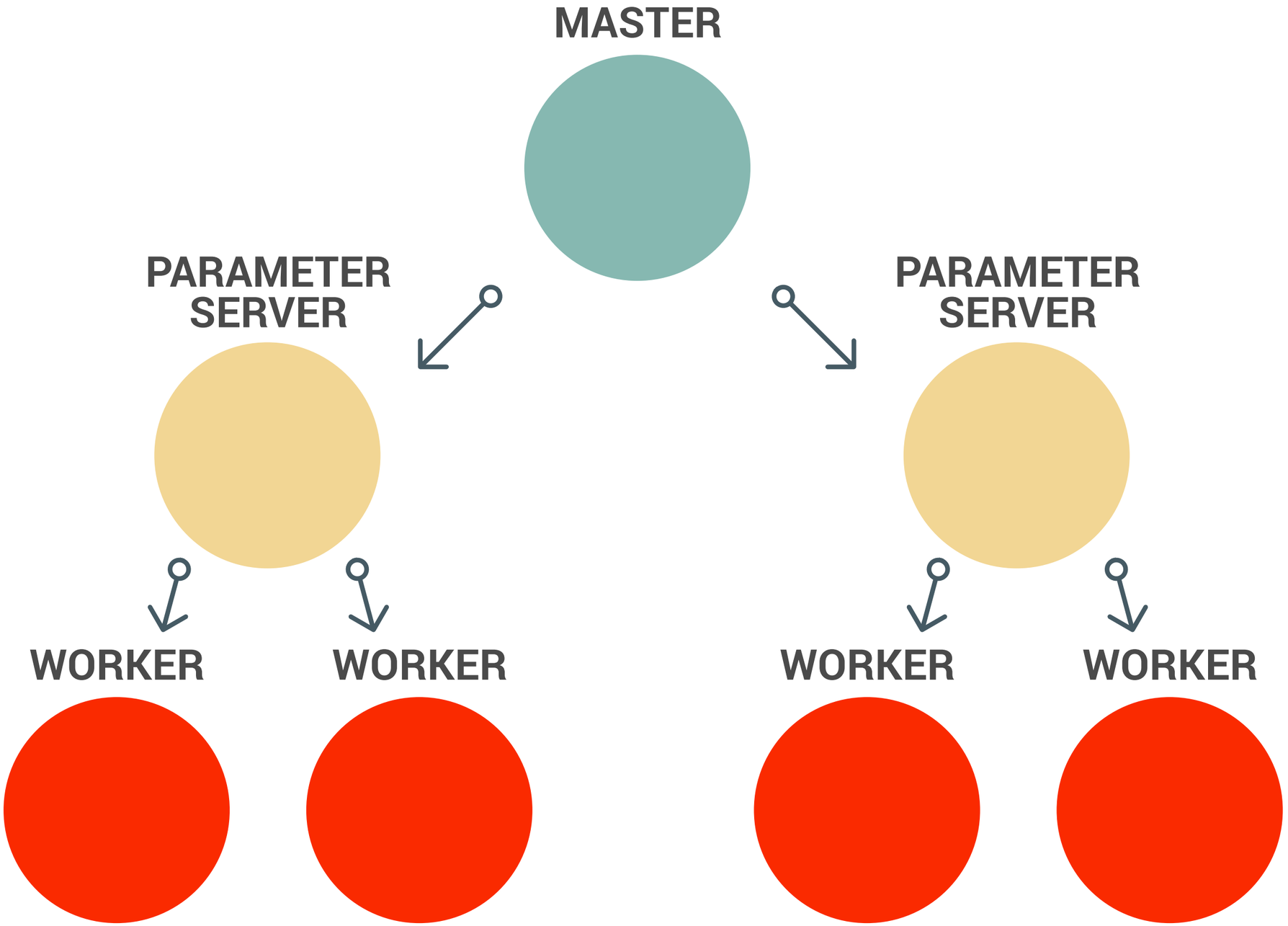

MASTER
PARAMETER SERVER
PARAMETER SERVER


MASTER
PARAMETER SERVER
PARAMETER SERVER

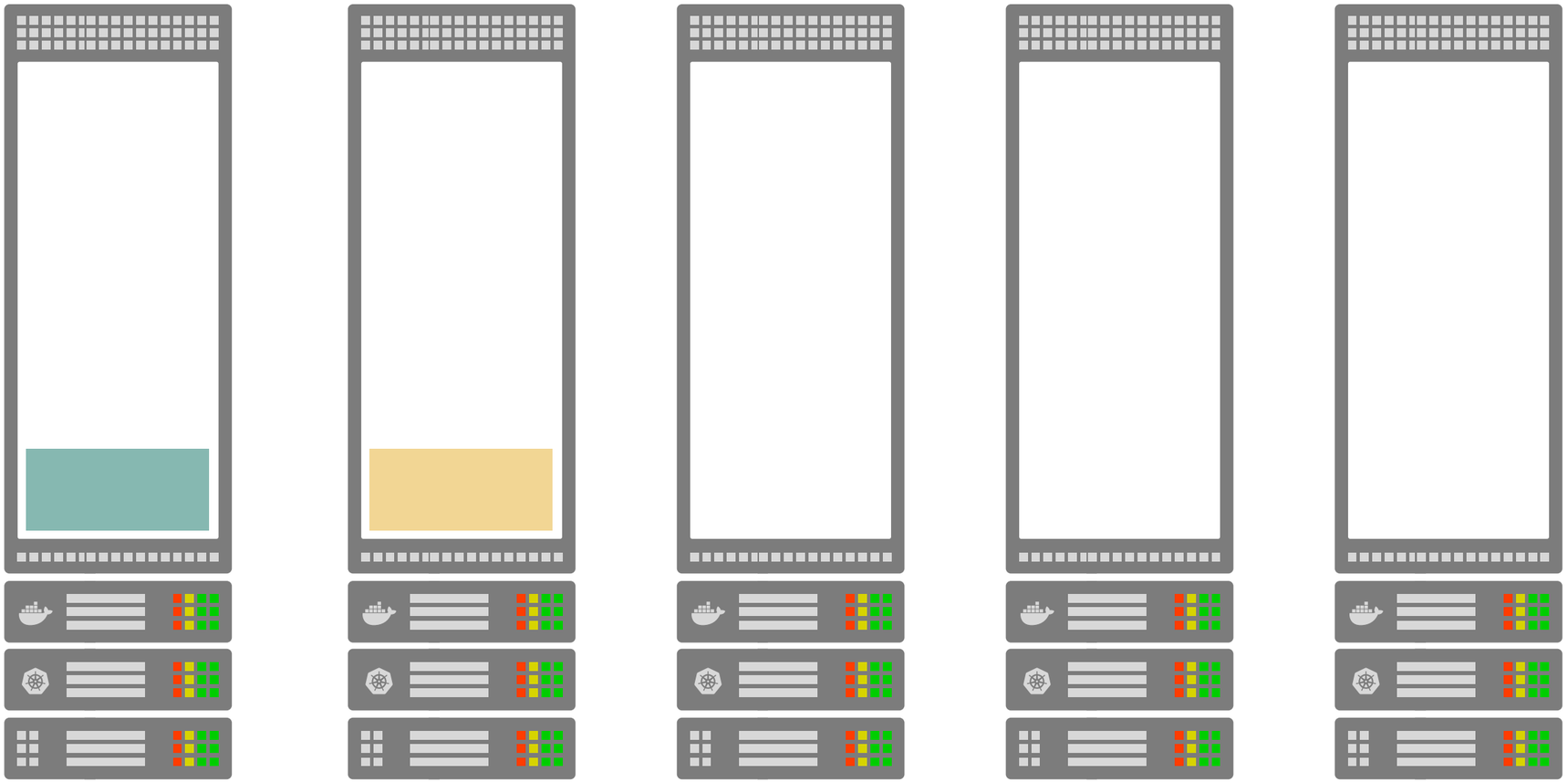

MASTER
PARAMETER SERVER
PARAMETER SERVER
WORKER
WORKER

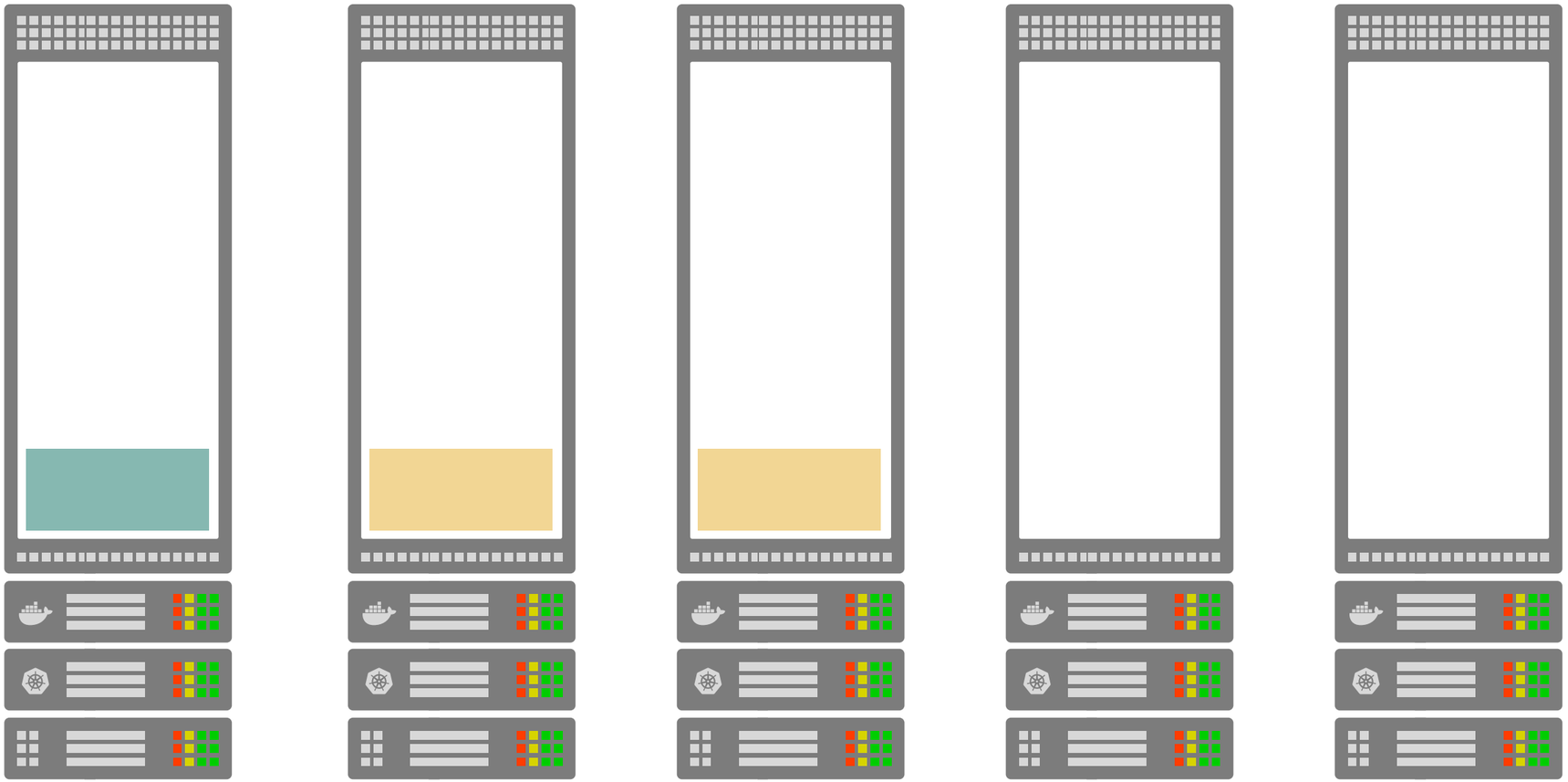

MASTER
PARAMETER SERVER
PARAMETER SERVER
WORKER
WORKER
WORKER
WORKER

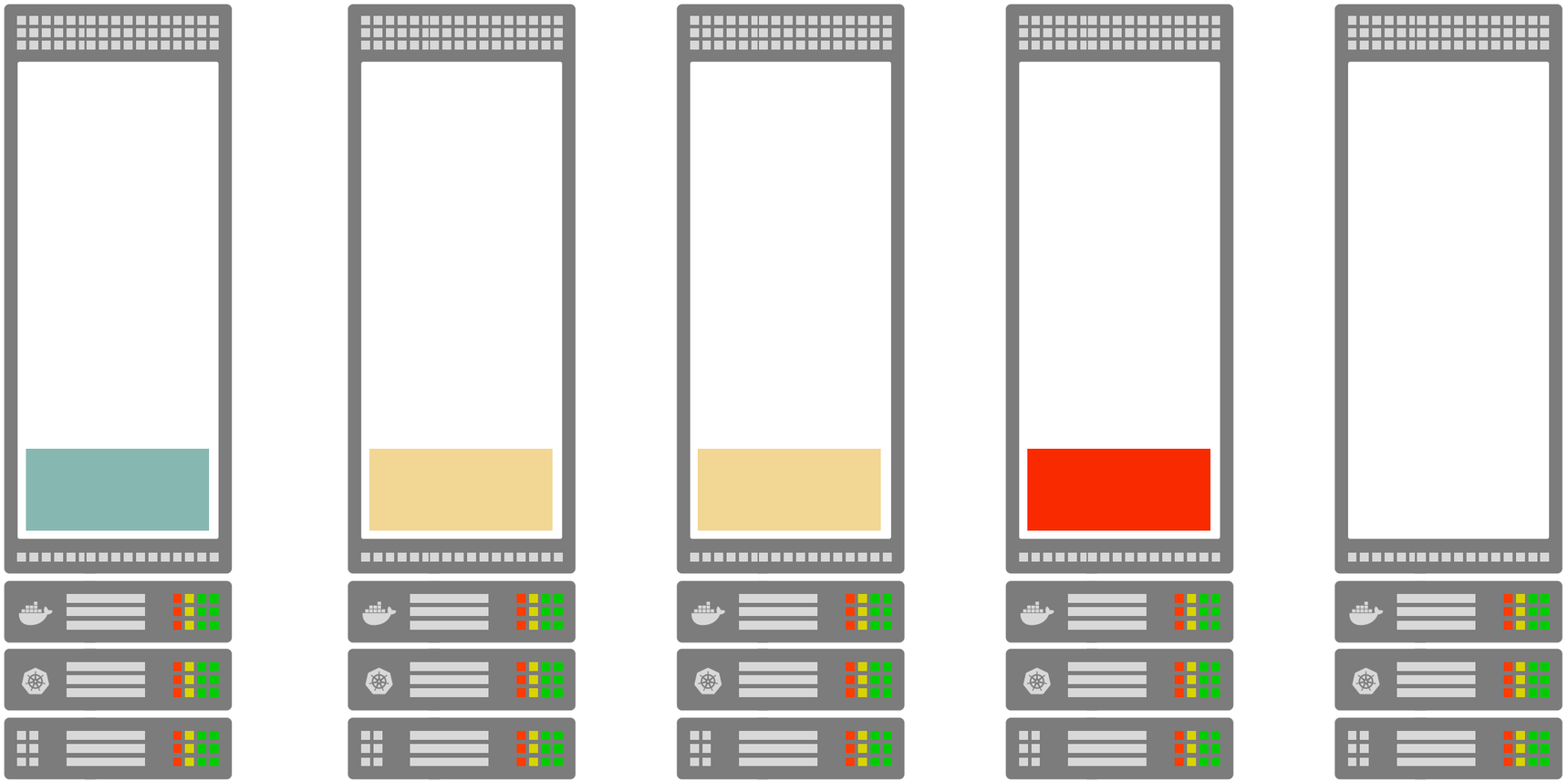

