Universidade Federal de Minas Gerais
Programa de Pós-Graduação em Engenharia Elétrica

Aproximação para sistemas de filas M/M/c com servidores heterogêneos

Frederico Samartini Queiroz Alves

Dissertação de mestrado submetida ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do Título de Mestre em Engenharia Elétrica.

Orientador Prof. Dr. Hani Camille Yehia

Belo Horizonte
Setembro/2007

Dedico esta dissertação, especialmente, a meu pai e minha mãe, a minha irmã, a toda a minha família e a todos aqueles que me apoiaram e incentivaram para que eu alcançasse este ideal.

Agradecimentos

Gostaria de expressar sincera gratidão ao Profº Luís Antônio Capanema Pedrosa, por ter me incentivado ao estudo da Teoria de Filas e por ter sido meu conselheiro, ajudando-me em várias etapas deste trabalho.

Ao meu orientador e Professor, Drº Hani Camille Yehia por ter me dado a oportunidade de seguir dentro da carreira acadêmica e, também, pelos conselhos que, ao longo do meu caminho, fizeram mudar o meu percurso.

Agradecimento especial e carinhoso faço a meu pai e minha mãe que me deram conselhos úteis para a construção desta dissertação e pela compreensão que tiveram durante esse período, suportando as ansiedades e as angústias.

Agradeço aos amigos que colaboraram com seus conselhos e que, também, souberam compreender minha ausência quando precisaram. Em especial a meu amigo Rabino, que acompanhou com mais interesse todo o processo e ajudou-me em vários momentos a buscar soluções.

Sou grato à minha namorada Luciana por ter me apoiado e ter me encorajado a manter minhas metas, além de ter contribuido diretamente na criação desta dissertação me ajudando com as simulações.

Deixo meus agradecimentos, também, aos professores, colegas de curso de engenharia elétrica e do CEFALA que me ajudaram na realização deste trabalho.

Agradeço à Fundação de Amparo a Pesquisa do Estado de Minas Gerais - FAPEMIG por ter financiado os estudos necessários para a conclusão desta dissertação.

Resumo

Este trabalho se aplica, especificamente, ao caso de C servidores heterogêneos, exponencialmente distribuídos e que atendem a uma fila única, com disciplina de atendimento *First-Come,First-Served*, formada por apenas uma classe de cliente. É apresentada uma formulação matemática para se obter um limite superior para as medidas de desempenho. Tal formulação é obtida, basicamente, através de uma expansão do espaço de estados, resultante da heterogeneidade dos servidores, e, em seguida, através de uma redução desse espaço de estados. Essa redução é viável, pois apenas as possibilidades de maior probabilidade foram consideradas. Com isso, é possível encontrar o pior caso para o tempo médio de espera na fila e para o número médio de serviços na fila, como, também, para o tempo médio no sistema e o número médio de pessoas no sistema.

Essa formulação se torna atraente por ser capaz de aproximar o comportamento real desses sistemas de servidores heterogêneos com um erro, na maioria dos casos, menor do que se fosse calculado utilizando a aproximação existente para uma MMc tradicional. Resultados de simulações, feitas em GPSS (General Purpose Simulation System), são apresentados com o intuito de validar a formulação criada e de comparar o erro relativo dela com o de outras aproximações.

O Índice de Gini é usado para facilitar a comparação entre sistemas, pois, através desse, é viável classificá-los quanto à heterogeneidade e, então, avaliar qual é o efeito resultante dessa sobre as medidas de desempenho de um sistema de filas qualquer. É apresentada, também, uma análise sobre a influência que alguns tipos de alocação têm sobre os resultados.

Abstract

This work is specifically applied to the case of C heterogeneous servers, exponentially distributed, which assist a single FCFS line formed for just one type of customer class. A mathematical formulation is presented to obtain an upper bound for the performance measures. Basically, such formulation is obtained through an expansion of the state space resulting from the heterogeneity of the servers and afterwards through a reduction of that state space. That reduction is feasible because only the possibilities of larger probability are considered. With that, it is possible to find the worst case for the average waiting time in queue and for the average number of people in the queue, as well as for the average waiting time in the system and the average number of people in the system.

In most of the cases that formulation becomes attractive as it is capable to approximate the real behavior of those systems of heterogeneous servers with an error that is smaller than if it was calculated using the traditional MMc.

Results from simulations, which were run in GPSS, are showed with the intention of validating the created formulation and to compare the relative error resulted from it with the error from other approaches. The Gini's Index is used to make the comparison among systems possible as it makes viable the classification of the systems according to their heterogeneity. As a result of that it is possible to evaluate which is the resultant effect on the performance measures of those queueing systems. In addition, some analyses on the influence which different allocation polices have over the results are also presented.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

ANOVA	Analysis of Variance
ATM	Asynchronous Transfer Mode
BD	Processo de Nascimento e Morte (Birth-Death process)
B-ISDN	Broadband Integrated Services Digital Network
BPI	Batch Poisson Input
CDF	Função Distribuição Cumulativa
CPDEE	Centro de Pesquisa e Desenvolvimento em Engenharia Elétrica
DSL	Digital Subscriber Line
EBIT	Earnings Before Interests and Tax
FCFS	First-Come, First-Served
FSF	Fastest Servers First
GAMS	General Algebraic Modeling System
GPSS	General Purpose Simulation System
IP	Internet Protocol
MAC	Medium Access Control
MCMS	Multi Class Multi Server
M/M/c	Chegadas Markovianas/Saídas Markovianas/c servidores - Notação de Kendall
MPLS	Multiprotocol Label Switching
MSE	Mean Squared Error
PB	Processo de Puro Nascimento (Pure-Birth)
PDF	Função Densidade de Probabilidade(Probability density function)
PP	Processo de Poisson
PSTN	Public Switched Telephone Network

QED	Quality and Efficiency Driven
QoS	Quality of Service
RP	Processo de Renovação (Renewal process)
RW	Random Walk process
SHE	Servidores Heterogêneos exponenciais
SLA	Service-Level-Agreements
SMP	Processo Semi-Markoviano (Semi-Markov process)
SQMS	Single Queue Multi Server
UPC	Usage Parameter Control
UT	Unidade de Tempo
VA	Variável Aleatória
VoD	Video on Demand
VoIP	Voice over IP
WSF	Web Server Farm

Lista de Símbolos

C	Número de servidores no sistema
$C(t)$	Confiabilidade do sistema
Δ(t)	Intervalo de tempo t
E_i	Estado i
$F(x)$	Função de densidade de probabilidade acumulada
$f(x)$	Função densidade de probabilidade de x
f_i	Probabilidade de o sistema sempre voltar a E_i
$f_i^{(n)}$	Probabilidade de o sistema voltar a E_i em n transições
$\Phi(t)$	Função geradora de momento
γ	Período de recorrência para uma cadeia periódica de Markov
L	Número médio de pessoas no sistema
L_q	Número médio de pessoas na fila
λ	Taxa constante de chegada de clientes
λ_i	Taxa de Nascimento para um processo Markoviano no estado i
m_c	Somatório das capacidades de processamento dos c servidores
M_i	Tempo médio de recorrência
m_i	Somatório das capacidades de processamento dos i primeiros servidores
μ_i	Taxa de Morte para um processo Markoviano no estado i
μ_j	Taxa de processamento do servidor i
$N(t)$	Número de eventos no intervalo t
p_0	Probabilidade de se ter 0 jobs no sistema
P_i	Distribuição estacionária de probabilidade de se estar em E_i
p_i	Probabilidade de se estar no estado i em um processo Markoviano ou a probabilidade de se ter i jobs em um sistema de filas
p_{ii}	Probabilidade do sistema permanecer no estado i

p_{ik}	Probabilidade de transição do estado i para o estado k
π_i	Probabilidade limite de se estar em E_i
π_i^n	Probabilidade de um sistema se encontrar em E_i na n-ésima transição
ρ	Utilização do sistema, ou intensidade de tráfego
$r(t)$	Função taxa de falha
$S(i)$	Tempo de espera até o i-ésimo evento
$V^{(f)}$	Sistema qualquer em um estado futuro
$V^{(n)}$	Sistema qualquer em um ponto n qualquer no tempo
$V^{(p)}$	Sistema qualquer em um estado presente
W	Tempo total médio que cada job fica no sistema
W_q	Tempo médio que cada job espera na fila
y_q	Probabilidade condicional de se estar no estado $V^{(n)} = y$

Capítulo 1

Introdução

1.1 Introdução

As filas são fenômenos que acontecem a todo o momento em nossos dias. Pode-se ver fila em tudo o que fazemos e aonde vamos. Há filas de carros, para comprar um lanche ou quando se vai pagar uma conta. Em diversas e variadas situações, podem-se vê-las. As filas que acontecem no nosso dia-a-dia não são, apenas, as visíveis diretamente. Existem filas, por exemplo, nos aeroportos e nos portos, além de estarem presentes em todo o processo de produção e escoamento de produtos, fazendo com que os preços das mercadorias subam ou desçam conforme a eficiência na distribuição e na logística. Em telefonia, principalmente com as tecnologias de comutação de pacotes, Video sob Demanda e Voz sobre IP, as filas também aparecem. Desta forma, resolver de maneira adequada os problemas de filas pode ser um fator de redução de custos e de maximização da eficiência de um sistema.

As filas acontecem a todo o momento em vários sistemas e elas poderão se tornar instáveis quando a quantidade de trabalho demandada ao sistema é maior ou igual à capacidade de processar essa quantidade de serviço. Quando a oferta de processamento é menor que o fluxo de trabalho que chega nesse sistema, as filas, inevitavelmente, surgirão e poder-se-ão tornar um entrave intransponível. Em alguns casos, a solução é aumentar a capacidade de atendimento, mas isso pode requerer vultosas despesas de capital. Em outros casos, as filas podem ser reduzidas através de uma alocação mais eficiente dos recursos existentes.

A investigação de sistemas de filas é, portanto, de extrema importância. Dentro deste contexto são considerados os casos em que as filas são processadas por servidores que não

têm as mesmas taxas de trabalho e, por isso, são chamados de Servidores Heterogêneos. A necessidade de estudar os sistemas atendidos por esses servidores vem do fato de que, na prática, o número de sistemas nos quais os servidores trabalham com taxas diferentes vem crescendo. Considerando, por exemplo, seres humanos como servidores, sendo cada um encarregado de realizar um serviço mesmo que esse serviço seja igual para todos, cada um terminará o trabalho em um tempo diferente. Afinal de contas, seres humanos são diferentes e têm capacidades diferentes. Ainda que os servidores sejam máquinas, é provável que o mesmo ocorra. No caso da Internet, é possível que o processador de uma máquina ou o roteador seja mais rápido que outros. Principalmente porque os sistemas de comunicação enfrentam um processo rápido e permanente de renovação de tecnologia, o que faz equipamentos velhos serem normalmente mais lentos. Considerando os processos industriais, é de se esperar que haja máquinas com diferentes preços e, portanto, com diferentes taxas de produção, ou mesmo, serem de origens diferentes o que, também, afeta as respectivas capacidades. Mais ainda, as máquinas podem, apenas, ser de idades diferentes e, consequentemente, trabalharem com capacidades díspares.

A Teoria de Filas é um ramo da Probabilidade Aplicada com a Pesquisa operacional que procura modelar, matematicamente, esses sistemas e, dessa forma, revelar a natureza probabilística por trás desses. Através desses modelos matemáticos, procura-se, de forma detalhada, obter respostas ótimas para problemas de gerenciamento de várias naturezas. Para isso, a Teoria de Filas aparece como ferramenta para a obtenção de parâmetros de desempenho, como, por exemplo, tempo de espera médio e número médio de pessoas tanto na fila quanto no sistema. Com a criação de formulações fechadas é possível tomar decisões de forma mais rápida e confiável, já que não é mais preciso utilizar apenas simulações, as quais podem ser imprecisas e muitas vezes demoradas.

1.2 Objetivo

O objetivo desta dissertação é desenvolver uma formulação matemática que represente, de forma aproximada, os sistemas de filas correspondentes aos casos em que existam servidores heterogêneos exponencialmente distribuídos, alimentados por uma fila única. Tal fila tem disciplina de atendimento FCFS (First-Come First-Served) e é formada por *jobs* de uma só classe, cujas chegadas ao sistema acontecem de acordo com um processo de Poisson. Pretende-se, com essa formulação, obter as medidas de desempenho, tempo médio de espera por *job* na fila, tempo médio de permanência de um *job* no sistema,

número médio de *jobs* no sistema e número médio de *jobs* na fila.

Especificamente, pretende-se avaliar, através da formulação desenvolvida, a influência da heterogeneidade dos servidores dentro do sistema. O Índice de Gini será utilizado para mensurar essa heterogeneidade entre tais sistemas e, então, tornar possível a classificação e a comparação desses.

Com base em tais medições, pode-se, então, determinar como o sistema se comporta quando se varia o número de servidores. Deseja-se com isso avaliar o quanto os resultados obtidos com a formulação proposta são afetados com o aumento da quantidade de servidores. Para essa avaliação, são comparados valores obtidos com valores simulados.

Em seguida, pretende-se definir limites superiores e inferiores para cada configuração de servidores. Isto é feito, através de comparações do erro resultante em relação às variações que ocorrem na formulação proposta quando são utilizadas as seguintes políticas de alocação de servidores: (i) escolhendo os servidores mais lentos primeiro; (ii) escolhendo os servidores de forma aleatória; e (iii) escolhendo os servidores mais rápidos primeiro.

Finalmente, pretende-se analisar o erro acarretado nos resultados calculados com a formulação proposta quando se varia o coeficiente ρ de utilização do sistema. Espera-se que o erro seja menor quando o ρ tende à unidade, pois, a influência das políticas de alocação nas medidas de performance para esse caso é menor.

A análise realizada permite a identificação de aplicações práticas nas quais a formulação proposta possa ser utilizada. Dessa forma, este estudo torna-se útil para o gerenciamento de quaisquer processos de filas que ocorrem no dia-a-dia, se esses puderem ser representados pelo modelo aqui proposto.

1.3 Motivação

A motivação para este trabalho vem da importância de se desenvolver uma ferramenta que possibilite uma melhora nas avaliações feitas sobre vários processos de filas e, assim, melhorar o gerenciamento dos mesmos. Como conseqüência de uma melhor gestão dos processos, tem-se: melhor qualidade dos serviços, maior capacidade de trabalho e economia financeira através de redução de custos.

São encontrados, na literatura, muitos modelos que se utilizam da Teoria de Filas para descrever matematicamente a realidade. Entretanto, quase todos, quando é o caso de mais de um servidor, consideram os servidores como sendo homogêneos. Esta é apenas uma das simplificações feitas ao tentar aproximar do que realmente acontece. Nesta dissertação,

procura-se dar um passo a mais em direção a uma melhor representação do que acontece na realidade. A possibilidade de criação de uma formulação relativamente simples e que seja sensível à heterogeneidade dos servidores quando esses são exponencialmente distribuídos representa uma contribuição suficientemente significativa para justificar a realização deste trabalho.

A gama de sistemas com natureza distinta em que a formulação proposta poderá ser aplicada é mais um estímulo à investigação da mesma. Esta gama estende-se desde sistemas de telecomunicações, nas quais agrupam-se tecnologias de video sob demanda e voz sobre Ip,passando por redes de computadores em geral, e indo até sistemas de manufatura, logística e análise financeira.

1.4 Revisão Bibliográfica

Com a ampla utilidade encontrada para a Teoria de Filas em várias áreas, observa-se que, nos últimos anos, vários autores vêm estudando extensivamente o assunto. Devido, portanto, à quantidade de trabalhos publicados neste campo, um leitor interessado pode se perder entre tantos caminhos possíveis. Neste trabalho os esforços são divididos de maneira a evitar esse erro. Para isso os seguintes passos são seguidos: (i) estudo sobre teoria de filas descrita em Gross e Harris (1985), Kleinrock (1976a), Kleinrock (1976b), Ross (1993), El-Taha e Stidham (1999), Wolff (1989), Saaty (1961), Feller (1957), Feller (1966), Cooper (1981) e Allen (1990) onde foram encontrados os fundamentos necessários para desenvolver este estudo; (ii) pesquisa e conseqüente estudo sobre trabalhos que desenvolvem a teoria sobre servidores heterogêneos necessária para este; e (iii) investigação sobre possíveis aplicações.

Existe mais de um tipo de sistemas com servidores heterogêneos, e esses são tratados de formas diferentes. O primeiro tipo de sistema consiste em servidores individuais, onde cada um deles "enxerga" à sua frente uma fila única à qual só ele atende. Esse sistema está representado na Figura 1.1. O segundo tipo de sistema, mostrado na Figura 1.2, é o caso em que os servidores atendem à mesma fila, ou seja, existe somente uma fila "alimentando" os servidores quando esses ficam vazios.

O primeiro tipo de tratamento, na verdade, considera múltiplas filas paralelas. Esse tipo de sistema é caracterizado, portanto, por vários servidores que trabalham em paralelo onde cada um tem uma fila própria como se fosse um sistema de apenas um servidor. Nesse tipo de sistema, normalmente é considerada a troca entre filas (jockeying), o que ocorre

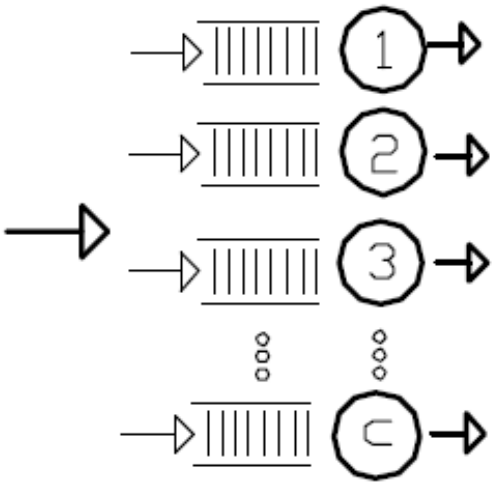


Figura 1.1: Sistema com servidores Heterogêneos atendidos por múltiplas filas paralelas.

quando um *job* sai de sua fila e vai para uma mais curta. Para se calcular as medidas de performance dos sistemas com múltiplas filas considera-se cada fila individualmente e depois tira-se a média ponderada de cada uma dessas filas. Essa ponderação é feita através de pesos probabilísticos de roteamento (The Routing Probabilities).

Muitos pesquisadores vêm estudando esses tipos de sistemas de filas paralelas e esses consideram ou não a heterogeneidade dos servidores. Alguns artigos sobre essas filas e que de alguma forma contribuem como base para o modelo desenvolvido neste trabalho, são: Levine e Finkel (1990), Banawan e Zahorjan (1989), Haight (1958), Kingman (1961), Borst (1990), Koenigsberg (1966), Lee (1994), Whitt (1980), Whitt (1992), Wein (1991), Disney e Mitchell (1971), Elsyed e Bastani (1985), Blanc (1987), Blanc (1992), Schwartz (1974), Zhao e Grasmann (1990), Grassmann e Zhao (2004), Zhao e Grasmann (1990).

O segundo tipo de sistema com servidores heterogêneos é o que é tratado neste trabalho. Nesse caso existe apenas uma fila onde os *jobs* aguardam para serem atendidos por algum servidor que fique livre, o que acontece quando esse termina o processamento do trabalho que estava realizando. Essa é uma fila FCFS (First-Come, First-Served), na qual o primeiro *job* que chega ao sistema é o primeiro a ser atendido. Obviamente, o fato do primeiro *job* que chegar ser atendido primeiro não resulta no fato de que esse irá também sair primeiro do sistema, uma vez que, a taxa de processamento de cada servidor é diferente. Outra razão para que isso ocorra vem da consideração de que o tempo de processamento de trabalho em cada servidor é governado pela distribuição exponencial.

No sistema de uma fila compartilhada por diversos servidores, não é possível calcular as medidas de performance fazendo as ponderações como feito para os sistemas com filas paralelas. Assim sendo, a complexidade analítica para os cálculos de medidas de desempenho desse tipo de sistema faz com que hoje na literatura ainda existam menos trabalhos publicados, que para os casos de filas paralelas e de servidores homogêneos. Como muitas aplicações necessitam da formulação de filas como forma de obtenção de conhecimento sobre os sistemas, simplificações são empregadas para que se contorne a dificuldade analítica. Uma forma simplificada de se tratar esses sistemas é considerar todos os servidores homogêneos reduzindo o sistema a uma fila G/G/C. Outra forma de simplificação é considerar as chegadas como sendo de acordo com o processo de Poisson e o tempo de serviço como sendo exponencialmente distribuído, o que reduziria o sistema a uma fila[1] M/M/c.

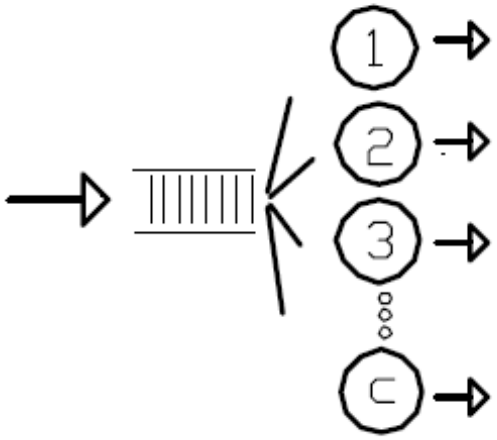


Figura 1.2: Sistema com servidores Heterogêneos atendidos por uma única fila.

Nesses sistemas, pode-se considerar os *jobs* como pertencendo a mais de uma classe diferente. Esse é o caso dos sistemas com múltiplos servidores heterogêneos com múltiplas classes de *jobs*. Neste trabalho, particularmente, os *jobs* são considerados como sendo pertencentes a uma só classe, ou seja, todos serão atendidos pelo mesmo tipo de servidor.

Até o momento não é do conhecimento do autor nenhum trabalho que trate do assunto de servidores heterogêneos com a generalidade proposta aqui dentro do mesmo foco

[1] Entrada e saída markovianas com c servidores - notação de modelos de filas de Kendall (Kleinrock, 1976a)

tratado. Porém, para que se tenha uma visão geral das pesquisas já realizadas sobre o tema são descritos a seguir trabalhos que serviram de base para esta dissertação.

- Shimkin e Mandelbaum (2004) consideram o caso em que clientes (por impaciência) abandonam a fila e deixam o sistema. Os clientes são considerados heterogêneos quando analisados pelo ponto de vista de seus parâmetros de utilização, pois esses variam dentro da população. A fila é assumida como sendo invisível para o cliente em espera, pois este não obtém nenhuma informação sobre o sistema. Esse tipo de fila acontece em centros de telefonia ou em qualquer tipo de serviço remoto. Eles modelam o sistema através de uma fila atendida por múltiplos servidores com clientes impacientes. E, então, definem pontos de equilíbrio para o sistema, nos quais este é mantido estável e o cliente não abandona a fila. As chegadas são de Poisson e o tempo de serviço é exponencialmente distribuído. Entretanto, eles não consideram os servidores heterogêneos. Medidas de performance também são feitas para esse modelo.

- Chao e Luh (2004) mostram que em um sistema com fila M/M/C/N, ou seja, com C servidores e com capacidade limitada em N (é considerado um buffer com N lugares para os *jobs* incluindo os que estão em serviço), a probabilidade de se barrar os *jobs* é convexa para (λ, μ), onde λ é a taxa de chegada no sistema e μ é a taxa de serviço. Eles levam em conta então, um sistema de fila com servidores heterogêneos com o tempo entre chegadas e os tempos de processamento de serviços não-estacionários. As taxas de chegada e de serviço alternam entre dois níveis (λ_1, μ_1) e (λ_2, μ_2) e ficam nos níveis i (i=1,2) por um tempo exponencialmente distribuído com um taxa $c(\alpha)_i$. Mostra-se que o número de *jobs* barrados dentro de um tempo t é decrescente em c em uma ordem estocástica convexa para qualquer $t \geq 0$. Esse trabalho é uma extensão do trabalho Fond e Ross (1978) onde é estudado o caso de C=N=1.

- Grassmann e Zhao (2004) analisam um sistema de filas com servidores heterogêneos com entradas generalizadas. Mostra-se como obter as probabilidades de estado permanente para esse sistema em pontos aleatórios no tempo e para um ponto no tempo que precede alguma chegada. Para determinar as probabilidades para os pontos no

tempo em que ocorrem chegadas, determinam-se as probabilidades para todos os estados em que o cliente não encontra fila, e depois determinam as probabilidades para quando já havia clientes esperando na hora de uma chegada. As probabilidades nos tempos de chegada permitem achar a probabilidade de não ter espera no sistema, a distribuição do tempo de espera e a distribuição do número de pessoas no sistema em pontos aleatórios do tempo. Através de argumentos heurísticos e cálculos numéricos, mostra-se que a importância da alocação é inversamente proporcional à intensidade de trafego no sistema, à variância entre os intervalos entre chegadas e diretamente proporcional ao número de servidores. Eles também definem regras para a decisão de alocação quando há mais de um servidor livre.

- Harten e Sleptchenko (2003) desenvolvem um estudo sobre sistemas de filas MCMS (Multi-Class Multi-server). Considera-se, então, um sistema de fila M/M/k com k classes e servidores exponencialmente distribuídos, mas, com taxas de processamento diferentes. Desenvolve-se um procedimento para a construção de soluções exatas para as equações de estado estacionário. Nesse procedimento faz-se uma redução de parte do problema para uma equação diferencial de segunda ordem. Mostra-se que a solução exata pode ser achada através de decomposição dos autovalores dessas. Com as soluções exatas para as equações de estado, computam-se as medidas de performance para o sistema para depois compará-las com aproximações heurísticas encontradas na literatura. Em seguida, ilustram-se os métodos desenvolvidos no artigo com resultados numéricos e demonstram-se algumas aplicações úteis para esses.

- Boxma, Deng, e Zwart (2002) estudam uma fila heterogênea M/M/2 e definem o tempo de serviço do primeiro servidor como sendo exponencialmente distribuído, enquanto que, no segundo servidor o tempo de serviço é representado por uma distribuição genérica B(.). Os autores apresentam uma análise matemática precisa para o tamanho da fila e para a distribuição do tempo de espera para os casos em que B(.) tenha a transformada de Laplace-Stieltjes. Mostra-se que a distribuição do tempo de espera é semi-exponencial se a taxa de chegada de *jobs* no sistema for menor que a taxa de processamento do servidor exponencial.

- Gall (1998b) generaliza o método de fatorização criado por ele mesmo em (Gall, 1998a) para as filas G/G/s, só que agora considerando os servidores como sendo heterogêneos. Apresentam-se três propriedades simples, as quais permitem a construção de um método numérico para os cálculos. Comparam-se os resultados achados com os determinados através de métodos Markovianos clássicos para o caso de um sistema de fila M/G/s simétrica. Compara também o atraso médio em fila encontrado na análise com resultados simulados.

- Singh (1971) em seu trabalho considera um sistema de fila $M/M_i/3$, onde os três servidores são heterogêneos. Encontram-se as seqüências de melhores taxas de serviço, através de investigações numéricas que minimizam as medidas de performance dos sistemas. Singh mostra que para $\rho = \frac{\lambda}{\mu_1+\mu_2+\mu_3}$ existe uma combinação de μ_i que otimiza o sistema.

- Singh (1970) desenvolve um sistema Markoviano com dois servidores heterogêneos que atendem a uma única fila. Os *jobs* são todos de um só tipo e no modelo é permitido que se decidam se os *jobs* vão ou não se juntar ao sistema. A fila considerada no artigo é do tipo $M/M^{[j]}/2/(\beta)$, onde β é a probabilidade de um *job* se juntar ao sistema se no momento da chegada os dois servidores estiverem ocupados. Se os dois servidores estiverem vazios, há duas possibilidades: i) o *job* é alocado para o servidor mais rápido; e ii) o *job* tem uma probabilidade de ser alocado para qualquer um dos servidores. Como o espaço de estados para um sistema de dois servidores é tratável analiticamente, o autor desenvolve uma formulação para o tempo médio de espera e o compara com o tempo médio de espera obtido para um sitema $M/M/2/(\beta)$ com servidores homogêneos.

- Gumbel (1960) caracteriza um sistema de fila M/M/c com c servidores heterogêneos que atendem a uma única classe de *jobs*. Ele assume que para mais de um servidor vazio, o que irá ser alocado para servir um cliente que chega será aleatoriamente escolhido. O autor prova que um sistema equivalente com servidores homogêneos não existe. Calcula-se, então, o erro que existe entre o sistema com servidores

heterogêneos quando este é comparado com um sistema de servidores homogêneos. Considerando o sistema em regime permanente, uma expressão para as probabilidades de estado é apresentada em uma formulação fechada e, a partir dessa, é desenvolvida a formulação para o tamanho esperado da fila.

Dentro da área de teoria de filas surgiu a necessidade de se controlar os sistemas de alocação e de entrada de *jobs*. A teoria de filas controláveis foi amplamente estudada com o intuito de se achar um ponto de controle ótimo para a admissão, o agendamento, o serviço e o roteamento de *jobs* em filas ou em redes de filas ((Rykov, 1975), (Stidham, 1985) e (Stidham e Weber, 1993)). "O principal objetivo da teoria desses processos é provar a propriedade Markoviana de se ter uma estratégia ótima, a qual permita a construção de um política ótima usando métodos numéricos"
(Rykov, 2001). Algumas publicações nessa área serviram de embasamento para este trabalho e são apresentadas a seguir:

- Marmony (2005) propõe um regra de roteamento para um sistema de larga escala com múltiplos servidores heterogêneos e apenas uma classe. A regra proposta é do tipo FSF (Fast Server First), ou seja, aloca o servidor mais rápido livre primeiro. Mostra-se que essa regra minimiza assintoticamente, para o sistema em estado permanente, o tamanho da fila e o tempo virtual de espera. Considera-se um regime com muitos servidores que atendem a um alto tráfego de *jobs*, o qual é chamado regime QED (Quality and Efficiency Driven). O sistema proposto atinge um alto nível de qualidade de serviço e de eficiência fazendo-se um balanceamento entre os dois. A análise feita mostra que o modelo com servidores heterogêneos funciona melhor do que o modelo com servidores homogêneos.

- Rykov e Efrosinin (2004) desenvolvem uma descrição numérica que prova a política de alocação ótima que minimiza o custo operacional do sistema. A escolha do servidor mais rápido é feita como política de alocação, e os autores mostram que esta política reduz o tamanho da fila. Considera-se um sistema de uma única fila controlável e múltiplos servidores heterogêneos.

- Rykov (2001) desenvolve um modelo de fila controlável para múltiplos servidores heterogêneos. O autor expande as propriedades desenvolvidas para dois servidores, generalizando o trabalho de Lin e Kumar (1984). Considera-se o sistema de filas M/M/K/N-K ($K \leq N \leq \infty$) que tem um buffer com apenas N-K lugares. A chegada segue o modelo de um processo de Poisson com taxa λ e o sistema tem K servidores heterogêneos exponencialmente distribuídos. Considera-se um custo por unidade de tempo em que cada *job* utiliza o serviço e também um custo para o tempo em que o *job* fica parado esperando na fila. No sistema considerado por eles há um controlador que decide se manda ou não um *job* para um servidor que está livre, e clientes são rejeitados no sistema se o buffer estiver cheio.

- Nobel e Tijms (2000) analisam um modelo de filas com chegadas de Poisson em bando BPI (Batch Poisson Input) em um sistema com dois servidores exponenciais heterogêneos. O servidor mais rápido está sempre em serviço e o mais lento é usado só quando a fila atinge um tamanho limite, pois, esse incorre em custos fixos para ser ativado. Desenvolve-se um algoritmo para achar a melhor regra para gerar um nível ótimo para a decisão de ativar ou não o servidor lento.

- Righter (2000) estuda um sistema com fila M/M/2, onde os servidores são heterogêneos. No sistema há múltiplas classes de clientes e é associado aos clientes de cada classe uma taxa de recompensa (reward rate) e um custo por ficar no sistema. Determinam-se prioridades e os clientes com a prioridade mais alta podem tomar o lugar (preempt) dos clientes de classe mais baixa que já estão sendo servidos. Definem-se dois modelos onde a política de alocação ótima e a distribuição de prioridades ótima maximizam o lucro através de uma estrutura de limiar que é dependente do número de clientes de cada classe que existe no sistema. Eles mostram que o limiar ótimo não depende do valor numérico especifico da taxa de remuneração e nem do valor de cada um deles.

