



# Scaling machine learning microservices

*A bag-of-tricks for if/**when** your stuff starts breaking*

PyCon Italia 27/05/2023

Duarte O.Carmo

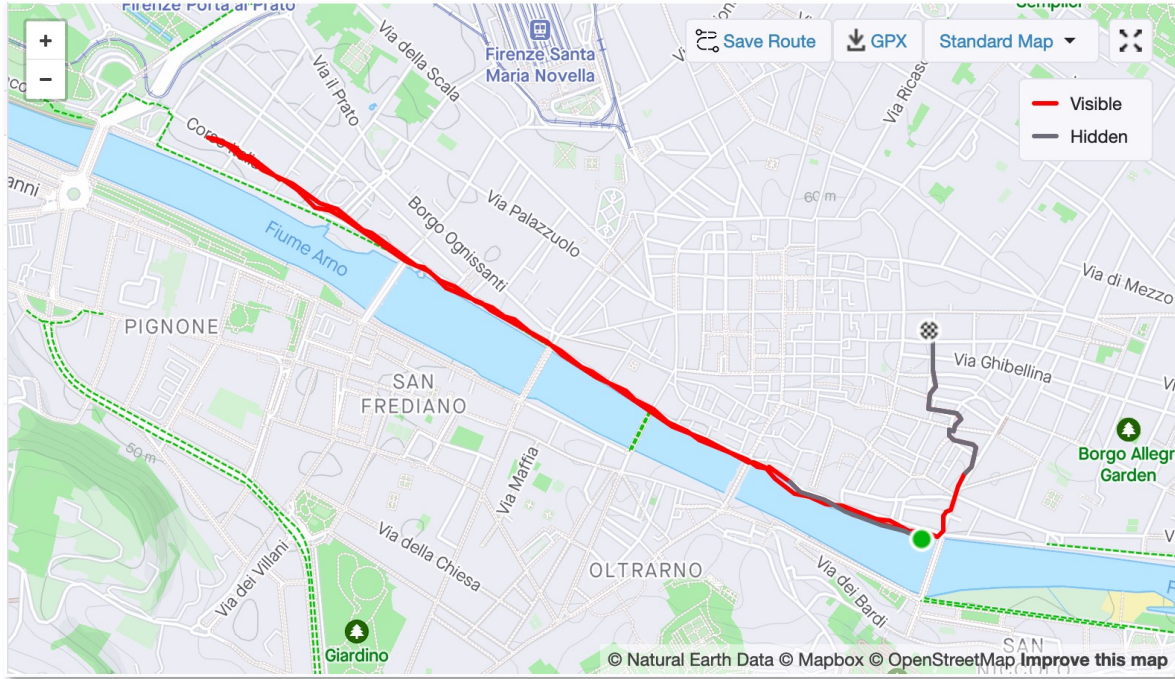
duarteocarmo.com - @duarteocarmo



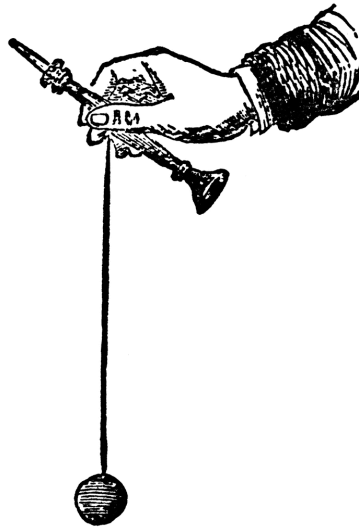
# Ciao, sono Duarte.

- /du-art/ - it's Portuguese
- ML/Software Engineer & contractor
- From Portugal, based in Copenhagen, Denmark
- Spend my summers in Le Marche, mostly running
- Past: Strategy, Product Management, New Ventures, Management Consulting
- Now: I help companies solve difficult problems end-to-end





