

סמינר בחקר ביצועים – אביב תשס"ו

Skill and Value Based Routing

מציגים:

קוסטה אלקין

ירון פידלר

מנחה: פרופסור אבישי מנדלבאום

מבוסס על:

A Staffing Algorithm for Call Centers with Skill-Based Routing (Whitt & Wallace)

A value-based routing and preference-based routing in customer contact centers
(Sisselman & Whitt)

Staff Scheduling for Inbound Call Centers and Customer Contact Centers (Fukunaga,
Hamilton, Fama, Andre, Matan, Nourbakhsh)

Skill Based Routing

במה עוסק Skill Based Routing?

- השיחות מתחלקות לסוגים ונושאים שונים.
- לכל סוג פנייה יש להכשיר שרת.
- ניתוב השיחות יעשה לפי סוגן לשרת המתאים.

השאלות העיקריות:

- מהי מדיניות ניתוב השיחות?
- לכמה יכולות להכשיר כל שרת (על כמה סוגי שיחות יוכל כל שרת לענות)?
- כמה שרתים מכל סוג?
- כמה קווי טלפון להתקין במוקד?
- הפתרון ימצא על ידי סימולציה ונראה שהוא קרוב מאוד לפתרון אופטימאלי.

הנחות וסימונים

- C שרתים
- K מקומות בתור
- n סוגי שיחות ($k = 1 \dots n$)
- הוספת שרת יקרה הרבה יותר מהוספת קו טלפון
- W_k תוחלת זמן המתנה של שיחה מסוג k במצב היציב
- Q_k מספר פניות מסוג k במערכת במצב היציב (בשירות+תור)
- (C, K) מתאר מוקד עם C שרתים ו- $K + C$ קווי טלפון
- זמני השירות וזמנים בין הגעות מתפלגים אקספוננציאלית
- שיחות המגיעות כאשר כל ה- $K + C$ קווים תפוסים נחסמות ונאבדות (Erlang B).
- מדיניות FCFS לכל סוג שיחה

הנחות וסימונים- המשך

- אין התייחסות לנטישות וחיוגים חוזרים (לכן גם אין התייחסות לסבלנות הלקוחות)
- לשרת יש את יכולת k אם הוא מסוגל לטפל בפנייה מסוג k
- מדיניות השירות תגדיר איזה שרת מקבל איזה פנייה
- אילוצים: ישנם שתי קבוצות אילוצים שבהם נצטרך לעמוד

הסימון בעייתי.
הכוונה היא שלכל
סוג מופע נדרוש
הסתברות סף
שונה לחסימה

1. הסתברות לזמן המתנה קטן מ- τ_k לסוג שיחה k :

$$P(W_k \leq \tau_k | Q_k < C + K) \geq \delta_k$$

2. הסתברות להגעה למערכת תפוסה: $P(Q_k = C + K) \leq \varepsilon_k$

כאשר ε_k ו- δ_k הם הסתברויות הסף

מדיניות ניתוב השיחות – ייצוג יכולות השרתים

- לכל מוקדן מספר יכולות בעדיפויות שונות.
- נציג את היכולות והעדיפויות של כל המוקדנים במטריצת הסוכנים A בה יהיו C שורות (שורה לכל סוכן) ו- n עמודות (לכל יכולת).
- $A_{i,j} = k \Leftrightarrow$ לסוכן i ישנה יכולת k בעדיפות j
- $A_{i,j} = 0 \Leftrightarrow$ לסוכן i אין יכולת בעדיפות j
- אם לא קיים j עבורו $A_{i,j} = k$ אז לסוכן i אין את יכולת k
- לכל סוכן מקסימום יכולת אחת בעדיפות אחת (אין באותה רמת עדיפות יותר מיכולת אחת)
- לכל סוכן, יכולת יכולה להופיע רק בעדיפות אחת.

דוגמא למטריצת שרתים

$$A_{6,4} = \begin{pmatrix} 3 & 4 & 1 & 0 \\ 1 & 4 & 0 & 0 \\ 2 & 3 & 0 & 0 \\ 4 & 0 & 0 & 0 \\ 3 & 1 & 2 & 4 \\ 1 & 0 & 4 & 0 \end{pmatrix}$$

במטריצה מיוצג מוקד עם 6 שרתים ו 4 סוגי שיחות.
עמודה = היכולת של מוקדן לטפל בשיחה כלשהי
(עמודה ראשונה = היכולות הראשיות של המוקדנים)
שורה = מוקדנים

מדיניות ניתוב השיחות – ייצוג יכולות

השרתים - המשך

- $A_{i,1}$ היא היכולת העיקרית של שרת i
- לכל שרת ישנה יכולת עיקרית אחת
- נגדיר קבוצת עבודה $G_k = \{i: 1 \leq i \leq C, A_{i,1} = k\}$ - כל השרתים בעלי יכולת ראשית k . כלומר כל שרת שייך לקבוצת עבודה אחת בלבד
- C_k הוא גודל קבוצת G_k , כלומר מספר השרתים עם יכולת עיקרית k
- נדרוש שלכל יכולת k יהיה ייצוג במטריצה, אבל לא חובה שלכל יכולת k יהיה שרת עם יכולת ראשית k .

.17

ניתוב שיחה

• ניתוב שיחה מתחלק לשני מקרים:

1. כאשר שיחה חדשה מסוג k מגיעה למוקד:

ראשית נעבור על כל מוקדן מקבוצת עבודה G_k ונפנה את השיחה לשרת פנוי לפי מדיניות Liar. (longest-idle-agent- routing)

במידה ואין שרת פנוי מקבוצת עבודה G_k נחפש שרת שיכולתו המשנית היא k וכך הלאה. אם אין אף שרת שמסוגל לטפל בפנייה השיחה תועבר לתור.

2. כאשר סוכן מתפנה:

אם אין פניות בתור השרת מתבטל.

אם ישנם שיחות בתור השרת "מחפש" שיחה בהתאם ליכולות שלו. המעבר על השיחות יעשה בהתאם לסדר העדיפויות של היכולות, כך שקודם כל השרת יחפש שיחה השייכת ליכולתו הראשית. במידה ואין שיחה בה הוא יכול לטפל הוא חוזר למצב בטלה.

לכמה יכולות להכשיר שרת?

הכשרת שרת לטיפול בפנייה מסוג מסוים עולה כסף וזמן.
על מנת למצוא תשובה לשאלה זו ערכו וולאס ו-וויט סימולציה
במודל של מוקד שירות.

