

Multi-Server Queues

“Our” model of a service station	M / M / m / B	+M: Birth & Death; 4CallCenters
	M / GI / m	+GI: Research Challenge (Whitt, Dec. 2003)
e.g.	M / M / m	+GI: Current Ph.D. (Zeltyn)
Parameters:	Markovian	$\lambda \quad \mu \quad \theta \Rightarrow$ practical (Palm, Garnett)
	General	$C_a^2 \quad C_s^2 \quad ? \Rightarrow$ Efficiency-Driven (Kingman)
		(heavy-tails) current, in telecommunication

G/G/m Stability $\Leftrightarrow \rho = \frac{\lambda}{m\mu} < 1$ (fluid-logic, but subtle); ρ utilization factor, via Little.

[With Finite Patience: Always Stable (via Abandonment).]

GI/GI/m: Approximate Analysis of Exact Model (Efficiency-Driven)

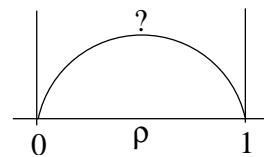
[M/M/m, M/M/m+GI: Either Exact Analysis of Approximate Model, or Asymptotics, with Many Servers (QED Call Centers)]

Natural *extensions*: heterogeneous servers (there exists some theory; networks)
 heterogeneous customers (important - CRM)
 heterogeneous both (important - SBR)

Importance:

Phenomena: Servers “help” each other (Pooling)
 Few fast vs. many slow?
 Economies of Scale (EOS) - Stochastic

Kleinrock’s Cycle
 Tools: Staffing (offline)
 Congestion Curves



In this teaching note: Focus on GI/GI/m (E-Driven Approximations, for practical staffing) and M/M/m (Exact Analysis, to demonstrate Phenomena)

Reducing Delay Through Changes in the Service Process

Hall, Chapter 7 (pg. 208–269); see also Chapter 5.

Types of queues:	Perpetual	(All customers Always wait)
	Predictable	(Queueing at known times)
	Stochastic	(Queueing at random)

Typically, eliminating a perpetual queue exposes predictable queues, and eliminating a predictable queue exposes stochastic queues, which is our focus here.

Managing Stochastic Q's:	λ, C_a^2	arrivals \leftarrow Chapter 8: how?
	μ, C_s^2	services \leftarrow Chapter 7, here: how?
	m, b	resources & facility.

Table 7.1, page 213: Ways to reduce service time (increase service rate).
 E.g., Team service (idle \Rightarrow help out, as in a garage).
 Automate, standardize,...

Add servers: Staffing: who, when and, for how long, how many?
 \downarrow \downarrow \downarrow
 Tough work shift HW, based on lectures

Inspirational useful reading: Case study on pg. 257–266:

Buffa, E.S., M.J. Cosgrove and B.J. Luce. “An integrated work shift scheduling system”, from Decision Sciences, 7, 1976, pages 620630.

First systematic hierarchical staffing of a telephone exchange (call center), which is still very useful as it describes current practice:

- **Forecasting:** Forecast load, namely $\lambda(t)$, $0 \leq t \leq T = 1$ day, via Time-Series Analysis;
- **Staffing:** Determine (desired) number of agents, during say each 1/2 hour, namely m of M/M/m *, based on MOP's in steady state.
- **Shift Scheduling:** Determine shifts (timing, duration, structure) via Optimization (LP/IP, as in HW)
- **Rostering:** Assign “servers” to shifts (heuristics, AI)

*Erlang-C dominates practice. We are capable of using Erlang-A, and sometimes more.

הה שמי קומן אבד

זו השיטה, והיא עובדת תמיד: מתי שלא תבוא, תomid יהיה לך, כי מישחו כבר יזאג לך. איך הוא יעשה את זה? הוא פשוט יפעיל רק קופת אחת מבודק השש הקיימות.

Hall, Chapter 5, Section 5.7

Ancillary Activities

(Implicit:
servers' flexibility)

◇ Add a server whenever

$$\frac{\# \text{Queued Customers}}{\# \text{Servers}} > K$$

(K = decision variable).

- ◇ Remove to ancillary activity when queue dissipates

eg. Stock Shelves vs.
Checkout (Cashier).

- Manage fairly customers' new queues;
- Manage sensibly servers' interruptions.
- $K \approx 3$
works well in practice.
(See Hall's Figures.)

eg. Telephone and Emails.
(inflicting mental setup)

eg. Dynamic Staffing
at BK, Bank, CC

נראה את עניין הסופרמרקט ריק כווגונט. כמשמעותו שיכל ליזיג בשבלוט את היחסים הדרי עכורים, לפחות מ- 20 שנותיו. בין זו שהשלמת את הרכמי ולכון יכול לשירות הולם, לבין זו שמקבל את הכסף ולהן מושביה לשירות גולם. נניח שאנו נורו מוחליטים טמפלרים. יש לנו יומם עמוס במלחמות, ושהעומס זהה עומד. בינו גור משוער לriskנות ש- 20 שנות. הרמסקה היא, שחייבים לילכט סופרמרקט לעשרות קילומטרים, ועוד יותר יתיר דמיים לישות ואזורי כביש מודר שאותו שידועים שפיעין כי יוק באנור הולם. בשבע בוכרי כרמי, שעת הפתוחות, מזמינים ליר ולחנות הופר, מוזהרים ש- 20 המקרים דרכן יוצאים, וושטפים במוחירות מה שצרכין, רבע שעה אורה. זה מתיידגים בטור ל��וף, ואז באנגר שולחן בכל תרואה טענת. אך זה יכול ללהות שבסופרmarket אין כמעט אגשין, ובכל זאת ליד הפה יש חור מתחבר.

מג' קאצ' צ'וּרַ מִעֵמֶק מִסְבֵּרַ אֶת הַעֲבֵרַן, אֲםַכְיַיְלַ אַלְעַזְעַמְתַּן כְּבֵדַה זֶה הַשְׁתָּוֹן, וְהַיאָעֲוֹבָתַה בְּבֵדַה: בְּבֵדַה שְׁלָאַ תְּבָאַ, תְּבָאַ תְּהַרְבֵּשׁ, כִּי מִשְׁׁוֹרַ בְּכַדְרָמְתִּיקַ כְּבֵדַה יְהִיָּגֵל לְכָךְ. אֲנַן זֶה אָעַשְׂה אֶת זֶה וְזֶה אָפָוֶשׁ יְפִילֵלַ קְרַפְתָּה אֶת תְּמִזְוִינָה שֶׁהַקְּיִמְתָּן. וְזֶה אָעַשְׂה אֶת זֶה, כִּי מִה שְׁוֹרְבָּבַלְתַּן הַלְּאַ לְהַלְתַּן לְנוּ שְׁוֹרְתָּתָן. טָבַל אַלְאַ לְחַסְדַּן לְעַצְמַתְךָ הַאֲגַם, וְלִכְמַן עֲדֵי שְׁאַנְטוֹן גַּעַמְרוֹ רַבְעַ שְׁעָה בְּחָנָדַכְיָה שְׁדָאַ שְׁלָמַן עַד כְּמָה קְלִקְטָאַת נְסָפָתָן, אַזְלִי אַפְּלִי לְבַדְבַּרְתָּה.

זה לא רק בוטה

ההתקשרות על חזיאר האזורייאלי של חלץ העזיריים הבנויות בגדלוות לא-נעים כדי לשפר שטח את האזור. זה עניין כמעט אבד. והוא נועד כדי להציג שבעיה של מטבח אקלטני, וארכיטקטנים אוחזים בדים למילויו של מבנה מושך בין מטבחים אוכלי קר-בשאולן ציר לחיותם, ואנרכיסטיים ציר לחיותם קר – והתקתק. כמה פעמים כתם באזורי ארכז'ה במאסז'ו, מדרסים לבניין דמויים שאותם נזקקם להחלה יותר מ晦י' שעבד בנה במנזר, בחרהותם למלט שטם לא-וושינגטוניק הילטם איז' מבריסק – בלב שטח ארכז'ה פעים מלאה היה מושגה אפרה. השיריה באנז'ו והקברן.

שוכת בפיניקיא בפיות כהו וארם שלז' וככשה קעה לאלתגרון; ואנרכו מושלים בהיריה על שכם; ואנדים לו כבשנש הכל ברוך, וממשיכים לחותן. י. עמיים נא茫然: תנאים אסוציאטיביים ומטחזקנטז' מלבד, כמו שאמרתו ל' גודה, כשים'ה שעה רורה בחירופת מס'ם, למילדי שהבאי לו המגנה הפהה, לגותם'ה יונק'ה מודידות, וגוחור ליגע זמ'יקוסטס בדורותם מהיר, בדורותם מהיר, וגעותם בשענ'ה ליקוט, שנותם את זבב'תם שליכם לעל נועז, ורטובות המלולך בבדצ'ים נודדים, ואטם בער'ה, ששות' ניק' עבורהם את זרכ'ם ההודו'תוב הו', ז'בו רגע, אומת' הקופאי'ת זדר'ה להע'ז וה'יכ'ת נ'ז'ו'ן נ'ז'ו'ן, הדיא מוח'ר'ה את סגנ'ת'ה ג'ב'ל, שמחפה'ת את ג'ב'ל, וב'לום זיד' מוח'ר'ים נ'ז'ו'ן. אפס'ל להמש'ג'ן נ'ז'ו'ן. מט'ב'ת' שע'ל אוד' מוח'ר'ים. שנותם'ה עבורה' אן לא עדר' ולא מהדר. מיד' של'ות' את'ם הקופאי'ת לבד' את' המהדר. אם' אטם' כב'ר' משל'ם'ה ג'ו' כט'י. כרא' ש'ג'ב'ן' קפ'צ'ת עבדו'. לא בז'וק' בחר' ל'ט' איד' ל'כדר' את' המהדר, אבל לא נ'גרא. הדיא תס'ר' ל'ט' מה' אטם' ז'ר'יכם' לעש'ב, כי'

שבועה שוכנה ומצלצת מטופלים כל התהילה, והשכון מוגש להם, ואתמים מהתפקידים לשלים במומנו: קורתה. כדי לסת לכם עוזר של אדריכלים שטי', אבל בקופה אין עוזר. למה שבמקרים שבעוד, בין השאר, לחתת לכם עוזר, יהיה סוף ואם לא דיה, אנהנו נבשיק לכוא לשכם אונטו נאבר לימי שציד און מה שאנטו וושכרים על השירית שזרה נוון ל'ז' ז'דים רגע, אומתת הקפאיית. אני ומלכת פטרות 100 ש"ר. שוכן מוחשיים את הסוגיות, שמחפש את דבנהל, וכטב' יוד בחפשם מטבחו של עוזה ש"ר. עכשוי נחבר רך לעובו שאוזיא על עוזה לאוטר, כדי שייסיע לאבב' און הצלחות, אם הצלחות לא ליעשות משלות, ולכן אין טעם להתייאש. אן מהם העובו' הות ישב במחוץ ובז'וד, למלות שואה זיך לשכט ליר' ז'קופות' ז'טמד'ה כהה', מסכרים לבן, ואונטו, נחומרם שכםון, אם ה'תמי' כהה' - כי

לא התחשקל לזרם

בשבוע שUPER נזכר נסoston לוחות גדרה מואן של טקסטיל מטל-אביב ערכו שם קנית ניכרת, בכמה מאות שקלין, מונזרות, לי' בחד גדרה ליד אחת הקופות, וביציטי לטוב עשר רകה אוח' שגענותי שם והודעה לי' והקופות עכל' לחשך לא' קופה אהורה, כי היא יוצאת להפקה, "מה את זהב", שאמי לא אצא להפקה כבר מזמן הגיע לי' לא, אך לא מזדהה שזיה לא' תא' להפקה, אג' רך רוזה, שכשיה יוצאת תחיה יוצאת תחוליה אהורה, או אני רוזה גרדן את מגלה הגורשה לקופה טמונה, ושם כבר דודת הקופאות ית' נזמרה, שטבוח השדי' והזיה לא' מקלחת ית' ג'ר' כב' מזרת' את העלה, שוב עמרותי, עשר דקוט בטור, וא', כהונת' את הדרים, השוכן הדרים הדועה לי' שהיא לא' מקלחת ית' ג'ר' כב' בסומן, קיבלתי גוזאות מזרות איך לברוך את המהו, למי לגשת, מה' לשאל, ואיזו תשופע נקלבל, הרוחה שפריט איך לא' שיכנה לבצע את העברה. תזרמה גדרה אהזה את זו שמלו'. צרךן שלא מוכן לחשם לזרהראת. איפה נשמעו...

“או אל תחקי את הפהיות הזה”, הצעיטה קבאיית. גם לך סדרתני. בשבילו נגנשתי בכל לחנות. מה עוזבטי? בלבונתיהם והור מאהורי מתאריך ואין מוצא. שאללה הקפואת את והולגאות שלל, לא דעת. בסופו של דבר, מפאת הדברים כל המוחלים מאזור, ומפני שאנו צרכנית יcordalit טיפושית, ויתרתו על הפורר, ובמיוחד ויתרתו על הנטזון שאשוו במעלה כל זרנחת נוכבתת תמורה בעקבות הנקה של. שירות כוה זהה אנטיגטומה. ולא הלאינו להתגלה עלי הנקפהית, וא Tatkeshtiy על הוכחתה של. שורות לא נקמתי נסבכהה דיהירה העומחת לרשויות: לעולמים לא לחייכנס למקומם הוה נשית, גם? הוזהר את כל תבר. ומה לא פשיתי את זה כזיהה שרב, לא זהה כי כה לעמוד על שלוי, ולא התחשך לי לדרכו.

3N8N-50-581
19/5/96 2:12N

Ancillary Activities / Dynamic Staffing

(Trading-off Customers' Waiting-time vs. Servers' Interruption)

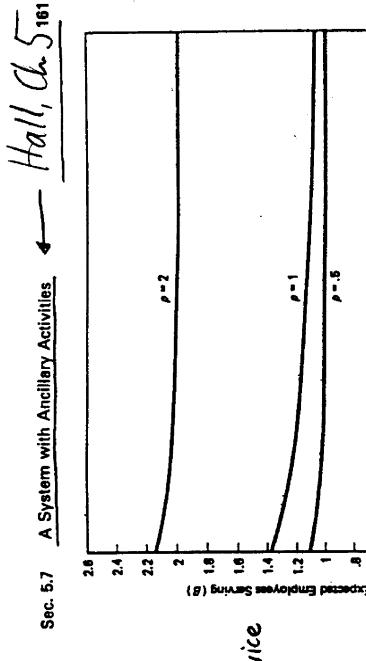


Figure 5.12 The expected number of employees serving declines as the maximum queue size per server increases. For $p > 1$, the limit is p . For $p \leq 1$, the limit is one.

