Anais do Congresso de Matemática Aplicada e Computacional
CMAC Centro-Oeste 2013
ISSN 2318-471X
ISBN: 978-85-8215-047-4

Multiobjective Optimization of Finite Queueing Networks

F. R. B. Cruz
Departamento de Estatística, Universidade Federal de Minas Gerais,
31270-901, Belo Horizonte, MG
E-mail: fcruz@est.ufmg.br

A. R. Duarte
Departamento de Matemática, Universidade Federal de Ouro Preto,
35400-000, Ouro Preto, MG
E-mail: anderson@iceb.ufop.br

N. L. C. Brito
Departamento de Ciências Exatas, Universidade Estadual de Montes Claros,
39401-089, Montes Claros, MG
E-mail: nilson.brito@unimontes.br

Abstract: *We aim at studying a multi-objective algorithm to simultaneously optimize the total number of buffers, the overall service rate, and the throughput of a general-service finite queueing network. These conflicting objectives are optimized by means of a multi-objective genetic algorithm, designed to produce solutions for more than one objective. Computational experiments are shown, in order to determine the efficacy and efficiency of the approach. Instigating news insights are given.*
Keywords: *Queues, Networks, Performance evaluation, Optimization*

1 Introduction

Our focus here is on single-server queueing networks with exponentially distributed inter-arrival times and generally distributed service times, configured in an arbitrary acyclic topology (see Fig. 1). More specifically, the focus is on networks of $M/G/1/K$ queues, which in Kendall notation stands for **M**arkovian arrivals, **G**enerally distributed service times, a single server, and the total capacity of K items, *including* the item in service.

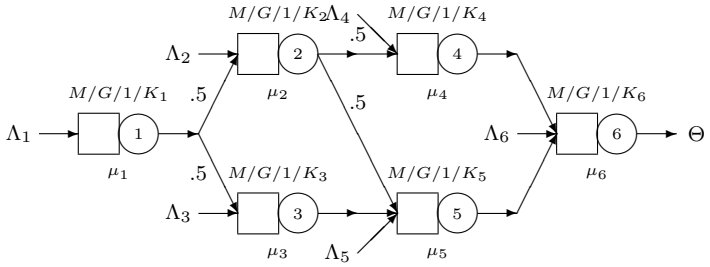

Figura 1: An $M/G/1/K$ queueing network

Given the topology and the external arrival rates ($\mathbf{\Lambda} = \{\Lambda_1, \Lambda_2, \ldots, \Lambda_n\}$), our goal is to obtain the maximum throughput (Θ) by means of the minimum number of buffers ($\mathbf{K} = \{K_1, K_2, \ldots, K_n\}$) and the minimum service rates ($\boldsymbol{\mu} = \{\mu_1, \mu_2, \ldots, \mu_n\}$). Potential users of these queueing models include computer scientists and engineers. Indeed, these models may help to understand and to improve various real-life systems, including manufacturing, production and health systems, urban or pedestrian traffic, computer and communication systems, and

ISSN 2318-471X
ISBN: 978-85-8215-047-4

web-based applications. There is a trade-off between the overall number of buffers, the service rates, and the resulting throughput. Because buffers and services can be very expensive, the overall buffer and service capacity should not be large. On the other hand, the highest possible network throughput should be reached. Unfortunately, the throughput is directly affected by the number of buffers allocated and the service rates. Indeed, if the buffer and service capacity reduces there will be in general an undesirable reduction in the throughput in a network of queues. Previous results show that such a surface is smooth and probably convex but since the top of the surface is flat, traditional optimization algorithms may have trouble converging. Indeed, results are reported for a successful optimization algorithm coupled with multiple starts to avoid premature convergence to local optima.

In this paper, an optimization approach to simultaneously optimize the total number of buffers, the overall service rate, and the throughput of networks of $M/G/1/K$ queues is discussed. The algorithm produces a set of efficient solutions for more than one objective in the objective function [1]. With the algorithm, the decision maker is able to evaluate the effect of solution replacement. Moreover, the multi-objective approach also allows the user to increase one objective (*e.g.*, throughput) while simultaneously reducing another objective (*e.g.*, buffer and service rate allocation).

The organization is as follows. A multi-objective evolutionary algorithm specifically developed to multi-objective optimization is presented in Sec. 2, along with the GEM, a performance evaluation tool used to approximate the throughput. In Sec. 3, the results of a comprehensive set of computational experiments are presented to show the efficiency of the approach. Finally, Sec. 4 concludes this paper with final remarks and suggestions for future research in the area.

