IEOR 8100: Spring 2012, Professor Whitt

The Performance of Single-Server Queues, Wednesday, January 25, 2012

0. Individual Queues within a More Complex Setting

(a) Individual queues typically arise as components of a more complex system. Often a queue is one of many in a **network of queues**. Often there are multiple **customer classes** each with their own **route** or path through the network. See the QNA paper [7].

A variant of this approach can be used to analyze service systems, just like manufacturing facilities, communication networks and computer systems. See slides 11, 18 and 23 in the first introductory lecture.

In some most manufacturing and service systems, there is a natural starting point. In communication networks there may be multiple starting points, with flows in different directions.

(b) Analysis of such a complex queueing network often is done by a process of **aggregation** and **disaggregation**. Aggregation is done to combine all the contributions to the performance of each queue. After the performance of the queues has been analyzed, disaggregation is done to describe the performance of each customer class.

For the customer classes and routes, see §2.3 of [7] and §2.1 of [4]. For the aggregated model, see Figure 1 on p. 2781 of [7].

(c) There is a familiar **base model for each individual queue**, which is a **frame of reference**. It is good to know this base model, but to be aware that the key assumptions may not hold in applications. It helps to look at a complex system through the lens of the base model. Given that we know the base model, we can see if the assumptions are satisfied or not. Given that we have expectations, we are prepared to notice exceptions. It is important not to blindly assume that the model assumptions are satisfied.

See the six assumptions on p. 2781 of [7]. To those assumptions should be added that the system is approximately stationary over the time interval under consideration. We are assuming that we are trying to understand the steady-state performance of a stationary system. The situation changes if we should be considering the dynamic performance of a nonstationary system, e.g., with strongly time-varying arrival rate.

It is good to approach the queueing model with the understanding that the mathematical assumptions are likely to be valid only roughly. Given that the queueing model is roughly appropriate, we want to capture the first-order performance.

1. Many-Server Queues versus Single-Server Queues

Experience shows that for individual queues there are two classes that are (i) relatively tractable and (ii) have different performance. There is a relatively well-developed theory for individual queues that either have one server or have many (a large number of) servers. The intermediate case, e.g., with 2-10 servers, might be well-approximated by the theory of either of these two "ideal" cases. Key elements of the theory for a single-server queue extends to multi-server queues, but it remains to determine whether the single-server perspective or the many-server perspective is most useful.

How do many-server queues differ from single-server queues?

(a) What are typical server utilizations (proportion of time each server is busy) or, equivalently, traffic intensities?

(b) When is scheduling (service discipline) more important?

(c) How should the ratio EW/ES, the mean waiting time divided by the mean service time, depend on the model?

(d) What are typical waiting time distributions, in each case?

2. Steady-State Performance of the Single-Server Queue

Key Question: How does the expected steady-state waiting time, $E[W]$, depend on (1) the expected service time, $\tau \equiv E[S]$, (2) the traffic intensity $\rho \equiv \lambda\tau$, where λ is the arrival rate, and (3) the variability? How should the variability be quantified?

Simple answers: All three model components are important. With some model structure, we can conclude that: (1) The mean waiting time is *proportional* to the mean service time. (2) The mean waiting time explodes as ρ approaches 1. (3) The variability is crucial as well; if there would be no variability, i.e., in the $D/D/1$ model, then we would have no waiting at all, i.e., $E[W] = 0$, provided that the model is stable, i.e., if $\rho < 1$.

Quantifying the Impact of the Model Parameters

(a) The Pollaczek-Khintchine formula (exact) for the mean waiting time in **the** $M/GI/1$ **queue**

$$E[W] = \left(\frac{\tau\rho}{1-\rho}\right)\left(\frac{1+c_s^2}{2}\right) \tag{1}$$

where τ is the mean service time, λ is the arrival rate, $\rho \equiv \lambda\tau$ is the traffic intensity, S is a random service time, $c_s^2 \equiv Var(S)/E[S]^2$ is the squared coefficient of variation (SCV) of a service time S. We use the SCV because it is a scale-free measure of variability, i.e., $SCV(aS) = SCV(S)$ for $a > 0$. The second term in (1) captures all the variability impact.

(b) approximation formula for **the** $GI/GI/1$ **queue** (i.i.d interarrival times independent of i.i.d. service times, each with general distributions)

$$E[W] \approx \left(\frac{\tau\rho}{1-\rho}\right)\left(\frac{c_a^2+c_s^2}{2}\right) \tag{2}$$

where $c_a^2 \equiv Var(U)/E[U]^2$ is the squared coefficient of variation (SCV) of an interarrival time U.

(i) exact for $M/GI/1$ by part (a).

(ii) asymptotically correct in heavy traffic for $GI/GI/1$ as $\rho \uparrow 1$.

(iii) The range of possible values of $E[W]$ for given λ, τ and c_a^2 in the $GI/M/1$ model is not too large; see 1984 papers by WW, e.g. [3].