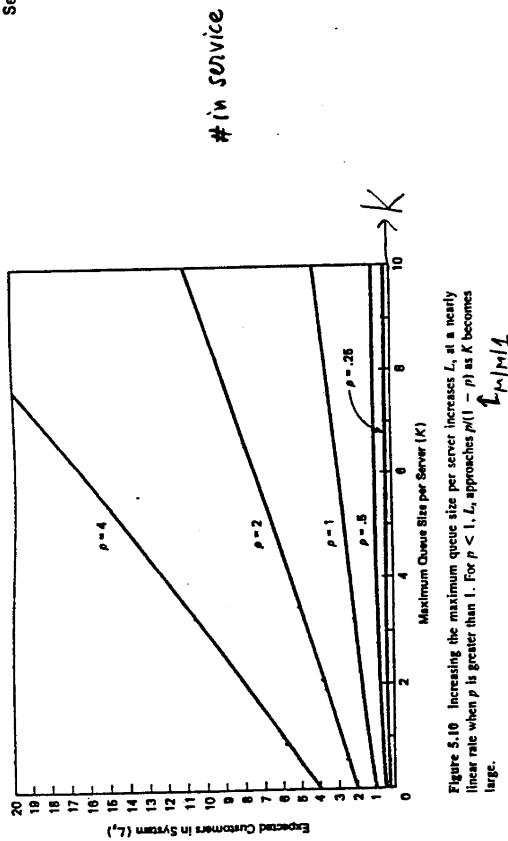


Figure 5.10 Increasing the maximum queue size per server increases L_s , at a nearly linear rate when p is greater than 1. For $p < 1$, L_s approaches $p/(1-p)$ as K becomes large.

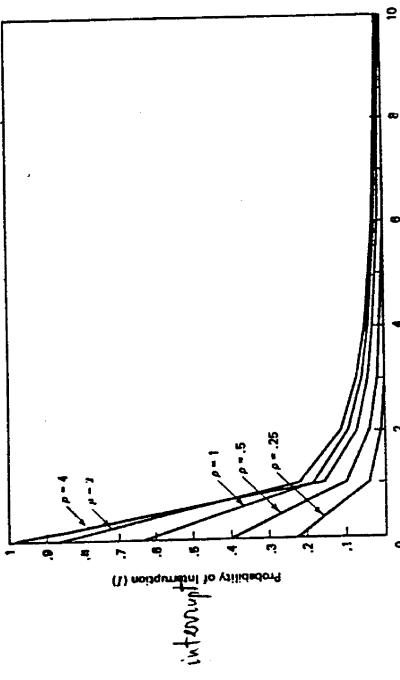


Figure 5.11 The probability of an interruption declines toward zero as the maximum queue size per server increases. For $K \geq 3$, the probability of interruption does not exceed .1.

Add servers
when the number of customers in queue per server exceeds three should provide good results in most instances.

By comparison to the $M/M/m/oo$ queue, the same number of "busy" servers provides far less delay. But, perhaps more importantly, the ancillary policy is far more robust with respect to changes in the arrival rate. With fixed servers, even small changes in p can produce enormous increases in delay. This is not true with ancillary servers, for the time in queue never exceeds K/p . In fact, for large values of p , the expected time in queue is approximately:

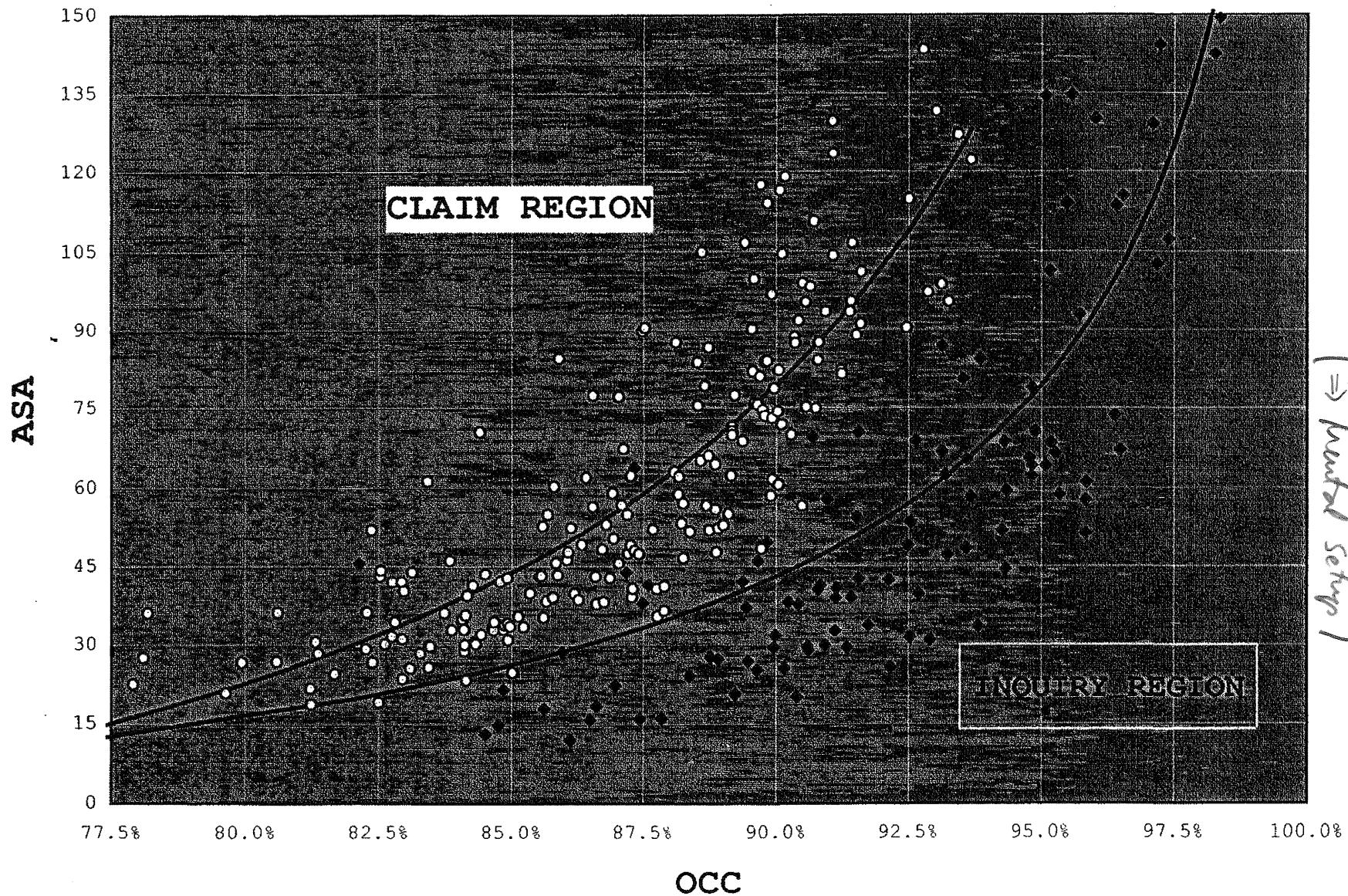
$$W_q \approx \frac{K}{2p} \quad (p \geq 2) \quad (5.81)$$

or just one-half K multiplied by the average service time. This is predicated on having an "infinite" reserve of ancillary employees from which to draw upon. The realism of this assumption depends on the ability to find things for employees to do when they are not serving customers and the ease at which employees can alternate between tasks. So long as there are plenty of employees in the "back room," the ancillary policy is far superior to the $M/M/m/oo$ system. In fact, if ancillary activities are available, there is little reason why a queue should ever become large.

The main obstacle to eliminating queues is the difficulty associated with diverting servers back and forth between activities. This is especially problematic (though not

- Thresholds ensure that help obtained when truly needed, yet not too frequently.
- Visible Q's: Manage fairly by opening new Q's for the longest-waiting customers.

K-P/A-C Law (2 moments, averages)



Health Insurance: Alternative between
Inquiries and Claims
(=> Mental Setup)

$$\frac{\bar{w}_q}{s} = \frac{1}{N} \cdot \frac{p}{1-p} \cdot \bar{w} \quad ?$$

index

Bottleneck Analysis : Short-Run Approximations Time-State Dependent Q_{-net}

TOUR F / A WORKER-PACED LINE FLOW PROCESS AND A SERVICE FACTORY 155

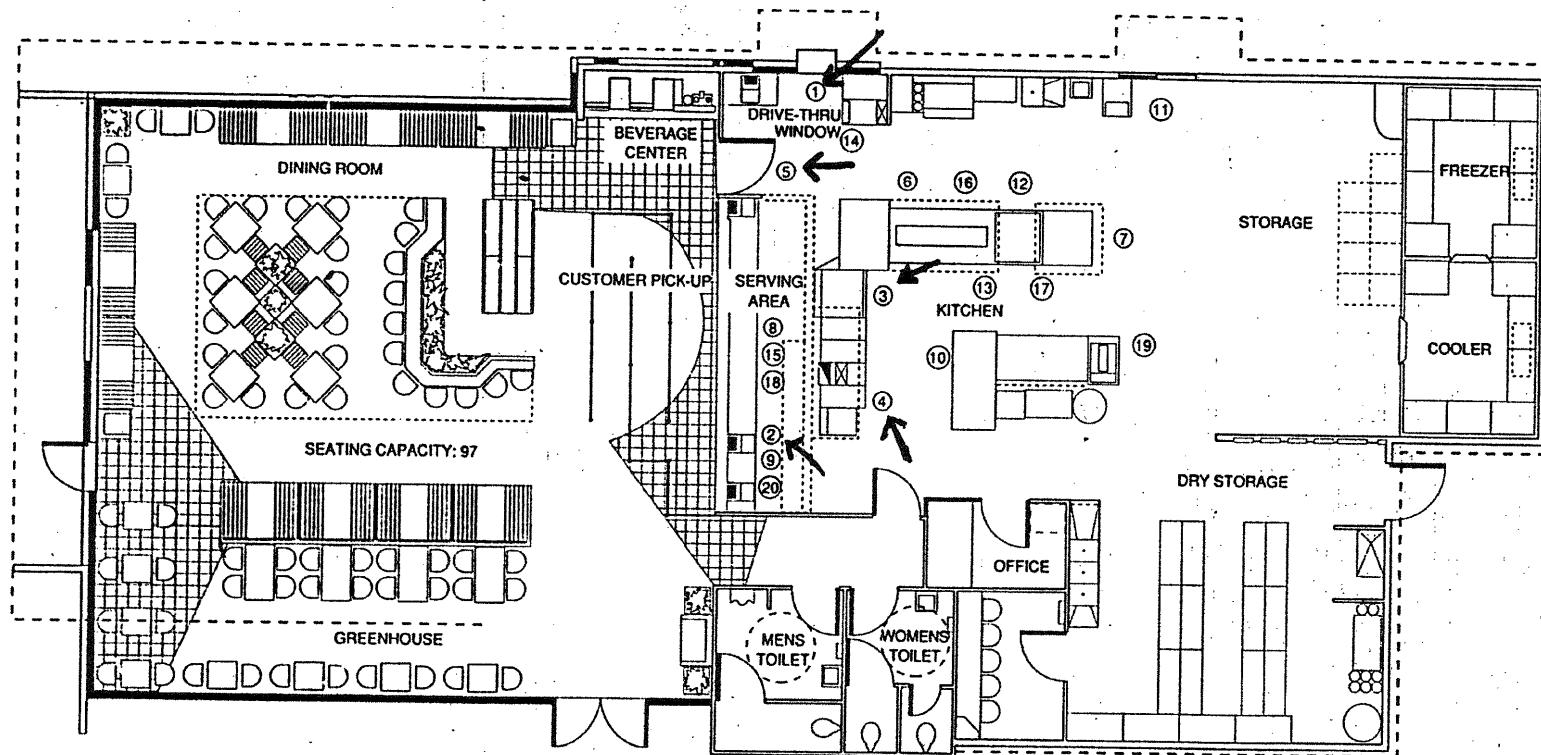


FIGURE F1 Layout of the Noblesville Burger King. The circled numbers indicate the sequence of additions of workers to the kitchen as demand increases.

3 minimal: Drive-thru
Counter
Kitchen

Add #4 Kitchen

#5 Help Drive-thru

Mapping Offered Load (Branch of a Bank)

Department	Business Services		Private Banking	Banking Services	
	Tourism	Teller		Teller	Comprehensive
Time					
8:30 – 9:00					
9:00 – 9:30					
9:30 – 10:00					
10:00 – 10:30					
10:30 – 11:00					
11:00 – 11:30					
11:30 – 12:00					
12:00 – 12:30					
Break					
16:00 – 16:30					
16:30 – 17:00					
17:00 – 17:30					
17:30 – 18:00					

Legend:

	Not Busy
	Busy
	Very Busy

Note: What can / should be done at 11:00 ?

Conclusion: Models are not always necessary but measurements are !

טבלה 2: ניתוח מצב קיימים -חקל ביצועים

Technical General Accounts

ה	מועד	מועד	מועד	
ה	ברוריים	ברוריים	אישוריים	
	■■■	■■■	■■■	
	אי	אי	אי, ג'	ימי עומס בשבוע
10-20	8-14 ; 2-3		12	ימי עומס בחוץ
1762	2476		4136	מספר פניות ביום
167	193		253.6	מופע שעתי ממוצע
<u>9:00-10:00</u>	<u>10:00-11:00</u>	<u>11:00-12:00</u>		שעות עומס
230	313	422		מופע בשעות עומס
55.9	20.0	10.9		זמן המתנה (שניות)
143.2	131.3	83.5		זמן שירות (שניות)
0.72	0.87	0.88		אינדקס שירות
11.2	5.6	2.7		אחוז נטישה
43.2	16.8	9.7		זמן המתנה ממוצע עד לנטישה (שניות)
5.2	10.3	9.7		רמת אישום ממוצעת בפועל
-	25	12		יעד - זמן המתנה

Peak
Hour

Steady-State Analysis

(From Hall, Chapter 5, page 144.)

