Universidade Federal do Paraná

Departamento de Informática

Marco Antonio Jonack
Cristina Duarte Murta

Avaliação do Impacto da Variabilidade da Carga no Desempenho dos Sistemas de Fila

Relatório Técnico
RT-DINF 001/2003

Curitiba, PR
2003

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Durante muitos anos a carga dos sistemas computacionais foi representada e modelada pelas distribuições exponencial e de Poisson. No entanto, durante a segunda metade da década passada, vários resultados publicados sobre caracterização de carga de sistemas computacionais mostraram que estas cargas apresentavam grande variabilidade, que não poderia ser modelada pelas distribuições citadas. Distribuições que apresentam grande variabilidade como, por exemplo, a distribuição de Pareto, se mostraram mais adequadas para a representação. Sabe-se da teoria de filas que a variabilidade nos parâmetros das cargas interfere negativamente nas métricas de desempenho dos sistemas. Esta monografia apresenta um estudo do impacto da variabilidade das cargas no desempenho dos sistemas computacionais. O estudo é feito através de simulação de sistemas de filas. O objetivo é avaliar quantitativamente o impacto de diferentes graus de variabilidade no desempenho dos sistemas. Os resultados mostram que a variabilidade é prejudicial às métricas de desempenho do sistema, principalmente se a variabilidade for inserida no parâmetro que representa o tempo de serviço da carga. Também foi constatado que aumentos na variabilidade dos tempos de serviço causam maior impacto ao desempenho do sistema do que variações na intensidade do tráfego a que este sistema está submetido.

CAPÍTULO 1

INTRODUÇÃO

O desempenho de um sistema computacional depende fundamentalmente da carga submetida a ele. Os projetistas de sistemas devem, portanto, fazer estudos detalhados das cargas de trabalho dos sistemas que são objeto dos seus projetos. Em geral, estas cargas são analisadas estatisticamente e são feitos modelos para sua descrição. Nestes modelos, cada característica da carga é representada por uma função de distribuição de probabilidade. A representação de cargas através de modelos estatísticos apresenta várias vantagens, como a facilidade de modificação dos parâmetros que determinam o comportamento da carga e a portabilidade devida à concisão do modelo [Jain, 1991].

Durante muito tempo a carga dos sistemas computacionais em geral foi representada pelas distribuições exponencial ou de Poisson [Greiner et al., 1995, Jain, 1991]. A maioria absoluta dos trabalhos de avaliação de desempenho que utilizavam modelos de carga baseavam-se nestas distribuições para representar a carga. A partir de 1994, vários trabalhos apresentaram novas caracterizações de carga, realizadas em vários ambientes, que mostraram que as cargas dos sistemas analisados seriam melhor representadas por distribuições de cauda pesada. Dentre as cargas com estas características estão o tráfego Ethernet [Leland et al., 1994, Greiner et al., 1995], os tamanhos e os tempos de transferência dos arquivos da Web [Crovella and Bestavros, 1996, Crovella et al., 1996] e os tamanhos dos processos em sistemas Unix [Harchol-Balter and Downey, 1996].

Distribuições de cauda pesada compõem uma classe ampla de distribuições denominadas em geral de *heavy-*, *fat-*, ou *long-tailed*, e são caracterizadas por seu comportamento para valores grandes de x, onde x é uma observação de uma variável aleatória X. Nas distribuições de cauda pesada, a probabilidade de ocorrer observações grandes (valores grandes de x) decai lentamente para 0, muito mais lentamente do que qualquer distribuição exponencial. Devido à probabilidade não nula de ocorrer observações muito grandes, a variabilidade observada nestas distribuições é extremamente alta. Em contraste, distribuições tais como a exponencial e a normal são denominadas distribuições de cauda leve.

Este trabalho apresenta um estudo sobre o impacto da variabilidade observada nas cargas no desempenho dos sistemas computacionais. O objetivo do trabalho é avaliar quantitativamente este impacto. Para isso, é utilizada a distribuição exponencial para representar uma carga pouco variável, e uma distribuição de cauda pesada, a distribuição de Pareto, para representar uma carga com grande variabilidade. As alterações no de-

sempenho do sistema são avaliadas separadamente para os dois parâmetros da carga, a saber, variabilidade no processo de chegada e nos tempos de serviço.

A avaliação proposta é realizada através de simulação de sistemas de filas e validada por resultados da teoria de filas [Jain, 1991, Molloy, 1989].

Conhecer o impacto da variabilidade das cargas atualmente experimentadas pelos sistemas computacionais auxilia o entendimento do comportamento do desempenho destes sistemas e incentiva a proposta de novos projetos de sistemas. Esta é a principal motivação para este trabalho.

O restante deste trabalho é organizado como descrito a seguir. O Capítulo 2 apresenta conceitos de modelagem de desempenho de sistemas computacionais e alguns trabalhos relacionados. No Capítulo 3 é descrito o projeto dos experimentos de simulação e uma breve discussão sobre os resultados esperados. O Capítulo 4 detalha o projeto e implementação do simulador utilizado neste trabalho e o Capítulo 5 analisa e mostra os resultados obtidos. Finalmente, o Capítulo 6 conclui o trabalho e apresenta algumas propostas para trabalhos futuros.

CAPÍTULO 2

CONCEITOS E TRABALHOS RELACIONADOS

Este capítulo apresenta os conceitos estatísticos e os conceitos da teoria de filas necessários ao entendimento deste trabalho, bem como alguns artigos que discutem questões relacionadas a variabilidade em modelos de filas.

2.1 Distribuições de Probabilidade e Variabilidade

Um dos conceitos fundamentais relacionados às distribuições de probabilidade é o conceito de *variável aleatória*. Uma variável aleatória é uma função que mapeia cada possível observação ocorrida em um espaço amostral a um número real [Papoulis, 1991]. Logo, é impreciso afirmar que uma variável aleatória é igual a um valor, mas sim que a função que a representa aplicada à observação é igual a um valor.

Dada a definição acima, a densidade de probabilidade (ou taxa de variação da função de distribuição) de uma variável aleatória é comumente chamada de *distribuição de probabilidade*. A caracterização definitiva de uma distribuição de probabilidade é realizada a partir da equação ou conjunto de equações que a definem e que determinam a sua representação gráfica.

Dependendo da natureza de uma distribuição, a variabilidade observada pode ser muito alta. Neste sentido, distribuições que apresentam uma variabilidade considerável e cuja cauda declina como uma função de potência são chamadas de distribuições de cauda pesada (*heavy-tailed*) [Crovella and Bestavros, 1996].

Uma distribuição é considera de cauda pesada se

$$P[X > x] \sim x^{-\alpha}, \quad x \to \infty, \quad 0 < \alpha < 2.$$

Isto significa que a probabilidade da variável aleatória X ter observações maiores que x é proporcional à razão $1/x^{\alpha}$, onde α caracteriza, de maneira inversa, o grau de variabilidade da distribuição, isto é, quanto menor o valor de α, maior a variabilidade.

A relação acima também expressa que, independente do comportamento da distribuição para valores pequenos de sua variável aleatória, se a curva assintótica da distribuição é hiperbólica, então ela é de cauda pesada, e a probabilidade de observações extremamente grandes é não desprezível [Crovella et al., 1996]. Em particular, X tem variância infinita e, se $\alpha \leq 1$, X tem média infinita, o que reflete sua grande variabilidade.

A distribuição de cauda pesada mais conhecida é a de *Pareto*. A distribuição de Pareto é muito utilizada na modelagem de fenômenos no ambiente Web e de propriedades da topologia da Internet, os quais são caracterizados por uma variabilidade extrema [Murta, 1999, Menascé and Almeida, 2002]. Pareto possui a seguinte função de distribuição de probabilidade (pdf):

$$f(x) = P[X = x] = \alpha k^{\alpha} x^{-\alpha-1}, \quad \alpha, k > 0, \quad x \geq k;$$

e função de distribuição cumulativa (cdf) correspondente a

$$F(x) = P[X \leq x] = 1 - (k/x)^{\alpha},$$

onde k representa o menor valor, ou limite inferior, da variável aleatória.

Em oposição às distribuições de cauda pesada, onde a cauda declina à direita com uma função de potência (isto é, mais lentamente), estão as distribuições mais comuns e conhecidas na modelagem de fenômenos, tais como normal, exponencial e Poisson, cuja cauda declina exponencialmente, apresentando uma variabilidade menor. A Figura 2.1 faz uma comparação entre as caudas das distribuições exponencial e Pareto, onde a cauda é o complemento da função $F(x)$, isto é, $\overline{F}(x) = 1 - F(x) = P[X > x]$.

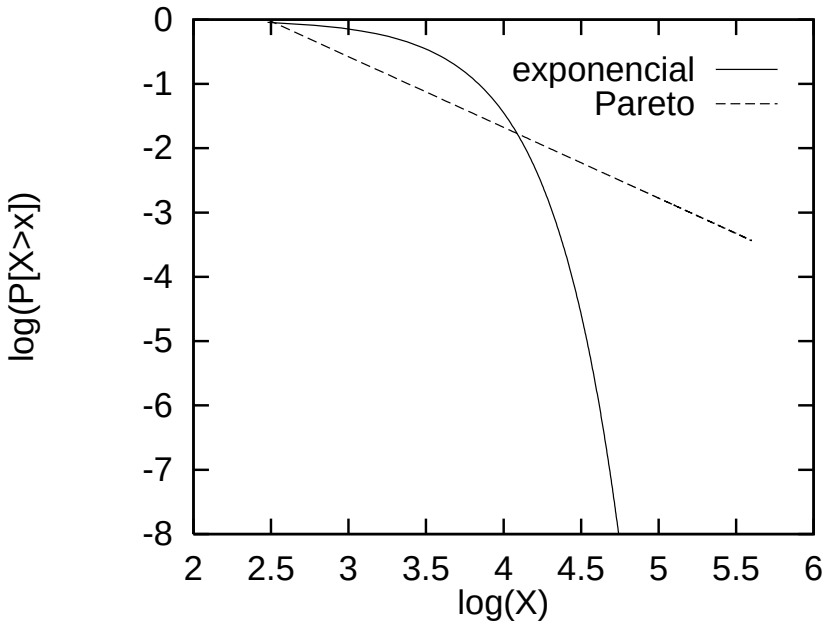

Figura 2.1: Visualização de cauda para as distribuições exponencial e Pareto.

A Tabela 2.1 compara as probabilidades das distribuições exponencial e Pareto com

mesma média (3000) para instâncias da variável aleatória X. Enquanto a probabilidade dos valores da distribuição exponencial se concentra em torno da média e decai rapidamente para valores grandes de X, a probabilidade dos valores da distribuição de Pareto apresenta um decréscimo relativamente pequeno, até mesmo para os maiores valores de X. Isto demonstra a propriedade das distribuições de cauda pesada de gerar números grandes com probabilidades não desprezíveis.

	$P[X > x]$	
x	Exponencial	Pareto ($\alpha = 1.1$ e $k = 300$)
300	0.90	1.00
600	0.82	0.47
1500	0.61	0.17
3000	0.37	0.079
6000	0.135	0.037
10000	0.036	0.021
50000	10^{-8}	0.0036
100000	10^{-17}	0.0017

Tabela 2.1: $P[X > x]$ para as distribuições exponencial e de Pareto com média 3000.

Outra diferença entre as distribuições de cauda pesada com relação às de cauda leve é a necessidade de grandes conjuntos de observações para a convergência dos valores estatísticos, fato que, dependendo da grau de variabilidade da distribuição, pode efetivamente não acontecer ou ser posto em dúvida [Crovella and Lipsky, 1997].

2.2 Teoria de Filas

A teoria de filas é uma área da teoria de probabilidades aplicada à avaliação de desempenho de sistemas computacionais [Jain, 1991]. Modelos analíticos provêem uma maneira compacta e flexível para a representação e estudo de sistemas reais.

A abstração de sistemas na teoria de filas é feita através dos chamados *sistemas de filas*. Entre os elementos mais importantes em um sistema de filas estão os *clientes*, que são os elementos que ocupam os recursos do sistema, e os *servidores*, que executam as tarefas requisitadas pelos clientes[1].

Um sistema de filas pode ser representado por seis parâmetros. A notação utili-

[1]Muitas vezes, dependendo do sistema computacional que está sendo modelado, o termo cliente da teoria de filas é referenciado como processo, tarefa ou *job*. Da mesma forma, o termo servidor é usado como recurso ou *device*. Este trabalho usa indistintamente os termos cliente/tarefa e servidor/recurso.

zada para representar estes parâmetros é chamada de *notação Kendall* e possui a forma $A/S/m/B/K/D$, onde as letras têm os seguintes significados:

A Distribuição do tempo entre as chegadas de clientes ao sistema;

S Distribuição do tempo de serviço;

m Número de servidores;

B Número de *buffers* (ou capacidade do sistema);

K Tamanho da população;

D Disciplina da fila.

As distribuições dos tempos entre as chegadas e dos tempos de serviço são freqüentemente denotados por símbolos de uma letra. Assim, M é usado para representar a distribuição exponencial, E_k denota a distribuição de Erlang com parâmetro k, H_k a distribuição hiperexponencial com parâmetro k, D é utilizado para distribuições determinísticas e G para representar um distribuição genérica.

A distribuição exponencial é denotada por M, devido a sua propriedade *memoryless*, que é a propriedade das observações correntes não serem afetadas por estados anteriores. Uma distribuição determinística é aquela em que os tempos são constantes e a variância é nula. A distribuição geral representa uma distribuição não especificada onde seus resultados são válidos para todas as distribuições.

Por exemplo, a notação $M/M/3/20/1500/FCFS$ denota um sistema de fila única com os seguintes parâmetros:

1. O tempo entre duas chegadas sucessivas é exponencialmente distribuído.
2. Os tempos de serviço são exponencialmente distribuídos.
3. O sistema possui três servidores.
4. A fila tem *buffers* para 20 clientes, três em atendimento pelos servidores e 17 em espera na fila.
5. Há um total de 1500 clientes que podem ser atendidos, que representa o tamanho da população.
6. A disciplina da fila é *First Come First Served* (FCFS).

Em geral, as filas são definidas tendo a capacidade do *buffer* infinita, tamanho da população infinito e FCFS como disciplina da fila. Dessa forma apenas os primeiros três

parâmetros são suficientes para indicar o tipo da fila. Assim a fila $M/M/1/\infty/\infty/FCFS$ é denotada como $M/M/1$.

A Figura 2.2 mostra o diagrama de um sistema de fila única e suas respectivas variáveis. Tais variáveis podem ser divididas em duas categorias: parâmetros do sistema e métricas.

Os parâmetros do sistemas são compostos por:

$\tau =$ tempo entre duas chegadas consecutivas.

$\lambda =$ taxa média de chegada, sendo igual a $\frac{1}{E[\tau]}$.

$s =$ tempo de serviço por cliente.

$\mu =$ taxa média de serviço por servidor, dada por $\frac{1}{E[s]}$.

$m =$ número de servidores.

As métricas do sistema são formadas por:

$n =$ número de clientes no sistema.

$n_q =$ número de clientes em espera ou tamanho da fila.

$n_s =$ número de clientes em atendimento nos servidores.

$r =$ tempo de resposta ou tempo no sistema.

$w =$ tempo de espera ou tempo de fila.

Uma métrica não definida acima, mas de grande importância para a análise de um sistema de filas, é a chamada *slowdown*. O *slowdown* de uma tarefa é expresso pela razão do seu tempo de resposta pelo seu tempo de serviço. Esta métrica mede o tempo de espera de uma tarefa em relação ao seu tempo de serviço, sendo importante na avaliação de "injustiças" em políticas de escalonamento [Bansal and Harchol-Balter, 2001]. Por exemplo, se todas as tarefas de um sistema experimentam valores pequenos e pouca variância no *slowdown*, então a política de escalonamento do sistema é considerada "justa". Em outras palavras, é considerado justo que uma tarefas com um grande tempo de serviço espere mais para ser executada do que uma tarefa com um pequeno tempo de serviço.

