Equity-Effectiveness Tradeoff in the Allocation of Flows in Closed Queueing Networks

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—Tradeoffs between multiple dimensions of performance are inherent in the allocation resources in many systems, particularly those with multiple stakeholders. This paper presents numerical results for the case of allocating flows in central-server closed queueing networks considering several inequity measures over many network configurations. These results show the conflict between server and customer perspectives of equity using multiple measures (server utilization, flow, wait time, and queue length). This paper first compares the effectiveness of the most equitable allocation relative to the most effective allocation for many network configurations. The complete efficient tradeoff is then generated using an optimization methodology. The results indicate that for low levels of server rate heterogeneity, all equity measures provide zero inequity allocations with high levels of effectiveness. However, as server rate heterogeneity increases, the total system effectiveness decreases and significant differences between the inequity measures are evident. Further, inequity with respect to wait time, server utilization, and queue length can be eliminated with relatively small impact on total system throughput (i.e., system effectiveness). In contrast, reductions in inequity with respect to customer flow incur large decreases in total system throughput.

I. INTRODUCTION

System performance generally has many dimensions. One common conceptualization includes effectiveness, efficiency, and equity. These are often in conflict, requiring system designers to make tradeoffs between performance along these dimensions. It is common for allocations which are optimal in some sense to be unfair, or suboptimal, in another sense. For example, in parallel queueing networks, it is well-known that the minimum wait time allocations overload faster servers generally [1]–[4]. For the case where these servers are people working in a call center, this minimum wait time allocation is not feasible. As a result, some recent work has sought to add constraints on server inequity [5]. This concept of inequity has been formalized and studied for many years in facility location and a review of these methods found over twenty structural forms for calculating inequity in the literature [6].

This paper considers the equity of flow assignment in closed queueing networks from the perspective of servers as well as customers. There are two main research questions of interest for this work. First, how do different inequity measures affect the effectiveness of the system (i.e., how does the total throughput of the most effective allocation compare to that of the most equitable allocation)? And secondly, how does the system configuration influence these differences? In addition, an efficient frontier will be presented to illustrate the tradeoff between effectiveness and equity for each inequity measure.

The measure of interest for this work is Coulter’s measure [9] of total inequality, \( I \), given generally by

\[
I = \sqrt{\frac{1}{N} \sum_i \left[ \frac{E_i}{E} - \frac{A_i}{A} \right]^2}
\]  

where \( E_i \) and \( A_i \) are the effect and relevant attribute for group \( i \) while \( E \) and \( A \) are the mean values across all \( N \) groups. Equation (1) provides a measure of the match between the effect (e.g., work being done) and some group attribute (e.g., importance of group), while penalizing higher levels of inequity, and providing a unit-less, normalized measure in a way which adheres to accepted guidelines — in particular scale invariance and the principle of transfers [6], [10]. It should be noted that equity measures using different groupings (e.g., customer vs. server) can be in conflict [11], and further that different equity measures (e.g., politeness vs. fairness) can conflict using the same groups [12].

While equity has been studied extensively in facility location [6] and resource allocation problems (e.g., [10], [13]), relatively little work has been done using queueing system...
models. Known work includes evaluating sequencing policies [12], [14], allocation of voting machines [15], [16], and location and dispatching of emergency medical services [11], [17].

The prior work finds that modest decreases in effectiveness can lead to dramatic increases in equity. For example, in one case presented in [10], an effectiveness decrease of only 1.4% allowed for removal of all inequity in the allocation. In the traveling salesman context, a particular equity objective has been shown to provide solutions having distances which are upper bounded by 3/2 of the minimum distance (i.e., effectiveness only) solution [18].

Multi-objective optimization methods are commonly used which explicitly include equity in the objective function along with effectiveness [10] or both effectiveness and efficiency [13]. Other work uses single objectives designed to increase equity [18] or equity constraints [11]. In contrast, this paper takes an implicit approach in order to identify the tradeoff between equity and effectiveness for several measures of inequity using an exploratory numerical study.

III. PROBLEM FORMULATION AND METHODOLOGY

The central-server network structure of interest is shown in Fig. 1. Here entities repeatedly travel from any number, \( n \), of parallel servers to a single central server. Each server has an associated congestion (latency) function which is increasing with customer flow. There are a fixed number of entities, \( N \), cycling in this network. Let \( \mathcal{P} \) be the set of parallel nodes (\(|\mathcal{P}| = n\)). Then, let \( i \in \mathcal{P} \) be a parallel node.

\[
\begin{array}{c}
\lambda_1 \\
\mu \\
\lambda_n \\
\mu_1 \\
\ldots \\
\mu_n \\
\end{array}
\]

Fig. 1. General central-server network with \( n \) parallel nodes with flows, \( \lambda_i \) and mean service rates, \( \mu_i \) shown.

