Improved Performance and Stability in Overflow Loss Systems via Exchange of Congestion Information

Eric W. M. Wong
Department of Electronic Engineering
City University of Hong Kong, Kowloon Tong
Hong Kong
Email: eeewong@cityu.edu.hk

Yin-Chi Chan
Institute for Manufacturing
University of Cambridge, Cambridge CB3 0FS
United Kingdom
Email: ycc39@cam.ac.uk

***Abstract*—In systems and networks where overflow is penalized, careful planning is required to prevent a small increase in the offered load from suddenly triggering a system collapse. In this work, we consider an overflow loss system where the service time distribution of a request is dependent on the number of overflows. We propose a novel admission control policy for improving the stabilization and performance of such a system based on a new information exchange mechanism (IEM) for congestion information, in which only a small amount of congestion information is carried by each request and is propagated via the overflow of requests. This avoids the need for an external mechanism for propagating congestion information. Furthermore, IEM is fully compatible with trunk reservation, a simple, classical, yet effective method of admission control in overflow loss systems and networks. Despite the very limited amount of communication overhead required by IEM (only a few bits per request), numerical results demonstrate that the combination of IEM with trunk reservation provides greater performance and stability than trunk reservation alone. We also present a computationally efficient analytical performance evaluation method based on IEM and demonstrate numerically that it is asymptotically exact and generally quite accurate. Therefore, the analytical method can be used for system design and planning.**

I. Introduction

In overflow loss systems (OLSs), requests which cannot be served by a primary server may overflow to other servers until an available server is found, or until all available servers have been exhausted, upon which the request is blocked and cleared. In certain cases, overflow is associated with additional resource use. For example:

- In wireless communication networks, the maximum channel capacity C is decreasing with respect to the distance between the transmitter and the receiver [1]. This increases the time required to complete a transmission, assuming a fixed payload size.
- In the emergency vehicular dispatch model of [2], a vehicle's response times may vary according to whether the vehicle is a first-choice or second-choice of a given geographical "atom". This approximates the effect of distance on vehicular travel times.
- In applications with human agents (e.g., call centers), the required time for a task may depend on whether the agent is a specialist (can only perform one task) or a generalist (may perform several tasks, but with less proficiency). In general, preferring specialists to generalists maximizes system efficiency [3], [4].
- As an alternative to consuming resources for an additional amount of time, overflow requests may instead require additional resource units, e.g., in wired communication networks with alternate routing where overflow paths require additional "hops" compared to the direct (first-choice) path [5]–[7]. This is known to cause instability in such networks unless admission control is applied [8], [9].

Due to the cascading effect of overflow calls requiring more resources, in turn further reducing the network's carrying capacity, a small increase in the offered load of a network may suddenly trigger system collapse, with high levels of system congestion and a large proportion of blocked requests. Therefore, careful planning is required in systems and networks where overflow is penalized, whether in the form of additional resource usage or increased service times.

In this paper, we consider an OLS in which each request requires the service of a single server, with a service time distribution dependent on the number of attempted server groups before an available server is found. We show that without admission control, our overflow system exhibits similar instability behavior to that observed for overflow networks in [5]–[7].

A. Admission control for OLSs

The primary purpose of admission control in OLSs is to avoid a vicious cycle of overflow requests using more resources than fresh requests, reducing the carrying capacity of the system and in turn yielding more overflow. Admission control thus involves answering the following question each time an incoming request overflows from a server group: should the request keep trying to obtain service from the system, or should it give up and make room for future fresh (i.e. less resource-intensive) requests? In principle, a request should keep trying if system congestion is mild, but give up immediately when system congestion is high, in order to avoid the vicious cycle.

Ideally, one would have complete system state information in order to make optimal admission control decisions, but this leads to a large overhead and is infeasible for large systems (e.g. telecommunications systems) with high throughput. In contrast, the classical method for admission control in overflow loss systems and networks, trunk reservation [7]–[9], uses only local congestion information without any knowledge of the global congestion information of the system. In this paper, we propose a mechanism that uses an *approximate* view of global congestion, but with low information overhead.