A *carga* ou *intensidade do tráfego* de um sistema de filas, denotada por ρ, é calculada como $\rho = \lambda/m\mu$. Também é definida a *condição de estabilidade* do sistema de filas. Se o número de tarefas em um sistema cresce continuamente e torna-se infinito, o sistema é dito instável. Para o sistema se manter estável, a taxa média de chegada deve ser menor que a taxa média de serviço, isto é, $\lambda < m\mu$. Dessa forma, pode-se afirmar que o sistema permanecerá estável enquanto $\rho < 1$.

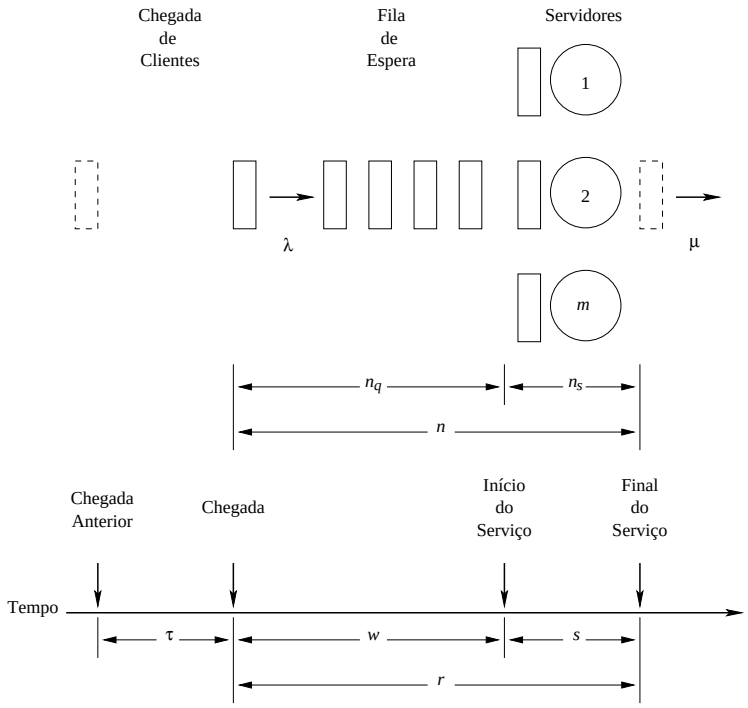

Figura 2.2: Diagrama de um sistema de fila única com m servidores e suas variáveis.

Para um sistema que cumpre a condição de estabilidade, a teoria de filas provê um conjunto de resultados e regras observadas experimentalmente que são válidos para todas as filas. Estas postulações são chamadas de *leis operacionais* e estabelecem diversos relacionamentos entre as variáveis do sistema.

Uma das regras mais conhecidas dentre as leis operacionais é a chamada *lei de Little*, que relaciona o número médio de tarefas no sistema com o tempo médio de resposta. Se as tarefas não são perdidas devido a falta de *buffers* no sistema, então

$$E[n] = \lambda E[r],$$

e, da mesma forma,

$$E[n_q] = \lambda E[w].$$

Em sistemas com número finito de *buffers*, esta lei fornece a taxa de chegada efetiva, isto é, a taxa em que as tarefas realmente são aceitas pelo sistema.

Outros resultados falam a respeito do número de tarefas e do tempo de resposta experimentados no sistema, são eles:

$$E[n] = E[n_q] + E[n_s],$$

e

$$E[r] = E[w] + E[s].$$

Em conjunto com as leis operacionais, várias outras fórmulas, particulares a determinado tipo de fila, são utilizadas na avaliação de um sistema modelado pela teoria da filas. A Tabela 2.2 mostra o conjunto de fórmulas da média e variância para uma fila $M/M/1$.

Métrica	Média $E[X]$	Variância $Var[X]$
Clientes no sistema (n)	$\rho/(1-\rho)$	$\rho/(1-\rho)^2$
Tamanho da fila (n_q)	$\rho^2/(1-\rho)$	$\rho^2(1+\rho-\rho^2)/(1-\rho)^2$
Tempo de resposta (r)	$(1/\mu)/(1-\rho)$	$(1/\mu^2)/(1-\rho)^2$
Tempo de fila (w)	$\rho(1/\mu)/(1-\rho)$	$(2-\rho)\rho/[\mu^2(1-\rho)^2]$

Tabela 2.2: Fórmulas da média e variância para uma fila $M/M/1$.

Da mesma forma, a fila $M/G/1$ possui seu próprio conjunto de fórmulas para a avaliação do sistema. Tais fórmulas são exibidas na Tabela 2.3, onde C_s denota o coeficiente de variação do tempo de serviço, dado por $\frac{\sqrt{Var[s]}}{[E[s]}$ e seu quadrado por $\frac{Var[s]}{(E[s])^2}$.

Métrica	Média $E[X]$	Variância $Var[X]$
Clientes no sistema (n)	$\rho + \frac{\rho^2(1+C_s^2)}{2(1-\rho)}$	$E[n] + \lambda^2 Var[s] + \frac{\lambda^3 E[s^3]}{3(1-\rho)} + \frac{\lambda^4 (E[s^2])^2}{4(1-\rho)^4}$
Tamanho da fila (n_q)	$\rho^2(1+C_s^2)/[2(1-\rho)]$	$Var[n] - \rho + \rho^2$
Tempo de resposta (r)	$E[s] + \frac{\rho E[s](1+C_s^2)}{2(1-\rho)}$	$Var[s] + \frac{\lambda E[s^3]}{3(1-\rho)} + \frac{\lambda^2 (E[s^2])^2}{4(1-\rho)^2}$
Tempo de fila (w)	$\rho E[s](1+C_s^2)/[2(1-\rho)]$	$Var[r] - Var[s]$

Tabela 2.3: Fórmulas da média e variância para uma fila $M/G/1$.

Apesar da teoria de filas prover uma maneira relativamente rápida e barata para a avaliação de desempenho de sistemas computacionais, ela não dispõe de resultados para

uma ampla gama de situações. Além disso, tais resultados não detém o nível de precisão muitas vezes exigido para análise de um sistema. Logo, duas ou mais formas de avaliação podem ser utilizadas simultaneamente [Jain, 1991, Menascé and Almeida, 2002], como por exemplo simulação ou experimentação, tendo os seus resultados comparados a fim de obter a correção e uma maior confiabilidade dos mesmos.

2.3 Trabalhos Relacionados

Um dos principais problemas em simular sistemas que utilizam distribuições de cauda pesada é a dificuldade de alcançar a estabilidade dos resultados durante a simulação, reflexo da grande variabilidade observada neste tipo de distribuição. Esta característica exige um grande número de ciclos de geração de amostras para a convergência dos valores estatísticos e obtenção de um estado de equilíbrio, aumentando consideravelmente o tempo das simulações ou até mesmo as tornando inviáveis.

Um estudo detalhado sobre as condições transientes de distribuições de cauda pesada em simulações é mostrado em [Crovella and Lipsky, 1997]. Neste trabalho, os autores alertam para a dificuldade de obter estabilidade em simulações baseadas em distribuições de cauda pesada. A dificuldade é devida ao fato de que, nestas distribuições, a probabilidade de ocorrer observações muito grandes não é desprezível, o que torna difícil a convergência da média e dos demais momentos da distribuição. Os autores mostram que estas simulações apresentam duas características de destaque. A primeira refere-se à convergência lenta da simulação para um estado estacionário. A segunda refere-se à grande variabilidade observada neste estado. A principal implicação destes resultados é quanto ao tamanho da simulação, em termos do número de observações geradas e tratadas no sistema. Simulações que utilizam distribuições de cauda pesada são muito mais longas do que simulações tradicionais.

O impacto da variabilidade do tempo de serviço no desempenho de sistemas de filas $M/G/1$ foi avaliado em [Heyman, 2000] através de modelagem analítica. O autor mostra que, quando as variâncias das distribuições são muito grandes, embora finitas, o efeito sobre o desempenho é similar à situação de variância infinita. O autor mostra também que, na situação de grande variância dos tempos de serviço, a convergência para o estado estacionário é tão lenta que provavelmente as medidas de desempenho neste estado não são de interesse. O autor ressalta a necessidade de definir outros critérios de desempenho que não tenham como assunção o estado estacionário. Em outras palavras, isto significa que os sistemas devem, na prática, trabalhar com a hipótese de estarem sempre em condições transientes, pelo menos em escalas de tempo significativas em relação à sua carga média de trabalho.

CAPÍTULO 3

PROJETO DOS EXPERIMENTOS

3.1 Objetivos

Os experimentos realizados têm o objetivo de avaliar o impacto da variabilidade dos parâmetros da carga sobre as métricas de um sistema de fila única e recurso único. Assim, as métricas selecionadas para a análise do sistema são:

Tempo de resposta

Tamanho da fila

Tempo de fila

Slowdown

Uma consideração deve ser feita com relação as métricas tamanho da fila e tempo de fila. Neste trabalho, ambas as métricas consideram a tarefa em atendimento como integrante de seus valores. Assim, todos os cálculos feitos para obter os resultados destas métricas levam em conta a tarefa sendo atendida pelo servidor.

3.2 Experimentos

A técnica de avaliação de desempenho escolhida para o estudo do sistema proposto foi simulação, com os resultados validados pela teoria de filas. O impacto da variabilidade é analisado comparando-se os resultados das métricas do sistema para cargas com pequena e grande variabilidade.

As cargas com pouca variabilidade são modeladas pela distribuição exponencial, cuja representação pela teoria de filas é feita pela letra M. Já as cargas com grande variabilidade são modeladas pela distribuição de Pareto, identificada pela letra G. Ambas as distribuições são aplicadas sobre os parâmetros da carga (tempo entre chegadas e tempo de serviço), permitindo a avaliação particular de cada um destes.

Conforme a configuração das distribuições sobre os tempos entre chegada e de serviço, ou seja, dado o tipo de fila no formato $A/S/1$ (ver Seção 2.2), obtém-se diversos tipos de carga. Dessa forma, os seguintes tipos de fila foram simulados:

$M/M/1$ – carga pouco variável.

$M/G/1$ – carga com grande variabilidade no tempo de serviço.

$G/M/1$ – carga com grande variabilidade no tempo entre chegadas.

$G/G/1$ – carga com grande variabilidade nos dois parâmetros do sistema.

Os tipos de filas descritos acima utilizam $FCFS$ como política de escalonamento, com número de *buffers* e população infinitos. A política $FCFS$ é muito utilizada em sistemas de supercomputação, onde os trabalhos (*jobs*) são processados em lotes (*batch*). Além disso, o estudo de sistemas de filas com esta política é clássico na teoria de filas e ponto de referência para a avaliação dos demais sistemas. Outra razão para a escolha de $FCFS$ é quanto à validação dos resultados da simulação. A maior parte dos resultados analíticos refere-se às filas com política $FCFS$, o que é certamente um auxílio para a verificação dos resultados deste trabalho e sua credibilidade.

O número de chegadas tratado por cada simulação é de um bilhão de tarefas, isto é, a simulação se desenvolve até que um bilhão de tarefas tenham sido aceitas pelo sistema (ver Seção 5.1 para detalhes sobre a escolha deste valor). Em termos práticos, supondo um sistema computacional que atenda 100 requisições por segundo, tal sistema levaria aproximadamente nove meses para processar um bilhão de requisições. Já em um sistema que atenda 1000 requisições por segundo, seriam necessários 11 dias para completar todas as tarefas.

Para efeitos de comparação e coerência dos resultados das simulações, a taxa de serviço é fixada e apenas a taxa de chegada sofre variações. Assim, para cada tipo de fila, as simulações produzem resultados onde a intensidade do tráfego (ρ) varia de 0.1 a 0.9 pontos (respectivamente, a carga mínima e máxima do sistema). Isto implica que diferentes distribuições atribuídas ao processo de chegada (A) têm médias aproximadamente iguais para um dado ρ, diferindo apenas em relação às suas variâncias.

A Tabela 3.1 mostra as comparações feitas entre os resultados das simulações executadas.

Resultados Comparados		
$M/M/1$	*versus*	$M/G/1$
$M/M/1$	*versus*	$G/M/1$
$G/G/1$	*versus*	$M/G/1$
$G/G/1$	*versus*	$G/M/1$
$M/G/1$	*versus*	$G/M/1$

Tabela 3.1: Relação das comparações feitas entre os tipos de filas simulados.

Com relação à distribuição de Pareto, os valores do parâmetro α considerados para as filas que a utilizam são 1.3, 1.5, 1.7 e 1.9, sendo o primeiro destes o que produz maior variabilidade e, o último, menor variabilidade. Especialmente para simulações com a fila $G/G/1$, o mesmo valor de α é atribuído as distribuições referentes ao tempo entre chegadas (primeiro G) e ao tempo de serviço (segundo G).

Uma discussão mais detalhada sobre a escolha dos parâmetros utilizados nas simulações pode ser encontrada na Seção 5.1, e a relação completa de parâmetros no Apêndice B.

3.3 Resultados Esperados

Em linhas gerais, é esperado que o desempenho de um sistema submetido à cargas muito variáveis seja pior do que o desempenho do mesmo sistema submetido à cargas pouco variáveis. Este resultado é conhecido na teoria de filas para as filas $M/G/1$, em comparação com o modelo $M/M/1$.

As métricas de desempenho da fila $M/G/1$ estão em função da variância do tempo de serviço. No entanto, não são conhecidos resultados das métricas para os sistemas $G/M/1$ e $G/G/1$. Também não são conhecidos resultados analíticos para a métrica *slowdown*.

Tendo em vista estas questões, a abordagem adotada por este trabalho é a de realizar a simulação de um sistema com as diversas filas propostas ($M/M/1$, $M/G/1$, $G/M/1$ e $G/G/1$), validando os resultados da simulação com os resultados analíticos conhecidos para as filas $M/M/1$ e $M/G/1$, e com a lei de Little para as outras filas.

Os demais resultados da simulação poderão ajudar a responder algumas perguntas, tais como:

1. Qual é o impacto quantitativo da variabilidade nos tempos de serviço?
2. Qual é o impacto da variabilidade nos tempos entre chegadas?
3. Como se comporta o *slowdown* para os diferentes tipos de carga?
4. O que é mais prejudicial para o sistema avaliado: a variabilidade nos tempos de serviço ou no processo de chegada?

Como este trabalho, espera-se responder todas estas perguntas.

CAPÍTULO 4

O SIMULADOR DE SISTEMAS DE FILA ÚNICA

Dentre as características essenciais que um programa que realiza simulações de sistemas deve possuir estão a sua correção, isto é, o programa deve gerar resultados precisos e confiáveis, e a eficiência, para que simulações de sistemas de carga elevada sejam realizadas em tempos aceitáveis. De acordo com esta proposta, foi desenvolvido o *Single queue Simulator* (`SqS`), um simulador de fila simples baseado em eventos. Este capítulo apresenta uma descrição do projeto e da implementação do `SqS`.

4.1 Projeto do Simulador

O projeto do `SqS`, visando modelar um sistema de fila única genérico com um servidor, contempla os principais conceitos da teoria de filas e outros conceitos necessários à simulação, tais como o tempo da simulação[1], que é o valor do tempo observado na operação do sistema real que está sendo representado computacionalmente, e os eventos, que são ações que causam mudanças de estado no modelo do sistema.

O fluxo de controle implementado na estrutura básica do `SqS`, que representa um simulador baseado em eventos, é apresentando na Figura 4.1. A execução de um dos laços do simulador denota um evento ocorrido e, a cada volta do laço, o tempo da simulação é avançado para o momento em que o próximo evento irá acontecer.

