Содержание

Список сокращений и условных обозначений

- CAS - Compare And Swap
- MWCAS - Multi Word Compare And Swap
- ОС - Операционная Система
- MPMC - Multiple Producer Mulptiple Consumer

Введение

На сегодняшний день в основе большинства разрабатываемых приложений: веб-серверов, кластеров обработки данных и.т.п. лежит инфрастуктурная основа в виде надёжной и производительной среды предоставления асинхронного исполнения - runtime. От него требуется обеспечение эффективной обработки задач, предоставления инструментов композиции и коммуникации между ними. Такая основа может быть встроена в сам язык, как например в Elixir и Go, или использоваться как отдельная библиотека. Примером такого подхода являются библиотеки Kotlin Coroutines и Rust Tokio.

Любой такой инструмент строит свои абстракции исполнения поверх процессов или потоков, предоставляемых операционной системой. Поэтому неминуемо возникает потребность в синхронизации между ними. Синхронизация по своей природе может быть нескольких типов. Каждый из них предоставлет как преимущества, так и недостатки. Современные архитектуры имеют тенденцию распараллеливать вычислетельные процессы. Однако в большом числе случаев для синхронизации до сих пор используются инструменты крайне неэффектино масштабирующиеся вслед за архитектурой ЭВМ. В большинстве случаев - это блокирующая синхронизация, несмотря на то, что существуют способы производить операции в неблокирующем режиме, который открывает возможности к существенному масштабированию вычислений. Связано это с тем, что неблокиющая синхронизация довольно сложна в реализации, в отличие от блокирующей, а для комплексных структур данных эффективная и простая реализация становится практически невозможной.

В рамках данной работы рассмотрена возможная оптимизация для примитивов синхронизации фреймворка Tokio в среде языка Rust

Цели и задачи работы

Целью работы является оптимизация модуля широковещательной рассылки фреймворка Tokio.

Для выполнения цели были выделены следующие задачи:

- Определить оптимальный дизайн очереди
- Реализовать полученную модель
- Провести сравнительное тестирование

Введение в предметную область

1 Синхронизация

Синхронизация между потоками по своей природе может быть разных видов, в понимании того, что они предоставляют пользователю разные гарантии прогресса многопоточного исполнения. Каждый вид на сегодняшний день имеет как свои достоинтсва, так и недостатки.

1.1 Блокирующая синхронизация

Первый тип представляет собой блокирование прогресса всех исполняемых задач на время выполнения одной или нескольких выбранных задач. У такого типа синхронизации есть большое преимущество - он простой и не требует специального подхода к построению структуры данных над которой производятся операции.

Самый простой пример такой синхронизации - использование мьютекса:

```
// Располагается в общей для потоков памяти
let mutex = Mutex::<State>
```

```
//   Вызывается   разными   потоками
fn   doCriticalSection(){
      {
      mutex.lock()
            //   Критическая   секция
            mutex.unlock()
      }
}
```

Однако данный подход крайне неэффективно масштабируется, ведь в единицу времени может выполняться только одна критическая секция. Остальные потокам необходимо ждать освобожнение блокировки. Ожидание может быть сопряжено с дополнительными системными вызовами, например futex syscall. Или же затраты на переключение контекста и координацию в очередях ожидиания, в случае использования корутин поверх потоков.

Если присутствует большое число потоков оперирующих над критической секцией, появляется большое число накладных расходов, накладных расходов. В худшем случае такое исполнение может показать производительность сравнительно худшую чем обычное последовательное исполнение. Существует также проблема, при которой поток или процесс захвативший блокировку и исполняющий критическую секцию, будет временно снят с исполнения планировщиком задач операционной системы. В данном случае возникает риск полной остановки прогресса исполнения системы до разблокировки.

1.2 Неблокирующая синхронизация

Для избавления от проблем присущих блокирующей синхронизации, существует альтернативный подход, при котором, операции над данными осуществляются в “неблокирующем режиме”.

