E-Companion for "A Queueing-Theoretic Framework for Evaluating Transmission Risks in Service Facilities During a Pandemic"

EC.1 Model Implications: Comparing the R_0^{sys} and $\mathbb{E}[N(N-1)]$ metrics

Perlman and Yechiali (2020) propose $\mathbb{E}[N(N-1)]$ (where N is the number of customers in the system) as a measure of public health risk in a queueing model of a grocery store. This metric gives the time average number of *pairs* of customers in the system. Each pair represents a possible transmission, as one customer in the pair may be infectious and the other susceptible. While this metric represents a reasonable candidate for studying public health risks due to congestion in queueing systems (especially, as existing methods exist for its computation across a variety of system settings), our R_0^{sys} metric captures certain features of disease transmission (and therefore distinguish between varying levels of risk) in ways that the $\mathbb{E}[N(N-1)]$ metric does not. In particular, there exist systems with identical $\mathbb{E}[N(N-1)]$ values but considerably different R_0^{sys} values, which makes $\mathbb{E}[N(N-1)]$ unable to assess the efficacy of some interventions.

The following example, although artificial, illustrates the primary shortcoming of the $\mathbb{E}[N(N-1)]$ metric, which is that it does not account for transmission times and thresholds. Consider a $D^m/D^m/1$ system with deterministic arrival and service processes where m customers arrive simultaneously every $1/\lambda$ time units (i.e., exactly at times $t=0,1/\lambda,2/\lambda,\ldots$), and a server serves the m customers as a batch after they have been in the system for $1/\mu$ time units (where $\mu>\lambda$ in order to ensure system stability), so that the number of customers in the system at time t is given by $N(t)=m$ during time intervals $(0,1/\mu)$, $(1/\lambda,1/\lambda+1/\mu)$, $(2/\lambda,2/\lambda+1/\mu)$, etc., and $N(t)=0$ during the time intervals $(1/\mu,1/\lambda)$, $(1/\lambda+1/\mu,2\lambda)$, $(2/\lambda+1/\mu+3/\lambda)$, etc. It is straightforward to obtain $\mathbb{E}[N(N-1)]=\lambda m(m-1)/\mu$ for this system.

Meanwhile, we can compute R_0^{sys} as follows: if we assume that one arrival is infectious, while others are susceptible, then the infectious arrival's sojourn will coincide with the $m-1$ other customers who arrived (and will depart) with that customer. Their sojourn will not overlap with that of any other customers. Hence, by the linearity of expectation, the infectious customer will infect $R_0^{\mathsf{sys}}=(m-1)\mathbb{P}(\theta\geq 1/\mu)$ other customers on average. Now assume that each arrival is infectious, independent of all others, with some probability $p\ll 1/m$, allowing us to safely ignore the possibility of there ever being more than one infectious arrival in the system. Then, since this system features an infectious customer arrival rate of $pm\lambda$, the rate at which customers become infected (under our transmission assumptions) is

$$pm\lambda R_0^{\mathsf{sys}}=p\lambda m(m-1)\mathbb{P}(\theta\geq 1/\mu)=p\mu\mathbb{P}(\theta\geq 1/\mu)\mathbb{E}[N(N-1)].$$

Hence, under our transmission assumptions, within this context, $\mathbb{E}[N(N-1)]$ is an appropriate choice for evaluating the efficacy of various interventions if and only if $p\mu\mathbb{P}(\theta \geq 1/\mu)$ remains (at least nearly) unchanged as a result of these interventions. If we further assume that $\theta \sim \mathrm{Exp}(\alpha)$, the aforementioned quantity becomes $p\mu\left(1-e^{-\alpha/\mu}\right)$. Presumably, most *operational* interventions will leave p and α unchanged, and moreover $p\mu(1-e^{-\alpha/\mu}) \approx p\alpha$ when $\alpha \ll \mu$. So, in a regime where transmission takes a very long time on average relative to the service rate, $\mathbb{E}[N(N-1)]$ can even be used to assess interventions that change μ (so long as it is kept in this regime). But, when α/μ is not negligibly small, then interventions that change μ may not be adequately assessed by the $\mathbb{E}[N(N-1)]$ metric. For example, an intervention which leads to both doubling λ and μ, which leaves $\mathbb{E}[N(N-1)]$ unchanged, can actually have a significant effect on reducing transmissions, as individual customers cut their exposure times in half.

In fact under the exponential dose response model (i.e., when $\theta \sim \mathrm{Exp}(\alpha)$), what the $\mathbb{E}[N(N-1)]$ metric captures is a measure of risk that measures the number of times infection events would have occurred under the assumption that a customer who has already become infected can become infected again (during the same sojourn in the service facility). This type of framework would suggest that spending 100 time units with one infected individual is 100 times worse (in terms of some expected healthcare risk) than spending 1 unit of time with the same individual. Note however that the likelihood of becoming infected more than once in this hypothetical sense is equal to the likelihood of experiencing two arrivals in a Poisson process with rate α in an interval of length $1/\mu$ which is $o(\alpha/\mu)$, and hence, negligible when $\alpha \ll \mu$, which explains why an $\mathbb{E}[N(N-1)]$-driven analysis agrees with our R_0^{sys}-driven analysis in this regime.

While we have focused on an artificial model in our discussion, it is important to note that the differences between the $\mathbb{E}[N(N-1)]$ and R_0^{sys} metrics persist across a variety of settings (although their quantitative relationships can be setting-dependent). We will discuss one other crucial difference between these metrics is that due to its abstraction of transmission dynamics, the $\mathbb{E}[N(N-1)]$ can exhibit insensitivity to scheduling policies in the presence of exponentially distributed service requirements, making it unsuitable for assessing scheduling-based interventions (see Section 5.2).

EC.2 Proofs

EC.2.1 Proof of Proposition 1

Given that the IC arrives to a system seeing state s, we find the expected number of customers who will become infected by the IC *among those present in the system upon the IC's arrival.* In our model, the IC infects customer $i \in \{1, 2, \ldots, n(s)\})$ if and only if $W_i^{(s)} \geq \theta_i$. Hence, it follows from the linearity of expectation that the expected number of customers that IC will infect among the $n(s)$ customers follows:

$$\mathbb{E}\left[\sum_{i=1}^{n(s)} I\left\{W_i^{(s)} \geq \theta_i\right\}\right] = \sum_{i=1}^{n(s)} \mathbb{E}\left[I\left\{W_i^{(s)} \geq \theta_i\right\}\right] = \sum_{i=1}^{n(s)} \mathbb{P}\left(W_i^{(s)} \geq \theta_i\right).$$

Next, we assume that the system is in steady state (rather than assuming it is in a given state s). We then condition on the state $s \in \mathcal{S}$ observed by the IC upon their arrival, which allows us to deduce that the expected number of (pre-existing) customers that will be infected by the IC is

$$\sum_{s \in \mathcal{S}} \pi(s) \sum_{i=1}^{n(s)} \mathbb{P}\left(W_i^{(s)} \geq \theta\right). \tag{EC.1}$$

We prove Eq. 1 by arguing that R_0^{sys} is precisely equal to twice the quantity given above in Display (EC.1) by way of a symmetry argument, where we show that the distribution of the number of customers that the IC infects among those who arrived to the system *before* the IC is equal to (but not necessarily independent of) to the distribution of customers that the IC infects among those who arrived to the system *after* the IC. For further details, see Appendix EC.2. □

For any arbitrary value of $t > 0$, consider the number of other customers—who must all be susceptible by assumption—$\mathbf{Ov}(t)$ that the IC's sojourn overlaps with for a nonzero duration of time that is less than or equal to t. If we further let $\mathbf{OvB}(t)$ and $\mathbf{OvA}(t)$ be the set of such customers that arrived *before* and *after* the IC, respectively, then $\mathbf{Ov}(t) = \mathbf{OvB}(t) + \mathbf{OvA}(t)$ (since we are assuming Poisson arrivals, no other customer's arrival coincided with the instant of the IC's arrival with probability 1, so each other customer arrived either before or after the IC). Moreover, since the system is ergodic (and assumed to be in steady state) and since infectious customers are *functionally indistinguishable* from their susceptible counterparts, $\mathbf{OvB}(t)$ and $\mathbf{OvA}(t)$ are distributed in the same way (although not necessarily independent); this is because given any pair of customers, where one is susceptible and the other is infected, it is equally likely that that infected one arrived before or after the other, and the same is true when conditioning on the pair's sojourn time overlap.

