$$P\{S_1\} = \frac{24}{16}P\{S_0\} = \frac{6}{19} \simeq 0.31579 \tag{5.21b}$$

$$P\{S_2\} = \frac{18}{16}P\{S_0\} = \frac{9}{38} \simeq 0.23684 \tag{5.21c}$$

$$P\{S_3\} = \frac{9}{16}P\{S_0\} = \frac{9}{76} \simeq 0.11842 \tag{5.21d}$$

$$P_Q = \sum_{i=4}^{\infty} P\{S_i\} = \frac{9}{16}P\{S_0\} \simeq 0.11842 \tag{5.21e}$$

Using these results, we can now write the hypercube equilibrium equations for nonsaturated system states:

1. **Empty state**

$$P_{000} = P\{S_0\} = \tfrac{4}{19} = 0.2105 \tag{5.22}$$

2. **Full state (no queue)**

$$P_{111} = P\{S_3\} = \tfrac{9}{76} = 0.11842 \tag{5.23}$$

3. **First hyperplane from the origin**

$$P_{001} + P_{010} + P_{100} = P\{S_1\} = \tfrac{6}{19} = 0.31579 \tag{5.24}$$

4. **Second hyperplane from the origin**

$$P_{110} + P_{101} + P_{011} = P\{S_2\} = \tfrac{9}{38} = 0.23684 \tag{5.25}$$

5. **Balance of flow about state 001**

$$P_{001}\{\underset{\substack{\uparrow\\ \text{total upward flow rate}}}{\tfrac{3}{2}} + \underset{\substack{\uparrow\\ \text{total downward flow rate}}}{1}\} = \underset{\substack{\uparrow\\ \text{flow rate from 000 to 001}}}{P_{000}(0.75)} + 1\cdot\underbrace{(P_{101} + P_{011})}_{\substack{\text{total downward flow into state 001}}} \tag{5.26}$$

$$\underbrace{\qquad\qquad}_{\text{total flow out of state 001}} \qquad\qquad \underbrace{\qquad\qquad\qquad\qquad}_{\text{total flow into state 001}}$$

6. **Balance of flow about state 100**

$$P_{100}(\tfrac{3}{2} + 1) = P_{000}(0.40) + 1\cdot(P_{110} + P_{101}) \tag{5.27}$$

7. **Balance of flow about state 011**

$$P_{011}\Big(\underset{\substack{\uparrow\\ \text{total upward flow rate}}}{\tfrac{3}{2}} + \underset{\substack{\uparrow\\ \text{total downward flow rate}}}{2}\Big) = \underset{\substack{\uparrow\\ \text{flow rate downward from state 111}}}{1\cdot P_{111}} + P_{001}\Big(\underset{\substack{\uparrow\\ \text{upward flow from unit 2's primary response area}}}{0.35} + \underset{\substack{\uparrow\\ \text{overflow from part of unit 1's primary response area}}}{0.50}\Big) + P_{010}\Big(\underset{\substack{\uparrow\\ \text{upward flow from unit 1's primary response area}}}{0.75} + \underset{\substack{\uparrow\\ \text{overflow from unit 2's primary response area}}}{0.35}\Big)$$

$$\underbrace{\qquad\qquad}_{\substack{\text{total flow out of state 011}}} \qquad\qquad \underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{\substack{\text{total flow into state 011}}}$$

$$\underbrace{\qquad\qquad\qquad\qquad}_{\substack{\text{total upward flow from state 001}}} \qquad\qquad \underbrace{\qquad\qquad\qquad\qquad}_{\substack{\text{total upward flow from state 010}}}$$

(5.28)

8. **Balance of flow about state 101**

$$P_{101}(\tfrac{3}{2} + 2) = 1\cdot P_{111} + P_{001}(0.4 + 0.25) + P_{100}(0.75 + 0.40) \tag{5.29}$$

We solve this set of equations by eliminating the following variables, in order, by use of the designated equations:

Variable Eliminated	*Using Equation Number*
P_{000}	(5.22)
P_{111}	(5.23)
P_{001}	(5.26)
P_{100}	(5.27)
P_{010}	(5.24)
P_{110}	(5.25)
P_{011}	(5.28)
P_{101}	(5.29)

After about 15 or 20 minutes with an electronic hand calculator, we arrive at the following values for the state probabilities:

$$P_{000} = 0.21053$$
$$P_{001} = 0.13669$$
$$P_{010} = 0.08863$$
$$P_{100} = 0.09047$$

$$P_{110} = 0.05301$$
$$P_{101} = 0.08894$$
$$P_{011} = 0.09489$$
$$P_{111} = 0.11842$$

5.4.4 System Performance Measures (Infinite Line Capacity)

Now that we know how to obtain the steady-state probabilities of the hypercube model, it is natural to inquire how to use these probabilities to obtain values of useful system performance measures.

Workloads. We can immediately obtain the workloads of the individual servers. The workload ρ_n of server n, which is the fraction of time that server n is busy, is equal to the sum of all steady-state probabilities having the state of server n equal to 1 (rather than 0) plus the fraction of time that a queue exists (during which time all servers are working). Thus, for our three-server example,

$$\rho_1 = P_{001} + P_{101} + P_{011} + P_{111} + P_Q = 0.5574 \quad (5.30a)$$
$$\rho_2 = P_{010} + P_{110} + P_{011} + P_{111} + P_Q = 0.4734 \quad (5.30b)$$
$$\rho_3 = P_{100} + P_{110} + P_{101} + P_{111} + P_Q = 0.4693 \quad (5.30c)$$

These results check (to four significant figures) with the requirement that the average workload $\rho = \lambda/3\mu = 0.5$. Note that the workload sharing among response units caused the workloads of the units to be more evenly distributed than the workloads of the primary response areas; if each unit served only the customers of its own response area, the workloads would have been $\rho_1 = 0.75$, $\rho_2 = 0.35$, $\rho_3 = 0.40$. In fact, it is possible for a particular primary response area to generate more work than one unit could handle, and workload sharing would facilitate the overflow (assuming, of course, that the total system is not saturated, which in this case would require that the sum of the λ_i's be less than 3).

Interatom dispatch frequencies. For virtually all non-workload-oriented performance measures it is necessary to compute

$f_{nj} \equiv$ fraction of *all* dispatches that send unit n to geographical atom j ($\sum_{n,j} f_{nj} = 1$)

Let

$E_{nj} \equiv$ set of states in which unit n is to be assigned any service request from atom j (assigning any ties arbitrarily)

For instance, in our three-server example, $E_{21} = \{001, 101\}$ and $E_{11} = \{000, 010, 100, 110\}$. If the system is in *any* state in the set E_{nj}, then the fraction of service requests that result in the dispatch of unit n to atom j is λ_j/λ. So, *excluding delays in queue*, the fraction of all dispatches that send unit n to atom j is (λ_j/λ), multiplied by the sum of probabilities of states in E_{nj}. However, unit n can also be dispatched to atom j from a queue of waiting service requests. Thus, it is convenient to write

$$f_{nj} = f_{nj}^{[1]} + f_{nj}^{[2]} \quad (5.31)$$

where $f_{nj}^{[1]} =$ fraction of all dispatches that send unit n to atom j *and* incur *no* queue delay

$f_{nj}^{[2]} =$ fraction of all dispatches that send unit n to atom j *and* incur a *positive* queue delay

As argued above, for nonsaturated states we have

$$f_{nj}^{[1]} = \frac{\lambda_j}{\lambda} \sum_{B_i \in E_{nj}} P_{B_i} \quad (5.32)$$

The term $f_{nj}^{[2]}$ is equal to the product of three terms: (1) the probability that a randomly arriving service request incurs a queue delay; (2) the conditional probability that the request originated from atom j, given it incurs a queue delay; and (3) the conditional probability that the request results in the dispatch of unit n, given that it originates from atom j and incurs a queue delay. Clearly, a queue delay will be incurred by any request arriving while all N response units are busy, and thus the first term is

$$P'_Q \equiv P_Q + P_{B_{2N-1}} \quad (5.33)$$

The second term is equal to the fraction of calls (λ_j/λ) that are generated from atom j, and is not dependent on the fact that a queue exists. To obtain the third term we use the fact the queued calls are handled in a FCFS manner, or in some other manner that ignores the location of the service request and the location of the responding unit (at the scene of the previous service request); thus, any of the N busy units is equally likely to be assigned to any particular queued request, yielding a conditional probability of $1/N$. Summarizing, we have

$$f_{nj}^{[2]} = \frac{\lambda_j P'_Q}{\lambda N} \quad (5.34)$$

and thus

$$f_{nj} = \frac{\lambda_j}{\lambda} \sum_{B_i \in E_{nj}} P_{B_i} + \frac{\lambda_j}{\lambda} \frac{P'_Q}{N} \quad (5.35)$$

Given that we now know how to calculate the f_{nj}'s from the state probabilities, we can immediately obtain several related performance measures of interest:

1. **Fraction of total dispatches that are interresponse area**

$$f_I = \sum_{n=1}^{N} \sum_{\substack{j \notin \text{response} \\ \text{area } n}} f_{nj} \tag{5.36}$$

2. **Fraction of dispatches of unit n that are interresponse area**

$$f_{I_n} = \frac{\sum\limits_{\substack{j \notin \text{response} \\ \text{area } n}} f_{nj}}{\sum\limits_{j=1}^{N_A} f_{nj}} \tag{5.37}$$

3. **Fraction of response area i requests that require other than unit i**

$$f'_{I_i} = \frac{\sum\limits_{n \neq i} \sum\limits_{\substack{j \in \text{response} \\ \text{area } i}} f_{nj}}{\sum\limits_{n=1}^{N} \sum\limits_{\substack{j \in \text{response} \\ \text{area } i}} f_{nj}} \tag{5.38}$$

We illustrate each of these computations with our three-server example. First computing the additive term associated with queued requests, we note that $P'_Q = 0.23684$ and thus $f_{nj}^{[2]} = \lambda_j(0.052631)$. Using this fact, we obtain for f_{11}, for instance,

$$\begin{aligned} f_{11} &= \frac{\lambda_1}{\lambda}(P_{000} + P_{010} + P_{100} + P_{110}) + \lambda_1(0.052631) \\ &= 0.25[0.6667(0.4426) + 0.052631] \\ &= 0.08692 \end{aligned}$$

In other words, about 8.7 percent of all service requests are the type that result in unit 1 being sent to atom 1. The full set of f_{nj}'s is displayed in Table 5-5.

Employing (5.36), the fraction of responses that are interresponse area is 0.43529. This figure checks with our intuition which might argue, roughly, that the fraction of *intra*response area responses would be equal to the average availability (50 percent) plus one-third the probability of incurring a queue delay ($P'_Q = 0.23684$), yielding a fraction of *intra*response area responses equal to 0.57895, or a fraction of *inter*response area responses equal to 0.42105. The actual figure of 0.435 is larger than that conjectured due to workload imbalances, as we will argue in Section 5.6.

TABLE 5-5 Matrix of interatom dispatch frequencies.[1,2]

	Unit Number (n)			
Atom Number (j)	*1*	*2*	*3*	*Total*
1	0.07376	0.03760	0.01581	0.1272
	0.08692	0.05076	0.02897	0.1667
2	0.07376	0.03760	0.01581	0.1272
	0.08692	0.05076	0.02897	0.1667
3	0.00944	0.03511	0.00633	0.0509
	0.01470	0.04037	0.01159	0.0667
4	0.07376	0.01482	0.03860	0.1272
	0.08692	0.02798	0.05176	0.1667
5	0.01416	0.05266	0.00949	0.0763
	0.02205	0.06056	0.01738	0.1000
6	0.00944	0.03511	0.00633	0.0509
	0.01470	0.04037	0.01159	0.0667
7	0.00957	0.00593	0.03538	0.0509
	0.01483	0.01119	0.04064	0.0667
8	0.00957	0.00593	0.03538	0.0509
	0.01483	0.01119	0.04064	0.0667
9	0.00957	0.00593	0.03538	0.0509
	0.01483	0.01119	0.04064	0.0667
10	0.00957	0.00593	0.03538	0.0509
	0.01483	0.01119	0.04064	0.0667
Total	0.2926	0.2366	0.2339	0.7632
	0.3715	0.3157	0.3128	1.000

[1] Upper figure is $f_{nj}^{[1]}$; lower figure is $f_{nj} = f_{nj}^{[1]} + f_{nj}^{[2]}$.
[2] Column and row sums may not check exactly because of rounding.

Employing (5.37), the fraction of dispatches of unit 1 that are interresponse area is 0.11077/0.3715 = 0.2982.

Employing (5.38), the fraction of responses into response area 1 that are interresponse area responses is 0.23919/0.500 = 0.4784.

