

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	λ	μ	$\theta \Rightarrow$ practical (Palm, Garnett)
	General	C_a^2	C_s^2	?
		(heavy-tails		Efficiency-Driven (Kingman) 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)

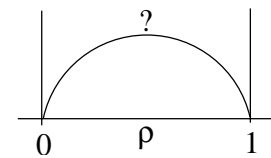
[M/M/m, M/M/m+GI: Either Exact Analysis of Approximate Model, or Asymptotics, with Many Servers (QED Call Centers)]

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: <u>Staffing</u> :	who,	when and,	for how long,	how many?
	\downarrow	\downarrow	\downarrow	
	Tough		work shift	<u>HW</u> , 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.

זה עניין כמעט אבוד

זו השיטה, והיא עובדת תמיד: מתי שלא תבוא, תמיד יהיה תור, כי מישוהו כבר ידאג לכך. איך הוא יעשה את זה? הוא פשוט יפעיל רק קופה אחת מתוך השש הקיימות

Hall, Chapter 5, Section 5.7
Ancillary Activities
(Implicit:
servers' flexibility)

◇ Add a server whenever
 $\frac{\#QueuedCustomers}{\#Servers} > K$
(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.

ניקח את עניין הסופרמרקס רק כדוגמה. כמשהו שיכול לייצג בשבילנו את היחסים הדי עכורים, לפעמים, בין הערכים לנתון השירות. בין זה שמשלם את הכסף ולכן וכאי לשירות הולם, לבין זה שמקבל את הכסף ולכן מחויב לשהיה הולם. נביח שאנחנו מחליטים שמחיר יש לנו יום עמוס במיוחד, ושהעומס הזה עומד בניגוד משוער לריקנות של המסדר. המסקנה היא, שהיחידים ללכת לסופרמרקס לעשות קניות, ועוד יותר חייבים לעשות זאת הכי מוזר שאפשר, בשעות שיהיה שיעורן די ריק באזור הזה. בשבע בבוקר בדיק, שעת הפתיחה, מתייצבים ליד דלתות הסופר, מוחצים כאן המקום ריק יחסית, וחסמים במהירות מזהירים מה שצריך, רבע שעה אחרי זה מתייצבים בתור לקופה. ואז כשהתור שחלה ככל הנראה טעות. איך זה יכול להיות שבסופר אין כמעט אנשים, ובכל זאת ליד הקופה יש תור מתארך. מבט קצת יותר מעמיק מסביר את העניין, אם כי לא צריך אפילו להעזיף מבט. זהו זה השיטה, והיא עובדת ובה: כתי שלא תבוא, תמיד יהיה תור, כי מישוהו בסופרמרקס כבר ידאג לכך. איך הוא יעשה את זה? הוא פשוט יפעיל רק קופה אחת מתוך שש הקיימות. והוא יעשה את זה, כי מה שחשוב לו זה לא לתת לנו שירות טוב, אלא לחסוך לעצמו בה אדם, ולכן עדיף שאנחנו נעמוד רבע שעה בתור כאשר שהוא ישלם עוד כמה שקלים לקופאית נוספת, אולי אפילו לבתים.

זה לא רק בסופר

זו השיטה היא עובדת בפלא גם בסניפי בנקים, למשל. גם שם, איך שראים שיש בתור דק חמישה אדם, מיד סוגרים דלפק אחד, בשביל שבצליה מזה כאור להרכיב מניין גם אם הוא לא כשה השיטה הזו טובה לחניה רבה, בעיקר הגרורות. אף פעם לא מאיישים שם את כל הקופות. אף פעם אין מצב בו הקופאית מחכה לנו ולא אכזרית לנו. והרבה. אם מישהו רואה שיש קופאית שיושבת חמש שניות ללא תעסוקה, מיד הוא שולח אותה להפסקה הקטנה, או לעשות עבודה אחרת. למה לבזבז ולקחת את הדי יורדים שהם צריכים תמיד לעצמם בתור, או למה להתאמץ בשבילם מה, הם לא יקו פה יותר אם הם לא יקבלו את השירות המגיע להם? האנשים יקנו מי אינו מכיר את הערכים הישראליים - הוא יעמוד בתור, יקטר קצת בינו לביתו, אולי יגיד לשכניו מתחרים, אליהם הצליח להתמודד רי סה בכשר חזק, שהוא לא בסדר ולמה הם לא מותחים עוד קופות, וכו' יסתוים העניין. התעקשות על התיאור הסטטיסטי של מהלך העניינים בתנוחת גדולות, לא צורך כדי לשפר שם את ההתנהגות. עניין כמעט אבוד. הוא נוצר כדי להסביר שהבעיה כולה מונחת אצלנו, הצרכנים. אנחנו באים למסעדות, בתיים לפעמים שעה בין מנה למנה, מקבלים אוכל קר, בשעה צריך להיות חם, או חם בשעה צריך להיות קר - ושתקום כמה פעמים קצתם באמצע אוזניהם במסעדה, והדעתם לבדלי המקום שאנחנו לא מוכנים לחכות יותר מזה. שעה בין כמה למנה, הסברים להם שסבם לא עושים צורך, והלכנו. אני מבטיחה לכם שאחריהם פועלים כאלה היה משתה אפי השירות באותו המקום. אבל אנחנו, הצרכנים, המסוימים שאנחנו עושים זה להעזיף לבדלי הבית. שאנחנו מחכים כבר וזהה זמן, אז הוא שוטח בפנינו את בעיות כוח האדם שלנו ובעיה קשה לו להתארגן, ואנחנו סופחים בחיבה על שכנו, אומרים לו שבעד הכל ברור. וממשיכים לחכות. לפעמים אנחנו תופשים אסטרטגיות רמתחצית למלצר, כמו שאמרה לי חברה, כשטיפרה שהעירה בחריפת מסימת, למלצר שהביא לה הזמנה הפכה לחוליה מוקדמת. ובחוזר. לרגע למיקרוקוסמוס של הסופרמרקס, כדוגמה למקור. הגעתם בשעה טובה לקופה, שמתם את הכספים שלכם על הסדר הנע, הרטוב והמלוכלך מבצרים קדמים, ואתם מבקשים שקצת ינקו עבורכם את הדבר הסודי והרטוב הזה. "חכו רגע", אומרת הקופאית ורצה לחפש חתיכת נייר כדי לנקות. היא מחפשת את סגנית המנהל, שמחפשת את המנהל, וכולם יחד מחפשים נייר. יש אפשר להמשיך ואז מסתבר שעל אחד מהמזכרים ששמתם בעגלה אין לא קד ולא מחיר. מייד שולחת אתכם הקופאית לבד את המחיר. אם אתם כבר משלמים לנו כסף - כדאי שתעבדו קצת עבודה. לא בדיק בדרך לכם איך לבדד את המחיר, אבל לא נורא. היא תסביר לכם מה אתם צריכים לעשות, כדי שתבצע את העבודה כהלכה. בשעה טובה ומצולחת מסתיים כל התהליך, החשבון מוגש לכם, ואתם מזהלים שלם במזומן. קורה. צריך לתת לכם עודף של איכזרים ש"ה, אבל בקופה אין עודף. למה שבמקום שנועד, בין השאר, לתת לכם עודף, יהיה כסף ואם לא יהיה, אנחנו נפסיק לכוה לשלם אנחנו נאבד למי שצריך את מה שאנחנו חושבים על השירות שהוא נותן לנו? לכן זהה רגע, אמרת הקופאית. אני רוצה לפרוט 100 ש"ח. שוב מחפשים את הסגנית, שמחפשת את המנהל, וכולם יחד מחפשים מטבעות של עשרה ש"ח. עכשיו נותן רק לקרוא לעובד שאחראי על "עזרה לאוסר", כדי שייסיע לכם להגלה את החבילות, אם התלפתם לא לעשות משלוח, ולכן אין טעם להתייאש. אז מה אם העובד הזה יושב במחסן והיה, למרות שהוא צריך לשבת ליד הקופות? תמיד זה ככה, מסבירים לנו, ואנחנו, נוסחים שכמונו, אם זה תמיד ככה - מי אנחנו שגשנה מדי עולם.

לא התחשק לי לריב

בשבוע שעבר נכנסתי לחנות גדולה מאוד של טקסטיל בול-אביב. ערסתי שם קניות גיכרות, בכמה מאות שקלים, הכתורתי לי בתור הגדול ליד אחת הקופות, וציפיתי לטוב. עשר דקות אחרי שנעמרת שם הודיעה לי הקופאית שקלי לחפש לי קופה אחרת, כי היא יוצאת להפסקה. "מה את רוצה, שאני לא אצא להפסקה כבר מזמן הגיע לי". לא, א"ר לא רוצה שהוא לא תצא להפסקה. אני רק רוצה, שכשהיא יוצאת תהיה מישוה שתחליף אותה. אז אני רוצה גרדת את העגלה הגדולה לקופה סמוכה, ושם כבר היתה קופאית יותר נחמדה, שמראש הודיעה לי שהיא לא מקבלת יותר סוב גרדתי את העגלה, שוב עמדתי עשר דקות בתור, ואז, כשהנחתי את הדברים ליד הקופה, הסתבר שפריט אחד לא מסומן. קיבלתי הוראות מדויקות איך לבדוק את המחיר, למי לגשת, מה לשאול. ואיוו תשובה לקבל. הודעתי שא"ר לא כוכנה לבצע את העבודה. תודהמה גדולה אחזה את זו שמולי. צרכן שלא מוכן להישמע להוראות. איפה נשמע ב"ר הזה?

"אז אל תקחי את הפריט הזה", הציעה קפאית. גם לכך סירבתי. בשבילנו נכנסתי בכלל לחנות. מה עושים ובינתיים התור מתארך ואין מוצא. שאלה הקופאית את הקולגות שלה, לא ידעו. בסופו של דבר, מפאת חובים על המחכים מאחורי, ומפני שאני צרכנית יסראלית טיפוסית, ויתרתי על הפריט, ובעיקר ויתרתי על העסקה ולא הלכתי להתלונן בפני המנהל על הקופאית, ולא התעקשתי על הוסבת לקבל שירות, וגם לא נקטתי בסנקציה ההיררכית העומדת לרשותי. לעולם לא להיכנס למקום הזה בשנית, וגם להזהר את כל חברי. למה לא עשיתי את זה כי היה טרם, ולא היה לי כוח לעמוד על שלי, ולא התחשק לי לריב. סתם. בעיקר מפני שעל לקחת כמוני בנזיה השיטה. ■

יזהר ספ-מלך
19/5/96 יזיז

Ancillary Activities / Dynamic Staffing (Trading-off Customers' Waiting-time vs. Servers' Interruption)

Sec. 5.7 A System with Ancillary Activities ← Hall, Ch 5.181

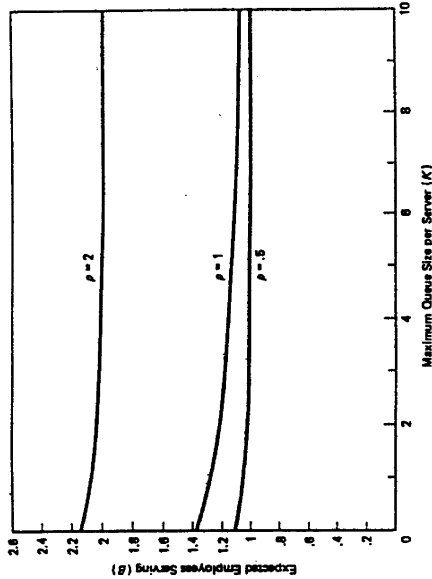


Figure 5.12 The expected number of employees serving declines as the maximum queue size per server increases. For $p > 1$, the limit is p . For $p \leq 1$, the limit is one.

