International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 01 | Jan -2020 ISSN: 2582-3930

Some Aspects of Queueing Models in Manufacturing Systems

Dr. Deepa Chauhan

Axis Institute of Technology and Management

Abstract

Queueing phenomenon occurs in several real-life congestion situations when resources (machines at a factory, elevators, telephone lines, traffic lights) cannot immediately render the amount or the kind of service required by their users. Queueing theory is an extremely active area of research. One of the key reasons for its strong viability is that, time and again, interesting new questions from production and manufacturing give rise to new and challenging queueing problems.

Manufacturing systems consist of extensive interactions between people, information, materials and machines. Typically one of the main problem is to deal with the diversity of the products. A complicated feature arising in manufacturing systems is that of synchronized resource possession. The manufacturing of a part requires a machine and an operator simultaneously. There is a considerable need to develop a methodology based on queueing theory to handle the delay and blocking issues of manufacturing systems. A successful implementation depends on the ability to eliminate all forms of waste, continuous improvement; employee involvement, disciplined implementation, supplier participation, reorganization of production floor, modular designs, cell layout, process control and total quality creation, etc..

Introduction

The basic processes assumed by most queueing models are the following. Arrivals requiring service are generated over time by an input source. These customers enter the queueing system and join a queue. At certain times, a member of the queue is selected for service by some rule known as the queue discipline. The required service is then performed for the customer by the service mechanism, after which the customer leaves the queueing system. This process is depicted in Fig. 1. Many alternative assumptions can be made about various elements of the queueing process.

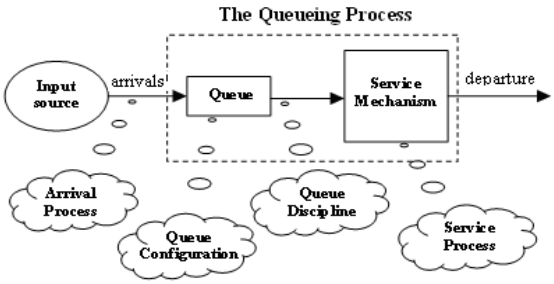


Figure 1: Components of a basic queueing process

We divide a queueing operation broadly into four parts: (i) the arrival, (ii) the waiting line, (iii) the service facility and (iv) the output. In order to study the congestion problems, we can characterize the queueing system as follows:

(A) Arrival process depicts the manner in which customers arrive and join the system. We can describe the arrival process in terms of random variable which can represent either the number of arrivals during a time interval or the time interval between successive arrivals. If customers arrive in groups, their size can be a random variable as well, this case is normally referred to as bulk queues. In queueing models, the customer's arrivals are summarized in terms of probability distributions.

The arrival process depends on the nature of arrivals, capacity of the system and the customer's behavior. In addition to understand customer arrival, one should also understand customer's reneging, balking and jockeying.

Reneging is the act of leaving a queue before being served. **Balking** is the act of not joining a queue upon arrival. Conceptually, the two concepts are the same. The only difference is the timing of when customer leaves, either immediately in the case of balking or later as in the case of reneging.

Jockeying is the act of switching from one queue to another. Reneging, balking and jockeying are three of the most difficult aspects of a queueing system to measure because the customer may never be recorded by the system.

(B) The **service mechanism** refers to the pattern according to which the arrivals are served in a queue. The service mechanism are described by the rate, which is defined by the number of arrivals served per unit time or as a time required to serve the arrival. If the system is empty the service facility is idle. The service time may also be deterministic or probabilistic. The specification of service mechanism includes the number of servers in service facility. The queueing system may consist of a single server, multi server or infinite servers. The service may be provided in batches of fixed or varying batch size. Generally, single server provides service to the customers at a time. There may arise some queueing situations wherein the arrivals are served simultaneously by the same server. The phenomenon in which server depends on the number of arrivals waiting in the system is referred to as **state dependent service.**

The time elapsed from the commencement of the service to its completion for an arrival at a service facility is referred as the **service time** (holding time). A model of a particular queueing system must specify the probability distribution of service times for each server (and possibly for different types of customers), although it is common to assume the same distribution for all the servers. The service-time distribution that is most frequently assumed in practice (largely because it is far more tractable than any other) is the exponential distribution discussed. Other important service-time distributions are the degenerate distribution (constant service time) and the Erlang (gamma) distribution.

