

Online Appendix to “Location-Allocation of Emergency Service Systems in Light Traffic”

In Appendix A, we provide proof for Theorem 1. In Appendix B, we provide proofs for Theorem 2 and present Corollary 1 together with its proof. In Appendix C, we discuss the design of ESSs with collaborative servers by developing a hypercube queueing model. In Appendix D, we prove Proposition 1. In Appendix E, we provide supplementary materials and proofs to §3 of the main paper. We present supplementary materials to §4 and §5 in Appendix F and Appendix G, respectively.

In the following proofs, we let $\mathbb{N}_0$ and $\mathbb{N}$ denote the set of non-negative integers and positive integers, respectively. We use $\xrightarrow{d}$ and $\xrightarrow{p}$ to represent convergence in distribution and in probability, respectively. For a known positive integer M , we use $[M] := \{1, 2, \dots, M\}$ to denote the set of running indices up to M . For two random variables X and Y , we let $X \leq_{s.t.} Y$ denote that X is less than Y in the usual stochastic order, i.e., $\Pr(X > x) \leq \Pr(Y > x)$ for all x , and we write $X \stackrel{d}{=} Y$ to mean that X equals to Y in distribution. For a function $f: \mathbb{R} \mapsto \mathbb{R}$ that is n -times differentiable, we let $f^{(n)}$ be its n th order derivative, and let $f^{(0)} = f$ denote the function itself. We use “cdf” and “pdf” as the shorthands for “cumulative distribution function” and “probability density function”, respectively.

Appendix A: Proof of Theorem 1

In this single-server spatial queue, recall that call arrivals from demand node $i \in [I] := \{1, 2, \dots, I\}$ follow a Poisson process with rate $\pi_i \lambda$, where λ is the total arrival rate, and $\pi_i \geq 0$ is the fraction of arrival from node i satisfying $\sum_{i \in [I]} \pi_i = 1$. Recall that the travel time and on-scene service time for node i are denoted by $S_{tr,i}$ and $S_{os,i}$, respectively, and the mission duration time S_i satisfies $S_i \stackrel{d}{=} S_{tr,i} + S_{os,i}$, which has a cdf G_i . Let g_i be the corresponding density function. Because a demand comes from node i with probability π_i , the distribution of mission duration $\tilde{G}$ for this random demand is a mixture of G_i , i.e., $\tilde{G} := \sum_{i \in [I]} \pi_i G_i$. Let $\tilde{g} := \sum_{i \in [I]} \pi_i g_i$ be the density of $\tilde{G}$, $\tilde{S}$ denote a generic random variable following distribution $\tilde{G}$ with a finite first moment $\mathbb{E}\tilde{S} := \sum_{i \in [I]} \pi_i \mathbb{E}S_i$, and $\tilde{\mathbb{S}}(s) := \mathbb{E}[e^{s\tilde{S}}]$ be the moment generating function of $\tilde{S}$. Furthermore, let $\bar{S}$ denote the random variable $\max_{i \in [I]} \{S_i\}$. It is clear that $\tilde{S} \leq_{s.t.} \bar{S}$. Similarly, let $\tilde{S}_{tr}$ denote the random travel time to respond to a call, which follows a mixture distribution. That is, with probability π_i , $\tilde{S}_{tr} = S_{tr,i}$.

For the n th call that arrives to the system, where $n \in \mathbb{N}$, let K_n denote the number of calls waiting in queue upon the call’s arrival, B_n denote the busy indicator where $B_n = 1$ if the call arrives to find the system busy; and $B_n = 0$ otherwise, and L_n denote the residual time to complete the on-hand mission upon the call’s arrival if $B_n = 1$; and $L_n = 0$ if $B_n = 0$. Finally, let R_n denote the response time of the n th call. Under a non-idling and FCFS rule, R_n is a sum of the residual mission duration L_n , the time to deplete the K_n calls already in queue, and the travel time to reach the call itself (distributed as $\tilde{S}_{tr}$). Thus, we have the relation $R_n \stackrel{d}{=} L_n + \tilde{S}_{tr}$ if $K_n = 0$ and $R_n \leq_{s.t.} L_n + k\bar{S} + \tilde{S}_{tr}$ if $K_n = k$ for all $k \in \mathbb{N}$. The process $\{(K_n, B_n, L_n, R_n) : n \in \mathbb{N}\}$ is a regenerative process, with the epoch on which $B_n = 0$, i.e., a call finds the system idle, as a regeneration epoch. By Theorem 2.2.4 of Tijms (2003), the limit of the sequence of probabilities $P[K_n \leq k, B_n \leq b, L_n \leq l, R_n \leq r]$ exists for any fixed k, b, l, r . The B_n sequence is bounded and thus is tight. The sequences of K_n, L_n and R_n are also tight as argued in Iglehart (1973). The marginal tightness is equivalent to joint tightness for a sequence of random vectors. These facts together imply that the limit

is a probability measure, and thus there exists a vector of limiting random variables $(K_\infty, B_\infty, L_\infty, R_\infty)$ such that $(K_n, B_n, L_n, R_n) \xrightarrow{d} (K_\infty, B_\infty, L_\infty, R_\infty)$ as $n \rightarrow \infty$. For each $k \in \mathbb{N}_0$, let $p_k := P[K_\infty + B_\infty = k]$ denote the limiting probability that there are k calls in the system, $p_{\mathcal{B}} := \sum_{k \in \mathbb{N}} p_k = 1 - p_0$ denote the limiting probability that the server is busy, and $q_k := p_{k+1}/p_{\mathcal{B}}$ be the limiting conditional probability that there are $k \in \mathbb{N}_0$ calls waiting in queue, given that the server is busy. It holds that $\sum_{k \in \mathbb{N}_0} q_k = 1$. For each $k \in \mathbb{N}_0$, we define two non-negative random variables $L(k)$ and $R(k)$ by $L(k) = [L_\infty | B_\infty = 1, K_\infty = k]$ and $R(k) = [R_\infty | B_\infty = 1, K_\infty = k]$. Here, $L(k)$ can be understood as the limiting variable for the residual mission duration at the moment that a new call arrives, conditional on the event that the new call finds the server busy with $k \in \mathbb{N}_0$ calls waiting in queue. Similarly, $R(k)$ can be interpreted as the conditional limit distribution for the response time of a call, conditional on the event that the call arrives to find the server busy with $k \in \mathbb{N}_0$ calls waiting in queue. From the relation between $R_n(k)$ and $L_n(k)$, the limiting random variables $R(k)$ and $L(k)$ satisfy the recursion

$$\begin{aligned} R(0) &\stackrel{d}{=} L(0) + \tilde{S}_{\text{tr}} \\ R(k) &\leq_{s.t.} L(k) + k\tilde{S} + \tilde{S}_{\text{tr}}, \quad k \in \mathbb{N}. \end{aligned} \tag{9}$$

Our interest centers on the loss of those calls that find the system busy. Over the first n calls, these are $\sum_{l=1}^n B_l$ such calls, and their empirical average cost is

$$\frac{\sum_{l=1}^n h(R_l) B_l}{\sum_{l=1}^n B_l} = \frac{n}{\sum_{l=1}^n B_l} \times \frac{1}{n} \sum_{l=1}^n h(R_l) B_l.$$

The first term on the right hand side can be readily seen to converge to $1/P[B_\infty = 1]$. According to the renewal reward theorem, the second term on the right hand side and its expectation converge to the same constant as $n \rightarrow \infty$. A routine calculation shows that the limit is $\mathbb{E}[h(R_\infty)B_\infty]$ (to see this, we can use the standard argument that first considers a simple function for h , then a non-negative function, and finally a general integrable function). By writing $h(R_\infty)B_\infty = \sum_{k \in \mathbb{N}_0} h(R_\infty) \cdot B_\infty \cdot \mathbb{I}(K_\infty = k)$, this limit is equal to $\mathbb{E} \sum_{k \in \mathbb{N}_0} h(R_\infty) B_\infty \mathbb{I}(K_\infty = k)$, which is further equal to $\sum_{k \in \mathbb{N}_0} \mathbb{E}[h(R_\infty) B_\infty \mathbb{I}(K_\infty = k)]$ by Fubini's theorem because $\mathbb{E}|h(R_\infty) B_\infty| \leq \mathbb{E}|h(R_\infty)|$ which is assumed finite by our Assumption 1 on h . This means that the long-run average loss of a call that finds the system busy equals

$$\frac{\sum_{k \in \mathbb{N}_0} \mathbb{E}[h(R_\infty) B_\infty \mathbb{I}(K_\infty = k)]}{P(B_\infty = 1)} = \frac{\sum_{k \in \mathbb{N}_0} \mathbb{E}[h(R_\infty) | B_\infty = 1, K_\infty = k] P(B_\infty = 1, K_\infty = k)}{P(B_\infty = 1)}.$$

Conditional on $B_\infty = 1$ and $K_\infty = k$, R_∞ has the same distribution as $R(k)$. In addition, by our definition of the limiting probabilities, it holds that $P(B_\infty = 1) = p_{\mathcal{B}}$ and $P(B_\infty = 1, K_\infty = k) = p_{k+1}$ for $k \in \mathbb{N}_0$. Therefore, this long-run average loss can be rewritten as

$$\lim_{n \rightarrow \infty} \frac{\sum_{l=1}^n h(R_l) B_l}{\sum_{l=1}^n B_l} = \frac{\sum_{k \in \mathbb{N}_0} p_{k+1} \mathbb{E} h(R(k))}{p_{\mathcal{B}}} = \mathbb{E} [h(L(0) + \tilde{S}_{\text{tr}})] q_0 + \sum_{k \geq 1} \mathbb{E} [h(R(k))] q_k, \tag{10}$$

where the second identity follows from (9) and the definition of q_k . This expression will serve as the basis for our asymptotic analysis of the long-run average loss of a call that finds the system busy in light traffic. We divide the proof into three steps. As the overall arrival rate $\lambda \downarrow 0$, the first step shows $q_0(\lambda) = 1 - \lambda \frac{\mathbb{E}\tilde{S}^2}{2\mathbb{E}\tilde{S}} + O(\lambda^2)$, the second shows $\mathbb{E} [h(L(0) + \tilde{S}_{\text{tr}})] - \mathbb{E} [h(L + \tilde{S}_{\text{tr}})] = O(\lambda)$, while the third step shows $\sum_{k \geq 1} \mathbb{E} [h(R(k))] q_k$ is $O(\lambda)$. Combining these pieces yields the conclusion of the theorem.

Step 1: Show $q_0(\lambda) = 1 - \lambda \frac{\mathbb{E}\tilde{S}^2}{2\mathbb{E}\tilde{S}} + O(\lambda^2)$ as the overall arrival rate $\lambda \downarrow 0$. Recall $q_0(\lambda)$ is the conditional probability of having zero call waiting in queue, given that the server is busy. To show its convergence, let D_n be the number of

calls in the system right after the completion of the n th call. It is readily checked that $D_n = [D_{n-1} - 1]_+ + A_n$, where A_n is the random number of arrivals during the service of the n th demand with $P(A_n = j) = a_j := \frac{1}{j!} \int_0^\infty e^{-\lambda x} (\lambda x)^j d\tilde{G}(x)$ for $j \in \mathbb{N}_0$. Because $\{A_n\}$ is an iid sequence of random variables and A_n is independent of $\{D_l : l \leq n-1\}$, the process $\{D_n, n \in \mathbb{N}\}$ evolves as a time-homogeneous Markov chain, and the limiting probabilities of this process, i.e., $d_k = \lim_{n \rightarrow \infty} P(D_n = k)$, satisfy the flow-balance equation $d_k = d_0 a_k + \sum_{l=1}^k d_l a_{k+1-l} + d_{k+1} a_0$ for each $k \in \mathbb{N}_0$ (e.g., [Adan and Haviv \(1978\)](#), Eq.(4)), where the right-hand side gives the limiting inflow rate into state k . To solve this system of equations, we introduce generating functions $\mathbb{D}(s) := \sum_{k \in \mathbb{N}_0} d_k s^k$ and $\mathbb{A}(s) := \sum_{k \in \mathbb{N}_0} a_k s^k$ for $s \geq 0$. Multiply both sides of the flow-balance equation by s^k and sum over each $k \in \mathbb{N}_0$ to see $\mathbb{D}(s) = d_0 \mathbb{A}(s) + (\mathbb{D}(s) - d_0) \mathbb{A}(s)/s$. Solve this equation for $\mathbb{D}(s)$ to yield $\mathbb{D}(s) = (s-1)d_0 \mathbb{A}(s)/\{s - \mathbb{A}(s)\}$. To compute d_0 , we let $s \uparrow 1$ on both sides of this equality and use the fact that $\mathbb{A}(1) = \mathbb{D}(1) = 1$ to obtain $1 = d_0 \times \lim_{s \uparrow 1} \frac{s-1}{s - \mathbb{A}(s)}$, where the limit equals $(1 - \mathbb{A}'(1))^{-1}$ by an appeal to L'Hospital's rule. This yields $d_0 = 1 - \lambda \mathbb{E}\tilde{S}$ because $\mathbb{A}'(1) = \mathbb{E}A_1$ and this expectation equals $\lambda \mathbb{E}\tilde{S}$ by direct calculation. To compute d_1 , first note that $a_0 = \int_0^\infty e^{-\lambda x} d\tilde{G}(x)$, which equals $\tilde{\mathbb{S}}(-\lambda)$, the moment generating function of $\tilde{S}$ evaluated as $-\lambda$. With this fact, we take $k=0$ in the flow-balance equation to see $d_0 = d_0 a_0 + d_1 a_0$, and solve it for d_1 to obtain $d_1 = d_0 \{1 - \tilde{\mathbb{S}}(-\lambda)\} / \tilde{\mathbb{S}}(-\lambda)$. By the up and down crossing argument, d_k is equal to long-run proportion of calls that arrive to find k calls (including the one in service, if any) in the system, that is, $d_k = p_k$ for all $k \in \mathbb{N}_0$. Therefore, $q_0(\lambda)$, defined as the ratio between p_1 and $p_{\mathcal{B}}$ when the overall arrival rate of demands is λ , can be computed as

$$q_0(\lambda) = \frac{p_1}{p_{\mathcal{B}}} = \frac{p_1}{1 - p_0} = \frac{d_1}{1 - d_0} = \frac{(1 - \lambda \mathbb{E}\tilde{S})\{1 - \tilde{\mathbb{S}}(-\lambda)\}}{\lambda \mathbb{E}\tilde{S} \tilde{\mathbb{S}}(-\lambda)},$$

Let $m_1 := \mathbb{E}\tilde{S}$, $m_2 := \mathbb{E}\tilde{S}^2$. By the mgf assumption, all moments exist and a Taylor expansion of $\tilde{\mathbb{S}}(-\lambda) = \mathbb{E}e^{-\lambda\tilde{S}}$ at 0 gives $\tilde{\mathbb{S}}(-\lambda) = 1 - \lambda m_1 + \frac{\lambda^2}{2} m_2 + O(\lambda^3)$, and thus $1 - \tilde{\mathbb{S}}(-\lambda) = \lambda m_1 - \frac{\lambda^2}{2} m_2 + O(\lambda^3)$. Write the numerator and denominator of $q_0(\lambda)$ as $\mathfrak{N}(\lambda) := (1 - \lambda m_1)\{1 - \tilde{\mathbb{S}}(-\lambda)\}$ and $\mathfrak{D}(\lambda) := \lambda m_1 \tilde{\mathbb{S}}(-\lambda)$. Using the above expansion of $1 - \tilde{\mathbb{S}}(-\lambda)$, the numerator can be expanded as

$$\mathfrak{N}(\lambda) = (1 - \lambda m_1) \left(\lambda m_1 - \frac{\lambda^2}{2} m_2 + O(\lambda^3) \right) = \lambda m_1 - \lambda^2 \left(m_1^2 + \frac{m_2}{2} \right) + O(\lambda^3).$$

Factoring out the common term λm_1 , the above can be rewritten as

$$\mathfrak{N}(\lambda) = \lambda m_1 \left\{ 1 - \lambda \left(m_1 + \frac{m_2}{2m_1} \right) + O(\lambda^2) \right\}.$$

Similarly, we can use the expansion of $\tilde{\mathbb{S}}(-\lambda)$ to expand the denominator as

$$\mathfrak{D}(\lambda) = \lambda m_1 \left(1 - \lambda m_1 + \frac{\lambda^2}{2} m_2 + O(\lambda^3) \right) = \lambda m_1 \left\{ 1 - \lambda m_1 + O(\lambda^2) \right\}.$$

Take the ratio to see

$$q_0(\lambda) = \frac{\mathfrak{N}(\lambda)}{\mathfrak{D}(\lambda)} = \frac{1 - \lambda \left(m_1 + \frac{m_2}{2m_1} \right) + O(\lambda^2)}{1 - \lambda m_1 + O(\lambda^2)}.$$

It is known that for small λ , $\{1 - \lambda m_1 + O(\lambda^2)\}^{-1} = 1 + \lambda m_1 + O(\lambda^2)$. Use this result to expand the above ratio as

$$q_0(\lambda) = \left\{ 1 - \lambda \left(m_1 + \frac{m_2}{2m_1} \right) + O(\lambda^2) \right\} \left\{ 1 + \lambda m_1 + O(\lambda^2) \right\} = 1 - \lambda \frac{m_2}{2m_1} + O(\lambda^2).$$

Recalling $m_1 = \mathbb{E}\tilde{S}$ and $m_2 = \mathbb{E}\tilde{S}^2$, we obtain $q_0(\lambda) = 1 - \lambda \frac{\mathbb{E}\tilde{S}^2}{2\mathbb{E}\tilde{S}} + O(\lambda^2)$.

Step 2: Show $\mathbb{E}[h(L^\lambda(0) + \tilde{S}_{tr})] - \mathbb{E}[h(L + \tilde{S}_{tr})] = O(\lambda)$ **as the arrival rate** $\lambda \downarrow 0$. Here, we will add a superscript ‘ λ ’ to a random variable that depends on the arrival rate λ to explicitly reveal the dependency. By Theorem 2.2.1 of [Kerner \(2008\)](#) (specifically, setting $\lambda_1 = \lambda$ in equation (3) therein), the pdf of the conditional residual service time $L^\lambda(0)$ is given by

$$f_{L^\lambda(0)}(x) = \frac{\lambda}{1 - \tilde{\mathbb{S}}(-\lambda)} e^{\lambda x} \int_x^\infty e^{-\lambda u} \tilde{g}(u) du, \quad x \geq 0, \quad (11)$$

where $\tilde{g} := \sum_{i \in [I]} \pi_i g_i$ is the density of the service time $\tilde{S}$ and

$$I(\lambda) = \frac{\lambda}{1 - \tilde{\mathbb{S}}(-\lambda)}, \quad II(\lambda, x) = \int_x^\infty e^{\lambda(x-u)} \tilde{g}(u) du.$$

For completeness, let $I(0) := \lim_{\lambda \downarrow 0} I(\lambda) = 1/\mathbb{E}[\tilde{S}]$, and $II(0, x) := \lim_{\lambda \downarrow 0} II(\lambda, x) = \int_x^\infty \tilde{g}(u) du$. Then it is immediate that $f_L(x)$ as defined in the theorem equals $I(0) \times II(0, x)$. Based on our assumption on $\tilde{S}$, we have $\tilde{\mathbb{S}}(-\lambda) \rightarrow 1$ and $\tilde{S}'(-\lambda) \rightarrow \mathbb{E}[\tilde{S}]$ as $\lambda \downarrow 0$. Hence for $0 < \lambda \leq \lambda_0$ small enough, $1 - \tilde{\mathbb{S}}(-\lambda)$ stays bounded away from 0, and so both $|I(\lambda)|$ and

$$I'(\lambda) = \frac{1}{1 - \tilde{\mathbb{S}}(-\lambda)} + \frac{\lambda \tilde{\mathbb{S}}'(-\lambda)}{(1 - \tilde{\mathbb{S}}(-\lambda))^2}$$

are both bounded for all $0 < \lambda \leq \lambda_0$. We can manually compute the right derivative $I'(0)$ and show that $I'(\lambda)$ converges to this right derivative as $\lambda \downarrow 0$. Next, we bound $II(\lambda, x)$ and its derivative. It is obvious that $II(\lambda, x) \leq II(0, x)$. For fixed $x \geq 0$ and λ , the integrand $e^{\lambda(x-u)} \tilde{g}(u)$ is differentiable in λ with derivative $(x-u)e^{\lambda(x-u)} \tilde{g}(u)$, which is upper bounded by $(u-x)_+ \tilde{g}(u)$, and thus by $u \tilde{g}(u)$ when $u \geq x$ and $\lambda \in [0, \lambda_0]$. Since $\mathbb{E}[\tilde{S}] < \infty$, we have $\int_0^\infty u \tilde{g}(u) du = \mathbb{E}[\tilde{S}] < \infty$. by dominated convergence,

$$\partial_\lambda II(\lambda, x) = \int_x^\infty (x-u) e^{\lambda(x-u)} \tilde{g}(u) du,$$

where the derivative at $\lambda = 0$ should be interpreted as the right derivative. In addition, $|\partial_\lambda II(\lambda, x)| \leq \int_x^\infty (u-x)_+ \tilde{g}(u) du$. Integrating in x and using Fubini's theorem,

$$\int_0^\infty |\partial_\lambda II(\lambda, x)| dx \leq \int_0^\infty \int_x^\infty (u-x) \tilde{g}(u) du dx = \int_0^\infty \tilde{g}(u) \int_0^u (u-x) dx du.$$

The inner integral equals $u^2/2$, and thus the double integral on the right hand side equals $\mathbb{E}[\tilde{S}^2]/2$, which is finite. Thus $\sup_{0 \leq \lambda \leq \lambda_0} \int_0^\infty |\partial_\lambda II(\lambda, x)| dx < \infty$. Based on the above set-up, we show $f_{L^\lambda(0)}$ is differentiable in L^1 . By the product rule,

$$\partial_\lambda f_{L^\lambda(0)}(x) = I'(\lambda) II(\lambda, x) + I(\lambda) \partial_\lambda II(\lambda, x).$$

Integrating and using the bounds above,

$$\int_0^\infty |\partial_\lambda f_{L^\lambda(0)}(x)| dx \leq |I'(\lambda)| \int_0^\infty |II(\lambda, x)| dx + |I(\lambda)| \int_0^\infty |\partial_\lambda II(\lambda, x)| dx.$$

A direct calculation shows $\int_0^\infty II(0, x) dx = \mathbb{E}\tilde{S}$, and

$$\begin{aligned} \int_0^\infty II(\lambda, x) dx &= \int_0^\infty \left(\int_x^\infty e^{\lambda(x-u)} \tilde{g}(u) du \right) dx = \int_0^\infty \tilde{g}(u) \left(\int_0^u e^{\lambda(x-u)} dx \right) du \\ &= \int_0^\infty \tilde{g}(u) \frac{1 - e^{-\lambda u}}{\lambda} du = \frac{1 - \tilde{\mathbb{S}}(-\lambda)}{\lambda}, \end{aligned}$$

which converges to $\int_0^\infty II(0, x) dx$ as $\lambda \downarrow 0$. Therefore, $\int_0^\infty |II(\lambda, x)| dx$ is finite and uniformly bounded for $0 \leq \lambda \leq \lambda_0$. Combining the bounds on $I'(\lambda)$, $II(\lambda, x)$, and $\partial_\lambda II(\lambda, x)$, we obtain

$$\sup_{0 \leq \lambda \leq \lambda_0} \int_0^\infty |\partial_\lambda f_{L^\lambda(0)}(x)| dx < \infty,$$

where the derivative at $\lambda = 0$ should be interpreted as the right derivative. For notational convenience, denote this right derivative at $\lambda = 0$ as $r(x)$. Let $Y = \tilde{S}_{\text{tr}}$, independent of $L^\lambda(0)$. Define

$$\varphi(\lambda) := \mathbb{E}[h(L^\lambda(0) + Y)] = \iint h(x + y) f_{L^\lambda(0)}(x) f_Y(y) dx dy, \quad 0 \leq \lambda \leq \lambda_0.$$

