Queue Length Modeling and Analysis of M/G/1 Queues with Heavy-Tailed Interarrival Times

Edgar Derricho

Georgia Gwinnett College

4/25/2024

What is Queuing Theory?

- Queuing Theory is the analysis of waiting lines. It answers questions such as:
  - How long will a customer have to wait in a line?
  - When do customers arrive in a queue?
  - What rules will the queue use to serve the customers?
- Examples of processes that utilize queuing theory:
  - Voice or data traffic in communication systems
  - Vehicles requiring service in a garage
  - The boarding line in an airport
- Components of a Queue:
  - Customer - unit demanding a service
  - Server - unit providing a service
  - Note: The terms **customer** and **server** are used in a generic sense.

Kendall's Notation and Queue Discipline

- The standard model for describing a queuing system is Kendall's notation first coined by D.G. Kendall in 1953
  - The formula for Kendall's notation is: input/service/number of servers
  - $G/G/1$: General interarrival distribution/General service distribution/1 server
  - $M/G/1$: Markovian (Poisson or exponential) inerarrival distribution/General service distribution/1 server
- Queue Dicipline
  - Queue discipline is the manner in which a queue provides a service. Some examples of queue disciplines are:
  - FCFS - First Come First Served
  - LCFS - Last Come First Served
  - PS - Processor Sharing
  - RS - Random Selection of Service
  - Therefore, M/G/1 FCFS is a queue with markovian interarrival times, general service distribution, 1 server and is first come first served.

Common Queue Metrics and Little's Law

- Common Queue Metrics
  - Queue Size at given time t
  - Sojourn Time (the entire time a customer is in the queue)
  - Busy Period - when customers are either in the queue or being served
  - Idle Time - when no customers are in the queue
- Little's Law
  - D.C. Little gave the first formal proof relating the long term mean queue length and the mean amount of time a customer is in the queue.
  - The formula for Little's Law is:

$$L = \lambda W$$

  - Where L is the queue length, λ is the arrival rate and W is the waiting time.

Introduction to Stochastic Processes and the Markov Property

Definition

A stochastic process is a collection of random variables indexed by time.

Definition

A process has the Markovian property if
$P[X_{t+1} = j | X_0 = k_0, X_1 = k_1, ..., X_{t-1}, X_t] = i = P[X_{t+1} = j | X_t = i]$

- i.e. This is a process where the future is independent of the past given the present

The $M/G/1$ Queue

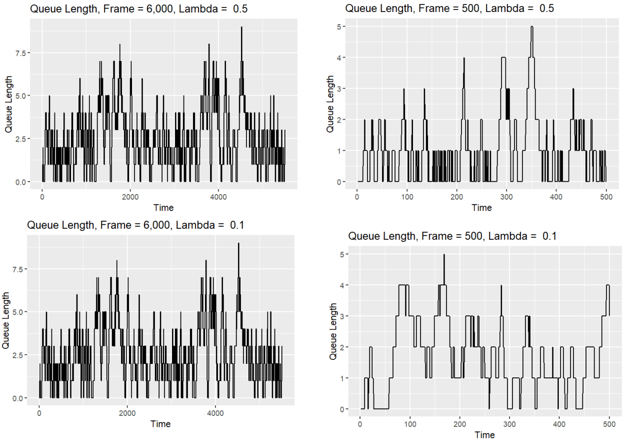
- Conversion into Markovian Queue
  - The $M/G/1$ queue is a queue with exponential arrival times and service times with independent and identically distributed random variables with an unspecified distribution
    Note: A true markovian queue must have the form M/M/n where both the arrival times and the service times are Poisson or exponential.
  - However, Kendall developed a procedure to convert the queue length process of a $M/G/1$ queue into Markov chains making analysis substantially easier.
- Conversion Process
  - Let $[Q(t), R(t)]$ be a vector where $Q(t)$ is the number of customers in the queue at time t and $R(t)$ is the remaining service time for any given customer at time t.
  - By considering the queue length only during times of departure or arrival (i.e. when $R(t) = 0$), we can reduce this vector to $[Q(t)]$ which is markovian

Burstiness Property

- The autocorrelation function is the correlation of a signal with a delayed copy of itself as a function of delay.
- When autocorrelations of random variables do not decay exponentially over time, the phenomena is called 'Burstiness'
  In other words, when observing a random variable, if the same sporadic behavior occurs through different time scales, the resulting variable is bursty. This is very common in network traffic.
- The burstiness property is closely related to long-range depenednece and can help us visualize the phenomena.