Неблокирующая синхронизация предоставляет следующие преимущества по сравнению с блокирующей:

- Гарантия прогресса системы в целом - означает, что при любом многопоточном исполнении, всегда есть поток или потоки успешно завершающие свои операции. Решения планировщика ОС теперь не могут привести к полной остановке системы.
- Более высокая масштабируемость - при использовании неблокирующей синхронизации потоки не обязаны ждать друг друга, поэтому операции могут работать в параллель. Вся координация между потоками образуется в специальных точках синхронизации. В большинстве языков программирования - это атомарные переменные.
- Меньшие накладные расходы на синхронизацию. Использование атомарных переменных не требует обращения к ядру операционной системы. Операции над атомарными переменными напрямую упорядочивают исполнение через L2 кэш ядер процессора, за счёт чего достигается синхронизация памяти ядер.

1.3 Общий подход к неблокирущей синхронизации

При использовании неблокирующей синхронизации образуется общий подход при построении структуры даннных и операций над ней. В большинстве случаев необходима модификация структуры на основе её текущего состояния. Выделяется общее состоянии, которое становится атомарной ячейкой памяти, в том плане, что все операции над ней линеаризуемы и образуют некоторый порядок обращений.

Все операции абстрактно разбиваются на три этапа:

1. Копирование текущего состояния (snapshot).
2. Локальная модификация полученного состояния.
3. Попытка замена общего состояния на модифицированную копию, в случае, если общее состояние за время модификации не изменилось. Если состояние успело измениться - начать заного с шага №1.

На такое поведение можно смотреть как на транзакцию над одной ячейкой памяти.

В псеводокоде это можно представить так:

```
// Располагается в общей для потоков памяти
let state = Atomic<State>

// Вызывается из разных потоков
fn doLockFreeOperation(){
while(true){
      let old_state = state.atomic_read()

      let modified_state = modify(old_state)

      if(state.atomic_cas(old_state,modified_state)){
            // За время транзакции состояние не
```

```
изменилось
            //  Поток  успешно  завершает  транзакцию
            break;
      }  else{
            //  Операция  по  замене  неуспешна
            //  Поток  повторяет  цикл
            continue;
      }
}
}
```

Несмотря на свои плюсы, такой подход обладает одним серъездным недостатком. Необходимая линеаризуемость и следующая из неё синхронизация, образуется лишь вокруг одной ячейки памяти. Однако в большинстве структур данных чаще всего требуется атомарная замена сразу нескольких ячеек памяти. Любые прямые изменения хотя бы двух ячеек памяти влекут за собой потерю порядка исполнения, так как между двумя атомарными операциями может произойти сколько-угодно сторонних событий, в зависимости от решения приоритета исполнения планировщика операционной системы.

Для решения описанной проблемы была представлена транзакционная память. Её можно реализовать как на уровне процессора, так и программно. Так как сейчас процессоры не поддерживают подобную опцию, будет рассмотрено использование программной реализации в виде примитива MWCAS.

2 Обзор MWCAS

MWCAS обобщает подход транзакции над одной ячейкой памяти до произвольного их числа.

3 Принцип рыботы MWCAS

Пример исполнения транзакции может выглядеть следующим образом:

```
// Ячейки располагаются в общей для потоков памяти
let state_1 = Atomic<State>
let state_2 = Atomic<State>

// Вызывается из разных потоков
fn doAtomicTransaction(){
while(true){
      let old_state1 = state_1.atomic_read()
      let old_state2 = state_2.atomic_read()

      let modified_state_1 = modify(old_state_1)
      let modified_state_2 = modify(old_state_2)

      let mwcas = new Mwcas

      // Транзакция атомарно заменяет ожидаемые значения
            // на модифицированные копии.
            // В случае, если наблюдаемое старое
значение изменилось
            // Транзакция помогает завершиться другой
возможной транзакции
            // и сообщает о неуспешном завершении
      if(mwcas.transaction(
      memory_cell = [state_1, state_2]
      expected_states = [old_state1, old_state2],
      new_states = [modified_state_1, modified_state_2],
      )){
            break;
      } else{
               // Транзакция прошла неуспешно
               // Поток повторяет цикл
```

```
            continue;
        }
}
}
```

Обзор модуля широковещательной рассылки Tokio

Модуль широковещательной рассылки фреймворка tokio представляет собой канал для пересылки сообщений между между потоками программы в режиме доступа Multiple Producer Multiple Consumer (MPMC): в канал конкуррентно могут одновременно отправлять сообщения сразу несколько потоков. При этом каждое значение доступно для чтения всем подписавшимся на момент отправки сообщения потокам читателям, достигается это за счёт дополнительного счётчика, устанавливаемого вместе с сообщением. Счётчик обновляется с добавлением или удалением потоков-читателей. При определённых обстоятельствах медленный поток-читатель может пропустить некоторые сообщения. В таком случае он переходит на последние актуальные сообщения. В случае если сообщений нет, поток-читатель становится в очередь ожидания значения.