For each fixed (x, y) , the map $\lambda \mapsto h(x + y) f_{L^\lambda(0)}(x) f_Y(y)$ is differentiable with derivative $h(x + y) \partial_\lambda f_{L^\lambda(0)}(x) f_Y(y)$. Our previous analysis on $I(\lambda)$ and $II(\lambda, x)$ has shown that $\partial_\lambda f_{L^\lambda(0)}(x)$ is upper bounded by $II(0, x) + \int_x^\infty (u - x) \tilde{g}(u) du$ up to a multiplicative constant, which is an integrable function of x . Together, the integrand in ψ is upper bounded by $C \bar{h}(x + y) (II(0, x) + \int_x^\infty (u - x) \tilde{g}(u) du) f_Y(y)$ for some constant C . Define $H(x) := \int_0^\infty \bar{h}(x + y) f_Y(y) dy = \mathbb{E}[\bar{h}(x + Y)]$. Since Y is stochastically smaller than $\max_{i \in [I]} S_{\text{tr}, i}$, it also has an mgf in a neighbourhood of 0. Assumption 1 yields an envelope $\bar{h}(t) \leq C_0 + C_1 t^\theta$, $t \geq 0$. Using the inequality $(x + y)^\theta \leq 2^\theta (x^\theta + y^\theta)$, there exist constants C'_0 and C'_1 such that $H(x) \leq C'_0 + C'_1 x^\theta$ for all $x \geq 0$. Hence, the integral of the envelop equals

$$\iint \bar{h}(x + y) \left(II(0, x) + \int_x^\infty (u - x) \tilde{g}(u) du \right) f_Y(y) dx dy,$$

which is upper bounded by

$$\int \{C'_0 + C'_1 x^\theta\} \left(\int_x^\infty \tilde{g}(u) du + \int_x^\infty (u - x) \tilde{g}(u) du \right) dx.$$

Using the fact $\int_0^\infty x^b \int_x^\infty u^a \tilde{g}(u) du dx = \frac{1}{b+1} \mathbb{E}[\tilde{S}^{a+b+1}]$, we can readily see that the above integral is finite. This provides an integrable envelope for $h(x + y) \partial_\lambda f_{L^\lambda(0)}(x) f_Y(y)$, uniform in $0 \leq \lambda \leq \lambda_0$, and justifies differentiation under the integral sign in the definition of $\varphi(\lambda)$. By the theorem for differentiation under the integral sign (see, e.g., Theorem 2.2.7 in Chapter 2 of [Folland \(1999\)](#)),

$$\varphi'(\lambda) = \frac{d}{d\lambda} \mathbb{E}[h(L^\lambda(0) + Y)] \Big|_{\lambda=0} = \iint h(x + y) r(x) f_Y(y) dx dy.$$

and φ is differentiable at 0. Use Taylor expansion to see $\varphi(\lambda) - \varphi'(0) = \lambda \varphi''(0) + o(\lambda)$. That is,

$$\mathbb{E}[h(L^\lambda(0) + Y)] - \mathbb{E}[h(L + Y)] = \lambda \varphi'(0) + o(\lambda).$$

Multiply this display with $q_0(\lambda) = 1 - \lambda \frac{\mathbb{E}\tilde{S}^2}{2\mathbb{E}\tilde{S}} + O(\lambda^2)$ to see that

$$q_0(\lambda) \mathbb{E}[h(L^\lambda(0) + Y)] = \mathbb{E}[h(L + Y)] + O(\lambda).$$

Step 3: Show that $\sum_{k \geq 1} \mathbb{E}[h(R(k))] q_k$ is $O(\lambda)$. Let $N(\lambda)$ be the stationary number in system at an arbitrary time in the $M/G/1$ queue with arrival rate λ and service time $\tilde{S}$. Recall that our assumption 1 implies $h(R(k)) \leq C_0 + C_1 R(k)^\theta$. At stationarity, we have $N(\lambda) = 0$ if and only if $B_\infty = 0$, and $N(\lambda) = K_\infty + 1$ on the event that $\{B_\infty = 1\}$. As such,

$$\mathbb{E}[K_\infty^\theta | B_\infty = 1] = \sum_{k \geq 0} k^\theta q_k(\lambda) = \mathbb{E}[(N(\lambda) - 1)^\theta | N(\lambda) \geq 1].$$

We first control the tail of $N(\lambda)$. By the Pollaczek–Khinchine PGF formula for $M/G/1$, e.g., Equation (2.5.8) in [Tijms \(2003\)](#), and the assumption that $\tilde{S}$ has an mgf in a neighbourhood of 0, the stationary PGF $\tilde{N}_\lambda(z) = \mathbb{E}[z^{N(\lambda)}]$

is analytic in a disc $\{z : |z| < 1 + \delta\}$ uniformly for $0 < \lambda \leq \lambda_0$. By Cauchy's estimates this implies an exponential tail bound $P(N(\lambda) \geq n) \leq C_1 e^{-C_2 n}$ for all large enough n and for some $C_1, C_2 > 0$ independent of λ . On $\{B_\infty = 1\}$ we have $K_\infty = N(\lambda) - 1$. This implies that for every integer $\theta \geq 1$, $K_\infty^\theta \leq N(\lambda)^\theta$ on $\{K_\infty \geq 1, B_\infty = 1\}$. By the exponential tail bound of $N(\lambda)$, $\mathbb{E}[K_\infty^\theta | K_\infty \geq 1, B_\infty = 1] < \infty$. Combining the above arguments to see

$$\mathbb{E}[K_\infty^\theta | B_\infty = 1] = \mathbb{E}[K_\infty^\theta \mathbf{1}_{\{K_\infty \geq 1\}} | B_\infty = 1] = \mathbb{E}[K_\infty^\theta | K_\infty \geq 1, B_\infty = 1] \mathbb{P}(K_\infty \geq 1 | B_\infty = 1).$$

The first term on the right is $O(1)$, while the second term equals $1 - q_0 = \lambda \frac{\mathbb{E}\bar{S}^2}{2\mathbb{E}\bar{S}} + O(\lambda^2)$. Combine to see the above display is $O(\lambda)$ as $\lambda \downarrow 0$. Hence, for every integer $\theta \geq 1$,

$$\sum_{k \geq 0} k^\theta q_k(\lambda) = \mathbb{E}[K_\infty^\theta | B_\infty = 1] = O(\lambda), \text{ as } \lambda \downarrow 0.$$

Following a similar line of reasoning, it is routine to show that $\sum_{k \geq 1} q_k(\lambda) = O(\lambda)$. Combining these estimates, we obtain

$$\sum_{k \geq 1} q_k(\lambda) \mathbb{E}[h(R(k))] = O(\lambda), \text{ as } \lambda \downarrow 0.$$

Proof for node-level loss. We now focus on arrivals from node $i \in [I]$. Denote $R^{(i)}$ as the limiting response time of a generic call from node i . For each $k \in \mathbb{N}_0$, let $R_i(k) := [R^{(i)} | B_\infty = 1, K_\infty = k]$. This can be interpreted as the response time of a call from node i to a *stationary* queue conditional on the event that the call finds the queue busy with k calls waiting in the queue. Under the non-idling FCFS rule, conditioning on $\{B_\infty = 1, K_\infty = k\}$, $R_i(k)$ is a sum of the residual mission duration $L(k)$, the time to serve the k calls already waiting in queue, and its own travel time $S_{\text{tr},i}$. Similar to (9), $R_i(k)$ satisfies

$$R_i(0) \stackrel{d}{=} L(0) + S_{\text{tr},i}, \text{ and } R_i(k) \leq_{s.t.} L(k) + k\bar{S} + S_{\text{tr},i}, k \in \mathbb{N},$$

Conditioning on K_∞ and B_∞ as in (10), we can similarly get the node-level decomposition

$$\gamma_{\mathcal{B},i}(\lambda) = q_0(\lambda) \mathbb{E}\left[h\left(L^\lambda(0) + S_{\text{tr},i}\right)\right] + \sum_{k \geq 1} q_k(\lambda) \mathbb{E}[h(R_i(k))].$$

We can repeat Steps 1-3 above to show that $\gamma_{\mathcal{B},i}(\lambda) = \mathbb{E}[h(L + S_{\text{tr},i})] + O(\lambda)$ as $\lambda \downarrow 0$.

Appendix B: Proofs of Theorem 2 and Corollary 1

We describe the setup for this multi-server queueing problem before proceeding to the proof, which will be presented in §B.1 and §B.2.

In this multi-server spatial queue, the arrival of calls from demand node $i \in [I]$ follows a Poisson process with rate $\pi_i \lambda$, with $\pi_i \geq 0$ and $\sum_{i \in [I]} \pi_i = 1$. There are $J \geq 2$ servers in the system. The set of eligible servers $\mathcal{J}_i \subseteq [J]$ that are allowed to service calls from node i is nonempty for each $i \in [I]$. We also define $\mathbf{I}_j := \{i \in [I] : j \in \mathcal{J}_i\}$ as the set of serviceable nodes for server $j \in [J]$. The travel time for server $j \in \mathcal{J}_i$ to service a call from node i is $S_{\text{tr},ij}$, and the mission duration S_{ij} satisfies $S_{ij} = S_{\text{tr},ij} + S_{\text{os},ij}$, where $S_{\text{os},ij}$ represents the on-scene service time for server j to handle a call from node i . Recall that we have required the *dispatching rule* of the multi-server spatial queue to be first-come-first-serve (FCFS), non-idling and stationary. That is, when multiple calls are waiting in queue, these calls are served in the order of their positions in the queue, while adhering to the eligible server constraint $\mathcal{J}_i, i \in [I]$; no

server can be idle when there are serviceable calls waiting in queue (thus, the call waiting at the head of the queue will be served by the eligible server that first completes its on-hand mission); and the dispatching decisions are independent of system time (but can depend on system states).

Let $x \mapsto f_{\text{Er}}(x; k, \mu) = \mu^k x^{k-1} e^{-\mu x} / (k-1)!$ be the pdf of an Erlang distribution with rate $\mu > 0$ and shape $k \in \mathbb{N}_+$. Given a V -dimensional weight vector $\mathbf{w} := (w_1, \dots, w_V)$ satisfying $w_v \geq 0$ and $\sum_{v \in [V]} w_v = 1$, a V -phase mixed-Erlang distribution with weight $\mathbf{w}$ and rate μ can be defined by the following mixed pdf:

$$f_{\text{Mix}}(x; \mathbf{w}, \mu) := \sum_{v \in [V]} w_v f_{\text{Er}}(x; v, \mu), \quad x \geq 0.$$

Let Z_{Mix} denote a random variable having this mixed pdf. We next draw a useful connection between Z_{Mix} and the random absorption time of a continuous-time Markov chain. Let $\{X(t) : t \geq 0\}$ denote a continuous-time Markov chain with state space $\{1, 2, \dots, V, V+1\}$, where the first V states are transient and called the *phases*, while the last state $V+1$ is absorbing. The transition rate from any state $s \in [V]$ to state s' equals μ if $s' = s+1$; and zero otherwise. Let the probability distribution of the initial state $X(0)$ be given by $P[X(0) = v] = w_{V-v+1}$, $v \in [V]$, and let $Z_{\text{Ab}} = \inf\{t \geq 0 : X(t) = V+1\}$ denote the random time to reach the absorbing state $V+1$. Then, it holds that Z_{Ab} equals in distribution to Z_{Mix} .

Recall that we have assumed S_{ij} —the mission duration for server $j \in \mathcal{J}_i$ to serve a call from node $i \in [I]$ —to follow a V_{ij} -phase mixed Erlang distribution with weight $\mathbf{w}_{ij} := (w_{ij1}, w_{ij2}, \dots, w_{ijV_{ij}})$ and rate $\mu_{ij} > 0$, where V_{ij} is a positive integer and $\mathbf{w}_{ij}$ satisfies $w_{ijv} > 0$ and $\sum_{v \in [V_{ij}]} w_{ijv} = 1$. Let $\bar{S}$ denote a random variable following the Erlang distribution with shape parameter $\max_{i \in [I], j \in \mathcal{J}_i} \{V_{ij}\}$ and rate parameter $\min_{i \in [I], j \in \mathcal{J}_i} \{\mu_{ij}\}$, which has a cdf $\bar{G}(\cdot)$ and a pdf $\bar{g}(\cdot)$, and let $\underline{S}$ denote a random variable following the exponential distribution with rate $\max_{i \in [I], j \in \mathcal{J}_i} \{\mu_{ij}\}$, which has a cdf $\underline{G}(\cdot)$ and a pdf $\underline{g}(\cdot)$. In addition, define $\bar{S}^*$ as the minimum of $\min_{i \in [I]} \{|\mathcal{J}_i|\}$ independent copies of $\bar{S}$, and define $\underline{S}^*$ as the minimum of $\max_{i \in [I]} \{|\mathcal{J}_i|\}$ independent copies of $\underline{S}$. Let $\bar{S}_{\text{tr},i}$ and $\underline{S}_{\text{tr},i}$ be random variables such that $\bar{S}_{\text{tr},i} \geq_{s.t.} S_{\text{tr},ij}$ and $\underline{S}_{\text{tr},i} \leq_{s.t.} S_{\text{tr},ij}$ for all $j \in \mathcal{J}_i$, respectively. Finally, let $\bar{S}_{\text{tr}}$ and $\underline{S}_{\text{tr}}$ denote random variables such that $\bar{S}_{\text{tr}} \geq_{s.t.} \bar{S}_{\text{tr},i}$ and $\underline{S}_{\text{tr}} \leq_{s.t.} \underline{S}_{\text{tr},i}$ for all $i \in [I]$, respectively.

For the n th call that arrives to the system, where $n \in \mathbb{N}$, let $i_n \in [I]$ be the demand node of the call, $K_n \in \mathbb{N}_0$ represent the number of calls waiting in queue upon the call's arrival, and $\mathbf{i}_n := (i_{n,1}, i_{n,2}, \dots, i_{n,J})$ denote the demand nodes that the J servers are working on upon the call's arrival, where $i_{n,j} = i \in \mathcal{J}_j$ if server j is servicing a call from node i and $i_{n,j} = 0$ if server j is idle. In addition, let $B_{n,j} := \mathbb{I}(i_{n,j} \neq 0)$ denote the busy indicator that equals 1 if the n th call arrives to find server $j \in [J]$ busy; and $B_{n,j} = 0$ otherwise, and let $B_n^i := \prod_{j \in \mathcal{J}_i} B_{n,j}$, $i \in [I]$ denote the $\mathcal{J}_i$ -busy indicator that equals 1 if the call arrives to find all servers that are allowed to service node i busy; and $B_n^i = 0$ otherwise. Note that if $\mathcal{J}_i = [J]$ for all i , i.e., allowing all servers to serve any node, then it is enough to define just one indicator $B_n := \prod_{j \in [J]} B_{n,j}$, which equals one only if the n th call arrive to find all servers busy. We use $\mathbf{I}_{\mathcal{B}}^i := \{(i_1, i_2, \dots, i_J) : i_j \in \mathcal{J}_j, j \in \mathcal{J}_i, i \in [I]\}$ to denote the set of server states $\mathbf{i}_n$ in which all servers in $\mathcal{J}_i$ are busy. We note that the two events, $\{B_n^i = 1\}$ and $\{\mathbf{i}_n \in \mathbf{I}_{\mathcal{B}}^i\}$, are equivalent for each $i \in [I]$, and they both indicate that all servers in $\mathcal{J}_i$ are busy upon the n th call's arrival. Again if $\mathcal{J}_i = [J]$ for all i , then it suffices to only define the set $\mathbf{I}_{\mathcal{B}} := \{(i_1, i_2, \dots, i_J) : i_j \in [I], j \in [J]\}$, which denotes the set of server states $\mathbf{i}_n$ in which all servers are busy. Moreover, let $L_{n,j}$ denote the residual time of server j 's ongoing mission upon the n th call's arrival if $B_{n,j} = 1$; and $L_{n,j} = 0$ if $B_{n,j} = 0$, and let $L_n^i := \min_{j \in \mathcal{J}_i} \{L_{n,j}\}$ denote the residual time to the first mission completion among servers in $\mathcal{J}_i$.

if all servers in $\mathcal{J}_i$ are busy, i.e., $B_n^i = 1$; and $L_n^i = 0$ if $B_n^i = 0$. If the n th call arrives to find all servers in $\mathcal{J}_i$ busy ($B_n^i = 1$), we let $j_n^i := \arg \min_{j \in \mathcal{J}_i} \{L_{n,j}\}$ denote the servers among $\mathcal{J}_i$ that first completes its on-hand mission. Under a non-idling dispatching rule, if all servers in $\mathcal{J}_i$ are busy upon the n th call's arrival, i.e., $B_n^i = 1$, then the n th call will be served by server j_n^i . Otherwise if $B_n^i = 0$, we let $j_n^i \in \{j \in \mathcal{J}_i : L_{n,j} = 0\}$ denote one of the idle servers in $\mathcal{J}_i$ that would be dispatched by the dispatching rule (based on the system states) to service the n th call, pretending that the n th call is from node i . Since the dispatching rule is stationary, $j_n^i, i \in [I]$ does not depend on the system time (and n) but is allowed to depend on system states including K_n and $\mathbf{i}_n$. By the property of Poisson arrivals, we have that i_n is independent of $(B_n^1, B_n^2, \dots, B_n^I, L_n^1, L_n^2, \dots, L_n^I, K_n, \mathbf{i}_n)$ with $P[i_n = i] = \pi_i$ for $i \in [I]$, and that i_n is independent of $[j_n^i | B_n^i = 1]$ for $i \in [I]$ due to the non-idling dispatching rule. We also note that the residual time to the first mission completion among servers in $\mathcal{J}_i$, given that all servers in $\mathcal{J}_i$ are busy, satisfies the following stochastic relation.

LEMMA 3. *It holds that $\underline{S}^* \leq_{s.t.} [L_n^i | i_n = i, j_n^i = j, K_n = k, B_n^i = 1, \mathbf{i}_n = \mathbf{i}] \leq_{s.t.} \bar{S}^*$ for all $n \in \mathbb{N}, i \in [I], j \in \mathcal{J}_i, \mathbf{i} \in \mathbf{I}_{\mathcal{B}}^i$ and $k \in \mathbb{N}_0$.*

Proof. See Appendix B.3. □

Finally, let R_n denote the response time of the n th call. Under a non-idling and FCFS dispatching rule, R_n is a sum of the residual time to the first mission completion among servers in $\mathcal{J}_i$, i.e., L_n^i , the time to deplete other competing calls already in queue (at most K_n), and the travel time to reach the call itself. This travel time is distributed as $S_{\text{tr}, i_n, j_n^i}$ if $K_n = 0$, because server j_n^i will service the call in this case, while if $K_n > 0$, we only know that this travel time is stochastically upper bounded by $\bar{S}_{\text{tr}, i_n}$. Thus, we have the relation $R_n \stackrel{d}{=} L_n^i + S_{\text{tr}, i_n, j_n^i}$ if $K_n = 0$ and $R_n \leq_{s.t.} L_n^i + k\bar{S} + \bar{S}_{\text{tr}, i_n}$ if $K_n = k$ for all $k \in \mathbb{N}$. From the relation between R_n and L_n , the fact that $\underline{S}_{\text{tr}} \leq_{s.t.} \underline{S}_{\text{tr}, i} \leq_{s.t.} S_{\text{tr}, i, j} \leq_{s.t.} \bar{S}_{\text{tr}, i} \leq_{s.t.} \bar{S}_{\text{tr}}$ and Lemma 3, we have that

$$\underline{S}^* + \underline{S}_{\text{tr}} \leq_{s.t.} \underline{S}^* + \underline{S}_{\text{tr}, i} \leq_{s.t.} [R_n | i_n = i, j_n^i = j, K_n = 0, B_n^i = 1, \mathbf{i}_n = \mathbf{i}] \leq_{s.t.} \bar{S}^* + \bar{S}_{\text{tr}, i} \leq_{s.t.} \bar{S}^* + \bar{S}_{\text{tr}}$$

and $[R_n | i_n = i, j_n = j, K_n = k, B_n^i = 1, \mathbf{i}_n = \mathbf{i}] \leq_{s.t.} \bar{S}^* + k\bar{S} + \bar{S}_{\text{tr}, i}$ if $k \in \mathbb{N}$ for all $i \in [I], j \in \mathcal{J}_i$ and $\mathbf{i} \in \mathbf{I}_{\mathcal{B}}^i$.

The process $\{(i_n, j_n^1, j_n^2, \dots, j_n^I, \mathbf{i}_n, K_n, B_n^1, B_n^2, \dots, B_n^I, L_n^1, L_n^2, \dots, L_n^I, R_n) : n \in \mathbb{N}\}$ is a regenerative process under a stationary dispatching rule, with the epoch n on which $B_{n,j} = 0$ for all $j \in [J]$, i.e., a call arrives to find all servers idle, as a regeneration epoch. By Theorem 2.2.4 of Tijms (2003), the limit of the sequence of probabilities $P[i_n = i, j_n^1 = j^1, \dots, j_n^I = j^I, \mathbf{i}_n = \mathbf{i}, K_n \leq k, B_n^1 = b^1, \dots, B_n^I = b^I, L_n^1 \leq l^1, \dots, L_n^I \leq l^I, R_n \leq r]$ exists for any fixed $i, (j^i)_{i \in [I]}, \mathbf{i}, s, k, (b^i)_{i \in [I]}, (l^i)_{i \in [I]}, r$. As in the proof of Theorem 1, the sequences of $i_n, (j_n^i)_{i \in [I]}, \mathbf{i}_n, K_n, (B_n^i)_{i \in [I]}, (L_n^i)_{i \in [I]}$ and R_n are tight. The marginal tightness is equivalent to joint tightness. These facts implies that there exists a vector of limiting random variables $(i_\infty, j_\infty^1, \dots, j_\infty^I, \mathbf{i}_\infty, K_\infty, B_\infty^1, \dots, B_\infty^I, L_\infty^1, \dots, L_\infty^I, R_\infty)$ such that $(i_n, j_n^1, \dots, j_n^I, \mathbf{i}_n, K_n, B_n^1, \dots, B_n^I, L_n^1, \dots, L_n^I, R_n) \xrightarrow{d} (i_\infty, j_\infty^1, \dots, j_\infty^I, \mathbf{i}_\infty, K_\infty, B_\infty^1, \dots, B_\infty^I, L_\infty^1, \dots, L_\infty^I, R_\infty)$ as $n \rightarrow \infty$. By the independence between i_n and $(B_n^1, \dots, B_n^I, K_n, \mathbf{i}_n)$ and between i_n and $[j_n^i | B_n^i = 1]$ for $n \in \mathbb{N}$ and $i \in [I]$, it holds that i_∞ is independent of $(B_\infty^1, \dots, B_\infty^I, K_\infty, \mathbf{i}_\infty)$ and $[j_\infty^i | B_\infty^i = 1]$ with $P[i_\infty = i] = \pi_i, i \in [I]$. We also have $P[B_\infty^i = 1] = \sum_{\mathbf{i} \in \mathbf{I}_{\mathcal{B}}^i} P[\mathbf{i}_\infty = \mathbf{i}, B_\infty^i = 1]$ by the equivalence between events $\{B_n^i = 1\}$ and $\{\mathbf{i}_n \in \mathbf{I}_{\mathcal{B}}^i\}$ for $n \in \mathbb{N}$. In addition, we have $\underline{S}^* \leq_{s.t.} [L_\infty^i | i_\infty = i, j_\infty^i = j, K_\infty = k, B_\infty^i = 1, \mathbf{i}_\infty = \mathbf{i}] \leq_{s.t.} \bar{S}^*$ for all $i \in [I], j \in \mathcal{J}_i, \mathbf{i} \in \mathbf{I}_{\mathcal{B}}^i$ and $k \in \mathbb{N}_0$ since Lemma 3 holds for any $n \in \mathbb{N}$. For convenience, we further define two sets of quantities.

