IEOR 8100: Spring 2012, Professor Whitt

Staffing with Time-Varying Arrival Rates, February 22, 2012

1 Direct Use of Infinite-Server (IS) Models

1.1 The Offered Load Approach

Look at how many resources (servers) would be used if there were no limit on their availability. For many-server queues with time-varying arrival rates, that means looking at the associated IS model. For the IS model, we look at $N(t)$, the number of busy servers at time t. To a large extent, the staffing can be done with the mean $m(t) \equiv E[N(t)]$, which is called the **offered load**, but it is also helpful to use the time-varying variance. It can also be helpful to do refinements, such as the **modified offered load (MOL) approximations**.

For more general offered load approaches to communication networks, see [1, 10, 13, 15]. These fall outside the many-server queue paradigm.

1.2 Nonhomogeneous Poisson Processes (NHPP's) as Arrival Processes

For the $M_t/GI/\infty$ model, with NHPP arrival process, $N(t)$ has a Poisson distribution with a mean $m(t)$ that can be conveniently computed; see [2]. The formula for the mean help explain the physics of the system. The exact Poisson distribution implies that a normal approximation for $N(t)$ is appropriate. It suggests staffing by the classical **square-root staffing (SRS) formula**

$$s(t) = \lceil m(t) + \beta\sqrt{m(t)} \rceil, \tag{1}$$

where $\lceil x \rceil$ is the least integer greater than or equal to x and β is a Quality-of-Service (QoS) parameter; see [8] and the survey [6].

The analysis extends to networks of queues and other complex systems, where the arrival process at each queue can be regarded as a Poisson process, see [2, 3, 4, 8, 10, **?**, 13, 22] for systems with M_t NHPP arrival processes.

1.3 Non-Poisson Arrival Processes

The same idea applies to systems with non-Poisson G_t arrival processes. To obtain tractable formulas, we can use heavy-traffic limits for IS models as the arrival rate is allowed to grow. For the more general $G_t/G/\infty$ model, the heavy-traffic limits show that $N(t)$ is again approximately approximately Gaussian, where the mean $m(t)$ is the same as for the $M_t/GI/\infty$ model, but the variance is different. See [18], which draws on [**?**]. See [19, 20, 21] for results which allows for dependence among successive service times as well as complex dependence within the arrival process.

We get a new SRS formula

$$s(t) = \lceil m(t) + \beta\sqrt{v(t)} \rceil, \tag{2}$$

where $v(t)$ is the variance at time t, which no longer is simply the mean $m(t)$.

2 Alternative Approaches

2.1 Modified-Offered-Load (MOL) Approximations

The MOL approximations are useful refinements; see the survey [6]. The MOL approximations use associated stationary models, such as Erlang A, B or C, in a time-varying way. When the stationary model matches the structure of the real system, the MOL approximations tend to perform better than direct IS approximations. For more, see [4, 5, 7, 8, 14, 16, 22].

2.2 The Simulation-Based Iterative Staffing Algorithm (ISA)

For a simulation approach that can be applied to much more general systems, see [4].

2.3 Stabilizing the Abandonment Probability

The method to stabilize the delay probability in does not work to stabilize abandonment probabilities under heavy loads. A new method to stabilize abandonment probabilities under all loadings is developed in [11].

3 Practical Issues

3.1 Staffing Constraints

References

[1] N. G. Duffield, W. A. Massey, W. Whitt. 2001. A Nonstationary Offered-Load Model for Packet Networks. *Telecommunication Systems*, vol. 13, Nos. 3,4, March-April, pp. 271-296.

[2] Eick, S. G., W. A. Massey, W. Whitt. 1993a. The physics of the $M_t/G/\infty$ queue. *Oper. Res.* **41** 731–742.

[3] Eick, S. G., W. A. Massey, W. Whitt. 1993b. $M_t/G/\infty$ queues with sinusoidal arrival rates. *Management Sci.* **39** 241–252.

[4] Feldman, Z., A. Mandelbaum, W. A. Massey, W. Whitt. 2008. Staffing of time-varying queues to achieve time-stable performance. *Management Sci.* **54** 324–338.

[5] R. Hampshire, O. B. Jennings, W. A. Massey. 2008. A time-varying call center design via Lagrangian mechanics. Princeton.

[6] L. V. Green, P. J. Kolesar, W. Whitt, Coping with time-varying demand when setting staffing requirements for a service system. *Production and Operations Management* **16** (2007) 13–39.

[7] Jagerman, D. L. 1975.Nonstationary blocking in telephone traffic. *Bell System Technical Journal* **54** 625–661.

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

[9] E. V. Krichagina, A. A. Puhalskii, A heavy-traffic analysis of a closed queueing system with a GI/∞ service center. *Queueing Systems* **25** (1997) 235–280.

[10] K. Leung, W. A. Massey, W. Whitt. Traffic Models for Wireless Communication Networks. *IEEE Journal on Selected Areas in Communication*, vol. 12, No. 8, 1994.

[11] Y. Liu, W. Whitt, Stabilizing customer abandonment in many-server queues with time-varying arrivals. Columbia University, NY, NY (2011b) http://www.columbia.edu/∼ww2040/allpapers.html

[12] C. McCalla, W. Whitt. 2002. A time-dependent queueing-network model to describe life-cycle dynamics of private-line telecommunication services. *Telecommunication Systems* **17** 9–38

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

[14] W. A. Massey, W. Whitt. 1994. An analysis of the modified offered load approximation for the nonstationary Erlang loss model. *Ann. Appl. Probabil.* **4** 1145–1160.

[15] W. A. Massey, W. Whitt. 1994. A Stochastic Model to Capture Space and Time Dynamics in Wireless Communication Systems. Probability in the Engineering and Informational Sciences, vol. 8, 1994, pp. 541-569.

[16] W. A. Massey, W. Whitt. 1997. Peak congestion in multi-server sewrvice systems with slowly varying arrival rates. *Queueing Systems* **25** 157–172.

[17] G. Pang, R. Talreja, W. Whitt, Martingale proofs of many-server heavy-traffic limits for Markovian queues. *Probability Surveys* **4** (2007) 193–267.

[18] G. Pang, W. Whitt, Two-parameter heavy-traffic limits for infinite-server queues. *Queueing Systems* **65** (2010) 325–364.

[19] G. Pang, W. Whitt, Two-Parameter Heavy-Traffic Limits for Infinite-Server Queues with Dependent Service Times. *Queueing Systems*, to appear.

[20] G. Pang, W. Whitt, The Impact of Dependent Service Times on Large-Scale Service Systems. To appear in *Manufacturing and Service Operations Management*.

[21] G. Pang, W. Whitt, Infinite-Server Queues with Batch Arrivals and Dependent Service Times. To appear in *Probability in the Engineering and Information Sciences*.

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