Dynamic Allocation of Entities in Closed Queueing Networks: An Application to Debris Removal

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

Koushik Kar
Electrical, Computer,
and Systems Engineering
Rensselaer Polytechnic Institute
Troy, NY 12180
Email: kark@rpi.edu

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

***Abstract*—This work describes a novel method for allocating entities to routes in a closed queueing network to maximize system throughput. Results are presented which compare this method with known prior work and known optimal solutions to provide an empirical optimality gap. Further, because the system of interest, debris removal following natural disasters, is under the control of rational agents, optimality conditions are derived in order to determine to efficacy of a flat market context in inducing optimal behavior.**

I. Introduction

Debris removal following natural disasters involves a set of hauling vehicles which cyclically deliver debris from a potentially large number of pickup sites (typically at curbsides) to a much fewer number of temporary debris storage and reduction (TDSR) sites, forming a network. Improvement in the throughput of this system (i.e., the average number of loads delivered per hour) translates directly to the ability of a community to return to normal operation after the disaster event. This work presents the underlying theoretical models and methodology for a decision support system which dynamically reallocates a fixed number of hauling vehicles to routes in this network (i.e., between pickup-TDSR pairs in this domain) in order to maximize system throughput as the system evolves.

This system is well-suited to modeling using queueing theory [1]. In particular, the system can be modeled as a network of queues in which a fixed number of entities repeatedly move between pickup and drop-off sites, with each site requiring a random service time. Arriving entities which find their server busy wait in line and are processed on a first come, first served basis. The routing of entities in the network to maximize system throughput is the focus of this work.

Because of the geographic extent of the system, a routing (allocation) policy which does not require full knowledge of the location of all entities is desired (i.e. a state-independent policy). While some limited prior work has considered optimal state-independent routing probabilities in closed queueing networks, no work is known to provide a method of partitioning the entities into *classes*, each with its own deterministic route, as is known to be optimal [2]. This work first presents a novel method for generating optimal routing probabilities in closed queueing networks using reasonable modeling approximations and then a method for determining an entity partition, each with its own deterministic-routing, from these probabilities.

This partition strategy is used to determine the optimal allocation of hauling vehicles among the pickup and TDSR sites given the current system state (e.g., differing loading rates and travel times). Simulation results show that the system throughput is significantly higher under the partition strategy than the purely probabilistic one. Optimality conditions for the network flows are presented for the exponential service distribution case which show that when travel times are equal, pickup sites that are able to load vehicles faster should be utilized disproportionally less than slower sites compared to a rational equilibrium (i.e., a *Wardrop equilibrium* in which total travel times of utilized paths are equal [3]). These results suggest that a payment policy which compensates each entity equally (and thus encourages a Wardrop equilibrium) does not support an efficient rational equilibrium in general.

The next section describes related prior work and motivates the approach used. The formulation for the general allocation problem in closed queueing networks is then discussed in detail, using a small example to illustrate the derivation before presenting the general result. The partition algorithm is then described before numerical results are presented. These results compare the new method with both existing methods and provably optimal allocations to obtain an estimate of the empirical optimality gap.

II. Related Work

Much work has been done in the area of optimal design of queueing systems [4]. However, very little work in routing of entities in closed networks has been done. In one notable example, Kobayashi and Gerla present an iterative method for finding optimal routing probabilities based on mean value analysis [5]. The primary drawback of this work is that no closed-form optimality conditions can be derived using this methodology. Further, while it is known that the optimal routing strategy is one in which each entity follows a deterministic routing policy [2], [6], [7], there is no known method for finding these individual routing strategies. Others have looked at the dual problem: that of designing the network capacities to maximize throughput given fixed routing probabilities (e.g., [8], [9]).

To evaluate the effectiveness of a payment policy (which influences the resulting routing behavior of rational agents operating in this network), closed-form optimality conditions are required. It is apparent that these existing iterative methods do

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

not satisfy these requirements and thus a system approximation and alternate method are needed. The work presented here provides a methodology for determining routing probabilities which exhibits optimality conditions and for determining a deterministic routing policy from these probabilities.

