ELSEVIER

European Journal of Operational Research 88 (1996) 586–591

EUROPEAN JOURNAL OF OPERATIONAL RESEARCH

Theory and Methodology

Optimizing flow rates in a queueing network with side constraints

B. Pourbabai [a], J.P.C. Blanc [b], F.A. van der Duyn Schouten [b,*]

[a] *Department of Mechanical Engineering, The University of Maryland, College Park, MD 20742, USA*
[b] *Department of Econometrics, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, Netherlands*

Received May 1992; revised July 1994

Abstract

In this paper, modified versions of the classical deterministic maximum flow and minimum cost network flow problems are presented in a stochastic queueing environment. In the maximum flow network model, the throughput rate in the network is maximized such that for each arc of the network the resulting probability of finding congestion along that arc in excess of a desirable threshold does not exceed an acceptable value. In the minimum cost network flow model, the minimum cost routing of a flow of given magnitude is determined under the same type of constraints on the arcs. After proper transformations, these models are solved by Ford and Fulkerson's labeling algorithm and out-of-kilter algorithm, respectively.

Keywords: Queueing network; Stochastic optimization; Labeling algorithm; Out-of-kilter algorithm

1. Introduction

In this paper, stochastic versions of *the deterministic maximum flow model* and of *the deterministic minimum cost flow model* in a single commodity, directed, and capacitated network are presented. We consider a network of queues. On each arc in the network there is a service unit. Jobs enter the network at a source and leave the network at a sink, and they require service at each arc that they pass on their way through the network. In order to be able to apply standard results for so-called open queueing networks with product-form solution, cf. Walrand (1988), we make the following assumptions:

1) Jobs are generated at the sources according to Poisson (arrival) processes;

2) at each arc the service unit consists of a fixed number of servers, and the service times are independent, identically, negative exponentially distributed random variables;

3) the routing of jobs through the network occurs according to controllable random mechanisms, i.e., at each node the flows of jobs arriving at that node are superposed, while the departing flow can be split into separate streams over the outgoing arcs according to fixed but controllable probabilities (actually, these splitting probabilities are decision variables in our model);

4) no blocking occurs, i.e., the buffer spaces at the arcs are unbounded;

5) the network is in statistical equilibrium.

* Corresponding author.

SSDI 0377-2217(94)00217-7

Under these conditions, it is well known that the flow of jobs at each arc behaves as that at an infinite capacity multi-server Markovian (i.e., $M/M/S/\infty$) queueing system with a work-conserving non-anticipating queueing discipline.

For each arc in the network a threshold capacity is specified, together with an acceptable probability of finding congestion in excess of this threshold. *The aim of these models is either to maximize the throughput rate in the network or to minimize the cost of a given flow in the network such that the resulting probability of finding congestion along each arc of the network in excess of the given threshold does not exceed an acceptable value.* The decision variables are the intensities of the flows of jobs along the arcs, or, equivalently, the intensities of the Poisson arrival streams at the sources and the routing probabilities at the nodes. Without loss of generality we may assume that the network contains a single source and a single sink. In case of multiple sources (sinks) one can add an artificial node acting as a single source (sink) and connected to all real sources (sinks) by arcs containing service units with service capacity larger than that at any other arc, and with an acceptable probability of congestion equal to one. For the minimum cost problem we assume that each arc has a cost associated with it representing the cost of processing one job along that arc.

The queue on arc representation has been chosen for conformity with deterministic flow problems. It should be noted that in a queueing context it is more usual to visualize queues as nodes of a network and flows of jobs from one queue to another as arcs of a network.

These models can be applied for optimization of throughput or routing in a single product flexible manufacturing system, where each item has to be processed through various manufacturing phases (the nodes). The source (sink) represents the starting (finishing) phase in the process. Due to the flexibility of the system there exist various ways to process an item from one stage to another (the arcs). The processing time of an item between two neighbouring stages is represented by a random variable with known exponential distribution. Alternatively, an arc may represent a transportation phase, with exponential travel time distribution and a limited number of transport units. The processing or transport cost per item between two neighbouring stages is deterministic.

For a discussion of the deterministic versions of network flow problems, see, e.g. Murty (1976). In spite of the abundant amount of literature on performance analysis of stochastic queueing networks, the notion of optimization of flows is not extensively dealt with in literature. Kleinrock (1976, Chapter 5) considers a traffic flow assignment problem with the aim of minimizing the total average delay for networks of single-server queues. Several static optimization problems in the context of the design of manufacturing systems are discussed in Buzacott and Shanthikumar (1993). Dynamic optimization of flows, using dynamic programming arguments, in simple network structures is addressed by several authors (see Walrand, 1988, for references). The main issue of this note is that the static optimization of routing a single commodity within a stochastic environment is translated into well studied deterministic flow optimization problems. For a similar perspective regarding the transportation model we refer to Pourbabai (1990).

