Spring 2024

# Queueing network models

Network queues and delays

**Cathy Wu**

1.041/1.200 Transportation: Foundations and Methods

# Readings

1. (Optional) Tom Slater. Queueing Networks. 2000. URL.

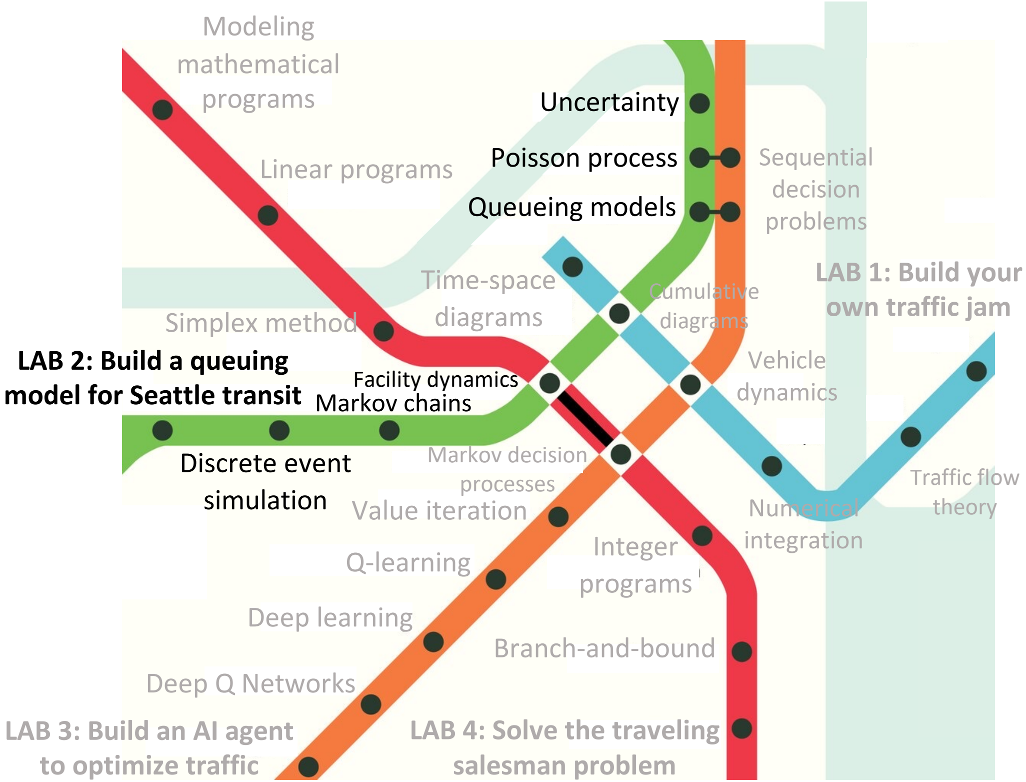
# Unit 2: Queuing systems

○

Unit 2

Modeling

**Stochastic**



# Outline

1. Queueing networks

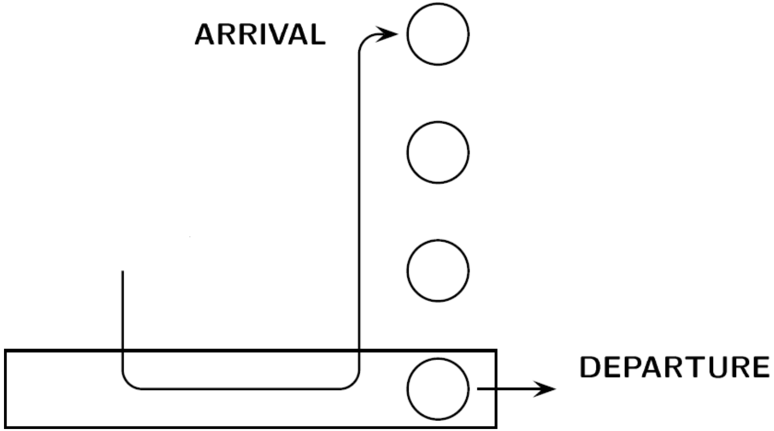
# Facilities in transportation

- So far, we have studied stochastic models for a single transportation facility:
- Examples:
    - A single bus stop
    - A single intersection
    - A single road (this lecture)
    - A single gas station (may have multiple gas pumps)
    - A single subway station (may have multiple trains)
    - A single airport (may have multiple runways)
    - A single port (may have many docking areas)

- This lecture: multiple facilities, that affect one another (queuing networks)
- Examples
    - A road network
    - A bus system
    - A rail / subway system