With this new notation and the fact that $\mathbf{OvA}(t)$ and $\mathbf{OvB}(t)$ are distributed in the same way, we can rewrite R_0^{sys} as follows:

$$R_0^{\mathsf{sys}} = \mathbb{E}\left[\int_0^\infty \mathbb{P}(t \geq \theta)\, d\mathbf{Ov}(t)\right] = \mathbb{E}\left[\int_0^\infty \mathbb{P}(t \geq \theta)\, d(\mathbf{OvA}(t) + \mathbf{OvB}(t))\right] = 2\mathbb{E}\left[\int_0^\infty \mathbb{P}(t \geq \theta)\, d\mathbf{OvB}(t)\right],$$

where the expectation operators are needed as for any given value of t, the $\mathbf{Ov}(t)$, $\mathbf{OvA}(t)$, and $\mathbf{OvB}(t)$ are random variables. Now observe that the last equality describes twice the expected number of customers that the IC will infect among those already present in the system, which is precisely twice the quantity given in Display (EC.1), completing the proof.

EC.2.2 Proof of Corollary 1

Eq. (2) follows directly from Eq. (1) noting that (i) for any random variable X for which a Laplace transform $\widetilde{X}$ exists and any random variable $\theta \sim \mathrm{Exp}(\alpha)$ independent of X, we have $\mathbb{P}(X < \theta) = \widetilde{X}(\alpha)$ (see exercise 25.7 of Harchol-Balter 2013) and (ii) $\mathbb{E}[N] = \sum_{s\in\mathcal{S}} n(s)\pi(s)$.

EC.2.3 Proof of Proposition 2

This result is derived by combining traditional M/M/1 analysis together with an examination of $W_i^{(s)}$. Since the service policy is FCFS, $W_i^{(s)}$ denotes the *remaining* sojourn time of customer i given that there are s customers in the system when the infected customer arrives, and so, $W_i^{(s)} \sim \mathrm{Erlang}(i,\mu)$ for all $i \in \{1,2,\ldots,s\}$. Hence, recalling that $\eta \equiv \alpha/\mu$

$$\widetilde{W_i^{(s)}}(\alpha) = \left(\frac{\mu}{\alpha+\mu}\right)^i = \left(\frac{1}{1+\eta}\right)^i,$$

which together with the fact that $\pi(s) = (1-\rho)\rho^s$ and $\mathbb{E}[N] = \rho/(1-\rho)$ in an M/M/1 system, let us use Eq. (2) to prove the claim as follows:

$$\begin{aligned} R_0^{\mathsf{sys}} &= 2\left(\frac{\rho}{1-\rho} - (1-\rho)\sum_{s=0}^{\infty}\rho^s \sum_{i=1}^{s}\left(\frac{1}{\eta+1}\right)^i\right) \\ &= 2\left(\frac{\rho}{1-\rho} - \left(\frac{1-\rho}{\eta}\right)\sum_{s=0}^{\infty}\rho^s\left(1-\left(\frac{1}{1+\eta}\right)^s\right)\right) = 2\left(\frac{\rho}{1-\rho}\right)\left(\frac{\eta}{\eta+1-\rho}\right). \end{aligned}$$

The claims that R_0^{sys} and $\lambda R_0^{\mathsf{sys}}$ are convex increasing, convex decreasing, and concave increasing in λ, μ, and α, respectively, can be verified in a straightforward manner by taking first and second derivatives.

EC.2.4 Proof of Proposition 3

It is known that for an M/M/c system

$$\pi(s) = \begin{cases} \dfrac{c!(1-\rho)}{s!(c\rho)^{c-s}}C(c,\rho) & 0 \le s \le c \\ \dfrac{1-\rho}{\rho^{c-s}}C(c,\rho) & s > c, \end{cases}$$

and $\mathbb{E}[N] = \dfrac{\rho}{1-\rho}C(c,\rho) + c\rho$ (Harchol-Balter 2013, chapter 14). These two facts, together with the claimed values of $\widetilde{W_i^{(s)}}(\alpha)$ and Eq. (2), yield

$$\begin{aligned} R_0^{\mathsf{sys}} &= 2\left(\mathbb{E}[N] - \sum_{s=0}^{\infty}\pi(s)\sum_{i=1}^{s}\widetilde{W_i^{(s)}}(\alpha)\right) \\ &= 2\left(\mathbb{E}[N] - \sum_{s=0}^{c}\pi(s)\sum_{i=1}^{s}\widetilde{W_i^{(s)}}(\alpha) - \sum_{s=c+1}^{\infty}\pi(s)\left(\sum_{i=1}^{c}\widetilde{W_i^{(s)}}(\alpha) + \sum_{i=c+1}^{s}\widetilde{W_i^{(s)}}(\alpha)\right)\right) \\ &= 2\left(\left(\frac{\rho}{1-\rho}\right)C(c,\rho) + c\rho - \frac{1}{\eta+2}\left(C(c,\rho)\left(\frac{2c\rho - c\eta}{\eta+c-c\rho}\right) + 2c\rho\right)\right), \end{aligned}$$

as claimed.

It remains only to prove that the sojourn time overlap between the IC and customer i is distributed according to

$$W_i^{(s)} \sim \begin{cases} \text{Exp}(2\mu) & 1 \le s < c \\ \min(\text{Erlang}(s-c+1,(c-1)\mu)+\text{Exp}(\mu),\text{Exp}(\mu)) & i \le c \le s \\ \text{Erlang}(i-c,c\mu)+W_c^{(s-(i-c))} & c < i \le s \end{cases},$$

where all component distributions above are independent of one other, and (as claimed in Proposition 3) takes the following Laplace Transform:

$$\widetilde{W}_i^{(s)}(\alpha) = \begin{cases} 2/(\eta+2) & 1 \le s < c \\ \left(\eta\left(\dfrac{c-1}{\eta+c}\right)^{s-c+1}+\eta+2\right)\left(\dfrac{1}{\eta^2+3\eta+2}\right) & i \le c \le s \\ \left(\dfrac{c}{\eta+c}\right)^{i-c}\left(\eta\left(\dfrac{c-1}{\eta+c}\right)^{s-i+1}+\eta+2\right)\left(\dfrac{1}{\eta^2+3\eta+2}\right) & c < i \le s \end{cases},$$

for all $s \in \mathcal{S}$ and $i \in \{1,2,\ldots,s\}$,

We address case 1 (i.e., when $1 \le s < c$), case 2 (i.e., when $i \le c \le s$), and case 3 (i.e., when $c < i \le s$), separately and sequentially.

Under case 1 (i.e., when $1 \le s < c$), the IC's service starts upon arrival, seeing some customer i who is already in service at that time. Therefore, the IC's sojourn overlaps with that of customer i for an amount of time that is whichever is less of the service time of the IC (call this X) or the *remaining* service time of customer i (call this X_i), i.e., $W_i^{(s)} = \min(X, X_i)$. Clearly, $X \sim \text{Exp}(\mu)$, but we must also have $X_i \sim \text{Exp}(\mu)$, due to the memoryless property of the exponential distribution. Moreover, since X and X_i are independent, we must have $W_i^{(s)} = \min(X, X_i) \sim \text{Exp}(2\mu)$ and $\widetilde{W}_i^{(s)}(\alpha) = 2/(\eta+2)$ as claimed.

Under case 2 (i.e., when $i \le c \le s$), the IC arrives at position $s-c+1$ of the *queue* (i.e., so that the IC will enter service at one of the c servers after the system experiences $s-c+1$ service departures), while customer i is already in service. In this case, the IC's sojourn overlaps with that of customer i for an amount of time equal to whichever is less of the *sojourn* time of the IC or the *remaining* service time of customer i (call this X_i; $X_i \sim \text{Exp}(\mu)$ as in the previous case). The sojourn time of the IC is $Y+X$ (so that $W_i^{(s)} = \min(Y+X, X_i)$), where Y and X are the durations of time the IC spends in the *queue* and *in service*, respectively. The random variable Y (the distribution of which depends on s) corresponds to the time it takes for $s-c+1$ successive departures (from any of the c servers), while $X \sim \text{Exp}(\mu)$ as in the previous case. Note that the IC is in the queue during the entire time it takes for these $s-c+1$ successive departures (that makes up Y) to take place, and hence, all c servers are busy during this time, and so these $s-c+1$ successive departures will each take up an amount of time that is drawn from the Exp$(c\mu)$ distribution, and since all such "inter-departure" times are independent (as service times are independent and exponentially distributed), we have $Y \sim \text{Erlang}(s-c+1, c\mu)$. Note, however, that Y and X_i are *not* independent (while X is independent of both Y and X_i), because a departure from the system may actually be due to customer i being served. We can alternatively view $W_i^{(s)} = \min(Y' + X, X_i)$, where $Y' \sim \text{Erlang}(s-c+1, (c-1)\mu)$ is the time it takes for $s-c+1$ successive departures to occur assuming only $c-1$ servers are running (i.e., ignoring the server on which job i is running); in this case Y' and X_i are independent, and we have $W_i^{(s)} \sim \min(\text{Erlang}(s-c+1, (c-1)\mu) + \text{Exp}(\mu), \text{Exp}(\mu))$, as claimed.

