



מרצה : פרופסור אבישי מנדלבאום
מתרגלת : גלית יום-טוב

תאריך הבחינה : 11.2.2010

מס' סטודנט _____

שם _____

הנדסת מערכות שירות – 096324

מועד א' – סמסטר חורף תש"ע 2010

פתרון

- המבחן נמשך שלוש שעות. באחריותך להחזירו במועד.
- המבחן בחומר סגור.
- מותר להשתמש במחשבון.
- המבחן כולל 26 עמודים, כולל עמוד זה.
- כאשר נדרשים חישובים, ניתן לעגלם באופן שאינו פוגע בהבנת התשובה.
- מלאו את הפרטים הנדרשים בראש העמוד, בכתב ברור בבקשה.
- על התשובות יש לענות במקומות המיועדים, שאמורים להספיק: מלל מיותר יגרע ניקוד.
- יש לתת הסברים או הוכחות רק אם התבקשתם במפורש לעשות זאת.

הרצאות/תרגולים/בחינות קודמות	מתוך 10 נקודות	שאלה 1	ניקוד:
תרגילי בית	מתוך 10 נקודות	שאלה 2	
יישום	מתוך 20 נקודות	שאלה 3	
תיאוריה	מתוך 10 נקודות	שאלה 4	
	מתוך 50 נקודות	סה"כ	

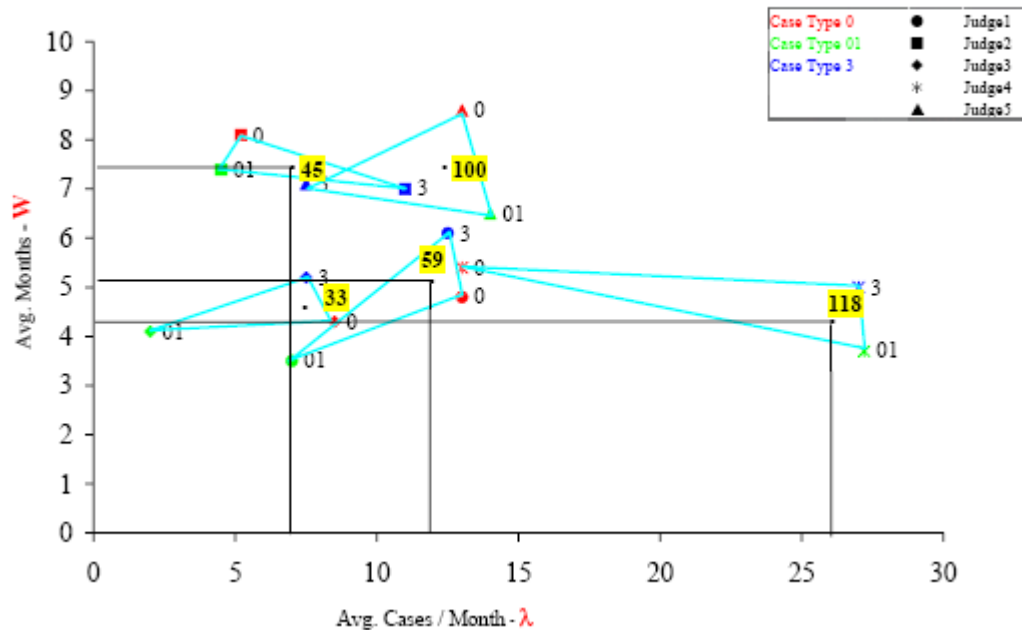
בהצלחה !

שאלה 1 - תרגילי כיתה/הרצאות/מבחנים קודמים

חלק 1.

גרף 1

Judges: The Best/Worst (Operational) Performer



התרשים הנ"ל שהוצג בהרצאה מתאר את רמת הביצועים התפעוליים של חמישה שופטים בבית הדין לעבודה.

1.1.1. מה מתארים שטחי 3 המלבנים שבתרשים? בחרו את התשובה הנכונה (אחת מארבע האפשרויות).

1. מלבן מתאר את סך זמן ההמתנה של לקוחות לכל סוג שופט בתקופות שונות.

2. מלבן מתאר את ממוצע מספר התיקים שנמצאים בטיפול אצל שופט בזמן נתון, וזאת על-פי חוק Little.

3. מלבן מתאר את סך זמן ההמתנה של לקוחות לכל אחד מסוגי התיקים בהם השופט מטפל.

4. אף תשובה לא נכונה.

תשובה: 2

1.1.2. ניתן לנסות להשוות בין רמות הביצועים התפעוליים של שלושה שופטים על סמך שטחי שלושת מלבנים שאנו רואים בתרשים (השטחים שווים ל-45, 59 ו-118). מהי מהות הקשר בין שטח המלבן ורמת הביצועים של השופט? בחרו את התשובה הנכונה (אחת מארבע האפשרויות).

1. שופט עם שטח המלבן גדול יותר הוא שופט עם רמת הביצועים הגבוהה יותר.

2. שופט עם שטח המלבן הגדול יותר הוא שופט עם רמת הביצועים נמוכה יותר.

3. לא נכון להשוות בין רמות הביצועים התפעוליים של שופטים על סמך שטחי המלבנים. לשופט עם רמת הביצועים גבוהה צריך להיות λ גבוה ו- W נמוך.

4. אף תשובה לא נכונה.

תשובה: 3

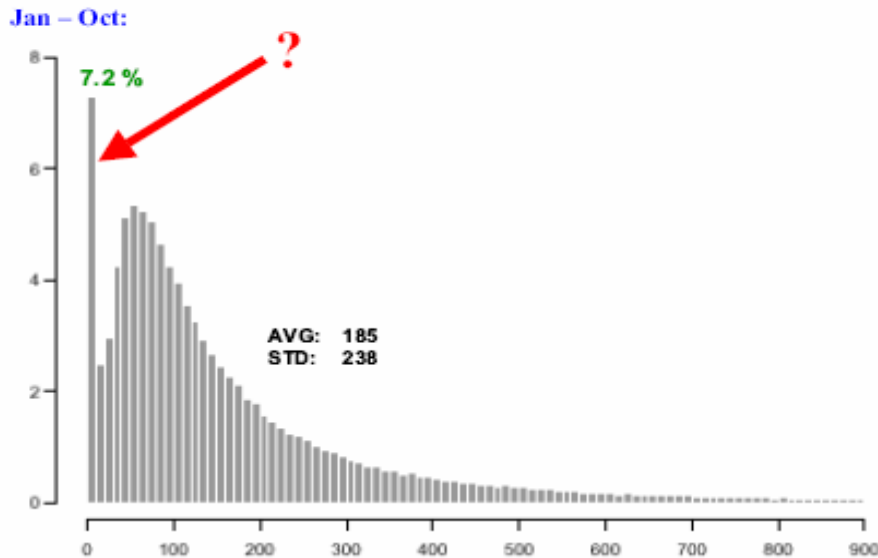
1.1.3. מבין שלושת השופטים, מיהו השופט עם רמת הביצועים התפעוליים הגבוהה ביותר? הסבירו בקצרה.