For each $i \in [I], j \in \mathcal{J}_i, k \in \mathbb{N}_0$ and $\mathbf{i} \in \mathbf{I}_{\mathcal{B}}^i$, let $p_{k,j,\mathbf{i}}^i := P[j_\infty^i = j, K_\infty = k, \mathbf{i}_\infty = \mathbf{i}, B_\infty^i = 1]$ denote the limiting probability that all servers in $\mathcal{J}_i$ are busy with k calls waiting in queue, the status of servers is given by $\mathbf{i}$, and server j is the server among $\mathcal{J}_i$ to first complete the on-hand mission. Let $p_k^i = \sum_{j \in \mathcal{J}_i} \sum_{\mathbf{i} \in \mathbf{I}_{\mathcal{B}}^i} p_{k,j,\mathbf{i}}^i$ denote the limiting probability that all servers in $\mathcal{J}_i$ are busy with k calls waiting in queue, and $p_{\mathcal{B}}^i = \sum_{k \in \mathbb{N}_0} p_k^i$ be the limiting probability that all servers in $\mathcal{J}_i$ are busy. Because $P[B_\infty^i = 1] = \sum_{\mathbf{i} \in \mathbf{I}_{\mathcal{B}}^i} P[\mathbf{i}_\infty = \mathbf{i}, B_\infty^i = 1]$, it holds that $p_{\mathcal{B}}^i = P[B_\infty^i = 1]$. We also define $\tilde{\pi}_i = \pi_i p_{\mathcal{B}}^i / \sum_{i \in [I]} \pi_i p_{\mathcal{B}}^i$, which can be interpreted as the limiting conditional probability that a call is from node i , given that the call arrives to find all its eligible servers busy. It is obvious that $\sum_{i \in [I]} \tilde{\pi}_i = 1$. Moreover, let $q_{k,j,\mathbf{i}}^i = p_{k,j,\mathbf{i}}^i / p_{\mathcal{B}}^i$ denote the limiting conditional probability that the status of servers is given by $\mathbf{i}$, there are k calls waiting in queue, and server j is the server to first complete the on-hand mission among $\mathcal{J}_i$, given that all servers in $\mathcal{J}_i$ are busy, and let $q_k^i = \sum_{j \in \mathcal{J}_i} \sum_{\mathbf{i} \in \mathbf{I}_{\mathcal{B}}^i} q_{k,j,\mathbf{i}}^i$ denote the limiting conditional probability of having k calls waiting in queue, given that all servers in $\mathcal{J}_i$ are busy. It is clear that $\sum_{k \in \mathbb{N}_0} q_k^i = 1$ for all $i \in [I]$. We note that if $\mathcal{J}_i = [J]$ for all $i \in [I]$, then the quantities $p_{k,j,\mathbf{i}}^i, p_k^i, p_{\mathcal{B}}^i, q_{k,j,\mathbf{i}}^i$ and q_k^i are identical across $i \in [I]$, and $\tilde{\pi}_i$ would simply equal π_i , i.e., the fraction of arrival from node i . The following result provides upper bound for q_k^i when the total arrival rate λ is small.

LEMMA 4. For any fixed $u > 0$, there exists a constant $C^\lambda = \frac{e^{\lambda u} - 1}{\bar{G}(u)(1 - \bar{G}(u))}$, such that $q_k^i \leq C^\lambda (C^\lambda + 1)^{k-1} q_0^i$ for all $i \in [I], k \in \mathbb{N}$.

Proof. See Appendix B.4. □

For each $i \in [I], j \in \mathcal{J}_i, k \in \mathbb{N}_0$ and $\mathbf{i} \in \mathbf{I}_{\mathcal{B}}^i$, we define two non-negative random variables $L_j^i(k, \mathbf{i})$ and $R_j^i(k, \mathbf{i})$ by $L_j^i(k, \mathbf{i}) = [L_\infty^i | i_\infty = i, j_\infty^i = j, B_\infty^i = 1, K_\infty = k, \mathbf{i}_\infty = \mathbf{i}]$ and $R_j^i(k, \mathbf{i}) = [R_\infty | i_\infty = i, j_\infty^i = j, B_\infty^i = 1, K_\infty = k, \mathbf{i}_\infty = \mathbf{i}]$. Here, $L_j^i(k, \mathbf{i})$ can be understood as the conditional limit distribution for the residual time to the first mission completion at the moment that a new call arrives, conditional on the event that the call arrives from node i to find all servers in $\mathcal{J}_i$ busy with k calls waiting in queue, the status of servers is given by $\mathbf{i}$, and server j is the server among $\mathcal{J}_i$ to first complete the on-hand mission. Similarly, $R_j^i(k, \mathbf{i})$ can be interpreted as the conditional limit distribution for the response time of a call, conditional on the event that the call arrives from node i to find all servers in $\mathcal{J}_i$ busy with k calls waiting in queue, the status of servers is given by $\mathbf{i}$, and server j is the server among $\mathcal{J}_i$ to first complete the on-hand mission. From the stochastic relation satisfied by L_n and R_n , it holds that the limiting random variable $L_j^i(k, \mathbf{i})$ and $R_j^i(k, \mathbf{i})$ satisfies

$$\begin{aligned} R_j^i(0, \mathbf{i}) &= L_j^i(0, \mathbf{i}) + S_{\text{tr},i,j} \leq_{s.t.} \bar{S}^* + \bar{S}_{\text{tr},i} \leq_{s.t.} \bar{S}^* + \bar{S}_{\text{tr}} \\ R_j^i(0, \mathbf{i}) &= L_j^i(0, \mathbf{i}) + S_{\text{tr},i,j} \geq_{s.t.} \underline{S}^* + \underline{S}_{\text{tr},i} \geq_{s.t.} \underline{S}^* + \underline{S}_{\text{tr}} \\ R_j^i(k, \mathbf{i}) &\leq_{s.t.} \bar{S}^* + k\bar{S} + \bar{S}_{\text{tr},i} \leq_{s.t.} \bar{S}^* + k\bar{S} + \bar{S}_{\text{tr}} \end{aligned} \quad (12)$$

for all $i \in [I], j \in \mathcal{J}_i, k \in \mathbb{N}$, and $\mathbf{i} \in \mathbf{I}_{\mathcal{B}}^i$.

B.1. Proof of Theorem 2.

We focus on the loss of those calls that arrive to find all eligible servers busy (and thus have to wait in queue). Over the first n calls, there are $\sum_{l=1}^n B_l = \sum_{l=1}^n \sum_{i \in [I]} B_l^i \mathbb{I}(i_n = i)$ such calls. Their empirical average loss is

$$\frac{\sum_{l=1}^n \sum_{i \in [I]} h(R_l) B_l^i \mathbb{I}(i_n = i)}{\sum_{l=1}^n \sum_{i \in [I]} B_l^i \mathbb{I}(i_n = i)} = \frac{n}{\sum_{l=1}^n \sum_{i \in [I]} B_l^i \mathbb{I}(i_n = i)} \times \frac{1}{n} \sum_{l=1}^n \sum_{i \in [I]} h(R_l) B_l^i \mathbb{I}(i_n = i).$$

The first term on the right hand side can be readily seen to converge to $1/\sum_{i \in [I]} P[B_\infty^i = 1, i_\infty = i]$. According to the renewal reward theorem, the second term on the right hand side and its expectation converge to the same constant as $n \rightarrow \infty$. A routine calculation shows that the limit is $\mathbb{E}[\sum_{i \in [I]} h(R_\infty) B_\infty^i \mathbb{I}(i_\infty = i)]$ (to see this, we can use the standard argument that first considers a simple function for h , then a non-negative function, and finally a general integrable function). By writing

$$\sum_{i \in [I]} h(R_\infty) B_\infty^i \mathbb{I}(i_\infty = i) = \sum_{i \in [I]} \sum_{j \in \mathcal{J}_i} \sum_{k \in \mathbb{N}_0} \sum_{\mathbf{i} \in \mathbf{I}_{\mathcal{B}}^i} h(R_\infty) B_\infty^i \mathbb{I}(i_\infty = i, j_\infty^i = j, K_\infty = k, \mathbf{i}_\infty = \mathbf{i}),$$

this limit is equal to

$$\sum_{i \in [I]} \sum_{j \in \mathcal{J}_i} \sum_{k \in \mathbb{N}_0} \sum_{\mathbf{i} \in \mathbf{I}_{\mathcal{B}}^i} \mathbb{E}[h(R_\infty) B_\infty^i \mathbb{I}(i_\infty = i, j_\infty^i = j, K_\infty = k, \mathbf{i}_\infty = \mathbf{i})]$$

by an application of Fubini's theorem. This means that the long-run average loss of a call that finds the system busy equals

$$\frac{\sum_{i \in [I]} \sum_{j \in \mathcal{J}_i} \sum_{k \in \mathbb{N}_0} \sum_{\mathbf{i} \in \mathbf{I}_{\mathcal{B}}^i} \mathbb{E}[h(R_\infty) B_\infty^i \mathbb{I}(i_\infty = i, j_\infty^i = j, K_\infty = k, \mathbf{i}_\infty = \mathbf{i})]}{\sum_{i \in [I]} P[B_\infty^i = 1, i_\infty = i]},$$

which further equals

$$\frac{\sum_{i \in [I]} \sum_{j \in \mathcal{J}_i} \sum_{k \in \mathbb{N}_0} \sum_{\mathbf{i} \in \mathbf{I}_{\mathcal{B}}^i} \mathbb{E}[h(R_j^i(k, \mathbf{i}))] P[B_\infty^i = 1, i_\infty = i, j_\infty^i = j, K_\infty = k, \mathbf{i}_\infty = \mathbf{i}]}{\sum_{i \in [I]} P[B_\infty^i = 1, i_\infty = i]}$$

by the law of total expectation, and the fact that conditional on $[B_\infty^i = 1, i_\infty = i, j_\infty^i = j, K_\infty = k, \mathbf{i}_\infty = \mathbf{i}]$, R_∞ has the same distribution as $R_j^i(k, \mathbf{i})$. In addition, since i_∞ is independent of $(B_\infty^i, K_\infty, \mathbf{i}_\infty, s_\infty^{\text{cr}})$ and $[j_\infty^i | B_\infty^i = 1]$ with $P[i_\infty = i] = \pi_i, i \in [I]$, we have $P[B_\infty^i = 1, i_\infty = i] = \pi_i P[B_\infty^i = 1]$ and

$$P[B_\infty^i = 1, i_\infty = i, j_\infty^i = j, K_\infty = k, \mathbf{i}_\infty = \mathbf{i}] = \pi_i P[B_\infty^i = 1, j_\infty^i = j, K_\infty = k, \mathbf{i}_\infty = \mathbf{i}].$$

Therefore, by our definition of the limiting probabilities above, the long-run average loss of a call that finds the system busy can be rewritten as

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{\sum_{l=1}^n \sum_{i \in [I]} h(R_l) B_l^i \mathbb{I}(i_n = i)}{\sum_{l=1}^n \sum_{i \in [I]} B_l^i \mathbb{I}(i_n = i)} &= \frac{\sum_{i \in [I]} \sum_{j \in \mathcal{J}_i} \sum_{k \in \mathbb{N}_0} \sum_{\mathbf{i} \in \mathbf{I}_{\mathcal{B}}^i} \pi_i \mathbb{E}[h(R_j^i(k, \mathbf{i}))] p_{k,j,\mathbf{i}}^i}{\sum_{i \in [I]} \pi_i p_{\mathcal{B}}^i} \\ &= \frac{\sum_{i \in [I]} \sum_{j \in \mathcal{J}_i} \sum_{k \in \mathbb{N}_0} \sum_{\mathbf{i} \in \mathbf{I}_{\mathcal{B}}^i} \pi_i p_{\mathcal{B}}^i \mathbb{E}[h(R_j^i(k, \mathbf{i}))] p_{k,j,\mathbf{i}}^i / p_{\mathcal{B}}^i}{\sum_{i \in [I]} \pi_i p_{\mathcal{B}}^i} \\ &= \sum_{i \in [I]} \sum_{j \in \mathcal{J}_i} \sum_{\mathbf{i} \in \mathbf{I}_{\mathcal{B}}^i} \tilde{\pi}_i \mathbb{E}[h(R_j^i(0, \mathbf{i}))] q_{0,j,\mathbf{i}}^i \\ &\quad + \sum_{i \in [I]} \sum_{j \in \mathcal{J}_i} \sum_{k \in \mathbb{N}} \sum_{\mathbf{i} \in \mathbf{I}_{\mathcal{B}}^i} \tilde{\pi}_i \mathbb{E}[h(R_j^i(k, \mathbf{i}))] q_{k,j,\mathbf{i}}^i. \end{aligned} \quad (13)$$

Since $\sum_{k \in \mathbb{N}_0} q_k^i = 1, q_k^i = \sum_{j \in \mathcal{J}_i} \sum_{\mathbf{i} \in \mathbf{I}_{\mathcal{B}}^i} q_{k,j,\mathbf{i}}^i$ and $q_{k,j,\mathbf{i}}^i \geq 0$, Lemma 4 implies that $q_{k,j,\mathbf{i}}^i \rightarrow 0$ as $\lambda \downarrow 0$ for all $i \in [I], k \in \mathbb{N}$. Hence, from (13) we obtain that

$$\lim_{\lambda \downarrow 0} \frac{\sum_{l=1}^n \sum_{i \in [I]} h(R_l) B_l^i \mathbb{I}(i_n = i)}{\sum_{l=1}^n \sum_{i \in [I]} B_l^i \mathbb{I}(i_n = i)} = \lim_{\lambda \downarrow 0} \sum_{i \in [I]} \sum_{j \in \mathcal{J}_i} \sum_{\mathbf{i} \in \mathbf{I}_{\mathcal{B}}^i} \tilde{\pi}_i \mathbb{E}[h(R_j^i(0, \mathbf{i}))] q_{0,j,\mathbf{i}}^i \quad (14)$$

Because we have $\underline{S}^* + \underline{S}_{\text{tr}} \leq_{s.t.} R_j^i(0, \mathbf{i}) \leq_{s.t.} \bar{S}^* + \bar{S}_{\text{tr}}$ for all $i \in [I], j \in \mathcal{J}_i, \mathbf{i} \in \mathbf{I}_{\mathcal{B}}^i$ from (12), that $\sum_{i \in [I]} \tilde{\pi}_i = 1$, and $\sum_{j \in \mathcal{J}_i} \sum_{\mathbf{i} \in \mathbf{I}_{\mathcal{B}}^i} q_{0,j,\mathbf{i}}^i = q_0^i \rightarrow 1$ as $\lambda \downarrow 0$ by Lemma 4, it holds that

$$\mathbb{E}[h(\underline{S}^* + \underline{S}_{\text{tr}})] \leq \lim_{\lambda \downarrow 0} \lim_{n \rightarrow \infty} \frac{\sum_{l=1}^n \sum_{i \in [I]} h(R_l) B_l^i \mathbb{I}(i_n = i)}{\sum_{l=1}^n \sum_{i \in [I]} B_l^i \mathbb{I}(i_n = i)} \leq \mathbb{E}[h(\bar{S}^* + \bar{S}_{\text{tr}})].$$

Here, unlike the proof of Theorem 1, we do not need to establish the uniform integrability of the residual time to the first mission completion for all small positive λ . This is because the mission duration is assumed to follow a mixed Erlang distribution, so that the stochastic relation in Lemma 3 holds for any $\lambda > 0$. This stochastic relation together with the non-decreasing assumption of h leads to the stated bounds.

Now if we additionally have $\mathcal{J}_i = [J]$ for all $i \in [I]$, then it holds that $\tilde{\pi}_i = \pi_i$ for $i \in [I]$ and the values of $q_{0,j,\mathbf{i}}^i$ are identical across $i \in [I]$. Moreover, we have $\underline{S}^* + \underline{S}_{\text{tr},i} \leq_{s.t.} R_j^i(0, \mathbf{i}) \leq_{s.t.} \bar{S}^* + \bar{S}_{\text{tr},i}$ for all $i \in [I], j \in \mathcal{J}_i, \mathbf{i} \in \mathbf{I}_{\mathcal{B}}^i$ from (12). Then it follows from (14) and Lemma 4 that

$$\sum_{i \in [I]} \pi_i \mathbb{E}[h(\underline{S}^* + \underline{S}_{\text{tr},i})] \leq \lim_{\lambda \downarrow 0} \lim_{n \rightarrow \infty} \frac{\sum_{l=1}^n \sum_{i \in [I]} h(R_l) B_l^i \mathbb{I}(i_n = i)}{\sum_{l=1}^n \sum_{i \in [I]} B_l^i \mathbb{I}(i_n = i)} \leq \sum_{i \in [I]} \pi_i \mathbb{E}[h(\bar{S}^* + \bar{S}_{\text{tr},i})].$$

B.2. Corollary 1 and its Proof.

COROLLARY 1. *Continue with Theorem 2. Assume further that $\mathcal{J}_i = [J]$ for all $i \in [I]$.*

- *Suppose that for $i \in [I], j \in [J]$, the mission durations S_{ij} are independent and exponentially distributed with a common rate parameter μ , and the travel times $S_{\text{tr},ij}$ are independent and identically distributed as S_{tr} . Letting $S^* \sim \exp(J\mu)$, we have*

$$\gamma_{\mathcal{B}}(0+) = \mathbb{E}[h(S^* + S_{\text{tr}})].$$

- *Suppose that for each $j \in [J]$, the mission durations S_{ij} are independent and exponentially distributed with rate parameters μ_j for all $i \in [I]$, and the travel times $S_{\text{tr},ij}$ are independent and identically distributed as $S_{\text{tr},j}$ for all $i \in [I]$. Letting $S^* \sim \exp(\sum_{j \in [J]} \mu_j)$, we have*

$$\gamma_{\mathcal{B}}(0+) = \sum_{j \in [J]} \frac{\mu_j}{\sum_{j' \in [J]} \mu_{j'}} \mathbb{E}[h(S^* + S_{\text{tr},j})].$$

- *Suppose that for each $i \in [I]$, the mission durations S_{ij} are independent and exponentially distributed with rate parameters μ_i for all $j \in [J]$, and the travel times $S_{\text{tr},ij}$ are independent and identically distributed as $S_{\text{tr},i}$ for all $j \in [J]$. Then, $\gamma_{\mathcal{B}}(0+)$ has a closed-form expression given by (16) of Appendix B.2.1.*

B.2.1. Proof of Corollary 1. We follow the setups at the beginning of Appendix B. When the mission duration S_{ij} follows an exponential distribution, it holds that $V_{ij} = 1$ and $w_{ij1} = 1$. Moreover, under the assumption that $\mathcal{J}_i = [J]$ for all $i \in [I]$, we simply write $\tilde{\pi}_i$ as π_i , $q_{0,j,\mathbf{i}}^i$ as $q_{0,j,\mathbf{i}}$, q_0^i as q_0 , B_l^i as B_l , and $\mathbf{I}_{\mathcal{B}}^i$ as $\mathbf{I}_{\mathcal{B}}$ for all $i \in [I]$ in the following. Given any state $\mathbf{i} \in \mathbf{I}_{\mathcal{B}}$ in which all servers are busy, the residual time to the first mission completion becomes the minimum of J independent exponential distributions with rates $\mu_{i,j}$. We let $S^*(\mathbf{i})$ denote this minimum. It is well known that $S^*(\mathbf{i})$ follows an exponential distribution with rate $\sum_{j \in [J]} \mu_{i,j}$, which we denote by $S^*(\mathbf{i}) \sim \exp(\sum_{j \in [J]} \mu_{i,j})$. In the following, we discuss the three special cases in Corollary 1 one by one.

Special case 1: For this case of colocated bases and colocated demands, the service rates satisfy $\mu_{ij} = \mu$ for all $i \in [I]$ and $j \in [J]$, and we use S_{tr} to denote the common travel time. We name all demand nodes as node 1. Then the set of feasible state $\mathbf{i}$, given that all servers are busy, becomes $\mathbf{I}_{\mathcal{B}} = \{(1, 1, \dots, 1) \in \mathbb{Z}^J\}$, representing that all the J servers are servicing node 1. From (12), the response time $R_j^i(0, (1, \dots, 1))$ equals in distribution to $S^* + S_{tr}$ for all $i \in [I]$, where $S^* \sim \exp(J\mu)$ represents the residual time to the first mission completion. From (14) we have

$$\begin{aligned} \lim_{\lambda \rightarrow 0} \lim_{n \rightarrow \infty} \frac{\sum_{l=1}^n h(R_l) B_l}{\sum_{l=1}^n B_l} &= \lim_{\lambda \rightarrow 0} \sum_{i \in [I]} \sum_{j \in [J]} \sum_{\mathbf{i} \in \mathbf{I}_{\mathcal{B}}} \pi_i \mathbb{E}[h(R_j^i(0, \mathbf{i}))] q_{(0, j, \mathbf{i})} \\ &= \lim_{\lambda \rightarrow 0} \sum_{i \in [I]} \sum_{j \in [J]} \pi_i \mathbb{E}[h(S^* + S_{tr})] q_{(0, j, (1, \dots, 1))} \\ &= \mathbb{E}[h(S^* + S_{tr})], \end{aligned}$$

where the last identity follows from Lemma 4 that $\lim_{\lambda \downarrow 0} \sum_{j \in [J]} q_{(0, j, (1, \dots, 1))} = \lim_{\lambda \rightarrow 0} q_0 = 1$.

Special case 2: In this case of dispersed bases and colocated demands, the service rates satisfy $\mu_{ij} = \mu_j$ for all $i \in [I]$ and $j \in [J]$, and we use $S_{tr,j}$ to denote the travel time of server j . By naming all demand nodes as node 1, we again have $\mathbf{I}_{\mathcal{B}} = \{(1, \dots, 1)\}$. From (12), the response time $R_j^i(0, (1, \dots, 1))$ equals in distribution to $S^* + S_{tr,j}$ for all $i \in [I], j \in [J]$, where $S^* \sim \exp(\sum_{j \in [J]} \mu_j)$ represents the residual time to the first mission completion. Moreover, by the property of exponential distributions, it holds that $q_{0, j, (1, 1, \dots, 1)} / \sum_{j \in [J]} q_{0, j, (1, 1, \dots, 1)} = \mu_j / \sum_{j \in [J]} \mu_j$ for $j \in [J]$, which gives the probability that server $j \in [J]$ is the server to first finish its on-hand mission. By (14) and Lemma 4, we have

$$\begin{aligned} \lim_{\lambda \rightarrow 0} \lim_{n \rightarrow \infty} \frac{\sum_{l=1}^n h(R_l) B_l}{\sum_{l=1}^n B_l} &= \lim_{\lambda \rightarrow 0} \sum_{i \in [I]} \sum_{j \in [J]} \pi_i \mathbb{E}[h(R_j^i(0, (1, \dots, 1)))] q_{(0, j, (1, \dots, 1))} \\ &= \lim_{\lambda \rightarrow 0} \sum_{i \in [I]} \sum_{j \in [J]} \pi_i \mathbb{E}[h(S^* + S_{tr,j})] \frac{\mu_j}{\sum_{j \in [J]} \mu_j} \sum_{j \in [J]} q_{0, j, (1, 1, \dots, 1)} \\ &= \sum_{i \in [I]} \sum_{j \in [J]} \pi_i \mathbb{E}[h(S^* + S_{tr,j})] \frac{\mu_j}{\sum_{j \in [J]} \mu_j}. \end{aligned}$$

Special case 3: In this case of colocated servers and dispersed demands, the service rates satisfy $\mu_{ij} = \mu_i$ for all $i \in [I]$ and $j \in [J]$, and we use $S_{tr,i}$ to denote the travel time to node i . Due to dispersed demands, we have $\mathbf{I}_{\mathcal{B}} = \{(i_1, i_2, \dots, i_J) : i_j \in [I], j \in [J]\}$. From (12), the response time $R_j^i(0, \mathbf{i})$ equals in distribution to $S^*(\mathbf{i}) + S_{tr,i}$ for all $\mathbf{i} \in \mathbf{I}_{\mathcal{B}}, i \in [I]$ and $j \in [J]$, where $S^*(\mathbf{i}) \sim \exp(\sum_{j \in [J]} \mu_{i_j})$ represents the residual time to the first mission completion. Moreover, under a non-idling dispatching rule and with identical servers (Tijms 2003), we have for all $\mathbf{i} \in \mathbf{I}_{\mathcal{B}}$ that

$$\frac{\sum_{j \in [J]} q_{0, j, \mathbf{i}}}{q_0} = \prod_{j \in [J]} \frac{\pi_{i_j} \mathbb{E} S_{i_j}}{\sum_{i \in [I]} \pi_i \mathbb{E} S_i}. \quad (15)$$

Again by (14) and Lemma 4, we have that

$$\begin{aligned} \lim_{\lambda \rightarrow 0} \lim_{n \rightarrow \infty} \frac{\sum_{l=1}^n h(R_l) B_l}{\sum_{l=1}^n B_l} &= \lim_{\lambda \rightarrow 0} \sum_{i \in [I]} \sum_{j \in [J]} \sum_{\mathbf{i} \in \mathbf{I}_{\mathcal{B}}} \pi_i \mathbb{E}[h(R_j^i(0, \mathbf{i}))] q_{0, j, \mathbf{i}} \\ &= \lim_{\lambda \rightarrow 0} \sum_{i \in [I]} \sum_{\mathbf{i} \in \mathbf{I}_{\mathcal{B}}} \pi_i \mathbb{E}[h(R_j^i(0, \mathbf{i}))] \left(\prod_{j \in [J]} \frac{\pi_{i_j} \mathbb{E} S_{i_j}}{\sum_{i \in [I]} \pi_i \mathbb{E} S_i} \right) q_0 \\ &= \sum_{i \in [I]} \sum_{\mathbf{i} \in \mathbf{I}_{\mathcal{B}}} \pi_i \mathbb{E}[h(S^*(\mathbf{i}) + S_{tr,i})] \prod_{j \in [J]} \frac{\pi_{i_j} \mathbb{E} S_{i_j}}{\sum_{i \in [I]} \pi_i \mathbb{E} S_i}, \end{aligned} \quad (16)$$

where the second identity follows from (15).

