МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ТЕРНОПІЛЬСЬКИЙ НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ІВАНА ПУЛЮЯ
ФАКУЛЬТЕТ КОМП'ЮТЕРНО-ІНФОРМАЦІЙНИХ СИСТЕМ І ПРОГРАМНОЇ ІНЖЕНЕРІЇ

ХУДОБА ВІКТОР ВОЛОДИМИРОВИЧ

УДК 004.9:504:519.6

АЛГОРИТМІЧНЕ, ПРОГРАМНЕ ТА АПАРАТНЕ ЗАБЕЗПЕЧЕННЯ КОМП'ЮТЕРНИХ СИСТЕМ ПАРАЛЕЛЬНОГО ОПРАЦЮВАННЯ ВЕЛИКИХ ДАНИХ НА ПЛАТФОРМІ JAVA

123 «Комп'ютерні системи та мережі»

Автореферат
дипломної роботи на здобуття освітнього ступеня «магістр»

Тернопіль
2018

Роботу виконано на кафедрі комп'ютерних систем та мереж Тернопільського національного технічного університету імені Івана Пулюя Міністерства освіти і науки України

Керівник роботи: кандидат технічних наук, доцент кафедри комп'ютерних систем та мереж
Луцків Андрій Мирославович,
Тернопільський національний технічний університет імені Івана Пулюя,

Рецензент: кандидат технічних наук, доцент, зав.кафедри фізики
Скоренький Юрій Любомирович,
Тернопільський національний технічний університет імені Івана Пулюя,

Захист відбудеться 27 грудня 2018 р. о $9^{.00}$ годині на засіданні екзаменаційної комісії №34 у Тернопільському національному технічному університеті імені Івана Пулюя за адресою: 46001, м. Тернопіль, вул. Руська, 56, навчальний корпус №1, ауд.1-603

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми роботи. Опрацювання великих даних, на сьогодні, є актуальною задачею. В основі переважної більшості інструментів для опрацювання великих даних лежить технологія Java. Важливо вміти адекватно її застосовувати для реалізації різноманітних задач. Сфера великих даних об'єднує цілу низку галузей комп'ютерної інженерії та інших прикладних галузей IT.

На сьогодні існує велика кількість програмних бібліотек, фреймворків та сервісів, які реалізовані на Java: Hadoop, Spark (Scala), HBase, Hive та багато інших, які надають зручні API-інтерфейси для Java-розробників. Враховуючи, що раніше Java утримувала традиційно ніші фінансового сектору, енергетики і суміжних галузей, то сьогодні Java-розробники набувають компетенцій у сферах DataScience та аналітики, штучного інтелекту та машинного навчання, роботи з InternetOfThings. Враховуючи, що сфера великих даних дедалі простіше інтегрується в різні сфери народного господарства потреба в таких компетенціях тільки зростатиме. Проблематикою створення ефективних програм на платформі Java займались Doug Lea, Josh Bloch, Rajat Mehta, Brian Goetz, David A. Patterson, John L. Hennessy та багато інших

Актуальною є задача створення коректного програмного забезпечення, яке б було достатньо ефективним з точки зору часу його виконання та кількості використовуваної пам'яті, вибору адекватних алгоритмів та структур даних, роботи з введенням/виведенням. Важливою є й задача коректного оцінювання ефективності написаного коду з можливістю визначення характеристик його виконання.

Доцільність дослідження даної теми зумовлена необхідністю розробки високоефективних систем опрацювання великих даних для різнотипних задач. Оскільки, технологія Java є найпоширенішою, кросплатформовою й за останні 20 років продемонструвала достатній рівень зрілості технології, то очевидно, що її використання буде тільки зростати. Попри гостру конкуренцію з іншими технологіями програмування: Python, JavaSript, C# та інших, Java як мова й платформа є гнучкою, недорогою, надійною, добре документованою та ефективною. З урахуванням нового підходу до підготовки релізів нововведення та кращі техніки програмування з'являтимуться в Java. Проте, загальні принципи залишаються тими ж. Багато концептуальних засад мови пройшли випробування часом й зарекомендували себе з кращого боку, зокрема, принцип багатопотоковості мови, який на сьогодні набув багатьох корисних фреймворків та бібліотек.

В контексті великих даних важливу прикладну цінність має обґрунтування вибору тих чи інших парадигм програмування високопродуктивних обчислювальних систем, які можуть бути використані в інших предметних областях. Потрібно обирати універсальні парадигми (патерни) паралельного та розподіленого програмування у рамках певних бібліотек і технологій програмування.