תשובה:

רמת ביצועים גבוהה מאופיינת בקצב מופע גבוה וזמן שהיה נמוך, ולכן רמת הביצועים של שופט 4 עברו $26 \approx \lambda$ ו- $4.2 \approx W$ היא הגבוהה מבין השלושה.

גרף 2

Beyond Data Averages Short Service Times



1.1.4. גרף 2 שהוצג בהרצאה מתאר משכי שירות במרכז שירות טלפוני. אנו רואים ש- 7.2% מהשיחות משכן קצר ביותר (עד 10 שניות). בחרו פיסקה שמסכמת בצורה הסבירה ביותר את הסיבות לתופעה ואת השלכותיה הניהוליות.

1. התופעה אינה בעייתית. יש לקוחות רבים שבעצם הגיעו למוקדן הלא נכון, ומשם הם מועברים הלאה, ולכן משך השירות שלהם קצר ביותר.

2. התופעה היא בעייתית. הגרף מראה כי תמריץ לפי מספר שיחות מטופלות גורם למוקדנים לנתק ללקוח מיד לאחר קבלת השיחה.

3. התופעה אינה בעייתית. השיחות הקצרות הן לא שיחות טלפוניות, כי אם שרות בעזרת צ'אט. בשירות זה השירות ניתן למקוטעין, וכל נותן שירות מטפל בשניים / שלושה לקוחות במקביל. ולכן משך השיחה מחולק למקטעים קצרים בהם השרת עסוק עם הלקוח, וחלק מהם הם אפילו פחות מ- 10 שניות.

4. התופעה נובעת מבעיה טכנית. מדובר כפי הנראה בתקלה במערכת הטלפונים של החברה.

תשובה: 2

1.1.5 למדנו בהרצאה שאת חווית ההמתנה של לקוח במרכז טלפוני ניתן לסכם באופן הבא :

- E - הזמן שלקוח מצפה להמתין
- τ - הזמן שלקוח מוכן להמתין
- V - הזמן שלקוח נדרש להמתין
- W - הזמן שלקוח למעשה המתין
- F - הזמן שלקוח הרגיש שהמתין

בחרו שוויון או אי שוויון שתמיד מתקיים לפי הגדרות שהוצגו בהרצאה.

1. $F \geq W$

2. $W = \min(\tau, V)$

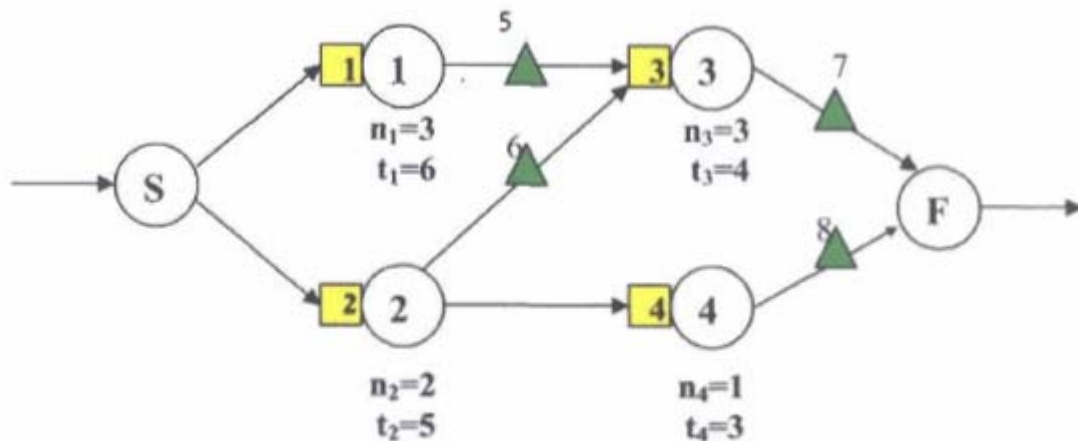
3. $W = \max(\tau, V)$

4. $W = E$

תשובה : 2

חלק 2. רשתות דינמיות סטוכסטיות.

נתונה רשת PERT/CPM דינמית סטוכסטית, כמתואר בגרף 1 וכפי שנלמד בכיתה: t_i הוא ממוצע זמן השירות בתחנה i , ו- n_i הוא מספר השרתים בתחנה i , $i = 1, 2, 3, 4$. פרויקטים חדשים מופיעים למערכת לפי תהליך פואסון עם ממוצע זמן בין מופעי של 3.5 ימים. נניח שהרשת במצב יציב.



Graph Number 1

1.2.1 לרשת 8 תורים הממוספרים מ-1 עד 8. מיהם התורים הנובעים מפערי סינכרון ומי הנובעים ממשאבים מוגבלים?

תשובה:

תורי סינכרון: 5,6,7,8

תורי משאבים: 1,2,3,4

1.2.2 מהו המספר המינימלי של שרתים הנדרשים לטפל בפעילות 3 כדי שנצילות כל אחד לא תעלה על 33%?

תשובה:

מספר מינימלי: $\lceil 4 / (3.5 * n_3) \rceil \leq 0.33$. התוצאה היא 4 שרתים.

1.2.3 נסמן ב- T_{DS} את משך ביצוע פרויקט שמגיע למערכת במצב יציב. הסבירו במדויק ככל שתוכלו, אבל בשני משפטים לכל היותר, מדוע $ET_{DS} < 10$.

תשובה:

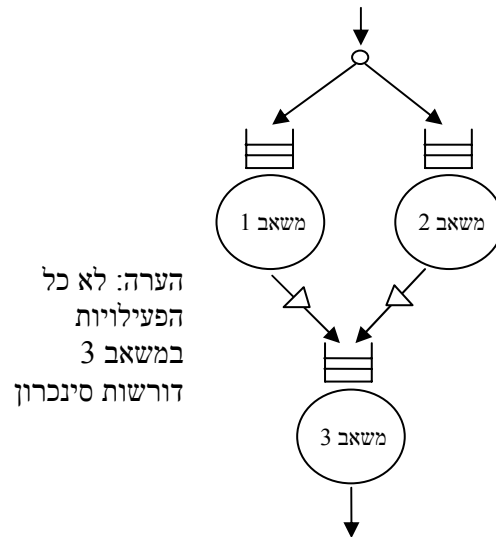
ההסבר כולל שני חלקים. מכיוון שהרשת סטוכסטית ועל סמך מבנה הקדימויות, אנחנו כבר יודעים ש- $10 \leq ET_{DS}$ (ראו הוכחה בהמשך). בנוסף, הדינאמיות של המערכת (ריבוי הפרויקטים) גורמת לתורי משאבים שמאריכים עוד יותר את משך ביצוע הפרויקט.

$$E\left(\max_j \sum_{i \in j} T_i\right) \geq \max_j E\left(\sum_{i \in j} T_i\right) = \max_j \sum_{i \in j} ET_i = 10$$