We obtain $\widetilde{W}_i^{(s)}$ using first-step analysis by observing that $W_i^{(s)}$ is in fact distributed according to a Coxian phase-type distribution (See Harchol-Balter 2013, Chapter 21.1, for details) with $s-c+2$ phases, all but the last of which have a rate of $c\mu$ as they correspond to a service at any of the c servers; at the conclusion of each of these phases the entire process terminates with probability $1/c$ (corresponding to customer i's service, as this would conclude the sojourn overlap) or continue to the next phase (correspond to a departure due to any customer in service other than customer i, as this would advance the IC one position in the queue, or bring them into service if they previously at the head of the queue). The last phase has a rate of 2μ as it corresponds only to the service of either the IC or server i (either of which would conclude the sojourn overlap). Denote by U_m the remaining duration of such a distribution given that we currently have m phases left to go after the current phase (assuming the process does not terminate early), so that $U_m = X_m + (c-1)U_{m-1}/c$, for all $m \geq 1$ where $U_m \sim \text{Exp}(c\mu)$, while $U_0 \sim \text{Exp}(2\mu)$. Clearly, $W_i^{(s)} = U_{s-c+1}$. Using standard manipulations of Laplace Transforms (See Harchol-Balter 2013, Chapter 25, for details) and recalling that $\eta \equiv \alpha/\mu$, we have

$$\widetilde{U}_m(\alpha) = \left(\frac{1}{\eta+c}\right)\left(1 + (c-1)\widetilde{U}_{m-1}(\alpha)\right)$$

for all $m \geq 1$, and $\widetilde{U}_0(\alpha) = 2\mu/(\alpha+2\mu) = 2/(\eta+2)$. Solving this linear recursion (and recalling that $\eta \equiv \alpha/\mu$) yields

$$\widetilde{U}_m(\alpha) = \left(\eta\left(\frac{c-1}{\eta+c}\right)^m + \eta + 1\right)\left(\frac{1}{\eta^2+3\eta+2}\right),$$

which coincides with the claimed value of $\tilde{W}_i^{(s)}(\alpha)$ for case 2, when we set $m = s-c+1$, thus verifying the claim.

Under case 3 (i.e., when $c < i \leq s$), the IC arrives at position $s-c+1$ of the queue, and finds customer i in position $i-c$ of the queue. We break up the sojourn time overlap between the IC and customer i into two parts: the duration of time their sojourns overlap while both the IC and customer i are present in the *queue* (call this Q, which depends on i), and the remaining portion of the sojourn time overlap, which corresponds to the duration of time their sojourns overlaps while customer i is in service (call this V, and note that the IC may—but need not necessarily be—in service during some of this time). The first of these durations corresponds to the time it takes for the system to experience $i-c$ consecutive departures, so $Q \sim \text{Erlang}(i-c, c\mu)$. Meanwhile, when customer i enters service, the IC will be in position $s-c+1-(i-c) = s-i+1$ of the queue, and hence, the remaining sojourn time overlap, V, will be the same as the total sojourn time overlap between a customer who had arrived to position $s-i+1$ of the queue (i.e., who had arrived to a system with $s-(i-c)$ *other* customers already present in the system), while customer i was in service (e.g., in position c). That is, $V \sim W_c^{(s-(i-c))}$; also note that Q and V are independent. It follows that in this case $W_i^{(s)} = Q + V \sim \text{Erlang}(i-c, c\mu) + W_c^{(s-(i-c))}$ as claimed. Moreover, it follows that $\widetilde{W}_i^{(s)}(\alpha) = \widetilde{Q}(\alpha)\widetilde{V}(\alpha) = (c/(\eta+c))^{i-c}\widetilde{W}_c^{(s-(i-c))}(\alpha)$. Substituting in the expression for $\widetilde{W}_c^{(s-(i-c))}$ from case 2 shows that $\widetilde{W}_i^{(s)}(\alpha)$ is also as claimed in case 3.

EC.2.5 Proof of Proposition 4

The buffer size does not affect the distribution of the sojourn time overlap under the FCFS policy. Therefore, $\widetilde{W}_i^{(s)}(\alpha)$ remains the same as in the case of Proposition 3. Also, the steady-state probability distribution of an M/M/c/k system is known to follow Eq. (4) (Shortle et al. 2018, Chapter 3). Using these in Eq. (2) we can establish the claimed result.

EC.2.6 Proof of Proposition 5

We first outline the expressions for the values appearing in Proposition 5, and then we prove the proposition.

In the setting considered in Proposition 5, let $\pi(h, \ell, \tau)$ be the limiting probability distribution of the number of high- and low-risk customers and the type ($\tau \in \{\mathsf{H}, \mathsf{L}\}$) of customer who is currently in service under steady state of the system (see Marks (1973) for the algorithm to compute it), the expressions for the five terms are given in the following Proposition:

Proposition EC.2.1 *In the M/M/1 system with non-preemptive priorities described above, we have*

$$R_0^{\mathsf{H}\xrightarrow{\mathbf{B}}\mathsf{H}} = \sum_{h=1}^{\infty}\sum_{\ell=0}^{\infty}\left[\pi(h,\ell,\mathsf{H})\sum_{i=1}^{h}\left(1-\left(\frac{1}{1+\eta}\right)^{i}\right) + \pi(h,\ell,\mathsf{L})\mathbf{1}_{\ell\geq 1}\sum_{i=1}^{h}\left(1-\left(\frac{1}{1+\eta}\right)^{i+1}\right)\right] \tag{EC.2}$$

$$R_0^{\mathsf{H}\xrightarrow{\mathbf{B}}\mathsf{L}} = \sum_{h=0}^{\infty}\sum_{\ell=1}^{\infty}\left[\pi(h,\ell,\mathsf{H})\mathbf{1}_{h\geq 1}\sum_{i=1}^{\ell}\left(1-\left(\frac{1}{1+\eta}\right)^{h+1}\right) + \pi(h,\ell,\mathsf{L})\left(1-\frac{1}{1+\eta}+\sum_{i=2}^{\ell}\left(1-\left(\frac{1}{1+\eta}\right)^{h+2}\right)\right)\right] \tag{EC.3}$$

$$R_0^{\mathsf{L}\xrightarrow{\mathbf{B}}\mathsf{H}} = R_0^{\mathsf{H}\xrightarrow{\mathbf{B}}\mathsf{H}} \tag{EC.4}$$

$$
\begin{aligned}
R_0^{\mathsf{L}\overset{\mathbf{B}}{\to}\mathsf{L}} = & \sum_{h=0}^{\infty}\sum_{\ell=1}^{\infty}\left[\pi(h,\ell,\mathsf{H})\mathbf{1}_{h\geq 1}\sum_{i=1}^{\ell}\left(1-\widetilde{W}_h(d)\Big(\frac{1}{1+\eta}\Big)^i\Big(\widetilde{W}(d)\Big)^{i-1}\right)\right] \\
& + \sum_{h=0}^{\infty}\sum_{\ell=1}^{\infty}\left[\pi(h,\ell,\mathsf{L})\left(1-\frac{1}{1+\eta}+\sum_{i=2}^{\ell}\left(1-\widetilde{W}_{h+A_S}(d)\Big(\frac{1}{1+\eta}\Big)^i\Big(\widetilde{W}(d)\Big)^{i-2}\right)\right)\right] \qquad \text{(EC.5)}
\end{aligned}
$$

$$
R_0^{\mathsf{L}\overset{\mathbf{A}}{\to}\mathsf{H}} = \sum_{h=0}^{\infty}\sum_{\ell=1}^{\infty}[\pi(h,\ell,\mathsf{H})\mathbf{1}_{h\geq 1}A_1 + \pi(h,\ell,\mathsf{L})A_2]. \qquad \text{(EC.6)}
$$

where $d=\alpha+\lambda_{\mathsf{H}}-\lambda_{\mathsf{H}}\widetilde{B}(\alpha)$, $\widetilde{B}(\alpha)=\frac{1}{2\lambda_{\mathsf{H}}}\left(\lambda_{\mathsf{H}}+\mu+\alpha-\sqrt{(\lambda_{\mathsf{H}}+\mu+\alpha)^2-4\lambda_{\mathsf{H}}\mu}\right)$, *and*

$$
A_1=\sum_{n=1}^{\infty}\left[V(h,n)\left(1-\Big(\frac{1}{1+\eta}\Big)^{n+1}\right)+(\ell-1)V(1,n)\left(1-\Big(\frac{1}{1+\eta}\Big)^{n+1}\right)\right]+\frac{\lambda_H}{\mu}\left(1-\frac{1}{1+\eta}\right)
$$

$$
A_2=\sum_{n=1}^{\infty}\left[V(h+1,n)\left(1-\Big(\frac{1}{1+\eta}\Big)^{n+1}\right)+(\ell-2)V(1,n)\left(1-\Big(\frac{1}{1+\eta}\Big)^{n+1}\right)\right]+\frac{\lambda_H}{\mu}\left(1-\frac{1}{1+\eta}\right).
$$

Also

$$
\widetilde{W}(s)=\hat{A}_S\left(\widetilde{S}(s)\right)=\widetilde{S}\left(\lambda_{\mathsf{H}}\left(1-\frac{\mu}{\mu+s}\right)\right)=\frac{\mu(\mu+s)}{\mu(\mu+s)+\lambda_{\mathsf{H}}s}
$$

$$
\widetilde{W}_{h+A_S}(s)=\left(\frac{\mu}{\mu+s}\right)^h\widetilde{S}\left(\lambda_{\mathsf{H}}\left(1-\frac{\mu}{\mu+s}\right)\right)=\frac{\mu\Big(\frac{\mu}{\mu+s}\Big)^h}{\mu+\lambda_{\mathsf{H}}\left(1-\frac{\mu}{\mu+s}\right)}
$$

note that when $h=0$, $\widetilde{W}_{h+A_S}(s)=\widetilde{W}(s)$.

