Published online ahead of print January 28, 2013

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Articles in Advance, pp. 1–4
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs

http://dx.doi.org/10.1287/msom.1120.0428
© 2013 INFORMS

Copyright: INFORMS holds copyright to this *Articles in Advance* version, which is made available to subscribers. The file may not be posted on any other website, including the author's site. Please send any questions regarding this policy to permissions@informs.org.

OM Forum

Offered Load Analysis for Staffing

Ward Whitt
Department of Industrial Engineering and Operations Research, Columbia University,
New York, New York 10027, ww2040@columbia.edu

This essay, based on my 2012 MSOM Fellow Lecture, discusses an idea that has been useful for developing effective methods to set staffing levels in service systems: offered load analysis. The main idea is to tackle a hard problem by first seeking an insightful simplification. For capacity planning to meet uncertain exogenous demand, offered load analysis looks at the amount of capacity that would be used if there were no constraints on its availability. This simplification is helpful because the stochastic model becomes much more tractable. Offered load analysis can be especially helpful when the demand is not only uncertain but also time varying, as in many service systems. Given the distribution of the stochastic offered load, we often can set staffing levels to stabilize performance at target levels, even in face of a strongly time-varying arrival rate, long service times, and network structure.

Key words: offered load analysis; capacity planning; server staffing; time-varying arrival rates; infinite-server queues
History: Received: July 12, 2012; accepted: December 1, 2012. Published online in *Articles in Advance*.

1. Introduction

I thank my colleagues in the MSOM Society for appreciating my research and for providing this opportunity for reflection. I am pleased to belong to a society dedicated to the discovery of principles and methods to produce goods and services more efficiently, and to doing so with integrity. Meeting those goals necessarily requires an understanding of both theory and practice. The beauty of mathematics led me to choose that undergraduate major, while the value of its application led me to choose, first, to study for a doctorate in the relatively new field of operations research and, second, to leave academia before a tenure decision to join Bell Laboratories. For me, those opportunities and choices were crucial. Somewhat ironically, those unconventional choices, moving from theory toward practice, in the end brought me closer to theory, helping me to spend a lifetime doing research. The increased exposure to practice enhanced the impact of the theory.

Early on, I learned that it was surprisingly difficult for me to do tasks when I was unmotivated, but that I could surprise myself when my interest was aroused. I think that we can safely conduct research, guided mostly by what fascinates us, if we develop good taste in research, but that requires time and effort. Good taste in research emerges gradually from studying broadly and deeply, listening to good speakers, and reading good papers. The profession plays a key role through its conferences and journals.

It is important to understand the challenges being faced in current practice, but it is also important to understand how to see the issues in their essential form, for which mathematics often plays a central role. For research, it is also important to know what others have done before. Practice points us toward the problems of greatest concern and invites us to honestly address these pressing problems, but another important role of research is to properly place new ideas in perspective, exposing the connections with all that has been done before. Hopp (2012) eloquently makes the case for practice as a source for problems, but in Hopp and Spearman (2004) he also demonstrates the importance of careful scholarship.

Cachon (2012) has done remarkably well in directly answering the question: What is interesting research in operations management? I will try to complement that by showing how my research on offered load (OL) analysis for staffing fits his template: *What is thought to be X is really Y*.

Here is the case for offered load analysis: *What is thought to be trivial and useless* (*because it seems to ignore the main problem*) *is really somewhat subtle and useful*. Here is the case for staffing a service system in face of time-varying demand: *What is thought to be complex and beyond analysis is really manageable, if not actually simple.* And even better is the combination: What is thought to be complex (staffing a service system in face of time-varying demand) is really manageable, if

Copyright: INFORMS holds copyright to this *Articles in Advance* version, which is made available to subscribers. The file may not be posted on any other website, including the author's site. Please send any questions regarding this policy to permissions@informs.org.

not actually simple, by applying that which is thought to be trivial and useless (offered load analysis).

This short overview aims to support the claims above, but not to provide a thorough review. Indeed, the main ideas presented in §§2–4 can be considered a subset of Jennings et al. (1996), which in turn follows a long tradition in teletraffic engineering, as in Palm (1943) and Jagerman (1975). Reviews appear in Green et al. (2007) and Massey (2002). Even though the main ideas are not new, I remain interested in pushing these ideas forward, as can be seen from Liu and Whitt (2012). To give some idea about the exciting problems that remain, I briefly describe the current state of the art in the e-companion (available at http://dx.doi.org/10.1287/msom.1120.0428) by giving a brief overview of recent research.

Much of my work on offered load analysis has been done with my former Bell Laboratories colleague Bill Massey, now of Princeton University, but other colleagues have contributed as well, including my students since I returned to academia. I thank all of these colleagues for their important contributions and wish them success in their future research.

