

История одной оптимизации производительности Node.js библиотеки

Андрей Печкуров



О докладчике

- Пишу на Java (10+ лет), Node.js (5+ лет)
- Интересы: веб, архитектура, распределенные системы, производительность
- Можно найти тут:
 - <https://twitter.com/AndreyPechkurov>
 - <https://github.com/puzpuzpuz>
 - <https://medium.com/@apechkurov>

О докладе

- Тема:
 - Подход к оптимизации производительности Node.js библиотек
- Подопытный:
 - Клиентская Node.js библиотека Hazelcast IMDG
- Аудитория:
 - Все, кто разрабатывает сетевые приложения на Node.js

План

1. Знакомство с подопытным
2. Цели и общий подход
3. Бенчмарки и инструменты анализа
4. Оптимизация: замеры, гипотезы, эксперименты
5. Планы на будущее

1. Знакомство с подопытным

Hazelcast IMDG

- <https://hazelcast.org/>
- Hazelcast In-Memory Data Grid (IMDG)
- Большой набор распределенных структур данных
(AP и CP согласно CAP теореме)
- Написана на Java, умеет embedded и standalone режимы
- Хорошо масштабируется вертикально и горизонтально
- Часто используется в high-load и low-latency приложениях

Возможности Hazelcast IMDG

| | Java | Scala | C++ | C#/.NET | Python | Node.js | Go |
|---------|---|--|--|--|--|-------------------|--|
| Clients | Near Cache  | | Near Cache | | | | |
| | REST | Memcached | Clojure | Open Client Network Protocol (Backward & Forward Compatibility, Binary Protocol) | | | |
| | Serialization (Serializable, Externalizable, DataSerializable, IdentifiedDataSerializable, Portable, Custom) | | | | | | |
| APIs | java.util.concurrent | | Web Sessions (Tomcat/Jetty/Generic)  | | Hibernate 2 nd Level Cache  | | JCache   |
| | Map   | MultiMap | Replicated Map | Set | List  | Queue | ReliableTopic |
| | Topic | Ringbuffer | Continuous Query | HyperLogLog | Flake ID Gen. | CRDT PN Counter | |
| | SQL Query | Predicate & Partition Predicate  | | Entry Processor | | Executor Services | Aggregation |
| | AP Subsystem | | | | | | CP Subsystem |

Hazelcast IMDG Node.js client

- <https://github.com/hazelcast/hazelcast-nodejs-client>
- Node.js 4+
- Стек: TypeScript, promisified API (bluebird)
- Первый стабильный релиз - май 2019

Особенности библиотеки

- "Умная" клиентская библиотека
- Общается с нодами кластера по [открытыму бинарному протоколу](#) поверх TCP
- Поддерживает множество распределенных структур данных
- Умеет near cache, retry on failure, client stats и многое другое

Пример использования

```
const Client = require('hazelcast-client').Client;

const client = await Client.newHazelcastClient();
const cache = await client.getMap('my-awesome-cache');

await cache.set('foo', 'bar');
const cached = await cache.get('foo');
console.log(cached); // bar
```

2. Цели и общий подход

Начальные цели

- Анализ текущей производительности перед стабильным релизом
- Включение в релиз "быстрых" правок (при необходимости)
- Постановка планов по дальнейшему анализу и оптимизации
- Спойлер: на сегодня большая часть из этих планов уже реализована

Оптимизация?



Оптимизация? Рецепт приготовления

0. Определить метрики производительности и их желаемые значения
1. Реализовать бенчмарк
2. Сделать замеры производительности
3. Проблема? Подобрать инструменты анализа
4. Найти узкие места, выдвинуть гипотезы и провести эксперименты
5. Сделать замеры
6. goto 0 .

Возможные метрики

- Сетевая клиентская библиотека
- I/O bound нагрузка
- Основные метрики:
 - Операции в секунду (throughput)
 - Время выполнения операции (условно, latency)
- Вспомогательные метрики:
 - Загрузка процессора
 - Потребление памяти

Выбор метрик?

- Оптимизируем throughput
- Желаемые значения: $\forall (\forall)$

Выбор метрик!



3. Бенчмарки и инструменты анализа

Старый бенчмарк

```
// ...
run: function () {
  var key = Math.random() * ENTRY_COUNT;
  var opType = Math.floor(Math.random() * 100);
  if (opType < GET_PERCENTAGE) {
    this.map.get(key).then(this.increment.bind(this));
  }
  // ...
  setImmediate(this.run.bind(this));
}
// ...
```

Старый бенчмарк: минусы

- Все операции стартуют через рекурсивный `setImmediate()`
- Нет ограничений по количеству операций (concurrency limit, backpressure)
- Операции и входные данные выбираются случайным образом
- Все это снижает результат и ухудшает детерминированность