Thus, we see that the unit-specific frequency of interresponse area responses can be significantly different from the response-area-specific frequency. In this case, fully (1.0 − 0.4784) × 100 percent = 52.16 percent of response area 1's workload is handled by unit 1, and thus 47.84 percent is handled by units 2 and 3. But (1.0 − 0.2982) × 100 percent = 70.18 percent of unit 1's workload is within response area 1.

Travel times. Travel time is a central performance measure computed by the hypercube model. All travel times are computed from a travel-time matrix whose generic element is τ_{ij}, the mean travel time from atom i to atom j. The numerical values of the τ_{ij}'s may reflect complications to travel such as

one-way streets, barriers, traffic conditions, and so on. Thus, the time required to travel from i to j may not be the same as the time to travel from j to i and thus we allow $\tau_{ij} \neq \tau_{ji}$. If no matrix of τ_{ij}'s is obtainable empirically for the period of time under study, then by specifying the centroid $(\bar{x}_i, \bar{y}_i)$ of each atom, one can approximate τ_{ij} to be $(|\bar{x}_i - \bar{x}_j| + |\bar{y}_i - \bar{y}_j|)/v$, where v is the effective response speed. One might wish to selectively override this equation, particularly for the case $i = j$, in which case a simple area square-root law (see Chapter 3) might be used to estimate mean intraatom travel time. For our 3-unit example, for simplicity we have set $v = 1$ unit of distance per minute and have used the right-angle distance metric (without overrides) to obtain the travel-time matrix shown in Table 5-6.

TABLE 5-6 Interatom travel-time matrix for 3-unit example.

Atom of Origin	Atom of Destination									
	1	2	3	4	5	6	7	8	9	10
1	0	3	5	2	5	7	5	5	7	7
2	3	0	2	5	2	4	8	6	10	8
3	5	2	0	7	4	2	10	8	12	10
4	2	5	7	0	3	5	3	3	5	5
5	5	2	4	3	0	2	6	4	8	6
6	7	4	2	5	2	0	8	6	10	8
7	5	8	10	3	6	8	0	2	2	4
8	5	6	8	3	4	6	2	0	4	2
9	7	10	12	5	8	10	2	4	0	2
10	7	8	10	5	6	8	4	2	2	0

To compute system mean travel times, we also require knowledge of the location of a unit when dispatched to a request. In the hypercube framework, the geographical depiction of the "location" of a response unit is general enough to model the fixed locations of ambulances and emergency repair vehicles and the mobile locations of police patrol units. This is accomplished by specifying a location matrix $L = (l_{nj})$, where l_{nj} is the *probability that response unit n is located in atom j while available for dispatch* (or, equivalently,

the fraction of available or idle time that response unit n spends in atom j). We require that L be a stochastic matrix (i.e., for all n, $\sum_{j=1}^{N_A} l_{nj} = 1$). A fixed-location unit would have $l_{nj} = 1$ for some (small) atom j and $l_{nk} = 0$ for $k \neq j$. In fact, if one wished to be precise about the fixed location, the atom j having $l_{nj} = 1$ could be defined to be a point in the city (having zero area and zero workload). A mobile location unit would most likely have several l_{nj}'s nonzero. Note that within this structure it is very natural to allow mobile units to have overlapping patrol areas; for instance, atom k would belong to overlapping patrol areas if $l_{n_1 k} \neq 0$ and $l_{n_2 k} \neq 0$ for some n_1 and $n_2 \neq n_1$. A unit n's patrol area contains all atoms j for which $l_{nj} > 0$, whereas unit n's primary response area contains all atoms for which unit n is the first preferred unit to dispatch. In certain applications a unit's patrol area and response area are identical, but in many they are not.

To obtain travel-time performance measures, it is necessary to compute

$$t_{nj} \equiv \text{mean time required for unit } n, \text{ when available, to travel to atom } j; \quad n = 1, 2, \ldots, N; \quad j = 1, 2, \ldots, N_A$$

Since unit n will be located in atom k with probability l_{nk}, we can write

$$t_{nj} = \sum_{k=1}^{N_A} l_{nk} \tau_{kj} \tag{5.39}$$

To include the case of assignments from queues, we define the mean "queued request travel time,"

$$\bar{T}_Q \equiv \sum_{i=1}^{N_A} \sum_{j=1}^{N_A} \frac{\lambda_i \lambda_j}{\lambda^2} \tau_{ij} \tag{5.40}$$

To interpret this expression, consider any particular service request which incurs a queue delay. The request is generated from atom j with probability λ_j/λ. Each one of the N busy response units (due to equal service rates) is equally likely to be dispatched to the request. The probability that the dispatched unit must travel from atom i is λ_i/λ, the fraction of region-wide workload generated from atom i, and the dispatch assignment is made independently of the particular values assumed by i and j. Thus, with conditional probability $\lambda_i \lambda_j / \lambda^2$, the travel time to a queued request will be the travel time τ_{ij} from atom i to atom j. Hence, (5.40) is the mean travel time to a request that incurs a queue delay.

The expression for the region-wide unconditional mean travel time can now be written

$$\bar{T} = \sum_{n=1}^{N} \sum_{j=1}^{N_A} f_{nj}^{[1]} t_{nj} + P_Q' \bar{T}_Q \tag{5.41}$$

If we define

$$\bar{T}_j = \text{average travel time to atom } j$$

following reasoning analogous to that above, we write

$$\bar{T}_j = \frac{\sum_{n=1}^{N} f_{nj}^{[1]} t_{nj}}{\sum_{n=1}^{N} f_{nj}^{[1]}}(1 - P'_Q) + \sum_{i=1}^{N_A} \left(\frac{\lambda_i}{\lambda}\right)\tau_{ij} P'_Q \tag{5.42}$$

If we are interested in the mean travel time to requests in a particular primary response area, we define

$$\overline{TRA}_n \equiv \text{average travel time to requests in primary response area } n$$

Assuming response area n includes at least one atom j with $\lambda_j \neq 0$, we have (following the reasoning above)

$$\overline{TRA}_n = \frac{\sum_{\substack{j \in \text{response} \\ \text{area } n}} \sum_{m=1}^{N} f_{mj}^{[1]} t_{mj}}{\sum_{\substack{j \in \text{response} \\ \text{area } n}} \sum_{m=1}^{N} f_{mj}^{[1]}}(1 - P'_Q) + \frac{\sum_{\substack{k \in \text{response} \\ \text{area } n}} \sum_{j=1}^{N_A} \lambda_j \lambda_k \tau_{jk}/\lambda^2}{\sum_{\substack{k \in \text{response} \\ \text{area } n}} \lambda_k/\lambda}(P'_Q) \tag{5.43}$$

Unfortunately, there is no known (exact) expression for the average travel time of unit n (assuming infinite queue capacity). The problem in deriving such an expression arises from the fact that the unit's position when dispatched for the first time back-to-back (with no idle time) during a system busy period is not selected from the probability distribution of request locations λ_j/λ; such a unit was most probably assigned to its current service request when *several* units were available, and thus its location tends to be near its home location (or patrol area). As an approximation, we estimate the mean travel time for unit n as follows:

$$\overline{TU}_n = \frac{\sum_{j=1}^{N_A} f_{nj}^{[1]} t_{nj} + (T_Q P'_Q/N)}{\sum_{j=1}^{N_A} f_{nj}^{[1]} + (P'_Q/N)} \tag{5.44}$$

We must recognize that the term reflecting travel time to queued requests is an overestimate which becomes asymptotically exact as the system utilization factor $\rho = \lambda/N \rightarrow 1$.

Returning to our 3-unit example, substitution into (5.40) yields a mean queued request travel time $\bar{T}_Q = 4.34$ minutes. To compute the remainder of the (conditional) mean travel times, we must specify the location matrix $L = (l_{nj})$. For simplicity in this example, suppose that $l_{14} = l_{25} = 1$ and $l_{37} = l_{38} = \frac{1}{2}$:

$$L = (l_{nj}) = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \end{bmatrix}$$

From this information, we can compute that the overall system mean travel time, from (5.41), is $\bar{T} = 3.26$ minutes. The area-specific and unit-specific mean travel times are displayed in Table 5-7.[9] Note the wide variation in atom-specific mean travel times, ranging from 1.75 minutes to 5.26 minutes, while the higher aggregation averages ($\overline{TRA}$ and $\overline{TU}$) exhibit much less variability. This type of behavior is consistent with the analytical models we studied in Chapters 2 and 3.

TABLE 5-7 Mean travel times for 3-unit example.[1]

j	*Average Travel Time to Atom j,* $\bar{T}_j$ *(minutes)*	*Average Travel Time to Response Area n,* $\overline{TRA}_n$ *(minutes)*	*Average Travel Time of Unit n,* $\overline{TU}_n$ *(minutes) [Estimated from (5.44)]*
1	3.37	3.13	3.14
2	4.29	3.30	3.18
3	5.26	3.42	3.45
4	1.75		
5	1.77		
6	3.64		
7	2.64		
8	2.31		
9	4.55		
10	4.19		

[1]Overall system mean travel time $\bar{T} = 3.258$ minutes.

5.4.5 Extensions to the Basic Hypercube Model

We have now completed our description of one basic form of the hypercube queueing model, using the 3-unit example as a means for developing the

[9]If the computed mean travel times in Table 5-7 are found to be inconsistent with the assumption that the mean total service time of each unit is the same known constant, then one must enter these travel times into a new computation for (unit-specific) mean service times and execute the model again to accomplish mean service time calibration (see Sec. 5.3.1). Several such executions may be necessary to achieve convergence.

general model structure. Briefly reviewing the restrictions placed on this form of the model, they were as follows:

1. Mean service times, including travel time, on-scene time, and related off-scene time, were identical for all servers.
2. A queue of unserviced requests was allowed to form; this queue, of potentially infinite capacity, was depleted in a FCFS manner (or other manner that ignored the location of the request, the identity of the unit, and the anticipated service time).
3. Dispatching was based on fixed preferences with no "ties."
4. Response areas were not optimized in any sense, such as was done for $N = 2$ in Section 5.3.4.

The removal of the equal-service-times assumption poses few problems. Instead of the downward transition rates on the hypercube all being equal, they would in general be different, with a downward rate for server n being μ_n (where μ_n^{-1} would be the mean service time of server n). Minor complications arise in computing system performance measures, particularly for the case of queued requests, and these are discussed in Problem 5.13.

The queue capacity of the hypercube model can, in general, assume any value from 0 to infinity. In fact, Problem 5.7 asks you to develop results analogous to (5.31)–(5.44) for the zero-capacity situation. The restriction on the options available for queue depletion remains an obstacle to obtaining more realistic geographically-based queue disciplines.

The upward transition rates on the hypercube model need not represent simple fixed preference dispatching with no ties. Allowing ties is relatively straightforward, as demonstrated in the publicly available version of the hypercube model [LARS 75b]. Of greater interest is the generalization to non-fixed-preference dispatching. The major problem encountered with state-dependent policies occurs with large N in the computation of the system performance measures and even in the initial determination of the upward transition rates [JARV 75]. Recently, the mathematics have been worked out (for efficient computer solution) for a dispatch policy that always dispatches the real-time closest unit, even if some or all of the units are non-fixed-position units [LARS 78]. This policy models a dispatcher utilizing an AVL (automatic vehicle locator) system. We ask you to develop such a model in an analytically tractable $N = 3$ setting in Problem 5.8.

The generalization of the Carter–Chaiken–Ignall optimization procedure (for optimal design of response areas) has been worked out for arbitrary N by Jarvis. The procedure is algorithmic; that is, the end result is a set of optimal response areas for a particular set of model parameters. Jarvis finds the same insensitivity of mean travel time to shifts from equal travel time boundaries to optimal state-dependent boundaries. However, the utility of the method for balancing server workloads is still high [JARV 75].

In implementing the basic model of this section (or any of its generalizations) on a computer, one immediately confronts substantial problems of computer storage and execution time. For an $N = 10$ problem, there are $2^{10} = 1{,}024$ states, and thus 1,024 simultaneous linear equations must be solved to obtain the hypercube state probabilities. The addition of one server *doubles* the size of the problem to $2^{11} = 2{,}048$ equations. For $N = 15$, $2^{15} = 32{,}768$ and the problem quickly gets out of hand, even for today's very large computers. For instance, just to store the state-transition matrix (if we did not take advantage of its special properties such as sparcity) would require $2^{15} \times 2^{15} = 2^{30} = 1{,}073{,}741{,}824$ elements of storage. Thus, as discussed in [LARS 74a], much time has been invested in using the special structure of the hypercube model to derive computer-efficient means for developing and solving its equations. We expect to see more work in this direction in the future, that is, the use of queueing theory to develop large sets of simultaneous equations which have a special structure that can be exploited for efficient computer solution. Much work in queueing networks has already had this flavor [KLEI 75, 76; FRAN 71].

