

EE/PEP 345

Modeling and Simulation

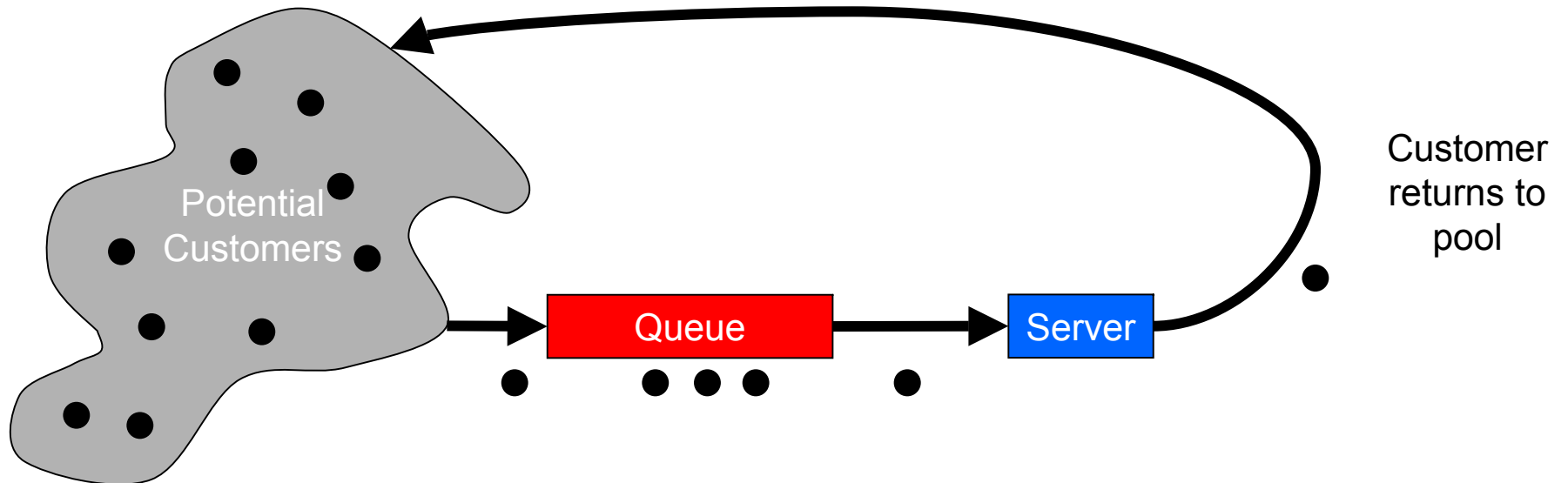
Spring 2004

Class 5

Today's Topic

- Ch 6 Queuing Models

Queuing Models



Performance Measures		
Delay	Queue length	server utilization

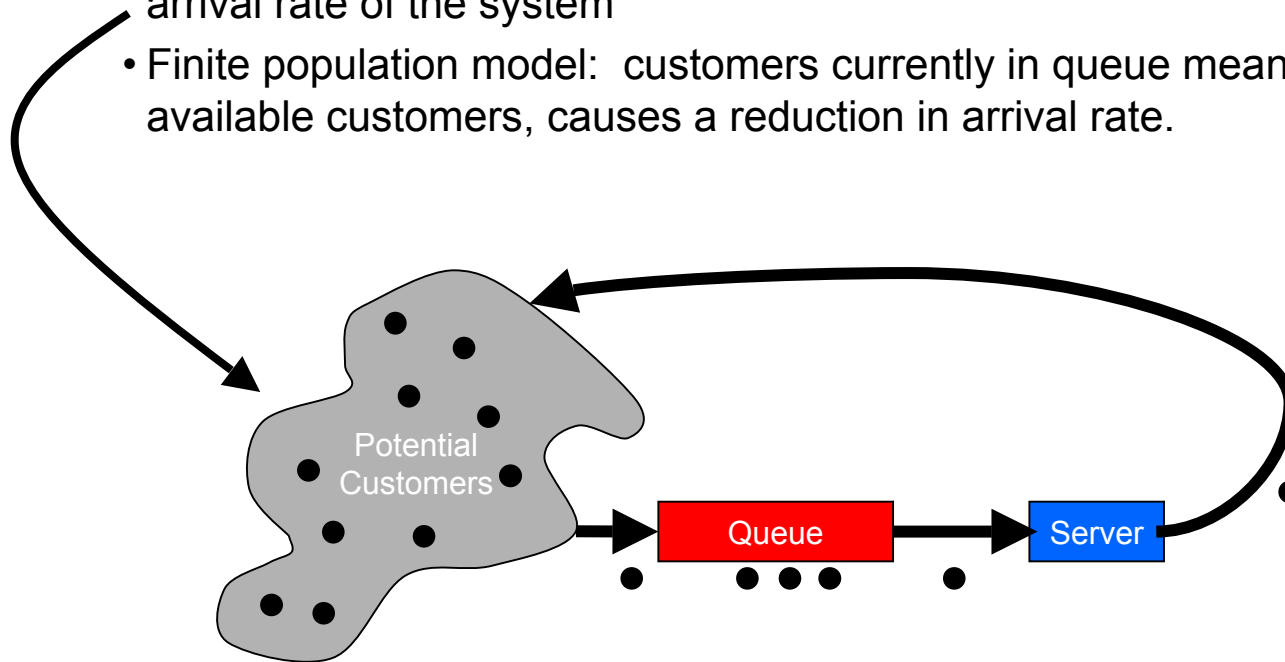
$$utilization \propto \frac{1}{delay}$$

Analysis vs. Simulation

- Simple queuing systems (e.g., single server queues) can be analyzed
- Complex queuing systems (multiple servers, tandem queues, etc.) are best simulated

