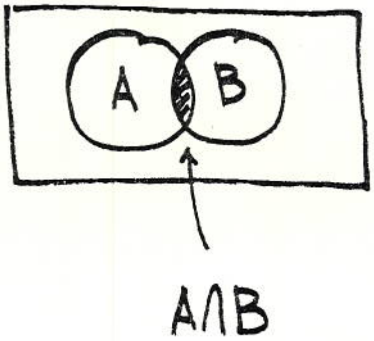
TG Gutowski and SB Gershwin

Principles needed to derive M/M/1 Queue Result.*

1. CONDITIONAL PROBABILITIES



$$P(A/B) = \frac{P(A\cap B)}{P(B)} \tag{1}$$

OR

$$P(A\cap B) = P(A/B)\,P(B) \tag{2}$$

now let B be the universal set, and let it be partitioned by $\mathcal{E}_j$, then

$$\boxed{P(A) = \sum_j P(A|\mathcal{E}_j)\,P(\mathcal{E}_j)} \tag{3}$$

here we have used

$$P(B) = \sum P(\mathcal{E}_j) = 1 \tag{4}$$

$$P(A\cap B) = P(A) \tag{5}$$

* Ref. "Mfg Sys Eng" by Stan Gershwin Prentice Hall 1994

2. MARKOV CHAINS (Interpretation of Eq 3)

$$\underset{\substack{\uparrow \\ \text{state of the} \\ \text{world at time} \\ \text{"t+1"}}}{P(A)} = \sum_j \underset{\substack{\uparrow \\ \text{transition} \\ \text{probabilities}}}{P(A \mid \varepsilon_j)} \underset{\substack{\uparrow \\ \text{state of the} \\ \text{world at} \\ \text{time "t"}}}{P(\varepsilon_j)}$$

Ex. Consider the breaking down of a machine. There are 2 states: machine up or machine down.

$$P(\text{down}, t+1) = P(\text{down now}/\text{was up})\,P(\text{was up}, t) + P(\text{down now}/\text{was down})\,P(\text{was down}, t) \quad \text{—(6)}$$

Mathematical representation: 0 = machine down, 1 = machine up, let P(down now/ was up) = p and P(down now/ was down) = 1, i.e. once down the machine stays down. Then Eq(6) can be written as:

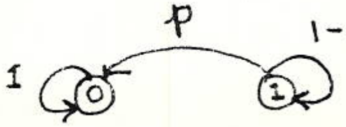
$$P(0, t+1) = p\,P(1,t) + 1\,P(0,t) \quad \text{—(7)}$$

also

$$P(1, t+1) = (1-p)\,P(1,t) + 0\,P(0,t) \quad \text{—(8)}$$

Eqs 7&8 represent the two possible states for the machine at time t+1. This is a rather powerful result, if we know "p" and the current state we can estimate the future! The underlying assumption here is that "p" tells it all. That is, the transition does not depend upon past events. There is no memory effect here. Eqs 7&8 can be written as a single matrix equation

$$\begin{bmatrix} P(0,t+1) \\ P(1,t+1) \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & p \\ 0 & 1-p \end{bmatrix}}_{\text{transition matrix}} \begin{bmatrix} P(0,t) \\ P(1,t) \end{bmatrix} \quad -(9)$$



Note that the columns in the transition matrix add to 1, as well as the column vectors.

3. CONTINUOUS TIME VS DISCRETE TIME

p(t) expand about t for a very small increment δt

Replace t+1 by $t+\delta t$, also the probability that a transition will occur in time interval δt is now written as $p\,\delta t + o(\delta t)$

↖ h.o.t.

h.o.t. = higher order terms are ignored

Also note

$$\frac{dP}{dt} = \lim_{\delta t \to 0} \frac{P(t+\delta t) - P(t)}{\delta t} \quad -(10)$$

Now Eqs 7&8 become

$$P(0,t+\delta t) = p\,\delta t\,P(1,t) + P(0,t) + o(\delta t) \qquad \text{—(11)}$$

$$P(1,t+\delta t) = (1-p\delta t)\,P(1,t) + 0\cdot P(0,t) + o(\delta t) \qquad \text{—(12)}$$