O `SqS` trata explicitamente de dois eventos: *chegada de tarefa* e *finalização de tarefa*. O evento de chegada de tarefa acontece quando uma nova tarefa vinda do meio externo encontra o serviço ocupado e tem que ser enfileirada. O evento de finalização de tarefa ocorre quando o servidor termina a execução da tarefa corrente. O sistema considera que, ao ser finalizada uma tarefa, o atendimento da próxima tarefa (que pode ser a última tarefa gerada ou uma tarefa da fila de espera) é iniciado imediatamente, não havendo um evento formal para representar o acontecimento desta ação.

A identificação de um evento implica no cômputo de estatísticas para este evento. As estatísticas para a chegada de tarefa compreendem o acúmulo dos valores observados para as variáveis tempo entre chegadas, tempo de serviço, tempo de resposta, tamanho da fila, tempo de fila e *slowdown*. O cálculo do tamanho da fila e do tempo de fila levam em conta a tarefa sendo atendida pelo serviço. No momento da finalização de tarefa, apenas

[1] O tempo de simulação não deve ser confundido com o tempo de execução da simulação, que é o valor de tempo observado durante a execução do simulador. O tempo de execução também é chamado de *wall clock time*.

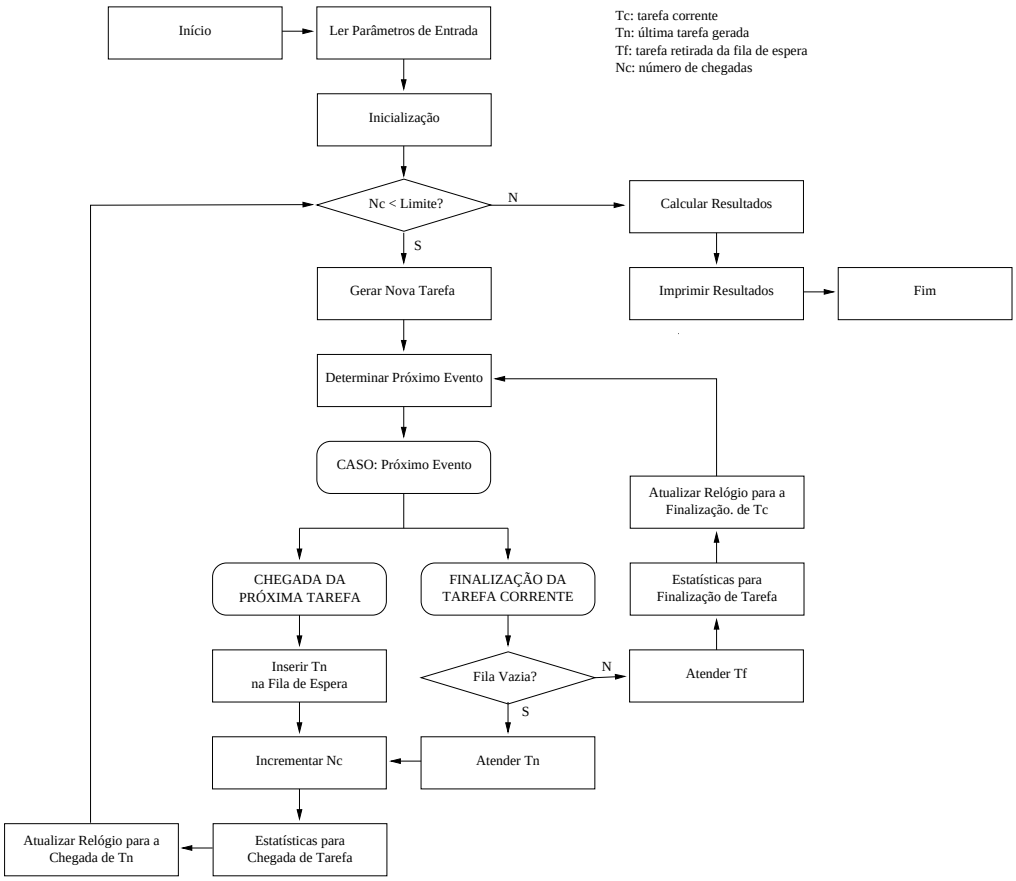

Figura 4.1: Fluxo básico de execução do `SqS`.

a contagem de tarefas finalizadas é feita.

Após o acontecimento de um evento, o tempo de simulação é avançado para o tempo da chegada da última tarefa gerada ou para o tempo da finalização da tarefa corrente, o que ocorrer primeiro.

Este procedimento continua até que um número pré-determinado de tarefas tenha chegado ao sistema. Quando este número é alcançado, o `SqS` realiza os cálculos finais (média, desvio padrão e variância) sobre as estatísticas coletadas durante a simulação e as exibe, juntamente com outros resultados e parâmetros do sistema, em um formato previamente escolhido, finalizando a execução do programa.

A próxima seção explica com todos esses elementos foram reunidos na implementação do `SqS`.

4.2 Implementação do Simulador

Com o objetivo de alcançar o quesito de eficiência, o `SqS` foi totalmente implementado sobre a linguagem *C++*, que provê suporte à programação baseada no paradigma orientado a objetos [Stroustrup, 1993, Coplien, 1994]. O código fonte completo do `SqS`, sobre as condições da licença GPL, pode ser obtido em [Jonack, 2002].

A passagem do diagrama mostrado na seção anterior para a implementação não é direta, apenas o fluxo é seguido rigorosamente. A implementação resultou no total de sete classes principais (entre outras auxiliares) que representam os conceitos da teoria de filas e controles necessários à execução do programa. Estas classes são:

- Tarefa (`Task`) – principal conceito do simulador. Um objeto tarefa engloba os atributos que representam o tempo decorrido desde a chegada da última tarefa no sistema (`elapsed_time`) e o tempo que a tarefa irá ocupar o serviço fornecido pelo sistema (`service_time`).
- Fluxo de Tarefas (`TaskStream`) – gerador de tarefas que simula a chegada aleatória de tarefas ao sistema. Possui funções de distribuição de probabilidade (exponencial e Pareto) necessárias à formação dos tempos entre chegadas e de serviço.
- Fila de Tarefas (`TaskQueue`) – armazena as tarefas geradas que estão aguardando serviço. A atual implementação provê FCFS e LCFS como disciplinas de escalonamento disponíveis.
- Serviço (`Service`) – recurso do sistema que oferece um serviço.
- Relógio da Simulação (`SimulationClock`) – computa o tempo total de simulação.
- Analisador de Eventos (`EventAnalyser`) – com base na nova tarefa gerada e na tarefa atualmente sendo atendida, determina qual será o próximo evento do sistema: a chegada da nova tarefa (`TASK_ARRIVAL`) ou a finalização da tarefa corrente (`TASK_TERMINATION`).
- Processador de Opções (`OptionParser`) – avalia as opções passados pela linha de comando ao `SqS`.

Como o processo de simulação do `SqS` requer a criação de várias tarefas, os objetos da classe tarefa são tratados dinamicamente durante a simulação através de ponteiros e funções de gerenciamento de memória (`new` e `delete`). Todas as outras classes listadas acima possuem uma única instância que as representam e não são declaradas como ponteiros.

Além dos tempos decorrido e de serviço, a classe tarefa possui dois outros importantes atributos (`arrv_time_remaining` e `serv_time_remaining`) que representam o tempo restante para a chegada e finalização da tarefa no sistema. Ambos recebem, no memento de sua inicialização, os valores do tempo decorrido e de serviço, respectivamente.

O primeiro deles controla o tempo restante para a tarefa chegar efetivamente ao sistema. Este é usado quando o momento de chegada da tarefa está num futuro distante o suficiente para que algumas ou até mesmo todas as tarefas existentes na fila do sistema sejam atendidas. Logo, à medida que as tarefas são atendidas, o tempo de restante para a tarefa chegar ao sistema é decrementado. O segundo atributo refere-se ao tempo de serviço restante da tarefa sendo atendida. Enquanto uma tarefa é tratada pelo serviço, várias outras tarefas podem chegar ao sistema, implicando na atualização do relógio da simulação para o momento da chegada destas tarefas. Cada vez que o relógio é incrementado, o tempo restante de serviço da tarefa também é atualizado.

Como descrito acima, os valores dos atributos tempo decorrido e de serviço de uma tarefa permanecem inalterados durante toda a existência da tarefa no sistema. Apenas os atributos que representam o tempo restante para a chegada e finalização da tarefa são atualizados, sendo estes utilizados pela classe analisador de eventos para determinar o próximo evento tratado pelo simulador.

Durante a execução da simulação, as tarefas são produzidas por demanda pela classe fluxo de tarefas. A Figura 4.2 mostra uma visão interna desta classe. Para gerar os valores dos atributos tempos decorrido e de serviço, esta classe utiliza dois "métodos virtuais"[2] (`generateElapsedTime` e `generateServiceTime`) ou ponteiros para funções que, dependendo do tipo de fila a ser simulado, irão apontar para uma das duas funções de distribuição de probabilidade (`exponencial` ou `pareto`) que efetivamente produzem os valores pseudo-aleatórios utilizados pela classe tarefa.

A implementação das funções de distribuição de probabilidade faz uso de uma rotina de geração de números pseudo-aleatórios com distribuição uniforme no intervalo $[0, 1]$, denotada por $U(seed)$, onde *seed* é a semente utilizada pela rotina no seu processo interno de geração de valores. No `SqS`, esta rotina é representada pela função `erand48` da biblioteca padrão C, e as distribuições exponencial e Pareto são codificadas, respectivamente, como mostrado abaixo:

$$\texttt{exponencial}(seed,\ t) = -\left(\frac{1}{t}\right) * \log(U(seed))$$

[2] O rótulo *método virtual* utilizado no `SqS` não se relaciona com o conceito de funções virtuais pertencente à linguagem *C++*, cuja semântica é totalmente diferente.

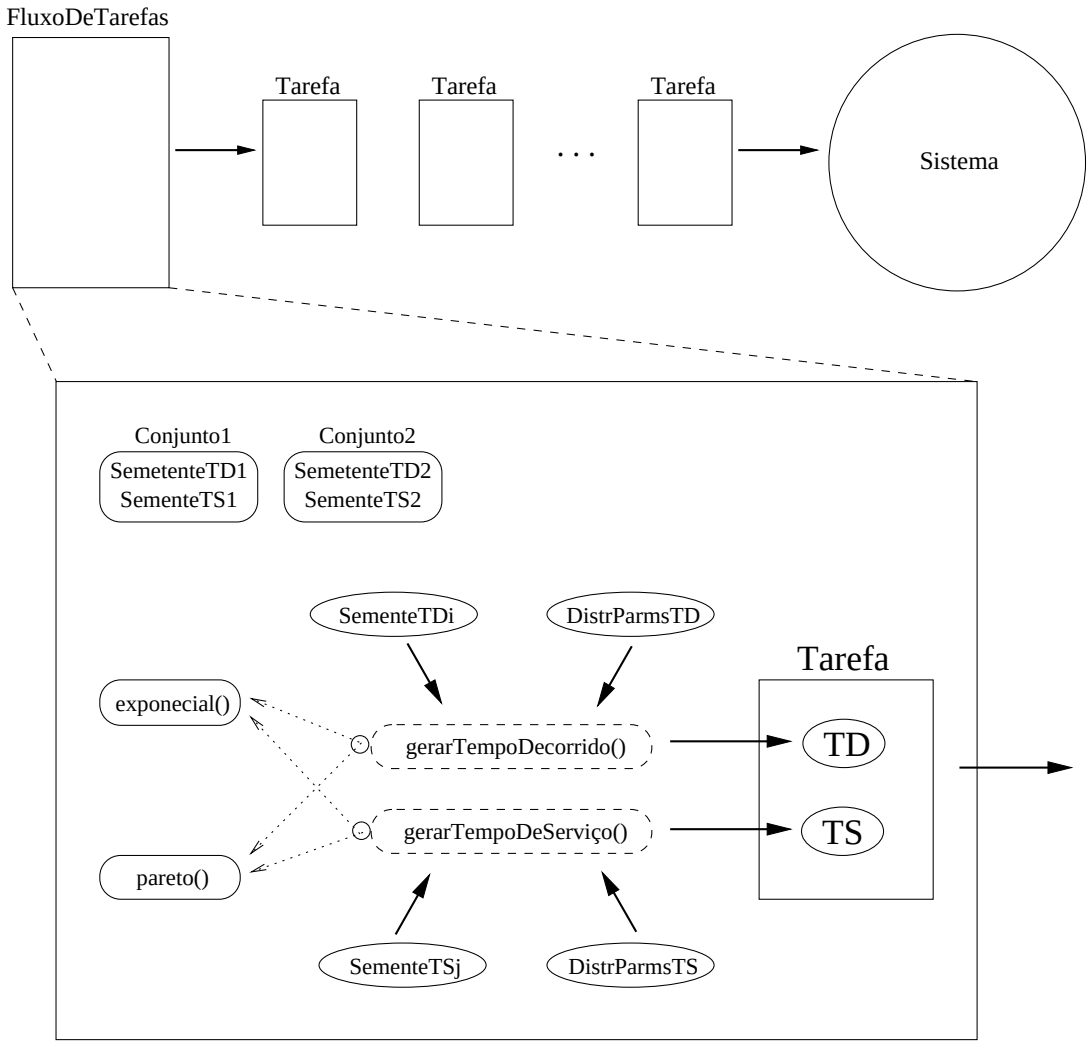

Figura 4.2: Visão interna da classe fluxo de tarefas.

$$\texttt{pareto}(seed,\ \alpha,\ min,\ max) = \frac{min}{\left(\left(\frac{min}{max}\right)^{\alpha} + U(seed) * \left(1 - \left(\frac{min}{max}\right)^{\alpha}\right)\right)^{\frac{1}{\alpha}}}$$

A função `exponencial` retorna o valor de uma variável pseudo-aleatória com distribuição exponencial negativa e média $1/t$, com t representado uma taxa de chegada λ ou uma taxa de serviço μ. Já a função `pareto` retorna o valor de uma variável pseudo-aleatória de acordo com a distribuição de Pareto, probabilidade $P[X > x] = (min/x)^{\alpha}$ e média $(min * \alpha)/(\alpha - 1)$, chamada de Pareto limitada, onde min e max são seus limites inferior e superior, respectivamente.

Para garantir que haja geração de seqüências distintas de números aleatórios, a

classe fluxo de tarefa implementa o conceito de conjunto de sementes, onde cada conjunto é formado por duas sementes que são passadas individualmente a cada um dos métodos de geração dos tempos decorrido e de serviço. A atual implementação dispõe de dois conjuntos de sementes (`SEEDS_SET1` e `SEEDS_SET2`), totalizando quatro sementes geradoras distintas.

A classe fila de tarefas é responsável pelo armazenamento e ordenação das tarefas que chegam ao sistema. A disposição das tarefas na fila de espera é determinada pela disciplina de fila escolhida, isto é, sempre que uma nova tarefa deve ser enfileirada, a função que representa a disciplina de fila decide a posição de inserção desta tarefa. O controle da disciplina de fila selecionada na classe fila de tarefas também utiliza o esquema de ponteiros para funções descrito anteriormente. Nesta classe, existe apenas um método de inserção genérico (`enqueueTask`) que irá apontar para uma das funções que implementam as disciplinas de fila.

Para a retirada de elementos da fila (método `dequeueTask`), a convenção adotada é que sempre o elemento apontado como o primeiro da fila (`first`) será desenfileirado. Dessa forma, a função que implementa uma disciplina de fila deve garantir que o primeiro elemento resultante da aplicação desta disciplina coincida com o apontador de primeiro elemento da fila.

