Queues

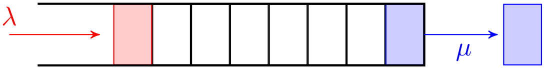
Alejandro Ribeiro & Cameron Finucane
Dept. of Electrical and Systems Engineering
University of Pennsylvania
aribeiro@seas.upenn.edu
http://www.seas.upenn.edu/users/~aribeiro/

October 28, 2015

- Queuing theory is concerned with the (boring) issue of waiting
- Waiting is boring, queuing theory not necessarily so
- "Customers" arrive to receive "service" by "servers"
- Between arrival and start of service wait in queue

- Quantities of interest (for example)
  - $\Rightarrow$ Number of customers in queue $\Rightarrow L$ (for length)
  - $\Rightarrow$ Time spent in queue $\Rightarrow W$ for (wait)

- Queues are a pervasive applications of CTMCs



Where do queues appear?

- Queues are fundamental to the analysis of (public) transportation
- Wait to enter a highway $\Rightarrow$ Customers = cars
- Subway travel times, Subway or buses? Infrequent big buses or frequent small buses?

- Packet traffic in communication networks
- Route determination, congestion management, real time requirements, resource management

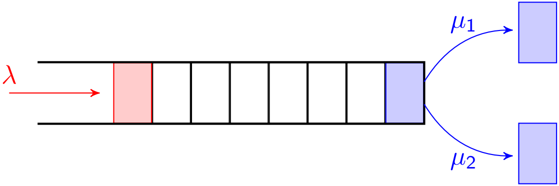
- Logistics
- Customers = raw materials, components, final products
- Customers in queue = products in storage = inactive capital

- Customer service
- How many representatives in a call center? Call center pooling

Examples of queues

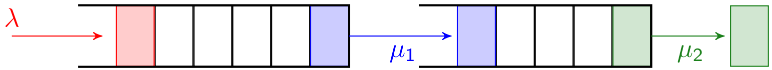
- Simplest manifestation ⇒ Single queue, single server, infinite spots
- Simpler if arrivals and services are Poisson ⇒ M/M/1 queue
- Limiting number of spots not difficult ⇒ losses appear

- Multi-server queues ⇒ Single queue, many servers
- M/M/c queue ⇒ c Poisson servers

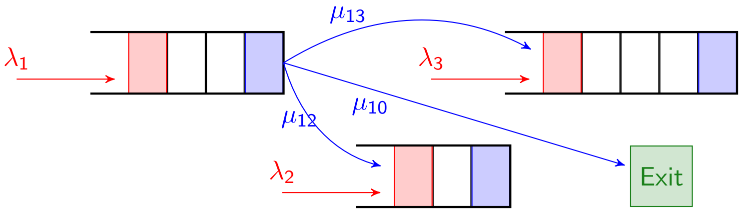


Networks of queues

- Groups of interacting queues $\Rightarrow$ Applications become interesting
- E.g. a queue tandem



- Can have arrivals at different points and random re-entries



- Batch service and arrivals, loss systems (not considered)

M/M/1 queue

Queuing theory

M/M/1 queue

Multiserver queues

Networks of queues

M/M/1 queue

- Arrival and service processes are Poisson $\Rightarrow$ Birth & death process
- Customers arrive at an average rate of λ per unit time
- Customers are serviced at an average rate of μ per unit time
- Interarrival and inter-service time are exponential and independent

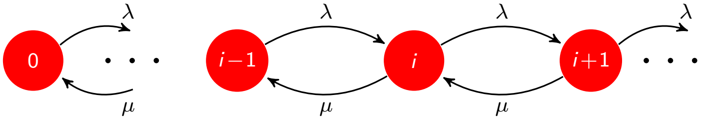
- Hypothesis of Poisson arrivals is reasonable
- Hypothesis of Poisson service times not so reasonable

- Steady state behavior (systems operating for a long time)
  $\Rightarrow$ Limit probabilities

CTMC model

- Define CTMC by identifying states with queue lengths
- Transition rates $q_{i,i+1} = \lambda$, for all i and $q_{i,i-1} = \mu$ for $i \neq 0$
- Recall that first of two exponential times is exponentially distributed

  $\Rightarrow$ Mean transition times are $\nu_i = \lambda + \mu$ for $i \neq 0$ and $\nu_0 = \lambda$



- Limit distribution equations (Rate into j = rate out of j)

