UNIVERSIDADE FEDERAL DO MARANHÃO - UFMA
CENTRO TECNOLÓGIO - CT
CURSO: CIÊNICIA DA COMPUTAÇÃO
DISCIPLINA : TEORIA DAS FILAS E SIMULAÇÃO

Teoria das Filas

Luciano Cajado Costa[1]

[1] Professor Substituto da Universidade Federal do Maranhão

1. TEORIA DAS FILAS

Todas as pessoas já passaram pelo aborrecimento de ter que esperar em filas. Nós esperamos em fila quando estamos num engarrafamento, quando estamos no supermercado aguardando para pagar nossas compras, nos bancos e em muitas outras situações.

As formações de filas ocorrem porque a procura pelo serviço é maior do que a capacidade do sistema de atender a esta procura. A razão pelo qual os gerentes dos estabelecimentos e o poder público não aumentam suas capacidades de atendimento podem ser resumidas basicamente por dois motivos : inviabilidade econômica e limitação de espaço.

Dessa forma, a Teoria das Filas tenta através de análises matemáticas detalhadas encontrar um ponto de equilíbrio que satisfaça o cliente e seja viável economicamente para o provedor do serviço.

1.1 Descrição do Problema de Filas

Um sistema de filas pode ser descrito como clientes chegando, esperando pelo serviço, se não forem atendidos imediatamente, e saindo do sistema após serem atendidos. O termo cliente é usado de maneira geral e não implica necessariamente num cliente humano, como por exemplo, um processo esperando para utilizar a CPU. A figura 1.1 mostra um processo de filas típico.

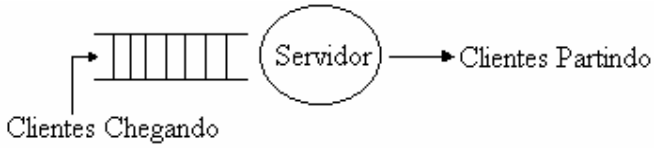

Figura 1.1 – Um processo de filas típico

A teoria das filas foi desenvolvida para prover modelos que retratem previamente o comportamento de um sistema que forneça serviços que possuam demandas que aumentem aleatoriamente

Existem muitas aplicações respeitáveis da teoria, a maioria das quais tem sido documentadas na literatura de probabilidade, pesquisa operacional e engenharia industrial. Alguns exemplos são fluxo de Tráfego (veículos, aeronaves, pessoas, comunicações), escalonamento (pacientes em hospitais, jobs em máquinas, programas

em computadores) e projetos de atendimentos à serviços (bancos, correios, parques de diversão, restaurantes fast-food).

1.2 Características dos Processos de Filas

Na maioria dos casos, seis características básicas de processos de filas fornecem uma descrição adequada de um sistema de filas: (1) padrão de chegada dos clientes, (2) padrão de serviço dos servidores, (3) disciplina de filas, (4) capacidade do sistema, (5) número de canais de serviço e (6) número de estágio de serviços.

1.2.1 Padrão de Chegada dos Clientes

Nos processos de filas comuns, os processos de chegadas são estocásticos, ou seja, desenvolvem-se no tempo e no espaço conforme leis de probabilidade. Assim, é necessário conhecer a distribuição de probabilidade descrevendo os tempos entre as sucessivas chegadas dos clientes (tempos de interchegada). Também é necessário saber se os clientes podem chegar simultaneamente (chegada batch), e se assim, qual a distribuição de probabilidade do tamanho do batch.

A reação do cliente ao entrar no sistema também é importante, de maneira que, um cliente pode decidir esperar sem problema, independente do tamanho da fila, ou, por outro lado, o cliente pode decidir não entrar no sistema caso a fila esteja muito grande. Assim, se o cliente decide não entrar na fila após a chegada, ele é conhecido como *decepcionado*. Um cliente pode, por sua vez, entrar na fila, mas depois de um tempo perder a paciência e decidir partir. Nos eventos que existem duas ou mais linhas paralelas, os clientes podem mudar de uma fila para outra. Estas situações são exemplos de filas com *clientes impacientes*.

Um fator final que pode ser considerado apesar do padrão de chegada é a maneira no qual o padrão muda com o tempo. Um padrão de chegada que não muda com o tempo (ou seja, que a distribuição de probabilidade descrevendo o processo de chegada é independente do tempo) é chamado padrão de chegada *estacionário*. Um que não é independente do tempo é chamado *não-estacionário*.

1.2.2 Padrões de Serviço

A maior parte da discussão mencionada nos padrões de chegada é valida para discussão dos padrões de serviço. A mais importante é que uma distribuição de

probabilidade é necessária para descrever a seqüência de tempos de serviços dos clientes. Os serviços também podem ser simples ou batch.

O processo de serviço pode depender do número de clientes esperando pelo serviço. Um servidor pode trabalhar mais rápido se a fila estiver aumentando, ou, caso contrário, pode ser tornar confuso é ficar mais lento. A situação na qual o serviço depende do número de clientes na fila é conhecida como serviço *dependente do estado*. Embora este termo não seja usado na discussão de padrões de chegada, o problema dos clientes impacientes podem ser considerados como chegadas dependentes do estado, desde que o comportamento da chegada depende da quantidade de congestionamento no sistema.

Serviços, como chegadas, podem ser estacionários ou não estacionários com respeito ao tempo. Por exemplo, o aprendizado pode ser considerado um fator de produtividade, de forma que, o serviço pode se tornar mais eficiente quando experiência é obtida, ou seja, não importa o número de clientes na fila (dependência do estado) e sim o período de tempo em atividade (dependência do tempo). Claro, que um sistema pode ser ao mesmo tempo não-estacionário e dependente do estado.

1.2.3 Disciplina de Filas

A disciplina de filas refere-se a maneira como os clientes são escolhidos para entrar em serviço após uma fila ser formada. A maioria das disciplinas comuns que podem ser observadas na vida diária é FCFS (First-Come-First-Served), ou seja, o primeiro a chegar é o primeiro a ser servido. Entretanto, existem outras disciplinas, tais como, LCFS(Last-Come-First-Served), aplicável em sistemas de controle de estoque onde o item mais recente é mais fácil de ser apanhado, e diversas outras disciplinas baseadas em esquemas de prioridade.

Existem duas situações gerais em disciplinas de prioridade. No primeiro caso, que é chamado de *preemptivo*, o cliente com a mais alta prioridade é permitido entrar em serviço independentemente de outro cliente com menor prioridade estar sendo servido, de forma que, o cliente com menor prioridade é interrompido e tem seu trabalho reiniciado mais tarde. Quando reiniciado, ele pode iniciar do ponto onde parou ou reiniciar todo o processo. Na segunda situação de prioridade, chamado caso *não-preemptivo*, os clientes com mais alta prioridade vão para o início da fila, mas só

entram em serviço quando o cliente sendo atendido deixa o sistema, mesmo que ele tenha uma prioridade baixa.

1.2.4 Capacidade do Sistema

Em alguns processos de filas existe uma limitação física da quantidade de espaço na fila, de modo que, se as filas alcançarem um certo comprimento, nenhum novo cliente poderá entrar no sistema até que espaço disponível seja obtido com o atendimento de um cliente e a conseqüente diminuição do tamanho da fila. Estas situações são referidas como sistemas de filas finitos, ou seja, existe um limite finito do tamanho máximo do sistema.

1.2.5 Número de Canais de Serviço

Quando o número de canais de serviço são definidos, tipicamente estão sendo determinados o número de estações de serviços paralelos que podem servir os clientes simultaneamente. A figura 1.1 ilustra um sistema com canal simples, enquanto a figura 1.2 mostra duas variações dos sistemas multicanais. Os dois sistemas multicanais diferem pelo fato que o primeiro possui uma única fila, enquanto o segundo possui uma fila para cada canal. Uma barbearia com várias cadeiras é um exemplo do primeiro tipo de multicanal, assumindo que não exista um estilo particular de corte de cabelo. Por outro lado, um supermercado e um restaurante fast-food preenche a segunda espécie de multicanal. É geralmente assumido que os mecanismos de canais paralelos operam independentemente um do outro.

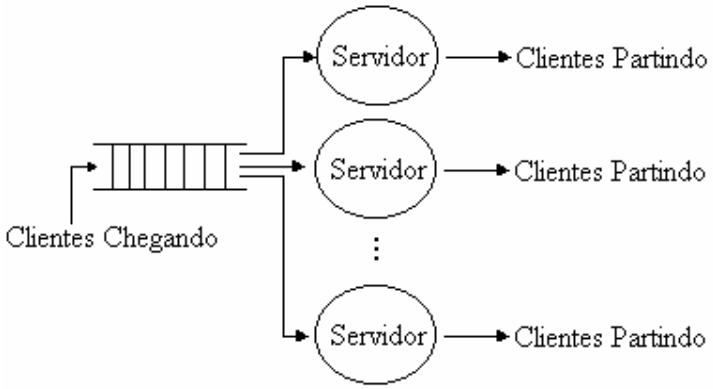

Figura 1.2(a) – Sistema Multicanal com Fila Única

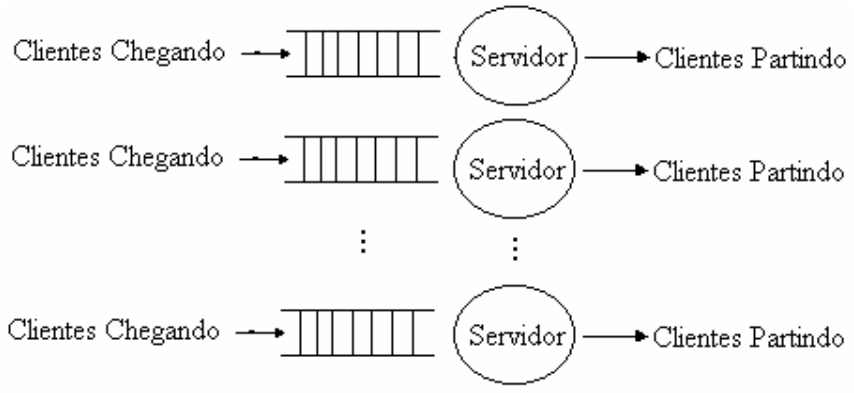

Figura 1.2(b) – Sistema Multicanal com Fila Individual.

1.2.6 Estágios de Serviço

Um sistema de filas pode ter um único estágio de serviço, como no caso da barbearia, ou pode ter vários estágios. Um sistema de multi-estágio pode ser exemplificado como um procedimento de exame físico, onde cada paciente passa por diversos exames, tais como: sangue, vista, urina e etc. Em alguns sistemas multi-estágio reciclagem ou retorno podem ocorrer. Reciclagem é comum em processos de manufatura, onde inspeções de controle de qualidade são realizadas sendo que se alguma peça não se adequa ele deve ser reprocessada. Um exemplo de sistema multi-estágio com retorno é mostrado na figura 1.3.

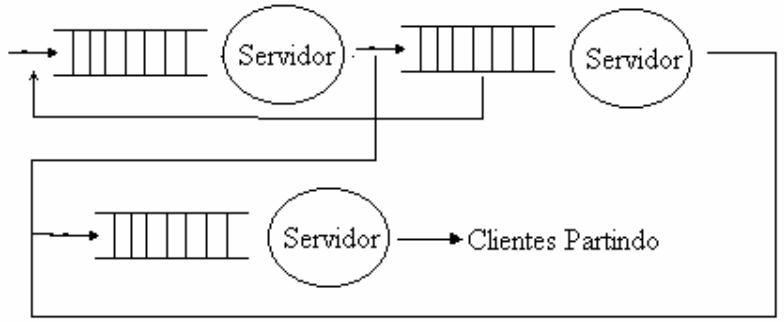

Figura 1.3 – Sistema de Filas Multi-Estágio com Retorno

As seis características de sistemas de filas discutidas nesta seção são suficientemente gerais para descrever um processo sob estudo. Assim, antes de realizar qualquer análise matemática, é necessário descrever adequadamente o processo sendo modelado, de forma que, o conhecimento das seis características básica são essenciais nesta tarefa.

1.3 Notação

A notação de processos de filas mais utilizada atualmente foi proposta por Kendall, em 1953, e é descrita por um série de símbolos, tais como, A/B/m/k/M, onde A indica a distribuição de interchegada dos clientes, B o padrão de serviço de acordo com uma distribuição de probabilidade para o tempo de serviço, m o número de canais de serviços paralelos (servidores), k a capacidade do sistema e M a disciplina de filas. Alguns símbolos padrões para estas características são mostradas na Tabela 1.1. Por exemplo, a notação M/D/2/∞/FCFS indica um processo de filas com tempos de interchegada exponenciais, tempos de serviço determinísticos, dois servidores paralelos, capacidade ilimitada e disciplina de fila First-Come-First-Served.

Características	Símbolo	Explicação
Distribuição de Tempo de Interchegada (A) e Distribuição de Tempo de Serviço (B)	M	Exponencial
	D	Determinístico
	E_k	Tipo k-Erlang (k = 1,2,...)
	H_k	Mistura de k exponenciais
	PH	Tipo Fase
	G	Geral
Número Paralelo de Servidores (m)	1,2,..., ∞	
Restrição na capacidade do sistema (k)	1,2,..., ∞	
Disciplina da fila (M)	FCFS	First Come First Served
	LCFS	Last Come First Served
	RSS	Seleção Aleatória por Serviço
	PR	Prioridade
	GD	Disciplina Geral

Tabela 1.1 – Notação de Fila – A/B/m/k/M

Em muitas situações só os três primeiros símbolos são utilizados, de maneira que, é assumido que o sistema tem capacidade ilimitada e possui uma disciplina FCFS. Neste caso, M/D/2/∞/FCFS poderia ser indicado apenas por M/D/2.

Os símbolos na tabela 1.1 são auto-explicativos, entretanto, alguns deles merecem algum complemento. Por exemplo, o símbolo G representa uma distribuição de probabilidade geral, isto é, resultados nestes casos são aplicáveis para qualquer distribuição de probabilidade.

Pode parecer estranho que o símbolo M seja usado como exponencial. O uso de E pode ser confundido com E_k que é usado para representar uma distribuição Erlang tipo k. Assim, M é usado ao invés disso, onde M é originado da propriedade sem memória ou Markoviana da distribuição exponencial.

1.4 Medindo a Performance do Sistema

Geralmente existem três tipos de respostas de interesse do sistema : (1) alguma medida de tempo de espera que um cliente típico pode ser forçado a encarar; (2) uma indicação da maneira pelo qual os clientes podem ir se acumulando; e (3) uma medida de tempo ocioso dos servidores. Desde que a maioria dos sistemas de filas tem elementos estocásticos, estas medidas freqüentemente são variáveis aleatórias e possuem distribuições de probabilidade, ou pelo menos valores esperados.

Existem dois tipos de tempo de espera de clientes, o tempo que o cliente gasta na fila e o tempo total do cliente no sistema (tempo na fila + tempo de serviço). A importância desses dois tipos de tempo de espera depende do estudo que está sendo realizado. Por exemplo, num parque de diversão, o tempo de espera na fila é que deixa o cliente infeliz. Por outro lado, no reparo de uma máquina é o tempo total no sistema que se deseja minimizar, de maneira a ter a máquina o mais rápido possível em produção.

Correspondentemente, existem duas medidas de acumulos de clientes : o número de clientes na fila e o número de clientes no sistema. Estas medidas são importantes na definição do tamanho do espaço reservado para os clientes esperarem. As medidas de ociosidade dos serviços podem incluir a porcentagem de tempo que um servidor particular está ocioso, ou o tempo que o sistema está desprovido de clientes.

A tarefa do analista de filas é determinar as medidas apropriadas de efetividade de um dado processo, ou projetar um sistema ótimo. No projeto de um sistema o analista pode querer balancear o tempo de espera dos clientes contra o tempo de ociosidade do servidor de acordo com alguma estrutura de custos. Se os custos de serviço ocioso e tempo de espera podem ser obtidos diretamente, eles podem ser usados para determinar o número ótimo de canais e as taxas de serviço nos quais esses canais devem operar. Também para projetar a sala de espera é necessário definir o tamanho da fila, que pode ser calculado pelo atraso do cliente na fila e pelo tempo de ociosidade do servidor. Em qualquer situação, o analista tentará resolver o problema através de métodos analíticos, caso isto falhe, o analista deverá proceder com uma simulação.

1.5 Alguns Resultados Gerais

Nesta seção serão apresentados alguns resultados gerais e relacionamentos para filas G/G/1 e G/G/c.

Denotando a taxa média de clientes chegando no sistema de filas como λ e a taxa média de serviço como μ, sendo que uma medida de congestionamento para um sistema c-server é $\rho \equiv \lambda/c\mu$. Quando $\rho > 1$ ($\lambda > c\mu$), o número médio de chegadas no sistema excede a taxa média de serviço do sistema, de forma que, quando o tempo avança a fila se tornará cada vez maior, impedindo qualquer situação de equilíbrio e fazendo com que novos clientes não possam mais se juntar a fila.

Assim, para obter uma situação de equilíbrio, ρ deve ser estritamente menor do que um. Quando $\rho = 1$, ao menos que as chegadas e os serviços sejam determinísticos e bem escalonados, nenhum estado de equilíbrio existe, desde que a aleatoriedade impedirá a fila de se esvaziar, de forma que, a mesma crescerá sem limites. Por essa razão, se a taxa média de serviço e a taxa média de chegada são conhecidas, o número médio de servidores paralelos requeridos para garantir a solução de estado de equilíbrio pode ser imediatamente calculado encontrando o menor c que satisfaça a equação $\lambda/c\mu < 1$.

