

Crushing Latency with Vert.x

Paulo Lopes 

Principal Software Engineer

 @pml0pes

 <https://www.linkedin.com/in/pmlopes/>

 pmlopes



收获国内外一线大厂实践 与技术大咖同行成长

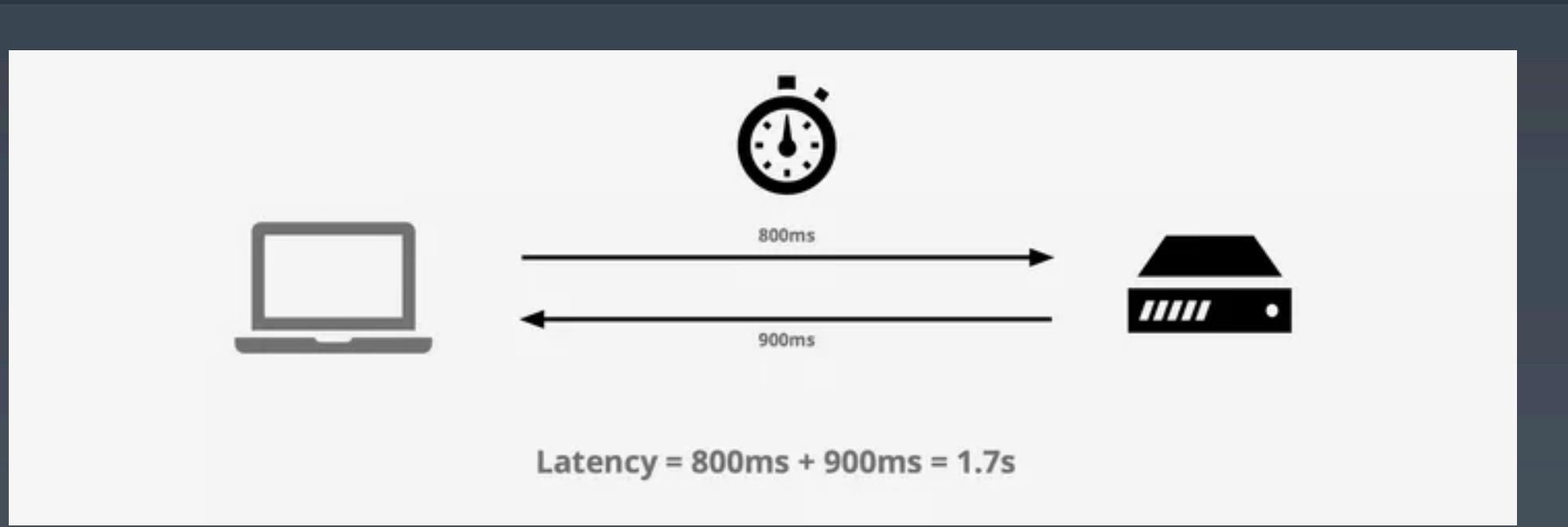
⌚ 演讲视频 ⌚ 干货整理 ⌚ 大咖采访 ⌚ 行业趋势





latency noun

/la·ten·cy | \ 'lā-tən(t)-sf \



Network latency is the term used to indicate any kind of delay that happens in data communication over a network.

(techopedia.com)

Latency by the numbers

- **Amazon:** every 100ms of latency costs 1% in sales
<http://home.blarg.net/~glinden/StanfordDataMining.2006-11-29.ppt>
- **Google:** an extra 0.5 seconds in search page generation time dropped traffic by 20%
<http://glinden.blogspot.com/2006/11/marissa-mayer-at-web-20.html>
- **A broker:** could lose \$4 million in revenues per millisecond if their electronic trading platform is 5 milliseconds behind the competition
<http://www.tabbgroup.com/PublicationDetail.aspx?PublicationID=346>

Latency is not the problem
it's the symptom!

2007: Dan Pritchett

- Loosely Couple Components
- Use Asynchronous Interfaces
- Horizontally Scale from the Start
- Create an Active/Active Architecture
- Use a **BASE** instead of **ACID** Shared Storage Mode

www.infoq.com/articles/pritchett-latency

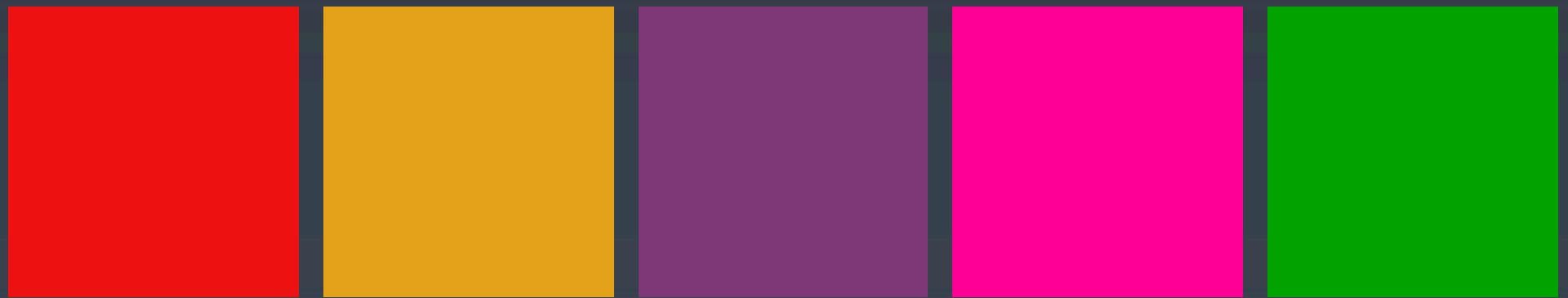
2011 (Tim Fox): Vert.x

- Loosely Couple Components (event bus)
- Use Asynchronous Interfaces (non blocking I/O)
- Horizontally Scale from the Start (clustered)

Eclipse Vert.x is a tool-kit for building reactive applications on the JVM. <https://vertx.io/>

Why Non-Blocking I/O?

5ms / req time



```
# In optimal circumstances  
  
1 Thread => 200 req/sec  
8 Cores => 1600 req/sec
```

