**Министерство образования и науки Российской Федерации**
**Московский физико-технический институт (государственный университет)**

Факультет управления и прикладной математики
Кафедра информатики

Выпускная квалификационная работа бакалавра по направлению
010900 «Прикладные математика и физика»

# Исследование различных методов разработки неблокирующих структур данных

Студент 576 группы
Пьянков С. А.

Научный руководитель
Бабичев С. Л.

Долгопрудный
2019

# Содержание

**1. Введение** **1**

**2. Постановка задачи** **3**

**3. Обзор источников** **4**

**4. Методология исследования** **13**

**Список литературы** **15**

**Благодарности** **17**

# Глава 1

# Введение

Современные процессоры уже весьма сложно ускорить так же, как это делалось десятки лет назад. Закон Мура, столь успешно описывавший увеличение числа транзисторов в процессорах с 1975 года, перестал соответствовать действительности уже к концу нулевых годов в силу атомарной природы вещества и ограничения скорости света. Возникает вопрос — неужели заставить работать вычислительную технику уже невозможно?

Производители процессоров подготовили свой ответ на этот вопрос, представив в начале нулевых первые многоядерные процессоры: IBM представила Power4 в 2001, Sun Microsystems представила UltraSparc IV в 2004 году, Intel и AMD представили свои двухъядерные процессоры в 2005 году. Параллелизм на уровне инструкций позволял даже старым однопоточным программам исполняться на новых процессорах быстрее без изменений в коде.

Таким образом, для получения всей выгоды от возросшей производительности ЦП алгоритмы должны разрабатываться соответствующим образом. Межпроцессное взаимодействие (`IPC`) в таких программах организуется двумя типами: через разделяемую память (`Shared Memory`) и с помощью сообщений (`Message passing`). Программы с разделяемой памятью имеют наибольший потенциал увеличения производительности, поскольку основной метод реализации межпроцессного взаимодействия с помощью сообщений — переключение контекста, которое занимает много времени.

Алгоритмы, работающие с разделяемой памятью, могут быть блоки-

рующими (с использованием мьютексов, семафоров и т.п.) и неблокирующими. Последние работают быстрее, но требуют корректной работы с памятью, поэтому важно умение правильным образом разрабатывать структуры данных для таких алгоритмов.

Данная работа является исследованием различных способов построения неблокирующих структур данных в параллельном программировании.

# Глава 2

# Постановка задачи

В этой работе изучаются разные реализации некоторых параллельных структур данных, сранивается их эффективность и отказоустойчивость:

- Изучение структур данных Стек (`Stack`) и Декартово дерево (`Cartesian tree` или `Treap`)
- Построение оптимистичной неблокирующей реализации Стека и Декартового дерева и анализ возможных ошибок в работе
- Использование `Hazard pointer` для построения корректно работающего неблокирующего стека и анализ применимости данного подхода для построения декартового дерева
- Анализ транзакционной память (`Intel TSX`) и использование `RTM (Redtricted Transactional Memory)` для построения неблокирующих структур данных
- Тестирование построенных структур данных и сравнение полученных результатов

# Глава 3

# Обзор источников

В книге Э.Таненбаума «Современные операционные системы» [8] описана история компьютеров и операционных систем. Среди прочего, Таненбаум рассказывает о том, как разработчики повышали производительность процессоров, уменьшая размеры транзисторов и увеличивая их количество. Автор подчеркивает, что со временем уменьшению размеров транзистора препятствуют законы физики — в силу вступают законы квантовой механики. Таким образом, Эндрю Таненбаум приходит к выводу о необходимости нового шага в развитии. Если раньше увеличивалось число функциональных блоков (собственно транзисторов), то теперь этого уже недостаточно; нужно дублировать и части управляющей логики. Это и есть неотъемлемая часть современных процессоров — многопоточность (`Multithreading` или `Hyperthreading` по версии `Intel`). В книге описаны различные проблемы, связанные с взаимодействием потоков (например, `race conditions` — состояние гонок потоков), в том числе сложности с совместной работой с данными, отсюда была осознана всю важность успешного построения структур данных в параллельных программах.