in service

Add servers

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

By comparison to the $M/M/1$ 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/μ . In fact, for large values of p , the expected time in queue is approximately:

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

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 $M/M/1$ 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

Policy:

- Reductress

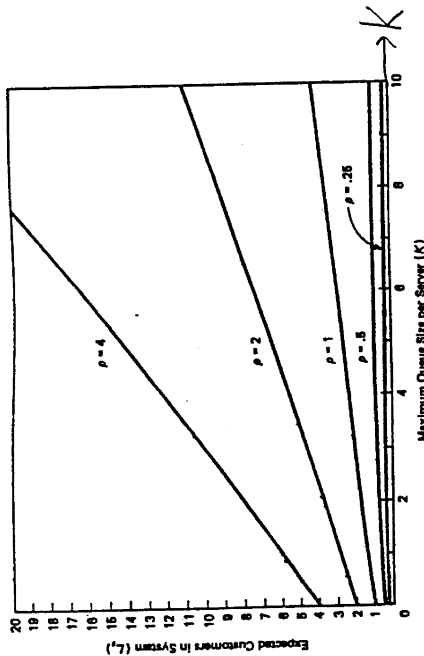


Figure 5.10 Increasing the maximum queue size per server increases L_s at a nearly linear rate when p is greater than 1. For $p < 1$, L_s approaches $p/(1-p)$ as K becomes large.

L_s

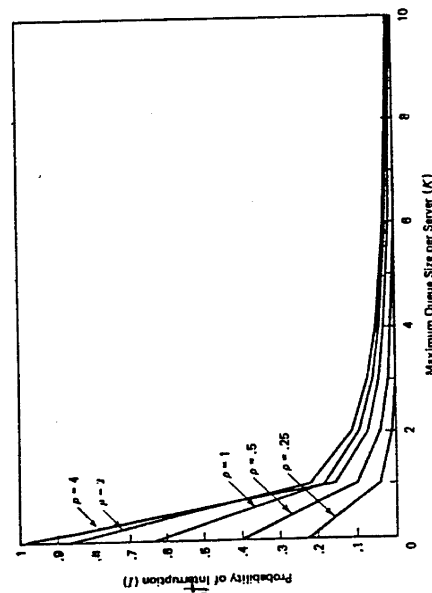
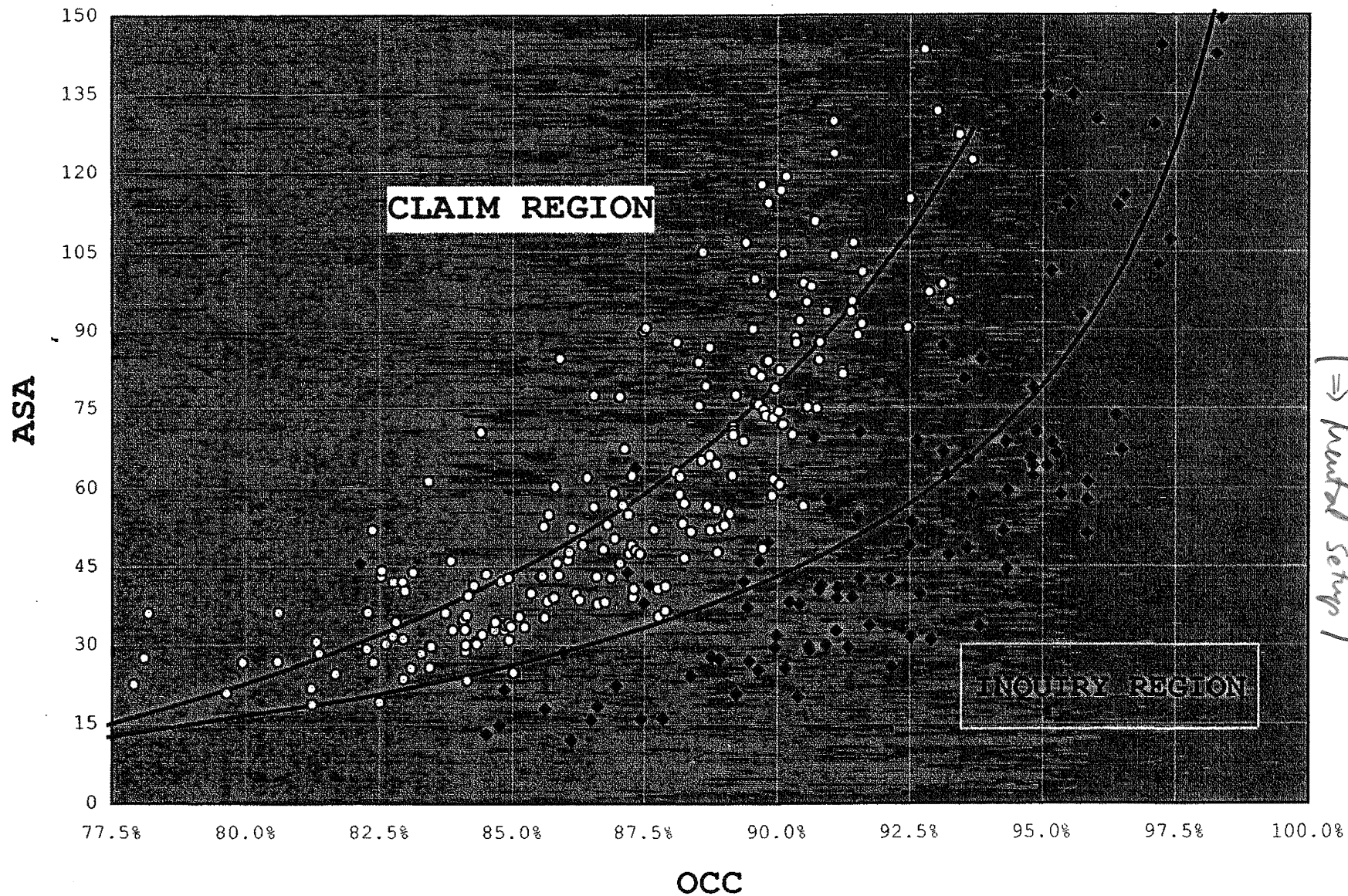


Figure 5.11 The probability of an interruption declines toward zero as the maximum queue size per server increases. For $K \approx 3$, the probability of interruption does not exceed .1.

interruption

- 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; ^{proportion}averages)



Health Insurance: Alternate between
Inquiries and Claims
(\Rightarrow nested setup)

$$\frac{\overline{Wq}}{\overline{S}} \approx \frac{1}{N} \cdot \frac{p}{1-p} \cdot \bullet \rightarrow ?$$

index additive

8-9
6

7.1

Bottleneck Analysis : Short-Run Approximations Time-State Dependent Q-net

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

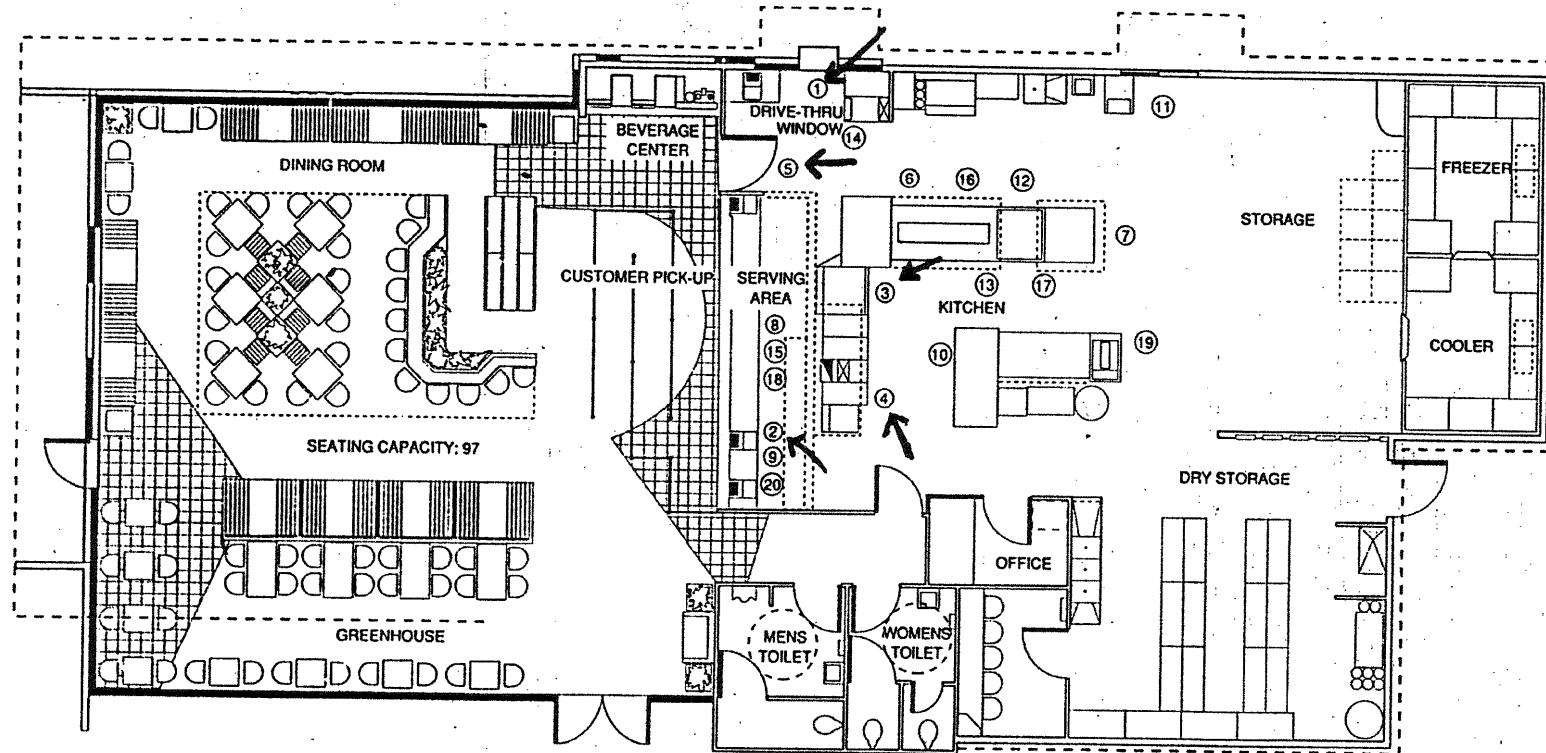


FIGURE F1 Layout of the Noblesville Burger King. The circled numbers indicate the sequence of additions of workers to the kitchen as demand increases.