4 Внутреннее состояние

С точки зрения программной реализации канал представляет собой общее состояние в памяти, обращения к которому совершаются с помощью интерфейсов структур двух типов: Sender и Reciever. Каждая структура предоставляет лёгковесный способ управления каналом через определённый интерфейс.

```
pub struct Sender<T> {
    // Ссылка на общее состояние
        shared: Arc<Shared<T>>,
}

pub struct Receiver<T> {
    // Ссылка на общее состояние
```

```
            shared:   Arc<Shared<T>>,

            // Локальная позиция для следующего чтения
            next:   u64,
}
```

Общее состояние содержит в себе:

1. Память самой очереди: она представляется в виде обычного массива слотов (структура типа Slot), логически представленного в виде кольцевого буффера (buffer). Каждому слоту в соотетсвие ставится RwLock - блокирующий примитив синхронизации позволяющий производит параллельное чтение, при отсутвии записи.

2. Основная информация для координации операций над очередью, представлена структурой Slot, это:
   1. Логическая позиция нового следующего сообщения
   2. Очередь (связный список) ожидания следующего значения
   3. Текущее число потоков-читателей
   4. Состояние канала (открыт / закрыт)

Синхронизация доступа к этим полям осуществяется с помощью мьютекса. Вокруг этой точки происходит осно

3. Поля необходимые для закрытия канала и определения текущей длины. В рамках оптимизации рассматриваться не будут.

```
struct   Shared<T>   {
            buffer:   Box<[RwLock<Slot<T>>]>,

            mask:   usize,
```

```
            tail:   Mutex<Tail>,

            num_tx:   AtomicUsize,

            notify_last_rx_drop:   Notify,
}

struct   Tail   {
            pos:   u64,

            rx_cnt:   usize,

            closed:   bool,

            waiters:   LinkedList<Waiter,   <Waiter   as
linked_list::Link>::Target>,
}

struct   Slot<T>   {
            rem:   AtomicUsize,

            pos:   u64,

            val:   UnsafeCell<Option<T>>,
}
```

5 Алгоритм канала

Массив слотов ограничен в длине значением, задаваемым пользователем. В этом возникает необходимость из-за проблемы следующей из архитектуры канала: так как значение должно быть доставлено всем получателям, оно должно всё это время оставаться в памяти, и, если для каждого нового сообщения будет выделяться

новая память, наличие лишь одного медленного потока-читателя может привести к переполнению памяти.

Две основные операции канала - отправление и получение сообщений.

Операция отправления - синхронная. Потоки-писатели не обязаны ждать потоков-получателей. В связи упомянутой раньше проблемой, при которой выделение памяти может привести к её неконтролируемогу росту, буффер очереди ограничен, и при этом новые сообщения, превышающие длину очереди перезаписывают старые. Потоки-читатели, не успевшие получить отправленные ранее сообщения, с помощью своего собственного локального счётчика позиции сообщений, обнаруживают неконсистентность и переводят счётчик на текущую актуальную позицию в очереди, добавляя к нему один круг длины очереди.

Псевдокод операции:

```
fn send(newValue){
    // Захват общей блокировки очереди
    tail.lock()

    // Захват блокировки слота, в который будет произведена запись
    buffer[nextId].lockWrite()

    // Запись нового значения
    buffer[nextId].value = newValue

    // Обновление мета-информации и необходимых счётчиков
    ...
```

```
        // Освобождение блокировки слота
        buffer[nextId].unlock()
        // Запуск на исполнение всех ожидающих задач потоков-читателей
        queue.notify_all()

        // Общая разблокировка очереди
        tail.unlock()
}
```

Операция получения - асинхронная. При попытке получить сообщение, поток-читатель может обнаружить ситуацию при которой готовых сообщений в очереди нет. Это может произойти как сразу при совпадении локального счётчика потока-отправителя и счётчика позиции слота, так и после добавления дополнительного круга длины, в случае, описанном ранее. В этом случае поток-читатель добавляет задачу на очередную проверку очереди в специальную очередь ожидания. Потоки-отправители после отправки сообщений проверяют эту очередь и запланирую на исполнение все оставшиейся в ней задачи.

