MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 1523-4614 | EISSN 1526-5498 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Offered Load Analysis for Staffing

Ward Whitt
Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027
ww2040@columbia.edu, http://www.columbia.edu/~ww2040/

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: July 14, 2012; revision September 17, 2012

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 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 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 the single paper by 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 (2013). To give some idea about the exciting problems that remain, I briefly describe the current state of the art in the e-Companion 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 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* and its expected value, simply called the *offered load* (OL).

A key assumption underlying offered load analysis is that the demand is indeed *exogenous*; i.e., 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 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 face of known demand. However, in practice there usually is uncertainty about the demand, so that 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 in order 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 that 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. Since 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 multi-server 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 independently in parallel, unlimited waiting space, the first-come first-served service discipline and customer abandonment from 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 stochastic OL (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, not yet known in full generality), whereas the SOL itself is a Poisson random variable, with variance equal to the mean m. Since 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. Since staffing primarily involves service representatives (people), it is usually relatively flexible; i.e., 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 $P(X(t) \ge s(t)) \le \alpha$ 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 $P(X(t) \ge s(t)) \approx \alpha$ 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 EC.

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,

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

where the S and S_e are random variables with the fixed service-time cdf and the associated stationary-excess cdf, i.e., $P(S_e \le x) \equiv (1/E[S]) \int_0^x P(S > u)\,du$ 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

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

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),

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

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

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

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 $2\pi \approx 6.3$ 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

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

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 $\beta \equiv 1$

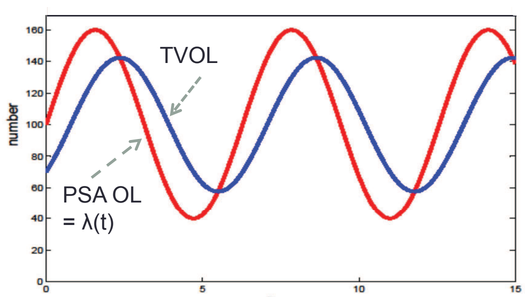

Figure 1 **The time-varying OL in (2) and (6) compared to the PSA approximate OL in (4), which coincides with the sinusoidal arrival rate function $\lambda(t)$ in (5).**

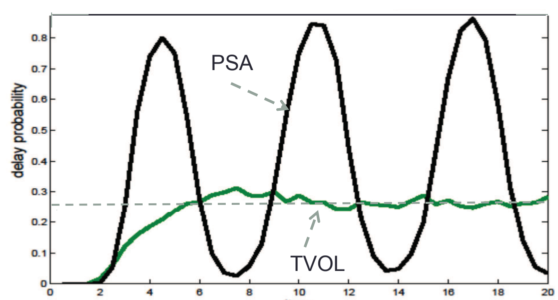

Figure 2 **Simulation estimates of the delay probability when staffing according to the SRS formula in (3) with $\beta = 1$ based on the OL and PSA approximation in Figure 1.**

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. ■

Example 1 shows that it is important to understand the OL $m(t)$ in (2) and its relation to the PSA approximation $m_{PSA}(t)$ in (4). Careful study has revealed many insights, e.g., 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, while 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.

References

Cachon, G. P. 2012. What is interesting in Operations Management? *Manufacturing and Service Oper. Management* **14**(2) 166–169.

Green, L. V., P. J. Kolesar, W. Whitt. 2007. Coping with time-varying demand when setting staffing requirements for a service system. *Production Oper. Management* **16** 13–29.

Hopp, W. J. 2012. MSOM Forum Fellows' statement. URL `http://msom.society.informs.org/fellows/wallace-hopp/`.

Hopp, W. J., M. L. Spearman. 2004. To pull or not to pull: what is the question? *Manufacturing and Service Oper. Management* **6** 133–148.

Jagerman, D. L. 1975. Nonstationary blocking in telephone traffic. *Bell System Tech. J.* **54** 625–661.

Jennings, O. B., A. Mandelbaum, W. A. Massey, W. Whitt. 1996. Server staffing to meet time-varying demand. *Management Sci.* **42** 1383–1394.

Liu, Y., W. Whitt. 2013. Stabilizing customer abandonment in many-server queues with time-varying arrivals. *Oper. Res.* **61**.

Massey, W. A. 2002. The analysis of queues with time-varying rates for telecommunication models. *Telecommunication Systems* **21** 173–204.

Palm, C. 1943. Intensity variations in telephone traffic. *Ericsson Technics* **44** 1–89.

Ross, S. M. 2010. *Introduction to Probability Models*. 10th ed. Academic Press, New York.

Yom-Tov, G., A. Mandelbaum. 2010. The Erlang-R queue: time-varying QED queues with re-entrant customers in support of healthcare staffing. Working paper, the Technion.

MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 1523-4614 | EISSN 1526-5498 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Offered Load Analysis for Staffing: e-Companion

Ward Whitt
Department of Industrial Engineering and Operations Research, Columbia University, New York, NY 10027
ww2040@columbia.edu, http://www.columbia.edu/∼ww2040/

This e-companion to the essay based on my 2012 MSOM Fellow Lecture briefly describes how the main ideas have already been pushed forward. Hopefully this brief description of the current state of the art will help researchers discover some of the exciting problems that remain.

