תורים עם עדיפויות של לקוחות בעלי סבלנות סופית:

ניתוח מערכות עם קצב הגעות המשתנה בזמן והשוואה למצב יציב

חיבור על מחקר

לשם מילוי חלקי של הדרישות לקבלת התואר

מגיסטר למדעים בחקר ביצועים וניתוח מערכת

ליובוב רוזנשמידט

הוגש לסנט הטכניון - מכון טכנולוגי לישראל

אייר תשס''ח, חיפה מאי 2008

המחקר נעשה בהנחייתו של פרופסור אבישי מנדלבאום בפקולטה להנדסת תעשייה וניהול. ברצוני להודות לו על הנחייתו ותמיכתו בכל שלב של עבודתי.

אני מודה להוריי כי בלעדיהם לא הייתי משיגה את מה שהשגתי. האמון שלהם בי ואהבתם נתנו לי כוחות להמשיך ועל זה אני אסירת תודה.

ברצוני להודות לדוד שלי, פרופסור יהודה חריט, ולאשתו לנה על כך שהציגו בפני את פירוש המילה "משפחה" וגרמו לי להרגיש בבית בימים הכי קשים שהיו לי.

תודות לחבר שלי, אריה, על נכונותו להתמודד עם השעות הארוכות של הבודה ועל יכולתו למצוא את המילים הנכונות כדי להרגיע ולשמח אותי.

אני מודה לדוקטור סרגיי זלטין על הדיונים הרבים שעזרו לי במחקרי.

תודה מיוחדת מגיעה לשמרית ממן שעזרה לי כל כך הרבה שקשה לי לדמיין איך הייתי מסיימת את התואר בלעדיה.

ולבסוף, אני מודה לטכניון על התמיכה הכספית הנדיבה בהשתלמותי.

תקציר מורחב

מערכות שירות מודרניות מורכבות לעתים קרובות ממספר סוגי לקוחות הנבדלים בסוג השירות הנדרש עבורם. אחת הדוגמאות החשובות למערכת שירות היא מוקדי שירות טלפוניים. במוקד טלפוני מוקדנים רבים מטפלים במספר סוגי שיחות הנבדלים זה מזה בסוג השירות הנדרש, שפת המתקשר וכדומה. מתן עדיפויות שונות לסוגים שונים של לקוחות בהתאם לחשיבותם לארגון הנו מקובל מאוד בתעשייה ומאפשר השגת מטרות תפעוליות של ארגון. עבודה זו מוקדשת למודלים מתמטיים וסימולציות התומכים בהנדסת שירות טלפוני.

העבודה מאורגנת באופן הבא. פרק 3 מציג טכניקה אנליטית שבעזרתה נוכל להציג את הסתברות ההמתנה של תור מרקובי כפונקציה של תקופות התעסוקה והבטלה שלה.

פרק 4 פותח את הדיון על תורים עם עדיפויות ומציג תוצאות ידועות עבור עדיפויות מסוג Preemptive ו- Non-Preemptive. לאחר מכן, בפרק 5, אנו מרחיבים את התוצאות לתורים עם נטישות ומבצעים ניתוח אסימפטוטי של תורים עם שני סוגים של לקוחות.

פרקים 6 ו- 7 מוקדשים לתורים עם קצבי הגעות משתנים. בפרק 6, אנו מבצעים סימולציות של 4 מוקדים שונים עם קצב הגעות, זמן שירות וסבלנות אמפיריים. אנו קובעים את רמת האיוש לפי כלל השורש ומשווים את התוצאות עם שתי שיטות איוש המקובלות כיום, שהן *PSA* ו-*Lagged PSA*.

פרק 7 מציג את תוצאות ניתוח של מוקד שירות עם שני סוגי לקוחות. פרק זה הנו הכללה של התוצאות ב- [13].

בפרק 8 אנו משווים תור *M/M/N+M* עם *M/G/N+M* תוך הדגשה של ההבדלים הנגרמים כתוצאה מהבדלים בהתפלגות זמן שירות ומכלילים את התוצאות של שוורץ [32] לתורים עם נטישות.

