

Managing Capacity and Inventory Jointly for Large-Scale Manufacturing Systems: Square-Root Rules with Refinements

Josh Reed
Stern School of Business
New York University
New York, NY

Bo Zhang
School of Industrial
and Systems Engineering
Georgia Tech
Atlanta, GA

March 15, 2011

Abstract

We study the problem of jointly optimizing capacity and inventory levels for manufacturing systems where the production facility is modeled as a parallel-server queue. Our results provide scalable, square-root algorithms which provide approximate optimal solutions whose relative error is shown to diminish to zero as the arrival rate to the system grows large. Our analysis provides insights into the setting of the optimal capacity and inventory levels and some of the tradeoffs involved in so doing. We also study various extensions of our model.

1 Introduction

In a manufacturing system, production capacity sizing and inventory level decisions are important determinants of operating cost and customer experience. Moreover, it is important to make both of these decisions jointly in order to minimize costs. In a recent paper, Bradley and Glynn [6] consider the problem of minimizing costs by jointly optimizing capacity and inventory levels. In [6], the production facility is modeled as a $GI/M/1$ queue. Capacity decisions in [6] correspond to increasing or decreasing the service rate of the server in a $GI/M/1$ queue.

In many situations, the capacity of an individual server may itself be limited and so large increases in system capacity may only be achieved by the acquisition of additional servers. Although perhaps easy to overlook at first, modeling a production facility as c separate servers each operating at a capacity of μ as opposed to a single server with a rate of $c\mu$ can have significant effects on the insights gained into overall system performance. These effects become even more pronounced when one considers high-volume systems requiring a large number of servers [16] in order to meet customer demand. In the present paper, we consider the problem of jointly optimizing the capacity

and inventory levels of a production facility where the production facility consists of c individual servers. Increases in the capacity of the production facility correspond to adding additional servers.

Our main goal of the present paper is to provide optimal inventory and capacity levels for a production system modeled as c individual servers. However, the joint decision problem that we study leads to a computationally complex optimization problem whose exact solution suffers from lack of insights and computational scalability in terms of the customer demand rate (or the offered workload to the production facility). In particular, the cost objective function is not convex and so finding its global minimum relies on a numerical search solution procedure. The computational complexity of the exact numerical search approach grows with the magnitude of the problem data, in particular, as noted above, with the demand rate.

In order to circumvent these difficulties, we propose an approximate approach that is both insightful and computationally scalable. Our proposed approach is based on considering systems where the customer demand rate is high. We show that for such systems it is optimal to closely match system capacity with the incoming demand rate. In particular, to adopt a phrase from the call center literature, it is optimal to operate in the QED regime in which the number of servers is not far off from the offered load and, in fact, is off by at most a square root factor. As part of our results, we show that solving for the optimal capacity and inventory levels using diffusion approximations for the $M/M/c$ queue in the QED regime, one obtains capacity and inventory levels whose difference from the true optimal levels remain bounded, independent of the system size. Moreover, using new results on corrected diffusion approximations for the elementary $M/M/c$ queue [18, 24], we are able to show that the difference between our proposed capacity and inventory levels and the optimal levels actually converges to a constant as the system size grows to infinity, and, we are able to explicitly calculate this constant, thus further refining our proposed capacity and inventory prescriptions.

There is a large body of literature on setting optimal inventory levels for systems with either a fixed or unlimited capacity. Clark and Scarf [9] show that base stock policies are optimal for multi-echelon, uncapacitated systems with unlimited capacity. Federgruen and Zipkin [11, 12] consider capacitated, single echelon systems and show that a base stock policy is optimal for both average-cost and the discounted-cost cost criteria. Glasserman [13] considers single-stage and multi-stage systems and obtains bounds and asymptotics for key performance measures as the target level increases. Glasserman and Liu [14] obtain corrected diffusion approximations for several performance measures for multi-stage, capacitated systems operating under base stock policies. Rubio and Wein [23] treat the production facility in the more general terms and derive optimal base stock levels for multi-product systems whose production facilities have product form stationary distributions. In an interesting, recent paper, Rubino and Ata [2] also consider a parallel-server system in a make-to-order context with order cancellations.

Our work also contributes to the literature on square-root staffing policies for parallel-server queues and, specifically, on refined square-root staffing policies. Borst, Mandelbaum and Reiman [5] show that under a general cost structure on the number of servers and the waiting times, the optimal number of servers to staff for the $M/M/c$ queues is within a square root factor of the

offered load. This work was followed up upon by Janseen, van Leeuwaarden and Zwart [19] in which refined square root staffing formulas were given. The work of Zhang, van Leeuwaarden and Zwart [30] presents refined square root staffing for parallel server queues with impatience. In the present paper, we extend these techniques to provide refined square root capacity and inventory levels for a two-dimensional optimization problem. An interesting observation is that our two-dimensional refinement may be characterized as the solution to a system of linear equations, whose coefficients we may explicitly compute.

2 Model description and problem statement

The model we consider is a single-product, make-to-stock manufacturing system with the production facility modeled by a parallel-server queue. Later in the paper, we extend our results to the case of multiple products.

Customer orders arrive according to a Poisson process with rate λ and are fulfilled from the finished-goods inventory immediately if a product is available. If no on-hand inventory is available at the time of an order arrival, then the order is backlogged and we assume that backlogged orders are fulfilled on a first-come-first-serve basis as finished-goods inventory is replenished from the production facility. Each unit of finished-goods inventory incurs a holding cost per unit time of h and each order backlogged incurs a penalty cost per unit time of p .

The production facility is modeled as c parallel servers. The amount of time to process an order on an individual server is exponentially distributed with rate μ and we assume that the servers operate independently of one another. The cost per unit time for running one server is denoted by d . Ultimately, in our model, it will be up to the system manager to choose the appropriate number of servers c to operate based on cost-efficiency tradeoffs. We also denote by w the work-in-process (WIP) cost per unit time for inventory at the production facility.

In [11, 12], it is shown that for a fixed capacity level c , the optimal inventory control policy is a base-stock policy. We also assume in our model that the system manager follows a base-stock inventory control policy, i.e., for a given base-stock or target-level s the system manager chooses to operate the production facility in such a manner as to bring the total units of finished-goods inventory up to the level of s . However, note that in our case, since there are multiple servers working in parallel at the production facility, the system manager must be careful not to overproduce and bring the amount of finished-goods inventory above the target-level s . The system manager will therefore operate the system so that the sum of the amount of finished-goods inventory plus the amount of WIP at the production facility is always equal to s . Backorders are represented as negative finished-goods inventory.

As in [6, 13, 25], we model the amount of finished-goods inventory using the shortfall process which keeps track of the difference between the target-level s and the actual total amount of finished-goods inventory. Let $N_A = \{N_A(t), t \geq 0\}$ and $N_S = \{N_S(t), t \geq 0\}$ be rate one, Poisson processes which are independent of one another. Also, let $Q = \{Q(t), t \geq 0\}$ denote the shortfall

process. For a fixed number of servers c , the dynamics of the shortfall process are then given by

$$Q(t) = Q(0) + N_A(\lambda t) - N_S \left(\mu \int_0^t (Q(s) \wedge c) ds \right), \quad (1)$$

for any $t \geq 0$. Note that the shortfall process is independent of the target-level s selected. Moreover, in terms of the shortfall process Q and the target level s , the total amount of finished-goods inventory at any point in time t is given by $(s - Q(t))^+$ and the backorder level is given by $(Q(t) - s)^+$, where $a^+ = \max\{a, 0\}$ for any real number a throughout the paper. The amount of WIP at the production facility at time t is simply $Q(t)$.

Now observe that by (1) the dynamics of the shortfall process are identical to that of an $M/M/c$ queue with an arrival rate of λ and a service rate μ (see for instance [21]). Hence, defining $R := \lambda/\mu$ to be the offered load of the system and assuming that $R < c$, it follows by standard results [15, 22] that $Q(t)$ converges in distribution to $Q(\infty)$ as $t \rightarrow \infty$, where

$$\mathbb{P}_c\{Q(\infty) = k\} = (R^k/k!)\eta, \text{ for } k \leq c, \quad (2)$$

and

$$\mathbb{P}_c\{Q(\infty) = k\} = (c^c(R/c)^k/c!)\eta, \text{ for } k \geq c, \quad (3)$$

where

$$\eta = \left(R^c/(c!(1 - R/c)) + \sum_{k=0}^{c-1} R^k \right)^{-1}. \quad (4)$$

Note that we use a subscript c on the probability operators in (2) and (3) in order to emphasize the dependence of the limiting shortfall distribution $Q(\infty)$ on the chosen capacity level c and also later on we do the same with expectation operators (e.g., in (5)).

Assuming now that the system capacity is greater than the offered load, $R < c$, the average cost per unit time of the system when operating under a base-stock policy with target-level s and with c servers is given by

$$\Pi(c, s, R) = d \cdot c + w \cdot \mathbb{E}_c[Q(\infty)] + h \cdot \mathbb{E}_c[(s - Q(\infty))^+] + p \cdot \mathbb{E}_c[(Q(\infty) - s)^+]. \quad (5)$$

The system manager's objective is now to determine the production capacity size c and the inventory base-stock level s in order to minimize the total long-run average cost per unit time as given by (5). Note that for a fixed capacity level c and offered load R , the objective function in (5) reduces to a standard newsvendor problem in terms of the target-level s . Its optimal solution is a critical fractile solution given by $F_{c, Q(\infty)}^{-1}(p/(p+h))$, where $F_{c, Q(\infty)}^{-1}$ denotes the inverse of the CDF of $Q(\infty)$, assuming a capacity level of c . In theory, this solution may be computed using (2)-(4). However, inverting $F_{c, Q(\infty)}$ does not scale well with the offered load of the system. Moreover, in order to

find a global minimum, one also needs to optimize over capacity levels c . We therefore suggest two alternative approaches which are computationally invariant to the size of the system R and provide solutions that, respectively, are within a constant factor of and coincide with the optimal ones as $R \rightarrow \infty$.

3 Main results

In this section, we develop a square-root approximation rule for solving the joint optimization (5) that is accurate up to a constant factor asymptotically. We then present a refined algorithm which further corrects the asymptotically constant error and hence, finds solutions that are asymptotically exact. We also discuss the insights yielded by the square-root rule.

3.1 Square-root rule and refinement algorithm

In this subsection, we present both our square-root type algorithm and its refinement for setting the optimal capacity and base-stock level. In order to state our result, we must first introduce the following notation. Throughout the paper, we let $\Phi(\cdot)$ and $\phi(\cdot)$ denote the standard normal cumulative distribution and density functions, respectively. A real-valued function $f(R)$ is said to be $\mathcal{O}(g(R))$ if $\limsup_{R \rightarrow \infty} |f(R)/g(R)| < \infty$, and an \mathbb{R}^2 -valued function $(f_1(R), f_2(R))$ is said to be $\mathcal{O}(g(R))$ if $f_1(R) = \mathcal{O}(g(R))$ and $f_2(R) = \mathcal{O}(g(R))$.

Our first result provides a second order approximation for the average cost function $\Pi(c, s, R)$ in (5) in terms of the offered load R . Although c and s in (5) are integer-valued, our result is stated in terms of an analytic continuation of $\Pi(c, s, R)$ to real-valued c and s , with $c \in (R, \infty)$ and $s \in [0, \infty)$. This continuation is based on the continued Erlang B formula [17] and its derivation may be found in the appendix. We continue to use the notation $\Pi(c, s, R)$ in order to denote the analytic continuation.