Ранее были упомянуты проблемы, возникающие при взаимодействии потоков. Для того, чтобы подробнее разобраться в этой теме, была изучена книга Э.Уильямса «Параллельное программирование на C++ в действии. Практика разработки многопоточных программ» [9]. Одна из тем, затронутых в книге, это разработка структур данных (с блокировками и без).

Для построения потокобезопасной структуры данных необходимо ис-

пользование атомарных типов данных и операций. Атомарные операции — неделимые, то есть операция либо выполнена окончательно, либо не выполнена совсем. В C++ атомарные операции возможно совершать только с данными атомарных типов (`std::atomic`). Кроме них, содержимое структуры данных защищено мьютексами и блокировками. В зависимости от строения структуры данных, может иметь место крупная и мелкая гранулярность (`coarse-grained` и `fine-grained`). Крупногранулярные структуры данных (например, список с блокировкой его головы) устроены так, что блокируются большие участки памяти. В мелкогранулярных структурах данных (например, список с отдельной блокировкой каждого элемента списка) блокируются более мелкие участки памяти, что несколько ускоряет параллельный доступ к объекту. Однако, и такие структуры данных работают медленно и параллельный код дает слабый выигрыш в производительности.

В структуре данных без блокировок не используются мьютексы или семафоры. Это позволяет избежать ряда проблем, присущих структурам данных с блокировками, однако усложняет написание структуры данных. Неблокирующая структура данных становится открытой для одновременного доступа со стороны сразу нескольких потоков (при этом необязательно эти потоки будут совершать одинаковые операции). Корректно построить такую структуру данных довольно сложно, на каждом шаге требуется, чтобы хоть какой-то поток продвигался вперед. Среди неблокирующих структур данных выделяют структуры данных, свободные от ожидания. На них наложено еще более серьезное ограничение: каждый поток должен завершить свою работу со структурой данных за ограниченное число шагов, независимо от работы других потоков. То есть, на каждом шаге должны продвигаться все потоки, ни один из них не ждет других.

Однако, написание неблокирующих структур данных — процесс достаточно тяжелый, к тому же у таких структур данных имеются свои недостатки. Несмотря на то, что они позволяют лучше распараллелить операции и сократить время ожидания, ничто не гарантирует нам повышение производительности программы. Это связано с тем, что атомарные операции, которые будут часто использоваться в неблокирующей

структуре данных, исполняются достаточно долго, к тому же возможно взаимодействие между потоками описано неоптимально. Опираясь на эти выводы, было принято решение рассмотреть две структуры данных — стек и декартово дерево (`Treap`), и построить их неблокирующие реализации.

В книге Херлихая «Искусство многопроцессорного программирования» [3] исследованы параллельные структуры данных очередь и стек. Наиболее наивная неблокирующая реализация структур данных основана на оптимистическом подходе. Поток, который будет работать с данными, делает снимок текущего состояния структуры данных, модифицирует данные, проверяет перед записью, не изменилась ли за это время структура, и, если не поменялась (в надежде на это и состоит оптимизм), то записываем модифицированные данные. В случае, если сторонний поток за это время уже внес какие-либо изменения, начинаем работу заново, то есть снова делаем снимок текущего состояния и т.д. Чтобы проверить, изменилась ли структура, используется операция `CAS (Compare And Swap)`. В качестве аргументов операции передаются сохраненные копии изначальных данных, указатель на структуру данных (то есть на текущее ее состояние) и модифицированные текущим потоком данные, которые надо вставить в структуру данных.

Но такой подход несет в себе опасность — структура данных будет подвержена ошибке `ABA`. Это ошибка, которая возникает при работе с памятью в структуре данных, когда ячейка памяти читается дважды и дважды считывается одно и то же значение, что интерпретируется будто изменения не были внесены. Однако второй поток мог между двумя считываниями изменить значение, работать и восстановить предыдущее значение в ячейке памяти.

