ТИТУЛЬНЫЙ ЛИСТ

СОДЕРЖАНИЕ

[ВВЕДЕНИЕ 5](#_Toc104017647)

[ОБЗОР АНАЛОГОВ 8](#_Toc104017648)

[1 ОПИСАНИЕ HASH ARRAY MAPPED TRIE 11](#_Toc104017649)

[1.1 Описание Trie 11](#_Toc104017650)

[1.2 Описание однопоточной Hash Array Mapped Trie 15](#_Toc104017651)

[1.3 Описание lock-free Hash Array Mapped Trie 23](#_Toc104017652)

[2 РАЗРАБОТКА 29](#_Toc104017653)

[2.1 Настройка среды разработки 30](#_Toc104017654)

[2.2 Реализация однопоточной версии HAMT 33](#_Toc104017655)

[2.3 Реализация блокирующей многопоточной версии HAMT 38](#_Toc104017656)

[2.4 Реализация lock-free HAMT 40](#_Toc104017657)

[2.5 Решение проблемы ABA 42](#_Toc104017658)

[2.6 Внедрение в библиотеку libcds 42](#_Toc104017659)

[3 ТЕСТИРОВАНИЕ 44](#_Toc104017660)

[4 ЗАКЛЮЧЕНИЕ 45](#_Toc104017661)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 46](#_Toc104017662)

# 

# ВВЕДЕНИЕ

Идея создания многопоточных подходов появилась и использовалась ещё в середине прошлого века [4]. Но её бурное развитие началось с повсеместным распространением многоядерных процессов [5]. Это раскрыло их потенциал через использование многопоточного кода, где задействуется вся мощность компьютера.

В многопоточных приложениях потокам (отдельные единицы исполнения кода) зачастую необходима синхронизация через передачу друг другу данных, обмен сигналами, сообщений. Потоки могут делать это посредством как *блокирующей синхронизации*, так и *неблокирующей*.

Блокирующая синхронизация использует специальные примитивы синхронизации (мьютекс, семафор, условные переменные), чтобы в любой момент только один поток мог исполняться в *критической секции*. Для обеспечения эксклюзивности выполнения другие потоки переводятся операционной системой в режим ядра, где они ожидают исполнения. Такой подход может привести к серьёзным проблемам: дедлок, ухудшение быстродействия из-за затрат на перевод режим ядра.

Альтернативой служит неблокирующий подход, который не переводит потоки в режим ядра, а заставляет каждый беспрерывно исполнять код. В этом случае понятие критической секции, в которой в любой момент времени может находится лишь один поток, уходит. Точнее, теперь блокирующая секция сужается лишь до одной ассемблерной инструкции *compare-and-swap*. И только этот код может исполняться потоком эксклюзивно. Данный подход во многих случаях эффективнее по времени, так как не требует временных затрат на блокировку.

В свою очередь, неблокирующие алгоритмы могут быть поделены на три типа в зависимости от гарантий, которые они дают: [6]

|  |  |
| --- | --- |
| Тип | Гарантии |
| Без препятствий | Любой поток при любой начальной конфигурации завершает свою работу через ограниченное количество шагов. |
| Без блокировок | Хотя бы один из потоков совершает прогресс в любой момент времени. |
| Без ожиданий | Любой поток завершается через конечное количество шагов. И это число не зависит от других потоков |

Таблица 1 – Типы неблокирующих алгоритмов и их гарантии

Из трёх вышеприведённых типов наиболее используемый тип «без блокировок (то есть lock-free). Он является компромиссом между не всегда возможным типом «без ожиданий» и типом «без препятствий», который даёт слишком слабые гарантии. Именно на основе типа «без блокировок» создано большое число *lock-free структур данных*, которые позволяют производить с ними операции многопоточно.

Одной из таких lock-free структур данных является Hash Array Mapped Trie. Впервые Hash Trie была описана Брандисом и Фредкином. [2] Позже, она была модифицирована Бегвилом [2] в Hash Trie Mapped Trie – подтип префиксных деревьев, поддерживающие операции поиска значения по ключу, удаление ключа, добавление значения по ключу. Также, операции производятся по хешированному ключу. Каждый узел префиксного дерева хранит ссылки на поддеревья внутри массива, которые индексируется с помощью *bitmap*. Это делает структуру эффективной по памяти и использованию кэша.

В 2017 году трое учёных из Швейцарии [1] предложили многопоточную lock-free реализацию *Hash Array Mapped Trie*. В данном алгоритме, они избавились от проблемы stop-the-word, которая свойственна хеш-таблицам; а также сделали структуру данных эффективной по памяти – с динамическим c глубины.

В данный момент реализация lock-free Hash Array Mapped Trie есть на языке Java, но отсутствует для С++. Так, в библиотеке lock-free структур данных libcds [3] необходима реализация этой структуры данных на языке С++. Именно это и является главной целью данной работы.

В качестве задач, шагов к выполнению были предложены следующие:

1. Анализ Hash Array Mapped Trie структуры данных
2. Реализация однопоточной версии
3. Реализация блокирующей многопоточной версии
4. Переход к неблокирующей многопоточной версии HAMT
5. Решение проблемы ABA
6. Внедрение в библиотеку libcds
7. Добавление в libcds

# ОБЗОР АНАЛОГОВ

Напомним, какая проблема решает при внедрении lock-free Hash Array Mapped Trie. Существует базовая задача поддержания *словаря* [5], где каждый элемент (значение) идентифицируется уникальным ключом. Под поддержанием имеется ввиду возможность производить некоторые операции. Обычно, в этот список входят:

* Вставка значения по ключу
* Удаление по ключу
* Поиск значения по ключу

Особо важно, чтобы каждая операции работала за наименьшее время, а также затраты на хранения словаря должны быть минимальны.

Данная прикладная задача решается повсеместно. Начиная от самой популярной [3] базы данных Redis и заканчивая таблицей символов в большинстве компиляторов [1]. Причём, в зависимости от специфики области применения некоторыми требованиями можно пренебречь в пользу других. Например, пожертвовать временем исполнения для сокращения затрачиваемой памяти или памятью для того, чтобы предотвратить снижения производительности.

Задача поддержания такого словаря в свою очередь посредством одной из следующих структур данных.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Добавление | Удаление | Поиск |
| Хеш-Таблица | O(1)\* | O(1) | O(1) |
| Красно-Чёрное Дерево | O(log n) | O(1) | O(1) |
| Список | O(n) | O(n) | O(n) |

Таблица 2 – Структуры данных и амортизационное время операций

Из таблицы 2 можно сделать вывод, что хеш-таблица является наиболее предпочтительной структурой данных. Она выигрывает по всем параметрам. Но стоит отметить:

* во-первых, линейное время исполнения достигается не всегда, это лишь амортизационное время;
* во-вторых, хеш-таблица время от времени необходим рефазинг, то есть время, когда её необходимо расширить.
* в-третьих, нельзя удалить элемент в хеш-таблице
* в-четвёртых, хеширование является затратным по времени. Более того, оно необходимо при каждой операции с деревом.

Второе замечание ведёт к понижению производительности в случайный момент времени. Это может быть недопустимы в некоторых системах реального времени. Третье замечание показывает неэффективность хеш-таблиц по памяти (хранятся лишние элементы).

В тоже время, красно чёрное дерево проигрывает хеш-таблице по операции добавление. Зато не имеет проблем с рефазингом и неэффективности по памяти.

Из всех структур данных из Таблицы 2 список является самым эффективным по памяти, но скорость работы неприемлемо низкая.

Хотело иметь структуру данных, которая является компромиссом между быстрой хеш-таблицей и эффективным по памяти красно-чёрным деревом. И если определить к структуре данных следующие требования:

* Отсутствие время на рефазинг
* Быстродействие по сравнению с аналогами
* Эффективность по памяти по сравнению с аналогами

То подходящей является Hash Array Mapped Trie. Она сочетает в себе свойства хеш-таблиц и префиксных деревьев (trie). Все операции у это структуры данных являются выполняются за O(1). Также необходимости в рефазинге – структура динамически при каждой новой операции меняет свой размер (сокращается, увеличивается), а не откладывает это как в хеш-таблице. То есть, структура ещё эффективна по памяти. И главное, Hash Array Mapped Trie может быть реализована как lock-free структура. Это даёт мощный прирост к производительности её базовых операций.

# 1 ОПИСАНИЕ HASH ARRAY MAPPED TRIE

1.1 Описание Trie

Идея внедряемого алгоритма была описана в статье Cache Aware Lock-Free Concurrent Hash Tries [1]. Перед тем как изложить соответствующие устройства lock-free Hash Array Mapped Trie, необходимо понять, как работает однопоточная версия. Это, в свою очередь, требует знаний о структуре данных Trie.