O que é mais freqüentemente desejado na solução de modelos de filas é encontrar a distribuição de probabilidade para o número total de clientes no sistema no tempo t, N(t), que é obtido através do número de clientes na fila no tempo t, $N_q(t)$, mais aqueles em serviço, $N_s(t)$. Assim, seja $p_n(t) = Pr\{N(t) = n\}$, e $p_n = Pr\{N=n\}$ no estado de equilíbrio. Considerando filas c-servidor no estado de equilíbrio, duas medidas de valores esperados de maior interesse são o número médio no sistema,

$$L = E[N] = \sum_{n=0}^{\infty} n p_n,$$

e o número esperado na fila,

$$L_q = E[N_q] = \sum_{n=1}^{\infty} (n-c)\, p_n$$

1.5.1 Fórmulas de Little

Um dos mais poderosos relacionamentos em teoria das filas foi desenvolvido por John Little no início dos anos 60. Little relacionou o tamanho médio do sistema em estado de equilíbrio com o tempo médio de espera dos clientes, também em estado de equilíbrio, da seguinte forma. Uma vez que T_q é o tempo que um cliente gasta esperando na fila antes de entrar em serviço e T o tempo total que um cliente gasta

no sistema ($T = T_q + S$, onde S é o tempo de serviço, e T, T_q e S são variáveis aleatórias), duas medidas de p erformance do sistema freqüentemente utilizadas com respeito aos clientes são $W_q = E[T_q]$ e $W = E[T]$, ou seja, o tempo médio de espera na fila e o tempo médio de espera no sistema, respectivamente. Partindo desse princípio, as fórmulas de Little são

$$L = \lambda W \qquad (1.1a)$$

$$L_q = \lambda W_q \qquad (1.1b)$$

Assim, só é necessário encontrar uma dos quatro valores esperados, em vistas das fórmulas de Little e do fato de que $E[T] = E[T_q] + E[S]$, ou, equivalentemente, $W = W_q + 1/\mu$, onde μ, como antes, é a taxa média de serviço.

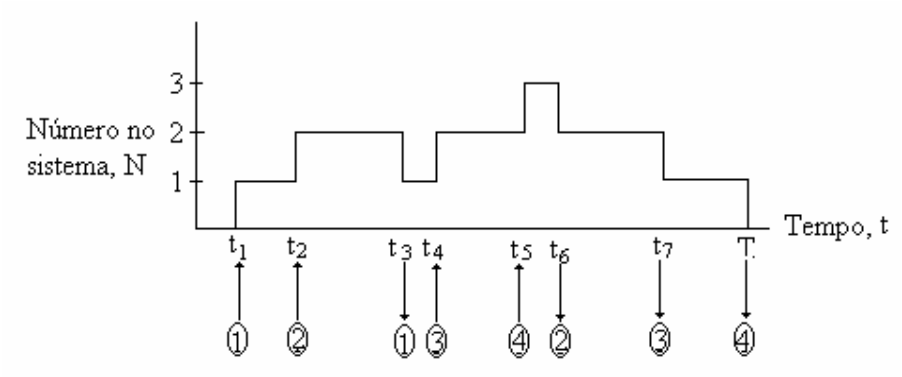

Figura 1.4 – Percurso do período de ocupação

O exemplo da figura 1.4 ilustra os conceitos das fórmulas de Little por considerar um intervalo de ocupação de um servidor. O número de clientes (N_c) que chega no período (0,T) é 4.

Os cálculos de L e W são

$$L = [1(t_2-t_1) + 2(t_3-t_2) + 1(t_4-t_3) + 2(t_5-t_4) + 3(t_6-t_5) + 2(t_7-t_6) + 1(T-t_7)]/T =$$

$$= (\text{área sob a curva})/ T$$

$$= (T + t_7 + t_6 - t_5 - t_4 + t_3 - t_2 - t_1)/T \qquad (1.2a)$$

e $$W = [(t_3-t_1) + (t_6-t_2) + (t_7-t_4) + (T- t_5)]/4 =$$

$$= (T + t_7 + t_6 - t_5 - t_4 + t_3 - t_2 - t_1)/T$$

$$= (\text{área sob a curva})/ Nc \qquad (1.2b)$$

Dessa forma, das equações 1.2a e 1.2b verifica-se que a área sob a curva é $LT = WNc$, que resulta $L = WNc/T$. A fração Nc/T é o número de clientes chegando no período T, ou seja, λ, de modo que, $L = \lambda W$.

Um resultado interessante que pode ser derivado das fórmulas de Little é que a relação entre W e W_q é

$$L - L_q = \lambda (W - W_q) = \lambda(1/\mu) = \lambda/\mu. \quad (1.3)$$

Mas, $L - L_q = E[N] - E[N_q] = E[N-N_q] = E[S]$, de modo que, o número de clientes em serviço no estado de equilíbrio é λ/μ, que será denotado por r. Note que num sistema de servidor simples $r = \rho$ e também segue que

$$L - L_q = \sum_{n=0}^{\infty} np_n - \sum_{n=1}^{\infty}(n-1)p_n = \sum_{n=1}^{\infty} p_n = 1 - p_0$$

Disto, pode-se derivar a probabilidade de um servidor estar ocupado em um sistema multiservidor no estado de equilíbrio, sendo esta probabilidade denotada por p_b. Desde que já foi mostrado que o número esperado presente no sistema em estado de equilíbrio é r. segue da simetria dos c servidores que o número esperado presente e um servidor é r/c. Então, por um simples argumento de valor esperado, pode-se mostrar que $p_b = \rho$, uma vez que,

$$r/c = \rho = 0(1- p_b) + 1p_b$$

Para uma fila com servidor simples (G/G/1), a probabilidade do sistema estar ocioso (N=0) é a mesma da probabilidade do servidor estar ocioso. Assim, $p_0 = 1- p_b$, neste caso, e $p_0 = 1- \rho = 1 - r = 1 - \lambda/\mu$. A quantidade $r = \lambda/\mu$, número esperado de clientes em serviço, tem uma outra conotação interessante. Na média cada cliente precisa de $1/\mu$ unidades de tempo de serviço e o número médio de clientes chegando no sistema é λ, de modo que, o produto $\lambda(1/\mu)$ é quantidade de trabalho chegando para o sistema por unidade de tempo. Dividindo isto pelo número de servidores c (que resulta ρ) obtém-se a quantidade média de trabalho chegando para cada servidor por unidade de tempo. A tabela 1.2 sumariza o resultado desta seção.

$\rho = \lambda/c\mu$	Intensidade de Tráfego; Taxa de Carga de Trabalho oferecida para um servidor
$L = \lambda W$	Fórmulas de Little
$L_q = \lambda W_q$	Fórmulas de Little
$W = W_q + 1/\mu$	Argumento de valor esperado
$p_b = \lambda/c\mu = \rho$	Probabilidade de ocupação de um servidor qualquer
$r = \lambda/\mu$	Número esperado de clientes em serviço; Taxa de carga de trabalho oferecida
$L = L_q + r$	Resultado combinado
$p_0 = 1 - \rho$	Probabilidade de um sistema G/G/1 estar vazio.

$L = L_q + (1 - p_0)$	Resultado combinado para G/G/1

Tabela 1.2 - Sumário dos Resultados Gerais para Filas G/G/c

1.6 Exemplo de um Modelo de Filas Orientado à Eventos

Nesta seção será mostrado como eventos aleatórios de chegadas e serviços completos interagem em um sistema com um único servidor para formar uma fila. Este exemplo será iniciado no tempo 0 com a chegada do primeiro cliente e então o sistema terá seu estado atualizado toda vez que uma chegada ou partida ocorrer.

Considere o caso elementar de uma taxa de chegada constante para um canal simples com taxa de serviço constante. Assuma que o sistema possua uma disciplina FCFS e que no tempo t = 0 não exista nenhum cliente esperando e o canal esteja vazio. Seja λ o número de chegadas por unidade de tempo, e $1/\lambda$ o tempo constante entre sucessivas chegadas. Similarmente, se μ é a taxa de serviço em termos de serviços completos por unidade de tempo quando o servidor está ocupado, então $1/\mu$ é o tempo de serviço constante. É desejável calcular o número de clientes no sistema num tempo arbitrário t, denotado por n(t), e o tempo que o n-ésimo cliente deve esperar na fila para obter serviço, denotado por $W_q^{(n)}$.

Supondo que tão logo um serviço seja completado um outro será iniciado, o número no sistema (incluindo clientes em serviço) no tempo t é determinado pela equação

$$n(t) = \{\text{número de chegadas no intervalo } (0,t]\} - \{\text{número de serviços completos no intervalo } (0,t]\} \quad (1.4)$$

Geralmente existem dois tempos de espera de interesse – o tempo gasto pelo n-ésimo cliente esperando pelo serviço (ou atraso na fila), que será denotado por $W_q^{(n)}$; e o tempo do n-ésimo cliente gasto no sistema, que será denotado por $W^{(n)}$.

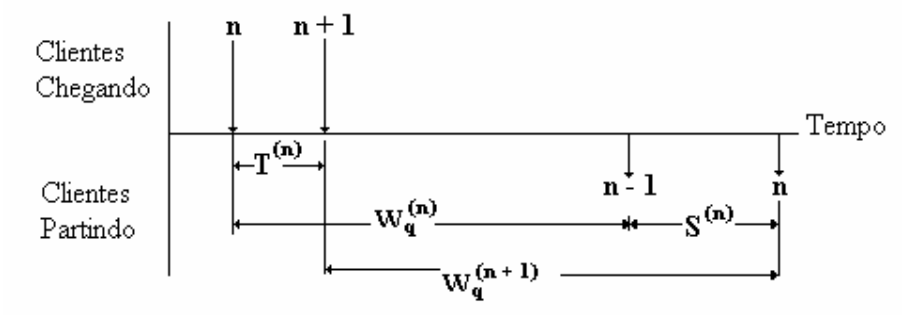

Figura 1.6 – Tempos de Esperas Sucessivas G/G/1

Para encontrar o tempo de espera na fila até que o serviço comece, observa-se que as filas de espera $W_q^{(n)}$ e $W_q^{(n+1)}$ de dois clientes sucessivos em qualquer fila com servidor simples são relacionados pela simples relação de recorrência

$$W_q^{(n+1)} = \begin{cases} W_q^{(n)} + S^{(n)} - T^{(n)} & (W_q^{(n)} + S^{(n)} - T^{(n)} > 0), \\ 0 & (W_q^{(n)} + S^{(n)} - T^{(n)} <= 0), \end{cases} \qquad (1.5)$$

onde $S^{(n)}$ é o tempo de serviço do n-ésimo cliente $T^{(n)}$ é o tempo de interchegada entre o n-ésimo cliente e o n-ésimo primeiro cliente. Isto pode ser visto por um simples diagrama mostrado na figura 1.6.

Tabela 1.3 Entrada de Dados

i.	1	2	3	4	5	6	7	8	9	10	11	12
Tempos de interchegada entre os clientes *i*+1 e *i*	2	1	3	1	1	4	2	5	1	4	2	-
Tempo de serviço do cliente *i*.	1	3	6	2	1	1	4	2	5	1	1	3

Neste exemplo o status do sistema é atualizado quando eventos ocorrem, registrando itens de interesse, e calculando medidas de efetividade. Modelos de filas orientados à eventos atualizam o estado do sistema só quando eventos (chegadas ou partidas) ocorrem. Desde de que não existe uma unidade de tempo básica para todos os eventos, o relógio mestre é incrementado por um quantidade de tempo variável, a invés de uma quantidade de tempo fixa, como seria em modelos de filas orientados por tempo. A abordagem orientada à evento será ilustrada aqui utilizando os dados de chegada e serviço mostrados na Tabela 1.3.

Tabela 1.4 – Modelo de Filas orientado à eventos.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Tempo do Relógio Mestre	Chegada/ Partida do Cliente i	Tempo que a chegada i entre em serviço	Tempo que a chegada i parte.	Tempo na Fila	Tempo no sistema	No. na Fila após o relógio mestre ser atualizado	No. no sistema após o relógio mestre ser atualizado
0	1-A	0	1	0	1	0	1
1	1-D					0	0
2	2-A	2	5	0	3	0	1
3	3-A	5	11	2	8	1	2
5	2-D					0	1
6	4-A	11	13	5	7	1	2
7	5-A	13	14	6	7	2	3
8	6-A	14	15	6	7	3	4
11	3-D					2	3
12	7-A	15	19	3	7	3	4
13	4-D					2	3

14	8-A;5-D	19	21	5	7	2	3
15	6-D					1	2
19	9-A;7-D	21	26	2	7	1	2
20	10-A	26	27	6	7	2	3
21	8-D					1	2
24	11-A	27	28	3	4	2	3
26	12-A;9-D	28	31	2	5	2	3
27	10-D					1	2
28	11-D					0	1
31	12-D					0	0

Tirando-se a média das colunas (5) e (6) na tabela 1.4 obtém-se que o atraso médio na fila dos 12 clientes foi 40/12 = 10/3, enquanto seu tempo de espera médio no sistema foi 70/12 = 35/6. Além disso, a taxa de chegada média é igual a 12/31 clientes por unidade de tempo, desde que existem 12 clientes no intervalo de 31 unidades de tempo observadas. Assim, aplicando-se a lei de Little para estes números tem-se que o tamanho médio do sistema L sob o tempo completo foi

$L = \lambda W = (70/12)/(31/12) = 70/31.$

O tamanho médio da fila pode calculado similarmente.

1.7 Processo de Poisson e Distribuição Exponencial

Os mais comuns modelos de filas estocásticos assumem que os tempos de interchegada e serviço obedecem uma distribuição exponencial ou, equivalentemente, que a taxa de chegada e a taxa de serviço seguem uma distribuição de Poisson. Desse modo, nesta seção será dado ênfase ao Processo de Poisson, bem como, sua relação com a distribuição exponencial. Contudo, inicialmente será considerado um processo de chegada geral.

1.7.1 Processo de Chegada

Assuma um sistema onde os clientes cheguem com intervalos de tempo aleatórios entre chegadas. Estas chegadas são eventos discretos assíncronos que dirigem o sistema, ou seja, o sistema só faz uma transição de um estado para outro no tempo que um evento ocorre. O primeiro cliente chega no tempo T1, o segundo no tempo T2, e assim por diante. Assume-se que os tempos de interchegada X1, X2, ... são variáveis aleatórias contínuas independentes com a mesma distribuição F_X (IID). Esta propriedade é referida como *estacionaridade do tempo*. Para manter a trilha do estado do sistema, isto é, quantos clientes chegaram, poderia-se por exemplo definir

S_n = Números de chegadas até o tempo T_n.

Então $\{S_n : n = 1,2,...\}$ é um processo estocástico de tempo discreto. Sob reflexão vê-se que S_n é igual a n para todo n, de modo que, este processo não é muito interessante, pois, ele apenas mantém-se incrementando por um a cada passo. Realmente, a informação interessante não é quantas chegadas tem existido, mas quando as chegadas ocorreram, implicando que um processo estocástico contínuo pode ser o modelo apropriado. Assim, define-se $N(t) = S_n$ para todo $T_n <= t < T_{n+1}$ como número de chegadas até o tempo t>0. Chama-se $\{N(t) : t>=0\}$ um processo de contagem de chegada, ou um processo de chegada, ou um processo de contagem.

1.7.2 O Processo de Poisson

Para um processo de chegada arbitrário o tempo da n-ésima chegada é

$$T_n = X_1 + X_2 + ... + X_n \qquad (1.6)$$

Quando $X_i \sim \exp(\lambda)$ é chamado um processo de Poisson. O parâmetro λ é chamado de a taxa do processo de Poisson. Para analisar este processo será explorado o fato que a soma das variáveis aleatórias exponenciais tem uma distribuição chamada distribuição de Erlang, isto é, T_n tem uma distribuição n-Erlang(λ). O fato importante sobre a distribuição de T_n é que ela completamente determina os caminhos de amostragem dos seus processos de chegada, Isto é facilmente visto por observar que o evento $\{ T_n <= t\}$ ocorre se e somente se o evento $\{N(t) >= n\}$ ocorre, isto é

$$\{ T_n <= t\} \Leftrightarrow \{N(t) >= n\}$$

Em outras palavras, o tempo da n-ésima chegada é menor que ou igual a t, se e somente se, tenha existido pelo menos n-chegadas até o tempo t. Usando FDC de Erlang pode-se portanto obter

$$P\{N(t) >= n\} = P\{ T_n <= t\} = 1 - \sum_{i=0}^{n-1} e^{-\lambda t}(\lambda t)^j/j!.$$

Aqui utilizou-se que para uma variável aleatória de Erlang X ~ n-Erlang(λ) a função de distribuição $F_X(x) = 1 - \sum_{i=0}^{\lfloor x \rfloor -1} e^{-\lambda t}(\lambda t)^j/j!$. Agora pode-se calcular a FDC de N(t):

$$F_{N(t)}(n) = P\{N(t) <= n\}= 1 - P\{N(t) >= n + 1\} = \sum_{j=0}^{n} e^{-\lambda t}(\lambda t)^j/j! \ .(1.7)$$

Portanto, a função máxima de N(t) é

$$p_{N(t)}(n) = P\{N(t) = n\} = F_{N(t)}(n) - F_{N(t)}(n-1) = e^{-\lambda t}(\lambda t)^n/n!. \quad (1.8)$$

Isto é a função máxima de uma variável aleatória com uma distribuição de Poisson com parâmetro λt, de onde vem a razão para o fato deste processo ser chamado de Processo de Poisson. Isto é, para um processo de chegada de Poisson com taxa de chegada λ, a distribuição do número de chegadas em t unidades de tempo é Poisson com parâmetro λt.

O processo de Poisson tem muitas propriedades interessantes. As duas mais fundamentais são que os incrementos das chegadas são ambos independentes e de tempo estacionário. *Incrementos independentes* significa que para um tempo arbitrário $t>=0$ e um intervalo de tempo $\Delta t >= 0$, o número de chegadas no período de tempo $(t, t + \Delta t]$ é independente do número de chegadas até o tempo t, isto é,

$$P\{N(t + \Delta t) - N(t) = m, N(t) = k\} \quad (1.9)$$

$$= P\{N(t + \Delta t) - N(t) = m\} * P\{N(t) = k\}$$

Quanto ao fato que as chegadas tem incrementos estacionários significa que a distribuição do número de chegadas não muda com o tempo, isto é, novamente para um tempo arbitrário $t >=0$ e um intervalo de tempo $\Delta t >=0$,

$$P\{N(t + \Delta t) - N(t) = m\} = P\{N(0 + \Delta t) - N(0) = m\} = P\{N(\Delta t) = m\} \quad (1.10)$$

Pela equação (1.8) está é uma distribuição de Poisson com parâmetro c, de modo que, o número esperado de novas chegadas em qualquer período de comprimento Δt é

$$E[N(t + \Delta t) - N(t)] = \lambda \Delta t. \quad (1.11)$$

Em palavras, o número esperado de chegadas em um intervalo de tempo fixo é proporcional a taxa de chegada e o comprimento do intervalo de tempo, que é a razão porque o parâmetro λ é referido como a taxa do processo de chegada.