Approximations of queueing networks include system approximations (i.e., to obtain a product-form network), decomposition methods (i.e., assume nodes of network are independent), and process approximations (i.e., continuous fluid flows for heavy traffic) [1]. Of particular interest for this work is Whitt's *finite population mean* (FPM) method [10]. In this closed network approximation, the nodes are assumed to be independent queues in an open network in which the sum of the mean number of entities at each node is equal to the number of entities in the closed network being approximated. This FPM approximation along with optimization has been previously used in the context of workload allocation among server pools [11].

III. Problem Formulation and Methodology

The general problem of interest for this work is that of allocating entities circulating in a bipartite closed network to maximize system throughput. The particular network structure of interest is shown in Figure 1. Here entities repeatedly travel from any number of parallel nodes (i.e., pickup sites) to a (fewer) number of central nodes (i.e., TDSR sites), here called a parallel-cycle network. Each node has an associated congestion (latency) function which is increasing with traffic flow and each path has some fixed travel delay.

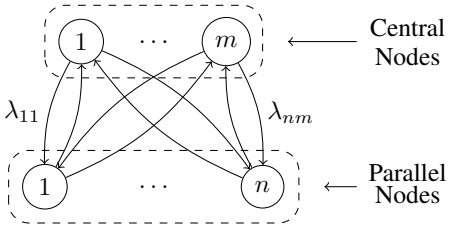

Fig. 1. General parallel-cycle network with m central nodes and n parallel nodes.

Let $\mathcal{C}$ be the set of central nodes ($|\mathcal{C}| = m$) and $\mathcal{P}$ be the set of parallel nodes ($|\mathcal{P}| = n$). Then, $i \in \mathcal{P}$ be a parallel node and $j \in \mathcal{C}$ be a central node. Consider the following parallel-cycle partition (PCP) problem formulation in which the continuous flows, λ_{ij}, are the decision variables, the constraint is the number of entities in the system, and the objective is to maximize the total system throughput. The relevant variables are shown in Table I along with additional notation.

TABLE I. Variables for PCP problem.

Variable	*Description*
λ_{ij}	Flow assigned to cycle $i-j$ (entities/hr)
$\hat{\lambda}_j$	Total flow to central node j $\left(\sum_i \lambda_{ij}\right)$
$\bar{\lambda}_i$	Total flow to parallel node i $\left(\sum_j \lambda_{ij}\right)$
μ_i	Average service rate of node i
N	Total number of entities in the system
ρ_i	Average utilization of node i
d_{ij}	Round trip mean travel delay for cycle $i-j$

Now, let these nodes be single server queues with known service time distributions. Note that the terms node and server will be used interchangeably throughout this paper. Using the FPM approximation, Little's Law (expected number of entities at each server equals average arrival rate times the expected wait time [1]) can be used to form a constraint on the total number of entities. The PCP problem can then be written as the following nonlinear program:

$$\max \quad T = \sum_{i=1}^{n}\sum_{j=1}^{m} \lambda_{ij} \tag{1}$$

$$\text{s.t.} \quad \sum_{j=1}^{m} \hat{\lambda}_j C_j(\hat{\lambda}_j) + \sum_{i=1}^{n} \bar{\lambda}_i D_i(\bar{\lambda}_i) + \sum_{i=1}^{n}\sum_{j=1}^{m} d_{ij}\lambda_{ij} \leq N \tag{2}$$

$$\lambda_{ij} \geq 0. \tag{3}$$

Here $D_j(\lambda), C_j(\lambda)\colon \mathcal{R}^+ \mapsto [0,\infty]$ are general mean congestion (i.e., latency) functions at the parallel and central servers, respectively and d_{ij} is the mean round-trip travel delay between i and j.

Note that we can write the first constraint as an inequality because of the form of the objective (i.e., sum of positive terms strictly increasing in λ_{ij}). Further, this formulation assumes no travel congestion (i.e., d_{ij} is not a function of λ_{ij}). This is reasonable in early disaster recovery situations as the number of vehicles on the road network is small relative to the capacity of the network as normal traffic is minimal.

For open networks (i.e., those with external Poisson arrivals and departures), these congestion functions have a complete closed-form representation since each server is known to be independent [1]. On the other hand, for closed networks a normalization constant must be found. The two most common approaches to approximating this constant is through Buzen's convolution algorithm [12] and mean value analysis [13]. Both of these are non-polynomial algorithms and neither admits a general closed-form congestion function suitable for use in solving the PCP problem in the programming framework.