The general approach taken is to compute the most equitable allocation of flows in the network and compare with the most effective allocation for a wide range of system configurations (i.e., service rates). To do so, measures of inequity and computation methods for both zero inequity and maximum effectiveness allocations are required. These items are covered in turn in the next three sections.

A. Inequity Measures

Equity from the perspectives of both servers and customers are considered. In both cases, the deviation from the average effect is of interest (i.e., inequity). Consider the case where each parallel server is compensated for its individual flow. The most equitable allocation would then prescribe equal flows among the parallel servers and equal attributes (i.e., importance). The total inequity, as given by Coulter’s measure [9], is then

\[
I_{\text{flow}} = \sqrt{\frac{1}{n} \sum_{i \in \mathcal{P}} \left( \frac{\lambda_i}{\lambda} - 1 \right)^2}.
\]

An alternate measure from the parallel server perspective is to consider equal utilizations as equitable, making the total inequity measure

\[
I_{\text{utilization}} = \sqrt{\frac{1}{n} \sum_{i \in \mathcal{P}} \left( \frac{\rho_i}{\rho} - 1 \right)^2}.
\]

From the customer perspective, equal average wait times may be considered best, leading to a total inequity of

\[
I_{\text{wait}} = \sqrt{\frac{1}{n} \sum_{i \in \mathcal{P}} \left( \frac{W_i}{W} - 1 \right)^2}.
\]

Alternatively, equal average queue lengths may be considered equitable by customers, giving a total system inequity of

\[
I_{\text{lengths}} = \sqrt{\frac{1}{n} \sum_{i \in \mathcal{P}} \left( \frac{L_i}{L} - 1 \right)^2}.
\]

B. Computing the Maximum Equity Allocation

In order to find the most equitable (i.e., zero inequity) allocation, an approximation is made as exact closed queueing networks analysis is difficult [19]. Whitt’s finite population mean (FPM) method [20] is a closed network approximation in which the nodes in the network are assumed to be independent queues in an open network thereby providing closed-form solution methods. This open network is then constrained such that 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 approximation has been shown to be asymptotically correct for large number of customers or nodes [20]. Further, the ratio of server utilizations in the approximate open network is identical to the same ratio in the closed network, motivating its appropriateness for the present study.

This FPM approximation reduces the problem of finding the zero inequity allocation to a simple line search in which the appropriate relationship between flows is fixed according to the inequity measure and they are increased together until the total number of entities in the system reaches the specified number. For example, for the flow inequity measure, all flows are required to be equal (\( \lambda_i = \lambda/n \) where \( \lambda \) is the total system throughput). For the queue length (or utilization) inequity measure, the flows are in the same proportion as the relative service rates (i.e., \( \lambda_i = \lambda \mu_i / \sum_{j \in \mathcal{P}} \mu_j \)). This ensures that all utilizations (and thus queue lengths) are equal at \( \rho = \lambda / \sum_{i \in \mathcal{P}} \mu_i \). In contrast, for the wait time inequity measure, each flow is related to the corresponding service rate by \( \lambda_i = \mu_i - \alpha \) where \( \alpha \leq \min_i \mu_i \) is uniform across all parallel servers, giving a uniform wait time of \( W = 1/\alpha \).

C. Computing the Maximum Effectiveness Allocation

In order to compare these zero inequity (i.e., fair) allocations with the maximum effectiveness allocation (i.e., the one giving the best possible total system throughput), an approximation method first presented in [21] is used. This method is briefly summarized here. Consider the following parallel-cycle partition (PCP) problem formulation in which
the continuous flows, $\lambda_i$, are the decision variables, the constraint is the number of entities in the system, and the objective is to maximize the total (weighted) system throughput (i.e., effectiveness). The relevant variables are shown in Table I along with additional notation.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda_i$</td>
<td>Flow assigned to parallel server $i$ (entities/time)</td>
</tr>
<tr>
<td>$\mu_i$</td>
<td>Average service rate of parallel server $i$ (entities/time)</td>
</tr>
<tr>
<td>$\rho_i$</td>
<td>Average utilization of parallel server $i$</td>
</tr>
<tr>
<td>$N$</td>
<td>Total number of entities in the system</td>
</tr>
</tbody>
</table>