2. Notation

Let M denote the set of all nodes in the network, and let A denote the set of all arcs in the network, i.e., the set of all queueing stations. There is a single source $s \in M$ and a single sink $t \in M$. For each node $k \in M$ the set I_k contains all incoming arcs, and the set O_k contains all outgoing arcs at node k, i.e.,

$$I_k = \{i \in M : i \text{ such that } (i, k) \in A\},$$

$$O_k = \{j \in M : j \text{ such that } (k, j) \in A\}.$$

For each arc $(i, j) \in A$ the following quantities are given:

- S_{ij}: The number of parallel servers at the arc.
- μ_{ij}: The processing rate of a job by a server at the arc.
- K_{ij}: The desirable threshold for the number of jobs present at the arc; it is assumed that the

threshold K_{ij} is larger than or equal to the number of servers S_{ij}.

- α_{ij}: The acceptable probability of finding congestion along the arc in excess of K_{ij}, $\alpha_{ij} > 0$.
- c_{ij}: The cost of processing one job along the arc.

For each arc $(i, j) \in A$, λ_{ij} will denote the flow rate of jobs along the arc. These rates are the decision variables of the optimization problems to be considered.

Let N_{ij} be the random variable denoting the number of jobs present at arc $(i, j) \in A$. The probability mass function of N_{ij} is the same as the queue length distribution of an $M/M/S_{ij}$ queueing system which is a function of the fixed quantities μ_{ij} and S_{ij}, and of the flow rate λ_{ij}, see, e.g. Walrand (1988) for the explicit formulas. The flow rate λ_{ij} is not allowed to exceed the service capacity $S_{ij} \times \mu_{ij}$ at any arc $(i, j) \in A$ in order to maintain stability of the network.

3. Maximum flow rate problem

The *maximum flow rate problem* for a directed single commodity capacitated Markovian open queueing network problem can be stated as the following stochastic (i.e., chance constrained) optimization problem.

$$\text{Maximize } \gamma \tag{1}$$

subject to

$$\sum_{j \in O_k} \lambda_{kj} - \sum_{i \in I_k} \lambda_{ik} = \begin{cases} \gamma & \text{if } k = s, \\ -\gamma & \text{if } k = t, \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

$$\Pr(N_{ij} \geq K_{ij}) \leq \alpha_{ij} \quad \text{for all } (i, j) \in A, \tag{3}$$

$$\lambda_{ij} \geq 0 \quad \text{for all } (i, j) \in A. \tag{4}$$

In the above model, the objective function (1) expresses that the throughput rate γ in the network has to be maximized; constraints (2) are the flow rate conservation equations for the nodes of the network; constraints (3) express that the probability of finding at least K_{ij} units along arc (i, j) is not allowed to exceed α_{ij} for any arc $(i, j) \in A$; and constraints (4) ensure that the flow rate along each arc is non-negative. The aim of constraints (3) is to avoid queues to exceed the regular buffer space K_{ij} for each arc $(i, j) \in A$ as much as possible. Stability of the queueing systems at the arcs of the network is implicitly guaranteed by constraints (3).

4. Minimum cost flow rate problem

The *minimum cost flow rate problem* for a directed single commodity capacitated Markovian open queueing network problem is stated as follows:

$$\text{Minimize } \sum_{(i,j) \in A} c_{ij} \lambda_{ij} \tag{5}$$

subject to

$$\sum_{j \in O_k} \lambda_{kj} - \sum_{i \in I_k} \lambda_{ik} = \begin{cases} \gamma & \text{if } k = s, \\ -\gamma & \text{if } k = t, \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

$$\Pr(N_{ij} \geq K_{ij}) \leq \alpha_{ij} \quad \text{for all } (i, j) \in A, \tag{7}$$

$$\lambda_{ij} \geq 0 \quad \text{for all } (i, j) \in A. \tag{8}$$

In this model, the objective function (5) represents the total processing cost rate which has to be minimized; constraints (6) are the flow rate conservation equations which reflect the requirement that the flow through the network should be of a given size γ. Finally, constraints (7) and (8) have similar interpretations as constraints (3) and (4) in the maximum flow rate problem.