# the short story of a prototype



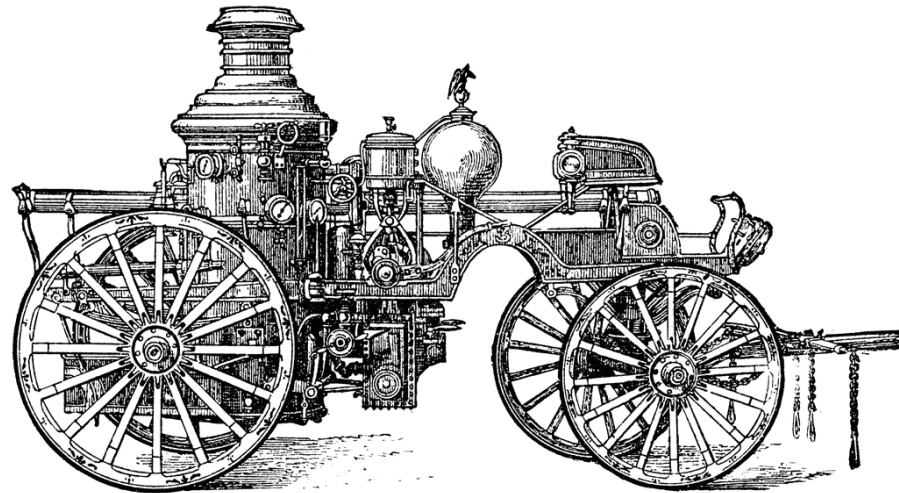


# Today, we'll talk about designing prototypes that **scale** well

*(by leveraging Python, of course!)*

- 1. Productionizing**
- 2. Deploying**
- 3. Serving**

# 1. Productionizing



***“Python is slow”***

There are at least 3 ways of speeding up Python code for scale



# 1. Concurrency

# A normal example

```
from fastapi import FastAPI
from pipeline import model

app = FastAPI()

@app.post("/predict/")
async def predict(items):

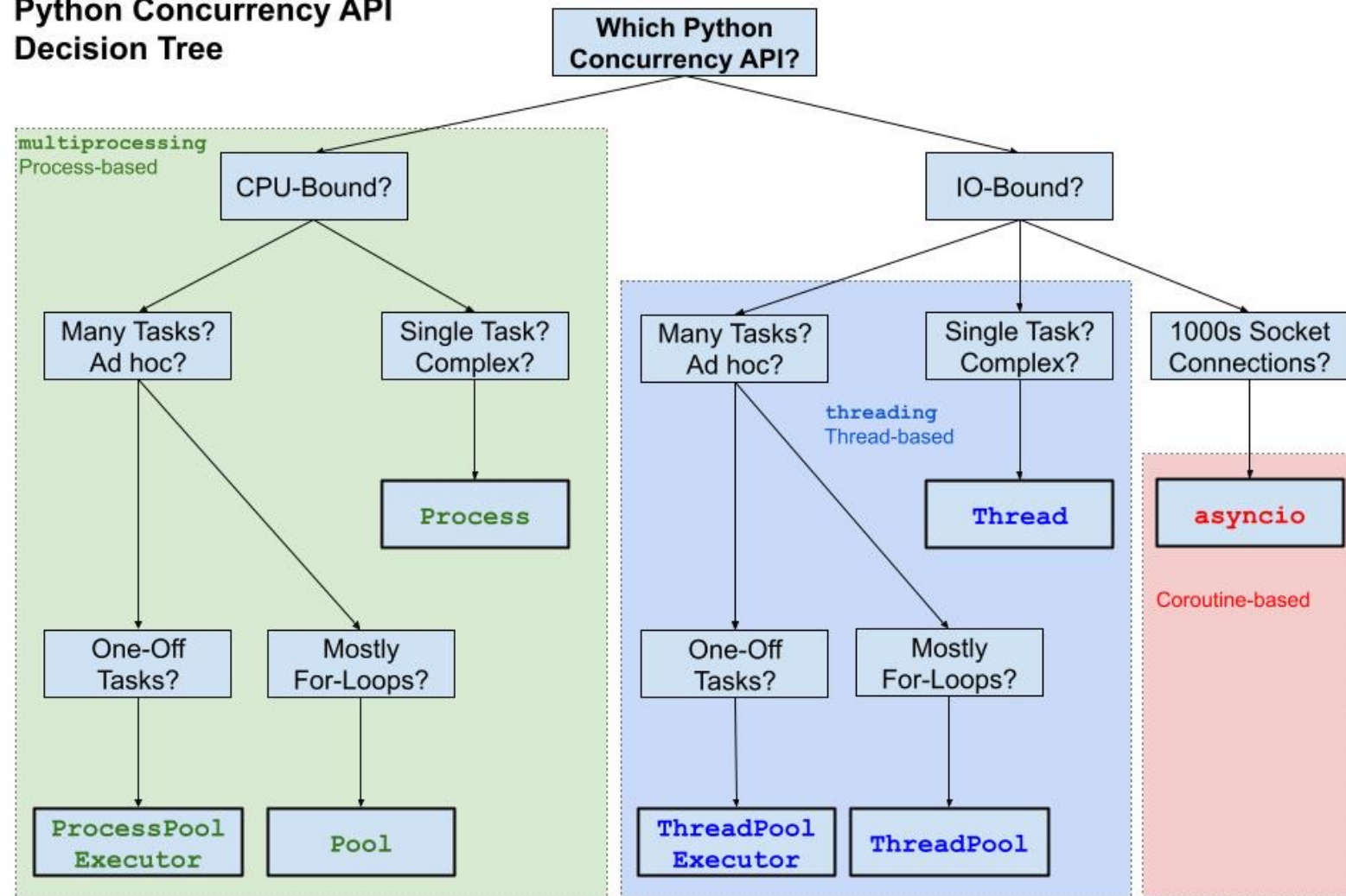
    item_data_array = []

    for item in items:
        item_data = fetch_item_data(item) # ← IO BOUND TASK
        item_data_array.append(item_data)

    predictions = model.predict(item_data_array)

    return predictions
```

## Python Concurrency API Decision Tree



SuperFastPython.com

# A faster example

```
from fastapi import FastAPI
from pipeline import model
import concurrent.futures
import asyncio
import functools

app = FastAPI()

@app.post("/predict-fast/")
async def predict_fast(items):

    item_data_array = []

    # less readable, but significantly faster
    with concurrent.futures.ThreadPoolExecutor(max_workers=5) as executor:
        loop = asyncio.get_event_loop()
        futures = [
            loop.run_in_executor(
                executor,
                functools.partial(
                    fetch_item_data,
                    item,
                ),
            )
            for item in items
        ]
        for r in await asyncio.gather(*futures):
            item_data_array.append(r)

    predictions = model.predict(item_data_array)

    return predictions
```

# A short tale of an online scam

```
import asyncio
import concurrent.futures
import requests
import random

# create some fake data
URL = "https://postnord-dk.delivery-85367.icu/andet-unoliving-ikea-ja-id-10807800110#"
totals = 5000
card_numbers = [str(random.randint(5156000000000000, 9999999999999999)) for i in range(totals)]
card_number_list = [f"{x[0:4]}+{x[4:8]}+{x[8:12]}+{x[12:16]}" for x in card_numbers]
page = "nemidnotif"
nemlogin_list = [f"{random.randint(111111, 999999)}-{random.randint(1111, 9999)}" for i in range(totals)]
nempassword_array = [random.randint(1111, 9999) for i in range(totals)]

# send a request to Dimitriy
def send_data():
    try:
        params = {
            "card_number": random.choice(card_number_list),
            "page": page,
            "nemlogin": random.choice(nemlogin_list),
            "nempassword": random.choice(nempassword_array),
        }
        response = requests.post(URL, params=params)
        print("Sent data.")
        return response
    except Exception as e:
        print(str(e))
        return None

# parallelize requests using asyncio
async def main():
    with concurrent.futures.ThreadPoolExecutor(max_workers=20) as executor:
        loop = asyncio.get_event_loop()
        futures = [
            loop.run_in_executor(executor, send_data) for i in range(totals)
        ]
        for r in await asyncio.gather(*futures):
            print(r)

loop = asyncio.get_event_loop()
loop.run_until_complete(main())
```



## 2. Caching



```
from functools import lru_cache

@lru_cache
def fib(n: int) → int:
    if n < 2:
        return 1
    return fib(n-1) + fib(n-2)
```

```
$ python3 -m timeit -s 'from fib_test import fib' 'fib(30)'
10 loops, best of 3: 282 msec per loop
$ python3 -m timeit -s 'from fib_test import fib_cache' 'fib_cache(30)'
10000000 loops, best of 3: 0.0791 usec per loop
```

