Syllabus
Course Program

Distributed and parallel computing

Specialty
113 Applied mathematics

Institute
Educational and Scientific Institute of Computer Science and Information Technology.

Educational program
Intelligent Data Analysis

Department
Computer Mathematics and Data Analysis.

Level of education
Bachelor's level

Course type
Special (professional), Mandatory.

Semester
6

Language of instruction
Ukrainian

Lecturers and course developers

Andrii Podorozhniak

andrii.podorozhniak@khpi.edu.ua
Candidate of Technical Science, docent, Professor of Computer Engineering and Programming Department.

Work experience – more than 20 years. The author more than 200 scientific, educational, and methodological works. Leading lecturer in the course: « Distributed and parallel computing», etc.

More about the lecturer on the department's website

General information

Summary

The course provides theoretical and practical training in the field of parallel and distributed computing, mastering the concepts of modern programming within the framework of parallel and distributed programming paradigms. The study is based on approaches to programming in multi-threaded systems, distributed systems, parallel computing systems, the problems of interoperability of parallel program processes and their synchronization are considered.

Course objectives and goals

The course aims to study the theoretical foundations and practical aspects of using parallel and distributed computing systems to solve complex applied problems with a large volume of calculations.

Format of classes

Lectures, laboratory classes, consultations, self-study. Final control in the form of an exam.

Competencies

SC 5. Ability to develop algorithms and data structures, software tools and program documentation.

SC 8. Ability to operate and maintain software of automated and information systems for various purposes.
SC 9. Ability to use modern technologies for programming and testing software.
SC 21. Ability to develop and operate software tools for processing large amounts of data based on information technologies of distributed and cloud computing.

Learning outcomes

LO 11. Be able to apply modern programming technologies and software development, software implementation numerical and symbolic algorithms.
L0 25. Be able to apply modern information technologies and software for processing large amounts of data based on distributed and cloud services.

Student workload

The total volume of the course is 90 hours (3 ECTS credits): lectures – 16 hours, laboratory classes – 16 hours, self-study – 58 hours.

Course prerequisites

To successfully pass the course, you must have knowledge and practical skills in the following courses: 'Computer discrete mathematics', 'Probability theory', 'Object-oriented programming', 'Theory and design of algorithms'.

Features of the course, teaching and learning methods, and technologies

Lectures are conducted interactively using multimedia technologies. Laboratory classes use a project approach to learning, game methods, and focus on the application of information technologies. Study materials are available to students through OneNote Class Notebook and Teams.

Program of the course

Topics of the lectures

Topic 1. An introduction to distributed and parallel computing.
Concepts of parallel and distributed computing. Sequential computing. Parallel computing. Means for performing parallel computing
Topic 2. Introduction to distributed computing. Model, specializations of distributed computing.
Concept of distributed computing and distributed system. Objectives of building distributed systems. Requirements for distributed systems.
Topic 3. Understanding big data.
Concepts and terminology. Characteristics of big data. Challenges of using big data. Definition of reference architecture. Justification of the use of architectural patterns. Overview of tools for working with big data and their analysis.
Topic 4. Big data storage and processing concepts.
Clusters. File systems and distributed file systems. NoSQL database. Sharding. Replication. A combination of sharding and replication. CAP theorem. ACID database design principle. BASE principle. Parallel data processing. Distributed data processing. Processing work tasks. Cluster. Processing in batch mode. Real-time processing.
Topic 5. Distributed batch and stream processing of unstructured and loosely structured data.
Apache Spark framework. Features, limitations, architecture and principle of operation of Apache Spark components. Working with data in Spark - RDD, DataFrame, DataSet. Spark RDD operations. Spark SQL DataFrame.
Topic 6. Development of data processing applications running in a distributed computing environment.
Apache Hadoop. The Hadoop ecosystem. Architectural concept of Hadoop. Hadoop daemons. Hadoop deployment methods. Hadoop Distributed File System. HDFS architecture. Data replication. HDFS file operations – write operation, read operation, delete operation.
Topic 7. Cluster resource management and task planning.