B.3. Proof of Lemma 3

We will prove the result by utilizing the connection between the mixed Erlang distribution and the random absorption time of its embedded continuous-time Markov chain. Given that all servers in $\mathcal{J}_i$ are busy with status given by $\mathbf{i} \in \mathbf{I}_{\mathcal{B}}^i$, the residual mission duration of server j (i.e., $L_{n,j}$) must be in one of the transient phases $s_j \in [V_{i,j}]$ of the embedded Markov chain for all $j \in \mathcal{J}_i$. Hence, $L_{n,j}$ follows an Erlang distribution with shape parameter $(V_{i,j} - s_j + 1)$ and rate parameter $\mu_{i,j}$. We let $E_j(i_j, s_j)$ be a generic random variable that follows this Erlang distribution. It then holds that the residual time to the first mission completion, given that all servers in $\mathcal{J}_i$ are busy with status given by $\mathbf{i}$, satisfy $[L_n^i | B_n^i = 1, \mathbf{i}_n = \mathbf{i}] \stackrel{d}{=} \min_{j \in \mathcal{J}_i} \{E_j(i_j, s_j)\}$. By our definition of the random variables $\underline{S}$ and $\bar{S}$, it is clear that $\underline{S} \leq_{s.t.} E_j(i_j, s_j) \leq_{s.t.} \bar{S}$ for all feasible $s_j \in [V_{i,j}]$. This implies that $\underline{S}^* \leq_{s.t.} \min_{j \in \mathcal{J}_i} \{E_j(i_j, s_j^{\text{cr}})\} \leq_{s.t.} \bar{S}^*$ for all feasible $s_j \in [V_{i,j}]$ by our definition of the random variables $\underline{S}^*$ and $\bar{S}^*$ and the fact that $\min_{i \in [I]} \{|\mathcal{J}_i|\} \leq |\mathcal{J}_i| \leq \max_{i \in [I]} \{|\mathcal{J}_i|\}$. The proof can be completed by noting that the above arguments remains true when conditioning on the joint event $[i_n = i, j_n^i = j, K_n = k, B_n^i = 1, \mathbf{i}_n = \mathbf{i}]$ for any $i \in [I], j \in \mathcal{J}_i, k \in \mathbb{N}_0$ and $\mathbf{i} \in \mathbf{I}_{\mathcal{B}}^i$. $\square$

B.4. Proof of Lemma 4

We prove the result for any fixed $i \in [I]$. Define a system state indicator $B^i(t)$ such that $B^i(t) = 1$ if all servers in $\mathcal{J}_i$ are busy at time $t \geq 0$; and $B^i(t) = 0$ otherwise. Denote the number of calls waiting in queue at time $t \geq 0$ as $N_q(t)$. By the time symmetry arguments for a stable queueing system (Burman and Smith 1983), it holds for any $u > 0$ and $k \in \mathbb{N}$ that

$$\begin{aligned} \lim_{r \rightarrow \infty} \Pr(B^i(r+u) = 1, N_q(r+u) < k, N_q(r) \geq k) = \\ \lim_{r \rightarrow \infty} \Pr(N_q(r+u) \geq k, B^i(r) = 1, N_q(r) < k). \end{aligned} \quad (17)$$

To proceed, we derive upper and lower bounds for the above limiting probabilities.

When the queue is nonempty at time $t \geq 0$, i.e., $N_q(t) > 0$, we let $\mathcal{J}(t) \subseteq [J]$ be the collection of eligible servers of calls waiting in the queue, so that if a server in $\mathcal{J}(t)$ completes its on-going mission, a call in the queue will enter service. We also let $\mathcal{J}_{\mathcal{B}}(t) \subseteq [J]$ denote the set of busy servers at time $t \geq 0$. It is clear that $\mathcal{J}(t) \subseteq \mathcal{J}_{\mathcal{B}}(t)$ if $N_q(t) > 0$ due to the non-idling dispatching rule. For $j \in \mathcal{J}_{\mathcal{B}}(t)$, we denote $L_j(t)$ as the residual mission duration of the ongoing service of server j at time t . We let $S_{\text{next}}(t, \mathcal{J})$ denote the time interval from the completion of the first on-going service among servers in $\mathcal{J}$ after time t to the next mission completion by servers among $\mathcal{J}$, where $\mathcal{J} \subseteq [J]$. Let $\text{Arr}((a, b])$ and $\text{Ser}((a, b], \mathcal{J})$ denote the number of arrivals and mission completions among servers in $\mathcal{J}$ during $(a, b]$ for any $0 < a < b$ and $\mathcal{J} \subseteq [J]$. For each $r > 0$, the prelimit of the left-hand side of (17) can be bounded from below as follows:

$$\begin{aligned} & \Pr(B^i(r+u) = 1, N_q(r+u) < k, N_q(r) \geq k) \\ & \geq \Pr(B^i(r+u) = 1, N_q(r+u) < k, N_q(r) = k) \\ & = \Pr(N_q(r+u) < k | B^i(r+u) = 1, N_q(r) = k) P[B^i(r+u) = 1, N_q(r) = k] \\ & \geq \Pr(\text{Ser}((r, r+u], \mathcal{J}(r)) = 1 | B^i(r+u) = 1, N_q(r) = k) \Pr(\text{Arr}((r, r+u]) = 0) P[B^i(r+u) = 1, N_q(r) = k] \\ & \geq \Pr\left(\min_{j \in \mathcal{J}(r)} \{L_j(r)\} \leq u | B^i(r+u) = 1, N_q(r) = k\right) \Pr(S_{\text{next}}(r, \mathcal{J}(r)) > u | N_q(r) = k) \Pr(\text{Arr}((r, r+u]) = 0) \\ & \quad P[B^i(r+u) = 1, N_q(r) = k] \\ & \geq \Pr(\bar{S} \leq u | B^i(r+u) = 1, N_q(r) = k) (1 - \underline{G}(u))^J \Pr(\text{Arr}((r, r+u]) = 0) P[B^i(r+u) = 1, N_q(r) = k] \\ & = \bar{G}(u)(1 - \underline{G}(u))^J e^{-\lambda u} P[B^i(r+u) = 1, N_q(r) = k] \end{aligned} \quad (18)$$

The second inequality follows from comparison with the event that no call arrives and only one service is completed among servers in $\mathcal{J}(r)$ during the time interval $(r, r+u]$. The third inequality is due to the fact that if the minimum residual mission time among servers in $\mathcal{J}(r)$ is less than u , and the time interval from this earliest mission completion to the next mission completion among servers in $\mathcal{J}(r)$ is greater than u , then it results in a single mission completion among servers in $\mathcal{J}(r)$ within the timeframe $(r, r+u]$. The fourth inequality follows from two facts. First, because $L_j(r) \leq_{s.t.} \bar{S}$ for all $j \in \mathcal{J}(r)$ due to the mixed-Erlang assumption, we have $\min_{j \in \mathcal{J}(r)} \{L_j(r)\} \leq_{s.t.} \bar{S}$, and thus $\Pr\left(\min_{j \in \mathcal{J}(r)} \{L_j(r)\} \leq u \mid N_q(r) = k\right) \geq \Pr\left(\bar{S} \leq u \mid N_q(r) = k\right)$. Second, consider the moment of the first mission completion among servers in $\mathcal{J}(r)$. The server who completes the mission then initiates the next service by taking a call from the queue. At this moment, any residual mission duration of the ongoing services (including the newly initiated one) must be stochastically greater than $\underline{S}$, and there are at most J ongoing services. Hence, the probability of the event $S_{\text{next}} > u$ is lower bounded by $(1 - \underline{G}(u))^J$. The last equality follows from the property of Poisson distribution.

For each $r > 0$, the prelimit of the right-hand side of (17) can be bounded from above as follows

$$\begin{aligned}
 & \Pr(N_q(r+u) \geq k, B^i(r) = 1, N_q(r) < k) \\
 &= \sum_{j=0}^{k-1} \Pr(N_q(r+u) \geq k \mid B^i(r) = 1, N_q(r) = j) P[B^i(r) = 1, N_q(r) = j] \\
 &\leq \sum_{j=0}^{k-1} \sum_{i=k-j}^{\infty} e^{-\lambda u} \frac{(\lambda u)^i}{i!} P[B^i(r) = 1, N_q(r) = j] \\
 &= e^{-\lambda u} \sum_{j=0}^{k-1} \left(e^{\lambda u} - \sum_{i=0}^{k-j-1} \frac{(\lambda u)^i}{i!} \right) P[B^i(r) = 1, N_q(r) = j] \\
 &\leq e^{-\lambda u} \sum_{j=0}^{k-1} (e^{\lambda u} - 1) P[B^i(r) = 1, N_q(r) = j],
 \end{aligned} \tag{19}$$

where the first inequality follows by comparison with the event of having at least $k-j$ arrivals, and the last inequality is obtained by discarding all $i > 0$ terms in the summation. Combining the lower bound (18) and the upper bound (19) and sending $r \rightarrow \infty$, we have for any $k \geq 1$ and $u > 0$:

$$\lim_{r \rightarrow \infty} P[B^i(r) = 1, N_q(r) = k] \leq \frac{e^{\lambda u} - 1}{\bar{G}(u)(1 - \underline{G}(u))^J} \sum_{j=0}^{k-1} \lim_{r \rightarrow \infty} P[B^i(r) = 1, N_q(r) = j].$$

By Poisson Arrival See Time Average (Tijms 2003), it holds that $p_k^i = \lim_{r \rightarrow \infty} P[B^i(r) = 1, N_q(r) = k]$ for $k \in \mathbb{N}_0$.

Using this relation and dividing $p_{\mathcal{B}}^i$ in both sides of the above display yields

$$q_k^i \leq \frac{e^{\lambda u} - 1}{\bar{G}(u)(1 - \underline{G}(u))^J} \sum_{j=0}^{k-1} q_j^i = C^\lambda \sum_{j=0}^{k-1} q_j^i.$$

We next prove that $q_k^i \leq C^\lambda (C^\lambda + 1)^{k-1} q_0^i$ for all $k \in \mathbb{N}$ by induction. Firstly, for $k = 1$, it holds that $q_k^i \leq C^\lambda q_0^i$ for all $i \in [I]$. We further assume that it holds for all $k = 1, \dots, n$ that $q_k^i \leq C^\lambda (C^\lambda + 1)^{k-1} q_0^i$. Then for $k = n+1$:

$$q_{n+1}^i \leq C^\lambda \sum_{j=0}^n q_j^i \leq C^\lambda \left(q_0^i + \sum_{j=1}^n C^\lambda (C^\lambda + 1)^{j-1} q_0^i \right) = C^\lambda (C^\lambda + 1)^n q_0^i,$$

where the last equality is from the geometric series formula. Therefore, we have

$$q_k^i \leq C^\lambda (C^\lambda + 1)^{k-1} q_0^i$$

for all $k \geq 1$ and $i \in [I]$. □

Appendix C: Design of ESSs with Collaborative Servers

This part demonstrates the use of Theorem 2 in location-allocation problems for ESSs with collaborative servers. As before, the arrival of calls from demand node $i \in [I]$ follows a Poisson process with rate $\pi_i \lambda$, with $\pi_i \geq 0$ and $\sum_{i \in [I]} \pi_i = 1$. There are $J \geq 2$ servers. We assume that the mission duration for server $j \in [J]$ to service a call from node $i \in [I]$ follows an exponential distribution. In the collaborative setting, each demand node $i \in [I]$ is serviced by a set of servers $\mathcal{J}_i \subseteq [J]$. An arriving call from node i does not need to enter the waiting line if any server $j \in \mathcal{J}_i$ is idle. We use the term “waiting line” interchangeably with “queue”. If multiple servers are idle, one of the server $j \in \mathcal{J}_i$ is dispatched to serve the call based on predefined probabilities that sum to one, where the probability to dispatch a busy server is set to zero. Hence, these dispatching probabilities are dependent on system states (which we formally introduce in §C.1). This is a generalization of the commonly adopted dispatching rules in the existing literature. For instance, dispatching based on a predetermined preference list as in Larson (1974) aligns with this framework by assigning probabilities of one to the designated server and zero to others. Similarly, the inverse distance weighting (IDW) rule (Burrough et al. 2015, Section 5.4) conforms by making the dispatching probability to an idle server from $j \in \mathcal{J}_i$ proportional to the inverse of the distance—or more generally, a decreasing function of the distance—between the demand node i and the server j ’s base location. Finally, the first call in the waiting line, assuming it is from node i , will be addressed by the first server in $\mathcal{J}_i$ that completes its current mission. Taking the dispatching probability as given, the problem is to optimize the base station locations of the J servers with the objective of minimizing the long-run average performance loss. We adhere to the notation in §3, where $\mathbf{y} \in \mathcal{Y}$ represents a feasible location scheme with servers located at occupied bases $m_1(\mathbf{y}), \dots, m_J(\mathbf{y}) \in \mathcal{O}(\mathbf{y})$. We will omit the dependence on $\mathbf{y}$ when there is no confusion. Unlike single-server spatial queues, multi-server systems with diverse demand and server classes lack closed-form busy probabilities, necessitating a model to calculate limiting probabilities for any specified location scheme $\mathbf{y}$.

To this end, we introduce the stacked hypercube queueing model that accounts for varying service times among different server-demand pairs, which extends the original model in Larson (1974). Our hypercube model aims to calculate the long-run average performance loss associated with a given location scheme $\mathbf{y}$ based on two approximations. First, we *approximate the infinite waiting line capacity of the actual system by a finite capacity of one call*. This approximation is motivated by our findings in Theorems 1 and 2 that in an ESS under light traffic conditions, the performance loss for queued calls mainly stems from the only call in queue. It is also supported by our numerical study on a lift-trap rescue system under light traffic conditions in §C.4, where we found that the simulated fraction of calls that arrive to find the queue length greater than 1, given that all servers are busy, is less than 0.03%. We wish to point out that it is possible to replace the capacity of 1 with other finite values, albeit at the cost of an increase in the state space and a complication of the system dynamics. This finite-capacity approximation allows us to compute the stationary distribution of the system states associated with the given location scheme $\mathbf{y}$. Second, when using this stationary distribution to compute the long-run average performance loss, we approximate the expected performance loss of queued calls by a constant $\hat{\gamma}_{\mathcal{B}}$ that is estimated from the simulation. This second approximation is supported by our asymptotic Theorem 2. Details for this approximation will become clear in §C.1.2.

In §C.1, we introduce stacked hypercube model and outline how we employ the model in conjunction with Theorem 2 to evaluate the performance loss for a specific location scheme $\mathbf{y}$. This evaluation enables the optimization of the location scheme, which is discussed in §C.2. In §C.3, we discuss an approach to reduce the size of the hypercube state space when $\mathcal{J}_i = [J]$ for all $i \in [I]$, where all the J servers collaboratively serve the whole ESS. Finally, we apply the methodology to a lift-trap rescue system in §C.4.

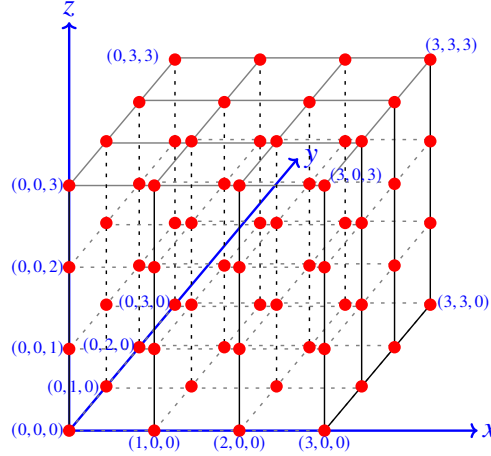


Figure 9 State space of a hypercube queueing model when the number of servers $J = 2$ and the number of demand nodes $I = 3$. The coordinate of each vertex in each unit hypercube (i_2, i_1, i_0) represents a system state, in which servers 1,2 are busy with demand node i_1, i_2 , respectively; and a call from node i_0 is waiting in queue.

C.1. Performance Evaluation: Stacked Hypercube Queueing Model

Throughout this section, we focus on evaluating the long-run average performance loss of a fixed location scheme $\mathbf{y}$ with base locations $\mathcal{O}(\mathbf{y}) = \{m_1, \dots, m_J\}$. For simplicity, we refer to the server based at m_j as server $j \in [J]$, and the waiting line as server 0 in the following. The exponential mission durations and Poisson arrival process enable us to develop a continuous-time Markov chain (CTMC) to model the spatial queueing system. We denote s_j as the state of server j , i.e., the server based at $m_j \in \mathcal{O}(\mathbf{y})$, and s_0 as the state of the waiting line. Specifically, we let $s_j = 0$ if server j is idle at its base, and $s_j = i, i \in [I]$ if it is serving demand node i . We let $s_0 = 0$ if the queue is empty, and $s_0 = i, i \in [I]$ if a call from node i is waiting in queue. In line with the notation system in [Larson \(1974\)](#), the state space of this spatial queue is denoted $\mathcal{S} \triangleq \{s = (s_J, \dots, s_1, s_0) \mid s_J, \dots, s_1, s_0 \in \{0\} \cup [I]\}$. There are $(I+1)^{J+1}$ states in total. Each state $s \in \mathcal{S}$ can be regarded as a base- $(I+1)$ number with $J+1$ bits and digits from 0 to I . Similar to Larson's hypercube queueing model, it can be coded using a decimal number $v(s)$ as

$$v(s) = s_J(I+1)^J + s_{J-1}(I+1)^{J-1} + \dots + s_1(I+1) + s_0.$$

There is a one-to-one correspondence between each state $s \in \mathcal{S}$ and decimal number $v \in \Omega \triangleq \{0, 1, \dots, (I+1)^{J+1} - 1\}$. As such, we will use $s(v) = (s_J(v), \dots, s_1(v), s_0(v))$ and $v(s)$ to denote the same state interchangeably. The classic hypercube queueing model deals with binary states for each server ([Larson 1974](#)), in which the system states can be represented by the collection of all vertices of a unit hypercube in the positive quadrant of the m -dimensional space. In comparison, the state of each server in our model takes integer values from 0 to I , and the system state space corresponds to the vertices of $I^{(J+1)}$ unit hypercubes stacked in the positive quadrant of the $(J+1)$ -dimensional space. A schematic of such correspondence is illustrated in [Figure 9](#) with $J = 2$ and $I = 3$, in which red dots represent system states. We thus call the model a *stacked hypercube queueing* model in this paper.

C.1.1. Transition rates and limiting distribution of the CTMC. To obtain the long-run average performance loss of the system for the given $\mathbf{y}$, we next derive the limiting distribution of the CTMC. More notation is needed for this purpose. We let $\hat{\mathcal{J}}_i \triangleq \mathcal{J}_i \cup \{0\}$, such that a call from node i can enter the queueing system if the state of

any server $j \in \hat{\mathcal{J}}_i$ is 0. We also define $\mathcal{I}_j = \{i \in [I] : j \in \mathcal{J}_i\}$ as the set of demand nodes that server $j \in [J]$ is allowed to serve. At a time point, consider that a call from node i arrives at the system. For each state $\nu \in \Omega$, let $\mathbb{B}_i(\nu) \subseteq \mathcal{J}_i$ be the collection of busy servers in $\mathcal{J}_i$ under state ν , and $\bar{\mathbb{B}}_i(\nu) \triangleq \mathcal{J}_i - \mathbb{B}_i(\nu)$ be the collection of idle servers in $\mathcal{J}_i$. For each state ν in which at least one server eligible to serve node i is idle, i.e., $\bar{\mathbb{B}}_i(\nu) \neq \emptyset$, let D_{i,m_j}^ν denote the dispatching probability that a call from node $i \in [I]$ is served by server j in state ν , which satisfies $D_{i,0}^\nu = 0$, $D_{i,m_j}^\nu = 0$ for $j \notin \bar{\mathbb{B}}_i(\nu)$, $D_{i,m_j}^\nu \geq 0$ for $j \in \bar{\mathbb{B}}_i(\nu)$, and $\sum_{j \in \bar{\mathbb{B}}_i(\nu)} D_{i,m_j}^\nu = 1$. For state ν where $\bar{\mathbb{B}}_i(\nu) = \emptyset$, we set $D_{i,m_j}^\nu = 0$ for all $j \in [J]$ and set $D_{i,0}^\nu = \mathbb{I}(s_0(\nu) = 0)$. The service rate of server $j \in \mathcal{J}_i$ serving the call is given by $\mu_{i,m_j} \triangleq 1/\mathbb{E}S_{im_j}$, where S_{im_j} is the total mission duration.

With the above set-up, we are in a position to derive the transition rates Q_{ν_1, ν_2} for $\nu_1, \nu_2 \in \Omega$ of the CTMC. Align with [Larson \(1974\)](#), we categorize the transition of the CTMC into two types based on the directional change in the number of calls within the system: (i) upward transitions, where the number of calls increases by one, and (ii) downward transitions, where it decreases by one. We now write the transition probabilities.

Downward transition. A downward transition happens when a busy server completes its on-hand service. Consider a state ν_1 where server j is busy at node $s_j(\nu_1)$. Upon completing its service, server j either becomes idle if there are no queued calls it can respond to, i.e., $s_0(\nu_1) \notin \mathcal{I}_j$, or it immediately turn to serve the queued call if $s_0(\nu_1) \in \mathcal{I}_j$. In the former case, the server's state changes from $s_j(\nu_1)$ to 0. In the latter, server j transitions to $s_0(\nu_1)$, and the state of the waiting line resets to 0. This modifies the system's state to $\nu_2 = \nu_1 - s_j(\nu_1) \times (I+1)^j + \mathbb{I}(s_0(\nu_1) \in \mathcal{I}_j) \times s_0(\nu_1) \times [(I+1)^j - 1]$. Therefore, for any $\nu_1 \in \{\nu \in \Omega : s_j(\nu) \neq 0\}$, the downward-transition rate from ν_1 to ν_2 is

$$q_{\nu_1, \nu_2} = \mu_{s_j(\nu_1), m_j}. \quad (20)$$

Upward transition. An upward transition happens when an arrival call enters the system. Specifically, for state ν_1 , a call from node i can enter the system if any sever $j \in \mathcal{J}_i$ is idle ($s_j(\nu_1) = 0$), or the waiting line is empty ($s_0(\nu_1) = 0$). The call will then be responded to by server j with probability $D_{i,m_j}^{\nu_1}$, or enter the waiting line with probability $D_{i,0}^{\nu_1}$ as we defined before. Consequently, the j -th bit of ν_1 changes from 0 to i for $j \in \hat{\mathcal{J}}_i$, transforming the state to $\nu_2(i) = \nu_1 + i \times (I+1)^j$ for $i \in [I]$. Therefore, the upward-transition rate from a state ν_1 , where $s_j(\nu_1) = 0$ for $j \in \hat{\mathcal{J}}_i, i \in [I]$, to $\nu_2(i) = \nu_1 + i \times (I+1)^j$ is

$$q_{\nu_1, \nu_2(i)} = \pi_i \lambda D_{i,m_j}^{\nu_1}. \quad (21)$$

The transition rate matrix $\mathbf{Q} = [q_{\nu_1, \nu_2}]_{(I+1)^J \times (I+1)^J}$ of the CTMC is the collection of the transition rates above, where we have defined the diagonal elements $q_{\nu, \nu}$ by $q_{\nu, \nu} := -\sum_{\nu' \neq \nu} q_{\nu, \nu'}$ for $\nu \in \Omega$. The limiting distribution $\mathbf{p} := (p_\nu)_{\nu \in \Omega}$ of the CTMC can then be obtained by solving $\mathbf{p}\mathbf{Q} = \mathbf{0}$, where $\mathbf{p}$ is interpreted as a row vector.

