Министерство образования и науки Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«Уральский федеральный университет
имени первого Президента России Б.Н. Ельцина»

Институт естественных наук и математики
Департамент математики, механики и компьютерных наук

Исследование потокобезопасных неблокирующих структур данных

«Допущен к защите»

Директор департамента
к.ф.-м.н., доцент
Асанов М. О.

«___»_____2017 г.

Квалификационная работа на соискание степени бакалавра наук по направлению «Фундаментальная информатика и информационные технологии»
студента группы ФТ-401 (МЕН-430802)
Сваловой А. А.
Научный руководитель:
Ассистент департамента к.ф.-м.н.
Плинер Ю. А.

Екатеринбург

2017 год

Содержание

1 Введение 3

2 Глава 1. Основные определения 7

2.1 Атомарная операция 7

2.2 Atomic markable reference 8

2.3 Set 9

3 Глава 2. Реализации алгоритмов 11

3.1 Односвязный список 11

3.2 Улучшенный односвязный список 15

3.3 Список с пропусками 17

3.4 Хэш-таблица 17

4 Глава 3. Тестирование 19

4.1 Модульное тестирование 19

4.2 Тестирование производительности 20

4.3 Результаты 21

5 Заключение 22

Список литературы 23

6 Приложения 24

1 Введение

С ростом прогресса многие электронные устройства становятся многоядерными, появляются многопроцессорные устройства. Поэтому задача программиста, как человека, который пытается максимально хорошо использовать предоставленные ресурсы, - писать программы, способные масштабироваться и параллелиться. Поэтому сейчас все чаще и чаще пишут многопоточные программы и используют многопоточные структуры данных. Такие программы могут выполнять сразу несколько инструкций на каждом процессоре или ядре, однако такой код обладает рядом проблем, связанных с доступом к общим ресурсам разными потоками или процессами.

Если два или более потока захотят изменить один и тот же участок памяти, то они попытаются сделать это единовременно. После выполнения операции неизвестно, как будет выглядеть этот участок памяти, так как порядок выполнения инструкций разных потоков неопределен. Возникает вопрос: как в таком случае контролировать доступ к этому ресурсу? Хочется чтобы в каждый момент времени, способом, очевидным для разработчика, ресурсом владел только один поток, а все остальные каким-то образом ждали своей очереди. Такое поведение можно осуществить несколькими способами.

Самый простой из них - блокировка. Каждый раз, когда поток хочет сделать что-то с ресурсом, он проверяет, нет ли блокировки на этот ресурс. Если есть, поток ждет, пока блокировка не освободится, если нет, то он пытается первым захватить блокировку. В случае успешного захвата он осуществляет все операции с ресурсом и освобождает блокировку. В это время все остальные потоки ждут этот и ничего не делают.

Такой механизм синхронизации очень прост в понимании и реализации,

учитывая существование встроенных блокировок в большинство современных ОС. Также этот способ, очевидно, позволяет только одному потоку единовременно получить доступ к ресурсу. Однако, в данном способе существует и масса проблем, которые сводят на нет все преимущества. Во-первых, при большом количестве потоков, желающих получить доступ к ресурсу, возникает «узкое горлышко», т. е. место в программе, которое тормозит выполнение программы в целом. Во-вторых, при существовании больших участков программы с блокировкой теряется весь смысл многопоточности. В эти участки все равно может заходить только один поток, как и в однопоточном программировании. В-третьих, существуют некоторые особенности операционной системы: переключение потоков - дорогая операция. При долгом ожидании освобождения ресурса происходит очень большое количество переключений, следовательно, большое количество времени тратится на бесполезные операции. В-четвертых, возможны ситуации, когда один поток захватил первый ресурс и ждет освобождение второго ресурса, в то время как второй поток захватил второй ресурс и ждет освобождения первого. Такая ситуация называется взаимная блокировка (deadlock). Программа в таком случае останавливает свое выполнение совсем и не может без каких-либо вмешательств извне разрешить эту ситуацию.

