Module 7: Introduction to Queueing Theory (Notation, Single Queues, Little's Result)

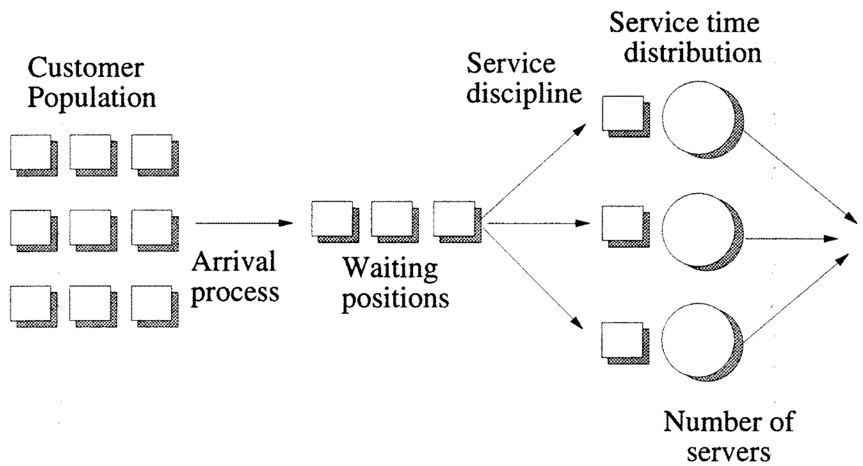
(Slides based on Daniel A. Reed, ECE/CS 441 Notes, Fall 1995, used with permission)

Outline of Section on Queueing Theory

1. Notation
2. Little's Result
3. Single Queues
4. Solutions for networks of queues - Product Form Results (on blackboard, not slides)
5. Mean value analysis (if time permits)

Queueing Theory Notation

- Queuing characteristics
  - Arrival process
  - Service time distribution
  - Number of servers
  - System capacity
  - Population size
  - Service discipline
- Each of these is described mathematically
  - Descriptions determine tractability of (efficient) analytic solution
  - Only a small set of possibilities are solvable using standard queueing theory



Arrival Processes

- Suppose jobs arrive at times t_1 , t_2 , ... ,t_j
  - Random variables $\tau_j = t_j - t_{j-1}$ are *inter-arrival times*
  - There are many possible assumptions for the distribution of the τ_j
  - Typical assumptions for the τ_j:
    - Independent
    - Identically distributed
  - Many other possible assumptions:
    - Bulk arrivals
    - Balking
    - Correlated arrivals
- For *Poisson* arrival, the inter-arrival times are:
  - IID (independent and identically distributed)
  - exponentially distributed (i.e., CDF $F(x) = 1 - e^{-x/a}$)
- Other common arrival time distributions include
  - Erlang, Hyper-exponential, Deterministic, General (with a specified mean and variance)

Other Queue Features

- Service time
  - Interval spent actually receiving service (exclusive of waiting time)
  - As with arrival processes, there are many possible assumptions
  - Most common assumptions are
    - IID random variables
    - exponential service time distribution
- Number of servers
  - Servers may or may not be identical
  - Service discipline determines allocation of customers to servers
- System capacity
  - Maximum number of customers in the system (including those in service)
  - May be finite or infinite
- Population size
  - Total number of potential customers
  - May be finite or infinite

Other Queue Features (Continued)

- Service discipline
  - The order waiting customers are serviced
  - Many possibilities, including
    - First-come-first-serve (FCFS), the most common
    - Last-come-first-serve (LCFS)
    - Last-come-first-serve preempt resume (LCFS-PR)
    - Round robin (RR) with finite quantum size
    - Processor sharing (PS) --- RR with infinitesimal quantum size
    - Infinite server (IS)
  - Almost anything might be used, depending on the the total state of the queue
  - As expected, service discipline affects the nature of the stochastic process that represents the behavior of the queueing system

Queueing Discipline Specification

- Queueing discipline is typically specified using Kendall's notation ($A/S/m/B/K/SD$), where
  - Letters correspond to six queue attributes
    - A: interarrival time distribution
    - S: service time distribution
    - m: number of servers
    - B: number of buffers (system capacity)
    - K: population size
    - SD: service discipline
- Interarrival and service time specifiers
  - M exponential
  - E_k Erlang with parameter k
  - H_k hyperexponential with parameter k
  - D deterministic
  - G general (any distribution, mean and variance used in the solution)
- Bulk arrivals or service are denoted by a superscript
  - $M^{[x]}$ denotes exponential arrivals with group size x
  - x is generally a random variable with separately specified distribution
- Omitted specifiers assume certain defaults
  - infinite buffer capacity
  - infinite population size
  - FCFS service discipline

Example Queueing Discipline Specifications

- $M/D/5/40/200/FCFS$
  - Exponentially distributed interarrival times
  - Deterministic service times
  - Five servers
  - Forty buffers (35 for waiting)
  - Total population of 200 customers
  - First-come-first-serve service discipline
- $M/M/1$
  - Exponentially distributed interarrival times
  - Exponentially distributed service times
  - One server
  - Infinite number of buffers
  - Infinite population size
  - First-come-first-serve service discipline

An Introductory Example

- Given these descriptions, what are examples of their application?
- Consider a typical bank
  - 5 tellers
  - Customers form a single line and are serviced FCFS
  - Excluding a run on the bank, the waiting room is effectively infinite
  - For a large bank, the population is effectively infinite
  - Bulk arrivals are possible if friends arrive together for service
- What about service time and inter-arrival time distributions?
  - We can go measure them with a watch at the bank
  - Or, we can make mathematically simplifying assumptions
  - Latter is most common and exponential distribution is typical
- Combining these facts and assumptions
  - $M/M/1$ queue
  - As we shall see, the mean queue length (including one in service) for an $M/M/1$ queue is

$$\frac{\lambda}{\mu - \lambda}$$

  - Where
    - λ is the mean inter-arrival time
    - μ is the mean service time

Notation and Basic "Facts"

- Standard variable names
  - τ is job interarrival time
  - $\lambda = 1/E[\tau]$ mean job arrival rate
  - s is service time per customer (job)
  - m is number of servers
  - $\mu = 1/E[s]$ is mean service rate per server
  - $n = n_q + n_s$ is number of jobs in the system
  - n_q is number of jobs waiting to receive service
  - n_s is number of jobs in service
  - r is response time (service time plus queueing delay)
  - w is waiting time (queueing delay only)
- System must be "stable" to have an interesting steady state solution
  - Number of jobs in the system is finite
  - Requires the relation $\lambda < m\mu$ hold unless
    - the population is finite (queue length is bounded)
    - the buffer capacity is finite (arrivals are lost when queue is full)
    - (in these cases, system is always stable)

Notation and Basic "Facts"

- Number of jobs in the system
  - $n = n_q + n_s$ (jobs are either waiting or in service)
  - $E[n] = E[n_q] + E[n_s] \quad (\text{or } \bar{n} = \bar{n}_q + \bar{n}_s)$
  - and, if the service rates are independent of queue length
    - $Cov(n_q, n_s) = 0$
    - $Var[n] = Var[n_q] + Var[n_s]$
- Number and time
  - $r = w + s$ (response time is the sum of queueing delay and service)
  - but, r, w, and s are random variables, so $\bar{r} = \bar{w} + \bar{s}$
  - and, if the service rates are independent of queue length
    - $Cov(w,s) = 0$
    - $Var[r] = Var[w] + Var[s]$

Little's Law

- Very important result -- Part of the queueing folk literature for the past century
- Formal proof due to J. D. C. Little (1961)
- Relates mean queue length to arrival rate and mean response time
- Mathematically (in seady state),

$$\bar{n} = \lambda \bar{r}$$

- Applies to any "black box" queue under the following assumptions
  - System is work conserving
  - Number of jobs entering is same as number leaving (system is stable)
- Also applies to any transient interval, without requirement that system be stable.
- Note that these are very general conditions, and can apply for any system ("black box") in which customers leave and enter subject to the above constraints.
- An intuitive proof...

