E-Companion — "Cross-Trained Fire-Medics Respond to Medical Calls and Fire Incidents: A Fast Algorithm for a Three-State Spatial Queuing Problem"

(Authors' names blinded for peer review)

EC.1. Miscellaneous Proofs

EC.1.1. Proof of Lemma 1

Proof: Define W_i^y as the event that unit i is busy at a type-y call, and define W_i as the event that unit i is busy. W_i^c is the event that unit i is free, which is the complement of W_i. Define V_i^y as the event that unit i is in state 0 in subsystem y. Let $-y$ denote the subsystem that is not y. Thus, $V_i^y = W_i^{-y} \cup W_i^c$. By Bayes' Theorem, we have

$$\alpha_i^y = \Pr(W_i^{-y} | V_i^y) \tag{EC.1}$$

$$= \frac{\Pr(W_i^{-y} \cap V_i^y)}{\Pr(V_i^y)} \tag{EC.2}$$

$$= \frac{\Pr(V_i^y \mid W_i^{-y}) \cdot \Pr(W_i^{-y})}{\Pr(W_i^{-y} \cup W_i^c)} \tag{EC.3}$$

Note that in the above formula, $\Pr(V_i^y \mid W_i^{-y}) = 1$ because $W_i^{-y} \subset V_i^y$. Also note that $\rho_i^{-y} = \Pr(W_i^{-y})$. We thus have in the numerator

$$\Pr(V_i^y \mid W_i^{-y}) \cdot \Pr(W_i^{-y}) = 1 \cdot \rho_i^{-y} = \rho_i^{-y} \tag{EC.4}$$

In the denominator, since W_i^{-y} and W_i^c are mutually exclusive events, we have

$$\Pr(W_i^{-y} \cup W_i^c) = \Pr(W_i^{-y}) + \Pr(W_i^c) = \rho_i^{-y} + \Pr(W_i^c) \tag{EC.5}$$

Using De Morgan's law, we have

$$W_i^c = (W_i^y \cup W_i^{-y})^c = (W_i^y)^c \cap (W_i^{-y})^c \tag{EC.6}$$

By the definition of conditional probability, we have

$$\Pr(W_i^c) = \Pr\left((W_i^y)^c \cap (W_i^{-y})^c\right) \tag{EC.7}$$

$$= \Pr\left((W_i^{-y})^c\right) \cdot \Pr\left((W_i^y)^c | (W_i^{-y})^c\right) \tag{EC.8}$$

We have $(W_i^{-y})^c = V_i^{-y}$ because $W_i^{-y} \cup V_i^{-y} = W_i^{-y} \cup W_i^y \cup W_i^c = U$, the universal set. Therefore,

$$\Pr\left((W_i^y)^c | (W_i^{-y})^c\right) = \Pr\left((W_i^y)^c | V_i^{-y}\right) \tag{EC.9}$$

$$= 1 - \Pr\left(W_i^y | V_i^{-y}\right) \tag{EC.10}$$

$$= 1 - \alpha_i^{-y} \tag{EC.11}$$

Also note that $\Pr\left((W_i^{-y})^c\right) = 1 - \Pr(W_i^{-y}) = 1 - \rho_i^{-y}$. We thus conclude that

$$\alpha_i^y = \frac{\rho_i^{-y}}{\rho_i^{-y} + (1-\rho_i^{-y})(1-\alpha_i^{-y})} \tag{EC.12}$$

□

EC.1.2. Proof of Theorem 1

Proof: Based on the Contraction Mapping Theorem, as shown in Theorem 10.6 of Burden and Faires (2011), function F has a fixed point in D if F is a continuous function from D into $\mathbb{R}^{N_1+N_c}$ such that whenever $\alpha^y \in D$, $F(\alpha^y) \in D$. We begin the proof by showing that F is a continuous function in D.

Let matrix Q^y be the transition matrix of the continuous-time Markov Chain of subsystem y characterized by the balance equations (3). We substitute the last row of Q^y by $\mathbf{1}^T$, the row vector of all 1's, and denote this matrix by A^y. We construct another matrix A_i^y by replacing the last row of Q^y by the row vector $\mathbf{e}_i$, where $e_{in} = 1$ if $B_n^y(i) = 1$, and 0 otherwise.

Without loss of generality, we only show the proof for subsystem $y = 1$. We first show that $F(\alpha^1) = (f_1(\alpha^1), \cdots, f_{N_1+N_c}(\alpha^1))$ is continuous where

$$f_i(\alpha^1) = \frac{\rho_i^1 \rho_i^2 + \rho_i^1(1-\rho_i^1)(1-\alpha_i^1)}{\rho_i^1 \rho_i^2 + (1-\rho_i^1)(1-\alpha_i^1)}, \tag{EC.13}$$

and

$$\rho_i^y = \frac{det(A_i^y)}{det(A^y)}. \tag{EC.14}$$

Since the determinant is continuous, both matrices are continuous, and the composition of continuous functions is also continuous, we conclude that f_i is a continuous function for all i. Thus, F is continuous.

We next show that $F(\alpha^1) \in D$ if $\alpha^1 \in D$. This is true because $\rho_i^y = \sum_{n:B_n^y(i)=y} P\{B_n^y\} \in [0,1]$, and if $\alpha_i^1 \in [0,1]$, we always have

$$f_i(\alpha^1) = \frac{\rho_i^1\rho_i^2 + \rho_i^1(1-\rho_i^1)(1-\alpha_i^1)}{\rho_i^1\rho_i^2 + (1-\rho_i^1)(1-\alpha_i^1)} \tag{EC.15}$$

$$= 1 - \frac{(1-\rho_i^1)^2(1-\alpha_i^1)}{\rho_i^1\rho_i^2 + (1-\rho_i^1)(1-\alpha_i^1)} \in [0,1]. \tag{EC.16}$$

Therefore, we conclude that F has a fixed point in D, and thus HDA has a fixed point. □

EC.1.3. Proof of Lemma 2

Proof: At convergence, substituting the expressions for α_i^{-y} into α_i^y using Lemma 1, we have

$$\alpha_i^y = \frac{\rho_i^{-y}}{\rho_i^{-y} + (1-\rho_i^{-y})(1 - \frac{\rho_i^y}{\rho_i^y + (1-\rho_i^y)(1-\alpha_i^y)})} \tag{EC.17}$$

$$= \frac{\rho_i^{-y}(\rho_i^y + (1-\rho_i^y)(1-\alpha_i^y))}{\rho_i^{-y}(\rho_i^y + (1-\rho_i^y)(1-\alpha_i^y)) + (1-\rho_i^{-y})(1-\rho_i^y)(1-\alpha_i^y)} \tag{EC.18}$$

$$= \frac{\rho_i^{-y}\rho_i^y + \rho_i^{-y}(1-\rho_i^y)(1-\alpha_i^y)}{\rho_i^{-y}\rho_i^y + (1-\rho_i^y)(1-\alpha_i^y)} \tag{EC.19}$$