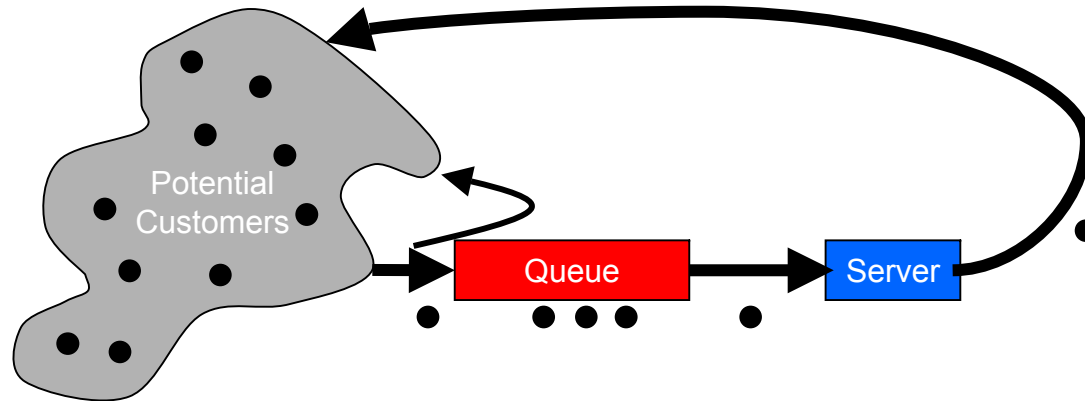
Characteristics of queuing systems

- Calling population - potential customers for service
 - Assume finite or infinite population
 - Infinite population model: customers currently in queue do not influence arrival rate of the system
 - Finite population model: customers currently in queue means fewer available customers, causes a reduction in arrival rate.



Characteristics of queuing systems

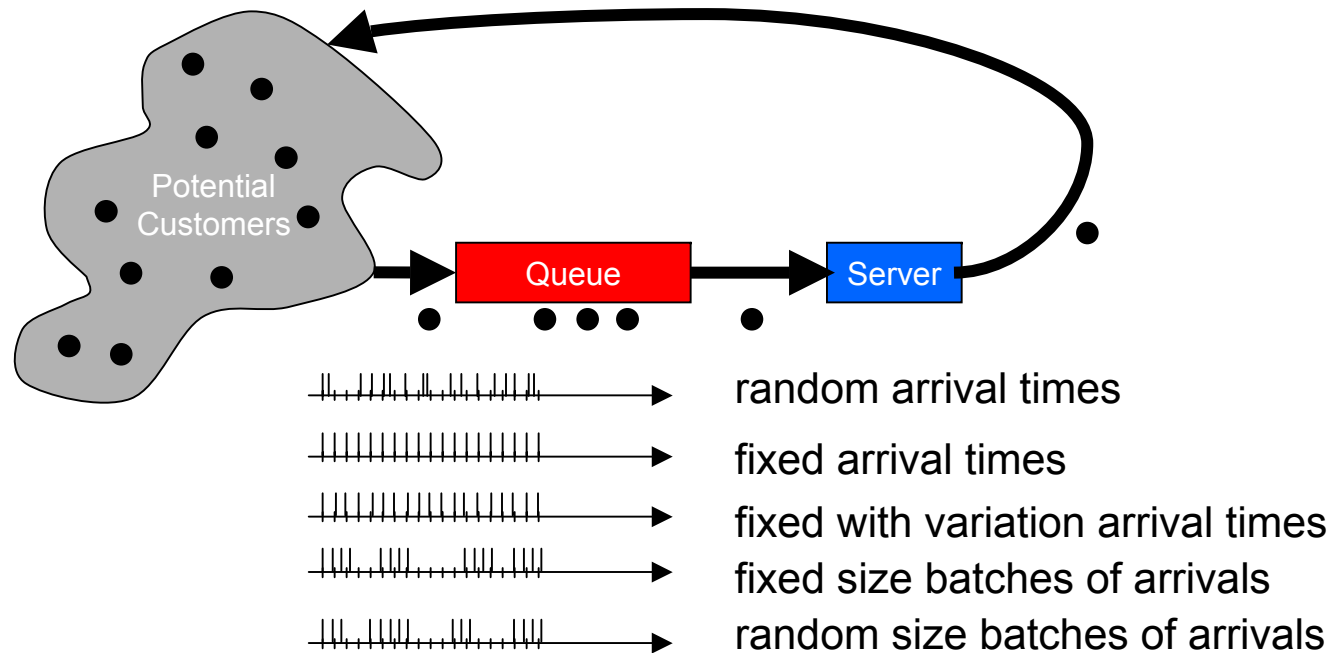
- System capacity
 - There may be a limit to the number of customers who can wait in queue
 - If an arriving customer finds the queue full, they return to the calling population
- Arrival rate: number of arrivals per unit time
- Effective arrival rate: number of customers who arrive and enter the system per unit time



