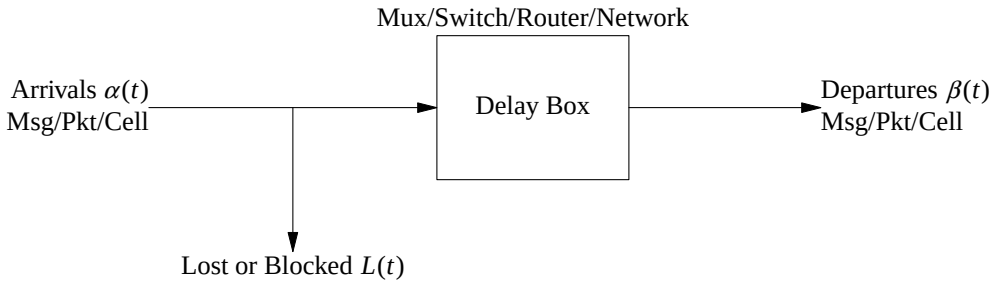
Chapter V: Analysis of Packet Switching Networks

1. Introduction

1.1 Basic Model of delay/loss system



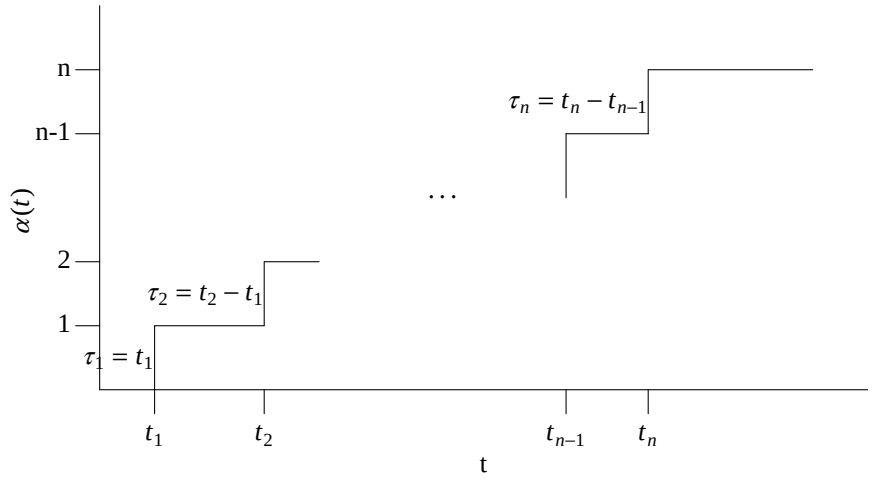
1. interested in calculating
   - — time spent in the system T
   - — number of customers in the system N(t)
   - — fraction of arriving customers that are lost P_b
   - — avg number of messages per second passing through the system - throughput
2. How do we determine the number of customers in the system at time t, $N(t)$
   - — If $\alpha(t)$ is the number of customers that arrive from 0->t
   - — $L(t)$ is the number of customers that are lost because they are blocked from entering from 0->t,

     $$\bar{L} = \lim_{t\to\infty} \frac{L(t)}{t}, \text{ and}$$

   - — $\beta(t)$ is the number of customers that depart from the system from 0->t then
   - — $N(t) = \alpha(t) - L(t) - \beta(t)$
3. The arrival rate into the system -- avg number of customers per second that arrive is $\lambda = \lim_{t\to\infty} \frac{\alpha(t)}{t}$

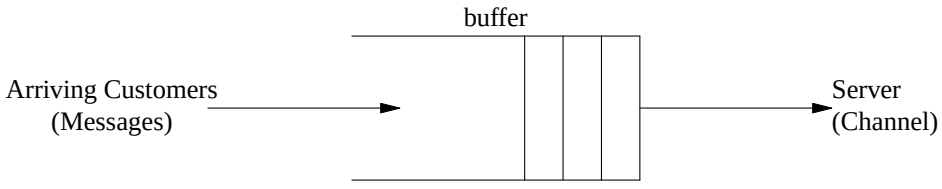
   If the interarrival time between customer $i-1$ and i is τ_i, then $\lambda = \lim_{n\to\infty} \frac{n}{\tau_1 + \tau_2 + \cdots + \tau_n} = \frac{1}{E(\tau)}$

Arrival times of Customers



4. The probability of a lost or blocked packet is
   $P_B = \frac{\bar{L}}{\lambda}$
5. The throughput is the average number of customers per second that depart the system
   $$\begin{aligned} T &= \lim_{t\to\infty} \frac{\beta(t)}{t} \\ &= \lim_{t\to\infty} \left[\frac{\alpha(t)}{t} - \frac{L(t)}{t} \right] \\ &= \lambda - \bar{L} \\ &= \lambda - \lambda P_B \\ &= \lambda(1 - P_B) \end{aligned}$$
6. The average number of customers in the system is $\bar{N} = \lim_{t\to\infty} \frac{1}{t} \int_0^t N(t')dt'$

Definitions and Description of Queue



W - Waiting time

S - Service time

T - Total queueing delay $T = W + S$

λ - arrival rate (messages or customers per second)

$\frac{1}{\mu}$ - average message length (bits)

C - capacity - bits xmitted per second

$\bar{S} = \frac{1}{\mu C}$ - avg time to xmit a message

ρ - link utilization = $\lambda \bar{S} = \frac{\lambda}{\mu C}$

if $\rho > 1$ queue becomes infinite

N - Number of messages in the queue, including customer being served

1.2 Kendall's Notation

A/B/C

A - Arrival Distribution

M - Poisson arrival Process

D - Deterministic - Voice Packets

GI - General, Independent

B - Service time distribution

M - exponential

D - fixed - ATM cells - be careful

G - general

C - # of servers

Extended Notation

(A/B/C):(D/E/F)

D - Service Process

FCFS - first come, first served

LIFO - last in, first out

SIRO - serve in random order

GD - general

E - max allowable customers in the queue

Buffer size

F - Number of customers

as opposed to an infinte user population

Example

(M/M/3):(FCFS/100/∞)

2. Little's Law

Reference [1], section 3.2

Average operation of the system in *steady-state*
steady state -> system isn't growing in an unbounded way

$N(t)$ = Number of customers in syst at time t - *Instantaneous*

Average customers in the syst $N_t = \frac{1}{t}\int_0^t N(\tau)d\tau$

steady state $N = \lim_{t\to\infty} N_t$

$\alpha(t)$ = Number of customers who arrive from 0 to t - *Cumulative*

Average arrival rate $\lambda_t = \frac{\alpha(t)}{t}$

steady state $\lambda = \lim_{t\to\infty} \lambda_t$

T_i = Time spent in the system by the i^{th} customer - *Instantaneous*

The avg time in the syst. is $T_t = \frac{\sum_{i=0}^{\alpha(t)} T_i}{\alpha(t)}$

At time t:
$\alpha(t)$ is the number of customers that have arrived
$\sum_{i=0}^{\alpha(t)} T_i$ is the sum of the times that they spend in the system.

steady state $T = \lim_{t\to\infty} T_t$

Little's Law $N = \lambda T$

The power of Little's law is that it holds for all distributions of arrivals and service times