# Vertical queues

- Vertical queues, also called point queues
- Fail to capture:
  - time needed to reach the physical queue
  - upon service completion, time needed for the physical queue to advance (i.e. for the newly available slot to appear upstream)
  - Example: backward/negative/jam wave in congested urban traffic

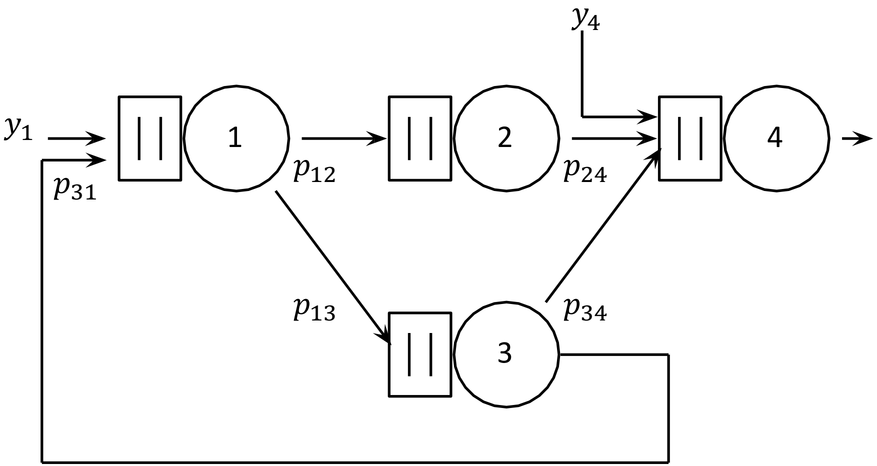


# Outline

1. **Queueing networks**
   a. Departure process
   b. Burke’s theorem
   c. Tandem infinite capacity networks
   d. Jackson networks
   e. Finite capacity queueing networks

# Queueing networks

- System composed of several subsystems that have different service/capacity characteristics: queueing network.
- Open - closed – mixed
  - External arrivals
- Finite population - infinite.
- Vehicle classes/categories
- Routing
  - deterministic
  - stochastic ($p_{ij}$)

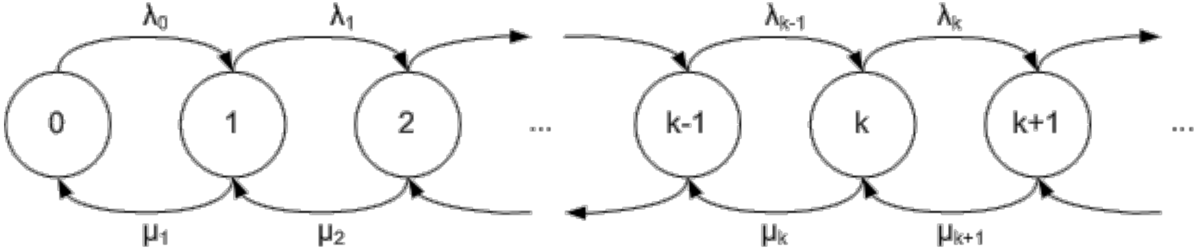


# Departure process

- In queueing networks, the departure process of one queue is the arrival process of another.
- Example: $M/M/c$ stable queue with arrival rate $\lambda$
  - In the stationary regime, the departure rate is equal to the arrival rate.
  - This flow conservation property is true for all systems in a stationary regime.
- Additionally, for an $M/M/c$ queue:
  - The departure process is also Poisson with rate $\lambda$
  - At time $t$, the number of customers in the queue is independent of the departure process prior to time $t$.
- Significance: the inter-arrival and inter-departure distributions are the same!
  - Known as the *equivalence property*
  - Allows the stationary analysis of $M/M/c$ queues to be carried out as independent queues
  - Among all FIFO $M/G/c$ queues, only $M/M/c$ have this property
  - So, analysis of networks with a single queue with non-exponential service time  distribution becomes intricate

# Recall: Birth-death process

- Special case of continuous-time Markov process where the state transitions are of only two types:
  - "births", which increase the state variable by one
  - "deaths", which decrease the state by one
- Can also stay in the same state
  - Example (M/M/1): with probability $1 - (\lambda + \mu)$
- Examples (for small time $\Delta t$)
  - $M/M/1$
  - $M/M/c$
  - $M/M/c/K$