Multiplying both sides of (EC.19) by its denominator, we have

$$\alpha_i^y\rho_i^{-y}\rho_i^y + \alpha_i^y(1-\rho_i^y)(1-\alpha_i^y) = \rho_i^{-y}\rho_i^y + \rho_i^{-y}(1-\rho_i^y)(1-\alpha_i^y) \tag{EC.20}$$

and collecting terms with $\rho_i^{-y}\rho_i^y$, we have

$$\alpha_i^y(1-\rho_i^y)(1-\alpha_i^y) = (1-\alpha_i^y)\rho_i^{-y}\rho_i^y + \rho_i^{-y}(1-\rho_i^y)(1-\alpha_i^y). \tag{EC.21}$$

Dividing both sides by $1-\alpha_i^y$, we obtain

$$\alpha_i^y(1-\rho_i^y) = \rho_i^{-y}\rho_i^y + \rho_i^{-y}(1-\rho_i^y). \tag{EC.22}$$

After collecting all terms that contain α_i^y, we obtain $\alpha_i^y = \frac{\rho_i^{-y}}{1-\rho_i^y}$. □

EC.1.4. Proof of Lemma 3

Proof: Define W_i^y as the event that unit i is busy at a type-y call, and W_i as the event that unit i is busy. We have $W_i = W_i^1 \cup W_i^2$. Also define W_i^c as the event that unit i is idle, the complement of W_i. Note that $P\{W_i^y\} = \rho_i^y$, the probability that unit i is busy at a type-y call, and $P\{W_i^c\} = 1-\rho_i$, the probability of being idle, where $\rho_i = \rho_i^1 + \rho_i^2$.

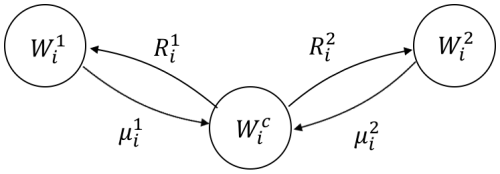


Figure EC.1 Transition Graph for Each Unit in a 3-state System

Let R_i^y be the rate at which unit i is assigned to type-y calls, given that unit i is idle. By definition, we have $S_i^y = R_i^y(1-\rho_i)$. From the transition diagram in Figure EC.1, by conservation of flow, we have the following balance equation for state W_i^y,

$$P\{W_i^y\}\mu_i^y = P\{W_i^c\}R_i^y. \tag{EC.23}$$

Rearranging terms in (EC.23), we have

$$R_i^y = \frac{P\{W_i^y\}\mu_i^y}{P\{W_i^c\}} = \frac{\rho_i^y \mu_i^y}{1-\rho_i}. \tag{EC.24}$$

Since $S_i^y = R_i^y(1-\rho_i)$, we have

$$S_i^y = R_i^y(1-\rho_i) = \frac{\rho_i^y \mu_i^y}{1-\rho_i}(1-\rho_i) = \rho_i^y \mu_i^y, \tag{EC.25}$$

which completes the proof and we obtain an expression for S_i^y. $\square$

EC.1.5. Proof of Theorem 2

Proof: To simplify notation, we omit the superscript y in the proof. We aggregate the states B_m according to the set $\mathscr{S}_k$. The aggregated states form a birth and death chain with upward transition rates $\lambda(k)$ and downward transition rates $\mu(k)$, where k is the number of busy servers.

Define $P(B_m|k)$ as the conditional probability of being in state B_m given k busy servers. Let r_m be the transition rate out of state B_m by an arrival. We have For $k=0,\ldots,N$,

$$\lambda(k) = \sum_{m\in\mathscr{S}_k} P(B_m|k) r_m, \tag{EC.26}$$

$$\mu(k) = k\mu, \tag{EC.27}$$

where μ is the common service rate if all units have the same service rate, or $1/\mu = \sum_{j=1}^{J} (\lambda_j/\lambda) \cdot \sum_{i=1}^{N} q_{ij}/\mu_i$, as developed by Jarvis (1985). In the surrogate queuing system with server vacations, given k busy units, the probability that each unit is busy is assumed to be the same. The server identity is accounted for by the α_i values for each unit i in the transition rate r_m. There are in total $\binom{N}{k}$ elements in set $\mathscr{S}_k$, so we have $P(B_m|k) = \frac{1}{|\mathscr{S}_k|} = \frac{k!(N-k)!}{N!}$. We next obtain an expression for r_m. Applying (4) in the body of the paper, we have

$$r_m = \sum_{n:d^+_{mn}=1} \bar{r}^y_{mn} = \sum_{n:d^+_{mn}=1} \sum_{j=1}^{J} \lambda_j \prod_{k:B_m(\gamma_{jk})=0}^{\eta_j(i_{mn})} \alpha_{\gamma_{jk}} (1-\alpha_{i_{mn}}). \tag{EC.28}$$

$$= \sum_{j=1}^{J} \lambda_j \underbrace{\sum_{n:d^+_{mn}=1} \prod_{k:B_m(\gamma_{jk})=0}^{\eta_j(i_{mn})} \alpha_{\gamma_{jk}} (1-\alpha_{i_{mn}})}_{=:\beta_j}. \tag{EC.29}$$