5.5 HYPERCUBE APPROXIMATION PROCEDURE (INFINITE LINE CAPACITY)

As just discussed, the storage and execution time required to solve the hypercube model equations roughly doubles with each additional server. Thus, one is motivated to find an efficient approximation procedure that computes to reasonable accuracy all the hypercube performance measures. Such a procedure has been developed that requires solution to only N simultaneous equations rather than 2^N, as in the exact model [LARS 75a].

The approximation has been found to compute results to within 1 or 2 percent of the exact results. In most practical situations such approximate solutions should suffice. For instance, data inaccuracies may not justify use of a highly precise model. Or the system planner may not have access to a sophisticated computer system necessary to perform calculations with the exact model. Or certain nonquantifiable concerns, perhaps involving political, legal, spatial, or administrative constraints, may play an important role in system design, thereby making precise estimates of quantifiable performance measures unnecessary. Finally, the approximation procedure can be carried out on a computer for any reasonable number of servers N, whereas the exact model is impractical for N greater than 15.

In this section we will simply sketch out the key ideas of the approximation procedure, as applied to the basic infinite queue capacity hypercube

model of the previous section. Full details are found in [LARS 75a]. In complex urban service systems, we expect such approximations to play an increasingly important role in bringing computationally practical models to the aid of the urban decision maker.

The following are the main features of the approximation procedure:

1. As with the exact model, one assumes that the dispatcher has a rank-ordered list of preferred units to dispatch to service requests from each geographical atom and that he always dispatches the most preferred available (free) unit (i.e., fixed-preference dispatching).
2. In addition, if ρ_n is the fraction of time that unit n is busy, we use $(1 - \rho_n)$ as the probability that unit n will be available to be dispatched to a service request when all units more preferred than unit n are busy. A correction factor is used as a multiplier to make this approximation as exact as possible. For instance, if unit 5 is the third preferred for a particular request, and units 7 and 2 are preferred to unit 5, the probability that unit 5 will be dispatched is equal to $\rho_7 \cdot \rho_2 \cdot (1 - \rho_5) \cdot$(correction factor).
3. The correction factor, which can deviate significantly from 1, is derived to account for the fact that the statuses of servers are not independent (as would be assumed if the correction factor were always 1).
4. Given features 2 and 3, we can write N simultaneous equations relating the N unknowns (the workloads) to the dispatch policy and the arrival rates from the various geographical atoms.
5. The N simultaneous equations are solved iteratively, thereby yielding estimates of the workloads of the units.
6. If we desire other performance measures of the system (e.g., the mean travel time to each geographical atom or the fraction of dispatches that are interresponse area), then the values of the utilization factors found in feature 5 may be used to estimate the fraction of dispatches that send unit n to atom j, for all n and j. These fractions are then entered into simple equations to obtain estimates of the values of the desired performance measures.

The derivation of the approximation procedure is carried out assuming the $M/M/N$ infinite-capacity model with indistinguishable (homogeneous) servers. As usual, the arrival rate is λ, the mean service time (for each server) is μ^{-1}, and $\rho \equiv \lambda/N\mu < 1$. The approximation arises from the fact that in the context of an urban service system, the servers are distinguishable and thus have differing performance characteristics.

5.5.1 Correction Factor

In an $M/M/N$ queueing system with infinite queueing capacity, suppose that we start randomly *sampling* servers in the system until we find the first server who is available or free (if there is one). Let

$$B_j \equiv \text{event that } j\text{th selected server is busy (not available)}$$

$$F_j \equiv B_j^c = \text{event that } j\text{th server selected is free (or available)}$$

$$P\{B_1 B_2 \ldots B_j F_{j+1}\} \equiv \text{probability that the first free server is the } (j+1)\text{st server selected}$$

The server selection process here is a strictly random sampling without replacement. In effect, we are selecting servers in a "blindfolded" manner.

Now the probability $P\{B_1 B_2 \ldots B_j F_{j+1}\}$ corresponds, in an urban services context, to a *dispatch probability*, that is, the probability of assigning the $(j+1)$st preferred server to a service request. If the nth server selected, say server s_n, is busy a fraction of time ρ_{s_n}, then a naive approach would be to assume that servers operate independently. In that case, we would have

$$P\{B_1 B_2 \ldots B_j F_{j+1}\} \stackrel{?}{=} \rho_{s_1}\rho_{s_2} \ldots \rho_{s_j}(1 - \rho_{s_{j+1}}) \qquad j < N$$

$$P\{B_1 B_2 \ldots B_N\} \stackrel{?}{=} \prod_{i=1}^{N} \rho_{s_i}$$

We can see by inspection that this is mathematically incorrect since $P\{B_1 B_2 \ldots B_N\}$ corresponds to the probability of saturation of the $M/M/N$ system (i.e., all servers are simultaneously busy) and, as shown by (4.44), this is not equal to the product of utilization factors.

We can still utilize the multiplicative concept if we incorporate a suitable "correction factor." To do this, we write

$$P\{B_1 B_2 \ldots B_j F_{j+1}\} = P\{F_{j+1} | B_1 B_2 \ldots B_j\} P\{B_j | B_1 B_2 \ldots B_{j-1}\} \ldots P\{B_1\} \qquad (5.45)$$

Multiplying both numerator and denominator of the right-hand side by $\rho^j(1 - \rho)$, we have

$$P\{B_1 B_2 \ldots B_j F_{j+1}\} = \left\{\left[\frac{P\{F_{j+1} | B_1 B_2 \ldots B_j\}}{1 - \rho}\right]\left[\frac{P\{B_j | B_1 B_2 \ldots B_{j-1}\}}{\rho}\right] \ldots \left[\frac{P\{B_1\}}{\rho}\right]\right\} \rho^j(1 - \rho) \qquad (5.46)$$

If we define $Q(N, \rho, j)$ to be the term preceding $\rho^j(1 - \rho)$, we have

$$P\{B_1B_2 \ldots B_jF_{j+1}\} = Q(N, \rho, j)\rho^j(1 - \rho) \tag{5.47}$$

The factor $Q(N, \rho, j)$ indicates the extent to which the result of the server-independence argument must be "corrected" to obtain the exact result. Each of the $j + 1$ terms in the product comprising $Q(N, \rho, j)$ can be considered to be a separate correction factor indicating the relative amount by which ρ or $1 - \rho$ overestimates (or underestimates) the respective conditional probabilities of being busy or free. As shown in Problem 5.9, the exact expression for $Q(N, \rho, j)$ can be derived by conditioning on each possible number of servers busy in an $M/M/N$ system, and then combining results by laws of conditional probability. The result is

$$Q(N, \rho, j) = \frac{\sum_{k=j}^{N-1} \left\{ \frac{(N - j - 1)!\,(N - k)}{(k - j)!} \right\} \frac{N^k}{N!} \rho^{k-j}}{(1 - \rho)\left[\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i\right] + \frac{N^N\rho^N}{N!}}$$

$$j = 0, 1, \ldots, N - 1 \tag{5.48}$$

In checking the result for a limiting case, we find that $Q(N, \rho, 0) = 1$, indicating that the probability that the first selected server is free is exactly $1 - \rho$ (a result in agreement with intuition, since $\rho = \lambda/N\mu$ is the average fraction of time a server is busy). Additional work shows that $Q(N, \rho, 1) < 1$, which implies that $\rho(1 - \rho)$ is an overestimate of $P\{B_1F_2\}$. This is true since, given that the first selected server is busy, the probability $P\{F_2 \mid B_1\}$ that the second selected server is free is less than $(1 - \rho)$, the value predicted by the independence argument. Here, discovering that the first selected server is busy shifts the system state probabilities in the direction of busier states. One can show that if

$$\rho > 1 - \frac{2}{N} \qquad (N \geq 2) \tag{5.49}$$

then $Q(N, \rho, j)$ is a monotonically decreasing function of j; otherwise, it is a unimodal function of j, reaching a minimum for some value of j, say j^0, and then increasing for all j greater than j^0. For illustrative purposes, plots of $Q(8, \rho, j)$ are given in Figure 5.18.

Even though the correction factor $Q(N, \rho, j)$ was derived assuming homogeneity of servers, one could conjecture that the same factor could be used as an approximation in an $M/M/N$ system with distinguishable servers. This would seem especially appropriate for systems in which the workloads of servers do not differ too greatly and in which many different dispatch preference vectors (motivated by different customer locations) effectively simulate

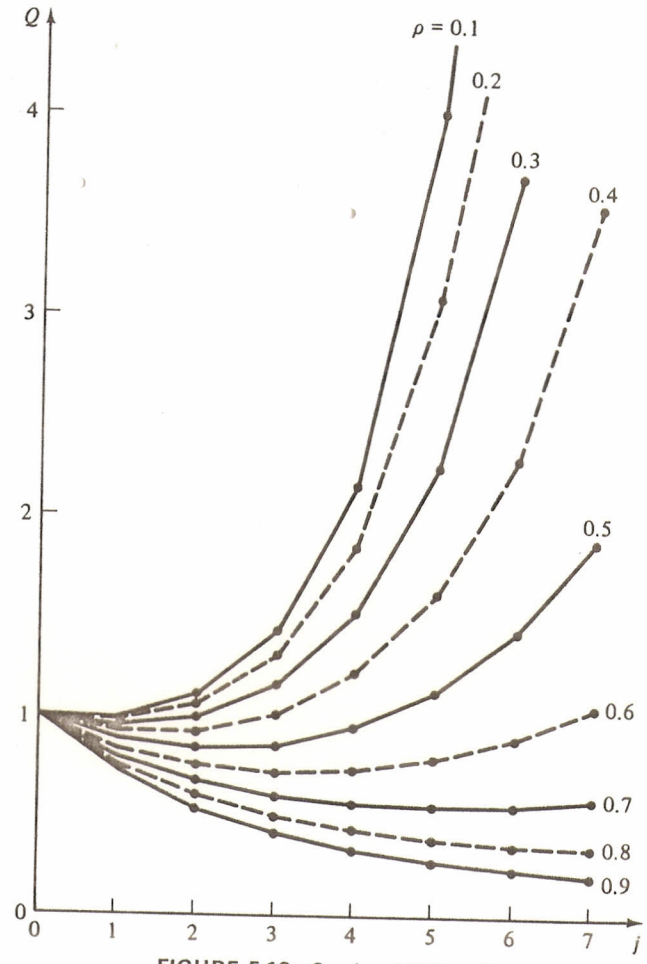


FIGURE 5.18 Graphs of $Q(8, \rho, j)$.

in the aggregate "blindfolded" selection of servers. Implementing this conjecture has yielded numerical results that are typically within 1 or 2 percent of the exact "hypercube" results.

5.5.2 Workload Estimation

The last conjecture brings us to feature 4 of the entire approximation procedure, in which we obtain a set of N simultaneous equations relating the N unknown ρ_n's to the dispatch policy and the arrival rates from the various geographical atoms. Suppose that we again equate the mean service time to unity (i.e., $\mu^{-1} = 1$), thus making the unit of time equal to one mean service time unit. Then, the net *rate* R_n at which server n is assigned to customers,

measured over a very long time period, is equal to the workload ρ_n of server n. Thus, if we can determine a value for R_n, we have determined a value for ρ_n. We describe mathematically the dispatch preference policy by defining the following:

$G_n^j \equiv$ set of geographical atoms for which server n is the jth preferred dispatch alternative, $n, j = 1, 2, \ldots, N$

$m_{aj} \equiv$ identification number of the jth preferred server for atom a

For instance, if $G_2^3 = \{7, 8, 10, 12\}$, server 2 is the third preferred dispatch alternative for atoms 7, 8, 10, 12 and, for instance, $m_{73} = 2$. Now an exact expression for R_n can be written as follows:

$$\begin{aligned} R_n = &\sum_{a \in G_n^1} \lambda_a (1 - \rho_n) + \sum_{a \in G_n^2} \lambda_a P\{B_{m_{a1}} F_n\} \\ &+ \sum_{a \in G_n^3} \lambda_a P\{B_{m_{a1}} B_{m_{a2}} F_n\} + \cdots \\ &+ \sum_{a \in G_n^N} \lambda_a P\{B_{m_{a1}} B_{m_{a2}} \cdots B_{m_{a(N-1)}} F_n\} + \lambda_D \end{aligned} \tag{5.50}$$

where the term λ_D accounts for delayed dispatches from a queue, which are apportioned equally among all N servers; thus,

$$\lambda_D = \frac{\lambda P_Q'}{N} \qquad \text{(assuming an infinite-capacity queue)}$$

In words, (5.50) states that the net rate at which server n is assigned to customers is equal to (1) the rate at which customers arrive from server n's primary response area, multiplied by the fraction of time server n is available to start servicing such customers immediately; *plus* (2) the rate at which customers arrive from server n's second-preferred areas, multiplied by the fraction of time that *both* the first preferred server is busy and server n is available to start servicing such customers immediately; *plus* (3) similar terms for the remaining j-preferred areas; *plus* (4) a term that apportions equally to all servers a fraction $(1/N)$ of customers delayed in queue.