חקירת תורים בעזרת "נדידות"

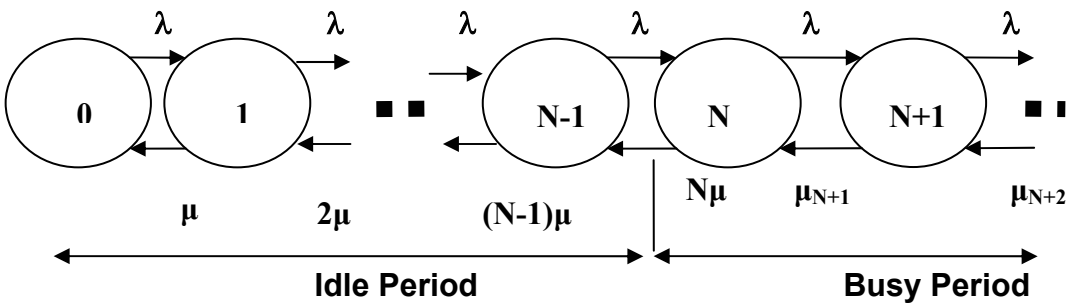
בפרק 3, אנו מציגים תור *M/M/N(+M)* עם קצב הגעה λ, קצב שירות μ בעזרת תקופת תעסוקה *(Busy Period)* ותקופת בטלה *(Idle Period)*.

באופן פורמאלי, נגדיר:

$L_+ = \{L_+(t), t \geq 0\}$ - מספר הלקוחות במערכת כאשר כל השרתים תפוסים.

$L_- = \{L_-(t), t \geq 0\}$ - מספר הלקוחות במערכת כאשר לפחות שרת אחד פנוי.

הגדרה מדויקת ניתנת בעמוד 18.



נסמן

$T_{N,N-1}$ - תוחלת משך תקופת התעסוקה.

$T_{N-1,N}$ - תוחלת משך תקופת הבטלה.

לפי PASTA, ניתן להציג את הסתברות ההמתנה כפונקציה של תקופות תעסוקה ובטלה:

$$P\left(W_q > 0\right) = \frac{T_{N,N-1}}{T_{N,N-1} + T_{N-1,N}} = \left[1 + \frac{T_{N-1,N}}{T_{N,N-1}}\right]^{-1} \qquad (1)$$

נגדיר עומס מוצע (Offered Load) באופן הבא:

$$R = \lambda / \mu ,$$

כאשר λ הנו קצב הגעות ו- μ הנו קצב שירות.

אנו מתרכזים בשלושה תחומי הפעילות הבאים:

- **מכוון יעילות (Efficiency Driven, ED)** $N \approx R - \varepsilon R, \quad 0 < \varepsilon < 1$

תחום זה מאופיין ע"י נצילות גבוהה מאוד של השרתים, זמני המתנה ארוכים והסתברות ההמתנה קרובה מאוד ל-1.

- **מכוון איכות (Quality Driven, QD)** $N \approx R + \varepsilon R, \quad \varepsilon > 0$

בתחום פעילות זה, כמעט אף אחד לא ממתין, זמני המתנה זניחים ונצילות השרתים נמוכה.

- **מכוון איכות ויעילות (Quality and Efficiency Driven, QED)** $N \approx R + \beta\sqrt{R}$

תחום זה משלב בהצלחה את התכונות החיוביות של שני התחומים הקודמים ומצליח להימנע מהחסרונות שלהם. נצילות השרתים כאן גבוהה (בסביבות 95%), זמני ההמתנה קצרים והסתברות ההמתנה הנה קבועה $0 < \alpha < 1$.

הצגה (1) מאפשרת ניתוח של ההסתברות להמתין בתחומי פעילות שונים. ניתן להסיק מסקנות לא רק על התוצאה הסופית שהיא ידועה היטב (הסתברות זו מתכנסת ל- 1 בתחום מכוון יעילות, ל-0 בתחום

מכוון איכות ולקבוע $0<\alpha<1$ בתחום המכוון איכות ויעילות). בנוסף, בעזרת ההצגה של ההסתברות להמתין לעיל, נוכל לחשב ולהבין גם את אופי ההתכנסות בתחומי הפעילות השונים.