Power iteration algorithm. We next describe an approach for fast solution of the equation $\mathbf{p}\mathbf{Q} = \mathbf{0}$. Let $q_{\max} := \max\{|q_{\nu, \nu'}| : \nu \in \Omega\}$ denote the maximum transition rate out of a state. Using $q_{\max}$ as the uniformization constant, the transition probability matrix $\mathbf{Q}_D$ of the uniformized discrete-time Markov chain of the CTMC $\mathbf{Q}$ is given by $\mathbf{Q}_D := (\mathbf{Q} + q_{\max}\mathbf{I})/q_{\max}$, where $\mathbf{I}$ represents the identity matrix. It can be verified that $\mathbf{Q}_D$ is a nonnegative square matrix with each row summing to 1, and that the equation $\mathbf{p}\mathbf{Q} = \mathbf{0}$ is equivalent to the equation $\mathbf{p}\mathbf{Q}_D = \mathbf{p}$. Hence, the limiting distribution $\mathbf{p}$ can be viewed as the eigenvector of the matrix $\mathbf{Q}_D$ corresponding to the eigenvalue 1. Since $\mathbf{Q}_D$ is a very sparse matrix with at most $J(I+1)$ out of $(I+1)^J$ nonzero entries in each row, an efficient algorithm to solve this eigenvector $\mathbf{p}$ is the power iteration algorithm. This algorithm initializes $\mathbf{p}_0$ as the uniform distribution over Ω , and then iteratively updates $\mathbf{p}_n$ using $\mathbf{p}_n = \mathbf{Q}_D \mathbf{p}_{n-1}$ until $\|\mathbf{p}_n - \mathbf{p}_{n-1}\|$ is small enough. In implementation, we propose using the infinity norm and a stopping criterion of $\|\mathbf{p}_n - \mathbf{p}_{n-1}\|_\infty \leq 0.001$.

C.1.2. Long-run average performance loss for given base locations. Having had the limiting distribution of the given ESS with collaborative response and J servers based at $m_1, m_2, \dots, m_J$, we derive its long-run average performance loss of calls in this section. When a call arrive from node i to find at least one server in $\mathcal{J}_i$ idle, i.e., for state ν in which $\bar{\mathbb{B}}_i(\nu) \neq \emptyset$, the expected performance loss of the call is $\mathbb{E}[h(S_{\text{tr},im_j})]$ if the server $j \in \mathcal{J}_i$ located at m_j is dispatched. On the other hand, when a call from node i finds all servers in $\mathcal{J}_i$ busy and the waiting line empty, i.e., for state ν in which $\bar{\mathbb{B}}_i(\nu) = \emptyset$ and $s_0(\nu) = 0$, the call enters the queue and its response time is the residual time of server j 's ongoing mission plus the travel time S_{tr,im_j} , suppose that server j is the server among $\mathcal{J}_i$ to first complete its on-hand mission. Under exponential mission duration assumption, the residual time to the first mission completion follows an exponential distribution with rate $\mu_i(\nu) \triangleq \sum_{j \in \mathcal{J}_i} \mu_{s_j(\nu),m_j}$, and the probability that server j is the first server among $\mathcal{J}_i$ to complete its mission is $\mu_{s_j(\nu),m_j} / \mu_i(\nu)$. Therefore, the expected loss for a queued call from node i in state ν is

$$\gamma_i(\nu) = \sum_{j \in \mathcal{J}_i} \frac{\mu_{s_j(\nu),m_j}}{\mu_i(\nu)} \int_0^\infty \mu_i(\nu) e^{-\mu_i(\nu)t} \mathbb{E}[h(S_{\text{tr},im_j} + t)] dt.$$

Putting together the above pieces, the long-run average performance loss for the ESS with servers based at $O(\mathbf{y}) = \{m_1, m_2, \dots, m_J\}$ can be calculated as

$$\sum_{i \in [I]} \pi_i \left[\sum_{\nu: \bar{\mathbb{B}}_i(\nu) \neq \emptyset} p_\nu \sum_{j: j \in \bar{\mathbb{B}}_i(\nu)} D_{i,m_j}^\nu a_{i,m_j} + \sum_{\nu: \bar{\mathbb{B}}_i(\nu) = \emptyset, s_0(\nu) = 0} p_\nu \gamma_i(\nu) \right].$$

The expression of each $\gamma_i(\nu)$ in the above display is very complex, involving $|\mathcal{J}_i|$ convolutional computations. Additionally, for a given location scheme $\mathbf{y}$, we need to calculate $\gamma_i(\nu)$ for each $i \in [I]$ and each ν where $\bar{\mathbb{B}}_i(\nu) = \emptyset, s_0(\nu) = 0$. Given this computation burden, Theorem 2 is essential for approximating the average loss for each queued call. Specifically, it motivates us to estimate $\gamma_{\mathcal{B}}(0+)$ by simulating the performance loss of queued calls under a fixed location scheme in light traffic situations. Let $\hat{\gamma}_{\mathcal{B}}$ denote this estimate. Then, we propose approximating the weighted average $\sum_{\nu: \bar{\mathbb{B}}_i(\nu) = \emptyset, s_0(\nu) = 0} p_\nu \gamma_i(\nu)$, normalized by the total weight $\sum_{\nu: \bar{\mathbb{B}}_i(\nu) = \emptyset, s_0(\nu) = 0} p_\nu$, by this $\hat{\gamma}_{\mathcal{B}}$, and thus avoids the intensive convolutional computations. Such an approximation is reasonable because the above normalized weighted average is exactly the expected performance loss of queued calls. Consequently, the long-run average performance loss of calls for the ESS with servers based at $O(\mathbf{y}) = \{m_1, m_2, \dots, m_J\}$ can be approximated as

$$\sum_{i \in [I]} \pi_i \left[\sum_{\nu: \bar{\mathbb{B}}_i(\nu) \neq \emptyset} p_\nu \sum_{j: j \in \bar{\mathbb{B}}_i(\nu)} D_{i,m_j}^\nu a_{i,m_j} + \hat{\gamma}_{\mathcal{B}} \sum_{\nu: \bar{\mathbb{B}}_i(\nu) = \emptyset, s_0(\nu) = 0} p_\nu \right], \quad (22)$$

where we have defined a_{i,m_j} by $a_{i,m_j} = \mathbb{E}[h(S_{\text{tr},im_j})]$.

C.2. Performance Optimization

The long-run average performance loss of the stacked hypercube queueing model above is based on a given feasible solution $\mathbf{y}$ with base locations $O(\mathbf{y}) = \{m_1, m_2, \dots, m_J\}$. This average performance loss is a function of $\mathbf{y}$, because the dispatching probabilities D_{i,m_j}^ν , service rates μ_{i,m_j} , and parameters a_{i,m_j} change with $\mathbf{y}$. The optimal location decision variable $\mathbf{y}$ can be obtained by solving the following program combined with the stacked hypercube queueing model:

$$\min_{\mathbf{y}} C(\mathbf{y}) \quad (23a)$$

$$\text{s.t. } C(\mathbf{y}) = \sum_{i \in [I]} \pi_i \left[\sum_{\nu: \bar{\mathbb{B}}_i(\nu) \neq \emptyset} p_\nu(\mathbf{y}) \sum_{j: j \in \bar{\mathbb{B}}_i(\nu)} D_{i,m_j}^\nu(\mathbf{y}) a_{i,m_j}(\mathbf{y}) + \hat{\gamma}_{\mathcal{B}} \sum_{\nu: \bar{\mathbb{B}}_i(\nu) = \emptyset, s_0(\nu) = 0} p_\nu(\mathbf{y}) \right]; \quad (23b)$$

$$m_j = \arg \min_{m \in [M]} \left\{ \sum_{m'=1}^m y_{m'} \geq j \right\} \quad j \in [J]; \quad (23c)$$

$$\sum_{m \in [M]} y_m = J; \quad (23d)$$

$$y_m \in \{0, 1\} \quad m \in [M], \quad (23e)$$

where the stationary distribution $p_\nu(\mathbf{y})$ is established from (20)-(21).

The above program is obviously highly nonlinear and even non-convex, which can hardly be solved directly. One way to solve the program is to enumerate every feasible solution $\mathbf{y}$ such that $\sum_{m \in [M]} y_m = J$ to calculate the average performance loss for each call $C(\mathbf{y})$. This requires $\binom{M}{J}$ times of enumeration in total with computation cost $O(M^J)$. For our case study on the Singapore lift-trap rescue system, we have $J = 2$, so the enumeration can be done efficiently. Solutions with the lowest cost rate $C(\mathbf{y})$ are the optimal solutions, i.e.,

$$\mathbf{y}^* \in \arg \min_{\mathbf{y} \in \{y_j=0 \text{ or } 1, \sum_{j \in [M]} y_j = m\}} C(\mathbf{y}), \quad (24)$$

where $C(\mathbf{y})$ is determined by Equation (22).

C.3. State Space Reduction

When each of the J servers is authorized to service all the I demand nodes, i.e., $\mathcal{J}_i = [J]$ for all $i \in [I]$, the state space of the stacked hypercube model introduced in §C.1 can be reduced. This makes the stationary distribution $\mathbf{p}$ easier to solve. Now, we introduce the approach for state space reduction in this special case. Recall that in §C.1 we have used s_0 to track the status of the queue, where $s_0 = 0$ if the queue is empty and $s_0 = i, i \in [I]$ if a call from node i is waiting in queue. Under the condition that $\mathcal{J}_i = [J]$ for all $i \in [I]$, we next show that letting $s_0 \in \{0, 1\}$, which tracks the queue length (rather than the specific demand node of the awaiting call), will be sufficient to set up hypercube model. Hence, the total number of states in the hypercube model can be reduced by a factor of $(I+1)/2$. The key insights behind this reduction is that, under Poisson arrivals of call, the probability that a call waiting in the queue (if any) comes from a specific demand node $i \in [I]$ is equal to π_i . In the following, we use a similar framework as in §C.1.1 to derive the transition probabilities of the CTMC under this reduced state space.

Downward transition. A downward transition happens when a busy server completes its on-hand service. Consider a state ν_1 where server j is busy at node $s_j(\nu_1)$. Upon completing its service, server j either becomes idle if there are no queued calls (i.e., $s_0 = 0$), or it begins to address the queued call if $s_0 = 1$. In the former case, the server's state changes from $s_j(\nu_1)$ to 0. In the latter, server j 's state transitions to i with probability π_i for $i \in [I]$, and the state of the waiting line s_0 resets to 0. This modifies the system's state to $\nu_2(i) = \nu_1 - s_j(\nu_1) \times (I+1)^J + s_0(\nu_1) \times [i \times (I+1)^J - 1]$. Therefore, for any $\nu_1 \in \{\nu \in \Omega : s_j(\nu) \neq 0\}$, the downward-transition rate of the CTMC from ν_1 to $\nu_2(i)$ for $i \in [I]$ is

$$q_{\nu_1, \nu_2(i)} = \pi_i \lambda \mu_{s_j(\nu_1), m_j}.$$

Upward transition. An upward transition happens when an arrival call enters the queueing system. Specifically, for state ν_1 , a call from node i can enter the system if any sever $j \in [J]$ is idle ($s_j(\nu_1) = 0$), or the waiting line is

empty ($s_0(v_1) = 0$). The call will then be responded to by server j with probability $D_{im_j}^{v_1}$, or enter the waiting line with probability $D_{i,0}^{v_1}$ as we defined before. Consequently, the j -th bit of v_1 changes from 0 to i for $j \in [J] \cup \{0\}$, transforming the state to $v_2(i) = v_1 + i \times (I+1)^j$ for $i \in [I]$. Therefore, the upward-transition rate from a state v_1 , where $s_j(v_1) = 0$ for $j \in [J]$, to $v_2(i) = v_1 + i \times (I+1)^j$ is

$$q_{v_1, v_2(i)} = \pi_i \lambda D_{i, m_j}^{v_1}.$$

The rest of the procedure (i.e., performance evaluation and optimization) follows similarly from the previous subsections.

C.4. Application to the lift-trap system

Next, we demonstrate our methodology using the lift-trap system as a real-world example. Although in practice, the two rescue teams in the day shift operate under disjoint working mode, we study the case where the two servers work collaboratively to serve the whole planning area. We use the data obtained in §5.1 and assume all the service times are exponentially distributed. We set the travel time used in the hypercube model as a weighted average of the time-varying travel times. We set $\mathcal{J}_i = [J]$ for both servers and invoke the state space reduction techniques described in §C.3. The size of the state space for the resulting hypercube model is $(I+1)^2 \times 2 = (111)^2 \times 2 = 24,642$. We employ a static closest-server dispatching rule that always dispatches the nearest available server. Specifically, $D_{i, m_1}^v = 1$ if $\mathbb{E}S_{im_1} \leq \mathbb{E}S_{im_2}$ and $=0$ otherwise, for states v where both $s_1(v) = 0$ and $s_2(v) = 0$. As such, no allocation decision variable is needed in this scheme, in contrast to §5.3. Therefore, we focus on optimizing the base station locations for the two servers using the methodology we discussed in this section.

We first decide the value of the constant $\hat{\gamma}_{\mathcal{B}}$ by deriving the simulated average overage loss for queued call where the location scheme is set as the optimal locations derived from MOL-AsmApX for the disjoint scheme in §5.3, i.e., $\hat{\gamma}_{\mathcal{B}} = 0.552$. Given the modest number of feasible solutions (i.e., $\binom{M}{J} = \binom{43}{2} = 903$), we proceed to enumerate each pair of locations to calculate their steady-state distribution using our stacked hypercube model and compute their average overage loss as per Equation (22). The location scheme yielding the lowest average overage loss is identified as optimal, and is presented in Figure 10. It can be seen that the optimal server location under the collaborative working mode is the same as that of the disjoint service region derived in §5.3. We compare the performances of the optimized schemes

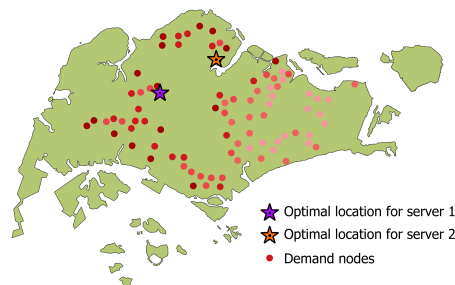


Figure 10 Optimal base locations for the two teams given by the proposed stacked hypercube model and the simulation. The two methods suggest the same optimal locations.

with the current schemes (assuming that the servers are collaborative) using the simulation described in §5.2. For this

comparison, we use 100 replications over 1 year each, discarding the first 10% of calls as warm-up. The results are summarized in Table 3. Compared to Table 2, it can be seen that the day-shift scheme optimized by our proposed model benefits from collaboration, reducing the overall performance loss from circa 30 to circa 25, whereas the current day-shift scheme does not benefit much, yielding a similar loss of circa 104.

Table 3 Comparison of simulated performance of the stacked hypercube model optimized scheme and the current scheme. The probabilities in the *Day* rows are the simulated probabilities where (i) both teams are idle, (ii) the first team is busy and the second is idle, (iii) the first team is idle and the second is busy, and (iv) both teams are busy. The probabilities in the *Night* and *Sum* rows indicate the simulated probabilities that a call enters the waiting line. Parentheses give the widths of 95% confidence intervals.

		Optimal Loc.	Total Loss	Queued Loss	Non-queued Loss	Probabilities
Stacked hypercube model	Day	16, 30	20.08(2.10)	0.39(0.26)	19.69(2.05)	.962(.0031), .027(.0025), .010(.0017), .0008(.0004)
	Night	16	4.87(0.83)	4.73(0.82)	0.14(0.14)	.034(.0041)
	Sum	–	24.95(2.32)	5.12(0.85)	19.83(2.08)	.011(.0013)
Current	Day	9,40	94.03(4.60)	0.59(0.32)	93.44(4.57)	.955(.0034), .014(.0017), .029(.0027), .001(.0006)
	Night	40	10.12(1.18)	5.88(1.01)	4.24(0.90)	.038(.0047)
	Sum	–	104.15(4.84)	6.47(1.07)	97.68(4.68)	.012(.0015)

Appendix D: Proof of Proposition 1

We first prove part (i) of the proposition. Consider the auxiliary System B' shown in Figure (11a), which is identical to System B^{NI} except that there are no node-2 call arrivals (i.e., $\lambda_2 = 0$). In System B', both server 1 and server 2 are dedicated to serving node-1 calls, where server 1 acts as the primary server and server 2 the backup. Let $\mathbb{E}[R_A^i]$, $\mathbb{E}[R_{B'}^i]$ and $\mathbb{E}[R_{B^{\text{NI}}}^i]$ denote the average response times for node- i calls for $i \in \{1, 2\}$ in Systems A, B' and B^{NI} , respectively. Our proof for Proposition 1(i) proceeds in the following two steps.

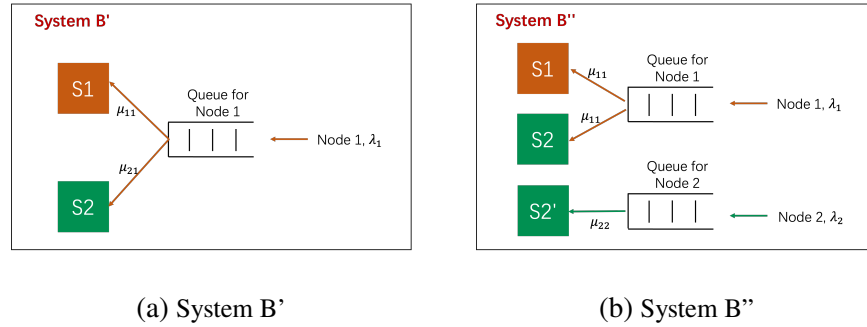


Figure 11 Auxiliary systems used in proof.

Step 1. Showing $\mathbb{E}[R_{B'}^1] \leq \mathbb{E}[R_{B^{\text{NI}}}^1]$ and $\mathbb{E}[R_A^2] \leq \mathbb{E}[R_{B^{\text{NI}}}^2]$. Let $N_A^i, N_{B^{\text{NI}}}^i$ and $N_{B'}^i$ denote the number of node- i calls for $i \in \{1, 2\}$ in Systems A, B^{NI} and B', respectively. The evolutions of $N_{B^{\text{NI}}}^1$ and $N_{B'}^1$ can be modeled as birth-and-death processes, in which both of them have identical birth rates. However, the death rate of $N_{B^{\text{NI}}}^1$ is no larger than that of $N_{B'}^1$, for all system states, because in System B^{NI} , server 2 can be occupied with a node-2 job while there are node-1 jobs in queue. By Theorem 5 in Smith and Whitt (1981), this implies that $N_{B'}^1 \leq_{s.t.} N_{B^{\text{NI}}}^1$, and thus $\mathbb{E}[N_{B'}^1] \leq \mathbb{E}[N_{B^{\text{NI}}}^1]$.

Applying Little's Law and noting that the on-scene service times are identical in both systems, we conclude that $\mathbb{E}[R_{B'}^1] \leq \mathbb{E}[R_{B^{\text{NI}}}^1]$. Similarly, modeling the evolutions of N_A^2 and $N_{B^{\text{NI}}}^2$ as birth-and-death processes, and applying Theorem 5 in [Smith and Whitt \(1981\)](#), we again conclude that $N_A^2 \leq_{s.t.} N_{B^{\text{NI}}}^2$. This implies $\mathbb{E}[R_A^2] \leq \mathbb{E}[R_{B^{\text{NI}}}^2]$ due to Little's Law and identical on-scene service times.

Step 2. Showing $\mathbb{E}[R_A^1] \leq \mathbb{E}[R_{B'}^1]$ under slow-server conditions. Next, we compare the average response time of node-1 calls in Systems A and B'—i.e., $\mathbb{E}[R_A^1]$ and $\mathbb{E}[R_{B'}^1]$. Astute readers may recognize that this constitutes a “slower server problem”, in which server 2 acts as the “slow server” in System B' ([Rubinovitch 1985](#)). Let $\mathbb{E}[T_A^1]$ and $\mathbb{E}[T_{B'}^1]$ denote the average sojourn time of node-1 calls in Systems A and B', respectively. By applying the results in Section 3 of [Rubinovitch \(1985\)](#), it holds that $\mathbb{E}[T_A^1] \leq \mathbb{E}[T_{B'}^1]$ if the following slow-server conditions hold:

$$r := \mu_{21}/\mu_{11} < 0.5 \quad \text{and} \quad \frac{\lambda_1}{\mu_{11} + \mu_{21}} \leq \frac{2 + r^2 - \sqrt{(2 + r^2)^2 + 4(1 + r^2)(2r - 1)(1 + r)}}{2(1 + r^2)}. \quad (25)$$

Since the on-scene service times are the same, it follows that $\mathbb{E}[R_A^1] \leq \mathbb{E}[R_{B'}^1]$ if (25) holds.

Steps 1 and 2 above together show that $\mathbb{E}[R_A^1] \leq \mathbb{E}[R_{B'}^1] \leq \mathbb{E}[R_{B^{\text{NI}}}^1]$ and $\mathbb{E}[R_A^2] \leq \mathbb{E}[R_{B^{\text{NI}}}^2]$ whenever (25) holds. This establishes Proposition 1(i) because the average response times over all customers in Systems A and B^{NI} are given by:

$$\mathbb{E}[R_A] = \frac{\lambda_1 \mathbb{E}[R_A^1]}{\lambda_1 + \lambda_2} + \frac{\lambda_2 \mathbb{E}[R_A^2]}{\lambda_1 + \lambda_2} \leq \frac{\lambda_1 \mathbb{E}[R_{B^{\text{NI}}}^1]}{\lambda_1 + \lambda_2} + \frac{\lambda_2 \mathbb{E}[R_{B^{\text{NI}}}^2]}{\lambda_1 + \lambda_2} = \mathbb{E}[R_{B^{\text{NI}}}] .$$

We now prove part (ii) of the proposition. Consider the auxiliary System B'' shown in Figure (11b). In System B'', server 1 and server 2 collaboratively serve node-1 calls, both operating with exponential service rates μ_{11} ; an additional independent server, server 2', exclusively serves node-2 calls with exponential rate μ_{22} . In effect, the node-1 subsystem forms an $M/M/2$ queue, and the node-2 subsystem forms an $M/M/1$ queue. With the additional server 2' and the faster service rate of server 2 to node-1 calls ($\mu_{21} \leq \mu_{11}$), it should be clear that the average response time of calls in System B'', denoted by $\mathbb{E}[R_{B''}]$, is no greater than that in System B^{NI} —i.e., $\mathbb{E}[R_{B''}] \leq \mathbb{E}[R_{B^{\text{NI}}}]$.