Эти проблемы привели исследователей к созданию других способов синхронизации. Один из них - неблокирующая синхронизация. Это способ, при котором каждый поток пытается применить низкоуровневые атомарные аппаратные примитивы, а не использовать блокировки. Таким образом в каждый момент времени выполняется только одна операция, только одного потока. Все остальные операции в других потоках либо завершаются ошибкой, либо

выполняются сразу следом за предыдущей. Такие алгоритмы обеспечивают общее продвижение программы в целом: даже если какой-то поток не смог выполнить операцию или завершился с ошибкой - значит, что какой-то другой поток успешно выполнил свою операцию. Не существует случаев, когда все потоки одновременно простаивают, и как частный случай этого, невозможно существование взаимных блокировок.

Однако, несмотря на все преимущества, данная область является до сих пор развивающейся. Нельзя просто взять и написать неблокирующую реализацию алгоритма, основанного на блокировках. В некоторых случаях это оказывается легко, в некоторых до сих пор не придумано неблокирующих аналогов. Причина: каждый раз нужно творчество, чтобы свести все операции над разделяемым ресурсом к последовательности независимых атомарных операций, т. е. не существует универсального способа написания неблокирующей реализаци. Однако сложность реализации и изобретения алгоритма часто стоит усилий. Пусть этот класс алгоритмов совсем не о скорости работы, а о гарантии продвижения системы в целом, но в итоге большинство неблокирующих алгоритмов имеют в среднем ожидаемую сложность меньше, чем блокирующие аналоги. Но это только в теории. На практике скорость работы зависит от конкретной реализации, области применения, часто встречающихся запросов и т. д.

Объект исследования данной работы - потокобезопасные неблокирующие алгоритмы структур данных. Цель - реализовать основные структуры данных, реализующие интерфейс ISet и на практике выявить являются ли неблокирующие алгоритмы эффективней блокирующих, какие алгоритмы вообще реально применимы, и выяснить, как адаптировать алгоритмы, разработан-

ные под языки программирования с неуправляемой памятью, к языкам с управляемой памятью.

В работе представлены структуры данных, реализующие интерфейс ISet. Данный интерфейс включает в себя добавление элемента в множество, удаление элемента, а также поиск и перечисление всех элементов в множестве. Этих сценариев достаточно, чтоб понять, как ведут себя различные реализации на практике. Для сравнения были выбраны следующие реализации: сортирующийся лист, хэш-таблица, скип-лист и дерево поиска. Также взяты готовые реализации всех этих структур из библиотеки языка C#, чтобы сравнить неблокирующие реализации с блокирующими. В приложении приведены различные результаты сравнений всех этих структур и вариации использования их в реальной жизни. Все алгоритмы адаптированы под язык C# и собраны в один общий модуль с внешним интерфейсом ISet.

2 Глава 1. Основные определения

2.1 Атомарная операция

Все неблокирующие алгоритмы можно разделить на три типа: Waitfree, Lockfree, Obstruction-free[1].

В первом типе каждый поток совершает каждую операцию за конечное число шагов, независимо от влияния других потоков. Это самое сильное требование из-за чего редко реализуемое. Такие алгоритмы обычно реализуют атомарный инкремент или атомарную замену ссылок.

Во втором типе система в целом двигается вперед, даже если какой-то поток стоит на месте. Если какой-то поток не смог выполнить операцию, значит, что какой-то другой поток смог выполнить свою операцию, следовательно, в целом система продвинулась. Эти алгоритмы обычно реализуют атомарное сравнение и замену.

В третьем типе каждый может выполнить каждую операцию за конечное количество шагов, если ничего ему не помешает. В данном случае может случиться ситуация, когда ни один из потоков не двигается вперед, однако ни один заблокированный поток не может мешать работе всех остальных потоков, следовательно, это все равно более сильная гарантия, чем блокирущие алгоритмы.

Каждая из этих реализаций использует абстракцию «атомарная операция» - это операция, которая либо не выполняется совсем, либо выполняется как единое целое. В данной работе используется атомарная операция Compare And Swap (CAS) (Рис 1). Эта операция сравнивает две ссылки и, если они равны, меняет одну из них на новую. Эта операция предоставляется

большинством операционных систем и уже встроена в язык C#.

```
CAS(ref reference, newReference, comporand)
```

Рис. 1: CAS сравнивает reference с comporand и, если они равны, заменяет reference на newReference

2.2 Atomic markable reference

Алгоритмы с неблокирующей синхронизацией оказываются в разы сложнее обычных алгоритмов. Зачастую, они зависят от конкретной реализации или конкретного языка программирования. Иногда они полагаются на сборщика мусора (абстракцию в языках программирования с управляемой памятью) или, наоборот, его отсутствие. Кроме того, доказать корректность таких алгоритмов бывает очень сложно. Поэтому со временем стали придумывать не только алгоритмы с неблокирующей синхронизацией, но и комбинации, где особо часто используемые операции производятся без блокировок, а некоторые операции производятся с блокировками, но на маленькие участки памяти. Такие алгоритмы оказываются проще в понимании и доказательстве, но не проигрывают в эффективности и применимости.

