An Introduction to Queuing Theory and Common Queuing Models

Dawson Berry
ELG5119/CSI5138B/EACJ5109O
Prof. Yongyi Mao

November 2025

Contents

1 Queuing Theory and Models

1.1 Queuing as a Random Process

Queuing theory is a well-studied area with applications in economics and engineering. Section 1 introduces queuing as a random process, explores fundamental concepts of queuing and derives common queuing models. Section 2 examines a particular trend in how queuing models are used and applied in wireless and wired communications. A specific queue model, $M/M/1$ with exponentially distributed inter-arrival and service times is commonly used in research literature. This model can be inaccurate and may fail to deliver appropriate analysis; Section 2 examines precisely this scenario in a very particular context. Within this blog the context of computer networks will be used to frame queuing. In queuing literature, customer is often used as an element in the system being serviced by a server. We will use packet in place of customer. For those unfamiliar with the concept a packet is a discrete atomic unit of data that must be served and queued in whole instead of a fraction.

Queuing theory follows quite naturally from many fundamental concepts in random processes. A queuing process is a random process $\{N(t) : t \geq 0\}$ on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$ indexed by time. All queuing models are constructed from the atomic unit of the single-server queue where we have:

- $A(t)$ an arrival counting process
- $D(t)$ a departure process
- $N(t) \in \mathbb{Z}_{\geq 0}$ a number-in-system process

$A(t)$ is a piecewise constant function. For each event or outcome ω in the probability space the sample path $t \to A(t, \omega)$ is the number of arrivals that have occurred in $[0, t]$. This function is constant since at some time $t \in [T_k, T_{k+1}]$ if $A(t) = k$ and an arrival happens in $t \in (T_{k+1}, T_{k+2}]$ then $A(t) = k + 1$. We talk about $A(t)$ as a function of time for a fixed random outcome (a sample path), not the expectation. Consider the example of a Poisson process $\mathbb{E}[A(t)] = \lambda t$, which is a linear function but each realized path of $A(t)$ is a stair function. The exact same logic applies to both $D(t)$, and $N(t)$ being integer-valued right-continuous piecewise constant step functions that change when a discrete event such as an arrival or departure happens. $D(t)$ is the number of service completions by time t. $N(t)$ is the number of packets in the queue plus the packets in service at time t. Intuitively, $N(t)$can be thought of as the state of the queue, since it counts how many packets are in the system. However, for many arrival and service distributions the process $\{N(t)\}$ by itself is not Markov. To get the Markov property we would need to augment the state with additional information, such as the residual service time of the job in service. In the special $M/M/1$ case, the exponential memoryless inter-arrival and service times ensure that the number-in-system process $\{N(t\}$ is a continuous-time Markov chain, so $N(t)$ alone is a sufficient state description. The relation between the number

of packets in the system and the arrival and departure processes at time t can be found by the path-wise relation of the conservation equation:

$$N(t) = N(0) + A(t) - D(t), \qquad t \geq 0$$

This is a sort of bookkeeping identity that holds for every realized trajectory and not just the average, rewritten in the same way we talked about $A(t)$ previously: $\forall \omega, \forall t \geq 0 : N(t, \omega) = N(0, \omega) + A(t, \omega) - D(t, \omega)$. Once again this is not an average or steady state but is a particular realization of the queue. This equation expresses conservation since packets are not created out of thin-air, nor do they disappear mysteriously the only way for the system size to change is via arrivals and departures. since departures only occur when $N(t) > 0$ as a queue cannot be serviced if there is nothing in the system, and arrivals cannot happen before time starts a natural boundary condition $\forall t \geq 0$ arises.

Let $\{S_n\}_{n \geq 1}$ be the i.i.d service times of jobs, and $\{T_n\}_{n \geq 1}$ be the i.i.d inter-arrival times. The waiting-time sequence $\{W_n\}$ of the n-th arrival under FIFO obeys Lindley's recursion

$$W_{n+1} = max\{0, W_n + S_n - T_{n+1}\}$$

This equation simply means the waiting time of the $(n+1)$-th packet is either 0 if the system was previously not busy, or a sum of how long it took for the n-th packet to depart minus how long it took for the $(n+1)$-th packet to arrive. The max(0, ...) argument also keeps the wait non-negative, as scenarios where packets exit a queue before they arrive is strictly not allowed. This recursion and the point processes $A(\cdot), D(\cdot)$ fully determine $N(\cdot)$. Queuing theory is interested in analyzing what happens to the number of packets in the system as the arrival and departure processes change. The arrival and departure processes are not required to remain the same throughout runtime, but in all common literature it is assumed whatever underlying process $A(\cdot)$ and $D(\cdot)$ take does not change at runtime. So for some arrival and departure process from the queue we are interested in what happens to the queue under different combinations of arrival and departure processes.