Given our approximation conjecture, we can simplify (5.50) by estimating the dispatch probabilities as products of utilization or availability factors and the appropriate correction term. For instance, we approximate $P\{B_3 B_6 F_5\} \simeq Q(N, \rho, 2) \cdot \rho_3 \cdot \rho_6 \cdot (1 - \rho_5)$. Given this approximation and substituting $R_n = \rho_n$, we can rewrite (5.50) as

$$\begin{aligned} \rho_n = &\sum_{a \in G_n^1} \lambda_a (1 - \rho_n) + \sum_{a \in G_n^2} \lambda_a Q(N, \rho, 1) \rho_{m_{a1}} (1 - \rho_n) \\ &+ \sum_{a \in G_n^3} \lambda_a Q(N, \rho, 2) \rho_{m_{a1}} \rho_{m_{a2}} (1 - \rho_n) + \cdots \\ &+ \sum_{a \in G_n^N} \lambda_a (N, \rho, N - 1) \rho_{m_{a1}} \rho_{m_{a2}} \cdots \rho_{m_{a(N-1)}} (1 - \rho_n) \\ &+ \lambda_D \end{aligned} \tag{5.51}$$

We notice that each of the summation terms on the right-hand side of (5.51) contains $(1 - \rho_n)$ as a factor. If we remove this factor, then the sum of these N terms is not a function of ρ_n, a desirable property in an iterative procedure whose purpose is to compute an estimate for ρ_n.

$$\begin{aligned} R_n^F \equiv &\sum_{a \in G_n^1} \lambda_a + \sum_{a \in G_n^2} \lambda_a Q(N, \rho, 1) \rho_{m_{a1}} + \sum_{a \in G_n^3} \lambda_a Q(N, \rho, 2) \rho_{m_{a1}} \rho_{m_{a2}} + \cdots \\ &+ \sum_{a \in G_n^N} \lambda_a Q(N, \rho, N - 1) \rho_{m_{a1}} \rho_{m_{a2}} \cdots \rho_{m_{a(N-1)}} \end{aligned} \tag{5.52}$$

has an intuitive interpretation. It is the (approximate) customer assignment rate for server n during periods when server n is free. Simple manipulation of (5.51) yields

$$\rho_n = \frac{R_n^F + \lambda_D}{1 + R_n^F} \qquad n = 1, 2, \ldots, N \tag{5.53}$$

Here we assume that $\lambda_D < 1$, implying that $\rho_n < 1$; otherwise, the system would be overloaded. Equation (5.53) [together with (5.52)] represent a set of N simultaneous nonlinear equations relating the N unknowns ρ_n, $n = 1, 2, \ldots, N$. They can be solved iteratively to obtain values for each of the ρ_n's. A specific iterative algorithm is displayed in Figure 5.19. As you will discover in Problem 5.12, (5.53) yields a very reasonable approximation for the workload even on the first iteration (in which all workloads on the right-hand side of the equation are set equal). Rarely are more than three or four iterations required for even very stringent convergence criteria.

5.5.3 Return to the Three-Server Example

We apply the workload approximation procedure to the three-server, 10-atom example of the previous section. Recall that $\lambda = 1.5$, $\mu = 1$, implying that $\rho = \lambda/(3 \times 1) = 0.5$. To initiate the procedure we need to specify the dispatch preference sets G_n^j, and these are shown in Table 5-8. From (5.48)

TABLE 5-8 Matrix of G_n^j [1] sets.

	Unit Number (n)		
Preference Number (j)	*1*	*2*	*3*
1	1, 2, 4	3, 5, 6	7, 8, 9, 10
2	3, 5, 6, 7, 8, 9, 10	1, 2	4
3	—	4, 7, 8, 9, 10	1, 2, 3, 5, 6

[1] $G_n^j \equiv$ set of geographical atoms for which server n is the jth preferred dispatch alternative, $n, j = 1, 2, 3$.

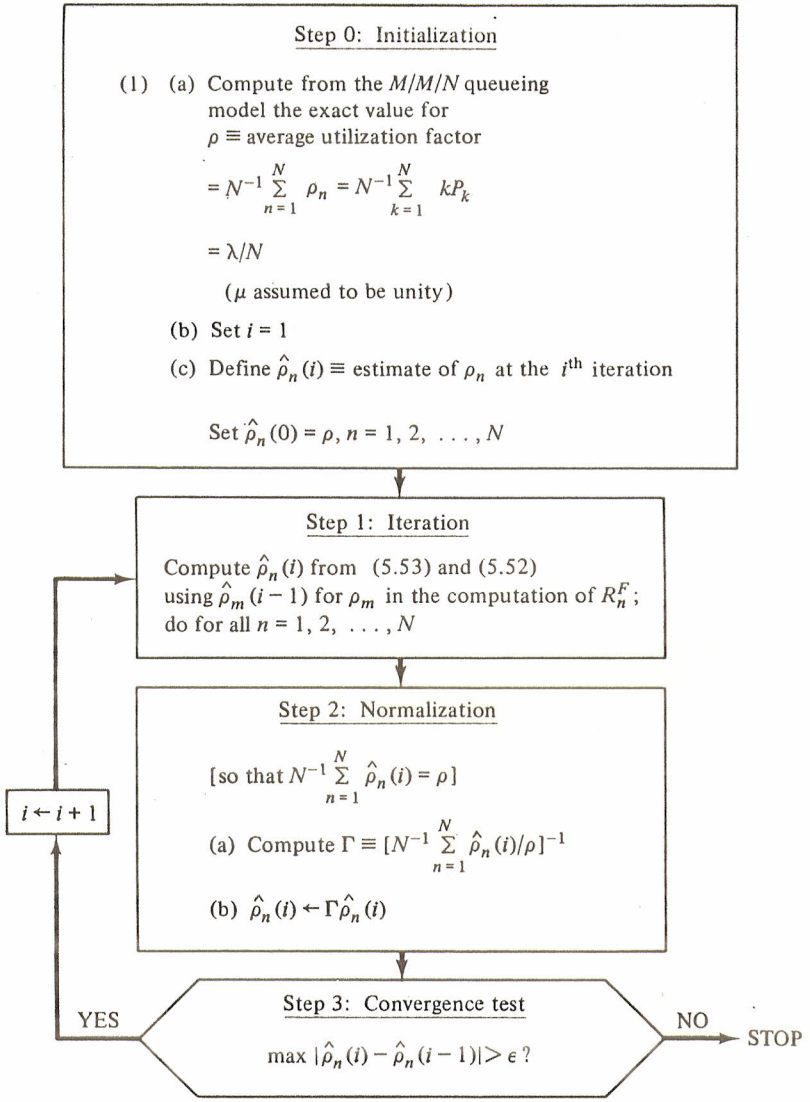


FIGURE 5.19 Algorithm for Estimating the N workloads.

we compute the numerical values for the three required correction factors:

$$Q(3, \tfrac{1}{2}, 0) = 1$$
$$Q(3, \tfrac{1}{2}, 1) = 0.73684$$
$$Q(3, \tfrac{1}{2}, 2) = 0.63158$$

We are now ready to proceed through the steps of the algorithm. One complete iteration of the algorithm is shown in each of Figures 5.20–5.22. As can

Step 0: Initialization

(a) $r = \frac{1}{2}$ $(\lambda_D = 0.1184)$

(b) $i = 1$

(c) $\hat{\rho}_n(0) = \frac{1}{2}$ $n = 1, 2, 3$

Step 1: Iteration

$$R_1^F = 0.75 + \tfrac{1}{2}(0.75)(0.73684) + 0 = 1.0263$$

$$\hat{\rho}_1(1) = \frac{1.0263 + 0.1184}{1 + 1.0263} = 0.5649$$

$$R_2^F = 0.35 + \tfrac{1}{2}(0.5)(0.73684) + \tfrac{1}{4}(0.65)(0.63158) = 0.6368$$

$$\hat{\rho}_2(1) = \frac{0.6368 + 0.1184}{1 + 0.6368} = 0.4614$$

$$R_3^F = 0.4 + \tfrac{1}{2}(0.25)(0.73684) + \tfrac{1}{4}(0.85)(0.63158) = 0.6263$$

$$\hat{\rho}_3(1) = \frac{0.6263 + 0.1184}{1 + 0.6263} = 0.4579$$

Step 2: Normalization

(a) $$\Gamma = \left[\frac{\frac{1}{3}(0.5649 + 0.4614 + 0.4579)}{\frac{1}{2}}\right]^{-1} = 1.01065$$

(b) $\hat{\rho}_1(1) = 0.5709$

$\hat{\rho}_2(1) = 0.4663$

$\hat{\rho}_3(1) = 0.4628$

Step 3: Convergence test

Test fails for $\epsilon < 0.01$; $i = 2$, return to Step 1.

FIGURE 5.20 Initialization and first iteration.

be seen, the procedure has converged for a very stringent convergence criterion ($\epsilon = 0.0011$ or greater) on the third iteration. The resulting workloads are estimated to be $\hat{\rho}_1 = 0.5627$, $\hat{\rho}_2 = 0.4716$, and $\hat{\rho}_3 = 0.4657$. The exact workloads, as computed from the exact hypercube model, are $\rho_1 = 0.5574$, $\rho_2 = 0.4734$, and $\rho_3 = 0.4693$. The maximum error is $|\hat{\rho}_1 - \rho_1| \simeq 0.0053$, or about 1 percent error. The average error is about 0.7 percent. These error magnitudes are typical of those achieved with the approximation procedure.

Once we have the ρ_n's, as derived above, it is relatively straightforward to obtain an estimate for the f_{nj}'s, where f_{nj} is the fraction of assignments that send unit n to atom j. The details of this are worked out in Problem 5.11. Once we have f_{nj} for all n, j, then all other hypercube performance measures

Step 1′: Iteration

$$R_1^F = 0.75 + 0.73684\,[0.35\,(0.4663 + 0.40\,(0.4628))] + 0 = 1.0067$$

$$\hat{\rho}_1(2) = \frac{1.0067 + 0.1184}{1 + 1.0067} = 0.5607$$

$$R_2^F = 0.35 + 0.73684\,(0.5709)\,(0.5) + 0.63158\,(0.65)\,(0.5709)\,(0.4628)$$

$$= 0.6688$$

$$\hat{\rho}_2(2) = \frac{0.6688 + 0.1184}{1 + 0.6688} = 0.4717$$

$$R_3^F = 0.4 + 0.73684\,(0.5709)\,(0.25) + 0.63158\,(0.85)\,(0.5709)\,(0.4663)$$

$$= 0.6481$$

$$\hat{\rho}_3(2) = \frac{0.6481 + 0.1184}{1 + 0.6481} = 0.4651$$

Step 2′: Normalization

(a) $\Gamma = [\frac{2}{3}\,(0.5607 + 0.4717 + 0.4651)]^{-1} = 1.0017$

(b) $\hat{\rho}_1(2) = 0.5616$

$\hat{\rho}_2(2) = 0.4725$

$\hat{\rho}_3(2) = 0.4659$

Step 3′: Convergence test

Test fails for any $\epsilon < 0.0093$; $i = 3$, return to Step 1.

FIGURE 5.21 Second iteration.

(including travel times) can be computed simply by substituting into the simple algebraic equations derived earlier for the exact hypercube model. Approximation errors for these measures, too, rarely exceed 2 percent and often are near 1 percent.

There are extensions to the basic approximation procedure described above, paralleling (but not as extensive as) those of the basic hypercube model. In particular, one can derive an approximation procedure for the zero-line-capacity queue and for the case of unequal service times. However, there is no known extension to dispatch policies other than fixed preference. Details can be found in [LARS 75a] and [JARV 75].

We have now completed our tour of N-server spatial queueing models for finite N. We conclude the chapter with two applications of many-server queues to derive analytically several important (rule-of-thumb) performance characteristics of server-to-customer queueing systems.

Step 1″: Iteration

$$R_1^F = 0.75 + 0.73684\,[0.35\,(0.4725 + 0.40\,(0.4659))] + 0 = 1.0092$$

$$\hat{\rho}_1(3) = \frac{1.0092 + 0.1184}{1 + 1.0092} = 0.5612$$

$$R_2^F = 0.35 + 0.73684\,(0.5616)\,(0.5) + 0.63158\,(0.65)\,(0.5616)\,(0.4659)$$

$$= 0.6643$$

$$\hat{\rho}_2(3) = \frac{0.6643 + 0.1184}{1 + 0.6643} = 0.4703$$

$$R_3^F = 0.4 + 0.73684\,(0.25)\,(0.5616) + (0.63158)\,(0.85)\,(0.5616)\,(0.4725)$$

$$= 0.6459$$

$$\hat{\rho}_3(3) = \frac{0.6459 + 0.1184}{1 + 0.6459} = 0.4644$$

Step 2″: Normalization

(a) $\Gamma = [\frac{2}{3}\,(0.5612 + 0.4703 + 0.4644)]^{-1} = 1.00274$

(b) $\hat{\rho}_1(3) = 0.5627$

$\hat{\rho}_2(3) = 0.4716$

$\hat{\rho}_3(3) = 0.4657$

Step 3″: Convergence test

Convergence test succeeds for any $\epsilon < 0.0011$. STOP.