הנחות המודל:

- n סוגי שיחות המגיעות לפי תהליכי פאוסון בלתי תלויים המגיעים בקצב λ_k , $n \leq k \leq 1$
- ניתן להניח ששיחות מגיעות למוקד בקצב: $\lambda = \lambda_1 + \dots + \lambda_n$ כאשר ההסתברות ששיחה תהיה מסוג k היא:

$$P_k = \frac{\lambda_k}{\lambda}$$

- זמן שירות מתפלג $\exp(\mu)$ ללא קשר לסוג השיחה. (הנחה מאד חזקה)
- שיחות שנחסמות לא משפיעות על הגעות עתידיות
- אין נטישות – כל הפניות המגיעות למערכת ישורתו בסופו של דבר

מהלך הסימולציה

- $n = 6$

- משתנה ההחלטה m מייצג מספר יכולות לשרת (m זהה לכל

- השרתים) $1 \leq m \leq 6$

- $(C, K) = (90, 30)$

- זמני שירות בלתי תלויים $\sim \exp(\frac{1}{10})$ לכל סוג שיחה (דקות)

- 3 עומסים שונים: בינוני- $\rho=0.933$, כבד- $\rho=1$ וקל- $\rho=0.86$

- (QED, ED, ו-QD, בהתאמה)

- משום שישנם 90 שרתים ונניח שכל השיחות מגיעות באותו קצב, גודל כל קבוצה עבודה יהיה זהה ושווה ל 15.

- הקצאת היכולות המשניות תעשה באופן מאוזן בין השרתים

מהלך הסימולציה – הקצאת יכולות

משניות באופן מאוזן

- כאשר $m=2$, עבור כל קבוצת עבודה (המונה 15 שרתים) יהיו 3 שרתים עם אותה יכולת משנית (חלוקה שווה בתוך הקבוצה)
- כל קומבינציה אפשרית של 2 יכולות תופיע בדיוק 3 פעמים.
- אם לכל שרת יותר מ 2 יכולות, נקצה את שאר היכולות באופן הבא: בכל שלב נקצה את המספר העוקב הקרוב ביותר שעוד לא הוקצה. דוגמא:

$(5,3,0,0,0,0) \rightarrow (5,3,4,0,0,0) \rightarrow (5,3,4,6,0,0) \rightarrow$

$(5,3,4,6,1,0) \rightarrow (5,3,4,6,1,2)$

מהלך הסימולציה - המשך

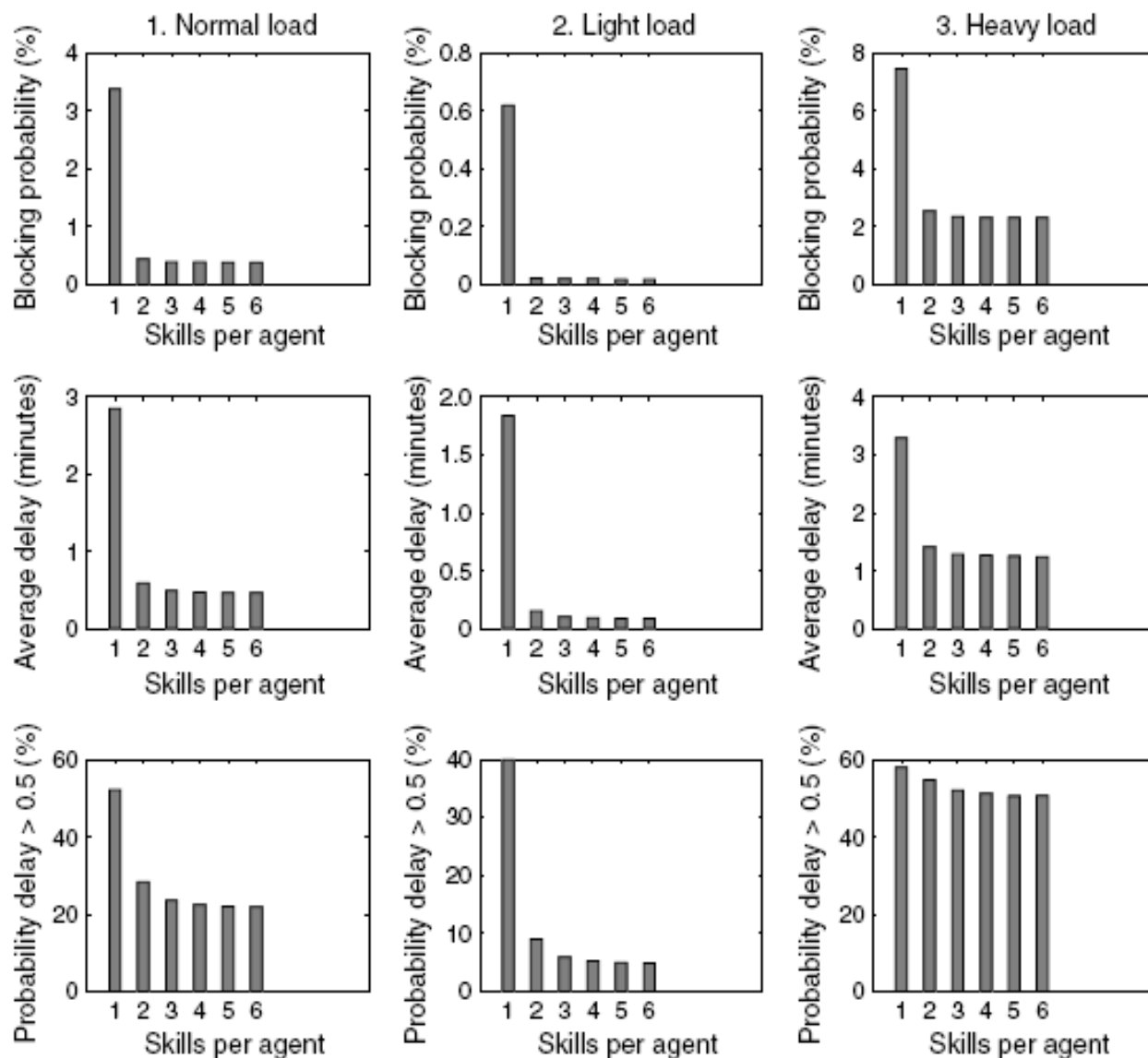
- הורצו 18 סימולציות שונות ($m = 1..6$, 3 עומסים שונים)
- בכל סימולציה בסביבות 800,000 פניות לאחר תקופת חימום של המכילה 20% - 24% מכלל הפניות.

נקודות חשובות

- כאשר $m = 6$ המערכת מתנהגת כמעט כמו תור M/M/90/30. (למה כמעט?)
- כאשר $m = 1$ המערכת לא מתנהגת בדיוק כ-6 מערכות נפרדות של M/M/15/5 בגלל "חדר ההמתנה" המשותף. נשים לב שכאשר חדר ההמתנה לא משותף זמן ההמתנה ישתפר אך הסיכוי להגיע למערכת תפוסה יעלה.
- נשתמש ב-M/M/90/30 כחסם עליון לביצועי המוקד וב-M/M/15/5 ו-M/M/15/30 כחסמים תחתונים.

