Novel Solutions for Closed Queueing Networks with Load-Dependent Stations*

Giuliano Casale, Peter G. Harrison, Ong Wai Hong
Department of Computing
Imperial College London, UK
{g.casale, p.harrison, wai.ong11}@imperial.ac.uk

1. INTRODUCTION

Load-dependent closed queueing networks are notoriously difficult to approximate since their analysis requires to consider state-dependent service demands. Commonly employed approximations for load-independent models, such as mean-value analysis, are not equally efficient in the load-dependent setting, since mean queue-lengths are no longer sufficient to determine the model performance.

For single-class load-dependent models, we provide the first explicit formula for the normalizing constant that applies to models with *arbitrary* load-dependent rates, while retaining $\mathcal{O}(1)$ complexity with respect to the total population size. From this result, we also derive corresponding closed-form solutions for the multiclass case. This further yields two novel integral forms for the normalizing constant of state probabilities, in the real and complex domains. The paper also illustrate practical improvements in computational properties resulting from these developments.

2. REFERENCE MODEL

We focus on closed queueing network models that satisfy the assumptions for a product-form solution [4, 15], such as closed networks including $-/GI/1$ processor sharing queues, $-/M/k$ first-come first-served queues with identical service rates, and $-/GI/\infty$ delay nodes. Service time distributions are assumed to have a rational Laplace-Stieltjes transform.

The queueing network model under study is assumed to have M single-server queues and R job classes. Indexes k, i are used to denote queues ($k, i = 1, \ldots, M$), while indexes r, s are used to indicate classes ($r, s = 1, \ldots, R$). Each class is populated by N_r jobs. Thus, the total number of jobs in the network is $N = N_1 + \ldots + N_R$.

For queue k, we denote by $\theta_{k,r}$ the mean service demand of class r, which is the ratio of the mean number of visits to the mean service rate of class r jobs at queue k. Without loss of generality, we assume that the first $M' \le M$ queues are distinct and denote by M_k, $1 \le k \le M'$, the multiplicity of queue k, i.e., the number of queues out of the M in the network having demands $\theta_{k,r}$, $1 \le r \le R$.

The state space of the Markov process underlying the network is $\mathcal{S}(\boldsymbol{N}) = \{\boldsymbol{n} = (n_{1,1}, \ldots, n_{M,R}) \mid n_{k,r} \ge 0, \sum_{k=1}^{M} n_{k,r} = N_r\}$, in which $n_{k,r}$ denotes the number of class-r jobs residing at queue k. From basic product-form theory, it is known that the equilibrium distribution of the network is

$$\pi(\boldsymbol{n}) = \frac{1}{H(\boldsymbol{N})} \prod_{k=1}^{M} \frac{n_k!}{\alpha_k(n_k)} \prod_{r=1}^{R} \frac{\theta_{k,r}^{n_{k,r}}}{n_{k,r}!} \qquad \boldsymbol{n} \in \mathcal{S}(\boldsymbol{N}) \quad (1)$$

where $\boldsymbol{N} = (N_1, \ldots, N_R)$, $N = \sum_{r=1}^{R} N_r$, is the population vector, $n_k = \sum_{r=1}^{R} n_{k,r}$ is the total number of jobs at queue k in state $\boldsymbol{n}$, and $\alpha_k(n_k)$ are load-dependent scaling factors, subject to $\alpha_k(0) = 1$. The normalizing constant in (1) ensures that state probabilities sum to unity.

Arbitrary load-dependent factors $\alpha_k(n_k)$ may be considered in our reference model. Some common assignments are the following: (i) Single-server node: $\alpha_k(n_k) = 1$; (ii) Station with s_k homogeneous servers: $\alpha_k(n_k) = \min(s_k, n_k)$; (iii) Delay node: $\alpha_k(n_k) = n_k$.

The rest of this paper is organized as follows. Section 3 introduces our exact results for load-dependent models. Integral forms stemming from these developments are obtained in Section 4. Section 5 gives conclusions.

3. EXACT SOLUTIONS

3.1 Single-class normalizing constant

We begin by giving an explicit solution for the normalizing constant of state probabilities in closed queueing networks with load-dependent stations.