Новый бенчмарк

```
const benchmark = new Benchmark({
    nextOp: () => map.get('foo'), // функция-фабрика для операций
    totalOpsCount: REQ_COUNT,     // общее число операций
    batchSize: BATCH_SIZE         // лимит на кол-во операций
});

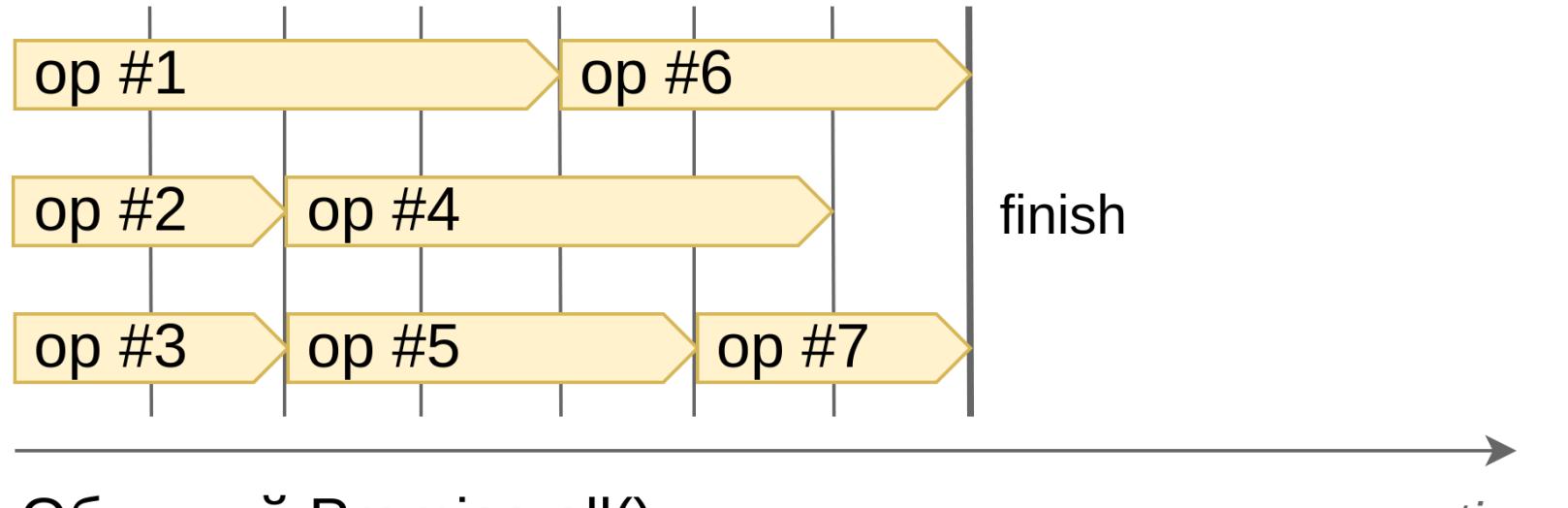
await benchmark.run();
```

Новый бенчмарк: плюсы

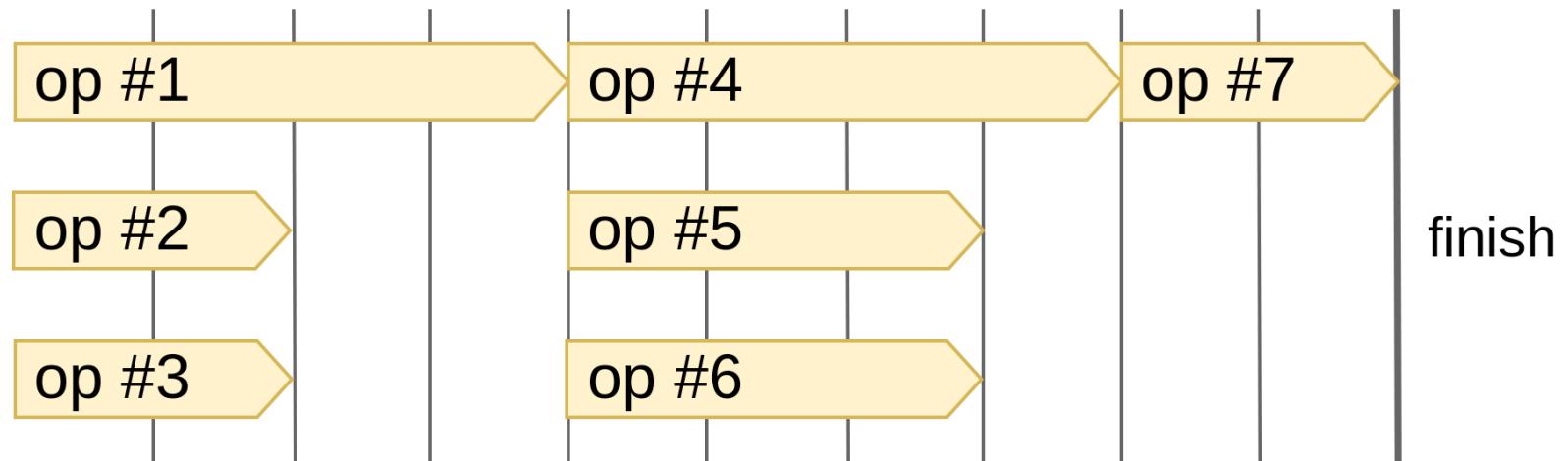
- Операции стартуют параллельно
- Общее число одновременно стартованных операций ограничено
- Операции и входные данные предопределены

Новый бенчмарк: визуализация

Бенчмарк с batchSize = 3, totalOpsCount = 7



Обычный Promise.all():



Сценарий бенчмарка

- Приложение-бенчмарк с клиентской библиотекой
- Кластер из одной ноды IMDG (Docker-контейнер)
- Локальная машина (loopback address)
- Фиксированные версии Linux, Node.js, IMDG и т.д.
- Операции: `Map#get()` и `Map#set()`
- Данные: фиксированные строки с ASCII-символами (3 В, 1 KB, 100 KB)
- Замер: несколько запусков и вычисление среднего результата
- Каждый запуск: 1 млн операций с лимитом 100

Не инструмент #1

- Proof of concept (PoC)
- Все средства хороши, но нужен весь функционал кода на горячем пути