A classe serviço representa um recurso genérico oferecido pelo sistema. Seu papel é receber um ponteiro para uma tarefa através do método `handleTask` e atualizar o tempo de serviço restante desta à medida que a simulação se desenvolve. A classe serviço também é responsável pela liberação do espaço de memória ocupado pela tarefa quanto esta é finalizada, oferecendo um método próprio para isto (`endTask`).

Devido a sua simplicidade e do fato de representarem conceitos globais ao sistema, todos os atributos e métodos das classes relógio da simulação e analisador de eventos são estáticos. Isto significa que não é preciso produzir instâncias destas classes para utilizar suas funcionalidades.

Como as opções de configuração do `SqS` são passadas via linha de comando, a classe processador de opção é responsável pela leitura e avaliação destas, utilizando para isto o método `parseOptions` que recebe os argumentos `argc` e `argv` do programa principal. Quando uma opção ou valor não válido são identificados, uma mensagem de erro é carregada no atributo `error_message` e um indicador de erros (`parse_error`) é ajustado como verdadeiro. Da mesma forma, quando a opção de ajuda é identificada, o indicador `help_solicited` é ajustado como verdadeiro.

Todos os atributos da classe processador de opções que representam as opções disponíveis são inicializados com valores padrão no próprio construtor da classe. Dessa forma, o programa principal solicita ao objeto desta classe os valores das opções para inicializar o sistema (método `getOptionValue`), apenas testando previamente a condição de erro nas

opções ou solicitação de impressão de ajuda, métodos `errorOccurred` e `helpSolicited` respectivamente. A mensagem de erro pode ser obtida pelo método `getErrorMessage` e a relação completa de opções para impressão na saída padrão pelo método `printUsage`.

Quando os objetos destas classes interagem para realizar a simulação do sistema, o programa principal é o responsável pela coleta dos dados para o computo das estatísticas e também pela impressão dos resultados.

A relação completa de opções disponíveis e resultados impressos pelo `SqS` são detalhados na próxima seção.

4.3 Entradas e Saídas do Simulador

A sintaxe de chamada do simulador é a seguinte:

*$ sqs [lista-de-opçõe*s],

onde *lista-de-opções* é formada por zero ou mais opções disponíveis no `SqS`. A chamada pura e simples do `SqS`, isto é, sem opções, provoca a execução da simulação com valores padrão. A relação de opções e seus respectivos valores padrão pode ser obtida com a opção `--help`, cuja saída e mostrada abaixo:

```
SqS Single queue Simulator.

 Usage:
   $ sqs [options-list]

 Options:
   --num-arrv=xxx         Number of arrivals. Default to 100000.
   --queue-dscpl=xxx      Queue discipline [fcfs* | lcfs].
   --queue-type=xxx       Queue Type [mm1* | mg1 | gm1 | gg1].
   --seeds-set=xxx        Seeds set to generate random numbers [ss1* | ss2].
   --exp-arrv-rate=xxx    Exponential arrival rate. Default to 0.00106665.
   --exp-serv-rate=xxx    Exponential service rate. Default to 0.002133.
   --prt-arrv-alpha=xxx   Pareto arrival alpha parameter. Default to 1.5.
   --prt-arrv-min=xxx     Pareto arrival min value. Default to 312.491.
   --prt-serv-max=xxx     Pareto arrival max value. Default to 1E+11.
   --prt-serv-alpha=xxx   Pareto service alpha parameter. Default to 1.5.
   --prt-serv-min=xxx     Pareto service min value. Default to 156.210.
   --prt-serv-max=xxx     Pareto service max value. Default to 1E+11.
   --full-rslt            Print intermediates results from statistics at the
                          end of main results.
   --plain-rslt           Print results in a plain format. Ideal for plot data.
```

```
   --help                 Show this help.

 * = Default, if the value is not explicitly provided.
```

As principais opções incluídas no `SqS` estão relacionadas com os parâmetros do sistema, como o número de chegadas, disciplina de fila, tipo de fila e conjunto de sementes, e com os parâmetros das distribuições, que são taxa de chegada (λ) e serviço (μ) para exponencial, e α, valor mínimo e valor máximo para Pareto.

O formato padrão com que o `SqS` apresenta os resultados da simulação é mostrado abaixo.

```
--------------------------------------------------------------------------------
                            Simulation Report
--------------------------------------------------------------------------------

--> SYSTEM STATUS

    1) Arrivals.....................: 100000
    2) Finalized tasks..............: 99998
    3) Queue count..................: 1
    4) Simulation time..............: 9.35553e+07
    5) Traffic intensity............: 0.499259

--> QUEUE/STREAM SETTINGS

    6) Queue type...................: M/M/1
    7) Queue discipline.............: FCFS
    8) Seeds Set....................: ss1

--> EXPONENTIAL DISTRIBUTION PARAMETERS

    9) Exponential arrival rate....: 0.00106665
    10) Exponential service rate...: 0.0021333

--> PARETO DISTRIBUTION PARAMETERS

    11) Pareto arrival alpha.......: 1.5
    12) Pareto arrival min.........: 312.491
    13) Pareto arrival max.........: 1e+11
    14) Pareto service alpha.......: 1.5
    15) Pareto service min.........: 156.21
    16) Pareto service max.........: 1e+11

--> MEAN
```

```
    17) Interarrival time..........: 935.553
    18) Service time..............: 467.084
    19) Response time.............: 941.077
    20) Queue count...............: 1.01454
    21) Queue time................: 473.994
    22) Slowdown..................: 14.2627

--> STANDARD DEVIATION

    23) Interarrival time..........: 933.588
    24) Service time..............: 467.39
    25) Response time.............: 947.938
    26) Queue count...............: 1.45512
    27) Queue time................: 825.184
    28) Slowdown..................: 774.245

--> VARIANCE

    29) Interarrival time..........: 871587
    30) Service time..............: 218453
    31) Response time.............: 898586
    32) Queue count...............: 2.11739
    33) Queue time................: 680928
    34) Slowdown..................: 599455

--------------------------------------------------------------------------------
```

Este formato pode ser modificado pelas opções `--full-rslt` e `--plain-rslt`. A primeira destas apenas inclui os resultados intermediários do cálculo das estatísticas no final do conjunto principal de dados, como a soma ao quadrado e a média ao quadrado das variáveis, sendo interessante apenas para a validação dos resultados principais. A outra opção imprime os resultados em um formato plano onde os valores aparecem em linha e na mesma ordem vista no formato padrão de saída.

Além dos valores da média, desvio padrão e variância sobre as métricas do sistema, os resultados trazem outros dados importantes, dentre eles o número de tarefas finalizadas, o número de tarefas na fila quando a simulação foi encerrada e o tempo de simulação.

CAPÍTULO 5

RESULTADOS

Este capítulo apresenta os resultados dos experimentos de simulação. A primeira seção apresenta a caracterização da carga de trabalho, mostrando as condições de variabilidade da carga testada. As demais seções mostram resultados comparativos dos diversos sistemas de filas modelados.

5.1 Caracterização da Carga

A carga de trabalho do sistema é representada por dois parâmetros, o tempo de serviço das tarefas e o tempo entre chegadas. Esta seção apresenta a caracterização da carga, através da análise estatística destes parâmetros.

Os dois parâmetros são modelados pelas distribuições exponencial e de Pareto, que representam, respectivamente, distribuições de cauda leve (pequena variabilidade) e de cauda pesada (grande variabilidade). Para avaliar apenas o impacto da variabilidade, a média das distribuições foi mantida constante.

A Tabela 5.1 apresenta as médias e as variâncias para as distribuições exponencial e de Pareto utilizadas na simulação para $\rho = 0.9$. Nota-se que as médias são aproximadamente iguais, com a exceção das médias para a distribuição de Pareto com $\alpha = 1.3$. O valor mais baixo da média para $\alpha = 1.3$ pode ser atribuído a convergência lenta desta distribuição devido à grande variabilidade. Também é observado que as variâncias apresentam ordem de grandeza decrescente em uma unidade para a distribuição de Pareto à medida que o parâmetro α vai aumentando.

	Tempo entre Chegadas		Tempo de Serviço	
Distribuição	Média	Variância	Média	Variância
Exponencial	520.818	271275	468.762	219731
Pareto ($\alpha = 1.3$)	517.041	3.79137e+09	465.906	3.22243e+09
Pareto ($\alpha = 1.5$)	520.342	1.39815e+08	468.305	1.21274e+08
Pareto ($\alpha = 1.7$)	520.713	1.21246e+07	468.553	1.04713e+07
Pareto ($\alpha = 1.9$)	520.781	2.23789e+06	468.601	1.89363e+06

Tabela 5.1: Média e variância do tempo entre chegadas e do tempo de serviço para as distribuições exponencial e de Pareto com $\rho = 0.9$.

A Tabela 5.2 apresenta os valores máximo e mínimo observados para cada distribuição simulada. As faixas nas quais os valores estão distribuídos se diferem muito para as distribuições exponencial e de Pareto, apresentando esta última valores mínimos e máximos bem acima dos apresentados pela distribuição exponencial. Para instâncias diferentes da distribuição de Pareto, representadas pelas variações de α, é observado que valores menores de α apresentam maiores intervalos de distribuição das observações, o que é consistente com o fato de que quanto menor o valor de α, mais variável é a distribuição. A faixa de valores em que estão distribuídas as observações dá uma dimensão da variabilidade do conjunto de dados.

	Tempo entre Chegadas		Tempo de Serviço	
Distribuição	Min	Max	Min	Max
Exponencial	1.82716e-07	11005.9	1.71945e-07	9862.65
Pareto ($\alpha = 1.3$)	120.189	1.37358e+09	108.146	1.15316e+09
Pareto ($\alpha = 1.5$)	173.606	2.27802e+08	156.21	1.92925e+08
Pareto ($\alpha = 1.7$)	214.455	5.36528e+07	192.966	4.57629e+07
Pareto ($\alpha = 1.9$)	246.703	1.66799e+07	221.983	1.43074e+07

Tabela 5.2: Valores mínimo e máximo do tempo entre chegadas e do tempo de serviço para as distribuições exponencial e de Pareto com $\rho = 0.9$.

As figuras 5.1 e 5.2 apresentam as funções cdf e pdf das distribuições simuladas (respectivamente, gráficos à esquerda e à direita). Os gráficos mostram que o perfis das distribuições de Pareto e exponencial diferem bastante. A distribuição de Pareto apresenta picos de probabilidade para valores pequenos, entre 100 e 1000, enquanto as probabilidades da distribuição exponencial são pouco variáveis na faixa entre 1 e 100, com decaimento a partir deste valor.

O decaimento da probabilidade para valores grandes pode ser melhor observado nos gráficos que mostram a cauda das distribuições (Figura 5.3). A probabilidade de ocorrência de valores grandes cai muito mais rapidamente para a distribuição exponencial do que para a distribuição de Pareto. Entre as distribuições de Pareto com diferentes valores de α, pode ser notado que a probabilidade decai mais rapidamente para valores maiores de α. Em outras palavras, quanto maior o valor de α, menor a probabilidade de ocorrer observações grandes.

A Figura 5.4 exibe gráficos que mostram a convergência da média para os tempos entre chegadas e tempos de serviço, segundo as distribuições exponencial e de Pareto. Observa-se a convergência para a média, para ambos os parâmetros da carga, a partir de um milhão de chegadas, nas distribuições exponencial e de Pareto com α igual a 1.5, 1.7 e 1.9. Para Pareto com $\alpha = 1.3$, a média começa a indicar sinais de estabilidade

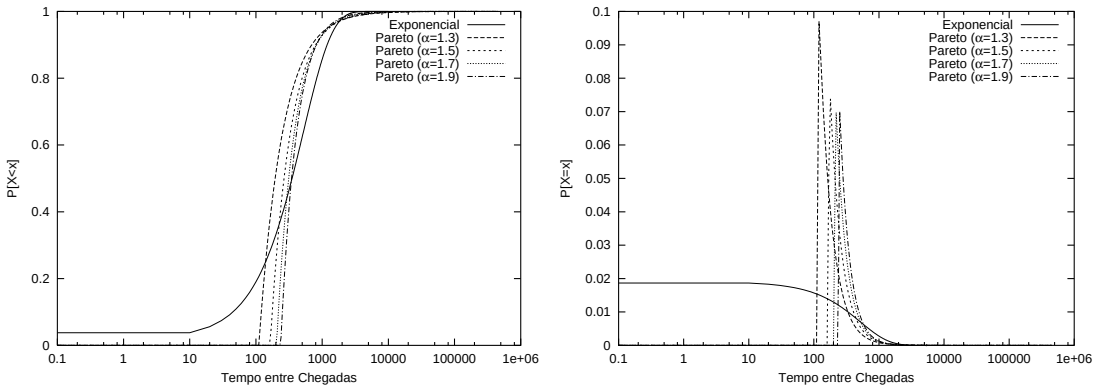

Figura 5.1: Curvas de probabilidade cdf e pdf do tempo entre chegadas para as distribuições exponencial e de Pareto.

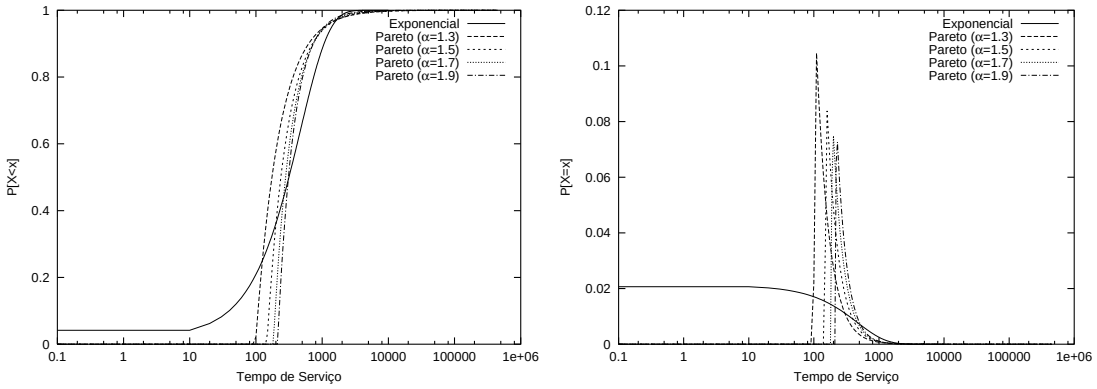

Figura 5.2: Curvas de probabilidade cdf e pdf do tempo de serviço para as distribuições exponencial e de Pareto.

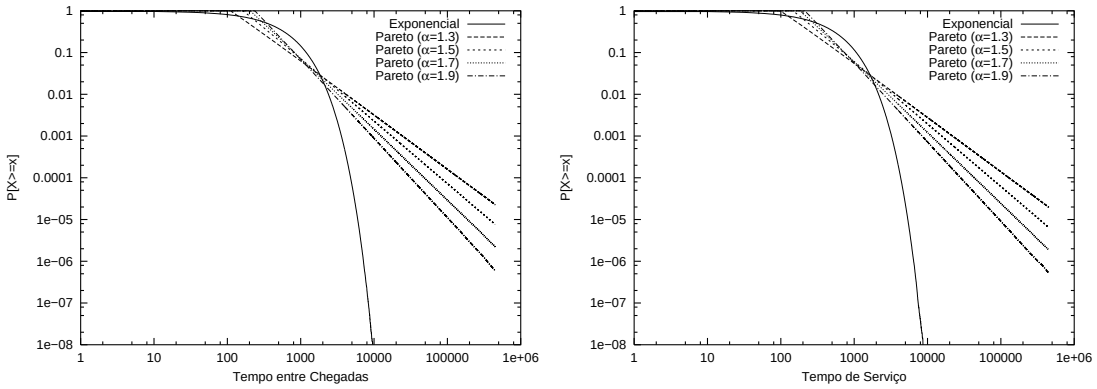

Figura 5.3: Visualização de cauda do tempo entre chegadas e tempo de serviço para as distribuições exponencial e de Pareto.

próximos a 100 milhões de chegadas, ainda não se mostrando totalmente estabilizada na faixa de 1 bilhão de chegadas, como mostrado na Tabela 5.1. Apenas para comparação, os gráficos também mostram a curva de convergência de Pareto com $\alpha = 1.1$. Neste caso, pode ser observada a convergência lenta, embora a estabilização não ocorra até um bilhão de chegadas.

