בקרה של מערכות תורים רבות שרתים בעומס גבוה

גנאדי שייחט

בקרה של מערכות תורים רבות שרתים בעומס גבוה

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת תואר

דוקטור לפילוסופיה

גנאדי שייחט

הוגש לסנט הטכניון - מכון טכנולוגי לישראל

ניסן תשס"ז חיפה אפריל 2007

המחקר נעשה בהנחייתו של פר' רמי אתר ופר' אבישי מנדלבאום בפקולתה להנדסת תעשיה וניהול.

כל תקופת ההשתלמות היתה חוויה נהדרת עבורי ועל כך אני מודה לרמי ולאבישי.

כמו כן ברצוני להודות לכל האנשים הנהדרים בפקולתה להנדסת תעשיה וניהול. היה לי מזל רב לפגוש את כולכם!

כתמיד, עם כל האהבה למשפחה שלי.

אני מודה לבית הספר ללימודי מוסמכים, רמי אתר ואבישי מנדלבאום על התמיכה הכספית הנדיבה בהשתלמותי.

תקציר

בעבודה זו אנחנו חוקרים מודלים תוריים - מערכות שירות המורכבות ממספר סוגי לקוחות ומספר תחנות שירות שונות. בכל תחנה יש מספר גדול של שרתים זהים סטטיסטית. סוגי הלקוחות נבדלים אלה מאלה בדרישות השירות שלהם. סוגי התחנות מאופיינים על ידי סוגי הלקוחות שהם יכולים לשרת ועל ידי קצב השירות שהם יכולים להעניק. לפי הנחת המודל, התפלגות משך השירות היא אקספוננציאלית ואילו הגעות הלקוחות נעשות לפי תהליכי חידוש.

אחת הסוגיות החשובות בתפעול המערכות הנ"ל היא **בקרה דינאמית**: כיצד לנתב את הלקוחות במערכת בזמן אמת, דהיינו איך לצוות לקוחות לשרתים כאשר מתפנה שרת או כאשר מופיע לקוח חדש במערכת. המסגרת העיקרית שלנו לניתוח היא משטר העומס הגבוה (*heavy traffic*) שהוצג על ידי Halfin and Whitt ב-1981. משטר זה מאופיין על ידי כך שמספר השרתים בכל תחנה וקצבי ההגעה של הלקוחות שואפים לאינסוף תוך כדי שמירה על רמה קריטית של עומס. נצילות השרתים מתכנסת ל-100%, כלומר המערכת פועלת בעומס גבוה. אחרי לקיחת גבולות דיפוזיה לתהליכים המייצגים את מספר הלקוחות של כל סוג במערכת מתקבל *מודל דיפוזיה* עם תנודות סביב *מודל נוזלים סטטי*, המתקבל על ידי לקיחת גבולות נוזלים. מודלים שכאלו קיבלו תשומת לב רבה בזמן האחרון, במיוחד בהקשר של מוקדים טלפוניים: מוקדים רבים מטפלים במספר רב של סוגי שיחות (הנבדלות בסוג השירות הנדרש, שפת המתקשר וכו') ומעסיקים מספר גדול של מוקדנים מסוגים שונים.

בעבודה זו גילינו התנהגות חדשה ובלתי רגילה של מערכת תורים במשטר עומס גבוה: תחת תנאים מסוימים קיימת מדיניות ניתוב דינמית, המאפשרת למערכת לתפקד כאילו היא בעומס נמוך.