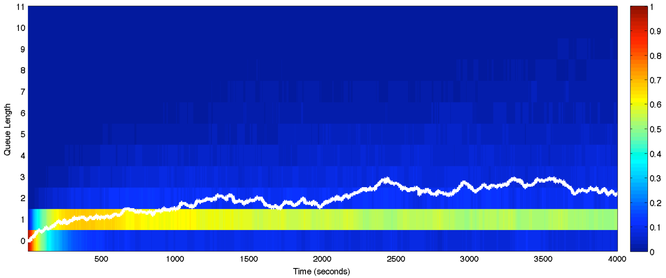
$$\lambda P_0 = \mu P_1, \qquad (\lambda + \mu) P_i = \lambda P_{i-1} + \mu P_{i+1}$$

Queue length as a function of time

- Simulation for $\lambda = 30$ customers/min, $\mu = 40$ services/min
- Probability distribution estimated by ensemble average with $M = 10^5$

$$\mathrm{P}\left[N(t) = k\right] \approx \frac{1}{M}\sum_{i=1}^{M} \mathbb{I}\left\{N_i(T) = k\right\}$$

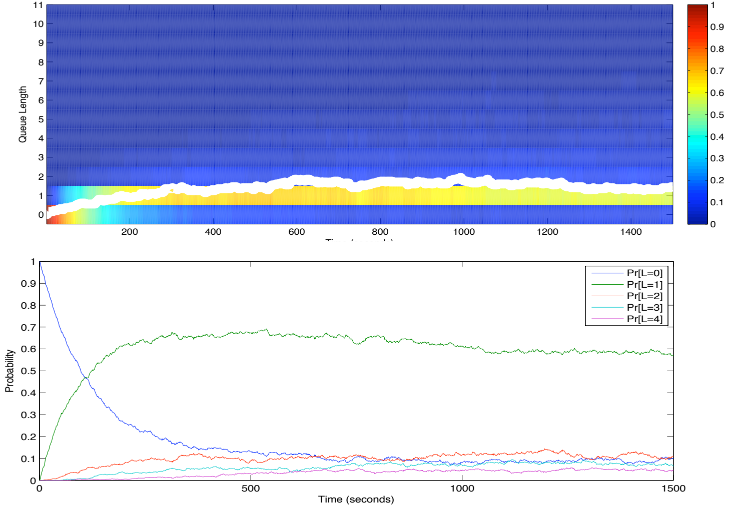
- Steady state (in a probabilistic sense) reached in around 10^3 mins.



- Queue length vs. time. Prob. color coded. Mean queue length in white

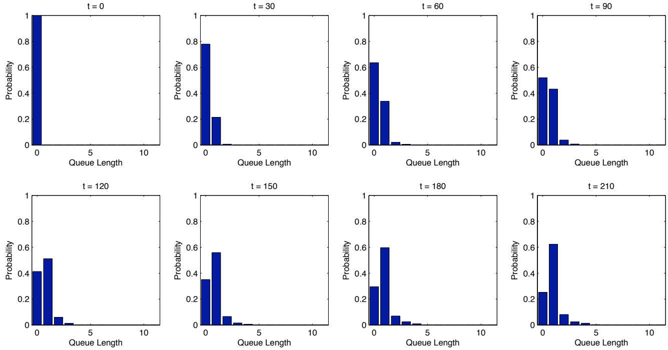
Close up on initial times

- Probabilities settle at their equilibrium values



Another view of queue length evolution

- Cross-sections of queue length probabilities at different times

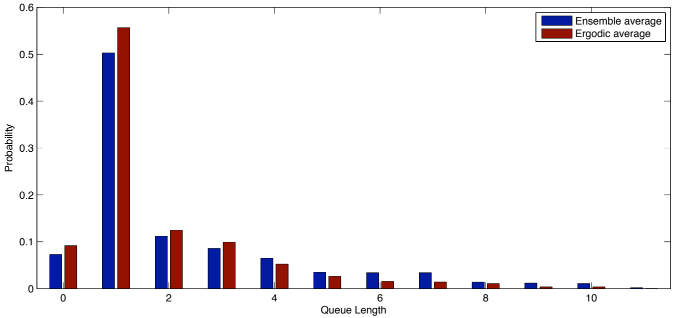


Ergodicity

- Compare ensemble averages for large t with ergodic averages

$$T_i(t) = \frac{1}{T} \int_{t=0}^{T} \mathbb{I}\{N_1(t) = k\}$$

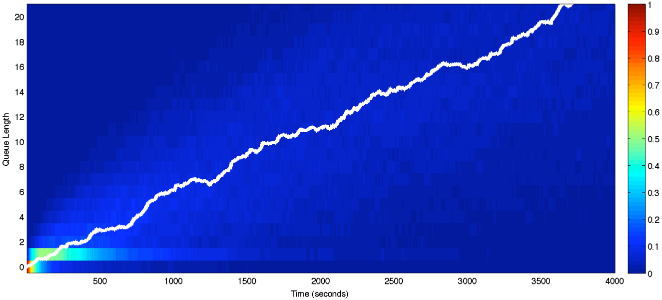
- They are approximately equal, as they should (equal as $t \to \infty$)