THEOREM 3.1. *Let s_i, $1 \le s_i \le N$, be the smallest index for which there exist a constant c_i such that $\alpha_i(n) = c_i$, $\forall n \ge s_i$. The normalizing constant of a single-class load-dependent closed queueing networks may then be written as*

$$h(N) = \sum_{\boldsymbol{0} \le \boldsymbol{v} < \boldsymbol{s}} g(N - v) \prod_{i=1}^{M} \phi_i(v_i) \quad (2)$$

where

$$\phi_i(v_i) = \begin{cases} \dfrac{\theta_i^{v_i}}{\prod_{j=1}^{v_i} \alpha_i(j)} \left(1 - \dfrac{\alpha_i(v_i)}{\alpha_i(s_i)}\right) & \text{if } v_i > 0 \\ 1 & \text{otherwise} \end{cases}$$

and $\boldsymbol{s} = (s_1, \ldots, s_M)$ and $g(N - v)$ is the single-class normalizing constant of a load-independent model with demands $\sigma_i = \theta_i / \alpha_i(s_i)$ and a population of $N - v$ jobs.

Note that the above result is explicit since a closed-form expression for the load-independent normalizing constants $g_{\boldsymbol{n}}(N - v)$ has been derived in [6, Eq.(3.12)] and can be conputed in $\mathcal{O}(1)$ time and space under a growth of the population N. As a result of this, to the best of our knowledge, equation (2) provides for the first time an explicit solution for load-dependent single-class closed queueing networks that is $\mathcal{O}(1)$ with respect to the population size.

*This work has been partially funded by the European Commission grant H2020-825040 (RADON).

3.2 Multiclass normalizing constant

We now derive a result showing that the normalizing constant $H(\boldsymbol{N})$ of a multiclass load-dependent model with class-independent scaling factors $\alpha_k(n_k)$ may be rewritten as a weighted sum of normalizing constant of single class models. In the theorem below, we use the definition of n-th order finite difference [9], i.e., $\Delta_k^n f_k = \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} f_k$.

THEOREM 3.2. *For a normalizing constant of a multiclass closed queueing network with load-dependent stations*

$$H(\boldsymbol{N}) = \frac{\Delta_{\boldsymbol{n}}^{\boldsymbol{N}} h_{\boldsymbol{n}}(N)}{N_1! \cdots N_R!} = \sum_{\boldsymbol{0} \leq \boldsymbol{n} \leq \boldsymbol{N}} \frac{(-1)^{N-n}}{N_1! \cdots N_R!} \prod_{r=1}^{R} \binom{N_r}{n_r} h_{\boldsymbol{n}}(N) \tag{3}$$

where $\boldsymbol{n} = (n_1, \ldots, n_R)$, $n = \sum_{r=1}^{R} n_r$, and in which $h_{\boldsymbol{n}}(N)$ is the normalizing constant of a single-class model with M load-dependent queueing stations, population $N = \sum_{r=1}^{R} N_r$, demands $\theta_{k,\boldsymbol{n}} = \sum_{r=1}^{R} n_r \theta_{k,r}$ and scaling factors $\alpha_k(n_k)$ as in the multiclass model.

From (3) we readily see that an explicit form can be obtained for $H(\boldsymbol{N})$ provided that we replace $h_{\boldsymbol{n}}(N)$ by (2). The computational complexity of (3), instantiated with (2), is $\mathcal{O}(\prod_{i=1}^{M} s_i N^R)$ time and $\mathcal{O}(1)$ space. This form offers computational advantages over the load-dependent mean value analysis (MVA-LD), which is the standard method to assess queueing networks with multi-server stations, since MVA-LD complexity is $\mathcal{O}(N_{max} N^R)$ time and $\mathcal{O}(N_{max}\sqrt{N^R})$ space [7], where $N_{max} = \max_r N_r$. Note in particular that space complexity is reduced to $\mathcal{O}(1)$ with (3).

4. INTEGRAL FORMS

In this section, we develop a integral form for $H(N)$ over the complex and real domains. The first form is based on the Norlünd-Rice integral and computes the normalizing constant $H(N)$ using R contour integrals in the complex domain. Conversely, the second integral form is on the M-dimensional unit simplex. The two integral forms have complementary computational properties, with the first form remaining efficient for large M, while the second form remaining efficient for large R.

4.1 Norlünd-Rice integral

A well-known property of divided differences is that they may be equivalently computed using the Nörlund–Rice integral [9]

$$\Delta_k^n f(k) = \frac{n!}{2\pi i} \oint \frac{f(z)}{z(z-1)(z-2)\cdots(z-n)} dz$$

where i denotes the imaginary unit and the contour of integration encircles the poles at $0, 1, \ldots, n$.

Since (3) computes $H(\boldsymbol{N})$ using R independent finite difference operators and the integrand is a normalizing constant, which being a multivariate polynomial in the demands does not have poles, we come to the integral form

$$H(\boldsymbol{N}) = \frac{1}{(2\pi i)^R} \oint \frac{h_{\boldsymbol{z}}(N)}{\prod_{r=1}^{R} z_r(z_r-1)(z_r-2)\cdots(z_r-N_r)} d\boldsymbol{z} \tag{4}$$

where $\boldsymbol{z} = (z_1, \ldots, z_R)$. This expression can be evaluated numerically for $H(\boldsymbol{N})$ by ensuring that integration circle for each variable z_r encircles the poles $0, \ldots, N_r$.