MASTER
PARAMETER SERVER
PARAMETER SERVER
WORKER
WORKER
WORKER
WORKER

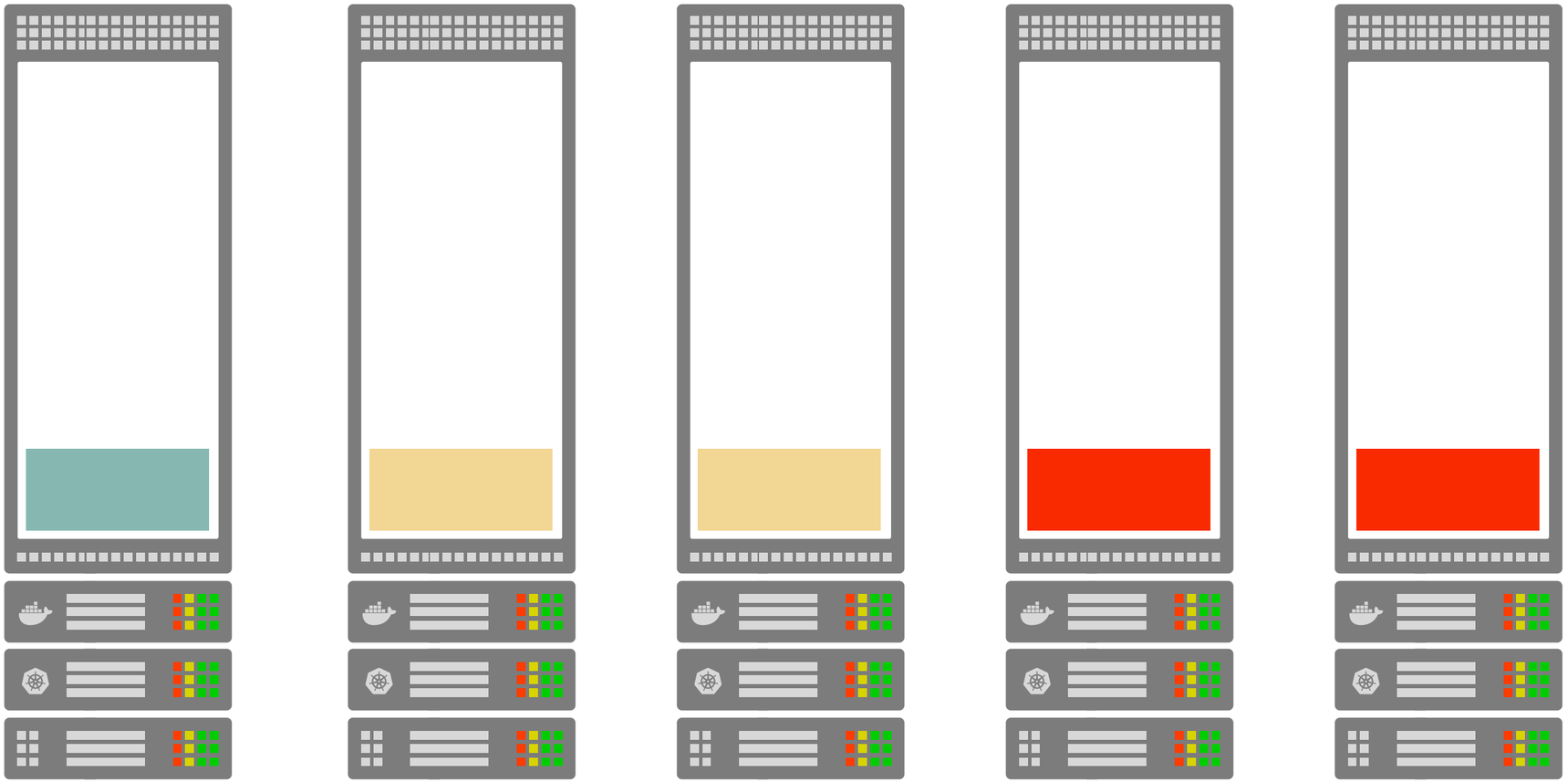

MASTER
PARAMETER SERVER
PARAMETER SERVER
WORKER
WORKER
WORKER
WORKER

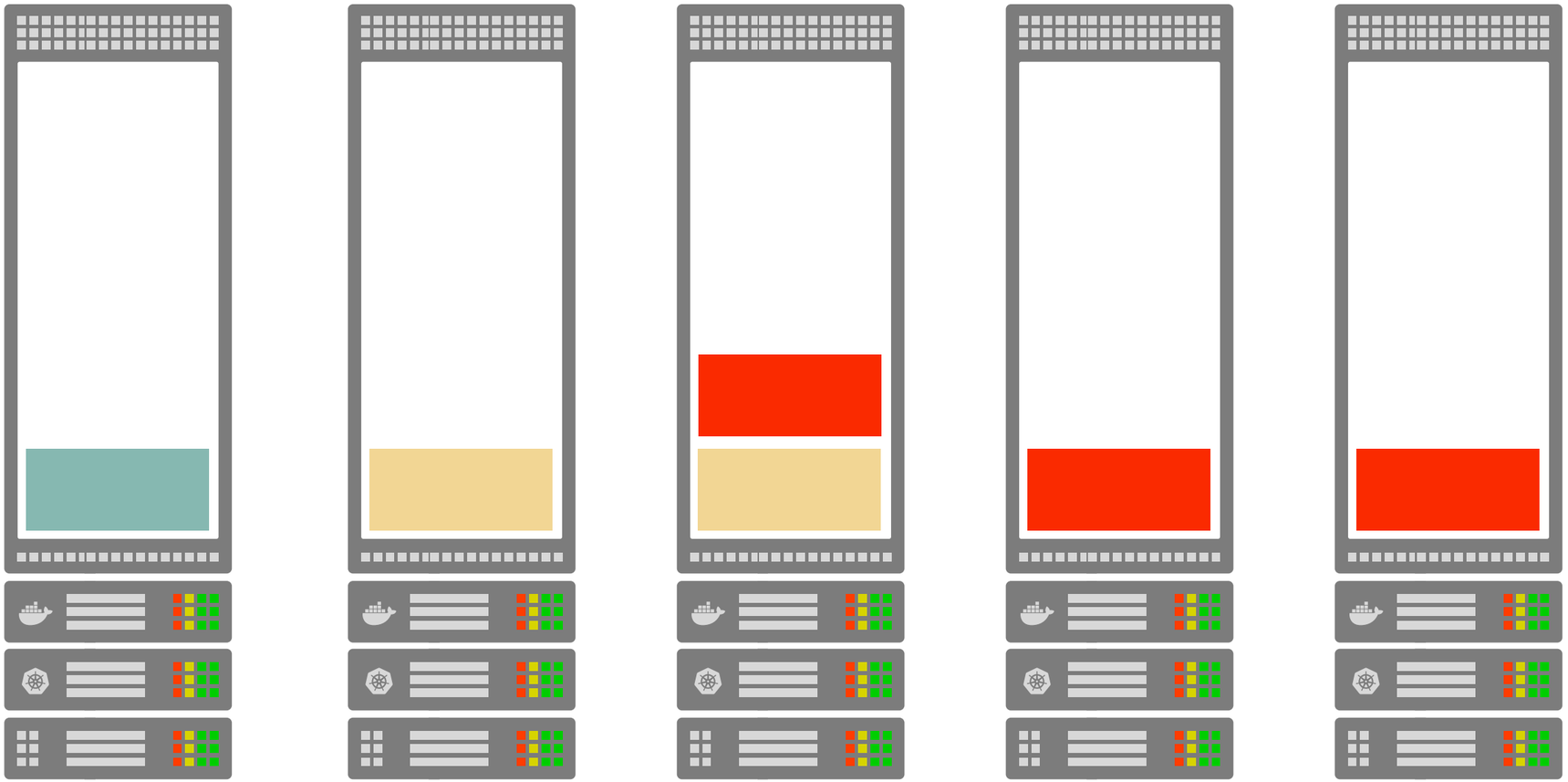
MEANWHILE 

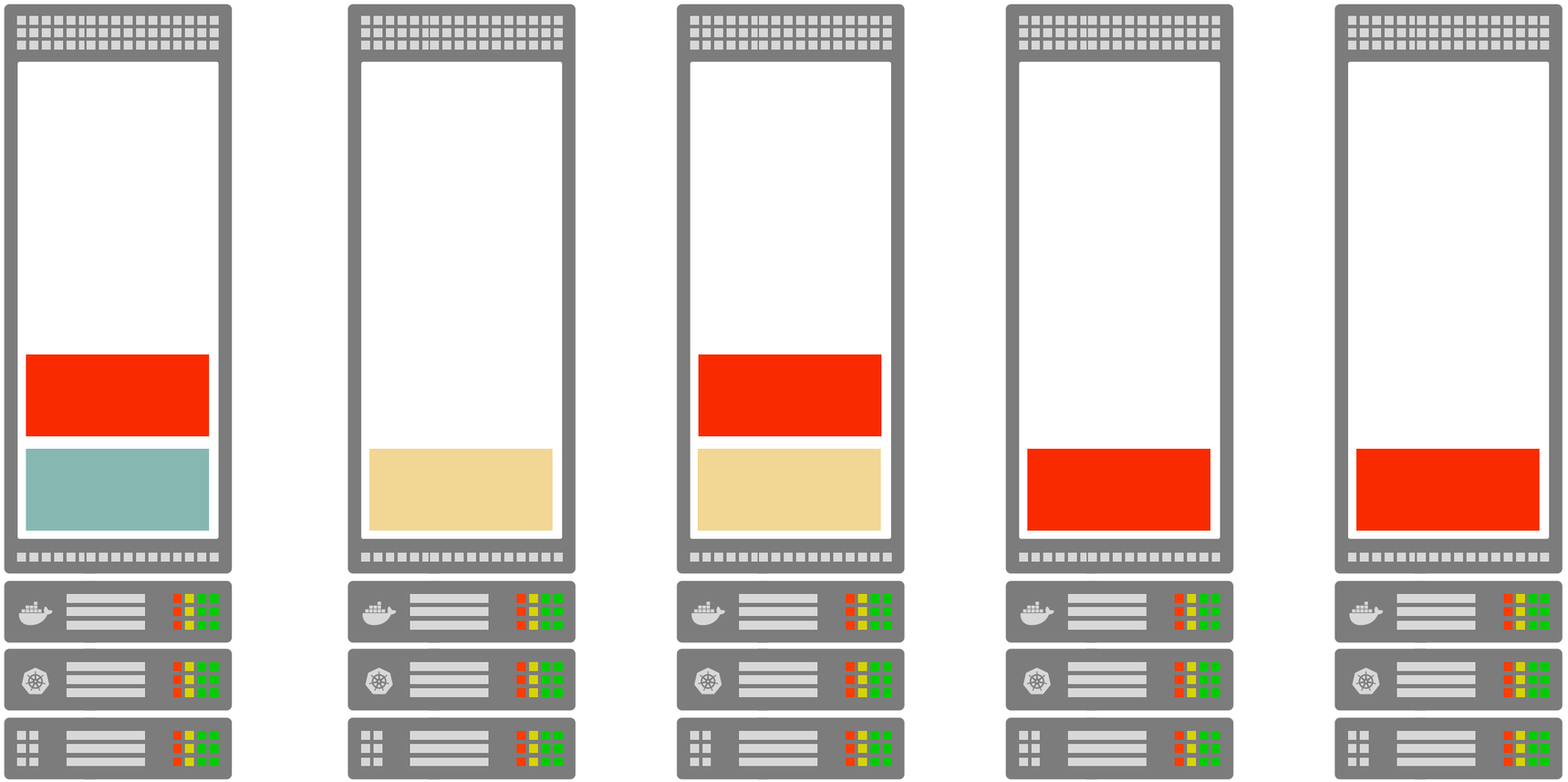










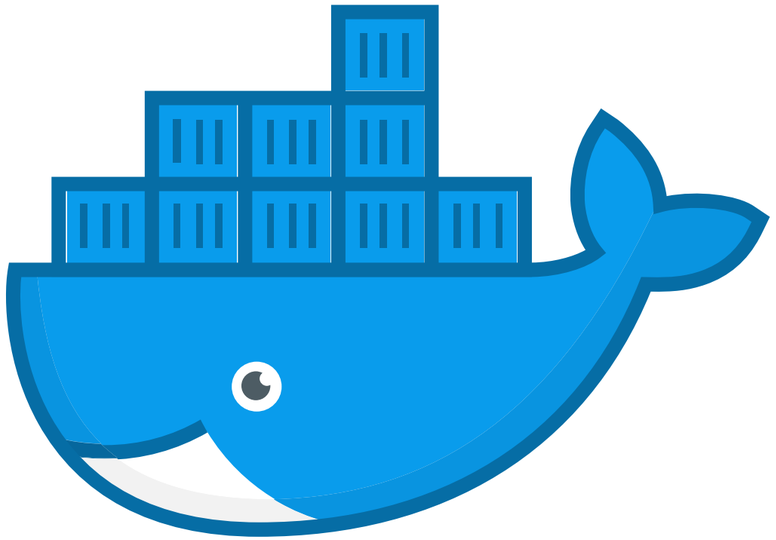


DOES K8S RUN  PYTHON  ?

NOPE!

