Stochastic Networks

A "system" such as a supermarket or a bank involves "customers" who arrive randomly over time and require service of random durations. Competition for service among the customers causes delays and customers typically must wait in *queues* (lines) before receiving service. Inside the system there may be only one service facility, yielding a *single-server* queue, or many facilities, yielding a complex *queueing network* in which customers move around among the facilities before departing. Service disciplines other than first-in-first-out (FIFO) might be employed such as last-in-first-out (LIFO), preemptive LIFO (P-LIFO), shortest-job-next (SJN), or processor sharing (PS). More generally, for the purpose of minimizing congestion or maximizing revenue, various *control policies* might be derived and then employed by the system, regulating both the arrival streams and servicing facilities. (Stochastic control theory)

Customers might really represent people or, in telecommunication models for example, might denote signals (requests to a WEB server) or travelling packets of information, or jobs sent to a printer. Arriving customers might also be partitioned apriori and along the way into classes and given different priorities for servicing; this yields *multi-class* models. In some situations, arrivals might even be turned away due to limited capacity or a finite waiting area; this yields *loss* models.

Customers $C_0, C_1, C_2, \ldots$ arrive to the system at random times $t_0 < t_1 < t_2 < \cdots$, require service of random durations, then depart at random times $t_0^d, t_1^d, t_2^d, \ldots$ (not necessarily in the same order as arrival). The total amount of time spent in the system by C_n is defined by $W_n = t_n^d - t_n$ and is called the *sojourn time*. Letting $N(t)$ denote the number of customer arrivals during the time interval $(0, t]$,

$$L(t) = \sum_{n=0}^{N(t)} I\{t_n^d > t\},$$

denotes the *total number of customers in the system at time* t ($I\{A\}$ denoting the indicator function for the event A, equal to 1 if A occurs, 0 if not)

The evolution of the system "state" over time is described by a stochastic process $\{X(t) : t \geq 0\}$ taking values in a general space, and it is usually of interest to also consider the embedded discrete-time process $\{X(t_n-) : n \geq 0\}$ where $X(t_n-)$ is the state as found upon arrival by C_n.

A famous law in queueing theory is *Little's law*,

$$l = \lambda w,$$

where

$$\begin{aligned} l &= \lim_{t\to\infty} \frac{1}{t}\int_0^t L(s)ds \text{ (average number in system)} \\ w &= \lim_{n\to\infty} \frac{1}{n}\sum_{j=0}^{n} W_j \text{ (average sojourn time)} \\ \lambda &= \lim_{t\to\infty} \frac{N(t)}{t} \text{ (average arrival rate).} \end{aligned}$$

These long-run averages are defined (when they exist) almost surely along sample paths, and can also be expressed as expected values with respect to stationary distributions.

Whenever both λ and w exist and are finite, l exists and $l = \lambda w$.

Congestion in a FIFO single-server queue

C_n arrives at time t_n, brings a service time S_n (denoting the length of time required of the server), and waits in the queue if the server is busy. The *delay* of C_n (time spent waiting in the queue (line) before entering service) is denoted by D_n and satisfies the recursion

$$D_{n+1} = (D_n + S_n - T_n)_+, \ n \geq 0, \tag{1}$$

where $T_n = t_{n+1} - t_n$ denotes the n^{th} interarrival time, and $a_+ = \max\{a, 0\}$ is the positive part of a.

Sojourn time: $W_n = D_n + S_n$

The delay and sojourn time processes $\{D_n : n \geq 0\}$, $\{W_n : n \geq 0\}$ *measure congestion from the customers' point of view.*

By imagining that service times represent work for the server and that the server processes the work at rate 1 unit per unit time, we can construct the *workload*
$V(t) =$ the sum of all remaining service times (whole or partial) in the system at time t. Under FIFO, D_n is the work found by C_n; $D_n = V(t_n-)$. Workload can also represent the amount of water in a reservoir in which at time t_n, the amount S_n of water is added to the reservoir, meanwhile water is drained out at constant rate 1 whenever the reservoir is not empty.

The worload process $\{V(t) : t \geq 0\}$ *measures congestion over time from the system's point of view.*

Of intrinsic interest is studying *steady-state* (limiting as time tends to ∞) quantities such as

$$P(D > x) = \lim_{n\to\infty} P(D_n > x),$$

where $P(D_n > x)$, $x \geq 0$ denotes the probability that the delay D_n exceeds x units of time.

Similarly

$$P(V > x) = \lim_{t\to\infty} P(V(t) > x),\ x \geq 0$$

$$P(L > k) = \lim_{t\to\infty} P(L(t) > k),\ k \geq 0$$

$$P(W > x) = \lim_{n\to\infty} P(W_n > x),\ x \geq 0.$$

(Then, $l = E(L) = \sum_{k=0}^{\infty} P(L > k), w = E(W) = \int_0^{\infty} P(W > x)dx$; l and w are expected values.)