- Shenker e Weinrib (1988) consideram um sistema com uma única fila e vários servidores heterogêneos. Eles identificam nesse trabalho uma política de alocação que

minimiza o atraso médio do sistema para o caso em que o número de servidores tende ao infinito. Analisando os servidores apenas com duas taxas diferentes de processamento eles mostram a convergência para o ponto ótimo quando usada essa política. Eles propõem também políticas para sistemas que têm uma grande quantidade, porém finita, de servidores com uma distribuição genérica para as taxas de processamento.

- Lin e Kumar (1984) mostraram que para dois servidores heterogêneos a política que minimiza o número de *jobs* no sistema tem uma propriedade de limiar (veja também Koole (1995) e Walrand (1984)) e essa usa o servidor mais rápido se necessário. Usando argumentos de programação estocástica dinâmica (veja Bellman (1954) e Bellman (1957)) eles provam que a regra de controle ótimo é do tipo limiar, o que era um resultado intuitivamente obvio. O controle criado por eles considera necessário a utilização do servidor mais lento só depois que a fila atinge um tamanho que ultrapasse certo valor de limiar.

Achamos pertinente citar aqui outros trabalhos analisados e que também lidam de alguma forma com sistema SQMS (single-queue, multi-server) e que não foram comentados acima. São eles: (Neuts e Takahashi, 1981), (Foss e Kovalevskii, 1999), (Viniotis e Ephremides, 1988). Ressaltamos aqui que nenhum desses autores desenvolveu uma formulação fechada relacionadas às medidas de desempenho de forma tão genérica ou para o caso de sistemas com presença de mais de três servidores heterogêneos, mostrando o quão este trabalho generaliza o estudo de filas para o caso de servidores heterogêneos e, portanto, garantindo o ineditismo deste.

1.5 Abordagem Proposta

Neste trabalho, pretende-se desenvolver uma formulação para o caso de um sistema com múltiplos servidores. Esses servidores não são tratados de forma homogênea como é o caso da M/M/c. O modelo abordado neste trabalho corresponde ao aprensentado na Figura 1.3. Os *jobs* chegam ao sistema de acordo com uma distribuição de Poisson com taxa λ. No modelo dessa dissertação é considerada apenas uma classe simples de *job*, logo, a taxa λ representa todas as chegadas no sistema. O número de servidores é c para

c = 1, 2, . . ., ∞. O tempo que cada *job* passa em cada servidor é exponencialmente distribuído e cada servidor no sistema tem uma capacidade de processamento específica que é representada pela taxa μ de trabalho. Por motivos de cálculos, os servidores são ordenados em uma ordem crescente de acordo com a taxa de processamento, ou seja, o servidor um será aquele com menor capacidade de trabalho o dois a segunda menor e assim por diante. Temos portanto:

λ Taxa de chegada de clientes no sistema

μ_j Taxa de processamento do servidor j, para j = 1, 2, ..., c. Sendo que, $\mu_1 \leqslant \mu_2 \leqslant \mu_3 \ldots \leqslant \mu_c$.

Cada *job* chega individualmente ao sistema, ou seja, não são consideradas as chegadas em bando (bulk arrivings) e também não é considerado nenhum tipo de prioridade. Além disso, o sistema tem capacidade infinita, ou seja, não há rejeição de *jobs* que chegam ao sistema. Finalmente, não é considerada a perda de *jobs* por nenhum tipo de desistência, troca ou impaciência por parte deles.

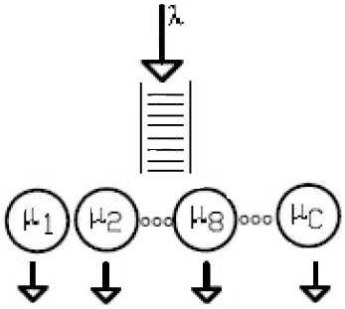


Figura 1.3: Sistema de fila única com c servidores Heterogêneos, modelo estudado neste trabalho.

Como a fila é do tipo FCFS (First-Come, First-Served), os *jobs* que chegam primeiro ao sistema são atendidos primeiro. No modelo, não é permitido que um servidor fique ocioso (livre) quando há trabalho a ser realizado ou seja, quando houver fila. Assim sendo, o primeiro *job* na fila será atendido assim que algum dos servidores terminar de processar o trabalho que estiver realizando.

Como os servidores são heterogêneos, o tipo de alocação utilizada influenciará nas medidas de desempenho desse sistema e consequentemente no modelo. Estuda-se aqui três tipos de alocação: (i) alocação rápida, em que os servidores mais rápidos são alocados primeiro sempre que estiverem disponíveis; (ii) alocação aleatória, onde o próximo *job* é mandado de forma aleatória para qualquer um dos servidores que estiverem livres;

(iii) alocação lenta, na qual os servidores mais lentos que estiverem livres são alocados primeiro. A última alocação considerada (lenta), ainda que não seja muito realista quando comparada com aplicações reais, é muito útil, pois, como será mostrado aqui, esta serve de aproximação para as outras.

1.6 Aplicações

Nesta seção, procura-se juntar casos em que se possa empregar o modelo proposto aqui. Para isso, buscam-se casos aplicáveis existentes na literatura, além de situações práticas. É muito ampla a quantidade de possibilidades existentes nas quais se pode relacionar com o trabalho. Servidores heterogêneos podem ser usados para representar vários tipos de sistemas e, às vezes, com naturezas completamente diferentes. Abaixo são exemplificados alguns modelos obtidos através de trabalhos já publicados.

1.6.1 VoD - Video Sob Demanda

Os sistemas de video sob demanda (VoD) são usados para se iniciar a exemplificação. VoD é um serviço pago de distribuição de vídeo eletrônico para usuários diversos. Esses sistemas permitem que os usuários selecionem e assistam determinados programas ou eventos em vídeo. Esses podem ser achados em um canal de televisão interativa ou em paginas da Web e consistem em envios de conteúdos em formato de vídeo, karaokê, jogos, etc. Para isso, pode-se fazer um "download"(quando todo o vídeo é enviado antes da exibição) ou por "streaming" (quando o vídeo é enviado constantemente durante a exibição). Essa é uma solução encontrada para as transmissões que utilizam tecnologia ADSLou outra tecnologia Banda Larga[2].

Para que todos os usuários sejam atendidos simultaneamente devem ser adotados múltiplos servidores heterogêneos. A heterogeneidade desses servidores se justifica por vários motivos, como, por exemplo, se um novo servidor for adicionado para expandir o sistema VoD ou se um servidor for substituído por outro que tenha falhado, é de se esperar que o novo servidor tenha mais rapidez e uma maior capacidade de armazenagem. A formulação criada aqui serve como ferramenta para a criação de modelos que representem esses sistemas.

[2]Banda larga é o nome usado para definir qualquer conexão acima da velocidade padrão dos modens de linha telefônica.

- Leung e Hou (2005) investigam como direcionar filmes para servidores heterogêneos para que a probabilidade de bloqueio diminua. O artigo trata das seguintes formas de lidar com o problema: (i) *Problem relaxation* - é determinada uma carga ideal com que cada servidor deva lidar e (ii) *Goal programming* - Através de iterações a carga é direcionada e redirecionada para cada servidor até se aproximar de um valor ideal. Desenvolve-se um modelo de fila que considera que os pedidos de vídeo chegam de acordo com um processo de Poisson e que cada servidor atende a vários usuários ao mesmo tempo dependendo da capacidade desse. A medida de performance analisada foi a probabilidade de bloqueio do sistema.

1.6.2 Redes de Comunicações

Um segundo exemplo para a aplicação do modelo é encontrado nas tecnologias para protocolos de Internet, no caso, a tecnologia de voz sobre IP (VoIP). VoIP torna possível estabelecer conversações telefônicas em uma rede IP, tornando a transmissão de voz mais um dos serviços suportados pelas redes de dados. Essa é uma tecnologia que cresce rapidamente e chama atenção não só pela habilidade do IP de carregar tráfego de voz mas, também, por ser capaz de carregar ao mesmo tempo tráfegos de fax e outros sobre a rede de dados. Isso tem implicações como a tomada pela Internet de espaços do mercado da PSTN (Public Switched Telephone Network).

- Liu, Squillante, e Wolf (2001) apresentam uma metodologia para maximizar lucros em ambientes de e-commerce. O modelo é baseado nas receitas que são geradas quando são garantidos níveis de QoS. Os critérios para QoS são derivados de acordos para níveis de serviço (SLAs) entre provedores e clientes. Usa-se um modelo de filas para achar medidas de performance, no caso: *throughput* e atraso médio. Modela-se a rede através de grupos de servidores da Web (Web server farm) que consistem em um sistema de computadores distribuídos que correspondem a servidores heterogêneos que executam classes de pedidos contínuos de dados (Figura 1.4). Os WSF's são modelados por filas únicas com múltiplas classes de clientes que são atendidos por grupos de servidores.

- Foo e Mercankosk (2005) também investigam analiticamente, através de modelos de filas, o efeito que o atraso total, devido a servidores heterogêneos gargalos (que

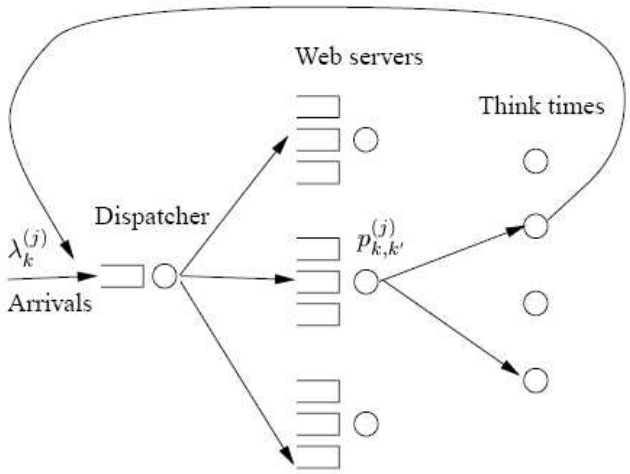


Figura 1.4: Modelo de rede de filas para grupos de servidores na Web.
Figura reproduzida de (Liu et al., 2001).

representam *buffers*), na rede tem sobre os pacotes. Com a definição do atraso total médio, usam-se métodos heurísticos para definir o "atraso" que deve ser dado na hora de se reorganizar os pacotes na última estação.

- Qian, Tipper, e Medhi (1996) apresentam análises comparativas para esquemas de controles de largura de banda sob a condição de trafego não-estacionário em redes B-ISDN. Utilizam-se métodos numéricos para resolver as equações de Chapman-Kolmogorov associadas e para determinar o comportamento não-estacionário. São apresentados resultados para várias medidas de desempenho do sistema e, para isso, faz-se uso de modelos de filas propostos em Eick, Massey, e Whitt (1990) e em Eick, Massey, e Whitt (1993). Relata-se que o processo de chegadas de pedidos para conexão para o tráfego de voz pode ser modelado como um Processo de Poisson não-estacionário e com essa consideração obtém-se aproximações de filas. Considera-se esse como sendo um sistema de fila $M_t/G/\infty$ onde as chegadas são dependentes no tempo e com apenas uma classe.

Outros autores estudam a utilização de *buffer* em vários pontos da rede. Através de formulações de filas Trajkovit e Halfin (1994) calculam possíveis tamanhos de buffer necessários para se garantir QoS com uma perda de pacotes arbitrariamente baixa. Em Wong, Mark, e Chua (2000) é feita também uma análise do tamanho de buffers para que

se tenha um taxa mínima de perda de pacotes em redes ATM sem fio. Considera-se um sistema que consiste em um protocolo MAC (Medium Acess Control) para o uplink e um UPC (Usage Parameter Control) para suporte a servidores heterogêneos na estação base.

O reseqüenciamento de pacotes descrito em Gogate e Panwar (1999) é também uma fonte de estudo e mais uma área onde o modelo desenvolvido aqui pode ser aplicado.

Finalmente, Fulton e Li (1998), entre outros, estudam o comportamento do Jitter[3] em nós da rede sob diferentes condições de tráfego.

Este breve levantamento bibliográfico ilustra a importância do uso de técnicas de filas para a análise da performance de redes de pacotes.

1.6.3 Telefonia

- Daigle (2005) mostra um modelo de fila para ser aplicado em sistemas de telefonia móvel. Para isso utiliza as seguintes premissas: a) um sistema de comunicação tem 832 freqüências e b) há apenas duas operadoras que fornecem os serviços. Com isso, há apenas 416 freqüências para cada operadora sendo que 21 delas têm que ser separadas para sinalização. Os canais são agrupados em grupos de 56, os quais são chamados de células. Para saber como direcionar essas células, é interessante que a operadora estime a probabilidade de bloqueio e, para isso, é interessante estudar o tráfego em hora de pico e, dessa maneira, tem-se que: λ é a taxa de ligações feitas por cliente durante o horário de pico e que $\frac{1}{\mu}$ é o tempo médio que cada cliente gasta por ligação durante a hora de pico. Sendo os tempos entre chegadas são representados por uma exponencial i.i.d., obtém-se a probabilidade de bloqueio e o tamanho médio da fila que representa o tanto que o canal é utilizado (o quanto do espectro de freqüência é aproveitado).

1.6.4 Multiplexação de Pacotes em Switches

Os dados na Internet passam por multiplexadores em roteadores e são mandados para as linhas de comunicação de dados fim-a-fim e depois de passar novamente em roteadores, passam por demultiplexadores. Filas são formadas em vários pontos nesse processo e a criação de modelos para a melhor compreensão desses é muito bem vinda.

[3] O Jitter (Delay Jitter) é uma flutuação no tempo de chegada dos pacotes e, essa é causa pela variação dos atrasos que ocorrem durante o percurso pela rede.

- Daigle (2005) exemplifica uma aplicação para um modelo que é usado para representar a fila que surge em roteadores durante a multiplexação de pacotes. Foca-se apenas nas filas formadas nas portas de saídas desses roteadores. Normalmente, um roteador tem N portas de entrada e N de saída e, quando os pacotes chegam a um processador de um roteador, eles são particionados em blocos de dados de tamanho fixo e, em seguida, esses blocos são comutados (switched). Daigle, estuda a influência das N portas e dos tipos de distribuição de chegadas no tamanho da fila, para um determinado tráfego. Faz-se a suposição de que o sistema é dividido em *slots* de tempo, sendo que cada um desses *slots* é o tempo para que um pacote entre ou saia do roteador. Portanto, N pacotes podem chegar nesse *switch* por *slot* de tempo e como só um pacote pode sair por slot, há uma fila que se forma e é armazenada em buffer.

Com o modelo criado aqui, pode-se acrescentar no estudo feito por Daigle(2005) a heterogeneidade relativa a esse processo.

1.6.5 Central de Atendimento (Call Center)

Desde 1878, quando a empresa de telefonia Bell (Bell Telephone Company) começou a usar operadores para conectar as ligações, o uso de centros de atendimento aumentou para vários fins, o que acarretou em um incrível crescimento de sua importância na economia em geral. Dados mostram que, nos Estados Unidos, três por cento da população trabalha em centros de atendimento, o que significa que tem mais gente trabalhando nessa área do que na agricultura por exemplo (Texto tirado de Pinedo, Seshadri, e Shanthikumar (2003)).

Os centros de atendimento têm fundamental importância para algumas indústrias, como:

1- indústria de telecomunicações;

2- indústrias da Aviação; e

3- indústria para vendas produtos e serviços em geral.

Os centros de atendimento, (Figura 1.5), têm diferentes propósitos dentro de uma companhia, os quais dependem de vários fatores. Esses podem ser usados, por exemplo, para informação, fazer reservas, fazer compras, pedir conselhos (médicos, por exemplo) ou para abrir uma conta. É muito usado em empresas de seguro, mercado financeiro e outros.

Vários autores vêm estudando modelos de filas que possam representar esses sistemas. Recentemente, pesquisadores têm focado seus esforços em modelos com processos de chegadas não-estacionárias. Entretanto, para que possamos usar o modelo desenvolvido aqui precisaríamos considerar intervalos de tempo que são estacionários. Uma boa referencia para esse tipo de sistema pode ser achada em (Melnick, Pinedo, Nayyar, e Seshardi, 1999).

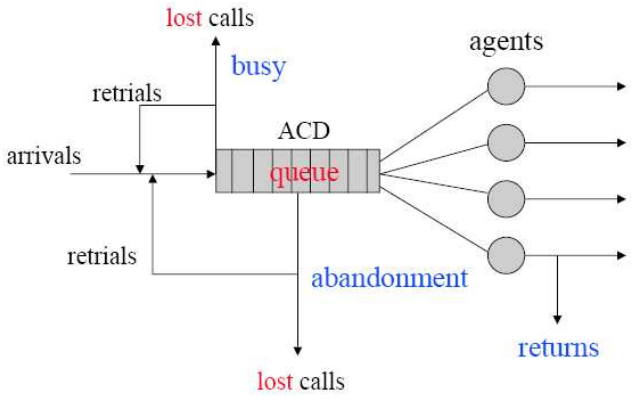


Figura 1.5: Modelo Operacional de um centro de atendimento.
Figura tirada de (Koole e Mandelbaum, 2001)

Normalmente os usuários estão interessados em medidas do tipo: probabilidade de atraso, quantidade de perda de clientes, tamanho da fila de espera e o atraso experimentado por clientes.

- Green e Kolesar (1991) considera um sistema em que o tempo de serviço dos clientes é independente e identicamente distribuído de acordo com uma distribuição exponencial com taxa média de serviço igual a μ. O número de servidores é constante e igual a S. O processo de chegadas é assumido como sendo um Processo não-homogêneo de Poisson com taxa de chegada $\lambda(t)$ no tempo t.

Green e Kolesar (1991) consideram que os servidores são homogêneos, o que na prática é muito difícil de acontecer. O modelo desenvolvido aqui permite criar modelos mais realistas para tais sistemas.

1.6.6 Modelos Financeiros

Em operações financeiras há riscos, e esses podem ser definidos como a incerteza associada aos retornos esperados. Não há como evitar o risco. Portanto, deve-se administrá-lo.

Entre as aplicações de filas encontradas para modelos financeiros a apresentada aqui é a que utiliza filas para se calcular riscos de crédito. O Risco de Crédito está associado às possíveis perdas que o credor tenha caso o devedor (contraparte) não honre com os seus compromissos. Toda empresa financeira deve, portanto, fazer um gerenciamento rigoroso de seus créditos de risco. Para isso, mensuram-se a probabilidade de inadimplência, taxa de recuperação, exposição em caso de inadimplência e perda inesperada. A partir de então, são criados modelos para que se estimem os riscos. Com o passar do tempo, as metodologias de modelagem de risco de crédito melhoraram e os bancos foram incorporando modelos aos processos de gradação de risco, precificação, gerenciamentos de carteira e tomadas de decisão. Com a relevância do papel dos modelos de risco de crédito, tornou-se importante compreender as diferentes opções de mensuração dos componentes individuais do risco de crédito e do relacionamento deles entre si. Obtém-se, assim, uma medida completa do risco de crédito.

- Schellhorn e Cossin (2004) criaram modelos de risco de crédito baseados em modelos representados por redes de filas. Precifica-se o débito de três firmas onde a firma A empresta para a firma B que empresta para a C, a qual, por sua vez, fecha o ciclo emprestando para a firma A. Mostra-se que esse ciclo contém efeitos complexos. O modelo criado por eles é equivalente a uma rede Jackson de várias filas do tipo $M/M/\infty$. Cada fila representa uma firma e o dinheiro é a variável de fluxo, o que é representado pelas receitas de cada firma. A saídas que ocorrem nas firmas são as despesas (financeiras ou não). Os servidores são no caso de três tipos: Acionistas, pessoas com dinheiro na firma, e gastos operacionais.

Apesar de terem identificado naturezas diferentes entre os servidores, no modelo matemático criado por Schellhorn e Cossin (2004), não foi feita a consideração dos servidores como sendo heterogêneos.

1.6.7 Outras Aplicações

É impossível pensar e ainda mais identificar todos os casos onde é possível aplicar o modelo de fila aqui proposto. Entre as diversas áreas de prováveis aplicações e que não foram citas acima, encontram-se:

- Logística - existe desde os tempos mais antigos. Na preparação das guerras, líderes militares desde os tempos bíblicos já se utilizavam de meios logísticos para alcançarem seus objetivos. Podemos dizer que a logística trata do planejamento, organização, controle e realização de outras tarefas associadas a armazenagem, transporte e distribuição de bens e serviços. Os modelos de filas são muito ricos quando usados para representar sistemas que se enquadram nessa área, pois, esses muitas vezes representam bem toda a natureza existente. Abaixo estão alguns exemplos simples, para mero efeito de ilustração, da utilização do modelo de filas proposto aqui para os seguintes objetivos:

    1. Armazenagem - Uma empresa petroquímica brasileira produz resinas termoplásticas, como o polietileno, o polipropileno e o PVC. O nafta, que é um derivado do petróleo, é a principal matéria-prima da cadeia produtiva dessa empresa. Em 2006 o preço do barril do petróleo chegou a quase 80 dólares, mas em janeiro de 2007 o preço caiu para 50 dólares o barril. De uma forma simplificada vamos considerar que a empresa passou a comprar a nafta com maior freqüência. As compras ocorriam dependendo da flutuação do preço do barril de petróleo e representam as chegadas no sistema. Considerando-se que a empresa consiga trabalhar toda a matéria-prima rapidamente, pode-se assumir que os servidores serão os clientes dessa empresa. Como esses fazem pedidos que variam em quantidade e em tipo de produto, podemos dizer que esses são heterogêneos. Consequentemente, a fila será o estoque e essa não poderá passar os limites permitidos de armazenagem.
    2. Transporte - Frutas e alimentos são transportados de áreas rurais para áreas urbanas. Pode-se modelar esse sistema da seguinte forma: para uma mesma fazenda, há pedidos feitos por comerciantes localizados em diferentes cidades. Neste exemplo as chegadas são representadas pelos pedidos. Os servidores são os caminhões, sendo que a taxa de serviço de cada um varia de caminhão para caminhão e de estrada para estrada.
    3. Distribuição de bens e serviços - (i) Um supermercado pode ser um exemplo para a distribuição de serviços, onde os caixas são os servidores e a fila pode ser única ou não. Esse tipo de sistema pode ser comparado com os caixas em agências bancárias; (ii) Outro exemplo de distribuição de bens e de serviços seria uma drogaria de Belo Horizonte. A empresa espalhou por toda a cidade

pontos de venda, porém mantém uma central onde é feita a estocagem de produtos que saem para entrega em domicílio. Pode-se considerar os pedidos feitos pelos clientes como sendo as chegadas e as formas de transporte para a entrega os servidores. Com o modelo é possível prever qual a diferença no tamanho da fila para o caso dos entregadores saírem da central e para o caso deles saírem de pontos distribuídos pela cidade.

- Gerenciamento de pessoas - Em um salão de corte de cabelo, os barbeiros são servidores heterogêneos. A taxa de clientes que chegam é tal que sempre há fila no salão. Vale ou não a pena o dono do salão contratar mais um funcionário? Às vezes, ao contratar outro funcionário, o salão passará a ter funcionários ociosos, fazendo com que, talvez, a margem de lucro diminua. Uma solução é usar o modelo para prever o tamanho médio da fila e a probabilidade de ócio no sistema para os casos i) com o funcionário extra e ii) sem o funcionário extra.

Há mais exemplos a serem dados, principalmente nos sistemas de manufatura e nas linhas de produção em indústrias. O interessante é que com o modelo de filas desenvolvido pode-se aproximar de forma mais eficaz esses sistemas com servidores heterogêneos e obter as medidas de desempenho desses, para que se possa gerenciar, tomar decisões e otimizá-los.

1.7 Organização do Trabalho

A organização do trabalho pode ser vista na Figura 1.6 que está explicada a seguir:

A teoria básica, necessária para a compreensão desta dissertação, encontra-se nos capítulos 2 e 3. No Capítulo 2, especificamente, é apresentada a teoria para compreender os processos estocásticos Markovianos, que são a base para o modelo proposto. Já no Capítulo 3, é mostrado de forma muito resumida os modelos de filas elementares que são, na verdade, governadas pelos processos Markovianos, descritos no Capítulo 2.

No Capítulo 4, é desenvolvida, analiticamente, uma formulação para se obter aproximadamente as medidas de performance para sistemas de fila única FCFS com servidores heterogêneos exponenciais. Para o desenvolvimento dessa formulação, é considerado o processo de Nascimento e Morte (BD) para representar os estados referentes ao número de pessoas no sistema. As chegadas são governadas por um processo de Poisson e os

tempos de serviço são modelados pela distribuição exponencial. Tanto a distribuição exponencial, quanto os processos de Poisson e de Nascimento e Morte, foram descritos no Capítulo 2.

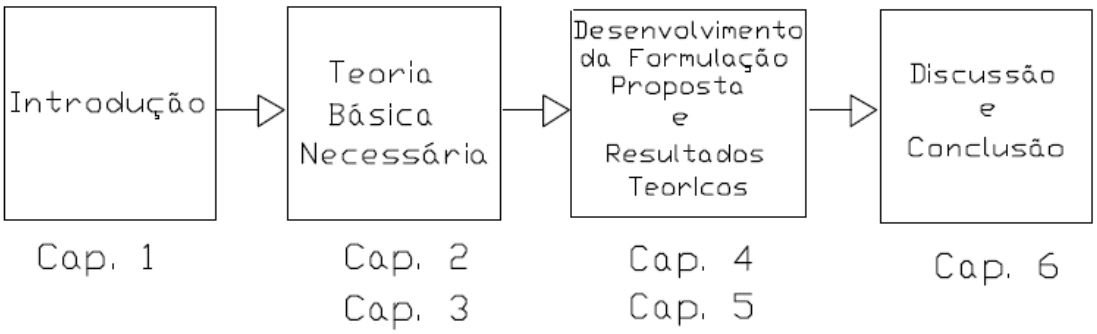


Figura 1.6: Organização dos capítulos da dissertação

No Capítulo 5, avalia-se, através de resultados obtidos com simulações, o comportamento da formulação desenvolvida. A partir das observações feitas e das conclusões tiradas, as fórmulas propostas no Capítulo 4 são comparadas com as fórmulas já existentes na literatura, que são as desenvolvidas para os modelos que consideram os servidores como homogêneos, no caso as filas M/M/c, já descritas no Capítulo 3. O intuito dessa comparação é validar a formulação criada e comparar o erro obtido entre a aproximação com o obtido com as fórmulas para as filas M/M/c. Ainda neste capítulo, são discutidas as capacidades e limitações do modelo desenvolvido no trabalho.

Finalmente, no Capítulo 6, são apresentadas as especulações para aplicações e pesquisas futuras e as conclusões desta dissertação.

Capítulo 2

Processos estocásticos Markovianos

2.1 Cadeias de Markov

Para se definir uma cadeia de Markov, precisa-se, primeiramente, definir eventos independentes. Esses podem ser descritos da seguinte maneira:

Definição 1. *Se X e Y são dois eventos e supondo que a probabilidade P(X) é > 0. Então o evento Y é considerado independente de X se:*

$$P(Y|X) = P(Y). \tag{2.1}$$

De forma geral, pode-se definir também a independência de n eventos, mas para isso é preciso garantir que a probabilidade de os n eventos acontecerem seja válida para a forma produto.

Definição 2. *Os eventos X_1, X_2, ..., X_n são considerados independentes se:*

$$P(X_{i1}X_{i2}...X_{ik}) = P(X_{i1})P(X_{i2})...P(X_{ik}). \tag{2.2}$$

para todo k = 2,3, ... n e todo $\{i_1, i_2, ..., i_k\} \subset \{1, 2, ..., n\}$ tal que $i_1 < i_2 < ... < i_k$.

Na teoria das Cadeias de Markov de primeira ordem[1] um novo evento depende somente do evento anterior e não mais de todos os eventos anteriores, como o mostrado acima. Portanto, se tivermos um evento X_k esse será associado com a probabilidade p_{ik}

[1] Ao longo deste texto, a expressão "cadeia de Markov" será usada para abreviar"cadeia de Markov de primeira ordem", uma vez que cadeias de ordem mais elevada não serão tratadas.

do par de eventos (X_i, X_k). Logo, dado que um evento X_i ocorreu, a probabilidade de um evento X_k ocorrer será p_{ik} vezes a probabilidade a_i de ocorrência do evento X_i na primeira rodada. Tem-se:

$$P(X_i, X_k) = a_i p_{ik} \qquad P\{(X_i, X_k, X_r)\} = a_i p_{ik} p_{kr}$$

$$P\{(X_i, X_k, X_r, X_s)\} = a_i p_{ik} p_{kr} p_{rs}$$

$$P\{(X_{i_0}, X_{i_1}, ..., X_{i_n})\} = a_{i_0} p_{i_0 i_1} p_{i_1 i_2} ... p_{i_{n-2} i_{n-1}} p_{i_{n-1} i_n}. \tag{2.3}$$

Definição 3. *Uma seqüência onde os eventos são X_1, X_2,..., é chamada Cadeia de Markov*[2] *se as probabilidades da seqüência amostrada são definidas pela Equação 2.3 em termos da distribuição de probabilidade a_k sendo E_k o evento ocorrido no estado inicial e a probabilidade condicional é p_{ik} dado que E_i ocorreu em seguida ao evento inicial.*

As probabilidades p_{ik} são chamadas de Probabilidades de Transição e essas serão arranjadas em uma matrix P de probabilidades de transição:

$$\begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots \\ p_{21} & p_{22} & p_{23} & \cdots \\ p_{31} & p_{32} & p_{33} & \cdots \\ . & . & . & \cdots \\ . & . & . & \cdots \\ . & . & . & \cdots \end{bmatrix}, \tag{2.4}$$

Onde as linhas são referenciadas pelo primeiro subscrito e as colunas pelo segundo. A matriz P acima é quadrada positiva e a soma das probabilidades em cada uma de suas linhas é um.

2.1.1 Processos Markovianos Discretos no Tempo

Em aplicações, colocam-se as cadeias de Markov em função de Variáveis Aleatórias, e daí surge o processo de Markov. O termo "Processo de Markov" é referente a uma classe

[2]Considerando aqui apenas uma classe das Cadeias de Markov. Aqui o termo é usado considerando somente os casos em que as probabilidades de Transição são Estacionárias

grande e importante dentre as muitas classes Estocásticas, as quais podem ser, discretas ou contínuas. Nos processos de Markov consideramos o número de eventos ocorridos, que é representado pelo inteiro k. O estado do sistema em um ponto n qualquer da linha de tempo é representado pela Variável Aleatória $V^{(n)}$ que assumirá com a probabilidade $a_k^{(n)}$ o valor k. A distribuição conjunta de $V^{(n)}$ com $V^{(n+1)}$ é dada por $P\{V^{(n)} = i, V^{(n+1)} = k\} = a_i^{(n)} p_{ik}$, e a distribuição conjunta de $(V^{(0)}, ..., V^{(n)})$ é dada pela Equação (2.3).

Nos processos Markovianos o presente determina a probabilidade do futuro, ou seja, o último evento é uma conseqüência do que foi feito no presente. Entretanto, observe que, o passado que foi o responsável pelo sistema estar em um estado $V^{(p)}$, onde p representa o tempo presente, não tem nenhuma influencia sobre o futuro. Conseqüentemente, pode-se afirmar apenas que se o sistema chegou a um estado futuro $V^{(f)}$ qualquer, este tem que ter passado pelo estado presente consecutivamente antes $V^{(p)}$ e nada mais poderá ser dito sobre os outros estados.