req time grows as threads fight for execution time



```
# PROCESS STATE CODES
#   D      Uninterruptible sleep (usually IO)

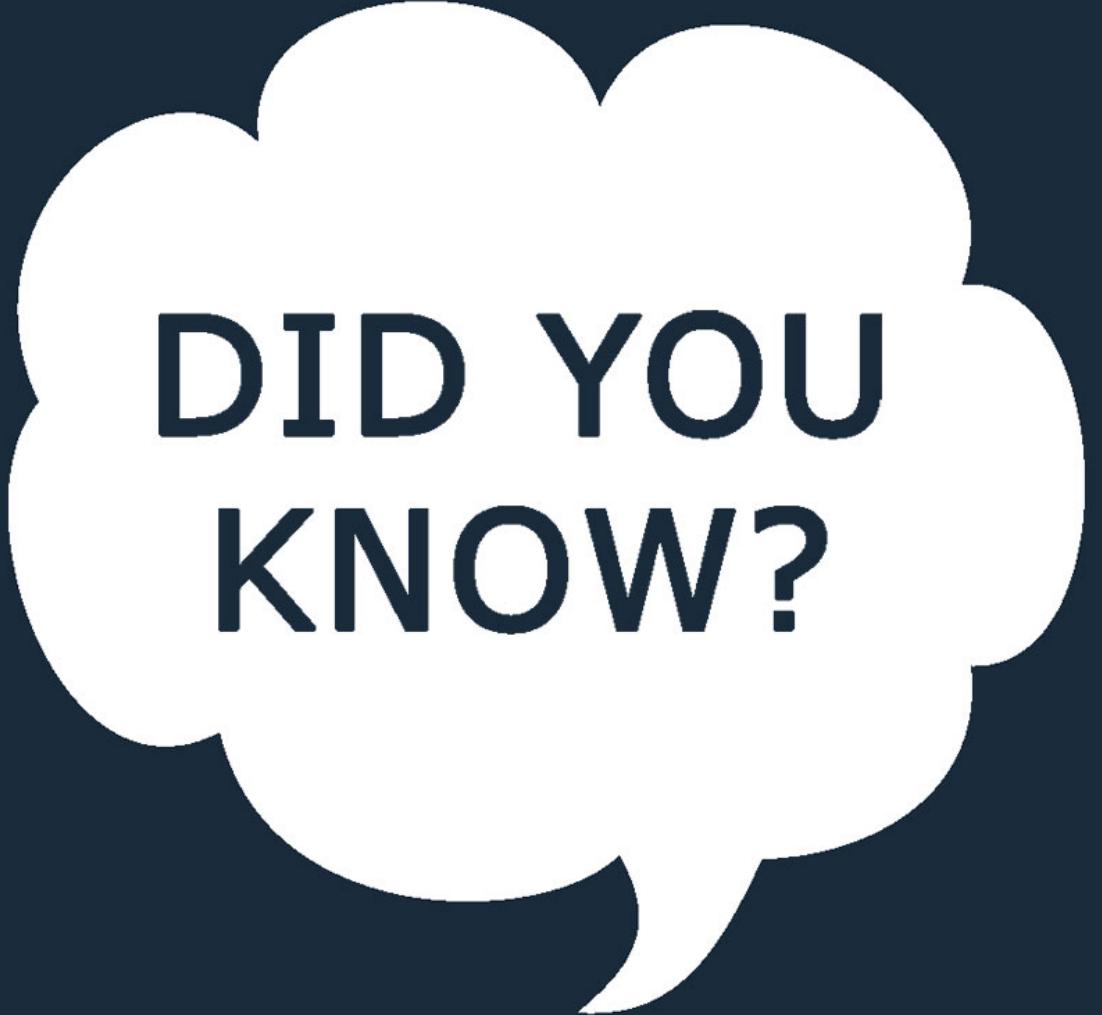
ps aux | awk '$8 ~ /D/ { print $0 }'
root 9324 0.0 0.0 8316 436 ? D< Oct15 0:00 /usr/bin/java...
```

when load is higher than max threads queuing builds up



...

```
# git@github.com:tsuna/contextswitch.git  
  
./cpubench.sh  
model name : Intel(R) Core(TM) i7-8650U CPU @ 1.90GHz  
1 physical CPUs, 4 cores/CPU, 2 hardware threads/core  
2000000 thread context switches in 2231974869ns  
                                (1116.0ns/ctxsw)
```



DID YOU
KNOW?

```
grep 'CONFIG_HZ=' /boot/config-$(uname -r)  
# CONFIG_HZ=1000
```

Practical example:

Tomcat 9.0

- Default maxThreads: **200**
- Avg req time: **5ms**
- Hypothetical High load: **1000 req**
- Wasted wait/queue time: $(1000 / 200 - 1) * 5 = \text{0~20ms}$