B. Contributions of this work

In this paper, we propose a new admission control mechanism that uses approximate global congestion information while limiting the amount of information required to be stored, processed, and propagated in the system. This is achieved by having each request carry a *congestion estimate* of the number of busy (i.e., refusing new requests) server groups in the system. As each request overflows, it increments its congestion estimate, reflecting increased knowledge of the global system state. In addition, overflowing requests may *exchange* their congestion estimates with requests already in service, further enhancing the propagation of global state information. Since information exchange only occurs during overflow, only relevant information is shared: the propagation of congestion information is directly linked to the propagation of congestion itself. This allows the admission control to respond quickly to system congestion. On the other hand, as an extreme example, if a system is separable into two independent parts, then congestion information will never be exchanged between the two parts.

Note that the information exchange mechanism (IEM) proposed in this paper is fully compatible with trunk reservation (TR) and can be used concurrently, allowing the system to make use of both local congestion information (e.g., whether the number of free servers at the current server group is smaller than the reservation threshold) and global congestion information (e.g., the estimated total number of busy server groups in the system). In addition, as congestion information in the proposed IEM is carried by the requests themselves, no external congestion information propagation mechanism is required, and the overhead of the network is only a few additional bits of information per request for carrying the request's congestion estimate.

We consider a model of an OLS with Poisson input traffic and global random routing. We demonstrate numerically that the combination of TR and our new IEM can yield *better performance and stability* in OLSs than TR alone, due to the way in which IEM favors fresh over overflow requests. Additionally, we consider the accuracy of decomposition-based approximations for estimating the request blocking probability of the system. The results demonstrate that decomposition of the system without admission control leads to a fixed-point problem with multiple solutions, reflecting the instability of the system itself, while decomposition of the system with IEM admission control *eliminates multiple solutions* and is generally quite accurate compared to simulation results.

II. Model

We consider an OLS with G groups of N identical servers each. Requests arrive to the system according to a Poisson process with rate $G\lambda$ and may each attempt up to k server groups, chosen at random from the G groups comprising the system. The service times of each request is exponentially distributed with mean μ_i^{-1}, where i is the number of overflows experienced by the request before obtaining service and $\mu_i \geq \mu_j$ for all $0 \leq i < j < k$. Overflows are assumed to be instantaneous, and any request that cannot obtain an available server after k overflows is blocked and cleared from the system. Under TR, an overflow request is admitted to a server group if and only if the number of available servers at that group is greater than r.

A. Proposed information exchange and admission control mechanism

In our proposed IEM, requests carry two attributes: Δ, the set of previously visited server groups, and Ω, an estimate of the level of congestion in the system. We shall use the term (Δ, Ω)-request to refer to a request with a specific Δ and Ω; fresh requests always start as $(\emptyset, 0)$-requests.

Consider a (Δ_1, Ω_1)-request arriving at an arbitrary server group g. Under TR, the request is admitted into group g if the number of free servers in group g is greater than x, where $x = 0$ if $\Delta = \emptyset$ and $x = r$ otherwise. Otherwise, the incoming request is compared to the most "senior" (highest Ω) request in service at group g, which we denote as an (Δ_2, Ω_2)-request:

- If $\Omega_1 \geq \Omega_2$, then the incoming request overflows as a $(\Delta_1 \cup \{g\}, \Omega_1 + 1)$-request.
- If $\Omega_1 < \Omega_2$, then the incoming request *exchanges* Ω values with the senior request in service before overflowing as an $(\Delta_1 \cup \{g\}, \Omega_2 + 1)$-request, whereas the request in service becomes a (Δ_2, Ω_1)-request.
- Finally, all overflowing requests immediately *abandon* the system if Ω reaches a predefined limit, denoted $\Omega^\star$, without attempting the remaining $k - |\Delta|$ server groups. For consistency with our definition of k, the maximum number of allowed server group attempts, we require that $\Omega^\star \geq k$.

Note that the proposed IEM has previously been introduced in [10], [11], where it was used to construct a *fictitious* surrogate model to facilitate the performance evaluation of a system *without* IEM. However, in this paper, IEM is applied as an admission control method to the *real* system under study.