Definição 4. *Uma seqüência de variáveis aleatórias discretas pode ser chamada de processo de Markov se: Para um grupo finito de números inteiros* $n_1 < n_2 < ... < n_r < n$ *a distribuição conjunta dos estados* $(V^{(n_1)}, V^{(n_2)}, ..., V^{(n_r)}, V^{(n)})$ *é definida de tal forma que a probabilidade condicional de se estar no estado* $V^{(n)} = y$ *dado que* $V^{(n_1)} = y_1, ..., V^{(n_r)} = y_r$ *é igual à probabilidade condicional de se estar no estado* $V^{(n)} = y$ *dado apenas que* $V^{(n_r)} = y_r$. *Sendo que* $y_1, ..., y_r$ *são valores arbitrários.*

É necessário lembrar que a Cadeia de Markov referida aqui é apenas uma classe das Cadeias genéricas de Markov, e é obviamente um processo Markoviano. Essa cadeia tem uma particularidade, que é o fato de suas Probabilidades de Transição $p_{ik} = P\{V^{(m+1)} = k | V^{(m)} = i\}$ serem independentes de m. Uma equação mais geral para essas probabilidades é dada por:

$$p_{ik}^{(n-m)} = P\{V^{(n)} = k | V^{(m)} = i\} \qquad (m < n). \tag{2.5}$$

Observa-se que estas dependem somente da diferença n - m. Por isso essas probabilidades de transição são chamadas de estacionárias (homogêneas no tempo).

Pode ser visto pela Equação 2.5 que o lado direito depende de m e n. E, por isso, para se determinar p_{ik}(m,n) é feita a consideração de que p_{ik}(n,n+1) é a probabilidade de transição de um passo. Tem-se então:

$$p_{ik}(m,n) = \sum_{v} p_{iv}(m,r)p_{vk}(r,n), \tag{2.6}$$

para todo m $<$ r $<$ n. A Equação 2.6 acima é chamada de Chapman-Kolmogorov.

A figura 2.1 mostra um diagrama de Venn para ajudar na visualização das relações entre os processos Semi-Markovianos, Markovianos e suas respectivas classes especiais que estão contidas dentro deles. Nesta dissertação não há detalhes para cada um dos processos apresentados no diagrama, pois, não é o foco do trabalho. Essa limita-se apenas ao processo Markoviano contínuo, e aos processos de Nascimento e Morte (BD), Puro Nascimento (PB) e de Poisson (PP).

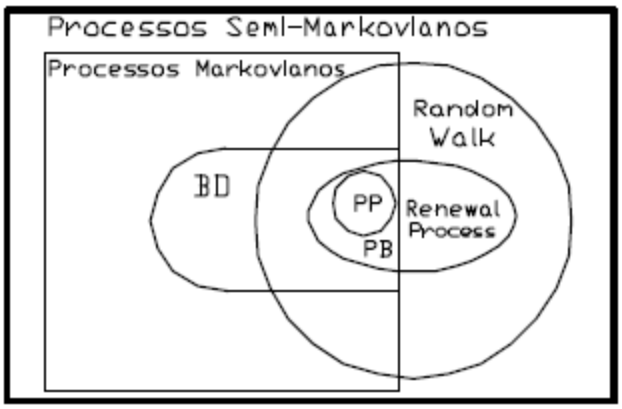


Figura 2.1: Diagrama de Venn para maior compreensão das relações entre os processos aleatórios: SMP (Semi-Markov process), MP (Markov process), RW (Random Walk), RP (Renewal process), BD (Birth-Death process), PB (Pure-Birth) e PP(Poisson process).

2.1.2 Classificações para as Cadeias Markovianas

Neste ponto, são definidas algumas probabilidades antes de se falar sobre as classificações possíveis para as Cadeias Markovianas.

Definição 5. *A probabilidade do sistema voltar pela primeira vez ao estado E_i em n transições após ter saído dele, é:*

$$f_i^{(n)} = P[Voltar\ em\ n\ transições\ pela\ 1^a\ vez\ para\ E_i\ depois\ de\ ter\ saido]$$

então a probabilidade de o sistema sempre voltar a E_i é:

$$f_i = \sum_{n=1}^{\infty} f_i^{(n)} = P[sempre\ voltar\ a\ E_i].$$

Definição 6. *Se para um estado i qualquer f_i=1, então o número médio de transições necessárias para se voltar ao estado i é:*

$$M_i = \sum_{n=1}^{\infty} n f_i^{(n)}.$$

M é também chamado de tempo médio de recorrência.

Entre os estados contáveis e discretos de uma cadeia de Markov, algumas peculiaridades podem ser identificadas e, a partir dessas, é possível classificá-los. Algumas características são referentes aos estados individualmente, enquanto outras são referentes ao conjunto de tipos de estados que existe dentro de uma cadeia de Markov. Sabendo as características desses conjuntos, pode-se também classificar essas cadeias. Listamos a seguir as classificações dos estados e das cadeias de Markov.

Classificação dos Estados:

- Conjunto Fechado - Um grupo X de estados é dito "fechado" se apenas com uma transição nenhum dos estados contidos no conjunto X conseguir ir para algum estado de X^c (onde c é o complemento de X).
- Estado Absorvente[3] - Se X é um conjunto Fechado, e se considerarmos que haja apenas um estado dentro de X, esse é chamado de estado absorvente. Sempre que um sistema entrar em um estado e não puder sair mais dele é porque esse entrou em um estado absorvente. Uma condição para esse estado é $p_{ii} = 1$.
- Estado Recorrente - Um estado E_i é chamado de recorrente quando $f_i = 1$, ou seja, o sistema sempre volta nele.
- Estado Recorrente Positivo - Se um estado i é recorrente, então ele é chamado também de positivo se o tempo necessário para o sistema voltar a esse estado é finito, ou seja, $M_i < \infty$. Em uma cadeia Markoviana com um número finito de estados, todos os estados recorrentes são também positivos.

[3] A nomenclatura usada aqui não é a mesma do inglês "absorbing", mas representa a mesma idéia.

- Estado Recorrente Sem Valor - Um estado recorrente i é dito sem valor quando o tempo necessário para o sistema voltar a esse estado é infinito, ou seja, $M_i = \infty$.
- Estado Transiente - Um estado E_i é chamado de transiente quando $f_i < 1$.
- Estado Periódico - Se o sistema só consegue voltar a um estado E_i em 2γ, 3γ, ...(onde gama é um número inteiro), então o estado i é dito periódico, com período γ. Portanto, o estado i tem período γ se p_{ii}^n=0 sempre que n não for divisível por γ.
- Estado Aperiódico - Se o sistema consegue voltar a um estado E_i em γ, 2γ, 3γ, ...só que aqui $\gamma = 1$.
- Estado Ergódigo - Um estado i é dito ergódico se esse for também recorrente positivo, e for aperiódico, ou seja, se $f_i = 1$, $M_i < \infty$, e $\gamma = 1$.

Classificação das Cadeias de Markov:

- Irredutível - Um processo de Markov é dito irredutível se, para qualquer estado i, esse possa ser alcançado de qualquer outro estado j. Para uma cadeia de Markov irredutível sempre há um inteiro n tal que:

$$p_{ik}^{(n)} > 0. \tag{2.7}$$

  Uma outra forma de se definir uma cadeia irredutível é através do conceito de Conjunto Fechado de estados. Se um Conjunto X é Fechado e contém todos os estados de uma cadeia de Markov e se dentro de X não houver nenhum outro conjunto Fechado, então, essa cadeia pode ser classificada com irredutível.
- Redutível - Se X é um conjunto fechado contendo todos os estados de uma cadeia de Markov, e se dentro de X há subconjuntos também fechados, então essa cadeia é redutível.
- Ergódiga - Quando todos os estados de uma cadeia de Markov são ergódigos, então, essa também o é. Uma cadeia de Markov também é dita ergódiga quando as suas probabilidades convergem com $n \to \infty$ para probabilidades estacionárias limites[4].

[4] As probabilidades limites vão ser estudadas na subseção seguinte.

2.1.3 Probabilidades Limite

As cadeias de Markov apresentam uma característica especial em relação às suas probabilidades de transição. Quando $n \to \infty$ na Equação 2.6 a probabilidade p_{ik}^n converge para um valor que passa a ser o mesmo valor independentemente de n. O que está sendo dito aqui é que a matriz P das probabilidades de transição apresentadas na Equação 2.4 converge para valores limites quando o número de transições tende para o infinito (no futuro distante). Com isso, a matriz P, depois de uma grande quantidade de transições, passa a conter essas probabilidades limites do sistema se encontrar em um estado k e, além disso, os valores dessas probabilidades não dependem do estado inicial.

Definição 7. *Uma Cadeia de Markov é dita Homogênea se as probabilidades de transição são independentes de n, como abaixo:*

$$p_{ik} = P\{V^{(n)} = k | V^{(n-1)} = i\} \qquad \forall n. \tag{2.8}$$

Além das probabilidades limites de transição, também existem no sistema as probabilidades limites do sistema se encontrar em um estado i qualquer.

Definição 8. *A probabilidade de um sistema se encontrar em um estado E_i na n-ésima transição é dada por:*

$$\pi_i^n = P\{V^{(n)} = i\}. \tag{2.9}$$

A seguir estão dois teoremas (sem provas). O primeiro é sobre grupos de estados para cadeias irredutíveis de Markov e o segundo é sobre a existência de uma probabilidade limite do sistema se encontrar em um estado i qualquer.

Teorema 2.1.1. *Os estados de uma cadeia irredutivel de Markov são todos transientes ou todos recorrentes positivos ou todos recorrentes sem valor. Se periódicos, então todos os estados têm o mesmo período γ.*

Para um sistema que passa por todos os seus estados o tempo todo, existe uma distribuição estacionária de probabilidade π_i que descreve a probabilidade desse sistema estar em E_i em um momento qualquer no futuro.

Definição 9. *Uma distribuição de probabilidade P_i é dita estacionária se, quando a escolhemos para ser a distribuição do estado inicial, ou seja, $\pi_i^0 = P_i$, para todo e qualquer n $\pi_i^n = P_i$.*

Teorema 2.1.2. *Para uma cadeia Markoviana, irredutível, homogênea e aperiódica as probabilidades-limite*

$$\pi_i = \lim_{n\to\infty} \pi_i^{(n)}, \tag{2.10}$$

sempre existem e são independentes do estado inicial. Além disso,

1. *Todos os estados são transientes ou todos são recorrentes sem valor, ou seja, $\pi_i = 0$ para todo i e não há uma distribuição estacionária, ou*

2. *Todos os estados são recorrentes positivos, ou seja, $\pi_i > 0$ para todo i, na qual o grupo $\{\pi_i\}$ é uma distribuição estacionária de probabilidade e*

$$\pi_i = \frac{1}{M_i}. \tag{2.11}$$

Sendo assim, π_i poderá ser determinada somente através das seguintes equações

$$1 = \sum_j \pi_j, \tag{2.12}$$

e

$$\pi_i = \sum_i \pi_i p_{ji}. \tag{2.13}$$

2.2 Distribuição Exponencial

Uma variável aleatória exponencialmente distribuída tem a propriedade de não ter memória, ou seja, as características probabilisticas dos sistemas representados por ela não se alteram ao longo do tempo. Essa propriedade faz dela uma V.A. de fácil analise. Uma variável aleatória é exponencialmente distribuída quando

$$f(x) = \begin{cases} 0, & se\ x \leqslant 0 \\ \lambda e^{-\lambda x}, & se\ x \geqslant 0. \end{cases} \tag{2.14}$$

A função distribuição cumulativa para uma variável aleatória exponencialmente distribuída é:

$$F(a) = \int_0^a \lambda e^{-\lambda x} dx = 1 - e^{-\lambda a}. \tag{2.15}$$

2.2.1 Função Geradora de Momento

A Função Geradora de Momentos $\Phi(t)$ de uma VA X é definida para todos os valores em t por:

$$\Phi(t) = E[e^{tx}] = \begin{cases} \sum_{x=-\infty}^{\infty} e^{tx} p(x), & \text{para X discreta;} \\ \int_{-\infty}^{+\infty} e^{tx} f(x) dx, & \text{para X continua.} \end{cases} \tag{2.16}$$

onde f(x) é a pdf da VA x e p(x) é a função distribuição de probabilidade da VA x.

A função Φ(t) é chamada de Função Geradora de momento porque todos os momentos de X podem ser obtidos através de derivadas sucessivas de Φ.

$$\Phi'(t) = \frac{d}{dt} E[e^{tX}] = E[\frac{d}{dt} e^{tX}] = E[Xe^{tX}] \tag{2.17}$$

$$\Rightarrow \quad \Phi'(0) = E[X],$$

$$\Phi''(t) = \frac{d}{dt}\Phi'(t) = \frac{d}{dt} E[Xe^{tX}] \ = E[\frac{d}{dt}(Xe^{tX})] \tag{2.18}$$

$$= \ E[X^2 e^{tX}]$$

$$\Rightarrow \quad \Phi''(0) = E[X^2],$$

$$\Phi^n(t) = E[X^n].$$

Na tabela 2.1 estão mostradas as funções geradoras de momentos, a média e a variância para as distribuições Gama, exponencial e Poisson, que são as distribuições usadas neste trabalho.

2.2.2 Propriedades da Distribuição exponencial

Para se modelar matematicamente certos fenômenos reais, é necessário fazer algumas simplificações para que a matemática fique tratável. Mas, por outro lado não se pode fazer tantas simplificações de modo que nosso modelo não seja mais aplicável.

Tabela 2.1: Funções Geradoras de Momentos, Médias e Variâncias

Distribuição	Φ	Média	Variância
Poisson	$\frac{\lambda}{\lambda-t}$	λ	λ
Exponencial	$\frac{\lambda}{\lambda-t}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Gama	$\left(\frac{\lambda}{\lambda-t}\right)^n$	$\frac{n}{\lambda}$	$\frac{n}{\lambda^2}$

Uma simplificação que é frequentemente usada é assumir algumas variáveis aleatórias sendo exponencialmente distribuídas, uma vez que a pdf exponencial tem propriedade de não se deteriorar com o tempo, simplificando a analise.

Com esta propriedade é possível dizer que se o tempo de vida de dois itens forem exponencialmente distribuídos com média $\frac{1}{\lambda}$, então o item que já está em uso por um período t, tem o mesmo tempo residual de vida que o item novo, ou seja, ela é uma distribuição sem memória, pois

$$P\{X > s+t | X > t\} = P\{X > s\} \tag{2.19}$$

para todo $s, t \geqslant 0$.

A única distribuição contínua que se enquadra nessas condições é a exponencial, como mostrado abaixo:

Se

$$\mathrm{P\{X{>}s{+}t|X{>}t\}{=}P\{X{>}s\}},$$

então

$$\frac{P\{X > s+t, X > t\}}{P\{x > t\}} = P\{X > s\}. \tag{2.20}$$

Para que $P\{X > s+t\}$ ocorra, $P\{X > t\}$ tem que ter ocorrido, logo:

$$\mathrm{P\{X{>}s{+}t,X{>}t\}{=}P\{X{>}s{+}t\}}.$$

Portanto a Equação 2.20 torna-se:

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

ou $P\{X > s+t\} = P\{X > s\}P\{X > t\}$. O que somente será verdade para a distribuição exponencial $e^{-\lambda(s+t)} = e^{-\lambda s}e^{-\lambda t}$.

2.2.3 Função taxa de Falha

Uma forma de se observar a ausência de memória da distribuição exponencial é através da função taxa de falha r(t). A função r(t) fornece a probabilidade de X não durar um tempo adicional além de t, sendo t a idade que ele já tem. Esse é um modelo utilizado para descrever o tempo de vida de um componente ou um sistema. A variável aleatória X tem densidade de probabilidade f(x) e distribuição cumulativa de probabilidade F(x), a qual representa, neste caso, tempo de vida. A probabilidade de um sistema não ter falhado até um instante qualquer t, será chamada *confiabilidade do sistema.*

Definição 10. *A confiabilidade de um sistema que tenha a distribuição cumulativa de probabilidade do tempo de vida F(t), é definida por*

$$C(t) = 1 - F(t). \tag{2.21}$$

A taxa de falha é uma função do tempo e corresponde à probabilidade de que aconteça uma falha num intervalo de tempo Δt, muito menor que t, após o instante t, dado que não houve falha até esse instante.

$$r(t) = lim_{\Delta t \to 0}\frac{P\{t < X \leq t + \Delta t | X > t\}}{\Delta t} = lim_{\Delta t \to 0}\frac{P\{t < X \leq t + \Delta t\}}{P\{X > t\}\Delta t}$$

$$= \frac{1}{1 - F(t)} lim_{\Delta t \to 0}\frac{F(t + \Delta t) - F(t)}{\Delta t}$$

$$= \frac{f(t)}{1 - F(t)}.$$

Definição 11. *A taxa de falha r(t) de um sistema cujo tempo de vida tem densidade de probabilidade f(t) e distribuição cumulativa de probabilidade F(t), é dada por:*

$$r(t) = \frac{f(t)}{1 - F(t)}. \tag{2.22}$$

Para a distribuição exponencial a função taxa de falha é constante igual a λ, pois, de acordo com as equações 2.14, 2.15 e 2.22, tem-se:

$$r(t) = \frac{\lambda e^{-\lambda x}}{1 - (1 - e^{-\lambda x})} = \frac{\lambda e^{-\lambda x}}{e^{-\lambda x}}$$

$$r(t) = \lambda \tag{2.23}$$

e a taxa de confiabilidade é dada por

$$C(t) = 1 - F(t) = e^{-\lambda t}$$

Observe, então, que a probabilidade de um item que tem o seu tempo de vida exponencialmente distribuído durar um tempo adicional dt, é constante. Esse fenômeno é uma conseqüência da ausência de memória da distribuição exponencial.

2.3 Distribuição de Poisson

Definição 12. *Uma Variável aleatória X, tomando os valores 0,1,2,... é chamada de variável aleatória de Poisson com parâmetro λ, para $\lambda > 0$ quando:*

$$p(i) = P[X = i] = \frac{e^{-\lambda}\lambda^i}{i!}. \tag{2.24}$$

A distribuição de Poisson na verdade é uma aproximação para a distribuição Binomial quando o número de ensaios de Bernolli tende a infinito e o produto np permanece constante igual a λ (Ross, 1993). Se uma variável aleatória X representa o número de sucessos em n tentativas, tem-se, então, uma distribuição Binomial com parâmetros n e p, onde p é a probabilidade de se ter sucesso:

$$P\{x = i\} = \binom{n}{i} p^i (1 - p)^{n-i} \tag{2.25}$$

Desenvolvendo a Equação 2.25 considerando que a probabilidade p=$\frac{\lambda}{n}$ é muito pequena e que o número de amostras n é muito grande temos:

$$P\{x=i\} = \frac{n!}{(n-i)!i!}p^i(1-p)^{n-i} = \frac{n!}{(n-i)!i!}(\tfrac{\lambda}{n})^i(1-\tfrac{\lambda}{n})^{n-i} =$$

$$= \frac{n(n-1)...(n-i+1)(n-i)!}{(n-i)!i!}(\frac{\lambda}{n})^i\frac{(1-\lambda/n)^n}{(1-\lambda/n)^i} =$$

$$= \frac{n(n-1)...(n-i+1)}{n^i}\frac{\lambda^i}{i!}\frac{(1-\lambda/n)^n}{(1-\lambda/n)^i}. \tag{2.26}$$

De acordo com as considerações feitas acima, quando $n \to \infty$ e $p \to 0$ pode-se afirmar que:

$$(1-\lambda/n)^n \simeq e^{-\lambda}; \tag{2.27}$$

$$(1-\lambda/n)^{-i} \simeq 1; \tag{2.28}$$

$$\frac{n(n-1)...(n-i+1)}{n^i} \simeq 1. \tag{2.29}$$

Subistituindo as equações 2.27, 2.28 e 2.29 na equação 2.26 chega-se à equação para a distribuição de Poisson:

$$P\{x=i\} = \frac{e^{-\lambda}\lambda^i}{i!}.$$

2.4 Processos Markovianos Contínuos no Tempo e o Processo de Poisson

Nas seção 2.1, as Cadeias de Markov foram definidas como processos estocásticos em que o estado futuro dependia simplesmente do estado presente, não tendo nenhuma relação com o estado passado ou com a forma em que o sistema chegou ao estado presente. Foi assumido que as mudanças de estado aconteciam em momentos discretos do tempo t = 0,1,... Além disso, o número de estados, ainda que elevado, eram contáveis, ou seja, havia uma quantidade finita de estados. Entretanto, nesta seção os sistemas representados sofrem mudanças de estado, que continuam tendo um número finito, mas, que ocorrem a qualquer instante de tempo. Tem-se, assim, eventos discretos que fazem um sistema variar de um estado para o outro, os quais dependem de um parâmetro que é contínuo no tempo. Por exemplo, se a variável aleatória X representa o evento chegada de um pacote vindo de uma rede ATM no seu destino final, é certo que X acontecerá a qualquer momento ao longo da linha do tempo.

A probabilidade de transição $P_{ik}^{(n)}$ usada nas Cadeias de Markov como sendo para o caso discreto, daqui para frente é representada por $P_{ik}(t)$, só que para o tempo t contínuo, e é chamada de probabilidade condicional. Isso significa que a probabilidade do sistema estar no estado E_k em t + s é condicionada ao fato de o sistema estar no estado E_i em um tempo s, sendo que s < (t + s). Observa-se que a probabilidade $P_{ik}(t)$ depende somente da duração de t, que nada mais é que um intervalo de tempo entre eventos, e não do momento em que esse intervalo está situado no tempo. Essa probabilidade de transição é chamada de estacionária ou de homogênea no tempo. Através dessa probabilidade, chega-se à identidade de Chapaman-Kolmogorov que é análoga à Equação 2.6 só que aqui considerando o tempo como sendo contínuo:

$$P_{ik}(s+t) = \sum_j P_{ij}(s)P_{jk}(t) \tag{2.30}$$

A Equação 2.4 pode ser interpretada da seguinte maneira: para uma época inicial, que é considerada como sendo o ponto 0 na linha do tempo, e para um sistema que esteja no estado E_i, resulta que, com a probabilidade $P_{ij}(s)$ o sistema vai para um estado intermediário E_j dentro de um intervalo s. E, com probabilidade $P_{jk}(t)$, ele sai do estado intermediário e, dentro de um intervalo de tempo t, vai para o estado E_k. Se for feita a soma em j de todas as possibilidades do sistema passar por qualquer um dos estados

intermediários possíveis ao sair de i e antes de chegar a k tem-se que em algum momento s + t, para s $>$ 0 e t $>$ 0, de acordo com a equação , existe uma probabilidade P_{ik} do sistema sair de i e chegar ao estado k.

Matematicamente, o processo de Poisson é um processo de Markov homogêneo no tempo. Alguns exemplos de eventos aleatórios que podem ser modelados com a ajuda do processo de Poisson como regente de seus comportamentos probabilísticos, são: pacotes que chegam a um roteador durante um intervalo de tempo, desintegração de partículas, chamadas telefônicas e quebra de cromossomos através de radiação. Esses exemplos foram tirados de (Feller, 1957).

Todas essas ocorrências, a princípio, têm a mesma natureza, e o processo de Poisson está relacionado com o número total de eventos N(t) que ocorrem dentro de um intervalo de tempo t. Observa-se que o descrito acima refere-se aos pontos que ocorrem na linha do tempo e que são governados por um processo que se mantém constante ao longo de todo o período. Em outras palavras, a probabilidade de um determinado evento acontecer permanece constante ao longo da linha do tempo de forma uniforme.

O processo de Poisson tem grande importância por ser um processo de fácil analise, pois os tempos entre chegadas são exponencialmente distribuídos. Com isso se faz uso da propriedade da distribuição exponencial não possuir memória. Além disso, esse é um processo que tem aplicabilidade em vários modelos práticos, sendo muitas vezes usado para representar eventos de natureza aleatória como os citados acima.

N (t), t $\geq$ 0 é um processo estocástico que é chamado de processo de contagem se N(t) representar o número total de eventos que acontecem no tempo t.

Um processo de contagem é definido de forma a atender as seguintes condições:

(i) N(t) $\geq$ 0;
(ii) N(t) é um valor inteiro;
(iii) se s $<$ t, então N(s) $\leq$ N(t);
(iv) para s $<$ t, N(t) - N(s) é igual ao número de eventos que ocorrem no intervalo (s, t).

Outros exemplos de processos de contagem, são:

- número de pedidos por produtos que uma loja de pronto entrega recebe em um período de tempo t,
- número de pessoas que nascem no intervalo de tempo t,

- número de ligações telefônicas feitas durante o tempo t.

O Processo de Poisson é um tipo de processo de contagem, cuja definição segue abaixo.

Definição 13. *Um processo estocástico N (t), $t \geq 0$ é dito ser de Poisson, com taxa λ, para $\lambda > 0$ se:*

1. *N(t) $\geq$ 0;*
2. *os incrementos são independentes e estacionários;*
3. *O número de eventos no intervalo de tempo t segue uma distribuição de Poisson com media λ. Ou seja, para s, t não-negativos, temos:*

$$P_i(t) = P\{N(t) = i\} = e^{-\lambda t}\frac{(\lambda t)^i}{i!}. \tag{2.31}$$

Cabe aqui apresentar as seguintes definições:

- Processo com incrementos Independentes: processo em que os números de eventos que ocorrem em intervalos disjuntos de tempo são independentes. Isso acontece quando o número de eventos N(t) que ocorreram no intervalo t é independente do número resultante de eventos N(t + s) - N(t).
- Processo com incrementos Estacionários: processo em que a distribuição de eventos que ocorrem em cada intervalo de tempo depende somente do tamanho do intervalo. Para isso, $N(t_2+s)$ - $N(t_1+s)$ tem a mesma distribuição de $N(t_2)$ - $N(t_1)$, sendo $t_1 < t_2$.

Um processo de Poisson é dito estar em um estado E_i em um tempo $t > 0$ se houve i mudanças de estado. Assim, a probabilidade $P\{N(t) = i\}$ deve ser descrita como sendo a probabilidade de transição referente a uma mudança de estado de um sistema que está em um estado arbitrário E_j em uma época s e que vai para um E_{j+i} até a época s + t. Em um intervalo de tempo s fracionado em i subintervalos de duração $k = \frac{1}{i}$, a probabilidade de sair (haver mudança) do estado j=1,2,...,i qualquer, é $1 - P\{N(k) = 0\}$ e, entã, o número esperado de subintervalos que sofreram mudanças será:

$$i(1 - P\{N(k) = 0\}) = \frac{1 - P\{N(k) = 0\}}{k}. \tag{2.32}$$

Observa-se que, para $k \to 0$, a Equação 2.32 converge para o número esperado de mudanças dentro do intervalo de tempo t, e espera-se que esse seja igual a λ. O processo de Poisson requer também que sempre que haja uma mudança, esta tenha que ocorrer entre um estado E_j e um estado E_{j+1}, o que significa que o número esperado de subintervalos de duração k que contêm mais de uma mudança deve tender a 0 quando $k \to 0$:

$$lim_{k\to 0}\frac{1 - P\{N(k) = 0\} - P\{N(k) = 1\}}{k} = 0. \tag{2.33}$$

Portanto, $P\{N(k) = 0\} = 1 - \lambda k + o(k)$ onde o(k) é uma quantidade de ordem menor que k.

Definição 14. *Uma função qualquer $f(\cdot)$ é dita ser o(k) se:*

$$lim_{k\to 0}\frac{f(k)}{k} = 0. \tag{2.34}$$

Agora é possível definir o Processo Poisson da seguinte maneira:

Definição 15. *Um processo estocástico N (t), $t \geq 0$ é dito ser de Poisson, com taxa λ, para $\lambda > 0$ se:*

1. *N(t) $\geq$ 0;*
2. *os incrementos são independentes e estacionários;*
3. *$P\{N(k) = 1\} = \lambda k + o(k)$;*
4. *$P\{N(k) \geq 2\ \} = o(k)$.*

2.4.1 Distribuição do Tempo entre Chegadas

Para se obter a distribuição dos tempos entre chegadas, tem-se que k_1 é o tempo do primeiro evento e k_i é o tempo decorrido entre o evento (i-1)-ésimo e o i-ésimo evento.

A probabilidade do primeiro evento não acontecer no intervalo k_1 é:

$$P\{k_1 > t\} = P\{N(t) = 0\} = e^{-\lambda t}. \tag{2.35}$$

Pela Equação 2.35 pode-se ver que, o primeiro intervalo entre chegadas é exponencialmente distribuído com taxa $\frac{1}{\lambda}$. O mesmo pode ser provado para k_2, como é mostrado a seguir:

$$P\{k_2 > t\} = E[P\{k_2 > t|k_1\}]$$

$$P\{k_2 > t|k_1 = s\} = P\{N(s, s+t] = 0|k_1 = s\}$$

$$= \ P\{N(s, s+t] = 0\} = e^{-\lambda t}. \tag{2.36}$$

Conclui-se que o tempo entre chegadas de um processo de Poisson tem distribuição exponencial. Com este resultado podemos afirmar que um próximo evento de Poisson acontecerá dentro de um intervalo exponencialmente distribuído. Portanto, esse é um processo sem memória e que tem um tempo médio $\frac{1}{\lambda}$ de intervalo entre os eventos. Como o Processo de Poisson tem incrementos independentes e estacionários, é correto dizer que, a qualquer ponto do tempo, suas probabilidades de transição se reiniciam, ou seja, ele terá a qualquer momento a mesma distribuição de probabilidade que em t = 0.

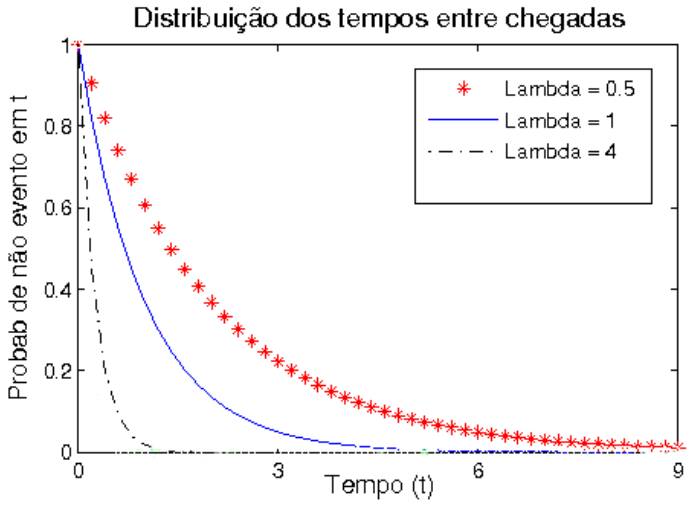


Figura 2.2: Intervalo entre chegadas. A probabilidade de um evento seguinte não ter acontecido diminui exponencialmente com t. O gráfico foi plotado para as taxas de chegadas λ = 0.5, 1 e 4 eventos por U.T.

A Figura 2.2 mostra que a probabilidade de um evento ainda não ter acontecido no tempo t = 0 é máxima e igual à unidade. Ela mostra também que esta probabilidade cai com o tempo, pois é mais provável que aconteça um evento à medida que o tempo passa. Logo, quando o tempo t cresce muito, a probabilidade de nenhum evento ocorrer cai exponencialmente até chegar a 0. Para plotar o gráfico da figura 2.2 foram ajustadas as taxas de chegadas da seguinte forma: $\lambda = 0.5$, 1 e 4 eventos por unidade de tempo.

2.4.2 Distribuição do Tempo de Espera

Para se saber quando um evento específico irá acontecer, é preciso determinar a função de densidade de probabilidade que representa o comportamento desse evento no tempo. Considerando que S_i é o tempo necessário para a chegada do i-ésimo evento a partir de t = 0 ou, em outras palavras, o tempo de espera até o i-ésimo evento, tem-se que:

$$S_i = \sum_{j=1}^{i} k_j, \qquad Para \qquad i \geqslant 1 \tag{2.37}$$

De acordo com a Equação 2.37, o tempo de espera S_i nada mais é que a soma de todos os intervalos k_j entre eventos até que ocorra o evento i.