Applying Eq (10) to Eq(12) we can write

$$\boxed{\frac{dP(1,t)}{dt} = -p\,P(1,t)} \qquad \text{—(13)}$$

This is a first order differential eq'n in $P(1.t)$ with solution:

$$P(1,t) = Ae^{-pt}$$

Assuming the machine was originally up

$$P(1,t=0) = 1 = A$$

∴ Our solution is:

$$\boxed{\begin{aligned} P(1,t) &= e^{-pt} \\ P(0,t) &= 1-e^{-pt} \end{aligned}} \qquad \begin{aligned} &\text{—(14)} \\ &\text{—(15)} \end{aligned}$$

The coefficient of $\mathbf{p}(0,t)$ is 0 because, in this system, there is no way of going from 0 to 1. Then,

$$\frac{d\mathbf{p}(1,t)}{dt} = -p\mathbf{p}(1,t). \tag{2.31}$$

The solution to (2.27), (2.31) is

$$\begin{aligned} \mathbf{p}(0,t) &= 1 - e^{-pt}, & (2.32)\\ \mathbf{p}(1,t) &= e^{-pt}. & (2.33) \end{aligned}$$

Recall that once the system makes the transition from 1 to 0 it can never go back, in this model. The probability that the transition takes place in $[t, t+\delta t]$ is

$$\text{prob}\ [\alpha(t+\delta t) = 0 \text{ and } \alpha(t) = 1] = e^{-pt}p\delta t.$$ first term on right Eq 11

The time of the transition from 1 to 0 is said to be *exponentially distributed* with rate p. The expected transition time is $1/p$. The exponential distribution is widely used because of its analytic tractability. In later sections, we typically assume that an operational machine's time to failure is exponentially distributed with parameter p, and that a failed machine's time to repair is exponentially distributed with parameter r.

2.3.4 The meaning of it all

Before we go further, now is a good time to reflect on the meaning of the models and assumptions that we have introduced, and to establish the connection with the real world. Many people find such models to be overly simplistic, not representative of the real complexities that are found in factories and elsewhere. How useful will all these mathematical calisthenics prove to be?

Figure 2.7 is a graph of $e^{-pt}p$, and, superimposed on it, is a set of samples of failure times of a machine. They are organized as a set of bars, so that the height of each bar represents the number of times the failure time fell within the width of that bar. Note that the bars generally follow the exponential curve, but some are above and some are below. Had we

p38 from Mfg Systems Engineering by Stan Gershwin

from Gershwin

taken more samples, and performed sophisticated statistical tests, we could say that, with some confidence, the machine fails according to an exponential distribution, or that it does not.

Note that the exponential density function decreases but does not go to zero as t goes to ∞. That is, prob$[t > T] > 0$ for all $t > 0$. There is a very small probability of a very large outcome.

For most purposes, the details of the shape of the curve are not as important as its gross features. The most important features of a probability distribution are its mean, its variance[3], and its general shape: whether the density has one, two, or more local maximums. If we made a small change to the curve in Figure 2.7 without changing its mean, variance, and general shape, the samples would fit about equally well. Consequently, the effect of the failures of this machine on other machines, material flow, and the overall performance of the system would not be greatly affected.

It is important to remember that we have not postulated any reason why a distribution should be exponential. Certainly it is desirable that it is, because it greatly simplifies the mathematics. Certainly, however, not everything is exponentially distributed. The practitioner must observe the system, and see whether this or any other distribution is at least a plausible representation of the observations. In many cases, the memoryless property of the exponential distribution is a statement of ignorance: the fact that the system has gone so long without a transition gives us no information about whether a transition is any more likely. In many cases, there may be many independent reasons for a transition to occur. Each may be distributed according to some non-exponential distribution, but it may be uncertain whether each will actually occur, or, among those that do, which will occur first. The effect of all this uncertainty may be to produce a distribution close to exponential.

2.3.5 Example: unreliable machine

This is similar to the discrete time, discrete state unreliable machine. A machine can be in two states: up or down. The probability that an operation is completed during an interval $[t, t + \delta t]$ while the machine is up is $\mu\delta t$. The probability that a failure occurs during an interval $[t, t + \delta t]$ while the machine is up is $p\delta t$. The probability that a repair is completed during an interval $[t, t + \delta t]$ while the machine is down is $r\delta t$. What is the long run average production rate of the machine?

