Operations Research Letters 44 (2016) 359–365

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

Optimality gaps in asymptotic dimensioning of many-server systems

Jaron Sanders *, S.C. Borst, A.J.E.M. Janssen, J.S.H. van Leeuwaarden

Department of Mathematics & Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

ARTICLE INFO

Article history:
Received 5 November 2015
Received in revised form
26 February 2016
Accepted 4 March 2016
Available online 11 March 2016

Keywords:
QED regime
Halfin–Whitt regime
Queues in heavy traffic
Asymptotic analysis
Asymptotic dimensioning
Optimality gap

ABSTRACT

The Quality-and-Efficiency-Driven (QED) regime provides a basis for solving asymptotic dimensioning problems that trade off revenue, costs and service quality. We derive bounds for the *optimality gaps* that capture the differences between the true optimum and the asymptotic optimum based on the QED approximations. Our bounds generalize earlier results for classical many-server systems. We also apply our bounds to a many-server system with threshold control.

1. Introduction

The theory of square-root staffing in many-server systems ranks among the most celebrated principles in applied probability. The general idea behind square-root staffing is as follows: a finite server system is modeled as a system in heavy traffic, where the number of servers s is large, whereas at the same time, the system is critically loaded. Under Markovian assumptions, and denoting the load on the system by λ, this can be achieved by setting $s = \lambda + \beta\sqrt{\lambda}$ and letting $\lambda \to \infty$ while keeping $\beta > 0$ fixed, or alternatively setting $\lambda = s - \gamma\sqrt{s}$ and letting $s \to \infty$ while keeping $\gamma > 0$ fixed. In both cases, the system reaches the desirable Quality-and-Efficiency-Driven (QED) regime.

The QED regime refers to mathematically defined conditions in which both customers and system operators benefit from the advantages that come with systems that operate efficiently at large scale, which is particularly relevant for systems in e.g. health care, cloud computing, and customer services. Such conditions manifest themselves in a low delay probability and negligible mean delay, despite the fact that the system utilization is high. Properties of this sort can be proven rigorously for systems such as the $M/M/s$ queue by establishing stochastic-process limits under the aforementioned QED scalings [2]. The QED regime also creates a natural environment for solving dimensioning problems that achieve an acceptable trade-off between service quality and capacity. Quality is usually formulated in terms of some target service level. Take for instance the probability that an arriving customer experiences delay. The target could be to keep the delay probability below some value $\epsilon \in (0, 1)$. The smaller ϵ, the better the offered quality of service. Once the target service level is set, the objective from the operator's perspective is to determine the highest load λ such that the target ϵ is still met.

For the $M/M/s$ queue, it was shown by Borst et al. [1] that such dimensioning procedures combined with QED approximations have certain asymptotic optimality properties. To illustrate this, consider the case of linear costs, i.e. waiting cost is b per customer per unit time, and staffing cost is c per server per unit time. Denoting the total cost by $K_\lambda(s)$, it can be shown that when $s = \lambda + \beta\sqrt{\lambda}$ and $\beta > 0$,

$$K_\lambda(s) = b\lambda \frac{C_\lambda(s)}{s-\lambda} + cs = c\lambda + \sqrt{\lambda}\Big(c\beta + \frac{b}{\beta}C_\lambda(s)\Big) \quad (1)$$

with $C_\lambda(s)$ the delay probability in the $M/M/s$ queue. The first term $c\lambda$ represents the cost of the minimally required capacity λ, while the second term gathers the cost factors that are all $O(\sqrt{\lambda})$. Halfin and Whitt [2] showed that in the QED regime $C_\lambda(s)$ converges to a nondegenerate limit $C_0(\beta) \in (0, 1)$, so that in the QED regime one only needs to determine $\beta_0 = \arg\min_\beta\{c\beta + bC_0(\beta)/\beta\}$, and then set $s_0 = [\lambda + \beta_0\sqrt{\lambda}]$ as an approximation for the optimal number of servers $s^{\mathrm{opt}} = \arg\min_s\{K_\lambda(s)\}$. Borst et al. [1] called this procedure *asymptotic dimensioning*.