4.1.1 Example

For a model with R classes we can carry out contour integration by evaluating $h_{\boldsymbol{z}}(N)$ at k equi-spaced points around the R circles that surround the poles. For example, if we consider a model with $R = 2$ classes, where we can denote $h(z_1, z_2) = h_{\boldsymbol{z}}(N)$, with $\boldsymbol{z} = (z_1, z_2)$, the circle integral may be rewritten as the double integral

$$H(\boldsymbol{N}) = \frac{1}{(2\pi i)^2} \int_0^{2\pi} \int_0^{2\pi} h(\gamma(t_1), \gamma(t_2)) \times \frac{\gamma'(t_1)}{\gamma(t_1)^{N_1+1}} \frac{\gamma'(t_2)}{\gamma(t_2)^{N_2+1}} dt_1 dt_2$$

where $\gamma(t) = \cos(t) + i\sin(t)$ and $\gamma'(t) = -\sin(t) + i\cos(t)$. We have verified that for smaller integration steps this formula converges to the exact value computed using the convolution algorithm.

4.2 Simplex integral

Note first that the single-class normalizing constant $g(N)$ can be written as the divided difference [8]

$$g(N) = [\theta_1, \ldots, \theta_M] x^{N+M-1} \tag{5}$$

Because divided differences admit an integral form over the unit simplex via the Hermite-Genocchi formula, the last expression may also be rewritten as

$$g(N) = \frac{(N+M-1)!}{N!} \int_{T_M} (\theta_1 u_1 + \ldots + \theta_M u_M)^N d\boldsymbol{u} \tag{6}$$

where $T_M = \{(u_1, \ldots, u_M) : u_1 + \ldots + u_M = 1, u_i \geq 0\}$.

If we let $\boldsymbol{s} = (s_1, ..., s_K)$ specify the number of servers at each node, we can give a new integral form for $H(\boldsymbol{N})$.

THEOREM 4.1. *In a network with multiclass load-dependent queues the normalizing constant can be written as*

$$H(\boldsymbol{N}) = \frac{1}{\prod_{r=1}^{R} N_r!} \int_{T_K} \sum_{\boldsymbol{0} \leq \boldsymbol{v} < \boldsymbol{s}} a_{\boldsymbol{v}} \boldsymbol{\Delta}_{t_0}^{N-v} \boldsymbol{\Delta}_{\boldsymbol{t}}^{\boldsymbol{v}} f(t_0, \boldsymbol{t}, \boldsymbol{u})\, d\boldsymbol{u} \tag{7}$$

where $\sigma_{kr} = \theta_{kr}/\alpha_k(s_k)$, $\boldsymbol{v} = (v_1, \ldots, v_K)$, $v = \sum_{i=1}^{K} v_i$, $\boldsymbol{t} = (t_1, \ldots, t_K)$, $a_{\boldsymbol{v}} = \frac{(N+K-1-v)!}{(N-v)!} \prod_{i=1}^{K} \phi_i(v_i)$, and we define

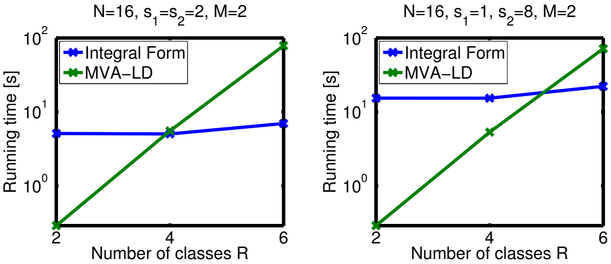
$$f(t_0, \boldsymbol{t}, \boldsymbol{u}) = \prod_{r=1}^{R} \left(\sum_{k=1}^{K} \sigma_{k,r}(t_k + t_0 u_k) \right)^{N_r}$$

4.2.1 Example

In this example, we consider a set of models with $M = 2$ queues and $R \in \{2, 4, 6\}$ classes. The demands are set to $\theta_{k,r} = k \cdot r$. Two scenarios are explored: two balance multi-server stations ($s_1 = s_2 = 2$) and two multi-server stations with unbalanced number of servers ($s_1 = 1, s_2 = 8$).

We use Grundmann-Möller (GM) cubature rules as a way to exactly integrate over the unit simplex [8]. The total number of jobs is $N = 16$ and we set $N_r = \lceil N/R \rceil$. The integral form (7) is evaluated with an exact GM rule of degree N.