*Trie* – дерево поиска (оно же *префиксное дерево* или бор) позволяет хранить, удалять, искать ключи, а также привязанные к ним значения. Обычно, ключами выступают строки, но могут быть и числа. Чтобы добавить элемент в дерево, ключ разбивается на части. Если это строка, то разбиение зачастую идёт по символам. Если число, то на биты. Далее для каждой части создаётся репрезентирующий её узел. Например, на рисунке 1.1.1 показано разбиение строки «hton»

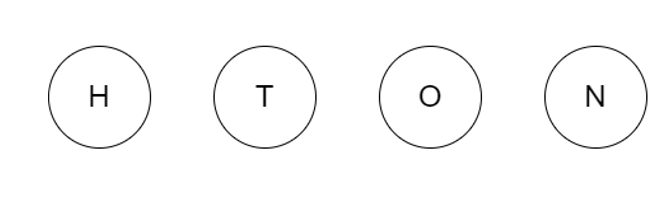


Рисунок 1.1.1 – Узлы из «hton»

А число 9 будет предварительно переведено в двоичную систему разбито на биты 1001 как показано на рисунке 1.X

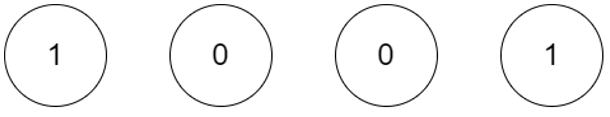


Рисунок 1.1.2 – Узлы из ключа 9

После разбития ключа, необходимо поочерёдно вставить узлы в дерево. Вставка всегда происходит из корня. Её алгоритм может быть описан так: пусть нам необходимо вставить ключ k, который разбивается на , … ... узлы, мы находимся в узле дерева и уже вставили все узлы до .

Тогда,

* если уже имеет ребёнка (назовём его ), то мы пропускаем вставку (так, как он уже есть = ) и вставляем начиная c .
* если не имеет ребёнка , то мы добавляем в ребёнка (назовём его ) и вставляет начиная с .

В обоих случаях, если был обработан последний узел (), то помечается как терминальный. Это указывает, что данный узел является в дереве концом какого-то ключа.

Что касается значений, ассоциированных с ключом, то они могут быть привязаны к терминальным узлам. Таким образом, не каждый узел в префиксном дереве связан с определённым значением.

Как пример, префиксное дерево, включающее в себя строки «kafka», «karma», «kafkaesque», «jail» показано на рисунке 1.1.3. Узлы красного цвета являются терминальными, белого – промежуточными, зелёный узел соответствует корню.

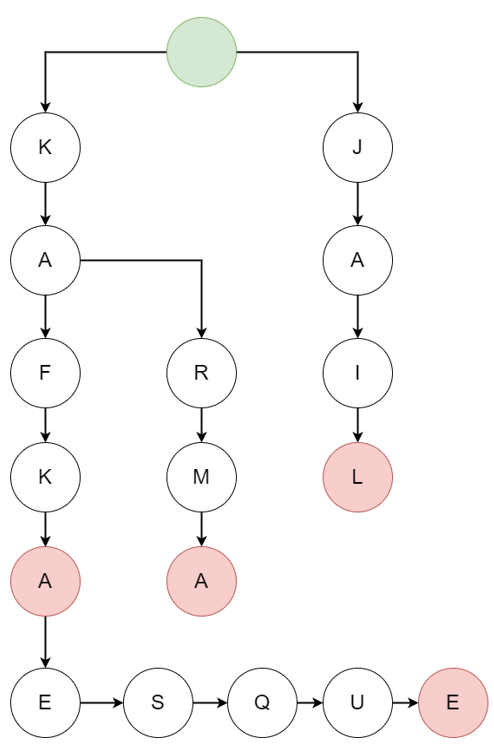


Рисунок 1.1.3 – Префиксное дерево, хранящее строки

Стоит заметить, trie компактно хранит ключи: узлы строк с одинаковыми префиксами не дублируются, а хранятся в единственном экземпляре («kafka» и «karma» имеют общий префикс «ka»). Это и делает эту структуру данных эффективной по памяти. То же можно сказать и про Hash Array Mapped Trie, которая использует свойства префиксных деревьев.

Trie может быть реализовано несколькими способами. Первый способ – наивный. Пусть у нас есть определённый алфавит, состоящий из символов (обычно = 256, что соответствует размерности таблицы ASCII). То есть, всего может быть различных видов узлов в дереве и у каждого узла может быть не больше, чем различных детей. Тогда пусть каждый узел хранит ссылки на своих детей в собственном фиксированном массиве на элементов (). Если , значит данный узел не имеет ребёнка, узла с типом . На рисунке 1.1.4 изображён узел A, который имеет только трёх детей B, C, D, но хранит информацию об отсутствии 253 других возможных детей (например, , то есть отсутствует узел C).

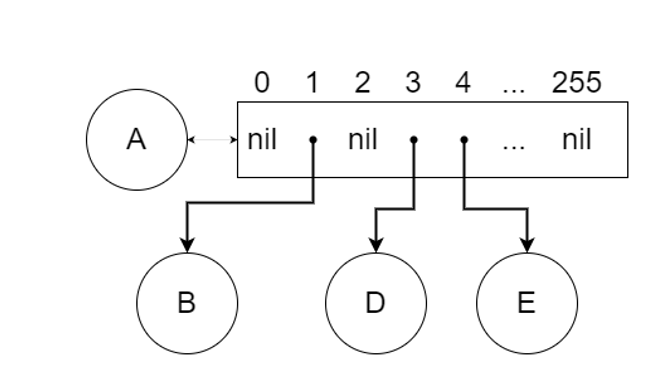


Рисунок 1.1.4 – Узел в наивной реализации префиксного дерева

Данный подход имеет плюс в скорости поиска следующего узла (асимптотика – Вместе с тем, огромное количество памяти расходуется впустую на хранение .

Закономерным решением проблемы с памятью было бы использование списков вместо массивов. Хотя это и увеличит время операции до , где m – количество детей у узла, но зато позволит избавиться от лишнего расхода памяти. На рисунке 1.1.5 показано префиксное дерево, реализующее данных подход: узел A хранит связный список (linked list) из трёх элементов B, D, E, что существенно меньше, чем в наивной реализации.

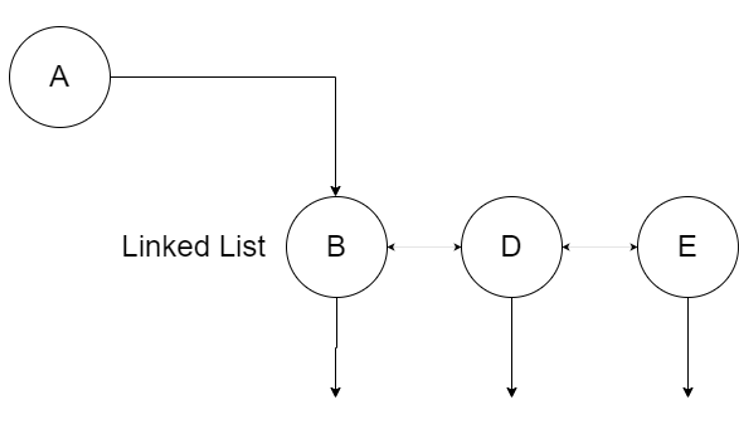


Рисунок 1.1.5 – Узел в реализации префиксного дерева со связным списком

Hash Array Mapped Trie снова находится на середине между двумя подходами. Для хранения детей использует массив как в наивном методе (но не статический, а динамический. И так же, как и реализации со связным списком, не хранит лишние элементы.

Основные положения Trie:

* Позволяет поддерживать словарь (операции: вставка, удаление, поиск)
* Ключ делится на части, которые репрезентируются в узлы.
* Узлы бывают трёх видов: терминальные (хранят в себе значение), корень (представляет особый узел, у которого нет родителей), промежуточные
* При вставке узлы ключа последовательно добавляются в дерево. Если два ключа имеют общий, префикс, их узлы, соответствующие этому префиксу, не дублируются.
* Каждый узел хранит детей в связном списке, что позволяет экономить память (в отличии от статического массива)

1.2 **Описание однопоточной Hash Array Mapped Trie**

После детального рассмотрения префиксного дерева, можно приступить к устройству Hash Array Mapped Trie. Hash Array Mapped Trie (HAMT) представляет собой такое же дерево поиска для хранения, удаления, поиска ключей/значений. Как было сказано ранее, она сочетает в себе свойства хеш-таблиц и префиксных деревьев. Но также включает отличные идеи для повышения быстродействия.

Как и для префиксных деревьев, HAMT разбивает свой ключ на части. Эти части определяют путь до ключа/значения. Ключи могу иметь общие префиксы, что позволяет экономить память. Но, если в префиксных деревьях разбиение определялось для каждого типа ключа по-разному (для строки – на символы, для числа – на биты), то Hash Array Mapped Trie разбивает единообразно: получает 64-битный хеш ключа, а затем делит его части по пять битов. Итого, частей . Каждые пять бит – это отдельное число. Например, если в дерево добавляется ключ, который хешируется в число 2172 (двоичную форму будем обозначать так ), то он будет разбит на (рисунок 1.2.1)

Части репрезентируются в узлах 1, 31, 0, 0 и так далее.