Como é de se esperar que o i-ésimo evento ocorra antes de um tempo t qualquer, então, o número de eventos em um intervalo de tempo t tem que ser maior ou igual a i. Consequentemente, a soma de todos os intervalos k_j tem que ser menor ou igual ao tempo total t,

$$N(t) \geqslant i \longleftrightarrow S_i \leq t. \tag{2.38}$$

De acordo com a propriedade mostrada na Equação 2.38, é desenvolvida a seguinte formulação para achar a densidade de probabilidade $f_{s_i}(t)$.

$$F_{S_i}(t) = P\{S_i \leq t\} = P\{N(t) \geqslant i\} = \sum_{j=i}^{\infty} e^{-\lambda t} \frac{(\lambda t)^j}{j!} \tag{2.39}$$

$$f_{s_i}(t) = \frac{dF_{S_i}(t)}{dt} = \frac{d\sum_{j=i}^{\infty} e^{-\lambda t}\frac{(\lambda t)^j}{j!}}{dt} \tag{2.40}$$

$$= \sum_{j=i}^{\infty} -\lambda e^{-\lambda t}\frac{(\lambda t)^j}{j!} + \lambda e^{-\lambda t}\frac{j(\lambda t)^{j-1}}{j!}$$

$$= \lambda e^{-\lambda t}(\sum_{j=i}^{\infty} \frac{(\lambda t)^{j-1}}{(j-1)!} - \sum_{j=i}^{\infty} \frac{(\lambda t)^{j}}{j!})$$

$$= \lambda e^{-\lambda t}(\frac{(\lambda t)^{i-1}}{(i-1)!} + \sum_{j=i+1}^{\infty} \frac{(\lambda t)^{j-1}}{(j-1)!} - \sum_{j=i}^{\infty} \frac{(\lambda t)^{j}}{j!})$$

$$f_{s_i}(t) = \lambda e^{-\lambda t} \frac{(\lambda t)^{i-1}}{(i-1)!} \tag{2.41}$$

Com as manipulações acima chegamos à pdf $f_{s_i}(t)$ do tempo de espera s_i e, como pode ser visto na Equação 2.41, essa tem distribuição Gama com parâmetros i e λ (Ross, 1993).

2.4.3 Distribuição Condicional dos tempos de chegadas

Supondo que um evento pertencente a um processo de Poisson aconteça em um ponto qualquer na linha do tempo e se queira determinar a distribuição de probabilidade que rege o intervalo de tempo t em que esse evento ocorreu. Como o processo de Poisson tem incrementos independentes e estacionários, é de se esperar que, em intervalos de tempo de mesmo tamanho [0, t], ainda que esses sejam disjuntos, a probabilidade de conter um evento é igual em todos eles. Com isso, é correto dizer que o tempo de ocorrência de um evento de Poisson é uniformemente distribuído no intervalo [0, t], como mostrado abaixo:

Para $s \leq t$, tem-se:

$$P\{k_1 < s | N(t) = 1\} = \frac{P\{k_1 < s, N(t) = 1\}}{P\{N(t) = 1\}} \tag{2.42}$$

$$= \frac{P\{N(s) = 1), N(t-s) = 0\}}{P\{N(t) = 1\}}$$

$$= \frac{P\{N(s) = 1)\}P\{N(t-s) = 0\}}{P\{N(t) = 1\}}$$

como, em um processo de Poisson,

$$P\{N(t) = i\} = e^{-\lambda t}\frac{(\lambda t)^i}{i!},$$

$$P\{k_1 < s | N(t) = 1\} = \frac{\lambda s e^{-\lambda s} e^{-\lambda(t-s)}}{\lambda t e^{-\lambda t}} = \frac{s}{t}. \tag{2.43}$$

2.5 Processo de Nascimento e Morte

O processo de Nascimento e Morte BD (birth-death) é também uma classe especial dentro dos processos de Markov. No processo BD, sempre que ocorrer uma mudança no sistema, esse tem que ir necessariamente para um estado vizinho antecessor ou para um estado vizinho sucessor. Assim sendo, o sistema só pode sair de um estado E_i para ir para algum dos estados E_{i-1} ou E_{i+1}[5]. O processo de Nascimento-Morte é extremamente importante para se dar continuidade, nos capítulos seguintes, aos estudos de filas e para que se entenda a teoria proposta neste trabalho.

O processo BD pode ser tanto discreto no tempo, quanto contínuo. Vamos considerar neste trabalho apenas os casos em que esse é contínuo no tempo e, por isso, é necessário que o sistema fique num estado E_i apenas por um intervalo finito de tempo. Ao fim desse intervalo o sistema tem que mudar para algum dos estados vizinhos. O leitor deve estar atento ao fato de que o processo BD considerado aqui, apesar de ser contínuo no tempo, tem o espaço de estados discreto.

O Processo BD é muito utilizado para o estudo de populações. O sistema em um estado E_i qualquer, ou seja, N(t) = i, significa que a população é de tamanho i. Para essa população i, o sistema terá uma taxa de Nascimento λ_i que fará com que o sistema suba para o estado E_{i+1} e, ainda no estado E_i, o sistema terá uma taxa de Mortalidade μ_i que fará com que o sistema volte para o estado E_{j-1}. Os processos em que essa as taxas μ_i são constantes iguais a zero, são chamados de Processos de Nascimento (Pure-Birth) e os processos em que as taxas λ_i são constantes iguais a zero, são chamados de Processos de Morte (Pure-Death).

Portanto, se em algum momento o sistema estiver no estado E_i, a probabilidade condicional de ocorrer uma transição de $E_i \rightarrow E_{i+1}$ durante o intervalo de tempo (t,t + k) será

[5]Esta sendo considerado aqui apenas o processo BD unidimensional

igual a $\lambda_i k + o(k)$, quando k$\rightarrow$ 0 e a probabilidade condicional de $E_i \rightarrow E_{i-1}$ será igual a $\mu_i k + o(k)$ quando k$\rightarrow$ 0. Se forem somadas todas as probabilidades,

$$P\{N(t+k) = i\} = \sum_{j=0}^{\infty} P\{N(t+k) = i | N(t) = j\} P\{N(t) = j\} \tag{2.44}$$

Se fizermos $k \rightarrow 0$, então:

$$P\{N(t+k) = i | N(t) = j\} = \begin{cases} \lambda_{i-1} k + o(k), & para\ j = i-1 \\ \mu_{i+1} k + o(k), & para\ j = i+1 \\ o(k), & para\ |j-i| \geqslant 2. \end{cases} \tag{2.45}$$

Considerando o caso em que $k \rightarrow 0$, tem-se:

$$P\{N(t+k) = i | N(t) = i\} = 1 - (\lambda_i + \mu_i)k + o(k). \tag{2.46}$$

Chamando $P\{N(t+k) = i\} = P_i(t+k)$ e considerando as seguintes condições:

$$[k \rightarrow 0; \quad j = 0, 1, ...; \quad \lambda_{-1} = \mu_0 = P_{-1}(t) = 0] \tag{2.47}$$

A equação 2.44 pode ser reescrita da seguinte forma:

$$P_i(t+k) = \lambda_{i-1} k P_{i-1}(t) + \mu_{i+1} k P_{i+1}(t) + [1 - (\lambda_i + \mu_i)k] P_i(t) + o(k).$$

Rearranjando e dividindo por k:

$$\begin{aligned} \frac{P\{N(t+k) = i\} - P\{N(t) = i\}}{k} &= \lambda_{i-1} P_{i-1}(t) + \mu_{i+1} P_{i+1}(t) \\ &\quad -(\lambda_i + \mu_i) P_i(t) + \tfrac{o(k)}{k}. \end{aligned} \tag{2.48}$$

Como $k \to 0$, o termo esquerdo da equação 2.48 nada mais é do que $dP_i(t)/dt$, o que resulta na equação diferencial de variação para o processo BD.

$$\frac{d}{dt}P_j(t) = \lambda_{i-1}P_{i-1}(t) + \mu_{i+1}P_{i+1}(t) - (\lambda_i + \mu_i)P_i(t), \tag{2.49}$$

onde as condições iniciais em t = 0 são,

$$P_j(0) = \begin{cases} 1 & i = j, \\ 0 & i \neq j. \end{cases} \tag{2.50}$$

O leitor deve observar que, para o caso de processos puros de Nascimento ou de Morte, a equação diferencial de variação 2.49 pode ser resolvida de forma recursiva. No estudo desenvolvido nessa dissertação, é considerado um processo de Nascimento e Morte. Entretanto, a taxa λ_i é constante e igual λ, o que representa uma taxa constante no tempo de chegadas de jobs no sistema. A taxa μ_i continua variável e dependente do número de servidores ocupados no sistema. Todavia, se o processo de chegada de jobs no sistema for considerado de forma isolada, então, esse será um processo Puro de Nascimento com taxa $\lambda_i = \lambda$. Considerando ainda que não houve chegadas e, portanto, o processo esteja no estado E_0, as equações 2.49 e 2.50 se transformam em:

$$\frac{d}{dt}P_j(t) = \lambda P_{i-1}(t) - \lambda P_i(t) \tag{2.51}$$

e

$$P\{N(0) = i\} = \begin{cases} 1 & i = 0, \\ 0 & i \neq 0. \end{cases} \tag{2.52}$$

Desenvolvendo as equações 2.51 e 2.52 por sucessivas substituições chega-se a seguinte formula:

$$P_i(t) = \frac{(\lambda t)^i}{i!}e^{-\lambda t} \tag{2.53}$$

A equação 2.53 achada, nada mais é do que a formula da distribuição de Poisson com média λt. Logo, N(t), t ≥ 0 é um Processo de Poisson que, por sua vez, respeita a condição necessária de normalização $\sum_{i=0}^{\infty} P_i(t) = 1$ quando t ≥ 0.

2.6 Sumário

Até aqui, vimos os principais conceitos referêntes aos processos Markovianos e às cadeias Markovianas. No capítulo 3 e no capítulo 4, utilizaremos os resultados obtidos aqui como base para a análise de modelos de filas.

Capítulo 3

Sistemas de Filas Markovianas

Neste capítulo são discutidos alguns modelos básicos de filas. Estes são os mais simples modelos encontrados. Entretanto, são também a base para todo o desenvolvimento posterior. Essas filas são chamadas de Markovianas porque os modelos que as utilizam são desenvolvidos através de análises de processos Markovianos, como será mostrado mais à frente.

Os modelos de filas mais básicos podem ser definidos, normalmente, através de três características: (i) o processo de chegada de jobs; (ii) a disciplina de fila utilizada; e (iii) o mecanismo de serviço. O processo de chegada descreve a seqüência de pedidos por serviço e aqui nos limitaremos aos casos que esse é representado por um processo de Poisson. A disciplina utilizada depende da fila e será tratada de forma específica de acordo com cada tipo de modelo. Por fim, o mecanismo de serviço inclui características como o número de servidores a taxa de processamento. Assim como neste capítulo, toda a formulação desenvolvida nesta dissertação é baseada na suposição de que os tempos de serviços são independentes e identicamente distribuídos de acordo com uma distribuição exponencial.

3.1 Soluções Gerais: Equações de Equilíbrio

Nesta seção outros caminhos são apresentados para que se possa buscar um maior conhecimento sobre os processos BD sem precisar solucionar as probabilidades dependentes no tempo $P_i(t)$. A Equação 2.49 se torna intratável quando é aumentada a quantidade de informação necessária para se compreender um determinado sistema. Para contornar esse obstáculo é preciso parar de olhar para o sistema quando esse está em um estado transiente e passar a analisá-lo apenas quando todas as probabilidades já estão estáveis.

Definição 16. *P_i é a probabilidade de um sistema conter i jobs em algum momento distante no tempo t:*

$$p_i = \lim_{t\to\infty} P\{N(t) = i\}. \tag{3.1}$$

Com o limite na Equação 3.1, a probabilidade p_i deixa de ser dependente do tempo t. Contudo, deve-se atentar para o fato de que, no sistema, ainda há mudanças entre estados, embora essa probabilidade não esteja mais relacionada com o tempo. Portanto, p_i somente descreve a probabilidade de se achar i jobs no sistema quando esse estiver estável.

Com o limite na Equação 3.1 a equação de Kolmogorov 2.49 passa a ter o seu lado esquerdo nulo, pois, $\lim \frac{P_i(t)}{dt} \to 0$ quando $t \to \infty$. Considerando também as condições iniciais apresentadas na equação 2.47, tem-se:

$$0 = -(\lambda_i + \mu_i)p_i + \lambda_{i-1}p_{i-1} + \mu_{i+1}p_{i+1}. \tag{3.2}$$

Neste ponto, é interessante usar equações de equilíbrio para tratar a equação 3.2. Essas equações de equilíbrio são fundamentadas na conservação dos fluxos, onde todo o fluxo entrante tem que ser igual ao fluxo de saída. Pelo diagrama de transição de estados para um processo BD apresentado na Figura 3.1, podem-se ver os i estados e as suas respectivas taxas de transmissão (Nascimento + Morte). Concentrando-se apenas no estado i, e montando as equações de equilíbrio para esse estado, tem-se:

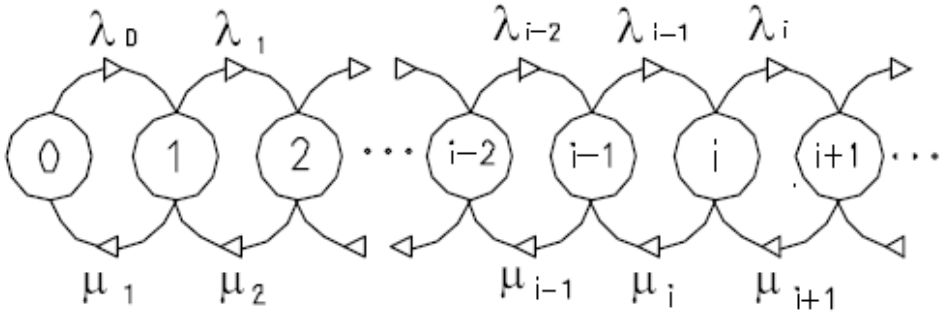


Figura 3.1: Diagrama de transição de estados para um processo BD.

$$Fluxo\ entrando\ em\ E_i = \lambda_{i-1}p_{i-1} + \mu_{i+1}p_{i+1},$$

e

$$Fluxo\ saindo\ de\ E_i = (\lambda_i + \mu_i)p_i.$$

Como, em equilíbrio as duas equações acima devem ser iguais, tem-se que:

$$(\lambda_i + \mu_i)p_i = \lambda_{i-1}p_{i-1} + \mu_{i+1}p_{i+1}. \tag{3.3}$$

As equações 3.3 e 3.2 são iguais, como era de se esperar. Elas são as equações de diferença em equilíbrio que representam as mudanças de estado quando o sistema está em estado perene. Para se achar uma equação geral para p_i pode-se, através de um método recursivo[1], realizar consecutivas substituições, sempre colocando p_0 em evidência e, assim, chegar em uma formulação geral para essa. A seguir é ilustrado esse método e, para isso, inicia-se a busca analisando as equações de equilíbrio para o estado E_0, ou seja, para i = 0 na equação 3.3:

$$(\lambda_0 + \mu_0)p_0 = \lambda_{-1}p_{-1} + \mu_1 p_1$$

Reescrevendo levando em conta as condições iniciais, $\mu_0 = \lambda_{-1} = 0$ mostradas em 2.47, obtém-se:

$$p_1 = \frac{\lambda_0}{\mu_1}p_0 \tag{3.4}$$

Se for feito o mesmo para o estado E_1, portanto, ajustando i=1 na equação 3.3, chega-se à seguinte equação de equilíbrio:

$$p_2 = \frac{\lambda_0\lambda_1}{\mu_1\mu_2}p_0 \tag{3.5}$$

[1]Em (Kleinrock, 1976a) página 93 foi desenvolvido também um outro método para se chegar na equação geral.

Com as sucessivas substituições e sempre deixando p_0 em evidencia obtém-se a seguinte equação produto:

$$p_i = \frac{\lambda_0\lambda_1...\lambda_{i-1}}{\mu_1\mu_2...\mu_i}p_0 = p_0\prod_{j=0}^{i-1}\frac{\lambda_j}{\mu_{j+1}}, \qquad i = 0, 1, 2, ... \tag{3.6}$$

Como o somatório em i de todas as possibilidades do sistema se encontrar em um estado E_i qualquer deve ser igual à unidade, segue-se que:

$$\sum_{i=0}^{\infty} p_i = 1 \tag{3.7}$$

e, com isso, chega-se à seguinte equação para p_0:

$$p_0 = \frac{1}{1 + \sum_{i=1}^{\infty}\prod_{j=0}^{i-1}\frac{\lambda_j}{\mu_{j+1}}} \tag{3.8}$$

As equações 3.6 e 3.8 são a base para o desenvolvimento proposto neste trabalho. Com elas são determinadas as probabilidades p_i e p_0 para os sistemas com servidores heterogêneos. Para isso, expandem-se as taxas λ_i e μ_i que são variáveis dependentes dos respectivos estados correntes i.

3.2 Lei de Little

Em 1961 J. D. C. Little (Little, 1961) desenvolveu um teorema que consiste na equação L=λW que foi chamada de Lei de Little. Essa equação é uma das mais úteis dentro da Teoria de Filas e, particularmente, para este trabalho.

$$L = \lambda W \tag{3.9}$$

onde

L	é o número médio de jobs no sistema,
W	é o tempo médio de espera e,
λ	é a taxa média de chegada de jobs no sistema

Stidham (1974) fez uma simplificação, sem perda de rigorosidade, da prova para a equação 3.9 proposta por Little.O teorema demonstrado em Stidham(1974) é apresentado a seguir,

Teorema 3.2.1. *Se L(x) é o número de jobs presente em um momento x, L é o número médio de jobs presente ao longo de todo o tempo [0,∞], como mostrado na equação abaixo:*

$$L = lim_{t\to\infty}\frac{1}{t}\int_0^t L(x)dx. \tag{3.10}$$

Definindo N(t) como o número de jobs que chegam no intervalo [0,t], define-se a taxa de chegada λ como

$$\lambda = \lim_{t\to\infty}\frac{N(t)}{t}; \tag{3.11}$$

Finalmente, definindo W_i como o tempo de espera do i-ésimo job, define-se o tempo médio de espera como

$$W = lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} W_i \tag{3.12}$$

Se λ e W existirem e forem finitos, então também será L, e eles serão relacionados de acordo com a equação 3.9: $L = \lambda W$.

A equação 3.9 pode ser usada também com o foco na fila do sistema. Ela pode ser usada também como a relação entre o número médio de pessoas na fila com o tempo médio que cada cliente fica na fila.

$$L_q = \lambda W_q \tag{3.13}$$

L_q	Número médio de pessoas na fila
W_q	Tempo médio de espera na fila
λ	Taxa média de chegada de pessoas à fila

3.3 Fila M/M/1

A fila M/M/1 é a mais simples existente entre os modelos propostos até hoje. Esse tipo de fila é de fácil análise e consiste em um processo markoviano de Nascimento e Morte, mas com os parâmetros $\lambda_i = \lambda$ para i=0,1,2,... e $\mu_i = \mu$ para i=1,2,3... Essa fila M/M/1 é chamada dessa maneira pelos seguintes motivos: a distribuição do tempo entre chegadas neste sistema é exponencial, o que é indicado pelo primeiro M do nome. A distribuição do tempo de serviço também é exponencial e por isso o segundo M no nome. E o último motivo é o fato de essa ser uma fila com apenas um servidor, o que é indicado pelo número "1" na terceira casa. Suponha, então, que um job chegue a este sistema. Esse vai chegar de acordo com um processo de Poisson com taxa λ e, se não houver pessoas no sistema, vai ser atendido pelo servidor único e ficará em serviço durante um tempo $\frac{1}{\mu}$ que é exponencialmente distribuído. Porém, se o suposto job chegar ao sistema e encontrar um job ou mais, esse terá que se juntar à fila e esperar por sua vez de ser atendido.

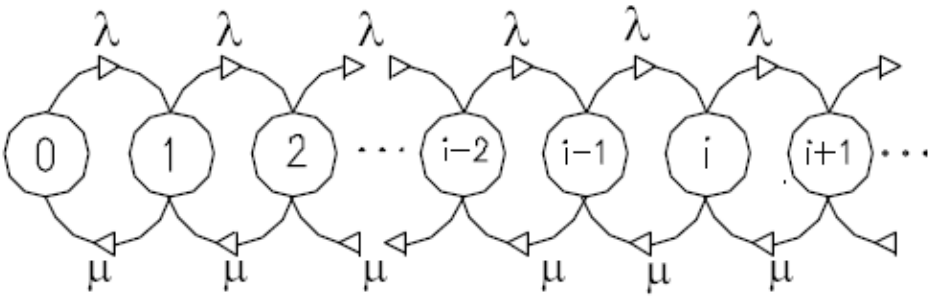


Figura 3.2: Diagrama de transição de estados para um sistema de fila M/M/1.

,

A disciplina de atendimento da fila é FiFo (First-in, First-out), o quer dizer que, se um job chegar e encontrar apenas um outro job no sistema, que estará com certeza em serviço, pois, nesse tipo de fila não há ociosidade de servidores quando há no sistema trabalho a ser processado, esse terá que iniciar uma fila e será o próximo a ser atendido e portanto, o primeiro a sair do sistema após o que já estava em serviço.

Um diagrama de transição de estados é apresentado para este sistema, Figura 3.2, onde os estados representam o número de pessoas no sistema. O sistema faz uma transição sempre que chegar um job ou sempre que um servidor termina de processar um trabalho. Utilizando-se do mesmo método empregado para achar as probabilidades p_i e p_0 no processo BD, chega-se às mesmas equações de equilíbrio 3.6 e 3.8 desenvolvidas

através dos sistemas de conservações de fluxo, mas com $\lambda_i = \lambda$ e $\mu_i = \mu$. Portanto, tem-se que:

$$p_i = p_0 \prod_{j=0}^{i-1} \frac{\lambda_j}{\mu_{j+1}} = p_0 \left(\frac{\lambda}{\mu}\right)^i, \qquad i = 1, 2, ... \tag{3.14}$$

$$p_0 = \frac{1}{1 + \sum_{i=1}^{\infty} \prod_{j=0}^{i-1} \frac{\lambda_j}{\mu_{j+1}}} = \frac{1}{1 + \sum_{i=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^i} \tag{3.15}$$

Considerando que $\frac{\lambda}{\mu} < 1$ para que o sistema seja ergódigo (Kleinrock, 1976a), o termo $\sum_{i=1}^{\infty} \left(\frac{\lambda}{\mu}\right)^i$ define a soma de uma série geométrica, a qual é igual a $\frac{\frac{\lambda}{\mu}}{1-\frac{\lambda}{\mu}}$. Logo a equação 3.15 de p_0 torna-se

$$p_0 = \frac{1}{\frac{\lambda}{\mu\left(1-\frac{\lambda}{\mu}\right)} + 1}. \tag{3.16}$$

3.4 Fila M/M/c

Nesta seção é considerado um sistema com chegadas de jobs que acontecem de acordo com o processo de Poisson e com c servidores homogêneos, todos com tempo médio de serviço exponencialmente distribuídos e com taxa $\frac{1}{\mu}$.

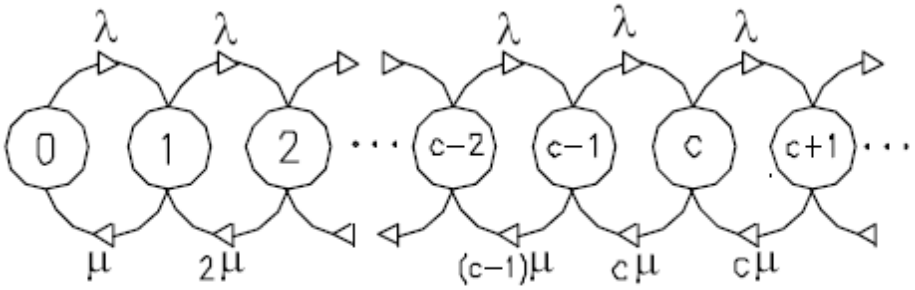


Figura 3.3: Diagrama de transição de estados para um sistema de fila M/M/c

,

Com as mesmas equações 3.6 e 3.8 são achados p_i e p_0 para a fila M/M/c. Para isso, precisa-se considerar $\lambda_i = \lambda$ para i=0,1,2,..., da mesma forma que foi feito para as

filas M/M/1. A taxa de pessoas (imaginando que são pessoas chegando) neste sistema é constante ao longo do tempo e, por isso, independentemente do estado em que esse se encontre a taxa será λ.

Contudo, a taxa de saída dessas pessoas para fora deste sistema já não é mais constante. Observe que, para o estado E_0 (nenhuma pessoa no sistema) não há nenhum trabalho sendo realizado e $\mu_0 = 0$. Para o estado E_1 (uma pessoa) há um servidor trabalhando com taxa μ, e por conseguinte, $\mu_1 = \mu$. Mas, para os estados E_i para 1≤i $<$ c teremos i pessoas no sistema, ou seja, i servidores trabalhando com taxa μ. Logo, $\mu_i = i\mu$. Quando o número de pessoas no sistema passa a ser maior ou igual a c, ou seja, o sistema se encontra no estado E_c, E_{c+1}, E_{c+2} ..., todos os servidores vão estar trabalhando e o excesso de pessoas ficará em uma fila. Assim sendo, mesmo que o sistema fique com um número enorme de pessoas a capacidade máxima de processamento vai ser o número de servidores vezes a taxa μ. Logo, $\mu_i = c\mu$ para i ≥c. A Figura 3.3 mostra o diagrama de transição de estados para a fila M/M/c.

Com a análise do diagrama de estado e das Equações 3.6 e 3.8, segue:

$$\mu_i = \begin{cases} i\mu & 0 \leq i \leq c \\ c\mu & c \leq i \end{cases} \tag{3.17}$$

Para $i \leq c$

$$p_i = p_0 \prod_{j=0}^{i-1} \frac{\lambda}{(j+1)\,\mu} = p_0 \left(\frac{\lambda}{\mu}\right)^i \frac{1}{i!} \tag{3.18}$$

Para $i \geqslant c$

$$p_i = p_0 \prod_{j=0}^{c} \frac{\lambda}{(j+1)\,\mu} \prod_{z=c+1}^{k} \frac{\lambda}{c\mu} = p_0 \left(\frac{\lambda}{\mu}\right)^i \frac{1}{c!c^{i-c}} \tag{3.19}$$

Depois de achada a probabilidade p_i pode-se, através da equação 3.7, achar p_0,

$$p_0 = \left[\sum_{i=0}^{c-1} \frac{\lambda^i}{i!\mu^i} + \left(\frac{\lambda^c}{c!\mu^c}\right)\left(\frac{c\mu}{c\mu - \lambda}\right)\right]^{-1} \tag{3.20}$$

3.5 Fator ρ de Utilização

O fator de utilização ρ é uma medida de capacidade do sistema, que é imprescindível para sua avaliação. Esse é um fator que representa a fração entre a taxa de entrada de trabalho no sistema e a capacidade total máxima com que esse consegue processar esse trabalho.

Definição 17. *O trabalho que um job trás para o sistema é igual ao número de unidades de tempo de serviço que deve ser usado para se processá-lo.*

$$\rho = \frac{\textit{taxa de chegada de trabalho por unidade de tempo}}{\textit{Capacidade máxima total de processamento por unidade de tempo}} \tag{3.21}$$

Para o caso de uma M/M/1, a utilização do sistema é a taxa média de chegada de jobs vezes o tempo médio de serviço. Neste caso a capacidade total do sistema é a capacidade do único servidor e, como a taxa de processamento é igual ao inverso do tempo de serviço, tem-se:

para uma M/M/1

$$\rho = \frac{\lambda}{\mu}, \tag{3.22}$$

e para uma M/M/c

$$\rho = \frac{\lambda}{c\mu}. \tag{3.23}$$

As duas equações acima são aplicadas para os casos em que a taxa máxima de serviço no sistema é independente do estado. A taxa de chegada de trabalho no sistema é também chamada de intensidade de tráfego. Essa medida de intensidade também é medida em

relação à utilização ρ do sistema. Para que o sistema se mantenha estável é necessário que $0 \leq \rho < 1$. Se a utilização for maior ou igual a "1" significa que no sistema chega mais trabalho por unidade de tempo do que se consegue processar e, consequentemente, a fila cresce infinitamente.

Capítulo 4

Aproximação Proposta: $M/M/c_{Heterogêneos}$

4.1 Desenvolvimento

No modelo proposto nesta dissertação, é considerado um sistema com C servidores heterogêneos e sem limite para o número de jobs na fila. Os clientes (todos de uma mesma classe) são atendidos de acordo com a chegada, ou seja, o primeiro a chegar é o primeiro a ser atendido (FCFS). Para desenvolver as equações probabilísticas que governam esse sistema, as mudanças de estado são tratadas como sendo um processo de Nascimento e Morte (BD), as chegadas de jobs como sendo governadas pelo processo de Poisson e o tempo de serviço como sendo exponencialmente distribuído.

4.1.1 Taxa de Nascimento e Taxa de Morte

Através do diagrama de estados mostrado na Figura 4.1, é possível montar as equações de equilíbrio que são fundamentadas na conservação dos fluxos, onde todo o fluxo entrante tem que ser igual ao fluxo de saída. Como é considerado que esse sistema está em um período longe no futuro, ou seja, $t \rightarrow \infty$, é usada a Equação de Kolmogorov 3.2. O objetivo é achar as equações probabilísticas 3.6 e 3.8 desenvolvidas para o processo de nascimento e morte, mas, considerando a taxa de Nascimento $\lambda_i = \lambda,\ para\ i = 0, 1, 2, ...,$ e que a taxa de Morte variável, dependente do estado em que o sistema se encontra.

Define-se então um $\mu_{equivalente}$ como sendo a taxa de Morte:

$$\mu_{eq} = \begin{cases} 0, & Para\ i = 0 \\ \mu_1, & Para\ i = 1 \\ \mu_1 + \mu_2, & Para\ i = 2 \\ . & . \\ . & . \\ . & . \\ \mu_1 + \mu_2 + \mu_3 + ... + \mu_c = \sum_{i=1}^{c} \mu_i = m_c, & Para\ i \geqslant c \end{cases} \tag{4.1}$$

Observa-se que, de acordo com a Equação 4.1 acima, μ_{eq} pode ser representado de duas maneiras. A primeira representação é feita para o caso em que o sistema contém *menos de c jobs* e a segunda, para o caso desse conter *c ou mais jobs*. Abaixo é definido m_i e m_c com o intuito de facilitar o desenvolvimento da formulação e para melhorar a visualização da formulação.

Para i $<$ c:

$$m_i = \sum_{j=1}^{i} \mu_j \tag{4.2}$$

Para i $\geqslant$ c:

$$m_i = m_c = \sum_{j=1}^{c} \mu_j \tag{4.3}$$

Neste ponto, é necessário definir nas equações 4.1, 4.2 e 4.3, quem são as taxas μ_1, μ_2, μ_3, ... μ_i. Entretanto, essa é uma escolha que não é óbvia e nem direta, e sua escolha influencia muito o sistema. Quando se faz referencia a μ_1, está sendo na verdade feita uma referencia à taxa de Morte do sistema quando esse se encontra no estado E_1, o que acontece quando há apenas um job no sistema. Quando isso acontece, tem-se apenas um servidor processando trabalho e, obviamente, a taxa de morte nesse caso será a taxa de processamento desse servidor que ainda está ocupado (supondo que os outros já acabaram de processar o que estavam fazendo).

A dificuldade ao definir μ_1 é intuitiva, pois, neste ponto, a seguinte pergunta surge: qual é o servidor que está ocupado nesse momento? A pergunta se repete quando se tem dois, três, quatro ou mais servidores ocupados: quais são os servidores que estão ocupados

em um tempo t qualquer se há i (para i < c) jobs no sistema?

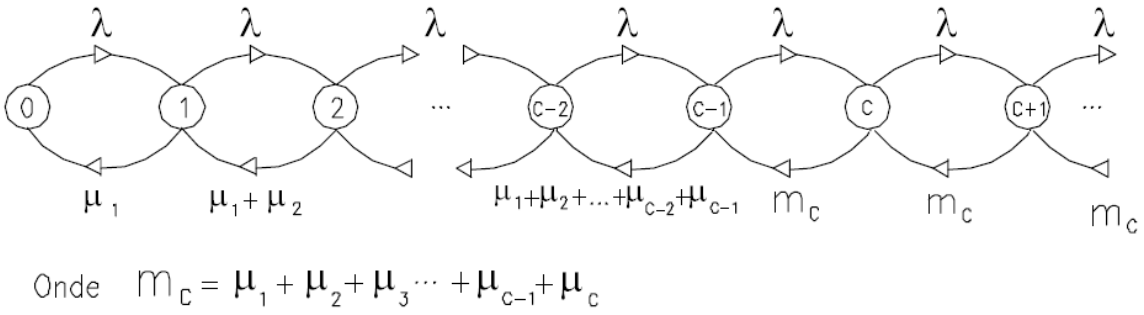


Figura 4.1: Diagrama de transição de estados para o modelo proposto de Servidores Heterogêneos.

