**Parallel Bubble Sort**

*(Паралелно изчисление на Bubble Sort - статично балансиране и настройваема грануларност)*

**

Изготвен под ръководството на:

Проф. Д-р Васил Цунижев

Ас. Христо Христов

Изготвил: Божидар Абаджиев

ФН: 9MI0600004, група 4, СИ

Съдържание:

1. Въведение:
   1. Сортирания
      1. Case - worst, average, best
      2. Space complexity
   2. Цел – паралелна версия
2. Анализ
   1. Какво представлява Bubble Sort
   2. Сортираща мрежа
   3. Паралелно сортиране Odd-Even
   4. Паралелен Bubble Sort – заключване на всички елементи
   5. Паралелен Bubble Sort
3. Проектиране
   1. Функционално проектиране
      1. Представяне чрез диаграми
   2. Представено чрез код
   3. Технологично проектиране
4. Тестови резултати
5. Списък с източници
6. Въведение
   1. Сортирания

Нека да започнем като дадем дефиниция на сортиране - Сортирането е процесът на нареждане на елементи от множество в определен ред. Този ред обикновено е възходящ (от по-малки към по-големи) или низходящ (от по-големи към по-малки), в зависимост от изискванията на конкретната задача. Сортирането има широко приложение в програмирането и информатиката, като помага за по-бързо търсене, сравнение и обработка на данни.

Различните видове сортирания имат различни критерии за сравнение. Един от тях е ефективността на алгоритъма - колко бързо може да сортира списъка. Някои алгоритми са по-бързи от други в най-лошите случаи или имат по-голяма степен на ефективност при специфични видове данни. Друг важен критерий е стабилността на сортирането. Стабилен сортиращ алгоритъм запазва реда на равнозначните елементи във входния списък, докато нестабилният алгоритъм не гарантира това. Този аспект е особено важен при сортиране на данни с множество ключове.

Допълнителната памет, която изисква алгоритмът за сортиране, също е важен фактор за сравнение. Някои алгоритми изискват допълнително място за сортиране на данните, докато други работят върху входния списък. И накрая, сложността по време е съществен аспект за оценка на ефективността на алгоритмите за сортиране. Тя показва колко време е необходимо на алгоритъма да сортира входния списък и се измерва в зависимост от броя на сравненията и обмените, които извършва.

* + 1. Сравнение по сложност (worst, average, best case)

В таблицата(фиг 1) представих различни алгоритми за сортиране - bubble sort, selection sort, insertion sort, merge sort, quick sort и heap sort. Всеки от тези алгоритми разполага със свой уникален метод за сортиране на елементите в списъка. За всеки алгоритъм съм посочил времевата сложност в най-лошия, средния и най-добрия случай. Изборът на подходящ алгоритъм за конкретна задача често зависи от различни фактори като размерът на входните данни, наличието на допълнителна памет и желаната ефективност.

Фиг 1. Времеви сложности при различните алгоритми за сортиране



* + 1. Space complexity



Фиг 2. Времеви сложности при различните алгоритми за сортиране

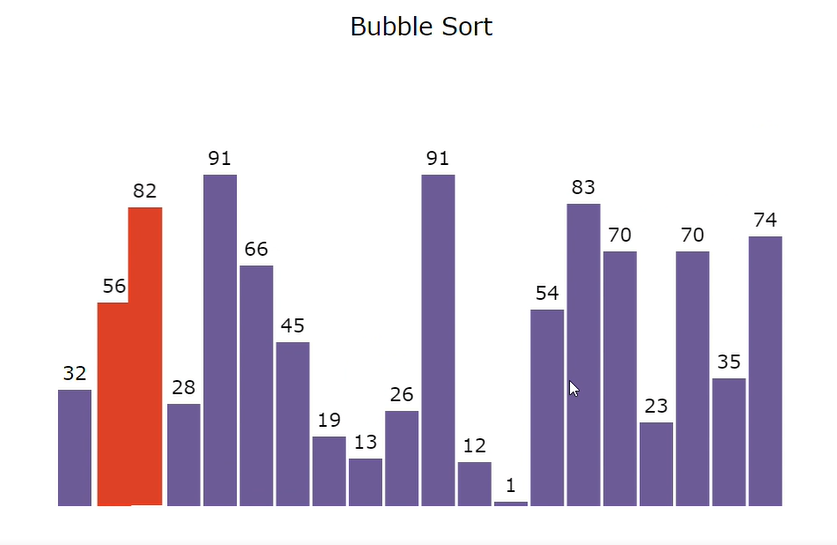
Тук нагледно може да видите и информация за пространствената сложност в най-лошия случай.