2. Offered Load Analysis

In capacity planning (resource allocation) to meet uncertain exogenous demand, *offered load analysis* estimates the required capacity by estimating how much capacity would be used if there were no limit on its availability. Given that there is uncertainty, we model that uncertainty and distinguish between the *stochastic offered load* (SOL) and its expected value, simply called the *offered load*.

A key assumption underlying offered load analysis is that the demand is indeed *exogenous*, that is, that the level of demand is given and independent of the capacity being provided, or at least approximately so. First, that assumption ignores the important insight of revenue management that it can be important, even essential, to actively manage the demand. That might be done through careful management of the product offered, as in assortment planning, or it might be through pricing, and often through both. In focusing on offered load analysis, we assume that step has been adequately addressed, and that indeed the demand can be regarded as exogenous.

Operations research has traditionally focused on capacity planning in the face of constraints, because resources are inevitably limited. That is the perspective of stochastic models as well as mathematical programming. For example, queueing theory is largely concerned with waiting, blocking, and reneging due to limited capacity. We emphasize the value of more elementary queueing models without these traditional features.

A major issue in the application of queueing models is actually determining the level of demand. Queueing theory is primarily concerned with the behavior of the models in the face of known demand. However, in practice there usually is uncertainty about the demand, so it must be estimated. For existing systems, it is thus essential to have appropriate system data. Then statistical forecasting methods can be applied and tested. It is also important to be familiar with the systems to understand and appreciate special features.

Even though much is needed besides queueing models in a successful application, the models nevertheless can play an important role, because they provide a useful abstract representation of the system, which can help clarify thinking. We emphasize the important role models can play in properly understanding the offered load.

First, stochastic models and their analysis can not only explain why it is natural to model an arrival process of demand requests as a Poisson process, but also when that might be inappropriate. For example, in telecommunication systems, requests for service that cannot gain access on a first-choice path often are offered alternative paths. Similarly, customers seeking a hotel room who cannot get a reservation at one hotel are often offered reservations at alternative hotels. Thus the demand for capacity at each facility (link in a communication network or hotel in a city) includes overflows from demand originally offered to other facilities. Because these overflows only occur when the initial facility is full, they tend to occur in clumps, and thus make the arrival process of requests at each facility more "bursty" than direct arrival processes, which are often well modeled as Poisson processes.

3. Staffing in a Service System

One concrete setting for offered load analysis is staffing in a service system. Service systems often have complicated network structure (e.g., they may be distributed service centers, each with multiple pools of service representatives, serving multiple classes of customers), but we will consider the basic case of a single facility with a single pool of agents serving a single class of customers. Then the capacity is simply the number of servers.

A standard model for a basic service system is a multiserver queue. Then the demand consists of service requests with random duration (the service times) arriving at random times (the arrival process). The basic model is the $M/GI/s+GI$ queue, which has a Poisson arrival process (the M) with arrival rate λ, independent and identically distributed (i.i.d.) service times with some general distribution (the first GI) having mean $E[S]$, s homogeneous servers working

Copyright: INFORMS holds copyright to this *Articles in Advance* version, which is made available to subscribers. The file may not be posted on any other website, including the author's site. Please send any questions regarding this policy to permissions@informs.org.

independently in parallel, unlimited waiting space, the first-come, first-served service discipline, and customer abandonment from the queue with i.i.d. times to abandon, also according to a general distribution (the $+GI$).

In this setting, it is common to think of the demand as the arrival process, but the demand should be regarded as the arrival process together with the service times. Indeed, we contend that the demand should be represented by the SOL, the steady-state number of busy servers in an associated $M/GI/\infty$ infinite-server (IS) model with the same arrival process and service times. By Little's law applied to that IS model, the OL is simply $m \equiv \lambda E[S]$.

To determine an appropriate level of staffing, the stochastic OL helps greatly, because the steady-state number of customers $N(m)$ in the $M/GI/s+GI$ model with OL m has a complicated distribution (in fact, one not yet known in full generality), whereas the SOL itself is a Poisson random variable, with variance equal to the mean m. Because a Poisson distribution is approximately normal, provided that the mean is not too small, offered load analysis leads to the approximation $N(m) \approx m + \sqrt{m}N(0,1)$, where $N(0,1)$ is a standard normal random variable. If we seek the minimum staffing level s subject to the constraint $P(N(m) \geq s) \leq \alpha$ for a target α, then the offered load approach leads to the classical *square root staffing* (SRS) formula:

$$s = m + \beta\sqrt{m}, \tag{1}$$

where β is a quality-of-service (QoS) parameter, satisfying $P(N(0,1) > \beta) = \alpha$. (We round to an integer, typically using the least integer greater than or equal to $s(t)$ in (1).) The constraint on the probability that all servers are busy, which coincides with the steady-state delay probability of an arrival, is convenient, but it is also possible to consider alternative constraints by further analysis.

