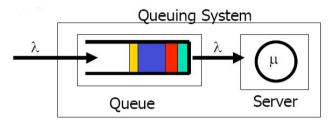


#### **Introduction to Queuing Systems**





# **Queuing Theory**

- View network as collections of queues
  - FIFO data-structures
- Queuing theory provides probabilistic analysis of these queues
- Examples:
  - Average length
  - Probability queue is at a certain length
  - Probability a packet will be lost



# Little's Formula (aka. Little's Law)

- Little's Law:
  - Mean number tasks in system = arrival rate x mean residence time
- Observed before, Little was first to prove
- Applies to any system in equilibrium, as long as nothing in black box is creating or destroying tasks





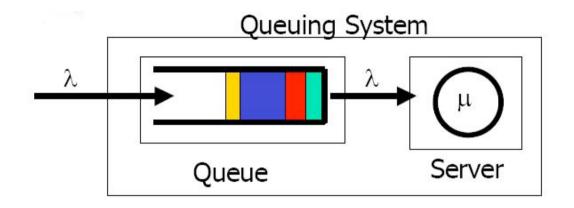
#### **Example using Little's Formula**

- Consider 40 customers/hour visit a pub
  - Average customer spend 15 minutes in the pub (a few quick shots)
- What is average number of customers at at given time?

 $r = \lambda T_r = 40 cust / hr \times 0.25 hr / cust = 10$ 



#### **Queuing System Model**



 Use Little's formula on complete system and parts to reason about average time in the queue



# **Kendall Notation**

- Six parameters in shorthand (x/x/x/x/x/x)
  - First three typically used, unless specified
- 1. Arrival Distribution
- 2. Service Distribution
- 3. Number of servers
- 4. Total Capacity (infinite if not specified)
- 5. Population Size (infinite)
- 6. Service Discipline (FCFS/FIFO)



#### **Distributions**

- M: Exponential
- D: Deterministic (e.g. fixed constant)
- E<sub>k</sub>: Erlang with parameter k
- H<sub>k</sub>: Hyperexponential with param. k
- G: General (anything)
  - M/M/1 is the simplest 'realistic' queue



#### **Kendal Notation Examples**

- M/M/1:
  - Exponential arrivals and service, 1 server, infinite capacity and population, FCFS (FIFO)
- M/M/n
  - Same as above, but *n* servers
- G/G/3/20/1500/SPF
  - General arrival and service distributions, 3 servers, 17 queue slots (20-3), 1500 total jobs, Shortest Packet First



#### M/M/1 Queue

#### Single Isolated Link

•Assume messages arrive at the channel according to a Poisson process at a rate  $\lambda$  messages/sec.

•Assume that the message lengths have a negative exponential distribution with mean 1/v bits/message.

•Channel transmits messages from its buffer at a constant rate *c* bits/sec

•The buffer associated with the channel can considered to be effectively of infinite length.

#### => an M/M/1 queue.



### M/M/1

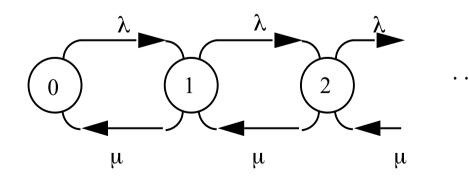
- The first M says that the inter-arrival times to the queue are negative-exponential (ie Markovian or memoryless);
- The second M says that the service times are negative-exponential (ie Markovian or memoryless again);
- The 1 says that there is a single server at the queue.
- Shorthand notation for "a queue with Poisson arrivals, negative exponentially distributed message lengths, a single server, and infinite buffer space".



- Arrival rate  $\lambda$  messages/sec,
- Message lengths neg exp, mean 1/v bits/message,
- Transmission rate *c* bit/sec,

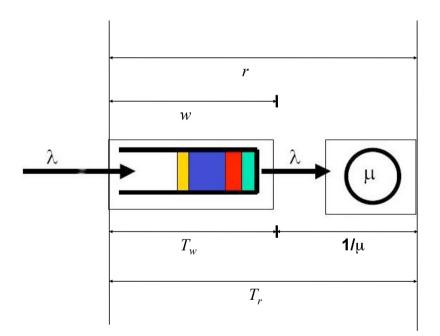
- i.e. messages transmitted at rate  $\mu = cv$  messages/sec.