* 1. Цел на проекта

Parallel Bubble Sort се стреми към конкретна паралелна реализация на сортирането по метода на мехурчето. Създаването на на паралелно такова сортиране, би подобрило времевата сложност и би използвало същата пространствена сложност(понеже няма да създава нов масив). Главната му цел е да подобри ефективността на сортирането на случаен масив чрез намаляване на времето, необходимо за тази операция. Сортиращият алгоритъм ще бъде тестван, чрез: въвеждане на различни входни данни, тестване на различни устройства, пускане на различен брой нишки и постоянство на връщане на правилни отговори.

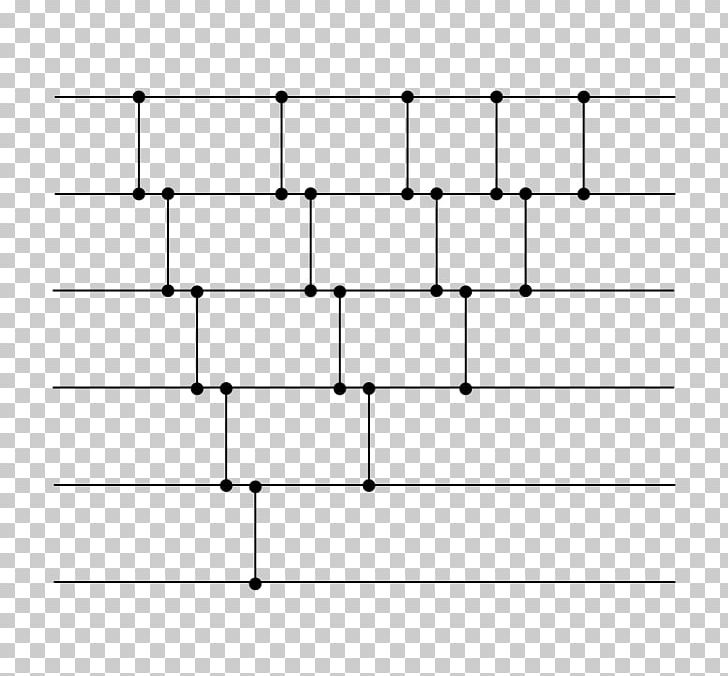
1. Анализ
   1. Представяне на Bubble sort

Алгоритъмът работи като започва от началото на списъка и сравнява всеки два последователни елемента. Ако текущият елемент е по-голям от следващия, те се разменят. Този процес се повтаря над списъка, като всяка итерация "избутва" по-голямите елементи към края на списъка. Този процес продължава докато не се извършат итерации, в които не се правят никакви размени, което означава че списъкът вече е сортиран.

Фиг 3. Изпълнение на Bubble Sort

Както се вижда на фигура 3, елемент 2 и елемент 3 са били разменени като се вижда, че това е първата итерация по започване на сортиране на масива, защото последният не е най-голямото число.