Один из способов такой локальной блокировки - это добавление особого маркера в ссылку на объект. Так, если один поток смог атомарно изменить ссылку на некий объект, пометив ее этим маркером, все остальные потоки понимают, что данный объект используется в какой-то операции и его нельзя изменять.

В языках с неуправляемой памятью такой способ легко осуществим благодаря выравниванию указателей на объект. При выделении памяти компилятор, обычно, выравнивает длину указателя на максимально большой тип

данных. Поэтому в указателе остаются реально неиспользуемые биты, которые можно как раз и использовать в качестве маркера.

В языках с управляемой памятью разработчик не имеет доступа к ссылке, поэтому стоит придумывать способы сымитировать эту ссылку с помощью объектов. В данной работе реализован примитив маркируемой ссылки (Atomic Markable Reference). Он будет хранить в себе текущее состояние этой ссылки (Рис 2). Само же состояние будет состоять из непосредственно ссылки и описанного выше маркера. В данном случае маркер как раз отвечает за неиспользуемые биты указателя. Теперь можно атомарно изменять ссылку на состояние, что равносильно изменению помеченного указателя. Изменяя этот примитив, можно сымитировать изменение указателя, что позволяет также использовать алгоритмы с локальными блокировками.

```
public class AtomicMarkableReference<TReference, TMark>
{
        State state;

        private class State
        {
                TReference Reference;
                TMark Mark;
        }
}
```

Рис. 2: AtomicMarkableReference

2.3 Set

Set - это коллекция для хранения неупорядоченного множества уникальных объектов. Это аналог математического понятия множество Эта структура данных позволяет добавлять элементы, удалять элементы и быстро проверять, существует ли уже такой элемент (Рис 3).

```
public interface ISet<TElement>
{
        bool Add(TElement element);
        bool Contains(TElement element);
        bool Remove(TElement element);
}
```

Рис. 3: Интерфейс ISet

Чаще всего set применяют для объединения объектов с какими-то общими свойствами. Примером использования сета может служить множество уникальных пользователей форума, множество допустимых тегов или множество запрешенных логинов пользователей.

Хотя операции в set могут быть реализованы произвольным способом, однако чаще всего его используют для быстрой проверки принадлежности элемента (черные, белые списки, базы данных). Поэтому в данной работе будет сделан акцент на быстром поиске. Быстрое добавление тоже будет играть роль, поэтому придется пожертвовать скоростью удаления.

3 Глава 2. Реализации алгоритмов

В данной главе будут приведены краткие описания неблокирующих алгоритмов, проблемы, которые они решают и сложности реализации. Полное описание алгоритмов можно найти в списке литературы [2], [3], [4], [5], [6]. Стандартные реализации однопоточных алгоритмов общедоступны, поэтому не будут описаны в данной работе. Познакомиться с ними можно, например, в книге «Структуры данных и алгоритмы» [7].

3.1 Односвязный список

Пусть односвязный список состоит из элементов, в каждом из которых есть значение этого элемента и ссылка на следующий элемент (Рис 4). Поиск элемента не будет рассмотрен, так как он совпадает с поиском в обычном сортирующимся списке.

```
public class LinkedListNode<TElement>
{
        TElement Element;
        LinkedListNode<TElement> Next;
}
```

Рис. 4: Вершина списка

Если реализовать добавление как в однопоточном варианте, то возможна проблема при одновременном добавлении двух последовательных элементов. Пусть, есть список с элементами 1-2-5 (Рис 5(a)). Поток А хочет вставить элемент 3, поток Б - элемент 4. Поток А понимает, что ему нужно вставить элемент между 2 и 5. Он запоминает ссылку на предыдущий и следующий элементы и в этот момент операционная система передает управление потоку Б (Рис 5(b)). Поток Б также находит место для вставки и тоже запомина-