The FPM approximation assumes that these servers are indeed independent and uses the open network congestion functions. This approximation is asymptotically correct for large number of nodes and entities and the ratio of any two server utilizations in the approximate model is equal to that in the actual closed system [10]. Thus, we expect the ratio of optimal flows to also be equal for large number of nodes and entities (as $\rho = \lambda/\mu$).

A locally optimal solution to the general NLP

$$\begin{aligned} &\text{maximize} && f(x) \\ &\text{subject to} && g(x) \leq 0 \end{aligned}$$

satisfies the Karush-Kuhn-Tucker (KKT) first-order conditions [14] given by

$$\begin{aligned} -\nabla f(x^*) + u\nabla g(x^*) &= 0 \\ g(x^*) &\leq 0 \\ ug(x^*) &= 0 \\ u &\geq 0. \end{aligned}$$

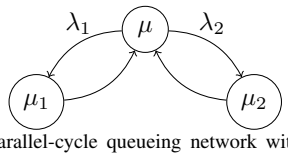

Fig. 2. Example parallel-cycle queueing network with two parallel nodes and a single central node.

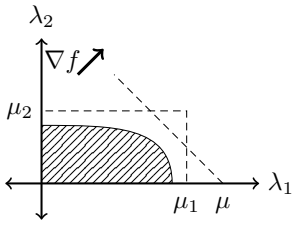

Fig. 3. Feasible region along with stability constraints and objective gradient for two parallel cycles.

This solution is also globally optimal when the feasible region and objective function are convex [14]. The objective is linear (and thus also convex) while the feasible region using M/M/c congestion functions is known to be convex [15], [16]. Therefore, the first-order KKT conditions are both necessary and sufficient to find globally optimal flows under this Markovian assumption. To illustrate the derivation of these conditions, a small example which can be easily visualized is described before providing the general result.

A. Two Parallel Cycles Example

Consider the system composed of a single central node and two parallel nodes as shown in Figure 2. Let these nodes have exponentially distributed service times which are known to produce central and parallel node delay functions given by

$$C(\lambda_1, \lambda_2) = \frac{1}{\mu - \lambda_1 - \lambda_2} \text{ and } D_i(\lambda_i) = \frac{1}{\mu_i - \lambda_i}, i \in \{1, 2\}.$$

The constraints are shown generically in Figure 3 along with the stability boundaries ($\lambda < \mu$) and objective gradient. Taking the derivatives of the objective ($\lambda_1 + \lambda_2$) and the three constraints (Constraint 2 and Constraint 3 for each decision variable) with respect to both decision variables, the gradients are given by

$$\nabla f = \begin{bmatrix} 1 & 1 \end{bmatrix}$$
$$\nabla g = \begin{bmatrix} \frac{\mu}{(\mu-\lambda_1-\lambda_2)^2} + \frac{\mu_1}{(\mu_1-\lambda_1)^2} + d_1 & 1 & 0 \\ \frac{\mu}{(\mu-\lambda_1-\lambda_2)^2} + \frac{\mu_2}{(\mu_2-\lambda_2)^2} + d_2 & 0 & 1 \end{bmatrix}^T.$$

These gradients then give the following two equations from the KKT first-order condition:

$$u_2 \left(\frac{\mu}{(\mu - \lambda_1 - \lambda_2)^2} + \frac{\mu_1}{(\mu_1 - \lambda_1)^2} + d_1 \right) + u_{3,1} = 1$$

and

$$u_2 \left(\frac{\mu}{(\mu - \lambda_1 - \lambda_2)^2} + \frac{\mu_2}{(\mu_2 - \lambda_2)^2} + d_2 \right) + u_{3,2} = 1$$

where u_2 is the Lagrange multiplier for Constraint 2 and $u_{3,1}, u_{3,2}$ are the multipliers for Constraint 3. For both of these equations to hold with non-trivial flows (i.e., $u_{3,1} = u_{3,2} = 0$), it must be true that

$$\frac{\mu_1}{(\mu_1 - \lambda_1)^2} - \frac{\mu_2}{(\mu_2 - \lambda_2)^2} = d_2 - d_1$$

while the first constraint is at equality (i.e., $u_2 > 0$). Further, one can see that this takes the form of a hyperbola in the parallel congestion (i.e., wait) times, given by

