E-Companion—"Approximating Vehicle Dispatch Probabilities for Emergency Service Systems with Location-Specific Service Times and Multiple Units per Location" by Susan Budge, Armann Ingolfsson, and Erhan Erkut.

This online companion contains the following additional material:

- Appendix A: Literature Review
- Appendix B: Comparison of the AH Model to Simulation and the Exact Hypercube Model
- Appendix C: Derivation and Magnitude of Station-Specific Correction Factors
- Appendix D: Proof of Theorem 1
- Appendix E: Sensitivity to Service Time Distribution
- Appendix F: Additional Computational Results

Appendix A: Literature Review

Two main streams of literature are relevant to the problem of considering server unavailability in emergency response systems. The first is that on the development of analytical models that allow for the calculation of measures related to server availability. The second is that related to location models for emergency service systems that incorporate such measures. Table A-1 summarizes the methods that will be discussed here in terms of the assumptions made about the four areas of the system outlined in the paper. Note that for the first three characteristics, the column heading is the characteristic and for each model we state the assumption made about that characteristic, but for the last characteristic (server cooperation) we focus on a specific aspect of the characteristic (server dependence) and only provide information about that aspect within the table. The reason for this is that all of the models incorporate some information about server cooperation, typically in the form of a "closest available ambulance" dispatch rule, and the main differences between the models in terms of this characteristic are in the way that they model the server dependence aspect. Additionally, we have attempted to list the models in order of increasing realism, although given the variety of assumptions, in some cases the order is admittedly subjective.

A major development in the first stream is the hypercube queueing model (Larson, 1974), which models server cooperation and dependence between servers in spatial queueing systems. This model allows the exact calculation of server-specific busy fractions and dispatch probabilities. For an s-server system, it involves the solution of 2^s simultaneous equations, and as a consequence it is

computationally intensive for large systems. Larson (1975) and Jarvis (1985) calculate server-specific busy fractions and dispatch probabilities with dependence using approximations to the hypercube model that assume servers are sampled randomly without replacement from an $M/M/s/\infty$ system (Larson, 1975) or an $M/M/s/s$ system (Larson, 1975, and Jarvis, 1985). In addition to the improvement in tractability offered by these approximate models, Jarvis' model allows one to consider service times that depend on the server and the customer, so that variations in the components of the service time that depend on the call location can be taken into account. Larson and McKnew (1982) extend both the exact and approximate hypercube model to allow for three possible server states, corresponding to idle, on patrol, or busy, in a police context. In a similar vein, Birge and Pollock (1989) formulate a method in which a system of non-linear equations is solved iteratively in order to approximate a much larger exact linear equation system. Like Larson and McKnew's approach, Birge and Pollock's method is not restricted to binary server states.

Table A-1: Summary of model assumptions of previous literature involving methods for estimating busy fractions and dispatch probabilities.

Reference	No. of vehicles per station	Average workload	Average service time	Server dependence
Daskin (MEXCLP, 1983)	Multiple	Constant	Constant	None
ReVelle and Hogan (1988)	Single	Allowed to vary	Constant	Within small regions only
Birge and Pollock (1989)	Single	Allowed to vary	Dependent on server location and call location	None
Goldberg and Szidarovszky (1991, 1991b)	Single	Allowed to vary	Dependent on server location and call location	None
Goldberg and Paz (1991)	Multiple	Allowed to vary	Dependent on server location and call location	None
Goldberg and Szidarovszky (1991c)	Multiple	Allowed to vary	Dependent on server location and call location	Within a station only
Larson (Approximate Hypercube, 1975)	Single	Allowed to vary	Constant	Yes – modeled approximately
Larson (Exact Hypercube, 1974)	Multiple	Allowed to vary	Dependent on server	Yes – modeled exactly
Jarvis (1985)	Single	Allowed to vary	Dependent on server location and call location	Yes – modeled approximately
Larson and McKnew (Exact version, 1982)	Multiple	Allowed to vary	Dependent on server	Yes – modeled exactly
Larson and McKnew (Approximate version, 1982)	Single	Allowed to vary	Dependent on server	Yes – modeled approximately
Goldberg and Benitez (1990)	Single	Allowed to vary	Dependent on server location and call location	Yes – modeled approximately
Burwell, Jarvis, and McKnew (1993)	Multiple	Allowed to vary	Dependent on server location and call location	Yes – modeled approximately

When there is more than one server located at a particular station it would usually be desirable to distribute the station's workload evenly between those servers and so these ambulances should be dispatched with equal probability to incoming calls. When two or more servers are equally preferred in the dispatch order for a particular demand location, it is referred to as a preference tie. Burwell, Jarvis, and McKnew (1993) extend the hypercube approximations by providing ways to account for preference ties and co-located servers. They suggest a "modified internal stacking method," that computes server-specific utilization and dispatch probabilities in the presence of arbitrary preference ties, making use of the correction factors developed by Larson (1975).

