A Network Flow Model Analysis of a Multiqueue Operator Service System with Priority

Jason Liu and David A. Kettler
Bell Telephone Laboratories, Inc., Holmdel, New Jersey, U.S.A.

ABSTRACT

A multiqueue operator service system with server priority is formulated as a bipartite congestive network flow problem. The intensity of the intraflow on each arc is dependent on the congestive conditions at each node. Feedback equations are introduced which relate the congestive conditions at each node to the intensity of the intraflow on each arc. This network flow model is transformed to a fixed point problem. It is further shown that if the congestive functions associated with each node of the network are continuous with respect to load, then a fixed point always exists. For specific load and server parameters, the fixed point is interpreted as the flow intensity on each arc under an equilibrium state.

This mathematical model is applied to the traffic characterization of a large call distributor system handling directory assistance calls. This particular distributor allows a limited number of calls which are blocked at the preferred server group to intraflow to nonpreferred groups with an idle server. A preferred group of servers primarily handles calls from a geographical cluster of customers. Thus the preferred server has a shorter servicing time than a nonpreferred server. Yet, because of the intraflow capability, the efficiencies of large team operation are retained. Empirical data support the intraflow traffic characteristics predicted by the mathematical model.

Through the use of this model the amount of intraflow for each cluster to nonpreferred servers becomes a predictable quantity for a forecasted offered load and specified server parameters. Thus, results generated by the model become an integral component in the determination of the minimum number of operators in each group required to provide objective service to the customer.

INTRODUCTION

A multiqueue operator service system with server group priority is formulated as a bipartite congestive network flow problem. This network flow model is transformed to a fixed point problem. This mathematical model is then applied to the traffic characterization of a large call distributor system which employs server (operator) group priority in handling directory assistance calls. Empirical data are presented which verify the intraflow traffic characteristics predicted by the mathematical model for this call distributor. Results generated by the model become an integral component in the determination of the minimum number of operators in each group required to provide objective service to the customer.

THE FLOW PROCESS

The multiqueue service system under consideration is characterized by an incoming traffic route and preferred server group for each queue. This particular distributor allows a limited number of calls which are blocked at the preferred server group to intraflow (overflow within the call distributor) to nonpreferred groups with an idle server. When all servers are busy the calls are placed in the queue associated with the originating route. A limit is placed on the number of calls served by each group at a point in time which have intraflowed to that group. The limit is the same for all groups. Groups and routes reciprocate calls with each other through intraflow. Also, server group holding times are heterogeneous and holding times for calls which have intraflowed to a nonpreferred server group are different (generally higher) from the holding time for the same call served by its preferred server group.

THE BIPARTITE NETWORK FLOW MODEL

A bipartite network [1] represents the flow process. This network contains two columns of nodes (Figure 1). Each node is represented by a circle. The column of G nodes on the right represents G server groups. Each node on the right represents a server group. The number of servers (or operators) allocated to each server group j is denoted by the parameter \( S_j \) (\( j=1,2,\ldots,G \)). For each server group node there is a corresponding incoming traffic route node on the left as its image node. Also, the server parameter \( S_j \) (\( j=1,2,\ldots,G \)) associated with each route node is identical to that of its associated group node. To the left of each route node, there is an "input" arc. Associated with each input arc, there is a number \( r_i \) which represents the intensity of calls offered to route \( i \). It is referred to as the route call intensity. To the right of each server group there is an "output" arc. Associated with each output arc, there is...
a number $q_i$; $q_i$ represents the intensity of calls carried by group node $i$.

Between the column of route nodes and column of group nodes, there are $G^2$ "intraflow arcs" connecting each route node to each group node. Among these $G^2$ intraflow arcs there are $G$ of them which connect route nodes and their associated group nodes. These arcs are called home arcs, while the remaining $G(G-1)$ intraflow arcs are called "overflow" arcs. Associated with each intraflow arc is an intraflow intensity value $\lambda_{ij}$. $\lambda_{ij}$ represents the number of calls initially offered as route $i$ calls but served by group $j$ (per unit time). $\lambda_{ii}$ represents the number of calls on route $i$ served by its preferred group per unit time. Associated with each intraflow arc is also a parameter $h_{ij}$, which represents the intraflow average holding time of a route $i$ call served by group $j$. $h_{ii}$ is the average holding time of a route $i$ call served by its preferred group. The function $C_i(S_i, V)$ is defined later.