Para o gráfico do tempo de serviço, com α igual a 1.1 e 1.3, são observados picos nas curvas, próximos a 10 milhões de chegadas, provavelmente ocasionados por um ou mais valores extremamente grandes, que elevaram a média acima do esperado, fato que pode ser explicado pelas próprias características da distribuição de Pareto, cuja probabilidade de gerar valores grandes é não desprezível.

Tal comportamento da convergência das distribuições foi determinante para a decisão de não execução de simulações com o valor α igual a 1.1 para a distribuição de Pareto, e também para a escolha do número de chegadas utilizado.

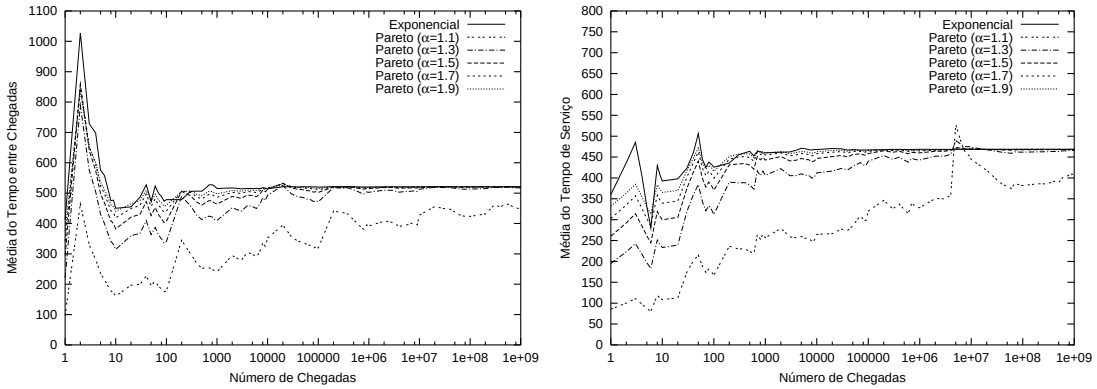

Figura 5.4: Convergência das médias do tempo entre chegadas e do tempo de serviço para as distribuições exponencial e de Pareto.

5.2 Comparação entre as Filas M/M/1 e M/G/1

Esta estudo refere-se ao impacto da variabilidade nos tempos de serviço, comparado através das duas distribuições em questão. Nesta seção, a taxa de chegadas não é alterada, isto é, segue a distribuição exponencial em ambos os casos (primeiro M).

A Figura 5.5 apresenta curvas para a média (gráfico à esquerda) e a variância (gráfico à direita) do tamanho da fila, ambas em função da intensidade do tráfego (ρ). Observa-se que o tamanho médio da fila, conforme esperado, cresce com a intensidade do tráfego, para todas as distribuições simuladas.

A distribuição exponencial apresenta a menor média para o tamanho da fila, segui-

da pelas distribuições de Pareto, que aparecem de acordo com seu grau de variabilidade (quanto maior a variabilidade, maior o tamanho da fila).

Observa-se claramente o impacto da variabilidade no tamanho da fila. Por exemplo, para $\rho = 0.5$, a média do tamanho da fila $M/M/1$ tem valor 1.0, enquanto a média do tamanho da fila $M/G/1$ com $\alpha = 1.9$ tem valor 2.893 e, para $\alpha = 1.3$, o valor é 3610.780.

As curvas para a variância do tamanho da fila, são qualitativamente similares às da média, porém observa-se que as variâncias para $M/G/1$ são bem maiores do que as da fila $M/M/1$, em relação aos valores observados para as médias.

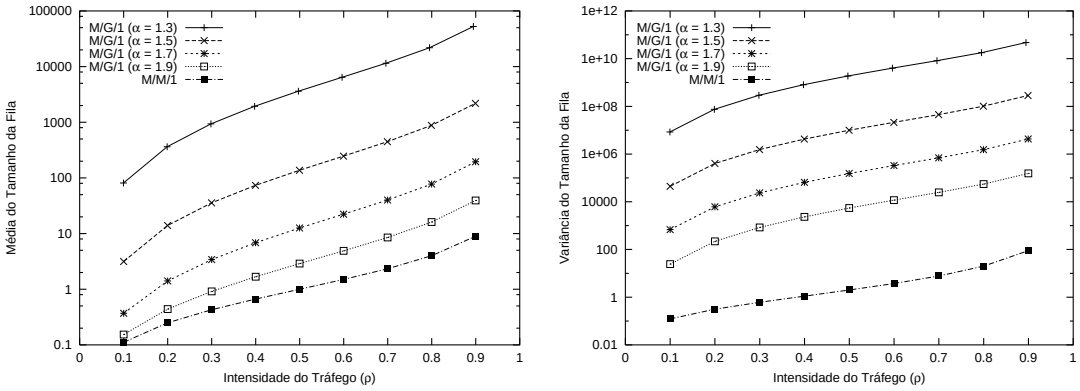

Figura 5.5: Média e variância do tamanho da fila para as filas $M/M/1$ e $M/G/1$.

Da mesma forma, a média e a variância do tempo de resposta (Figura 5.6) também experimentam crescimento com o aumento da intensidade do tráfego do sistema. Novamente a variabilidade, representada pela fila $M/G/1$, tem grande impacto tanto na média como na variância, comparadas com as curvas da fila $M/M/1$, que marcam valores bem menores do que as curvas da fila $M/G/1$. No caso da média, as tarefas da fila $M/G/1$ com $\alpha = 1.7$ têm tempo médio de resposta igual a aproximadamente 10 vezes o tempo de resposta da fila $M/M/1$ no ponto de $\rho = 0.5$.

Outra observação importante refere-se ao comportamento das métricas citadas acima sobre a fila $M/G/1$. Nota-se que os valores das médias para as instâncias da fila $M/G/1$ com α iguais a 1.3 e 1.5 no ponto de $\rho = 0.5$ são maiores do que os valores das médias para as instâncias com α iguais a 1.7 e 1.9 no ponto de $\rho = 0.9$, bem como para a fila $M/M/1$. Isto significa que a variabilidade inserida nos tempos de serviço causa maior impacto ao sistema do que variações na intensidade do tráfego.

Os gráficos para a métrica *slowdown* são mostrados na Figura 5.7. Esta métrica comporta-se de maneira diferente das outras métricas já discutidas. A média do *slowdown* para a fila $M/M/1$ é superior à média da fila $M/G/1$ com $\alpha = 1.9$. Isto pode ser explicado pelo fato de que, na distribuição de Pareto, a maioria das observações é pequena (ver

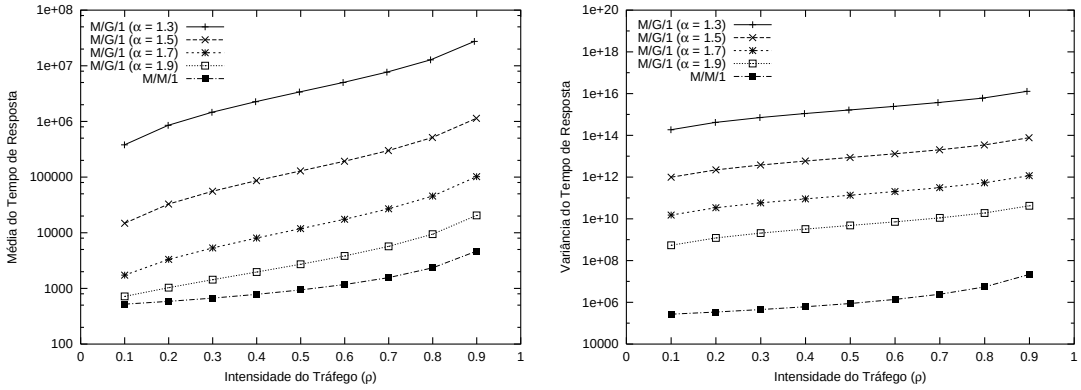

Figura 5.6: Média e variância do tempo de resposta para as filas $M/M/1$ e $M/G/1$.

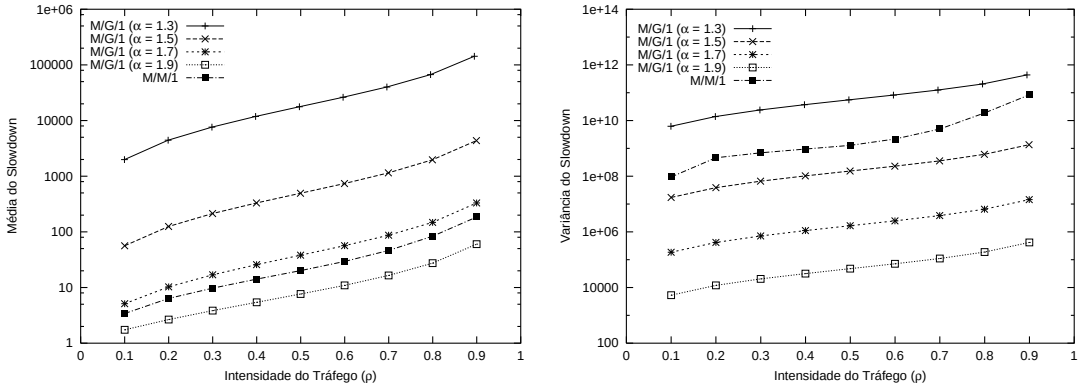

Figura 5.7: Média e variância do *slowdown* para as filas $M/M/1$ e $M/G/1$.

Tabela 2.1 da Seção 2.1), gerando muitos tempos de resposta com valores pequenos que diminuem a média do *slowdown*.

Ainda analisando as particularidades do *slowdown*, a curva da variância da fila $M/M/1$ é inferior apenas à curva da fila $M/G/1$ com $\alpha = 1.3$, que é a instância da distribuição de Pareto com maior variabilidade dentre as simuladas.

Portanto, é observado que a fila $M/M/1$ apresenta os melhores resultados quanto ao tamanho da fila e tempo de resposta, fato que não ocorre para a métrica *slowdown*. Desse forma, a variabilidade da distribuição de Pareto é prejudicial às métricas tradicionais, mas pode ser benéfica (até certo ponto) para a métrica de justiça.

5.3 Comparação entre as Filas M/M/1 e G/M/1

Nesta comparação, a variabilidade é analisada no processo de chegadas, isto é, o tempo entre chegadas é modelado tanto pela distribuição exponencial quanto pela distribuição

de Pareto. O tempo de serviço segue apenas a distribuição exponencial.

Uma visão geral das figuras 5.8, 5.9 e 5.10 mostra que o desempenho das métricas do sistema não difere consideravelmente para ρ entre 0.3 e 0.7, especialmente comparando-se a distribuição exponencial com as distribuições de Pareto com α iguais a 1.7 e 1.9.

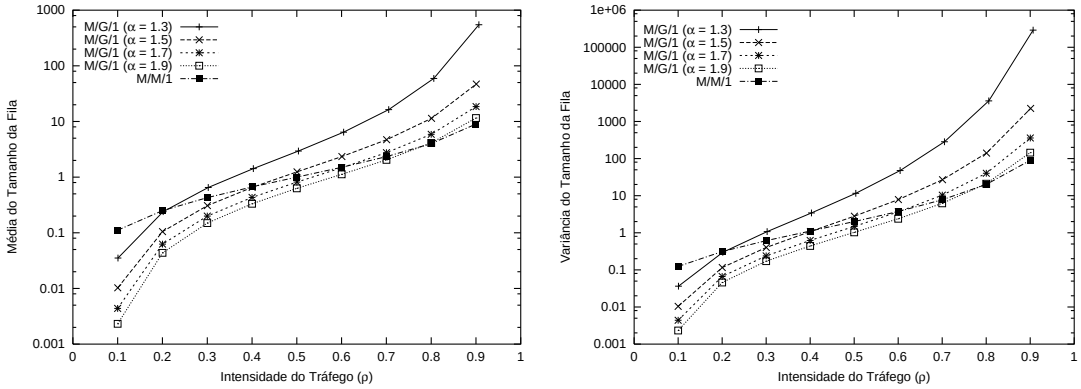

Figura 5.8: Média e variância do tamanho da fila para as filas $M/M/1$ e $G/M/1$.

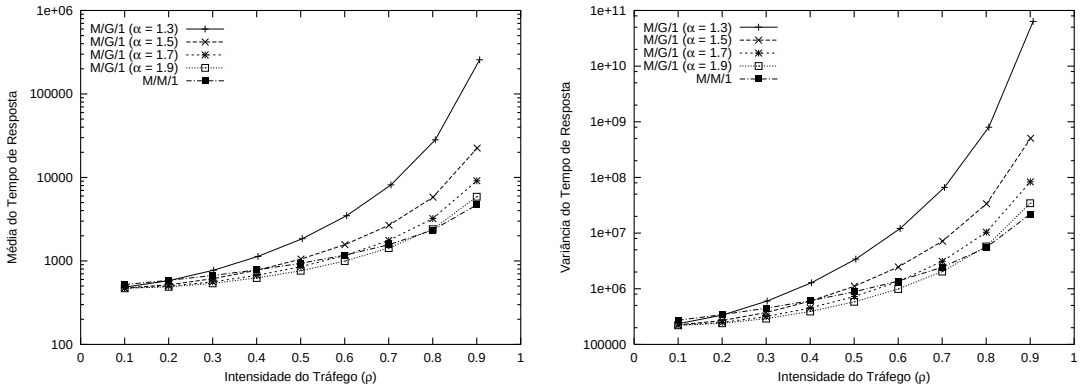

Figura 5.9: Média e variância do tempo de resposta para as filas $M/M/1$ e $G/M/1$.

Para as métricas tamanho da fila e *slowdown* (figuras 5.8 e 5.10), tanto a média e a variância de $G/M/1$ mostram-se inferiores aos valores de $M/M/1$ nas intensidades de tráfego equivalentes a 0.1 e 0.2. Uma situação contrária é observada nos maiores valores de ρ do sistema (0.8 e 0.9), onde Pareto supera a distribuição exponencial.

As estatísticas para o tempo de resposta (Figura 5.9) mantêm-se com pouca diferença entre as filas até ρ igual a 0.5. A partir deste ponto, a média e a variância da filas que possuem distribuição de Pareto com α iguais a 1.3 e 1.5 começam a apresentar um crescimento mais acentuado do que as demais filas.

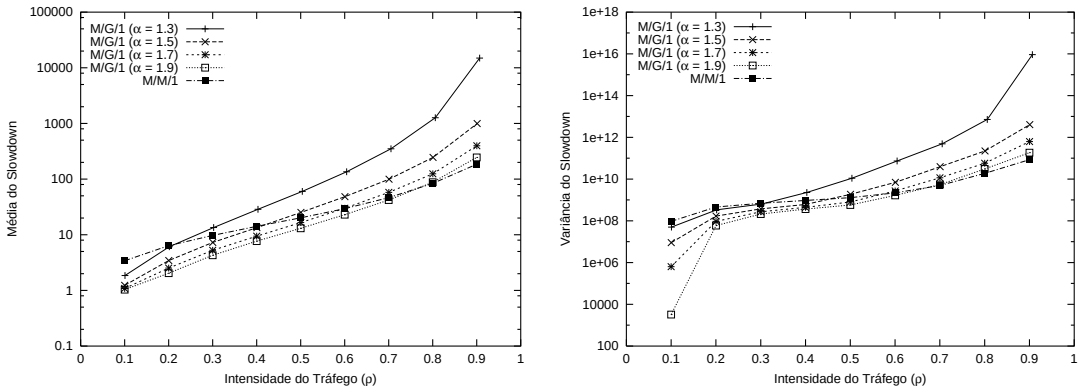

Figura 5.10: Média e variância do *slowdown* para as filas $M/M/1$ e $G/M/1$.