(iv) exact value of the mean $E[W]$ and the entire distribution can be computed for the $GI/GI/1$ model; see [1]

(c) What happens in **the more general** $G/G/1$ **model**, with stationary processes but without the independence assumptions? And is that really important?

(i) **The Stationary Interval Approach**: Ignore the dependence and approximate directly by $GI/GI/1$. If necessary, measure the mean and variance of the interarrival time and service time, and use the resulting parameters λ, c_a^2, τ and c_s^2.

(ii) **The Asymptotic Approach**: Heavy-traffic limit for the $G/G/1$ model (actually for the more general $G/G/s$ model)was established by Iglehart and WW (1970). For, $G/G/1$, a modification of (2) still holds in the heavy traffic limit. (See [**?**].)

$$E[W] \approx \left(\frac{\tau\rho}{1-\rho}\right)\left(\frac{c_A^2 - 2c_{AS}^2 + c_S^2}{2}\right) \tag{3}$$

where c_A^2, c_S^2 and $c_{A,S}^2$ are asymptotic variability parameters, i.e., the normalization constant in the central limit theorems (CLT's) for the arrival and service processes, with $c_{A,S}^2$ capturing correlations. See WW 2002 book, §9.6.1, especially (6.11) on p. 308. See Fendick, Saksena and WW (1989).

(d) What happens in **the multi-server $GI/GI/s$ model**?

rough approximation:

$$E[W; GI/GI/s] \approx E[W; M/M/s]\left(\frac{c_a^2 + c_s^2}{2}\right) \tag{4}$$

Different story when s is large.

3. The Queueing Network Analyzer (QNA): A Parametric-Decomposition Approximation; see 1983 WW QNA paper.

(a) A convenient fiction (approximation). At a minimum, this view corresponds to starting with a process flow diagram.

(b) a network calculus for transforming the variability parameters of the flows; see §9.9 of [8], as well as [7].

(i) flow through a queue: from arrival process to departure process

For departure process from $GI/GI/1$ queue, see (38) in §4.5 of [7]:

$$c_d^2 = \rho^2 c_s^2 + (1-\rho^2)c_a^2. \tag{5}$$

(ii) merging streams: superposition processes

For superposition of independent renewal processes, see (33) of [7]:

$$c_H^2 = w\sum_{i=1}(\lambda_i/\lambda)c_i^2 + (1-w). \tag{6}$$

(iii) splitting streams

For independent splitting of renewal process, get renewal process, see (36) in §4.4 of [7]:

$$c_i^2 = p_i c^2 + (1-p) \tag{7}$$

(c) Motivates Measurements

(i) What are common service-time processes? What are common service-time distributions?

low-variability in manufacturing, heavy-tail in file sizes, lognormal

(ii) What are common arrival processes?

Poisson or?

4. Supporting Limits for Superposition Processes

(a) Convergence of superposition processes to a Poisson Process (number n of processes goes to infinity)

(b) The CLT: no convergence to Poisson; see [8], Theorem 9.4.1 (time t goes to infinity for any given number of processes)

(c) The two iterated limits do not agree: $\lim_{t\to\infty}\lim_{n\to\infty} \neq \lim_{n\to\infty}\lim_{t\to\infty}$

(d) Use Double Limit, see [8], §9.8.

5. Important Role of Measurements

indices of dispersion, see [5] and [2].

References

[1] J. Abate, G. L. Choudhury, W. Whitt, Calculation of the GI/G/1 Steady-State Waiting-Time Distribution and its Cumulants from Pollaczek's Formula. *Archiv fur Elektronik und Ubertragungstechnik*, vol. 47, No. 5/6, 1993, pp. 311-321.

[2] K. W. Fendick, V. R. Saksena, W. Whitt, Dependence in Packet Queues. *IEEE Transactions on Communications*, vol. 37, No. 11, 1989, pp. 1173-1183.

[3] J. G. Klincewicz, W. Whitt, On Approximations for Queues, II: Shape Constraints. *AT&T Bell Laboratories Technical Journal*, vol. 63, No. 1, January 1984, pp. 139-161.

[4] M. Segal, W. Whitt, A Queueing Network Analyzer for Manufacturing. *Teletraffic Science for New Cost-Effective Systems, Networks and Services*, Proceedings of ITC 12 (ed. M. Bonatti), North-Holland, Amsterdam, 1989, pp. 1146-1152.

[5] K. Sriram, W. Whitt, Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data. *IEEE Journal on Selected Areas in Communications*, vol. SAC-4, No. 6, September 1986, pp. 833-846.

[6] W. Whitt, Approximating a Point Process by a Renewal Process: Two Basic Methods. *Operations Research*, vol. 30, No. 1, January-February 1982, pp. 125-147.

[7] W. Whitt, The Queueing Network Analyzer. *Bell System Technical Journal*, vol. 62, No. 9, November 1983, pp. 2779-2815.

[8] W. Whitt, *Stochastic-Process Limits*, Springer, New York, 2002.