Не инструмент #2

- Микробенчмарки позволяют быстро проверить гипотезу и/или обосновать результаты РоС
- Использовался фреймворк [Benchmark.js](#) (+ node-microtime)
- *Предупреждение:* могут показывать температуру в Антарктиде

Инструмент #1

- Стандартный профилировщик Node.js
- Основан на V8 sample-based profiler
- Учитывает JS и C++ код
- `node --prof app.js`
- Можно получить человекочитаемое представление:
`node --prof-process isolate-0xffffffffffff-v8.log > processed.txt`

Простой пример

```
const Benchmark = require('benchmark');
const suite = new Benchmark.Suite();

suite
  .add('awesome microbenchmark', () => cpuIntensiveFn(...))
  .on('cycle', function (event) {
    console.log(String(event.target))
  })
  .run();

function cpuIntensiveFn (...) {
  // какие-то тяжелые вычисления
}
```

Пример вывода

[JavaScript]:

| ticks | total | nonlib | name |
|-------|-------|--------|---|
| 2806 | 55.2% | 55.5% | LazyCompile: *suite.add /home/puzpuzpuz/app.js:5:34 |
| 1631 | 32.1% | 32.3% | LazyCompile: *<anonymous> :1:20 |
| 35 | 0.7% | 0.7% | Eval: ~<anonymous> :1:20 |
| 12 | 0.2% | 0.2% | Builtin: InterpreterEntryTrampoline |
| 9 | 0.2% | 0.2% | LazyCompile: *cpuIntensiveFn /home/puzpuzpuz/app.js:11:25 |

...

[Summary]:

| ticks | total | nonlib | name |
|-------|-------|--------|------------------|
| 4577 | 90.0% | 90.5% | JavaScript |
| 475 | 9.3% | 9.4% | C++ |
| 15 | 0.3% | 0.3% | GC |
| 29 | 0.6% | | Shared libraries |
| 5 | 0.1% | | Unaccounted |

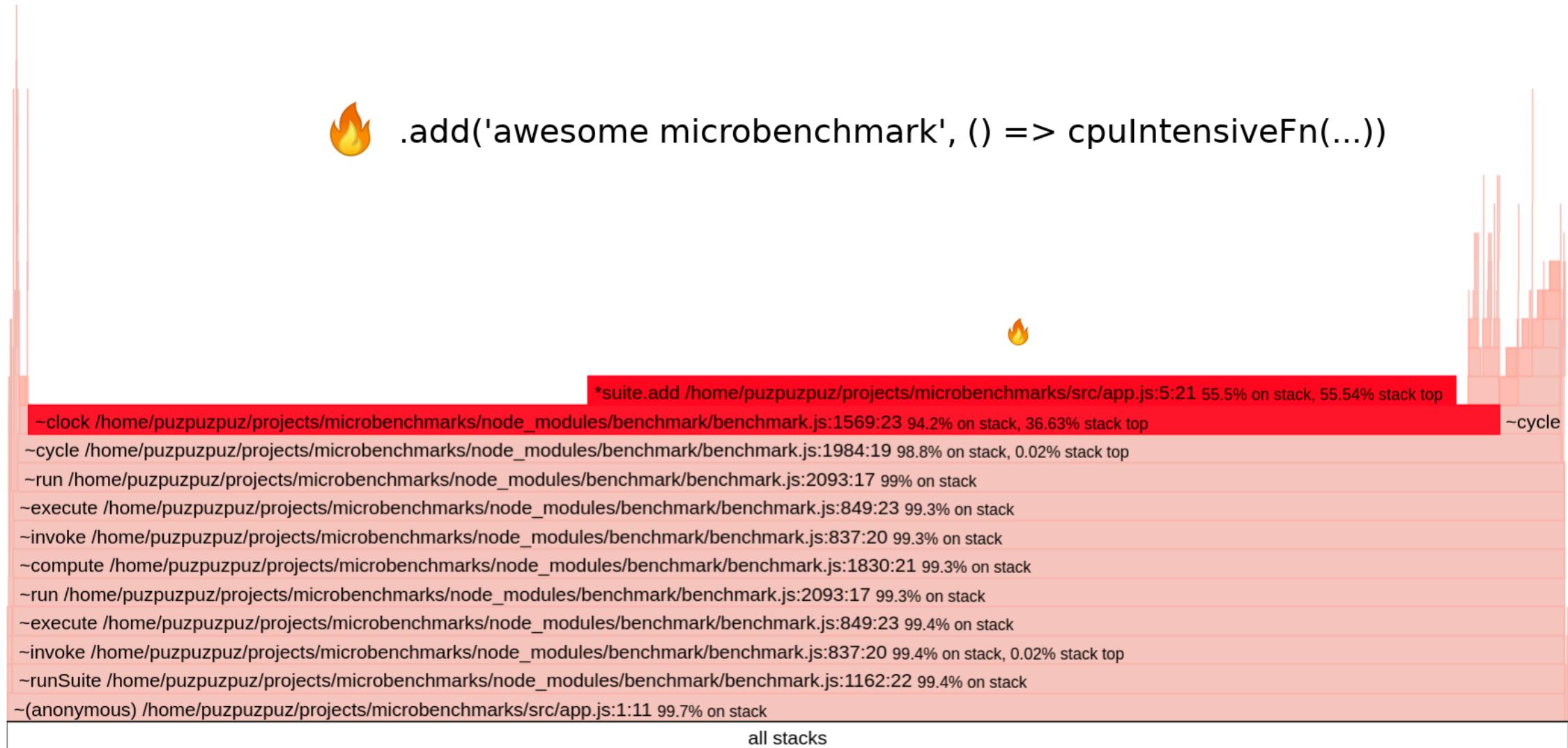
...