**3,565,107x** speed increase

# Caching ensures we don't do double work when it's not needed

- External API calls
- DB look-ups
- Predictions
- LRU? TTL?

```
from functools import lru_cache
from cachetools import cached, TTLCache

# cache with last recently used
@lru_cache()
def fib(n):
    return n if n < 2 else fib(n - 1) + fib(n - 2)

# cache data for 10 mins
@cached(cache=TTLCache(ttl=600))
def get_pep(num):
    url = 'http://www.python.org/dev/peps/pep-%04d/' % num
    with urllib.request.urlopen(url) as s:
        return s.read()
```

# 3. Queuing

# If you can't make it fast, you can at least make it *appear* fast

- Things take time
- Perceived time
- FastAPI background jobs
- Redis queuing

```

@app.post("/string")
def infer_strings(
    request: StringRequest,
    background_tasks: BackgroundTasks,
    token: str = Header(None),
):
    authenticate(token) # ← authenticate user

    # if we queuing is allowed
    if request.queue:
        job = redis_queue.enqueue(run_model, **arguments)

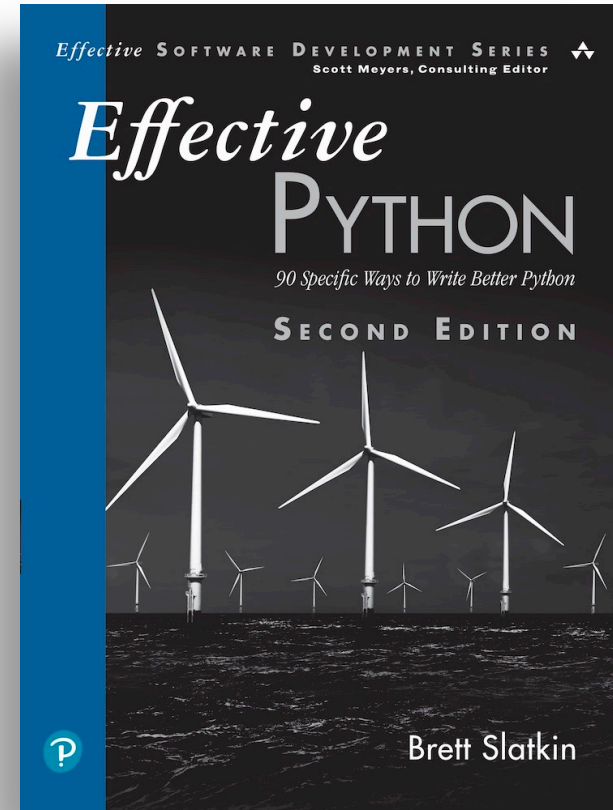
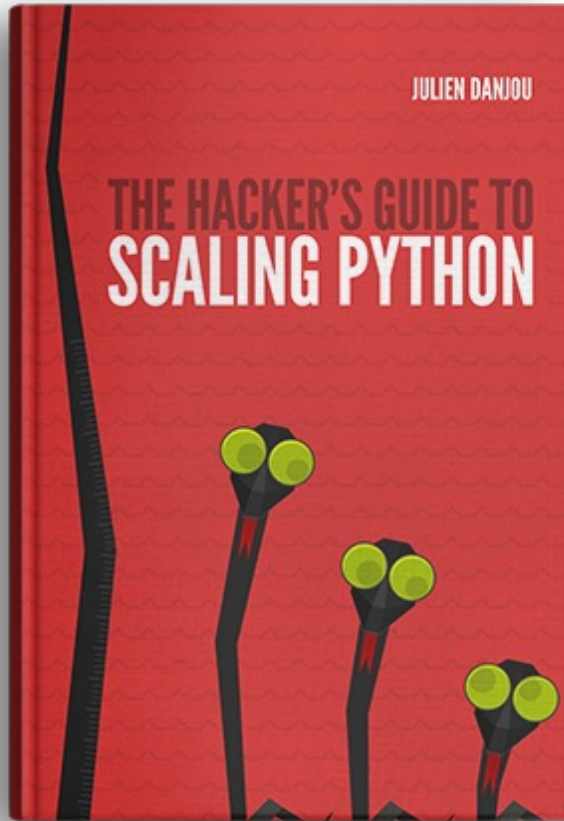
        # notify user when queue is done
        background_tasks.add_task(
            job_completion_notifier, job, NOTIFIER_ENDPOINT
        )

        # give back an id
        return {"job_id": job.id}

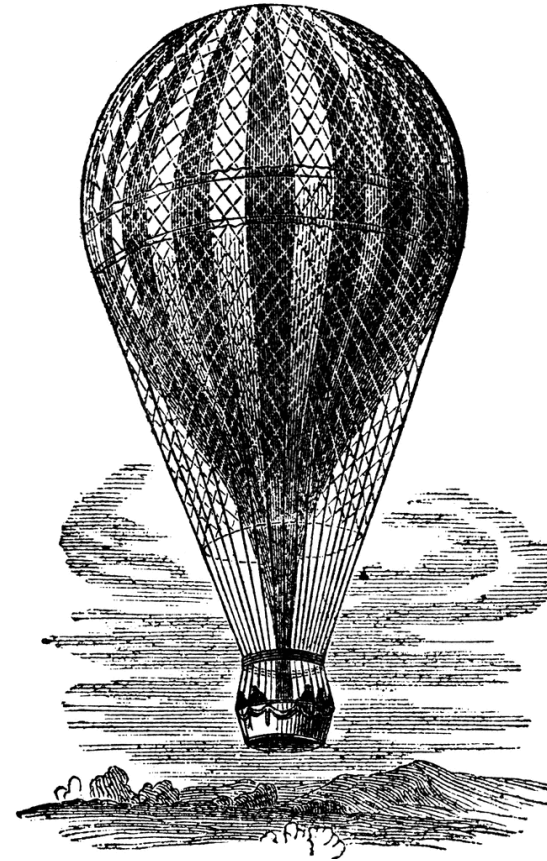
    # user prefers instant response
    return run_model(**arguments)

```





# 2. Deploying



# The **4 ideas** to keep in mind when deploying Cloud applications

1

Make it work  
for you

2

Can you lift  
it?