Псевдокод выполнения операции:

```
// taskHandler - структура, отвечающая за планирование
// задачи, вызвавшей recv, на исполнение

async fn recv(taskHandler) -> T {

        // Следующий слот на чтение значения
        let slot = buffer[nextId]

        // Блокировка слота на чтение
```

```
        // Допускается параллельное чтение
        slot.lockRead()

        // Поток-читатель отстал от актуальной информации
        // как минимум на один круг очереди
        if slot.pos != self.next {

            // Блокировка состояния канала
            tail.lock()

            // Добавление одного круга очереди
            let next_pos = slot.pos + buffer.len()

            // Позиция потока-читателя совпадает
            // с суммой позициии слота и длины одного круга очереди
            // В таком случае готового значения ещё нет.
            // Поток оставляет задачу в очереди ожидания
            if self.next == next_pos{
                qeueu.park(taskHandler)
            }

            // Поток-читатель утанавливает указатель
            // на самое старое текущее значение в канале
            // Все остальные значения считаются пропущенными
            let next = tail.pos - buffer.len()
            let missed = next - self.next
            let self.next = next

            // Блокировка состояния канала и слота
            slot.unlockRead()
            tail.unlock()
```

```
        } else{
            //  Если  первая  проверка  показала,
            //  что  текущее  значение  self.next  совпадает
с  позицией  слота,
            //  это  означает,  что  поток  читает
актуальную  информацию,
            //  В  таком  случае  он  инкрементирует
локальный  счётчик  и  возвращает  значение
            let  result  =  slot.value
            slot.unlock()
            self.next++
            return  result
        }


}
```

Новый способ синхронизации позволит полностью избавиться от блокирующих примитивов Mutex и RwLock