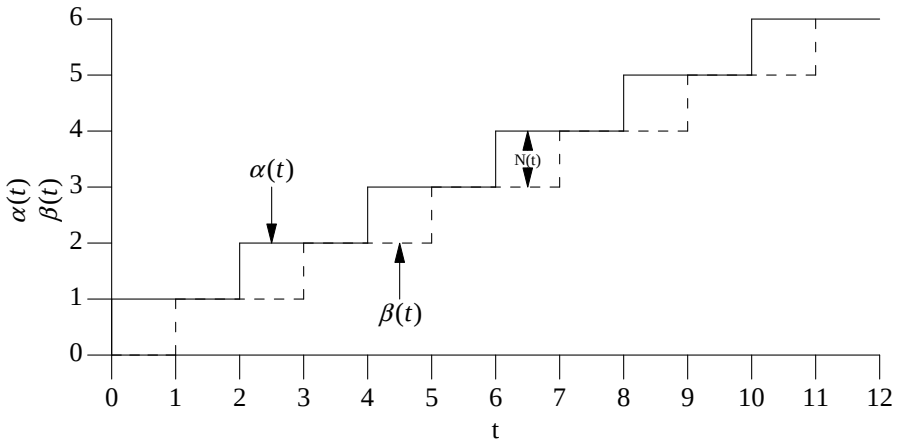
We can apply Little's Law to the entire system or parts of the system

2.1 Simple Examples

2.1.1 Example 1

if there is an arrival every 2 seconds - arrival rate λ= 1/2

and each user spends 1 second in the system T = 1

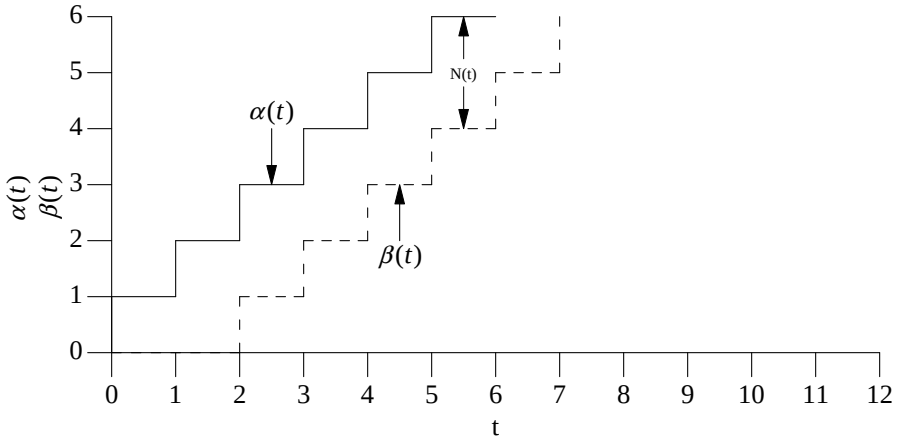


the average number of users in the system is N = 1/2 * 1 = 1/2

2.1.2 Example 2

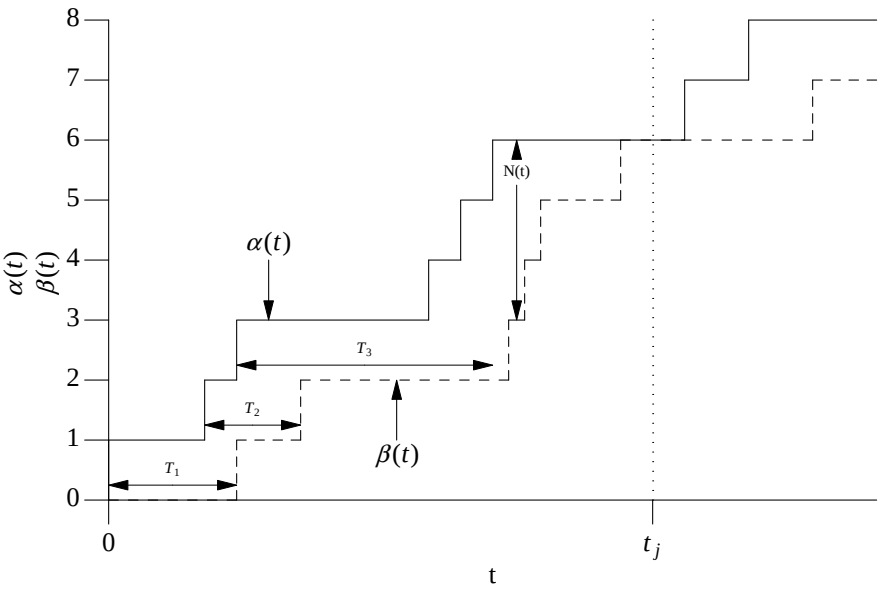
if there is an arrival every second - arrival rate λ = 1

and each user spends 2 seconds in the system T = 2



the average number of users in the system is N = 1 * 2 = 2

2.2 Graphical Proof of Little's Law



Assumptions:

- The system is empty at time t=0
- At a future time t_j, the system is also empty, $\alpha(t_j) = \beta_{t_j}$

  This is true infinitely often when $\lambda T < 1$

  If $\lambda T > 1$, the system is unstable

  This assumption is removed as $t \to \infty$
- The customers depart in the order that they arrive

The number in the syst. is $N(t) = \alpha(t) - \beta(t)$

the area between $\alpha(t)$ and $\beta(t)$ (up to t_j) is $A(t_j) = \int_0^{t_j} N(\tau)d\tau$

$A(t_j)$ is also $\sum_{i=1}^{\alpha(t_j)} T_i$

$$A(t_j) = \sum_{i=1}^{\alpha(t_j)} T_i = \int_0^{t_j} N(\tau)d\tau$$

Dividing by t_j, then multiply the left hand side by $\frac{\alpha(t_j)}{\alpha(t_j)}$

$$\frac{1}{t_j}\int_0^{t_j} N(\tau)d\tau = \frac{1}{t_j}\sum_{i=1}^{\alpha(t_j)} T_i = \frac{\alpha(t_j)}{t_j}\,\frac{\sum_{i=1}^{\alpha(t_j)} T_i}{\alpha(t_j)}$$

Take the average values up to time t_j

$$N(t_j) = \frac{1}{t_j}\int_0^{t_j} N(\tau)d\tau$$

$$\lambda(t_j) = \frac{\alpha(t_j)}{t_j}$$

$$T(t_j) = \frac{\sum_{i=1}^{\alpha(t_j)} T_i}{\alpha(t_j)}$$

$$N(t_j) = \lambda(t_j)T(t_j)$$

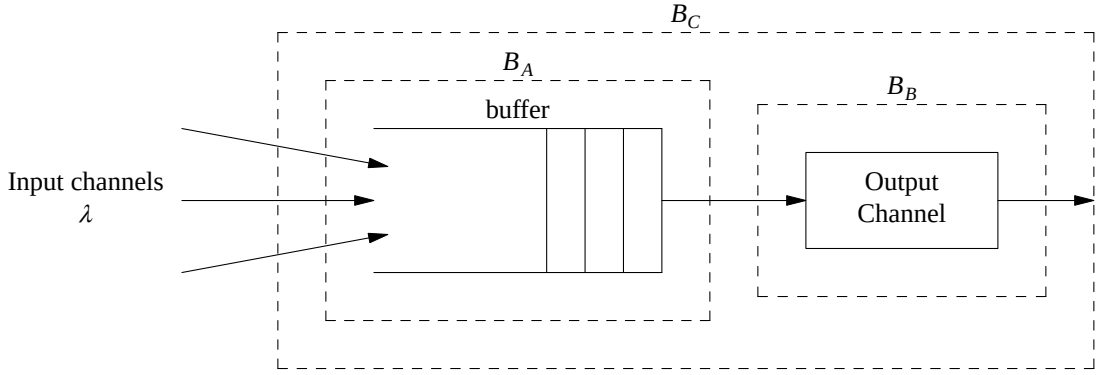
And $\lim_{t\to\infty}\left(N(t_j) = \lambda(t_j)T(t_j)\right) \to (N = \lambda T)$

In reference [1], fig 3.2 it is shown that the customers don't have to depart in the order they arrive. Therefore Little's law can be applied to LIFO, SIRO and Priority systems, as well as FIFO systems

2.3 Applications of Little's law

2.3.1 Single queue

Determine the relationship between the parameters that specify a queue



A. Draw box around queue, but not xmission line

λ = arrival rate from xmission lines

N_Q = avg # of packets in queue, not being xmitted.