FIGURE 5.22 Third and final iteration.

5.6 FRACTION OF DISPATCHES THAT ARE INTERRESPONSE AREA DISPATCHES

Often an agency administrator will want to know the number of dispatches that take units outside of their primary response areas. Such responses increase travel time and they may result in degraded service due to unfamiliarity with the neighborhood (its geography, people, and traditions). While we could compute this quantity exactly with the hypercube model, it is instructive to obtain some approximate "back-of-the-envelope" results for situations with many servers N. In particular, we wish to see how the number of interresponse area dispatches varies with average system-wide workload, workload imbalances, and time-varying demands for service.

For a given region with many response units N, define

$\lambda_n(t) \equiv$ average rate (requests/hour) at which requests for service are generated in a Poisson manner from response area n at time t, $n = 1, 2, \ldots, N$; $0 \leq t \leq T$

$\rho_n(t) \equiv$ probability that response unit n is *unavailable* for dispatch at time t, $n = 1, 2, \ldots, N$; $0 \leq t \leq T$

$$\lambda \equiv \text{"average service request rate"} = \frac{1}{T}\int_0^T \sum_{n=1}^{N} \lambda_n(t)\, dt \tag{5.54}$$

$\mu^{-1} \equiv$ average total time required for a response unit to service a request

Given a request that arrives from primary response area n, we assume that the dispatching strategy is as follows:

1. Dispatch unit n, if available.
2. Otherwise, dispatch some unit m, $m \neq n$, where the particular choice depends on the state of the system (and, perhaps, other factors).

Invoking a large-N (or low-workload) assumption, we assume that the probability that all units are simultaneously busy is negligibly small.

Suppose that a service request arrives from response area n at time t. The probability that it will result in an interservice area dispatch is equal to the probability that unit n is busy [i.e., $\rho_n(t)$]. Given a random request that arrives in the interval $[0, T]$, the likelihood that it arrives in the interval t to $t + dt$ *and* is from response area n is

$$\frac{\lambda_n(t)\, dt}{\lambda T}$$

Thus, the likelihood that a random request arrives in t to $t + dt$, is from response area n, and results in an interresponse area dispatch is

$$\rho_n(t)\frac{\lambda_n(t)\, dt}{\lambda T}$$

Summing over all N response areas and integrating over $[0, T]$ we obtain the probability that a random dispatch that occurs in $[0, T]$ will be an interresponse area dispatch,

$$f_I = \frac{1}{T}\int_0^T \sum_{n=1}^{N} \rho_n(t)\frac{\lambda_n(t)}{\lambda}\, dt \tag{5.55}$$

If the system is non-time-varying, with $\lambda_n(t) = \lambda_n$, $\rho_n(t) = \rho_n$, (5.55) reduces to

$$f_I = \sum_{n=1}^{N} \rho_n \frac{\lambda_n}{\lambda} \tag{5.56}$$

In Problem 5.14 we explore some special cases of (5.56). We will find, in general, that f_I increases as workload imbalances increase. If workloads are fairly well balanced, one can often approximate f_I to equal (or slightly exceed) the average fraction of time units are busy, averaged over all units, provided that the system is non-time-varying.

What if the system is time-varying? We can show that this usually "makes things worse," assuming that interresponse area dispatches are undesirable. That is, given one further assumption, we can obtain a bound for f_I in a time-varying system which states that the amount of interresponse area dispatching is at least as great as that which would occur in the "equivalent" non-time-varying system. To describe this equivalent system, we simply replace $\lambda_n(t)$ and $\rho_n(t)$ with their time averages,

$$\lambda_n = \frac{1}{T}\int_0^T \lambda_n(t)\, dt$$

$$\rho_n = \frac{1}{T}\int_0^T \rho_n(t)\, dt$$

The additional assumption we require is that $\rho_n(t) \geq \rho_n$ whenever $\lambda_n(t) \geq \lambda_n$ and that $\rho_n(t) \leq \rho_n$ whenever $\lambda_n(t) \leq \lambda_n$. This says that a unit's workload should be above (or below) average whenever the service request rate from its primary response area is above (or below) average. Is this reasonable? Can you think of counterexamples?

We wish to prove, given the foregoing assumptions, that f_I for the time-varying system as computed from (5.55) satisfies the following inequality:

$$f_I \geq \sum_{n=1}^{N} \rho_n \frac{\lambda_n}{\lambda} \tag{5.57}$$

Proof: Let

$$\rho_n(t) = \rho_n + \rho_n^{\Delta}(t)$$
$$\lambda_n(t) = \lambda_n + \lambda_n^{\Delta}(t)$$

Clearly, the perturbation terms on the right-hand side integrate to zero; that is,

$$\frac{1}{T}\int_0^T \rho_n^{\Delta}(t)\, dt = 0$$

$$\frac{1}{T}\int_0^T \lambda_n^{\Delta}(t)\, dt = 0$$

From (5.55) we have

$$f_I = \frac{1}{\lambda T}\sum_{n=1}^{N}\int_0^T [\rho_n\lambda_n + \rho_n\lambda_n^{\Delta}(t) + \lambda_n\rho_n^{\Delta}(t) + \rho_n^{\Delta}(t)\lambda_n^{\Delta}(t)]\,dt$$

Since the second and third terms in the integrand integrate to zero,

$$f_I = \sum_{n=1}^{N}\rho_n\frac{\lambda_n}{\lambda} + \frac{1}{\lambda T}\sum_{n=1}^{N}\int_0^T \rho_n^{\Delta}(t)\lambda_n^{\Delta}(t)\,dt$$

Now, since

$$\text{sgn}\,[\rho_n^{\Delta}(t)] = \text{sgn}\,[\lambda_n^{\Delta}(t)]$$

then

$$\int_0^T \rho_n^{\Delta}(t)\lambda_n^{\Delta}(t)\,dt \geq 0$$

and thus (5.57) must be true.

Problem 5.14 asks you to reexamine this analysis for systems that do not always give first preference to the primary response area's unit (e.g., a system incorporating an automatic vehicle locator system, which would allow the dispatcher to assign the vehicle closest to the scene of the service request).

5.7 SPATIAL DISTRIBUTION OF BUSY SERVERS

In some situations we would like to know the spatial distribution of servers busy at the scene of service requests. For instance, we may wish to analyze a dispatching policy which may interrupt a server busy on low-priority service in order to send him or her to a nearby higher-priority request; in such a case we would need to know the distribution of travel time to the nearest busy server. Or, in police applications, since police presence is said to deter crime, we may wish to know the spatial distribution of busy servers (as well as available servers) because a parked patrol car also acts as a visible deterrent; we may wish to alter this distribution, if possible, by adjusting our spatial prepositioning policies (e.g., beat designs).

We examine this question using the theory of $M/G/\infty$ queues (cf. Section 4.8). The assumption of an infinite number of servers implies that the actual number of servers is sufficiently large or that workload is sufficiently small so that queues almost never form.

Consider, then, a spatially distributed service system in which:

1. Service requests are generated in a Poisson manner from a region of area A_0 at a rate λA_0 requests per hour.
2. Service requests are distributed uniformly in space.
3. μ^{-1} = average time to service a request (general service time pdf).
4. There are infinitely many servers.

Then, according to (4.87), in the steady state the probability that there are k busy service units in any subregion of area $A \leq A_0$ is

$$P_k(A) = \frac{[\lambda A/\mu]^k e^{-(\lambda A/\mu)}}{k!} \qquad k = 0, 1, 2, \ldots \tag{5.58}$$

This result says that, regardless of the method of prepositioning the units, the *busy* servers are distributed as a spatial Poisson process with parameter (λ/μ) busy servers per unit area. This simple result allows us to use "nearest neighbor theory" of spatial Poisson processes to develop probability laws of the travel times to the kth closest busy server (see Example 15 in Chap. 3).

5.8 EXPECTED TRAVEL DISTANCES AND EXPECTED TRAVEL TIMES, REVISITED

In Chapter 3 we derived several expressions for the expected travel distance and the expected travel time of urban service units responding to calls. In this section we shall extend our results to account, under certain conditions, for the effects of congestion (i.e., for the fact that some response units may be busy and unavailable at the time when a call for service occurs).

Consider a large region of a city of area A_0 and assume that response-unit stations have been located in this region according to a spatial Poisson process with intensity of γ stations per square mile. Assume that each of the N_0 stations holds exactly one response unit. In the following, we shall assume that:

1. Calls for service are independently and uniformly distributed in the city region and are generated in a Poisson manner at a rate of λ calls per hour per unit area.
2. Each call is handled by a single response unit and service times are independent and approximately identically distributed random variables with mean μ^{-1} (this includes travel time to the call, time spent on the scene, and travel time back to the response unit's station).

3. A nearest-response-unit dispatching policy is used [i.e., the response unit dispatched to a call is always the *available* (nonbusy) unit which is closest to the location of that call at the time when the call is received].

Note that assumption 2 implies that

$$E[\text{time spent on the scene of a call}] \gg E[\text{travel time to a call}]$$

a condition that is true for many urban service systems.

When A_0 is sufficiently large so that the effects of the boundaries of the region can be ignored and when all N_0 response units are available, it was shown in Chapter 3 [cf. (3.104a) and (3.105)] that, for the system described, the expected travel distance to a call is given by

$$E[D] = c\sqrt{\frac{A_0}{N_0}} \tag{5.59}$$

Here, c is a constant that depends on the travel metric in use and possibly on other geographical characteristics of the region in question.

In general, the number of available response units, N, will fluctuate with workload and, over a long period of time, will take values ranging from 0 to N_0. It would clearly be very useful if we could find a simple relationship between $E[D]$ and the average number $E[N]$ of available response units in the area. Such a relationship would be helpful for planning purposes since it would take into account the effects of system workload—as reflected in $E[N]$.

To develop such a relationship, we shall assume that rarely, if ever, are all N_0 units in the region busy. We have already shown in the previous section that under these circumstances one can use the $M/G/\infty$ system results, and we proved that the *busy* response units at any instant in time are distributed as a spatial Poisson process with parameter λ/μ busy units per unit area. Since the N_0 unit stations are distributed as a spatial Poisson process as well, it follows that, at any instant, the *available* response units are approximately distributed as a spatial Poisson process.[9] Hence, given any value of N, we have, as in (5.59), that

$$E[D \mid N = k] \simeq c\sqrt{\frac{A_0}{k}} \qquad k = 1, 2, 3, \ldots, N_0 \tag{5.60}$$

[9]This assumption is an approximation, since the dispatch policy that assigns the closest available server creates "holes in coverage"; these holes are somewhat larger than those ordinarily found in a spatial Poisson process. The net effect is that assignment of servers to busy status does not constitute "random erasures" of an existing spatial Poisson process, but rather "correlated erasures," yielding a residual (unerased) process that is not strictly a Poisson process.

For the case $N = 0$, which is very unlikely anyway, we can assume that a response unit from outside the region responds to calls and that the expected travel distance is then given by a constant D_0. In the steady state, we therefore have

$$E[D] = P_0 D_0 + \sum_{k=1}^{N_0} P_k c\sqrt{\frac{A_0}{k}} \tag{5.61}$$

where

$$P_k = \frac{[N_0\rho]^{(N_0-k)}/(N_0 - k)!}{\sum_{i=0}^{N_0} [N_0\rho]^i/i!} \qquad k = 0, 1, 2, \ldots, N_0 \tag{5.62}$$

are the steady-state probabilities of having k units available, or $(N_0 - k)$ units busy, as given by an $M/G/N_0$ queueing model with no waiting space [cf. (4.85)]. This approach was first utilized by Larson in analyzing overlapping police beats [LARS 72a, Eqs. (7.40), (7.42)].

Since P_0 is very small by assumption, we can safely ignore the first term in (5.61) and write

$$E[D] \approx c\sqrt{A_0}\, E\left[\frac{1}{\sqrt{N}}\right] \tag{5.63}$$

where the expectation is taken over the state probabilities P_k.

It is much more convenient, however, to use a further simplifying approximation, by writing

$$E[D] \approx c\sqrt{\frac{A_0}{E[N]}} \tag{5.64}$$

In truth (5.64) provides only a lower bound for (5.63) since $E[(\sqrt{N})^{-1}] < (E[N])^{-1/2}$. This follows from the fact that if $h(X)$ is a convex function of a nonnegative random variable X [such as $h(X) = (\sqrt{X})^{-1}$], then[10]

$$E[g(X)] \geq g(E[X]) \tag{5.65}$$

However, we show in Problem 5.5 that the substitution of $1/\sqrt{E[N]}$ for $E[1/\sqrt{N}]$ is quite reasonable for $E[N]$ sufficiently large and N "compactly distributed" about its mean.