* 1. Сортираща мрежа



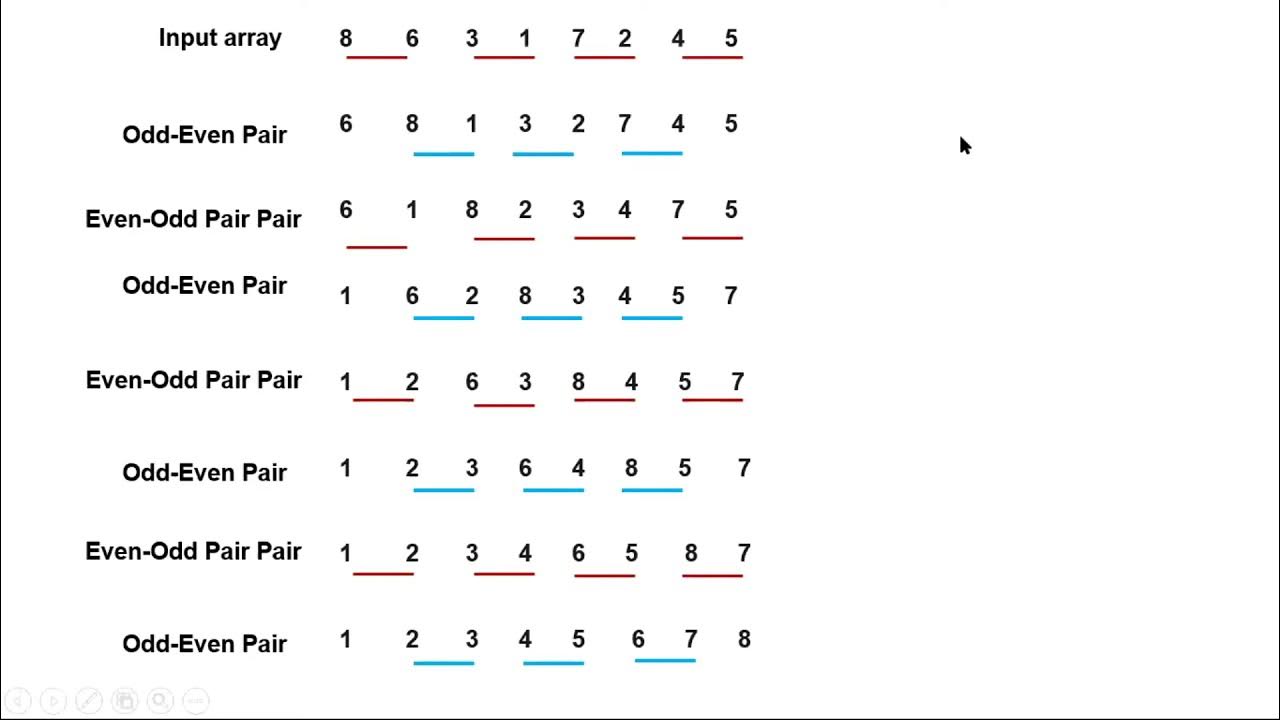
При неизвестен брой елементи една итерация(минаване през всички елементи и сравняването им, даден елемен с неговия следващ) ще се извършва от един процес.

Фиг 4. Сортираща мрежа

Ако разпределим работата на няколко процеса и примерно първия отговаря за намирането на 2та най-големи елемента, това ще е много повече работа отколкото друг процес, която ще трябва да намери 2та най-големи от вече останалите 2/3/4… елемента. Така ако изразим Bubble Sort визуално, ще видим веднага, че се забелязва неравномерно разпределената работа на процесите, докато първия процес ще трябва да извършва много сравнения, то работата на последния е почти нищожна.

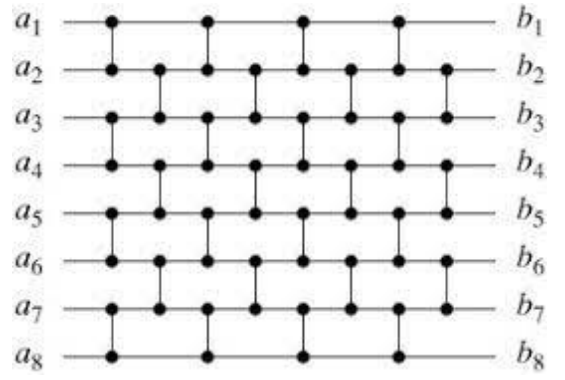
* 1. Паралелно сортиране Odd-Even

Един от популярните подходи за паралелно сортиране е методът Odd-Even. Този метод често се свързва с паралелната версия на сортирането по метода на мехурчета. В същността си, Odd-Even методът използва сортираща мрежа, където при сравнение на две числа можем да видим два възможни сценария: числата да запазят позициите си или да се разместят, ако първото е по-голямо от второто. Процесът на сортиране се изпълнява като групираме масива в четни и нечетни двойки. Всяка итерация сравнява всички нечетни или четни двойки, като всяка част от процеса е отговорна за определен брой четни или нечетни двойки.