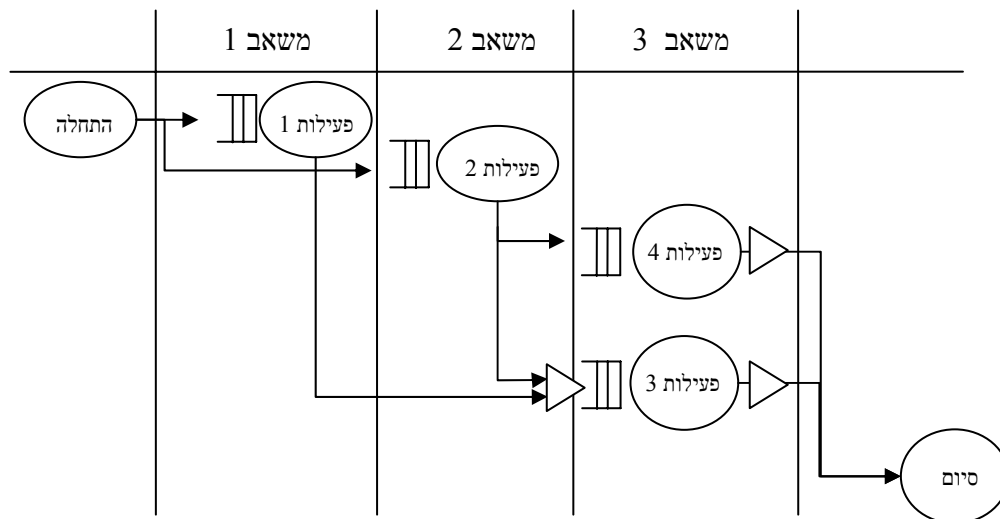
כאשר j מונה את המסלולים האפשריים ברשת.

1.2.4 מנהל החברה הכשיר את השרתים שמבצעים את פעילות 3 ו-4 לבצע את שתי הפעילויות. צייר תרשים משאבים ותרשים משולב משאבים ופעילויות לפרויקט תחת תרחיש חדש זה.

תרשים משאבים:



תרשים משולב פעילויות-משאבים:



שאלה 2 – תרגילי בית

איוש בעזרת 4CC

ארבעה עשר מוקדנים עובדים במרכז השירות של GazolCo. רוב השיחות הן של לקוחות המתקשרים לשלם את חשבון הגז שלהם. מנתוני שיחות שניתחנו עולה כי זמן שירות טלפוני ממוצע הוא 4 דקות, קצב השיחות הנכנסות הוא 180 שיחות לשעה בממוצע, והסבלנות הממוצעת היא 3 דקות. אנו רוצים לעזור למנהל השירות לנתח את המערכת ולקבוע איוש מתאים בעזרת תוכנת 4CC.

2.1. מנהל המערכת רצה לבדוק מה יקרה אם קצב השיחות הנכנסות יגדל (בעוד מספר השרתים נשאר קבוע). להלן שתי טבלאות המתארות מה יקרה במצבים אלו. הטבלאות מבוססות על פלטים של התוכנה, אחד הופק בעזרת מודל M/M/N והשני בעזרת מודל M/M/N+M. איזה מהטבלאות מתאימה לאיזה מודל? הסבירו תשובתכם.

טבלה 1

Calls Per Hour	Average Speed of Answer (secs)	Agent's Occupancy (%)	P(Wait>0)
180	57.8	85.7%	$1-0.518=0.482$
185	79.6	88.1%	$1-0.447=0.553$
190	113.5	90.5%	$1-0.37=0.63$
195	171.3	92.9%	$1-0.286=0.714$
200	289.2	95.2%	$1-0.197=0.803$
205	647.0	97.6%	$1-0.101=0.899$
210	n/a	n/a	

טבלה 2

Calls Per Hour	Average Speed of Answer (secs)	Agent's Occupancy (%)	P(Wait>0)
180	9.2	80.7%	$1-0.699=0.300$
185	10.6	82.2%	$1-0.666=0.333$
190	12.0	83.7%	$1-0.632=0.367$
195	13.5	85.1%	$1-0.598=0.401$
200	15.1	86.3%	$1-0.564=0.435$
205	16.8	87.5%	$1-0.530=0.469$
210	18.6	88.7%	$1-0.497=0.502$

תשובה:

טבלה 1 הופקה בעזרת מודל M/M/N וטבלה 2 בעזרת מודל עם נטישות M/M/M+M. אנו מגלים זאת כי המודל עם הנטישות תמיד יציב, בעוד המודל ללא נטישות לא יציב אם קצב הגעת שיחות גדול מ- 210 שיחות לשעה.

2.2. הסבירו מדוע בשורה האחרונה של טבלה 1 מופיע הערך n/a . מהו קצב ההגעה המקסימלי בו תתקבל תשובה מספרית.

תשובה:

מודל Erlang-C יציב כל עוד $\eta\mu > \lambda$, ולכן $60/4 * 14 = 210$ עבור כל קצב הגעה קטן ממש מ-210 לקוחות לשעה נקבל שהמערכת יציבה והתוכנה תחשב מדדי שירות, ועבור כל ערך גדול או שווה ל-210 נקבל שהמערכת לא יציבה.

2.3. ציינו מה אתם מצפים שיקרה למדדים הבאים, במידה ופרמטר הסבלנות θ יעלה: אחוז הנוטשים, ממוצע זמן ההמתנה לשירות (ASA), ממוצע אורך התור, נצילות השרתים. סמנו את התשובה הנכונה והסבירו תשובתכם בקצרה.

תשובה: כאשר פרמטר הסבלנות עולה:

אחוז הנוטשים: יעלה / ירד

ממוצע זמן ההמתנה לשירות (ASA): יעלה / ירד

ממוצע אורך התור: יעלה / ירד

נצילות השרתים: יעלה / תרד

הסבר: כאשר פרמטר הסבלנות עולה, הסבלנות הממוצעת של לקוח יורדת ולכן יותר לקוחות נוטשים. כתוצאה מכך פחות לקוחות יישארו לחכות לשירות מה שיגרום לתור לקטון ולעומס לרדת. לכן, גם זמן ההמתנה ונצילות השרתים ירדו.