É difícil responder às perguntas acima de forma analítica. As possibilidades aumentam com o aumento do número de servidores e as probabilidades ficam mais difíceis de serem encontradas, pois passam a ser influenciadas também pelas políticas de alocação usadas no sistema específico. Para o caso de um sistema com 4 servidores (c=4), tem-se: $\binom{4}{1} = 4$ taxas diferentes de serviço possíveis para o estado E_1, $\binom{4}{2} = 6$ taxas diferentes de serviço possíveis para o estado E_2 e $\binom{4}{3} = 4$ taxas diferentes de serviço possíveis para o estado E_3. Portanto, para um sistema com quatro servidores, o número de possibilidades aumenta e com isso a formulação passa a ter uma natureza probabilística difícil de ser modelada analiticamente de forma fechada.

Para se entender melhor a influencia que a escolha dos μ's tem sobre o sistema, são desenvolvidas as equações de equilíbrio de fluxo para os estados i $\leq$ c, e, para isso, utiliza-se o diagrama apresentado na Figura 4.1.

- $\mu_1 p_1 = p_0 \lambda \quad \Leftrightarrow \quad p_1 = \frac{\lambda}{\mu_1} p_0$
- $\lambda p_0 + (\mu_{1+}\mu_2)\, p_2 = (\mu_1 + \lambda)\, p_1 \quad \Leftrightarrow \quad p_2 = \frac{\lambda^2}{\mu_1(\mu_1+\mu_2)} p_0$
- $\lambda p_1 + (\mu_1 + \mu_2 + \mu_3)\, p_3 = (\lambda + \mu_{1+}\mu_2)\, p_2 \quad \Leftrightarrow \quad p_3 = \frac{\lambda^3}{\mu_1(\mu_1+\mu_2)(\mu_1+\mu_2+\mu_3)} p_0$

.
.
.

$$\circ\ \lambda p_{c-2} + (m_c)\, p_c = (\lambda + m_{c-1})\, p_{c-1} \qquad \Leftrightarrow \qquad p_{c-1} = \frac{\lambda^{c-1}}{\mu_1(\mu_1+\mu_2)...(\mu_1+\mu_2+...+\mu_{c-1})} p_0$$

$$\circ\ \lambda p_{c-1} + (m_c)\, p_{c+1} = (\lambda + m_c)\, p_c \qquad \Leftrightarrow \qquad p_c = \frac{\lambda^{c}}{\mu_1(\mu_1+\mu_2)(\mu_1+\mu_2+\mu_3)...(m_c)} p_0$$

Observando, por exemplo, p_{c-1}, pode-se ver como μ_1 aparece em todos os termos do denominador causando grande influencia sobre a fórmula. Seguido por ele vem a taxa μ_2, que tem um segundo maior peso, depois μ_3 e assim por diante até μ_c, que só aparece no último termo do denominador de p_{c-1} mostrado acima. Conseqüentemente a escolha de cada μ_i não é obvia.

Prestando atenção ao significado da probabilidade p_i, pode-se começar a entender como essa escolha poderia ser feita. p_1 é a probabilidade de haver um job no sistema e, olhando para a sua fórmula, é visível a dependência dela pelos μ's escolhidos. É intuitivo que se houver uma pessoa no sistema e o tráfego de jobs for intenso, essa terá maior probabilidade de estar no servidor mais lento, pois, ela ficará mais tempo sendo atendida nesse do que as outras que estarão em servidores mais rápidos.

Outra razão para imaginar que a única pessoa no sistema está no servidor mais lento é a propriedade da distribuição exponencial não possui memória. Se em um momento t qualquer acontecer de todos os c servidores estarem ocupados e se eventualmente fosse necessário saber qual dos servidores tem a maior probabilidade de ser o último a terminar o trabalho, devido à falta de memória da distribuição exponencial essa probabilidade não depende da informação de qual servidor iniciou o trabalho primeiro. Essa probabilidade depende apenas da taxa de serviço μ, não importando, portanto, se algum servidor começou o trabalho antes ou não.

Levando em conta o descrito acima, pode-se fazer a seguinte escolha para a atribuição dos μ's:

$$\mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4 \leq \ ... \ \leq \mu_c$$

Esta configuração, onde μ_1 é a taxa de atendimento do servidor mais lento, μ_2 a taxa de atendimento do segundo servidor mais lento e assim por diante até μ_c, que é a taxa de atendimento do servidor mais rápido, modela o nosso sistema como se reduzíssemos todas as possibilidades de escolha para apenas uma, que é a mais provável. O que está sendo realizado na verdade é uma simplificação, pois é feita uma aproximação para esse sistema usando apenas as possibilidades de maior probabilidade e desprezando as outras.

Outra coisa é que estão sendo definidos limites superiores para o nosso modelo, pois as probabilidades usadas são as de que o sistema está sempre no pior caso, que é o servidor mais lento. Logo, é feita uma aproximação pelo pior caso, e isso pode ser considerado como um limite superior quando calculadas as medidas de performance.

4.1.2 Definição de p_i

Como definido anteriormente, μ_1 é a taxa de atendimento do servidor mais lento, μ_2 a taxa de atendimento do segundo servidor mais lento e assim por diante até μ_c que é a taxa de atendimento do servidor mais rápido.

Das equações de fluxo, chega-se à seguinte expressão para a probabilidade p_i de se achar i pessoas no sistema:

Para $i \leqslant c$ temos:

$$p_i = p_0 \frac{\lambda^i}{\prod_{k=1}^{i} (\sum_{j=1}^{k} \mu_j)}$$

ou

$$\boxed{\mathrm{p}_i = p_0 \prod_{j=1}^{i} \left(\frac{\lambda}{m_j}\right)} \tag{4.4}$$

Para $i \geqslant c$ temos:

$$p_i = p_0 \prod_{j=1}^{c} \left(\frac{\lambda}{m_j}\right) \left[\prod_{j=c+1}^{i} \left(\frac{\lambda}{m_c}\right)\right]$$

Depois de desenvolvida fica:

$$\boxed{\mathrm{p}_i = p_0 \frac{\lambda^i}{(m_c)^{i-c} \prod_{j=1}^{c} m_j}} \tag{4.5}$$

4.1.3 Definição de p_0

Para se achar p_0, coloca-se esse termo em evidencia nas equações 4.4 e 4.5. E usa-se a restrição

$$\sum_{i=0}^{\infty} p_i = 1,$$

chegando à expressão

$$p_0 = \left\{ \sum_{i=0}^{c-1} \left[\frac{\lambda^i}{\prod_{j=1}^{i} m_j} \right] + \sum_{i=c}^{\infty} \left[\frac{\lambda^i}{\left(\prod_{j=1}^{c} m_j \right) (m_c)^{i-c}} \right] \right\}^{-1} =$$

$$= \left\{ \sum_{i=0}^{c-1} \left[\frac{\lambda^i}{\prod_{j=1}^{i} m_j} \right] + \left[\frac{(m_c)^c}{\left(\prod_{j=1}^{c} m_j \right)} \sum_{i=c}^{\infty} \left(\frac{\lambda}{m_c} \right)^i \right] \right\}^{-1} \tag{4.6}$$

Para que o sistema se mantenha estável, ou seja, para que a fila não cresça indefinidamente, é necessário que a grandeza ρ definida por

$$\boxed{\rho = \frac{\lambda}{\sum_{j=1}^{c} \mu_j} = \frac{\lambda}{m_c}} \tag{4.7}$$

satisfaça a restrição $\rho < 1$.

Subistituindo ρ na equação 4.6 obtém-se a seguinte expressão para p_0:

$$p_0 = \left\{ \sum_{i=0}^{c-1} \left[\frac{\lambda^i}{\prod_{j=1}^{i} m_j} \right] + \left[\frac{(m_c)^c}{\left(\prod_{j=1}^{c} m_j \right)} \sum_{i=c}^{\infty} (\rho)^i \right] \right\}^{-1} \tag{4.8}$$

Como $|\rho| < 1$ para se garantir que o sistema seja estável, tem-se:

$$\sum_{i=c}^{\infty} (\rho)^i = (1-\rho)^{-1} (\rho^c) \tag{4.9}$$

Substituindo a equação 4.9 na equação 4.8 a formula de p_0 torna-se:

$$p_0 = \left\{ \sum_{i=0}^{c-1} \left[\frac{\lambda^i}{\prod_{j=1}^{i} m_j} \right] + \left[\frac{(m_c)^c}{\left(\prod_{j=1}^{c} m_j \right)} \frac{1}{\left(1-\left(\frac{\lambda}{m_c}\right)\right)} \frac{\lambda^c}{(m_c)^c} \right] \right\}^{-1}$$

e, finalmente

$$\boxed{p_0 = \left[\left(\sum_{i=0}^{c-1} \frac{\lambda^i}{\prod_{j=1}^{i} m_j} \right) + \frac{\lambda^c}{(1-\rho) \prod_{j=1}^{c} m_j} \right]^{-1}} \tag{4.10}$$

4.2 Medidas de Perfomance

Achadas as probabilidades p_i e p_0 é possível obter as formulações necessárias para a análise de performance de um sistema como esse. A intenção é desenvolver as equações para: i) número médio L de jobs no sistema; ii) tempo médio W que cada job fica no sistema; iii) número médio L_q de jobs na fila; e iv) tempo médio W_q que cada job fica na fila.

4.2.1 L_q e W_q

Para se achar o número médio L_q de pessoas na fila é preciso achar a esperança de se encontrar uma ou mais pessoas na fila, ou seja, mais que c pessoas no sistema.

$$L_q = \sum_{i=c}^{\infty} (i-c) p_i. \tag{4.11}$$

Substituindo a equação 4.5 na equação acima, tem-se:

$$L_q = \sum_{i=c}^{\infty} (i-c) p_i = \sum_{i=c}^{\infty} (i-c) p_0 \frac{\lambda^i}{\left(\prod_{j=1}^{c} m_j \right) (m_c)^{i-c}} =$$

$$= p_0 \frac{(m_c)^c \rho^c}{\prod_{j=1}^{c} m_j} \sum_{i=c}^{\infty} (i-c) (\rho)^{i-c}$$

Para $i = k + c$

$$L_q = p_0 \frac{(m_c)^c \rho^c}{\prod_{j=1}^{c} m_j} \sum_{k=0}^{\infty} (k) (\rho)^k =$$

$$= p_0 \frac{(m_c)^c \rho^{c+1}}{\prod_{j=1}^{c} m_j} \sum_{k=0}^{\infty} (k) (\rho)^{k-1} =$$

$$= p_0 \frac{(m_c)^c \rho^{c+1}}{\prod_{j=1}^{c} m_j} \frac{d}{d\rho} \left[\sum_{k=0}^{\infty} (\rho)^k \right] =$$

$$= p_0 \frac{(m_c)^c \rho^{c+1}}{\prod_{j=1}^{c} m_j} \frac{d}{d\rho} \left[\frac{\rho}{1-\rho} \right] = p_0 \frac{(m_c)^c \rho^{c+1}}{\left(\prod_{j=1}^{c} m_j \right)} (\rho - 1)^{-2}$$

tem-se então

$$\boxed{\mathrm{L}_q = p_0 \frac{(m_c)^c \rho^{c+1}}{\left(\prod_{j=1}^{c} m_j \right) (\rho - 1)^2}}. \tag{4.12}$$

Através da Lei de Little Equação 3.13

$$W_q = \frac{L_q}{\lambda}$$

Obtem-se

$$\boxed{\mathrm{W}_q = p_0 \frac{(m_c)^c \rho^{c+1}}{\left(\prod_{j=1}^{c} m_j \right) (\rho - 1)^2} \frac{1}{\lambda}}. \tag{4.13}$$

4.2.2 L e W

Para se achar o número médio L de pessoas no sistema é preciso achar a esperança de se encontrar uma ou mais pessoas no sistema:

$$L = \sum_{i=0}^{\infty} i p_i. \tag{4.14}$$

$$L = \sum_{i=0}^{\infty} i\ p_i = \sum_{i=0}^{c-1} i \left[p_0 \prod_{j=1}^{i} \left(\frac{\lambda}{m_j} \right) \right] + \sum_{i=c}^{\infty} i \left[p_0 \frac{\lambda^i}{(m_c)^{i-c} \prod_{j=1}^{c} m_j} \right]$$

$$= p_0 \sum_{i=0}^{c-1} i \left[p_0 \prod_{j=1}^{i} \left(\frac{\lambda}{m_j} \right) \right] + \left[p_0 \frac{(m_c)^c \rho}{\left(\prod_{j=1}^{c} m_j \right)} \sum_{i=c}^{\infty} i \left[\rho^{i-1} \right] \right] =$$

$$L = p_0 \sum_{i=0}^{c-1} i \left[p_0 \prod_{j=1}^{i} \left(\frac{\lambda}{m_j} \right) \right] + \frac{p_0 (m_c)^c \rho}{\left(\prod_{j=1}^{c} m_j \right)} \frac{d}{d\rho} \sum_{i=c}^{\infty} \rho^i$$

Substituindo na equação acima a Equação 4.9, tem-se:

$$L = p_0 \sum_{i=0}^{c-1} i \left[\prod_{j=1}^{i} \left(\frac{\lambda}{m_j} \right) \right] + \frac{p_0 (m_c)^c \rho}{\left(\prod_{j=1}^{c} m_j \right)} \frac{d}{d\rho} \left((1-\rho)^{-1} (\rho^c) \right) =$$

$$\boxed{\mathrm{L}{=}\mathrm{p}_0 \left[\sum_{i=1}^{c-1} i \prod_{j=1}^{i} \left(\frac{\lambda}{m_j} \right) + \frac{(m_c)^c}{\left(\prod_{j=1}^{c} m_j \right)} \frac{(c + \rho - c\rho)(\rho^c)}{(\rho - 1)^2} \right]} \tag{4.15}$$

Finalmente, para se obter o tempo médio W que cada job fica no sistema é usada novamente a Lei de Little, o que resulta em:

$$\boxed{\mathrm{W}= \frac{p_0}{\lambda}\left[\sum_{i=1}^{c-1} i \prod_{j=1}^{i}\left(\frac{\lambda}{m_j}\right) + \frac{(m_c)^c}{\left(\prod_{j=1}^{c} m_j\right)}\frac{(c+\rho-c\rho)\,(\rho^c)}{(\rho-1)^2}\right]}. \tag{4.16}$$

4.3 Comparação da Aproximação Proposta com Outros Modelos

Para avaliar se a aproximação é consistente, é feita uma comparação com formulações confiáveis já conhecidas. Para isso foram escolhidos os modelos de fila M/M/1 e M/M/c. O modelo dessa dissertação é uma aproximação que tenta captar o comportamento do sistema quando nesse há servidores heterogêneos. Entretanto, ele tem que ser capaz de modelar também: i) o caso em que há apenas um servidor no sistema e ii) o caso em que as taxas de serviço dos servidores são iguais, ou seja, quando os servidores são homogêneos, M/M/c.

1. M/M/1 - Servidor único: Considerando aqui a formulação desenvolvida para se obter o número médio L de pessoas no sistema. Para o caso de c=1 na Equação 4.15.

   Para $\mu_1 = \mu$

   Para i $<$ c

   $$m_i = \sum_{j=1}^{1} \mu_j = \mu$$

   Para i $\geqslant$ c

   $$m_c = \sum_{j=1}^{1} \mu_j = \mu$$

   Com isso, chega-se à seguinte equação:

$$L = p_0 \left[\frac{\sum_{i=1}^{c-1} i \prod_{j=1}^{i} \left(\frac{\lambda}{\mu}\right) + \frac{(\mu)^c}{\left(\prod_{j=1}^{c} \mu\right)} \frac{(c+\rho-c\rho)\,(\rho^c)}{(\rho-1)^2}}{} \right]$$

$$L = p_0 \left[\frac{\sum_{i=1}^{0} i \prod_{j=1}^{i} \left(\frac{\lambda}{\mu}\right) + \frac{\mu}{\left(\prod_{j=1}^{1} \mu\right)} \frac{(1+\rho-\rho)\,(\rho)}{(\rho-1)^2}}{} \right]$$

Como na expressão acima, $\prod_{j=1}^{1} \mu = \mu$ e $\sum_{i=1}^{0} i \prod_{j=1}^{i} \left(\frac{\lambda}{\mu}\right) = 0$, tem-se:

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

$$L = p_0 \frac{\rho}{(\rho-1)^2} \tag{4.17}$$

Com a equação 3.16 desenvolvida para se obter $p_{0_{M/M/1}}$, chega-se à seguinte equação:

$$L = \frac{\rho}{(\rho-1)^2} \frac{1}{\rho\frac{1}{(1-\rho)}+1}$$

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

$$L = \frac{\rho}{\rho-1} \tag{4.18}$$

A equação 4.18 desenvolvida para se obter o número médio de jobs no sistema quando c=1, como é esperado, é a mesma que para uma M/M/1 tradicional ((Kleinrock, 1976a)).

2. M/M/c - Todos os servidores são iguais:

$$\mu_1 = \mu_2 = \mu_3 = ... = \mu_c = \mu$$

Com isso as equações 4.2 e 4.3 se reduzem a $m_i = \sum_{j=1}^{i} \mu_j = i\mu$ e $m_c = \sum_{j=1}^{c} \mu_j = c\mu$. Substituindo esses valores na Equação 4.10, tem-se:

$$p_0 = \left[\left(\sum_{i=0}^{c-1} \frac{\lambda^i}{\prod_{j=1}^{i} j\mu} \right) + \frac{\lambda^c}{(1-\rho) \prod_{j=1}^{c} j\mu} \right]^{-1}$$

Como $\prod_{j=1}^{i} j\mu = i!\mu^i$ e $\prod_{j=1}^{c} j\mu = c!\mu^c$, segue que:

$$p_0 = \left[\left(\sum_{i=0}^{c-1} \frac{\lambda^i}{i!\mu^i} \right) + \frac{\lambda^c}{c!\mu^c} \frac{1}{(1-\rho)} \right]^{-1} =$$

$$p_0 = \left[\left(\sum_{i=0}^{c-1} \frac{(c\rho)^i}{i!} \right) + \frac{(c\rho)^c}{c!} \frac{1}{(1-\rho)} \right]^{-1} \tag{4.19}$$

Que é a mesma equação obtida para uma M/M/c tradicional apresentada na equação 3.20 (Pode ser encontrado desenvolvimento para essa em (Kleinrock, 1976a)).

É importante deixar claro que, o fato de o modelo deduzido poder ser aplicado com sucesso aos casos M/M/1 e $M/M/c_{homogneo}$ corrobora para, mas não garante que o modelo $M/M/c_{heterogneo}$ aqui deduzido represente corretamente processos reais.

4.4 Sumário

Neste capítulo foi desenvolvida uma formulação para se definir limites superiores para as medidas de performance do modelo de filas com servidores heterogêneos proposto neste trabalho. Observa-se que tal formulação serve, também, para aproximar o valor dessas medidas, independentemente, do tipo de alocação utilizada. Ainda, foi mostrado que quando a aproximação desenvolvida é utilizada para modelar sistemas com apenas 1 servidor ou com c servidores homogêneos, essa se reduz respectivamente a uma M/M/1 ou a uma M/M/c. No capítulo 5, mostramos com resultados númericos o comportamento da aproximação desenvolvida.

Capítulo 5

Resultados Numéricos, Capacidades e Limitações

5.1 Resultados Numéricos

Nesta seção são apresentados resultados numéricos que servem como fonte para a investigação da validade e da utilidade da formulação proposta quando empregada ao modelo desejado (veja Seção 1.5) de servidores heterogêneos. Já que não há nenhum outro modelo existente na literatura (até onde vai o conhecimento do autor) que possa ser usado para se obter medidas que sirvam como comparação para os mesmos casos estudados aqui, simulações (veja Apendice A) são usadas para que se tenha os dados necessários para avaliar os resultados obtidos com a formulação desenvolvida.

Os esforços foram concentrados na medida de performance dado pelo tempo médio de espera em fila Wq, pois ela representa o mesmo comportamento probabilístico que as outras medidas de performance. Dessa maneira, é esperado que as avaliações dos resultados para tal medida sirvam para entender o comportamento do sistema em relação às outras também.

Vários sistemas de filas foram criados, onde variou-se:

1. o número c de servidores existentes;
2. a divisão da capacidade total de processamento entre esses servidores;
3. a taxa λ de chegada de jobs.

Para cada um desses sistemas foram calculados os valores para o tempo médio de espera W_q através da formulação criada e através da formulação existente, M/M/c, que considera os servidores como sendo homogêneos. Em seguida simulamos para cada um desses sistemas os casos de alocação rápida, aleatória e lenta. Portanto, para cada sistema, foram encontrados três valores simulados para W_q. O referente ao caso em que foi alocado sempre o servidor livre mais rápido primeiro, o referente ao caso em que alocou-se sempre o servidor livre mais lento primeiro e o correspondente ao caso em que os servidores foram alocados de forma aleatória.

As simulações foram feitas na linguagem GPSS (General Purpose Simulation System) e o "compilador" usado foi o GPSS World Program. O GPSS World foi escolhido, pois permite obter rapidamente as respostas, permite a visualização da simulação enquanto essa está sendo executada e também fornece ferramentas para o tratamento estatístico dos dados. Para cada um dos sistemas[1] foram feitas duzentas replicações que serviram para definir um intervalo mínimo de 95% de confiança. A partir das simulações, foram extraídas a média dos tempos médios de espera na fila, ou seja, obteve-se um $\bar{W}_q$ simulado para cada alocação de cada sistema.

Para classificar os sistemas foi necessário empregar o índice de Gini. Esse índice varia entre 0 e 1, onde 0 é o caso homogêneo e 1 é o caso mais heterogêneos. Neste trabalho o índice de Gini serviu para medir a diferença entre as capacidades de processamento que existiam entre os servidores de um determinado sistema. Essas medidas, na verdade, indicam a heterogeneidade existente entre os servidores para que se possa classificar e comparar os sistemas em questão.

Com base na descrição acima, pretende-se avaliar a influência da heterogeneidade dos servidores na formulação proposta, determinar como o sistema se comportava quando o número de servidores é variado, considerar o controle que o coeficiente ρ tem sobre o sistema e definir o erro resultante que ocorre na formulação proposta quando são mudadas as políticas de alocação de servidores.

Os gráficos a seguir contêm os valores de Wq simulados e calculados para o caso de dois servidores heterogêneos. Primeiramente, simulou-se e calculou-se W_q para o sistema que tem a seguinte distribuição das capacidades de processamento, $\mu_1 = 0.02$ e $\mu_2 = 0.98$. Esse sistema é representado pelo ponto 1 de cada curva nas Figuras 5.1, 5.2 e 5.3, as quais representam os mesmos sistemas só que com taxas λ de chegada diferentes. Em seguida,

[1] O número total de casos é multiplicado por três, já que para cada sistema foram considerados três tipos de alocação diferente.

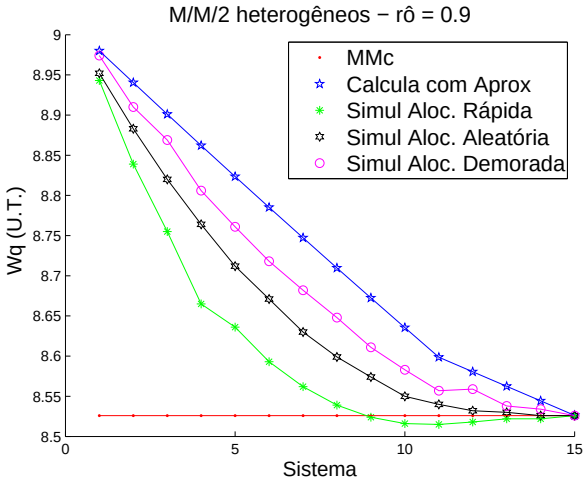
nessas mesmas figuras, são plotados os valores de W_q obtidos para outras distribuições de μ's. Essas foram sendo mudadas gradualmente de forma a diminuir a heterogeneidade entre os servidores, até o caso em que os servidores fossem iguais.

Nas Figuras 5.1 até 5.12 é mostrado o erro existente entre os valores obtidos através da aproximação criada e através da M/M/c em relação aos valores simulados para cada uma das alocações. Esse erro foi plotado em função do índice de Gini para que se possa observar como esse se comporta quando é variada a heterogeneidade dos sistemas. Usou-se a seguinte formula para o calculo do erro:

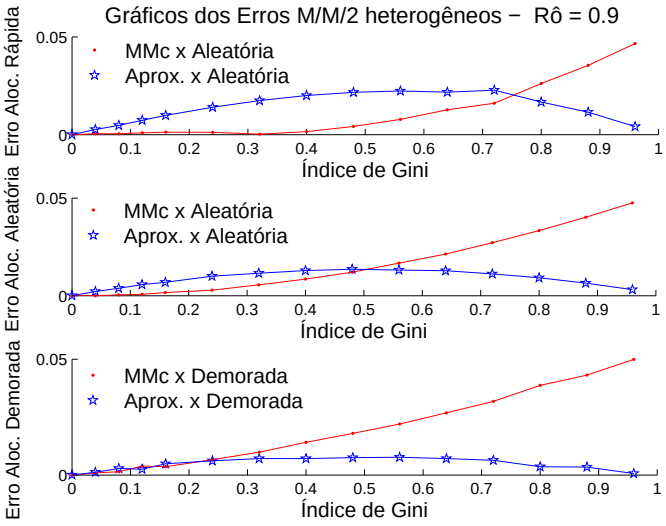
$$Erro = \frac{|| \ Simulado \ - \ Calculado \ ||}{Simulado} \tag{5.1}$$

A equação de erro 5.1 acima, nada mais é do que a normalização do erro em relação aos valores simulados. Essa normalização é necessária para que se possa comparar o erro para os diferentes sistemas e para os casos em que as taxas de chegadas variam.

Apenas com a análise das Figuras 5.1, 5.2 e 5.3, as quais foram feitas para sistemas que contêm apenas dois servidores heterogêneos já é possível observar que o erro normalizado obtido tanto pela formulação criada como pela M/M/c tradicional é inversamente proporcional a ρ. Através dessas figuras, pode-se ver que para $\rho = 0.9$ o erro máximo obtido foi aproximadamente 0.049 para a curva da M/M/c e de 0.022 para a formulação criada. Já para $\rho = 0.75$ e $\rho = 0.6$ foram respectivamente 0.1349 e 0.2363 para a curva da M/M/c e 0.0691 e 0.1408 para a curva obtida com a formulação criada. O erro aumenta muito para os dois casos calculados quando diminui-se o ρ do sistema. Isso se deve ao fato de que quando o sistema tem utilização mais baixa (ρ menor), a variação dos estados de probabilidade se torna maior. Em outras palavras, na fórmula de p_i, equações 4.5 e 4.4, o denominador gera um erro maior devido à diminuição do ρ, pois a probabilidade do servidor ocupado ser o de taxa de processamento mais lenta, como proposto na seção 4.1, diminui. E, para a M/M/c, o erro aumenta, pois a diminuição de ρ faz com que o peso da heterogeneidade dos servidores aumente, distanciando-se, portanto, dos resultados obtidos com a consideração da homogeneidade. Esse fenômeno acontece também para mais servidores, o que pode ser visto pelas Figuras 5.7, 5.8 e 5.9 que foram feitas para modelos com 6 servidores e pelas Figuras 5.10, 5.11 e 5.12 que foram feitas para 12 servidores.



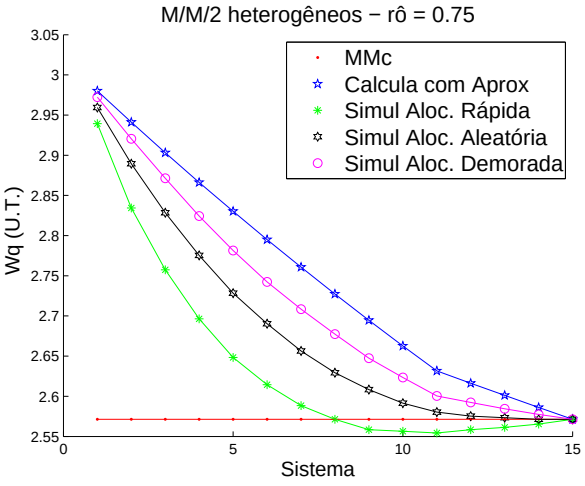
(a)



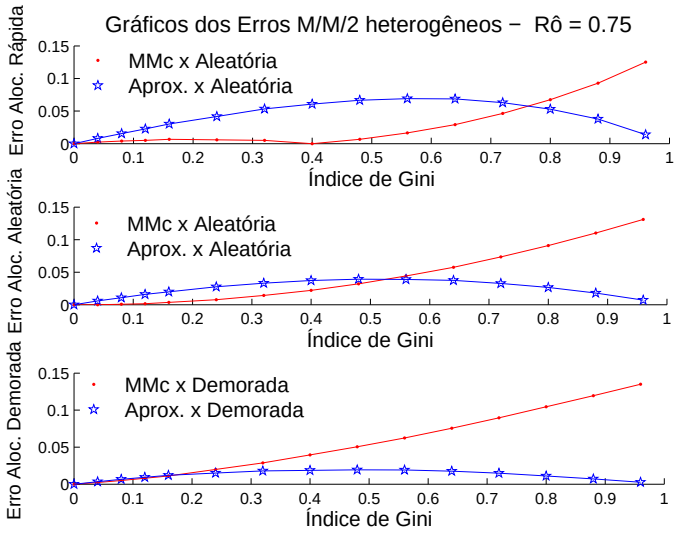
(b)

Figura 5.1: Sistemas com 2 servidores para $\rho = 0.9$ (a) Tempo médio de espera na fila (W_q) para 15 sistemas com diferentes μ's. Os pontos foram obtidos através dos calculados com MMc e com aproximação criada e através de simulações para as políticas de alocação rápida, aleatória e lenta. (b)Erro obtido com a aproximação criada e com a M/M/c em relação às simulações.

Na Tabela 5.1 estão os erros máximos obtidos para cada modelo. Através dela pode-se ver, entre outras coisas, a influência que o coeficiente de utilização ρ tem sobre os sistemas. Com ela, observa-se, também, que nos casos em que calculávamos pela formulação criada, os erros maiores ocorrem quando é considerada a alocação rápida, enquanto que, nos casos em que é usada a M/M/c, o erro é maior para a alocação lenta.



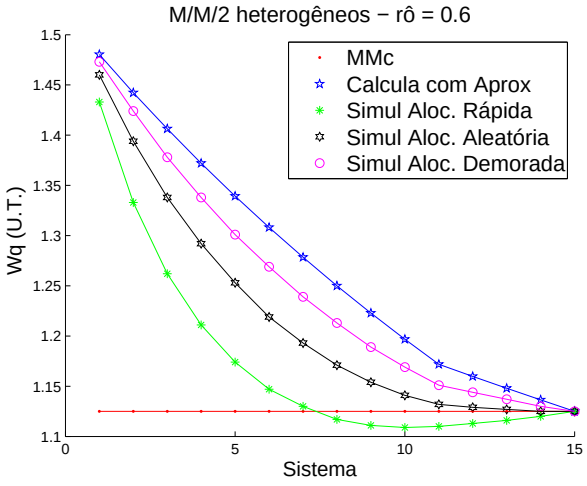
(a)



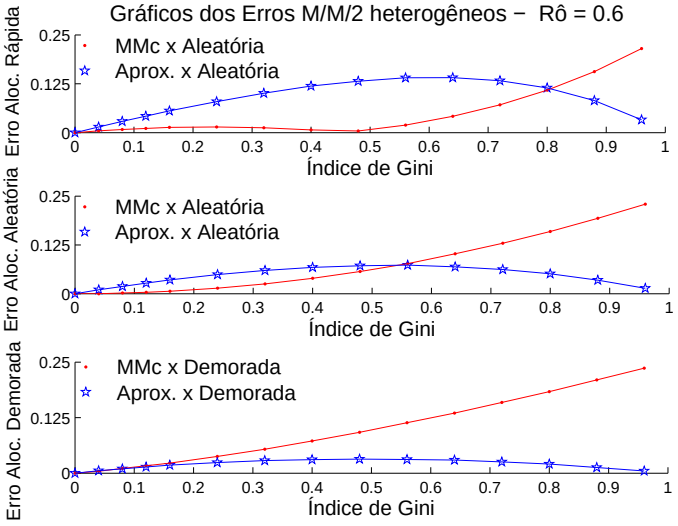
(b)

Figura 5.2: Sistemas com 2 servidores para $\rho = 0.75$ (a) Tempo médio de espera na fila (W_q) para 15 sistemas com diferentes μ's. Os pontos foram obtidos através dos calculados com MMc e com aproximação criada e através de simulações para as políticas de alocação rápida, aleatória e lenta. (b)Erro obtido com a aproximação criada e com a M/M/c em relação às simulações.

