Queuing models

Stochastic throughput

Cathy Wu

1.041/1.200 Transportation: Foundations and Methods

Readings

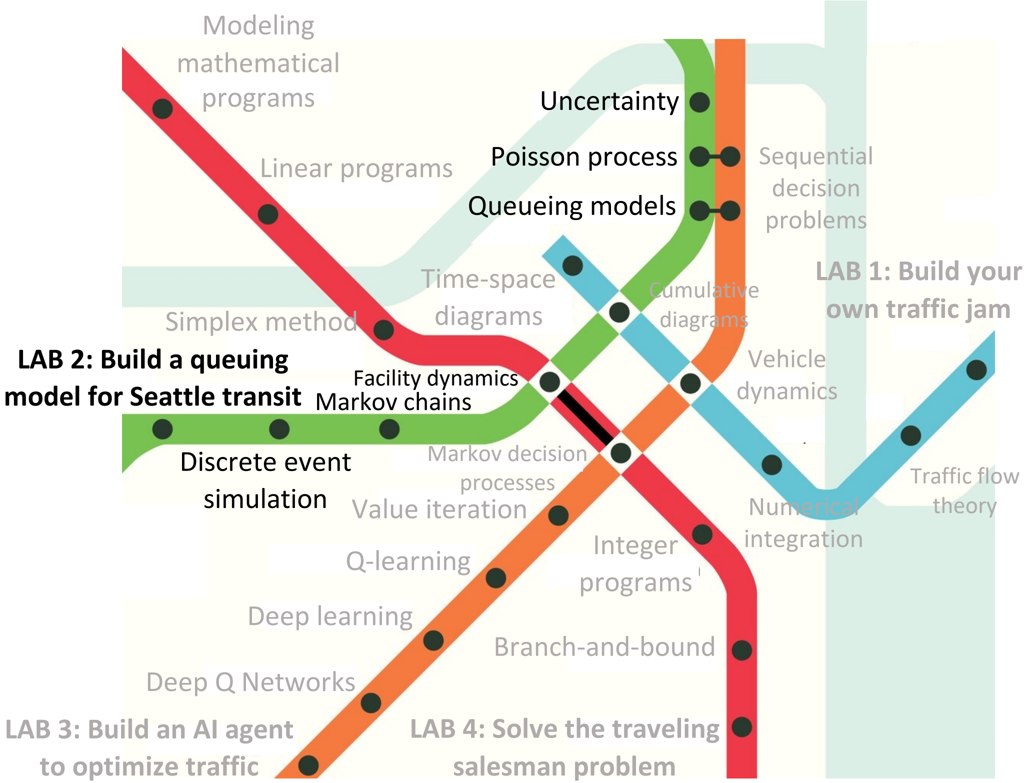
1. Larson, Richard C. and Amedeo R. Odoni. **Urban Operations Research**. Prentice-Hall (1981). Chapter 4: Queueing Theory. URL.

Unit 2: Queuing systems

Unit 2

Modeling

Stochastic



Outline

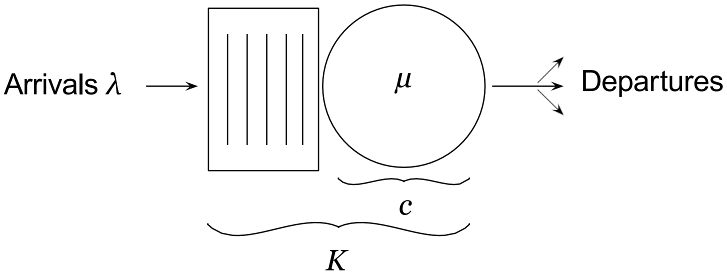
1. Fundamental queueing models
2. Stationary analysis and Little’s law
3. M/M/1: Detailed analysis
4. More queues

Outline

1. **Fundamental queueing models**
2. Stationary analysis and Little's law
3. M/M/1: Detailed analysis
4. More queues

