IEOR 8100: Spring 2012, Professor Whitt

Traffic Measurements and Many-Server Queues, February 8, 2012

1 Today's Topics

There are two topics for today:

1. traffic measurements for single-server queues, emphasizing the dependence
2. many-server queues, the important role of infinite-server queues

We first continue the previous two lectures and discuss traffic measurements related to stationary single-server queues. We then consider many-server queues, discussing how to quantify performance. A key issue is how many-server queues differ from single-server queues. A key model to understand the performance of a many-server queue is the associated infinite-server model. We will then combine the two topics and discuss traffic measurements for many-server queues. The peakedness helps to understand the impact of the traffic on performance.

We will next discuss time-varying arrivals, contrasting common behavior for single-server queues and many-server queues. Successful analysis of many-server queues, with time-varying arrival rates and without, draws heavily on the analysis of infinite-server queues.

2 Traffic Measurements for Stationary Single-Server Queues

2.1 Familiar Traffic Measurements

By traffic measurements for a stationary single-server queue, we primarily mean measurements of the arrival and service processes, but we also want to directly observe how the system performs. So we want to also observe the performance statistics that we would use the queueing models to predict. It is helpful to think of traffic measurements related to underlying stochastic queueing models. We here are considering the special case of stationary models.

As reviewed in the previous lecture, the arrival process in a stationary queueing model is a stationary point process on the positive half line $[0, \infty)$. Hence, the arrival process can be represented in three related ways: (i) in terms of $A(t)$, the number of arrivals in the interval $[0, t]$ as a function of $t \geq 0$, (ii) in terms of A_n, the epoch (time) of arrival n as a function of $n \geq 1$, with $A_0 \equiv 0$, and (iii) in terms of U_k, the interval between arrival $k-1$ and arrival k for $k \geq 1$. In terms of the intervals U_k, we write

$$A_k \equiv U_1 + \cdots + U_k, \quad k \geq 1, \quad \text{and} \quad A(t) \equiv \max\{k \geq 0 : A_k \leq t\}, \quad t \geq 0. \tag{1}$$

It may be appropriate to associate service times with arrivals or with servers. For a single-server queue with the first-come first-served (FCFS) service discipline, these two perspectives are equivalent, but not more generally. Let V_k be the service time associated with arrival k.

In the $GI/GI/1$ model the two sequences $\{U_k : k \geq 1\}$ and $\{V_k : k \geq 1\}$ are independent sequences of i.i.d. random variables, distributed as generic random variables U and V with general distributions, say F and G. For the $M/M/1$ model the distributions F and G are exponential. We suppose that these distributions have finite means $E[U] = \lambda^{-1}$ and $E[V] = \mu^{-1}$. The traffic intensity is $\rho \equiv E[V]/E[U] = \lambda/\mu$. To have a proper steady-state in the queue, we assume that $\rho < 1$.

There are two ways to look at traffic measurements: (i) the discrete-time perspective, focusing on the interarrival times and service times (U_k, V_k), $1 \le k \le n$, and (ii) the continuous-time perspective, focusing on the counting process $A(t)$ and associated service times, e.g., $V_k : 1 \le k \le A(t)$, for some fixed t.

Discrete-Time Traffic Measurements: the finite segment $(U_k, V_k) : 1 \le k \le n$. In this representation, U_k is the interval between arrival $k-1$ and arrival k, while V_k is the service time associated with arrival k. There are no stochastic model assumptions, but we have in mind the $M/M/1$ and $GI/GI/1$ queueing models as reference cases.

Continuous-Time Traffic Measurements: the finite segments $\{A(s) : 0 \le s \le t\}$ **and** $\{V_k : 1 \le k \le A(t)\}$. In this continuous-time representation, $A(s)$ is the number of arrivals in the interval $[0, s]$ while V_k is the service time associated with arrival k. There are no stochastic model assumptions, but we have in mind the $M/M/1$ and $GI/GI/1$ queueing models as reference cases.

The first thing to check is the **stationarity assumption**. Assume that has been done for the following discussion.

Thinking of the $M/M/1$ model model, we want to estimate the parameters λ and μ by finite averages, e.g., for the two perspectives, we use

$$\bar{\lambda}(t) \equiv \frac{A(t)}{t} \quad \text{or} \quad \bar{U}_n \equiv \frac{A_n}{n}. \tag{2}$$

Clearly, $1/\bar{\lambda}(t) \approx \bar{U}_{A(t)}$. (Recall the inverse relation between the processes $\{A_n : n \ge 1\}$ and $\{A(t) : t \ge 0\}$ discussed in the last class.)

