Department of Systems Engineering
George Mason University

SYST 302: Systems Methodology and Design II #6

Kuo-Chu Chang
Fairfax, Virginia

Queuing Theory and Analysis

- Concepts and Introduction
- Monte Carlo Analysis of Queuing
- Single-Channel Queuing Models
- Multiple-Channel Queuing Models
- Finite Population Queuing Models

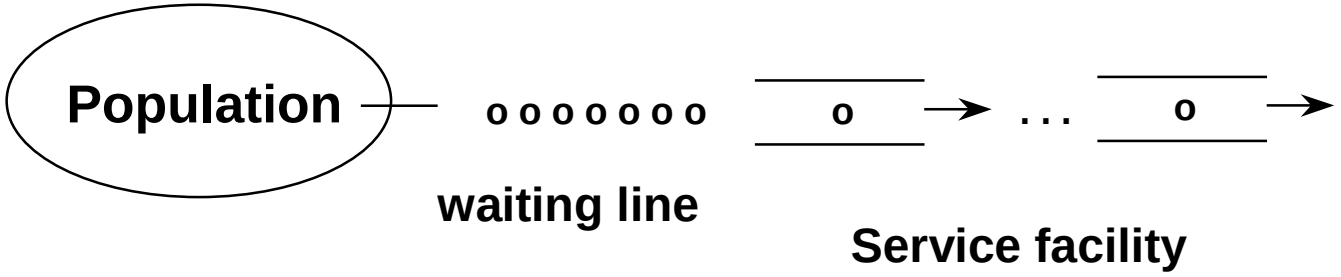
Introduction

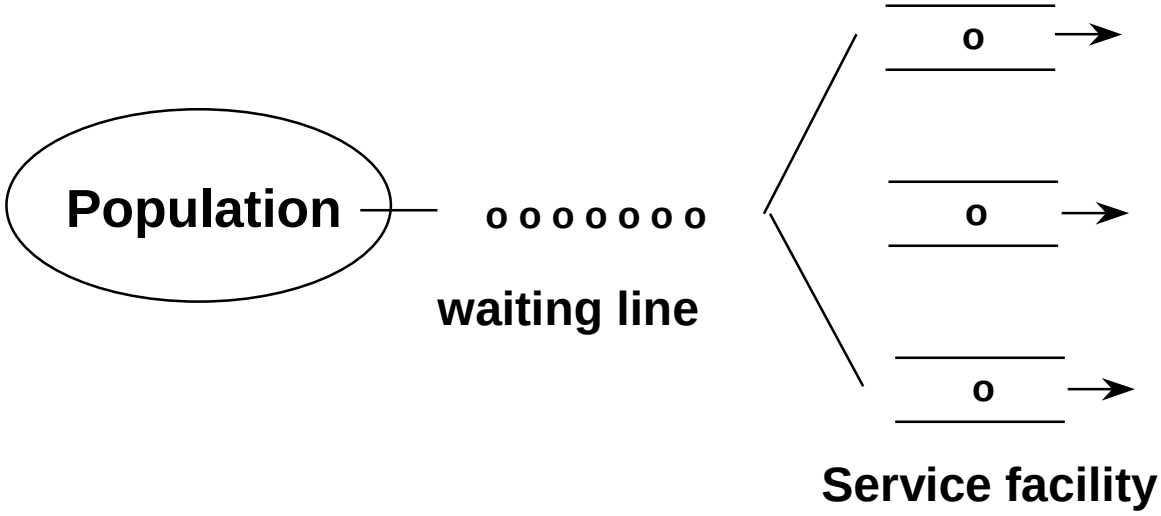
- **The Queuing System (Waiting-Line)**
  - A facility or a group of facilities for service
  - A population of individuals or units which form a line to be served with a pre-determined waiting-line discipline
  - Single-channel or multiple-channel
  - Single-stage or multiple-stage
- **Common Examples**
  - The public forms waiting-lines at grocery stores, theaters, doctor's office, etc.
  - Items in process produces a waiting-line at each machine center in a production system
  - Automobiles form waiting-lines at toll-gates, traffic signals, docks in a transportation system