Theorem 3.1 (Diffusion approximation for the average cost function). *For any $\beta > 0$ and $b \in (-\infty, \infty)$,*

$$\Pi(R + \beta\sqrt{R}, R + b\sqrt{R}, R) = (d + w)R + K_*(\beta, b)R^{1/2} + O(1),$$

where

$$C_*(\beta) := \left[1 + \frac{\beta\Phi(\beta)}{\phi(\beta)} \right]^{-1},$$

and

$$K_*(\beta, b) := \begin{cases} d\beta - pb + \frac{w+p}{\beta} \cdot C_*(\beta) + (p+h)D_*(\beta, b) \left[b + \frac{\phi(b)}{\Phi(b)} \right], & \text{if } b \leq \beta \\ d\beta + hb + \frac{w-h}{\beta} \cdot C_*(\beta) + (p+h)[1 - C_*(\beta)] \frac{\phi(\beta)}{\beta^2\Phi(\beta)} \cdot e^{-\beta(b-\beta)}, & \text{if } b > \beta \end{cases}, \quad (6)$$

with

$$D_*(\beta, b) := \begin{cases} [1 - C_*(\beta)]\Phi(b)/\Phi(\beta) = \beta\Phi(b)[\phi(\beta) + \beta\Phi(\beta)]^{-1}, & \text{if } b \leq \beta \\ 1 - C_*(\beta)e^{-\beta(b-\beta)}, & \text{if } b > \beta. \end{cases} \quad (7)$$

Theorem 3.1 suggests as an approximation to the optimal capacity and base-stock levels setting

$$c_* = R + \beta_* \sqrt{R} \quad \text{and} \quad s_* = R + b_* \sqrt{R}, \quad (8)$$

where β_* and b_* jointly minimize the function $K_*(\beta, b)$. The joint minimization of $K_*(\beta, b)$ may be achieved in the following way. First note that for $\beta > 0$ fixed, the form of $K_*(\beta, b)$ in (6) is identical to that of the cost function in the newsvendor problem. Indeed, one may write

$$K_*(\beta, b) = h \cdot \mathbb{E}_\beta[(b - \tilde{Q}(\infty))^+] + p \cdot \mathbb{E}_\beta[(\tilde{Q}(\infty) - b)^+],$$

where $\tilde{Q}(\infty)$ is a random variable with distribution $\mathbb{P}_\beta\{\tilde{Q}(\infty) \leq x\} = D_*(\beta, x)$. Hence, using a critical fractile solution, one has that for a given β , the optimal value of b is given by

$$w_*(\beta) := \begin{cases} \beta + \frac{1}{\beta} \ln \left[\frac{C_*(\beta) \cdot (p+h)}{h} \right], & \text{if } C_*(\beta) > \frac{h}{p+h}, \\ \Phi^{-1} \left(\frac{p\Phi(\beta)}{(p+h)[1-C_*(\beta)]} \right), & \text{if } C_*(\beta) \leq \frac{h}{p+h}. \end{cases} \quad (9)$$

Optimizing over β , one may then set $\beta_* := \arg \min_{\beta > 0} K_*(\beta, w_*(\beta))$ and, subsequently, $b_* := w_*(\beta_*)$. We refer to β_* and b_* obtained in this way as the square-root rule.

Given that Theorem 3.1 provides a characterization of the cost function which is correct up to a $O(1)$ term, it is natural to suspect that the proposed capacity and base-stock levels derived from Theorem 3.1 should be correct up to a $O(1)$ term as well. Our next result shows that indeed this is the case. Let

$$(c_{\text{opt}}, s_{\text{opt}}) := \arg \min_{(c,s) \in (R,\infty) \times [0,\infty)} \Pi(c, s, R)$$

jointly minimize $\Pi(c, s, R)$. We refer to the difference between (c_*, s_*) and $(c_{\text{opt}}, s_{\text{opt}})$ as the *optimality gap* of (c_*, s_*) .

Theorem 3.2.

$$(c_{\text{opt}}, s_{\text{opt}}) - (c_*, s_*) = \mathcal{O}(1). \quad (10)$$

Theorem 3.2 states that as the offered load of the system increases, the optimality gap of (c_*, s_*) remains bounded. Hence, for a problem instance with a high demand rate, the relative error made by (c_*, s_*) becomes negligible. Therefore, for large manufacturing systems, one may learn the behavior of the optimal capacity-inventory level by examining (c_*, s_*) , or simply (β_*, b_*) , and we shall perform such analysis in the next subsection.

However, the optimality gap, or the absolute error, of (c_*, s_*) does not diminish as $R \rightarrow \infty$. Also, for a specific problem with a small or moderate value of R , the asymptotic notion of $\mathcal{O}(1)$ accuracy may not be sufficiently strong to deliver a reliable solution. We therefore next derive a more detailed version of of Theorem 3.1 which explicitly identifies the $O(1)$ term and in turn allows us to provide refinements to c_* and s_* in (8).

Theorem 3.3 (Corrected diffusion approximation for the average cost function). *For any $\beta > 0$ and $b \in (-\infty, \infty)$,*

$$\Pi(R + \beta\sqrt{R}, R + b\sqrt{R}, R) - [(d + w)R + K_*(\beta, b)R^{1/2}] = K_\bullet(\beta, b) + \mathcal{O}(R^{-1/2}),$$

where the quantity $K_\bullet(\beta, b)$ is given by,

$$K_\bullet(\beta, b) := \begin{cases} \frac{w+p}{\beta} \cdot C_\bullet(\beta) + (p+h)L_{1,\bullet}(\beta, b), & \text{if } b \leq \beta \\ \frac{w-h}{\beta} \cdot C_\bullet(\beta) + (p+h)L_{2,\bullet}(\beta, b), & \text{if } b > \beta \end{cases}, \quad (11)$$

where

$$D_\bullet(\beta, b) = \begin{cases} [1 - C_*(\beta)]g_1(\beta, b) - C_\bullet(\beta)\Phi(b)\Phi(\beta)^{-1}, & \text{if } b \leq \beta \\ [1 - C_*(\beta)]g_2(\beta, b) - C_\bullet(\beta) [1 + (1 - e^{-\beta(b-\beta)})\beta\phi(\beta)\Phi(\beta)^{-1}], & \text{if } b > \beta \end{cases}, \quad (12)$$

$$g_1(\beta, b) = \frac{1}{6\Phi(\beta)} \left[\frac{\Phi(b)\phi(\beta)(\beta^2 + 2)}{\Phi(\beta)} - \phi(b)(b^2 - 4) \right], \quad (13)$$

$$g_2(\beta, b) = [1 - e^{-\beta(b-\beta)}] g_3(\beta)\beta^{-1} - \frac{\phi(\beta)}{\beta\Phi(\beta)} \left[\frac{1}{2}\beta^2(b - \beta) - \beta \right] e^{-\beta(b-\beta)}, \quad (14)$$

$$g_3(\beta) = \frac{1}{2}\beta \frac{\phi(\beta)}{\Phi(\beta)} + \frac{1}{6}\beta^2 \left(\frac{\phi(\beta)}{\Phi(\beta)} \right)^2 + \frac{1}{6}\beta^3 \frac{\phi(\beta)}{\Phi(\beta)} + \frac{1}{3} \left(\frac{\phi(\beta)}{\Phi(\beta)} \right)^2, \quad (15)$$

$$C_\bullet(\beta) = \beta C_*(\beta)^2 \left[\frac{1}{3} + \frac{\beta^2}{6} + \frac{\Phi(\beta)}{\phi(\beta)} \left(\frac{\beta}{2} + \frac{\beta^3}{6} \right) \right]. \quad (16)$$

$$L_{1,\bullet}(\beta, b) = \left[D_*(\beta, b)g_4(b) + D_\bullet(\beta, b) \left(b + \frac{\phi(b)}{\Phi(b)} \right) \right] \quad (17)$$

$$g_4(b) = \frac{1}{6}b^2 \left(\frac{\phi(b)}{\Phi(b)} \right)^2 + \frac{1}{6}b^3 \frac{\phi(b)}{\Phi(b)} - \frac{1}{2}b \frac{\phi(b)}{\Phi(b)} - \frac{2}{3} \left(\frac{\phi(b)}{\Phi(b)} \right)^2. \quad (18)$$

$$L_{2,\bullet}(\beta, b) = \beta^{-2}e^{-\beta(b-\beta)} \left[[1 - C_*(\beta)] \cdot g_3(\beta) - C_\bullet(\beta) \frac{\phi(\beta)}{\Phi(\beta)} + \frac{1}{2}[1 - C_*(\beta)] \cdot \frac{\phi(\beta)}{\Phi(\beta)} \beta^2(b - \beta) \right]. \quad (19)$$

Using Theorem 3.3 one may further refine the capacity and base-stock levels in (8). Indeed, we propose the following algorithm to determine the refined capacity and base-stock levels.

Algorithm 1 Refined square-root algorithm

1: Solve for

$$\beta_* := \arg \min_{\beta > 0} K_*(\beta, w_*(\beta)). \quad (20)$$

 2: Set $b_* = w_*(\beta_*)$

 3: Let $(\beta_\bullet, b_\bullet)$ be the solution to

$$\frac{\partial D_*(\beta_*, b_*)}{\partial \beta} \beta_\bullet + \frac{\partial D_*(\beta_*, b_*)}{\partial b} b_\bullet = -D_\bullet(\beta_*, b_*), \quad (21)$$

$$\frac{\partial^2 K_*(\beta_*, b_*)}{\partial \beta^2} \beta_\bullet + \frac{\partial^2 K_*(\beta_*, b_*)}{\partial \beta \partial b} b_\bullet = -\frac{\partial K_\bullet(\beta_*, b_*)}{\partial \beta}. \quad (22)$$

 4: Calculate $c_\bullet = R + \beta_* \sqrt{R} + \beta_\bullet$ and $s_\bullet = R + b_* \sqrt{R} + b_\bullet$.

 5: Set the capacity level to $\lceil c_\bullet \rceil$ and the base stock level to $\lceil s_\bullet \rceil$.

Steps 1 and 2 in our algorithm simply compute the values of β_* and b_* of the square-root rule. Next, step 3 in Algorithm 1 calculates a refinement $(\beta_\bullet, b_\bullet)$, which explicitly characterizes the dominating term of the $\mathcal{O}(1)$ optimality gap in (10) and, as we shall see shortly (cf. Theorem 3.4), improves it to $\mathcal{O}(R^{-1/2})$. Note that step 3 involves solving a system of linear equations, and that the explicit solution is given by

$$(\beta_\bullet, b_\bullet) = J^{-1} \cdot \left(D_\bullet(\beta_*, b_*) \frac{\partial^2 K_*(\beta_*, b_*)}{\partial \beta \partial b} - \frac{\partial D_*(\beta_*, b_*)}{\partial b} \frac{\partial K_\bullet(\beta_*, b_*)}{\partial \beta}, \quad (23)$$

$$\frac{\partial D_*(\beta_*, b_*)}{\partial \beta} \frac{\partial K_\bullet(\beta_*, b_*)}{\partial \beta} - \frac{\partial^2 K_*(\beta_*, b_*)}{\partial \beta^2} D_\bullet(\beta_*, b_*) \right), \quad (24)$$

where

$$J := \frac{\partial D_*(\beta_*, b_*)}{\partial b} \frac{\partial^2 K_*(\beta_*, b_*)}{\partial \beta^2} - \frac{\partial D_*(\beta_*, b_*)}{\partial \beta} \frac{\partial^2 K_*(\beta_*, b_*)}{\partial \beta \partial b}. \quad (25)$$

Although the coefficients in (21) and (22) in Step 3 do not appear simple, they are explicit and, perhaps more importantly, need only be calculated once regardless of the offered load R .

Our derivation of the refinement $(\beta_\bullet, b_\bullet)$ extends, to our two-dimensional joint optimization problem, the one-dimensional refined square-root staffing rule originally proposed by [19] and later formalized by [30]. Specifically, [19] and [30] consider constraint satisfaction problems, for which the objective is to determine the minimal staffing level necessary to reach a desired performance level. In our case, the exact optimal solution pair $(c_{\text{opt}}, s_{\text{opt}})$ must satisfy

$$\left(\frac{\partial \Pi(c_{\text{opt}}, s_{\text{opt}}, R)}{\partial c}, \frac{\partial \Pi(c_{\text{opt}}, s_{\text{opt}}, R)}{\partial s} \right) = (0, 0), \quad (26)$$

which simply can be viewed as a two-dimensional version of the constraint satisfaction problems considered by [19] and [30]. A detailed derivation leading to our refinements may be found in the appendix.

A salient feature of our algorithm is its computational scalability. In the above procedure, the most computationally difficult part, steps 1 through 3, are completely independent of the offered load R and do not involve testing candidate solutions (c, s) , whose magnitude are proportional to R . Such independence makes this algorithm particularly desirable for solving problems for large-scale systems, where R and the solution pair (c, s) take on large values. For different problem instances with different values of R , steps 1 through 3 need not be repeated and the same set of (β_*, b_*) and $(\mathbf{c}_\bullet, \mathbf{s}_\bullet)$ can be used to calculate the approximate solution $(\mathbf{c}_\bullet, \mathbf{s}_\bullet)$, as specified by Step 4 in Algorithm 1.

We complete this section with a result justifying our proposed algorithm. It states that our refined approximation solution $(\mathbf{c}_\bullet, \mathbf{s}_\bullet)$ has an optimality gap that decreases at a rate proportional to the square-root of the offered load.

Theorem 3.4.

$$(c_{\text{opt}}, s_{\text{opt}}) - (\mathbf{c}_\bullet, \mathbf{s}_\bullet) = \mathcal{O}(R^{-1/2}). \quad (27)$$

The $\mathcal{O}(R^{-1/2})$ optimality gap in (27) indicates a clear improvement over the $\mathcal{O}(1)$ performance of (c_*, s_*) . Specifically, the refinement algorithm leads to more accurate solutions than the square-root rule, especially for problem instances with small or moderate R values.