Рассмотрим проблему `ABA` применительно к неблокирующему стеку. Изначально стек содержит $head \longrightarrow A \longrightarrow B \longrightarrow C$. Первый поток выполняет операцию `pop`, которая возвращает A, head должен указывать на B. Одновременно с этим операцию `pop` дважды выполняет второй поток, при этом он успевает сделать `CAS` раньше первого потока, поэтому в стеке сначала будет $head \longrightarrow B \longrightarrow C$, затем $head \longrightarrow C$. Затем второй поток добавляет A в стек, превращая его в $head \longrightarrow A \longrightarrow C$.

Затем уже первый поток продолжает работу (стоит напомнить, что первому потоку осталось выполнить `CAS`, заменив указатель head с A на B. Поскольку в текущем состоянии стек указывает на A, то операция `CAS` успешно исполнится, хотя сам стек уже изменился и следует откатиться к предыдущему состоянию и повторить выполнение операции `pop` первым потоком. Вместо этого элемент A будет заменен на B, стек будет иметь вид $head \longrightarrow B \longrightarrow C$, что не соответствует действительности.

Более подробно проблема `ABA` описана в книге . В книге А.Г. Тормасова «Параллельное программирование многопоточных систем с разделяемой памятью» [7] помимо основных аспектов работы с разделяемой памятью, таких как причины условий гонок и необходимость использования атомарных операций, более подробно разобраны проблемы, возникающие при работе с разделяемой памятью, в том числе проблема `ABA`. Кроме описания проблемы, в книге приведен один из способов ее решения — `RCU`. `RCU (Read, Copy, Update)` — алгоритм чтения, копирования и обновления. Для изменения данных писатель делает себе полную копию, меняет эту копию и атомарно меняет указатель на свою копию. Писатель сначала находится в фазе `Removal phase`, в рамках которой происходит изменение структуры данных без непосредственного удаления элемента; затем переходит в `Grace Period start phase`, которая объявляет о начале `Grace Period`, промежутка, в рамках которого все потоки не находятся в критической секции; следующая фаза — `Grace Period end waiting phase`, в которой происходит ожидание окончания `Grace Period`; все заканчивается `Reclamation Phase` — фазой, в которой писатель коммитит изменения. Однако, замена указателя должна происходить «в удобное для всех потоков время», то есть нужно дождаться выполнения кода на каждом потоке и быть уверенным, что нигде нет неправильной ссылки на старую область памяти. При большом числе потоков замена указателя может откладываться очень надолго, что приводит в дальнейшем к долгой потере работоспособности системы. На уровне ядра операционной системы проблема решается — управлением занимается планировщик потоков, а поток-писатель находится в ожидании команд от процессора. Но структуры данных, требующие использование `RCU`, находятся в пользовательском адресном пространстве (`user-space`), в ко-

тором явный вызов примитивов планировщика недоступен. Приходится прибегать к использованию так называемых user-space RCU, которые представляют собой совершенно нетривиальные структуры с большим объемом кода и использование такого механизма приводит к большой потере времени исполнения, делая борьбу с блокировками непрактичной.