Фиг 5. Изпълнение на Odd-Even Sort

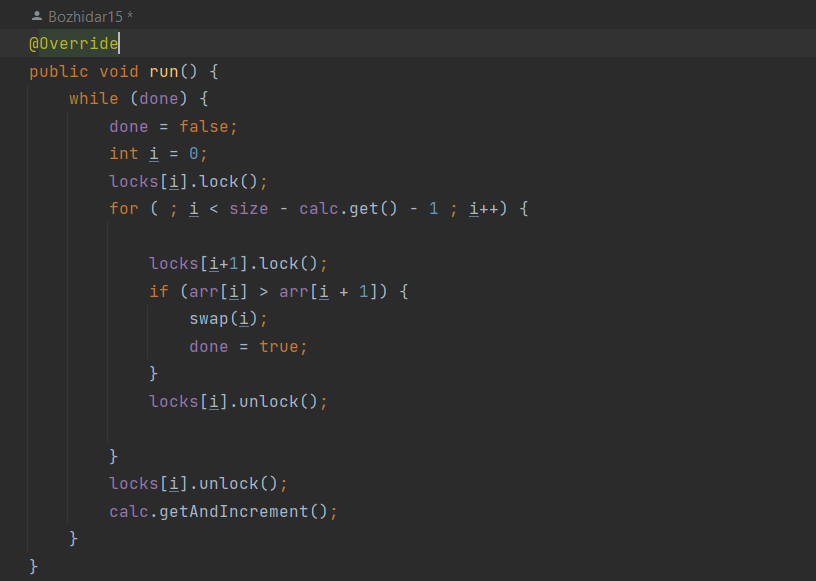
Този метод е добре структуриран, като всяка стъпка от процеса е ясно дефинирана. Балансът между работата на отделните процеси е оптимален, а методът е сравнително лесен за изпълнение. Гъвкавостта на метода позволява адаптиране към размера на кеша и броя на процесорите. При по-малък брой процеси се предпочита по-фина гранулация, докато при по-голям брой процеси се използва по-едра гранулация.

Фиг 6. Сортираща мрежа на odd-even sort

* 1. Паралелен Bubble Sort – заключване на всички елементи

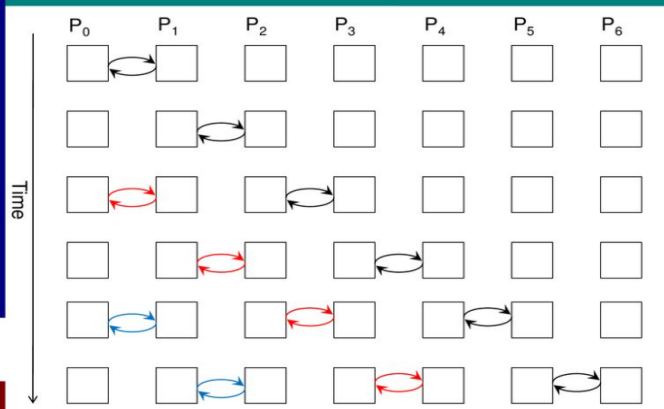
След като разгледахме горните две идеи, стигаме до извода, че Bubble Sort може да бъде оптимизиран, ако създадем комбинирана идея между двете. Забелязва се, че смените се случват в пъти повече в началните елементи отколкото в крайните елементи. Което ни подсказва, че трябва да обработваме елементите постоянно.

Стигаме до идеята, че има ключалка(критична секция) за всеки елемент. Всяка нишка ще започва от първия елемент на масива като ще обработва два елемента от масива(ел1 и ел2) и след като приключи работа с тях, запазва ключа за 2рия, освобождава първия и взима следващият в масива. Като за новата двойка елементи действието по сравняване и размяна се извършва отново. А при освобождаването на елемент1 друга нишка вече може да го вземе и да го използва за ново сравнение и размяна.

Фиг 7. Идея със заключване на всеки елемент

Правейки няколко теста и замервания стигаме до извода, че при голям брой елементи в масива процеса по създаване на същия брой ключалки и постоянното им отключване и заключване е доста скъп и бавен процес.