- Define the state of the system as the total number of messages at the link (waiting + being transmitted).
- The system is therefore a <u>Markov chain</u>, since both the arrival and message length distributions are memoryless.





#### M/M/1 queue model



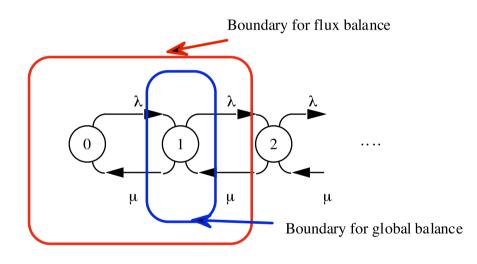


# Solving queuing systems

- Given:
  - $-\lambda$ : Arrival rate of jobs (packets)
  - $-\mu$ : Service rate of the server (output link)
- Solve:
  - r: average number of items in queuing system
  - $-T_r$ : avg. time an item spends in whole system
  - $-T_w$ : avg. time an item waits in the queue



#### **Equilibrium conditions**



$$\lambda p_{i-1} = \mu p_i \ i = 1, 2, \cdots$$



# Solving the flux balance equations recursively we obtain:

$$p_i = \frac{\lambda}{\mu} p_{i-1} = \left(\frac{\lambda}{\mu}\right)^2 p_{i-2} = \dots = \left(\frac{\lambda}{\mu}\right)^i p_0$$

**Define** occupancy 
$$\rho = \frac{\lambda}{\mu} \implies p_i = \rho^i p_0$$



(normalising condition)

$$p_0 = \left(\sum_{i=0}^{\infty} \rho^i\right)^{-1}$$

(geometric distribution)

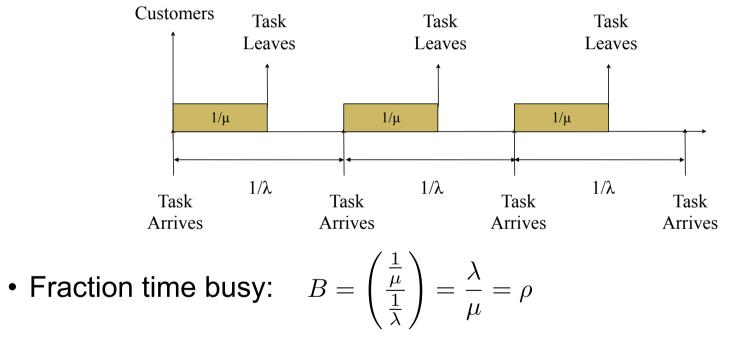
$$p_i = (1 - \rho)\rho^i$$
  $i = 0, 1, 2, \cdots$ 

# Note: Solution valid only for $\rho < 1$ (stability condition) For $\rho \ge 1$ , there is no steady state!



#### **Alternative explanation**

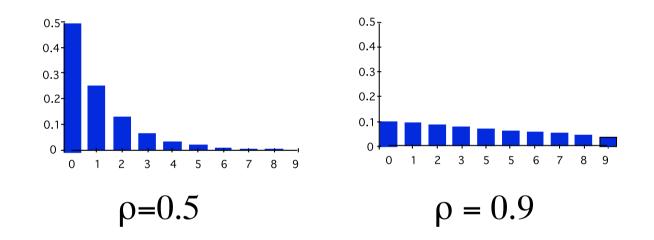
- Consider a D/D/1 system with arrival rate  $\lambda$  and service rate  $\mu$ 



- Fraction time idle:  $\Omega = 1 \Longrightarrow p_0 = (1 \rho)$
- Valid for the general case G/G/1 with averages  $\lambda$  and  $\mu$