<https://tomcat.apache.org/tomcat-9.0-doc/config/executor.html>

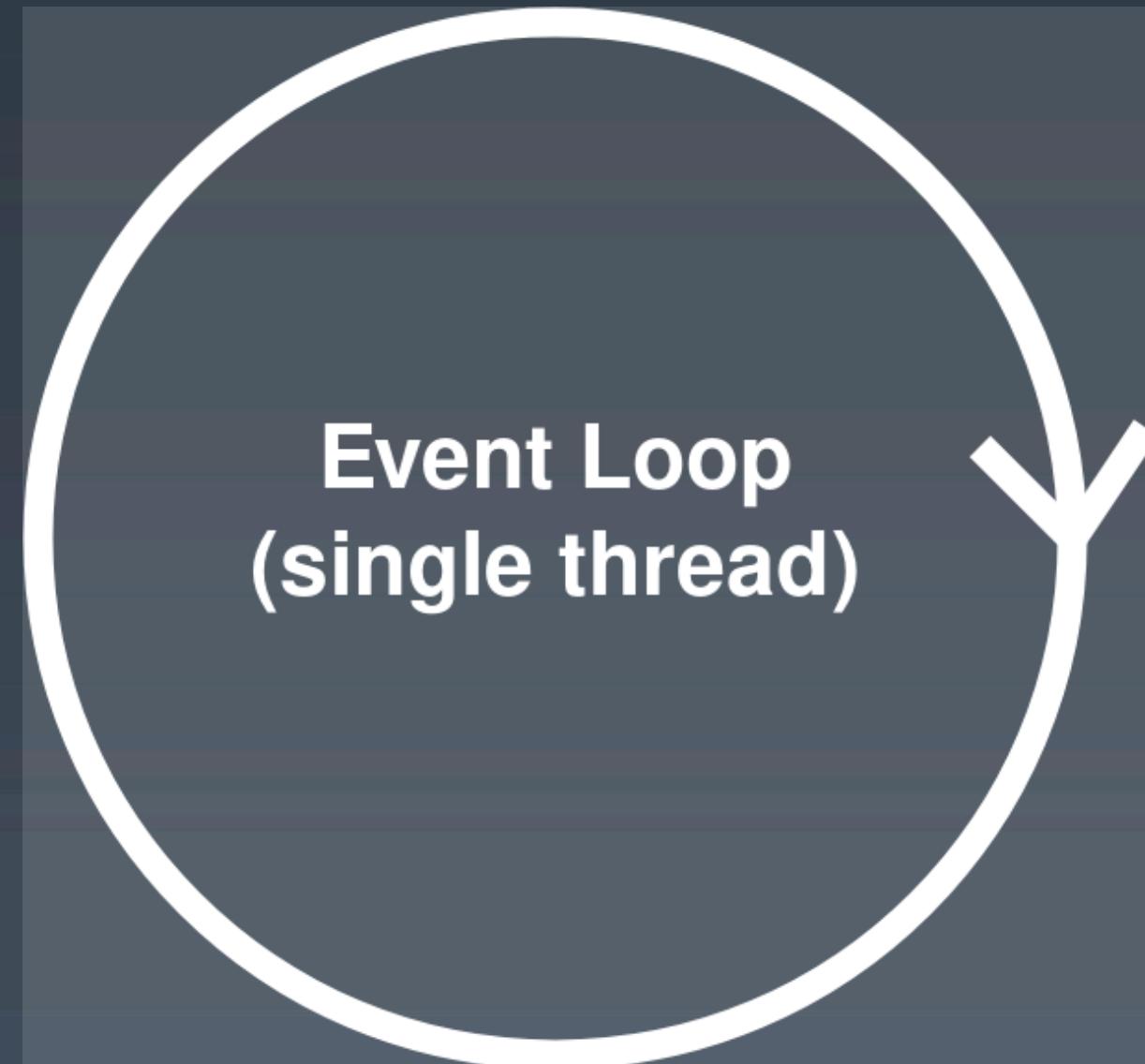
at max utilization

CPU is mostly waiting

Non-Blocking I/O

Vert.x

Events



Request handler

AUTH handler

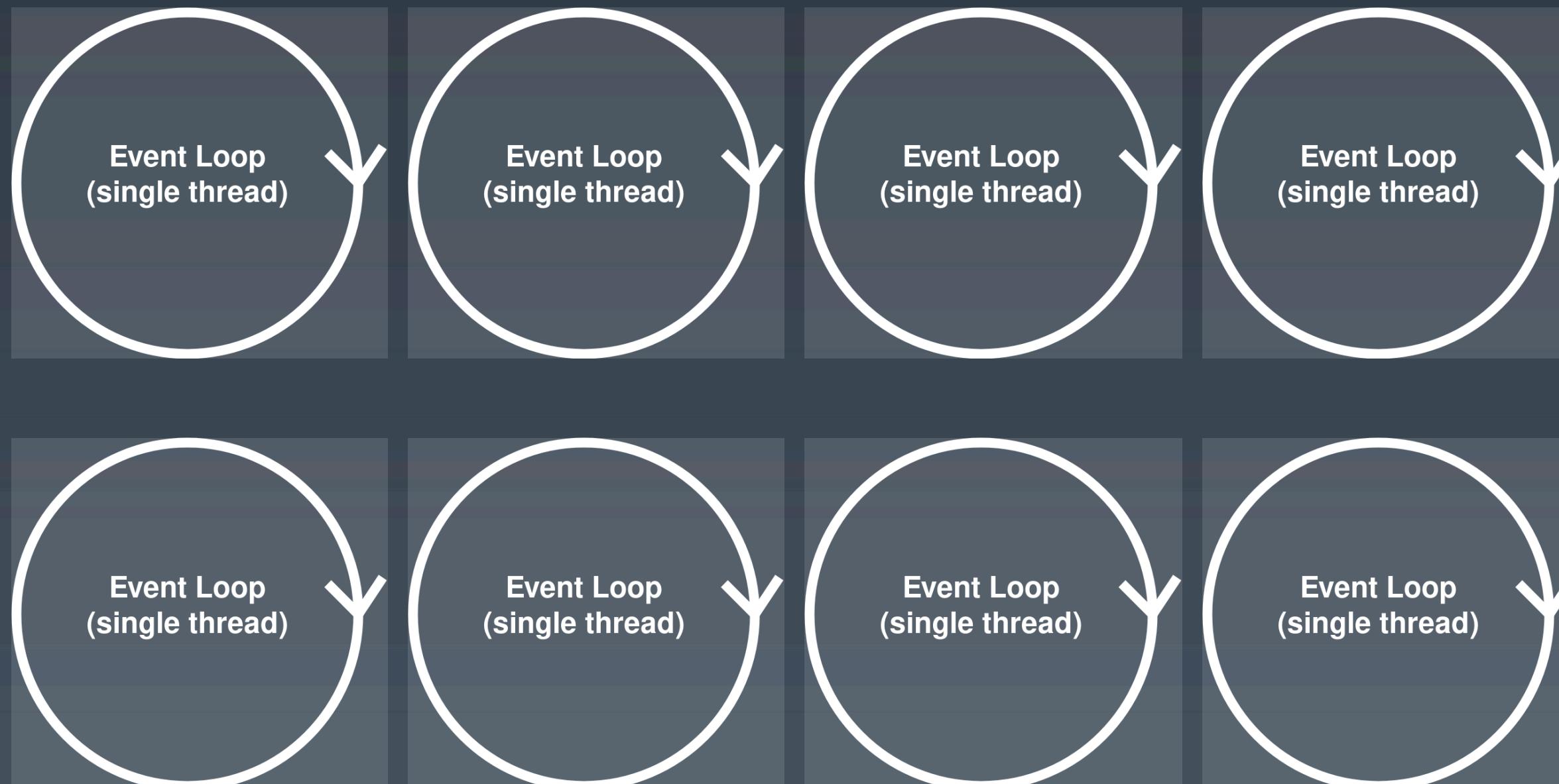
DB handler

JSON handler

1 CPU core fully used!

Vert.x

Intel(R) Core(TM) i7-8650U CPU @ 1.90GHz



100% CPU cores used!

Benchmarking is hard

- *Meaningful benchmarks are even harder*
- TechEmpower Framework Benchmarks
 - Contributors: **528**
 - Pull Requests: **4022**
 - Commits: **11095**

<https://github.com/TechEmpower/FrameworkBenchmarks>

Baseline: JAX-RS

- Blocking API
- Thread Based
- Java

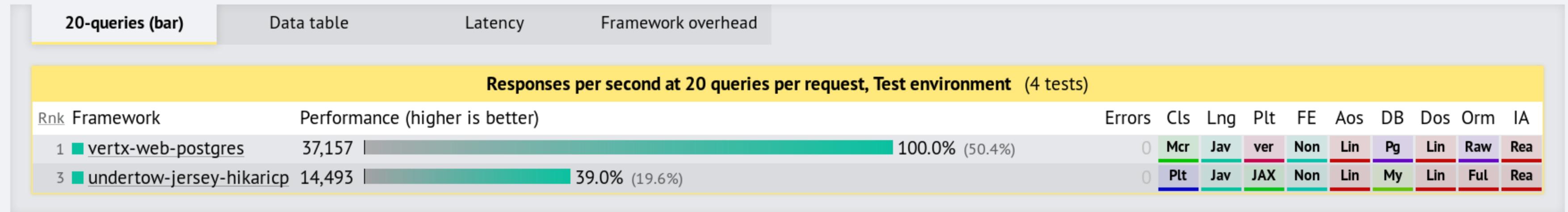
jax-rs

```
@GET  
@Path("/queries")  
World[]  
queries(@QueryParam("queries") String queries)  
{  
    World[] worlds = new World[queries];  
    Session session = emf.createEntityManager();  
  
    for (int i = 0; i < queries; i++) {  
        worlds[i] = session  
            .byId(World.class)  
            .load(randomWorld());  
    }  
  
    return worlds;  
}
```

vert.x

```
void  
queriesHandler(final RoutingContext ctx) {  
  
    World[] worlds = new World[getQueries(ctx)];  
    AtomicInteger cnt = new AtomicInteger();  
  
    for (int i = 0; i < getQueries(ctx); i++) {  
        db.preparedQuery(FIND_WORLD, ..., res -> {  
            final Row row = res.result()  
                .iterator()  
                .next();  
  
            worlds[cnt.incrementAndGet()] =  
                new World(row);  
  
            if (cnt.get() == queries) {  
                ctx.response()  
                    .end(Json.encodeToBuffer(worlds));  
            }  
        });  
    }  
}
```

Multiple queries



Requirements summary

In this test, each request is processed by fetching multiple rows from a simple database table and serializing these rows as a JSON response. The test is run multiple times: testing 1, 5, 10, 15, and 20 queries per request. All tests are run at 512 concurrency.

Example response for 10 queries:

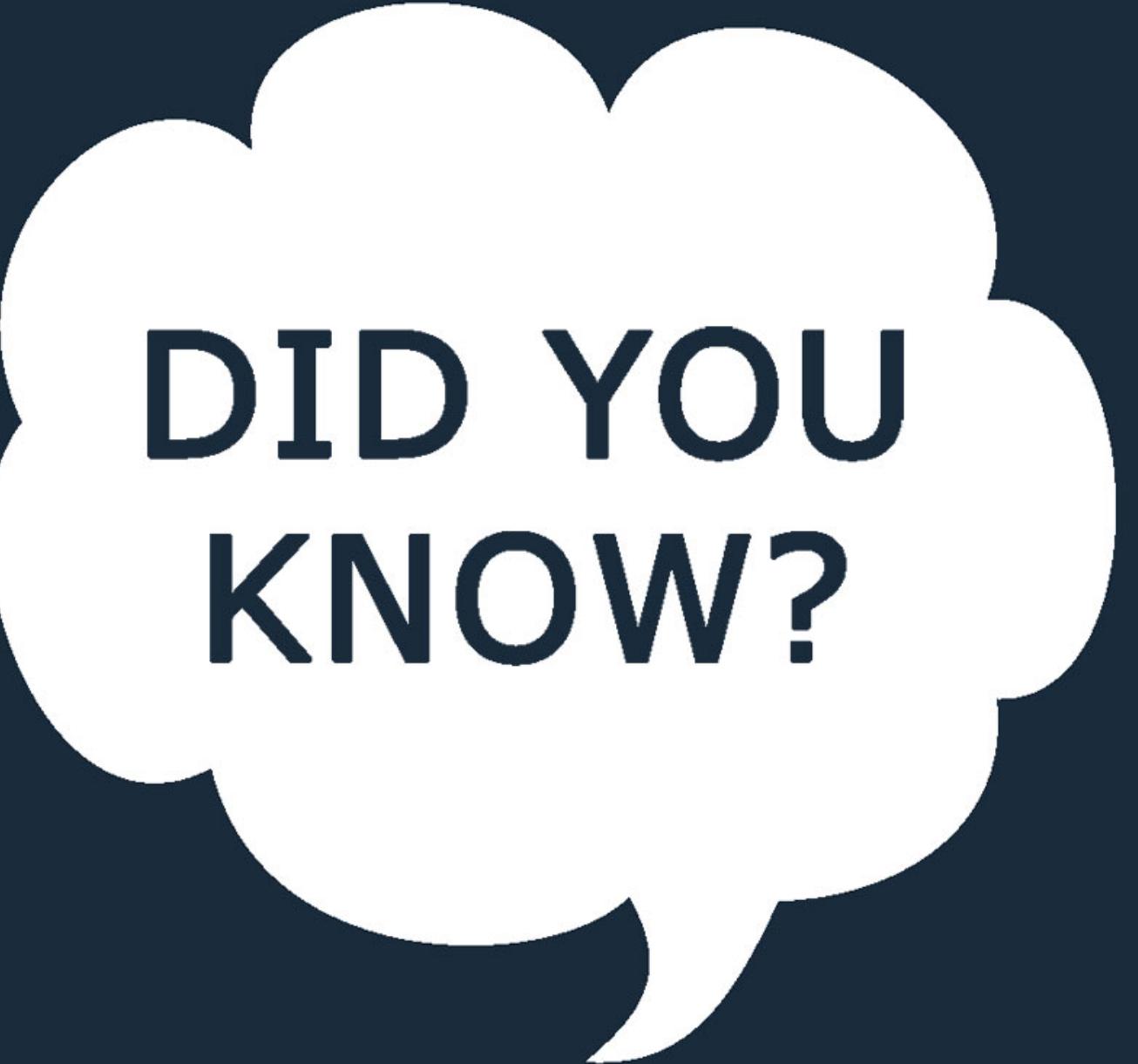
```
HTTP/1.1 200 OK
Content-Length: 315
Content-Type: application/json
Server: Example
Date: Wed, 17 Apr 2013 12:00:00 GMT
```

```
[{"id":4174,"randomNumber":331}, {"id":51,"randomNumber":6544}, {"id":4462,"randomNumber":952}, {"id":2221,"randomNumber":532}, {"id":9276,"randomNumber":3097},
 {"id":3056,"randomNumber":7293}, {"id":6964,"randomNumber":620}, {"id":675,"randomNumber":6601}, {"id":8414,"randomNumber":6569},
 {"id":2753,"randomNumber":4065}]
```

For a more detailed description of the requirements, see the [Source Code and Requirements](#) section.

Simple results

- Vert.x: **37,157** req/s
- Jax-RS: **14,493** req/s

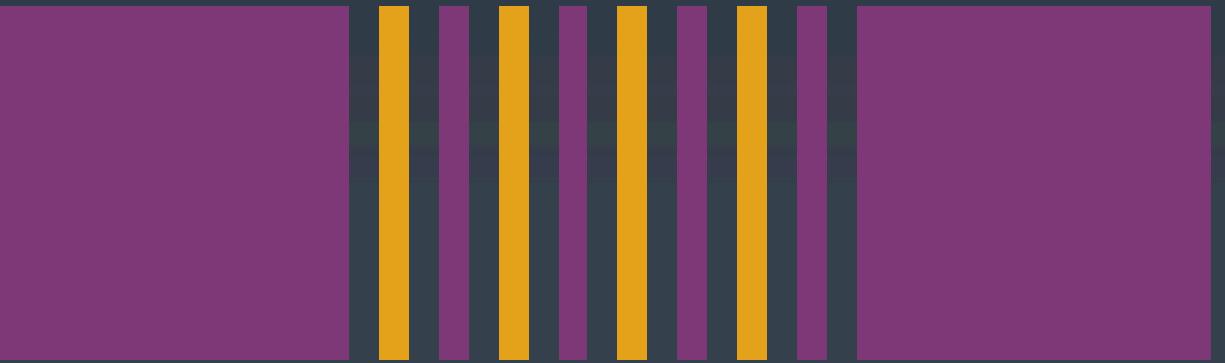


DID YOU
KNOW?

Polyglot

English - 简体中文 - Português

What happens when you say Hello?



```
function handler (context) {  
  // the exchange context  
  context  
  // get the response object  
  .response()  
  // send the message and end  
  // the response  
  .end('你好');  
}
```

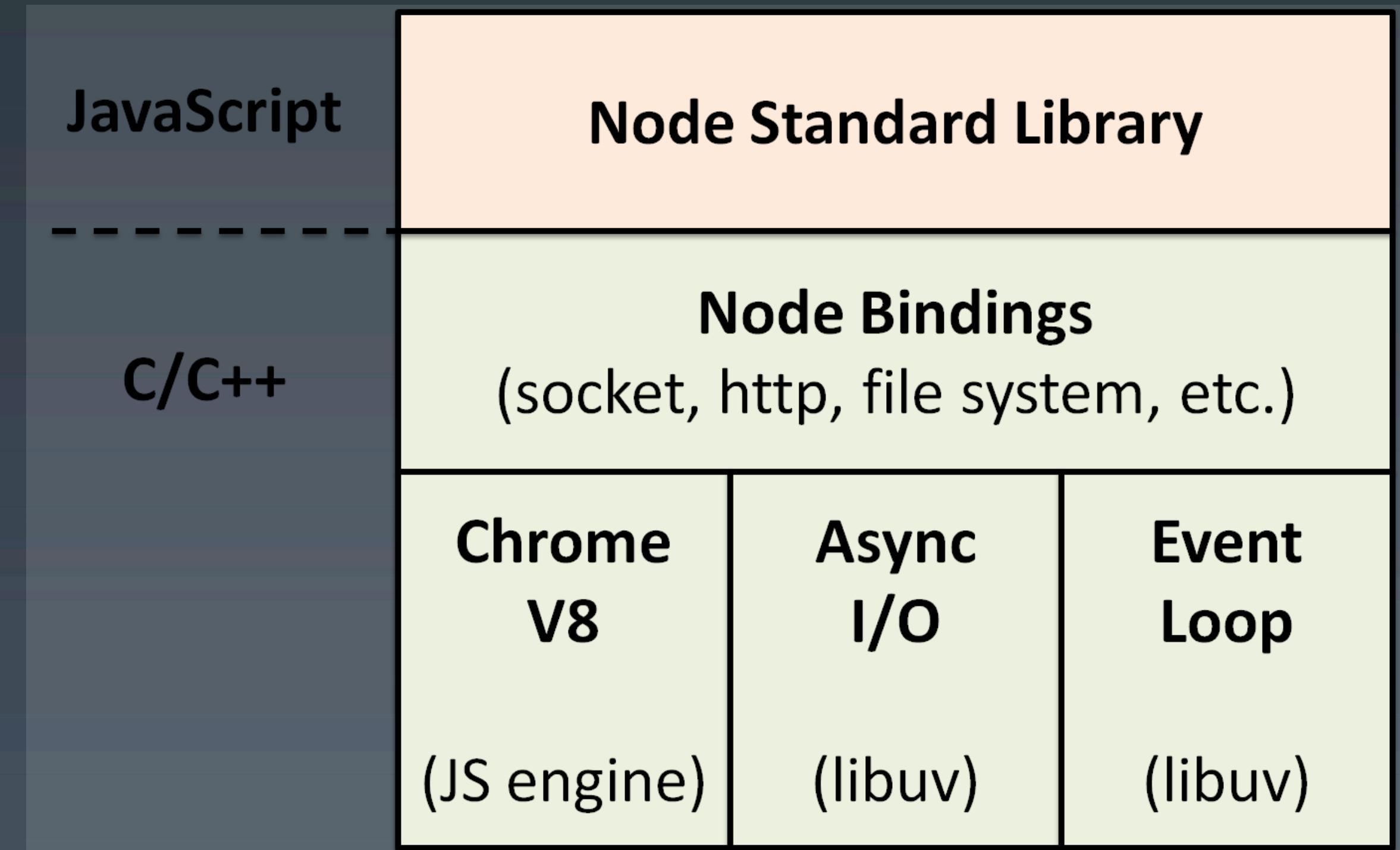
Getting the response object

```
1 this.response = function() {
2   var __args = arguments;
3   if (__args.length === 0) {
4     if (that.cachedresponse == null) {
5       that.cachedresponse = utils.convReturnVertxGen(
6         HttpServerResponse,
7         j_routingContext["response"]());
8     }
9     return that.cachedresponse;
10 } else if (typeof __super_response != 'undefined') {
11   return __super_response.apply(this, __args);
12 }
13 else throw new TypeError('invalid arguments');
14 };
```