W = avg waiting time in the queue *(The time before we start to transmit the packet)*

$$N_Q = \lambda W \Rightarrow W = N_Q/\lambda$$

B. Draw box around xmission line

$\bar{X}$ = avg packet xmission time = $\frac{1}{\mu C}$

From Little's Law: $N_C = \lambda\bar{X} = \frac{\lambda}{\mu C}$

Since the channel xmits 0 or 1 message.

$N_C = \rho$

The average number of customers on the channel is the fraction of the time that the channel is busy, ρ.

Therefore, the channel utilization is: $\rho = \frac{\lambda}{\mu C}$

C. Draw the box around the entire system

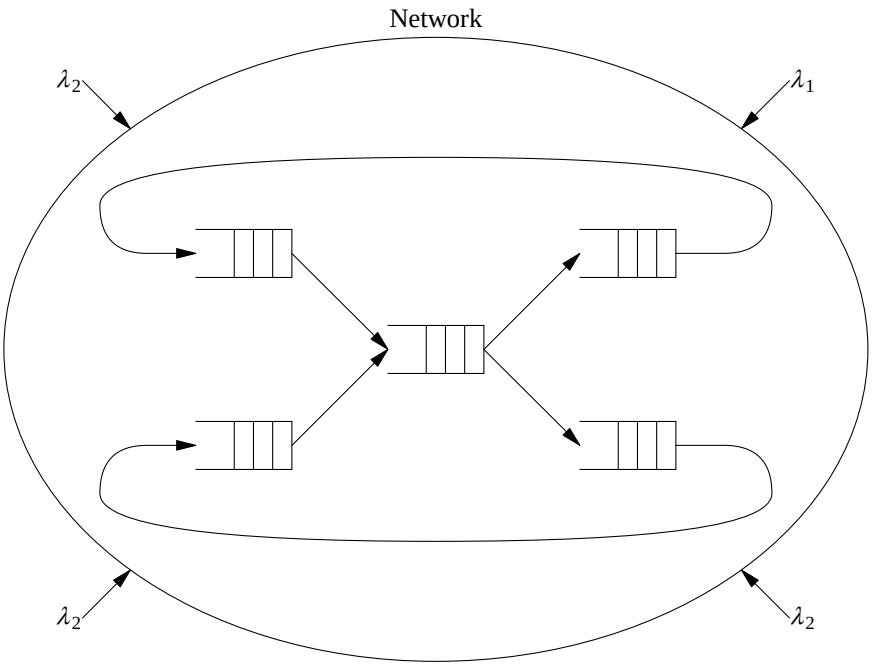
T = total delay = $W + \bar{X}$

N = total number in system = $N_Q + \rho$

$N = \lambda T = \lambda(W + \bar{X})$

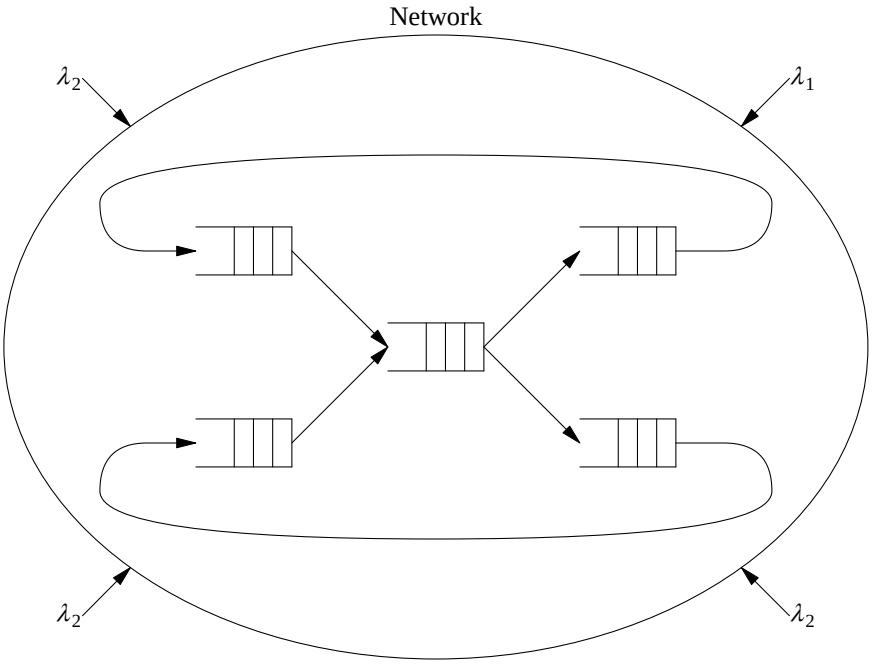
$$T = \frac{N}{\lambda} = \frac{N_Q + \rho}{\lambda}$$

2.3.2 Network Delay



$T = \frac{N}{\sum_{i=1}^{n} \lambda_i}$ independent of packet length distribution or routing

2.3.3 Networks of Queues



1. Apply Little's law to the entire network
   $\bar{T}_{net} = \bar{N}_{net}/\lambda_{net}$
2. Apply Little's Law to each queue
   $\bar{N}_i = \lambda_i \bar{T}_i$
   The λ_i depends on the routing
3. The average number of messges in the network is sum of the average numbers of messages in the queues
   $$\bar{N}_{net} = \sum_i \bar{N}_i = \sum_i \lambda_i \bar{T}_i$$
4. The average delay in the network can be found from the average delays in queues
   $$\bar{T}_{net} = \frac{\sum_i \lambda_i \bar{T}_i}{\lambda_{net}} = \sum_i \frac{\lambda_i}{\lambda_{net}} \bar{T}_i$$
   The delay in each queue is weighted by the proportion of the flow through the queue

This result will be used with Kleinrock's independence assumption.
We will calculate the delays in the individual queues, then use this result to find the average network delay

2.3.4 The use of windows for flow control - TCP

The window constrains the number of packet from a source to W

W is the maximum number of packets that the source can have in the network

Therefore, $\lambda T \leq W$ - Little's Law
Where T is the delay that the packet experiences
Assumes that the ACK delays are small

When the network is congested, T increases, and eventually, λ must decrease

Note: If the network is congested, and can only provide a throughput of λ for each user, $W \approx \lambda T$.

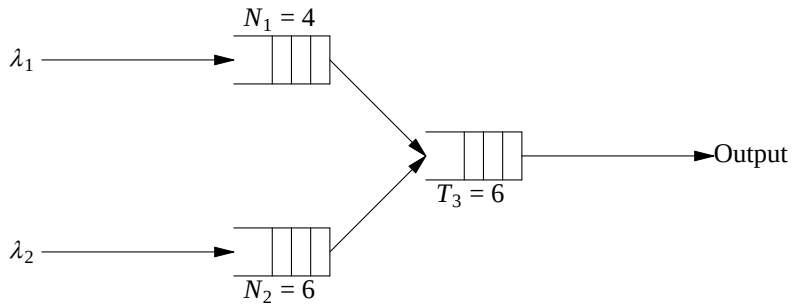
If each user increases their window size, they increase their delays.

Therefore, to decrease the delay in a congested network, each user should decrease their window size

In TCP the window size is decreased when an acknowledgement is received later than expected.