YARN is a Hadoop resource scheduler. YARN architecture. Key components of Hadoop YARN.
Topic 8. A programming model for parallel and distributed processing of large data sets.
Apache Hadoop MapReduce. MapReduce stages. Map and Reduce tasks. MaprReduce examples.

Topics of the workshops

The course does not include workshops

Topics of the laboratory classes

Topic 1. Familiarity with the main services of Apache Hadoop. Software installation. INTELLIJ IDEA development environment. Programming operations with resilient distributed datasets (RDD).
Topic 2. Aggregation in Spark. RDD. Creating and working with DataFrames using Spark.
Topic 3. Working with Spark GraphX and Spark MLlib. Work with graphs using Spark GraphX. Implementation of machine learning algorithms using Spark MLlib and the spark.ml package.
Topic 4. Implementation of machine learning algorithms. Statistical calculations and hypothesis testing. Hadoop installation.
Topic 5. Working with Hadoop Distributed File System (HDFS) and using YARN. The solution to the problem of writing data in HDFS. Features of resource management in Hadoop cluster mode with YARN. Hadoop and Spark testing.
Topic 6. Scala programming language. Familiarity with the syntax and basic capabilities of the Scala language. Introduction to data analysis using Scala and Spark.
Topic 7. MapReduce model of distributed computing. Familiarization with construction features and capabilities of MapReduce.
Topic 8. Solving problems of MapReduce in Spark. Solution problems of MapReduce in Hadoop.

Self-study

The course involves the implementation of individual calculation tasks on each topic. The results of calculations are drawn up in a written report. Students are also recommended additional materials (videos, articles) for independent study and analysis.

Non-formal education

Within the framework of non-formal education according to the relevant Regulation (http://surl.li/pxssv), the educational component or its separate topics can be taken into account in case of independent completion of professional courses/trainings, obtaining civic education, online education, professional internship, etc.
In particular, individual topics of this component may be taken into account upon successful completion of such courses:
– Topic 1. An introduction to distributed and parallel computing.
Concepts of cloud computing https://www.coursera.org/learn/cloud-computing?specialization=cloud-computing
Scala2 parallel programming https://www.coursera.org/learn/scala2-parallel-programming
– Topic 2. Introduction to distributed computing. Model, specializations of distributed computing.
Concepts of cloud computing https://www.coursera.org/learn/cloud-computing?specialization=cloud-computing
– Topic 3. Understanding big data.
Introduction to big data with Spark Hadoop https://www.coursera.org/learn/introduction-to-big-data-with-spark-hadoop.
– Topic 4. Big data storage and processing concepts.
Concepts of cloud computing https://www.coursera.org/learn/cloud-computing?specialization=cloud-computing
– Topic 5. Distributed batch and stream processing of unstructured and loosely structured data.
Introduction to big data with Spark Hadoop https://www.coursera.org/learn/introduction-to-big-data-with-spark-hadoop
– Topic 6. Development of data processing applications running in a distributed computing environment.
Introduction to big data with Spark Hadoop https://www.coursera.org/learn/introduction-to-big-data-with-spark-hadoop

– Topic 7. Cluster resource management and task planning.
Hadoop Platform and Application Framework https://www.coursera.org/learn/hadoop
– Topic 8. A programming model for parallel and distributed processing of large data sets.
Hadoop Platform and Application Framework https://www.coursera.org/learn/hadoop

