On Optimal Service Capacity Allocations for Fork-Join Open Queueing Networks via Second Order Cone Programming

01008610 上智大学 *石塚 陽 ISHIZUKA, Yo
01703040 東北大学 山下英明 YAMASHITA, Hideaki
01605610 電気通信大学 村松正和 MURAMATSU, Masakazu

1. Introduction

For queueing network systems, we consider a problem which finds an optimal service capacity allocation to the servers so as to maximize the throughput of the system.

For a serial queueing network system in which servers are connected in series, some studies have been devoted to solving optimal *mean service time* allocation problems[1, 3]. For general queueing network systems, however, such kind of studies have not been reported.

A natural interpretation of the mean service time allocation for serial queueing system is a decomposition of a job into a series of "pipelined processors." This study assumes a different situation; we consider fork-join type queueing network systems, and formulate an optimal *service capacity* allocation problem. Our problem can be interpreted as an optimal server performance allocation for the servers each of which has to process a certain amount of jobs.

We apply the idea of "sample-path optimization" to the throughput maximization problem, and show the (approximate) optimization problem can be formulated as a second order cone programming problem (SOCP) which can be solved effectively by the interior point methods.

2. Model

We consider an M sever synchronized fork-join open queueing network system. Let us define

- S : Set of all servers ($=\{1,2,\cdots,M\}$);
- $I \subset S$: Set of input servers to which jobs enter;
- $O \subset S$: Set of output servers from which jobs leave the system;
- P_i : Set of the preceding (upstream) servers of server i;
- Q_i : Set of the succeeding (downstream) servers of server i
- $S_{i,j}$: Service time of j-th job at server i;
- $D_{i,j}$: Completion time of j-th departure at server i,

We impose the following assumption on topology of the network under consideration.

(A1) Network is connected without any closed loop, $I \cap O = \emptyset$, input servers accept jobs only from the outside of the system, and that all jobs from output servers leave the system.

Here are our basic assumptions.

(A2)

- There are infinitely many jobs waiting in front of each input servers so that the input servers are never starved;
- At server i, completed job can leave the server if at least one buffer space is available at each downstream server $q \in Q_i$, $i \in S \setminus O$;
- The service at server i can be started only if there is a set of jobs from P_i in its buffer and the server is empty, $i \in S \setminus I$;
- After completion of a service at server i, the next service at the server can be started after all departures to the downstream servers in Q_i are completed, $i \in S \setminus O$;
- At each output server $i \in O$, the departures of completed jobs are never blocked.

3. Formulation

We suppose that, at server i, the service times $S_{i,j}, j = 1, 2\cdots$ are i.i.d. with mean $1/\mu_i$, and that we can control values of these means. Let $TH_i(\boldsymbol{\mu})$ be the throughput from server i. Given total service capacity C, our problem then is to find an optimal allocation $\boldsymbol{\mu} = (\mu_1 \cdots \mu_M)^T$ which maximizes the throughput TH_r for arbitrarily chosen $r \in S$:

$$P(C)\begin{cases} \max\limits_{\mu} \; TH_r(\boldsymbol{\mu}) \\ \text{subj. to } \sum\limits_{i=1}^{M} \mu_i \le C \\ \qquad\qquad \boldsymbol{\mu} \ge \mathbf{0} \end{cases}$$

However, since it is difficult to obtain exact values of $TH_r(\boldsymbol{\mu})$, adopting idea of the so-called "sample-path optimization[2]", we approximate them by a simulation run under fixed sample (random numbers) as follows. Let $\boldsymbol{\omega}$ be a sample (a series of random numbers) in a sample space, and a method for generating a sample-path (realized values) $\overline{S}_{i,j}(1/\mu_i)$ with any fixed mean value $1/\mu_i$ of $S_{i,j}$ from $\boldsymbol{\omega}$ be given. Then, with the sample $\boldsymbol{\omega}$ being fixed, departure times $\overline{D}_{i,j}(\boldsymbol{\mu})$ under service capacity allocation $\boldsymbol{\mu}$ can be calculated by

$$\overline{D}_{i,j}(\boldsymbol{\mu}) = \max\Big\{ \max_{p\in P_i} \overline{D}_{p,j}(\boldsymbol{\mu}) + \overline{S}_{i,j}(1/\mu_i), \quad \overline{D}_{i,j-1}(\boldsymbol{\mu}) + \overline{S}_{i,j}(1/\mu_i), \ \max_{q\in Q_i} \overline{D}_{q,j-B_q}(\boldsymbol{\mu}) \Big\} \quad (1)$$

We thus can approximate $TH_i(\boldsymbol{\mu})$ by the value $TH_{r,N}(\boldsymbol{\mu}) = N/D_{r,N}(\boldsymbol{\mu})$ for large N, and obtain the following approximate optimization problem.

$$P_N(C) \left\{ \begin{array}{l} \max\limits_{\theta} \dfrac{N}{\overline{D}_{r,N}(\boldsymbol{\theta})} \\ \text{subj. to } \displaystyle\sum_{i=1}^{M} \frac{1}{\theta_i} \le C \\ \qquad\qquad \boldsymbol{\theta} > \mathbf{0} \end{array} \right.$$

where $\boldsymbol{\theta} = (\theta_1 \cdots \theta_M)^T = (1/\mu_1 \cdots 1/\mu_N)^T$. We assume that

(A3) $\overline{S}_{i,j}(\theta_i) = \overline{S}_{i,j}(1/\mu_i)$ is linear in θ_i.