Um processo de Poisson Simples

Considere um processo de Poisson $\{N(t) : t >= 0\}$ com taxa $\lambda = 2$. Pode-se usar a propriedade independência, a propriedade tempo-estacionário e a função máxima da distribuição de Poisson para calcular várias quantidades. Por exemplo,

$$P\{N(5) = 4 | N(4) = 2\} = \frac{P\{N(5) = 4, N(4) = 2\}}{P\{N(4) = 2\}}$$

$$= \frac{P\{N(4+1) - N(4) = 2, N(4) = 2\}}{P\{N(4) = 2\}}$$

$$= \frac{P\{N(4+1) - N(4) = 2\} * P\{N(4) = 2\}}{P\{N(4) = 2\}}$$

$$= P\{N(4+1) - N(4) = 2\} = P\{N(1) = 2\}$$

$$= \frac{(1.\lambda)^2}{2!} e^{-1.\lambda} = \frac{(1.2)^2}{2!} e^{-1.2}$$

$$= 2e^{-2} = 0.27.$$

Note que primeiro foi utilizado a definição de probabilidade condicional, então foi utilizada a simples observação que $\{N(5) = 4, N(4) = 2\}$, se e somente se $\{N(4 + 1) - N(4) = 2, N(4) = 2\}$, e depois que os incrementos independentes e a estacionaridade do tempo reduzem o problema, de forma que, a função máxima da distribuição pode ser utilizada.

1.7.3 Distribuição Exponencial

Como dito anteriormente, desde que os tempos de interchegada do processo de Poisson são exponencialmente distribuídos é interessante olhar esta distribuição um pouco mais de perto. Para tanto, seja $X \sim \exp(\lambda)$ exponencial com parâmetro λ. Assumindo que o tempo é iniciado em zero e mantendo as aplicações de fila em mente, pode-se para propósitos de ilustração pensar em X como o tempo da chegada do próximo cliente.

Assuma que em algum tempo fixo $t > 0$ foi checado se alguma chegada ocorreu e descobriu-se que a chegada ainda não ocorreu. Denote o tempo de t até o momento em que a chegada efetivamente ocorreu como R_t. Para determinar a distribuição de R_t simplesmente calcula-se a função de distribuição

$P\{R_t <= r\} = 1 - P\{R_t > r\}$.

Optou-se por olhar o evento $R_t > r$, haja visto que torna-se mais fácil obter

$$P\{R_t > r\} = \{X > t + r \mid X > t\} \tag{1.12}$$

$$= \frac{P\{X > t + r , X > t\}}{P\{X > t\}}$$

$$= \frac{P\{X > t + r\}}{P\{X > t\}}$$

$$= e^{-\lambda(t + r)} / e^{-\lambda t}$$

$$= e^{-\lambda r}$$

$$= P\{X > r\}.$$

Assim, pode-se obter a função de distribuição $P\{R_t <= r\} = 1 - P\{R_t > r\} = 1 - e^{-\lambda r} = P\{X <= r\}$, que implica que R_t é exponencialmente distribuído com parâmetro λ, o mesmo que X. O que isso significa é que sempre que para-se para checar se um cliente chegou e ele não chegou, então o tempo restante até ele chegar tem a mesma distribuição que o tempo anterior a checagem. Esta propriedade, que é única para a distribuição exponencial em relação a todas as distribuições contínuas , é conhecida como propriedade markoviana ou falta de memória da distribuição exponencial.

1.8 Processos Estocásticos e Cadeias de Markov

Um processo estocástico é a abstração matemática de um processo empírico cujo desenvolvimento é governado por leis probabilísticas (o processo de Poisson é um exemplo). Do ponto de vista da teoria da probabilidade matemática, um processo estocástico é melhor definido como um conjunto de variáveis aleatórias, $\{X(t), t \varepsilon T\}$, definido sobre algum conjunto de índices ou espaço de parâmetro T. O conjunto T é algumas vezes também chamado de faixa de tempo, e X(t) denota o estado do processo no tempo t. Dependendo da natureza da faixa de tempo, o processo é classificado como um processo de parâmetro contínuo ou de parâmetro discreto como a seguir:

1. Se T é uma seqüência contável, por exemplo, $T =\{..., -2,-1,0,1,2,...\}$ ou $T = \{0,1,2,...\}$, então o processo estocástico $\{X(t), t \varepsilon T\}$ é dito ser um processo de parâmetro discreto definido no conjunto de índices de T.
2. Se T é um intervalo ou uma combinação de intervalos algébricos, por exemplo, $T = \{t:-\infty < t < +\infty\}$ ou $T = \{t:0 < t < +\infty\}$, então o processo estocástico $\{X(t), t \varepsilon T\}$ é chamado de processo de parâmetro contínuo definido no conjunto de índices de T.

1.8.1 Processo de Markov

Um processo estocástico de parâmetro discreto $\{X(t), t=0,1,2,...\}$ ou processo estocástico de parâmetro contínuo $\{X(t), t > 0\}$ é dito ser um processo de Markov se, para qualquer conjunto dos n pontos de tempo $t_1 < t_2 < ... < t_n$ no conjunto de índices ou faixa de tempo do processo, a distribuição condicional de $X(t_n)$, dados os

valores de $X(t_1)$, $X(t_2)$, $X(t_3)$, ..., $X(_{tn-1})$, depender só do valor imediatamente precedente, $X(t_{n-1})$; mais precisamente, para quaisquer números reais x_1, x_2, ..., x_n,

$$P\{X(t_n) <= x_n \mid X(t_1) = x_1, ..., X(_{tn-1}) = x_{n-1}\}$$
$$= P\{X(t_n) <= x_n \mid X(t_{n-1}) = x_{n-1}\}.$$

Em uma linguagem não matemática pode-se dizer que, dado a condição "presente" do processo, o "futuro" é independente do "passado", ou seja, o processo está sem memória.

Os processos de Markov são classificados de acordo com;

1. A natureza do conjunto de índices do processo (se de espaço de parâmetro discreto ou contínuo), e
2. A natureza do espaço de estado do processo (se de espaço de estado discreto ou contínuo).

Um número real x é dito ser estado do processo estocástico $\{X(t), t \varepsilon T\}$ se existe um ponto de tempo t em T tal que a $P\{x - h < X(t) < x + h\}$ é positivo para todo $h > 0$. O conjunto de estados possíveis constitui o espaço de estado do processo. Se o espaço de estado é discreto, o processo de Markov é geralmente chamado de cadeia de Markov, sendo que os processos de Markov de parâmetro discreto com espaço de estado discreto são conhecidos como cadeias de Markov planas, e os processos de Markov de parâmetro contínuo com espaço de estado discreto são conhecidos como cadeias de Markov de tempo contínuo.

Tabela 1.5 Classificação do Processo de Markov

	Tipo de Parâmetro	
Espaço de Estado	Discreto	Contínuo
Discreto	(Parâmetro Discreto) Cadeia de Markov	Cadeia de Markov de Parâmetro Contínuo
Contínuo	Processo de Markov de Parâmetro Discreto	Processo de Markov de Parâmetro Contínuo

Quando o processo de Markov tem um espaço de estado contínuo e um espaço de parâmetro discreto, ele é conhecido como processo de Markov de parâmetro discreto. Por outro lado, se ambos espaço de estado e espaço de parâmetro forem contínuos então o processo de Markov será chamado de processo de Markov de parâmetro contínuo. A tabela 1.5 sumariza estas classificações do processo de Markov.

1.8.2 Cadeias de Markov de Parâmetro Discreto

Considere uma cadeia de Markov com dois estados. Um dos estados denotado por 1 representará um "sucesso", enquanto o outro estado representará uma "fracasso" sendo denotado por 0.

Suponha que se a n-ésima tentativa resultar em fracasso então a probabilidade de fracasso na n-ésima primeira tentativa é $1 - \alpha$ e a probabilidade de sucesso nesta mesma tentativa é igual a α. Similarmente, se a n-ésima tentativa resultar em sucesso então as probabilidades de sucesso e fracasso na n-ésima primeira tentativa são $1 - \beta$ e β, respectivamente. Estas probabilidades são chamadas de *probabilidades de transição*, podendo serem escritas na seguinte forma matricial, também conhecida como matriz de transição :

$$P = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{bmatrix} \end{array}$$

O elemento da matriz na posição (j,k) denota a probabilidade condicional de uma transição para um estado k no tempo n + 1 dado que o sistema está no estado j no tempo n.

Exemplo 1.1 Chuva em Tel Aviv.

Em um estudo de chuvas em Tel Aviv, Gabriel e Nenmann encontraram que uma cadeia de Markov de dois estados daria uma boa descrição da ocorrência de dias chuvosos e secos durante o período chuvoso de Dezembro, Janeiro e Fevereiro. Para tanto, eles denominaram dia seco como estado 0 e dia chuvoso como estado 1. Realizando a amostragem de um período de 37 anos, eles descobriram que a probabilidade α de ocorrência de um dia chuvoso após um dia seco era de 0.250 e a probabilidade β de ocorrência de um dia seco após um dia chuvoso era de 0.338, formando a matriz de transição mostrada abaixo:

$$P = \begin{array}{c} \\ S \\ C \end{array} \begin{array}{c} \begin{array}{cc} S & C \end{array} \\ \begin{bmatrix} 0.750 & 0.250 \\ 0.338 & 0.662 \end{bmatrix} \end{array}$$

Denotando a linha do vetor $P^{(n)} = (P_0^{(n)}, P_1^{(n)})$ como a probabilidade de encontrar o sistema no estado 0 ou 1 no tempo n quando as probabilidades iniciais dos dois estados são dados por $P^{(0)} = (P_0^{(0)}, P_1^{(0)})$. Este evento pode ocorrer de duas maneiras mutuamente exclusivas. O estado 0 foi ocupado no tempo n –1 e nenhuma transição de estado ocorreu no tempo n, de forma que este evento tem a probabilidade $P_0^{(n-1)}(1 - \alpha)$. Alternativamente, o estado 1 foi ocupado no tempo n - 1 e uma transição do estado 1 para o estado 0 ocorreu no tempo n, de forma que este evento tem a probabilidade $P_1^{(n-1)}\beta$. Considerações, tais como estas levam as seguintes relações de recorrência:

$$P_0^{(n)} = P_0^{(n-1)}(1 - \alpha) + P_1^{(n-1)}\beta,$$

$$P_1^{(n)} = P_0^{(n-1)}\alpha + P_1^{(n-1)}(1-\beta),$$

que em notação de matriz podem ser escritas como

$$P^{(n)} = P^{(n-1)} P, \tag{1.13}$$

e em iteração

$$P^{(n)} = P^{(n-2)} P^2 = P^0 P^n. \tag{1.14}$$

Assim, dado as probabilidades iniciais P^0 e a matriz de transição de probabilidade P, pode-se encontrar qualquer probabilidade de ocupação de estado em qualquer tempo n usando a equação 1.14.

Denotando o elementos (j,k) de P^n por $p_{jk}^{(n)}$. Se o sistema está inicialmente no estado 0, então $P^{(0)} = (1,0)$ e $P^{(n)} = (p_{00}^{(n)}, p_{01}^{(n)})$. Por outro lado, se o sistema está inicialmente no estado 1, então $P^{(0)} = (0,1)$ e $P^{(n)} = (p_{10}^{(n)}, p_{11}^{(n)})$.

Assim,

$$p_{jk}^{(n)} = P\{X_n = k \mid X_0 = j\}$$

Estas quantidades $p_{jk}^{(n)}$ são chamadas *probabilidades de transição n-passos*. Esta é uma matriz equivalente as conhecidas equações de Chapman-Kolmogorov (CK) para este processo de Markov, mostrada abaixo:

$$p_{jk}^{(n)} = \sum_r p_{ir}^{(n-k)} p_{rj}^{(k)}$$

Exemplo 1.2 Chuva em Tel Aviv.

Utilizando a matriz do exemplo 1.1 pode-se calcular P5, a partir de P^2, $P^3 = P^2P$, $P^5 = P^3P^2$, de maneira que o resultado encontrado é mostrado na matriz abaixo:

$$P = \begin{array}{c} \\ S \\ C \end{array} \begin{array}{c} \begin{array}{cc} S & C \end{array} \\ \begin{bmatrix} 0.580 & 0.420 \\ 0.568 & 0.432 \end{bmatrix} \end{array}$$

Assim, por exemplo se 1 de Janeiro é um dia seco a probabilidade que 6 de Janeiro seja um dia seco é 0.580, enquanto que se 1 de Janeiro é um dia chuvoso, a probabilidade de 6 de Janeiro ser um dia seco é de 0.568.

Uma questão que eleva-se, é que após um período suficientemente longo de tempo o sistema se estabilizará para um situação de equilíbrio estatístico na qual as probabilidades de ocupação de estado serão independentes das condições iniciais. Se isto for verdade existe uma distribuição de probabilidade de equilíbrio $\pi = (\pi_0, \pi_1)$ e, quando $n \rightarrow \infty$, π satisfará claramente a condição

$$\pi = \pi P,$$

ou

$$\pi(I - P) = 0. \qquad (1.15)$$

A condição $\pi = \pi P$ será demonstrada na próxima seção.

Assim,

$$\pi_0\alpha - \pi_1\beta = 0, \qquad -\pi_0\alpha + \pi_1\beta = 0.$$

Este é um sistema de equações *homogêneas* e só terá uma solução não zero se o determinante $|I - P|$ não existir. Claramente, $|I - P|$ existe e pode-se obter a solução única utilizando a seguinte condição

$$\pi_0 + \pi_1 = 1$$

Note que se a distribuição de probabilidade é π, então

$$P^{(1)} = \pi P = \pi, \; P^{(2)} = P^{(1)}P = \pi P = \pi,$$

e

$$P^{(n)} = \pi \; (n= 1,2,...)$$

Assim, a distribuição $P^{(n)}$ é estacionária se $P^{(0)} = \pi$, ou seja, não muda com o tempo.

Exemplo 1.3 Chuva em Tel Aviv.

Usando-se a matriz P do exemplo 1.1, obtém-se que as probabilidades de equilíbrio são

$\pi_0 = 0.575$, $\pi_1 = 0.425$, como mostrado abaixo:

$$P_{10} = \begin{array}{c} \\ S \\ C \end{array} \begin{array}{c} \begin{array}{cc} S & C \end{array} \\ \begin{bmatrix} 0.575 & 0.425 \\ 0.575 & 0.425 \end{bmatrix} \end{array}$$

1.8.3 Propriedades de Longa Execução do Processo de Markov

Nesta seção será abordada as propriedades de longa execução das cadeias de Markov, com ênfase principal as propriedades do limite de movimentação de um estado para outro,

$$\lim_{n\to\infty} Pij^{(n)} = \pi_j, \ \forall j \in S,$$

que quando existir, refere-se a distribuição limítrofe para a cadeia de Markov.

Para tanto, considere uma cadeia de Markov e suponha que

$$\lim_{n\to\infty} pij^{(n)} = \pi_j \qquad \text{(para todo i);}$$

isto é, para um longo período, a probabilidade que o processo esteja no estado j, dado que ele iniciou no estado i é independente do estado inicial i. Isto significa que P^n possui um limite quando n tende para infinito, de forma que, todas as linhas de P^n se tornam idênticas. π_j ,neste caso, é chamado de probabilidades de *estado de equilíbrio* ou *limítrofe* da cadeia de Markov.

Considere agora a probabilidade incondicional após n passos de $\pi^{(n)} = \pi^{(0)}P^n$, isto é,

$$\pi_j^{(n)} = \sum_i \pi_i^{(0)} p_{ij}^{(n)}$$

Então

$$\lim_{n\to\infty} \pi_j^{(n)} = \lim_{n\to\infty} \sum_i \pi_i^{(0)} p_{ij}^{(n)} = \sum_i \pi_i^{(0)} \lim_{n\to\infty} p_{ij}^{(n)}$$
$$= \sum_i \pi_i^{(0)} \pi_j = \pi_j \sum_i \pi_i^{(0)} = \pi_j ,$$

e consequentemente $\pi_j^{(n)}$ tende para o mesmo limite que $pij^{(n)}$ e é independente das probabilidades iniciais e do parâmetro de tempo n.

Partindo do mesmo princípio de que $P^{(n)} = P^{(n-1)}P$, demonstrado anteriormente, tem-se que

$$\lim_{n\to\infty} \pi^{(n)} = \lim_{n\to\infty} \pi^{(n-1)} = \pi,$$

de forma que,

$$\pi = \pi P.$$

Estas equações bem conhecidas são chamadas de *equações estacionárias* da cadeia de Markov, e suas soluções são chamadas de *distribuições estacionárias*.

É possível, em alguns casos, obter soluções para (1.15) mesmo quando nenhuma distribuição do limite existe. Assim, quando existe uma distribuição do limite, isto implica uma solução para (1.15) , e a distribuição estacionária resultante é a distribuição do limite. Mas, o contrário não é verdadeiro, isto é, a solução para (1.15) não implica na existência de uma distribuição limítrofe.

Dessa forma, faz-se necessário encontrar outras maneiras de verificar a existência da distribuição limítrofe. Uma das maneiras mais utilizadas é identificar os tipos de estados pertencentes a uma cadeia de Markov e utilizar sua propriedades para determinar a existência da distribuição.