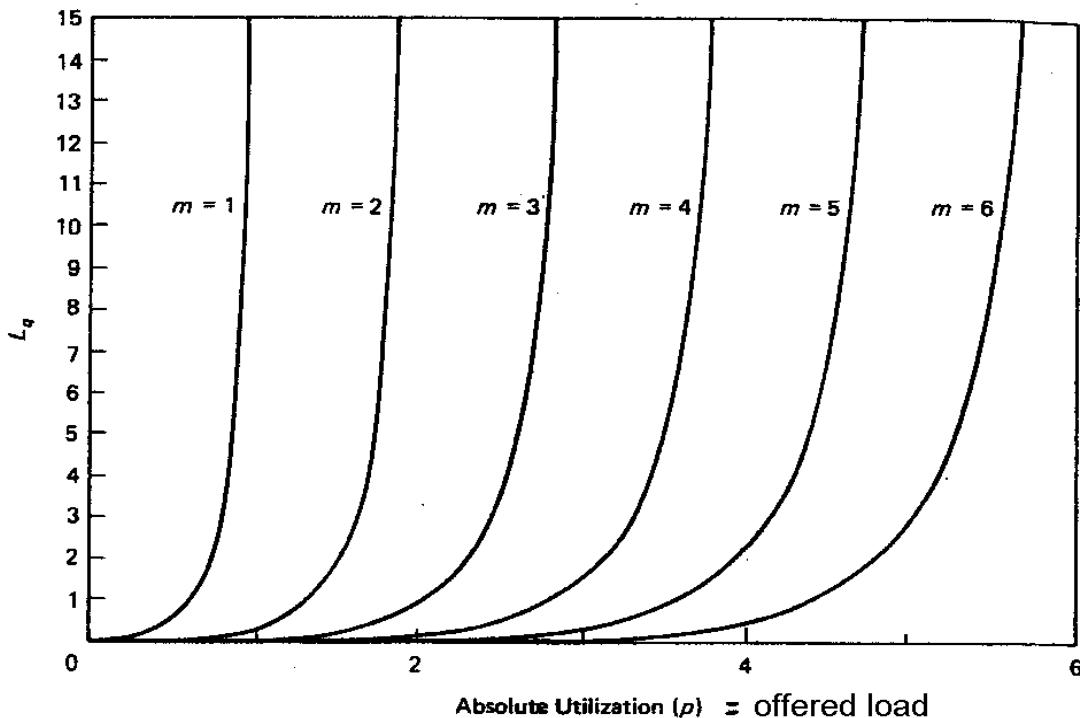


Figure 5.3 Expected queue length for the $M/M/m/\infty$ queue. L_q will be small when the number of servers equals or exceeds $\rho + \sqrt{\rho}$.

Theory ?!

be estimated by adding ρ (not ρ/m) to L_q , and W_s and W_q can be estimated in the usual manner from Little's formula. This approximation is most accurate for large values of ρ/m , close to 1 (which is to say it is a *heavy traffic approximation*. See Kölleström 1974.).

Figure 5.3 also illustrates the fundamental property that adding servers reduces waiting time. However, note that adding servers does not always provide an appreciable benefit. For ρ less than 1, L_q is nearly the same for any number of servers greater than or equal to 2, implying that one or two servers is all that is ever needed.

Example

The Wayout Arena (see example in Sec. 5.3.2) would like to evaluate the benefits of adding a second server. If $m = 2$, $\mu = 120$ customers/hour per server, $\lambda = 105$ customers/hour, and $\rho = .875$:

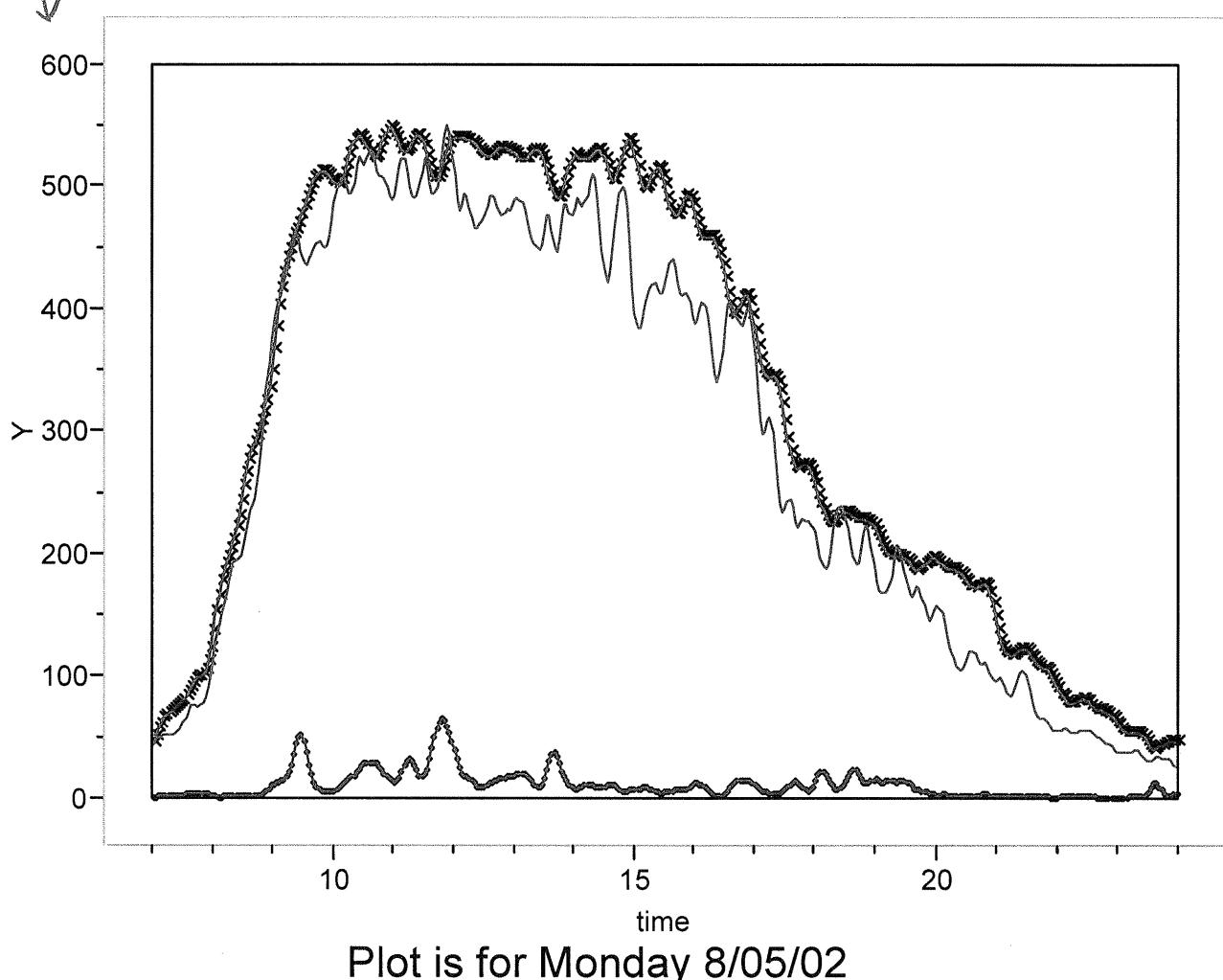
$$P_0 = \frac{1}{1 + \rho + \frac{\rho^2/2}{1 - \rho/2}} = \frac{1 - \rho/2}{1 + \rho/2} = .391$$

$$L_q = \frac{.875^3/2}{2!(1 - .875/2)^2} .391 = .529 \cdot .391 = .206 \text{ customer}$$

$$L_s = .206 + .875 = 1.08 \text{ customers}$$

agents

Efficiency Plots Showing Load and Staffing

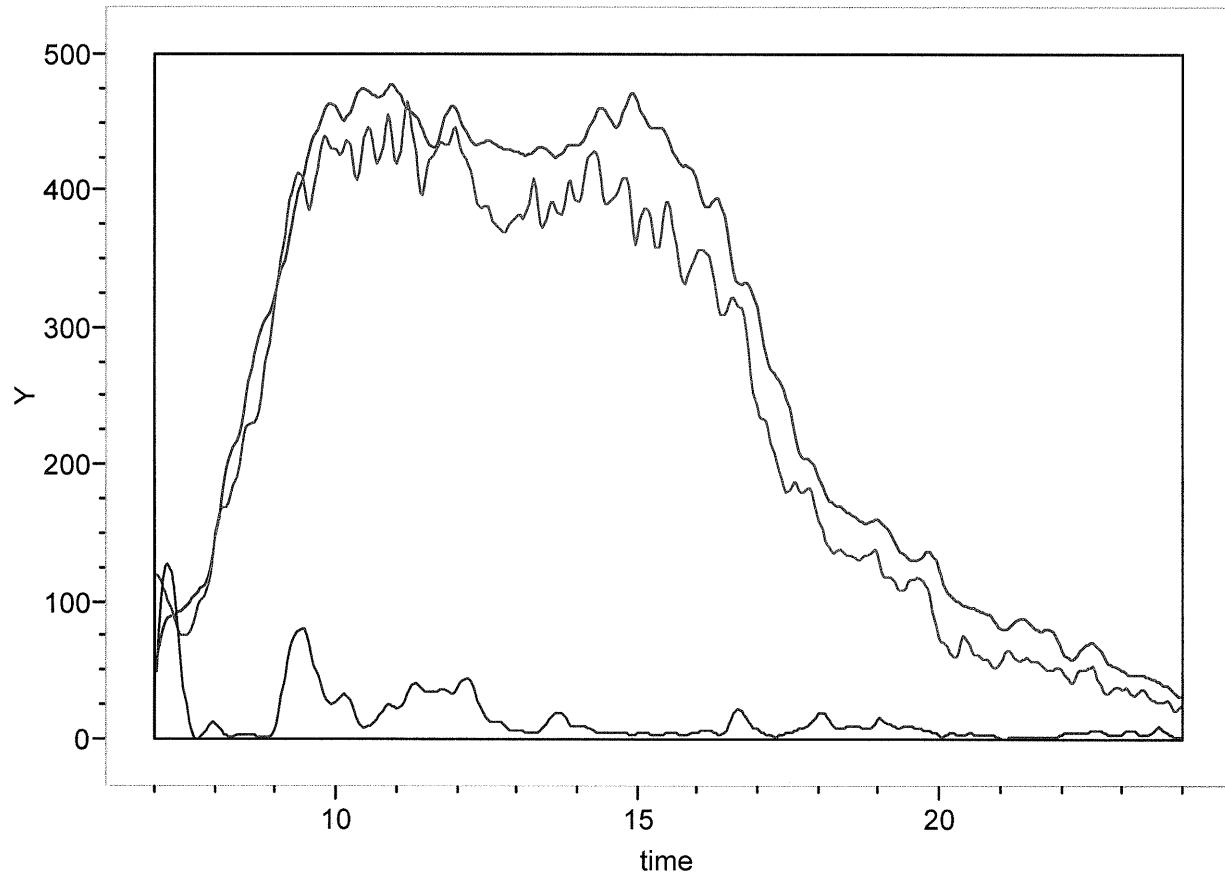


Y — NumberAgents (s)
— load (s)
— AvgQueueWaitAll (s)

“Agents” = *Estimate* of number of agents on-duty at that time.
[In each 150 second interval an agent is estimated to be on-active-duty for the entire interval if (s)he is on the phone sometime in that interval.]

Staffing Matters

Efficiency Plots, cont



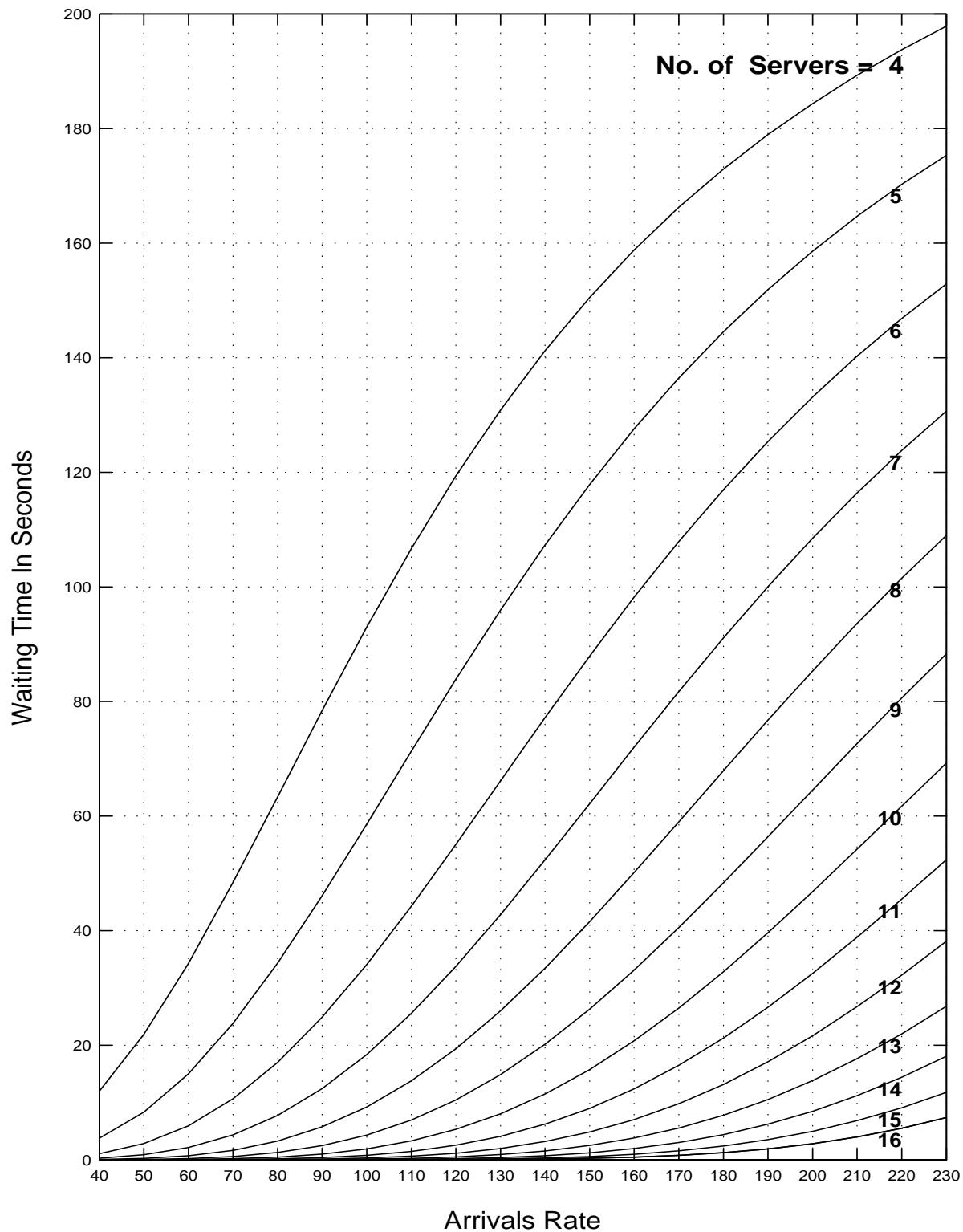
Y — NumberAgents (s)
— load (s)
— AvgQueueWaitAll (s)

Note increased usage from 7-7:30 am (typical of Fridays).
Note increased average Queue-Wait during this time.
(Accompanied by a rise in abandonments to about 10%.)