Burstiness Examples and Long-Range Dependence

Burstiness occurs in the bottom two figures where $\lambda = 0.1$.



Exponential Distribution with Heavy-Tailed Characteristics

Heavy-tailed distributions are characterized by the phenomena of long range dependence and self similarity.

Definition

Let $c(t) = Cov\{X(s), X(s+t)/VarX(s)\}$ be the autocorrelation function for a time series, χ is short range dependent if $\int_0^\infty |c(t)|dt < \infty$. If $\int_0^\infty |c(t)|dt = \infty$, then χ is long-range dependent.

Definition

A Stochastic process $\chi = X(t), t \geq 0$ is (strictly) self similar with parameter H if $X(t), t \geq 0$ and $\gamma^{-H} X(\gamma t), t \geq 0$ have the same finite-dimensional distributions for any $\gamma \geq 0$.

Checking Queue for Long Range Dependence

- The M/G/1 queue shows signs of both long range dependence and self similarity in its queue length.
- Long Range Dependence
  The method for checking for long range dependence is by analyzing the autocorrelation function. Observe the following definition:

Definition

X is long-range dependent if the autocorrelation function $c(.)$ shows a particular type of power law behavior. In other words: $c(t)\ c_0 t^{-\alpha}$

- Fortunately, we can observe this power law behavior by graphing the autocorrelation function for different values of our parameter λ

Checking Queue for Self-Similarity

- An approximation of this quality can be done through the analysis of the mean and variance at various windows of our queue length. However, there is also an alternative formula:

  $$y(t) \equiv^d a^\alpha y(\frac{t}{a})$$

  Note: $y(t)$ is our instantaneous queue length.
  For self similarity, both sides of this equation must have the same probability distribution.
- Conducting a Kolmogov-Smirnov(KS) test at various time scales shows us that the two distributions indeed stem from the same distribution for $a = 2$ and $\alpha = -0.053$.

Average Queue Length Analysis

- We will analyze the $M/G/1$ queue by the embedded Markov chain method
- Let customers arrive in a Poisson process with parameter λ and are served by a single server. Let the service times of these customers be independent, identically distributed random variables $S_n, n = 1, 2, 3, ...$ with $P(S_n \geq x) = B(x)$

$$k_i = P(X_n = j) = \int_0^\infty e^{-\lambda t}(\frac{\lambda t^j}{j!})dB(t)$$

- We will define the Laplace-Stieltjes transform of the service time distribution

$$\psi(\theta) = \int_0^\infty e^{-\lambda t} dB(t), Re(\theta) \geq 0$$

- One property of the Laplace-Stieltjes transforms and PGF's is

$$E(X_n) = K'(1)$$

Average Queue Length Analysis

- The probability generation function of the customers arriving during a service time $K(z)$:

$$K(z) = \sum_{j=0}^{\infty} k_j z^j, |z| \geq 0$$

- $\lambda b =$ (arrival rate)*(mean service time)
  We will call this number ρ Note: The Chapman-Kolmogrov relations connect joint probability distributions of different sets of coordinates on a stochastic process.