By exchanging congestion information only upon the overflow of a request, IEM links the propagation of congestion information with the propagation of congestion itself. In addition, IEM requires very little overhead as congestion information is carried by the requests themselves, eliminating the need for an external mechanism for monitoring system state. Since each request needs to store Δ regardless of whether IEM is employed, to facilitate routing, only the Ω attribute is unique to IEM, meaning that the communication overhead required by IEM is only $\lceil \log_2 \Omega^\star \rceil$ bits per request, irrespective of the value of G. This makes IEM scalable to systems with a large number of server groups.

The admission control element of the above mechanism, in which all overflowing requests immediately *abandon* the system if Ω reaches $\Omega^\star$, can be explained intuitively as follows: we consider it safe for an overflowing request to attempt additional server groups if its estimate of system congestion is low, but force the request to give up immediately when congestion is estimated to be high. This admission control element helps avoid the vicious cycle mentioned in Section I-A which can cause system instability.

III. Analytical Approximation

Whereby the Erlang Fixed-Point approximation [6] is a decomposition-based method for approximating the request blocking probability in our OLS model without IEM, in this section we propose a similar decomposition-based method for our OLS model with IEM, based on the Information Exchange Surrogate Approximation (IESA) framework [10]–[14]. For consistency with [10], we will consider a *surrogate* model in which, in addition to requests immediately abandoning the system if Ω reaches $\Omega^\star$, such requests will also abandon with probability $P_{n,\Omega}$ for all $\Omega < \Omega^\star$, where $n = |\Delta|$ and

$$P_{n,\Omega} = \begin{cases} \frac{\binom{\Omega-n}{k-n}}{\binom{\Omega^\star-n}{k-n}}, & \Omega \geq k \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

This can be interpreted as the probability that when $k-n$ server groups are selected from a pool of $\Omega^\star - n$ server groups, all of them are full, given that $\Omega - n$ of them are full in the entire pool and assuming independence among the groups.

One key feature of IESA is the existence of a closed-form expression for request blocking probability, unlike EFPA, due to the hierarchical traffic structure created by IEM in which the overflow behavior of requests are unaffected by the presence of requests with a higher Ω value. In addition, this traffic hierarchy is able to capture state dependencies between server groups in a manner that preserves these dependencies when decomposition is applied. Therefore, the approximation error caused by decomposing an OLS with IEM is much less than that caused for the corresponding system without IEM.

A. Derivation for current model

In this paper, we extend previous work on IESA [10], [11] to handle multiple service time distributions, as required for the current system model. Consider an arbitrary server group and let:

- $\lambda_{j,n}$ denote the arrival rate of all (Δ, j)-requests to the group where $|\Delta| = n$.
- $\Lambda_{j,n}$ denote the arrival rate of all (Δ, Ω)-requests to the group where $|\Delta| = n$ and $\Omega \leq j$.
- $\vec{\Lambda}_j = (\Lambda_{j,0}, \ldots, \Lambda_{j,k-1})$.
- $b_{j,n}$ denote the blocking probability of requests with $|\Delta| = n$ in the jth level of the IEM hierarchy, containing all requests where $\Omega \leq j$.
- $\vec{b}_j = (b_{j,0}, \ldots, b_{j,k-1})$.
- $\vec{\mu} = (\mu_0, \ldots, \mu_{k-1})$.

To simplify our derivations, we assume that the arrival processes of (Δ, Ω)-requests to a server group, $|\Delta| = n$, is Poisson for all $n = 0, \ldots, k-1$. This allows us to compute $\vec{b}_j$ from $\vec{\Lambda}_j$ and $\vec{\mu}$ by solving a system of steady-state equations. We shall write $\vec{b}_j = B\left(\vec{\Lambda}_j, \vec{\mu}, r\right)$. The algorithm for approximating the blocking probability of the system is given in Fig. 1.

IV. Numerical Results

A. Results for a small OLS

We consider an OLS with $G = 8$ groups of $N = 20$ servers each, $k = 4$, $r = 0$, and

$$\mu_i^{-1} = \begin{cases} 1, & i = 0 \\ 1 + x, & i = 1 \\ 1 + 2x & 2 \leq i < k. \end{cases} \quad (2)$$

In other words, x represents the degree of the overflow penalty; a higher value of x means longer service times for overflow requests.