It is possible to use the server-specific approximation approaches developed by Larson (1975) and Jarvis (1985) to compute station-specific performance measures for systems with multiple vehicles at some stations using an approach that Burwell (1986) termed *post-averaging*. Burwell (1986) proves the validity of this approach for a special case where average service times depend only on the server and only one station has multiple servers, and he provides a counter-example that demonstrates that the approach is not always valid. We consider it plausible that post-averaging is a correct approach in the setting that we consider, where several stations may have multiple vehicles, preference ties occur only because of co-location, and average service times depend both on the server location and the call location, but this has not been proven, to our knowledge. When the input data is organized by station, the post-averaging approach requires pre- and post-processing. The pre-processing involves arbitrarily breaking preference ties, and it can be done as shown in Figure A-1. The post-processing procedure averages performance measures over servers that, in reality, are given equal preference by all demand locations.

Goldberg et al. (1990) describe a method for calculating server-specific busy fractions in order to calculate expected coverage in the objective function of their optimization problem. A number of related papers present extensions to this model (including allowing co-located servers) (Goldberg and Paz, 1991, Goldberg and Szidarovszky 1991c), and provide a focus on estimation of the server busy fractions (Goldberg and Szidarovszky, 1991, 1991a-c). Many of these works include an assumption of independence between servers and in one (Goldberg and Szidarovszky, 1991c), the authors suggest a way to improve the accuracy of the estimated busy fractions by including correction factors similar to those of Jarvis, but state that these had not been developed for the extensions in that paper (multiple vehicles per station and multiple vehicles responding to a call). The model for multiple vehicles per station does account for dependence among the vehicles in each station (using Erlang's loss formula) but assumes server independence between stations. One paper (Goldberg and Benitez, 1990) presents a

method (the "Decomposition method") for approximately calculating server busy fractions without assuming independence and compares the results to those obtained using Jarvis' approximate method as well as an approximation that assumes independence between servers. The results indicated that both the Decomposition method and Jarvis' method performed better than the method with the independence assumption, and that as the system load increased the differences became more pronounced. They also found that the Decomposition method and Jarvis' method performed equally well for low loads, but that Jarvis' method performed better for higher loads.

Some lessons from the papers of Goldberg and Szidarovszky (1991, 1991a-c) are relevant to our work. The first is that for estimating the server utilization, a Seidel iterative process converges at a faster rate than a Normal iterative process over a broad range of cases. The next is that it is valuable to formulate the problem in such a way that the server busy fractions at each step of the iterative process will always stay in the range [0, 1]. Finally, they suggest a way to deal with incorporation of correction factors to correct for the assumption of independence, without affecting the convergence results. In particular, they were able to provide theoretical convergence guarantees (i.e., a set of sufficient conditions that guarantee convergence) under the independence assumption, but could only extend these to the approximate hypercube procedure by assuming a single server at each station and that the correction factors were pre-specified constants, independent of the system load.

```
for each demand node j in N do:
    k ← 1, i ← 1, n ← 1
    while k ≤ Σ_i s_i
            τ'_kj ← τ_ij
            a'_kj ← a_ij + (n − 1)/Σ_i s_i
            n ← n + 1
            if n > s_i then
                    n ← 1
                    i ← i + 1
            end if
            k ← k + 1
    end while
    sort {a'_1j, a'_2j, …} in ascending order, replace each entry with its rank.
end do
```

Figure A-1: Pre-processing of station-level data for use with a server-based approximate hypercube model. This procedure expands the station-demand node preference matrix into a server-demand node preference matrix (a'_{kj}) and similarly expands the average service time matrix to create (τ'_{kj}), by replacing the column for each station with multiple vehicles with columns for each vehicle.

The second stream of literature, location models for emergency service systems that incorporate methods of modeling server unavailability, is relevant in terms of motivating this work. Taken together, these papers highlight the importance of modeling server unavailability and specifically, the need for models that take into account the aspects of emergency service systems that we focus on (demand variation by station, multiple servers per station, customer/server dependant service times, and server cooperation). Brotcorne, Laporte, and Semet (2003) provide a recent survey of this stream. Here, we mention only a few papers that are representative of how models of ambulance availability have been embedded in optimization models for facility location: Daskin (1983), who incorporated a system-wide busy fraction into the maximal covering location problem of Church and ReVelle (1974), and ReVelle and Hogan (1988, 1989), who incorporate local estimates of ambulance unavailability (region-specific busy fractions).

Our paper's main contribution is to adapt the approximate hypercube model for use with station-specific data and to prove that a restricted version of the model is guaranteed to converge.

Appendix B: Comparison of the AH Model to Simulation and the Exact Hypercube Model

The system we deal with can be described as a multi-server loss system with distinguishable servers. Analysts who would like to evaluate the performance of such systems have three main options: (1) exact numerical solution, (2) discrete event simulation, and (3) an approximate approach, such as the approximate hypercube model. These three approaches have different strengths and weaknesses and all of them deserve to be in an analyst's toolkit. The advantage of an exact approach is obvious: it provides exact answers. The main drawback is that the size of the state space (and, therefore, the required storage space) needed to model the system as a Markov process grows rapidly with system size and the level of detail that the model includes, i.e., the "curse of dimensionality." The advantages of simulation include the flexibility to represent as much detail as desired and the availability of powerful and relatively easy-to-use software. The drawbacks of simulation are that all answers are subject to sampling error and run lengths required to achieve acceptable accuracy are highly dependent on such system characteristics as the number of servers, the average system utilization, and even which performance measure is to be estimated. Koopman (1972) discusses the pros and cons of an exact approach versus simulation in greater detail, in the context of queueing systems with time-varying parameters.