Little's Law (Continued)

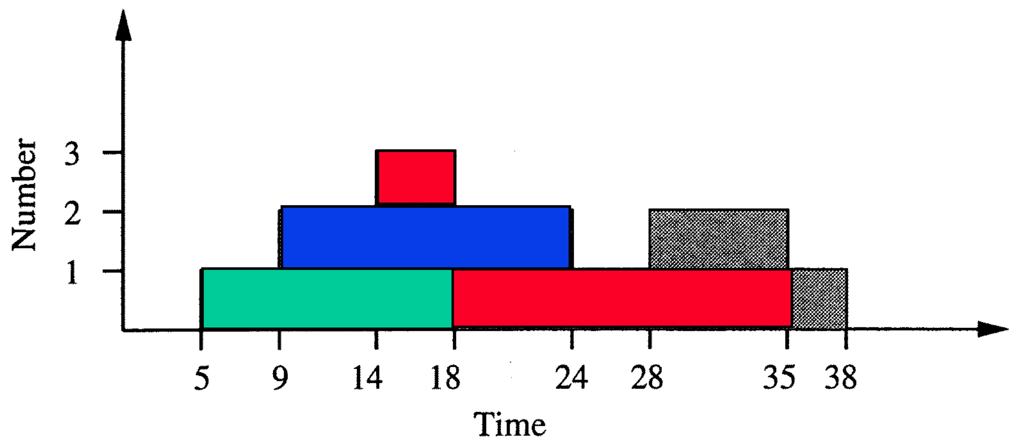
- Sketch of proof (of steady-state case):
  - During a long interval, arrivals ≈ departures (else no stability)
  - Area under the curve is total job time units (jobs x time)
  - Mean queue length $\bar{n}$ is average curve height (area/time)
  - Mean time in system $\bar{r}$ is area/arrivals
  - Mean arrival rate is arrivals/time
- A very general result:
  - No assumptions about arrival or service processes
  - Holds for any queueing discipline (simply charge the area differently)

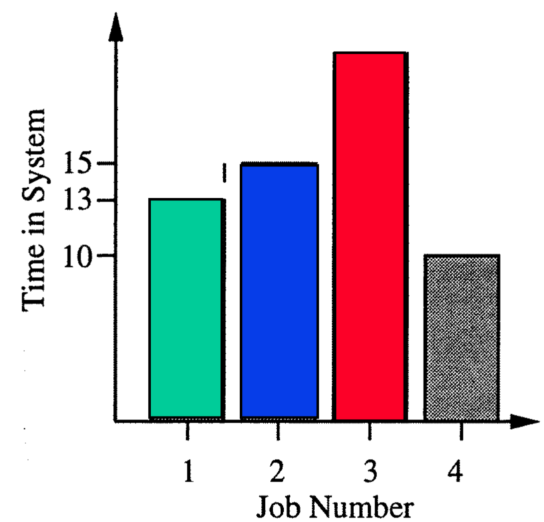
arrival rate

$$\frac{\text{jobs x time}}{\text{time}} = \frac{\text{jobs}}{\text{time}} \times \frac{\text{jobs x time}}{\text{jobs}}$$

Avg number in system

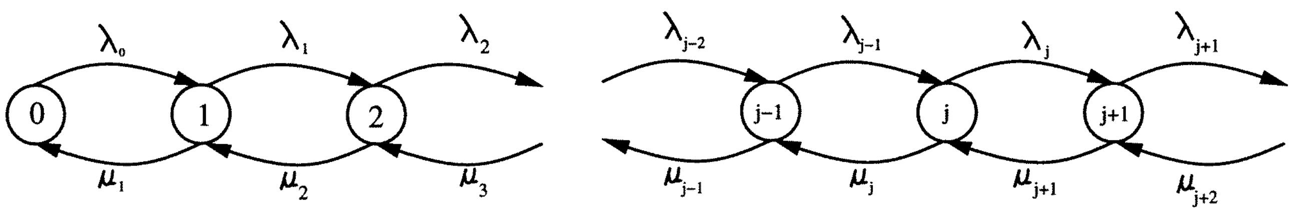
Avg time in system





Analysis of Single Queues

- Plan:
  - Start with one of the simplest queues, an $M/M/1$
  - Model as a "birth-death" process
  - Generalize result to other types of queues



- A *birth-death process* is a Markov process in which states are numbered a integers, and transitions are only permitted between "neighboring" states.
- Steady state solution of a birth death process (Kleinrock, Queueing Systems, vol. 1):
  - (*Theorem*) steady state probability p_n of being in state n is

$$p_n = \frac{\lambda_0 \lambda_1 \ldots \lambda_{n-1}}{\mu_1 \mu_2 \ldots \mu_n} p_0 \qquad\qquad n = 1, 2, \ldots, \infty$$

  - where p_0 is the probability of being in state 0
- Now for a proof ...

Birth-Death (Steady-State) State Occupancy Proof

- If stable, in the steady state (by Markov process solution described earlier)

$$0 = \lambda_{j-1} p_{j-1} - (\mu_j + \lambda_j) p_j + \mu_{j+1} p_{j+1}$$ Flow balance at state j

or

$$p_{j+1} = \left(\frac{\mu_j + \lambda_j}{\mu_{j+1}} \right) p_j - \frac{\lambda_{j-1}}{\mu_{j+1}} p_{j-1} \qquad j = 1,2,3,\ldots$$

and

$$p_1 = \frac{\lambda_0}{\mu_1} p_0$$

- And the solution is...

$$p_n = \frac{\lambda_0 \lambda_1 \ldots \lambda_{n-1}}{\mu_1 \mu_2 \ldots \mu_n} p_0$$

$$= p_0 \prod_{j=0}^{n-1} \frac{\lambda_j}{\mu_{j+1}} \qquad n = 1,2,\ldots,\infty$$

Birth-Death (Steady-State) State Occupancy Proof, cont.