This mathematical model assumes statistical equilibrium and no call abandonment. The latter assumption reflects the conservation of flow at each node.

**INTRA flowing INTENSITY RELATIONS**

Now, given the values of $h_{ij}$, $r_i$, and $S_i$ can the values of $\lambda_{ij}$ for $i, j = 1, 2, \ldots, G$ be determined? Heuristically, one can argue that during a unit time period the number of calls served by nonpreferred server groups is proportional to the degree of congestion at the preferred server group and is also proportional to the degree of availability of an idle server in nonpreferred server groups.

A large class of functions can be used to express the degree of congestion in a service system. These functions have the following properties in common:

1. It is a real valued function $C(S, V)$ of two real non-negative arguments; namely, the supply argument $S$ and the demand argument $V$.
2. $0 \leq C(S, V) \leq 1$ for $S, V$.
3. $C(S, V) \leq C(S_2, V)$ when $S_1 \geq S_2$ for fixed $V$.
4. $C(S_1, V) \leq C(S_1, V_2)$ when $V_1 \leq V_2$ for fixed $S$.

Any function with these properties shall be called a "congestive function." The Erlang C delay probability function [2] is an example of such a congestive function. The first arguments are the number of servers or trunks, and the second arguments are the loads in units of Erlangs. The degree of availability can be expressed in terms of congestive functions. For instance, let $C(S, V)$ be a certain congestive function, then $1 - C(S, V)$ can be used to express the degree of availability.

The bipartite congestive network can now be specified by the $G$-tuple $(G, a, h, r, S, C, A)$ where $G$, $h$, $r$, and $S$ are defined above. The parameter $a_{ij}$ is defined as $a_{ij} = 1$ or $a_{ij} = 0$. The $i$-th incoming node to the $j$-th server node, $C$ is the congestive function space of the $G$ congestive functions $C_i(S_i, V)$ where $C_i(S_i, V) = (1 - C_i(S_i, V))$ and $C_i$ is the congestive function associated with incoming group node $i$. $A$ is the availability function space which contains $G$ availability functions $f_{ij}(S_i, V)$ where $f_{ij}$ is a congestive function. $C_i$ and $C_i$ are not necessarily the same for $i = 1, 2, \ldots, G$.

$V_i$ is the carried load of server group node $i$. Therefore, $V_i = \sum_{k=1}^{G} \beta_{ik}^i h_{ik} V_i$ where $V_i = 1, 2, \ldots, G$ and

$$Z_i = \sum_{k=1}^{G} S_k a_{ik} Y_{ik}$$

is the effective number of nonpreferred servers for route node $i$ and

$$U_i = \sum_{k=1}^{G} V_k a_{ik} Y_{ik}$$

is the effective nonpreferred server load for route node $i$. $\gamma$ is a discount coefficient matrix which weights the effectiveness of nonpreferred servers. In this paper $Y_{ij} = 1 - \gamma_{ij}$ where $\gamma$ is the $G 

proportionality coefficient matrix which represents the overflow controller characteristics of the system. $0 \leq \gamma_{ij} \leq 1$ for $i, j = 1, 2, \ldots, G$ for $i \neq j$.

The overflow call intensity is determined by the congestive function of the route node $i$ and the availability function of nonpreferred servers for the route node $i$. The overflow call intensity to a specific nonpreferred server group $j$ is then determined by the ratio of the specific nonpreferred server group $j$ availability to the total nonpreferred server availability $Y_i$.

$$Y_i = \sum_{k=1}^{G} \frac{1}{\gamma_{ik}}$$

Thus, the intraflow mapping $F(A)$ associated with the bipartite congestive network assumes the form:

$$F(A) = \gamma_{ij} C_i(S_i, V_i) \left[1 - C_i(S_i, V_i)\right] \frac{1 - C(S_i, V_i)}{Y_i}$$

for $Y_i \neq 0$.

$$F(A) = 0$$

$F(A)$ is a mapping from $R^d = R^G$ where $d$ is the degree of freedom of the underlying network as defined by

$$d = \sum_{j=1}^{G} a_{ij} - G$$

$F(A)$ is not defined. However, $\lambda_{ij}$ is determined by the conservation of flow at route node $i$ as follows:

$$\lambda_{ij} = r_i - \sum_{k=1}^{G} \lambda_{ik} a_{ik}$$

**EXISTENCE OF FIXED POINT**

One basic property of this bipartite congestive network is that the associated intraflow mapping $F(A)$ always has
a fixed point $\lambda$ for the fixed point problem $\lambda = F(\lambda)$ under very general conditions. Thus, by the use of the Brouwer Fixed Point Theorem [3,4] we have the following existence theorem:

**THEOREM**

If the congestive functions in $C$ and $A$ are continuous functions with respect to the second arguments and let

$$B = \{(\lambda, 0 \leq \lambda_{ij} \leq \gamma_i, 1 \leq i \leq G, 0 \leq j \leq G, i \neq j \mid \lambda_{ij} \neq 0, \lambda \in R^2 \}$$

and if for any $\lambda$ in $B$ such that $Y_i(\lambda) = 0$, $1 \leq i \leq G$ implies that $C^T(\lambda^i_U, U_j) = 1$, then the associated intraflow mapping $F(\lambda)$ always has a fixed point $\lambda^F$ in $B$. (Proof: See Appendix I.

If $C$ and $A$ consist of only the Erlang C function, then it is easy to show as a corollary to the above theorem that the associated fixed point problem always has a fixed point (see Appendix I).

Therefore, by making the assumption on the congestive function at each node, and by making the assumption on the intraflow functional relationship, which relates the intraflow intensity on each arc with congestive conditions at each node, every bipartite network flow problem intrinsically defines a fixed point problem. The fixed point is interpreted as the flow intensity on each arc under an equilibrium state.

The solution of the fixed point problem can be obtained by the use of a modified Merrill's Algorithm [5,6] and other methods of numerical calculation.

**DIRECTORY ASSISTANCE APPLICATION**

A major effort in achieving the minimization of traffic operating force expenses is the consolidation of small operator offices into large offices to realize the benefits of the efficiencies of large team operation. The volume of Directory Assistance traffic and the associated operating expenses of handling it continue to increase at a high rate. Thus, large operator team operation (i.e., increased operator utilization to achieve an objective delay) becomes increasingly attractive. However, as the team grows larger more records are placed at the operator position. And the operator now answers calls from a very large geographical area requesting Directory Assistance for a large geographical area. Therefore, the average operator work load per call (server holding time) increases and thereby conflicts with the goal of minimizing operating expenses.

Geographic grouping is a concept envisioned to resolve this conflict, i.e., to have simultaneously both large team efficiency and low average work times. By organizing the geographical area into several smaller territories and associating a preferred group of operators to serve calls primarily from each geographical cluster of customers the average operator work time per call can be maintained at a low level due to increased operator familiarity with the locality and possibly more efficient record configuration. Thus, the preferred operator has a shorter serving time than a nonpreferred operator. Yet because of the intraflow capability the efficiencies of large team operation are retained. For the automatic call distributor under consideration all server groups and routes reciprocate intraflow calls with a limit established as discussed earlier. This limit will hereafter be referred to as the foreign call allowance (FCA).

Through simple manipulation of Equation (1) and the use of Erlang C as the congestive function for all nodes the theoretical proportion of intraflow to nonpreferred operator groups for each route can be predicted by the following formula:

$$\frac{\min(G, FCA)}{S_i} - 2 \frac{\min(G, FCA)}{S_i}$$

The proportionality coefficient matrix $S$ must be determined. It is of interest to note that as the utilization of the operator's position time spent handling calls increases the intraflow to nonpreferred operator groups for a particular route increases and then decreases as shown in Figure 2. Prior to actual implementation limited simulation results and engineering judgment are used to establish the necessary relationship of $S$ to independent parameters.

In the application of this approach to the geographic grouping concept an a priori relationship for $S_i$ was determined to be

Data was then collected on the system in service in East Orange, New Jersey, USA. Theoretical proportions of call intraflow to nonpreferred operator groups compared to the measured values are shown in Figures 3 and 4 for two of the five geographic grouping routes employed in this system. A foreign call allowance (limit) of $h$ was used during this measurement period. An almost linear relationship exists between the theoretical value and actual value for all routes of the automatic call distributor, and, in fact, appears as a $45^\circ$ line in the figures. Thus, the model well represents the actual process.
The foreign call allowance (FCA) was changed from 4 to 10 for another measurement period. Based on a comparison made on data from this period a rotational adjustment factor was introduced for \( \theta \). Thus,

\[
\theta_i = (1-k) \left[ \frac{\min(S_i, FCA)}{S_1} \right] \frac{\min(S_i, FCA)}{S_1}
\]