Under this assumption the objective function of $P_N(C)$ is a concave function, and hence, the optimal solution of $P_N(C)$ converges to a true optimal solution of $P(C)$[2].

4. Conversion to SOCP

Let us define a network $\mathcal{N}(\boldsymbol{\theta})$ consisting of $(MN+1)$ nodes $\{(0)\} \cup \{(i,j) \mid j = 1,2,\cdots,N, i \in S\}$ such that node (i,j) corresponds to departure time $\overline{D}_{i,j}(\boldsymbol{\theta})$ and the weights of arcs are given as follows:

Arc	Weight
$(0) \to (i,1),\ i \in I$	$\overline{S}_{i,1}(\theta_i)$
$(p,j) \to (i,j),\ p \in P_i, i \in S$	$\overline{S}_{i,j}(\theta_i)$
$(i,j-1) \to (i,j),\ i \in S$	$\overline{S}_{i,j}(\theta_i)$
$(q,j-B_q) \to (i,j),\ q \in Q_i, i \in S$	0

where (0) is a dummy node. Then, it is clear that $\overline{D}_{i,N}(\boldsymbol{\theta})$ coincides with the length of the longest path from node (0) to node (i,N). Thus, defining

$$\mathcal{P}_{i,N} = \text{Set of paths from (0) to } (i,N) \text{ in } \mathcal{N}(\boldsymbol{\theta})$$
$$d_i(P,\boldsymbol{\theta}) = \text{Length of } P \in \mathcal{P}_{i,N}$$

we can rewrite problem $P_N(C)$ as:

$$\left\{ \begin{array}{l} \min\limits_{\theta,\sigma} \ \sigma \\ \text{subj. to } d_r(P,\boldsymbol{\theta}) - \sigma \le 0 \ \forall P \in \mathcal{P}_{r,N} \\ \qquad\qquad \boldsymbol{\theta} \in \Theta(C) \end{array} \right.$$

where $\Theta(C) = \{\boldsymbol{\theta} \mid \sum_{i=1}^{M} 1/\theta_i \le C,\ \boldsymbol{\theta} > \mathbf{0}\}$. It should be noted that the first constraint condition consists of many linear inequalities. Second constraint can be converted into a set of second order cone constraints as follows. Introducing new variables η_i and $\boldsymbol{\xi}_i = (\xi_{i0}\, \xi_{i1}\, \xi_{i2})^T$, we have

$$\boldsymbol{\theta} \in \Theta(C)$$
$$\Leftrightarrow \left\{ \begin{array}{l} \sum_{i=1}^{M} \eta_i \le C, \\ \boldsymbol{\xi}_i \in K(3),\ \xi_{i0} = \dfrac{\eta_i + \theta_i}{2},\ \xi_{i1} = \dfrac{\eta_i - \theta_i}{2}, \\ \xi_{i2} = 1,\ \theta_i \ge 0,\ \eta_i \ge 0 \ \ \forall i \in S \end{array} \right.$$

where $K(3)$ is the 3-dimensional second order cone. We thus have shown that the problem $P_N(C)$ is equivalent to the following second order cone programming problem (SOCP).

$$SOCP_N(C) \left\{ \begin{array}{l} \min\limits_{\boldsymbol{\theta},\sigma,\boldsymbol{\xi}} \ \sigma \\ \text{subj. to} \\ \quad d_r(P,\boldsymbol{\theta}) - \sigma \le 0 \ \forall P \in \mathcal{P}_{r,N} \\ \quad \sum_{i=1}^{M} \eta_i \le C \\ \quad \theta_i \ge 0,\ \eta_i \ge 0 \ \forall i \in S \\ \quad \xi_{i0} = \dfrac{\eta_i + \theta_i}{2} \ \ \forall i \in S \\ \quad \xi_{i1} = \dfrac{\eta_i - \theta_i}{2} \ \ \forall i \in S \\ \quad \xi_{i2} = 1 \ \ \forall i \in S \\ \quad \boldsymbol{\xi}_i \in K(3) \ \ \forall i \in S \end{array} \right.$$

For a fixed path $P \in \mathcal{P}_{r,N}$, $d_r(P,\boldsymbol{\theta})$ is a linear function in $\boldsymbol{\theta}$, and hence, $SOCP_N(C)$ is an SOCP with a huge number of linear constraints. Since it is hard to deal with all of these constraints simultaneously, we take an approach of "relaxation method", and the sub-problems (relaxed problems) can be solved by interior point methods[4] effectively.

5. Conclusion

Concrete examples with Coxian service time distributions and numerical results will be shown at the presentation.

References

[1] Hillier F. S. and Boling R. W.: "On the optimal allocation of work in symmetrically unbalanced production line systems with variable operation times," Management Science, Vol.25, pp.721-728, (1979).

[2] Robinson S. M.: "Analysis of sample-path optimization," Mathematics of Operations Research, Vol.21, No.3, (1996).

[3] Plambeck E. L., Fu B.-R. Robinson S. M. and Suri R.: "Sample-path optimization of convex stochastic performance functions," Mathematical Programming, 75, pp.137-176, (1996).

[4] Tsuchiya, T.: "A Polynomial Primal-dual Path-following Algorithm for Second-order Cone Programming," *Research Memorandum No. 649*, The Institute of Statistical Mathematics, Tokyo, Japan (1997)