- Finally, because

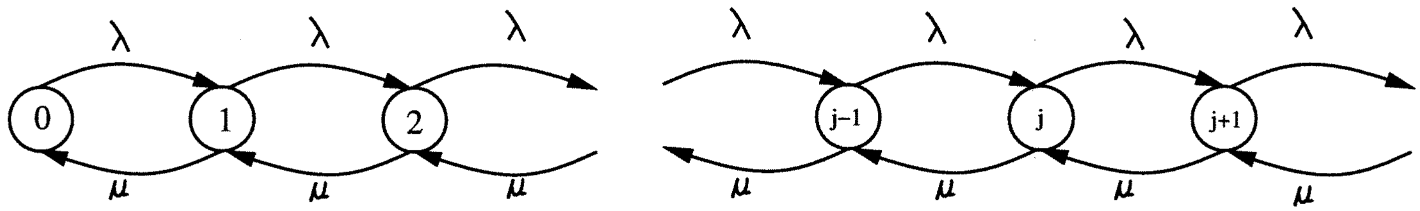
$$\sum_{j=0}^{\infty} p_j = 1$$

we have

$$p_0 = \frac{1}{1 + \sum_{n=1}^{\infty} \prod_{j=0}^{n-1} \frac{\lambda_j}{\mu_{j+1}}}$$

M/M/1 Queue Analysis

- $M/M/1$ is a special case of a birth - death process
  - $\lambda_i = \lambda_j$ for all i and j
  - $\mu_i = \mu_j$ for all i and j



- By simplification

$$p_n = \left(\frac{\lambda}{\mu}\right)^n p_0 \qquad n = 1,2,\ldots,\infty$$

- By tradition, the ratio

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

is called the "traffic intensity" and

$$p_n = \rho^n p_0$$

and

$$p_0 = \frac{1}{1+\rho+\rho^2+\ldots+\rho^\infty} = 1-\rho$$

- By substitution

$$p_n = (1-\rho)\rho^n \qquad n = 0,1,2,\ldots\infty$$

M/M/1 Queue Analysis (Continued)

- Utilization U is simply $1 - p_0 = \rho$
- Mean queue length $E[n]$ (or $\overline{n}$)

$$\overline{n} = \sum_{n=1}^{\infty} n p_n$$

$$= \sum_{n=1}^{\infty} n(1-\rho)\rho^n$$

“almost” mean of a geometric random variable---factor out a rho first

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

- Variance of number of jobs in the system

$$Var[n] = E[n^2] - (E[n])^2$$

$$= \left(\sum_{n=1}^{\infty} n^2 (1-\rho)\rho^n \right) - (E[n])^2$$

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

- Probability of n or more jobs in the system

$$\sum_{j=n}^{\infty} p_j = \sum_{j=n}^{\infty} (1-\rho)\rho^j = \rho^n$$

M/M/1 Queue Analysis (Continued)

- Mean response time $\bar{r}$ (or R) via Little's Law

  $$\bar{n} = \lambda \bar{r}$$

  yields

  $$\bar{r} = \frac{\bar{n}}{\lambda} = \frac{\rho}{(1-\rho)\lambda} = \frac{1}{\mu - \lambda}$$

  where the response time approaches ∞ as $\lambda \to \mu$

- CDF of response time is

  $$F(r) = 1 - e^{-r\mu(1-\rho)}$$

- Mean number of jobs in the queue $E[n_q]$ (or $\bar{n}_q$)

  $$\bar{n}_q = \sum_{n=1}^{\infty} (n-1) p_n = \frac{\rho^2}{1-\rho}$$

M/M/1 Queue Example

- Consider the following queue
  - $\lambda = 0.3$
  - $\mu = 0.5$
- We can calculate the following statistics
  - utilization U

$$U = \rho = \frac{\lambda}{\mu} = \frac{0.3}{0.5} = 0.6$$

  - mean number of jobs in the system $\overline{n}$

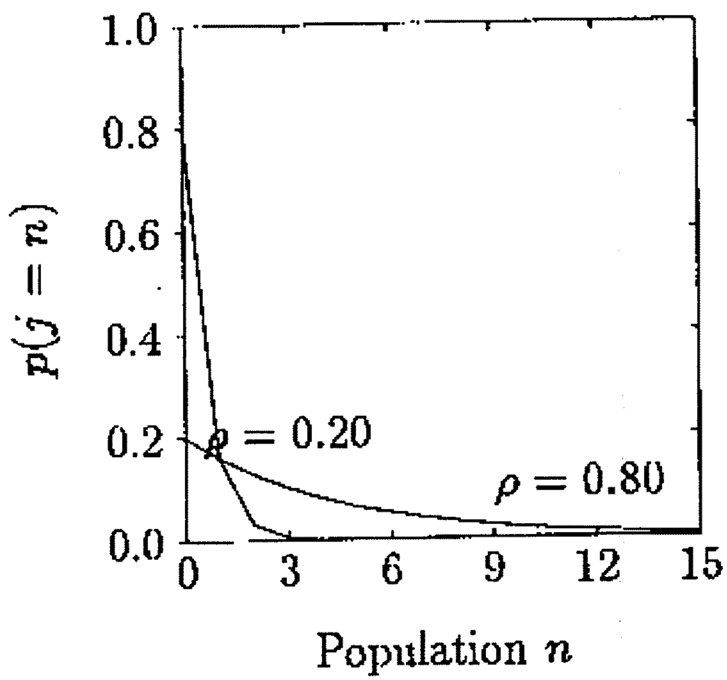
$$\overline{n} = \frac{\rho}{1-\rho} = \frac{0.6}{0.4} = 1.5$$

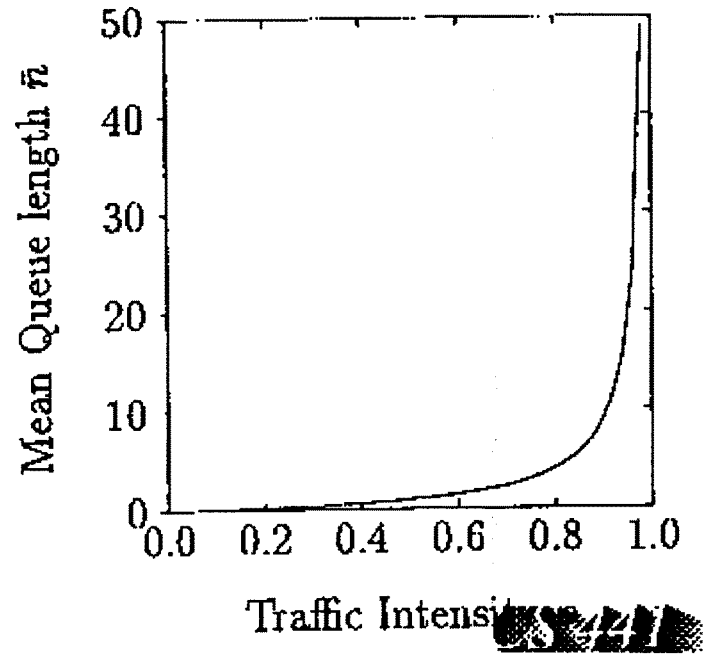
  - mean response time $\overline{r}$

$$\overline{r} = \frac{1}{\mu - \lambda} = \frac{1}{0.2} = 5.0$$

M/M/1 Queue Example (Continued)

- Consider changing λ
  - hold μ fixed at 0.5
  - examine changes in performance metrics