טבלא מס׳ 1: קצבי התכנסות וגבולות של תקופות תעסוקה/בטלה בתחומי פעילות שונים

	תקופת תעסוקה		**תקופת בטלה**	
	קצב התכנסות	גבול	קצב התכנסות	גבול
QED $N \approx R+\beta\sqrt{R}$	$1/\sqrt{N}$	**0**	$1/\sqrt{N}$	**0**
QD $N \approx R+\varepsilon R$	$1/N$	**0**	$\left[\sqrt{N}\,\phi\left(\varepsilon\sqrt{\lambda/\theta}\right)\right]^{-1}$	∞
ED $N \approx R-\varepsilon R$	$\left[\sqrt{N}\,\phi\left(\varepsilon\sqrt{\lambda/\theta}\right)\right]^{-1}$	∞	$1/N$	**0**

טבלא מס׳ 1 מציגה את קצבי ההתכנסות והגבולות של תקופות הבטלה והתעסוקה תחת שלושה תחומי הפעילות השונים. באמצעות טבלא זו ו- (1), ניתן להבין את הגבולות הידועים של הסתברויות ההמתנה. כך למשל, עבור QED הסתברות ההמתנה מתכנסת לקבוע שהוא ממש גדול מ-0 וממש קטן מ-1. מכיוון שתקופות תעסוקה ובטלה מתכנסות ל-0 באותו קצב, תוצאה זו הנה הגיונית.

תורים עם עדיפויות

אנו מתייחסים לשתי דרכים של מתן עדיפויות ללקוחות: עדיפות מסוג Preemptive ועדיפות מסוג Non-Preemptive. ההנחה כאן שקצב שירות μ הנו זהה לכל סוגי הלקוחות. כמו כן, קצב הנטישות θ זהה לכל הסוגים.

עדיפות מסוג Preemptive: כאשר לקוחות מעדיפויות גבוהות אינם מודעים לקיומם של לקוחות בעדיפויות נמוכות. במדיניות זו לקוחות מעדיפויות נמוכות לא ייכנסו לשירות כל עוד ישנם לקוחות מעדיפויות גבוהות שממתינים בתור. בנוסף, כאשר לקוח מעדיפות גבוהה מגיע ומוצא את כל השרתים תפוסים וחלק מהם עסוק בשירות לקוחות מעדיפויות נמוכות, אזי אחד הלקוחות בעדיפות נמוכה מייד נשלח לראש התור, והלקוח מהעדיפות הגבוהה נכנס לשירות.

תוחלת זמן ההמתנה של לקוח מסוג k נתונה ע״י ביטוי הבא:

$$E_{pr}\left(W_q^k\right)=\left[\lambda_{1\to k}E_{pr}\left(W_q^{(1\to k)}\right)-\lambda_{1\to(k-1)}E_{pr}\left(W_q^{(1\to k-1)}\right)\right]\lambda_k^{-1} \qquad (2)$$

כאשר

λ_k הנו קצב הגעה של לקוחות מסוג k;

$E_{pr}\left(W_q^k\right)$ הנה תוחלת זמן ההמתנה של לקוח מסוג k כאשר מדיניות מתן העדיפות הנה מסוג Preemptive;

$E_{pr}\left(W_q^{(1\to k)}\right)$ הנה תוחלת זמן ההמתנה של לקוחות מעדיפות גבוהה המגיעים לפי קצב $\lambda_{1\to k}=\sum_{j=1}^{k}\lambda_j$ כאשר מדיניות מתן העדיפות הנה מסוג Preemptive.