Based on standard results for $M/M/c$ queues (see §7.3 in [Kulkarni \(2016\)](#)),

$$\mathbb{E}[R_{B''}] = \frac{\lambda_1}{\lambda_1 + \lambda_2} \cdot \frac{4\mu_{11}}{4\mu_{11}^2 - \lambda_1^2} + \frac{\lambda_2}{\lambda_1 + \lambda_2} \cdot \frac{1}{\mu_{22} - \lambda_2} - \mathbb{E}[S_{\text{os}}]$$

where $\mathbb{E}[S_{\text{os}}]$ is the average on-scene service time. In System A, the average response time is:

$$\mathbb{E}[R_A] = \frac{\lambda_1}{\lambda_1 + \lambda_2} \cdot \frac{1}{\mu_{11} - \lambda_1} + \frac{\lambda_2}{\lambda_1 + \lambda_2} \cdot \frac{1}{\mu_{22} - \lambda_2} - \mathbb{E}[S_{\text{os}}] .$$

Therefore,

$$\mathbb{E}[R_A] - \mathbb{E}[R_{B^{\text{NI}}}] \leq \mathbb{E}[R_A] - \mathbb{E}[R_{B''}] = \frac{\lambda_1}{\lambda_1 + \lambda_2} \cdot \left(\frac{1}{\mu_{11} - \lambda_1} - \frac{4\mu_{11}}{4\mu_{11}^2 - \lambda_1^2} \right) = \frac{\lambda_1}{\lambda_1 + \lambda_2} \cdot \frac{\lambda_1(4\mu_{11} - \lambda_1)}{(4\mu_{11}^2 - \lambda_1^2)(\mu_{11} - \lambda_1)} .$$

Taking the limit as $\lambda_1 \rightarrow 0$, the RHS of the above bound approaches zero. If $\lambda_2 > 0$, then $\mathbb{E}[R_A]$ converges to the positive constant $\frac{1}{\mu_{22} - \lambda_2} - \mathbb{E}[S_{\text{os}}]$; if $\lambda_2 = 0$, then $\mathbb{E}[R_A]$ converges to the positive constant $\frac{1}{\mu_{11}} - \mathbb{E}[S_{\text{os}}] = \mathbb{E}[S_{\text{tr},11}]$.

Therefore,

$$\limsup_{\lambda_1 \rightarrow 0} \frac{\mathbb{E}[R_A] - \mathbb{E}[R_{B^{\text{NI}}}]}{\mathbb{E}[R_A]} \leq 0 .$$

We finally prove part (iii). Because the optimal dispatching rule in System $B^\star$ may keep server 2 idle when a node-1 call is waiting, it can reproduce the disjoint dispatching rule in System A by never assigning server 2 to node-1 calls. Therefore,

$$\mathbb{E}[R_{B^\star}] \leq \mathbb{E}[R_A]. \quad (26)$$

Moreover, when $\lambda_2 > 0$, the same coupling argument as in Step 1 gives

$$\mathbb{E}[R_A^2] \leq \mathbb{E}[R_{B^\star}^2], \quad (27)$$

because node-2 calls can be served only by server 2 in both systems, while System $B^\star$ may also assign server 2 to node-1 calls.

For a node-1 call in System $B^\star$, its response time is the sum of its nonnegative waiting time and the travel time of the dispatched server. Because server 1 is closer to node 1 than server 2 and the dispatching rule is nonanticipative, the average response time of node-1 calls under any dispatching rule is bounded below by the expected travel time from server 1 to node 1. Hence,

$$\mathbb{E}[R_{B^\star}^1] \geq \mathbb{E}[S_{tr,11}] = \frac{1}{\mu_{11}} - \mathbb{E}[S_{os}]. \quad (28)$$

In System A, the node-1 subsystem is an M/M/1 queue, so

$$\mathbb{E}[R_A^1] = \frac{1}{\mu_{11} - \lambda_1} - \mathbb{E}[S_{os}]. \quad (29)$$

Combining (27)–(29), when $\lambda_2 > 0$ we obtain

$$\begin{aligned} \mathbb{E}[R_A] - \mathbb{E}[R_{B^\star}] &\leq \frac{\lambda_1}{\lambda_1 + \lambda_2} \left(\mathbb{E}[R_A^1] - \mathbb{E}[S_{tr,11}] \right) \\ &= \frac{\lambda_1^2}{(\lambda_1 + \lambda_2)\mu_{11}(\mu_{11} - \lambda_1)}. \end{aligned}$$

When $\lambda_2 = 0$, the average response time is simply the node-1 average response time, and (28)–(29) give the same bound, with $\lambda_1/(\lambda_1 + \lambda_2) = 1$. Together with (26), this proves

$$0 \leq \mathbb{E}[R_A] - \mathbb{E}[R_{B^\star}] \leq \frac{\lambda_1^2}{(\lambda_1 + \lambda_2)\mu_{11}(\mu_{11} - \lambda_1)}. \quad (30)$$

It remains to establish the relative comparison. If $\lambda_2 > 0$ is fixed, the upper bound in (30) is $O(\lambda_1^2)$, while

$$\mathbb{E}[R_A] \longrightarrow \frac{1}{\mu_{22} - \lambda_2} - \mathbb{E}[S_{os}] > 0.$$

If $\lambda_2 = 0$, the upper bound is $O(\lambda_1)$, whereas

$$\mathbb{E}[R_A] \longrightarrow \frac{1}{\mu_{11}} - \mathbb{E}[S_{os}] = \mathbb{E}[S_{tr,11}] > 0.$$

Therefore, in either case, (26) and (30) imply

$$\lim_{\lambda_1 \downarrow 0} \frac{\mathbb{E}[R_A] - \mathbb{E}[R_{B^\star}]}{\mathbb{E}[R_A]} = 0,$$

which completes the proof of part (iii).

Appendix E: Supplementary materials to §3

E.1. Numerical investigation: Allowing multiple servers on one base

Next, we conduct small-scale experiments that permit server co-location using the synthesized systems in §4, and we use brute-force enumeration to identify optimal location-allocation schemes. Starting from the three demand scenarios in Figure 2 (each with $I = 80$ demand nodes and distinct demand ratios $(\pi_i)_{i \in [I]}$), we consider locating $J = 2$ servers, with an objective to minimize the average response time (i.e., $h(x) = x$). To keep enumeration tractable, we downsample to $I' = 10$ demand nodes chosen at even spacings from $\{1, 2, \dots, 80\}$ and re-normalize (π_i) to sum to one. We vary the total arrival rate λ over three levels—namely, $0.05J/\bar{\tau}$, $0.075J/\bar{\tau}$, $0.1J/\bar{\tau}$ —to probe different traffic intensities, where $\bar{\tau}$ is the expected mission duration averaged over the 20×10 selected base–node pairings. For each (scenario, λ) pair, we enumerate all feasible location-allocation schemes (including co-location) and evaluate each scheme’s mean response time. When no co-location occurs, each server operates an independent $M/G/1$ queue and mean response times are computed via the Pollaczek–Khinchine formula using the primitives in §4.1.1; when co-location occurs, the resulting two-server station is evaluated by discrete-event simulation using 100 replications of 10^6 arrivals per base, discarding the first 10% of arrivals, under the lognormal mission-duration setting.

Figure 12 compares the average response times of the best co-located and non-co-located designs across all nine (scenario, λ) pairs. In every setting, the optimal non-co-located design achieves a shorter average response time than its co-located counterpart, indicating that the best-performing scheme never co-locates servers. Intuitively, under light-traffic conditions, the queueing benefit from pooling two servers at the same base is small, while the increase in travel time from concentrating capacity dominates, so placing servers at distinct bases yields shorter response times.

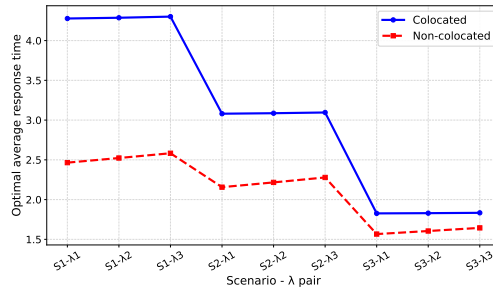


Figure 12 Comparison of average response times between the best co-located and non-co-located designs across all (Scenario, λ) pairs. Non-co-located designs consistently outperform co-located ones. The widths of the 95% confidence intervals for the plotted co-located simulation means are all below 0.0015.

E.2. Proof of Lemma 1

By the definition of $C_m^{\text{RT}} = (\sum_{i \in \mathcal{R}_m} \pi_i) \hat{\gamma}_m^{\text{RT}}$, we have

$$\begin{aligned}
 C_m^{\text{RT}} &= \left(\sum_{i \in \mathcal{R}_m} \pi_i \right) \hat{\gamma}_m^{\text{RT}} \\
 &\stackrel{(a)}{=} (1 - \rho_m) \sum_{i \in \mathcal{R}_m} \pi_i \mathbb{E}[S_{\text{tr},im}] + \rho_m \sum_{i \in \mathcal{R}_m} \pi_i \mathbb{E}[L_m + S_{\text{tr},im}]
 \end{aligned}$$

$$\begin{aligned}
&\stackrel{(b)}{=} \sum_{i \in \mathcal{R}_m} \pi_i \mathbb{E}[S_{\text{tr},im}] + \rho_m \sum_{i \in \mathcal{R}_m} \pi_i \mathbb{E}[L_m] \\
&\stackrel{(c)}{=} \sum_{i \in \mathcal{R}_m} \pi_i \mathbb{E}[S_{\text{tr},im}] + \rho_m \sum_{i \in \mathcal{R}_m} \pi_i \int_0^\infty u \frac{\sum_{i' \in \mathcal{R}_m} \pi_{i'} (1 - G_{im}(u))}{\sum_{i' \in \mathcal{R}_m} \pi_{i'} \mathbb{E}S_{im}} du \\
&\stackrel{(d)}{=} \sum_{i \in \mathcal{R}_m} \pi_i \mathbb{E}[S_{\text{tr},im}] + \lambda \sum_{i \in \mathcal{R}_m} \sum_{i' \in \mathcal{R}_m} \pi_i \pi_{i'} \int_0^\infty u (1 - G_{i'm}(u)) du \\
&\stackrel{(e)}{=} \sum_{i \in [I]} \pi_i \mathbb{E}[S_{\text{tr},im}] x_{im} + \lambda \sum_{i \in [I]} \sum_{i' \in [I]} \pi_i \pi_{i'} \int_0^\infty u (1 - G_{i'm}(u)) x_{im} x_{i'm} du.
\end{aligned}$$

Here, equality (a) follows from the expression of $\hat{\gamma}_m^{\text{RT}}$ displayed above Lemma 1; equality (b) is due to the linearity of expectation; equality (c) is obtained by integrating over the pdf of L_m given by $f_{L_m}(u) = \sum_{i \in \mathcal{R}_m} \pi_i (1 - G_{im}(u)) / \sum_{i \in \mathcal{R}_m} \pi_i \mathbb{E}S_{im}$; equality (d) utilizes the relation that $\rho_m = \sum_{i \in \mathcal{R}_m} \pi_i \lambda \mathbb{E}S_{im}$, and the last equality (e) follows from our definition of the service region that $\mathcal{R}_m(\mathbf{X}) = \{i \in [I] : x_{im} = 1\}$ for $m \in \mathcal{O}(\mathbf{y}) = \{m \in [M] : y_m = 1\}$. By defining parameters $a_{im}^{\text{RT}} := \mathbb{E}[S_{\text{tr},im}]$ and $b_{im}^{\text{RT}} := \int_0^\infty u (1 - G_{im}(u)) du$ for all $i \in [I], m \in [M]$, we recover the expression in Lemma 1.

E.3. Proof of Lemma 2

By the definition of $C_m^{\text{EL}} = (\sum_{i \in \mathcal{R}_m} \pi_i) \hat{\gamma}_m^{\text{OL}}$ and the expression of $\hat{\gamma}_m^{\text{OL}}$ displayed above Lemma 2, we have

$$\begin{aligned}
C_m^{\text{OL}} &= (1 - \rho_m) \sum_{i \in \mathcal{R}_m} \pi_i \mathbb{E}[\mathbb{I}(S_{\text{tr},im} \geq \alpha)] + \rho_m \sum_{i \in \mathcal{R}_m} \pi_i \mathbb{E}[\mathbb{I}(L_m + S_{\text{tr},im} \geq \alpha)] \\
&= (1 - \rho_m) \sum_{i \in \mathcal{R}_m} \pi_i \Pr(S_{\text{tr},im} \geq \alpha) + \rho_m \sum_{i \in \mathcal{R}_m} \pi_i \Pr(L_m + S_{\text{tr},im} \geq \alpha) \\
&= \sum_{i \in \mathcal{R}_m} \pi_i \Pr(S_{\text{tr},im} \geq \alpha) - \lambda \sum_{i \in \mathcal{R}_m} \sum_{i' \in \mathcal{R}_m} \pi_i \pi_{i'} \mathbb{E}S_{i'm} \Pr(S_{\text{tr},im} \geq \alpha) \\
&\quad + \lambda \left(\sum_{i \in \mathcal{R}_m} \pi_i \mathbb{E}S_{im} \right) \sum_{i \in \mathcal{R}_m} \pi_i \Pr(L_m + S_{\text{tr},im} \geq \alpha), \tag{31}
\end{aligned}$$

where the last equality uses the identity that $\rho_m = \lambda (\sum_{i \in \mathcal{R}_m} \pi_i \mathbb{E}S_{im})$. Recall that L_m is a nonnegative random variable with pdf $f_{L_m}(u) = \sum_{i \in \mathcal{R}_m} \pi_i (1 - G_{im}(u)) / \sum_{i \in \mathcal{R}_m} \pi_i \mathbb{E}S_{im}$. The distribution of L_m depends on the service region $\mathcal{R}_m$.

To gain additional tractability, we rewrite the pdf of L_m as follows:

$$\begin{aligned}
f_{L_m}(u) &= \frac{\sum_{i \in \mathcal{R}_m} \pi_i (1 - G_{im}(u))}{\sum_{i \in \mathcal{R}_m} \pi_i \mathbb{E}S_{im}} \\
&= \frac{\sum_{i \in \mathcal{R}_m} \pi_i \mathbb{E}S_{im} \frac{1 - G_{im}(u)}{\mathbb{E}S_{im}}}{\sum_{i \in \mathcal{R}_m} \pi_i \mathbb{E}S_{im}} \\
&= \sum_{i' \in \mathcal{R}_m} \frac{\pi_{i'} \mathbb{E}S_{i'm}}{\sum_{i \in \mathcal{R}_m} \pi_i \mathbb{E}S_{im}} \frac{1 - G_{i'm}(u)}{\mathbb{E}S_{i'm}}.
\end{aligned}$$

Hence, the random variable L_m can be viewed as a mixture of the random variables $L_{i'm}, i' \in \mathcal{R}_m$ with mixture probability given by $\pi_{i'} \mathbb{E}S_{i'm} / \sum_{i \in \mathcal{R}_m} \pi_i \mathbb{E}S_{im}$ and the pdf of $L_{i'm}$ given by $f_{L_{i'm}}(u) = (1 - G_{i'm}(u)) / \mathbb{E}S_{i'm}$. In view of this interpretation, we rewrite the last term in (31) as follows:

$$\begin{aligned}
&\lambda \left(\sum_{i \in \mathcal{R}_m} \pi_i \mathbb{E}S_{im} \right) \sum_{i \in \mathcal{R}_m} \pi_i \Pr(L_m + S_{\text{tr},im} \geq \alpha) \\
&= \lambda \left(\sum_{i \in \mathcal{R}_m} \pi_i \mathbb{E}S_{im} \right) \sum_{i \in \mathcal{R}_m} \pi_i \left[\sum_{i' \in \mathcal{R}_m} \frac{\pi_{i'} \mathbb{E}S_{i'm}}{\sum_{i \in \mathcal{R}_m} \pi_i \mathbb{E}S_{im}} \Pr(L_{i'm} + S_{\text{tr},im} \geq \alpha) \right] \\
&= \lambda \sum_{i \in \mathcal{R}_m} \pi_i \left[\sum_{i' \in \mathcal{R}_m} \pi_{i'} \mathbb{E}S_{i'm} \Pr(L_{i'm} + S_{\text{tr},im} \geq \alpha) \right].
\end{aligned}$$

Replacing this expression into (31), we have

$$\begin{aligned}
C_m^{\text{OL}} &= \sum_{i \in \mathcal{R}_m} \pi_i \Pr(S_{\text{tr},im} \geq \alpha) - \lambda \sum_{i \in \mathcal{R}_m} \sum_{i' \in \mathcal{R}_m} \pi_i \pi_{i'} \mathbb{E} S_{i'm} \Pr(S_{\text{tr},im} \geq \alpha) \\
&\quad + \lambda \sum_{i \in \mathcal{R}_m} \sum_{i' \in \mathcal{R}_m} \pi_i \pi_{i'} \mathbb{E} S_{i'm} \Pr(L_{i'm} + S_{\text{tr},im} \geq \alpha) \\
&= \sum_{i \in \mathcal{R}_m} \pi_i \Pr(S_{\text{tr},im} \geq \alpha) + \lambda \sum_{i \in \mathcal{R}_m} \sum_{i' \in \mathcal{R}_m} \pi_i \pi_{i'} \mathbb{E} S_{i'm} \left[\Pr(L_{i'm} + S_{\text{tr},im} \geq \alpha) - \Pr(S_{\text{tr},im} \geq \alpha) \right] \\
&= \sum_{i \in [I]} \pi_i \Pr(S_{\text{tr},im} \geq \alpha) x_{im} + \lambda \sum_{i \in [I]} \sum_{i' \in [I]} \pi_i \pi_{i'} \mathbb{E} S_{i'm} \left[\Pr(L_{i'm} + S_{\text{tr},im} \geq \alpha) - \Pr(S_{\text{tr},im} \geq \alpha) \right] x_{im} x_{i'm},
\end{aligned}$$

where the last equality follows from our definition of the service region that $\mathcal{R}_m(\mathbf{X}) = \{i \in [I] : x_{im} = 1\}$ for $m \in \mathcal{O}(\mathbf{y}) = \{m \in [M] : y_m = 1\}$. Defining parameters $a_{im}^{\text{OL}} := \Pr(S_{\text{tr},im} \geq \alpha)$ and $b_{ii'm}^{\text{OL}} := \Pr(L_{i'm} + S_{\text{tr},im} \geq \alpha)$ for all $i, i' \in [I], m \in [M]$, we recover the expression in Lemma 2.

E.4. Proof of Proposition 2

We first show that constraints (6b)-(6e) collectively guarantee that $z_{ii'm} = x_{im}x_{i'm}$ for all $i \in [I], i' \in [I], m \in [M]$. Because x_{im} is binary, it suffices to enumerate all the possible cases. If $x_{im} = 1$ and $x_{i'm} = 1$, we have $z_{ii'm} \geq 1$ from (6d). Since $z_{ii'm}$ is binary by (6e), this ensures that $z_{ii'm} = 1$, matching the product of x_{im} and $x_{i'm}$. At the same time, constraints (6b) and (6c) are not violated. Next, we consider the case in which $x_{im} = 0$ or $x_{i'm} = 0$. In this case, we have $z_{ii'm} = 0$ from constraints (6b) or (6c), so the value of $z_{ii'm}$ again matches the product $x_{im}x_{i'm}$. Meanwhile, constraint (6d) is satisfied in the case because we have $1 - z_{ii'm} = 1$ and $2 - x_{im} - x_{i'm} = 2$, thus constraint $2 - x_{im} - x_{i'm} \geq 0$ is clearly satisfied.

With the above knowledge that $z_{ii'm} = x_{im}x_{i'm}$, we next show that constraint (5a) in formulation (5) is equivalent to constraint (6a) in MRT-AsmApX. Consider an arbitrary $m \in [M]$. If $y_m = 1$, then the equivalence is obvious by replacing $x_{im}x_{i'm}$ in (5a) by $z_{ii'm}$. Otherwise if $y_m = 0$, constraint (5a) stipulates that $C_m^{\text{RT}} = 0$. It suffices to show that the constraint (6a) also ensures that $C_m^{\text{RT}} = 0$ in this case. By the feasibility of x_{im} in (2), it holds that $x_{im} \leq y_m = 0$ for all $i \in [I]$. As such, we have $x_{im} = 0$ and $z_{ii'm} = x_{im}x_{i'm} = 0$ for all $i, i' \in [I]$. Therefore, constraint (6a) ensures that

$$C_m^{\text{RT}} = \sum_{i \in [I]} \pi_i a_{im}^{\text{RT}} x_{im} + \lambda \sum_{i \in [I]} \sum_{i' \in [I]} \pi_i \pi_{i'} b_{ii'm}^{\text{RT}} z_{ii'm} = 0.$$

Putting all the pieces together, we have established that constraint (5a) is equivalent to constraints (6a)-(6e). This in turn proves the equivalence of formulation (5) and formulation MRT-AsmApX.

E.5. Proof of Theorem 3

By optimality of $(\mathbf{y}_\star^h, \mathbf{X}_\star^h)$ for the ground-truth objective γ^h , we have $0 \leq \gamma^h(\mathbf{y}_{\text{Asm}}^h, \mathbf{X}_{\text{Asm}}^h) - \gamma^h(\mathbf{y}_\star^h, \mathbf{X}_\star^h)$. Moreover, by telescoping terms, we have

$$\begin{aligned}
\gamma^h(\mathbf{y}_{\text{Asm}}^h, \mathbf{X}_{\text{Asm}}^h) - \gamma^h(\mathbf{y}_\star^h, \mathbf{X}_\star^h) &= [\gamma^h(\mathbf{y}_{\text{Asm}}^h, \mathbf{X}_{\text{Asm}}^h) - \hat{\gamma}_{\text{Asm}}^h(\mathbf{y}_{\text{Asm}}^h, \mathbf{X}_{\text{Asm}}^h)] + [\hat{\gamma}_{\text{Asm}}^h(\mathbf{y}_{\text{Asm}}^h, \mathbf{X}_{\text{Asm}}^h) - \hat{\gamma}_{\text{Asm}}^h(\mathbf{y}_\star^h, \mathbf{X}_\star^h)] \\
&\quad + [\hat{\gamma}_{\text{Asm}}^h(\mathbf{y}_\star^h, \mathbf{X}_\star^h) - \gamma^h(\mathbf{y}_\star^h, \mathbf{X}_\star^h)].
\end{aligned}$$

Because $(\mathbf{y}_{\text{Asm}}^h, \mathbf{X}_{\text{Asm}}^h)$ minimizes $\hat{\gamma}^h$, the middle term is nonpositive. Therefore, we have

$$0 \leq \gamma^h(\mathbf{y}_{\text{Asm}}^h, \mathbf{X}_{\text{Asm}}^h) - \gamma^h(\mathbf{y}_\star^h, \mathbf{X}_\star^h) \leq \gamma^h(\mathbf{y}_{\text{Asm}}^h, \mathbf{X}_{\text{Asm}}^h) - \hat{\gamma}_{\text{Asm}}^h(\mathbf{y}_{\text{Asm}}^h, \mathbf{X}_{\text{Asm}}^h) + \hat{\gamma}_{\text{Asm}}^h(\mathbf{y}_\star^h, \mathbf{X}_\star^h) - \gamma^h(\mathbf{y}_\star^h, \mathbf{X}_\star^h).$$

By the triangle inequality, the right-hand side is bounded by $2 \max_{\mathbf{y} \in \mathcal{Y}, \mathbf{X} \in \mathcal{X}(\mathbf{y})} |\gamma^h(\mathbf{y}, \mathbf{X}) - \hat{\gamma}_{\text{Asm}}^h(\mathbf{y}, \mathbf{X})|$. Therefore, it suffices to show that $\max_{\mathbf{y} \in \mathcal{Y}, \mathbf{X} \in \mathcal{X}(\mathbf{y})} |\gamma^h(\mathbf{y}, \mathbf{X}) - \hat{\gamma}_{\text{Asm}}^h(\mathbf{y}, \mathbf{X})| = O(\lambda^2)$. The maximum is taken over finitely many pairs of $\mathbf{y}$ and $\mathbf{X}$, and thus it suffices to show that, for every feasible $(\mathbf{y}, \mathbf{X})$, $|\gamma^h(\mathbf{y}, \mathbf{X}) - \hat{\gamma}_{\text{Asm}}^h(\mathbf{y}, \mathbf{X})| = O(\lambda^2)$. For clarity, fix a feasible $(\mathbf{y}, \mathbf{X})$ and suppress its dependence in the notation. For each occupied base $m \in \mathcal{O}$, let $\gamma_{\mathcal{B},m}^h$ denote the true long-run average performance loss of queued calls in region $\mathcal{R}_m$. Following the same arguments that lead to (7), and averaging over $m \in \mathcal{O}$ with weight $\sum_{i \in \mathcal{R}_m} \pi_i$, we obtain

$$\gamma^h = \sum_{m \in \mathcal{O}} \left(\sum_{i \in \mathcal{R}_m} \pi_i \right) \left[(1 - \rho_m) \frac{\sum_{i \in \mathcal{R}_m} \pi_i \mathbb{E}[h(S_{\text{tr},im})]}{\sum_{i \in \mathcal{R}_m} \pi_i} + \rho_m \gamma_{\mathcal{B},m}^h \right].$$

Our asymptotic approximation replaces $\gamma_{\mathcal{B},m}^h$ by its light-traffic limit $\hat{\gamma}_{\mathcal{B},m}^h := \frac{\sum_{i \in \mathcal{R}_m} \pi_i \mathbb{E}[h(L_m + S_{\text{tr},im})]}{\sum_{i \in \mathcal{R}_m} \pi_i}$, where the random L_m are carried over from §3.2. Thus, the objective of our model is given by

$$\hat{\gamma}^h = \sum_{m \in \mathcal{O}} \left(\sum_{i \in \mathcal{R}_m} \pi_i \right) \left[(1 - \rho_m) \frac{\sum_{i \in \mathcal{R}_m} \pi_i \mathbb{E}[h(S_{\text{tr},im})]}{\sum_{i \in \mathcal{R}_m} \pi_i} + \rho_m \hat{\gamma}_{\mathcal{B},m}^h \right].$$

Subtracting the two expressions yields

$$|\gamma^h - \hat{\gamma}^h| = \sum_{m \in \mathcal{O}} \left(\sum_{i \in \mathcal{R}_m} \pi_i \right) \rho_m \left| \gamma_{\mathcal{B},m}^h - \hat{\gamma}_{\mathcal{B},m}^h \right| = \sum_{m \in \mathcal{O}} \left(\sum_{i \in \mathcal{R}_m} \pi_i \right) \left(\sum_{i \in \mathcal{R}_m} \pi_i \lambda \mathbb{E}S_{im} \right) \left| \gamma_{\mathcal{B},m}^h - \hat{\gamma}_{\mathcal{B},m}^h \right|,$$

where the last equality uses the definition $\rho_m = \sum_{i \in \mathcal{R}_m} \pi_i \lambda \mathbb{E}S_{im}$. By Theorem 1, we have $|\gamma_{\mathcal{B},m}^h - \hat{\gamma}_{\mathcal{B},m}^h| = O(\lambda)$ as $\lambda \downarrow 0$, and so $|\gamma^h - \hat{\gamma}^h| = O(\lambda^2)$. This bound holds for every feasible $(\mathbf{y}, \mathbf{X})$, completing the proof.

