

Python for Distributed Systems

PyData Madrid. April 8th 2016

Guillem Borrell i Nogueras

<https://github.com/guillemborrell/pydata-madrid>

Me

- Aerospace Engineer. PhD.
- I worked with supercomputers.
- I dealt with lots of data.
- In some cases I actually built the supercomputers I used.
- I build high-performance distributed tools for a living.
- @guillemborrell
- <http://guillemborrell.es>

High performance distributed systems





Infrastructure

NoSQL Databases

FOUNDATION DB, DATASTAX, mongoDB, COUCHBASE, KEROPIKE, boshio, HYPERTABLE, apex, CLOUDANT, OpenData, Neo4j, series

Big Data

HADOOP On Prem, HADAPT, cloudera, Zettaset, amazon, MAPR, Microsoft, Pivotal, Hortonworks, MORTAR, Infochimps, Qubole, alfiscale, amazon

NewSQL Databases

MarkLogic, TRANSATLITE, RAIN, paradigm, montagi, deep-db, skySQL, NUODB, citusdata, Clustrix VoltDB, SOLPine

MPP Databases

TERADATA, ParticScreen, InfiniDB, Kognitio, N, NETEZZA, VERTICA, SQL Server, Pivotal, PARACCEL

Graph Databases

Neo4j, aster data, InfiniteGraph

Crowd-sourcing

microtask, CROWD COMPUTING, SERVO, mobileworks, amazon

Data

TRIFACTA, Pakata, dataTamer, KALIDO, everlytix, TRANSFORM, IRON, syncsift

Security

DATAGUISH, Stormpath, IMPERA

Storage

Cleversafe, Panosys, e, alekster, Compusera

App Dev

CONJUGITY, CONCURRENT, wibedata

Analytics

Analytics Platforms

Databricks, QuantCast, PERSASIVE, QUAVUS, Datasphere, KARMA, FRECOLO, dataSpora

For Business Analysts

STAT WING, CIRRO, TREPAREL, OrigamiLogic, ClearStory, DataGravity

Data Science Platforms

domino, Alpine, Sense, MORTAR, T tools, COHERENT, plotly, yhat, MOORE

Unstructured Data

BASIS, ATTIVO, GENERAL SENTIMENT, semantrio, CRIMSON HEXAGON, DIGITAL EXPLORING, Quid, Palantir

BI Platforms

bime, pentaho, GoodData, StSense, platforma

Machine Learning

SKYTREE, bigmi, TITANUM, wise.io, context relevant

Location/People/Events

RADIUS, FLIPLOP, LOCUS, PlaceIQ

Big Data Search

hp, LOCKWORKS, ONTOLOGY

Crowd-Sourced

kaggle, DataKind

Real-Time

METAMARKETS, omio, causata

Statistical Computing

REVOLUTION, SSAS, MATLAB

Log Analytics

splunk, loggly, sumologic, Kibana

Analysis

THINK BIG, VALUANCE, DATA SCIENCE, MU SIGMA

Statistical Computing

ssas, REVOLUTION, MATLAB

Log Analytics

splunk, loggly, sumologic, Kibana

Applications

Ad Optimization

aggregate knowledge, rocketfuel, TAPAD, ai Match, thetradebook, 33 SCORE

Publisher Tools

Chartbeat, Yieldex, yieldbot

Marketing

LATTICE ENGINES, Sailthru, spinnaker, gainsight, Kontera, Q RelateIQ, Telaport, persado, bloomreach, CLICKFOX, Pursway

Finance

BillGuard, LendUp, KENSHO, OnDeck

Human Capital

evolv, centelo, gild, JUDICATA, RAVEL, Lex Machina

Government / Regulation

mark43, enigma, FortScale, feedzai

Educational Learning

KNEWTON, geclara, PANORAMA, Clever

Industries

Recombine, tubular, NEXT BIG SOUND, OPOWER, NIGHT MACHINE, THE CLIMATE CORPORATION, FLATIRON, COURSEIFY

Health

23andMe, Ginger.io, GENE

Cross Infrastructure /

SAP, ssas, IBM, Google, Microsoft, vmware, amazon, 1010data, talend, TERAdata, hp, NetApp

Open Source

Framework

Spark, Hadoop, YARN, HDFS

Query / Data Flow

Cassandra, SciDB, ORACLE, HBASE, mongoDB, riak, Sqoop

Data Access

Coordin -ation / Work-flow, ZooKeeper, Real-Time, Storm

Stat Tools

BoPly, (p)