We obtain the above equality by switching the summation order since λ_j does not depend on n. In what follows, we simplify r_m by simplifying β_j. We have

$$\beta_j = \sum_{n:d^+_{mn}=1} \prod_{k:B_m(\gamma_{jk})=0}^{\eta_j(i_{mn})} \alpha_{\gamma_{jk}} (1-\alpha_{i_{mn}}) \tag{EC.30}$$

$$= \sum_{l:B_m(\gamma_{jl})=0} \prod_{k:B_m(\gamma_{jk})=0}^{l} \alpha_{\gamma_{jk}} (1-\alpha_{\gamma_{jl}}) \tag{EC.31}$$

$$= 1 - \alpha_{\gamma_{j1}} + \alpha_{\gamma_{j1}}(1-\alpha_{\gamma_{j2}}) + \alpha_{\gamma_{j1}}\alpha_{\gamma_{j2}}(1-\alpha_{\gamma_{j3}}) + \cdots \tag{EC.32}$$

$$= 1 - \prod_{i:B_m(i)=0} \alpha_i \tag{EC.33}$$

In going from (EC.30) to (EC.31), we change from summing over states n to summing over preferences l at node j. We expand the summation and product terms to obtain (EC.32). Cancelling terms leads to (EC.33). Also, because β_j does not depend on j, we have

$$r_m = \sum_{j=1}^{J} \lambda_j \Big(1 - \prod_{i:B_m(i)=0} \alpha_i\Big) = \lambda \Big(1 - \prod_{i:B_m(i)=0} \alpha_i\Big). \tag{EC.34}$$

In (EC.34), r_m is the product of the region-wide arrival rate λ and the probability that at least one unit is available. We obtain the probability $P(k)$ by the birth-and-death process, for $k=1,\ldots,N$,

$$P(k) = \frac{1}{M}\prod_{r=1}^{k}\frac{\lambda(r-1)}{\mu(r)} = \frac{1}{M}\prod_{r=1}^{k}\frac{\sum_{m\in\mathscr{S}_{r-1}}P(B_m|r-1)r_m}{r\mu} \tag{EC.35}$$

$$= \frac{1}{M}\prod_{r=1}^{k}\frac{(r-1)!(N-r+1)!}{r\mu N!}\sum_{m\in\mathscr{S}_{r-1}}\Big(1-\prod_{i:B_m(i)=0}\alpha_i\Big)\lambda \tag{EC.36}$$

where M is the normalizing factor. Also note that $P(0)=\frac{1}{M}$. $\square$

EC.1.6. Proof of Lemma 4

Proof: In a two-server pooled system, the service time is no longer exponential unless the service rates of the two services are the same. However, the blocking probability for the Erlang loss system depends only on the mean service time, not its distribution, as shown in Smith and Whitt (1981). In the pooled system in which all units are cross-trained joint units, the total arrival rate is $\lambda^1+\lambda^2$, and the mean service time is $\frac{\lambda^1+\lambda^2}{\lambda^1/\mu^1+\lambda^2/\mu^2}$. Thus, the offered load is a_1+a_2. We obtain the blocking probabilities by applying the Erlang B formula. The blocking probabilities for the two types of services are the same. $\square$

EC.1.7. Proof of Lemma 5

Proof: We first prove part i) where separate units are the primary server group. Calls that find all separate units busy are served by joint units, and calls that find all joint units busy are lost. The overflow processes of the two separate groups have intensities b_1 and b_2 respectively, where $b_y = a_y B(N_y, a_y)$ for $y\in\{1,2\}$ and $B(N_y,a_y)$ is the loss probability of each separate units group. This overflow in the joint units group is not Poisson as it depends on the number of busy servers in the primary group. Using a probability-generating function, Cooper (1981) shows that the variance of each overflow process is

$$v_y = b_y(1-b_y+\frac{a_y}{N_y+1-a_y+b_y}), \tag{EC.37}$$

which is known as the Riordan formula. The arrival process in the joint units group is the superposition of two overflows, each with intensity b_y and variance v_y, which is not a renewal process.

Under *Hayward's assumption*, we describe this overflow traffic by an equivalent Poisson traffic. As shown in Fredericks (1980), the peakedness of this process is $z = \frac{\sum_y v_y}{\sum_y b_y}$, and the blocking probability of this aggregated process is $B(\frac{N_c}{z}, \frac{b_1+b_2}{z})$, where $B(n,r)$ is the interpolated Erlang loss function, given by

$$B(n,r) = \left[r^{-n} e^r \int_r^\infty e^{-x} x^n dx \right]^{-1}. \tag{EC.38}$$

Thus, the blocking probability P_{b1}^y is given by

$$P_{b1}^y = \frac{b_y B(\frac{N_c}{z}, \frac{b_1+b_2}{z})}{a_y} = \frac{a_y B(N_y, a_y) B(\frac{N_c}{z}, \frac{b_1+b_2}{z})}{a_y} = B(N_y, a_y) B\Big(\frac{N_c}{z}, \frac{b_1+b_2}{z}\Big). \tag{EC.39}$$

We next prove ii) where joint units are the primary server group. The calls that enter the primary group are the two aggregated Poisson processes each with intensity a_y. From Lemma 3, the blocking probability of this process is $\frac{(a_1+a_2)^{N_c}}{\sum_{k=0}^{N_c}(a_1+a_2)^k}$, which equals $B(N_c, a_1+a_2)$. The overflow process, with a total intensity $(a_1+a_2)B(N_c, a_1+a_2)$, consists of two separate processes, one for each type of call, each with intensity $\tilde{b}_y = a_y B(N_c, a_1+a_2)$. Each overflow process flows into the separate units group with N_y units. The variance of each process is

$$\tilde{v}_y = \tilde{b}_y (1 - \sum_y \tilde{b}_y + \frac{\sum_y a_y}{N_c + 1 - \sum_y a_y + \sum_y \tilde{b}_y}), \tag{EC.40}$$

and the peakedness is $\tilde{z}_y = \frac{\tilde{v}_y}{\tilde{b}_y}$. Thus, the blocking probability of the separate units under Hayward's assumption is $B(\frac{N_y}{\tilde{z}_y}, \frac{\tilde{b}_y}{\tilde{z}_y})$, and the total blocking probability P_{b2}^y is given by

$$P_{b2}^y = \frac{\tilde{b}_y B(\frac{N_y}{\tilde{z}_y}, \frac{\tilde{b}_y}{\tilde{z}_y})}{a_y} = \frac{a_y B(N_c, a_1+a_2) B(\frac{N_y}{\tilde{z}_y}, \frac{\tilde{b}_y}{\tilde{z}_y})}{a_y} = B(N_c, a_1+a_2) B\Big(\frac{N_y}{\tilde{z}_y}, \frac{\tilde{b}_y}{\tilde{z}_y}\Big). \tag{EC.41}$$

□

In Figure EC.2 we show the values of P_{b1}^y and P_{b2}^y, $y \in \{1,2\}$, are close in six different settings, indicating that the approximation $P_b^y = \frac{1}{2}(P_{b1}^y + P_{b2}^y)$ is accurate.