Fig. 2 shows the blocking probability of the system with respect to λ for various values of x and $\Omega^\star$, as evaluated using simulation and IESA. For comparison, we also include the EFPA estimate of blocking probability for the corresponding system without IEM admission control. It is demonstrated that a smaller value of $\Omega^\star$ leads to higher blocking when λ is low, but lower blocking when λ is high. In other words, $\Omega^\star$provides a trade-off between system performance and robustness (with respect to increases in the offered load). The results also demonstrate that increasing x, the overflow penalty, decreases the robustness of the system to changes in the offered load.

With regards to the approximation methods, IESA is demonstrated to be quite accurate except when x and $\Omega^\star$ are both large. On the other hand, EFPA is quite inaccurate and even produces multiple solutions in many cases, with the range of λ for which this occurs becoming wider as the overflow penalty x increases. This reflects the instability issues known to exist in such systems as previously described in [7]–[9]. Finally, since EFPA does not model the IEM-based admission control system used in the actual OLS, it cannot reflect changes in system performance and robustness due to changes in $\Omega^\star$.

We also consider the case of $x = 0.5$ and $r = 2$ for various values of $\Omega^\star$, with the results shown in Fig. 3. IESA is demonstrated to be quite accurate in all cases considered. Similar results were obtained for $r = 1$ and $r = 3$, with increasing trunk reservation demonstrated to slightly decrease the blocking probability when λ is high, at the cost of increased blocking when λ is low.

function IEA($\lambda, k, \vec{\mu}, N, r, \Omega^\star$)
    $\omega \leftarrow 0$ ▷ Total blocked/abandoned traffic
    $\lambda_{0,0} \leftarrow \lambda$; $\Lambda_{0,0} \leftarrow \lambda$; $\vec{b}_0 \leftarrow B\left(\vec{\Lambda}_0, \vec{\mu}, r\right)$
    for $j \leftarrow 1$ **to** $\Omega^\star$ **do**
        for $n \leftarrow 1$ **to** $\min(k, j)$ **do**
            $x \leftarrow \lambda_{j-1,n-1} b_{j-1,n-1} + \Lambda_{j-2,n-1}\left(b_{j-1,n-1} - b_{j-2,n-1}\right)$ ▷ Overflow (n, j)-request traffic
            if $n < k$ **and** $j < \Omega^\star$ **then** ▷ Possible retry
                $\lambda_{j,n} \leftarrow x\left(1 - P_{n,j}\right)$; $\Lambda_{j,n} \leftarrow \Lambda_{j,n} + \lambda_{j,n}$ ▷ See (1)
            end if
            $\omega \leftarrow \omega + x P_{n,j}$
        end for
        if $j < \Omega^\star$ **then**
            $\vec{b}_j = B\left(\vec{\Lambda}_j, \vec{\mu}, r\right)$
        end if
    end for
    return ω/λ
end function

Fig. 1. Algorithm for approximating the blocking probability of the overflow loss model with IEM. All variables are assumed to be zero until assigned a value, and all variables with negative indices are always zero.