CONTAINER ORCHESTRATOR

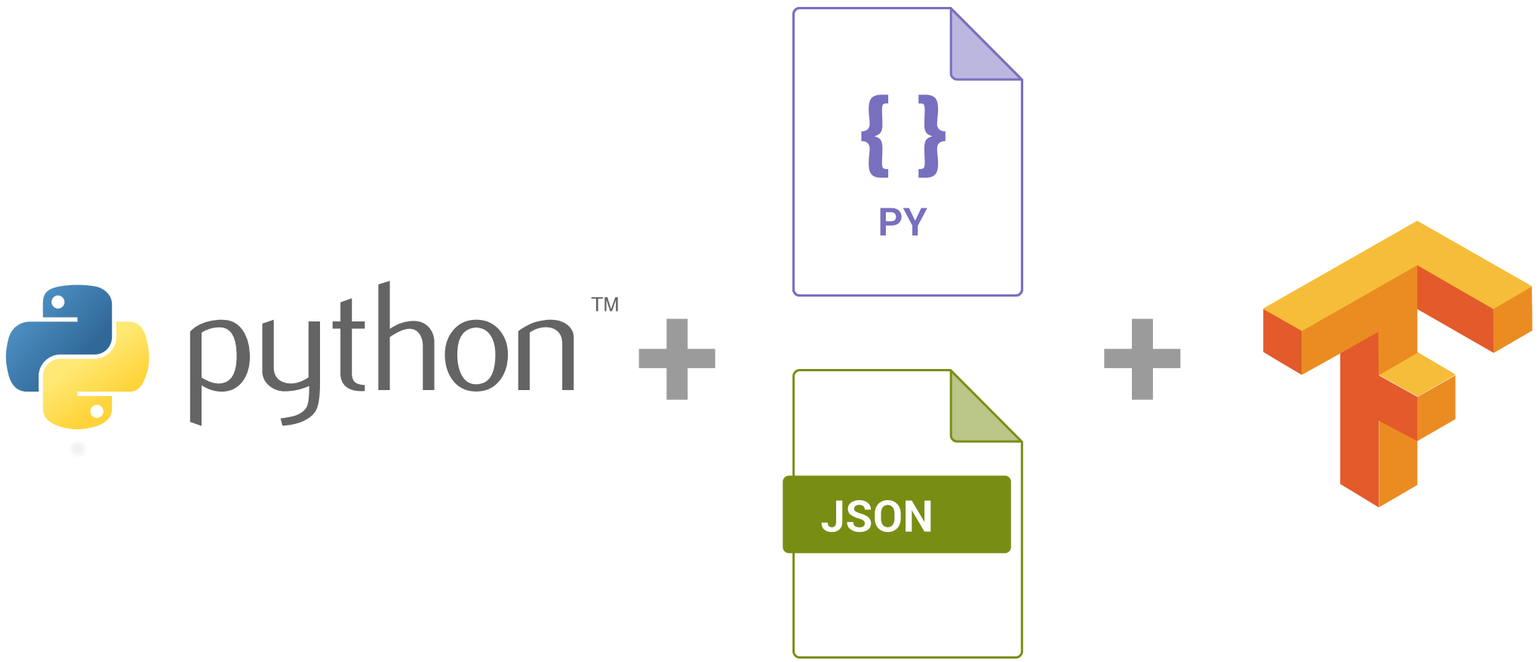
CONTAINER



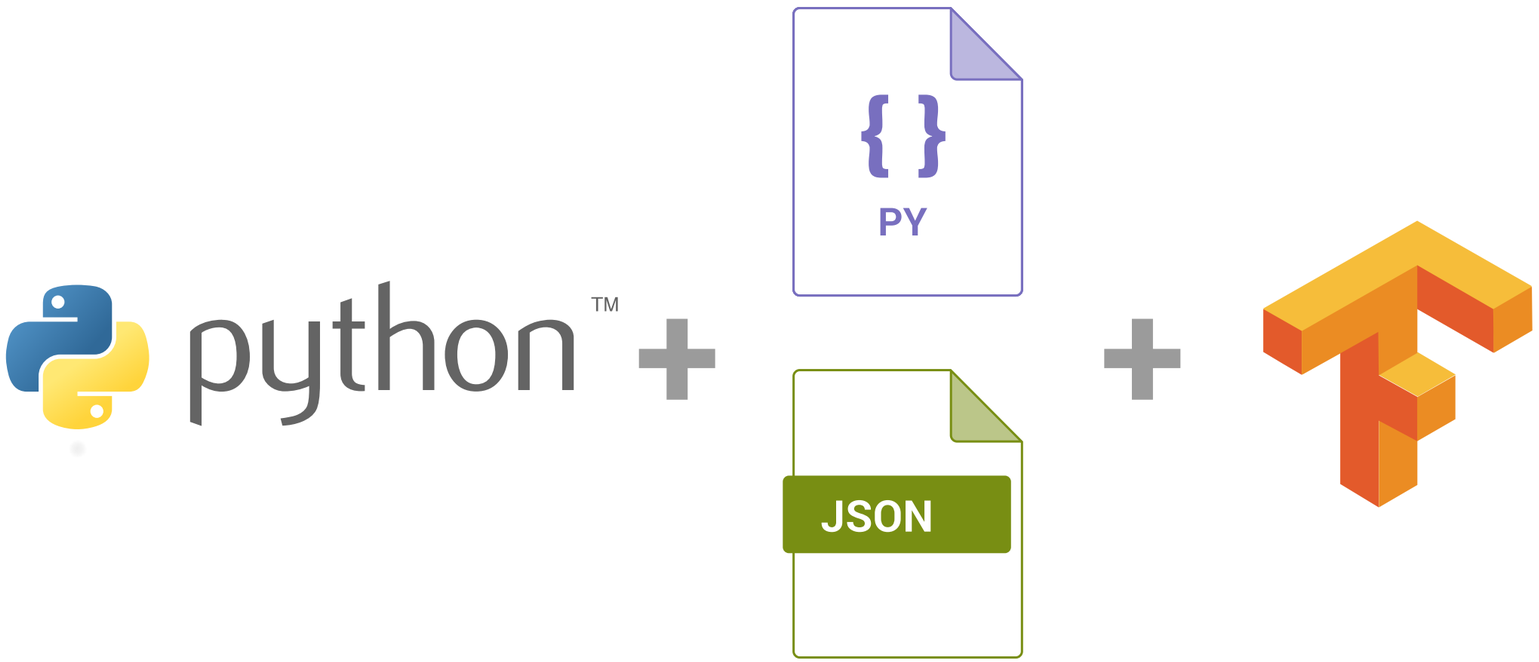
docker®


python™

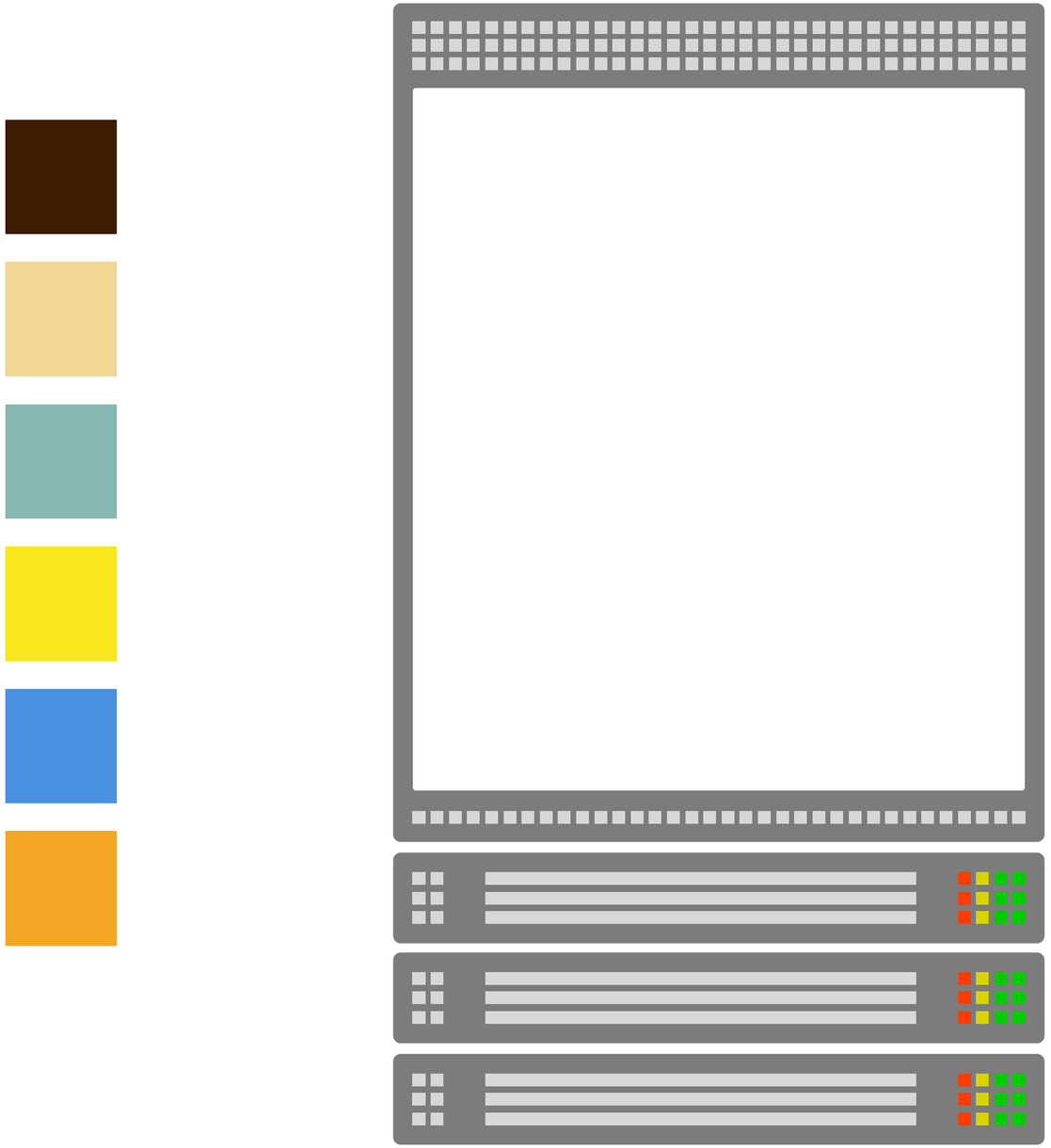

python™
+
{ }
PY
JSON

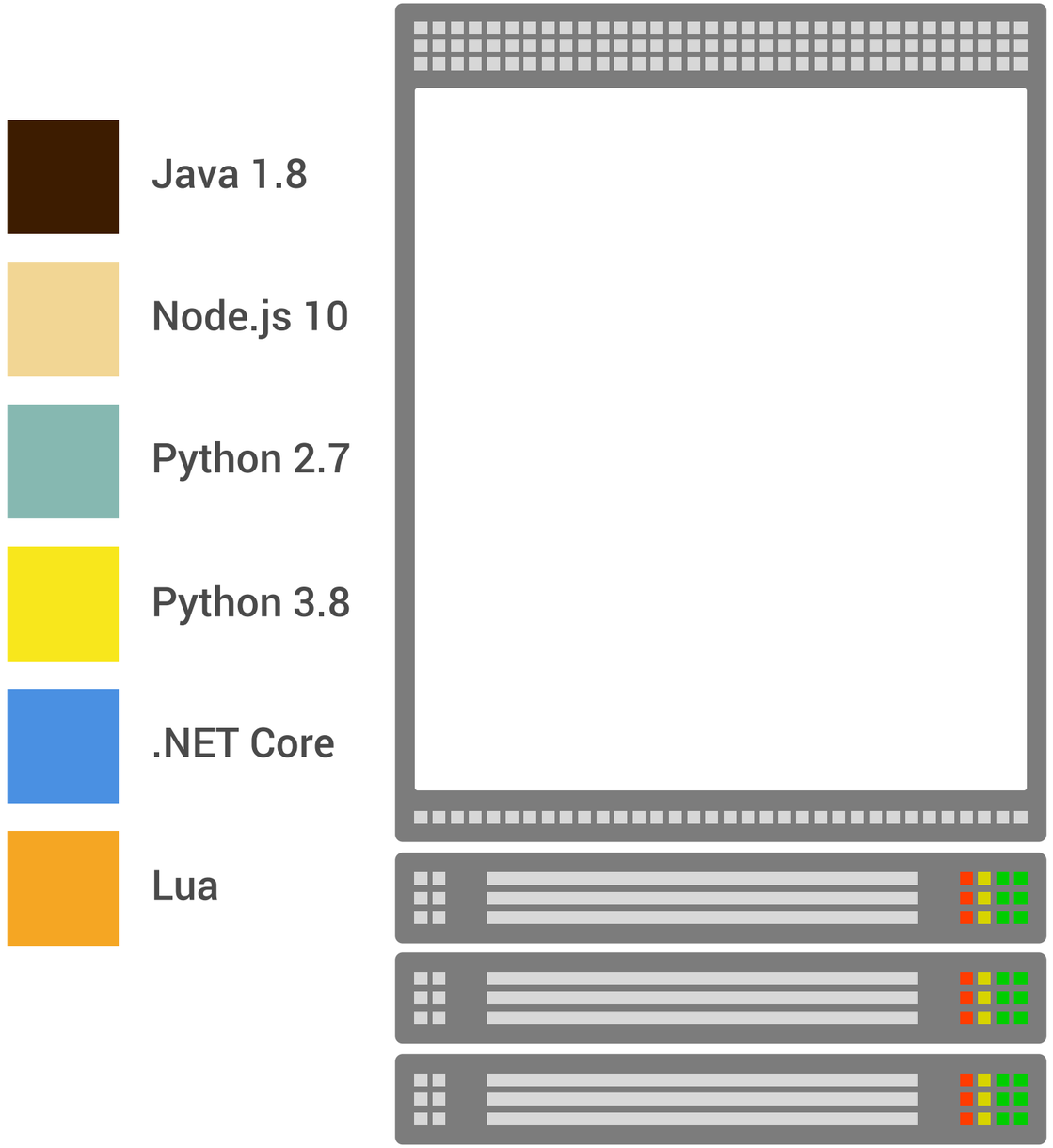

python™
{ }
PY
JSON

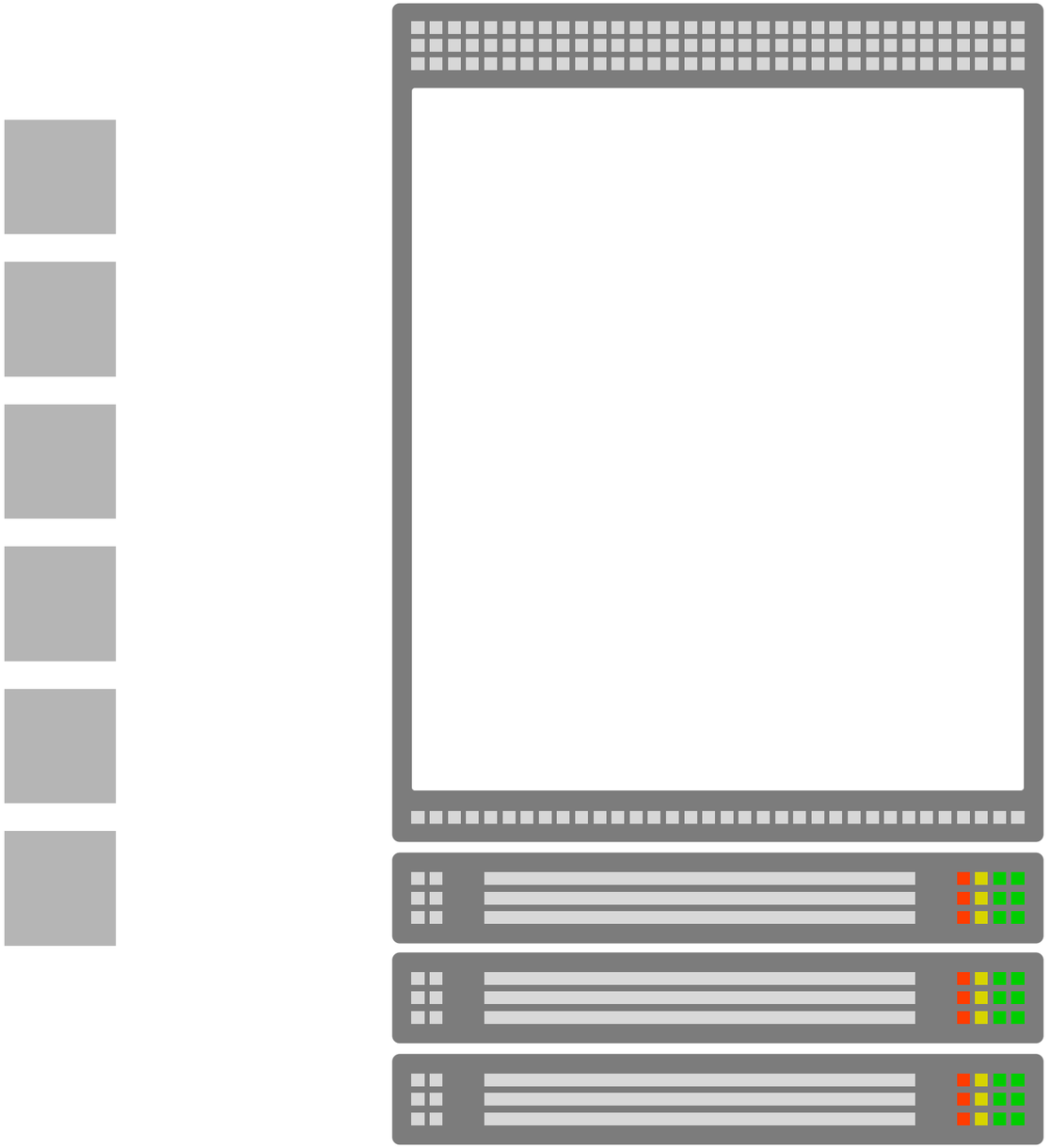
CONTAINER