Against this backdrop, the advantages of the approximate hypercube model are easily stated: it is fast, with computation times that grow slowly with system size, and it is sufficiently accurate for most practical purposes.

To make matters more concrete, consider how large an instance of the exact hypercube model is solvable with current computing technology. The size of the state space for this model is 2^s, where s is the number of servers. At a minimum, one must store the entries in the vector of steady state probabilities (the entries in the transition rate matrix could be computed as needed, rather than stored). With single precision floating point arithmetic (requiring 8 bytes per entry) and a 1,000 Gb hard drive (the largest capacity currently available), s can be at most $\left\lfloor \log_2 \left(1000 \times 10^9 / 8\right) \right\rfloor = 36$. In contrast, we have used the approximate hypercube model for systems with as many as 57 servers (the Calgary Health Region, which includes the city of Calgary and adjacent rural communities).

For simulation, Srikant and Whitt (1996) provide approximate formulas for run lengths to estimate blocking probabilities in loss systems with specified precision. They argue that the computational effort to simulate a loss system is approximately proportional to the arrival rate λ times the simulation run length t, or in other words, to the expected number of simulated calls for service. From equations (9), (14), (15), (17), and (20) in their paper, we obtain the following approximation for the simulation computational effort for a lightly loaded loss system (one with $\lambda\tau < s$) with a Poisson arrival process and i.i.d. service times with coefficient of variation equal to 1:

$$\lambda t \approx \frac{z_{\beta/2}^2}{\varepsilon^2} \sqrt{\frac{2}{\pi\gamma^2}} e^{-\gamma^2/2}.$$

Here, $z_{\beta/2}$ is the upper $\beta/2$ percentile of a standard normal distribution, ε is the desired $1-\beta$ confidence interval half width for the blocking probability, and $\gamma = (\lambda\tau - s)/\sqrt{\lambda\tau}$. Figure A-1 shows the computational effort as a function of the number of servers, for $\rho = \lambda\tau/s$ equal to 30%, 50%, and 70%, assuming that a 95% confidence interval with a half width of 0.01% for the blocking probability is desired.

As Figure B-1 makes clear, the simulation effort increases rapidly with ρ and decreases rapidly with the number of servers. Obtaining a reliable estimate for the blocking probability, however, may not be sufficient. More important would be to obtain reliable estimates of the dispatch probabilities for each demand node, for the most preferred servers. Estimating the simulation effort required to obtain

reliable estimates of these probabilities is a topic worthy of further research. However, the available estimates for run lengths required to estimate blocking probabilities suggest that simulation is most efficient when utilization is low and the number of servers is large.

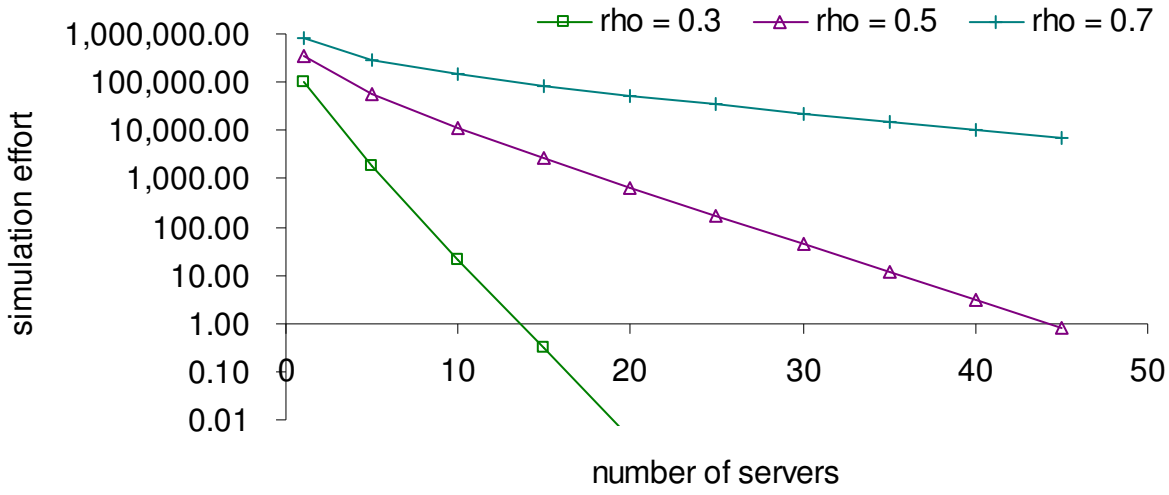

Figure B-1: Simulation effort required to estimate blocking probabilities in a loss system to 0.01% with 95% confidence, as a function of number of servers and offered load per server (rho).