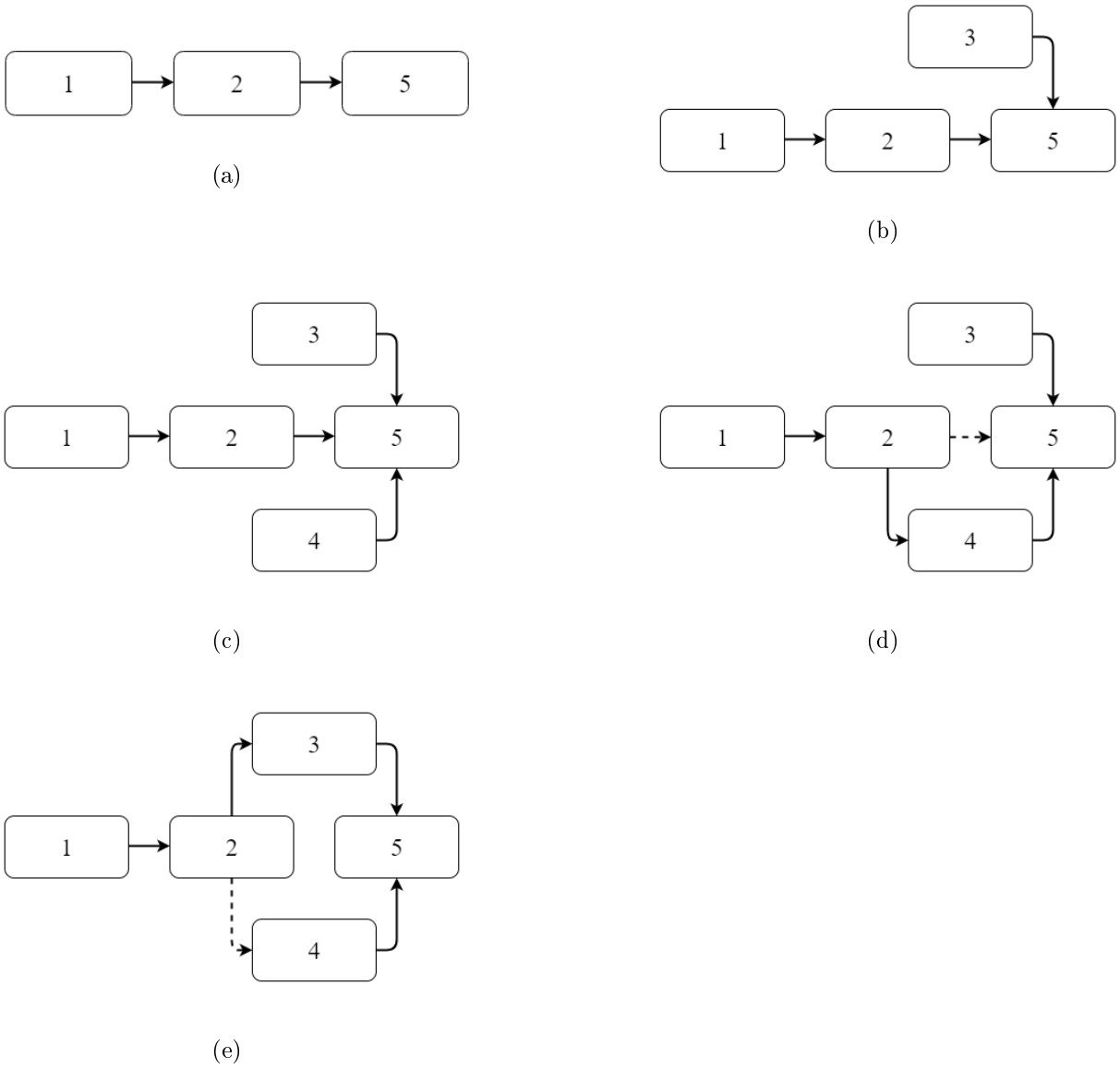

Рис. 5: Одновременное добавление в 2 потоках: a) начальное состояние, b) А поставил ссылку на 5, c) Б поставил ссылку на 5, d) А сменил ссылку «Next» у 2, e) Б сменил ссылку «Next» у 2.