(C) Queue discipline is the manner of choosing customers for service from the queue. Arrivals in the queue receive service according to a queueing discipline; some possible queueing disciplines are:

- **First Come First Served (FCFS):** The most common and probably the fairest is the FCFS/FIFO (First In First Out). Arrivals receive service in the order of their arrival. This service discipline may be seen at cinema ticket window, at a patrol pump etc..
- **Last Come First Served (LCFS):** This discipline may be seen in big godowns where the arrivals which come last are taken out (served) first.
- **Round Robin (RR):** Customer receives service in cyclic fashion; each customer receives a fixed amount of service in each cycle unless they complete its total service demand.
- **Random:** When one customer completes his service and other waiting, each of waiting customer is equally likely to be the next one to receive service.
- **Priority:** Some customers are served before the others without considering their order of arrival i.e. some customers are served on priority basis**.**

(D) A queue is characterized by the maximum permissible number of customers that it can contain, known as **system**

capacity. Queues are called infinite or finite, according to whether this number is infinite or finite.

(E) The **service mechanism** consists of one or more service facilities, each of which contains one or more parallel **service channels**, called servers**.** If there are more than one service facility, the customer may receive service from a sequence of these (service channels in series). At a given facility, the customer enters one of the parallel service channels and is completely serviced by that server. A queueing model must specify the arrangement of the facilities and the number of servers (parallel channels) at each one. Most elementary models assume one service facility with either one server or a finite number of servers.

Queueing Process

Performance indices for each model indicate how the corresponding queueing system should perform, including the average amount of waiting that will occur, under a variety of circumstances. Some random variables or families of random variables that arise in the study of queueing system provide important measures of performance and effectiveness of a stochastic queueing system. One needs to have their complete probabilistic description. It will be seen that many queueing processes that we would come across are stochastic in nature. The queueing tactics implement to create a manufacturing system is primarily based on stochastic modeling approach. It is worthwhile to discuss some basic concepts used in queueing modeling.

(a) Stochastic Process

The scope of applications of random variables which are functions of time or space or both has been on the increase. Families of random variables which are functions of time are known as stochastic process (or random process, or random functions). A stochastic process is the counterpart to a deterministic process in probability theory.

A stochastic process is a family of random variables $\{ X\{t\} : t \in T \}$, defined on a given probability space, indexed by the parameter t, where t varies over an index set T.

The values assumed by the random variable X(t) are called states and the set of all possible values forms the state space of the process. Process can be classified in general into the following four types of process:

(i) Discrete time, discrete state space
(ii) Discrete time, continuous state space
(iii) Continuous time, discrete state space
(iv) Continuous time, continuous state space

There are certain important classes of stochastic process that plays important role in the modeling of a manufacturing system such as Markov process Markov chain, birth death process etc..

(b) Markov Process

A Markov process is a probabilistic dynamic system of states in which the future state depends only on the present situation and is independent of the past history. The states of Markov process may be discrete or continuous. More specifically for a discrete state space and discrete parameter stochastic process, X(t) is a Markov process if

$$\Pr\{X(t) = x / X(t_n) = x_n, X(t_{n-1}) = x_{n-1}, \ldots, X(t_0) = x_0\} = \Pr\{X(t) = x / X(t_n) = x_n\}$$

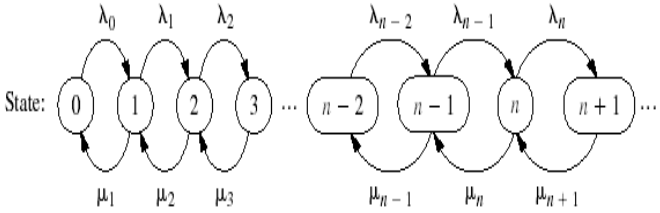
for $t_0 < t_1 < t_2 < t_3 < \ldots\ldots\ldots < t_n < t$.