תוצאות הסימולציה:

Figure 1 Typical Performance Measures as a Function of the Number of Skills per Agent and the Offered Load for an $M_6/M/90/30$ SBR Call Center with $1/\mu_1 = \dots = 1/\mu_6 = 10$ Minutes



Note. The specific performance measures are the blocking probability, the mean conditional delay given entry, and the conditional probability that the delay is greater than 0.5 given entry.

תוצאות הסימולציה - המשך

Table 1 **A Comparison of Simulation Estimates of Performance Measures for the SBR Model with Six Skills, Two Skills, and One Skill for Three Different Loadings: Light, Normal, and Heavy**

Loading	Model	Performance measure			
		Blocking	$E[W]$	$P(W \leq 0.5)$	Utilization
Light	<i>M/M/90/30</i>	0.00017	0.08	0.942	0.860
	Six skills	0.00018	0.09	0.951	0.860
	Two skills	0.00023	0.15	0.910	0.860
	One skill	0.00620	1.84	0.600	0.854
	<i>M/M/15/5</i>	0.03880	0.59	0.737	0.854
	<i>M/M/15/30</i>	0.00007	2.15	0.575	0.827
Normal	<i>M/M/90/30</i>	0.0036	0.45	0.733	0.930
	Six skills	0.0038	0.46	0.781	0.929
	Two skills	0.0044	0.59	0.716	0.928
	One skill	0.0336	2.85	0.478	0.897
	<i>M/M/15/5</i>	0.0650	0.82	0.641	0.927
	<i>M/M/15/30</i>	0.0066	4.94	0.344	0.872
Heavy	<i>M/M/90/30</i>	0.024	1.24	0.387	0.977
	Six skills	0.023	1.24	0.493	0.976
	Two skills	0.025	1.40	0.453	0.974
	One skill	0.075	3.29	0.419	0.918
	<i>M/M/15/5</i>	0.095	1.05	0.555	0.972
	<i>M/M/15/30</i>	0.028	8.97	0.153	0.905

Notes. Also included are the corresponding exact numerical results for the *M/M/90/30*, *M/M/15/5*, and *M/M/15/30* models. The waiting time is conditional upon entry.

מסקנות ביניים

- המוקד פועל בצורה כמעט זהה כאשר לכל מוקדן 2 יכולות או 3-6 יכולות! בכל מקרה ישנו שיפור ניכר מהמצב בו לכל מוקדן יכולת אחת בלבד.
- ככל שהעומס יותר כבד ההבדלים בביצועי המוקד בין 2 יכולות לשרת ו-6 יכולות לשרת יותר גדולים, אך עדיין אינם מהותיים.
- המעבר מיכולת אחת לשרת לשתי יכולות לשרת משפרת משמעותית את ביצועי המוקד. הוספת יכולות מעבר לכך אינה תורמת תרומה משמעותית לביצועים.

אלגוריתם לאיוש המוקד

- המטרה: מציאת צמד (C, K) ומטריצת השרתים במודל המוקד $M / M / C / K$ כך שהאילוצים על תוחלת זמן המתנה והסתברות להגעה למערכת מלאה ישמרו.
- לאלגוריתם 5 שלבים:
 1. מציאת צמד (C, K) התחלתי
 2. קביעה התחלתית של גודל קבוצות העבודה
 3. קביעת יכולות משניות (מילוי מטריצת השרתים)
 4. מציאת פתרון התחלתי פיזיבילי
 5. שיפור הפתרון הפיזיבילי

מציאת צמד (C,K) התחלתי

- נניח שלכל השרתים יש את כל היכולות.
- המודל מתנהג כמעט כמודל M/M/C/K, כאשר קצב ההגעה הוא: $\lambda = \lambda_1 + \dots + \lambda_n$
- במידה וישנם אילוצים שונים על סוגי שיחות שונות, נבחר את האילוצים המשוחררים ביותר כאילוצים על המערכת.
- דרכים למציאת C ו K אופטימאליים תחת אילוצים אלו פותחו ופורסמו על ידי Massy and Wallase (2005)
- נשים לב שהעלאת K אמנם תוריד את ההסתברות להגעה למערכת מלאה, אך תפגע בזמן ההמתנה הממוצע. לכן השינוי באלגוריתם יעשה על C שבעקבותו יעודכן K, כך שהסכום C+K יישאר קבוע.

קביעה התחלתית של גודל קבוצות העבודה

• המטרה: עבור כל k לקבוע את גודל קבוצת העבודה G_k

• שימוש בכלל השורש

• לכל C_k נקבע R_k כך ש: $R_k = \alpha_k + x\sqrt{\alpha_k}$ ו- $\alpha_k = \frac{\lambda_k}{\mu_k}$ עומס של פניות מסוג k .

• $x = \frac{(C - \alpha)}{\sum_{i=1}^n \sqrt{\alpha_i}}$, קל לראות ש $R_k > 0$ וכמו כן ש $R_1 + \dots + R_n = C$ * בתנאי

ש $C > \alpha$, מה שמתקיים תמיד אם ההסתברות להגיע למערכת

מלאה מספיק קטנה.

$$* R_k = \alpha_k + x\sqrt{\alpha_k} = \alpha_k + \frac{(C - \alpha)}{\sum_{i=1}^n \sqrt{\alpha_i}} \sqrt{\alpha_k}$$

$$\sum_{k=1}^n R_k = \sum_{k=1}^n \alpha_k + \frac{(C - \alpha)}{\sum_{i=1}^n \sqrt{\alpha_i}} \sum_{k=1}^n \sqrt{\alpha_k} = \frac{(C - \alpha)}{\alpha} \sum_{k=1}^n \alpha_k + \sum_{k=1}^n \alpha_k = (C - \alpha) + \alpha = C$$

קביעה התחלתית של גודל קבוצות העבודה - המשך

- נעגל כל R_k למספר שלם C_k תוך שמירה על כך ש

$$C_1 + \dots + C_n = C$$

- הדרך בה נעגל לא חשובה כרגע משום שבהמשך נשנה את גודל הקבוצות בהתאם לתוצאות הסימולציה.