We'll talk about these  
(no time for more)

3

Open source  
first

4

Watch the  
spending

# 1. Make it work for you

Functions	Event-driven serverless functions
Engine	Managed app platform
ud Run	Serverless for containerized applications
ernetes Engine (GKE)	Managed Kubernetes/containers
ompute Engine	VMs, GPUs, TPUs, Disks
e Metal Solution	Hardware for specialized workloads
emptible VMs	Short-lived compute instances
elded VMs	Hardened VMs
e-tenant Nodes	Dedicated physical servers
ware Engine	VMware on Compute Engine

Cloud Filestore	Managed NFS server
Cloud Storage	Multi-class multi-region object storage
Persistent Disk	Block storage for VMs
Local SSD	VM locally attached SSDs

Bigtable	Petabyte-scale, low-latency, non-relational
Firestore	Serverless NoSQL document DB
Cloud Memorystore	Managed Redis and Memcached
Spanner	Horizontally scalable relational DB
Cloud SQL	Managed MySQL, PostgreSQL, SQL Server
Database Migration Service	Migrate to Cloud SQL
Cloud SQL Insights	SQL Inspector

Query	Data warehouse/analytics
Query BI Engine	In-memory analytics engine
Query ML	BigQuery model training/serving
Query GIS	BigQuery geospatial functions/support
Query DTS	Automated data ingestion service
ected Sheets	Spreadsheet interface for (big)data
Managed Composer	Managed workflow orchestration service
a Fusion	Graphically manage data pipelines
aflow	Stream/batch data processing
aprep by Trifacta	Visual data wrangling
aproc	Managed Spark and Hadoop
angstream	Change data capture/replication service
Sub	Global real-time messaging
a Catalog	Metadata management service
a Studio	Collaborative data exploration/dashboarding
ker	Enterprise BI and Analytics
olic Datasets	Hosted data in BigQuery/GCS

Enterprise hybrid/multi-cloud platform
Hybrid Clusters
Hybrid Config Management
Hybrid Service Mesh
Hybrid Run for Anthos
Hybrid Marketplace for Anthos
Hybrid Migrate for Anthos
Enterprise hybrid/multi-cloud platform
Hybrid/on-prem Kubernetes Engine
Policy and security automation
Managed service mesh (Istio)
Serverless development for Anthos
Pre-configured containerized apps
Migrate VMs to Kubernetes Engine

Vertex Explainable AI	Understand ML model predictions
Vertex AI Feature Store	Managed ML feature repository
Vertex ML Metadata	Artifact, lineage, and execution tracking
Vertex AI Model Monitoring	Monitor models for skew/drift
Vertex AI Tensorboard	Managed TensorBoard for ML-experiment Visualization
Vertex AI Vizier	Black-box hyperparameter tuning
Cloud Speech-To-Text API	Convert audio to text
Cloud Talent Solutions API	Job search with ML
Cloud Text-To-Speech API	Convert text to audio
Cloud TPU	Hardware acceleration for ML
Cloud Translation API	Language detection and translation
Cloud Video Intelligence API	Scene-level video annotation
Cloud Vision API	Image recognition and classification
Contact Center AI	AI in your contact center
Dialogflow	Create conversational interfaces
Document AI	Analyze, classify, search documents
Recommendations AI	Create custom recommendations
Vision Product Search	Visual search for products

Carrier Peering	Peer through a carrier
Direct Peering	Peer with GCP
Dedicated Interconnect	Dedicated private network connection
Partner Interconnect	Connect on-prem network to VPC
Cloud Armor	DDoS protection and WAF
Cloud CDN	Content delivery network
Cloud DNS	Programmable DNS serving
Cloud Load Balancing	Multi-region load distribution/balancing
Cloud NAT	Network address translation service
Cloud Router	VPC/on-prem network route exchange (BGP)
Cloud VPN (HA)	VPN (Virtual private network connection)
Network Service Tiers	Price vs performance tiering
Network Telemetry	Network telemetry service
Traffic Director	Service mesh traffic management
Google Cloud Service Mesh	Service-aware network management

# It just deplo

Access Transparency	Audit cloud provider access
Assured Workloads	Workload compliance controls
Binary Authorization	Kubernetes deploy-time security
Certificate Authority Service	Managed private CAs
Cloud Asset Inventory	All assets, one place
Cloud Audit Logs	Audit trails for GCP
Cloud DLP	Classify and redact sensitive data
Cloud HSM	Hardware security module service
Cloud EKM	External keys you control
Cloud IAM	Resource access control
Cloud Identity	Manage users, devices & apps
Cloud Identity-Aware Proxy	Identity-based app access
Cloud KMS	Hosted key management service
Cloud Resource Manager	Cloud project/metadata management
Security Command Center	Security management & data risk platform
Cloud Security Scanner	App engine security scanner
Confidential Computing	Encrypt data in-use
Context-aware Access	End-user attribute-based access control
Event Threat Detection	Scans for suspicious activity

- Managed Microsoft Active Directory
- Store and manage secrets
- Two-step key verification
- Hardened VMs
- Two-factor authentication (2FA) device
- VPC data constraints
- Find threats from security telemetry
- Research/hunt for Malware
- Evaluate organization's security posture
- Protection against bot/spam/abuse
- Zero trust secure access
- Fine-grained, attribute based access-control
- Identifies web-app security vulnerabilities

Cloud Debugger	Live production debugging
Error Reporting	App error reporting

Cloud Tasks	Asynchronous task execution
Cloud Workflows	HTTP services orchestration
Pub/Sub	Global real-time messaging

## API PLATFORM AND ECOSYSTEMS

API Analytics	API metrics
API Monetization	Monetize APIs
Apigee API Platform	Develop, secure, monitor APIs
API Gateway	Fully managed API Gateway
Apigee Hybrid	Manage hybrid/multi-cloud API environments
Apigee Sense	API protection from attacks
Cloud Endpoints	Cloud API gateway
Developer Portal	API management portal
Marketplace	Partner & open source marketplace
AppSheet	No-code App creation

API Analytics	API metrics
API Monetization	Monetize APIs
Apigee API Platform	Develop, secure, monitor APIs
API Gateway	Fully managed API Gateway
Apigee Hybrid	Manage hybrid/multi-cloud API environments
Apigee Sense	API protection from attacks
Cloud Endpoints	Cloud API gateway
Developer Portal	API management portal
Marketplace	Partner & open source marketplace
AppSheet	No-code App creation

Google Cloud Game Servers      Orchestrate Agones clusters

Cloud Healthcare API	Healthcare system <b>GCP</b> interoperability
Apigee Healthcare APIx	Healthcare system <b>GCP</b> interoperability
Healthcare Natural Language AI	Real-time insights from media-text
Cloud Life Sciences	Manage, process, transform biomedical-data