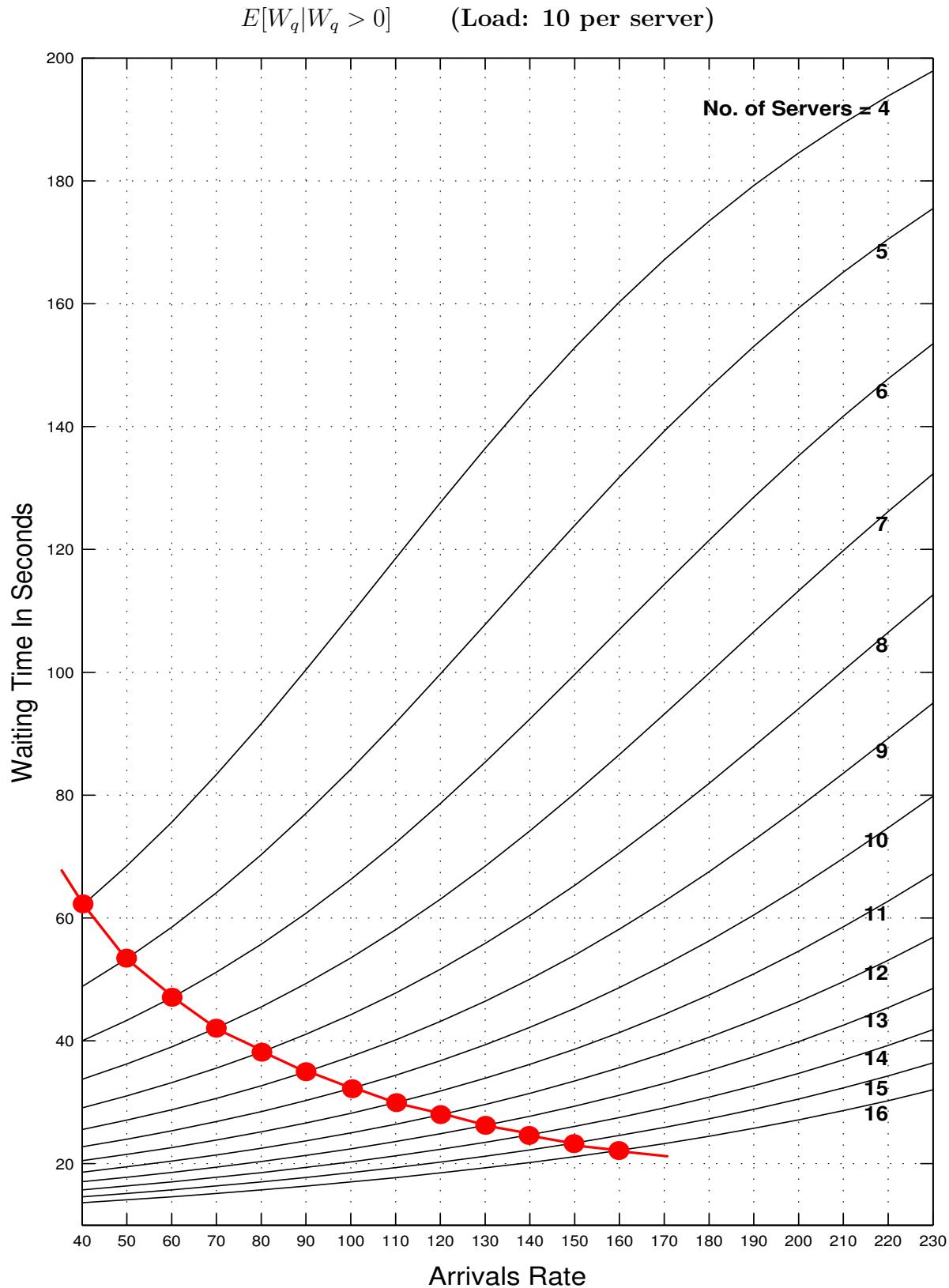
Overall Utilization: 8/02/02 = 88%
 8/05/02 = 89%

Case Study: A Large Utility Company Average Waiting Time

Commonly used MOP: $E(W_q)$
Total Service Time = 3.3 min.

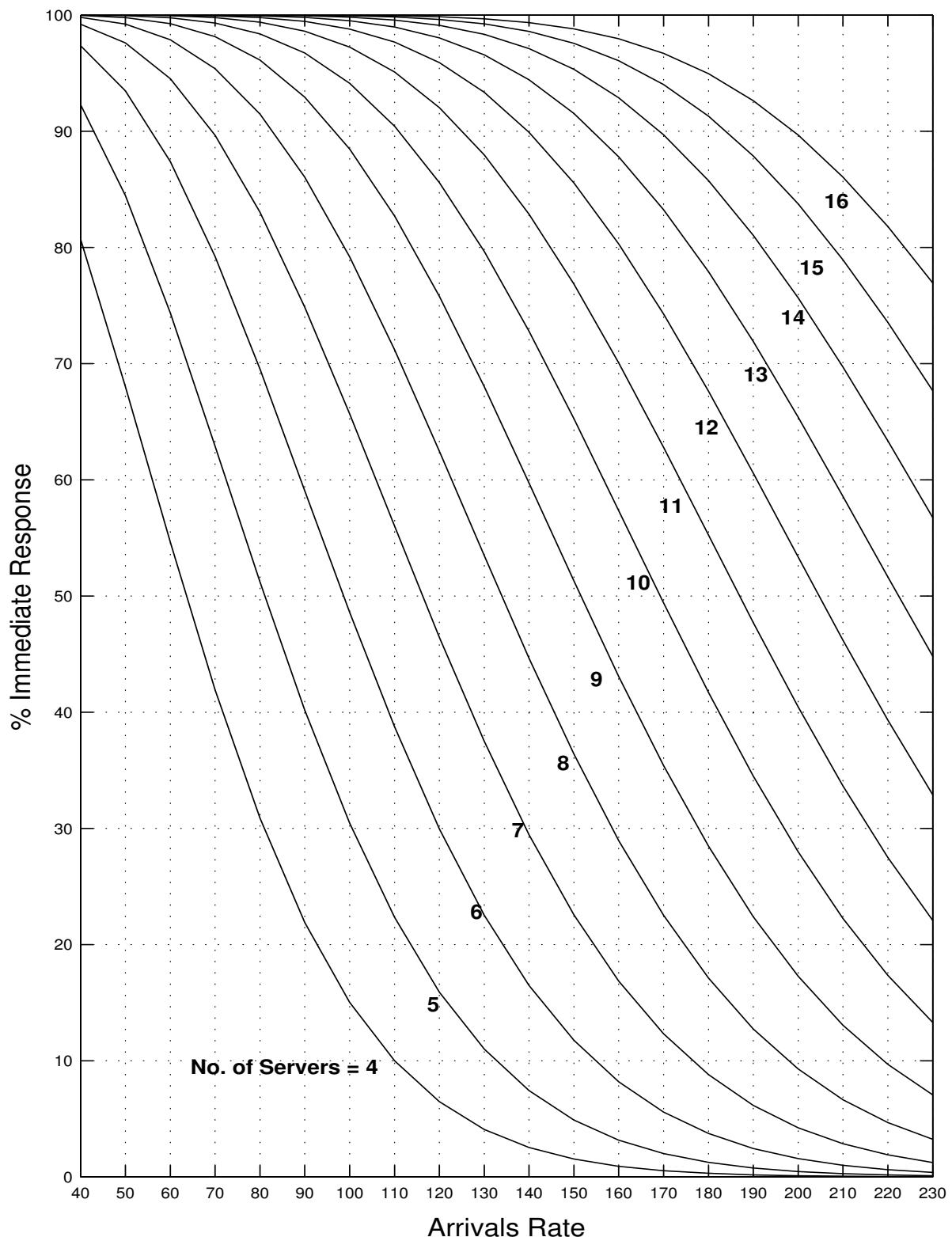


Economies of Scale
Average Waiting Time - But Only of Those Who Wait

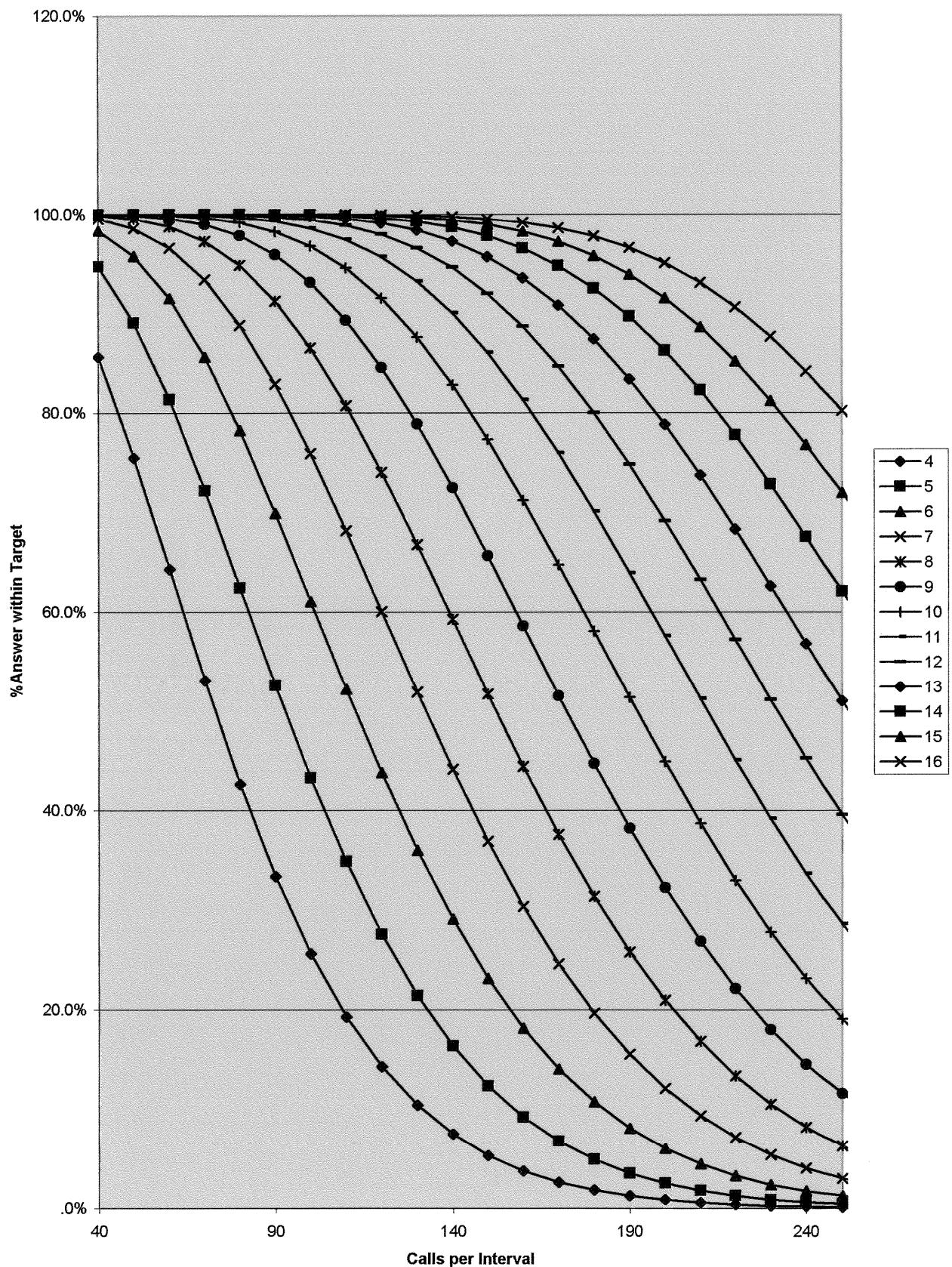


% Immediate Response (Often Not Measured)

$$P(\text{Wait} = 0)$$

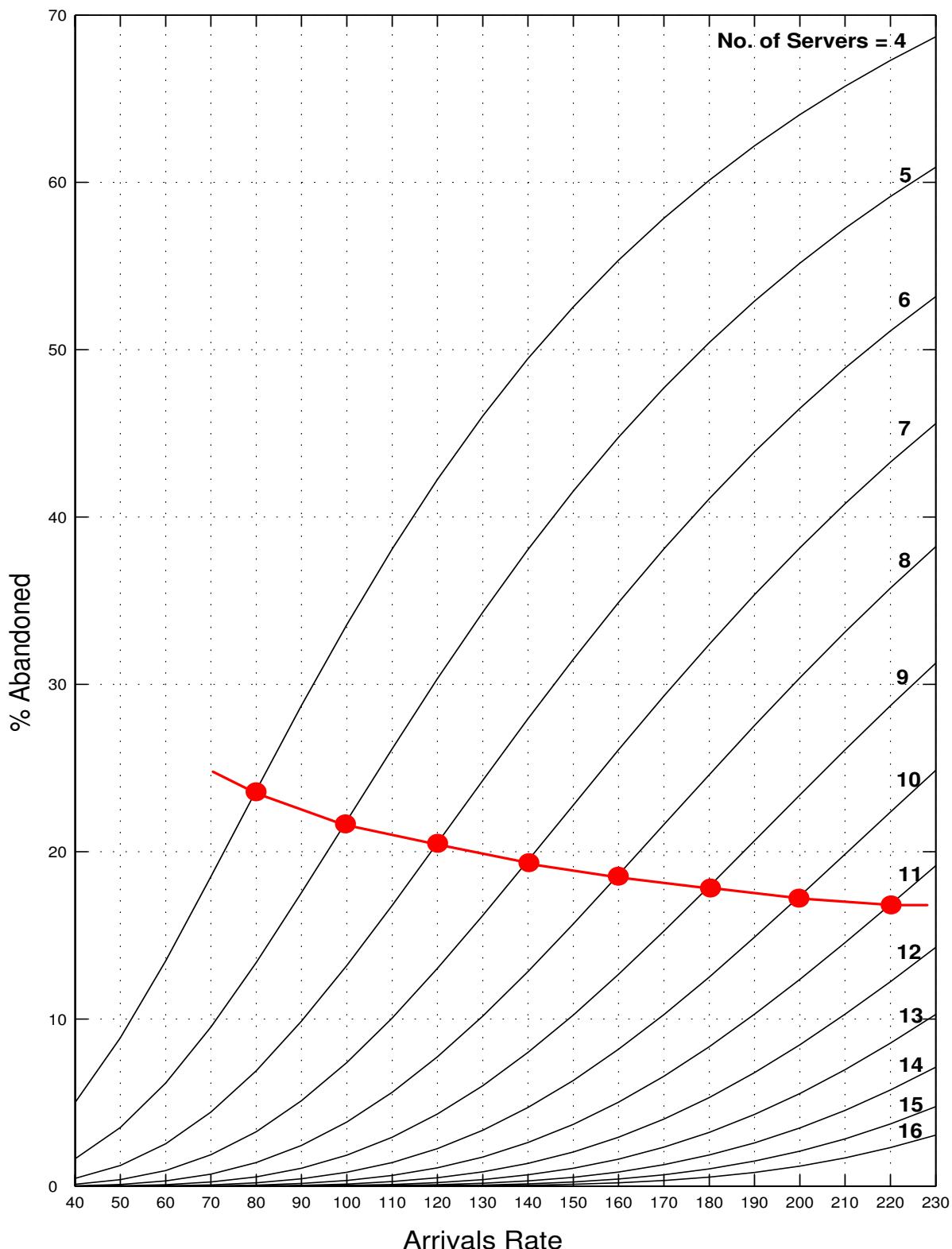


20 seconds
 %Answer within Target vs. Calls per Interval for various Number of Agents

