Optimizing Hauling Vehicle Mix for Debris Removal: A Queueing Theory Approach

James D. Brooks
Industrial and Systems Engineering
Rensselaer Polytechnic Institute
Troy, NY 12180
Email: brookj7@rpi.edu

David Mendonça
Industrial and Systems Engineering
Rensselaer Polytechnic Institute
Troy, NY 12180
Email: mendod@rpi.edu

***Abstract*—This work explores one aspect of debris removal system design, the optimal mix of hauling vehicles, from two perspectives: efficiency and equity. First, the vehicle mix which maximizes the efficiency of the system is determined (i.e., that which minimizes the average wait time). Secondly, the equity of the most efficient mix is considered (i.e., the ratio of wait times between different vehicles). The results suggest that the solution to the optimal mix of hauling vehicles is nominally the same for a wide range of arrival rates for some system configurations. Further, the analysis shows that the optimal solution permanently disadvantages some vehicles for the majority of service rate combinations, however equitable optimal solutions (i.e., those for which all hauling vehicles have the same expected wait times in the system wait time minimizing solution) do exist, though the locus is very small. System configurations which lead to these equitably optimal solutions are highlighted. Policy implications are also presented.**

I. Introduction

Debris removal following natural disasters necessitates rapid design of the operational structure and network of physical assets to support the cleanup efforts. As a result, many aspects of system configuration and design "just happen", creating little feedback between system configuration decisions and assessment of resulting performance effects. A fundamental building block in a debris removal system are the loops of activity between the many locations where debris is picked up (curbside) and the central drop-off sites. Experience has shown that these drop-off sites, called temporary debris storage and reduction (TDSR) sites, are often a bottleneck for system throughput. As a result, effort should be expended to understand how to best design and utilize these facilities. This work provides guidance on system configuration (i.e., TDSR configuration and vehicle mix) to achieve two potentially conflicting objectives, maximizing system efficiency (i.e., reducing average wait time at TDSR sites) while providing performance equity between vehicles (and thus the contractors involved in the work), by exploring optimal solutions generated using a queueing systems model of a typical TDSR site.

These TDSR sites are commonly situated in large, open areas such as a ball field, parking lot, or open field. The purpose of these sites is to serve as centralized staging areas where debris can be collected from the nearby area, and reduced as much as possible (through chipping, burning, crushing, etc.) prior to long haul transportation to the final disposal site. Arriving loads first enter a queue at the inspection tower where the volume of the load is estimated for payment. The load then proceeds to the appropriate area depending on the capability of the hauling vehicle (dump only or grapple loader for self-loading/unloading) and the debris type (e.g., vegetative, construction and demolition, or hazardous material). After each hauling vehicle is finished unloading, it cycles back to the inspection tower to verify it is empty before returning to a pickup site. The reduced debris is often piled in rows and is subsequently loaded into much larger hauling vehicles for final disposal.

Because the TDSR site layout is traditionally static, any inefficiencies will be compounded significantly over time. As a result, performance gains at these bottleneck sites are expected to have dramatic impacts on both system and individual team effectiveness. This analysis will also highlight any structural effects which would permanently disadvantage a particular subset of hauling vehicles.

This system is naturally suited to modeling and analysis using queueing theory [1]: services are performed by discrete entities taking random service times (either a chipper or self-unloading areas) and arriving vehicles which find their server busy wait in line and are processed on a first come, first served basis. The capabilities of the various vehicles used to haul debris from the initial site (curbside) to the TDSR sites are modeled as two *customer classes* in the queueing model. Each class is then served by its own server station: standard dump trucks are served by a chipper and auxiliary loading equipment, while vehicles equipped with their own grapple loaders are served by a separate area which allows some number of these vehicles to unload themselves. The desired proportion of these two basic vehicle capabilities is the focus of this work.

While much work in system-level optimization of multi-class queueing systems has been done, little attention has been given to any resulting performance differences between customer classes (i.e., system inequity). This work presents both the optimal solution over many possible states of nature (i.e., differing service rates, μ, and arrival rates, λ) and an exploration of any resulting differences in expected performance between customer classes (i.e., systemic or structural inequity).