#### M/M/1 Queue Length Distribution



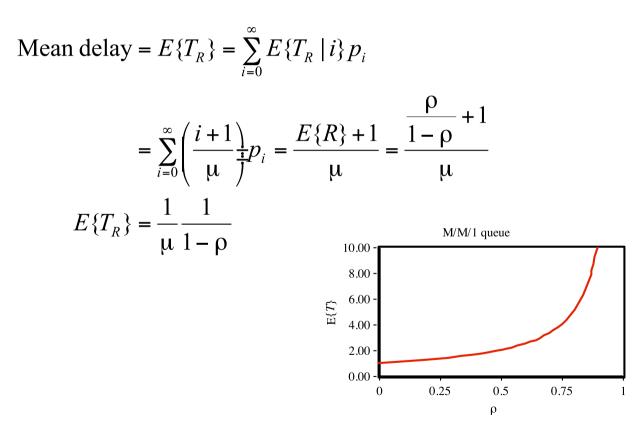


#### M/M/1 Mean Number in the System

$$E\{R\} = \sum_{i=0}^{\infty} ip_i = (1-\rho) \sum_{i=0}^{\infty} i\rho^i$$
$$= (1-\rho) \frac{\rho}{(1-\rho)^2}$$
$$= \frac{\rho}{1-\rho}$$



#### M/M/1 Mean Delay





#### M/M/1 Mean Waiting Time

• The waiting time is defined as the time in the queue, excluding service time:

$$E\{T_W\} = E\{T_R\} - \frac{1}{\mu} = \frac{1}{\mu} \left(\frac{1}{1-\rho}\frac{1}{j} - \frac{1}{\mu} = \frac{1}{\mu} \left(\frac{\rho}{1-\rho}\frac{1}{j}\right)$$



#### Summary of M/M/1 (infinite buffer)

$$\rho = \frac{\lambda}{\mu} < 1$$

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

$$E\{T_R\} = \frac{1}{\mu} \frac{1}{1-\rho}$$

$$E\{T_W\} = \frac{1}{\mu} \frac{\rho}{1-\rho}$$



#### Little's Formula

• If  $\lambda$  messages/sec enter a "system",

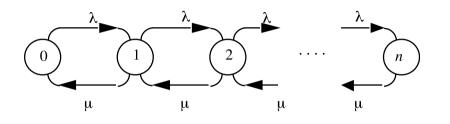
and each spends, on average, W seconds there, then the total number in the "system" must be given by  $\lambda$  W by a simple flow argument.

• Proof for M/M/1:

$$E\{T_R\} = \frac{1}{\mu} \frac{1}{1-\rho} = \frac{1}{\lambda} \frac{\lambda/\mu}{1-\rho} = \frac{1}{\lambda} E\{R\}$$
$$\implies E\{R\} = \lambda E\{T_R\}$$



#### M/M/1/n – Finite Buffer



$$p_{i} = \rho^{i} p_{0} i = 0, 1, \dots, n$$

$$p_{0} = \left(\sum_{i=0}^{n} \rho^{i} \frac{1}{j}^{-1}\right)^{-1} = \frac{1 - \rho}{1 - \rho^{n+1}}$$

$$p_{i} = \frac{1 - \rho}{1 - \rho^{n+1}} \rho^{i} i = 0, 1, \dots, n$$



#### M/M/1/n

• Probability that the buffer will be blocked:

$$P_B = \Pr\{\text{message is blocked}\} = p_n = \frac{1-\rho}{1-\rho^{n+1}}\rho^n$$



#### M/M/1/n example

*Example* : 
$$\rho = 0.5$$

- Assume that we want  $P_B \le 10^{-3}$
- Choose *n* so that

$$\frac{1-0.5}{1-0.5^{n+1}} 0.5^n = \frac{0.5^{n+1}}{1-0.5^{n+1}} \le 10^{-3}$$
$$0.5^{n+1} \le \frac{10^{-3}}{1+10^{-3}}$$
$$n+1 \ge 9.96$$

• i.e. choose n = 9



#### M/M/1/n – validity of solution

- Solution for the finite buffer queue is valid for <u>all</u>  $\rho$ - not just for  $\rho < 1$  as was the case for the infinite buffer.
- This is because the finite buffer copes with overloads by blocking messages, rather than by creating a backlog of messages.



