Screaming Fast API Clients

Moshe Zadka – https://cobordism.com

2020

Acknowledgement of Country

San Francisco Bay Area Peninsula
Ancestral homeland of the Ramaytush Ohlone

Latency is the Site Killer

Every 100ms of latency in your site lose more customers

(Micro)service Architecture

Layers

(Micro)service Architecture

Fan-out

Lognormal Black Swans

- Lognormal: $1/x$ (kinda)
- Normal: e^{-x^2}

Averages Lie

Only good for normal distributions

Your Backend is Slow

Lognormal, not normal

Multiplicity Magnifies Outliers

With 5 queries:

- P90 becomes P50
- P99 becomes P90

Measure

Histograms, not averages

Measure

All layers

Let's Write Some Code

```
@app.route('/')
def hello_world():
    all_values = sum(
        CLIENT.get(URL).json()["value"]
        for x in range(FANOUT)
    )
    return json.dumps(dict(total=all_values))
```

Let's Write Some Code

```
@app.route('/')
async def hello_world(request):
    all_values = await defer.gatherResults([
        CLIENT.get(URL).addCallback(treq.json_conte
        for x in range(FANOUT)
    ])
    total = sum(res["value"] for res in all_values)
    return f'Total {total}'
```

Let's Simulate

With fanout of 10:

- P50: each: 0.04 seq: 0.82 par 0.3
- P90: each: 0.23 seq: 1.8 par 0.98
- P99: each: 1.04 seq: 4.33 par 3.05

Timing Out and Retry

Temporary slow-downs

Let’s Write Some Code

```
def get_with_timeout(url):
    def try(_ign=None):
        return CLIENT.get(URL).addCallback(treq.jso
    d = try()
    d.addTimeout(0.1)
    d.addErrback(try)
    return d
```

Let's Simulate

- P50: 0.18
- P90: 0.51
- P99: 1.66

Let's Simulate

Retried requests: 25

Let's Write Some Code

```
def get_with_timeout(url):
    def try(_ign=None):
        return CLIENT.get(URL).addCallback(treq.jso
    d = try()
    d.addTimeout(0.1)
    d.addErrback(try)
    d.addTimeout(0.4)
    return d
```

Let’s Simulate

- P50: 0.19
- P90: 0.53
- P99: 0.6

Summary

- Latency
- Backend latency
- SLA
- Measurement
- Simulation