Para tanto, suponha que uma cadeia esteja inicialmente em um dado estado j e defina a probabilidade de j voltar a ser ocupado após n transições como

$$f_{jj} = \sum_{n=1}^{\infty} f_{jj}^{(n)}.$$

Se f_{jj} é igual a 1 então o estado é chamado *recorrente*. Por outro lado, se f_{jj} é menor que 1 então j é um estado *transiente*. Quando f_{jj} =1,

$$m_{jj} = \sum_{n=1}^{\infty} n f_{jj}^{(n)}$$

é o tempo de recorrência médio, ou seja, o tempo médio para que o estado j volte a ser visitado. Se $m_{jj} < \infty$, então j é conhecido como um *estado recorrente positivo*, enquanto se $m_{jj} = \infty$, então j é um *estado recorrente nulo*.

Suponha que quando a cadeia ocupa o estado j, subseqüentes ocupações do estado j só ocorram nos tempos $t, 2t, 3t, \ldots$ quando t é um inteiro maior que 1. Quando isto ocorre, o estado j é chamado de *periódico* com período t. Um estado que não é periódico é conhecido como *aperiódico*, sendo que se além de ser aperiódico ele for recorrente positivo então ele é conhecido como ergódico.

Para exemplificar observe a cadeia de Markov com espaço de estado S = 0,1 e matriz de transição:

$$P = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \end{array}$$

é periódica com período 2, pois, assumindo que o estado inicial seja $X_0 = 0$ ele só poderá probabilisticamente voltar a ser ocupado após 2 transições de estados, ou seja, em P_3, P_5, P_7, ...

Em particular, para qualquer $j \in S$,

$$\lim_{n \to \infty} Pij^{(n)} = \pi_j,$$

existe, se somente se, j é aperiódico.

Um estado j é dito se comunicar com um estado k se é possível alcançar k a partir de j num número finito de transições. Por outro lado, se j se comunica com k e k se comunica com j então esta relação é conhecida como intercomunicação entre j e k. Além disso, uma cadeia de Markov onde todos os estados se comunicam é chamado *irredutível*, caso contrário, a cadeia é dita *redutível*. Um subconjunto de estados onde a cadeia de Markov entre e nunca mais saia é chamado de *conjunto fechado*.

Identificar todos os subconjuntos irredutíveis habilita a classificação dos estados em recorrentes ou transientes:

- Estados que pertencem a um subconjunto fechado irredutível são recorrentes.
- Estados que não pertencem a um subconjunto fechado irredutível são transientes.

Note também que se dois estados se comunicam então ambos são recorrentes ou transientes.

Para fixar a idéia de como os estados são classificados observe o exemplo abaixo:

$$P = \begin{bmatrix} 0.1 & 0.3 & 0.6 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0.8 & 0.2 & 0 \\ 0 & 0 & 0.7 & 0.3 & 0 \\ 0.5 & 0 & 0 & 0.5 & 0 \end{bmatrix}$$

Em P existem dois conjuntos fechados irredutíveis R1 = {2} e R2 = {3,4}, de modo que, estes estados são recorrentes. Os estados transientes são, portanto, por exclusão, T = {1,5}. A explicação para o conjunto R1 ser fechado é que se o estado 2 for alcançado, nunca mais ocorrerá uma transição de estado. O mesmo acontece para o conjunto R2, pois, se 3 ou 4 forem alcançados, as transições de estado ficarão permanentemente entre esses dois estados. Já nos estados 1 e 5, existem probabilidades de transição diretamente ou indiretamente para os conjuntos fechados R1 e R2, de forma

que, caso isso aconteça 1 e 5 nunca mais serão revisitados. Além disso, todos os estados são aperiódicos.

Exemplo da Falha em uma Máquina

Seja,

$$Xn = \begin{cases} 0, \text{ se a máquina for reparada no dia n,.} \\ 1, \text{ se a máquina estiver operando no dia n} \end{cases}$$

uma cadeia de Markov com a matriz de transição,

$$P = \begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{bmatrix} 0.2 & 0.8 \\ 0.1 & 0.9 \end{bmatrix} \end{array}$$

Utilizando do fato mostrado anteriormente que $\pi = \pi P$ tem-se que

$$\pi = \lim_{n \to \infty} \pi^{(n)} = (1/9, 8/9)$$

é um vetor de probabilidade de estado de longa execução. Se a execução for iniciada com $\pi^{(0)} = \pi = (1/9, 8/9)$ então o vetor de probabilidade de estado no primeiro passo será

$$\pi^{(1)} = (1/9, 8/9) \begin{pmatrix} 0.2 & 0.8 \\ 0.1 & 0.9 \end{pmatrix} = (1/9, 8/9)$$

Em outras palavras, quando uma cadeia é iniciada com uma distribuição limítrofe π como vetor de probabilidade de estado inicial, então após uma transição o vetor de probabilidade de estado será o mesmo.

Isto significa que quando o estado de longa execução for alcançado a cadeia de Markov entrará num estado de *equilíbrio* ou *estacionário*, isto é, num sentido probabilístico nada muda deste ponto em diante. Por este motivo, que esta distribuição é conhecida como *distribuição estacionária*.

1.8.4 Cadeias de Markov de Parâmetro Contínuo

Os modelos discretos são úteis para responder a questões pontuais, no que diz respeito ao tempo, entretanto, quando se deseja saber, por exemplo, qual o tempo esperado exato do reparo de uma máquina ou o tempo exato até que duas máquinas em funcionamento sejam desligadas para reparo, o modelo discreto não é satisfatório. Assim, faz-se necessário a inclusão de um processo estocástico de tempo contínuo {

X(t) : t > 0}, onde X(t) é o estado do sistema no tempo t>=0. Se o modelo contínuo é adotado, o interesse geralmente recai sobre

- quanto tempo o sistema fica em um dado estado, e
- dado seu estado atual, para qual estado o processo se moverá.

Inicialmente, será tentado modelar a segunda dessas questões. Se for assumido que a probabilidade de mover de um estado para outro só depende do estado atual e não do comportamento passado, isto é, se o processo não possui memória, então as transições de estado podem ser descritas como uma cadeia de Markov,

$$P = \begin{bmatrix} 0 & \gamma_1 & 1 - \gamma_1 \\ \gamma_2 & 0 & 1 - \gamma_2 \\ \gamma_3 & 1 - \gamma_3 & 0 \end{bmatrix}$$

onde γ_1 é a probabilidade de uma máquina quebrar quanto existem duas operando, $1 - \gamma_1$ é a probabilidade de duas máquinas quebrarem simultaneamente, e assim por diante. Esta é uma matriz de transição de uma cadeia de Markov embutida que descreve as transições entre estados para este processo de tempo contínuo. Note que como o processo é contínuo não existe nenhuma transição de um estado para ele mesmo. Esta cadeia de Markov embutida é metade do que é necessário para descrever o processo acima. A outra metade é quanto tempo um processo fica em um dado estado, ou *tempo de permanência*. Para este exemplo, o tempo de permanência T_i no estado i = 0,1,2 é

T_0 = Tempo até que um ou mais máquinas falhem.

T_1 = Tempo até que uma máquina falhe ou seja reparada.

T_0 = Tempo até que um ou mais máquinas sejam reparadas.

Um processo estocástico de tempo contínuo, onde as transições podem ser escritas com a cadeia de Markov e o tempo de permanência é uma variável aleatória com uma distribuição arbitrária é chamado de *processo semi-Markov*. Quando um modelo de uma cadeia de Markov é construído, o foco esta no processo de mudança de estado, ao invés de quanto tempo o sistema ficou em um dado estado, que é de interesse em muitas aplicações. Por exemplo, quanto tempo uma máquina está operando ? Quanto tempo ela está ociosa ? Quanto tempo um cliente tem que esperar por um serviço? Se essas questões são de interesse existe a necessidade de modelar o sistema como um processo estocástico de tempo contínuo

Assim, seja { Y(t) : t >= 0} um processo estocástico de tempo contínuo com espaço de estado S. Então, Y(t) : t >= 0 é uma cadeia de Markov de tempo contínuo (CMTC). Se para todos os tempos s,t >= 0,

1. A propriedade de markov

   $P\{Y(t) = j \mid Y(s) = i, Y(u) = y_u, 0 <= u < s\} = P\{Y(t) = j \mid Y(s) = i\}$ se mantém, e (1.16)

2. O tempo estacionário for dado por,

   $P\{Y(t) = j \mid Y(s) = i\} = P\{Y(t - s) = j \mid Y(0) = i\}$ (1.17)

Isto é equivalente aos processos que descrevem as transições de estado de uma cadeia de Markov com o tempo gasto em cada estado sendo exponencialmente distribuído. Em outras palavras o CMTC é um processo semi-Markov com tempos de permanência exponencialmente distribuídos. As cadeias de Markov que descrevem as transições de estado são conhecidas como *cadeias de Markov embutidas*.

Exemplo de uma CMTC simples.

Assuma que o tempo até que um cliente chegue seja exponencialmente distribuído com taxa λ e o tempo que ele leva para ser servido e parta tenha a mesma distribuição com taxa μ. Assim, se existem i > 0 clientes então o tempo até que um cliente chegue ou parta, e o estado do sistema mude, é exponencialmente distribuído com taxa $\lambda + \mu$.

Um CMTC com espaço de estado S = {1,2, ..., n}é caracterizado pelo vetor de estado inicial $\pi^{(0)} = (\pi_1^{(0)}\ \pi_2^{(0)} \cdots \pi_n^{(0)})$, onde $\pi_i^{(0)} = P\{Y_0 = 1\}$ e sua matriz geradora, também conhecida como *matriz de intensidade* ou *gerador infinitesimal* possui a seguinte forma:

$$Q = \begin{bmatrix} -q_0 & q_{01} & q_{02} & q_{03}\cdots \\ q_{10} & -q_1 & q_{12} & q_{13}\cdots \\ q_{20} & q_{21} & -q_2 & q_{23}\cdots \end{bmatrix}$$

Aqui q_{ij} é a taxa na qual o processo de move do estado i para o estado j, e $q_{ii} = -\sum_{j \neq i} q_{ij}$.

A matriz geradora para esta CMTC simples é

$$Q = \begin{matrix} [-\lambda_0 & \lambda_0 & 0 & 0 & 0\cdots \\ \mu_1 & -(\lambda_1 + \mu_1) & \lambda_1 & 0 & 0\cdots \\ 0 & \mu_2 & -(\lambda_2 + \mu_2) & \lambda_2 & 0\cdots \end{matrix}$$

$$\begin{matrix} \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots] \end{matrix}$$

que pode ser facilmente memorizada da seguinte forma:

$q_{i,i+1} = \lambda_i$,
$q_{i,i-1} = \mu_i$ $(\mu_i \neq 0)$
q_{rj} (caso contrário)
$q_i = \lambda_i + \mu_i$ $(q_0 = \lambda_0)$

Para o CMTC simples, a matriz de transição P da cadeia de Markov embutida pode ser obtida a partir da matriz geradora Q, usando as seguintes regras:

$$p_{ij} = \begin{matrix} \lambda_i / (\lambda_i + \mu_i) & (j = i+1, i >= 1), \\ \mu_i / (\lambda_i + \mu_i) & (j = i-1, i >= 1), \\ 1 & (i=0, j=1), \\ 0 & \text{(caso contrário)}. \end{matrix}$$

Assim,

$$P = \begin{matrix} [& 0 & 1 & 0 & 0 & 0\dots \\ & \mu_1 / (\lambda_1 + \mu_1) & 0 & \lambda_1 / (\lambda_1 + \mu_1) & 0 & 0\dots \\ & 0 & \mu_2 / (\lambda_2 + \mu_2) & 0 & \lambda_2 / (\lambda_2 + \mu_2) & 0\dots \\ & \dots & \dots & \dots & \dots & \dots] \end{matrix}$$

Calcular a cadeia de Markov torna possível a classificação dos estado, haja visto que o estado $i \in S$ é transiente/recorrente para a CMTC, se e somente se, é transiente/recorrente para a cadeia de Markov embutida. Quando $\mu >= \lambda$ todos os estados são recorrentes e transientes, em caso contrário.

A distribuição estacionária π pode ser calculada através das equações abaixo:

$$0 = \pi Q, \qquad (1.17)$$
$$1 = \pi 1.$$

As mesmas interpretações anteriores são mantidas, mas não existe nenhum problema com a periodicidade, isto é, tão longo seja o tempo que se esteja observando em um conjunto irredutível de estados recorrentes, ou sabendo o conjunto iniciado, a distribuição estacionária pode ser interpretada como:

- A fração de tempo de longa execução gasta em cada estado.
- A probabilidade de encontrar o processo em um dado estado depois de um longo tempo.

Um processo de poisson é um caso especial de um processo de nascimento-puro, onde tal processo apresenta a seguinte matriz geradora e cadeia de Markov embutida, sendo que no processo de Poisson $\lambda_i = \lambda$.

$$Q = \begin{bmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & 0... \\ 0 & -\lambda_1 & \lambda_1 & 0 & 0... \\ 0 & 0 & -\lambda_2 & \lambda_2 & 0... \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \end{bmatrix},$$

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0... \\ 0 & 0 & 1 & 0 & 0... \\ 0 & 0 & 0 & 1 & 0... \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \end{bmatrix}$$

1.9 Processo de Nascimento – Morte

Agora será estudado uma cadeia de Markov de tempo contínuo chamada *processo de nascimento - morte*. Como anteriormente, seja $\{Y(t) : t >= 0\}$ uma CMTC com espaço de estado $S = \{0,1,2,...\}$. Agora, assuma que no estado 0 só existe um evento $e_n^{(0)}$ que pode ocorrer, e quando ocorre então o estado do sistema é incrementado de um. Em qualquer outro estado $i > 0$, existem dois eventos possíveis: um nascimento $e_n^{(i)}$ e uma morte $e_m^{(i)}$. Se $e_m^{(i)}$ ocorre o estado do sistema é decrementado de um e se $e_n^{(i)}$ ocorre o estado do sistema é incrementado de um. λ_i e μ_i serão chamados de taxas de nascimento e morte, no estado i, respectivamente. A matriz geradora será dada por

$$Q = \begin{bmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & 0... \\ \mu_1 & -(\lambda_1 + \mu_1) & \lambda_1 & 0 & 0... \\ 0 & \mu_2 & -(\lambda_2 + \mu_2) & \mu_2 & 0... \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ... \end{bmatrix},$$

e a cadeia de Markov embutida será dada por

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0... \\ \mu_1/(\lambda_1 + \mu_1) & 0 & \lambda_1/(\lambda_1 + \mu_1) & 0 & 0... \\ 0 & \mu_2/(\lambda_2 + \mu_2) & 0 & \lambda_2/(\lambda_2 + \mu_2) & 0... \\ ... & ... & ... & ... & ... \\ ... & ... & ... & ... & ...] \end{bmatrix}$$

Note que este processo já foi olhado anteriormente onde todos os $i \in S$ satisfazem $\lambda_i = \lambda$, $\mu_i = \mu$ (a CMTC simples); $\mu_i = 0$ (processo de nascimento puro); $\lambda_i = \lambda$, $\mu_i = 0$ (o processo de Poisson).

Dessa forma, se for desenhado um diagrama com as transições de estados será verificado que o fluxo de entrada em um estado e igual ao seu fluxo de saída.

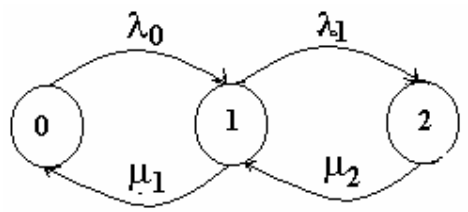

Figura 1.7 – Diagrama de Balanço

Assim, com as condições apropriadas em μ_i e λ_i, uma solução de estado de equilíbrio existe e pode ser determinada de 0 = pQ, de forma que, equações de balanço para o processo de nascimento e morte podem ser geradas como a seguir :

$$0 = -(\lambda_j + \mu_j)p_j + \lambda_{j-1}p_{j-1} + \mu_{j+1}p_{j+1} \quad (j >= 1), \qquad (1.18)$$

$$0 = -\lambda_0 p_0 + \mu_1 p_1$$

Rescrevendo, obtém-se que

$$p_{j+1} = \frac{(\lambda_j + \mu_j)}{\mu_{j+1}}p_j - \frac{\lambda_{j-1}}{\mu_{j+1}}p_{j-1} \qquad (j >= 1),$$

$$p_1 = (\lambda_0 / \mu_1)p_0.$$

$$p_2 = \frac{(\lambda_1 + \mu_1)}{\mu_2}p_1 - \frac{\lambda_0}{\mu_2}p_0$$

$$= \frac{(\lambda_1 + \mu_1)}{\mu_2}\frac{\lambda_0}{\mu_1}p_0 - \frac{\lambda_0}{\mu_2}p_0$$

$$= \frac{\lambda_1\lambda_0}{\mu_2\mu_1}p_0$$

$$p_3 = \frac{(\lambda_2 + \mu_2)}{\mu_3}p_2 - \frac{\lambda_1}{\mu_3}p_1$$

$$= \frac{(\lambda_2 + \mu_2)}{\mu_3} \frac{\lambda_1\lambda_0 p_0}{\mu_2\mu_1} - \frac{\lambda_1\lambda_0 p_0}{\mu_3\mu_1}$$

$$= \frac{\lambda_2\lambda_1\lambda_0 p_0,}{\mu_3\mu_2\mu_1}$$

.

.

.

O padrão que parece estar emergindo é que

$$p_n = \frac{\lambda_{n-1}\lambda_{n-2} \ldots \lambda_0 p_0,}{\mu_n\mu_{n-1} \ldots \mu_1} \qquad (n >=1) \qquad (1.19)$$

$$= p_0 \prod_{i=1}^{n} (\lambda_{i-1}/\mu_i)$$

e por indução pode-se comprovar que a fórmula esta correta. Entretanto, para os propósitos deste curso não será necessário a demonstração.