- In switched telephone network terms, this is equivalent to the distinction between “Blocked Calls Cleared” and “Blocked Calls Held”

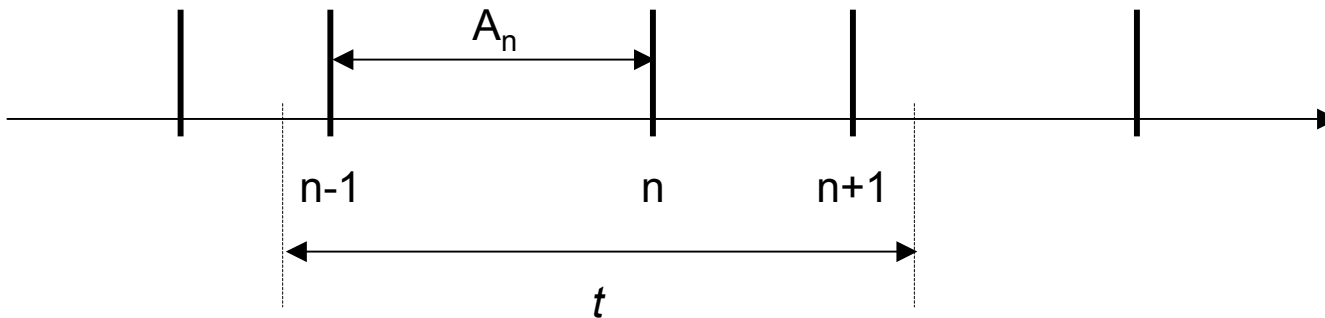
Characteristics of queuing systems

- Arrival process
 - For infinite customer population, arrival process normally characterized by interarrival times of successive customers
 - Arrivals may occur at random times, at scheduled times, or at scheduled times with random variation
 - Arrivals may occur individually or in batches



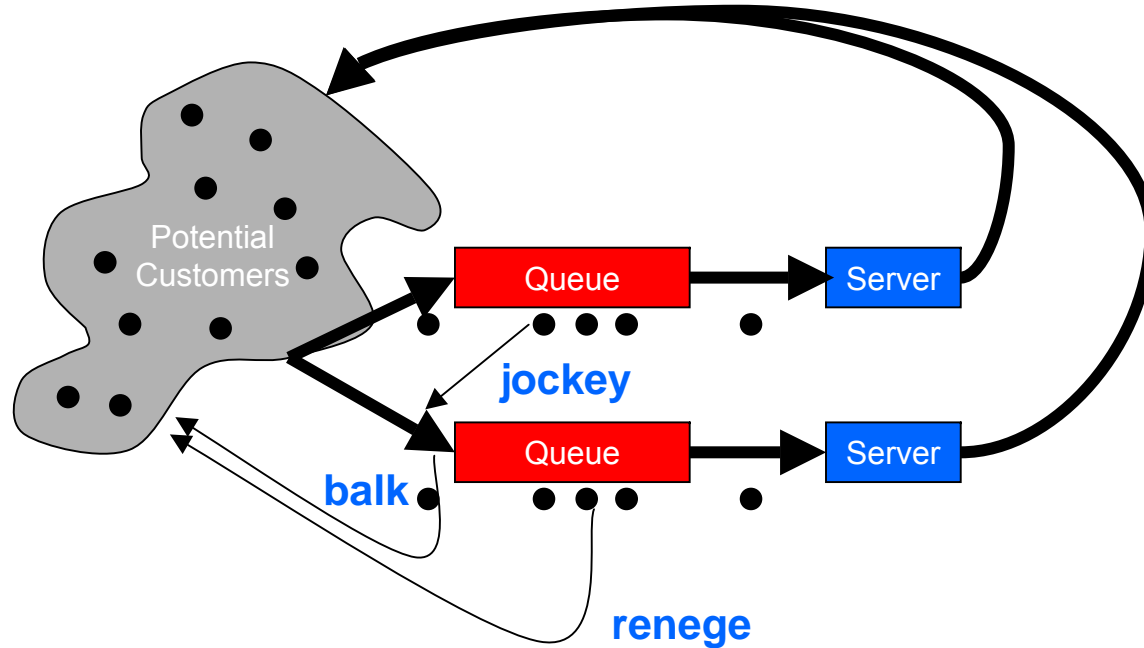
Poisson Arrival Process

- The interarrival time, A_n , is exponentially distributed with mean $1/\lambda$
- Arrival rate is λ
- $N(t)$ is the number arrivals in time interval of length t
- $N(t)$ is Poisson distributed with mean λt