Инструмент #2

- Визуализация профиля в виде flame graph
- Отлично работает для event loop'a Node.js
- Действительно помогает обнаруживать узкие места
- Спасибо Brendan Gregg, Netflix, [придумавшему подход](#) в 2013
- Наиболее популярный инструмент - [0x](#) (умеет V8, perf, DTrace)

```
$ npm install -g 0x
$ 0x -o app.js
```

Flame graph для простого примера

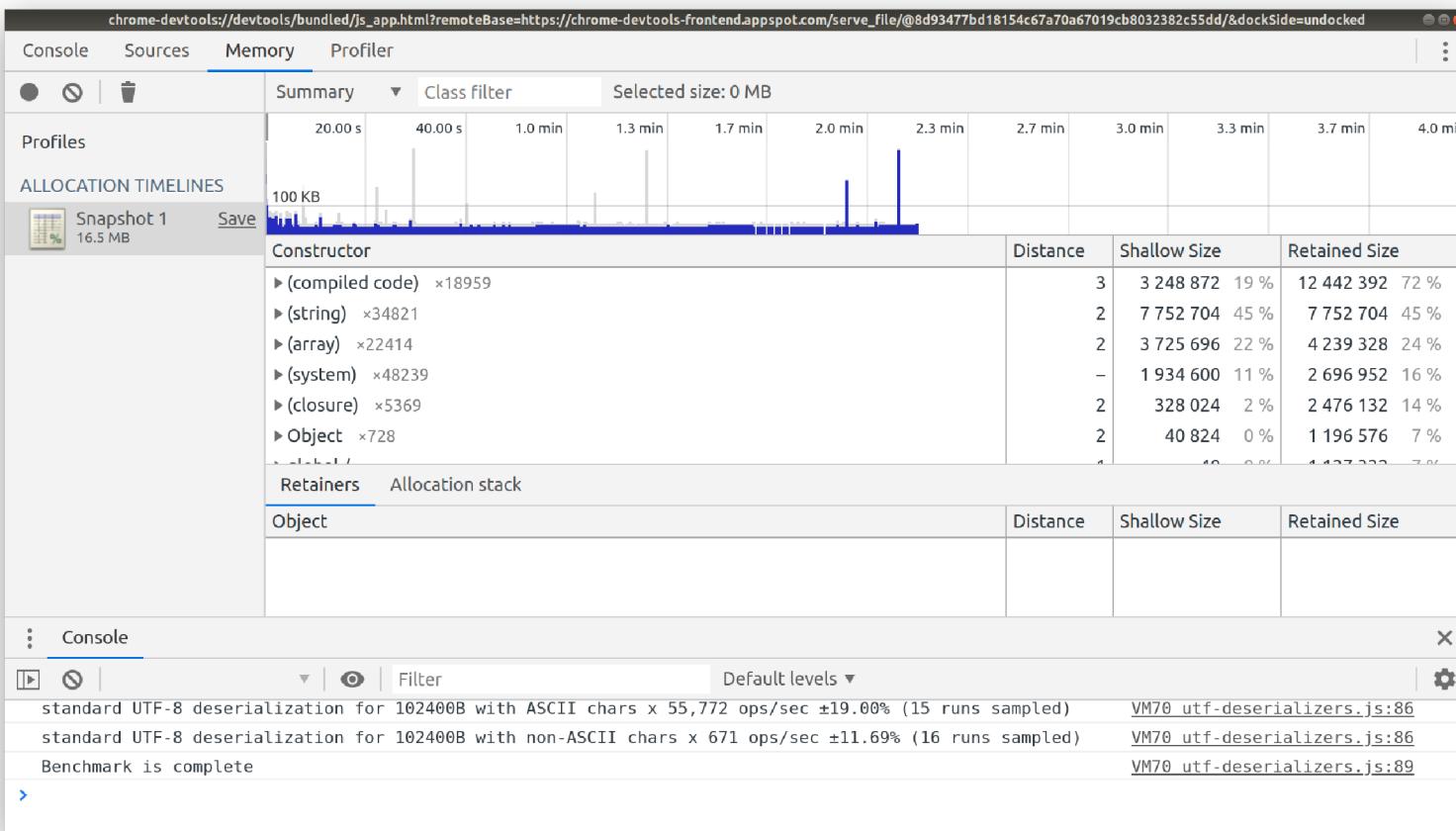


Пример flame graph из реального мира



Инструмент #3

- Профилировщик памяти из Chrome DevTools (Node.js)
- Умеет делать heap snapshot, отслеживать аллокации и не только



Проверяем чеклист

- [X] Метрики
- [X] Бенчмарк
- [X] Инструменты анализа
- [] Оптимизация

4. Оптимизация: замеры, гипотезы, эксперименты

Горячий путь

1. Старт операции (создание `Promise`)
2. Сериализация сообщения в бинарный формат
3. Отправка в сеть в `socket.write(...)`
4. Чтение фрейма в `socket.on('data', ...)`
5. Десериализация ответного сообщения
6. Вызов `resolve()` у `Promise`'а операции