E.6. Proof of Proposition 3

We first prove part (i) by constructing a problem instance such that $\gamma^h(\mathbf{y}_{\text{Bmk}}, \mathbf{X}_{\text{Bmk}}) - \gamma^h(\mathbf{y}_\star^h, \mathbf{X}_\star^h) > \mathbb{C}\lambda$. Consider an instance with $I = 1$ demand node, $M = 2$ candidate bases, and $J = 1$ server. The location-allocation problem reduces to choosing which base to locate the single server that serves the single demand node. Assume the travel times are such that $\mathbb{E}S_{\text{tr},11} = \mathbb{E}S_{\text{tr},12}$. Then the benchmark MRT-Bmk1, which only accounts for mean travel time, is indifferent between base 1 and base 2. Suppose MRT-Bmk1 chooses base 1; an analogous argument applies if it randomizes between the two bases. Under $h(x) = x$, the ground-truth objective is the mean response time. If the server is located at base $m \in \{1, 2\}$, the mission time is $S_{1m} = S_{\text{tr},1m} + S_{\text{os},1}$, where the travel time $S_{\text{tr},1m}$ and on-scene service time $S_{\text{os},1}$ are independent. By the P-K formula for the mean waiting time of an M/G/1 queue, the average response time when the server is at base m is

$$\frac{\lambda \mathbb{E}S_{1m}^2}{2(1 - \lambda \mathbb{E}S_{1m})} + \mathbb{E}S_{\text{tr},1m} = \frac{\lambda (\mathbb{E}S_{\text{os},1}^2 + \mathbb{E}S_{\text{tr},1m}^2 + 2\mathbb{E}S_{\text{tr},1m}\mathbb{E}S_{\text{os},1})}{2(1 - \lambda \mathbb{E}S_{1m})} + \mathbb{E}S_{\text{tr},1m},$$

where the equality uses independence of $S_{\text{tr},1m}$ and $S_{\text{os},1}$. We choose the travel time distributions so that $\mathbb{E}S_{\text{tr},11}^2 > \mathbb{E}S_{\text{tr},12}^2$, which implies $\mathbb{E}S_{11}^2 > \mathbb{E}S_{12}^2$, while maintaining $\mathbb{E}S_{\text{tr},11} = \mathbb{E}S_{\text{tr},12}$. Then

$$\Delta(\lambda) := \left(\frac{\lambda \mathbb{E}S_{11}^2}{2(1 - \lambda \mathbb{E}S_{11})} + \mathbb{E}S_{\text{tr},11} \right) - \left(\frac{\lambda \mathbb{E}S_{12}^2}{2(1 - \lambda \mathbb{E}S_{12})} + \mathbb{E}S_{\text{tr},12} \right) = \lambda \frac{\mathbb{E}S_{11}^2 - \mathbb{E}S_{12}^2}{2(1 - \lambda \mathbb{E}S_{11})} = \lambda \phi(\lambda),$$

where the second equality uses $\mathbb{E}S_{11} = \mathbb{E}S_{12}$ and the function $\phi(\lambda) := \frac{\mathbb{E}S_{11}^2 - \mathbb{E}S_{12}^2}{2(1 - \lambda \mathbb{E}S_{11})}$ is strictly increasing in λ on $(0, (1 - \epsilon)/\max_m \{\mathbb{E}S_{1m}\})$. Therefore, there exist a constant $\mathbb{C} = \phi(0) > 0$ such that $\phi(\lambda) > \mathbb{C}$ for all $\lambda \in (0, (1 - \epsilon)/\max_m \{\mathbb{E}S_{1m}\})$. Hence, $\gamma^h(\mathbf{y}_{\text{Bmk}}, \mathbf{X}_{\text{Bmk}}) - \gamma^h(\mathbf{y}_\star^h, \mathbf{X}_\star^h) = \Delta(\lambda) > \mathbb{C}\lambda$. In this instance, the ground-truth optimal solution $(\mathbf{y}_\star^h, \mathbf{X}_\star^h)$ locates the server at base 2, whereas MRT-Bmk1 chooses base 1, and the resulting optimality gap is of order λ , establishing part (i).

Next, we prove part (ii). Since $h(x) = x$, the performance measure is the average response time. As before, fix a feasible $(\mathbf{y}, \mathbf{X})$ and omit it from the notation. For each $m \in \mathcal{O}$, let γ_m^h , $\hat{\gamma}_{\text{Asm},m}^h$, and $\hat{\gamma}_{\text{Bmk},m}^h$ denote the average response time of calls in region $\mathcal{R}_m$ under the ground-truth model, our asymptotic approximation model, and the benchmark MRT-Bmk1, respectively. Suppose we show that, for each $m \in \mathcal{O}$,

$$\frac{\gamma_m^h - \hat{\gamma}_{\text{Asm},m}^h}{\gamma_m^h - \hat{\gamma}_{\text{Bmk},m}^h} = \rho_m. \quad (32)$$

This will complete the proof because: (i) the overall errors $\gamma^h - \hat{\gamma}_{\text{Asm}}^h$ and $\gamma^h - \hat{\gamma}_{\text{Bmk}}^h$ are weighted averages of the region-wise errors, with weights $\sum_{i \in \mathcal{R}_m} \pi_i$; and (ii) each region-wise error is positive for $\lambda > 0$ (shown below). We now establish (32). Define the effective arrival rate, mission duration, and travel time for region $\mathcal{R}_m$ by $\lambda_m := \sum_{i \in \mathcal{R}_m} \pi_i \lambda$, $S_m := \frac{\sum_{i \in \mathcal{R}_m} \pi_i S_{im}}{\sum_{i \in \mathcal{R}_m} \pi_i}$ and $S_{\text{tr},m} := \frac{\sum_{i \in \mathcal{R}_m} \pi_i S_{\text{tr},im}}{\sum_{i \in \mathcal{R}_m} \pi_i}$, respectively. By the P-K formula, the average response time in region $\mathcal{R}_m$ under the ground-truth model is

$$\gamma_m^h = \frac{\lambda_m \mathbb{E} S_m^2}{2(1 - \lambda_m \mathbb{E} S_m)} + \mathbb{E} S_{\text{tr},m}.$$

Using (3) and the definition of L_m therein, our asymptotic approximation model yields

$$\hat{\gamma}_{\text{Asm},m}^h = \mathbb{E} S_{\text{tr},m} + \lambda_m \mathbb{E} S_m \frac{\mathbb{E} S_m^2}{2 \mathbb{E} S_m}.$$

On the other hand, the benchmark MRT-Bmk1 assumes the server is always idle, so its average response time is simply the average travel time, that is, $\hat{\gamma}_{\text{Bmk},m}^h = \mathbb{E} S_{\text{tr},m}$. With the above, we have $\gamma_m^h - \hat{\gamma}_{\text{Asm},m}^h = \frac{\lambda_m^2 \mathbb{E} S_m \mathbb{E} S_m^2}{2(1 - \lambda_m \mathbb{E} S_m)} > 0$ and $\gamma_m^h - \hat{\gamma}_{\text{Bmk},m}^h = \frac{\lambda_m \mathbb{E} S_m^2}{2(1 - \lambda_m \mathbb{E} S_m)} > 0$. As a result, we have

$$\frac{\gamma_m^h - \hat{\gamma}_{\text{Asm},m}^h}{\gamma_m^h - \hat{\gamma}_{\text{Bmk},m}^h} = \lambda_m \mathbb{E} S_m,$$

which establishes (32) since $\lambda_m \mathbb{E} S_m = \sum_{i \in \mathcal{R}_m} \pi_i \lambda \frac{\sum_{i \in \mathcal{R}_m} \pi_i \mathbb{E} S_{im}}{\sum_{i \in \mathcal{R}_m} \pi_i} = \rho_m$.

E.7. Extension to Heterogeneous Servers

We next consider heterogeneous servers in the ESS design problem. Let $S_{\text{tr},imj}$ be the time required by server $j \in [J]$ to travel from base $m \in [M]$ to node $i \in [I]$, and $S_{\text{os},ij}$ be the on-scene service time for server j to handle a call from node i . The distributions of these variables reflect both the spatial structure of ESS and the characteristics of the heterogeneous servers. Let S_{imj} denote the mission duration for server j located at base m to service a call from node i . Its distribution G_{imj} can be computed from a convolution of the variables $S_{\text{tr},imj}$ and $S_{\text{os},ij}$. Compared with the inputs in the previous sections, an extra dimension on server information is introduced in the above inputs.

Similarly, we augment the decision variables with one more dimension to index servers. Specifically, define a zero-one indicator y_{mj} for the m th candidate base station and the j th server, where $m \in [M]$ and $j \in [J]$, such that $y_{mj} = 1$ if base m is occupied by server j , and zero otherwise. Moreover, let x_{imj} be a zero-one indicator that equals one only if server j is at base m and designated to serve demand node i . These indicators fully determine the locations of the servers at the bases and the allocation of demand nodes to servers. Since each server must be located at one base and each base can accommodate at most one server, the location indicators y_{mj} satisfy

$$y_{mj} \in \{0, 1\}, \sum_{m \in [M]} y_{mj} = 1 \text{ and } \sum_{j \in [J]} y_{mj} \leq 1 \quad \text{for } m \in [M], j \in [J]. \quad (33)$$

On the other hand, because each demand node is allocated to exactly one candidate base that is occupied by one of the J servers, the allocation indicators x_{imj} satisfy

$$\begin{aligned} x_{imj} \in \{0, 1\}, \sum_{m \in [M], j \in [J]} x_{imj} = 1 \text{ and } x_{imj} \leq y_{mj} \quad \text{for } i \in [I], m \in [M], j \in [J], \\ \sum_{i \in [I]} x_{imj} \pi_i \lambda \mathbb{E} S_{imj} \leq 1 - \epsilon \quad \text{for } m \in [M], j \in [J]. \end{aligned} \quad (34)$$

As in §3.3, we evaluate the performance loss associated with a feasible location-allocation scheme $\{y_{mj} : m \in [M], j \in [J]\}$ and $\{x_{imj} : i \in [I], m \in [M], j \in [J]\}$ by partitioning the I demand nodes into J service regions $\mathcal{R}_m$ for occupied bases $m \in \mathcal{O}$. In this setting, we have $\mathcal{O} := \{m \in [M] : \sum_{j \in [J]} y_{mj} = 1\}$ and $\mathcal{R}_m := \{i \in [I] : \sum_{j \in [J]} x_{imj} = 1\}$. With tedious algebra similar to but more complicated than that in the preceding section, minimizing the long-run average performance loss under a general loss function h can be formulated as the MILP as follows:

$$\begin{aligned} \min_{\mathbf{y}, \mathbf{x}, \mathbf{z}, \mathbf{C}_{mj}^h} \quad & \sum_{m \in [M]} \sum_{j \in [J]} C_{mj}^h \\ \text{subject to} \quad & C_{mj}^h = \sum_{i \in [I]} \pi_i a_{imj}^h x_{imj} + \lambda \sum_{i \in [I]} \sum_{i' \in [I]} (b_{ii'mj}^h - a_{i'i'mj}^h) \pi_i \pi_{i'} \mathbb{E} S_{im} z_{ii'mj}, \forall m \in [M], j \in [J]; \\ & z_{ii'mj} \leq x_{imj}, z_{ii'mj} \leq x_{i'mj}, \quad \forall i \in [I], i' \in [I], m \in [M], j \in [J]; \\ & 1 - z_{ii'mj} \leq 2 - x_{imj} - x_{i'mj}, \quad \forall i \in [I], i' \in [I], m \in [M], j \in [J]; \\ & z_{ii'mj} \in \{0, 1\}, \quad \forall i \in [I], i' \in [I], m \in [M], j \in [J]; \\ & y_{mj} \text{ subject to (33), } x_{imj} \text{ subject to (34)}. \end{aligned} \quad (35)$$

Here, we have defined parameters $a_{imj}^h := \mathbb{E}[h(S_{\text{tr},imj})]$ and $b_{ii'mj}^h := \mathbb{E}[h(L_{imj} + S_{\text{tr},i'mj})]$, where L_{imj} is a nonnegative random variable with pdf $f_{L_{imj}}(x) := (1 - G_{imj}(x))/\mathbb{E} S_{imj}$. These parameters can be computed from the problem inputs, i.e., the distributions of $S_{\text{tr},imj}$ and $S_{\text{os},ij}$.

Appendix F: Supplementary Materials to §4

F.1. Equivalence of benchmarks in §3.4 and §4.1 to well-known location models

We introduce the p -median location problem following the setup in the review paper of Ahmadi-Javid et al. (2017), with adaptation to our notation system. Let $[I] = \{1, \dots, I\}$ denote the set of demand points and $[M] = \{1, \dots, M\}$ denote the set of candidate locations. Let d_{im} be the travel time from demand point $i \in [I]$ to candidate location $m \in [M]$, π_i be the demand at point $i \in [I]$ and J be the number of candidate locations to be established. Let y_m be a binary decision variable that equals 1 if a facility is established at candidate location $m \in [M]$; and $y_m = 0$ otherwise. Let x_{im} be a binary decision variable that equals 1 if demand point $i \in [I]$ is assigned to a facility at candidate location $m \in [M]$. Then, the p -median location model that minimizes the total demand-weighted travel distance takes the following form:

$$\begin{aligned} \text{minimize} \quad & \sum_{i \in [I]} \sum_{m \in [M]} \pi_i d_{im} x_{im} \\ \text{subject to} \quad & \sum_{m \in [M]} x_{im} = 1, i \in [I] \\ & \sum_{m \in [M]} y_m = J; \quad x_{im} \leq y_m, i \in [I], m \in [M] \\ & x_{im} \in \{0, 1\}, i \in [I], m \in [M]; \quad y_m \in \{0, 1\}, m \in [M]. \end{aligned}$$

The location and assignment constraints above coincide with those in (1) and (2), except that the latter also impose stability. Hence, the benchmark model MRT-Bmk1 can be viewed as an instance of the above p-median location model where the travel time parameter d_{im} is replaced by $a_{im}^{\text{RT}} = \mathbb{E}[S_{\text{tr},im}]$ when the stability constraints are redundant.

As for the maximum partial covering location model, we introduce the parameter $z_{im} \in [0, 1]$, which represents the coverage level provided by candidate location $m \in [M]$ to demand node $i \in [I]$. This coverage level is expected to decrease as the distance between candidate location $m \in [M]$ and demand node $i \in [I]$ increases. The maximum partial covering location model that maximizes the total demand-weighted coverage level is as follows:

$$\begin{aligned} & \text{maximize} && \sum_{i \in [I]} \sum_{m \in [M]} \pi_i z_{im} x_{im} \\ & \text{subject to} && \sum_{m \in [M]} x_{im} = 1, i \in [I] \\ & && \sum_{m \in [M]} y_m = J; \quad x_{im} \leq y_m, i \in [I], m \in [M] \\ & && x_{im} \in \{0, 1\}, i \in [I], m \in [M]; \quad y_m \in \{0, 1\}, m \in [M]. \end{aligned}$$

By defining the coverage level z_{im} as the probability that the travel time between candidate location $m \in [M]$ and demand node $i \in [I]$ is less than α , it holds that $z_{im} = 1 - a_{im}^{\text{OL}}$, where $a_{im}^{\text{OL}} = \Pr(S_{\text{tr},im} \geq \alpha)$. Since the maximization of an objective function is equivalent to the minimization of the same objective function multiplied by -1 , the benchmark model MOL-Bmk1 is equivalent to the maximum partial covering location model when the stability constraints are redundant, up to a constant of $\sum_{i \in [I]} \pi_i$ (i.e., the total demand) in the optimal objective value.

F.2. Investigation results for Corollary 1 in §4.2

In this section, we investigate the special cases mentioned in Corollary 1. Similar to §4.2, we compare the simulated average loss of queued calls in these special cases with the respective asymptotic limit provided by Corollary 1 under a given location-allocation scheme. Specifically, we consider a multi-server spatial queue with $J = 3$ servers located at coordinates (6, 6), (9, 9), and (12, 6). The travel times and mission durations are generated according to the three special cases. Servers are authorized to service all the demand nodes. When a call arrives to find multiple servers idle, the nearest one is dispatched. We focus on the demand distribution in Scenario 1 of Figure (2a). The comparison results are presented in Figure 13, where we use the simulated probability that all servers are busy as the x-axis, because for a given arrival rate λ , the traffic intensity in a multi-server spatial queue is not obvious. Compared to the single-server setting, the relative error is more sensitive to the arrival rate λ . Nonetheless, the simulated values approach the asymptotic limit in Corollary 1.

F.3. Robustness check using assumption-relaxed simulation for results in §4.2

In this section, we repeat the investigation in §4.2 using a simulation model that relaxes the assumption that servers always travel from base stations to demand nodes. To this end, we consider both forward (from base to demand) and return (from demand to base) travels of servers in the simulation. When a server finishes the on-scene service, it travels back to its base station if no call is waiting; otherwise, it proceeds directly to the next queued call. We assume this return travel time has the same distribution as the forward one, with an independent draw. If a call arrives in the middle of the return travel of its assigned server, the server would abort its return travel and travel to the call. Since it is infeasible to obtain the travel time distribution from an arbitrary point in the return travel to a demand node, we approximate this

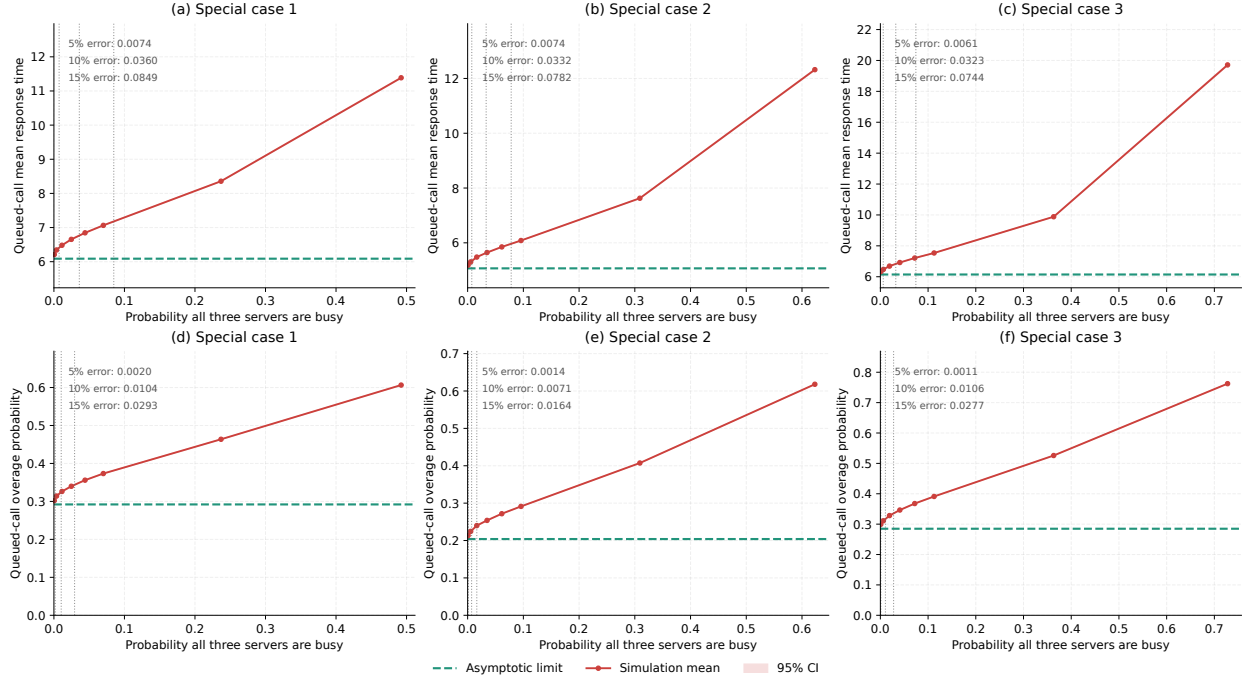


Figure 13 Comparison of simulated average performance loss of queued calls under varying busy probabilities with the asymptotic limit indicated by Corollary 1. The top row corresponds to $h(x) = x$, while the bottom row represents $h(x) = \mathbb{I}(x \geq \alpha)$. Columns, from left to right, depict special cases 1-3, respectively. Red lines are the mean values of 100 simulation runs, where the widths of 95% CI are less than 0.093 for response time and 0.0092 for overage loss.

“cut travel time” by the geometric relationship of the available travel times, as illustrated in Figure 14. Consider the scenario that a call from node j arrives at time t and is assigned to a server who has just completed its service at node i at time $t_{\text{comp}} < t$ and on the way back to its base station, denoted by node 0 in Figure 14. To calculate the cut travel time x , we sample travel time realizations from node i to node j , denoted by $t_{i,j}$, travel time from the base station to node j , denoted by $t_{0,j}$, and use $t_{i,0}$ for the already realized duration of the ongoing return from node i to the base station. If the three values $t_{i,j}$, $t_{0,j}$ and $t_{i,0}$ can form a triangle as shown in Figure 14, we use the geometric relation to get the cut travel time x . Otherwise, we set $x = \min\{t - t_{\text{comp}} + t_{i,j}, t_{i,0} - (t - t_{\text{comp}}) + t_{0,j}\}$, which is the minimum of the travel time of two paths, where the first path connects the current position of the server with node j via node i , while the second path connects the two via node 0.