בפרק 2 הראינו כי התופעה קשורה להצגה של מודל דיפוזיה של המערכת הנידונה כדיפוזיה מבוקרת עם מרכיב בקרה סינגולרי. המרכיב הסינגולרי יכול להגביל את הדיפוזיה לאזורים מסוימים של המרחב. נקרא לדיפוזיה *מבוקרת אפס* (*null controllable*) אם היא יכולה להיות מוגבלת לאזור המבטיח תורים ריקים במערכת השירות המקורית. אנחנו מציגים תנאים מספיקים לבקרת אפס במונחים של הגרף המקודד את מבנה מערכת התורים. תחת תנאים אלה, אנו מראים כי תופעה אסימפטוטית דומה מתקיימת גם עבור המודל המקורי: נבנתה מדיניות דינמית כזו שעבור כל $0 < \varepsilon < T < \infty$, כל התורים במערכת נשמרים ריקים באינטרוול $[\varepsilon, T]$ בהסתברות המתכנסת ל-1.

בפרק 3 אנו מציגים וחוקרים את המושג של תת - אופטימליות של תפוקת מודל הנוזלים (*throughput sub-optimality*). נניח שמודל הנוזלים נמצא בעומס קריטי. בפרט מתקיימים התנאים הבאים: (1) ניתן לסדר את השרתים כך שעבור כל סוג i של לקוחות, קצב ההגעה λ_i שווה לקצב השירות הכולל שלהם, (2) תכונה (1) לא מתקיימת אם לפחות לסוג לקוחות אחד קצב ההגעה גדל ל- $\lambda_i < \hat{\lambda}_i$. התכונה הבאה יכולה להתקיים במודלים כאלה: ניתן לסדר את השרתים כך שקצב העזיבה הכולל של הלקוחות (שכבר שורתו) יהיה גדול יותר מקצב ההגעה הכולל שלהם. מודל נוזלים עם תכונה כזאת נקרא תת-אופטימלי ביחס לתפוקה. בפרק זה הראינו כי עבור מערכות תורים עם מודלים נוזלים כאלה קיימת מדיניות ניתוב המאפשרת את הדבר הבא: לכל T, משך הזמן הכולל לפני T, שבו יש במערכת תורים, מתכנס ל-0 בהסתברות. כמו כן, הראינו כי תת- אופטימליות ביחס לתפוקה מקבילה לתנאי מסוים על הגרף המקודד את מבנה המערכת. תנאי זה הוא חלש יותר מאשר התנאים המוצגים בפרק 2, ולכן יש הבדלים בין תוצאות שני הפרקים 2 ו-3. ראשית, פרק 3 מאפשר ללקוחות להיות בתור, אך לזמנים קצרים. ההבדל הבא הוא שבפרק 2 מוצגות מדיניות משני הסוגים: *Preemptive (P)* , מדיניות לפיה שירות של לקוח יכול להיות מופסק ויחודש אחרי זמן כלשהו, לא בהכרח אצל אותו שרת ולא בהכרח באותה תחנת שירות, ו- *Non-Preemptive (NP)* , לפיה לא ניתן להפסיק שירות של לקוח. פרק 3 מטפל רק במדיניות P. אנחנו מאמינים (אך לא מוכיחים) כי התוצאות של פרק 2 אינן אפשריות תחת התנאים המוצגים בפרק 3. הכלי שבו אנו משתמשים בפרק 3 הוא מודל נוזלים דינמי דטרמיניסטי שבו, באופן גס, התנודות הסטוכסטיות הוחלפו לדטרמיניסטיות.

פרק 4 מוקדש למודל דיפוזיה מבוקרת. ניסוח בעיית בקרה למודל התורים המקורי מוביל לבעיית בקרה סטוכסטית. לעיתים קרובות, פתרון של בעיית דיפוזיה משמש כבסיס לבניית מדיניות הבקרה האופטימלית למודל המקורי. אנו מסתכלים על מודלים תוריים עם נטישות וקצבי שירות שהם תלויי-תחנה. מסתבר כי במקרים האלה מודל הדיפוזיה עובר כמה רדוקציות המובילות למודל חד-מימדי. כתוצאה מכך בעיית הבקרה הסטוכסטית נעשית חד-ממידית, דבר שמפשט את חיפוש הפתרון וכמו כן מאפשר לקבל פתרון מדויק במקרים מסוימים. זה איננו המקרה הראשון שבו מתגלית תופעת הרדוקציה הנ״ל, כאשר עבודות רלוונטיות טיפלו או במודלים עם נטישות בעלות 2 סוגי לקוחות ושתי תחנות שירות, או במודלים בלי נטישות. כאן אנחנו מבצעים הרחבה של התוצאות הידועות.