Home work

Homework: Find the average delay T_{net} for the following network



3. Analysis of M/M/1 Queues

Reference [1], section 3.3

The arrival of messages to the queue is described by a Poisson process with an average arrival rate λ.

The time it takes to service a customer in the queue is exponentially distributed with an average time $\frac{1}{\mu C}$. $\frac{1}{\mu}$ is the average message length, and C is the transmission rate of the channel.

Kleinrock's approximation is an approximate analysis of network's of M/M/1 queues and has been used to predict the performance of the ARPAnet/Internet.

3.1 Some Characterisitics of the Poisson Process and the Exponential distribution

3.1.1 Poisson Arrival Process

The probability of n arrivals in any interval τ is

$P_n(\tau) = e^{-\lambda\tau}\frac{(\lambda\tau)^n}{n!}$, for n=0,1,...

3.1.1.1 The interarrival times between messages are exponentially distributed

1. The probability distribution function of the n^{th} interarrival is:
   $P(\tau_n \le s) = 1 - P_0(s) = 1 - e^{-\lambda s}$, for $s \ge 0$,
   Where:
   t_n is the time of the n^{th} arrival, and
   $\tau_n = t_n - t_{n-1}$
   The probability distribution is independent of n, the message number.
2. The probability density function of interarrivals is: $p(\tau) = \lambda e^{-\lambda\tau}$
   $E(\tau) = 1/\lambda$, $Var(\tau) = 1/\lambda^2$
   The arrival rate is λ messages/second

Note: Both the service time and interarrival times in an M/M/1 queue are exponentially distributed

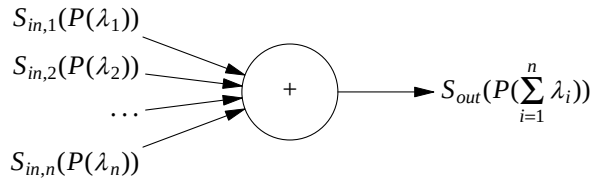
3.1.1.2 There are 3 approximations of the arrival process for small intervals, δ, that we need to analyze an M/M/1 queue:

The properties are found by taking the Taylor series expansion of $P_N(\tau)$

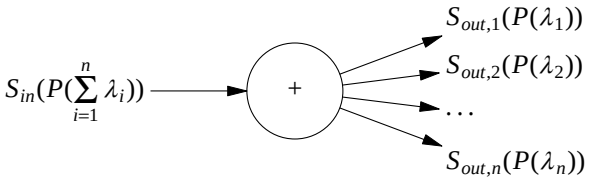
1. The probability of no arrivals in δ is:
   $\lim_{\delta\to 0} P_0(\delta) = \lim_{\delta\to 0} e^{-\lambda\delta} = 1 - \lambda\delta + o(\delta) \approx 1 - \lambda\delta$
2. The probability of 1 arrival in δ is:
   $\lim_{\delta\to 0} P_1(\delta) = \lim_{\delta\to 0} \lambda\delta e^{-\lambda\delta} = \lambda\delta + o(\delta) \approx \lambda\delta$
3. And the probability of $n > 1$ arrivals is:
   $\lim_{\delta\to 0} P_n(\delta) = \frac{(\lambda\delta)^n}{n!} e^{-\lambda\delta} = o(\delta) \approx 0$ for $n \ge 2$

3.1.1.3 There are 2 properties of the Poisson process that we will use in Kleinrock's approximate analysis of networks of queues

1. The sum of Poisson processes is a Poisson process with the output rate equal to the sum of the input rates



2. When a Poisson process is split by independently assigning an input to an output, each process is a Poisson process with the sum of the output rates equal to the input rate.



This is not true when the splitting is deterministic - ie: arrivals are sequentially assigned to outputs

3.1.2 The exponential service time distribution

$P(s_n \le s) = 1 - e^{-\mu C s}$, $s \ge 0$

Where s_n = service time of n^{th} customer

$$E(s_n) = \frac{1}{\mu C},\ Var(s_n) = \frac{1}{(\mu C)^2}$$

3.1.2.1 The exponential distribution is memoryless:

$$P(x > r + t \ / \ x > t) = \frac{P(x > r + t)}{P(x > t)} = \frac{e^{-\lambda(r+t)}}{e^{-\lambda t}} = e^{-\lambda r} = P(x > r)$$

1. The completion of the service process is independent of when the service process started
   $P(s_n > r + t \ / \ s_n > t) = P(s_n > r)$
2. Since the distribution of interarrival times is also exponential, the next arrival is independent of the number of customers in the system or when the last arrival occurred
   $P(\tau_n > r + t \ / \ \tau_n > t) = P(\tau_n > r)$
3. The memoryless property allows us to use a Markov chain formulation of a queue in which the next state is only dependent on the current state and does not depend on when the current state was entered.

3.1.2.2 There is an approximation of the service process for small intervals, δ, that we need to analyze an M/M/1 queue:

The probability that the service completes in δ is: $\lim_{\delta \to 0} P(s_n \le \delta) = \lim_{\delta \to 0}\left(1 - e^{\mu C s}\right) = \mu C\delta + o(\delta) \approx \mu C\delta$

3.2 Discrete Markov chain formulation of an M/M/1 queue

Partition time into small steps $0, \delta, 2\delta, \cdots, k\delta, \cdots$

N_k = number of customers in the queue at time $k\delta$

$P_{ij} = P(N_{k+1} = j \ / \ N_k = i)$

When $\delta \to 0$, there is only one departure or arrival in δ

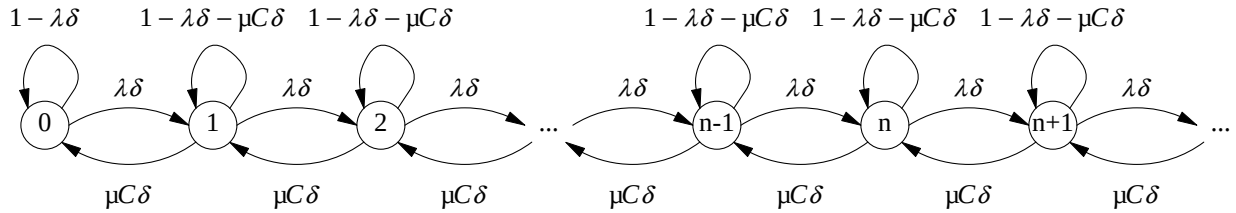
$P_{00} = 1 - \lambda\delta$ -- no arrival

$P_{ii} = 1 - \lambda\delta - \mu C\delta$ for $i > 0$ - no arrival or departure during δ

$P_{i,i+1} = \lambda\delta$, for $i \geq 0$ -- one arrival

$P_{i,i-1} = \mu C\delta$, for $i \geq 1$ -- one departure

$P_{i,j} = 0$ for $j \neq i, i+1, i-1$



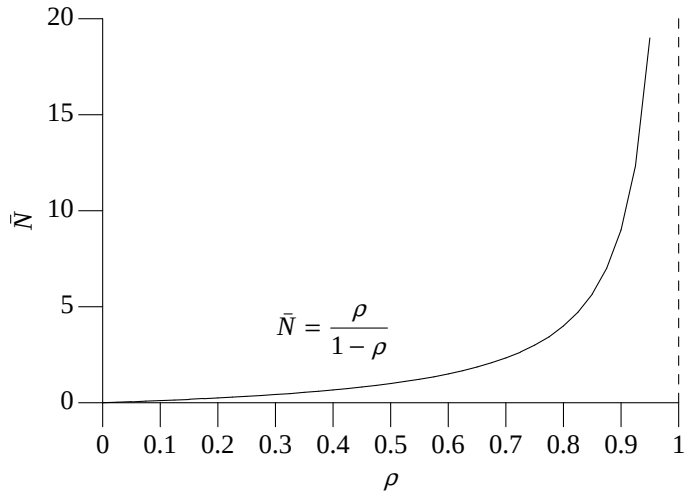
1. Derivation of the stationary distribution of p_i, the probability that there are i customers in the queue.