Базовый замер

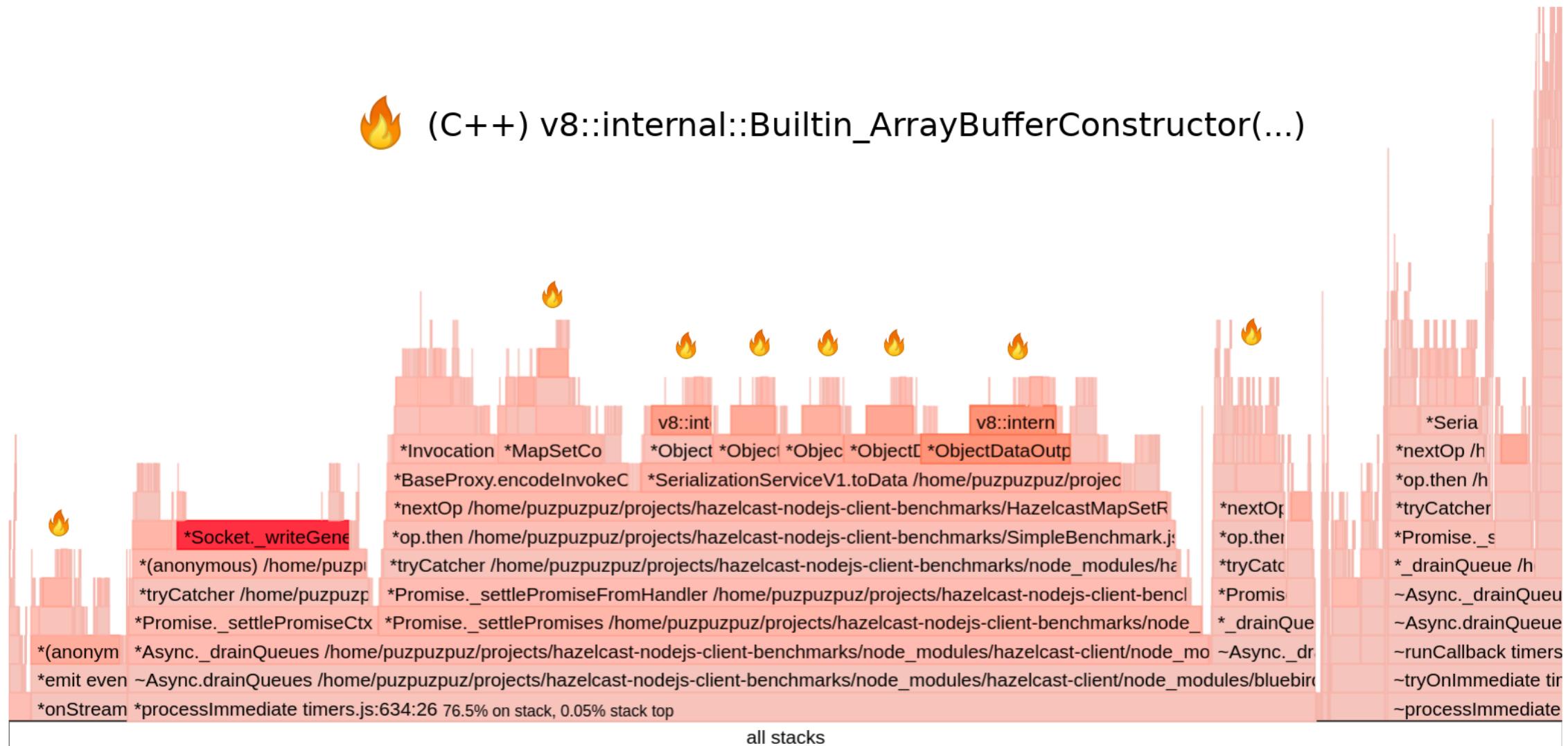
| | get() 3 В | get() 1 KB | get() 100 KB | set() 3 В | set() 1 KB | set() 100 KB |
|---------|--------------|---------------|-----------------|--------------|---------------|-----------------|
| v0.10.0 | 90 933 | 23 591 | 105 | 76 011 | 44 324 | 1 558 |

* Абсолютные значения не важны (замеры сделаны на моем ноутбуке)

Видны проблемы?

- Производительность практически линейно зависит от размера данных
- Java-клиент для `get('foo', 'bar')` быстрее примерно в 5 раз
(конечно, сравнение заведомо некорректное)

Профилировщик, приди! (запись 3 В)



Профилировщик, приди! (запись 3 В)

```
...
[C++ entry points]:
  ticks    cpp    total   name
  2775    37.8%  21.8%  v8::internal::Builtin_ArrayBufferConstructor(...)
   991    13.5%   7.8%  __libc_write
   329     4.5%   2.6%  v8::internal::Builtin_ArrayConcat(....)
...
```

Хьюстон, у нас аллокации

- Для работы с бинарными данными, конечно, используется `Buffer`
- В на горячем пути много `Buffer#alloc()/#allocUnsafe()`, а это "дорогая" операция
- Во время сериализации одной операции происходит несколько аллокаций, а затем буферы копируются в финальный
- Это упрощает код, но производительность страдает
- Сначала пробуем РоС с полумерой

РоС с полумерой

```
export class ObjectDataOutput implements DataOutput {  
  
    protected buffer: Buffer;  
    private pos: number;  
  
    constructor() {  
        // пробуем аллоцировать жадно  
-        this.buffer = Buffer.allocUnsafe(1);  
+        this.buffer = Buffer.allocUnsafe(1024);  
  
        // ...  
    }  
}
```