A non stable queue

- All former observations valid for stable queues ($\lambda < \mu$)
- Simulation for $\lambda = 60$ customers/min and $\mu = 40$, customers/min
- Queue grow unbounbded

  $\Rightarrow$ probability of small number of customers in queue vanishes



- Queue length vs. time. Prob. color coded. Mean queue length in white

Solution of limit distribution equations

- Start expressing all prob. in terms of P_0
- Repeat process done for birth and death process

- Equation for P_0 $\Rightarrow$ $\Rightarrow$ $\lambda P_0 = \mu P_1$
- Sum eqs. for P_1 and P_0 $\Rightarrow$

$$\begin{aligned} \lambda P_0 &= \mu P_1 \\ (\lambda+\mu)P_1 &= \lambda P_0 + \mu P_2 \end{aligned} \quad \Rightarrow \quad \lambda P_1 = \mu P_2$$

- Sum result and eq. for P_2 $\Rightarrow$

$$\begin{aligned} \lambda P_1 &= \mu P_2 \\ (\lambda+\mu)P_2 &= \lambda P_1 + \mu P_3 \end{aligned} \quad \Rightarrow \quad \lambda P_2 = \mu P_3$$

- Sum result and eq. for P_i $\Rightarrow$

$$\begin{aligned} \lambda P_{i-1} &= \mu P_i \\ (\lambda+\mu)P_i &= \lambda P_{i-1} + \mu P_{i+1} \end{aligned} \quad \Rightarrow \quad \lambda P_i = \mu P_{i+1}$$

- From where it follows $\Rightarrow P_{i+1} = (\lambda/\mu)P_i$ and $P_i = (\lambda/\mu)^i P_0$

Solution of limit distribution equations (continued)

- The sum of all probabilities is (geometric sum)

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

- Now use that the sum of probabilities is 1 and solve for P_0 to obtain

$$P_0 = 1-\lambda/\mu, \qquad P_i = (1-\lambda/\mu)(\lambda/\mu)^i$$

- Result valid for $\lambda/\mu < 1$, if not CTMC is transient

- Expression coincides with non concurrent queue in discrete time
- Not surprising $\Rightarrow$ continuous time $\approx$ discrete time with small Δt
- For small Δt non concurrent hypothesis is accurate

- Present derivation "much cleaner," though

Steady state expected queue length

- Expected queue length is $L = i$. To compute $\mathbb{E}[L]$ use limit probs.

$$\mathbb{E}[L] = \sum_{k=0}^{\infty} iP_i = \sum_{k=0}^{\infty} i(1 - \lambda/\mu)(\lambda/\mu)^i$$

- Latter is derivative of geometric sum $(\sum_{i=0}^{\infty} ix^i = x/(1-x)^2)$. Then

$$\mathbb{E}[L] = (1 - \lambda/\mu)\frac{\lambda/\mu}{(1 - \lambda/\mu)^2} = \frac{\lambda}{\mu - \lambda}$$

- Recall $\lambda < \mu$.
- If $\lambda \approx \mu$ queue is stable but average queue length becomes very large

Expected wait

- Packet arrives when there are L customers in queue
- Time spent in queue is time required to service these L customers
- Plus time until arriving customer is served
- Let $T_1, T_2, \ldots, T_{L+1}$ be these times. Queue wait $\Rightarrow W = \sum_{i=1}^{L+1} T_i$
- Expected value (first conditioned on $L = l$, then with respect to L)

$$\mathbb{E}[W] = \mathbb{E}\left[\sum_{i=1}^{L+1} T_i\right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{L+1} T_i \,\Big|\, L = l\right]\right]$$

- $L = l$ "not random" in inner expectation $\Rightarrow$ interchange with sum

$$\mathbb{E}[W] = \mathbb{E}\left[\sum_{i=1}^{L+1} \mathbb{E}[T_i]\right] = \mathbb{E}[(L+1)\mathbb{E}[T_i]] = \mathbb{E}[L+1]\,\mathbb{E}[T_i]$$