# Burke’s Theorem (1956)

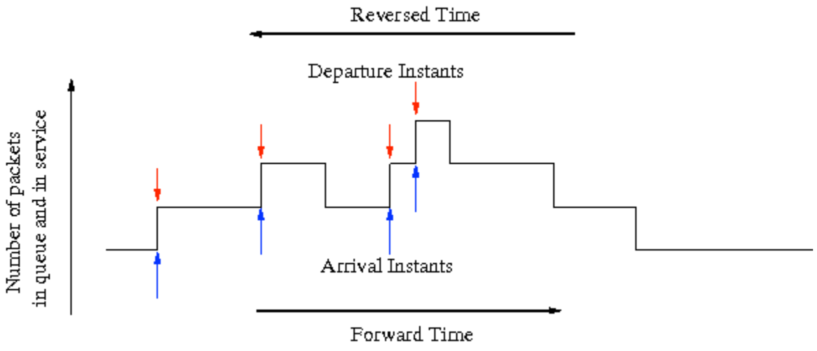
## Theorem

If an $M/M/c$ queue with arrivals according to a Poisson process with rate parameter $\lambda$ is in the steady state, then:

1. The departure process is a Poisson process with rate parameter $\lambda$.
2. At time $t$ the number of customers in the queue is independent of the departure process prior to time $t$.

**Proof sketch**

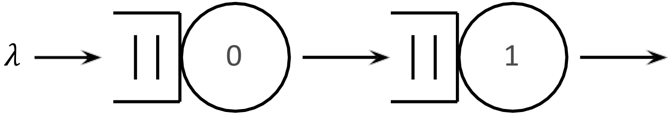
- By Kolmogorov's criterion for reversibility, any birth-death process is a reversible Markov chain, including $M/M/c$ queues.
- Note that the arrival instants in the forward Markov chain are the departure instants of the reversed Markov chain. Thus the departure process is a Poisson process of rate $\lambda$.
- Moreover, in the forward process the arrival at time $t$ is independent of the number of customers after $t$. Thus in the reversed process, the number of customers in the queue is independent of the departure process prior to time $t$.



# Analysis of queueing networks

- Exact methods
  - Analytical
    - Global balance equations
    - Closed-form expressions available: For some simple networks we can analyze the queues as if they were independent (product form joint stationary distribution)
  - Numerical (e.g. global balance equations)
- Approximation methods
  - Analytical: use of bounds or decomposition methods
  - Simulation

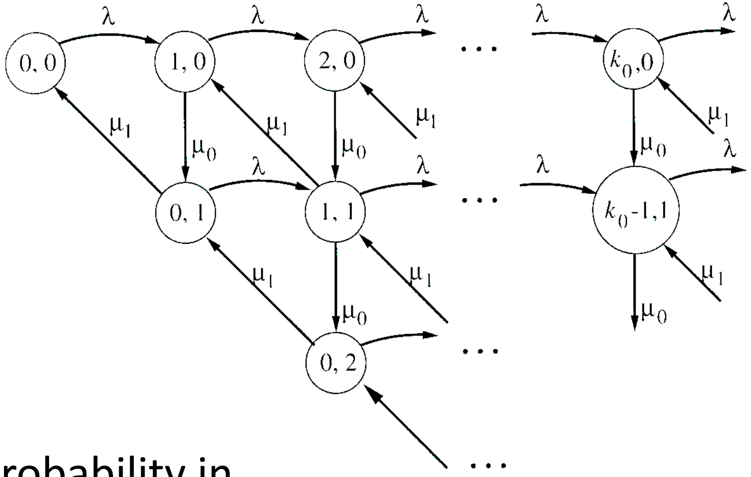
# Global balance equations



- Tandem infinite capacity network
- Two queues, each with one server and infinite capacity, in a tandem network

Steady state?

$$P(n_0, n_1) = ?$$



Utilization rate, traffic density

$$\rho_0 \coloneqq \frac{\lambda}{\mu_0},\ \ \rho_1 \coloneqq \frac{\lambda}{\mu_1}$$

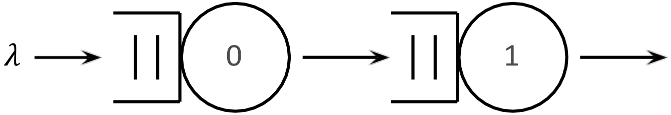
Global balance: probability out = probability in

$$Q_{i_1 j_1} \sum_{(i_2,j_2)\neq(i_1,j_1)} p_{i_1,j_1\to i_2,j_2} = \sum_{(i_2,j_2)\neq(i_1,j_1)} Q_{i_2 j_2} p_{i_2,j_2\to i_1,j_1}$$