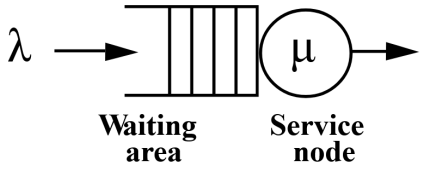

Figure 1: single-server Queue with arrival rate λ and service rate μ

1.2 Properties of Tractable Models

To construct our queuing model, first we should take a look at some related and useful concepts in probability theory to build some intuition for what is going on under the hood. The Markov property is satisfied by a random process when the future state depends only on the present state but not past states. Consider a state of a queuing system where there are k packets in the queue with an arrival process $A(\cdot)$ and departure process $D(\cdot)$. It does not matter how the system ended up in this state, but the next state the system takes can only depend on the interaction between the current state and the arrival and departure processes. This is an incredibly helpful property when analyzing queuing behavior since the future evolution of the queue can be predicted with only a single snapshot of the system. The entire history or previous states are not required. Formally, a Markov process has the Markov property if $X(t)$, the system at time t, follows:

$$\mathbb{P}(X(t+h)|X(s), s \leq t) = \mathbb{P}(X(t+h)|X(t))$$

This trick to simplify queuing behavior comes at a cost to generality. One can imagine only some arrival and departure processes support this property. First, we should explore the memoryless property, where the memoryless axiom implies a Markov Process with the Markov property if every RV of the random process is memoryless. A nonnegative RV X is memoryless if $P(X > s+t|X > s) = P(X > t) \ \ \forall s, t \geq 0$. This reads the probability of having to wait t additional time given that you have waited s is the same as if you had just started waiting. The future remaining time of the system is independent of the elapsed time given the current state. Recall that for a Markov process, given the current state X_t, the distribution of all future states $\{X_{t+h} : h \geq 0\}$ does not depend on the past. The memoryless axiom is a one-step single variable version, that is no memory in a single waiting time, of the much larger Markov property, no memory of the entire time-evolving state process. This has only one valid continuous solution for $P(X > t)$ but we should seek to build some intuition about what is happening first. Effectively, any $D(\cdot)$ that depends on the previous departure or service time for the last departure breaks the Markov

property. In a similar manner, any $A(\cdot)$ that depends on a previous arrival also breaks this property. Without any derivations or explanation done so far, consider the most intuitive example a deterministic arrival time $\frac{1}{\lambda}$ as well as a deterministic service time $\frac{1}{\mu}$. If at any state $N(t)$ at time t we know there is 1 packet in the system this means there is one packet being served, but there is no indication of how far along the service is. The packet's time under service is totally unknown and the packet in the system could be anywhere from 0 time in the system or almost exactly $\frac{1}{\mu}$ time in system. If the service is almost finished then a departure will happen soon, and if the service just started there will be a wait until the next departure. Now the system does not only depend on $N(t)$ but also how long ago the last arrival occurred where this is called time-phase memory. This system is neither memoryless nor Markovian. Although a fully deterministic system is easy to analyze mathematically, it can be inappropriate for modeling real scenarios. The only continuous solution that satisfies the two properties above is exponential.

$$\mathbb{P}(X > t) = e^{-\theta t}, \quad \theta > 0$$

There are many other ways to prove this with one of the simplest and intuitive proofs given as:

The conditional property gives the survival function of X: $\bar{F}(t) = P(X > t)$, which satisfies $\bar{F}(s+t) = \bar{F}(s)\bar{F}(t)$ for a memoryless distribution. Continuity at 0 with $\bar{F}(0) = 1$ implies Cauchy's exponential equation. Hence $\bar{F}(t) = e^{-\theta t}$ is the only continuous solution. Recall that $N(\cdot)$ is fully specified by two processes an arrival and departure process. If both the arrival and departure processes, specifically the inter-arrival times and departure times are exponential, then the state $N(t)$ is a continuous-time Markov chain with generator:

$$Q_{n,n+1} = \lambda, \qquad Q_{n,n-1} = \mu, (n \geq 1), \qquad Q_{n,n} = -(\lambda + 1_{n \geq 1\mu})$$

Here Q is an nxn matrix and this generator gives the transitions for moving states or staying at the same state. λ is the mean arrival rate to the system and μ is the mean service rate of the system. This is standard notation in queuing theory and will be used throughout the rest of this work. The Markov property is the main source of tractability, and no other continuous service-time models enjoy this property.

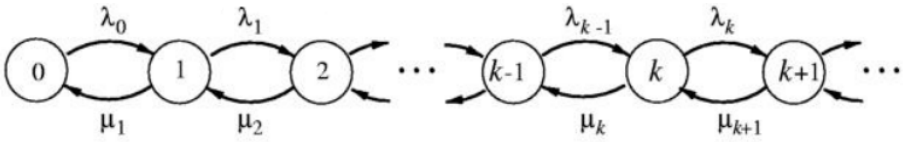

Figure 2: CTMC - Birth-Death Process

1.3 Building a Queuing Model

Now we can construct our queuing model and look at key results in analyzing it. We will use Kendall's notation: $A/S/c/K/N/D$ fully specifies the queuing model:

- A arrival process distribution.
- S service time distribution.
- c number of servers serving the queue.
- K system capacity being queue size + packets in service.
- N population size bounding the total potential number of packets.
- D service discipline such as FIFO or Priority.