The exponential distribution plays a fundamental role for representing the distributions of inter arrival and service time to represent queueing system as a continuous time Markov chain.

(c) Birth and Death Process

Most elementary queueing models assume that the inputs (arriving customers) and outputs (leaving customers) of the queueing system occur according to the birth-and-death process. In the context of queueing theory, the term birth refers to the arrival of a new customer into the queueing system, and death refers to the departure of a served customer. Let N(t) be the number of customers in the queueing system at time t. The birth-and-death process describes probabilistically how N(t) changes as t increases. Broadly speaking, it says that individual births and deaths occur randomly, where their mean occurrence rates depend only upon the current state of the system. Because of its assumptions, the birth-and-death process is a special type of continuous time Markov chain. A continuous parameter homogeneous Markov chain {X(t), t≥0} with the state space {0,1,2,….} is known as a birth- death process if there exist constant λ_i (i=0,1,…) and μ_i (i=0,1,…) such that the transition rates are given by

$$q_{i,i+1} = \lambda_i;\ q_{i,i-1} = \mu_i;\ q_i = \lambda_i + \mu_i;\ q_{ij} = 0\ for\ |i-j| > 1$$



The birth and death process

The birth rate λ_i is the rate at which births occur in state i, and the death rate μ_i is the rate at which deaths occur in state i. These rates are assumed to depend only on state and are independent of the time. In a given state births and deaths occur independently of each other and only nearest neighbor transitions are allowed. Such a process is a useful model of many situations in queueing theory and reliability theory.

(e) Poisson Process

Continuous time and discrete state space stochastic processes play an important role in the study of a congestion phenomenon. One such process is Poisson process. The Poisson process is a pure birth process in which the arrival rates are all equal to constant. It is the collection of arrivals for which the inter arrival times are independent and exponentially distributed.

We define the Poisson process as follows. Suppose that events occur successively in time, so that the intervals between successive events are independent and identically distributed according to exponential distribution $F(x) = 1 - e^{-\lambda x}$. Let the number of events in the interval (0, t] be denoted by N(t), then the stochastic process $\{N(t) : t \geq 0\}$ is a Poisson process with mean rate λ and is given by

$$P_n(t) = \Pr[N(t) = n] = \frac{e^{-\lambda t}(\lambda t)}{n!},\ n = 0,1,2,...$$

The Poisson process is an extremely useful process for modeling queueing situations in many practical applications. It is empirically found that in many circumstances the arising stochastic process can be well approximated by Poisson process.

(f) Exponential Process

The operating characteristics of queueing systems are determined largely by two statistical properties, namely, the probability distribution of inter arrival times and the probability distribution of service times. To formulate a queueing model, it is necessary to specify the assumed form of each of these distributions. The most important probability distribution in queueing theory is the exponential distribution.

Suppose that a random variable T represents either inter arrival or service times. This random variable

is said to have an exponential distribution with parameter α if its probability density function is

$$f_T(t) = \begin{cases} \alpha e^{-\alpha t}, for\, t \geq 0 \\ 0 \quad\quad , for\, t < 0 \end{cases}$$

The CDF of exponential distribution is

$$F_T(t) = P\{T \leq t\} = 1 - e^{-\alpha t}$$

The expected value and variance of T are, respectively

$$E(T) = \frac{1}{\alpha}, Var(T) = \frac{1}{\alpha^2}.$$

Performance Measures

Queueing theory studies congestion situations by formulating mathematical model and then using these models to derive various measures of performance. Queueing analysis provides a vital information for effective designing of queueing systems in different frameworks. There are various generic performance measures using which the performance of a service system can be described. Some performance measures for the analysis of queueing models developed in our present study are described as follows:

- **Queue Length**

  Queue length can be defined by the number of customers waiting for service to begin and the number of customers being served.