The Queuing System

- **Objective:** **determine the capacity of the service facility in the light of the relevant costs and the characteristics of the arrival patterns so that the overall cost associated with the queuing system is minimized**
  - **The arrival mechanism: finite or infinite population, arrival pattern may be time-dependent**
  - **The waiting line: form a queue with a certain discipline such as first come/first serve, relative urgency, or first come/last serve. Waiting cost is incurred.**
  - **The service mechanism: discrete process provides service for units in line, can be single channel or multiple channel with a certain capacity. The cost is facility dependent.**
  - **The decision model: decide a policy of service capacity to meet the demand at a minimum cost**

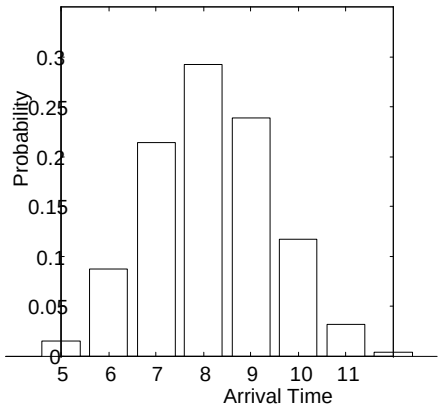
Some Examples

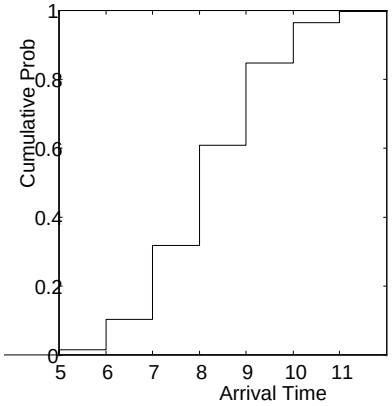


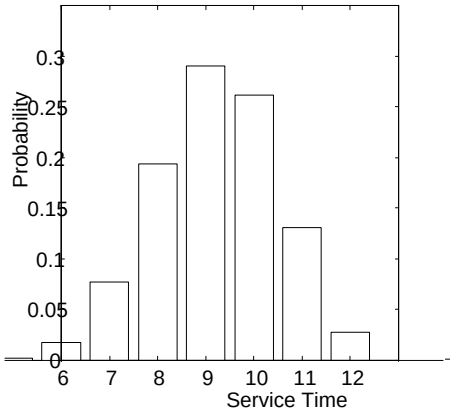


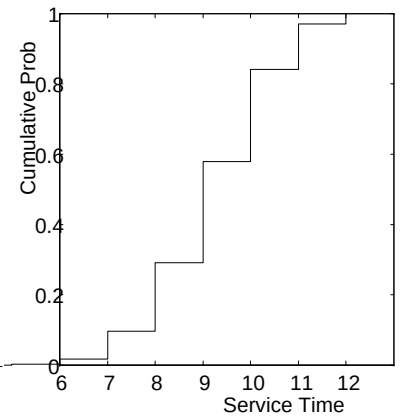
Monte Carlo Analysis

- **The Mechanism**
  - **Uncertain Waiting-line decision models usually based on assumptions of the arrival and service-time distributions**
  - **Formal mathematic solutions may be difficult or impossible in general**
  - **Monte Carlo analysis does not require these distributions to obey certain theoretical forms**
  - **Provides analysis tool through simulation**