It is natural to ask: What are the exciting research problems that remain? Our goal here is to provide a partial answer by describing how the ideas in the main paper, Whitt (2013), have been extended, thus revealing the current state of the art. We briefly discuss six extensions: (i) more general arrival and service processes, (ii) the modified offered load approximation, (iii) a theoretical explanation, (iv) more complex stochastic offered load models, (v) routing and scheduling and (vi) alternative approaches with inflexible staffing.

1. More General Arrival and Service Processes

Offered load analysis extends to more general models than the $M_t/GI/s_t + GI$ queueing model discussed so far. First, the service times can also be time-dependent; then a modification of formula (2) of Whitt (2013) still holds, namely,

$$m(t) \equiv E[X_\infty(t)] = \int_{-\infty}^{t} P(S_{t-s} > s)\lambda(t-s)\,ds. \tag{1}$$

but it is harder to analyze. Even when the service-time distribution varies over time, it often does so relatively slowly, so that formula (2) of Whitt (2013) can still be applied over subintervals.

Second, the offered load analysis extends to the more general $G_t/G/s_t + GI$ model with non-Poisson arrival process having a time-varying arrival rate $\lambda(t)$ and a general stationary sequence of service times, allowing dependence among successive interarrival times and among successive service times, e.g., see Pang and Whitt (2012). When the arrival process is no longer Poisson, the time-varying number of busy servers $X_\infty(t)$ itself no longer has a Poisson distribution, but a heavy-traffic limit theorem can still generate a normal approximation supporting the SRS formula. For the associated $G_t/G/\infty$ IS model, the approximation

$$X_\infty(t) \approx N(m(t), v(t)) \tag{2}$$

is asymptotically correct, where the mean $m(t)$ remains as in formula (2) of Whitt (2013), while the variance $v(t)$ satisfies

$$v(t) \equiv \int_0^\infty \lambda(t-s)V(s)\,ds, \quad V(s) \equiv G^c(s) + (c_a^2 - 1)G^c(s)^2 + \Gamma(s),$$
$$\Gamma(s) \equiv 2\sum_{k=1}^{\infty}(H_k^c(s,s) - G^c(s)^2), \quad G^c(s) \equiv P(S > s), \quad H_k(s_1,s_2) \equiv P(S_j \le s_1, S_{j+k} \le s_2), \quad (3)$$

with c_a^2 being the asymptotic variability parameter of the arrival process, which is 1 for a Poisson process, and S_j and S_{j+k} being service times separated by k indices (independent of j because of stationarity). The second term in the integral including the three terms of $V(s)$ drops out if the arrival process is Poisson, while the third term drops out if the service times are i.i.d.

With this generalization, a time-varying square-root-staffing (SRS) formula still holds. Instead of (3) in Whitt (2013), we now have

$$s(t) = m(t) + \beta\sqrt{v(t)}, \quad (4)$$

for $v(t)$ in (3) above, again using the least integer above that value.

2. The Modified Load Approximation

The offered load approach to staffing reviewed in §4 of Whitt (2013) succeeds in stabilizing the performance, but by itself cannot closely match specified performance targets. Given that the performance can indeed be stabilized, the appropriate constant level is not difficult to find by simulation, but the desired performance target also can often be met by applying the the *modified offered load* (MOL) approximation. The MOL approximation uses the steady-state performance of the corresponding stationary model with capacity constraints and other details, e.g., customer abandonment, but in a nonstationary way. The stationary OL for the approximation at time t is made to agree with the time-varying OL in formula (2) of Whitt (2013) by letting the arrival rate in the stationary model to be used for the MOL approximation at time t be

$$\lambda_{MOL}(t) \equiv m(t)/E[S]. \quad (5)$$

Moreover, it often suffices to apply many-server heavy-traffic approximations for that steady-state distribution. The MOL approximation can stabilize all standard performance measures at desired targets with a suitably high quality of service, but not all performance measures simultaneously at a low quality of service; see Liu and Whitt (2012b).

The MOL approximation also applies with more general arrival and service processes. Historically, the offered load analysis and the MOL approximation for non-Poisson arrival processes trace their roots to approximations for the performance of loss and delay models based on the peakedness

$$z(t) \equiv v(t)/m(t), \tag{6}$$

as discussed in Li and Whitt (2012), Pang and Whitt (2012) and references therein.

3. A Theoretical Explanation

It is natural to wonder why it should be possible to stabilize the performance using the SRS formula in (4) or equation (3) of Whitt (2013) when the offered load $m(t)$ is time-varying. Valuable insight can be gained from many-server heavy-traffic limits, because these SRS formulas correspond to the scaling in the Quality-and-Efficiency-Driven (QED) many-server heavy-traffic regime. Indeed, the QED regime is defined by the scaling

$$\frac{s-m}{\sqrt{m}} \to \beta \quad \text{as} \quad m \to \infty. \tag{7}$$

As the offered load increases with these SRS formulas held fixed for given QoS parameter β, in considerable generality the delay probability converges to a nondegenerate limit. Hence, at least for larger offered loads, we should anticipate that the delay probability should be independent of the offered load, provided that the SRS formula holds for all t with fixed QoS parameter β.

