After The Race

The Boston Marathon is a local institution with over a century of history and tradition. The race is run on Patriot's Day, starting on the Hopkinton green and ending at the Prudential Center in Boston's Back Bay, 26.2 miles later. Key milestones along the route are Wellesley College, Heartbreak Hill, Kenmore Square and Commonwealth Avenue. In a typical year as many as ten thousand runners will participate.

Planning for the race presents several logistical challenges for the race organizers. One major challenge is at the finish line. Runners as they cross the line may require a variety of services.

A few will need medical attention for problems ranging from blisters and leg cramps to severe dehydration and possibly heat exhaustion. The less severe problems, like blisters, can be handled by nurse practitioners, and do not demand immediate attention. The more severe can often not wait. A runner with cramps and dehydration will need a cot and oral liquids; a runner in worse shape may need intravenous fluids and possibly even a trip to a nearby emergency room.

Runners who are not in need of medical attention will need services so as to avoid the subsequent need for medical resources. At the finish runners need an area to continue to walk so that their muscles don't cramp. After a while, they will want a place to sit and rest, where they can start the process of replenishing the nutrients and liquids that their bodies lost during the race. They may often need to use rest rooms, as well as phones. Finally, they will need to retrieve their clothes from the storage vans that have traveled from Hopkinton to the Prudential center, and then to find a place to change.

In 1978, 4391 runners started; there were 3872 male finishers and 186 female finishers within 4 hours of the start time, as shown below

Time Window	Male Finishers	Female Finishers	Total Finishers
2:10 – 2:20	33	0	33
2:20 – 2:30	129	0	129
2:30 – 2:40	316	0	316
2:40 – 2:50	629	5	634
2:50 – 3:00	942	24	966
3:00 – 3:10	507	22	529
3:10 – 3:20	475	43	518
3:20 – 3:30	448	33	481
3:30 – 3:40	217	33	250
3:40 – 3:50	117	15	132
3:50 – 4:00	59	11	70

The challenge is to design a service system to handle the runners as they finish the race.

Basic Queueing Theory

Notation

We categorize single-stage queuing systems by a three-component descriptor, A/B/m, where A denotes the distribution of inter-arrival times, B denotes the distribution of service times, and m is the number of servers. The notation M denotes an exponential distribution (Memory-less or Markovian) and G denotes a general distribution (for i. i. d. times, we denote this by GI).

When there is a maximum limit on the number of customers that can be in the system, we describe the queuing system by a four-component descriptor, A/B/m/k, for k being the limit.

Assumptions and Analysis for M/M/1 queue

- Single server

- Memory-less or Markovian arrival process: At any time t, the probability of an arrival in the next instant is independent of all past history (i. e., memory-less). That is, for "small" Δt, the probability that the next arrival occurs in the time interval $(t, t+\Delta t)$ equals $\lambda\Delta t$, where λ is the arrival rate.

  The number of arrivals in a time interval of length τ is a Poisson random variable with mean $\lambda\tau$, and with variance $\lambda\tau$. The interarrival time between successive arrivals is an exponentially-distributed random variable with mean $1/\lambda$, and with variance $(1/\lambda)^2$.

- Memory-less or Markovian service process: Suppose the service process for a job starts at time 0 and has not completed by time t. Then, for "small" Δt, the probability that the service completes in the time interval $(t, t+\Delta t)$ equals $\mu\Delta t$, where μ is the service rate.

  The service time is an exponentially-distributed random variable with mean $1/\mu$, and with variance $(1/\mu)^2$.

- An analysis of this queueing system is based on solving the "equations for motion" or probability transition equations for this system. Let Pr[n, t] denote the probability that there are n jobs or customers in the system at time t. Then the transition equations can be derived from the following general form (for small Δt):

  $$Pr[n, t + \Delta t] = Pr[n, t] * (1 - \lambda\Delta t - \mu\Delta t) + \mu\Delta t * Pr[n+1, t] + \lambda\Delta t * Pr[n-1, t]$$

- Let $\Pi(n) = Pr[n, t = \infty]$ be the steady state probability for queue lengths. Then, key results are as follows:

  $$\Pi(0) = 1 - \rho$$

  $$\Pi(n) = \rho^n (1 - \rho)$$

  where $\rho = \lambda/\mu$ = the utilization level for the queue.

- Steady-state performance measures:

  L = expected number in system $= \rho/(1 - \rho)$

  Q = expected number in queue (not in service) $= \rho^2/(1 - \rho)$

  W = expected waiting time in system $= 1/(\mu - \lambda) = [1/(1 - \rho)] * [1/\mu]$

  D = expected waiting time in queue $= \lambda/\mu(\mu - \lambda) = [\rho/(1 - \rho)] * [1/\mu]$

  Note that $L = \lambda W$, and $Q = \lambda D$; i. e., Little's Law.