Среди других способов решения проблемы ABA — Garbage Collector, Reference Counting, Double-length CAS, LL/SC. Garbage Collector — весьма простой подход к решению проблемы ABA, заключается в хранении указателей, по которым «в удобный момент времени», то есть когда только никто не будет использовать данные по указателям, будет высвобождаться память. Без блокировок такой механизм не реализуется, следовательно для написания неблокирующих структур данных он не подходит. Часто в Garbage Collector используются Reference Counter — счетчики ссылок на удаляемый объект. Такие счетчики можно использовать и без Garbage Collector, но без блокировок операции с ними трудно реализуются и эффективность lock-free структуры данных резко снижается. Замена CAS на Double-length CAS достаточно элегантно решает проблему ABA, храня во второй части счетчик. Сравнение тогда ведется по предыдущему значению и счетчику с текущим значением и счетчиком. При совпадении — происходит обмен (swap). В случае возникновения проблемы ABA значения совпадут, а счетчики — нет, то есть в таком случае CAS не выполнится. Если для 32-битных машин Double-length CAS - это 64-битный CAS и современные устройства способны обеспечить их поддержку, то большинство 64-битных машин не поддерживают 128-битный CAS, что не позволяет считать Double-length CAS эффективным решением проблемы ABA. В отличие от другого примитива аппаратно синхронизации — LL/SC (Load linked/Store conditional). Это пара инструкций, первая из которых возвращает значение ячейки памяти, второе записывает туда свое значение тогда и только тогда, когда между выполнением второй и первой инструкций не было других записей в ячейку памяти. При всех плюсах такого механизма (в отличие от CAS, LL/SC не подвержен проблеме ABA), эти инструкции не реализованы в архитектуре Intel.

В качестве оптимального решения проблемы ABA был выбран механизм, который называется «Опасные указатели» (Hazard pointers). Он был предложен в статье Maged M. Michael «Safe memory reclamation for lock-free objects» [4]. Основная идея метода состоит в следующем: каждый поток сохраняет фиксированное число опасных указателей (как правило один или два), указывающий на данные, которые, возможно, будут изменены. Каждый такой опасный указатель может быть записан только владеющим им потоком, однако читать данные по этому указателю могут все потоки (режим «один писатель — много читателей»). Если какой-то поток хочет удалить указатель, то этот указатель помещается в особый список, из которого указатели удаляются в тот момент, когда они перестанут быть опасными для всех потоков. Эффективность данного подхода заключается в том, что он занимает дополнительно всего лишь константный фиксированный промежуток времени для каждого изменяемого объекта. Hazard pointers не только работают без блокировок (lock-free), но и без ожидания (wait-free), то есть гарантируется прогресс каждого потока.

Применение опасных указателей в построении неблокирующих структур данных довольно подробно разобрано в статье А. Александреску «Неблокирующие структуры данных с опасными указателями» [1]. В качестве примера структуры данных в статье взята WRRMMap (класс, в котором хранится указатель на однопоточный std::map и обеспечен многопоточный неблокирующий доступ к нему). Автор описывает саму структуру опасных указателей, устройство списка указателей «на удаление» и алгоритм по сканированию опасных указателей. Затем, с помощью операций (Update) и (Lookup), Александреску встраивает опасные указатели в структуру данных. Автор доказывает, что данные операции неблокирующие (lock-free); кроме того, в статье приведены указания, как надо изменить операции, чтобы они стали свободными от ожидания (wait-free). Пользуясь материалами из данной статьи, можно смело приступать к построению произвольных неблокирующих структур данных.

Однако непосредственно реализация опасных указателей не входила в планы данной работы, поэтому было принято решение искать готовые

реализации данного механизма. В библиотеке `libcds` [2] представлены методы безопасного освобождения памяти, в том числе опасные указатели (`Hazard pointers`). Они реализованы в виде синглтона, к которому нужно подключать все создаваемые потоки. При помощи методов этой библиотеки была создана неблокирующая реализация стека, не подверженная проблеме `ABA`.