- **Waiting Time**

  The time spent by a customer in waiting before taken into service including service time is identified as waiting time. Let W_q is the waiting time in the queue excluding service time for each individual customer and assume that mean service time is constant 1/μ, for all n≥1, then waiting time can be given by

$$W = W_q + \frac{1}{\mu}$$

- **Average Queue Length**

  The number of customers in the queue per unit time signifies the average queue length. Let P_n be the probability that there are exactly n customers in the queueing system then average queue length is given by

$$L = \sum_{n=0}^{\infty} nP_n$$

- **Busy and Idle Period**

  Busy period of a server is the time during which he remains busy in servicing. Thus it is the time between the start of service of the customer to the end of service of the last customer in the queue.

  When all the customers in the queue are served, the idle period of the server begins and it continues up to the time of arrival of the customer. The idle period of a server is the time during which he remains free because there is no customer present in the system.

- **Reliability**

  Reliability is mainly concerned with the determination of the probability that a system will operate adequately for a given period of time in its intended application. Let the random variable T be the life time or time to failure of a component. The probability that the component survives until some time t is called the reliability (or the survival function) R(t) of the component. Thus reliability can be defined as

$$R(t) = P(T > t) = 1 - F(t)$$

  where F(t) is the distribution function of the component life time. It is a measure that the device is working during the interval (0,t].

  The reliability of the device is also defined by

$$R(t) = \Pr\{X(u) = 1, 0 \le u \le t : X(0) = 1\}.$$

- **Availability**

  Availability is a combined measure of reliability and maintainability. We define the availability A(t) of a component or a system as the probability that the component is properly functioning at time t. In the absence of a repair or replacement, availability is simply equal to reliability R(t)=1-F(t) of the component. Limiting or stationary availability is defined by $A = \lim_{t\to\infty} A(t)$.

- **Mean Time to Failure**

  Let X denotes the life time of a component so that its reliability $R(t) = P(X > t)$ and $R'(t) = -f(t)$. Then the expected life or the mean time to failure (MTTF) of the component is given by

$$E[X] = \int_0^\infty t f(t)dt = -\int_0^\infty t R'(t)dt.$$

Since R(t) approaches zero more faster than t approaches ∞, we have

$$E[X] = \int_0^\infty R(t)dt$$

- **Throughput**

  Throughput can be defined as mean number of customers whose service is completed in single unit of time.

Some performance indices are model specific. For M/M/m/N/N machine repair problem, let m be the number of identical repairmen. Let P(n) be the steady state probability that there are n failed machines in the system and λ denotes the failure rate. Then certain performance measures can be defined as

(i) Expected number of failed machines in the system is

$$L_s = \sum_{n=0}^{N} nP(n)$$

(ii) Expected number of failed machines in the queue is

$$L_q = \sum_{n=m}^{N} (n-m)P(n)$$

(iii) Expected waiting time in the system is

$$W_s = \frac{L_S}{\lambda_{eff}} = \frac{\sum_{n=0}^{N} nP(n)}{\lambda_{eff}}$$ where

$\lambda_{eff} = \lambda(1 - P(N))$.

(iv) Expected waiting time in the queue is

$$W_q = \frac{L_q}{\lambda_{eff}} = \frac{\sum_{n=m}^{N} (n-m)P(n)}{\lambda_{eff}}$$

(v) Expected number of machines being repaired is

$$N_S = L_S - L_q = \sum_{n=0}^{N} nP(n) - \sum_{n=m}^{N} (n-m)P(n)$$

The performance measures would facilitate us to compute and investigate various system metrics that would help in improving the efficiency of the manufacturing system.

Some Aspects of Queueing Model

Queueing theory is concerned with mathematical modeling and analysis of system that provide service to random demands. A queueing model is an abstract description of such a system. A queueing model represents the system's physical configuration and arrangement of the servers, which provide service to the customers and stochastic (i.e. probabilistic or statistical) nature of the demand, by specifying the variability in the arrival process and in the service process. The art of applied queueing theory is to construct a model that is simple enough so that

it yields to mathematical analysis, yet contains sufficient detail so that its performance measures reflect the behavior of the real system. In this thesis, we have developed several queueing models of manufacturing systems in different frameworks. The topics covered in our investigations are machine repair problems (see chapter 5), bulk arrival queues (chapters 6-8), phase repair (chapters 2 and 5), optional services (chapters 5, 6 and 7), Bernoulli schedule vacation (chapter 8), inventory management (chapter 9), flexible manufacturing system (chapter 10), etc.. It is worthwhile to discuss some aspects of these congestion problems from modeling perspective view points.