EC.2. Figures and Tables in the Accuracy Experiments

In Section 4.1 of the paper, we summarized the results of experiments to test the accuracy of the approximation algorithms. Here we present tables and figures that show the detailed results. In the

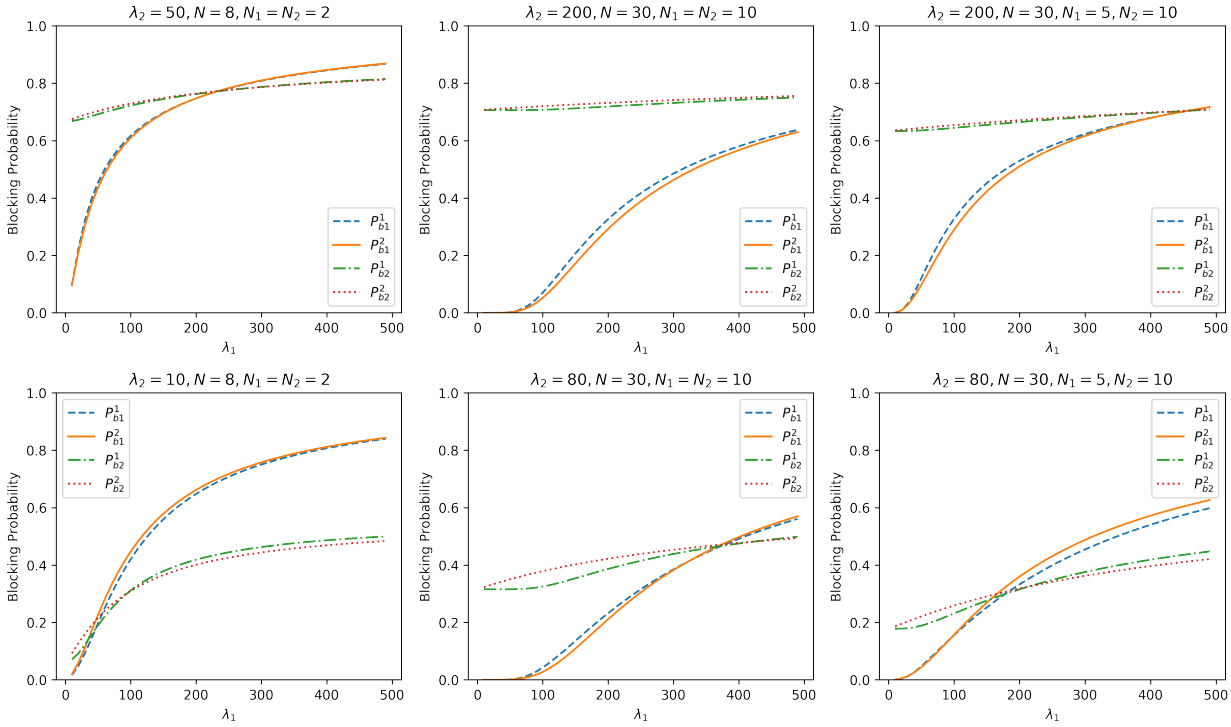


Figure EC.2 **Blocking Probabilities P^y_{b1} and P^y_{b2} for Six Different Settings with Fixed λ_2 and Varying λ_1. λ_1 and λ_2 are the arrival rates of type-1 and type-2 calls, respectively. N is the total number of units, and N_1 and N_2 are the numbers of type-1 and type-2 units, respectively.**

following tables and figures, we define the symbols as follows: N is the total number of units, N_1 is the number of separate type-1 units (emergency medical), N_2 is the number of separate type-2 units (fire), $t(s)$ is the run time in seconds, ϵ_ρ and ϵ_{MRT} are the absolute percentage errors of unit utilizations and mean response times, respectively. In the tables, the best performing algorithm are highlighted in bold.

Table EC.1 and Figure EC.3 show the results of experiments on small and medium sized systems. The exact model is feasible computationally and we compare it to the results of the LHDA, with and without normalization, the HDA, and a simulation with 200,000 runs.

Table EC.2 shows the results of experiments on large systems. In these experiments both the exact model and HDA were not feasible computationally. Hence, we compared the LHDA approximation to a simulation with 10 million arrivals as a proxy for the exact model base case result.

We also tested cases with both joint and separate units. The first experiments were for small systems. In these experiments, we fixed the total number of units $N = 8$ and the number of joint

units $N_c = 2$. This choice ensures the feasibility of running the exact hypercube model. We varied the number of type-1 units from 1 to 2 to 3, and varied the number of type-2 units from 5 to 4 to 3. The results are shown in Figure EC.4 and Table EC.3. We observe that LHDA does not always perform best in terms of accuracy. This is because the blocking probability in this case is an approximate value, which introduces an error in the normalization step. The average error of unit utilization is 0.662% (±0.385%) and the average error of mean response time is 0.431% (±0.304%), where the numbers in parentheses are standard deviations. The run times of LHDA are also consistently much lower than for the other methods.

We also experimented with large systems having joint and separate units. The results are shown in Table EC.4. There were 40 total units, 30 joint units, 4 emergency medical units, and 6 fire units. We compared LHDA to a simulation with 10 million runs as a proxy for the exact model.

Table EC.1 Algorithm Accuracy and Computation Times for Small and Medium Sized Systems