3.2 Insights and discussion

We now discuss some insights gained from the square-root rule presented in the previous subsection. Our approach is to examine the behavior of (β_*, b_*) . We adopt the term variability hedge from [3] by referring to $\beta_*\sqrt{R}$ as the (optimal) capacity variability hedge and $b_*\sqrt{R}$ as the (optimal) inventory variability hedge. We also refer to β_* and b_* as the (optimal) capacity and inventory safety factors, respectively.

We begin by separating the approximate cost function $K_*(\beta, b)$ appearing in Theorem 3.1 into four parts, each part representing the expected value of one of our four costs: capacity costs, WIP costs, finished-goods holding costs and backordering costs. Doing so, we obtain

$$K_*(\beta, b) = d \cdot K_d(\beta, b) + w \cdot K_w(\beta, b) + p \cdot K_p(\beta, b) + h \cdot K_h(\beta, b), \quad (28)$$

where $K_d(\beta, b) = \beta$, $K_w(\beta, b) = C_*(\beta)/\beta$,

$$K_p(\beta, b) = \begin{cases} \frac{C_*(\beta)}{\beta} - b + D_*(\beta, b) \left[b + \frac{\phi(b)}{\Phi(b)} \right], & \text{if } b \leq \beta, \\ [1 - C_*(\beta)] \frac{\phi(\beta)}{\beta^2 \Phi(\beta)} \cdot e^{-\beta(b-\beta)}, & \text{if } b > \beta, \end{cases}$$

and

$$K_h(\beta, b) = \begin{cases} D_*(\beta, b) \left[b + \frac{\phi(b)}{\Phi(b)} \right], & \text{if } b \leq \beta, \\ b - \frac{C_*(\beta)}{\beta} + [1 - C_*(\beta)] \frac{\phi(\beta)}{\beta^2 \Phi(\beta)} \cdot e^{-\beta(b-\beta)}, & \text{if } b > \beta. \end{cases}$$

One may easily verify that since at each point in time

$$\text{WIP} + \text{finished-goods inventory} = \text{base stock level } s + \text{backlog level},$$

it follows that

$$K_w(\beta, b) + K_h(\beta, b) = b + K_p(\beta, b). \quad (29)$$

We now take a closer look at each term appearing in (28). Figure 1 shows that for a fixed base-stock level b , the average number of orders backordered, $K_p(\beta, b)$, is decreasing in β . This observation should not be surprising since an increase in production capacity will improve the system's responsiveness to customer demand and result in a lower average backorder level. Less obvious, however, is the fact that increasing capacity reduces the total amount of inventory in the system (WIP plus finished-goods inventory), as can be immediately deduced from the fact that $K_p(\beta, b)$ is decreasing in β and relation (29).

In addition, the production capacity level affects the way in which inventory is distributed in the pipeline. Specifically, $K_w(\beta, b)$ can be shown to be decreasing in β easily, while $K_h(\beta, b)$ is increasing in β , as illustrated in Figure 2. Therefore, since the sum of $K_w(\beta, b)$ and $K_h(\beta, b)$ is decreasing in β , it follows that increasing capacity pushes a larger percentage of the total inventory towards the retailer.

The above findings suggest that the approximate optimal safety factor β_* tends to increase as w increases, p increases or h decreases, or any combination thereof. Also, one may easily see from $K_d(\beta, b) = \beta$ that β_* tends to increase as the server operating cost d decreases.

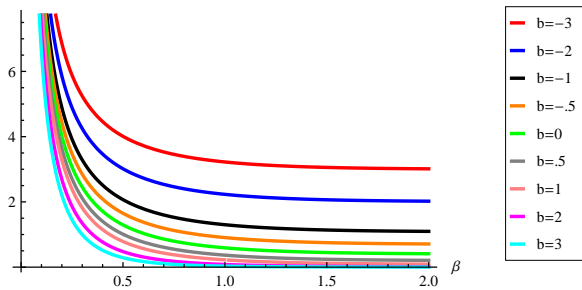


Figure 1: $K_p(\beta, b)$ with fixed b values.

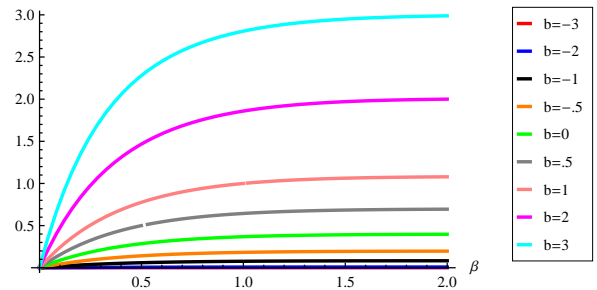


Figure 2: $K_h(\beta, b)$ with fixed b values.

Recall next the definition of $w_*(\cdot)$ in (9) as the optimal base-stock safety factor b corresponding to a particular capacity safety factor β . In order to better understand the dependence of the

optimal base stock level on the corresponding production capacity level, we plot $w_*(\cdot)$ in Figure 3 for varying values of h/p . For a fixed value of h/p , $w_*(\cdot)$ is clearly decreasing in β and hence the optimal base stock level is decreasing in the capacity level. Moreover, one has that

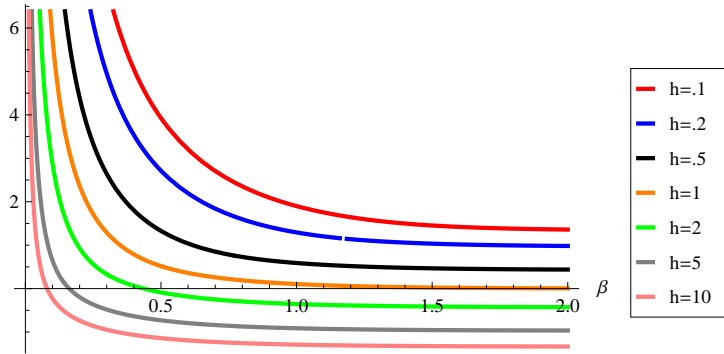


Figure 3: $w_*(\beta)$ with $p = 1$ and h taking on different values.

$$\lim_{\beta \rightarrow 0} \left[w_*(\beta) - \frac{1}{\beta} \ln \left(\frac{p+h}{h} \right) \right] = -\frac{\sqrt{2\pi}}{2}, \quad (30)$$

which follows since

$$C_*(\beta) = 1 - \frac{\sqrt{2\pi}}{2} \beta + o(\beta), \quad (31)$$

where a function $f(\beta)$ is said to be $o(\beta)$ if $\lim_{\beta \rightarrow 0} f(\beta)/\beta = 0$. Hence, from (30) we observe that the optimal inventory and capacity safety factors are inversely related to one another when the capacity level is low. On the other hand, if the capacity variability hedge is high, then the c -server facility behaves similarly to an infinite-server queue and the actual capacity level becomes irrelevant. In this case, the steady-state shortfall distribution is approximately normal. That is, $D_*(\beta, b) \approx \Phi(b)$. Hence, we obtain that $w_*(\beta) \approx \Phi^{-1} \left(\frac{p}{p+h} \right)$ and, in fact, one may easily show that

$$\lim_{\beta \rightarrow \infty} w_*(\beta) = \Phi^{-1} \left(\frac{p}{p+h} \right). \quad (32)$$

Figure 3 further demonstrates that the rate of convergence in (32) is very fast. In particular, as long as β is not too small, the infinite-server normal approximation seems to be very accurate.

Economies of Scale. We next point out that the square-root form of the optimal solution carries with it important insights. For example, the square-root rule, often referred to as *square-root (safety) staffing* in the context of service systems (e.g., see the discussion on p. 200 to 201 in

[26]), quantifies the economies of scale (EOS) that can be achieved by combining several isolated systems into a larger one. In order to illustrate this point, consider two identical manufacturing systems both with the same set of cost parameters (d, w, p, h) and hence, by the results of the previous section, the same safety factor (β_*, b_*) , where β_* must be strictly positive in order for the system to be stable. Moreover, assume that both systems have an offered load of R . If these two systems are operated independently, then, by Theorem 3.2, the optimal total capacity level is approximately $2R + 2\beta_*\sqrt{R}$. On the other hand, if they are combined together the optimal capacity size is approximately $2R + \beta_*\sqrt{2R}$, which equates to saving of $(2 - \sqrt{2})\beta_*\sqrt{R}$ servers. More generally, an n -fold increase in the offered load R requires that the capacity variability hedge increase by only \sqrt{n} times.

Although the pooling of two systems together will always result in a decrease in the optimal capacity level, it may lead to an increase in the base stock level in order to compensate. Specifically, this is the case when b_* is negative. This can for instance happen if $\frac{p}{p+h} < 1/2$ and β_* is sufficiently large (see Figure 3). In such a case, combining two systems together, each with an offered load R and identical costs parameters, will result in an increase in the cost-optimal total base stock level by approximately $(2 - \sqrt{2})|b_*|\sqrt{R}$ units. Nevertheless, even though the base stock level may rise due to pooling, the average amount of finished-goods inventory level will always decrease. To see this, note that even when $b_* < 0$ and thus $s_* < R < c_*$, the average finished-goods inventory level as given by Lemma A.3 can be shown to have the following form (see (109) in the Appendix):

$$\mathbb{E}_{c_*}[(s_* - Q(\infty))^+] = L_1(c_*, s_*, R) = D_*(\beta_*, b_*) \left[b_* + \frac{\phi(b_*)}{\Phi(b_*)} \right] R^{1/2} + \mathcal{O}(1). \quad (33)$$

One can also easily show that $D_*(\beta_*, b_*) > 0$ for any (β_*, b_*) and that for any $b_* < 0$,

$$b_* + \frac{\phi(b_*)}{\Phi(b_*)} \geq 0, \quad (34)$$

which is due to the relation that $\Phi(-x) \leq \phi(x)/x$ for any $x > 0$ (see Theorem 1.2.3 on p. 11 in [10]). We therefore have that the coefficient in front of $R^{1/2}$ in (33) is non-negative and consequently economies of scale holds for the average inventory level (or cost).

We next discuss an important measure of service level in our context, namely, the fill rate [4, 20], which measures the fraction of demand met from on-hand finished-goods inventory without delay. By definition, the fill rate is given by

$$\mathbb{P}_c\{Q(\infty) \leq s - 1\} = \mathbb{P}_c\{Q(\infty) \leq s\} - \mathbb{P}_c\{Q(\infty) = s\}.$$

However, because $\mathbb{P}_c\{Q(\infty) = s\} = \mathcal{O}(R^{-1/2})$ for any s (see Proposition 2 in [16]), the fill rate in a system with a high offered load approximately equals $\mathbb{P}_c\{Q(\infty) \leq s\}$. Therefore, we have that under the optimal capacity-inventory prescription,

$$\text{fill rate} \approx \mathbb{P}_{c_{\text{opt}}}\{Q(\infty) \leq s_{\text{opt}}\} = \frac{p}{p+h}.$$

Similarly, under the approximate solution (c_*, s_*) ,

$$\mathbb{P}_{c_*}\{Q(\infty) \leq s_* - 1\} = \mathbb{P}_{c_*}\{Q(\infty) \leq s_*\} + \mathcal{O}(R^{-1/2}) \approx D_*(\beta_*, b_*) + \mathcal{O}(R^{-1/2}) = \frac{p}{p+h} + \mathcal{O}(R^{-1/2}).$$

Therefore, both the exact and approximate cost-optimal solutions deliver a stable service level, in terms of the fill rate

The previous observation also has an interesting implication for the *service degradation*, the degradation in the fill rate caused by delegation of control (see [20]). Specifically, suppose that the retailer and the manufacturer act independently. The manufacturer, as the agent, first determines the number of servers, say, c_{pa} , based on its own utility function (for example, to minimize its costs, consisting of server operating cost $d \cdot c$ and the work-in-process (WIP) cost $w \cdot \mathbb{E}_c[Q(\infty)]$). Then the retailer, as the principal, with the knowledge of the number of servers installed by the manufacturer c_{pa} , determines the base stock level s_{pa} to minimize its own costs, including inventory holding cost $h \cdot \mathbb{E}_{c_{\text{pa}}}[(s - Q(\infty))^+]$ and backorder cost $p \cdot \mathbb{E}_{c_{\text{pa}}}[(Q(\infty) - s)^+]$. Then the resulting service degradation is given by

$$\begin{aligned} \mathbb{P}_{c_{\text{opt}}}\{Q(\infty) \leq s_{\text{opt}} - 1\} - \mathbb{P}_{c_{\text{pa}}}\{Q(\infty) \leq s_{\text{pa}} - 1\} &= \mathbb{P}_{c_{\text{opt}}}\{Q(\infty) \leq s_{\text{opt}}\} - \mathbb{P}_{c_{\text{pa}}}\{Q(\infty) \leq s_{\text{pa}}\} + \mathcal{O}(R^{-1/2}) \\ &= \frac{p}{p+h} - \frac{p}{p+h} + \mathcal{O}(R^{-1/2}) = \mathcal{O}(R^{-1/2}). \end{aligned} \quad (35)$$

Hence, we conclude that the service degradation is asymptotically proportional to $R^{-1/2}$, as $R \rightarrow \infty$.