Os efeitos da variabilidade, inserida apenas no processo de chegada, sobre as métricas podem ser melhor notados nos momentos em que o sistema apresenta maior intensidade de tráfego. Nesta situação, a métrica mais afetada é o tempo de resposta, onde a média e a variância no ponto de intensidade de tráfego máxima do sistema ($\rho = 0.9$) para a fila $G/M/1$ ($\alpha = 1.3$) chegam, respectivamente, aos valores de 256584 e 6.37311e10, contra 4686.620 e 2.19312e7 da fila $M/M/1$.

5.4 Comparação entre as Filas G/G/1 e M/G/1

Esta seção apresenta a avaliação dos efeitos da variabilidade sobre o processo de chegada das tarefas do sistema, estando estas submetidas à uma grande variabilidade nos tempos de serviço.

As figuras 5.11, 5.12 e 5.13 mostram os gráficos que comparam o desempenho das métricas das filas $G/G/1$ e $G/M/1$. Para todas as métricas, ambas as filas apresentam comportamento similar, com valores muito próximos tanto para a média quanto para a variância. As maiores diferenças são encontradas na média do tamanho da fila (Figura 5.11), onde as instâncias da fila $G/G/1$ com α iguais a 1.7 e 1.9 são, respectivamente, 22% e 64% maiores que as mesmas instâncias da fila $G/M/1$ no ponto de intensidade de tráfego com o valor de 0.1. Esta diferença não é relevante pois os valores do tamanho da fila são muito pequenos em $\rho = 0.1$ (menores que 0.1), sendo prejudicados por problemas de precisão.

Assim, quando o tempo de serviço é muito variável, a variabilidade inserida no processo de chegada não afeta significativamente o desempenho das métricas do sistema. O mesmo comportamento não pode ser verificado quando o tempo de serviço detém pouca variabilidade, como foi mostrado na Seção 5.3.

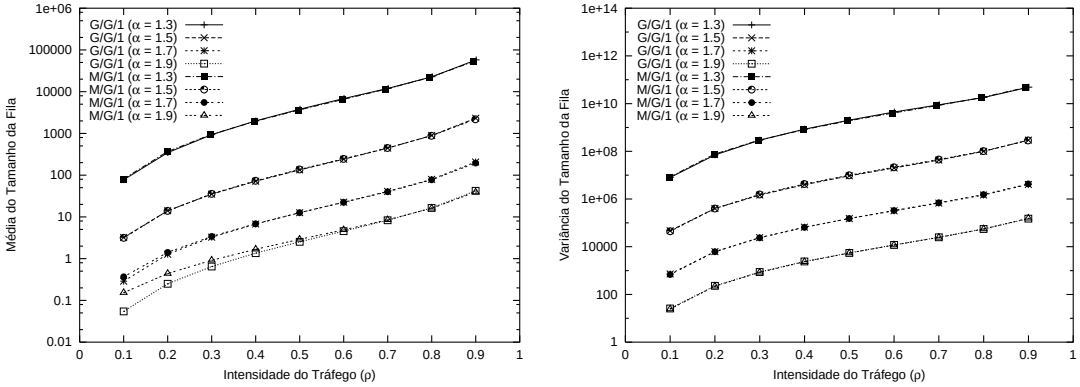

Figura 5.11: Média e variância do tamanho da fila para as filas $G/G/1$ e $M/G/1$.

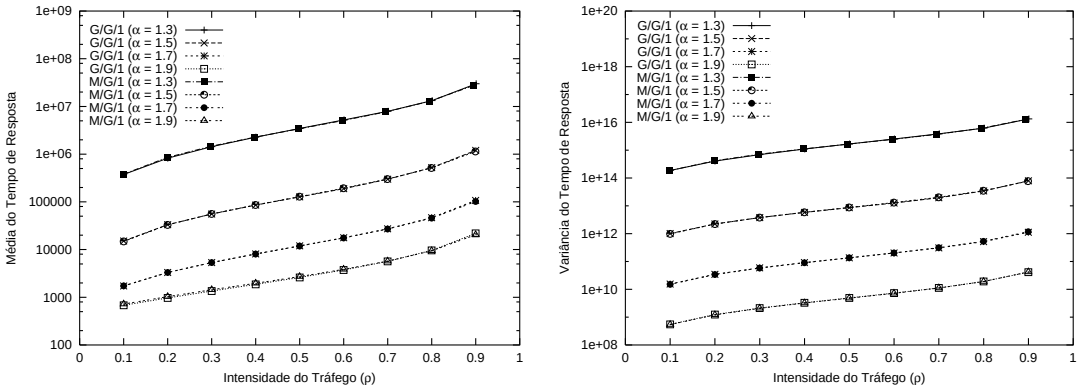

Figura 5.12: Média e variância do tempo de resposta para as filas $G/G/1$ e $M/G/1$.

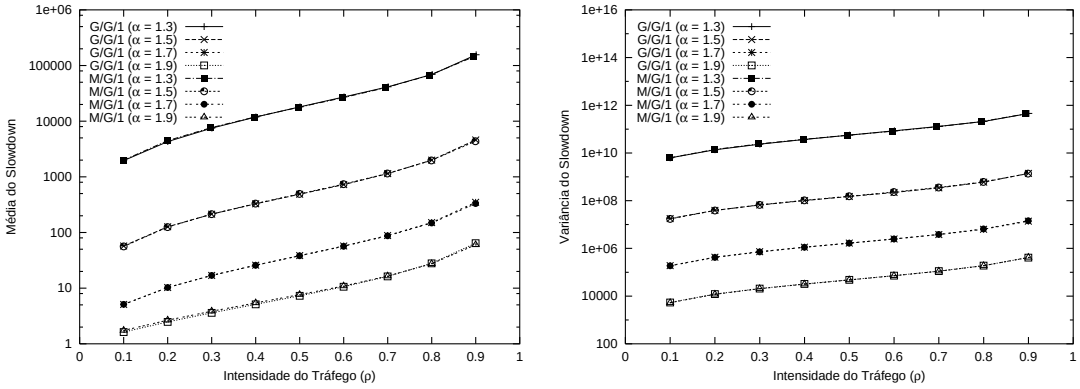

Figura 5.13: Média e variância do *slowdown* para as filas $G/G/1$ e $M/G/1$.

5.5 Comparação entre as Filas G/G/1 e G/M/1

A comparação feita nesta seção estuda o comportamento das métricas do sistema quando submetidas a cargas com muita e pouca variabilidade nos tempos de serviço. Os valores do processo de chegada para ambas as filas são gerados pela distribuição de Pareto.

Tanto para o tamanho da fila quanto para o tempo de resposta, vistos, respectivamente, nas figuras 5.14 e 5.15, as curvas da fila $G/G/1$ para a média e a variância superam as curvas da fila $G/M/1$ para um mesmo valor do parâmetro α. Por exemplo, no ponto de $\rho = 0.1$, a média do tempo de resposta para a instância da fila $G/G/1$ com $\alpha = 1.3$ é 700 vezes maior que a média da mesma instância da fila $G/M/1$. Esta diferença diminui consideravelmente no ponto onde ρ é igual a 0.9, com $G/G/1$ sendo apenas 100 vezes maior que $G/M/1$, mostrando o grande crescimento desta última fila quando o sistema é submetido à uma maior intensidade de tráfego.

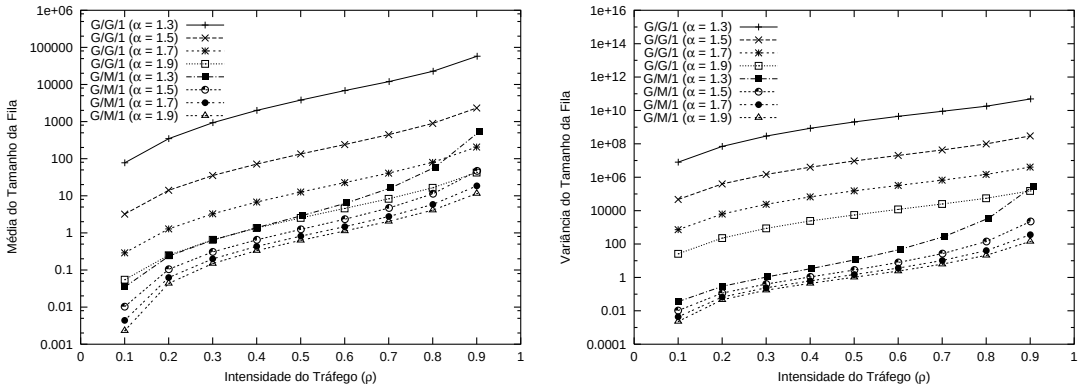

Figura 5.14: Média e variância do tamanho da fila para as filas $G/G/1$ e $G/M/1$.

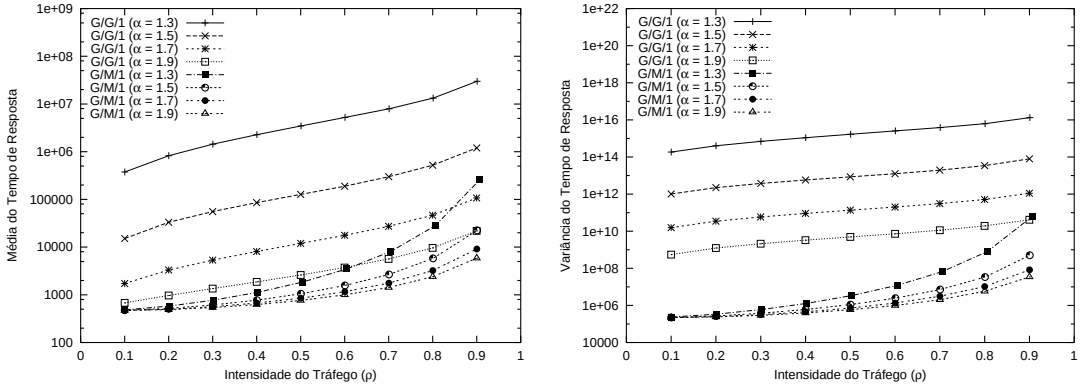

Figura 5.15: Média e variância do tempo de resposta para as filas $G/G/1$ e $G/M/1$.

Observa-se que a média do *slowdown* (Figura 5.16) é similar para as filas com valores de α na faixa de 1.5 a 1.9. Quanto à variância desta métrica, os resultados da fila $G/M/1$ para os menores valores de α (1.3 e 1.5) permanecem muito próximos dos resultados da fila $G/G/1$ com $\alpha = 1.3$, sendo maiores que este último nos pontos de ρ iguais a 0.8 e 0.9.

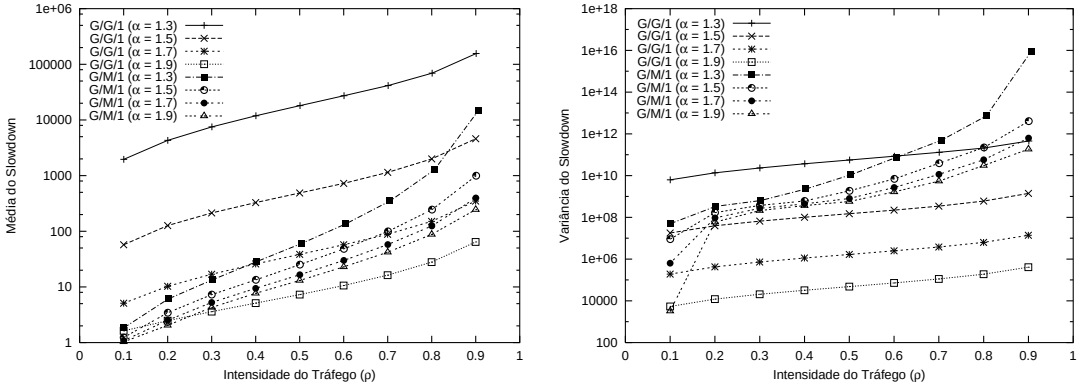

Figura 5.16: Média e variância do *slowdown* para as filas $G/G/1$ e $G/M/1$.

De forma geral, as curvas da fila $G/G/1$ têm os valores mais altos, sofrendo variações mais suaves do que as curvas da fila $G/M/1$, as quais apresentam picos nos pontos de maior intensidade de tráfego do sistema.

5.6 Comparação entre as Filas M/G/1 e G/M/1

Esta seção apresenta a comparação entre filas que detêm grande variabilidade em apenas um dos parâmetros da carga. Os gráficos são apresentados em função do parâmetro α da distribuição de Pareto. Tal análise permite avaliar sobre qual parâmetro a variabilidade causa maior impacto: no processo de chegadas ou nos tempos de serviço.

Observando a média e a variância das métricas tamanho da fila e tempo de resposta para uma mesma intensidade de tráfego (figuras 5.17 e 5.18), nota-se que a fila $M/G/1$ supera a fila $G/M/1$ em todos os casos, isto é, à medida que o valor de α vai diminuindo (maior variabilidade), a variabilidade inserida no tempo de serviço causa um efeito negativo sobre a média destas métricas muito maior que a variabilidade no processo de chegada.

Os gráficos para o *slowdown* são mostrados na Figura 5.19. A variabilidade nos tempos de serviço das tarefas não provoca grandes diferenças entre as médias do *slowdown* até o valor 1.7 do eixo do parâmetro α, ponto onde as filas com mesma intensidade de tráfego apresentam desempenho semelhante. Entretanto, os menores valores do parâmetro

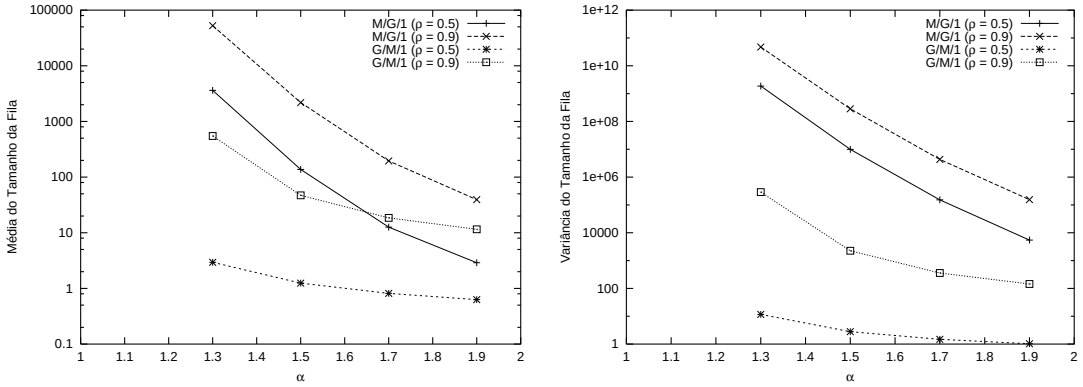

Figura 5.17: Média e variância do tamanho da fila para as filas $M/G/1$ e $G/M/1$.