Related work in optimal customer allocation in similar queueing systems is first presented. The problem formulation is then described along with the equity measure used and its relation to prior work. Numerical results are then presented for the optimal proportion of customer classes (i.e., the pro-

978-1-4799-1535-4/13/$31.00 ©2013 IEEE

portion of vehicle capabilities which minimizes the expected average delay) given a common TDSR site layout over a range of arrival and service rates (i.e., system configurations). The equity of these solutions is then considered and system configurations which exhibit equitable, optimal solutions are highlighted. Finally, conclusions and potentially fruitful areas of future work are discussed.

II. Related Work

Prior work in routing of customers optimally between parallel servers is extensive. Only the most relevant known work is discussed here. In a parallel server topology, arriving external customers choose from among a number of parallel servers and then exit the system. The objective is to minimize the expected average delay (or equivalently the expected total number of entities in the system). Both dynamic (control) and static (design) policies have been explored. The join-shortest-queue policy was first shown to be optimal for exponential homogeneous servers [2]. This result was further shown to be asymptotically optimal for the multi-server case as well [3]. If the parallel system has only two heterogeneous servers, it is optimal to send all traffic to the faster server until the number of customers waiting exceeds a threshold, at which point the slower servers should be utilized [4], [5]. Under heavy traffic (i.e., $\lambda \approx \sum_i \mu_i$), these state-dependent threshold policies converge to a static threshold policy [6]. Other policies explored in this system topology include minimum expected delay [7] and round robin [8]. For the problem of interest, the routing is determined entirely by the hauling vehicle capability, thus only static policies are of interest.

For multiple parallel servers, it is generally found that faster servers are disproportionally higher utilized in the optimal routing solution [9], [10]. When multiple servers are available to an incoming customer, the faster-server-first policy is known to be asymptotically optimal while allowing slower servers to idle [11]. This inequity common to optimal routing policies often goes unmentioned. In one exception in the known literature, Armony and Ward consider policies which minimize customer wait times in a call center while adhering to *server* equity constraints [12].

Other work has considered optimal proportion of hauling vehicles and workload assigments, particularly in the mining and waste management industries. Muduli and Yegulalp consider the best allocation of truck capacities given fixed routing probabilities to maximize system throughput using mean value analysis [13]. Queueing delays have also been minimized at central processing sites by small modifications to workload assignment which lead to differences in arrival processes [14].

The contributions of the present work are that the equity of *customers* in the optimal efficiency solution is considered. Further, the approach considers customer equity in a post-hoc way, rather than as a constraint, in order to identify system configurations which are *simultaneously* efficient and equitable. The robustness of these results over a wide range of system configurations is also discussed along with related policy considerations.

III. Problem Formulation

The two classes of customers — self-unloading trucks and dump trucks — are routed deterministically to two separate service areas as discussed previously. Dump trucks are routed to a chipper while self-unloading trucks are routed to a separate area where multiple trucks can unload themselves. A diagram of a typical layout is shown in Figure 1.

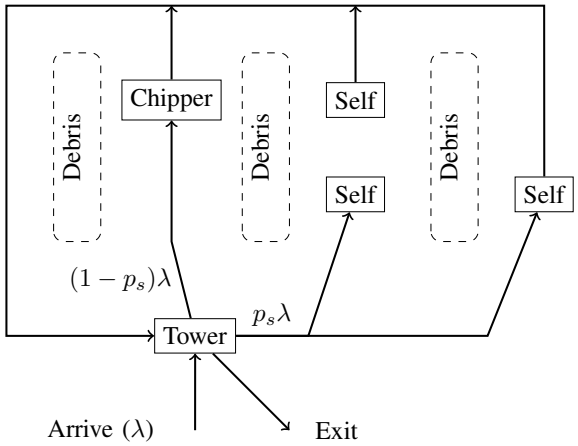

Fig. 1. Typical temporary debris storage and reduction (TDSR) site layout.

The chipper is assumed to have a constant average service rate, μ_c, while the service times are exponentially distributed (i.e., it is a simple M/M/1 queue). The self-unloading areas are assumed to allow three vehicles to simultaneously unload with exponentially distributed service times with a constant mean, μ_s, (i.e., this area is an M/M/3 queue). Notice that this implicitly assumes that all vehicles of the same type have the same amount of debris to unload and that all operators have equivalent skill levels so that no systematic difference in service rates exists between vehicles.

