17 M/D/1 and M/G/1 Queues

17.1 Motivation

In this simulation experiment, we will study a model that is important to understand the queuing and delay phenomena in packet communication links. Let us consider the network shown in **Figure 17-1**. Wired_Node_1 is transmitting UDP packets to Wired_Node_2 through a router. Link 1 and Link 2 are of speed 10 Mbps. The packet lengths are 1250 bytes plus a 54-byte header, so that the time taken to transmit a packet on each 10 Mbps link is $\frac{1304\times8}{10}$ μsec =1043.2 μsec. In this setting, we would like answers to the following questions:

1. We notice that the maximum rate at which these packets can be carried on a 10 Mbps link is $\frac{10^6}{1043.2} = 958.59$ packets per second. Can the UDP application send packets at this rate?
2. The time taken for a UDP packet to traverse the two links is $2 \times 1043.2 = 2086.4$ μsec. Is this the time it actually takes for a UDP packet generated at Wired_Node_1 to reach Wired_Node_2.

The answer to these questions depends on the manner in which the UDP packets are being generated at Wired_Node_1. If the UDP packets are generated at intervals of 1043.2 μsec then successive packets will enter the Link 1, just when the previous packet departs. In practice, however, the UDP packets will be generated by a live voice or video source. Depending on the voice activity, the activity in the video scene, and the coding being used for the voice and the video, the rate of generation of UDP packets will vary with time. Suppose two packets were generated during the time that one packet is sent out on Link 1, then one will have to wait, giving rise to queue formation. This also underlines the need for a buffer to be placed before each link; a buffer is just some dynamic random-access memory in the link interface card into which packets can be stored while waiting for the link to free up.

Queuing models permit us to understand the phenomenon of mismatch between the service rate (e.g., the rate at which the link can send out packets) and the rate at which packets arrive. In the network in **Figure 17-1**, looking at the UDP flow from Wired_Node_1 to Wired_Node_2, via Router 1, there are two places at which queueing can occur. At the interface between Wired_Node_1 and Link 1, and at the interface between Router 1 and Link 2. Since the only flow of packets is from Wired_Node_1 to Wired_Node_2, all the packets entering Link 2 are from Link 1, and these are both of the same bit rate. Link 2,

therefore, cannot receive packets faster than it can serve them and, at any time, only the packet currently in transmission will be at Link 2. On the other hand at the Wired_Node_1 to Link 1 interface, the packets are generated directly by the application, which can be at arbitrary rates, or inter-packet times.

Suppose that, at Wired_Node_1, the application generates the successive packets such that the time intervals between the successive packets being generated are statistically independent, and the probability distribution of the time intervals has a negative exponential density, i.e., of the form $\lambda\, e^{-\lambda x}$, where λ (packets per second) is a parameter, called the *rate* parameter, and x (seconds) is the argument of the density. The application generates the entire packet *instantaneously,* i.e., all the bits of the packet arrive from the application together, and enter the buffer at Link 1, to wait behind the other packets, in a first-in-first-out manner. The resulting random process of the points at which packets enter the buffer of Link 1 is called a Poisson Process of rate λ packets per second. The buffer queues the packets while Link 1 serves them with *service time* $b = 1043.2\ \mu$sec. Such a queue is called an M/D/1 queue, where the notation is to be read as follows.

- The M before the first slash (denoting "Markov") denotes the Poisson Process of instants at which packets enter the buffer.
- The D between the two slashes (denoting "Deterministic") denotes the fixed time taken to serve each queued packet.
- The 1 after the second slash denotes that there is just a single server (Link 1 in our example)

This way of describing a single server queueing system is called Kendall's Notation.

In this experiment, we will understand the M/D/1 model by simulating the above-described network on NetSim. The M/D/1 queueing model, however, is simple enough that it can be mathematically analyzed in substantial detail. We will summarize the results of this analysis in the next section. The simulation results from NetSim will be compared with the analytical results.

17.2 Mathematical Analysis of the M/D/1 Queue

The M/D/1 queueing system has a random number of arrivals during any time interval. Therefore, the number of packets waiting at the buffer is also random. It is possible to mathematically analyze the *random process* of the number of waiting packets. The procedure for carrying out such analysis is, however, beyond the scope of this document.

We provide the final formulas so that the simulation results from NetSim can be compared with those provided by these formulas.

As described earlier, in this chapter, the M/D/1 queue is characterized by two parameters: λ (packets per second), which is the arrival rate of packets into the buffer, and μ (packets per second), which is the rate at which packets are removed from a nonempty queue. Note that $1/\mu$ is the service time of each packet.