If we randomly choose a customer C_n way out in the future, then this customer's delay D_n is distributed as D. If we randomly choose a time t way out in the future, then workload $V(t)$ at this time is distributed as V.

Classical Stochastic Assumptions: Exponential

Arrival times $\{t_n : n \geq 0\}$ form a Poisson process at rate λ (equaivalently, the interarrival-time sequence $\{T_n\}$ is independent and identically distributed (iid) with an exponential distribution: $A(x) = P(T_n \leq x) = 1 - e^{-\lambda x},\ x \geq 0$), and, independently, the service times at a given node $\{S_n\}$ are iid with an exponential distribution at rate μ; $G(x) = P(S \leq x) = 1 - e^{-\mu x},\ x \geq 0$.

$E(T) = 1/\lambda$, $E(S) = 1/\mu$, and $\rho \overset{\text{def}}{=} \lambda/\mu$.

ρ represents the long-run rate at which work arrives to the system.

In the single-server case, this is called the M/M/1 queue. When $\rho < 1$ (stability), then $\{L(t)\}$ forms a positive-recurrent continuous-time *Markov chain*, and

$$P(L \geq k) = \rho^k, \; k \geq 0,$$

and

$$P(D > x) = \rho e^{\mu(1-\rho)x}, \; x \geq 0.$$

Also, V has the same distribution as D via *Poisson arrivals see time averages* (PASTA);

$$P(V > x) = \rho e^{\mu(1-\rho)x}, \; x \geq 0.$$

L has a geometric tail and both $D$and V have an exponential tail.

($W = D + S$ and so $P(W > x) = e^{\mu(1-\rho)x}$, $x \geq 0$, sojourn time is exactly exponential.)

More generally, consider a network with $J \geq 2$ nodes in which customers completing service at node i, independent of the past next join the queue at node j with probability $P_{i,j}$ (Markovian routing), and service times at node i are iid exponential with rate μ_i, and arrivals (from the outside) to node i are Poisson with rate λ_i. ($P_{i,0}$ = probability of departing the system.) $P(L_i > k) = \lim_{t \to \infty} P(L_i(t) > k)$, steady-state number of customers at node i.

Total arrival rates Λ_i (to each node i) can be computed via solving

$$\Lambda_i = \lambda_i + \sum_{j=1}^{J} \Lambda_j P_{j,i},$$

and if $\rho_i \overset{\text{def}}{=} \Lambda_i/\mu_i < 1, \ i = 1, \ldots, J$, (stability) then

$$P(L_1 \geq k_1, L_2 \geq k_2, \ldots, L_J \geq k_J) = \rho_1^{k_1} \rho_2^{k_2} \cdots \rho_J^{k_J},$$

product form geometric steady-state distribution for Jackson Networks.

Classical Stochastic Assumptions: iid finite MGF

Here , arrival times $\{t_n : n \geq 0\}$ form a renewal process at rate λ, that is, the interarrival-time sequence $\{T_n\}$ is independent and identically distributed (iid) with a general distribution: $A(x) = P(T_n \leq x)$. Independently, the service times at a given node $\{S_n\}$ are iid with a general distribution: $G(x) = P(S \leq x)$. But S is assumed to have a finite MGF: for all $\epsilon > 0$ sufficiently small

$$E(e^{\epsilon S}) = \int_0^\infty e^{\epsilon x} dG(x) < \infty.$$

This condition ensures that the tail of S is approximately exponential (or better yet, S might be bounded; e.g. $P(S \leq b) = 1$ for some b):

$$P(S > x) \leq ce^{-\epsilon x},\ x \geq 0,$$

where $c = E(e^{\epsilon S})$. (Chebychev's inequality)

We refer to such an S as being *light-tailed* because $P(S > x)$ tends to 0 fast, at worst like an exponential.

System performance it turns out is also light-tailed whenever S is. For the FIFO single-server queue, for example, it can be shown that $E(e^{\epsilon D}) < \infty$ and

$$P(D > x) \leq ce^{\epsilon x}, \ x \geq 0;$$

D has a tail that is bounded by an exponential tail whenever S is light-tailed.

Similarly there exists a $b > 0$ and $0 < c < 1$ such that

$$P(L > k) \leq ba^{k}, \ k \geq 0;$$

L has a tail that is bounded by a geometric tail whenever S is light-tailed.

With some further regularity conditions enforced, these approximations can be made asymptotically exact:

$$
\begin{aligned}
P(D > x) &\sim ce^{\epsilon x},\ x \to \infty \\
P(L > k) &\sim ba^{k},\ k \to \infty.
\end{aligned}
$$

(Large deviations theory, Cramér-Lundberg approximation, etc.)

where $a(x) \sim b(x)$ means that $a(x)/b(x) \to 1$ as $x \to \infty$.

Finally: Even in complex single-class networks with $J \geq 2$ nodes, system performance such as W is light-tailed whenever all service times are so.