The total arrival process is assumed to be Poisson with a constant mean rate of λ vehicles per hour. The mean rate for the two customers classes is further assumed to be proportional to the population proportion (i.e., there is no systematic difference between the operating characteristics of these two classes). Suppose the proportion of the hauling vehicle population which has self-loading capability is p_s. The arrival rate of vehicles destined for the chipper servers is then $\lambda_c = (1 - p_s)\lambda$ while the rate of self-unloading vehicles is $\lambda_s = p_s\lambda$.

Finally, the system interaction between TDSR wait times and the arrival rate as well as travel times within the TDSR site itself are neglected. A wide range of arrival rates is considered to account for this system interaction. Due to the relatively small size of TDSR sites, it is reasonable to assume travel times will be small relative to wait and service times.

Queueing theory provides expressions for several mean performance measures [1]. For the chipper model, an M/M/1 queue with arrival rate of $(1-p_s)\lambda$ and service rate of μ_c, the expected waiting time is given by

$$\mathbb{E}\{W_c\} = \frac{1}{\mu_c - (1-p_s)\lambda}.$$

The expression for a queue with $c > 1$ servers (i.e., the self-unloading area model) is less straightforward in that the probability of all servers being free must first be calculated as

$$p_0 = \frac{1}{\sum_{k=1}^{c-1} \frac{(c\rho)^k}{k!} + \frac{\rho^c}{c!(1-\rho/c)}}$$

where $\rho = p_s\lambda/\mu_s$ and $c = 3$. The expected wait time for the self-unloading vehicles is then

$$\mathbb{E}\{W_s\} = \left(\frac{1}{c\mu_s - p_s\lambda}\right)\left(\frac{\rho^c}{c!(1-\rho/c)}\right)p_0 + \frac{1}{\mu_s}.$$

The optimization problem is then to minimize the average wait time (or equivalently, the number of vehicles waiting and being serviced in the TDSR) which is given by

$$\begin{aligned} \text{minimize} \quad & p_s\mathbb{E}\{W_s\} + (1-p_s)\mathbb{E}\{W_c\} \\ \text{subject to} \quad & 0 \le p_s \le 1. \end{aligned}$$

A. *Equity*

In addition to efficiency concerns, because the system is one of public service, it is important that the configuration of the system not favor a particular vehicle capability over another — that is, the system should be *equitable*. One influential review of equity in facility location problems describes twenty-four functional forms for equity (or more commonly, inequity) measurement [15]. More recently, Wierman outlines equity measures for evaluating sequencing policies for queueing systems [16]. There is little consensus about the best measure and in fact some are provably conflicting [16].

Some desired properties of equity measures have, however, been identified. These include analytic tractability, appropriateness, impartiality, principle of transfers, scale invariance, Pareto optimality, and normalization [15]. This work follows other authors and justifies the measure used with respect to these desired properties. As only two classes of vehicles are considered in the present study, the ratio of expected mean waiting times of the two customer classes is used as an equity measure. This is given by

$$I = \frac{\mathbb{E}\{W_c\}}{\mathbb{E}\{W_s\}}.$$

In a perfectly equitable system, $I = 1$. A system configuration which favors dump trucks has $I < 1$, while $I > 1$ for a configuration which favors self-unloading trucks.

This measure is tractable, appropriate for the domain (i.e., wait times are clearly understood), and impartial (i.e., simply based on measured or predicted wait time). The measure also satisfies principle of transfers as the equity measure is directly tied to the difference (here the ratio) of performance measures. It also is scale invariant as a uniform scaling of wait times does not effect the equity measure. Pareto optimality is met because as the equity measure approaches one, the efficiency effects on both groups are the same. Finally, the measure does not strictly satisfy the normalization property in that it is not restricted to a unit interval and it differentiates between inequity of the two groups as described previously.

In the next section, the optimal efficiency solutions for a range of likely system configurations is presented. Additionally, the equity of these solutions is shown along with an approximation of the configuration (i.e., chipper and self-unloading rates) for systems whose optimal efficiency solution is equitable for a range of arrival rates.