Стек — относительно простая структура данных, имеющая всего один указатель. При построении сложных структур данных, состоящих из нескольких указателей, атомарное обновление каждого указателя по отдельности не имеет смысла (при отсутствии блокировок), а совместное обновление затруднительно. В связи с этим встает вопрос о разработке неблокирующей реализации сложной структуры данных. В качестве примера подобной структуры данных было взято декартово дерево (`cartesian tree` или `Treap`). Эта структура данных была описана в статье Seidel, Aragon «Randomised search trees» [5]. Эта структура данных объединяет в себе бинарное дерево поиска и бинарную кучу. Говоря более строго, эта структура данных хранит пары $(x, y)$ так, что является бинарным деревом по элементам $x$ и бинарной пирамидой по элементам $y$. Если некоторый элемент дерева содержит $(x_0, y_0)$, то у всех элементов в левом поддереве $x \leq x_0$, а у всех элементов в правом поддереве $x \geq x_0$, а также и в левом, и в правом поддереве $y \leq y_0$. Если выбирать приоритеты случайно, то декартово дерево станет рандомизированным бинарным деревом поиска.

Основные операции, через которые реализуется работа с декартовым деревом, это `split` и `merge`. Операция `split` разбивает дерево на два поддерева по ключу так, чтобы в первом поддереве были элементы с меньшим ключом, а во втором — элементы с большим ключом. Реализация такой функции будет, очевидно, основана на рекурсии: если ключ совпадает с ключом корня, то результатом будут левое и правое поддеревья. Если ключ больше ключа корня, то корнем первого поддерева будет корень исходного дерева, а корень второго поддерева появится в результате применения операции `split` к правому поддереву. Все выполняется симметрично, если ключ меньше ключа корня.

Операция `merge` сливает два дерева в одно, сохраняя при этом струк-

туру (по ключам и приоритетам). Предполагается, что оба дерева, которые передаются функции, обладают соответствующим порядком, то есть все ключи в первом дереве меньше, чем ключи во втором. Тогда сравниваются приоритеты корней первого и второго деревьев, если приоритет первого выше, то он и будет являться корнем. Тогда выполняем операцию `merge`, передав в качестве аргументов правое поддерево первого дерева и второе дерево. Если приоритет корня второго дерева выше, то выполняем все симметрично.

Поскольку каждый элемент декартова дерева содержит два указателя, которые, при изменении структуры данных, необходимо одновременно атомарно обновлять, то опасные указатели (`Hazard Pointers`) не могут обеспечить корректную работу. В поиске подлходящих механизмов снова пришлось обратиться к книге Тормасова [7]. В ней упоминается подход, называемый транзакционной памятью, суть которого в следующем: любые операции с памятью, выделенные в специальный блок, выполняются транзакционно, то есть или выполняются все, или не выполняется ни одна. Существенное ограничение в работе с транзакционной памятью - все операции внутри транзакции должны быть откатываемы. Выделение памяти, обмен данными со внешними устройствами или с другими процессами, вывод на экран — подобные этим операции не могут выполняться внутри транзакции.

`G++` поддерживает транзакционную память (`STM`), начиная с версии 4.7. Идея программного подхода к транзакционной памяти состоит в том, чтобы выделить участок кода, который будет внутри транзакции (для `g++` код нужно писать внутри фигурных скобок `__transaction_atomic{}`). Однако, максимальную эффективность дает аппаратная поддержка транзакционной памяти, которую `Intel` встроила в свои процессоры в 2013 году (`Intel TSX`). Документация Intel [6] предлагает два набора команд: `HLE` и `RTM`. `HLE (Hardware Lock Elision)` предоставляет две инструкции - `XACQUIRE` и `XRELEASE`. В случае, если начать транзакцию не получается, система автоматически пробует сделать это снова (выполнить `XACQUIRE`), таким образом, нельзя задать алгоритм действий на случай, если транзакция не завершилась успешно. В `RTM` такой проблемы нет, поэтому в данной работе будет использоваться именно этот набор команд. `RTM`

включает в себя три инструкции: `XBEGIN`, `XEND`, `XABORT`. `XBEGIN` и `XEND` отмечают начало и конец транзакции, а `XABORT` прерывает транзакцию, при этом в регистре `eax` содержится статус транзакции. Проверить этот статус можно с помощью инструкции `XTEST`.