* 1. Паралелен Bubble Sort



Фиг. 8 Parallel Bubble Sort

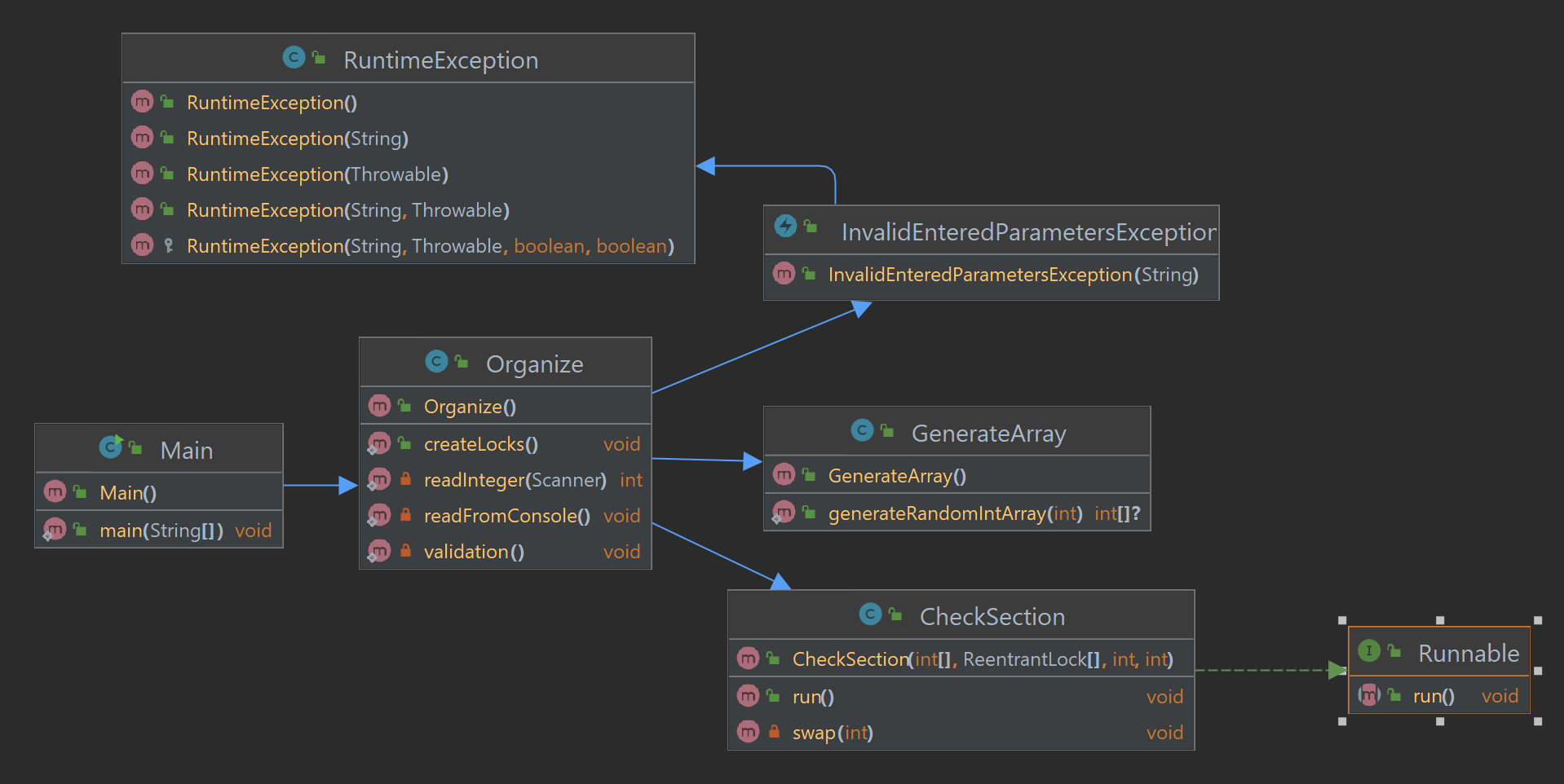
Всичко започва с работата на main нишката, която пуска всички останали(сортиращи нишки). Всяка нишка ще обработва даден сегмент(точно определено множество от елементи) от масива и след като приключи работа по него минава на следващия, като току-що обработената част я предава на следващата нишка. Разбира се правят се проверки най-големият елемент досега да продължава напред по масива за да може да стигне на правилното място.

Какво очакваме да се случи при развитието на идеята – ускорение при увеличаване на нишките.

