Queuing Theory

Gonzalo Mateos
Dept. of ECE and Goergen Institute for Data Science
University of Rochester
gmateosb@ece.rochester.edu
http://www.ece.rochester.edu/~gmateosb/

November 16, 2016
Queueing theory

M/M/1 queue

Multiserver queues

Networks of queues
Queues

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

- Queues are a pervasive application of CTMCs
Where do queues appear?

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

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

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

- Customer service
  - Q: How many representatives in a call center? Call center pooling
Examples of queues

- **Simplest rendition** ⇒ 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 (i.e., exp. service times)
Networks of queues

- Groups of interacting queues ⇒ Applications become interesting

Ex: 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
Arrival and service processes are Poisson \( \Rightarrow \) Birth & death process

- Customers arrive at an average rate of \( \lambda \) per unit time
- Customers are serviced at an average rate of \( \mu \) per unit time
- Interarrival and inter-service time are exponential and independent

Hypothesis of Poisson arrivals is reasonable

Hypothesis of exponential service times not so reasonable \( \Rightarrow \) Simplifies the analysis. Otherwise, study a M/G/1 queue

Steady-state behavior (systems operating for a long time) \( \Rightarrow \) Q: Limit probabilities for the M/M/1 system?
Define CTMC by identifying states $Q(t)$ 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

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

Limit distribution equations (Rate out of $j =$ Rate into $j$)

$$\lambda P_0 = \mu P_1, \quad (\lambda + \mu)P_i = \lambda P_{i-1} + \mu P_{i+1}$$
Queue length as a function of time

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

$$P (Q(t) = k) \approx \frac{1}{M} \sum_{i=1}^{M} \mathbb{I} \{Q_i(t) = k\}$$

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

- Queue length vs. time. Probabilities are color coded
  ⇒ Mean queue length shown 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

$$T_i(t) = \frac{1}{t} \int_0^t \mathbb{I}\{Q(\tau) = i\} d\tau$$

- 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 length grows unbounded
  - Probability of small number of customers in queue vanishes
  - Actually CTMC transient, $P_i \to 0$ for all $i$

- Queue length vs. time. Probabilities are color coded
  - Mean queue length shown in white
Solution of limit distribution equations

- Start expressing all prob. in terms of $P_0$. Define traffic intensity $\rho := \frac{\lambda}{\mu}$
- Repeat process done for birth and death process

- Equation for $P_0$ \( \Rightarrow \) \( \lambda P_0 = \mu P_1 \) \( \Rightarrow \lambda P_0 = \mu P_1 \)

- Sum eqs. for $P_1$ and $P_0$ \( \Rightarrow \) \( (\lambda + \mu)P_1 = \lambda P_0 + \mu P_2 \) \( \Rightarrow \lambda P_1 = \mu P_2 \)

- Sum result and eq. for $P_2$ \( \Rightarrow \) \( \lambda P_1 = \mu P_2 \) \( (\lambda + \mu)P_2 = \lambda P_1 + \mu P_3 \) \( \Rightarrow \lambda P_2 = \mu P_3 \)

- Sum result and eq. for $P_i$ \( \Rightarrow \) \( \lambda P_{i-1} = \mu P_i \) \( (\lambda + \mu)P_i = \lambda P_{i-1} + \mu P_{i+1} \) \( \Rightarrow \lambda P_i = \mu P_{i+1} \)

- From where it follows \( \Rightarrow \) $P_{i+1} = (\lambda/\mu)P_i = \rho P_i$ and recursively $P_i = \rho^i P_0$
Solution of limit distribution equations (continued)

- The sum of all probabilities is 1 (use geometric series formula)

\[
1 = \sum_{i=0}^{\infty} P_i = \sum_{i=0}^{\infty} \rho^i P_0 = \frac{P_0}{1 - \rho}
\]

- Solve for \( P_0 \) to obtain

\[
P_0 = 1 - \rho, \quad \Rightarrow \quad P_i = (1 - \rho)\rho^i
\]

\( \Rightarrow \) Valid for \( \lambda/\mu < 1 \), if not CTMC is transient (queue unstable)

- Expression coincides with non-concurrent queue in discrete time

\( \Rightarrow \) Not surprising. Continuous time \( \approx \) discrete time with small \( \Delta t \)

\( \Rightarrow \) For small \( \Delta t \) non-concurrent hypothesis is accurate

- Present derivation “much cleaner,” though
To compute expected queue length $E[L]$ use limit probabilities

$$E[L] = \sum_{i=0}^{\infty} iP_i = \sum_{i=0}^{\infty} i(1 - \rho)\rho^i$$

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

$$E[L] = (1 - \rho) \times \frac{\rho}{(1 - \rho)^2} = \frac{\rho}{1 - \rho}$$

Recall $\lambda < \mu$ or equivalently $\rho < 1$ for queue stability

$\Rightarrow$ If $\lambda \approx \mu$ queue is stable but $E[L]$ becomes very large
Steady-state expected wait

- Customer arrives, \( L \) in queue already. Q: Time spent in queue?
  - 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 \( W = \sum_{i=1}^{L+1} T_i \)

- Expected value (condition on \( L = \ell \), then expectation w.r.t. \( L \))

\[
\mathbb{E}[W] = \mathbb{E}\left[ \sum_{i=1}^{L+1} T_i \right] = \mathbb{E}\left[ \mathbb{E}\left[ \sum_{i=1}^{\ell+1} T_i \mid L = \ell \right] \right]
\]