Proof of Proposition 5. We first consider the expected number of high-risk customers who were already in the system being infected when the IC arrives. No matter being which type of customer, this IC will be served after all the high-risk customers who were already in the system, so each pair of sojourn time overlap ends when the high-risk SC within the pair leaves the system. Note that when the IC arrives, if a low-risk customer is currently in service, then all the high-risk SCs have to experience one more service duration for their remaining sojourn time due to the non-preemptive policy. Therefore, Eqs. (EC.2) and (EC.4) follow directly after conditioning on different system states and applying Eq. (EC.1). Then we consider the expected number of low-risk SCs who were already in the system being infected when the high-risk IC arrives, similarly, we have Eq. (EC.3) since if a low-risk SC is currently in service, then this SC's sojourn time overlap with the IC will end when she finishes the service while that of other SCs will end when the IC leaves the system since they have lower priority and the SCs will experience one more service duration in this situation for their sojourn time overlaps with the IC; if the customer currently in service is high-risk, then all sojourn time overlaps with low-risk SCs end when this high-risk IC leaves. In the case of considering the expected number of low-risk SCs who were already in the system being infected when a low-risk IC arrives, the sojourn time overlaps with the low-risk SCs will only depend on the departure time of the SCs. We have two situations: (i) the customer in service is high-risk, then for the i-th low-risk SC, the time until she leaves the system consists of three parts: the busy period of high-risk customers started by the first h high-risk customers, all the service times of the i low-risk SCs (including herself), and $(i-1)$ busy periods started by all the work created during each low-risk SC's service time (i.e. the time gaps between any two consecutive low-risk SCs of receiving the service); (ii) the customer in service is low-risk, then the pair of sojourn time overlap with the low-risk SC in service will end when the SC finished the service while the other overlap times of the i-th low-risk SC will include three parts: the busy period started by the work created by h high-risk customers and the arrivals of high-risk customers during the first low-risk SC's service time, all the service times of the i low-risk SCs (including herself), and $(i-2)$ busy periods started by all the work created during each low-risk SC's service time (i.e. the time gaps between any two consecutive low-risk SCs of receiving the service). So we let W, W_h, and W_{h+A_S} denote the length of busy period started by the work created by the number of high-risk arrivals in one service duration, the length of busy period started by the work of h number of service duration, and the length of busy period started by the work of h high-risk customers and the number of high-risk arrivals in one service duration respectively. Conditioning and rearranging terms yield Eq. (EC.5). We finish the proof by figuring out $R_0^{\mathsf{L}\xrightarrow{\mathbf{A}}\mathsf{H}}$ with the use of $V(x,y)$ introduced in the proof of Lemma 2 in Section EC.5. Recall that $V(x,y)$ gives the expected number of arrivals to an M/M/1 system with $\rho=\rho_H$ in state y during current busy period given the initial state x. In this case that there is ℓ low-risk customers in the system and the high-risk is currently in service when the low-risk IC arrives, we define the first busy period as the amount of time until the first low-risk customer enters the service, then the next busy period will be defined as the time until the second low-risk customer enters the service (note that at this moment the system will have no high-risk customer) and so on. Except the first busy period, the following busy periods will start when the low-risk customer enters the service and end when there is no more high-risk customer. Hence, the first busy period will contribute $\sum_{n=1}^{\infty} V(h,n)\left(1-\left(\frac{1}{1+\eta}\right)^{n+1}\right)$ and each following $(\ell-1)$ busy period except the last one will each contribute $\sum_{n=1}^{\infty} V(1,n)\left(1-\left(\frac{1}{1+\eta}\right)^{n+1}\right)$. For the last busy period, since the sojourn time overlaps will end earlier when the IC leaves the system, so it will contribute $\frac{\lambda_H}{\mu}\left(1-\frac{1}{1+\eta}\right)$ instead. Similarly, when there is ℓ low-risk customers in the system and the low-risk customer is currently in service when the low-risk IC arrives, the first busy period will contribute $\sum_{n=1}^{\infty} V(h+1,n)\left(1-\left(\frac{1}{1+\eta}\right)^{n+1}\right)$ (with one more low-risk customer) and the next $(\ell-2)$ following busy periods except the last one will each contribute $\sum_{n=1}^{\infty} V(1,n)\left(1-\left(\frac{1}{1+\eta}\right)^{n+1}\right)$ and the last one contributes $\frac{\lambda_H}{\mu}\left(1-\frac{1}{1+\eta}\right)$ which completes the proof.

EC.2.7 Proof of Proposition 6

We let $q = \dfrac{\lambda_{\mathsf{H}}}{\lambda_{\mathsf{H}} + \lambda_{\mathsf{L}}}$ $(0 < q < 1)$, for any set of parameters $0 < \lambda < \mu$ and $\forall\ 0 < f < 1$ which satisfy the stability conditions, we claim that $f = q$ yields the optimal (minimum) value of R_0^{sys}.

We can write the R_0^{sys} in this setting as follows:

$$R_0^{\mathsf{sys}} = 2q\left(\frac{\dfrac{q\lambda}{f\mu}}{1-\dfrac{q\lambda}{f\mu}}\right)\left(\frac{\eta}{\eta+1-\dfrac{q\lambda}{f\mu}}\right) + 2(1-q)\left(\frac{\dfrac{(1-q)\lambda}{(1-f)\mu}}{1-\dfrac{(1-q)\lambda}{(1-f)\mu}}\right)\left(\frac{\eta}{\eta+1-\dfrac{(1-q)\lambda}{(1-f)\mu}}\right).$$

Taking the first and second order derivatives gives:

$$\begin{aligned}\frac{\partial R_0^{\mathsf{sys}}}{\partial f} = 2\alpha\lambda\left(\frac{q^2}{(\lambda q - f\mu)(\lambda q - f(\alpha+\mu))} + \frac{f\mu q^2}{(\lambda q - f\mu)^2(\lambda q - f(\alpha+\mu))} + \frac{fq^2(\alpha+\mu)}{(\lambda q - f\mu)(\lambda q - f(\alpha+\mu))^2}\right)\\
-\frac{2\alpha\lambda(q-1)^2}{(-f\mu+\mu+\lambda(q-1))(\alpha+\alpha(-f)-f\mu+\mu+\lambda(q-1))}\\
-\frac{2\alpha\lambda(f-1)\mu(q-1)^2}{(-f\mu+\mu+\lambda(q-1))^2(\alpha+\alpha(-f)-f\mu+\mu+\lambda(q-1))}\\
-\frac{2\alpha\lambda(f-1)(q-1)^2(\alpha+\mu)}{(-f\mu+\mu+\lambda(q-1))(\alpha+\alpha(-f)-f\mu+\mu+\lambda(q-1))^2},\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^2 R_0^{\mathsf{sys}}}{\partial f^2} = \frac{4\alpha\lambda(1-f)(q-1)^2(\alpha+\mu)^2}{((q-1)\lambda+(1-f)\mu)\,((1-f)\alpha+(1-f)\mu-(1-q)\lambda)}\\
+\frac{(1-f)(q-1)^2\mu(\alpha+\mu)-(1-f)\mu(q-1)^2(\alpha+\mu)+(1-q)\lambda(q-1)^2(\alpha+\mu)}{((1-f)\mu-(1-q)\lambda)^2((1-f)\alpha+(1-f)\mu-(1-q)\lambda)^2}\\
+\frac{(1-f)(q-1)^2\mu^2-(1-f)\mu^2(q-1)^2+(1-q)\lambda(q-1)^2\mu}{((1-f)\mu-(1-q)\lambda)^3\,((1-f)\alpha+(1-f)\mu-(1-q)\lambda)}\\
+\frac{fq^2(\alpha+\mu)^2}{(q\lambda-f\mu)\,(q\lambda-f(\alpha+\mu))^3} + \frac{fq^2\mu(\alpha+\mu)+q^2(\alpha+\mu)(q\lambda-f\mu)}{(q\lambda-f\mu)^2(q\lambda-f(\alpha+\mu))^2}\\
+\frac{fq^2\mu^2+q^2\mu(q\lambda-f\mu)}{(q\lambda-f\mu)^3\,(q\lambda-f(\alpha+\mu))}.\end{aligned}$$