Machine Learning

Cloud Deploy

Search

elasticsearch, Solr, LUCENE

Data Sources

Data Mkts

Windows Azure, knoema, DataMarket, factual

Data Sources

DATA.GOV, premise, YODLEE, xignite, plaid, quandl, STANDARD TREASURY, human/apl

Sensor Data

kinsa, SKYWATCH, STREETLINE, fitbit, RunKeeper, BASIS, Jawbone, LUMASENSE, Withings

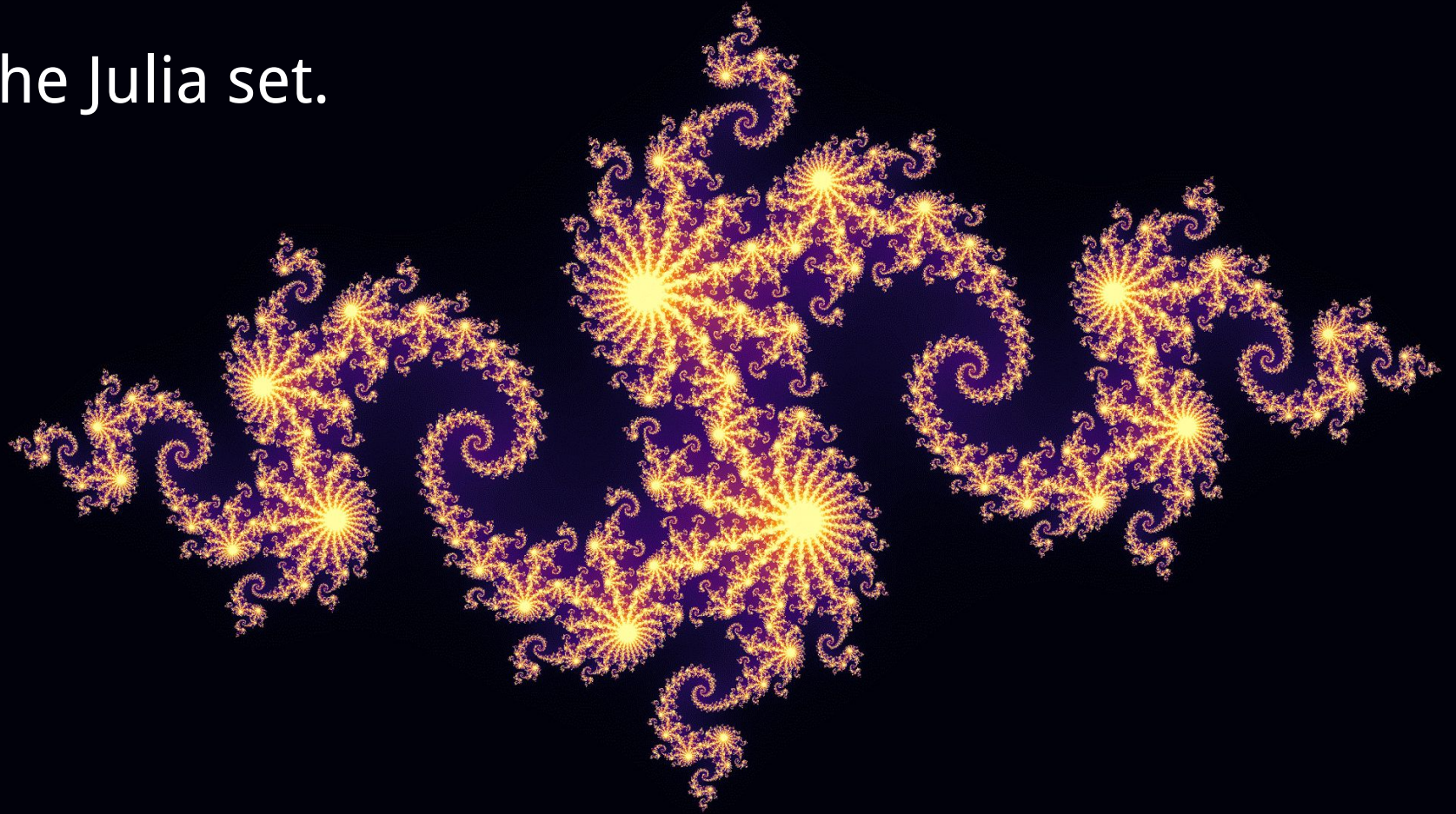
Incubators & School

zipfian, CA, INSIGHT, DataLine

Let's start from ground up.
And let's be ambitious.

The Unicorn idea.

The Julia set.





Pierre Fatou
1878 - 1929



Gaston Julia
1893 - 1978

$$z_{n+1} = z_n^2 + c$$





Let's make a Julia set.