Expected wait (continued)

- Use expression for $\mathbb{E}[L]$ to evaluate $\mathbb{E}[L+1]$ as

$$\mathbb{E}[L+1] = \mathbb{E}[L] + 1 = \frac{\lambda}{\mu - \lambda} + 1 = \frac{\mu}{\mu - \lambda}$$

- Substitute expressions for $\mathbb{E}[L+1]$ and $\mathbb{E}[T_i] = 1/\mu$

$$\mathbb{E}[W] = (1/\mu)\frac{\mu}{\mu - \lambda} = \frac{1}{\mu - \lambda}$$

- Recall λ=arrival rate. Former may be written as

$$\mathbb{E}[W] = (1/\lambda)\frac{\lambda}{\mu - \lambda} = (1/\lambda)\mathbb{E}[L]$$

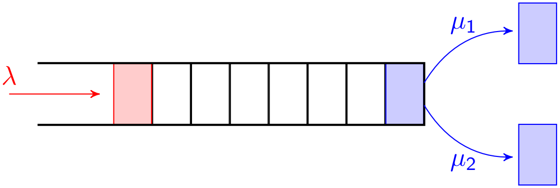
Little's law

- For M/M/1 queue have just seen $\Rightarrow \mathbb{E}[L] = \lambda \mathbb{E}[W]$
- Expression referred to as Little's law
- True even if arrivals and departures are not Poisson (not proved)

- Expected nr.customers in queue = arrival rate × expected wait

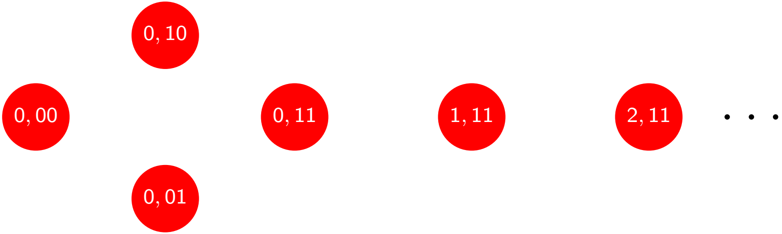
M/M/2 queue

- Arrivals are Poisson with rate λ
- Service offered by two Poisson servers with service rates μ_1 and μ_2
- When a server finishes serving a customer, it starts serving next customer in queue.
- If queue is empty the server waits for the next customer
- If both servers are waiting when customer arrives, service is performed by server 1



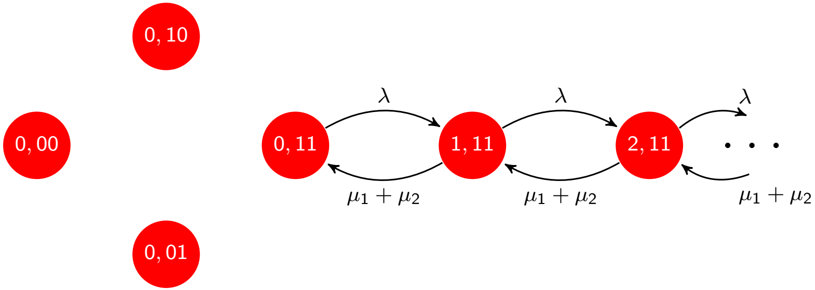
CTMC model: States

- When no customers are queued, need to distinguish servers' states
  - State $0, 00$ = no customers in queue, no customers being served
  - State $0, 10$ = no customers in queue, 1 customer served by server 1
  - State $0, 01$ = no customers in queue, 1 customer served by server 2
  - State $0, 11$ = no customers in queue, 2 customers in service
- When there are packets in queue, packets are already being served
  - State $i, 11$ = i customers in queue and 2 customers in service
  - States $i, 01$, $i, 10$ and $i, 00$ are not possible