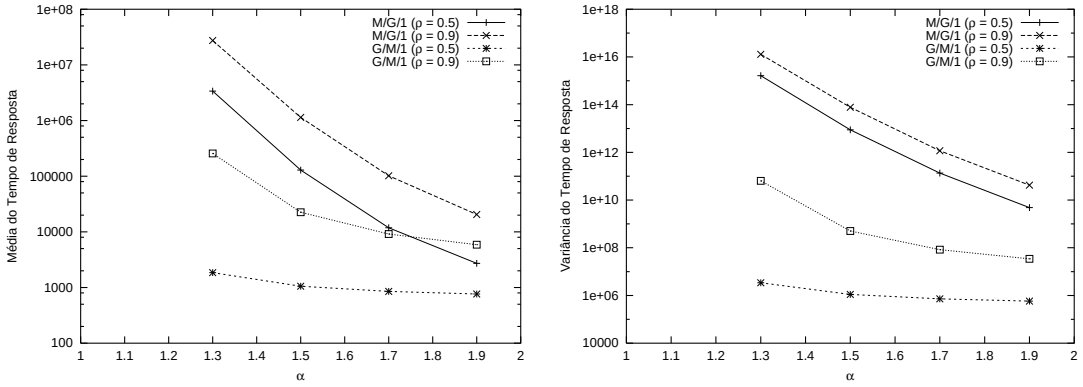

Figura 5.18: Média e variância do tempo de resposta para as filas $M/G/1$ e $G/M/1$.

α (1.3 e 1.5) fazem com que a fila $M/G/1$ produza um crescimento bem maior do que o visto na fila $G/M/1$. Por exemplo, a fila $M/G/1$ mostra um aumento de mais de 432 vezes entre os valores 1.7 e 1.3 de α, enquanto a fila $G/M/1$ experimenta um crescimento de 37 vezes.

Quanto à variância do *slowdown*, a fila $M/G/1$ continua apresentando um crescimento mais acentuado do que a fila $G/M/1$, mas não o suficiente para ultrapassar os valores desta última com $\rho = 0.9$, que chega ao valor máximo de 9.372e15, contra apenas 4.224e11 de $M/G/1$.

Torna-se evidente que, à medida que os valores do parâmetro α vão diminuindo, os efeitos da variabilidade sobre os tempos de serviço causam maior impacto negativo às métricas do sistema do que a variabilidade inserida no processo de chegada.

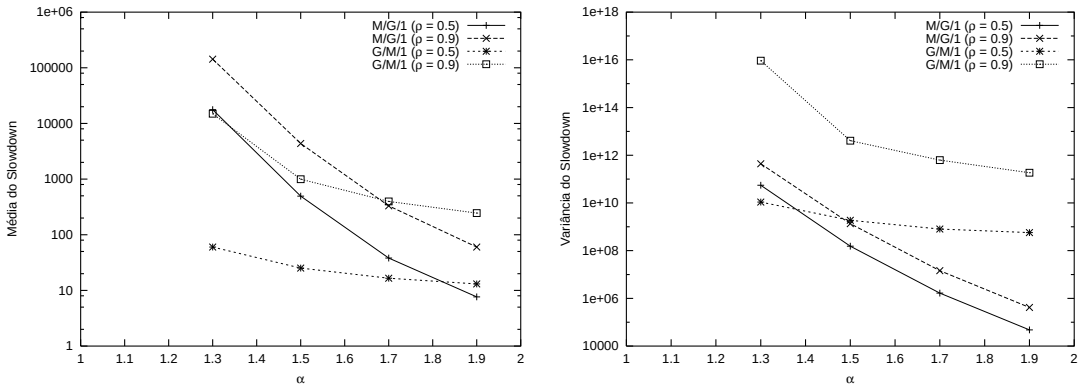

Figura 5.19: Média e variância do *slowdown* para as filas $M/G/1$ e $G/M/1$.

5.7 Análise do Impacto da Variabilidade

As seções anteriores mostraram o comportamento das métricas do sistema sobre cargas com pouca e muita variabilidade. Esta seção faz uma avaliação geral das efeitos desta variabilidade.

Nota-se que nas comparações onde um dos parâmetros da carga está fixado em uma única distribuição (seções 5.2, 5.3, 5.4 e 5.5), a variabilidade não se mostra tão prejudicial para a métrica *slowdown*, principalmente em relação a variância. Uma exceção para este quadro ocorre quando a variabilidade é extremamente alta como, por exemplo, na distribuição de Pareto com $\alpha = 1.3$. Os maiores valores para *slowdown*, tanto para a média como para a variância, são obtidos quando a variabilidade é inserida nos tempos de serviço, independente da distribuição (exponencial ou Pareto) atribuída ao processo de chegada.

Nas métricas tamanho da fila e tempo de resposta, novamente as comparações mostram que a variabilidade sobre os tempos de serviço é mais prejudicial do que a variabilidade sobre o processo de chegada. Em outras palavras, a diferença entre os valores das filas do tipo $M/M/1$ e $M/G/1$ é bem maior do que a diferença entre os valores das filas do tipo $M/M/1$ e $G/M/1$. A mesma observação é valida se a fila $M/M/1$ for substituída pela fila $G/G/1$.

Os resultados das comparações entre as filas $M/G/1$ e $G/M/1$ (Seção 5.6) comprovam que a variabilidade sobre os tempos de serviço afeta o desempenho das métricas do sistema, sendo consideravelmente pior do que a mesma variabilidade inserida no processo de chegada. O reflexo deste fato pode ser visto nos gráficos da Seção 5.4, onde a variação das distribuições sobre o processo de chegada nas filas com grande variabilidade no tempo de serviço não causa efeitos significativos sobre os valores das métricas.

Outra característica apresentada por sistemas que detêm grande variabilidade no

parâmetro da carga referente ao tempo de serviço, como as filas $M/G/1$ e $G/G/1$, é o efeito desta variabilidade comparado à variações da intensidade do tráfego que o sistema experimenta. Neste caso, o aumento da variabilidade sobre os tempos de serviço (ou menores valores do parâmetro α da distribuição de Pareto), é mais prejudicial ao desempenho do sistema do que variações na intensidade de tráfego do sistema.

A validação dos resultados das simulações pode ser encontrada no Apêndice A, bem como uma discussão sobre a mesma.

CAPÍTULO 6

CONCLUSÕES

Este trabalho apresentou um estudo sobre os efeitos da variabilidade da carga em um sistema de fila única e recurso único. O estudo foi feito através da simulação deste sistema, onde os resultados de filas com diferentes graus de variabilidade nos parâmetros da carga foram comparados. O simulador implementado utilizou as distribuições exponencial e de Pareto para representar, respectivamente, conjuntos de valores com pequena e grande variabilidade.

A dificuldade da convergência para um estado de equilíbrio em simulações com a distribuição de Pareto foi comprovada, principalmente para os menores valores do parâmetro α.

Quanto ao impacto da variabilidade, nota-se que esta tem efeitos negativos sobre as métricas tradicionais (tamanho da fila e tempo de resposta), mas pode ser benéfica à métrica *slowdown*, ou métrica de justiça, em casos onde a variabilidade é moderada, como na distribuição de Pareto com α iguais a 1.5, 1.7 e 1.9.

Em destaque, foi observado que a variabilidade inserida no parâmetro da carga referente aos tempos de serviço é bem mais prejudicial ao desempenho do sistema do que a mesma variabilidade sobre o processo de chegada. Outra conclusão importante é que a variabilidade causa maiores impactos ao desempenho do sistema do que o impacto de variações na intensidade do tráfego. Este é um resultado de grande valia para a teoria de filas, pois podemos ter sistemas com tráfegos razoáveis ($\rho = 0.5$), cujo desempenho pode ser seriamente prejudicado pela variabilidade.

A avaliação do impacto da variabilidade sobre sistemas de filas que utilizem outras políticas de escalonamento é uma continuação importante para este trabalho. Duas políticas de interesse imediato seriam *Round-Robin* (RR) e *Shortest Remaining Processing Time* (SRPT), além de *Shortest Job First* (SJF).

REFERÊNCIAS

[Bansal and Harchol-Balter, 2001] Bansal, N. and Harchol-Balter, M. (2001). Analysis of SRPT Scheduling: Investigating Unfairness. In *SIGMETRICS/Performance*, pages 279–290.

[Coplien, 1994] Coplien, J. O. (1994). *Advanced C++: Programming Styles and Idioms*. Addison-Wesley.

[Crovella and Bestavros, 1996] Crovella, M. E. and Bestavros, A. (1996). Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes. In *ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pages 160–169.

[Crovella and Lipsky, 1997] Crovella, M. E. and Lipsky, L. (1997). Long-Lasting Transient Conditions in Simulations With Heavy-Tailed Workloads. In *Winter Simulation Conference*.

[Crovella et al., 1996] Crovella, M. E., Taqqu, M. S., and Destavros, A. (1996). Heavy-Tailed Probability Distributions in the World Wide Web. *A Practical Guide to Heavy Tails: Statistical Techniques for Analyzing Heavy Tailed Distributions*.

[Greiner et al., 1995] Greiner, M., Jobmann, M., and Lipsky, L. (1995). Importance of Power-tail Distributions For Telecommunication Traffic Models. Technical Report TUM-I9521, Institute für Informatik, Technische Universität München, Gemany.

[Harchol-Balter and Downey, 1996] Harchol-Balter, M. and Downey, A. (1996). Exploiting Process Lifetime Distributions for Dynamic Load Balancing. In *ACM Sigmetrics International Conference on Measurement and Modeling of Computer Systems*, pages 13–24.

[Heyman, 2000] Heyman, D. P. (2000). Performance Implications of Vary Large Service-Times Variances. *Performance Evaluation*, (40):47–70.

[Jain, 1991] Jain, R. (1991). *The Art of Computer Systems Performance: Analysis, Techniques for Experimental Design, Measurement, Simulation, and Modeling*. John Wiley & Sons, New York.

[Jonack, 2002] Jonack, M. A. (2002). SqS Single queue Simulator. Disponível em http://www.inf.ufpr.br/~cristina/projects.html.

[Leland et al., 1994] Leland, W., Taqqu, M. S., Willinger, W., and Wilson, D. V. (1994). On the self-similar nature of ethernet traffic. *IEEE/ACM Transactions on Networking*, 2(1):1–15.

[Menascé and Almeida, 2002] Menascé, D. A. and Almeida, V. A. F. (2002). *Capacity Planning for Web Services: Metrics, Models, and Methods*. Prentice-Hall, New Jersey.

[Molloy, 1989] Molloy, M. K. (1989). *Fundamentals of Performance Modeling*. Macmillan Publishing.

[Murta, 1999] Murta, C. D. (1999). *Modelo de Particionamento de Espaço para Caches da WWW*. PhD thesis, Universidade Federal de Minas Gerais, Brasil, Departamento de Ciência da Computação.

[Papoulis, 1991] Papoulis, A. (1991). *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, third edition.

[Stroustrup, 1993] Stroustrup, B. (1993). *The C++ Programming Language*. Addison-Wesley, second edition.

APÊNDICE A

VALIDAÇÃO DOS RESULTADOS DAS SIMULAÇÕES

Este apêndice mostra a validação feita para alguns dos resultados obtidos nas simulações. O significado dos símbolos utilizados nas tabelas é descrito abaixo.

$\rho =$ intensidade do tráfego.

$r =$ tempo de resposta.

$n =$ tamanho da fila, incluindo a tarefa sendo atendida pelo servidor.

$w =$ tempo de fila, incluindo a tarefa sendo atendida pelo servidor.

$A =$ resultados obtidos através das fórmulas do modelo analítico do sistema de filas.

$S =$ resultados obtidos através da simulação do sistema de filas.

$E[x] =$ média da variável x.

$Var[x] =$ variância da variável x.

A validação foi feita comparando-se os resultados da média e variância das métricas provenientes da simulação com os resultados das fórmulas da teoria de filas.

Entretanto, deve-se observar que a modelagem analítica não provê meios para validar todos os resultados obtidos neste trabalho, como foi o caso da métrica *slowdown*. Também em alguns casos, os resultados da simulação não puderam ser testados, pois as fórmulas para estes exigiam valores que não podiam ser obtidos diretamente. Por exemplo, a variância do tempo de resposta da fila $M/G/1$ depende do terceiro momento da média do tempo de serviço, o qual não foi calculado a partir dos resultados da simulação. Assim, foram utilizados os resultados conhecidos e possíveis de serem calculados analiticamente para validar os resultados da simulação. Dessa forma, considerando a validação realizada e confirmada nestes casos, acredita-se que os demais resultados gerados pelo simulador também estejam corretos.

As fórmulas utilizadas na validação das filas $M/M/1$ e $M/G/1$ são as mesmas mostradas nas tabelas 2.2 e 2.3 da Seção 2.2. Para as filas $G/M/1$ e $G/G/1$, a validação utilizou as fórmulas da lei de Little, também descritas na Seção 2.2. As filas que utilizam a distribuição de Pareto (G) tiveram os resultados testados para o parâmetro α com valores 1.3 e 1.5.

As tabelas A.1 e A.2 mostram, respectivamente, a validação para a média e variância do tempo de resposta, tamanho da fila e tempo de fila do modelo $M/M/1$. Os

resultados da análise (A) e da simulação (S) são muito similares. Esta proximidade dos resultados pode ser explicada pela precisão dos valores de ρ obtidos para a fila $M/M/1$, visto que as fórmulas de validação desta fila utilizam massivamente tais valores.

A validação para a fila $M/G/1$ com instâncias da distribuição de Pareto referentes a $\alpha = 1.3$ e $\alpha = 1.5$ é mostrada nas tabelas A.3 a A.6. Nota-se que os resultados da simulação para as médias das métricas são sempre menores do que os resultados da análise, fato que pode ser atribuído aos valores de ρ da simulação, os quais são inferiores aos da análise. Também há o fato de que a média da distribuição de Pareto não alcançou o valor esperado dentro de 1 bilhão de amostras para $\alpha = 1.3$. Os resultados da variância mostram-se idênticos ao da análise, pois como se tratam de valores muito grandes, pequenas diferenças passam despercebidas.

As tabelas A.7 a A.10 mostram a validação das filas $G/M/1$ e $G/G/1$ para instâncias da distribuição de Pareto com α iguais a 1.3 e 1.5. Estas validações foram feitas através das fórmulas da lei de Little. Os resultados da simulação para a média e variância do tempo de resposta apresentam pouca diferença com relação aos resultados da análise. Já a média do tamanho da fila experimenta um comportamento diferente, com os resultados de maior valor sendo alternados entre a análise e a simulação, os quais merecem um estudo mais detalhado não feito neste trabalho.

Em geral, observa-se que os resultados da simulação que foram comparados com a análise são iguais ou muito similares, onde o erro relativo máximo é visto na Tabela A.7 (fila $G/M/1$), em $E[n]$ com $\alpha = 1.3$ e $\rho = 0.9$, chegando a 10%. Assim, pode-se considerar que os resultados da simulação foram validados.

ρ		$E[r]$		$E[n]$		$E[w]$	
A	S	A	S	A	S	A	S
0.1	0.100006	520.845	520.847	0.111119	0.111121	52.0876	52.0848
0.2	0.200011	585.955	585.962	0.250017	0.250049	117.197	117.2
0.3	0.300017	669.67	669.693	0.428606	0.42866	200.912	200.931
0.4	0.400022	781.291	781.328	0.666728	0.666825	312.534	312.566
0.5	0.500028	937.567	937.6	1.00011	1.00016	468.81	468.838
0.6	0.600033	1171.99	1172.03	1.50021	1.50025	703.233	703.264
0.7	0.700039	1562.73	1562.79	2.33377	2.33388	1093.97	1094.03
0.8	0.800044	2344.3	2344.26	4.0011	4.00103	1875.55	1875.5
0.9	0.90005	4689.92	4686.62	9.005	8.99836	4221.16	4217.86

Tabela A.1: Validação da média de r, n e w para a fila $M/M/1$.