$$\frac{D_1^2}{A^2} - \frac{D_2^2}{B^2} = 1$$

where $A = \sqrt{\frac{d_2 - d_1}{\mu_1}}$ and $B = \sqrt{\frac{d_2 - d_1}{\mu_2}}$. As $d_2 - d_1 \to 0$ (i.e., the travel times of each possible route are the same), the first quadrant solution tends to

$$D_2 = \sqrt{\frac{\mu_1}{\mu_2}}\, D_1. \tag{4}$$

When this is compared to the rational equilibrium for this case ($D_1 = D_2$, i.e., the expected wait times at each pickup site are the same), it is clear that for $\mu_1 \neq \mu_2$, the rational equilibrium is *not* optimal.

B. General Optimality Condition

For an arbitrary number of parallel cycles, n, and central servers, m, the general gradients are given by

$$\nabla f = \mathcal{E}_{nm}^T$$
$$\nabla g = \begin{bmatrix} \mathcal{A}_1 & \mathcal{I}_n & 0 & \cdots & 0 \\ \mathcal{A}_2 & 0 & \mathcal{I}_n & & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ \mathcal{A}_m & 0 & \cdots & 0 & \mathcal{I}_n \end{bmatrix}^T.$$

where $\mathcal{I}_n$ is an n-dimensional identity matrix, $\mathcal{E}_{nm}$ is a column vector of ones of length nm (the number of flow decision variables), and each element, i, in the column vector $\mathcal{A}_j$ is

$$\mu_i D_i^2 + \mu_j C_j^2 + d_{ij}, \quad i = 1, \ldots, n$$

The first-order KKT conditions require that all non-zero flows satisfy

$$\mu_i D_i^2 + \mu_j C_j^2 + d_{ij} = \mu_k D_k^2 + \mu_l C_l^2 + d_{kl},$$
$$\forall\ (i,j), (k,l) \in M \times N \text{ where } \lambda_{ij}, \lambda_{kl} > 0$$

and the optimal point intersects with the first constraint at equality. Note that this gives an over constrained system in general (i.e., $n + m$ unknowns, nm equations). As a result, for any system in which $d_{ij} \neq d_{kj} \quad \forall\ i, k \in \mathcal{P}$, the optimal solution must have some zero flows. For the case in which all travel times are equal,

$$\sqrt{\mu_j} C_j = \sqrt{\mu_l} C_l, \quad \forall\ j, l \in \mathcal{C} \text{ and}$$
$$\sqrt{\mu_i} D_i = \sqrt{\mu_k} D_k, \quad \forall\ i, k \in \mathcal{P}.$$

Again, it is clear that the rational equilibrium ($C_j = C_l$ and $D_i = D_k$, i.e., all possible routes have equal delays) is *not* optimal in general. Therefore, any market context which encourages this kind of equilibrium where wait times are equal among all possible options will lead to suboptimal performance for cases in which the servers are heterogeneous.

C. Partition Allocation Method

The solution to the problem as posed gives continuous flows, λ_{ij}, for each cycle. These flows now need to be converted to routing probabilities, to allow for comparison of the FPM method with prior work, and a partition of the entities into deterministic routes, which is known to be optimal as discussed previously. Additionally, note that a partition strategy is also more practical in the case of debris removal as this corresponds to the creation of teams with fixed assignments. A probabilistic routing strategy can be obtained by normalizing the flows:

$$p_{ij} = \frac{\lambda_{ij}}{\sum_j \lambda_{ij}} \quad \text{and} \quad p_{ji} = \frac{\lambda_{ij}}{\sum_i \lambda_{ij}} \quad \forall\ i \in \mathcal{P},\ j \in \mathcal{C}.$$

Meanwhile, a novel rounding strategy is required to obtain the deterministic routing assignments (i.e., a partition) which is described next.

Obtaining an optimal integer solution from a relaxed solution is commonly done by either branch-and-bound or by adding cutting planes when the decision variables themselves are desired to take integer values [17]. However, because the decision variables (mean flows) themselves can remain fractional with an integral entity assignment, these standard methods are not appropriate. The method proposed to obtain a partition routing strategy is shown in Algorithm 1. First, calculate the mean lengths from the optimal flows (lines 2–4), and then calculate the initial integral solution by rounding down (flooring) the number of entities on each path thus calculated (lines 5–8). The number of entities unallocated, k, is then calculated (line 9). Finally, increase the entity count of the k partitions with the largest non-integer parts by one (line 10) to complete the allocation.