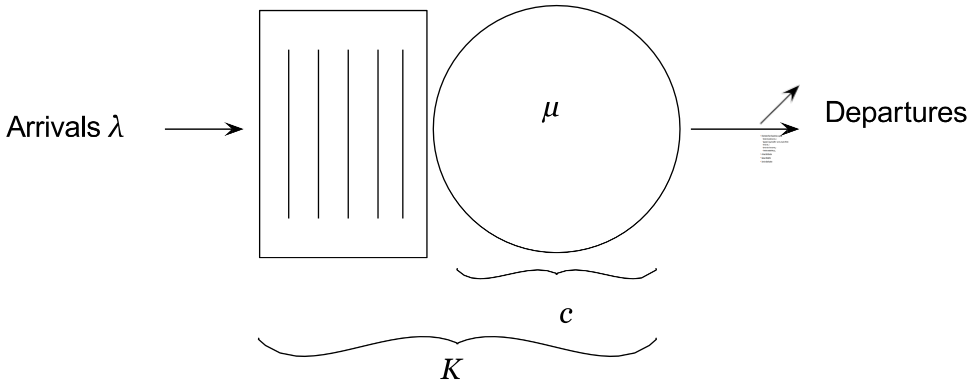
Queueing models

- Customers requiring service are generated over time by an input source
- These customers enter the *queueing system* and join a queue
- At certain times, a member of the queue is selected for service by some rule known as the *queue discipline*.
- The required service is then performed for the customer by the *service mechanism*, after which the customer leaves the queueing system.



Queueing models

- Parameters that characterize a queue
  - Number of parallel servers, c
  - Capacity, K (equal to buffer + servers, may be infinite)
  - Arrival rate, λ
  - Service rate of one server, μ
  - Transition probabilities, p_{ij}
- Arrival distribution
- Queue discipline
- Service distribution



Complete Kendall notation

$$A \;/\; S \;/\; c \;/\; K \;/\; P \;/\; QD$$

- A: inter-arrival time distribution
- S: service time distribution
- c: number of servers
- K: total system size (∞)
- P: population size (∞)
- QD: Queue discipline (FIFO)

Kendall notation

$$A\ /\ S\ /\ c\ /\ K\ /\ P\ /\ QD$$

- Arrival (A) / Service (S) Process
  - Assumption: i.i.d
- Some standard code letters for A and S:
  - M: Exponential (M stands for memoryless/Markovian)
  - D: Deterministic
  - E_k : kth-order Erlang distribution
  - G: General distribution
- Examples:
  - $D/D/1$, lends itself to a graphical analysis (Unit 1)
  - $M/M/c$

Number of servers

- Single server
  - One server for all queued customers
- Multiple server
  - Finite number of “identical” servers operating in a parallel configuration
- Infinite-server
  - A server for every customer

Kendall notation

$$A \,/\, S \,/\, c \,/\, K \,/\, P \,/\, QD$$

- K: total system size, i.e. buffer size + number of servers
- Referred to as “capacity” in queueing theory
- $K < \infty$: finite capacity queues

Queue discipline

- Refers to the order in which members of the queue are selected for service
- FIFO: first-in first-out (a.k.a. FCFS)
  - first customer to arrive is first to depart, no passing
  - Single road lane, airport check-in counters
- LIFO: last-in first-out
  - last customer into queue is first to leave
  - Unboarding cars from a ferry, unboarding a bus from behind
- Priority
  - Customers get served in order of priority (highest to lowest)
  - Flight departures along a runway, priority seating when boarding flights
  - Yields / intersections: priority between approaches
- SIRO: service in random order
- PS: processor sharing
- FIFO is the most common discipline for most transportation applications

Queueing theory - keep in mind

- Queueing theory can provide **insights** and approximation of the main system performance measures.
  - Can enable identification of the location of bottlenecks in networks,
  - Give indications on how to improve the system's performance.
- Most closed-form results involve stationary regime (steady-state) and low-order moments (mean, variance) of the inter-arrival and service time distributions
- Trade-off: realistic model (few available results) vs. tractability (assumptions are questionable)

Outline

1. Fundamental queueing models
2. **Stationary analysis and Little's law**
3. M/M/1: Detailed analysis
4. More queues

Stationary analysis

- State of system: number n of customers in the system
- Steady state condition: system is independent of initial state and has reached its long-term equilibrium characteristics
  - A.k.a. steady state regime, stationary regime
- Given:
  - λ = arrival
  - μ = service rate per server
  - c = number of servers (parallel service channels)