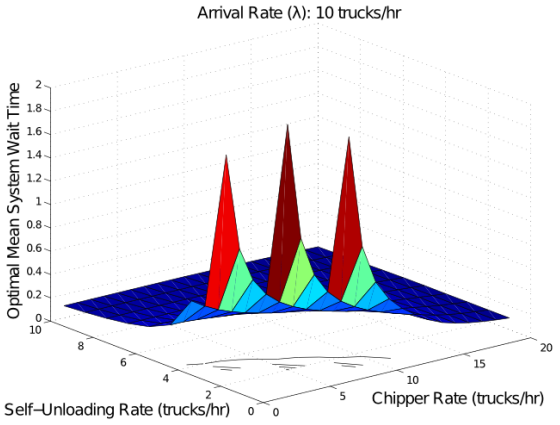

Fig. 2. Optimal mean system wait time for $\lambda = 10$ for varying service rates.

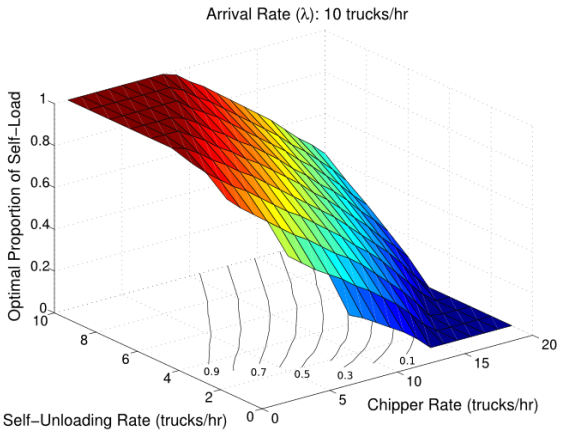

Fig. 3. Optimal proportion of self-loading trucks for $\lambda = 10$ for varying service rates. The contours for the various proportions are also shown.

IV. Numerical Results

Because the solution space is not complex, the optimal point is found by numerically evaluating the system expected wait time over uniformly spaced increments of p_s for a wide variety of possible service rates. These service rates are likely to vary based on equipment and other site- and event-specific considerations. As a result, the results consider a broad range of service rates. Further, the arrival rate will vary throughout the mission making the robustness of the prescribed proportion important (i.e., the amount of variation in the optimal solution with changes in other system parameters). Thus, results are presented for a wide range of system parameters to aid in system planning.

In one recent debris removal mission following tornadoes in the state of Alabama in 2011, TDSR mean arrival rates were found to be 3.25 – 13.25 for vegetative sites and 1.75 – 6.08 trucks per hour for construction and demolition sites while approximate service rates were 5.80 – 15.30 for vegetative sites and 3.26 – 8.32 trucks per hour for construction and demolition

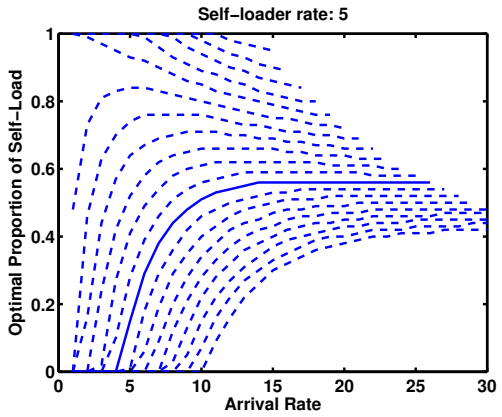

Fig. 4. Optimal proportion of self-loading trucks for $\mu_s = 5$ and $c = 3$ for varying arrival rates. Each curve represents a different $\mu_c = \{1, \ldots, 20\}$ with $\mu_c = 12$ highlighted.