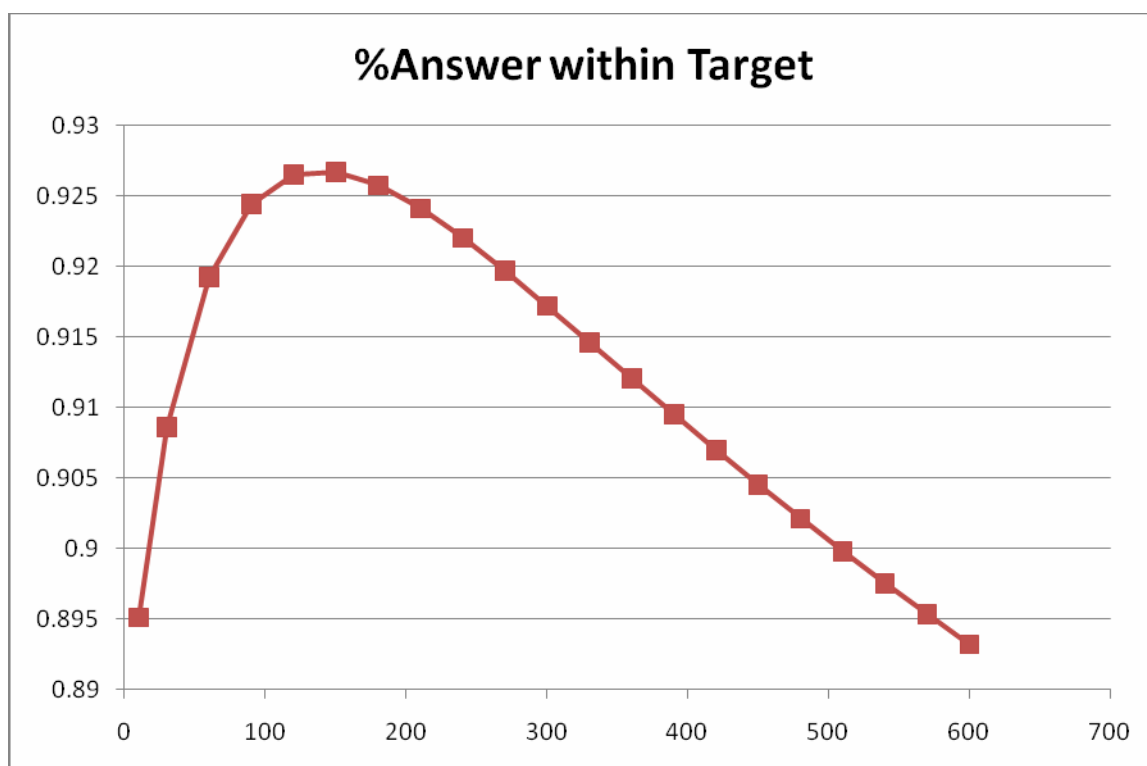
2.4. מה יקרה לאחוז הנוטשים אם סבלנות הלקוחות תעלה לאינסוף או תרד ל-0? לאיזה מודל המערכת תשאף במקרים אלו? הסבירו תשובתכם.

תשובה:

$\theta \rightarrow \infty$: כאשר פרמטר הסבלנות עולה לאינסוף, ממוצע הסבלנות שואף ל-0, ולכן כל לקוח שמגיע כאשר אין שרת פנוי עוזב מיד את המערכת. כלומר, % הנוטשים בגבול יהיה שווה לאחוז הלקוחות המגיעים לשרת עסוק - $P(Abandon) \rightarrow P(W > 0)$. במצב זה המערכת שואפת למערכת M/M/N/N (Erlang-B).

$\theta \rightarrow 0$: כאשר פרמטר הסבלנות יורד ל-0, ממוצע הסבלנות שואף לאינסוף, אף לקוח לא נוטש (כלומר, $P(Abandon) \rightarrow 0$), ולכן המערכת שואפת למערכת M/M/N (Erlang-C).

2.5. כאשר בודקים את התנהגות של "אחוז הנענים תוך 1.5 דקות" כפונקציה של ממוצע הסבלנות מקבלים את הגרף הבא:



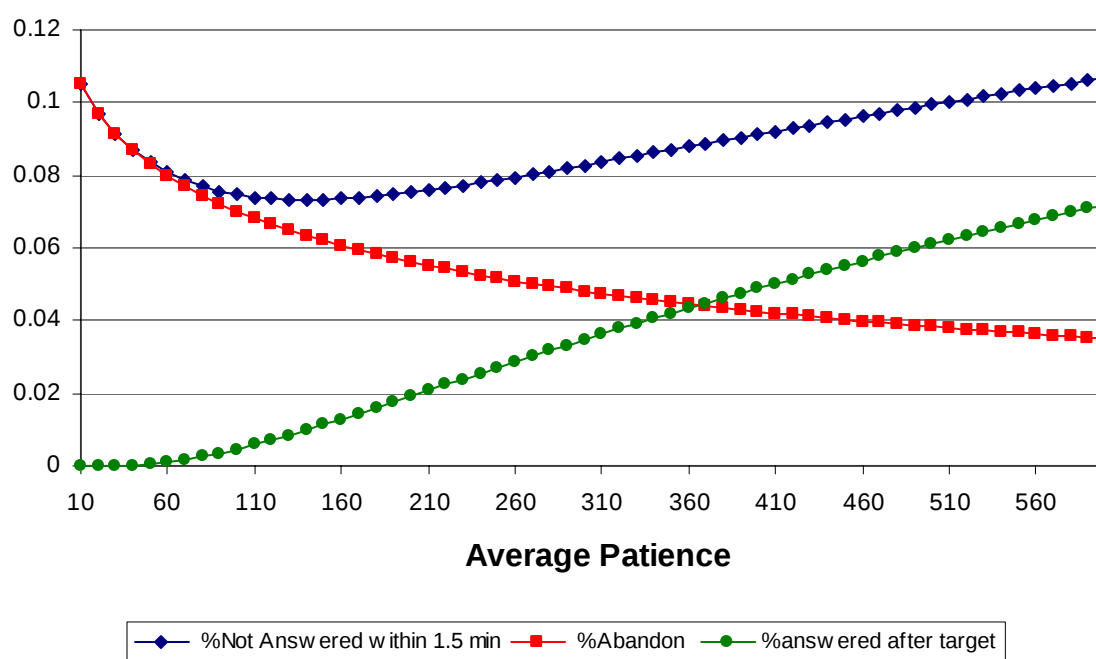
הסבירו מדוע לגרף צורה כזו.

רמז: ניתן לחלק את הלקוחות שלא נענים תוך 1.5 דקות לשתי קבוצות. הגדירו מהן, וכיצד הן עוזרות להסברת התופעה הנ"ל.

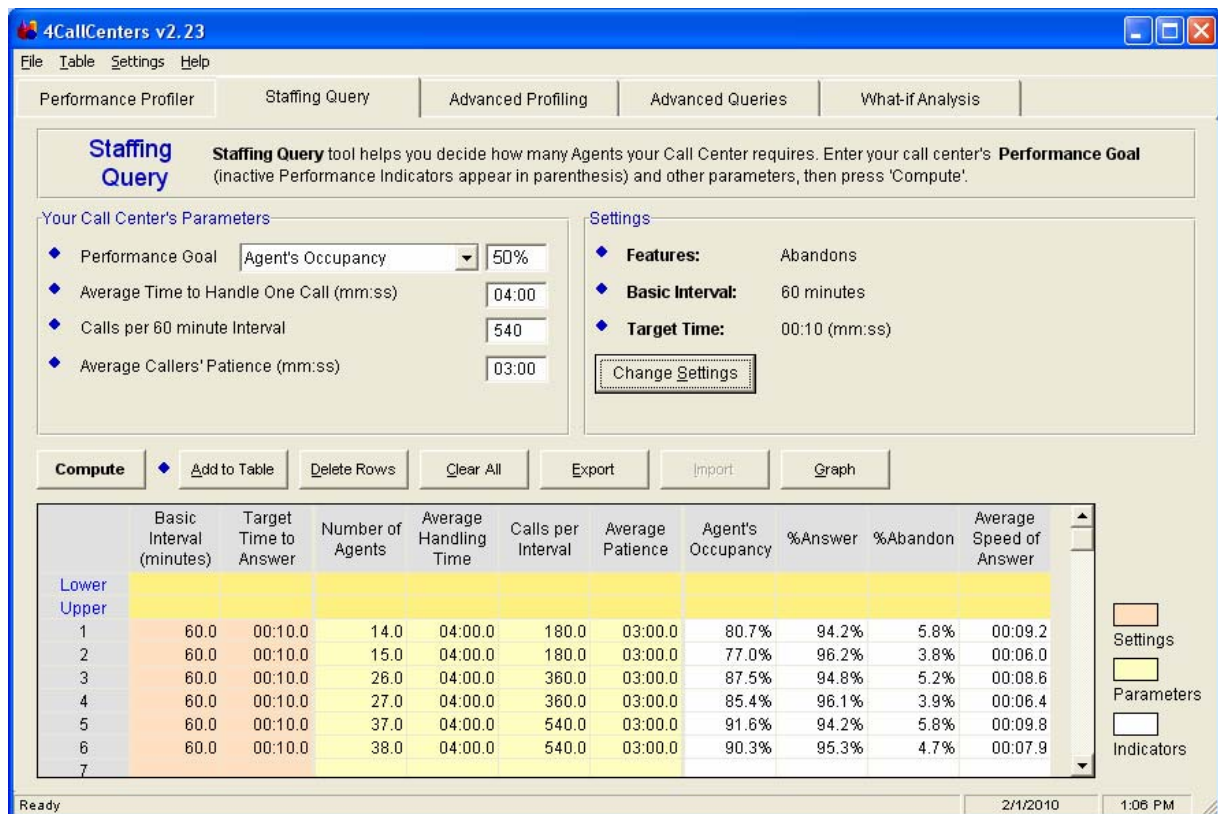
תשובה:

ניתן לחלק את הלקוחות שלא נענים תוך 1.5 דקות, ללקוחות שנטשו (לפני או אחרי 1.5 דקות) וללקוחות שלא נטשו וקיבלו שרות לאחר יותר מ- 1.5 דקות.

אחוז הלקוחות הנוטשים יורד כפונקציה של ממוצע הסבלנות. אחוז הלקוחות שלא נטשו ומקבלים שרות לאחר יותר מ- 1.5 דקות עולה ככל שממוצע הסבלנות עולה (כיוון שאם ללקוחות יש יותר סבלנות יהיו יותר לקוחות בתור $\leq$ זמני ההמתנה יעלו $\leq$ אחוז אלו שלא נוטשים ומקבלים שרות לאחר יותר מ- 1.5 דקות עולה). לכן, חיבור שתי הפונקציות המונוטוניות הנ"ל גורם לכך שאחוז הלקוחות שלא קיבלו שירות תוך 1.5 דקות כפונקציה של ממוצע הסבלנות יראו לפי הגרף למטה. מכאן נובע שאחוז הלקוחות שכן קיבלו שרות תוך 1.5 דקות היא פונקציה כפי שרואים בשאלה.



2.6. מנהל המוקד גילה שבימים מסוימים (לאחר שליחת חשבונות ללקוחות), קצב השיחות למוקד גדול פי כמה מימים רגילים. ליתר דיוק, ביום לאחר שליחת החשבונות קצב ההגעה הוא 540 לקוחות לשעה (פי 3 מיום רגיל) וביום למחרת 360 לקוחות לשעה (פי 2 מיום רגיל); לאחר מכן הקצב חוזר לרמתו הרגילה. מנהל המוקד רוצה לחשב את כמות המוקדנים הדרושה כך שאחוז הנוטשים ישאר ברמתו של יום רגיל (קרי כ- 5%). לשם כך הוא יצר בתוכנת 4CC את הפלט הבא:



הוא טוען שניתן לראות שבמערכת שלנו קיים יתרון לגודל באיזושהי האם הוא צודק? הסבירו.

תשובה:

המנהל צודק. כפי שניתן לראות מהנתונים, כאשר קצב ההגעות עולה פי 2 מספר השרתים עולה רק פי 1.8; כאשר קצב ההגעות עולה פי 3, מספר השרתים הדרוש על מנת לשמור על אותה רמת שירות עולה רק פי 2.5. זוהי בדיוק התופעה של "יתרון לגודל".

שאלה 3 – יישום

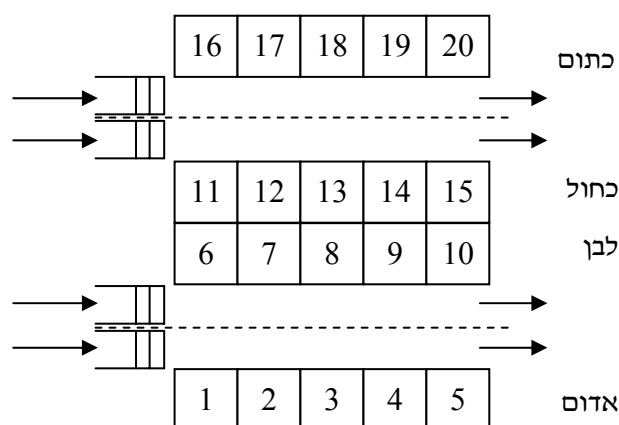
להלן תקציר של מכתב שקיבלנו מסטודנטית לשעבר בקורס הנדסת מערכות שירות:
 "רציתי לשתף אותך בתמונות שצילמתי במהלך החופשה בסופרמרקט בניו-יורק, ברשת הסופרמרקטים
 "Whole Foods", שנמצאת במנהטן ב-Union Square. זאת רשת סופרמרקטים יוקרתית מאוד, גדולה
 מאוד, שמוכרת הרבה מוצרי בריאות יקרים יותר ופחות.

מצורפות שתי תמונות מהתור לקופות ביציאה.

הקופות ממוקמות במקום אחד בחנות. יש ארבעה תורים במתחם הקופות. לכל אחד מארבעת התורים יש
 צבע: כתום, כחול, לבן, אדום. כשקופה מתפנה, מופיע מספר הקופה אליה יש לגשת, על גבי מסך מעל
 התור, וזאת לפי צבע התור בו עומדים הלקוחות (כפי שרואים בתמונה): התור הכתום לקופה 9, התור
 האדום לקופה 12.

הקופות מסודרות בשתי קבוצות גדולות, כך שבכל קבוצה עשר קופות עם מעבר ביניהן – חמש מכל צד, לפי
 האיור המצורף.

חיכיתי בתור פחות מ-3 דקות, לדעתי, למרות שהחנות הייתה עמוסה."



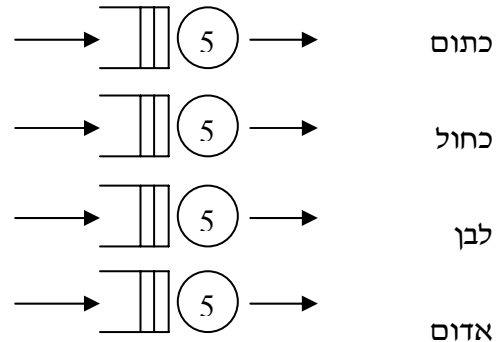
בשאלה זו ננתח תפעולית את תכנון הקופות בסופרמרקט זה.

נניח כי תהליך הגעת לקוחות לתור הוא תהליך פואסון עם קצב λ , ושמשך השירות בקופה מתפלג מעריכית עם קצב μ . כמו כן נניח, בשלב זה, כי תוחלת משך השירות היא 5 דקות. בנוסף נניח כי הזמן שלוקח ללקוח לאתר את הקופה הפנויה ולהגיע אליה הוא זניח.