Using this assumption-relaxed simulation, we repeat the comparison in §4.2 between the simulated average response time of queued calls, with $h(x) = x$, and the asymptotic limit from Theorem 1. The results are shown in Figure 15. We observe that the simulated loss exhibit the same pattern as in §4.2, approaching the asymptotic loss as the arrival rate approaches zero. This indicates that the asymptotic loss derived in Theorem 1 can still provide a good approximation for the performance loss of queued calls in real-world scenarios where servers could potentially travel from any point of the return travel to a demand node.

F.4. Location-allocation optimization results for average overage loss in §4.3.2

Following the setting in §4.3.2, we solve the proposed model MOL-AsmAp $\times$ and the two benchmark models MOL-Bmk1 and MOL-Bmk2 to obtain optimal location-allocation schemes for the three demand scenarios. The optimization results

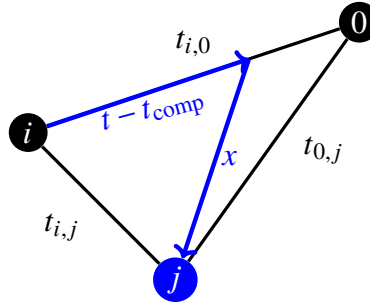


Figure 14 Illustration of calculating the cut travel time. A call from node j arrives at time t and is assigned a server who has completed its service at node i at time $t_{\text{comp}} < t$ and is on the way back to its base station at node 0. The cut travel time is x .

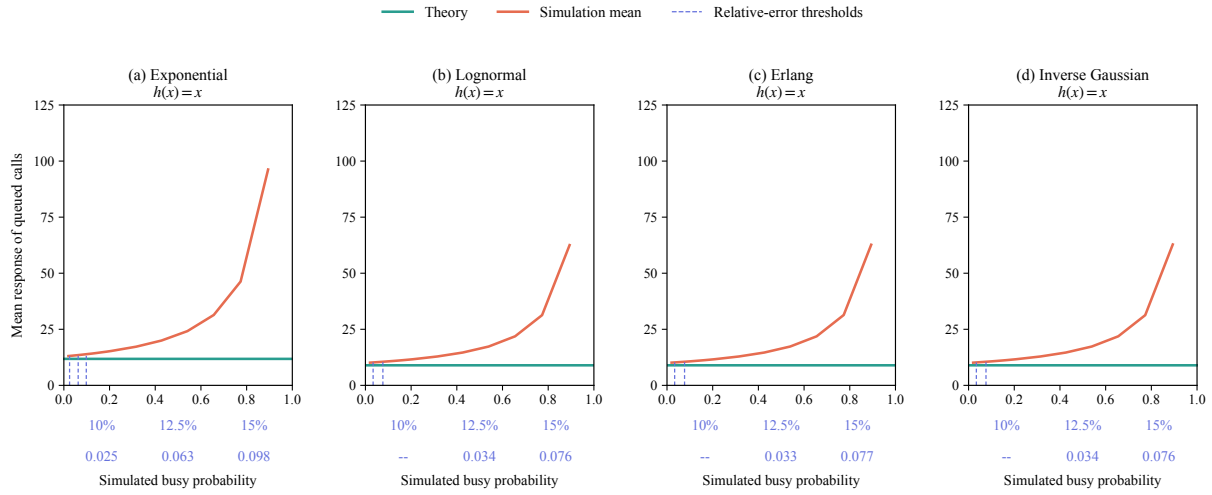


Figure 15 Comparison of simulated average response times of queued calls under varying traffic intensity conditions produced by the assumption-relaxed simulation, with the theoretical values from Theorem 1. The columns, from left to right, present cases where mission durations follow exponential, lognormal, Erlang, and inverse Gaussian distributions. The widths of the 95% CIs for the simulated means are at most 0.901. The figures demonstrate that Theorem 1, based on the assumption of instantaneous return, can still provide a good approximation for the loss rate of queued calls in real-world scenarios where the returning travel of servers exists.

are presented in Figure 16. Moreover, we compare the simulated overage loss of these location-allocation schemes by conducting 100 simulation runs. The results are reported in Figure 17. We also compute the P -values for the hypotheses $H_0 : \mathbb{E}[C_{\text{MOL-Bmk1}}^{\text{OL}}] \leq \mathbb{E}[C_{\text{MOL-AsmApX}}^{\text{OL}}]$ and $H_0 : \mathbb{E}[C_{\text{MOL-Bmk2}}^{\text{OL}}] \leq \mathbb{E}[C_{\text{MOL-AsmApX}}^{\text{OL}}]$. Both P -values are less than 10^{-3} , indicating strong rejections to the null hypotheses. The results show that our proposed model MOL-AsmApX outperforms both benchmarks MOL-Bmk1 and MOL-Bmk2 in all three demand scenarios, yielding a consistently lower mean overage loss.

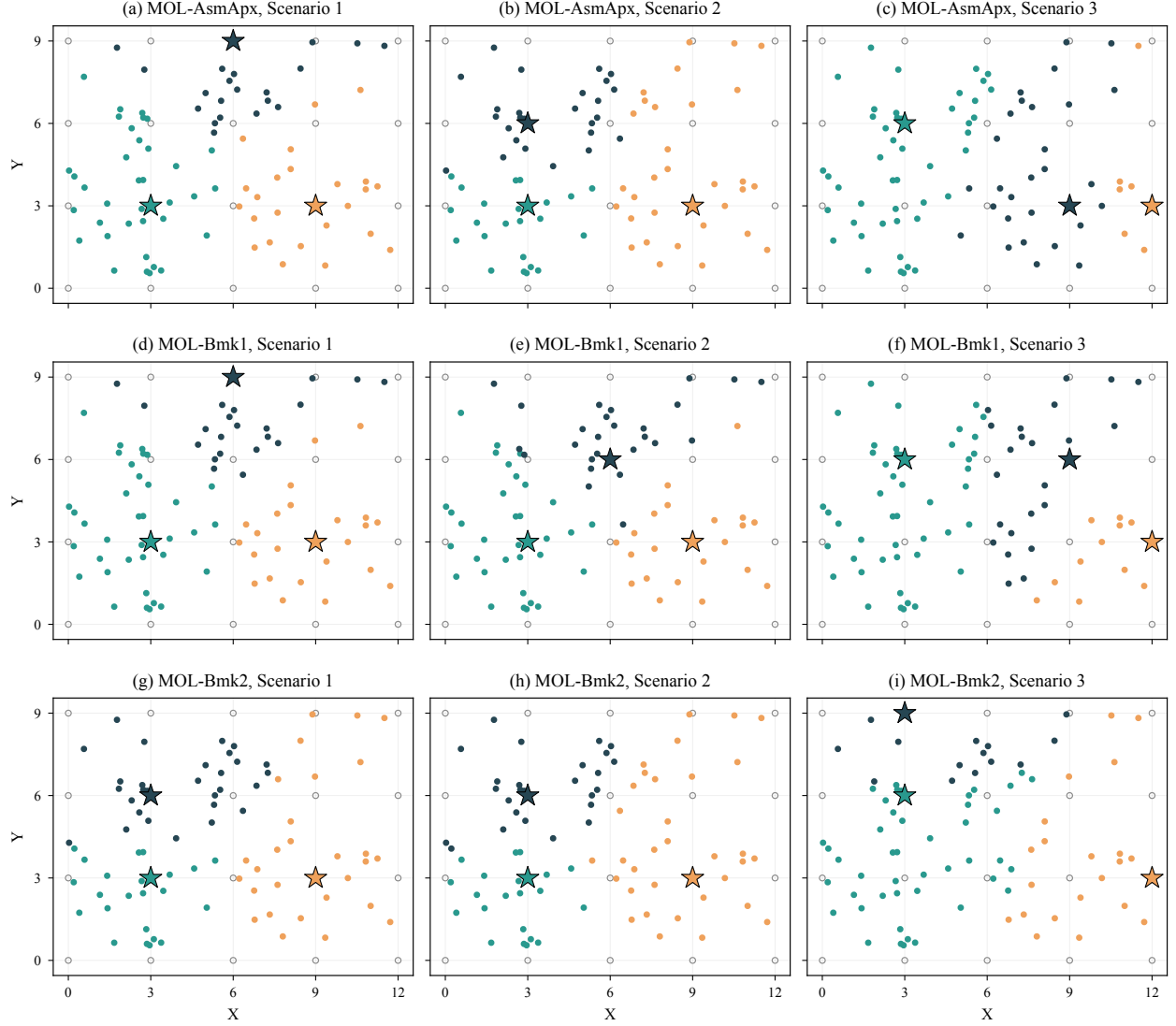


Figure 16 Optimal location-allocation schemes obtained from the proposed MOL-AsmApX (first row), the benchmark MOL-Bmk1 (second row) and MOL-Bmk2 (third row) under the three demand scenarios. The star symbols describe the optimal server locations. The color of demand nodes encode the optimal allocations.

F.5. Larger Scale Location-Allocation Optimization in §4.3.2

In this section, we revisit the location-allocation optimization in §4.3.2 with a larger problem scale. Specifically, in addition to the 20 candidate bases shown in Figure 2, we expand the candidate locations to include the centers of each 3×3 square. This increases the total number of candidate bases to $M = 32$. In this setting, we are interested in optimally locating $J = 10$ servers and allocating demand nodes to these servers. As an illustration, we focus on the demand Scenario 1 with a goal to minimize the average response time. Following the procedure in §4.3.2, we solve the proposed model MRT-AsmApX and the benchmark MRT-Bmk1 to obtain the optimal location-allocation schemes, and we compare the simulated average response time of these schemes from 100 simulation runs. The optimization results are shown in Figure 18, and the simulation comparison is provided in Figure 19. The optimal scheme from our proposed

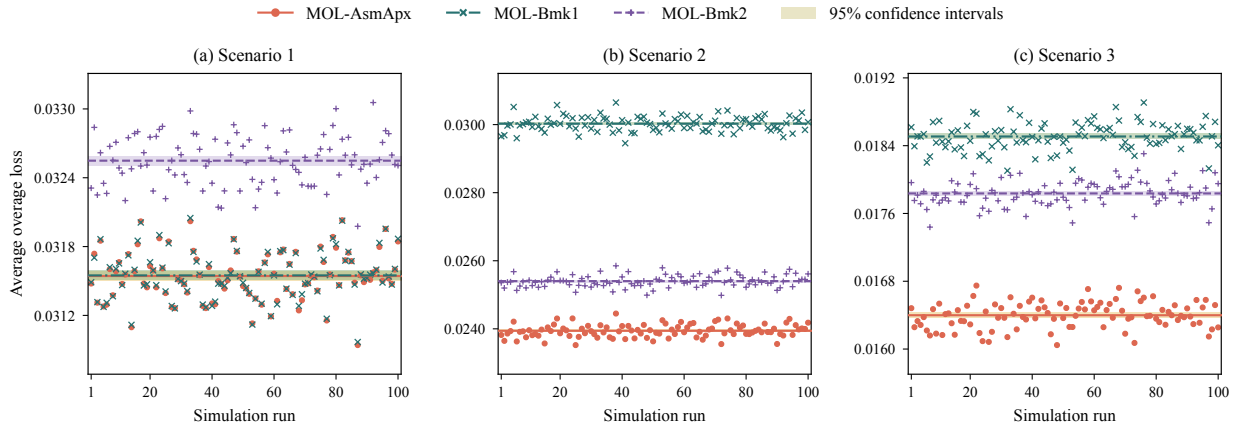


Figure 17 Simulated overage loss associated with location-allocation schemes derived from the proposed MOL-AsmAp_x and benchmarks MOL-Bmk1 and MOL-Bmk2 under three demand scenarios from 100 paired simulation runs. The figures depict the 100 simulated values, the mean of the 100 values and 95% CI for the mean.

model MRT-AsmAp_x yields consistently faster average response times than that of MRT-Bmk1. The results from this increased problem scale reaffirm the conclusions drawn in §4.3.2.

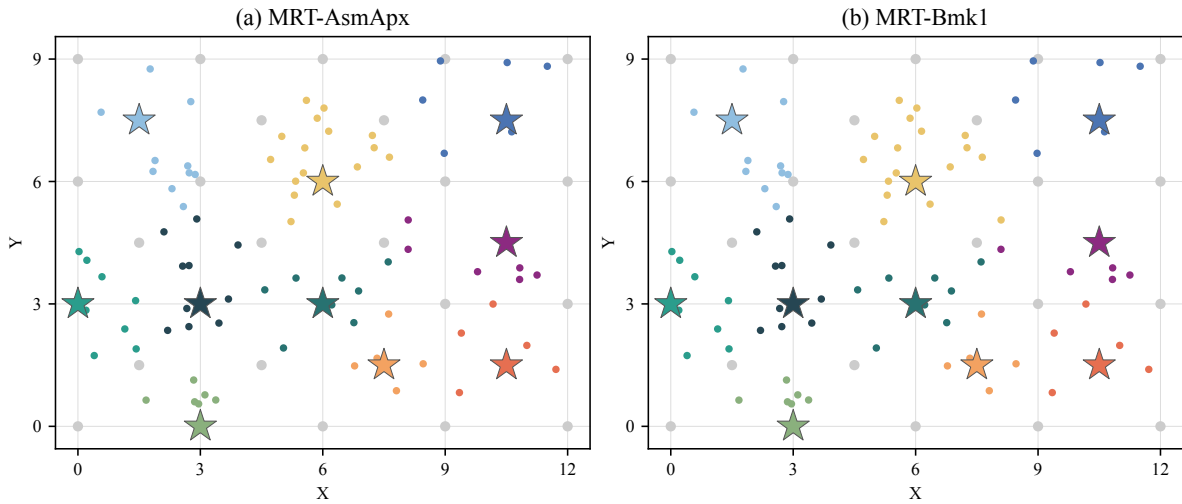


Figure 18 Optimal location-allocation schemes obtained from the proposed MRT-AsmAp_x (left) and the benchmark MRT-Bmk1 (right) in the larger-scale problem. The star symbols describe the optimal server locations. The colors of demand nodes encode the optimal allocations. Both methods select the same ten base locations but differ in the allocation of demand node 77.

F.6. Additional numerical results for §4.4 (Comparison with brute-force)

This subsection provides supplementary numerical results for §4.4 of the main paper. Figure (20a) reports the maximum, minimum, and average busy probabilities of the three servers in the base example under the optimal brute-force scheme

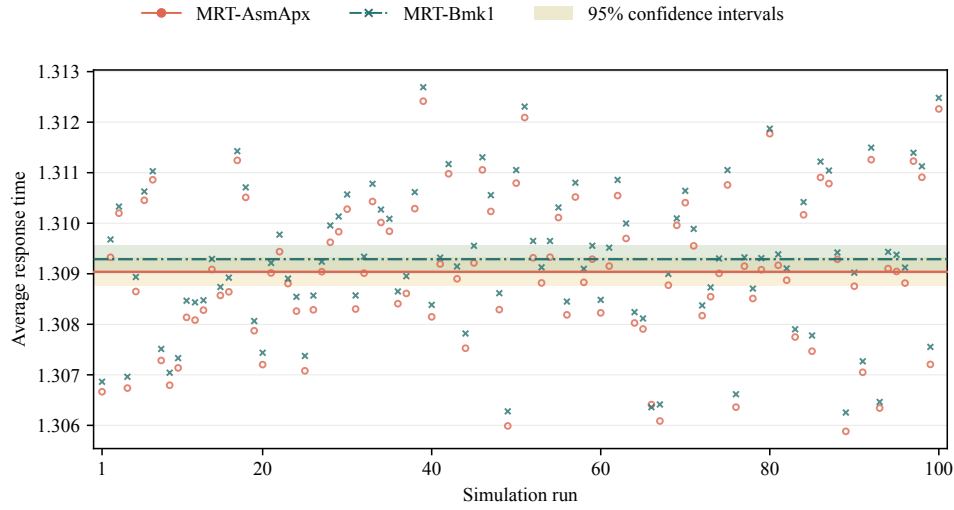


Figure 19 Simulated response time of the optimal location-allocation schemes derived from the proposed MRT-AsmApX and the benchmark MRT-Bmk1 in the larger-scale problem over 100 paired simulation runs.

as the total arrival rate λ varies. The busiest server's utilization approaches 0.8 at the highest λ , indicating that our parameter settings span a wide range of traffic intensities. Figure (20b) compares the location schemes produced by the proposed MRT-AsmApX, the benchmark MRT-Bmk1, and the brute-force optimum as λ increases. Consistent with Figure (6a) in the main paper, MRT-AsmApX closely tracks the brute-force solution across all traffic levels, whereas MRT-Bmk1 fails to adjust appropriately as λ grows. Finally, Figure (20c) complements Figure (6c) of the main paper by comparing the actual average response times of MRT-AsmApX and the brute-force solutions as the number of demand nodes I' increases. The results confirm that MRT-AsmApX continues to deliver near-optimal performance even for larger problem instances.

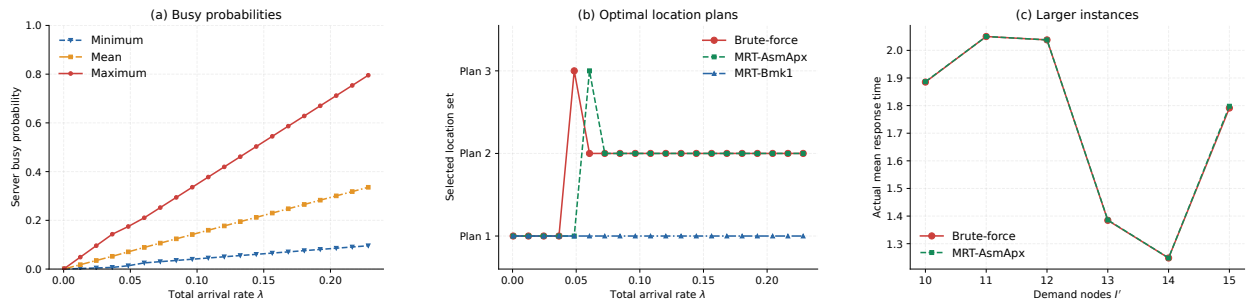


Figure 20 Performance comparison between the proposed MRT-AsmApX, benchmark MRT-Bmk1, and brute-force schemes. (a) Busy probabilities of the three servers under the brute-force optimum as the total arrival rate λ varies. (b) Location schemes obtained from the three methods for varying λ . (c) Average response times of MRT-AsmApX and brute-force solutions as the number of demand nodes I' increases to 15.

Table 4 KS tests of Unif(0, 1) for within-window call arrival times by two-hour interval.

Interval	n	KS stat.	p -value
00:00–02:00	36	0.142	0.426
02:00–04:00	24	0.164	0.489
04:00–06:00	48	0.0942	0.753
06:00–08:00	186	0.0640	0.415
08:00–10:00	174	0.0934	0.0901
10:00–12:00	192	0.0755	0.213
12:00–14:00	300	0.0429	0.623
14:00–16:00	234	0.0643	0.276
16:00–18:00	282	0.0649	0.178
18:00–20:00	420	0.0281	0.886
20:00–22:00	204	0.0816	0.125
22:00–24:00	168	0.0824	0.193

Appendix G: Supplementary materials for §5

G.1. Tests of Poisson arrivals and robustness checks on mission-duration distributions

Tests of Poisson arrivals. To assess whether call arrivals can be well approximated by a homogeneous Poisson process within each two-hour window, we follow the approach in [Reynaud-Bouret et al. \(2014\)](#). Specifically, we pool all call arrival times within each two-hour window across days, rescale them to the unit interval, and apply a one-sample Kolmogorov–Smirnov (KS) test against Unif(0, 1). Under a homogeneous Poisson process within the window, rescaled times should be i.i.d. uniform. Table 4 reports, for each two-hour interval, the number of calls (n), the KS statistic, and the p -value. Across the 12 windows, all p -values exceed 0.05 (and are far above a Bonferroni-adjusted threshold of $0.05/12 \approx 0.0042$), providing no evidence against within-window homogeneity. We therefore proceed using a homogeneous Poisson model within each two-hour interval.

Sensitivity of mission-duration distributions. We assess sensitivity of our case study results with respect to the mission-duration distributions by perturbing (i) the on-scene release time by ± 2 minutes and (ii) the *realized* travel times by $\pm 20\%$ in simulation. Concretely, we consider four stress-test scenarios combining the two perturbations: (-2 min, -20%), (-2 min, +20%), (+2 min, -20%), and (+2 min, +20%). For each scenario, we re-run the simulation using the location-allocation scheme produced by our proposed MOL-AsmApX and the current scheme, as shown in Figure 8, for 100 replications. We then report the annual overage loss with the widths of 95% CIs in parentheses. As summarized in Table 5, the scheme recommended by our model consistently and substantially outperforms the current scheme in all four cases, confirming the robustness of our results.

Table 5 Robustness checks under joint perturbations of on-scene time and realized travel times. Entries report annual overage loss (units consistent with Table 2 in the main text). The widths of 95% confidence intervals are reported in parentheses.

	(-2 min, -20%)	(-2 min, +20%)	(+2 min, -20%)	(+2 min, +20%)
Proposed MOL-AsmApX	8.18(1.03)	72.55(4.59)	11.95(1.51)	77.03(3.96)
Current scheme	32.82(2.53)	241.04(7.61)	35.46(2.90)	242.66(9.19)

G.2. Endogenous Determination of server number and validation of light-traffic regimes

In the following, we show that cost-minimizing staffing can *endogenously* induce light traffic in the lift-trap case study (§5). For day-shift operations, we solve the proposed model MOL-AsmApX under a staffing constraint of $J \in \{1, 2, 3, 4, 5\}$ rescue teams. For each J , we evaluate the resulting location–allocation scheme via 100 simulation replications under the same setting as in §5 and record: (i) the *annual penalty* from late rescues in day shift, and (ii) the *average busy probability* of servers (averaged across teams). Table 6 summarizes these performance metrics by J , showing a sharp drop in annual penalties when increasing staffing from $J = 1$ to $J = 2$.

To examine cost trade-offs, one can combine Table 6 with a per-team annual staffing cost c and choose J^* to minimize total cost, namely, cJ plus annual penalties. Using the table’s values, staffing $J = 2$ teams—the current practice in the case study of §5—is more cost-effective than $J = 1$ so long as $c < 149,300$ Singapore dollars ($\approx 115,067$ US dollars). For comparison, the median gross annual income of full-time employed residents in Singapore in 2024 was 66,000 Singapore dollars. Thus, the cost-minimizing J^* would plausibly place the system in the light-traffic regime, with an average busy probability below 2.5%. This outcome arises because late-service penalties are substantial in the lift-trap setting (\$2,500 Singapore dollars per case) and call arrival rates are low. These observations support our statement in §1 that, in emergency settings, light traffic can emerge endogenously when staffing is chosen optimally to minimize costs.

Table 6 Day shift: simulated annual penalties (in Singapore dollars) and average busy probability under $J \in \{1, 2, 3, 4, 5\}$ rescue teams. The case study setting is described in §5 in details. The widths of 95% confidence intervals are reported in parentheses.

	$J=1$	$J=2$	$J=3$	$J=4$	$J=5$
Annual penalties	250,700 (10,138)	101,400 (6,478)	48,250 (4,105)	31,575 (3,780)	17,250 (2,461)
Average busy probability	0.0538 (0.0028)	0.0247 (0.0018)	0.0136 (0.0016)	0.0095 (0.0012)	0.0074 (0.0011)

References

- Iglehart DL (1973) Weak convergence in queueing theory. *Advances in Applied Probability* 5(3):570–594.
- Adan I, Haviv M (2009) Conditional ages and residual service times in the $M/G/1$ queue. *Stochastic Models* 25(1):110–128.
- Kerner Y (2008) The conditional distribution of the residual service time in the $M_n/G/1$ queue. *Stochastic Models* 24(3):364–375.
- Sigman K, Yechiali U (2007) Stationary remaining service time conditional on queue length. *Operations Research Letters* 35(5):581–583.