Then, using the fact that P_0 is very small and that, consequently, the $M/G/N_0$ model is virtually indistinguishable from an $M/G/\infty$ model, we finally have

$$E[N] = N_0 - E[\text{number of busy response units}] \approx N_0 - N_0\rho \tag{5.66}$$

[10] Inequality (5.65) is also known as Jensen's inequality (cf. Problem 3.5).

from which it follows by substitution into (5.64) that

$$E[D] \approx c\sqrt{\frac{A_0}{N_0(1-\rho)}} \tag{5.67}$$

These results can now be extended in a straightforward way to travel times. Using, for instance, the approximate acceleration/cruising speed model of Chapter 3 [cf. (3.93)],

$$E[T] = \frac{E[D]}{v_c} + \frac{v_c}{a} \tag{5.68}$$

we have

$$E[T] \approx \frac{c}{v_c}\sqrt{\frac{A_0}{N_0(1-\rho)}} + \frac{v_c}{a} \tag{5.69}$$

Equations (5.64), (5.67), and (5.69) are often referred to as *square-root laws*, since they relate $E[D]$ and $E[T]$ to the square root of the density of available response units in an urban area [KOLE 75a].

5.8.1 Extensions and Empirical Evidence

Square-root laws for $E[D]$ and $E[T]$ can also be derived for cases other than the one described above. For example, in Chapter 3 it was shown that, when response units are arranged in a symmetric pattern at the centers of equal squares rotated by 45° with respect to the directions of (right-angle) travel, $E[D] = 0.47\sqrt{A_0/N_0}$ [cf. (3.107)]. This expression can be used as the starting point for deriving a square-root law for this situation. The same applies when other regular patterns for response unit locations are in effect (and, as we have seen in Chapter 3, the constant c is likely to be quite insensitive to the precise shape of these patterns).

Another extension would involve the case in which more than one response unit could be placed in some or all of the N_0 stations in the region, as in fire department operations. Then (5.67) must be modified to

$$E[D] = c\sqrt{A_0}(N_0 - E[\text{number of empty stations}])^{-1/2} \tag{5.67a}$$

A third extension, also applicable to fire departments, concerns situations in which some requests may be serviced by more than one response unit. It has been shown [CHAI 71] that (5.67) can still be used if the mean service time, μ^{-1}, is adjusted to reflect the total number of "service time units" spent on requests. For example, if three fire engines spend 20 minutes each at the site of a particular fire alarm, then the service time for that alarm must be set equal to 60 minutes.

The New York City Rand Institute accumulated an impressive amount of data showing that expressions such as (5.64), (5.67), and (5.68) are valid in practice under a considerable variety of conditions [KOLE 75a, KOLE 75b], including dispatching policies that do not always send the nearest available response unit to a call but may instead dispatch the second or third nearest unit (for reasons such as those discussed in Section 5.3). For several different urban regions the constant c has been found to fall in the range 0.55 to 0.61 for fire department operations. This is not surprising in view of the fact that, for right-angle travel, $c \approx 0.63$ if stations are located completely randomly and $c \approx 0.47$ if stations are at the corners of a perfectly regular lattice. For most cities the actual pattern of firehouse locations is somewhat between these two extremes.

References

[CART 72] Carter, G. M., J. M. Chaiken, and E. Ignall, "Response Areas for Two Emergency Units," *Operations Research*, *20*, 571–594 (1972).

[CHAI 72] Chaiken, J. M. and E. Ignall, "An Extension of Erlang's Formulas which Distinguishes Individual Servers," *J. Appl. Probab.*, *9*, 192–196, (1972).

[CHAI 71] Chaiken, J., "Number of Emergency Units Busy at Alarms which Require Multiple Servers," New York City Rand Institute, R-531-NYC/HUD, (March 1971).

[CHEL 80] Chelst, K. R., and Z. Barlach, "Multiple Unit Dispatches in Emergency Services: Models to Estimate System Performance," to appear in *Management Science*.

[FRAN 71] Frank, H., and I. T. Frisch, *Communication, Transmission, and Transportation Networks*, Addison-Wesley, Reading, Mass., 1971.

[HALP 77] Halpern, J., "The Accuracy of Estimates for the Performance Criteria of Certain Emergency Service Queueing Systems," *Transportation Science*, *11*, 227–242 (1977).

[JARV 75] Jarvis, J. P., *Optimization in Stochastic Service Systems with Distinguishable Servers*, TR-19-75, MIT Operations Research Center, Cambridge, Mass., June 1975.

[KLEI 75] KLEINROCK, L. *Queueing Systems, Volume 1: Theory*, Wiley, New York; 1975.

[KLEI 76] KLEINROCK, L., *Queueing Systems, Volume 2: Computer Applications*, Wiley, New York, 1976.

[KOLE 75a] KOLESAR, P., "A Model for Predicting Average Fire Engine Travel Times," *Operations Research, 23* (4), 603–613 (July–August 1975).

[KOLE 75b] KOLESAR, P., W. WALKER, AND J. HAUSNER, "Determining the Relation between Fire Engine Travel Times and Travel Distances in New York City," *Operations Research, 23* (4), 614–627, (July–August 1975).

[LARS 72] LARSON, R. C., AND K. A. STEVENSON, "On Insensitivities in Urban Redistricting and Facility Location," *Operations Research, 20*, 595–612 (1972).

[LARS 74a] LARSON, R. C., "A Hypercube Queuing Model for Facility Location and Redistricting in Urban Emergency Services," *Computers and Operations Research, 1* (1), 67–95 (1974).

[LARS 74b] LARSON, R. C., "Illustrative Police Sector Redesign in District 4 in Boston," *Urban Analysis, 2*, 51–91 (1974).

[LARS 75a] LARSON, R. C., "Approximating the Performance of Urban Emergency Service Systems," *Operations Research, 23* (5), 845–868 (1975).

[LARS 75b] LARSON, R. C., "Computer Program for Calculating the Performance of Urban Emergency Service Systems: User's Manual (Batch Processing)," *Innovative Resource Planning in Urban Public Safety Systems*, Report TR-14-75, Massachusetts Institute of Technology, Cambridge, Mass., 1975.

[LARS 78] LARSON, R. C., AND E. A. FRANCK, "Evaluating Dispatching Consequences of Automatic Vehicle Location in Emergency Services," *Computers and Operations Research, 5*, 11–30 (1978).

[LARS 80] LARSON, R. C., AND M. A. MCKNEW, "Approximating Heterogeneous Server Queueing Systems Having Server-Initiated and Centrally Assigned Tasks," submitted to *Management Science*.

[ODON 69] ODONI, A. R., "An Analytical Investigation of Air Traffic in the Vicinity of Terminal Areas," Ph.D. dissertation (unpublished), Massachusetts Institute of Technology, Cambridge, Mass., 1969.

Problems

5.1 *M/G/1 modified queue* Apply (5.2) for the case of a rectangular (X_0-by-Y_0) service region with directions of travel parallel to the sides of the rectangle and with a single emergency repair unit garaged at the region's center, ($x = X_0/2$, $y = Y_0/2$). The emergency repair unit operates as in the third paragraph of Section 5.2.

a. Suppose that $\lambda = 1$ call per hour. Let us examine how several alternative service region designs having equal area (i.e., $X_0 Y_0 = A$) can affect system performance. Let $A = 4$ square miles and response speed be 10 miles/hr. Further, suppose that the mean and variance of on-scene service time are 45 minutes and $(45)^2$ minutes^2, respectively. Find the mean time from calling until arrival of a service unit for $X_0 = Y_0$, $X_0 = 2Y_0$, and $X_0 = 20Y_0$ (assuming that we constrain $X_0 Y_0 = A = 4$ square miles). Can the system be saturated (i.e., $\rho > 1$) for some values of X_0, Y_0, and unsaturated for others?

b. Verify that with the constraint $X_0 Y_0 = A$, minimum response time is always achieved by setting $X_0 = Y_0 = \sqrt{A}$.

5.2 *Infinite array of linear concatenated sectors* One infinite server spatially distributed queueing system that has provided certain physical insights into alternative dispatching procedures is a linear concatenated sector system. On the x-axis, assume that sector i covers the interval from $x = i/2$ to $x = (i/2) + 1$ for i even and from $x = -(i+1)/2$ to $x = -(i-1)/2$ for i odd. Response unit i is assigned to patrol uniformly sector i when it is available for dispatch assignment. Each unit is assumed to be available with probability $(1 - \rho)$, *independently* of the status of all other units. (It should be clear that the independence assumption is an approximation.) The position of each available unit is selected from a uniform distribution over the length of the unit's sector. The random variable indicating the position of unit i is X_i; a particular experimental value of the random variable is x_i.

Assume that an incident is reported from some point x in sector 0 ($0 \leq x \leq 1$) and that the dispatcher must select an available unit to assign to the incident. The incident position x is drawn from a uniform probability density over [0, 1]. The dispatcher may use any one of the following three selection criteria:

1. Strict center of mass $\underset{\substack{i \in \text{set of} \\ \text{available units}}}{\text{Min}} \{E[|X_i - \frac{1}{2}|]\}$

2. Modified center of mass $\underset{\substack{i \in \text{set of} \\ \text{available units}}}{\text{Min}} \{E[|X_i - x|]\}$

3. Closest car $\underset{\substack{i \in \text{set of} \\ \text{available units}}}{\text{Min}} \{|x_i - x|\}$

Let $\bar{D}_i(\rho)$ = expected travel distance for strategy i, given a utilization factor of ρ ($i = 1, 2, 3$).

a. Prove the following:

i. $\bar{D}_1(\rho) = \frac{1}{3}(1 - \rho) + \frac{\rho}{1 - \rho^2}$.

ii. $\bar{D}_2(\rho) = \frac{1}{3} - \frac{1}{12}\rho + \frac{\rho/2}{1 - \rho}$.

iii. $\bar{D}_3(\rho) = \frac{7}{24} + \frac{11}{24}\rho + \frac{1}{2}\frac{\rho^2}{1 - \rho}$.

b. Let $\epsilon_{ij}(\rho) = \bar{D}_i(\rho) - \bar{D}_j(\rho)$.
Verify the following:

$$\epsilon_{12}(\rho) = \rho\frac{1 - \rho}{4(1 + \rho)}$$

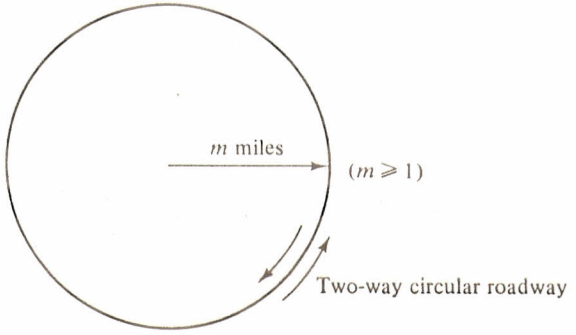
$$\epsilon_{13}(\rho) = \frac{1 + 6\rho - 7\rho^2}{24(1 + \rho)}$$

$$\epsilon_{23}(\rho) = \frac{1}{24}(1 - \rho)$$

Do these results make intuitive sense for limiting values of ρ? What practical significance do they have?

5.3 *Coverage problem* Consider a circular roadway in an urban region as shown in Figure P5.3. The roadway has many shops, entertainment spots, and other potential targets for crime. The patrol cars of the city's police department are all unmarked and thereby provide no visible police coverage (protection) while patrolling. However, when they are stopped at a scene of a call for service, they flash a beacon that is visible (in either direction) for exactly a miles ($a \ll 1.0$). Thus, each parked patrol car generates a region of *visible police coverage* of length $2a$ miles.

The total number of patrol cars is so great, and the priority of calls on the circular roadway so urgent, that no call for service from the circular roadway is ever delayed in queue. And travel time to the scene for a call for service is insignificant compared to the service time at the scene, which has a mean $1/\mu$.



Calls for police service along the circular roadway arrive as a Poisson process with an average rate of λ calls per linear mile. We assume that the system is operating in the steady state.