Algorithm 1 Partition Assignment Method

1: **procedure** ASSIGN(λ, μ, N)
2: $L_i \leftarrow \frac{\hat{\lambda}_i}{\mu_i - \hat{\lambda}_i}$ ▷ Mean parallel length
3: $L_j \leftarrow \frac{\hat{\lambda}_j}{\mu_j - \hat{\lambda}_j}$ ▷ Mean central length
4: $L_{ij} = d_{ij}\lambda_{ij}$ ▷ Mean travel length
5: **for** $i \leftarrow 1, n; j \leftarrow 1, m$ **do**
6: $\hat{n}_{ij} \leftarrow L_{ij} + \frac{L_j \hat{\lambda}_j}{\lambda_{ij}} + \frac{L_i \hat{\lambda}_i}{\lambda_{ij}}$
7: $n_{ij} \leftarrow \lfloor \hat{n}_{ij} \rfloor$
8: **end for**
9: $k \leftarrow N - \sum_i \sum_j n_{ij}$
10: $n_{ij} = n_{ij} + 1$ for k largest $\hat{n}_{ij} - n_{ij}$
11: **end procedure**

IV. Numerical Results

To assess the performance of the throughput optimization method using the FPM approximations, two hundred random instances with 2 central servers, 6 parallel servers, and zero travel delays were generated using parameters chosen from uniform distributions ($N \sim U(1, 40)$, $\mu_{i\in\mathcal{P}} \sim U(10, 25)$, $\mu_{j\in\mathcal{C}} \sim U(5, 45)$) Travel delays were omitted so that the FPM approximation could be evaluated in isolation. The network configuration and parameters are in line with previous analysis of actual debris removal missions [18].

The problem formulation is implemented in MATLAB and solved with `fmincon`'s interior-point method. This continuous solution was first converted into a probability routing matrix and partition (i.e., fixed routing groups) assignment according to the methods described earlier and then simulated for 5000 time units (with a 1000 unit warm-up period).

First, the mean throughput observed in the simulation is compared with the approximate throughput, both relative to the maximum capability of the system, $\lambda_{\max} = \min\left(\sum_i \mu_i, \sum_j \mu_j\right)$, in the left panel of Figure 4. These results show that the FPM approximation consistently underestimates the simulated throughput (as the approximation predicted result (stars) is below both probabilistic simulation (squares) and partition simulation (triangles)). This result can further be seen in that all points in the right panel are greater than one. As the approximation assumes that an infinite number of entities would result from high utilization, while in reality the number of entities is fixed, this result is to be expected. Further, the partition strategy uniformly outperforms the probabilistic strategy as expected. The ratio of simulated throughput to the FPM predicted throughput for each of the two routing strategies is shown for each case in the right panel of Figure 4. These data also support Whitt's conjecture that the approximation throughput error is bounded above by 1/2 with probabilistic routing [10] as all probabilistic simulation ratios (squares) are less than 1.5. It is also clear from the right panel that the approximation generally improves with increasing number of entities as expected (i.e., the points approach one as N.

1) Comparison with Prior Work: In this portion of the numerical results, the routing probabilities resulting from proposed method are compared with those presented in Kobayashi and Gerla's Example 1 [5]. This example is a central server system similar to that shown in Figure 2 with three parallel cycles where $\mu = 4$, $\mu_1 = 2$, $\mu_2 = 1$, and $\mu_3 = 0.5$. The resulting routing probabilities from both the proposed FPM method and Kobayashi and Gerla's MVA method are shown in Figure 5. All FPM routing probabilities are within 0.133 of those determined by the MVA method. The FPM method tends to under-allocate to the first parallel server while over-allocating to the remaining two. The differences are generally decreasing as the number of entities increases. Finally, as N increases, the proposed method converges to the *balanced* probabilities in which the utilizations are equalized (i.e., the well-known *proportional routing* asymptotic result [19]).