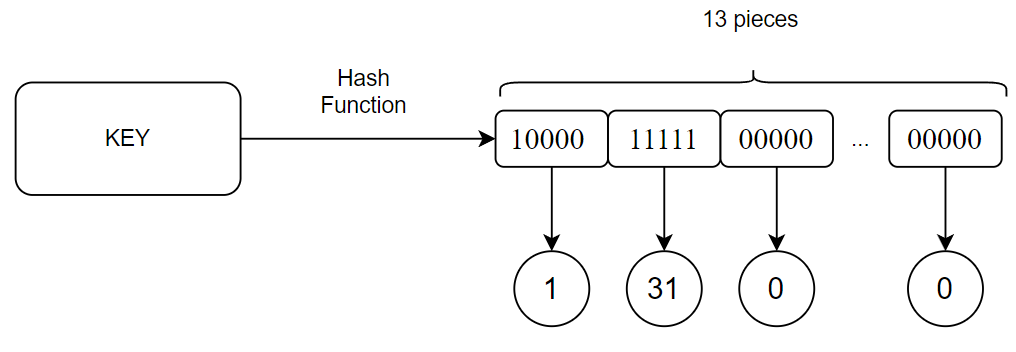
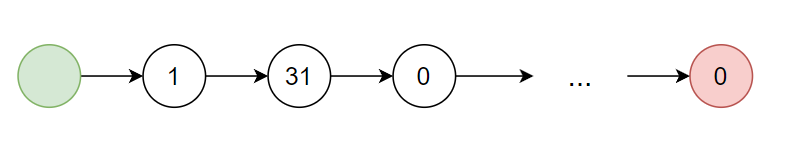


Рисунок 1.2.1 – Разбиение хеша ключа на части

Таким образом, данный ключ в дереве будет располагаться по пути, который показан на рисунке 1.2.2. Зелёным цветом обозначен корень дерева, белым – промежуточные вершины, красным – терминальная вершина, где хранится значение для ключа.

Рисунок 1.2.2 – Путь в дерева для ключа с хешом 2172

На самом деле, для хеша 2172 не обязательно хранить все узлы в контексте дерева. Нужно хранить ровно такое количество, чтобы можно было уникальное идентифицировать путь до ключа в дереве. Приведём несколько примеров.

Пусть у нас есть пустое дерево (рисунок 1.2.3)



Рисунок 1.2.3 – Hash Array Mapped Trie не содержащий ни одного ключа

Добавим в дерево ключ с хешом 2172 (рисунок 1.2.4). Как говорилось ранее, нужно добавлять минимальное количество узлов, чтобы можно было уникально идентифицировать путь до ключа. В данном случае в пустом дереве хватит лишь одного узла. Этот узел нужно сделать терминальным и положить в него значение и полный ключ (это понадобиться для проверки при поиске).

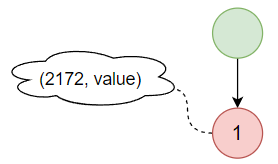


Рисунок 1.2.4 – Hash Array Mapped Trie c единственным ключом.

Теперь, если нам необходимо будет найти ключ 2172, мы проходить не по всем узлам (ведь дерево их не содержит), а лишь по части и них. И когда поиск приведёт к терминальной вершине, необходимо провести проверку на полное совпадение ключей. Ниже будет описан более строгий алгоритм поиска. При добавлении ключа с хешом 65 появляется неоднозначность. Действительно, у обоих хешей первая часть совпадает. Значит, узел 1 () уже нельзя сделать терминальным (неоднозначно какому ключу он будет принадлежать: 65 или 2172). Поэтому необходимо сделать узел 1 промежуточным, и отвести от него два узла для 65 и 2172 (рисунок 1.2.5). Эти узлы вычисляются уже по второй части хеша.

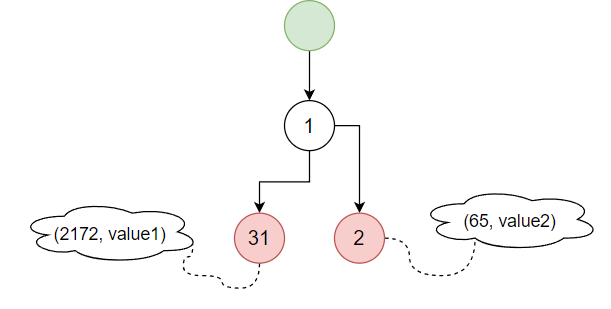


Рисунок 1.2.5 – Hash Array Mapped Trie c двум ключами с общим префиксом у хешей.

Может случиться так, что и вторые части хеша будут совпадать. Тогда производим те же самые операции (создаём общий узел и отводим узлы для детей). В случае, если все части хешей совпадают, то на последнем уровне вместо терминальной вершины создаём *терминальный узел-список* из двух ключей. Терминальный узел-список, это лист, который хранит себе множество пар ключ/значение. На рисунке 1.2.6 показан процесс разрешения коллизии, при добавлении двух ключей с идентичными первыми 12 частями хеша.

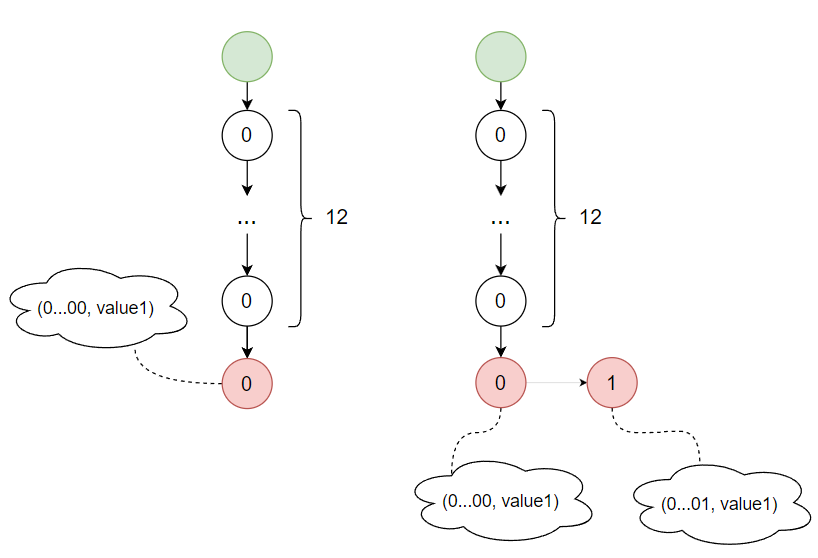


Рисунок 1.2.6 – Разрешение коллизии

Теперь осуществим удаление. Оно будет производится рекурсивно, начиная с терминальной вершины, соответствующей ключу: если текущая вершина не имеет детей, то она подлежит удалению. То есть, HAMT сжимает дерево сразу же после удаления, поэтому не хранит лишних узлов.

Более строго операции вставки, поиска, удаления можно описать следующим образом.

*Вставка*: пусть нам необходимо вставить пару , где , … ... , мы находимся в узле дерева и уже вставили все узлы до . Тогда:

* если у текущего узла дерева нет ребёнка , мы добавляем его, помечаем терминальным и вкладываем
* если у текущего узла есть нетерминальный ребёнок , то пропускаем вставку и производим добавление следующей части хеша , но уже c узла
* если (то есть, мы находимся ниже предпоследнего уровня глубины дерева) и у текущего узла есть терминальный ребёнок с ключом и значением , то сначала проверяем не совпадает ли он с . Если да, то заменяем внутри этого ребёнка на. Иначе, делаем из него промежуточный узел и добавляем , и
* если (то есть, мы находимся на предпоследнем уровне глубины дерева) и у текущего узла есть терминальный ребёнок с ключом и значением , то мы заменяем его на терминальный узел-список, состоящий из
* если (то есть, мы находимся на предпоследнем уровне глубины дерева) и у текущего узла есть терминальный узел-список ребёнок, то проверяем есть ли в этом списке пара , где . Если имеется, то заменяем на . Иначе, добавляем в список новую

*Поиск:* пусть нам необходимо найти ключ , где , … ... , мы находимся в узле дерева и уже спустились на уровень . Тогда:

* если у текущего узла дерева нет ребёнка , то ключ/значение не найдены.
* если у текущего узла есть нетерминальный ребёнок , то производим поиск следующей части хеша , но уже с узла .
* если у текущего узла есть терминальный ребёнок с ключом и значением, то проверяем есть ли в этом списке пара , где . Если да, то и есть искомая пара. Иначе, ключ/значения не найдены.