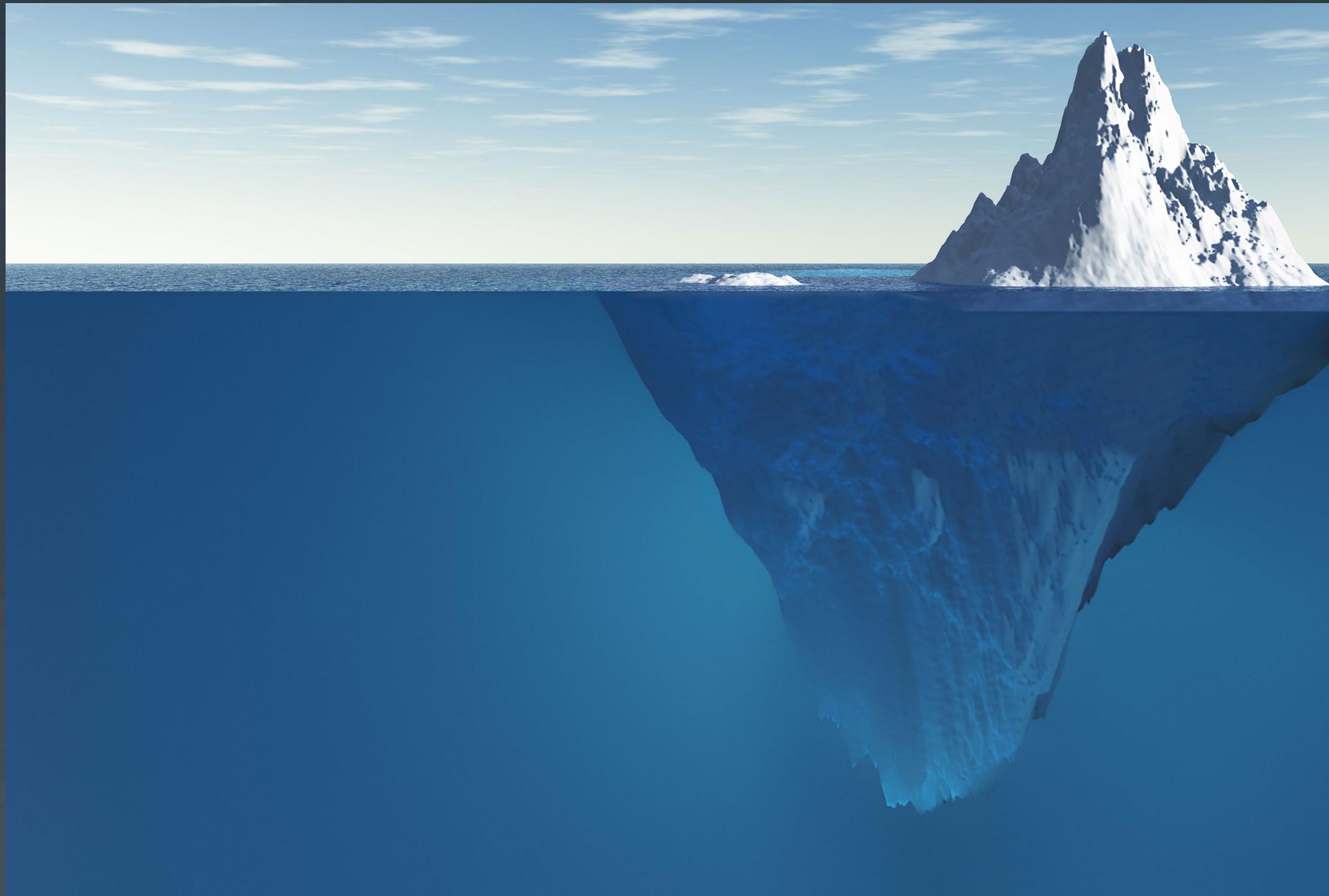
In a nutshell

- Lots of conversions (GC)
- Constant switch from JS engine to Java code
(somehow similar to context switching)
- Not suited for performance
- **JIT optimization will stop at language cross**