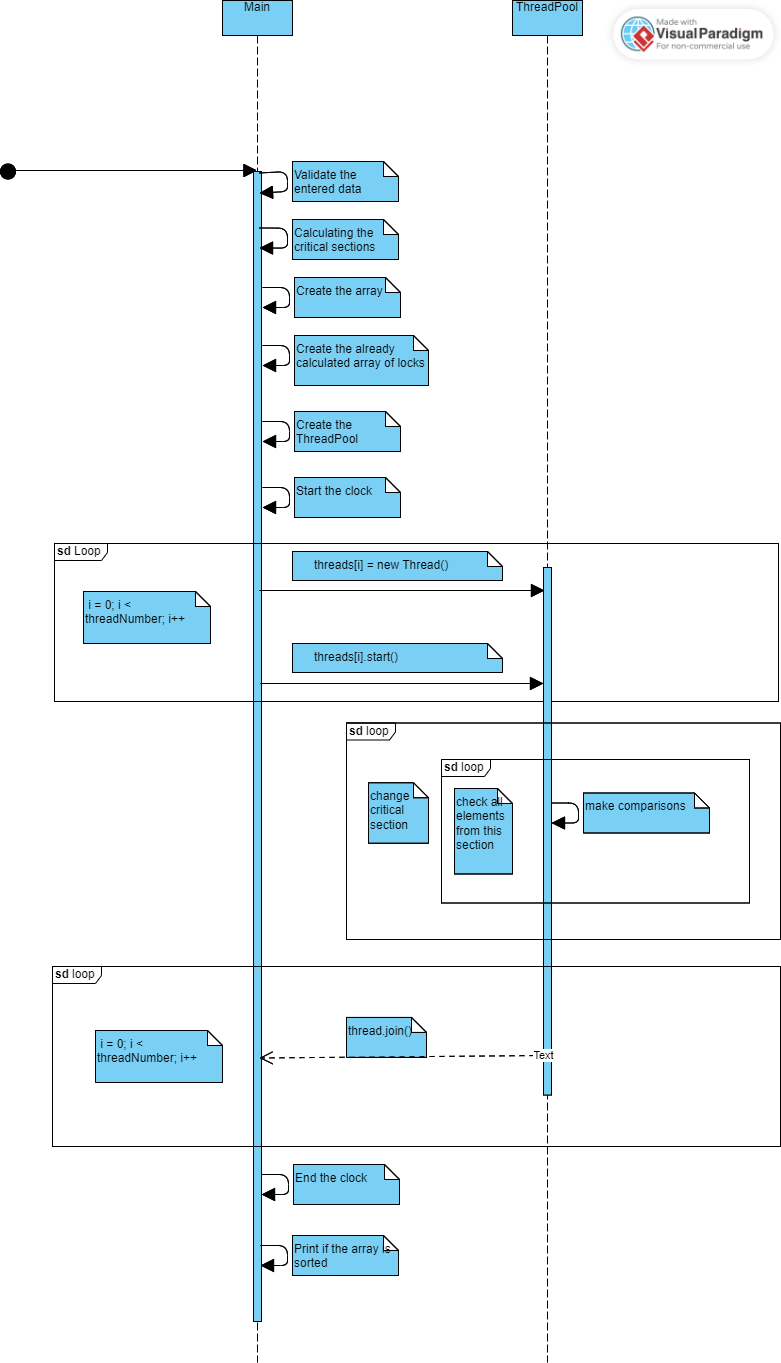
1. Проектиране
   1. Функционално проектиране

Първо започваме с проверка и валидация на въведените/подадени данни от потребителя за броя на нишките и размера на масива. Следва една от най-важните проверки – дали размера на масива се дели без остатък или не. От това зависи колко критични секции(ключалки ще имаме). Ако няма остатък и се дели точно създаваме толкова критични секции, колкото са нишките. Ако има остатък обаче създаваме толкова критични секции, колкото са нишките плюс една(последната секция няма да е със същия брой като всички останали).

Идеята е че когато достигнем границата на текущия блок, разблокираме текущата критична секция и увеличаваме брояча на критичните секции, за да преминем към следващата критична секция. След това генерираме нова точка за освобождаване. При завършване на цикъла, последната нишка освобождава и последната критична секция. За да се уверим, че целият масив е сортиран, проверяваме дали има смяна в блока от масива и ако има, правим още едно минаване през алгоритъма. Ако няма смяна никъде, значи масивът е сортиран. Важно е да се избягва ситуацията на race condition или deadlock. Затова е от съществено значение да се следи реда на заключване и отключване на критичните секции и да се поддържа баланс между броя на работещите нишки и сегментите на масива, като се намалява броят на работещите нишки при нарастване на техния брой.