Замер производительности PoC

| | get() 3 B | get() 1 KB | get() 100 KB | set() 3 B | set() 1 KB | set() 100 KB |
|---------|--------------|---------------|-----------------|--------------|---------------|-----------------|
| v0.10.0 | 90 933 | 23 591 | 105 | 76 011 | 44 324 | 1 558 |
| PoC | 104 854 | 24 929 | 109 | 95 165 | 52 809 | 1 581 |
| | +15% | +5% | +3% | +25% | +19% | +1% |

Промежуточные итоги

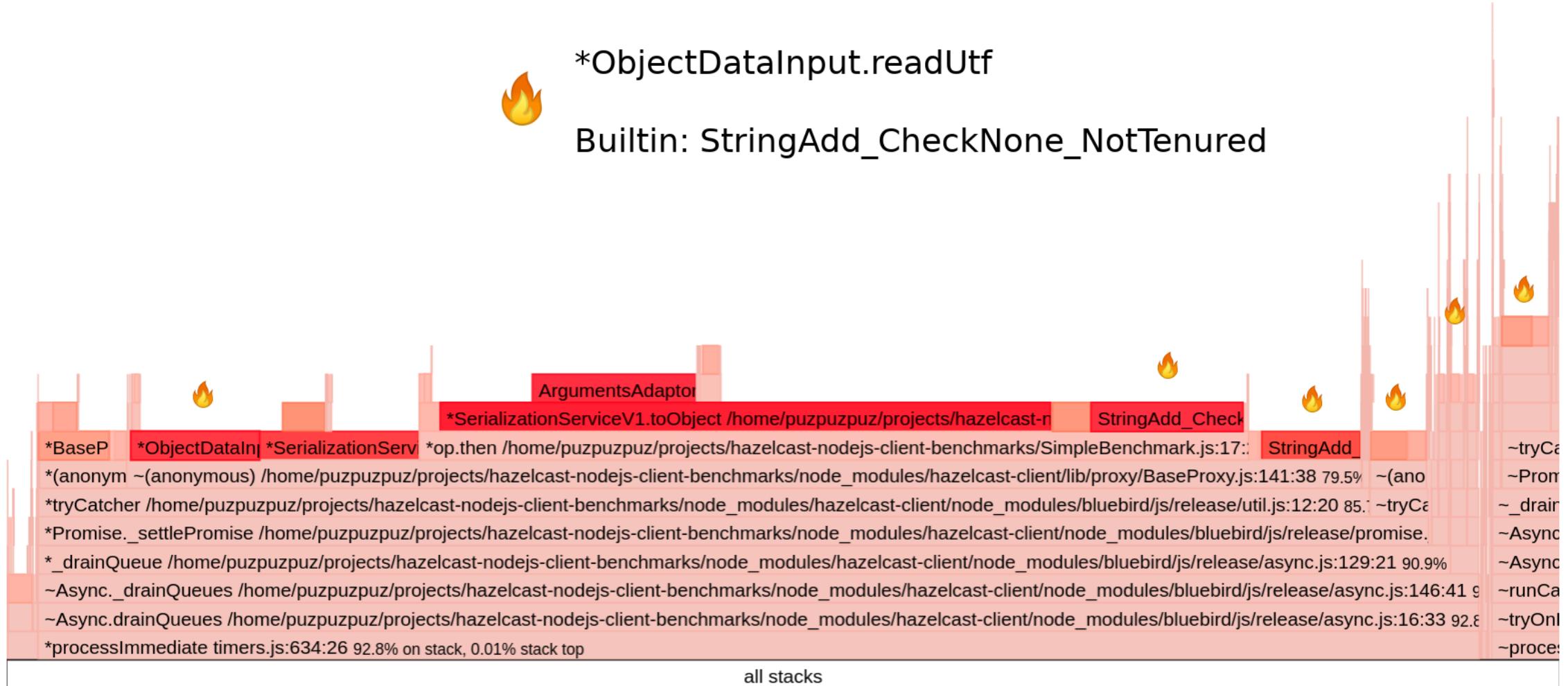
- Гипотеза верна и правка идет в ближайший релиз
- Нужно избавиться от оставшихся избыточных аллокаций в будущих релизах
- Результаты для больших размеров все равно оставляют желать лучшего
- Так что же у нас с зависимостью от размера данных?

Профильтровщик, приди! (чтения 100 КВ)

*ObjectDataInput.readUtf



Builtin: StringAdd_CheckNone_NotTenured



А ЧТО ЭТО У НАС ТАМ?

```
private readUTF(pos?: number): string {
    const len = this.readInt(pos);
    // ...
    for (let i = 0; i < len; i++) {
        let charCode: number;
        leadingByte = this.readByte(readingIndex) & MASK_1BYTE;
        readingIndex = this.addOrUndefined(readingIndex, 1);
        const b = leadingByte & 0xFF;
        switch (b >> 4) {
            // ...
        }
        result += String.fromCharCode(charCode);
    }
    return result;
}
```

Предварительная оптимизация?