Pelas figuras, pode-se notar também que o erro em geral aumenta quando aumenta-se o número de servidores. Ao considerar-se, por exemplo, a Tabela 5.1 de erro máximo pode-se ver que esse aumento no erro de forma generalizada realmente ocorre. Isso se deve ao fato de que, ao aumentar-se o número de servidores, aumenta, também, o número de possibilidades na qual o job em serviço poderá se encontrar.



(a)



(b)

Figura 5.3: Sistemas com 2 servidores para $\rho = 0.6$ (a) Tempo médio de espera na fila (W_q) para 15 sistemas com diferentes μ's. Os pontos foram obtidos através dos calculados com MMc e com aproximação criada e através de simulações para as políticas de alocação rápida, aleatória e lenta. (b)Erro obtido com a aproximação criada e com a M/M/c em relação às simulações.

Com as Figuras 5.1(b) a 5.12 (b), pode-se observar a influência da heterogeneidade dos servidores no erro resultante. Nessas figuras estão plotados os erros obtidos em cada sistema em função do índice de Gini. Observa-se que a formulação desenvolvida se comporta melhor para os casos em que a heterogeneidade é maior, enquanto que a M/M/c comporta-se melhor para os casos em que a heterogeneidade é menor. Entretanto, é difícil

definir exatamente para qual faixa do índice de Gini a formulação criada se comporta de melhor maneira. Essa dificuldade pode ser ilustrada, por exemplo, com as Figuras 5.3(b) e 5.9(b). Na primeira, para $\rho = 0.6$ e alocação aleatória, a partir de um IG=0.58, a formulação criada apresenta um erro menor do que o apresentado pela M/M/c e, já na segunda figura, o comportamento da primeira passa a ser melhor a partir de um IG=0.15. Com isso, observa-se que com o aumento do número de servidores o erro obtido com a formulação criada passa a ser menor que o erro obtido com a M/M/c para um IG mais baixo. Em outras palavras, pode-se dizer que, quanto maior o número de servidores, melhor a formulação criada fica em relação à M/M/c.

Ainda na Tabela 5.1 pode-se ver que, para a formulação desenvolvida, os erros de maior valor ocorrem para os casos de alocação rápida. Um resultado importante obtido aqui é que, para todos os casos, o erro máximo encontrado para a formulação criada é menor que os erros máximos encontrados com a formulação da M/M/c. Isso é um resultado importante, pois escolhendo usar a formulação desenvolvida nesta dissertação para aproximar algum sistema de servidores heterogêneos, diminui-se a intensidade do erro máximo que tal aproximação pode acarretar.

A Tabela 5.2 mostra a soma dos erros obtidos para cada um dos sistemas, dividida pelo número de amostras (sistemas) feitas dentro de cada modelo, ou seja, a média dos erros. O erro médio é referente a cada uma das curvas apresentadas nas Figuras 5.1(b) até 5.12(b), e consiste na soma dos pontos obtidos em cada uma dividido pelo número de pontos. Através dessa, pode-se avaliar e comparar para os vários casos o erro médio obtido entre as curvas conseguidas com a fórmula para M/M/c com as curvas conseguidas com a fórmula proposta.

Com a Tabela 5.2 pode-se mais uma vez observar a influência das alocações nos resultados. Para a formulação aqui proposta, o erro médio é menor do que o achado com a formulação da M/M/c quando a política de alocação é a lenta, ou a aleatória. Já para o caso da alocação rápida, não se sabe a que apresenta o menor erro entre as duas formulações, pois esse depende do ρ utilizado e do número c de servidores empregados.

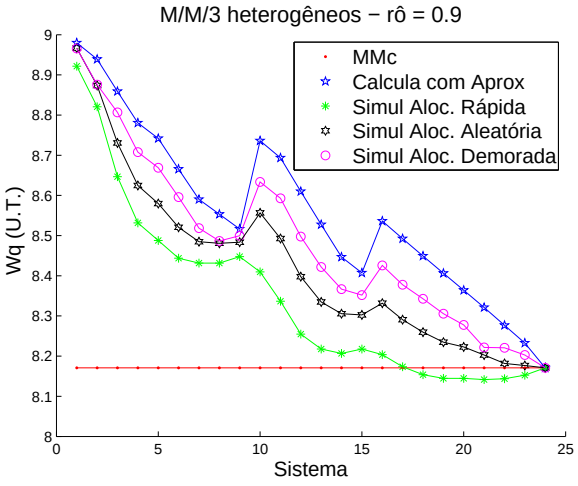
Na Tabela 5.3 observa-se o efeito que o aumento do número de servidores no sistema tem sobre o tempo médio W_q de espera em fila. Para uma mesma taxa total de processamento, foi calculado o tempo de espera para vários ρ's e para vários valores de c. Chega-se à conclusão de que, quando é aumentado o número de servidores, W_q diminui. Isso acontece porque a probabilidade de fila diminui, ou seja, menos jobs encontraram fila ao chegar ao sistema e, portanto, menos jobs esperam, fazendo com que o número médio

Tabela 5.1: Erro máximo encontrado nos resultados. Essa tabela compara os erros de maior intensidade gerados pela M/M/c com os gerados pela Aproximação criada.

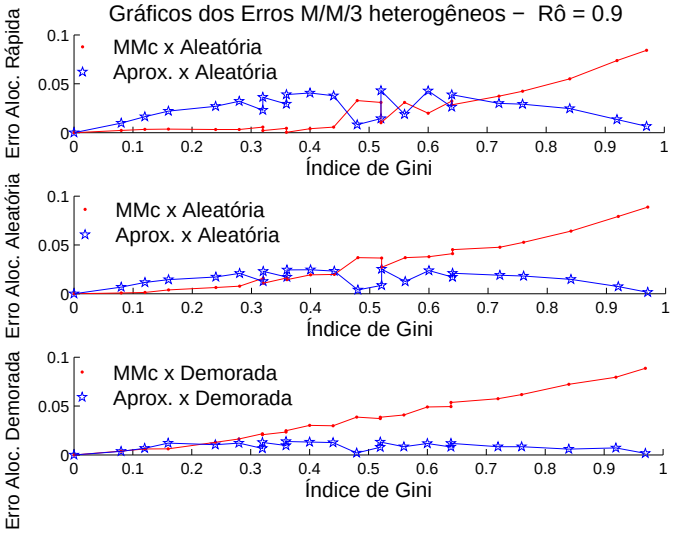
Erro Máximo Encontrado				
Para 2 servidores		**Aloc. Rápida**	**Aloc. Aleatória**	**Aloc. Demorada**
ρ = 0.9	Aproximação criada	0.0227	0.0136	0.0077
	M/M/c	0.0466	0.0476	0.0499
ρ = 0.75	Aproximação criada	0.0691	0.0393	0.0193
	M/M/c	0.1252	0.1311	0.1349
ρ = 0.6	Aproximação criada	0.1408	0.0731	0.0318
	M/M/c	0.2149	0.2295	0.2363
Para 3 servidores		**Aloc. Rápida**	**Aloc. Aleatória**	**Aloc. Demorada**
ρ = 0.9	Aproximação criada	0.043	0.0252	0.0137
	M/M/c	0.0842	0.0888	0.0887
ρ = 0.75	Aproximação criada	0.137	0.074	0.0351
	M/M/c	0.2262	0.2333	0.2359
ρ = 0.6	Aproximação criada	0.2956	0.1464	0.0588
	M/M/c	0.3797	0.3921	0.3979
Para 6 servidores		**Aloc. Rápida**	**Aloc. Aleatória**	**Aloc. Demorada**
ρ = 0.9	Aproximação criada	0.0689	0.047	0.0245
	M/M/c	0.1637	0.1657	0.1716
ρ = 0.75	Aproximação criada	0.2358	0.1182	0.0543
	M/M/c	0.4041	0.4181	0.4236
ρ = 0.6	Aproximação criada	0.6275	0.2264	0.0864
	M/M/c	0.5797	0.6493	0.6566
Para 12 servidores		**Aloc. Rápida**	**Aloc. Aleatória**	**Aloc. Demorada**
ρ = 0.9	Aproximação criada	0.0863	0.0435	0.0222
	M/M/c	0.2748	0.2796	0.2828
ρ = 0.75	Aproximação criada	0.3177	0.1369	0.0585
	M/M/c	0.6209	0.6317	0.6353
ρ = 0.6	Aproximação criada	0.9933	0.2706	0.0946
	M/M/c	0.8553	0.8654	0.8687

de jobs na fila diminua e, conseqüentemente, fazendo com que o tempo médio de espera diminua.

Entretanto, para servidores heterogêneos, um outro fenômeno acontece. Quando servidores heterogêneos são considerados, dependendo do valor do índice de Gini do sistema, o W_q real se comporta como se o sistema tivesse menos servidores, ou mais, dependendo também da alocação escolhida. Observa-se que um sistema com $\rho = 0.95$ e dois servidores, c=2, o valor de W_q é 18.513. Entretanto, para um índice de Gini próximo da unidade, esse sistema com dois servidores se comporta mais como um sistema de um servidor. Ao aumentar o número de servidores, como pode ser visto pela Tabela 5.3, o tempo de espera em fila diminui. Para o exemplo com $\rho = 0.95$, ao considerar-se c=20, tem-se que $W_q = 15.108$. Ao comparar-se o valor obtido com c=1, com o obtido com c=20, observa-se uma variação resultante de aproximadamente 20%. Isso justifica o porquê dos erros nas Figuras 5.1(b) até 5.12(b) aumentarem com o aumento de c. Isso justifica, também, o



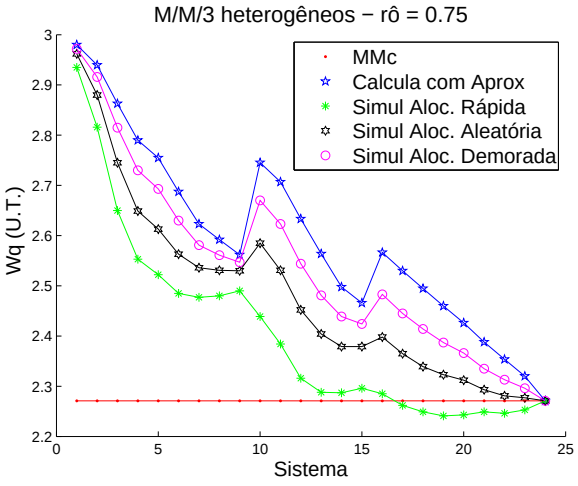
(a)



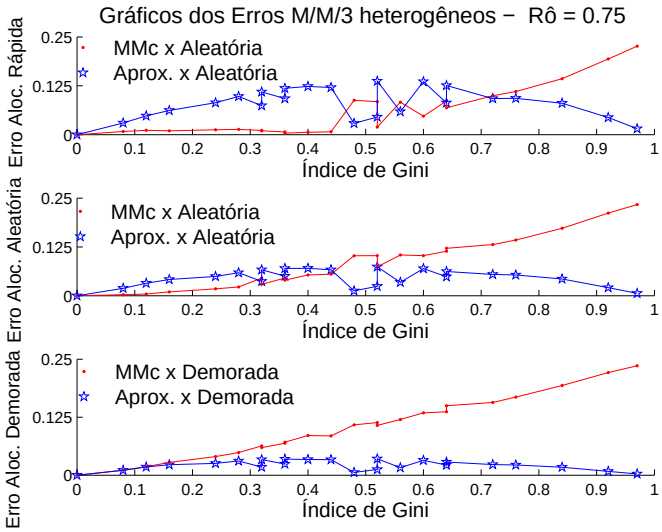
(b)

Figura 5.4: Sistemas com 3 servidores para $\rho = 0.9$ (a) Tempo médio de espera na fila (W_q) para 24 sistemas com diferentes μ's. Os pontos foram obtidos através dos calculados com MMc e com aproximação criada e através de simulações para as políticas de alocação rápida, aleatória e lenta. (b)Erro obtido com a aproximação criada e com a M/M/c em relação às simulações.

fato de que a formulação criada nessa dissertação se torna cada vez melhor, ao longo de todo o intervalo do índice de Gini, do que a formulação da M/M/c, quando aumenta-se c, pois a primeira representa melhor o sistema quando o número de servidores é grande.



(a)

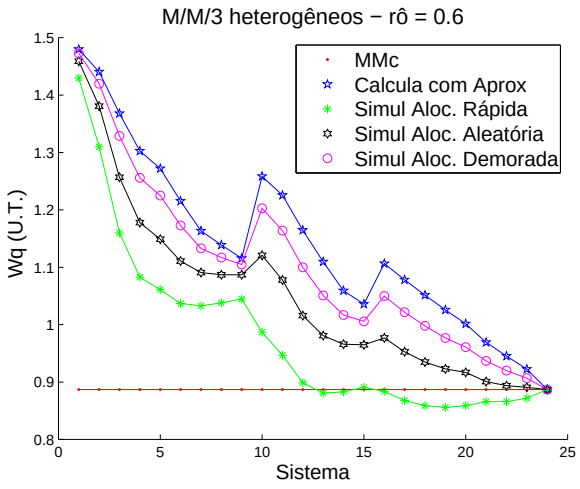


(b)

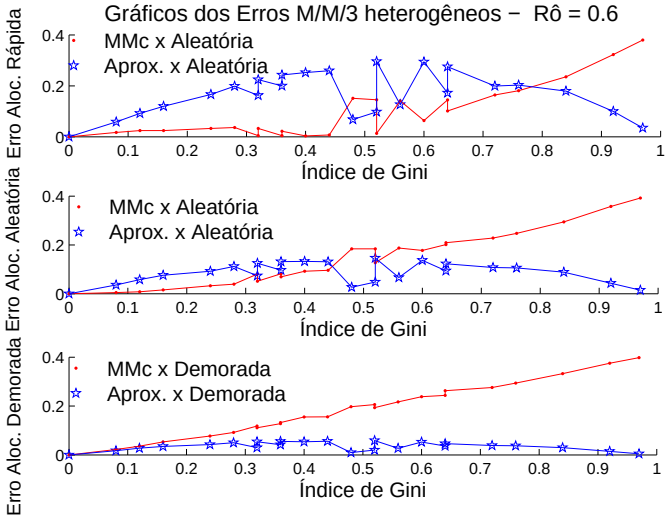
Figura 5.5: Sistemas com 3 servidores para $\rho = 0.75$ (a) Tempo médio de espera na fila (W_q) para 24 sistemas com diferentes μ's. Os pontos foram obtidos através dos calculados com MMc e com aproximação criada e através de simulações para as políticas de alocação rápida, aleatória e lenta. (b)Erro obtido com a aproximação criada e com a M/M/c em relação às simulações.

5.2 Capacidades e Limitações

A formulação desenvolvida nesta dissertação fornece um limite superior para as medidas de desempenho tempo médio de espera na fila e no sistema e pelo número médio de jobs na fila e no sistema. Com esse limite, pode-se estimar, para vários casos, o valor



(a)



(b)

Figura 5.6: Sistemas com 3 servidores para $\rho = 0.6$ (a) Tempo médio de espera na fila (W_q) para 24 sistemas com diferentes μ's. Os pontos foram obtidos através dos calculados com MMc e com aproximação criada e através de simulações para as políticas de alocação rápida, aleatória e lenta. (b)Erro obtido com a aproximação criada e com a M/M/c em relação às simulações.

máximo para essas medidas. Com isso é possível gerir melhor vários sistemas de filas que dependam de servidores heterogêneos.

Uma vantagem para o uso das fórmulas desenvolvidas aqui é que estas incorrem em um erro máximo menor do que o obtido pela formulação que considera os servidores homogêneos. Outra vantagem é que as fórmulas desenvolvidas aqui são mais simples de

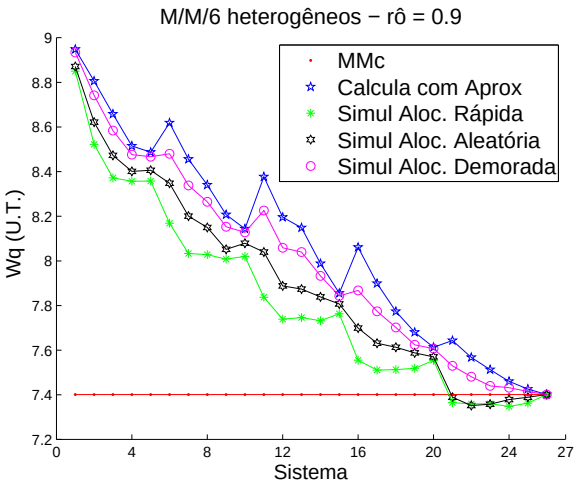
Tabela 5.2: Erro Médio encontrado nos resultados para cada modelo. Essa tabela mostra a soma dos erros ocorridos dividido pelo número de amostras feitas com a M/M/c e com a Aproximação criada.

Erro Médio				
Para 2 servidores		**Aloc. Rápida**	**Aloc. Aleatória**	**Aloc. Demorada**
ρ = 0.9	Aproximação criada	0.0131	0.0082	0.0046
	M/M/c	0.0103	0.0145	0.018
ρ = 0.75	Aproximação criada	0.0402	0.0233	0.0116
	M/M/c	0.0275	0.0393	0.0501
ρ = 0.6	Aproximação criada	0.081	0.0428	0.0191
	M/M/c	0.0458	0.0691	0.0899
Para 3 servidores		**Aloc. Rápida**	**Aloc. Aleatória**	**Aloc. Demorada**
ρ = 0.9	Aproximação criada	0.0254	0.0153	0.0087
	M/M/c	0.0215	0.0296	0.0359
ρ = 0.75	Aproximação criada	0.0789	0.0443	0.0211
	M/M/c	0.0562	0.0807	0.1009
ρ = 0.6	Aproximação criada	0.1675	0.086	0.0349
	M/M/c	0.0938	0.14	0.1798
Para 6 servidores		**Aloc. Rápida**	**Aloc. Aleatória**	**Aloc. Demorada**
ρ = 0.9	Aproximação criada	0.0315	0.0212	0.0088
	M/M/c	0.0535	0.062	0.0718
ρ = 0.75	Aproximação criada	0.1074	0.0517	0.0249
	M/M/c	0.1368	0.1721	0.1924
ρ = 0.6	Aproximação criada	0.2619	0.1027	0.0422
	M/M/c	0.2172	0.2939	0.3301
Para 12 servidores		**Aloc. Rápida**	**Aloc. Aleatória**	**Aloc. Demorada**
ρ = 0.9	Aproximação criada	0.0401	0.0175	0.0084
	M/M/c	0.1343	0.1527	0.16
ρ = 0.75	Aproximação criada	0.1551	0.0575	0.024
	M/M/c	0.3326	0.3856	0.4032
ρ = 0.6	Aproximação criada	0.4991	0.1183	0.0433
	M/M/c	0.483	0.6044	0.6251

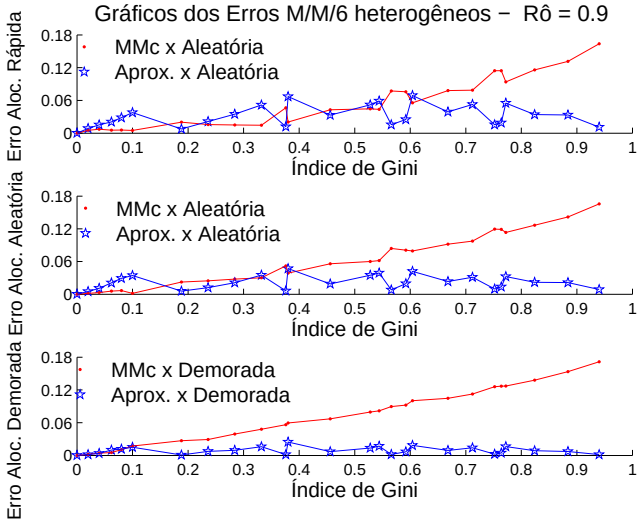
serem usadas e podem ser incorporadas facilmente em vários modelos já existentes na literatura.

A formulação se torna mais interessante ainda, quando é aumentado o número de servidores heterogêneos, pois o erro gerado com esse aumento não cresce significativamente para ela. Isso é uma grande vantagem, já que as outras aproximações existentes sofrem com um aumento excessivo do erro quando é elevado o número de servidores.

A limitação mais evidente para essa formulação é a de não conseguir captar muito bem o comportamento ótimo dos modelos de servidores heterogêneos nos quais usa-se a política de alocação que ocupa primeiro os servidores livres que trabalham com taxas de processamento maiores. Ao alocarmos os servidores mais rápidos, como pode ser visto pelos resultados obtidos com simulação, há uma região de heterogeneidade ótima para a



(a)

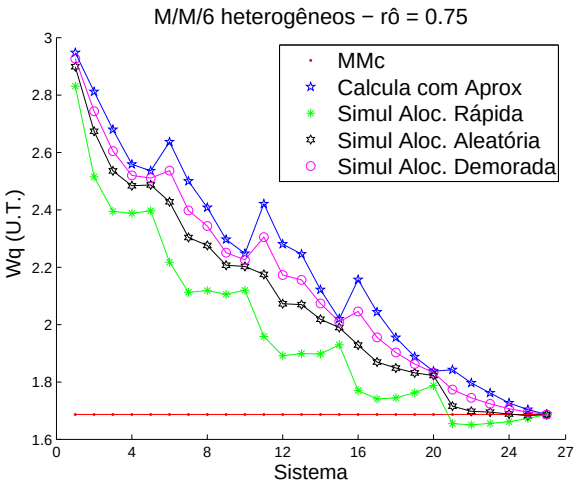


(b)

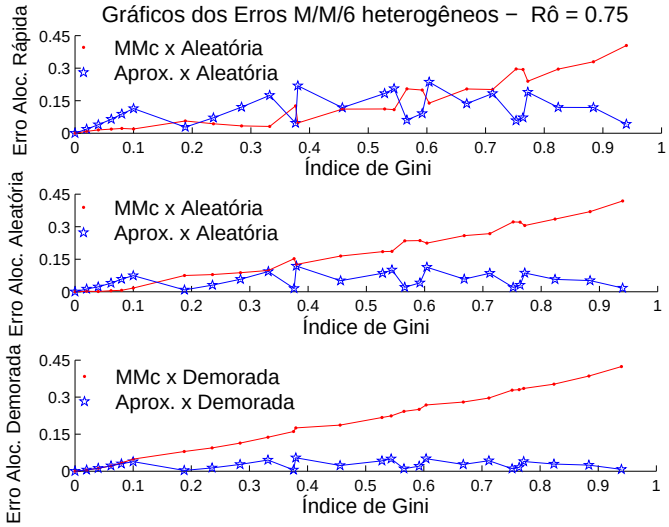
Figura 5.7: Sistemas com 6 servidores para $\rho = 0.9$ (a) Tempo médio de espera na fila (W_q) para 26 sistemas com diferentes μ's. Os pontos foram obtidos através dos calculados com MMc e com aproximação criada e através de simulações para as políticas de alocação rápida, aleatória e lenta. (b)Erro obtido com a aproximação criada e com a M/M/c em relação às simulações.

qual as medidas de performance para esses sistemas são melhores do que para o caso onde todos os servidores são iguais. Entretanto, a formulação desenvolvida aqui, por considerar só os casos mais prováveis, não consegue acompanhar essa região.

A formulação é uma aproximação, já que não fornece os resultados exatos e, por isso gera um erro. Entretanto, para o gerenciamente esse erro é menos ruim, pois é



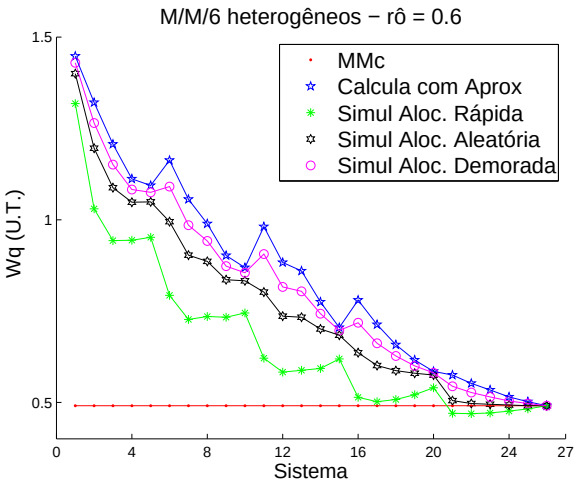
(a)



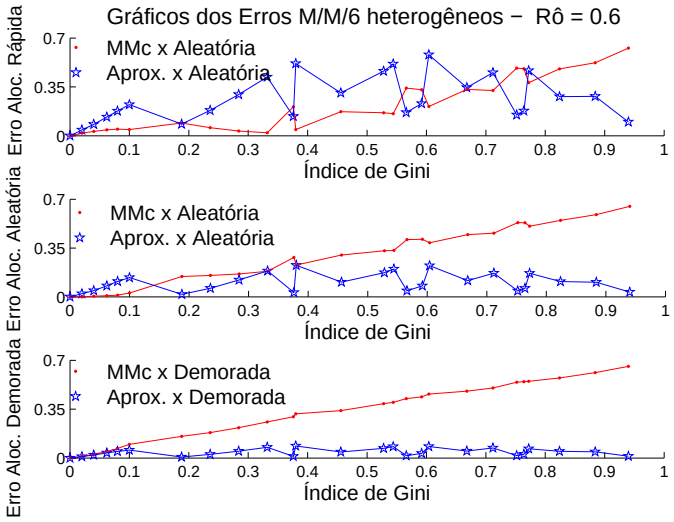
(b)

Figura 5.8: Sistemas com 6 servidores para $\rho = 0.75$ (a) Tempo médio de espera na fila (W_q) para 26 sistemas com diferentes μ's. Os pontos foram obtidos através dos calculados com MMc e com aproximação criada e através de simulações para as políticas de alocação rápida, aleatória e lenta. (b)Erro obtido com a aproximação criada e com a M/M/c em relação às simulações.

resultante do fato de que a aproximação superestima os sistemas. Isso significa que, a aproximação identifica o pior caso, que é, exatamente, o "ponto" que se quer evitar, através de gerenciamento, que o sistema atinja.

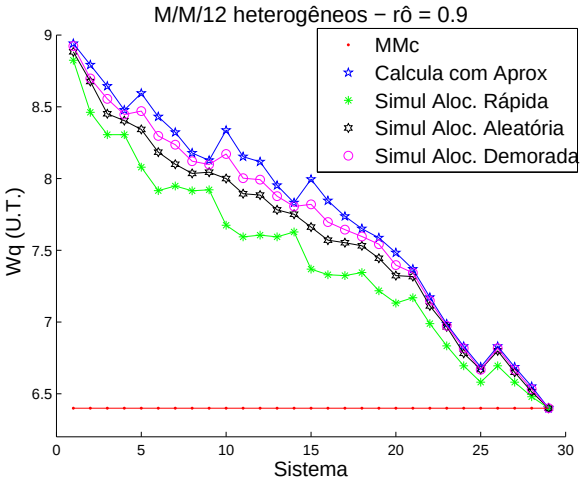


(a)

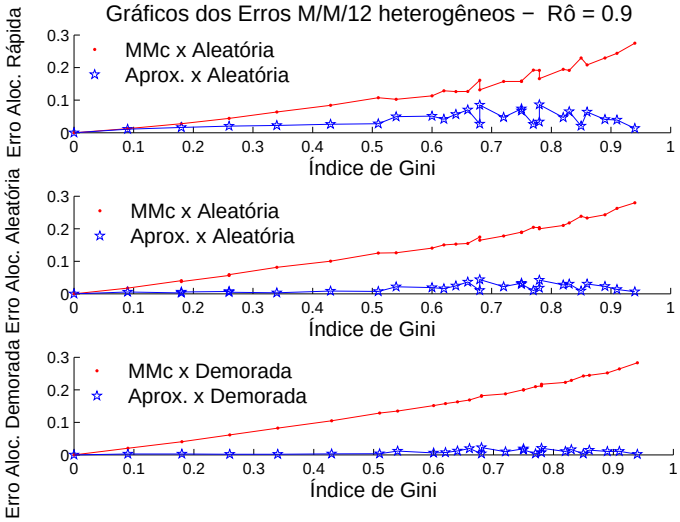


(b)

Figura 5.9: Sistemas com 6 servidores para $\rho = 0.6$ (a) Tempo médio de espera na fila (W_q) para 26 sistemas com diferentes μ's. Os pontos foram obtidos através dos calculados com MMc e com aproximação criada e através de simulações para as políticas de alocação rápida, aleatória e lenta. (b)Erro obtido com a aproximação criada e com a M/M/c em relação às simulações.

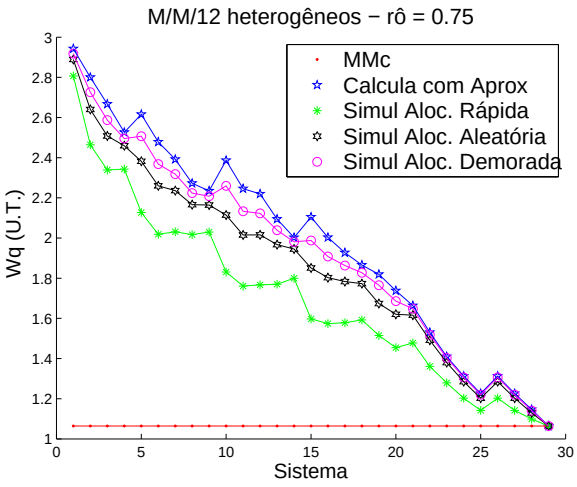


(a)

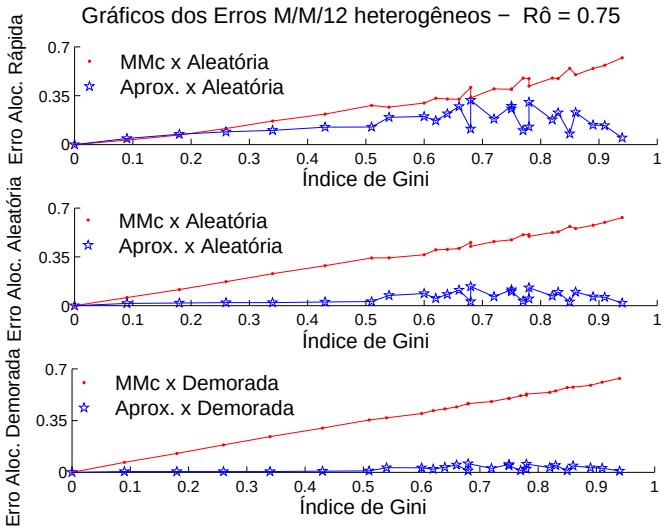


(b)

Figura 5.10: Sistemas com 12 servidores para $\rho = 0.9$ (a) Tempo médio de espera na fila (W_q) para 29 sistemas com diferentes μ's. Os pontos foram obtidos através dos calculados com MMc e com aproximação criada e através de simulações para as políticas de alocação rápida, aleatória e lenta. (b)Erro obtido com a aproximação criada e com a M/M/c em relação às simulações.