Given that we use data to estimate arrival rates and mean service times, there is the usual issue of statistical precision. As usual, it is important to estimate confidence intervals to understand the statistical precision of the direct averages. Within operations research, that issue is usually discussed as a fundamental topic of stochastic simulations. Confidence intervals are routinely estimated in simulation output analysis.

Thinking of the $GI/GI/1$ model, we want to estimate the cdfs F and G, which is naturally done by the empirical cdfs, i.e.,

$$\bar{F}_n(t) \equiv \frac{1}{n} \sum_{i=1}^{n} 1_{\{U_i \le t\}}, \quad t \ge 0, \tag{3}$$

where 1_A is the usual indicator function of the set A, equal to 1 on A and 0 otherwise. We look at the histogram and possibly fit various distributions to it.

Since the performance of a $GI/GI/1$ queue primarily depends on the distributions F and G beyond their means through their variances or SCV's (see Lecture of 012512), we estimate those. Just like the mean, we can estimate the k^{th} moment of F, $m_k \equiv E[U^k]$ by the sample average

$$\bar{m}_k(n) \equiv \frac{1}{n} \sum_{i=1}^{n} U_i^k. \tag{4}$$

Then we can estimate the variance of F, $\sigma_U^2 \equiv Var(U)$, by

$$\bar{\sigma}_U^2(n) \equiv \bar{m}_2(n) - (\bar{m}_1(n))^2. \tag{5}$$

We then can estimate the two SCV's c_a^2 and c_s^2 as functions of the corresponding estimators; e.g., we estimate c_a^2 by $\bar{\sigma}_U^2(n)/(\bar{m}_1(n))^2$.

If the random variables actually are dependent, then the traditional sample variance is not such a good estimator, i.e., we might refrain from using

$$s_U^2(n) \equiv \frac{1}{n-1}\sum_{i=1}^{n}(U_i - \bar{m}_1(n))^2, \tag{6}$$

just as we should avoid using the sample variance if we were estimating the variance of the waiting-time distribution. Even in the $M/M/1$ model, the successive waiting times are typically dependent.

2.2 Measurements of the Dependence

As discussed in the last class, a real queueing system, even if it is approximately stationary, is often not well modeled by the $GI/GI/1$ model, because there is dependence among the interarrival times and service times. An arrival process is rarely renewal if it is not Poisson. Thinking of more general $G/G/1$ models, where dependence plays a role, we also want to quantify the dependence among successive interarrival times and service times. A convenient partial characterization of dependence can be thorough the second-order properties, i.e., through covariances and correlations as well as variances. (It is important to be aware that this is indeed only a partial characterization, just as the mean and variance are only a partial characterization of a probability distribution.)

In choosing appropriate measurements, we are guided by what we know about the queueing model. Key insights are obtained from the asymptotic behavior of the $G/G/1$ queue in light and heavy traffic. The story is especially clean and clear mathematically in heavy traffic. The heavy-traffic behavior of the $G/G/1$ queue is determined by the arrival process and service process beyond the parameters λ and μ by the variability constants in the CLT's. These in turn are the asymptotic variances, including all the covariances; see §4.4 of [12].

It is important to avoid the error of looking only at the marginal distribution and the lag-1 correlation. Even when the dependence matters a lot, as with superposition processes, the marginal distribution can be nearly exponential and the lag-1 correlation can be nearly 0. Checking only those properties is not enough to be sure that an $M/M/1$ model is appropriate. There can be many small correlations over many lags. This is not a rare pathology, but something that occurs frequently with superposition arrival processes; see [7, 1].

2.2.1 Three Indices of Dispersion

The **index of dispersion for intervals (IDI)** of a stationary arrival process is

$$IDI(k) \equiv c_a^2(k) \equiv \frac{kVar(A_k)}{(E[A_k])^2} = \frac{Var(A_k)}{k(E[U_k])^2} = c_a^2(1) + \frac{\sum_{i,j=1;i\neq j}^{k} cov(U_i, U_j)}{k(E[U_1])^2},\ k \geq 1. \tag{7}$$

The limit as $k \to \infty$ (if it is finite) is the asymptotic approximation, that characterizes the heavy-traffic behavior of the queue. For the IDI of the arrival process, see formula (1) of [7]. For the IDI of the arrival and service process combined, see (15) of [1], where we focus on the functions of n before letting $n \to \infty$. Two extreme perspectives are the stationary-interval method, using variability parameter $c_a^2(1)$, which all ignores dependence, and the asymptotic method, using $c_a^2(\infty) \equiv \lim_{n\to\infty} c_a^2(n)$, which counts all the covariances (the limit as $n \to \infty$); see [9]. Indices of dispersion consider all intermediate values.