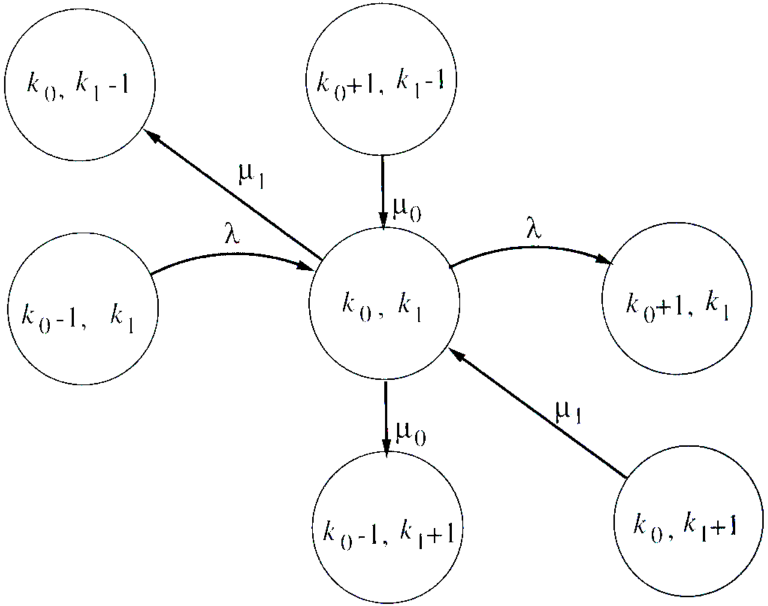
Queue 0, Queue 1

Source: Trivedi (2002)

# Global balance equations

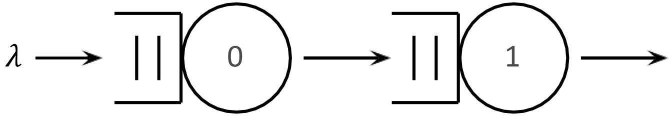


- Tandem infinite capacity network
- Two queues, each with one server and infinite capacity, in a tandem network



Source: Trivedi (2002)

# Example: tandem infinite capacity network



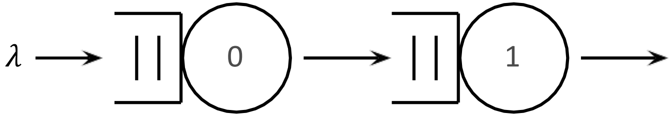
- We can show that this process is still Markovian / memoryless, because all processes are exponential

- Finding the global balance equation (the infinitesimal generator deriving from the “Chapman-Kolmogorov differential equations”) is not difficult, but…
- … this is the simplest queuing network we can imagine!

- Can we find a more general and simpler way to solve this system?
- We know that the steady-state solution of an Markov Chain is unique, thus to find the solution of an Markov Chain, we only have to find one solution

Adapted from: Renato Lo Cigno

# Example: tandem infinite capacity network

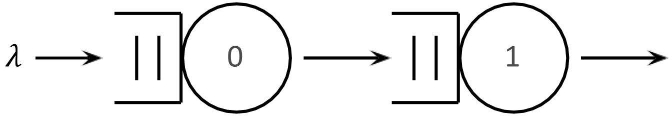


- … yes, but guessing a solution at random …

- Indeed, we can observe that the first queue is entirely independent from the second, thus its behavior should be like a normal $M/M/1$
- But at steady state, jobs arrive randomly, and they are served by the first queue with exponential time services, thus the second queue "sees" as arrival process, the service process of the first one
- We can guess (not really at random) that this second queue behaves like another $M/M/1$ with some $\rho_1$

Adapted from: Renato Lo Cigno

# Example: tandem infinite capacity network



- The two queues are clearly not independent (the input to the second one is the output of the first one!!)
- But what if they behave “as if they were” independent? Then the solution would be the product of the two solutions

$$P(n_0, n_1) = (1-\rho_0)\rho_0^{n_0}(1-\rho_1)\rho_1^{n_1}$$

- Plugging this simple solution in the global balance equations solves them … Bingo!
- This is called a **product form solution**.