3.1. תארו גרפית את תצורת תורי הקופות בסופרמרקט.

תשובה:

בהנחה שלקוחות לא עוברים בין התורים, התרשים יראה כך:



נתמקד מעתה באחד התורים (לדוגמה התור ה"כחול"). נסמן ב- $Q(t)$ את מספר הלקוחות במערכת השירות בזמן t . $Q(t)$ כולל את הלקוחות הממתינים בתור הכחול, כמו גם את הלקוחות שמקבלים כרגע שירות בעמדות הקשורות לתור זה). אזי, $Q = \{Q(t), t \geq 0\}$ הוא תהליך לידה ומוות תורי, תחת הנחות מתאימות, בפרט זמן שירות מעריכי. נניח מעתה שהנחות אלו אכן מתקיימות.

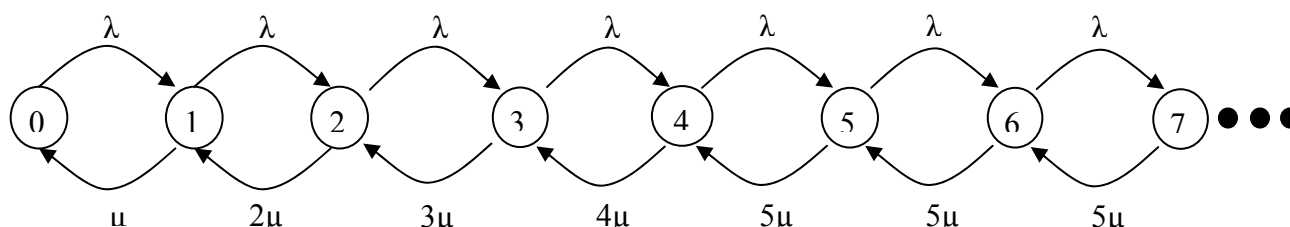
3.2. איזה מודל תורי שנלמד בכיתה מתאים לתיאור התור הנ"ל.

תשובה:

בהנחה שלקוחות לא נוטשים את התור (לא כדי לעזוב את הסופר ולא כדי לעבור לתור אחר), כל אחד מהתורים יכול להיות מתואר ע"י מודל Erlang-C, כלומר תור M/M/5.

3.3. שרטטו את דיאגרמת קצבי המעבר של Q . היעזרו בסימונים הבאים: λ – קצב מופע לקוחות לתחנה. μ – קצב שירות בתחנה. N – מספר קופאים הנותנים שירות לתור הכחול.

תשובה:



3.4. מדוע לדעתכם נוצר הצורך לקבוע לכל תור צבע ולהכריז במסך על מספר העמדה שהתפנתה? כיצד מערכת זו משפיעה על תפעול המערכת?

תשובה:

הגדרת התורים והמסך אמורים לשפר את זרימת הקונים. כך שלקונה יהיה קל יותר לשים לב שהתפנתה עמדה ואיזו עמדה פנויה כרגע. באופן זה זמן הבטלה של הקופאים במעבר בין קונה אחד למשנהו יורד.

3.5. הסידור המקובל של קופות בסופרמרקטים בישראל הוא שלכל קופה יש תור משלה. נניח כעת כי לסופרמרקט הישראלי מגיעים אותו מספר אנשים לשעה, משך השירות בקופה דומה ומספר הקופאות זהה לסופרמרקט האמריקאי. האם בתצורה האמריקאית צפויה רמת שרות טובה יותר ללקוחות? הסבירו. בפרט, האם יהיה משך ההמתנה לשירות קצר יותר, ובכמה?

תשובה:

בסופר האמריקאי צפויה רמת שירות טובה יותר. זה נובע מתופעת ה-pooling. ברגע שלכל קבוצת קופות יש תור אחד במקום תורים נפרדים זמני ההמתנה בתור יקטנו, וההסתברות להמתין תקטן. לדוגמה: זמן

ההמתנה הממוצע בתור יקטן פי 5. (ליתר דיוק ב- $\frac{5\rho}{E_{2,5}}$, אבל ρ ו- $E_{2,5}$ הם באותו סדר גודל).

הערה: זה נכון אם משך ההליכה לקופות זניח, במידה ולא, בדיקה יותר מעמיקה צריכה להתבצע, וההפרש בין התצורות יקטן.

3.6. האם בתצורה האמריקאית צפוי להיות בממוצע מספר שונה של לקוחות הממתינים בתור (בכל המערכת)? הסבירו.

תשובה:

כן. ע"ס חוק Little, אם תוחלת זמן ההמתנה קטן, גם תוחלת מספר האנשים בתור קטן.

מידות שבוצעו בסופר הראו את הנתונים הבאים.

Report 1

Time	Arrivals	Average Service Time (sec)	Offered Load	Number of Agents	Average Wait (sec)	Served Immediately (%)
7:00	58	410	13.2	15.3	107	55.8
7:30	42	450	10.5	13.9	42	72.4
8:00	51	433	12.3	14.5	78	57.9

Report 2

Time	Arrivals	Average Service Time (sec)	Offered Load	Number of Agents	Average Wait (sec)	Served Immediately (%)
18:00	61	582	19.7	20.0	2089	4.7
18:30	58	613	19.7	20.0	2153	5.8
19:00	56	603	18.8	19.6	1578	9.2

Report 3

Time	Arrivals	Average Service Time (sec)	Offered Load	Number of Agents	Average Wait (sec)	Served Immediately (%)
22:00	15	407	3.4	10.8	0.0	100.0
22:30	17	357	3.4	9.6	5.1	99.1
23:00	12	312	2.1	8.9	0.0	100.0

3.7. הסבירו כיצד מחושב ה- Offered Load:

תשובה:

ה- offered-load מחושב ע"י הנוסחה $R = \frac{\lambda \cdot E(S)}{1800}$, כאשר $E[S]$ הוא ממוצע זמן השירות ו- 1800 הוא מספר השניות באינטרוול של חצי שעה.

3.8. התאימו לכל דו"ח את התחום התפעולי שהתקיים באותו זמן. נמקו בקצרה.

תשובה:

דו"ח 1: QED. כיוון שתוחלת זמן ההמתנה היא בסדר גודל אחד פחות ממשך השירות, ניצולת השרתים סבירה, וההסתברות להמתנה בין 25% ל- 75%.

דו"ח 2: ED. כיוון שתוחלת זמן ההמתנה היא באותו סדר גודל כמו זמן השירות, ניצולת השרתים גבוהה, וההסתברות להמתנה שואפת ל-1.

דו"ח 3: QD. כיוון שתוחלת זמן ההמתנה שואפת ל-0, ניצולת השרתים נמוכה, וההסתברות להמתנה שואפת ל-0.

3.9. אפיינו לכל דו"ח את כלל האיוש הרלוונטי (עפ"י התחום התפעולי שהגדרתם). חשבו את פרמטר האיכות (β, δ , או γ) לשורה האמצעית בכל דוח.