2 Algorithms

The throughput maximization problem can be defined in a digraph $G(N, A)$, where N is a finite set of nodes (queues) and A is a finite set of arcs (pair of connected queues) by the following mixed-integer mathematical programming formulation:

$$\text{minimize } F(\mathbf{K}, \boldsymbol{\mu}), \tag{1}$$

subject to

$$K_i \in \{1, 2, \ldots\}, \ \forall i \in N, \tag{2}$$

$$\mu_i \geq 0, \ \forall i \in N, \tag{3}$$

where the decision variables K_i and μ_i indicate the total capacity of the service and the service rate for the ith $M/G/1/K$ queue, respectively. The objective functions, $F(\mathbf{K}, \boldsymbol{\mu}) \equiv \Big(f_1(\mathbf{K}), f_2(\boldsymbol{\mu}), -f_3(\mathbf{K}, \boldsymbol{\mu})\Big)$, are the total buffer allocation, $f_1(\mathbf{K}) = \sum_{\forall i \in N} K_i$, the overall service allocation, $f_2(\boldsymbol{\mu}) = \sum_{\forall i \in N} \mu_i$, and the overall throughput, $f_3(\mathbf{K}, \boldsymbol{\mu}) = \Theta(\mathbf{K}, \boldsymbol{\mu})$. Formulation (1)–(3) has been used before with success [2]. However, notice that in the literature the throughput is commonly modeled as a constraint that must be greater than a threshold Θ_τ value rather than as an objective that must be maximized. The problem is that to solve this single-objective version of the problem the throughput constraint must be relaxed and to establish an appropriate Θ_τ is not a trivial task.

When the interest is on single queues, the throughput $\Theta(\mathbf{K}, \boldsymbol{\mu})$ is:

$$\Theta(\mathbf{K}, \boldsymbol{\mu}) = \lambda(1 - p_K), \tag{4}$$

where λ is the external arrival rate and p_K is the blocking probability, which is the probability that an item finds the system full (that is, the number of items in the systems is equal to the total capacity K). Thus, the problem of finding $\Theta(\mathbf{K}, \boldsymbol{\mu})$ reduces to determining p_K. In particular, it has been shown in a previous paper that a two-moment approximation based on the Markovian expression is quite effective [2]. For networks of queues, the estimation of the throughput is

ISSN 2318-471X
ISBN: 978-85-8215-047-4

made by means of the generalized expansion method (GEM), which is an algorithm that has been successfully used to estimate the performance of arbitrarily configured, finite queueing, acyclic networks [6].

For the problem under consideration, multiobjetive evolutionary algorithms (MOEAs) seem to be a suitable choice since they are optimization algorithms that perform an approximate global search based on information obtained from the evaluation of several points in the search space. The population of points that converge to an optimal value are obtained through the application of the genetic operators, *mutation*, *crossover*, *selection*, and *elitism*. Each one of these operators characterizes an instance of a MOEA and can be implemented in several different ways. The instance of MOEA used in this study is based upon the elitist non-dominated sorting genetic (NSGA-II) algorithm of Deb et al. [4]. Elitism is based on the concept of dominance. Point $\mathbf{x}_i = (x_{i_1}, x_{i_2}, \ldots, x_{i_n})$ dominates point $\mathbf{x}_j = (x_{j_1}, x_{j_2}, \ldots, x_{j_n})$ if $\mathbf{x}_i$ is superior to $\mathbf{x}_j$ in one objective ($f_k(\mathbf{x}_i) < f_k(\mathbf{x}_j)$, for minimization) and is not inferior in any other objective ($f_\ell(\mathbf{x}_i) \not> f_\ell(\mathbf{x}_j)$, for minimization). To perform elitism, the fast non-dominated sorting algorithm is employed [4]. By sequentially choosing points from each non-dominated front $(\mathcal{F}_1, \mathcal{F}_2, \ldots)$, selection is performed until the number of required individuals for the next iteration is obtained. Some decision must be made if the maximum number of individuals is exceeded after the addition of a group of individuals from front $\mathcal{F}_i$. One possibility is to compute a measure of diversity, such as the crowding distance defined by [4], to ensure the highest diverse population. For the problem at hand, the *uniform crossover* mechanism was selected, which is popular in multivariable encodings due to its efficiency in identifying, inheriting, and protecting common genes, as well as re-combining non-common genes [5]. In this mechanism, crossover is performed for each variable with a probability (`rateCro`), in accordance with the crossover operator. The crossover operator used in the algorithm is the *simulated binary crossover operator* (SBX) [3]. SBX is quite convenient for real-coded GAs because of its ability to simulate *binary crossover operators* avoiding re-encoding the variables. For each individual gene, the mutation scheme occurs with a specific probability (`rateMut`) and as suggested by Deb and Agrawal [3] Gaussian perturbations were added to the decision variables, $K_i + \varepsilon_i$ and $\mu_i + \varepsilon_{N+i}$, for all $i \in N$, with $\varepsilon_i \sim \mathcal{N}(0, 1)$, $i \in \{1, 2, \ldots, 2N\}$. Finally, to ensure feasibility of constraints (2) and (3) after crossover and mutation, the integer variables values must be readjusted by applying reflection operators

$$K_{\text{rfl}_i} = K_{\text{lowlim}} + |K_i - K_{\text{lowlim}}|, \text{ and } \mu_{\text{rfl}_i} = \mu_{\text{lowlim}_i} + |\mu_i - \mu_{\text{lowlim}_i}|, \tag{5}$$

where K_{lowlim} is the lower limit of buffer allocation (*i.e.*, $K_{\text{lowlim}} = 1$) and μ_{lowlim_i} is the lower limit of service allocation (to ensure that $\rho < 1$ holds).