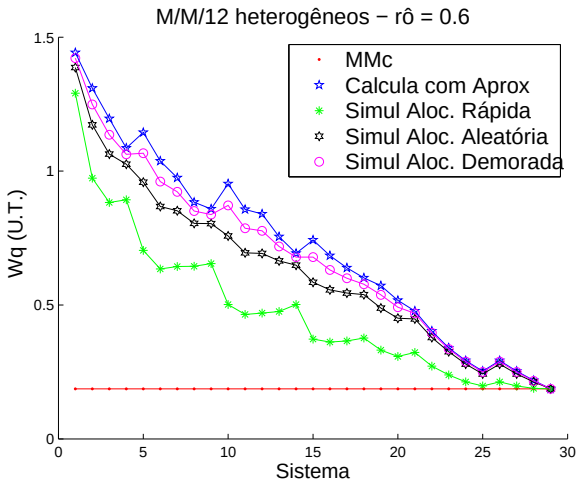


(a)

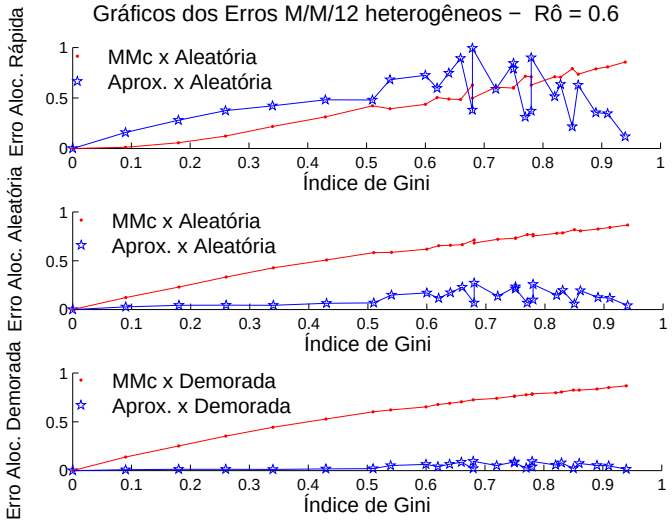


(b)

Figura 5.11: Sistemas com 12 servidores para $\rho = 0.75$ (a) Tempo médio de espera na fila (W_q) para 29 sistemas com diferentes μ's. Os pontos foram obtidos através dos calculados com MMc e com aproximação criada e através de simulações para as políticas de alocação rápida, aleatória e lenta. (b)Erro obtido com a aproximação criada e com a M/M/c em relação às simulações.



(a)



(b)

Figura 5.12: Sistemas com 12 servidores para $\rho = 0.6$ (a) Tempo médio de espera na fila (W_q) para 29 sistemas com diferentes μ's. Os pontos foram obtidos através dos calculados com MMc e com aproximação criada e através de simulações para as políticas de alocação rápida, aleatória e lenta. (b)Erro obtido com a aproximação criada e com a M/M/c em relação às simulações.

Tabela 5.3: Tempo médio de espera em fila para M/M/c. Os cálculos foram feitos para diferentes ρ's e para vários números c de servidores Homogêneos.

RESULTADO DOS CÁLCULOS COM **M/M/C** - Tempo médio de espera na fila				
Homogêneos	**ρ = 0,95**	**ρ = 0,90**	**ρ = 0,75**	**ρ = 0,6**
MM1	19	9.0	3	1.5
MM2	18.513	8.526	2.5714	1.125
MM3	18.140	8.171	2.271	0.8869
MM4	17.828	7.877	2.0377	0.7176
MM5	17.556	7.624	1.8472	0.5904
MM6	17.312	7.401	1.687	0.491
MM7	17.089	7.199	1.5485	0.4126
MM8	16.883	7.015	1.4279	0.3489
MM9	16.692	6.845	1.3214	0.2966
MM10	16.512	6.687	1.2264	0.2532
MM12	16.181	6.400	1.0641	0.1867
MM14	15.880	6.144	0.9304	0.1392
MM16	15.604	5.913	0.8183	0.1047
MM18	15.348	5.702	0.7232	0.0793
MM20	15.108	5.507	0.6417	0.0603

Capítulo 6

Discussões

Neste capítulo, novos rumos a serem seguidos são discutidos com o intuito de aprofundar o conhecimento dentro do campo da Teoria de Filas e, mais especificamente, para o caso de servidores heterogêneos. Pretende-se, também, identificar um novo tipo de aplicação em telecomunicações para o modelo proposto nesta dissertação.

6.1 Identificação de Região Ótima

A formulação desenvolvida nesta dissertação não consegue aproximar bem o modelo de servidores heterogêneos (Seção 1.5) quando a política de alocação é a que utiliza o servidor mais rápido primeiro e a heterogeneidade é baixa (Indice de Gini entre 0 e 0.5). Observa-se que a dificuldade do modelo é maior quando o número de servidores é pequeno, aproximadamente para c menor que dez. Quando se utiliza a alocação rápida há uma região na qual o tempo médio de espera em fila é ótimo, ou seja, menor possível. Nessa região ótima, que pode ser observada, por exemplo, nas Figuras 6.1 e 6.2, o valor do tempo de espera em fila é menor do que o obtido quando todos os servidores são homogêneos. Com isso, demonstra-se, através de simulações, que a heterogeneidade dos servidores tem uma distribuição ótima para a qual as medidas de desempenho do sistema são melhores do que se esse fosse composto apenas por servidores homogêneos.

Seria interessante se o estudo da presente dissertação fosse continuado com a finalidade de se criar uma aproximação que conseguisse prever a região ótima que surge quando é alocado primeiro o servidor mais rápido. Se tal aproximação fosse criada, não apenas o erro de aproximação para sistemas que se utilizam dessa política de alocação diminuiria, mas, também, novas possibilidades para projetos surgiriam. Com uma fórmula fechada

que aproxime o tempo de espera em fila em mãos, por meio de métodos de otimização tais como: algoritmo do Gradiente, método de Newton, etc., seria possível definir qual é o ponto ótimo para a heterogeneidade dos servidores. Assim, através dessa fórmula que aproxima o caso em que a política de alocação é a rápida e através de ferramentas de otimização, será possível projetar sistemas ótimos, nos quais o tempo de espera será o menor possível para uma determinada capacidade de trabalho e um determinado número de servidores.

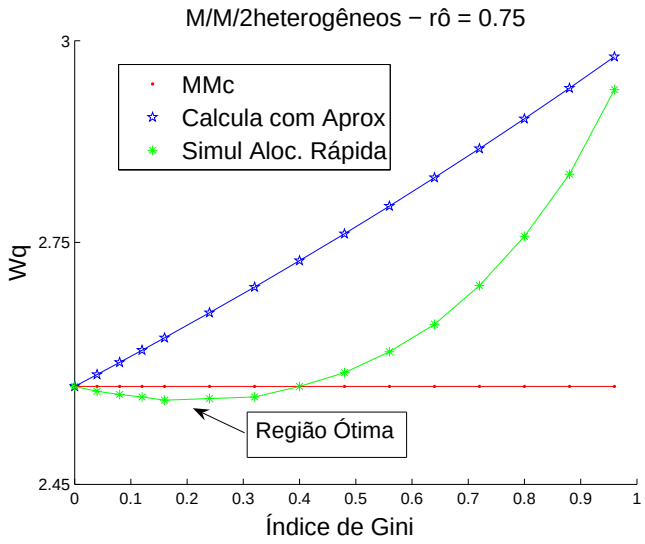


Figura 6.1: Sistemas com 2 servidores para $\rho = 0.75$ e política de alocação rápida. A figura mostra a região ótima onde o tempo médio de espera na fila (W_q) é menor do que para o caso em que os servidores são homogêneos. Isso só ocorre para o caso de alocação rápida e servidores heterogêneos.

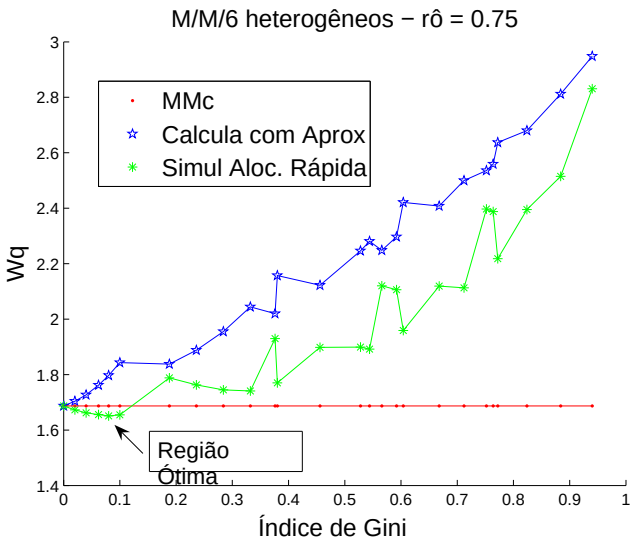


Figura 6.2: Sistemas com 6 servidores para $\rho = 0.75$ e política de alocação rápida. A figura mostra a região ótima onde o tempo médio de espera na fila (W_q) é menor do que para o caso em que os servidores são homogêneos. Isso só ocorre para o caso de alocação rápida e servidores heterogêneos. Observa-se que tal região é bem menor para c=6 do que para o caso da Figura 6.1 na qual c=2.

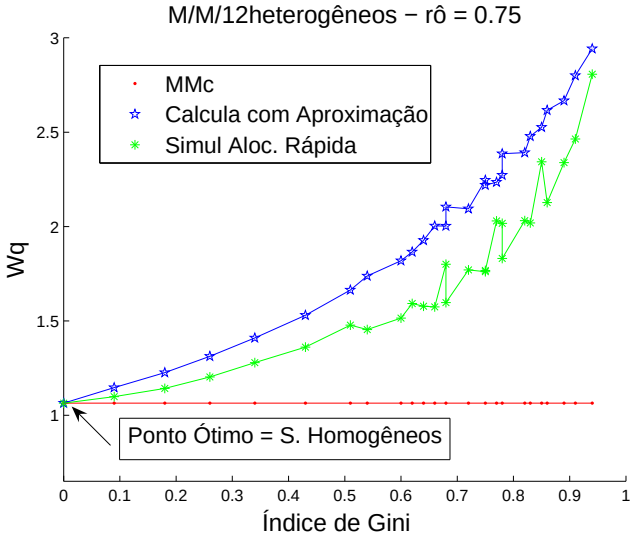


Figura 6.3: Sistemas com 12 servidores para $\rho = 0.75$ e política de alocação rápida. Para o caso de 12 servidores já não há uma região ótima como a que pode ser vista nas Figuras 6.1 e 6.2. Neste caso o ponto ótimo onde o tempo médio de espera na fila (W_q) é o menor possível ocorre quando os servidores são homogêneos.

6.2 Modelo Proposto x Tecnologia MPLS

A tecnologia chamada de MPLS (Multiprotocol Label Switching) atua dentro de uma rede IP. Tal tecnologia serve para aumentar o desempenho das redes IP, através do controle de trafego. Com a tecnologia MPLS é possível desviar os pacotes de dados de links que estejam congestionados ou falhos de uma forma dinâmica. Com uma gestão do tráfego da rede é possível melhorar a qualidade dos serviços e a confiabilidade dessas redes. Essa tecnologia utiliza rótulos (labels) para os pacotes IP, rótulos esses que determinam qual rota os pacotes seguirão.

Como a tecnologia MPLS suporta também o tráfego de voz e vídeo, surge uma oportunidade para o uso do modelo criado. A idéia é alocar os pacotes de voz para os servidores (rotas mpls) mais rápidos, pois esses pacotes necessitam de maior velocidade. Já os pacotes de dados são alocados para os servidores mais lentos, pois esses requerem mais confiabilidade do que velocidade. O modelo proposto nesta dissertação poderia ser usado para ajudar a encaminhar os pacotes. Para isso, o tempo esperado em fila (para entrar na rede) e o tempo esperado em serviço (transporte através das rotas) seriam calculados frequentemente. Com essas medidas em mãos seria possível definir quais são os melhores percursos que cada pacote deve seguir. Com isso a rede poderia ganhar em qualidade e em rapidez.

Capítulo 7

Conclusão

7.1 Conclusão

Em diferentes campos do conhecimento, a teoria de filas vem sendo aplicada. A criação de modelos analíticos mais realistas é uma rica oportunidade para várias aplicações. Particularmente, para a área de Pesquisa Operacional, existe um grande número de casos em que esses modelos podem ser aplicados. Com o advento da Internet, onde ocorrem vários processos de filas, e com a possibilidade de modelar-se o comportamento de vários outros sistemas, nos quais as filas também estão presentes, o desenvolvimento de modelos mais complexos e que representam melhor a realidade de todos esses casos se torna ainda mais relevante.

Neste trabalho, foi criada uma formulação para que se possa obter as medidas de performance, tempo médio de espera no sistema e em fila e número médio de pessoas no sistema e na fila, em sistemas com servidores heterogêneos. Essa formulação é uma generalização da formulação existente para M/M/c, que consiste, inicialmente, na expansão dos estados de probabilidade possíveis e, finalmente, na redução desses estados, através da escolha de manter apenas os estados de maior probabilidade. Do resultado, obteve-se um limite superior para as medidas de performance dos sistemas em questão. Logo, pode-se prever o pior caso no qual o desempenho desses sistemas pode resultar.

Com o intuito de avaliar a aplicabilidade dessa formulação, comparou-se essa com os casos já desenvolvidos na Teoria de Filas, que não consideram a heterogeneidade dos servidores. Devido à dificuldade encontrada para que se pudesse fazer a comparação direta entre o modelo desenvolvido aqui e o já existente na literatura, partiu-se para o uso de simulações com o intuito de calcular a distância entre essa e os casos a serem compara-

dos.A utilização da formulação criada é uma alternativa muito útil à tradicional M/M/c, pois, além de ser capaz de captar o comportamento desses sistemas mais efetivamente, a primeira se reduz à segunda quando os servidores são homogêneos e quando a utilização do sistema é tão grande que os efeitos das políticas de alocação se tornam desprezíveis. Chega-se, portanto, à conclusão de que o modelo desenvolvido aqui é um avanço em direção ao entendimento do comportamento de tais sistemas de filas.

É esperado que vários sistemas, que antes eram avaliados através de formulações que consideravam os servidores iguais e que, consequentemente, subestimavam essas medidas, possam ser recalculados e revistos com a nova formulação. Espera-se, também, que haja, a partir dessa, uma abertura para a continuação de vários trabalhos existentes na literatura, visto que, esses foram baseados somente em modelos que consideravam os servidores como sendo homogêneos. Entretanto, a falta de modelos exatos e que sejam aplicáveis na realidade, continua sendo um obstáculo real para aqueles que procuram modelar os sistemas de servidores heterogêneos através de teoria de filas.

Apêndice A

Simulação

A simulação é uma ferramenta usada para que, de certa forma, se possa imitar a realidade. Com a ajuda de simulações é possível experimentar modelos, observar as características resultantes deles, analisá-los de acordo com nossas estratégias e então tomar as devidas decisões.

Ela é útil quando se quer prever algum fenômeno cuja medição não seja fácil de ser obtida, principalmente para vários pontos ou regiões. Particularmente, para o campo de pesquisa operacional, as simulações são também muito importantes, já que difíceis decisões podem ser tomadas com mais firmeza quando se as tem para auxiliar. Além disso, elas também ajudam na elaboração de estratégias que conduzam nossas tomadas de decisões para um ponto ótimo.

Os Modelos simulados podem se aproximar muito da realidade, pois, esses capturam as características de um sistema sem ter que fazer tantas simplificações que distanciam o modelo dessa realidade, o que quase sempre é necessário quando são usadas equações fechadas.

Vantagens de se usar a simulação, como a não necessidade da existência previa do modelo na prática, e a capacidade de compressão do tempo, permitindo a obtenção, em poucas horas, de resultados que demorariam demasiadamente se tivessem que ser medidos, foram imprescindíveis para a elaboração deste trabalho.

Entretanto, deve-se deixar claro as limitações encontradas, como a dificuldade de obter resultados exatos, pois, apenas se conseguiu obter parâmetros estimados. Outra dificuldade foi a de se generalizar os parâmetros obtidos e, portanto, definir comportamentos gerais, pois, cada simulação se aplica somente ao caso simulado e às vezes, o número de casos necessários para que se chegue a uma conclusão genérica é muito grande.

Essa ferramenta teve grande importância neste trabalho, pois, ela ajudou a calibrar os modelos matemáticos, visto que, não existiam dados reais disponíveis. Usou-se a simulação nesta dissertação como ferramenta para ajudar na criação dos modelos matemáticos e para a calibração da formulação criada. A linguagem usada foi a GPSS (General Purpose Simulation System) e o "compilador"/ usado foi o GPSS World Program, que permite obter rapidamente as respostas, além de permitir a visualização da simulação enquanto essa está "rodando"/ e de fornecer ferramentas para o tratamento estatístico dos dados.

Através do GPSS as filas M/M/1 e M/M/c (para vários valores de c) tradicionais e depois as $M/M/c_{heterogneos}$ criadas (também para vários valores de c) foram modeladas. Primeiramente fez-se uma replicação para as filas MMc e se comparou os resultados com os obtidos por cálculo (usando as equações já conhecidas, veja seção 3.4) com o intuito de calibrar o sistema. Devido ao fato de que o intervalo de confiança de 95% resultante foi muito grande, houve uma variância muito grande nos resultados, por isso, o erro entre o tempo em fila médio encontrado com as simulações foi muito grande quando comparado com o calculado. Para se diminuir o intervalo de confiança e aumentar o número de amostras aumentou-se gradativamente o número de replicações e fez-se a média das médias dos tempos de espera em fila. Com 200 replicações e com 700 000 unidades de tempo para cada sistema, o erro já variava na faixa entre 4% a menos de 1% de acordo com o coeficiente de utilização ρ escolhido. A variância nos resultados e consequentemente o erro de simulação, aumentaram para o ρ mais próximo de um. Com o modelo calibrado se fez várias simulações para os casos em que se tinha servidores heterogêneos, objetivando confirmar a veracidade das formulas desenvolvidas, como também para entender a influencia que os diferentes tipos de alocações tinham no sistema.

Em Schriber (1991), foi achado o material inicial necessário para que se pudesse adquirir conhecimento sobre a linguagem GPSS e também para entender mais sobre simulações. O compilador GPSS World Student pode ser achado no site do minuteman: http://www.minutemansoftware.com

Apêndice B

Resultados Numéricos Calculados e Simulados

Esta seção foi separada para se colocar as tabelas criadas com os valores de simulações e os valores obtidos através dos cálculos com a formulação da MMc e dos obtidos com a formulação criada neste trabalho. Todas as tabelas apresentadas aqui mostram os resultados para a medida de performance tempo médio na fila por job.

A coluna 1 de cada tabela mostra a numeração dos sistemas criados. Quando foi aumentado o número de servidores aumentou-se também o número de sistemas criados. Cada sistema se diferencia pela divisão da capacidade de processamento entre os servidores, o que está representado nas colunas seguintes.

As colunas chamadas de IG representam o índice de Gini referente à distribuição das taxas μ's definidas para cada sistema. No topo de cada figura há os dados de simulação, número de replicações e quantidade de unidades de tempo utilizadas em cada uma das replicações. E ρ é o coeficiente de utilização ao qual cada sistema foi submetido. Abaixo estão as tabelas:

Tabela B.1: Resultados para M/M/2heterogênea para $\rho = 0.9$. Nesta tabelas, assim como nas tabela B.2 e B.3 a coluna 1 mostra os 15 sistemas criados para os casos com 2 servidores. Cada sistema se diferencia pela divisão da capacidade de processamento entre os servidores, o que está representado nas colunas 2-3. Para cada ponto simulado foram realizadas 200 replicações onde se deixou o sistema rodar por 500000 unidades de tempo. A coluna IG representa o índice de Gini.

Fila M/M/2 - ρ = 0,9				**Tempo Médio de Espera em Fila** Simulado COM 200 REPLICAÇAO) - (500 000) UT				
				Alocação rápida	Alocação aleatoria	Alocação demorada	MMc	Calculado Com Aproximação
Sistema	μ1 (%)	μ2 (%)	IG	ρ = 0,9 simulado	ρ = 0,9 simulado	ρ = 0,9 simulado		
1	98	2	0.96	8.92	8.935	8.958	8.526	8.980
2	94	6	0.88	8.816	8.866	8.894	8.526	8.940
3	90	10	0.8	8.732	8.803	8.853	8.526	8.901
4	86	14	0.72	8.642	8.747	8.79	8.526	8.862
5	82	18	0.64	8.613	8.695	8.745	8.526	8.824
6	78	22	0.56	8.57	8.654	8.702	8.526	8.785
7	74	26	0.48	8.539	8.613	8.666	8.526	8.747
8	70	30	0.4	8.516	8.582	8.632	8.526	8.710
9	66	34	0.32	8.501	8.557	8.595	8.526	8.672
10	62	38	0.24	8.493	8.533	8.567	8.526	8.635
11	58	42	0.16	8.492	8.523	8.541	8.526	8.599
12	56	44	0.12	8.495	8.515	8.543	8.526	8.581
13	54	46	0.08	8.499	8.513	8.522	8.526	8.562
14	52	48	0.04	8.499	8.509	8.518	8.526	8.544
15	50	50	0	8.503	8.509	8.51	8.526	8.526

Tabela B.2: Resultados para M/M/2heterogênea para $\rho = 0.75$.

Fila M/M/2 - ρ = 0,75				**Tempo Médio de Espera em Fila** Simulado COM 200 REPLICAÇAO) - (500 000) UT				
				Alocação rápida	Alocação aleatoria	Alocação demorada	MMc	Calculado Com Aproximação
Sistema	μ1 (%)	μ2 (%)	IG	ρ = 0,75 simulado	ρ = 0,75 simulado	ρ = 0,75 simulado		
1	98	2	0.96	2.937	2.957	2.97	2.571	2.980
2	94	6	0.88	2.832	2.887	2.918	2.571	2.941
3	90	10	0.8	2.755	2.826	2.869	2.571	2.903
4	86	14	0.72	2.694	2.773	2.822	2.571	2.866
5	82	18	0.64	2.646	2.726	2.779	2.571	2.830
6	78	22	0.56	2.612	2.688	2.74	2.571	2.795
7	74	26	0.48	2.586	2.654	2.706	2.571	2.761
8	70	30	0.4	2.569	2.627	2.675	2.571	2.727
9	66	34	0.32	2.556	2.606	2.645	2.571	2.695
10	62	38	0.24	2.554	2.589	2.621	2.571	2.663
11	58	42	0.16	2.552	2.578	2.598	2.571	2.632
12	56	44	0.12	2.556	2.573	2.59	2.571	2.616
13	54	46	0.08	2.559	2.571	2.582	2.571	2.601
14	52	48	0.04	2.563	2.569	2.575	2.571	2.586
15	50	50	0	2.569	2.569	2.569	2.571	2.571

Tabela B.3: Resultados para M/M/2heterogênea para $\rho = 0.6$.

Fila M/M/2 - ρ = 0,6				Tempo Médio de Espera em Fila Simulado COM 200 REPLICAÇAO) - (500 000) UT				
				Alocação rápida	Alocação aleatoria	Alocação demorada	MMc	Calculado Com Aproximação
Sistema	μ1(%)	μ2(%)	IG	ρ = 0,6 simulado	ρ = 0,6 simulado	ρ = 0,6 simulado		
1	98	2	0.96	1.433	1.459	1.473	1.125	1.480
2	94	6	0.88	1.333	1.393	1.424	1.125	1.442
3	90	10	0.8	1.262	1.337	1.378	1.125	1.406
4	86	14	0.72	1.211	1.291	1.338	1.125	1.372
5	82	18	0.64	1.174	1.252	1.301	1.125	1.339
6	78	22	0.56	1.147	1.218	1.269	1.125	1.308
7	74	26	0.48	1.13	1.192	1.239	1.125	1.278
8	70	30	0.4	1.117	1.17	1.213	1.125	1.250
9	66	34	0.32	1.111	1.153	1.189	1.125	1.223
10	62	38	0.24	1.109	1.14	1.169	1.125	1.197
11	58	42	0.16	1.11	1.131	1.151	1.125	1.172
12	56	44	0.12	1.113	1.128	1.144	1.125	1.160
13	54	46	0.08	1.116	1.126	1.137	1.125	1.148
14	52	48	0.04	1.12	1.124	1.13	1.125	1.136
15	50	50	0	1.125	1.124	1.125	1.125	1.125

Tabela B.4: Resultados para M/M/2heterogênea para $\rho = 0.9$. Nesta tabela, assim como nas tabela B.5 e B.6 a coluna 1 mostra os 24 sistemas criados para os casos com 3 servidores. Cada sistema se diferencia pela divisão da capacidade de processamento entre os servidores, o que está representado nas colunas 2-4. Para cada ponto simulado foram realizadas 200 replicações onde se deixou o sistema rodar por 700000 unidades de tempo. A coluna IG representa o índice de Gini.

Fila M/M/3 - ρ = 0,9					Tempo Médio de Espera em Fila Simulado (200 replicações - 700 000 UT)				
					Alocação rápida	Alocação aleatoria	Alocação demorada	MMc	Calculado Com Aproximação
Sistema	μ3(%)	μ2(%)	μ1(%)	IG	ρ = 0,9 simulado	ρ = 0,9 simulado	ρ = 0,9 simulado		
1	98	1	1	0.97	8.901	8.956	8.945	8.171	8.980
2	94	4	2	0.92	8.8	8.863	8.855	8.171	8.939
3	86	12	2	0.84	8.626	8.72	8.786	8.171	8.859
4	78	20	2	0.76	8.511	8.614	8.687	8.171	8.781
5	74	24	2	0.72	8.467	8.569	8.648	8.171	8.742
6	66	32	2	0.64	8.423	8.51	8.575	8.171	8.665
7	58	40	2	0.56	8.411	8.474	8.498	8.171	8.590
8	54	44	2	0.52	8.411	8.47	8.466	8.171	8.553
9	50	48	2	0.48	8.427	8.473	8.479	8.171	8.516
10	76	12	12	0.64	8.389	8.546	8.613	8.171	8.736
11	72	16	12	0.6	8.316	8.482	8.572	8.171	8.694
12	64	24	12	0.52	8.234	8.387	8.477	8.171	8.610
13	56	32	12	0.44	8.197	8.324	8.401	8.171	8.528
14	48	40	12	0.36	8.186	8.295	8.346	8.171	8.447
15	44	44	12	0.32	8.197	8.292	8.331	8.171	8.407
16	60	20	20	0.4	8.183	8.321	8.405	8.171	8.536
17	56	24	20	0.36	8.153	8.28	8.357	8.171	8.493
18	52	28	20	0.32	8.133	8.249	8.322	8.171	8.449
19	48	32	20	0.28	8.124	8.224	8.285	8.171	8.406
20	44	36	20	0.24	8.124	8.212	8.257	8.171	8.364
21	44	28	28	0.16	8.121	8.192	8.201	8.171	8.321
22	40	32	28	0.12	8.123	8.171	8.2	8.171	8.277
23	36	36	28	0.08	8.132	8.166	8.182	8.171	8.233
24	33.333	33.333	33.333	0	8.15	8.16	8.15	8.171	8.171

Tabela B.5: Resultados para M/M/3heterogênea para $\rho = 0.75$.

Fila M/M/3 - ρ = 0,75					Tempo Médio de Espera em Fila Simulado (200 replicações - 700 000 UT)				
					Alocação rápida	Alocação aleatoria	Alocação demorada	MMc	Calculado Com Aproximação
Sistema	μ3(%)	μ2(%)	μ1(%)	IG	ρ = 0,75 simulado	ρ = 0,75 simulado	ρ = 0,75 simulado		
1	98	1	1	0.97	2.93	2.958	2.967	2.271	2.980
2	94	4	2	0.92	2.811	2.876	2.911	2.271	2.940
3	86	12	2	0.84	2.645	2.741	2.81	2.271	2.863
4	78	20	2	0.76	2.548	2.645	2.725	2.271	2.790
5	74	24	2	0.72	2.517	2.609	2.688	2.271	2.755
6	66	32	2	0.64	2.48	2.559	2.625	2.271	2.687
7	58	40	2	0.56	2.472	2.532	2.576	2.271	2.623
8	54	44	2	0.52	2.475	2.527	2.556	2.271	2.592
9	50	48	2	0.48	2.485	2.526	2.542	2.271	2.562
10	76	12	12	0.64	2.434	2.581	2.665	2.271	2.745
11	72	16	12	0.6	2.379	2.527	2.618	2.271	2.707
12	64	24	12	0.52	2.311	2.448	2.539	2.271	2.633
13	56	32	12	0.44	2.283	2.4	2.476	2.271	2.564
14	48	40	12	0.36	2.282	2.375	2.434	2.271	2.498
15	44	44	12	0.32	2.291	2.375	2.419	2.271	2.466
16	60	20	20	0.4	2.28	2.394	2.478	2.271	2.567
17	56	24	20	0.36	2.257	2.361	2.44	2.271	2.530
18	52	28	20	0.32	2.244	2.335	2.409	2.271	2.495
19	48	32	20	0.28	2.236	2.319	2.382	2.271	2.460
20	44	36	20	0.24	2.238	2.308	2.361	2.271	2.426
21	44	28	28	0.16	2.244	2.289	2.33	2.271	2.388
22	40	32	28	0.12	2.241	2.277	2.308	2.271	2.354
23	36	36	28	0.08	2.248	2.273	2.291	2.271	2.320
24	33.333	33.333	33.333	0	2.266	2.267	2.266	2.271	2.271

Tabela B.6: Tabela com resultados para M/M/3heterogênea para $\rho = 0.6$.

Fila M/M/3 - ρ = 0,6					Tempo Médio de Espera em Fila Simulado (200 replicações - 700 000 UT)				
					Alocação rápida	Alocação aleatoria	Alocação demorada	MMc	Calculado Com Aproximação
Sistema	μ3(%)	μ2(%)	μ1(%)	IG	ρ = 0,6 simulado	ρ = 0,6 simulado	ρ = 0,6 simulado		
1	98	1	1	0.97	1.429	1.459	1.472	0.887	1.4799
2	94	4	2	0.92	1.309	1.381	1.419	0.887	1.4405
3	86	12	2	0.84	1.159	1.257	1.328	0.887	1.3681
4	78	20	2	0.76	1.082	1.178	1.255	0.887	1.3026
5	74	24	2	0.72	1.060	1.149	1.224	0.887	1.2721
6	66	32	2	0.64	1.036	1.111	1.172	0.887	1.2153
7	58	40	2	0.56	1.032	1.091	1.132	0.887	1.1634
8	54	44	2	0.52	1.037	1.087	1.116	0.887	1.139
9	50	48	2	0.48	1.044	1.087	1.104	0.887	1.1157
10	76	12	12	0.64	0.9860	1.121	1.202	0.887	1.2584
11	72	16	12	0.6	0.9460	1.078	1.163	0.887	1.2255
12	64	24	12	0.52	0.8980	1.016	1.099	0.887	1.1646
13	56	32	12	0.44	0.8800	0.981	1.050	0.887	1.1095
14	48	40	12	0.36	0.8820	0.966	1.016	0.887	1.0593
15	44	44	12	0.32	0.8900	0.965	1.005	0.887	1.0359
16	60	20	20	0.4	0.8830	0.977	1.049	0.887	1.1066
17	56	24	20	0.36	0.8670	0.953	1.021	0.887	1.0783
18	52	28	20	0.32	0.8580	0.935	0.997	0.887	1.0514
19	48	32	20	0.28	0.8550	0.923	0.976	0.887	1.0258
20	44	36	20	0.24	0.8580	0.917	0.960	0.887	1.0015
21	44	28	28	0.16	0.8650	0.901	0.936	0.887	0.9693
22	40	32	28	0.12	0.8650	0.894	0.919	0.887	0.9454
23	36	36	28	0.08	0.8710	0.891	0.906	0.887	0.9226
24	33.333	33.333	33.333	0	0.8860	0.887	0.886	0.887	0.8869

Tabela B.7: Resultados para M/M/6heterogênea para $\rho = 0.9$. Nesta tabela, assim como nas tabela B.11 e B.12 a coluna 1 mostra os 26 sistemas criados para os casos com 6 servidores. Cada sistema se diferencia pela divisão da capacidade de processamento entre os servidores, o que está representado nas colunas 2-7. Para cada ponto simulado foram realizadas 200 replicações onde se deixou o sistema rodar por 700000 unidades de tempo. A coluna IG representa o índice de Gini.