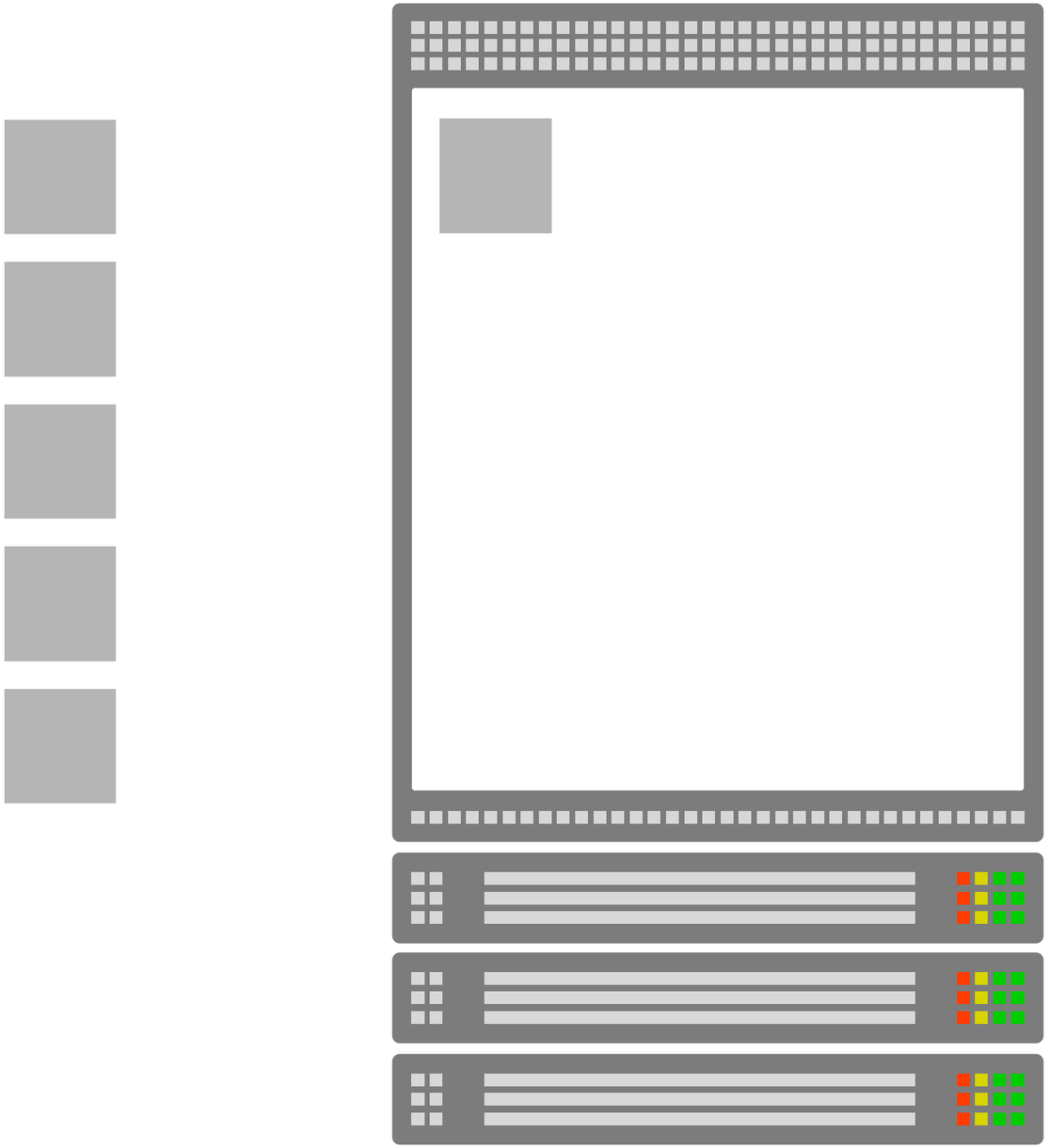


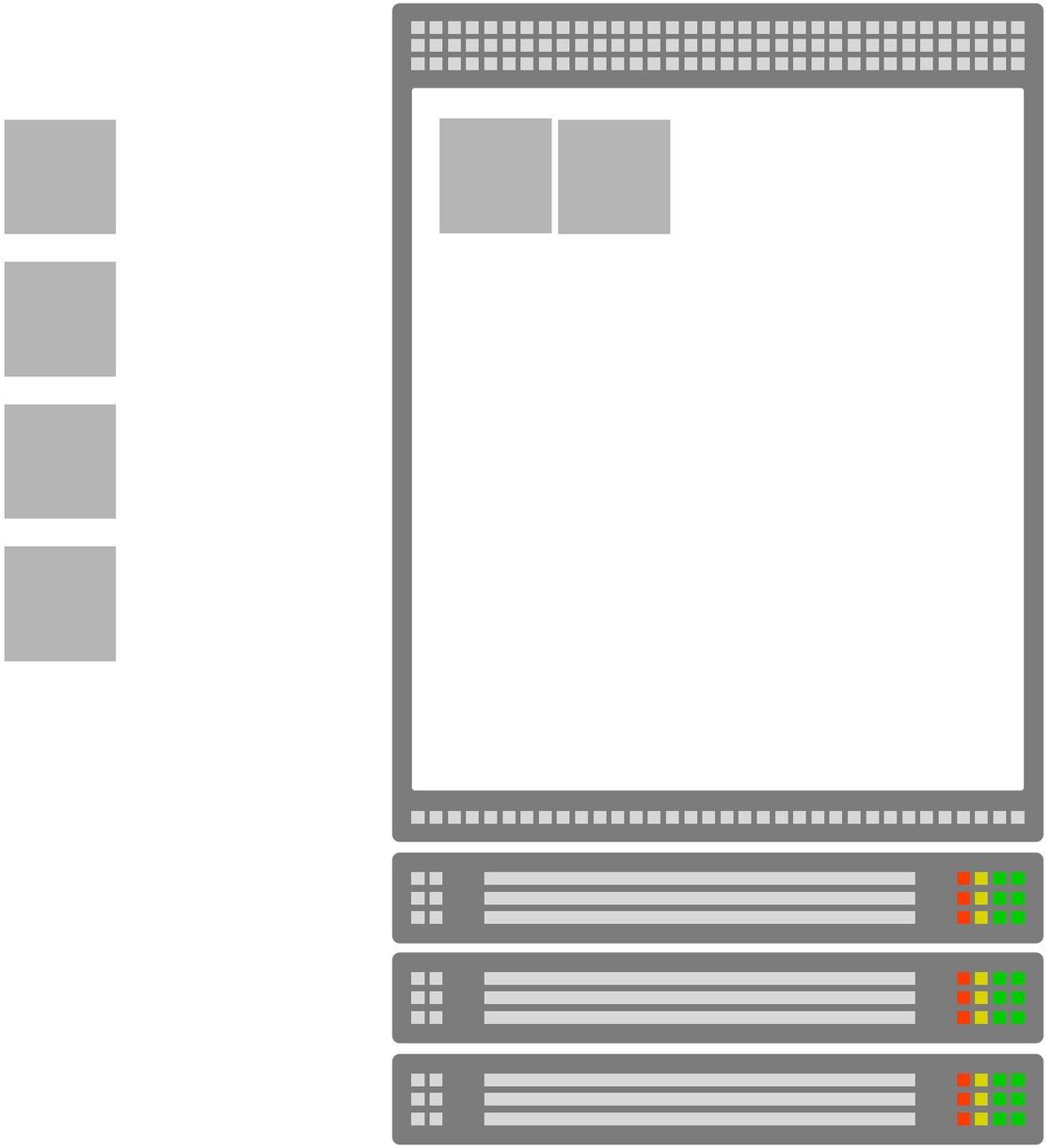
LANGUAGE AGNOSTIC

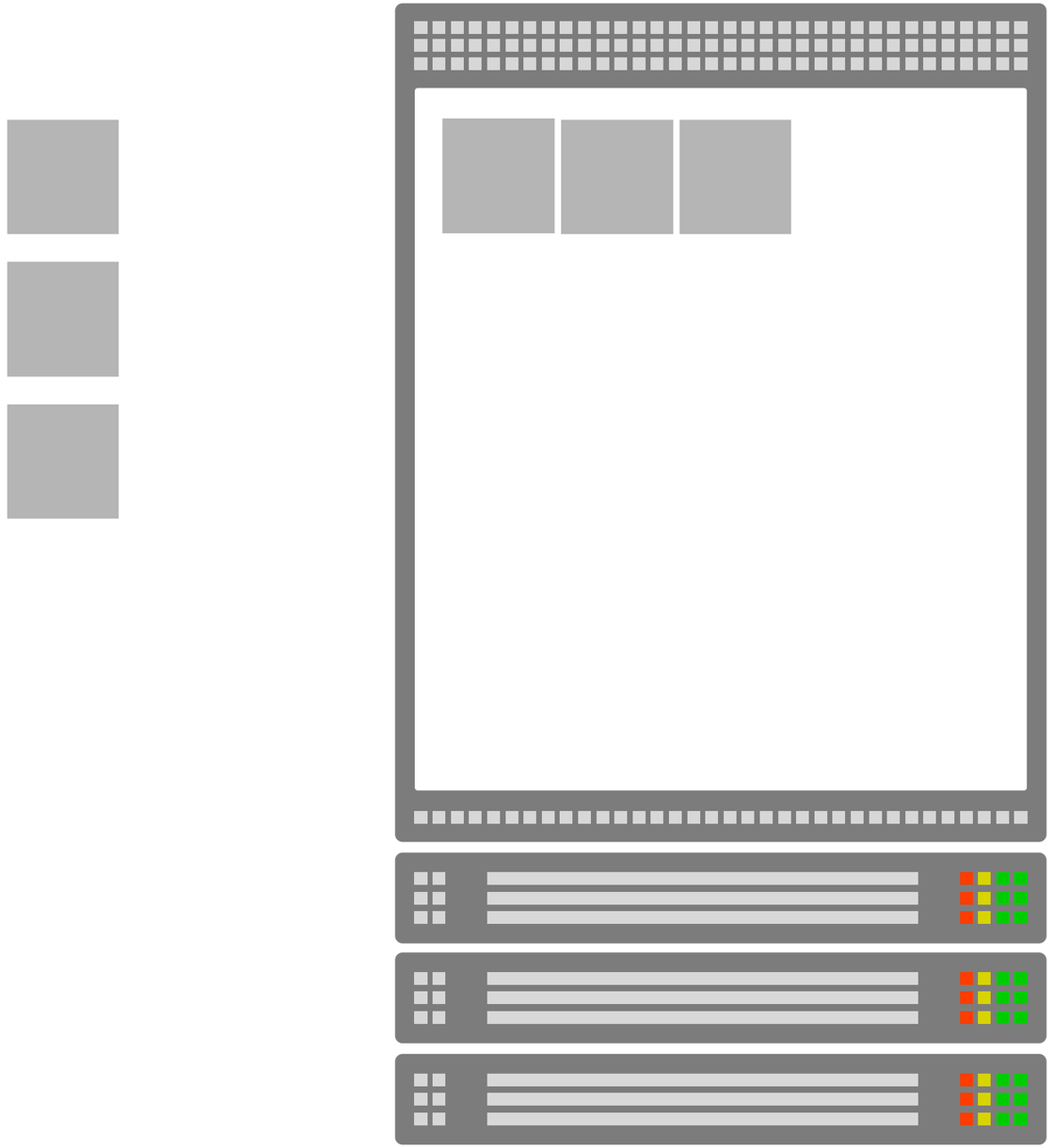


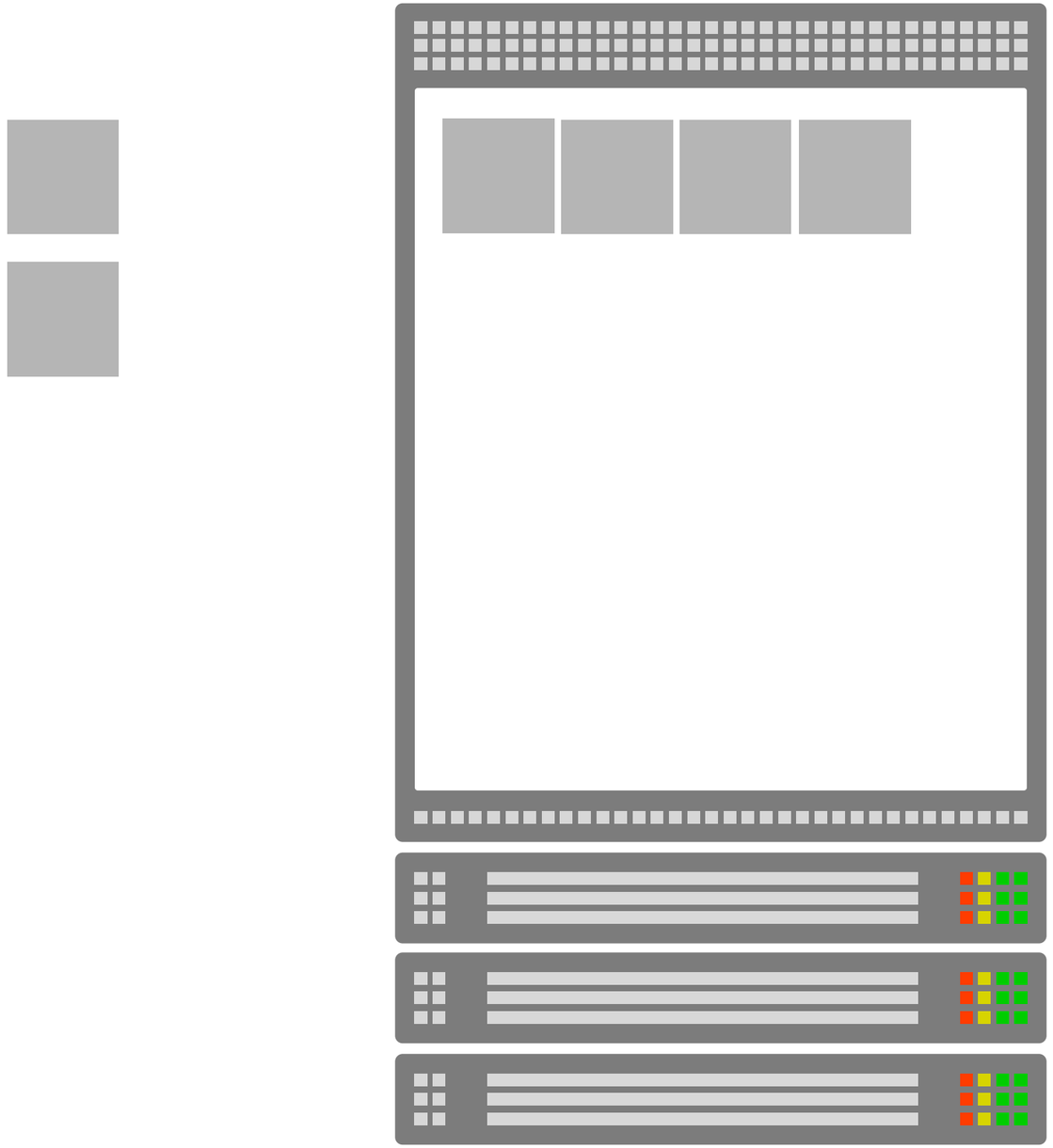

Java 1.8
Node.js 10
Python 2.7
Python 3.8
.NET Core
Lua

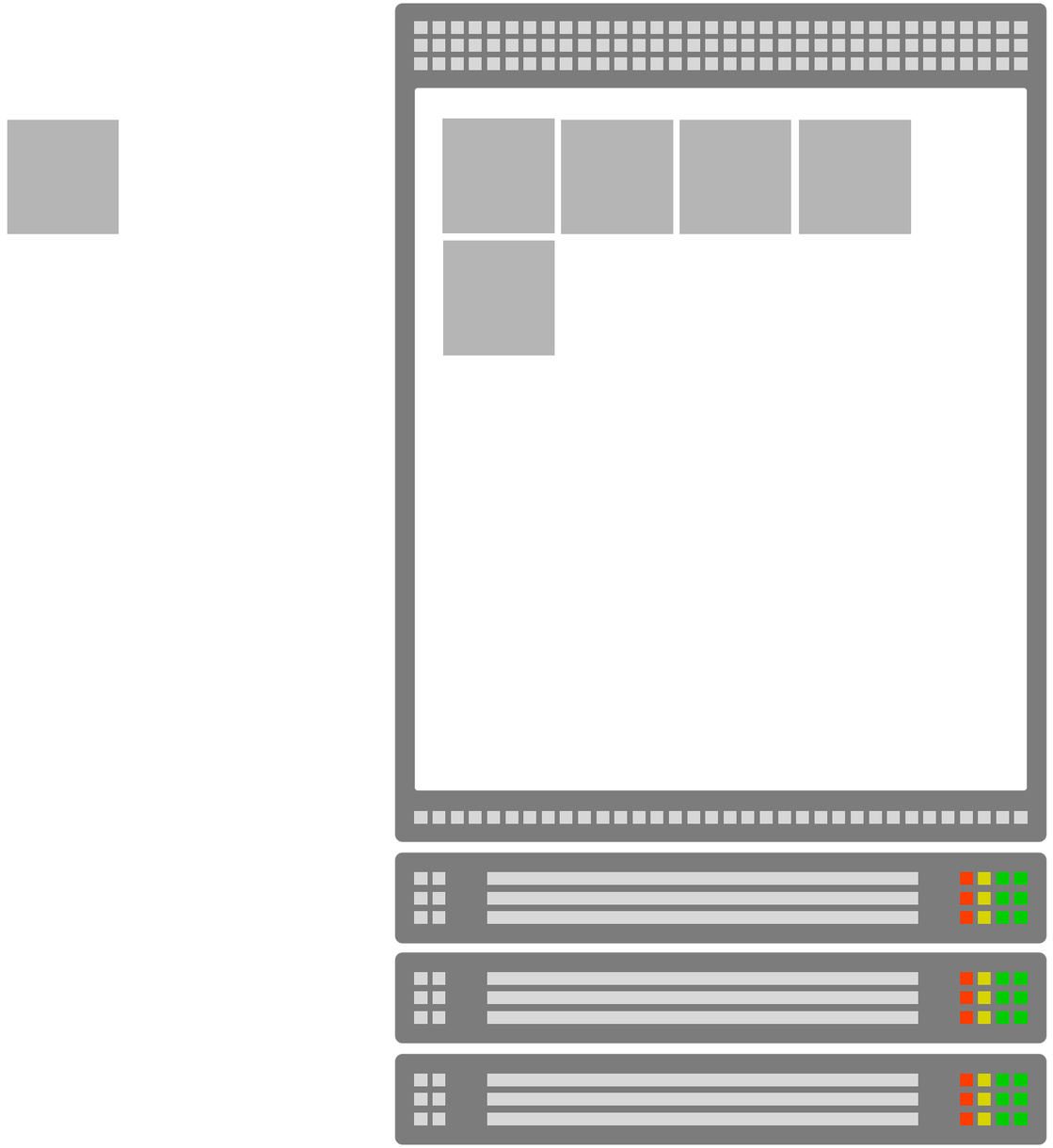


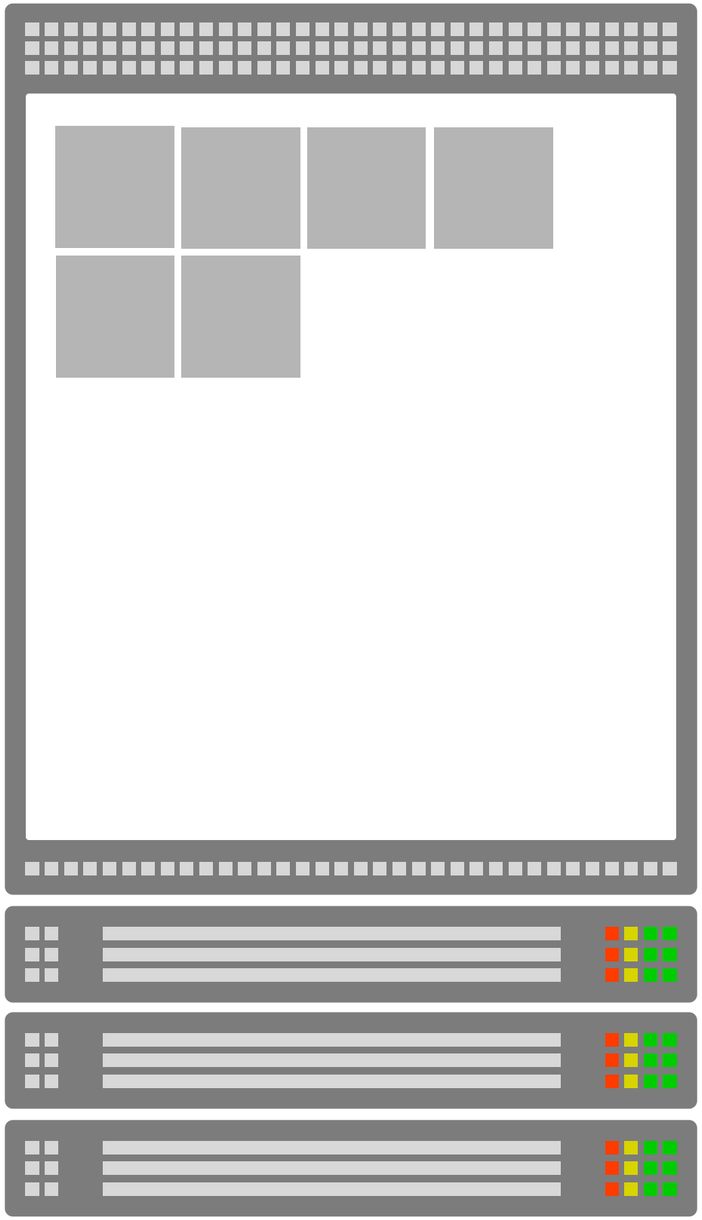












Demo...