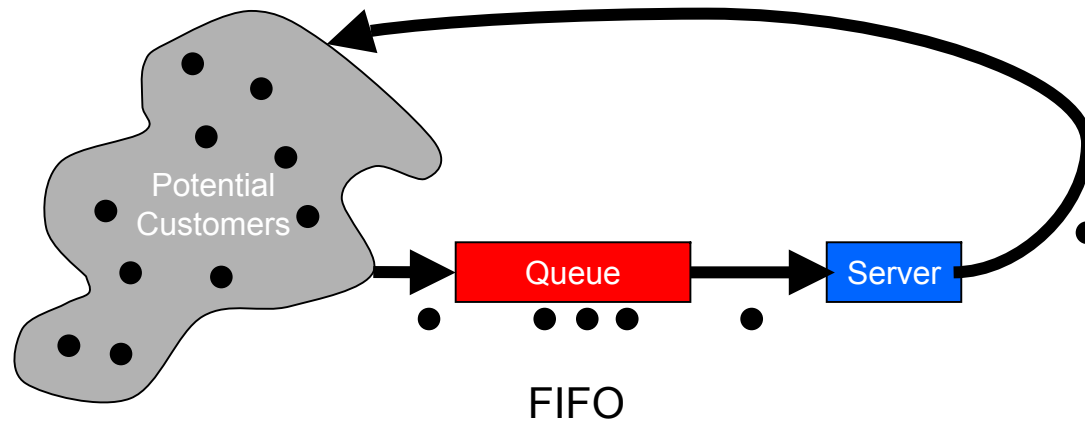
Characteristics of queuing systems

- Customer behavior



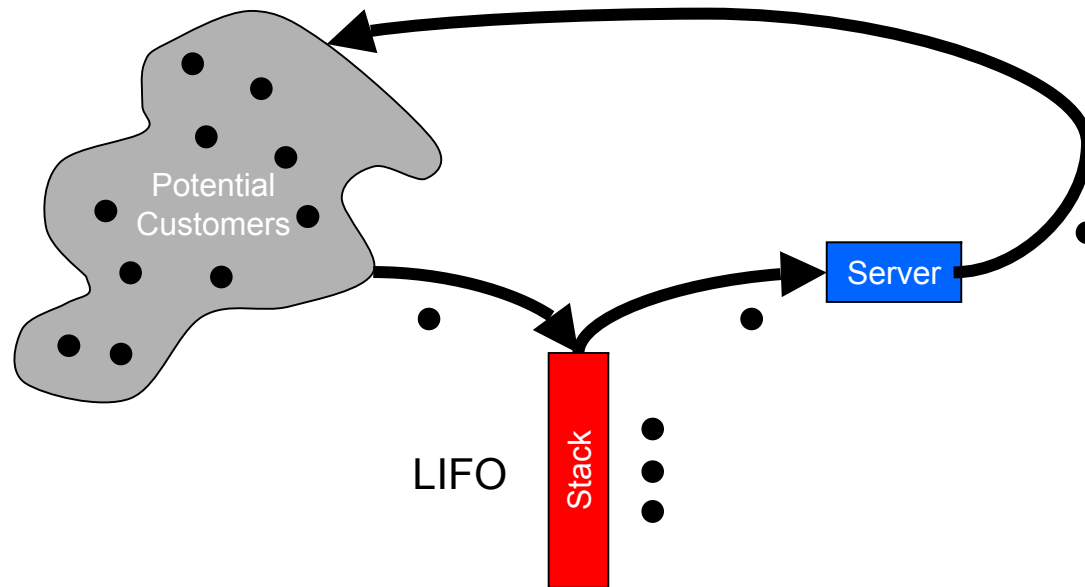
Characteristics of queuing systems

- Queue discipline
 - First in First Out



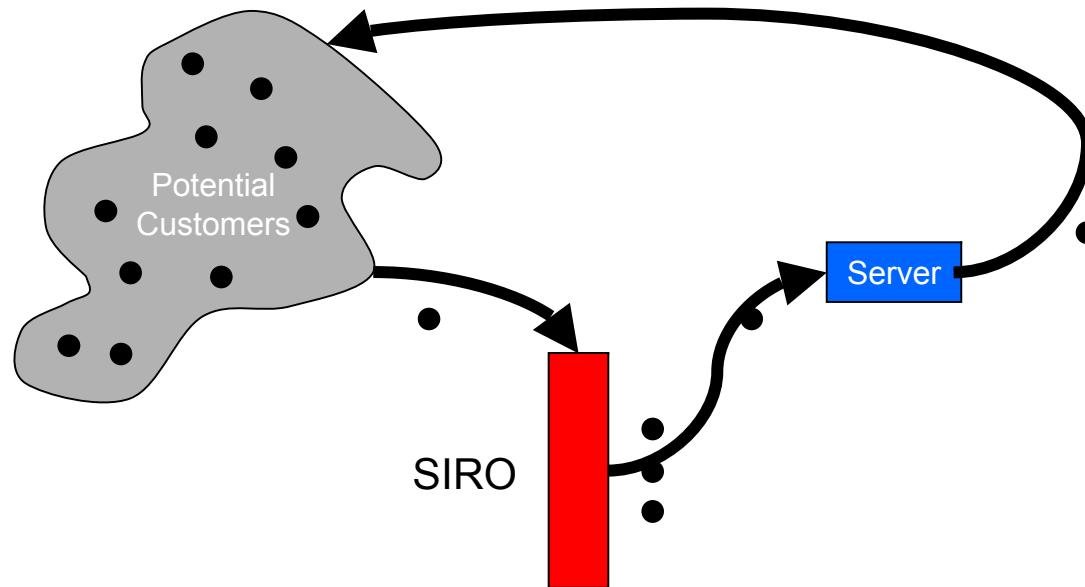
Characteristics of queuing systems

- Queue discipline
 - Last in First Out



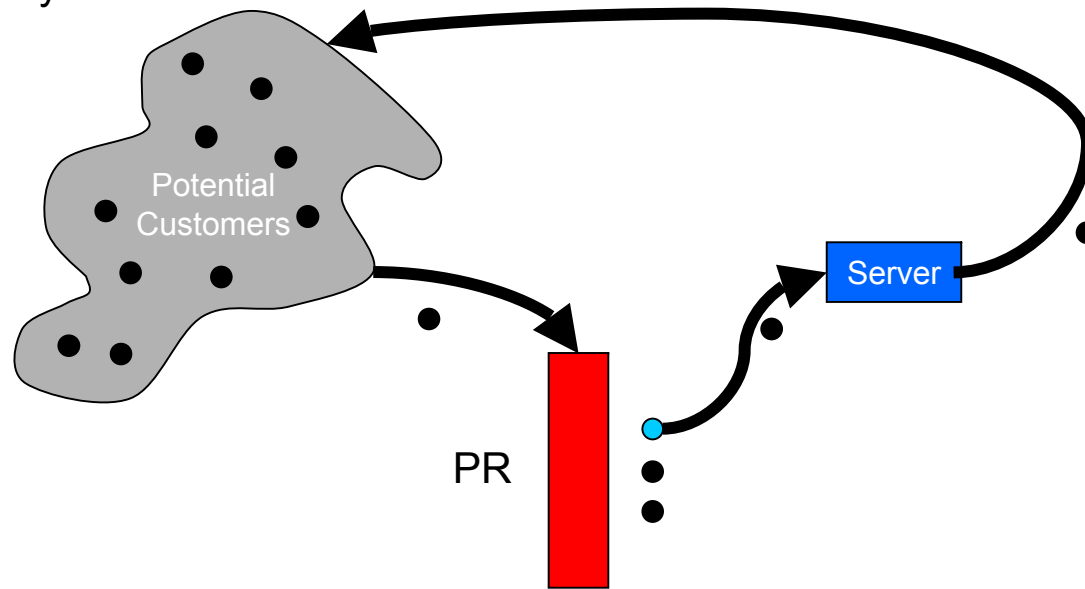
Characteristics of queuing systems

- Queue discipline
 - Service In Random Order



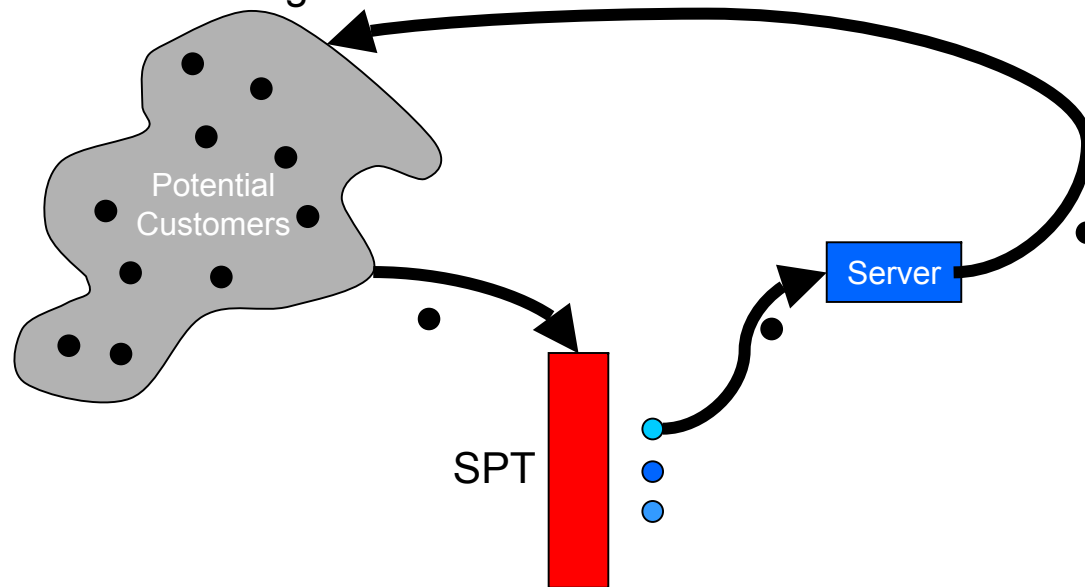