ет 2, как предыдущий элемент, 5, как следующий (Рис 5(c)). После этого он переписывает ссылку «Следующий» у элемента 2 на новосозданный элемент 4, а у элемента 4 на 5 (Рис 5(d)). Управление возвращается к потоку А. Он перезаписывает ссылку «Next» предыдущего элемента (2) на новосозданный элемент 3, а ссылку элемента 3 на следующий элемент (5) (Рис 5(e)). В результате элемент 4 «потеряется», т. е. не будет ни одной ссылки, указывающей

на него.

Реализация неблокирующего доступа использует типичный прием для неблокирующих алгоритмов - вечный цикл с операцией CAS. На каждом шаге цикла алгоритм пытается найти два элемента а и б, между которыми должен быть вставлен новый, и атомарно перезаписать ссылку «Next» с предыдущего элемента (а) на новый, при этом сравнивая, является ли эта ссылка до сих пор ссылкой на следующий (б). Алгоритм выходит из цикла, когда попытка замены ссылки происходит успешно (Рис 6). Такая реализация полностью решает вышеописанную проблему. При попытке перезаписать ссылку элемента 2 с 5 на 3 (Рис 5(e)), CAS не проходит, потому что ссылка уже не на 5, а на 4. Алгоритм заново находит соседние элементы, и следующий уже не 5, а 4. На этом шаге цикла CAS уже выполняется успешно. Оба элемента вставлены правильно.

```
while (true)
{
        1) (predsessor, subsessor) = FindPlace(newElement)
        2) newElement.Next = subsessor
        2) if (CAS(ref predsessor.Next, newElement, subsessor))
               break
}
```

Рис. 6: Шаг цикла. 1) находим соседние элементы, между которыми нужно вставить новый, 2) ссылку Next у нового элемента поместим на subsessor 3) пытаемся атомарно вставить

Еще одна проблема может возникнуть при одновременным удалении и вставке двух последовательных элементов. Пусть есть список 1-2-4 (Рис 7(a)). Поток А хочет добавить элемент 3, поток Б удалить элемент 2. Поток Б запоминает, что предыдущий элемент 1, следующий 4. Управление передается потоку А. Поток А вставляет элемент 3, как это было описано ранее (Рис 7(b)). Управление возвращается к потоку Б. Он атомарно заменяет ссылку «Next»

у элемента 1 на элемент 4 (Рис 7(c)). В итоге элемент 3 «потерялся».