   - $P_{i,i+1} = \lambda\delta$
   - $P_{i+1,i} = \mu C\delta$
   - In equilibrium $p_i P_{i,i+1} = p_{i+1} P_{i+1,i}$ otherwise the state of the system is either increasing or decreasing. (the probability of moving up to a higher state equals the probability of moving down from that higher state). Therefore,
     $p_i \lambda\delta = p_{i+1} \mu C\delta$
   - $p_{i+1} = \rho p_i$
     $\rho = \dfrac{\lambda}{\mu C}$, the utilization of the output line from Little's law
   - $p_{i+1} = \rho^{i+1} p_0$, by iteration.
   - $1 = \sum_{i=0}^{\infty} p_i = \sum_{i=0}^{\infty} \rho^i p_0 = \dfrac{p_0}{1-\rho}$
   - $p_0 = 1 - \rho$
   - $p_i = \rho^i (1-\rho)$, i=0,1,...

2. Calculate $\bar{N}$, the average number of customers in the queue.

$$\bar{N} = \sum_{i=0}^{\infty} i p_i = \sum_{i=0}^{\infty} i \rho^i (1-\rho)$$
$$= \rho(1-\rho) \sum_{i=0}^{\infty} i \rho^{i-1} = \rho(1-\rho) \frac{d}{d\rho} \sum_{i=0}^{\infty} \rho^i$$
$$= \rho(1-\rho) \frac{d}{d\rho} \left(\frac{1}{1-\rho} \right) = \rho(1-\rho) \frac{1}{(1-\rho)^2}$$
$$= \frac{\rho}{1-\rho} = \frac{\lambda}{\mu C - \lambda}$$

Note that $\bar{N} \to \infty$ as $\rho \to 1$



3. From Little's Law

- The average delay is:
  $T = \frac{N}{\lambda} = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu C(1-\rho)}$
- The average waiting time is:
  $W = \frac{1}{\mu C(1-\rho)} - \frac{1}{\mu C} = \frac{\rho}{\mu C(1-\rho)}$
- The average number waiting in the queue is:
  $N_Q = \lambda W = \frac{\rho^2}{1-\rho}$.

3.3 Examples

3.3.1 Dispersity routing

K times as many messages that are $\frac{1}{k}^{th}$ as long

Increase the arrival rate from λ to $\lambda' = K\lambda$

Decrease the average message size from $\frac{1}{\mu}$ to $\frac{1}{\mu'} = \frac{1}{K\mu}$

$\rho = \frac{\lambda}{\mu C}$, $\rho' = \frac{K\lambda}{K\mu C}$, therefore $\rho' = \rho$ remains the same.

$N = \frac{\rho}{1-\rho} = N'$ remains the same

$T = \frac{N}{\lambda}$, $T' = \frac{N'}{\lambda'} = \frac{N}{K\lambda} = \frac{T}{K}$, is reduced by K

3.3.2 Dividing a channel into smaller pieces increases the delay

Given arrival rate λ and service time $1/\mu C$

$$T = \frac{1}{\mu C(1-\rho)}$$

Partition arrivals into M channels with arrival rate $\lambda' = \frac{\lambda}{M}$ and channel capacity $C' = \frac{C}{M}$.

The utilization remains the same: $\rho' = \frac{\lambda/M}{\mu C/M} = \frac{\lambda}{\mu C} = \rho$

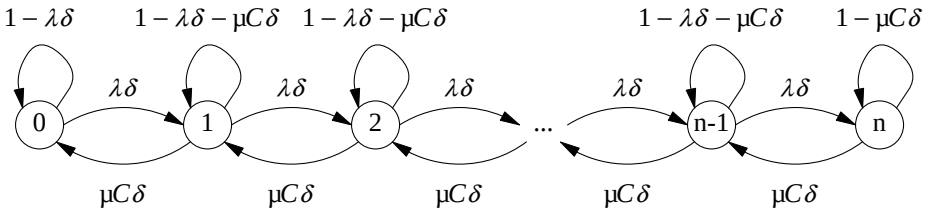
The service time increases to $\frac{M}{\mu C}$

$$T' = \frac{M}{\mu C(1-\rho)} = MT$$

This is an argument for statistical multiplexing rather than TDM for Poisson traffic

3.4 Finite Buffers

— (M/M/1):(GD/N/∞)



— Steady State Equations

$p_i(n)$ = probability of being in state i at the n^{th} step.
In steady state, $p_i(n) = p_i$

$p_0(n) = (1-\lambda\delta)p_0(n-1) + \mu C\delta\, p_1(n-1)$.
In steady state: $\rho p_0 = p_1$

Similarly, $(1+\rho)p_i = p_{i+1} + \rho p_{i-1}$ for $0 < i < N$ and, $p_N = \rho p_{N-1}$

Solving this set of equations: : $p_0 = \frac{1-\rho}{1-\rho^{N+1}}$

— Blocking probability - The probability that all of the buffers are full when a message arrives. The message is discarded.

input rate: $\lambda(1-P_B)$

output rate: $(1-p_0)\mu C$ *In p_0 queue is empty, channel is unoccuppied*

input rate = output rate

$$P_B = \frac{(1-\rho)\rho^N}{1-\rho^{N+1}}$$

3.5 What about other Service disciplines? LIFO, SIRO, Priorities

1. The average delay, T, is the same for all service disciplines

   The state model looks at the probability that any user is serviced - not necessarily the first to arrive.

   p_n is the same for any service discipline

$\bar{N}$ remains the same

Therefore, from Little's law, T remains the same

2. The average delay, T, is the same for all *Work Conserving* priority disciplines.

3.6 M/G/1 Queues

Pollaczek-Khintchine formula

Reference [1] section 3.5

M/G/1 Queue

$$W = \frac{\lambda \overline{X^2}}{2(1-\rho)}$$

Where $\overline{X^2} = E(X^2)$ - the second moment of the service time dist.

$$T_P = W + \bar{X} = W + \frac{1}{\mu C}$$

Apply the P-K formula to the exponential service time distribution

$$\overline{X^2} = \mu C \int_0^\infty x^2 e^{-\mu C x} dx = \frac{2}{(\mu C)^2}$$

$$W = \frac{2\lambda/(\mu C)^2}{2(1-\rho)} = \frac{\rho}{\mu C(1-\rho)},\ T = W + \frac{1}{\mu C} = \frac{1}{\mu C(1-\rho)}.$$

Home work

1. The Department of Motor Vehicles (DMV) has an office with 2 rooms. 14 customers per hour arrive at the DMV (with a Poisson arrival process).

   In the first office a clerk processes applications for drivers licenses. The clerk can process an average of 24 customers per hour. The processing time is an exponential distribution.

   The clerk in the first office rejects the applications from 1/4 of the customers that arrive at the desk. Half of the customers with rejected applications give up and go home. The other half take an average of 15 minutes to correct the application, and return to the end of the line in the first office. *(Assume that the combined arrivals from new customers and repeat customers form a Poisson arrival process.)*

   In the second office, photographs are taken and the drivers licenses are issued. There are 2 waiting lines. The customers that enter this room are randomly directed to one of the two lines, and are not allowed to switch to the other line. *(The arrival process to these queues is Poission - Burke's lemma)* The service time in each of the queues is exponentially distributed. In the first line the average service time is 8 minutes and in the second line the average service time is 9 minutes.