5. The solution algorithm

To solve the above optimization problems the stochastic constraints (3) and (7) will be replaced by equivalent deterministic upperbounds for the flow rates λ_{ij}. More specifically, we will show that constraints (3) and (7) are equivalent to constraints of the form

$$\lambda_{ij} \leq \lambda_{ij}^* \quad \text{for all } (i, j) \in A, \tag{9}$$

in which, for each arc $(i, j) \in A$, λ_{ij}^* is the unique

solution of the following non-linear programming problem:

$$\lambda_{ij}^{*} = \max\{\lambda_{ij} \in (0, S_{ij}\mu_{ij}):$$
$$\Pr(N_{ij} \geqslant K_{ij}) \leqslant \alpha_{ij}\}. \quad (10)$$

For brevity the indices (i, j) will be ignored in the next part of this section. The excess probability $\Pr(N \geqslant K)$ for an M/M/S system, with threshold $K \geqslant S$, can be written as

$$\Pr(N \geqslant K) = \frac{\frac{S^S}{S!}\left(\frac{\lambda}{S\mu}\right)^K}{1 + \sum_{m=1}^{S-1} \frac{1}{m!}\left(\frac{\lambda}{\mu}\right)^m \left(1 - \frac{m}{S}\right)}. \quad (11)$$

In order to prove that constraints (9) are well defined and equivalent to constraints (3) and (7) it is sufficient to show that $\Pr(N \geqslant K)$ is a strictly increasing function of λ with range the interval (0, 1). This property can be proved directly by showing that the derivative of this probability with respect to λ is positive. It also follows from the fact that N is stochastically increasing in λ, cf., e.g., Shaked and Shanthikumar (1988). To reduce the search area for the solution of the non-linear programming problems (10) we note that the denominator in (11) satisfies the following inequalities for $\lambda < S\mu$:

$$1 \leqslant 1 + \sum_{m=1}^{S-1} \frac{1}{m!}\left(\frac{\lambda}{\mu}\right)^m \left(1 - \frac{m}{S}\right)$$
$$\leqslant 1 + \sum_{m=1}^{S-1} \frac{S^m}{m!}\left(1 - \frac{m}{S}\right) = \frac{S^S}{S!}. \quad (12)$$

The foregoing discussion leads to the following results.

Lemma. *The probability* $\Pr(N \geqslant K)$, *with threshold* $K \geqslant S$, *in an* M/M/S *queueing system is a strictly increasing function of the arrival rate* λ *for fixed service rate* μ, *for* $0 < \lambda < S\mu$. *Moreover, this probability satisfies the following inequalities for* $0 < \lambda < S\mu$:

$$\left(\frac{\lambda}{S\mu}\right)^K \leqslant \Pr(N \geqslant K) \leqslant \min\left\{1, \frac{S^S}{S!}\left(\frac{\lambda}{S\mu}\right)^K\right\}. \quad (13)$$

Because $\Pr(N \geqslant K)$ increases from 0 to 1 when λ increases from 0 to $S\mu$, this lemma implies:

Corollary. *There exists for every* α, $0 < \alpha < 1$, *a unique arrival rate* $\lambda = \Lambda(\mu, S, K, \alpha)$ *such that*

$$\Pr(N \geqslant K) = \frac{\frac{S^S}{S!}\left(\frac{\lambda}{S\mu}\right)^K}{1 + \sum_{m=1}^{S-1} \frac{1}{m!}\left(\frac{\lambda}{\mu}\right)^m \left(1 - \frac{m}{S}\right)} = \alpha. \quad (14)$$

This arrival rate $\Lambda(\mu, S, K, \alpha)$ *satisfies the following bounds*:

$$S\mu\sqrt[K]{S!\alpha/S^S} \leqslant \Lambda(\mu, S, K, \alpha) \leqslant S\mu\sqrt[K]{\alpha}. \quad (15)$$

Having solved the non-linear programming problems (10) corresponding to each arc, which means having solved numerically the non-linear equations (14) for each arc of the network, constraint sets (3) and (7) can be replaced by the following constraint set:

$$\lambda_{ij} \leqslant \lambda_{ij}^{*} = \Lambda(\mu_{ij}, S_{ij}, K_{ij}, \alpha_{ij})$$
$$\text{for all } (i, j) \in A. \quad (16)$$

Then, the resulting problem (1), (2), (16), (4) transforms into the classical deterministic, directed, capacitated, single commodity *maximum flow problem*, which can be solved by Ford and Fulkerson's labeling algorithm, while (5), (6), (16), (8) represents the deterministic, directed, capacitated, single commodity *minimum cost flow problem*, which can be solved by the out-of-kilter algorithm. For further discussions, see Murty (1976, Chapter 12).