#### M/M/1/n – Little's formula

To use Little's formula for the finite buffer case, note that the rate of non-blocked messages is given by:

$$\gamma = \lambda (1 - P_B)$$

and it is *only these messages*\_which contribute to the buffer size. This implies that:

 $E\{R\} = \gamma E\{T_R\} = \lambda(1 - P_B)E\{T_R\}$ 



#### Summary of M/M/1/n

$$\rho = \frac{\lambda}{\mu} \quad (\operatorname{can be} \ge 1 \operatorname{ or} < 1)$$

$$E\{R\} = \lambda(1 - P_B)E\{T_R\}$$

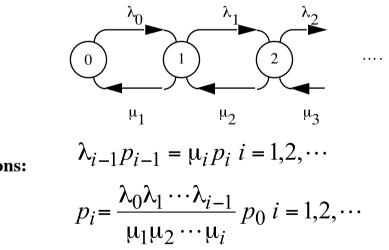
$$P_B = \frac{(1 - \rho)\rho^n}{1 - \rho^{n+1}}$$

$$E\{R\} = \sum_{i=0}^n ip_i = \frac{(1 - \rho)}{1 - \rho^{n+1}}\sum_{i=0}^n i\rho^i$$



#### **State dependent Queues**

Let  $\lambda_i$  = arrival rate when the system is in state *i*  $\mu_i$  = service rate when the system is in state *i* 



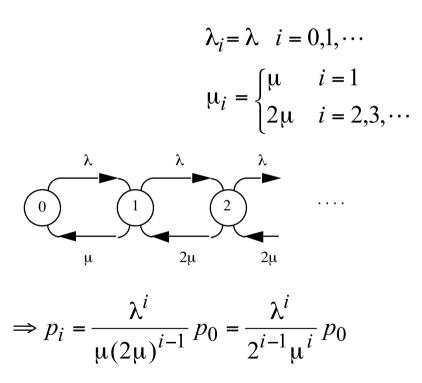


**Balance Equations:** 

with solution

#### M/M/2

• State dependent with





#### M/M/1 vs M/M/2

Compare M/M/2 (2 servers – each at rate  $\mu$ ) to that of a single link at rate  $2\mu$ .

$$E\{T_R\}_{1 \text{ server, rate } 2\mu} = \frac{1}{2\mu} \frac{1}{1-\rho} \qquad (M/M/1 \text{ result})$$

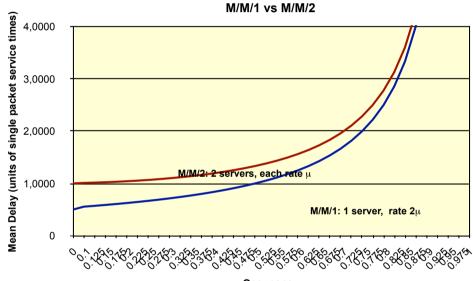
$$E\{T_R\}_{2 \text{ servers, each rate}\mu} = \frac{1}{\mu} \frac{1}{(1+\rho)(1-\rho)} \qquad (M/M/2 \text{ result})$$

$$\frac{E\{T_R\}_{1 \text{ server, rate } 2\mu}}{E\{T_R\}_{2 \text{ servers, each rate}\mu}} = \frac{1+\rho}{2} \le 1$$

#### **Conclusion: Single fast server is always more efficient!**



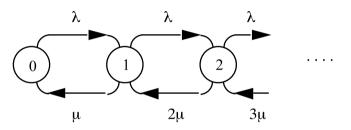
#### M/M/1 vs M/M/2



Occupancy



#### $M/M/\infty$



$$p_i = \frac{\lambda^i}{\mu^i i!} p_0 = e^{-A} \frac{A^i}{i!} \left( \text{with } A = \frac{\lambda}{\mu} \right)^{\frac{1}{2}}$$

#### (a Poisson distribution)



#### M/M/1 delay distribution

 $p_{T_R}(t) = \mu(1-\rho)e^{-\mu(1-\rho)t}$ 

 $\rightarrow$  (an exponential distribution)

• Probability of delay within certain limit:

$$\Pr(T_R \le t) = 1 - e^{-\mu(1-\rho)t}$$