We now analyze the capacity-inventory tradeoff assuming that the system manager wishes to achieve a target fill rate of $0 \leq \delta \leq 1$. In general, for a capacity level $c = R + \beta\sqrt{R}$ and base-stock level $s = R + b\sqrt{R}$, the fill rate may be approximated by the function $D_*(\beta, b)$ given by (7) (see Theorem B.1 in the Appendix). It is therefore instructive to characterize those (β, b) such that $D_*(\beta, b) = \delta$. First note that as $\beta \rightarrow \infty$, capacity becomes unlimited and the production facility behaves as if it were an infinite server queue. Hence, letting $b_\delta(\beta)$ be such that $D_*(\beta, b_\delta(\beta)) = \delta$, one expects that b_δ should tend to a constant as β tends to ∞ . Indeed, since $\beta/(\phi(\beta) + \beta\Phi(\beta)) \rightarrow 1$ as $\beta \rightarrow \infty$, we obtain from (7) that

$$b_\delta(\beta) \rightarrow \Phi^{-1}(\delta) \quad \text{as } \beta \rightarrow \infty.$$

Next, suppose that the capacity safety factor β tends to 0. As the capacity level decreases, the system manager must increase the base-stock level in order to maintain the desired fill rate δ . Indeed, from (7) one has that for β sufficiently small

$$b_\delta(\beta) = \frac{\ln(C_*(\beta)) - \ln(1 - \delta)}{\beta} + \beta.$$

Hence using (31) and the fact that $\ln(1 + x) = x + o(x)$ for $|x| < 1$, it follows that

$$\lim_{\beta \rightarrow 0} \left(b_\delta(\beta) - \frac{-\ln(1 - \delta)}{\beta} \right) = -\frac{\sqrt{2\pi}}{2}.$$

Thus, the required inventory and capacity safety factors are approximately inversely proportional to one another when the capacity hedge is low. Figure 4 plots level curves of the function $D_*(\beta, b)$.

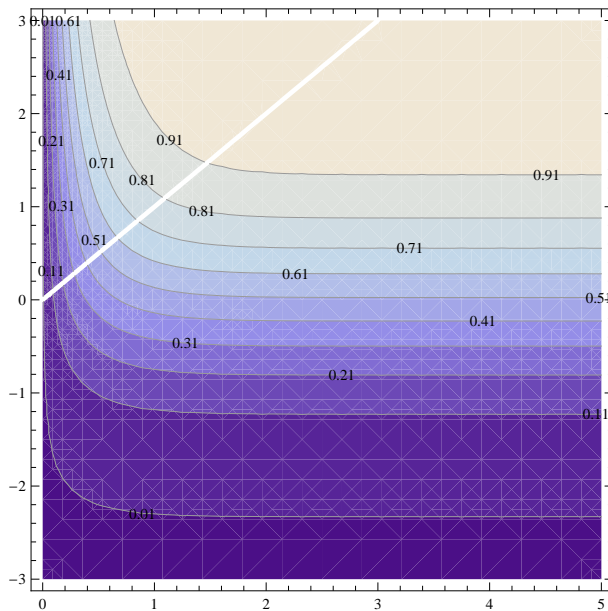


Figure 4: Contour plot of $D_*(\beta, b)$ at values from 0.01 to 0.91, for $(\beta, b) \in (0, 5) \times (-3, 3)$.

We conclude this subsection by comparing our results to the case in which the production facility is modeled as a single-server with a capacity of $c\mu$. First note that the foregoing discussion implies that in the parallel-server model, the average WIP and finished-goods inventory are on the order of R and \sqrt{R} , respectively. Thus, the distribution of inventory in the system tends to be more heavily weighted towards the production facility. On the other hand, the discussion in [6] implies that for the single-server model, the amount average amount of inventory at the production facility and the retailer are of the same order of magnitude and so in the single-server model there tends to be a more equitable distribution of inventory in the pipeline.

4 Extensions

In this section, we consider two extensions of our basic model in the previous two sections. We first consider a model in which the system manager may turn away customer orders if the backlog is too large. In this case, the system manager must determine the maximum permissible backlog. In our second extension, we allow for multiple product classes and provide a method to determine the approximate optimal safety stock level for each class.

4.1 Partially Backlogged Demand

In this subsection, we consider a model in which customer orders are rejected if the backlog reaches a certain level (see Section 2.3 in [8]) and present the square-root rule in this model. Specifically, as in our original model, customer orders arrive to the system according to a Poisson process with rate $\lambda > 0$ and are fulfilled from the finished-goods inventory which is managed under a base-stock policy with base-stock level s . The production facility consists of c servers in parallel and each order's processing time has an exponential distribution with mean $1/\mu$. At most $r \in [0, \infty)$ units of backorders are allowed and any customer order that arrives when there are r customer orders backlogged is lost.

In this case, it is straightforward to show that the shortfall process, $Z_{s+r} = \{Z_{s+r}(t), t \geq 0\}$, is identical to the number of customers process in an $M/M/c/(s+r)$ queue with arrival rate λ and service rate μ , where $s+r$ indicates the maximum number of customers allowed in the queueing system. Also, without loss of generality, we assume $c \leq s+r$, because if $c > s+r$, then $c - (s+r)$ of the servers can be idled at all times without affecting the system dynamics and so the system manager would want to simply install $s+r$ servers instead, which reduces this case to that of $c = s+r$.

Now note that the shortfall process Z_{s+r} process is simply a truncation of the birth-death process Q defined by (1) to the set $\{0, 1, \dots, s+r\}$. We therefore have (see Proposition 5.6.3 in [22]) that

$$Z_{s+r}(\infty) \stackrel{d}{=} Q(\infty) | Q(\infty) \leq s+r, \quad (36)$$

where $\stackrel{d}{=}$ means "equal in distribution to." Now denote by l the penalty cost for each unit of lost sale. The average cost objective function is then given by

$$\begin{aligned} T(c, s, r, R) &= d \cdot c + w \cdot \mathbb{E}_c[Z_{s+r}(\infty)] + h \cdot \mathbb{E}_c[(s - Z_{s+r}(\infty))^+] \\ &\quad + p \cdot \mathbb{E}_c[(Z_{s+r}(\infty) - s)^+] + l \cdot \lambda \cdot \mathbb{P}_c\{Z_{s+r}(\infty) = s+r\}, \end{aligned} \quad (37)$$

where the last term in (37) represents the long-run average penalty rate for lost sales. We further write

$$T(c, s, r, R) = (d + w) \cdot R + R^{1/2} \cdot K_T(c, s, r, R),$$

with

$$\begin{aligned} K_T(c, s, r, R) &:= d \cdot \frac{c - R}{\sqrt{R}} + w \cdot \frac{\mathbb{E}_c[Z_{s+r}(\infty)] - R}{\sqrt{R}} + h \cdot \frac{\mathbb{E}_c[(s - Z_{s+r}(\infty))^+]}{\sqrt{R}} \\ &\quad + p \cdot \frac{\mathbb{E}_c[(Z_{s+r}(\infty) - s)^+]}{\sqrt{R}} + l \cdot \frac{\lambda \cdot \mathbb{P}_c\{Z_{s+r}(\infty) = s+r\}}{\sqrt{R}}. \end{aligned} \quad (38)$$

Our goal now is to choose the 3-tuple (c, s, r) that minimizes $T(c, s, r, R)$ or, equivalently, that minimizes $K_T(c, s, r, R)$.

First note that

$$\mathbb{E}_c[(Z_{s+r}(\infty) - s)^+] = \mathbb{E}_c[Z_{s+r}(\infty)] - s + \mathbb{E}_c[(s - Z_{s+r}(\infty))^+], \quad (39)$$

and, it follows from (36) that

$$\mathbb{E}_c[(s - Z_{s+r}(\infty))^+] = \frac{\mathbb{E}_c[(s - Q(\infty))^+]}{\mathbb{P}_c\{Q(\infty) \leq s + r\}}, \quad (40)$$

$$\mathbb{E}_c[Z_{s+r}(\infty)] = s + r - \frac{\mathbb{E}_c[(s + r - Q(\infty))^+]}{\mathbb{P}_c\{Q(\infty) \leq s + r\}}, \quad (41)$$

and

$$\mathbb{P}_c\{Z_{s+r}(\infty) = s + r\} = \frac{\mathbb{P}_c\{Q(\infty) = s + r\}}{\mathbb{P}_c\{Q(\infty) \leq s + r\}}. \quad (42)$$

Because we have developed diffusion approximations for all of the terms on the right-hand sides of expressions (40)-(42), a second-order approximation for $K_T(c, s, r, R)$ can be readily obtained. Specifically, for any $(\beta, b, \nu) \in S_3 := \{(a_1, a_2, a_3) \in (0, \infty) \times (-\infty, \infty) \times [0, \infty) : a_2 + a_3 \geq a_1\}$, one may easily derive a series expansion

$$K_T(R + \beta\sqrt{R}, R + b\sqrt{R}, \nu\sqrt{R}, R) = K_{T,*}(\beta, b, \nu) + \mathcal{O}(R^{-1/2}),$$

where

$$K_{T,*}(\beta, b, \nu) = d\beta + w \cdot \mathbb{E}_\beta[\tilde{Z}_{b+\nu}(\infty)] + h \cdot \mathbb{E}_\beta[(b - \tilde{Z}_{b+\nu}(\infty))^+] + p \cdot \mathbb{E}_\beta[(\tilde{Z}_{b+\nu}(\infty) - b)^+] + l \cdot (\beta + \mathbb{E}_\beta[\max(0, -\tilde{Z}_{b+\nu}(\infty))]), \quad (43)$$

where $\tilde{Z}_{b+\nu}(\infty)$ is a random variable with distribution given by

$$\mathbb{P}_\beta\{\tilde{Z}_{b+\nu} \leq x\} = \frac{D_*(\beta, x)}{D_*(\beta, b + \nu)},$$

for $x \leq b + \nu$. The approximate optimal capacity-inventory-backlog prescription is then given by

$$(c_*, s_*, r_*) := (R + \beta_*\sqrt{R}, R + b_*\sqrt{R}, \nu_*\sqrt{R}),$$

where $(\beta_*, b_*, \nu_*) := \arg \min_{(\beta, b, \nu) \in S_3} K_{T,*}(\beta, b, \nu)$.

In order to solve for (β_*, b_*, ν_*) , note that if we fix $\beta > 0$ and further fix the sum of b and ν at a certain value $y \geq \beta$, then it can be seen from (43) that the optimal base-stock safety factor b (i.e., the minimizer of $K_{T,*}(\beta, b, y - b)$) must satisfy

$$\frac{D_*(\beta, b)}{D_*(\beta, y)} = \frac{p}{p + h}. \quad (44)$$

This optimality characterization corresponds to a critical fractile solution for the optimal base-stock level s , given c and $s + r$ for the objective function $T(c, s, r, R)$, namely,

$$\mathbb{P}_c\{Z_{s+r}(\infty) \leq s\} = \frac{p}{p+h},$$

because

$$\mathbb{P}_c\{Z_{s+r}(\infty) \leq s\} = \frac{\mathbb{P}_c\{Q(\infty) \leq s\}}{\mathbb{P}_c\{Q(\infty) \leq s+r\}}, \quad (45)$$

and the $D_*(\cdot, \cdot)$ function is the diffusion approximation for the CDF of $Q(\infty)$. Inverting the left-hand side of (44) as a function of b , we find that the optimal b with fixed β and $b + \nu = y$ is simply given by $v_*(\beta, y)$, where

$$v_*(\beta, y) := \begin{cases} \beta + \frac{1}{\beta} \ln \left[\frac{C_*(\beta) \cdot (p+h)}{p+h-pD_*(\beta, y)} \right], & \text{if } C_*(\beta) > \frac{p+h-pD_*(\beta, y)}{p+h} \\ \Phi^{-1} \left(\frac{pD_*(\beta, y)\Phi(\beta)}{(p+h)[1-C_*(\beta)]} \right), & \text{if } C_*(\beta) \leq \frac{p+h-pD_*(\beta, y)}{p+h}. \end{cases} \quad (46)$$

Therefore, instead of searching for (β_*, b_*, ν_*) in the three-dimensional set S_3 , one can first find the minimizer of the two-variable function $K_{T,*}(\beta, v_*(\beta, y), y)$, for $\beta > 0$ and $y \geq \beta$, say (β_*, y_*) , and then set $(b_*, \nu_*) := (v_*(\beta_*, y_*), y_* - v_*(\beta_*, y_*))$.

