Queuing Theory

SYS-611: Simulation and Modeling

Paul T. Grogan, Ph.D.
Assistant Professor
School of Systems and Enterprises

Agenda

1. Queuing Theory

Reading: J.V. Farr, “Queuing Theory,” Ch. 2 in *Simulation of Complex Systems and Enterprises,* Stevens Institute of Technology, 2007, pp. 2-1–2-15.

R.C. Larson and A.R. Odoni, “Introduction to Queuing Theory and Its Applications,” Ch. 4 in *Urban Operations Research*, 2007, pp. 182-211. (Web Version Available)

Queuing Theory

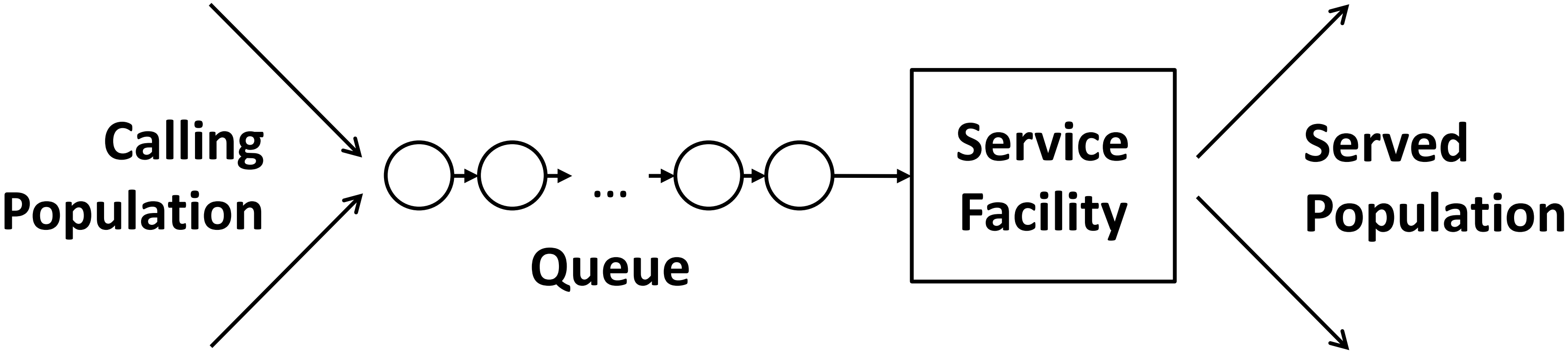
Queuing Theory

Queuing Theory provides analytical models to compute statistics for simple queuing systems

- Single/multiple servers, finite/infinite population/capacity
- Well-defined behaviors (lane switching/reneging)

- Useful when little information is available
- There exist entire graduate courses on queuing
- Simulation handles more general systems with context-specific information, but does not provide an equivalent generalizable “theory”

