IEOR 6711: Stochastic Models I

Fall 2013, Professor Whitt

Topics for Discussion, Thursday, September 26

Infinite-Server Queues and Staffing

We continued discussing the papers posted last Tuesday. We emphasized Theorem 1 of the 1993 "Physics" paper. It describes the $M_t/GI/\infty$ infinite-server queueing model, with a nonhomogeneous Poisson arrival process (NHPP). See Prop. 2.3.2, Example 2.3(B), Example 2.3(C) and Section 2.4 of the Ross textbook for related material.

Last time we discussed Theorem 9 and formula (14) of that same paper, which gives an approximation for the time lag and the space shift, assuming that the arrival-rate function is quadratic. Such a quadratic function can arise by taking a Taylor series approximation.

The following are the main topics discussed on Thursday.

1. **not a Poisson process**

   Even though $Q(t)$, the number of busy servers at time t in the $M_t/G/\infty$ model, has a Poisson distribution for each t, the stochastic process $\{Q(t) : t \geq 0\}$ is *not* a nonhomogeneous Poisson process (NHPP). An NHPP is a counting process. It has sample paths that are nondecreasing. That is not true of $\{Q(t) : t \geq 0\}$.

2. **the covariance**

   Theorem 2 describes $Cov(Q(t), Q(t + u))$. The idea is to exploit the random measure representation and the picture, here Figure 3. We see that $Q(t) = X+Y$, while $Q(t+u) = Y + Z$, where X, Y and Z are independent. Hence

   $$Cov(Q(t), Q(t + u)) = Cov(Y, Y) = Var(Y) = E[Y],$$

   where $E[Y]$ has a simple integral formula, like the mean $m(t) = E[Q(t)]$.

3. **the departure process and the departure rate**

   Theorem 1 in the physics paper includes a description of the departure process and the departure rate. Note that the departure process *is* an NHPP. It is easy to see that it has independent increments; see Figure 2.

   Note that the departure rate has a formula closely related to the mean $m(t) \equiv E[Q(t)]$.

4. **ODE with M service**

   Theorem 6 and Corollary 4 show that the mean $m(t)$ satisfies an ODE when the service-time distribution is exponential. That reveals how the peaks of m and λ are related. In particular, that explains why the curve for $m(t)$ crosses the curve for $\lambda(t)E[S]$ where the derivative $\dot{m}(t) = 0$, e.g., where $m(t)$ assumes its maximum. This tends to be approximately true in corresponding models with finitely many servers.

5. **relaxation time: approach to steady state**

   For a stationary model, it is important to understand how the system approaches steady state as time evolves, starting with various typical special initial conditions, such as starting empty. A very simple revealing formula exists for the $M_t/GI/\infty$ model; formula (20). It shows how $m(t)$ approaches the steady state value $m(\infty) = \lambda ES$ when the system starts empty. That is achieved by letting $\lambda(t) = \lambda$ for $t \geq 0$, but letting $\lambda(t) = 0$ for $t < 0$.. In particular,

$$\frac{m(t)}{m(\infty)} = G_e(t) = P(S_e \leq t) < \quad t \geq 0.$$

Thus we might say that ES_e is approximately the time required to approach steady state. Since steady state is approached gradually, some simplification is needed in the definition.

6. **sinusoidal and other periodic arrival rates**

   The sine paper describes results for periodic arrival rate functions,

   $$\lambda(t) = \bar{\lambda} + \beta \sin{(\gamma t},$$

   as in (6), which we assume holds for all t into the infinite past.

   The key fact is that m inherits structure from λ. The function $m(t)$ is also sinusoidal with the same frequency, but there is a time lag and space shift there too. Hence revealing formulas are available. In particular, a general formula for the mean $m(t)$ is given in Theorem 4.1 when the arrival rate function is given in (6). The case of exponential service times yields a convenient explicit formula; see (15) in Section 5:

   $$m(t) = \bar{\lambda} + \frac{\beta}{1+\gamma^2} \left(\sin{(\gamma t)} - \gamma \cos{(\gamma t)}\right).$$

7. **starting in the infinite past**

   It is important to point out and emphasize that the simple formulas for the sinusoidal arrival rate function depend on starting in the infinite past. Starting at time 0 is covered as the special case in which $\lambda(t) = 0$ for $t < 0$. But that assumption makes the formulas for $m(t)$.

8. **staffing: the 1996 paper**

   We briefly discussed the 1996 staffing paper. The infinite-server (IS) approximation, or offered-load approximation, we have been discussing should be contrasted with the pointwise-stationary approximation (PSA) and the simple stationary approximation (SSA) there, for the case of a sinusoidal arrival-rate function. The PSA approximates the time-dependent performance at time t by the performance of the stationary model with constant arrival rate equal to $\lambda(t)$, The SSA approximates the performance at all times t by the performance of the stationary model with average arrival rate. When the arrival rate changes slowly compared to the mean service time, the PSA performs well. When the arrival rate changes very rapidly compared to the mean service time, the SSA performs well. We can understand by looking at the explicit formula for $m(t)$ in the case of a sinusoidal arrival rate, given above. The rate of change depends on the scaling parameter γ there.

Related topics not discussed in class:

1. **Erlangs: the concept of offered load**

   An Erlang is a dimensionless quantity indicating the average amount of load in a stochastic system. It is the mean number of busy servers in an associated infinite-server model.

That notion is defined for a stationary model. We want to extend it to a nonstationary model. The first idea is to go beyond arrival rate and include the service requirements. The second idea is to adjust for nonstationarity. We regard $m(t)$ above as the time-varying offered load.

2. **the modified offered load approximation**

   An important concept and method is the modified offered load (MOL) approximation. The idea is to use a stationary model at each time t but with the arrival rate depending on the time-varying offered load $m(t)$ instead of the arrival rate function $\lambda(t)$. In particular, at time t we use the stationary model with arrival rate function

$$\lambda_{MOL}(t) \equiv \frac{m(t)}{ES}.$$

   This tends to work well.

3. **networks of infinite-server queues**

   The theory extends to networks of infinite-server queues; e.g., see W. A. Massey and WW, Networks of Infinite-Server Queues with Nonstationary Poisson Input. *Queueing Systems*, vol. 13, No. 1, 1993, pp. 183-250.

4. **many subsequent developments**

   See Y. Liu and WW, Stabilizing Customer Abandonment in Many-Server Queues with Time-Varying Arrivals. *Operations Research*, vol. 60, No. 6, November-December 2012, pp. 1551-1564.