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	
Time	Tourism	Teller	Teller	Teller	Comprehensive
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 קיים -חקר ביצועים

Technical General Accounts

מוקד מוקד מוקד	מוקד ברורים	מוקד אישורים	
א'	א'	א', ר'	ימי עומס בשבוע
10-20	8-14 ; 2-3	12	ימי עומס בחודש
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 איוש ממוצעת בפועל
-	25	12	יעד - זמן המתנה

Peak
Hour

Steady-State Analysis

(From Hall, Chapter 5, page 144.)

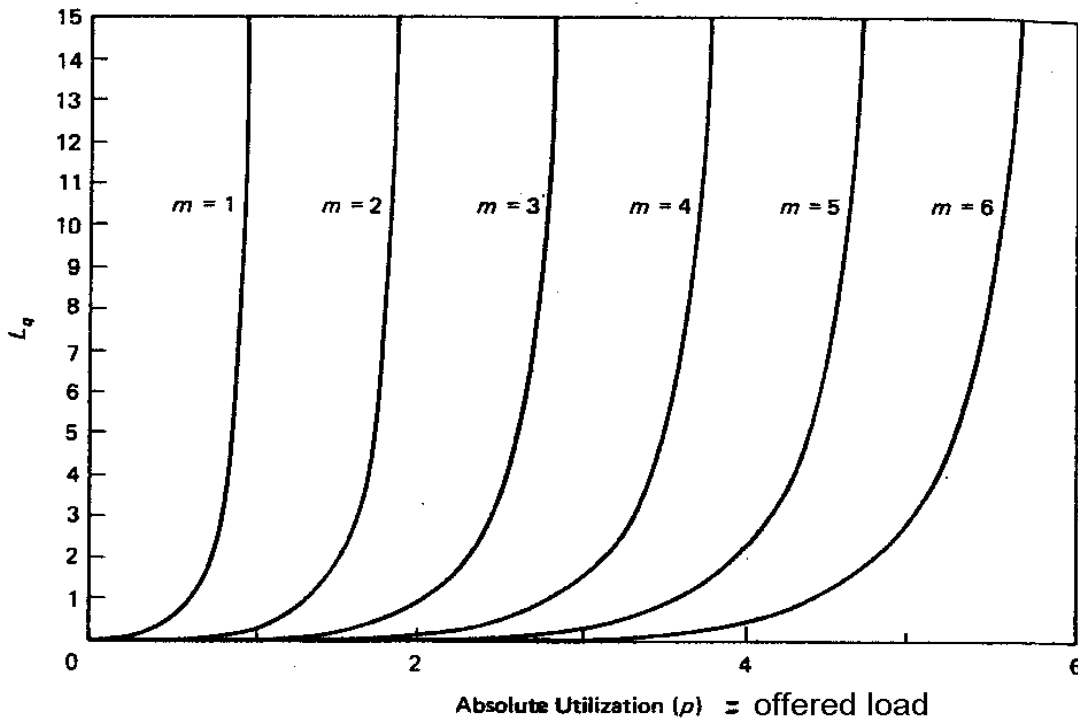


Figure 5.3 Expected queue length for the $M/M/m/\infty$ queue. L_q will be small when the number of servers equals or exceeds $\rho + \sqrt{\rho}$.

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 Köllerström 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 $\rho = .875$:

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

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

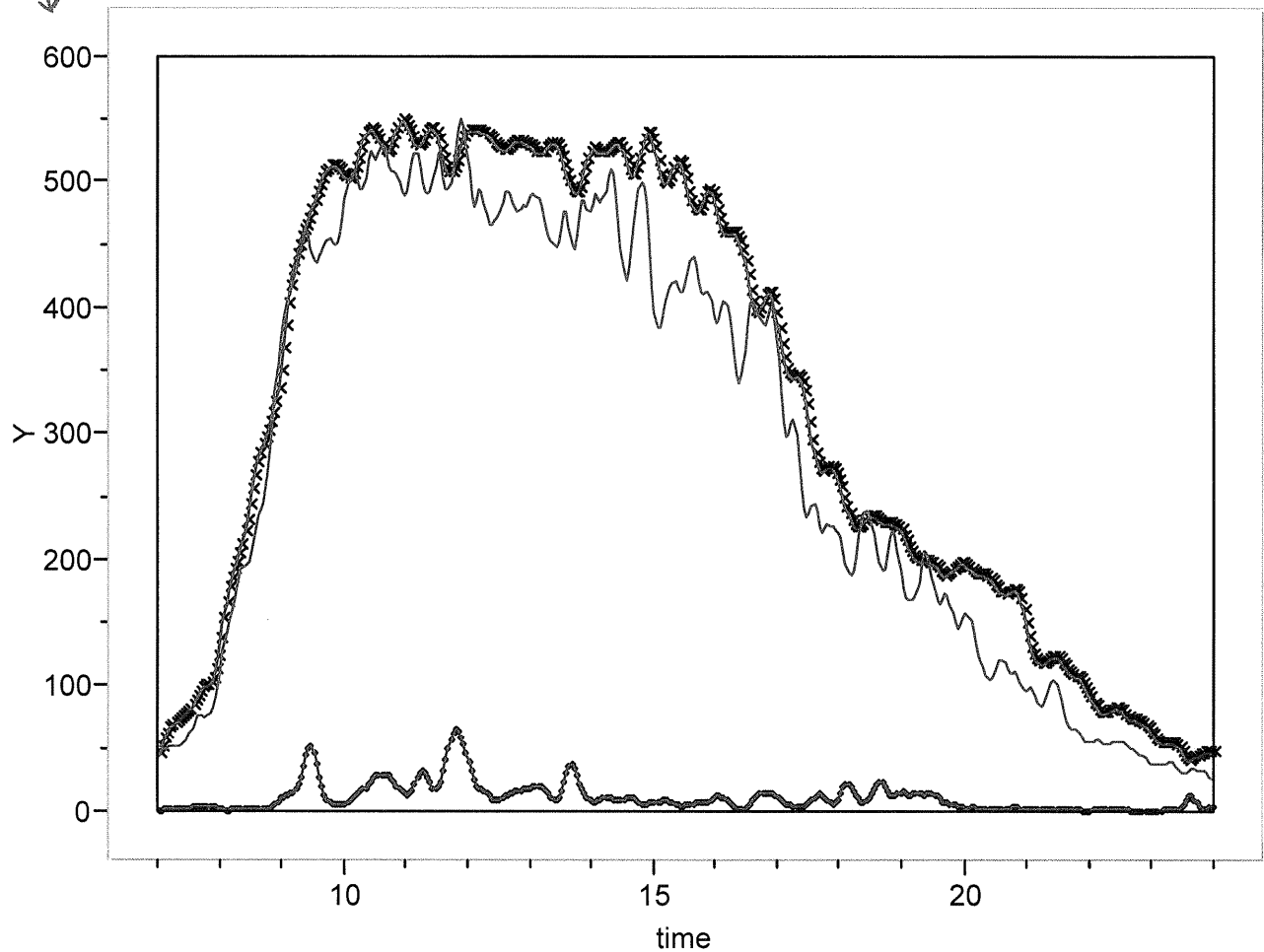
$$L_s = .206 + .875 = 1.08 \text{ customers}$$

Staffing a Large Call Center

agents



Efficiency Plots Showing Load and Staffing

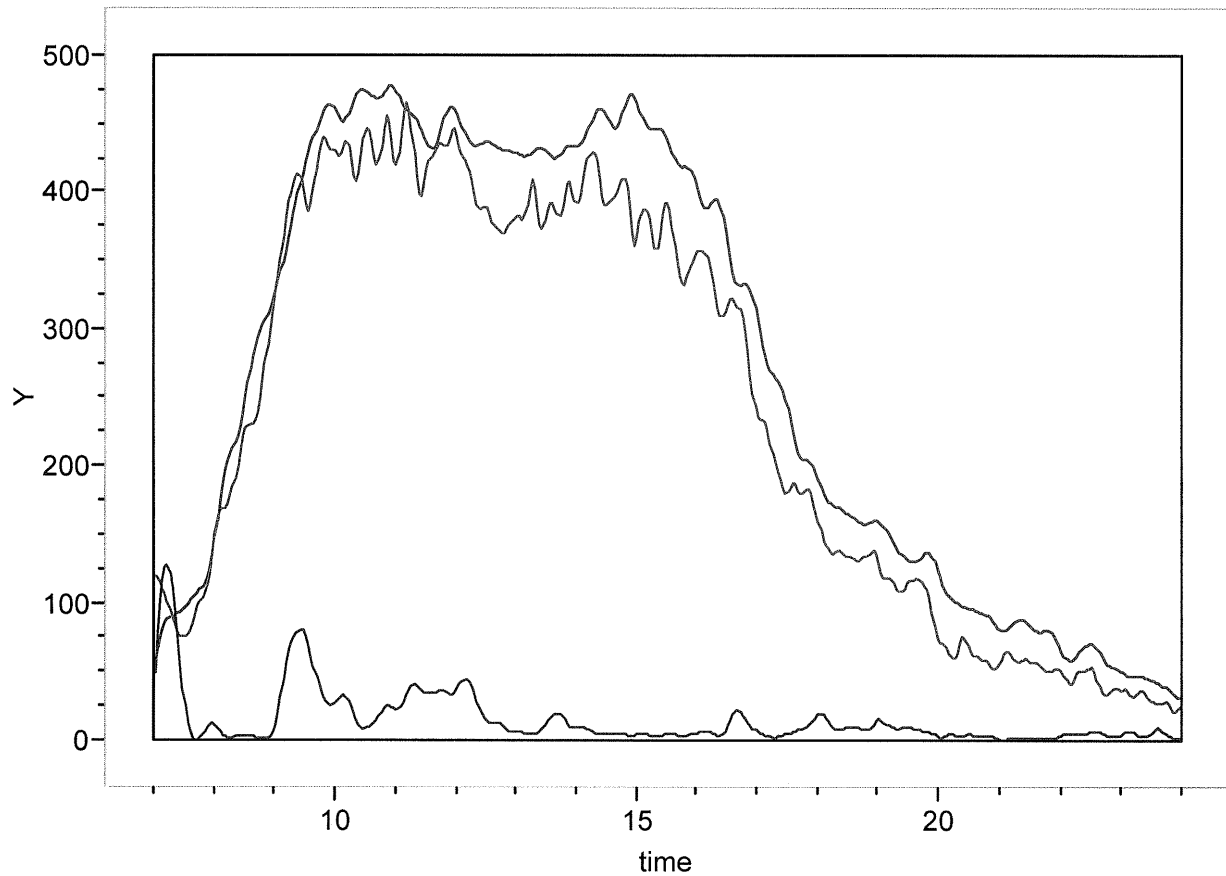


Y — NumberAgents (s)
— **load** (s)
— AvgQueueWaitAll (s)