Generic Queuing Model



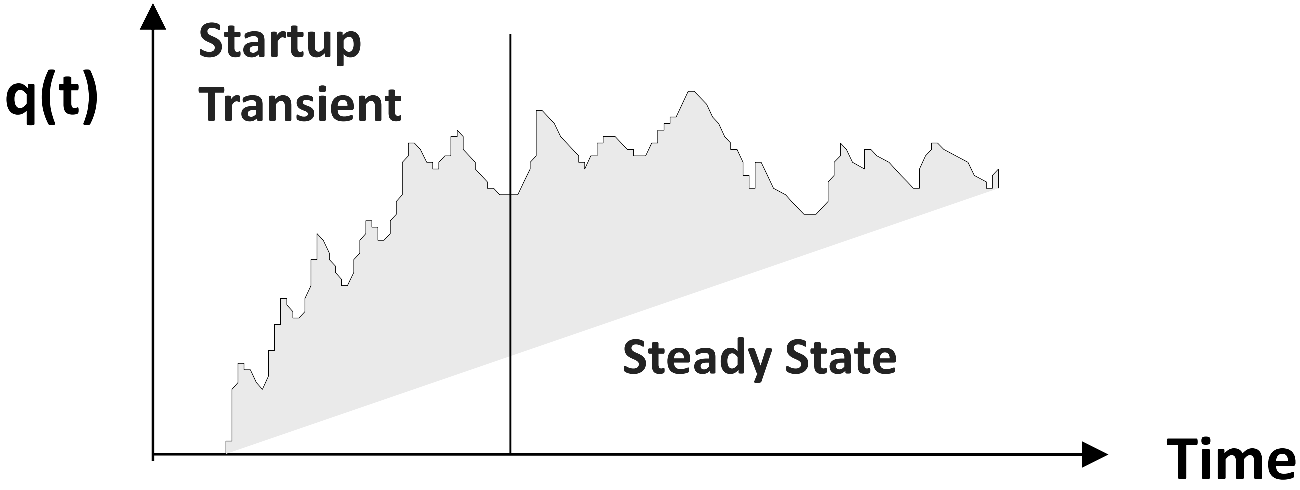
- Size
- Arrival Pattern
- Attitude

- Length

- Structure
- Distribution
- Discipline

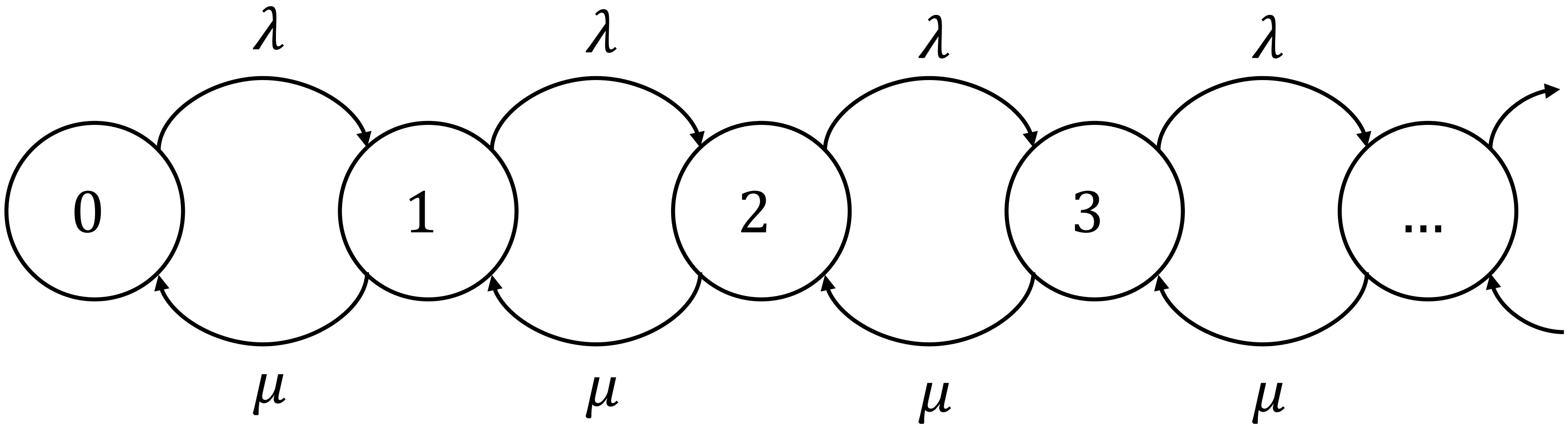
Queuing Theory Statistics

- Queuing theory reports steady-state statistics
  - **Transient**: system behavior is a function of time, e.g. starting from zero initial conditions
  - **Steady-state**: system behavior is no longer a function of time

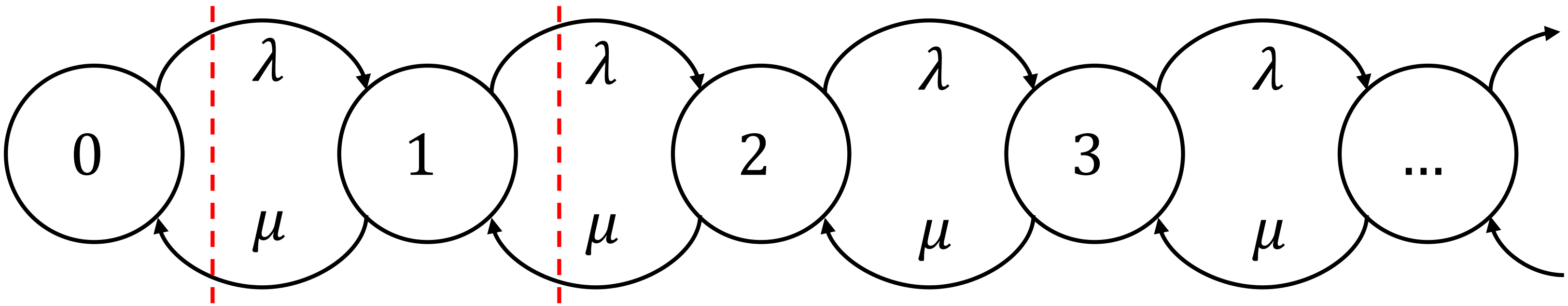


M/M/1 Queuing Model

- State represents number of customers in queue
- Exponentially-distributed inter-arrival period (X) with rate with rate λ cust./min.
- Exponentially-distributed service time (Y) with rate μ cust./min.



