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.