קביעת מטריצת השרתים

- לפי המסקנות של הניסוי שבדק לכמה יכולות להכשיר כל שרת, ניתן לכל שרת יכולת משנית אחת.
- $C_{i,k}$ - מספר השרתים בקבוצת עבודה i עם יכולת משנית k
- $R_{i,k}$ - אמד התחלתי ל $C_{i,k}$ (לא בהכרח שלם) כך ש: $R_{i,k} = \frac{C_i C_k}{C - C_i}$
- $\sum_k R_{i,k} = C_i$ כך שמספר השרתים בקבוצת עבודה i עם יכולת משנית k יהיה פרופורציונאלי גם ל C_i וגם ל C_k
- נעגל כל $R_{i,k}$ למספר שלם $C_{i,k}$
- כאשר לכל שרת יש אכן רק 2 יכולות תהליך זה מגדיר מטריצת שרתים במלואה. במקרה ולכל שרת יותר מיכולת אחת נעזר בשיטה אותה הגדרנו בשקף " מהלך הסימולציה – הקצאת יכולות משניות באופן מאוזן"

מציאת פתרון פיזיבילי

מציאת הפתרון יעשה על ידי הגדלת C ו- K .

חשוב: קיים Trade Off בין הורדת ההסתברות הגעה למערכת מלא על ידי הגדלת K , והקטנת זמני ההמתנה.

לאחר בניית מטריצת השרתים יש בידינו פתרון מועמד אך לא בהכרח פיזיבילי. תזכורת: יש לעמוד בשתי קבוצות האילוצים:

1. הסתברות לזמן המתנה קטן מ- τ_k : $P(W_k \leq \tau_k | Q_k < C+K) \geq \delta_k$

2. הסתברות להגעה למערכת תפוסה: $P(Q_k = C+K) \leq \varepsilon_k$

כאשר δ_k ו- ε_k הם הסתברויות הסף

במידה והפתרון כעת איננו פיזיבילי:

1. צעד 1: הגדלת C

0 לכל קבוצה בה האילוץ מופר נחשב: $D_k = \delta_k - P(W_k \leq \tau_k | Q_k < C+K)$
(הסטייה מהאילוץ)

0 נגדיר: $k_1^* = \max\{D_k : 1 \leq k \leq n\}$ היכולת עם רמת הביצועים הנמוכה ביותר.

$k_2^* = \max\{D_k : 1 \leq k \leq n, k \neq k_1^*\}$ היכולת עם רמת הביצועים הנמוכה

ביותר לאחר k_1^*

מציאת פתרון פיזיבילי - המשך

- o נוסף שרת (שורה במטריצת השרתים) עם יכולת ראשית k_1^* ויכולת משנית k_2^* : $(k_1^*, k_2^*, 0, 0, 0, 0)$
 - o נחזור על אלגוריתם זה עד שלכל יכולת k יתקיים האילוץ על זמן ההמתנה.
 - o עבור כל איטרציה נוריד את K ב-1 כדי לשמור על $C+K$ קבוע.
2. צעד 2: הגדלת K
- o אם האילוץ על הסתברות להגעה למערכת מלאה אינו מתקיים, נעלה את K ב-1, אחרת סיימנו.
 - o שינוי K אמנם הוריד את ההסתברות להגעה למערכת מלאה, אך העלה את זמני ההמתנה, לכן אם אחד האילוצים לא מתקיים נחזור לצעד 1. נמשיך כך עד ששני האילוצים יתקיימו.
 - o האלגוריתם חייב להסתיים משום שבכל פעם שנחזור לצעד 1, נחזור עם $C+K$ גדול ב-1.

שיפור הפתרון הפיזיבילי

- לאחר מציאת פתרון פיזיבילי ננסה לשפר אותו תוך שימוש בשיפור שולי. גם כאן האלגוריתם יכלול 2 צעדים:

1. הסרה – נסיר שרת (שורה מהמטריצה). אם הפתרון עדיין פיזיבילי נחזור על צעד זה.

2. החלפה – אם הפתרון אחרי צעד 1 איננו פיזיבילי ננסה להחליף יכולות לשרת מסוים עד שנקבל פתרון פיזיבילי. אם קיבלנו פתרון פיזיבילי נחזור לצעד 1. אם לא נפסיק את החיפוש.

- חסם תחתון – נזכור שידוע לנו החסם התחתון על הזוג (C, K) , מפתרון מודל $M/M/C/K$

איך להסיר או להחליף שרת?

• הסרת שרת - נגדיר:

– $k_1^{**} = \min_k \{D_k : 1 \leq k \leq n, A_{i,1}=k, \text{ for some } i\}$ היכולת עם הביצועים הטובים ביותר מבחינת זמן המתנה.

– $i^* = \min_{A_{i,2}} \{D_{A_{i,2}} : 1 \leq i \leq C, A_{i,1}=k_1^{**}\}$ השרת (השורה) מבין כל השרתים להם יכולת ראשית k_1^{**} בעל יכולת משנית בעלת הביצועים הטובים ביותר.

– כעת נסיר את שורה i^* . אם לאחר ההסרה הפתרון עדיין פיזיבילי, נהפוך אותו לפתרון הטוב ביותר עד כה ונחזור על צעד 1.

– אם לאחר ההסרה האילוש על זמני ההמתנה לא מתקיים אך האילוש על ההסתברות למערכת מלאה כן מתקיים, ניתן לנסות להפוך את הפתרון לפיזיבילי על ידי הקטנת K . אם נסיון זה נכשל נעבור לצעד 2 (החלפה) בנסיון להפוך אותו לפיזיבילי

איך להסיר או להחליף שרת? (המשך)

- החלפת יכולות
 - נחליף את השרת בעל היכולות (A_{i*1}, A_{i*2}) בשרת בעל יכולות (k_1^*, k_2^*) – החלפת השרת עם צמד היכולות בעלות הביצועים הטובים ביותר בשרת עם צמד היכולות בעלות הביצוע הרע ביותר.
 - אם האילוצים תקפים, נכריז על פתרון זה כפתרון הפיזיבילי הטוב ביותר. אחרת, אם ניתן, ננסה להגדיל את K כך שהפתרון יהפוך לפיזיבילי.
 - אם הגדלת K גם כן לא מביאה את הפתרון לפיזיבילי, נחזור ל K ממנו התחלנו ונבצע החלפה נוספת.

איך להסיר או להחליף שרת? (המשך)

- משום שיתכן שאלגוריתם זה לעולם לא יסתיים, נגדיר מספר החלפות מקסימלי (X) . אם לאחר X החלפות לא מצאנו פתרון פיזיבילי נסיים את החיפוש.
- ברגע שפתרון פיזיבילי חדש נמצא הוא הופך לפתרון הטוב ביותר וחוזרים לצעד 1 (הסרת שרת).