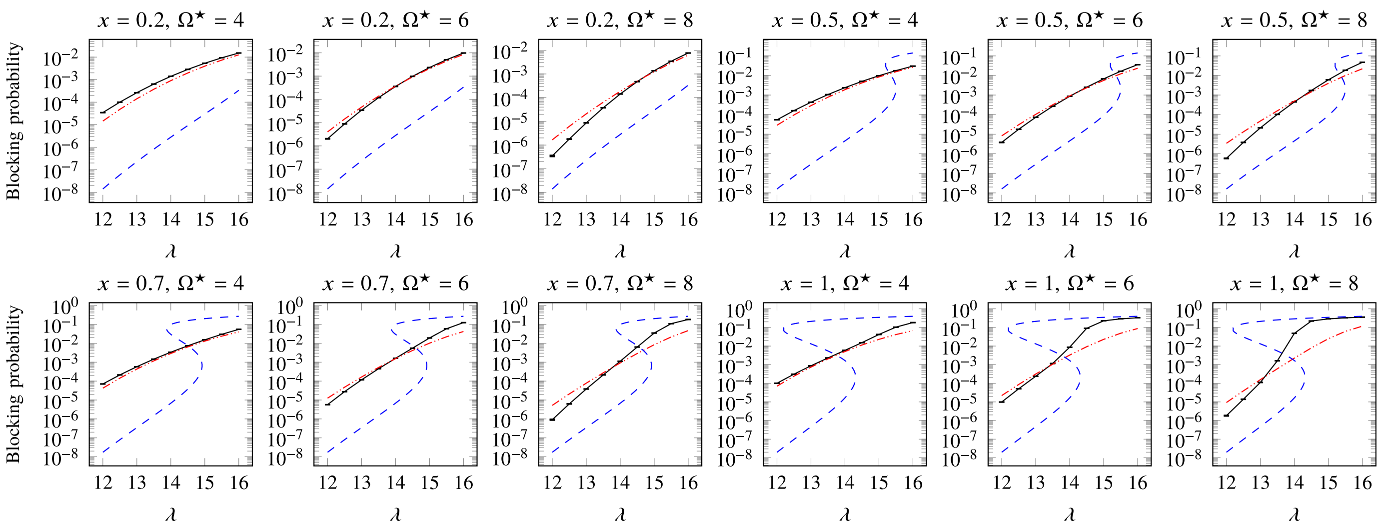

Fig. 2. Blocking probability for an OLS with $G = 8$, $N = 20$, $k = 4$, and $r = 0$ (solid = simulation, dash-dot-dotted = IESA, dashed = EFPA).

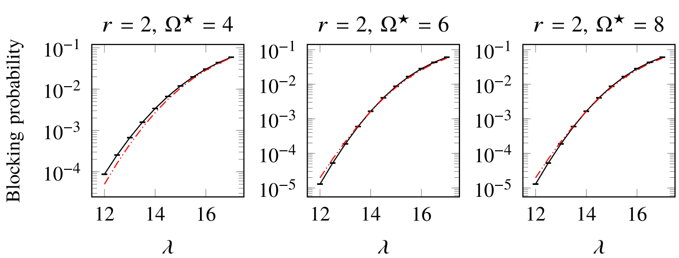

Fig. 3. Blocking probability for an OLS with $G = 8$, $N = 20$, $k = 4$, and $r = 2$ (solid = simulation, dash-dot-dotted = IESA, dashed = EFPA).

B. Behavior as $\Omega^\star = G \to \infty$

We consider the case where $\Omega^\star = G$, i.e., requests only block when all the server groups in the OLS are predicted to be busy, with $N = 20$, $k = 4$,$r = 0$ (IEM only with no trunk reservation), and μ_i as in (2). Fig. 4 shows the blocking probability of the system with respect to λ for various values of x and $\Omega^\star$, as evaluated using simulation and IESA.

The results demonstrate that as $\Omega^\star = G \to \infty$ (i.e. a large system with effectively no admission control), IESA becomes exact. This is consistent with previous work [11] which proved for the case of $x = 0$ that IESA is asymptotically exact as $\Omega^\star = G \to \infty$. In addition, as G increases, the blocking probability decreases for low values of λ, but the sensitivity of the system to the offered load increases, with a vertical jump in the blocking probability in the limit as $G \to \infty$. Thus, for high loads, the blocking probability actually increases with respect to G. In other words, the value of G provides a trade-off between system efficiency and robustness.

C. Optimization of a large OLS

In this subsection, we try to optimize system planning for large OLSs using IESA. We consider an OLS with $G = 500$ groups of $N = 20$ servers each, $\lambda = 14.7$, $k = 3$, and $x = 1$ for various values of r and $\Omega^\star$, with results shown in Fig. 5. It is demonstrated that a lower blocking probabiliity can be obtained using a combination of trunk reservation and IEM–based admission control than by using trunk reservation alone, as confirmed using simulation.