According to the stability conditions we know that $(1-f)\mu-(1-q)\lambda > (1-f)^2\mu-(1-q)\lambda > 0$ and $q\lambda - f\mu < q\lambda - f^2\mu < 0$. Clearly, it is straightforward to check $\dfrac{\partial^2 R_0^{\mathsf{sys}}}{\partial f^2} > 0$ with these conditions and $\left.\dfrac{\partial R_0^{\mathsf{sys}}}{\partial f}\right|_{f=q} = 0$. Hence, R_0^{sys} is convex in f and when $f = q$, R_0^{sys} achieves its minimum. Moreover, when $f = q$, it is clear that $\rho_{\mathsf{T}} = \dfrac{\lambda_{\mathsf{H}}+\lambda_{\mathsf{L}}}{\mu}$, therefore, the R_0^{sys} in this case is identical to that under the FCFS scheduling policy which completes the proof.

EC.2.8 Proof of Proposition 7

Before proving Proposition 7, we include the R_0^{sys} expression under PLCFS policy as well, the proof follows by the similar analysis in the proof of Proposition 2 and the distribution of $W_i^{(s)}$ is identically distributed as an M/M/1 busy period started by $\mathrm{Exp}(\mu)$ amount of work with system load ρ for all $s \in \mathcal{S}$ and $i \in \{1, 2, \ldots, s\}$ under PLCFS policy, we summarize the results in the following Proposition:

Proposition EC.2.2 *In an M/M/1/PLCFS system with arrival rate λ, service rate μ, load $\rho \equiv \lambda/\mu$, transmission threshold $\theta \sim \mathrm{Exp}(\alpha)$ and* normalized transmission rate $\eta \equiv \alpha/\mu$, $W_i^{(s)}$ *is identically distributed for all $s \in \mathcal{S}$ and $i \in \{1, 2, \ldots, s\}$, such that*

$$\widetilde{W}_i^{(s)}(\alpha) = \frac{\eta + 1 + \rho}{2\rho}\left(1 - \frac{\sqrt{(\eta + 1 - \rho)^2 + 4\eta\rho}}{\eta + 1 + \rho}\right),$$

while

$$R_0^{\mathsf{sys}} = \frac{-(\eta + 1 - \rho) + \sqrt{(\eta + 1 - \rho)^2 + 4\eta\rho}}{1 - \rho}.$$

Moreover, for an M/M/1 queue with exponentially distributed transmission thresholds.

Then we proceed to start the proof of Proposition 7 by considering any two work-conserving scheduling policies P_1 and P_2. Given any sample path of arrival times and service requirements, the scheduling policies P_1 and P_2 each yield their own sequence of departure times. Moreover, since we are assuming that service requirements are exponentially distributed (and independent of the arrival process), then we can couple sample paths with the same arrival sequences that also generate the same departure time sequences under P_1 and P_2 (although, the order in which jobs depart may differ across the two policies); this coupling is valid due to the memoryless property of exponential distributions.

Now observe that under the assumption that only a single infectious customer will arrive to the system, the infectious customer must arrive during some busy period. Therefore, whenever it is the case that during *any* busy period, featuring *any* number of jobs, with *any* sequence of departure times (coupled across P_1 and P_2) P_1 yields an *expected* number of transmissions no greater than that yielded by P_2, then P_1 yields an R_0^{sys} no greater than that yielded by P_2. Note that the *expectation* of the number of transmissions is taken over the randomness associated with the transmission thresholds, randomness associated with the sequence in which jobs will be served under the policy, and the identity of the infectious customer, which is equally likely to be any of the customers that arrived during that busy period. Given the fact that comparisons across sample paths are sufficient, the proof of the claimed result reduces to the proof of the following lemma.

Lemma 1 *For any integer $n \geq 1$ and any real numbers $a_1, a_2, \dots, a_n, d_1, d_2, \dots, d_n \geq 0$ such that (i) $a_1 < a_2 < \cdots < a_n$, (ii) $d_1 < d_2 < \cdots < d_n$, (iii) $d_k > a_{k+1}$ for all $k \in \{1, 2, \dots, n-1\}$, and (iv) $d_n > a_n$, consider a busy period of an M/M/1 system with n customers, where arrivals occur at times $a_1, a_2, \dots, a_n$, and departures occur at times $d_1, d_2, \dots, d_n$. If we assume that exactly one customer among these n customers is infectious and each of the n customers is equally likely to be the infectious one, and we further assume i.i.d. transmission thresholds drawn from the* $\mathrm{Exp}(\alpha)$ *distribution, then—among all work-conserving scheduling policies—the expected number of transmissions during this busy period is minimized under the* preemptive-last-come-last-served *(PLCFS) scheduling policy and maximized under the* first-come-first-served *(FCFS) scheduling policy; note that the departure times are fixed regardless of the choice of scheduling policy.*

The result is trivially true in the case where $n = 1$, as no infections are possible in busy periods with only one customer arrival and furthermore, scheduling policies are irrelevant in such busy periods, so we will henceforth assume that $n \geq 2$.

First, see that whenever we run the system under some policy, at the conclusion of the busy period, the process yields a unique bijective function $\sigma \colon \{1, 2, \dots n\} \to \{1, 2, \dots, n\}$ (i.e., a unique permutation on n elements, sigma) where σ designates the order in which customers depart the system, that is $d_{\sigma(1)} < d_{\sigma(2)} < \cdots < d_{\sigma(n)}$. Since scheduling policies can make use of randomness (or, e.g., information about past busy periods), *a priori* scheduling policy results in a probability distribution over such permutations (of which there are finitely many), although scheduling policies such as FCFS and PLCFS will result in a single specific permutation with probability 1.

Now consider a policy, P that does not behave exactly like PLCFS during this busy period. It follows that P that yields a specific permutation σ with some probability $p_\sigma > 0$ such that there exist $j, k \in \{1, 2, \dots, n\}$ where $j < k$ and $a_j < a_k < d_{\sigma(j)} < d_{\sigma(k)}$, i.e., under σ some arrival (the j-th) departs after some later arrival. Fix such a j and k and construct a new policy, $\mathsf{P}^{\sigma}_{j,k}$ that yields the same permutation of P, except whenever P would yield the permutation σ, $\mathsf{P}^{\sigma}_{j,k}$ instead yields the permutation $\sigma_{j,k}$, where $\sigma_{j,k}$ is defined as follows:

$$\sigma_{j,k}(i) = \begin{cases} i & i \in \{1, 2, \dots, n\} \backslash \{j, k\} \\ k & i = j \\ j & i = k. \end{cases}$$

That is, $\sigma_{j,k}$ acts like σ, except it applies the transposition that swaps the departure times of the j-th and k-th jobs that would be completed under σ. We will show that $\mathsf{P}^{\sigma}_{j,k}$ yields an expected number of transmissions no greater than that yielded by P.

Now let $O_{\ell,m}$ and $O^*_{\ell,m}$ denote the sojourn time overlap of the customers arriving at times a_ℓ and a_m under departure orders represented by the permutations σ and $\sigma_{j,k}$, respectively. That is, let $O_{\ell,m} \equiv \min(d_{\sigma(\ell)}, d_{\sigma(m)}) - \max(a_\ell, a_m)$ and $O^*_{\ell,m} \equiv \min(d_{\sigma_{j,k}(\ell)}, d_{\sigma_{j,k}(m)}) - \max(a_\ell, a_m)$. Since any *pair* of customers arriving during the busy period will consist of one IC and one SC with probability $2/n$ (as there is exactly one SC, and it is equally likely to be any of the customers), it follows that the expected number of transmissions during this busy period under the σ departure order is

$$\frac{2}{n}\sum_{\ell=1}^{n-1}\sum_{m=\ell+1}^{n}\left(1-e^{-\alpha\cdot O_{\ell,m}}\right) = n-1-\frac{2}{n}\sum_{\ell=1}^{n-1}\sum_{m=\ell+1}^{n} e^{-\alpha\cdot O_{\ell,m}}. \tag{EC.7}$$