על סמך פתרון בעיית הדיפוזיה אנחנו מציעים מדיניות ניתוב למודל המקורי המשוערות (אך לא מוכחות) להיות אסימפטוטית אופטימליות. נדגים זאת על שני מקרים פשוטים:

- **מקרה 1**: נניח כי נתונים מחירי המתנה $c_1 \geq c_2 \geq ... \geq c_I > 0$ ונניח, באופן גס, כי אנחנו מעוניינים למזער את הקומבינציה הליניארית של אורכי התורים (תחת סקלה מתאימה) $c_1 Y_1 + ... + c_I Y_I$, כאשר קצבי הנטישות מקיימים $\theta_1 \leq \theta_2 \leq ... \leq \theta_I$. אזי, מדיניות NP אסימפטוטית אופטימלית מכתיבה את שיטת הניתוב הבאה: (1) כאשר לקוח מגיע למערכת ויש שרתים פנויים היכולים לשרת אותו – אזי הלקוח מכוון לשרת המהיר מבין הפנויים, אחרת הוא מצטרף לתור; (2) כאשר מתפנה שרת היכול לשרת סוג I ויש לקוחות מסוג I המחכים בתור- אזי לקוח מסוג I יתקבל לשירות אם ורק אם אין לקוחות מסוגים אחרים היכולים לקבל שירות אצל אותו שרת. כל המקרים האחרים נפתרים שרירותית. במלים אחרות, המערכת נוטה "לשמור" את כל התור רק בסוג I ובאותו אופן נוטה להיות במצב בו שרתים פנויים, אם ישנם כאלה, צריכים להיות רק בתחנה עם קצב השירות הנמוך ביותר.

- **מקרה 2**: נניח כי $\theta_1 = \theta_2 = ... = \theta_I$ ונניח כי אנו מעוניינים למזער את $C_1(Y_1) + ... + C_I(Y_I)$ כאשר $C_1(\cdot),...,C_I(\cdot)$ פונקציות קמורות עולות ממש. אזי, מדיניות NP אסימפטוטית אופטימלית היא כדלהלן: (1) אותה מדיניות ניתוב עבור לקוח חדש כמו במקרה 1; (2) כאשר שרת מתחנה j מתפנה בזמן t כלשהו ויש כמה סוגי לקוחות הממתינים לשירות, הוא ייקח לקוח מסוג k המקיים $k = \arg\max\{i \sim j : C_i'(Y_i(t))\}$. כאן $i \sim j$ מסמן כי לקוחות מסוג i יכולים לקבל שירות בתחנה j.

חשוב לציין שמדיניות P אופטימלית, גם במקרה 1 וגם במקרה 2, שקולה אסימפטוטית למדיניות NP

בסיום העבודה, אנו מתארים כיווני מחקר אפשריים הנובעים ממנה. נתאר אותם בקצרה. הופעת המרכיב הסינגולרי בניסוח מודל הדיפוזיה בפרק 2 מהווה חידוש בפני עצמו ופרט למקרה המתואר בפרק 2 (שימוש במרכיב סינגולרי להגבלת הדיפוזיה לאזור מסוים) לא נעשה שום טיפול במקרים כלליים. כיוון נוסף קשור למודל הדיפוזיה מפרק 4 וניתוח מקרים מעניינים בהם הפתרון לא ידוע. כיוון מעניין נוסף הוא לנסות להכליל את תוצאות הפרקים 2 ו-3 למודלים עם התפלגות משך זמן שירות לא אקספוננציאלית.