By similar reasoning, we can understand why it is not possible to stabilize the abandonment probability at high targets (low QoS) using the SRS formulas, while it is by the new method in Liu and Whitt (2012b). That behavior is to be expected because, with SRS scaling, the abandonment probability tends to be asymptotically negligible. A positive abandonment probability is consistent with a many-server heavy-traffic limit in the ED regime, in which $s = \beta m + o(m)$ as $m \to \infty$ for constant β, but not according to the SRS formula and the QED regime. In contrast, the new procedure for stabilizing the abandonment probability in Liu and Whitt (2012b) should be effective for high abandonment probabilities, because it is consistent with many-server heavy-traffic limits in the ED regime.

4. More Complex Offered Load Models

Another way offered load analysis can be generalized is to introduce network structure. That can arise when service does not take place continuously, but is occasionally interrupted, as in web chat, or when patients return to a medical unit of a hospital after being elsewhere, e.g., to take tests. Such features can be modeled by considering networks of queues and we can apply results for networks of IS queues, as in Massey and Whitt (1993), as has been done by Yom-Tov and Mandelbaum (2010).

The network structure can be viewed as adding a spatial component as well as a time component to the OL. Other spatial OL models are the Poisson arrival location model (PALM) in Massey and Whitt (1994) and the variant in Leung et al. (1994) introduced to represent wireless communication taking place in mobiles moving through space. General non-integer-valued OL models were introduced by Duffield et al. (2001) to represent the demand for bandwidth in communication networks. Capacity can be set in a similar way with these more general OL models.

5. Routing and Scheduling

For more complex systems with multiple classes and network structure, it is important to consider routing and scheduling and the design of the network, all of which can have important implications for staffing. However, in some cases, routing and scheduling policies can be developed that permit the overall problem to be decomposed, so that staffing can be considered separately, as for the single-class system considered above; e.g., see Gurvich and Whitt (2009).

6. Inflexible Staffing

Staffing to stabilize performance with time-varying arrival rates requires suitable flexibility in staffing, but in some cases, as in many hospitals, staffing is actually quite inflexible. Then the system inevitably must alternate between periods of overloading and underloading. Then offered load analysis as described above cannot be applied, because the periods of overloading significantly alter system performance. Nevertheless, infinite-server (IS) models can often be used to analyze the performance in both overloaded and underloaded intervals and thus select appropriate inflexible staffing levels. For this purpose, fluid and diffusion approximations resulting from many-server heavy-traffic limit theorems can be applied, as in Liu and Whitt (2012a) and references therein.

7. Conclusions

For staffing to meet time-varying exogenous demand in service systems, much depends on the length of the service times. When the service times are relatively short, as in many telephone call centers, it usually suffices to apply traditional stationary models in a nonstationary way in order to set staffing. On the other hand, when the service times are relatively long, as in many healthcare systems, that approach fails, while offered load analysis can often produce stable performance over time at target performance levels, even in face of strongly time-varying demand. Moreover, the scope of offered load analysis can be extended, e.g., by including space as well as time, as in a network of queues; see §4.

References

Duffield, N. G., W. A. Massey, W. Whitt. 2001. A nonstationary offered-load model for packet networks. *Telecommunication Systems* **13**(3-4) 271–296.

Gurvich, I., W. Whitt. 2009. Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing and Service Oper. Management* **11**(2) 237–253.

Leung, K. K., W. A. Massey, W. Whitt. 1994. Traffic models for wireless communication networks. *IEEE J. Selected Areas in Communication* **12**(8) 1353–1364.

Li, A., W. Whitt. 2012. Loss models with dependent service times. Columbia University, http://www.columbia.edu/∼ww2040.

Liu, Y., W. Whitt. 2012a. The $G_t/GI/s_t + GI$ many-server fluid queue. *Queueing Systems* **71**(4) 405–444.

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

Massey, W. A., W. Whitt. 1993. Networks of infinite-server queues with nonstationary Poisson input. *Queueing Systems* **13**(1) 183–250.

Massey, W. A., W. Whitt. 1994. A stochastic model to capture space and time dynamics in wireless communication systems. *Prob. in the Engineering and Informational Sciences* **8**(4) 541–569.

Pang, G., W. Whitt. 2012. The impact of dependent service times on large-scale service systems. *Manufacturing and Service Oper. Management* **14**(2) 262–278.

Whitt, W. 2013. Offered load analysis for staffing. *Manufacturing and Service Oper. Management* .

Yom-Tov, G., A. Mandelbaum. 2010. The Erlang-R queue: time-varying QED queues with re-entrant customers in support of healthcare staffing. Working paper, the Technion.