*Удаление*: пусть нам необходимо удалить ключ . Произведём поиск этого ключа. Если ключ отсутствует, то завершаем удаление. Иначе, существует терминальная вершина соответствующая ему. Начнём удаление с вершины . Тогда:

* если текущая промежуточная вершина не имеет детей, то удаляем её (удаляем ссылку на неё у родителя)
* если текущая терминальная вершина не имеет детей, то удаляем её, если она хранит единственный ключ. Если это узел-список, то удаляем из неё такую пару , что .
* иначе, прекращаем процесс.

Следующий важный момент в Hash Array Mapped Trie – это использование связки битмапа и динамический массив. Идея заключается в том, что узел хранит ссылки только на существующих детей в динамическом массиве. Поиск по такому массиву занимает *O(n),* как это было в одной из реализаций trie ранее. Но HAMT также хранит битмапу, по которой можно равен определить индекс необходимого элемента в массиве. Индекс вычисляется с использованием побитовых операций, что значительно быстрее предыдущей версии.

Структура данных битмапа (bitmap) – это число с определённой разрядностью , где каждый бит говорит об присутствии (если равен единице) или отсутствии (если равен нулю) чего-либо. Битмапа идёт всегда в связке с другим объектом и описывает его состояние. В случае Hash Array Mapped Trie битмапа есть у каждого узла, её разрядность 32, а характеризует она наличие или отсутствие детей определённого. Например, битмапа 104 (или в двоичном виде ) у узла P (рисунок 1.2.7) говорит о наличии трёх детей: 0, 1, 3.

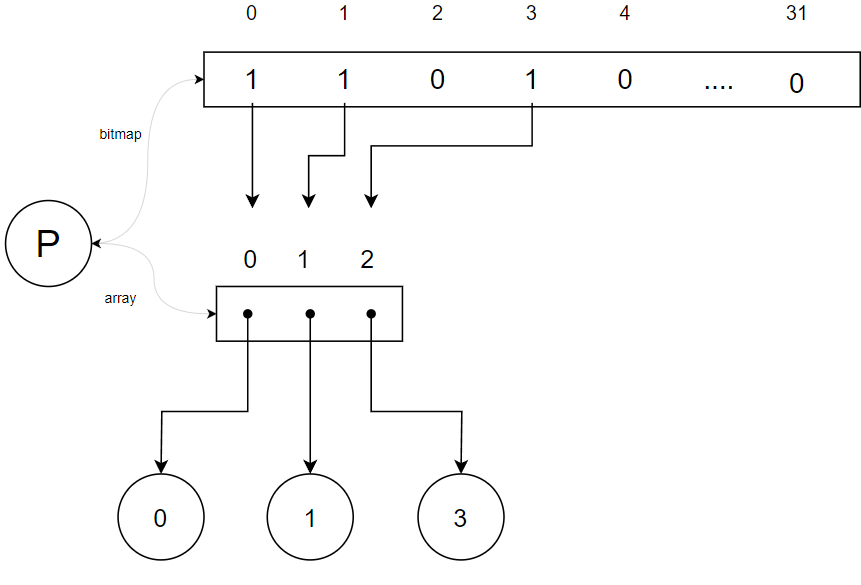


Рисунок 1.2.7 – Узел с битмапой и массивом(array), содержащий трёх детей.

Зная индекс битмапы, где поднят флаг единицы, можно вычислить индекс в массиве.

Формула искомого индекса вычисляется по формуле (1):

, (1)

где popcount – функция подсчёта поднятых битов в числе, – позиция в битмапе с установленным битом, & - операция побитового «И», – побитовый сдвиг числа 1 раз влево. Получения бита в битмапе производится тоже через побитовые операции: . Итого, чтобы получить элемент массива по битмапе, необходимо сделать два шага:

1. Проверить, установлен ли бит на определённой позиции (используются только побитовые операции)
2. Если в предыдущем шаге бит не установлен, возратить ничего. Иначе, вычисляем индекс массива (используются только побитовые операции)

Благодаря битмапе поиск ребёнка в узле происходит не за время O(n), а за O(1).

Основные положения о Hash Array Mapped Trie:

* Ключ хешируется и разбивается на части.
* У каждого узла может быть не больше, чем 32 ребёнка
* Максимальная высота дерева равна 13
* При добавлении ключей с одинаковыми префиксами хешей, общая часть не дублируется.
* Добавляется минимальное количество узлов, при котором можно уникально идентифицировать ключ
* При удалении, дерево сжимается и не хранит лишних узлов.
* Каждый узел хранит детей в динамическом массиве и битмапе. Через побитовые операции можно за время O(1) вычисляет ребёнок узла или его отсутствие.

1.3 **Описание lock-free Hash Array Mapped Trie**

Идея алгоритма [1] построена на введение 3 типов узлов (рисунок 1.3.1):

1. SNode – узел, который содержит список пар ключ/значение
2. CNode – узел с битмапой и массивом детей. Дети могу двух типов: SNode и INode
3. INode – узел, который хранит в себе ссылку на CNode. Это искусственная обёртка над CNode для поддержания lock-free.

Все три типа унаследованы от абстрактного узла Node.

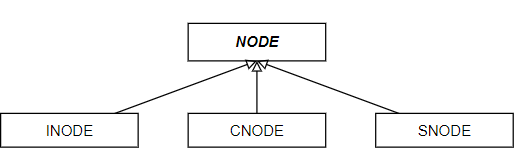


Рисунок 1.3.1 – UML-диаграмма для узлов

Начнём объяснения используя SNode и CNode. В дальнейших рассуждениях будем усложнять структуру данных. Пусть у нас есть дерево с узлом SNode(рисунок 1.3.2). Надпись на направленном ребре обозначает тип ребёнка.

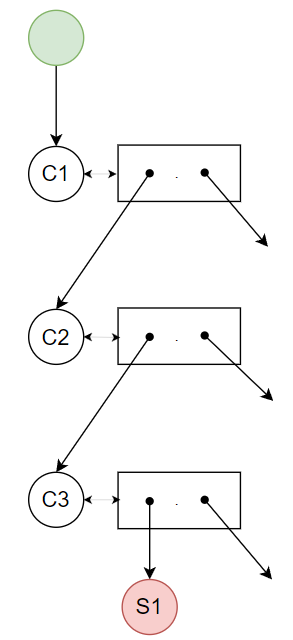


Рисунок 1.3.2 – Hash Array Mapped Trie с узлом S1

Тогда, если мы хотим втавить в С3 ещё одного ребёнка, нам необходимо будет совершить два действия:

1. Обновить битмапу, чтобы она указывала и на новый элемент
2. Вставить новый узел в массив C3

Операции 1 и 2 происходят неатомарно и это приводит к логической гонке данных. Например, если два потока T1 и T2 одновременно пытаются добавить соответсвенно новые узлы S2 и S3, то может произойти следующее:

1. T1 копирует битмапу и меняет её локально.
2. T2 копирует битмапу (которая указывает только на S1) меняет её локально.
3. T1 меняет битмапу в C3 на изменённую локальную версию
4. T2 перезаписывает битмапу C3 на свою изменённую локальную версию.

В результате, битмапа будет создержать версию потока T2, который не учивает изменения T1. Решение этой проблемы заключается в том, что обновляют не битмапу и массив, а целые узлы. Так, в данном случае T1 не меняет содержимое C3, а копирует этот узел , меняет его локально, и уже потом обновляет в С2 указатель на С3 на

Для того, чтобы избежать техническую гонку данных при изменении указателя, необходимо воспользоваться операцией *compare-and-swap* (cas).

Это функция принимает три аргумента:

1. указатель на старый узел (назовём это )
2. значение, которое лежало по этому указателю (назовём это )
3. обновлённый узел (назовём это )

Если , то мы меняем содержимое указателя на и возвращаем , как показатель того, что значение было обновлено. Иначе возвращаем .

В нашем случае, поток T1 сначала зафиксирует локально чему равно С3. Потом сделает его копию , начнёт менять (добавит новый узел). И воспользуется сas, куда передаст указатель на C3 в качесте p, зафиксироованный ранее в качестве и в качестве . Если cas прошёл успешно, то есть никакой другой поток не поменял C3, то наш поток обновил его. Иначе, производим вставку заново с самого начала. Операция cas происходит за одну ассемблерную иструкцию, что позволяет считать её lock-free, то есть неблокирующей.

Даже при таком решении с атомарны compare-and-swap на целым узлом, отстаются проблемы с гонками данных. Пусть есть некоторое дерево в следующей конфигурации ( рисунок 1.3.3) и пусть есть поток T1, который хочет вставить в C3 новый узел S2. Для этого он копирует ноду С3 (жёлтым цветом обозначена копия).