Fila M/M/6 - ρ = 0,9								Tempo Médio de Espera em Fila Simulado COM 200 REPLICAÇAO) - (700 000) UT				
								Alocação rápida	Alocação aleatoria	Alocação demorada	MMc	Calculado Com Aproximação
Sistema	μ6(%)	μ5(%)	μ4(%)	μ3(%)	μ2(%)	μ1(%)	IG	ρ = 0,9 simulado	ρ = 0,9 simulado	ρ = 0,9 simulado		
1	95	1	1	1	1	1	0.94	8.822	8.879	8.906	7.401	8.948
2	81	15	1	1	1	1	0.88	8.494	8.63	8.714	7.401	8.8056
3	66	30	1	1	1	1	0.82	8.344	8.481	8.556	7.401	8.6579
4	51	45	1	1	1	1	0.76	8.329	8.409	8.448	7.401	8.5151
5	48	48	1	1	1	1	0.75	8.33	8.414	8.439	7.401	8.4871
6	67	15	15	1	1	1	0.77	8.141	8.356	8.452	7.401	8.6186
7	52	30	15	1	1	1	0.71	8.005	8.209	8.31	7.401	8.4558
8	41	41	15	1	1	1	0.67	8.000	8.158	8.237	7.401	8.3402
9	37	30	30	1	1	1	0.59	7.98	8.06	8.125	7.401	8.2076
10	32.5	32.5	32	1	1	1	0.57	7.992	8.087	8.1	7.401	8.1436
11	53	15	15	15	1	1	0.6	7.809	8.048	8.198	7.401	8.377
12	38	30	15	15	1	1	0.54	7.711	7.896	8.031	7.401	8.1959
13	34	34	15	15	1	1	0.53	7.718	7.882	8.012	7.401	8.1489
14	28	28	27	15	1	1	0.46	7.704	7.847	7.905	7.401	7.9882
15	24.5	24.5	24.5	24.5	1	1	0.38	7.735	7.815	7.815	7.401	7.8552
16	39	15	15	15	15	1	0.38	7.527	7.707	7.84	7.401	8.0609
17	27	27	15	15	15	1	0.33	7.483	7.639	7.746	7.401	7.8988
18	23	23	23	15	15	1	0.28	7.485	7.621	7.674	7.401	7.7737
19	21	21	21	21	15	1	0.24	7.491	7.596	7.596	7.401	7.6806
20	19.8	19.8	19.8	19.8	19.8	1	0.19	7.526	7.578	7.579	7.401	7.6129
21	25	15	15	15	15	15	0.1	7.337	7.397	7.502	7.401	7.6435
22	20	20	15	15	15	15	0.08	7.331	7.36	7.453	7.401	7.5675
23	18.5	18.5	18	15	15	15	0.06	7.332	7.366	7.412	7.401	7.5127
24	17.5	17.5	17.5	17.5	15	15	0.04	7.32	7.387	7.404	7.401	7.4602
25	17	17	17	17	17	15	0.02	7.335	7.397	7.387	7.401	7.4261
26	16.667	16.667	16.667	16.667	16.667	16.667	0	7.373	7.409	7.373	7.401	7.4011

Tabela B.8: Resultados para M/M/6heterogênea para $\rho = 0.75$.

Fila M/M/3 - ρ = 0,75								Tempo Médio de Espera em Fila Simulado (200 replicações - 700 000 UT)				
								Alocação rápida	Alocação aleatoria	Alocação demorada	MMc	Calculado Com Aproximação
Sistema	μ6(%)	μ5(%)	μ4(%)	μ3(%)	μ2(%)	μ1(%)	IG	ρ = 0,75 simulado	ρ = 0,75 simulado	ρ = 0,75 simulado		
1	95	1	1	1	1	1	0.94	2.829	2.895	2.925	1.687	2.948
2	81	15	1	1	1	1	0.88	2.513	2.670	2.742	1.687	2.812
3	66	30	1	1	1	1	0.82	2.393	2.532	2.603	1.687	2.680
4	51	45	1	1	1	1	0.76	2.386	2.480	2.518	1.687	2.559
5	48	48	1	1	1	1	0.75	2.395	2.483	2.509	1.687	2.536
6	67	15	15	1	1	1	0.77	2.216	2.424	2.535	1.687	2.637
7	52	30	15	1	1	1	0.71	2.111	2.300	2.396	1.687	2.500
8	41	41	15	1	1	1	0.67	2.117	2.272	2.342	1.687	2.408
9	37	30	30	1	1	1	0.59	2.104	2.203	2.249	1.687	2.297
10	32.5	32.5	32	1	1	1	0.57	2.118	2.199	2.224	1.687	2.248
11	53	15	15	15	1	1	0.6	1.957	2.171	2.303	1.687	2.421
12	38	30	15	15	1	1	0.54	1.890	2.069	2.171	1.687	2.281
13	34	34	15	15	1	1	0.53	1.897	2.066	2.154	1.687	2.246
14	28	28	27	15	1	1	0.46	1.896	2.015	2.072	1.687	2.122
15	24.5	24.5	24.5	24.5	1	1	0.38	1.928	1.986	2.008	1.687	2.020
16	39	15	15	15	15	1	0.38	1.768	1.925	2.044	1.687	2.157
17	27	27	15	15	15	1	0.33	1.739	1.866	1.954	1.687	2.044
18	23	23	23	15	15	1	0.28	1.743	1.845	1.901	1.687	1.955
19	21	21	21	21	15	1	0.24	1.761	1.828	1.861	1.687	1.888
20	19.8	19.8	19.8	19.8	19.8	1	0.19	1.786	1.819	1.831	1.687	1.838
21	25	15	15	15	15	15	0.1	1.653	1.712	1.772	1.687	1.843
22	20	20	15	15	15	15	0.08	1.649	1.694	1.743	1.687	1.797
23	18.5	18.5	18	15	15	15	0.06	1.654	1.691	1.722	1.687	1.762
24	17.5	17.5	17.5	17.5	15	15	0.04	1.660	1.685	1.705	1.687	1.727
25	17	17	17	17	17	15	0.02	1.671	1.679	1.693	1.687	1.704
26	16.667	16.667	16.667	16.667	16.667	16.667	0	1.685	1.683	1.685	1.687	1.687

Tabela B.9: Resultados para M/M/6heterogênea para $\rho = 0.6$.

Fila M/M/6 - ρ = 0,6								Tempo Médio de Espera em Fila Simulado (200 replicações - 700 000 UT)				
								Alocação rápida	Alocação aleatoria	Alocação demorada	MMc	Calculado Com Aproximação
Sistema	μ6(%)	μ5(%)	μ4(%)	μ3(%)	μ2(%)	μ1(%)	IG	ρ = 0,6 simulado	ρ = 0,6 simulado	ρ = 0,6 simulado		
1	95	1	1	1	1	1	0.94	1.319	1.399	1.431	0.491	1.448
2	81	15	1	1	1	1	0.88	1.031	1.195	1.266	0.491	1.321
3	66	30	1	1	1	1	0.82	0.944	1.087	1.152	0.491	1.207
4	51	45	1	1	1	1	0.76	0.945	1.047	1.084	0.491	1.112
5	48	48	1	1	1	1	0.75	0.953	1.048	1.076	0.491	1.094
6	67	15	15	1	1	1	0.77	0.794	0.994	1.092	0.491	1.163
7	52	30	15	1	1	1	0.71	0.728	0.902	0.986	0.491	1.056
8	41	41	15	1	1	1	0.67	0.736	0.885	0.943	0.491	0.989
9	37	30	30	1	1	1	0.59	0.734	0.835	0.874	0.491	0.902
10	32.5	32.5	32	1	1	1	0.57	0.746	0.832	0.856	0.491	0.869
11	53	15	15	15	1	1	0.6	0.622	0.801	0.907	0.491	0.981
12	38	30	15	15	1	1	0.54	0.584	0.735	0.817	0.491	0.883
13	34	34	15	15	1	1	0.53	0.589	0.732	0.805	0.491	0.860
14	28	28	27	15	1	1	0.46	0.594	0.700	0.744	0.491	0.775
15	24.5	24.5	24.5	24.5	1	1	0.38	0.620	0.683	0.698	0.491	0.705
16	39	15	15	15	15	1	0.38	0.515	0.635	0.719	0.491	0.780
17	27	27	15	15	15	1	0.33	0.503	0.600	0.663	0.491	0.713
18	23	23	23	15	15	1	0.28	0.509	0.586	0.628	0.491	0.658
19	21	21	21	21	15	1	0.24	0.522	0.579	0.601	0.491	0.616
20	19.8	19.8	19.8	19.8	19.8	1	0.19	0.541	0.574	0.582	0.491	0.585
21	25	15	15	15	15	15	0.1	0.471	0.504	0.545	0.491	0.575
22	20	20	15	15	15	15	0.08	0.470	0.496	0.528	0.491	0.552
23	18.5	18.5	18	15	15	15	0.06	0.472	0.494	0.516	0.491	0.534
24	17.5	17.5	17.5	17.5	15	15	0.04	0.477	0.492	0.505	0.491	0.515
25	17	17	17	17	17	15	0.02	0.483	0.490	0.498	0.491	0.502
26	16.667	16.667	16.667	16.667	16.667	16.667	0	0.492	0.490	0.492	0.491	0.491

Tabela B.10: Resultados para M/M/12heterogênea para $\rho = 0.9$. Nesta tabela, assim como nas tabela B.11 e B.12 a coluna 1 mostra os 29 sistemas criados para os casos com 12 servidores. Cada sistema se diferencia pela divisão da capacidade de processamento entre os servidores, o que está representado nas colunas 2-13. Para cada ponto simulado foram realizadas 250 replicações onde se deixou o sistema rodar por 700000 unidades de tempo. A coluna IG representa o índice de Gini.

Fila M/M/12 ρ = 0,9														Tempo Médio de Espera em Fila Simulado (250 REPLICAÇÕES - 700 000 UT)				
														Alocação rápida	Alocação aleatória	Alocação demorada	MMc	Calculado Com Aproximação
Sistema	μ12	μ11	μ10	μ9	μ8	μ7	μ6	μ5	μ4	μ3	μ2	μ1	IG	ρ = 0,9 simulado	ρ = 0,9 simulado	ρ = 0,9 simulado		
1	94.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.94**	8.8	8.856	8.899	6.400	8.942
2	80.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.91**	8.437	8.649	8.673	6.400	8.793
3	65.00	30.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.89**	8.281	8.424	8.531	6.400	8.644
4	47.50	47.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.85**	8.281	8.376	8.422	6.400	8.477
5	65.50	15.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.86**	8.055	8.315	8.446	6.400	8.595
6	50.50	30.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.83**	7.891	8.156	8.273	6.400	8.431
7	40.25	40.25	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.82**	7.923	8.072	8.211	6.400	8.322
8	35.50	30.00	30.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.78**	7.89	8.008	8.096	6.400	8.178
9	32.00	32.00	31.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.77**	7.897	8.017	8.073	6.400	8.127
10	51.00	15.00	15.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.78**	7.65	7.972	8.147	6.400	8.337
11	36.00	30.00	15.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.75**	7.569	7.867	7.978	6.400	8.153
12	33.00	33.00	15.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.75**	7.581	7.857	7.968	6.400	8.117
13	27.00	27.00	27.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.72**	7.569	7.754	7.853	6.400	7.953
14	24.00	24.00	24.00	24.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.68**	7.603	7.724	7.781	6.400	7.832
15	36.50	15.00	15.00	15.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.68**	7.344	7.634	7.796	6.400	7.995
16	25.75	25.75	15.00	15.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.66**	7.305	7.543	7.672	6.400	7.846
17	22.50	22.00	22.00	15.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.64**	7.299	7.525	7.619	6.400	7.737
18	20.50	20.50	20.50	20.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.62**	7.32	7.504	7.572	6.400	7.651
19	19.50	19.25	19.25	19.25	19.25	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.60**	7.193	7.417	7.518	6.400	7.587
20	18.50	18.50	15.00	15.00	15.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	**0.54**	7.107	7.296	7.372	6.400	7.483
21	16.25	16.25	16.25	16.25	16.00	16.00	0.50	0.50	0.50	0.50	0.50	0.50	**0.51**	7.145	7.289	7.319	6.400	7.369
22	14.00	14.00	14.00	14.00	14.00	14.00	13.50	0.50	0.50	0.50	0.50	0.50	**0.43**	6.964	7.085	7.124	6.400	7.172
23	12.25	12.25	12.25	12.25	12.25	12.25	12.25	12.25	0.50	0.50	0.50	0.50	**0.34**	6.809	6.939	6.949	6.400	6.987
24	11.00	11.00	11.00	11.00	11.00	11.00	11.00	11.00	10.50	0.50	0.50	0.50	**0.26**	6.67	6.754	6.793	6.400	6.831
25	10.00	10.00	10.00	10.00	10.00	10.00	9.75	9.75	9.75	9.75	0.50	0.50	**0.18**	6.556	6.642	6.643	6.400	6.687
26	11.00	11.00	11.00	11.00	11.00	11.00	11.00	11.00	10.50	0.50	0.50	0.50	**0.26**	6.67	6.771	6.793	6.400	6.831
27	10.00	10.00	10.00	10.00	10.00	10.00	9.75	9.75	9.75	9.75	0.50	0.50	**0.18**	6.556	6.622	6.643	6.400	6.687
28	9.25	9.25	9.00	9.00	9.00	9.00	9.00	9.00	9.00	9.00	9.00	0.50	**0.09**	6.455	6.486	6.508	6.400	6.551
29	8.333	8.333	8.333	8.333	8.334	8.334	8.334	8.334	8.333	8.333	8.333	8.333	**0.000**	6.375	6.372	6.375	6.400	6.400

Tabela B.11: Resultados para M/M/12heterogênea para $\rho = 0.75$.

Fila M/M/12 ρ = 0,75														Tempo Médio de Espera em Fila Simulado (250 REPLICAÇÕES - 700 000 UT)				
Sistema	μ12	μ11	μ10	μ9	μ8	μ7	μ6	μ5	μ4	μ3	μ2	μ1	IG	Alocação rápida ρ = 0,75 simulado	Alocação aleatória ρ = 0,75 simulado	Alocação demorada ρ = 0,75 simulado	MMc	Calculado Com Aproximação
1	94.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.94**	2.805	2.887	2.916	1.064	2.9422
2	80.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.91**	2.462	2.638	2.724	1.064	2.8001
3	65.00	30.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.89**	2.337	2.506	2.585	1.064	2.6668
4	47.50	47.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.85**	2.341	2.457	2.493	1.064	2.5265
5	65.50	15.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.86**	2.125	2.38	2.505	1.064	2.616
6	50.50	30.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.83**	2.017	2.258	2.366	1.064	2.4781
7	40.25	40.25	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.82**	2.029	2.234	2.316	1.064	2.3919
8	35.50	30.00	30.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.78**	2.015	2.164	2.222	1.064	2.2724
9	32.00	32.00	31.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.77**	2.028	2.162	2.206	1.064	2.2349
10	51.00	15.00	15.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.78**	1.829	2.112	2.258	1.064	2.3863
11	36.00	30.00	15.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.75**	1.759	2.013	2.131	1.064	2.2457
12	33.00	33.00	15.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.75**	1.765	2.014	2.121	1.064	2.2195
13	27.00	27.00	27.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.72**	1.768	1.965	2.037	1.064	2.0942
14	24.00	24.00	24.00	24.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.68**	1.798	1.943	1.979	1.064	2.0024
15	36.50	15.00	15.00	15.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.68**	1.595	1.849	1.986	1.064	2.1045
16	25.75	25.75	15.00	15.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.66**	1.572	1.801	1.906	1.064	2.0033
17	22.50	22.00	22.00	15.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.64**	1.576	1.781	1.861	1.064	1.9268
18	20.50	20.50	20.50	20.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.62**	1.59	1.771	1.825	1.064	1.8655
19	19.50	19.25	19.25	19.25	19.25	0.50	0.50	0.50	0.50	0.50	0.50	0.50	**0.60**	1.513	1.672	1.763	1.064	1.819
20	18.50	18.50	15.00	15.00	15.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	**0.54**	1.452	1.618	1.684	1.064	1.7379
21	16.25	16.25	16.25	16.25	16.00	16.00	0.50	0.50	0.50	0.50	0.50	0.50	**0.51**	1.476	1.614	1.646	1.064	1.6636
22	14.00	14.00	14.00	14.00	14.00	14.00	13.50	0.50	0.50	0.50	0.50	0.50	**0.43**	1.359	1.489	1.516	1.064	1.5302
23	12.25	12.25	12.25	12.25	12.25	12.25	12.25	12.25	0.50	0.50	0.50	0.50	**0.34**	1.277	1.378	1.401	1.064	1.4096
24	11.00	11.00	11.00	11.00	11.00	11.00	11.00	11.00	10.50	0.50	0.50	0.50	**0.26**	1.201	1.283	1.304	1.064	1.3123
25	10.00	10.00	10.00	10.00	10.00	10.00	9.75	9.75	9.75	9.75	0.50	0.50	**0.18**	1.14	1.201	1.218	1.064	1.2261
26	11.00	11.00	11.00	11.00	11.00	11.00	11.00	11.00	10.50	0.50	0.50	0.50	**0.26**	1.201	1.283	1.304	1.064	1.3123
27	10.00	10.00	10.00	10.00	10.00	10.00	9.75	9.75	9.75	9.75	0.50	0.50	**0.18**	1.14	1.201	1.218	1.064	1.2261
28	9.25	9.25	9.00	9.00	9.00	9.00	9.00	9.00	9.00	9.00	9.00	0.50	**0.09**	1.097	1.127	1.139	1.064	1.1464
29	8.333	8.333	8.333	8.333	8.334	8.334	8.334	8.334	8.333	8.333	8.333	8.333	**0.000**	1.062	1.062	1.062	1.064	1.0641

Tabela B.12: Resultados para M/M/12heterogênea para $\rho = 0.6$.

Fila M/M/12 ρ = 0,6														Tempo Médio de Espera em Fila Simulado (250 REPLICAÇÕES - 700 000 UT)				
Sistema	μ12	μ11	μ10	μ9	μ8	μ7	μ6	μ5	μ4	μ3	μ2	μ1	IG	Alocação rápida ρ = 0,6 simulado	Alocação aleatória ρ = 0,6 simulado	Alocação demorada ρ = 0,6 simulado	MMc	Calculado Com Aproximação
1	94.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.94	1.291	1.386	1.422	0.187	1.4424
2	80.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.91	0.974	1.172	1.249	0.187	1.3097
3	65.00	30.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.89	0.883	1.063	1.136	0.187	1.1958
4	47.50	47.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.85	0.893	1.025	1.063	0.187	1.0858
5	65.50	15.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.86	0.704	0.958	1.067	0.187	1.1444
6	50.50	30.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.83	0.635	0.867	0.961	0.187	1.0374
7	40.25	40.25	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.82	0.644	0.851	0.923	0.187	0.9752
8	35.50	30.00	30.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.78	0.645	0.804	0.851	0.187	0.8839
9	32.00	32.00	31.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.77	0.655	0.803	0.838	0.187	0.8585
10	51.00	15.00	15.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.78	0.502	0.757	0.872	0.187	0.9527
11	36.00	30.00	15.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.75	0.465	0.694	0.787	0.187	0.8569
12	33.00	33.00	15.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.75	0.470	0.692	0.777	0.187	0.8401
13	27.00	27.00	27.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.72	0.476	0.664	0.719	0.187	0.7552
14	24.00	24.00	24.00	24.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.68	0.502	0.648	0.679	0.187	0.6928
15	36.50	15.00	15.00	15.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.68	0.373	0.584	0.679	0.187	0.7429
16	25.75	25.75	15.00	15.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.66	0.362	0.556	0.631	0.187	0.6844
17	22.50	22.00	22.00	15.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.64	0.366	0.544	0.600	0.187	0.6386
18	20.50	20.50	20.50	20.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.62	0.377	0.538	0.578	0.187	0.6012
19	19.50	19.25	19.25	19.25	19.25	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.60	0.332	0.488	0.538	0.187	0.572
20	18.50	18.50	15.00	15.00	15.00	15.00	0.50	0.50	0.50	0.50	0.50	0.50	0.54	0.308	0.450	0.492	0.187	0.5172
21	16.25	16.25	16.25	16.25	16.00	16.00	0.50	0.50	0.50	0.50	0.50	0.50	0.51	0.323	0.447	0.469	0.187	0.4777
22	14.00	14.00	14.00	14.00	14.00	14.00	13.50	0.50	0.50	0.50	0.50	0.50	0.43	0.272	0.378	0.396	0.187	0.4026
23	12.25	12.25	12.25	12.25	12.25	12.25	12.25	12.25	0.50	0.50	0.50	0.50	0.34	0.239	0.325	0.336	0.187	0.3396
24	11.00	11.00	11.00	11.00	11.00	11.00	11.00	11.00	10.50	0.50	0.50	0.50	0.26	0.213	0.279	0.289	0.187	0.2923
25	10.00	10.00	10.00	10.00	10.00	10.00	9.75	9.75	9.75	9.75	0.50	0.50	0.18	0.198	0.242	0.250	0.187	0.253
26	11.00	11.00	11.00	11.00	11.00	11.00	11.00	11.00	10.50	0.50	0.50	0.50	0.26	0.213	0.279	0.289	0.187	0.2923
27	10.00	10.00	10.00	10.00	10.00	10.00	9.75	9.75	9.75	9.75	0.50	0.50	0.18	0.198	0.242	0.250	0.187	0.253
28	9.25	9.25	9.00	9.00	9.00	9.00	9.00	9.00	9.00	9.00	9.00	0.50	0.09	0.189	0.212	0.217	0.187	0.2185
29	8.333	8.333	8.333	8.333	8.334	8.334	8.334	8.334	8.333	8.333	8.333	8.333	0.000	0.187	0.186	0.187	0.187	0.1867

Referências Bibliográficas

Allen, Arnold O. *Probability, Statistics and Queueing Theory,with Computer Science Applications*. Academic Press, second edition, 1990.

Banawan, S. A. e Zahorjan, J. Load sharing in hererogeneous queueing systems. *IEEE*, 1989.

Bellman, R. E. The theory of dynamic programming. *Bull. Amer. Math. Soc.*, 60:503–516, 1954.

Bellman, R. E. *The Theory of Dynamic Programming*. Princeton University Press,Princeton,NJ., 1957.

Blanc, J. P. C. A note on waiting times in systems with queues in parallel. *Applied Probabilities*, 24:540–546, 1987.

Blanc, J. P. C. The power-series algorithm applied to the shortest-queue model. *Operations Research*, 40:157–167, 1992.

Borst, S. C. Optimal probabilistic allocation of customer types to servers. *CWI Report BS-R94*, 1990.

Boxma, O. J., Deng, Q., e Zwart, A. P. Waiting-time asymptotics for the $m/g/2$ queue with heterogeneous servers. *Queueing Systems*, 40, 2002.

Chao, Xiuli e Luh, Hsing Paul. A stochastic directional convexity result and its application in comparison of queues. *Queueing Systems*, 48:399–419, 2004.

Cooper, R. B. *Introduction to Queueing Theory*. North Holland, New York - Oxford, second edition, 1981.

Daigle, John N. *Queueing Theory With Applications To Packet Telecommunication*. Springer Science + Business Media, Inc., 2005.

Disney, R. L. e Mitchell, W. E. A solution for queues with instantaneous jockeying and other customer selection rules. *Naval Research Logistics*, 17:315–325, 1971.

Eick, S. G., Massey, W. A., e Whitt, W. Nonstationarity in offered traffic to the att long distance network. *ATT Joint Symposium on Performance Analysis and Teletraffic Restoration Theory and Application,*, 1990.

Eick, S. G., Massey, W. A., e Whitt, W. $mt/g/\infty$ queues with sinusoidal arrival rates,. *Management Science*, 39:241–252, 1993.

El-Taha, M. e Stidham, S. Jr. *Sample-Path Analysis of Queueing Systems*. Kluwer Academic Publishers, 1999.

Elsyed, E. e Bastani, A. General solutions of the jockeying problem. *European Journal of Operations Research*, 22:387–396, 1985.

Feller, W. *An Introduction to Probability Theory and Its Applications.*, volume I. Wiley, New York., (3rd),1968 edition, 1957.

Feller, W. *An Introduction to Probability Theory and Its Applications.*, volume II. Wiley, New York., 1966.

Fond, S. e Ross, S.M. A heterogeneous arrival and service queueing lossmodel. *Naval Res. Logistics*, 25:483–488, 1978.

Foo, Justin e Mercankosk, Guven. Waiting time for packet-based networks with slow and fast servers. *Asia-Pacific Conference on Communications, IEEE*, 2005.

Foss, Serguei e Kovalevskii, Artyom. A stability criterion via fluid limits and its application to a polling system. *Queueing Systems*, 32:131–168, 1999.

Fulton, C. A. e Li, S. Delay jitter first-order and second-order statistical functions of general traffic on high-speed multimedia networks. *IEEE/ACM Transactions on Networking*, 6, 1998.

Gall, Pierre Le. The stationary $g/g/s$ queue. *Journal of Applied Mathematics and Stochastic Analysis*, 11:1:59–71, 1998a.

Gall, Pierre Le. The stationary g/g/s queue with non-identical servers. *Journal of Applied Mathematics and Stochastic Analysis*, 11:2:163–178, 1998b.

Gogate, N. R. e Panwar, S. S. Assigning customers to two parallel servers with resequencing. *IEEE Communications Letters*, 3, 1999.

Grassmann, K. W. e Zhao, Q. Y. Heterogeneous multiserver queues with general input. Technical report, University of Winnipeg, 2004.

Green, L. e Kolesar, P. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Science*, 37:84–97, 1991.

Gross, D. e Harris, M. H. *Fundamentals of Queueing Theory*. Wiley Series in Probability and Mathematical Statistics, second edition, 1985.

Gumbel, H. Waiting lines with heterogeneous servers. *Operations Research*, 8(4):504–511, 1960.

Haight, F. A. Two queues in parallel. *Biometrika*, 45:401–410, 1958.

Harten, A. V. e Sleptchenko, A. On markovian multi-class, multi-server queueing. *Queueing Systems*, 43:307–328, 2003.

Kingman, J. F. C. Two similar queues in parallel. *Ann. Math. Stat.*, 2:1314–1323, 1961.

Kleinrock, L. *Queueing Systems Vol. I.* Wiley and Sons, New York, NY, 1976a.

Kleinrock, L. *Queueing Systems Vol. II.* Wiley and Sons, New York, NY, 1976b.

Koenigsberg, E. On jockeying in queues. *Management Science*, 12:412–436, 1966.

Koole, G. A simple proof of the optimality of a threshold policy in two server queueing system. *Systems and Control Lett.*, 26:301–303, 1995.

Koole, G. e Mandelbaum, A. Queueing models of call centers an introduction. Technical report, faculteit der exacte wetenschappen, 2001.

Lee, H. Simultaneous determination of capacities and load in parallel $m/m/1$ queues. *European Journal of Operational Research*, 73:95–102, 1994.

Leung, Yiu-Wing e Hou, Ricky Yuen-Tan. Assignment of movies to heterogeneous video servers. *IEEE Transactions on Systems, Man and Cybernetics*, 35, 2005.

Levine, A. e Finkel, D. Load balancing in a multi-server queuing system. *Computers and Operations Research*, 17(1):17–25, 1990.

Lin, W. e Kumar, P. R. Optimal control of a queueing system with two heterogeneous servers. *IEEE Transaction Automat. Control*, 29:696–703, 1984.

Little, J. D. C. A proof for the queueing formula: L=λw. *Operations Research*, 9, 1961.

Liu, Zhen, Squillante, Mark S., e Wolf, Joel L. On maximizing service-level-agreement profits. *ACM*, 2001.

Marmony, M. Dynamic routing in large-scale service systems with heterogeneous servers. *Queueing Systems*, 51, 2005.

Melnick, E., Pinedo, M., Nayyar, P., e Seshardi, S. *Creating Value in the Financial Services, Strategies, Operations and Technologies*. Kluwer Academic Pres, 1999.

Neuts, M. F. e Takahashi, Y. Asymptotic behavior of the stationary distributions in the $gi/ph/c$ queue with heterogeneous servers. *Wahrscheinlichkeitstheorie verw*, 57: 441–452, 1981.

Nobel, R. D. e Tijms, H. C. Optimal control of a queueing system with heterogeneous servers and setup costs. *IEEE Transactions on Automatic Control*, 45, 2000.

Pinedo, Michael, Seshadri, Sridhar, e Shanthikumar, J. George. *Call Centers in Financial Services:Strategies, Technologies, and Operations*, chapter 18. Kluwer Academic Pres, 2003.

Qian, Y., Tipper, D., e Medhi, D. A nonstationary analysis of bandwidth access control schemes for heterogeneous traffic in b-isdn. *IEEE*, 1996.

Righter, Rhonda. Expulsion and scheduling control for multiclass queues with heterogeneous servers. *Queueing Systems*, 34:289300, 2000.

Ross, M. S. *Introduction to Probability Models*. Academic Press, Inc., sixth edition, 1993.

Rykov, V. e Efrosinin, D. Optimal control of queueing systems with heterogeneous servers. *Queueing Systems*, 46, 2004.

Rykov, V. V. Controllable queueing systems. *Itogi Nauki i Techniki. Teoria Verojatn. Matem. Statist. Teoretich. Kibern*, 12:45–152, 1975. in Russian(English translation in J. Soviet Math.).

Rykov, V. V. Monotone control of queueing systems with heterogeneous servers. *Queueing Systems*, 37, 2001.

Saaty, T. *Elements of Queueing Theory with Applications*. Dover Publications, New York, NY, 1961.

Schellhorn, Henri e Cossin, Didier. Credit risk in a network economy. *FAME - International Center for Financial Asset Management and Engineering*, 106, 2004.

Schriber, T. J. *An Introduction to Simulation Using GPSS/H*. John wiley Sons, Inc, 1991.

Schwartz, B. L. Queueing models with lane selection. *Operations Research*, 22:331–339, 1974.

Shenker, Shenker e Weinrib, Abel. Asymptotic analysis of large heterogeneous queueing systems. *ACM*, 1988.

Shimkin, Nahum e Mandelbaum, Avishai. Rational abandonment from tele-queues: Nonlinear waiting costs with heterogeneous preferences. *Queueing Systems*, 47:117–146, 2004.

Singh, V. Two-server markovian queues with balking: Heterogeneous vs. homogeneous servers. *Operations Research*, 18(1):145–159, 1970.

Singh, Vijendra P. Markovian queues with three heterogeneous servers. *AIIE Transaction*, 3:46–48, 1971.

Stidham, S. Jr. A last word on l=λw. *Operations Research*, 22:417–421, 1974.

Stidham, S. Jr. Optimal control of admission to a queueing system,. *IEEE Trans. Automat. Control*, 30:705–713, 1985.

Stidham, S. Jr. e Weber, R. A survey of markov decision models for control of networks of queues. *Queueing Systems*, 13:291–314, 1993.

Trajkovit, L. e Halfin, S. Buffer requirements in atm networks with leaky buckets. *IEEE*, 1994.

Viniotis, I. e Ephremides, A. Extension of the optimality of the threshold policy in heterogeneous multiserver queueing systems. *IEEE Transactions on Automatic Control*, 33, 1988.

Walrand, J. A note on optimal control of a queueing system with two heterogeneous servers. *Syst. Contr. Lett.*, 4:131134, 1984.

Wein, L. M. Due-date setting and priority sequencing in a multiclass $m/g/1$ queue. *Management Science*, 37(7):834–850, 1991.

Whitt, W. The effect of variability in the $gi/g/s$ queue. *Journal of Applied Probability*, 17:1062–1071, 1980.

Whitt, W. Understanding the efficiency of multi-server service systems. *Management Science*, 38(5):708–723, 1992.

Wolff, R. W. *Stochastic Modeling and Theory of Queues.* Prentice Hall, 1989.

Wong, T. C., Mark, J. W., e Chua, K. C. Access and control in a cellular wireless atm network. *IEEE*, 2000.

Zhao, Y. e Grasmann, W. K. The shortest queue model with jockeying. *Naval Research Logistics*, 37:773–787, 1990.