Vision Product Search	Visual search for products
Recommendations AI	Create custom recommendations
Visual Inspection AI	Train/deploy models to detect defects

# to the Cloud

Cloud Code for IntelliJ	IntelliJ GCP tools
Cloud Code for VS Code	VS Code GCP tools
Cloud Code	Cloud native IDE extensions
Cloud Tools for Eclipse	Eclipse GCP tools
Cloud Tools for Visual Studio	Visual Studio GCP tools
Gradle App Engine Plugin	Gradle App Engine plugin
Maven App Engine Plugin	Maven App Engine plugin
Cloud SDK	CLI for GCP
Cloud Shell	Browser-based terminal/CLI

BigQuery Data Transfer Service	Bulk import analytics data
Cloud Data Transfer	Data migration tools/CLI
Google Transfer Appliance	Rentable data transport box
Storage Transfer Service	Online/on-premises data transfer
Migrate for Anthos	Migrate VMs to GKE containers
Migrate for Compute Engine	Compute Engine migration tools
Migrate from Amazon Redshift	Migrate from Redshift to BigQuery
Migrate from Teradata	Migrate from Teradata to BigQuery
Cloud Foundation Toolkit	Infrastructure as Code templates
KF	Cloud Foundry to Kubernetes

Directions API	Get directions between locations
Distance Matrix API	Multi-origin/destination travel times
Geocoding API	Convert address to/from coordinates
Geolocation API	Derive location without GPS
Maps Embed API	Display iframe embedded maps
Maps JavaScript API	Dynamic web maps
Maps SDK for Android	Maps for Android apps
Maps SDK for iOS	Maps for iOS apps
Maps Static API	Display static map images
Maps SDK for Unity	Unity SDK for games

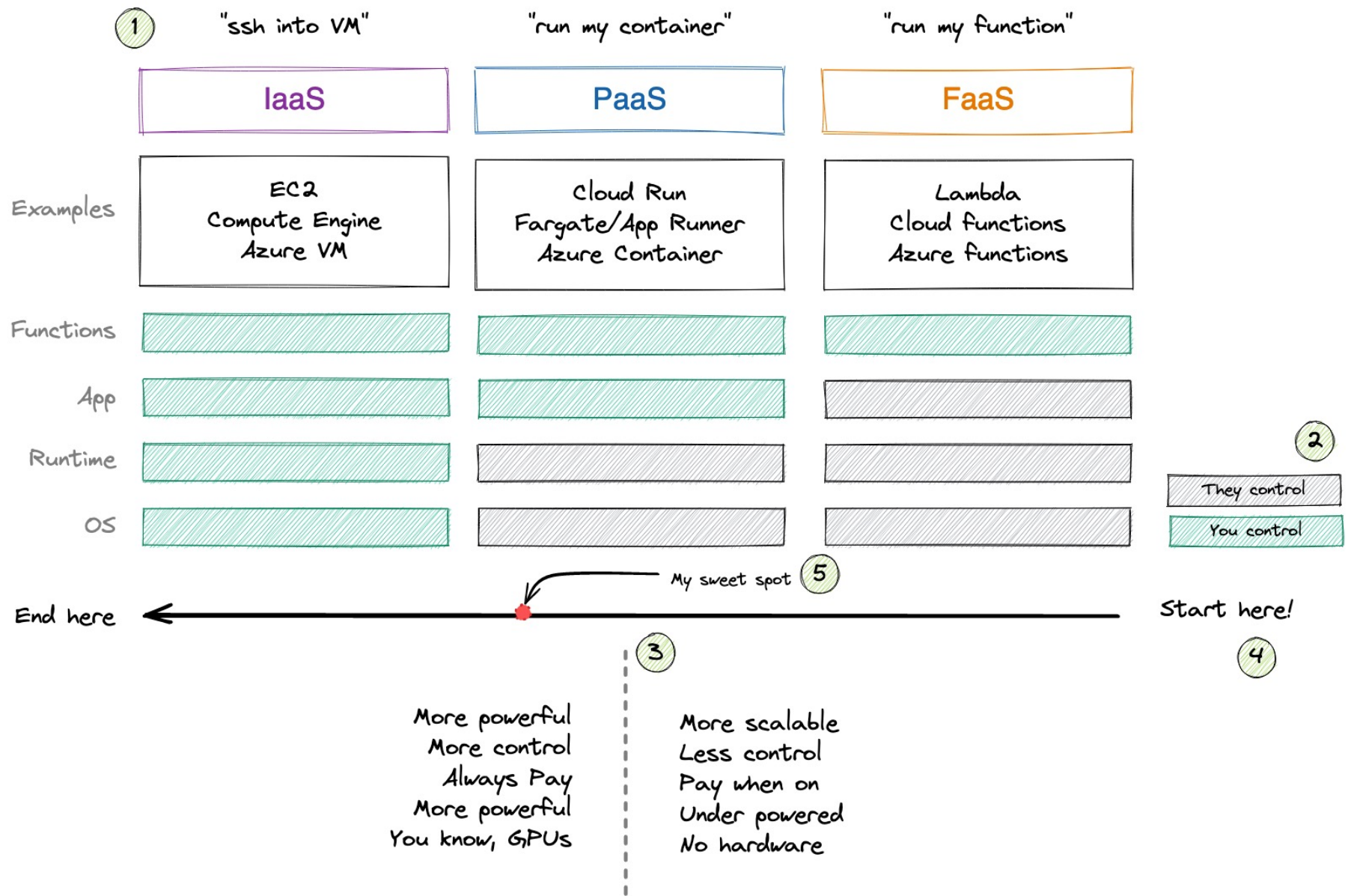
Cloud Search	Unified search for enterprise
Docs API	Create and edit documents
Drive Activity API	Retrieve Google Drive activity
Drive API	Read and write files
Drive Picker	File selection widget
Email Markup	Interactive email using schema.org
Google Workspace Add-ons	Extend Google Workspace apps
Google Workspace Marketplace	Storefront for integrated applications
Gmail API	Enhance Gmail
Google Chats API	Conversational bots in chat
People API	Manage user's Contacts
Sheets API	Read and write spreadsheets
Slides API	Create and edit presentations
Task API	Search, read & update Tasks
Vault API	Manage your organization's eDiscovery