- Quantities of interest:
  - $\bar{N}$: expected number of users in queueing system ($\bar{N} = E[N]$)
  - $\bar{N}_q$: expected number of users in queue ($\bar{N}_q = E[N_q]$)
  - $\bar{T}$: expected time in queueing system per user ($\bar{T} = E[T]$)
  - $\bar{T}_q$ = expected waiting time in queue per user ($\bar{T}_q = E[T_q]$)
- 4 unknowns $\Longrightarrow$ need 4 equations
- Also of interest: (P_n): stationary queue length distribution
  - $\sum_{i=0}^{\infty} P_i = 1$

Stability

- A system is said to be **stable** if its long run averages (N, T) exist and are finite
- Consider an infinite capacity queue:
  - Traffic intensity (also called utilization factor):

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

    - $c\mu$: queue service rate.
  - The queue is stable if and only if $\rho < 1$
  - If a system is unstable, its long run measures are meaningless
  - Note:
    - This is necessary only for infinite capacity queues
    - Finite capacity queues have bounded queue lengths, and are therefore always stable
    - Stable systems $\rightarrow$ a steady state condition exists

Little's law

- John Little, MIT Institute Professor
- Proof in: "A proof for the queuing formula: $L = \lambda W$" (1961), Operations Research
- [Little's Law as viewed on its 50th Anniversary]() (INFORMS)
- $\bar{N} = \lambda \bar{T}$ (1)
  - $\bar{N}$: expected number of vehicles in the system
  - λ: system arrival rate
  - $\bar{T}$: expected time in the system
- Assumption: The system is in a stationary regime

- No assumptions/restrictions on the:
  - inter-arrival and service time distributions
  - queue discipline
  - number of servers

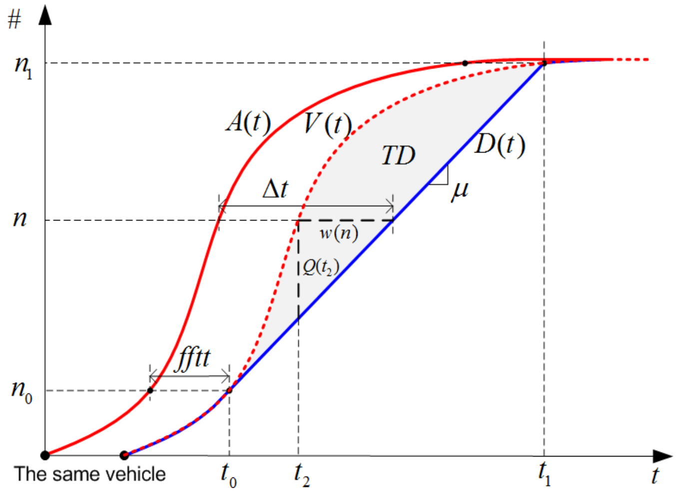
- For several classes/categories of users, Little's law applies to each category
- If you consider a finite time horizon (i.e. $\tau < \infty$) then stationarity is not required.

Little's Law (1961) – deterministic version [Unit 1]

- Simple relationship between arrival rate, average queue length, and average delay (waiting time).

  - Definition (Average arrival rate): $\lambda = \frac{n_1 - n_0}{t_1 - t_0}$
  - The delay of vehicle n: $w(n)$
  - Queue at t_2: $Q(t_2)$
  - Total Delay: $TD = \int_{t_0}^{t_1} [V(t) - D(t)]dt = \int_{t_0}^{t_1} Q(t)dt$

- Assumption 1: Finite time window & vehicles
- Assumption 2: Conservation of vehicles (all arriving vehicles eventually depart)
- Then: $\bar{Q} = \lambda \bar{w}$



1961, John Little, MIT Institute Professor; See "Little's Law as Viewed on its 50th Anniversary" (INFORMS)

Little's law

$$\bar{N} = \lambda \bar{T} \quad (1)$$

- $\bar{N}$: expected number of vehicles in the system
- λ: system arrival rate
- $\bar{T}$: expected time in the system

$$\bar{N}_q = \lambda \bar{T}_q \quad (2)$$

- $\bar{N}_q$: expected number of vehicles in the buffer
- λ: system arrival rate
- $\bar{T}_q$: expected time in the buffer