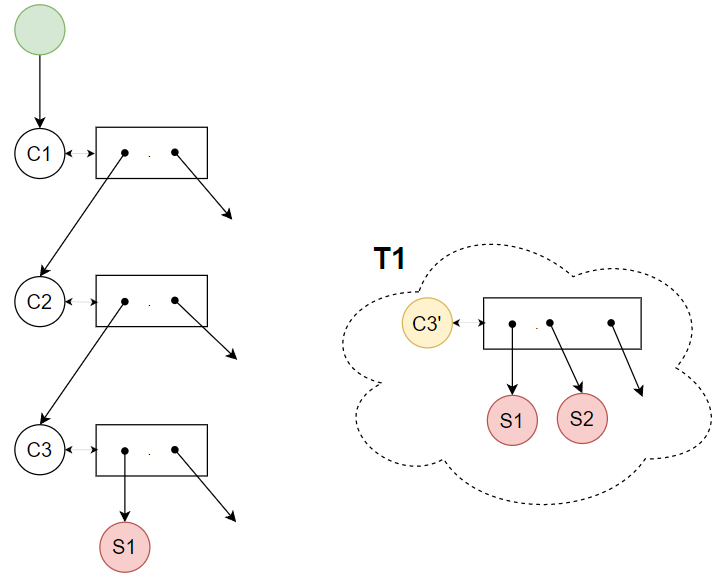


Рисунок 1.3.3 – поток T1 копирует C3 и добавляет в копию новый узел S2

Пока T1 занимается копирование и локальным добавлением нового узла, другой поток T2 решает добавить новый узел S3 в C2. Для этого он также создаёт копию , добавляет узел S2 (рисунок 1.3.4)

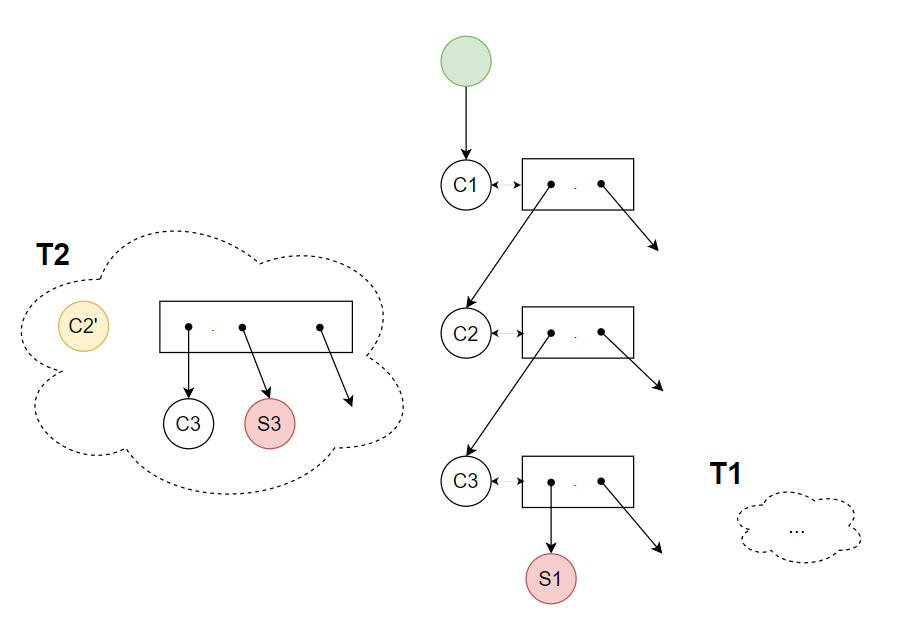


Рисунок 1.3.4 – поток T1 копирует C3 и добавляет в копию новый узел S3

И если сначала T1 сделает cas-операцию, а потом T2, то узел S2 потеряется (рисунок 1.3.5).

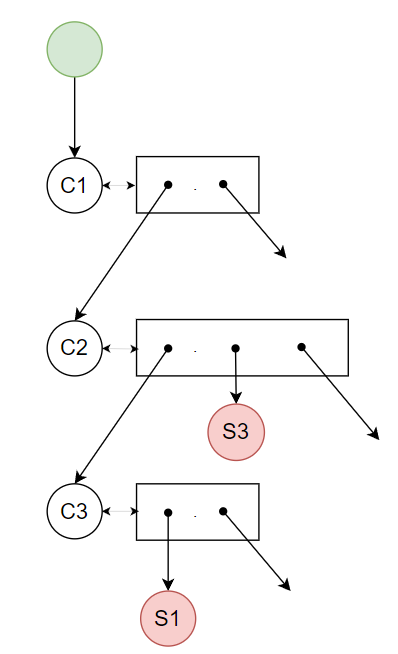


Рисунок 1.3.5 – дерево после обновление C3 потоком T1 и С2 потоком T1

Для решения данной проблемы, вводится новый тип узла INode. Он является обёрткой над каждым CNode. То есть, для каждого CNode узла существует родитель INode, который ссылается только на него. Тогда CNode в качестве ребёнка может иметь только INode или SNode. Теперь, любая cas-операция производится над содержимым промежуточного узла INode, что позволяет избавить от проблемы, описанной ранее.

На данном этапе, lock-free Hash Array Mapped Trie работает корректно для операций вставка и поиск. Но есть ещё операция удаления. Она происходит следующим образом: удаляется пара с искомым ключом из соответствующего узла. Если в этом узле SNode не осталось ключей, то необходимо удалить его и произвести сокращение дерева. Сокращение дерева случается, когда существует INode, ребёнок которого имеет одного ребёнка. Тогда этот ребёнок должен заменить INode. Во время этого процесса может возникнуть проблема. Представим дерево, которое, должно сократиться (рисунок 1.3.6) потоком T1. Пока T1 копирует и изменяет C1(меняет ребёнка I2 на S2), другой поток T2 добавляет в C2 новое значение S3. Для этого также начнёт копировать C2 и обновлять копию. В этот момент, T1 совершает сокращение (I2 заменяется на S2). Затем поток T1 совершит cas и изменит ребёнка I2. Но I2 уже удалён. Поэтому, в дереве не будет вершины S3.

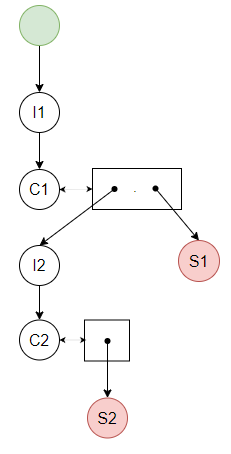


Рисунок 1.3.6 – дерево, перед тем, как сократиться

Решение этой проблемы – это помечать сокращаемые узлы CNode как гробовые (tomb-node). Такие узлы все потоки могут только сокращать. Если CNode уже нельзя сократить, то необходимо его воскресить (из гробового узла сделать обычный).

Основные положения lock-free Hash Array Mapped Trie, описанные в этой главе:

* вводятся три типа узлов INode, CNode, SNode
* при помощи compare-and-swap атомарно происходит изменение целого узла, а не битмапы и массива отдельно
* после удаления узла, дерево начинает сжиматься
* CNode, который может быть сокращён помечаться как гробовой. В него нельзя добавлять или удалять детей.

# 2 РАЗРАБОТКА

В предыдущей главе подробнейшим образом были описаны теоретические аспекты алгоритма lock-free Hash Array Mapped Trie и также заложены предпосылки для внедрения. Внедрение такого алгоритма — процесс трудозатратный. Он включает в себя множество пунктов: постоянная отладка программы, ручное и автоматическое тестирование, анализ кода, проверка гипотез, исправление ошибок и так далее. Чтобы облегчить эти процессы, необходимо иметь чёткий план действий: что реализовывать в первую очередь, а что можно отложить на потом.

План:

* Настройка среды разработки
* Реализация однопоточной версии HAMT
* Реализация блокирующей многопоточной версии
* Переход к неблокирующей многопоточной версии HAMT
* Решение проблемы ABA
* Внедрение в библиотеку libcds

2.1 Настройка среды разработки  
 В качестве контроля версий был выбран Git. Он является свободно распространяемым, наиболее популярным и имеет необходимый функционал:

* создание веток разработки
* слияние одной ветки в другую
* просмотр истории файла
* создание именованных коммитов
* удобное взаимодействие через командную строку
* множество интеграций с другими сервисами

Для того, чтобы код не хранился в одном месте, а также была возможность поделиться с ним, использовался GitHub. Это бесплатный сервис, который работает с системой версией Git и позволяет хранить, анализировать код, производить Git команды, совместно редактировать код и тестировать его благодаря GitHub Actions.

Благодаря GitHub Actions можно настроить систему так, чтобы при каждом коммите тестировался код.

В основной репозитории был положен файл Dockerfile, который запускает докер контейнер при коммите. В него устанавливается операционная система, копируется код из текущего.