- M/M/1 model can be extended to multiple, parallel servers; to finite waiting rooms; to queue-dependent arrival or service rates.

- Model is useful for **gross-level understanding** of congestion effects, and for **examining design or planning tradeoffs**.

Approximation for General Arrival Process, General Service Times (GI/G/1)

- For a GI/G/1 queue, we assume we know the mean and coefficient of variation (std. deviation/mean) for the interarrival times, and the mean and coefficient of variation for the service times (the "inter-completion" time):

$1/\lambda$	mean interarrival time
$1/\mu$	mean service time
SCV_a	squared coefficient of variation for interarrival times
SCV_S	squared coefficient of variation for service times

  Note: SCV = 1 for exponential random variables (as occurs in M/M/1 model)

- We assume that utilization is less than one: $\rho = \lambda/\mu < 1$; and we assume no limit on the queue size.

- A useful approximation is as follows:

  D = expected waiting time in queue = $[\ \rho/(1-\rho)\ ] * [1/\mu] * (SCV_a + SCV_S)/2$

  From this approximation we get the other common performance measures:
  W = expected waiting time in system = $D + 1/\mu$

  L = expected number in system = $\lambda W = \lambda * [D + 1/\mu\ ]$

  Q = expected number in queue (not in service)

  $= \lambda D = [\rho^2/(1-\rho)] * [(SCV_a + SCV_S)/2\ ]$

- To model networks of queues, this model is often used as a building block, where the departure stream from one queue is the arrival stream for another. Here we need approximate the SCV for the departure stream:

  $SCV_d = (1 - \rho^2)\ SCV_a + \rho^2\ SCV_S$

Assumptions and Analysis for M/G/∞ Queue

- Poisson arrivals with arrival rate λ; <u>any distribution</u> for service times, where τ is the mean service time; and an unlimited number of servers.

- The steady-state number of customers in the system (and in service) is a Poisson random variable with mean equal to $\lambda\tau$, and with variance equal to $\lambda\tau$. That is the steady-state probabilities are as follows:

$$\Pr[\#\text{ in system} = n] = (\lambda\tau)^n * e^{-\lambda\tau}/n! \qquad \text{for } n = 0, 1, 2, \ldots.$$

- For $\lambda\tau > 20$ or 30, this distribution is well approximated by a normal distribution.

Assumptions and Analysis for M/G/k/k Queue

- Poisson arrivals with arrival rate λ; <u>any distribution</u> for service times, where τ is the mean service time; k servers; and a limit of at most k customers in the system at any point in time. That is, there is no waiting room, and a customer either enters service upon arrival or is lost due to the system being full (busy).

- The steady-state probabilities for the number of customers in the system (and in service) are as follows:

$$\Pr[\#\text{ in system} = n] = \frac{(\lambda\tau)^n / n!}{\sum_{i=0}^{k} (\lambda\tau)^i / i!} \qquad \text{for } n = 0, 1, 2, \ldots. k$$

- Pr [# in system = k] is the "loss" probability: that is, the probability that a customer arrives and finds the system full.

- If the system is designed to have a very small loss probability, then the M/G/∞ model can be used to approximate the M/G/k/k system.

Assumptions and Analysis for M/M/k Queue

Poisson arrivals with arrival rate λ; exponential service times with mean service rate μ (mean service time = $1/\mu$); k servers. The system utilization is defined as $\rho = \lambda/k\mu$

Expected time in queue,

$$E[D] = \left(\frac{1}{k\mu - \lambda}\right)\left(\frac{(k\rho)^k}{(1-\rho)k!}\right)\pi_0$$

$$= \left(\frac{\rho}{1 - \rho}\right)\left(\frac{1}{\mu}\right)\left(\frac{(k\rho)^{k-1}}{(1-\rho)k!}\right)\pi_0$$

where,

$$\pi_0 = \frac{1}{\frac{(k\rho)^k}{(1-\rho)k!} + \sum_{i=0}^{k-1}\frac{(k\rho)^i}{i!}}$$

Approximation for M/G/k Queue

- A useful approximation is as follows:

  D = expected waiting time in queue

  = expected waiting time in queue for M/M/k queue * $[(1 + SCV_S)/2]$

  From this approximation we can get the other common performance measures.

Applications

- Facility planning for human service systems , e. g., hospitals

- Inventory management for one-for-one policies, e. g., spares provisioning.

- Capacity planning for telemarketing centers

- Fleet sizing for rail cars

- Crew planning for tending multiple machines (e.g., for repair, setup, loading)