Finally, we remark that if we require that $r = 0$, then no backorders are allowed in the system and we have a pure lost sales model. In this case,

$$T(c, s, 0, R) = d \cdot c + w \cdot \mathbb{E}_c[Z_s(\infty)] + h \cdot \mathbb{E}_c[s - Z_s(\infty)] + l \cdot \lambda \cdot \mathbb{P}_c\{Z_s(\infty) = s\},$$

where $0 \leq Z_s(\infty) \leq s$. In this case, the solution procedure for the joint optimization problem specified above can be simplified since only two decision variables are involved.

4.2 Multiple Customer Classes

In this subsection, we consider the case in which there are J customer classes, where customers of class $j = 1, \dots, J$ arrive according to a Poisson process with rate λ_j . Moreover, we assume that the arrival streams of each customer class are independent of one another. We also denote by h_j and p_j the holding costs and backordering costs, respectively, for orders of class j . Each class may be thought of as representing a different product, or, it may simply represent a separate customer class with differing cost parameters. We again use c to denote the number of servers, and we use d and w to denote the capacity and WIP holding costs, respectively. Also, we let $\lambda = \lambda_1 + \dots + \lambda_J$.

The inventory level for class j is assumed to be managed according to a base-stock policy with base-stock level s_j . Moreover, production requests at the production facility are served on a first-come-first-served (FCFS) basis. This setup is similar to that considered by Benjaafar, ElHafsi and de Véricourt [4], however, we model the production facility as a parallel-server queue. Wein [27] and Zheng and Zipkin [31] consider multi-product make-to-stock systems with inventory control policies at the production facility that are not necessarily of a FCFS nature.

Let $Q_j(\infty)$ denote the steady-state number of class j orders at the production facility under the base-stock policy described above. We then have, similar to as in the previous sections, that the long-run average steady-state cost may be written as

$$\Pi(c, s, R) = d \cdot c + w \cdot \mathbb{E}_c \left[\sum_{j=1}^J Q_j(\infty) \right] + \sum_{j=1}^J h_j \cdot \mathbb{E}_c [(s_j - Q_j(\infty))^+] + p_j \cdot \mathbb{E}_c [(Q_j(\infty) - s_j)^+]. \quad (47)$$

The random variable $Q(\infty) = Q_1(\infty) + \dots + Q_J(\infty)$ represents the total number of orders at the production facility and has the same distribution as given by (2)-(4) and is therefore independent of the base-stock levels s_j .

Now note that given there are $Q_j(\infty)$ orders pending at the production facility, the distribution of the number of orders of each type of customer class is a multinomial distribution. In particular, the marginal distribution of the number of orders from customer class j is a binomial distribution with probability of success $v_j = \lambda_j/\lambda$. Therefore, conditioning on the value of $Q(\infty)$, it follows that

$$\mathbb{P}_c\{Q_j(\infty) = m | Q(\infty) = k\} = \binom{k}{m} v_j^m (1 - v_j)^{k-m}. \quad (48)$$

Now let $\rho_j = \lambda_j/(c\mu)$ and let $\rho = \lambda/(c\mu)$. Also, let $q_j = \rho_j/(\rho - \rho_j)$. By (2)-(4) and (48), one then obtains after some algebra that for $m \geq c$,

$$\mathbb{P}_c\{Q_j(\infty) = m\} = \frac{(c^c/c!)}{1 - \rho} (1 - q_j) q_j^m. \quad (49)$$

For $m \leq c$, one has that

$$\begin{aligned} \mathbb{P}_c\{Q_j(\infty) = m\} &= \sum_{k=m}^n \mathbb{P}_c\{Q_j(\infty) = m | Q(\infty) = k\} \mathbb{P}_c\{Q(\infty) = k\} \\ &+ \sum_{k=n+1}^{\infty} \mathbb{P}_c\{Q_j(\infty) = m | Q(\infty) = k\} \mathbb{P}_c\{Q(\infty) = k\}, \end{aligned} \quad (50)$$

where

$$\begin{aligned} &\sum_{k=n+1}^{\infty} \mathbb{P}_c\{Q_j(\infty) = m | Q(\infty) = k\} \mathbb{P}_c\{Q(\infty) = k\} \\ &= \eta(c^c/n!) \left(\frac{\lambda_j}{\lambda - \lambda_j} \right)^m \sum_{k=c+1}^{\infty} \binom{k}{m} (\rho - \rho_j)^k \end{aligned} \quad (51)$$

and

$$\sum_{k=m}^n \mathbb{P}_c\{Q_j(\infty) = m | Q(\infty) = k\} \mathbb{P}_c\{Q(\infty) = k\} = \eta \left(\frac{\lambda_j}{\mu} \right)^m \frac{1}{m!} \sum_{p=0}^{n-m} \left(\frac{\lambda - \lambda_j}{\mu} \right)^p \frac{1}{p!}. \quad (52)$$

The formulas given by (49)-(52) imply that the distribution of Q_j is independent of the base-stock levels $s_k, k = 1, \dots, J$, chosen. It therefore follows by (47) that each base-stock levels s_j may be managed individually. In particular, for a given capacity level c , the optimal base-stock level for customers class j is given by a critical ration solution $s_j^* = F_{c, Q_j(\infty)}^{-1}(p_j/(p_j h_j))$, where $F_{c, Q_j(\infty)}^{-1}$ denotes the inverse of the CDF of $Q_j(\infty)$. Although this solution is exact for a chosen capacity level c , it is again not computationally scalable in the demand rate. We therefore now turn towards to approximating $Q_j(\infty)$ when the demand rate is high.

Let $\{X_i^j, i \geq 1\}$ be a sequence of i.i.d. random variables, independent of $Q(\infty)$, and such that $P(X_1^j = 1) = 1 - P(X_1^j = 0) = v_j$. Also, let $R = \lambda/\mu$ denote the offered load of the system and assume that the capacity level $c = R + \beta\sqrt{R}$ for some $\beta > 0$. By (48), we have that we may write

$$Q_j(\infty) = \sum_{i=1}^{Q(\infty)} X_i^j, \quad (53)$$

which, after rearranging terms, yields

$$Q_j(\infty) = v_j R + \sum_{i=1}^R (X_i^j - v_j) + v_j(Q(\infty) - R) + \sum_{i=R}^{R+(Q(\infty)-R)} (X_i^j - v_j). \quad (54)$$

We now consider each of the four terms appearing on the righthand side of (54). We begin by noting that the first three terms on the righthand side of (54) are independent of one another. Next, by the central limit theorem,

$$\sum_{i=1}^R (X_i^j - v_j) = \sqrt{v_j(1-v_j)}\mathcal{N}(0, 1)R^{1/2} + o(R^{1/2}), \quad (55)$$

where $\mathcal{N}(0, 1)$ denotes a standard normal random variable and we use the notation $o(R^{1/2})$ to represent a random variable such that $o(R^{1/2})/R^{1/2} \Rightarrow 0$ as R tends to ∞ . Next note that by the results of the previous section,

$$v_j(Q(\infty) - R) = v_j\tilde{Q}(\infty)R^{1/2} + o(R^{1/2}), \quad (56)$$

where $\tilde{Q}(\infty)$ is a random variable with distribution given by $\mathbb{P}_\beta\{\tilde{Q}(\infty) \leq x\} = D_*(\beta, x)$, with $D_*(\beta, x)$ defined as in (7). Finally, by the central limit theorem and (56),

$$\sum_{i=R}^{R+(Q(\infty)-R)} (X_i^j - v_j) = o(R^{1/2}). \quad (57)$$

Placing (55),(56) and (57) together, one obtains by (54) that

$$Q_j(\infty) = v_j R + \left(\sqrt{v_j(1-v_j)}\mathcal{N}(0, 1) + v_j\tilde{Q}(\infty) \right) R^{1/2} + o(R^{1/2}). \quad (58)$$

Letting F_j be the CDF of the random variable $\sqrt{v_j(1-v_j)}\mathcal{N}(0,1)+v_j\tilde{Q}(\infty)$, (58) suggests that for a fixed capacity level β , one should set the base-stock level for product class j to be $s_{j,*} = R + b_{j,*}\sqrt{R}$, where $b_{j,*} = F_j^{-1}(v_j/(v_j + h_j))$. In order to determine the optimal capacity level, one may then optimize over β .

5 Conclusions

In this paper, we have analyzed the problem of jointly optimizing capacity and inventory levels for manufacturing systems where the production facility is modeled as a parallel-server queue. Our results provide scalable, square-root algorithms which provide approximate optimal solutions whose relative error is shown to diminish to zero as the arrival rate to the system grows large. Our analysis also presents important insights to the setting of the optimal capacity and inventory levels and some of the tradeoffs involved in so doing. The objective function criterion in our analysis is an average cost criterion. However, in some cases, especially when the discount factor is large, it may be more appropriate to consider a discounted cost criterion. We intend to address this problem in future work.

A Preliminary results

In the appendix, we provide the proofs of the results found in the paper. We begin with some preliminary results with regard to the analytic continuation of various performance functions. As discussed in Section 3.1, we consider the analytic continuation of each performance function, allowing c and s to take on any non-negative real value. The same approach is taken in [19] and [30] and is based on results from earlier work such as [17].

First, the analytic continuation of the steady-state shortfall distribution to real-valued arguments is given by the following lemma.

Lemma A.1. *For any $c \in \mathbb{Z}_+ \cup (R, \infty)$ and $s \in \mathbb{Z}_+$, $D(c, s, R) = \mathbb{P}_c\{Q(\infty) \leq s\}$, where $D(c, s, R)$ is defined for all $c \in (R, \infty)$ and $s \in [0, \infty)$:*

$$D(c, s, R) := \begin{cases} [1 - C(c, R)] \cdot R^{s-c+1} B(c-1, R) \Gamma(c) / [B(s, R) \Gamma(s+1)], & \text{if } s \leq c \\ [1 - C(c, R)] \cdot [1 + \rho(1 - \rho^{s-c+1}) B(c-1, R) / (1 - \rho)], & \text{if } s > c \end{cases}, \quad (59)$$

with

$$C(c, R) := \left[R \int_0^\infty t e^{-Rt} (1+t)^{c-1} dt \right]^{-1}, \quad \text{for all } c \in (R, \infty)$$

$$B(x, R) := \left[R \int_0^\infty e^{-Rt} (1+t)^x dt \right]^{-1}, \quad \text{for all } x \in [0, \infty) \quad (60)$$

It is easy to verify that $D(c, s, R)$ is continuous at $s = c$ and also $D(c, 0, R) = \mathbb{P}_c\{Q(\infty) = 0\}$.

Proof of Lemma A.1. For any integer $c > R$, the analytic continuation of the Erlang C formula reads (see [19])

$$\mathbb{P}_c\{Q(\infty) \geq c\} = C(c, R), \quad (61)$$

and therefore

$$\mathbb{P}_c\{Q(\infty) \leq c - 1\} = 1 - C(c, R). \quad (62)$$

For any $s \in \mathbb{Z}_+ \cup [0, c]$, it follows from the exact formula for the distribution of $Q(\infty)$ that

$$\frac{\mathbb{P}_c\{Q(\infty) \leq s\}}{\mathbb{P}_c\{Q(\infty) \leq c - 1\}} = \frac{\sum_{i=0}^s R^i/i!}{\sum_{i=0}^{c-1} R^i/i!} = \frac{R^{s-c+1}B(c-1, R)\Gamma(c)}{B(s, R)\Gamma(s+1)}. \quad (63)$$

Multiplying (62) by (63) yields the desired result for $s \leq c$. In the case of $s > c$, we have that

$$\frac{\mathbb{P}_c\{Q(\infty) \leq s\}}{\mathbb{P}_c\{Q(\infty) \leq c - 1\}} = 1 + \frac{\sum_{i=c}^s \mathbb{P}_c\{Q(\infty) = i\}}{\mathbb{P}_c\{Q(\infty) \leq c - 1\}} = 1 + \frac{c^c \sum_{i=c}^s \rho^i/c!}{\sum_{i=0}^{c-1} R^i/i!} = 1 + \frac{\rho(1 - \rho^{s-c+1})B(c-1, R)}{(1 - \rho)}. \quad (64)$$

which combined with (62) completes the proof. \square

We next show that a recursive relation that is known to hold for the Erlang B loss formula (see [28]) remains valid for its analytic continuation.