Благодаря GitHub Actions можно настроить систему так, чтобы при каждом коммите тестировался код. В основной репозитории был положен файл Dockerfile, который запускает докер контейнер при коммите. В него устанавливается операционная система, копируется код из текущего репозитория GitHub и запускается скрипт entrypoints.sh

# Container for building and testing cmake projects v3.16

FROM ubuntu:18.04

# get the latest cmake release

RUN apt-get -y update

RUN apt-get -y install software-properties-common git wget clang-10

RUN wget -O - https://apt.kitware.com/keys/kitware-archive-latest.asc 2>/dev/null | apt-key add -

RUN apt-add-repository 'deb https://apt.kitware.com/ubuntu/ bionic main'

RUN apt-get -y install build-essential cmake qtbase5-dev

COPY entrypoint.sh /entrypoint.sh

RUN echo "$GITHUB\_WORKSPACE"

COPY ./ /hamt

RUN chmod +x entrypoint.sh

ENTRYPOINT ["/entrypoint.sh"]

В скрипте entrypoints.sh производятся приготовления для сборки и запуска исполняемого файла для тестирования.

#!/usr/bin/env bash  
cd "$GITHUB\_WORKSPACE" || exit 1  
cd test  
cmake -DCMAKE\_C\_COMPILER=clang-10 -DCMAKE\_CXX\_COMPILER=clang++-10 .  
make  
./run\_tests

Для сборки был написан CmakeFileLists.txt. Этот файл используется утилитой CMake, которая генерирует Makefile. А Makefile, в свою очередь, используется утилитой Make, которая создаёт на его основе исполняемый файл для тестирования.

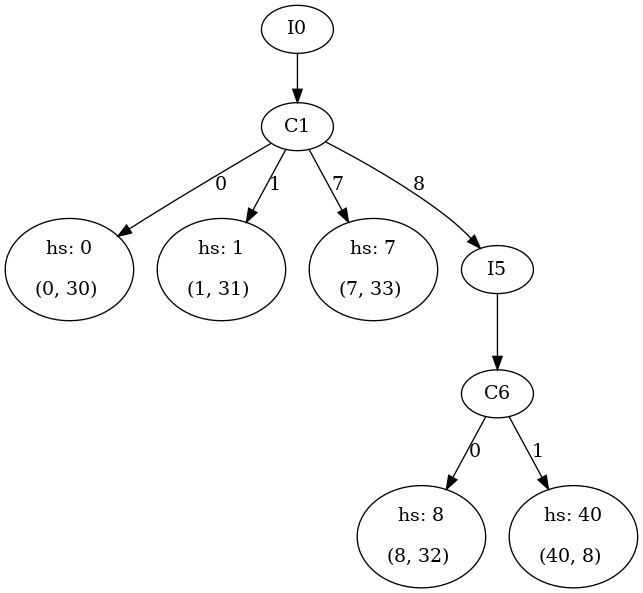
cmake\_minimum\_required(VERSION 3.16)  
project(tests)  
  
set(CMAKE\_CXX\_STANDARD 14)  
  
add\_subdirectory(../lib/googletest googletests)  
  
# TEST TARGET  
add\_executable(  
 run\_tests test.cpp ../src/utils.cpp  
)  
  
target\_link\_libraries(  
 run\_tests PRIVATE  
 gtest gtest\_main  
)

Cmake для сборки был выбран так как прост в использовании и описания действий выражаются проще и короче, чем в утилите Make. В качестве интегрированной среды разработки был выбран Clion от компании JetBrains. Помимо встроенного отладчика, статических анализаторов кода, Clion предоставляет удобный графический интерфейс к Git, а также интегрируется с GitHub. Одна из особенностей лок-фри алгоритмов, что ошибка на конкретных тестах может появляться не постоянно. Поэтому было необходимо тестировать код как можно чаще, чтобы была больше вероятность воспроизвести потенциальные ошибки в коде. GitHub action позволяет запускать тесты, не только при каждом коммите, но и по расписанию. Так, в специальный файл, отвечающий за запуск докер контейнера, было добавлена следующая строка:

schedule:  
 - cron: '\*/5 \* \* \* \*'

В поле cron в синтаксисе подобном crontab находится указание запускать докер контейнер с тестами по расписанию «каждые пять минут». В случае, если тесты проходят неуспешно, на почту отсылается соответствующее уведомление. Автоматическое тестирование при коммитах и по расписанию облегчает внесения изменений в код. Также это ускоряет разработку, так как избавляет от необходимости ручного тестирования и ожидания результатов.

Ещё одна проблема разработке — отладка программы. Clion предоставляет прекрасные возможности для этого, но этого недостаточно. Зачастую необходима визуализация самой структуры данных Hash Array Mapped Trie. Для этого в исходный репозитория был добавлен код для визуализации (visualize.h), в котором была реализована логика сбора информации о переданном HAMT, её обработка, а затем передача библиотеке boost graphviz hpp. Данная библиотека умеет визуализировать графы в формате в png.

  
Рисунок 2.1.1 – Дерево, построенное при помощи visualize.h

2.2 Реализация однопоточной версии HAMT

Приступая к описанию однопоточной версии HAMT на языке C++, необходимо сказать, что с самого код писался с учётом будущей трансформации под lock-free версию HAMT. Это было сделано, чтобы для ускорения разработки и для логического структурирования всей работы.

Начнём с перечисления сущностей (классов, структур), опишем их взаимодействия друг с другом. А затем перейдём к трём основным операциям insert, remove, lookup.

Есть класс Hamt, представляющий контейнер c Hash Array Mapped Trie структурой (рисунок 2.2.1). В нём есть поле root, в котором хранится корневой узле и три публичных метода: вставка, удаление и поиск. K и V означают типы один из типов int или string.

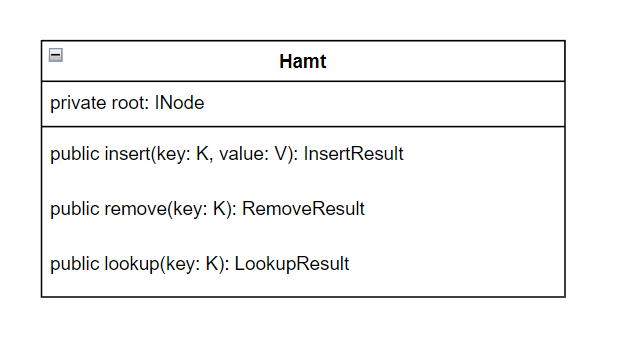


Рисунок 2.2.1 – UML-диаграмма для Hamt класса

Так, при создании контейнера, необходимо указать какого типа буду ключи и значения (рисунок 2.2.2)



Рисунок 2.2.2 – Создание контейнера с ключами и значениями типа int

Все три метода insert, remove, lookup возвращают соотвественно структуры InsertResult, RemoveResult, LookupResult. Это необходимо для того, чтобы хранить информацию о том было ли значение заменено при вставке, удалилось ли ключ. Например, на рисунке 2.2.3 показана UML-диаграмма для структуры RemoveResult. Тип Status – тип перечисления. Он может хранить в себе три значения:

* Removed (если удаление прошло успешно, тогда в поле value будет хранится значение у удаляемой пары)
* Failed (заготовка для lock-free версии, где удаление может не получиться)
* NotFound (в дереве нет такого ключа, которого хотя удалить)

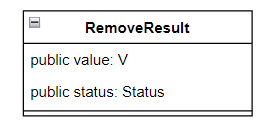


Рисунок 2.2.3 – UML-диаграмма для узлов

Внутри дерева есть вида узлов, наследующиеся от базового абстактного типа Node (рисунок 2.2.4).

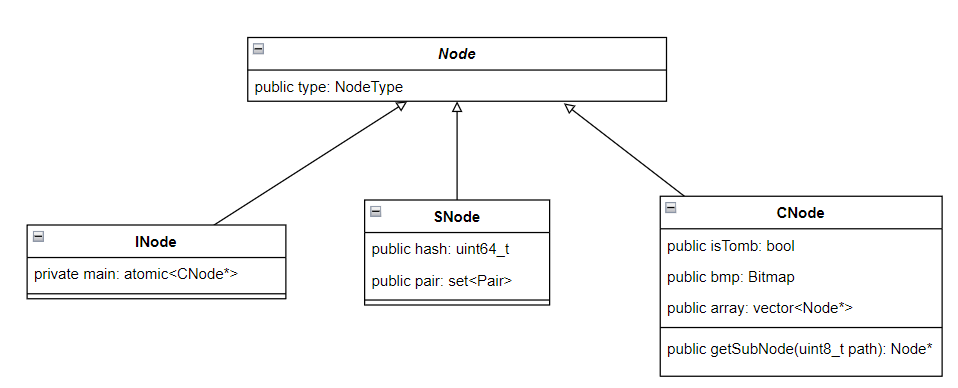


Рисунок 2.2.4 – UML-диаграмма для узлов

INode, как говорилось в прошлой главе, является обёрткой на CNode. Это позволяет обеспечить потокобезопасные обновление узлов. Именно поэтому у него есть одно поле main типа atomic<CNode\*>. С таким типом можно совершать атомарные операции над указателем на CNode (том числе и cas операцию). Сейчас это лишь заготовка для будущей lock-free версии, и она не противоречит однопоточной.