MiniKF

by Aarikto

Chicago Taxi

Chicago Taxi

100M Trips

Chicago Taxi

100M Trips

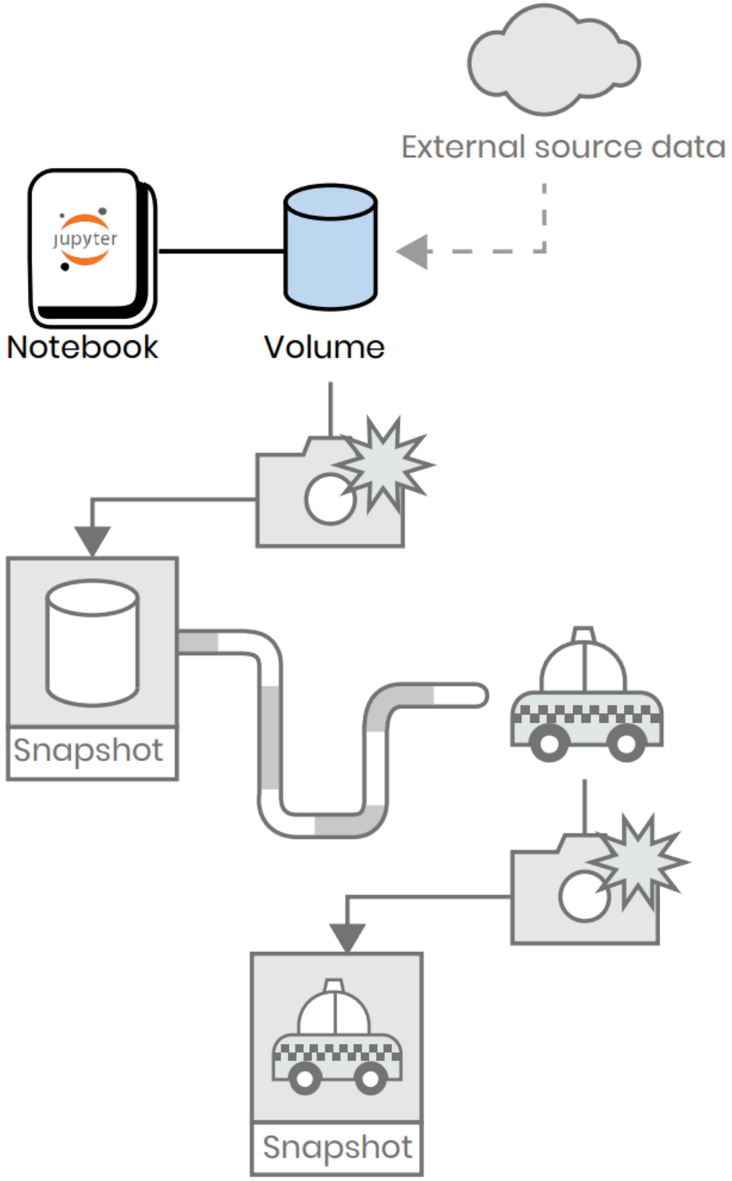
fare, tripstartmonth, tripstarthour

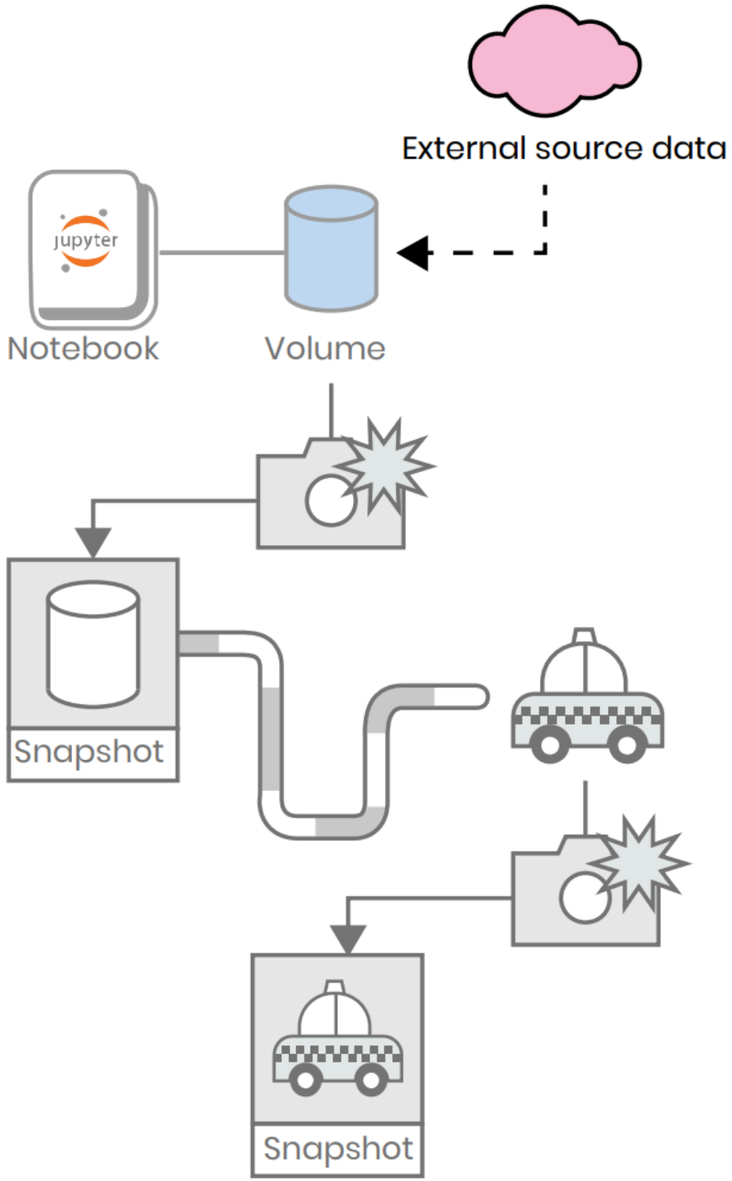
Chicago Taxi

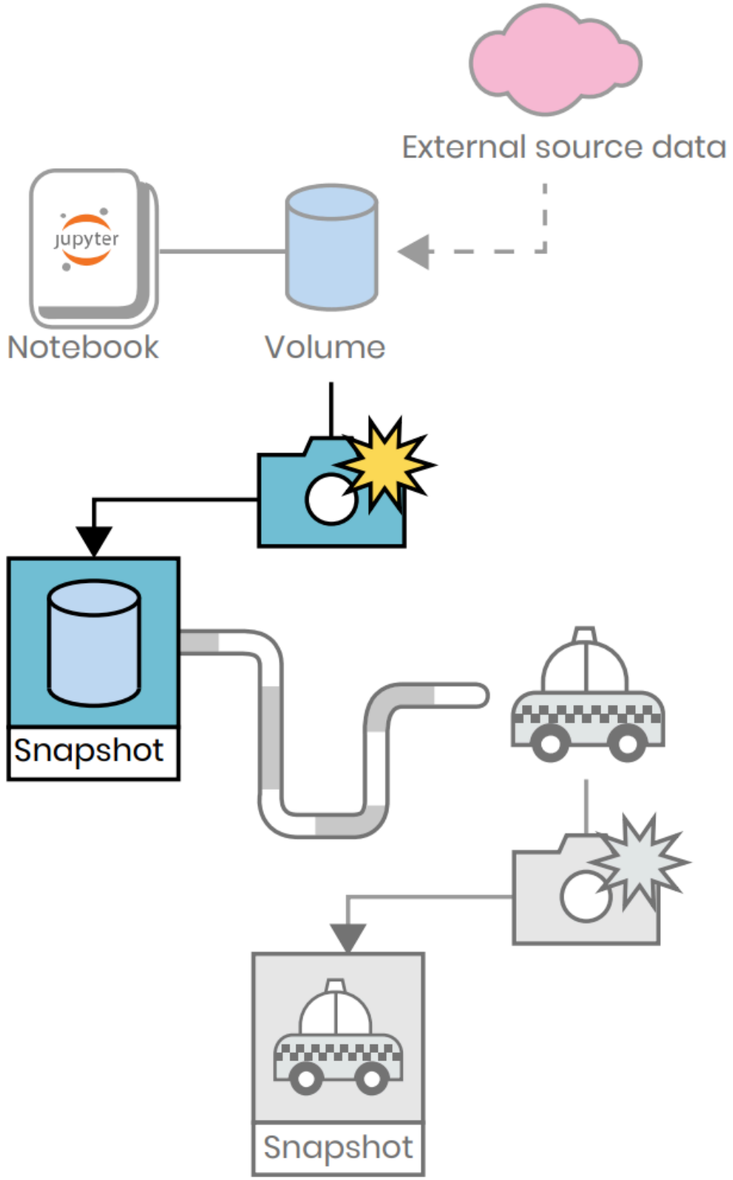
100M Trips

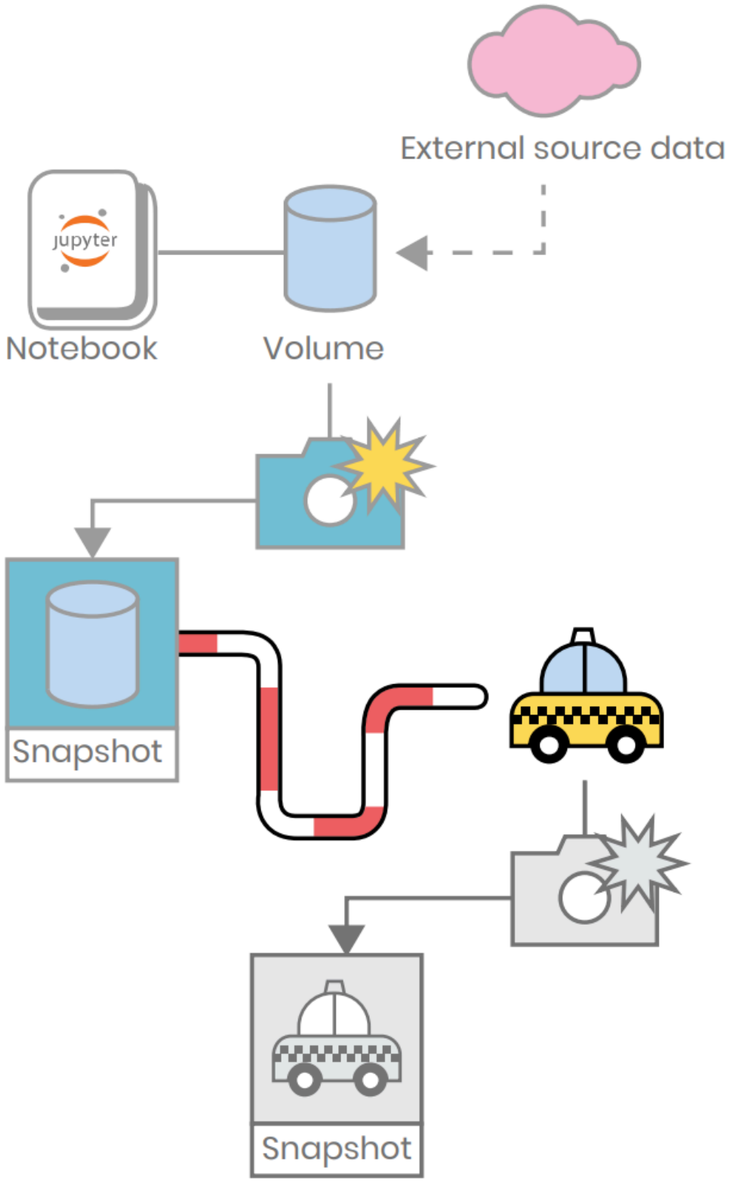
fare, tripstartmonth, tripstarthour

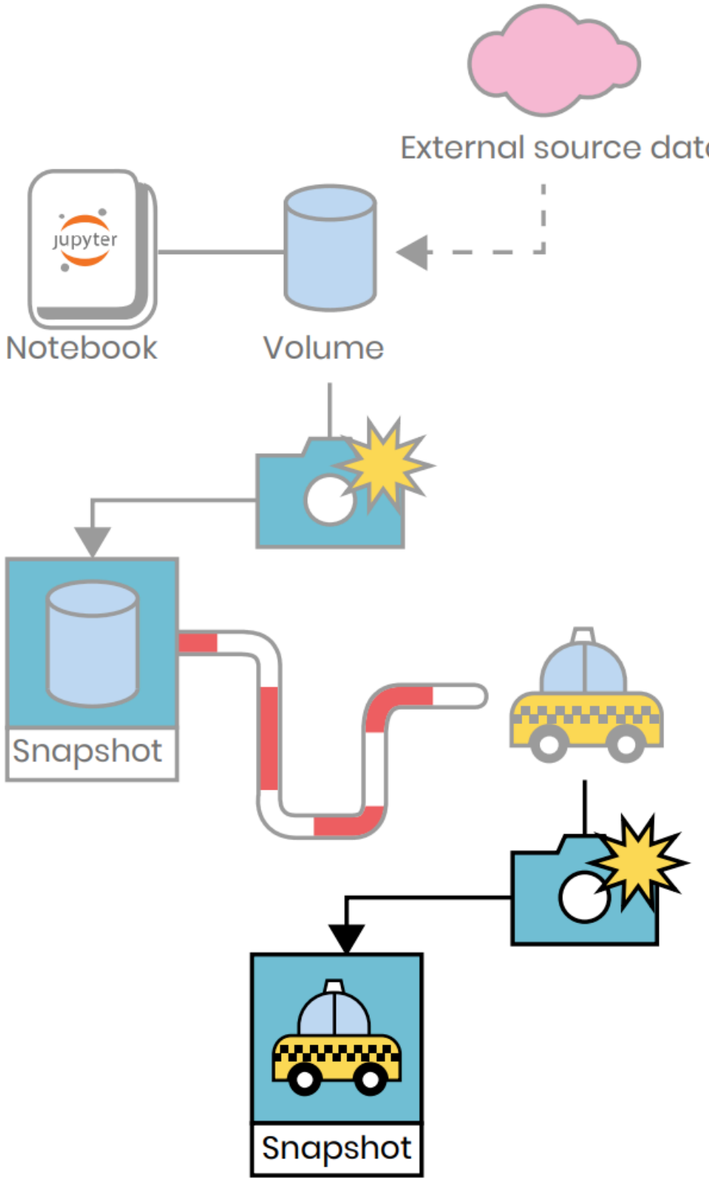
Out: tips > 20%


External source data
jupyter
Notebook
Volume
Snapshot
Snapshot


External source data
jupyter
Notebook
Volume
Snapshot
Snapshot


External source data
jupyter
Notebook
Volume
Snapshot
Snapshot


External source data
jupyter
Notebook
Volume
Snapshot
Snapshot


External source data
jupyter
Notebook
Volume
Snapshot
Snapshot

Pipeline