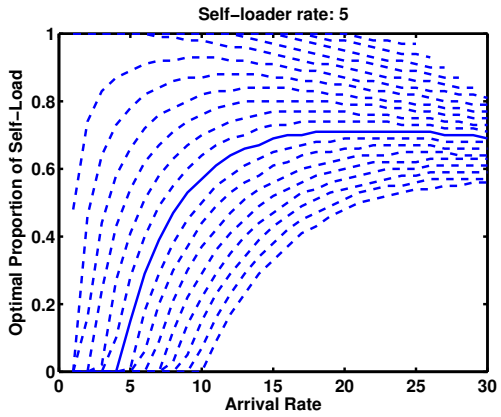

Fig. 5. Optimal proportion of self-loading trucks for $\mu_s = 5$ and $c = 5$ for varying arrival rates. Each curve represents a different $\mu_c = \{1, \ldots, 20\}$ with $\mu_c = 12$ highlighted.

TDSR sites [17]. The ranges of possible service rates in this work are chosen to be $1 \leq \mu_c \leq 20$ and $1 \leq \mu_s \leq 10$. Most of the results presented here focus on a single arrival rate of 10 trucks per hour for simplicity of presentation.

The optimal average wait time for the parameter ranges considered is shown in Figure 2 for a mean arrival rate of $\lambda = 10$ trucks per hour. The optimal system efficiency is nearly uniform for a large portion of the parameter space. However, for slow server speeds ($\mu_s \leq 4$ and $\mu_c \leq 10$), the wait times increase rapidly as expected.

The corresponding surface of the optimal proportion of self-unloading vehicles, p_s, is shown in Figure 3. There are small regions of operations where the optimal proportion is 100% (fast self-unload and slow chipper) or 0% self-loading trucks (slow self-unload and fast chipper). However, the majority of optimal operating points have an optimal proportion in between. The contours shown indicate the system configurations for each optimal proportion value. Each arrival rate produces a different optimal proportion surface, making robustness of the optimal point the next item of inquiry.

In order to look at the changes in optimal proportion with changes in arrival rate to assess robustness, another view of the optimal proportions is provided in Figure 4. This figure shows the optimal proportion of self-loading trucks under one particular self-loading service rate, $\mu_s = 5$, for varying arrival rates (x-axis) and chipper service rates (different curves).

Each curve in Figure 4 represents a fixed TDSR configuration (i.e., equipment selection). The optimal proportion varies the most for each configuration under small arrival rates. This is the same region of operation in which queue lines are the shortest on average. Thus, performance gain (i.e., the difference in performance between optimal proportion and all other proportions) is smaller than in the higher arrival rate regions. For the cases wherein the self-unloading rate is slower than the chipper rate ($\mu_s < \mu_c$), the optimal proportion of self-unloading vehicles falls to zero for small arrival rates as expected. Similarly, for the $\mu_s \geq \mu_c$ cases, the proportion approaches unity. For higher arrival rates, the optimal proportion tends to converge to some value in between.

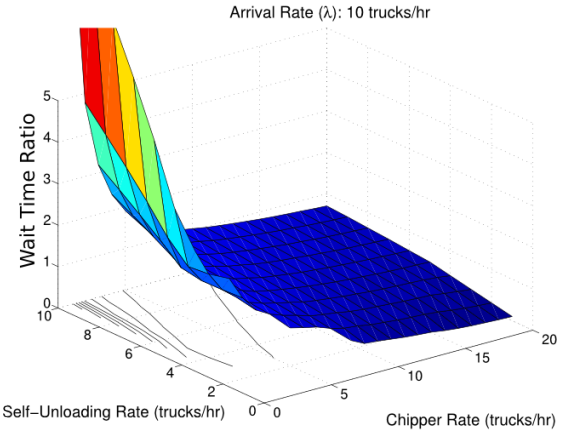

Fig. 6. Equity measure of mean wait times between vehicles for $\lambda = 10$ for varying service rates.

The region in the top right of the figure which contains no curves is where the system is unstable (i.e., no steady-state operating point exists). Thus, it becomes clear that in order to support increased total throughput (i.e., higher arrival rates) it is necessary to concurrently increase chipper speed and reduce the proportion of self-loading trucks for a fixed self-unloading rate. Figure 5 show the effect of adding two additional unloading spots for the self-loading trucks. As expected, the optimal proportion of self-loading trucks increases and the unstable region decreases.