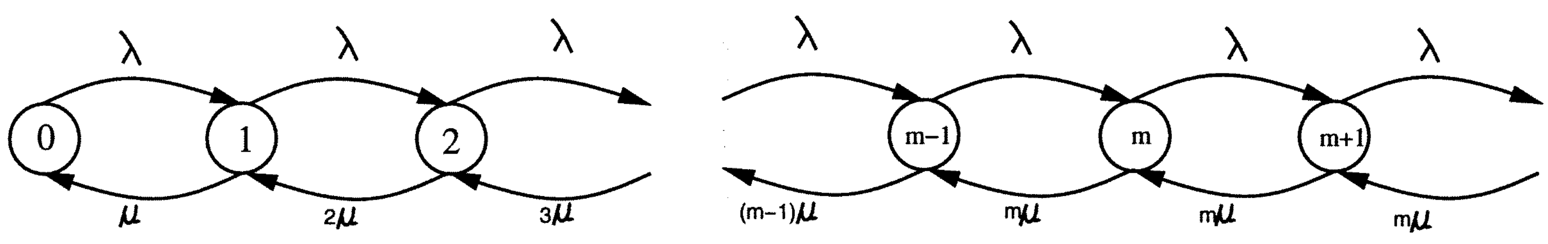
M/M/m Queues

- $M/M/m$ queues
  - m servers rather than one server
  - Reasonable model of
    - a bank queue with multiple tellers
    - a shared memory multiprocessor
- Assumptions
  - m servers
  - All servers have the same service rate μ
  - Single queue for access to the servers
  - Arrival rate λ
  - Formally

$$\lambda_n = \lambda \qquad n = 0, 1, \ldots, \infty$$

$$\mu_n = \begin{cases} n\mu & n = 0, 1, \ldots m-1 \\ m\mu & n = m, m+1, \ldots \infty \end{cases}$$

- What are the state occupancy probabilities?



M/M/m Queues (Continued)

- State occupancy probabilities
  - Just another birth - death process
  - Recall general form of the probability occupancies earlier

$$p_n = \frac{\lambda_0 \lambda_1 \ldots \lambda_{n-1}}{\mu_1 \mu_2 \ldots \mu_n} p_0 \qquad n = 1,2,\ldots,\infty$$

- By simple substitution of the λ_j and μ_j, we have

$$p_n = \begin{cases} \dfrac{\lambda^n}{n!\mu^n} p_0 & n = 1,2,\ldots,m-1 \\ \dfrac{\lambda^n}{m!m^{n-m}\mu^n} p_0 & n = m, m+1,\ldots,\infty \end{cases}$$

$M/M/m$ Queues (Continued)

or equivalently (with $\rho = \lambda/(m\mu)$)

$$p_n = \begin{cases} \dfrac{(m\rho)^n}{n!} p_0 & n = 1, 2, \ldots, m-1 \\ \dfrac{\rho^n m^m}{m!} p_0 & n = m, m+1, \ldots, \infty \end{cases}$$

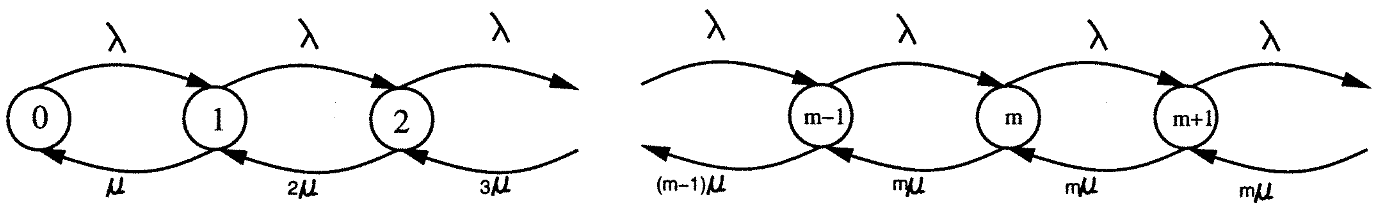
- And, because

$$\sum_{n=0}^{\infty} p_n = 1$$

we have

$$p_0 = \left[1 + \frac{(m\rho)^m}{m!(1-\rho)} + \sum_{n=1}^{m-1} \frac{(m\rho)^n}{n!} \right]^{-1}$$

M/M/m Queues (Continued)



- In a similar manner to that for the M/M/1 queue,
  - We can derive the "standard" measures (queue length, utilization, response time, etc.)
  - You should do these derivations yourself
- Mean number of jobs in the system $\bar{n} = \bar{n}_q + \bar{n}_s$

$$\bar{n} = m\rho + \frac{\rho\varsigma}{1-\rho}$$

M/M/m Queues (Continued)

where

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

$$\varsigma = P(\geq m\ jobs) = \sum_{n=m}^{\infty} p_n = \frac{(m\rho)^m}{m!(1-\rho)} p_0$$

observe that ς is

- the probability an arriving job must queue

- also known as Erlang's C formula

- Expected number of jobs in service $\overline{n}_s$

$$\overline{n}_s = \sum_{n=1}^{m-1} np_n + \sum_{n=m}^{\infty} mp_n = m\rho$$

M/M/m Queues (Continued)

- Utilization of each server
  - m servers
  - $m\rho$ mean jobs in service
  - individual server utilization must be ρ
- Mean response time $\overline{r} = \overline{w} + \overline{s}$ (just apply Little's law)

$$\overline{r} = \frac{\overline{n}}{\lambda}$$

$$= \frac{1}{\mu}\left(1 + \frac{\varsigma}{m(1-\rho)}\right)$$

- Mean waiting time $\overline{w}$ (Little's law again)

$$\overline{w} = \frac{\overline{n}_q}{\lambda}$$

$$= \frac{\overline{n} - \overline{n}_s}{\lambda}$$

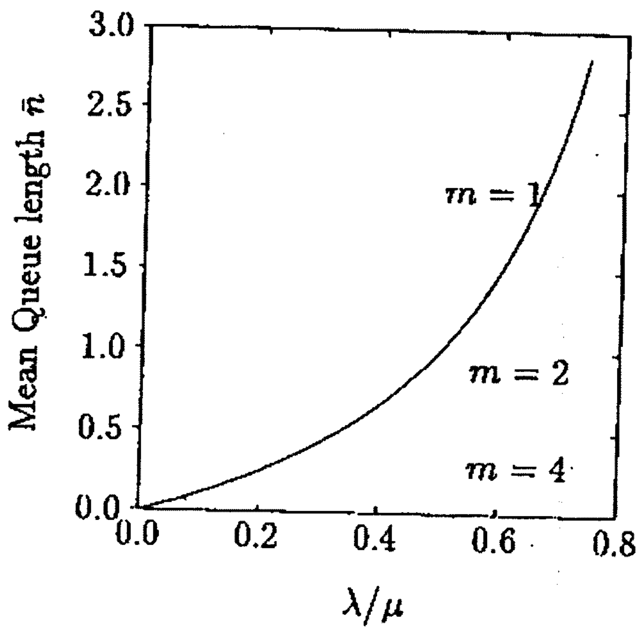
$$= \frac{\rho}{m\mu(1-\rho)}$$

- r_q (q percentile of waiting time)

$$r_q = \max\left(0, \frac{\overline{w}}{\varsigma}\ln\frac{100\varsigma}{100-q}\right)$$

M/M/m Queue Example

- Consider changing m
  - hold λ and μ fixed
  - examine changes in performance metrics
- Observations
  - $M/M/m$ queue has asymptote at $\dfrac{\lambda}{m\mu}$
  - substantial performance gains with even two servers