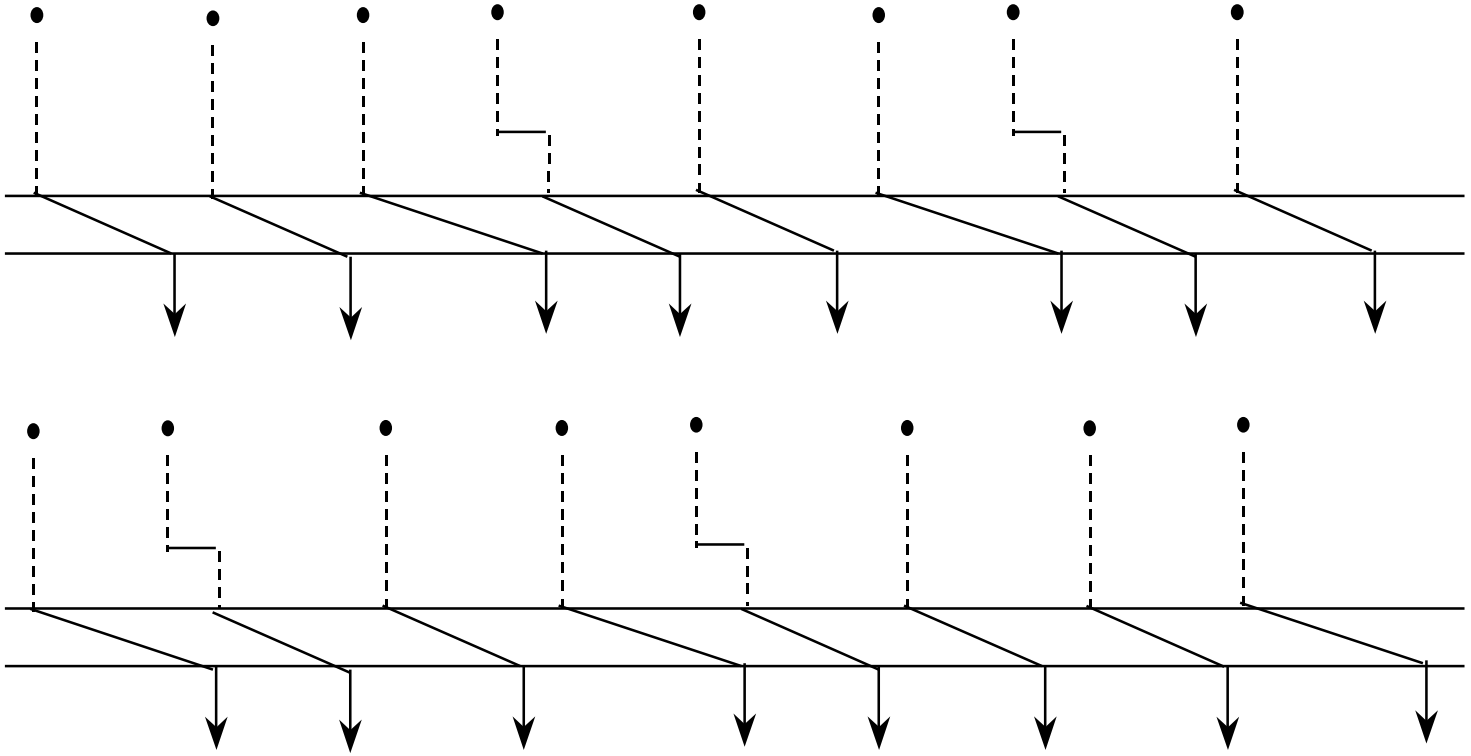








Single Channel Queuing MC Analysis



Economic Analysis:

Total cost = unit-waiting-cost-per-time-period * (waiting-time-in-queue + waiting-time-in-service)
+ service-cost-per-time-period * service-time
Ex: $9.6 * (23 + 337) + $16.10 * 400 = $9,896

Queuing Models

- **Queuing Models**
  - **A/B/C: Arrival/Service/Number of Servers**
  - **M/M/1: Markov (Poisson) / Markov (Exponential) / one server**
  - **M/M/S: Multiple servers**
  - **M/G/1: General service time distribution**
  - **M/D/1: Deterministic service time**

Single-Channel Queuing Models

- **Model Assumptions**
  - **Infinite population pool**
  - **Arrival per period is a Poisson distribution**
  - **Service time is an exponential distribution**
  - **Provides theoretical analysis tool**
- **System Parameters**

λ : expected number of arrival per time period

μ : expected number of service completions per time period

n : number of units in the system at time t

$P_n(t)$: probability of n units in the system at time t

$\lambda\Delta t$: probability that an arrival occurs between time t and $t+\Delta t$

$\mu\Delta t$: probability that a service completion occurs between t and $t+\Delta t$

$$\text{Note}: \int_t^{t+\Delta t} \mu e^{-\mu t} dt = -e^{-\mu t}\Big|_t^{t+\Delta t} = (1-\mu t) - (1-\mu t - \mu\Delta t) = \mu\Delta t$$