נקודות חשובות

- פתרון אופטימאלי לא מובטח.
- ניתן לזהות שפתרון קרוב לאופטימאלי ב 2 דרכים:
 1. הזוג (C, K) קרוב מאד לזוג (C^*, K^*) שמצאנו על ידי פתרון המודל $M/M/C/K$.
 2. כל האילוצים מתקיימים עם Slack קטן.
- הוספת אילוץ על העלות הכוללת של המוקד עלולה להפוך את הבעיה ללא פיזיבילית.

דוגמא

- כמו בסימולציה הראשונה גם כאן יהיו 6 סוגי שיחות.
- העומסים של כל סוג שיחה הם:

$$\alpha_1=\alpha_2=4.25, \alpha_3=105, \alpha_4=1375, \alpha_5=1925, \alpha_6=, 305$$

- העומס הכולל הוא 82.5 דקות עבודה לדקה.
- אילוצי רמת שירות שונים לסוגי השיחות.
- לאחר פתרון M/M/C/K התקבל הזוג (90,21) כפתרון האופטימאלי.
- השרתים חולקו לקבוצות עבודה **שונות בגודלן** לפי ההסבר בשקופית "קביעה התחלתית של גודל קבוצות העבודה"
- בשקופיות הבאות מוצגות תוצאות הסימולציות השונות והתפתחות הפתרון

Table 5 Performance Measures in Seven Later Iterations of the SBR Heuristic Algorithm in the Case of Different Service-Level Targets and Unbalanced Offered Loads: $\rho_1 = \rho_2 = 4.25$, $\rho_3 = 10.50$, $\rho_4 = 13.75$, $\rho_5 = 19.25$, and $\rho_6 = 30.50$

SBR heuristic resource provisioning algorithm							
Performance measure	Number of iterations (Number of agents, extra waiting space)						
	8 (93, 18)	9 (93, 18)	10 (93, 18)	11 (93, 18)	12 (92, 19)	15 (93, 15)	16 (93, 14)
Blocking (%)	0.31	0.32	0.32	0.32	0.38	0.44	0.503
Mean delay (minutes)	0.20	0.20	0.20	0.21	0.26	0.19	0.18
Mean delay 1	0.40	0.45	0.34	0.38	0.44	0.34	0.34
Mean delay 2	0.34	0.35	0.33	0.36	0.40	0.31	0.32
Mean delay 3	0.13	0.12	0.12	0.11	0.14	0.10	0.09
Mean delay 4	0.28	0.28	0.29	0.25	0.31	0.24	0.23
Mean delay 5	0.12	0.11	0.10	0.11	0.12	0.09	0.09
Mean delay 6	0.18	0.19	0.22	0.24	0.32	0.22	0.20
$\mathcal{P}(\text{delay} \leq 0.5 \mid \text{entry})$ (%)	88.2	887.9	87.6	87.5	84.9	88.2	88.6
$\mathcal{P}(\text{delay}_1 \leq 0.5 \mid \text{entry})$	81.4	79.9	82.9	81.8	79.5	83.0	82.6
$\mathcal{P}(\text{delay}_2 \leq 1/3 \mid \text{entry})$	80.5	80.2	80.5	80.0	78.3	81.2	81.4
$\mathcal{P}(\text{delay}_3 \leq 1/3 \mid \text{entry})$	89.8	90.2	90.1	90.7	88.9	91.6	91.7
$\mathcal{P}(\text{delay}_4 \leq 1/3 \mid \text{entry})$	81.0	80.9	80.0	81.9	78.6	82.6	82.8
$\mathcal{P}(\text{delay}_5 \leq 1/3 \mid \text{entry})$	89.4	90.1	90.6	90.6	89.0	91.4	91.8
$\mathcal{P}(\text{delay}_6 \leq 0.5 \mid \text{entry})$	88.3	87.3	86.2	85.3	81.3	85.6	86.4
Average agent utilization (%)	88.4	88.4	88.4	88.4	89.3	88.3	88.2
Work group 1 utilization	80.8	81.3	79.8	79.9	81.2	79.6	79.8
Work group 2 utilization	80.8	81.0	80.9	80.9	82.4	80.8	80.6
Work group 3 utilization	85.4	85.4	85.7	85.9	87.1	85.7	85.6
Work group 4 utilization	88.9	89.0	89.3	88.5	89.5	88.6	88.5
Work group 5 utilization	88.6	88.2	88.3	88.5	89.3	88.3	88.2
Work group 6 utilization	92.6	92.8	93.3	93.6	94.3	93.5	93.5
Work group 1 primary utilization	51.4	52.5	48.2	48.9	48.6	48.6	49.3
Work group 2 primary utilization	50.8	51.5	50.8	51.1	51.0	51.0	50.8
Work group 3 primary utilization	58.7	58.0	58.0	57.7	56.8	57.8	57.6
Work group 4 primary utilization	71.4	71.3	71.1	69.1	68.6	69.3	68.9
Work group 5 primary utilization	69.1	67.6	66.9	67.1	66.0	66.7	66.6
Work group 6 primary utilization	83.1	84.1	85.4	86.4	87.4	86.4	86.1

Notes. The mean service times are $1/\mu_1 = \dots = 1/\mu_6 = 10.0$ minutes, and the target blocking is $\epsilon = 0.005$. The service-level targets for call types 1–6 are (80, 30), (80, 20), (90, 20), (80, 20), (90, 20), and (80, 30), respectively, where (80, 30) means $\mathcal{P}(\text{delay} \leq 0.5 \mid \text{entry}) \geq 0.80$.

נקודות פתוחות

- זמני שירות שונים – כאשר זמני השירות שונים לכל סוג שיחה (מתפלגים אקספוננציאלית עם תוחלות שונות) ותלויים גם בשרת וגם בלקוח, יש לצפות להתנהגות בעייתית של האלגוריתם (חלוקה לקבוצות, הודרות שרתים ושינוי יכולות)
- הכנסת סבלנות של לקוחות וזמני שירות לא אקספוננציאליים – מחקרים על מרכזי שירות הראו שהתפלגות סבלנות הלקוחות וזמני השירות בעולם האמיתי רחוקים מלהיות אקספוננציאליים לעיתים קרובות (Brown et al.(2005)). יש לחקור את ביצועי האלגוריתם למקרים אלו.
- חשוב לשים לב שבשלב הראשון של האלגוריתם ניתן להחליף את מודל M/M/C/K במודל M/GI/C/K+GI אם זמני השירות והסבלות הם בלתי תלויים.