ρ		$Var[r]$		$Var[n]$		$Var[w]$	
A	S	A	S	A	S	A	S
0.1	0.100006	271279	271261	0.123466	0.123474	51546	51534.8
0.2	0.200011	343343	343310	0.312526	0.312602	123609	123590
0.3	0.300017	448457	448429	0.612309	0.612421	228724	228709
0.4	0.400022	610415	610402	1.11125	1.1115	390682	390682
0.5	0.500028	879032	878886	2.00034	2.00016	659299	659169
0.6	0.600033	1.37356e+06	1.37322e+06	3.75083	3.74987	1.15383e+06	1.1535e+06
0.7	0.700039	2.44212e+06	2.44125e+06	7.78023	7.77716	2.22238e+06	2.22154e+06
0.8	0.800044	5.49575e+06	5.49202e+06	20.0099	19.9966	5.27602e+06	5.27233e+06
0.9	0.90005	2.19953e+07	2.19312e+07	90.0951	89.8375	2.17756e+07	2.17117e+07

Tabela A.2: Validação da variância de r, n e w para a fila $M/M/1$.

ρ		$E[r]$		$E[n]$		$E[w]$	
A	S	A	S	A	S	A	S
0.1	0.0993963	382165	380013	81.531	80.9479	381699	379547
0.2	0.198793	858572	851205	366.336	362.947	858106	850739
0.3	0.298189	1.46992e+06	1.45903e+06	940.777	934.011	1.46945e+06	1.45856e+06
0.4	0.397585	2.28301e+06	2.26586e+06	1948.22	1934.01	2.28254e+06	2.26539e+06
0.5	0.496981	3.41743e+06	3.38468e+06	3645.36	3610.78	3.41696e+06	3.38422e+06
0.6	0.596378	5.11059e+06	5.01627e+06	6541.75	6420.96	5.11012e+06	5.0158e+06
0.7	0.695774	7.9101e+06	7.68338e+06	11812.8	11473.3	7.90963e+06	7.68291e+06
0.8	0.79517	1.34266e+07	1.28085e+07	22915.4	21860.4	1.34261e+07	1.2808e+07
0.9	0.894566	2.93443e+07	2.74018e+07	56342.6	52611.4	2.93438e+07	2.74013e+07

Tabela A.3: Validação da média de r, n e w para a fila $M/G/1$ com $\alpha = 1.3$.

ρ		$Var[w]$	
A	S	A	S
0.1	0.0993963	1.85704e+14	1.85704e+14
0.2	0.198793	4.1497e+14	4.1497e+14
0.3	0.298189	7.09307e+14	7.09307e+14
0.4	0.397585	1.1023e+15	1.1023e+15
0.5	0.496981	1.64538e+15	1.64538e+15
0.6	0.596378	2.43831e+15	2.4383e+15
0.7	0.695774	3.70883e+15	3.70883e+15
0.8	0.79517	6.10073e+15	6.10072e+15
0.9	0.894566	1.28426e+16	1.28426e+16

Tabela A.4: Validação da variância de w para a fila $M/G/1$ com $\alpha = 1.3$.

ρ		$E[r]$		$E[n]$		$E[w]$	
A	S	A	S	A	S	A	S
0.1	0.099908	14866.5	14818.9	3.17161	3.14638	14398.2	14350.6
0.2	0.199816	32860	32762.1	14.0207	13.9734	32391.7	32293.8
0.3	0.299724	55987.8	55734.9	35.8333	35.6305	55519.5	55266.6
0.4	0.399632	86813.1	86256.3	74.0827	73.5041	86344.8	85788
0.5	0.49954	129946	128796	138.613	137.289	129478	128328
0.6	0.599448	194595	192366	249.089	246.076	194127	191898
0.7	0.699356	302213	299584	451.318	447.3	301744	299116
0.8	0.799264	516954	514164	882.295	877.526	516486	513696
0.9	0.899172	1.15726e+06	1.13569e+06	2222	2180.53	1.15679e+06	1.13522e+06

Tabela A.5: Validação da média de r, n e w para a fila $M/G/1$ com $\alpha = 1.5$.

ρ		$Var[w]$	
A	S	A	S
0.1	0.099908	9.86351e+11	9.86351e+11
0.2	0.199816	2.2152e+12	2.2152e+12
0.3	0.299724	3.78961e+12	3.78961e+12
0.4	0.399632	5.87357e+12	5.87357e+12
0.5	0.49954	8.76621e+12	8.76621e+12
0.6	0.599448	1.30582e+13	1.30581e+13
0.7	0.699356	2.02462e+13	2.02462e+13
0.8	0.799264	3.48491e+13	3.48491e+13
0.9	0.899172	7.73135e+13	7.73135e+13

Tabela A.6: Validação da variância de w para a fila $M/G/1$ com $\alpha = 1.5$.

ρ		$E[r]$		$E[n]$		$Var[r]$	
A	S	A	S	A	S	A	S
0.1	0.100757	485.244	485.243	0.1043	0.0351746	235448	235440
0.2	0.201488	581.087	581.086	0.249768	0.239643	337616	337603
0.3	0.302222	774.752	774.752	0.499502	0.652805	600107	600090
0.4	0.402954	1134.3	1134.3	0.975063	1.41977	1286081	1.28606e+06
0.5	0.503688	1848.18	1848.19	1.98589	2.94266	3413771	3.41377e+06
0.6	0.604423	3475.71	3475.71	4.4816	6.4147	12078931	1.20788e+07
0.7	0.705159	8138.3	8138.3	12.2425	16.3617	66272331	6.62722e+07
0.8	0.805894	28280.6	28280.5	48.6197	59.3335	800900731	8.009e+08
0.9	0.906624	256584	256584	496.255	546.406	6.37311e+10	6.37311e+10

Tabela A.7: Validação da média de r e n, e da variância de r para a fila $G/M/1$ com $\alpha = 1.3$.

ρ		$E[r]$		$E[n]$		$Var[r]$	
A	S	A	S	A	S	A	S
0.1	0.100097	473.578	473.578	0.101126	0.010276	224268	224264
0.2	0.200194	517.926	517.926	0.22119	0.104899	268216	268203
0.3	0.300292	612.765	612.765	0.392541	0.307226	375429	375414
0.4	0.400389	775.103	775.103	0.662045	0.653554	600634	600614
0.5	0.500485	1053.21	1053.22	1.1245	1.24683	1108722	1.10869e+06
0.6	0.600582	1568.43	1568.43	2.00949	2.34586	2458491	2.45848e+06
0.7	0.700678	2675.89	2675.89	3.99977	4.70855	7155851	7.1558e+06
0.8	0.800775	5805.88	5805.88	9.91805	11.3859	33708531	3.37084e+07
0.9	0.900873	22482.9	22482.9	43.2079	46.9647	505538731	5.05538e+08

Tabela A.8: Validação da média de r e n, e da variância de r para a fila $G/M/1$ com $\alpha = 1.5$.

ρ		$E[r]$		$E[n]$		$Var[r]$	
A	S	A	S	A	S	A	S
0.1	0.100143	376110	376109	80.8419	77.675	1.86018e+14	1.86019e+14
0.2	0.200261	822856	822856	353.688	346.476	4.04298e+14	4.04298e+14
0.3	0.300381	1.43639e+06	1.43638e+06	926.069	935.874	6.94416e+14	6.94416e+14
0.4	0.400499	2.27981e+06	2.27981e+06	1959.76	2008.51	1.09857e+15	1.09857e+15
0.5	0.50062	3.46816e+06	3.46815e+06	3726.55	3817.39	1.67607e+15	1.67607e+15
0.6	0.600741	5.23017e+06	5.23016e+06	6743.79	6891.02	2.54018e+15	2.54018e+15
0.7	0.700863	7.93367e+06	7.93366e+06	11934.6	12004.9	3.87788e+15	3.87788e+15
0.8	0.800984	1.32738e+07	1.32738e+07	22820.3	22818.1	6.22785e+15	6.22786e+15
0.9	0.901101	2.98783e+07	2.98783e+07	57787.1	57787.9	1.3302e+16	1.3302e+16

Tabela A.9: Validação da média de r e n, e da variância de r para a fila $G/G/1$ com $\alpha = 1.3$.

ρ		$E[r]$		$E[n]$		$Var[r]$	
A	S	A	S	A	S	A	S
0.1	0.0999996	15136.4	15136.4	3.23216	3.21229	1.0115e+12	1.0115e+12
0.2	0.199999	33125.4	33125.4	14.1468	14.0549	2.24267e+12	2.24267e+12
0.3	0.299999	55704.6	55704.6	35.6847	35.1767	3.77882e+12	3.77882e+12
0.4	0.399998	85393.6	85393.6	72.938	71.629	5.78497e+12	5.78497e+12
0.5	0.499997	127021	127022	135.618	134.101	8.56253e+12	8.56254e+12
0.6	0.599996	189358	189358	242.607	240.885	1.26254e+13	1.26254e+13
0.7	0.699995	297867	297868	445.236	445.086	1.95915e+13	1.95915e+13
0.8	0.799994	521320	521321	890.561	894.548	3.42234e+13	3.42235e+13
0.9	0.899995	1.19888e+06	1.19888e+06	2304.02	2318.27	7.93679e+13	7.93679e+13

Tabela A.10: Validação da média de r e n, e da variância de r para a fila $G/G/1$ com $\alpha = 1.5$.

APÊNDICE B

PARÂMETROS DAS DISTRIBUIÇÕES EXPONENCIAL E PARETO UTILIZADOS NAS SIMULAÇÕES

As tabelas B.1 até B.5 listam, respectivamente, os parâmetros utilizados nas simulações das filas $M/M/1$, $M/G/1$, $G/M/1$ e $G/G/1$. Abaixo é feita a relação dos símbolos e nomes utilizados nas tabelas com os parâmetros do simulador:

$\lambda =$ `--exp-arrv-rate`.

$\mu =$ `--exp-serv-rate`.

$\alpha =$ `--prt-arrv-alpha` ou `--prt-serv-alpha`.

$min =$ `--prt-arrv-min` ou `--prt-serv-min`.

$max =$ `--prt-arrv-max` ou `--prt-serv-max`.

Os parâmetros α, min e max se relacionam com os parâmetros do simulador de acordo com a posição do G na nomenclatura do tipo de fila ($A/S/1$), isto é, o G na primeira posição refere-se à distribuição do tempo entre chegadas, enquanto o G na segunda posição refere-se à distribuição do tempo de serviço.

A intensidade do tráfego (ρ) é incluída nas tabelas apenas para referência, significando que uma simulação executada com dado conjunto de parâmetros irá produzir um valor aproximado à carga indicada pela primeira coluna da tabela.

$M/M/1$		
ρ	λ	μ
0.1	0.00021333	0.0021333
0.2	0.00042666	0.0021333
0.3	0.00063999	0.0021333
0.4	0.00085332	0.0021333
0.5	0.00106665	0.0021333
0.6	0.00127998	0.0021333
0.7	0.00149331	0.0021333
0.8	0.00170664	0.0021333
0.9	0.00191997	0.0021333

Tabela B.1: Parâmetros utilizados na simulação da fila $M/M/1$.

$M/G/1$									
ρ	λ	α							
		1.3		1.5		1.7		1.9	
		min	max	min	max	min	max	min	max
0.1	0.00021333	108.146	1e+11	156.21	1e+11	192.966	1e+11	221.983	1e+11
0.2	0.00042666	108.146	1e+11	156.21	1e+11	192.966	1e+11	221.983	1e+11
0.3	0.00063999	108.146	1e+11	156.21	1e+11	192.966	1e+11	221.983	1e+11
0.4	0.00085332	108.146	1e+11	156.21	1e+11	192.966	1e+11	221.983	1e+11
0.5	0.00106665	108.146	1e+11	156.21	1e+11	192.966	1e+11	221.983	1e+11
0.6	0.00127998	108.146	1e+11	156.21	1e+11	192.966	1e+11	221.983	1e+11
0.7	0.00149331	108.146	1e+11	156.21	1e+11	192.966	1e+11	221.983	1e+11
0.8	0.00170664	108.146	1e+11	156.21	1e+11	192.966	1e+11	221.983	1e+11
0.9	0.00191997	108.146	1e+11	156.21	1e+11	192.966	1e+11	221.983	1e+11

Tabela B.2: Parâmetros utilizados na simulação da fila $M/G/1$.

$G/M/1$									
ρ	μ	α							
		1.3		1.5		1.7		1.9	
		min	max	min	max	min	max	min	max
0.1	0.0021333	1081.7	1e+11	1562.45	1e+11	1930.09	1e+11	2220.33	1e+11
0.2	0.0021333	540.849	1e+11	781.227	1e+11	965.045	1e+11	1110.16	1e+11
0.3	0.0021333	360.565	1e+11	520.817	1e+11	643.362	1e+11	740.108	1e+11
0.4	0.0021333	270.425	1e+11	390.613	1e+11	482.522	1e+11	555.082	1e+11
0.5	0.0021333	216.34	1e+11	312.491	1e+11	386.018	1e+11	444.066	1e+11
0.6	0.0021333	180.283	1e+11	260.409	1e+11	321.681	1e+11	370.054	1e+11
0.7	0.0021333	154.528	1e+11	223.208	1e+11	275.727	1e+11	317.19	1e+11
0.8	0.0021333	135.212	1e+11	195.307	1e+11	241.261	1e+11	277.541	1e+11
0.9	0.0021333	120.189	1e+11	173.606	1e+11	214.455	1e+11	246.703	1e+11

Tabela B.3: Parâmetros utilizados na simulação da fila $G/M/1$.

$G/G/1$ – Tempo entre Chegadas								
ρ	α							
	1.3		1.5		1.7		1.9	
	min	max	min	max	min	max	min	max
0.1	1081.7	1e+11	1562.45	1e+11	1930.09	1e+11	2220.33	1e+11
0.2	540.849	1e+11	781.227	1e+11	965.045	1e+11	1110.16	1e+11
0.3	360.565	1e+11	520.817	1e+11	643.362	1e+11	740.108	1e+11
0.4	270.425	1e+11	390.613	1e+11	482.522	1e+11	555.082	1e+11
0.5	216.34	1e+11	312.491	1e+11	386.018	1e+11	444.066	1e+11
0.6	180.283	1e+11	260.409	1e+11	321.681	1e+11	370.054	1e+11
0.7	154.528	1e+11	223.208	1e+11	275.727	1e+11	317.19	1e+11
0.8	135.212	1e+11	195.307	1e+11	241.261	1e+11	277.541	1e+11
0.9	120.189	1e+11	173.606	1e+11	214.455	1e+11	246.703	1e+11

Tabela B.4: Parâmetros utilizados na simulação da fila $G/G/1$ relativos à distribuição atribuída ao tempo entre chegadas (primeiro G).

$G/G/1$ – Tempo de Serviço								
ρ	α							
	1.3		1.5		1.7		1.9	
	min	max	min	max	min	max	min	max
0.1	108.146	1e+11	156.21	1e+11	192.966	1e+11	221.983	1e+11
0.2	108.146	1e+11	156.21	1e+11	192.966	1e+11	221.983	1e+11
0.3	108.146	1e+11	156.21	1e+11	192.966	1e+11	221.983	1e+11
0.4	108.146	1e+11	156.21	1e+11	192.966	1e+11	221.983	1e+11
0.5	108.146	1e+11	156.21	1e+11	192.966	1e+11	221.983	1e+11
0.6	108.146	1e+11	156.21	1e+11	192.966	1e+11	221.983	1e+11
0.7	108.146	1e+11	156.21	1e+11	192.966	1e+11	221.983	1e+11
0.8	108.146	1e+11	156.21	1e+11	192.966	1e+11	221.983	1e+11
0.9	108.146	1e+11	156.21	1e+11	192.966	1e+11	221.983	1e+11

Tabela B.5: Parâmetros utilizados na simulação da fila $G/G/1$ relativos à distribuição atribuída ao tempo de serviço (segundo G).