SNode – узел состоящий из двух полей: хеш и множество из пар ключ/значение с таким хешом. Поле hash имеет тип uint64\_t, что занимает 64 бита. Ровно столько, чтобы вместить в себя максимальное значение хеша. Это позволяет не хранить лишние биты и не тратить память. Pair представляет собой класс с двумя публичными полями (один для ключа другой для значения) и перегруженным оператором меньше (это необходимо для создания set на его основе).

CNode – хранит в себе поле isTomb (заготовка для lock-free версии, указывает гробовой это узел или обычный), поле типа Bitmap битмапа у данного узла, и динамический массив указателей на узлы типа Node. Один из важнейших методов – getSubNode, который принимает тип ребёнка (число от 0 до 31) и возвращает указатель на ребёнка или nil, если такого ребёнка нет.

Подробнее стоит остановиться на вспомогательной структуре данных Bitmap и методов, работающих с ней (рисунок 2.2.5). Bitmap характеризует состояние массива array. То есть, указывает на отсутствие или наличие определённых детей. При удалении или добавлении новых узлов в массив, поле типа Bitmap у CNode должно изменяться. Для у это у Bitmap есть публичные методы set, unset и isSet.

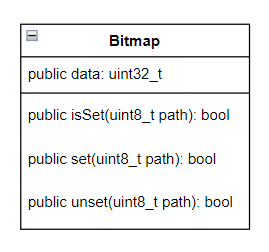


Рисунок 2.2.5 – UML-диаграмма для класса Bitmap

Все три метода (рисунок 2.2.6) используют побитовые операции «И» (&), сдвига вправо (<<), инвертирование (~).

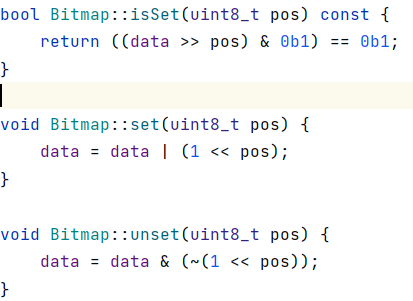


Рисунок 2.2.6 – Реализация методов Bitmap

На основе данных битмапы также вычисляет индекс массива (рисунок 2.2.7). Эта логика реализована в отдельном методе класса CNode.

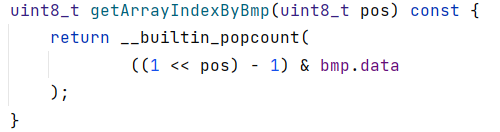


Рисунок 2.2.7 – Вычисление индекса массива для установленного бита

Команда вычисляет количество установленные битов в числе, bmp.data – значение битмапы в соответствующем узле CNode.

Теперь, когда были описаны основные классы и структуры, можно перейти к реализации insert, remove, lookup операций. Они определены в публичных методах класса Hamt и каждый ссылается на соответствующий непубличный метод. Так было сделано по двум причинам:

* Декомпозиция метода (в публичном создаются вспомогательные структуры, которые идут в приватные)
* Необходимость повторного вызова вставки, удаления и поиск (это заготовка для lock-free версии)
* Такая структура легко трансформируется в блокирующую многопоточную версию.

Для примера разберём операцию insert (рисунок 2.2.8). На вход в публичную функцию поступают ключ и значение. В строчка 282-284 идёт проверка особого случая (проверка на пустоту дерева). В строчке 285 создаётся новый узел SNode, который нужно будет вставить. Далее, вызывается приватный метод insert (рисунок 2.2.9).

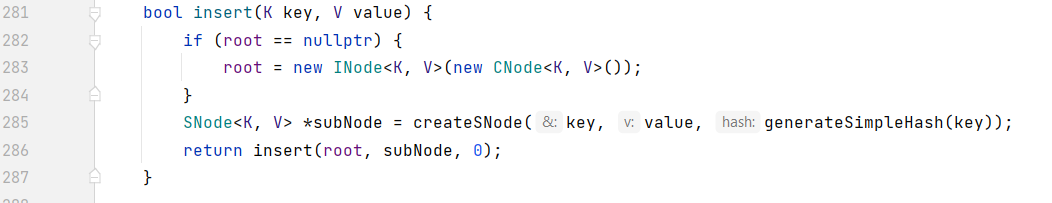


Рисунок 2.2.7 – Вычисление индекса массива для установленного бита

В приватном методе, чтобы определить по какому пути идти, вычисляет часть хеша (строчка 347). Далее вычисляет ребёнок, на которого мы, возможно перейдём на следующем шаге. И затем обрабатывается одна из следующих ситуаций, которая была описана в предыдущей главе.

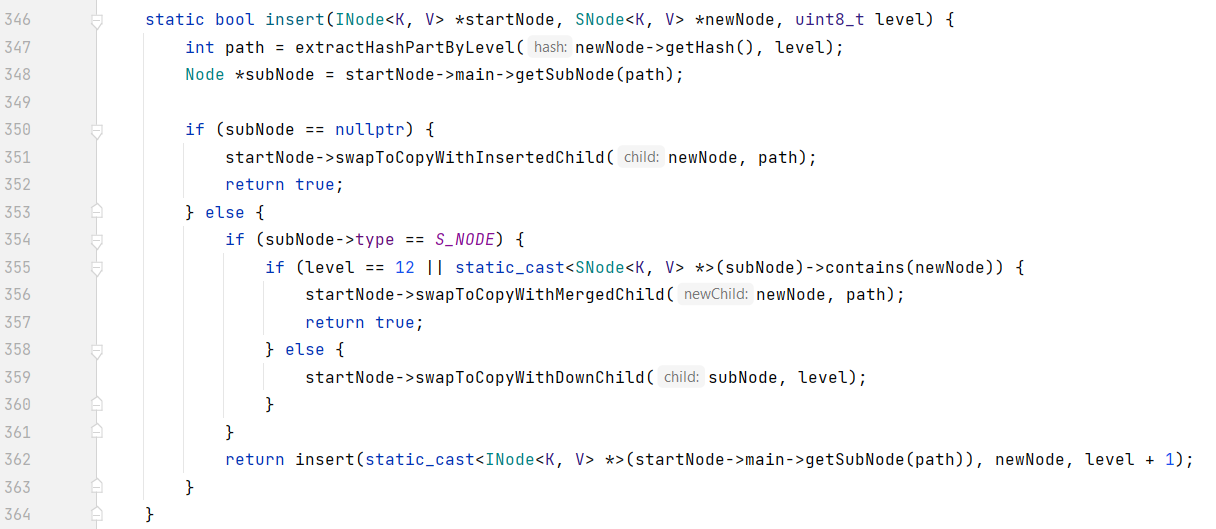


Рисунок 2.2.7 – Вычисление индекса массива для установленного бита

Так, например, если по данному пути нет ребёнка, то есть (строка 350), то для текущего узла вызывается специальный метод swapToCopyWithInsertedChild, который вставляет ребёнка по пути path в текущий узел (строка 351).

Методы swapToCopyWithMergedChild, swapToCopyWithDownChild, swapToCopyWithInsertedChild занимаются обработкой одной из возможных ситуаций. Все их перечисленных изменяют состояние текущего узла.

2.3 Реализация блокирующей многопоточной версии HAMT

Код, описанный в предыдущей главе, легко трансформируется в многопоточную блокирующую версию. Всё что, необходимо поменять – это добавить примитив синхронизации в класс Hamt (рисунок 2.3.1) и определить, в каких местах он будет использоваться.

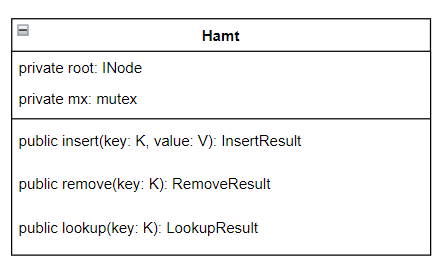


Рисунок 2.3.1 – UML-диаграмма класса Hamt для блокирующе многопоточной версии

В качестве примитива синхронизации был выбран мьютекс. Это наиболее простой примитив, позволяющий создавать критические секции. Чтобы попытаться захватить мьютекс, потоку необходимо воспользоваться методов lock(). Если это удаётся сделать, то он единолично входит в критическую секцию. Иначе, ожидает. После завершения работы в критической секции потоку необходимо отпустить притив при помощи метода unlock().

Как говорилось ранее, методы вставки, удаления, поиска логически разделены на публичные и приватные. Такая структура позволяет легко добавить использование мьютекса. Идея в том, что потоконебезопасный код внутри приватных метода необходимо окружить взятием и отпускание мьютекса (рисунок 2.3.2).