Define $\rho = \lambda \times \frac{1}{\mu} = \lambda/\mu$. We note that ρ is the average number of packets that arrive during the service time of a packet. Intuitively, it can be expected that if $\rho > 1$ then packets arrive faster than the rate at which they can be served, and the queue of packets can be expected grow without bound. When $\rho < 1$ we can expect the queue to be "stable." When $\rho = 1$, the service rate is exactly matched with the arrival rate; due to the randomness, however, the queue can still grown without bound. The details of this case are beyond the scope of this document.

For the k^{th} arriving packet, denote the instant of arrival by a_k, the instant at which service for this packet starts as s_k, and the instant at which the packet leaves the system as d_k. Clearly, for all $k, d_k - s_k = \frac{1}{\mu}$, the deterministic service time. Further define, for each k,

$W_k = s_k - a_k$

$T_k = d_k - a_k$

i.e., W_k is called the *queuing delay,* i.e., time from the arrival of the k^{th} packet until it starts getting transmitted, whereas T_k is called the *total delay,* i.e., the time from the arrival of the k^{th} packet until its transmission is completed. Considering a large number of packets, we are interested in the average of the values $W_1, W_2, W_3, \cdots$, i.e., the *average queueing time* of the packets. Denote this average by W. By mathematical analysis of the packet queue process, it can be shown that for an M/D/1 queueing system,

$$W = \frac{1}{2\mu} \times \frac{\rho}{1-\rho}$$

Denoting by T, the average total time in the system (i.e., the average of $T_1, T_2, T_3, \cdots$), clearly

$$T = W + \frac{1}{\mu}.$$

Observe the following from the above formula:

1. As ρ approaches 0, W becomes 0. This is clear, since, when the arrival rate becomes very small, and arriving packet sees a very small queue. For arrival rate approaching 0, packets get served immediately on arrival.
2. As ρ increases, W inreases.
3. As ρ approaches 1 (from values smaller than 1), the mean delay goes to ∞.

We will verify these observations in the NetSim simulation.

17.3 The Experimental Setup

The model described at the beginning of this chapter is shown in **Figure 17-1.**

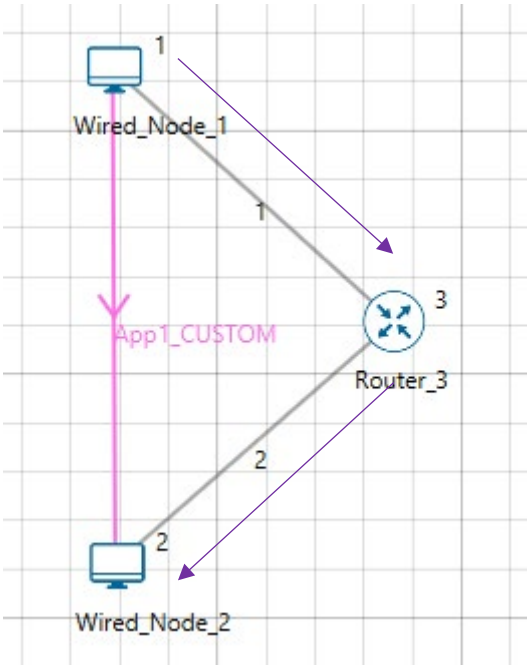

Figure 17-1: A single wired node (Wired_Node_1) sending UDP packets to another wired node (Wired_Node_2) through a router (Router 3). The packet interarrival times at Wired_Node_1 are exponentially distributed, and packets are all of the same length, i.e., 1250 bytes plus UDP/IP header.

Open NetSim and Click on **Examples > Experiments > MD1-and-MG1-Queues > Sample-1** as shown below **Figure 17-2**.