4. Time-Varying Arrival Rates

Offered load analysis becomes even more useful when we try to address additional complexity commonly found in practice, in particular, the fact that the arrival rate of service requests in a service system usually varies significantly over time. When service takes place during a single day, the arrival rate typically increases at the beginning of a day and decreases at the end of the day. Because staffing primarily involves service representatives (people), it is usually relatively flexible; that is, it too can be made time-varying in response to the time-varying demand.

The natural generalization of the previous queueing model to cover time-varying arrival rates is the $M_t/GI/s_t + GI$ model, where the arrival rate $\lambda(t)$ is now time varying. Paralleling the previous section, now the goal is to select the minimum staffing function $s(t)$ such that $P(X(t) \geq s(t)) \leq \alpha$ for all t, where $X(t)$ is the random number in the system at time t. Ideally, we would like to achieve the stable performance $P(X(t) \geq s(t)) \approx \alpha$ for all t. A key assumption is that the staffing is suitably flexible; we discuss what happens if it is not in §6 of the e-companion.

Even though the $M_t/GI/s_t + GI$ model with time-dependent parameters is even more complicated, the associated $M_t/GI/\infty$ IS model remains remarkably tractable. (The story here is available in elementary textbooks, e.g., §5.3.5 of Ross (2010), but it is much less common knowledge than §3 here.) Indeed, the SOL $X_\infty(t)$ again has a Poisson distribution for each t with a tractable mean,

$$\begin{aligned} m(t) \equiv E[X_\infty(t)] &= \int_{-\infty}^{t} P(S > s)\lambda(t-s)\,ds \\ &= E[\lambda(t - S_e)]E[S], \end{aligned} \tag{2}$$

where the S and S_e are random variables with the fixed service-time cumulative distribution function (cdf) and the associated stationary-excess cdf, that is, $P(S_e \leq x) \equiv (1/E[S])\int_0^x P(S > u)\,du$ with $E[S_e^k] = E[S^{k+1}]/(k+1)E[S]$. Reasoning exactly as in §3, we can use the time-varying SRS formula

$$s(t) = m(t) + \beta\sqrt{m(t)}, \quad t \geq 0, \tag{3}$$

where β again is the QoS parameter.

The final expression in Equation (2) shows that the time-varying OL $m(t)$ has the same form as the stationary OL $m \equiv \lambda E[S]$ in §3 except for the random time lag S_e in $\lambda(t)$. If the arrival rate function changes relatively slowly compared to the random variable S_e, which tends to occur for given arrival rate functions when the mean service time $E[S]$ is relatively short, then the random time lag can be ignored, and we obtain the *pointwise stationary approximation* (PSA),

$$m_{\mathrm{PSA}}(t) \equiv \lambda(t)E[S], \tag{4}$$

which is the basis for effective traditional staffing methods in call centers. However, the PSA fails with longer service times.

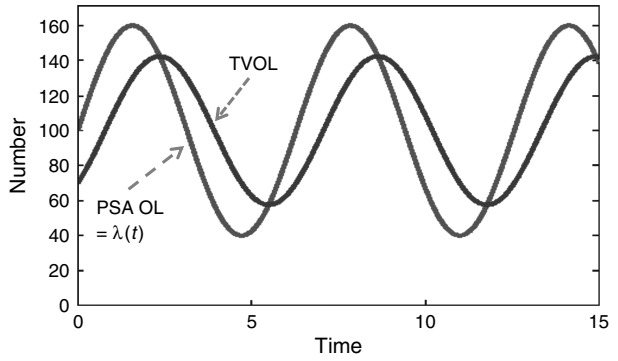
Example 1 (A Sinusoidal Example with Longer Service Times). To illustrate the advantage of the OL in (2) with longer service times, suppose that we consider the sinusoidal arrival rate function

$$\begin{aligned} \lambda(t) \equiv \bar{\lambda}(1 + \nu\sin(\gamma t)) \quad &\text{for all } t, \\ &\text{where } (\bar{\lambda}, \nu, \gamma) \equiv (100, 0.60, 1), \end{aligned} \tag{5}$$

with service times exponentially distributed, so that S_e is distributed as S, with mean service time $E[S] = 1$. If we measure time in hours, a full sine cycle is

Copyright: INFORMS holds copyright to this *Articles in Advance* version, which is made available to subscribers. The file may not be posted on any other website, including the author's site. Please send any questions regarding this policy to permissions@informs.org.