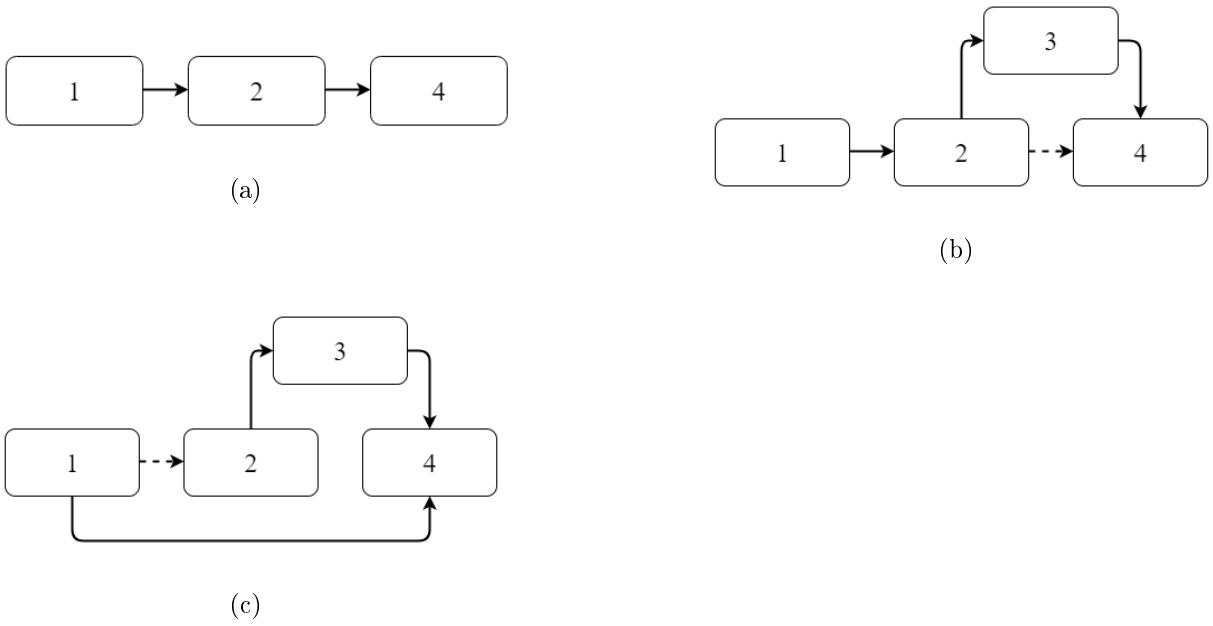

Рис. 7: Одновременное добавление и удаление в 2 потоках: a) начальное состояние, b) А поставил ссылку на 4, c) Б сменил ссылку с 2 на 4.

Для устранения этой проблемы можно ввести дополнительный флаг в ссылку на объект (Рис 8).

```
public class LinkedListNode<TElement>
{
        TElement Element;
        AtomicMarkableReference<LinkedListNode<TElement>, Flag> Next;
}
```

Рис. 8: AtomicMarkableReference - структура, описанная в предыдущей главе, Flag может быть никаким или помеченным

Теперь объект удаляется в два шага:

- пометить как удаленный, но не удалить
- физически удалить.

Можно заметить, что при помечивании ссылки на удаленный элемент, ситуация, изображенная на рисунке 2, существенно не изменится. Однако, при по-

мечивании ссылки «Next» у удаляемого объекта, можно избежать петерь элементов. Теперь при вставке тройки из предыдущего примера ссылка «Next» у 2 уже будет помеченной. Это будет сообщать о том, что элемент в данный момент удаляется, а значит, манипулировать этой ссылкой пока что нельзя, надо заново перейти на новый виток в цикле и заново определить соседей. В итоге проблемы, описанной ранее при одновременной вставке и удалении не случится.

3.2 Улучшенный односвязный список

Вышеописанная реализация односвязного списка является неблокирующей, что, возможно, может ускорить работу программы, однако у нее до сих пор существует недостаток: если операции удаления происходят достаточно часто, то операции вставки будут также часто заканчиваться не успехом, из-за чего они каждый раз начинать сначала. В результате в худшем случае может получиться, что программа каждый раз заново пробегает весь список.

Чтобы устранить эту проблему можно ввести еще две дополнительных абстракции. В ссылку «Next» добавить еще один флаг, который будет свидетельствовать, что следующий элемент в данный момент на стадии удаления. В сам элемент нужно добавить поле «Backlink», который будет указывать на предыдущий элемент, который еще не участвует в удалении (Рис 9). Теперь

```
public class LinkedListNode<TElement>
{
        TElement Element;
        LinkedListNode<TElement> Backlink;
        AtomicMarkableReference<LinkedListNode<TElement>, Flag> Next;
}
```

Рис. 9: AtomicMarkableReference - структура, описанная в предыдущей главе, Flag может быть никаким, или помеченным на удаление, или помеченным на невозможность удаления

операция удаления будет проходить не в два, а в три этапа. Между двумя этапами из предыдущего алгоритма появится новый этап. Теперь после помечивания удаляемой вершины на удаление (Рис 10(a)) алгоритм добавляет в ссылку «Next» у предыдущей вершины новый флаг, который будет обозначать, что в данный момент вершина участвует в удалении, и ее саму удалять нельзя. У удаляемой вершины алгоритм устанавливает ссылку «Backlink» на ближайшую предыдущую вершину, которая еще не помечена новым флагом (Рис 10(b)). Теперь каждый раз, когда вставка не может завершиться успехом, поток будет по ссылкам «Backlink» возвращаться не в самое начало, а в первую вершину, следующая за которой еще не удаляется. Это позволяет еще немного ускорить работу программы, так как при каждой неудачной вставке, возможно, больше не нужно проходить лист полностью заново.