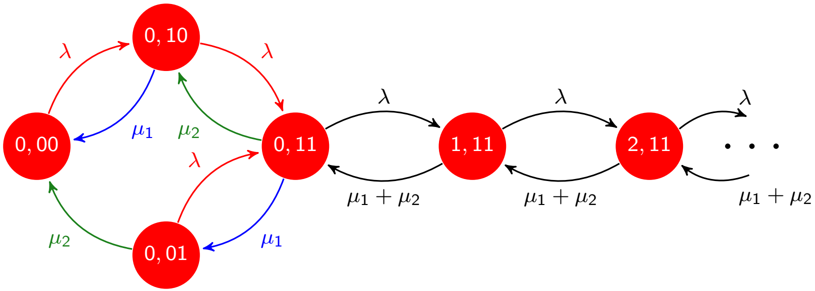
CTMC model: Transition rates

- Transition from $i, 11$ to $(i+1, 11)$ when arrival $\Rightarrow q_{i,11;(i+1),11} = \lambda$
- Transition from $i, 11$ to $(i-1, 11)$ when either server 1 or 2 finishes
- First service completion by either server 1 or 2
- Min. of two exponentials is exponential $\Rightarrow q_{i,11;(i+1),11} = \mu_1 + \mu_2$

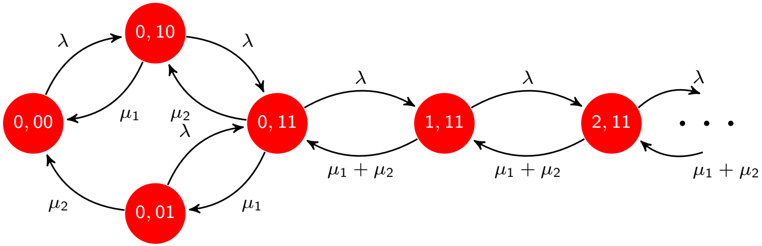


CTMC model: Transition rates (continued)

- From 0, 00 move to 0, 10 on arrival $\Rightarrow q_{0,00;0,10} = \lambda$
- From 0, 10 move to 0, 11 on arrival $\Rightarrow q_{0,10;0,11} = \lambda$
- From 0, 01 move to 0, 11 on arrival $\Rightarrow q_{0,01;0,11} = \lambda$
- From 0, 10 to 0, 00 when server 1 finishes $\Rightarrow q_{0,01;0,00} = \mu_1$
- From 0, 11 to 0, 01 when server 1 finishes $\Rightarrow q_{0,11;0,01} = \mu_1$
- From 0, 01 to 0, 00 when server 2 finishes $\Rightarrow q_{0,01;0,00} = \mu_2$
- From 0, 11 to 0, 10 when server 2 finishes $\Rightarrow q_{0,11;0,10} = \mu_2$



Limit distribution equations



- For states $i, 11$ with $i \geq 1$ eqs. are analogous to M/M/1 queue

$$(\lambda + \mu_1 + \mu_2) P_{i,11} = \lambda P_{(i-1),11} + (\mu_1 + \mu_2) P_{(i+1),11}$$

- For states $0, 11$, $0, 10$, $0, 01$ and $0, 00$ we have

$$\begin{aligned}
(\lambda + \mu_1 + \mu_2)\, P_{0,11} &= \lambda P_{0,10} \;+ \lambda P_{0,01} \;+ (\mu_1 + \mu_2) P_{1,11} \\
(\lambda + \mu_1)\, P_{0,10} &= \lambda P_{0,00} \;+ \mu_2 P_{0,11} \\
(\lambda + \mu_2)\, P_{0,01} &= \mu_1 P_{0,11} \\
\lambda\, P_{0,00} &= \mu_1 P_{0,10} + \mu_2 P_{0,01}
\end{aligned}$$

- System of linear equations $\Rightarrow$ solve numerically to find probabilities

Closing comments

- For large i behaves like M/M/1 queue with service rate $(\mu_1 + \mu_2)$
- Not identical, though, states with no queued packets are important

- M/M/c queue $\Rightarrow$ c servers with rates $\mu_1, \ldots, \mu_c$
- More cumbersome to analyze but no fundamental differences

Networks of queues

Queuing theory

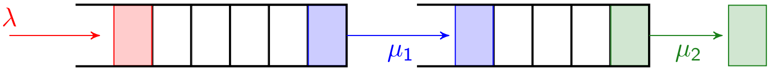
M/M/1 queue

Multiserver queues

Networks of queues

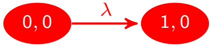
A queue tandem

- Customers arrive at system to receive two services
- They arrive at a rate λ and wait in queue 1 for service 1
- Service 1 is performed at a rate μ_1
- After completions of service 1 customers move to queue 2
- Service 2 is performed at a rate μ_2