ניתן להראות כי את התוחלת ניתן להציג בעזרת ביטוים של מערכות M/M/N(+M) **ללא עדיפויות**:

$$E_{pr}\left(W_q^k\right)=\left[\lambda_{1\to k}E\left(W_q^{(1\to k)}\right)-\lambda_{1\to(k-1)}E\left(W_q^{(1\to k-1)}\right)\right]\lambda_k^{-1} \qquad (3)$$

כאשר

$E\left(W_q^{(1\to k)}\right)$ הנה תוחלת זמן ההמתנה במערכת M/M/N(+M) עם קצב הגעות $\lambda_{1\to k}=\sum_{j=1}^{k}\lambda_j$.

עבור מערכות ללא נטישות, את הביטוי המדויק של תוחלת זמן ההמתנה ללקוח מסוג k ניתן לרשום בצורה הבאה:

$$E_{pr}\left(W_q^k\right)=\frac{E_{2,N}\left(\lambda_{1\to k}\right)}{\lambda_k\left(1-\sigma_k\right)}-\frac{E_{2,N}\left(\lambda_{1\to(k-1)}\right)}{\lambda_k\left(1-\sigma_{(k-1)}\right)} \qquad (4)$$

כאשר

$E_{2,N}\left(\lambda\right)$ הנה ההסתברות להמתנה ב- *M/M/N* עם קצב הגעות λ (נוסחת Erlang-C);

$\sigma_k=\sum_{j=1}^{k}\rho_j$ הנה פרופורציית הזמן ששרת עסוק עם לקוחות מסוגים $1,\ldots,k$;

$\rho_k=\frac{\lambda_k}{N\mu}$ הנה פרופורציית הזמן ששרת עסוק עם לקוח מסוג k.

עדיפות מסוג Non-Preemptive: תחת מדיניות מתן עדיפויות זו, לקוחות מעדיפויות נמוכות לא יכנסו לשירות כל עוד ישנם לקוחות מעדיפויות גבוהות שממתינים. אך לאחר שהשירות התחיל, לא ניתן להפסיקו.

מבנה תוחלת זמן ההמתנה כאשר עדיפות הנה מסוג Non-Preemptive: ניתן לראות כי לתוחלת זמן ההמתנה של לקוח מסוג k, כאשר העדיפות הנה מסוג Non-Preemptive, ישנו מבנה משותף עבור תורים עם ובלי נטישות. תוחלת זו מורכבת משלושת המרכיבים הבאים:

1. הסתברות ההמתנה שווה לכל סוגי לקוחות ושווה להסתברות ההמתנה ב- *M/M/N(+M)* עם קצב הגעות λ (נוסחת Erlang-C או Erlang-A)
2. תוחלת משך תקופת התעסוקה של המערכת M/M/1(+M) עם קצב הגעות $\lambda_{1\to(k-1)}$, קצב שירות $N\,\mu$ וקצב נטישות θ (אם יש).
3. תוחלת מספר הלקוחות במערכת בהינתן המתנה במערכת M/M/1(+M) עם קצב הגעות $\lambda_{1\to k}$, קצב שירות $N\,\mu$ וקצב נטישות θ.

חישוב של תוחלת זמן ההמתנה בתורים M/M/N+M כאשר העדיפות הנה מסוג Non-Preemptive: אנו מראים כי עבור עדיפות מסוג Non-Preemptive, ניתן לחשב את תוחלת זמן ההמתנה של לקוח מעדיפות גבוהה באופן הבא:

$$E_{np}\left(W_q^{1}\right)=P_{\lambda}\left(W_q>0\right)\bullet E_{\lambda_1}\left(W_q\mid W_q>0\right) \qquad (5)$$

כאשר

$E_{np}\left(W_q^{1}\right)$ הנה תוחלת זמן ההמתנה של הלקוחות מסוג ראשון כאשר מדיניות מתן העדיפות הנה מסוג Non-Preemptive;

$P_{\lambda}\left(W_q>0\right)$ הנה הסתברות ההמתנה במערכת *M/M/N+M* עם קצב הגעות $\lambda=\sum_{j=1}^{K}\lambda_j$;

$E_{\lambda_1}\left(W_q\mid W_q>0\right)$ הנה תוחלת זמן ההמתנה במערכת *M/M/N+M* עם קצב הגעות λ_1.