“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



Plot is for Friday 8/02/02

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

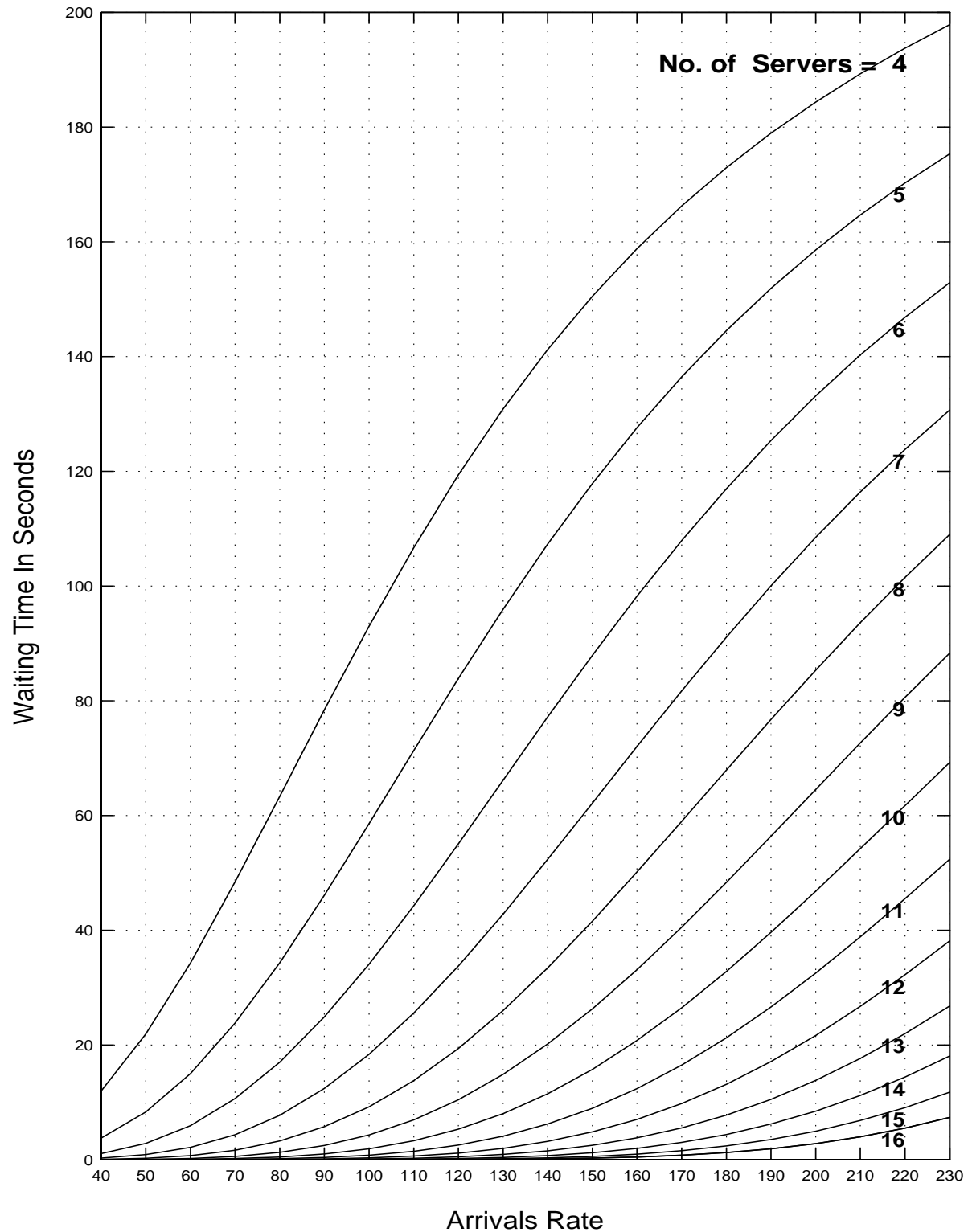
Note increased usage from 7-7:30 am (typical of Fridays).
Note increased average Queue-Wait during this time.
(Accompanied by a rise in abandonments to about 10%.)

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

Case Study: A Large Utility Company

Average Waiting Time

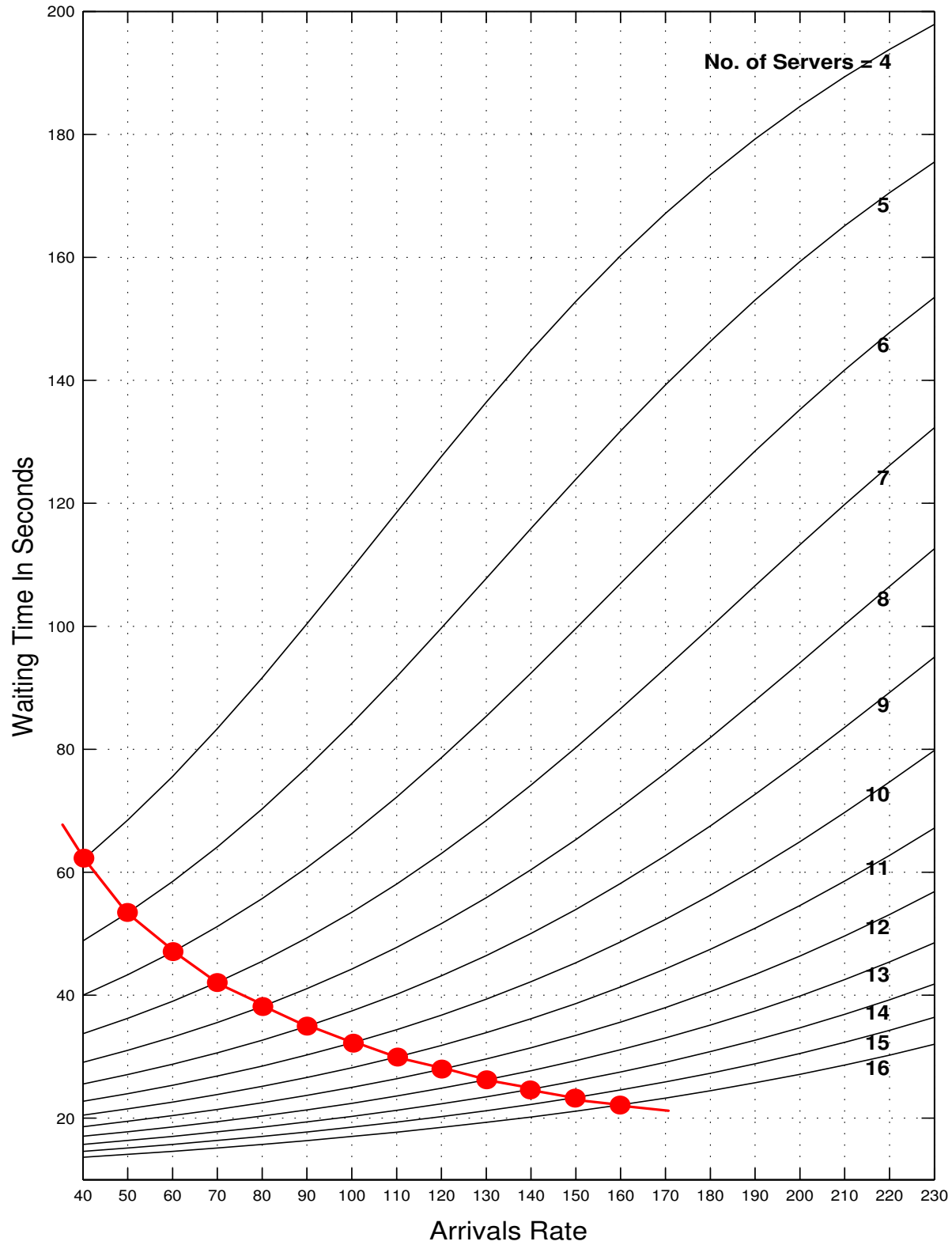
Commonly used MOP: $E(W_q)$
Total Service Time = 3.3 min.



Economies of Scale

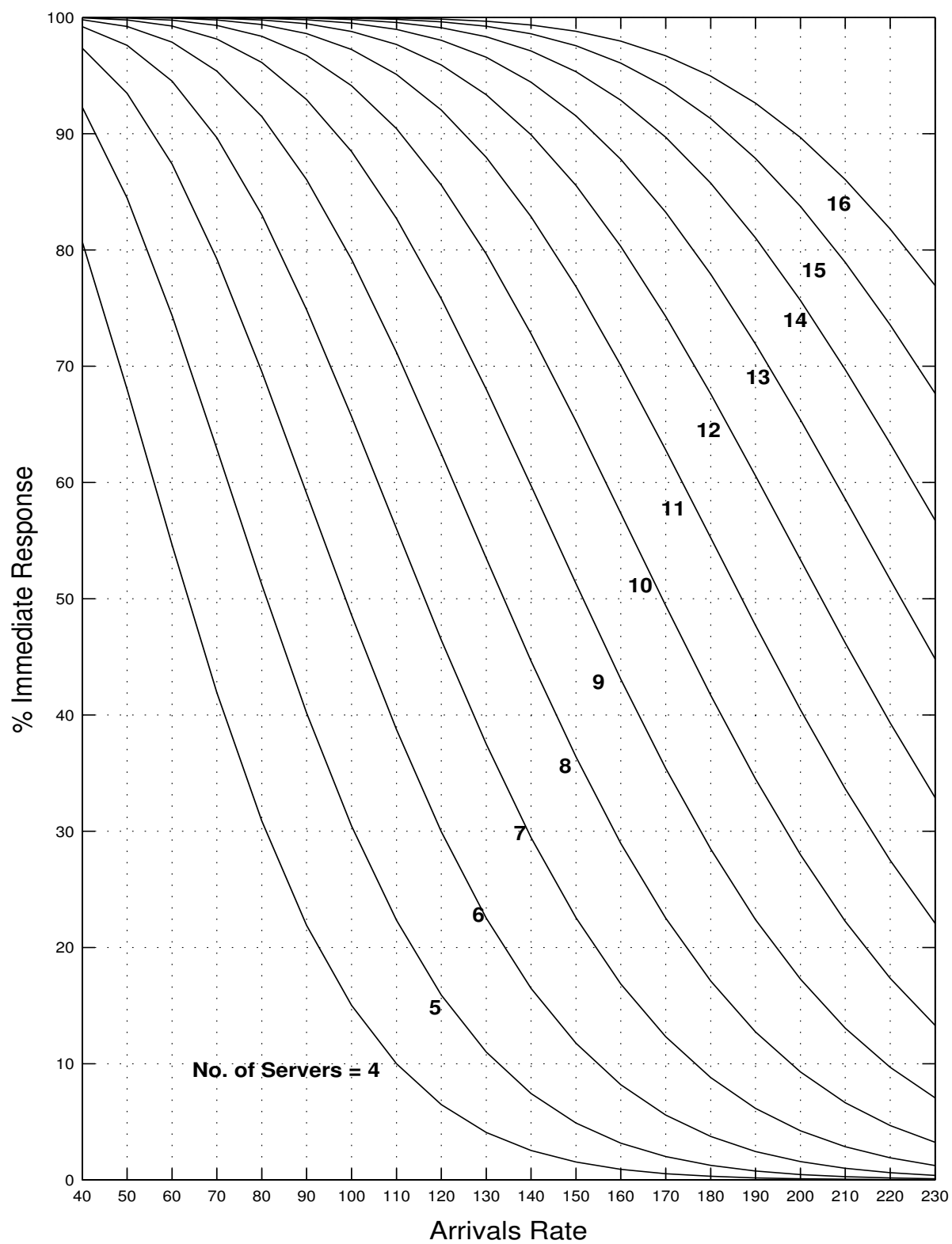
Average Waiting Time - But Only of Those Who Wait