Machine Repair Problem

It is essential for manufacturing companies to manage their resources and obtain high percentage of availability of resources efficiently. Whenever a machine breaks down, it will result in a great loss to the organization. So the machine repair problems are very important problems in the queueing theory. When a machine breaks down, the mechanic/ repairman starts its repairing. If a machine breaks down and repairman is not available, then it will be attended after the repair of other failed machine is completed. Thus the broken machines form a queue and wait for their repair. The production system may not operate with full capacity during the period of breakdown and this may lead to the loss of production. Availability of the manufacturing system can be increased by having spare machines in order to replace the failed machines.

A machine repair problem consisting of 'M' online machines, 'Y' cold spares and 'S' warm spares under the supervision of unreliable server who repairs the machines as they fails, can be modeled as follows:

The following notations are used in the formulation of the machine repair problem.

λ_0 (α_0)	Failure rate of the operating (warm spare)units.
b	Service rates of repairman
ε_0	The rate at which server goes for vacation on finding the system empty
μ_0, μ_1, μ_2	Service rate of the server when both cold and warm spares are available, the cold spares are exhausted and both cold as well as warm spares are exhausted respectively
a_0	Rate of server breakdown

Let 'i' represent the number of failed machines in the system and 'j' represent the status of the server such that, 0,1,2 denote the server states when server is idle, busy, broken down, respectively.

$P_{i,j}$=Pr{there are i; 0<i<L (L=M+Y+S), failed units in the system with server being in the j^{th} state}.

The differential equation governing the model can be formulated as by considering the state transition flow as shown in fig. 1.6.

$$P_t'(0,0) = -[M\lambda_0 + S\alpha_0]P(0,0) + \varepsilon P(0,1)$$

$$P_t'(i,0) = -[M\lambda_0 + S\alpha_0]P(i,0) + [M\lambda_0 + S\alpha_0]P(i-1,0), \quad 1 \le i \le Y$$

$$P_t'(i,0) = -[M\lambda_0 + (Y+S-i)\alpha_0]P(i,0) + [M\lambda_0 + (Y+S-\overline{i-1})\alpha_0]P(i-1,0), \; Y+1 \le i \le Y+S-1$$

$$P_t'(Y+S,0) = -[M\lambda_0]P(Y+S,0) + [M\lambda_0 + \alpha_0]P(Y+S-1,0), \quad i = Y+S$$

$$P_t'(i,0) = -[(L-i)\lambda_0]P(i,0) + [(L-\overline{i-1})\lambda_0]P(i-1,0), \quad Y+S+1 \le i \le L-1$$

$$P_t'(L,0) = -P(L,0) + \lambda_0 P(L-1,0)$$

$$P_t'(0,1) = -[M\lambda_0 + S\alpha_0 + a_0 + \varepsilon]P(0,1) + \mu_0 P(1,1) + bP(0,2)$$

$$P_t'(Y,1) = -[M\lambda_0 + S\alpha_0 + a_0 + \mu_1]P(Y,1) + [M\lambda_0 + S\alpha_0]P(Y-1,1) + \mu_1 P(Y+1,1) + bP(Y,2)$$

$$P_t'(i,1) = -[M\lambda_0 + S\alpha_0 + a_0 + \mu_0]P(i,1) + [M\lambda_0 + S\alpha_0]P(i-1,1) + \mu_0 P(i+1,1) + bP(i,2), \quad 1 \le i \le Y-1$$

$$P_t'(i,1) = -[M\lambda_0 + (Y+S-i)\alpha_0 + a_0 + \mu_1]P(i,1) + [M\lambda_0 + (Y+S-\overline{i-1})\alpha_0]P(i-1,1) + \mu_1 P(i+1,1) + bP(i,2)$$