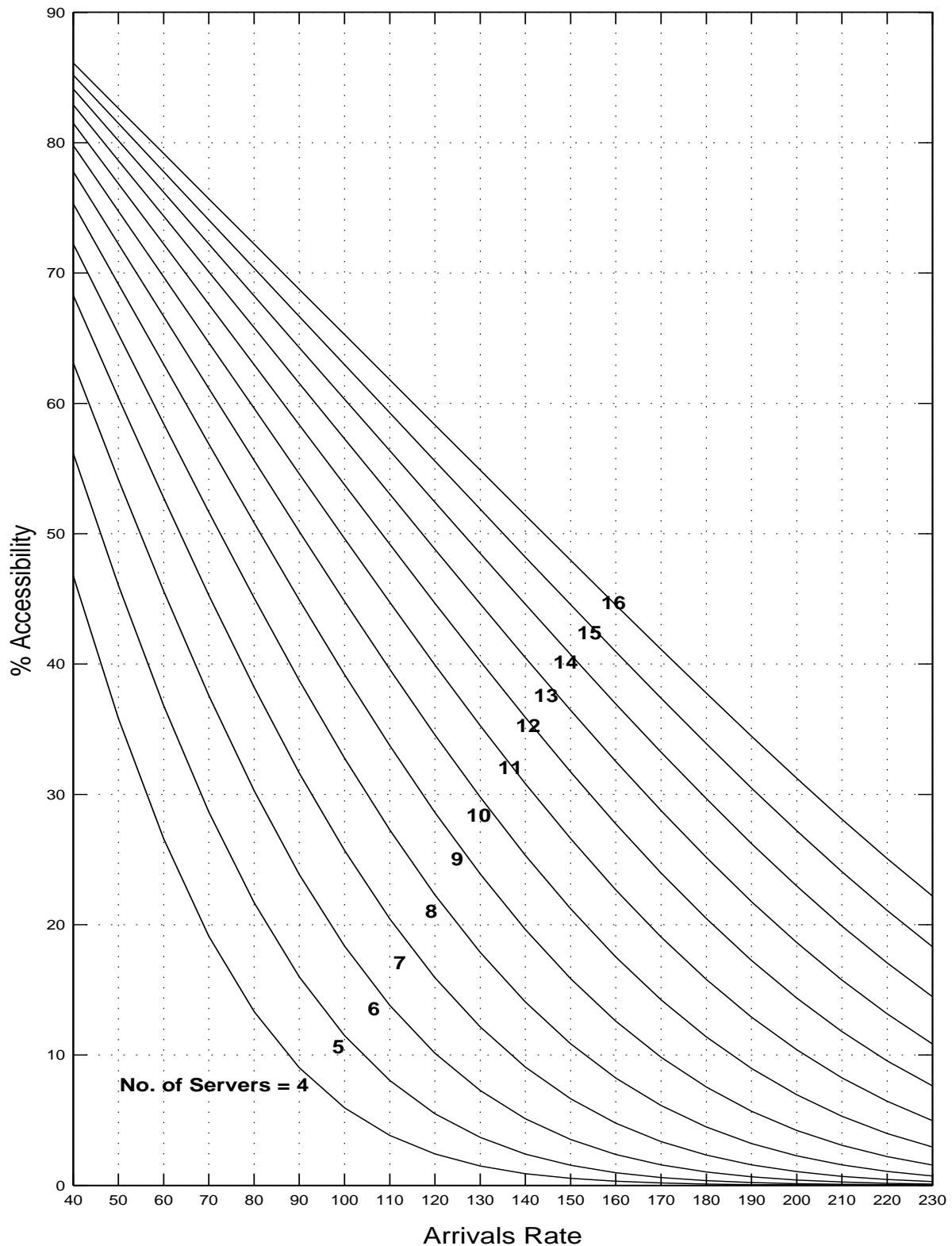


$$\begin{pmatrix}
 \text{ES} = 3:18 \\
 1/\theta = 3:30
 \end{pmatrix}$$

% Abandoned
(My #1 MOP: Subjective; Determines Operational Regime - Later.)



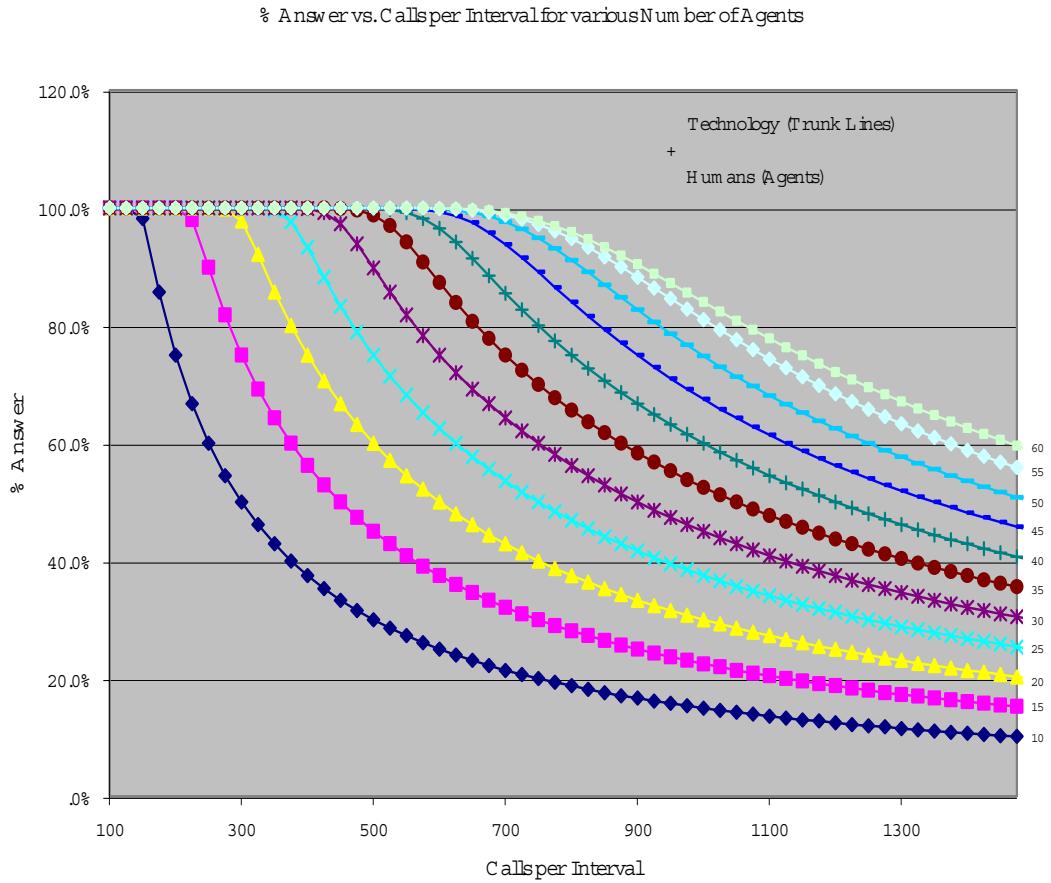
% Accessibility (Fraction of “Idle” Time)



High utilization (low accessibility), combined with high-pressure, results in very-high turnover rates (perhaps the most significant call-center management challenge).

Case Study: A Cable Company

% Calls Encountering a Busy-Tone



Combining the fraction of “busy-tone calls” - i.e., calls that arrive when all the lines/trunks are busy (hence the caller receives a busy-tone), with the amount of requests that are handled in an hour, allows one to estimate the total number of calls (successful or not) performed during, say, an hour in order to access the call center. This could be done using the following formula:

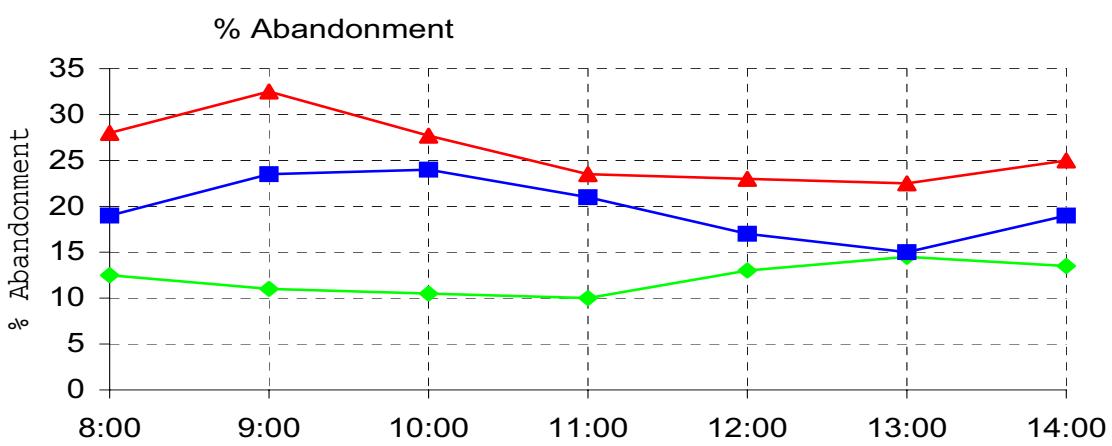
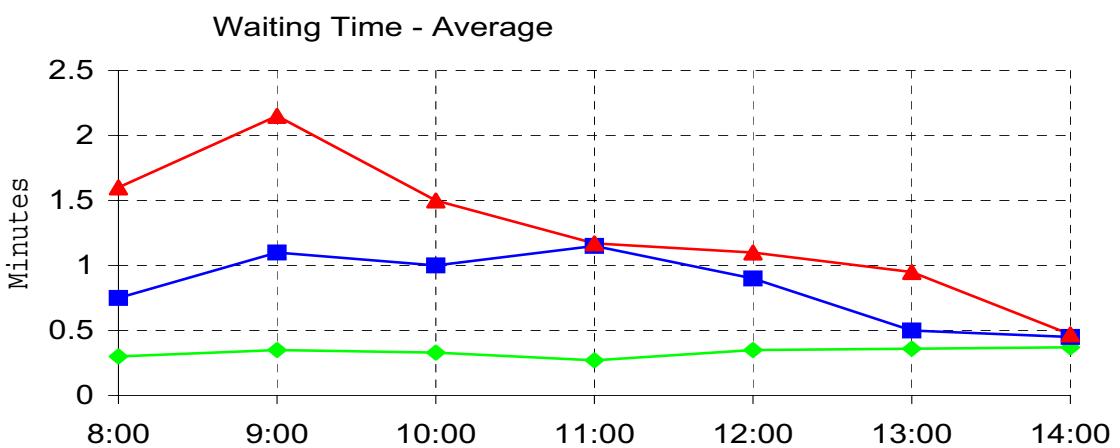
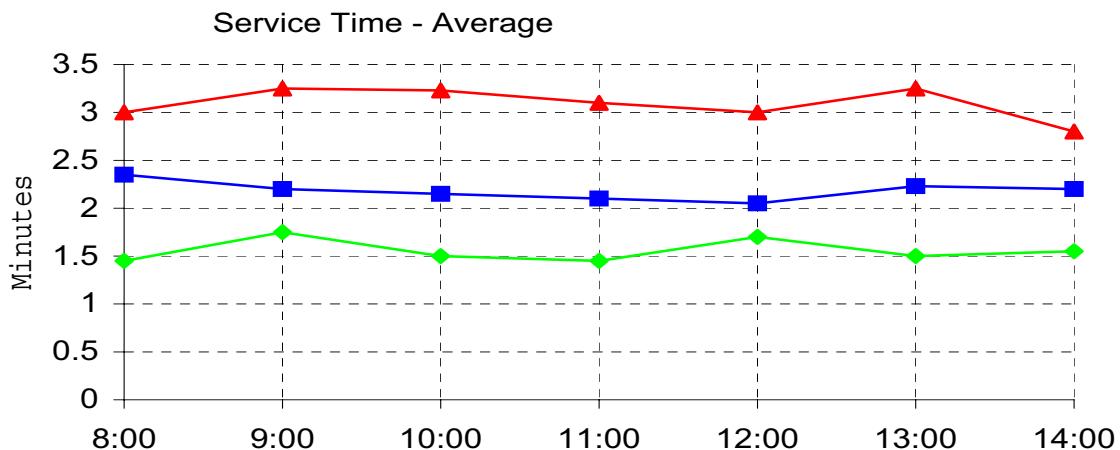
$$\text{Number of calls} = \frac{\text{Handled requests}}{100 - \% \text{ “Busy-tone calls”}} \times 100$$

The percentage of “busy-tone calls” increases as the amount of calls in an hour increases. Also, for a fixed number of handled requests, the percentage of “busy-tone calls” decreases as the number of operators increases. This is clearly manifested in the above congestion curve, trading off the fraction busy-signals with arrival rates.