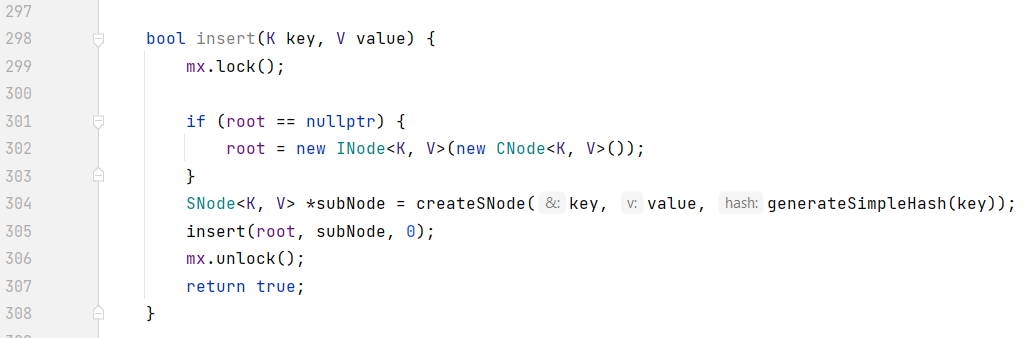


Рисунок 2.3.2 – Реализация публичного метода вставки в блокирующей версии

2.4 Реализация lock-free HAMT

Lock-free версия HAMT является закономерным продолжением блокирующей многопоточной версии, которая, в свою очередь, была трансформирована из однопоточной.

Поэтому при реализации lock-free версии необходимо было постепенно сужать критические секции в блокирующей версии и вводить cas-операции.

Главным отличием lock-free является то, что узлы дерева изменяются только при cas операции. Поток всегда сначала копирует узел, который он хочет изменить, меняет эту локальную копию и уже потом пытается вставить свою версию. Если вставка своей версии (операция cas) проходит неуспешно, значит другой поток уже заменил соответствующий узел и наши изменения неактуальны. Поэтому необходимо повторить операцию.

Код для вставки проходит через следующие этапы (рисунки 2.4.1, 2.4.2):

1. Запускается публичный метод, в бесконечном цикле происходит попытка вставить новый узел. Либо вставка происходит в корне, либо вызывается приватный метод insert для вставки ниже корня.
2. В приватном методе фиксируется текущая CNode и родитель, если он есть. (строки 398-399)
3. Попытка сократить дерево (строки 441-443). Метод производит сокращение текущего узла, если в этом есть необходимость. Возвращаемое значение сигнализирует о том, необходимо ли заново провести операцию вставки или нет.
4. Получение копии текущего CNode (строка 445)
5. В зависимости от ребёнка, произвести изменения локальной копии (строки 450-475)
6. Попытаться вставить при помощи новую версию CNode при помощи cas-операции (строки 453, 463, 467, 453)
7. Если не получилось, то повторяем вставку через вызов публичного метода (строка 474)

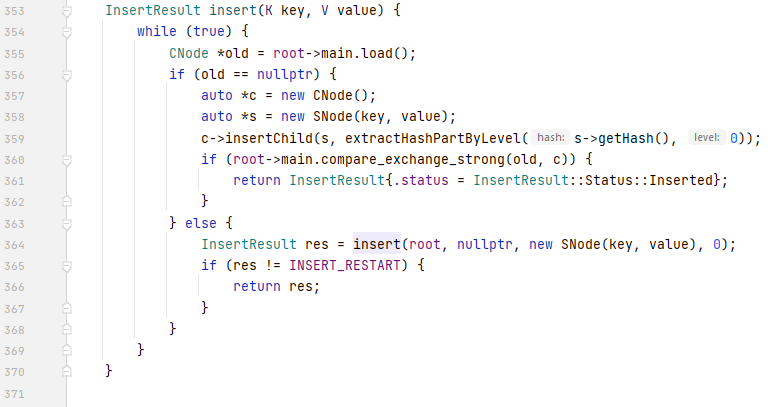
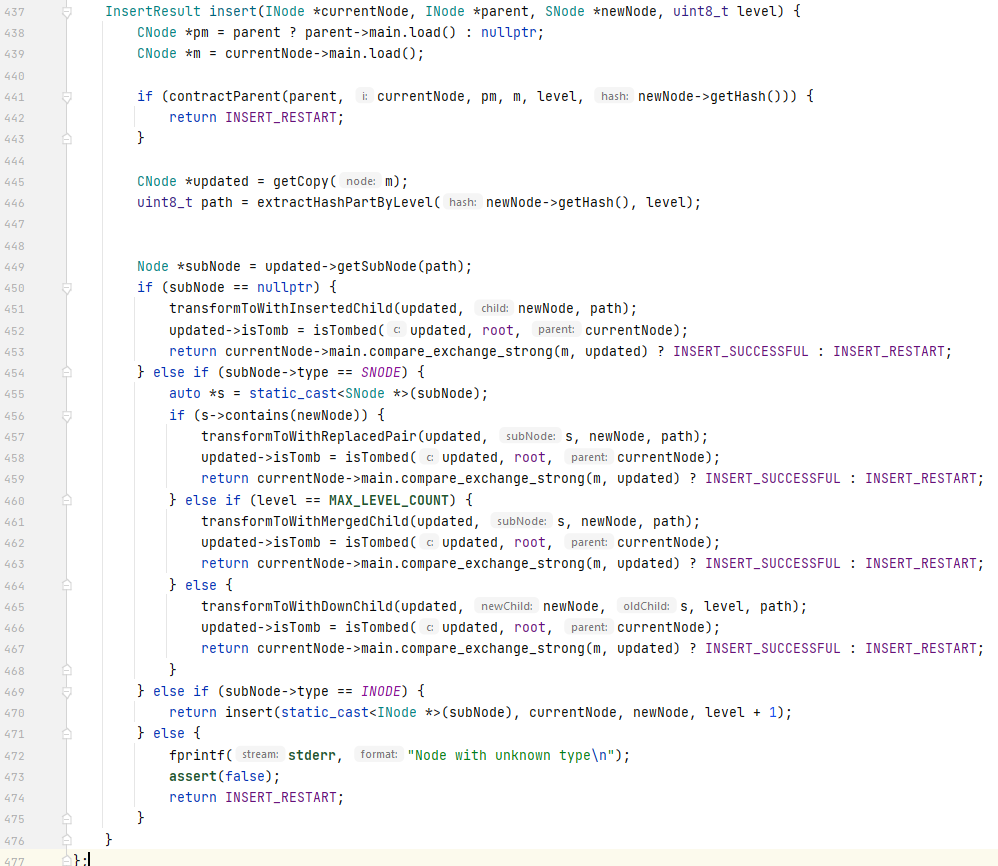


Рисунок 2.4.1 – Реализация приватного метода вставки в lock-free версии

 Рисунок 2.4.2 – Реализация приватного метода вставки в lock-free версии

Стоит обратить внимание, что рекурсивному алгоритму был предпочтён не рекурсивный через бесконечный цикл. Это помогает избежать проблемы переполнения стека при большом количестве неуспешных попыток произвести вставку, а значит большом количестве рекурсивных вызовов функций.

2.5 Решение проблемы ABA

2.6 Внедрение в библиотеку libcds

# 3 ТЕСТИРОВАНИЕ

# 4 ЗАКЛЮЧЕНИЕ

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Aleksandar Prokopec, Phil Bagwell, Martin Odersky Cache Aware Lock-Free Concurrent Hash Tries [Электронный ресурс] – 2011. – URL: https://www.researchgate.net/publication/313432754\_Cache-Aware\_Lock-Free\_Concurrent\_Hash\_Tries (дата обращения 15.04.2022)
2. Fredkin E. Trie memory. Communications of the ACM 3 [Электронный ресурс] – 1960 – URL: https://dl.acm.org/doi/10.1145/367390.367400 (дата обращения 15.04.2022)
3. Smotherman M. History of multithreading [Электронный ресурс] – 2005. – URL:   
   https://people.cs.clemson.edu/~mark/multithreading.html#:~:text=History%20of%20Multithreading&text=Summary%3A%20Multithreading%20first%20appeared%20in,investigated%20by%20IBM%20in%201968.&text=Most%20attempts%20at%20a%20history,PPUs%20in%20the%20CDC%206600 (дата обращения 15.04.2022)
4. Venu B. Multi–core processor – An overview [Электронный ресурс] – 2011. – URL: https://www.researchgate.net/publication/51945986\_Multi-core\_processors\_-\_An\_overview (дата обращения 15.04.2022)
5. Phil Bagwell. Ideal Hash Trees [Электронный ресурс] – 2001. – URL: https://www.researchgate.net/publication/2378571\_Ideal\_Hash\_Trees (дата обращения 15.04.2022)
6. Time and Space Lower Bounds for Non-Blocking Implementation [Электронный ресурс] – 1996 – URL: https://dl.acm.org/doi/10.1145/248052.248105 (дата обращения 15.04.2022)