To conclude, an exact approach is feasible with a small number of servers and simulation is likely to be effective for systems with low utilization and a large number of servers. In contrast to these more accurate approaches, the performance of the approximate hypercube model is far less sensitive to the number of servers and the system utilization.

Appendix C: Derivation and Magnitude of Station-Specific Correction Factors

Consider an $M/M/s/s$ system with arrival rate λ, average service time τ (we discuss how to estimate τ later), and let $\rho = \lambda\tau/s$. We will refer to this $M/M/s/s$ system (which approximates the real system) as the *parallel system*. To simplify notation we suppress the dependence of various quantities on the node j. We establish a correspondence between the parallel system and the real system as follows. When a call arrives from node j, the dispatcher in the real system first checks whether any of the $s_{(1)}$ ambulances at the most preferred station for that node are available. If none are available, then the dispatcher checks whether any of the $s_{(2)}$ ambulances at the second-most preferred station are

free, and so on, until a station is found with at least one free ambulance. The corresponding sequence of events in the parallel system is to first select $s_{(1)}$ servers at random and check whether at least one of them is idle. If not, then select $s_{(2)}$ servers at random from the $s - s_{(1)}$ servers that have not been checked already (i.e., sampling without replacement) and continue in this manner until a station with at least one free server is found.

With this correspondence in mind, we define the following events for the parallel system:

S_n : exactly n servers are busy
B_k : all servers at k^{th} preferred station are busy
F_k : the k^{th} preferred station has at least one free server

Additionally, we define $B_{1,n} \equiv B_1 \cap B_2 \cap \cdots \cap B_n$.

Using the law of total probability, we can express the probability that the first free server is found at the k^{th} preferred station as

$$\begin{aligned} \Pr\{B_{1,k-1} \cap F_k\} &= \sum_{n=1}^{s} \Pr\{B_{1,k-1} \cap F_k \mid S_n\} P_n \\ &= \sum_{n=1}^{s} \Pr\{B_{1,k-1} \mid S_n\} \Pr\{F_k \mid B_{1,k-1} \cap S_n\} P_n \end{aligned} \tag{C-1}$$

Letting $z_{(k-1)} = s_{(1)} + s_{(2)} + \ldots + s_{(k-1)}$ be the total number of ambulances at the $k-1$ most preferred stations, consider the probability that all of these ambulances are busy, given that a total of n servers are busy, i.e., $\Pr\{B_{1,k-1} \mid S_n\}$. If u ambulances have been checked and found to be busy, then the chances that the $(u+1)^{\text{st}}$ ambulance checked is busy are $(n-u)/(s-u)$. It follows that

$$\Pr\{B_{1,k-1} \mid S_n\} = \begin{cases} 0 & \text{if } k = 1 \text{ or } z_{(k-1)} > n \\ \prod_{u=0}^{z_{(k-1)}-1} \dfrac{n-u}{s-u} & \text{if } k > 1 \text{ and } z_{(k-1)} \le n \end{cases}. \tag{C-2}$$

The probability that the k^{th} preferred station has at least one free ambulance, given that all ambulances at the $k-1$ most preferred stations are busy and a total of n ambulances are busy is

$$\Pr\{F_k \mid B_{1,k-1} \cap S_n\} = 1 - \Pr\{B_k \mid B_{1,k-1} \cap S_n\}$$
$$= \begin{cases} 1 & \text{if } z_{(k)} > n \\ 1 - \prod_{u=0}^{s_{(k)}-1} \dfrac{n-(z_{(k-1)}+u)}{s-(z_{(k-1)}+u)} & \text{if } z_{(k)} \le n \end{cases} \qquad \text{(C-3)}$$

Combining (C-1) – (C-3) and substituting $P_n = (\rho s)^n P_0 / n!$ results in

$$\Pr\{B_{1,k-1} \cap F_k\} = P_0 \sum_{n=z_{(k-1)}}^{s-1} \frac{(\rho s)^n}{n!} \prod_{u=0}^{z_{(k-1)}-1} \frac{n-u}{s-u} \left[1 - \prod_{u=0}^{s_{(k)}-1} \frac{n-(z_{(k-1)}+u)}{s-(z_{(k-1)}+u)} \right]$$
$$= P_0 \sum_{n=z_{(k-1)}}^{s-1} \frac{(\rho s)^n}{n!} \left[\prod_{u=0}^{z_{(k-1)}-1} \frac{n-u}{s-u} - \prod_{u=0}^{z_{(k)}-1} \frac{n-u}{s-u} \right] \qquad \text{(C-4)}$$

Now it is necessary to relate $\Pr\{B_{1,k-1} \cap F_k\}$ to the dispatch probabilities f_{ij} of the real system. In the parallel system, the fraction of time each server is busy is $r = \rho(1 - P_s)$. Therefore, we set

$$\Pr\{B_{1,k-1} \cap F_k\} = Q_j(\{s_{(k)j}\}, \rho, k) r^{z_{(k-1)}} \left(1 - r^{s_{(k)}}\right) \qquad \text{(C-5)}$$