The graph of this Markov process is shown in Figure 2.8. Note that the directed links are labeled with probability rates, not probabilities, and that the so-called self-loops, the links leading from a state back to itself, are not drawn. The probability distribution satisfies

$$\mathbf{p}(0, t + \delta t) = \mathbf{p}(0, t)(1 - r\delta t) + \mathbf{p}(1, t)p\delta t + o(\delta t)$$

[3] The *variance* of a scalar random variable x is $E(x - E(x))^2$, where E is the expectation operation. The *standard deviation* is the square root of the variance. In a graph like Figure 2.7, the variance and standard deviation are indicators of how widely spread the density function is. Variance and standard deviation are very important concepts, but they are defined in a footnote because they are not used very much in this book. This is unfortunate because variation and deviation (and their reduction) are very important to manufacturers. They play a small role here because, frankly, the variances of the processes described in this book have been little studied. This neglect should be remedied. See Section 3.2.

→ μ has units operations / time or 1 / mean cycle time

from Gershwin

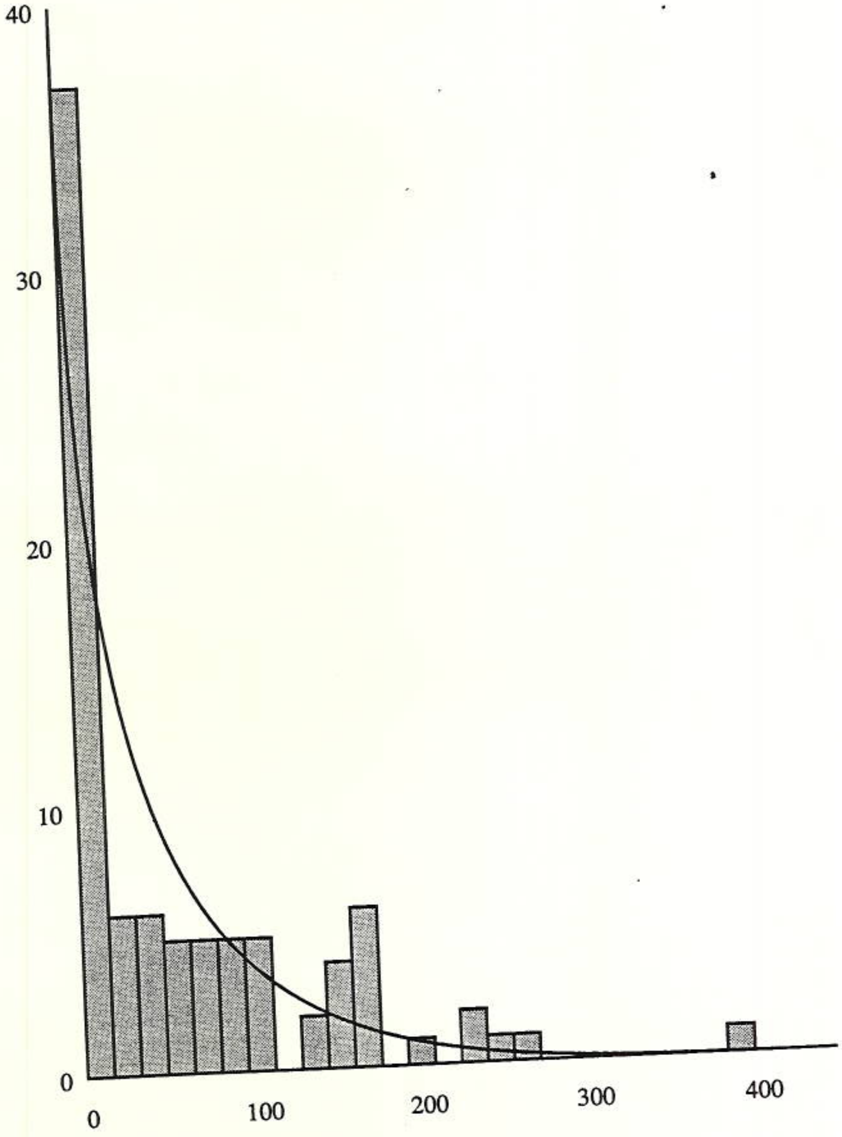


Figure 2.7: Exponential Density Function and Samples

from Gershwin

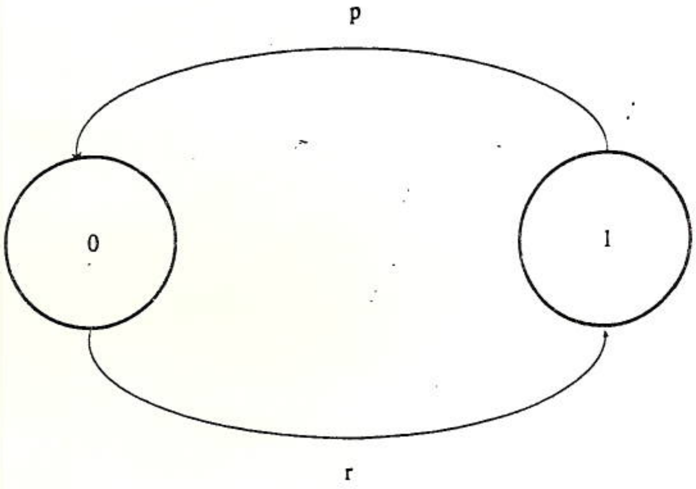


Figure 2.8: Graph of Markov Chain for Continuous Time Unreliable Machine Model

$$\mathbf{p}(1, t+\delta t) = \mathbf{p}(0,t) r \delta t + \mathbf{p}(1,t)(1 - p\delta t) + o(\delta t)$$

or

$$\frac{d\mathbf{p}(0,t)}{dt} = -\mathbf{p}(0,t)r + \mathbf{p}(1,t)p$$

$$\frac{d\mathbf{p}(1,t)}{dt} = \mathbf{p}(0,t)r - \mathbf{p}(1,t)p.$$

The solution is

$$\begin{aligned} \mathbf{p}(0,t) &= \frac{p}{r+p} + \left[\mathbf{p}(0,0) - \frac{p}{r+p}\right] e^{-(r+p)t} && (2.34) \\ \mathbf{p}(1,t) &= 1 - \mathbf{p}(0,t). && (2.35) \end{aligned}$$

As $t \to \infty$, we have

$$\mathbf{p}(0) = \frac{p}{r+p}; \mathbf{p}(1) = \frac{r}{r+p}.$$

The average production rate is $\mathbf{p}(1)\mu$ or $\frac{r\mu}{r+p}$. $= \frac{\mu}{1 + \frac{MTTR}{MTTF}}$

where $r = \frac{1}{MTTR}$, $p = \frac{1}{MTTF}$

from Gershwin

2.3.6 The $M/M/1$ queue