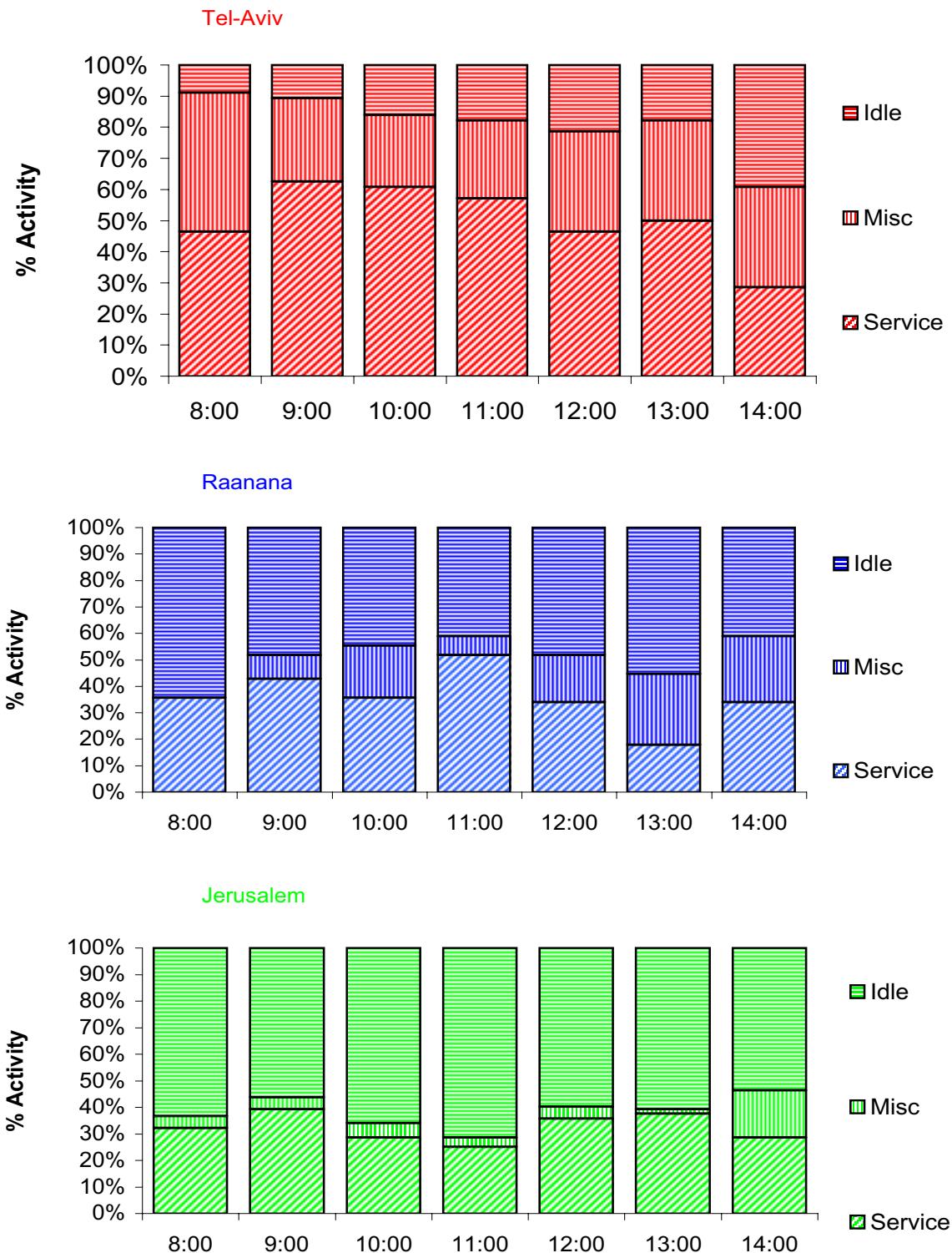
Note: In order to generate the above graph, we used an average call time of 3.8 minutes, which is 3.52 minutes (inferred from that data) multiplied by 1.08.

What is “Service Time”? or “Managing Accessibility”

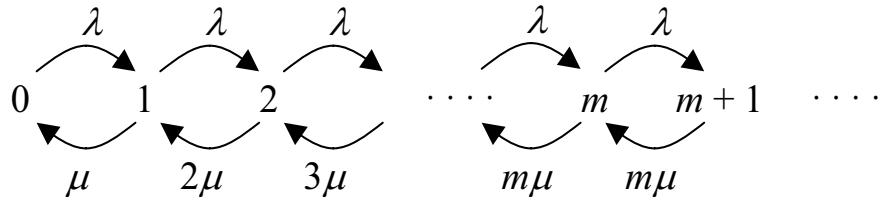
MOP's in Three Call Centers that are Doing the Same Thing !



Utilization Profiles



M/M/m Hall, Section 5.4.1 ; Whitt “Approx...” Sections 2.3 and 4.1.



Birth & Death rates:

$$\lambda_n = \lambda, \quad \mu_n = \mu(m \wedge n) , \quad \quad n = 0, 1, 2, \dots$$

Offered load: $\frac{\lambda}{\mu}$ both service work arriving per unit of time and average number of busy servers $(L = \lambda \cdot \frac{1}{\mu})$.

Traffic intensity $\rho = \frac{\lambda}{m\mu}$, also each server's utilization. Assume $\rho < 1$ for stability.

(Careful: Hall denotes $\rho = \frac{\lambda}{\mu}$, unlike here.
I have used R , and sometimes a , for the offered load.)

Steady-state equations (via “cuts”):

$$\lambda\pi_k = \mu((k+1) \wedge m)\pi_{k+1}, \quad k \geq 0.$$

Recursion:

$$\pi_{k+1} = \frac{\lambda}{\mu((k+1) \wedge m)} \pi_k, \quad k \geq 0.$$

Solution:

$$\begin{aligned}\pi_k &= \frac{m^k}{k!} \rho^k \pi_0 & 0 \leq k \leq m \\ &= \frac{m^m}{m!} \rho^k \pi_0 & k \geq m\end{aligned}$$

where

$$\pi_0 = \left[\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \frac{(m\rho)^m}{m!} \cdot \frac{1}{1-\rho} \right]^{-1}.$$

Erlang-C Formula: (Erlang Delay Formula)

$$P(\text{Wait} > 0) = \frac{(m\rho)^m}{m!(1-\rho)} \cdot \pi_0 \quad , \text{ denoted } E_{2,m}.$$

Proof: $P(\text{Wait} > 0) = P(L(\infty) \geq m) = \sum_{k \geq m} \frac{m^m}{m!} \rho^k \pi_0$

↑
PASTA

$$= \frac{m^m}{m!} \frac{\rho^m}{1 - \rho} \pi_0. \quad \text{q.e.d.}$$

Additional MOP: $E(L_q) = \frac{\rho}{1 - \rho}$ $P(\text{Wait} > 0)$,
as in Hall (5.38)–(5.40), but a nicer representation is:

M/M/m	$\frac{E(W_q)}{E(S)} = \frac{1}{m} \frac{P(\text{Wait} > 0)}{1 - \text{servers' utilization}}$
--------------	--

In fact[†],

$$\frac{1}{E(S)} W_q | W_q > 0 \sim \exp \left(\text{mean} = \frac{1}{m} \frac{1}{1 - \rho} \right)$$

(compare with M/M/1), which “suggests” the following

Kingman's Exponential Law of Congestion [‡] for **GI/GI/m**: as $\rho \uparrow 1$,

GI/GI/m	$\frac{1}{E(S)} W_q \approx \begin{cases} 0 & \text{with probability } P(\text{Wait} = 0) \\ \exp \left[\text{mean} = \frac{1}{m} \frac{1}{1 - \rho} \frac{C_a^2 + C_s^2}{2} \right] & \text{otherwise.} \end{cases}$
----------------	--

It is left to approximate $P(\text{Wait} > 0)$. See Whitt[§] (and later) for details. A reasonable approximation is to simply use Erlang-C ($E_{2,m}$). In particular, for the special case M/G/m (Poisson arrivals), one gets the following expression for the “tail” of the waiting time:

M/G/m	$\Pr \{W_q > x \cdot E(S)\} \approx P\{\text{Wait} > 0\} \cdot \exp \left[-x \cdot \frac{2m(1 - \rho)}{1 + C_s^2} \right]$
--------------	---

where $P\{\text{Wait} > 0\} = \frac{(m\rho)^m}{m!(1-\rho)} \pi_0 = E_{2,m}$, as given for the M/M/m model.

[†]Recall Gazolco: What happens if $\sqrt{m}(1 - \rho_m) \sim \beta > 0$, m large?
or equivalently $m \approx R + \beta\sqrt{R}$, with R being the offered load?

[‡]Invariance Principle (with respect to Distributions). This provides a 2nd moment approximation for Efficiency-Driven services, namely those in which essentially all customers are delayed prior to service. (With m large, this necessitates $\sqrt{m}(1 - \rho_m) \sim 0$; an example is $m(1 - \rho_m) \sim \gamma > 0$, or equivalently $m = R + \gamma$, as will be discussed below.)

[§] Whitt, W.: Recent Book (2002)

Paper: Approx G/G/m
Internet site (at Columbia)

Erlang's Formulae

(Exact Results for M/M/m = Erlang-C, and M/M/m/m = Erlang-B)

$$R = \text{offered load } (= \lambda/\mu = m \cdot \rho ; \rho = \frac{R}{m})$$

$$\text{Erlang B: } E_{1,m} = \frac{\frac{R^m}{m!}}{\sum_{k=0}^m \frac{R^k}{k!}} \quad \text{Probability of } \textit{blocking}/\textit{loss}$$

$$\text{Erlang C: } E_{2,m} = \frac{\frac{R^m}{m!} \frac{1}{1-\rho}}{\sum_{k=0}^{m-1} \frac{R^k}{k!} + \frac{R^m}{m!} \frac{1}{1-\rho}} \quad \text{Probability of } \textit{delay}$$

Relations (Palm, 1943?)

- Some observations on the Erlang formulae. pg. 18
- Contributions to the Theory of Delay Systems pg. 37

$$1. \quad E_{2,n} = \frac{nE_{1,n}}{(n-R) + RE_{1,n}} = \frac{E_{1,n}}{(1-\rho) + \rho E_{1,n}} \quad \text{for } \begin{cases} \rho < 1 \\ \left(\rho = \frac{R}{n}\right) \end{cases}$$

$$E_{2,n} > E_{1,n} ; \quad \frac{d}{dR} E_{2,n}(n) = \frac{1}{nE_{1,n}(n)}$$

$$2. \quad E_{2,n} = \frac{R(n-1-R)E_{2,n-1}}{(n-1)(n-R) - RE_{2,n-1}} \quad \text{for } R < n-1.$$

(Must have $R < 1$ to start with $E_{2,1} = \rho$)

$$3. \quad E_{1,n} = \frac{RE_{1,n-1}}{n + RE_{1,n-1}} = \frac{\rho E_{1,n-1}}{1 + \rho E_{1,n-1}} ; \quad E_{1,0} = 1.$$

Recursions are useful for calculations.

For example, to calculate $E_{2,n}$, it is convenient to calculate recursively $E_{1,n}$ via 3. and then calculate $E_{2,n}$ via 1.

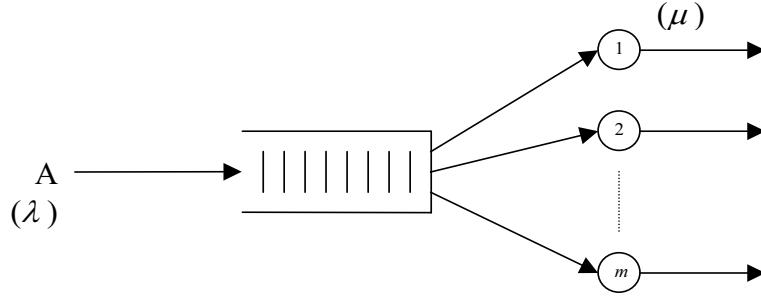
They will also be useful for us in asymptotic analysis of systems with many servers.

For example, to analyze the behavior of $E_{2,n}$, as $n \uparrow \infty$, it is convenient to analyze first $E_{1,n}$, and then use 1.

Recall: Erlang B/C/A formulae, and much more, are implemented in 4CallCenters that you have been using.

GI/GI/m/ ∞ (or $G/G/m$ for simplicity)

Recall: m servers, statistically identical and independent, attending to a single queue:



Primitives: distributions of an inter-arrival time and a service-duration.

Stability (subtle, since not always “periodically empty”).

Via Fluid view: Stability iff $\lambda < m\mu$ (load less than capacity), or equivalently $\rho = \frac{\lambda}{m\mu} < 1$ (servers’ utilization).

(Recall: Hall denotes $\rho = \lambda/\mu$ absolute utilization, which we have been referring to as *offered load*, and denoting by R and sometimes a .)

G/G/m defies exact analysis: One thus resorts to “Approximate Analysis of an Exact Model,” in an operational regime that is **Efficiency-Driven**: essentially all customers are delayed prior to service.

An E-driven operation prevails, for example, when a few-to-moderate number of servers are highly-utilized ($\rho \uparrow 1$).

Approximations, in the E-Driven Regime:

(Whitt; Hall, Chapter 5; Congestion-Laws Handout)

Fundamental: **Kingman’s Exponential-Invariance Law** (on page 20 of the present note), using only first and second moments. This implies the **Allen-Cunneen** 2nd moment approximations for average congestion measures:

$$\begin{aligned}
 E[L_q(G/G/m)] &\approx E[L_q(M/M/m)] \frac{C_a^2 + C_s^2}{2} ; \\
 E(L_q) &= \lambda E(W_q) ; \\
 \Rightarrow E[W_q(G/G/m)] &\approx E[W_q(M/M/m)] \cdot \frac{C_a^2 + C_s^2}{2} \\
 &\equiv \frac{1}{m} E(S) \frac{E_{2,m}}{1-\rho} \frac{C_a^2 + C_s^2}{2} \\
 &\approx \frac{1}{m} E(S) \frac{\rho}{1-\rho} \frac{C_a^2 + C_s^2}{2} ;
 \end{aligned}$$

$$E(W_s) = E(W_q) + 1/\mu ; E(L_s) = E(L_q) + E(\# \text{ busy servers}) = E(L_q) + \lambda/\mu.$$

“Strategic” Q-Theory

- $$L = \lambda \cdot W$$

\uparrow
manager

\uparrow
server

\nwarrow
customer

- Laws of congestion: parameters $\lambda, C_a^2 ; \mu, C_s^2 ; m, b$

\uparrow
arrivals

\uparrow
services

\uparrow
technology

Human resources
↓
(Role of the Normal distribution ? later)

- Congestion curves:
 - Determine (operational) service quality.
 - Deduce parameter values, typically m (Staffing).
 - Cross-Check MOP's (eg. Sufficient Idleness)
 - Tradeoff: Efficiency vs. Quality.
 - Continuous improvement/management control
- Economies of Scale/Scope (Mass customization, Flexible specialization).
Information Technology is the enabler (reduce “friction”).
- Service/Process design: Pooling Queues and Resources (Today);
- Pooling Tasks/Services (Later).

Recall M/M/m (Erlang-C):

$$\begin{aligned}
 P\{\text{Wait} > 0\} &= E_2(m, \rho) \\
 W_q | \text{Wait} > 0 &= \text{Exponentially Distributed} \\
 E[W_q | \text{Wait} > 0] &= E(S) \cdot \frac{1}{m} \cdot \frac{1}{1 - \rho} \\
 E[L_q | \text{Wait} > 0] &= \frac{\rho}{1 - \rho}, \quad \text{independent of } m.
 \end{aligned}$$

$$(E(W_q) = E(W_q | \text{Wait} > 0) P(\text{Wait} > 0), \quad E(L_q) = \lambda E(W_q))$$

Economies of Scale: First Observations

- Simple Example: Increase m , together with λ , while $\rho = \text{servers' utilization}$ **fixed**. Total queue unchanged (on average), hence queue per-server *and* average wait (for those waiting) “shrink” by the same factor that m increases in.
- GO TO Congestion Curves, e.g., $E[W_q/W_q > 0]$.