Important reference points are Poisson processes, deterministic processes and renewal processes. For a Poisson process, $IDI(k) = 1$ for all k; for a renewal process with SCV c^2,

$IDI(k) = c^2$ for all k. Dependence can be detected by seeing non-constant IDI's in estimates of the IDI.

The associated **index of dispersion for counts (IDC)** of a stationary arrival process is a continuous-time analog of the IDI, in particular,

$$IDC(t) \equiv c_a^2(t) \equiv \frac{Var(A(t))}{E[A(t)]}, \quad t \geq 0. \tag{8}$$

Again, important reference points are Poisson processes. For a Poisson process, $IDI(t) = 1$ for all t. However, the IDC of a renewal process is actually complicated, so dependence is not so easy to see directly. The IDC has the advantage over the IDI that the time scale of the dependence can be seen directly in the behavior of the IDC as a function of time t

For the single-server queue (especially the steady-state workload), it is useful to look at the input of workload in required service time. Thus in [3] we introduced the **index of dispersion for work (IDW)**. To define the IDW, let

$$X(t) \equiv \sum_{i=1}^{A(t)} V_i, \quad t \geq 0. \tag{9}$$

The IDW is defined by

$$IDW(t) \equiv \frac{Var(X(t))}{\tau E[X(t)]}, \quad t \geq 0, \tag{10}$$

where $\tau \equiv E[V]$ is the mean service time. When the service times are independent of the arrival process and i.i.d., the IDW in (10) can be expressed simply in terms of the SCV of the service times SCV c_s^2 and the IDC in (8), in particular,

$$IDW(t) = c_s^2 + IDC(t), \quad t \geq 0; \tag{11}$$

(See Example 3.19 of Green Ross.) See [3] for more discussion.

In the more complex queueing systems, the indices of dispersion will be non-constant functions. That means that the arrival and service processes tend to exhibit different levels of variability at different time scales. Very roughly, in heavy (light) traffic the values of the indices of dispersion at large arguments (small arguments) tend to describe performance better. The indices of dispersion show the levels of variability in the arrival process and in the service process (or in the arrival and service) processes combined) as a function of continuous time t or discrete time n. Very roughly, if the queue length frequently assumes value like k, then the arrivals separated by k tend to interact, so that the IDI is relevant for $n \approx k$. Similarly, if the waiting time assumes values like x, then arrivals time x apart tend to interact, so that the IDC is relevant for $t = x$. See [1, 2, 3, 7] for more discussion.

2.2.2 Estimating the Indices of Dispersion

With data, from actual systems or simulations, these indices of dispersion must be estimated. This primarily reduces to estimating variances. In simulations it is convenient to work with independent replications. For the IDC, we would have m independent samples of the segment $\{A(s) : 0 \leq s \leq t\}$; for the IDW, we would have m independent samples of the segment $\{X(s) : 0 \leq s \leq t\}$. We can estimate $Var(A(s))$ and $Var(X(s))$ for a finite set of s, $0 \leq s \leq t$, all at one time, but of course these estimates will be dependent.

However, it is also possible to estimate these variances using overlapping intervals, assuming that we have processes with stationary increments. To estimate $Var(A(s)$, we estimate the first

two moments of $A(s)$ be looking that the sample average of estimates taken from finitely many intervals $[t_i, t_i + s]$, which can be overlapping. We estimate the variance by subtracting the square of the estimate of the mean from the estimate of the second moment. It is convenient to use independent replications to estimate confidence intervals.

The same estimation procedure applies approximately with system data if we can obtain from multiple days under the same or similar conditions. We then could estimate $Var(A(s))$ by the sample variance

$$s^2_{A(s)} \equiv \frac{1}{m-1}\sum_{i=1}^{m}(A_i(s) - \bar{m}_{A(s)})^2, \tag{12}$$

where $\bar{m}_{A(s)}$ is the usual estimate of the mean $E[A(s)]$, i.e.,

$$\bar{m}_{A(s)} \equiv \frac{1}{m}\sum_{i=1}^{m} A_i(s). \tag{13}$$

We can then estimate $IDC(s)$ by $s^2_{A(s)}/\bar{m}_{A(s)}$.