M/M/1 and M/M/m Queue Comparison

- Which is better?
  - m queues each with an arrival rate λ/m
  - one queue with m servers and an arrival rate of λ
- Suppose we use mean response time as our metric...
  - $m\ M/M/1$ queues

$$\bar{r} = \frac{1}{\mu - \lambda/m}$$

  - one $M/M/m$ queue

$$\bar{r} = \frac{1}{\mu}\left(1 + \frac{\varsigma}{m(1-\rho)}\right)$$

where

$$\varsigma = \frac{(\lambda/\mu)^m}{m!\left(1 - \lambda/(m\mu)\right)} p_0$$

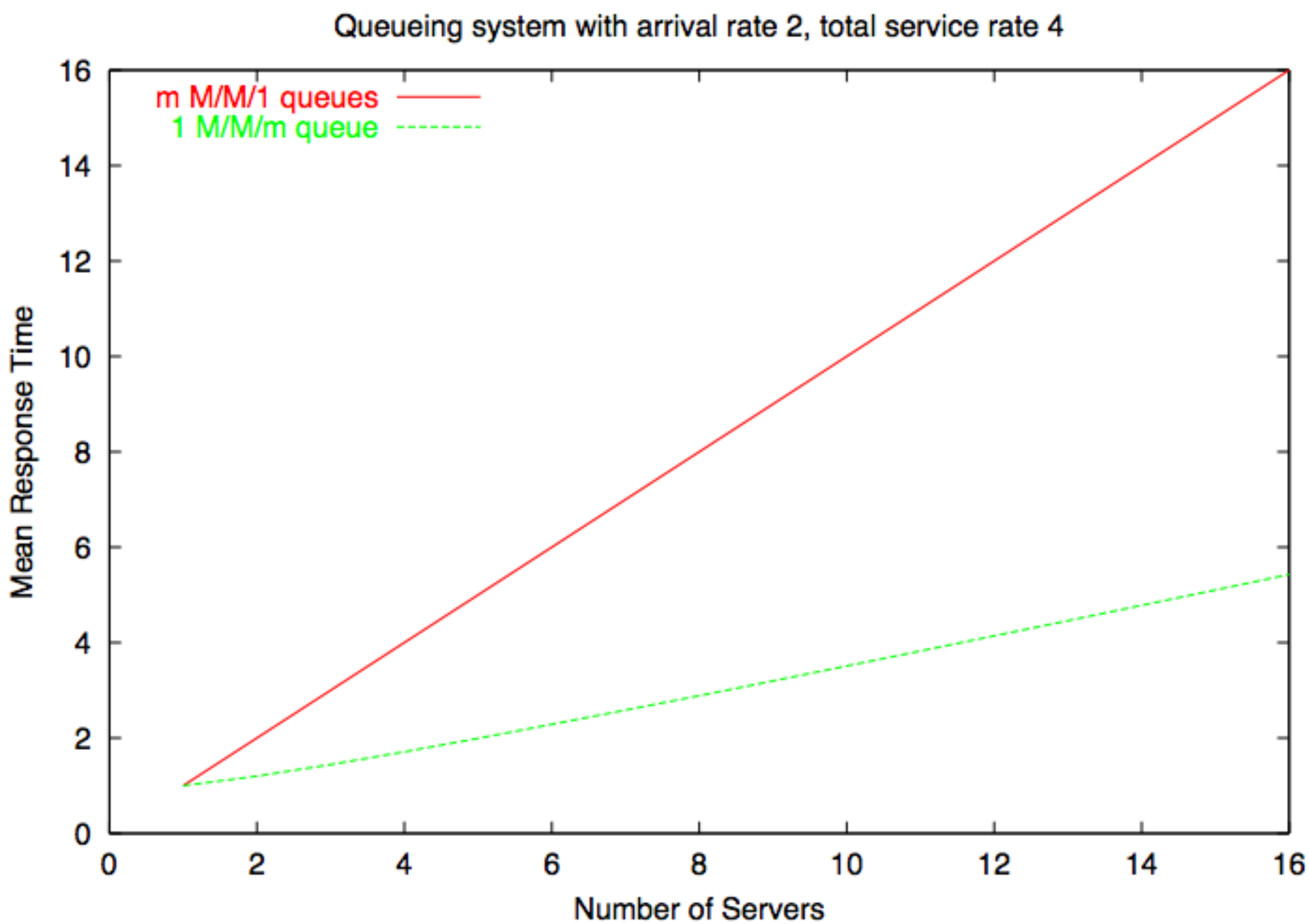
and

$$p_0 = \left[1 + \frac{(\lambda/\mu)^m}{m!\left(1 - \lambda/(m\mu)\right)} + \sum_{n=1}^{m-1} \frac{(\lambda/\mu)^n}{n!}\right]^{-1}$$

Queueing Comparison

- Consider the following
  - service rate μ fixed at 4, divided evenly among *m* servers
  - fix $\lambda = 2$
  - *m* *M*/*M*/1 queues (arrival rate to each is λ/m)
  - One *M*/*M*/*m* queue (total arrival rate is λ)
  - Increase *m*
- What happens to response time in both queues? Why?