General EOS: “Cost/Quality” changes in a favorable direction as scale increases.

$$\text{Simple example: } \frac{\text{Fixed cost} + \text{Variable cost} \times \text{Scale}}{\text{Scale}} = \frac{F}{S} + V \downarrow \text{in } S.$$

$$\text{Subtle example: Poisson } (\lambda m) \text{ has } CV = \frac{\sqrt{\lambda m}}{\lambda m} = \frac{1}{\sqrt{\lambda m}} \downarrow 0, \quad \text{hence SLLN !.}$$

Another Subtle Example: Given ρ fixed, how does $E_2(m, \rho)$ vary as m increases?

Why EOS ?

1. Servers help each other (load shared dynamically) **1st order**
2. Stochastic variability decreases with scale **2nd order**

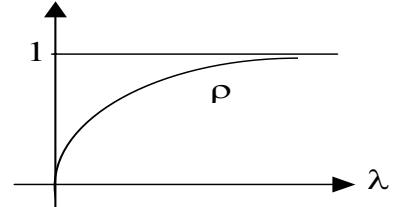
Additional simple manifestations of EOS:

$$\text{M/M/1: } EW_q = \frac{1}{\mu} \frac{\rho}{1-\rho} , \quad EL_q = \frac{\rho^2}{1-\rho} ; \quad EW = \frac{1}{\lambda} \frac{\rho}{1-\rho}$$

- $\rho = \frac{\lambda}{\mu} = \frac{n\lambda}{n\mu}$, unchanged as $n \uparrow \infty$.

Hence, EL_q unchanged with n , but $EW_q = \frac{1}{n} \frac{1}{\mu} \frac{\rho}{1-\rho} \downarrow 0$, $\forall \rho$!

- Fix EW_q , or EW . Then $\rho = \frac{\lambda}{\lambda + 1/E(W)}$
As $\lambda \uparrow \infty$, $\rho \uparrow 1$ regardless of EW .
 EW achievable at higher ρ (efficiency), as $\lambda \uparrow$.



Numerical demonstration (ALWAYS necessary for understanding))

$\lambda = 282$ customers per hour, arrive (on average) to an airport terminal.

$\mu = 1$ per min. = 60 per hour.

$m = 5$ separate M/M/1, without jockeying:

$$\rho = \frac{282/5}{60} = 0.94 \text{ very busy!}$$

$$W_q = 15.7 \text{ min.}, L_q = 14.73 \text{ customers per queue.}$$

$$\text{M/M/5: } W_q = 2.85 \text{ min.}, L_q = 13.4 \text{ (close to the previous case)}$$

Note: With $\lambda = 150$, $\rho \approx 0.5$ (utilization halved), but $W_q = 3$ seconds, $L_q = 0.13$ (performance 50 times “better”).

Pooling in a Q-Net (Part I)

- Pooling **queues** : geographic pooling (virtual service center)
- servers** : capacity pooling (fast vs. slow)
- tasks** : job design (later)

Recall Rafaeli's lecture: Operational + Psychological Aspects.

Rothkopf & Beth, 1987:

Common belief: combining queues is beneficial ...

e.g. banks and other counter systems.

But many operations do not combine queues

e.g. supermarkets, toll booths, rabbinate, doctors,...

In favour:

- (1) $m \times M/M/1 (\lambda, \mu)$ vs. $M/M/m (m\lambda, \mu)$
 $\forall \lambda, \forall m > 1$, the latter has smaller average wait + variance.
- (2) Share equipment
- (3) Fairness perception: no slips or skips

Against

Homogeneous services

- (1) $m \times M/M/1$ not always the “right” alternative to $M/M/m$;
human (intelligent) customers jockey, join shortest queue, renege
- (2) Often alternative to $m \times M/M/1$ is $M/M/m$ with overhead,
namely $M/M/m (m\lambda, \mu - \delta)$.
- (3) Physically or psychologically prohibitive:
e.g., lines too long scare customers: cars, customers with luggage,
 \Downarrow snake-like queues \Uparrow airports' customs.

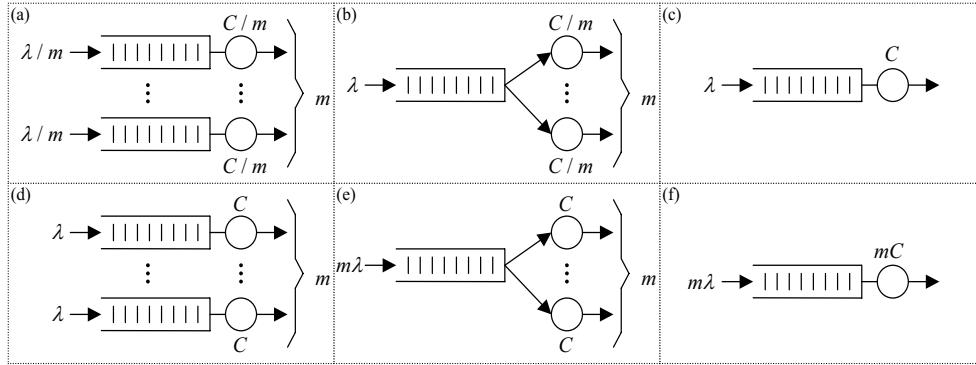
Heterogeneous customers/servers (To discuss *later*)

- (4) Depersonalization (doctors, rabbinate)
- (5) Think of combining the express-lines with the rest
- (6) Flexible servers expensive to hire, train, maintain.

Question: Design that “mixes” efficiency and fairness (physical queues)?

Business Growth (Strategic Q-Theory): Kleinrock's cycle. (1976, Classic)

Resource Sharing



Simplest is Best! Do *not* model complicated undesirable scenarios!

$$m \times M/M/1 \xrightarrow{\text{scale-up}} M/M/m \xrightarrow{\text{technology}} M/M/1$$

$$\lambda, \mu \qquad \qquad \qquad m\lambda, \mu \qquad \qquad \qquad m\lambda, m\mu$$

Combine: queues
Saved inefficiency idleness servers
(1 long queue, 2 idle) lost capacity
(rate $m\mu$ at all times)

Remark $EW_q \left(m, \lambda, \frac{\mu}{m} \right) \leq EW_q(1, \lambda, \mu)$

while $EW_s \left(m, \lambda, \frac{\mu}{m} \right) \geq EW_s(1, \lambda, \mu)$
 \uparrow
individual server's capacity

(Explain, via $P_m(\text{Wait} > 0)$, noting $W_q \mid W_q > 0$.)

Summary (pg. 287)

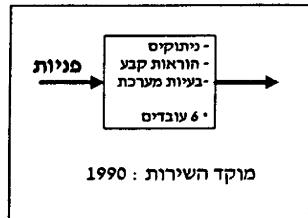
Large systems (scaling up input rate and system capacity) yield improvements (in average response-time) that are proportional to the scaling factor.

For a given scale factor, the single-server (fast) system is superior to the multiple-server (slow) system, as far as total time a system is concerned. The opposite is true, however, when restricting to only waiting time. (See Homework).

Appendix: Service System Design
 Technion IE&M Project 1996

תהליכי התהווות של מוקד שירות לקוחות גדול בארץ

המוקד עליו מדבר (שמפעילו בקשר לשמרן על אוניברסיטת) במהותו הינו מוקד שירות ללקוחות של חברת תקשורת ישראלית גדולה. החברה זו התחילה בשוק מכשירי תקשורת בישראל בשנת 1986. בשנים הראשונות לפעילות החברה, עבדה המערכת ללא מוקד שירות לקוחות. פניות לקוחות

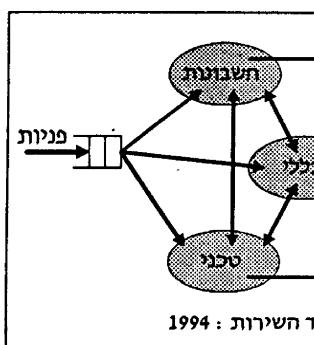


לשירות, נענו על ידי עובדי החברה הרגילים, בKOVI העבודה הרגילים של החברה.

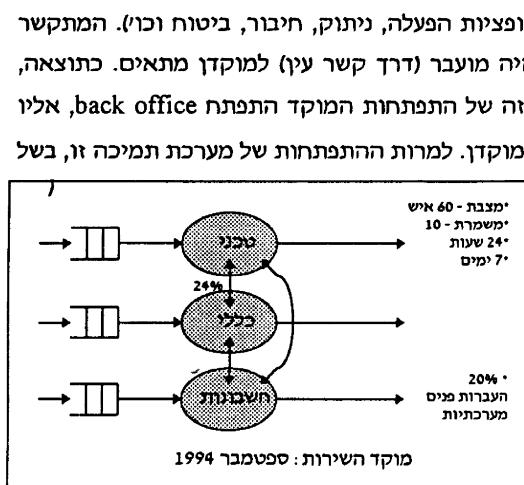
عقب גידול קהל המנוים, בינואר 1990 הוקם מוקד שירות לקוחות ראשון של החברה, בשל הצורך בשירות יותר ממושך ומסודר. המוקד אויש ביום אחד 7:00 עד 22:00 ובימים וביום 7:00 עד 14:00, ונתן מענה לפניות ובקשות

בנושא ניוקים, הוראות קבוע, וביעות מערכת. המוקד אויש על ידי 6 עובדים, ללא ACD (Automatic Call Distributor) ללא מערכת בקרה ולא תיעוד. המוקד הרושם את הבעה או הפעולה הנדרשת, גם היה מוזרים אותו הלהה ומטפל בבקשתו.

עם הזמן, וגדול במספר הפניות לשירות, גדל המוקד, והווסף ACD, עד אשר בשנת 1994 מוגבר כוח האדם של המוקד הגיע לכ- 60 איש. המוקד הופעל במשמרות של 10-15 איש, 24 שעות



ביממה, שבעה ימים בשבוע. הפניות למוקד בשלב זה היו עדין דרך מספן אחד בלבד, והמתקשרים המתוינו לשירות בטור יחיד. עם זאת, הונגה במוקד חלוקה מקצועית פנימית, וזאת מכיוון שמנגנון השאלות והבעיות עליהם ניתן גודל משמעותי, והמוקדים אינם יכולים להתמצא בכל הנושאים. לא יכולו להשתתף לשולש תחומים: חשבונות, נושאים טכניים, ושירותים כלליים (אופציות הפעלה, ניוק, חיבור, ביטוח וכו'). המתקשר היה מנותב באקרים למוקד כלשהו ואז היה מועבר (דרך קשר עין) למוקד מתאים. כתוצאה, העברות פנים מערכתיות היו רבות. בשלב זה של התפתחות המוקד התפתחה back office, אליו זרמו הפעולות לביצוע כפי שנרשמו על ידי המוקד. למרות ההתפתחות של מערכות תמיכה זו, בשל



ההעברות הרבות במערכת ובזבזז המשאים שנוצר, בספטמבר 1994 פרסם החברה שני מספרים נוספים לפניות שירות. כל אחד משלושת המספרים יועד לפניה מסווג אחד בלבד (טכנני, כללי, וחשבונתי). פרסום המספרים וייעודם נעשה דרך עיתון המנוים של החברה, ובמספר המדריך של

החברה ללקוחות חדשים. מတקירים שפנו למספר הלא-נכון, והעברו למחלקה המתאימה לאחר שהובהר להם המספר הנכון אליו עליהם לפנות בעתיד. עם זאת, עדין היו ההצעות במערכת כ-24% מכלל נפח הפעולות, ועל כן הוחלט להפסיק להעביר מתקקרים שהתפקידו למספר לא נכון. הדבר יצר בעיות בתקשות עם הלקוחות, שכן החלוקה לאنفسה, למעט המספר הנפרד של חשבונות. העבודה נמשכה במתכונת זאת עד מרץ 1996, כאשר הוחלט כי יש צורך למצוא פתרון אחר. בנוסף, עד מרץ 1996, חל גידול משמעותי במספר המנויים במספר השירותים הניתנים. חוסר הייעול של מבנה המוקד הנוכחי וקשיי יכולת להתמודד היו מאי ברורים למול כמות השירותים הייסטרית שהחלה לזרום מדצמבר 1995, כשהחברה יצאה במצב מיוחד, בה הצעה שורת תקשות מזול בשעות הערב והסופשבוע. במציע זה בחרה החברה לשנות את גישתה ללקות, ופנתה לשוק מטרה חדשה למזרי, אותו לא התעסקה בעבר. המוקד אmons היה ערוץ למבחן מבחינת עדכוני מידע, אך לא היה ערוץ למעבר מוקד שירותים למוקד טלמרקטיינג. 70% מנהל פעילות המוקד עבר להיות מידע שיווקי ותפעולי, ויתר הפעולות שהוויה את מירב נפח הפעולות עד לשלב זה, הצטמכם ל-30% הנוטר. המבצע הביא מנגנום חדש משכבה סוציאלית אחרת ממנייה החברה עד כה, שכן קהל היעד העיקרי של החברה בעבר היה קהילתי, ובמציע זה מדובר היה בקהל יעד המעניין במכשורי תקשות ניידים לשימוש פרטי בשעות הפנאי. השינוי באוכלוסיית היעד גורמה גם לעומסים כבדים בשעות הערב והסופשבוע, בשונה מההתפלגות עומס השירותים עד עתה, שהתרכזה בעיקר בשעות היום.

בעקבות הנלמד מתפקיד המוקד בחודשים שלאחר המבצע, נעשו שינויים מערכתיים, והמוקד עבר לפעול שוב עם תור מאוחד לשירותים טכני וככללי, בשיטת חזית ווורף.

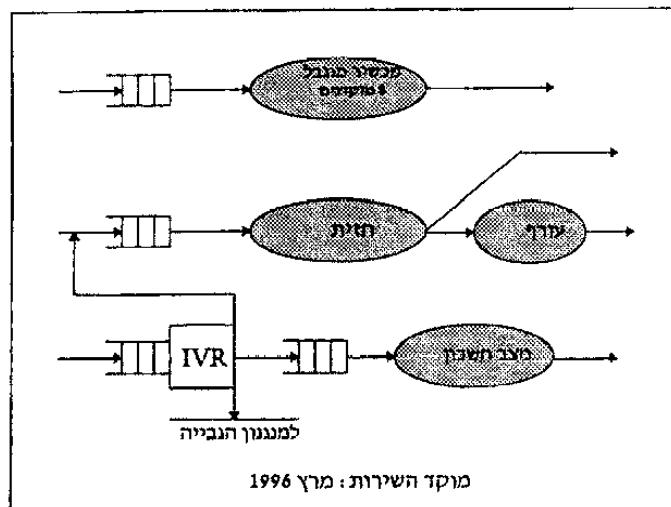
המוקד הנוכחי

המוקד הנוכחי פועל בשני תורים, האחד לבקשתו של לקוחות כללי וטכנאי, והשני לחשבונות. בטור החשבונות, מופעל מענה קולי (IVR) משלב ACD, המכادر למתוך לבוחר מראש את סוג השירות הנדרש לו, ועל ידי כך מועבר למקום המתאים במערכת. ל��וחות המתקקרים לחשבונות, אך מעוניינים בשירותים טכני או כללי, יכולים לבחר באופציה המתאימה ולעבור לטור השירות הכללי.