Show that, in the steady state, the expected length (in miles) of the circular roadway that is given visible police coverage is equal to

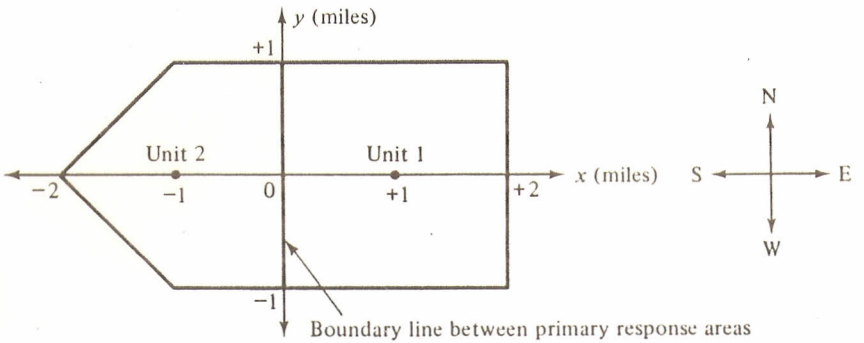
$$2\pi m(1 - e^{-2a\lambda/\mu})$$

5.4 *Testing your understanding of the two-server model* Consider the city depicted in Figure P5.4. In this city emergency repair service is provided by two response units, prepositioned at $(1, 0)$ and $(-1, 0)$, respectively. Travel distance is right-angle, with distance d between two points (x_1, y_1) and (x_2, y_2) equal to

$$d = |x_1 - x_2| + |y_1 - y_2| = d_x + d_y$$

where

$$d_x = |x_1 - x_2| \quad \text{and} \quad d_y = |y_1 - y_2|$$



Repairs occur according to a spatial and temporal Poisson process with

$$\lambda(x, y)\,dx\,dy\,dt \equiv P\{\text{an emergency occurs in the infinitesimal rectangle } (x, x + dx), (y, y + dy) \text{ during the infinitesimal time interval } (t, t + dt)\}$$

For points (x, y) within the city we are given that

$$\lambda(x, y) = 2\lambda_0 \qquad \text{for } x \geq 0$$
$$\lambda(x, y) = \lambda_0 \qquad \text{for } x < 0$$

Unit 1's primary response area consists of all points east of the north-south boundary line that partitions the city. Unit 2's primary response area consists of all points west of that line. Note that the boundary, as drawn, is an equal travel-time boundary line. Whenever an emergency repair is reported from response area i ($i = 1, 2$), unit i is assigned, if available; otherwise, the other unit is assigned, if

available. If neither unit is available, the call is lost (i.e., no queueing is allowed).

Repair units travel at 100 miles/hr to and from the scene. On-scene service time is negative exponential with mean $\mu^{-1} = 10$ hours. Upon completion of service, a unit always returns to its home location. Finally, $\lambda_0 = 2 \times 10^{-5}$ (emergencies per hour per square mile). Assume the system is operating in steady state. Average system-wide travel time to an emergency is $\bar{T}$.

a. Using suitable approximations, or exact analysis if you prefer, find approximate nonzero values for ρ_1 and ρ_2, where $\rho_m \equiv$ workload of unit m.

b. The random variable D_x is defined to be the system-wide east-west (or west-east) travel distance for a random emergency. Sketch the pdf for D_x, again making suitable approximations.

c. Approximately, what is the fraction of dispatches that are interresponse area dispatches?

For parts (d)–(g) only, suppose that $\lambda_0 = 10^{-2}$. This yields new values for ρ_1, ρ_2, $\bar{T}$, and so on. In considering each of these questions [parts (d)–(g)], assume that the total service-time distribution remains the same as that assumed in parts (a)–(c).

d. Suppose the equal-travel-time boundary line is shifted a distance ϵ (ϵ small) toward unit 1. Which is the appropriate response?

   i. As a result of the shift, $\bar{T}$ will increase but $|\rho_1 - \rho_2|$ will decrease.

   ii. As a result of the shift, $\bar{T}$ will increase and $|\rho_1 - \rho_2|$ will increase.

   iii. As a result of the shift, $\bar{T}$ will decrease and $|\rho_1 - \rho_2|$ will decrease.

   iv. As a result of the shift, $\bar{T}$ will decrease but $|\rho_1 - \rho_2|$ will increase.

   v. As a result of the shift, we cannot tell which of the four possibilities above will apply.

e. Now suppose that the equal-travel-time boundary line is shifted a distance ϵ (ϵ small) toward unit 2. D_y is defined to be the system-wide north-south (or south-north) travel distance for a random emergency. As a result of the shift,

   i. $E[D_y]$ will stay the same.

   ii. $E[D_y]$ will increase.

   iii. $E[D_y]$ will decrease.

   iv. The behavior of $E[D_y]$ cannot be determined.

f. As in part (e), suppose that the equal-travel-time boundary line is shifted a distance ϵ (ϵ small) toward unit 2. As a result of the shift:

   i. $(\rho_1 + \rho_2)$ will stay the same.

   ii. $(\rho_1 + \rho_2)$ will increase.

   iii. $(\rho_1 + \rho_2)$ will decrease.

   iv. The behavior of $(\rho_1 + \rho_2)$ cannot be determined.

g. Suppose that a third unit is added as a backup unit and is located at $(x = 0, y = 0)$. This unit is assigned to any emergency that occurs when both units 1 and 2 are busy and unit 3 is available. Emergencies that arrive when all three units are busy are lost. Determine the workload of unit 3, ρ_3.

5.5 *More on the square-root law* We wish to explore the reasonableness of the approximation $E[1/\sqrt{N}] \approx 1/\sqrt{E[N]}$. We know from Section 5.8 that, in fact, $E[1/\sqrt{N}] > 1/\sqrt{E[N]}$.

a. Suppose that N is uniformly distributed over the integers $1, 2, \ldots, 10$. Verify that $E[1/\sqrt{N}] \approx 0.502$, while $1/\sqrt{E[N]} = 1/\sqrt{5.5} \approx 0.426$.

b. Now suppose that N's distribution is clustered more around its mean:

$$P_i = P\{N = i\} = \begin{cases} \dfrac{i}{30} & i = 1, 2, 3, 4, 5 \\ \dfrac{10 - i}{30} & i = 6, 7, 8, 9, 10 \end{cases}$$

Verify in this case that $E[1/\sqrt{N}] \approx 0.466$, a result much closer to the desired approximation.

c. Now suppose that $E[N]$ is large and N's distribution is fairly symmetric about its mean $E[N]$, which for simplicity we assume to be an integer. Using the square-root approximation

$$(y + \epsilon)^{1/2} \approx y^{1/2} + \tfrac{1}{2}y^{-1/2}\epsilon$$

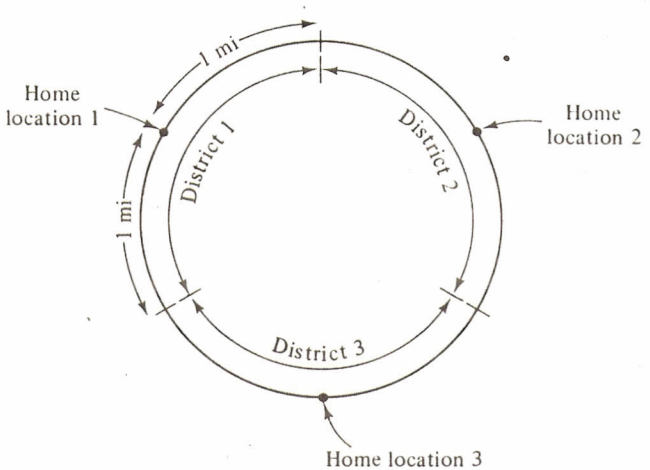
for $|\epsilon|$ considerably smaller than $y > 0$, write $E[1/\sqrt{N}]$ as a series of terms symmetrically expanded about the mean

$$\left[\ldots + (P_{E[N]-1})\frac{1}{\sqrt{E[N]-1}} + P_{E[N]}\frac{1}{\sqrt{E[N]}} + (P_{E[N]+1})\frac{1}{\sqrt{E[N]+1}} + \ldots\right]$$

Argue why $E[1/\sqrt{N}] \approx 1/\sqrt{E[N]}$ in this case.

5.6 *Ring city* Consider a ring city of circumference 6 miles which is serviced by three emergency response units. When not servicing an emergency call, a response unit is stationed at its "home location," located at the middle of the unit's district. There are three home locations (and districts) positioned symmetrically around the city as shown in Figure P5.6.

Calls for service occur as a homogeneous Poisson process (in time) at average rate two calls per hour. Locations of calls are uniformly, independently distributed over the circle's circumference. The total service time per call is negatively exponen-



tially distributed with mean 1 hour. Travel time can be ignored when computing service times, since units respond (traveling along the circle) at 60 miles/hr (1 mile/min).

The dispatching strategy is as follows for a call from district i ($i = 1, 2, 3$):

1. Assign unit i to the call for service, assuming that unit i is available.
2. Otherwise, randomly select one of the other two units, assuming that at least one is available.
3. If no unit is available, enter the call in a queue that is to be depleted in a first-come, first-served manner. (That is, no calls are "lost"—all are eventually serviced.)

As soon as a unit has completed servicing a call, we assume that it returns nearly instantaneously to its home location (even if it is then dispatched to a call waiting in queue).

Assume that the system is operating in the steady state.

a. Compute the workload ρ_i of unit i ($i = 1, 2, 3$), where $\rho_i \equiv$ fraction of time unit i is busy servicing calls.

b. Determine the mean travel time to calls for service.

c. Determine the probability density function of the travel time to calls for service.

d. A point x on the circle is said to be "covered" if at least one response unit is within $\frac{1}{2}$ mile of the point. It makes no difference whether or not the response unit is busy servicing a call. Find the average amount of the city (measured in miles) that is covered at a random time.

5.7 *Hypercube performance measures for the zero-line-capacity case* Suppose that we wish to derive hypercube performance measures for the case in which customers cannot enter a queue; they are either lost or, more likely, are handled by a backup system if they arrive when all N servers are busy.

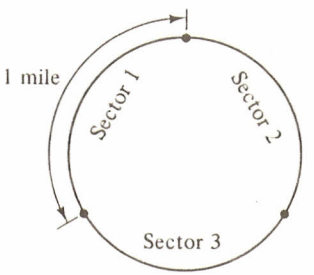
a. Argue that in this case, (5.35) is replaced by

$$f_{nJ} = \frac{\sum_{B_i \in E_{nJ}} \frac{\lambda_J}{\lambda} P\{B_i\}}{1 - P\{B_{2^N-1}\}} \qquad (*)$$

b. Argue further that the quantities found in (5.36)–(5.38), and (5.41)–(5.44) can be determined for the zero-line-capacity case by setting $P_Q = P'_Q = 0$ in these equations and by replacing $f_{nJ}^{[1]}$ with f_{nJ} from (*).

Note: The resulting equations can be used either with the exact or the approximate hypercube model.

5.8 *Three-server queue: evaluating a new technology* A certain circular highway is patrolled by three public safety cars. Each car patrols a 1-mile sector of the 3-mile highway (see Figure P5.8). Calls for assistance occur along the highway. A dispatcher assigns a car to each call, if at least one is available. We wish to examine various operating properties of this system.



The system operates as follows:

1. Call positions are uniformly, independently distributed over the circular highway.
2. The call arrival process is a homogeneous Poisson process with rate parameter λ calls per hour.
3. Service time at the scene of the call has a negative exponential distribution with mean $\mu^{-1} = \frac{1}{2}$ hour.
4. Travel time is negligibly small compared to service time at the scene.
5. Speed of response is always 30 miles/hr.
6. U-turns are permissible everywhere.

For parts (a)–(c), assume that the dispatching strategy is as follows. Given a call from sector i ($i = 1, 2, 3$):

1. Assign car i, if available.
2. Otherwise, randomly choose some car j ($j \neq i$), and assign it, if at least one other car is available.

3. Otherwise, the call is lost.

   a. Find the steady-state probability that i cars are busy ($i = 0, 1, 2, 3$).

   b. Find the steady-state probability that car 1 is busy and car 2 is free.

   c. Find the average travel time to calls for this system. Evaluate for $\lambda \approx 0$, $\lambda = 3$, $\lambda = 1{,}000$.

   d. It has been proposed that the public safety bureau should purchase a perfect resolution car locator system. With such a system, the dispatching strategy is changed as follows:

      Given a call from sector i ($i = 1, 2, 3$):

      1. Assign the *closest available car*, if at least one is available;
      2. Otherwise, the call is lost.

      Find the average travel time to a call for this system. Evaluate for $\lambda \approx 0$, $\lambda = 3$, $\lambda = 1{,}000$. (This part will utilize your knowledge of geometrical probability concepts.)