It is desirable from a policy perspective to be able to recommend a single target proportion which is robust (i.e., supports optimal performance over a wide range of arrival rates). Different system parameters support this desire better than others. For example, if $\mu_s = 5, c = 3$, and $\mu_c = 12$, the optimal proportion is 56% for all arrival rates above 12 trucks per hour (shown by solid line in Figure 4). This proportion increases to 71% with five self-unloading servers. For other configurations, however, the optimal proportion can vary considerably.

In many cases, the optimal operating point from a system perspective is highly non-optimal for some vehicles. Thus,

TABLE I. LINEAR FIT OF SYSTEM PARAMETERS ADMITTING EFFICIENT, EQUITABLE SOLUTIONS FOR $c = 3$.

Arrival Rate	*Slope*	*Intercept*	R^2
1	1.006	-0.043	1.000
5	1.093	-0.926	0.998
10	1.274	-3.267	0.994
15	1.678	-8.890	0.902
20	1.99	-15.105	0.950

$$\mu_s = (\text{slope})\mu_c + (\text{intercept})$$

another important consideration is how fair the system optimal points are for individual trucks, that is, how different are the expected wait times for self-loading and dump trucks in the most efficient solution. Indeed if the system optimal point is such that all dump trucks have a wait time which is always twice as much as the expected wait time for self-loading trucks, the dump truck individuals are permanently disadvantaged by the system in a way that they cannot overcome. This can be remedied in two ways: either by considering the equity of an operating point in determining optimality directly, or by compensating the disadvantaged truck type.

To assess the equity of the optimal solution, Figure 6 shows the inequity measure, I, the ratio of dump truck wait time to self-loading truck wait time, for the same arrival rate as before (10 trucks/hour). When the chipper service rate is small and self-unloading rate is large, the dump trucks experience more than five times the wait of self-unloading trucks on average. Conversely, when the chipper rate is large and self-unloading rate is small, the self-unloading trucks experience more than five times the wait of dump trucks. This general result holds over a wide range of arrival rates which are not shown for brevity.

Recall that the desired value for the equity measure, I, is one. Looking at the contours shown in Figure 6, one can see that the unity contour can be well-fit by a linear function. These system parameters whose maximum efficiency solution has an equity measure of one are next characterized by a linear fit of this contour. The system parameters (i.e., service rates) which admit equitable optimal solutions for an arrival rate of $\lambda = 10$ are given approximately by

$$\mu_s = 1.274\mu_c - 3.267.$$

The slope and intercept of this line for a variety of arrival rates is shown in Table I. These system parameters are also shown graphically in Figure 7. For faster self-unloading rates, the range of chipper rates which admit equitable, efficient solutions is reduced. This is then a more robust system configuration as the rates differ less among the range of arrival rates.

V. CONCLUSIONS AND FUTURE WORK

This work has provided an initial exploration of the equity, efficiency tradeoff present in arrival process control for heterogeneous parallel server systems. It has furthermore introduced the concept of *robustness* with respect to these two dimensions of performance. System configurations which admit equitable, optimal solutions were also highlighted.

This paper has shown evidence that queueing theory can be used to gain important insight into the design of a critical component of the debris removal system: the temporary debris storage and reduction sites. The results show that the design of the TDSR sites can dramatically effect both the performance of the site and the equity of the system. Currently no target for proportion self-loading trucks is provided by debris contracts or TDSR site design guides. The intuitive perceptions of informally interviewed subcontractors are that 100% self-loading trucks would be optimal. However, for most system parameters, a proportion of self-loading trucks substantially different from 100% has been shown to be robustly optimal over a wide range of arrival rates.

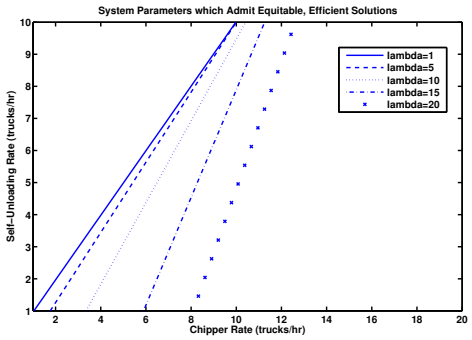

Fig. 7. Linear fit of system parameters which exhibit equitable, optimal solutions.