   A. What is the arrival rate at each of the 3 queues in the DMV?

   B. What is the average number of customers in the DMV?

   C. What is the average amount of time that an arriving customer spends in the DMV?

2. Assume that half the messages are fixed size 4000 byte packets and the other half are fixed size 20 byte acknowledgement messages. The transmission link transmits 10,000 bits/sec.

   a. Plot T vrs ρ using P-K formula

   b. Assume that the messages are from a single exponential distribution with the same mean service time, $(4000 + 20)/2$. Plot T vrs ρ on the same graph.

   Does the exponential distribution still give a worse case answer?

4. Kleinrock's Independence approximation

Applies to communications networks:

1. With nodes connected by transmission lines.
2. Each node has several communications lines arriving at and leaving the node.
3. Each communication line leaving a node has a queue where messages wait for service.
4. The arrivals into the network have Poisson distribution.
5. And, the messages have an exponential message length distribution.

Kleinrock showed that:
The average delays in this type of a network can be approximately calculated by assuming that the delays in the queues are independent.

The approximation was verified by simulations in Kleinrock's PhD dissertation at MIT in 1964.
He showed that the approximation is accurate as long as there are 4 or more communications links entering and leaving each node and the link utilizations are ≤ .8.

This approximation was used to analyze the delays, determine the routing, and design the topology of the original ARPAnet.

4.1 Correlation Between Queues

Consider 2 tandem M/D/1 queues

1. When the first queue has messages waiting, the departures from the first queue, and hence the arrivals at the second queue are evenly spaced and are not Poisson.
   Example: ATM networks
2. If the xmission line at the output of the second queue has the same xmission rate as the first queue, then there is NO waiting time in the second queue. The separation between arrivals at the second queue is ≥ to the time to service the message that preceded it.
   The service times in tandem queues are not independent.

Consider 2 tandem M/M/1 queues, with equal rate transmission lines.

— For a particular message, the service time in both queues is the same.

— If the second queue is empty, and the first queue has a short message, followed by a long message. The messages follow one another on the line between the queues. When the second message arrives at the second queue, the queue is empty and the service of the first message is complete. The service of the second message starts immediately.

— If a short message follows a long message in the first queue, when the second message arrives at the second queue, the server is still busy, and the second message must wait for service.

— The delay for messages in the second queue, depends on the service time of the messages in the first queue, so the queueing delays are **NOT** independent.

4.2 The mathematical basis for making Kleinrock's approximation

1. **Burke's Lemma** which proves that:
   The interdeparture time for messages from an M/M/1 queue is a Poisson process, with the same λ as the arrival process.
   This lemma is proven in 1.

2. In the analysis of **Jackson Networks** where it is proven that:
   The average delay in a network of M/M/1 queues can be exactly calculated by independently calculating the average delays in each of the queues, as long as the service times in successive queues are independent.
   The analysis of Jackson Networks follows directly from Burke's Lemma and the properties of Poisson processes.

   — The departures from a queue is a Poisson process, from Burke's Lemma.
   Therefore, the arrivals on the channel, at the next node, is a Poisson process.

   — When a channel reaches the next node, the arrivals are split and sent to several queues at the node.
   As long as the arrivals are split randomly, the arrivals at a queue from a single input channel is also a Poisson process.

   — The total arrivals at a queue come from several channels.
   The total arrivals is also a Poisson process, because the sum of Poisson processes is a Poisson process.

The difference between a Jackson Network and a communication network is that messages retain their lenght as they pass through a communication network. This results in correlated delays in successive queues, as demonstrated earlier.

Kleinrock's approximation is accurate in networks with multiple channels at each node because:

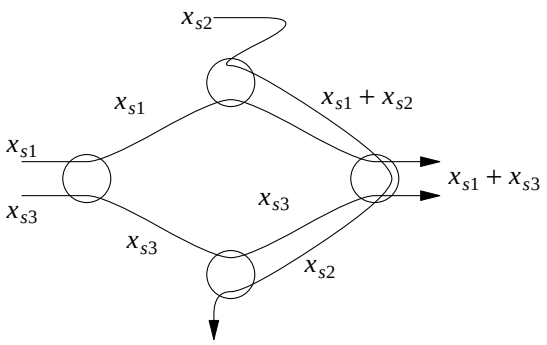
1. Messages arriving on each channel are split among several queues and the destination at the node, and
2. Messages arriving from several incoming channels and the source at the node are mixed together as they enter the queue.

The mixing reduces the liklihood that messages follow one another in successive queues.

The approximation is more accurate at low utilizations because the space between messages is also a Poisson process, and is independent of the message lengths.

4.3 Procedure for applying Kleinrock's independence approximation

1. Use the routing rule to determine the flows (arrival rates) on each link.



— The figure shows the flows in a portion of a communication network.

— The arrival rate on a link is the sum of the arrival rates from the sources that are routed over the link.
$\lambda_{i,j} = \sum_{p \in i,j} x_p$, is the arrival rate on a link, where x_p is the arrival rate from a source that is routed on link i,j
The arrival rate on a link is the arrival rate at the queue at the input to that link.

— $\gamma = \sum_{p} x_p$ is all of the flows traversing the network

2. Use the M/M/1 queueing analysis to calculate the average delay in the individual queues.

— The utilization is $\rho_{ij} = \frac{\lambda_{ij}}{\mu C}$

— The average number of messages waiting or being served at each queue is $N_{ij} = \frac{\rho_{ij}}{1-\rho_{ij}}$

— The average queueing delay is $T_{i,j} = \frac{N_{ij}}{\lambda_{ij}} = \frac{1}{\mu C(1-\rho_{ij})}$

3. Use Little's Law to find the average delay in the entire network

— The total number of messages in the network is $N = \sum_{ij} N_{ij} = \sum_{ij} \frac{\rho_{ij}}{1-\rho_{ij}}$

— The average network delay is $\bar{T} = \frac{N}{\gamma} = \frac{1}{\gamma} \sum_{ij} \frac{\rho_{ij}}{1-\rho_{ij}}$

4. If the propagation on the links is not negligible

— From Little's law, the average number of messages on a link is $N_{x,ij} = \lambda_{ij} d_{ij}$, where d_{ij} is the propagation delay on the link.

— The total number of messages in the system is $N' = \sum_{ij} \left[\frac{\rho_{ij}}{1-\rho_{ij}} + \lambda_{ij} d_{ij} \right]$

— The average network delay is: $\bar{T}' = \frac{1}{\gamma} \sum_{ij} \left[\frac{\rho_{ij}}{1-\rho_{ij}} + \lambda_{ij} d_{ij} \right]$

5. The average delay on a path from a source to a destination is the sum of the delays on the path

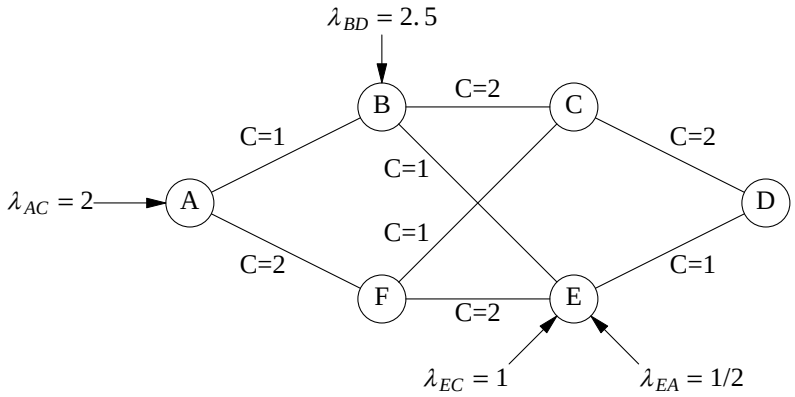
$T_p = \sum_{all\ i,j\ on\ the\ path} T_{ij} = \sum_{all\ i,j\ on\ the\ path} \left[\frac{1}{\mu C(1-\rho_{ij})} + d_{ij} \right]$

The result of Kleinrock's approximation is that the network delay calculation becomes an arithmetic procedure that we can use for future optimizations.