* Corresponding author.
E-mail addresses: jaron.sanders@tue.nl (J. Sanders), s.c.borst@tue.nl (S.C. Borst), a.j.e.m.janssen@tue.nl (A.J.E.M. Janssen), j.s.h.v.leeuwaarden@tue.nl (J.S.H. van Leeuwaarden).

http://dx.doi.org/10.1016/j.orl.2016.03.003

Based on the QED limiting regime, one expects that such approximate solutions are accurate for relatively large loads λ. For the *optimality gaps* $|s_0 - s^{\text{opt}}|$ and $|K_\lambda(s_0) - K_\lambda(s^{\text{opt}})|$, i.e. inaccuracies that arise from the fact that the actual system is of finite size, Borst et al. [1] showed through numerical experiments that the approximation s_0 performs exceptionally well in almost all circumstances, even when systems are only moderately sized. A rigorous underpinning for these observations was provided by Janssen et al. [5], who used *refined* QED approximations to quantify the optimality gaps. The delay probability, for instance, was shown to behave as $C_0(\beta) + C_1(\beta)/\sqrt{\lambda} + O(\lambda^{-1})$, which in turn was used to estimate the optimality gaps for the dimensioning problem in (1). Zhang et al. [8] obtained similar results for optimality gaps in the context of the $M/M/s + M$ queue, in which customers may abandon before receiving service.

Motivated by the results in [5,8], Randhawa [6] took a more abstract approach to quantify optimality gaps of asymptotic dimensioning problems. He showed under general assumptions that when the approximation to the objective function is accurate up to $O(1)$, the prescriptions that are derived from this approximation are $o(1)$-optimal. The optimality gap thus becomes zero asymptotically. This general setup was shown in [6] to apply to the $M/M/s$ queues in the QED regime, which confirmed and sharpened the results on the optimality gaps in [5,8] by implying that $|K_\lambda(s_0) - K_\lambda(s^{\text{opt}})| = o(1)$. The abstract framework in [6], however, can only be applied if refined approximations as in [5,8] are available.

Such refined approximations were recently developed in [4,7] for a broad class of many-server systems operating in the QED regime with $\lambda = s - \gamma\sqrt{s}$, and equipped with an admission control policy and a revenue structure. For a wide range of performance metrics, $M_s(\gamma)$ say, these refinements are of the form $M_s(\gamma) = M_0(\gamma) + M_1(\gamma)/\sqrt{s} + \cdots$. The method in [4,7] can deliver as many higher-order terms as needed, and generate all the refinements obtained in [5,8,6].

In the present paper, we demonstrate how the results in [4,7] can be leveraged to determine the optimality gaps of novel asymptotic dimensioning problems for a large class of many-server systems. Our main result (Theorem 1) provides generic bounds for the optimality gaps that become sharper when more terms in the QED expansion for $M_s(\gamma)$ are included.

2. Model description

2.1. Service systems with admission control and revenues

We consider many-server systems with s parallel servers, to which customers arrive according to a Poisson process with rate λ. Every customer requires an exponentially distributed service time with mean one. If a customer arrives and finds $k - s \geq 0$ customers waiting, the customer is allowed to join the queue with probability $p_s(k-s)$ and is rejected with probability $1 - p_s(k-s)$. The total number of customers in the system evolves as a birth–death process $\{X_s(t)\}_{t\geq 0}$ and has a stationary distribution

$$\pi_s(k) = \begin{cases} Z^{-1}, & k = 0, \\ Z^{-1}\dfrac{(s\rho)^k}{k!}, & k = 1, 2, \ldots, s, \\ Z^{-1}\dfrac{s^s\rho^k}{s!}\displaystyle\prod_{i=0}^{k-s-1} p_s(i), & k = s+1, s+2, \ldots, \end{cases} \tag{2}$$

where $\rho = \lambda/s$, $Z = \sum_{k=0}^{s}(s\rho)^k/k! + ((s\rho)^s/s!)F_s(\rho)$, and $F_s(\rho) = \sum_{n=0}^{\infty} p_s(0)\cdot\cdots\cdot p_s(n)\rho^{n+1}$. The stationary distribution in (2) exists if and only if the relative load ρ and the *admission control policy* $\{p_s(k)\}_{k\in\mathbb{N}_0}$ are such that $F_s(\rho) < \infty$.

Next, we assume that the system generates revenue at rate $r_s(k) \in \mathbb{R}$ when there are k customers in the system. The sequence $\{r_s(k)\}_{k\in\mathbb{N}_0}$ will be called the *revenue structure*. The stationary rate at which the system generates revenue is then given by

$$R_s(\gamma) = \sum_{k=0}^{\infty} r_s(k)\pi_s(k), \tag{3}$$

which depends via the equilibrium distribution on the admission control policy. Ref. [7] discusses the problem of maximizing the revenue rate over the set of all admission control policies.