Створення ефективних програмних компонентів комп'ютерних систем опрацювання великих даних на базі платформи Java дає змогу шляхом споживання меншої кількості системних ресурсів отримати результати опрацювання даних за менший час.

Мета роботи: Метою магістерського дослідження є обґрунтування вибору ефективних методів та засобів опрацювання великих даних на платформі Java.

Досягнення цієї мети вимагає розв’язання таких завдань:

1. Проведення аналізу предметної області програмного забезпечення для опрацювання великих даних, з метою формулювання основних проблем та вимог, які ставляться до такого програмного забезпечення.

2. Обґрунтувати вибір того чи іншого підходу (фреймворку), який би давав змогу ефективно використовувати ресурси SMP-систем.

3. Провести дослідження типових алгоритмів та структур даних, які найчастіше використовуються у відповідних обчислювальних задачах.

4. На основі аналізу хороших практик узагальнити рекомендації щодо створення паралельних програм опрацювання великих даних для платформи Java.

5. Створити та впровадити комп’ютерні програми для апробації запропонованих фреймворків та методологій створення відповідних програм, й, таким чином, провести апробацію основних положень магістерської роботи.

Об’єкт дослідження: обчислювальні процеси у комп'ютерних системах опрацювання великих даних.

Предмет дослідження: алгоритми, патерни паралельного програмування, методи декомпозиції обчислювальних задач, паралельні та розподілені обчислювальні системи, віртуальна машина Java.

Методи дослідження: моделювання комп’ютерних систем та програм, теорія алгоритмів та обчислювальних методів, теорія компіляторів, теорія побудови обчислювальних систем.

Наукова новизна отриманих результатів:

- вперше, на основі аналізу алгоритмічного та програмного забезпечення запропоновано методики проведення обчислювальних експериментів з метою визначення реальної ефективності виконання коду програми;
- узагальнено рекомендації по розробленню паралельного програмного забезпечення опрацювання великих даних для платформи Java.

Практичне значення отриманих результатів.

Урахування принципів та рекомендацій щодо розроблення програмного забезпечення паралельного опрацювання великих даних дають змогу створювати ефективне з точки зору використання системних ресурсів програмне забезпечення. Шляхом проведення обчислювального експерименту апробовано методи запропоновані в дипломній роботі магістра.

Апробація результатів дипломної роботи магістра. Результати дипломної роботи магістра апробовано на двох конференціях:

- міжнародній науково-технічний конференції молодих учених та студентів «Актуальні задачі сучасних технологій» (Тернопіль, ТНТУ, 2018);
- VI науково-технічній конференції «Інформаційні моделі, системи та технології» Тернопільського національного технічного університету імені Івана Пулюя (2018).

Структура роботи. Робота складається з розрахунково-пояснювальної записки та графічної частини. Розрахунково-пояснювальна записка складається з вступу, 6 розділів, висновків, переліку посилань та додатків. Обсяг роботи:

розрахунково-пояснювальна записка – 133 арк. формату А4, графічна частина – 10 аркушів формату А1

ОСНОВНИЙ ЗМІСТ РОБОТИ

У вступі обґрунтовано актуальність й важливість даного дослідження та здійснено короткий огляд сучасного стану проблем у галузі розроблення програмного забезпечення комп'ютерних систем опрацювання великих даних. Охарактеризовано основні завдання, які необхідно вирішити у дипломній роботі магістра.

В розділі 1 «Аналіз предметної області комп'ютерних систем паралельного опрацювання великих даних на платформі JAVA» проведено аналіз предметної області розроблення програмного забезпечення комп'ютерних систем опрацювання великих даних. Проаналізовано задачі з опрацювання великих даних та їх характер. Проаналізовано основні задачі, які стосуються предметної області великих даних в контексті платформи Java. Виокремлено базові проблеми сфери опрацювання великих даних та високопродуктивних обчислювальних систем. Сформульовано основні задачі дипломної роботи магістра.

В розділі 2 «Алгоритмічне забезпечення комп'ютерних систем паралельного опрацювання великих даних на платформі JAVA» проведено аналіз алгоритмічної складності задач опрацювання великих даних, належність їх до відповідних класів складностей. Розглянуто алгоритмічну складність: час та пам'ять. Проаналізовано структури даних у мові Java, а також кількість пам'яті, який замають базові конструкції мови з урахуванням особливостей 64-бітних архітектур.

Показано доцільність використання патернів паралельних обчислень у SMP-системах у рамках багатопоточності технології Java. Розглянуто ефективність фреймворків ThreadPool та Fork-Join. Наведені переваги та недоліки даного фреймворку в контексті задач опрацювання великих даних.