תשובה:

דו"ח 1: QED - $\beta = \frac{N-R}{\sqrt{R}} = 1.05, N = R + \beta\sqrt{R}, \beta > 0$

דו"ח 2: ED - $N = R + \gamma, \gamma = N - R = 0.3$

דו"ח 3: QD - $\delta = N/R - 1 = 1.82, N = R + \delta R$

3.10. אמדו בעזרת הגרף המצורף מה יהיה אחוז הלקוחות המשורתים מיידית עבור אינטרוול זמן שבו : קצב הגעה – 60 לקוחות בשעה, זמן שרות ממוצע 10 דקות, 12 קופאיות נותנות שירות.

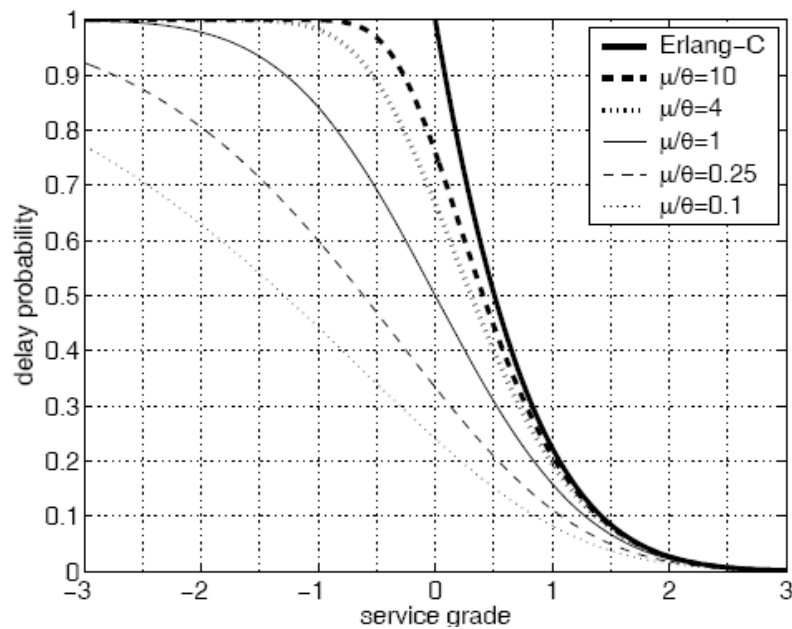
תשובה:

תחילה נחשב את ה- Offered Load : $R=60/6=10$

עתה נחשב את פרמטר השירות :

$$\beta = \frac{N - R}{\sqrt{R}} = \frac{12 - 10}{\sqrt{10}} = 0.63$$

על פי הגרף ההסתברות להמתנה תהיה כ- 0.4, ולכן אחוז המקבלים שרות מיידית הוא 60%.



3.11. מנהל הסופר רצה לחשב תחזית למספר הקונים בחנות בכל שעה. מניסיונו הוא יודע שמספר האנשים משתנה בין שעות היום ובין ימות השבוע. המנהל יכול לאסוף בקלות מידע מהקופות על מספר הקונים שרכשו בחנות בכל שעה בשבוע ברבעון האחרון. הוא החליט לחלק את היממה לאינטרוולים של חצי שעה. נגדיר: $N_{i,j}$ – הגעות בפועל ביום i באינטרוול j , ו- $F_{i,j}$ – תחזית ליום i באינטרוול j . המנהל רוצה להשתמש בשיטת הממוצע הנע, על פני 4 המדידות האחרונות. הגדירו כיצד יש לבצע את החישוב לחיזוי מספר הקונים הצפוי בכל שעה על פי מודל זה. השתמשו בסימונים: $F_{i,j}$ $N_{i,j}$.

תשובה:

בשיטת הממוצע הנע (4) משתמשים בנוסחה הבאה:

$$F_{i,j} = \frac{N_{i-3,j} + N_{i-14,j} + N_{i-21,j} + N_{i-28,j}}{4}$$

כדי לחשב את מספר הקונים בשעה יש לסכום שני אינטרוולים.

3.12. בכיתה הוגדרו שני מדדים לקביעת טיב התחזית RMSE (Root Mean-Square Error) ו- APE (Average Percent Error). תארו במשפט אחד כיצד מסיקים באיזו שיטת תחזית לבחור, בהתבסס על מדדים אלו. איזה מדד מהשניים יותר אינטואיטיבי לשימוש? הסבירו מדוע.

תשובה:

ככל שהמדדים קטנים יותר כך התחזית טובה יותר, ולכן נבחר במודל עם הערכים המינימליים. המדד APE יותר אינטואיטיבי, כיוון שקל יחסית להגדיר מהו אחוז הטעות שאנו מוכנים לסבול במערכת כלשהי.

3.13. מנהל הסופר שוקל לשנות את התורים כך שכל שני תורים (סמוכים) ישורתו ע"י 10 קופות. הניחו כי מדיניות תזמון התורים היא כזו שעדיין נשמרת תכונת ה-work conservation, כלומר, לא יתכן שיש לקוחות באחד התורים וגם שרת פנוי בו זמנית. בנוסף, אין הבדל בין הלקוחות בשני התורים ומדיניות התזמון ביניהם תתבצע בצורה הוגנת. האם כתוצאה משינוי זה יתקצרו משכי ההמתנה?

תשובה:

מה שמנהל החנות מציע הוא פחות או יותר pooling של המערכת משני תורים $M/M/5$ עם קצב כניסה λ , לתור $M/M/10$ עם קצב כניסה 2λ . כאשר מבצעים pooling זמני ההמתנה קטנים. ולכן נצפה שגם כאן יתקצרו זמני ההמתנה בתור, בממוצע פי 2 בערך.

3.14. כמדיניות תזמון הוגנת, שוקל מנהל החנות מדיניות Round-Robin, שפירושה שירות שני התורים לסירוגין: לקוח מתור 1, הבא מתור 2 (אם התור לא ריק), חזרה ל-1, וחוזר חלילה. האם מדדי הביצוע של מדיניות זו יהיו שונים מתור Erlang-C רגיל (עם 10 שרתים), בה יש תור אחד לכל הלקוחות? בפרט, ציינו אילו מהמדדים הבאים ישתנו ואילו לא: ממוצע סך האנשים במערכת (בתור+שירות), ממוצע מספר הממתנים בתור, נצילות השרתים, ההסתברות להמתין, וההסתברות להמתין יותר מחמש דקות.

תשובה:

למבנה זה ולתור $M/M/10$ לא יהיו באופן כללי אותם מדדי ביצוע. המדדים שלא ישתנו הם מדדים התלויים בהתפלגות סך האנשים במערכת (בתור + שירות), כיוון שהתפלגות זו תלויה רק בקצב המופע והשירות ולא במדיניות התור. כיוון שהתפלגות זו נקבעת על ידי אותו תהליך לידה ומוות. לכן, בפרט המדדים הבאים לא ישתנו: ממוצע מספר האנשים במערכת, ממוצע אורך התור, ממוצע זמן המתנה בתור, נצילות השרתים וההסתברות להמתין. מאידך, התפלגות מספר הממתנים בכל אחד מתורי ה-Round-Robin דורשת מידול נפרד (בפרט תהליך קפיצה מרקובי דו-מימדי). לכן, חישוב כל מדד שתלוי במידע זה, עשוי לתת תשובה שלא ניתן לגזור אותה ממודל Erlang-C. לדוגמה, התפלגות זמן ההמתנה מושפעת ממדיניות התזמון, מאחר והיא תלויה במספר האנשים שהיו בתור המסוים אליו מצטרף הלקוח. לכן, מדדים כמו $P(W>T)$ יהיו שונים בין המודלים.