- Итак, у нас нестандартная (де)серIALIZАЦИЯ UTF-8 строк
- Похоже на предварительную оптимизацию
- Почему бы не сравнить со стандартным API?

```
// сериализация
buf.write(inStr, start, end, 'utf8');
// десериализация
const outStr = buf.toString('utf8', start, end);
```

Микробенчмарк

| | 100 В ASCII | 100 KB ASCII | 100 В UTF | 100 KB UTF |
|----------|------------------------|-------------------------|----------------------|-----------------------|
| custom | 1 515 803 | 616 | 1 093 390 | 613 |
| standard | 11 297 821 | 68 721 | 1 311 610 | 794 |
| | +645% | +11 056% | +20% | +29% |

* Результаты для десериализации в ops/sec

Проваливаемся в кроличью нору

- Buffer#toString()
- node:buffer.js#stringSlice()
- node:node_buffer.cc#StringSlice()
- node:StringBytes#Encode()
- v8:String#NewFromUtf8()
- v8:Factory#NewStringFromUtf8()
- v8:Factory#NewStringFromOneByte()

Что там, в hope?

```
// v8:Factory#NewStringFromUtf8()
MaybeHandle<String> Factory::NewStringFromUtf8(
    Vector<const char> string,
    PretenureFlag pretenure
) {
    // Check for ASCII first since this is the common case.
    const char* ascii_data = string.start();
    int length = string.length();
    int non_ascii_start = String::NonAsciiStart(ascii_data, length);
    if (non_ascii_start >= length) {
        // If the string is ASCII, we do not need to convert
        // the characters since UTF8 is backwards compatible with ASCII.
        return
            NewStringFromOneByte(
                Vector<const uint8_t>::cast(string), pretenure);
    }
    // ...
}
```

РоС для сериализации

| | get() 3 B | get() 1 KB | get() 100 KB | set() 3 B | set() 1 KB | set() 100 KB |
|---------|--------------|---------------|-----------------|--------------|---------------|-----------------|
| v0.10.0 | 90 933 | 23 591 | 105 | 76 011 | 44 324 | 1 558 |
| РоС | 122 458 | 104 090 | 7 052 | 110 083 | 73 618 | 8 428 |
| | +34% | +341% | +6 616% | +45% | +66% | +440% |

Промежуточные итоги

- Гипотеза про (де)сериализацию верна и правка идет в ближайший релиз

Первый публичный релиз

| | get() 3 B | get() 1 KB | get() 100 KB | set() 3 B | set() 1 KB | set() 100 KB |
|---------|--------------|---------------|-----------------|--------------|---------------|-----------------|
| v0.10.0 | 90 933 | 23 591 | 105 | 76 011 | 44 324 | 1 558 |
| v3.12 | 132 855 | 120 670 | 8 756 | 127 291 | 94 625 | 10 617 |
| | +46% | +411% | +8 239% | +67% | +113% | +581% |

Editor's Cut

1. Эксперимент с пулом буферов для сериализации
 - Неудачный. К тому же, `Buffer#allocUnsafe()` и так [использует](#) пул (8 KB by default)
2. Эксперимент с Write Queue (aka Automated Pipelining)
 - Оказался перспективным и был отложен до следующего релиза.
Простой РоС показал увеличение write throughput на 20-25%

Оптимизация Automated Pipelining

- Подобная оптимизация использована в классе `WriteQueue` в DataStax Node.js Driver for Apache Cassandra
- Идея в объединении сообщений перед записью в сокет
- В результате делается меньше "дорогих" вызовов `Socket#write()`

Pipelining

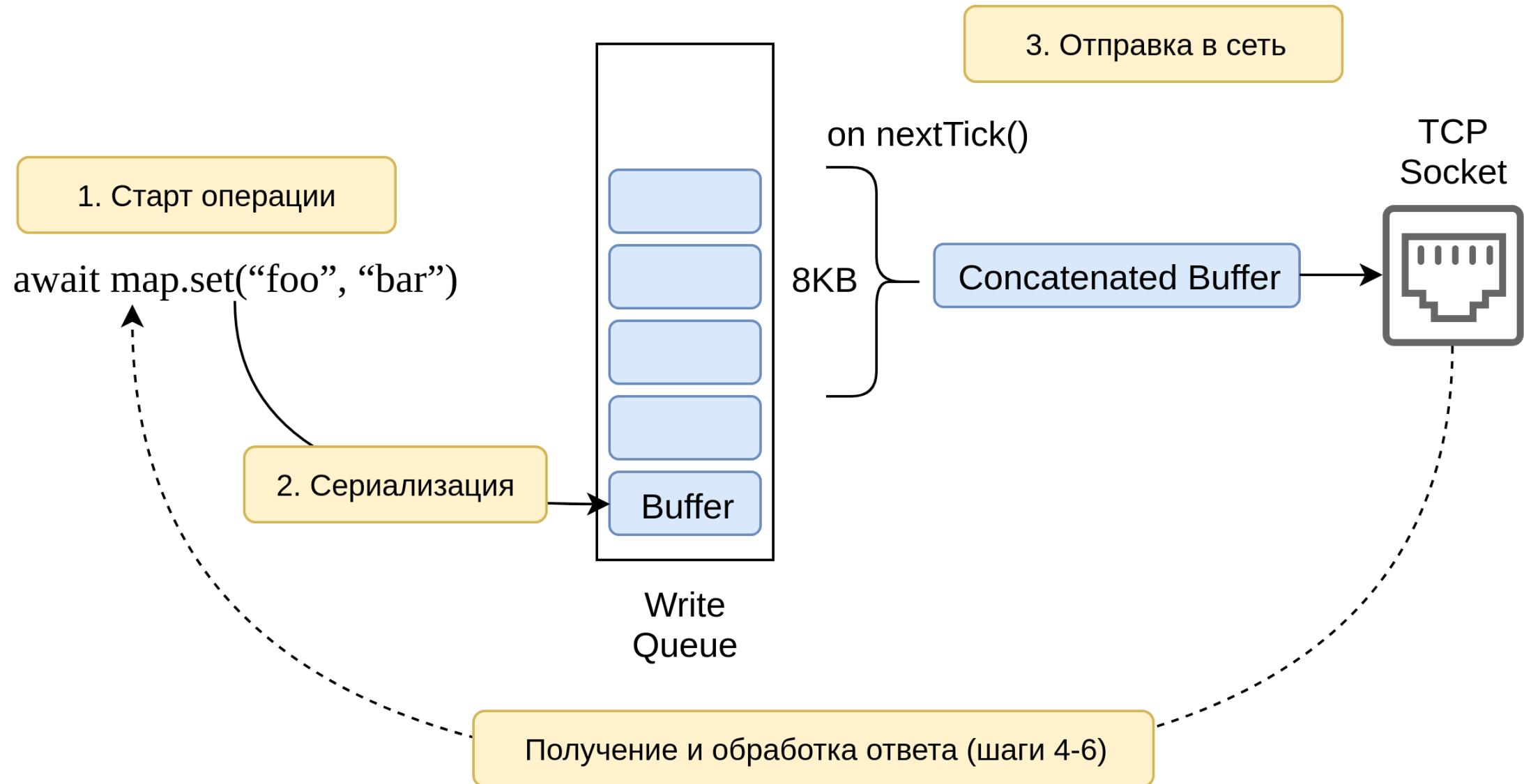
- В блокирующихся клиентах схожий прием называют **pipelining**
- Пример pipelining API в Java:

```
Pipelining<String> pipelining = new Pipelining<>(10);
for (int i = 0; i < 100; i++) {
    pipelining.add(map.getAsync(i));
}
// блокируемся и ждем результат
pipelining.results();
```