Solving for the correction factor, substituting (C-4), and re-introducing the subscript, *j*, where appropriate to show the dependence on the demand node, gives

$$Q_j(\{s_{(k)j}\}, \rho, k) = \frac{P_0 \sum_{n=z_{(k-1)j}}^{s-1} \frac{(\rho s)^n}{n!} \left[\prod_{u=0}^{z_{(k-1)j}-1} \frac{n-u}{s-u} - \prod_{u=0}^{z_{(k)j}-1} \frac{n-u}{s-u} \right]}{r^{z_{(k-1)j}} \left(1 - r^{s_{(k)j}}\right)} \qquad \text{(C-6)}$$

Figure C-1 illustrates how $Q_j(\{s_{(k)j}\}, \rho, k)$ varies with ρ and *k* and for different scenarios $\{s_{(k)}\}$. In the first scenario (Panel 1), each of ten stations has one server. In this case, (C-6) reduces to the correction factor formula (5) from Jarvis (1985). The second scenario (Panel 2) is identical to the first except that the fourth preferred station has two servers. The third scenario (Panel 3) has two servers at the second preferred station and three servers at the fourth preferred station. In Panel 4, the correction factors for a number of different scenarios, or $\{s_{(k)}\}$, are shown as identified in the legend, all for $\rho = 0.4$. By comparing the graphs, one can see that increasing the number of servers at a particular station results in steeper functions beyond that station (towards the less preferred stations), and that the impact is much larger for lower values of system load. It is also evident from the graph in Panel 3 that as the number of servers increases, the linearly interpolated curves will not necessarily remain convex.

An additional insight from Panel 4 is that the correction factors are the same for the same number of total more preferred servers at a given k, even if the $\{s_{(k)}\}$ vector up to that k is not the same (e.g., scenarios $\{1,3,1,1,1,1,1,1,1,1\}$ and $\{1,1,1,1,1,3,1,1,1,1\}$). Finally, note that although the vertical axis is truncated (at a value of 4 for the first 3 panels, and a value of 30 in the fourth panel), the values of Q can be much higher than this, especially at low loads and when there are multiple servers in the most preferred stations (at the lower values of k).

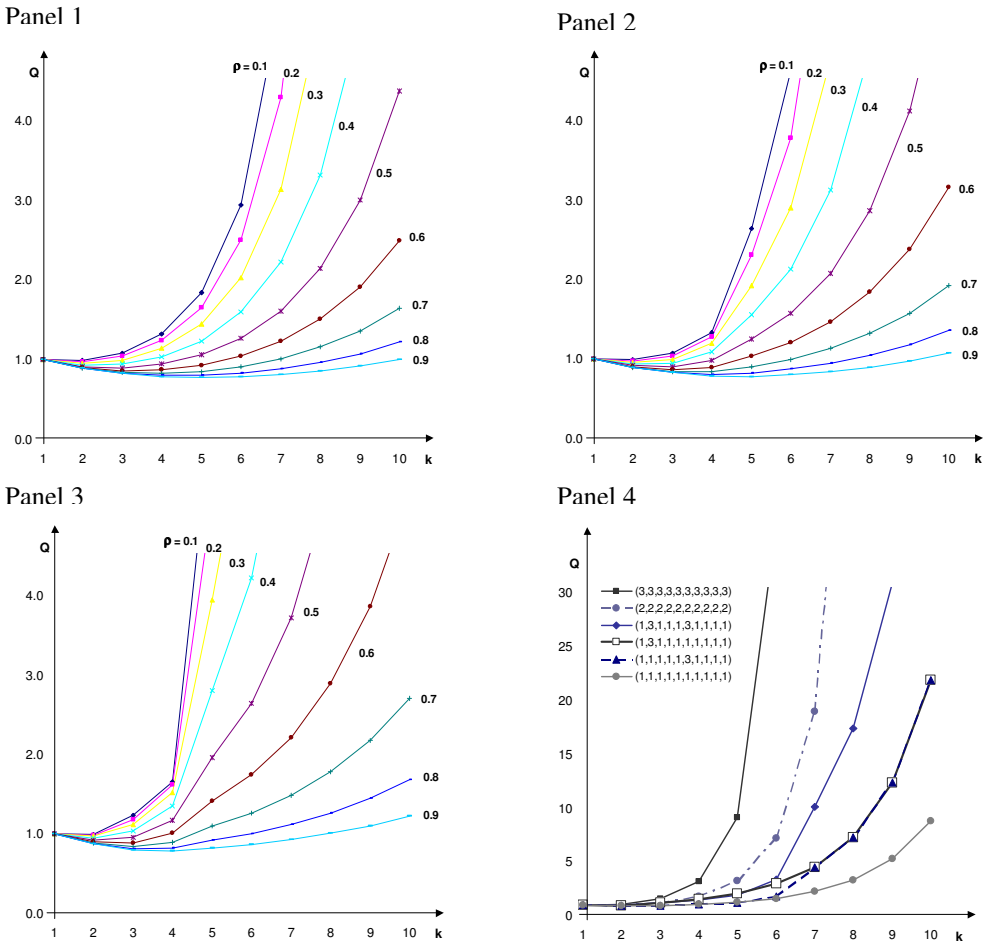

Figure C-1: Graphs of $Q_j\left(\{s_{(k)j}\},\rho,k\right)$. The first panel is for one server per station, the second panel has an additional server at the fourth preferred station, and the third panel has an additional server at the second preferred station and two additional servers at the fourth preferred station. Panel 4 is for $\rho = 0.4$ and gives a number of different scenarios, or $\{s_{(k)}\}$, as identified in the legend.