1. DATA VALIDATION

1. DATA VALIDATION

2. DATA PREPROCESSING

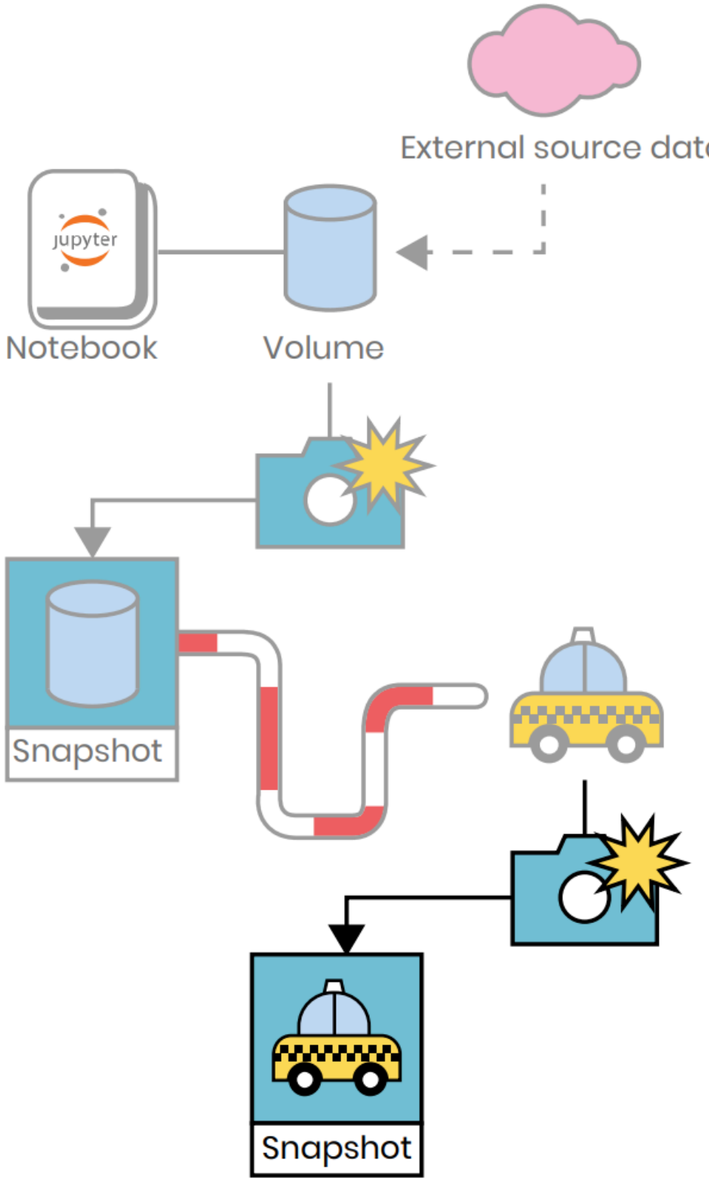
1. DATA VALIDATION

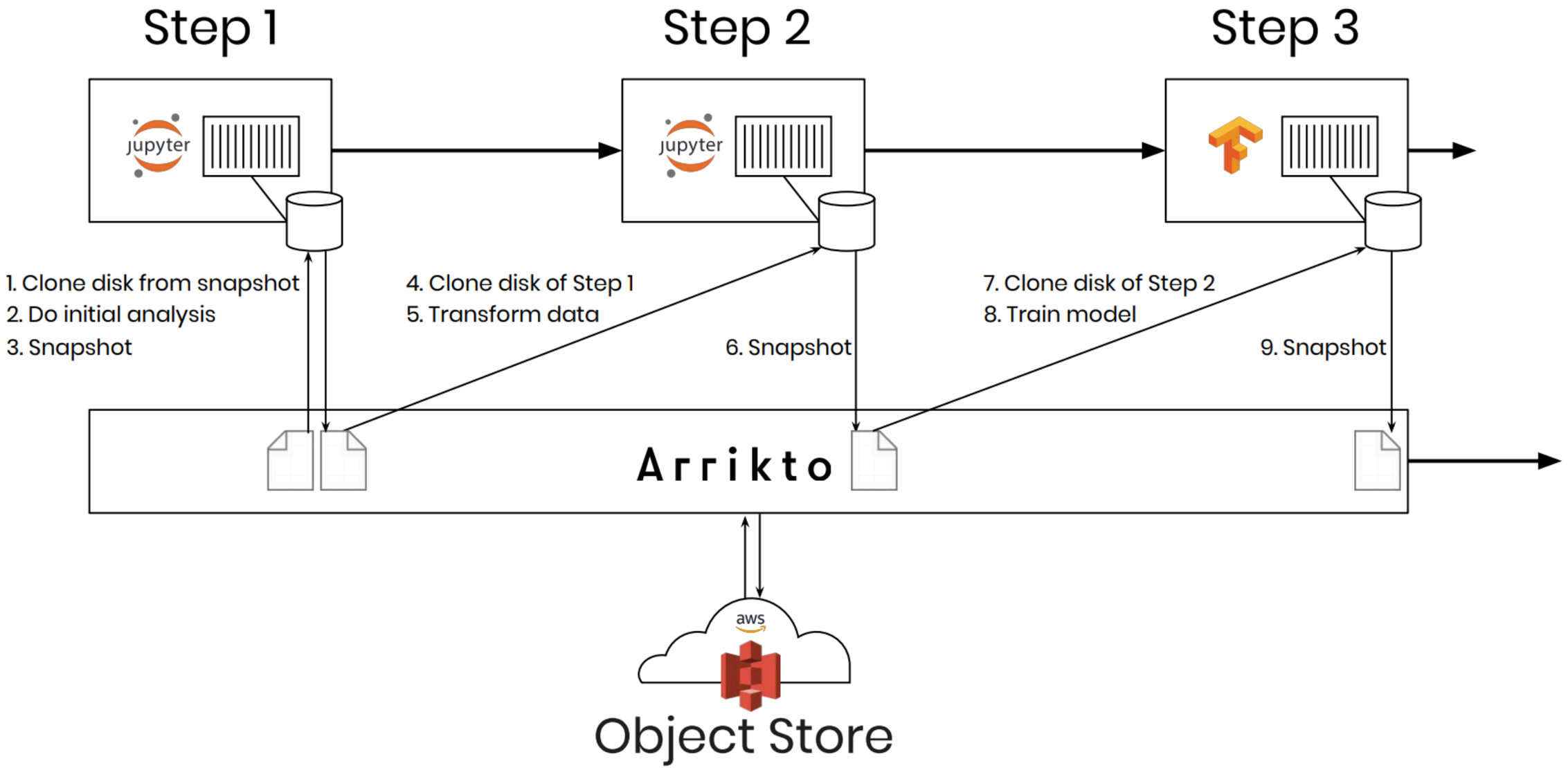
2. DATA PREPROCESSING

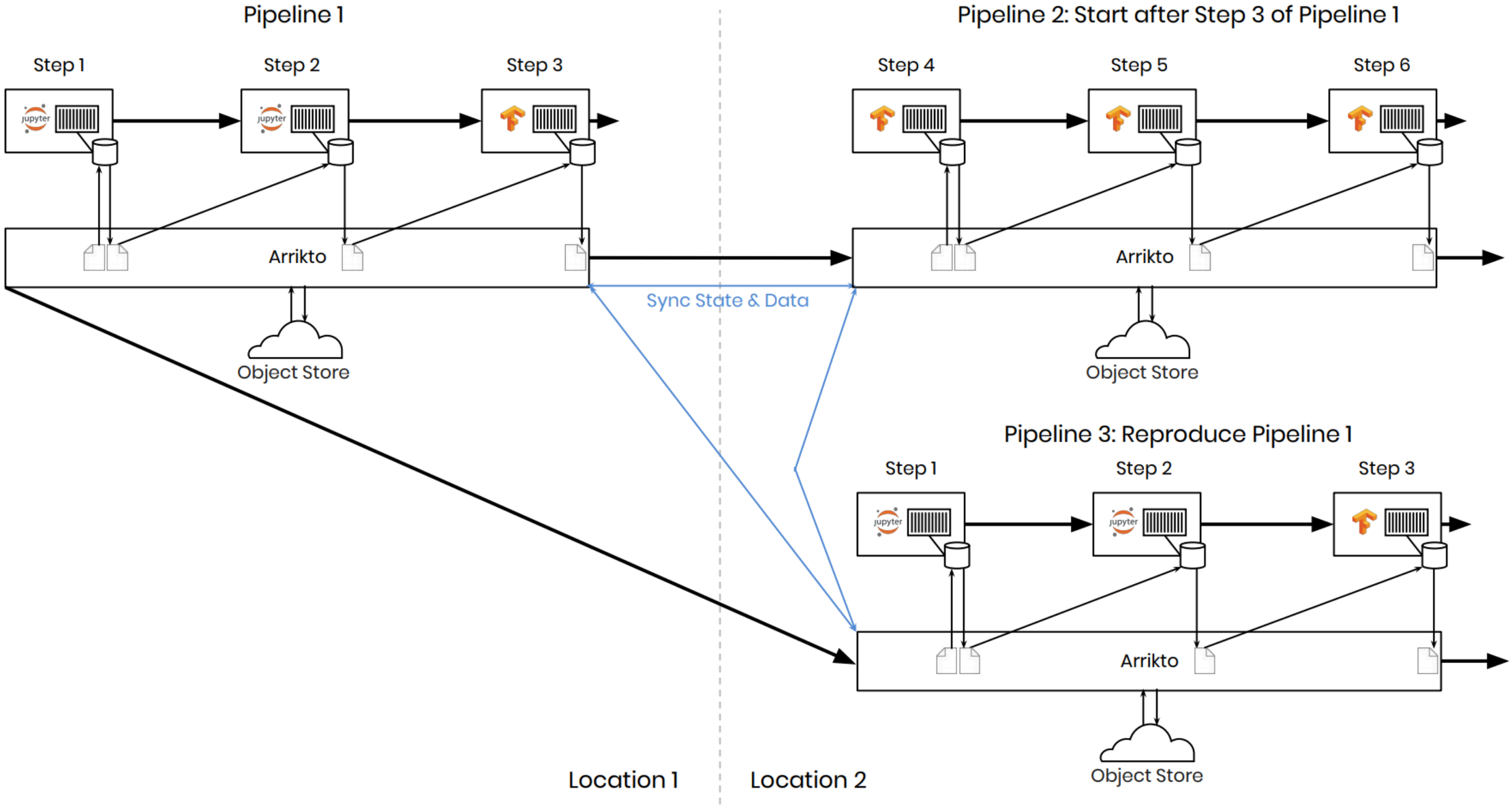
3. MODEL TRAINING

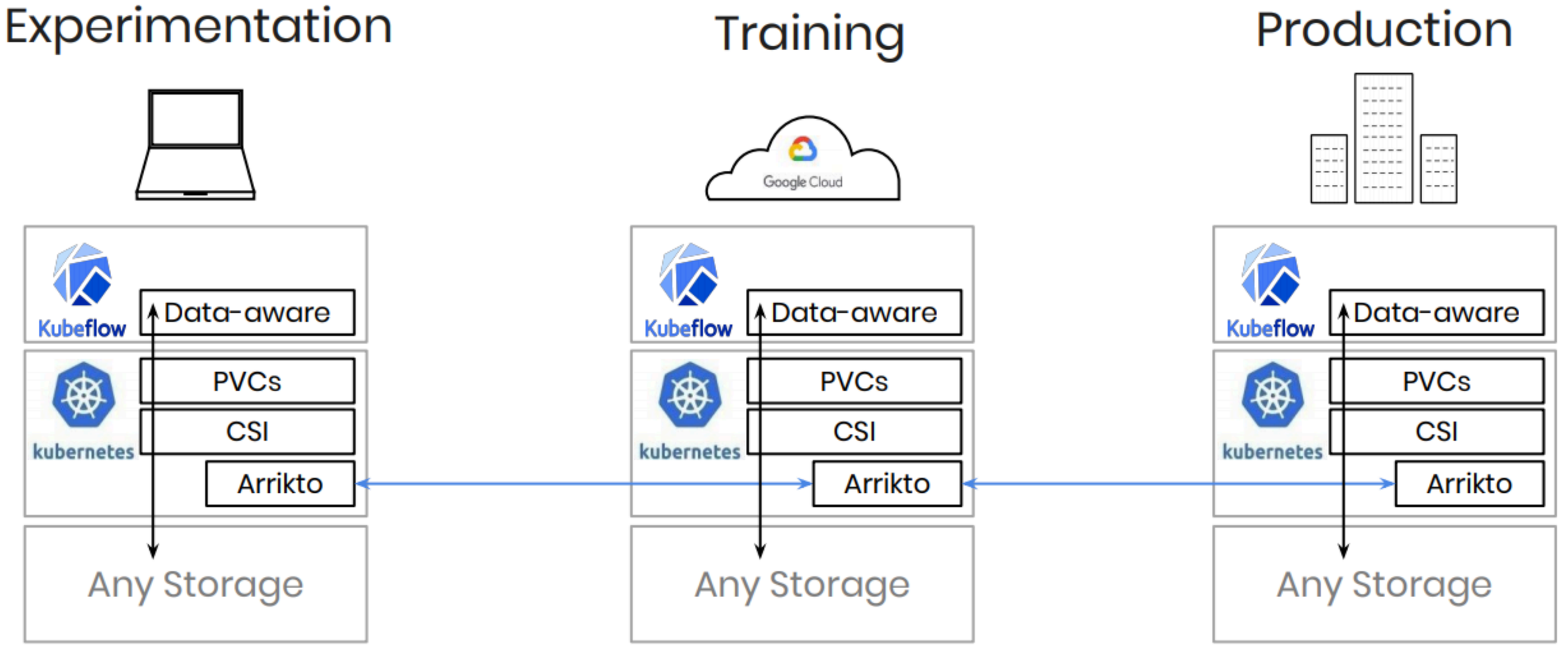
1. DATA VALIDATION
2. DATA PREPROCESSING
3. MODEL TRAINING
4. MODEL ANALYSIS

1. DATA VALIDATION
2. DATA PREPROCESSING
3. MODEL TRAINING
4. MODEL ANALYSIS
5. PREDICTION


External source data
jupyter
Notebook
Volume
Snapshot
Snapshot


Step 1
Step 2
Step 3
jupyter
jupyter
1. Clone disk from snapshot
2. Do initial analysis
3. Snapshot
4. Clone disk of Step 1
5. Transform data
6. Snapshot
7. Clone disk of Step 2
8. Train model
9. Snapshot
Arrikto
aws
Object Store