Agora, desde que o somatório das probabilidades tem que ser igual a um, segue que

$$p_0 = (1 + \sum_{n=1}^{\infty} \prod_{i=1}^{n} (\lambda_{i-1}/\mu_i))^{-1} \qquad (1.20)$$

Estes tipos de equações são muito úteis para gerar uma variedade de modelos de filas. Por exemplo, adotando-se $\lambda_n = \lambda$ e $\mu_n = \mu$ obtém-se uma fila M/M/1.

2. MODELOS DE FILAS APLICADOS A PROCESSOS MARKOVIANOS DE NASCIMENTO E MORTE SIMPLES

O propósito deste capítulo é desenvolver um conjunto de modelos de filas pela aplicação direta dos resultados obtidos no capítulo anterior para o processo de nascimento e morte.

2.1 Soluções de Estado de Equilíbrio para Modelos M/M/1

As funções densidade para os tempos de interchegada e tempos de serviço para as filas M/M/1 são dadas, respectivamente, como

$$a(t) = \lambda e^{-\lambda t},$$

$$b(t) = \mu e^{-\mu t},$$

onde $1/\lambda$ então é o tempo de interchegada médio e $1/\mu$ é o tempo de serviço médio. Tempos de interchegada, tanto quanto os tempos de serviço, são assumidos serem estatisticamente independentes. Note que, os tempos de interchegada e serviço são exponenciais, e as taxas de chegada e serviço condicionais obedecem uma distribuição de Poisson.

Além disso, o problema M/M/1 é um processo de nascimento e morte com $\lambda_n = \lambda$ e $\mu_n = \mu$, para todo n. Chegadas podem ser consideradas como nascimentos para o sistema, desde que, se o sistema está no estado n, ou seja, se existem n clientes no sistema e uma chegada ocorre, o estado do sistema é atualizado para n + 1. Por outro lado, a ocorrência de uma partida pode ser considerada como uma morte, de maneira que, quando este evento ocorre, o sistema será levado para o estado n – 1, estando ele no estado n. Note que isto vale para todos os valores onde n é maior ou igual a 1. Consequentemente, utilizando-se as equações 1.18, obtém-se as equações de estado de equilíbrio como

$$0 = -(\lambda + \mu)p_n + \lambda p_{n-1} + \mu p_{n+1} \qquad (n >= 1), \qquad (2.1a)$$

$$0 = \lambda p_0 + \mu p_1,$$

ou

$$p_{n+1} = \frac{(\lambda + \mu)}{\mu} p_n - \frac{\lambda}{\mu} p_{n-1} \qquad (n >= 1), \qquad (2.1b)$$

$$p_1 = (\lambda / \mu) p_0.$$

2.1.1 Balanço Estocástico

Um atalho para obter diretamente as equações estacionárias (2.1) é por meio das equações de balanço mostrada anteriormente, onde o fluxo de entrada em um estado tem que ser igual ao fluxo de saída deste mesmo estado, quando as condições de estado de equilíbrio existem. Tais procedimentos de balanço de fluxo são conhecidos como *balanço estocástico*. Para reforçar, considere um estado n (n>=1) em uma fila M/M/1. Então um diagrama de estado pode ser construído como na figura 2.1.

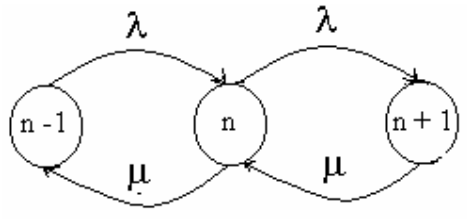

Figura 2.1 Diagrama de Transição de Taxa – M/M/1.

Neste diagrama, do estado n, o sistema vai para o estado n - 1 se um serviço é completado ou para n + 1 se uma chegada ocorre. O sistema pode ir para o estado n de n – 1 se uma chegada ocorre ou para o estado n de n + 1 se um serviço é completado. As taxas de fluxos são em termos de λ e μ, de modo que, o fluxo total médio para fora do estado n é igual a $\lambda p_n + \mu p_n$, enquanto o fluxo total médio para dentro do estado n é igual a $\lambda p_{n-1} + \mu\, p_{n+1}$.

Pode-se olhar p_n como a porcentagem de tempo no estado de equilíbrio que o sistema permanece no estado n. As equações 2.1 são chamadas de equações de balanço *global*, desde que, elas equacionam o fluxo total médio de entrada no estado com o fluxo total médio de saída deste mesmo estado. Contudo, numa análise mais além, pode-se referir a uma equação de balanço *detalhada* informando o fluxo médio dentro e fora de estados adjacentes. Para ilustrar, coloque uma linha artificial na figura 2.1 entre os estados n e n + 1. O fluxo médio através desta barreira deve ser também igual a 0 para condições de estado de equilíbrio, isto é, o fluxo para a direita deve ser igual ao fluxo para a esquerda. Isto resulta que a equação $\lambda p_n = \mu p_{n+1}$. Da mesma forma, se for colocada uma barreira entre os estados n e n-1, obtém-se $\lambda p_{n-1} = \mu p_n$. Note que somando as duas equações resulta nas equações de balanço global de (2.1a).

Esta análise de balanço de fluxo entre dois estados trabalha neste caso por causa da característica de nascimento e morte, onde só estados adjacentes podem diretamente se comunicar, com taxas de transição constantes. Para modelos Markovianos mais gerais, isto não é necessariamente verdadeiro, de modo que, a análise de balanço detalhada pode ser muito mais complicada.

2.2 Métodos de Solução de Equações Diferenciais de Estado de Equilíbrio

2.2.1 Método Iterativo de Solução de Equações Diferenciais de Estado de Equilíbrio para {pn}

Com isso quer-se dizer que utilizando-se iterativamente as equações de balanço global dada na equação (2.1) ou as equações de balanço local (detalhada) obtém-se uma seqüência de probabilidades de estado, p_1, p_2, p_3, ..., cada uma em termos de p_0. Assim, desde que o sistema M/M/1 é realmente uma processo de nascimento e morte com taxas de nascimento e morte constantes, pode-se diretamente aplicar a equação (1.19) com $\lambda_n = \lambda$ e $\mu_n = \mu$ para todo n. Disto segue que

$$p_n = p_0 \prod_{i=1}^{n} \lambda/\mu \qquad (n >= 1)$$

$$= p_0(\lambda/\mu)^n, \qquad (2.2)$$

Para obter p_0, utiliza-se do fato de que a soma das probabilidades deve ser igual a 1, ou seja,

$$1 = \sum_{n=0}^{\infty} (\lambda/\mu)^n p_0,$$

Chamando λ/μ de ρ para filas de servidor simples, pode-se rescrever a equação acima da seguinte maneira

$$p_0 = 1 / \sum_{n=0}^{\infty} \rho^n$$

Agora $\sum_{n=0}^{\infty} \rho^n$ é a série geométrica $1 + \rho + \rho^2 + \rho^3 + \ldots$ e converge, se e somente se, $\rho < 1$. Assim, para a existência de uma solução de estado de equilíbrio, $\rho = \lambda/\mu$ deve ser menor que 1, ou equivalentemente, λ deve ser menor μ. Isto faz sentido, pois, se $\lambda > \mu$, a taxa média de chegada é maior que a taxa média de serviço, de maneira que, o servidor não conseguirá atender aos clientes na mesma proporção em que eles chegam, fazendo com que ocorra a formação de filas. Já para $\lambda = \mu$, não existe nenhuma resposta intuitiva, pois, nenhuma solução de estado de equilíbrio existe quando $\lambda = \mu$,

desde que, é improvável que as chegadas ocorram exatamente nos tempos em que os clientes em serviço terminam de serem atendidos.

Fazendo uso de uma expressão bem conhecida para a soma dos termos de uma progressão geométrica,

$$\sum_{n=0}^{\infty}\rho^n = 1 / (1 - \rho) \qquad (\rho < 1),$$

tem-se que

$$p_0 = (1 - \rho) \qquad (\rho = \lambda/\mu < 1). \tag{2.3}$$

Assim, a solução de estado de equilíbrio completa para um sistema M/M/1 é a função de probabilidade geométrica

$$p_n = (1 - \rho)\rho^n \qquad (\rho = \lambda/\mu < 1). \tag{2.4}$$

2.2.2 Medidas de Efetividade

As distribuições de probabilidade de estado de equilíbrio para o tamanho do sistema permiti-nos calcular as medidas de efetividade do sistema. Duas de interesse imediato são o número esperado no sistema e o número esperado na fila em um estado de equilíbrio. Para derivar isto, seja N a variável aleatória "número de clientes no sistema em estado de equilíbrio" e L represente seu valor esperado. Assim,

$$L = E[N] = \sum_{n=0}^{\infty} np_n$$

$$= (1 - \rho) \sum_{n=0}^{\infty} n\rho^n \tag{2.5}$$

Considerando o somatório

$$\sum_{n=0}^{\infty} n\rho^n = \rho + 2\rho^2 + 3\rho^3 + \ldots$$

$$= \rho(1 + 2\rho + 3\rho^2 + \ldots)$$

$$= \rho\sum_{n=1}^{\infty} n\rho^{n-1}.$$

Pela equação acima, observa-se que $\sum_{n=1}^{\infty} n\rho^{n-1}$ é simplesmente a derivada do $\sum_{n=1}^{\infty}\rho^n$ com respeito a ρ, dado que o somatório e operações de diferenciação podem ser trocados. Como

$$\sum_{n=0}^{\infty}\rho^n = 1/(1-\rho),$$

consequentemente,

$$\sum_{n=1}^{\infty}n\rho^{n-1} = d[1/(1-\rho)]/d\rho = 1/(1-\rho)^2.$$

Assim, o número esperado no sistema em estado de equilíbrio é então

$$L = (\rho(1-\rho)) / (1-\rho)^2 = \rho/(1-\rho) = \lambda/(\mu-\lambda). \qquad (2.6)$$

Agora, adotando a variável aleatória "número na fila em estado de equilíbrio" como N_q e seu valor esperado por L_q, tem-se pela fórmula de Little apresentada anteriormente que

$$L_q = \sum_{n=1}^{\infty}(n-1)p_n = \sum_{n=1}^{\infty}np_n - \sum_{n=1}^{\infty}p_n$$

$$= L - (1-p_0) = \rho/(1-\rho) - \rho.$$

Isto porque, $\sum_{n=0}^{\infty}p_n = 1$, consequentemente, $\sum_{n=1}^{\infty}p_n = 1 - p_0$.

Note que $L_q = L - (1-p_0)$ se mantém para filas com um único servidor onde os clientes chegam um de cada vez. Assim, o comprimento médio da fila é

$$L_q = \rho^2/(1-\rho) = \lambda^2/(\mu(\mu-\lambda)). \qquad (2.7)$$

Existe também o interesse em conhecer o tamanho da fila esperado onde a fila nunca está vazia, que será denotado aqui por L'_q . Assim, pode-se escrever

$$L'_q = E[N_q | N_q \neq 0]$$

$$= \sum_{n=1}^{\infty}(n-1)p'_n = \sum_{n=2}^{\infty}(n-1)p'_n,$$

onde p'_n é a distribuição de probabilidade condicional de n no sistema dado que a fila não está vazia, ou $p'_n = \Pr\{n \text{ no sistema} \mid n >= 2\}$. Das leis de probabilidade condicional,

$$p'_n = \frac{\Pr\{n \text{ no sistema e } n >= 2\}}{\Pr\{n >= 2\}}$$

$$= p_n / \sum_{n=2}^{\infty} p_n \qquad (n>=2)$$

$$= \frac{p_n}{1-(1-\rho)-(1-\rho)\rho}$$

$$= p_n / \rho^2,$$

pois, se $\sum_{n=0}^{\infty} np_n = (1 - \rho)\sum_{n=0}^{\infty} n\rho^n$ (2.5), então, retirando-se o multiplicador n das duas parcelas, tem-se que $\sum_{n=2}^{\infty} p_n = (1 - \rho)\sum_{n=2}^{\infty} \rho^n$, menos as parcelas referentes a n = 0, que é igual a (1 - ρ) e n = 1, que é igual a (1 - ρ)ρ. Desse modo, a distribuição de probabilidade $\{p'_n\}$ é a distribuição de probabilidade $\{p_n\}$ normalizada por omitir os casos n = 0 e 1. Assim,

$$L'_q = \sum_{n=2}^{\infty} (n-1)\,(p_n / \rho^2)$$

$$= \frac{L - p_1 - (1 - p_0 - p_1)}{\rho^2},$$

Consequentemente,

$$L'_q = 1/(1-\rho) = \mu/(\mu - \lambda). \qquad (2.8)$$

Note que a generalização para Pr{n>=2} é intuitiva, de maneira que, $\Pr\{N>=n\} = \rho^n$

Para completar a parte básica desta apresentação, pode-se utilizar das equações de Little apresentadas anteriormente que o tempo de espera W e o atraso na fila W_q, dadas as condições de estado equilíbrio, para uma fila M/M/1 são

$$W = L / \lambda = \frac{\rho}{\lambda(1-\rho)} = 1/(\mu - \lambda) \qquad (2.9)$$

$$W_q = L_q / \lambda = \frac{\rho^2}{\lambda(1-\rho)} = \rho/(\mu - \lambda) \qquad (2.10)$$

Para exemplificar todas as medidas de efetividade apresentadas até o momento observe o exemplo abaixo:

Exemplo 2.1 – A Sra. Cutt possui uma salão unissex. Ela não marca hora, de forma que, o atendimento ocorre numa disciplina FCFS. Ela acredita que está muito ocupada aos sábados pela manhã, de modo que, ela está estudando a possibilidade de contratar uma assistente e possivelmente mudar-se para um estabelecimento maior. Entretanto, antes de tomar essa decisão ela deseja saber qual o número médio de clientes no salão e qual o número médio de clientes esperando para serem atendidos.

Para tanto, a Sra. Cutt observou que os clientes chegavam ao salão de acordo com um processo de Poisson com uma taxa média de chegada de 5/h. Além disso, foi observado que o tempo de atendimento era exponencialmente distribuído com

uma taxa média de 10 minutos. De posse destes dados, tem-se que $\lambda = 5/h$ e $\mu = 1/10$ min = 6/h. Isto dá um ρ de 5/6. De (2.6) e (2.7), a Sra. Cutt encontra que L = 5 e $L_q = 4^{1/6}$. O número médio esperando quando existe pelo menos uma pessoa esperando é obtido de (2.8) como $L'_q = 6$. Ela também está interessada em saber a porcentagem de tempo que os clientes são atendidos sem terem que esperar em fila, ou seja, a probabilidade que uma chegada ocorra e o sistema esteja vazio. A probabilidade disto é $p_0 = 1 - \rho = 1/6$. Desse modo, 16.7% do tempo Cutt está ociosa, fazendo com que o próximo cliente a chegar possa ser atendido sem ter que esperar. Devido ao fato do processo de Poisson governar sua chegada e por causa de sua propriedade completamente aleatória a porcentagem de clientes que entrarão em serviço diretamente sem ter que esperar é 16.7% também. Por outro lado, 83.3 % dos clientes devem esperar antes de serem atendidos.

A sala de espera de Cutt tem quatro assentos, de forma que, ela está interessada em saber a probabilidade de um cliente, após chegar, não ter lugar para sentar. Isto pode ser facilmente calculado como

Pr{encontrar nenhum assento} = Pr{N>=5} = ρ^5 = 0.402. Isto implica que em pouco mais de 40% do tempo, os clientes não encontrarão lugar para sentar.

Por fim, utilizando-se as equações (2.9) e (2.10) obtém-se que o tempo de atendimento no salão e o atraso na fila são, respectivamente, W = 1h e W_q = 5/6h, resultados não muito agradáveis.

2.2.2.1 Distribuições de Tempo de Espera

Apesar da Sra. Cutt já possuir informações que justifiquem a ampliação de seu negócio, ela não tem condições de responder, dadas as condições atuais, qual a probabilidade que o atraso na fila seja superior a 45 minutos. Dessa forma, as seguintes formulações matemáticas devem ser levantadas para atender a tal requisito de informação.

Seja T_q a variável aleatória "tempo gasto esperando na fila" e $W_q(t)$ represente sua distribuição de probabilidade cumulativa. Até agora, a disciplina de filas não tem tido nenhum efeito em nossas derivações. Contudo, ao levar em consideração tempos de espera individuais, a disciplina deve ser especificada, sendo neste caso, assumido que ela é do tipo FCFS. A variável aleatória tempo de espera tem uma propriedade interessante em que ela é parte discreta e parte contínua. O tempo de espera

é, para maior parte do tempo, uma variável contínua, exceto que existe uma probabilidade não zero que o atraso seja zero, isto é, um cliente entrando em serviço imediatamente após sua chegada. Assim, tem-se que

$$W_q(0) = Pr\{T_q <= 0\} = Pr\{T_q = 0\} = Pr\{\text{sistema vazio na chegada}\} = q_0.$$

Pode-se denotar a probabilidade condicional de n no sistema dado que um chegada está próxima a ocorrer por q_n. Entretanto, estas probabilidades nem sempre são iguais a probabilidade p_n com o qual se está trabalhando. Isto porque, p_n são probabilidade incondicionais de n em um ponto qualquer no tempo. Para encontrar a distribuição do tempo de espera *virtual*, ou seja, o tempo que o cliente teria que esperar caso chegasse num tempo qualquer, será utilizado p_n. Contudo, para a entrada de Poisson, $q_n = p_n$. Assim,

$$W_q(0) = p_0 = 1 - \rho.$$

Resta, portanto, encontrar, $W_q(t)$ para t >0.