Appendix D: Proof of Theorem 1

Proof: Under the assumptions of the theorem, the algorithm can be described as follows:

$$r_i^0 = 1, \quad i = 1,2,\ldots,I$$

$$V_i^h = f_i(r_1^{h-1}, r_2^{h-1}, \ldots, r_I^h) \equiv \sum_{j=1}^{J} c_{ij} \prod_{l=1}^{a_{ij}-1} \left(r_{(l)j}^{\;h-1} \right)^{s_{(l)j}}, \quad h = 1,2,\ldots, \quad i = 1,2,\ldots,I$$

$$r_i^h = g_i(V_i^h, r_i^{h-1}) \equiv \left(\frac{V_i^h}{V_i^h + s_i \Big/ \left(r_i^{h-1} \right)^{s_i - 1}} \right)^{\frac{1}{s_i}}, \quad h = 1,2,\ldots, \quad i = 1,2,\ldots,I$$

where $c_{ij} = \lambda_j \tau_{ij} Q_j(\{s_{(a_{ij})j}\}, \rho, a_{ij})$ are positive constants. We will use induction to prove that the sequence $\{r_i^h\}_{h=0}^{\infty}$ is non-increasing and bounded in the range [0, 1], for each *i*, which implies that these sequences converge. Suppose that $1 \geq r_i^{h-1} \geq r_i^h \geq 0, i = 1,2,\ldots,I$. This implies that $V_i^h \geq V_i^{h+1} \geq 0, i = 1,2,\ldots,I$, because f_i is non-decreasing in all of its arguments. In turn, $1 \geq r_i^{h-1} \geq r_i^h \geq 0$ and $V_i^h \geq V_i^{h+1} \geq 0$ taken together imply that $1 \geq r_i^h \geq r_i^{h+1}$ because g_i is non-decreasing in its two arguments. Furthermore, it is easy to verify that $r_i^{h+1} \geq 0$. Finally, direct substitution demonstrates that $1 = r_i^0 > r_i^1 \geq 0, i = 1,2,\ldots,I$ and this completes the induction proof. Q.E.D.

Appendix E: Sensitivity to Service Time Distribution

The parallel $M/M/s/s$ system assumes exponentially distributed service times and the exact model that is closest to the system that we model (the exact hypercube model) also assumes exponentially distributed service times, with a mean that depends on the responding server (but not on the call location). In reality, the service time distributions are likely to have a coefficient of variation (CV) that is considerably smaller than 1 (the value for an exponential distribution) because the service time is the sum of components including the chute time, the travel time, and the on-scene time, all of which have CVs less than one. Fortunately, the occupancy probabilities in the $M/M/s/s$ model are insensitive to the shape of the service time distribution beyond the mean (see, e.g., Gross and Harris, 1998). Computational experiments by Jarvis (1975) indicate that although results for the loss version of the exact hypercube model are not completely insensitive to the shape of the service time distribution, its impact is extremely small.

We will use two examples, adapted from Jarvis (1975), to test how sensitive the steady state probabilities for a loss system with distinguishable servers are to the shape of the service time distribution. Jarvis uses distributions with a CV of 1 (an exponential distribution) and strictly between 0 and 1 (a convolution of two exponential distributions). We complement Jarvis' exact numerical results with simulation results for a system with deterministic service times (CV = 0), which allows us to see the range of results when the CV varies from 0 and 1. Both examples assume two stations with one server each and two demand nodes. In one example, the mean of the service time distribution depends only on the server and in the second example, the mean depends both on the server and the demand node.

We implemented the simulation with the SSJ library (L'Ecuyer, 2004). We simulated approximately 3×10^8 arrivals for each system, which resulted in 95% confidence interval half-widths (computed using a batch means approach) of less than 10^{-4} for all estimated probabilities shown below.

The data for the first example is shown in Table D-1. Jarvis computed the steady state probabilities of four system states (empty, unit 1 busy, unit 2 busy, both units busy) for both an exponential service time distribution and a convolution of two exponential distributions. Table D-2 compares his results to simulation results for deterministic service times (to validate our simulation model, we also simulated the two systems that Jarvis solved. The results agree to three digits.) We see that the impact of the service time distribution on the steady state probabilities is quite small, appearing only in the third digit. Note that the system is so heavily loaded that almost 50% of the calls are lost, so it represents an extreme situation that is unlikely to occur in reality. We would expect to see even less sensitivity to the shape of the service time distribution for more realistic examples.

Table D-1: Data for first example.

Demand node	Arrival rate	Server	Mean service time	Mean of 1st exp. component	Mean of 2nd exp. component
1	1	1	3/2	1	1/2
2	2	2	7/12	1/3	1/4

Table D-2: Results for first example.