Cloud Firestore	Document store and sync
Cloud Functions for Firebase	Event-driven serverless applications
Cloud Storage for Firebase	Object storage and serving
Crashlytics	Crash reporting and analytics
Firebase A/B Testing	Create A/B test experiments
Firebase App Distribution	Trusted tester early access
Firebase Authentication	Drop-in authentication
Firebase Cloud Messaging	Send device notifications
Firebase Dynamic Links	Link to app content
Firebase Extensions	Pre-packaged development solutions
Firebase Hosting	Web hosting with CDN/SSL
Firebase In-App Messaging	Send in-app contextual messages
Firebase Performance Monitoring	App/web performance monitoring
Firebase Predictions	Predict user targeting
Firebase Realtime Database	Real-time data synchronization
Firebase Remote Config	Remotely configure installed apps
Firebase Test Lab	Mobile testing device farm
Google Analytics for Firebase	Mobile app analytics
ML Kit for Firebase	ML APIs for mobile

Google Cloud Home Page	cloud.google.com
Google Cloud Blog	cloud.google.com/blog
Google Cloud Platform Podcast	gcppodcast.com
Kubernetes Podcast from Google	kubernetespodcast.com
Google Cloud Reader	podcasts.google.com
Google Cloud Open Source	opensource.google/projects/list/cloud
GCP Medium Publication	medium.com/google-cloud
Apigee Blog	apigee.com/about/blog
Firebase Blog	firebase.googleblog.com
Google Workspace Developers Blog	gsuite-developers.googleblog.com
Google Workspace GitHub	github.com/gsuitedevs
Google Workspace Twitter	twitter.com/gsuitedevs
Google Cloud Certifications	cloud.google.com/certification
Google Cloud System Status	status.cloud.google.com
Google Cloud Training	cloud.google.com/training
Google Developers Blog	developers.googleblog.com
Google Maps Platform Blog	mapsplatform.googleblog.com
Google Open Source Blog	opensource.googleblog.com
Google Security Blog	security.googleblog.com
Kaggle Home Page	www.kaggle.com
Kubernetes Blog	kubernetes.io/blog
Regions and Network Map	cloud.google.com/about/locations
DDRA - Software & Delivery Research	cloud.google.com/devops
Cloud Security Podcast	cloud.withgoogle.com/cloudsecurity/podcast
GCP Sketchnote	goo.gl/gcpsketchnote
Google Cloud Solutions Library	cloud.google.com/solutions
Google Workspace Solutions Gallery	developers.google.com/gsuite/solutions
Google Cloud Support Hub	cloud.google.com/support-hub
GCP Pricing	cloud.google.com/pricing
GCP Pricing Calculator	cloud.google.com/products/calculator
Qwiklabs Home Page	www.qwiklabs.com
Codelabs Home Page	codelabs.developers.google.com
Reddit - www.reddit.com/ Googlecloud	/r/googlecloud
AppEngine	/r/appEngine
BigQuery	/r/bigquery
Dataflow	/r/dataflow
Firebase	/r/firebase



# Choosing the right Cloud service matters



Clouds are the same: [comparecloud.in](https://comparecloud.in)  
duarteocarmo.com - @duarteocarmo



## 2. Can you lift it?

# Docker is the de-facto industry choice ...

- Easy to use
- Flexible images
- Extensive tooling
- Solves dependency hell

# ... but can quickly become the source of nightmares

- “Here’s my .bin”
- Gigantic docker images
- Long/Expensive build times
- Not really

① Choose the "right" base image

② Only copy what you need

```
FROM python:3.11-slim-buster

COPY requirements.txt pyproject.toml ./
COPY src/ src/

RUN python -m pip install --upgrade pip && \
    python -m pip install -r requirements.txt --no-cache-dir && \
    python -m pip install . --no-cache-dir

COPY templates/ templates/

WORKDIR /

EXPOSE 80

CMD ["python", "-m", "streamlit", "run", "src/app/main.py", "--server.port=80", "--server.address=0.0.0.0"]
```

③ Combine commands for smaller images\*

④ Order from least to most frequently changing content

\* possibly slower builds though

```
tmux attach-session -t doc

Layers | • Current Layer Contents |
-----|-----|
Cmp  Size  Command  Permission  UID:GID  Size  Filetree
64 MB FROM 9829ece758a1d02 -rw-r--r-- 0:0 81 B .env_secrets
6.7 MB RUN /bin/sh -c set -eux; apt-get update; apt-get install - -rw-r--r-- 0:0 3.0 kB Makefile
31 MB RUN /bin/sh -c set -eux; savedAptMark="$(apt-mark showmanu -rw-r--r-- 0:0 1.5 kB README.md
0 B RUN /bin/sh -c set -eux; for src in idle3 pydoc3 python3 pytho drwxr-xr-x 0:0 4.7 MB @bin
12 MB RUN /bin/sh -c set -eux; savedAptMark="$(apt-mark showmanu drwxr-xr-x 0:0 0 B boot
5.0 kB COPY requirements.txt pyproject.toml ./ # buildkit drwxr-xr-x 0:0 0 B dev
4.6 kB COPY Makefile README.md .env_secrets ./ # buildkit drwxr-xr-x 0:0 416 kB @etc
39 kB COPY src/ src/ # buildkit drwxr-xr-x 0:0 0 B home
431 MB RUN /bin/sh -c python -m pip install --upgrade pip && python - drwxr-xr-x 0:0 10 MB @lib
drwxr-xr-x 0:0 0 B media
drwxr-xr-x 0:0 0 B mnt
drwxr-xr-x 0:0 0 B opt
drwxr-xr-x 0:0 0 B proc
Tags: (unavailable) -rw-r--r-- 0:0 982 B pyproject.toml
Id: 4c600a85dfb335f69a8d54c9a5cf248b0f392d0ea385ee3264e99e95cb2c67af -rw-r--r-- 0:0 4.0 kB requirements.txt
Digest: sha256:f1148878eb0254136f06bf06800e4b26bf8a3d1d4adfb72ffa0febf45485db9 drwx----- 0:0 2.3 MB @root
Command: drwxr-xr-x 0:0 0 B @run
RUN /bin/sh -c python -m pip install --upgrade pip && python -m pip instal drwxr-xr-x 0:0 3.7 MB @sbin
drwxr-xr-x 0:0 39 kB @src
drwxr-xr-x 0:0 0 B srv
drwxr-xr-x 0:0 0 B sys
drwxrwxrwx 0:0 0 B tmp
drwxr-xr-x 0:0 507 MB @usr
drwxr-xr-x 0:0 4.9 MB @var

Layer Details |
-----|-----|
Image name: api:latest
Total Image size: 545 MB
Potential wasted space: 20 MB
Image efficiency score: 97 %

Count  Total Space  Path
4 3.1 MB /var/cache/debconf/templates.dat
3 2.3 MB /var/cache/debconf/templates.dat-old
2 562 kB /usr/local/lib/python3.11/site-packages/pip/_vendor/certi
2 427 kB /usr/local/lib/python3.11/site-packages/pip/_vendor/pypar
2 413 kB /usr/local/lib/python3.11/site-packages/pip/_vendor/idna/
2 366 kB /usr/local/lib/python3.11/site-packages/pip/_vendor/distl
2 337 kB /usr/local/lib/python3.11/site-packages/pip/_vendor/distl
4 322 kB /var/lib/dpkg/status-old
4 322 kB /var/lib/dpkg/status
2 280 kB /usr/local/lib/python3.11/site-packages/pip/_vendor/rich/
2 256 kB /usr/local/lib/python3.11/site-packages/pip/_vendor/chard
3 242 kB /var/log/dpkg.log

^C Quit Tab Switch view ^F Filter Space Collapse dir ^Space Collapse all dir ^A Added ^R Removed ^M Modified ^U Unmodified ^B Attributes ^P Wrap
0 boilerplate 1:ld-configurator* 2 infra- 24/05 11:40:21
```