The expected number of transmission during this busy period under the $\sigma_{i,j}$ departure order is naturally given by equivalent expressions in display (EC.7) if we replace $O_{\ell,m}$ by $O^*_{\ell,m}$ in each of the equivalent expressions. We note that $O_{\ell,m} = O^*_{\ell,m}$ whenever $\{\ell,m\} = \{j,k\}$ and whenever $\{\ell,m\} \subseteq \{1,2,\ldots,n\}\backslash\{j,k\}$. With this fact in mind, we proceed measure Δ the decrease in the mean number of transmissions when implementing policy $\mathsf{P}^\sigma_{j,k}$ rather than P (recalling that the two policies yield different departure orders with probability p_σ when the latter yields the departure order $\sigma_{j,k}$ rather than σ):

$$\begin{aligned}
\Delta &= p_\sigma\left(\left(n-1-\frac{2}{n}\sum_{\ell=1}^{n-1}\sum_{m=\ell+1}^{n} e^{-\alpha\cdot O_{\ell,m}}\right)-\left(n-1-\frac{2}{n}\sum_{\ell=1}^{n-1}\sum_{m=\ell+1}^{n} e^{-\alpha\cdot O^*_{\ell,m}}\right)\right)\\
&= \frac{2p_\sigma}{n}\sum_{i=1}^{n-1}\sum_{m=\ell+1}^{n}\left(e^{-\alpha\cdot O^*_{\ell,m}} - e^{-\alpha\cdot O_{\ell,m}}\right)\\
&= \frac{2p_\sigma}{n}\sum_{i=1}^{n}\left(e^{-\alpha\cdot O^*_{i,j}} + e^{-\alpha\cdot O^*_{i,k}} - e^{-\alpha\cdot O_{i,j}} - e^{-\alpha\cdot O_{i,k}}\right).
\end{aligned}$$

We will argue that every term of Δ is non-negative, and therefore, $\Delta \geq 0$, establishing that implementing $\mathsf{P}^\sigma_{j,k}$ reduces (or leaves unchanged) the mean number of transmissions as compared to implementing P. Before doing so, we will establish three important results. First, observe that for all $i \in \{1,2,\ldots,n\}$, we must have

$$\begin{aligned}
O_{i,j} + O_{i,k} &= \left(\min(d_{\sigma(i)}, d_{\sigma(j)}) - \max(a_i, a_j)\right) + \left(\min(d_{\sigma(i)}, d_{\sigma(k)}) - \max(a_i, a_k)\right)\\
&= \left(\min(d_{\sigma_{j,k}(i)}, d_{\sigma_{j,k}(k)}) - \max(a_i, a_j)\right) + \left(\min(d_{\sigma_{j,k}(i)}, d_{\sigma_{j,k}(j)}) - \max(a_i, a_k)\right)\\
&= \left(\min(d_{\sigma_{j,k}(i)}, d_{\sigma_{j,k}(j)}) - \max(a_i, a_j)\right) + \left(\min(d_{\sigma_{j,k}(i)}, d_{\sigma_{j,k}(k)}) - \max(a_i, a_k)\right)\\
&= O^*_{i,j} + O^*_{i,k},
\end{aligned}$$

from which it follows that $O^*_{i,j} = O_{i,j} + \delta_i$ and $O^*_{i,k} = O_{i,k} - \delta_i$ where (for fixed σ, j, and k) we define $\delta_i \equiv O^*_{i,j} - O_{i,j} = O_{i,k} - O^*_{i,k}$ for each $i \in \{1,2,\ldots,n\}$. We now have

$$\Delta = \frac{2p_\sigma}{n}\sum_{i=1}^{n}\left(e^{-\alpha(O_{i,j}+\delta_i)} + e^{-\alpha(O_{i,k}-\delta_i)} - e^{-\alpha\cdot O_{i,j}} - e^{-\alpha\cdot O_{i,k}}\right)$$

Second, we must have

$$\begin{aligned}\delta_i &= O^*_{i,j} - O_{i,j}\\ &= \big(\min(d_{\sigma_{j,k}(i)}, d_{\sigma_{j,k}(j)}) - \max(a_i, a_j)\big) - \big(\min(d_{\sigma(i)}, d_{\sigma(j)}) - \max(a_i, a_j)\big)\\ &= \min(d_{\sigma_{j,k}(i)}, d_{\sigma_{j,k}(j)}) - \min(d_{\sigma(i)}, d_{\sigma(j)})\\ &= \min(d_{\sigma(i)}, d_{\sigma(k)}) - \min(d_{\sigma(i)}, d_{\sigma(j)}) \geq 0,\end{aligned}$$

because $d_{\sigma(j)} < d_{\sigma(k)}$. Third, for all $i \in \{1, 2, \dots, n\}$, we must have

$$\begin{aligned}\delta_i + O_{i,j} - O_{i,k} &= O^*_{i,j} - O_{i,k}\\ &= \big(\min(d_{\sigma_{j,k}(i)}, d_{\sigma_{j,k}(j)}) - \max(a_i, a_j)\big) - \big(\min(d_{\sigma(i)}, d_{\sigma(k)}) - \max(a_i, a_k)\big)\\ &= \big(\min(d_{\sigma(i)}, d_{\sigma(k)}) - \max(a_i, a_j)\big) - \big(\min(d_{\sigma(i)}, d_{\sigma(k)}) - \max(a_i, a_k)\big)\\ &= \max(a_i, a_k) - \max(a_i, a_j) \geq 0,\end{aligned}$$

as $a_j < a_k$, from which it follows that $\delta_i \geq O_{i,k} - O_{i,j}$.

Having established the above observations, we now express the i-th term of Δ (for any $i \in \{1, 2, \dots, n\}$) as

$$\Delta_i(x) \equiv \frac{2p_\sigma}{n}\left(e^{-\alpha(O_{i,j}+x)} + e^{-\alpha(O_{i,k}-x)} - e^{-\alpha\cdot O_{i,j}} - e^{-\alpha\cdot O_{i,k}}\right)$$

evaluated at $x = \delta_i$. Now observe that for each $i \in \{1, 2, \dots, n\}$, $\Delta_i(x)$ is a convex function in x, as

$$\frac{\partial^2 \Delta_i(x)}{\partial x^2} = \frac{2\alpha^2 p_\sigma}{n}\left(e^{-\alpha(O_{i,j}+x)} + e^{-\alpha(O_{i,j}-x)}\right) > 0,$$

so $\Delta_i(x)$ has at most two roots; by inspection, those roots are at $x = 0$ and $x = O_{i,k} - O_{i,j}$ (and it is easily seen that $x = 0$ is th only root when $O_{i,j} = O_{i,k}$); moreover, $\lim_{x\to\infty} \Delta_i(x) = \lim_{x\to-\infty} \Delta_i(x) = +\infty$, hence, $\Delta_i(x) \geq 0$ at all values of x that do not lie in between the aforementioned roots; as $\delta_i > \max(0, O_{i,k} - O_{i,j})$, $\Delta_i(\delta_i) \geq 0$, as desired, and hence, $\Delta \geq 0$.

Therefore it follows that any policy except PLCFS can be (weakly) improved by using an "operation" where some σ that is used some nonzero probability of the time is improved to some $\sigma_{j,k}$. Hence, PLCFS minimizes the expected number of transmissions. Using "inverse operations" of this kind on any non-FCFS policy where we replace some σ with some $\sigma_{j,k}$ where j and k satisfy $j < k$ and $a_j < a_k < d_{\sigma(k)} < d_{\sigma(j)}$, we can a policy yielding a higher mean transmission rate. Hence, any policy except FCFS can be (weakly) worsened by such (inverse) operations, so FCFS maximizes the expected number of transmissions.

EC.3 Hyperexponential Transmission Thresholds

When transmission thresholds are hyperexponentially distributed, we have the following decomposition result, in terms of systems with exponentially distributed transmission thresholds.

Proposition EC.3.1 *If transmission thresholds are hyperexponentially distributed so that there exists some set of infectious-susceptible customer pair types $\mathcal{J}$ such that $\theta \sim \text{Exp}(\alpha_j)$ with probability q_j and $\sum_{j\in\mathcal{J}} q_j = 1$, then*

$$R_0^{\mathsf{sys}} = \sum_{j\in\mathcal{J}} q_j R_0^{\mathsf{sys}}[j],$$

where $R_0^{\mathsf{sys}}[j]$ is the R_0^{sys} when $\theta \sim \text{Exp}(\alpha_j)$ for all infectious-susceptible customer pairs.