Assuming that $A(s)$ is approximately a Gaussian random variable with unknown mean and variance, we would regard $(\bar{m}_{A(s)} - E[A(s)])/\sqrt{s^2_{A(s)}/m}$ as distributed approximately according to the Student-t distribution with $m-1$ degrees of freedom. Thus, when $m = 10$, a 95% confidence interval for $E[A(s)]$ would be

$$\bar{m}_{A(s)} \pm 2.26\sqrt{s^2_{A(s)}/10}. \tag{14}$$

Considering the variance, we would use the fact that $(m-1)s^2_{A(s)}/Var(A(s))$ has the chi-square distribution with $m-1$ degrees of freedom. Thus we can form corresponding confidence intervals for $Var(A(s))$ in terms of $s^2_{A(s)}$.

2.2.3 Queue Measurements of Traffic Processes

An alternative way to characterize the traffic is to see how various test queueing systems behave when this traffic is offered to that system. This is the idea of peakedness approximations for many-server queues; then the test system is an infinite-server queue; see

However, the test system could be an alternative model, such as a single-server model; see [8, 11].

3 Many-Server Queues

3.1 Insensitivity in the $M/GI/s/0$ and $M/GI/\infty$ Models

The steady-state number in system is independent of the service-time distribution beyond its mean in the $M/GI/s/0$ and $M/GI/\infty$ models. This is an important theoretical reference point for many-server queueing models. However, that insensitivity property does not hold in corresponding $M/GI/s/\infty$ and $M/GI/s/r$ delay models, with or without abandonment, it does not hold in $GI/GI/s/0$ and $GI/GI/\infty$ models with non-Poisson arrival processes, and it does not hold for the time-dependent performance in the associated $M_t/GI/s/0$ and $M_t/GI/\infty$ models with time-varying arrival rates. See [18, 14, 15].

3.2 The Important Role of Infinite-Server Queues

First consider the stationary $G/G/s$ model, with unlimited waiting space. Let B be the steady-state number of busy servers. By Little's law, applied to the service facility,

$$E[B] = \lambda\tau = \frac{\lambda}{\mu}. \tag{15}$$

In the associated infinite-server (IS) model, $E[B]$ is also the expected steady-state number of customers in the system.

3.2.1 Characterizing Traffic Via the Peakedness

Consider the $G_a/G/s/r$ model with r waiting spaces, so that the loss model corresponds to $r = 0$, while the delay model corresponds to $r = \infty$. Assume that the service process is independent of the arrival process. The associated $G_a/G/\infty$ system with $s = \infty$ serves as a direct approximation, which can be useful if $E[B]$ is not too high. More generally, we can partially characterize the general G_a arrival process in the $G_a/G/s/r$ model by looking at the congestion produced in the associated $G_a/G/\infty$ system, which differs only by letting $s = \infty$. This is a useful device because the IS system tends to much easier to analyze.

The **peakedness** of the given G_a arrival process relative to the fixed G service process is defined as the ratio of the variance to the mean of the number of busy servers, i.e.,

$$z \equiv z(G) \equiv z_a(G) \equiv \frac{Var(B)}{E[B]}, \tag{16}$$

where B is the steady-state number of busy servers in the $G_a/G/\infty$ model; see [13, 16].

As a frame of reference, we observe that B has a Poisson distribution in the $M/GI/\infty$ special case with mean $E[B] = \lambda\tau$, independent of the service-time distribution beyond its mean. Since the variance of a Poisson distribution is equal to its mean, we have $z = 1$ in the M/GI case. For a renewal arrival process in which the interarrival-time has cdf F with pdf f and Laplace transform

$$\phi(s) \equiv \int_0^\infty e^{-st} dF(t) = \int_0^\infty e^{-st} f(t)\, dt, \tag{17}$$

and an exponential service time with mean μ^{-1}, i.e., in the GI/M case, the peakedness turns out to be

$$z(\mu; G) = \frac{1}{1 - \phi(\mu)} - \frac{\lambda}{\mu}. \tag{18}$$

Let $B(s, a)$ be the Erlang blocking formula in the $M/M/s/0$ Erlang Loss model with s servers and $a \equiv \lambda/\mu$. The Hayward approximation for the blocking in a $G_a/G/s/0$ system is

$$B(s, a; G_a, G) \equiv B(s/z, a/z), \tag{19}$$

which is applicable because $B(s, a)$ can be extended to non-integer s; see [13].

3.2.2 The Heavy-Traffic Peakedness

Now consider the $G/GI/\infty$ model, where the service times are independent of the arrival process and are themselves i.i.d. Assume that the arrival process satisfies a FCLT with scaling factor λc_a^2, so that c_a^2 is the asymptotic variability parameter (limit of IDC and IDI). In heavy

traffic (as $\lambda \to \infty$), the distribution of B is asymptotically Gaussian with mean in (15) and variance

$$Var(B) \approx \lambda\tau + \lambda(c_a^2 - 1)\int_0^\infty G^c(s)^2\, ds; \tag{20}$$

see [15] and the many references cited there. See [14] for simple case of D service.