Сравнительный анализ производительности

Описание тестов …

Блокирующая реализации broadcast

contention/10

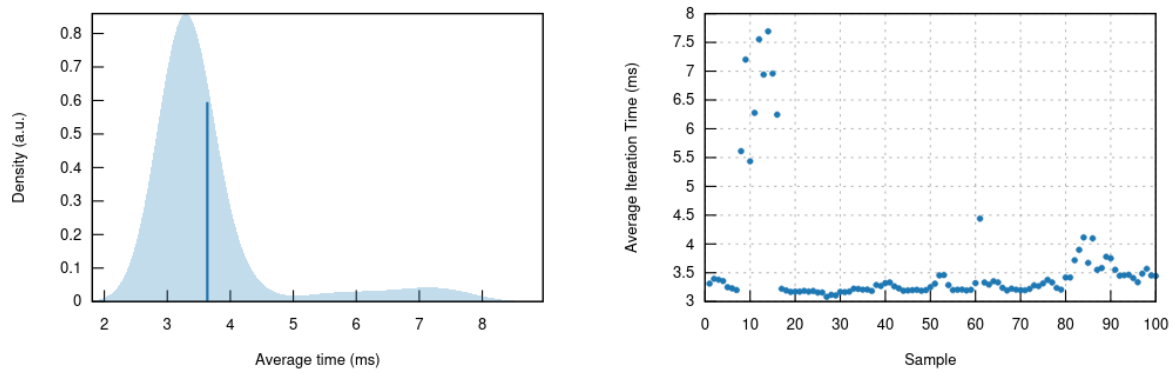

Рисунок 1 — пример изображения

contention/100

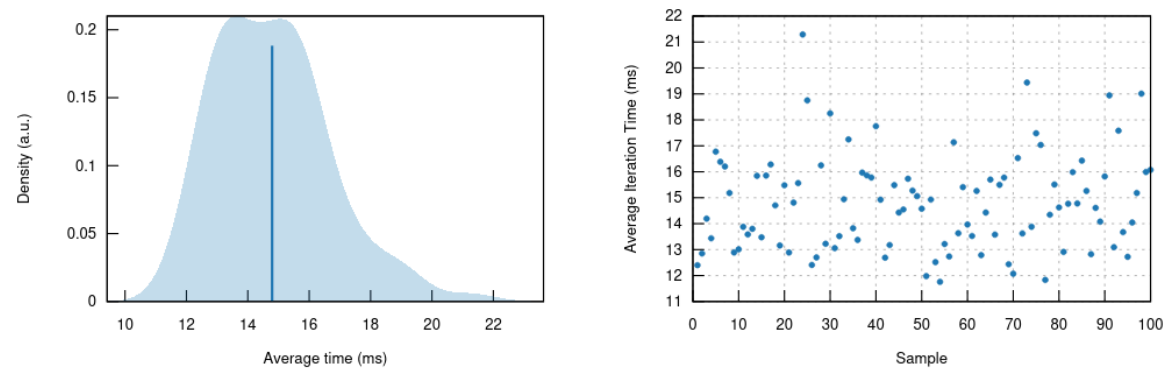

Рисунок 1 — пример изображения

contention/1000

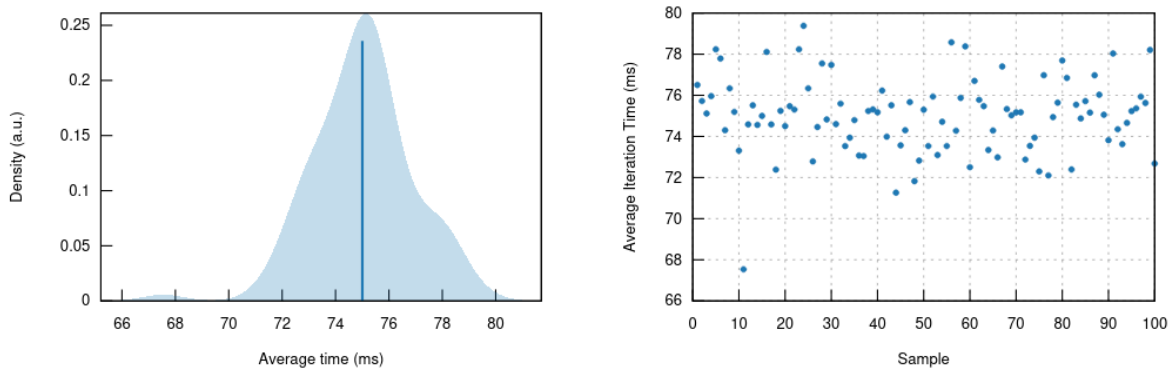

Рисунок 1 — пример изображения

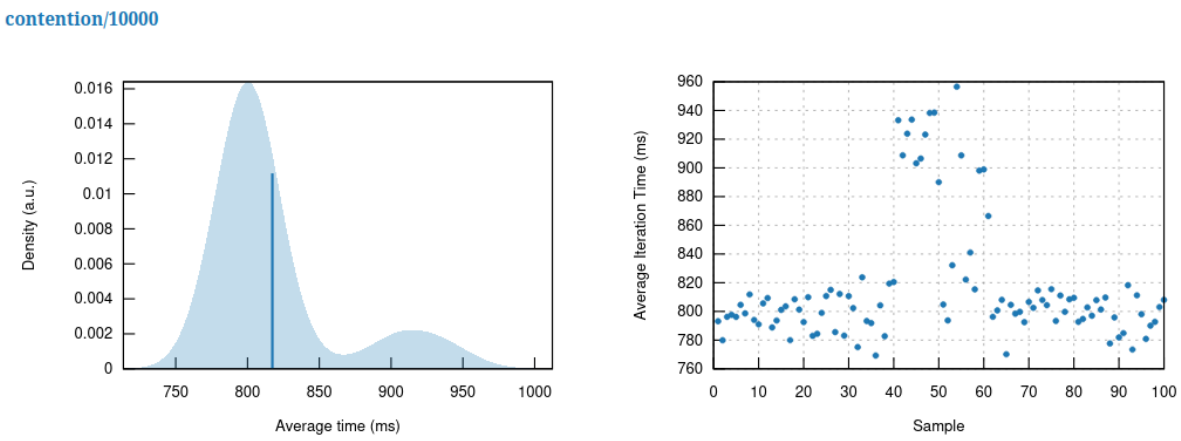

Рисунок 1 — пример изображения

6 Актуальность темы исследования

6.1 Актуальность темы исследования. Часть 1

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri.

6.2 Цель и Задачи

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri.

Заключение

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri.

Таблица 1 — Таблица

Таблица	**для**	**примера**
item1	$\sum_{k=0}^{n} k = 1 + ... + n$	description1
item2	$\sqrt{2}$	description2

Список использованных источников

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri.

Приложение

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim aeque doleamus animo, cum corpore dolemus, fieri.