Automated Pipelining

- В нашем случае, объединение происходит неявным образом, отсюда название - automated pipelining
- Достоинство:
 - Большее число приложений оказываются в выигрыше
- Недостаток:
 - Из обычного pipelining можно "выжать" лучшую производительность

Логика работы



Особенности

- Для больших сообщений, записываемых в сокет, производительность может упасть
- Поэтому добавлена настройка библиотеки, включающая "обычный" режим
- Порог записи в сеть тоже вынесен в настройки
- *Примечание:* данная оптимизация не отменяет появление batch API в будущем

Снова релиз?

- Automated Pipelining вошла в ближайший релиз
- Кроме этого, туда вошли миорные оптимизации для горячего пути:
 - Убрали мусор от `new Date().getTime()` (спасибо, `Date.now()`)
 - Убрали мусор от лишних конвертаций `number <-> Long` (используется [long.js](#))
 - Избавились от аллокаций буферов в edge cases, например, при чтении больших сообщений (> 128 KB)
- Спойлер: заметной прибавки миорные оптимизации не дали, но они все равно полезны

Второй публичный релиз

| | get() 3 В | get() 1 KB | get() 100 KB | set() 3 В | set() 1 KB | set() 100 KB |
|---------|--------------|---------------|-----------------|--------------|---------------|-----------------|
| v3.12 | 132 855 | 120 670 | 8 756 | 127 291 | 94 625 | 10 617 |
| v3.12.1 | 173 611 | 161 812 | 10 879 | 172 028 | 82 747 | 8 208 |
| | +30% | +34% | +24% | +35% | -13% | -23% |

* Замеры с включенным Automated Pipelining

А как же аллокации?

- Поскольку правки объемные, было решено начать с РоС
- РоС реализует оптимизацию только для `Map#get()` and `Map#set()`
- Он позволит оценить, оправдает ли оптимизация вложенные усилия

РоС без избыточных аллокаций

| | get() 3 B | get() 1 KB | get() 100 KB | set() 3 B | set() 1 KB | set() 100 KB |
|---------|--------------|---------------|-----------------|--------------|---------------|-----------------|
| v3.12.1 | 173 611 | 161 812 | 10 879 | 172 028 | 82 747 | 8 208 |
| РоС | 222 172 | 192 122 | 12 594 | 205 254 | 109 051 | 11 630 |
| | +28% | +19% | +16% | +19% | +32% | +42% |

* Замеры с включенным Automated Pipelining

Сравним с тем, что было в начале

| | get() 3 B | get() 1 KB | get() 100 KB | set() 3 B | set() 1 KB | set() 100 KB |
|---------|--------------|---------------|-----------------|--------------|---------------|-----------------|
| v0.10.0 | 90 933 | 23 591 | 105 | 76 011 | 44 324 | 1 558 |
| PoC | 222 172 | 192 122 | 12 594 | 205 254 | 109 051 | 11 630 |
| | +144% | +714% | +11 894% | +170% | +146% | +646% |

* Замеры с включенным Automated Pipelining

Наконец-то!



5. Планы на будущее

Текущие планы

1. Избавиться от избыточных аллокаций буферов
2. Включить замеры производительности в релиз цикл
3. Продолжить оптимизацию: периодические замеры, гипотезы, эксперименты

Совет #1

Не пытайтесь оптимизировать все и сразу

Совет #2

**Бенчмарки должны быть детерминистичными и
"справедливыми"**

Совет #3

Семь раз замерь, один релизни

Совет #4

Выбирайте инструменты под задачу, а не наоборот

Совет #5

Помните, что оптимизация производительности - это процесс

Полезные ссылки

- <https://hazelcast.org/>
- <https://github.com/hazelcast/hazelcast-nodejs-client>
- <https://nodejs.org/en/docs/guides/simple-profiling/>
- <https://nodejs.org/en/docs/guides/dont-block-the-event-loop/>
- <https://blog.insiderattack.net/event-loop-and-the-big-picture-nodejs-event-loop-part-1-1cb67a182810>

Спасибо за внимание!

Время для Q&A