We again focus on the ratio of the variance to the mean, which is called **the heavy-traffic (HT) peakedness**. it is defined by

$$z_{HT}(G) \equiv \lim_{\lambda\to\infty}\left(\frac{Var(B)}{E[B]}\right). \tag{21}$$

In the $M/GI/\infty$ model, B has a Poisson distribution, so that $z = 1$. The heavy-traffic limit for the peakedness (with respect to the service-time cdf G is obtained from (20):

$$z_{HT}(G) = 1 + (c_a^2 - 1)\frac{\int_0^\infty G^c(s)^2\, ds}{\tau}. \tag{22}$$

The final ratio always falls in the interval $[0, 1]$ and assumes the value 1 for a deterministic (D) distribution, 1/2 for an exponential M distribution, and converges to 0 as the variability increases. (In particular, under regularity conditions, $G^c(s)^2$ decreases and $\int_0^\infty G^c(s)^2\, ds$ gets small as variability increases, even though the mean $\tau = \int_0^\infty G^c(s)\, ds$ is held fixed.)

4 Time-Varying Arrival Rates

Upcoming topic.

4.1 Coping with Time-Varying Arrival Rates

See [19]. IS models yield approximations, e.g., drawing on [18].

4.2 Time-Varying Offered-Load Models

The mean number of busy servers in the IS model serves as the time-varying offered load (OL). The formula for the $M_t/GI/\infty$ model in [18] is actually valid for more general G_t arrival processes.

References

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

[2] K. W. Fendick, V. R. Saksena, W. Whitt, Investigating Dependence in Packet Queues with the Index of Dispersion for Work. *IEEE Transactions on Communications*, vol. 39, No. 8, 1991, pp. 1231-1244

[3] K. W. Fendick, W. Whitt, Measurements and Approximations to Describe the Offered Traffic and Predict the Average Workload in a Single-Server Queue. *Proceedings of the IEEE*, vol. 77, No. 1, 1989, pp. 171-194,

[4] Leland, W. E., M. S. Taqqu, W. Willinger, D. Wilson. On the self-similar nature of Ethernet traffic (extended version) *IEEE/ACM Transactions on Networking*, vol. 2, 1994, 1–15.

[5] Pang, G., W. Whitt. Two-Parameter Heavy-Traffic Limits for Infinite-Server Queues. *Queueing Systems*, vol. 65, 2010, pp. 325-364.

[6] Paxson, V., S. Floyd. Wide area traffic: the failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, vol. 3, 1995, 226-224.

[7] 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.

[8] W. Whitt, Approximating a Point Process by a Renewal Process: The View Through a Queue, An Indirect Approach. *Management Science*, vol. 27, No. 6, June 1981, pp. 619-636

[9] 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.

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

[11] W. Whitt, Queue Tests for Renewal Processes. *Operations Research Letters*, vol. 2, No. 1, April 1983, pp. 7-12

[12] W. Whitt, *Stochastic-Process Limits*, Springer, New York, 2002. Available at: http://www.columbia.edu/ ∼ww2040/jumps.html

[13] A. A. Fredericks, Congestion in blocking systems – a simple approximation technique. The Bell System Technical Journal, vol. 59, No. 6, July-August 1980, pp. 805-827.

[14] P. W. Glynn, W. Whitt, A New View of the Heavy-Traffic Limit for Infinite-Server Queues. *Advances in Applied Probability*, vol. 23, No. 1, 1991, pp. 188-209.

[15] G. Pang, W. Whitt, Two-Parameter Heavy-Traffic Limits for Infinite-Server Queues. *Queueing Systems*, vol. 65, 2010, pp. 325-364. (with Guodong Pang)

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

[17] W. Whitt, Understanding the Efficiency of Multi-Server Service Systems. *Management Science*, vol. 38, No. 5, 1992, pp. 708-723.

[18] S. G. Eick. W. A. Massey, W. Whitt, The Physics of The Mt/G/infty Queue. *Operations Research*, vol. 41, No. 4, 1993, pp. 731-742.

[19] 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*, vol. 16, No. 1, January-February 2007, pp. 13-39.

[20] Y. Liu, W. Whitt, The $G_t/GI/s_t + GI$ Many-Server Fluid Queue. Submitted to Queueing Systems. Available at: http://www.columbia.edu/∼ww2040/allpapers.htm