M/M/1 Steady-State Probabilities



Definition of probability: $P\{q=i\}=P_i,\ \sum_{i=0}^{\infty} P_i=1$

Steady-state flow between states 0 and 1:

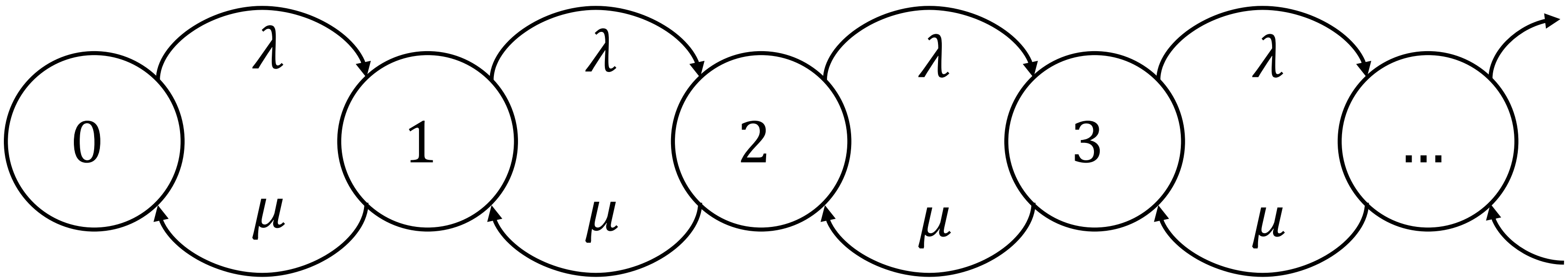
$$\lambda P_0=\mu P_1 \Rightarrow P_1=(\lambda/\mu)P_0$$

Steady-state flow between states 1 and 2:

$$\lambda P_1=\mu P_2 \Rightarrow P_2=(\lambda/\mu)P_1=(\lambda/\mu)^2 P_0$$

$$\Rightarrow P_i=(\lambda/\mu)^i P_0$$

M/M/1 Steady-State Probabilities



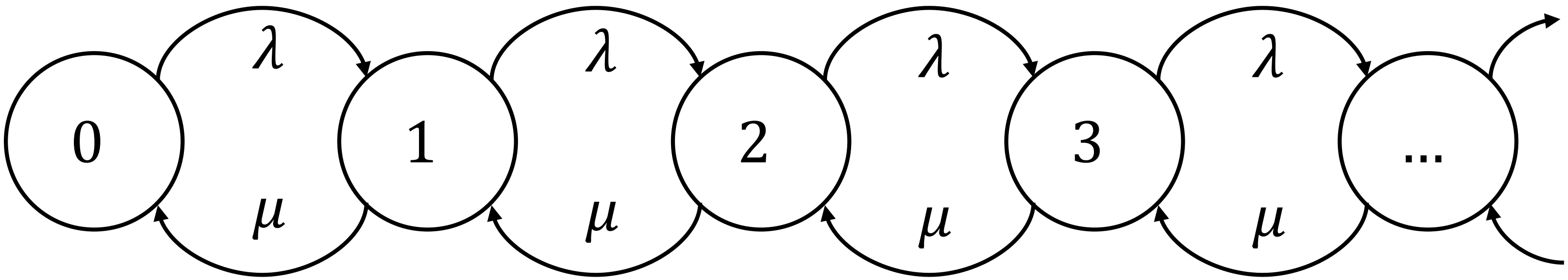
Substituting all cases: $\sum_{i=0}^{\infty} P_i = \sum_{i=0}^{\infty} (\lambda/\mu)^i P_0 = 1$

Recognizing a geometric series for $\rho = \lambda/\mu < 1$:

$$P_0 = \frac{1}{\sum_{i=0}^{\infty} (\lambda/\mu)^i} = 1 - \lambda/\mu$$

$$P_i = (\lambda/\mu)^i (1 - \lambda/\mu)$$

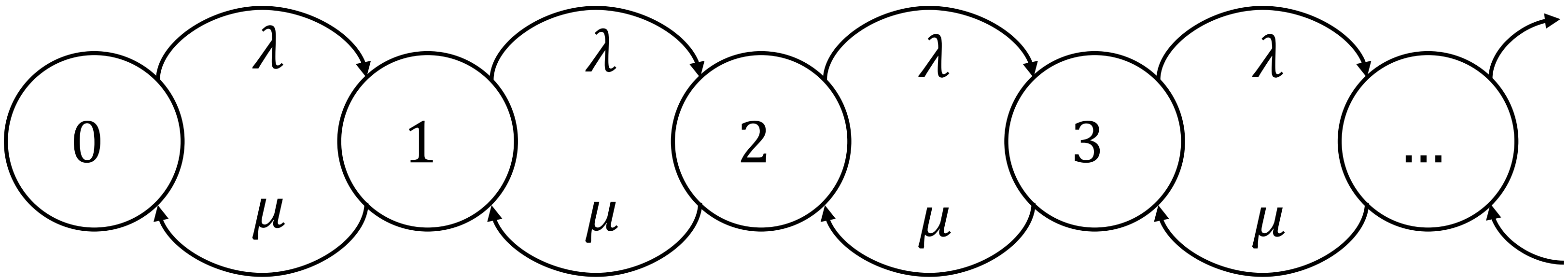
M/M/1 Users in System



$$\bar{L} = \sum_{i=0}^{\infty} i \cdot P_i = \sum_{i=0}^{\infty} i \cdot \left(\frac{\lambda}{\mu}\right)^i \left(1 - \frac{\lambda}{\mu}\right) = \cdots = \frac{\lambda}{\mu - \lambda}$$

$$\bar{L}_q = \sum_{i=1}^{\infty} (i-1) \cdot P_i = \cdots = \frac{\lambda^2}{\mu(\mu - \lambda)}$$

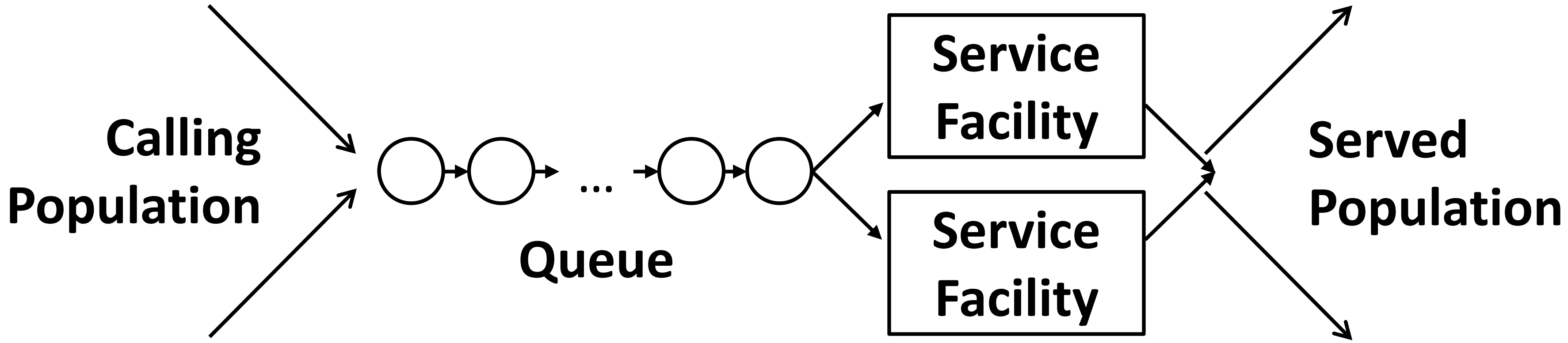
M/M/1 Wait Time in System



$$\bar{W} = \sum_{i=0}^{\infty} \frac{i+1}{\mu} P_i = \cdots = \frac{1}{\mu - \lambda}$$

$$\bar{W}_q = \sum_{i=0}^{\infty} \frac{i}{\mu} P_i = \cdots = \frac{\lambda}{\mu(\mu - \lambda)}$$

Queuing Theory Extensions



- Multiple servers
- Finite calling population
- Finite queue capacity
- General arrival times (non-Markov)
- General service times (non-Markov)
- *See Larson and Odoni (2007) for more detail*