Average Queue Length Analysis - Limiting Distribution

- One property of aperiodic positive recurrent irreducible Markov chains is the results of $\lim_{n\to\infty} P^n$ becoming a Markov chain with identical rows. Using the Chapman-Kolmogrov relations we may write:

$$\pi P = \pi = \pi P^n$$

- Let $Q(t)$ be the queue length. We want to know $\lim_{n\to\infty} Q_n$. Define

$$\pi_j = \sum_{j=0}^{\infty} \pi_i P_{ij}, j = 0, 1, 2... \text{ and } K(z) = \sum_{j=0}^{\infty} \pi_j z^j, |z| \leq 1$$

- Therefore

$$\pi(z) = \pi_0 K(z) + \pi_1 K(z) + \pi_2 z K(z) + ...$$

- After rearranging the terms and using the normalizing condition $\sum_0^\infty \pi_j = 1$, we have the following probability generating function (PGF)

$$\pi(z) = \frac{(1-\rho)(z-1)K(z)}{z - K(z)} \text{ and } L = E(Q(t)) = \pi'(1)$$

Simulation of $M/G/1FCFS$ Queue

- Method: Using the R programming language, I developed a M/G/1 FCFS queue where $\lambda = 0.1$
  Note: Heavy-tailed characteristics of exponential distributions occur at $\lambda < 0.1$
- The results show that the model does not work for heavy-tailed interarrival times!

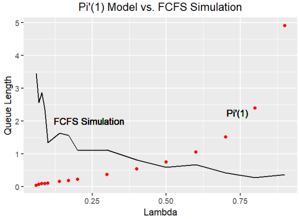


Figure: Pi'(1) Model vs FCFS Simulation

Simulation of $M/G/1PS$ Queue

- Question: Is there a scalar or a function of lambda that can "normalize" the model to account for heavy-tailed behavior in the queue?
- Proposal: Using Little's law, there is a scalar or function that will satisfy the following:

$$\pi'(1) + s(\lambda) = \lambda W = L$$

$$s(\lambda) = \lambda W - \pi'(1)$$

- Using the queue simulation for L and a nonlinear regression model, we have a proposed $s(\lambda)$:

$$s(\lambda) = 4.4005\lambda^2 - 0.5557\lambda + 1.8639$$

Frame Title

- This gives us the following results:

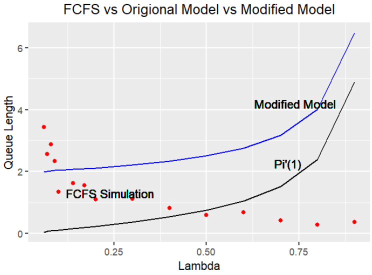


Figure: Pi'(1) vs FCFS Simulation vs Modified Model

- Now, we will use a technique similar to a logarithmic transformation by multiplying $s(\lambda)$ by $e^{-\lambda}$
  Now our new function is

$$s^*(\lambda) = e^{-\lambda}s(\lambda) = e^{-\lambda}(4.4005\lambda^2 - 0.5557\lambda + 1.8639)$$

:

Simulation of $M/G/1PS$ Queue

- Using $s^*(\lambda)$ gives us the following results:

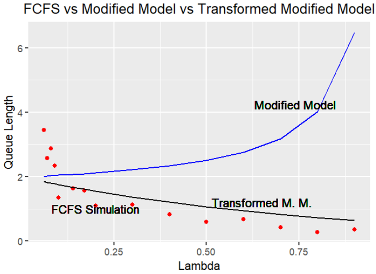


Figure: FCFS Simulation vs Modified Model vs Modified Model

- This new model fits well for exponential distributions that exhibit heavy-tailed behavior.

Conclusions

- Heavy-tailed behaviors can be exhibited in exponential distributions with parameter λ when $\lambda < 0.1$
- When interarrival times are heavy-tailed, a modification to the queue length model needs to be made.

Thank you!

Dr. Curry

References

1 Zwart, A. P. (2001). Queueing systems with heavy tails. [Phd Thesis 1 (Research TU/e / Graduation TU/e), Mathematics and Computer Science]. Technische Universiteit Eindhoven. https://doi.org/10.6100/IR547196

2 Bhat, U. N. (2015). An introduction to queueing theory: Modeling and analysis in applications. Birkhauser.

3 Montgomery, D. C., Jennings, C. L., Kulahci, M. (2008). Introduction to time series analysis and forecasting. Wiley-Interscience.