To understand the impact of these small differences, 100 simulation runs were performed for the $N = 5$ case with the two sets of routing probabilities shown in Figure 5. Each configuration was simulated for 5000 time units (with a 1000 unit warm-up period). The histograms of the resulting average system throughputs are shown in Figure 6 along with the results for the partition routing strategy. Little difference between the probabilistic routing cases resulting from the FPM approximation method and the MVA method are observed (means are 2.386 and 2.401, respectively), while the partition routing clearly out-performs with a mean total system throughput of 2.830, an increase of 17.6% over the prior literature for this particular case.

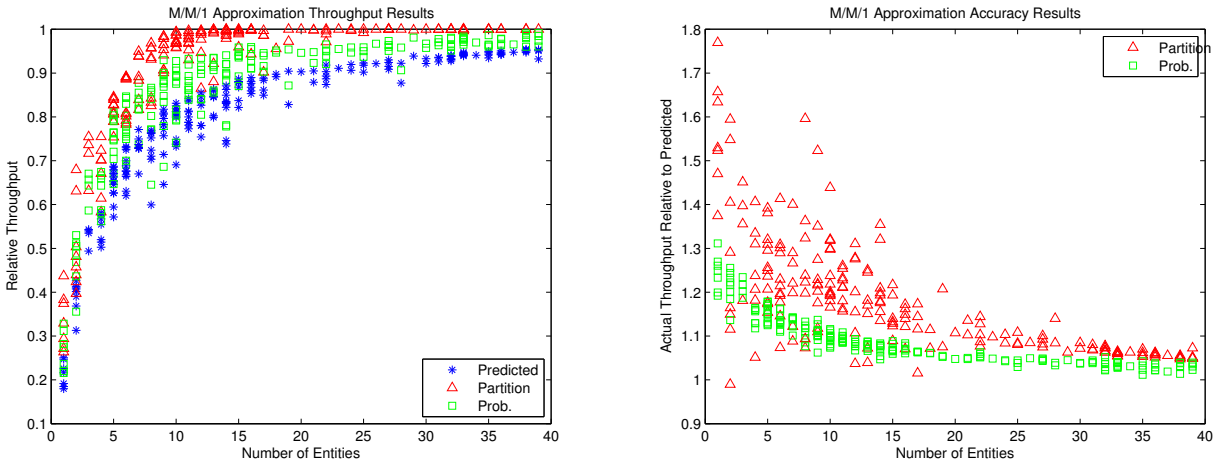

Fig. 4. Throughput of approximation prediction and actual simulation with both routing schemes (probabilistic and partition assignment) relative to maximum capability of the system and ratio of actual to predicted throughput (left and right, respectively) for FPM approximation.

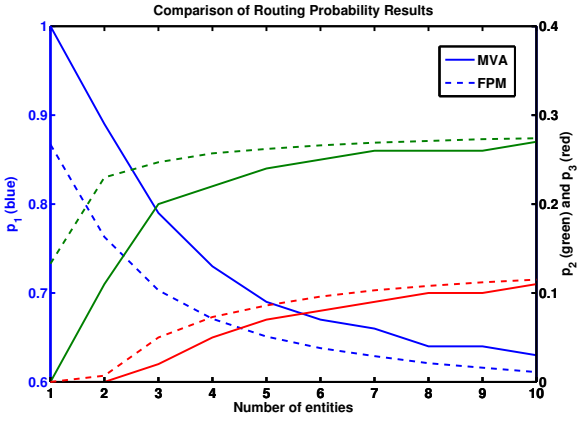

Fig. 5. Comparison of routing probility results. All MVA results are from Table V in [5]. Left axis shows routing probability to the first parallel server while the right axes show probability to the remaining two parallel servers.

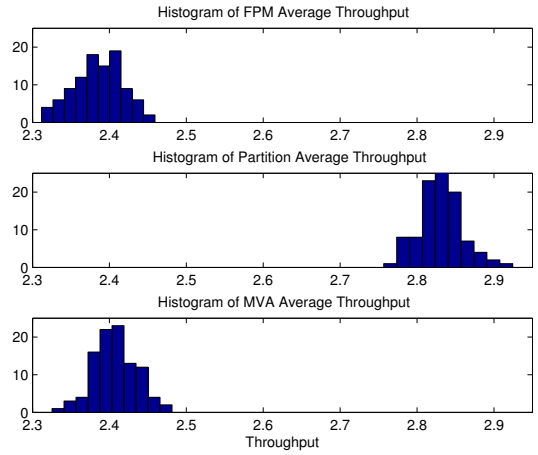

Fig. 6. Histogram of system throughput using FPM approximation and partition routing compared to the MVA method for 100 simulation runs.

