

ביה"ס למדעי המחשב

סמסטר א' תשפ"ו



מערכות אחסון לביג-דאטה (קוד: 142240)			שם הקורס בעברית
Storage Systems for Big Data			שם הקורס באנגלית
קורס בחירה 3 נ"ר	תואר ראשון שיעור	תשפ"ו סמסטר א' שלישי 16:15-18:45	
שעת קבלה בתיאום מראש	מייל dalit.naor@mta.ac.il		שם המרצה פרופ' דלית נאור

פרטי הקורס

התפוצצות כמות המידע הדיגיטלי – הן הארגוני והן האישי – הינו אחד מהמאפיינים הבולטים של תקופתנו, ובעקבותיו מתפתחות טכנולוגיות חדשות ומעניינות המאפשרות שמירת נפחים גדולים של נתונים בעלויות נמוכות ונצילות אנרגטית גבוהה. בנוסף, מהפיכת ה AI המבוססת על מאגרים עתירי מידע מחייבת לא רק שמירה כי אם יכולות עיבוד גבוהות למידע. על מנת לענות על צרכים אלה, התרחש בשני העשורים האחרונים שינוי מהותי באופן שבו מערכות מחשוב ואחסון גדולות (large-scale) נבנות, וזאת בעקבות המהפכה בחישוב מבוסס-ענן ומהפיכת המידע שהביא עולם ה Big data וה-AI. בקורס זה נסקור את העקרונות הבסיסיים של מערכות אחסון (storage) כגון, דיסקים, טייפים, SSD, מערכות קבצים, caching, RAID, ושיטת להקטנת נפחי דאטה (data reduction). נלמד על אחסון מבוזר בענן עבור BigData לדוגמה Object Stores ו-Distributed cloud file systems. נבחן מספר מערכות חדשות שניבנו ע"י חברות הענן הגדולות הנקראות hyper-scale companies על מנת לשמור ולנהל כמויות מידע גדולות. נבין גישות שונות לביזור החישוב, מודלים חדשים לעמידות ואמינות המערכות והמידע ועוד. תחום אחסון המידע הדיגיטלי (storage systems) הינו נושא מאד מרכזי וחשוב בעולם המחשוב. בישראל בפרט יש מספר גדול מאד של חברות בתחום (חברות startup ואחרות). מטרת הקורס היא להכשיר את הסטודנטים בכלים הבסיסיים הנדרשים להבנת מערכות אלה, ובאמצעות הקורס תלמדו להבין ולהכיר את הטכנולוגיות עליהן מבוססות חברות רבות וביניהן ישראליות בתחום אחסון המידע. הקורס מיועד לסטודנטים לתואר ראשון, אך פתוח לרישום גם לסטודנטים מתואר שני. עבור סטודנטים מהתואר השני, הקורס יוכר כנקודות פטור במסגרת הפטורים מקורסי תואר ראשון, בכפוף לאישורים הנדרשים.	תקציר בעברית
---	----------------------------

The exponential growth of digital data — often referred to as the “data tsunami” — is a defining characteristic of the modern computing era. This surge, driven by both personal and legacy data, has led to the development of new storage media and architectures capable of persistently and efficiently managing vast datasets while minimizing energy consumption. In parallel, the Big Data and AI revolutions depend critically on the ability to store, access, and analyze massive datasets in large-scale distributed environments, primarily hosted by major Cloud providers (e.g. Amazon, Google, and Microsoft). This course provides a comprehensive introduction to storage systems, from foundational concepts to modern large-scale architectures. We will cover traditional storage technologies (magnetic disks, tapes, and solid-state drives), file systems, caching strategies, RAID architectures, data compression, and deduplication. We will then explore cloud storage solutions, including object stores and distributed file systems, and examine approaches to distributed computing, data resiliency, and fault tolerance as applied to large-scale storage systems. The course will emphasize both theoretical foundations and real-world implementations. Israel is home to many innovative storage companies — from startups to global industry leaders. Students will acquire the skills to understand, analyze, and design storage solutions applicable to a wide range of applications and organizations in the high-tech industry, specifically in Israel. The course is open to both undergraduate and graduate students. M.Sc. students may, with approval, count the course towards their degree requirements.	<u>תקציר באנגלית</u>
---	------------------------------------

• עבור תלמידי תואר ראשון נידרש: לימוד במקביל של מערכות הפעלה ומערכות תקשורת נתונים. מערכות בסיסי נתונים – מומלץ. • עבור תלמידי תואר שני, נדרשת השלמת הקורסים: מערכות הפעלה, מערכות תקשורת נתונים.	דרישות קדם
הקורס יתנהל כשיעור פרונטלי של 3 שעות שבועיות ויילוו בשקפים. ההרצאות תינתנה ע"י המרצה וייתכנו מספר הרצאות אורח. מצגות הקורס יפורסמו בתום ההרצאה.	שיטת ההוראה
במהלך הקורס יינתנו מספר תרגילים, התרגילים הם חובה ואישיים. בסיום הקורס תערך בחינה.	דרישות הקורס/מטלות
• תרגילים: 50% • מבחן: 40% • השתתפות בשיעורים ותרומה: 10%	הרכב הציון
נוכחות חובה ב 9 מתוך 13 ההרצאות. חובת נוכחות בכל הרצאות האורח. מגויסים שלא יוכלו לעמוד בדרישת הנוכחות – יש לפנות למרצה.	נוכחות

הנוכחות גם ביתר ההרצאות והשתתפות חשובה ומומלצת מכיוון שאין מקור מסודר שמנגיש את החומר ושעליו ניתן להתבסס למטרות לימוד עצמי.

תכנית הקורס - נושאי הלימוד, בכפוף לשינויים ולהתקדמות הקורס

1. Data persistence and storage: review of the basic notions of I/O systems, file systems, caching.
2. The basics of traditional storage media: hard drive (disk), tape, RAID
3. Flash-based SSDs
4. Caching strategies
5. Storage for BigData and Cloud Based architectures
 - a. Cloud file systems - DFS, GFS, HDFS
 - b. The Cloud paradigm and Cloud storage: Block/File/Objects
6. Hashing and Bloom Filters for Cloud Caching
7. Distributed storage architectures
 - a. Dynamic Hash Tables (DHT), Consistent Hashing
8. Data Reduction in storage systems: data deduplication and data compression.
9. Other topics
 - a. Data Integrity: Merkle Trees
 - b. Log-Structured Merge-trees (LSM Trees)
 - c. Systems Sustainability
10. New storage media: DNA Storage, Glass Storage

רשימת קריאה ומקורות נוספים חלקיים (רשות) :

Books

1. *Operating Systems: Three Easy Pieces*. By Remzi H. Arpaci-Dusseau and Andrea C. Arpaci-Dusseau
2. *Guide to Reliable Distributed Systems: Building High-Assurance Applications and Cloud-Hosted Services*. Kenneth P Birman

3. *Designing Data-Intensive Applications*. Martin Kleppmann
4. *Information Storage and Management: Storing, Managing, and Protecting Digital Information in Classic, Virtualized, and Cloud Environments*, 2nd Edition. EMC Education Services (Editor)
5. *Storage Systems: Organization, Performance, Coding, Reliability, and Their Data Processing*, Alexander Thomasian.