$$P_t'(Y+S-1,1) = -[M\lambda_0 + \alpha_0 + a_0 + \mu_1]P(Y+S-1,1) + [M\lambda_0 + (S-2)\alpha_0]P(Y+S-2,1) + \mu_2 P(Y+S,1) + bP(Y+S-1,2)$$

$$P_t'(Y+S,1) = -[M\lambda_0 + a_0 + \mu_2]P(Y+S,1) + [M\lambda_0 + \alpha_0]P(Y+S-1,1) + \mu_2 P(Y+S+1,1) + bP(Y+S,2)$$

$$P_t'(i,1) = -[(L-i)\lambda_0 + a_0 + \mu_2]P(i,1) + [(L-\overline{i-1})\lambda_0]P(i-1,1) + \mu_2 P(i+1,1) + bP(i,2), \; Y+S+1 \le i \le L-1$$

$$P_t'(L,1) = -[\mu_2 + a_0]P(L,1) + \lambda_0 P(L-1) + bP(L,2)$$

$$P_t'(0,2) = -[M\lambda_0 + S\alpha_0 + b]P(0,2) + a_0 P(0,1)$$

$$P_t'(i,2) = -[M\lambda_0 + S\alpha_0]P(i,2) + [M\lambda_0 + S\alpha_0]P(i-1,2) + a_0 P(i,1), \quad 1 \le i \le Y$$

$$P_t'(i,2) = -[M\lambda_0 + (Y+S-i)\alpha_0 + b]P(i,2) + [M\lambda_0 + (Y+S-\overline{i-1})\alpha_0]P(i-1,2) + a_0 P(i,1), \quad Y+1 \le i \le Y+S-1$$

$$P_t'(Y+S,2) = -[M\lambda_0 + b]P(Y+S,2) + [M\lambda_0 + \alpha_0]P(Y+S-1,2) + a_0 P(Y+S,1)$$

$$P_t'(i,2) = -[(L-i)\lambda_0 + b]P(i,2) + [(L-\overline{i-1})\lambda_0 P(i-1,2) + a_0 P(i,1)$$

$$P_t'(L,2) = -(\lambda + b)P(L-1,2) + a_0 P(L,1)$$

After evaluating the transient probabilities using classical methods such as Laplace transform and matrix method or using numerical method (e.g. Runge-Kutta method), we can obtain the various system performance characteristics in terms of the transient probabilities as

- The expected number of failed units in the system E{S(t)}

$$E\{S(t)\} = \sum_{j=0}^{3} \sum_{i=0}^{L} iP_t(i, j)$$

- The expected number of failed units in the queue E{Q(t)}

$$E\{Q(t)\} = \sum_{j=0}^{3} \sum_{i=0}^{L} (i-1)P_t(i, j)$$

- Probability that the server is in idle state I(t)

$$I(t) = \sum_{i=0}^{L} P_t(i,0)$$

- Probability that the server is in breakdown state D(t)

$$D(t) = \sum_{i=0}^{L} P_t(i,2)$$

- Probability that the server is in busy state B(t)

$$B(t) = 1 - [I(t) + D(t)]$$

Bulk Arrival Queue

In the earlier years of the development of queueing theory, the studies were confined to single arrival and individual or personalized service system. The queueing systems in which pattern of arrival or service or both are in bulk (or batches or groups) are known as bulk queueing systems. The actual service time distribution frequently deviates greatly from exponential form, particularly when the service requirements of the customers are quite similar. Therefore, it is important to have other queueing model that uses alternative distribution. The following are the notations used in the construction of the bulk arrival queueing model.