$$E[W_q | W_q > 0] \quad (\text{Load: 10 per server})$$



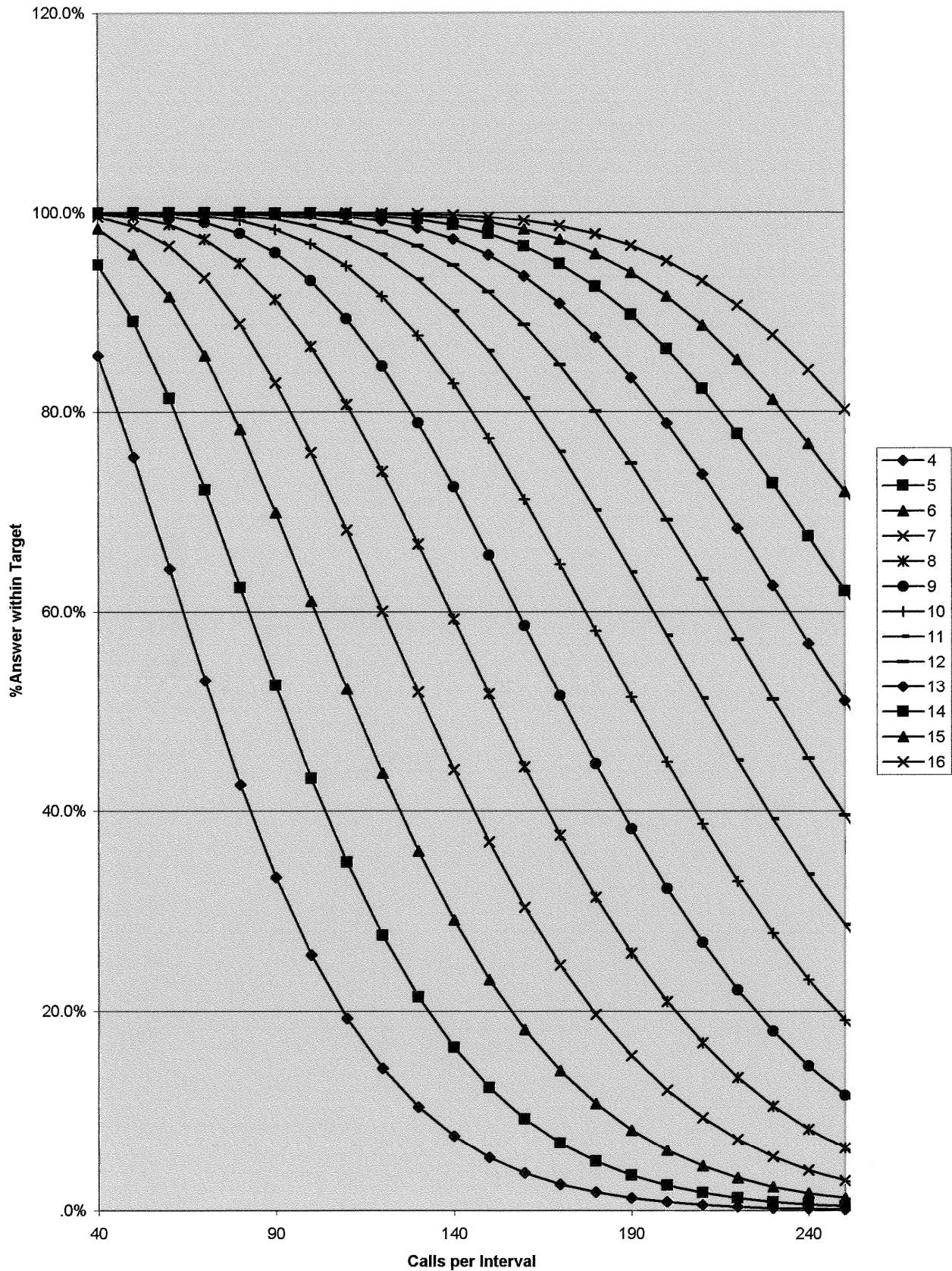
% Immediate Response (Often Not Measured)