This is the simplest queuing theory model. It has very few assumptions, and they are rarely satisfied in reality. It is a good way to get into the subject, however, because anything more realistic is much more complicated. In spite of its unreality, we can learn something from it.

Consider a queuing system with an infinite amount of storage space. Parts arrive according to a *Poisson process.* That is, the interarrival times are exponentially distributed, which means that if a part arrives at time s, the probability that the next part arrives during the interval $[s+t, s+t+\delta t]$ is $e^{-\lambda t}\lambda\delta t + o(\delta t)$. λ is the *arrival rate.* Similarly, the service times are exponentially distributed, which means that if an operation is completed at time s and the buffer is not empty, the probability that the next operation is completed during the interval $[s+t, s+t+\delta t]$ is $e^{-\mu t}\mu\delta t + o(\delta t)$. μ is the *service rate.*

Let $\mathbf{p}(n,t)$ be the probability that there are n parts in the system at time t. Then,

$$\mathbf{p}(n, t+\delta t) = \mathbf{p}(n-1,t)\lambda\delta t + \mathbf{p}(n+1,t)\mu\delta t + \mathbf{p}(n,t)(1-(\lambda\delta t + \mu\delta t)) + o(\delta t), \quad n > 0 \tag{2.36}$$

and

$$\mathbf{p}(0, t+\delta t) = \mathbf{p}(1,t)\mu\delta t + \mathbf{p}(0,t)(1-\lambda\delta t) + o(\delta t). \tag{2.37}$$

Eq 3

These equations are application of Equation (2.2). In (2.36), A is the event {there are n parts in the system at time $t+\delta t$}, $\mathcal{E}_1$ is the event {there are $n-1$ parts in the system at time t}, $\mathcal{E}_2$ is the event {there are $n+1$ parts in the system at time t}, and $\mathcal{E}_3$ is the event {there are n parts in the system at time t}. In (2.37), A is the event {there is 1 part in the system at time $t+\delta t$}, $\mathcal{E}_1$ is the event {there are no parts in the system at time t}, and $\mathcal{E}_2$ is the event {there is 1 part in the system at time t}. Equations (2.36) and (2.37) become

$$\frac{\partial \mathbf{p}(n,t)}{\partial t} = \mathbf{p}(n-1,t)\lambda + \mathbf{p}(n+1,t)\mu - \mathbf{p}(n,t)(\lambda+\mu), n > 0$$

and

$$\frac{\partial \mathbf{p}(0,t)}{\partial t} = \mathbf{p}(1,t)\mu - \mathbf{p}(0,t)\lambda.$$