5.9 *Deriving the Q factor* In this problem we derive $Q(N, \rho, j)$, as given in (5.48). By laws of conditional probability we can write

$$P\{B_1B_2 \ldots B_jF_{j+1}\} = \sum_{k=0}^{N} P\{B_1B_2 \ldots B_jF_{j+1} \mid S_k\}P\{S_k\}$$

where S_k = system state in which exactly k servers are busy
$P\{S_k\}$ = steady-state probability that the system is in state S_k

Note also that

$$P\{B_1B_2 \ldots B_jF_{j+1} \mid S_k\} = P\{F_{j+1} \mid B_1B_2 \ldots B_jS_k\}P\{B_j \mid B_1B_2 \ldots B_{j-1}S_k\} \ldots P\{B_1 \mid S_k\}$$

   a. Argue that $P\{B_1 \mid S_k\} = k/N$.

   b. For the general case, argue that

$$P\{B_i \mid B_1B_2 \ldots, B_{i-1}S_k\} = \frac{k - (i-1)}{N - (i-1)} \qquad i = 1, 2, \ldots, k+1$$

$$P\{F_{j+1} \mid B_1B_2 \ldots B_jS_k\} = \frac{N-k}{N-j} \qquad j = 0, 1, \ldots, k$$

   c. Combine the results above with the appropriate results for the $M/M/N$ infinite capacity queue, to obtain

$$P\{B_1B_2 \ldots B_jF_{j+1}\} = Q(N, \rho, j)\rho^j(1-\rho)$$

   where $Q(N, \rho, j)$ is given by (5.48).

5.10 *Hypercube approximation procedure for the zero-line-capacity case* Suppose that we wish to derive an approximate procedure for finding the performance measures of the hypercube model, analogous to that of Section 5.5, but assuming zero line capacity. To do this, we must develop a new Q factor and new workload normalization conditions.

   a. Verify that the appropriate steady-state probabilities for the corresponding $M/M/N$ zero line capacity queue are

$$P'\{S_k\} = \frac{N^k\eta^kP'\{S_0\}}{k!} \qquad k = 0, 1, \ldots, N$$

$$P'\{S_0\} = \left(\sum_{i=0}^{N} \frac{N^i\eta^i}{i!}\right)^{-1}$$

   where $\eta \equiv \lambda/N < +\infty$ (Assume that $\mu = 1$.)

   b. Confirm that the average utilization factor is

$$\rho = \frac{1}{N}\sum_{k=0}^{N} kP'\{S_k\} = \eta(1 - P'\{S_N\})$$

   c. Now we would like to develop a correction factor $Q'(N, \eta, j)$ that, when multiplied by $\rho^j(1-\rho)$, gives the exact probability $P\{B_1B_2 \ldots B_jF_{j+1}\}$ for the $M/M/N$ zero line capacity system. Following reasoning analogous to Problem 5.9, verify that

$$P\{B_1B_2 \ldots B_jF_{j+1}\} = Q'(N, \eta, j)\eta^j(1-\eta)$$

   where

$$Q'(N, \eta, j) = Q^*(N, \eta, j)\left(\frac{1}{1 - P'\{S_N\}}\right)^j \frac{1}{1 + \eta P'\{S_N\}/(1-\eta)}$$

   and where $Q^*(N, \eta, j)$ is equal to $Q(N, \eta, j)$ as computed for the $M/M/N$ infinite line capacity case, but with $P\{S_0\}$ replaced by $P'\{S_0\}$.

   d. Conclude that an appropriate workload approximation procedure for the zero-line-capacity case would utilize (5.52), (5.53), and the algorithm of Figure 5.19, with $Q(\ )$ replaced by $Q'(\ )$ and with the following other modifications:

      i. $\lambda_D = 0$.

      ii. *At Step 0:* $\hat{p}_n(0) = \rho = \left(\frac{\lambda}{N}\right)(1 - P'\{S_N\})$.

5.11 *Estimating interatom dispatch frequencies* Suppose that we have applied the workload approximation procedures summarized in Figure 5.19, resulting in workload estimates $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_N$.

a. Argue, that for the infinite-line-capacity case, a reasonable approximation for

$f_{nk}^{[1]} \equiv$ fraction of assignments that take server n to atom k and incur no queue delay

can be obtained by ordering on the dispatch preference list, using

$m_{aj} =$ identification number of the jth preferred server for atom a

and writing

$$\hat{f}_{m_{kj}k}^{[1]} = \left(\frac{\lambda_k}{\lambda}\right) Q(N, \rho, j-1) \left(\prod_{\ell=1}^{j-1} \hat{\rho}_{m_{k\ell}}\right)(1 - \hat{\rho}_{m_{kj}}) \qquad (*)$$

Argue further that the $\hat{f}_{nk}^{[1]}$'s must satisfy the following normalization conditions:

$$\sum_{n=1}^{N} \hat{f}_{nk}^{[1]} = (1 - P\{S_N\})\frac{\lambda_k}{\lambda} \qquad k = 1, 2, \ldots, N_A \qquad (**)$$

Thus, a suitable approximation procedure would utilize the results of (*), appropriately normalized (or scaled) to satisfy (**).

b. Show that, for the corresponding zero-line-capacity system, the procedures above apply with $Q(N, \rho, j-1)$ replaced by $Q(N, \rho, j-1)/(1 - P'_N)$ in (*) and the equation $\sum_{n=1}^{N} \hat{f}_{nk}^{[1]} = \lambda_k/\lambda$ replacing (**).

5.12 *Fire problem* Consider three fire stations, A, B, and C, located on the X-axis at -1, 0, and $+1$, respectively. The fire trucks at these stations service fires that are reported in the region $X = -2$ to $X = +2$. Fires within this region are independently uniformly distributed; they are generated at the rate λ per hour. Travel time to and from the fires is instantaneous. Service time at the scene is negatively exponentially distributed with mean $1/\mu$. There is one truck at each station and the dispatcher will assign the *closest available* truck. If no truck is available, reports of fires enter a queue that is depleted in a *first-come, first-served* manner.

a. Write down the equations whose solution would provide the utilization factor (fraction of time busy) of each of the trucks. You need not solve the equations.

b. Write down the equations whose solution would provide the fraction of calls generated in the interval $X = -\frac{1}{2}$ to $X = +\frac{1}{2}$ that are serviced by unit C. You need not solve the equations.

c. Find *approximate* solutions to parts (a) and (b) by employing the hypercube approximation procedure of Section 5.5 and Problem 5.11.

5.13 *Server-dependent mean service times* Suppose that in the three-server example of Section 5.4.3 we were told that the mean service times of the various servers were not identical, but were given as follows: $\mu_1^{-1} = \frac{2}{3}$, $\mu_2^{-1} = 1.0$, and $\mu_3^{-1} = 2.0$ (units of time). Here μ_n^{-1} includes both travel time and on-scene time.

a. Determine a requirement for λ such that the system is not saturated (i.e., that the total required service does not exceed the total available capacity).

b. Assuming that $\lambda = 1.5$, write a set of balance-of-flow equations analogous to (5.22)–(5.29) whose solutions yield the system equilibrium state probabilities.

Hint: This system, when unsaturated, does not collapse to an $M/M/3$ birth-and-death model, so one does not know the sums of probabilities along certain hyperplanes.

c. Suppose we are given a numerical value for P_{111}. Find P_Q in terms of P_{111} [see (5.21e)].

Hint: The system is a birth-and-death process for saturated system states.

d. Argue that (5.30) for server workloads and (5.32) for "unsaturated" interatom dispatch frequencies remain unchanged.

e. Argue that (5.34) should be replaced with

$$f_{nj}^{[2]} = \frac{\lambda_j}{\lambda} P'_Q \left(\frac{\mu_n}{\sum_{i=1}^{N} \mu_i}\right) \qquad (*)$$

where, as usual,

$$P'_Q = P_Q + P_{B_{2^N-1}}$$

Conclude that (5.36), (5.37), and (5.38) remain unchanged, assuming that $f_{nj}^{[2]}$, as given by (*), is substituted for $\lambda_j P'_Q/\lambda N$ in (5.35).

f. Mean travel times are more difficult and require approximations. Can you fill in the details?

5.14 *Interdistrict dispatching, revisited* In Section 5.6 we found that for a non-time-varying system with $\lambda_n(t) = \lambda_n$, $\rho_n(t) = \rho_n$, the fraction of dispatches that are interdistrict dispatches is

$$f = \sum_{n=1}^{N} \rho_n \frac{\lambda_n}{\lambda}$$

a. Examine the special case $\rho_n = \rho$, and physically interpret your result.

b. Examine the special case $\lambda_n =$ constant, and physically interpret your result.

c. In the text we developed (5.55), allowing for a time-varying system in which the sector car is always given first preference. However, for a system in which the dispatcher has car location information, he may prefer to assign an out-of-sector car that is closer to the scene than the sector car. Our previous analysis can be generalized to allow for this type of behavior. Let

$a_n(t)$ = probability that unit n is assigned to a call that arrives from sector n at time t, given that unit n is *available*

Derive the analogous result to (5.55) for this more general model. What are the physical implications of the result?

d. Does the practical significance of the results above change if we allow a queue to form? As a guide to answering this question, consider a non-time-varying system in near-saturation conditions (i.e., a queue almost always exists). A call that arrives when all N servers are busy is entered in queue. The queue is depleted in a first-come, first-served manner. Prove that

$$1 - f = \frac{1}{N}$$

5.15 *Spatial distribution of busy servers, revisited* Generalize (5.58) to show that if service requests are spatially distributed in some arbitrary way, then busy servers are distributed as a spatially nonhomogeneous Poisson process; identify the mean value of this process over any particular region $R' \subset R$.

5.16 *Linking travel times in finite regions to spatial Poisson processes* In this problem we wish to explore the validity of the spatial Poisson process model as we increase the number of independently, uniformly located response units in our area.

a. If a unit n is uniformly distributed over a square region of area N and if the incident is located at the center of the square, find the probability density function of D_n, the travel distance for unit n to reach the incident. (Assume that we have right-angle response parallel to the sides of the square.)

b. Assume that there are N such units in the region, indexed from $n = 1$ to $n = N$. The minimum travel time to an incident is

$$R_N = \text{Min}\,[D_1, D_2, \ldots, D_N]$$

But the spatial Poisson assumption implies that as N gets large, the pdf for R_N approaches a Rayleigh with parameter 2. The cumulative distribution function for a Rayleigh random variable with parameter 2 is

$$F_U(r) = 1 - e^{-2r^2} \qquad r \geq 0$$

Thus, if the spatial Poisson model is correct for large N, we must have

$$\lim_{N\to\infty} [\ln\{1 - F_{R_N}(r)\}] = \ln[1 - F_U(r)] = -2r^2 \qquad (*)$$

Prove (*).

c. Explain briefly how your analysis in part (b) is modified if you do not condition on the position of the incident being at the center of the square.

d. How might we use this result in developing an approximate model for travel time in a finite homogeneous region with N response units, demand rate λ (incidents/hour), average service time μ^{-1}, and a service discipline that assigns units completing service to the closest waiting call?

5.17 *Coverage problem for urban service systems* Consider a collection of response units in the plane whose positions are distributed according to a spatial Poisson process with parameter $\lambda A(S)$, where $A(S)$ denotes the area of the region S. Each unit is available for dispatch with probability $(1 - \rho)$, independent of the status of all other units. Units have different response speeds: the (Euclidean) distance that a randomly chosen available unit can travel in a time T is determined by the probability density function $f_R(r, T)$ (T fixed). Show that the number of available response units which can travel to a random incident in the time T is a Poisson random variable with parameter

$$\lambda(1 - \rho) \int_r^\infty \pi r^2 f_R(r, T)\, dr$$

Hint: Define the family of random variables $C(r, T) \equiv$ the number of available response units that can get to an incident in time T *and* that are located at a distance less than r from the incident. Show that the family $C(r, T)$ determines a time-varying Poisson process where the time variable is taken to be the distance r. To do this, prove that a "Poisson event" occurring in a ring between r and $r + dr$ centered at the incident has probability $\lambda 2\pi r\, dr \int_r^\infty f_R(x, T)\, dx$, and events occurring over disjoint intervals constitute independent random variables. Then show that $C(r, T)$ is a variable-time (nonhomogeneous) Poisson process with parameter

$$\lambda(r) = \lambda(1 - \rho)2\pi r \int_r^\infty f_R(x, T)\, dx$$

5.18 *How many police cars?* (This problem has been adopted from a real-world situation.) The police department of a medium-sized city recently decided to expand police services and, among other things, placed two additional patrol cars in the streets (over and above those they already had) on a 24-hour-a-day basis. Both before and after the addition of the two cars, it was found that each car on the streets was responding to about one call every 2 hours (apparently the total number of calls also increased somewhat) and that the mean service time for a call was about 30 minutes. It was also found that in the first 4 months after the addition of the two

new cars, the average travel time to a call was 5.28 minutes. Before the addition of the two cars it was 5.82 minutes. The police dispatcher, throughout this time, was using an approximate closest-available-car dispatching strategy ("approximate" because the dispatcher does not know the *exact* position of every police car at all times).

Based on this information can you guess approximately how many patrol cars this city's police department was fielding before the addition of the two new patrol cars?

Hint: One of the coauthors, with essentially the same information as you have (but not quite as neatly presented) guessed "12 or 13," using the techniques of Section 5.8. The right answer turned out to be 13!