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



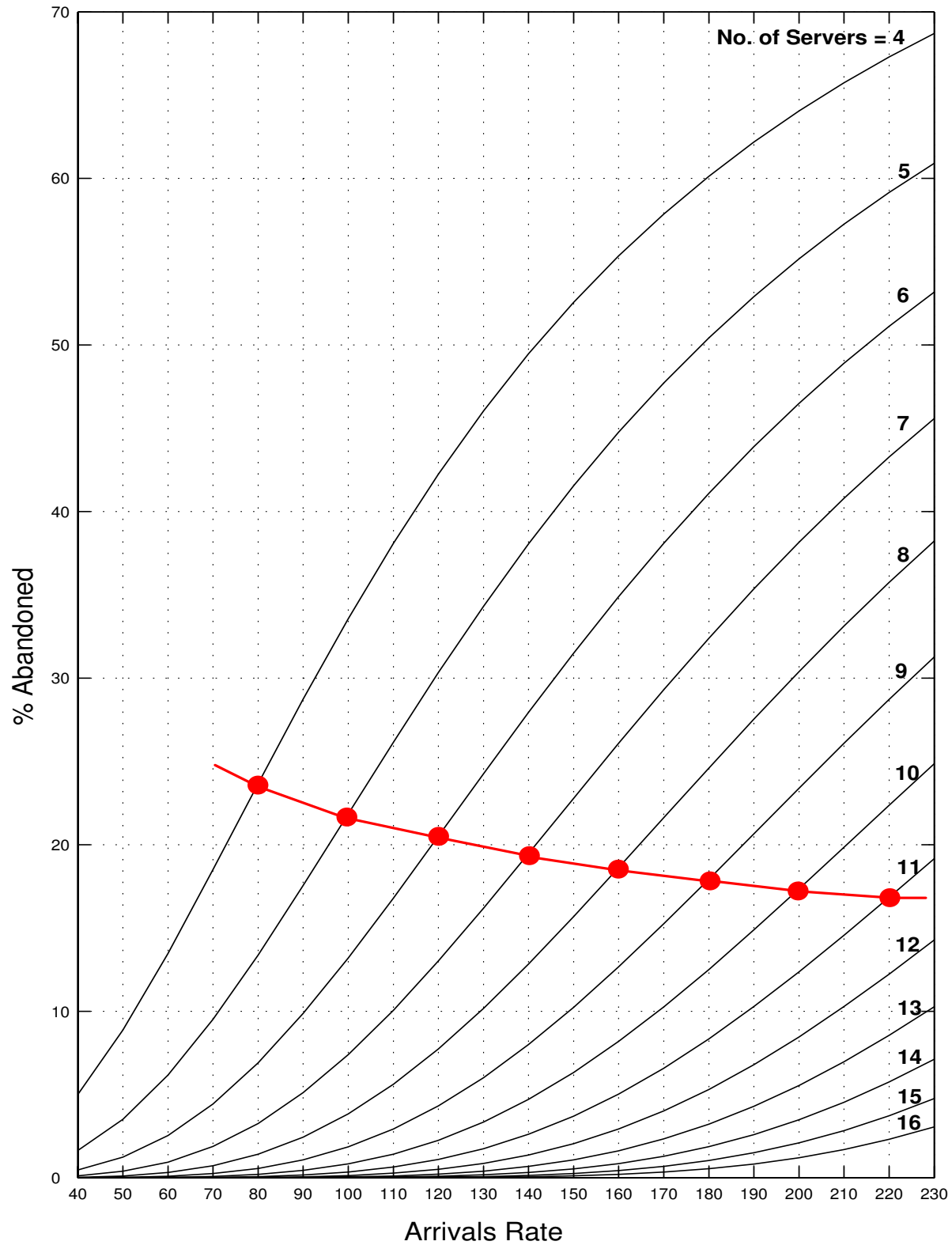
20 seconds

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

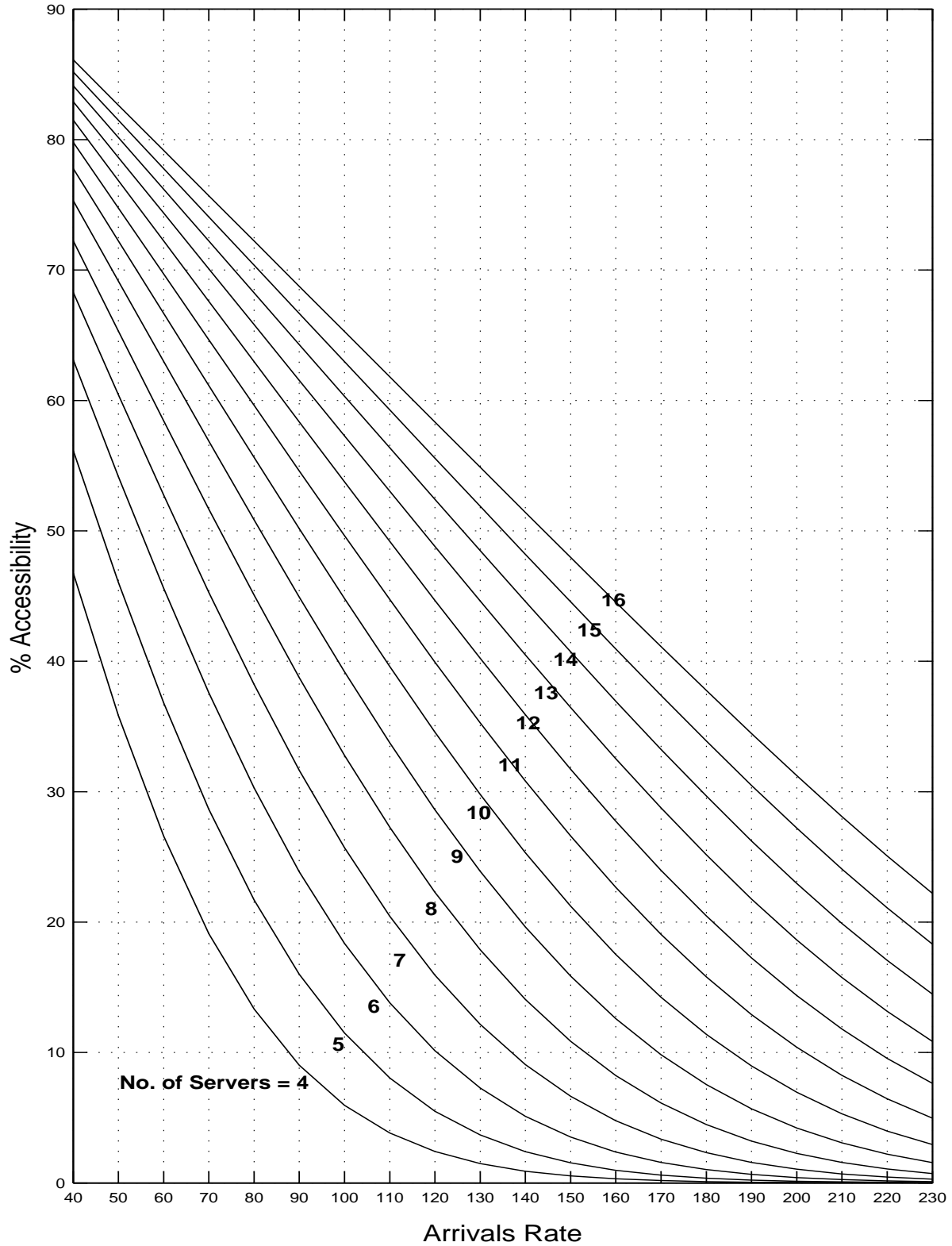


$(ES = 3:18)$
 $1/\theta = 3:30$

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



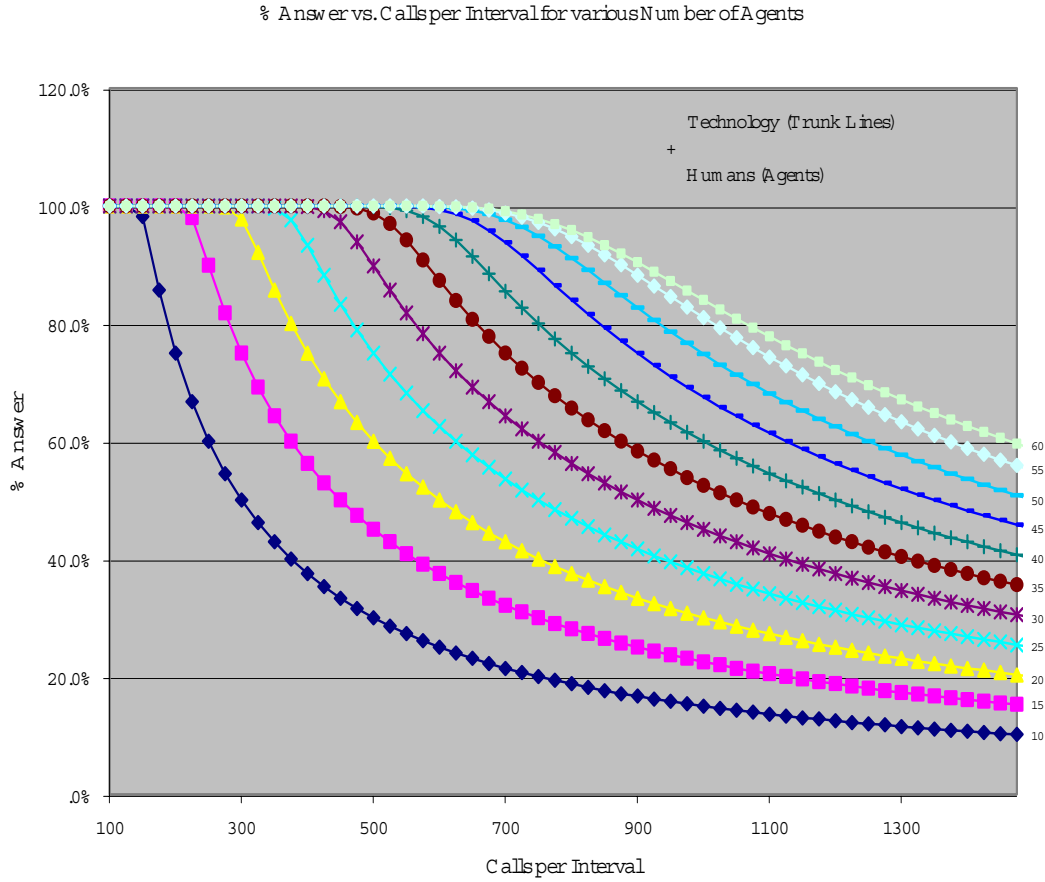
% Accessibility (Fraction of “Idle” Time)



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:

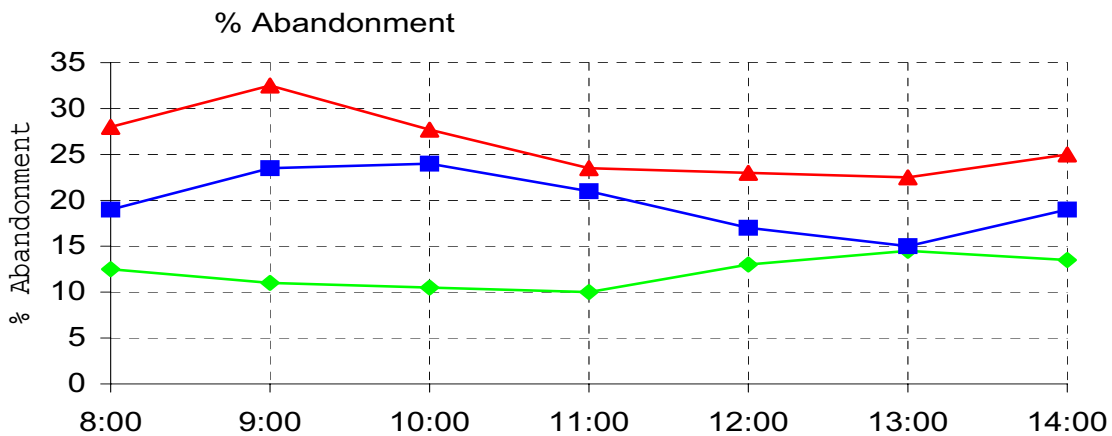
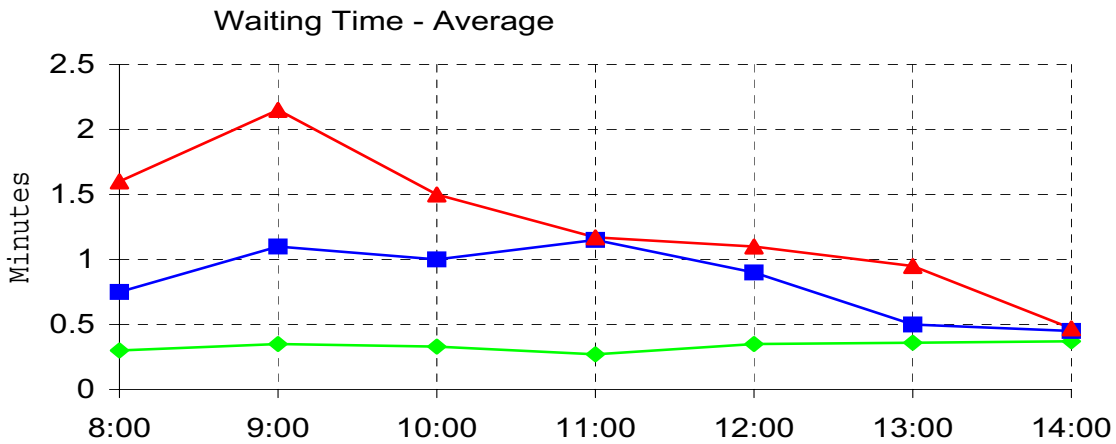
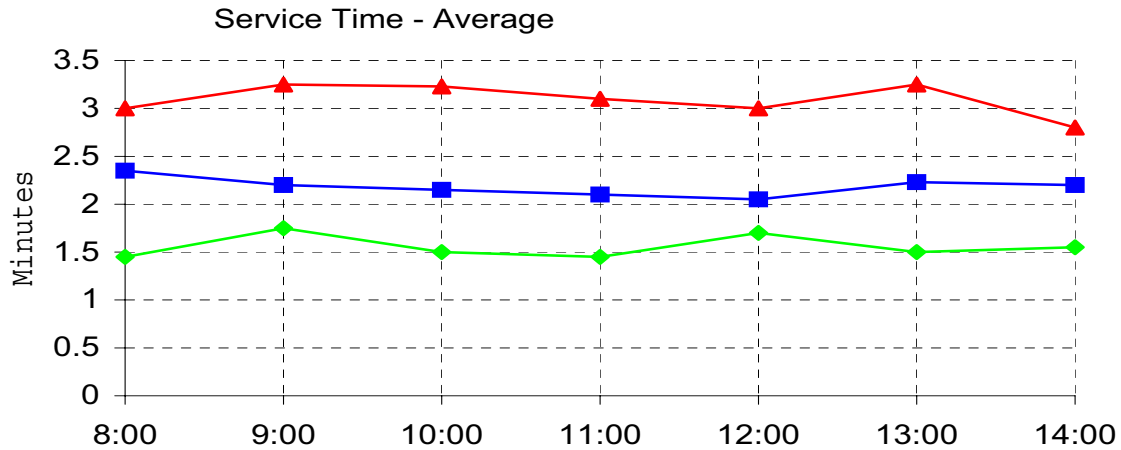
$$\text{Number of calls} = \frac{\text{Handled requests}}{100 - \% \text{ “Busy-tone calls”}} \times 100$$

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.