From an equity perspective, there is a significantly less robust result in that there is a very small region of system parameters for which the optimal proportion results in equal waiting time between the truck types. It appears as though for a given self-loader rate, there is on unique chipper rate for which this is true and vice versa. These system configurations have been shown to be well-approximated by a linear function. These approximations have been presented for a range of arrival rates.

This work can be expanded in several interesting ways. Field work is needed to verify the service time distribution and parameter(s). A sensitivity analysis with respect to modeling assumptions would further provide confidence in the applicability of these results to a real mission. Additional considerations for chipping the self-unloaded debris and associated overhead of long haul operations is also left as future work. Further, the site operations could also be made dynamic (e.g., by varying the number of self-loading spots available in response to arrival rate changes over time). Another possible area of future work is to explore competition between multiple contractors, each having some set of hauling vehicles from a game theoretic perspective.

ACKNOWLEDGMENT

The authors thank two reviewers for their comments. This material is based upon work supported by the National Science Foundation under Grant No. 1313589.

REFERENCES

[1] D. Gross, J. F. Shortle, J. M. Thompson, and C. M. Harris, *Fundamentals of Queueing Theory*. John Wiley & Sons, 2008.

[2] W. Winston, "Optimality of the shortest line discipline," *Journal of Applied Probability*, vol. 14, no. 1, pp. 181–189, 1977.

[3] D. J. Houck, "Comparison of policies for routing customers to parallel queueing systems," *Operations Research*, vol. 35, pp. 306–310, 1987.

[4] W. Lin and P. R. Kumar, "Optimal control of a queueing system with two heterogeneous servers," *IEEE Transactions on Automatic Control*, vol. 29, no. 8, pp. 696–703, 1984.

[5] J. Filipiak, "Dynamic routing in a queueing system with a multiple service facility," *Operations Research*, vol. 32, no. 5, pp. 1163–1180, 1984.

[6] Y.-C. Teh and A. R. Ward, "Critical thresholds for dynamic routing," *Queueing Systems*, vol. 42, pp. 297–316, 2002.

[7] J. Lui, R. Muntz, and D. Towsley, "Bounding the mean response time of the minimum expected delay routing policy: An algorithmic approach," *IEEE Transactions on Computers*, vol. 44, no. 12, pp. 1371–1382, 1995.

[8] Z. Liu and D. Towsley, "Optimality of the round-robin routing policy," *Journal of Applied Probability*, vol. 31, no. 2, pp. 466–475, 1994.

[9] W. F. Piepmeier, "Optimal balancing of I/O requests to disks," *Communications of the ACM*, vol. 18, no. 9, pp. 524–527, 1975.

[10] M. Delasay, B. Kolfal, and A. Ingolfsson, "Maximizing throughput in finite-source parallel queue systems," *European Journal of Operational Research*, vol. 217, no. 3, pp. 554–559, Mar. 2012.

[11] M. Armony, "Dynamic routing in large-scale service systems with heterogeneous servers," *Queueing Systems*, vol. 51, no. 3-4, pp. 287–329, Dec. 2005.

[12] M. Armony and A. R. Ward, "Fair dynamic routing in large-scale heterogeneous-server systems," *Operations Research*, vol. 58, no. 3, pp. 624–637, Feb. 2010.

[13] P. K. Muduli and T. M. Yegulalp, "Modeling truck-shovel systems as closed queueing network with multiple job classes," *International Transactions in Operational Research*, vol. 3, no. I, pp. 89–98, 1996.

[14] B. Wilson, B. Baetz, and F. Hall, "Reduction of queuing delays at waste management facilities," *Civil Engineering and Environmental Systems*, vol. 19, no. 4, pp. 311–331, 2002.

[15] M. Marsh and D. Schilling, "Equity measurement in facility location analysis: a review and framework," *European Journal of Operational Research*, vol. 74, pp. 1–17, 1994.

[16] A. Wierman, "Fairness and scheduling in single server queues," *Surveys in Operations Research and Management Science*, vol. 16, no. 1, pp. 39–48, Jan. 2011.

[17] J. D. Brooks and D. Mendonça, "Simulating market effects on boundedly rational agents in control of the dynamic dispatching of actors in network-based operations," in *Proceedings of the 2013 Winter Simulation Conference*, forthcoming.