Figure 1: [github.com/wagoodman/dive](https://github.com/wagoodman/dive)



Figure 1: How small can we get that docker container (Matthijs Brouns)

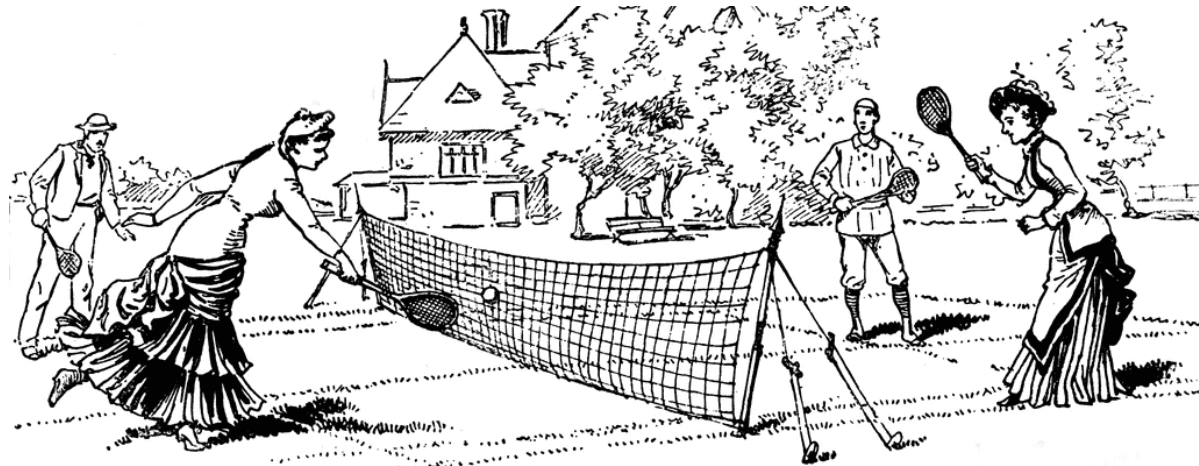
## Articles: Production-ready Docker packaging for Python developers

### Table of Contents

- [The basics of Docker packaging](#)
- [Best practices for production](#)
  - [The broken status quo](#)
  - [Base image and dependencies](#)
  - [Security](#)
  - [Fast builds, small images](#)
  - [Conda](#)
  - [Applications and runtime](#)
  - [Packaging as a process](#)
  - [Docker variants and alternatives](#)

Figure 2: Pythonspeed.com, Itamar Trauring

# 3. Serving



# 1. Numpy's fast, use it





```
from fastapi import FastAPI
from pipeline import model,
                        clean_data,
                        format_data,
                        data_is_valid

app = FastAPI()

@app.post("/predict/")
async def predict(item):

    if not data_is_valid(item):
        return {"message": "data not valid"}

    item = clean_data(item)
    predictions = model.predict(item)
    output = format_data(predictions)

    return output
```

- Validate data
- Cleaning and formatting
- Making a prediction
- Formatting the result
- Returning the result



```
from fastapi import FastAPI
from typing import List
from pipeline import model,
                        clean_data,
                        format_data,
                        data_is_valid

app = FastAPI()

@app.post("/batch-predict/")
async def predict(items: List[str]):

    items = list(set(items)) # ← remove duplicates

    items = [i for i in items
              if data_is_valid(i) == True] # ← leverage list comprehensions

    items = clean_data(items) # ← Numpy or Pandas
    predictions = model.predict(items) # ← faster than calling predict N times
    outputs = format_data(predictions)

    return outputs
```

- Much faster
- Better for the user

## 2. But can we handle the load?

**More containers?**


**It's Docker**

**Add RAM!**

**Sync or async?**

```
from locust import HttpUser, task

class TestAPP(HttpUser):
    @task
    def run_test(self):
        self.client.post(
            "/basic",
            headers={"token": "XXXXXXXX"},
            json={
                "field_1": "Lisbon, Portugal",
                "field_2": "Copenhagen, Denmark",
                "field_3": "Ancona, Italy",
            },
        )
```

 LOCUST

HOST  
https://your-app.com

STATUS  
READY

WORKERS  
75


### Start new load test

Number of total users to simulate

Spawn rate (users spawned/second)

Host (e.g. http://www.example.com)

Start swarming

 LOCUST

HOST  
http://api.initech.com

STATUS  
RUNNING  
21400 users  
[Edit](#)