We will only look at $A/S/1$ queues, those being queue models that implicitly do not bound the system capacity or total population size, use FIFO queuing disciplines and have 1 server. The first model we will construct is the $M/M/1$ queue that has an exponential inter-arrival time, exponential service time, and 1 server. Inter-arrivals are exponentially distributed across time, $T_n \sim \text{Exp}(\lambda)$. Likewise, Services are also exponentially distributed across time, $S_n \sim \text{Exp}(\mu)$. let $\rho = \lambda/\mu$ be the utilization and represents how busy the server is. Recall the CTMC process that models how the queue evolves. The $M/M/1$ queue is also a stationary process where $N(t)$ is a birth-death CTMC with birth rate λ and death rate μ for $n \geq 1$. A birth means $N(t)$ enters some specific state at that time t with rate λ and likewise a death means $N(t)$ leaves that state at some other time t with rate μ. The Birth-Death process can also be thought of as the probability that $N(t)$ enters or leaves a state from a pervious state with the previously mentioned rates.

The forward equations for $p_n(t) = P(N(t) = n)$ are

$$\dot{p}_0 = -\lambda p_0 + \mu p_1,$$

$$\dot{p}_n = \lambda p_{n-1} - (\lambda + \mu)p_n + \mu p_{n+1}, \quad (n \geq 1)$$

where a stationary distribution π exists iff $\rho < 1$. If $\rho > 1$, meaning packets enter the queue faster than they are serviced the queue length blows up to infinity and so the π only exists if the queue is stable. The balance equation is [1], p. 307:

$$\pi_n \lambda = \pi_{n+1} \mu \Rightarrow \pi_n = (1-\rho)\rho^n, \quad n \geq 0$$

This has moments, and variance:

$$\mathbb{E}[N] = \frac{\rho}{1-\rho}, \qquad Var(N) = \frac{\rho}{(1-\rho)^2}$$

1.4 Little's Law

Some of the niceties start to emerge here under an $M/M/1$ model as the equations that fall out of the random process end up being very simple and clean. We now know the expectation and variance for the state $N(\cdot)$ but we want to know how long packets stay in the system, what the average length of the queue is, and how long packets stay in the queue. The first set of tools to analyze queues is Little's law first derived in[2] which is a key result in queuing theory and is distribution independent. We know in some sense $N(t) = A(t) - D(t)$, and we are interested in the steady-state averages of the system as $t \to \infty$. Since the area under $N(t)$ on $[0, T]$ gives us the total packet time spent in the system:

$$\int_0^T N(t)dt$$

Then averaging the total time T by how long the system ran for gives us the time-averaged number of packets in the system L:

$$L = \lim_{T \to \infty} \frac{1}{T} \int_0^T N(t)dt$$

Similarly the average arrival rate λ and average departure rates μ are:

$$\lambda = \lim_{T \to \infty} \frac{A(t)}{T}, \qquad \mu = \lim_{T \to \infty} \frac{D(t)}{T}$$

Once again the only stable system is one where in the long run the arrival rate is at most the departure rate. Note how these tools to analyze queuing behavior do not yet depend on the type of queue model, Little's law holds for all queuing models. The sojourn time is the total time spent in the system by a packet. Let

the k-th packet arrive at time a_k and depart at time d_k.

$$T_k = d_k - a_k$$

So the total time spent by all packets in the system that have departed by time T is:

$$\sum_{k=1}^{D(T)} T_k$$

It follows that the area under $N(t)$ which is the total time spent by all packets in the system can be expressed as:

$$\int_0^T N(t)dt = \sum_{k=1}^{D(T)} T_k + R(T)$$

Where $R(T)$ is the total time in the system accumulated by packets that have arrived but not yet departed by time T, being time spent in the queue plus time spent being serviced. If we divide both sides by T the left-hand side becomes the result we showed earlier for L and so we can also express L_T as:

$$\frac{1}{T}\int_0^T N(t)dt = \frac{D(t)}{T}\left(\frac{1}{D(T)}\sum_{k=1}^{D(T)} T_k\right) + \frac{R(T)}{T}$$

$$\rightarrow L_t = \mu_T W_T + \frac{R(T)}{T}$$

With L_T being the time-averaged number of packets in the system over the observation interval $[0, T]$, μ_T being the average departure rate over this same observation interval, and W_T being the average sojourn time of departed packets on the interval. Finally, if we take the limit of this as $T \rightarrow \infty$ and assume the system is stable, that is $\lambda = \mu$ such that $N(t)$ does not blow up and $\frac{R(t)}{T} \rightarrow 0$ intuitively unfinished work at the horizon of the system contributes negligible time to the total time, we get:

$$L = \lim_{T\rightarrow\infty} L_T = \lim_{T\rightarrow\infty} \mu_T W_T = \lambda W$$

$$L = \lambda W$$

A slight modification of the above derivations can be used to find the average length of the queue $L_q = \lambda W_q$

