## 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 $\times$ 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


## Arrivals

## System

## Departures

## 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?

$$
r=\lambda T_{r}=40 \text { cust } / h r \times 0.25 h r / \text { cust }=10
$$

## 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)
- $\mathrm{E}_{k}$ : Erlang with parameter $k$
- $\mathrm{H}_{k}$ : 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.
$\Rightarrow$ 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 $c \mathrm{bit} / \mathrm{sec}$,
- i.e. messages transmitted at rate $\mu=c v$ 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 Markov chain, 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:

$$
p_{i}=\frac{\lambda}{\mu} p_{i-1}=\left(\frac{\lambda}{\mu}\right)^{2} p_{i-2}=\cdots=\left(\frac{\lambda}{\mu}\right)^{i} p_{0}
$$

Define occupancy $\rho=\frac{\lambda}{\mu} \quad \Rightarrow p_{i}=\rho^{i} p_{0}$
(normalising condition)

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

(geometric distribution)

$$
p_{i}=(1-\rho) \rho^{i} \quad i=0,1,2, \cdots
$$

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

## Alternative explanation

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

- Fraction time busy: $\quad B=\left(\frac{\frac{1}{\mu}}{\frac{1}{\lambda}}\right)=\frac{\lambda}{\mu}=\rho$
- Fraction time idle: $\Omega=1=>p_{0}=(1-\rho)$
- Valid for the general case $\mathrm{G} / \mathrm{G} / 1$ with averages $\lambda$ and $\mu$


## M/M/1 Queue Length Distribution




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

$$
\begin{aligned}
E\{R\} & =\sum_{i=0}^{\infty} i p_{i}=(1-\rho) \sum_{i=0}^{\infty} i \rho^{i} \\
& =(1-\rho) \frac{\rho}{(1-\rho)^{2}} \\
& =\frac{\rho}{1-\rho}
\end{aligned}
$$

## M/M/1 Mean Delay

Mean delay $=E\left\{T_{R}\right\}=\sum_{i=0}^{\infty} E\left\{T_{R} \mid i\right\} p_{i}$

$$
\begin{aligned}
&=\sum_{i=0}^{\infty}\left(\frac{i+1}{\mu} \frac{1}{j} p_{i}=\frac{E\{R\}+1}{\mu}=\frac{\frac{\rho}{1-\rho}+1}{\mu}\right. \\
& E\left\{T_{R}\right\}=\frac{1}{\mu} \frac{1}{1-\rho} \\
& \hline
\end{aligned}
$$

## M/M/1 Mean Waiting Time

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

$$
E\left\{T_{W}\right\}=E\left\{T_{R}\right\}-\frac{1}{\mu}=\frac{1}{\mu}\left(\frac{1}{1-\rho} \frac{\frac{1}{\dot{i}}}{j}-\frac{1}{\mu}=\frac{1}{\mu}\left(\frac{\rho}{1-\rho} \frac{\frac{\dot{\dot{i}}}{\bar{j}}}{}\right.\right.
$$

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

$$
\begin{aligned}
& \rho=\frac{\lambda}{\mu}<1 \\
& E\{R\}=\frac{\rho}{1-\rho} \\
& E\left\{T_{R}\right\}=\frac{1}{\mu} \frac{1}{1-\rho} \\
& E\left\{T_{W}\right\}=\frac{1}{\mu} \frac{\rho}{1-\rho}
\end{aligned}
$$

## 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 $\lambda \mathrm{W}$ by a simple flow argument.
- Proof for M/M/1:

$$
\begin{aligned}
E\left\{T_{R}\right\}= & \frac{1}{\mu} \frac{1}{1-\rho}=\frac{1}{\lambda} \frac{\lambda / \mu}{1-\rho}=\frac{1}{\lambda} E\{R\} \\
& \Rightarrow E\{R\}=\lambda E\left\{T_{R}\right\}
\end{aligned}
$$

## M/M/1/n - Finite Buffer



## M/M/1/n

- Probability that the buffer will be blocked:

$$
P_{B}=\operatorname{Pr}\{\text { message is blocked }\}=p_{n}=\frac{1-\rho}{1-\rho^{n+1}} \rho^{n}
$$

## M/M/1/n example

Example: $\quad \rho=0.5$

- Assume that we want

$$
P_{B} \leq 10^{-3}
$$

- Choose $n$ so that

$$
\begin{aligned}
\frac{1-0.5}{1-0.5^{n+1}} 0.5^{n} & =\frac{0.5^{n+1}}{1-0.5^{n+1}} \leq 10^{-3} \\
0.5^{n+1} & \leq \frac{10^{-3}}{1+10^{-3}} \\
n+1 & \geq 9.96
\end{aligned}
$$

- i.e. choose $n=9$


## M/M/1/n - validity of solution

- Solution for the finite buffer queue is valid for all $\rho$
- 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:

$$
\gamma=\lambda\left(1-P_{B}\right)
$$

and it is only these messages_which contribute to the buffer size. This implies that:

$$
E\{R\}=\gamma E\left\{T_{R}\right\}=\lambda\left(1-P_{B}\right) E\left\{T_{R}\right\}
$$

## Summary of M/M/1/n

$$
\begin{aligned}
& \rho=\frac{\lambda}{\mu} \quad(\text { can be } \geq 1 \text { or }<1) \\
& E\{R\}=\lambda\left(1-P_{B}\right) E\left\{T_{R}\right\} \\
& P_{B}=\frac{(1-\rho) \rho^{n}}{1-\rho^{n+1}} \\
& E\{R\}=\sum_{i=0}^{n} i p_{i}=\frac{(1-\rho)}{1-\rho^{n+1}} \sum_{i=0}^{n} i \rho^{i}
\end{aligned}
$$

## 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:

$$
\lambda_{i-1} p_{i-1}=\mu_{i} p_{i} i=1,2, \cdots
$$

with solution

$$
p_{i}=\frac{\lambda_{0} \lambda_{1} \cdots \lambda_{i-1}}{\mu_{1} \mu_{2} \cdots \mu_{i}} p_{0} i=1,2, \cdots
$$

## M/M/2

- State dependent with

$$
\begin{gathered}
\lambda_{i}=\lambda \quad i=0,1, \cdots
\end{gathered} \mu_{i}=\left\{\begin{array}{ll}
\mu & i=1 \\
2 \mu & i=2,3, \cdots
\end{array}\right\}
$$

## 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 $2 \mu$.

$$
\begin{aligned}
& E\left\{T_{R}\right\}_{1 \text { server, rate } 2 \mu}=\frac{1}{2 \mu} \frac{1}{1-\rho} \\
& E\left\{T_{R}\right\}_{2 \text { servers, each rate } \mu}=\frac{1}{\mu} \frac{1}{(1+\rho)(1-\rho)} \quad(\mathrm{M} / \mathrm{M} / 1 \text { result }) \\
& \frac{E\left\{T_{R}\right\}_{1 \text { server, rate } 2 \mu}}{E\left\{T_{R}\right\}_{2 \text { servers, each rate } \mu}}=\frac{1+\rho}{2} \leq 1
\end{aligned}
$$

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



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



## M/M/ $\infty$


(a Poisson distribution)

# M/M/1 delay distribution 

$$
\begin{gathered}
p_{T_{R}}(t)=\mu(1-\rho) e^{-\mu(1-\rho) t} \\
\rightarrow \text { (an exponential distribution) }
\end{gathered}
$$

- Probability of delay within certain limit:

$$
\operatorname{Pr}\left(T_{R} \leq t\right)=1-e^{-\mu(1-\rho) t}
$$