Course materials and recommended reading

1. Hirvonen J. Distributed Algorithms / J. Hirvonen, J. Suomela. – Aalto: Aalto University, 2023. – 221 p. https://jukkasuomela.fi/da2020/da2020.pdf
2. Минайленко Р.М. Паралельні та розподілені обчислення: навч. посіб. – Кропивницький: Видавець Лисенко В. Ф., 2021. – 153 с. https://dspace.kntu.kr.ua/server/api/core/bitstreams/396e02d2-725b-47b5-a1c0-ae07a9bec326/content
3. Коцовський В. М. К75 Теорія паралельних обчислень: навчальний посібник. Ужгород: ПП «АУТДОР-Шарк», 2021. 188 с. https://dspace.uzhnu.edu.ua/jspui/bitstream/lib/38994/1/%D0%9D%D0%B0%D0%B2%D1%87%D0%B0%D0%BB%D1%8C%D0%BD%D0%B8%D0%B9%20%D0%BF%D0%BE%D1%81%D1%96%D0%B1%D0%BD%D0%B8%D0%BA.pdf
4. Кузьма К .Т. К89 Паралельні та розподілені обчислення: навчальний посібник для вищих закладів освіти / К. Т. Кузьма, О. В. Мельник. – Миколаїв: ФОП Швець В. М., 2020. – 172 с. http://dspace.mdu.edu.ua/jspui/bitstream/123456789/860/1/%D0%9A%D1%83%D0%B7%D1%8C%D0%BC%D0%B0%2C%20%D0%9C%D0%B5%D0%BB%D1%8C%D0%BD%D0%B8%D0%BA_%D0%9F%D0%B0%D1%80%D0%B0%D0%BB%D0%B5%D0%BB%D1%8C%D0%BD%D1%96%20%D1%82%D0%B0%20%D1%80%D0%BE%D0%B7%D0%BF%D0%BE%D0%B4%D1%96%D0%BB%D0%B5%D0%BD%D1%96%20%D0%BE%D0%B1%D1%87%D0%B8%D1%81%D0%BB%D0%B5%D0%BD%D0%BD%D1%8F.pdf
5. Наконечна О. А., Ярмоленко Т. А., Алексеєнко В. В., Якимчук Б. М. Інструктивно-методичні рекомендації з дисципліни «Технології розподілених систем та паралельних обчислень» / уклад.: Оксана Наконечна, Тетяна Ярмоленко, Вікторія Алексеєнко, Богданна Якимчук. Житомир: Житомир: Вид-во ЖДУ ім. Івана Франка, 2023. 74 с. http://eprints.zu.edu.ua/35948/1/%D1%96%D0%BD%D1%81%D1%82-%D0%BC%D0%B5%D1%82%D0%BE%D0%B4_%D1%82%D0%B5%D1%85%D0%BD%D0%BE%D0%BB%D0%BE%D0%B3%D1%96%D1%97%20%D1%80%D0%BE%D0%B7%D0%BF%D0%BE%D0%B4%20%D1%81%D0%B8%D1%81%D1%82%D0%B5%D0%BC%20%284%29.pdf
6. Вербіцький В. В. Паралельне програмування з використанням технології OpenMP: метод. вказівки / В. В. Вербіцький, А. Л. Максимов. – Одеса : Одес. нац. ун-т ім. І. І. Мечникова, 2022 – 48 с. http://dspace.onu.edu.ua:8080/handle/123456789/34441

Assessment and grading

Criteria for assessment of student performance, and the final score structure

Description of the final score structure, course requirements, and necessary steps to earn points, especially paying attention to self-study and individual assignments.

Grading scale

Total points	National	ECTS
90–100	Excellent	A
82–89	Good	B
75–81	Good	C
64–74	Satisfactory	D
60–63	Satisfactory	E
35–59	Unsatisfactory (requires additional learning)	FX
1–34	Unsatisfactory (requires repetition of the course)	F

Norms of academic integrity and course policy

The student must adhere to the Code of Ethics of Academic Relations and Integrity of NTU «KhPI»: to demonstrate discipline, good manners, kindness, honesty, and responsibility. Conflict situations should be openly discussed in academic groups with a lecturer, and if it is impossible to resolve the conflict, they should be brought to the attention of the Institute's management.
Regulatory and legal documents related to the implementation of the principles of academic integrity at NTU «KhPI» are available on the website: http://blogs.kpi.kharkov.ua/v2/nv/akademichna-dobrochesnist/

Approval

Approved by	Date, signature 29.08.2024		Head of the Department Olena AKHIIEZER
	Date, signature 29.08.2024		Guarantor of the Educational Program Olena AKHIIEZER