Quantity	Description	M/M/1	M/M/m
λ	Mean arrival rate (cust. per time)	λ	λ
μ	Mean service rate (cust. per time)	μ	μ
ρ	Mean server utilization.	$\rho = \frac{\lambda}{\mu}$	$\rho = \frac{\lambda}{m\mu}$
P_0	Probability of having no customers in queuing system	$P_0 = 1 - \frac{\lambda}{\mu}$	$P_0 = \left[\left(\sum_{i=0}^{m-1} \frac{(\lambda/\mu)^i}{i!}\right) + \frac{(\lambda/\mu)^m}{m!} \frac{1}{1-\rho}\right]^{-1}$
P_i	Probability of having i customers in queuing system	$P_i = \left(\frac{\lambda}{\mu}\right)^i P_0$	$P_i = \begin{cases} \frac{(\lambda/\mu)^i}{i!} P_0, & 0 \le i \le m \\ \frac{(\lambda/\mu)^i}{m!\, m^{i-m}} P_0, & i \ge m \end{cases}$
$\bar{L}_q$	Average number of cust. in queue	$\bar{L}_q = \frac{\lambda^2}{\mu(\mu - \lambda)}$	$\bar{L}_q = \frac{P_0 (\lambda/\mu)^m \rho}{m!\,(1-\rho)^2}$
$\bar{L}$	Average number of cust. in system	$\bar{L} = \frac{\lambda}{\mu - \lambda}$	$\bar{L} = \bar{L}_q + (\lambda/\mu)$
$\bar{W}$	Average cust. waiting time in system	$\bar{W} = \frac{1}{\mu - \lambda}$	$\bar{W} = \bar{W}_q + (1/\mu)$
$\bar{W}_q$	Average cust. waiting time in queue	$\bar{W}_q = \frac{\lambda}{\mu(\mu - \lambda)}$	$\bar{W}_q = \bar{L}_q / \lambda$

Café Java (Class Problem 4-2*)

Café Java's manager is considering hiring a second cashier to handle the evening coffee rush hour.

$$\lambda = \frac{1}{1.5} = 0.67 \text{ cust/min}$$

$$\mu = \frac{1}{0.75} = 1.33 \text{ cust/min}$$

$$\rho = \frac{\lambda}{\mu} = \frac{0.67}{1.33} = 0.50$$

$$\bar{W} = \frac{1}{\mu - \lambda} = 1.50 \text{ min.}$$

$$\bar{W}_q = \frac{\lambda}{\mu(\mu - \lambda)} = 0.75 \text{ min.}$$

$$\bar{L} = \frac{\lambda}{\mu - \lambda} = 1.00 \text{ customers}$$

$$\bar{L}_q = \frac{\lambda^2}{\mu(\mu - \lambda)} = 0.50 \text{ customers}$$

→ **second cashier not needed**

Train Wash (Class Problem 2-1)

NJ Transit operates a single train wash as a part of its Hoboken facility. The service time of the machine is 30 minutes per train. The arrival time of trains can be approximated as a Poisson process with 1.4 trains per hour. The washing center operates 24 hours per day.

1. What is the appropriate queuing model?
2. How many trains (avg) are waiting to use the facility?
3. How much time (avg) does a train spend at the facility?
4. What is the utilization rate of the facility?
5. Is it worthwhile to build a second facility?

Train Wash (Class Problem 2-1)

1. What is the appropriate queuing model?
   - M/M/1 with $\lambda = 1.4$ trains/hour, $\mu = 1/0.5 = 2.0$ trains/hour
2. How many trains (mean) are waiting to use the facility?
   - $\bar{L}_q = \frac{\lambda^2}{\mu(\mu-\lambda)} = 1.63$ trains
3. How much time (mean) does a train spend at the facility?
   - $\bar{W} = \frac{1}{\mu-\lambda} = 1.67$ hours
4. What is the utilization rate of the facility?
   - $\rho = \frac{\lambda}{\mu} = 0.70$
5. Is it worthwhile to build a second facility?
   - Average utilization drops to 0.35, probably not worthwhile.

Hoboken Police (Problem 2-4)

The Hoboken police have 10 patrol cars. A patrol car requires service on average every 15 days. The police department has two repair workers, each of whom take an average of 3 days to repair a car.

A M/M/2 queuing model with calling population 10 yields the following steady-state probabilities:

$P_0 = 0.12$	$P_1 = 0.24$	$P_2 = 0.22$	$P_3 = 0.17$	$P_4 = 0.12$	$P_5 = 0.073$
$P_6 = 0.036$	$P_7 = 0.015$	$P_8 = 0.0044$	$P_9 = 0.0009$	$P_{10} = 0.0001$	

1. How many cars are working at any time?
2. What fraction of the time is at least one worker idle?

Hoboken Police (Problem 2-4)

$P_0 = 0.12$	$P_1 = 0.24$	$P_2 = 0.22$	$P_3 = 0.17$	$P_4 = 0.12$	$P_5 = 0.073$
$P_6 = 0.036$	$P_7 = 0.015$	$P_8 = 0.0044$	$P_9 = 0.0009$	$P_{10} = 0.0001$	

1. How many cars are working at any time?

   $\bar{L} = \sum_{i=0}^{10} i \cdot P_i = 2.4 \Rightarrow 10 - \bar{L} = 7.6$ cars not in repair

2. What fraction of the time is at least one worker idle?

$$P\{\text{worker idle}\} = P_0 + P_1 = 0.36$$