1.5 Building the M/G/1 Queue

Little's law holds for all queuing models we are concerned with, namely single-server unbounded queue length models. More formally it holds for $G/G/1$ which is generalized arrival and generalized departure processes. This generalized process may take any form or shape and some processes in literature are quite exotic. The $M/M/1$ exponential arrival and exponential departure processes queue we already looked at is within the family of $G/G/1$ queues. We will first examine equations for the $M/G/1$ exponential arrival generalized departure, queue which is the backbone for $M/M/1$, and $M/D/1$ queue, which we are interested in for Section 2. For the $M/G/1$ queue, assume Poisson arrivals with rate λ, general i.i.d service S with mean $m = \mathbb{E}[S] = 1/\mu$, second moment $\mathbb{E}[S^2]$, and utilization $\rho = \lambda/\mu < 1$. Without derivation, the Pollaczek-Khinchine (P-K) transform shown in [3] for the waiting time in the queue distribution is:

$$\tilde{W}_q(s) = \mathbb{E}\left[e^{-sW_q}\right] = \frac{(1-\rho)s}{s - \lambda + \lambda B^*(s)},$$

where

$$\rho = \lambda \mathbb{E}[S], \qquad B^*(s) = \mathbb{E}[e^{-sS}]$$

The P-K transform is the Laplace transform of the waiting-time distribution.As mentioned before ρ is the system utilization, $B^*(s)$ is the Laplace-Stieltjes (LST) transform of the service-time distribution, and $\tilde{W}_q(s)$ is the LST of the waiting time distribution being the P-K transform. Differentiating $\tilde{W}_q(s)$ at $s = 0$ gives the expectation for the wait in queue:

$$\mathbb{E}[W_q] = \frac{\lambda \mathbb{E}[S^2]}{2(1-\rho)}$$

The time in system $W = W_q + S$, since W_q and S are independent as service time doesn't depend on the waiting time. The Laplace transforms multiply:

$$\tilde{W}(s) = \tilde{W}_q(s)\tilde{S}(s)$$

With the no-wait probability in the system given by Poisson Arrivals See Time Averages (PASTA):

$$\mathbb{P}(W_q = 0) = 1 - \rho$$

Intuitively PASTA holds because a Poisson process, which our arrival process is, has independent increments and no memory. The proportion of arrivals that happen during periods where $N(t) = n$ is proportional to the total time spent in state n.

1.6 Building the M/M/1 Queue

The sojourn time W is exponentially distributed with rate $\mu - \lambda$.

$$f_W(t) = (\mu - \lambda)e^{-(\mu-\lambda)t}, t \geq 0$$

By using the P-K transform to find $\tilde{W}_q(s)$ and multiplying by $\tilde{S}(s) = \frac{\mu}{\mu+s}$, since there is an exponential service:

$$\tilde{W}_q(s) = (1-\rho)\frac{\mu+s}{\mu+s-\lambda} = (1-\rho) + \frac{(1-\rho)\lambda}{s+(\mu-\lambda)}$$

$$\tilde{W}(s) = \tilde{W}_q\tilde{S}(s) = \frac{\mu-\lambda}{\mu-\lambda+s}$$

and so $W \sim Exp(\mu - \lambda)$ consequently has:

$$\mathbb{E}[W]_{M/M/1} = \frac{1}{\mu-\lambda}, \qquad \mathbb{E}[W_q]_{M/M/1} = \frac{\rho}{1-\rho}\frac{1}{\mu}$$

By Little's law we get:

$$L_{M/M/1} = \frac{\rho}{1-\rho}, \qquad L_{qM/M/1} = \lambda\mathbb{E}[W_q] = \frac{\rho^2}{1-\rho}$$

The busy period B, the time to empty starting from one job has LST determined by:

$$\tilde{B}(s) = \tilde{S}(s + \lambda(1 - \tilde{B}(s))) \Rightarrow \lambda\tilde{B}(s)^2 - (\lambda + \mu + s)\tilde{B}(s) + \mu = 0$$

with the root in $(0, 1]$ the mean busy period is:

$$\mathbb{E}[B]_{M/M/1} = \frac{1}{\mu-\lambda}$$

where the busy period starts when the system is empty and a packet arrives such that $N(t) = 1$. Along the sample path of $N(t)$, the subtle difference is that each sojourn time is the horizontal distance between a single packet's arrival and departure, and the busy period is the horizontal span from first arrival to an empty system. The busy period can be thought of as a cluster of overlapping sojourn times for all packets along this horizontal span.

1.7 Building the M/D/1 Queue

We have defined the $M/G/1$ queue and used the properties to derive the $M/M/1$ queue. We will now do the same with the $M/D/1$ exponential arrival and deterministic departure queue. fix $S \equiv 1/\mu$, so $c_s^2 = 0$ and $\tilde{S}(s) = e^{-s/\mu}$. Once again using the P-K transform to find the :

$$\tilde{W}_q(s) = \frac{(1-\rho)s}{s - \lambda(1 - e^{-s/\mu})}$$

This inverts to a piecewise-exponential density with a point mass at 0 of size $1-\rho$. There is no closed-form elementary PDF but the transform moments are still explicit and defined:

$$\mathbb{E}[W_q]_{M/D/1} = \frac{\lambda(1/\mu^2)}{2(1-\rho)} = \frac{\rho}{1-\rho}\frac{m}{2}$$