The results in Figure 1 illustrate the much greater scalability of the integral form (7) compared to the standard MVA-LD algorithm as the number of classes increase. For larger population MVA-LD quickly experience memory bottlenecks, since memory usage is quadratic in the population size, whereas this is $\mathcal{O}(1)$ in the integral form, which is thus asymptotically more efficient than MVA-LD in space complexity.



(a) Balanced servers (b) Unbalanced servers

Figure 1: Example - GM cubature rule

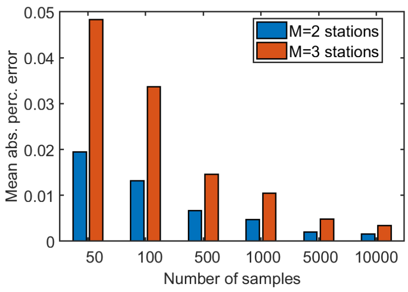


Figure 2: Results - logistic sampling

4.2.2 *Logistic sampling*

The unit simplex integral in (7) can be estimated numerically by adapting the logistic sampling in [8] to the integrand in (7). An open source implementation has been made available in the JMVA tool, part of the Java Modelling Tools suite [5]. The mode of the integrand is obtained, after an additive logistic transformation, by running a conjugate gradient search, implemented using multiprecision arithmetic to avoid numerical issues associated with finite differences and normalizing constants.

The effectiveness of logistic sampling has been tested against random instances with the following characteristics: number of classes $R \in [1,2]$; number of stations $M \in [2,3]$; class populations $N_r \in [1,5]$; number of servers in station 2 $s_i \in [1,2]$; random service demands $\theta_{kr} \in [0,1]$. The average errors are shown in figure 2, showing a rapid reduction of errors as the number of samples grows.

5. CONCLUSION

ZENODO COORDINATES This paper has used finite differences to obtain the first explicit solutions for multiclass load-dependent closed queueing networks. The novelty of this result is that traditional analysis methods are unable to obtain tractable explicit solutions. From a computational standpoint, our most achievement is the mapping of the normalizing constant for load-dependent closed queueing networks to an integral representation over the unit simplex. Integrals over the unit simplex enjoy a wide range of numerical methods for their evaluation [8], which can now also be exploited in performance analysis of load-dependent models.

6. REFERENCES

[1] J. Aitchison. *The Statistical Analysis of Compositional Data*. J. Royal Stat. Soc. Series B, 44(2), 139-177, 1982.

[2] K. E. Atkinson *An Introduction to Numerical Analysis*. John Wiley & Sons, 2nd ed., 1989.

[3] V. Baldoni, N. Berline, J. A. D. Loera, M. Köppe, and M. Vergne. How to integrate a polynomial over a simplex. *Mathematics of Computation*, 80:297–325, 2011.

[4] F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios. Open, closed, and mixed networks of queues with different classes of customers. *JACM*, 22:248–260, 1975.

[5] M. Bertoli, G. Casale, G. Serazzi. JMT: performance engineering tools for system modeling. *ACM PER*, 36(4), 10-15, March 2009.

[6] A. Bertozzi and J. McKenna. Multidimensional residues, generating functions, and their application to queueing networks. *SIAM Review*, 35(2):239–268, 1993.

[7] S. C. Bruell, G. Balbo, and P. V. Afshari. Mean value analysis of mixed, multiple class BCMP networks with load dependent service stations. *Perform. Eval*, 4:241–260, 1984.

[8] G. Casale. Accelerating Performance Inference over Closed Systems by Asymptotic Methods. *POMACS*, 1(1):1–25, 6 2017.

[9] P. Flajolet, R. Sedgewick. Mellin transforms and asymptotics: Finite differences and rice's integrals. *TCS: Theoretical Computer Science*, 144, 1995.

[10] J. J. Gordon. The evaluation of normalizing constants in closed queueing networks. *Operations Research* **38, 5** 863–869.

[11] H. W. Gould. Euler's formula for nth differences of powers. *The American Mathematical Monthly* **85**(6), 450–467, 1978.

[12] A. Grundmann, H. M. Möller. Invariant integration formulas for the n-simplex by combinatorial methods. *SIAM Journal on Numerical Analysis* **15**(2) 282–290, 1978.

[13] R. Kan. From moments of sum to moments of product. *Journal of Multivariate Analysis*, 99(3):542 – 554, 2008.

[14] E. Koenigsberg. Cyclic queues. *Operational Research Quarterly*, 9(1), 22–35, 1958.

[15] R. D. Nelson. The mathematics of product form queuing networks. *ACM Computing Surveys*, 25(3):339–369, 1993.