Proof First observe that in this setting we have

$$\mathbb{P}\left(W_i^{(s)} \ge \theta\right) = \sum_{j\in\mathcal{J}} q_j \mathbb{P}\left(W_i^{(s)} \ge \theta \middle| \theta \sim \text{Exp}(\alpha_j)\right) = \sum_{j\in\mathcal{J}} q_j \left(1 - \widetilde{W}_i^{(s)}(\alpha_j)\right),$$

and so, following Eqs. (1) and (2), the claim follows:

$$R_0^{\mathsf{sys}} = 2\sum_{s\in\mathcal{S}} \pi(s) \sum_{i=1}^{n(s)} \left(\sum_{j\in\mathcal{J}} q_j \left(1 - \widetilde{W}_i^{(s)}(\alpha_j)\right)\right) = \sum_{j\in\mathcal{J}} q_j \left(2\sum_{s\in\mathcal{S}} \pi(s) \sum_{i=1}^{n(s)} \left(1 - \widetilde{W}_i^{(s)}(\alpha_j)\right)\right) = \sum_{j\in\mathcal{J}} q_j R_0^{\mathsf{sys}}[j].$$

□

EC.4 A Supplemental Discussion on the Impact of Spatial Positioning on Transmission

Consider the setting where transmission thresholds are position-dependent as discussed in Section 6.2. In this setting, the following proposition gives R_0^{sys} for an M/M/1/FCFS system:

Proposition EC.4.1 *Consider an M/M/1/FCFS system where transmission rates $\alpha_{m,j}$ depend on positions as described in Section 6.2. Letting $\eta_{m,j} \equiv \alpha_{m,j}/\mu$, we have*

$$R_0^{\mathsf{sys}} = \left(\frac{2\rho}{1-\rho} - (1-\rho)\sum_{s=0}^{\infty} \rho^s \sum_{i=1}^{s} \left(2 - \prod_{j=0}^{i-1}\left\{\frac{1}{(\eta_{s+1-j,i-j})+1}\right\} - \prod_{j=0}^{i-1}\left\{\frac{1}{(\eta_{i-j,s+1-j})+1}\right\}\right)\right).$$

Proof of Proposition EC.4.1. This proof follows a similar argument to that presented in Proposition 2. The first crucial difference is that the probability that the IC infects the SC initially at position i is not $1-(\eta+1)^{-i}$ in this setting, but rather, it is and is given by $1 - \prod_{j=0}^{i-1}\left\{\frac{1}{(\eta_{s+1-j,i-j})+1}\right\}$. This is because initially the IC is in position $s+1$ while the SC is in position i, then the IC is in position s while the SC is position $i-1$, and so on, until the SC is in position $s+1-i$, while the IC is in position 1. The IC and SC (that was initially at position i) spend a duration of time that is distributed $\text{Exp}(\mu)$ in each of these i positional configuration. Hence, during the IC's sojourn in position $s+1-j$, given that the SC did not previously become infected, the IC *fails* to infect the SC (who is concurrently in position $i-j$), with probability $\mu/((\alpha_{s+1-j,i-j})+\mu) = 1/((\eta_{s+1-j,i-j})+1)$, from which the claimed infection probability follows. Following the proof of Proposition 2, the expected number of customers that the IC infects among those who arrived *before* the IC is given by

$$\frac{\rho}{1-\rho} - (1-\rho)\sum_{s=0}^{\infty}\rho^s \sum_{i=1}^{s}\left(1 - \prod_{j=0}^{i-1}\left\{\frac{1}{(\eta_{s+1-j,i-j})+1}\right\}\right). \tag{EC.8}$$

Given that the M/M/1 system is time-reversible (See chapters 9 and 13 of Harchol-Balter 2013, for details), together with the assumption that the IC is functionally indistinguishable from SCs, and the fact that we are considering a first-come-first-serve system, the expected number of customers that the IC infects among those who arrived *after* the IC is given by a modified version of the same formula given in Display (EC.8): the only modification is that $\eta_{s+1-j,i-j}$ is replaced by $\eta_{i-j,s+1-j}$ (i.e., we have reversed indices). This modification is due to the fact that the symmetry introduced by time-reversibility does not necessarily apply to the infection rates between pairs of positions (i.e., $\alpha_{i,j}$ need not be equal to $\alpha_{j,i}$, and hence, $\eta_{i,j}$ need not be equal to $\eta_{j,i}$). The claimed result then follows by summing these two expectations. □

We proceed to discuss a special case of position-dependent transmission rates where rates depend on distance. Assuming a queue proceeding in a straight line where distances between successive customers are the same, we consider a transmission model where $\alpha_{i,j} = \alpha I\{|i-j| \le d\}$, where $I\{\cdot\}$ denotes the indicator function. That is, an IC can only infect those customers who are waiting up to d positions in front of or behind them in the queue. This model would be reasonable, if, e.g., successive customers in the queue are spaced exactly 6 feet apart and we believe that there is a non-negligible transmission risk (occurring with rate α) when customers are spaced 6–18 feet, but the risk is assumed to be negligible when customers are spaced 24 or more feet apart; in this example, $d=3$. In the case of an M/M/1/FCFS queue, we can compute the R_0^{sys} value for this distance-based transmission model in closed form:

Proposition EC.4.2 *Consider the same M/M/1/FCFS system as in Proposition 2 where a susceptible customer can only be infected by an infected customer within d positions in the queue from themselves. Then we have*

$$R_0^{\mathsf{sys}} = \frac{2\rho\left(((1+\eta)\rho)^d(2\rho-1)+\eta^2(1+\eta)^d(\rho^2-1)+\rho^d\left((1-\rho)^2-(1+\eta)^d\left(\left(\frac{1}{1+\eta}\right)^d(1-\rho)^2+2\rho-1\right)\right)\right)}{\eta(1+\eta-\rho)(\rho-1)(1+\eta)^d}$$

Proof. Noting that a tagged IC overlaps with, at most, d customers at the back of the queue upon their arrival, we can again follow the same symmetry argument as in Proposition 2 and condition on whether or not the length of the queue on arrival exceeds the threshold amount d to find

$$\begin{aligned}
R_0^{\mathsf{sys}} &= 2\sum_{s=0}^{\infty}\pi(s)\sum_{k=1}^{\min\{s,d\}}\left(1-\widetilde{W}_{s-k+1}^{(s)}(\alpha)\right)\\
&= 2\left(\sum_{s=0}^{d}\pi(s)\sum_{k=1}^{s}\left(1-\widetilde{W}_{s-k+1}^{(s)}(\alpha)\right)+\sum_{s=d+1}^{\infty}\pi(s)\sum_{k=1}^{d}\left(1-\widetilde{W}_{s-k+1}^{(s)}(\alpha)\right)\right)\\
&= 2\left(\sum_{s=0}^{d}(1-\rho)\rho^s\sum_{k=1}^{s}\left(1-\left(\frac{1}{1+\eta}\right)^{s-k+1}\right)+\sum_{s=d+1}^{\infty}(1-\rho)\rho^s\sum_{k=1}^{d}\left(1-\left(\frac{1}{1+\eta}\right)^{s-k+1}\right)\right)\\
&= \frac{2\rho\left(((1+\eta)\rho)^d(2\rho-1)+\eta^2(1+\eta)^d(\rho^2-1)+\rho^d\left((1-\rho)^2-(1+\eta)^d\left(\left(\frac{1}{1+\eta}\right)^d(1-\rho)^2+2\rho-1\right)\right)\right)}{\eta(1+\eta-\rho)(\rho-1)(1+\eta)^d}.
\end{aligned}$$

EC.5 Analysis of Preemptive Priority Service Policies

For the system described in Section 5.2.1 except having preemptive priority service policy instead, we have the following proposition:

Proposition EC.5.1 *In the M/M/1 system with preemptive priorities, we have*

$$R_0^{\mathsf{sys}} = 2\left(q_{\mathsf{H}}\left(R_0^{\mathsf{H}\xrightarrow{\mathsf{B}}\mathsf{H}}+R_0^{\mathsf{H}\xrightarrow{\mathsf{B}}\mathsf{L}}\right)+q_{\mathsf{L}}\left(R_0^{\mathsf{L}\xrightarrow{\mathsf{B}}\mathsf{H}}+R_0^{\mathsf{L}\xrightarrow{\mathsf{B}}\mathsf{L}}\right)\right) \tag{EC.9}$$

$$R_0^{H} = 2q_{\mathsf{H}}\left(\frac{\rho_H}{1-\rho_H}\right)\left(\frac{\eta}{\eta+1-\rho_H}\right)+q_{\mathsf{L}}\left(R_0^{\mathsf{L}\xrightarrow{\mathsf{B}}\mathsf{H}}+R_0^{\mathsf{L}\xrightarrow{\mathsf{A}}\mathsf{H}}\right) \tag{EC.10}$$

$$R_0^{L} = R_0^{\mathsf{sys}} - R_0^{H}, \tag{EC.11}$$

where expressions for $R_0^{\mathsf{H}\xrightarrow{\mathsf{B}}\mathsf{H}}$, $R_0^{\mathsf{H}\xrightarrow{\mathsf{B}}\mathsf{L}}$, $R_0^{\mathsf{L}\xrightarrow{\mathsf{B}}\mathsf{H}}$, $R_0^{\mathsf{L}\xrightarrow{\mathsf{B}}\mathsf{L}}$, *and* $R_0^{\mathsf{L}\xrightarrow{\mathsf{A}}\mathsf{H}}$ *together with their derivations are given (in terms of the limiting probability distribution of the M/M/1 system with two priority classes) in Appendix EC.5.1.*

Proof. The first equation follows from the symmetry argument that for the whole system, the expected number of susceptible customers (SCs) that the IC infects among those who arrive before and after the IC are the same. Conditioning on the type of IC, either high-risk or low-risk SCs who were present in the system when the IC arrives will possibly be infected which yields Eq. (EC.9). Next, we obtain the second equation by conditioning on the type of the IC. If the IC is high-risk, then applying Proposition 2 with the load ρ_{H} gives the first half of the equation. Otherwise, the IC will be low-risk and infect $R_0^{\mathsf{L}\xrightarrow{\mathsf{B}}\mathsf{H}}+R_0^{\mathsf{L}\xrightarrow{\mathsf{A}}\mathsf{H}}$ high-risk SCs on average. Finally, the last equation follows from the fact that $R_0^{\mathsf{sys}} = R_0^H + R_0^L$.