- \( L = \ell \) “not random” in inner expectation \( \Rightarrow \) interchange with sum

\[
\mathbb{E}[W] = \mathbb{E}\left[ \sum_{i=1}^{L+1} \mathbb{E}[T_i] \right] = \mathbb{E}\left[ (L + 1)\mathbb{E}[T_i] \right] = \mathbb{E}[L + 1] \mathbb{E}[T_i]
\]
Use expression for $\mathbb{E}[L]$ to evaluate $\mathbb{E}[L + 1]$ as

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

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

$$\mathbb{E}[W] = \frac{1}{\mu} \times \frac{1}{1 - \rho} = \frac{1}{\mu - \lambda}$$

Recall $\lambda =$ arrival rate. Former may be written as

$$\mathbb{E}[W] = \frac{1}{\lambda} \times \frac{\rho}{1 - \rho} = \frac{1}{\lambda} \mathbb{E}[L]$$
Little’s law

- For M/M/1 queue have just seen \( \Rightarrow E[L] = \lambda E[W] \)
  \( \Rightarrow \) 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 \( \times \) expected wait
Multiserver queues

Queuing theory

M/M/1 queue

Multiserver queues

Networks of queues
M/M/2 queue

- Service offered by two Poisson servers with service rates $\mu_1$ and $\mu_2$
  - Arrivals are Poisson with rate $\lambda$ as in the M/M/1 queue

- When a server finishes serving a customer, serves next one in queue
  - If queue is empty the server waits for the next customer

- If both servers are idle when a new customer arrives
  - Service is performed by server 1 (simply by convention)
CTMC model: States

- When no customers are in line, 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 customers in line, both servers are busy
  - State $i,11 =$ $i > 0$ customers in queue and 2 customers in service
  - States $i,01$, $i,10$ and $i,00$ are not possible for $i > 0$
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 $\Rightarrow$ 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 \( i > 0 \), 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

\[
(\lambda + \mu_1 + \mu_2)P_{0,11} = \lambda P_{0,10} + \lambda P_{0,01} + (\mu_1 + \mu_2)P_{1,11}
\]

\[
(\lambda + \mu_1)P_{0,10} = \lambda P_{0,00} + \mu_2 P_{0,11}
\]

\[
(\lambda + \mu_2)P_{0,01} = \mu_1 P_{0,11}
\]

\[
\lambda P_{0,00} = \mu_1 P_{0,10} + \mu_2 P_{0,01}
\]

System of linear equations ⇒ Solve numerically to find probabilities
For large $i$ behaves like M/M/1 queue with service rate $(\mu_1 + \mu_2)$

⇒ Still, states with no queued packets are important

M/M/c queue ⇒ $c$ 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 $\lambda$ and wait in queue 1 for service 1
  $\Rightarrow$ Service 1 is performed at a rate $\mu_1$
- After completions of service 1 customers move to queue 2
  $\Rightarrow$ Service 2 is performed at a rate $\mu_2$
States \((i, j)\) represent \(i\) customers in queue 1 and \(j\) in queue 2

- If both queues are empty \((i = j = 0)\), only possible event is an arrival
  
  \[ q_{00,10} = \lambda \]

- If queue 2 is empty might have arrival or completion of service 1
  
  \[ q_{i0,0} = \lambda, \quad q_{i0,(i-1)1} = \mu_1 \]
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

\[ q_{ij,(i+1)j} = \lambda \]
\[ q_{ij,(i-1)(j+1)} = \mu_1 \]
\[ q_{ij,i(j-1)} = \mu_2 \]
Balance equations

- Rate at which CTMC enters state \((i, j)\) = rate at which CTMC leaves \((i, j)\)

- **State** \((0, 0)\) - Both queues empty
  - From \((0, 0)\) can go to \((1, 0)\)
  - Can enter \((0, 0)\) from \((0, 1)\)

\[
\lambda P_{00} = \mu_2 P_{01}
\]

- **State** \((i, 0)\) - Queue 2 empty
  - From \((i, 0)\) go to \((i + 1, 0)\) or \((i - 1, 1)\)
  - Into \((i, 0)\) from \((i - 1, 0)\) or \((i, 1)\)

\[
(\lambda + \mu_1)P_{i0} = \lambda P_{(i-1)0} + \mu_2 P_{i1}
\]
State $(0, j)$ - Queue 1 empty

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

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

\[(\lambda + \mu_2)P_{0j} = \mu_1P_{1(j-1)} + \mu_2P_{0(j+1)}\]
Balance equations (continued)

- **State** \((i, j)\) - Neither queue empty
- 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
⇒ It behaves as two independent M/M/1 queues
⇒ First queue has rates \( \lambda \) and \( \mu_1 \)
⇒ Second queue has rates \( \lambda \) and \( \mu_2 \)

Result can be generalized to networks of queues
⇒ Important in transportation networks
⇒ Also useful to analyze Internet traffic
Glossary

- Queuing theory
- Customers and servers
- Queue length
- Time spent in queue
- M/M/1 queue
- Finite-capacity queue
- Multi-server queue
- Network of queues
- Queue tandem
- Poisson arrivals

- Exponential service times
- Balance equations
- Stable queue
- Traffic intensity
- Expected queue length
- Expected waiting time
- Little’s law
- M/M/c queue
- Aggregate service rate
- Independent M/M/1 queues