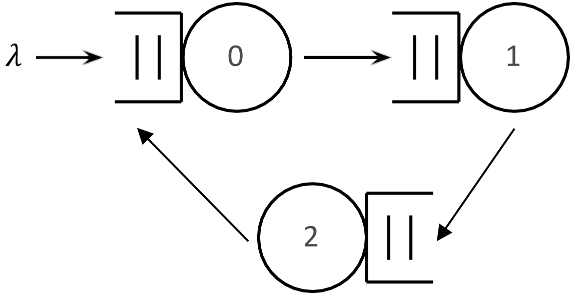
Recall stationary distribution of M/M/1 queue with utilization factor $\rho$:

$$P_n = \rho^n(1-\rho)$$

Adapted from: Renato Lo Cigno

# Jackson networks (1957)

- In 1957, Jim Jackson demonstrated that any network of Markovian queues without losses and FCFS serving discipline admits a product form solution
- Assumptions
  - External arrivals constitute a Poisson process
  - Service times, iid, follow an exponential dbn
  - Routing probabilities are fixed (exogenous)
  - Each queue has infinite capacity/size
  - Each queue is stable: $\forall i, \rho_i < 1$, and thus the overall system is stable
  - FCFS
- Feedback among queues is admitted



Example of queueing network with feedback

# Jackson networks

- Product form stationary distribution
- Network of $m$ queues:

$$P_n = \prod_{i=1}^{m} P_{n_i}^{i}$$

where $P_n$: probability that the network is in state $n$

$n$: network state, $n = (n_1, \ldots, n_m)$

$P^i$: marginal distribution for queue $i$

$P_{n_i}^{i}$ : probability that queue $i$ has $n_i$ vehicles

- Example: single server network:

$$P_n = \prod_{i=1}^{m} (1 - \rho_i)\rho_i^{n_i}$$

# Jackson networks

- Product form stationary distribution

$$P_n = \prod_{i=1}^{m} P_{n_i}^i$$

- Queue lengths behave **AS IF** they were independent $M/M/c$ queues
- Even though the arrival process to a queue within a network is not Poisson!
  - Remarkable aspect of Jackson’s result
  - The network does not decompose into independent M/M/c queues
- Note: departure/internal arrival process
  - Non-feedback (feedforward) networks: the departure process of each queue is Poisson
  - For networks with feedback loops, the actual internal arrival process to a queue is generally not Poisson
- Caution: Analysis of other performance measures (e.g. waiting time dbn) cannot be analyzed as if each queue were an $M/M/c$ queue

# BCMP networks (Baskett, Chandy, Muntz, Palacios, 1975)

- A significant extension to a Jackson network allowing virtually arbitrary customer routing and service time distributions, subject to particular service disciplines.
- Result: Product-form solution stationary distribution
- Assumptions (same as Jackson networks):
  - External arrivals constitute a Poisson process
  - Routing probabilities are fixed (exogenous)
- Relaxations
  - Each of the queues is of one of the following four types:
    - FCFS discipline where all customers have the same negative exponential service time distribution. The service rate can be state dependent, so write $\mu_j$ for the service rate when the queue length is j.
    - Processor sharing queues
    - Infinite-server queues
    - LCFS with pre-emptive resume (work is not lost)
  - In the final three cases, service time distributions must have rational Laplace transforms. This means the Laplace transform must be of the form $L(s) = N(s)/D(s)$

# Queueing networks

- Jackson networks
  - assume each queue is of infinite capacity,
  - this assumption is violated in practice

- It is the between-queue dependency that is at the origin of: blocking, spillbacks, and gridlocks (also referred to as deadlocks).

- To appropriately capture between-queue interactions and model network congestion: finite capacity queuing networks
  - Ex: $M/M/c/c$

# Finite capacity queuing networks

- Explicitly describe how congestion arises and propagates: blocking
  - ‘Blocking after service’ (BAS)
  - ‘Blocking before service’ (BBS)
  - ‘Repetitive service blocking’ (RS)
    - ‘Random destination’ (RS-RD)
    - ‘Fixed destination’ (RS-FD)

- Blocking after service:
  1. [queue]
  2. is served
  3. [blocked]
  4. departs
- Recall the container terminal example

# Finite capacity queuing networks

- Analytical expressions for the joint stationary queue length distribution is only possible for networks with:
  - 2 or 3 queues
  - and a simple topology (e.g. tandem)

- In practice for network analysis, simulation or decomposition methods are used.
- Decomposition methods:
  - decompose the network into subnetworks
  - obtain subnetwork performance measures

# Queueing networks

Overall, queueing network analysis provides:

- Detailed decomposition/description of congestion in a network
- Identify main sources of congestion
- Quantify the impact of congestion on the different network components (e.g. each link)
- Distinguish between congestion occurrence (i.e. probabilities) and its impact: rare events may have a strong impact on performance

# References

1. Tom Slater. Queueing Networks. 2000. https://homepages.inf.ed.ac.uk/jeh/Simjava/queueing/Networks/networks.html

2. Slides adapted from Carolina Osorio