Lemma A.2. *For any $x > 0$,*

$$\frac{1}{B(x, R)} = \frac{R}{(x+1)B(x+1, R)} - \frac{R}{x+1}$$

Proof. The recursion follows from

$$B(x, R) = \frac{e^{-R}R^x}{\Gamma(x+1, R)}, \quad \text{for all } x > 0$$

and a recursive relation for the incomplete gamma function (see [29])

$$\Gamma(x+1, R) = x\Gamma(x, R) + R^x e^{-R}.$$

\square

Our third preliminary result is the analytic continuation for the two performance functions that determine the inventory holding cost and backorder cost.

Lemma A.3. *If $c \geq s$, $\mathbb{E}_c[(s - Q(\infty))^+] = L_1(c, s, R)$, and if $c \leq s$, $\mathbb{E}_c[(Q(\infty) - s)^+] = L_2(c, s, R)$, where*

$$L_1(c, s, R) := D(c, s, R) \cdot [s - R(1 - B(s, R))], \quad (65)$$

$$L_2(c, s, R) := [1 - C(c, R)] \cdot B(c-1, R) \cdot \rho^{s-c} \cdot \frac{\rho^2}{(1 - \rho)^2}, \quad (66)$$

and both functions are defined for all $c \in (R, \infty)$ and $s \in [0, \infty)$.

Proof. First, by examining the state transition rates in the underlying birth-death process of the total number of customers in the $M/M/c$ queue, we observe that, if $c \geq s$, $Q(\infty)|Q(\infty) \leq s$ is equal in distribution to Q_B , where Q_B denotes the steady-state number of customers in the $M/M/s/s$ loss queue with offered load R , and if $c \leq s$, $Q(\infty) - s|Q(\infty) \geq s$ equal in distribution to $Q_{M/M/1}$, where $Q_{M/M/1}$ denotes the steady-state number of customers in the $M/M/1$ queue with traffic intensity ρ .

If $c \geq s$, we condition to obtain that

$$\mathbb{E}_c[(s - Q(\infty))^+] = \mathbb{P}_c\{Q(\infty) \leq s\} \cdot (s - \mathbb{E}_c[Q(\infty)|Q(\infty) \leq s]). \quad (67)$$

Substituting the formula for the expected steady-state queue length in the $M/M/s/s$ queue

$$\mathbb{E}[Q_B] = R(1 - B(s, R)) \quad (68)$$

into (67) then yields (65).

In the case of $c \leq s$,

$$\begin{aligned} \mathbb{E}_c[(Q(\infty) - s)^+] &= \mathbb{P}_c\{Q(\infty) \geq s\} \cdot \mathbb{E}_c[Q(\infty) - s|Q(\infty) \geq s] \\ &= \mathbb{P}_c\{Q(\infty) \geq s\} \cdot \mathbb{E}[Q_{M/M/1}] \\ &= \frac{\mathbb{P}_c\{Q(\infty) = c\} \rho^{s-c}}{1 - \rho} \cdot \frac{\rho}{1 - \rho} \\ &= L_2(c, s, R), \end{aligned} \quad (69)$$

because

$$\mathbb{P}_c\{Q(\infty) = c\} = [1 - C(c, R)] \cdot B(c - 1, R)\rho,$$

due to

$$\frac{\mathbb{P}_c\{Q(\infty) = c\}}{\mathbb{P}_c\{Q(\infty) \leq c - 1\}} = \frac{R^c/c!}{\sum_{i=0}^{c-1} R^i/i!} = B(c - 1, R)\rho.$$

□

We further note that

$$\mathbb{E}_c[(Q(\infty) - s)^+] - \mathbb{E}_c[(s - Q(\infty))^+] = \mathbb{E}_c[Q(\infty)] - s.$$

Also, since

$$\mathbb{E}_c[Q(\infty)] = C(c, R) \cdot \frac{\rho}{1 - \rho} + R,$$

we then have

$$\mathbb{E}_c[(Q(\infty) - s)^+] - \mathbb{E}_c[(s - Q(\infty))^+] = C(c, R) \cdot \frac{\rho}{1 - \rho} + R - s. \quad (70)$$

Therefore, combining (70) and Lemma A.3, we can have the continuation function for $\mathbb{E}_c[(s - Q(\infty))^+]$ and $\mathbb{E}_c[(Q(\infty) - s)^+]$, regardless of whether c is greater than or less than s , and this eventually leads to the following continued expression of $\Pi(c, s, R)$ as defined in (5). For notational convenience, we still use $\Pi(\cdot, \cdot, \cdot)$ to denote this continued cost objective function.

Lemma A.4.

$$\Pi(c, s, R) = (d + w)R + R^{1/2}K(c, s, R), \quad (71)$$

where

$$K(c, s, R) = \begin{cases} d \cdot \frac{c-R}{\sqrt{R}} - p \cdot \frac{s-R}{\sqrt{R}} + \frac{(w+p)\sqrt{R}}{c-R} \cdot C(c, R) + (p+h)L_1(c, s, R) \cdot R^{-1/2}, & \text{if } s \leq c \\ d \cdot \frac{c-R}{\sqrt{R}} + h \cdot \frac{s-R}{\sqrt{R}} + \frac{(w-h)\sqrt{R}}{c-R} \cdot C(c, R) + (p+h)L_2(c, s, R) \cdot R^{-1/2}, & \text{if } s > c \end{cases}. \quad (72)$$

One may easily verify that $K(c, s, R)$ is continuous at $s = c$ and thus so is $\Pi(c, s, R)$.

Proof of Lemma A.4. The representation simply follows by applying (70) and Lemma A.3 to (5). We provide the details below.

$$\Pi(c, s, R) = \begin{cases} d \cdot c + (w+p) \cdot \mathbb{E}_c[Q(\infty)] + (p+h) \cdot L_1(c, s, R) - p \cdot s, & \text{if } s \leq c \\ d \cdot c + (w-h) \cdot \mathbb{E}_c[Q(\infty)] + (p+h) \cdot L_2(c, s, R) + h \cdot s, & \text{if } s > c \end{cases}. \quad (73)$$

If $s \leq c$,

$$\begin{aligned} \Pi(c, s, R) &= d \cdot \left(R + \frac{c-R}{\sqrt{R}} R^{1/2} \right) + (w+p) \cdot \left(C(c, R) \cdot \frac{\rho}{1-\rho} + R \right) + (p+h) \cdot L_1(c, s, R) - p \cdot \left(R + \frac{s-R}{\sqrt{R}} R^{1/2} \right) \\ &= (d+w)R + \left(d \cdot \frac{c-R}{\sqrt{R}} - p \cdot \frac{s-R}{\sqrt{R}} \right) R^{1/2} + \frac{(w+p)\sqrt{R}}{c-R} \cdot C(c, R) R^{1/2} + (p+h) \cdot L_1(c, s, R) \\ &= (d+w)R + R^{1/2} \left[d \cdot \frac{c-R}{\sqrt{R}} - p \cdot \frac{s-R}{\sqrt{R}} + \frac{(w+p)\sqrt{R}}{c-R} \cdot C(c, R) + (p+h)L_1(c, s, R) \cdot R^{-1/2} \right] \end{aligned} \quad (74)$$

and if $s > c$,

$$\begin{aligned} \Pi(c, s, R) &= d \cdot \left(R + \frac{c-R}{\sqrt{R}} R^{1/2} \right) + (w-h) \cdot \left(C(c, R) \cdot \frac{\rho}{1-\rho} + R \right) + (p+h) \cdot L_2(c, s, R) + h \cdot \left(R + \frac{s-R}{\sqrt{R}} R^{1/2} \right) \\ &= (d+w)R + \left(d \cdot \frac{c-R}{\sqrt{R}} + h \cdot \frac{s-R}{\sqrt{R}} \right) R^{1/2} + \frac{(w-h)\sqrt{R}}{c-R} \cdot C(c, R) R^{1/2} + (p+h) \cdot L_2(c, s, R) \\ &= (d+w)R + R^{1/2} \left[d \cdot \frac{c-R}{\sqrt{R}} + h \cdot \frac{s-R}{\sqrt{R}} + \frac{(w-h)\sqrt{R}}{c-R} \cdot C(c, R) + (p+h)L_2(c, s, R) \cdot R^{-1/2} \right] \end{aligned} \quad (75)$$

□

B Corrected diffusion approximation

In this section we prove Theorems 3.1 and 3.3. We start by obtaining the corrected diffusion approximation for the steady-state shortfall distribution, which is equivalent to the steady-state queue-length distribution in the elementary $M/M/c$ queue. This corrected diffusion approximation refines the celebrated QED diffusion approximation, specifically Propositions 1 and 2 in [16], and may be of independent interest. Also, note that our result is on the analytic continuation $D(c, s, R)$ (see Lemma A.1).

Theorem B.1. *For any $\beta > 0$ and $b \in (-\infty, \infty)$,*

$$D(R + \beta\sqrt{R}, R + b\sqrt{R}, R) = D_*(\beta, b) + D_\bullet(\beta, b)R^{-1/2} + \mathcal{O}(R^{-1}), \quad (76)$$

It is easy to verify from (13) and (14) that

$$g_1(\beta, \beta) = g_2(\beta, \beta) = \frac{\phi(\beta)}{\Phi(\beta)},$$

and $D_\bullet(\beta, b)$ is continuous at $b = \beta$. Also because $D_*(\beta, \beta) = 1 - C_*(\beta)$,

$$D(c, c, R) = [1 - C_*(\beta)] + [(1 - C_*(\beta))\phi(\beta)\Phi(\beta)^{-1} - C_\bullet(\beta)]R^{-1/2} + \mathcal{O}(R^{-1}). \quad (77)$$

Proof of Theorem B.1. Throughout the proof, let $c := R + \beta\sqrt{R}$ and $s := R + b\sqrt{R}$. By Theorem 2 in [19],

$$1 - C(c, R) = 1 - C_*(\beta) - C_\bullet(\beta)R^{-1/2} + \mathcal{O}(R^{-1}). \quad (78)$$

We first consider the case of $s \leq c$ or $b \leq \beta$. Due to (59) and (78), it is sufficient to prove

$$\frac{R^s/[B(s, R)\Gamma(s+1)]}{R^{c-1}/[B(c-1, R)\Gamma(c)]} = \Phi(b)\Phi(\beta)^{-1} + g_1(\beta, b)R^{-1/2} + \mathcal{O}(R^{-1}). \quad (79)$$

In [19], it is shown that

$$B(s, R)^{-1} = \frac{\Phi(\alpha_s)}{\phi(\alpha_s)}s^{1/2} + \frac{2}{3} + \mathcal{O}(s^{-1/2}), \quad (80)$$

where

$$\alpha_s = \sqrt{-2s(1 - R/s + \ln(R/s))}, \quad \text{sign}(\alpha_s) = \text{sign}(1 - R/s), \quad (81)$$

a simple function of R and s with $\alpha_s \rightarrow b$ as $s \rightarrow \infty$. By letting $p(s) := s^s e^{-s} \sqrt{2\pi s} \Gamma(s+1)^{-1}$, we have

$$\frac{e^{-R}R^s}{\Gamma(s+1)} = \phi(\alpha_s)p(s)s^{-1/2}. \quad (82)$$

Multiplying (80) by (82) yields

$$\frac{e^{-R}R^s}{\Gamma(s+1)B(s,R)} = p(s)\Phi(\alpha_s) + \frac{2}{3}\phi(\alpha_s)p(s)s^{-1/2} + \mathcal{O}(s^{-1}). \quad (83)$$

To expand the first term in (83), we note from the proof of Theorem 2 in [19] that

$$\frac{\Phi(\alpha_s)}{\phi(b)} = \frac{\Phi(b)}{\phi(b)} - \frac{1}{6}b^2R^{-1/2} + \mathcal{O}(R^{-1})$$

and thus

$$\Phi(\alpha_s) = \Phi(b) - \frac{1}{6}b^2\phi(b)R^{-1/2} + \mathcal{O}(R^{-1}). \quad (84)$$

By the Stirling's approximation for the Gamma function (see p. 257 of [1]),

$$p(s) = 1 + \mathcal{O}(s^{-1}) = 1 + \mathcal{O}(R^{-1}). \quad (85)$$

We then multiply (84) by (85) and arrive at

$$p(s)\Phi(\alpha_s) = \Phi(b) - \frac{1}{6}b^2\phi(b)R^{-1/2} + \mathcal{O}(R^{-1}). \quad (86)$$