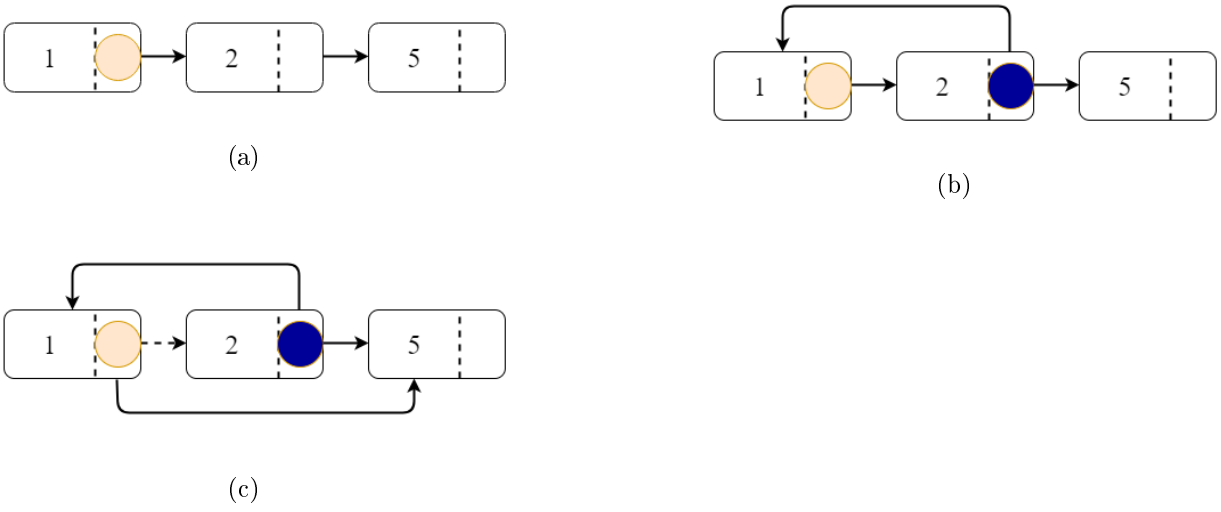

Рис. 10: Удаление в 3 этапа: a) помечивание на участие в удалении, b) помечивание на удаление, c) реальное удаление.

3.3 Список с пропусками

Список с пропусками в своей структуре содержит несколько самосортирующихся односвязных списков. Поэтому алгоритм неблокирующего списка с пропусками будет использовать все те же идеи, что и неблокирующий односвязный список. Остается разобраться, как применить все те же идеи, но вставляя и удаляя не 1 элемент, а сразу столбец.

Вставка, как и в однопоточном варианте осуществляется снизу-вверх. Однако, в данном случае на каждом уровне приходится искать заново, иначе можно запомнить элемент, который какой-то другой поток уже удалил.

Удаление тоже происходит, начиная с удаления вершины на самом нижнем уровне. Этого действия достаточно, чтобы весь столбец считался удаленным. При каждом следующем поиске по списку нужно проверять не удалена ли текущая вершина, а удалена ли ее вершина с первого уровня (вершины реально не удаляются из памяти, но на них больше никто не ссылается, поэтому можно считать, что они больше не принадлежат к списку, так как они недостижимы). Если вершина с первого уровня удалена, то нужно удалить и текущую вершину, а также больше не ссылаться на нее и не строить из нее ссылки на новые вершины.

Списки внутри списка с пропусками можно также улучшить с помощью второго алгоритма.

3.4 Хэш-таблица

Хеш-таблица - структура данных, основывающаяся на массиве с произвольным доступом. На данный момент не придумано алгоритма, как можно реализовать строгий параллельный доступ к одной ячейке памяти на запись.

Однако блокировать каждый раз весь массив, очевидно, неправильно. В таких случаях используют другой подход.

Весь массив разбивают на кусочки. Чаще всего используют куски одинаковой длины, распределяя их одновременно по массиву. Куски могут пересекаться или не пересекаться. Во время операции модификации высчитывается нужный хэш, находится место в массиве, где этот элемент должен быть изменен, и блокируется только тот кусок, которому принадлежит данный элемент. После модификации блокировка отпускается.

При увеличении количества элементов количество кусков остается неизменным, однако длина каждого из них увеличивается.