Settings		HDA			LHDA w/o Norm			LHDA			Simulation		
N	Offered Load	t(s)	ϵ_ρ(%)	ϵ_{MRT}(%)	t(s)	ϵ_ρ(%)	ϵ_{MRT}(%)	t(s)	ϵ_ρ(%)	ϵ_{MRT}(%)	t(s)	ϵ_ρ(%)	ϵ_{MRT}(%)
3	0.1	**0.01**	4.54	4.25	0.01	4.40	4.15	**0.01**	**0.21**	**0.04**	2.44	0.88	0.30
3	0.2	**0.01**	2.51	3.18	0.01	2.28	2.92	**0.01**	**0.08**	**0.02**	2.18	0.62	0.32
3	0.3	0.02	1.67	2.36	0.01	1.50	2.13	**0.01**	**0.04**	**0.01**	2.06	1.14	0.32
3	0.4	0.02	1.26	1.86	0.01	1.15	1.66	**0.01**	**0.05**	**0.01**	1.96	1.19	0.43
3	0.5	0.03	1.01	1.52	0.02	0.95	1.36	**0.02**	**0.07**	**0.00**	1.92	0.99	0.40
3	0.6	0.03	0.85	1.29	0.02	0.83	1.14	**0.02**	**0.09**	**0.00**	1.96	1.33	0.63
3	0.7	0.03	0.74	1.12	0.02	0.75	0.99	**0.02**	**0.10**	**0.00**	1.92	1.49	0.51
3	0.8	0.04	0.65	0.99	0.02	0.69	0.87	**0.02**	**0.13**	**0.00**	1.88	1.49	0.63
3	0.9	0.04	0.58	0.88	0.02	0.66	0.78	**0.02**	**0.15**	**0.00**	1.88	1.97	0.45
3	1.0	0.04	0.53	0.80	0.01	0.62	0.70	**0.03**	**0.17**	**0.01**	1.84	1.89	0.45
5	0.1	0.08	0.07	0.15	0.03	0.08	0.13	**0.02**	**0.05**	**0.02**	2.40	1.12	0.18
5	0.2	0.08	0.33	0.39	0.03	0.42	1.00	**0.02**	**0.12**	**0.04**	2.56	1.04	0.19
5	0.3	0.10	1.12	0.10	0.04	1.32	2.99	**0.03**	**0.17**	**0.05**	2.50	1.01	0.24
5	0.4	0.10	2.54	1.06	0.04	2.76	5.66	**0.03**	**0.20**	**0.04**	2.58	1.11	0.33
5	0.5	0.11	4.09	2.76	0.04	4.30	8.16	**0.03**	**0.20**	**0.04**	2.66	0.97	0.23
5	0.6	0.13	5.28	4.47	0.06	5.47	9.92	**0.04**	**0.18**	**0.03**	2.64	0.92	0.34
5	0.7	0.13	5.89	5.78	0.06	6.05	10.84	**0.04**	**0.16**	**0.03**	2.64	0.85	0.28
5	0.8	0.15	5.97	6.61	0.07	6.11	11.09	**0.05**	**0.14**	**0.02**	2.90	0.90	0.24
5	0.9	0.16	5.72	7.02	0.09	5.84	10.93	**0.06**	**0.12**	**0.02**	3.00	0.99	0.33
5	1.0	0.13	5.31	7.15	0.08	5.42	10.54	**0.05**	**0.10**	**0.02**	2.64	1.07	0.19
8	0.1	1.09	0.08	0.09	0.04	0.08	0.04	**0.03**	**0.06**	**0.03**	27.58	0.55	0.05
8	0.2	1.08	0.27	0.48	0.05	0.24	0.44	**0.04**	**0.17**	**0.08**	29.90	0.45	**0.08**
8	0.3	1.43	0.52	1.09	0.06	0.70	2.35	**0.05**	**0.24**	0.13	31.82	0.48	**0.08**
8	0.4	1.44	1.06	1.17	0.07	1.86	7.14	**0.06**	**0.27**	0.13	34.78	0.46	**0.07**
8	0.5	1.69	2.58	0.13	0.10	3.76	14.15	**0.07**	**0.26**	**0.11**	35.50	0.42	0.12
8	0.6	1.70	4.54	2.84	0.12	5.81	20.59	**0.08**	**0.24**	**0.09**	35.88	0.39	0.14
8	0.7	2.03	6.08	5.93	0.14	7.24	24.24	**0.09**	**0.21**	**0.06**	37.54	0.40	0.16
8	0.8	2.12	6.71	8.29	0.18	7.71	25.04	**0.11**	**0.18**	**0.05**	38.04	0.41	0.09
8	0.9	2.06	6.57	9.58	0.18	7.48	24.14	**0.11**	**0.15**	**0.04**	37.10	0.40	0.15
8	1.0	1.64	6.03	10.04	0.14	6.90	22.55	**0.10**	**0.13**	**0.03**	34.02	0.44	0.17

Table EC.2 Algorithm Accuracy for Large Systems ($N = 40$)

	Offered Load	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Simulation	t(s)	501.82	502.18	498.84	496.82	497.04	502.72	496.88	499.88	508.84	498.34
LHDA	t(s)	0.26	0.31	0.26	0.3	0.24	0.27	0.23	0.22	0.24	0.21
	ϵ_ρ(%)	0.38	0.41	0.41	0.36	0.42	0.36	0.31	0.4	0.54	0.42
	ϵ_{MRT}(%)	0.12	0.14	0.14	0.12	0.14	0.12	0.14	0.17	0.16	0.17

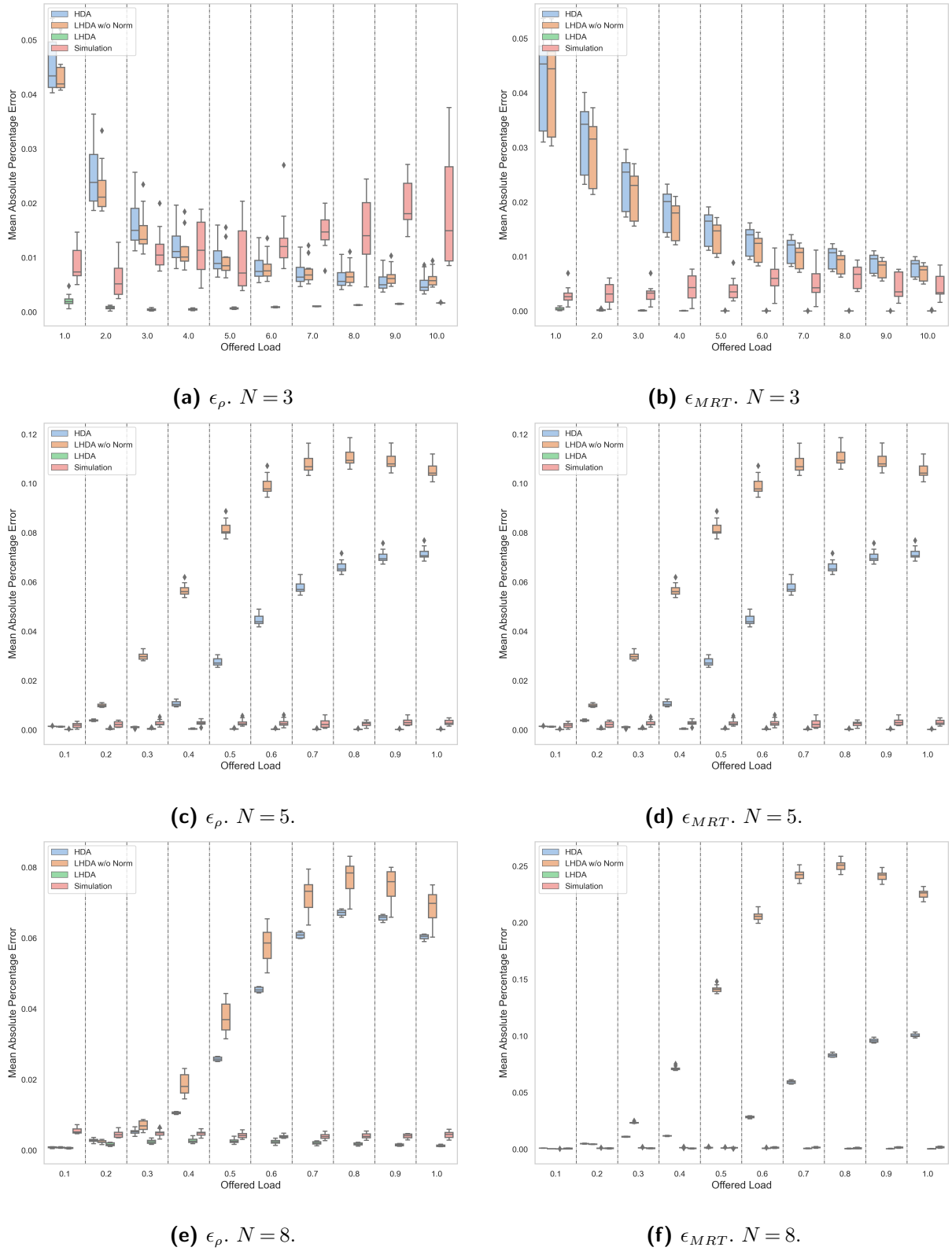
Table EC.3 Algorithm Accuracy and Computation Times for Medium-Sized Systems with Both Joint and Separate Units