Next, we expand the second term in (83). Simple computations show that

$$\phi(\alpha_s) = \phi(b) + \mathcal{O}(R^{-1/2})$$

and

$$s^{-1/2}p(s) = s^{-1/2} + \mathcal{O}(s^{-3/2}) = R^{-1/2} + \mathcal{O}(R^{-1}).$$

It then follows that

$$\frac{2}{3}\phi(\alpha_s)p(s)s^{-1/2} = \frac{2}{3}\phi(b)R^{-1/2} + \mathcal{O}(R^{-1}). \quad (87)$$

Substituting (86) and (87) into (83) yields

$$\frac{e^{-R}R^s}{\Gamma(s+1)B(s,R)} = \Phi(b) + \left[\frac{2}{3}\phi(b) - \frac{1}{6}b^2\phi(b) \right] R^{-1/2} + \mathcal{O}(R^{-1}). \quad (88)$$

This provides a power series expansion of the numerator in (79) times e^{-R} . We then turn to expanding the denominator of expression (79) times e^{-R} . By Lemma A.2,

$$\frac{1}{B(c-1,R)} = \frac{R}{cB(c,R)} - \frac{R}{c}$$

and therefore

$$\frac{e^{-R}R^{c-1}}{\Gamma(c)B(c-1,R)} = \frac{e^{-R}R^c}{\Gamma(c+1)B(c,R)} - \frac{e^{-R}R^c}{\Gamma(c+1)}. \quad (89)$$

The expansion of the first term of (89) is just the same as (88), with b replaced by β :

$$\frac{e^{-R}R^c}{\Gamma(c+1)B(c,R)} = \Phi(\beta) + \left[\frac{2}{3}\phi(\beta) - \frac{1}{6}\beta^2\phi(\beta) \right] R^{-1/2} + \mathcal{O}(R^{-1}). \quad (90)$$

For the second term of (89), following the same procedure as above (*i.e.*, from (82) to (87)), we obtain that

$$\frac{e^{-R}R^c}{\Gamma(c+1)} = \phi(\beta)R^{-1/2} + \mathcal{O}(R^{-1}). \quad (91)$$

Substituting (90) and (91) into (89) yields

$$\frac{e^{-R}R^{c-1}}{\Gamma(c)B(c-1,R)} = \Phi(\beta) - \left[\frac{1}{3}\phi(\beta) + \frac{1}{6}\beta^2\phi(\beta) \right] R^{-1/2} + \mathcal{O}(R^{-1}),$$

which upon inversion becomes

$$\left[\frac{e^{-R}R^{c-1}}{\Gamma(c)B(c-1,R)} \right]^{-1} = \Phi(\beta)^{-1} + \Phi(\beta)^{-2} \left[\frac{1}{3}\phi(\beta) + \frac{1}{6}\beta^2\phi(\beta) \right] R^{-1/2} + \mathcal{O}(R^{-1}). \quad (92)$$

Finally, we multiply (88) by (92) to get (79). This completes the proof for the case of $b \leq \beta$. We now turn to proving the theorem in the case of $b > \beta$. First, by Lemma A.2 and (80) (with s replaced by c),

$$B(c-1,R)^{-1} = \rho(B(c,R)^{-1} - 1) \quad (93)$$

$$= \frac{\Phi(\alpha_c)}{\phi(\alpha_c)} \cdot \rho c^{1/2} - \frac{1}{3}\rho + \mathcal{O}(c^{-1/2}), \quad (94)$$

which upon inversion becomes

$$B(c-1,R) = \frac{1}{\rho} \cdot \frac{\phi(\alpha_c)}{\Phi(\alpha_c)} \left(c^{-1/2} + \frac{1}{3} \cdot \frac{\phi(\alpha_c)}{\Phi(\alpha_c)} c^{-1} \right) + \mathcal{O}(c^{-3/2}). \quad (95)$$

Simple computations show that

$$\rho^{-1} = 1 + \beta R^{-1/2}, \quad (96)$$

$$c^{-1/2} = R^{-1/2} - \frac{1}{2}\beta R^{-1} + \mathcal{O}(R^{-3/2}), \quad (97)$$

and

$$\frac{\phi(\alpha_c)}{\Phi(\alpha_c)} = \frac{\phi(\beta)}{\Phi(\beta)} + \frac{1}{6}\beta^2 \left[\left(\frac{\phi(\beta)}{\Phi(\beta)} \right)^2 + \beta \frac{\phi(\beta)}{\Phi(\beta)} \right] R^{-1/2} + \mathcal{O}(R^{-1}). \quad (98)$$

Applying (96), (97) and (98) to (95), we have that

$$B(c-1, R) = \frac{\phi(\beta)}{\Phi(\beta)} R^{-1/2} + g_3(\beta) R^{-1} + \mathcal{O}(R^{-3/2}). \quad (99)$$

Next, we derive a refined approximation for ρ^{s-c+1} . We shall need the following result (see [7]): for any $x < -1$,

$$\left(1 + \frac{1}{x}\right)^x = e - \frac{e}{2} x^{-1} + \mathcal{O}(x^{-2}). \quad (100)$$

Also, we can express the traffic intensity as

$$\rho = 1 + \frac{1}{-(\beta^{-1} R^{1/2} + 1)} = 1 - \beta R^{-1/2} + \mathcal{O}(R^{-1}). \quad (101)$$

Applying (100) and the two expressions in (101), we have that

$$\begin{aligned} \rho^{s-c} &= \left[\left(1 + \frac{1}{-(\beta^{-1} R^{1/2} + 1)}\right)^{-(\beta^{-1} R^{1/2} + 1)} \cdot (1 - \beta R^{-1/2} + \mathcal{O}(R^{-1})) \right]^{-\beta(b-\beta)} \\ &= \left[\left(e + \frac{e}{2} (\beta^{-1} R^{1/2} + 1)^{-1} + \mathcal{O}((\beta^{-1} R^{1/2} + 1)^{-2})\right) \cdot (1 - \beta R^{-1/2} + \mathcal{O}(R^{-1})) \right]^{-\beta(b-\beta)} \\ &= \left[\left(e + \frac{e}{2} \beta R^{-1/2} + \mathcal{O}(R^{-1})\right) \cdot (1 - \beta R^{-1/2} + \mathcal{O}(R^{-1})) \right]^{-\beta(b-\beta)} \\ &= \left(e - \frac{e}{2} \beta R^{-1/2} + \mathcal{O}(R^{-1})\right)^{-\beta(b-\beta)} \\ &= \left(e^{-1} + \frac{1}{2} e^{-1} \beta R^{-1/2} + \mathcal{O}(R^{-1})\right)^{\beta(b-\beta)} \\ &= e^{-\beta(b-\beta)} \left(1 + \frac{1}{2} \beta R^{-1/2} + \mathcal{O}(R^{-1})\right)^{\beta(b-\beta)} \\ &= e^{-\beta(b-\beta)} \left(1 + \frac{1}{2} \beta^2 (b-\beta) R^{-1/2} + \mathcal{O}(R^{-1})\right). \end{aligned} \quad (102)$$

Combining (102) with (101) yields

$$\rho^{s-c+1} = e^{-\beta(b-\beta)} + \left(\frac{1}{2} \beta^2 (b-\beta) - \beta\right) e^{-\beta(b-\beta)} R^{-1/2} + \mathcal{O}(R^{-1}). \quad (103)$$

Finally, substituting (78), (99), (103) and $\rho(1-\rho)^{-1} = \beta^{-1} R^{1/2}$ into (59), we obtain the desired series expression in the case of $b > \beta$ and complete the proof of the theorem. \square

Finally, we provide the proof of Theorem 3.3 and 3.1.

Proof of Theorems 3.3 and 3.1. We first note that Theorem 3.1 immediately follows from Theorem 3.3. We next focus on proving Theorem 3.3. Throughout the proof, let $c := R + \beta\sqrt{R}$ and $s := R + b\sqrt{R}$. Due to Lemma A.4, the desired result is equivalent to

$$K(c, s, R) = K_*(\beta, b) + K_\bullet(\beta, b)R^{-1/2} + \mathcal{O}(R^{-1}). \quad (104)$$

First, [19] show that

$$C(c, R) = C_*(\beta) + C_\bullet(\beta)R^{-1/2} + \mathcal{O}(R^{-1}). \quad (105)$$

Inverting (80)

$$B(s, R) = \frac{\phi(\alpha_s)}{\Phi(\alpha_s)}s^{-1/2} - \frac{2}{3} \left(\frac{\phi(\alpha_s)}{\Phi(\alpha_s)} \right)^2 s^{-1} + \mathcal{O}(s^{-3/2}). \quad (106)$$

Applying (97) and (98) (with c replaced by s) to (106), we have that

$$B(s, R) = \frac{\phi(b)}{\Phi(b)}R^{-1/2} + g_4(b)R^{-1} + \mathcal{O}(R^{-3/2}), \quad (107)$$

where

$$g_4(b) = \frac{1}{6}b^2 \left(\frac{\phi(b)}{\Phi(b)} \right)^2 + \frac{1}{6}b^3 \frac{\phi(b)}{\Phi(b)} - \frac{1}{2}b \frac{\phi(b)}{\Phi(b)} - \frac{2}{3} \left(\frac{\phi(b)}{\Phi(b)} \right)^2.$$

From (65) and (66), we have that

$$L_1(c, s, R) = D(c, s, R) \cdot [bR^{1/2} + R \cdot B(s, R)] \quad (108)$$

$$= D_*(\beta, b) \left[b + \frac{\phi(b)}{\Phi(b)} \right] R^{1/2} + \left[D_*(\beta, b)g_4(b) + D_\bullet(\beta, b) \left(b + \frac{\phi(b)}{\Phi(b)} \right) \right] + \mathcal{O}(R^{-1/2}) \quad (109)$$

and

$$L_2(c, s, R) = [1 - C_*(\beta)] \frac{\phi(\beta)}{\beta^2 \Phi(\beta)} \cdot e^{-\beta(b-\beta)} R^{1/2} + g_5(\beta, b) + \mathcal{O}(R^{-1/2}), \quad (110)$$

where

$$g_5(\beta, b) = \beta^{-2} e^{-\beta(b-\beta)} \left[[1 - C_*(\beta)] \cdot g_3(\beta) - C_\bullet(\beta) \frac{\phi(\beta)}{\Phi(\beta)} + \frac{1}{2} [1 - C_*(\beta)] \cdot \frac{\phi(\beta)}{\Phi(\beta)} \beta^2 (b - \beta) \right].$$

Finally, applying (105), (109), and (110) to (74) and (75) leads to (104). \square

Using the expression of $C_*(\beta)$ and the fact that $D_*(\beta, \beta) = 1 - C_*(\beta)$, we can also verify the continuity of $K_*(\beta, b)$ at $b = \beta$.

C Refined square-root capacity-inventory prescriptions

This section is devoted to the proof of Theorems 3.2 and 3.4.