	Service time distribution		
State	Exponential*	Sum of exponentials*	Deterministic
Empty	0.128729	0.127742	0.125
Unit 1 busy	0.262	0.263	0.265
Unit 2 busy	0.123	0.124	0.125
Both units busy	0.486	0.486	0.485

* From Jarvis (1975)

In the second example, the arrival rates for the two demand nodes remain the same. Now, we assume that station 1 is the closest to demand node 1 and that station 2 is the closest to demand node 2. The average service times remain the same as before when the closest unit responds but we add 1/2 time unit to the average service time if the unit that responds is not the closest. Table D-3 shows simulation results for this example, with exponential and deterministic service time distributions. Here, we see larger differences than in the first example, but still, when rounded to two digits, all of the probabilities are the same for the two service time distributions.

Table D-3: Results for second example.

State	Exponential	Deterministic
Empty	0.103	0.097
Unit 1 busy	0.239	0.243
Unit 2 busy	0.114	0.117
Both units busy	0.544	0.544

These simulation results strengthen Jarvis' conclusion that the shape of the service time distribution beyond its mean has a small impact on steady state probabilities for loss systems with distinguishable servers.

Appendix F: Additional Computational Results

Magnitude of Correction Factors

Figure F-1 provides information about the correction factors for the subset of cases (648 in total) of the Edmonton dataset in which all stations had at least one ambulance. We tabulated the correction factors, Q, for these scenarios based on the value of *k*, the preference of the station in the response list for the demand node, using small bin sizes at the low end of the scale and larger bin sizes for higher values of Q (and hence use a log scale for the horizontal axis in the figure on the right). The graphs give, for each bin, the relative portion of the correction factors for various station preferences, *k*. As the graphs show, in general, less preferred stations have higher correction factors. However, even second and third preferred stations have correction factors that are in some cases considerably higher than one.

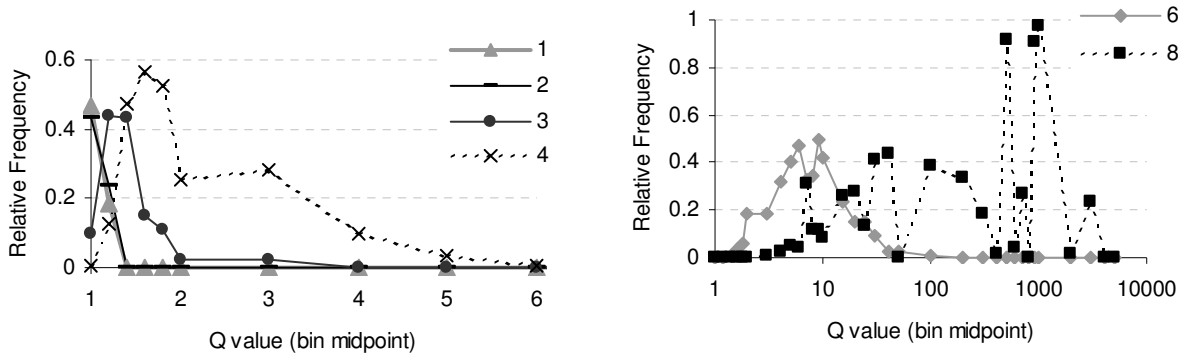

Figure F-1: Relative frequencies of correction factors by station preference (the first through fourth preferred stations are shown on the left graph, and the sixth and eighth preferred stations are shown on the right graph) for 648 scenarios of the Edmonton dataset. The horizontal axis for the right graph uses a log scale.

Measurement Error

In our experimental design to evaluate measurement error, we considered 4, 8, or 10 open stations, shown in Table F-1, together with the fraction of total demand from each station's district. We considered a station's "district" to consist of all demand nodes for which that station was closest. For each set of open stations, we used four different ambulance allocations, shown in Table F-2. This table shows, for each allocation and each station, the number of ambulances as well as the station's district demand divided by the number of ambulances. The latter number provides a first-order approximation of the workload imposed on each ambulance. In the rightmost column, we show the maximum difference between these workload estimates, which is a measure of how *balanced* an allocation is.

Table F-1: Fraction of total demand from each station's district.

Station number	1	2	3	4	5	6	7	8	9	10	Total
Using stations 6, 7, 9, and 10						24%	13%		22%	41%	100%
Using stations 1 and 4-10	8%			6%	11%	13%	13%	13%	8%	29%	100%
Using all stations.	8%	7%	3%	6%	8%	7%	12%	13%	8%	29%	100%

Table F-2: Ambulance allocations. Each cell shows the number of ambulances allocated to the station, with the station's district demand divided by the number of ambulances in parentheses. The rightmost column shows the maximum difference between the numbers in parentheses.