4 Глава 3. Тестирование

4.1 Модульное тестирование

Во время модульного тестирования проверяются максимально изолированные от внешнего мира части системы. Чаще всего такие тесты мелкие и целенаправленные. Такие тесты как раз определяют, изменилась ли функциональность программы или работает ли данная функция в целом.

Тестирование многопоточного приложения всегда трудно. Приходится каждый раз задумываться о том, как будут работать те или иные функциональности вместе. В данной работе использовано два основных подхода. Первый заключается в тщательном тестировании каждого метода в одном потоке. Тестирование многопоточного приложения всегда должно начинаться с проверки функциональности в однопоточном искусственно-созданном окружении. Далее проверяются простые сценарии в нескольких потоках, чтобы проверить, что они вообще корректно взаимодействуют между собой, т. е. делают то, что от них «ожидают» в каждом конкретном сценарии.

Второй подход: «тестирование грубой силой». В этом случае запускается большое количество потоков или выполняется большое количество операций одновременно. При увеличении числа операций вероятность ошибки увеличивается, в этом и заключается данный метод. Однако даже это не гарантирует, что программа работает правильно. В некоторых случаях сценарии неправильной работы кода настолько редки, что можно вообще никогда их не получить ошибку в тестировании, но получить в работе с пользователем.

Отсюда плавно вытекает третий подход: тестирование аналитически. Он заключается в тщательном продумывании всех возможных сценариев, про-

верки каждой строчки кода, попытки смоделировать выполнение программы и найти потенциальные ошибки. Однако большая проблема данного подхода: «человеческий фактор». Иногда такую проверку все же можно сделать формально и наглядно. В данной в ссылках на алгоритмы приведен подробный анализ корректности работы алгоритма, заключающийся просто в разборе всех возможных сценариев.

4.2 Тестирование производительности

Если тестирование работоспособности программы нужна для проверки, что программа работает корректно, то оценка производительности нужна для представления, реализуема ли вообще эта программа в жизни. Проверяется, соответствует ли время работы или количество используемой памяти теоретическим оценкам. Чаще всего такое тестирование проходит в сравнении с эталоном. В данной работе эталоном представлялся дополнительный алгоритм, основанный на блокировании структуры. Ожидается, что на большом количестве потоков, этот алгоритм будет работать в среднем хуже, чем неблокирующий алгоритм.

Для оценки производительности можно использовать различные метрики. В данной работе рассматривалось несколько различных тестовых окружений, в каждом из которых несколько различных структур данных и варианты работы в 1, 2 и 8 потоках.

Были произведены тесты на 1, 2 и 8 потоках всех вышеописанных структур данных а также их блокирующих аналогов. В качестве тестового окружения был выбран компьютер Intel Core i7-4790 CPU 3.60GHz (Haswell), ProcessorCount=8 Frequency=3507505 Hz, Resolution=285.1029 ns, Timer=TSC. C# Clr 4.0.30319.42000,

64bit LegacyJIT/clrjit-v4.6.127.1;compatjit-v4.6.1055.0. В качестве тестовых сценариев были выбраны такие как:

- вставка, удаление, поиск по-отдельности
- только вставка и поиск в соотношении 9:1
- вставка, поиск и удаление в соотношении 2:7:1

Результаты представлены в приложении 1.

4.3 Результаты

Из результатов

5 Заключение

Список литературы

[1] Herlihy M. Wait-free synchronization // ACM Transactions on Programming Languages and Systems. 1991. p. 124–149.

[2] Harris T. L. A Pragmatic Implementation of Non-Blocking Linked-Lists // Proceedings of the 15th International Symposium on Distributed Computing. 2001. P. 300–314.

[3] Mikhail Fomitchev E. R. Lock-Free Linked Lists and Skip Lists. 2003.

[4] Fomitchev M. Lock-free linked lists ans skip lists. 2003. p. 124–149.

[5] Michael M. M. High Performance Dynamic Lock-Free Hash Tables and List-Based Sets // IBM Thomas J. Watson Research Center.

[6] Maurice Herlihy N. S. The Art of Multiprocessor Programming. Morgan Kaufmann, 2012. 552 p.

[7] Alfred V. Aho Jeffrey D. Ullman J. E. H. Data Structures and Algorithms. Addison-Wesley publishing company, 2003. 382 p.

6 Приложения