Settings				HDA			LHDA			Simulation		
N	N_1	N_2	Offered Load	t(s)	ϵ_ρ(%)	ϵ_{MRT}(%)	t(s)	ϵ_ρ(%)	ϵ_{MRT}(%)	t(s)	ϵ_ρ(%)	ϵ_{MRT}(%)
8	1	5	0.1	0.17	**0.16**	0.06	**0.02**	0.17	**0.03**	2.76	1.13	0.16
8	1	5	0.2	0.22	0.44	0.30	**0.03**	**0.34**	**0.15**	2.84	1.07	0.17
8	1	5	0.3	0.23	**0.68**	0.56	**0.04**	0.72	0.30	2.92	1.14	**0.19**
8	1	5	0.4	0.22	**0.84**	0.73	**0.04**	1.00	0.44	3.06	1.01	**0.20**
8	1	5	0.5	0.22	0.96	0.88	**0.05**	1.15	0.55	3.12	**0.87**	**0.25**
8	1	5	0.6	0.23	1.09	1.13	**0.08**	1.25	0.64	3.12	**0.88**	**0.32**
8	1	5	0.7	0.22	1.21	1.34	**0.09**	1.32	0.73	3.14	**0.82**	**0.24**
8	1	5	0.8	0.22	1.29	1.49	**0.14**	1.39	0.80	3.18	**0.75**	**0.20**
8	1	5	0.9	0.23	1.33	1.59	**0.10**	1.45	0.87	3.22	**0.89**	**0.34**
8	1	5	1.0	0.24	1.36	1.64	**0.12**	1.50	0.91	3.30	**0.81**	**0.34**
8	2	4	0.1	0.09	0.13	0.06	**0.02**	**0.13**	**0.02**	2.76	1.11	0.19
8	2	4	0.2	0.12	0.37	0.14	**0.03**	**0.34**	**0.06**	2.94	1.03	0.19
8	2	4	0.3	0.12	0.66	0.36	**0.04**	**0.48**	**0.12**	3.22	0.86	0.18
8	2	4	0.4	0.12	0.93	0.55	**0.05**	**0.53**	**0.18**	3.26	1.05	0.31
8	2	4	0.5	0.12	1.21	0.98	**0.05**	**0.53**	**0.28**	3.38	0.84	0.34
8	2	4	0.6	0.12	1.45	1.39	**0.08**	**0.51**	0.41	3.48	0.77	**0.31**
8	2	4	0.7	0.12	1.58	1.69	**0.11**	**0.53**	0.54	3.54	0.65	**0.17**
8	2	4	0.8	0.12	1.60	1.89	**0.09**	**0.56**	0.66	3.54	0.69	**0.18**
8	2	4	0.9	0.12	1.55	1.98	**0.10**	**0.60**	0.76	3.40	0.77	**0.27**
8	2	4	1.0	0.11	1.46	2.01	**0.13**	**0.64**	0.83	3.32	0.72	**0.33**
8	3	3	0.1	0.05	0.13	0.07	**0.02**	**0.09**	**0.02**	2.38	1.02	0.22
8	3	3	0.2	0.08	0.38	0.16	**0.03**	**0.24**	**0.06**	2.60	1.02	0.20
8	3	3	0.3	0.08	0.64	0.06	**0.04**	**0.45**	**0.07**	2.86	0.85	0.25
8	3	3	0.4	0.08	0.97	0.44	**0.04**	**0.62**	**0.08**	2.96	0.85	0.27
8	3	3	0.5	0.08	1.41	0.98	**0.05**	**0.65**	**0.20**	3.06	0.91	0.35
8	3	3	0.6	0.08	1.70	1.46	**0.07**	**0.58**	0.37	3.26	0.67	**0.28**
8	3	3	0.7	0.08	1.82	1.81	**0.07**	**0.53**	0.53	3.32	0.65	**0.26**
8	3	3	0.8	0.08	1.80	2.02	**0.07**	**0.51**	0.67	3.20	0.66	**0.19**
8	3	3	0.9	0.08	1.72	2.12	**0.07**	**0.51**	0.78	3.26	0.69	**0.31**
8	3	3	1.0	0.08	1.59	2.14	**0.07**	**0.53**	0.86	3.36	0.74	**0.34**

Table EC.4 Algorithm Accuracy for Large Systems with Both Joint and Separate Units

($N = 40, N_c = 30, N_1 = 4, N_2 = 6$)

	Offered Load	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Simulation	t(s)	461.10	461.00	455.28	460.08	457.18	459.16	458.30	460.44	458.48	456.82
LHDA	t(s)	0.37	0.37	0.35	0.36	0.38	0.36	0.35	0.34	0.38	0.37
	ϵ_ρ(%)	0.87	0.94	0.94	0.92	1.04	0.86	0.86	0.74	1.12	0.91
	ϵ_{MRT}(%)	0.68	0.61	0.71	0.65	0.64	0.50	0.62	0.56	0.59	0.65