If a steady state distribution exists, it satisfies

$$0 = \mathbf{p}(n-1)\lambda + \mathbf{p}(n+1)\mu - \mathbf{p}(n)(\lambda+\mu), n > 0$$

and

$$0 = \mathbf{p}(1)\mu - \mathbf{p}(0)\lambda.$$

Let $\rho = \lambda/\mu$. These equations are satisfied by

$$\mathbf{p}(n) = (1-\rho)\rho^n, n \geq 0 \tag{2.38}$$

the probability that there are "n" parts in the system

- from Gershwin

if $\rho < 1$. The average number of parts in the system is

$$\bar{n} = \sum_n n\mathbf{p}(n) = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}. \tag{2.39}$$

From Little's law, the average delay experienced by a part is

$$W = \frac{1}{\mu - \lambda}.$$

Figure 2.9 is a graph of W as a function of λ, with $\mu = 1$.

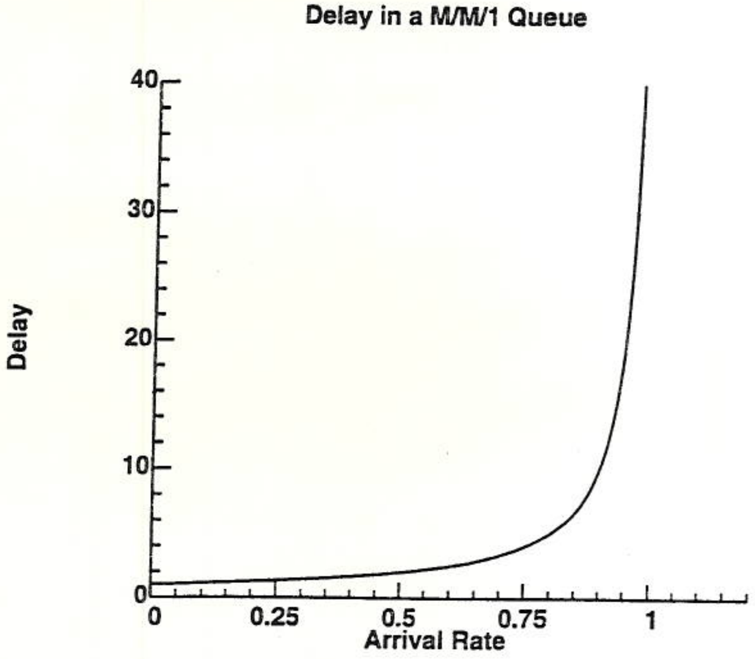


Figure 2.9: Delay *versus* Arrival Rate

2.3.7 Interpretation

The most important characteristic of this system is that the arrival process is not affected by the number of parts in the system, but the departure process is turned off when the buffer is empty.

The condition $\rho < 1$ or $\lambda < \mu$ means that the rate at which parts arrive is less than the rate at which parts can be processed. This means that if, at some time, there happen to be many parts in the system, that number will probably decrease over time. On the other

11

from Gershwin

hand, if the system is empty, a part will arrive sooner or later. If, by chance, more parts than average arrive during a period, the system may accumulate a few parts. Thus the number of parts in the system will increase and decrease, but not get very far from 0 very often.

On the other hand, if $\lambda > \mu$, parts will tend to accumulate in the system. Parts arrive faster than they can be processed, and the arrival mechanism is never turned off. In fact, if the system is started empty at time 0, the number of parts in the system at time t is close to $(\lambda - \mu)t$. The probability of finding the system empty approaches 0. In this case, there is no steady state probability distribution.

The *capacity* of this system is μ. This is the greatest rate at which parts can enter and leave the system. The delay and the average number of parts in the system increase dramatically as the arrival rate approaches the capacity. These quantities are much harder to calculate in other systems, but this behavior is characteristic of all systems with waiting.

There are only two ways of reducing delay: increase the capacity, or change the relationship between throughput and delay. The first approach involves changing the manufacturing process; the second involves scheduling.