Probability of n Units in the System

$$P_n(t+\Delta t) = \{P_n(t)[1-\lambda\Delta t][1-\mu\Delta t]\} + \{P_{n+1}(t)[1-\lambda\Delta t][\mu\Delta t]\} + \{P_{n-1}(t)\lambda\Delta t[1-\mu\Delta t]\}$$
$$= P_n(t) - (\lambda+\mu)P_n(t)\Delta t + \lambda\mu P_n(t)\Delta t^2 + \mu P_{n+1}(t)\Delta t - \lambda\mu P_{n+1}(t)\Delta t^2$$
$$+ \lambda P_{n-1}(t)\Delta t - \lambda\mu P_{n-1}(t)\Delta t^2$$

ignoring Δt^2 terms, we have

$$\lim_{\Delta t\to 0}\frac{P_n(t+\Delta t)-P_n(t)}{\Delta t} = \frac{d}{dt}P_n(t) = -(\lambda+\mu)P_n(t) + \mu P_{n+1}(t) + \lambda P_{n-1}(t)$$

but $P_0(t+\Delta t) = \{P_0(t)[1-\lambda\Delta t]\} + \{P_1(t)[1-\lambda\Delta t][\mu\Delta t]\}$

$$\lim_{\Delta t\to 0}\frac{P_0(t+\Delta t)-P_0(t)}{\Delta t} = \frac{d}{dt}P_0(t) = -\lambda P_0(t) + \mu P_1(t)$$

In steady state,

$$\left.\begin{aligned}\frac{d}{dt}P_n(t) = 0 = -(\lambda+\mu)P_n(t) + \mu P_{n+1}(t) + \lambda P_{n-1}(t)\\ \frac{d}{dt}P_0(t) = 0 = -\lambda P_0(t) + \mu P_1(t) \Rightarrow P_1(t) = \tfrac{\lambda}{\mu}P_0(t)\end{aligned}\right\} \Rightarrow P_n = \left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n$$

Probability of n Units in the System

Let $\rho = \frac{\lambda}{\mu}$ (average number of units being served)

$$\Rightarrow P_{n+1}(t) + (1+\rho)P_n(t) + \rho P_{n-1}(t) = 0 \Rightarrow P_n = c_1 + c_2\rho^n$$

But $P_1 = \rho P_0$

$$\Rightarrow c_1 + c_2 = P_0,\ c_1 + c_2\rho = P_1 = \rho P_0 \Rightarrow c_1 = 0,\ c_2 = P_0 \Rightarrow P_n = P_0\rho^n$$

$$\text{But } \sum_{n=0}^{\infty} P_n = 1 = P_0 \sum_{n=0}^{\infty} \rho^n = P_0 \frac{1}{1-\rho} \Rightarrow P_0 = 1-\rho$$

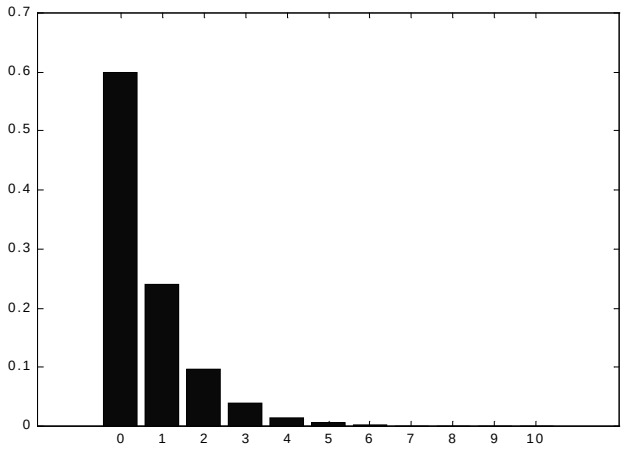
$$\Rightarrow P_n = (1-\rho)\rho^n = \left(1 - \frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n$$

Ex:

$\lambda = 0.1$

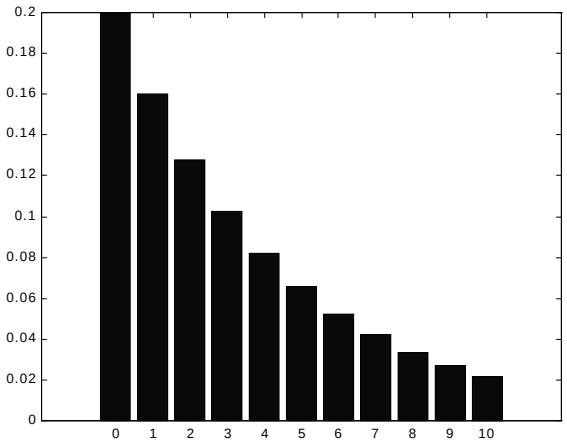
$\mu = 0.25$

$\rho = \frac{\lambda}{\mu} = 0.4$



Ex:

$\rho = 0.8$



Other Performance Measures

Expected number of units in the system:

$$\overline{n} = \sum_{n=0}^{\infty} nP_n = \sum_{n=0}^{\infty} n(1-\rho)\rho^n = (1-\rho)\sum_{n=0}^{\infty} n\rho^n = (1-\rho)\rho\frac{\partial}{\partial\rho}\sum_{n=0}^{\infty}\rho^n$$

$$= (1-\rho)\rho\frac{\partial}{\partial\rho}\frac{1}{1-\rho} = (1-\rho)\rho\frac{1}{(1-\rho)^2} = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}$$

Average length of the queue:

$\overline{m}$ = avg. number of units in the system - avg. numbers of units being serviced

$$= \frac{\lambda}{\mu-\lambda} - \frac{\lambda}{\mu} = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

Probability of more than k units in the system:

$$P[\geq k \text{ in system}] = \sum_{n=k}^{\infty} P_n = \sum_{n=k}^{\infty}(1-\rho)\rho^n = (1-\rho)\rho^k\left\{\sum_{n'=0}^{\infty}\rho^{n'}\right\} = \rho^k$$

Note: $\rho = \lambda/\mu$ is the average number of units being serviced

Example

$$P_n = \left(1-\frac{\lambda}{\mu}\right)\left(\frac{\lambda}{\mu}\right)^n = (1-\rho)\rho^n, \lambda = \frac{1}{10} = 0.1, \mu = \frac{1}{4} = 0.25$$

Expected number of units in the system :

$$\overline{n} = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda} = 0.667$$

Average length of the queue :

$$\overline{m} = \frac{\lambda^2}{\mu(\mu-\lambda)} = 0.267$$

Average length of the nonempty queue:

$$\overline{m}_{m>0} = \frac{\overline{m}}{P(m>0)} = \frac{\overline{m}}{P[\geq 2 \text{ in the system}]} = \frac{0.267}{\rho^2} = \frac{0.267}{0.16} = 1.667$$

Waiting Time

Distribution of waiting time w:

$P(w=0) = P\{0 \text{ unit in the system}\} = P_0 = 1-\lambda/\mu$

for w > 0, ICBST (Sec. 10.3)

$$f(w) = \lambda(1-\lambda/\mu)e^{-(\mu-\lambda)w} \Rightarrow \overline{w} = \int_0^\infty wf(w)dw = \frac{\lambda}{\mu(\mu-\lambda)}$$

Average total time spent in the system :

T = average waiting time + average service time

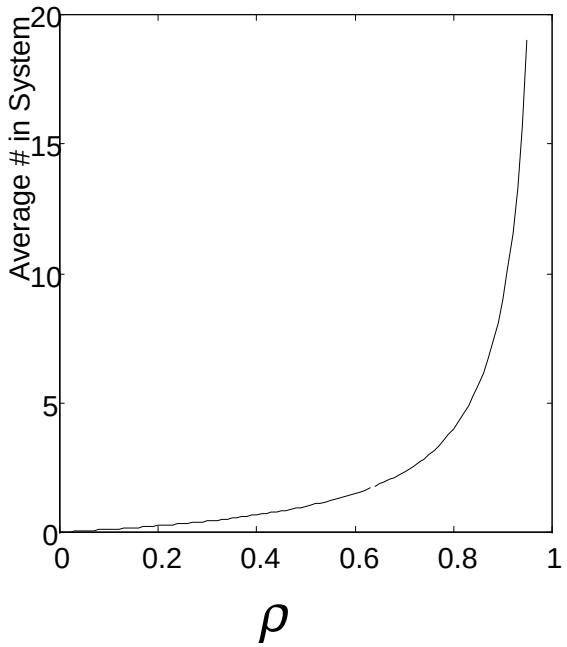
$$= \frac{\lambda}{\mu(\mu-\lambda)} + \frac{1}{\mu} = \frac{1}{\mu-\lambda}$$