Now we have an application

- Something that is useful, but takes a good amount of resources. Like CPU/IO intensive stuff.
- But applications are for cellphones and pirates!
- Services = \$, €, ¥...
- We need a name, an elevator pitch, and a REST API to hire a frontend developer.
- We will make something that works in the Angel round.

Fractal.ly

Find beauty in complexity...
Web scale!

The REST API

Is it scalable?
Is it responsive?
Is it efficient?

Let's make it a little more efficient.

Now we visit a business angel.
And we got money to rent servers!

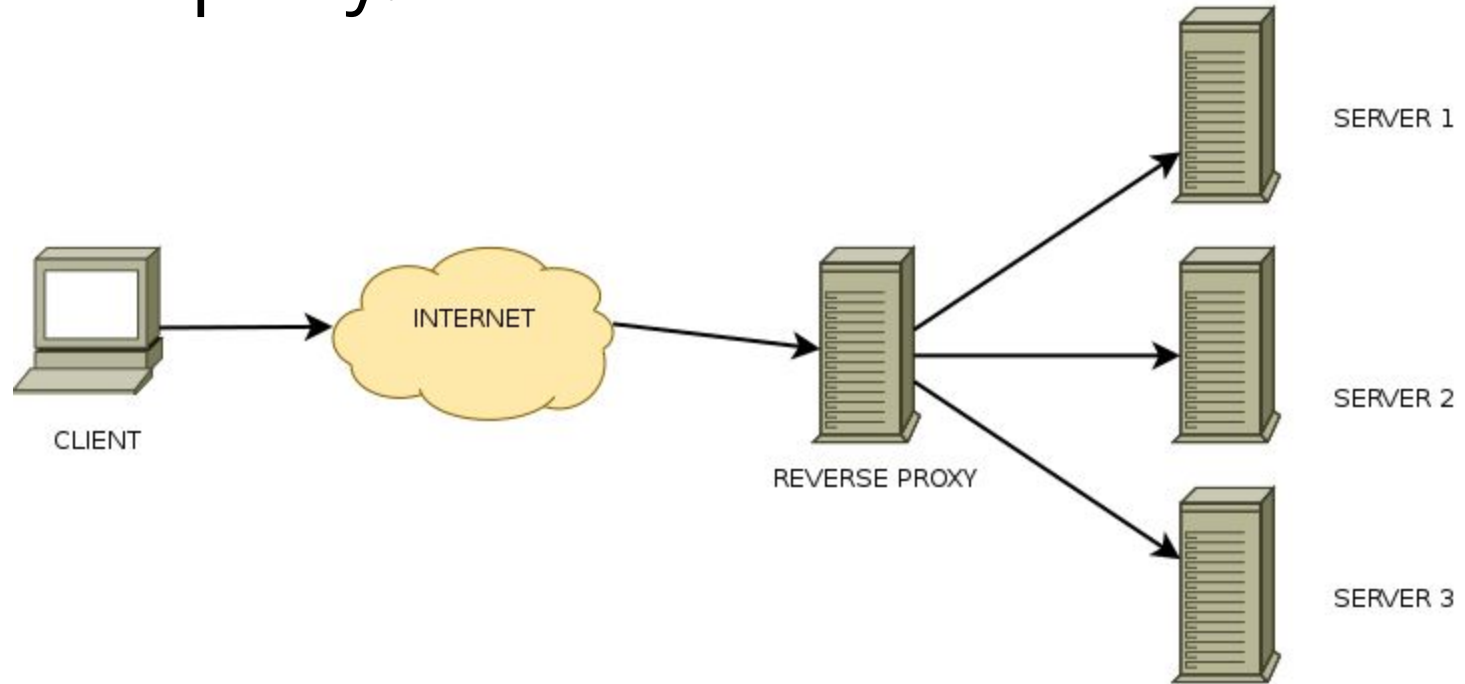
Let's make it parallel.

Parallelization is about domain decomposition

- Split the problem in separated portions.
- The portions should share the minimum amount of information. (CAP)
- Exploit low-hanging fruits.

It is fine to deal with each request in parallel. We have to be pragmatic, it's not about publishing a CS paper.

A reverse proxy.



Is it scalable?
Is it responsive?
Is it efficient?

We have our first users. They complain
because the application is not
responsive enough.

We hire a frontend developer.
And we make the application
asynchronous.

Polling is a good way to deal with
asynchronicity.