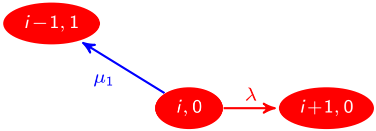
- States (i, j) represents i customers in queue 1 and j in queue 2
- If both queues are empty $i = 0$ $j = 0$ only possible event is an arrival

$$q_{00,10} = \lambda$$



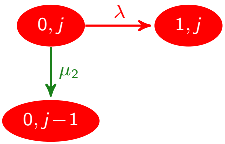
- If queue 2 is empty might have arrival or completion of service 1

$$q_{i0,(i+1)0} = \lambda$$
$$q_{i0,(i-1)1} = \mu_1$$



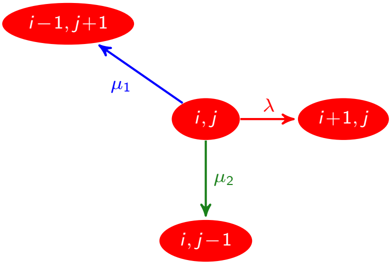
- If queue 1 is empty might have arrival or completion of service 2

$$q_{0j,1j} = \lambda$$
$$q_{0j,0(j-1)} = \mu_2$$



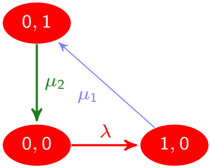
- If no queue is empty arrival, service 1 and service 2 possible

$$q_{ij,(i+1)j} = \lambda$$
$$q_{ij,(i-1)(j+1)} = \mu_1$$
$$q_{ij,i(j-1)} = \mu_2$$



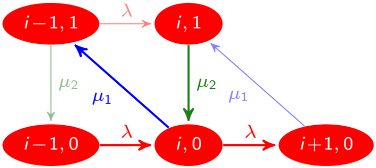
- Rate at which CTMC enters state (i, j) = rate at which CTMC leaves (i, j)
- **State $(\mathbf{0}, \mathbf{0})$**
- From $(0, 0)$ can go to $(1, 0)$
- Can enter $(0, 0)$ from $(0, 1)$

$$\lambda P_{00} = \mu_2 P_{01}$$



- **State $(\mathbf{i}, \mathbf{0})$**
- From $(i, 0)$ go to $(i+1, 0)$ or $(i-1, 0)$
- Into $(i, 0)$ from $(i-1, 0)$ or $(i, 1)$

$$(\lambda + \mu_1) P_{i0} = \lambda P_{(i-1)0} + \mu_2 P_{i1}$$

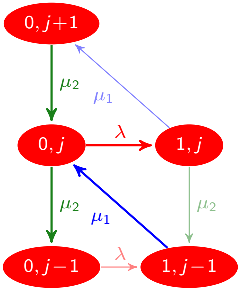


Balance equations (continued)

- **State (0, j)**

- From $(0, j)$ go to $(1, j)$ or $(0, j-1)$
- Into $(0, j)$ from $(1, j-1)$ or $(0, j+1)$

$$(\lambda + \mu_2)P_{0j} = \mu_1 P_{1(j-1)} + \mu_2 P_{0(j+1)}$$

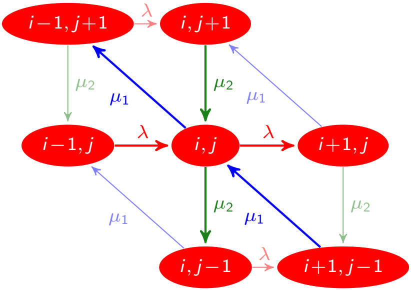


Balance equations (continued)

- **State** $(\mathbf{i}, \mathbf{j})$

- From (i, j) can go to $(i+1, j)$, $(i-1, j+1)$ or $(i, j-1)$
- Can enter (i, j) from $(i-1, j)$, $(i+1, j-1)$ or $(i, j+1)$

$$(\lambda + \mu_1 + \mu_2)P_{ij} = \lambda P_{(i-1)j} + \mu_1 P_{(i+1)(j-1)} + \mu_2 P_{i(j+1)}$$



Solution of balance equations

- Direct substitution shows that balance equations are solved by

$$P_{ij} = \left(1 - \frac{\lambda}{\mu_1}\right)\left(\frac{\lambda}{\mu_1}\right)^i \left(1 - \frac{\lambda}{\mu_2}\right)\left(\frac{\lambda}{\mu_2}\right)^j$$

- Compare with expression for M/M/1 queue
  - $\Rightarrow$ It behaves as two two independent M/M/1 queues
  - $\Rightarrow$ First queue has rates λ and μ_1
  - $\Rightarrow$ Second queue has rates λ and μ_2

- Result can be generalized to networks of queues
- Result important in transportation networks
- Also useful to analyze Internet traffic