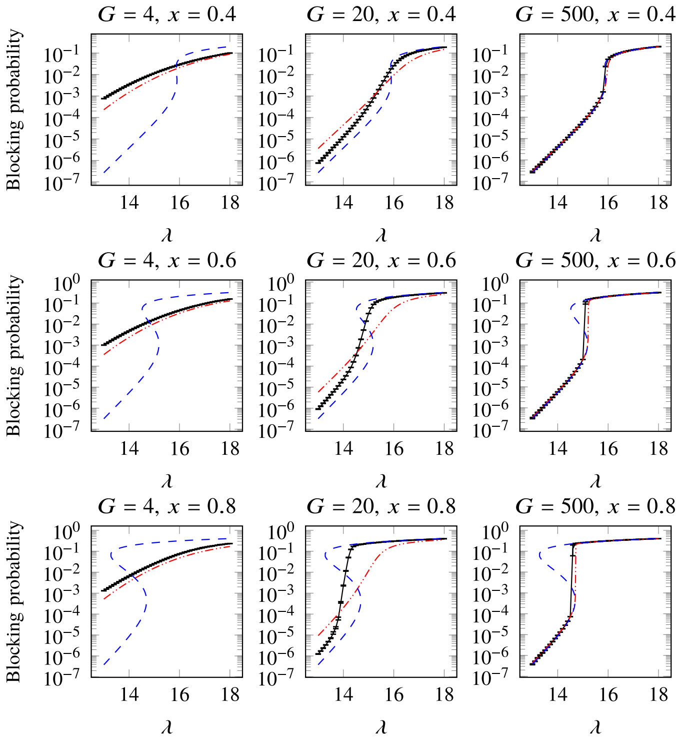

Fig. 4. Blocking probability for the OLS described in Section IV-B (solid = simulation, dash-dot-dotted = IESA, dashed = EFPA).

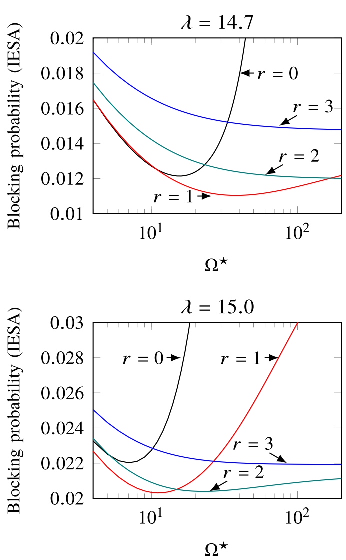

Fig. 5. Results for the OLS described in Section IV-C.

Furthermore, the IESA-optimal configuration of $r = 1$ and $\Omega^\star = 37$ yields a blocking probability 32% higher than that of the simulation-optimal configuration of $r = 1$ and $\Omega^\star = 6$. This suggests that IESA can form a quick heuristic method for finding near-optimal admission control configurations, due to IESA being generally one to two orders of magnitude faster than simulation. This may be particularly useful in heterogeneous systems where one might want to assign different admission control settings to each server group.

The above scenario is repeated for $\lambda = 15.0$, again with results shown in Fig. 5. Again, it is demonstrated that a combination of trunk reservation and IEM-based admission control outperforms trunk reservation alone, as confirmed by simulation. In this case, IESA is less accurate, with the IESA-optimal configuration of $r = 1$ and $\Omega^\star = 11$ is yielding a blocking probability 85% more than the simulation-optimal configuration of $r = 2$ and $\Omega^\star = 5$; however, this is still significantly better than not using IEM at all, demonstrating the usefulness of IESA as a quick heuristic. In particular, the IESA-optimal solution still provides a tenfold reduction in blocking probability compared to no admission control at all.

V. Concluding Remarks

In this paper, we propose IEM as a method for admission control in OLSs where overflow requests are penalized in the form of increased service time. The key feature of IEM is a simple exchange mechanism for the propagation of global system congestion information. This contrasts with the classical method of admission control in such systems, trunk reservation, which only uses local system congestion information. On the other hand, IEM is fully compatible with trunk reservation and can be used concurrently, leading to better performance than either of the two used alone.

IEM operates by attaching a value Ω to each request representing the estimated level of global system congestion, specifically the estimated number of busy server groups in the system, and allowing requests to exchange their Ω parameters according to a fixed set of rules. This means that IEM operates with a very low communication overhead (only a few bits per request) and does not require an external mechanism for propagating congestion information. In addition, the blocking probability of OLSs using IEM can be quickly estimated using an analytical performance evaluation method (i.e. IESA). Unlike the classical EFPA approximation for systems without IEM, IESA always results in a single solution with a closed-form expression.