(a) ϵ_ρ. $N = 3$

(b) ϵ_{MRT}. $N = 3$

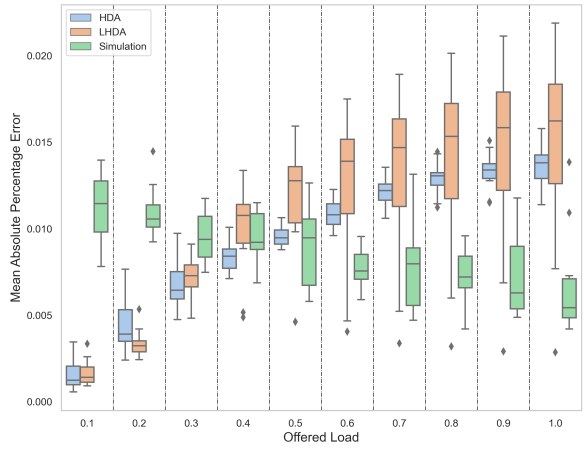
(c) ϵ_ρ. $N = 5$.

(d) ϵ_{MRT}. $N = 5$.

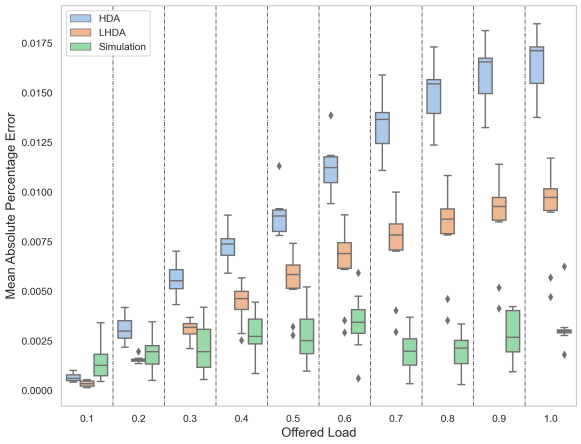
(e) ϵ_ρ. $N = 8$.

(f) ϵ_{MRT}. $N = 8$.

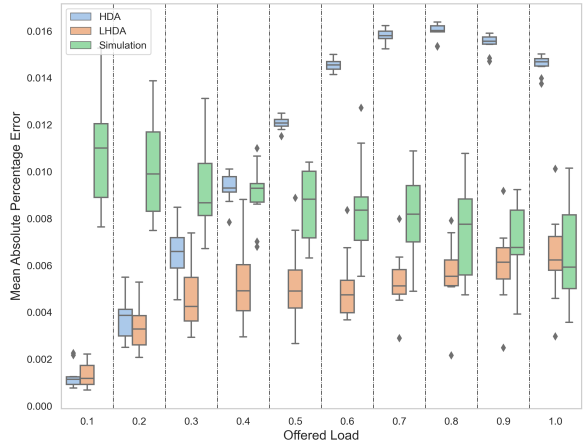
Figure EC.3 **(a) - (c) Show the Errors of Unit Utilizations for the Approximation Methods Compared to the Exact Three-State Hypercube Model. (d) - (f) Show the Corresponding Errors of the Mean Response Times. Within each tall rectangle in (a) – (f), the box plots are for HDA on the left, LHDA without the normalization step on the middle left, LHDA on the middle right, and the simulation on the right.**