Home work

Kleinrock's Independence Approximation and Proportional Routing

Given a network,



The network has:

bidirectional communications channels, with the capacity specified in each direction, and an M/M/1 queue for each channel leaving a node.

The propagation delay on the links is negligible.

The arrival processes at the nodes are Poisson with the arrival rates specified.

λ_{XY} is the flow from node X to node Y.

The message lenght distributions are exponential with mean $\frac{1}{\mu} = \frac{1}{2}$

The paths that the flows follow to the their destinations are:

λ_{AC}:	1/2 of the packets on A->B->C
	1/2 of the packets on A->F->C
λ_{BD}:	2/5 of the packets on B->C->D
	3/5 of the packets on B->E->D
λ_{EC}:	E->B->C
λ_{EA}:	E->B->A

A. Find the utilization on each link with a non-zero flow

B. Find the average network delay

5. Priorities in Queues - Conservation Laws

5.1 Service Disciplines

A. *Non-preemptive:*

If a message is being transmitted when a higher priority message is being transmitted, the lower priority message completes transmission before the higher priority message is transmitted.

This is the discipline that is used in most communications network applications

B. *Preemptive, resume:*

If a low priority message is being transmitted when a higher priority message arrives, the low priority message stops transmitting. When the higher priority messages complete transmission, the lower priority message resumes transmission where it left off.

The extra overhead associated with this approach has discouraged its use in communications networks. It has many practical applications, including modeling people wating in line at a copy machine.

C. *Preemptive, restart:*

If a low priority message is being transmitted when a higher priority message arrives, the low priority message stops transmitting. When the higher priority messages complete transmission, the lower priority message must start over again from the beginning of its transmission.

Of the 3 disciplines, this is the only discipline that is *Not* work conserving. In the first 2 disciplines, the link utilization is the same for systems with priorities as it was for a system without priorities because the same number of bits per second are transmitted. In this system, more bits are transmitted and the utilization increases. In effect, the average message length for low priority messages that are interrupted is longer than the average message length of messages that are not transmitted.

In communications networks, this service discipline is used for extremely important events, like alarms.

5.2 M/M/1 Queues with 2 priorities, non-preemptive service discipline

- **Special Case**: Both low and high priority users have the same service time distribution.
- In the derivation of M/M/1 queues we showed that the average delay for work conserving priority service disciplines the delay, T, remains the same.
- Therefore, if we decrease T_H for the high priority users, we must increase T_L for the low priority users.
- In this section, we will derive the average delay for low and high priority users without taking advantage of the fact the the total average delay remains the same. Taking advantage of the total average delay can simplify the derivation, but this derivation will lead to the average delays in a general case.

1. The average delay is

$$T_P = \frac{\rho_H}{\rho} T_H + \frac{\rho_L}{\rho} T_L$$

2. Find T_H, the delay for high priority traffic

A. $W_H = R + E(S_H)$

- W_H = The average time that high priority users spend waiting in the queue
- R = expected time for a message, that is being served when the high priority message arrives, to complete service.
- $E(S_H)$ = average time to service the high priority messages that are in the queue when the high priority message arrives.

a. Find R

i. R = (Prob that a message is being served when a high priority message arrives) X (average time remaining to complete the service)

ii. ρ = Prob that a message is being served when a high priority message arrives

iii. $\frac{1}{\mu C}$ = The time remaining to service a service that is in progress, because of the memoryless property of the exponential distribution.

iv. $R = \rho \frac{1}{\mu C}$

b. Find $E(S_H)$

i. $E(S_H) = \frac{N_H}{\mu C}$

- N_H = number of high priority messages waiting when the message arrives
- $\frac{1}{\mu C}$ = average time to service each of these messages

ii. $N_H = \lambda_H W_H$, by Little's Law.

iii. $E(S_H) = \frac{\lambda_H}{\mu C} W_H = \rho_H W_H$

B. $W_H = \frac{\rho}{\mu C} + \rho_H W_H = \frac{\rho}{\mu C(1-\rho_H)}$

C. $T_H = \frac{1}{\mu C} + W_H = \frac{1}{\mu C}\left[1 + \frac{\rho}{1-\rho_H}\right]$

3. Find T_L, the delay for the low priority traffic

A. $W_L = R + E(S_H) + E(S_L) + E(S_H')$

- W_L = the expected waiting time for a low priority message.
- R and $E(S_H)$ are defined as before.
- $E(S_L)$ = average time to service the low priority messages that are in the queue when the message arrives.
- $E(S_H')$ = average time to service the additional high priority messages that arrive while the low priority message is waiting for service.

a. $R = \rho \frac{1}{\mu C}$, as before.

b. $E(S_H) = \rho_H W_H$, as before.

c. $E(S_L) = \rho_L W_L$, similar to the derivation for high priority traffic.

d. Find $E(S_H')$

i. $E(S_H') = \frac{N_H'}{\mu C}$

- N_H' = number of high priority messages that arrive while the low priority message is wating for service
- $\frac{1}{\mu C}$ = average time to service each of those messages.

ii. $N_H' = \lambda_H W_L$

iii. $E(S_H') = \frac{\lambda_H W_L}{\mu C} = \rho_H W_L$

B.

$$W_L = \frac{\rho}{\mu C} + \rho_H W_H + \rho_L W_L + \rho_H W_L$$
$$= \frac{\frac{\rho}{\mu C} + \rho_H W_H}{1 - \rho_L - \rho_H}$$
$$= \frac{\rho}{\mu C} \frac{1}{(1 - \rho_L - \rho_H)(1 - \rho_H)} = \frac{\rho}{\mu C} \frac{1}{(1 - \rho)(1 - \rho_H)}$$

C. $T_L = \frac{1}{\mu C}\left[1 + \frac{\rho}{(1 - \rho)(1 - \rho_H)}\right]$

4. Find the relationship between T_L, T_H and $T_{M/M/1}$

- $T_{M/M/1} = \frac{1}{\mu C(1 - \rho)}$ = The delay in an M/M/1 queue without priorities.

A. $T_H = T_{M/M/1}\left[1 - \frac{\rho\rho_L}{1 - \rho_H}\right] \leq T_{M/M/1}$

The delay for high priority traffic decreases

B. $T_L = T_{M/M/1}\left[1 + \frac{\rho\rho_H}{1 - \rho_H}\right] \geq T_{M/M/1}$

The delay for low priority traffic increases.

C. $T_P = \frac{\rho_H}{\rho} T_H + \frac{\rho_L}{\rho} T_L = T$

The total delay is conserved.

5.3 Generalization of Conservation Laws to more than 2 priorities and M/G/1 queues

Objective: to show how priorities affect the service of other customers

Reference [1] pp 188-9, [2] pp. 63-5, [3] pp. 113-7.

System

- Class of users: $(\lambda_1, 1/\mu_1, \overline{X_1^2})$, $(\lambda_2, 1/\mu_2, \overline{X_2^2})$, $\cdots$, $(\lambda_p, 1/\mu_p, \overline{X_p^2})$
- Lower numbered classes are higher priorities.
- The total link utilization is $\rho = \sum_{i=1}^{p} \rho_i = \sum_{i=1}^{p} \frac{\lambda_i}{\mu_i C} < 1$.