3 Results and Discussion

As indicated by previous studies on GAs, a set of parameters to ensure rapid convergence with a minimal amount of computational effort may be determined without major trouble by trial and error. As a word concerning the best group of parameters for the algorithm, one could use the following combination: (i) combined use of SBX and mutation, with (ii) a mutation rate below 2%, (iii) although greater the better the population, 400 individuals seem to be enough, and (iv) the dispersion parameter, η, should not go above 8. To ensure a finite computation time, a maximum number of generations `numGen` was set to 4,000. Fortunately, MOEAs are robust enough to perform well in a broad range of problems, as confirmed by the experiments run.

The network presented in Fig. 1 was analyzed with the proposed method. Two different squared coefficients of variation were analyzed, $cv^2 = 0.5$ and 1.5, with arrival rate ($\Lambda_1 = 1.0$). Table 1 presents some Pareto efficient solutions for a more detailed analysis. Note that, with this multiobjective methodology, it is possible to identify points from which there is no more interest in increasing the spend on buffer sizes or service rates because the gain in the throughput will be rather narrow. For example, for a $cv^2 = 0.5$, keeping the overall service rate approximately

ISSN 2318-471X
ISBN: 978-85-8215-047-4

constant one had to increase the overall buffer size by 22% to produce a gain of only 0.01% in the throughput. Similarly, there may be a similar point for the overall service rate. In fact, it can be seen that for an increase of 12% in the overall service rate, an increase of only 0.01% is produced in the throughput, which may be considered negligible. Note also that with $cv^2 = 1.5$ such a phenomenon can occur even more pronounced. It is observed that it may be necessary to increase by 36% the overall buffer size to reach an increase of only 0.6% on the throughput. It is therefore more advantageous to maintain a system with an allocation that produces on output of 99.99% of the input (that is, 0.9999/1.000) than spending 70% more in service rate to raise the output by only 0.01% (ie, raising it to 100% of the arrival rate). These are just some examples of the analyzes that can be done in finite general service queueing networks via the multiobjective methodology.

Tabela 1: Pareto efficient solutions selected from the computational experiments

cv^2	$\sum_i K_i$	$\Delta\%$	$\sum_i \mu_i$	$\Delta\%$	Θ	$\Delta\%$
0.5	18	-	51.0	-	0.9999	-
	22	22%	51.4	0.8%	1.0000	0.01%
	20	-	46.6	-	0.9999	-
	20	0%	52.1	12%	1.0000	0.01%
1.5	14	-	60.4	-	0.9944	-
	19	36%	61.1	1.1%	0.9999	0.6%
	19	-	61.1	-	0.9999	-
	19	0%	104.0	70%	1.0000	0.01%

4 Conclusions

In order to optimize the throughput, the buffer sizes, and the service rates of single server, general-service queueing networks, a multi-objective approach was presented. The generalized expansion method (GEM) was coupled with a multi-objective genetic algorithm (MOGA) to make it possible to derive insightful Pareto curves displaying the trade-off between throughput and the allocation of buffers and service rates. Topics for future investigation in this area include extensions to networks of multi-server queues and networks of general-arrival queues. Also interesting is to consider different performance measures, such as the WIP, sojourn time, and so on. These are only few examples of possible topics for research.

5 Acknowledgments

This research is supported by the Brazilian agencies, CNPq, CAPES, and FAPEMIG.

Referências

[1] Chankong, V., Haimes, Y. Y., 1983. Multiobjective Decision Making: Theory and Methodology. Elsevier, Amsterdam, The Netherlands.

[2] Cruz, F. R. B., Kendall, G., While, L., Duarte, A. R., Brito, N. L. C., 2012. Throughput maximization of queueing networks with simultaneous minimization of service rates and buffers. Mathematical Problems in Engineering 2012 (Article ID 348262), 19 pages.

[3] Deb, K., Agrawal, R. B., 1995. Simulated binary crossover for continuous search space. Complex Systems 9, 115–148.

[4] Deb, K., Pratap, A., Agarwal, S., Meyarivan, T., 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation 6, 182–197.

[5] Hu, X.-B., Di Paolo, E., 2007. An efficient genetic algorithm with uniform crossover for the multi-objective airport gate assignment problem. In: IEEE Congress on Evolutionary Computation, CEC 2007. Singapore, pp. 55–62.

[6] Kerbache, L., Smith, J. M., 1987. The generalized expansion method for open finite queueing networks. European Journal of Operational Research 32, 448–461.