For any fixed λ, μ

$$\mathbb{E}[W_q]_{M/D/1} = \frac{1}{2}\mathbb{E}[W_q]_{M/M/1}$$

This is a rather interesting result as for the same mean departure time the wait in the queue with a deterministic service is half that of an exponentially distributed service. This strongly motivates the use of deterministic processors in computer networking from the hardware perspective. A deterministic processor services packets on a clock, where one packet takes exactly as many clock cycles as any other. This finite number of identical operations on a clock cycle implies deterministic service times. This subtly differs from general purpose x86 or ARM CPU's where the CPU's internal scheduler, how interrupts are handled, and CPU resource heterogeneity make no guarantees on service time and must be approximated with some distribution. By Little we find the length of the queue and total number in-system:

$$L_{qM/D/1} = \lambda\mathbb{E}[W_q]_{M/D/1} = \frac{\rho^2}{2(1-\rho)}, \qquad L_{M/D/1} = L_{qM/D/1} + \rho = \frac{\rho(2-\rho)}{2(1-\rho)}$$

1.8 Key Results and Remark

To recap: First we defined a queue as a random process whose state in time is defined by both an arrival, and departure process. Next we looked at three highly similar properties for a RV with different forms of time invariance. They were: 1) The memoryless property, where the future depends only on the present, 2) the Markov property, where the future state of the process only depends on the current state not on a sequence of past states, and 3) stationarity, where

the underlying distributions of the RV do not change over time. These properties are the foundation for closed-form analysis of $M/G/1$ queues, and in fact a Poisson arrival is enough to make queues with an exponential arrival and service-time distribution analytically tractable. Some reasons for this have already been mentioned but PASTA and CTMC are crucial ones. PASTA with Poisson input gives us the ability to interpret performance results and measurements. An intuitive example to demonstrate why: Imagine a router in a network where the arrivals are Poisson, that is completely random with no synchronization; the likelihood a packet arrives during a busy or idle period is exactly equal to how often the busy or idle periods occur. If the router is busy 80% of the time then 80% of packets will find it busy when they arrive meaning they will be enqueued and wait accumulating W_q^p for that packet p. If the arrivals are deterministic where packets arrive every 10 seconds and suppose the average service time is 8 seconds, the system alternates from busy to idle in a highly regular pattern. Every individual packet arrives just after an idle period starts so each arrival finds the system idle almost every time but looking at the time average the system is "busy" 80% of the time. $P(server - busy - at - arrival) \neq P(server - busy - at - random - time)$. Effectively taking measurements as seen by the packets where they yield different results than measurements taken across the system averaged on time, so the results are not directly interpretable. CTMC and closure for $M/M/1$ queues specifically creates favorable scenarios for downstream analysis. The exact Markov structure created by an exponential service time and exponential inter-arrival time makes $N(t)$ a one-dimensional CTMC meaning it has linear balance equations that give closed-forms. Non-exponential service requires higher-dimensional states that hold residual service greatly complicating analysis. This is highly related to closure of $M/M/1$ queues where exponential service time and exponential inter-arrival times imply reversibility. The key property here is that departures remain a Poisson process meaning downstream queues in the network can also correctly assume $M/M/1$ models and be interpretable and tractable. It is this key property under specific circumstances that Section 2 examines and challenges.

Result	$M/M/1$	$M/D/1$
$\mathbb{E}[L]$	$\frac{\rho}{1-\rho}$	$\frac{\rho(2-\rho)}{2(1-\rho)}$
$\mathbb{E}[L_q]$	$\frac{\rho^2}{1-\rho}$	$\frac{\rho^2}{2(1-\rho)}$
$\mathbb{E}[W_q]$	$\frac{\rho}{\mu-\lambda}$	$\frac{\rho}{2\mu(1-\rho)}$
$\mathbb{P}(W_q = 0)$	$1-\rho$	$1-\rho$
$\mathbb{E}[W]$	$\frac{1}{\mu-\lambda}$	$\frac{\rho}{2\mu(1-\rho)} + \frac{1}{\mu}$

Table 1: Key Results for Queuing Models

2 Applications of Queuing Models

2.1 Case Study: M/M/1 Failures

Consider a computer network with a source of packets $S1$ that has an exponential inter-packet time and two routers $R1$, $R2$ where $R2$ acts as a sink or destination for all the packets. $R1$ services packets as they arrive according to some service process. A key result from Section 1 was reversibility meaning that if $R1$'s departure process is also exponential then the ingress or arrival process to $R2$ will also be exponential. This implies if $R1$ is $M/M/1$, $R2$ is at least $M/G/1$ and if $R2$ also has an exponential departure process then it will likewise be $M/M/1$. What happens in the scenario where $R1$ is NOT $M/M/1$ but $M/G/1$ for any generalized departure process that is not exponential? What would the distribution of inter-arrival times λ^* look like at $R2$, and what properties would it have?