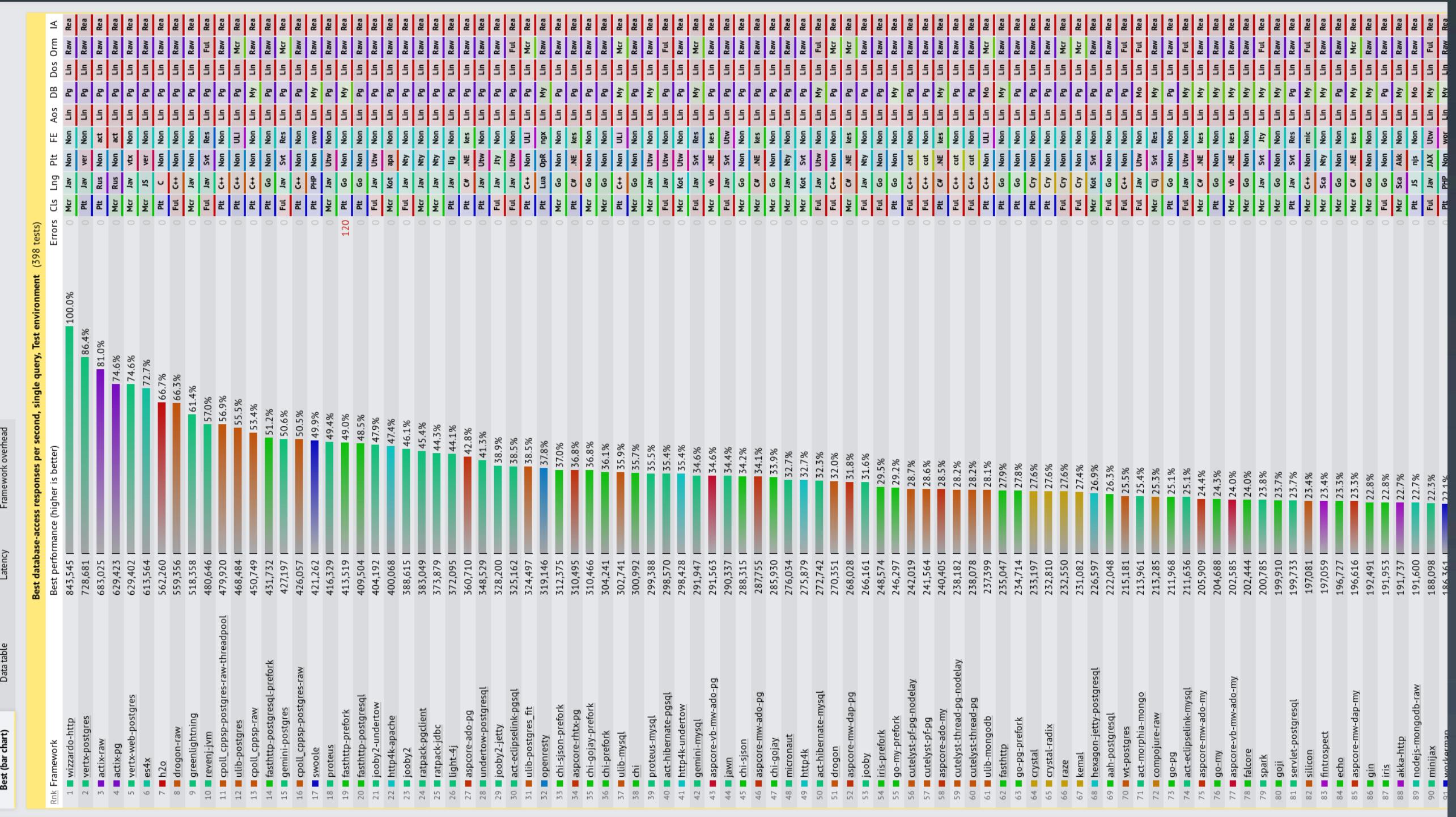
Node.js

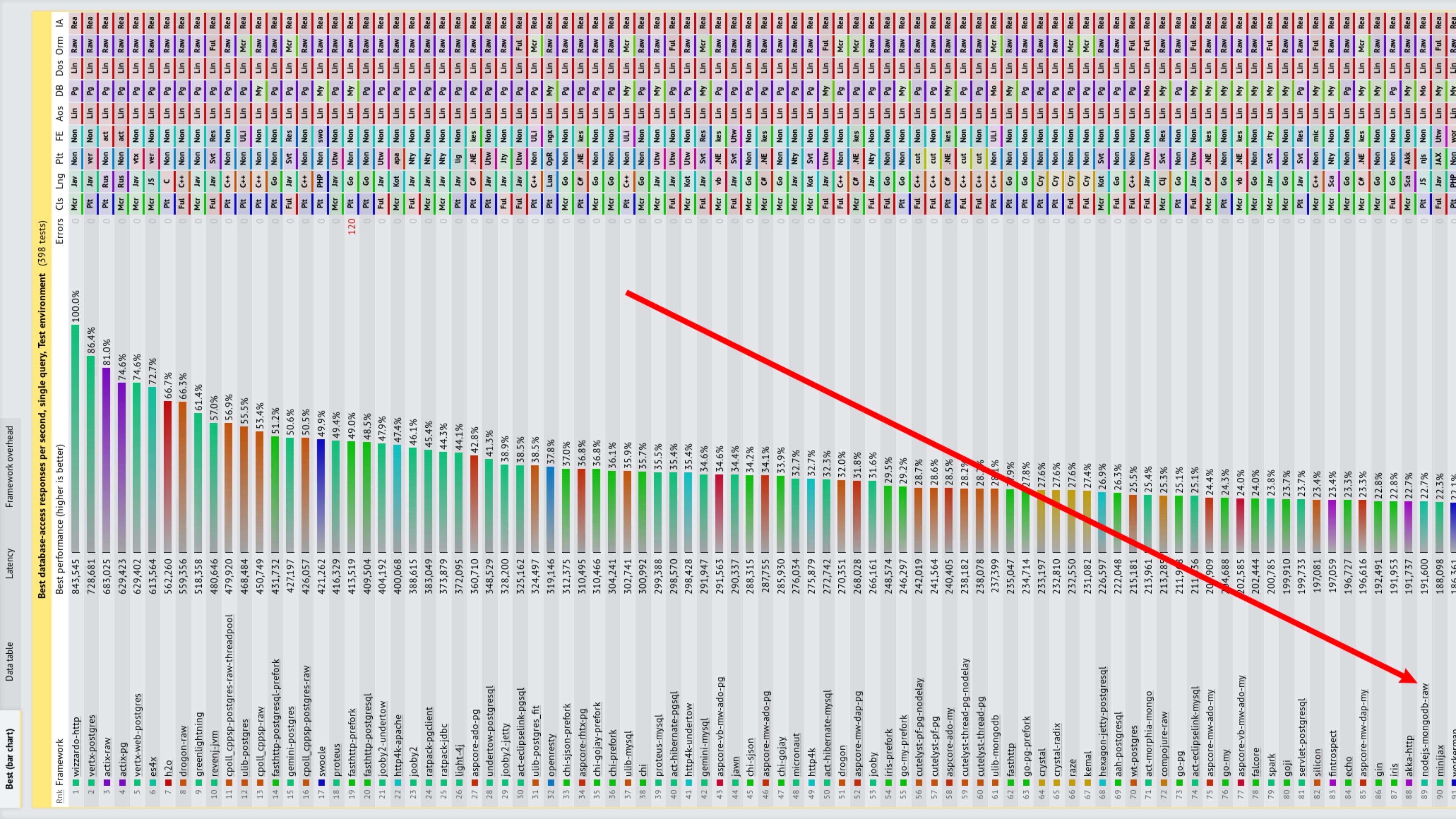


JIT can't optimize it all



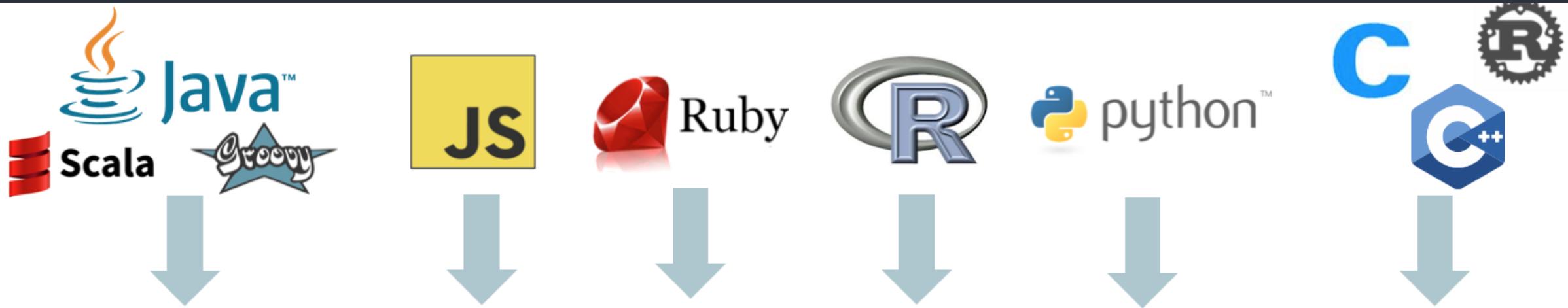
Where's Node?







Can we make polyglot
fast?



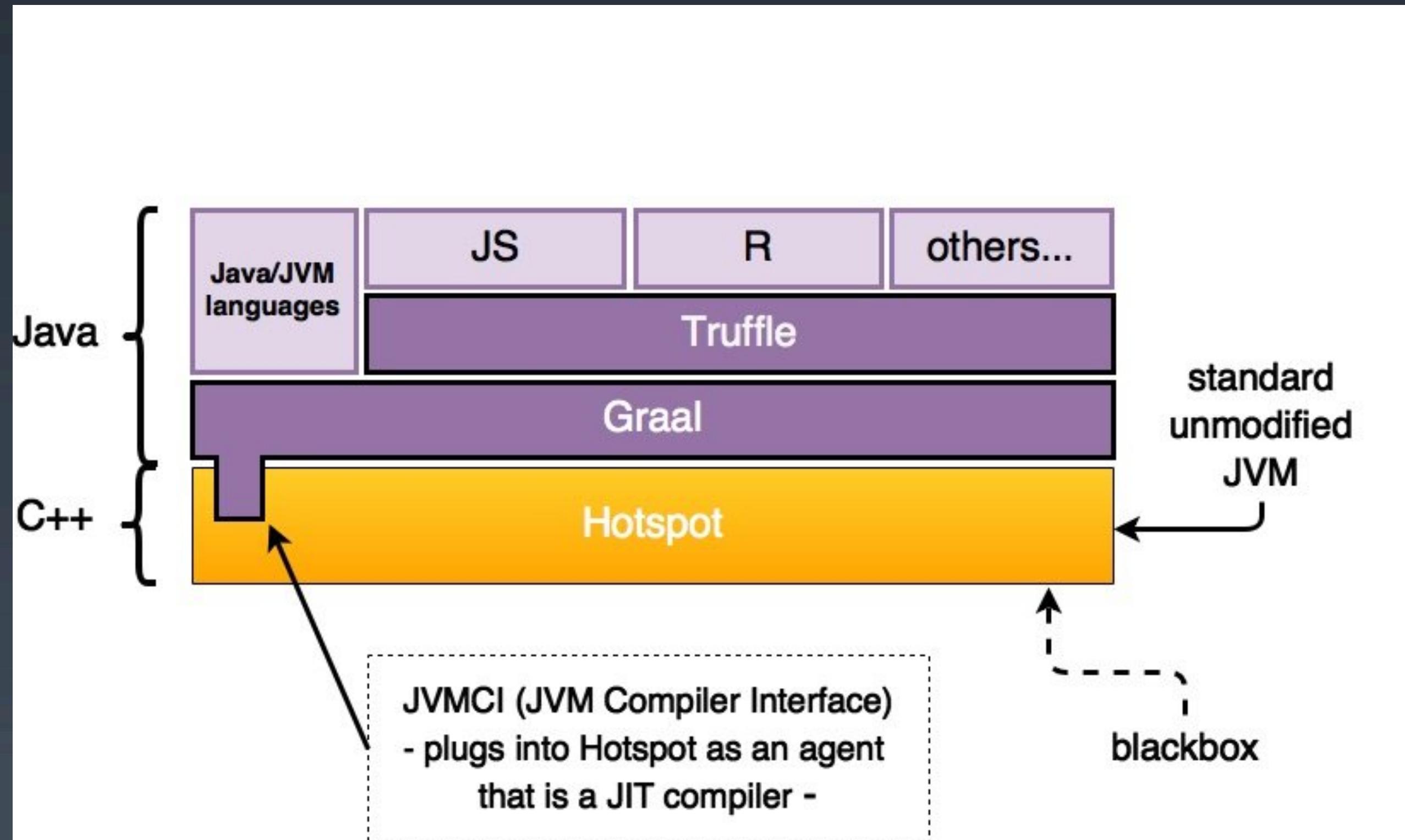
Automatic transform of interpreters to compiler