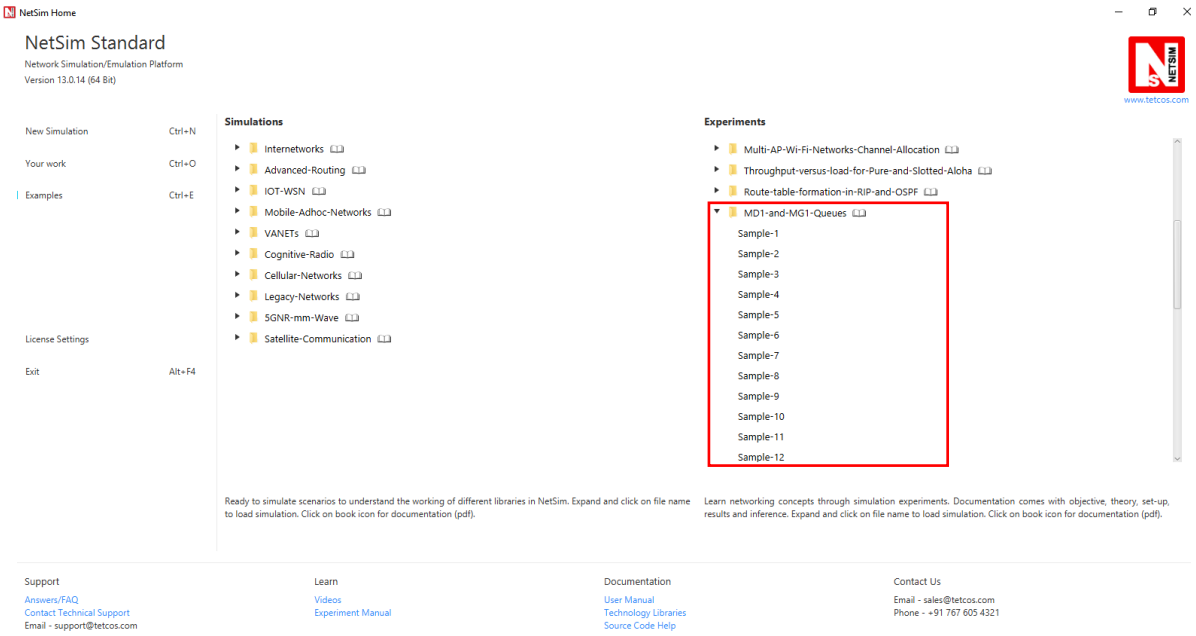

Figure 17-2: Experiments List

NetSim UI displays the configuration file corresponding to this experiment as shown above:

17.4 Procedure

Sample 1

The following set of procedures were done to generate this sample:

Step 1: A network scenario is designed in NetSim GUI comprising of 2 Wired Nodes and 1 Router in the **"Internetworks"** Network Library.

Step 2: Link Properties are set as per the table given below **Table 17-1**.

Link Properties	Link 1	Link 2
Uplink Speed (Mbps)	10	10
Downlink Speed (Mbps)	10	10
Uplink BER	0	0
Downlink BER	0	0
Uplink Propagation Delay (µs)	0	0
Downlink Propagation Delay (µs)	0	0

Table 17-1: Wired link properties

Step 3: Right click on the Application Flow **App1 CUSTOM** and select Properties or click on the Application icon present in the top ribbon/toolbar.

A CUSTOM Application is generated from Wired Node 1 i.e., Source to Wired Node 2 i.e., Destination. Transport Protocol is set to **UDP** with Packet Size set to 1250 Bytes and Inter Arrival Time set to 20863 µs and distribution to Exponential.

The Packet Size and Inter Arrival Time parameters are set such that the Generation Rate equals 0.096 Mbps. Generation Rate can be calculated using the formula:

$$Generation\ Rate\ (Mbps)\ =\ Packet\ Size\ (Bytes)\ *\ 8/Interarrival\ time\ (\mu s)$$

Step 4: Packet Trace is enabled in NetSim GUI. At the end of the simulation, a very large .csv file is containing all the packet information is available for the users to perform packet level analysis. Plots is enabled in NetSim GUI.

Step 5: Run the Simulation for 100 Seconds.

Similarly, the other samples are created by changing the Inter Arrival Time per the formula

$$IAT = \frac{10^6}{958.59 * \rho}$$

as per the table given below **Table 17-2.**

ρ	IAT (μs)
0.05	20863
0.10	10431
0.15	6954
0.20	5215
0.25	4172
0.30	3477
0.35	2980
0.40	2607
0.45	2318
0.50	2086
0.55	1896
0.60	1738
0.65	1604
0.70	1490
0.75	1390
0.80	1303
0.85	1227
0.90	1159
0.95	1098

Table 17-2: Inter Arrival Time Settings

Even though the packet size at the application layer is 1250 bytes, as the packet moves down the layers, overhead is added. The overheads added in different layers are shown in the below table and can be obtained from the packet trace:

Layer	Overhead (Bytes)
Transport Layer	8
Network Layer	20
MAC layer	26
Physical Layer	0

Total	54

Table 17-3: Overheads added to a packet as it flows down the network stack

17.5 Obtaining the Mean Queuing delay from the Simulation Output:

After running the simulation, note down the **"Mean Delay"** in the Application Metrics within the Results Dashboard. This is the average time between the arrival of packets into the buffer at Wired_Node_1, and their reception at Wired_Node_2.