The conservation law we will prove is that:

$$\sum_{j=1}^{p} \rho_j W_j = \begin{cases} \frac{\rho R}{1 - \rho} & \rho < 1 \\ \infty & \rho \geq 1 \end{cases}$$

1. $W_j = R + \sum_{k=1}^{j} E(S_k) + \sum_{k=1}^{j-1} E(S_k')$

- W_j = the expected waiting time for a customer in class j
- R = the time for the customer being served to complete service

- $\sum_{k=1}^{j} E(S_k)$ is the average time to service the customers already in the queue with priority *higher than or the same* as the current customer.
- $\sum_{k=1}^{j-1} E(S_k')$ is the average time to service customers with *higher priority*, who arrive while this customer is waiting.

Note that, except for the customer being served, a customer in class j is not affected by customers with priority > j

A. $E(S_k) = \dfrac{N_k}{\mu_k C}$, as before.

a. $N_k = \lambda_k W_k$, by Little's Law.

b. $E(S_k) = \rho_k W_k$

B. $E(S_k')$ = number of arrivals while waiting * service time for each of those arrivals $E(S_k') = \dfrac{\lambda_k W_j}{\mu_k C} = \rho_k W_j$, as before.

C. $W_j = R + \sum_{k=1}^{j} \rho_k W_k + W_j \sum_{k=1}^{j-1} \rho_k = \dfrac{R + \sum_{k=1}^{j-1} \rho_k W_k}{1 - \sum_{k=1}^{j} \rho_k}$

2. Find W_j by solving recursively:

$$W_1 = \frac{R}{1-\rho_1}$$

$$W_2 = \frac{R + \rho_1 W_1}{1-\rho_1-\rho_2} = \frac{R}{(1-\rho_1-\rho_2)(1-\rho_1)}$$

$$\cdots$$

$$W_j = \frac{R}{(1-\sum_{k=1}^{j} \rho_k)(1-\sum_{k=1}^{j-1} \rho_k)}$$

3. Find R

- For service time distributions, other than exponential, the remaining service time is not independent of the time in service.
- The residual service time R is found by integrating over the distribution of all service times.
- In reference [1], pg 188-9, this is shown to be:

$$R = \frac{1}{2}\lambda\overline{X^2} = \frac{1}{2}\sum_{1}^{p} \lambda_k \overline{X_k^2}$$

4. Find $\sum_{j=1}^{p} \rho_j W_j$ by recursively substituting the values for W_j

$$\rho_1 W_1 = \frac{\rho_1 R}{1-\rho_1}$$

$$\rho_1 W_1 + \rho_2 W_2 = \left(\frac{\rho_1}{1-\rho_1} + \frac{\rho_2}{(1-\rho_1-\rho_2)(1-\rho_1)}\right) R = \frac{\rho_1+\rho_2}{1-\rho_1-\rho_2} R$$

. . .

$$\sum_{j=1}^{p} \rho_j W_j = \frac{\sum_{j=1}^{p} \rho_j}{1 - \sum_{j=1}^{p} \rho_j} R = \frac{\rho}{1-\rho} R$$

5.4 Special cases

5.4.1 All of the priority classes have the same average message, $1/\mu_i = 1/\mu$

- The conservation law becomes $\sum_{j=1}^{p} \lambda_j W_j = \frac{\lambda R}{1-\rho}$
- From Little's law $\lambda_j W_j = N_j$.

  Therefore, $\sum_{j=1}^{p} N_j = \frac{\lambda R}{1-\rho}$
- $\sum_{j=1}^{p} N_j = N = \lambda W$

  Therefore, $W = \frac{R}{1-\rho} = constant$
- The average waiting time remains the same.

  If we decrease the delay of one group of users, we must increase the delay of another group of users by the same amount

  This is more general than the result in the earlier section because it applies to *ANY* service time distribution.

5.4.2 Two priority classes with different average message lengths, (λ_1, μ_1), (λ_2, μ_2)

- When the message length distributions are different, it is possible to reduce the average delay per customer.
- In general, the average delay is decreased when the customers with shorter service times are given priority

  Super markets - special counters for customers with fewer items

  Networks - higher priority to short ack messages

A. Find the delay T_P, with priorities:

a. $T_P = \frac{\lambda_1}{\lambda} T_1 + \frac{\lambda_2}{\lambda} T_2$, where $\lambda = \lambda_1 + \lambda_2$

b. $T_P = \frac{1}{C}\left(\frac{\lambda_1}{\lambda}\frac{1}{\mu_1} + \frac{\lambda_2}{\lambda}\frac{1}{\mu_2}\right) + \frac{\lambda_1}{\lambda} W_1 + \frac{\lambda_2}{\lambda} W_2$, since $T_i = W_i + \frac{1}{\mu_i C}$

c. $T_P = \frac{1}{\mu C} + \frac{\lambda_1}{\lambda} W_1 + \frac{\lambda_2}{\lambda} W_2$, since $\frac{1}{\mu} = \frac{\lambda_1}{\lambda}\frac{1}{\mu_1} + \frac{\lambda_2}{\lambda}\frac{1}{\mu_2}$

d. $W_1 = \frac{R}{1-\rho_1}$, $W_2 = \frac{R}{(1-\rho)(1-\rho_1)}$,

e. $R = \frac{1}{2}\sum_{1}^{2} \lambda_i \overline{X_i^2} = \frac{\lambda}{2}\sum_{1}^{2} \frac{\lambda_i}{\lambda} \overline{X_i^2} = \frac{\lambda}{2} \overline{X^2}$

f.

$$T_P = \frac{1}{\mu C} + \frac{\lambda_1}{2}\frac{\overline{X^2}}{1-\rho_1} + \frac{\lambda_2}{2}\frac{\overline{X^2}}{(1-\rho)(1-\rho_1)}$$
$$= \frac{1}{\mu C} + \frac{\overline{X^2}}{2(1-\rho)}\frac{\lambda_1(1-\rho)+\lambda_2}{(1-\rho_1)}$$

B. With no priorities:

$$T_{NP} = \frac{1}{\mu C} + \frac{\lambda\overline{X^2}}{2(1-\rho)}$$

C. The change in average delay is :

$$T_P - T_{NP} = \frac{\overline{X^2}}{2(1-\rho)}\left(\frac{\lambda_1(1-\rho)+\lambda_2}{(1-\rho_1)} - \lambda\right)$$
$$= \frac{\overline{X^2}}{2(1-\rho)}\left(\frac{\lambda_1 - \frac{\lambda_1^2}{\mu_1 C} - \frac{\lambda_1\lambda_2}{\mu_2 C} + \lambda_2 - \lambda_1 - \lambda_2 + \frac{\lambda_1^2}{\mu_1 C} + \frac{\lambda_2\lambda_1}{\mu_1 C}}{1-\rho_1}\right)$$
$$= \frac{\overline{X^2}}{2(1-\rho)}\frac{\lambda_1\lambda_2}{(1-\rho_1)C}\left(\frac{1}{\mu_1} - \frac{1}{\mu_2}\right)$$

D. Therefore $T_P < T_{NP}$ iff $\frac{1}{\mu_1} < \frac{1}{\mu_2}$

REFERENCES

[1] D. Bertsekas, R. G. Gallager, **Data Networks**, Prentice-Hall Inc, 1992.

[2] M. Schwartz, **Telecommunication Networks: Protocols, Modeling and Analysis,** Addison Wesley Publication, 1987.

[3] L. Kleinrock, **Queueing Systems -- Volume 2: Computer Applications,** John Wiley & Sons, 1976.