Проаналізовано методи оцінювання ефективності використання обчислювальних ресурсів шляхом створення пулів потоків певного розміру. Відповідні методи дають змогу реалізувати програмне забезпечення, яке буде адаптуватись до типу обчислювальної системи.

В розділі 3 «Засоби та методи розробки програмного забезпечення комп'ютерних систем паралельного опрацювання великих даних на платформі JAVA» проаналізовано апаратне забезпечення комп'ютерних систем в контексті опрацювання великих даних та сформульовано рекомендації по вибору структур даних оптимальних з точки зору часу доступу до елементу в пам'яті залежно від типу задачі.

Обґрунтовано доцільність поєднання об'єктно-орієнтованої та функційної парадигм програмування для задач опрацювання великих даних.

Проаналізовано та обґрунтовано ефективних шаблонів зчитування великих даних на Java, що дасть змогу зменшити час доступу до даних та кількість використовуваної пам'яті.

Проаналізовано ефективності базових конструкцій мови Java та колекцій з точки зору їх ефективності роботи. Проаналізовано та обґрунтовано методи оцінювання ефективності коду, які не базуються на оптимізаціях JIT-компілятора

Java, враховують «гарячий» та «холодний» старт JVM та базуються на статистичних експериментах.

Проведено серію обчислювальних експериментів з використанням базових конструкцій мови Java, структур даних та фреймворків. Експерименти підтвердили теоретичні положення, які висвітлені в роботі. Зокрема, ефективність при опрацюванні великих файлів.

На основі проаналізованого матеріалу, який узагальнено, сформульовано рекомендації по розробленню ефективного програмного для опрацювання великих даних на SMP-системах.

В розділі 4 «Обґрунтування економічної ефективності» показано доцільність проведення науково-дослідних робіт за даною тематикою і економічно обґрунтовано доцільність застосування запропонованих засобів. Розраховано вартість та ціну проведено науково-дослідної роботи. Розкрито питання нової ліцензійної політики компанії Oracle та новий цикл релізів платформ OpenJDK та OracleJDK.

В розділі 5 «Охорона праці та безпека в надзвичайних ситуаціях» розглянуто вимоги до охорони праці користувачів ЕОМ, до яких належать науковці, розробники програмного забезпечення, користувачі, а також розглянуті вимоги до умов праці відповідно до сучасних нормативних документів. Це дало змогу забезпечити належний рівень умов праці.

У контексті забезпечення безпеки в надзвичайних ситуаціях у відповідному підрозділі:

- розглянуто питання запобігання забрудненню повітря, виробничих приміщень небезпечними хімічними речовинами, наведені допустимі значення для основних речовин-забруднювачів, їх характеристика та засоби захисту.
- наведено методику оцінювання стійкості роботи об'єкту господарської діяльності до дії проникаючої радіації і радіоактивного забруднення на підприємствах.

В розділі 6 «Екологія» проаналізовано абсолютні показники екологічних явищ, а також розглянуто питання моніторингу довкілля та система спостережень за впливом на довкілля антропогенних факторів.

У загальних висновках щодо дипломної роботи описано прийняті в роботі технічні рішення і організаційно-технічні заходи, які забезпечують виконання завдання на проектування; оригінальні технічні рішення, прийняті автором в процесі роботи.

В додатках до пояснювальної записки наведено фрагменти вихідного коду програм для проведення обчислювального експерименту на мові програмування Java.

В графічній частині наведено структурні схеми, які відображають архітектурні особливості створеної системи, її компоненти, методології оцінювання системних ресурсів. А також наведено основні задачі роботи та основні результати розробки.

ВИСНОВКИ

У даній магістерській роботі проведено дослідження алгоритмічного, програмного та апаратного забезпечення комп'ютерних систем опрацювання великих даних на платформі Java. Основні результати та висновки проведених теоретичних та експериментальних досліджень такі:

1. Проаналізовано предметну область великих даних, апаратного та програмного забезпечення комп'ютерних систем опрацювання великих даних. Проаналізовано основні задачі, які стосуються предметної області великих даних в контексті платформи Java. Виокремлено базові проблеми сфери опрацювання великих даних та високопродуктивних обчислювальних систем.

2. Проведено аналіз алгоритмічної складності задач опрацювання великих даних, належність їх до відповідних класів складностей. Розглянуто алгоритмічну складність: час та пам'ять. Проаналізовано структури даних у мові Java, а також кількість пам'яті, який замають базові конструкції мови з урахуванням особливостей 64-бітних архітектур.

3. Проаналізовано методи оцінювання ефективності використання обчислювальних ресурсів шляхом створення пулів потоків певного розміру.