Relationships between $\bar{N}, \bar{N}_q, \bar{T}$, and $\bar{T}_q$

- Little's law:
  - $\bar{N} = \lambda \bar{T}$ (1)
  - $\bar{N}_q = \lambda \bar{T}_q$ (2)
- $\bar{T} = \bar{T}_q + \frac{1}{\mu}$ (3)
  - μ = service rate (Hz) $\Longrightarrow$ expected service time = $\frac{1}{\mu}$
- $\bar{N} - \bar{N}_q = \frac{\lambda}{\mu}$ (for M/M/1) (4)
  - which represents the expected number of vehicles under service (in steady-state)

- Obtain one of the performance measures, the other three can then be deduced
- Let's try to obtain $\bar{N}$.
  - The determination of $\bar{N}$ may be hard or easy depending on the type of queueing model at hand
  - It is easy for $M/M/1$ and quite easy for $M/M/s$ and for $M/G/1$
- In general: $\bar{N} = \sum_{n=0}^{\infty} nP_n$, where P_n is the probability that there are n customers in the system

Outline

1. Fundamental queueing models
2. Stationary analysis and Little’s law
3. **M/M/1: Detailed analysis**
4. More queues

Analysis of queueing models

- Closed-form expressions for the main performance measures typically involve:
  - stationary regime (i.e. steady state analysis)
  - specific distributional assumptions
- Computational techniques allow us to numerically evaluate performance measures for more general queues, and also for transient regime (i.e. dynamic analysis)
- $M/M/1$ queueing system: "simple" to analyze
- General strategy:
  - Compute steady state probabilities P_n
  - Compute $\overline{N} = \sum_{n=0}^{\infty} nP_n$
  - Obtain $\overline{N}_q$, $\overline{T}$, and $\overline{T}_q$

Detailed analysis of $M/M/1$ queueing system

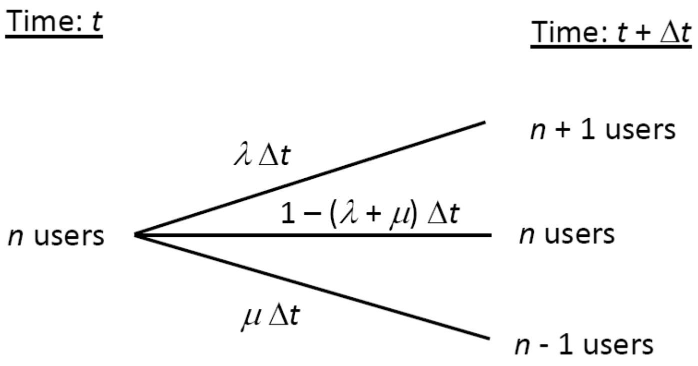
- (Recall) Inter-arrival times:

$$f_X(t) = \lambda e^{-\lambda t} \quad t \geq 0; \;\; E[X] = \frac{1}{\lambda}; \quad \sigma_X^2 = \frac{1}{\lambda^2}$$

- (Recall) Service times:

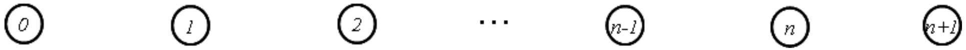
$$f_S(t) = \mu e^{-\mu \mathrm{t}} \quad t \geq 0; \quad E[S] = \frac{1}{\mu}; \quad \sigma_S^2 = \frac{1}{\mu^2}$$

- From the properties of exponential r.v.'s, the probabilities of transitions in the next Δt:

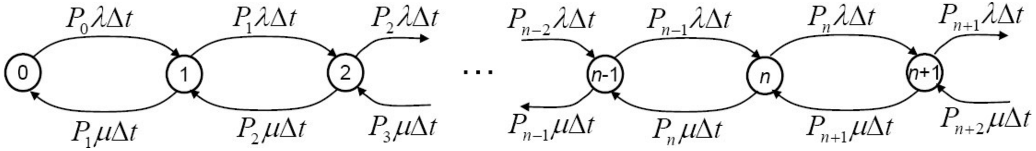


State transition diagram for $M/M/1$

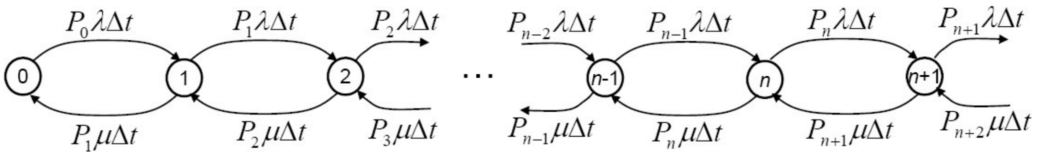
- States (number of "customers" in the system):