Using the FPM approximation, Little’s Law (expected number of entities at each server equals average arrival rate times the expected wait time [19]) can be used to form a constraint on the total number of entities. The PCP problem for central-server networks can then be written as the following nonlinear program:

$$
\text{max} \sum_{i=1}^{n} w_i \lambda_i \quad (2)
$$

s.t. $\sum_{i=1}^{n} \lambda_i C \left( \sum_{i=1}^{n} \lambda_i \right) + \sum_{i=1}^{n} \lambda_i D_i(\lambda_i) \leq N \quad (3)$

$$\lambda_i \geq 0. \quad (4)$$

Here $C(\lambda), D_i(\lambda_i): \mathbb{R}^+ \rightarrow [0, \infty]$ are general mean congestion (i.e., latency or wait time) functions for the central and parallel servers, respectively. For M/M/1 queues, these are given by [19]:

$$C(\lambda) = \frac{1}{\mu - \lambda}$$

and

$$D_i(\lambda_i) = \frac{1}{\mu_i - \lambda_i}.$$  

Note that (3) can be written as an inequality because of the form of the objective (i.e., sum of positive terms strictly increasing in $\lambda_i$). To obtain the maximum effectiveness allocation, $w_i = 1 \forall i \in P$.

IV. NUMERICAL RESULTS

The results are divided into two sections: the first shows the relative total throughput (i.e., system effectiveness) of the zero equity allocations according to all four measures relative to maximum effectiveness allocation while the second explores the continuous tradeoff between effectiveness and efficiency using weighted optimal flows for a single network configuration.

A. Most Equitable Allocations

For these tests, the central server service rate is fixed at 4 customers/hour while the other three are varied systematically from 0.1 to 3.6 customers/hour in 0.5 increments. This experimental design leads to 512 network configurations. For each configuration, the zero equity allocation is determined by the line search methodology described earlier for each of the four measures of inequity. Note that the length and utilization perspectives have the same zero equity allocation for M/M/1 queues because $L = \frac{1}{\rho_{\text{max}}}$ [19], thus only length is shown. In addition, the most effective allocation is found by implementing the PCP problem in MATLAB and solving using fmincon’s interior-point method. To see the tradeoff in total performance for the zero inequity allocations as a function of system configuration, the relative throughput of each inequity perspective is plotted against the inequity of parallel server rates in Fig. 2.

As one might expect, for cases in which all of the parallel servers have the same service rate (rate inequity of zero), there is no performance degradation in the zero inequity allocation for all equity perspectives. However, as parallel server heterogeneity (i.e., rate inequity) increases, dramatic differences between the total system throughput of the zero inequity allocations are observed among the inequity measures. As server rate inequity increases, maintaining zero flow inequity incurs much higher degradation of total system throughput, reaching nearly a factor of 12 less than the zero length inequity allocation. Additionally, all zero wait time inequity allocations have exactly the same total throughput as the corresponding zero length inequity allocation. Further, while zero length (utilization) and flow inequity allocations always exists, a zero wait time inequity allocation only exists for 37.9% of the configurations tested. This is due to the fact that the minimum wait time is lower bounded by $\frac{\text{max}(1/\mu_i)}{N}$. Therefore, an allocation may not exist which has such a large wait time for the faster servers when the number of customers is small.

Finally, the inequity of the maximum effectiveness allocation is considered with respect to each inequity measure. These measures are $I_{\text{flow}} = 1.23$, $I_{\text{flow}}^* = 0.48$, $I_{\text{utilization}} = 0.43$, and $I_{\text{length}} = 0.85$. Because these measures are normalized, they can be directly compared. Thus, the maximum effectiveness allocation has smallest utilization inequity and largest flow inequity.

In summary, the results of this section show that for low parallel server rate inequity, all measures of inequity produce zero inequity allocations which have close to the maximum
effectiveness. However, as rate inequity increases, allocations which have zero flow inequity show marked declines in total effectiveness relative to those zero inequity allocations with respect to the other inequity measures.

B. Equity-Effectiveness Tradeoff

The previous results showed that large performance differences exist between zero inequity allocations depending on the inequity measure used. In this section, the continuous tradeoff between equity and effectiveness is explored through systematically varying the weights, $w_i$, in the PCP problem in order to identify the set of Pareto optimal allocations for each inequity measure. To do so, a single network configuration is used in which $\mu_2 = 4$, $\mu_1 = 2$, $\mu_2 = 1$, and $\mu_3 = 0.5$ while $N = 5$. The rate inequity for this case is $I_{\mu} = 0.655$. The objective weights are chosen uniformly from the 3-simplex.

The resulting total system throughput has a high plateau for weights near the center of the simplex (i.e., equal weights) as expected (maximum throughput: 2.09). The complete surface can be seen in Fig. 3. The throughput drops off significantly for objective weights which give importance to servers 2 and 3 (the right hand side of the simplex).