4.1. בכיתה הוסבר המשפט הבא המכסה את מודל Erlang-B בתחום התפעול QED. בפרט הוכחנו שסעיף 3 במשפט (איוש לפי כלל השורש), גורר את 1. דהיינו

$$\lim_{n \rightarrow \infty} \sqrt{n} E_{1,n} = h(-\beta)$$

כאשר h היא פונקצית קצב הסיכון של ההתפלגות הנורמאלית הסטנדרטית. הוכיחו טענה זו תוך השלמת החלקים החסרים בהוכחה הנ"ל. (ההוכחה ניתנת בעמוד הבא).

The Erlang-B Queue in the QED-Regime

Recall the **Erlang-B Formula**:

$$E_{1,n} \triangleq P\{\text{Blocked}\} = \frac{R^n}{n!} \bigg/ \sum_{j=0}^n \frac{R^j}{j!}$$

Consider a sequence of **M/M/n/n** queues, indexed by **the number of servers $n = 1, 2, \dots$** .

- λ_n = arrival-rate, varies with n ;
- μ = service-rate, fixed (independent of n).
- $R_n = \lambda_n/\mu$ (Offered Load) ; $\rho_n = R_n/n$ (Load per Server);
We shall use R and ρ , without the subscript n , for simplicity.

Theorem (QED Erlang-B; Jagerman, 1974)

As $n \rightarrow \infty$, the following 3 statements are equivalent:

1. **Customers:** $E_{1,n} \approx \frac{\gamma}{\sqrt{n}}$, for some $\gamma > 0$;
2. **Servers:** $\rho \approx 1 - \frac{\beta}{\sqrt{n}}$, for some $-\infty < \beta < \infty$;
3. **Manager:** $n \approx R + \beta\sqrt{R}$ (square-root “staffing”);

in which case

$$\gamma = h(-\beta) = \frac{\phi(-\beta)}{\bar{\Phi}(-\beta)} = \frac{\phi(\beta)}{\Phi(\beta)},$$

where $\phi, \Phi, \bar{\Phi}$ and h are the density, cdf, survival function and hazard rate of $N(0, 1)$ (standard-normal), respectively.

Note: **Servers' Occupancy** $\approx 1 - \frac{\beta + \gamma}{\sqrt{n}}$, accounting for blocking.

QED Erlang-B: Proof

Proof:

2 $\iff$ **3** is straightforward algebra from the definitions.

3 $\Rightarrow$ **1**. Assume $n = R + \beta\sqrt{R}$. The key observation is a **Poisson-Representation** of the Erlang-B Formula:

$$E_{1,n} = \frac{\mathbb{P}\{X_R = n\}}{\mathbb{P}\{X_R \leq n\}},$$

where $X_R \stackrel{d}{=} \text{Poisson}(R)$.

$$\begin{aligned} \mathbb{P}\{X_R \leq n\} &= \mathbb{P}\left\{\frac{X_R - R}{\sqrt{R}} \leq \frac{n - R}{\sqrt{R}}\right\} \\ &\stackrel{CLT,3}{\approx} \mathbb{P}\{N(0,1) \leq \beta\} = \Phi(\beta). \end{aligned}$$

$$\begin{aligned} \mathbb{P}\{X_R = n\} &= \mathbb{P}\{n-1 < X_R \leq n\} \\ &= \mathbb{P}\left\{\frac{n-R-1}{\sqrt{R}} < \frac{X_R - R}{\sqrt{R}} \leq \frac{n-R}{\sqrt{R}}\right\} \\ &\approx \mathbb{P}\left\{\beta - \frac{1}{\sqrt{R}} \leq N(0,1) \leq \beta\right\} \\ &\approx \left[\frac{1}{\sqrt{R}} \cdot \phi(\beta)\right] \approx \frac{1}{\sqrt{n}} \cdot \phi(\beta). \end{aligned}$$

Finally, $\left[\frac{\phi(\beta)}{\Phi(\beta)} = \frac{\phi(-\beta)}{1-\Phi(-\beta)}\right] = h(-\beta)$, by the symmetry of $N(0,1)$.

4.2. נניח קצב מופע λ , וקצב שירות μ .
 בכיתה נאמר שחישוב נאיבי של איוש n , ממליץ :

$$n = \left\lceil \frac{\lambda}{\mu} \right\rceil = [R]$$

(עיגול של R כלפי מעלה לשלם)

4.2.1. למה תוביל המלצה זו במודל Erlang-C? בפרט מה תהיה (בערך) ההסתברות להמתין?

תשובה:

המלצה זו תוביל לאיוש בתחום ה-ED. בפרט, ההסתברות להמתין תשאף ל-1.
 אם R שלם בעצמו המערכת לא תהיה ארגודית, ולכן תתפוצץ. אם R אינו שלם, אזי
 $n = R + \gamma$, $\gamma = [R] - R$, ולכן ממוצע זמן ההמתנה יהיה $\gamma \cdot E[S]$.

4.2.2. למה תוביל המלצה זו במודל Erlang-A? בפרט, נניח $\theta = \mu$, מה תהיה (בערך) ההסתברות להמתין במקרה זה?

תשובה:

המלצה זו תוביל לאיוש בתחום ה-QED. בפרט, ההסתברות להמתין תשאף לערך בין 0 ל-1, הערך המדויק תלוי ביחס בין μ ל- θ . במקרה ש- $\mu = \theta$, הערך ישאף ל-0.5.
 הסבר ל-0.5: במקרה ש- $\mu = \theta$, התפלגות סך מספר הלקוחות במערכת במודל Erlang-A זהה להתפלגות סך מספר הלקוחות במערכת במודל $M | M | \infty$ עם קצב מופע λ וקצב שירות μ , כאשר התפלגות סך מספר הלקוחות במערכת במודל $M | M | \infty$ היא $Poisson(R = \lambda / \mu)$.
 מהחוק החזק של המספרים הגדולים, עבור R מספיק גדול, ניתן לקרב את ההתפלגות $Poisson(R)$ על-ידי התפלגות $Normal(R, R)$. ומכאן,

$$\begin{aligned} P(W_q = 0) &\stackrel{PASTA}{=} P(L \leq R) = P\left(\frac{L - R}{\sqrt{R}} \leq \frac{R - R}{\sqrt{R}}\right) \approx \\ &= P(N(0,1) \leq 0) = 0.5 \end{aligned}$$

4.2.3. בכיתה הוסבר למה שאלה 4.2.2 מהווה דוגמה למקרה שבו "מנהלים מקבלים תשובה נכונה מהסיבות הלא נכונות" – הסבירו במדויק ככל שתוכלו.

תשובה:

המנהל פועל בצורה נאיבית הוא מכפיל את מספר האנשים שמגיעים בשעה בממוצע זמן השירות וכך מחליט כמה מוקדנים לאייש. הסיבה שזה עובד טוב למרות השגיאה היא שיש נטישות, שמייצבות את המערכת ושומרות על רמת שרות טובה.