\( \forall i = 1, 2, \ldots, G \)

\[
K = \begin{cases} 
0.27 \times \ln(FCA+1) & \text{for } FCA \geq 4 \\
0 & \text{for } FCA < 4 
\end{cases}
\]

Based on this model the amount of intraflow for each cluster to nonpreferred servers becomes a predictable quantity for a forecasted offered load and specified server parameters. The proportion of intraflow for all clusters (routes) can be determined as shown in Figure 5 as a function of the mean route call value of all routes. The bands indicated account for differences in server group holding times and server group sizes from the mean for all server groups. Actual data is shown in Figure 6.
The decrease in intraflow is a result of the limit placed on intraflow. Results from a separate study of the operator work times showed that a call which was served by a nonpreferred operator required an operator holding time which exceeded the preferred operator holding time to serve that call by approximately 20 percent. This fact coupled with the retention of large team efficiency have proven the economic viability of this concept.

Thus, results generated by the model become an integral component in the determination of the minimum number of operators in each group required to provide objective service to the customer. Through the use of this approach operator force management is based on the theoretically predicted offered load. Divergence effects produced by characteristics of a multiqueue operator service system operator requirements for objective system operation.

The authors are indebted to Romesh Saigal of Bell Telephone Laboratories for his suggestions on the numerical assistance of the American Telephone and Telegraph Company for his cooperation on the numerical computations of the fixed point and to J. V. Fardelmann of the American Telephone and Telegraph Company for his assistance with the application study.

REFERENCES


APPENDIX I

THEOREM

If the congestive functions in C and A are continuous functions with respect to the second arguments and let

\[ B = \{ \lambda : 0 \leq \lambda_{ij} \leq r_{ij}, 1 \leq i \leq G, 1 \leq j \leq G, i \neq j, a_{ij} \neq 0, \lambda \in \mathbb{R}^d \} \]

and if for any \( \lambda \) in B such that \( Y_j(\lambda) = 0 \) implies that \( C^j(Z_{1j}, Y_j) = 0 \), \( 1 \leq i \leq G \), then the associated intraflow mapping \( F(\lambda) \) always has a fixed point \( \lambda^* \) in B.

PROOF

It can easily be shown that B is convex, compact, and nonempty in the finite dimensional vector space \( \mathbb{R}^d \).

Claim 1: \( F(\lambda) \) is a mapping from B into B. Let \( \lambda \) be any element in \( \mathbb{R}^d \) where d is the degree of freedom of the congestive network. Then the following are true:

1. \( 0 \leq F_{ij}(\lambda) \leq r_{ij} \) for those \( i, j \) with which \( a_{ij} \neq 0 \) and \( i \neq j \). \( V_i, j = 1, 2, \ldots, G \).

Claim 2: Group carried load \( V_j \) is a continuous function of intraflow intensity and intraflow holding times.

PROOF

\[ Y_j = \sum_{k=1}^{G} a_{kj} h_j \]

\[ Y_j = \sum_{k=1}^{G} a_{k_j} h_j \]

\[ \lambda_{ij} = r_i - \sum_{k=1}^{G} F_{ik}(\lambda) a_{ik} \]

\[ r_i \geq \sum_{k=1}^{G} \lambda_{ik} a_{ik} \]

\[ 0 \text{ by Claim 1} \]

Claim 3: \( F(\lambda) \) is a continuous function in B at all those points \( \lambda \) where \( Y_j(\lambda) \neq 0 \) for \( i = 1, 2, \ldots, G \).

PROOF

Let \( B_1 = \{ \lambda = \lambda e_B, Y_j(\lambda) \neq 0 \text{ for } i = 1, 2, \ldots, G \} \).

Since all the congestive functions \( C(S, V) \) in C and A are continuous functions with respect to the second arguments, \( V_j \) and \( V \) is a continuous function of \( \lambda \) by Claim 2.

\[ C_1(S, V) = [1 - C^j(Z_{1j}, Y_j)] [1 - C^j(V_j, Y_j)] \]

is a continuous function of \( \lambda \) for \( i, j = 1, 2, \ldots, G. \) \( j \neq i \) and \( a_{ij} \neq 0 \).