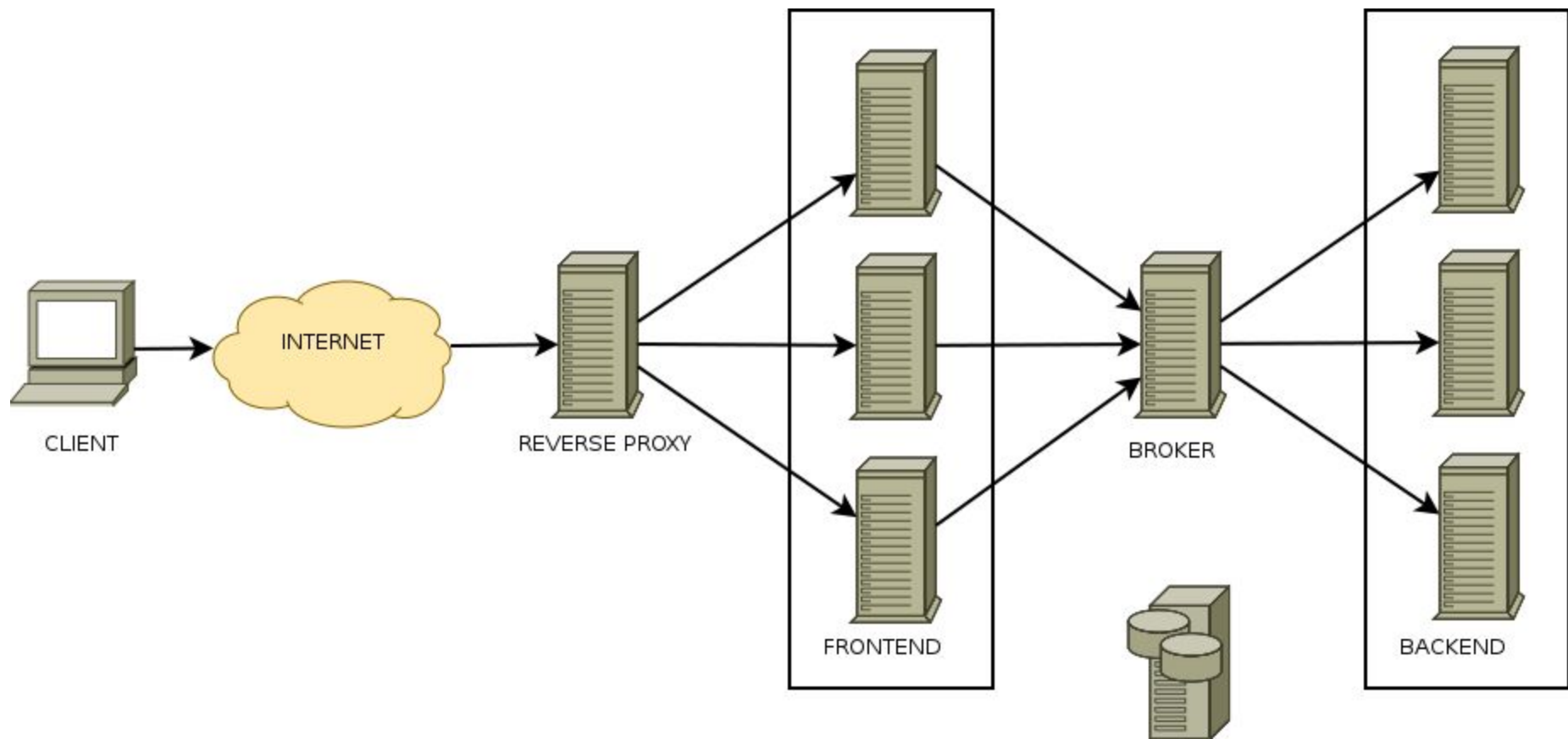
But the application is not as responsive
as we expected





The GIL is there because the alternatives
are quite frightening.

Split I/O intensive process from CPU
intensive processes.



We need message passing.

A message queue

~~Ø~~MQ

A serialization format

The logo for JSON (JavaScript Object Notation) is displayed. It features the characters '{', 'J', 'S', 'O', 'N', and '}' in a blue, sans-serif font. The letter 'O' is replaced by a 3D sphere with a black-to-white gradient and a shadow underneath, giving it a three-dimensional appearance. The entire logo is centered on a white background.

{JSON}

Is it scalable?
Is it responsive?
Is it efficient?

The logo for the Julia programming language. It features the word "julia" in a bold, dark grey, sans-serif font. The letter "j" has a blue circle above it. The letter "i" has a red circle above it. The letter "l" has a green circle above it. The letter "a" has a purple circle above it. The circles are arranged in a slightly overlapping, triangular pattern above the letters.

julia

We now have a parallel, distributed,
multi-threaded, multi-language,
responsive and efficient service to
obtain julia sets **WEB SCALE!**



Questions?