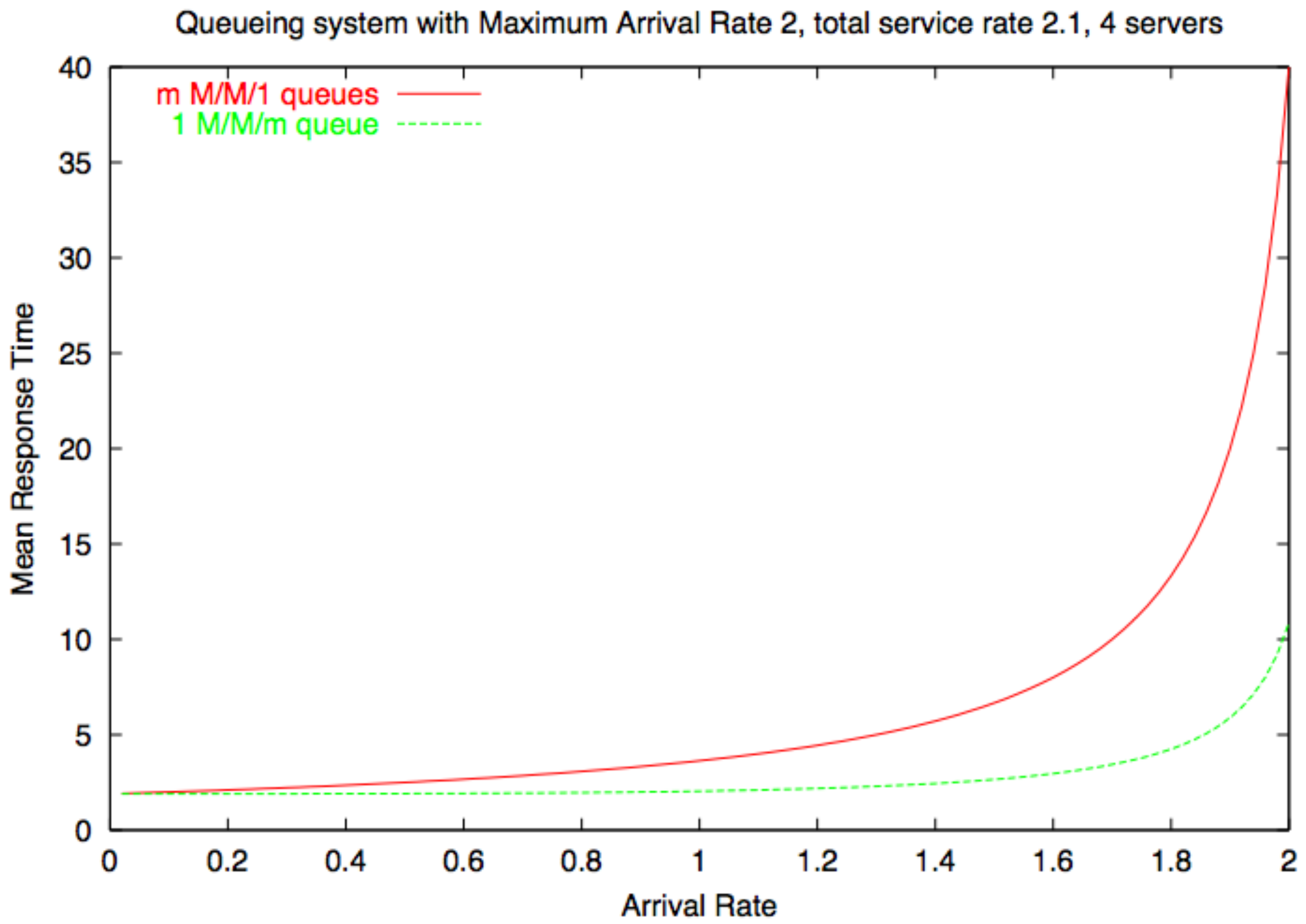
Mean Response Time as function of m



Queueing Comparison

- Consider the following
  - service rate μ fixed at 2.1, divided evenly among m servers
  - varying λ (subject to stability constraint)
  - m $M/M/1$ queues (arrival rate to each is λ/m)
  - One $M/M/m$ queue (total arrival rate is λ)
- What happens as λ approaches 2.1? Why?

Mean Response Time as a Function of Arrival Rate



Extrapolation Scenarios

- Given queueing formulae, standard questions include
  - Performance measures for different parameters
  - Parameters values needed to satisfy a particular performance constraint
- Examples:
  - What is the mean response time if arrival rate doubles?
  - What is the mean queue length if service rate decreases by one third?
  - What is the number of servers for mean response time less than five minutes?
- Approach:
  - Plug and crank
  - Repeated solution with different parameter values

Extrapolation Scenarios (Continued)

- Concrete example
  - multiprocessor system (two processors)
  - mean job service time is 15 seconds
  - mean job interarrival time is 12 seconds
- By inspection
  - mean service rate is 4.0 jobs/minute (per processor)
  - mean arrival rate is 5.0 jobs/minute

  and by plug and crank, we have mean response time $\bar{r}$

$$\bar{r} = \frac{1}{\mu}\left[1 + \frac{\varsigma}{m(1-\rho)}\right] = 0.41 \text{ minutes } (24.6 \text{ seconds})$$

- How many processors do we need to have $\bar{r} < 0.3$ minutes?
  - solve for $m = 3, 4, \ldots$
  - find smallest value of m such that $\bar{r} < 0.3$
  - here, $m = 3$ satisfies this constraint

M/M/m/B Queues

- Finite buffers
  - no more than B jobs in total can be
    - queued
    - *and* in service

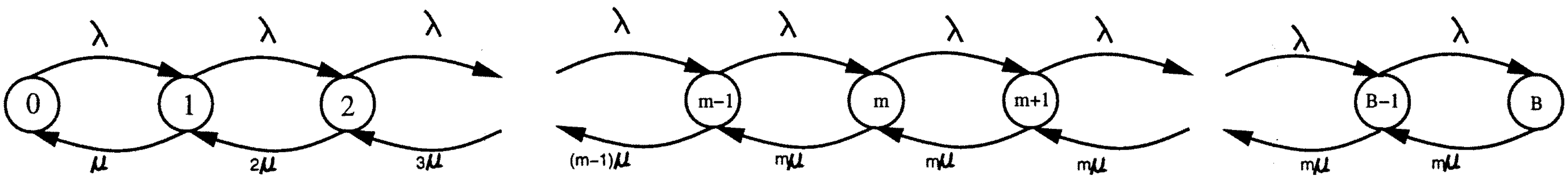
    (i.e., total number of jobs in the system must be less than B)
  - jobs arriving when B jobs are present are discarded
- More formally, this implies

$$\lambda_n = \lambda \qquad n = 1, 2, \ldots, B-1$$

and

$$\mu_n = \begin{cases} n\mu & n = 1, 2, \ldots, m-1 \\ m\mu & n = m, m+1, \ldots, B \end{cases}$$

- Observations
  - $B \geq m$ or servers are wasted
  - birth-death process
  - finite number of states



M/M/m/B Queues (Continued)

- Applying the state occupancy formula

$$p_n = \begin{cases} \dfrac{\lambda^n}{n!\,\mu^n} p_0 & n = 1, 2, \ldots, m-1 \\ \dfrac{\lambda^n}{m!\,m^{n-m}\mu^n} p_0 & n = m, m+1, \ldots, B \end{cases}$$

- And, because $\rho = \dfrac{\lambda}{m\mu}$

$$p_n = \begin{cases} \dfrac{(m\rho)^n}{n!} p_0 & n = 1, 2, \ldots, m-1 \\ \dfrac{\rho^n m^m}{m!} p_0 & n = m, m+1, \ldots, B \end{cases}$$

- Finally, the probability of zero jobs in the system is

$$p_0 = \sum_{n=0}^{B} p_n = \left[1 + \frac{\left(1 - \rho^{B-m+1}\right)(m\rho)^m}{m!(1-\rho)} + \sum_{n=1}^{m-1} \frac{(m\rho)^n}{n!} \right]^{-1}$$

- Now, we can use the state occupancy probabilities to compute
  - mean response time
  - mean queue lengths
  - effective arrival rates

$M/M/m/B$ Queues (Continued)

- Mean queue length $\bar{n}$ (queue plus service)

$$\bar{n} = \sum_{n=1}^{B} n p_n$$

and mean number in the queue

$$\bar{n}_q = \sum_{n=m+1}^{B} (n-m) p_n$$

- Arrivals are constrained by waiting space
  - effective arrival rate $\tilde{\lambda}$ is less than λ
  - jobs enter the system only when buffers are available

$$\tilde{\lambda} = \sum_{n=0}^{B-1} \lambda p_n = \lambda(1 - p_B)$$

and the difference λ - $\tilde{\lambda}$ is the loss rate

  - Because jobs are not lost after entry, the mean response time is

$$\bar{r} = \frac{\bar{n}}{\tilde{\lambda}} = \frac{\bar{n}}{\lambda(1 - p_B)}$$

by Little's law

  - Finally, the utilization U of each server is

$$U = \frac{\tilde{\lambda}}{m\mu} = \rho(1 - p_B)$$

Other Queues

- Other queues can be solved to varying degrees...
- Exact solutions are possible for
  - $M/E_r/1$ (Erlangian service)
  - $M/D/1$ (special case of $M/G/1$)
  - $M/M/1$ with bulk arrivals (restricted cases)
- Analysis is more difficulty for:
  - $G/M/1$
  - M/G/1
  - $G/G/1$