Since \( Y_j \) is also a continuous function of \( \lambda \) over B at those points \( \lambda \) where \( Y_j(\lambda) \neq 0 \), \( F_{ij} \) is a continuous function of \( \lambda \) over \( B_1 \).
Claim 4: \( F(\lambda) \) is a continuous function of inflow intensity \( \lambda \) over \( B \) at those points \( \lambda \) for some \( Y_1(\lambda) = 0 \).

**Proof**

By definition, \( F_{ij}(\lambda) = 0 \) for \( Y_1(\lambda) = 0 \). Now for arbitrary \( \epsilon > 0 \), we can find \( \delta_1 > 0 \) and \( \delta_2 > 0 \) such that whenever \( |w-\lambda| < \min(\delta_1, \delta_2) \),

\[
|Y_1(w) - Y_1(\lambda)| = |Y_1(w) - 0| \leq \epsilon/r_i S_{ij} ,
\]

\[
(1 - c_i^j(z_i^j, u_i^j(w))) - (1 - c_i^j(z_i^j, u_i^j(\lambda))) =
\]

\[
|1 - c_i^j(z_i^j, u_i^j(w))| - |1 - c_i^j(z_i^j, u_i^j(\lambda))| =
\]

because both \( Y_1 \) and \( c_i^j(z_i^j, u_i^j) \) are continuous functions of \( \lambda \) for \( \lambda \in B \).

Now for all those \( w \) such that \( |w-\lambda| < \delta \)

\[
|F_{ij}(w) - F_{ij}(\lambda)| = |F_{ij}(w) - 0|
\]

\[
= \sum_{k} C_1(S_1, V_1) \frac{1 - C_1(S_1, V_1)}{Y_1(w)} \cdot r_{ij}
\]

\[
\leq \sum_{k} \epsilon \frac{1 - C_1(S_1, V_1)}{Y_1(w)} \cdot r_{ij}
\]

\[
\leq \epsilon \frac{1 - C_1(S_1, V_1)}{Y_1(w)}
\]

because of the fact that

\[
|Y_1(w)| = \sum_{k} |1 - C_k(S_k, V_k)| a_{ik}
\]

\[
= \sum_{k} \frac{|1 - C_k(S_k, V_k)| a_{ik}}{k \neq i}
\]

\[
= \sum_{k} |1 - C_k(S_k, V_k)| a_{ik}
\]

\[
\geq |1 - c_j^i(S_j^i, V_j^i)|
\]

for all those \( j \) such that \( j \neq i, a_{ij} \neq 0 \). Therefore, \( F_{ij}(\lambda) \) is a continuous function over \( B \) even at those points where \( Y_1(\lambda) = 0 \). Since \( F_{ij} \) is continuous at all points in \( B \), \( F \) is a continuous mapping over \( B \) into \( B \). Now by the Brouwer fixed point theorem, a fixed point \( \lambda^* \) always exists such that \( \lambda^* = F(\lambda^*) \) and \( \lambda^* \in B \).

Corollary: If \( C \) and \( A \) consist of only the Erlang C delay probability function, then the associated fixed point problem always has a fixed point.

**Proof**

It can be shown that Erlang C function is a continuous function in its second argument. We claim it is also true that

\[
y_1 = 0 \implies c_i^j(z_i^j, u_i^j) = 1
\]

when all the congestive functions involved are the Erlang C delay probability function.

Now suppose

\[
y_1(\lambda) = 0
\]

then

\[
\sum_{k} |1 - C_k(S_k, V_k)| a_{ik} = 0.
\]

Therefore, \( C_k(S_k, V_k) = 1 \) for all those \( k \) such that \( a_{ik} \neq 0, k \neq i \). However, \( C_k(S_k, V_k) = 1 \) implies that \( S_k \neq S_i \). Since

\[
z_i = \sum_{k} S_k a_{ik} Y_i
\]

\[
u_i = \sum_{k} V_k a_{ik} Y_i
\]

we have

\[
u_i \geq z_i
\]

Therefore, \( c_i^j(z_i^j, u_i^j) = C(z_i^j, u_i^j) = 1 \).

Now that both the conditions in the theorem are satisfied the proof of this corollary is completed.

Note that the condition \( Y_1 = 0 \implies c_i^j(z_i^j, u_i^j) = 1 \) verified in this corollary is equivalent to the following: when the total availability of the nonpreferred server groups with respect to route \( i \) is zero then all the nonpreferred server groups, when viewed as a single entity, would produce complete congestion to overflow traffic from route \( i \).