Proof of Theorems 3.2 and 3.4. Define (β_R, b_R) as the solution to

$$\begin{cases} D_*(\beta_R, b_R) + D_\bullet(\beta_R, b_R)R^{-1/2} = \frac{p}{p+h} \\ \frac{\partial K_*(\beta_R, b_R)}{\partial \beta} + \frac{\partial K_\bullet(\beta_R, b_R)}{\partial \beta} R^{-1/2} = 0 \end{cases}. \quad (111)$$

Let $g(R) := (g_1(R), g_2(R)) = (\beta_R, b_R) - (\beta_*, b_*)$, and then (111) can be rewritten as

$$\begin{cases} D_*(\beta_* + g_1(R), b_* + g_2(R)) + D_\bullet(\beta_* + g_1(R), b_* + g_2(R))R^{-1/2} = \frac{p}{p+h} \\ \frac{\partial K_*(\beta_* + g_1(R), b_* + g_2(R))}{\partial \beta} + \frac{\partial K_\bullet(\beta_* + g_1(R), b_* + g_2(R))}{\partial \beta} R^{-1/2} = 0 \end{cases}. \quad (112)$$

It follows from the expressions for $D_*(\cdot, \cdot)$, $D_\bullet(\cdot, \cdot)$, $K_*(\cdot, \cdot)$, and $K_\bullet(\cdot, \cdot)$ that their first, second, and third (partial) derivative functions are all continuous and thus bounded in a small neighborhood of (β_*, b_*) . Therefore, a first-order Taylor expansion of (112) yields

$$\begin{cases} D_*(\beta_*, b_*) + \mathcal{O}(g_1(R) \vee g_2(R)) + D_\bullet(\beta_*, b_*)R^{-1/2} + \mathcal{O}([g_1(R) \vee g_2(R)]R^{-1/2}) = \frac{p}{p+h} \\ \frac{\partial K_*(\beta_*, b_*)}{\partial \beta} + \mathcal{O}(g_1(R) \vee g_2(R)) + \frac{\partial K_\bullet(\beta_*, b_*)}{\partial \beta} R^{-1/2} + \mathcal{O}([g_1(R) \vee g_2(R)]R^{-1/2}) = 0 \end{cases}. \quad (113)$$

Due to steps 1 and 2 in Algorithm 1 and the fact that $D_*(\beta, w_*(\beta)) = \frac{p}{p+h}$ for any $\beta > 0$, we have $D_*(\beta_*, b_*) = \frac{p}{p+h}$ and $\frac{\partial K_*(\beta_*, b_*)}{\partial \beta} = 0$. Therefore, it immediately follows that

$$g_1(R) \vee g_2(R) = \mathcal{O}(R^{-1/2}). \quad (114)$$

Then, we apply a second-order Taylor expansion to (112) to have that

$$\begin{aligned} D_*(\beta_*, b_*) + \frac{\partial D_*(\beta_*, b_*)}{\partial \beta} g_1(R) + \frac{\partial D_*(\beta_*, b_*)}{\partial b} g_2(R) + \mathcal{O}(g_1^2(R) \vee g_2^2(R) \vee g_1(R)g_2(R)) \\ + D_\bullet(\beta_*, b_*)R^{-1/2} + \mathcal{O}([g_1(R) \vee g_2(R)]R^{-1/2}) = \frac{p}{p+h}, \end{aligned} \quad (115)$$

and

$$\begin{aligned} \frac{\partial K_*(\beta_*, b_*)}{\partial \beta} + \frac{\partial^2 K_*(\beta_*, b_*)}{\partial \beta^2} g_1(R) + \frac{\partial^2 K_*(\beta_*, b_*)}{\partial \beta \partial b} g_2(R) + \mathcal{O}(g_1^2(R) \vee g_2^2(R) \vee g_1(R)g_2(R)) \\ + \frac{\partial K_\bullet(\beta_*, b_*)}{\partial \beta} R^{-1/2} + \mathcal{O}([g_1(R) \vee g_2(R)]R^{-1/2}) = 0. \end{aligned} \quad (116)$$

Using (114), $D_*(\beta_*, b_*) = \frac{p}{p+h}$ and $\frac{\partial K_*(\beta_*, b_*)}{\partial \beta} = 0$, we solve (115) and (116), respectively, and obtain that

$$g_1(R) = \left[D_\bullet(\beta_*, b_*) \frac{\partial^2 K_*(\beta_*, b_*)}{\partial \beta \partial b} - \frac{\partial D_*(\beta_*, b_*)}{\partial b} \frac{\partial K_\bullet(\beta_*, b_*)}{\partial \beta} \right] J^{-1} R^{-1/2} + \mathcal{O}(R^{-1}), \quad (117)$$

$$g_2(R) = \left[\frac{\partial D_*(\beta_*, b_*)}{\partial \beta} \frac{\partial K_\bullet(\beta_*, b_*)}{\partial \beta} - \frac{\partial^2 K_*(\beta_*, b_*)}{\partial \beta^2} D_\bullet(\beta_*, b_*) \right] J^{-1} R^{-1/2} + \mathcal{O}(R^{-1}), \quad (118)$$

with

$$J := \frac{\partial D_*(\beta_*, b_*)}{\partial b} \frac{\partial^2 K_*(\beta_*, b_*)}{\partial \beta^2} - \frac{\partial D_*(\beta_*, b_*)}{\partial \beta} \frac{\partial^2 K_*(\beta_*, b_*)}{\partial \beta \partial b}.$$

Therefore, (β_R, b_R) is well approximated by $(\beta_* + \beta_\bullet R^{-1/2}, b_* + b_\bullet R^{-1/2})$, up to $\mathcal{O}(R^{-1})$.

We next turn to proving the optimality gap results in (10) and (27). Let $\beta_{\text{opt}} = (c_{\text{opt}} - R)R^{-1/2}$ and $b_{\text{opt}} = (s_{\text{opt}} - R)R^{-1/2}$. The desired result is equivalent to

$$(\beta_{\text{opt}}, b_{\text{opt}}) - (\beta_*, b_*) = \mathcal{O}(R^{-1/2}), \quad (119)$$

$$(\beta_{\text{opt}}, b_{\text{opt}}) - (\beta_* + \beta_\bullet R^{-1/2}, b_* + b_\bullet R^{-1/2}) = \mathcal{O}(R^{-1}). \quad (120)$$

We first note that by Lemma A.4, the characterization (26) is equivalent to

$$\left(\frac{\partial K(c_{\text{opt}}, s_{\text{opt}}, R)}{\partial c}, \frac{\partial K(c_{\text{opt}}, s_{\text{opt}}, R)}{\partial s} \right) = (0, 0), \quad (121)$$

in which the second component is equivalent to

$$D(c_{\text{opt}}, s_{\text{opt}}, R) = p/(p+h). \quad (122)$$

From (121), (122), (104), and Theorem B.1, we have that

$$\frac{p}{p+h} = D(R + \beta_{\text{opt}}\sqrt{R}, R + b_{\text{opt}}\sqrt{R}, R) = D_*(\beta_{\text{opt}}, b_{\text{opt}}) + \mathcal{O}(R^{-1/2}), \quad (123)$$

$$0 = \frac{\partial K(R + \beta_{\text{opt}}\sqrt{R}, R + b_{\text{opt}}\sqrt{R}, R)}{\partial \beta} = \frac{\partial K_*(\beta_{\text{opt}}, b_{\text{opt}})}{\partial \beta} + \mathcal{O}(R^{-1/2}). \quad (124)$$

Let $g_{1,*}(R) := \beta_{\text{opt}} - \beta_*$ and $g_{2,*}(R) := b_{\text{opt}} - b_*$. Then applying a first-order Taylor expansion to (123), we obtain that

$$\frac{p}{p+h} = D_*(\beta_*, b_*) + \mathcal{O}(g_{1,*}(R) \vee g_{2,*}(R)) + \mathcal{O}(R^{-1/2}).$$

Since $D_*(\beta_*, b_*) = \frac{p}{p+h}$, $g_{1,*}(R) \vee g_{2,*}(R) = \mathcal{O}(R^{-1/2})$, and thus (119) holds. We next prove (120). First, it follows from the derivation of $(\beta_\bullet, b_\bullet)$ that

$$(\beta_R, b_R) - (\beta_* + \beta_\bullet R^{-1/2}, b_* + b_\bullet R^{-1/2}) = \mathcal{O}(R^{-1}).$$

Therefore, in order to conclude (120), it suffices to prove that

$$(\beta_{\text{opt}}, b_{\text{opt}}) - (\beta_R, b_R) = \mathcal{O}(R^{-1}). \quad (125)$$

Let $g_{1,\bullet}(R) := \beta_{\text{opt}} - \beta_R$ and $g_{2,\bullet}(R) := b_{\text{opt}} - b_R$. The rest of the proof is similar as above:

$$\begin{aligned} \frac{p}{p+h} &= D(R + \beta_{\text{opt}}\sqrt{R}, R + b_{\text{opt}}\sqrt{R}, R) = D_*(\beta_{\text{opt}}, b_{\text{opt}}) + D_\bullet(\beta_{\text{opt}}, b_{\text{opt}})R^{-1/2} + \mathcal{O}(R^{-1}) \quad (126) \\ &= D_*(\beta_R, b_R) + \mathcal{O}(g_{1,\bullet}(R) \vee g_{2,\bullet}(R)) + D_\bullet(\beta_R, b_R)R^{-1/2} + \mathcal{O}([g_{1,\bullet}(R) \vee g_{2,\bullet}(R)]R^{-1/2}) + \mathcal{O}(R^{-1}). \quad (127) \end{aligned}$$

Since $D_*(\beta_R, b_R) + D_\bullet(\beta_R, b_R)R^{-1/2} = p/(p+h)$ by (111), we find that $\mathcal{O}(g_{1,\bullet}(R) \vee g_{2,\bullet}(R)) = \mathcal{O}(R^{-1})$, which proves the assertion in (125). \square

References

- [1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, 1964.
- [2] B. Ata and M. Rubino. Dynamic control of a make-to-order parallel server system with cancellations, 2008. Forthcoming in. *Operations Research*.
- [3] A. Bassamboo, R. S. Randhawa, and A. Zeevi. Capacity sizing under parameter uncertainty: Safety staffing principles revisited. *Manage. Sci.*, 56:1668–1686, October 2010.
- [4] S. Benjaafar, M. ElHafsi, and F. de Véricourt. Demand allocation in multiple-product, multiple-facility, make-to-stock systems. *Manage. Sci.*, 50:1431–1448, October 2004.
- [5] S. Borst, A. Mandelbaum, and M. Reiman. Dimensioning large call centers. *Oper. Res.*, 52:17–34, 2004.
- [6] J.R. Bradley and P.W. Glynn. Managing capacity and inventory jointly in manufacturing systems. *Manage. Sci.*, 48(2):273–288, 2002.
- [7] H. J. Brothers and J. A. Knox. New closed-form approximations to the logarithmic constant e . *Math. Intell.*, 20(4):25–29, 1998.
- [8] R. Caldentey. *Analyzing the Make-to-Stock Queue in Supply Chain and e-Business Settings*. Ph.D. Thesis, 2001.
- [9] A.J. Clark and H. Scarf. Optimal policies for a multi-echelon inventory problem. *Management science*, 6(4):475–490, 1960.
- [10] R. Durrett. *Probability: Theory and examples*. *Online book*, 2010.

- [11] A. Federgruen and P. Zipkin. An inventory model with limited production capacity and uncertain demands I. The average-cost criterion. *Mathematics of Operations Research*, 11(2):193–207, 1986.
- [12] A. Federgruen and P. Zipkin. An inventory model with limited production capacity and uncertain demands II. The discounted-cost criterion. *Mathematics of Operations Research*, 11(2):208–215, 1986.
- [13] P. Glasserman. Bounds and asymptotics for planning critical safety stocks. *Operations Research*, 45(2):244–257, 1997.
- [14] P. Glasserman and T.W. Liu. Corrected diffusion approximations for a multistage production-inventory system. *Mathematics of Operations Research*, 22(1):186–201, 1997.
- [15] D. Gross and C. M. Harris. *Fundamentals of Queueing Theory (3rd Edition)*. Wiley-Interscience, 1998.
- [16] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers. *Oper. Res.*, 29:567–588, 1981.
- [17] A.A. Jagers and E.A. van Doorn. On the continued erlang loss function. *Operations Research Letters*, 5(1):43–46, 1986.
- [18] A.J.E.M. Janssen, J.S.H. van Leeuwen, and B. Zwart. Gaussian expansions and bounds for the Poisson distribution applied to the Erlang B formula. *Adv. in Appl. Probab.*, 40(1):122–143, 2008.
- [19] A.J.E.M. Janssen, J.S.H. van Leeuwen, and B. Zwart. Refining square root safety staffing by expanding Erlang C. *To appear in Oper. Res.*, 2008.
- [20] E.L. Plambeck and S.A. Zenios. Incentive efficient control of a make-to-stock production system. *Oper. Res.*, 51:371–386, May 2003.
- [21] J. Reed and T. Tezcan. Hazard rate scaling for the $gi/m/n_{gi}$ queue. *Operations Research*.
- [22] S.M. Ross. *Stochastic processes*. Wiley New York, 1996.
- [23] R. Rubio and L.M. Wein. Setting base stock levels using product-form queueing networks. *Management Science*, pages 259–268, 1996.
- [24] R. Spira. Calculation of the Gamma function by Stirling’s formula. *Math. Comp.*, 25(114):317–322, 1971.
- [25] S.R. Tayur. Computing the optimal policy for capacitated inventory models. *Stochastic Models*, 9(4):585–598, 1993.

- [26] H. C. Tijms. *Stochastic Models: An Algorithmic Approach*. John Wiley & Sons, 1995.
- [27] L. M. Wein. Dynamic scheduling of a multiclass make-to-stock queue. *Oper. Res.*, 40:724–735, July 1992.
- [28] W. Whitt. The Erlang B and C formulas: Problems and solutions. *Class notes*, 2002.
- [29] Wikipedia. Incomplete gamma function — Wikipedia, the free encyclopedia, 2010. [Online; accessed 10-Feb-2010].
- [30] B. Zhang, J.S.H. van Leeuwen, and B. Zwart. Staffing call centers with impatient customers: refinements to many-server asymptotics. *Operations Research*, 2011. forthcoming.
- [31] Y.S. Zheng and P. Zipkin. A queueing model to analyze the value of centralized inventory information. *Operations Research*, 38(2):296–307, 1990.