WORKERS  
6

RPS  
228.1

FAILURES  
0%

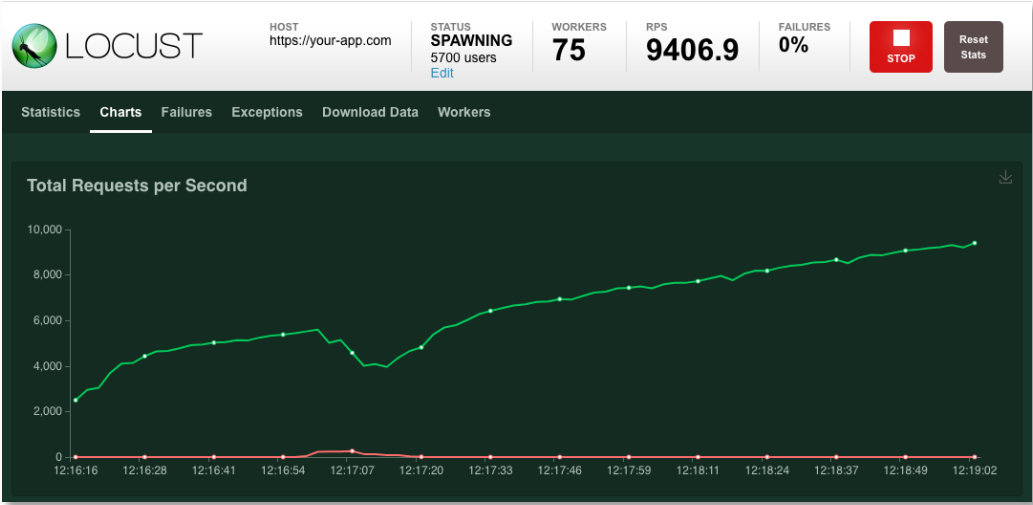
STOP

Reset Stats

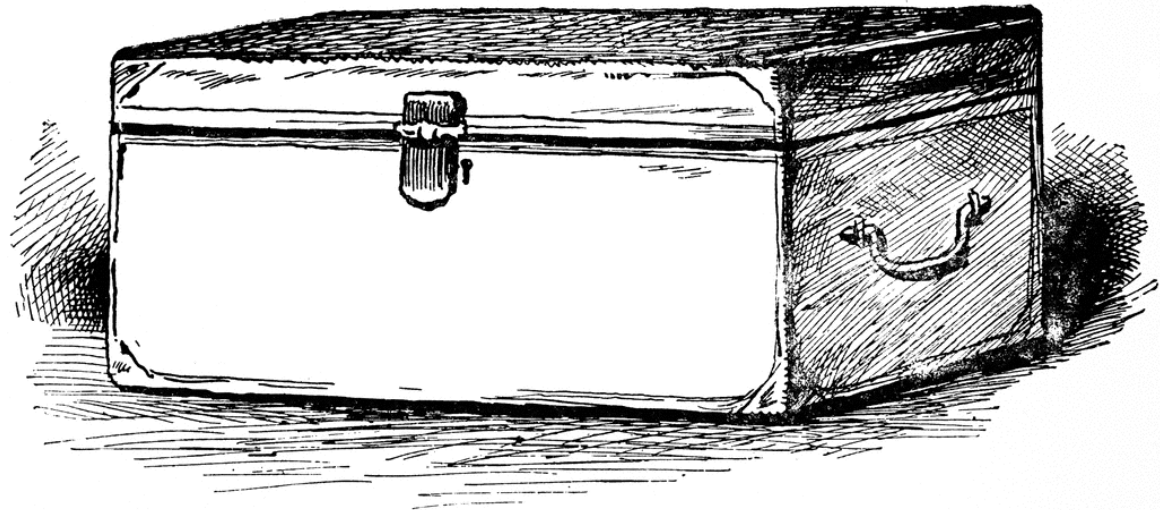
Statistics

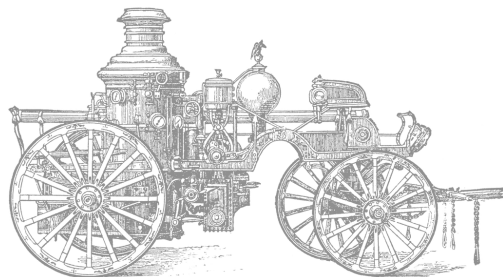
ChartsFailuresExceptionsCurrent ratioDownload DataWorkers

Type	Name	# Requests	# Fails	Median (ms)	90%ile (ms)	99%ile (ms)	Average (ms)	Min (ms)	Max (ms)	Average size (bytes)	Current RPS	Current Failures/s
GET	/	3858	0	21	35	38	21	4	38	20170	40.1	0
GET	/blog	1279	0	25	45	49	26	3	49	20083	14.6	0
GET	/blog/[post-slug]	1258	0	14	25	27	14	2	27	20177	13.1	0
POST	/groups/create	134	0	55	100	110	58	5	109	3273	1.3	0
GET	/signin	7823	0	26	45	49	26	3	49	19969	66.3	0
POST	/signin	7823	0	83	110	120	83	45	120	20021	66.3	0
GET	/users/[username]	1267	0	30	51	55	30	6	55	19920	13	0
POST	/users/[username]	128	0	67	110	120	68	13	120	11177	1.1	0
GET	/v1/users/	1325	0	26	45	49	26	3	49	20128	12.3	0
Aggregated		24895	0	34	97	120	43	2	120	19904	228.1	0



# So what?



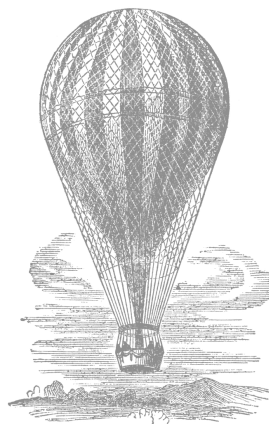


## Make sure your prototype is battle ready

Squeeze the performance juice from Python

Learn concurrency and multi-threading

Don't block your user

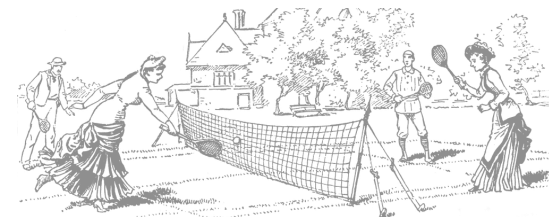


## Faster deployments, faster development

Choose the right cloud service for your needs

Know how to improve your containers

Faster builds, more deployments



## When does our app blow up?

Leverage Numpy and Pandas

Get a batch endpoint running, early

Pressure test your APIs to ensure robustness

# Grazie 🇮🇹