One advantage of considering general admission control policies and revenue structures is that one can study different service systems and steady-state performance measures through one unifying framework. For example, the equilibrium behavior of the canonical $M/M/s/s$, $M/M/s$, and $M/M/s + M$ systems can be recovered by choosing $p_s(k-s) = 0$, $p_s(k-s) = 1$, and $p_s(k-s) = 1/(1 + (k - s + 1)\theta/s)$, respectively. Here, θ corresponds to the rate at which waiting customers abandon from the $M/M/s + M$ system. Similarly, the delay probability $D_s(\gamma) = \sum_{k=s}^{\infty}\pi_s(k)$ can be recovered by setting $r_s(k) = \mathbb{1}[k \geq s]$, the mean queue length $Q_s(\gamma) = \sum_{k=s}^{\infty}(k-s)\pi_s(k)$ is recovered when considering $r_s(k) = (k-s)\mathbb{1}[k \geq s]$, and the average number of idle servers $I_s(\gamma) = \sum_{k=0}^{s-1}(s-k)\pi_s(k)$ follows from $r_s(k) = (s-k)\mathbb{1}[k < s]$.

As a primary example we will consider a scenario where besides the waiting cost $b > 0$ incurred per customer per unit time, a fee $a > 0$ is received for every served customer, and a penalty $d \geq 0$ is imposed for rejecting a customer. The latter cost accounts for the degree of revenue loss from the admission control policy. Denoting by $D_s^R(\gamma) = \sum_{k=s}^{\infty}(1 - p_s(k-s))\pi_s(k)$ the probability that an arriving customer is rejected, and by $W_s(\gamma) = \sum_{k=s}^{\infty}((k - s + 1)/s)p_s(k-s)\pi_s(k)$ the expected waiting time of an arriving customer, the total system revenue rate is given by

$$R_s(\gamma) = a\lambda(1 - D_s^R(\gamma)) - b\lambda W_s(\gamma) - d\lambda D_s^R(\gamma). \tag{4}$$

By virtue of Little's law $\lambda W_s(\gamma) = Q_s(\gamma)$ and $\lambda(1 - D_s^R(\gamma)) = s - I_s(\gamma)$, and since $\lambda = s - \gamma\sqrt{s}$, the revenue rate can equivalently be expressed as

$$R_s(\gamma) = as + d\gamma\sqrt{s} - (a+d)I_s(\gamma) - bQ_s(\gamma). \tag{5}$$

This scenario therefore corresponds to the revenue structure

$$r_s(k) = \begin{cases} ak + d\gamma\sqrt{s} - d(s-k) & k < s, \\ as + d\gamma\sqrt{s} - b(k-s), & k \geq s. \end{cases} \tag{6}$$

2.2. QED scaling and refinements

We now discuss how to apply the QED scaling to obtain an asymptotic expansion for $R_s(\gamma)$ for general revenue structures $\{r_s(k)\}_{k\in\mathbb{N}_0}$, which we will exploit in Section 3 to characterize the asymptotic optimality gap. We impose the following three conditions throughout this paper:

(i) The arrival rate and system size are coupled via the scaling $\lambda = s - \gamma\sqrt{s}$;
(ii) $\lim_{s\to\infty} |p_s(0)\cdots p_s(n) - f((n+1)/\sqrt{s})| = 0$ where $f(x)$ is either a continuous, nonincreasing function, or $f(x) = \mathbb{1}[x \leq \eta]$;
(iii) There exist sequences $\{n_s\}_{s\in\mathbb{N}_+}$, $\{q_s\}_{s\in\mathbb{N}_+}$ with $q_s > 0$, and a continuous function $r(x)$ that satisfy the scaling condition $\lim_{s\to\infty} |(r_s(k) - n_s)/q_s - r((k-s)/\sqrt{s})| = 0$.

It is proven in [4,7] that $\lim_{s\to\infty}(R_s(\gamma) - n_s)/q_s = R_0(\gamma)$ under conditions (i)–(iii), with

$$R_0(\gamma) = \frac{\int_{-\infty}^{0} r(x)\mathrm{e}^{-\frac{1}{2}x^2 - \gamma x}\mathrm{d}x + \int_0^{\infty} r(x)f(x)\mathrm{e}^{-\gamma x}\mathrm{d}x}{\frac{\Phi(\gamma)}{\phi(\gamma)} + \int_0^{\infty} f(x)\mathrm{e}^{-\gamma x}\mathrm{d}x}. \tag{7}$$