Pipeline 1
Step 1
Step 2
Step 3
Arrikto
Object Store
Pipeline 2: Start after Step 3 of Pipeline 1
Step 4
Step 5
Step 6
Arrikto
Object Store
Sync State & Data
Pipeline 3: Reproduce Pipeline 1
Step 1
Step 2
Step 3
Arrikto
Object Store
Location 1
Location 2


Experimentation
Training
Production
Google Cloud
Kubeflow
Data-aware
PVCs
CSI
kubernetes
Arrikto
Any Storage
Kubeflow
Data-aware
PVCs
CSI
kubernetes
Arrikto
Any Storage
Kubeflow
Data-aware
PVCs
CSI
kubernetes
Arrikto
Any Storage

MORE...

Network topology

K8s General Purpose Software

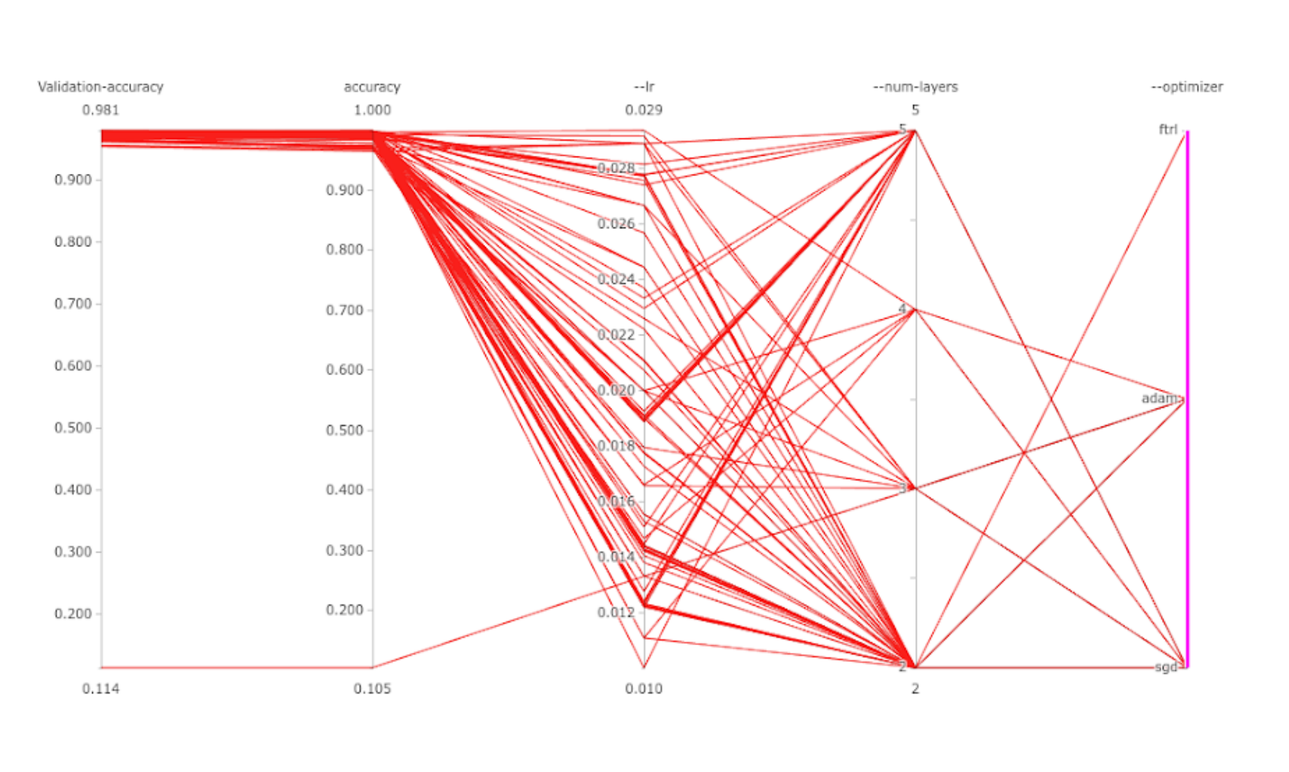
TF Job

```yaml
apiVersion: kubeflow.org/v1
kind: TFJob
metadata:
  generateName: tfjob
  namespace: kubeflow
spec:
  tfReplicaSpecs:
    PS:
      replicas: 1
      restartPolicy: OnFailure
      .
      .
    Worker:
      replicas: 3
      restartPolicy: OnFailure
      template:
        spec:
          containers:
          - name: tensorflow
            image: gcr.io/your-project/your-image
            command:
              - python
              - -m
              - trainer.task
              - --batch_size=32
              - --training_steps=1000
```

Fairing Attribute

Katib

Experiment Name: mnist-experiment



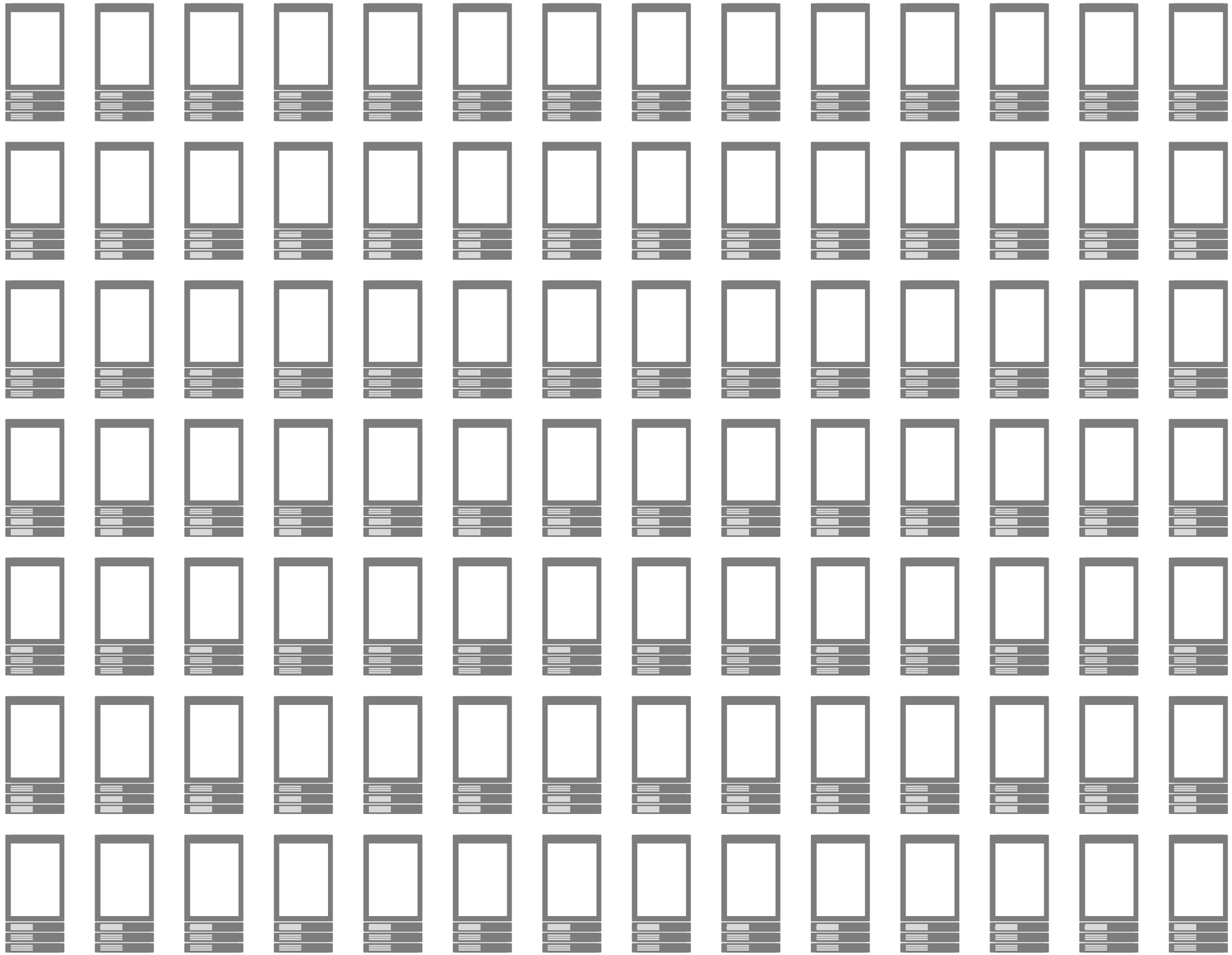
trialName	Validation-accuracy	accuracy	--lr	--num-layers	--optimizer
mnist-experiment-5p7tvdgn	0.979299	0.992031	0.01796154198777169	3	sgd
mnist-experiment-ccdgdp5v	0.981190	0.998906	0.014491188430762471	5	sgd
mnist-experiment-n24h6lwd	0.976712	0.992656	0.02915797314372763	3	sgd
mnist-experiment-lwf8dq49	0.962878	0.984062	0.017791967922146663	2	adam
mnist-experiment-2qv4mtsf	0.964869	0.974688	0.014689376618494194	4	adam
mnist-experiment-msdlqrjp	0.978603	0.994687	0.01775280397488328	2	sgd
mnist-experiment-pcr92pt2	0.957404	0.967969	0.022601601543350176	3	adam
mnist-experiment-njwhd2tr	0.977309	0.996719	0.029364679387401435	4	sgd

MACHINE LEARNING AT SCALE WITH K8S

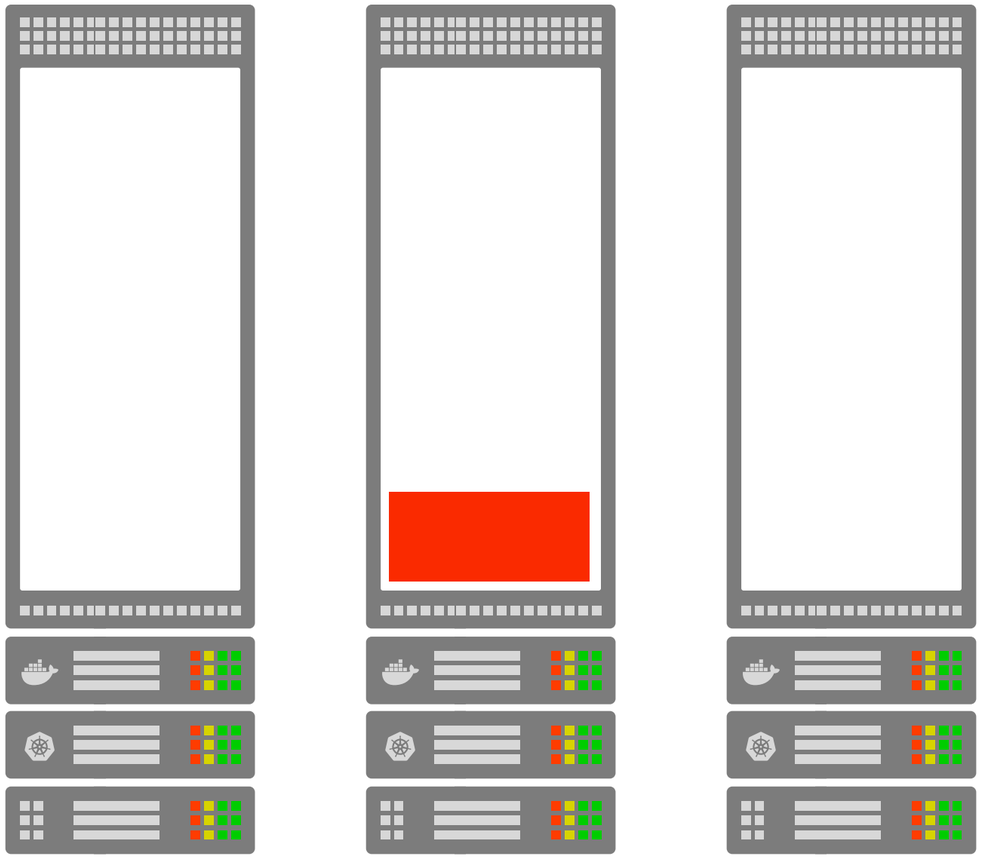
Bloomberg

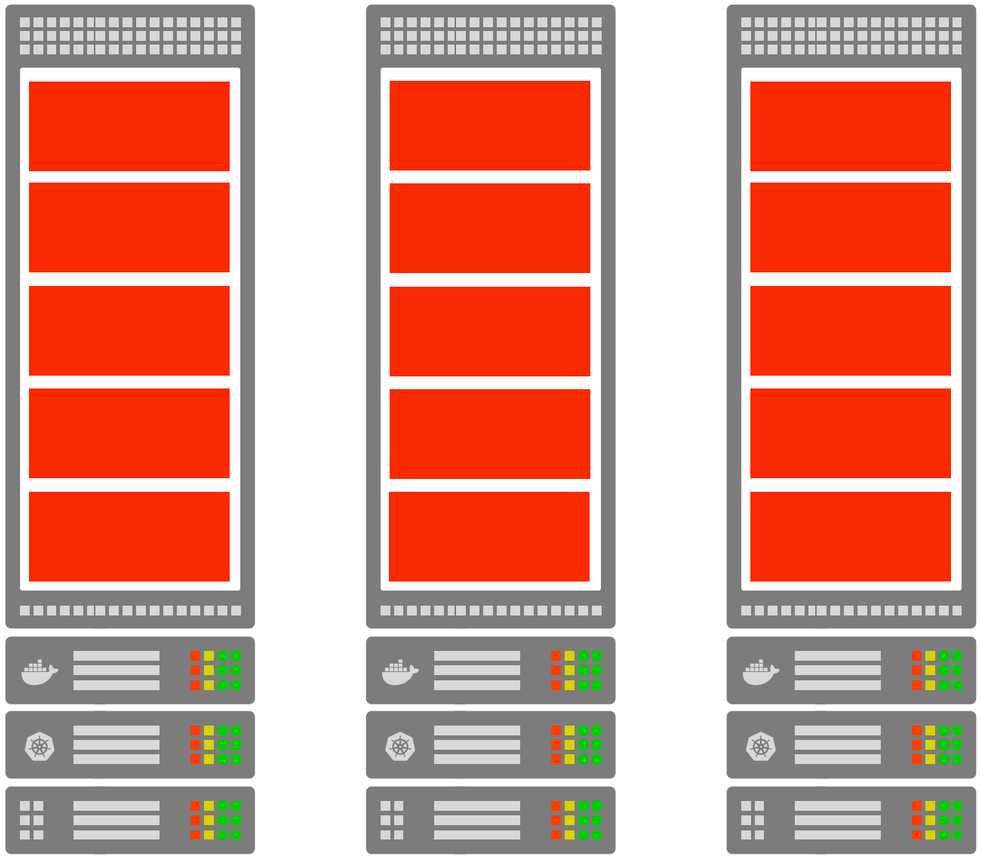
6000 INSTANCES OF APACHE SOLR ~ 1000 SERVERS