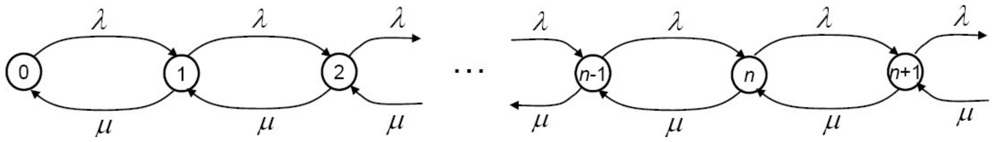
- The probability of observing a transition from state i to state j during the next Δt with the system in steady-state:



State transition diagram for $M/M/1$



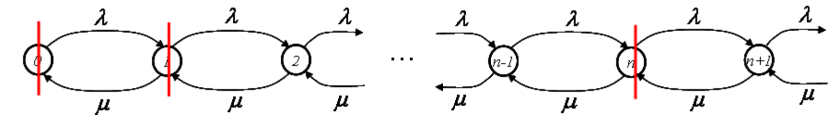
- Another way to represent this **State transition diagram**:
  - Nodes: states
  - Arcs: possible state transitions



Observing the diagram from two points

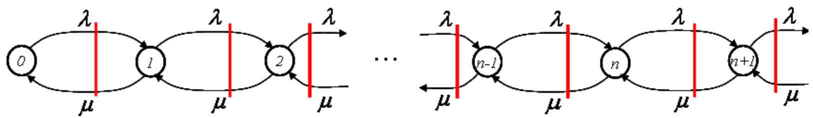
1. At a state:

$$\lambda P_0 = \mu P_1 \qquad (\lambda+\mu)P_1 = \lambda P_0 + \mu P_2 \qquad (\lambda+\mu)P_n = \lambda P_{n-1} + \mu P_{n+1}$$



2. Between states:

$$\lambda P_0 = \mu P_1 \qquad \lambda P_1 = \mu P_2 \qquad \lambda P_n = \mu P_{n+1}$$



- The two sets of equations yield the same solutions

$M/M/1$: deriving P_0 and P_n

1. $P_1 = \frac{\lambda}{\mu} P_0, \quad P_2 = \left(\frac{\lambda}{\mu}\right)^2 P_0, \ldots, \quad P_n = \left(\frac{\lambda}{\mu}\right)^n P_0$
2. $\sum_{n=0}^{\infty} P_n = 1, \ \Rightarrow P_0 \sum_{n=0}^{\infty} \left(\frac{\lambda}{\mu}\right)^n = 1, \ \Rightarrow P_0 = \frac{1}{\sum_{n=0}^{\infty}\left(\frac{\lambda}{\mu}\right)^n}$
3. For $|x| < 1, \ \sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$
4. Define: $\rho = \frac{\lambda}{\mu}$

$$P_0 = \frac{1}{\sum_{n=0}^{\infty} \rho^n} = 1 - \rho$$
$$P_n = \rho^n (1 - \rho)$$

$M/M/1$: deriving $\bar{N}, \bar{N}_q, \bar{T}$, and $\bar{T}_q$

$$\bar{N} = \sum_{n=0}^{\infty} nP_n$$

$$= \sum_{n=0}^{\infty} n\rho^n (1-\rho)$$

$$= (1-\rho)\sum_{n=0}^{\infty} n\rho^n$$

$$= (1-\rho)\rho\sum_{n=0}^{\infty} n\rho^{n-1}$$

$$= (1-\rho)\rho\frac{d}{d\rho}\sum_{n=0}^{\infty} \rho^n$$

$$= (1-\rho)\rho\frac{d}{d\rho}\left(\frac{1}{1-\rho}\right)$$

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

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

$$\bar{T} = \frac{\bar{N}}{\lambda} = \frac{\lambda}{\mu-\lambda}\cdot\frac{1}{\lambda} = \frac{1}{\mu-\lambda}$$

$$\bar{T}_q = \bar{T} - \frac{1}{\mu} = \frac{1}{\mu-\lambda} - \frac{1}{\mu} = \frac{\lambda}{\mu(\mu-\lambda)}$$