המוקד הכללי בניו בצורת חזית ווורף. החזית נותנת את מירב השירותים, כגון ניירוקים וחייברים, הוראות קבוע, מידע שוקי ותפעולי, תקלות פשעות, וכדומה. העורף מטפל בunosאים הדורשים מומחיות וידע רב, כגון תקלות ברמה מתקדמת, מכשירים מסובכים, שירות שבחן מבקשים מנהל או אחראי, ובירורים יותר מרכיבים הדורשים בדיקה אל מול מחלוקת אחרת. לעורף מגיעים כ-5% מכלל נפח השירות, אך מבחרנית כוח אדם, העורף מהויה 20%. עם זאת, השירותים המטופלים על ידי העורף מהויה יותר מ-5% מכלל זמן העבודה, וזאת מכיוון שהשירותים אלו ממוצעים אורכו יותר זמן משήה המטופל בחזית.

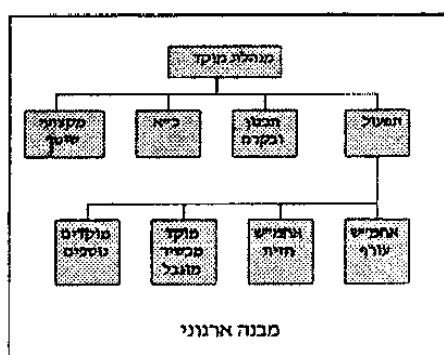
כיוון, העורף והחזה אינם מסוכרים בינהם, וمتקשר יכולים ללחוץ לשיחת עורף במשך זמן רב, אם הוא מתעטש להישאר על הקו, כולל שיחה ארוכה המטוגלת בעורף לפניו.



במשך השנים גדלה מזבון כוח האדם (ל-105%). עד 52 עמדות יכולות להיות מאושנות בו, ומנית (עורף וחוזית). באפריל 1996 יצאה חברת התקורת במכשור ותקשורת חדש וול. ביצועי המכשור מוגבלים יחסית לביצועי מושרי החברה הרגולים, ונמכר בתפוצה רחבה מאד. למוקד

השירותות זוטר עומס כבד של שירותים מכשור מוגבל. מכיוון שאוכלוסיית היעד הייתה שונה, ומכוון שהגבילות המכשור נתנו פיתוי גודל להתקשרות למספר השירות, שהינו שירות חינם, החלטה החברה להפריד בין מוקד שירותים למכשור המוגבל לבני מוקד השירותים הרגול. שיקול נסף היה הרצון לשמר על אפשרות למוקן אופי שונה ורמה שונה של שירות בכל אחד מהמוקדים. על אף זאת, בחודש הראשון למכירה, 1 מותך כל 3 לקוחות התקשר פעם ביום למוקד - 4500-5000 שירותים ביום רק למוקד המכשור המוגבל. עומסים אלו גרמו לסתימות מערכת המענה הקולי והניתוב (ACD). בעקבות זאת נקנה IVR (מענה קולי אינטראקטיבי) חדש, אשר הופעל במרכזיית למוקד המכשור המוגבל. אחת מטרות IVR זה היא להפחית את מספר שירותים השווה המגיעות למוקדים, ועל כן ב-IVR 9 אופציות, כאשר האחרונה מבינהם הוא לתגיאר לשירותים קוליים (מוקד).

המבנה הארגוני וכוח האדם

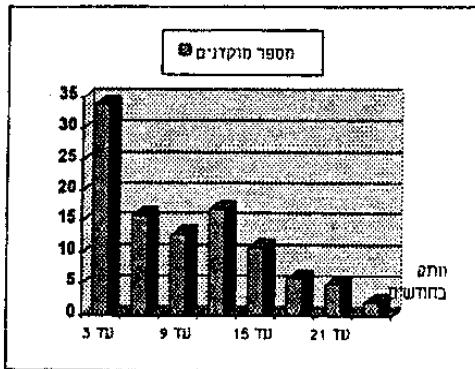


בראש המוקד עומדת הנהלת המוקד. מתחתיה ניצבים מנהלי היחידות הבאות: תפעול, מכון ובקраה, כ"א של המוקד, ומקצועי שוטף. תפעול המוקד מתרכו בידי מנהל התפעול, שהינו אחראי על סגן מנהל התפעול, מנהל מוקד המכשור המוגבל, אחראי משמרות חווית, עורף, וככלי, וכן על מוקדים נוספים של החברה. אנשים אלו אחראים על הפעלת מוקד השירותים ומוקד שירות

המכשור המוגבל, ומרכזים ומפקדים על צוותי העבודה. על כל צוות מפקד ראש צוות, ואיש המוקד מוחלק בצוותים של עובד יראה את ראש הצוות שלו לפחות פעמיים בשבוע.

העובדת במקדד היא במשמרות של 6.5 שעות, כאשר יש הפסקה של רבע שעה כל שעתיים. קביעת המשמרות נעשית בעורת תוכנה, המבנתה את הדורישה למספר המוקדנים לפי צפי שיחות. הקליטה לעובדה במקדד נעשית לאחר מבחני מין מצעיים וראיון אישי. העובדים הנקלטים עוברים קורס אינטנסיבי הכלול מבחנים וסימולציות. העובדים השורדים את הקורס נוכנסים לעובדה במקדד, ומוגדרים כמתלמידים למשך 3 החודשים הראשונים. לאחר חודשי התלמידות מקבלים העובדים החודשים ציוני ביצועים בדומה לעובדים הוותיקים, ולאחר חצי שנה של עבודה, באם עמדו במצוון מינימלי, נקליטים כעובדים קבועים.

כוח האדם מורכב בעיקר מסטודנטים וחייליסות משוחררים (כ- 30% סטודנטים, ו- 70% חיילים משוחררים). מעת מאייד העובדים הם עובדים זותרים וקבועים.



תהליכי השירות

שירותות כללי וטכני.

מנוי מתקשר למערכת באמצעות חיוג חינם ממכשיר התקשרות שברשותו, או על ידי חיוג רגיל למספר בזק. השיחה מגיעה ל-ACD, המעביר את השיחה למוקדן חזית פנוי, או בהעדר מוקדן פנוי, מכניס את השיחה למזר, עד אשר יתפנה מוקדן. מוקדן החזית מביר את בעיתת הליקות. כאשר מדובר בבקשת בה הוא מוכשר לטפל, מבצע המוקדן את הטיפול והקלדת המידע למערכת המחשב. באם מדובר בשירות, תקללה או שאלת אשר מוקדן החזית אינו מוכשר לטפל בו, מועבר המני למוקדן עורך, או ליחידה המתאימה בחברה (ימען התקנות, טלמרקטינג, יחס לקוחות). אין העברת שיחות לחשבון.

שירותות חשבוןונת.

מנוי מתקשר בעורת חיוג חינם ממכשיר שברשותו. השיחה מגיעה ל-ACD המשולב IVR ומענה קולי אינטראקטיבי, ולמנוי ניתנת האפשרות לבחור בין שלוש פעולות רצויות: גביה, בירורים, ומצב חשבון. השיחה מנוטבת בהתאם לבחירת המני: בחירות גביה מנוטבת את השיחה למתקנת הגביה של החברה (העברה לתוך החברה ומחוץ למערכת המוקדן), בחירות בירורים מנוטבת את השיחה ל-ACD של השירות הכללי והטכני, ובחירות מצב חשבון מנוטבת את השיחה לאחד ממקדני חשבונות של המוקדן, לשם בירורים על מצב חשבון המני; במקרה שאין מוקדן חשבונות פנוי, מנוטבת השיחה לתוך המתנה למוקדן כזה.

From a Stanford **MBA Exam** (A “True story” - my first encounter with the subject):

Question 12: QUTE & City Bank (20 points)

Consider the following quotation from the case “First National City Bank Operating group (A)” (HBS Case). (There is no need for you to consult the case itself; the quotation is all that is required to answer the question below.):

“By tradition, the method of meeting increased work load in banking was to increase staff. If an operation could be done at the rate of 800 transactions per day, and the load increased by 800 pieces per day, then the manager in charge of that operation would hire another person; it was taken for granted But, in the late 1960s, the work load began to rise faster than the hiring rate could keep up Backlogs of work to be done would pile up in one OPG department or another, and they could not be cleared away without overtime. Even with extensive reassignment of people and with major overtime efforts, some departments would periodically fall behind by two or even three weeks, generating substantial numbers of complaints from customers.”

Evaluate the above practice of meeting increased demand. In particular, explain why backlogs started to build up. Support your answer with facts acquired in class discussions, course readings or assignments. On the *next* page, there is a summary of some QUTE output, with parameters that fit the above quotation. (The time unit is “day’s work”, and the arrival rate is in 100’s of transactions per day.) Refer to this output in your answer: Either reason why the output supports your answer or explain why it does not.

Evaluation:

The QUTE program output summarizes an M/M/S model whose input represents transaction load (in 100’s per day), S is the number of workers, U is the utilization of a worker, W_q is the average time in queue for a transaction, L_q is the average backlog, L is the number of transactions in the system (queued and in-process).

QUTE tells us that “linear” response to increasing load has the following effects: Workers utilization increases with load (for example, 50% utilization with $S = 2$, 83% with $S = 32$, 98% with $S = 51$). The average waiting time for a transaction also increases, but not as dramatically as might be first expected from the high utilization rates. (The reason is the economies of scales, or pooling of resources, as observed in class and as exploited by the 411 directory in N.J.). In the bank operation during the 1960s, in contrast to 411 in the 80’s, pooling was not carried out electronically. Hence the actual performance, under heavy loading and a large number of workers, should be in fact worse than what the M/M/S model predicts (probably much worse). Add to that the high utilization rates per worker, which are likely to be impossible to sustain over a full day’s work, and you deduce the large backlogs which City Bank was lead to suffer.

QUTE Output M/M/S

 $\lambda = 8k, \quad k = 1, 2, \dots \quad (\text{i.e. } \lambda = 8, 16, 24, \dots),$
 $1/\mu = 1/8 = 0.125,$
 $n = k + 1, \quad k = 1, 2, \dots \quad (\text{i.e. } S = 2, 3, 4, 5, \dots),$

$\underline{\lambda}$	\underline{n}	\underline{U}	\underline{W}_q	\underline{L}_q	\underline{L}
8	2	50%	0.04	0.33	1.33
16	3	66.7%	0.056	0.89	2.89
24	4	75%	0.064	1.53	4.53
32	5	80%	0.069	2.22	6.22
40	6	83%	0.073	2.94	7.94
48	7	85.7%	0.076	3.68	9.68
56	8	87.5%	0.079	4.45	11.45
64	9	88.9%	0.082	5.23	13.23
120	16	93.7%	0.091	10.95	25.95
400	51	98%	0.097	41.93	91.91
640	81	98.8%	0.105	67.18	147.2
			↓		
			$E(S) = 0.125$		

Animation: - Bank

- Teller capable of handling 800 transactions per day.
- Policy: load increased by 800 per day \Rightarrow hire another person.

Analysis: In the n -th system ($m = n + 1$) we have

$$\rho_n = \frac{\lambda_n}{(n+1)\Delta} = \frac{\lambda_0 + n\Delta}{\Delta + n\Delta} = \frac{\frac{\lambda_0}{n} + \Delta}{\frac{\Delta}{n} + \Delta} \uparrow 1, \quad \text{as } n \uparrow \infty.$$

 Here we assume $\lambda_0 < \Delta$ (above: $\lambda_0 = 0$; $\Delta = 8$ for 800 transactions)

$$\text{Key observation: } n(1 - \rho_n) = \frac{n(\Delta - \lambda_0)}{\Delta + n\Delta} \rightarrow \frac{\Delta - \lambda_0}{\Delta}, \quad \text{as } n \uparrow \infty.$$

$$\Rightarrow \rho_n \rightarrow 1 \quad (\text{Efficient})$$

$$\Rightarrow P_n\{\text{Wait} > 0\} \rightarrow 1 \quad (\text{later: Efficiency-Driven regime})$$

$$\Rightarrow W_q \approx \exp \left(\text{mean} = \frac{E(S)}{(n+1)(1 - \rho_n)} \right) \xrightarrow{d} \exp \left(\text{mean} = E(S) \times \frac{\Delta}{\Delta - \lambda_0} \right): \text{congestion index}$$

4CallCenters
Garnett's Software (~~Internet Version~~)

Call Center iProfiler™

Performance Profiler Staffing Profiler Settings Edit Account Send Feedback

Performance Profiler Tool - Find out the Performance Level of your Call Center.

Performance Profiler Tool allows you to determine and optimize the Performance Level of your Call Center.
Please enter Your Call Center's parameters below.

Number of **Agents** in your call center Agents. Features: None Selected.

Basic Interval: 60 Minutes.

Average **Time to Handle** one call (mm:ss) : Target Time: 00:30 (mm:ss).

Number of **Calls** per 60 minutes Calls.

[Add To Table](#)

[Compute](#)

	Basic Interval	Target Time to Answer	Number of Agents	Average Handling Time	Calls per Interval	Agent's Occupancy	% Answered within Target	Average Speed of Answer	Average Queue Length
	60	00:30	451	07:30	3600	99.8%	11.8%	07:04.3	424.3
<input type="checkbox"/>	60	00:30	451	07:30	3600	99.8%	11.8%	07:04.3	424.3
<input type="checkbox"/>	60	00:30	321	07:30	2560	99.7%	12.7%	06:59.7	298.5
<input type="checkbox"/>	60	00:30	161	07:30	1280	99.4%	15.2%	06:47.9	145.0
<input type="checkbox"/>	60	00:30	81	07:30	640	98.8%	18.6%	06:31.8	69.6
<input type="checkbox"/>	60	00:30	51	07:30	400	98.0%	21.4%	06:17.9	42.0
<input type="checkbox"/>	60	00:30	16	07:30	120	93.8%	31.7%	05:28.5	11.0
<input type="checkbox"/>	60	00:30	10	07:30	72	90.0%	37.4%	05:00.9	6.0
<input type="checkbox"/>	60	00:30	7	07:30	48	85.7%	42.6%	04:36.2	3.7
<input type="checkbox"/>	60	00:30	5	07:30	32	80.0%	48.2%	04:09.4	2.2
<input type="checkbox"/>	60	00:30	4	07:30	24	75.0%	52.3%	03:49.2	1.5
<input type="checkbox"/>	60	00:30	3	07:30	16	66.7%	58.4%	03:20.0	0.9
<input type="checkbox"/>	60	00:30	2	07:30	8	50.0%	68.8%	02:30.0	0.3

[Delete All](#)

[Delete Selected](#)

- Current Result - Settings - Call Center Parameters - Performance Indicators