* + 1. Представяне чрез диаграми

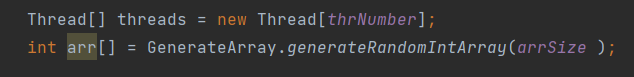
Фиг. 9 Class diagram



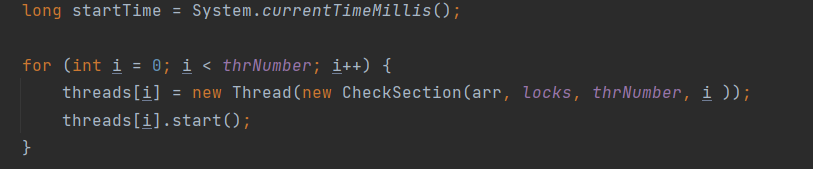
Фиг. 10 Sequence diagram

* 1. Представено чрез код

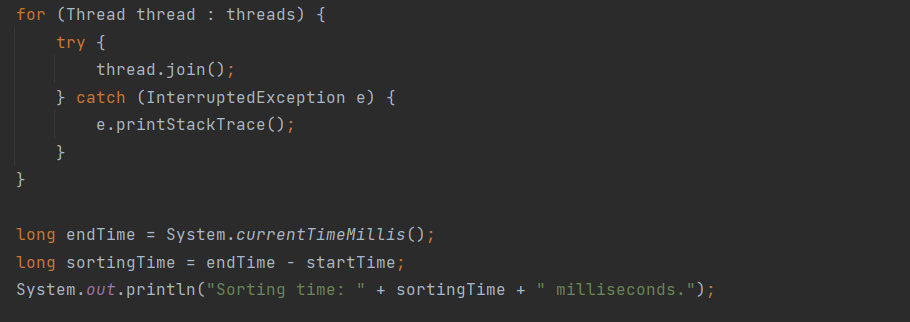
Първа стъпка е да създадем thread pool (банка/масив в която/който ще се съдържат нишките) и да се създаде масива с random числа.



Продължаваме със запазване на точния момент, който да използваме за стартов. С цикъл обхождаме thread pool и стартираме всяка една нишка като и подаваме самият масив с числа, масивът с критични секции, брой на нишките и позиция на съответната нишка.



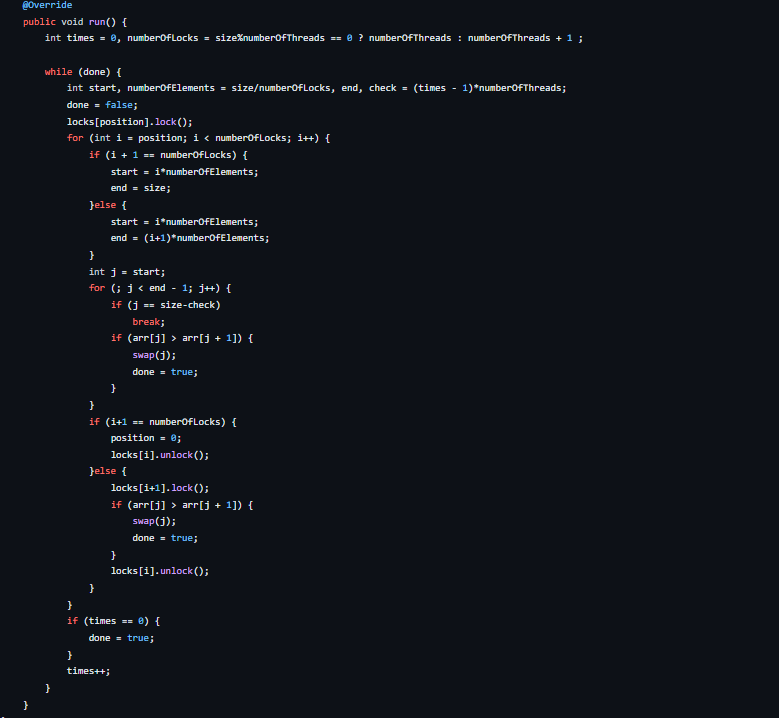
Първо ще разгледаме как main нишката ни продължава и после ще разгледаме метода run() на нишките. В main нишката правим цикъл, с който я караме да изчака всяка една друга стартирана сортираща нишка. Това се случва чрез метода join(). Щом main нишката мине и прочете името на избрана стартирана нишка и после join() след името – това приспива main докато избраната нишка не си изпълни целия run метод.



Взимаме времето след като всички сметки са приключили, вадим от крайното, началното и имаме за колко време се е случило всичко.

Сега нека разгледаме run() метода на сортиращите нишки:

Първо си създаваме променливите times – колко пъти сме минали масива досега, numberOfLock - критични секции имаме, Boolean done = true – създадена променлива с стойност истина, която да поддържа цикъла, при всяко влизане в него я правим false, но ако има swap (смяна на елементи) я променяме на true, така цикъла ще се върти, докато няма извършена смяна, тоест е нареден масива. При създаването на нишката последния параметър беше номера под който е пусната нишката - position. Следователно ако е стартирана първата нишка, тя ще е с позиция 0 и ще заключи съответно критична секция 0. След влизането във втория цикъл правим изчисления за съответната критична секция от кой до кой индекс на масива може да минава нишката. В третия цикъл сравняваме елементите само в тази секция. След приключването на третия цикъл, правим проверка дали досега сме били в последната секция или в някоя друга и предприемаме различни действия по отключване на секцията.



* 1. Технологично проектиране

 Фиг. 11 Таблица с технически характеристики

За тестването са използвани следните две тестови среди, като в следващата точка са представени резултатите и от тези двете тестови машини.