Figure 1 **Time-Varying OL in (2) and (6) Compared to the PSA Approximate OL in (4), Which Coincides with the Sinusoidal Arrival Rate Function $\lambda(t)$ in (5)**



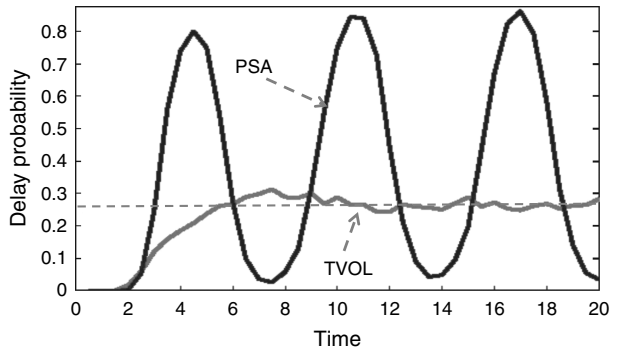
$2\pi \approx 6.3$ hours; if we think of 24-hour cycles, then a mean service time is $E[S] = 24/6.3 \approx 3.8$ hours. Even longer service times are common in healthcare; see Yom-Tov and Mandelbaum (2010).

For the sinusoidal arrival rate function in (5), the OL is

$$\begin{aligned} m(t) &= \bar{\lambda}E[S]\big(1+\nu\big(\sin(\gamma t)E[\cos(\gamma S_e)] \\ &\qquad -\cos(\gamma t)E[\sin(\gamma S_e)]\big)\big), \\ &= \bar{\lambda}E[S]\left(1+\frac{\nu}{1+\gamma^2}(\sin(\gamma t)-\gamma\cos(\gamma t))\right) \\ &\qquad \text{for} \quad S_e \stackrel{\mathrm{d}}{=} S, \\ &= 100+30(\sin(t)-\cos(t)) \\ &\qquad \text{for} \quad (\bar{\lambda},\nu,\gamma,E[S]) = (100,0.60,1.0,1.0). \quad (6) \end{aligned}$$

Figure 1 compares the OL in (2) and (6) to the PSA approximation in (4), which coincides with the arrival rate itself in (5). Figure 2 shows the implications for SRS staffing in (3) with $\beta \equiv 1$ using these two methods. The time-varying delay probability is stabilized in dynamic steady state (after the initial transient) using the OL, but not with PSA.

Figure 2 **Simulation Estimates of the Delay Probability When Staffing According to the SRS Formula in (3) with $\beta = 1$ Based on the OL and PSA Approximation in Figure 1**



Example 1 shows that it is important to understand the OL $m(t)$ in (2) and its relation to the PSA approximation $m_{\mathrm{PSA}}(t)$ in (4). Careful study has revealed many insights; for example, a Taylor series expansion shows that the main effect is captured by a deterministic time lag and space shift.

5. Conclusions

We have discussed offered load analysis for staffing because we think, first, that the benefits to operations of what has been learned largely remain to be realized in practice and, second, that there remain many exciting possibilities for more research along these lines. We think that some of the ideas belong in elementary textbooks, whereas others invite deeper analysis.

Hopefully, these methods will have more applications in the future, enabling us to produce goods and services more efficiently. Hopefully, the profession will continue to develop new and better methods that will benefit society.

Electronic Companion

An electronic companion to this paper is available as part of the online version at http://dx.doi.org/10.1287/msom.1120.0428.

Acknowledgments

The author's research has been supported by the National Science Foundation, most recently through Grant CMMI-1066372.

References

Cachon GP (2012) What is interesting in operations management? *Manufacturing Service Oper. Management* 14(2):166–169.

Green LV, Kolesar PJ, Whitt W (2007) Coping with time-varying demand when setting staffing requirements for a service system. *Production Oper. Management* 16(1):13–29.

Hopp WJ (2012) MSOM Forum Fellows' statement. Last accessed January 16, 2013, http://msom.society.informs.org/fellows/wallace-hopp/.

Hopp WJ, Spearman ML (2004) To pull or not to pull: What is the question? *Manufacturing Service Oper. Management* 6(2):133–148.

Jagerman DL (1975) Nonstationary blocking in telephone traffic. *Bell System Tech. J.* 54(3):625–661.

Jennings OB, Mandelbaum A, Massey WA, Whitt W (1996) Server staffing to meet time-varying demand. *Management Sci.* 42(10):1383–1394.

Liu Y, Whitt W (2012) Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Oper. Res.* 60(6):1551–1564.

Massey WA (2002) The analysis of queues with time-varying rates for telecommunication models. *Telecomm. Systems* 21(2–4): 173–204.

Palm C (1943) Intensity variations in telephone traffic. *Ericsson Technics* 44:1–89. [English translation by North-Holland, Amsterdam, 1988.]

Ross SM (2010) *Introduction to Probability Models*, 10th ed. (Academic Press, New York).

Yom-Tov G, Mandelbaum A (2010) The Erlang-*R* queue: Time-varying QED queues with re-entrant customers in support of healthcare staffing. Working paper, Technion, Haifa, Israel.