*M/G/*1 Queues

- $M/G/1$
  - General service time distribution
  - Otherwise, similar to $M/M/1$ queues
  - The most complex, readily solvable single queue
- Solution approach
  - First, some additional mathematical machinery
  - Then, comparisons with $M/M/1$ queues
- Service time distribution is general
  - Service history matters
  - Denote service time already received by $X_0(t)$
- Arrival distribution is negative exponential
  - Arrival history does not matter
  - But we do need to know the number of customers $N(t)$ present
  - $N(t)$ is non-Markovian because it depends on service time
- State-space description
  - States are $[N(t), X_0(t)]$
  - Mixed discrete/continuous, two-dimensional description
  - Analysis via this method (supplementary variables) is ugly
  - Use the method of embedded Markov chains...

$M/G/1$ Queues (Continued)

- What has changed from $M/M/1$?
  - Two-dimensional state space
  - State space is now continuous (due to $X_0(t)$)
- Ideally
  - Convert $[N(t), X_0(t)]$ to one-dimensional $N(t)$
  - Implicitly specify remaining service duration $X_0(t)$
- How do we do this?
  - Look only at selected points in time
  - Compute new metrics only at those points
  - Choose those points to implicitly carry $X_0(t)$
  - *departures instants* make great choices
    - Remaining (residual) service $X_0(t)$ is zero!
    - At that instant, we can treat the behavior like a Markov chain
    - $N(t)$ is the number of customers left behind
    - This is an *embedded Markov chain*; for details (see Kleinrock, vol. 1) but we haven't specified the distribution of departure instants

*M/G/*1 Queues (Continued)

- A informal derivation follows (see Kleinrock vol. 1 for details)...
- Notation
  - Arrival rate λ (Poisson process)
  - General service time distribution
    - mean $\bar{x}$
    - variance
- What is the expected time until a customer that arrives completes service?
  - Mean time needed to service customers already waiting
    - Mean time is $\bar{n}_q \bar{x}$
    - Note that this is independent of the distribution of x
  - *plus* the residual time for customer in service ...
- Residual life requires yet another aside...

Residual Life

- What is a "renewal"?
  - Informally, a point where random variables which describe a model are memoryless given current state, with respect to past state.
- Renewal example
  - Consider a queue with general service distribution, and Poisson arrival process
  - Most time points are not renewal points, since remaining service time depends on service time completed.
  - However, times at which service completes are renewal points, since arrival process is Poisson.
- Need to determine the residual lifetime of a customer in service:
  - Denote this random variable as R
  - Distribution of R depends on
    - Distribution of original variable A (the service time distribution) at its renewal point and some time expended after the renewal point

Residual Life (Continued)

- Suppose
  - $a(t)$ is the pdf of A (original variable)
  - the original lifetime has expended time t_e

  then $r(t)$, the pdf of R (the residual lifetime) is

$$r\left(\tau - t_e \middle| t_e\right) = a\left(\tau \middle| \tau > t_e\right)$$

$$= \frac{a(\tau)}{P(A > \tau_e)}$$

$$= \frac{a(\tau)}{1 - \int_0^{t_e} a(s)ds}$$

- Intuition
  - in general, knowing about the expended time helps
  - in short, knowledge changes the pdf
  - we saw that this was *not* true for the exponential distribution
  - the geometric distribution is the only case in the discrete domain
- Average residual lifetime $\bar{r}$ (claim without proof)
  - depends only on the first two moments of the original pdf $f(x)$
    - mean $\bar{f}$
    - second moment $\bar{f}^2$ (*not* the variance!)
    - mean residual lifetime is

$$\bar{r} = \frac{\bar{f}^2}{2\bar{f}}$$

Residual Life (Continued)

- Example (computer part)
  - suppose the pdf $b(t)$ of the failure time is uniform
  - and suppose the mean value is 10

$$b(t) = \begin{cases} \frac{1}{20} & 0 < \tau \le 20 \\ 0 & \text{otherwise} \end{cases}$$

  - if the part has been in use for 5 time units, then

$$b(t + t_e) = \begin{cases} \frac{1}{20} & t_e < t + t_e \le 20 \\ 0 & \text{otherwise} \end{cases}$$

and

$$1 - \int_0^5 b(s)ds = 1 - \frac{1}{20} \cdot 5 = 0.75$$

and finally

$$r(\tau - t_e | 5) = \begin{cases} \frac{1}{15} & 0 < \tau + t_e \le 15 \\ 0 & \text{otherwise} \end{cases}$$

  - notice that

$$\bar{r} = \frac{\bar{f}^2}{2\bar{f}} = \frac{133.33}{2 \cdot 10} = 6.67$$

Observe

- pdf of residual time is not the same as the original pdf
- Knowledge of past behavior changes the pdf
- There are only two exceptions
  - negative exponential distribution (continuous)
  - geometric distribution (discrete)

*M/G/*1 Queues (Continued)

- How long does a new arrival have to wait for service?
  - mean time needed to service customers already waiting
    - let $\bar{x}$ denote the mean service time
    - mean time is $\bar{n}_q \bar{x}$
    - note that this is independent of the distribution of x
  - *plus* the residual time for customer in service
    - recall that this is

$$\bar{t} = \frac{\overline{x^2}}{2\bar{x}}$$

assuming a customer is in service

    - the probability of a customer in service is ρ

- Combining items, the waiting time for a new arrival is

$$\bar{r}_q = \bar{n}_q \bar{x} + \rho \frac{\overline{x^2}}{2\bar{x}}$$

Little's Law again!

- Expected number of arrivals during this interval is $\lambda \bar{r}$, so

$$\bar{r}_q = \bar{r}_q \lambda \bar{x} + \rho \frac{\overline{x^2}}{2\bar{x}}$$

and by rearranging terms

$$\bar{r}_q = \frac{\lambda \overline{x^2}}{2(1-\rho)}$$

M/G/1 Queues (Continued)

- As we just saw, the mean time to receive service is is

$$\bar{r}_q = \frac{\lambda \overline{x^2}}{2(1-\rho)}$$

- Adding the mean service time yields the mean response time

$$\bar{r} = \bar{x} + \frac{\lambda \overline{x^2}}{2(1-\rho)}$$

- Normally, both $\bar{r}$ and $\bar{r}_q$ are expressed as (verify the math)

$$\bar{r} = \frac{1}{\mu} + \frac{\lambda(1+C_s^2)}{2\mu^2(1-\rho)}$$

and

$$\bar{r}_q = \frac{\lambda(1+C_s^2)}{2\mu^2(1-\rho)}$$

where C_s^2 is the coefficient of variation

$$C_s^2 = \frac{\sigma^2}{\bar{x}^2}$$

and

- σ^2 is the variance of the mean service time

- $\frac{1}{\mu}$ is the mean service time

- and by simplification (yields original formulation): $1 + C_s^2 = 1 + \frac{\sigma^2}{\bar{x}^2} = 1 + \frac{\overline{x^2} - \bar{x}^2}{\bar{x}^2} = \frac{\overline{x^2}}{\bar{x}^2} = \mu^2 \overline{x^2}$