1. Тестови резултати

Ще тестваме при различен размер на входния масив и различен брой нишки и ще наблюдаваме резултатите от:

• Size- размерът на входния масив

• p - брой нишки

• window– размерът на един блок от масива

• Тp (i) - времето за изпъленение на тест i с p нишки

• Tp = min(Tp (i))

• Sp = T1/Tp - ускорение

• Ep = Sp/p – ефективност

* 1. My PC



Фиг. 12 Таблица с данните от My PC

Отдолу ще видите Line Graph относно ускорение и ефективност съответно за различните случаи – с различен брой елементи в масива и изпълнено с различен брой нишки. Опитите ще бъдат показани графично и за двете машини. Изобразени са първо с size =10 000 / 100 000 / 200 000, p = 1 / 2 / 4 / 6 / 8.

Фиг. 13. Ускорение за n = 10 000 Фиг. 14. Ефективност за n = 10 000

Фиг. 15. Ускорение за n = 100 000 Фиг. 16. Ефективност за n = 100 000

Фиг. 17. Ускорение за n = 200 000 Фиг. 18. Ефективност за n = 200 000

По данните, представени в таблицата и графиките, се забелязва, че при изпълнението на трите експеримента има вариации в ускорението, като входните масиви са с различна големина – при 10 000 елемента ускорението не е високо колкото очакваме – при повече елементи – се получава по добро ускорение – със статична грануларност. При 100 000 елемента, поделени на 8 нишки се достига най-високо ускорение ~ 3,5. При 200 000 елемента пак може да наблюдаваме добро ускорение от около ~ 2,5.

* 1. Powerful PC



Фиг. 19 Таблица с данните от Powerful PC

Фиг. 20. Ускорение за n = 10 000 Фиг. 21. Ефективност за n = 10 000

Фиг. 22. Ускорение за n = 100 000 Фиг. 23. Ефективност за n = 100 000

Фиг. 24. Ускорение за n = 200 000 Фиг. 25. Ефективност за n = 200 000

Из анализа на данните можем да извлечем заключението, че за малки масиви паралелната обработка не е ефективна, но с нарастването на размера на масива се наблюдава поне някакво подобрение в ускорението. Стигаме до заключението, че общата последователна зависимост (например, при Bubble sort) пречи на паралелизма, тъй като не всички процеси могат да работят едновременно през цялото време, което допълнително забавя процеса. Всяка нишка често чака вместо да извършва работа при малък брой елементи. Едно възможно подобрение би било да стартираме не всяка нишка отначало, а само в определени блокове. Важно е да отбележим две неща: първо, прехвърлянето на данни от един блок в друг, и второ, да разберем кога масивът е сортиран. Алгоритъма може да се подобри, както е описано в една от книгите по-долу като след свършване на дадена секция се използва merge sort. По този начин нишките ще работят асинхронно и ще се ползва добрата сложност на merge sort.

1. Списък с източници

* [1] Peter Pacheco – “An Introduction to Parallel Programming” - Morgan Kaufmann Publisher - [http://www.e-tahtam.com/~turgaybilgin/2013-2014- guz/ParalelProgramlama/ParallelProg.pdf](http://www.e-tahtam.com/~turgaybilgin/2013-2014-%20guz/ParalelProgramlama/ParallelProg.pdf)
* [3] Kai Hwang, Geoffrey C. Fox, Jack J. Dongarra – “Distributed and Cloud Computing From Parallel Processing to the Internet of Things” - Morgan Kaufmann Publisher - <https://cutepooji.files.wordpress.com/2017/01/distributed-and-cloud-computing-fromparallel-processing-to-the-internet-of-things.pdf>
* [5] Article „Parallel Bubble Sort“ - <https://iq.opengenus.org/parallel-bubble-sort/>
* [8] Reem Saadeh and Mohammad Qatawneh - Performance Evaluation Of Parallel Bubble Sort Algorithm On Supercomputer Iman1- International Journal of Computer Science & Information Technology (IJCSIT) , 2019 - [https://www.academia.edu/39787015/PERFORMANCE\_EVALUATION\_OF\_PARALLEL\_BUBBL E\_SORT\_ALGORITHM\_ON\_SUPERCOMPUTER\_IMAN1](https://www.academia.edu/39787015/PERFORMANCE_EVALUATION_OF_PARALLEL_BUBBL%20E_SORT_ALGORITHM_ON_SUPERCOMPUTER_IMAN1)
* [9] Zaid Abdi Alkareem Alyasseri Kadhim Al-Attar, Mazin Nasser - Parallelize Bubble Sort Algorithm Using OpenMP – ISMAIL - <https://arxiv.org/ftp/arxiv/papers/1407/1407.6603.pdf>