Characteristics of queuing systems

- Queue discipline
 - Priority



Characteristics of queuing systems

- Queue discipline
 - Shortest Processing First

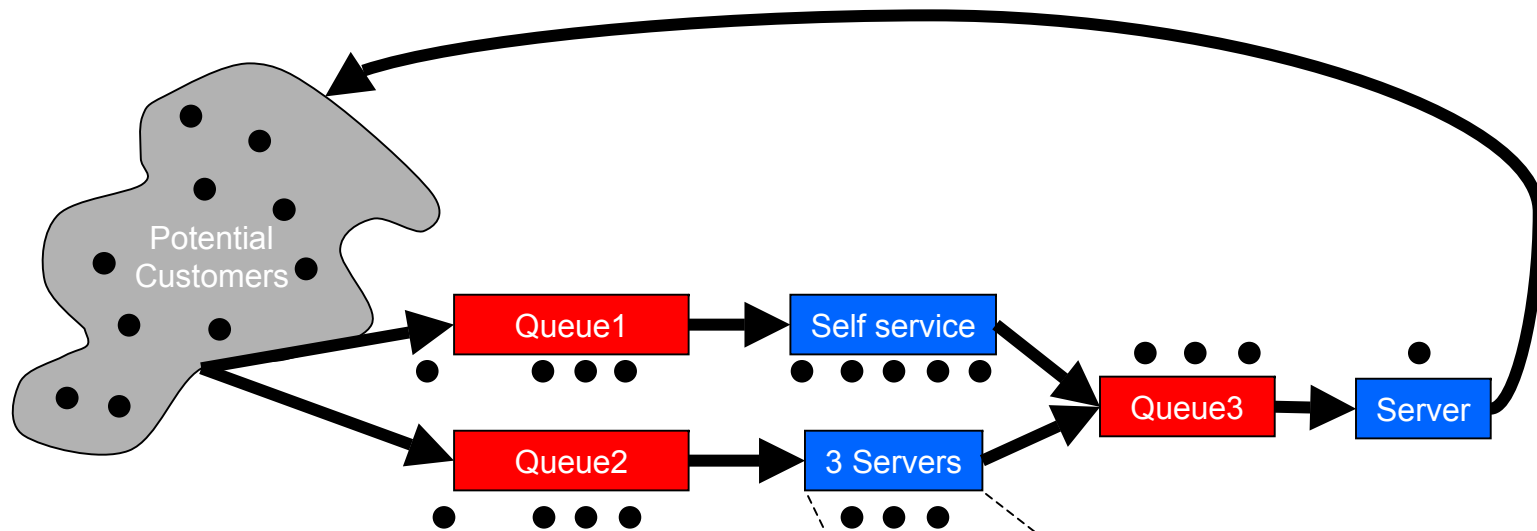


Characteristics of queuing systems

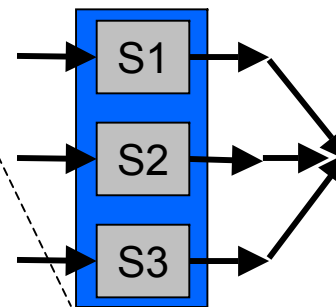
- Service times and the service mechanism
 - Service times of successive arrivals: S_1, S_2, S_3, \dots
 - Service times may be constant or random duration
 - $\{S_1, S_2, S_3, \dots\}$ is characterized as a sequence of independent, identically distributed random variables
 - Service times may be identically distributed for all customers of a given class, different class customers may have different service type distributions
(Consider the queues at an airline terminal for first class, coach customers)

Characteristics of queuing systems

- Service times and the service mechanism



- single server $c=1$
- multiple server $1 < c < \infty$
- self/unlimited servers $c = \infty$