In fact, $\overline{n} = T\lambda \Rightarrow T = \overline{n}/\lambda = \frac{\lambda}{\mu-\lambda}\frac{1}{\lambda} = \frac{1}{\mu-\lambda} = \frac{1/\mu}{1-\rho}$

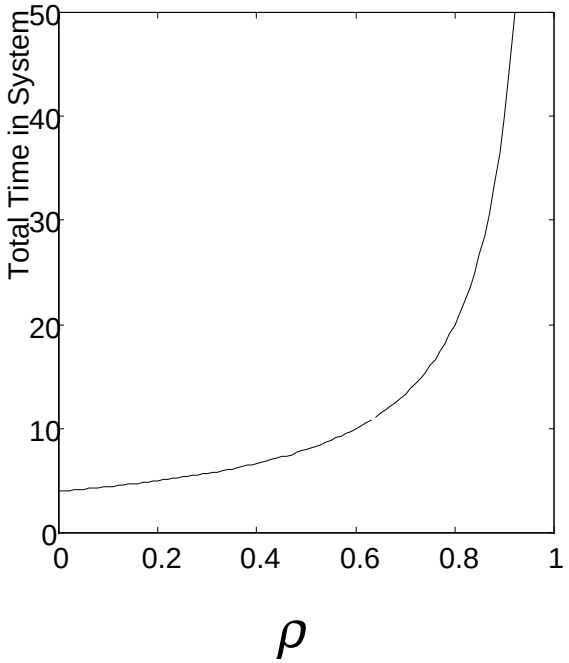
In the previous example, $T = 0.667/(1/10) = 6.67$

System Performance Curves

$$\bar{n} = \frac{\rho}{1-\rho}$$



$$T = \frac{1/\mu}{1-\rho}$$



Example

Q : A large processor handles transactions at a rate of $K\mu$ transactions per second. Suppose transactions arrival according to a Poisson process of rate $K\lambda$ transactions per second, and that transactions require an exponentially distributed amount of processing time. Suppose that a proposal is made to replace the large processor with K processors each with processing rate μ and arrival rate λ. Compare the mean delay performance of the existing and the proposed systems.

A : The large processor is an M/M/1 queue with arrival arte $K\lambda$, service rate $K\mu$, and the utilization $\rho = K\lambda / K\mu = \lambda / \mu$. The mean delay is given by

$$E[T] = \frac{1/K\mu}{1-\rho}$$

Each of the small processors is an M/M/1 system with arrival rate λ, service rate μ, and utilization $\rho = \lambda / \mu$. The mean delay is

$$E[T'] = \frac{1/\mu}{1-\rho} = KE[T]$$

$\Rightarrow$ The concentration of customer demand into a single system results in significant delay performance improvement.

Cost Analysis
per unit time

Expect total cost = expected waiting cost + expected facility cost

$$TC = C_w \bar{n} + C_f \mu = C_w \frac{\lambda}{\mu - \lambda} + C_f \mu$$

Minimize cost:

$$\frac{\partial TC}{\partial \mu} = 0 \Rightarrow \mu = \lambda + \sqrt{\frac{\lambda C_w}{C_f}}$$