# Глава 4

# Методология исследования

В данной работе разрабатываются и исследуются разные имплементации структур данных стек и декартово дерево (`Treap`). Реализованы эти структуры на языке `C++`, который является стандартом для системного программирования и для которого есть большое количество эффективных компиляторов и библиотек. В версию `C++11` была внедрена поддержка многопоточного программирования (`std::thread`), что является ключевым для данного исследования.

Компилятором для исходников в данной работе был выбран `g++`. Некоторые реализации структур данных компилируются только этим компилятором, например, `Software Transactional Memory` поддерживает «из коробки» только `g++`, начиная с версии 4.8. Используется компьютер с процессором `Intel Core i5 7267U, Kaby Lake`, который поддерживает аппаратную транзакционную память (`Intel TSX`).

Для наиболее точных замеров времени работы используется ассемблерная инструкция `rdtsc`, которая возвращает число тактов с момента последнего сброса процессора. Разница двух величин (замеров до начала тестирования структуры данных и после начала) дает время (в тактах процессора) работы структуры данных.

Тестирование структуры данных заключается во вставке и удалении большого количества данных. Для чистоты эксперимента, данные генерируются случайным образом и операции вставки и удаления тоже

вызываются случайно; чтобы минимизировать шанс удаления элемента из уже пустой структуры данных, изначально вставляется относительно большое число данных (в замерах времени эта порция вставок не участвует).

Для декартова дерева необходимо, чтобы ключи были уникальными, для этого заранее генерируется множество ключей (`unordered_set`). Это множество «тасуется», то есть используется алгоритм Фишера-Йетса, который перемешивает это множество ключей. Тогда выбирая ключ из этого множества и создавая случайный приоритет, получаем значение, которое вставляем в декартово дерево.

//ПРОВЕРКА СТАТИСТИКИ — критерий Пирсона? пользоваться готовым из электронной таблицы?

# Список литературы


[1] Andrei Alexandrescu, Maged Michael. “Lock-free data structures with hazard pointers”. *C++ User Journal* (2004), с. 17—20.

[2] *C++ concurrent data structure library*. URL: `http://libcds.sourceforge.net`.

[3] Maurice Herlihy, Nir Shavit. *Искусство многопроцессорного программирования*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2008. ISBN: 0123705916, 9780123705914.

[4] Maged M. Michael. “Hazard Pointers: Safe Memory Reclamation for Lock-Free Objects”. *IEEE Trans. Parallel Distrib. Syst.* **15** 6 (июнь 2004), с. 491—504. ISSN: 1045-9219. DOI: `10.1109/TPDS.2004.8`. URL: `http://dx.doi.org/10.1109/TPDS.2004.8`.

[5] R. Seidel, C. R. Aragon. “Randomized search trees”. *Algorithmica* **16** 4 (окт. 1996), с. 464—497. ISSN: 1432-0541. DOI: `10.1007/BF01940876`. URL: `https://doi.org/10.1007/BF01940876`.

[6] Richard M. Yoo и др. “Performance Evaluation of Intel&Reg; Transactional Synchronization Extensions for High-performance Computing”. *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*. SC ’13. Denver, Colorado: ACM, 2013, 19:1—19:11. ISBN: 978-1-4503-2378-9. DOI: `10.1145/2503210.2503232`. URL: `http://doi.acm.org/10.1145/2503210.2503232`.

[7] Тормасов А.Г. *Параллельное программирование многопоточных систем с разделяемой памятью*. Физматкнига, 2014. ISBN: 9785891552357.

[8] Э.С. Таненбаум. *Современные операционные системы: [перевод]*. Классика computer science: КС. Питер, 2010. ISBN: 9785498073064. URL: `https://books.google.ru/books?id=60p9kgEACAAJ`.

[9] Э. Уильямс. *Параллельное программирование на C++ в действии. Практика разработки многопоточных программ.* ЛитРес, 2017. ISBN: 9785457427020. URL: https://books.google.ru/books?id=1UXRAAAAQBAJ.

# Благодарности

Благодарности идут тут.