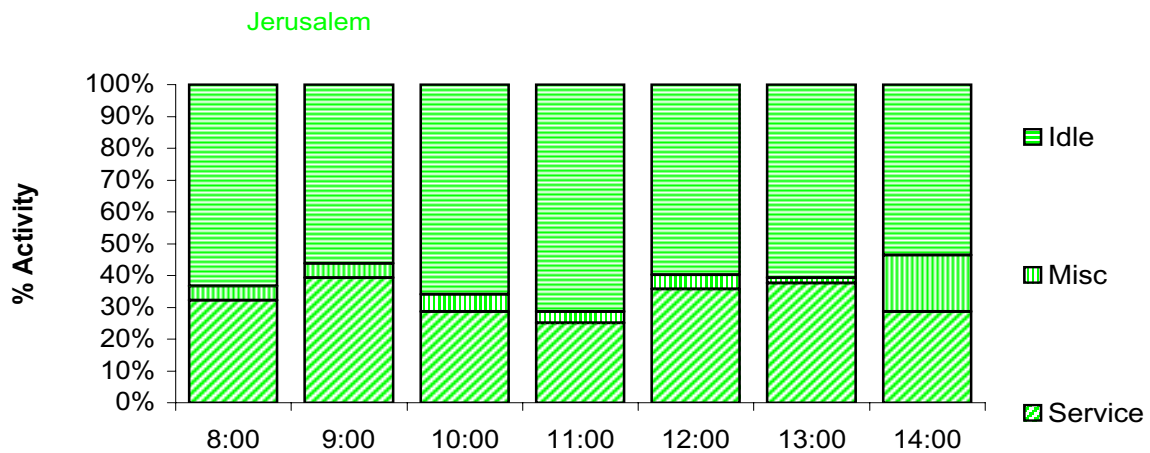
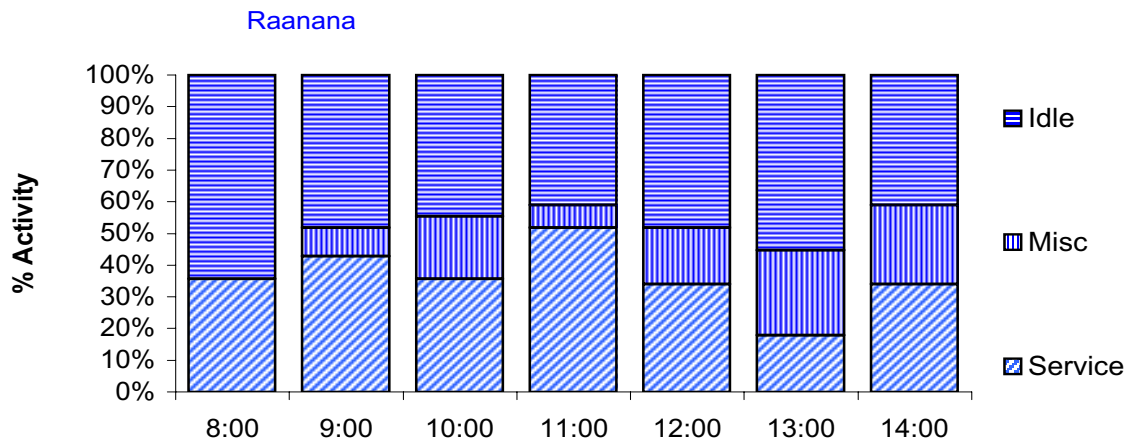
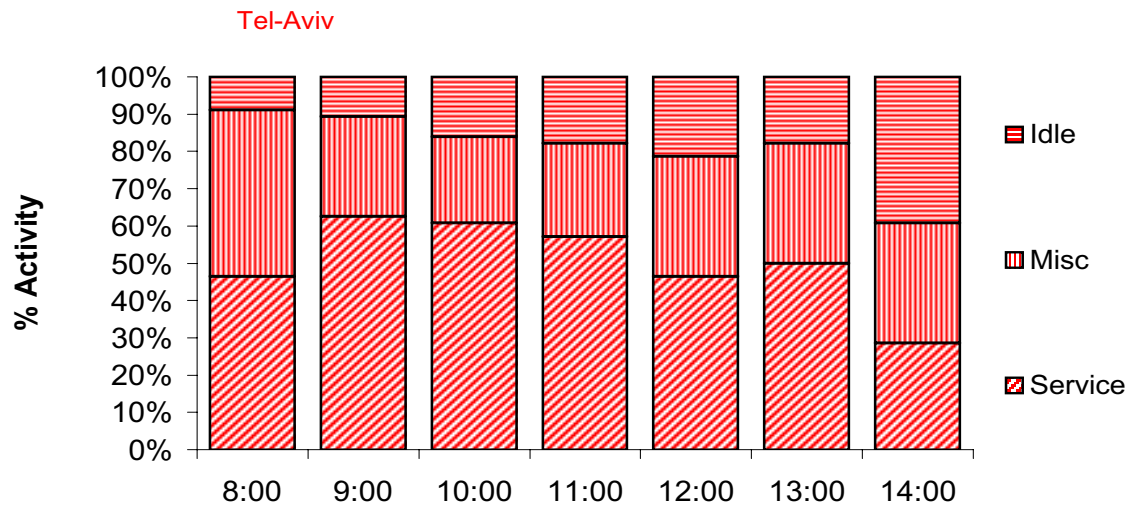
Note: In order to generate the above graph, we used an average call time of 3.8 minutes, which is 3.52 minutes (inferred from that data) multiplied by 1.08.

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

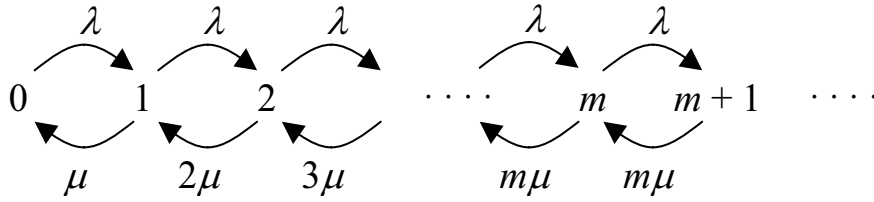
MOP's in Three Call Centers that are Doing the Same Thing !



Utilization Profiles



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



Birth & Death rates:

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

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)

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

Proof:
$$\begin{aligned} P(\text{Wait} > 0) &= \underset{\substack{\uparrow \\ \text{PASTA}}}{P(L(\infty) \geq m)} = \sum_{k \geq m} \frac{m^k}{k!} \rho^k \pi_0 \\ &= \frac{m^m}{m!} \frac{\rho^m}{1 - \rho} \pi_0. \end{aligned} \quad \text{q.e.d.}$$

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:

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

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 $\rho \uparrow 1$,

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

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:

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

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)

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

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

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

Relations (Palm, 1943?)

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

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

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

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

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

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

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.

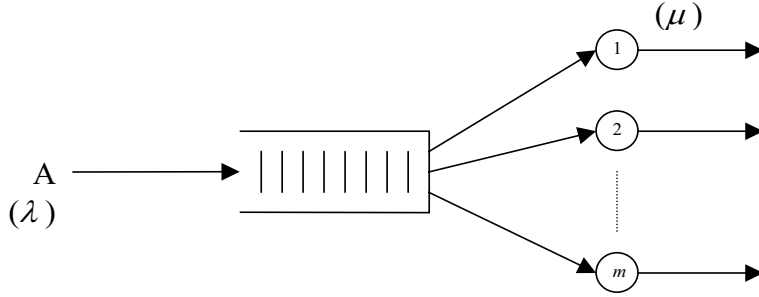
They will also be useful for us in asymptotic analysis of systems with many servers.

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: **Kingmans’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:

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

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

“Strategic” Q-Theory

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

\uparrow
 manager

\uparrow
 server

\nwarrow
 customer

- Laws of congestion: parameters

λ, C_a^2
 \uparrow
 arrivals

$;$

μ, C_s^2
 \uparrow
 services

$;$

m, b
 \uparrow
 technology

\downarrow
 Human resources

distributions: Exponential (small values, fat tails)
 (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 keeping $\rho = \text{servers' utilization fixed}$. 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.

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

Subtle example: Poisson (λm) has $CV = \frac{\sqrt{\lambda m}}{\lambda m} = \frac{1}{\sqrt{\lambda m}} \downarrow 0$, 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:

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

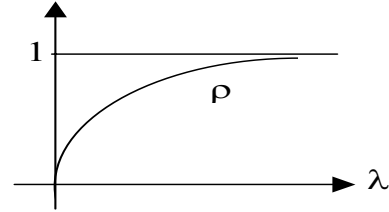
1. $\rho = \frac{\lambda}{\mu} = \frac{n\lambda}{n\mu}$, unchanged as $n \uparrow \infty$.

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

2. 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.

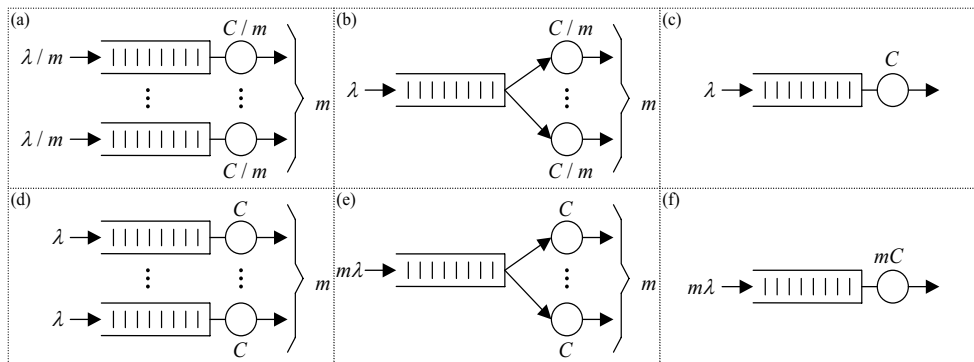
Heterogeneous customers/servers (To discuss *later*)

- (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!

$$m \times \begin{matrix} M/M/1 \\ \lambda, \mu \end{matrix} \xrightarrow{\text{scale-up}} \begin{matrix} M/M/m \\ m\lambda, \mu \end{matrix} \xrightarrow{\text{technology}} \begin{matrix} M/M/1 \\ m\lambda, m\mu \end{matrix}$$

Combine:

Saved inefficiency

queues

idleness

(1 long queue, 2 idle)

servers

lost capacity

(rate $m\mu$ at all times)

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

while $EW_s(m, \lambda, \frac{\mu}{m}) \geq EW_s(1, \lambda, \mu)$

↑
individual server's capacity

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

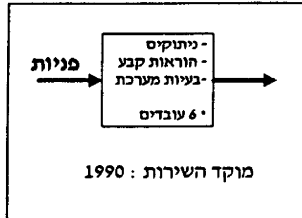
Summary (pg. 287)

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).

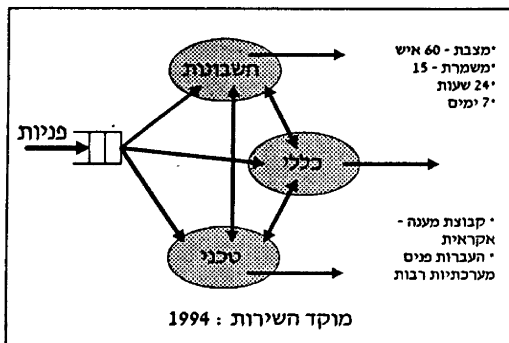
תהליך התהוותו של מוקד שירות לקוחות גדול בארץ

המוקד עליו מדובר (שמפעיליו בקשו לשמור על אנונימיות) במהותו הינו מוקד שירות לקוחות של חברת תקשורת ישראלית גדולה. חברה זו החלה בשיווק מכשירי תקשורת בישראל בשנת 1986. בשנים הראשונות לפעילות החברה, עבדה המערכת ללא מוקד שירות לקוחות. פניות לקוחות לשירות, נענו על ידי עובדי החברה הרגילים, בקווי העבודה הרגילים של החברה.

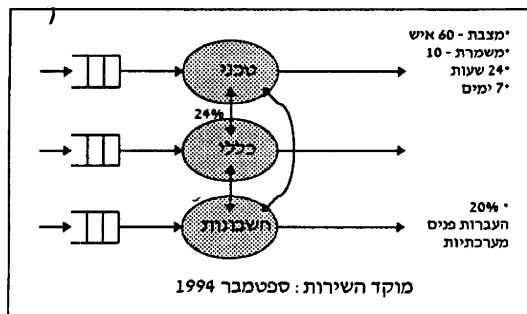


עקב גידול קהל המנויים, בינואר 1990 הוקם מוקד שירות לקוחות הראשון של החברה, בשל הצורך בשירות יותר ממוסד ומסודר. המוקד אויש בימים א'-ה' בשעות 7:00 עד 22:00 וביום ו' בשעות 7:00 עד 14:00, ונתן מענה לפניות ובקשות בנושאי ניתוקים, חוראות קבע, ובעיות מערכת. המוקד אויש על ידי 6 עובדים, ללא ACD (Automatic Call Distributor), ללא מערכת בקרה וללא תיעוד. המוקדן הרושם את הבעיה או הפעולה הנדרשת, גם היה מזרים אותו הלאה ומטפל בבקשה.