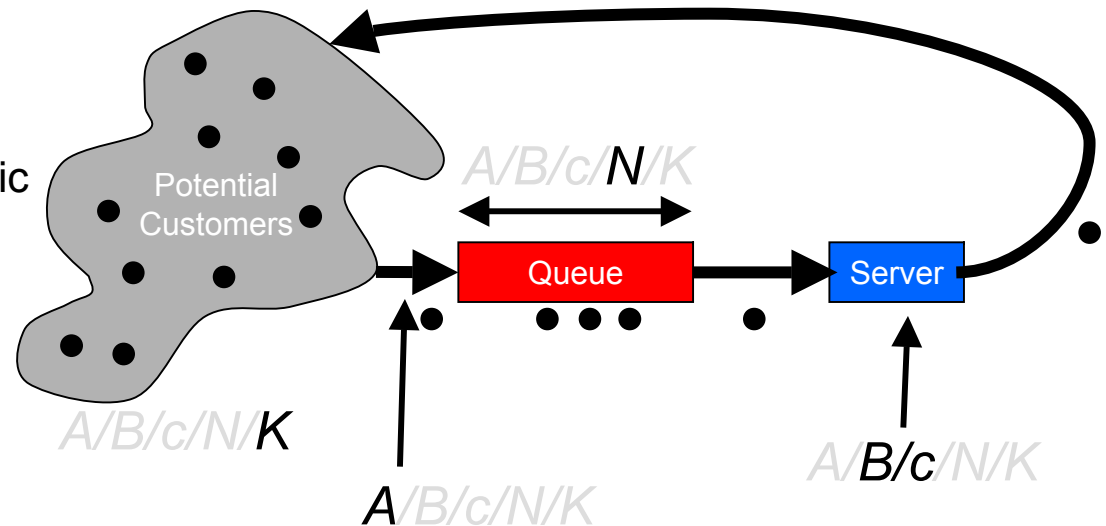
Queuing Notation

- $A/B/c/N/K$

- A : Interarrival-time distribution
- B : Service-time distribution
- c : The number of parallel servers
- N : System capacity
- K : The size of the calling population

- Notation for A and B :

- M : Exponential or Markov
- D : Constant or deterministic
- E_k : Erlang of order k
- PH : Phase-type
- H : Hyperexponential
- G : Arbitrary or general
- GI : General independent



Queuing Notation

- $M/M/1/\infty/\infty$:
 - Exponentially distributed interarrival times, exponentially distributed service times, single server, infinite queue capacity, infinite population of potential arrivals
 - If N and K are infinite, they can be dropped: $M/M/1$

Queuing Notation for Parallel Server Systems

P_n	Steady-state probability of having n customers in system
$P_n(t)$	Probability of n customers at time t
λ	Arrival rate
λ_e	Effective arrival rate
μ	Service rate of one server
ρ	Server utilization
A_n	Interarrival time between customers $n-1$ and n
S_n	Service time of n th arriving customer
W_n	Total time spent in system by customer n
W_n^Q	Total time spent in queue by customer n
$L(t)$	The number of customers in system at time t
$L_Q(t)$	The number of customers in queue at time t
L	Long-run time-average number of customers in system
L_Q	Long-run time-average number of customers in queue
w	Long-run average time spent in system per customer
w_Q	Long-run average time spent in queue per customer

Long-run measures of performance of queuing systems

P_n	Steady-state probability of having n customers in system
$P_n(t)$	Probability of n customers at time t
λ	Arrival rate
λ_e	Effective arrival rate
μ	Service rate of one server
ρ	Server utilization
A_n	Interarrival time between customers $n-1$ and n
S_n	Service time of n th arriving customer
W_n	Total time spent in system by customer n
W_n^Q	Total time spent in queue by customer n
$L(t)$	The number of customers in system at time t
$L_Q(t)$	The number of customers in queue at time t
L	Long-run time-average number of customers in system
L_Q	Long-run time-average number of customers in queue
w	Long-run average time spent in system per customer
w_Q	Long-run average time spent in queue per customer

Long-run measures of performance of queuing systems

- Time-weighted average number of customers in system:

$$\hat{L} = \frac{1}{T} \sum_{i=0}^{\infty} iT_i = \sum_{i=0}^{\infty} i \frac{T_i}{T}$$

- Estimated value for L is weighted average of number of customers in system, weighted by the fraction of the time that i customers are in the system

- Equivalently,

$$\hat{L} = \frac{1}{T} \sum_{i=0}^{\infty} iT_i = \frac{1}{T} \int_0^T L(t) dt$$

- For systems that exhibit long term stability,

$$\hat{L} = \frac{1}{T} \int_0^T L(t) dt \rightarrow L \quad \text{as } T \rightarrow \infty$$

- This measure can be applied to subsystem of a queuing system, as well as the overall system

Time Average Number in System

- $L_Q(t)$ is the number of customers waiting in line at time t

$$\hat{L}_Q = \frac{1}{T} \sum_{i=0}^{\infty} iT_i^Q = \frac{1}{T} \int_0^T L_Q(t) dt \rightarrow L_Q \text{ as } T \rightarrow \infty$$

- In the steady-state condition for a stable system, the expected number of customers waiting approaches a constant value, L_Q

Average time spent in system per customer, w

- N arrivals during $[0, T]$; The time each customer spends in the system is:
 W_1, W_2, \dots, W_N
- Average system time: The average time spent in system per customer:

$$\hat{w} = \frac{1}{N} \sum_{i=1}^N W_i$$

- For stable systems,

$$\text{As } N \rightarrow \infty, \hat{w} \rightarrow w$$

Average time spent in system per customer, w

- Consider the time spent in queue separately:

$$\hat{w}_Q = \frac{1}{T} \sum_{i=1}^N W^Q_i$$

- For stable systems,

$$\text{As } N \rightarrow \infty, \hat{w}_Q \rightarrow w_Q$$

The Conservation Equation

As $T \rightarrow \infty$, $N \rightarrow \infty$, $L = \lambda w$

- For almost all queuing systems or subsystems, independent of number of servers, queue discipline, etc.
- Little's Equation:
 - The average number of customers in the system at an arbitrary point in time is equal to the average number of arrivals per time unit, times the average time spent in the system
- Use this as a quick analytical tool, or to verify sanity of a simulation