נקודות פתוחות - המשך

- חלוקה לקבוצות עבודה – רב הדרכים בהם נחלק את העובדים לקבוצות עבודה יתנו תוצאות זהות. ישנן חלוקות מסוימות שיכולות להשפיע מאד על ביצועי המוקד. לדוגמא: עבור כל $1 \leq i \leq 3$ ניתן לכל עובד בקבוצת עבודה $2i-1$ יכולת משנית $2i$ וההפך. במקרה זה המוקד יתפקד כשלושה מוקדי שירות נפרדים עם 30 שרתים בכל אחד וביצועיו יפגעו מאד.
- מספר יכולות שונות לכל מוקדן – יש לחקור כיצד ישתנה המודל והאלגוריתם כאשר לכל מוקדן מספר יכולות שונות. מחקרים הראו שישנו יתרון להשאיר את השרתים היותר גמישים (בעלי יותר יכולות) פנויים כמה שאפשר.

Value Based Routing

הקדמה

- מוקדי שירות טלפוניים מטפלים במספר סוגי שיחות
- מוקדי השירות משתמשים ב SBR על מנת לנתב את השיחות למוקדנים
- המדדים לפיהם שופטים את ביצועי המוקד הם בד"כ אחוז נחסמים ואחוז הפניות שהמתינו יותר ממשך זמן מסויים (y_j)
- אין התייחסות למיקסום התועלת למוקד

במה עוסק Value Based Routing?

- VBR הינו אלגוריתם שיבוץ המבוסס על אלגוריתם ה SBR הנותן דגש למיקסום התועלת של המוקד
- תועלת יכולה להימדד רווח הנובע מן השיחות בארגונים למטרות רווח, או במספר פניות שטופלו כראוי בארגונים שלא למטרות רווח

הנחות וסימונים

- Erlang-A == <= מודל עם נטישות
- m – סוגי שיחות
- n – מספר שרתים
- $v_{i,j}$ - התועלת מטיפול בשיחה j ע"י שרת i
- מספר השרתים נשאר קבוע באינטרוולים של חצי שעה

בניית המודל

• $C_{L,j}$ - קנס על נטישה של לקוח j

• $C_{D,j}$ - קנס על כל לקוח שחיכה יותר מ y_j שניות בתור

• $N_{i,j}$ - מספר שיחות מסוג j שנענות ע"י סוכן i

• L_j - מספר שיחות מסוג j שהלכו לאיבוד עקב נטישות

בניית המודל - המשך

- D_j - מספר שיחות שנענו מסוג j וחיכו בתור יותר מ y_j שניות

- V - התועלת הכוללת של המוקד

- $N_{i,j}, L_j, D_j$ הינם מ"מ התלויים בהתפלגות קצב ההגעה (לא בהכרח פואסונית!) ובמדיניות הניתוב

פונקציית המטרה

- המטרה היא להביא למקסימום את התועלת הכללית - V

$$V \equiv \sum_{i=1}^n \sum_{j=1}^m (E[N_{i,j}] \cdot v_{i,j}) - \sum_{j=1}^m ([E[L_j] \cdot c_{L,j}] + [E[D_j] \cdot c_{D,j}]) .$$

אילוצים

$$\frac{E[L_j]}{E[N_j]} \leq a_j$$

• אילוץ על נטישות:

$$\frac{E[D_j]}{E[N_j]} \leq (1-x_j)$$

• אילוץ על זמני ההמתנה:

- איננו מנסים להגיע לפיתרון אופטימלי, אנו מחפשים ע"י סימולציות ערך גבוהה ל V תוך כדי עמידה באילוצים

תזכורת - מטריצת שרתים

$$A_{6,4} = \begin{pmatrix} 3 & 4 & 1 & 0 \\ 1 & 4 & 0 & 0 \\ 2 & 3 & 0 & 0 \\ 4 & 0 & 0 & 0 \\ 3 & 1 & 2 & 4 \\ 1 & 0 & 4 & 0 \end{pmatrix}$$