עם הזמן, וגידול במספר הפניות לשירות, גדל המוקד, והוסף ACD, עד אשר בשנת 1994 מצבת 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 עדיין דרך מספר אחד בלבד, והמתקשרים המתונו לשירות בתור יחיד. עם זאת, הונהגה במוקד חלוקה מקצועית פנימית, וזאת מכיוון שמגוון השאלות והבעיות עליהן ניתן מענה גדל משמעותית, והמוקדנים לא יכלו להתמצא בכל הנושאים. החלוקה נעשתה לשלושה תחומים:



חשבונות, נושאים טכניים, ושירות כללי (אופציות הפעלה, ניתוק, חיבור, ביטוח וכו'). המתקשר היה מנותב באקראי למוקדן כלשהו ואז היה מועבר (דרך קשר עין) למוקדן מתאים. כתוצאה, העברות פנים מערכתיות היו רבות. בשלב זה של התפתחות המוקד התפתח back office, אליו זרמו הפעולות לביצוע כפי שגרשמו על ידי המוקדן. למרות ההתפתחות של מערכת תמיכה זו, בשל

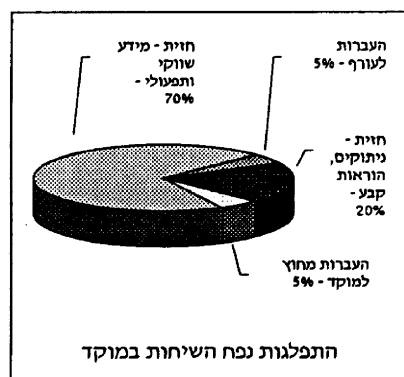


ההעברות הרבות במערכת ובזבוז המשאבים שנוצר, בספטמבר 1994 פרסם החברה שני מספרים נוספים לפניות שירות. כל אחד משלושת המספרים יועד לפניה מסוג אחד בלבד (טכני, כללי, וחשבונות). פרסום המספרים וייעודם נעשה דרך עיתון המנויים של החברה, ובספר המדריך של

החברה ללקוחות חדשים. מתקשרים שפנו למספר הלא-נכון, הועברו למחלקה המתאימה לאחר שהובהר להם המספר הנכון אליו עליהם לפנות בעתיד. עם זאת, עדיין היוו ההעברות במערכת כ-24% מכלל נפח הפעילות, ועל כן הוחלט להפסיק להעביר מתקשרים שהתקשרו למספר לא נכון. הדבר יצר בעיות בתקשורת עם הלקוחות, שכן החלוקה לא תפסה, למעט המספר הנפרד של חשבונות. העבודה נמשכה במתכונת זאת עד למרץ 1996, כאשר הוחלט כי יש צורך למצוא פתרון אחר. בנוסף, עד מרץ 1996, חל גידול משמעותי במספר המנויים ומספר השירותים הניתנים. חוסר היעילות של מבנה המוקד הנוכחי וקשיי יכולת להתמודד היו מאד ברורים למול כמות השיחות ההיסטרית שהחלה לזרום מדצמבר 1995, כשהחברה יצאה במבצע מיוחד, בה הציעה שרות תקשורת מוזל בשעות ערב והסופשבוע. במבצע זה בחרה החברה לשנות את גישתה ללקוח, ופנתה לשוק מטרה חדשה לגמרי, אתו לא התעסקה בעבר. המוקד אמנם היה ערוך למבצע מבחינת עדכוני מידע, אך לא היה ערוך למעבר ממוקד שירות למוקד טלמרקטינג. 70% מנפח פעילות המוקד עבר להיות מידע שיווקי ותפעולי, ויתר הפעילות שהוזה את מירב נפח הפעילות עד לשלב זה, הצטמצם ל-30% הנותר. המבצע הביאה מנויים חדשים משכבה סוציו כלכלית אחרת ממנויי החברה עד כה, שכן קהל היעד העיקרי של החברה בעבר היה קהל עסקי, ובמבצע זה מדובר היה בקהל יעד המעוניין במכשירי תקשורת ניידים לשימוש פרטי בשעות הפנאי. השינוי באוכלוסיית היעד גרמה גם לעומסים כבדים בשעות הערב והסופשבוע, בשונה מההתפלגות עומס השיחות עד עתה, שהתרכזה בעיקר בשעות היום. בעקבות הנלמד מתפקוד המוקד בחודשים שלאחר המבצע, נעשו שינויים מערכתיים, והמוקד עבר לפעול שוב עם תור מאוחד לשירות טכני וכללי, בשיטת חזית ועורף.

המוקד הנוכחי

המוקד הנוכחי פועל בשני תורים, האחד למבקשי שירות כללי וטכני, והשני לחשבונות. בתור החשבונות, מופעל מענה קולי (IVR משולב ACD), המאפשר למתקשר לבחור מראש את סוג השירות הנדרש לו, ועל ידי כך מועבר למקום המתאים במערכת. לקוחות המתקשרים לחשבונות, אך מעוניינים בשירות טכני או כללי, יכולים לבחור באופציה המתאימה ולעבור לתור השירות הכללי.



המוקד הכללי בנוי בצורת חזית ועורף. החזית נותנת את מירב השירותים, כגון ניתוקים וחיבורים, הוראות קבע, מידע שיווקי ותפעולי, תקלות פשוטות, וכדומה. העורף מטפל בנושאים הדורשים מומחיות וידע רב, כגון תקלות ברמה מתקדמת, מכשירים מסובכים, שיחות שבהן מבקשים מנהל או אחראי, ובירורים יותר מורכבים הדורשים בדיקה אל מול מחלקה אחרת. לעורף מגיעים כ-5% מכלל נפח השיחות, אך מבחינת כוח אדם, העורף מהווה 20%. עם זאת, השיחות המטופלות על ידי העורף מהוות יותר מ-5% מכלל זמן העבודה, וזאת מכיוון ששיחות אלו במוצע אורכות יותר זמן משיחה המטופלת בחזית.

מכשיר מרכז מוקד

חזית

עורף

מרכז חשבוני

IVR

למעגל הגבייה

מוקד השירות : מרץ 1996

השירות זרם עומס כבד של שיחות מכשיר מוגבל. מכיוון שאוכלוסיית היעד הייתה שונה, ומכיוון שהגבלות המכשיר נתנו פיתוי גדול להתקשר למספר השירות, שהינו שיחת חינם, החליטה החברה להפריד בין מוקד שירות למכשיר המוגבל לבין מוקד השירות הרגיל. שיקול נוסף היה הרצון לשמור על אפשרות למתן אופי שונה ורמה שונה של שירות בכל אחד מהמוקדים. על אף זאת, בחודש הראשון למכירה, 1 מתוך כל 3 לקוחות התקשר פעם ביום למוקד - 4500-5000 שיחות ביום רק למוקד המכשיר המוגבל. עומסים אלו גרמו לסתימת מערכת המענה הקולי והניתוב (ACD). בעקבות זאת נקנה IVR (מענה קולי אינטראקטיבי) חדש, אשר הופעל במרכזייה למוקד המכשיר המוגבל. אחת ממטרות IVR זה היא להפחית את מספר שיחות השווא המגיעות למוקדנים, ועל כן ב-IVR 9 אופציות, כאשר האחרונה מביניהם הוא להגיע לשירות קולי (מוקד).

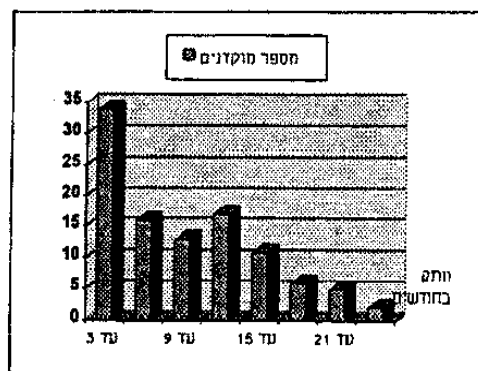
```

graph TD
    A[מנהל מוקד] --> B[מקדמי שטח]
    A --> C[כיה]
    A --> D[הבטח ובקרה]
    A --> E[המשל]
    E --> F[אח"כ שר]
    E --> G[אח"כ ח"ה]
    E --> H[מקד מבכיר מנהל]
    E --> I[מקד מיליטרי]
  
```

מבנה ארגוני

30

העבודה במוקד היא במשמרות של 6.5 שעות, כאשר יש הפסקה של רבע שעה כל שעתיים. קביעת המשמרות נעשית בעזרת תוכנה, המנבאת את הדרישה למספר המוקדנים לפי צפי שיחות. הקליטה לעבודה במוקד נעשית לאחר מבחני מיון מקצועיים וראיון אישי. העובדים הנקלטים עוברים קורס אינטנסיבי הכולל מבחנים וסימולציות. עובדים השורדים את הקורס נכנסים



לעבודה במוקד, ומוגדרים כמתלמדים למשך 3 החודשים הראשונים. לאחר חודשי ההתלמדות מקבלים העובדים החזשים ציוני ביצועים בדומה לעובדים הוותיקים, ולאחר חצי שנה של עבודה, באם עמדו בציון מינימלי, נקלטים כעובדים קבועים.

כוח האדם מורכב בעיקר מסטודנטים וחילאים/ות משוחררים/ות (כ- 30% סטודנטים, ו- 70% חיילים משוחררים). מעט מאד מהעובדים הם עובדים וותיקים וקבועים.

תהליך השירות

שירות כללי וטכני.

מנוי מתקשר למערכת באמצעות חיוג חינם ממכשיר התקשורת שברשותו, או על ידי חיוג רגיל למספר בוק. השיחה מגיעה ל- ACD, המעביר את השיחה למוקדן חזית פנוי, או בהעדר מוקדן פנוי, מכניס את השיחה לתור, עד אשר יתפנה מוקדן. מוקדן החזית מברר את בעיית הלקוח. כאשר מדובר בבקשה בה הוא מוכשר לטפל, מבצע המוקדן את הטיפול והקלדת המידע למערכת המחשב. באם מדובר בשירות, תקלה או שאלה אשר מוקדן החזית אינו מוכשר לטפל בו, מועבר המנוי למוקדן עורף, או ליחידה המתאימה בחברה (יומן התקנות, טלמרקטינג, יחסי לקוחות). אין העברת שיחות לחשבונות.

שירות חשבונות.

מנוי מתקשר למערכת בעזרת חיוג חינם מהמכשיר שברשותו. השיחה מגיעה ל- ACD המשולב IVR (מענה קולי אינטראקטיבי), ולמנוי ניתנת האפשרות לבחור בין שלוש פעולות רצופות: גביה, בירורים, ומצב חשבון. השיחה מנותבת בהתאם לבחירת המנוי: בחירת גביה מנותבת את השיחה למחלקת הגביה של החברה (העברה לתוך החברה ומחוץ למערכת המוקד), בחירת בירורים מנותבת את השיחה ל- ACD של השירות הכללי והטכני, ובחירת מצב חשבון מנותבת את השיחה לאחד ממוקדני חשבונות של המוקד, לשם בירורים על מצב חשבון המנוי; במקרה שאין מוקדן חשבונות פנוי, מנותבת השיחה לתור 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 backloads which City Bank was lead to suffer.

Strategic Q-Theory: **EOS**

QUTE Output M/M/S

$$\lambda = 8k, \quad k = 1, 2, \dots \quad (\text{i.e. } \lambda = 8, 16, 24, \dots),$$

$$1/\mu = 1/8 = 0.125,$$

$$n = k + 1, \quad k = 1, 2, \dots \quad (\text{i.e. } S = 2, 3, 4, 5, \dots),$$

λ	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)

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)

:30

Target Time: 00:30 (mm:ss).

Number of **Calls** per 60 minutes

Calls.

	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



- Current Result



- Settings



- Call Center Parameters



- Performance Indicators