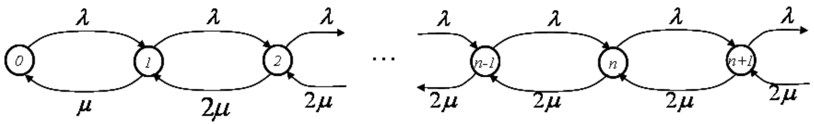
$$\bar{N}_q = \lambda\bar{T}_q = \frac{\lambda^2}{\mu(\mu-\lambda)}$$

Outline

1. Fundamental queueing models
2. Stationary analysis and Little's law
3. M/M/1: Detailed analysis
4. **More queues**
   a. $M/M/2$
   b. $M/M/c$
   c. $M/M/c/K$
   d. $M/M/c/c$ - Erlang loss model
   e. $M/D/1$

$M/M/2$ queueing system

- What happens if we have two parallel, independent servers, each with service rate μ and exponentially distributed service times?
- State transition diagram:



- Balance equations: $\begin{cases} \lambda P_0 = \mu P_1 \\ \lambda P_1 = 2\mu P_2 \\ \lambda P_2 = 2\mu P_3, \ldots \end{cases}$

$$\begin{cases} P_1 = \frac{\lambda}{\mu} P_0 \\ P_2 = \frac{\lambda}{2\mu} P_1 = \frac{1}{2}\left(\frac{\lambda}{\mu}\right)^2 P_0 \\ P_3 = \frac{\lambda}{2\mu} P_2 = \left(\frac{1}{2}\right)^2 \left(\frac{\lambda}{\mu}\right)^3 P_0 \\ P_n = \left(\frac{1}{2}\right)^{n-1} \left(\frac{\lambda}{\mu}\right)^n P_0 \end{cases}$$

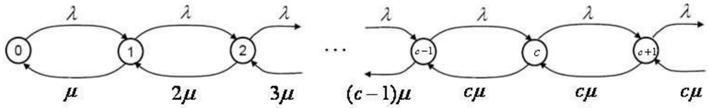
$M/M/2$ queue

$$P_0 = \frac{1-\frac{\lambda}{2\mu}}{1+\frac{\lambda}{2\mu}} \quad \text{for } \lambda < 2\mu$$

$$P_n = 2\left(\frac{\lambda}{2\mu}\right)^n P_0, \qquad n \geq 1$$

$$\ldots\ \bar{N} = \frac{\frac{\lambda}{\mu}}{\left(1+\frac{\lambda}{2\mu}\right)\left(1-\frac{\lambda}{2\mu}\right)}$$

$M/M/c$ queue



- This model is a reasonable assumption at toll booths on turnpikes or at toll bridges where there is often more than one toll booth open.
- Traffic intensity / utilization factor: $\rho = \frac{\lambda}{c\mu}$
- Stability: $\frac{\lambda}{c\mu} < 1$
- Stationary dbn:

$$P_k = \begin{cases} \frac{(\lambda/\mu)^k}{k!} P_0, & k = 1, 2, \dots, c-1 \\ \frac{(\lambda/\mu)^k}{c!\, c^{k-c}} P_0, & k = c, c+1, \dots \end{cases}$$

$$P_0 = \left[\frac{(\lambda/\mu)^c}{c!\,(1-\rho)} + \sum_{k=0}^{c-1} \frac{(\lambda/\mu)^k}{k!} \right]^{-1}$$

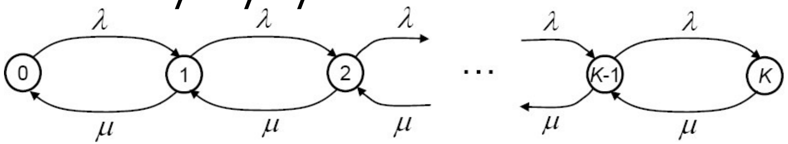
- Expected queue length (in the buffer)?