4. Обґрунтовано використання фреймворку Fork-Join технології Java для виконання багатопотокових програм на системах зі спільною пам'яттю.

5. Сформульовано рекомендації по розробленню ефективного програмного для опрацювання великих даних на SMP-системах.

6. Проведено обчислювальний експеримент й апробовано запропоновані підходи. Обґрунтовано методи оцінювання ефективності коду, які не базуються на оптимізаціях JIT-компілятора Java, враховують «гарячий» та «холодний» старт JVM та базуються на статистичних експериментах.

СПИСОК ОПУБЛІКОВАНИХ АВТОРОМ ПРАЦЬ ЗА ТЕМОЮ РОБОТИ

1. Луцків А. М. Ключові особливості створення багатопотокових програм для платформи JAVA/ А.М. Луцків, В. В. Худоба // Актуальні задачі сучасних технологій : зб. тез доповідей міжнар. наук.-техн. Конф. Молодих учених та студентів, (Тернопіль, 28–29 листоп. 2018.) в 3-х томах /М-во освіти і науки України, Терн. націон. техн. ун-т ім. І. Пулюя [та ін.]. –Тернопіль : ФОП Паляниця В. А., 2018 – Т. 2. – 105-106с. [Електронний ресурс] Режим доступу: URL: http://elartu.tntu.edu.ua/bitstream/lib/26291/1/Book%202-2018.pdf

2. Луцків А. М. Шляхи підвищення ефективності опрацювання великих даних у середовищі JVM / А. М. Луцків, В. В. Худоба // Збірник тез доповідей VI Науково-технічна конференція «Інформаційні моделі, системи та технології», 12-13 грудня 2018 року. — Т. : ТНТУ, 2018. —С.77.

АНОТАЦІЯ

Худоба В.В. Алгоритмічне, програмне та апаратне забезпечення комп'ютерних систем паралельного опрацювання великих даних на платформі Java

Дипломна робота магістра, 123 – Комп'ютерні системи та мережі. – Тернопільський національний технічний університет імені Івана Пулюя, Тернопіль, 2018.

В дипломній роботі магістра виконано дослідження алгоритмічного, програмного та апаратного забезпечення комп'ютерних систем паралельного опрацювання великих даних на платформі Java. Аналізувались шляхи оптимізації Java-програм при опрацюванні великих даних з точки зору ефективності використання алгоритмів та структур даних на апаратному забезпеченні. Запропоновано методику оцінювання ефективності (тестування) створеного коду без JIT-оптимізацій.

У роботі використовується архітектура паралельної та розподіленої комп'ютерної системи на базі доступних компонентів: багатоядерних x86_64 процесорів, типової пам'яті та комунікаційних інтерфейсів. Обчислювальні системи об'єднані комунікаційним каналом GigabitEthernet.

Проаналізовано особливості створення багатопотокових програм на мові Java, зокрема з використанням бібліотеки java.util.concurrent. На основі аналізу бібліотек програм, фреймворків та різноманітних літературних джерел, узагальнено рекомендації яких варто дотримуватись при створенні Java-програм.

У роботі наведені результати оцінювання ефективності використання відповідних технологій. Застосування технології Java дало змогу використати усі конкурентні переваги даної мови програмування, зокрема простоту, надійність та високу ефективність. Використано Java 8 фреймворк Fork-Join.

Ключові слова: високопродуктивні обчислення, оптимізація, Java, Big Data, JVM

ANNOTATION

Khudoba V. Algorithms, software and hardware of computer systems of Java-platform parallel processing of big data

Master diploma thesis, 123 – Computer systems and networks - Ternopil Ivan Puluj National Technical University, Ternopil, 2018.

Master's degree thesis deals with the algorithms, software and hardware of computer systems for the parallel processing of large data on the Java platform. The ways of optimization of Java-programs in the processing of large data in terms of the efficiency of the algorithms use and data structures on the hardware were analyzed. The method of estimating the effectiveness (testing) of the generated code without JIT optimizations is proposed.

The work uses the architecture of a parallel and distributed computer system based on the available components: multi-core x86_64 processors, typical memory and

communication interfaces. The computing systems are connected by a communication channel GigabitEthernet.

The peculiarities of multithreaded programs development in Java, including using the java.util.concurrent library are analyzed. Based on the analysis of program's libraries, frameworks and various information resources, recommendations for Java-program developers are suggested.

The results of the relevant technologies effectiveness evaluation are presented in the work. The use of Java technology has made it possible to take advantage of all the competitive advantages of this programming language, including simplicity, reliability and high efficiency. Used Java 8 Framework Fork-Join.

Key words: high-performance computing, optimization, Java, Big Data, JVM