GraalVM™

Engine integration native and managed



<https://graalvm.org>



<https://graalvm.org>

GraalVM: In a nutshell

- Lots of conversions (GC)
- Constant switch from JS engine to Java code
(somehow similar to context switching)
- Not suited for performance
- JIT optimization will stop at language cross

ES4X



- GraalVM based
- Vert.x for I/O
- commonjs and ESM loader
- debug/profile chrome-devtools

<https://reactiverse.io/es4x>

ES4X design principles

- GraalJS (for fast JS runtime)
- Vert.x (for fast I/O + event loops)
- GraalVM (for full JIT)
- .d.ts (for IDE support)

| github.com/AlDanial/cloc v 1.72 T=0.09 s (401.5 files/s, 51963.8 lines/s) | | | | |
|---|-------|-------|---------|------|
| Language | files | blank | comment | code |
| Java | 26 | 389 | 683 | 1778 |
| JavaScript | 9 | 201 | 253 | 1226 |
| SUM: | 35 | 590 | 936 | 3004 |

Node.js vs ES4X

```
const cluster = require('cluster'),
numCPUs = require('os').cpus().length,
express = require('express');

if (cluster.isMaster) {
  for (let i = 0; i < numCPUs; i++)
    cluster.fork();
} else {
  const app = module.exports = express();
  app.get('/plaintext', (req, res) =>
  res
    .header('Content-Type', 'text/plain')
    .send('Hello, World!'));
}
```

```
import { Router } from '@vertx/web';

const app = Router.router(vertx);

app.get("/plaintext").handler(ctx => {
  ctx.response()
    .putHeader("Content-Type", "text/plain")
    .end('Hello, World!');
});
```

Test types

Hardware

JSON serialization

Single query

Multiple queries

Fortunes

Data updates

Plaintext

Physical

Cloud

Single query

Best (bar chart)

Data table

Latency

Framework overhead

Best database-access responses per second, single query, Test environment (398 tests)

| Rnk | Framework | Best performance (higher is better) | Errors | Cls | Lng | Plt | FE | Aos | DB | Dos | Orm | IA |
|-----|-------------------------------------|-------------------------------------|--------|-----|-----|-----|-----|-----|----|-----|-----|-----|
| 1 | wizzardo-http | 843,545 100.0% | 0 | Mcr | Jav | Non | Non | Lin | Pg | Lin | Raw | Rea |
| 2 | vertx-postgres | 728,681 86.4% | 0 | Plt | Jav | ver | Non | Lin | Pg | Lin | Raw | Rea |
| 3 | actix-raw | 683,025 81.0% | 0 | Plt | Rus | Non | act | Lin | Pg | Lin | Raw | Rea |
| 4 | actix-pg | 629,423 74.6% | 0 | Mcr | Rus | Non | act | Lin | Pg | Lin | Raw | Rea |
| 5 | vertx-web-postgres | 629,402 74.6% | 0 | Mcr | Jav | vtx | Non | Lin | Pg | Lin | Raw | Rea |
| 6 | es4x | 613,564 72.7% | 0 | Mcr | JS | ver | Non | Lin | Pg | Lin | Raw | Rea |
| 7 | h2o | 562,260 66.7% | 0 | Plt | C | Non | Non | Lin | Pg | Lin | Raw | Rea |
| 8 | drogon-raw | 559,356 66.3% | 0 | Ful | C++ | Non | Non | Lin | Pg | Lin | Raw | Rea |
| 9 | greenlightning | 518,358 61.4% | 0 | Mcr | Jav | Non | Non | Lin | Pg | Lin | Raw | Rea |
| 10 | revenj-jvm | 480,646 57.0% | 0 | Ful | Jav | Svt | Res | Lin | Pg | Lin | Ful | Rea |
| 11 | cpoll_cppsp-postgres-raw-threadpool | 479,920 56.9% | 0 | Plt | C++ | Non | Non | Lin | Pg | Lin | Raw | Rea |
| 12 | ulib-postgres | 468,484 55.5% | 0 | Plt | C++ | Non | ULi | Lin | Pg | Lin | Mcr | Rea |
| 13 | snail_cppsp-raw | 450,740 57.4% | 0 | Plt | C++ | Non | Non | Lin | My | Lin | Raw | Rea |

Multiple queries

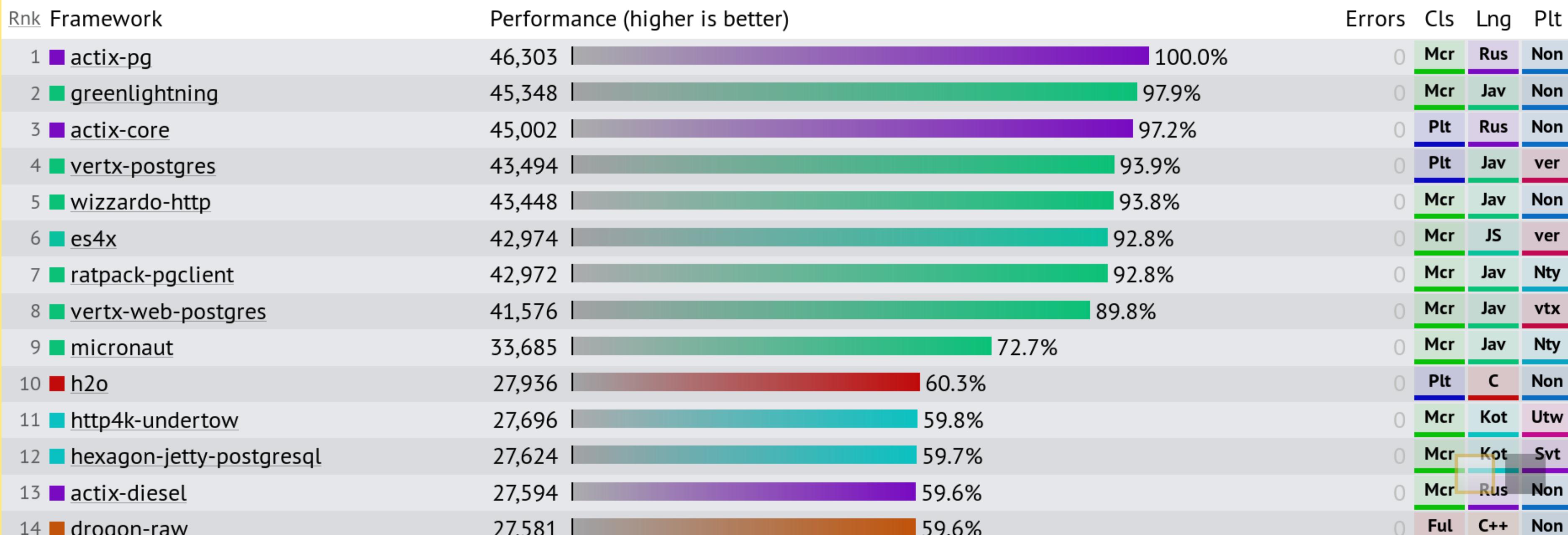
20-queries (bar)