בעזרת (5), העובדה שהסתברות ההמתנה שווה לכל הלקוחות וניתנת לחישוב כהסתברות ההמתנה עם קצב הגעות כולל, ואפשרות להתייחס ל-k סוגים ראשונים כאל סוג אחד בעדיפות גבוהה, אנו מראים שאת תוחלת זמן ההמתנה של לקוח מסוג כלשהו ניתן לחשב בעזרת נוסחא הבאה:

$$E_{np}\left(W_q^{k}\right)=\left[\lambda_{1\to k}E_{np}\left(W_q^{(1\to k)}\right)-\lambda_{1\to(k-1)}E_{np}\left(W_q^{(1\to k-1)}\right)\right]\lambda_k^{-1} \qquad (6)$$

כאשר

$E_{np}\left(W_q^{k}\right)$ הנה תוחלת זמן ההמתנה של לקוח מסוג k כאשר מדיניות מתן העדיפות הנה מסוג Non-Preemptive;

$E_{pr}\left(W_q^{(1\to k)}\right)$ הנה תוחלת זמן ההמתנה של לקוחות מעדיפות גבוהה המגיעים לפי קצב $\lambda_{1\to k}=\sum_{j=1}^{k}\lambda_j$ כאשר מדיניות מתן העדיפות הנה מסוג Non-Preemptive.

שקילות אסימפטוטית של המשטרים Preemptive ו- Non-Preemptive עבור לקוחות מעדיפות נמוכה: אשלגי [2] הראה עבור תורים בלי נטישות שכאשר מספר השרתים גדל לאינסוף, תוחלות זמן ההמתנה של לקוחות בעדיפות נמוכה תחת עדיפות Preemptive ו- Non-Preemptive הנן שקולות אסימפטוטית. אנו מסבירים את הפיזיקה של התופעה עבור כל תחומי הפעילות ומראים שככל שמספר השרתים גדל, ההסתברות של ה-Preemption שואפת ל- 0. בנוסף, אנו מבצעים ניתוח אסימפטוטי עבור תחום מכוון יעילות (ED) ותחום מכוון איכות ויעילות (QED) של תורים עם נטישות ועם שני סוגים של לקוחות.

טבלא מס׳ 2 מציגה את קצבי התכנסות של תוחלות זמן ההמתנה תחת שלושת תחומי הפעילות. ניתן לראות ממנה כי העדיפות הגבוהה, גם במדיניות Preemptive וגם ב- Non-Preemptive, אינה רגישה לשינוי תחום הפעילות. בנוסף לכך, הסוג הראשון של לקוחות בשני המקרים אינו רגיש מבחינת קצב התכנסות להתווספות של נטישות.

בהתאם לנאמר קודם, ניתן לראות כי עבור העדיפות הנמוכה, קצבי התכנסות של תוחלת זמן ההמתנה במדיניות Preemptive וב- Non-Preemptive תחת אותו תחום פעילות שווים.

טבלא מס׳ 2: קצבי התכנסות של תוחלת זמן המתנה בתחומי פעילות שונים

	QED $N = R + \beta\sqrt{R}$		ED $N = R - \gamma R$	
	M/M/N	*M/M/N+M*	*M/M/N*	*M/M/N+M*
$E_{np}\left(W_q^1\right)$	$\Theta\left(1/N\right)$	$\Theta\left(1/N\right)$	$\Theta\left(1/N\right)$	$\Theta\left(1/N\right)$
$E_{np}\left(W_q^2\right)$	$\Theta\left(1/\sqrt{N}\right)$	$\Theta\left(1/\sqrt{N}\right)$	$\Theta(1)$	$\Theta(1)$
$E_{pr}\left(W_q^1\right)$	$\Theta\left(\frac{\rho_1^N}{N\sqrt{N}}\right)$	$\Theta\left(\frac{\rho_1^N}{N\sqrt{N}}\right)$	$\Theta\left(\frac{\rho_1^N}{N\sqrt{N}}\right)$	$\Theta\left(\frac{\rho_1^N}{N\sqrt{N}}\right)$
$E_{pr}\left(W_q^2\right)$	$\Theta\left(1/\sqrt{N}\right)$	$\Theta\left(1/\sqrt{N}\right)$	$\Theta(1)$	$\Theta(1)$