M/G/1 Queues (Continued)

- Via Little's law, the mean number in the system is

$$\begin{aligned}\bar{n} &= \lambda \bar{r} \\ &= \frac{\lambda}{\mu} + \frac{\lambda^2 \left(1 + C_s^2\right)}{2\mu^2 (1-\rho)} \\ &= \rho + \frac{\rho^2 \left(1 + C_s^2\right)}{2(1-\rho)}\end{aligned}$$

- Observations
  - this is the famous Pollaczek - Khinchin (PK) formula
  - learn it, remember it, treasure it!
  - C_s is *one* for the negative exponential distribution, so

$$\bar{n} = \rho + \frac{\rho^2 (1+1)}{2(1-\rho)} = \rho + \frac{\rho^2}{1-\rho} = \frac{\rho}{1-\rho}$$

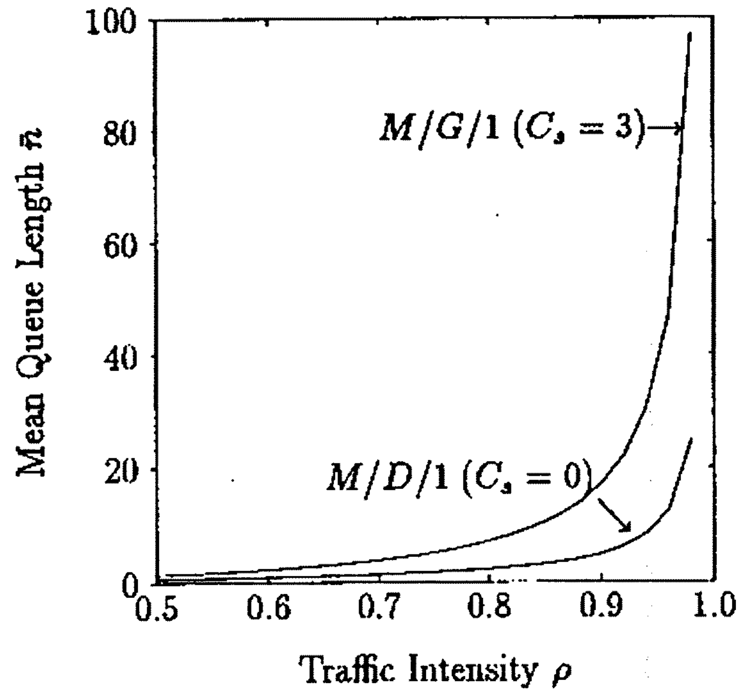
    as we knew before
  - C_s is *zero* for the deterministic distribution ($M/D/1$ queue)

$$\bar{n} = \rho + \frac{\rho^2 (1+0)}{2(1-\rho)} = \frac{\rho}{(1-\rho)} \left(\frac{2-\rho}{2} \right)$$

- The value of C_s has profound implications
  - larger C_s increases mean queue length and response time
  - values grow linearly with C_s

Queueing Comparison

- Consider the following
  - $M/D/1$ queue ($C_s = 0$)
  - $M/M/1$ queue ($C_s = 1$)
  - $M/G/1$ queue ($C_s > 1$)



Queueing Example

- Consider the following
  - arrival rate $\lambda = 0.6$
  - service rate $\mu = 1.0$
  - $M/D/1$, $M/M/1$, and $M/G/1$ queues

  and compare mean response times
- $M/M/1$

$$\bar{r} = \frac{1}{\mu - \lambda} = \frac{1}{1.0 - 0.6} = 2.5$$

- $M/D/1$

$$\bar{r} = \frac{1}{\mu} + \frac{\lambda\left(1 + C_s^2\right)}{2\mu^2(1-\rho)} = \frac{1}{1.0} + \frac{0.6(1+0)}{2(1.0)(1-0.6/1.0)} = 1.75$$

- $M/G/1$ $(C_s = 2.0)$

$$\bar{r} = \frac{1}{\mu} + \frac{\lambda\left(1 + C_s^2\right)}{2\mu^2(1-\rho)} = \frac{1}{1.0} + \frac{0.6(1+1)}{2(1.0)(1-0.6/1.0)} = 3.25$$

Queueing Example (Continued)

- Consider $M/M/1$ and $M/G/1$ queues
  - assume same arrival rates for both
  - desire same mean response times
  - must solve for ratio of service rates
- $M/M/1$

$$\bar{r} = \frac{1}{\mu_m - \lambda}$$

- $M/G/1$

$$\bar{r} = \frac{1}{\mu_g} + \frac{\lambda\left(1 + C_s^2\right)}{2\mu_g^2\left(1 - \lambda / \mu_g\right)}$$

- Equating, we have

$$\frac{1}{\mu_m - \lambda} = \frac{1}{\mu_g} + \frac{\lambda\left(1 + C_s^2\right)}{2\mu_g^2\left(1 - \lambda / \mu_g\right)}$$

- Let's look at some numerical solutions...

M/G/1 via Embedded DTMC

- M/G/1 can be analyzed from the point of view of an *embedded DTMC*

- Note : state can be defined as (n, r) where n is the number in system, and r is remaining time of the job in service.

- Future behavior depends only on n at instants when $r = 0$—job departures

- The state of the embedded DTMC is the number of jobs in system *at the time the last job left service*

M/G/1 via Embedded DTMC

- So what is P_{ij}? The probability that exactly $j-i-1$ jobs arrived while the last job received service.

- If the arrival process is Poisson, and the service time is x, then the number of arrivals during service is Poisson distributed with mean λx.

- Let $f(x)$ be the pdf for the service time distribution, then for $i > 0$ and $i-1 \leq j$

$$P_{ij} = \int_0^\infty f(x) \frac{(\lambda x)^{(j-i-1)}}{(j-i-1)!} \exp\{-\lambda x\}$$

and for $i = 0$, $P_{0,1} = 1$ and $P_{0,n} = 0$ for $n > 1$.

M/G/1 via Embedded DTMC

- Let $\{\pi_i^*\}$ be the equilibrium state probabilities for the embedded DTMC.

- The mean time (from the M/G/1 queue point of view) the DTMC is in a state i is
  - for $i > 1$, $1/\mu$, the mean of the general service time distribution.
  - for $i == 0$, $1/\lambda$ the mean time between arrivals

- From $\{\pi_i^*\}$ and mean occupancy times we can get the stationary distribution of the M/G/1 queue.
  - Define $G = (1/\lambda)\pi_0^* + (1 - \pi_0^*)/\mu$—weighted sum of state probabilities

M/G/1 via Embedded DTMC

– We have $\pi_0 = \pi_0^*/G$, and for $i > 0$ we have $\pi_i = \pi_i^*/G$

Queueing Example (Continued)

- Comparison Example (Continued)
  - arrival rate $\lambda = 0.6$
  - $M/M/1$ queue (service rate $\mu_m = 1.0$)
  - $M/G/1$ queue (service rate μ_g)