Queueing Syst (2011) 68:361–363
DOI 10.1007/s11134-011-9244-z

A lower bound for the Erlang C formula in the Halfin–Whitt regime

A.J.E.M. Janssen · Johan S.H. van Leeuwaarden · Bert Zwart

Received: 10 May 2011 / Revised: 10 May 2011 / Published online: 30 July 2011

Mathematics Subject Classification (2000) 60K25

One of the classical models of queueing theory is the $M/M/s$ queue or Erlang delay model. This model has s homogeneous servers working in parallel. Customers arrive according to a Poisson process with arrival rate λ, and the service times are independent and exponentially distributed with mean $1/\mu$. Let the offered load be $a = \lambda/\mu$ and assume $a < s$ to have a proper steady-state distribution. The most important performance characteristic for this system is the probability that a customer is delayed when arriving at the system in steady state. This probability is known as the Erlang C formula, given by (with $\rho = a/s$)

$$C(s,a) = \frac{\frac{a^s}{s!(1-\rho)}}{\sum_{k=0}^{s-1} a^k/k! + \frac{a^s}{s!(1-\rho)}} = \left(\rho + (1-\rho)\frac{\mathbb{P}[\mathrm{Pois}(a) \le s]}{\mathbb{P}[\mathrm{Pois}(a) = s]}\right)^{-1} \tag{1}$$

with $\mathrm{Pois}(a)$ a Poisson random variable with mean a.

A.J.E.M. Janssen
EURANDOM and the Department of Electrical Engineering, Eindhoven University of Technology,
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
e-mail: a.j.e.m.janssen@tue.nl

J.S.H. van Leeuwaarden
Department of Mathematics and Computer Science, Eindhoven University of Technology,
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
e-mail: j.s.h.v.leeuwaarden@tue.nl

B. Zwart (✉)
CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
e-mail: bertz@cwi.nl

The Halfin–Whitt regime refers to the scaling of the arrival rate λ and the number of servers s such that, while both λ and s increase toward infinity, the traffic intensity $\rho = \lambda/s$ approaches one, while

$$(1-\rho)\sqrt{s} \to \beta \tag{2}$$

for some $\beta \in (0,\infty)$. Halfin and Whitt [2] showed that setting $s = a + \beta\sqrt{a}$ (square-root staffing) for any fixed $\beta > 0$ yields

$$\lim_{a\to\infty} C\big(a+\beta\sqrt{a},a\big) = C_*(\beta) := \left(1+\beta\frac{\Phi(\beta)}{\phi(\beta)}\right)^{-1}, \tag{3}$$

with Φ, ϕ the distribution and density of the standard normal random variable. Hence, the scaling (2) combines large capacity with high utilization such that the probability of delay converges to a non-degenerate limit away from both zero and one. Halfin and Whitt went further and established a stochastic-process limit involving the convergence of the scaled queue-length process to a diffusion process. However, in this short note we just focus on the steady-state limit result in (3) and the suggested approximation

$$C(a+\beta\sqrt{a},a) \approx C_*(\beta) \quad \text{for large } a. \tag{4}$$

In recent years, such approximations have found application in many-server systems, like call centers. In particular, Borst, Mandelbaum and Reiman [1] developed an asymptotic framework that applies the square-root staffing principle to constraint satisfaction problems and cost minimization problems. The simplest constraint satisfaction problem is to determine the number of servers necessary to ensure that the fraction of customers that need to wait is below a certain threshold, say ϵ. Borst, Mandelbaum and Reiman proposed to determine the number of servers as a round-off of $s_* = \lambda + \beta_*(\epsilon)\sqrt{\lambda}$, with β_* the solution of $C_*(\beta(\epsilon)) = \epsilon$. A natural question is how well this approximation performs, and they observed that square-root staffing is accurate over a wide range of system parameters. A formal justification of this fact was presented in Janssen, Van Leeuwaarden and Zwart [5].

The formula $C(s,\lambda)$ in its basic form is only defined for integer values of s. The remaining statements in this paper will pertain to the analytic continuation of C given by (see [3])

$$C(s,a) = \left(a\int_0^\infty t\mathrm{e}^{-at}(1+t)^{s-1}\,\mathrm{d}t\right)^{-1}, \tag{5}$$

which is valid for all $0 < a < s$. We are now ready to state our two conjectures:

Conjecture 1 *$C(a+\beta\sqrt{a},a) \geq C_*(\beta)$ for all $\beta, a > 0$.*

Conjecture 2 *$C(a+\beta\sqrt{a},a)$ is strictly decreasing in a for each fixed $\beta > 0$.*

Both conjectures were verified by means of extensive numerical experiments. Note that the traffic intensity

$$\rho = \frac{a}{s} = \frac{a}{a+\beta\sqrt{a}}$$

monotonically *increases* in a for any fixed $\beta > 0$, while Conjecture 2 shows that the pre-limit steady-state delay probability *decreases* in a for any fixed $\beta > 0$, and hence decreases as a function of the load. This is somewhat surprising in view of the common wisdom that the performance of queueing systems deteriorates as the load increases. However, for the scaling at hand, not only the load increases in a, but the system size s increases at the same time. Hence, one way to provide intuition for Conjecture 2 is the classical economies-of-scale argument.

Apart from the fact that any new result for such a classical model is interesting, there is additional motivation for proving these conjectures:

(i) Conjecture 1 would imply that using the Halfin–Whitt approximation $C_*(\beta)$ in constraint satisfaction problems as in [1] *always* results in understaffing. This fact has been observed in [4].
(ii) The Erlang C formula is clearly decreasing in s and increasing in a. However, for the Halfin–Whitt scaling $s = a + \beta\sqrt{a}$ with $\beta > 0$ fixed, there is no obvious ordering between two systems (indexed by s), making a stochastic comparison difficult.

It is clear that Conjecture 2 implies Conjecture 1. Both conjectures are formulated as precise mathematical problems, but we do not see any outline for a potential proof, although the bounds on the Erlang C formula as provided in [4] may be instrumental as a starting point.

Acknowledgements We thank an anonymous referee for subjecting Conjecture 1 to close scrutiny and for suggesting Conjecture 2.

References

1. Borst, S., Mandelbaum, A., Reiman, M.: Dimensioning large call centers. Oper. Res. **52**, 17–34 (2004)
2. Halfin, S., Whitt, W.: Heavy-traffic limits for queues with many exponential servers. Oper. Res. **29**, 567–588 (1981)
3. Jagers, A.A., van Doorn, E.A.: On the continued Erlang loss function. Oper. Res. Lett. **5**, 43–46 (1986)
4. Janssen, A.J.E.M., van Leeuwaarden, J.S.H., Zwart, B.: Gaussian expansions and bounds for the Poisson distribution applied to the Erlang B formula. Adv. Appl. Probab. **40** (2008)
5. Janssen, A.J.E.M., van Leeuwaarden, J.S.H., Zwart, B.: Refining square root safety staffing by expanding Erlang C. Oper. Res. (2009, to appear)