Considere $W_q(t)$, a probabilidade de um cliente esperar um tempo menor ou igual a t para ser atendido. Se existem n clientes no sistema na chegada de um cliente, a ordem para que o cliente entre em serviço no tempo entre 0 e t é que todos os n clientes existentes no sistema tenham que ser servidos até o tempo t. Desde que a distribuição de serviço é markoviana, a distribuição do tempo requerido para os n clientes serem atendidos é independente do tempo da chegada atual e a convolução de n variáveis aleatórias exponenciais é do tipo n-Erlang. Complementando, desde que a entrada é Poisson, os pontos de chegada são uniformemente espaçados e consequentemente a probabilidade que uma chegada encontre n no sistema é idêntica a distribuição estacionária do tamanho do sistema. Portanto, pode-se escrever que

$$W_q(t) = Pr\{T_q <= t\} = W_q(0) + \sum_{n=1}^{\infty} Pr\{\text{n clientes atendidos} <= t \mid \text{chegada encontra n no sistema}\}.\ p_n.$$

$$= 1 - \rho + (1 - \rho)\sum_{n=1}^{\infty} \rho^n \int_o^t \frac{\mu(\mu x)^{n-1}}{(n-1)!} e^{-\mu x}\, dx$$

$$= 1 - \rho + \rho(1 - \rho)\int_o^t \mu e^{-\mu x} \sum_{n=1}^{\infty} \frac{(\mu x \rho)^{n-1}}{(n-1)!}\, dx$$

$$= 1 - \rho + \rho(1 - \rho)\int_o^t \mu e^{-\mu x(1 - \rho)}\, dx$$

$$= 1 - \rho e^{-\mu(1-\rho)t} \qquad (t > 0).$$

Assim, a distribuição de tempo de espera em fila é

$$W_q(t) = 1 - \rho e^{-\mu(1-\rho)t} \qquad (t >= 0) \tag{2.11}$$

Com o resultado de (2.11), a Sra. Cutt é capaz de calcular a probabilidade que um cliente chegando ter que esperar mais que 45 minutos como $5/6\ e^{-3/4} = 0.3936$.

Partindo do que foi demonstrado acima, pode-se definir o tempo total gasto pelo cliente num sistema M/M/1 (incluindo tempo de serviço) da seguinte maneira. Seja T a variável aleatória que representa o tempo total do cliente no sistema com W(t), w(t) e W sendo, respectivamente, a distribuição cumulativa, a função densidade e o valor esperado de T. Além disso, pode ser mostrado que a espera no sistema tem uma distribuição exponencial negativa, tal que,

$$\begin{aligned} W(t) &= 1 - e^{-(\mu-\lambda)t} \qquad (t >= 0) \\ w(t) &= (\mu - \lambda)e^{-(\mu-\lambda)t} \qquad (t > 0) \end{aligned} \tag{2.12}$$

2.3 Filas com canais paralelos (M/M/c)

Agora nossa atenção será direcionada para modelos multiservidores M/M/c no qual cada servidor possui uma distribuição de tempo de serviço exponencial distribuída identicamente e independentemente, com o processo de chegada de novo assumido ser Poisson. Dada as condições de que a entrada é Poisson, o serviço é exponencial e o número de servidores é igual a c, tem-se um processo de nascimento e morte. Assim, $\lambda_n = \lambda$ para todo n, faltando, portanto, determinar μ_n para que se possa utilizar a fórmula 2.2 aplicada a processos de nascimento e morte.

Se existem mais que c clientes no sistema, todos os c servidores devem estar ocupados com uma taxa média de serviço μ, com a taxa média de saída do sistema sendo igual a $c\mu$. Quando existem menos que c clientes no sistema, $n < c$, por exemplo, só n dos c servidores estão ocupados, fazendo com que o sistema passe a ter uma taxa média de saída de $n\mu$. Portanto, μ_n pode ser escrito como

$$\mu_n = \begin{cases} n\mu & (1<=n<c) \\ c\mu & (n>=c). \end{cases} \tag{2.13}$$

Utilizando (2.13) na equação (2.2) e o fato que $\lambda_n = \lambda$ para todo n, obtém-se

$$p_n = \begin{cases} \dfrac{\lambda^n}{n!\mu^n} p_0 & (1<=n<c), \\ \dfrac{\lambda^n}{c^{n-c}c!\mu^n} p_0 & (n>=c). \end{cases} \tag{2.14}$$

Para encontrar p_0, será levado em consideração o fato que as probabilidades devem somar um, que implica que

$$p_0 = [\sum_{n=0}^{c-1} \lambda^n/(n!\mu^n) + \sum_{n=c}^{\infty} \lambda^n/(c^{n-c}c!\mu^n)]^{-1}.$$

Adotando $r = \lambda/\mu$ e $\rho = r/c = \lambda/c\mu$, tem-se que

$$p_0 = [\sum_{n=0}^{c-1} r^n/n! + \sum_{n=c}^{\infty} r^n/(c^{n-c}c!)]^{-1}.$$

Fazendo uso de uma expressão bem conhecida para a soma dos termos de uma progressão geométrica,

$$\sum_{n=0}^{\infty} \rho^n = 1 / (1 - \rho) \qquad (\rho < 1),$$

tem-se que

$$\sum_{n=c}^{\infty} r^n/(c^{n-c}c!) = (r^c/c!) \sum_{n=c}^{\infty} (r/c)^{n-c} = (r^c/c!) \sum_{n=0}^{\infty} (r/c)^m =$$

$$= (r^c/c!)\frac{1}{(1 - \rho)} \qquad (r/c = \rho < 1).$$

Finalmente, pode-se escrever que

$$p_0 = (\sum_{n=0}^{c-1} r^n/n! + \frac{r^c}{c!(1 - \rho)})^{-1} \quad (r/c = \rho < 1) \qquad (2.15)$$

Note que a condição para a existência de uma solução de estado de equilíbrio aqui é $\lambda/(c\mu) < 1$; isto é, taxa média de chegada deve ser menor que a taxa média potencial máxima de serviço do sistema, que é intuitivamente o que seria esperado. Uma outra coisa que pode ser notada da equação (2.15) é que quando c = 1, (2.15) reduz-se para a equação do modelo M/M/1 mostrado anteriormente.

Baseando-se nas probabilidades de estado de equilíbrio dada nas equações (2.14) e (2.15) pode-se derivar medidas de efetividade para o modelo M/M/c, de maneira similar as utilizadas no modelo M/M/1.

Primeiro será considerado o tamanho da fila esperado L_q, haja visto que esta medida é computacionalmente mais fácil de determinar que L, desde que, só é necessário tratar-se com p_n para n >= c. Assim,

$$L_q = \sum_{n=c+1}^{\infty} (n-c)\, p_n = \sum_{n=c+1}^{\infty} (n-c)(r^n/c^{n-c}c!)\, p_0$$

$$= (r^c/c!)p_0 \sum_{n=c+1}^{\infty}(n-c)(r^{n-c}/c^{n-c}),$$

substituindo-se n-c por m, tem-se que

$$L_q = (r^c/c!)p_0\sum_{n=1}^{\infty}m\rho^m = (r^c/c!)\rho p_0\sum_{n=1}^{\infty}m\rho^{m-1} =$$

$$= (r^c/c!)\rho p_0 \frac{d}{d\rho}\left(\sum_{n=1}^{\infty}\rho^m\right) = (r^c/c!)\rho p_0 \frac{d}{d\rho}(1/(1-\rho) - 1)$$

$$= \frac{r^c\rho p_0}{c!(1-\rho)^2}.$$

Assim,

$$L_q = \frac{(r^c\rho)}{c!(1-\rho)^2} p_0 \qquad (2.16)$$

Para encontrar L agora, emprega-se a fórmula de Little para obter W_q, então utiliza-se W_q para encontrar $W = W_q + 1/\mu$, e finalmente emprega-se novamente a fórmula de Little para calcular $L = \lambda W$. Assim, obtém-se que

$$W_q = L_q/\lambda = \frac{(r^c)}{c!(c\mu)(1-\rho)^2} p_0 \qquad (2.17)$$

desde que, $\rho = \lambda/(c\mu)$.

$$W = 1/\mu + \frac{(r^c)}{c!(c\mu)(1-\rho)^2} p_0 \qquad (2.18)$$

e

$$L = r + \frac{(r^c\rho)}{c!(1-\rho)^2} p_0 \qquad (2.19)$$

Note que o resultado final para L poderia ser obtido diretamente L_q por utilizar $L = L_q + r$.

Embora as distribuições de probabilidade de tempos de espera, W(t) e $W_q(t)$, são sejam necessárias para obter W e W_q, para que questões a respeito das probabilidades de uma espera serem maiores que uma determinada quantidade possam ser respondidas, faz-se necessário a utilização dessas distribuições.

Dessa forma, definindo T_q como a variável aleatória "tempo gasto esperando na fila" e $W_q(t)$ sua FDC, tem-se que

$$W_q(0) = \Pr\{T_q = 0\} = \Pr\{<= c-1 \text{ no sistema}\} =$$

$$= \sum_{n=0}^{c-1}p_n = p_0\sum_{n=0}^{c-1}r^n/n!.$$

Agora para avaliar o $\sum r^n/n!$, tem-se da expressão (2.15) para p_0 que

$$= \sum_{n=0}^{c-1} r^n/n! = [1/p_0] - [r^c/(c!(1-\rho))].$$

Dessa forma, tem-se que

$$W_q(0) = p_0\left(\frac{1}{p_0} - \frac{r^c}{c!(1-\rho)}\right) = 1 - \frac{r^c p_0}{c!(1-\rho)}) \qquad (2.20)$$

Para $T_q > 0$ e assumindo uma disciplina FCFS,

$$W_q(t) = \Pr\{T_q <= t\} = W_q(0) + \sum_{n=c}^{\infty} \Pr\{n - c + 1 \text{ clientes atendidos} <= t \mid \text{chegada encontra n no sistema}\}.\ p_n.$$

$$= W_q(0) + p_0 \sum_{n=c}^{\infty} \frac{r^n}{c^{n-c}c!} \int_o^t \frac{c\mu(c\mu x)^{n-c}}{(n-c)!} e^{-c\mu x}\, dx$$

Cortando-se c^{n-c} de dentro da integral com c^{n-c} que está dividindo r^n ,cortando –se o c que multiplica $\mu(c\mu x)^{n-c}$ com o c correspondente de c! tem-se que

$$= W_q(0) + p_0 \sum_{n=c}^{\infty} \frac{r^n}{(c-1)!} \int_o^t \frac{\mu(\mu x)^{n-c}}{(n-c)!} e^{-c\mu x}\, dx$$

Agora, retirando-se c parcelas de r^n ($r^{n-c)}$) e passando-as para dentro da integral, com a troca do somatório com a integral, pode-se retirar $r^c/(c-1)!$ do somatório e da integral já que estes são constantes para equação e não interferiram no somatório. Assim,

$$= W_q(0) + \frac{r^c p_0}{(c-1)!} \int_o^t \mu e^{-c\mu x} \sum_{n=c}^{\infty} \frac{(\mu r x)^{n-c}}{(n-c)!}\, dx$$

Contudo, já foi visto anteriormente que $\sum_{n=c}^{\infty} \frac{(\mu r x)^{n-c}}{(n-c)!}\, dx = e^{\mu x r}$. Logo,

$$= W_q(0) + \frac{r^c p_0}{(c-1)!} \int_o^t \mu e^{-c\mu x} e^{\mu x r}\, dx = W_q(0) + \frac{r^c p_0}{(c-1)!} \int_o^t \mu e^{-\mu x(c-r)}\, dx$$

$$= W_q(0) + \frac{r^c p_0}{c!(1-\rho)} (1 - e^{-(c\mu - \lambda)t}).$$

Substituindo $W_q(0)$ de (2.20) tem-se que

$$W_q(t) = 1 - \frac{r^c p_0}{c!(1-\rho)} e^{-(c\mu - \lambda)t}). \qquad (2.21)$$

De (2.21) nota-se que

$$\Pr\{ T_q > t\} = 1 - W_q(t) = \frac{r^c p_0}{c!(1-\rho)} e^{-(c\mu - \lambda)t}).$$

de modo que, a probabilidade condicional $\Pr\{ T_q > t \mid T_q > 0\} = e^{-(c\mu - \lambda)t}$).

Para encontrar a fórmula para a FDC do tempo de espera no sistema, deve-se dividir a situação em duas possibilidades separadas: aquelas cujos clientes não tiveram nenhum atraso na fila [probabilidade $W_q(0)$]; e aquelas onde a espera no sistema é o atraso na fila mais o tempo de serviço [probabilidade $1 - W_q(0)$]. A primeira dessas duas classes de clientes tem uma FDC que é idêntica a distribuição de tempo de serviço exponencial, com média $1/\mu$; a segunda tem uma FDC obtida da convolução da distribuição do tempo de serviço com uma segunda distribuição exponencial tendo média $1/(c\mu - \lambda)$. Esta convolução também pode ser escrita como a diferença de duas funções exponenciais,

$$\Pr\{ T_q <= t\} = \frac{c(1-\rho)}{c(1-\rho) - 1}(1 - e^{-\mu t}) - \frac{1}{c(1-\rho) - 1}(1 - e^{-(c\mu - \lambda)t}).$$

Assim, a FDC da espera no sistema M/M/c pode ser escrita como

$$W(t) = W_q(0)[1 - e^{-\mu t}] + [1 - W_q(0)]$$

$$X \left[\frac{c(1-\rho)}{c(1-\rho) - 1}(1 - e^{-\mu t}) - \frac{1}{c(1-\rho) - 1}(1 - e^{-(c\mu - \lambda)t})\right]$$

$$= \frac{c(1-\rho) - W_q(0)}{c(1-\rho) - 1}(1 - e^{-\mu t}) - \frac{1 - W_q(0)}{c(1-\rho) - 1}(1 - e^{-(c\mu - \lambda)t}).$$

Exemplo 2.2 – Uma clínica de olhos da cidade oferece consultas grátis todas as quartas a tarde. Existem três oftalmologistas no prédio. Um teste, leva em média, 20min, e o tempo atual é achado ser aproximadamente exponencialmente distribuído ao redor desta média. Os clientes chegam de acordo com um processo de Poisson com média de 6/h, sendo que os pacientes são atendidos numa base FCFS. Os planejadores do hospital estão interessados em saber: (1) qual é o número médio de pessoas esperando; (2) a quantidade média de tempo que um paciente gasta na clínica; e (3) a porcentagem média de tempo ocioso de cada um dos oftalmologistas. Assim, deseja-se saber L_q, W e a porcentagem de tempo ocioso de um servidor.

Para responder a estas questões será calculado inicialmente p_0, desde que, ele aparece em todas as fórmulas derivada das medidas de efetividade. Dessa forma, tem-se que $c = 3$, $\lambda = 6/h$ e $\mu = 1/(20min) = 3/h$. Assim $r = \lambda/\mu = 2$, $\rho = 2/3$, e de (2.15) tem-se que

$$p_0 = (1 + 2 + 2^2/2! + \frac{2^3}{3!(1-2/3)})^{-1} = 1/9$$

Da equação (2.16) tem-se que

$$L_q = \frac{(2^3 2/3)}{3!(1-2/3)^2} 1/9 = 8/9,$$

e das equações (2.16) e (2.18) que

$$W = 1/\mu + L_q/\lambda = 1/3 + (8/9)/6 = 13/27h = 28.9min.$$

Para o último item, já foi visto (tabela 1.2) que a fração média de tempo ocioso de qualquer servidor em um sistema M/M/c é igual a 1 - ρ. Para este problema, portanto, casa oftalmologista está ocioso 1/3 do tempo, desde que a intensidade tráfego é ρ = 2/3. Dados os três servidores no prédio, dois deles estarão ocupado em qualquer tempo (em média), uma vez que r = 2. Além disso, a fração de tempo que existe pelo menos uma oftalmologista ocioso pode ser calculada aqui como $p_0 + p_1 + p_2 =$

$$\Pr\{W_q = 0\} = 1 - \frac{r^c p_0}{c!(1-\rho)} = 1 - \frac{2^3(1/9)}{3!(1-2/3)} = 1 - 4/9 = 5/9.$$

2.4 Filas com Canais Paralelos e Limitação na Capacidade do Sistema (M/M/c/K)

Nesta seção o estudo será feito em cima de um modelo de nascimento e morte com servidor paralelo M/M/c/K, na qual existe um limite K colocado no número permitido no sistema em qualquer tempo. A abordagem aqui é idêntica para modelos com capacidade infinita (M/M/c), exceto que a taxa de chegada λ_n agora deve ser 0 sempre que n >= K. Então segue da equação (2.14) que as probabilidades de tamanho do sistema em estado de equilíbrio são dadas por

$$p_n = \begin{cases} \dfrac{\lambda^n}{n!\mu^n} p_0 & (1<=n<c), \\ \dfrac{\lambda^n}{c^{n-c}c!\mu^n} p_0 & (c<= n <= K). \end{cases} \qquad (2.22)$$

Da condição de que as probabilidades devem somar 1 resultará p_0. Novamente, a computação é idêntica a utilizada no modelo M/M/c, exceto que agora ambas as séries na computação são finitas e assim não existira nenhum requerimento que a intensidade do tráfego ρ seja menor que 1. Assim,

$$p_0 = [\sum_{n=0}^{c-1} \lambda^n/(n!\mu^n) + \sum_{n=c}^{K} \lambda^n/(c^{n-c}c!\mu^n)]^{-1}.$$

Para simplificar, considere o segundo somatório acima, com $r = \lambda/\mu$ e $\rho = r/c$:

$$\sum_{n=c}^{K} r^n/(c^{n-c}c!) = (r^c/c!)\sum_{n=c}^{K}\rho^{n-c}$$

$$= \begin{cases} \dfrac{r^c}{c!}\,\dfrac{1-\rho^{k-c+1}}{1-\rho} & (\rho \neq 1), \\[2ex] \dfrac{r^c}{c!}(K-c+1) & (\rho = 1). \end{cases}$$

Assim,

$$p_0 = \begin{cases} \left(\sum_{n=0}^{c-1}(r^n/n!) + (r^c/c!)\,\dfrac{1-\rho^{k-c+1}}{1-\rho}\right)^{-1} & (\rho \neq 1), \\[2ex] \left(\sum_{n=0}^{c-1}(r^n/n!) + (r^c/c!)\,(K-c+1)\right)^{-1} & (\rho = 1). \end{cases} \qquad (2.23)$$

O próximo passo será encontrar o comprimento da fila esperado ($\rho \neq 1$):

$$L_q = \sum_{n=c+1}^{K}(n-c)p_n = (r^c/c!)p_0\sum_{n=c+1}^{K}(n-c)(r^{n-c}/c^{n-c})$$

$$= (r^c/c!)p_0\rho\sum_{n=c+1}^{K}(n-c)\rho^{n-c-1}$$

Substituindo-se n-c por i tem-se que

$$= \frac{p_0 r^c \rho}{c!}\sum_{i=1}^{K-c} i\rho^{i-1} = \frac{p_0 r^c \rho}{c!}\,\frac{d}{d\rho}\left(\frac{1-\rho^{K-c+1}}{1-\rho}\right),$$

ou

$$= \frac{p_0 r^c \rho}{c!\,(1-\rho)^2}\left[1-\rho^{K-c+1} - (1-\rho)(K-c+1)\,\rho^{K-c}\right]. \qquad (2.24)$$

Para $\rho = 1$, é necessário empregar a regra de L'Hôpital duas vezes, de forma que, para os propósitos deste curso não será demonstrado.