where C_w: waiting cost per time period per unit

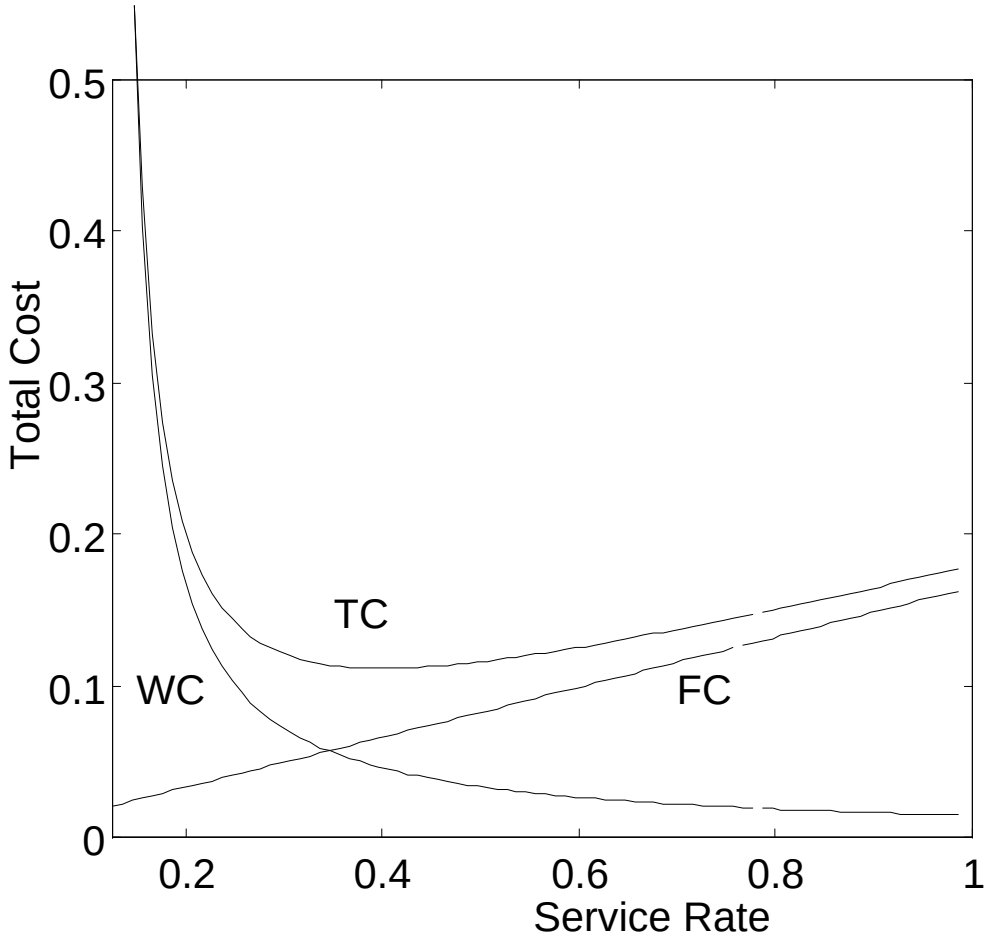
C_f: facility cost per unit serviced

Example:

$$\lambda = 1/8,\ C_w = \$0.10,\ C_f = \$0.165$$

$$\Rightarrow \mu = 0.4,\ TC = 0.1115$$

Cost Curves



Multiple-Channel Queuing Models

- **Model Assumptions**
  - **Same as single channel**
  - **Multiple service facility, # of channel = c**

Let $\rho = \lambda/c\mu$ ***,then***

$$P_{c,n} = P_{0,0}\left(\frac{\lambda}{\mu}\right)^{c}\frac{1}{c!}\rho^{n},\ P_{m,0} = P_{0,0}\left(\frac{\lambda}{\mu}\right)^{m}\frac{1}{m!}$$

$$P_{0,0} = \frac{1}{(\lambda/\mu)^{c}(1/c!)[1/(1-\rho)] + \sum_{r=0}^{r=c-1}(\lambda/\mu)^{r}(1/r!)}$$

other measures such $\overline{n}, \overline{m}$ ***,and*** $\overline{w}$ ***can also be obtained (see Sec. 10.4), where*** $P_{m,n}$ ***is the probability that there are n units waiting in queue and m channels are busy***

Note : $0 \le m \le c$ ***and*** $P_{m,n} = 0$ ***for*** $m \ne c$ ***and*** $n \ne 0$

Queuing with Nonexponential Service

- **Model Assumptions**
  - **Poisson arrival**
  - **Nonexponential service time**

For any service time distribution, if σ^2 is its variance, then

$$\bar{n} = \frac{(\lambda/\mu)^2 + \lambda^2\sigma^2}{2[1-(\lambda/\mu)]} + \frac{\lambda}{\mu} \quad \text{and} \quad T = \frac{(\lambda/\mu^2) + \lambda\sigma^2}{2[1-(\lambda/\mu)]} + \frac{1}{\mu}$$

For constant service time, $\sigma^2 = 0$

$$\bar{n} = \frac{(\lambda/\mu)^2}{2[1-(\lambda/\mu)]} + \frac{\lambda}{\mu} \quad \text{and} \quad T = \frac{(\lambda/\mu)}{2\mu[1-(\lambda/\mu)]} + \frac{1}{\mu}$$

For exponential service time, $\sigma^2 = (1/\mu)^2$

$$\bar{n} = \frac{\lambda}{\mu-\lambda} \quad \text{and} \quad T = \frac{1}{\mu-\lambda}$$