Data table

Latency

Framework overhead

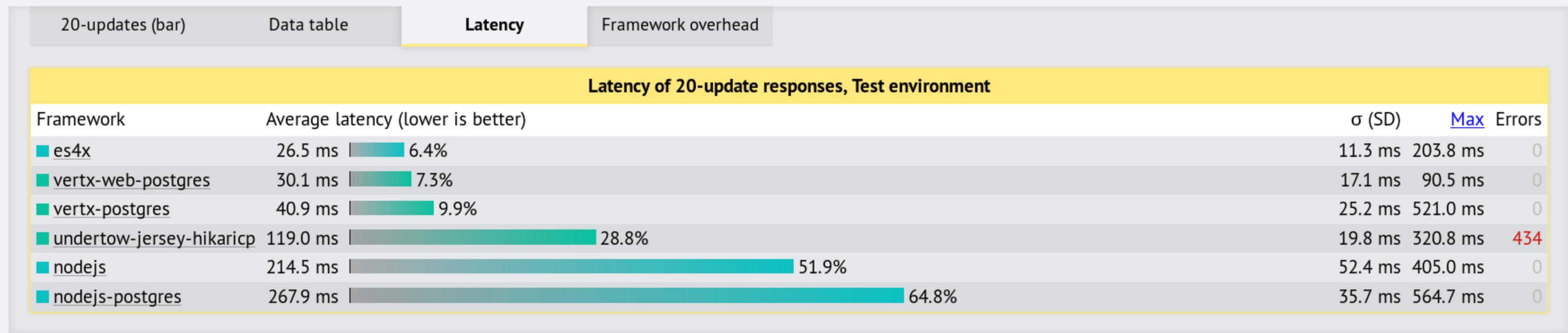
Responses per second at 20 queries per request, Test environment (405 tests)



Polyglot GraalVM is fast.

what about latency?

Data updates



Requirements summary

This test exercises database writes. Each request is processed by fetching multiple rows from a simple database table, converting the rows to in-memory objects, modifying one attribute of each object in memory, updating each associated row in the database individually, and then serializing the list of objects as a JSON response. The test is run multiple times: testing 1, 5, 10, 15, and 20 updates per request. Note that the number of **statements** per request is twice the number of updates since each update is paired with one query to fetch the object. All tests are run at 512 concurrency.

The response is analogous to the multiple-query test. Example response for 10 updates:

```
HTTP/1.1 200 OK
Content-Length: 315
Content-Type: application/json
Server: Example
Date: Wed, 17 Apr 2013 12:00:00 GMT
```

Conclusion

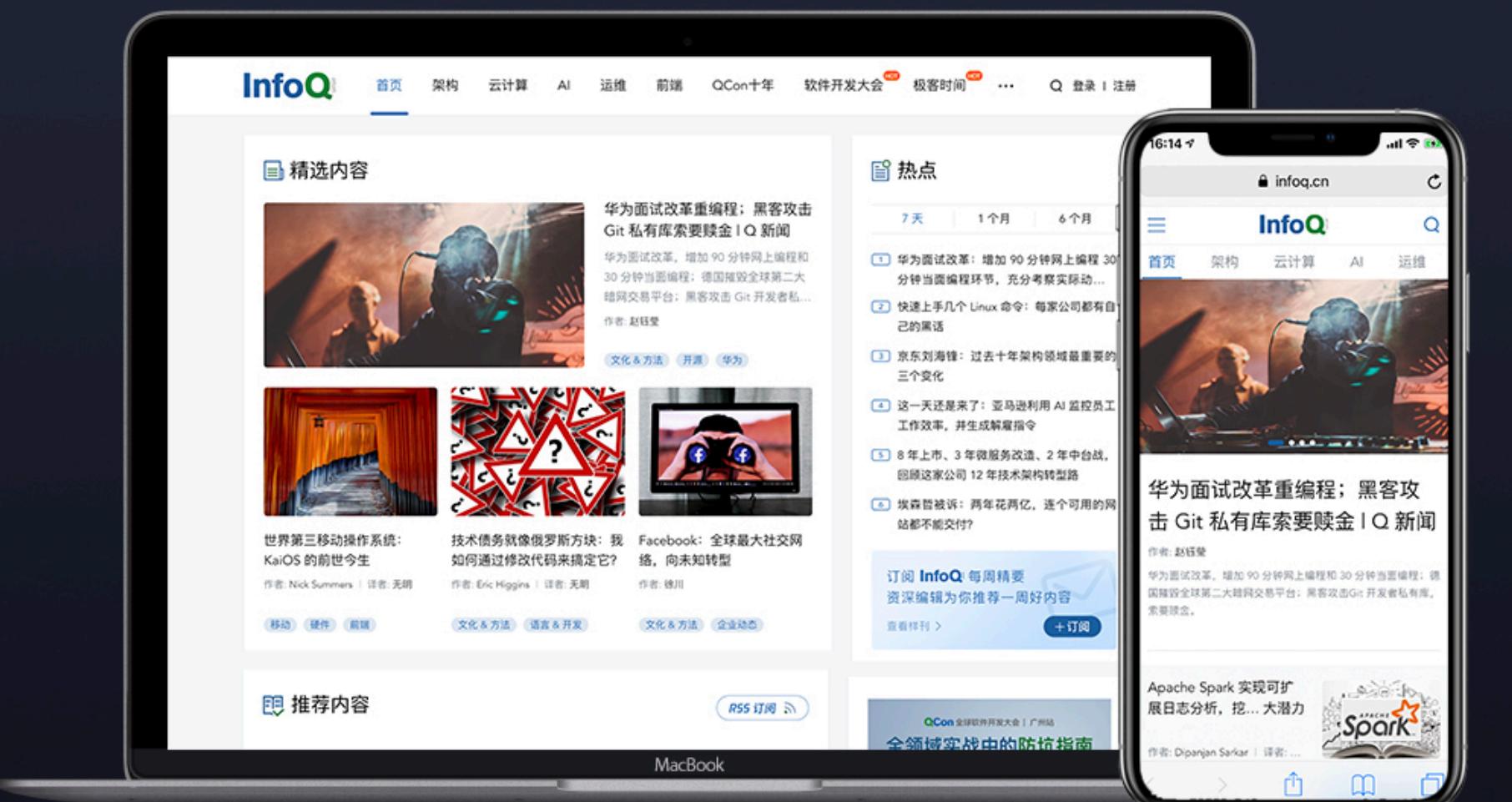
- latency is not a problem, it's a symptom
- use non-blocking to fully use the CPU
- **use vert.x ;-)**
- polyglot is fine
- use graalvm for polyglot JIT optimizations
- node can be slow for server applications
- **use ES4X ;-)**

Let's connect!

-  @pml0pes
-  pmlopes
-  <https://www.linkedin.com/in/pmlopes/>
- <https://reactiverse.io/es4x>
- <https://vertx.io>
- <https://graalvm.org>

InfoQ官网 全新改版上线

促进软件开发领域知识与创新的传播



关注InfoQ网站
第一时间浏览原创IT新闻资讯



免费下载迷你书
阅读一线开发者的技术干货

THANKS! | QCon th