	Station										
Allocation	1	2	3	4	5	6	7	8	9	10	Diff.
1						1 (24%)	1 (13%)		1 (22%)	1 (41%)	0.28
2						1 (24%)	1 (13%)		2 (11%)	2 (20%)	0.13
3						2 (12%)	1 (13%)		2 (11%)	3 (14%)	0.03
4						2 (12%)	2 (7%)		2 (11%)	2 (20%)	0.14
5	1 (8%)			1 (6%)	1 (11%)	1 (13%)	1 (13%)	1 (13%)	1 (8%)	1 (29%)	0.23
6	1 (8%)			1 (6%)	1 (11%)	2 (6%)	2 (7%)	1 (13%)	2 (4%)	2 (14%)	0.11
7	1 (8%)			1 (6%)	2 (6%)	2 (6%)	2 (7%)	2 (6%)	3 (3%)	3 (10%)	0.07
8	2 (4%)			2 (3%)	2 (6%)	2 (6%)	2 (7%)	2 (6%)	2 (4%)	2 (14%)	0.11
10	1 (8%)	1 (7%)	1 (3%)	1 (6%)	1 (8%)	1 (7%)	1 (12%)	1 (13%)	1 (8%)	1 (29%)	0.25
10	1 (8%)	2 (3%)	1 (3%)	1 (6%)	1 (8%)	1 (7%)	2 (6%)	2 (6%)	1 (8%)	2 (14%)	0.11
10	1 (8%)	2 (3%)	1 (3%)	2 (3%)	2 (4%)	2 (3%)	3 (4%)	2 (6%)	2 (4%)	3 (10%)	0.07
10	2 (4%)	2 (3%)	2 (2%)	2 (3%)	2 (4%)	2 (3%)	2 (6%)	2 (6%)	2 (4%)	2 (14%)	0.13

We simulated each allocation for system loads ($\rho = \lambda\tau / s$) ranging from 0.1 to 0.9 by varying the total arrival rate of calls to the system. The system load values are approximate because the average service time, τ, depends on the individual vehicle utilizations and so we estimated the average service time value (assuming an average system-wide utilization to calculate the dispatch probabilities). See the published version of the main paper for a comparison of the results of the simulation and the results of our approximation procedure. The relative error is generally under 2%.

References

Also see the list of references in the published version of the paper.

Birge J, S. Pollock (1989). Using Parallel Iteration for Approximate Analysis of a Multiple Server Queueing System. *Operations Research* **37** 769-779.

Brotcorne, L., G. Laporte, F. Semet (2003). Ambulance Location and Relocation Models. *European Journal of Operational Research* **147** 451-463.

Budge, S., A. Ingolfsson, E. Erkut (2006). Optimal Ambulance Location with Random Delays and Travel Times, working paper, available from http://www.business.ualberta.ca/aingolfsson/publications.htm

Church, R., C. ReVelle (1974). The Maximal Covering Location Problem. *Papers of the Regional Science Association* **32** 101-118.

Daskin, M. S. (1983). A Maximum Expected Covering Location Model: Formulation, Properties, and Heuristic Solution. *Transportation Science* **17** 48–70.

Goldberg, J., R. Benitez (1990). Evaluating Bias in Models to Approximate Performance in Emergency Vehicle Systems. Working Paper. Department of Systems and Industrial Engineering, University of Arizona, Tucson.

Goldberg, J., R. Dietrich, J. M. Chen, M. G. Mitwasi, T. Valenzuela, and E. Criss (1990). Validating and Applying a Model for Locating Emergency Medical Vehicles in Tucson, AZ. *European Journal of Operational Research* **49** 308–324.

Goldberg, J., L. Paz (1991). Locating Emergency Vehicle Bases when Service Time Depends on Call Location. *Transportation Science* **25** 264–280.

Goldberg, J., F. Szidarovszky (1991a). A General Model and Convergence Results for Determining Vehicle Utilization in Emergency Systems. *Communications in Statistics – Stochastic Models* **7** 137-160.

Goldberg, J., F. Szidarovszky (1991b). A Model for Determining Emergency Vehicle Utilization Under an Infinite Queue and Location Dependent Service Times. Working Paper. Department of Systems and Industrial Engineering, University of Arizona, Tucson.

Goldberg J., F. Szidarovszky (1991c). Extended Models for Determining Emergency Vehicle Busy Probabilities. Working Paper. Department of Systems and Industrial Engineering, University of Arizona, Tucson.

Jarvis, J. (1975). Optimization in Stochastic Service Systems with Distinguishable Servers. Ph.D. dissertation, Massachusetts Institute of Technology.

Koopman, B. O. (1972). Air-Terminal Queues under Time-Depentent Conditions. *Operations Research* **20** 1089–1114.

Larson, R. C., M. A. McKnew (1982). Police Patrol-Initiated Activities within a System Queueing Model. *Management Science* **28** 759-774

L'Ecuyer, P. (2004). SSJ: A Java Library for Stochastic Simulation, Software user's guide, Available at http://www.iro.umontreal.ca/~lecuyer.

ReVelle, C., K. Hogan (1988). A Reliability-Constrained Siting Model with Local Estimates of Busy Fractions. *Environment and Planning B: Planning and Design* **15** 143–152.

ReVelle, C., K. Hogan (1989). The Maximum Availability Location Problem. *Transportation Science* **23** 192–200.

Srikant, R., W. Whitt (1996). Simulation Run Lengths to Estimate Blocking Probabilities. *ACM Transactions on Modeling and Computer Simulation* **6** 7-52.