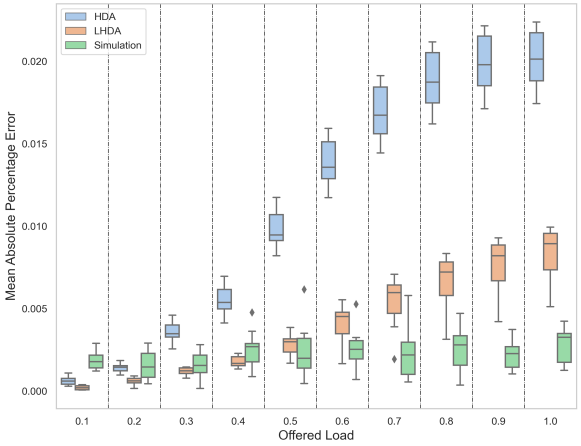
(a) ϵ_ρ. $N_1 = 1, N_2 = 5$



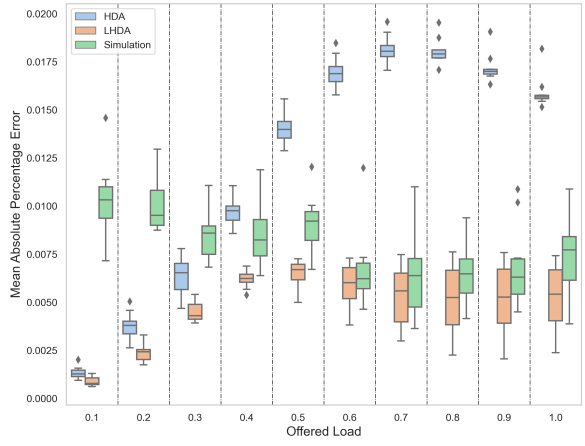
(b) ϵ_{MRT}. $N_1 = 1, N_2 = 5$



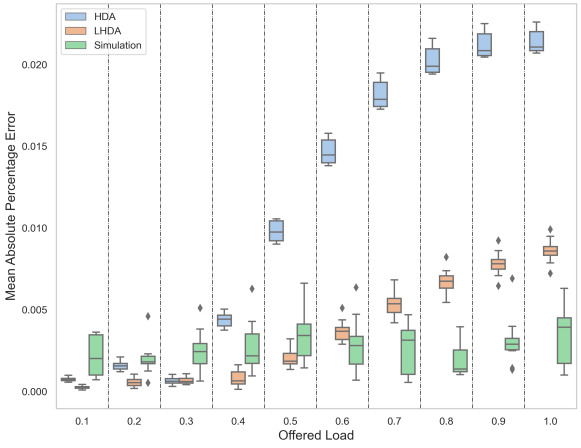
(c) ϵ_ρ. $N_1 = 2, N_2 = 4$



(d) ϵ_{MRT}. $N_1 = 2, N_2 = 4$



(e) ϵ_ρ. $N_1 = 3, N_2 = 3$



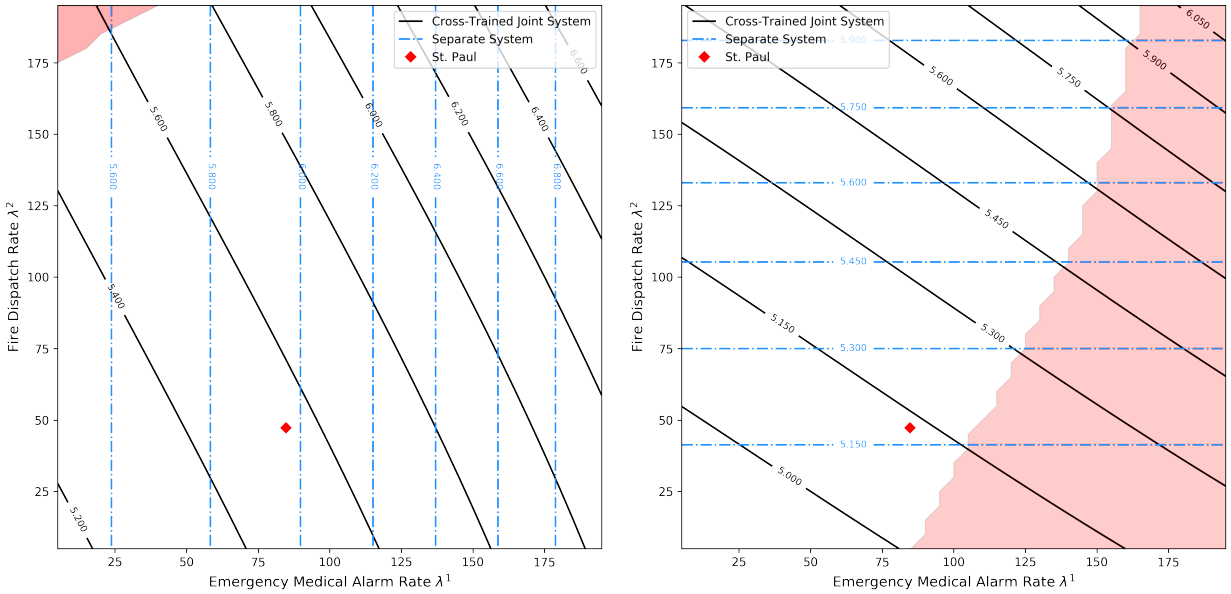
(f) ϵ_{MRT}. $N_1 = 3, N_2 = 3$

Figure EC.4 **(a) - (c) Show the Errors of Unit Utilizations for the Approximation Methods Compared to the Exact Three-State Hypercube Model. (d) - (f) Show the Corresponding Errors of the Mean Response Times. Within each tall rectangle in (a) – (f), the box plots are for HDA on the left, LHDA without normalization step on the middle left, LHDA on the middle right, and the simulation on the right.**

EC.2.1. Sensitivity Analysis: Varying the Fire and Medical Alarm Rates

In Section 5.6 of the text we summarize the results of this sensitivity analysis that varies St. Paul alarm rates. Here we provide the details of that analysis. In Figure EC.5 we compare the mean response times under the St. Paul joint fire-medic system with nine joint units to the traditional system with six separate engines and six separate emergency medical units that are optimally located, with both systems having the same number of positions. In both systems we keep the three super units and two separate engine units. The red diamonds in Figures EC.5a and EC.5b designate the fire and emergency medical call rates that we used in the St. Paul analysis. In both figures we vary these call rates from a low of 5 per day to a high of 200 per day with an increment of 5 per day. Figure EC.5a shows the mean response time to emergency medical calls and Figure EC.5b shows the mean response time to fire calls. The solid lines are the contour plots of equal mean response times for the joint system while the dash-dotted lines are the contour plots of equal mean response times for the separate system. In Figure EC.5a the dash-dotted lines are vertical because under the separate system, the response times to emergency medical calls are only affected by the medical call rates. In Figure EC.5b the dash-dotted lines are horizontal because under the separate system the response times to fire calls are only affected by the fire call rates. For completeness, we examined a very wide range of call rates, but recognize that in practical terms the area surrounding the red diamond in each of the two figures has the greatest practical significance.

Examining Figure EC.5a in detail, we observe that the fire-medic system has a lower mean response time to emergency medical calls for all combinations of fire and medical emergency call rates except in the very small, shaded area in the upper left corner of Figure EC.5a. In that area, the emergency medical call rate is less than about 30 calls per day, much less than the St. Paul rate in our study of 84.7 calls per day, and the fire incident call rate is more than 175 calls per day which is about three and a half times the actual call rate in our study. With this extremely high fire alarm rate, the number of available fire-medic units is greatly reduced, and the separate system performs better than the joint system. This is a theoretical result since actual fire rates



(a) EMS Response Time **(b)** Fire Response Time

Figure EC.5 Contour Plots of the Mean Response Times for both Joint System (Solid Lines) and Separate System (Dash-dotted Lines). Shaded Areas Show System Settings where Separate System Performs Better.

are declining and emergency medical rates are increasing as we described in the Introduction. The important conclusion from Figure EC.5a is that the fire-medic system is preferred for virtually all combinations of fire and emergency medical call rates.

Figure EC.5b shows that with the current fire call rate, increasing the emergency medical call rate beyond about 110 calls per day would result in the separate system having lower mean response times to fire calls. But from Figure EC.5a observe that increasing the emergency medical call rate would increase the mean response time to emergency medical calls and likely cause a self-correcting mechanism to come into play, leading the city to add units to reduce response times and the workload of units and thereby maintain the advantage of the fire-medic system.

Observe also from Figure EC.5b that the boundary of the shaded region has a positive slope. This occurs because for a fixed emergency medical call rate, an increase in the fire rate has a greater negative effect on the separate system because fire calls are spread over fewer units than in the fire-medic system.

References

Burden R, Faires J (2011) *Numerical Analysis 9th'ed* (Boston: Brooks/Cole).

Cooper R (1981) *Introduction to Queueing Theory, 2nd'ed* (North-Holland, Amsterdam).

Fredericks A (1980) Congestion in blocking systems—a simple approximation technique. *Bell System Technical Journal* 59(6):805–827.

Jarvis JP (1985) Approximating the equilibrium behavior of multi-server loss systems. *Management Science* 31(2):235–239.

Smith DR, Whitt W (1981) Resource sharing for efficiency in traffic systems. *Bell System Technical Journal* 60(1):39–55.