Para obter o tamanho esperado do sistema, será utilizado a fórmula de Little $L = L_q + r$ aplicada ao modelo M/M/c. Entretanto, para o caso do espaço de espera finito, será necessário ajustar a fórmula de Little, haja visto que existe uma fração p_k de chegadas que não se juntam ao sistema, pois, estas chegadas ocorreram quando não existe nenhum espaço disponível na fila. Assim, a taxa efetiva de chegada vista pelos

servidores é $\lambda(1 - p_k)$, de modo que, a esta taxa de ajuste para a entrada será denotada por λ_{eff}. O relacionamento entre L e L_q deve, portanto, ser restruturado para este modelo como $L = L_q + \lambda_{eff}/\mu = L_q + \lambda(1 - p_k)/\mu = L_q + r(1 - p_k)$. Sabe-se que a quantidade $r(1 - p_k)$ deve ser menor que c, desde que o número médio de clientes em serviço deve ser menor que o número total de servidores disponíveis. Isto sugere a definição de alguma coisa, tal como, $\rho_{eff} = \lambda_{eff}/\mu$, que deveria ser menor que 1 para qualquer modelo M/M/c mesmo que nenhuma restrição exista nos valores de $\rho = \lambda/c\mu$.

Os valores esperados para os tempos de espera podem ser prontamente obtidos das fórmulas de Little como

$$W = \frac{L}{\lambda_{eff}} = \frac{L}{\lambda(1 - p_k)}, \tag{2.25}$$

$$W_q = W - 1/\mu = L_q/\lambda_{eff}$$

Para modelos M/M/1/K, todas as medidas de efetividade reduzem-se para expressões consideravelmente mais simples, como a seguir

$$p_0 = \begin{cases} \dfrac{1 - \rho}{1 - \rho^{k+1}} & (\rho \neq 1), \\ 1/(K+1) & (\rho = 1). \end{cases} \tag{2.26}$$

$$p_n = \begin{cases} \dfrac{(1 - \rho)\rho^n}{1 - \rho^{k+1}} & (\rho \neq 1), \\ 1/(K+1) & (\rho = 1). \end{cases} \tag{2.27}$$

e

$$L_q = \begin{cases} \dfrac{\rho}{1 - \rho} - \dfrac{\rho(K\rho^k + 1)}{1 - \rho^{k+1}} & (\rho \neq 1), \\ \dfrac{K(K-1)}{2(K + 1)} & (\rho = 1). \end{cases} \tag{2.28}$$

com $L = L_q + (1 - p_0)$. Note que este relacionamento final implica que $1 - p_0 = \lambda(1 - p_k)/\mu$, de modo que, esta equação pode ser rescrita na forma $\mu(1 - p_0) = \lambda(1 - p_k)$ indicando que a taxa efetiva de saída do sistema deve ser igual a sua taxa efetiva de entrada.

A derivação da FDC do tempo de espera é um pouco complicada, desde que as séries são finitas, embora elas possam ser expressas em termos de somas de Poisson cumulativas, como será mostrado agora. Antes de mais nada é necessário derivar as

probabilidades de ponto de chegada $\{q_n\}$, desde que a entrada não é mais Poisson devido a limitação K no sistema e $q_n \neq p_n$.

Para determinar q_n será utilizado o teorema de Baye, de forma que

$$q_n \equiv \Pr\{\text{n no sistema} \mid \text{chegada acabou de ocorrer}\} =$$

$$= \frac{\Pr\{\text{chegada ocorrer} \mid \text{n no sistema}\}p_n}{\sum_0^k \Pr\{\text{chegada ocorrer} \mid \text{n no sistema}\}p_n} =$$

$$= \lim_{\Delta t \to 0} \frac{p_n[\lambda\Delta t + o(\Delta t)]}{\sum_{n=0}^{K-1} p_n[\lambda\Delta t + o(\Delta t)]}$$

$$= \frac{\lambda p_n}{\lambda \sum_{n=0}^{K-1} p_n} = p_n / 1 - p_k \qquad (n <= K - 1).$$

Utilizando-se a mesma formulação apresentada em 2.21 para $W_q(t)$ tem-se que

$$W_q(t) = \Pr\{T_q <= t\} = W_q(0) + \sum_{n=c}^{K-1} \Pr\{n - c + 1 \text{ serviços completados em } <= t \mid \text{a chegada encontrou n no sistema}\}.q_n$$

desde que não podem existir chegadas unindo-se a o sistema depois de ter encontrado K no sistema, segue que

$$W_q(t) = W_q(0) + \sum_{n=c}^{K-1} q_n \int_o^t \frac{c\mu(c\mu x)^{n-c}}{(n-c)!} e^{-c\mu x}\, dx$$

$$= W_q(0) + \sum_{n=c}^{K-1} q_n \left[1 - \int_t^\infty \frac{c\mu(c\mu x)^{n-c}}{(n-c)!} e^{-c\mu x}\, dx \right]$$

Agora, é conhecido que

$$\int_t^\infty \frac{\lambda(\lambda x)^m}{m!} e^{-\lambda x}\, dx = \sum_{i=0}^{M} \frac{(\lambda t)^i e^{-\lambda t}}{i!}$$

Assim, tomando $m = n - c$ e $\lambda = c\mu$ obtém-se

$$\int_t^\infty \frac{c\mu(c\mu x)^{n-c}}{n-c!} e^{-c\mu x}\, dx = \sum_{i=0}^{N-c} \frac{(c\mu t)^i e^{-c\mu t}}{i!}$$

e conseqüentemente

$$W_q(t) = W_q(0) + \sum_{n=c}^{K-1} q_n - \sum_{n=c}^{K-1} q_n \sum_{i=0}^{N-c} \frac{(c\mu t)^i e^{-c\mu t}}{i!}$$

$$= 1 - \sum_{n=c}^{K-1} q_n \sum_{i=0}^{N-c} \frac{(c\mu t)^i}{i!} e^{-c\mu t}.$$

Exemplo 2.3 – Considere uma estação de inspeção de emissão de gases poluentes de automóveis com três boxes de inspeção, onde cada boxe tem capacidade para apenas um carro. É razoável assumir que carros esperam de tal maneira que quando um boxe se torna livre, o carro no início da fila entra no boxe. A estação pode acomodar quatro carros esperando (sete na estação) a cada tempo. O padrão é Poisson com uma média de um carro a cada minuto durante os períodos de pico. O tempo de serviço é exponencial com média de 6 minutos. I. M. Fussy, o inspetor chefe, deseja saber o número médio no sistema nos períodos de pico, a espera média (incluindo serviço), e o número esperado por hora que não podem entrar na estação devido a lotação máxima ser atingida.

Usando minutos como unidade de tempo básica , $\lambda = 1$ e $\mu = 1/6$. Assim, tem-se que $r = 6$ e $\rho = 2$ para este sistema M/M/3/7. Primeiro será calculado p_0 da equação (2.23) encontrando

$$p_0 = \left(\sum_0^2 \frac{6^n}{n!} + \frac{6^3}{3!}\frac{1 - 2^5}{1 - 2}\right)^{-1} = [(1 + 6 + 18) + 36(31)]^{-1} =$$

$$p_0 = (1 + 6 + 18) + 36(31)]^{-1} = (25 + 1116)^{-1} = 1/1141.$$

Da equação (2.24) obtém-se que

$$L_q = \frac{p_0(6^3)2}{3!}[1 - 2^5 + 5(2^4)] = 72p_0[-31 + 80] = 3528/1141.$$

de modo que

$$L = L_q + r(1 - p_k) = \frac{3528}{1141} + 6\left(1 - \frac{6^7}{(3^4)(3!)(1141)}\right) = \frac{9606}{1141} = 6{,}06 \text{ carros}$$

Para encontrar a espera média nos períodos de pico, a Equação (2.25) é utilizada obtendo-se

$$W = \frac{L}{\lambda_{eff}} = \frac{L}{\lambda(1 - p_7)} = \frac{L}{1 - p_0 6^7(3^4 3!)} = 12{,}3 \text{minutos}.$$

O número esperado de carros por hora que não podem entrar na estação é dado por

$$60\lambda\, p_k = 60\, p_7 = \frac{60\, p_0 6^7}{3^4 3!} = 30{,}4 \text{ carros/h}.$$

Como dito anteriormente, a taxa efetiva do sistema é $\lambda(1 - p_k)$, pois, quando se fala em taxa de chegada, nós estamos nos referindo aquela parcela de clientes que estão fora do sistema e podem, probabbilísticamente falando, chegar no sistema em algum momento. Assim, se existem $(1 - p_k)$ clientes fora do sistema, então existem p_k clientes que já entrarem no sistema quando a capacidade máxima foi atingida, cada um com uma taxa de chegada λ. Dessa forma, para verificar-se o número de carros que não conseguiram entrar no sistema, multiplica-se o número de carros que podem chegar em uma hora pela taxa de chegada dos p_k carros que chegaram no sistema, antes da capacidade máxima ser atingida, dado por $\lambda\, p_k$. O que resultado que o número de carros que não conseguiram entrar no sistema é igual a 60carros/hora (1 a cada minuto) condicionado aos $\lambda\, p_k$ carros que já estão no sistema.

2.5 Fórmula de Erlang (M/M/c/c)

Um caso especial de fila com capacidade limitada M/M/c/K para o qual K = c, isto é, nenhuma fila é permitida ser formada, é originada de uma distribuição estacionária conhecida como *primeira fórmula de Erlang* que pode ser obtida das equações (2.22) e (2.23) com K = c como

$$p_n = \frac{(\lambda/\mu)^n}{n!} \Big/ \left(\sum_0^c \frac{(\lambda/\mu)^i}{i!}\right) \qquad (0 <= n <= c). \qquad (2.29)$$

A fórmula resultante para p_c é conhecida como *fórmula da perda de Erlang* e corresponde a probabilidade de um sistema completo em qualquer tempo em estado de equilíbrio, definido da seguinte forma

$$p_n = \frac{r^c / c!}{\sum_0^c r^i/i!} \qquad (r = \lambda/\mu).$$

Desde que a entrada para M/M/c/c é Poisson, a probabilidade de que uma chegada seja perdida é igual a probabilidade de que todos os servidores estejam ocupados. A situação física original que motivou Erlang para gerar este modelo foi uma rede de telefone simples. As chamadas telefônicas chegam no sistema com uma taxa de Poisson, os tempos de serviço são mutuamente independentes e exponenciais, e todas as chamadas que chegam e encontram a linha ocupada, isto é, obtém um sinal de ocupado são descartadas, de modo que, este modelo tem sempre sido, de fato, de muita utilidade no projeto de telecomunicações.

Entretanto, a grande importância da equação (2.29) é que ela é válida para qualquer modelo M/G/c/c, independente da forma da distribuição de tempo de serviço.

2.6 Filas com Serviço Ilimitado (M/M/∞)

Agora será tratado um modelo de filas no qual o serviço é ilimitado, isto é, existe um número infinito de servidores disponível. Este modelo é frenqüentemente referido como um problema de servidor amplo. Um exemplo de um self-service é um bom exemplo do uso deste modelo.

É feito uso do resultado de nascimento e morte geral com $\lambda_n = \lambda$ e $\mu_n = n\mu$, para todo n, que resulta

$$p_n = (r^n/n!)p_0, \qquad p_0 = (\sum_{n=0}^{\infty} (r^n/n!))^{-1}.$$

Mas, as séries infinitas na expressão para p_0 é claramente idêntica a representação de e^r. Assim, segue que

$$p_n = (r^n e^{-r})/n! \quad (n>=0) \tag{2.30}$$

de modo que, a distribuição de probabilidade de estado de equilíbrio de n no sistema é Poisson com parâmetro $r = \lambda/\mu$, sendo que está equação é válida para qualquer modelo M/G/∞, isto é, p_n depende só do tempo de serviço médio e não da distribuição de tempo de serviço.

O tamanho esperado no sistema é a média da distribuição de Poisson de (2.30), sendo obtido de $L = r = \lambda/\mu$. Desde que o número de servidores é infinito, $L_q = 0 = W_q$. O tempo médio no sistema se tornou meramente o tempo médio de serviço, de modo que, $W = 1/\mu$, e a função de distribuição de tempo de espera W(t) é idêntica a distribuição de tempo de serviço, ou seja, exponencial com média $1/\mu$.

Exemplo 2.4 – A emissora KCAD deseja saber qual o número de telespectadores que ela pode esperar num programa numa tarde de Sábado no horário principal. A KCAD obteve de fontes passadas que os telespectadores que ligam sua TVs nos Sábados à tarde no horário principal obedecem uma distribuição de Poisson com média de 100.000/h. Existem cinco grandes emissoras na área, e é acreditado que as pessoas escolhem uma das cinco aleatoriamente. Fontes também tem mostrado que os telespectadores mudam

de canal a cada 90 minutos e que os tempo de audiência são exponencialmente distribuídos.

Desde que a taxa média de chegada (pessoas sintonizando na KCAD durante o horário principal nos Sábados à tarde) é 100.000/5 = 20.000/h e o tempo de serviço médio é 90min ou 1,5h, segue que o número médio de telespectadores a cada hora é L = 20.000*(2/3) = 40.000/3 = 13334pessoas.

2.7 Filas de Fonte Finita

Nos modelos anteriores foi assumido que a população de onde as chegadas vinham era infinita, desde que o número de chegadas em qualquer intervalo de tempo é uma variável aleatória de Poisson com um espaço de amostragem incontável. Agora será tratado um problema onde a população chegando é finita, digamos de tamanho M, e a probabilidade de ocorrência de eventos futuros são funções do estado do sistema. Uma aplicação típica deste modelo é o caso do reparo de uma máquina, onde a população de chegadas possíveis são as máquinas, sendo que uma chegada corresponde a uma máquina quebrada e os técnicos são os servidores. Será assumido que existem c servidores disponíveis, com tempos de serviço exponencialmente distribuídos com média $1/\mu$. Já o processo de chegada segue a seguinte descrição. Se uma chegada não está no sistema no tempo t, a probabilidade de que ela tenha entrado no tempo $t + \Delta t$ é $\lambda\Delta t + o(\Delta t)$, ou seja, o tempo que uma chegada gasta fora do sistema é exponencial com média $1/\lambda$.

Por causa destas suposições, pode-se usar a teoria de nascimento e morte desenvolvida anteriormente, com as taxas de nascimento e morte dadas por

$$\lambda_n = \begin{cases} (M - n)\lambda & (0<=n<M), \\ 0 & (n>=M) \end{cases}$$

e

$$\mu_n = \begin{cases} n\mu & (0<=n<c), \\ c\mu & (n>=c). \end{cases}$$

usando (1.19) resulta, com $r = \lambda/\mu$,

$$p_n = \begin{cases} \dfrac{M!(M-n)!}{n!} r^n p_0 & (1<=n<c), \\ \dfrac{M!(M-n)!}{c^{n-c}c!} r^n p_0 & (c<=n<M). \end{cases} \qquad (2.31)$$

A forma algébrica de $\{p_n\}$ não permitem cálculos de forma fechada de p_0 (também verdadeiro para L_q, L, W e W_q). Assim, cada um dos coeficientes multiplicando p_0 em (2.31), onde estes coeficientes podem ser denotados por $\{a_n, n = 1,2, \ldots, M\}$ e então a computação pode ser completada por

$$p_0 = (1 + a_1 + a_2 + \ldots + a_M)^{-1}.$$

Para encontrar o número médio de clientes no sistema, ou seja, o número médio de máquinas em reparo, será usado a definição do valor esperado, obtendo-se

$$L = \sum_{n=1}^{M} np_n = p_0 \sum_{n=1}^{M} na_n.$$

Contudo, para obter L_q, W e W_q, deve-se primeiro encontrar a taxa efetiva média de chegadas no sistema. Como mostrado anteriormente, a taxa de chegada média quando o sistema está no estado n é $(M - n)\lambda$. Somando todos os valores possíveis de n depois de pesar cada termo pela probabilidade p_n apropriada, obtém-se que

$$\lambda_{eff} = \sum_{n=0}^{M-1} (M - n)\lambda\, p_n = \lambda(M - L). \qquad (2.32)$$

A equação (2.32) é certamente intuitiva, desde que, na média, L estão no sistema e, conseqüentemente, na média, M – L estão fora e cada um tem uma taxa média de chegada de λ. Agora, para L_q, é sabido de demonstrações anteriores que

$$L_q = L - \lambda_{eff}/\mu = L - r(M - L). \qquad (2.33)$$

e segue das fórmulas de Little que

$$W = \frac{L}{\lambda(M - L)} \text{ e } W_q = \frac{L_q}{\lambda(M - L)} \qquad (2.34)$$

Para a versão de servidor simples deste problema, a expressão para as probabilidades de estado do sistema encontradas em (2.31) reduz-se para

$$p_n = \frac{M!}{(M - n)!n!} r^n p_0 \quad (0<=n<M),$$

e o resto da análise é idêntica.