$M/M/c$ queue

- Little's formula: $T_q = \frac{N_q}{\lambda}$
- $T = T_q + \frac{1}{\mu}$
- To obtain N:
  1. Little's formula: $N = \lambda T$
  2. $N = N_q + \frac{\lambda}{\mu}$

$M/M/c/K$

- Finite capacity queue: $M/M/1/K$



$$P_n = \frac{1-\rho}{1-\rho^{K+1}}\rho^n \quad n = 0, 1, \dots, K$$

- The queue is always stable ($\forall\rho$), so steady state is always reached
- When system is full: arrivals are lost
- *Effective* arrival rate (i.e. rate of arrivals that actually enter the system): $\lambda(1 - P_K)$
- Careful when applying Little's Law! Count only the vehicles that actually join the system:

$$\lambda' = \lambda(1 - P_K)$$
$$\bar{N} = \lambda' T$$

$M/M/c/c$ - Erlang loss model

- $M/M/c/c$: Erlang loss model
- First queueing model to be investigated
- Agner Krarup Erlang
  - Danish telephone engineer, who investigated it in the early 1900's as a model for telephone switch which can handle only c calls
  - Queueing theory pioneer
  - The theory of stochastic processes was not yet developed at the time
  - Erlang derived a formula for the proportion of lost calls, Erlang loss formula:

$$P_c = \frac{(\lambda/\mu)^c/c!}{\sum_{i=0}^{c}(\lambda/\mu)^i/i!}$$

$M/M/c/c$ - Insensitivity

- Erlang assumed exponential service times, but conjectured that it would hold for generally distributed service times
- Insensitivity: the loss probability is insensitive to the form of the service time distribution; it depends only on its expectation.
- This was not proved until the 1960's

$$P_c = \frac{(\lambda/\mu)^c/c!}{\sum_{i=0}^{c}(\lambda/\mu)^i/i!}$$

- Loss probability holds for $M/G/c/c$ queues

$M/G/c/c$ - Erlang loss model

- Actually, the insensitivity property also holds for the stationary distribution
- Insensitivity: the stationary distribution is insensitive to the form of the service time distribution; it depends only on its expectation.

$$P_n = \frac{(\lambda/\mu)^n/n!}{\sum_{i=0}^{c}(\lambda/\mu)^i/i!}, \forall n \in [0, c]$$

- Insensitivity property → Erlang loss model is of wide interest
  - Model is commonly used for the analysis of telecommunication systems, also: urban service systems, inventory, reliability

$M/D/1$ queue

- Has been used to model vehicles on a lane at signalized urban intersections
- Exponentially distributed inter-arrival times
- Deterministic service distribution
- One server

- Recall the traffic intensity: $\rho = \frac{\lambda}{\mu}$
  - ρ: traffic intensity
  - λ: arrival rate [veh/unit time]
  - μ: service rate [veh/unit time]

$M/D/1$ queue

- For a stable queue ($\rho < 1$):
  - Expected number of vehicles in the buffer [veh]:
$$N_q = \frac{\rho^2}{2(1-\rho)}$$
  - Expected waiting time in the buffer (per veh)
$$T_q = \frac{\rho}{2\mu(1-\rho)}$$
  - Expected time in the system: sum of the expected waiting time and the expected service time:
$$T = \frac{2-\rho}{2\mu(1-\rho)}$$
- **Note**: traffic intensity: $\rho < 1$, then:
  - the $D/D/1$ queue predicts no queue formation,
  - models with probabilistic arrivals/departures (e.g. $M/D/1$) predict queue formations under such conditions.

References

1. Larson, Richard C. and Amedeo R. Odoni. **Urban Operations Research**. Prentice-Hall (1981). Chapter 4: Queueing Theory.

2. John Little, Little's Law as Viewed on its 50th Anniversary. Operations Research, vol 59. 2011. https://www.informs.org/Blogs/Operations-Research-Forum/Little-s-Law-as-Viewed-on-its-50th-Anniversary

3. Slides adapted from Carolina Osorio