Non-Preemptive - זמן המתנה של לקוחות בעדיפויות גבוהות בתורים M/M/N+M: בנוסף, אנו מראים כי תחת מדיניות מתן עדיפויות מסוג Non-Preemptive, בתחומי פעילות ED, QD ו-QED ובהנחה שהעדיפות הנמוכה אינה זניחה, תוחלת זמן ההמתנה בהינתן המתנה של הלקוחות מסוגים $k = 1,\ldots,K-1$ בתורים עם נטישות מתכנסת לזו שבתורים ללא נטישות.

מערכות תורים עם קצב הגעות לא קבוע בזמן: התפלגות הגעות אמפירית

בפרק 6 חקרנו 4 מוקדי שרות טלפוניים בגדלים שונים: ממוקד מאוד קטן עם עומס מוצע מאוד נמוך (לא יותר מ-5 שעות עבודה לשעה עבור הדוגמא השנייה) ועד מוקדים יחסית גדולים (עד עומס מוצע 250 שעות עבודה לשעה עבור הדוגמא הראשונה והאחרונה). עבור כל אחד מהמוקדים, בעזרת סימולציות, השוונו את תוצאות האיוש לפי ISA או Square-Root Safety Staffing, PSA ו- Lagged PSA. ההבדלים בין מערכות עם איוש הנקבע לפי ISA ו- Square-Root Safety Staffing היו זניחים ולכן עבור כל דוגמא בחרנו שרירותית להציג תוצאות של אחת משתי שיטות אלו.

מטרות הניסויים עם ארבע הדוגמאות הייתה לבדוק האם ניתן להגיע לרמת ביצועים הקבועה בזמן. כמו כן, רצינו לבדוק האם ניתן להשתמש במודל סטאציונרי על מנת לחזות מדדי ביצוע בסוף היום. לאחר ניתוח התוצאות, ניתן להסיק כי הנוסחאות האסימפטוטיות הנן מאוד רגישות לגודל של מוקד השירות. כך, עבור הדוגמא השנייה שניתחנו, בה העומס המוצע לא עבר 5 שעות עבודה לשעה, כמעט כל הנוסחאות נכשלו. המדד היחיד שהיה קרוב לערכו התיאורטי עבור דוגמא זו היה הסתברות ההמתנה. כמו כן, כל הקירובים למעט פונקצית גרנט מביאים לתוצאות טובות עבור מוקדים עם עומס מוצע גדול מכמה עשרות שעות עבודה לשעה. פונקצית גרנט לעומת זאת, משיגה תוצאות טובות גם עבור מוקדים עם עומסים נמוכים.

פונקצית גרנט: הקירוב של גרנט להסתברות ההמתנה הנו חסין מאוד ואפילו עבור הדוגמא השנייה קיבלנו דיוק סביר. גרף ההסתברות האמפירית אומנם לא התלכד עם הקו התיאורטי שנמצא לפי נוסחת גרנט, אך ההפרשים בין שני הגרפים לא היו גדולים מדי. עבור שאר הדוגמאות הקו התיאורטי והאמפירי היו כמעט זהים.

שיטות איוש שונים - ISA ,PSA, Lagged PSA: בדוגמאות שלנו, איוש לפי ISA או לפי Square-Root Safety Staffing אפשרו להגיע לרמת הביצועים היציבה ביותר. באופן כללי, הביצועים של Lagged PSA מאוד קרובים לאלו של ISA (או Square-Root Safety Staffing) ורק במעט פחות יציבים. לעומת זאת, PSA מוצלח פחות ומביא להעסקת שרתים מיותרים בחצי הראשון של היום, כאשר קצב ההגעות גדל, ולחוסר שרתים בחצי השני של היום, כאשר קצב ההגעות קטן.