Exemplo 2.5 – Uma fábrica de semicondutores utiliza cinco robôs na fabricação de placas de circuito. Os robôs quebram periodicamente, e a companhia possui dois técnicos para consertá-las. Quando uma é consertada, o tempo até a próxima quebra é pensado ser exponencialmente distribuído com uma média de 30h. A fábrica tem um backlog de encomendas, garantindo que todos os robôs em condições operacionais estarão trabalhando. O tempo de reparo para cada serviço é exponencialmente

distribuído com uma média de 3h. O gerente da fábrica deseja saber o número de robôs operacionais a qualquer hora, o tempo parado de cada robô que necessita de reparo e a porcentagem de tempo ocioso de cada técnico.

Para responder a qualquer destas questões, deve-se primeiro calcular p_0. Neste exemplo, M = 5, c = 2, λ = 1/30 e μ = 1/3, resultando em r = λ/μ = 1/10. Utilizando (2.31) obtém-se os cinco $\{a_n\}$ multiplicadores como (a_1 = 5/10 = ½; a_2 = 1/10; a_3 = 15/1000 = 3/200; a_4 = 15/10000 = 3/2000; a_5 = 15/200000 = 3/40000. E assim segue que

$$p_0 = (1 + ½ + 1/10 + 3/200 + 3/2000 + 3/40000)^{-1} = 40000/64663 = 0{,}619.$$

O número médio de robôs operacionais é dado por M – L, onde

$$L = p_0(1 \text{ x } 1/2 + 2 \text{ x } 1/10 + 3 \text{ x } 3/200 + 4 \text{ x } 3/2000 + 5 \text{ x } 3/40000) =$$
$$=30055/64663 = 0{,}465$$

Assim 5 – 0.465, ou 4.535, robôs estão em condições operacionais na média. O tempo para do pode ser encontrado de (2.34) e é

$$W = \frac{30055 / 64663}{1/30(5 - (30055/64663))} = 901650/29326 = 3{,}075\text{h}.$$

A fração média de tempo ocioso de cada servidor é

$$p_0 + 1/2\, p_0 = p_0(1 + ½) = 60000/64663 = 0{,}9278.$$

Assim, cada técnico está ocioso 92,78% do tempo. Partindo deste valor alto para o tempo ocioso, o gerente está interessado em saber qual seriam os resultados acima se o número de técnicos fosse decrementado em um. O resultados seria,

$$p_0 = 0{,}564, \quad L = 0{,}640,\ M - L = 4{,}360,\ W = 4{,}400\text{h}$$

Desde que quatro robôs estão ativos em qualquer uma das situações e o tempo de parada para reparo de uma máquina elevou-se para mais ou menos uma hora ao reduzir um funcionário, o gerente pode decidir mover um dos técnicos para outra área.

Este modelo pode ser generalizado para incluir o uso de robôs reservas. Assume-se agora que existem Y robôs reservas em mão, de modo que, quando uma robô em operação falha, ele é imediatamente substituído por um reserva. Entretanto, se um robô falhar e todas os reservas estiverem sendo utilizados então o sistema se tornará reduzido. Quando uma máquina for reparada, ela ficará como reserva (a menos que o sistema esteja reduzido, onde ela entrará imediatamente em serviço. Para este modelo, λ_n é diferente de antes e é dado por

$$\lambda_n = \begin{cases} M\lambda & (0<=Y<=M) \\ (M - n + Y) & (Y<=n<Y + M), \\ 0 & (n>=Y + M). \end{cases}$$

Novamente considerando c reparadores, tem-se que

$$\mu_n = \begin{cases} n\mu & (0<=n<c), \\ c\mu & (n>=c). \end{cases}$$

Assumindo, inicialmente, que c <=Y e usando a equação (1.19) mais uma vez, obtém-se

$$p_n = \begin{cases} \dfrac{M^n}{n!} r^n p_0 & (0<=n <c), \\ \dfrac{M^n}{c^{n-c}c!} r^n p_0 & (c<=n<Y) \\ \dfrac{M^Y M!}{(M- n + Y)!\, c^{n-c}c!} r^n p_0 & (Y<=n<Y+M). \end{cases} \tag{2.35}$$

Se Y é muito grande tem-se uma população infinita de clientes com taxa média de chegada $M\lambda$ para λ.

Quando c > Y, tem-se que

$$p_n = \begin{cases} \dfrac{M^n}{n!} r^n p_0 & (0<=n <=Y), \\ \dfrac{M^Y M!}{(M- n + Y)!n!} r^n p_0 & (Y+1<=n<c) \\ \dfrac{M^Y M!}{(M- n + Y)!\, c^{n-c}c!} r^n p_0 & (c<=n<Y+M). \end{cases} \tag{2.36}$$

Deve ser notado que quando Y = 0 (sem reservas), a Equação (2.36) reduz-se para (2.31). Observe que o uso direto da equação (2.36) para obter os coeficientes $\{a_n\}$pode ser meio bagunçado quando M e Y são grandes. Felizmente, pode-se evitar tal dificuldade usando a relação de recursão p_{n+1} para p_n, que está naturalmente fora das formulações de nascimento e morte, sendo uma conseqüência direta das equações de balanço local. As equações de balanço podem ser escritas em forma recursiva da seguinte forma

$$p_{n+1} =(\lambda_n/\mu_{n+1})\, p_n,$$

de modo que, para problema sem reservas segue que

$$p_{n+1} = \begin{cases} \dfrac{M - n}{n+1} r p_0 & (0<=n <c - 1), \\ \dfrac{M - n)}{c} r p_0 & (c<=n<M - 1). \end{cases}$$

A probabilidade do sistema vazio, p_0, pode ser encontrada como anteriormente para modelos sem reserva por mais uma vez usar o fato que as

probabilidades devem somar um, de modo que, a computação de p_0 é feita sobre somas finitas. O mesmo é verdadeiro para L e L_q. Para obter os resultados para W e W_q comparável a equação (2.34) é necessário obter novamente a taxa média efetiva de chegada λ_{eff}. Para obter λ_{eff}, pode-se usar a equação (2.33) diretamente ou pode-se usar lógica similar a utilizada na equação (2.32), como a seguir,

$$\lambda_{eff} = \sum_{n=0}^{y-1} M\lambda\, p_n + \sum_{n=Y}^{Y+M} (M- n + Y)\lambda\, p_n =$$

$$= \lambda(M - \sum_{n=Y}^{Y+M}(n - Y)p_n). \qquad (2.37)$$

Como um ponto final para fechar esta seção, suponha que se deseje encontrar a distribuição de tempo espera. O procedimento padrão empregado até agora tem sido encontrar o tempo de espera de um cliente chegando condicionado a existência de n no sistema no ponto de chegada, e então incondicionando com respeito a distribuição estacionária $\{q_n\}$, onde $\{q_n\}$ são as probabilidades de estado dado que uma chegada ocorreu. Como no caso para M/M/c/K, tem-se que $q_n \neq p_n$, de forma que, antes de obter a FDC do sistema ou tempo na fila, é necessário relacionar as probabilidades de tempo geral p_n com a probabilidade q_n de uma chegada encontrar n no sistema. Para o problema da fila com fonte finita (problema do reparo da máquina sem reservas), as duas probabilidades são relacionadas como

$$q_n = \frac{(M - n)\, p_n}{k},$$

onde k é uma constante normalizadora apropriada determinada pelo somatório de q_n para 1. Para provar isto será usado novamente o teorema de Baye como na situação M/M/c/K da seção 2.4:

$$\Pr\{n \text{ no sistema} \mid \text{chegada acabou de ocorrer}\} =$$

$$= \frac{\Pr\{n \text{ no sistema}\}\Pr\{\text{chegada acabou de ocorrer} \mid n \text{ no sistema}\}}{\sum_n(\Pr\{n \text{ no sistema}\}\Pr\{\text{chegada acabou de ocorrer} \mid n \text{ no sistema}\})} =$$

$$= \lim_{\Delta t \to 0} \frac{p_n[(M - n)\lambda\Delta t + o(\Delta t)]}{\sum_n p_n[M - n)\lambda\Delta t + o(\Delta t)]}$$

$$= \frac{(M - n)\, p_n}{\sum_n(M - n)p_n} = \frac{(M - n)\, p_n}{M - L}.$$

Uma propriedade interessante que pode ser provada é que $q_n(M)$, a probabilidade do ponto de chegada para o caso do reparo de máquinas sem reservas com M máquinas, é igual a probabilidade de tempo geral para o mesmo modelo só que com M – 1

máquinas, dada por $p_n(M-1)$. Isto não é necessariamente o caso quando existem máquinas reservas disponíveis. De fato, quando existem máquinas reservas, $q_n(M)$ pode ser mostrado ser

$$q_n = \begin{cases} \dfrac{M p_n}{M - \sum_{n=Y}^{Y+M} (n - Y) p_n} & (0 <= n < Y - 1), \\[2ex] \dfrac{(M - n + Y)\, p_n}{M - \sum_{n=Y}^{Y+M} (n - Y) p_n} & (Y <= n < Y + M - 1). \end{cases}$$

Isto não é igual a $p_n(M-1)$, mas é igual $p_n(Y-1)$, isto é, se for reduzido o tamanho da população por um, retirando-se uma máquina reserva, então as probabilidades de tempo geral da população reduzida é igual as probabilidades de ponto de chegada da população original.

As distribuições de tempo de espera serão de novo em termos de somas de Poisson cumulativas, como no modelo M/M/c/K. A análise procede como segue:

$$W_q(t) \quad Pr\{T_q <= t\} = W_q(0) + \sum_{n=c}^{Y+M-1} [Pr\{n - c + 1 \text{ serviços completados em } <= t \mid \text{la chegada encontrou n no sistema}\}.q_n]$$

$$= W_q(0) + \sum_{n=c}^{Y+M-1} q_n \int_o^t \frac{c\mu(c\mu x)^{n-c}}{(n-c)!} e^{-c\mu x}\, dx$$

$$= W_q(0) + \sum_{n=c}^{Y+M-1} q_n \left[1 - \int_t^\infty \frac{c\mu(c\mu x)^{n-c}}{(n-c)!} e^{-c\mu x}\, dx \right]$$

$$= 1 - \sum_{n=c}^{Y+M-1} q_n \sum_{n=0}^{N-c} \frac{(c\mu t)^i}{i!} e^{-c\mu t}.$$

2.8 Serviço Dependente do Estado

Nesta seção será tratado o caso de filas Markovianas com serviç dependentes do estado, isto é, a taxa de serviço médio depende do estado do sistema (número no sistema). Em muitas situações reais, os servidores podem acelerar quando ocorre uma longa formação de filas. Por outro lado, pode acontecer de um servidor com pouco tempo de atividade (sem experiência) ficar mais lento quando estas formações ocorrem, fazendo com que a taxa média de serviço caia tornando o sistema congestionado. São esses tipos de situações que serão considerados agora.

O primeiro modelo que será considerado é um no qual um servidor simples tem duas taxas médias, digamos *lento* e *rápido*. O trabalho é feito na taxa lenta até que existam k clientes no sistema, momento onde a taxa do servidor é acelerada. Este mecanismo de serviço pode ser uma máquina com duas velocidades. Será assumido que os tempos de serviço são Markovianos, mas a taxa média μ_n dependerá do estado n do sistema. Além disso, nenhum limite no sistema será imposto. Assim, μ_n é dado por

$$\mu_n = \begin{cases} \mu_1 \ (1<=n<k), \\ \mu \ (n>=k). \end{cases} \tag{2.39}$$

Assumindo que o processo de chegada é Poisson com parâmetro λ e utilizando a equação (1.19), tem-se que

$$p_n = \begin{cases} \rho_1^{\,n}\rho_0 & (0 <= n < k), \\ \rho_1^{\,k-1}\rho^{n-k+1}\rho_0 & (n >= k) \end{cases} \tag{2.40}$$

onde $\rho_1 = \lambda/\mu_1$ e $\rho = \lambda/\mu < 1$. Devido ao fato de que probabilidades devem somar 1, segue que

$$p_0 = \left(\sum_{n=0}^{k-1}\rho_1^{\,n} + \sum_{n=k}^{\infty}\rho_1^{\,k-1}\rho^{n-k+1}\right)^{-1},$$

de mo que,

$$p_0 = \begin{cases} \left(\dfrac{1-\rho_1^{\,k}}{1-\rho_1} + \dfrac{\rho\rho_1^{\,k-1}}{1-\rho}\right)^{-1} & (\rho_1 \neq 1, \rho < 1), \\ [\,k + \rho/(1-\rho)]^{-1} & (\rho_1 = 1, \rho < 1). \end{cases} \tag{2.41}$$

Note que se $\mu_1 = \mu$, as equações (2.40) e (2.41) reduzem-se para aquelas apresentadas no modelo M/M/1.

O tamanho esperado do sistema para $\rho_1 \neq 1$ é obtido da seguinte maneira:

$$L = E[N] = \sum_{n=0}^{\infty} n p_n = p_0\left(\sum_{n=0}^{k-1} n\rho_1^{\,n} + \sum_{n=k}^{\infty} n\rho_1^{\,k-1}\rho^{n-k+1}\right)$$

$$= p_0\left[\rho_1\sum_{n=0}^{k-1} n\rho_1^{\,n-1} + \rho_1(\rho_1/\rho)^{k-2}\sum_{n=k}^{\infty} n\rho^{n-1}\right]$$

$$= p_0\left[\rho_1\frac{d}{d\rho_1}\sum_{n=0}^{k-1}\rho_1^{\,n} + \rho_1(\rho_1/\rho)^{k-2}\frac{d}{d\rho}\sum_{n=k}^{\infty}\rho^{n}\right]$$

$$= p_0\left[\rho_1\frac{d}{d\rho_1}\left(\frac{1-\rho_1^{\,k}}{1-\rho_1}\right) + \rho_1(\rho_1/\rho)^{k-2}\frac{d}{d\rho}\left[\frac{1}{1-\rho} - \frac{1-\rho^{k}}{1-\rho)}\right]\right]$$

Assim, finalmente,

$$L = p_0\left(\frac{\rho_1[1 + (k-1)\rho_1^k - k\rho_1^{k-1}]}{(1-\rho_1)^2} + \frac{\rho\rho_1^{k-1}[k-(k-1)\rho]}{(1-\rho_1)^2}\right) \qquad (2.42)$$

Uma vez de posse de L pode-se obter L_q usando os dois últimos relacionamentos da Tabela 1.2 como

$$L_q = L - (1 - p_0),$$

e W e W_q das fórmulas de Little como $W = L/\lambda$ e $W_q = L_q/\lambda$.

Note que a relação $W = W_q + 1/\mu$ não pode ser usado aqui, desde que μ não é constante, mas depende do chaveamento do sistema no ponto k. Entretanto, pela combinação das equações acima, tem-se que

$$W = W_q + (1 - p_0)/\lambda$$

que implica que o tempo de serviço esperado é $(1 - p_0)/\lambda$.

Exemplo 2.6 – Um estabelecimento comercial possui uma máquina de polimento de carros com duas velocidades. Na velocidade baixa, a máquina leva 40 minutos, em média, para polir um carro. Na velocidade alta, leva só 20 minutos na média. Uma vez que o chaveamento da baixa velocidade para alta é feito, os tempos atuais podem ser assumidos seguirem uma distribuição exponencial. Os clientes são atendidos na base do primeiro a chegar é o primeiro a ser atendido, e uma vez que, o estabelecimento está numa região de baixa densidade populacional e área de tráfego, não existe nenhum limite no número de clientes que podem esperar.

É estimando que os clientes cheguem de acordo com um processo de Poisson com um tempo de interchegada médio de 30 minutos. Os donos do estabelecimento desejam calcular o efeito de duas medidas: chaveamento para alta velocidade quando existem quaisquer clientes esperando, ou seja, dois ou mais no sistema e chaveamento para ata velocidade só quando mais que um cliente estiver esperando, ou seja, três ou mais no sistema. É desejado saber o tempo de espera média sob as duas políticas.

É, portanto, necessário calcular W para o caso k = 2 e k = 3. Para isto é necessário calcular p_0 das equação (2.41) e L da equação (2.42) e finalmente das fórmulas de Little. Antes de realizar estes cálculos, primeiro deve-se calcular ρ e ρ_1. Assim,

$$\rho_1 = \lambda/\mu_1 = (1/30)/(1/40) = 4/3 \text{ e } \rho_1 = \lambda/\mu = (1/30)/(1/20) = 2/3.$$

Agora, para o caso 1, k = 2 tem-se que

$$p_0 = 1/5 = 0.2, \qquad L = 12/5 = 2.4 \text{ carros},$$

$W = L/\lambda = 2.4/(1/30) = 72$ minutos = 1 hora e 12 minutos.

Para o caso 2, k = 3 tem-se que

$p_0 = 3/23 = 0.13$, $L = 204/69 = 2.96$ carros,

$W = L/(1/30) = 89$ minutos = 1 hora e 29 minutos.

Os donos do estabelecimento sentem que a espera média de 17 minutos ou mais para realizar o chaveamento das velocidades em 3, aon invés de 2, não tem nenhum efeito adverso na clientela. Entretanto, custa mais executar a máquina na velocidade mais alta. De fato, é estimado que custa R$15,00 por hora para operar a máquina na baixa velocidade e R$24,00 para operar na alta velocidade. Assim, o custo de operação esperado ao chavear em k é dado por

$$C(k) = 15\sum_{n=1}^{k-1} p_n + 24\sum_{n=k}^{\infty} p_n$$

$$= 15\sum_{n=1}^{k-1} \rho_1^n p_0 + 24\left(1 - \sum_{n=0}^{k-1} \rho_1^n p_0\right).$$

Para o caso 1 tem-se que

$C(2) = 15(4/3)(1/5) + 24[1 - 1/5 - (4/3)(1/5)] =$ R$16,80/h,

enquanto para o caso 2 o custo de operação médio é

$C(3) = 15(3/23)[4/3 + 16/9] + 24[1 - 3/23 - (4/3)(3/23) - (16/9)(3/23)] =$ R$17,22/h.