6. Extensions

Several extensions of the basic queueing model as described in Section 2 are possible without

affecting the results. To mention just a few, consider the situation in which every service station is equipped with an automated inspection unit for detecting defective (but repairable) items. Upon identification of such an item the inspection unit reroutes it back to the associated workstation the item will travel another time along the same arc). This leads to a queueing network with instantaneous Bernoulli feedback along the individual arcs. Let q_{ii} denote the probability that an item has to be reprocessed along arc $(i, j) \in A$. The resulting maximum flow rate problem and minimum cost flow rate problem can be formulated as in Sections 3 and 4 with the only difference that μ_{ij} has to be replaced by $(1 - q_{ij})\mu_{ij}$ in Eq. (14) for each arc $(i, j) \in A$. In fact, we have introduced here a situation with a tandem queueing system at an arc (first a service unit, then an inspection unit). This concept can be generalized to arcs with an arbitrary number of service units arranged as an open Markovian sub-network, with fixed (non-controllable) routing probabilities. Suppose there is a chance constraint of the form (3) for each service unit in such a sub-network. For each service unit we can determine an upperbound on the flow through the sub-network by solving an equation of the form (14) in which the service rate μ should be replaced by μ divided by the average number of visits to the service unit by a job entering the sub-network. For the arc containing such a sub-network we finally obtain a single upperbound for the flow rate on that arc, being the minimum of the upper bounds of all service units at that arc.

Variants of our optimization problems are obtained when the constraint sets (3) and (7) are replaced (or supplemented by) any set of constraints of the form

$$E\{\phi_{ij}(N_{ij})\} \leqslant \alpha_{ij} \quad \text{for all } (i, j) \in A, \tag{17}$$

with functions $\phi_{ij} : \mathbb{N} \to \mathbb{R}$ such that $E\{\phi_{ij}(N_{ij})\}$ are increasing functions of λ. For all such functions the chance or moment constraints can be replaced by deterministic constraints of the form (9). As shown by Shaked and Shanthikumar (1988) functions of the type $E\{\phi(N)\}$, with N the number of jobs in a stable M/M/S system, are increasing in λ for every increasing function $\phi : \mathbb{N} \to \mathbb{R}$.

Another extension of our model is the introduction of losses (or gains) of items at arcs. If each item has a probability p_{ij} of getting lost at arc $(i, j) \in A$ (e.g., because of an irreparable defect), then a product-form solution remains available, while the chance constraint optimization problems can be reduced to deterministic *flows with gains problems*, cf. Gondran and Minoux (1984).

Further generalizations of our problems can be obtained by adding a finite number L of constraints on the total number of jobs in the system of the form

$$E\left\{\psi_k\left(\sum_{(i,j)\in A} N_{ij}\right)\right\} \leqslant \beta_k \quad \text{for } k = 1, 2, \ldots, L, \tag{18}$$

with $\psi_k : \mathbb{N} \to \mathbb{R}$, $k = 1, \ldots, L$, convex and increasing functions, or by replacing the objective function (5) by a separable convex and increasing function of the flow rates λ_{ij} at the arcs $(i, j) \in A$. The resulting optimization problems are less structured, but polynomial time algorithms exist for solving such problems, cf. Hochbaum and Shanthikumar (1990).

Finally, we note that the ideas as presented in this paper can also be used to translate stochastic multi-commodity network flow problems into their deterministic counterparts by considering multi-class queueing networks. However, the translation of a chance constraint on the total number of jobs of all classes present at an arc into a deterministic upper bound on the sum of the flow rates at the arc over all job classes can only be performed if the service rates are class-independent at every arc; otherwise, there exists no product-form solution to the network, cf. Walrand (1988).

References

Buzacott, J.A., and Shanthikumar, J.G. (1993), *Stochastic Models of Manufacturing Systems*, Prentice-Hall, Englewood Cliffs, NJ.

Gondran, M., and Minoux, M. (1984), *Graphs and Algorithms*, Wiley, Chichester.

Hochbaum, D.S., and Shanthikumar, J.G. (1990), "Convex separable optimization is not much harder than linear optimization", *Journal of the Association for Computing Machinery* 37, 843–862.

Kleinrock, L. (1976), *Queueing Systems, Vol. 2: Computer Applications*, Wiley, New York.

Murty, K.G. (1976), *Linear and Combinatorial Programming*, Wiley, New York.

Pourbabai, B. (1990), "A class of queueing optimization problems", *Applied Mathematics Letters* 3, 91–94.

Shaked, M., and Shanthikumar, J.G. (1988), "Stochastic convexity and its applications", *Advances in Applied Probability* 20, 427–446.

Walrand, J. (1988), *An Introduction to Queueing Networks*, Prentice-Hall, Englewood Cliffs, NJ.