The inequity surfaces according to each measure on the weight simplex can be seen in the top plots of Fig. 4 (flow), 6 (wait time), 5 (utilization), and 7 (length). When flow inequity is considered, the best allocations prefer parallel servers 2 and 3, as can be seen in Fig. 4. The total throughput of these low flow inequity allocations is relatively poor as expected from the earlier results. In contrast, the low wait inequity allocations have relatively good throughput since these are closer to the center of the simplex as shown in Fig. 6. Length and utilization inequity share a common minimum as expected from the previous zero inequity results (see Fig. 4). However, the region of low inequity is much greater for the utilization measure.

In order to identify the efficient frontier (i.e., the set of Pareto efficient allocations), these surfaces are compressed along with the corresponding effectiveness to show inequity as a function of throughput, as shown in the bottom plots of Fig. 4, 6, 5, and 7. Many of the allocations are Pareto inefficient in that there is another allocation which has either higher throughput for the same level of inequity or lower inequity for the same level of throughput. The zero inequity allocations from before can be seen where the points reach the horizontal axis, while the maximum effectiveness allocations are located at the point at which the maximum throughput is reached (far right point of each plot). The efficient frontier is then the lower right convex hull of each figure, encompassing the allocations from the maximum throughput allocation to the zero inequity allocation. For example, for the flow inequity measure (bottom of Fig. 4), the frontier extends from the maximum effectiveness allocation (approximately the point $(2.1, 1.2)$) down to the zero inequity allocation (approximately the point $(1.2, 0)$).

The differences between the efficient frontier for the length, wait time, and server utilization inequity perspectives are relatively small. Furthermore, the total throughput of the zero inequity allocations are equal, as expected from the previous results (total throughput: 2). While the total throughput is
Fig. 5. Server utilization inequity for the locus of allocations to the optimal parallel cycle partition problem with weights chosen uniformly over the entire simplex for $N = 5$ (top). Frontier of effectiveness and inequity (bottom).

equal, it is important to note that the underlying allocations are quite different ($\lambda = [1.14, 0.57, 0.29]$ for equal queue length/utilization while $\lambda = [1.5, 0.5, 0]$ for equal wait times).

The efficient frontier for the flow perspective, on the other hand, is significantly different from the others (see bottom of Fig. 4). The tradeoff between effectiveness and equity is much more severe in the sense that reductions in inequity produce larger decreases in total system effectiveness.

In summary, these results show the shape of the efficient frontier for all four inequity measures. Consistent with the first set of results, the flow inequity measure showed marked differences from the other three measures. In particular, reductions in this inequity measure result in larger reductions in system effectiveness. The zero inequity throughput for the other three inequity measures was also shown to be identical as expected from the previous results.

V. CONCLUSIONS

This paper has explored the effect of different measures of system inequity on the total effectiveness of closed central-server queueing systems through an exploratory numerical study. From the servers’ perspective, the more effective equity measure was found to be equal server utilizations. From the customers’ perspective, both wait and length equity measures provide similar effectiveness, with the wait time measure providing a lower measure of inequity for the most effective allocation. The results for wait time, length, and utilization inequity measures echo those previously found in the literature in that a small reduction in total system effectiveness can result in dramatic decreases in total system inequity [10]. However, a flow inequity measure provides more significant declines in effectiveness. Further, regardless of the inequity measure used, increased parallel server rate inequity was shown to decrease system effectiveness of the zero inequity allocation. While the length and utilization perspectives share the same zero inequity allocation, they have different tradeoffs with effectiveness. Similarly, length and wait time were shown empirically to share the same zero inequity effectiveness while having different underlying allocations. Thus, it was found that there are conflicts between the inequity measures in the sense that all four cannot be met concurrently by one allocation.

A customer flow inequity measure, however, was found to
be profoundly different than the other three in that the tradeoff with effectiveness was more severe in general and that system effectiveness degraded more quickly as server heterogeneity increases. This is particularly important for practice in cases where servers are compensated for their work according to the flow allocated to them. These results indicate that allocations which reduce flow inequity between heterogeneous servers suffer from particularly poor system-level effectiveness relative to the most effective allocation. In contrast, a server utilization inequity measure supports much better system effectiveness.

This work can be extended to more general bipartite networks in which multiple central servers exist. Additionally, future work is needed to identify means of choosing the objective weights in order to select the point on the efficiency frontier directly. Consider a function $w = f([\mu_1, \ldots, \mu_n], \alpha)$ where $a \in [0, 1]$ and $\alpha = 1$ gives the maximum throughput while $\alpha = 0$ gives the minimum inequity allocation. Finally, the effect of travel time between nodes was also not considered in this work.

**ACKNOWLEDGMENT**

This material is based upon work supported by the National Science Foundation under Grant No. 1313589.

**REFERENCES**