TRAINING ML WITH 2500 NODES

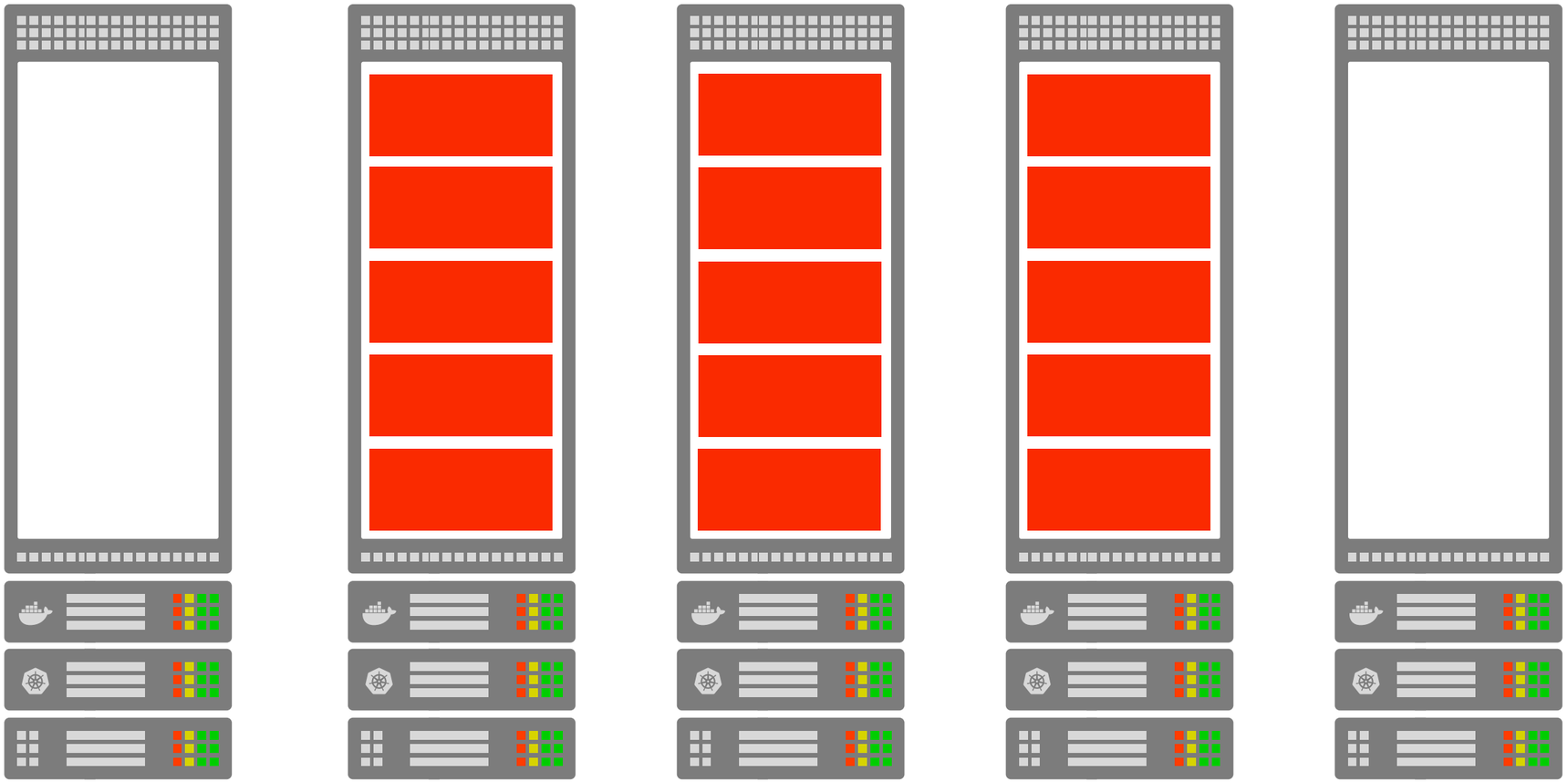


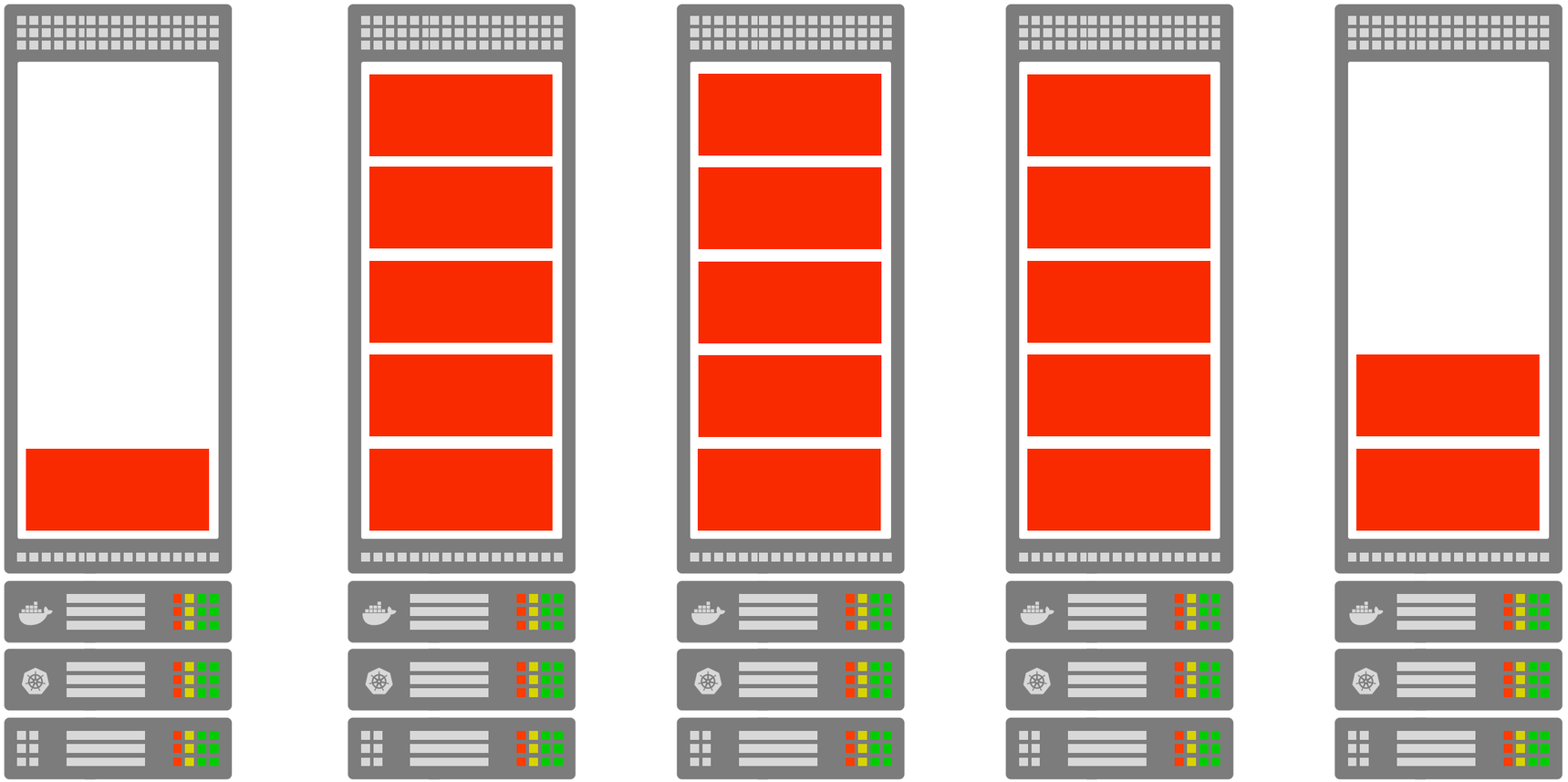
HORIZONTAL POD AUTOSCALER





CLUSTER AUTOSCALER





OpenAI


MVP
593
24
47:09
38
10
LGD
4955
GAME STATS
LGD.Sep.战旗TV
LGD.Agressif.战旗TV
LGD.xiao8.战旗TV
LGD.Maybe.战旗TV
LGD.MMY!.战旗TV
CLOCKWERK
16
1123/1700
1691 / 1955
+6.6
370 / 921
+2.3
111
8
-7
110
68+12
47+12
36+22
15
75
13
80
50
6
150
2378
SHOP
3
17
1
13
K/D/A 4/6/21
LH/D 91/16

256 P100

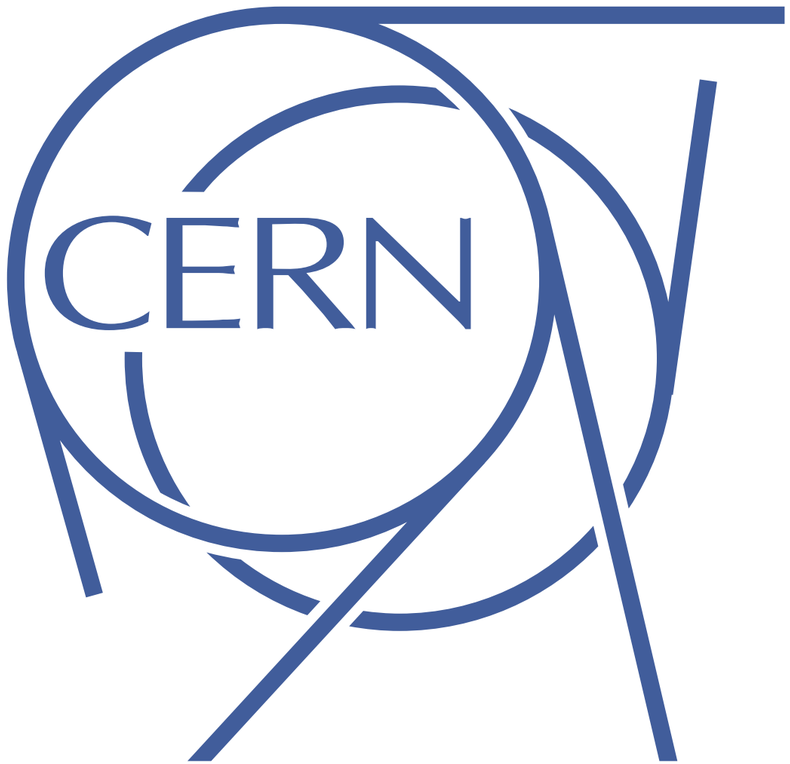
128000 PREEMPTIBLE VCPUS

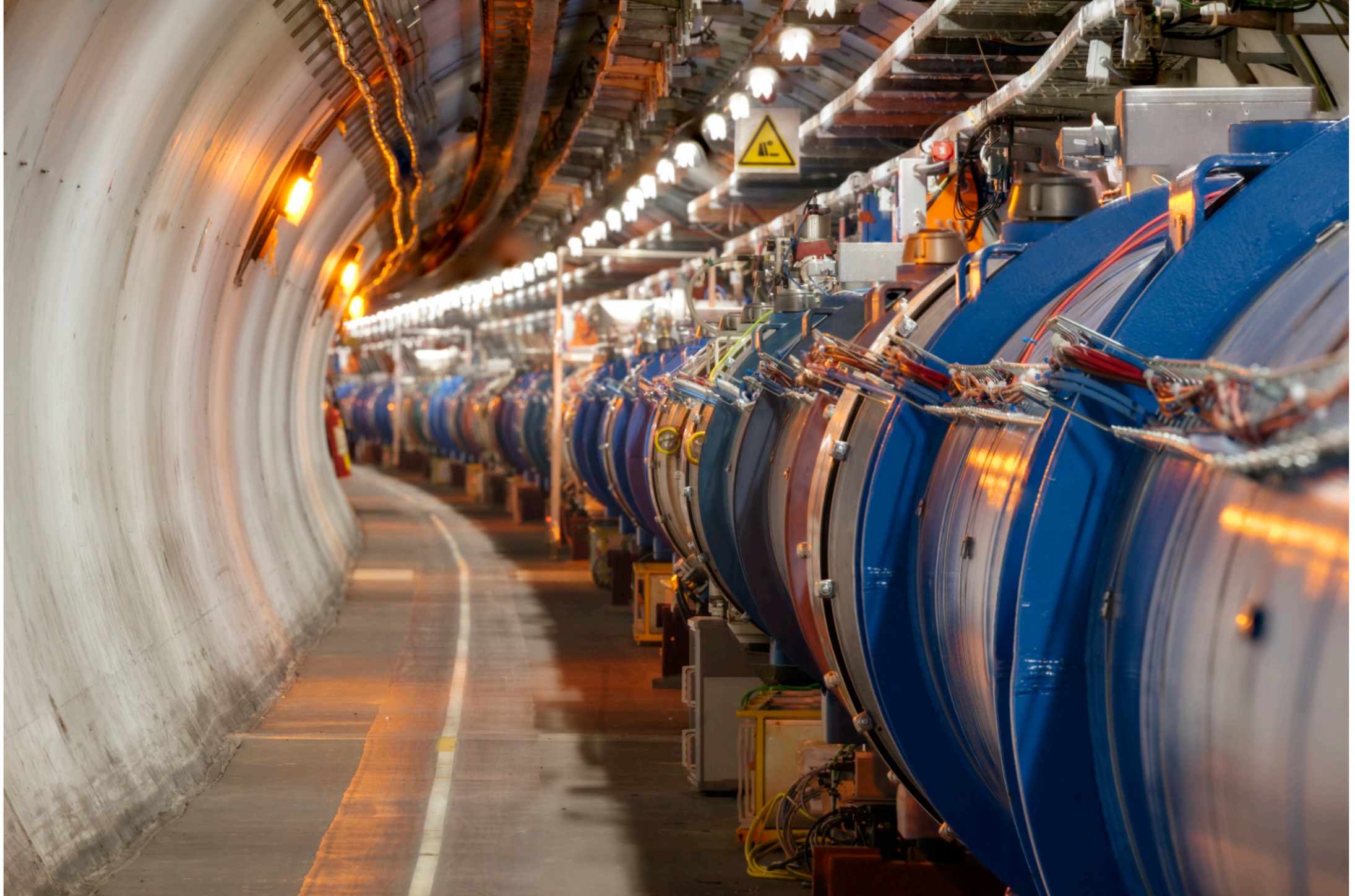
256 P100 $400/hr

128000 preemptible vCPUs $1280/hr

256 P100 $400/HR

128000 PREEMPTIBLE VCPUS $1280/HR


CERN



320,000 CORES

320,000 CORES

4,300 PROJECTS

320,000 CORES

4,300 PROJECTS

330 PETABYTES

3,300 USERS

3,300 USERS

10,000 HYPERVISORS

3,300 USERS

10,000 HYPERVISORS

210 KUBERNETES CLUSTERS

Not free 

Docker: ~950K LOC

Kubernetes: ~3.1M

Kubeflow: ~830K

Pipelines: ~220K

KFServing: ~3.1M ~

...

Kubernetes

Containers

Kubeflow - run anywhere

Demo

WHAT'S NEXT?

Resources ☸

- https://www.kubeflow.org/
- https://kubernetes.io/
- https://learnk8s.io/academy/
- https://deploy.kubeflow.cloud
- bit.ly/kubeflow-taxi
- https://www.arrikto.com/

Resources II ☸

- https://kubernetes.io/case-studies/cern/
- Kubernetes @ CERN
- Kubernetes @ CERN presentation

Thanks!

- @soulmaniqbal
- @learnk8s