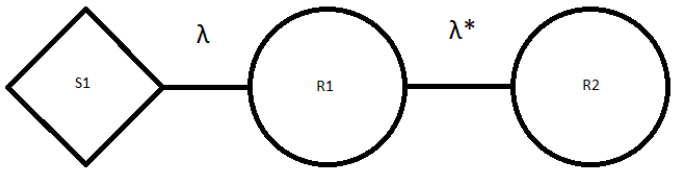

Figure 3: Two router line network with source $S1$

Let's consider a direct example for which we already have equations for and model $R1$ as $M/D/1$. The motivation for this is quite realistic, as real network equipment uses multi-stage pipeline packet processors that have a deterministic number of operations and therefor deterministic service time.

Let $\mu = 1/S$ and $\rho := \lambda/\mu = \lambda S,\ \rho \in (0,1)$. Let D_n be the time of the n-th departure from $R1$ and $Y_n := D_{n+1} - D_n$ is the inter-departure time. The arrival process to $R2$ is then $\{D_n\}_{n\geq 1}$.

Let Q_n be the number of packets in the system immediately after the n-th departure from $R1$. There are essentially two cases that Q_n can take: the system empties after the n-th departure, or the system is non-empty and a packet is still waiting to be serviced and so the system immediately starts servicing the $n+1$-th packet.

If $Q_n > 0 \rightarrow Y_n = S$

If $Q_n = 0$, the system empties out and idles until the next arrival. Since the arrival process is a Poisson process the idle time $I_n \sim Exp(\lambda)$ by memoryless property: $Y_n = I_n + S$. Let $p := \mathbb{P}(Q_n > 0)$ takes expectation:

$$\mathbb{E}[Y] = pS + (1-p)(S + \frac{1}{\lambda}) = S + \frac{1-p}{\lambda}$$

Throughput conservation gives $\mathbb{E}[Y] = 1/\lambda$ since in stable lossless queues, that is $R1$'s buffer size ≥ 2, the input rate equals the output rate as seen in Section 1. Hence:

$$p = \lambda S = \rho$$

The exact law of Y is:

$$\mathbb{P}(Y = S) = \rho, \qquad f_Y(y) = (1-\rho)\lambda e^{-\lambda(y-S)}\mathbf{1}_{\{y \geq S\}}.$$

This reads the probability that the inter-departure time random gap is the constant $S = \rho$. The immediate consequences of this are $Y \geq S$ almost surely, and $\mathbb{P}(Y = S) = \rho$, and the squared coefficient of variation (SCV) of Y is:

$$c_Y^2 = \frac{Var(Y)}{\mathbb{E}[Y]^2} = 1 - \rho^2 < 1$$

Therefore:

$$\mathbb{E}[Y] = 1/\lambda, \;\; \mathbb{E}[Y^2] = \frac{(2-\rho^2)}{\lambda^2}$$

2.2 Arrival Process Inconsistencies

This Arrival process Y at $R2$ from $M/D/1$ at $R1$ cannot be exponential, and it is easy to see why. Exponential distributions have no atoms, or point masses, and a positive probability on every interval $(0, \epsilon)$. $\mathbb{P}(Y = S) = \rho > 0$, which is a point mass at S, and $\mathbb{P}(Y < S) = 0$ which is not a positive probability within the interval. So the packets arriving to $R2$ with rate λ^* are not exponentially distributed. The SCV $c_Y^2 = 1 - \rho^2$ is not the same as the SCV identity of an exponential process $c^2 = 1$, again meaning the arrival process to $R2$ is not exponential. Another way to think about what is happening is that in very light traffic with a very high probability that $Q_n = 0$ after a departure Y is close to $S + Exp(\lambda)$. This reflects the anti-bunching behavior caused by the deterministic service time. Without the constant shift right the tail is exponential but as we have seen earlier there is a point mass at S. This can be thought of as the minimum lag in the system as a packet cannot possibly leave faster than the

time taken by the deterministic service where if the service was exponentially distributed there is no point mass anywhere as the service time can take any value. In heavy traffic scenarios, the arrival process at $R2$ looks very much like a deterministic stream of packets. We can reason that, as λ approaches μ, the tail of Y collapses and only a single point mass is left at S. This is caused by $R1$ being constantly busy with a nonzero queue length most of the time. In essence the packet stream leaving $R1$ will just look like $R1$ is generating packets on a deterministic cycle.