Note that although IESA is used in this paper to optimize OLSs that implement our proposed IEM congestion control method, it was originally designed for OLSs without any congestion control. More importantly, IESA can be used for detecting whether system congestion collapse or instability occur for such systems, and under what conditions, so that preventive measures such as IEM can be introduced beforehand. Numerical results show that IEM-based admission control is effective at improving stability of OLSs by reducing the sensitivity of an OLS to increases in the offered load, and that IESA is a fairly accurate approximation method for such OLSs, including for cases with trunk reservation and especially when the penalty for overflow requests is not too large.

Additionally, we presented scenarios in this paper where a combination of IEM-based admission control and trunk reservation provides improved performance over either method alone, in terms of minimizing the blocking probability of the system, and where IESA produces configurations that are fairly close to optimal in terms of minimizing blocking probability, suggesting the usefulness of IESA as a quick heuristic method for admission control policy design. We also show that in the limiting case of $\Omega^\star = G \to \infty$ and $r = 0$, IESA becomes asymptotically exact, the first demonstration of asymptotic exactness for IESA for a system where overflows are penalized in the form of increased resource usage duration.

Possibilities for future work include more extensive study on the effect of system blocking by various system parameters (such as k, G, and N), the comparison of different methods for information propagation and admission control (for example, the information *exchange* mechanism in this paper might be replaced by a *copy* mechanism), a wider variety of arrival processes (e.g. time-varying Poisson processes), and adaptive admission control methods.

REFERENCES

[1] J. Wu, M. Wang, Y.-C. Chan, E. W. M. Wong, and T. Kim, "Performance evaluation of 5G mmWave networks with physical-layer and capacity-limited blocking," in *Proc. Int. Conference on High Performance Switching and Routing (HPSR)*, IEEE, May 2020.

[2] J. B. Atkinson, I. N. Kovalenko, N. Kuznetsov, and K. V. Mykhalevych, "A hypercube queueing loss model with customer-dependent service rates," *European Journal of Operational Research*, vol. 191, pp. 223–239, Nov. 2008.

[3] P. Chevalier and N. Tabordon, "Overflow analysis and cross-trained servers," *International Journal of Production Economics*, vol. 85, pp. 47–60, July 2003.

[4] G. Koole and A. Pot, "Approximate dynamic programming in multi-skill call centers," in *Proc. 2005 Winter Simulation Conference*, pp. 576–583, Dec. 2005.

[5] Y. Nakagome and H. Mori, "Flexible routing in the global communication network," in *Proc. ITC 7*, pp. 426/1–8, 1973.

[6] F. P. Kelly, "Blocking probabilities in large circuit-switched networks," *Advances in Applied Probability*, vol. 18, pp. 473–505, June 1986.

[7] R. J. Gibbens, P. J. Hunt, and F. P. Kelly, "Bistability in communication networks," in *Disorder in Physical Systems* (G. Grimmett and D. Welsh, eds.), pp. 113–128, Oxford University Press, 1990.

[8] R. S. Krupp, "Stabilization of alternate routing networks," in *Proc. IEEE ICC*, 1982.

[9] L. G. Mason, "On the stability of circuit-switched networks with non-hierarchical routing," in *Proc. 25th IEEE CDC*, pp. 1345–1347, Dec. 1986.

[10] E. W. M. Wong, J. Guo, B. Moran, and M. Zukerman, "Information exchange surrogates for approximation of blocking probabilities in overflow loss systems," in *Proc. ITC 25*, Sept. 2013.

[11] Y.-C. Chan and E. W. M. Wong, "Blocking probability evaluation for non-hierarchical overflow loss systems," vol. 66, pp. 2022–2036, May 2018.

[12] J. Wu, E. W. M. Wong, Y.-C. Chan, and M. Zukerman, "Energy efficiency-QoS tradeoff in cellular networks with base-station sleeping," in *Proc. IEEE GLOBECOM 2017*, Dec. 2017.

[13] J. Wu, E. W. M. Wong, Y.-C. Chan, and M. Zukerman, "Power consumption and GoS tradeoff in cellular mobile networks with base station sleeping and related performance studies," *IEEE Transactions on Green Communications and Networking*, vol. 4, pp. 1024–1036, Dec. 2020.

[14] Y.-C. Chan, E. W. M. Wong, and C. S. Leung, "Evaluating non-hierarchical overflow loss systems using teletraffic theory and neural networks," vol. 25, pp. 1486–1490, May 2021.