2) Empirical Optimality Gap: To evaluate the optimality gap of the proposed partitioning method, a large set of small networks was created and optimal partitions found empirically via complete enumeration of the solution space and determination of the expected system throughput via mean value analysis. For this test, only two parallel cycles were present (as shown in Figure 2) and a fixed number of entities was used. All possible assignments were evaluated using mean value analysis and an optimal assignment was selected. The objective function was *system throughput* — i.e., the throughput of the central server. The number of entities was chosen to be twenty-four and the first parallel node service rate, μ_1, to be ten. The central node rate, μ, was varied from 5 to 40 in increments of 5 while the second parallel node rate, μ_2, was varied from 1 to 25 in increments of 0.2. This parameter space resulted in 968 total test cases. All test results were generated using the `queueing` Octave package [20].

The expected relationship between the two parallel server waiting times that characterize the first-order optimality condition given by Equation 4 can be seen from this empirical data in Figure 7. This figure shows a scatter plot which has axes equal to the scaled waiting times for each parallel node ($\sqrt{\mu_i} D_i$) for the known optimal solutions for these 968 cases. The optimality condition for the continuous flow case (i.e., solution to the PCP problem) is a straight line in this space. The observed dispersion is likely due to the approximation in the optimization problem, partition assignment, and the inclusion of cases which have multiple optima. The general trend, however, is consistent with the optimality conditions from the approximation.

Almost half of the cases produce a calculated partition which is identical to a known optimal partition. Over 90% of the calculated partitions are within 3 of the known optimal partition. Note that in the case of multiple known optimal solutions, the first one found in the search is used. Thus, these percentages should be considered lower bounds.

The empirical optimality gap itself is calculated by considering the total system throughput of the system under the calculated partition compared to the known optimal partition.

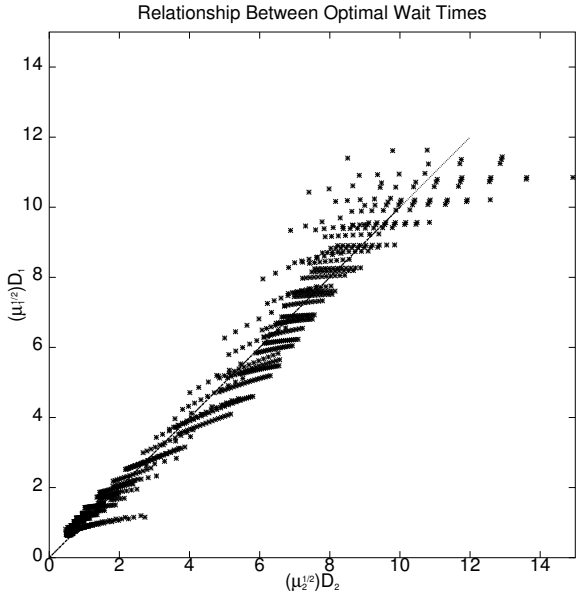

Fig. 7. Relationship of optimal parallel server wait times. The theoretical optimality condition given by Equation 4 is the straight line shown.

The mean optimality gap for the test cases considered is 0.0025%. The maximum optimality gap observed, and thus the lower bound on the worst-case optimality gap, is 0.079%.

V. Conclusions and Future Work

This paper presents a new methodology for allocating entities in a closed queueing network using a finite population mean approximation and math programming framework. Further, a method for determining a partition of the entities into deterministic routes is presented. The results show that the probabilistic routing results of this new method are as effective as the previous work while the new partition method clearly out-performs by producing a mean system throughput 17.6% higher in the instance examined. Further, the empirical optimality gaps with this partition strategy have been considered for a small class of problem instances in which provably optimal solutions could be obtained. The partition routing method proposed was observed to provide allocations whose throughput is within 0.079% of the known optimal throughput for the nearly one thousand instances considered. Future work includes extending this result to larger networks.