EC.5.1 Expressions for the values appearing in Proposition EC.5.1.

In the setting considered in Proposition EC.5.1, let $\pi(h,\ell)$ be the limiting probability distribution of the number of high- and low-risk customers in the system under steady state (see Marks (1973) for the exact solutions), the expressions for the five terms are given in the following Proposition:

Lemma 2 *In the M/M/1 system with preemptive priorities described above, we have*

$$R_0^{\mathsf{H}\overset{\mathbf{B}}{\to}\mathsf{H}} = \left(\frac{\rho_{\mathsf{H}}}{1-\rho_{\mathsf{H}}}\right)\left(\frac{\eta}{\eta+1-\rho_{\mathsf{H}}}\right) \tag{EC.12}$$

$$R_0^{\mathsf{H}\overset{\mathbf{B}}{\to}\mathsf{L}} = \sum_{h=0}^{\infty}\sum_{\ell=1}^{\infty}\pi(h,\ell)\sum_{i=1}^{\ell}\left(1-\left(\frac{1}{1+\eta}\right)^{h+1}\right) \tag{EC.13}$$

$$R_0^{\mathsf{L}\overset{\mathbf{B}}{\to}\mathsf{H}} = \left(\frac{\rho_{\mathsf{H}}}{1-\rho_{\mathsf{H}}}\right)\left(\frac{\eta}{\eta+1-\rho_{\mathsf{H}}}\right) \tag{EC.14}$$

$$R_0^{\mathsf{L}\overset{\mathbf{B}}{\to}\mathsf{L}} = \sum_{h=0}^{\infty}\sum_{\ell=1}^{\infty}\pi(h,\ell)\sum_{i=1}^{\ell}\mathbb{P}(W_{\mathsf{L}\to\mathsf{L}(i)}^{(h,\ell)} \geq \theta) \tag{EC.15}$$

$$R_0^{\mathsf{L}\overset{\mathbf{A}}{\to}\mathsf{H}} = \sum_{(h,\ell)\in\mathcal{S}}\pi(h,\ell)A \tag{EC.16}$$

where

$$A = \frac{1-\rho_{\mathsf{H}}-(1+\eta)^h\left(1-\rho_{\mathsf{H}}-\eta\left(h+\eta+h\eta+\ell\eta-h\rho_{\mathsf{H}}\right)\right)}{\eta(1-\rho)(1+\eta-\rho)(1+\eta)^h} - (1+\ell)\left(1-\frac{1}{1+\eta}\right)$$

and $W_{\mathsf{L}\to\mathsf{L}(i)}^{(h,\ell)}$ *denotes the length of a busy period started by* $(h+i)/\mu$ *amount of work in an M/M/1 system with arrival rate* λ_{H} *and service rate* μ.

Proof. Eq. (EC.12) and Eq. (EC.14) follow from Proposition 2 and the observation that in both cases, we only consider the SCs who are all high-risk customers, so we can treat it as an M/M/1/FCFS system which only has high-risk customers (load becomes ρ_{H}). We can get Eq. (EC.13) and Eq. (EC.15) by directly applying Proposition 1, the sojourn time overlap distribution in the case of Eq. (EC.13) follows Erlang$(h+1,\mu)$ while the sojourn time overlap $W_{\mathsf{L}\to\mathsf{L}(i)}^{(h,\ell)}$ in the case of Eq. (EC.15) not only depends on the present number of customers (both high- and low-risk customers) but also will be affected by the future arrival of high-risk customers (we defer the derivation of $\mathbb{P}\left(W_{\mathsf{L}\to\mathsf{L}(i)}^{(h,\ell)} \geq \theta\right)$ right after this proof). We finish this proof by finding the expression of $R_0^{\mathsf{L}\overset{\mathbf{A}}{\to}\mathsf{L}}$, when the low-risk IC arrives. Assuming the system state is (h,ℓ), there will be ℓ low-risk SCs and h high-risk SCs in the system, all of whom will leave the system before the IC leaves. Note that more high-risk customers may arrive before the IC leaves, and they will be served before the IC. Therefore, we define the first busy period (with h high-risk SCs) as the time until the first low-risk customer leaves the system. Each remaining busy period will end when the next low-risk customer leaves the system. Since only future high-risk SCs will affect the process, we define $V(x,y)$ as the expected number of arrivals to an M/M/1 system with $\rho=\rho_H$ in state y during current busy period before the next service completion of a low-risk customer given the initial state x. $V(x,y)$ can be solved by the following system:

$$\begin{cases} V(x,y) = \dfrac{\rho_H}{1+\rho_H}V(x+1,y) + \dfrac{1}{1+\rho_H}V(x-1,y), \quad \forall\ 1\leq x\leq y \\ V(1,y) = \dfrac{\rho_H}{1+\rho_H}V(2,y), \\ V(y,y) = \dfrac{\rho_H}{1+\rho_H}(1+V(y,y)) + \dfrac{1}{1+\rho_H}V(y-1,y) \end{cases}$$

which leads to

$$V(x,y) = \sum_{j=(y+1-x)^+}^{y}(\rho_H)^j. \tag{EC.17}$$

Hence, the first busy period will contribute $\sum_{n=1}^{\infty} V(h+1,n)\,(1-(1/1+\eta)^{n+1})$ to our risk metric while the other ℓ busy periods will contribute $\ell\sum_{n=1}^{\infty} V(1,n)\,(1-(1/1+\eta)^{n+1})$, together with Proposition 1 we get

$$R_0^{\mathsf{L}\overset{\mathsf{A}}{\to}\mathsf{H}} = \sum_{(h,\ell)\in\mathcal{S}} \pi(h,\ell)\left(\sum_{n=1}^{\infty}\left[V(h+1,n)+\ell V(1,n)\right]\left(1-\left(\frac{1}{1+\eta}\right)^{n+1}\right)\right). \tag{EC.18}$$

Substituting Eq. (EC.17) into Eq. (EC.18) and simplifying the formulas yield the claimed result in Eq. (EC.16). □

Next we proceed to derive the expression of $\mathbb{P}\left(W_{\mathsf{L}\to\mathsf{L}(i)}^{(h,\ell)} \geq \theta\right)$. According to the definition of $W_{\mathsf{L}\to\mathsf{L}(i)}^{(h,\ell)}$, we have (See Chapter 27 of Harchol-Balter 2013, for details on the Laplace transform of the busy period):

$$\mathbb{P}\left(W_{\mathsf{L}\to\mathsf{L}(i)}^{(h,\ell)} \geq \theta\right) = 1 - \widetilde{W}_{\mathsf{L}\to\mathsf{L}(i)}^{(h,\ell)}\left(\alpha + \lambda_{\mathsf{H}} - \lambda_{\mathsf{H}}\widetilde{B}(\alpha)\right)$$

where

$$\widetilde{W}_{\mathsf{L}\to\mathsf{L}(i)}^{(h,\ell)}(\mathbf{s}) = \left(\frac{\mu}{\mu+\mathbf{s}}\right)^{h+i},$$

and

$$\widetilde{B}(\alpha) = \frac{1}{2\lambda_{\mathsf{H}}}\left(\lambda_{\mathsf{H}} + \mu + \alpha - \sqrt{\left(\lambda_{\mathsf{H}} + \mu + \alpha\right)^2 - 4\lambda_{\mathsf{H}}\mu}\right).$$

References

Harchol-Balter M (2013) *Performance modeling and design of computer systems: Queueing theory in action* (Cambridge University Press).

Marks BI (1973) State probabilities of M/M/1 priority queues. *Operations Research* 21(4):974–987.

Perlman Y, Yechiali U (2020) Reducing risk of infection – The COVID-19 queueing game. *Safety Science* 104987.

Shortle JF, Thompson JM, Gross D, Harris CM (2018) *Fundamentals of Queueing Theory*, volume 399 (John Wiley & Sons).