- במטריצה מיוצג מוקד עם 6 שרתים ו 4 סוגי שיחות

- $A_{i,j}$ - היכולת של מוקדן i לטפל בשיחה מסוג j

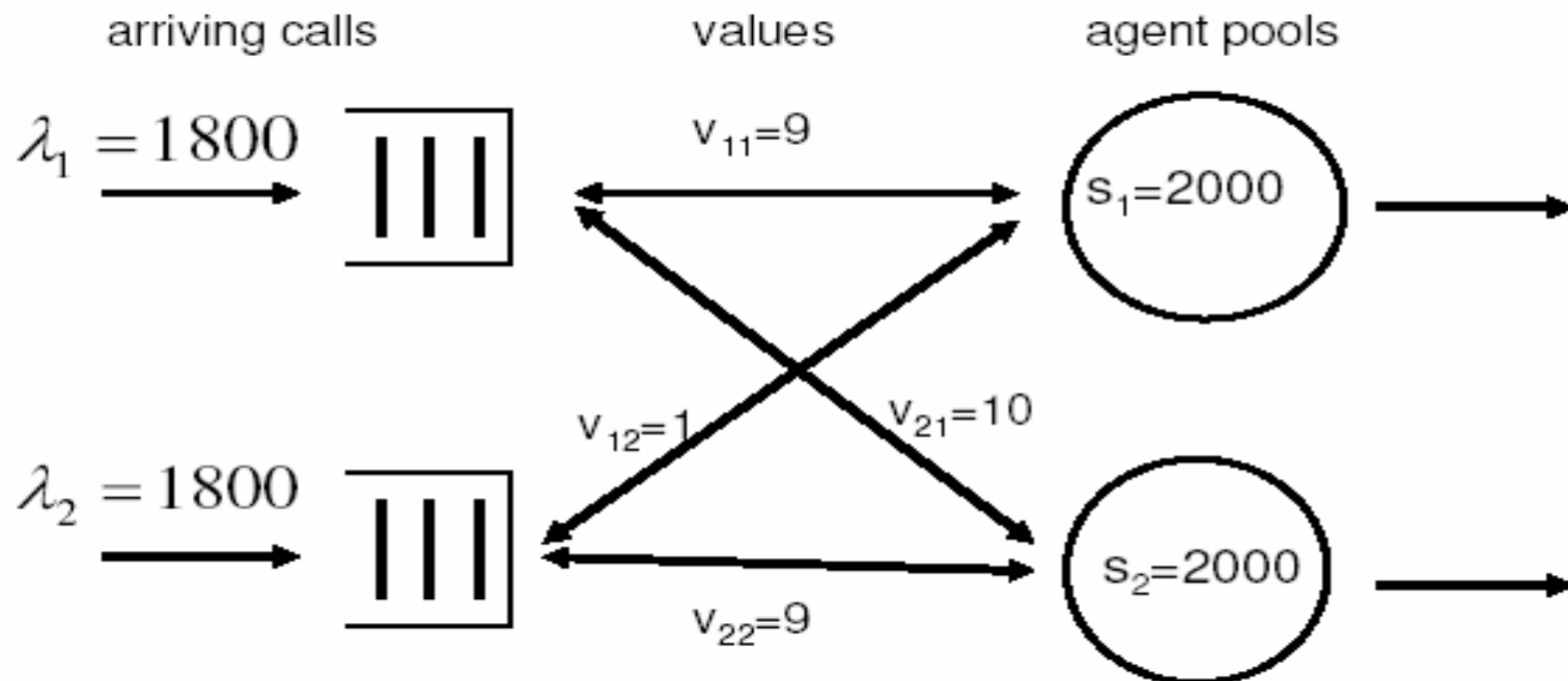
בניית הסימולציה

- ניקח 2 מוקדי שירות טלפוניים ו 2 סוגי שיחות
- גודל כל מוקד – 2000 מוקדנים
- משך השירות מתפלג אקספוננציאלי עם תוחלת של 10 דקות
- נמדוד את משך השירות במונחים של זמן שירות, ולכן ההתפלגות תהיה אקספוננציאלית עם תוחלת 1 (קונסיסטנטיות עם העומס)

בניית הסימולציה

- M/M/M+s
- קצבי הגעה – [שיחות למשך שירות ממוצע] $\lambda_1 = \lambda_2 = 1800$
- עומסים - $R_i = \lambda_i = 1800$
- עומס כולל – $R = 3600$
- זמני הנטישות מתפלגים אקס' עם פרמטר 1
(ביחידות של זמן שירות)
- מכיוון שקצב השירות שווה לקצב הנטישה המודל מתנהג בעצם כמו מודל M/M/ ∞

תרשים סכמטי



$$v = \begin{pmatrix} 9 & 1 \\ 10 & 9 \end{pmatrix}.$$

Direct VBR

- ב DIRECT VBR מטריצת התועלות זהה למטריצת השרתים
- המוקד השני מטפל בשני סוגי השיחות טוב יותר מהמוקד הראשון
- הנצילות של המוקד השני תהיה קרובה ל 100%, ובמוקד הראשון יהיו 300-500 מוקדנים "מובטלים" בכל רגע נתון.
- מכיוון שקצב ההגעה של שיחות מסוג 1 זהה לקצב ההגעה של שיחות מסוג 2, לפי PASTA ניתן לומר ששתי סוגי השיחות ישתמשו במספר שווה של מוקדנים מהמוקד השני

Direct VBR

• $\lambda_{i,j}^D$ - מספר שיחות מסוג j שמקבל סוכן ממוקד i

$$\lambda^D = \begin{pmatrix} 800 & 800 \\ 1000 & 1000 \end{pmatrix}$$

$$V^D = \sum_{i=1}^2 \sum_{j=1}^2 \lambda_{i,j}^D v_{i,j} = 27,000$$

Indirect VBR

- ניתן לקבוע שמוקד מס' אחד יטפל רק בשיחה מסוג 1, ומוקד מס' 2 יטפל רק בשיחות מסוג 2

$$\lambda^I = \begin{pmatrix} 1800 & 0 \\ 0 & 1800 \end{pmatrix}$$

$$V^I = \sum_{i=1}^2 \sum_{j=1}^2 \lambda_{i,j}^I v_{i,j} = 32,400$$

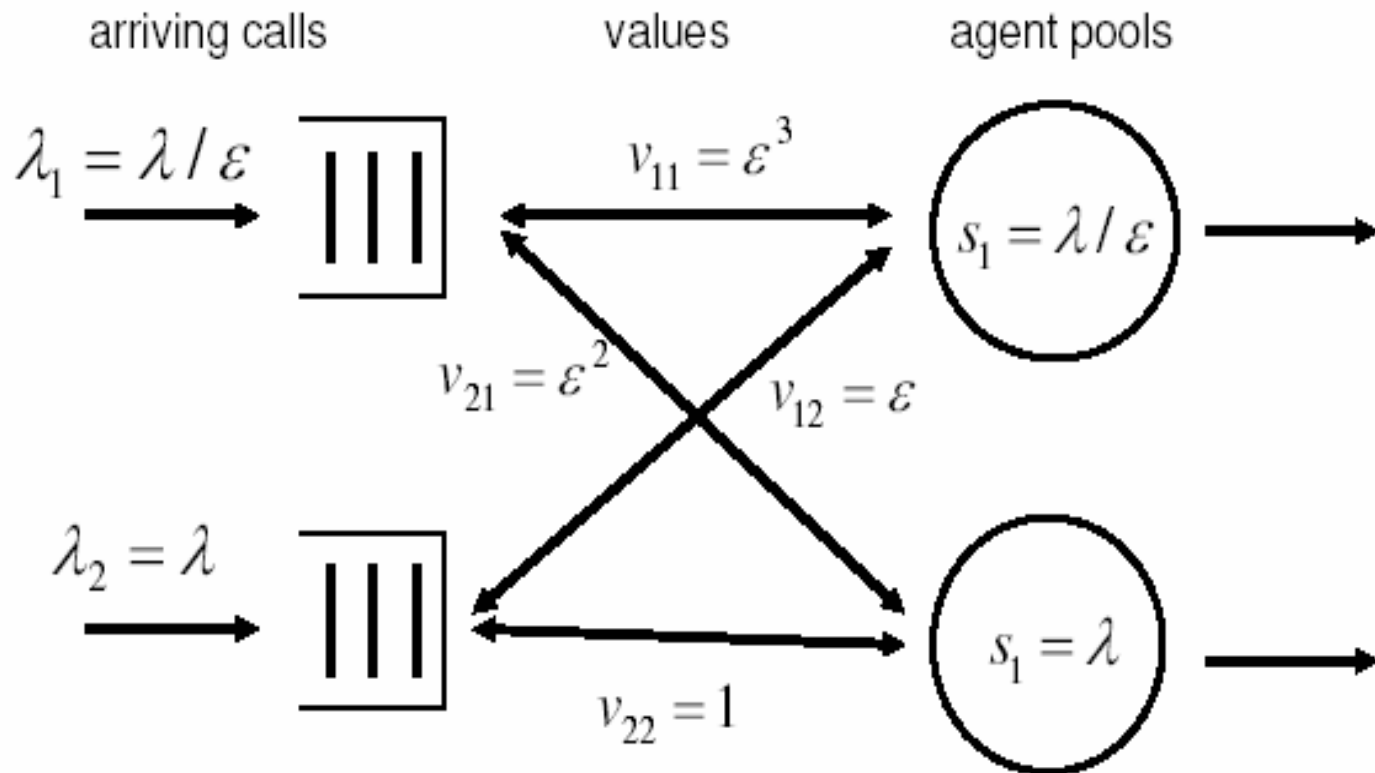
Optimal solution

- מכיוון שמוקד 2 מטפל בכל סוגי השיחות יותר טוב ממוקד 1, נרצה לנצלו כמה שיותר

$$\lambda^{opt} = \begin{pmatrix} 1600 & 0 \\ 200 & 1800 \end{pmatrix}$$

$$V^{opt} = 32,600$$

Indirect VBR vs Direct VBR



Indirect VBR מול Direct VBR

$$v = \begin{pmatrix} \epsilon^2 & \epsilon^3 \\ \epsilon & 1 \end{pmatrix}$$

$$\lambda^D = \begin{pmatrix} \frac{\lambda}{\epsilon(1+\epsilon)} & \frac{\lambda}{1+\epsilon} \\ \frac{\lambda}{1+\epsilon} & \frac{\lambda\epsilon}{1+\epsilon} \end{pmatrix}$$