Furthermore, the framework presented provide closed-form optimality conditions which allow for the evaluation of expected performance of payment policies under the assumption of rational agents. It was shown that for the case of equal travel times and exponential servers that a uniform policy is not efficient in general. Additional work is needed to design an appropriate market context which would induce good behavior for the general case wherein the system is under the control of boundedly rational agents.

The current method is developed for the case where each node can process any entity in the system. Extensions are required, for example, for the case where multiple vehicle/load types exist and central nodes are restricted in their ability to process certain vehicle/load types. The other closed network approximations proposed by Whitt (e.g., FPM with finite waiting room [10]) could also be explored along with any travel congestion effects and various service distributions.

While the motivating domain was debris removal following natural disasters, the framework is general and can be applied to any system for which a closed queueing network is appropriate. This includes mining, logging, and other material transport systems.

Acknowledgment

The authors thank three reviewers for their helpful comments on this paper. 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] S. K. Tripathi and C. M. Woodside, "A vertex-allocation theorem for resources in queuing networks," *Journal of the ACM*, vol. 35, no. 1, pp. 221–230, Jan. 1988.

[3] J. G. Wardrop, "Some theoretical aspects of road traffic research," *Proceedings of the Institution of Civil Engineers*, vol. 1, pp. 325–378, 1952.

[4] S. Stidham, *Optimal design of queueing systems*. CRC Press, 2009.

[5] H. Kobayashi and M. Gerla, "Optimal routing in closed queuing networks," *ACM Transactions on Computer Systems*, vol. 1, no. 4, pp. 294–310, 1983.

[6] W. C. Cheng and R. R. Muntz, "Optimal routing for closed queuing networks," *Performance Evaluation*, vol. 13, pp. 3–15, 1991.

[7] A. Hordijk and J. A. Loeve, "Optimal static customer routing in a closed queuing network," *Statistica Neerlandica*, vol. 54, no. 2, pp. 148–159, 2000.

[8] P. Whittle, "Optimal routing in Jackson networks," *Asia-Pacific Journal of Operational Research*, vol. 1, pp. 32–37, 1984.

[9] D. L. Bakuli and J. M. Smith, "Theory and Methodology Resource allocation in state-dependent emergency evacuation networks," *European Journal of Operational Research*, vol. 89, pp. 543–555, 1996.

[10] W. Whitt, "Open and closed models for networks of queues," *AT&T Bell Laboratores Technical Journal*, vol. 63, no. 9, pp. 1911–1979, 1984.

[11] J. Calabrese, "Optimal workload allocation in open networks of multiserver queues," *Management Science*, vol. 38, no. 12, pp. 1792–1802, 1992.

[12] J. P. Buzen, "Computational Algorithms for Closed Queueing Networks with Exponential Servers," *Communications of the ACM*, vol. 16, no. 9, pp. 527–531, 1973.

[13] M. Reiser and S. S. Lavenberg, "Mean-Value Analysis of Closed Multichain Queuing Networks," *Journal of the Association for Computing Machinery*, vol. 27, no. 2, pp. 313–322, 1980.

[14] O. L. Mangasarian, *Nonlinear programming*. Society for Industrial and Applied Mathematics, 1987, vol. 10.

[15] W. Grassmann, "The convexity of the mean queue size of the M/M/c queue with respect to the traffic intensity," *Journal of Applied Probability*, vol. 20, pp. 916–919, 1983.

[16] H. L. Lee and M. A. Cohen, "A note on the convexity of performance measures of M/M/c queueing systems," *Journal of Applied Probability*, vol. 20, pp. 920–923, 1983.

[17] G. L. Nemhauser and L. A. Wolsey, *Integer and combinatorial optimization*. Wiley New York, 1988, vol. 18.

[18] 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.

[19] J. Zahorj, K. C. Sevcik, D. L. Eager, and B. Galler, "Balanced Job Bound Analysis of Queueing Networks," *Communications of the ACM*, vol. 25, no. 2, pp. 134–141, 1982.

[20] M. Marzolla, "The `qnetworks` toolbox: A software package for queueing networks analysis," in *Analytical and Stochastic Modeling Techniques and Applications, 17th International Conference, ASMTA 2010, Cardiff, UK, Proceedings*, ser. Lecture Notes in Computer Science, K. Al-Begain, D. Fiems, and W. J. Knottenbelt, Eds., vol. 6148. Springer, Jun.14–16 2010, pp. 102–116.