Server Utilization

- Server utilization is the percentage of time that the server is busy serving a customer

- In a $G/G/1$ queue,

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

- For a stable system, $\lambda < \mu$ (the arrival rate must be less than service rate)
 - λ/μ is also called the offered load, a measure of workload on system
- If $\lambda/\mu > 1$, system is unstable and queue length grows without bound

Server Utilization

- In a $G/G/c$ queuing system,

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

- Likewise for or a stable multiserver system, $\lambda < c\mu$ (the arrival rate must be less than service rate)

Steady-State Behavior of Infinite-Population Markovian Models

- $M/M/c/N/\infty$, $M/G/c/N/\infty$
- Assumptions:
 - Infinite population
 - Arrivals: Poisson process with rate λ arrivals per time unit
 - Interarrival time: Exponentially distributed with mean $1/\lambda$
 - Service time: Exponentially distributed or arbitrary
 - Queue discipline: FIFO
- Markovian model: Exponentially distributed arrival process
- Static equilibrium (steady-state):
 - $P(L(t) = n) = P_n(t) = P_n$
 - System state (number of customers in systems) is independent of time
 - If system is stable, it is generally either
 - approaching static equilibrium
 - staying in static equilibrium

Steady-State Behavior of Infinite-Population Markovian Models

- Average number of customers in system, L :

$$L = \sum_{n=0}^{\infty} nP_n$$

– time average number of customers = average over number of customers

- Given L , apply Little's equation to find other parameters: $L = \lambda w$

– average customer time in system

$$w = \frac{L}{\lambda}$$

– average customer time in queue
(time in system-service time)

$$w_Q = w - \frac{1}{\mu}$$

– average number of customers in queue

$$L_Q = \lambda w_Q$$

- For **infinite** calling population, for system to be stable: $\rho = \frac{\lambda}{c\mu} < 1$

Single Server Queues with Poisson Arrivals and Unlimited Capacity: $M/G/1$

- Assume server has service times with mean $1/\mu$ and variance σ^2
- If $\rho = \lambda/\mu < 1$, system is stable and has steady-state characteristics:

ρ	$\frac{\lambda}{\mu}$
L	$\rho + \frac{\lambda^2 \left(\frac{1}{\mu^2} + \sigma^2 \right)}{2(1-\rho)} = \rho + \frac{\rho^2 (1 + \sigma^2 \mu^2)}{2(1-\rho)}$
W	$\frac{1}{\mu} + \frac{\lambda \left(\frac{1}{\mu^2} + \sigma^2 \right)}{2(1-\rho)}$
W_Q	$\frac{\lambda \left(\frac{1}{\mu^2} + \sigma^2 \right)}{2(1-\rho)}$
L_Q	$\frac{\lambda^2 \left(\frac{1}{\mu^2} + \sigma^2 \right)}{2(1-\rho)} = \frac{\rho^2 (1 + \sigma^2 \mu^2)}{2(1-\rho)}$
P_0	$1 - \rho$

Example 6.9: Able and Baker Again

- Able: $1/\mu=24$ minutes, $\sigma^2=400$ minutes² faster service with high variability
- Baker: $1/\mu=25$ minutes, $\sigma^2=4$ minutes² slower service with low variability
- $\lambda=1/30$ per minute
- Which has the shortest average queue length (L_Q)
- Which has highest $P(\text{no delay})$

	Able	Baker
ρ	.8	.833
L_Q	2.711	2.097
P_0	.2	.167

An M/M/1 Queue

- An M/G/1 queue with exponential service time is an M/M/1 queue
- Mean service time = $1/\mu$ variance, $\sigma^2=1/\mu^2$
- Steady state parameters:

L	$\frac{\lambda}{\mu - \lambda} = \frac{\rho}{1 - \rho}$
W	$\frac{1}{\mu - \lambda} = \frac{1}{\mu(1 - \rho)}$
W_Q	$\frac{\lambda}{\mu(\mu - \lambda)} = \frac{\rho}{\mu(1 - \rho)}$
L_Q	$\frac{\lambda^2}{\mu(\mu - \lambda)} = \frac{\rho^2}{1 - \rho}$
P_n	$(1 - \rho)\rho^n$

An $M/M/1$ Queue - Example 6.11

- An $M/M/1$ queue with service rate $\mu=10$ per hour
- Examine variation of L and w with increasing arrival rate $\lambda=\{5,6,7.2,8.64,9.99,10\}$

λ	5.0	6.0	7.2	8.64	9.99	10.0
ρ	.500	.600	.720	.864	.999	1.0
L	1.00	1.50	2.57	6.35	999	∞
w	.20	.25	.36	.73	100	∞

Homework 5

- Ch 6 – p 248, problem 5, 8