מערכות תורים עם קצב הגעות לא קבוע בזמן: התפלגות הגעות תיאורטית

בפרק 7, אנו מסבירים כיצד ניתן לזהות את רמות האיוש המובילות לביצועים קבועים של המערכת עבור כל סוגי לקוחות.

אנו משתמשים בפתרון המוצג ב- [10] ומפעילים את אלגוריתם ה- ISA על מערכות עם שני סוגי לקוחות המגיעים לפי תהליך פואסון לא הומוגני בזמן.

תוצאות הניסויים שלנו הן כדלקמן:

- הצלחה כללית בהשגת **רמת ביצועים קבועה בזמן**: ISA היה מאוד מוצלח בייצוב הסתברות ההמתנה α ורמת השירות β, וכמו כן השיג תוצאות סבירות ביציבות זמני המתנה, אורכי תור וקצבי נטישות, במיוחד עבור לקוחות בעדיפות גבוהה.
- רמת הביצועים הגלובאלית של המערכות עם קצבי הגעות לא קבועים מתאימה לזו של **מערכות סטאציונריות** מתאימות. ככל שהחלק היחסי של אחד הסוגים הנו גדול יותר, כך ההתאמה למודל סטאציונרי הנה מדויקת יותר.
- מספר השרתים שנדרש על מנת לספק הסתברות המתנה רצויה הקבועה בזמן הנו פונקציה של **קצב ההגעות הכללי** לעומת וקטור של קצבי הגעות. (ברור לנו, שזה קורה בגלל העובדה שזמני שירות הם בלתי תלויים ושוויי התפלגות). כתוצאה מכך, ניתן לצמצם את בעיית האיוש למציאת מספר שרתים עבור תור עם לקוחות הומוגניים.
- **השפעת נטישות:** לנטישות תפקיד מאוד חשוב והכנסתן למודל לא רק הופכת את המודל לריאליסטי יותר, אלא גם משפרת את ביצועי המודל. כך למשל, זמן הסימולציה על מנת לקבל מצב יציב ירד מ- 72 שעות ל- 24 בלבד. בנוסף, עבור Erlang-A, ISA התכנס עבור כל הערכים של α מ- 0.1 ל- 0.9, כאשר במקרה בלי נטישות האלגוריתם לא התכנס עבור α הגדול מ- 0.75.
נטישות לא רק גורמות להתכנסות יותר מהירה של המודלים, אלא גם מאפשרות חיסכון בכוח עבודה. אנו מראים, כי כתוצאה מלקיחת הנטישות בחשבון, החיסכון במספר השרתים למשמרת הנו 6, 4 ו- 114 עבור α = 0.1, 0.5 ו- 0.9 בהתאמה.

התפלגויות שונות של זמן השירות

בפרק 8 אנו משווים מערכת *M/M/N+M* עם *M/G/N+M* ובודקים את השפעת התפלגות זמן שירות על ביצועי מערכת. הדיון בפרק זה ממשיך ומרחיב את תוצאות הסימולציות עבור מערכות M/G/N המוצגות ב- [32]. מטרתנו בפרק 8 הייתה לבדוק האם השפעתה של התפלגות זמן השירות על תורים עם נטישות דומה לזו במערכות בלי נטישות [32].
לפי תיאוריה קונבנציונאלית של עומס גבוה, תוחלת זמן ההמתנה בכל M/G/N בעומס גבוה ניתנת לקירוב על-ידי נוסחת Kchinchine-Pollazcek. שוורץ [32] מראה כי הקירוב אינו מדויק בתחום הפעילות המכוון איכות ויעילות (QED).

אנו בדקנו את הקירוב לתוחלת זמן ההמתנה האנלוגי ל- Kchinchine-Pollazcek בתורים עם נטישות שפותח ע"י Ward et. al. [33] והראנו כי בתחום QED קיימים הבדלים משמעותיים בין התפלגויות זמן שירות שונות עם שני מומנטים ראשונים שווים. פירושו של דבר הוא שהקירוב אינו מדויק עבור תחום הפעילות QED.