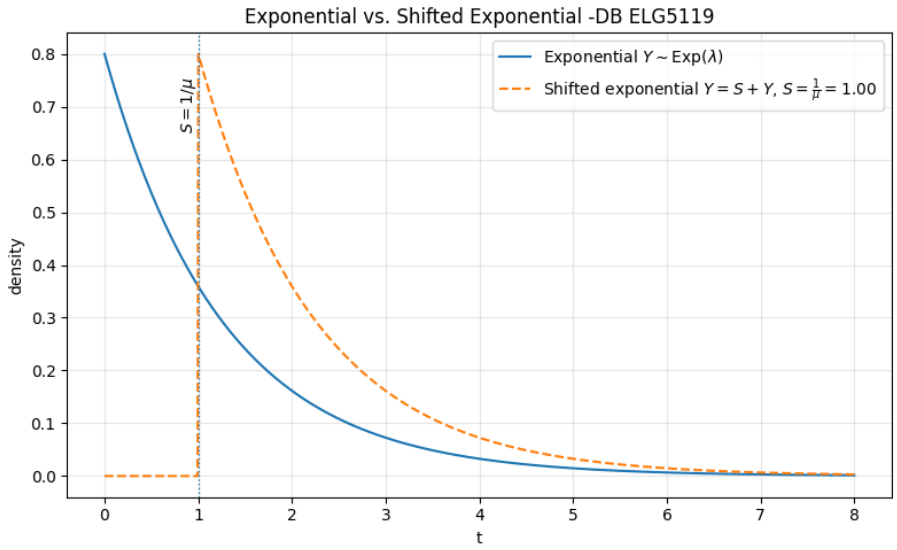

Figure 4: Point mass at S with $EXP(\lambda)$ tail

In any case, the random gap Y is some mixture and, more importantly breaks the memoryless and Markov properties at $R2$. From the exact law we have:

$$\mathbb{P}(Y = S) = \rho, \qquad \mathbb{P}(Y < S) = 0, \qquad c_Y^2 = 1 - \rho^2 < 1$$

so the gap created satisfies $Y \geq S$ almost surely and there is a positive point mass at S. An exponential inter-arrival has no such point mass and has $c^2 = 1$, so Y cannot be exponential. We can see that Y is also not memoryless. Memorylessness requires:

$$\mathbb{P}(Y > t + s | Y > t) = \mathbb{P}(Y > s) \;\; \forall s, t \geq 0$$

From the density of Y, for $y > S$ the tail is:

$$\mathbb{P}(Y > s) = (1-\rho)e^{-\lambda(y-S)}$$

While $\mathbb{P}(Y > S) = 1-\rho$ since $\mathbb{P}(Y = S) = \rho$. Take $t = S/2$ and $s = S$, because $Y \geq S$ almost surely $\mathbb{P}(Y > S/2) = 1$ and:

$$\mathbb{P}(Y > t+s|Y > t) = \mathbb{P}(Y > 3S/2) = (1-\rho)e^{-\lambda S/2}$$

For any $0 < \rho < 1$, we have $e^{-\lambda S/2} < 1$ so:

$$\mathbb{P}(Y > t+s|Y > t) = (1-\rho)e^{-\lambda S/2} < 1-\rho = \mathbb{P}(Y > S)$$

Therefore Y fails to hold the memoryless identity and the arrival process to R_2 is not Poisson. If the arrivals to $R2$ do not fit either our $M/D/1$ or $M/M/1$ models is there any reasonable way to model and analyze the queue at $R2$? Since the arrivals are sub-Poisson with less variability ($c_a^2 < 1$) than a Poisson process and a hard minimum spacing plus a point mass at S_1, a reasonable model for $R2$ is a $G/D/1$ queue. There are no closed-form equations for a $G/D/1$ so we will use Kingman's well-known approximation for the mean waiting time derived in [4] for $G/G/1$ queues:

$$\mathbb{E}[W_q^{R2}] \approx \frac{\rho_2}{1-\rho_2} \cdot \frac{c_{a,2}^2 + c_{d,2}^2}{2} S_2$$

Where $\rho_2 = \lambda_2 S_2 = \lambda_2/\mu_2, c_{d,2}^2 = 0$ since the departure time is deterministic, and $c_{a,2}^2$ is the SCV of the arrivals to $R2$. We can approximate this from the downstream Sub-Poisson from $R1$, $c_{a,2}^2 \approx 1-\rho_1^2$. Because $c_{a,2}^2 < 1$ modeling $R2$ as $M/D/1$ would overestimate the delay both $\mathbb{E}[W_q]$ and $\mathbb{E}[W]$. Arrivals to $R2$ are still a renewal (i.i.d inter-arrival times) with inter-arrival $Y \approx S_1 + EXP(\lambda_{effective})$. So with this we can now use Kingman with the approximated Y as a pure shifted-exponential:

$$c_{a,2}^2 \approx c_Y^2 = 1-\rho_1^2 < 1$$

So the wait in the queue at $R2$ is given by:

$$\mathbb{E}[W_q^{R2}] = \frac{\rho_2}{1-\rho_2} \cdot \frac{1-\rho_1^2}{2} S_2$$

This is rather arduous to deal with and arises out of a simple case where there are two queues strung together and the first is simply not an $M/M/1$ queue. If we modeled each as an $M/D/1$ queue ignoring the destruction of the Poisson

process that gives us memoryless and Markov properties, we would still overestimate the wait times at $R2$. For $M/D/1$ $c_a^2 = 1, c_d^2 = 0$ giving:

$$\mathbb{E}[W_q]_{M/D/1} = \frac{\rho}{1-\rho} \cdot \frac{1}{2} S$$

Which is exact for $M/D/1$, if arrivals are smoother such as in the above $G/G/1$ case where $c_a^2 < 1$:

$$\mathbb{E}[W_q]_{G/G/1} = \frac{\rho}{1-\rho} \cdot \frac{c_a^2}{2} S$$

is smaller by a factor of c_a^2. A quick example:

Suppose $\rho_2 = 0.8, S_2 = 1$ if $R1$ modeled as $M/D/1$ yields $c_{a,2}^2 = 0.64$ ($\rho_1 = 0.6,\ c_{a,2}^2 = 1 - \rho_1^2$) then $M/D/1$ at $R2$ predicts:

$$\mathbb{E}[W_q]_{M/D/1} = \frac{0.8}{0.2} \cdot \frac{1}{2} \cdot 1 = 2$$

Subsequently $M/M/1$ predicts:

$$\mathbb{E}[W_q]_{M/M/1} = \frac{0.8}{0.8 - 0.6} = 4$$

Using the correct SCV for arrivals and approximating $G/D/1$ at $R2$ predicts:

$$\mathbb{E}[W_q]_{G/D/1} \approx \frac{0.8}{0.2} \cdot \frac{0.64}{2} \cdot 1 = 1.28$$

The second router is being fed a smoother-than-Poisson stream, any model that assumes it sees Poisson traffic will systematically overstate the delay. In this case even $M/D/1$ overstates the delay by 56% which is beyond significant. Delay-aware routing algorithms will take vastly different paths through a network for a few % difference in wait times. Consider a more realistic or extensible case with n routers in a line where the service time of each is deterministic. Is there any easy way to approximate how long packets would spend in each queue or the entire network?

2.3 Queues in Series

If we initialize with Poisson into $R1$: $c_{a,1}^2 = 1$, for routers $i = 1, ..., n$ and assume a utilization $\rho_i = \lambda S_i$, meaning there are no losses or dropping of packets in the network we find the approximate arrival-variability recursion:

$$c_{a,i+1}^2 \approx (1 - \rho_i^2) c_{a,i}^2, \qquad c_1^2 = 1$$

This is a natural extension from the results of Section 2.1, where we found

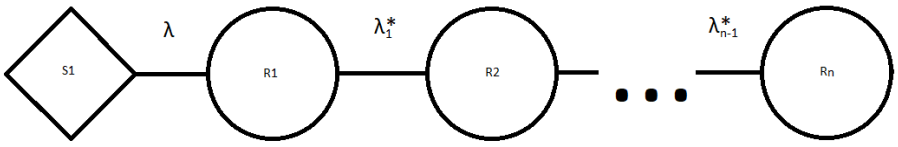

Figure 5: n router line network with source $S1$

that an $M/D/1$ queue with Poisson input $c^2_{a,1} = 1$ produces output with SCV $c^2_Y = 1 - \rho^2_1$. Applying the recursion repeatedly up to the n-th router gives:

$$c^2_{a,n} = \prod_{i=1}^{n-1}(1 - \rho^2_i)$$

This shows the squared covariance shrinks at each hop meaning traffic smoothes out through the network. This smoothing by the deterministic services is why end-to-end delays can be much lower than $M/M/1$ or $M/D/1$ suggest. The mean wait at router i:

$$\mathbb{E}[W^i_q] \approx \frac{\rho_i}{1 - \rho_i} \cdot \frac{c^2_{a,i}}{2} S_i$$

and the end-to-end mean delay of Queuing + Service at each router:

$$\mathbb{E}[T_{e2e}] \approx \sum_{i=1}^{n} \left(\mathbb{E}[W^i_q] + S_i\right)$$

These results are a simplified case of Whitt's Queuing Network Analyzer decomposition for queues in series in[5]. There are refinements for heavy traffic scenarios, but the core idea is that propagating the SCV for the arrival process conditioned by each previous queue (router) allows close approximation for what would otherwise be intractable with no closed-form solutions. You may think of the path a packet takes through a random network as a series of routers $\{1, ..., n\}$ and so the wait in the queue along this path could be closely approximated with a $G/D/1$ model, given that our previous assumptions on router behavior hold. In short, while $M/M/1$ greatly simplifies mathematical analysis of computer networks there are very strict assumptions placed on the underlying processes. It is common for service times to be highly deterministic in modern data networks and while this breaks many of the nice tractable properties of an appropriate queuing model, improperly modeling the queue in order to keep analysis simple can significantly reduce the accuracy of the analysis.

References

[1] R. G. Gallagher, *Stochastic Processes: Theory for Applications*, Springer, 2012.

[2] J. D. C. Little and S. C. Graves, "Little's law," in *Building Intuition*, International Series in Operations Research & Management Science, vol. 115, Springer, 2008, p. 81, doi: 10.1007/978-0-387-73699-0_5, ISBN 978-0-387-73698-3.

[3] L. Takács, "Review: J. W. Cohen, *The single-server Queue*," *Annals of Mathematical Statistics*, vol. 42, no. 6, pp. 2162–2164, Dec. 1971, doi: 10.1214/aoms/1177693087.

[4] J. F. C. Kingman, "The single-server queue in heavy traffic," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 57, no. 4, p. 902, Oct. 1961, doi: 10.1017/S0305004100036094.

[5] S. S. W. Whitt, "Arranging queues in series," AT&T Engineering Research Center, AT&T Bell Laboratories, Murray Hill, NJ, USA, Technical Report, June 21, 1988, Revision: April 21, 1989.