λ	Mean arrival rate of customers
X	Random variable denoting the batch size
X(z)	Generating function for batch size X
α	Failure rate of the server
μ	Service rate of the server
μ(x), β(y)	Hazard service rate and repair rate of the server
b(x), r(y)	Probability density functions for service time and repair time
B(x), R(y)	Distribution functions of service time and repair time
$P_n(t,x)$	Probability that there are n customers in the queue at time t and elapsed service time lies in (x, x+dx)
$R_n(t,x,y)$	Joint probability that there are n customers in the queue at time t when the server is in repair state and the elapsed service time for the customer under service is equal to x, elapsed repair time lies in (y, y+dy)

Hazard rates are

$$\mu(x)dx = \frac{dB(x)}{1-B(x)} \text{ and } \beta(y)dy = \frac{dR(y)}{1-R(y)}$$

We construct the partial differential equations governing the model for the system and assume the elapsed service time, elapsed setup time and the elapsed repair time as supplementary variables:

$$\left(\frac{d}{dt}+\lambda\right)Q(t)=\int_0^\infty P_n(t,x)\mu(x)$$

$$\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial x}+\mu(x)+\lambda+\alpha\right)P_n(t,x)=\lambda\sum_{i=1}^{n}a_iP_{n-i}(x)+\int_0^\infty R_n(t,x,y)\beta(y)dy$$

$$\left(\frac{\partial}{\partial t}+\frac{\partial}{\partial y}+\beta(y)+\lambda\right)R_n(t,x,y)=\lambda\sum_{i=1}^{n}a_iR_{n-i}(t,x,y)$$

The following boundary conditions are taken into consideration:

$$P_n(t,0)=\int_0^\infty P_{n+1}(t,x)\mu(x)dx+\lambda Q(t)$$

$$R_n(t,x,0)=\alpha P_n(t,x)$$

By applying supplementary variable approach and generating function method, the above set of equations can be solved to get steady state probabilities.

Steady state probabilities of server being idle, busy and under repair are $Q=1-\rho\left(1+\frac{\alpha}{\beta}\right)$, $P=\rho$ and $R=\frac{\rho\alpha}{\beta}$, respectively.

In order to analyze reliability indices, we consider breakdown states as absorbing states.

(i) The steady state availability of the server is given by

$$A=Q+P=1-\rho\frac{\alpha}{\beta}$$

(ii) The steady state failure frequency of the server is given by

$$M_f=\lim_{z\to 1}\int_0^\infty\left(\alpha P_q^*(s,z)\right)dx=\alpha\rho$$

(iii) The mean time to the first failure (MTTFF) of the server is given by

$$MTTFF=\int_0^\infty R(t)dt=R^*(s),\qquad \lim_{s\to 0}sQ^*(s)=Q$$

Therefore $MTTFF=Q^*(0)+\rho\frac{\alpha b^*(\alpha)}{(\alpha+\mu)b^*(\alpha)}$,

where $\rho=\frac{E[X]\lambda}{\mu}$.

References

1. Bitran, G.R. and Sarkar, D. (1994): Throughput analysis in manufacturing networks, Euro. J. Oper. Res., Vol. 74 (3), pp. 448-465.
2. Brandimarte, P. and Villa, A. (1996): Advance Models For Manufacturing Systems Management, Boca Raton, FL: CRC Press.
3. Cagliano, R. and Spina, G. (2000): Advance manufacturing technologies and strategically flexible production, J. Oper. Manag., Vol. 18 (2), pp. 169-190.
4. Coffman, E.G., Gelenbe, E. and Gilbert, E.N. (1988): Analysis of a conveyor queue in a flexible manufacturing system, Euro. J. Oper. Res., Vol. 35 (3), pp. 382-392
5. Diallo, M., Pierrevel, H. and Quilliot, A. (2001): Manufacturing cell design with flexible routing capability

in presence of unreliable machines, Int. J. Prod. Eco., Vol. 74 (1-3), pp. 175-182.

6. Jain, M., Rakhee and Maheshwari, S. (2004): N-policy for machine repair system with spares and reneging,

Appl. Math. Model., Vol. 28, pp. 513-531.

7. Jain, M., Sharma, G.C. and Sharma, R. (2008): Performance modeling of state dependent system with mixed standbys and two modes of failure, App. Math. Model., Vol. 32 (5), pp. 712-724