As explained in the beginning of this chapter, for the network shown in **Figure 17-1**, the end-to-end delay of a packet is the sum of the queueing delay at the buffer between the wired-node and Link_1, the transmission time on Link_1, and the transmission time on Link_2 (there being no queueing delay between the Router and Link_2). It follows that.

$$Mean\ Delay = \left(\frac{1}{2\mu} \times \frac{\rho}{1-\rho}\right) + \frac{1}{\mu} + \frac{1}{\mu}$$

17.6 Output Table

Sample	ρ	λ	Mean Delay (μs)	Queuing Delay (μs) (Simulation)	Queuing Delay (μs) (Theory)
1	0.05	47.93	2112.87	26.47	27.45
2	0.10	95.86	2144.01	57.61	57.96
3	0.15	143.79	2178.86	92.46	92.05
4	0.20	191.72	2218.09	131.69	130.40
5	0.25	239.65	2259.11	172.71	173.87
6	0.30	287.58	2309.49	223.09	223.54
7	0.35	335.51	2365.74	279.34	280.86
8	0.40	383.44	2435.65	349.25	347.73
9	0.45	431.37	2513.79	427.39	426.76
10	0.50	479.30	2608.38	521.98	521.60
11	0.55	527.22	2721.59	635.19	637.51
12	0.60	575.15	2864.88	778.48	782.40
13	0.65	623.08	3052.84	966.44	968.68
14	0.70	671.01	3304.58	1218.18	1217.07
15	0.75	718.94	3633.66	1547.26	1564.80
16	0.80	766.87	4160.39	2073.99	2086.40
17	0.85	814.80	5115.95	3029.55	2955.73
18	0.90	862.73	6967.16	4880.76	4694.39
19	0.95	910.66	12382.98	10296.58	9910.39

Table 17-4: Mean Delay, Queueing delay from Simulation and Queueing delay from analysis

Comparison Charts:

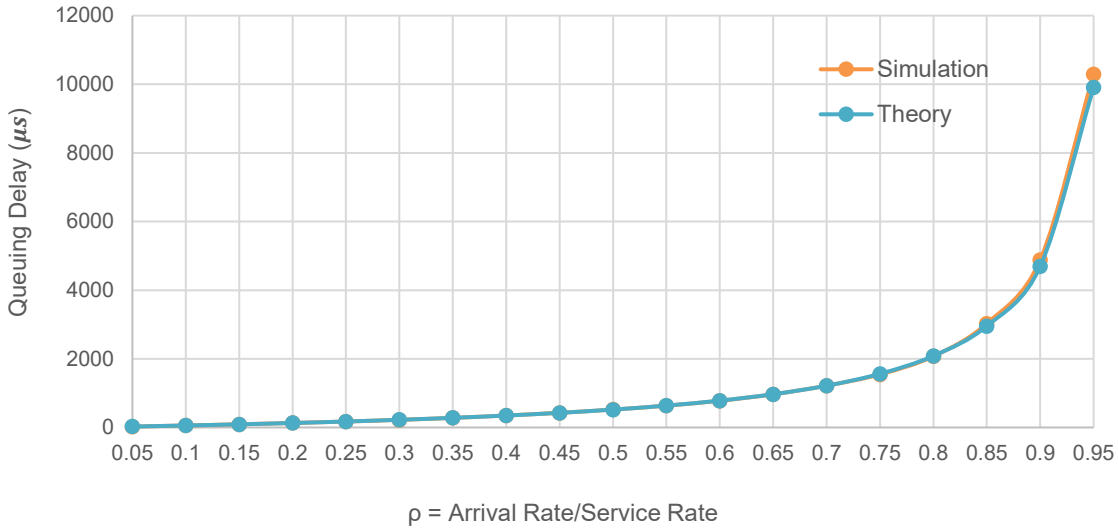

Figure 17-3: Comparison of queueing delay from simulation and analysis

17.7 Advanced Topic: The M/G/1 Queue

In Section 17.1, we introduced the M/D/1 queue. Successive packets were generated instantly at exponentially distributed time intervals (i.e., at the points of a Poisson process); this gave the “M” in the notation. The packets were all of fixed length; this gave the “D” in the notation. Such a model was motivated by the transmission of packetized voice over a fixed bit rate wireline link. The voice samples are packetized into constant length UDP packets. For example, typically, 20ms of voice samples would make up a packet, which would be emitted at the instant that the 20ms worth of voice samples are collected. A voice source that is a part of a conversation would have listening periods, and “silence” periods between words and sentences. Thus, the intervals between emission instants of successive UDP packets would be random. A simple model for these random intervals is that they are exponentially distributed, and independent from packet to packet. This, formally, is called the Poisson point process. With exponentially distributed (and independent) inter-arrival times, and fixed length packets we obtain the M/D/1 model. On the other hand, some applications, such as video, generate unequal length packets. Video frames could be encoded into packets. To reduce the number of bits being transmitted, if there is not much change in a frame, as compared to the previous one, then the frame is encoded into a small number of bits; on the other hand if there is a large change then a large number of bits would need to be used to encode the new information in the frame. This motivates variable packet sizes. Let us suppose that, from such an application, the packets arrive at the points of a Poisson

process of rate λ, and that the randomly varying packet transmission times can be modelled as independent and identically distributed random variables, $B_1, B_2, B_3, \cdots$, with mean b and second moment $b^{(2)}$, i.e., variance $b^{(2)} - b^2$. Such a model is denoted by M/G/1, where M denotes the Poisson arrival process, and G ("general") the "generally" distributed service times. Recall the notation M/D/1 (from earlier in this section), where the D denoted fixed (or "deterministic") service times. Evidently, the M/D/1 model is a special case of the M/G/1 model.

Again, as defined earlier in this section, let W denote the mean queueing delay in the M/G/1 system. Mathematical analysis of the M/G/1 queue yields the following formula for W

$$W = \frac{\rho}{1-\rho}\frac{b^{(2)}}{2b}$$

where, as before, $\rho = \lambda b$. This formula is called the Pollacek-Khinchine formula or P-K formula, after the researchers who first obtained it. Denoting the variance of the service time by $Var(B)$, the P-K formula can also be written as

$$W = \frac{\rho b}{2(1-\rho)}\left(\frac{Var(B)}{b^2} + 1\right)$$

Applying this formula to the M/D/1 queue, we have $Var(B) = 0$. Substituting this in the M/G/1 formula, we obtain.

$$W = \frac{\rho}{1-\rho}\frac{b}{2}$$

which, with $b = 1/\mu$, is exactly the M/D/1 mean queuing delay formula displayed earlier in this section.

17.8 A NetSim Exercise Utilising the M/G/1 Queue

In this section we demonstrate the use of the M/G/1 queueing model in the context of the network setup shown in **Figure 17-1**. The application generates exponentially distributed data segment with mean d bits, i.e., successive data segment lengths are sampled independently from an exponential distribution with rate parameter $\frac{1}{d}$. Note that, since packets are integer multiples of bits, the exponential distribution will only serve as an approximation. These data segments are then packetised by adding a constant length header of length h bits. The packet generation instants form a Poisson process of rate λ. Let us denote the link speed by

c. Let us denote the random data segment length by X and the packet transmission time by B, so that.

$$B = \frac{X + h}{c}$$

Denoting the mean of B by b, we have

$$b = \frac{d + h}{c}$$

Further, since h is a constant,

$$Var(B) = Var(X)/c^2$$

These can now be substituted in the P-K formula to obtain the mean delay in the buffer between Node 1 and Link 1.

We set the mean packet size to 100B or 800 bits, the header length h = 54B or 432 bits and $\lambda = 5000$

For a 10Mbps link, the service rate $\mu = \frac{10*10^6}{154\times8} = 8116.8$

Using the Pollaczek–Khinchine (PK) formula, the waiting time for a M/G/1 queuing system is

$$w = \frac{\rho + \lambda \times \mu \times Var(s)}{2(\mu - \lambda)}$$

Where $var(s)$ is the variance of the service time distribution S. Note that $var(s) = \frac{1}{(\mu')^2}$ where μ' is the mean service time of the exponential random variable (100B packets and not 154B)

$$\mu' = \frac{10 \times 10^6}{100 * 8} = 12500$$

Hence substituting into the PK formula, one gets

$$w = \frac{0.4 + \frac{(3467.7 \times 8116.8)}{12500^2}}{2\,(8116.8 - 3246.7)} = 59.5\ \mu s$$

By simulation the queuing delay is 60.5 μs.

The queuing delay is not available in the NetSim results dashboard. It can be got from the packet trace. It is the average of (PHY_layer_Arrival_time - APP_layer_arrival time) for packets being sent from Node_1.