$$\lambda^I = \begin{pmatrix} \frac{\lambda}{\epsilon} & 0 \\ 0 & \lambda \end{pmatrix}$$

$$V^D = \epsilon(3 + \epsilon^2)/(1 + \epsilon)$$

$$V^I = \lambda(1 + \epsilon)$$

$$\frac{V^I}{V^D} \text{ שואף לאינסוף כאשר } \mathcal{E} \text{ שואף ל } 0$$

INDIRECT VBR

- עבור כל סוג שיחה j קיימת קבוצת מוקדנים שבמטריצת המוקדנים השיחה נמצאת בעדיפות הראשונה שלהם
- למוקדנים ישנה גם יכולת משנית
- ראינו כי אין טעם שלמוקדן יהיו יותר משתי יכולות (מתוך SBR)

קביעת התועלות

- מטריצת התועלות נתונה
- התועלת מן הטיפול של מוקדן i בשיחה מסוג j
ידועה לכל $1 \leq i \leq n$ ולכל $1 \leq j \leq m$
- כעת צריך לקבוע את $V_{i,j,k}$ - התועלת המתקבלת
ממוקדן i בעל עדיפות ראשית j ועדיפות משנית k
($1 \leq k \leq m$)

קביעת התועלת

$$v_{i,j,k} = (1 - p_{i,j,k})v_{i,j} + p_{i,j,k}v_{i,k}$$

- $p_{i,j,k}$ - פרופורציית השיחות מסוג k מכלל השיחות שבהן מטפל מוקדן i .

$$0 \leq p_{i,j,k} \leq 1$$

קביעת המיומנויות - סימונים

- $n_{j,k}$ - סך המוקדנים עם מיומנות ראשית j ומיומנות משנית k

- $$\begin{cases} x_{i,j,k} = 1 & \text{מוקדן } i \text{ מקבל עדיפויות } (j,k) \\ x_{i,j,k} = 0 & \text{אחרת} \end{cases}$$

קביעת המיומנויות – בעיית האופטימיזציה

$$\text{Maximize Total Value} = \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1, k \neq j}^m v_{i,j,k} x_{i,j,k}$$

$$\sum_{j=1}^m \sum_{k=1, k \neq j}^m x_{i,j,k} = 1 \quad \text{for all } i$$

$$\sum_{i=1}^n x_{i,j,k} = n_{j,k} \quad \text{for all } (j,k) \text{ with } j \neq k .$$

$$x_{i,j,k} \geq 0 \quad \text{for all } i, j \text{ and } k .$$

הרחבה

- לעיתים נרצה להעסיק במוקד יותר מוקדנים מהדרוש על מנת שיעסקו בפעולות "לא יצרניות" כמו למידה וניירת, ובמקרה הצורך יעזרו למוקדנים הפעילים
- לשם כך נקבע גבול תחתון ועליון למספר המוקדנים מכל קבוצה

סימונים

- n - מספר המוקדנים במוקד
- n' - מספר המוקדנים שעסוקים בטיפול בשיחות
- n_j^L, n_j^U - גבולות מספר המוקדנים ששיחות מסוג j נמצאים בעדיפות הראשית שלהם
- $n_{j,k}^L, n_{j,k}^U$ - גבולות מספר המוקדנים ששיחות מסוג j נמצאים בעדיפות הראשית שלהם, ושיחות מסוג k בעדיפות המשנית שלהם

בעיית האופטימיזציה

$$\text{Maximize Total Value} = \sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^m v_{i,j,k} x_{i,j,k}$$

$$\sum_{i=1}^n \sum_{j=1}^m \sum_{k=1}^m x_{i,j,k} \leq n' \leq n$$

$$\sum_{j=1}^m \sum_{k=1}^m x_{i,j,k} \leq 1 \quad \text{for all } i$$

$$n_j^L \leq \sum_{i=1}^n \sum_{k=1}^m x_{i,j,k} \leq n_j^U \quad \text{for all } j$$

$$n_{j,k}^L \leq \sum_{i=1}^n x_{i,j,k} \leq n_{j,k}^U \quad \text{for all } j \text{ and } k$$

תמונת מצב של המוקדים כיום

- אינם עומדים ביעדים התפעוליים

- אין איזון עומס בין המוקדנים

- הוצאות גבוהות מן המתוכנן

- הכנסות נמוכות מן המתוכנן

מצב המוקדנים

- קצב תחלופה גבוה – יכולה להגיע עד ל - 200% בשנה
- יותר העדרויות מהעבודה ממה שמתוכנן
- הרבה חריגות מסידורי העבודה
- עייפות מוגברת לקראת סוף יום העבודה

מצב המוקדנים

- המסקנה – המוקדנים אינם מרוצים
- פתרון אפשרי – שיתוף המוקדנים בקביעת סוגי השיחות שעליהן המוקדנים יענו

מודל פשוט

• $v_{i,j}^m$ - ערך התועלת שנקבע ע"י ההנהלה

• $v_{i,j}^a$ - ערך התועלת שנקבע ע"י המוקדן

• $v_{i,j}^c$ - ערך התועלת המשוכלל

$$v_{i,j}^c = v_{i,j}^a * v_{i,j}^m \quad \forall i, j$$

מודל עם משקולות

• $w_{i,j}^m$ - משקל החלטת ההנהלה

• $w_{i,j}^a$ - משקל החלטת המוקדן

• $w_{i,j}^c$ - משקל הערך המשוקלל

$$v_{i,j}^w = w_{i,j}^m * v_{i,j}^m + w_{i,j}^a * v_{i,j}^a + w_{i,j}^c * v_{i,j}^c$$

$$w_{i,j}^m \geq 0, w_{i,j}^a \geq 0, w_{i,j}^c \geq 0$$

$$w_{i,j}^m + w_{i,j}^a + w_{i,j}^c = 1$$